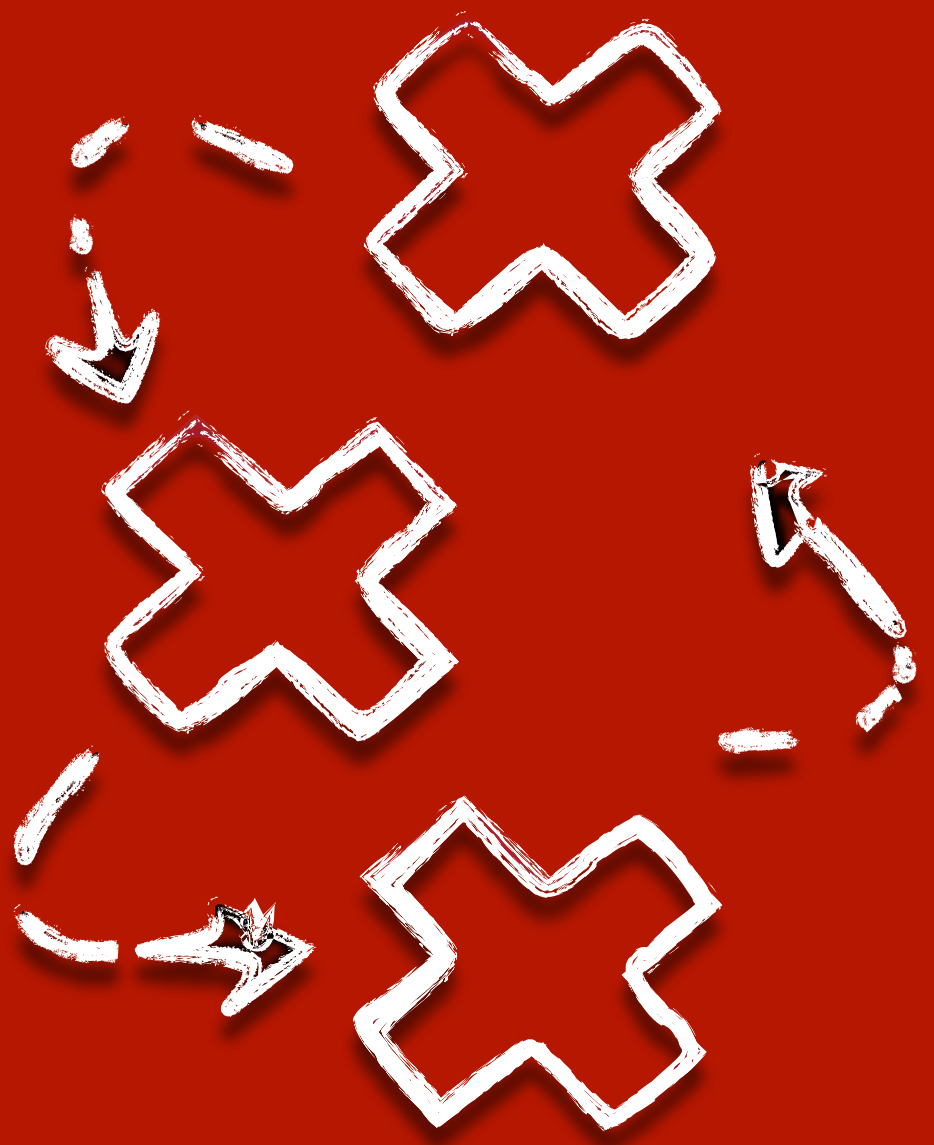
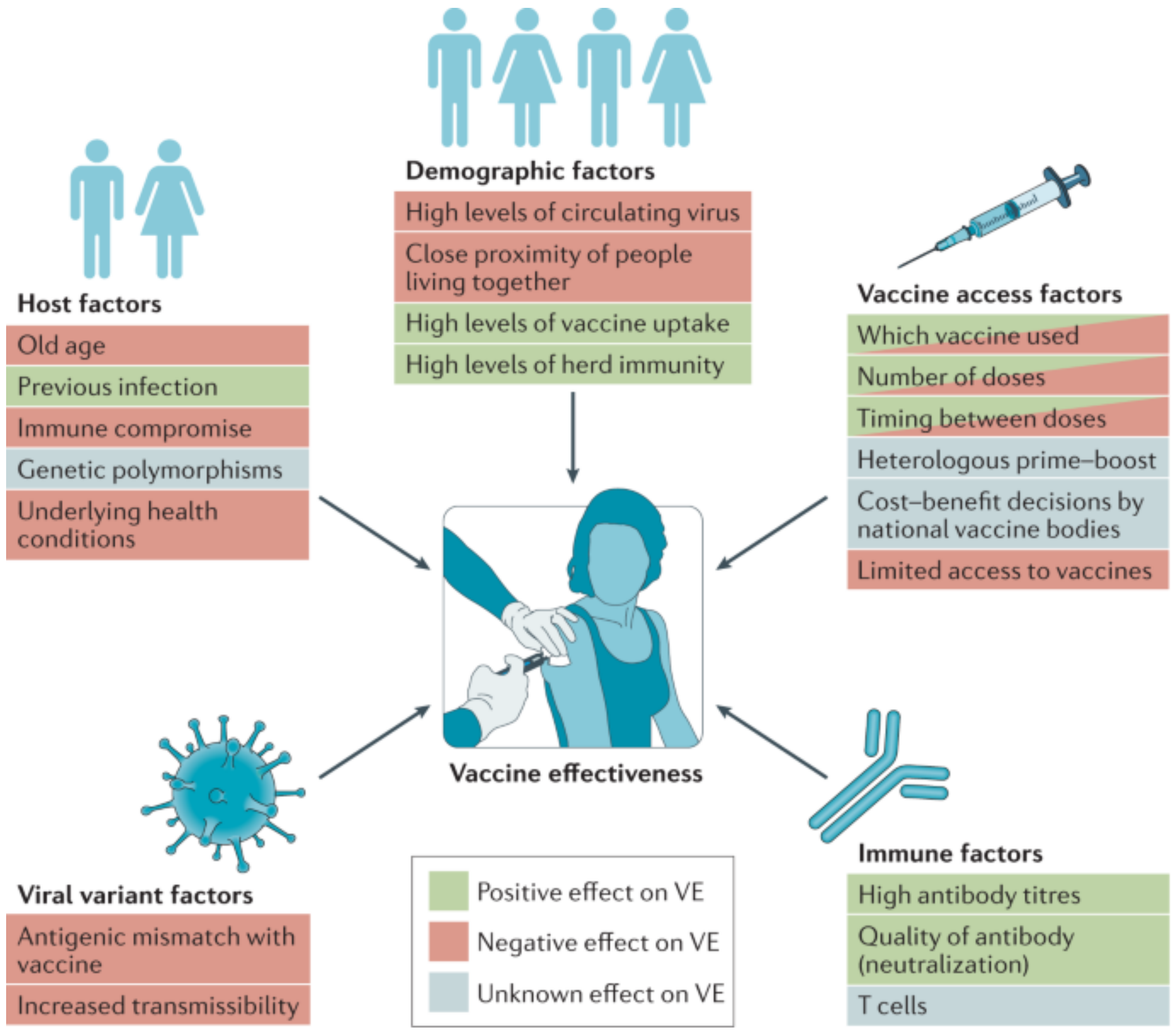


Causal vs causality-inspired representation learning

Sara Magliacane (University of Amsterdam, MIT-IBM Watson AI Lab

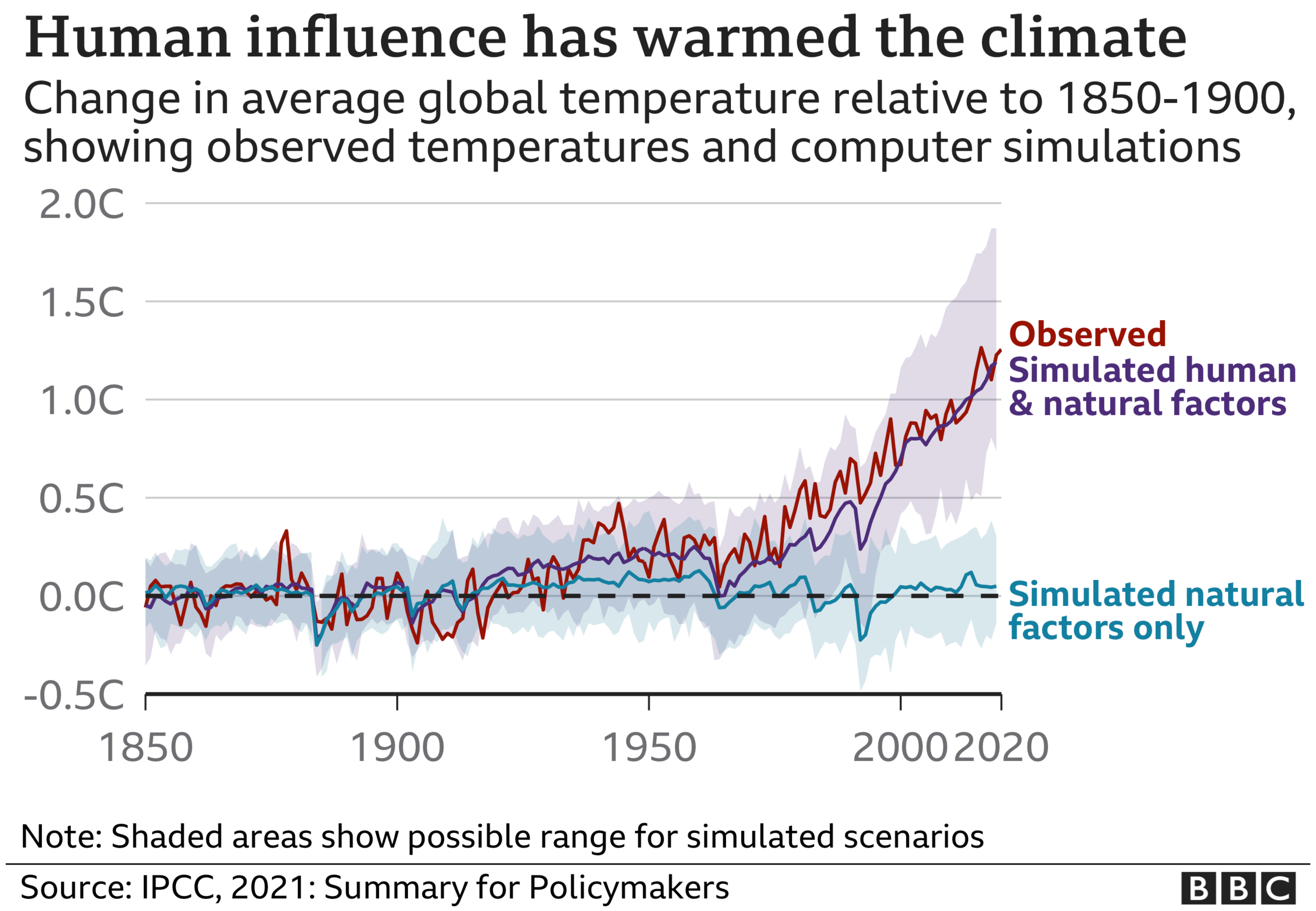
(joint work with Phillip Lippe, Sindy Löwe, Yuki Asano, Taco Cohen, Stratis Gavves, Biwei Huang, Fan Feng, Chaochao Lu and Kun Zhang)

Causal questions are ubiquitous: healthcare



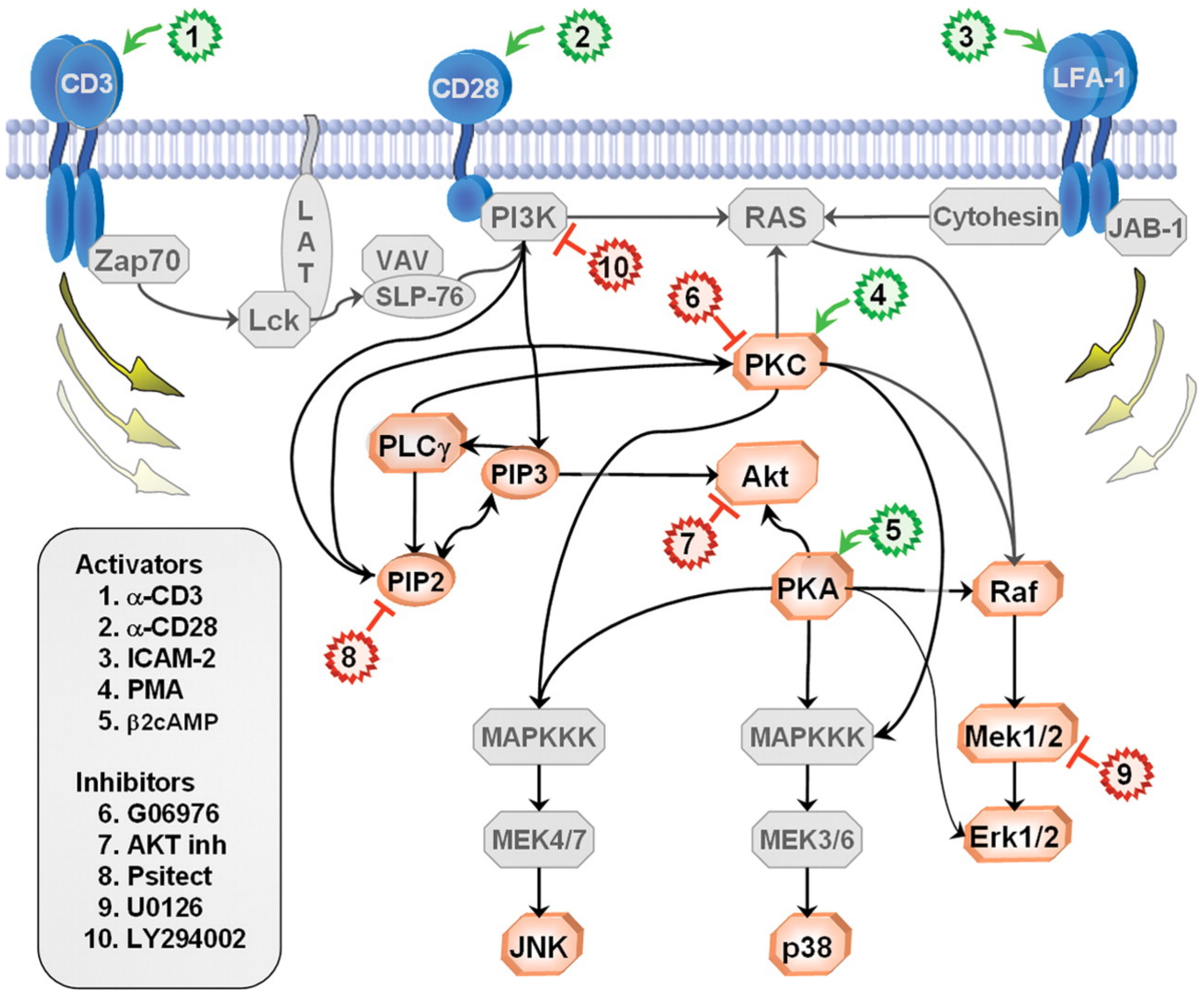
https://www.nature.com/articles/s41577-021-00592-1

Causal questions are ubiquitous: climate change



https://www.bbc.com/news/science-environment-58600723

Causal questions are ubiquitous: biology



https://www.science.org/doi/abs/10.1126/science.1105809

A working definition of causality in machine learning

Informal definition: A variable X causes another variable Y, if changing (the distribution of) X, e.g. by fixing its value, changes (the distribution of) Y

A working definition of causality in machine learning

Informal definition: A variable X causes another variable Y, if changing (the distribution of) X, e.g. by fixing its value, changes (the distribution of) Y

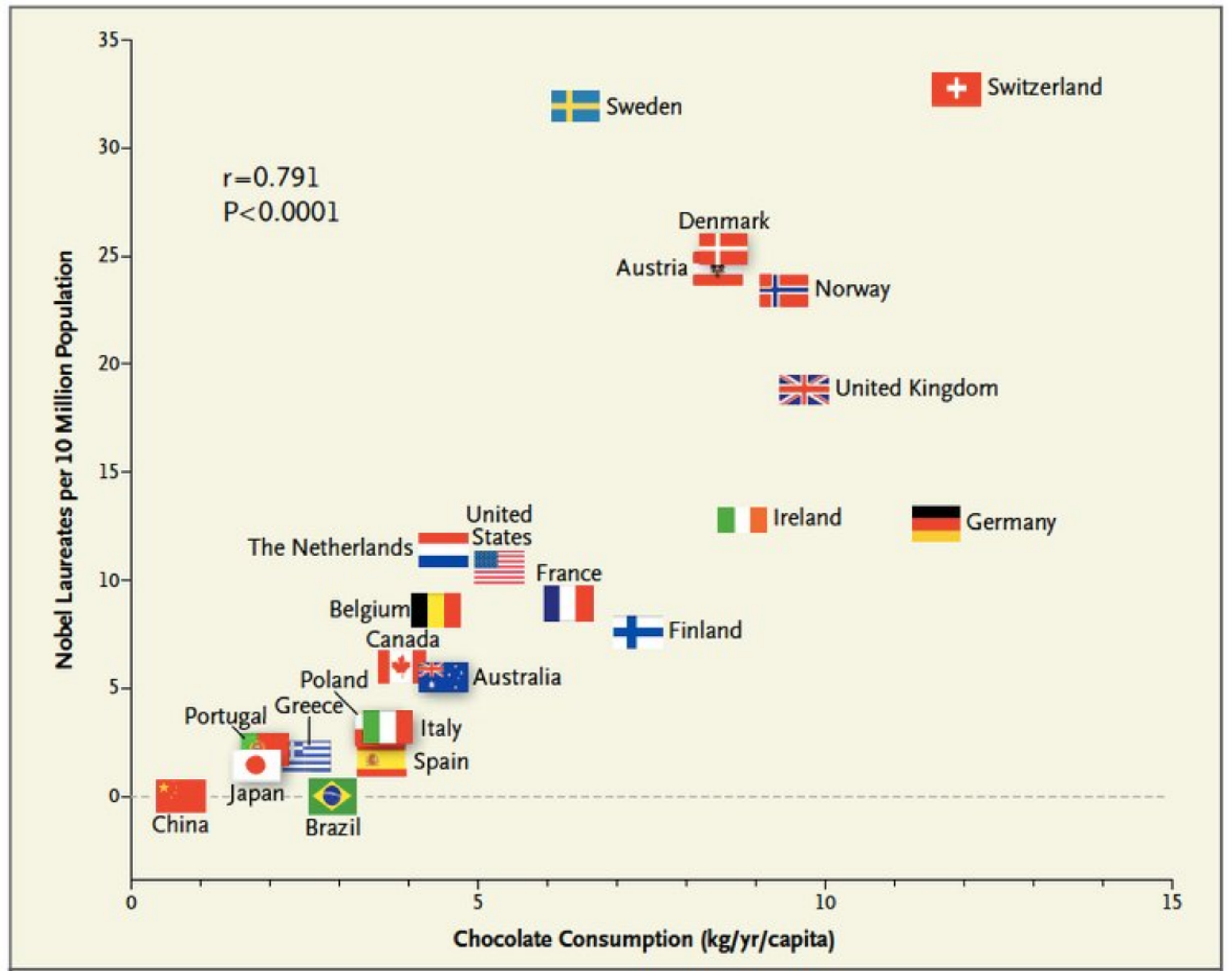


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

[Messerli, 2012] https://www.nejm.org/doi/full/10.1056/NEJMon1211064

A working definition of causality in machine learning

Informal definition: A variable X causes another variable Y, if changing (the distribution of) X, e.g. by fixing its value, changes (the distribution of) Y

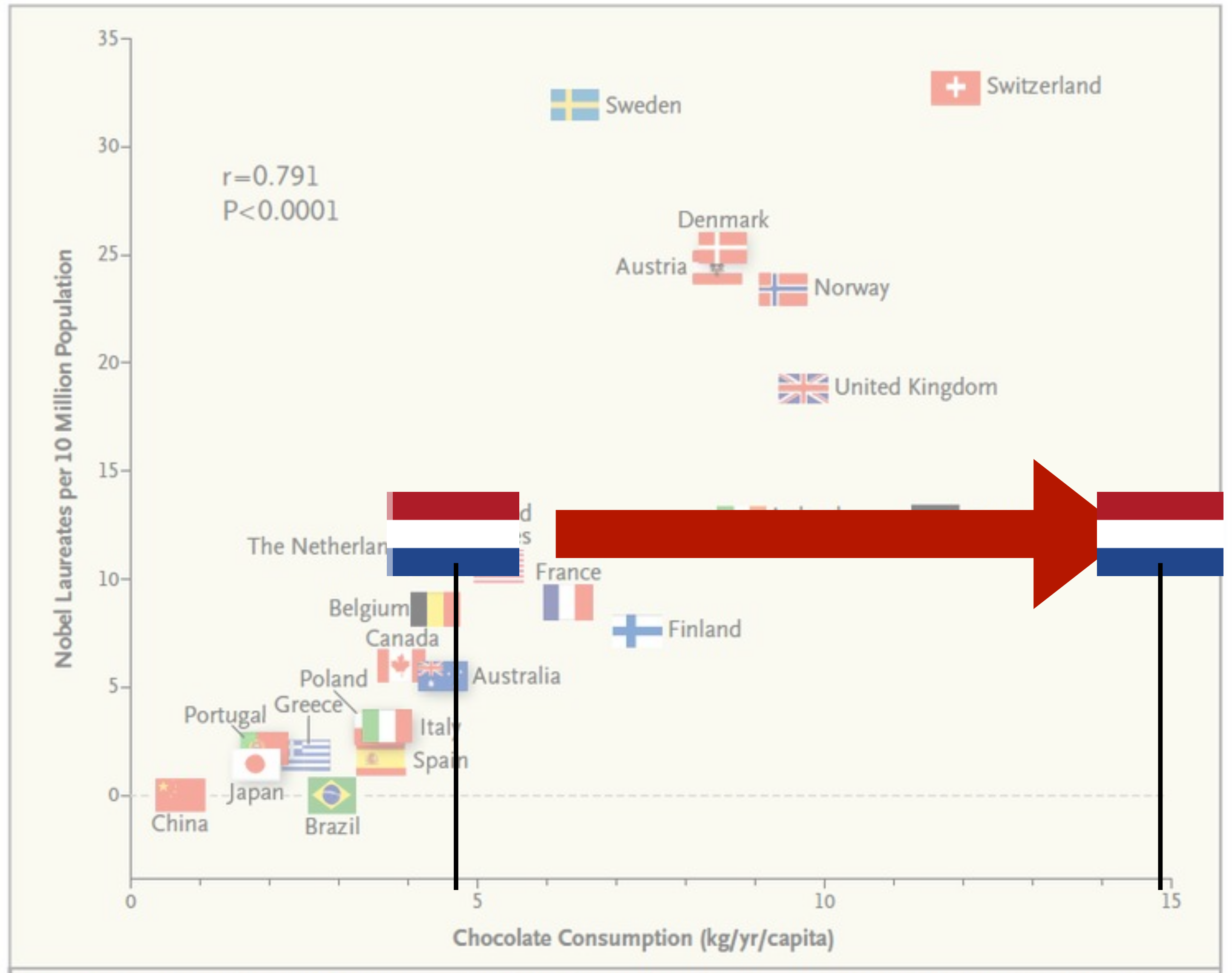


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

NL eats more chocolate => nothing changes

[Messerli, 2012] https://www.nejm.org/doi/full/10.1056/NEJMon1211064

A working definition of causality in machine learning

Informal definition: A variable X causes another variable Y, if changing (the distribution of) X, e.g. by fixing its value, changes (the distribution of) Y

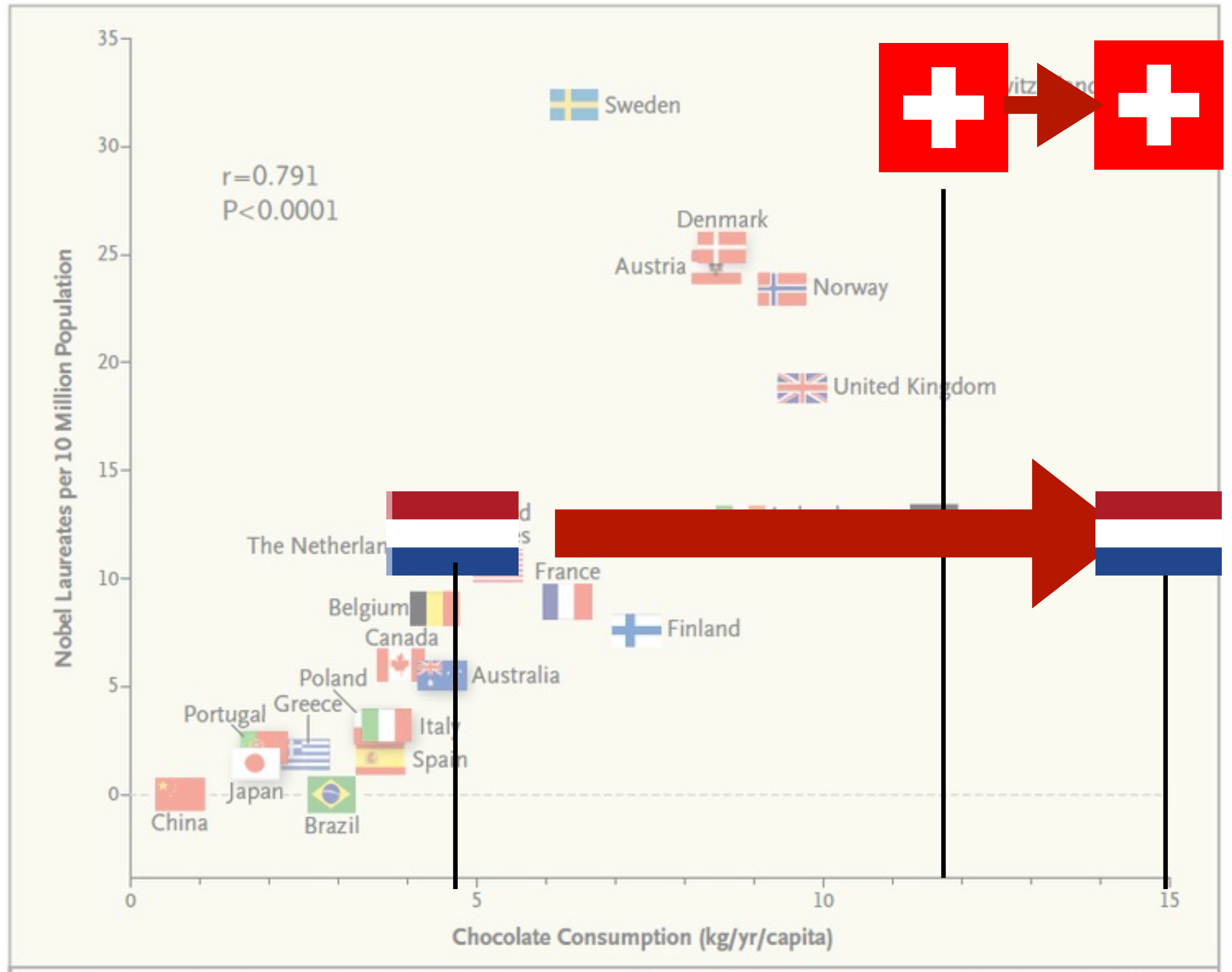


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

NL eats more chocolate => nothing changes

… and similarly for other countries (and other values)

Chocolate does not cause Nobel prizes

[Messerli, 2012] https://www.nejm.org/doi/full/10.1056/NEJMon1211064

A working definition of causality in machine learning

Informal definition: A variable X causes another variable Y, if changing (the distribution of) X, e.g. by fixing its value, changes (the distribution of) Y

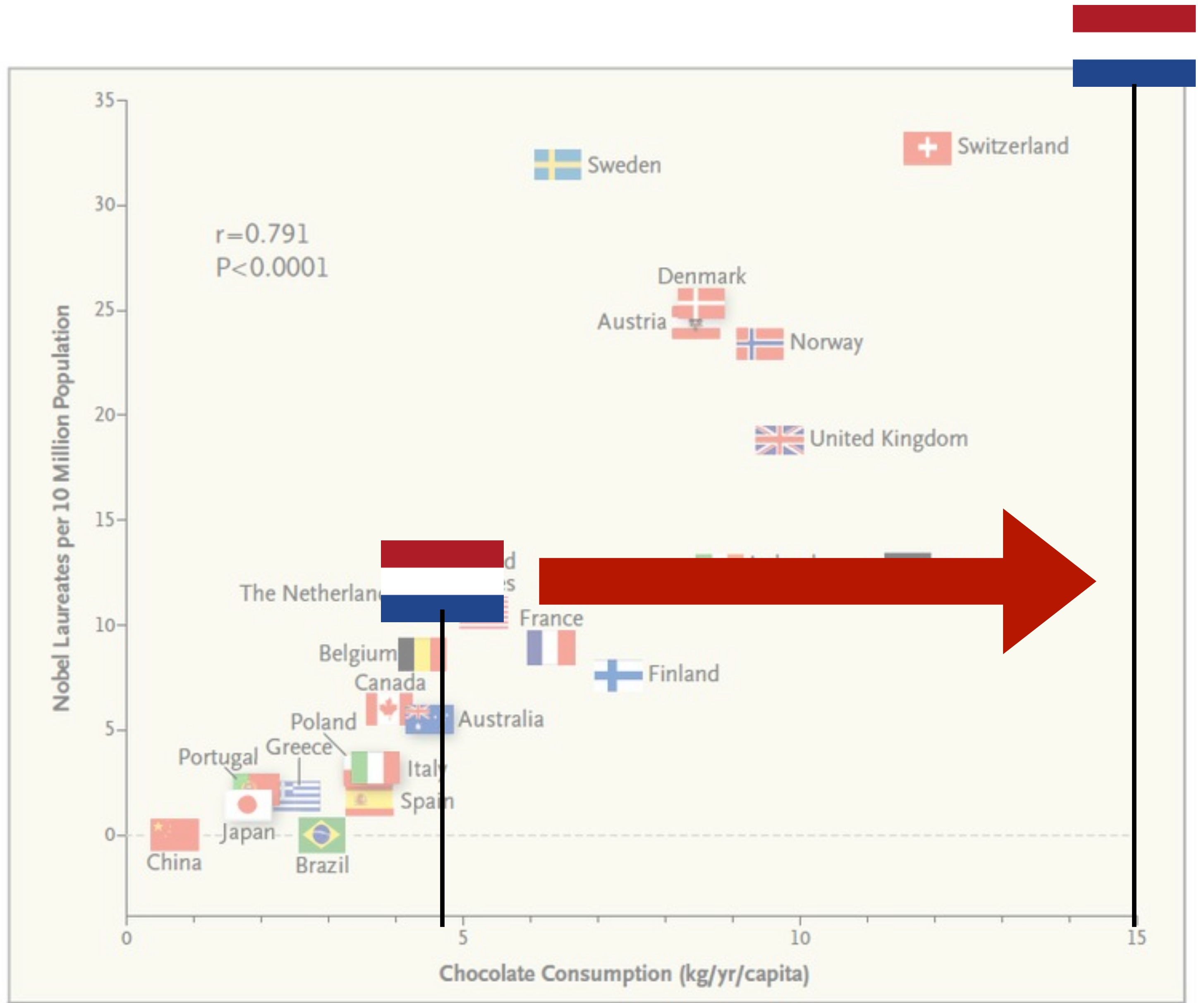


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

In a hypothetical universe:

NL eats more chocolate => more Nobel prizes

[Messerli, 2012] https://www.nejm.org/doi/full/10.1056/NEJMon1211064

A working definition of causality in machine learning

Informal definition: A variable X causes another variable Y, if changing (the distribution of) X, e.g. by fixing its value, changes (the distribution of) Y

