Causal Model Abstraction & Grounding via Category Theory

Taco Cohen
Principal Engineer
Qualcomm Technologies Netherlands B.V.

tacos@qti.qualcomm.com

Talk Outline

- Motivation
- Introduction to causal models
  - SCMs
  - Intervention vs conditioning
  - Causal learning problems
  - The grounding problem
- Causal model morphisms
  - Functorial semantics for causal models & equivariant networks
  - Various notions of morphism
- Causal Representation Learning
- Grounding causality in dynamical systems & MDPs
  - What's missing in current foundations
  - Interventions as transformations

CausaLand

- Contributions:
  - Major contributions to statistical methodology for empirical research
  - So far, limited impact on AI

- Strong claims:
  - “Deep Learning is just Curve Fitting”
  - “Deep Learning is stuck on rung-1 of the ladder of causation”
  - “Deep learning has succeeded primarily by showing that certain questions or tasks we thought were difficult are in fact not. It has not addressed the truly difficult questions that continue to prevent us from achieving human like AI”
    - Judea Pearl, The Book of Why

DeepLearnistan

- Contributions:
  - Vision: object recognition, segmentation, SLAM, image generation, …
  - Speech recognition & generation, text-to-speech
  - Language modelling, translation, QA, …
  - Game playing, long-term planning in world models
  - Protein folding, drug design, materials design
  - Program synthesis
  - Robot control
  - …

- Strong claims:
  - “Radiologists will be out of a job in 5 years”
  - “More profound than electricity & fire”
  - “AI will probably most likely lead to the end of the world, but in the meantime, there'll be great companies”
  - “Causal reasoning will emerge from training on all text on the internet”

Causal Models

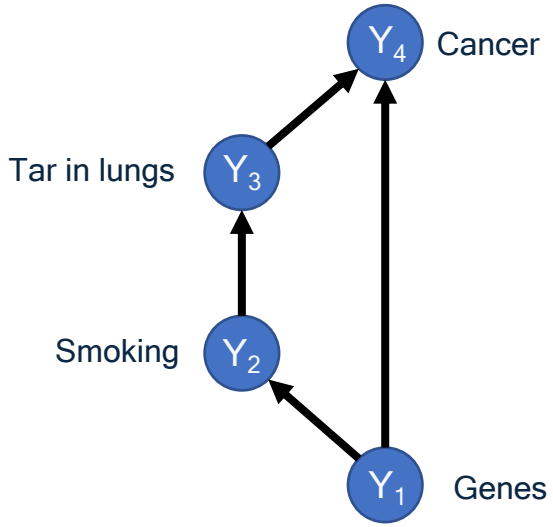
A brief introduction to the classical theory

Structural Causal Models

The classical definition

Definition: an SCM consists of:

- A finite DAG G
- A finite set of endogenous variables $Y = \{Y_i\}$ (one for each node in G)
- A finite set of exogenous variables $E = \{E_i\}$ (one for each node in G)
- Independent noise distributions for the exogenous variables $p(E_i)$
- Information on which endogenous variables are observed vs latent
- For each endogenous variable, a default causal mechanism $f_i : E_i \times Pa_i \rightarrow Y_i$
  - Where Pa_i are the parents of Y_i in G
- Sometimes: a set I of allowed interventions do(variable=value)
  - May also include mechanism changes



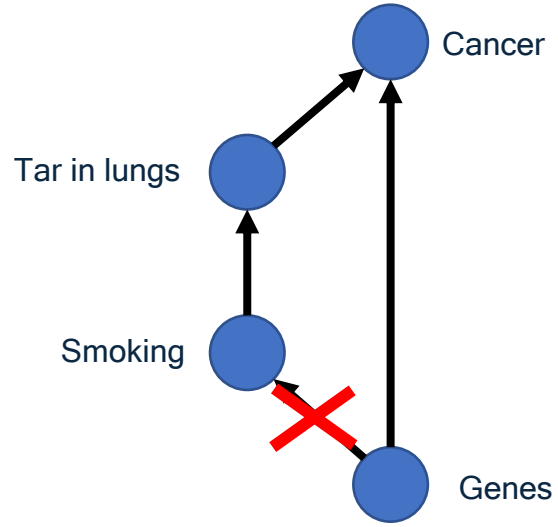
Causal Bayes Nets

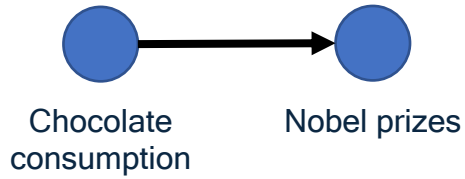
Definition: a Causal Bayesian Network is a Bayesian Network with the word “causal” prepended

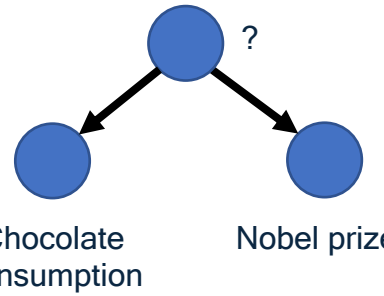
- BN & CBN are the same as mathematical objects
- CBNs & SCMs are “causal”: they are expected to correctly predict the effect of interventions
  - → If we want to actually define “causal” mathematically, we need to model more than just the CBN / SCM (grounding)
- An SCM induces a CBN:
  - A mechanism $f_i : E_i \times Pa_i \rightarrow Y_i$ and noise distribution $p(E_i)$ induce a conditional $p(Y_i \mid Pa_i)$
- CBNs cannot be used to reason about counterfactuals, whereas SCMs can

Intervention vs Conditioning

- Intervention do(Y_i = y): replace mechanism for Y_i by a constant y
  - Change the data generating process (must be done before the process)
- Conditioning: filter out data points that don't satisfy a condition
  - Conditioning happens after the data generating process
- Intervention != conditioning
  - For decision making rather than prediction, one is interested in the effect of interventions
- Intervention on one mechanism leaves invariant all other mechanisms
  - This is a different kind of invariance than what we study in geometric deep learning
  - Later we will argue that this is the only objective property that makes a map "causal"





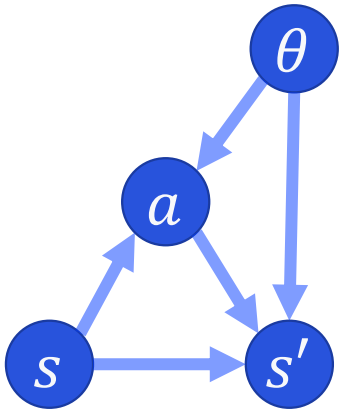


Deconfounded Imitation Learning

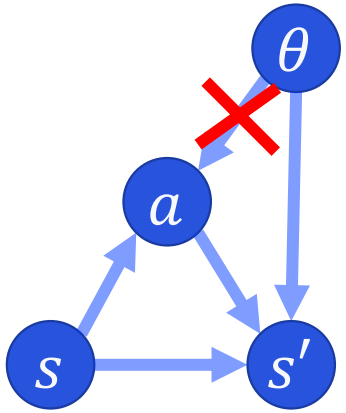
Expert

Imitator

Imitation learning can fail when the imitator does not observe the world the same way the expert does.



