Causal discovery and Deep Learning

Faria et. al. 2022; Pearl et. al. 2018; Neal et. al. 2022; Schölkopf et. al. 2019; Bengio et. al. 2020; Peters et. al. 2017, Ke et. al. 2020, Brouillard et. al. 2020

Gonçalo R. A. Faria
@goncalorafaria

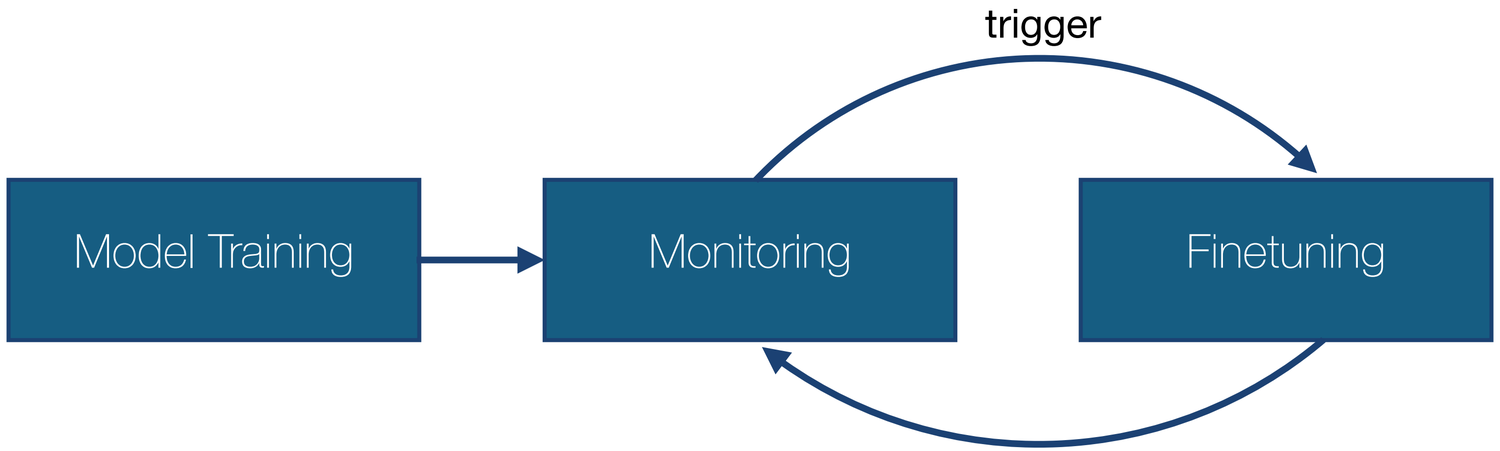
Deep Learning Sessions Portugal

Overview

- Limitations of current Deep Learning systems
- Causal Discovery
- Identifiability
- Interventions and Causal Discovery
- Latent interventions and Causal Discovery

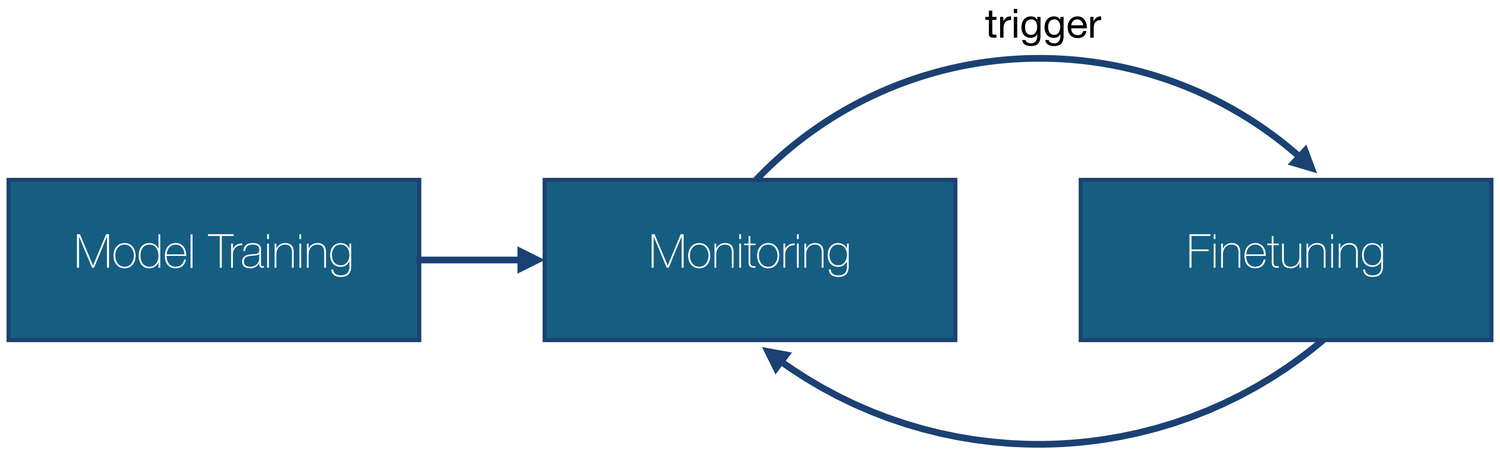
Limitations of current Deep Learning

- Distribution shifts.



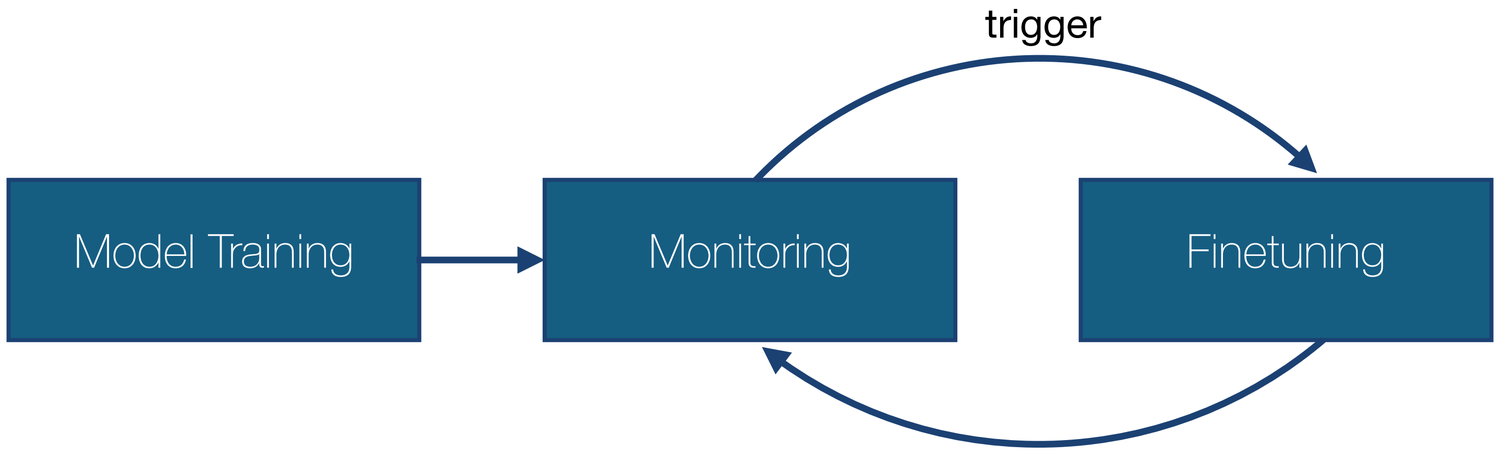
Limitations of current Deep Learning

- Distribution shifts.
- Model deployment drives distribution shifts.



Limitations of current Deep Learning

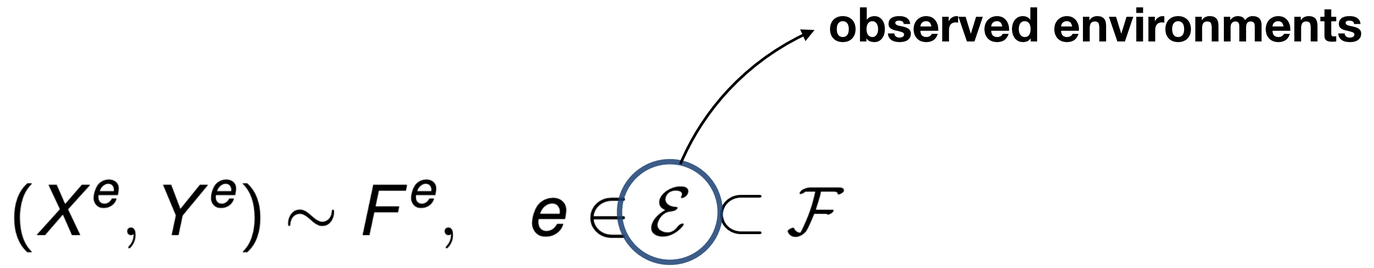
- Distribution shifts.
- Model deployment drives distribution shifts.
- Test-set performance vs generalization.
- Interpretability and scientific understanding



Benefits of causal predictors

- Causal predictors are optimally robust.

Consider the setting :



$$\hat{\beta} = \text{argmin}_\beta \max_{e \in \mathcal{F}} \mathbb{E}|Y^e - (X^e)^T \beta|^2$$

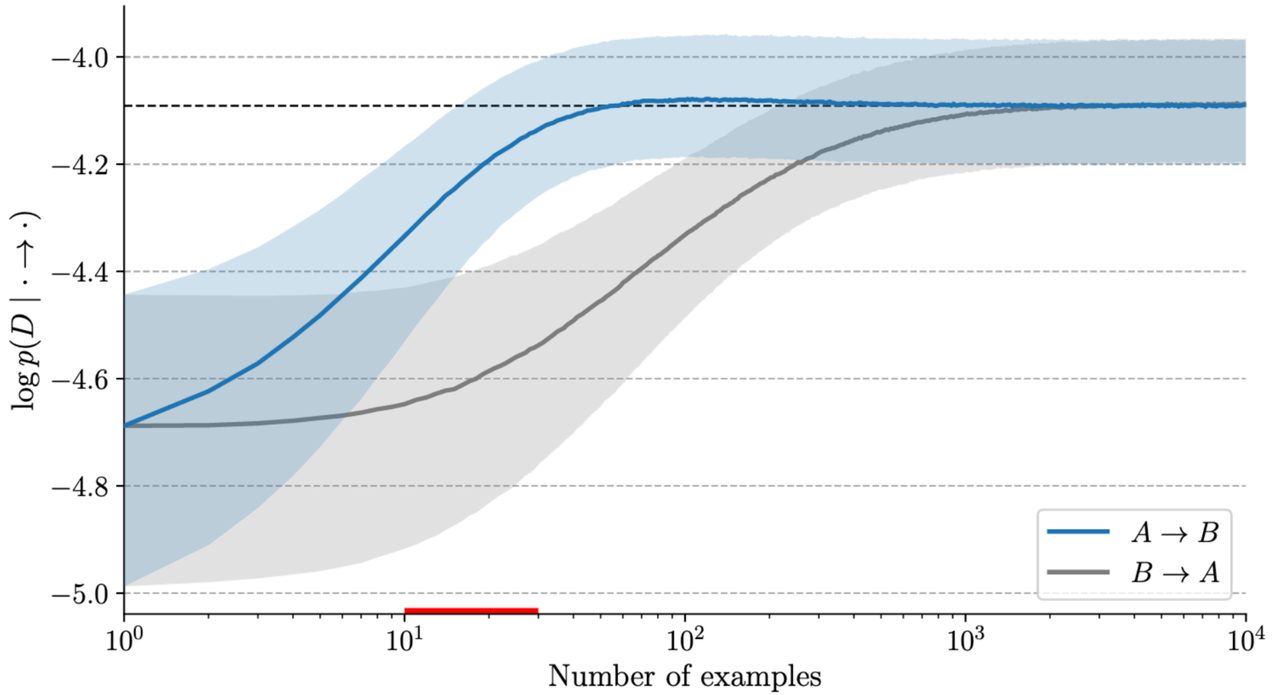
causal predictor

Causality $\Leftrightarrow$ **Prediction in heterogeneous scenarios/environments**

Causal inference using invariant prediction (Peters et. al. 2016)

Benefits of causal predictors

- Causal predictors are optimally robust.
- Fast and data efficient adaptation to transfer distribution.

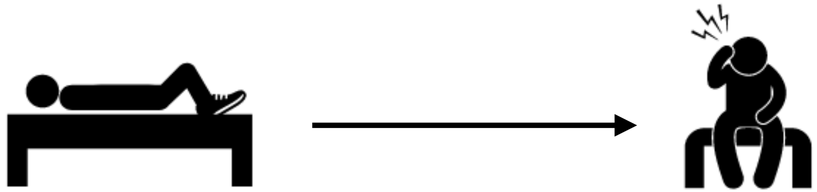


A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms (Bengio et. al. 2019)

What if we do not know the causal model ?