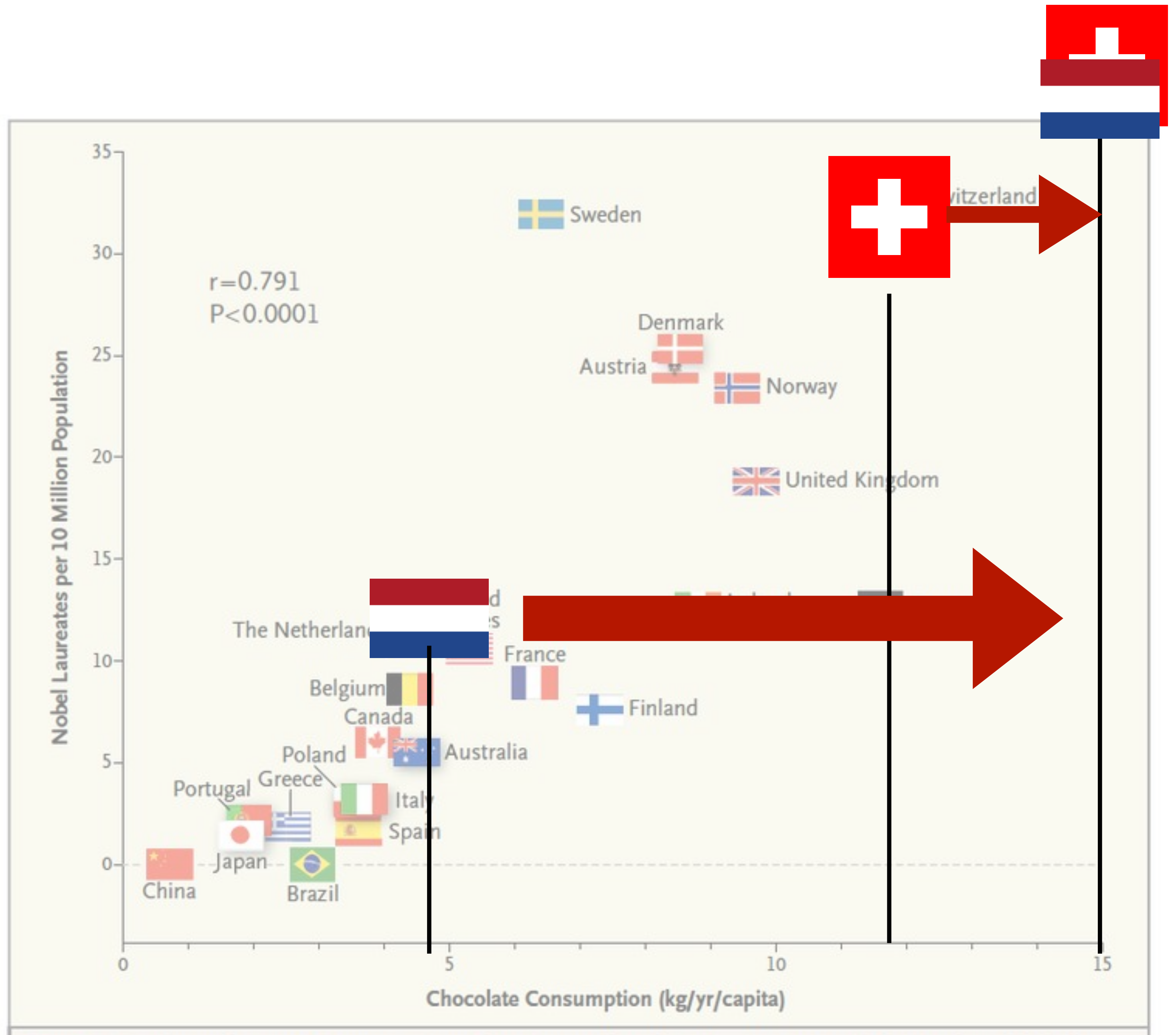


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

In a hypothetical universe:

NL eats more chocolate => more Nobel prizes

CH eats more chocolate => more Nobel prizes

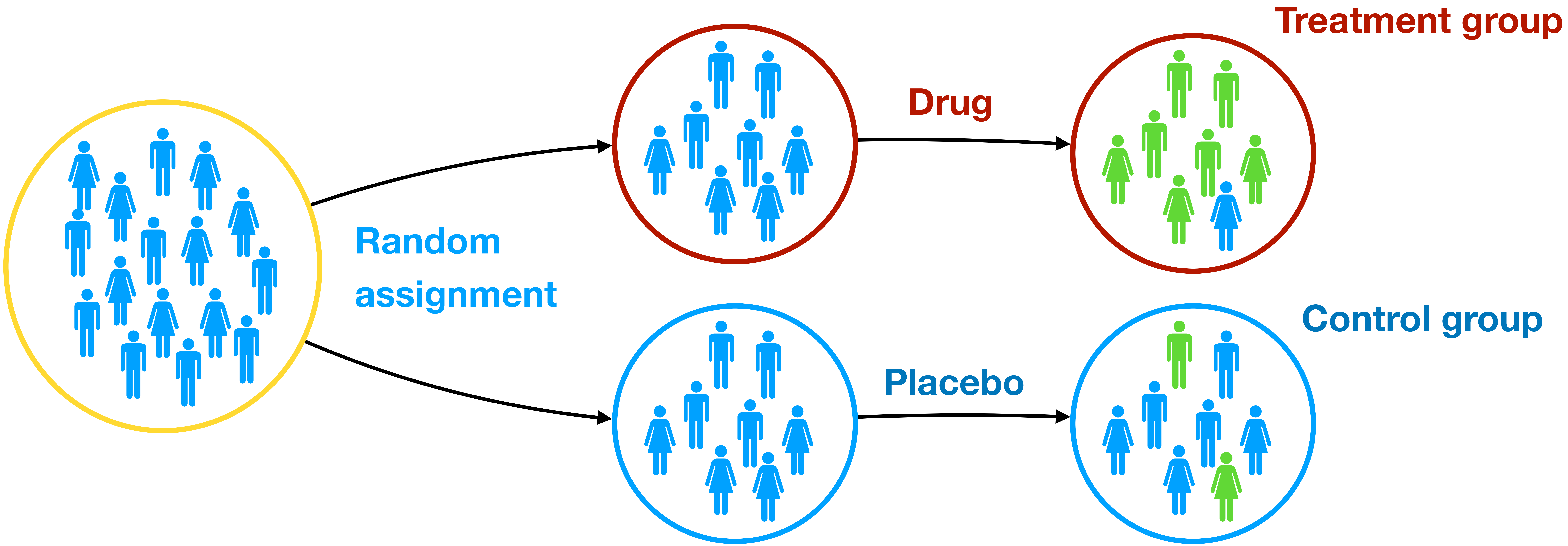
… and similarly for (some) other countries

Chocolate causes Nobel prizes

Based on experimental data

[Messerli, 2012] https://www.nejm.org/doi/full/10.1056/NEJMon1211064

Gold standard of experiments: Randomized Controlled Trials (RCTs)



A working definition of causality in machine learning

Informal definition: A variable X causes another variable Y, if changing (the distribution of) X, e.g. by fixing its value, changes (the distribution of) Y

Intervention

A working definition of causality in machine learning

Informal definition: A variable X causes another variable Y, if changing (the distribution of) X, e.g. by fixing its value, changes (the distribution of) Y

Intervention

Challenge: estimate the causal effect of an intervention, when we do not have (all possible) interventional data **(e.g. observational data)**

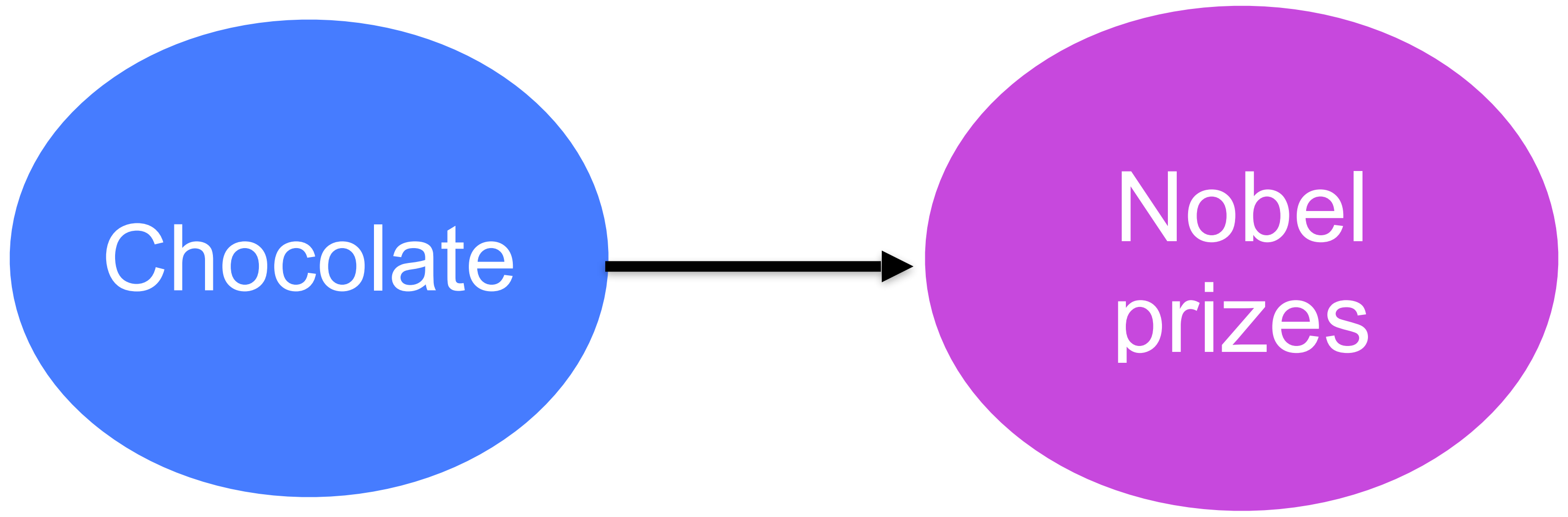
A working definition of causality in machine learning

Informal definition: A variable X causes another variable Y, if changing (the distribution of) X, e.g. by fixing its value, changes (the distribution of) Y

Intervention

Challenge: estimate the causal effect of an intervention, when we do not have (all possible) interventional data **(e.g. observational data)**

Representation: We can represent causal relations in **causal graphs:** nodes are random variables, edges causal relations



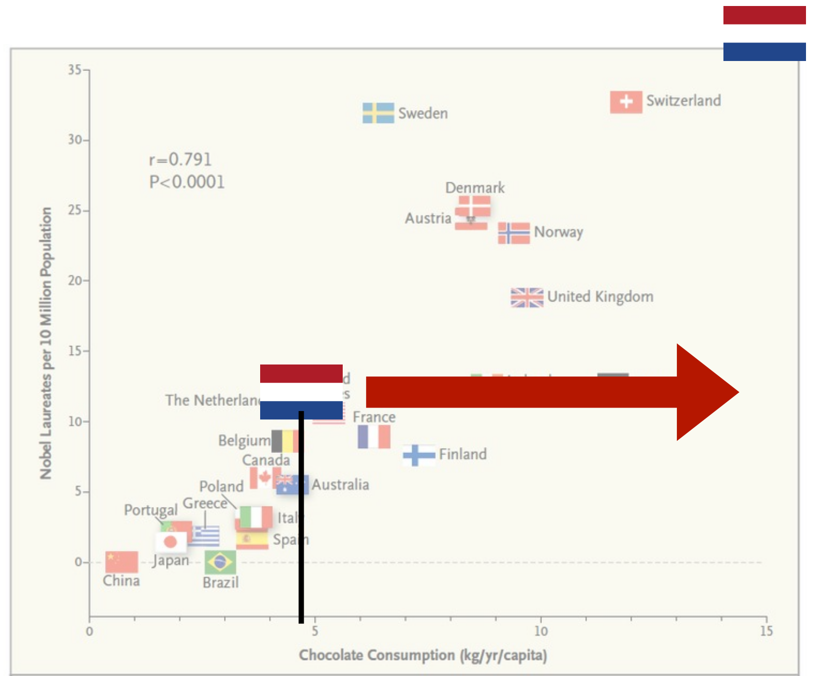
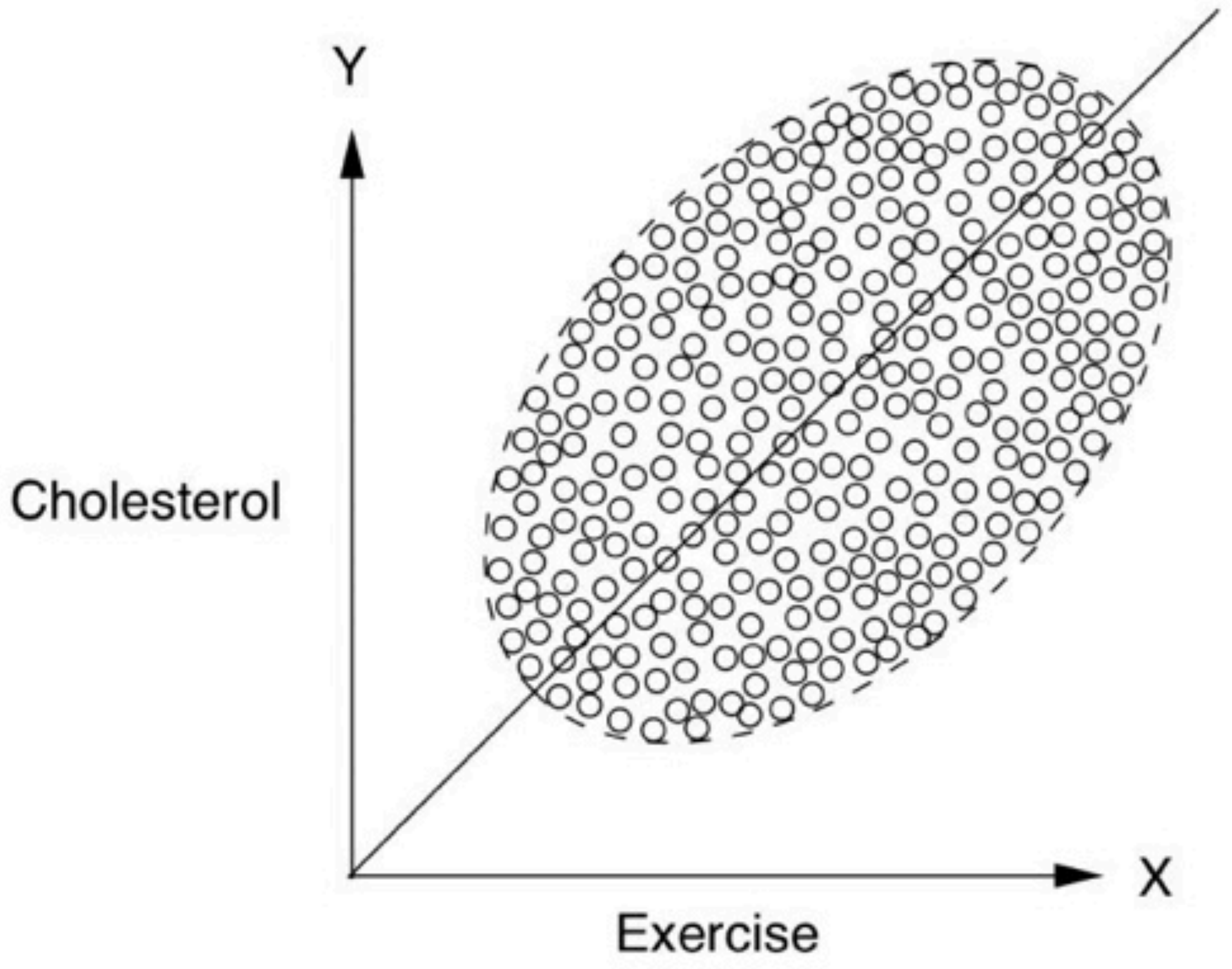


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Observational data: when we do not have RCTs

Let's assume we have **observational data** (e.g. data collected by hospitals)



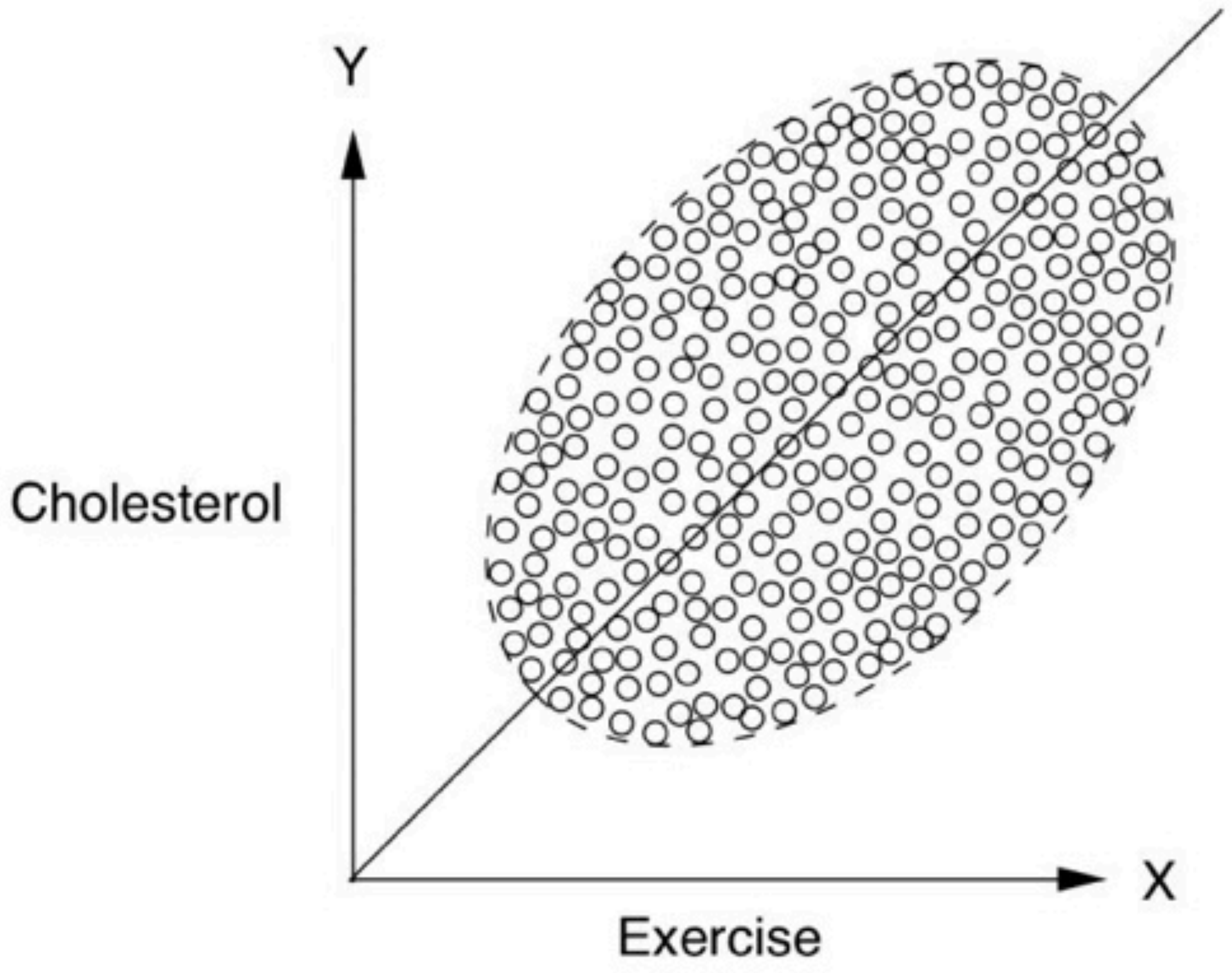
We don't know if there is a policy (and in case, which one) of how people decide to exercise.

We know they were **not** randomly split in two similar groups and randomly assigned exercise.

From the Book of Why [Pearl 2018]

Observational data: when we do not have RCTs

Let's assume we have **observational data** (e.g. data collected by hospitals)



We don't know if there is a policy (and in case, which one) of how people decide to exercise.

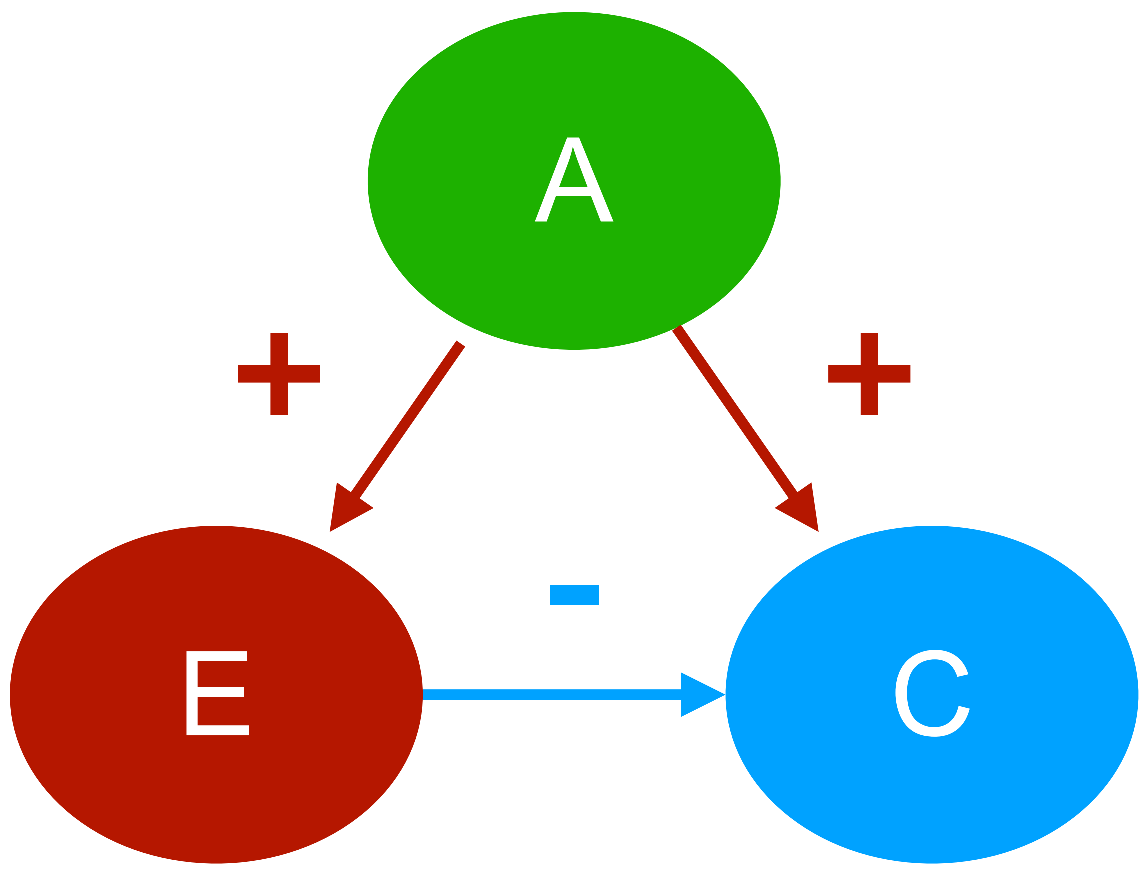
We know they were **not** randomly split in two similar groups and randomly assigned exercise.

Exercise increases cholesterol??

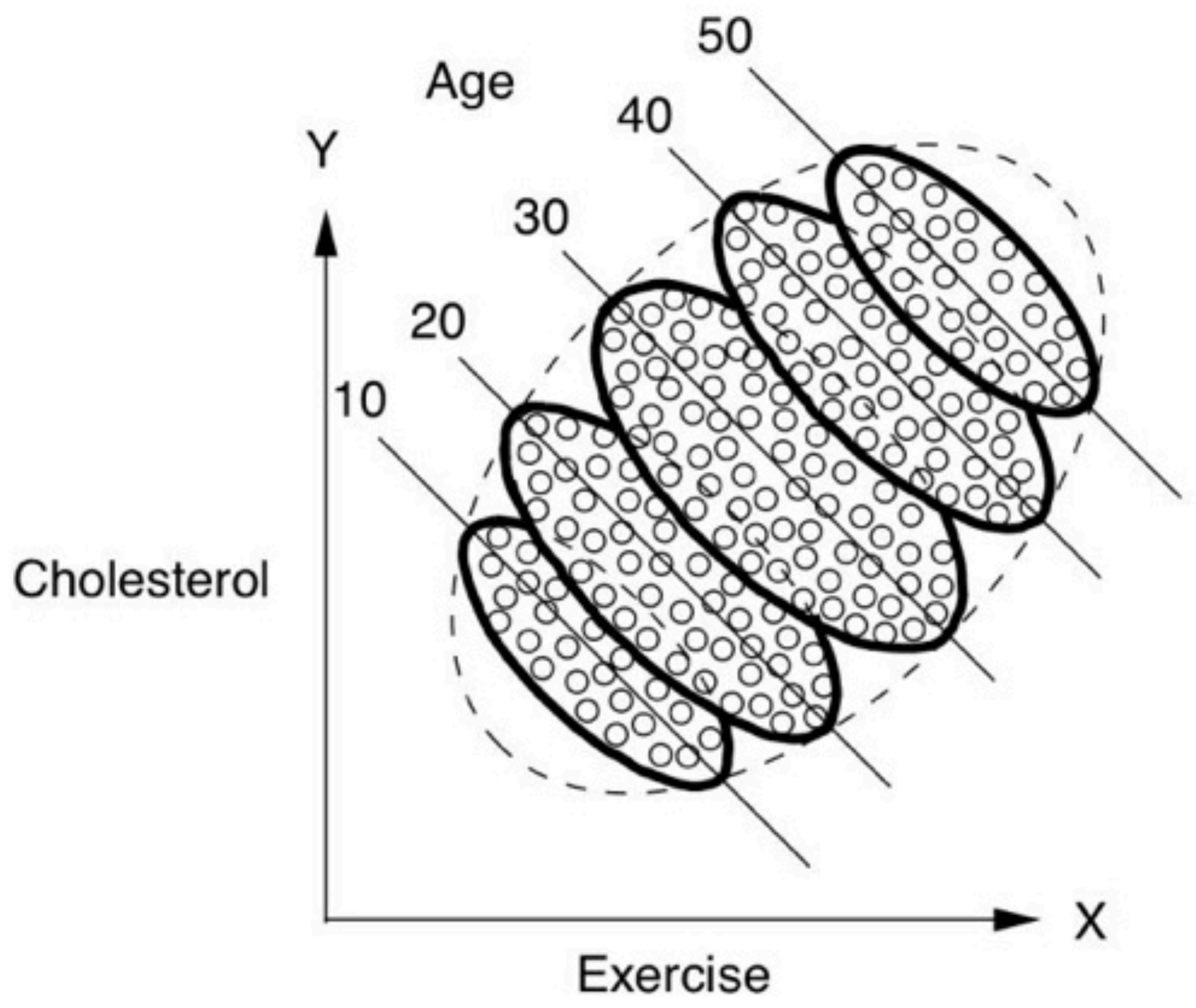
From the Book of Why [Pearl 2018]

What if we don't have an RCT? Opposite conclusion

Let's assume we have **observational data** (e.g. data collected by hospitals)



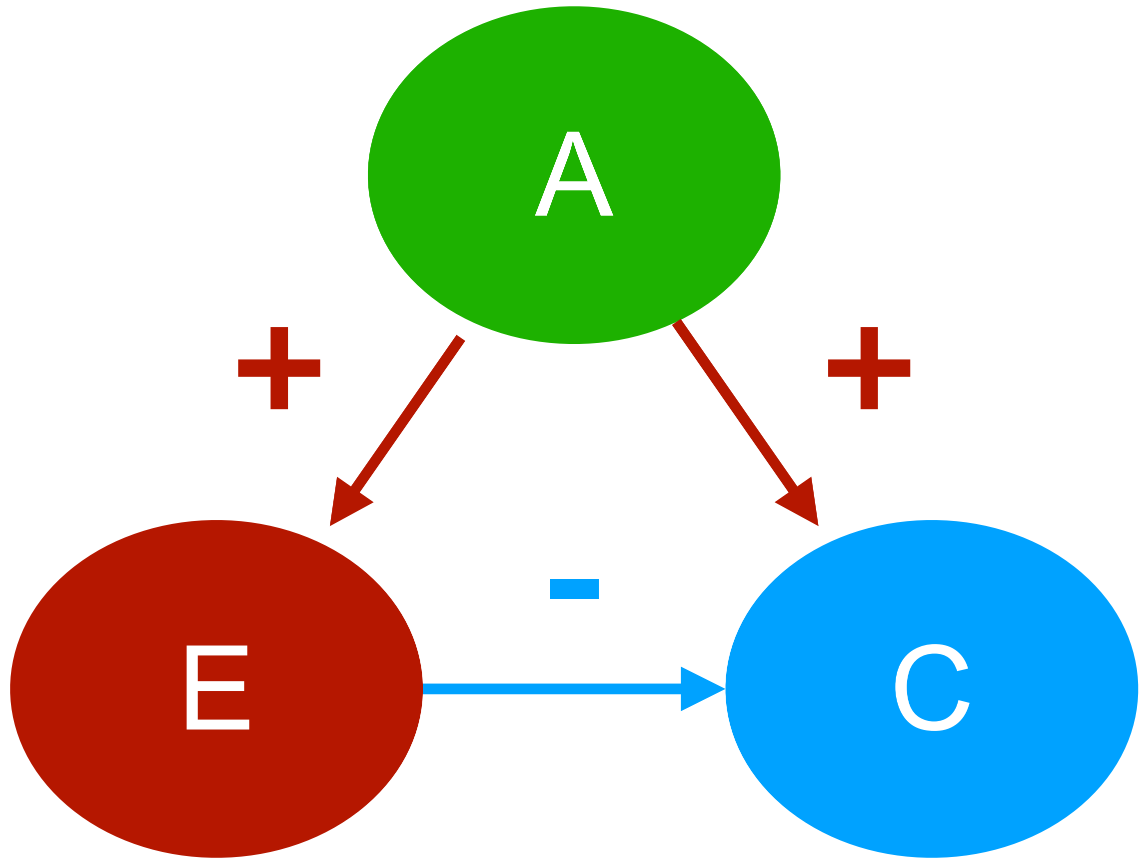
Exercise decreases cholesterol!



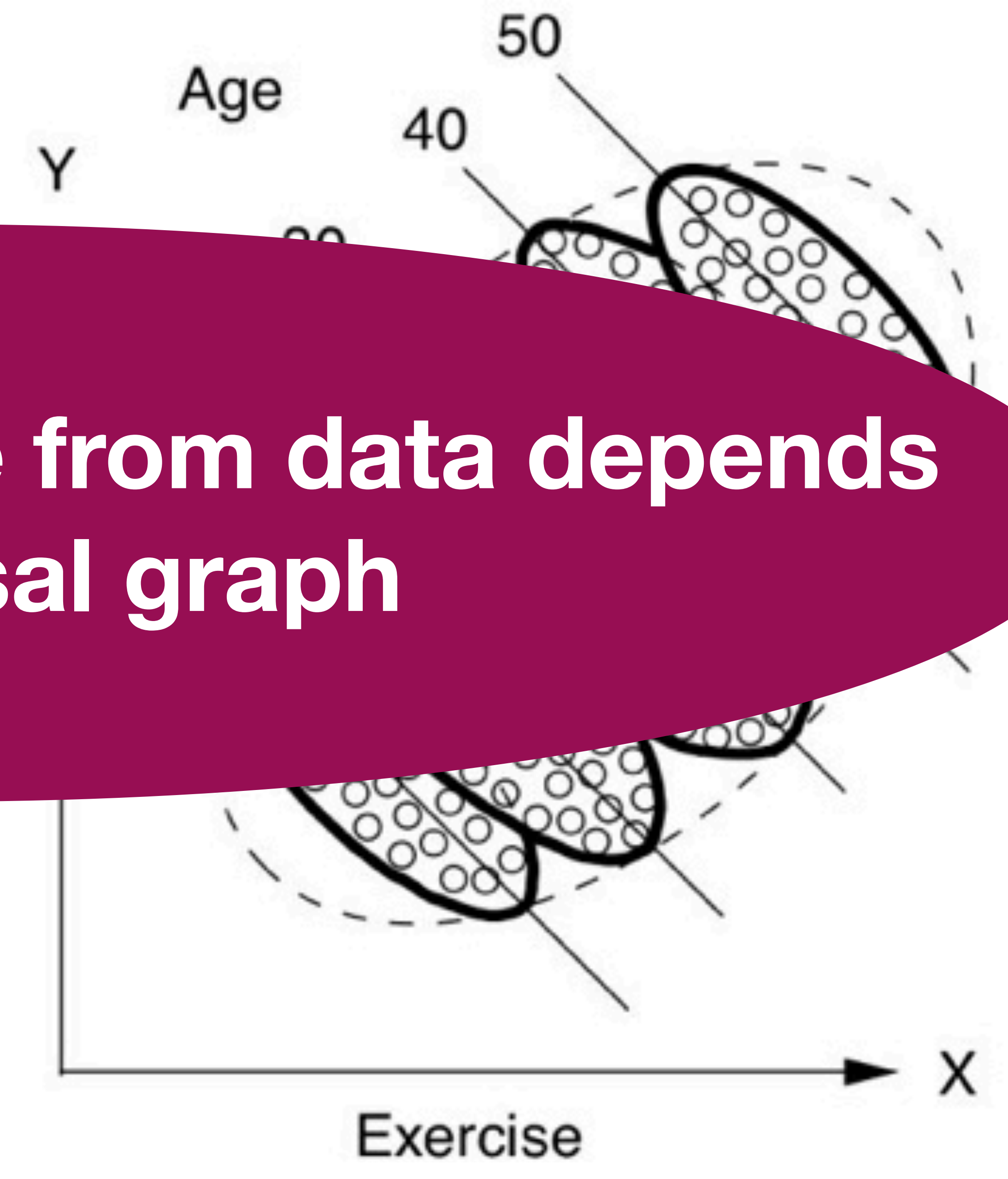
From the Book of Why [Pearl 2018]

What if we don't have an RCT? Opposite conclusion

Let's assume we have **observational data** (e.g. data collected by hospitals)



The effect we estimate from data depends on the causal graph



Exercise decreases cholesterol!