Expert



Imitator

This partial observability gives rise to hidden confounders in the causal graph.

R. Vuorio, J. Brehmer, H. Ackermann, D. Dijkman, T. Cohen, P. de Haan, *Deconfounded Imitation Learning*, 2022

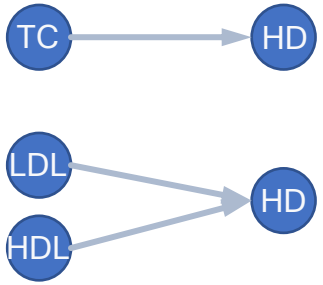
Causal Learning Problems

- Causal Discovery: infer causal graph from data, using conditional independencies
- Causal Inference: given a causal graph, infer mechanisms and causal effects from data
- Causal Representation Learning: learning to map sensory observations to latent causal variables with associated causal model
- Causal Intervention Learning: learning a behaviour policy to actually perform an intervention

Causal Models are Abstractions

Simplified models to facilitate simple reasoning

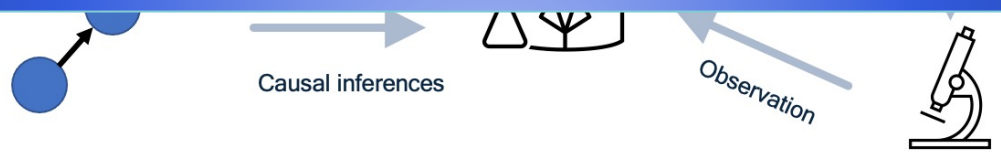
- Causal variables are usually far removed from sensory inputs
- The notion of “direct cause” is never fundamental
- The real-world implementation of an intervention is only vaguely defined



Intervention on Total Cholesterol has an ambiguous effect on Heart Disease

Causal Models are abstractions without grounding

- This is the primary cause of confusion in literature
- Grounding is essential for AI



Addressing the Grounding Problem

1. Causal Model Abstraction
   - Define a notion of causal model morphism
   - Grounding one causal model in another
2. Causal Representation Learning
   - Grounding variables only
   - Based on identifiability up to isomorphism
3. Grounded Theory of Causation
   - Grounds both variables & interventions
   - Based on admissibility; invariance

Causal Model Morphisms & Categorical Causal Models

Literature on Causal Model Morphisms / Abstraction

- Rubenstein, P.K., Weichwald, S., Bongers, S., Mooij, J.M., Janzing, D., Grosse-Wentrup, M., Scholkopf, B.: Causal Consistency of Structural Equation Models, UAI 2017
- Kissinger, A., & Uijlen, S. A categorical semantics for causal structure. 32nd Annual ACM/IEEE Symposium on Logic in Computer Science (LICS), 2017
- Beckers, S. and Halpern, J.: Abstracting Causal Models, AAAI 2019
- Beckers, S., Eberhardt, F., and Halpern, J.: Approximate Causal Abstraction, UAI 2019
- Brendan Fong. Causal Theories: A Categorical Perspective on Bayesian Networks. MSc thesis, 2012.
- Bart Jacobs, Aleks Kissinger, and Fabio Zanasi. Causal Inference by String Diagram Surgery. In: Foundations of Software Science and Computation Structures (2019),
- Eigil F Rischel. The category theory of causal models. MSc thesis, 2020.
- Eigil F Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. arXiv preprint arXiv:2103.15758, 2021
- Jun Otsuka, Hayato Saigo, On the equivalence of causal models: a category-theoretic approach, CCIeaR 2022
- Zennaro, F. M. Abstraction between Structural Causal Models: A Review of Definitions and Properties. 2022
- Yimu Yin, Jiji Zhang, Markov Categories, Causal Theories, and the Do-Calculus, 2022

Exact τ-transformations

Rubenstein et al., Causal Consistency of Structural Equation Models, UAI 2017

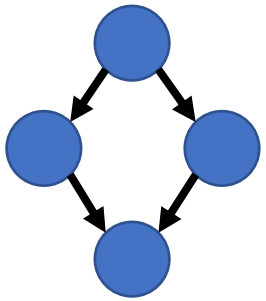
- Equip SCM with a set of allowed interventions I
  - This set forms a poset, reflecting the compositional structure of interventions
    - e.g. $do(Y_i = y_i) \leq_X do(Y_i = y_i, Y_j = y_j)$
    - Incompatible interventions are not ordered
- Exact transformation of SCMs consists of:
  - A map between joint state spaces $\tau : X \rightarrow Y$
  - A map between intervention sets $\omega : I_X \rightarrow I_Y$ (order-preserving and surjective)
  - Such that:

$$\tau\left(P_X^{do(i)}\right) = P_Y^{do(\omega(i))} \quad \forall i \in I_X$$

- Limitation: does not separate syntax & semantics

Syntax & Semantics

Classic theory



$$\longrightarrow \quad p(x) = \prod_i p(x_i \,|\, \mathrm{Pa}_i)$$

Graph G

Conditional distributions

Categorical theory

$$\mathrm{Syn}_G \qquad \longrightarrow \qquad \mathrm{Stoch}$$

Syntax Category

Functor

Semantics Category

Functorial Semantics

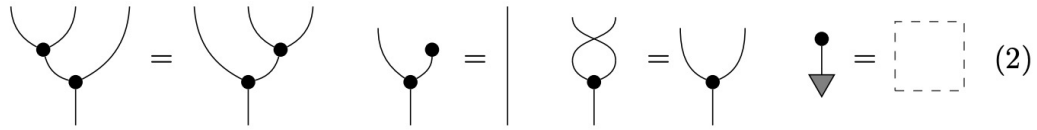
F. W. Lawvere, *Functorial Semantics of Algebraic Theories*, Ph.D. thesis, Columbia University, 1963

- Categorical approach to universal algebra
  - Lawvere theory L: syntax category with finite products encoding the generic idea of a group, ring, associative algebra, …
  - Model of the theory in C: a functor F : L → C (e.g. a particular group, ring, etc.); semantics
  - Morphism between models: a natural transformation between functors
- Key example in Geometric Deep Learning:
  - Group G as a one-object category where every morphism is an isomorphism
  - Functor to set / vect: group representation
  - Natural transformation: equivariant map
- Same approach should work for causal theories & models

