Towards Causal Representation Learning

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, Yoshua Bengio

Presented by Jason Hartford

Overview

- Fundamentals of Causal Inference (Section 2 and 3)
- Independent Causal Mechanisms (Section 4)
- Causal Discovery and Machine Learning (Section 5)
- Learning Causal Variables (Section 6)
- Implications for Machine Learning (Section 7)

Overview

- Fundamentals of Causal Inference (Section 2 and 3)
- Independent Causal Mechanisms (Section 4)
- Causal Discovery and Machine Learning (Section 5)
- Learning Causal Variables (Section 6)
- Implications for Machine Learning (Section 7)

Prediction, interventions, counterfactuals; observational vs interventional data, structural causal models (SCMs) & graphs

Overview

- Fundamentals of Causal Inference (Section 2 and 3)
- Independent Causal Mechanisms (Section 4)
- Causal Discovery and Machine Learning (Section 5)
- Learning Causal Variables (Section 6)
- Implications for Machine Learning (Section 7)

Independent Causal Mechanisms (ICM) Principle & Sparse Mechanism Shift (SMS) hypothesis

Overview

- Fundamentals of Causal Inference (Section 2 and 3)
- Independent Causal Mechanisms (Section 4)
- Causal Discovery and Machine Learning (Section 5)
- Learning Causal Variables (Section 6)
- Implications for Machine Learning (Section 7)

Overview

- Fundamentals of Causal Inference (Section 2 and 3)
- Independent Causal Mechanisms (Section 4)
- Causal Discovery and Machine Learning (Section 5)
- Learning Causal Variables (Section 6)
- Implications for Machine Learning (Section 7)

Overview

- Fundamentals of Causal Inference (Section 2 and 3)
- Independent Causal Mechanisms (Section 4)
- Causal Discovery and Machine Learning (Section 5)
- Learning Causal Variables (Section 6)
- Implications for Machine Learning (Section 7)

Semi-Supervised Learning, Adversarial Vulnerability, Robustness and Strong Generalization, Reinforcement Learning, Scientific Applications, Multi-Task Learning and Continual Learning

Overview

- *Fundamentals of Causal Inference (Section 2 and 3)*
- *Independent Causal Mechanisms (Section 4)*
- ~~Causal Discovery and Machine Learning (Section 5)~~
- **Learning Causal Variables (Section 6)**
- ~~Implications for Machine Learning (Section 7)~~

Some caveats upfront

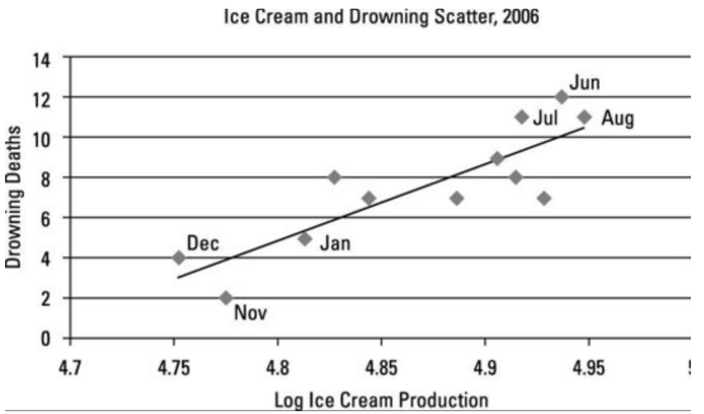
- This is a tough paper to present because the most interesting ideas for this group are speculative / aspirational…
- When I first read this paper I thought it was mostly a paper version of Bernhard Schölkopf's keynotes and tutorials…
- But then I started working on representation learning and reread it and thought it had some really useful insights & foreshadowed our progress on the problem.

Causality Fundamentals

- Prediction “fine” under IID assumption - e.g. from $P(X, Y)$ can predict,
  $P(Y = \text{drown} \mid X = \text{Ice cream})$

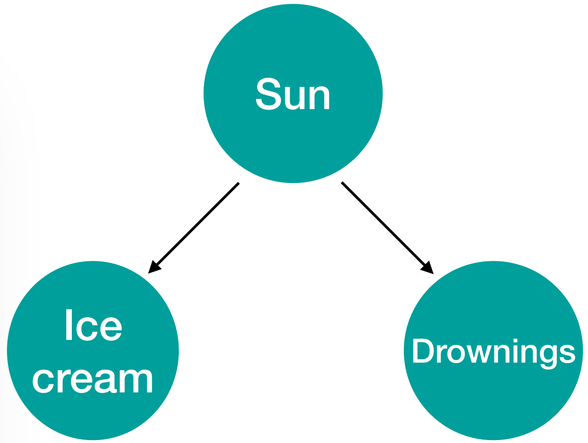
Ice cream

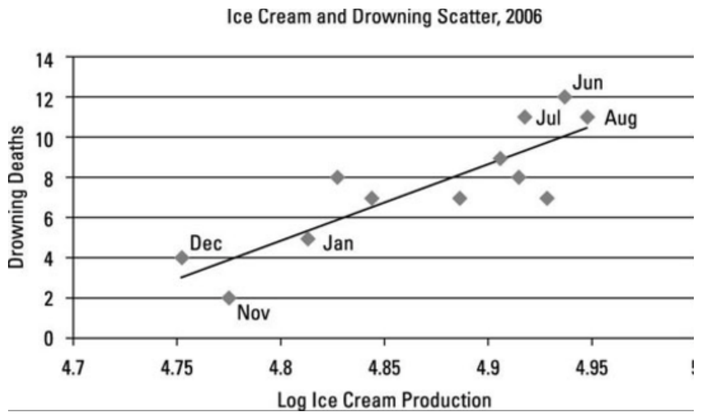
Drownings



Causality Fundamentals

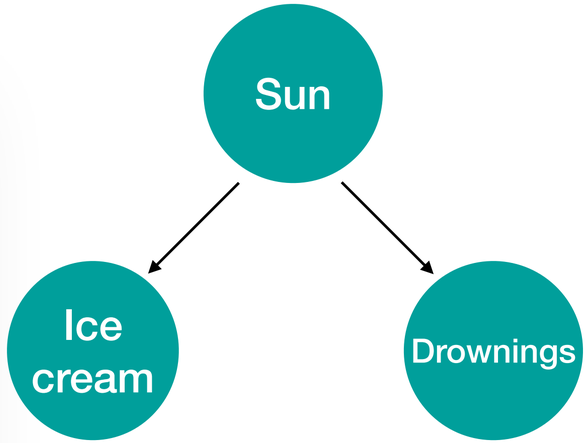
- Prediction “fine” under IID assumption - e.g. from $P(X, Y)$ can predict,
  $P(Y = \text{drown} \mid X = \text{Ice cream})$

