Counterfactual Realizability

Arvind Raghavan, Elias Bareinboim

Causal Artificial Intelligence Lab
Columbia University

13th International Conference on Learning Representations, Singapore

CS@CU COMPUTER SCIENCE

Preliminaries

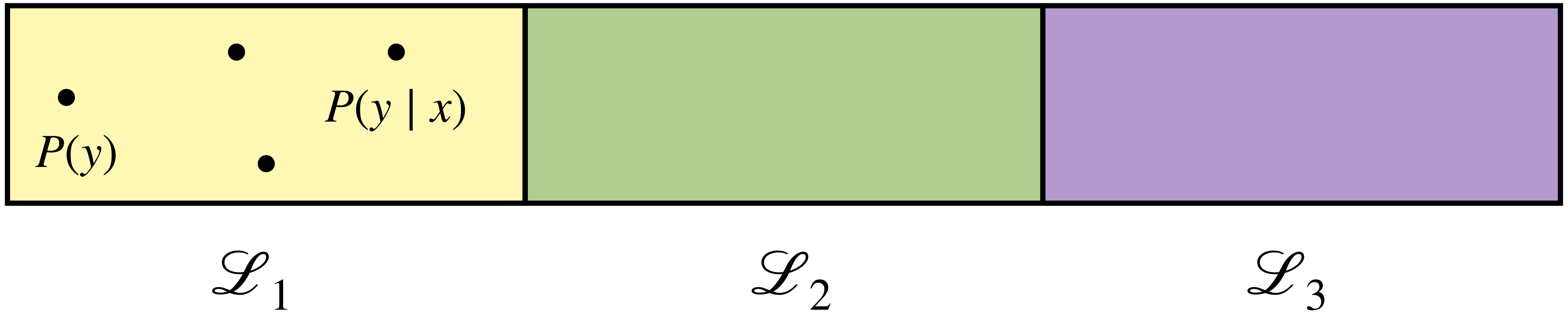
- We use *Structural Causal Models* (SCMs) to model the data-generating process in a real-world environment.[1]
- The *Pearl Causal Hierarchy* (PCH) describes the three ways an agent can interact with a system of interest:[2]
  - Layer 1 ($\mathscr{L}_1$) contains distributions from the *observational* regime
  - Layer 2 ($\mathscr{L}_2$) contains distributions from the *interventional* regime
  - Layer 3 ($\mathscr{L}_3$) contains distributions from the *counterfactual* regime

[1] Pearl (2009). Causality: Models, Reasoning, and Inference

[2] Bareinboim et al (2022). On Pearl's Hierarchy and the Foundations of Causal Inference

The limits of experimentation

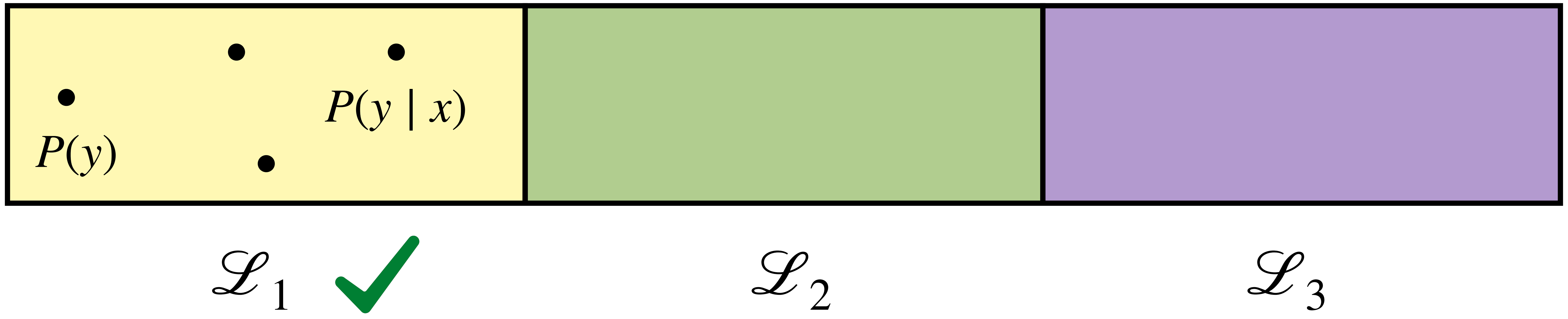
Question:
From which distributions is it possible to draw samples in the real world, in principle, where the SCM is unknown?



PCH induced by an (unknown) SCM

The limits of experimentation

Question:
From which distributions is it possible to draw samples in the real world, in principle, where the SCM is unknown?