Categorical Causal Models

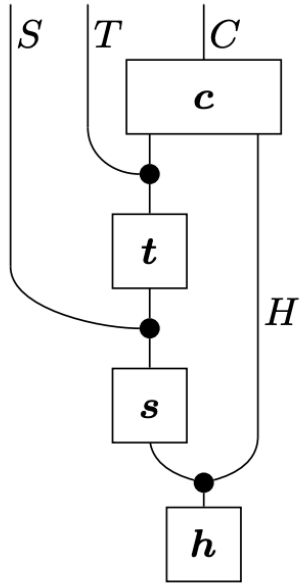
Brendan Fong. Causal Theories: A Categorical Perspective on Bayesian Networks. MSc thesis, 2012
Bart Jacobs, et al. Causal Inference by String Diagram Surgery. FSSCS, 2019

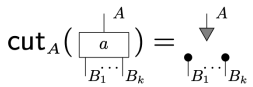
Definition 2.1. *A* CDU category *(for* **c***opy,* **d***iscard,* **u***niform) is a symmetric monoidal category* $(\mathsf{C}, \otimes, I)$ *where each object* A *has a copy map* $: A \to A \otimes A$*, a discarding map* $: A \to I$*, and a uniform state* $: I \to A$ *satisfying the following equations:*



(2)

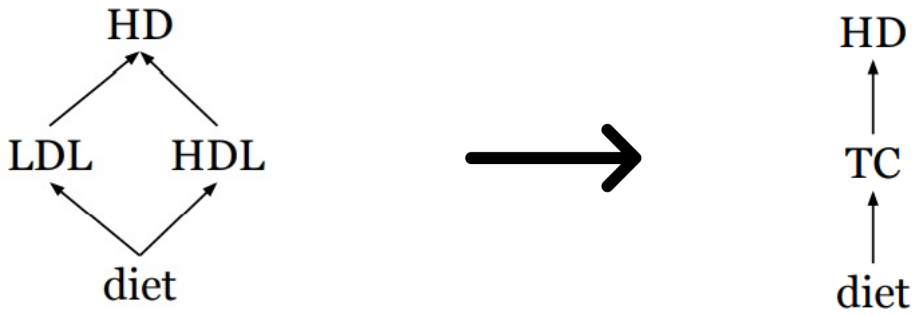
CDU functors *are symmetric monoidal functors between CDU categories preserving copy maps, discard maps and uniform states.*



- Causal Theory: “free CDU category on a DAG” Syn_G
- Causal Model: functor F : Syn_G → Stoch
- Intervention: functor Cut_A : Syn_G → Syn_G 
  - Limitation: cannot define intervention do(A=a) this way, because this is not syntactical!

Causal model morphisms as natural transformations

Jun Otsuka, Hayato Saigo, *On the equivalence of causal models: a category-theoretic approach*, CCleaR 2022

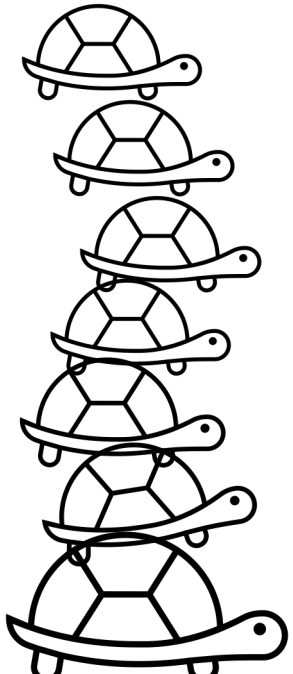


Definition 1 (Φ-abstraction) *Let $\phi : G \to H$ be a graph homomorphism; $\Phi : \mathsf{Syn}_G \to \mathsf{Syn}_H$ the induced functor; and F_G, F_H functors (causal models) to* **Stoch** *from Syn_G and Syn_H, respectively. We say that F_H is a* Φ-abstraction *of F_G if there is a natural transformation $\alpha : F_G \Rightarrow F_H\Phi$.*

$$\begin{array}{ccc} F_G(X) & \xrightarrow{F_G(f)} & F_G(Y) \\ {\scriptstyle \alpha_X}\downarrow & & \downarrow{\scriptstyle \alpha_Y} \\ F_H\Phi(X) & \xrightarrow{F_H\Phi(f)} & F_H\Phi(Y) \end{array}$$

Can Causal Model Abstraction solve the Grounding Problem?

- No
- Very valuable & exciting research direction, but:
- But only grounds causal models in other causal models
- Does not explain how causal models are grounded in physics or agent-centric frameworks (MDPs)