Exercise increases cholesterol??

From the Book of Why [Pearl 2018]

Causal Hierarchy [Pearl 2009, 2018]

Most ML

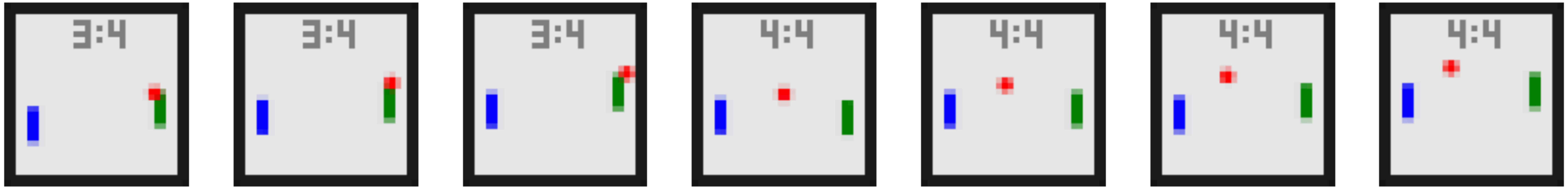
Causality

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

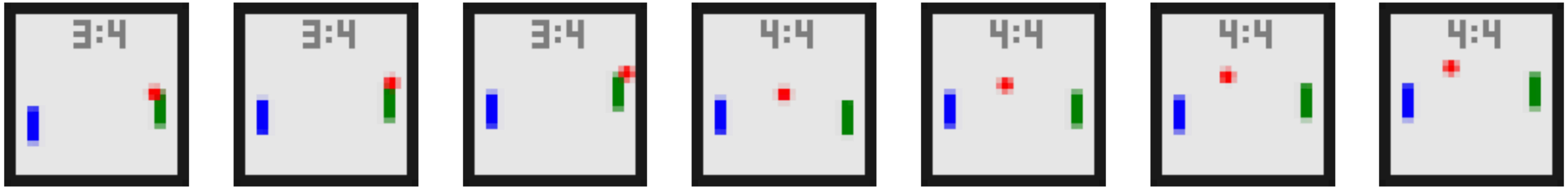
Model-based

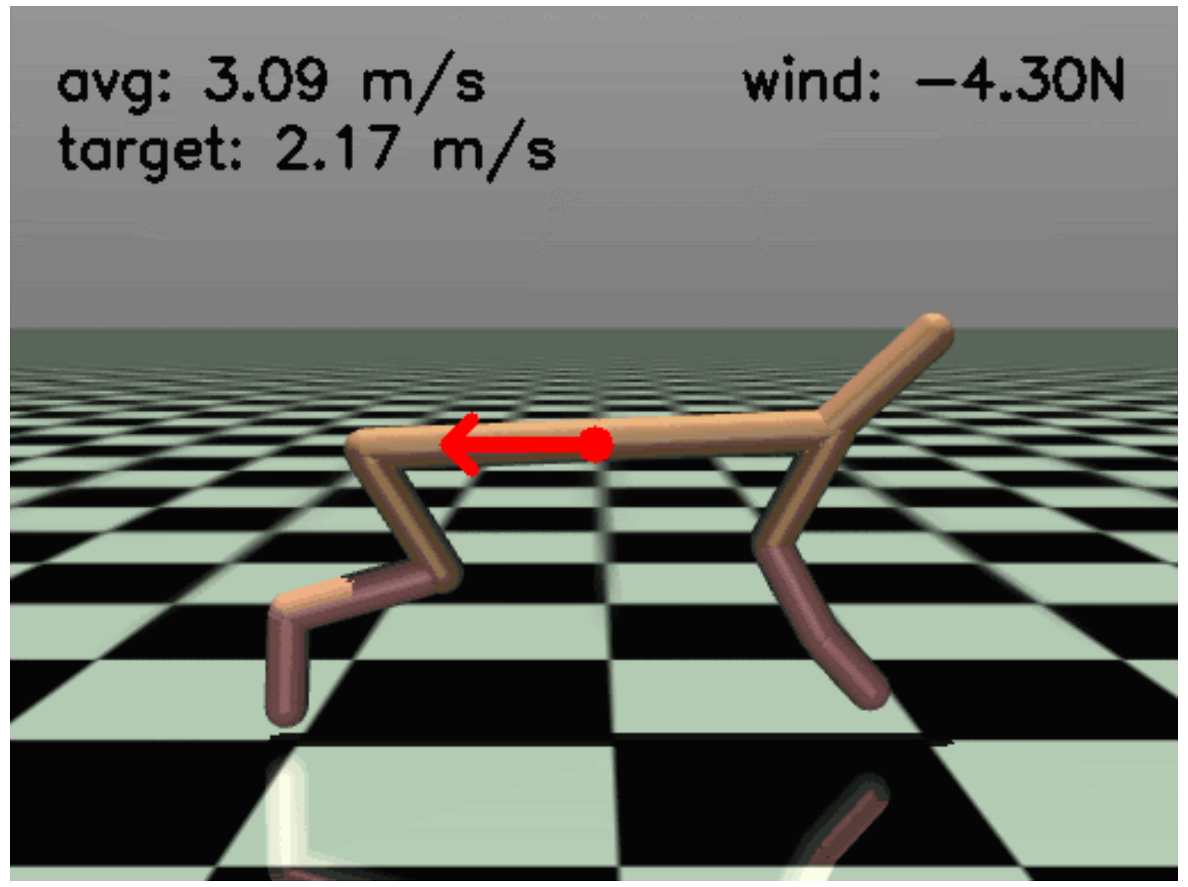
Can we learn causal variables from high-dimensional data?

Can we learn causal variables from high-dimensional data?

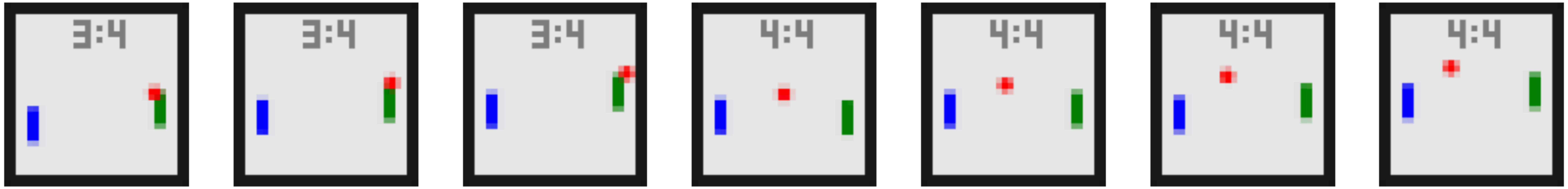


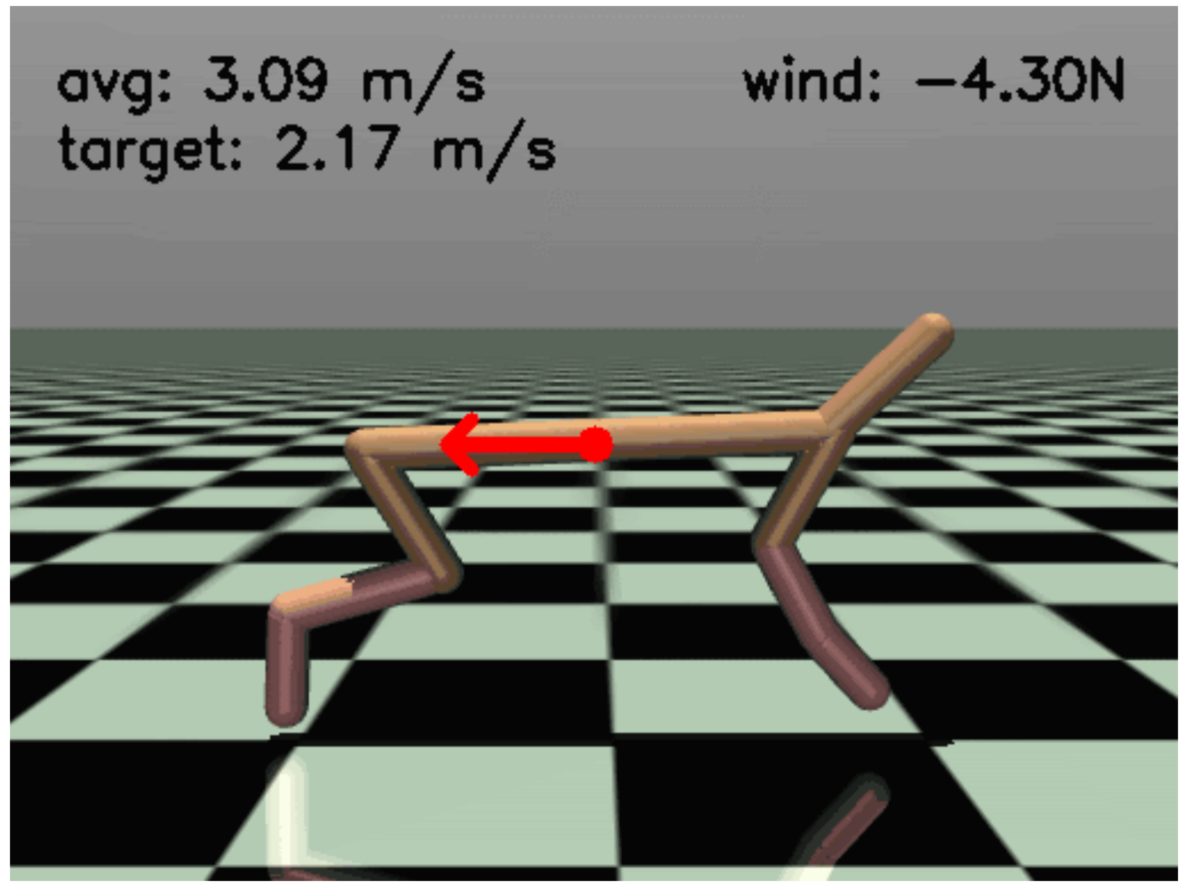
Can we learn causal variables from high-dimensional data?

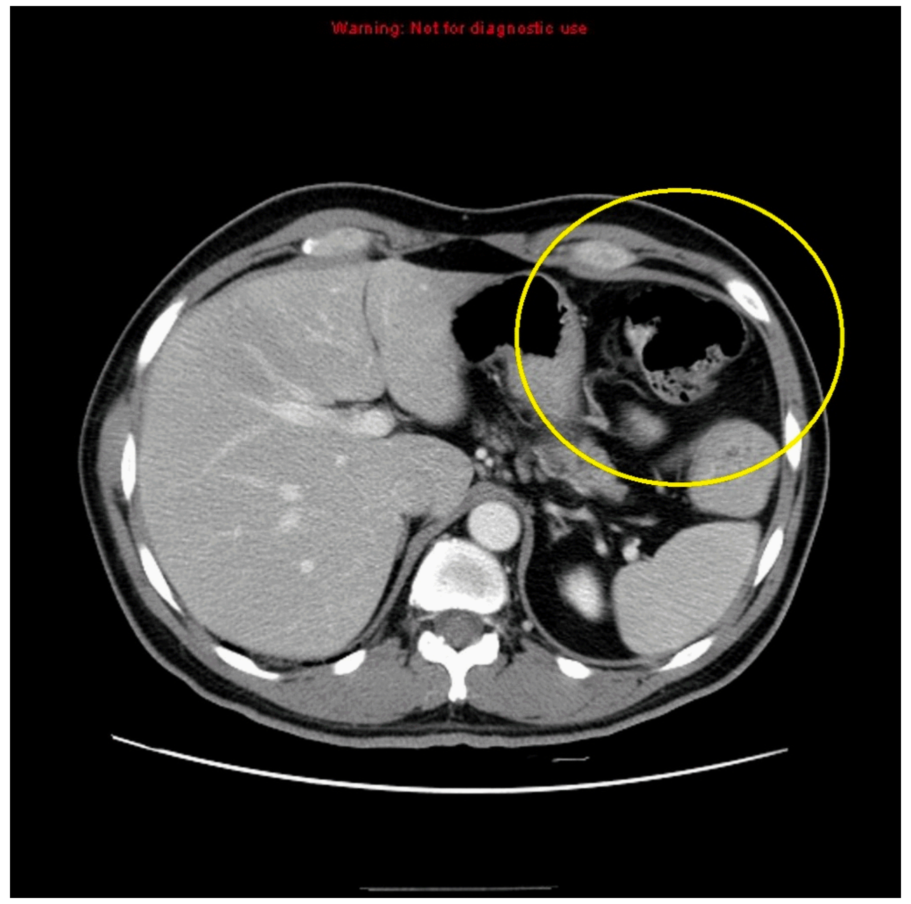




Can we learn causal variables from high-dimensional data?







Note: wishful thinking at this point of time…

https://www.wikidoc.org/index.php/File:CT_gastric_cancer.gif

Can we learn causal variables from high-dimensional data?

Towards Causal Representation Learning

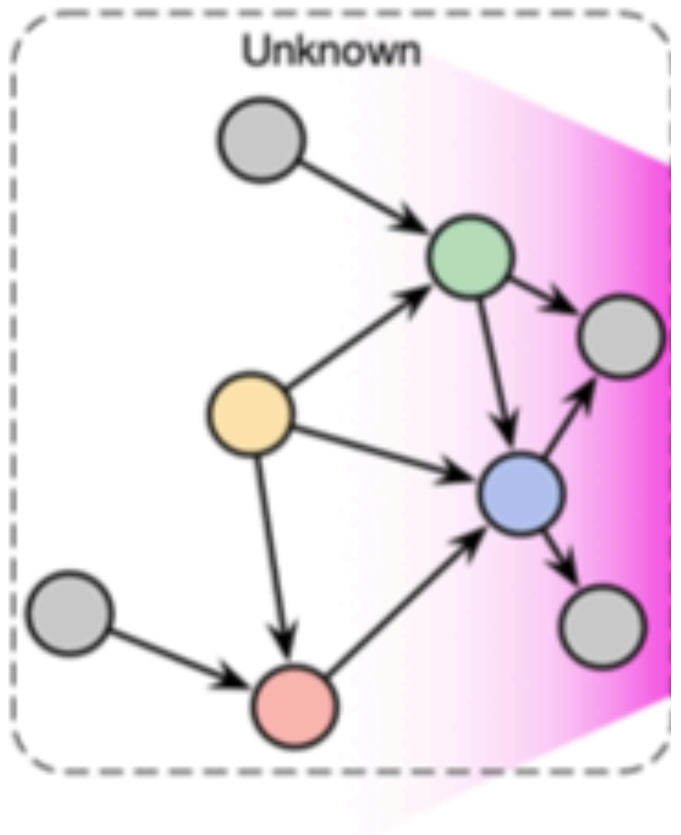
Bernhard Schölkopf †, Francesco Locatello †, Stefan Bauer *, Nan Rosemary Ke *, Nal Kalchbrenner
Anirudh Goyal, Yoshua Bengio

Abstract—**The two fields of machine learning and graphical causality arose and developed separately. However, there is now cross-pollination and increasing interest in both fields to benefit from the advances of the other. In the present paper, we review fundamental concepts of causal inference and relate them to crucial open problems of machine learning, including transfer and generalization, thereby assaying how causality can contribute to modern machine learning research. This also applies in the opposite direction: we note that most work in causality starts from the premise that the causal variables are given. A central problem for AI and causality is, thus, causal representation learning, the discovery of high-level causal variables from low-level observations. Finally, we delineate some implications of causality for machine learning and propose key research areas at the intersection of both communities.**

et al., 2018], and speech recognition [Graves et al., 2013], a substantial body of literature explored the robustness of the prediction of state-of-the-art deep neural network architectures. The underlying motivation originates from the fact that in the real world there is often little control over the distribution from which the data comes from. In computer vision [Geirhos et al., 2018, Shetty et al., 2019], changes in the test distribution may, for instance, come from aberrations like camera blur, noise or compression quality [Hendrycks and Dietterich, 2019, Karahan et al., 2016, Michaelis et al., 2019, Roy et al., 2018], or from shifts, rotations, or viewpoints [Azulay and Weiss, 2019, Barbu et al., 2019, Engstrom et al., 2017, Zhang, 2019]. Motivated by this, new benchmarks were proposed to

https://arxiv.org/abs/2102.11107

The causal representation learning problem

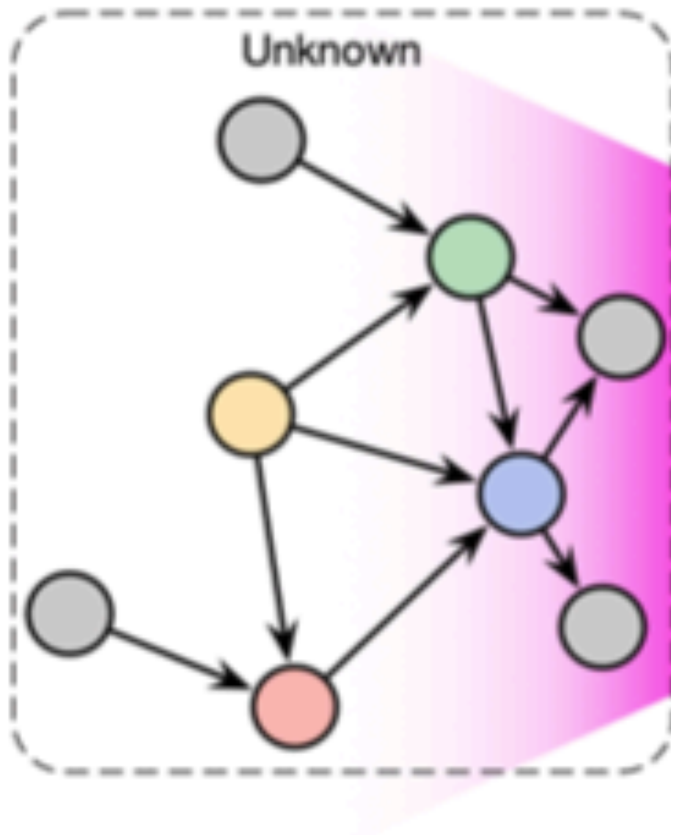


Unknown causal graph over unknown causal variables

$Z_1, \ldots, Z_d$

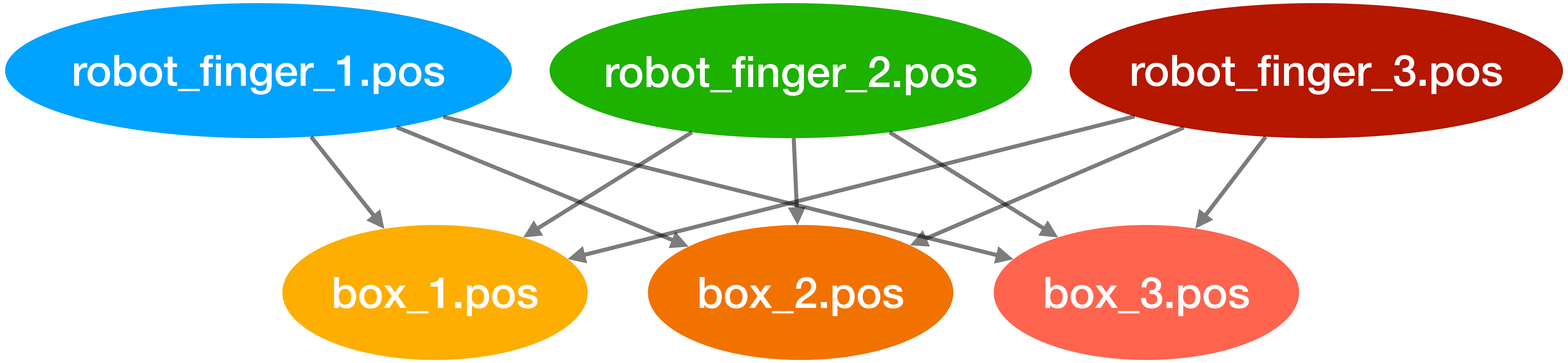
Adapted from [Schölkopf et al 2021]

The causal representation learning problem



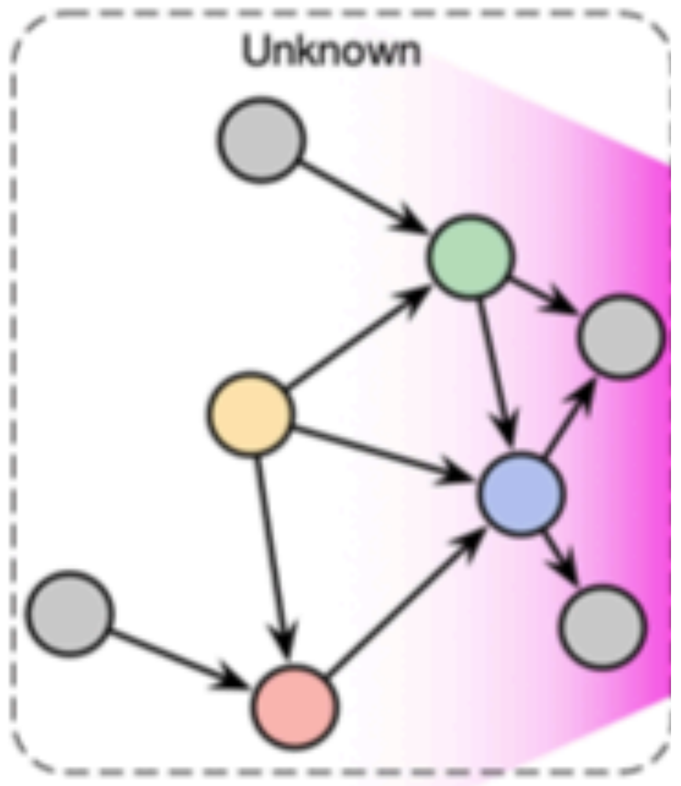
Unknown causal graph over unknown causal variables

$Z_1, \ldots, Z_d$



Adapted from [Schölkopf et al 2021]

The causal representation learning problem



Unknown causal graph over unknown causal variables

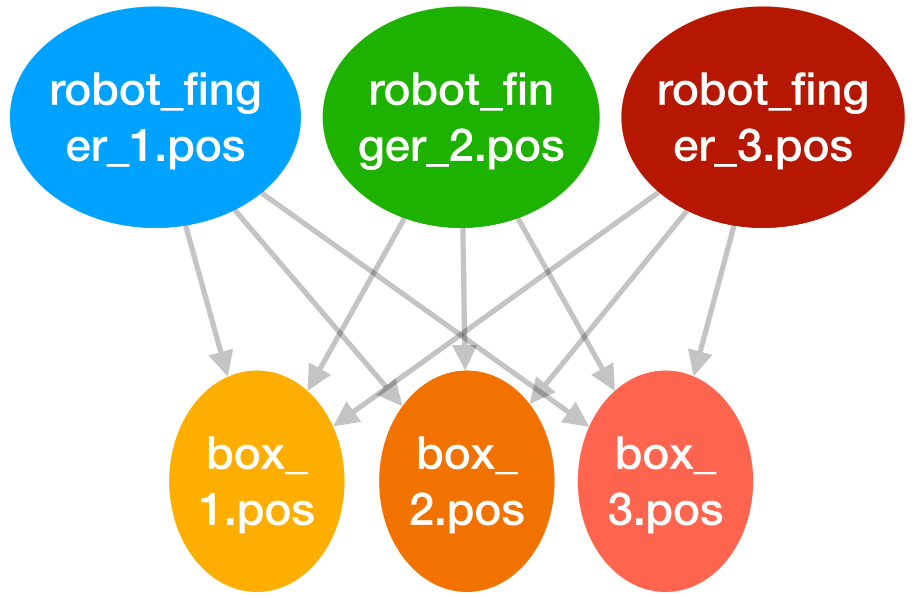
$Z_1, \ldots, Z_d$

Now we have only an entangled measurement of the causal variables (e.g. image)

Adapted from [Schölkopf et al 2021]

The causal representation learning problem (simplified)

Mixing function



Unknown causal graph over unknown causal variables

$Z_1, \ldots, Z_d$

Sensor measurements as an entangled view

$X_1, \ldots, X_p$

$p \gg d$

Adapted from [Schölkopf et al 2021]

The causal representation learning problem (simplified)

Mixing function **Encoder** **Decoder**

Unknown causal graph over unknown causal variables

$Z_1, \ldots, Z_d$

Sensor measurements as an entangled view

$X_1, \ldots, X_p$

$p \gg d$

Causal graph over reconstructed causal variables

$\hat{Z}_1, \ldots, \hat{Z}_n$

Reconstructed sensor measurements

$\hat{X}_1, \ldots, \hat{X}_p$

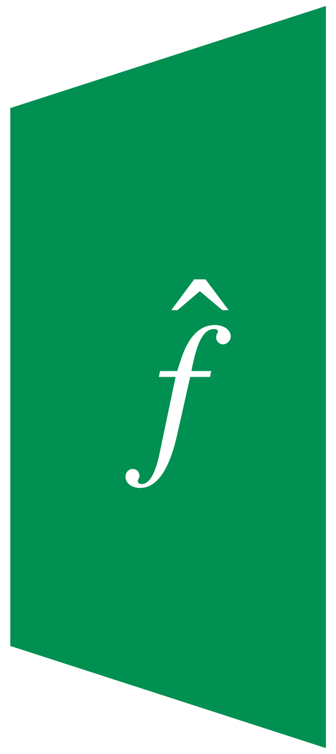
$p \gg d$

Adapted from [Schölkopf et al 2021]

The causal representation learning problem: issue

Mixing function **Encoder** **Decoder**

Issue: in general the latent space of a VAE does not disentangle the causal factors!



Unknown causal graph over unknown causal variables

$Z_1, \ldots, Z_d$

Sensor measurements as an entangled view

$X_1, \ldots, X_p$

$p \gg d$

Reconstructed sensor measurements

$\hat{X}_1, \ldots, \hat{X}_p$

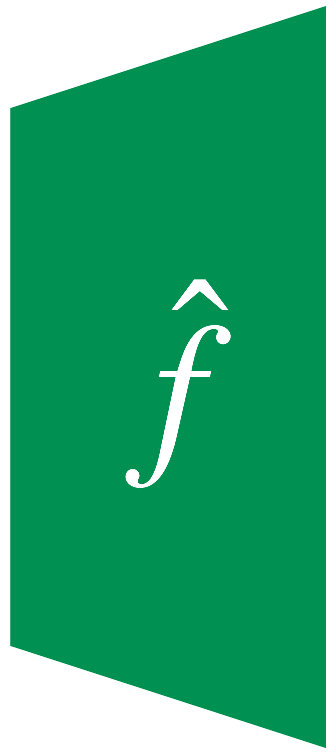
$p \gg d$

Adapted from [Schölkopf et al 2021]

The causal representation learning problem: issue

Mixing function | **Encoder** | **Decoder**

Issue: in general the latent space of a VAE does not disentangle the causal factors!



Unknown causal graph over unknown causal variables

$Z_1, \ldots, Z_d$

Sensor measurements as an entangled view

$X_1, \ldots, X_p$

$p \gg d$

We need extra assumptions to prove identifiability (and usually only up to some equivalence class)

Reconstructed sensor measurements

$\hat{X}_1, \ldots, \hat{X}_p$

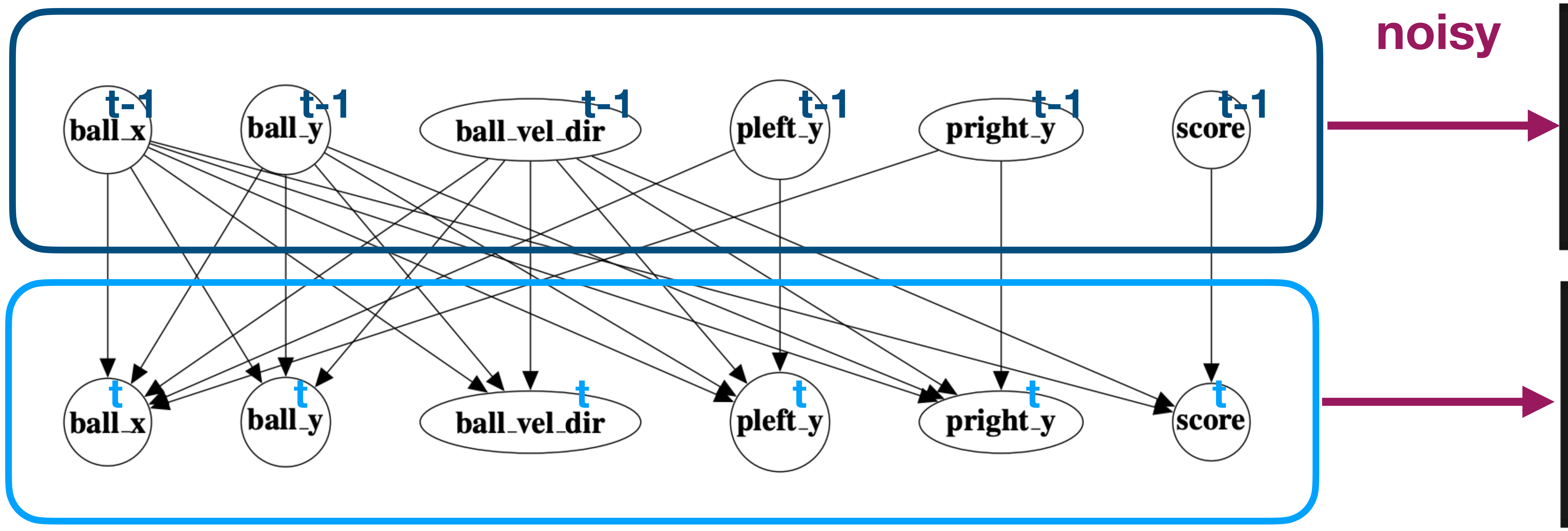
$p \gg d$

Adapted from [Schölkopf et al 2021]

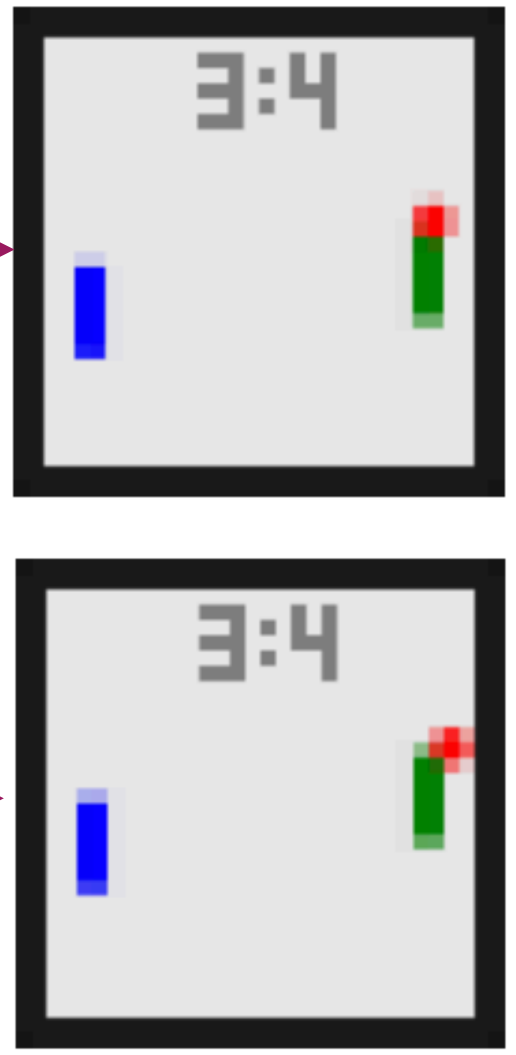
CITRIS: Causal Identifiability from TempoRal Intervened Sequences