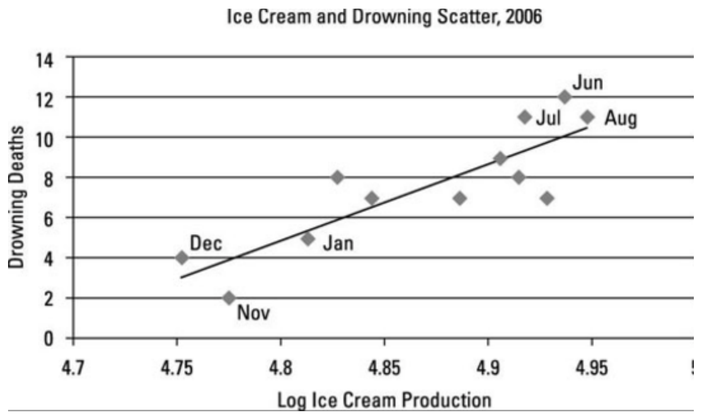


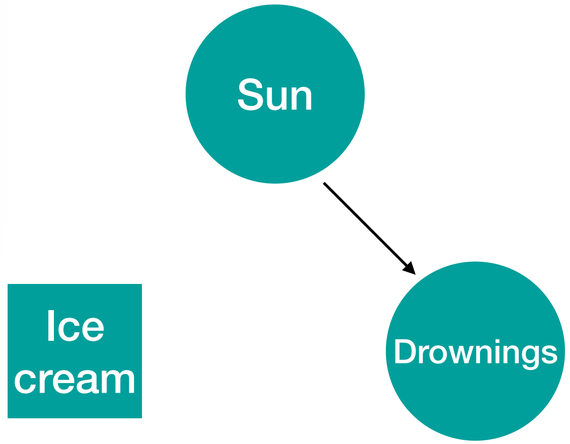


Causality Fundamentals

- Prediction “fine” under IID assumption - e.g. from $P(X, Y)$ can predict,
  $P(Y = \text{drown} \mid X = \text{Ice cream})$
- Interventions change the joint distribution so IID no longer applies.

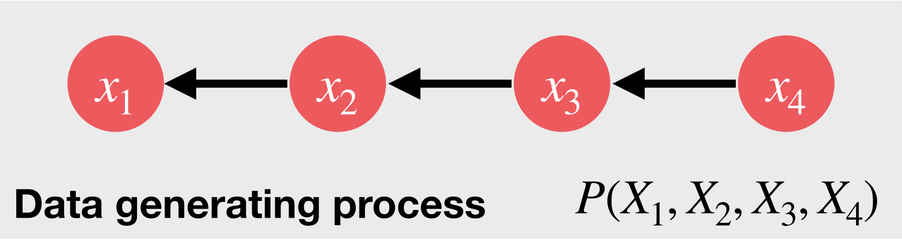




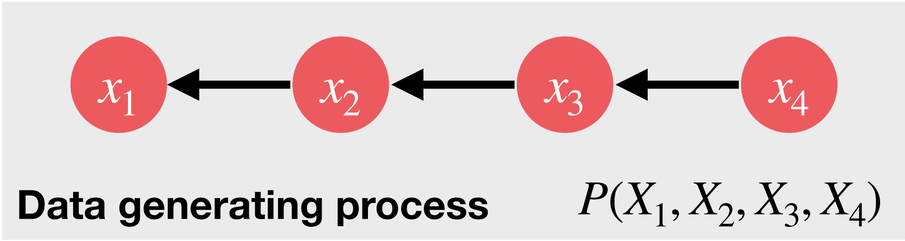


Intervene - e.g. what would happen if we ban ice cream sales?

Causality Fundamentals



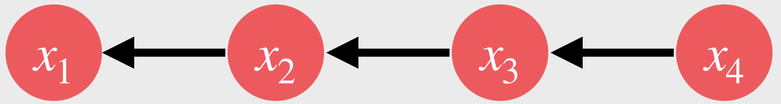
Causality Fundamentals



SCM representation

$$x_4 = f_4(u_4)$$
$$x_3 = f_3(x_4, u_3)$$
$$x_2 = f_2(x_3, u_2)$$
$$x_1 = f_1(x_2, u_1)$$

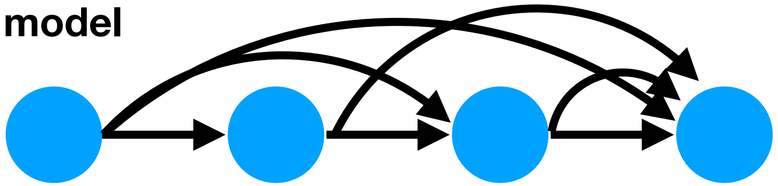
Causality Fundamentals



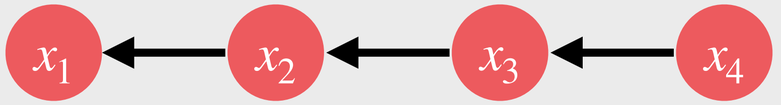
Data generating process $P(X_1, X_2, X_3, X_4)$