Causal Discovery

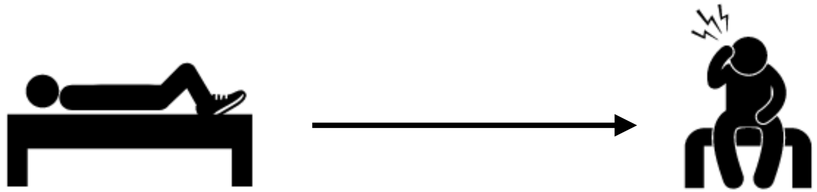
- Say we have a dataset.



Introduction to Causal Inference (Neal et. al. 2022)

Causal Discovery

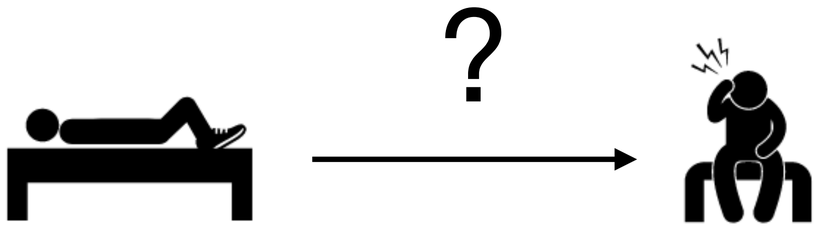
- Say we have a dataset.
- Cause-effect relation?



Introduction to Causal Inference (Neal et. al. 2022)

Causal Discovery

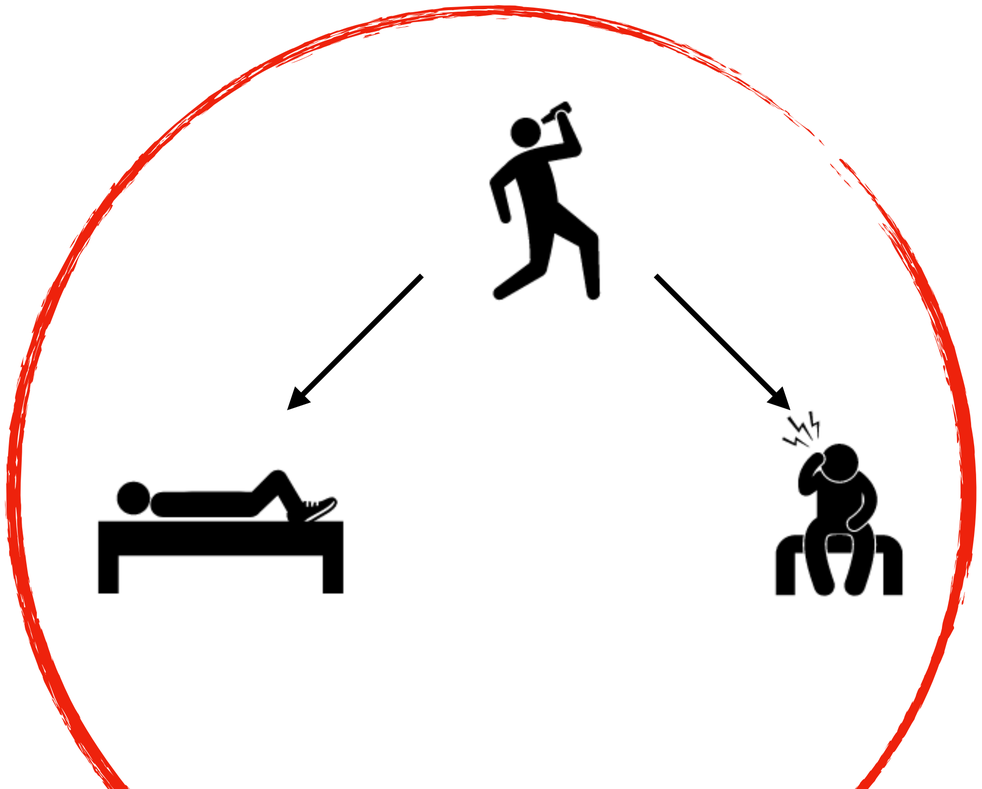
- Say we have a dataset.
- Cause-effect relation?
- **Correlation is not causation**



Introduction to Causal Inference (Neal et. al. 2022)

Causal Discovery

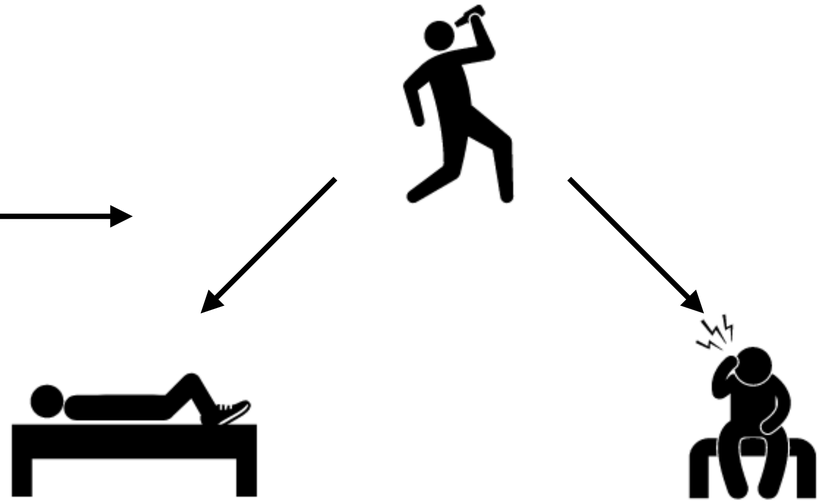
- Say we have a dataset.
- Cause-effect relation?
- **Correlation is not causation**
- Domain-expertise
- Hidden Confounding



Data-driven Causal Discovery

dataset (of observations)

causal discovery method



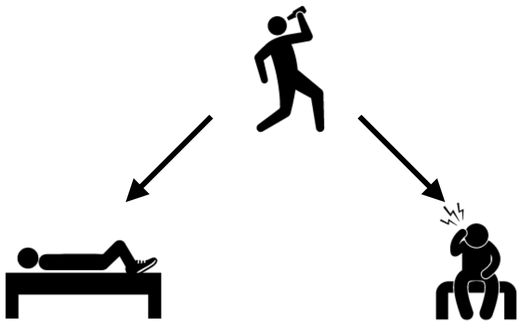
Identifiability

 independent from  given 

Identifiability

independent from

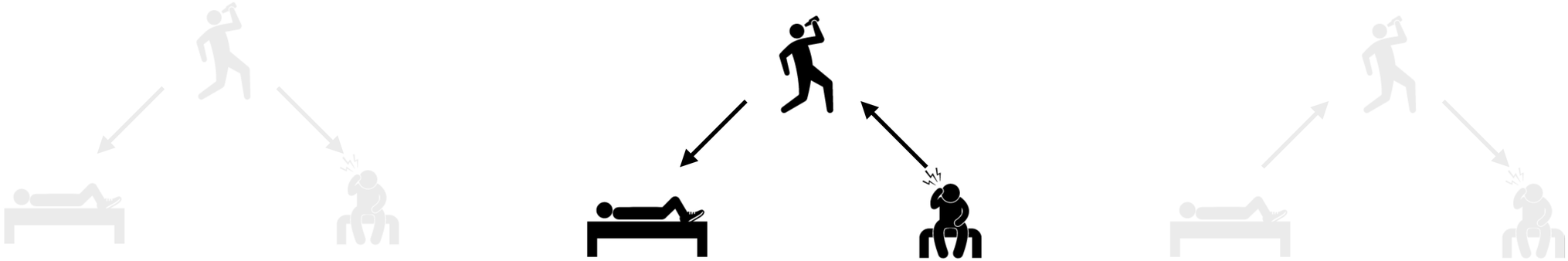
given



Identifiability

independent from

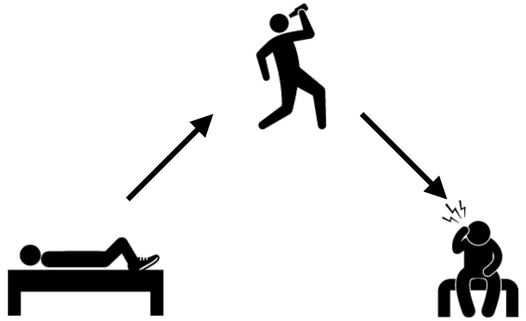
given



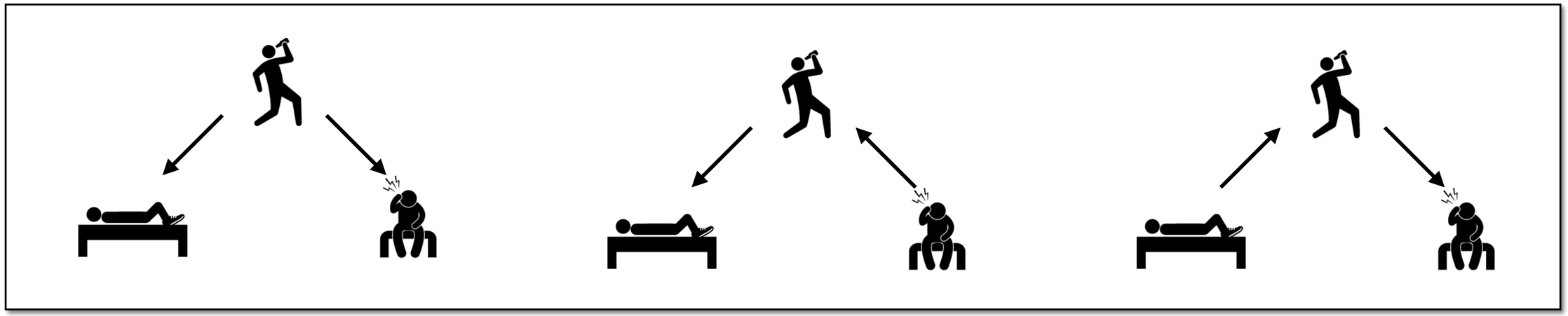
Identifiability

independent from

given



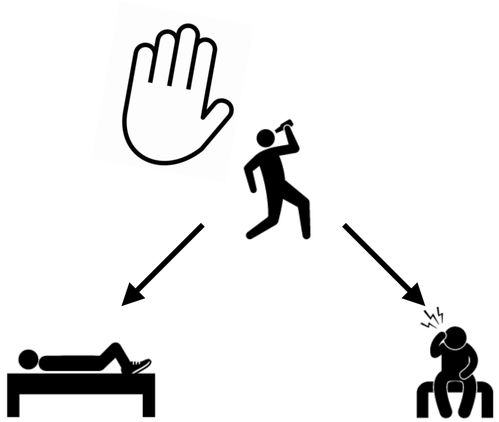
Identifiability



Markov Equivalence class

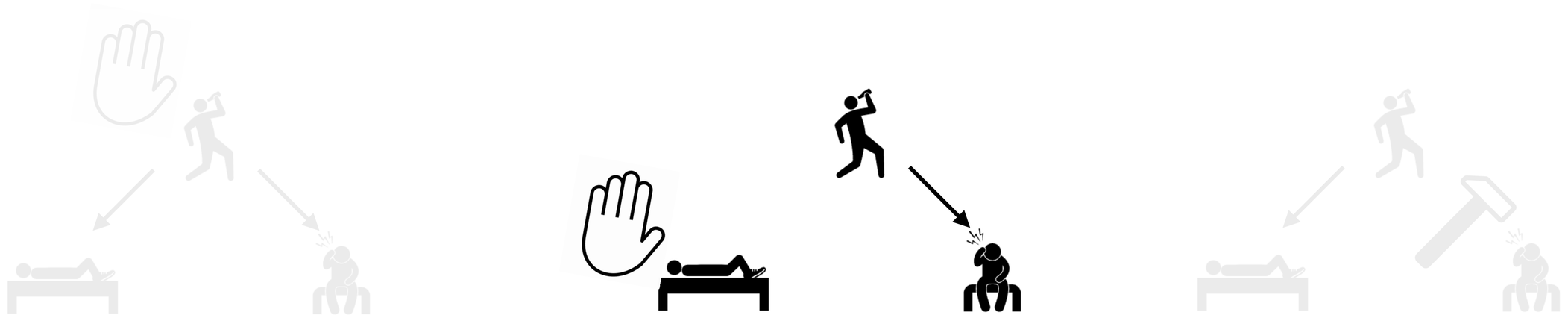
Interventions

- Science; hypotheses; experiments (interventions)



Interventions

- Science; hypotheses; experiments (interventions)