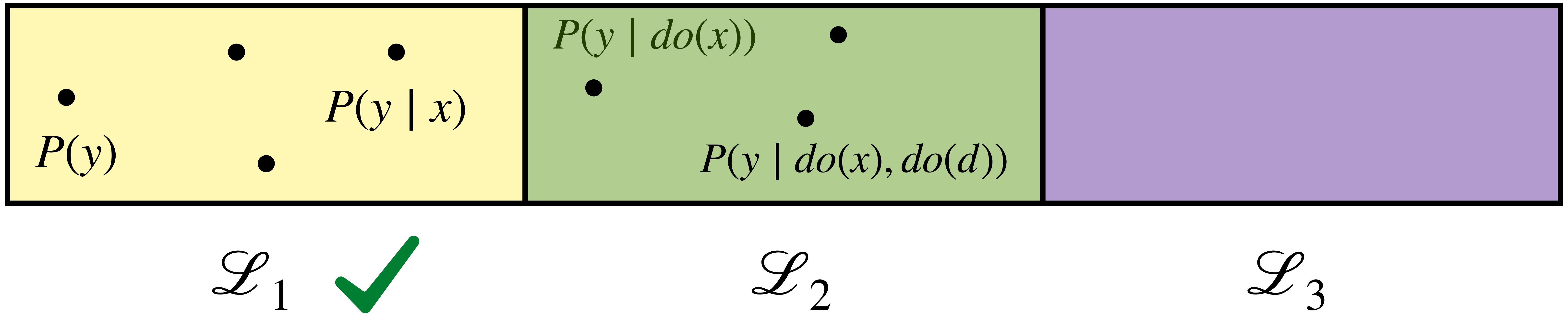
PCH induced by an (unknown) SCM

- Observe $\mathbf{V}$

The limits of experimentation

Question:

From which distributions is it possible to draw samples in the real world, in principle, where the SCM is unknown?

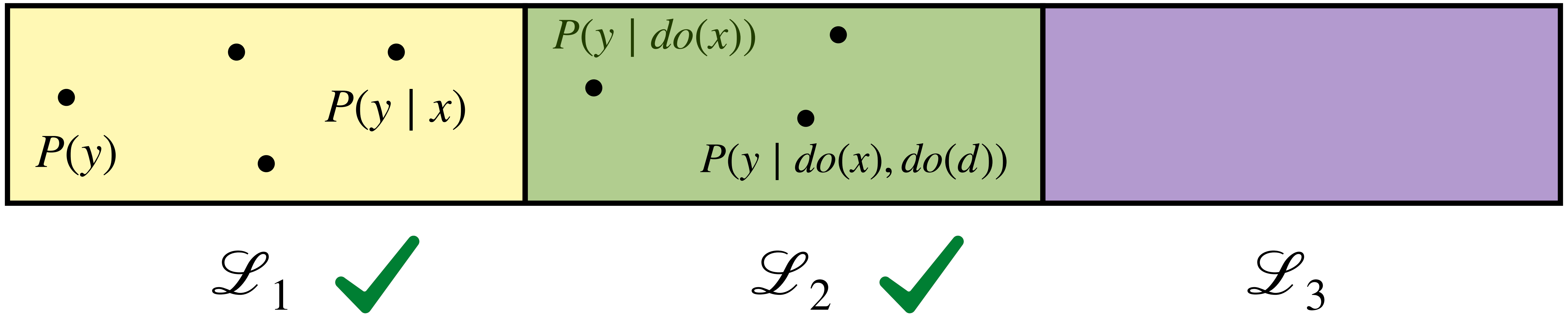


PCH induced by an (unknown) SCM

The limits of experimentation

Question:

From which distributions is it possible to draw samples in the real world, in principle, where the SCM is unknown?

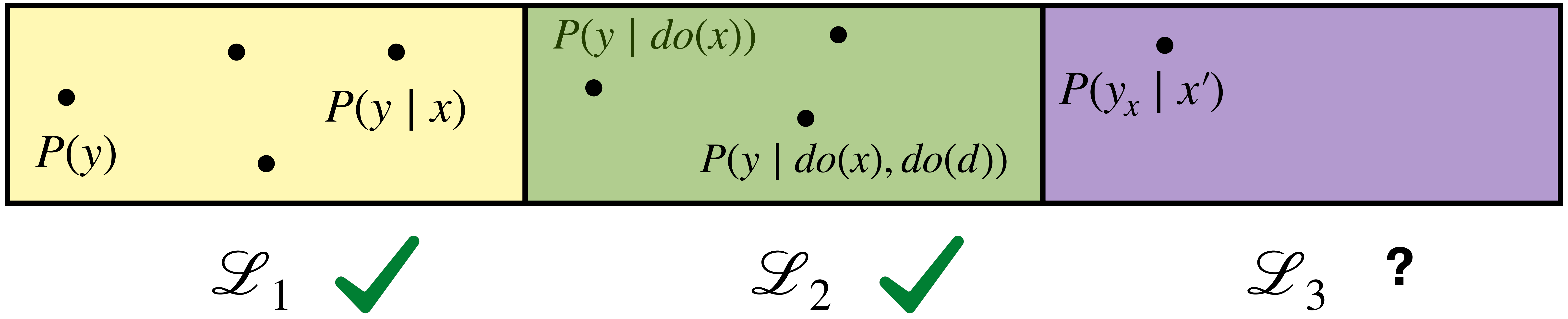


PCH induced by an (unknown) SCM

- Fisherian randomization of $\mathbf{X}$
- Observe $\mathbf{V}$ under $do(\mathbf{x})$

The limits of experimentation

Question:
From which distributions is it possible to draw samples in the real world, in principle, where the SCM is unknown?