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves

ICML 2022



Possibly noisy

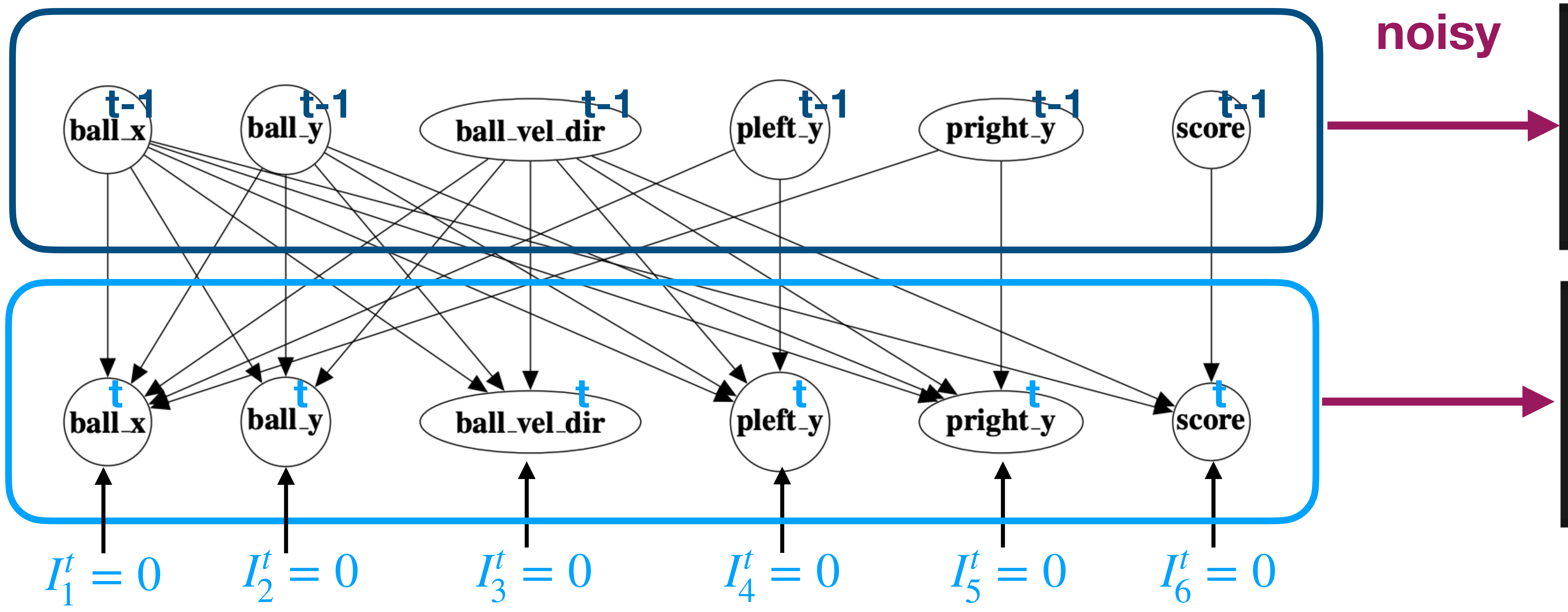


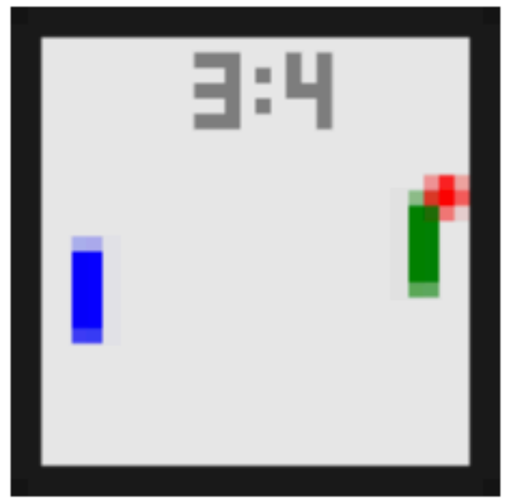
https://arxiv.org/abs/2202.03169

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves **ICML 2022**

Possibly noisy

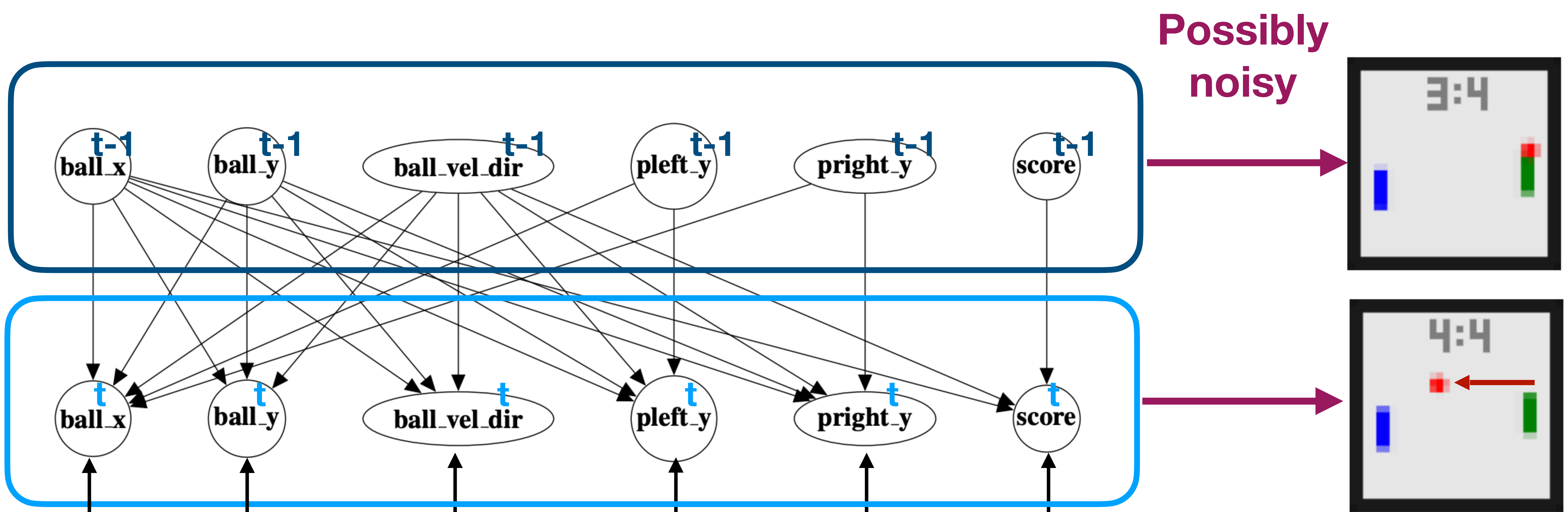




https://arxiv.org/abs/2202.03169

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves ICML 2022



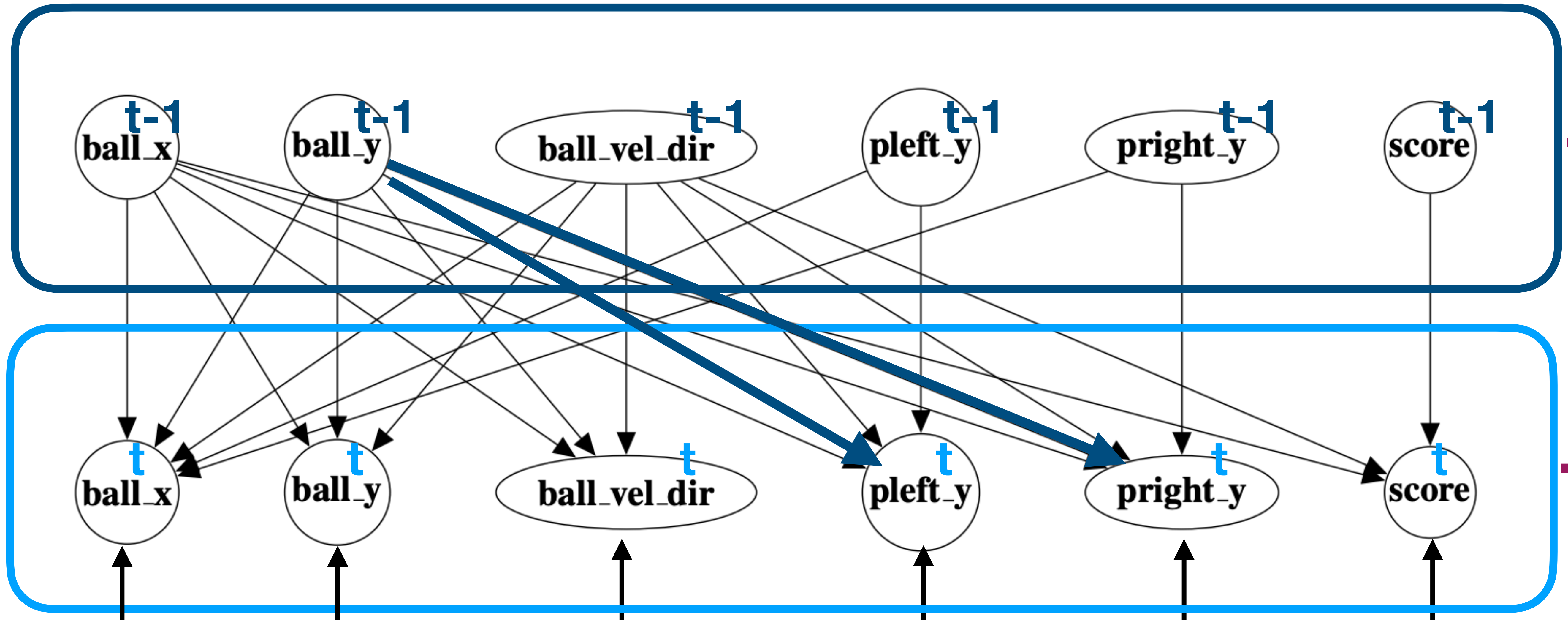
$I_1^t = 1$ $I_2^t = 0$ $I_3^t = 0$ $I_4^t = 0$ $I_5^t = 0$ $I_6^t = 0$

Stochastic intervention
(we don't know where the ball will be)

https://arxiv.org/abs/2202.03169

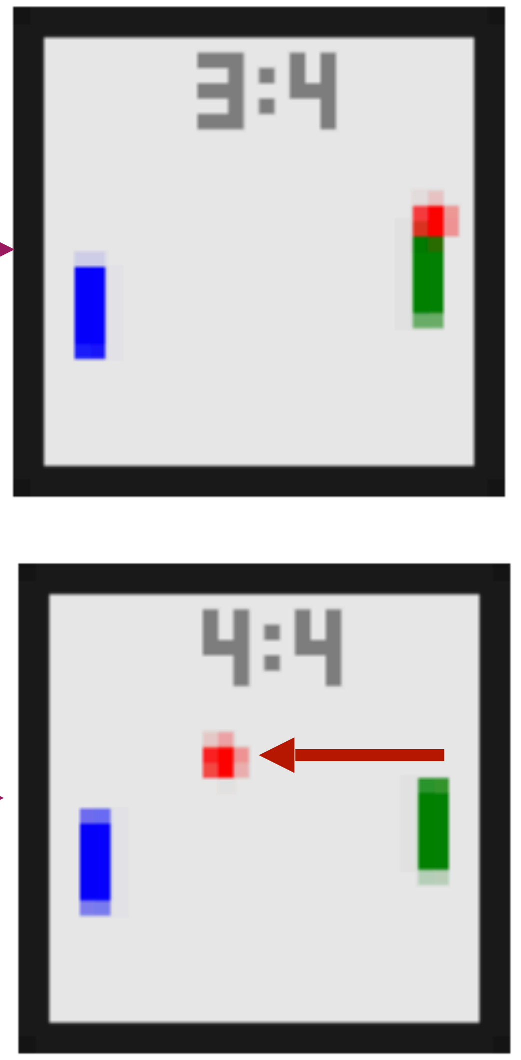
CITRIS: Causal Identifiability from TempoRal Intervened Sequences

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves ICML 2022



$I_1^t = 1$ $I_2^t = 0$ $I_3^t = 0$ $I_4^t = 0$ $I_5^t = 0$ $I_6^t = 0$

Possibly noisy



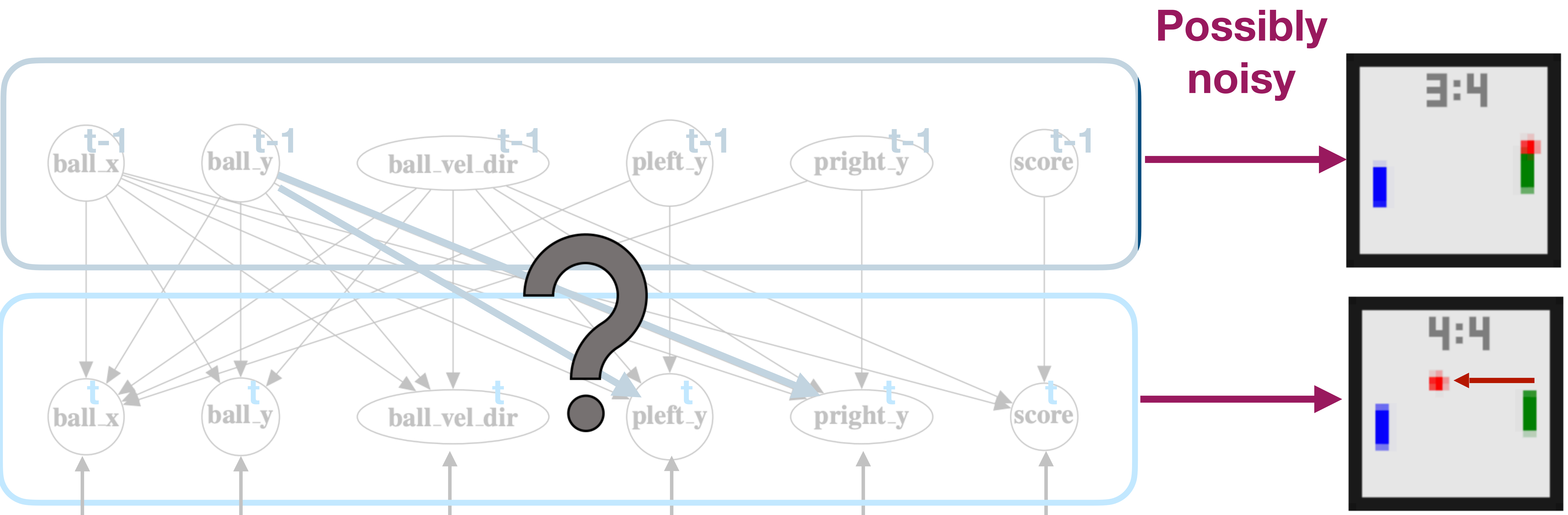
Stochastic intervention
(we don't know where the ball will be)

The paddles continue moving as usual (not counterfactual)

https://arxiv.org/abs/2202.03169

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves ICML 2022



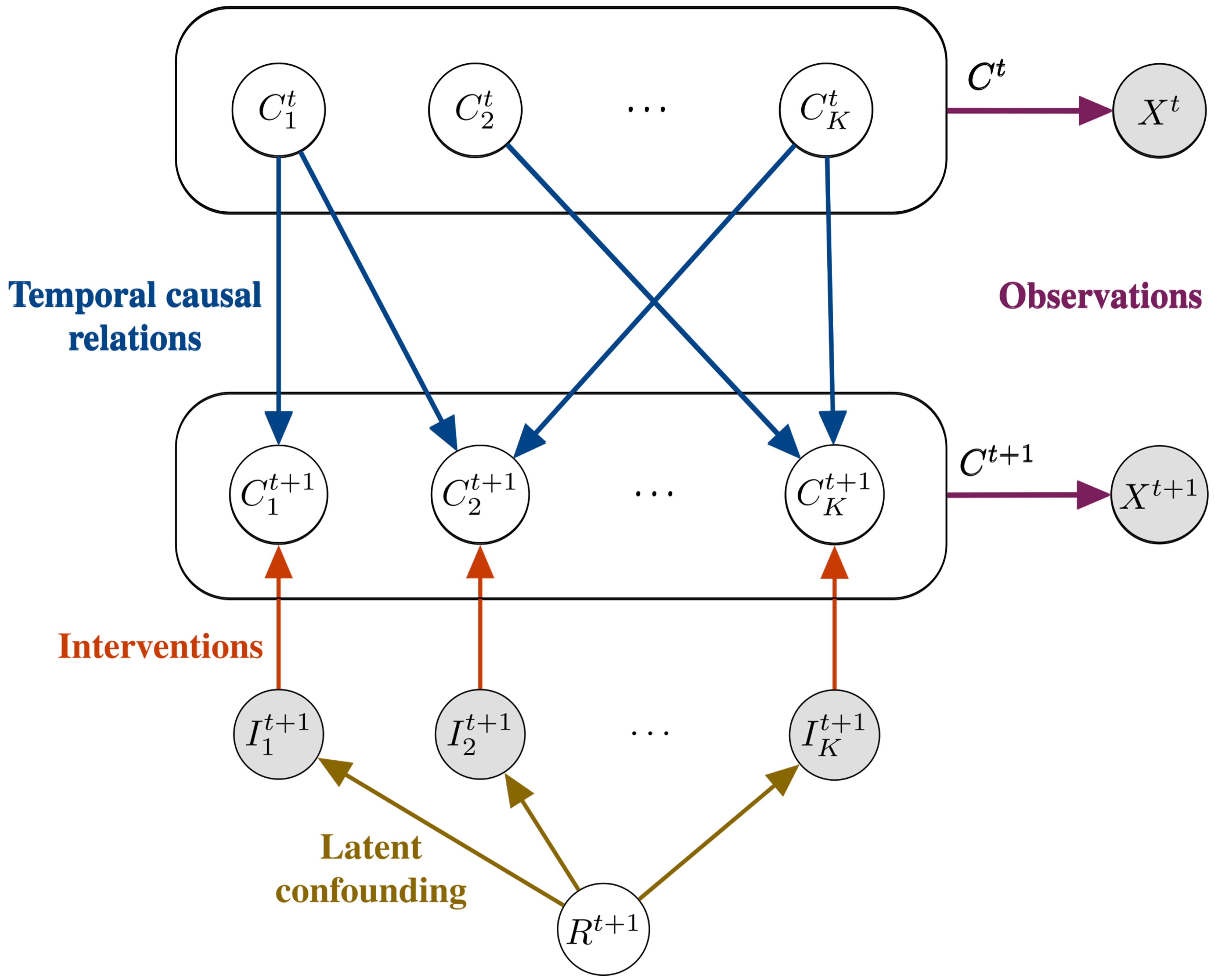
$I_1^t = 1$ $I_2^t = 0$ $I_3^t = 0$ $I_4^t = 0$ $I_5^t = 0$ $I_6^t = 0$

Stochastic intervention
(we don't know where the ball will be)

https://arxiv.org/abs/2202.03169

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

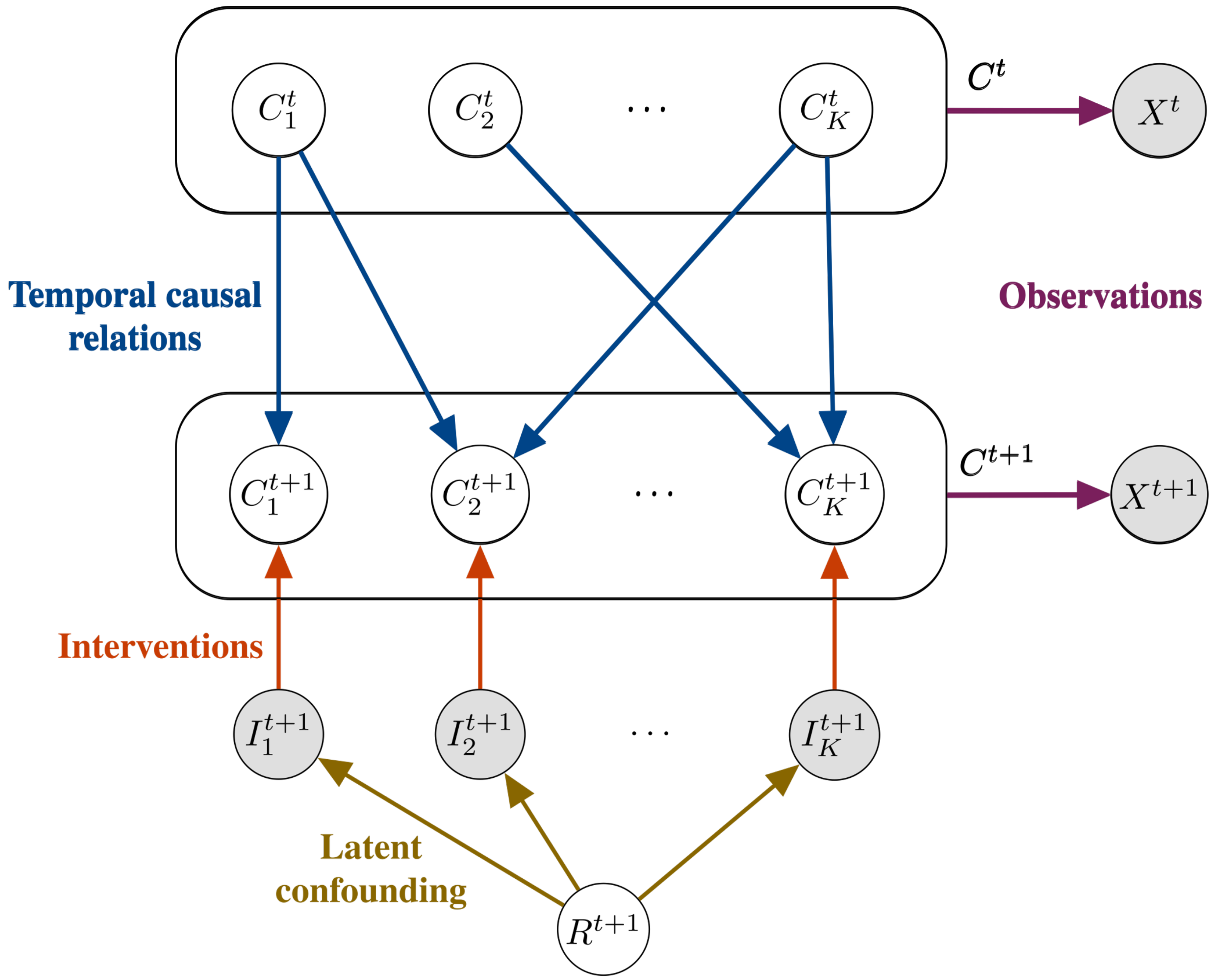
Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves **ICML 2022**



- We want to learn the underlying causal process from **temporal sequences** of **high-dimensional data** $\{X^t\}_{t=1}^T$, e.g. images

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves **ICML 2022**

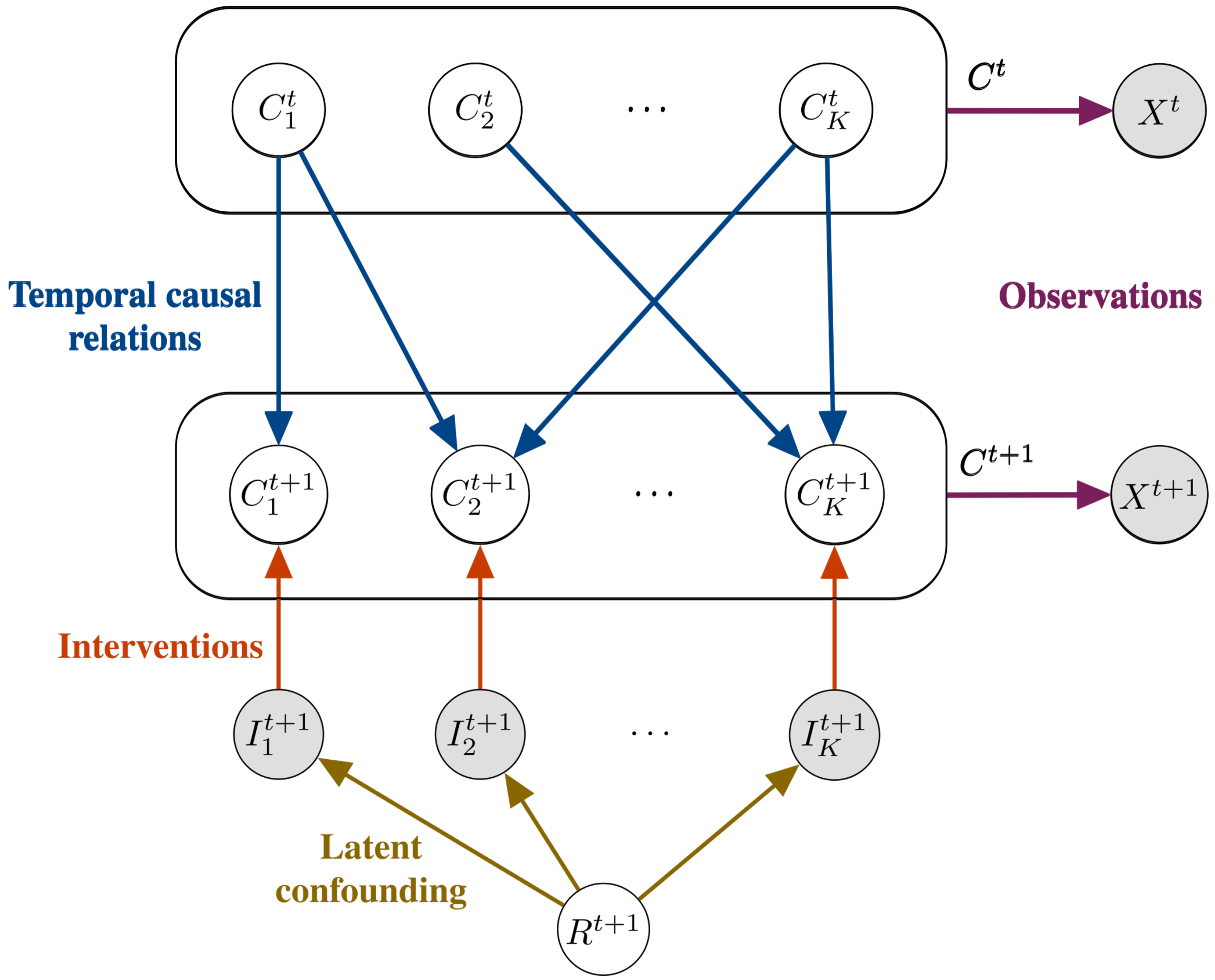


- We want to learn the underlying causal process from **temporal sequences** of **high-dimensional data** $\{X^t\}_{t=1}^T$, e.g. images
- We assume that the **latent** causal process is a **Dynamic Bayesian network** with **K multidimensional causal variables**

$$X^t = h(C_1^t, \ldots C_K^t, E_0^t)$$

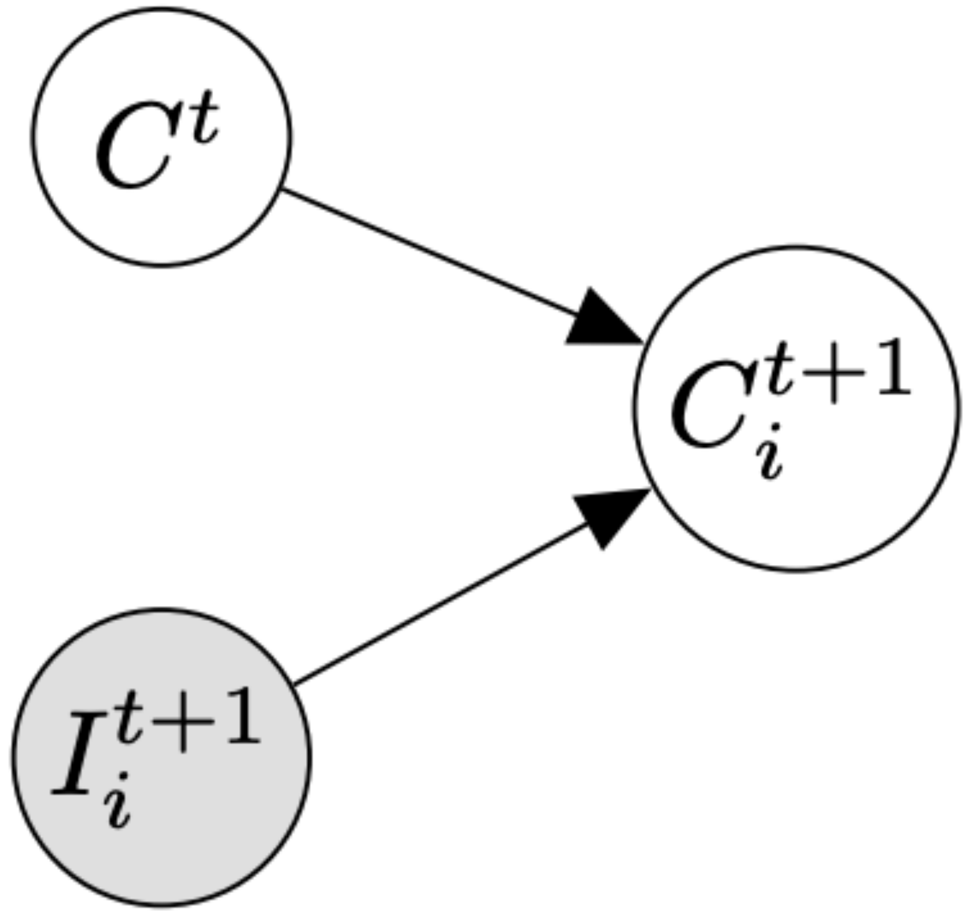
CITRIS: Causal Identifiability from TempoRal Intervened Sequences

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves **ICML 2022**



- We want to learn the underlying causal process from **temporal sequences** of **high-dimensional data** $\{X^t\}_{t=1}^T$, e.g. images
- We assume that the **latent** causal process is a **Dynamic Bayesian network** with **K multidimensional causal variables**
- We assume that **(soft or perfect) interventions** can happen on the underlying system and **we observe the targets** I_i^t
  - $I_i \to C_i$