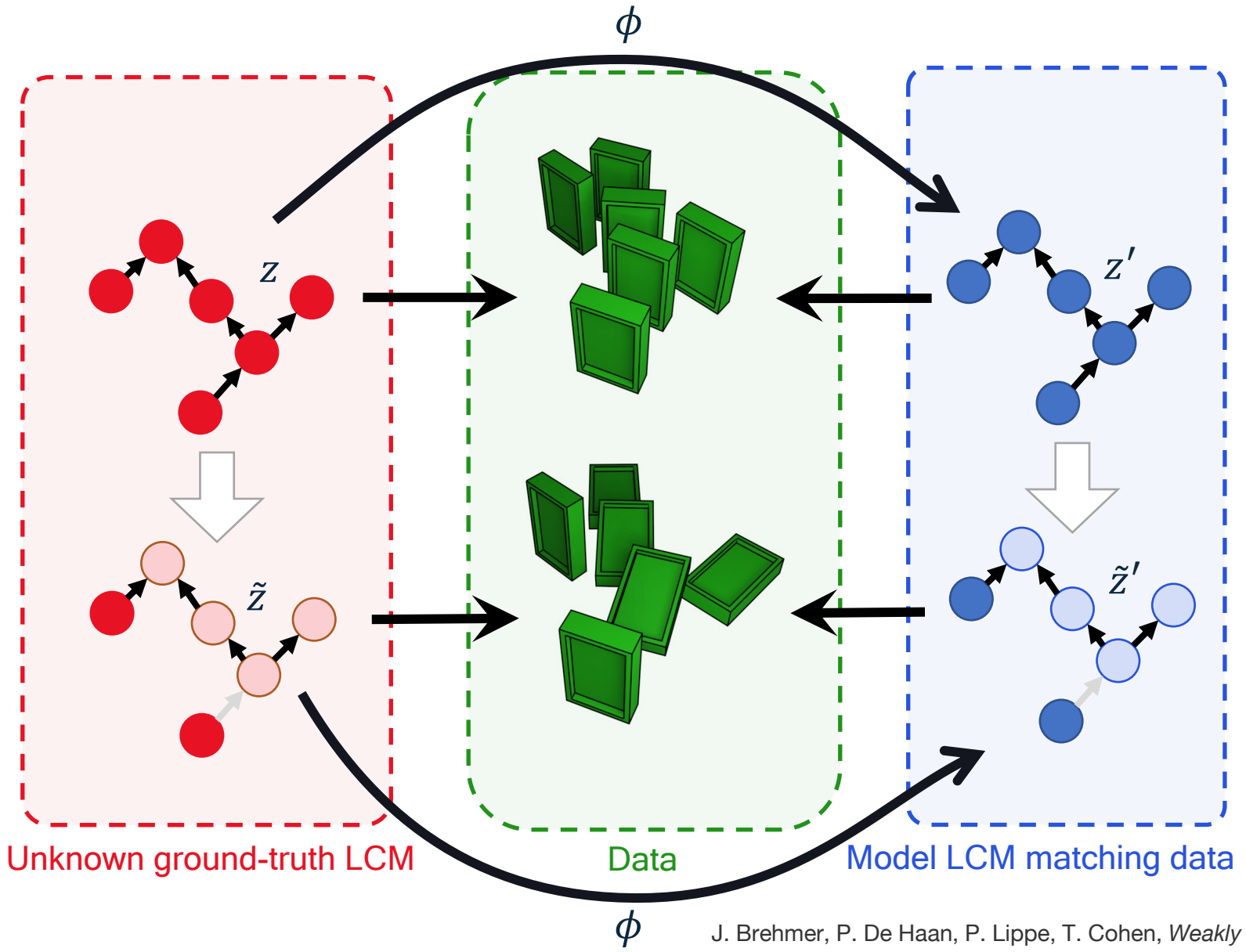


Causal Representation Learning

Literature on Causal Representation Learning

- Background:
  - Spirtes, P. Variable definition and causal inference. ICLMPS 2007
  - Eberhardt, F. Green and grue causal variables. Synthese, 2016
  - Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. Towards Causal Representation Learning. 2021
- Disentangling
  - Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. Neural Networks, 2000.
  - Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. ICML 2020
- Causal representation learning
  - Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. UAI 2014.
  - Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? NeurIPS 2021.
  - Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. NeurIPS 2021.
  - Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. CVPR 2021.
  - Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal Identifiability from Temporal Intervened Sequences. ICML 2022
  - J. Brehmer, P. De Haan, P. Lippe, T. Cohen, Weakly supervised causal representation learning, NeurIPS 2022

Causal Representation Learning



- Latent causal model should be identifiable up to isomorphism ϕ
- We need an appropriate notion of (iso) morphism

J. Brehmer, P. De Haan, P. Lippe, T. Cohen, *Weakly supervised causal representation learning*, NeurIPS 2022

A Critique of Pure CRL

- Mainstream approach to CRL:
  1. Make assumptions about data generating process,
  2. Prove identifiability (if the data was generated by a process of this kind, we can recover it from the data up to isomorphism)
  3. Train model by maximum likelihood and show that latents+mechanisms can actually be recovered in simulated environment
- Issues with this approach (for AI applications):
  1. Assumptions are often unrealistic
  2. We can represent systems at different levels of abstraction, so it is not clear why causal variables should be uniquely identifiable
  3. Many domains cannot be described by a single DAG, but CRL methods currently assume this
  4. DAGs are hard to learn (discrete)
  5. Causal representation + causal model is not enough: autonomous agent needs policies implementing interventions (skill learning)
  6. No guarantee that learned representations are task-relevant
  7. Not clear how to demonstrate an advantage of causal representations
- Alternative approach: from identifiability to admissibility
  - Instead of finding “the true causal variables + mechanisms”, find a set of (task-relevant) variables + mechanisms that are “causal”
  - We hope to enable this approach via our invariance-based definition of “causal mechanism”

Grounding Causal Models

T. Cohen, Towards a Grounded Theory of Causation for Embodied AI, UAI WS 2022

Causation: It Still Confounds Us

A fictional dialogue between student and teacher

Okay so basically Hume said in 1740 that causation is when one thing goes after another, like, all the time no matter what?

Oh yeah they were pretty confused back then! Fortunately we now have a mathematical definition of causation in the form of CBNs!

Indeed! I read in your lecture notes that a CBN is a Bayesian Network with the *causal edges assumption.*

Exactly. The causal edges assumption says that in a causal DAG, every parent is a direct cause of all their children.

Makes sense. And what was the definition of direct cause again?

See definiton 1.0.1: A variable X is said to be a cause of a variable Y if Y can change in response to changes in X

Okay, so Y is like a function of X? Or like a conditional distribution P(Y | X) in the stochastic case?

Uh yeah but no. X and Y should be related by a structural equation / assignment, which means there is a causal mechanism from X to Y.
Remember that functions and conditional distributions are epistemic, whereas mechanisms are ontological.

What do you mean? What is a mechanism and how is it different from a function?

Well, unlike functions and conditionals, causal mechanisms have no mutual algorithmic information.
And mechanisms are modular, stable, and invariant. They are aligned with the entropy gradient of the universe. Did you read chapter 7?

I did read it! But it doesn't say what it is that mechanisms should be invariant to! It's not like in geometric deep learning, where they clearly state what the symmetry group and group action are, to which the function should be invariant.

A mechanism for Y should be invariant to changes in X! We model these changhes as surgical interventions.
And don't get me started on GDL. Those people are forever stuck on rung-1 of Pearl's ladder of Causation.

You're right, and I know we shouldn't associate with folks who resist the causal revolution. But the mathematics is just so *crisp!*
Anyway, could you remind me of the definition of surgical intervention?

A surgical intervention is a change that causes one mechanism to be replaced by another, while leaving all other mechanisms invariant.
I have to go now. Next time please consult the lecture notes before asking me to repeat basic definitions.