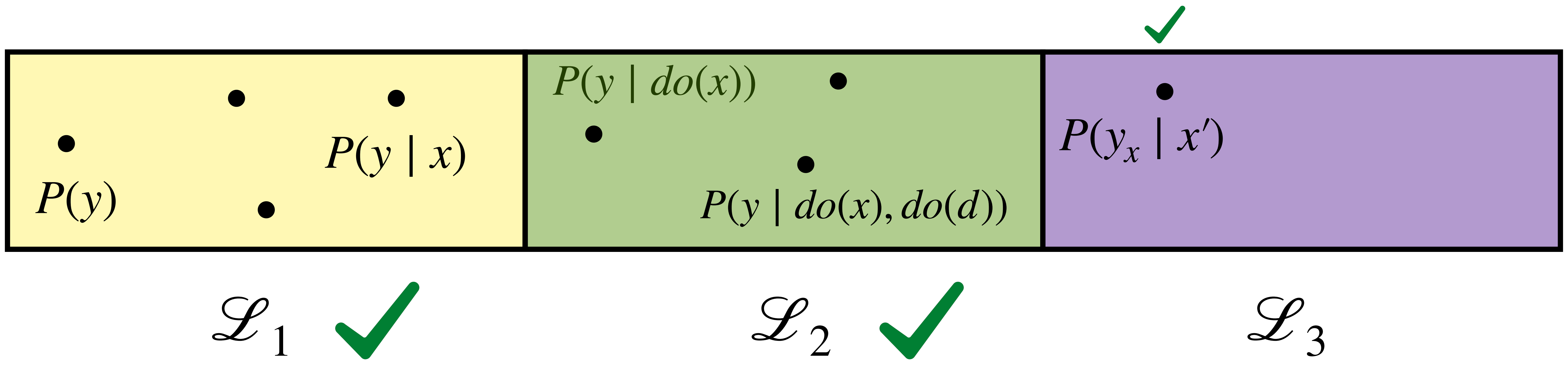


PCH induced by an (unknown) SCM

Generally believed to be inferred only by identification

The limits of experimentation

Question:
From which distributions is it possible to draw samples in the real world, in principle, where the SCM is unknown?



PCH induced by an (unknown) SCM

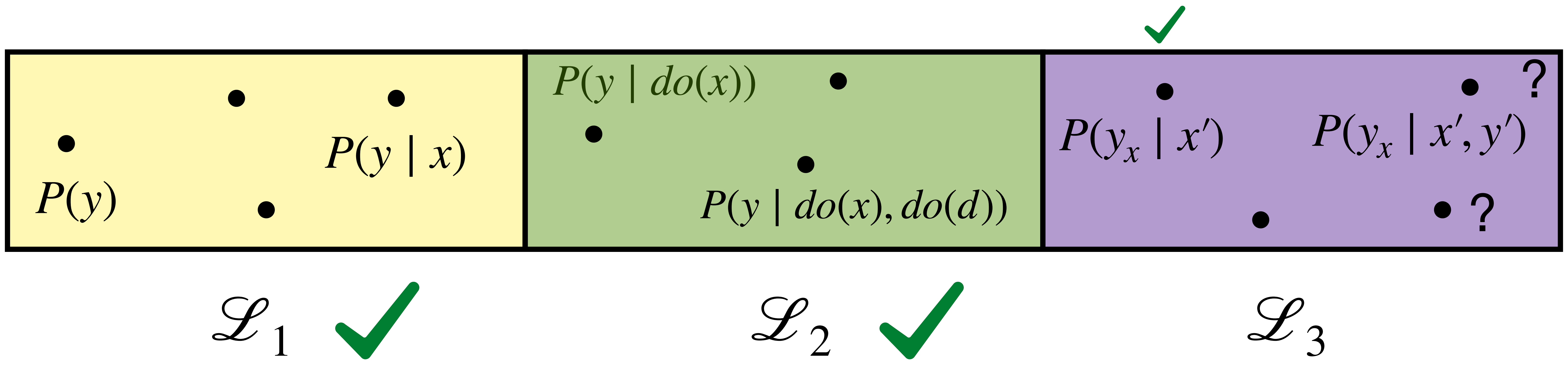
- There is at least one $\mathscr{L}_3$ distribution that can be experimentally realised: $P(Y_x \mid x')$
- Cf. Greedy Casino decision problem: randomly assign X given that unit *would have naturally performed $X = x'$ otherwise.*[3]

[3] Bareinboim, Forney, and Pearl (2015). Bandits with Unobserved Confounders: A Causal Approach

The limits of experimentation

Question:

From which distributions is it possible to draw samples in the real world, in principle, where the SCM is unknown?