Autoregressive model

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_{1:i-1})$$



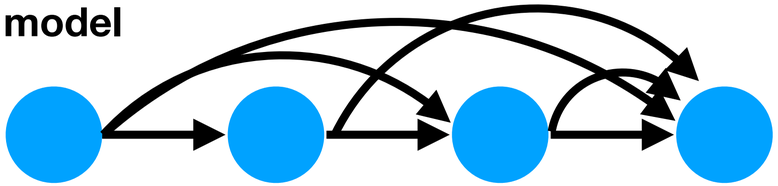
Causality Fundamentals



Data generating process $P(X_1, X_2, X_3, X_4)$

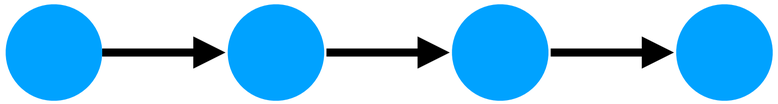
Autoregressive model

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_{1:i-1})$$

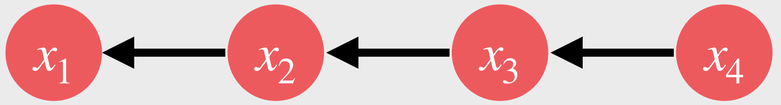


Markov equivalent model

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_{i-1})$$



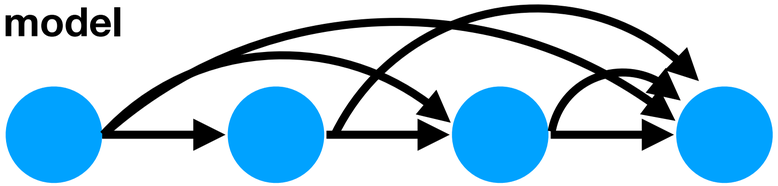
Causality Fundamentals



Data generating process $P(X_1, X_2, X_3, X_4)$

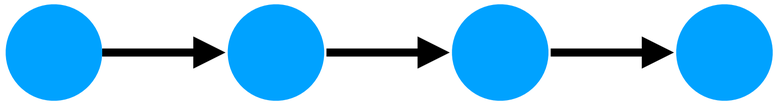
Autoregressive model

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_{1:i-1})$$



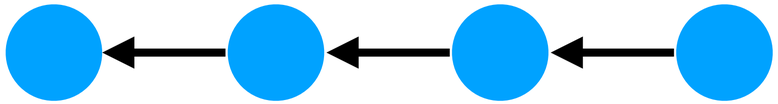
Markov equivalent model

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_{i-1})$$

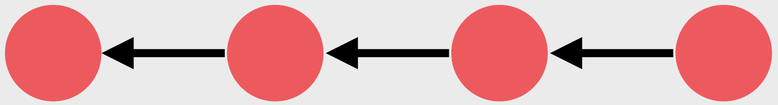


Causal model

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \mathbf{PA}_i)$$

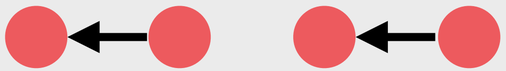


Causality Fundamentals



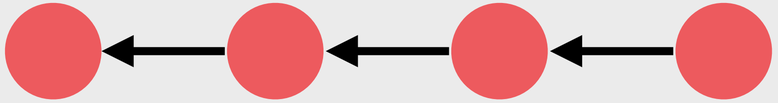
$P(X_1, X_2, X_3, X_4)$

$P(X_1, X_2, \mathbf{do}(X_3), X_4)$



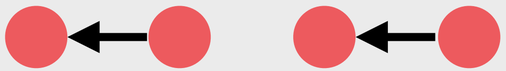
$P(X_1, \mathbf{do}(X_2), X_3, X_4)$

Causality Fundamentals



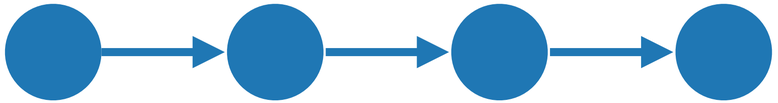
$P(X_1, X_2, X_3, X_4)$

$P(X_1, X_2, \mathbf{do}(X_3), X_4)$

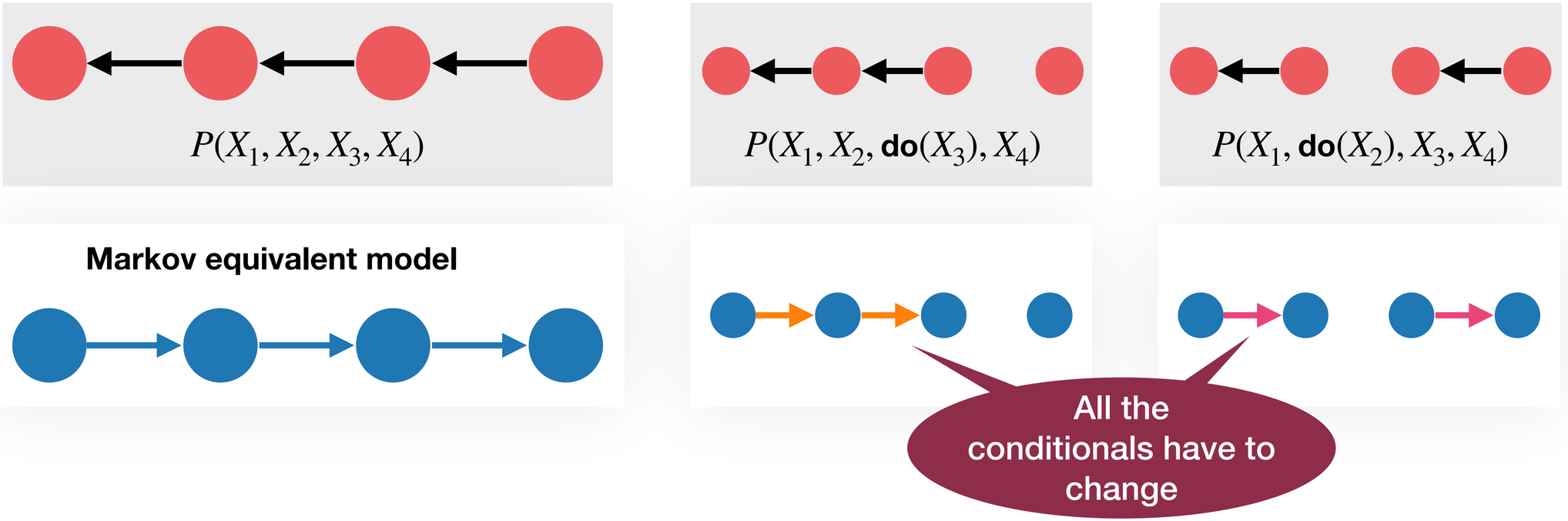


$P(X_1, \mathbf{do}(X_2), X_3, X_4)$

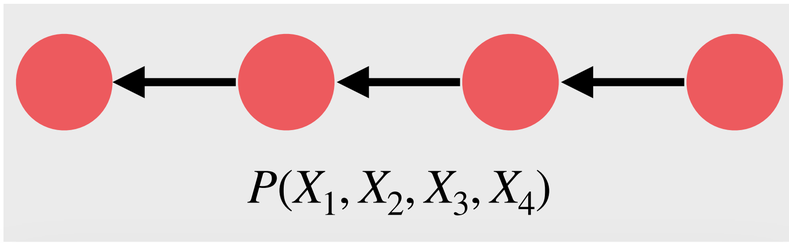
Markov equivalent model



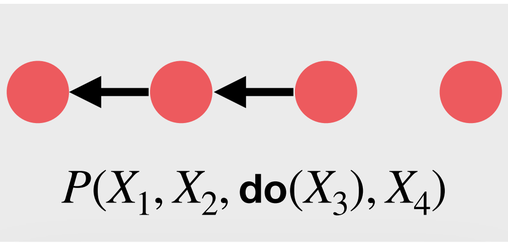
Causality Fundamentals



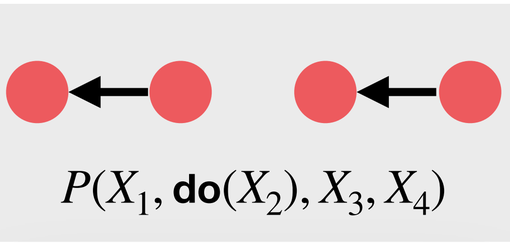
Causality Fundamentals



$P(X_1, X_2, X_3, X_4)$

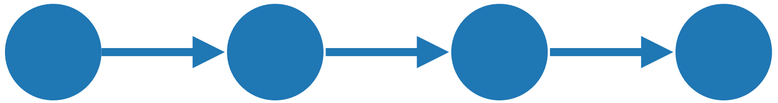


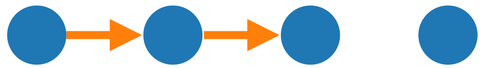
$P(X_1, X_2, \mathbf{do}(X_3), X_4)$



$P(X_1, \mathbf{do}(X_2), X_3, X_4)$