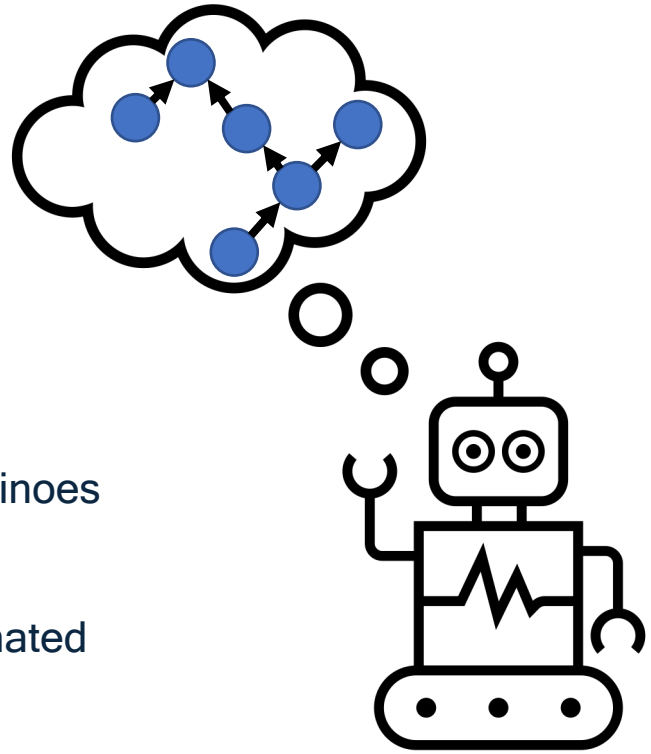
Guiding principles & goals

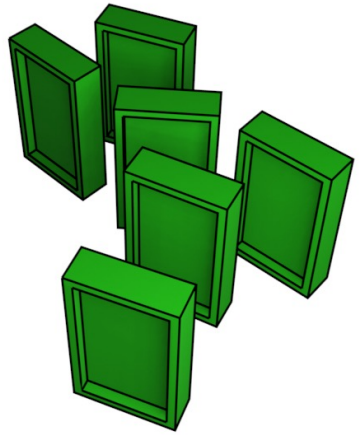
Towards a Grounded Theory of Causation

- Completely Mathematical
  - Evaluating the truth of a claim should never require intuition or common sense
    - At the same time, it should be completely clear what each mathematical object is supposed to model
  - Should fit within the broader mathematical landscape; causal models are just another mathematical gadget
- Embarassingly Simple
  - Trivial things should be trivially trivial
- Admissibility over Identifiability
- Actions are primary
  - Classical theory: define variables & mechanisms, then define interventions in terms of them
  - Our theory: define actions & variables as maps, give a criterion for when action behaves as intervention on a mechanism
- Variables are not real
  - Key lesson from physics: any choice of variables is equally valid
- Take (open) dynamical systems as ground truth
  - Covers classical physics as well as agent-centric frameworks
    - But don't focus on details like differential equations or details of MDP framework
  - Assume the arrow of time. No need to worry about the entropy, past hypothesis, quantum theory, etc.
    - Arrows are built into category theory

Causal reasoning for autonomous embodied AI

Running example: robo-dominoes





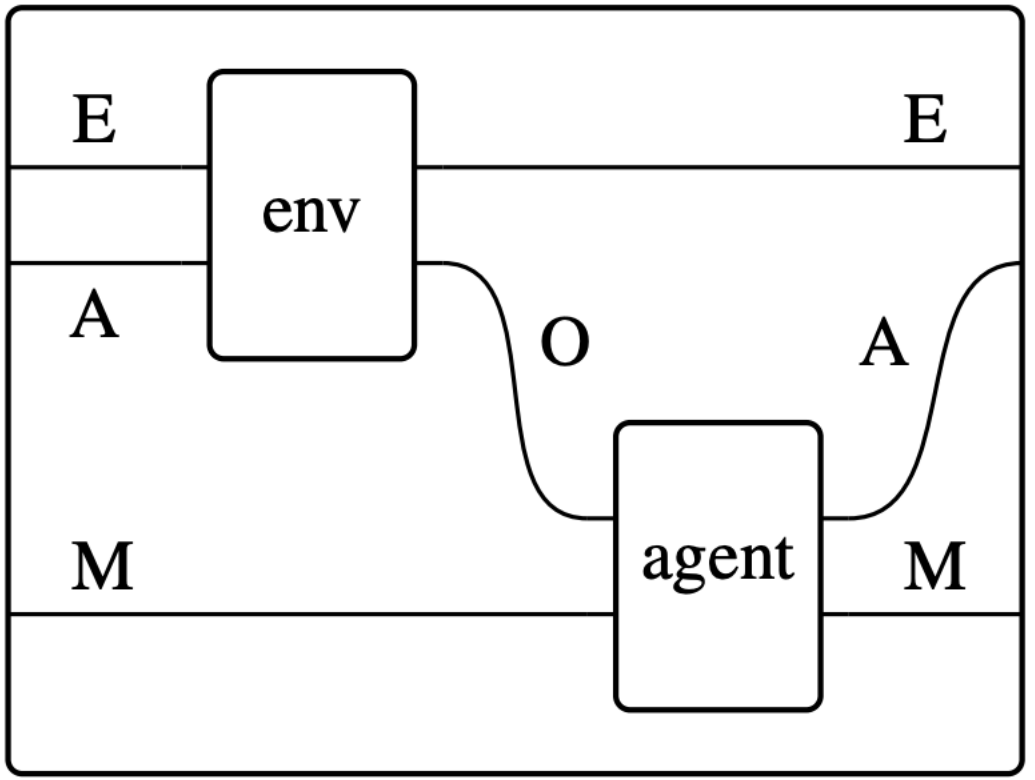
- Actions / Interventions:
  - Picking / placing dominoes
  - Placing a barrier
  - …
- Experiment: push a designated domino forward and wait

The Agent-Environment Model

Markov Decision Process (without reward)

- Assumption: all stochasticity comes from partial observability
  - → MDP can be modelled as deterministic maps env and agent with partially latent input
  - Similar to SCM → CBN, we can make it stochastic as well
- Composing the environment dynamics and agent policy, we obtain a map X → X
  - Where X = E x A x M is the complete state space
- Repeating this for a policy *a* until some termination condition is met, we obtain a map:

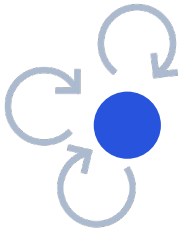
$$\mathrm{do}_X(a) : X \to X$$



One timestep induces an endomap on X = E x A x M

Monoids of Actions & Actions of Monoids

- Monoid:
  - Set M with associative binary operation & identity
  - Equivalent to a one-object category
- Generators & relations:
  - Elementary actions (generators): arrows X → X
  - Relations, like commutativity
- Examples:
  - All endomaps on any object in any category
  - All endomaps of a state space, generated by agent policies / options (and any maps freely generated by these)
  - All physically possible transformations of state space
- Monoid action
  - Define abstract monoid M, and a functor do : M → Set
  - Allows for multiple actions to be defined, e.g.:
    - On state space, implemented by policies $do_{\hat{X}}(a)$
    - On mental model used by the agent $do_X(a)$

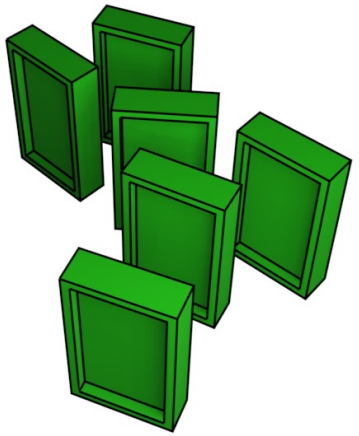


X

The Do-Monoid

"To be is to do"—Socrates.
"To do is to be"—Jean-Paul Sartre.
"Do be do be do"—Frank Sinatra.