PCH induced by an (unknown) SCM

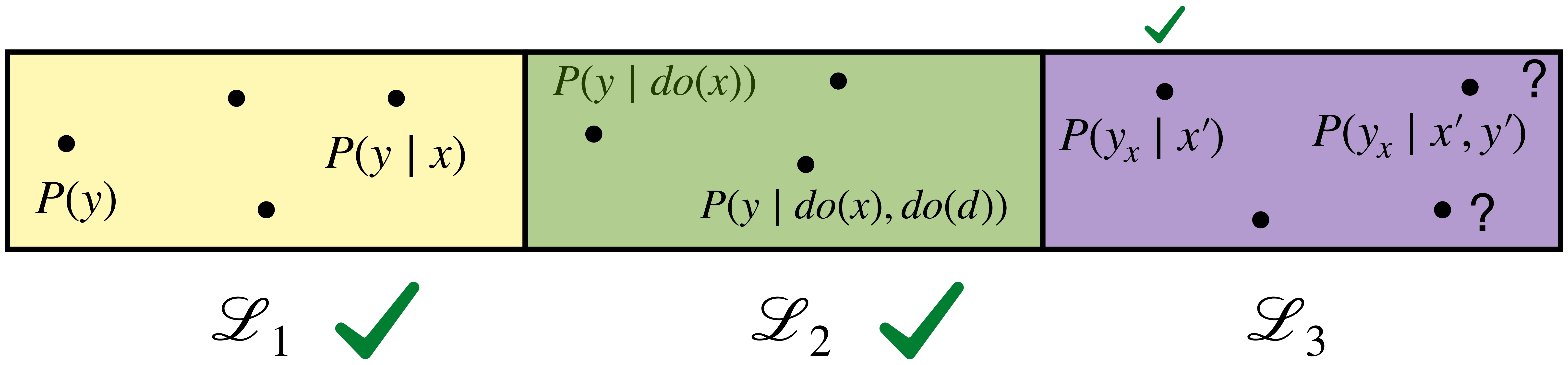
Is there a similar clever way to directly sample from other $\mathscr{L}_3$ distributions?

Is this the limit?

The limits of experimentation

Rephrasing the Question:

How far up the PCH can one go, in principle, via direct experimentation?



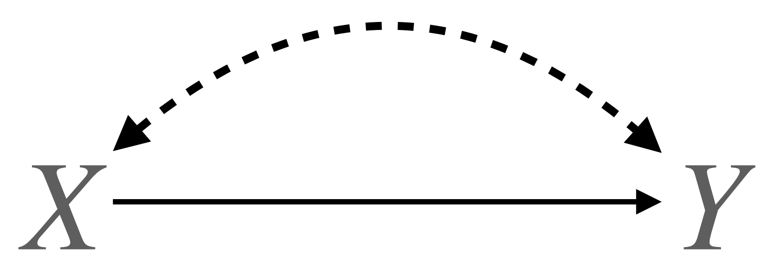
PCH induced by an (unknown) SCM

Note:

- Assume SCM and unit identity $\mathbf{U}$ are unknown
- Not considering environments like simulators or open-source LLMs

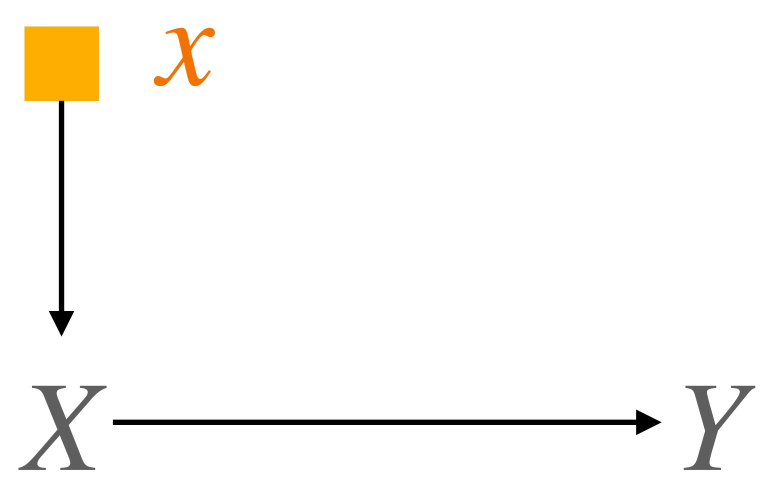
Contribution #1

We formalize a new physical action that an agent can perform in a real-world environment, represented by a causal diagram.



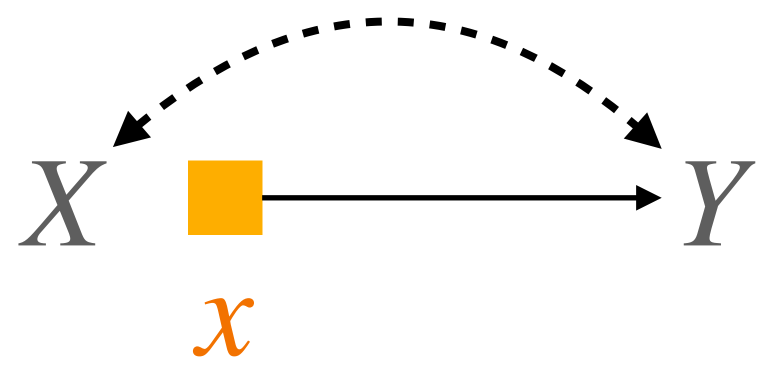
Fisherian Randomization:

- Override the unit's natural treatment value and randomly assigning x. Corresponds to a stochastic intervention (stochastic version of $do(x)$).
- Intervention on the node.



Counterfactual Randomization:

- Randomly fix the value of $X = x$ as perceived by a child variable Y.
- Intervention along the edge (examples in paper). Subsumes previous similar notions from the literature.
- Key differences: (a) does not erase the unit's natural treatment value; (b) does not necessarily affect all descendants of X.