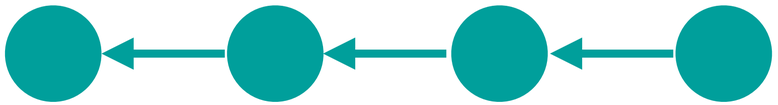
Markov equivalent model



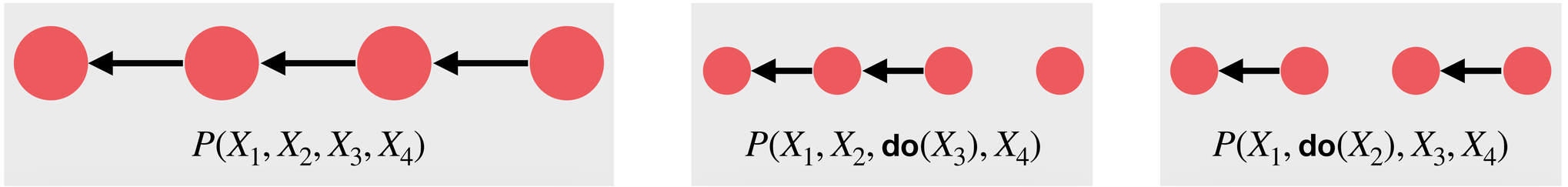


All the conditionals have to change

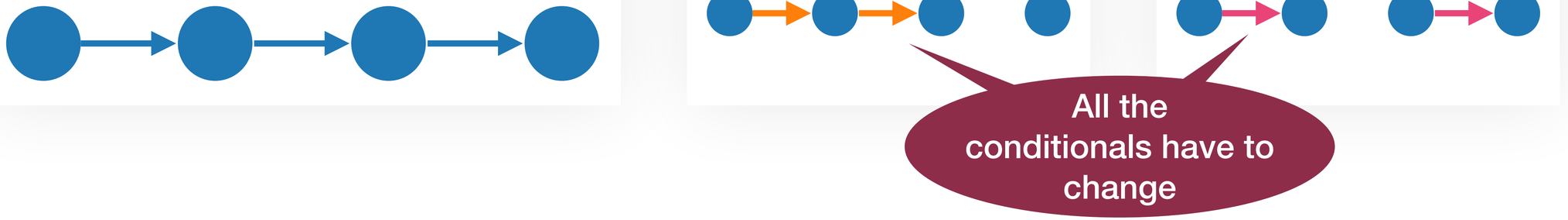
Causal model



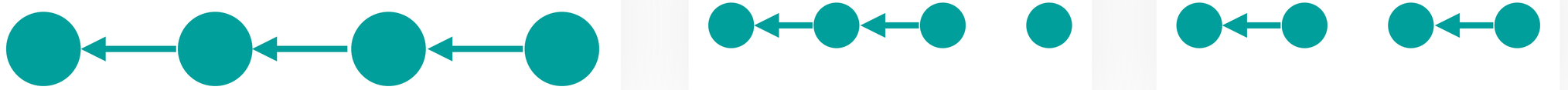
Causality Fundamentals



Markov equivalent model



Causal model

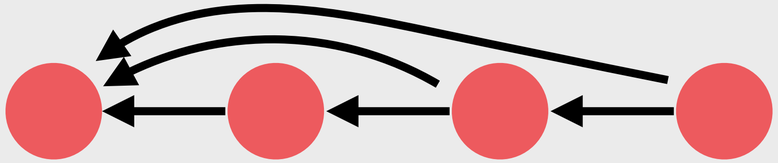


Independent causal mechanisms

Independent Causal Mechanisms (ICM) Principle. *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.*

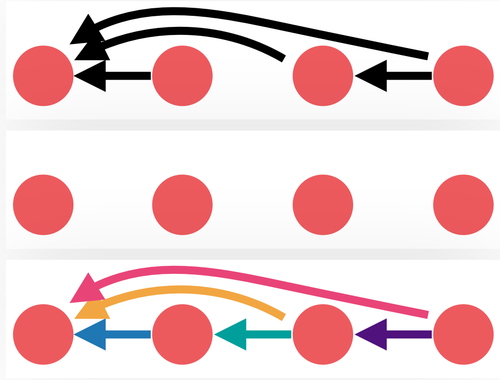
The generative distribution is **modular**; and if our models reflected this, we would inherit this modularity.

Sparse Mechanism Shift Hypothesis



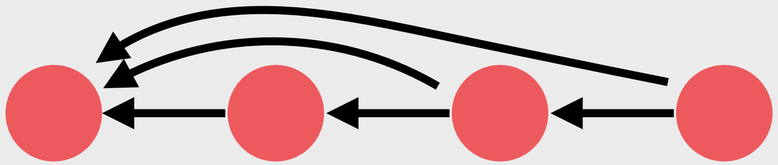
Data generating process

$$P(X_1, X_2, X_3, X_4) \rightarrow P'(X_1, X_2, X_3, X_4)$$



Sparse change

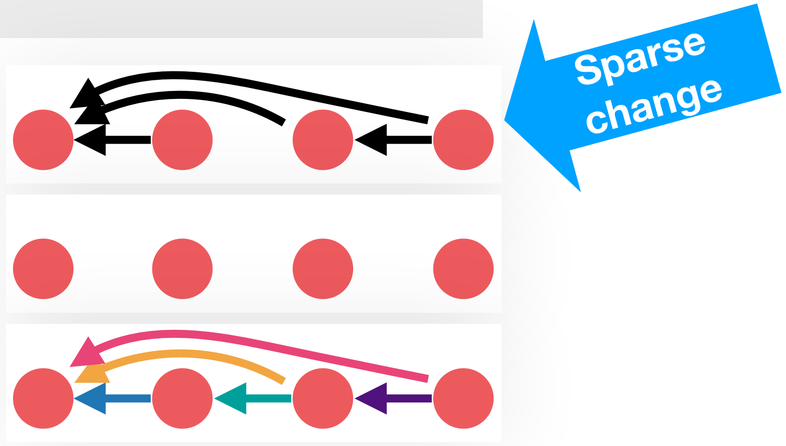
Sparse Mechanism Shift Hypothesis



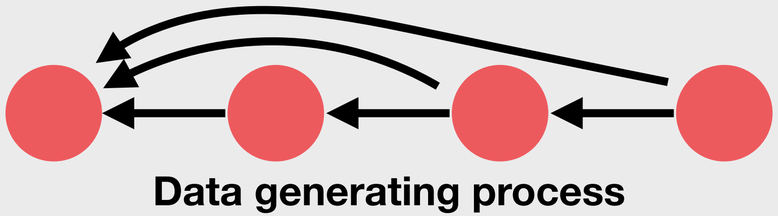
Data generating process

$$P(X_1, X_2, X_3, X_4) \rightarrow P'(X_1, X_2, X_3, X_4)$$

Sparse Mechanism Shift (SMS). *Small distribution changes tend to manifest themselves in a sparse or local way in the causal/disentangled factorization (4), i.e., they should usually not affect all factors simultaneously.*

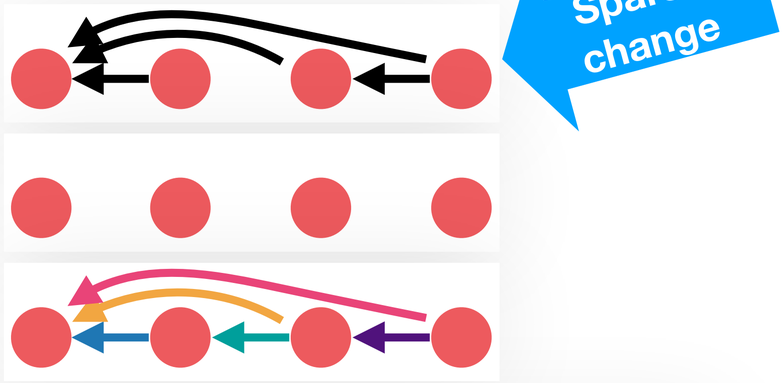


Sparse Mechanism Shift Hypothesis



$$P(X_1, X_2, X_3, X_4) \rightarrow P'(X_1, X_2, X_3, X_4)$$

Sparse Mechanism Shift (SMS). *Small distribution changes tend to manifest themselves in a sparse or local way in the causal/disentangled factorization (4), i.e., they should usually not affect all factors simultaneously.*



SMS hypothesis is an Occum's razor-style assumption: small distribution shifts are assumed to be the result of a small number of factors. I.e. there's a "simple" (sparse) explanation for the distribution shift.