- In SCMs, interventions are not ordered in time!
  - Q: How can we encode interventions in a monoid?
- The Do-Monoid on sets / variables Y_1, …, Y_n is the monoid generated by:
  - Generators: $(Y_i = y)$ for each Y_i and each y in Y_i
  - Subject to the relations:
    - $(Y_i = x)\ (Y_j = y) = (Y_j = y)\ (Y_i = x)$ (commutativity for i != j)
    - $(Y_i = y)\ (Y_i = y') = (Y_i = y)$ (annihilation)
  - Warning: mixing syntax & semantics…
- Captures abstractly the idea of "setting variables to values"
- Variables are assumed to be independently controllable[1]

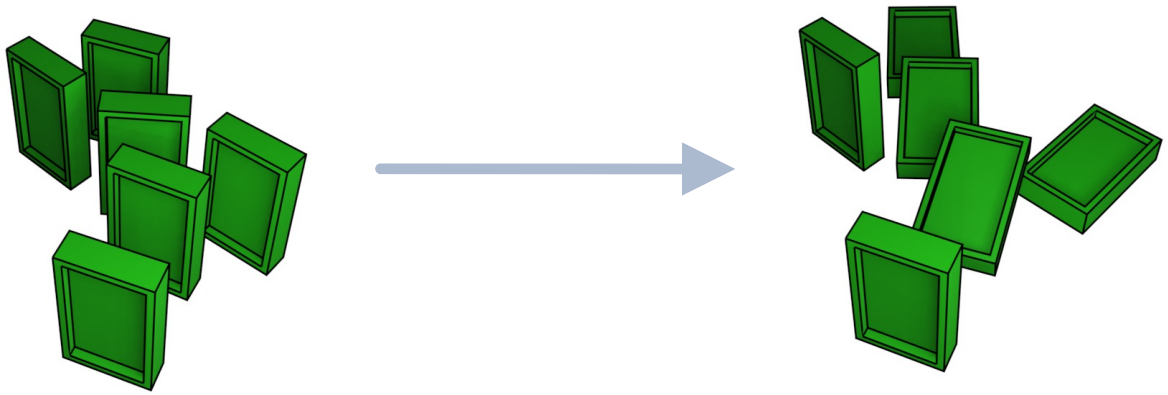


Picking / placing different dominos is commutative.
Actions are primary: these actions are possible, therefore it is reasonable to think of different dominos as different objects.

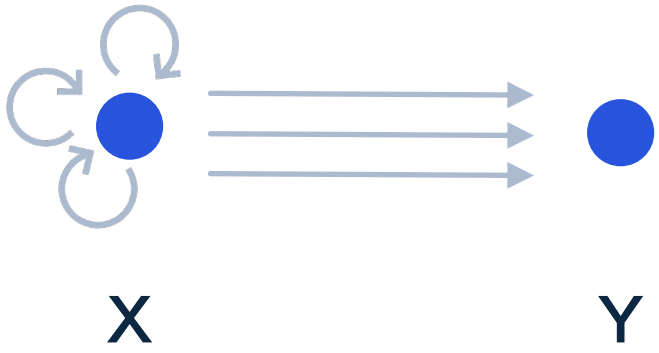
[1] Thomas, V., Pondard, J., Bengio, E., Sarfati, M., Beaudoin, P., Meurs, M.-J., Pineau, J., Precup, D., & Bengio, Y. Independently Controllable Factors. 2017

Experiments, processes, outcomes

- Interventions are compositional:
  - ✅ In the space of mechanisms
  - ❌ Not in the space of causal variables
  - Cannot define a monoid action of do-monoid on space of causal (endogenous) variables!
- ➡️ Need to introduce a process (mapping): $\mathrm{proc}_Y : X \to Y$
  - This represents the experiment of interest
  - For each state in X, we get an outcome in Y
- The outcome of an action is: $\mathrm{outcome}_Y^a = \mathrm{proc}_Y \, \mathrm{do}_X(a) : X \to Y$



Action Theories & Models



- Action Theory:
  - Small category with a state-space object X and outcome-space object Y
  - Actions a : X → X and a process proc : X → Y, and everything generated by these maps subject to chosen relations
  - Note: Many variations possible (e.g., let actions act on Hom(X, Y), use maps A x X → Y, …)
- Action model: functor to (e.g.) Set
- Follows the philosophy of Functorial Semantics, making these yet another mathematical gadget one can study,.and giving us a notion of morphism of action models for free.

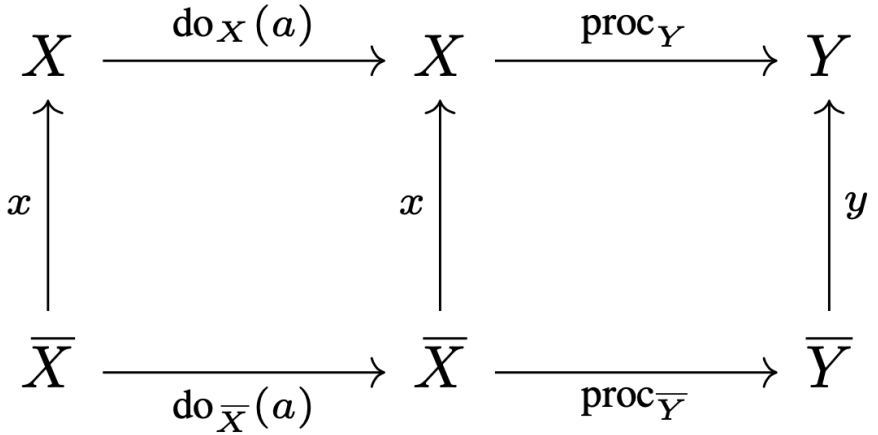
Natural Transformations between Action Models

- Consider two action models:

  $\overline{X}, \overline{Y}$: True underlying system; Do-operators generated by agent policies

  X, Y : Simplified / abstract mental model of the system; Do/effect maps are learned