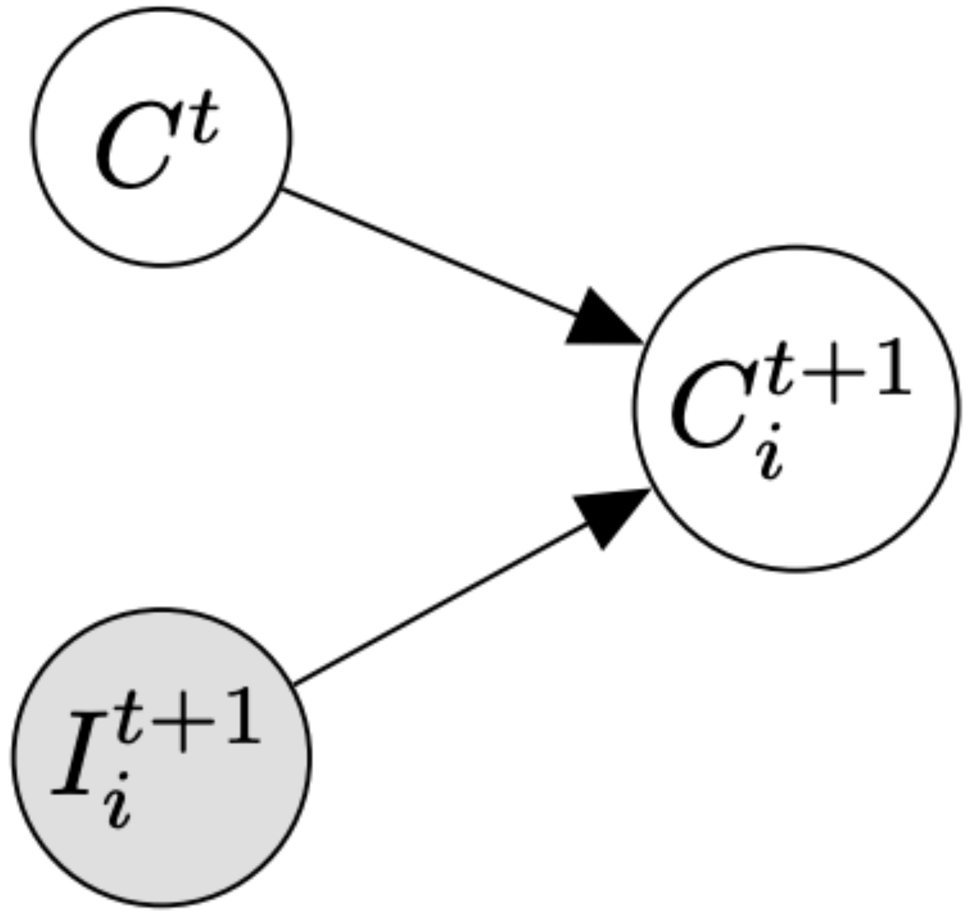
Minimal causal variables - theory

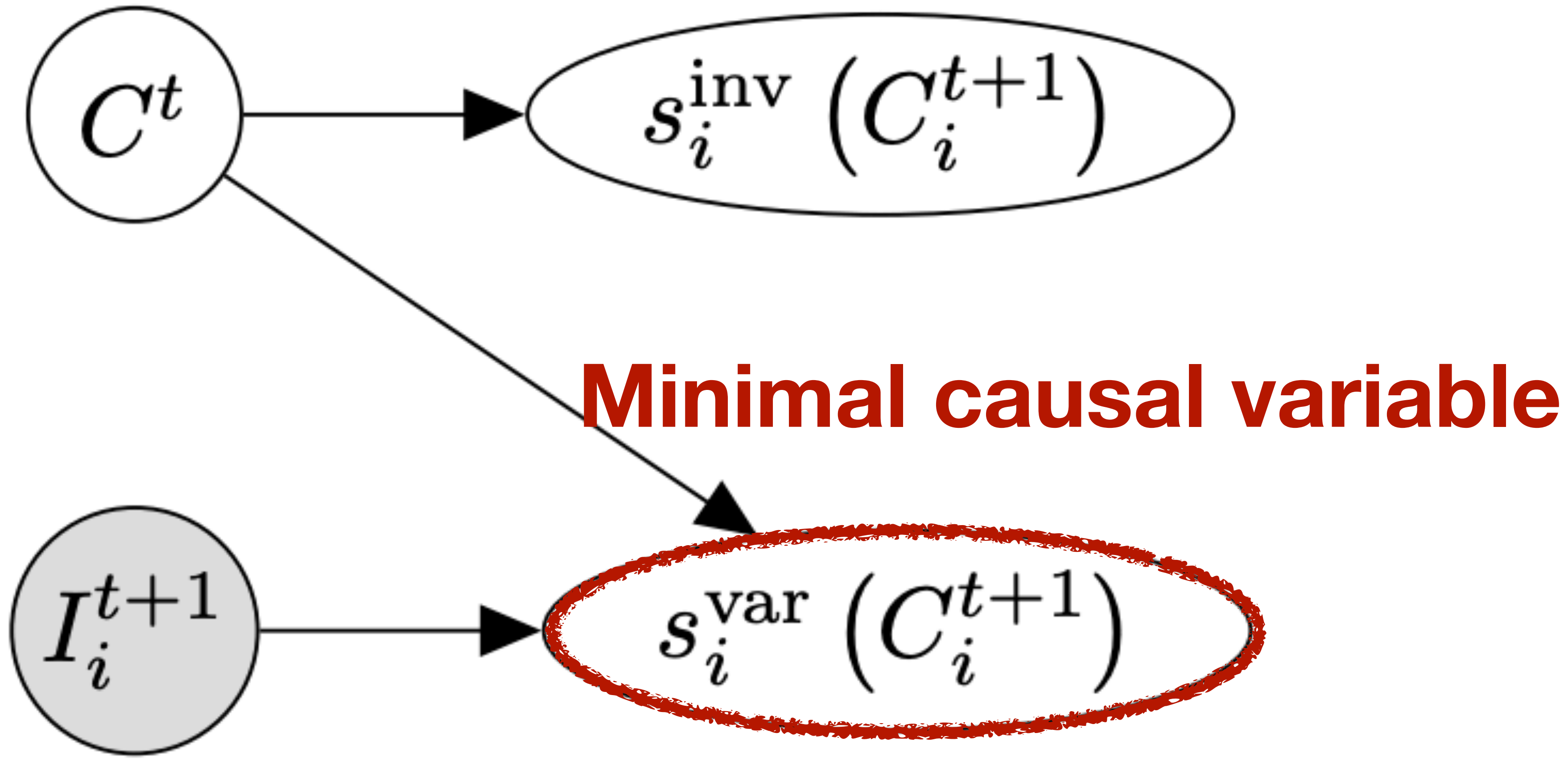


Minimal causal variables - theory

- We can define a split of a causal variable C_i^t

$$s_i(C_i^t) = (s_i^{\text{var}}(C_i^t), s_i^{\text{inv}}(C_i^t))$$

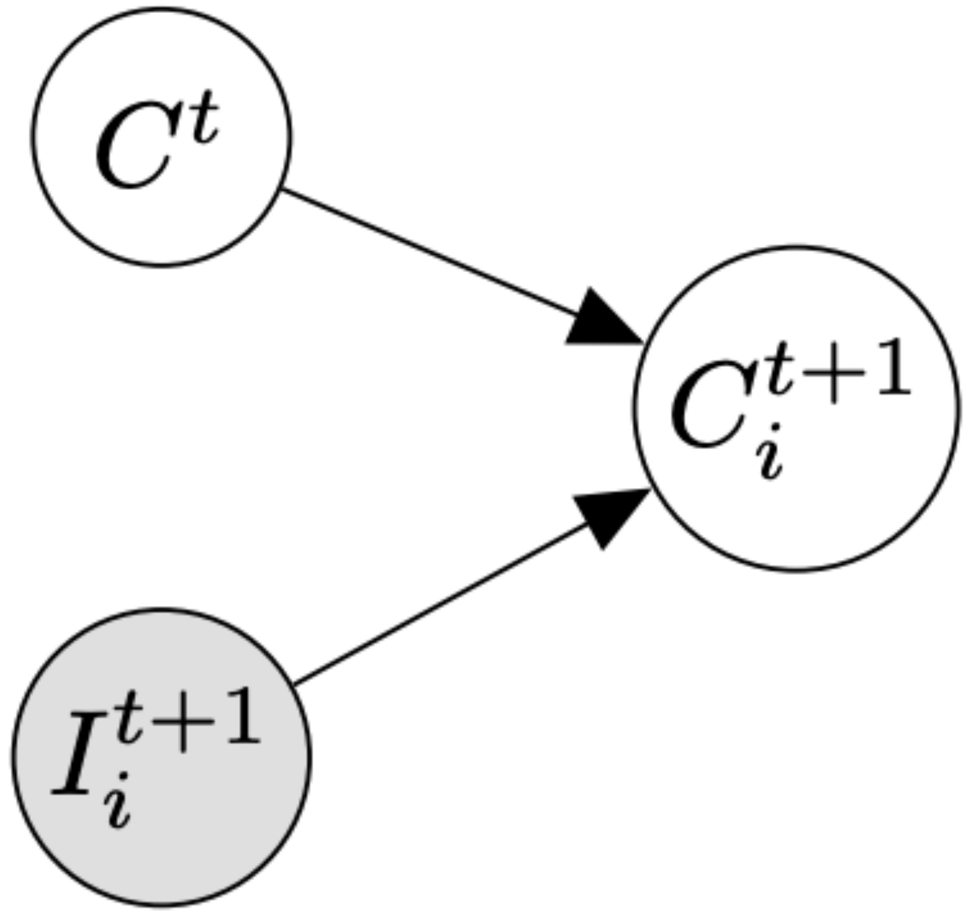


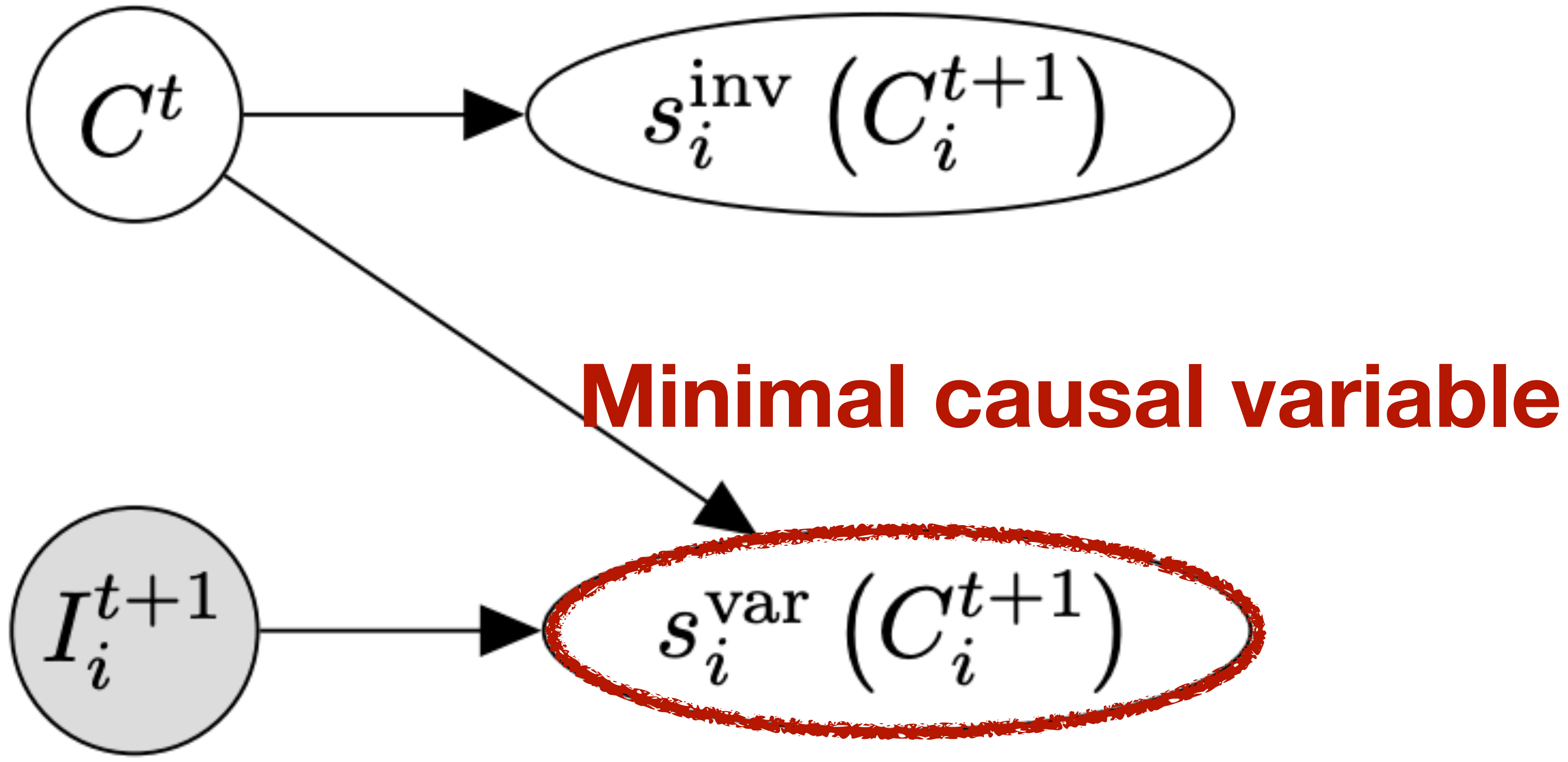


Minimal causal variables - theory

- We can define a split of a causal variable C_i^t

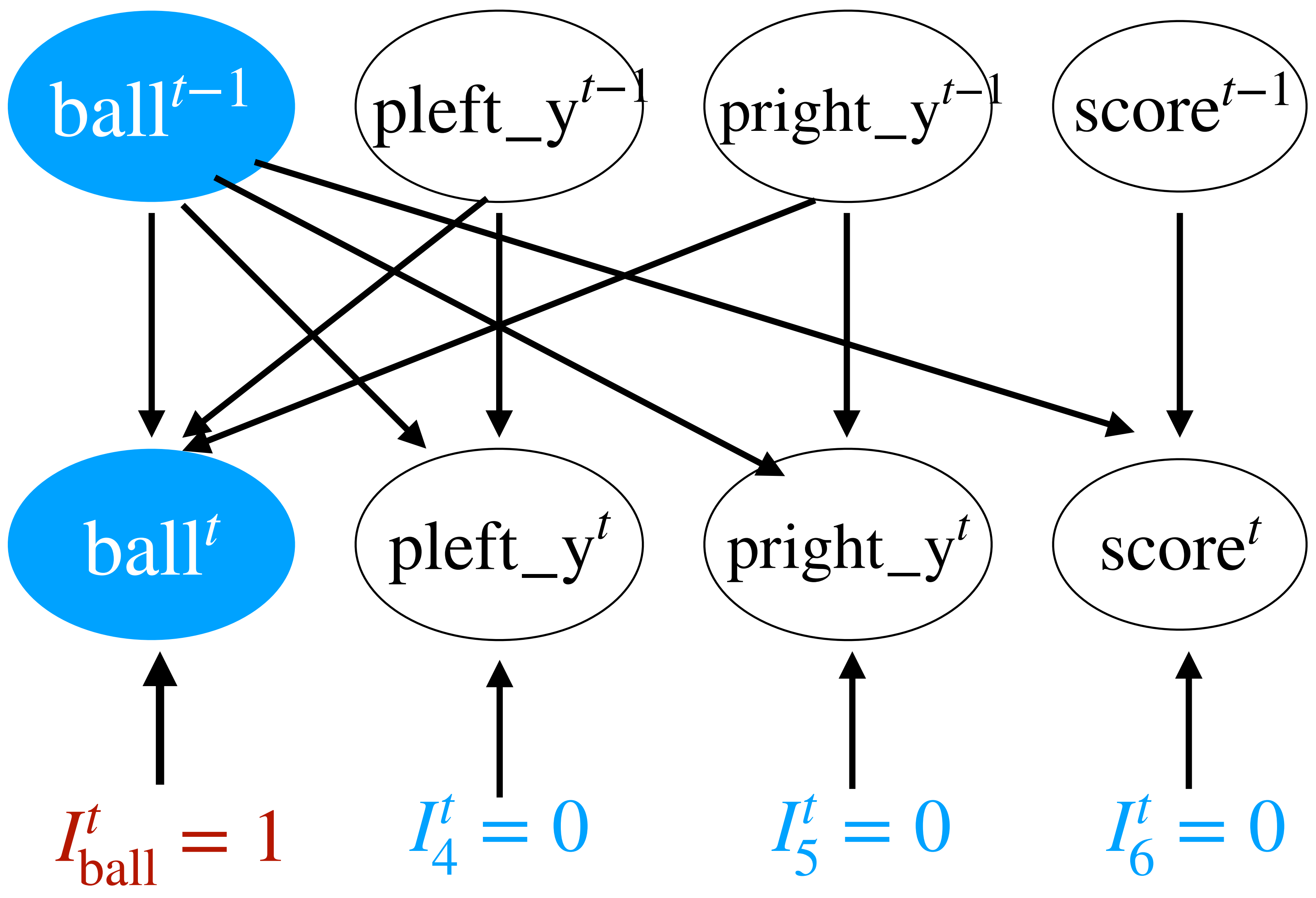
$$s_i(C_i^t) = (s_i^{\text{var}}(C_i^t), s_i^{\text{inv}}(C_i^t))$$





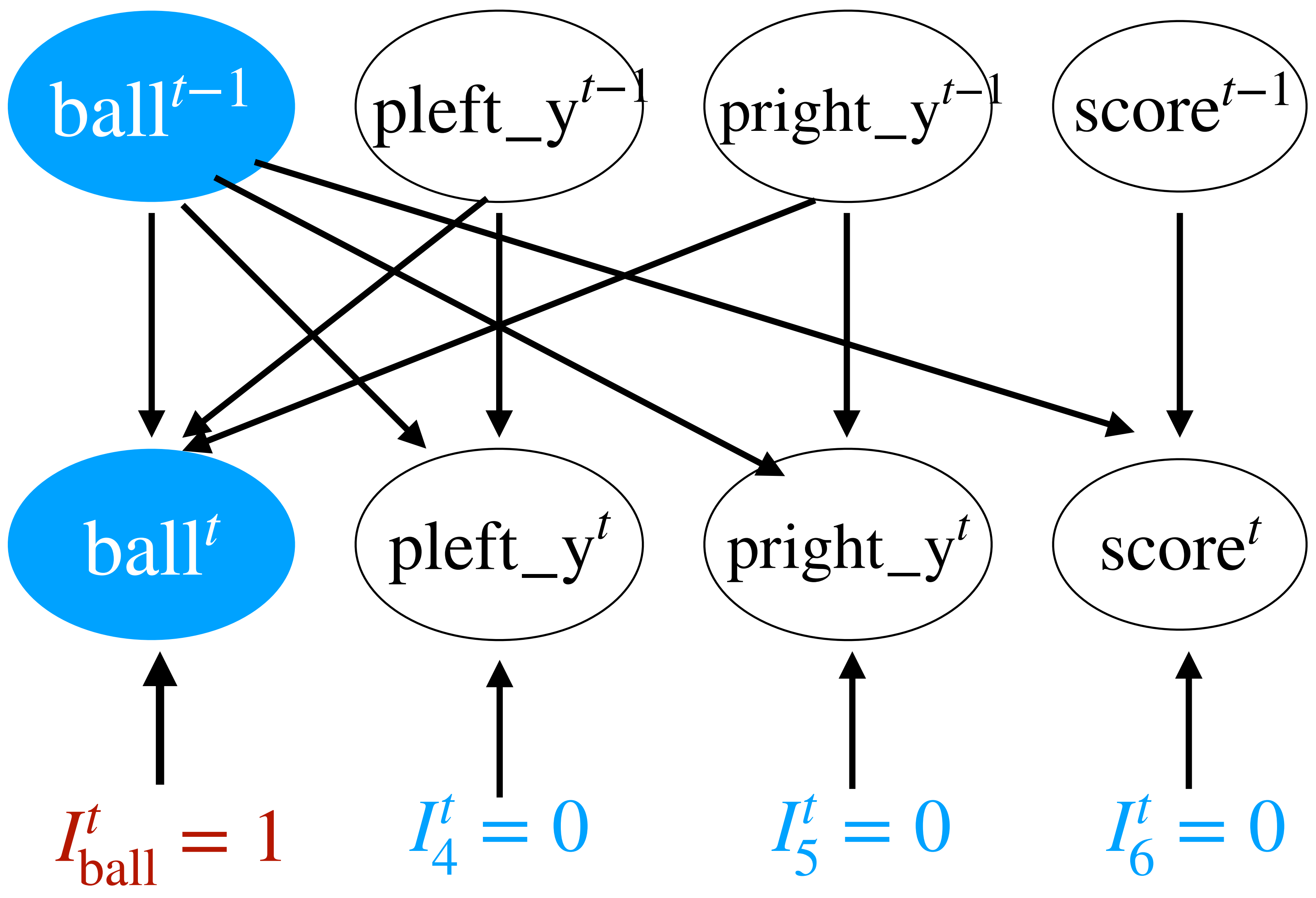
- We choose $s_i^{\text{var}*}$ that contains only the information that depends on I_i^{t+1}
- We can identify minimal causal variables **up to invertible component-wise transformations**, if I_i^t is not a deterministic function of I_j^t

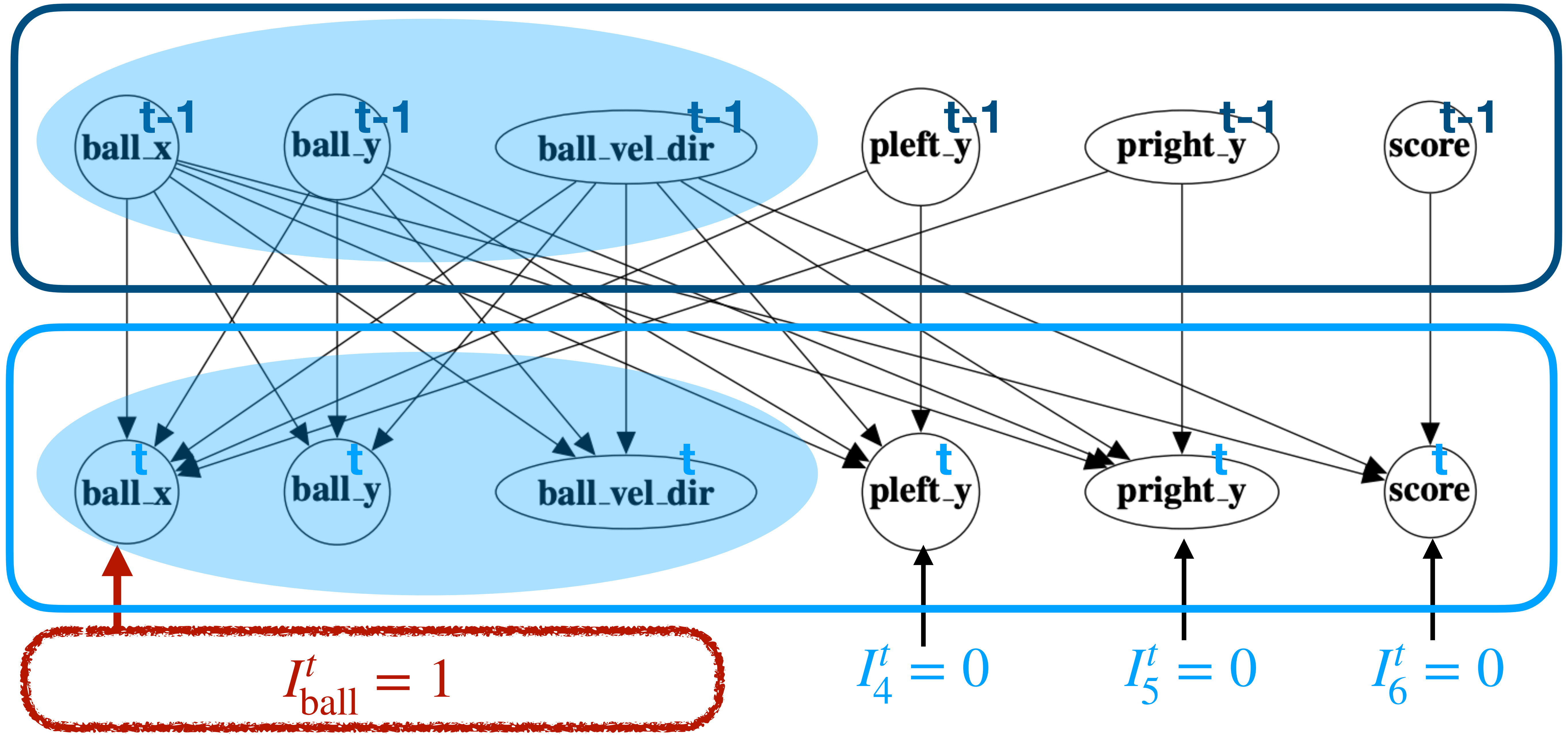
Minimal causal variables - example



The intervention only has an effect on a part of the variable
(e.g.only on ball_x)

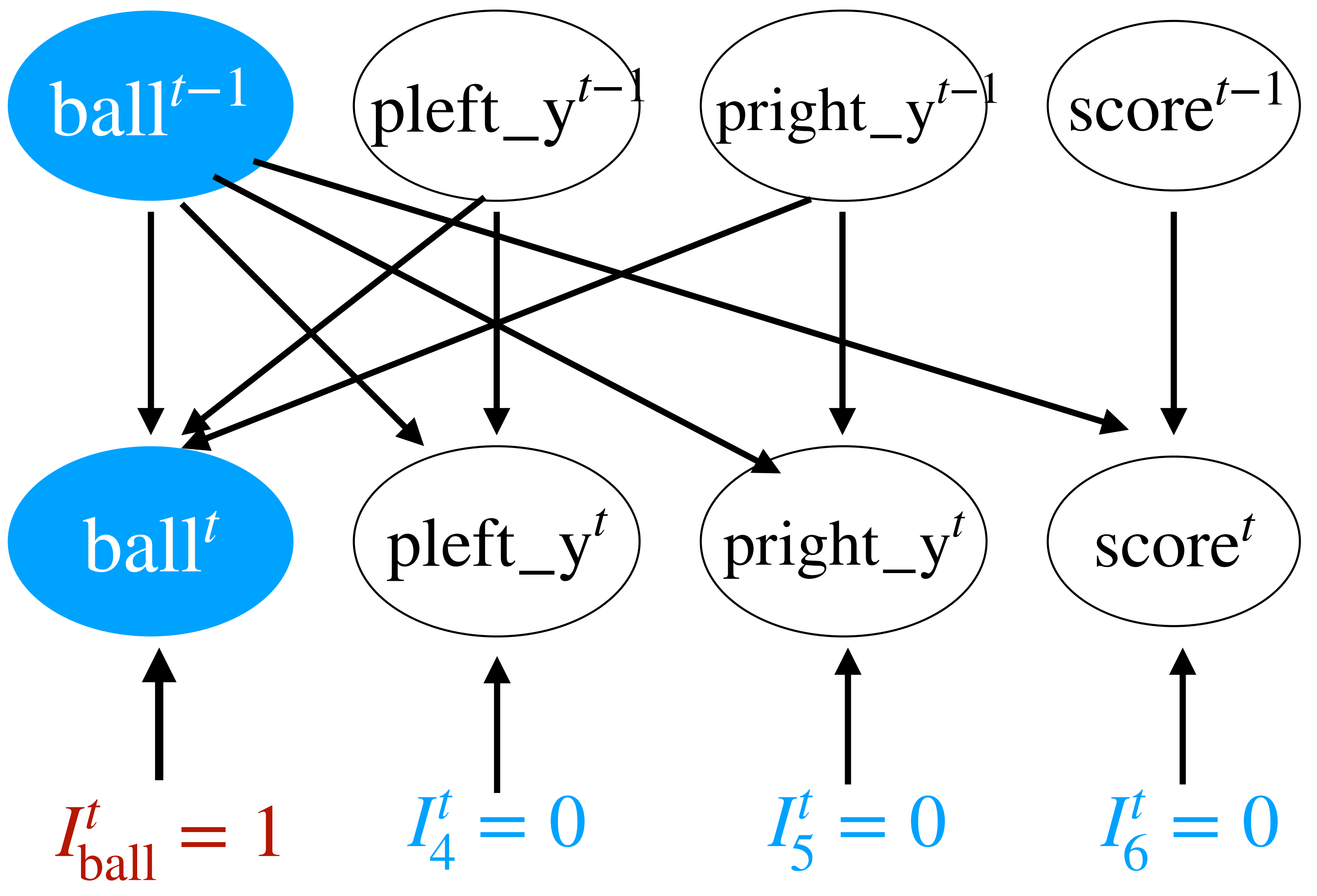
Minimal causal variables - example

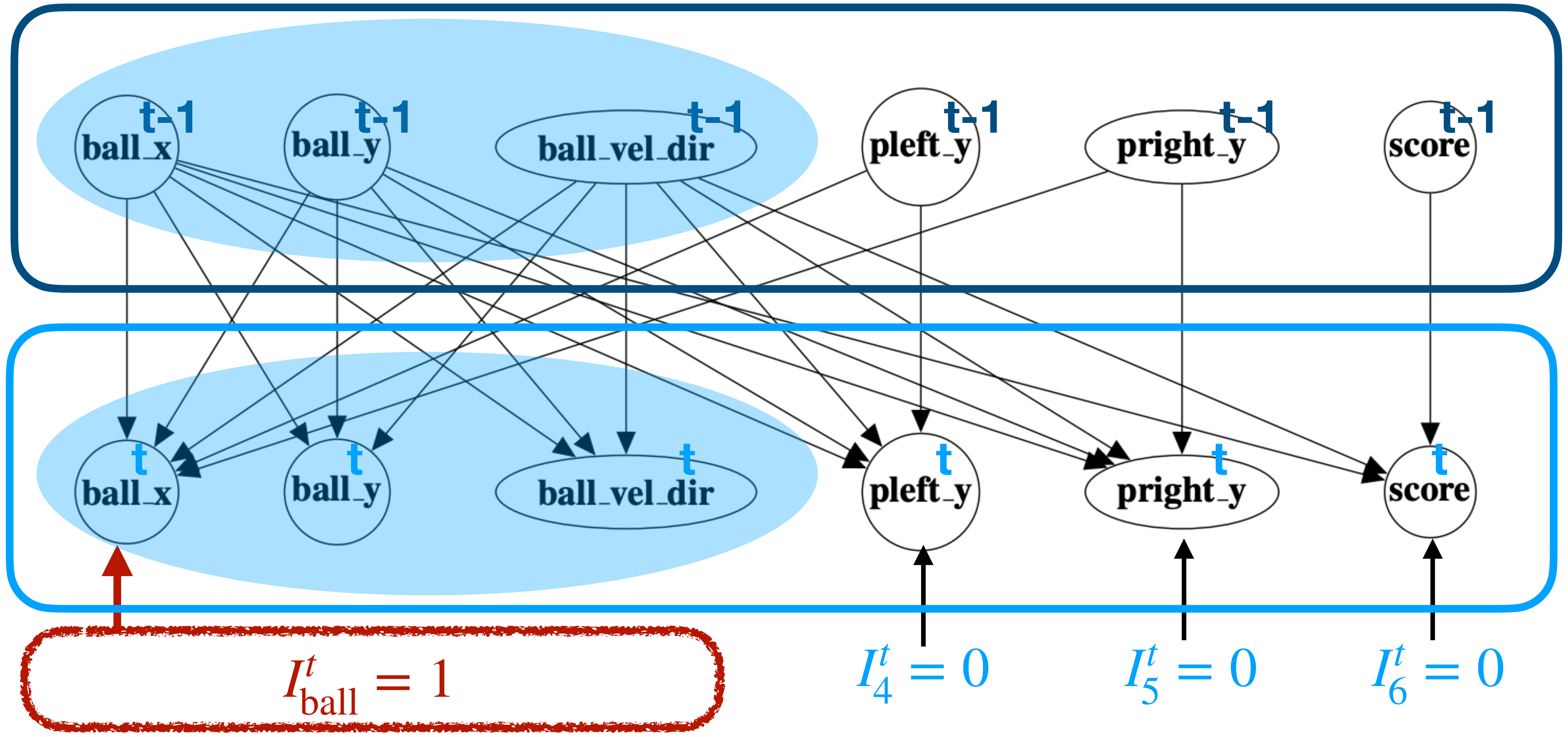




The intervention only has an effect on a part of the variable
(e.g.only on ball_x)

Minimal causal variables - example





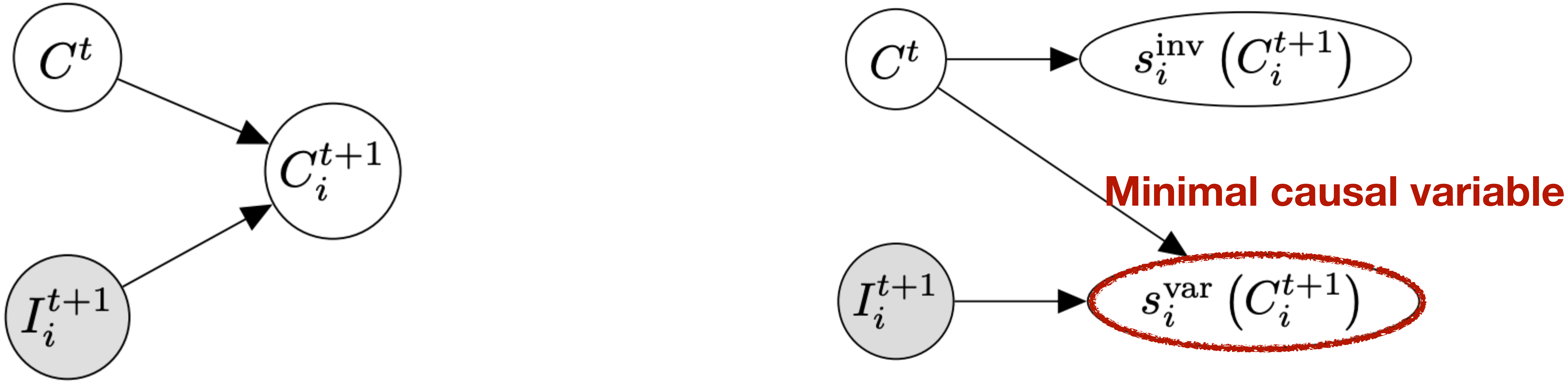
The intervention only has an effect on a part of the variable
(e.g.only on ball_x)

We can distinguish only x of the ball
(y and vel_dir are never intervened upon)