Causality Fundamentals

- Generalizing to interventional distributions isn't easy—*we could run a seminar course on this topic just using methods developed by people in this group!*—but relatively well-understood with **semantically meaningful variables** and a **causal graph**.
- No causal graph? Need *faithfulness,* and/or some interventions, or functional restrictions. See the section 5 on causal discovery.
- Unstructured variables? Today's topic.

Learning causal variables

- In ML we often work with unstructured data like images, audio, video, etc. where we don't directly observed the causal variables.

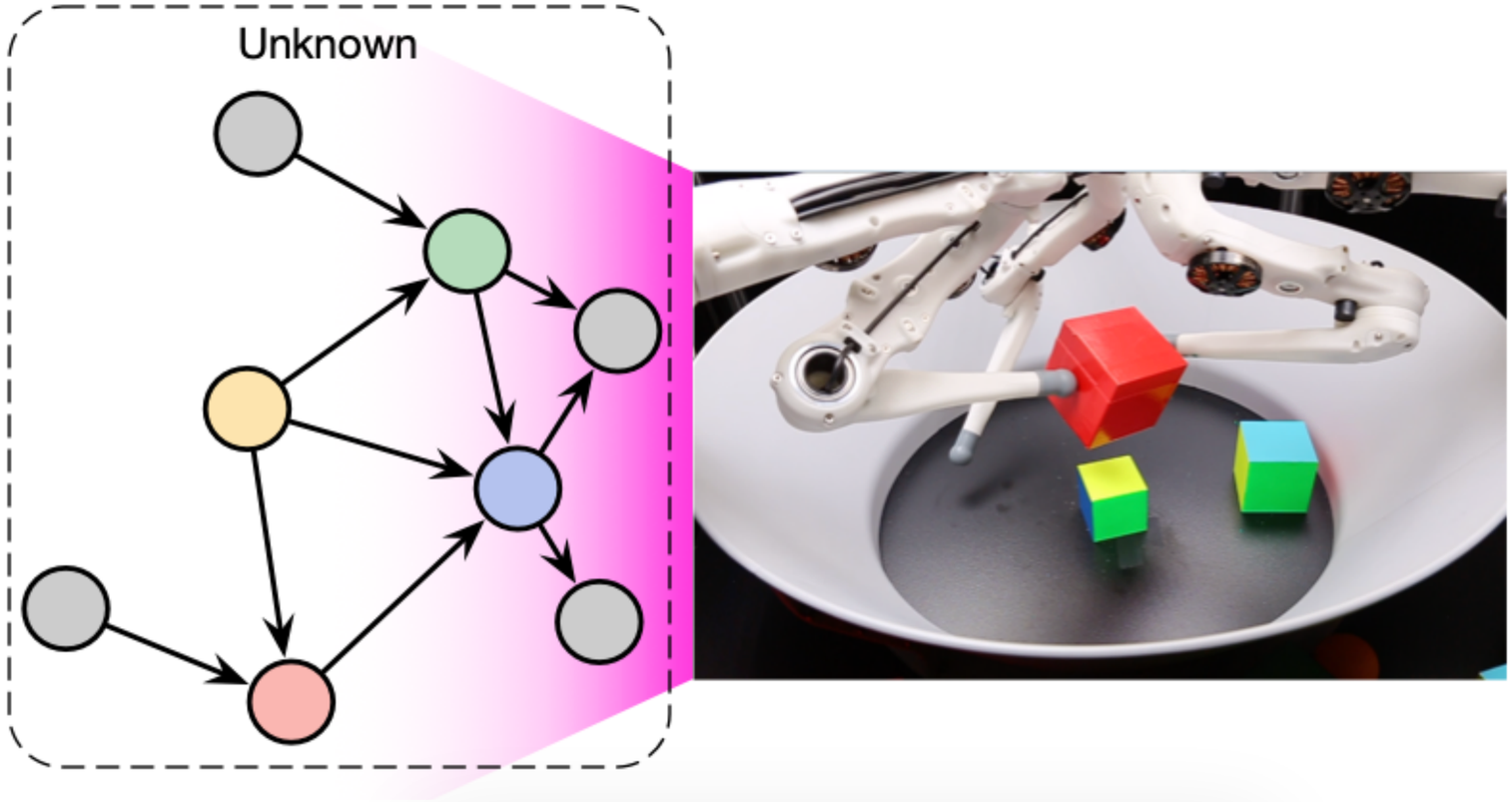
- This paper argues that…

    1. We should strive to embed an SCM into larger ML models with high-dim unstructured inputs and output & SCM-based inner workings.

    2. Resulting architecture has modular components, $f_i(\mathbf{PA}_i, u_i)$, which can be fine-tuned and repurposed for new tasks.