Contribution #2

Realizability: formal definition of the ability to physically draw samples from a distribution in an environment.

CTF-REALIZE algorithm:

- Input: causal diagram $\mathcal{G}$, list of feasible actions an agent can physically perform, $\mathcal{L}_3$ distribution $P(\mathbf{W}_\star)$.
- I.i.d sample from $P(\mathbf{W}_\star)$ if and only if the distribution is *realizable*.

Graphical criterion:

- If the feasible action set is "maximal", the distribution is realizable if and only if the following condition is met

$$\nexists W_{\mathbf{t}}, W_{\mathbf{z}} \in An_{\mathcal{G}}(\mathbf{W}_\star) \text{ where } \mathbf{t} \neq \mathbf{z}$$

Contribution #2

Realizability: formal definition of the ability to physically draw samples from a distribution in an environment.

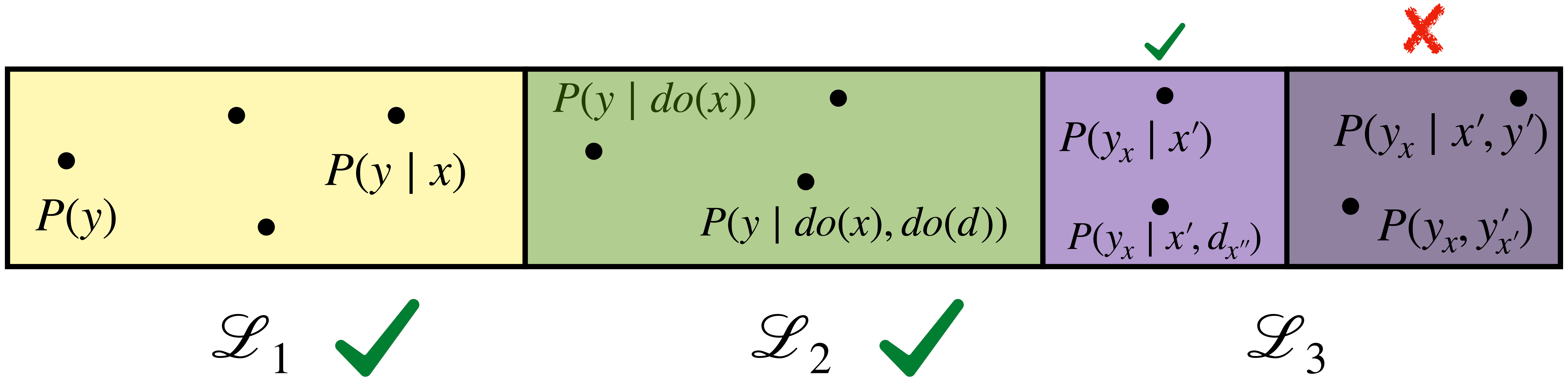
CTF-REALIZE algorithm:

- Input: causal diagram $\mathcal{G}$, list of feasible actions an agent can physically perform, $\mathcal{L}_3$ distribution $P(\mathbf{W}_\star)$.
- I.i.d sample from $P(\mathbf{W}_\star)$ if and only if the distribution is *realizable*.

Graphical criterion:

- If the feasible action set is "maximal", the distribution is realizable if and only if the following condition is met

$$\nexists W_{\mathbf{t}}, W_{\mathbf{z}} \in An_{\mathcal{G}}(\mathbf{W}_\star) \text{ where } \mathbf{t} \neq \mathbf{z}$$



Contribution #2

Realizability: formal definition of the ability to physically draw samples from a distribution in an environment.

CTF-REALIZE algorithm:

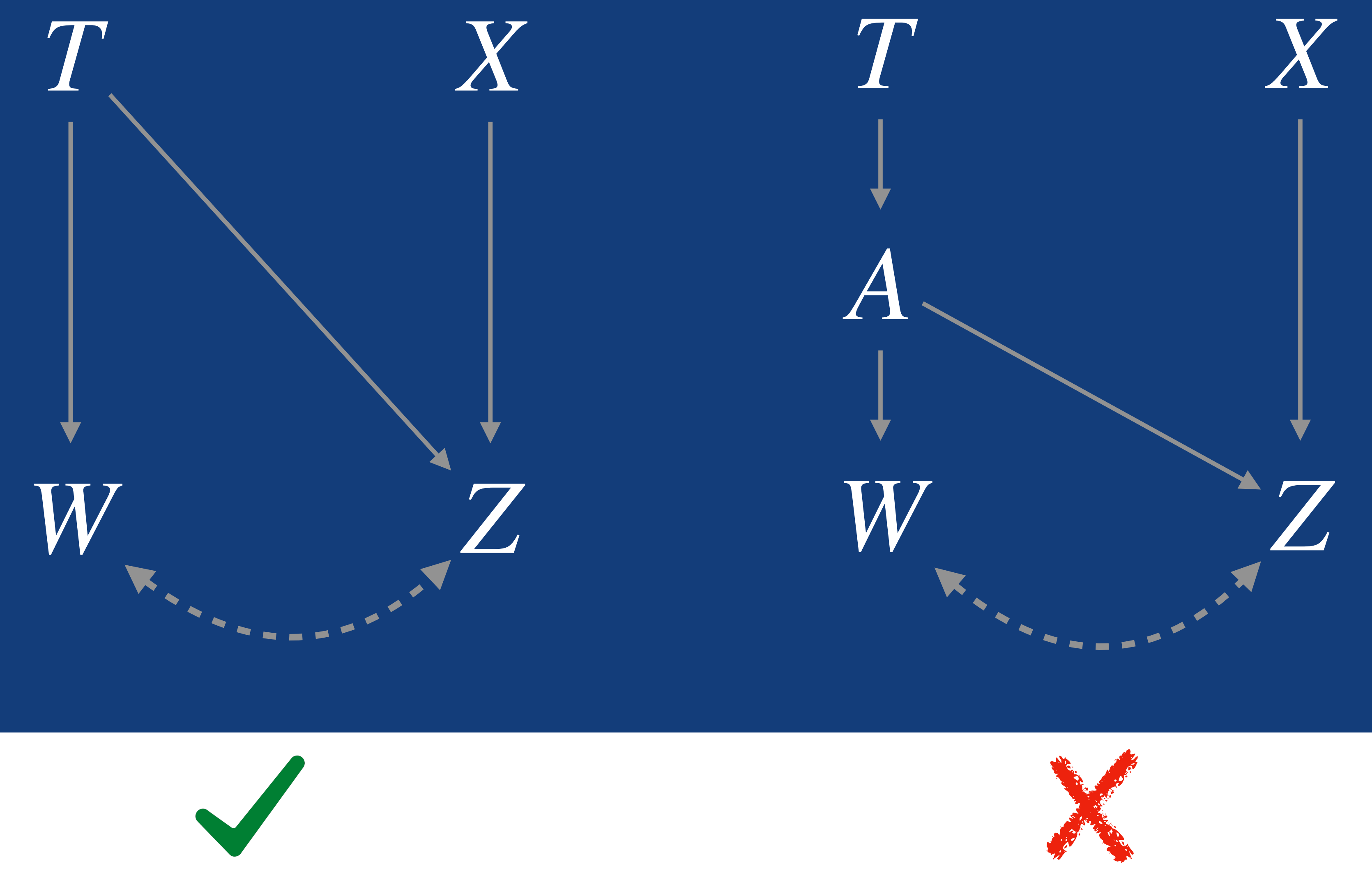
- Input: causal diagram $\mathcal{G}$, list of
- I.i.d sample from $P(\mathbf{W}_\star)$ if and

Graphical criterion:

- If the feasible action set is "ma

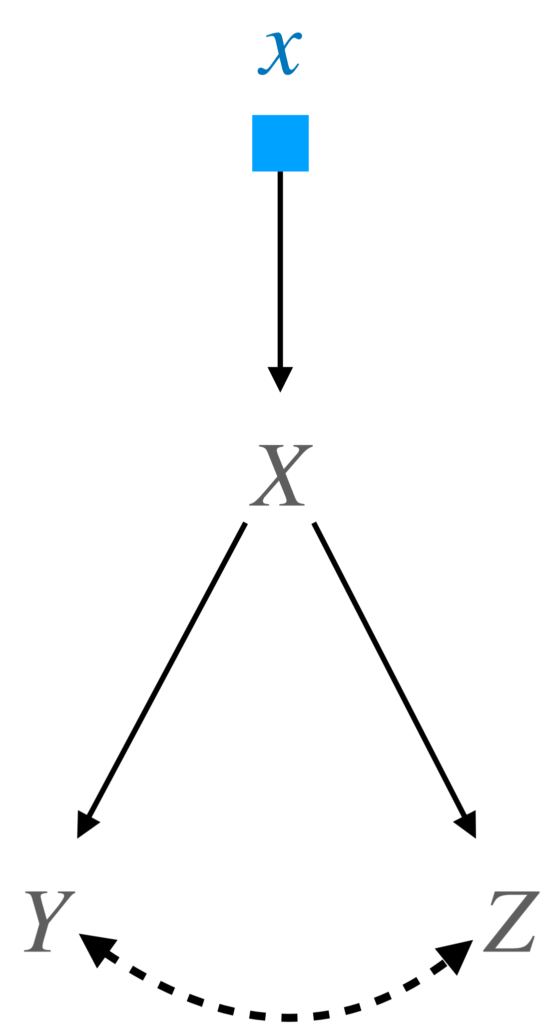
Subsumes the so-called *fundamental problem of causal inference, or FPCI* (Holland '86) as a special case.