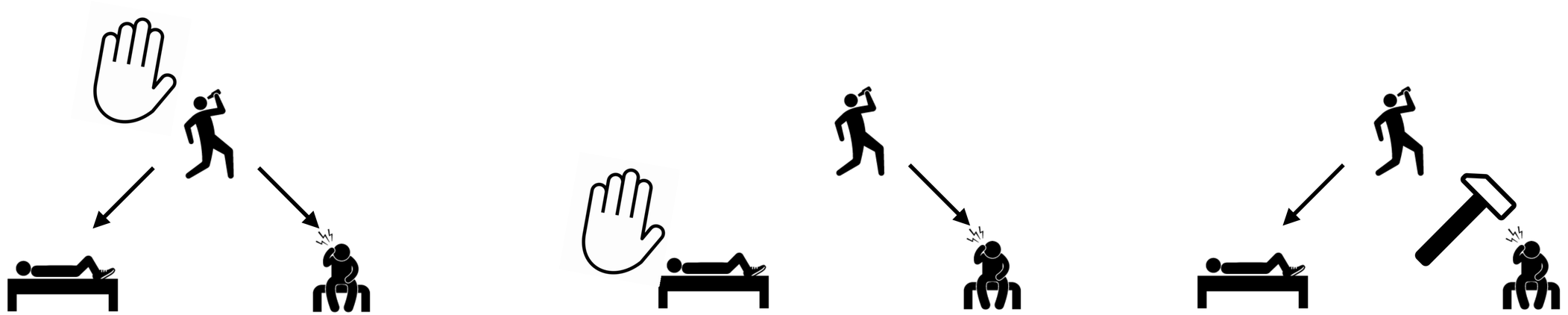


Interventions

- Science; hypotheses; experiments (interventions)

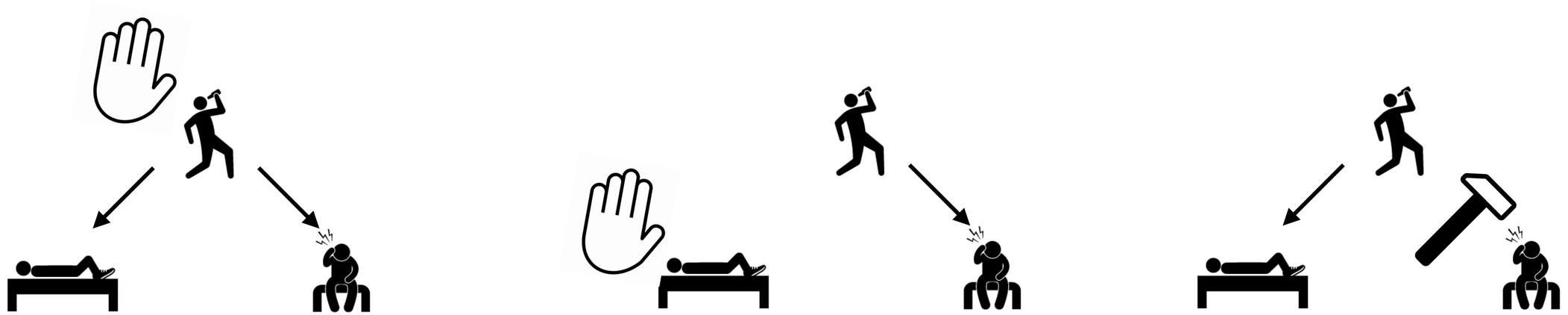
Interventions

- Science; hypotheses; experiments (interventions)
- Causal discovery with intervention data



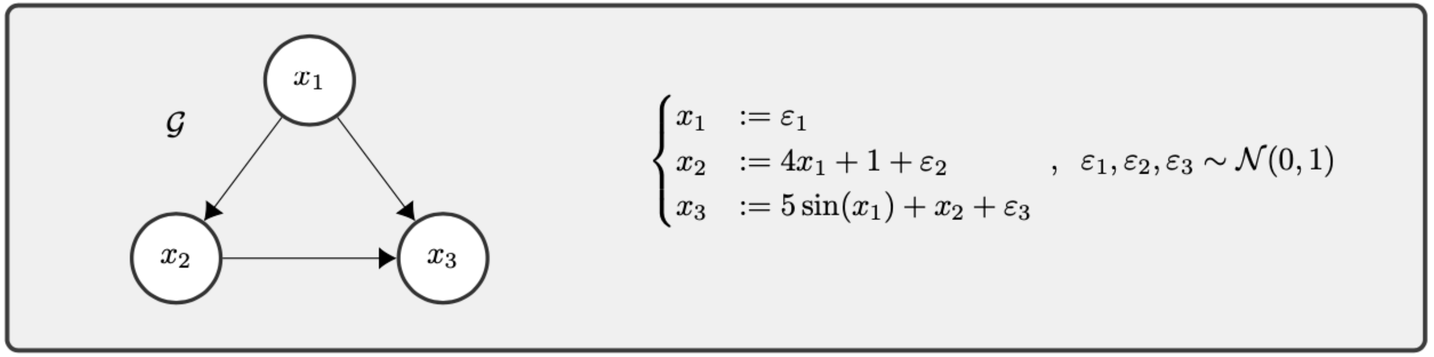
Interventions

- Science; hypotheses; experiments (interventions)
- Causal discovery with intervention data
- **Complex**; **expensive**; **unethical**.

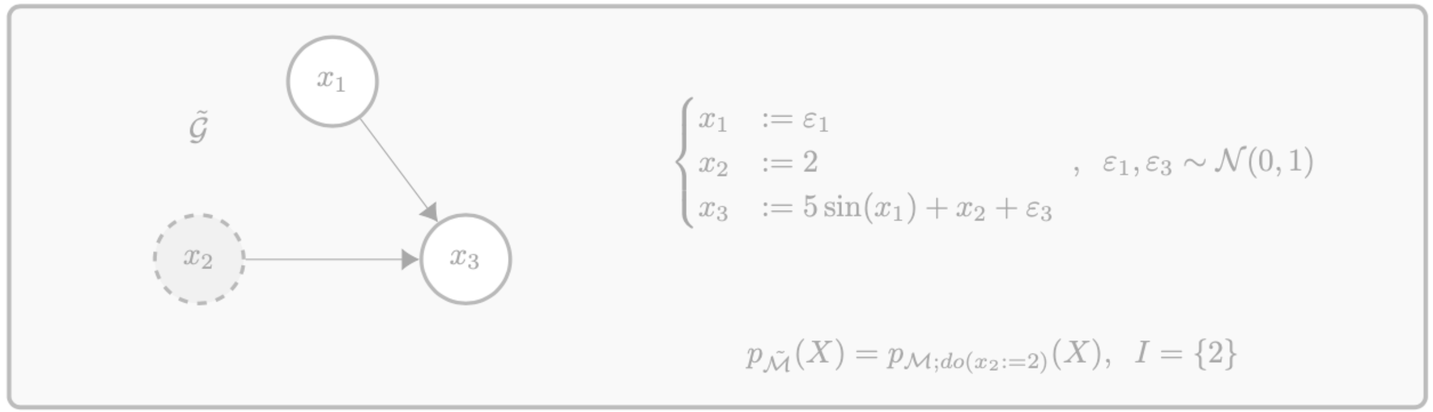


Structural Causal Models

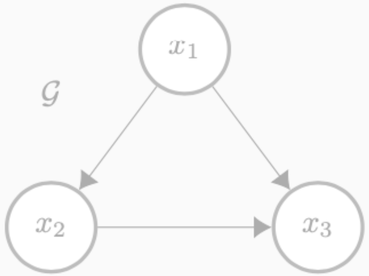
Structural causal models



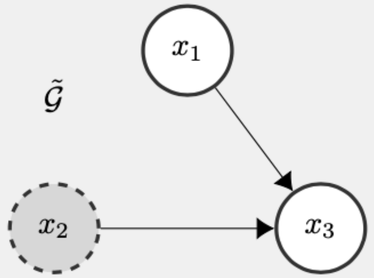
$$p_{\mathcal{M}}(X) = p_{\mathcal{M}}(x_3|x_2, x_1)p_{\mathcal{M}}(x_2|x_1)p_{\mathcal{M}}(x_1)$$



Structural causal models



$$\begin{cases} x_1 & := \varepsilon_1 \\ x_2 & := 4x_1 + 1 + \varepsilon_2 \\ x_3 & := 5\sin(x_1) + x_2 + \varepsilon_3 \end{cases}, \quad \varepsilon_1, \varepsilon_2, \varepsilon_3 \sim \mathcal{N}(0,1)$$



$$\begin{cases} x_1 & := \varepsilon_1 \\ x_2 & := 2 \\ x_3 & := 5\sin(x_1) + x_2 + \varepsilon_3 \end{cases}, \quad \varepsilon_1, \varepsilon_3 \sim \mathcal{N}(0,1)$$

$$p_{\tilde{\mathcal{M}}}(X) = p_{\mathcal{M};do(x_2:=2)}(X), \quad I = \{2\}$$

$$p_{\tilde{\mathcal{M}}}(X) = p_{\mathcal{M}}(x_3|x_2,x_1)p_{\mathcal{M};\mathrm{do}(x_2:=2)}(x_2)p_{\mathcal{M}}(x_1)$$

Structural causal models

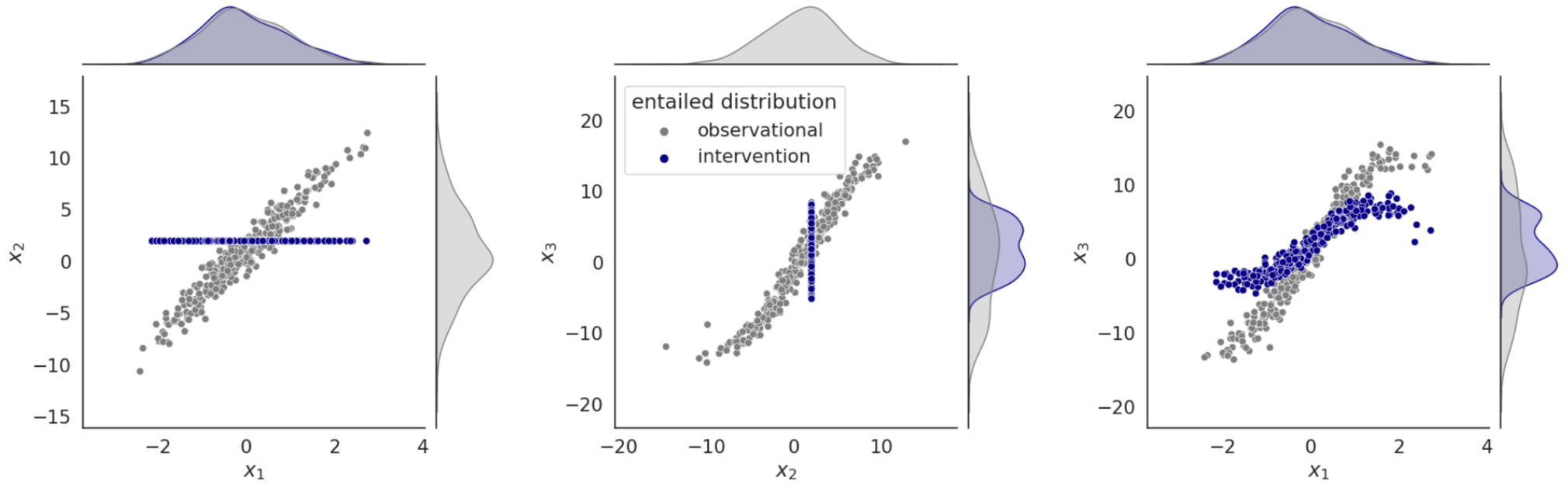
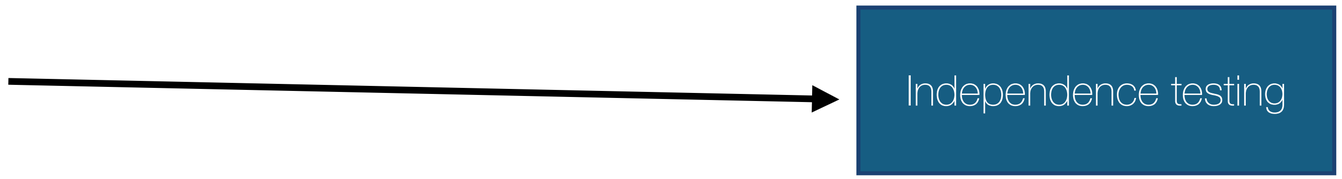


Figure 2.5: Samples from the observational SCM in Figure 2.2, and intervention SCM in Figure 2.4.