    3. SMS hypothesis can be used to enforce this structure.

Data generating process

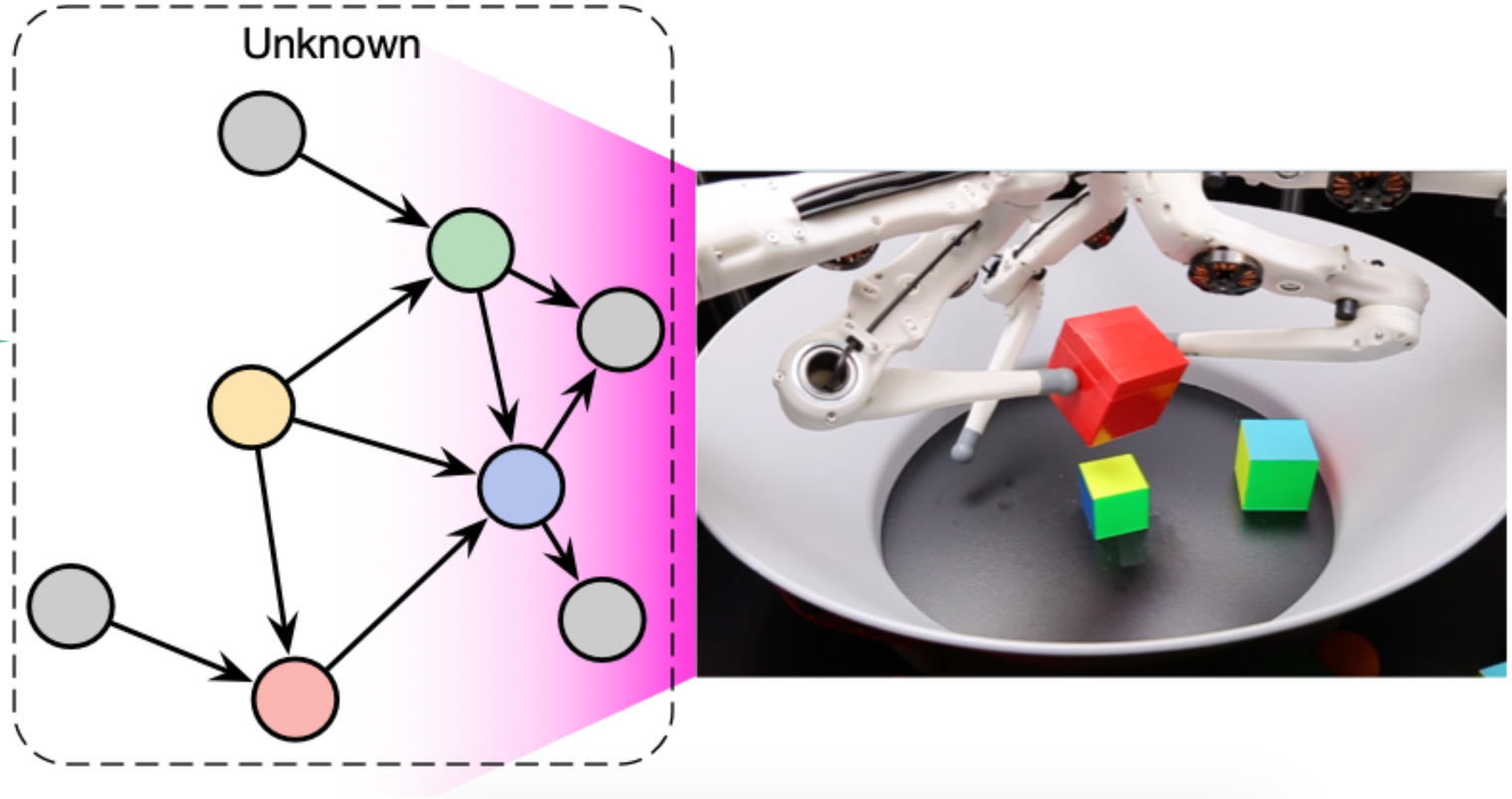


$$X = G(S_1, S_2, \ldots, S_n)$$

G is the function (“camera”) that maps from state variables S_i to an image X

Data generating process

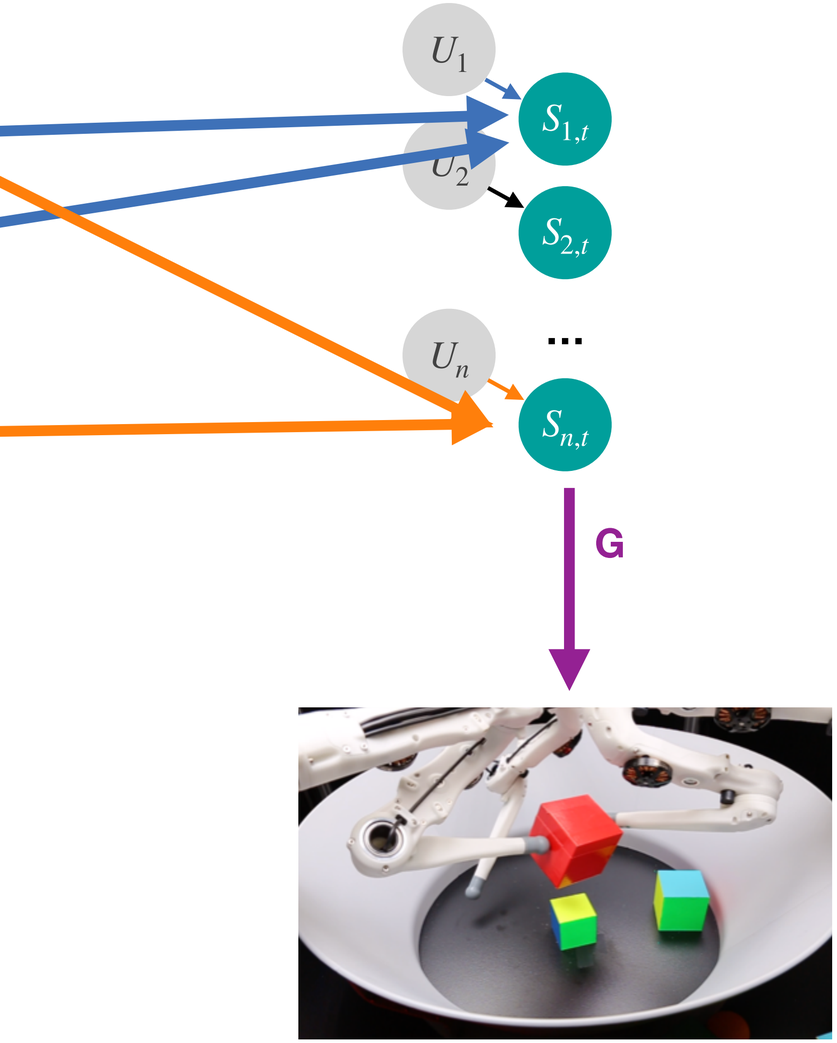
Not mentioned in the paper, but in physical systems like these it's probably more accurate to think of this graph as an unrolled dynamical system