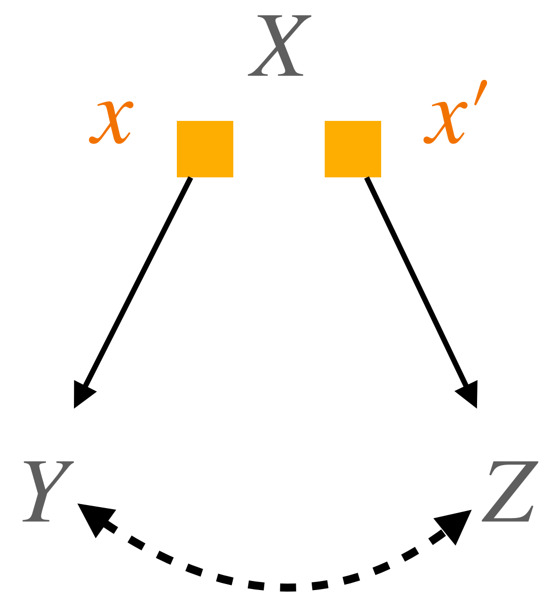
Example from paper:



Contribution #3

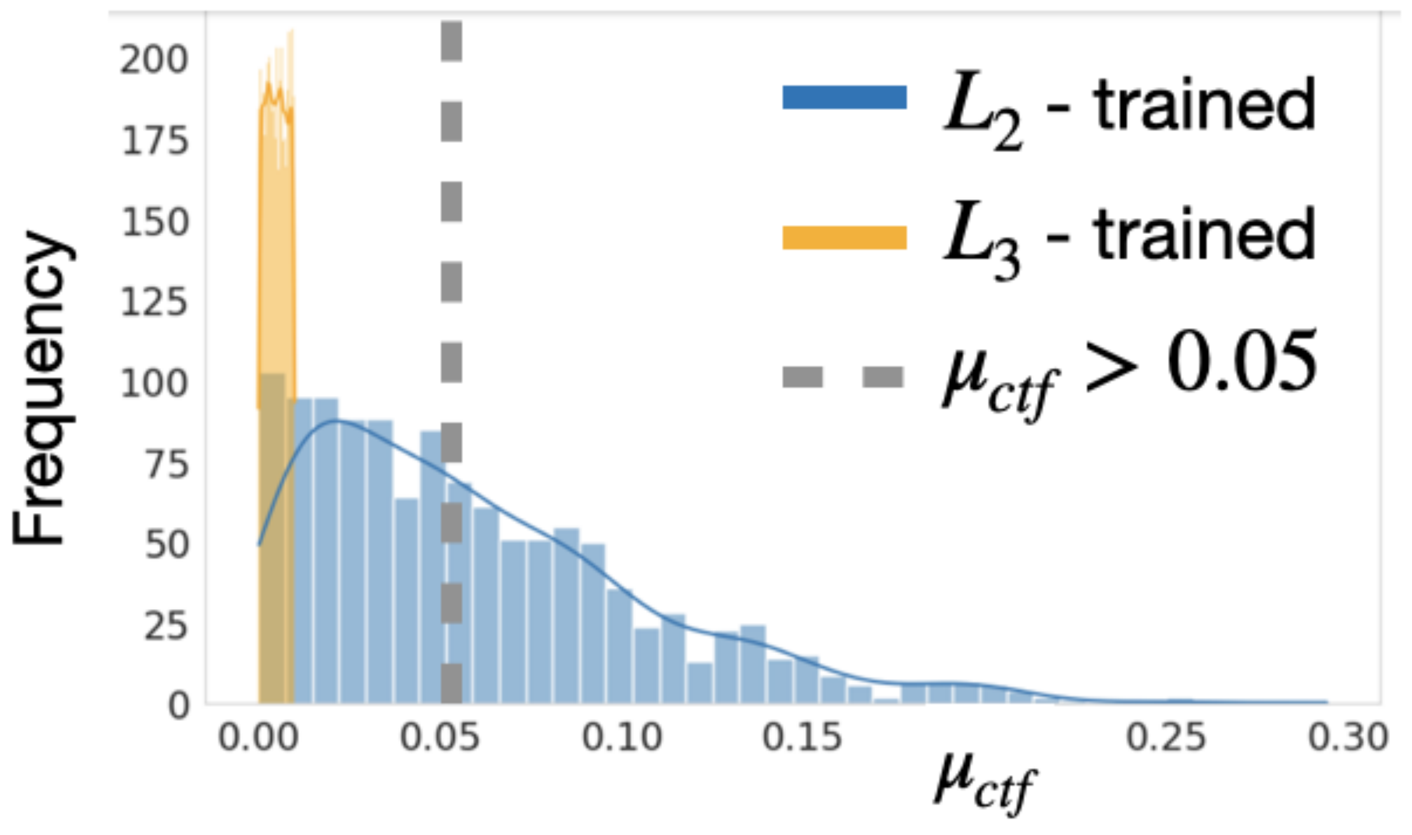
We demonstrate the relevance of counterfactual realizability using examples from **causal fairness** and **causal reinforcement learning**.





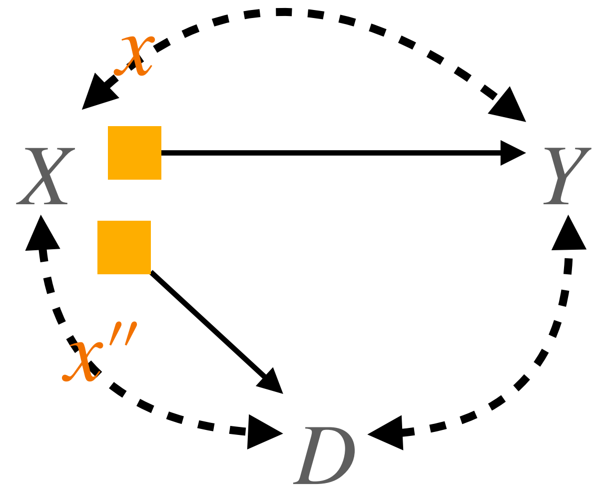
Causal Fairness Analysis:

- Experiment involving CVs being screened for college admission.
- Interventional method does not guarantee fairness in 50% of simulations.
- (Realizable) counterfactual method almost always guarantees fair outcomes.