Minimal causal variables - theory

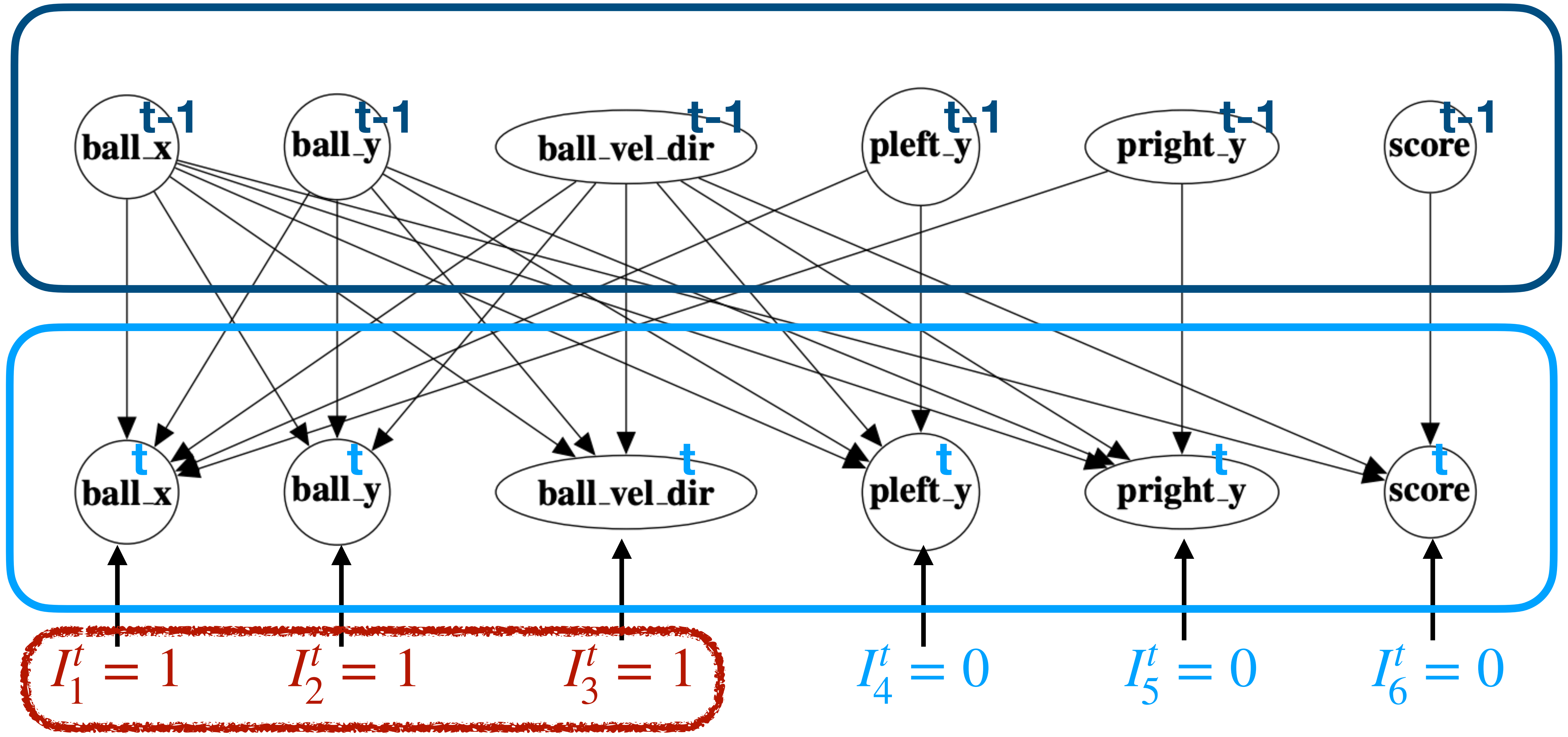
- We can define a split of a causal variable C_i^t

$$s_i(C_i^t) = (s_i^{\text{var}}(C_i^t), s_i^{\text{inv}}(C_i^t))$$



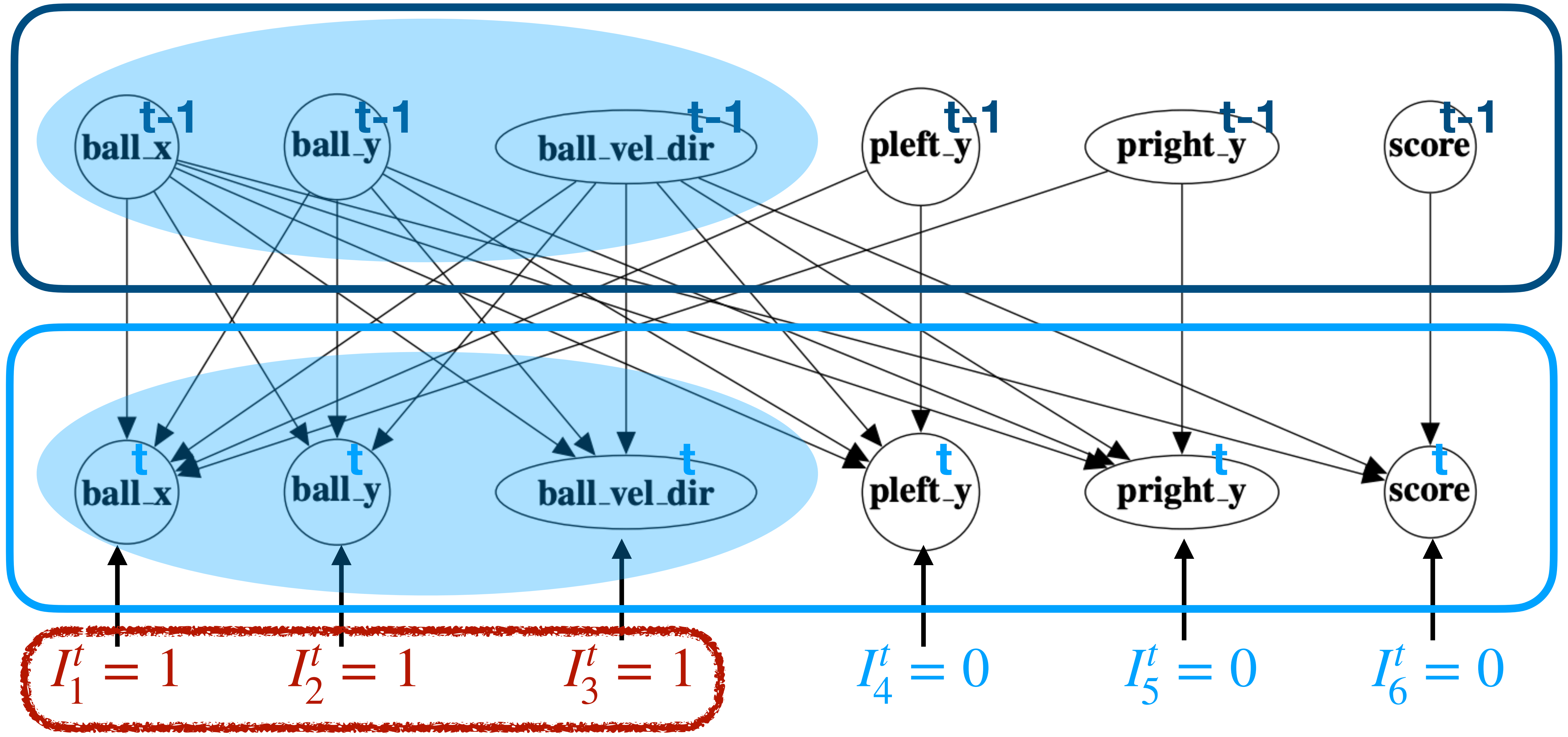
- We choose $s_i^{\text{var}*}$ that contains only the information that depends on I_i^{t+1}
- We can identify minimal causal variables up to invertible component-wise transformations, **if I_i^t is not a deterministic function of I_j^t**

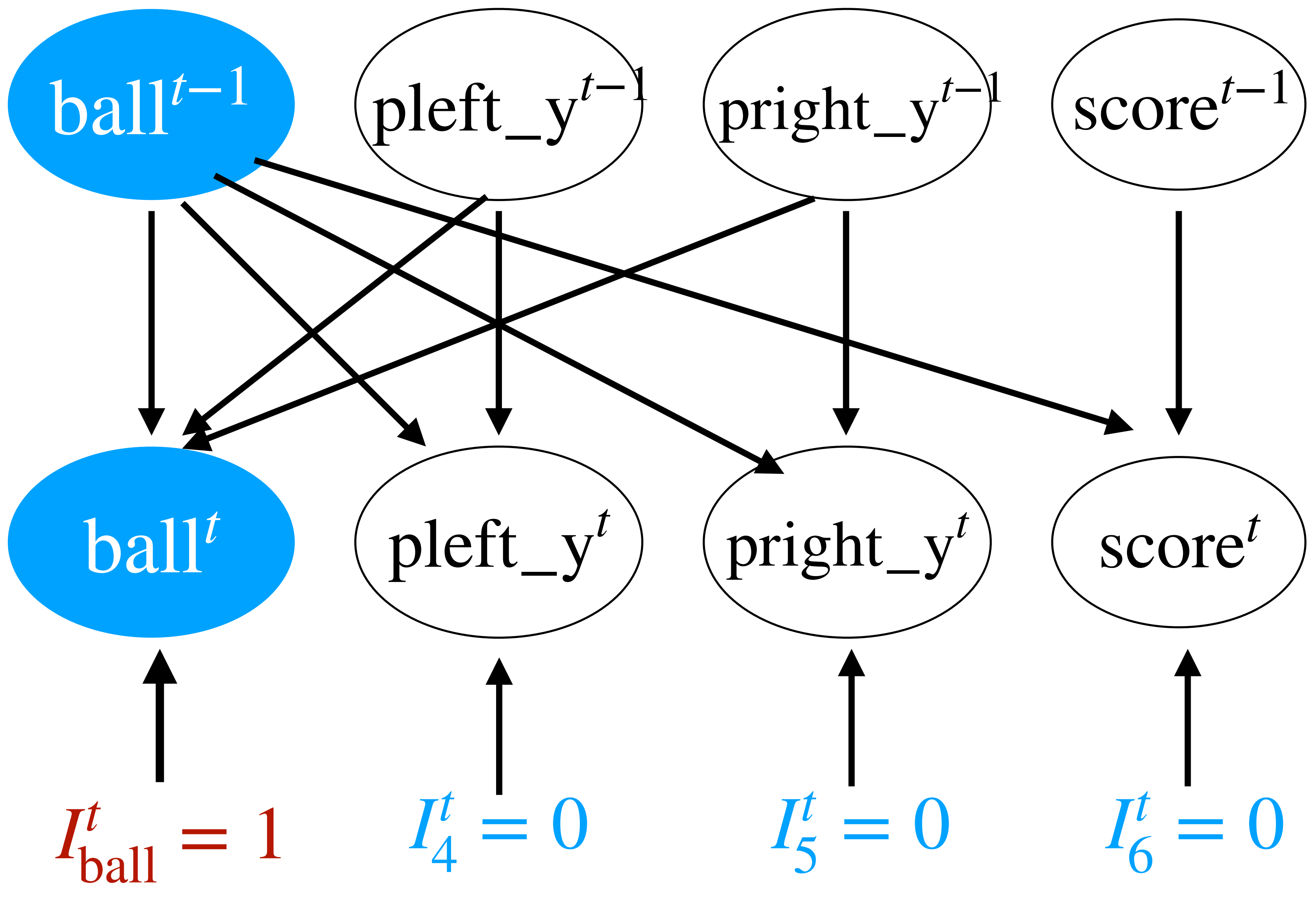
Deterministic relations between I_i



These are always intervened together
(or they are a deterministic function of each other)

Deterministic relations between I_i



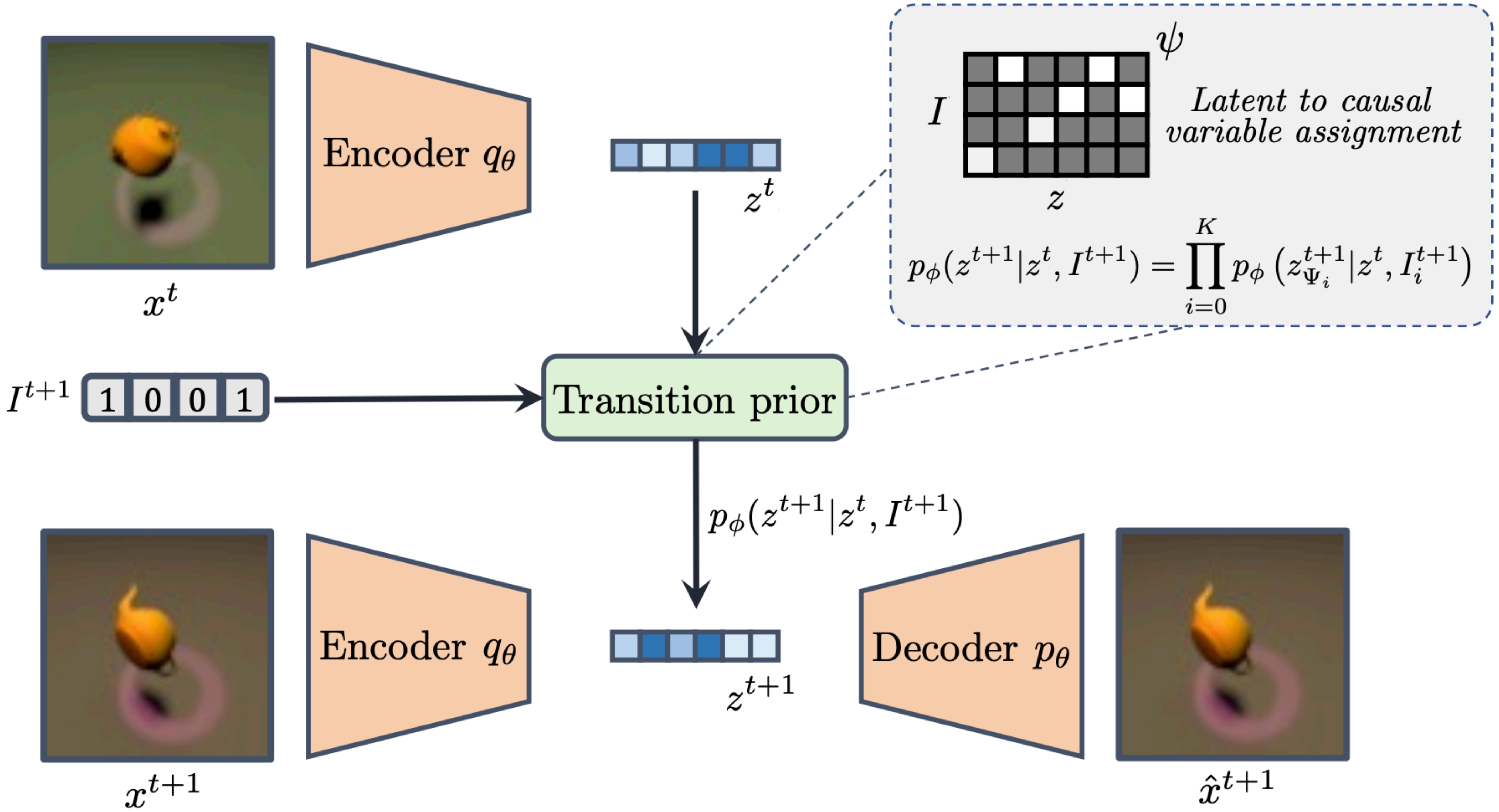


These are always intervened together (or they are a deterministic function of each other)

We cannot distinguish x, y and velocity direction of the ball

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves ICML 2022



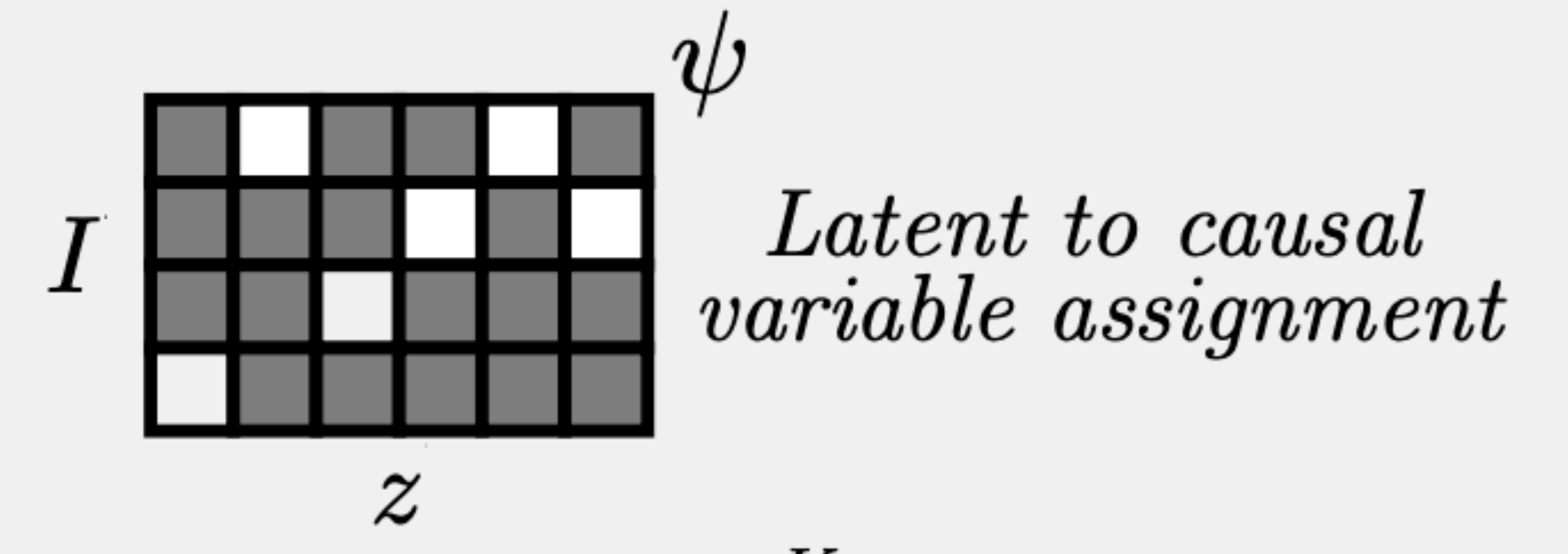
CITRIS-VAE

https://arxiv.org/abs/2202.03169

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves **ICML 2022**

- We have **multidimensional causal factors**, so we need to learn an assignment function ψ that matches each C_i with the assigned latents

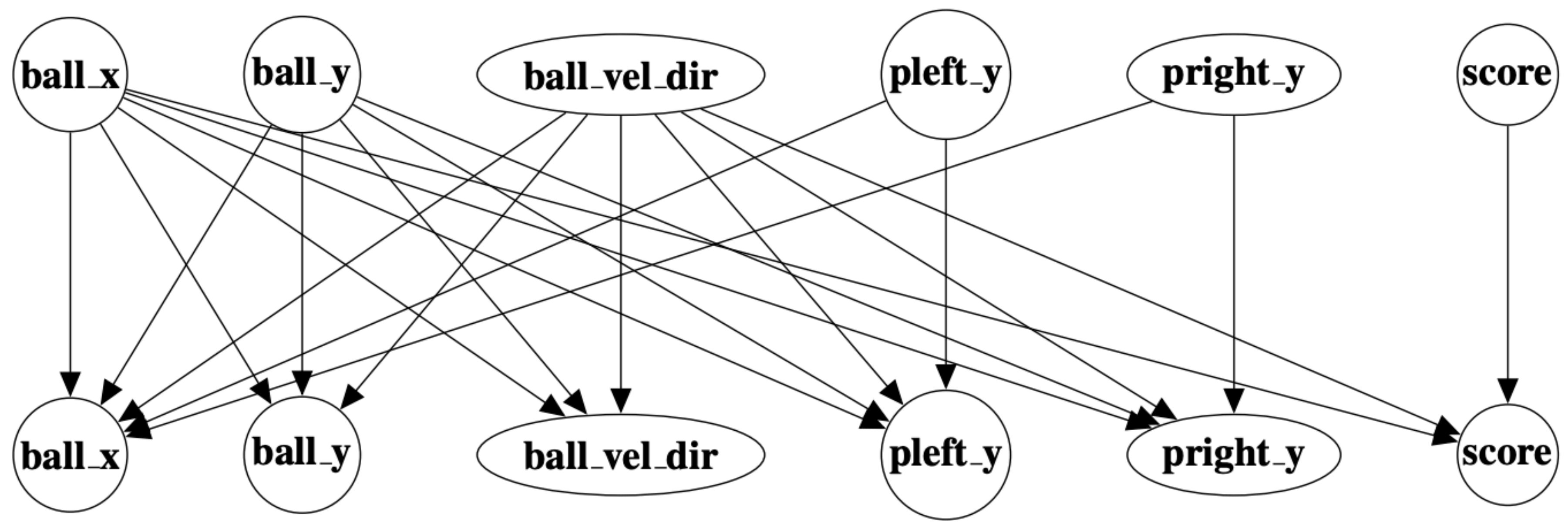


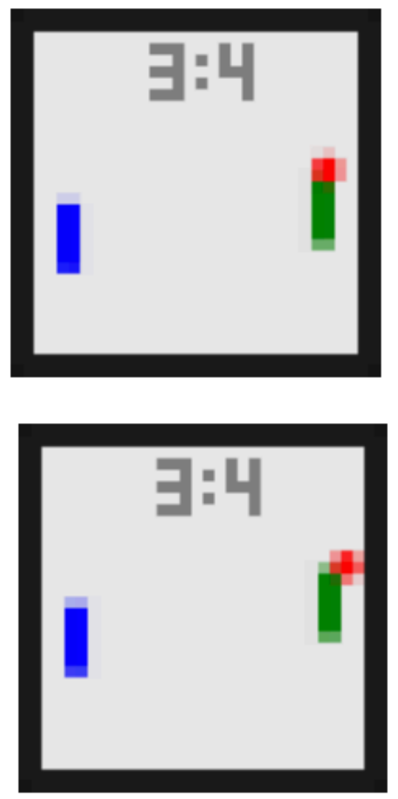
$$p_\phi(z^{t+1}|z^t, I^{t+1}) = \prod_{i=0}^{K} p_\phi\left(z^{t+1}_{\Psi_i}|z^t, I^{t+1}_i\right)$$