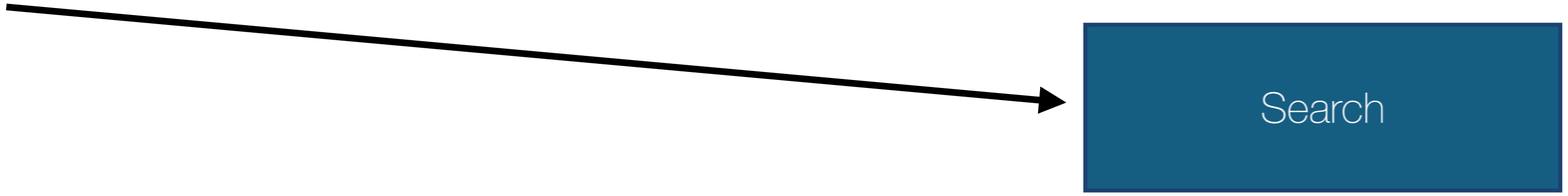
Causal Discovery Methods

- Constrained-Based
- Score based
- Hybrid
- Others



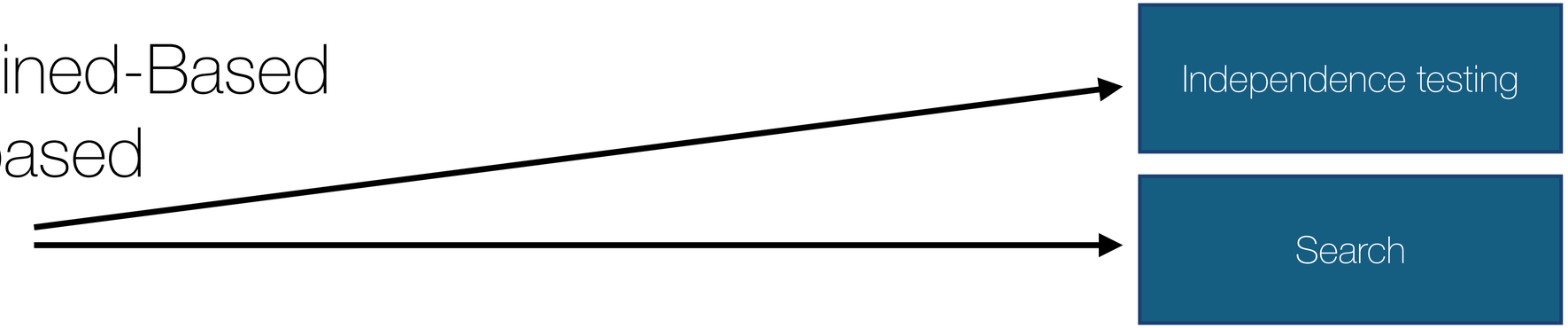
Causal Discovery Methods

- Constrained-Based
- Score based
- Hybrid
- Others



Causal Discovery Methods

- Constrained-Based
- Score based
- Hybrid
- Others



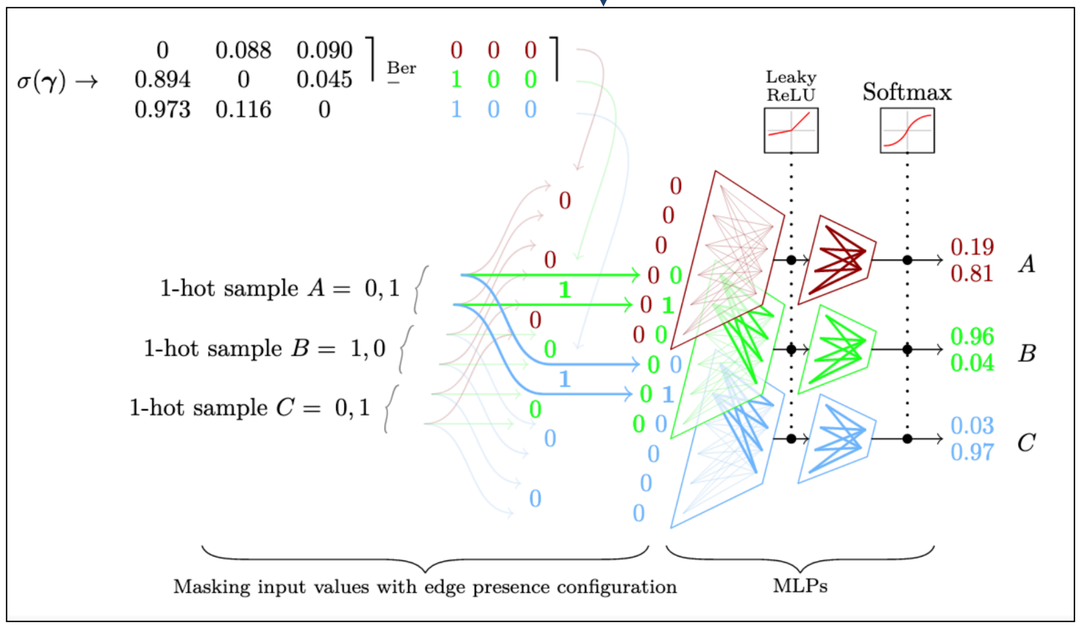
Score based methods with deep learning

$$\Lambda^* = \arg\max_{\Lambda} S(\Lambda)$$

$$S(\Lambda) = \max_{\theta} \mathbb{E}_{\mathcal{G}\sim \mathrm{DAG}(\mathcal{G};\Lambda)} \left[\log p(\mathcal{D}|\mathcal{G},\theta) + \log p(\mathcal{G}) \right]$$

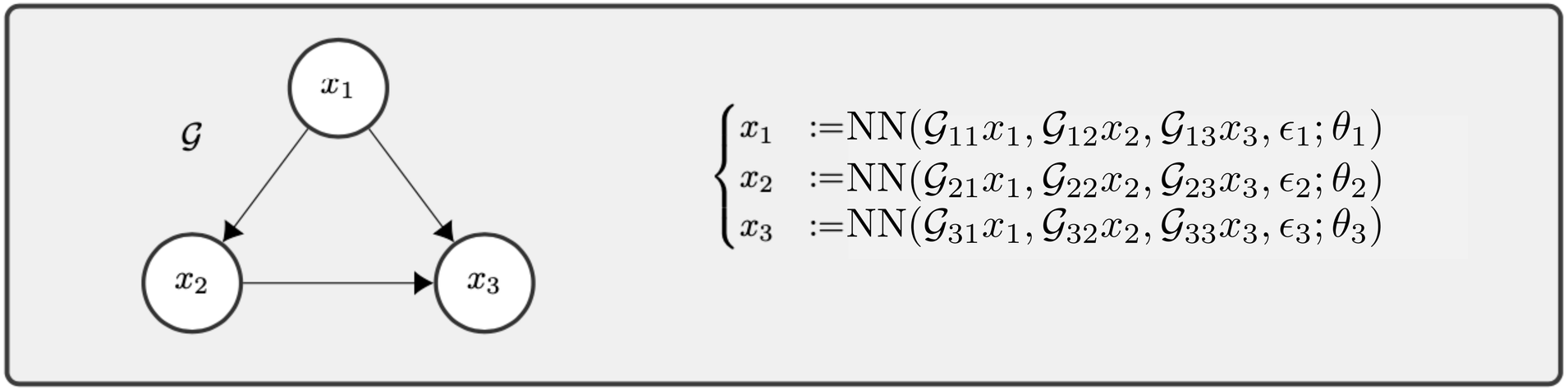
(most of the time) approximated w/ fully factorized Bernoulli

Penalize dense graphs/enforce acyclicity



Ke et. al. 2020

Score based methods with deep learning



Similar to doing LASSO for all of the variables at once w/ acyclicity constraints.

Ke et. al. 2020

Interventions and Causal Discovery

Causal discovery with intervention data

$$\Lambda^*, \tilde{\mathcal{M}}^* = \arg\max_{\Lambda, \tilde{\mathcal{M}}} S(\Lambda, \tilde{\mathcal{M}})$$

$$S(\Lambda, \tilde{\mathcal{M}}) = \max_{\theta, \phi} \mathbb{E}_{\mathcal{G} \sim \mathrm{DAG}(\mathcal{G}; \Lambda)} \Big[\mathbb{E}_{k \sim p(k)} \big[\log p(\mathcal{D}^k | \mathcal{G}, \theta^k, \tilde{\mathcal{M}}^{(k)}) + \log p(\mathcal{G}) \big] \Big]$$