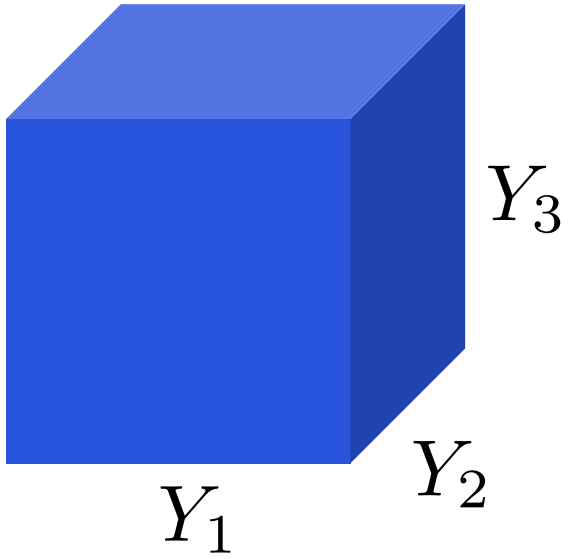
- A *natural transformation* between these models is a pair of maps x, y, satisfying for each a:

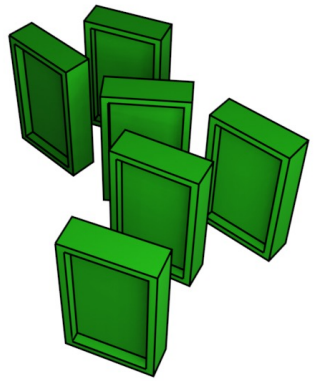


A model is *grounded / veridical* if there is a natural transformation like this

(Causal) variables

$$Y = \prod_i Y_i =$$





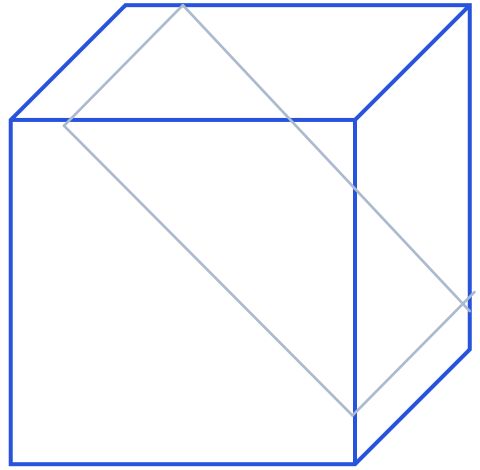
Example: Y_i indicates the state of domino i

Different ways to produce outcome variables:

- Do something in the system, then measure outcome variable $\overline{X} \xrightarrow{\mathrm{do}_{\overline{X}}(a)} \overline{X} \xrightarrow{\mathrm{proc}_{\overline{Y}}} \overline{Y} \xrightarrow{y} Y \xrightarrow{\pi_i} Y_i$
- Do something in the model, then predict outcome variable $X \xrightarrow{\mathrm{do}_X(a)} X \xrightarrow{\mathrm{proc}_Y} Y \xrightarrow{\pi_i} Y_i$

Possible and impossible outcomes

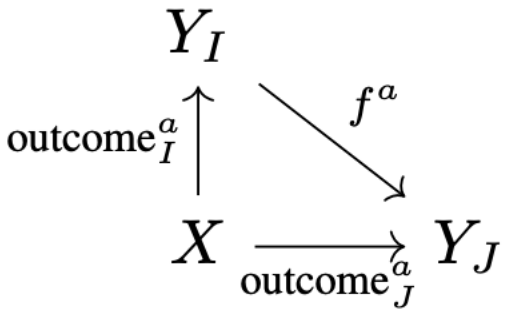
- For “disentangled” choices of variables, there can be “impossible outcomes”
  - All dominos can be in any position, but no two can be in the same position at once
  - Hence, if domino A has fallen forward then adjacent domino B must also have fallen
  - Ideal gas law says that only certain values of pressure, temperature and volume are jointly possible
  - Etc.
- The possible outcomes are the image of $y : \overline{Y} \to Y$
  - Same as the image of proc if the model is accurate
- Can also consider the possible outcomes (image) of any effect map



When there are impossible outcomes, we can learn something about one outcome variable from another, even in the absence of subjective probabilities!

Determination & Effectiveness

- Let a be an action (monoid element) and let I, J be (sets of) variables.
- We say that outcome^a_J is determined by outcome^a_I via $f^a : Y_I \to Y_J$ if the following commutes



$$i.e. \text{outcome}^a_J = f^a \, \text{outcome}^a_I$$

- Note: f^a may or may not be unique!
- Effective actions: outcome determined by 1 = {0} (i.e. no variables; the monoidal unit)
  - In other words, the outcome is a constant map
  - This is a property enjoyed by perfect interventions as modelled in SCMs

Invariant determination, mechanisms & surgical interventions

- Determination is not sufficient to call f a mechanism
  - For example: if we initiate the system to a particular state x, every variable determines every other one
- What we call “mechanism” is a determination relation that is highly robust / invariant
  - Holds not just for one action / effect map effect^a_Y but many
- Determination is always preserved under precomposition (“doing something before”)
- Determination is in general *not* preserved under postcomposition (“doing something after”)

Definition 3.3 (Invariance of Determination)**.** *Let* a, b *be actions,* $\bar{a}_J : Y_I \to Y_J$ *a mapping, and assume that the determination relation* $\text{outcome}^a_J = \bar{a}_J\, \text{outcome}^a_I$ *holds. If* $\text{outcome}^{ba}_J = \bar{a}_J\, \text{outcome}^{ba}_I$ *also holds, we say that* b *leaves the determination via* $\bar{a}_J$ *invariant.*

- Surgical intervention: changes mechanism for its target variable, while leaving all other mechanisms invariant

Action Theories with Invariant (causal) Mechanisms

- Extend an action theory with:
  - Variables Y_i and E_i (both are outcomes of proc)
  - Products
  - Mechanisms f : $Y_I \rightarrow Y_J$
- Encode all invariances we believe in in the theory
- Possible to encode SCMs, but much more as well
- Much work remains to flesh this out, find interesting special cases

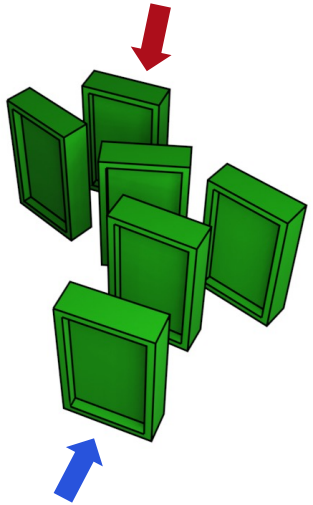
Encoding SCMs in this Framework

- Let (U, E, F) be an SCM, with endogenous variables U, exogenous E, and mechanisms F
- Define state X = E x M, where M is the space of mechanisms, defined as:
  - $M = \text{prod}_i M_i$, where $M_i = U_i + \{f_i\}$
  - The state X should somehow encode the exogenous variables and which mechanism is active; here we hard-code this
- Define outcomes Y = E x U
  - Note: we view both endogenous and exogenous variables as outcomes
- Define $\text{proc}_Y(e, m) = (e, Y_m(e))$, where $Y_m(e)$ is the potential response defined by the SCM in intervention condition m (i.e. solve structural equations given e)
- Define interventions $do(m_i)$ as setting one mechanism variable M_i while leaving others unchanged, and leaving E unchanged as well
  - These actions commute & annihilate as interventions do in SCMs. (i.e. this is the do-monoid on M_i)
- This defines a causal model in our framework that makes equivalent predictions to the SCM
  - The outcome variables $(e, Y_m(e))$ satisfy the functional/determination relations for all active mechanisms
- Is this useful?
  - Probably less intuitive in domains where we have domain expertise, but conceptually clearer & more expressive than SCM

Are SCMs all you need?