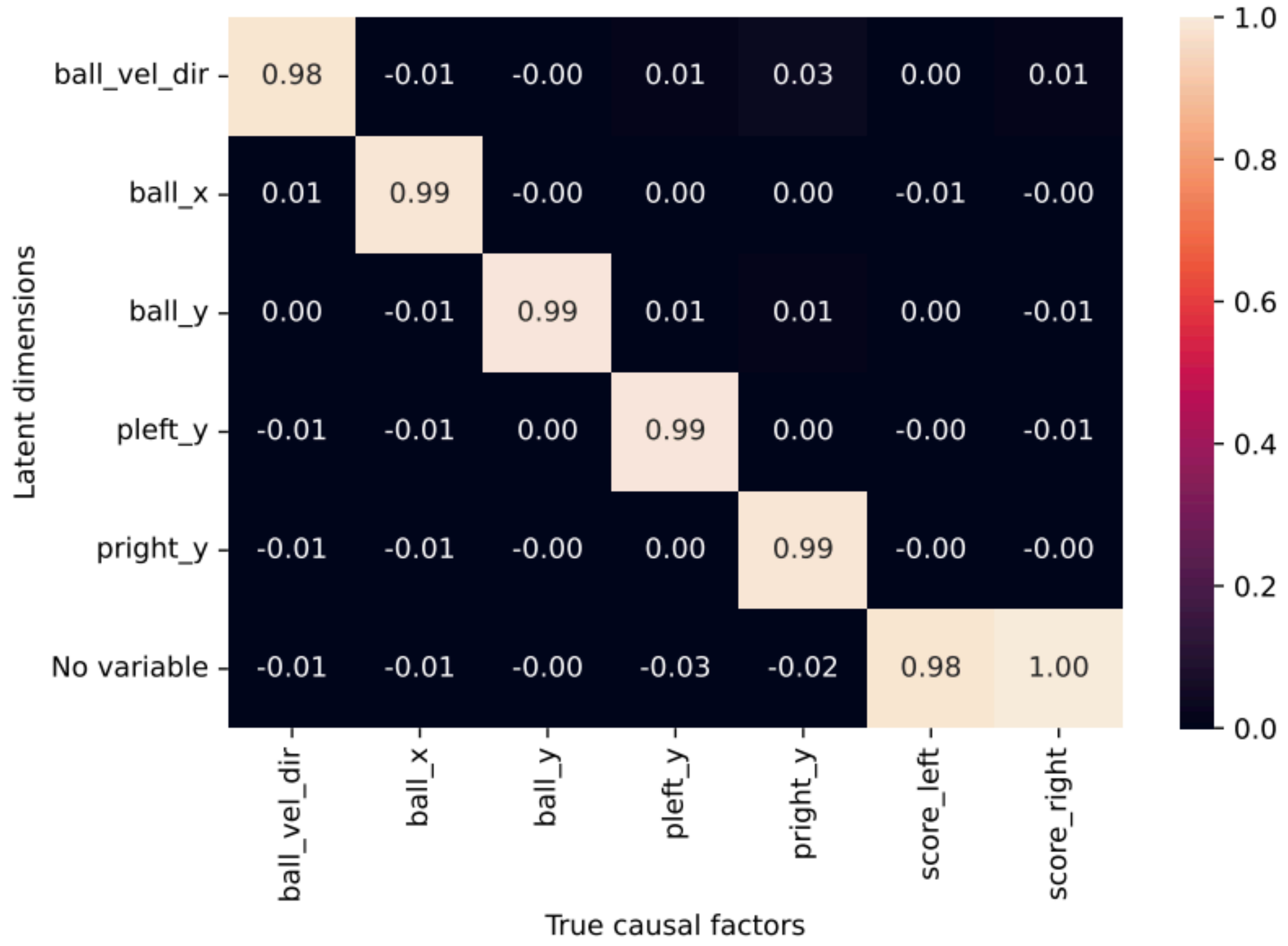
$$C_i \longrightarrow z_{\Psi_i}$$

z_{Ψ_0} “junk” variables

Experiments: Interventional Pong

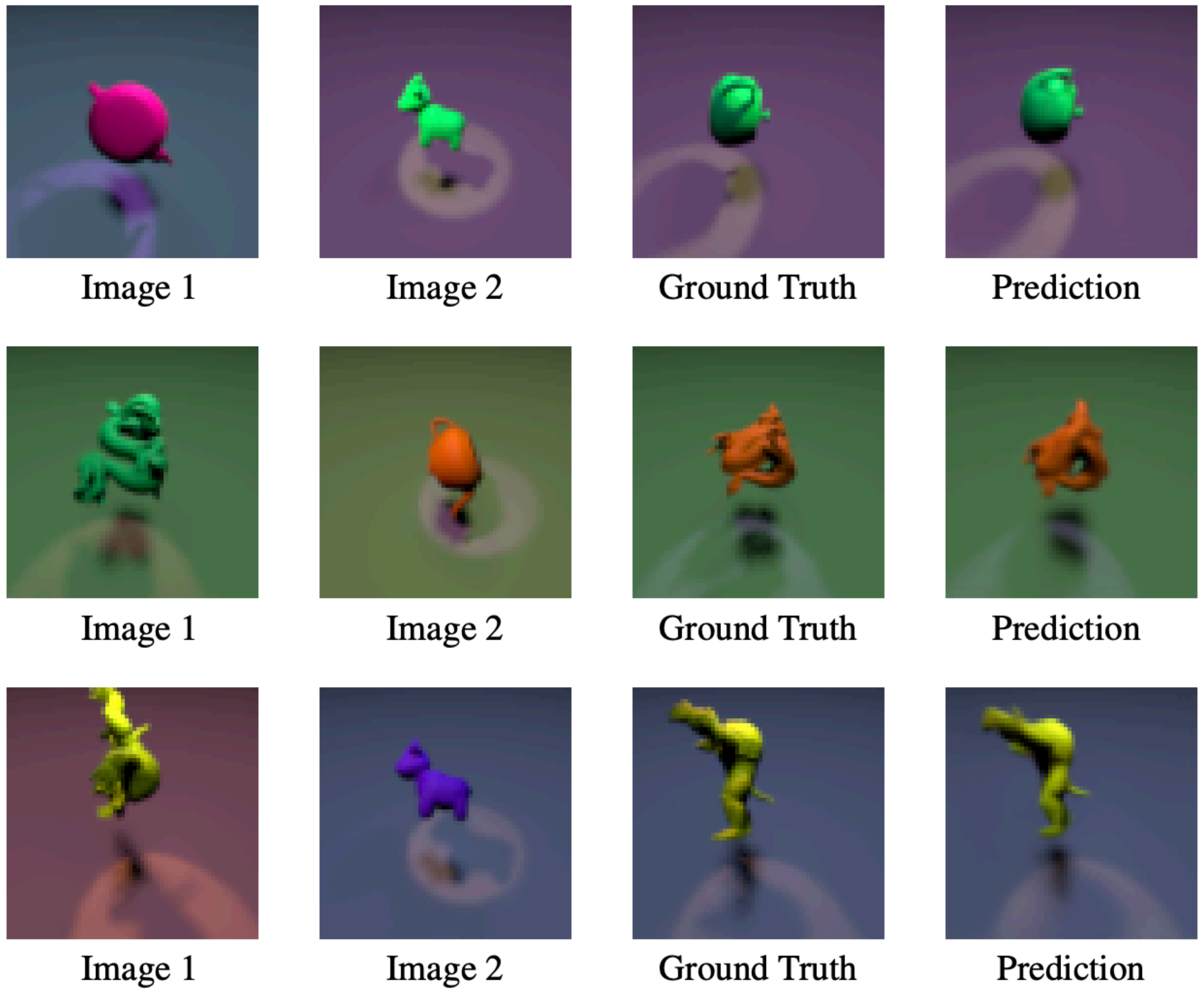


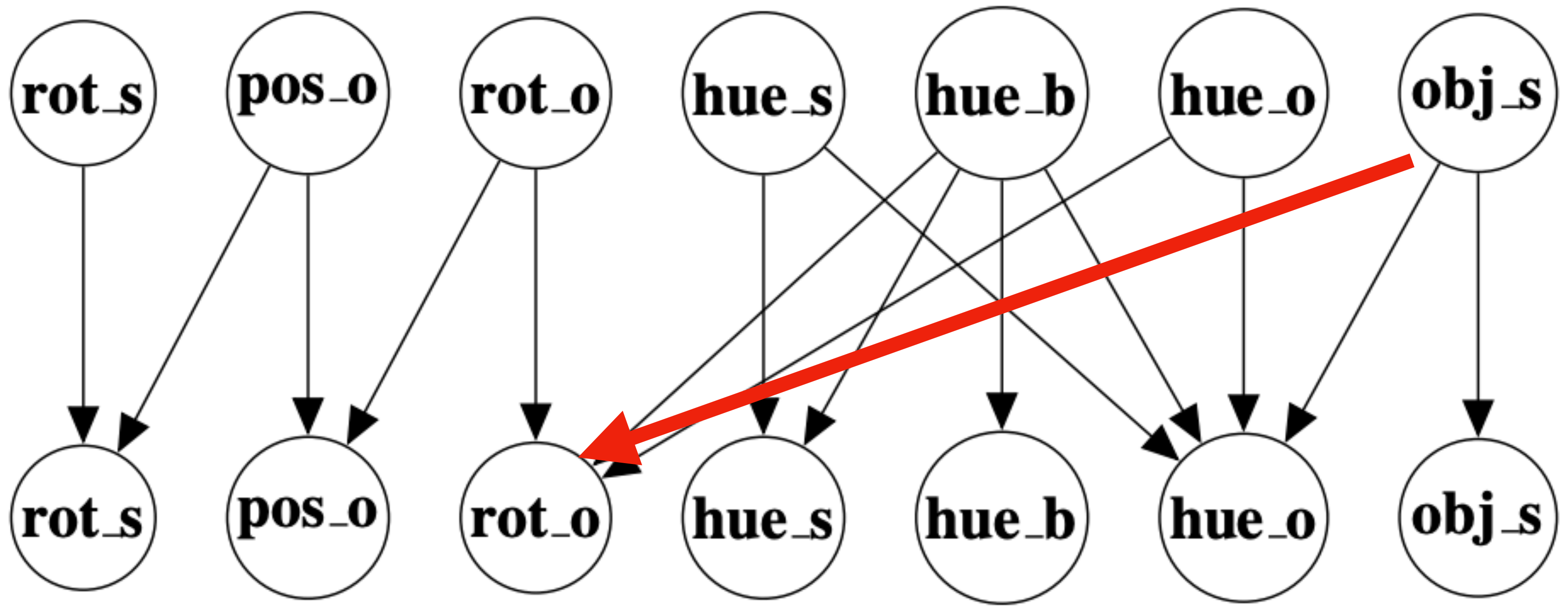




Experiments: Temporal Causal3DIdent

Temporal Causal3D Ident



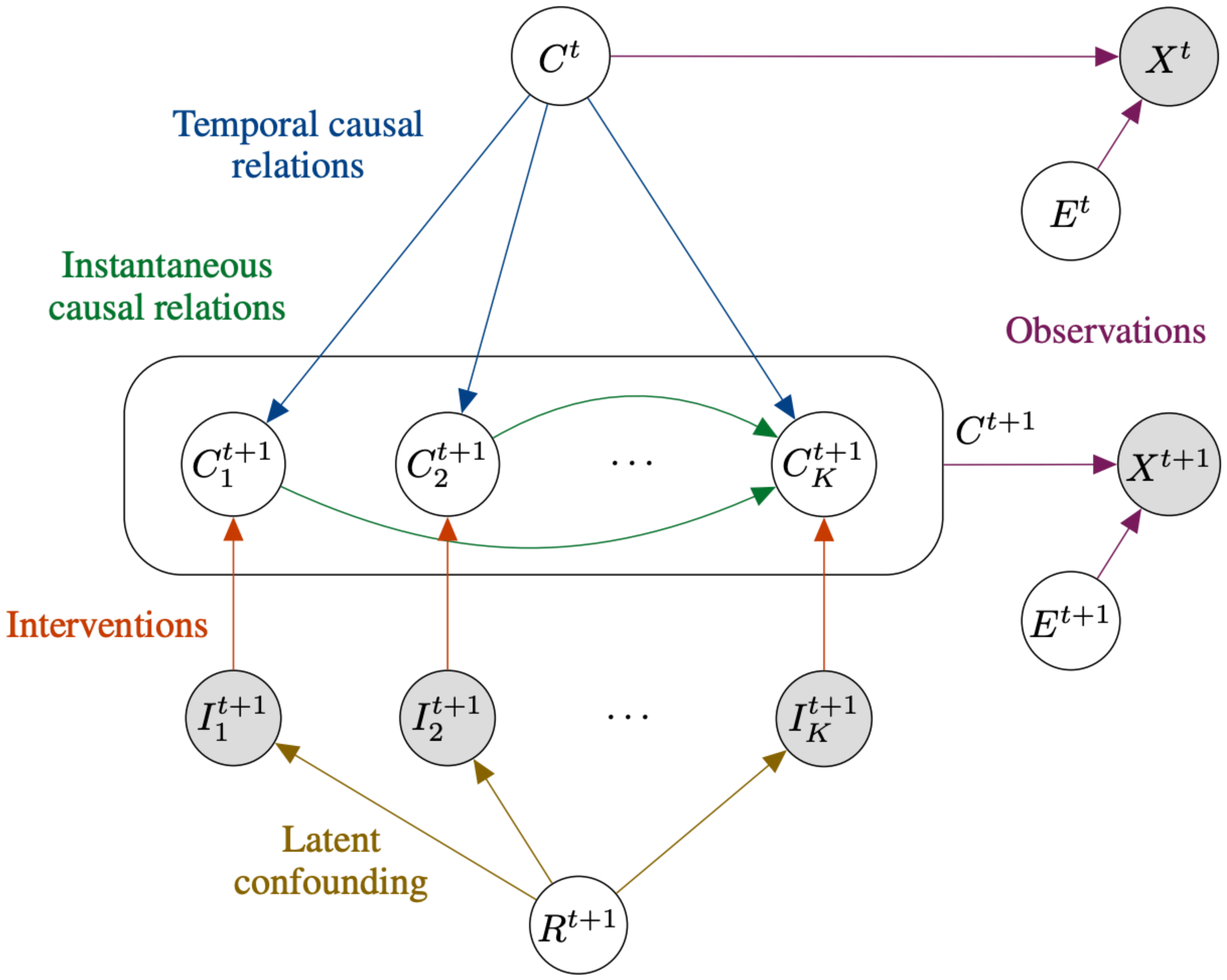


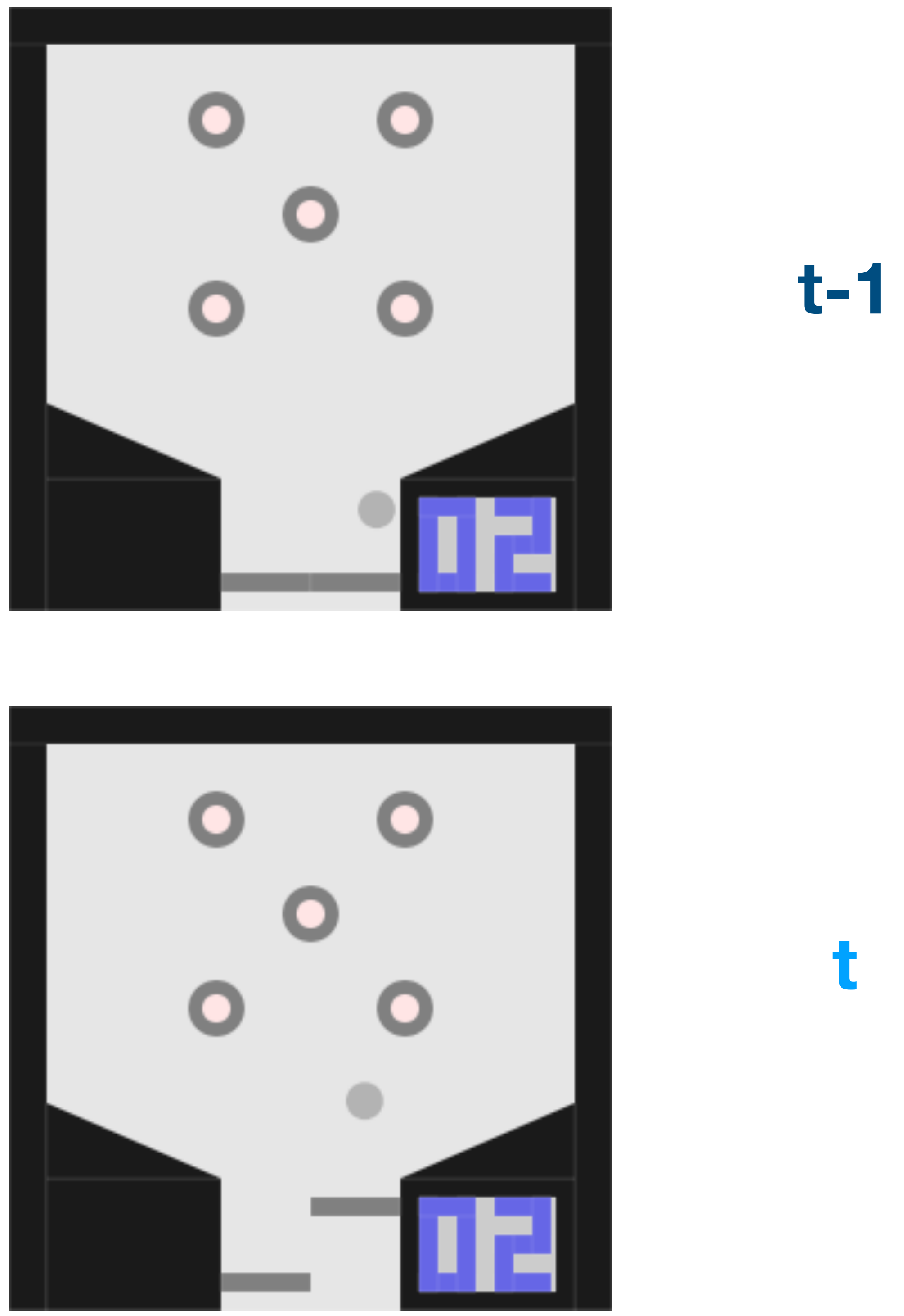
Causal graph learnt with CITRIS-NF

iCITRIS: Causal Representation Learning for Instantaneous Temporal Effects

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves

ICLR 2023





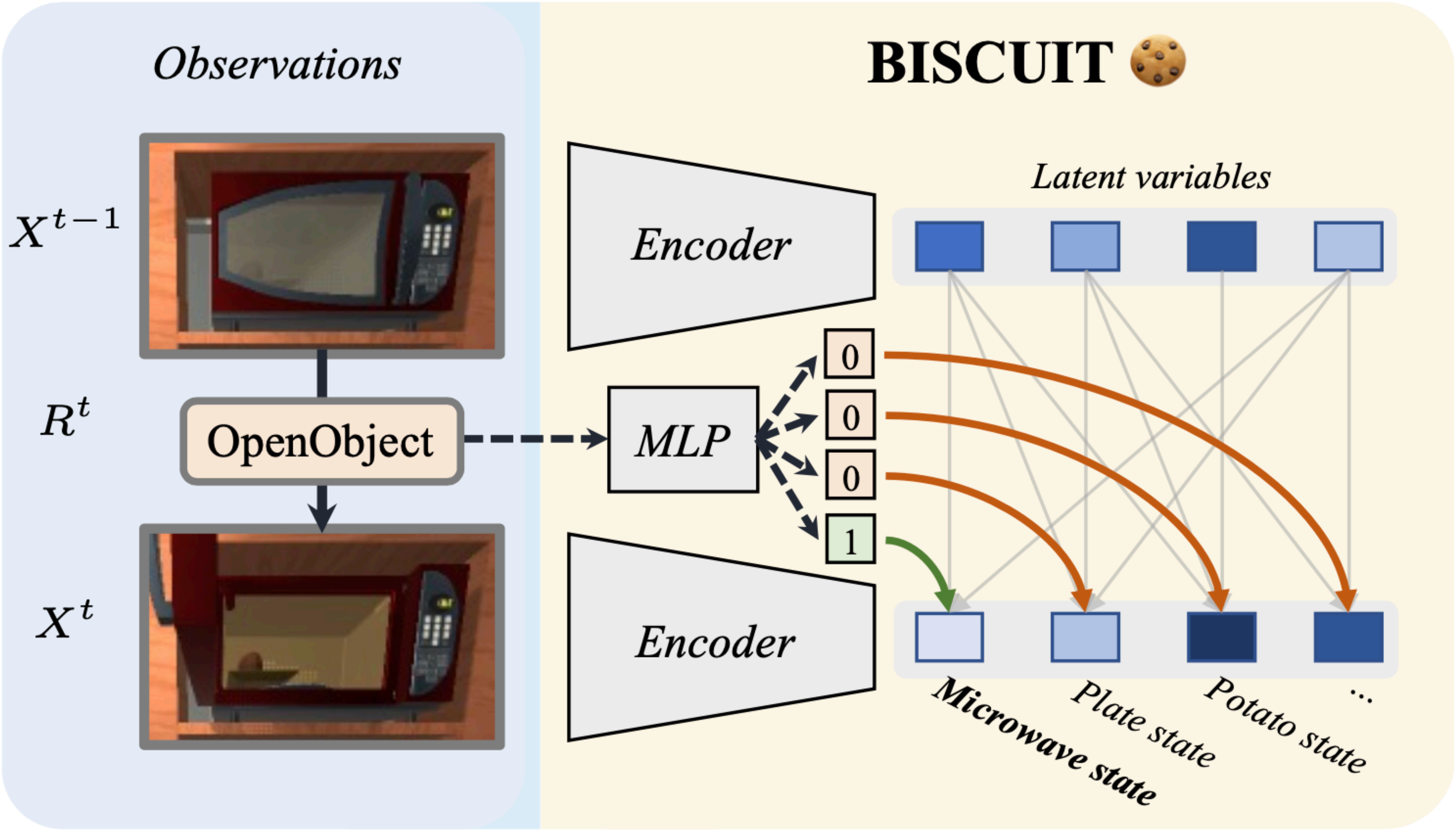
https://arxiv.org/abs/2206.06169

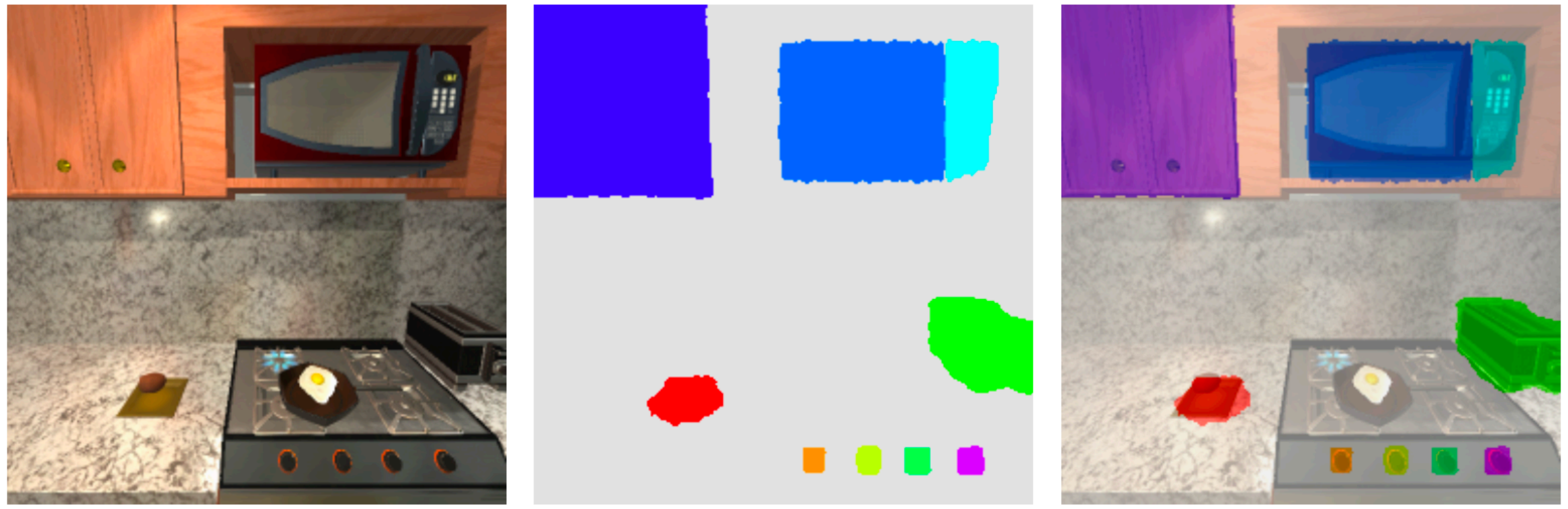
BISCUIT: Causal Representation Learning from Binary Interactions

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves

UAI 2023

COMING SOON





Input Image

Learned Interactions

Combined Image

Input Image 1

Input Image 2

Generated Output

Causal Hierarchy [Pearl 2009, 2018]

Most ML

Causality

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Causal Hierarchy [Pearl 2009, 2018]

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection		

Most ML

Causality

E.g. need many experiments or strong assumptions to identify the causal graph or the causal variables

"Full" causality can be **not necessary** or **too expensive** ->

Causal Hierarchy [Pearl 2009, 2018]

Most ML

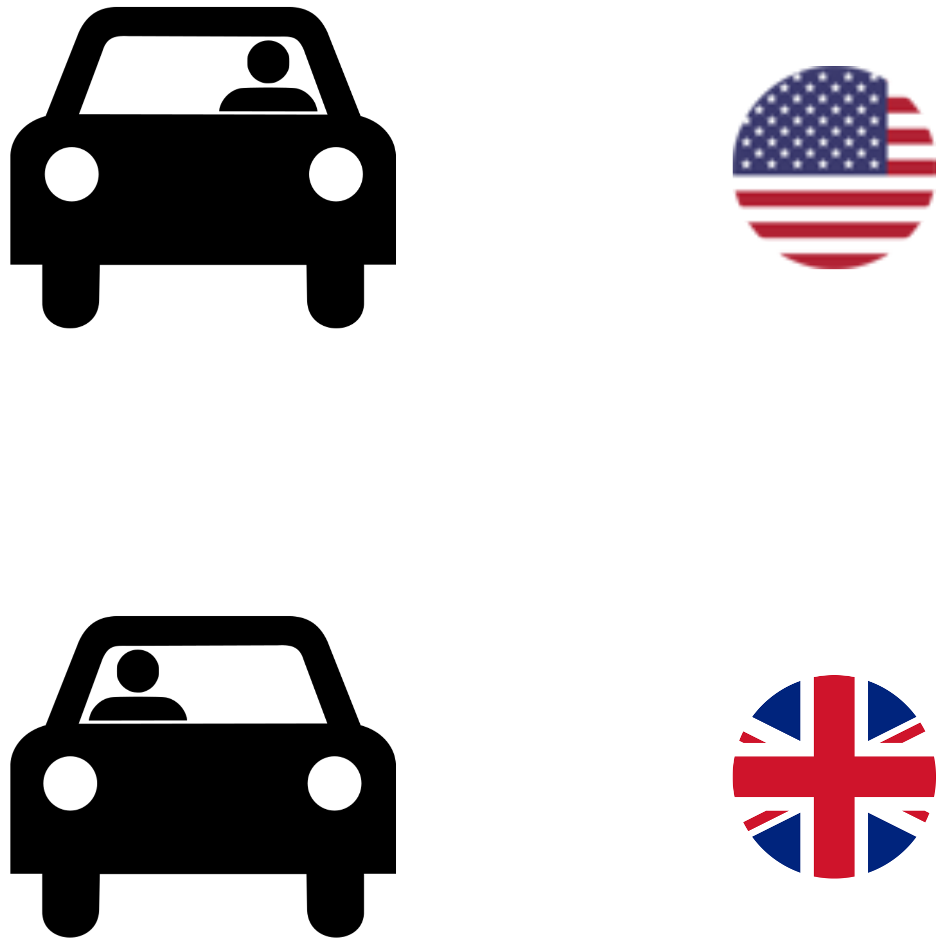
Causality

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

"Full" causality can be **not necessary** or **too expensive** -> *Causality-Inspired*

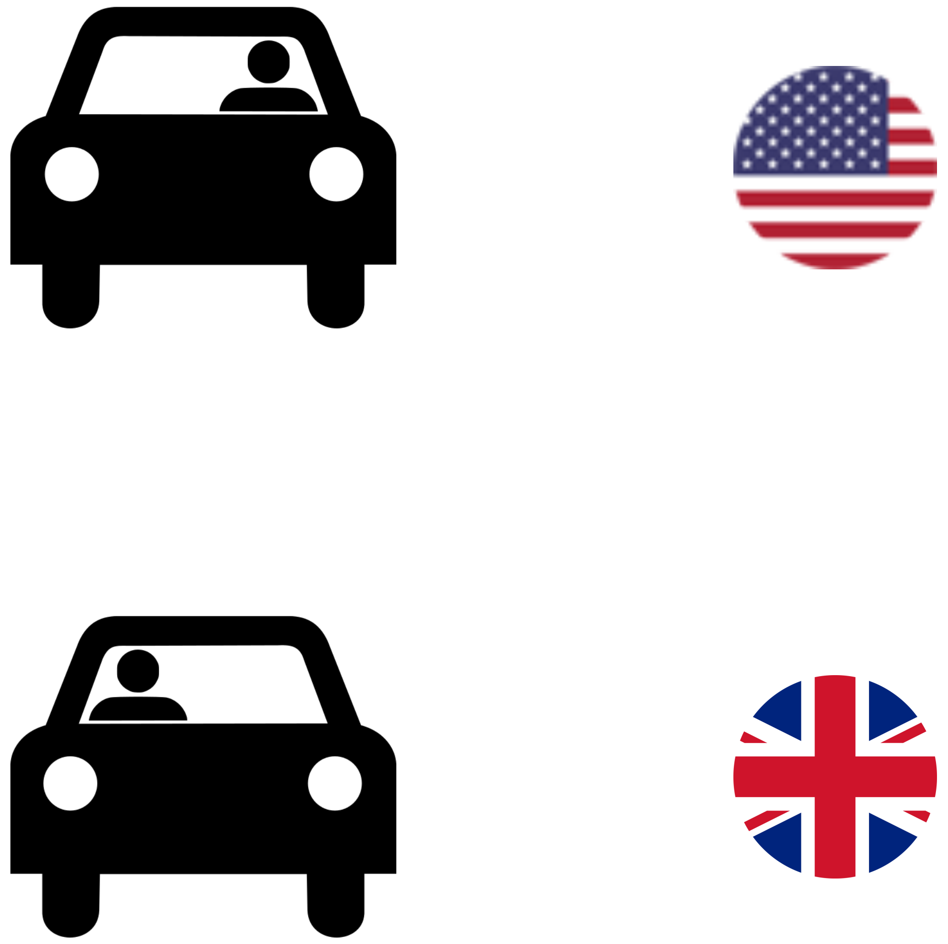
Causality vs Transfer learning

- Transfer learning:
    - How can I predict what happens when the distribution changes?



Causality vs Transfer learning

- Transfer learning:
  - How can I predict what happens when the distribution changes?



- Causal inference:
  - How can I predict what happens when the distribution changes after an intervention?
    - Perfect intervention do(X):
      - do-calculus [Pearl, 2009]
    - **Soft intervention on X** ≈ change of distribution of P(X| parents)

Causality vs Transfer learning

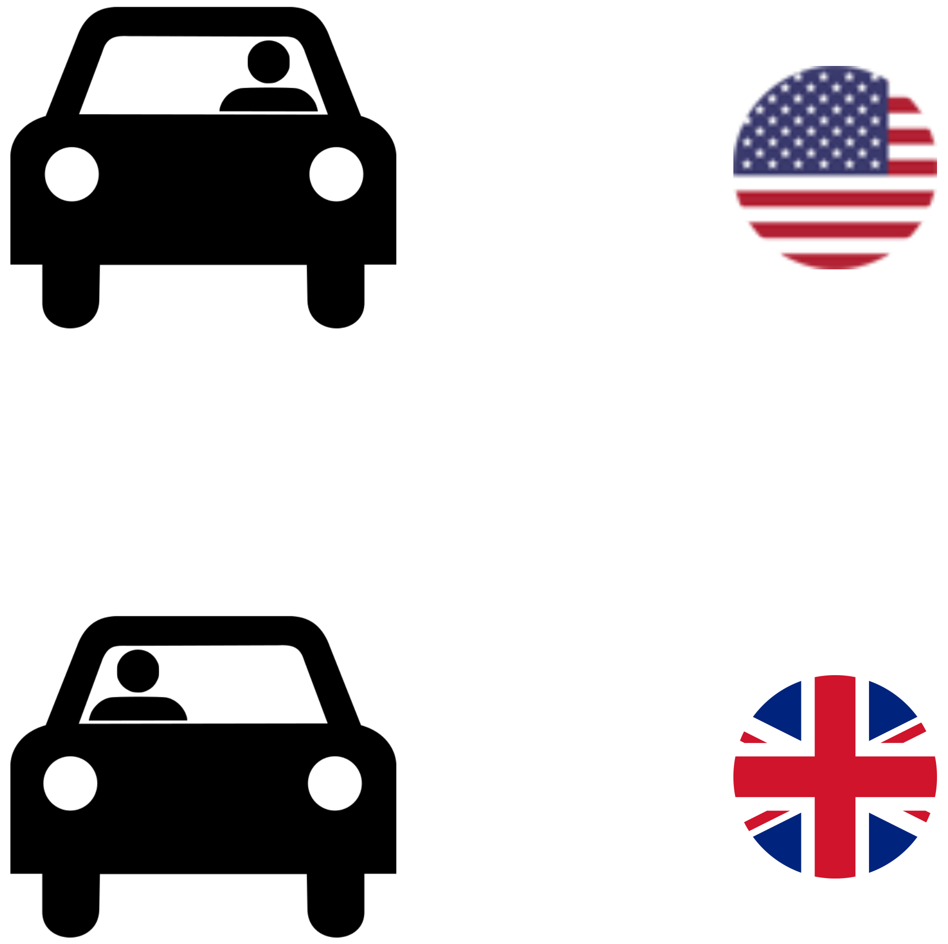
- Transfer learning:
  - How can I predict what hap… when the distribution chang…

…ens when … an

…t intervention do(X):

- do-calculus [Pearl, 2009]

Very general - can model also changes in distribution that are not from "real" interventions



- **Soft intervention on X** ≈ change of distribution of P(X| parents)

Causality allows us to reason systematically about distribution shifts, e.g. through graphs

On Causal and Anticausal Learning

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang FIRST.LAST@TUE.MPG.DE
Max Planck Institute for Intelligent Systems, Spemannstrasse, 72076 Tübingen, Germany
Joris Mooij J.MOOIJ@CS.RU.NL
Institute for Computing and Information Sciences, Radboud University, Nijmegen, The Netherlands

Domain Adaptation as a Problem of Inference on Graphical Models

Kun Zhang[1*], Mingming Gong[2*], Petar Stojanov[3]
Biwei Huang[1], Qingsong Liu[4], Clark Glymour[1]
[1] Department of philosophy, Carnegie Mellon University
[2] School of Mathematics and Statistics, University of Melbourne
[3] Computer Science Department, Carnegie Mellon University, [4] Unisound AI Lab
kunz1@cmu.edu, mingming.gong@unimelb.edu.au, liuqingsong@unisound.com
{pstojano, biweih, cg09}@andrew.cmu.edu

Anchor regression: heterogeneous data meet causality

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann and Jonas Peters

Invariant Risk Minimization

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, David Lopez-Paz

J. R. Statist. Soc. B (2016)
78, *Part* 5, *pp.* 947–1012

Causal inference by using invariant prediction: identification and confidence intervals

Jonas Peters
Max Planck Institute for Intelligent Systems, Tübingen, Germany, and Eidgenössische Technische Hochschule Zürich, Switzerland
and Peter Bühlmann and Nicolai Meinshausen
Eidgenössische Technische Hochschule Zürich, Switzerland

Invariant Models for Causal Transfer Learning

Mateo Rojas-Carulla MR597@CAM.AC.UK
Max Planck Institute for Intelligent Systems Tübingen, Germany
Department of Engineering Univ. of Cambridge, United Kingdom
Bernhard Schölkopf BS@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems Tübingen, Germany
Richard Turner RET26@CAM.AC.UK
Department of Engineering Univ. of Cambridge, United Kingdom
Jonas Peters* JONAS.PETERS@MATH.KU.DK
Department of Mathematical Sciences Univ. of Copenhagen, Denmark

Invariance, Causality and Robustness

2018 Neyman Lecture *

Peter Bühlmann †
Seminar for Statistics, ETH Zürich

Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests

Victor Veitch[1,2], Alexander D'Amour[1], Steve Yadlowsky[1], and Jacob Eisenstein[1]
[1]*Google Research*
[2]*University of Chicago*

Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions

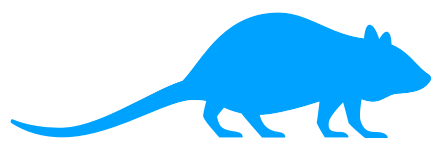
Sara Magliacane, IBM Research*, sara.magliacane@gmail.com
Thijs van Ommen, University of Amsterdam, thijsvanommen@gmail.com
Tom Claassen, Radboud University Nijmegen, tomc@cs.ru.nl
Stephan Bongers, University of Amsterdam, srbongers@gmail.com
Philip Versteeg, University of Amsterdam, p.j.j.p.versteeg@uva.nl
Joris M. Mooij, University of Amsterdam, j.m.mooij@uva.nl

A Causal View on Robustness of Neural Networks

Cheng Zhang *, Microsoft Research, Cheng.Zhang@microsoft.com
Kun Zhang, Carnegie Mellon University, kunz1@cmu.edu
Yingzhen Li *, Microsoft Research, Yingzhen.Li@microsoft.com

and many many more....

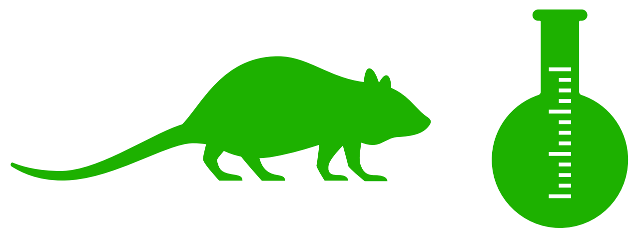
Domain adaptation from the graphical perspective



	X1	X2	Y
Wildtype	0,1	2	0
Wildtype	0,2	3	0
Wildtype	1,1	2	1
Wildtype	0,1	3	0

Source domain

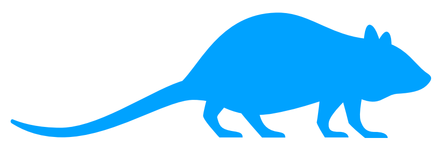
No labels in target



	X1	X2	Y
Gene A	3,1	2	?
Gene A	3,2	3	?
Gene A	4	2	?
Gene A	3,2	3	?

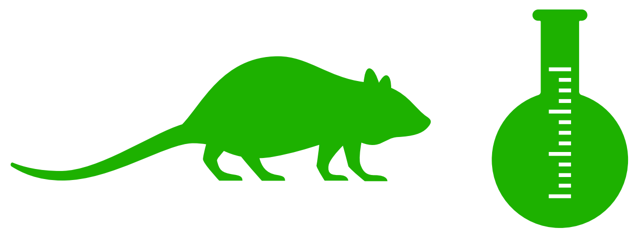
Target domain

Domain adaptation from the graphical perspective



D	X1	X2	Y
0	0,1	2	0
0	0,2	3	0
0	1,1	2	1
0	0,1	3	0

Source domain

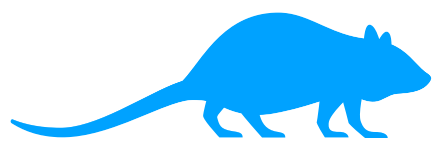


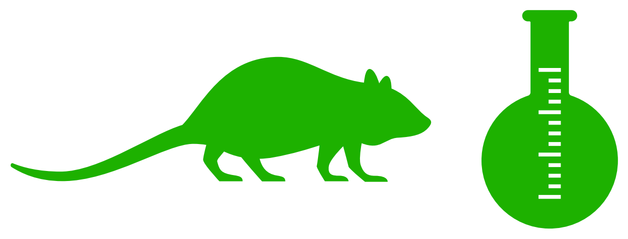
D	X1	X2	Y
1	3,1	2	?
1	3,2	3	?
1	4	2	?
1	3,2	3	?

Target domain

1. We add a domain variable D to distinguish the domains

Domain adaptation from the graphical perspective





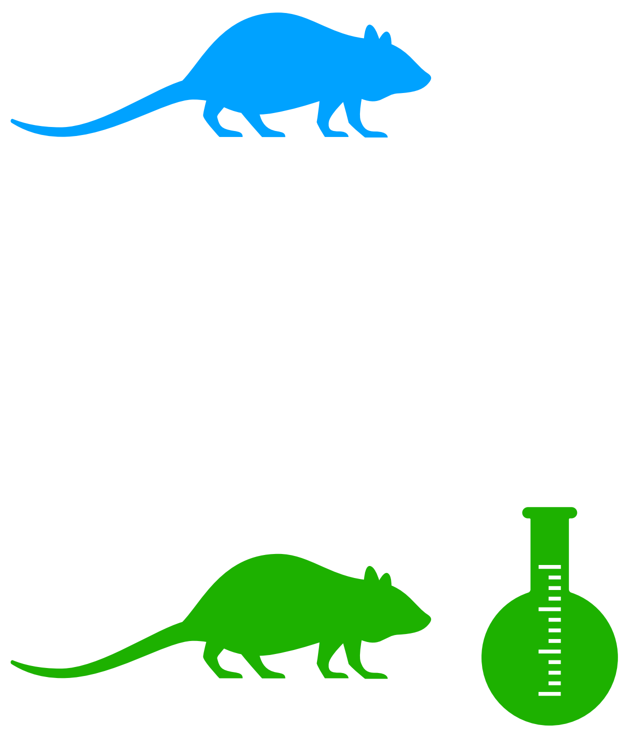
D	X1	X2	Y
0	0,1	2	0
0	0,2	3	0
0	1,1	2	1
0	0,1	3	0
1	3,1	2	?
1	3,2	3	?
1	4	2	?
1	3,2	3	?

Source domain

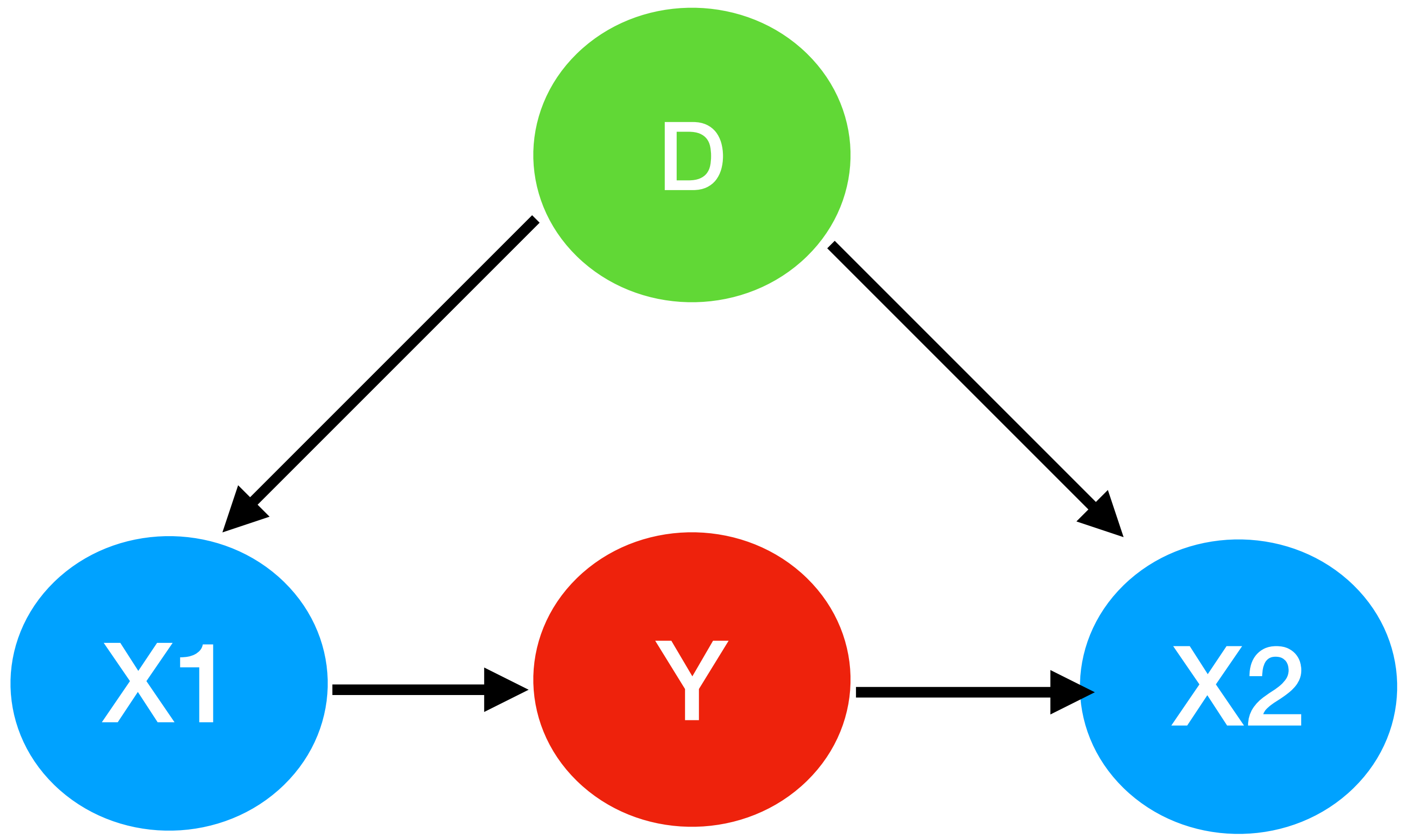
Target domain

1. We add a domain variable D to distinguish the domains
2. We consider the data as coming from a single distribution P(X1, X2,Y, D)

Domain adaptation from the graphical perspective



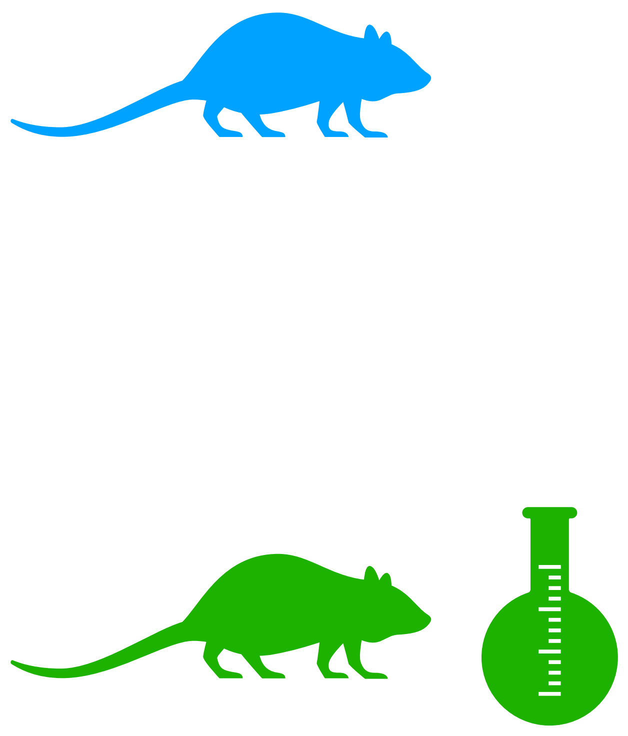
D	X1	X2	Y
0	0,1	2	0
0	0,2	3	0
0	1,1	2	1
0	0,1	3	0
1	3,1	2	?
1	3,2	3	?
1	4	2	?
1	3,2	3	?