Intervention specific parameters

Intervention variables

Brouillard et. al. 2020

Differentiable Causal Discovery Under Latent Interventions



Gonçalo R. A. Faria André F. T. Martins Mário A. T. Figueiredo

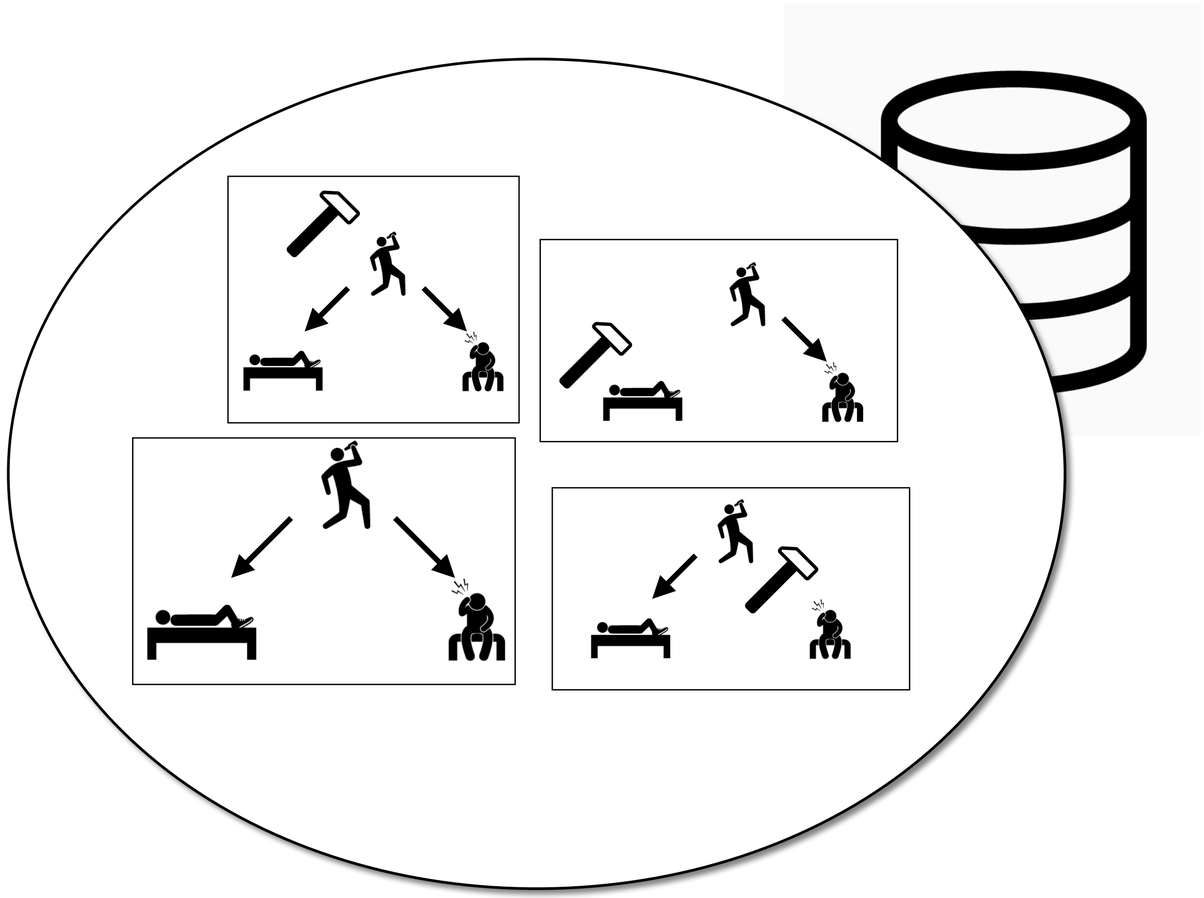
instituto de telecomunicações

Unbabel

TÉCNICO LISBOA

ellis

Latent interventions

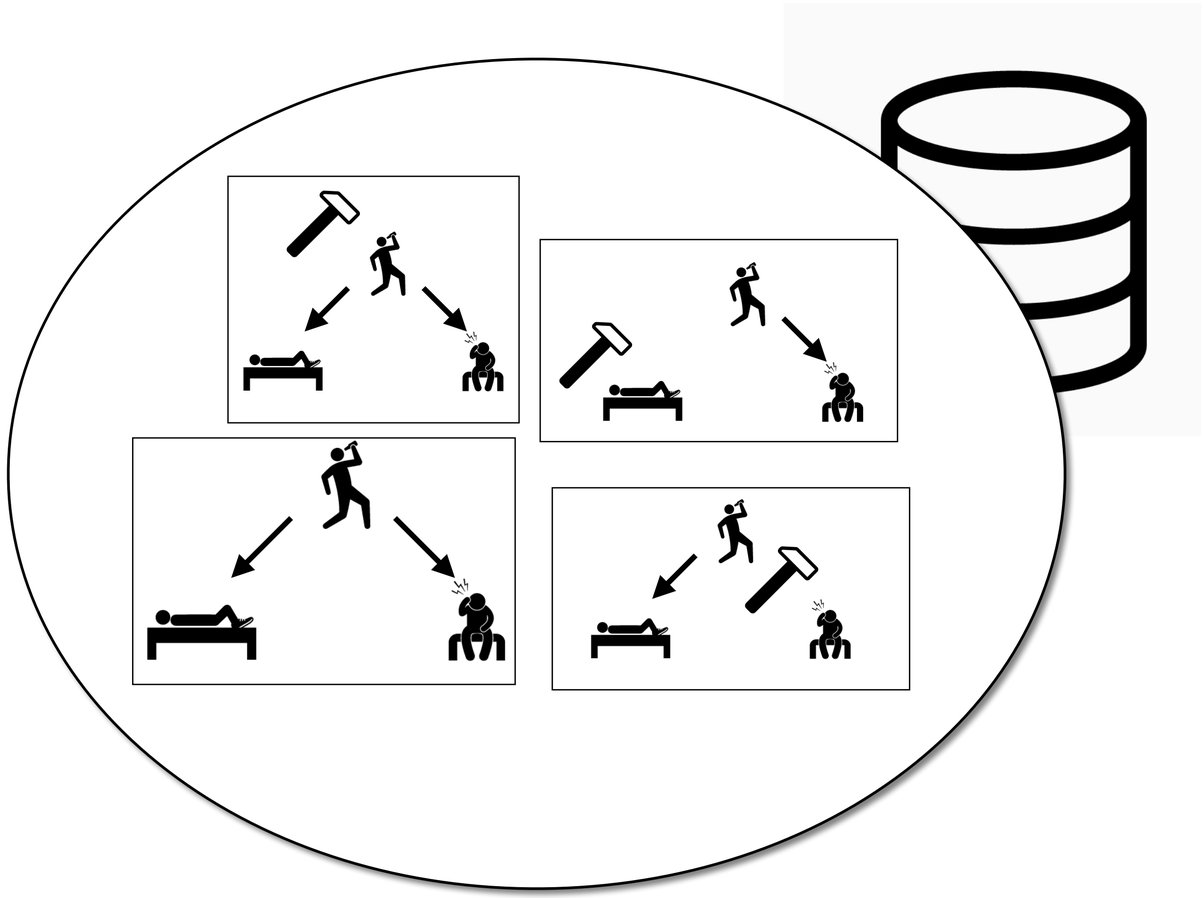


Mixture of experimental regimes.

For each sample:

- do not know correspondence to intervention regime;

Latent interventions



Mixture of experimental regimes.

For each sample:

- do not know correspondence to intervention regime;

For each intervention:

- do not know experimental conditions

Can we recover latent interventions?

Intervention recovery

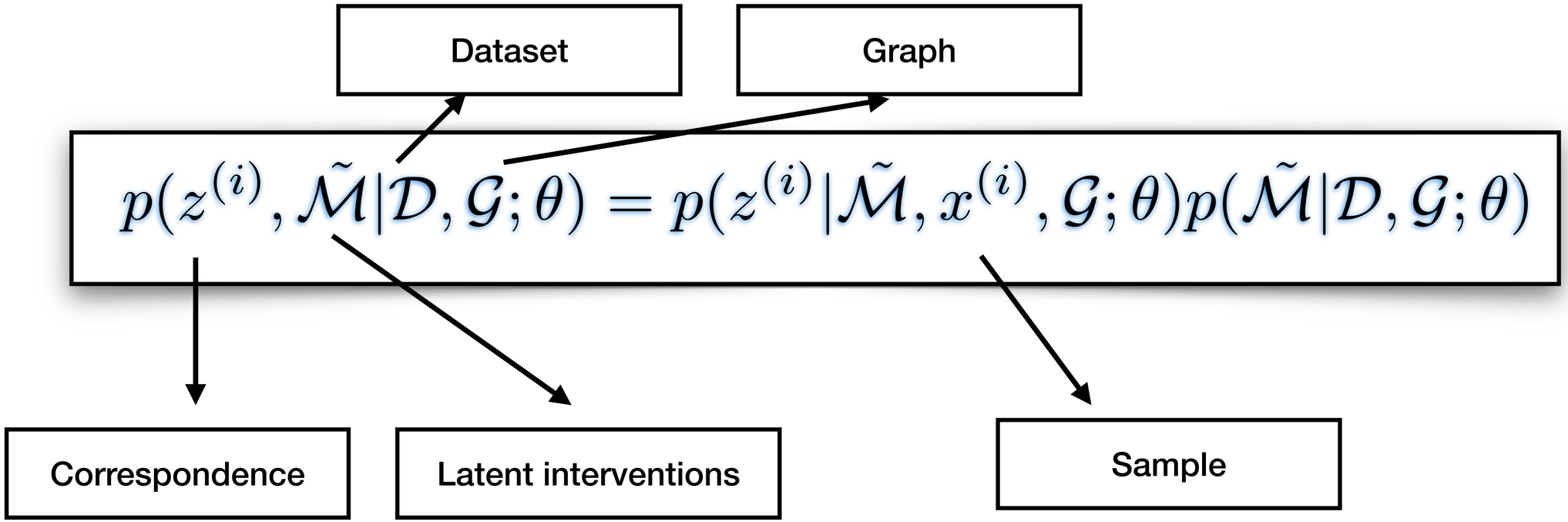
- Given the causal graph.
- Recover interventions and correspondences.
- Propose joint distribution

Infinite mixture of intervention SCMs

+

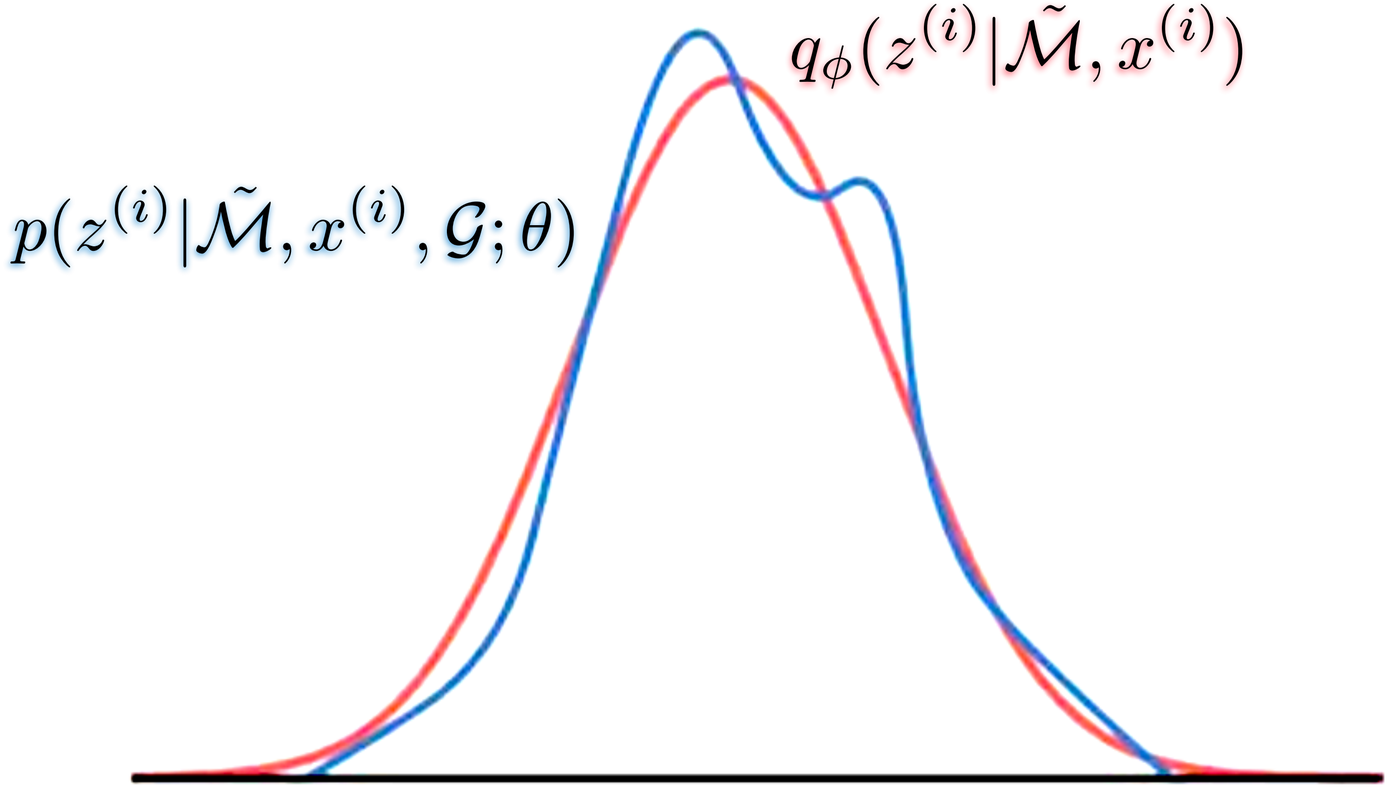
Prior distribution for the interventions
(Dirichlet process)