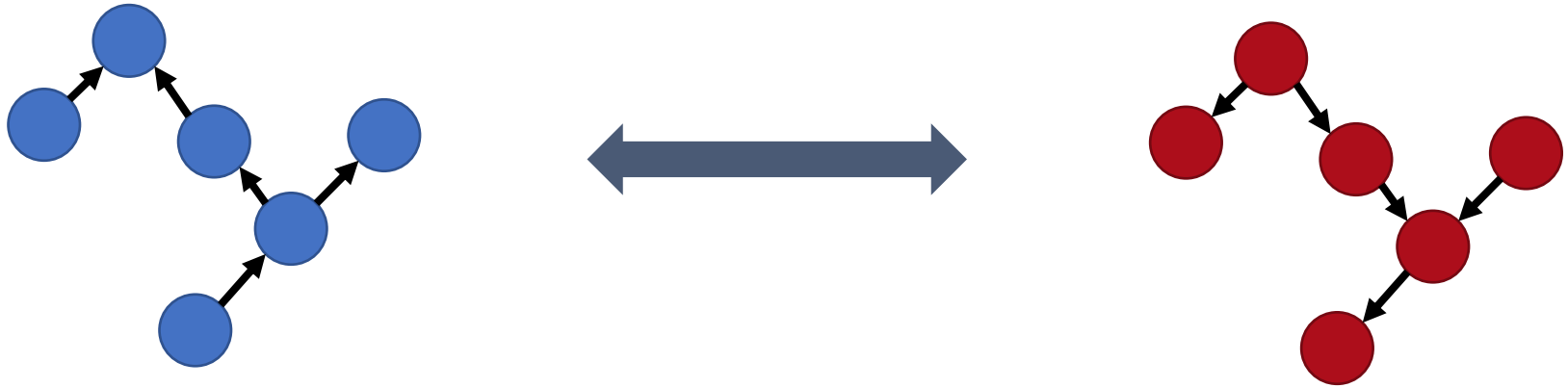
- Some facts about models derived from an SCM in this way:
  - We can dispense with all aspects of the state, except for noise E and mechanisms M
  - All interventions on different variables / mechanisms commute.
  - We can set any variable to any value (set constant mechanism), and do so jointly as well
    - this is not always true in reality. E.g. ideal gas law says only certain combinations of pressure/temperature/volume are possible. Or: physical objects can't occupy same space.
  - Actions leave invariant the exogenous variables
- These are nice properties, but they are just not true for every system
  - New language allows to express such deviations from the SCM ideal
  - E.g. non-commutativity of actions, irreversibility, mechanisms that are invariant to some but not all actions under consideration, are only approximately invariant or invariant only in a certain context, etc.
- Moreover, starting from a forward model proc_Y and then learning about or encouraging some invariant determination relations seems easier than learning a discrete graph
- Starting with actions + proc map also alleviates concerns about existence & uniqueness of solutions in (cyclic) SCMs
  - Of course the true process has a unique outcome, one for every starting state…

Inverting the Causal Direction



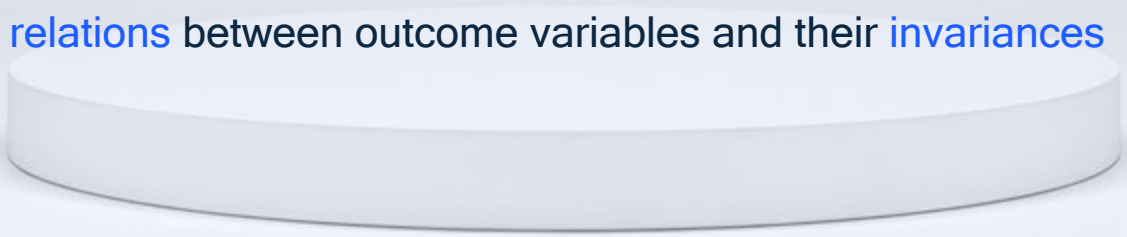
- Actions / Interventions:
  - Picking / placing dominoes
  - Placing a barrier
  - Choosing a domino & direction to push
  - …
- Experiment: push the chosen domino forward and wait

😱 Changing your mind can change the direction of Causation 😱



A Radical Proposition

- If we have a determination relation invariant to all actions in M, we could call it causal
- If we have an action that destroys exactly one determination relation while installing a new one, we could call it a surgical intervention
- But these concepts are not fundamental, as the word “Causal” would suggest:
  - Depends on which actions we consider, the process / experiment, and other context
  - Some actions can reverse causal direction, some can destroy all determination relations or change them arbitrarily
- Therefore it is more precise to speak of:
  - The composition structure of actions
  - Functional or probabilistic relations between outcome variables and their invariances



Conclusion

- New perspective on causal models
- All definitions based on maps and composition
  - Very little mathematical structure used
  - Discrete graphs may emerge implicitly
  - Map-based approach may be more suitable for DL
- Definition of intervention as a transformation of state space (in mental model or actual system)
- Grounds the theory of SCMs in actual behaviours or physically possible transformations
- Definition of mechanism as invariant predictor
- No metaphysical baggage; just assuming “states + time evolution”
- Beginnings of a theoretical foundation for causal representation & intervention skill learning
- Categorical treatment offers many possibilities for the mathematical study of causal models, model abstraction, etc.

Towards a Grounded Theory of Causation for Embodied AI

Taco Cohen[1]

[1]Qualcomm AI Research*

https://arxiv.org/abs/2206.13973

Further reading

- Books:
  - Fong & Spivak, *Seven Sketches in Compositionality: an invitation to applied category theory*
  - Lawvere & Schanuel, *Conceptual Mathematics: a first introduction to categories*
  - Lawvere, *Sets for Mathematics*
  - Goldblatt, *Topoi: the Categorical Analysis of Logic*
- Papers:
  - Gavranovic, https://github.com/bgavran/Category_Theory_Machine_Learning

Thank you

Snapdragon

Follow us on:

For more information, visit us at:

snapdragon.com & snapdragoninsiders.com