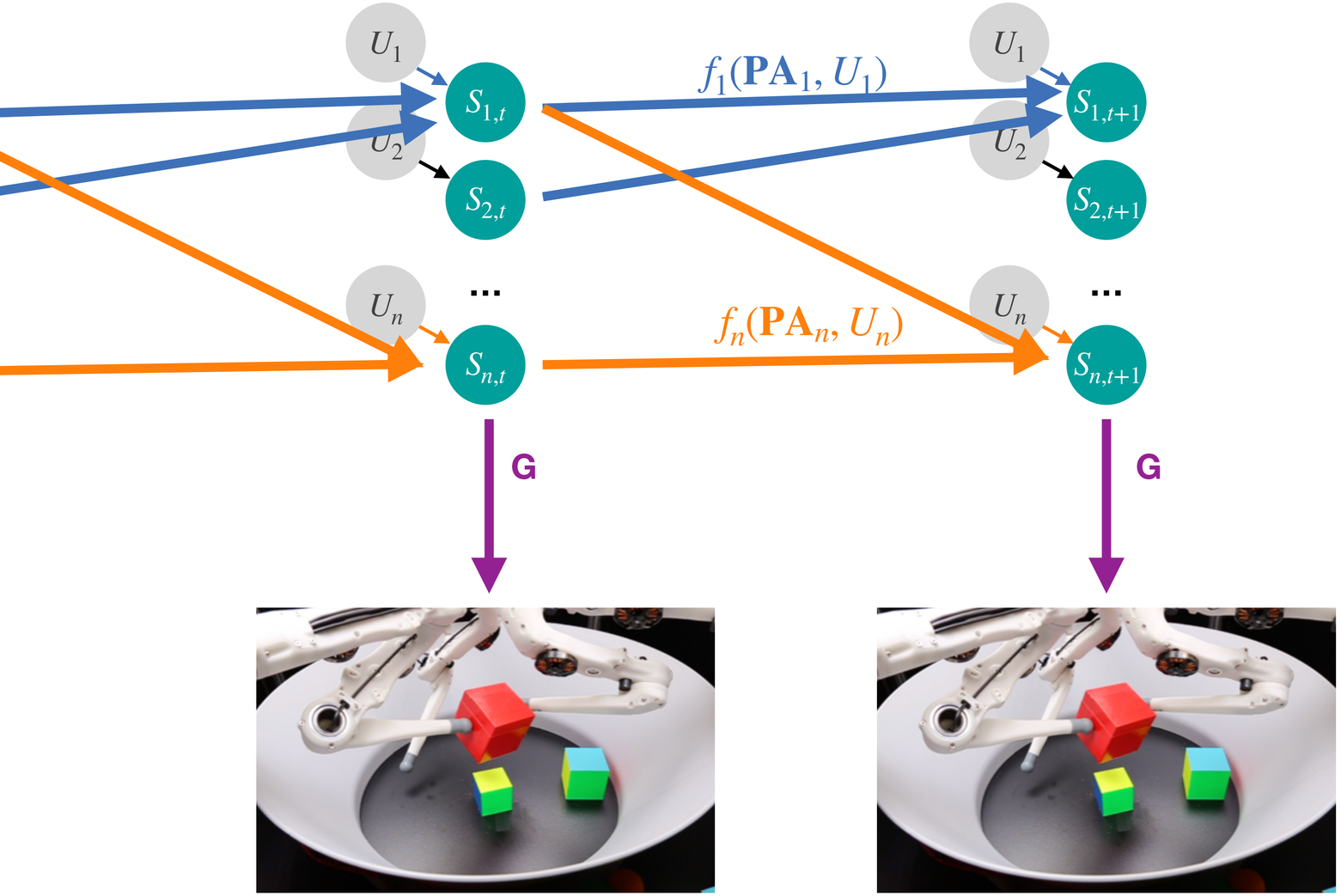
$$X = G(S_1, S_2, \ldots, S_n)$$

G is the function ("camera") that maps from state variables S_i to an image X

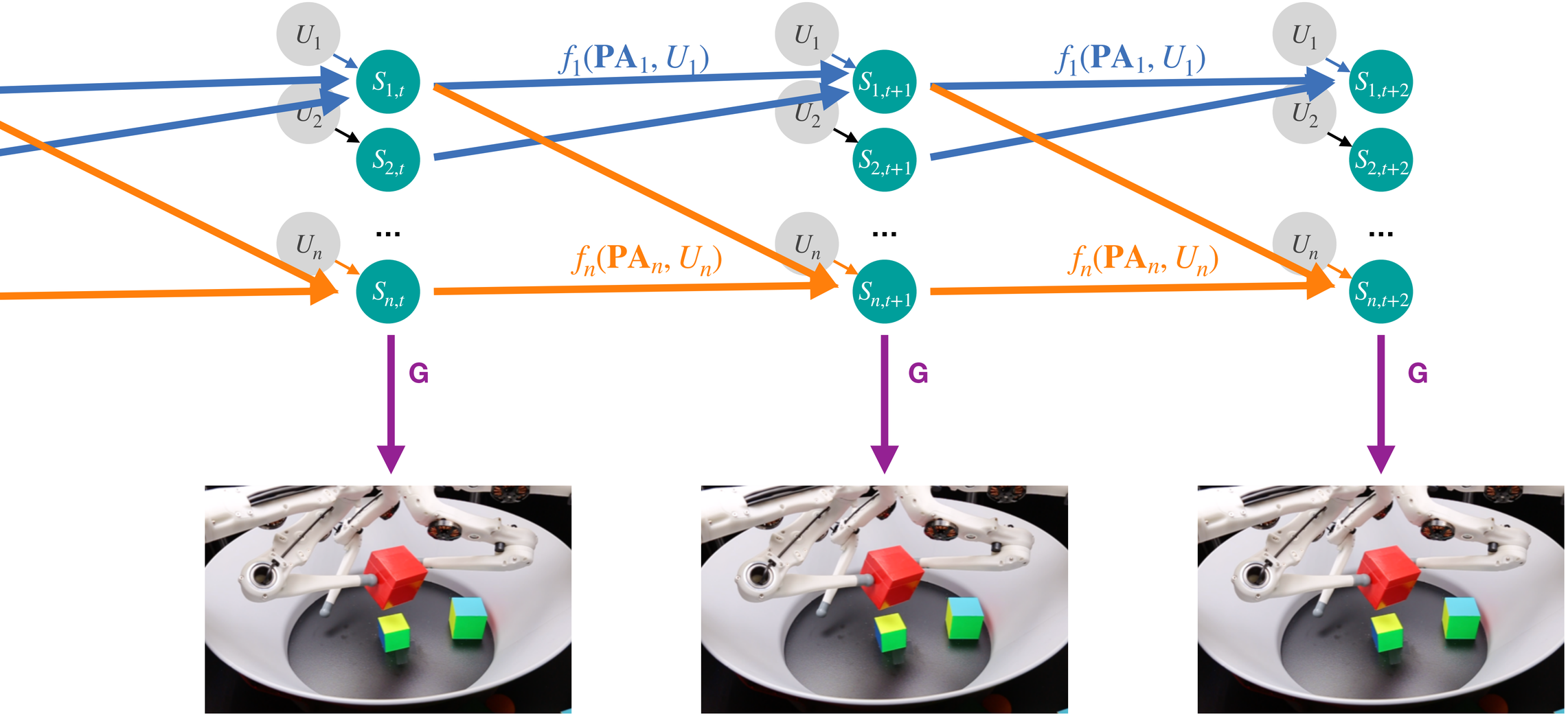
Data generating process



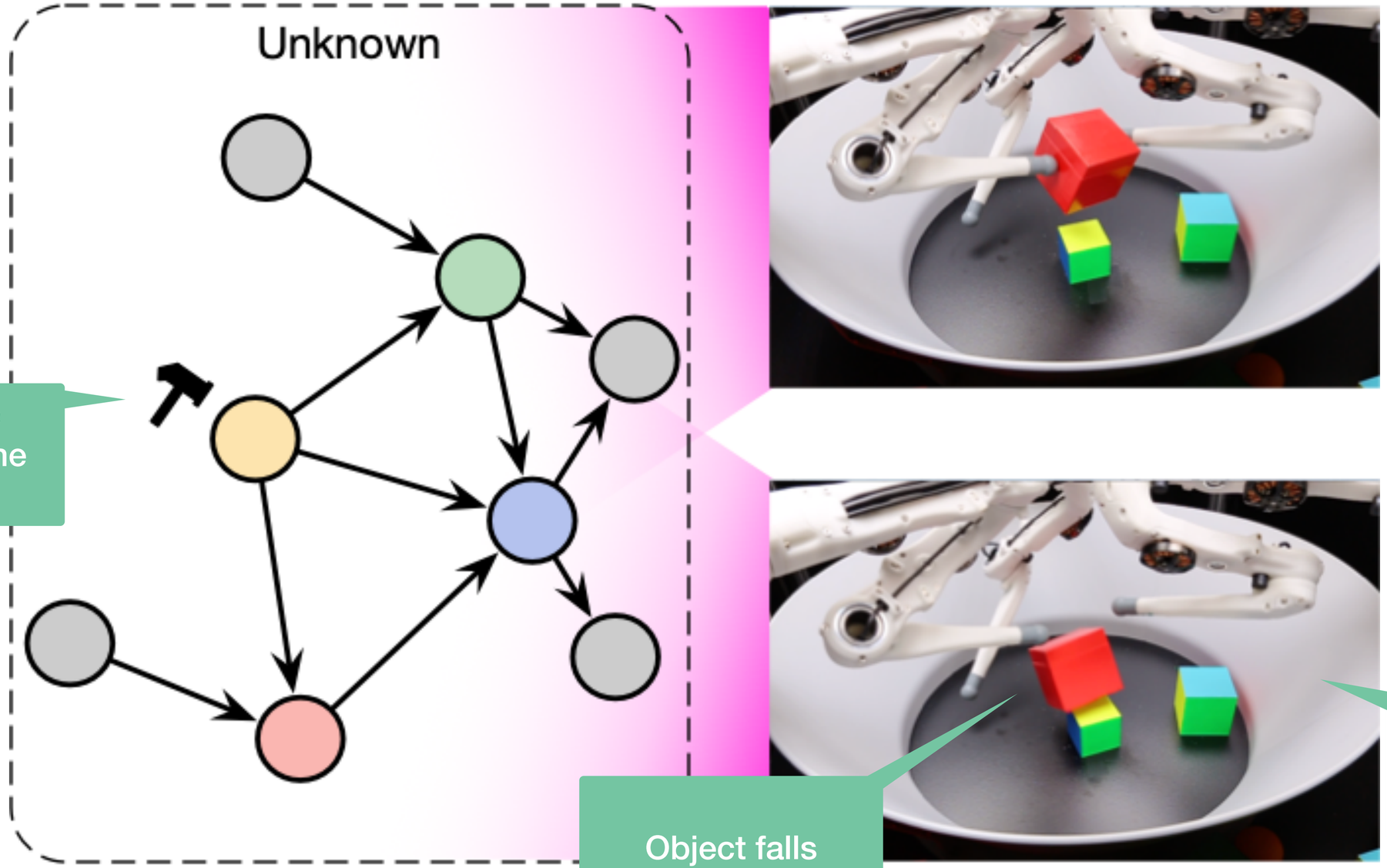
Data generating process



Data generating process



Sparse shifts



Problem 1: Learning disentangled representations

The causal factorization of the S_i's implies this disentangled representation,

Problem 1: Learning disentangled representations

The causal factorization of the S_i's implies this disentangled representation,

$$P(S_1, \ldots, S_n) = \prod_{i=1}^{n} P(S_i \,|\, \mathbf{PA}_i)$$

Problem 1: Learning disentangled representations

The causal factorization of the S_i's implies this disentangled representation,

$$P(S_1, \ldots, S_n) = \prod_{i=1}^{n} P(S_i \,|\, \mathbf{PA}_i)$$

This is more complicated than product distributions used in VAE-style model…
but if we could learn it we'd get two benefits:

Problem 1: Learning disentangled representations

The causal factorization of the S_i's implies this disentangled representation,

$$P(S_1, \ldots, S_n) = \prod_{i=1}^{n} P(S_i \mid \mathbf{PA}_i)$$

This is more complicated than product distributions used in VAE-style model… but if we could learn it we'd get two benefits:

- Conditionals, $P(S_i \mid \mathbf{PA}_i)$, are independently manipulable (ICM principle)

Problem 1: Learning disentangled representations

The causal factorization of the S_i's implies this disentangled representation,

$$P(S_1, \ldots, S_n) = \prod_{i=1}^{n} P(S_i \mid \mathbf{PA}_i)$$

This is more complicated than product distributions used in VAE-style model… but if we could learn it we'd get two benefits:

- Conditionals, $P(S_i \mid \mathbf{PA}_i)$, are independently manipulable (ICM principle)
- Conditionals are invariant across related problems

Problem 1: Learning disentangled representations

- We want to reconstruct this factorization from data, without knowing the S_i's.

Problem 1: Learning disentangled representations

- We want to reconstruct this factorization from data, without knowing the S_i's.
- To solve this, we'd need to go from $X \in \mathbb{R}^d$ to the causal variables and mechanisms, $S_i = f_i(\mathbf{PA}_i, U_i)$ for all i.