Approximate posterior inference



Approximate posterior inference

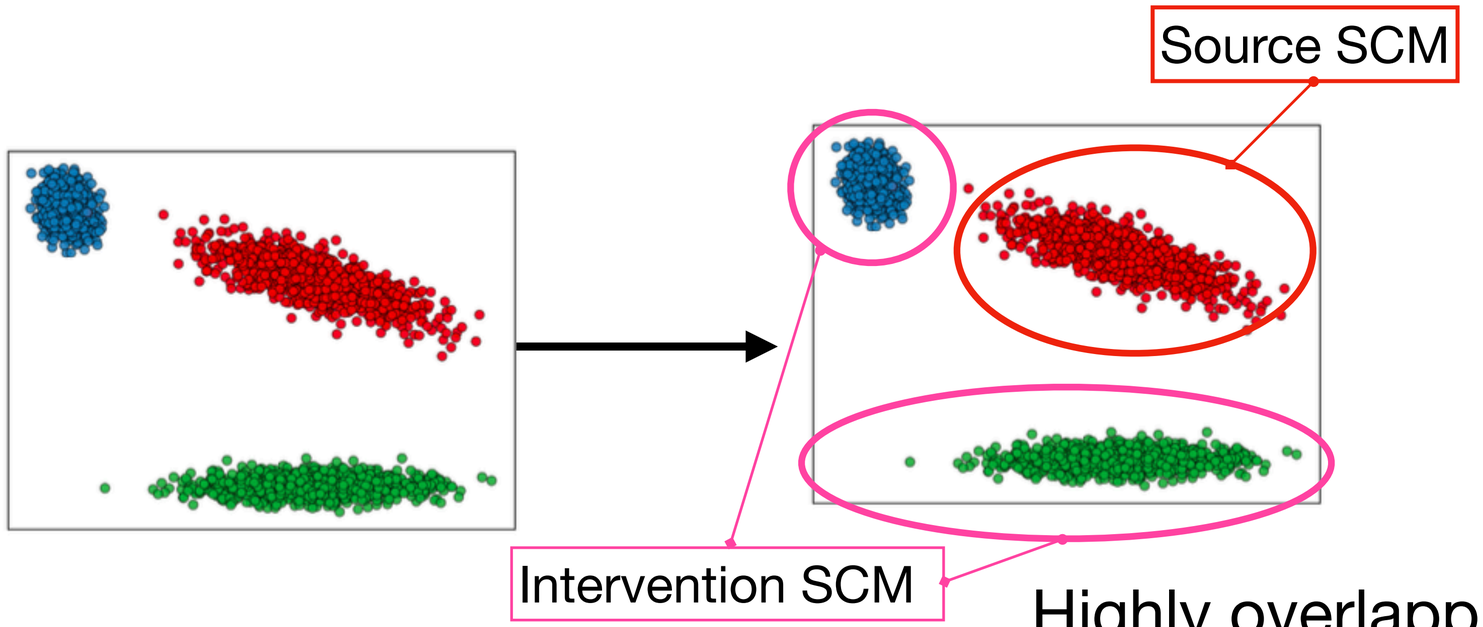
- Variational inference



Approximate posterior inference

- Variational inference

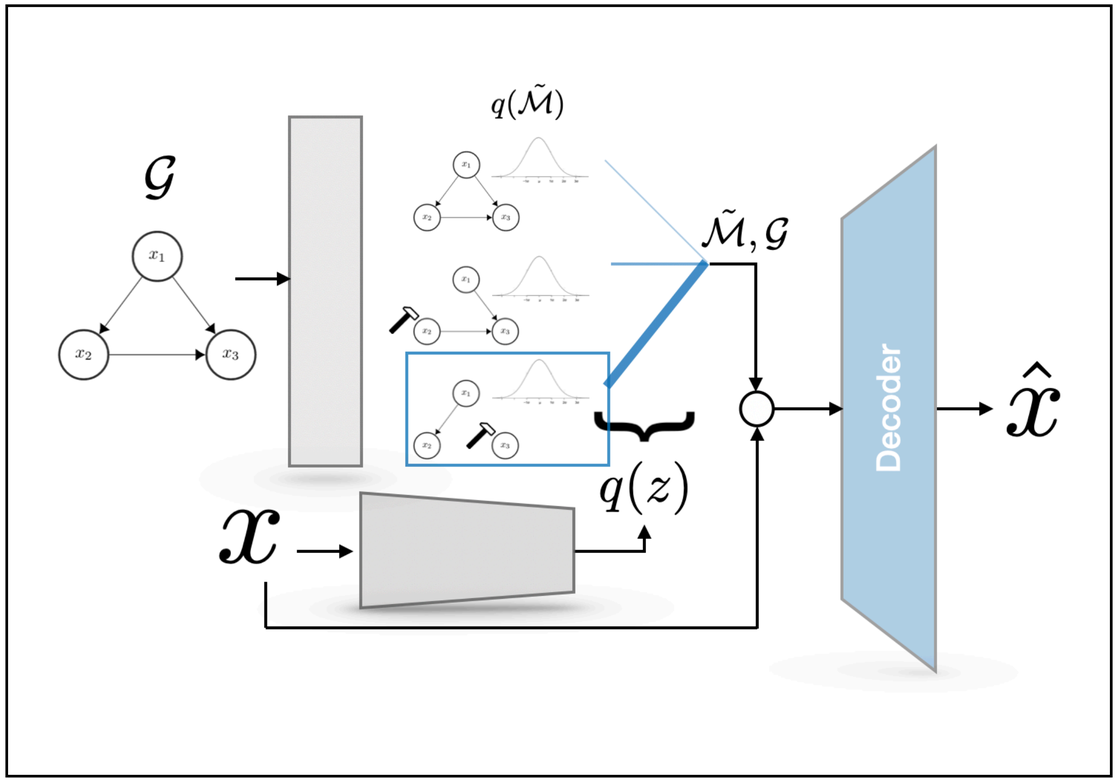
Similar to a clustering problem
(each cluster as few degrees of freedom)



Highly overlapping clusters
(not that noticeable in this 2D example)

Approximate posterior inference

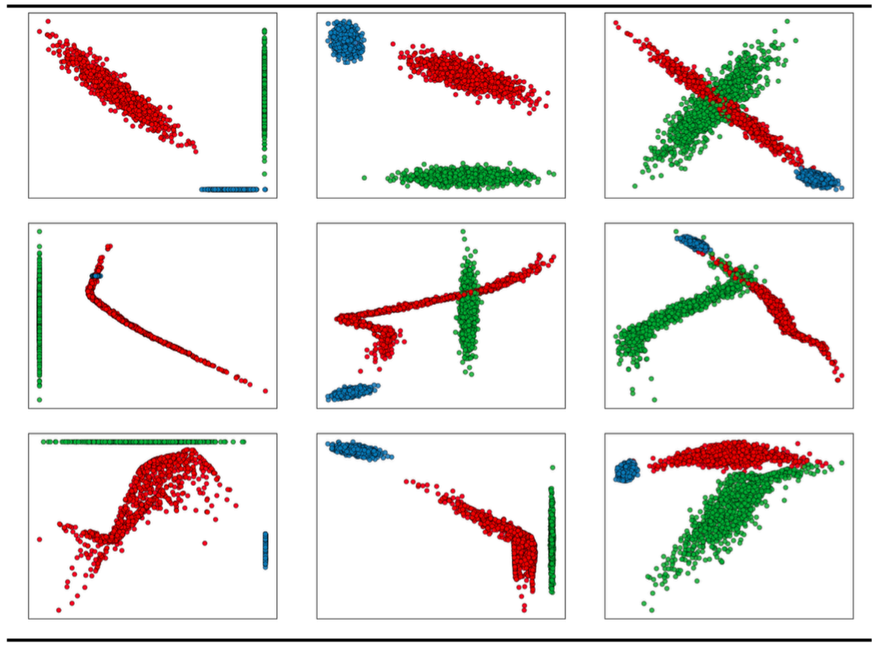
$$\max_{\theta,\phi} \sum_{i=1}^{N} \text{ELBO}_{q(z^{(i)},\tilde{\mathcal{M}};\phi)}\left(x^{(i)}, \mathcal{G}; \theta\right)$$



Experiments with single node interventions

Model Type	d = 10		d = 20	
	ER1	ER4	ER1	ER4
Stochastic Interventions:				
Linear Gaussian	0.9816 ± 0.019	0.9492 ± 0.007	0.9815 ± 0.005	0.9785 ± 0.006
Non-Linear Gaussian	0.9643 ± 0.015	0.9242 ± 0.006	0.9655 ± 0.005	0.9431 ± 0.003
Non-Linear Non-Gaussian	0.8314 ± 0.062	0.7812 ± 0.1074	0.8929 ± 0.027	0.8412 ± 0.087
Imperfect Interventions:				
Linear Gaussian	0.9262 ± 0.030	0.8802 ± 0.040	0.8072 ± 0.098	0.7819 ± 0.054
Non-Linear Gaussian	0.9541 ± 0.012	0.9205 ± 0.005	0.9670 ± 0.007	0.9434 ± 0.004
Non-Linear Non-Gaussian	0.9103 ± 0.004	0.9090 ± 0.004	0.9345 ± 0.005	0.9196 ± 0.005
Atomic Interventions:				
Linear Gaussian	0.9141 ± 0.0018	0.9148 ± 0.0021	0.9507 ± 0.0059	0.9461 ± 0.0030
Non-Linear Gaussian	0.9138 ± 0.0053	0.9195 ± 0.0041	0.9506 ± 0.0056	0.9435 ± 0.0032
Non-Linear Non-Gaussian	0.9100 ± 0.0056	0.9034 ± 0.0115	0.9314 ± 0.0104	0.9306 ± 0.0041

Table 3.2: Results for the single-node intervention experiment. The metric used is the average rand index.

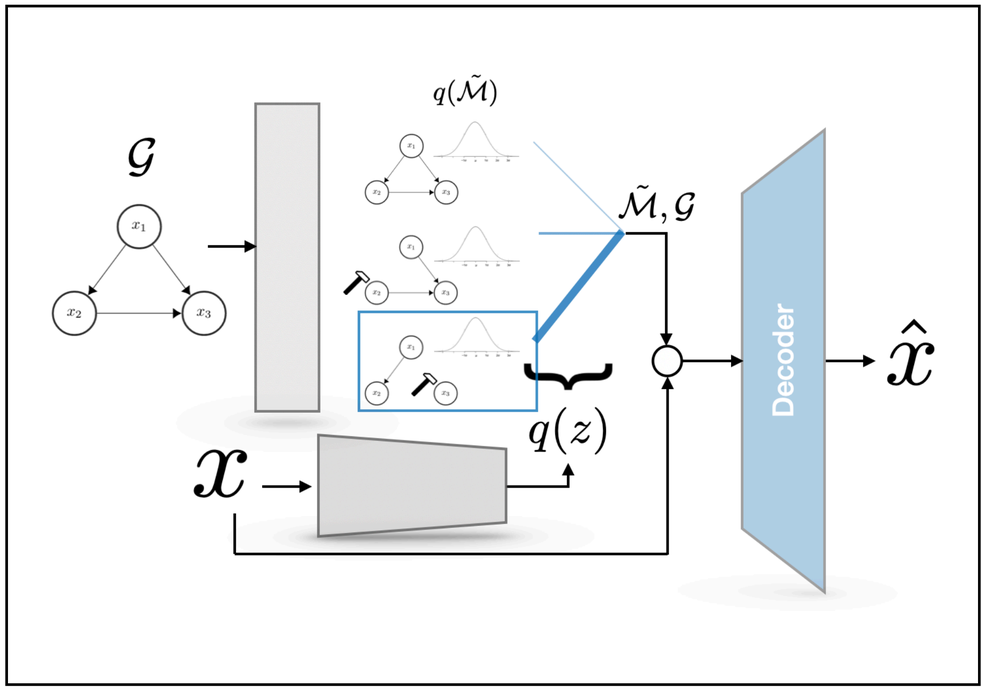


Can we recover the causal graph?

Score based method

$$\Lambda^* = \arg\max_{\Lambda} S(\Lambda)$$

$$S(\Lambda) = \max_{\theta,\phi} \mathbb{E}_{\mathcal{G}\sim \mathrm{DAG}(\mathcal{G};\Lambda)} \Big[\sum_{i=1}^{N} \mathrm{ELBO}_{q(z^{(i)},\tilde{\mathcal{M}};\phi)} \left(x^{(i)}, \mathcal{G}; \theta\right) + \log p(\mathcal{G}) \Big]$$



Model variants

- **“latent”** (new)

For each sample:

- do not know correspondence to intervention regime;

For each intervention:

- do not know experimental conditions

Model variants

- “latent” (new)
- **“unknown”**

For each sample:

- ~~do not~~ know correspondence to intervention regime;

For each intervention:

- do not know experimental conditions

Model variants

- “latent” (new)
- “unknown”
- **“known”**

For each sample:

- ~~do not~~ know correspondence to intervention regime;

For each intervention:

- ~~do not~~ know experimental conditions

Model variants

- “latent” (new)
- “unknown”
- “known”
- **“observational”**

assume there is no intervention data.

Model variants

- “latent” (new)
- “unknown”
- “known”
- “observational”
- **“semi-supervised”** (new)

Model variants

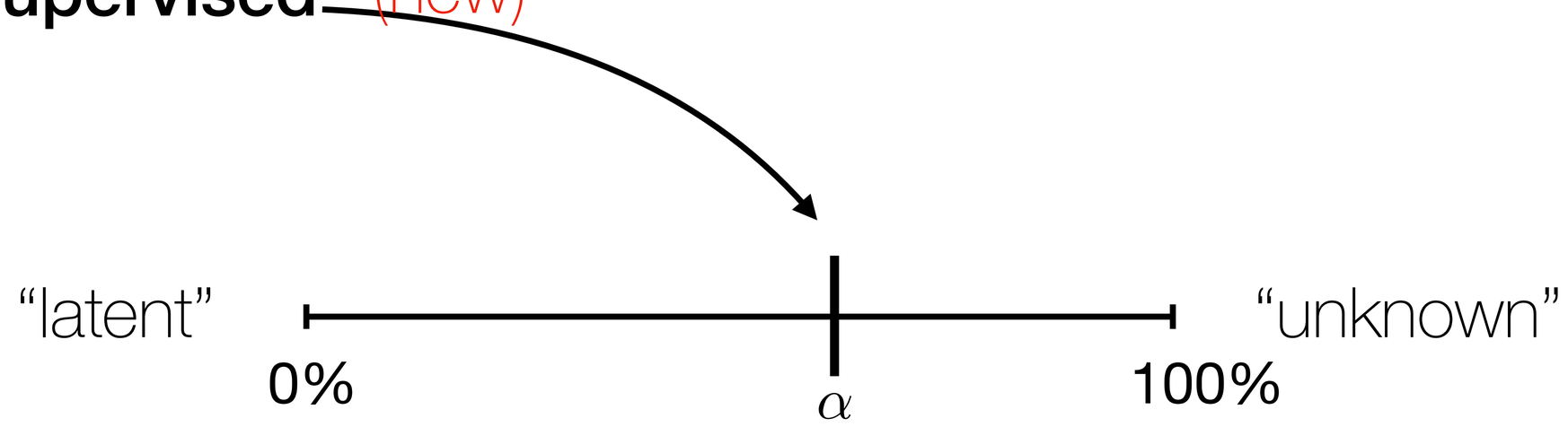
- “latent” (new)
- “unknown”
- “known”
- “observational”
- **“semi-supervised”** (new)

For each sample:

- (fraction) know correspondence to intervention regime;

For each intervention:

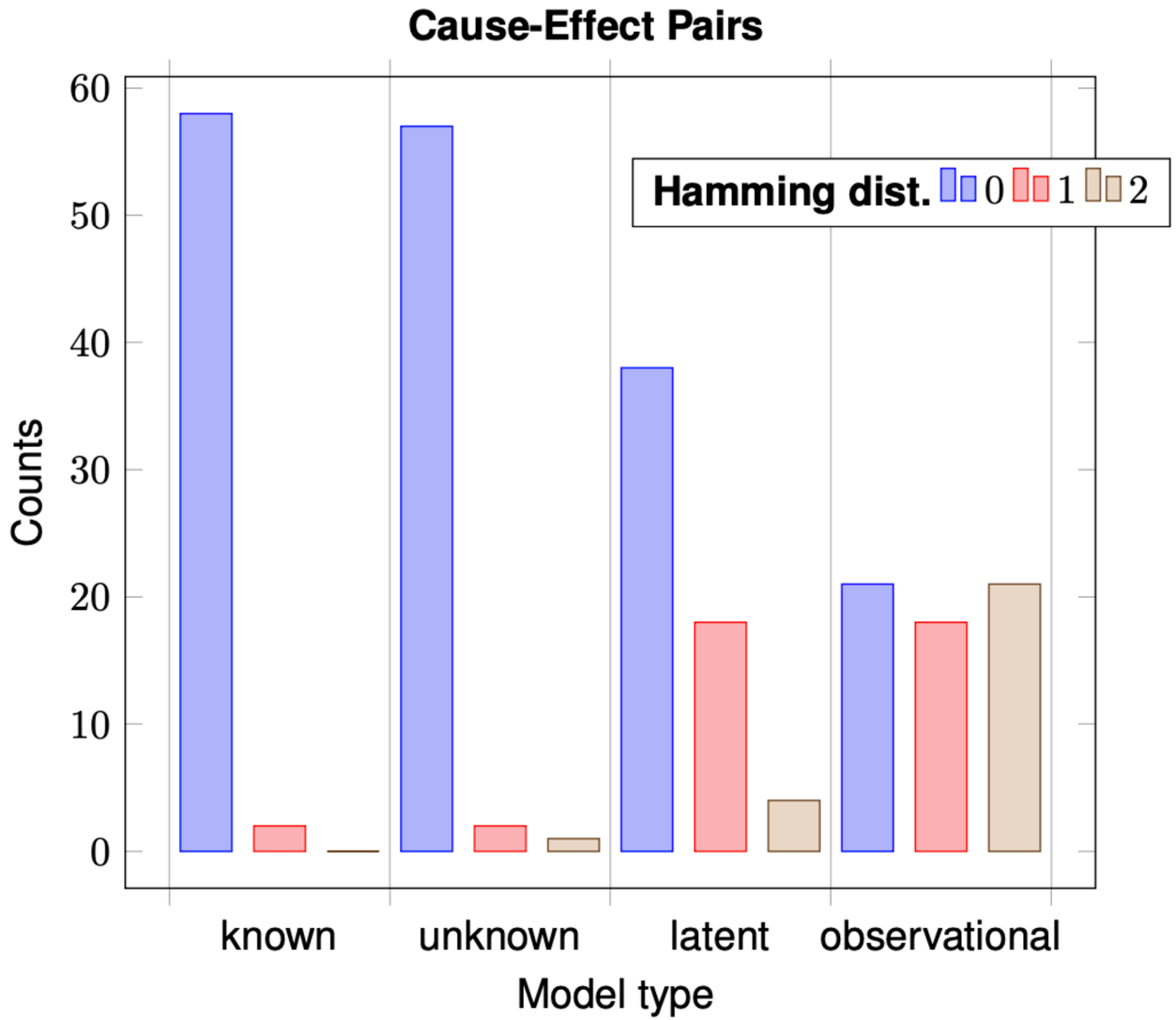
- do not know experimental conditions

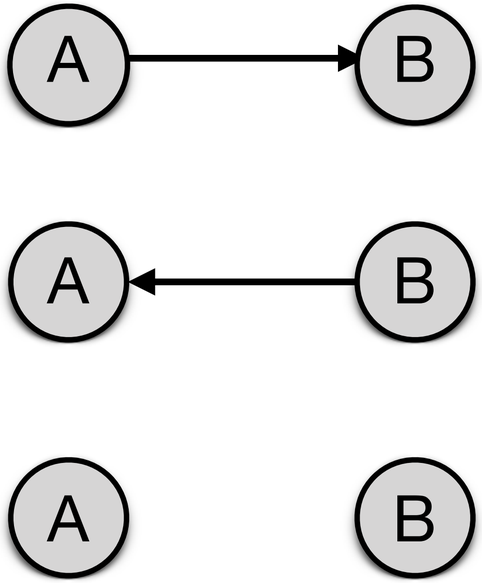


Fraction of samples with correspondences.

Experiments

Experiments with cause effect pairs

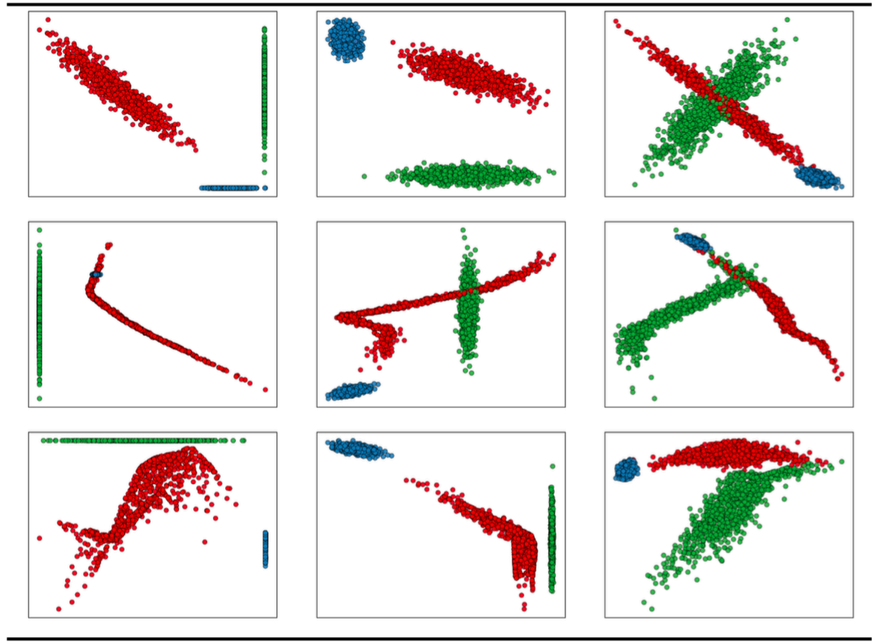




Experiments on synthetic data

Model Type	e	latent	unknown	known	observational
		Stochastic Interventions:			
Linear Gaussian		5.9 ± 6.2	3.4 ± 3.2	0.5 ± 1.3	10.3 ± 7.8
Non-Linear Gaussian	1	12.2 ± 3.9	10.3 ± 2.5	7.0 ± 3.6	13.7 ± 3.8
Non-Linear Non-Gaussian		8.7 ± 6.6	8.0 ± 2.7	6.6 ± 2.2	11.3 ± 5.0
Linear Gaussian		27.2 ± 6.2	24.1 ± 5.8	15.6 ± 6.0	39.6 ± 5.0
Non-Linear Gaussian	4	35.8 ± 3.8	30.3 ± 5.3	27.7 ± 4.3	37.5 ± 5.2
Non-Linear Non-Gaussian		36.1 ± 4.4	35.5 ± 8.1	31.5 ± 5.6	40.2 ± 6.9
		Imperfect Interventions:			
Linear Gaussian		5.8 ± 4.2	6.2 ± 3.06	4.7 ± 3.6	10.4 ± 2.9
Non-Linear Gaussian	1	9.3 ± 2.4	8.9 ± 2.5	7.8 ± 3.9	10.5 ± 2.8
Non-Linear Non-Gaussian		8.8 ± 3.0	9.1 ± 3.5	7.9 ± 1.4	11.5 ± 5.4
Linear Gaussian		35.9 ± 8.3	29.7 ± 5.6	17.7 ± 7.9	39.1 ± 9.1
Non-Linear Gaussian	4	32.1 ± 6.0	32.6 ± 5.8	32.8 ± 5.4	39.8 ± 9.3
Non-Linear Non-Gaussian		30.4 ± 12.2	30.2 ± 11.2	25.8 ± 3.9	36.7 ± 9.8

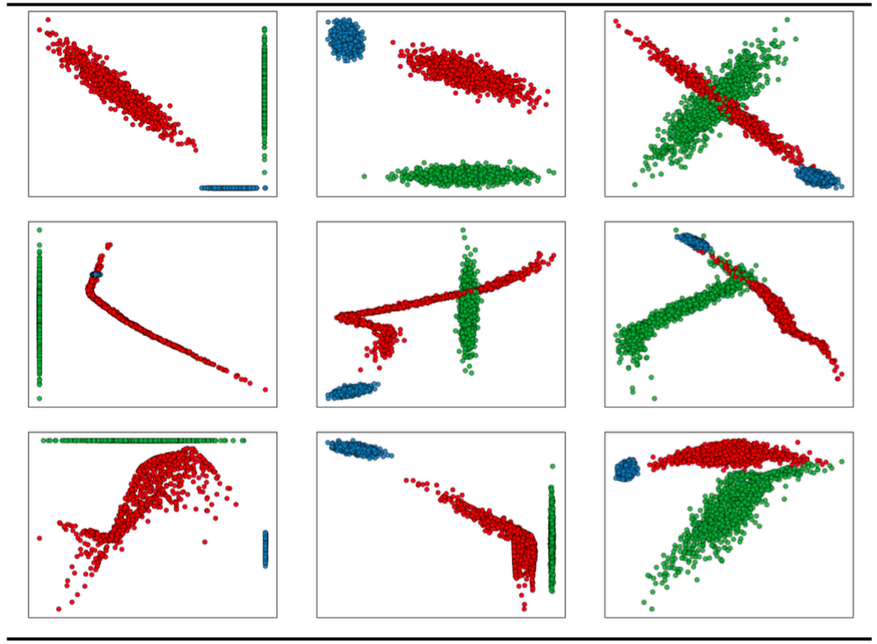
Table 4.1: Hamming distances on synthetic 10 variable SCMs.



Experiments on synthetic data

Model Type	e	latent	unknown	known	observational
		Stochastic Interventions:			
Linear Gaussian		5.9 ± 6.2	3.4 ± 3.2	0.5 ± 1.3	10.3 ± 7.8
Non-Linear Gaussian	1	12.2 ± 3.9	10.3 ± 2.5	7.0 ± 3.6	13.7 ± 3.8
Non-Linear Non-Gaussian		8.7 ± 6.6	8.0 ± 2.7	6.6 ± 2.2	11.3 ± 5.0
Linear Gaussian		27.2 ± 6.2	24.1 ± 5.8	15.6 ± 6.0	39.6 ± 5.0
Non-Linear Gaussian	4	35.8 ± 3.8	30.3 ± 5.3	27.7 ± 4.3	37.5 ± 5.2
Non-Linear Non-Gaussian		36.1 ± 4.4	35.5 ± 8.1	31.5 ± 5.6	40.2 ± 6.9
		Imperfect Interventions:			
Linear Gaussian		5.8 ± 4.2	6.2 ± 3.06	4.7 ± 3.6	10.4 ± 2.9
Non-Linear Gaussian	1	9.3 ± 2.4	8.9 ± 2.5	7.8 ± 3.9	10.5 ± 2.8
Non-Linear Non-Gaussian		8.8 ± 3.0	9.1 ± 3.5	7.9 ± 1.4	11.5 ± 5.4
Linear Gaussian		35.9 ± 8.3	29.7 ± 5.6	17.7 ± 7.9	39.1 ± 9.1
Non-Linear Gaussian	4	32.1 ± 6.0	32.6 ± 5.8	32.8 ± 5.4	39.8 ± 9.3
Non-Linear Non-Gaussian		30.4 ± 12.2	30.2 ± 11.2	25.8 ± 3.9	36.7 ± 9.8

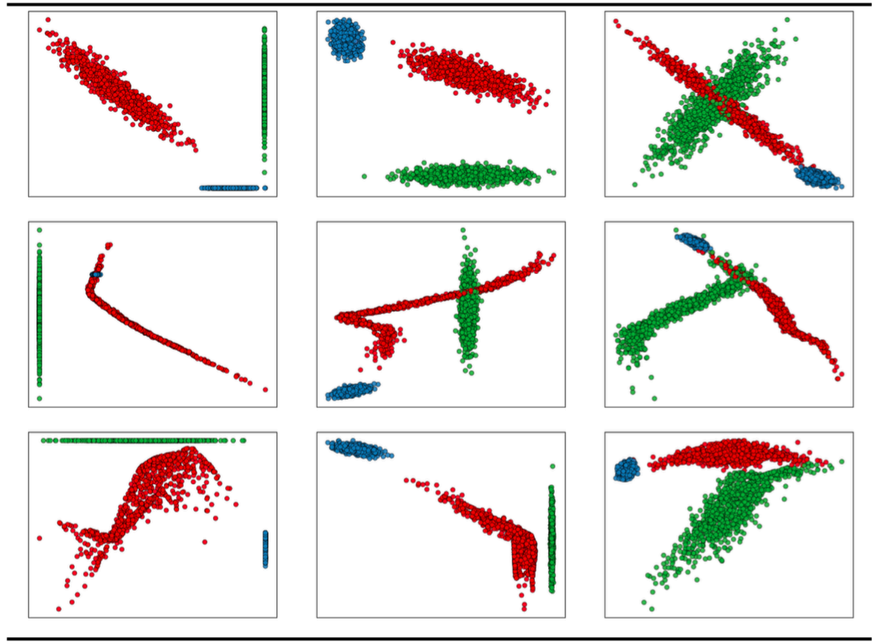
Table 4.1: Hamming distances on synthetic 10 variable SCMs.