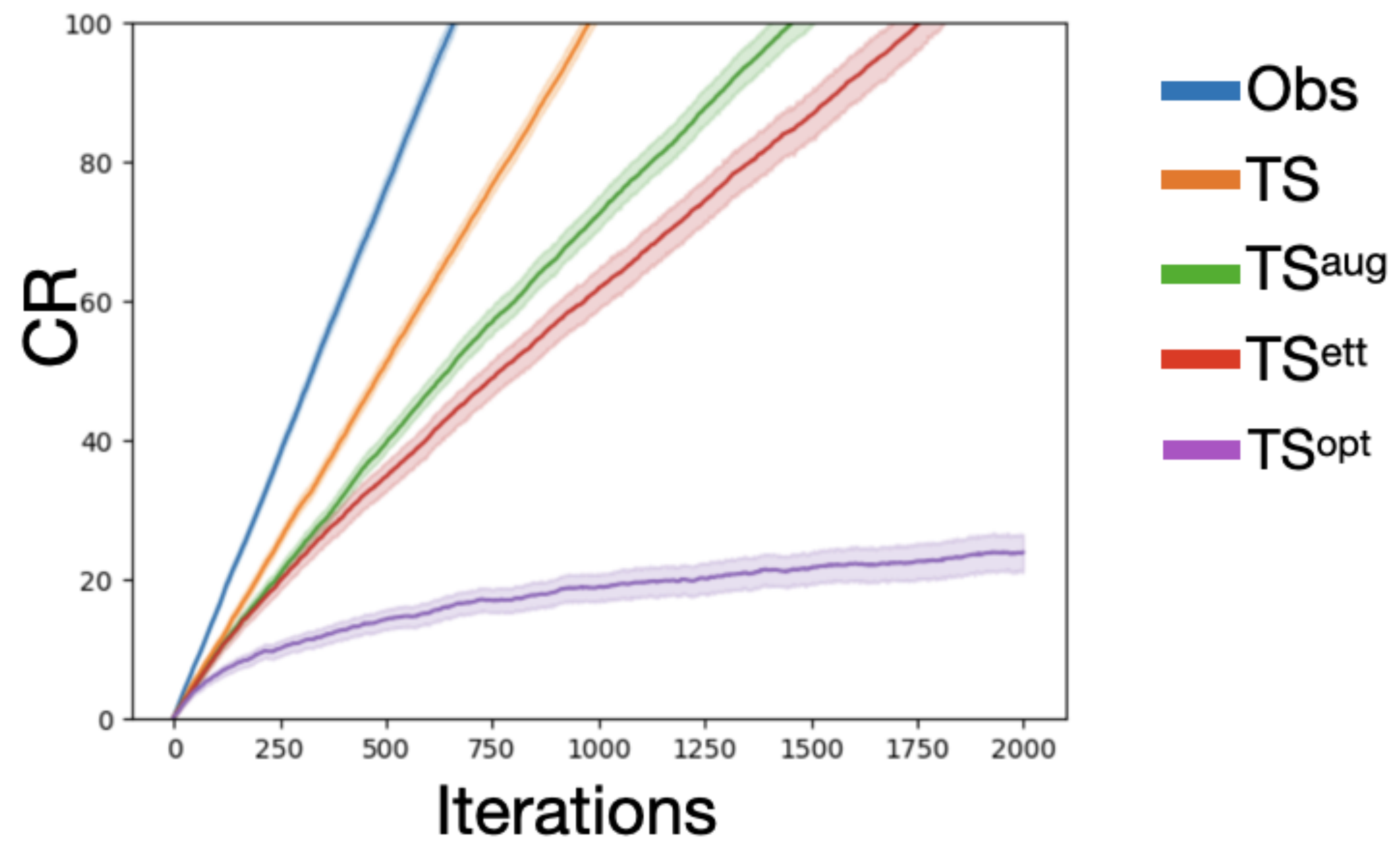
Contribution #3

We demonstrate the relevance of counterfactual realizability using examples from **causal fairness** and **causal reinforcement learning**.



Causal Reinforcement Learning:

- Bandit problem involving adversarial latent confounding.
- Counterfactual randomization provably outperforms observational, interventional, and previous counterfactual benchmarks (in terms of cumulative regret across all rounds).



Why is this important?

Fundamental limits:

- Reveals something foundational about the limit of our ability to learn about a system through black-box experimentation.

Sub-optimality:

- Ignoring this possibility could lead to sub-optimal performance (as seen in the cumulative regret of following an RL strategy without leveraging counterfactual randomisation).

Fairness, explanation, mediation:

- Counterfactual randomisation extends the reach of the experimenter in computing quantities like NDE (example in the paper), even when it is non-identifiable per the causal diagram.
- Relying solely on $\mathscr{L}_2$ data can also be misleading about the fairness of a system (as seen in the example of using only interventional measures to reason about fairness).

Thank You