Problem 1: Learning disentangled representations

- We want to reconstruct this factorization from data, without knowing the S_i's.
- To solve this, we'd need to go from $X \in \mathbb{R}^d$ to the causal variables and mechanisms, $S_i = f_i(\mathbf{PA}_i, U_i)$ for all i.
- So - we need an **encoder**, q, that maps to the noise variables U_i

Problem 1: Learning disentangled representations

- We want to reconstruct this factorization from data, without knowing the S_i's.
- To solve this, we'd need to go from $X \in \mathbb{R}^d$ to the causal variables and mechanisms, $S_i = f_i(\mathbf{PA}_i, U_i)$ for all i.
- So - we need an **encoder**, q, that maps to the noise variables U_i
- One then applies the f_i's (learned jointly) to get the variables S_i and decode to minimize reconstruction error.

Problem 1: Learning disentangled representations

$$\textbf{dec} \circ f \circ \textbf{enc}(x) \approx x$$

Encoder

U_1

U_2

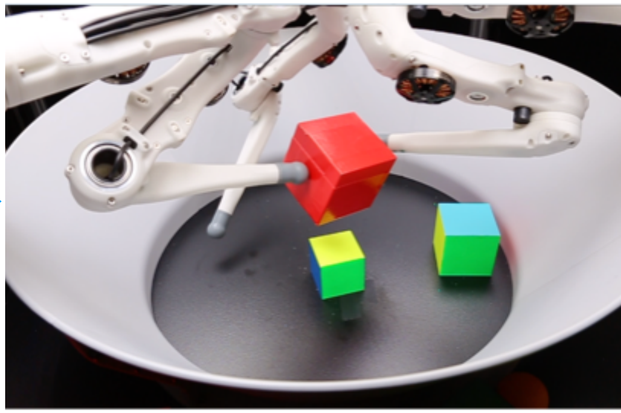
U_3

...

U_n

Unknown

Decoder



Note - if the graph is trivial this reduces to a VAE

(even this simple case is not identified with IID data)

Problem 2: Learning transferrable mechanisms

- Future AI models will need to **re-use components**, which requires them to be robust across tasks and environments.
- ICM principle says that **mechanisms are modular**, so if we can model them, then we can reuse them.
- The model structure needs to **reflect the causal structure** of the environment's structure even if the implementation is different.
- E.g. vision is invariant to lighting changes which suggests the need for a "gain control" mechanism which plays a corresponding role to the physical lighting mechanism in the true DGP.

Conclusions

- Causal structure of the world gives many invariances that we could exploit.
- If our models reflect this causal structure we gain modularity which should allow for adaption to distribution shifts
- We could also used causal world models to allow intervention, planning, and reasoning, realizing Konrad Lorenz' notion of *thinking as acting in an imagined space* and make RL more efficient by considering counterfactuals (modulo identifiability issues).