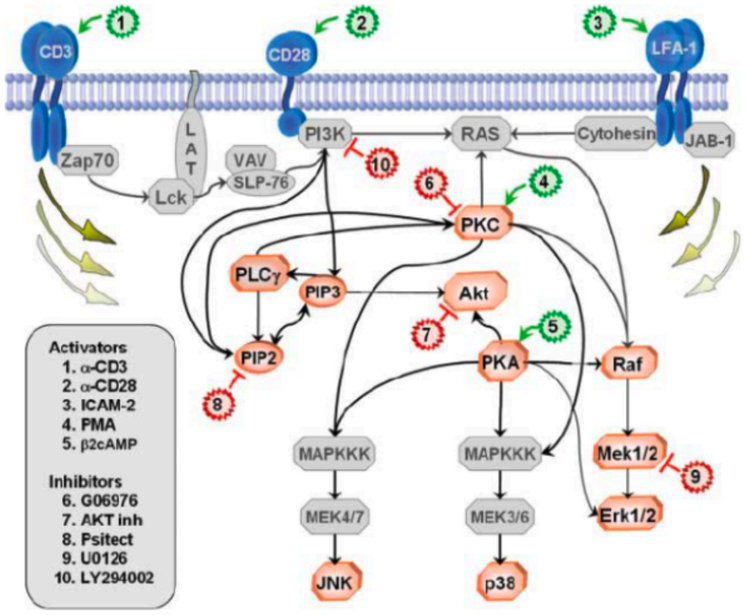
Experiments on synthetic data

Model Type	e	latent	unknown	known	observational
		Stochastic Interventions:			
Linear Gaussian		5.9 ± 6.2	3.4 ± 3.2	0.5 ± 1.3	10.3 ± 7.8
Non-Linear Gaussian	1	12.2 ± 3.9	10.3 ± 2.5	7.0 ± 3.6	13.7 ± 3.8
Non-Linear Non-Gaussian		8.7 ± 6.6	8.0 ± 2.7	6.6 ± 2.2	11.3 ± 5.0
Linear Gaussian		27.2 ± 6.2	24.1 ± 5.8	15.6 ± 6.0	39.6 ± 5.0
Non-Linear Gaussian	4	35.8 ± 3.8	30.3 ± 5.3	27.7 ± 4.3	37.5 ± 5.2
Non-Linear Non-Gaussian		36.1 ± 4.4	35.5 ± 8.1	31.5 ± 5.6	40.2 ± 6.9
		Imperfect Interventions:			
Linear Gaussian		5.8 ± 4.2	6.2 ± 3.06	4.7 ± 3.6	10.4 ± 2.9
Non-Linear Gaussian	1	9.3 ± 2.4	8.9 ± 2.5	7.8 ± 3.9	10.5 ± 2.8
Non-Linear Non-Gaussian		8.8 ± 3.0	9.1 ± 3.5	7.9 ± 1.4	11.5 ± 5.4
Linear Gaussian		35.9 ± 8.3	29.7 ± 5.6	17.7 ± 7.9	39.1 ± 9.1
Non-Linear Gaussian	4	32.1 ± 6.0	32.6 ± 5.8	32.8 ± 5.4	39.8 ± 9.3
Non-Linear Non-Gaussian		30.4 ± 12.2	30.2 ± 11.2	25.8 ± 3.9	36.7 ± 9.8

Table 4.1: Hamming distances on synthetic 10 variable SCMs.



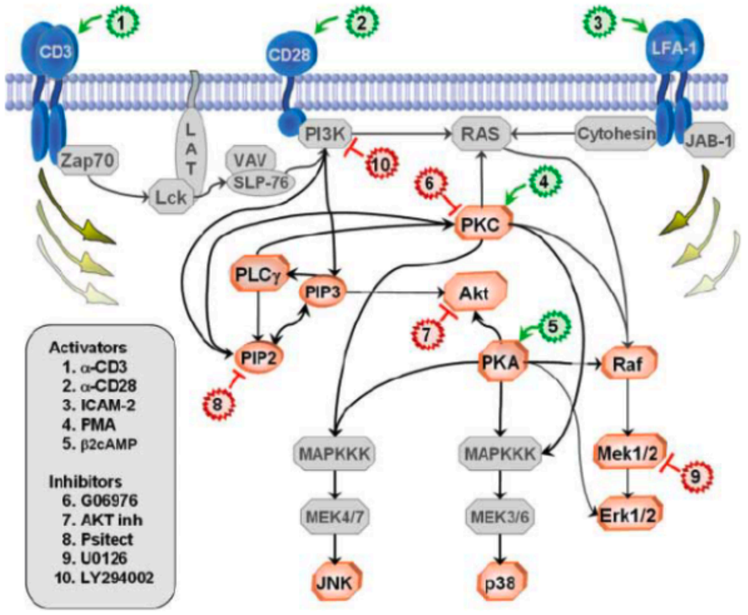
Experiments on real data



ground truth causal model.

	HD	tp	fn	fp	rev	F_1 score
GIES (Hauser and Bühlmann, 2012)	38	10	0	41	7	0.33
CAM (Bühlmann et al., 2014)	35	12	1	30	4	0.51
IGSP (Wang et al., 2017)	18	4	6	5	7	0.42
DCDI-G (Brouillard et al., 2020)	36	6	2	25	9	0.31
DCDI-DSF (Brouillard et al., 2020)	33	6	2	22	9	0.33
FCI (Spirtes et al., 1993)	35	4	12	21	5	0.22
Imperfect Linear Gaussian **(ours)**	33	7	11	22	3	0.30
Imperfect Non-Linear Gaussian **(ours)**	19	7	11	8	0	0.42
Imperfect Normalizing Flow **(ours)**	30	9	9	21	1	0.38
Perfect Linear Gaussian **(ours)**	23	8	10	13	3	0.41
Perfect Non-Linear Gaussian **(ours)**	24	11	7	17	1	0.48
Perfect Normalizing Flow **(ours)**	23	7	11	12	2	0.38

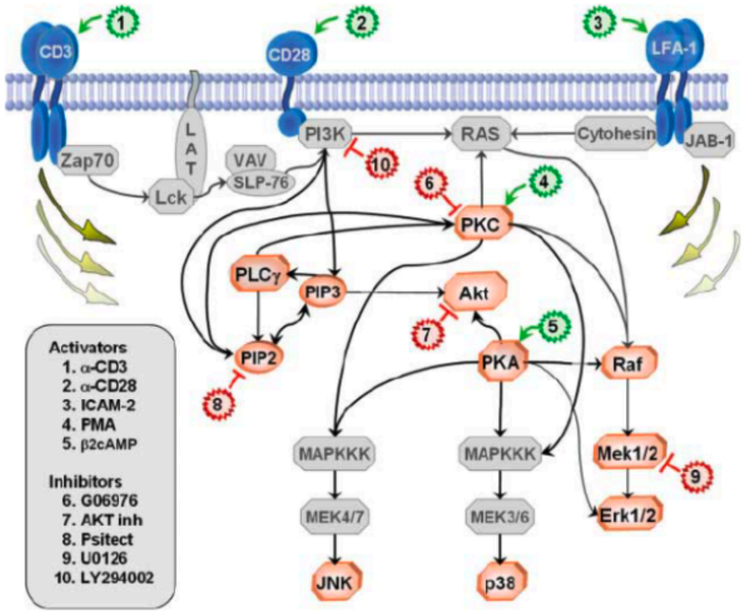
Experiments on real data



ground truth causal model.

	HD	tp	fn	fp	rev	F_1 score
GIES (Hauser and Bühlmann, 2012)	38	10	0	41	7	0.33
CAM (Bühlmann et al., 2014)	35	12	1	30	4	0.51
IGSP (Wang et al., 2017)	18	4	6	5	7	0.42
DCDI-G (Brouillard et al., 2020)	36	6	2	25	9	0.31
DCDI-DSF (Brouillard et al., 2020)	33	6	2	22	9	0.33
FCI (Spirtes et al., 1993)	35	4	12	21	5	0.22
Imperfect Linear Gaussian **(ours)**	33	7	11	22	3	0.30
Imperfect Non-Linear Gaussian **(ours)**	19	7	11	8	0	0.42
Imperfect Normalizing Flow **(ours)**	30	9	9	21	1	0.38
Perfect Linear Gaussian **(ours)**	23	8	10	13	3	0.41
Perfect Non-Linear Gaussian **(ours)**	24	11	7	17	1	0.48
Perfect Normalizing Flow **(ours)**	23	7	11	12	2	0.38

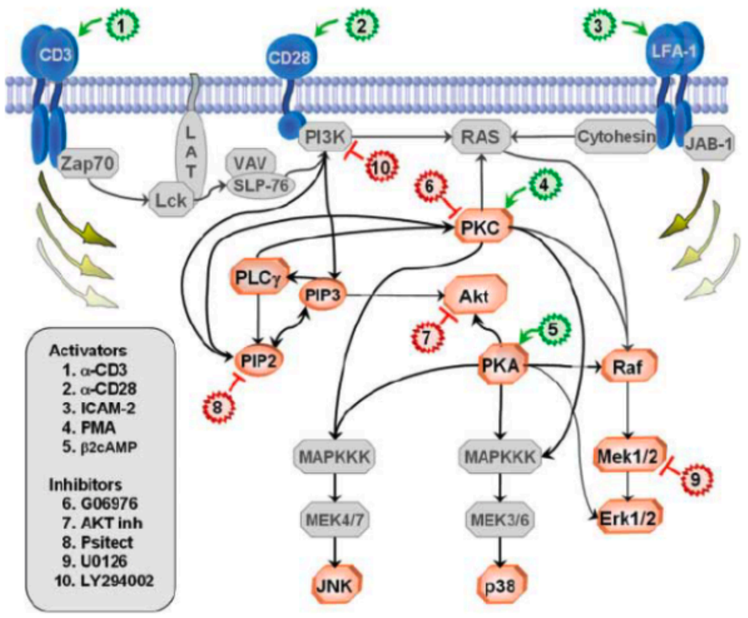
Experiments on real data



ground truth causal model.

	HD	tp	fn	fp	rev	F_1 score
GIES (Hauser and Bühlmann, 2012)	38	10	0	41	7	0.33
CAM (Bühlmann et al., 2014)	35	12	1	30	4	0.51
IGSP (Wang et al., 2017)	18	4	6	5	7	0.42
DCDI-G (Brouillard et al., 2020)	36	6	2	25	9	0.31
DCDI-DSF (Brouillard et al., 2020)	33	6	2	22	9	0.33
FCI (Spirtes et al., 1993)	35	4	12	21	5	0.22
Imperfect Linear Gaussian **(ours)**	33	7	11	22	3	0.30
Imperfect Non-Linear Gaussian **(ours)**	19	7	11	8	0	0.42
Imperfect Normalizing Flow **(ours)**	30	9	9	21	1	0.38
Perfect Linear Gaussian **(ours)**	23	8	10	13	3	0.41
Perfect Non-Linear Gaussian **(ours)**	24	11	7	17	1	0.48
Perfect Normalizing Flow **(ours)**	23	7	11	12	2	0.38

Experiments on real data



ground truth causal model.

Edges in the wrong direction

	HD	tp	fn	fp	rev	F_1 score
GIES (Hauser and Bühlmann, 2012)	38	10	0	41	7	0.33
CAM (Bühlmann et al., 2014)	35	12	1	30	4	0.51
IGSP (Wang et al., 2017)	18	4	6	5	7	0.42
DCDI-G (Brouillard et al., 2020)	36	6	2	25	9	0.31
DCDI-DSF (Brouillard et al., 2020)	33	6	2	22	9	0.33
FCI (Spirtes et al., 1993)	35	4	12	21	5	0.22
Imperfect Linear Gaussian **(ours)**	33	7	11	22	3	0.30
Imperfect Non-Linear Gaussian **(ours)**	19	7	11	8	0	0.42
Imperfect Normalizing Flow **(ours)**	30	9	9	21	1	0.38
Perfect Linear Gaussian **(ours)**	23	8	10	13	3	0.41
Perfect Non-Linear Gaussian **(ours)**	24	11	7	17	1	0.48
Perfect Normalizing Flow **(ours)**	23	7	11	12	2	0.38

Conclusion

- Make heterogeneity or non-stationarity your friend.
- Causal Discovery under Latent Interventions

Causal discovery and Deep Learning

Faria et. al. 2022; Pearl et. al. 2018; Neal et. al. 2022; Schölkopf et. al. 2019; Bengio et. al. 2020; Peters et. al. 2017, Ke et. al. 2020, Brouillard et. al. 2020

Gonçalo R. A. Faria
@goncalorafaria

Deep Learning Sessions Portugal