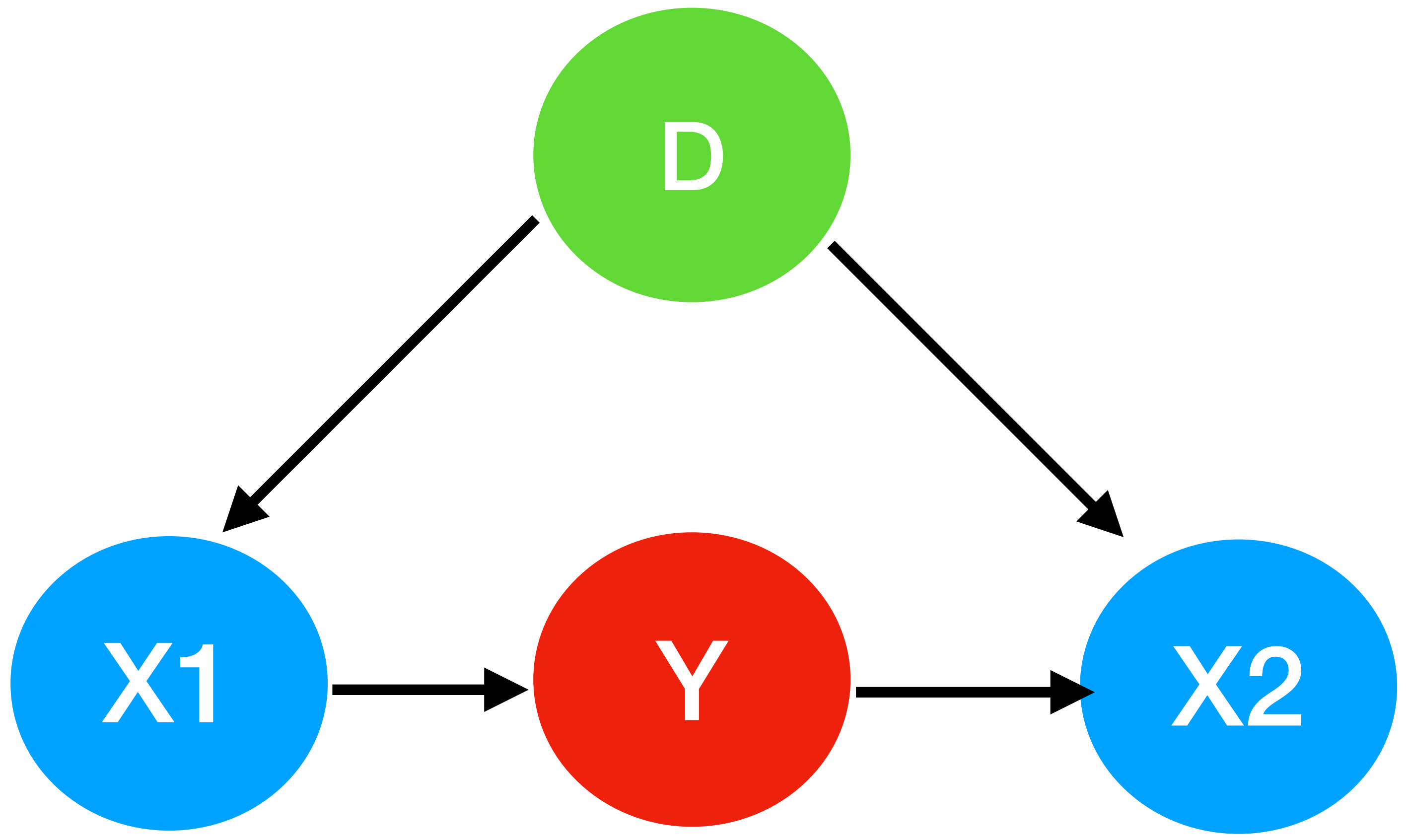
We can represent P(X1, X2,Y, D) with an **(unknown)** causal graph

1. We add a domain variable D to distinguish the domains
2. We consider the data as coming from a single distribution P(X1, X2,Y, D)

Domain adaptation from the graphical perspective



D	X1	X2	Y
0	0,1	2	0
0	0,2	3	0
0	1,1	2	1
0	0,1	3	0
1	3,1	2	?
1	3,2	3	?
1	4	2	?
1	3,2	3	?

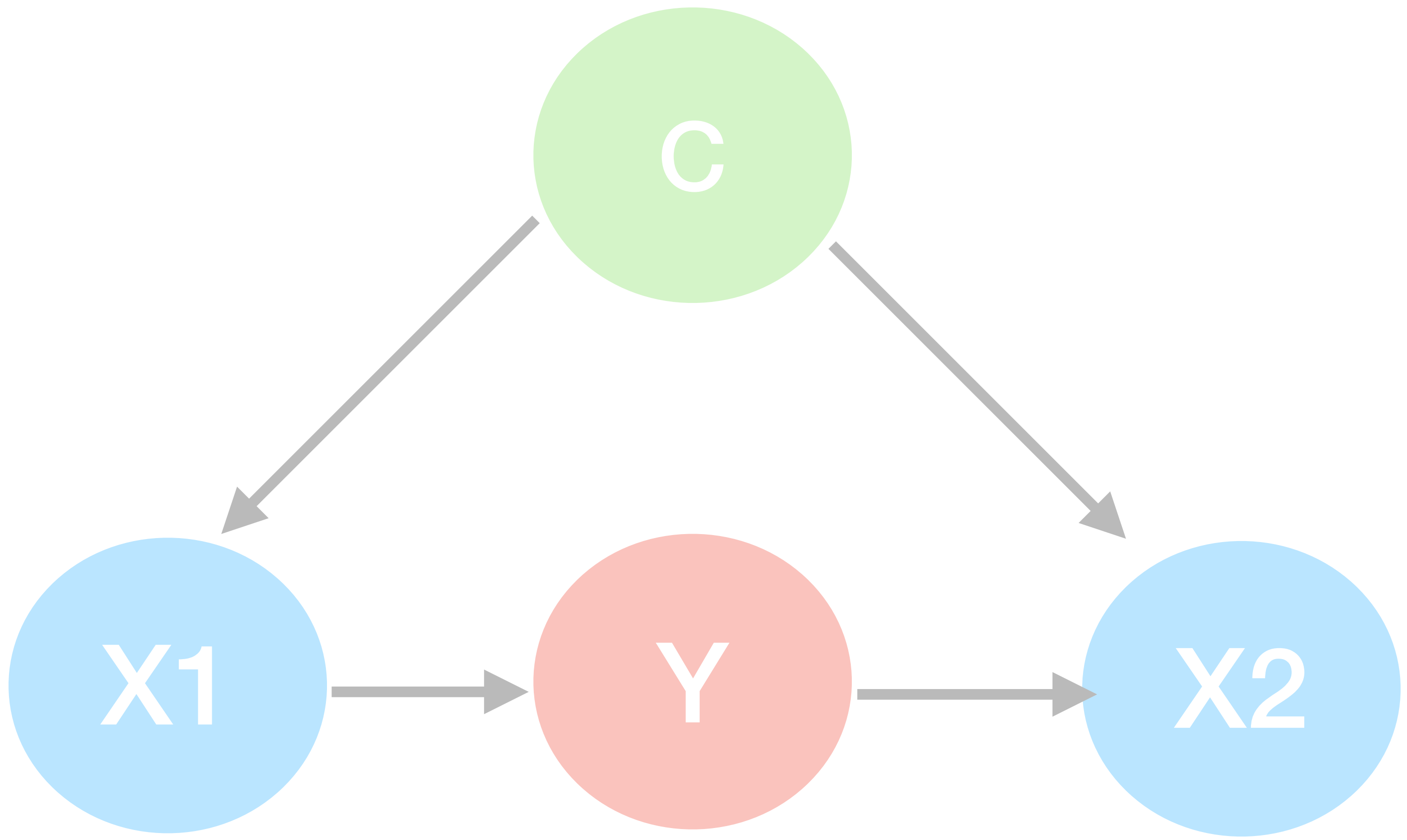


We can represent P(X1, X2,Y, D) with an **(unknown)** causal graph

- **Task**: find a subset of features X that predict Y robustly in the target domain
  - **Separating features $\mathbf{S} \subseteq \mathbf{X} : Y \perp_d D \,|\, \mathbf{S}$ … d-separation [Pearl 2009]**

Separating features = safe for (adversarial) domain adaptation

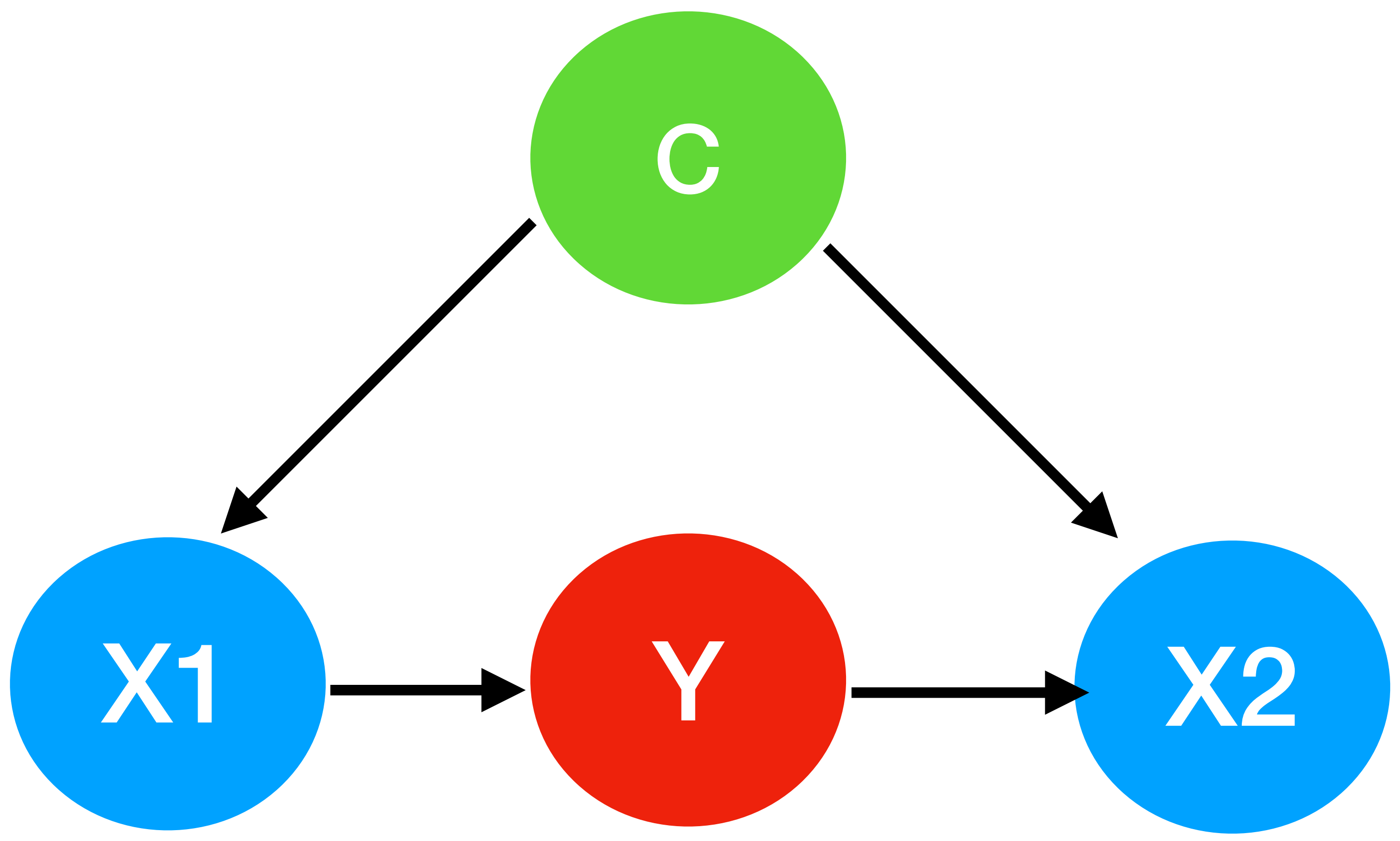
- **Separating features:** sets of features that d-separate Y from the context



Step 1: Known causal graph

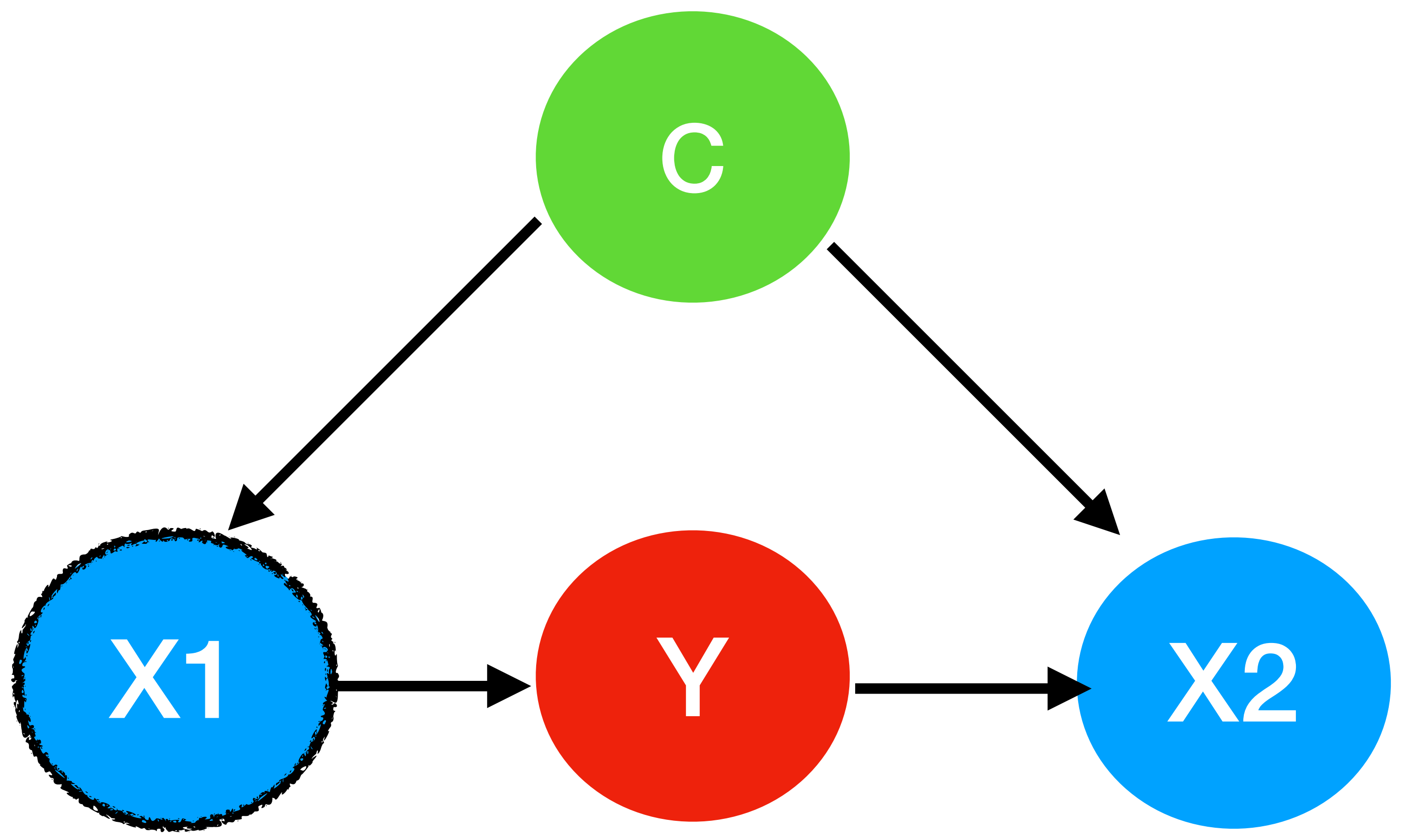
Separating features = safe for (adversarial) domain adaptation

- **Separating features:** sets of features that d-separate Y from the context



Separating features = safe for (adversarial) domain adaptation

- **Separating features:** sets of features that d-separate Y from the context

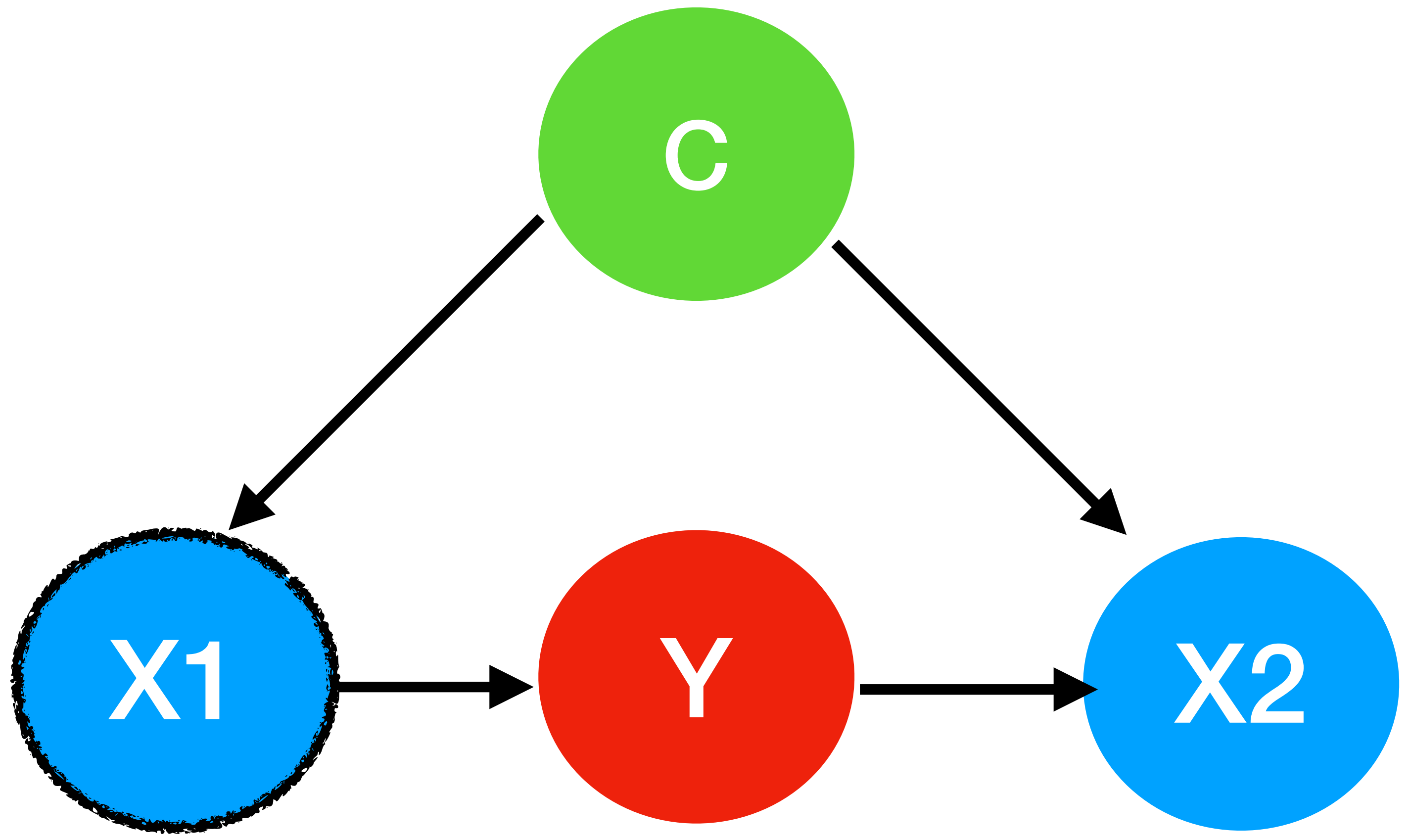


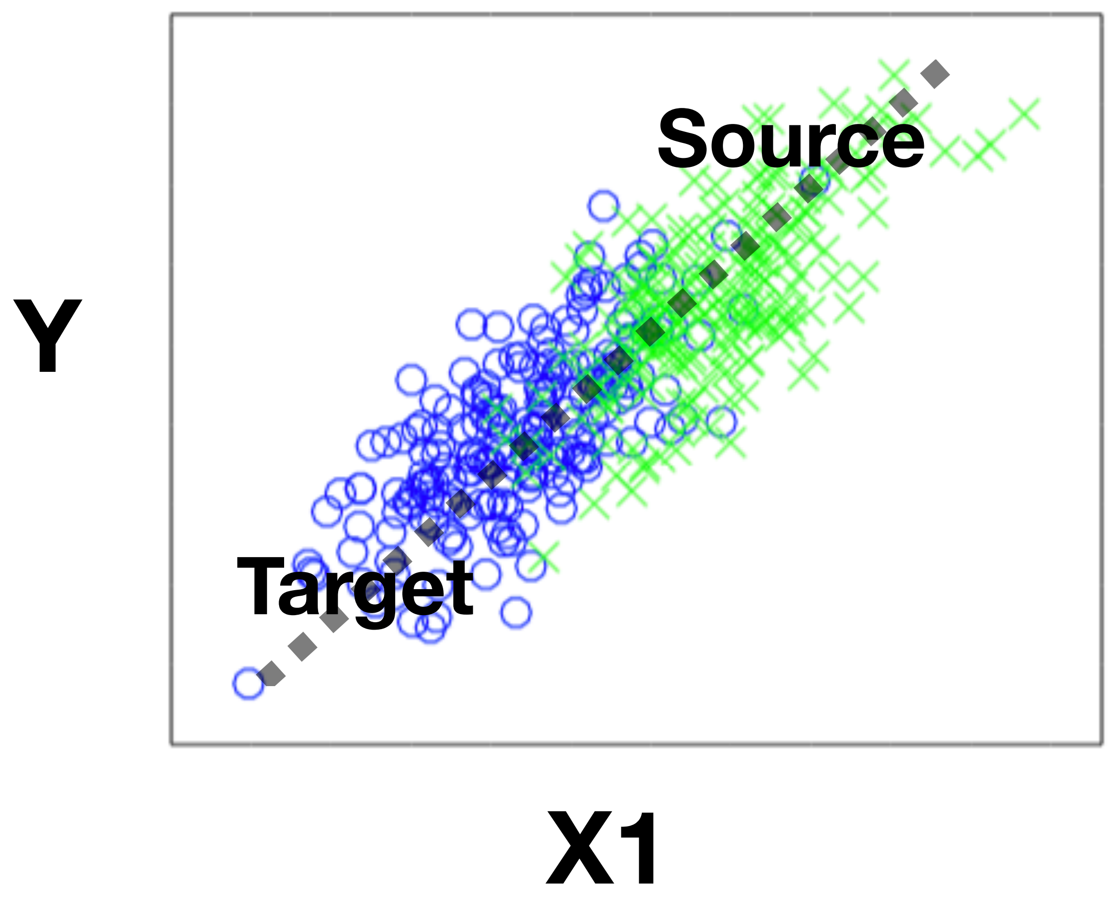
$$Y \perp\!\!\!\perp_d C \mid X_1 \iff Y \perp\!\!\!\perp C \mid X_1$$

(under Markov and faithfulness assumptions)

Separating features = safe for (adversarial) domain adaptation

- **Separating features:** sets of features that d-separate Y from the context





$$Y \perp\!\!\!\perp_d C \mid X_1 \iff Y \perp\!\!\!\perp C \mid X_1$$

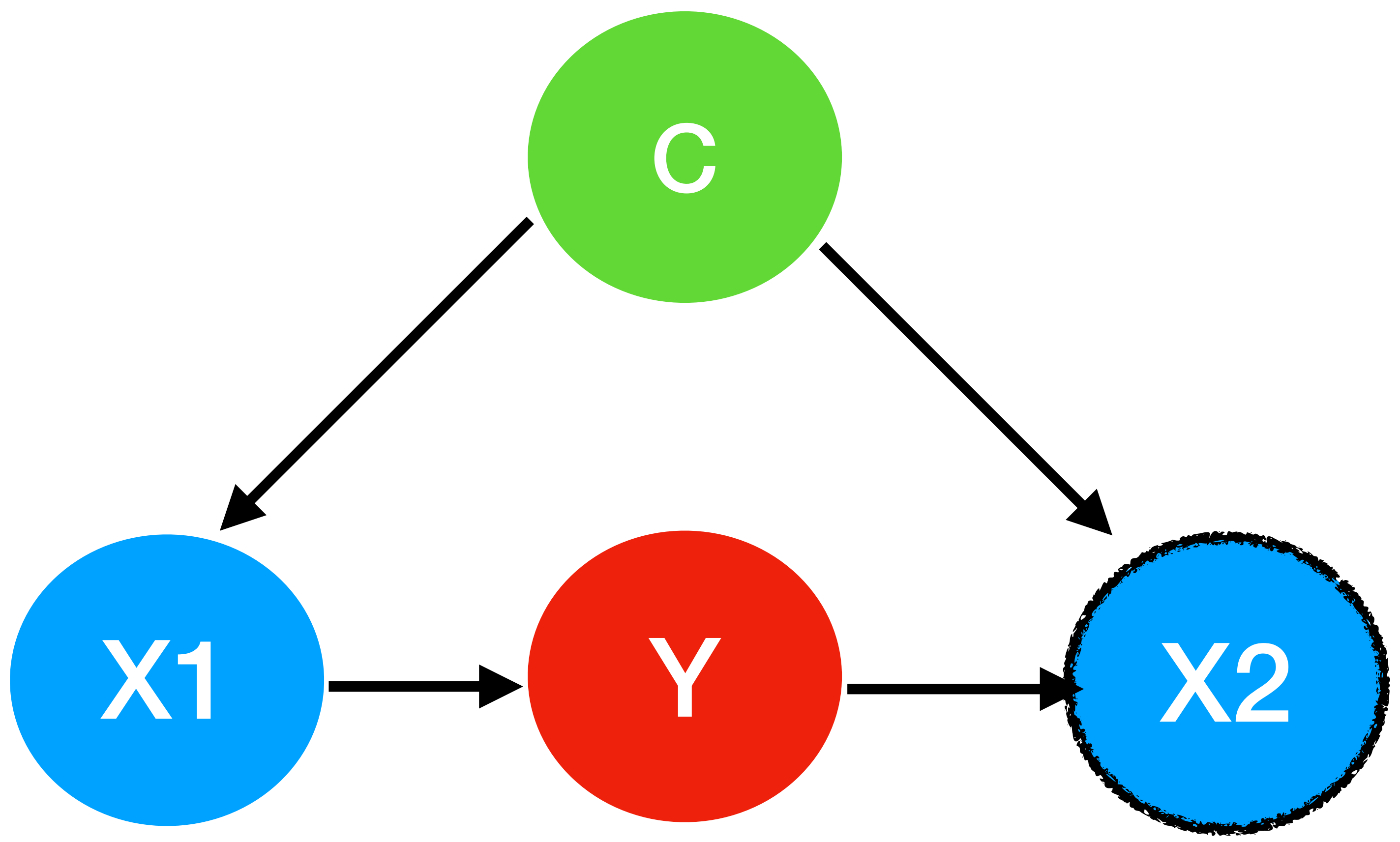
(under Markov and faithfulness assumptions)

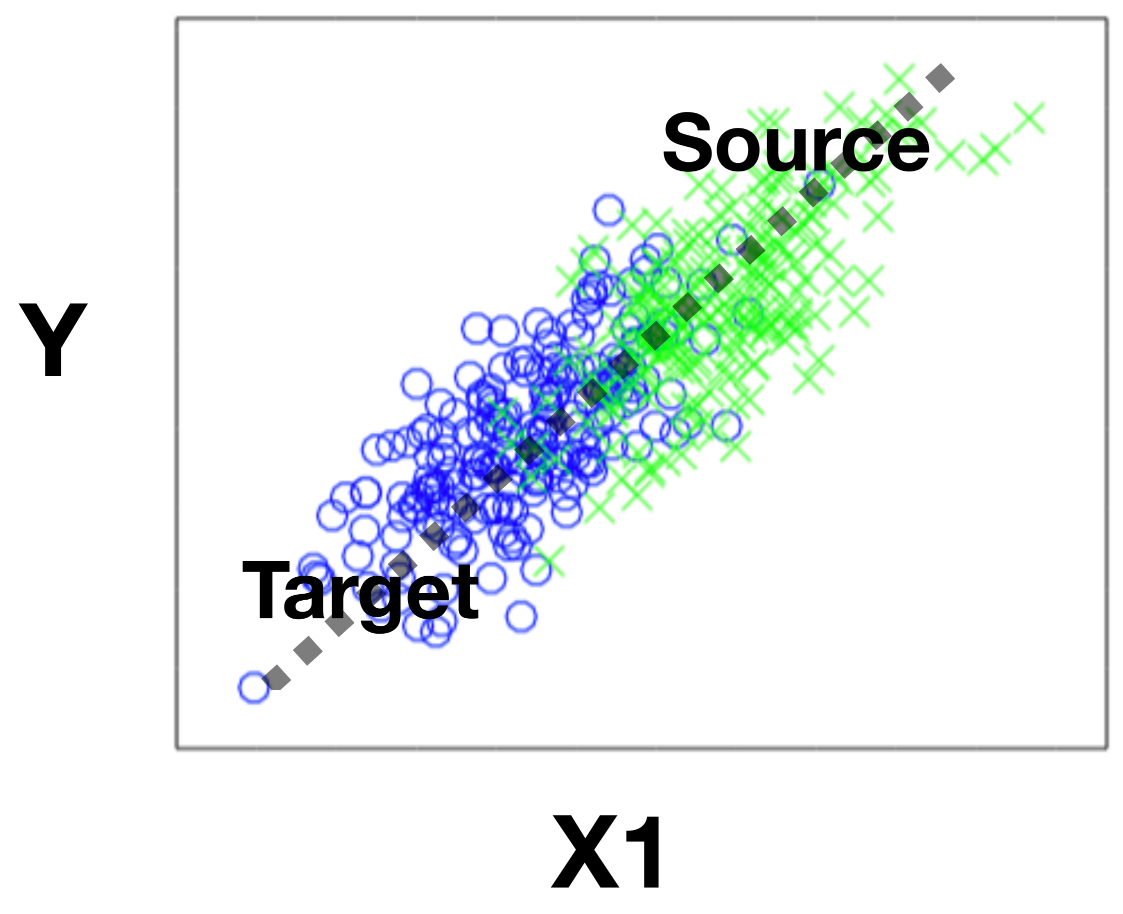
$$Y \perp\!\!\!\perp C \mid X_1 \equiv$$

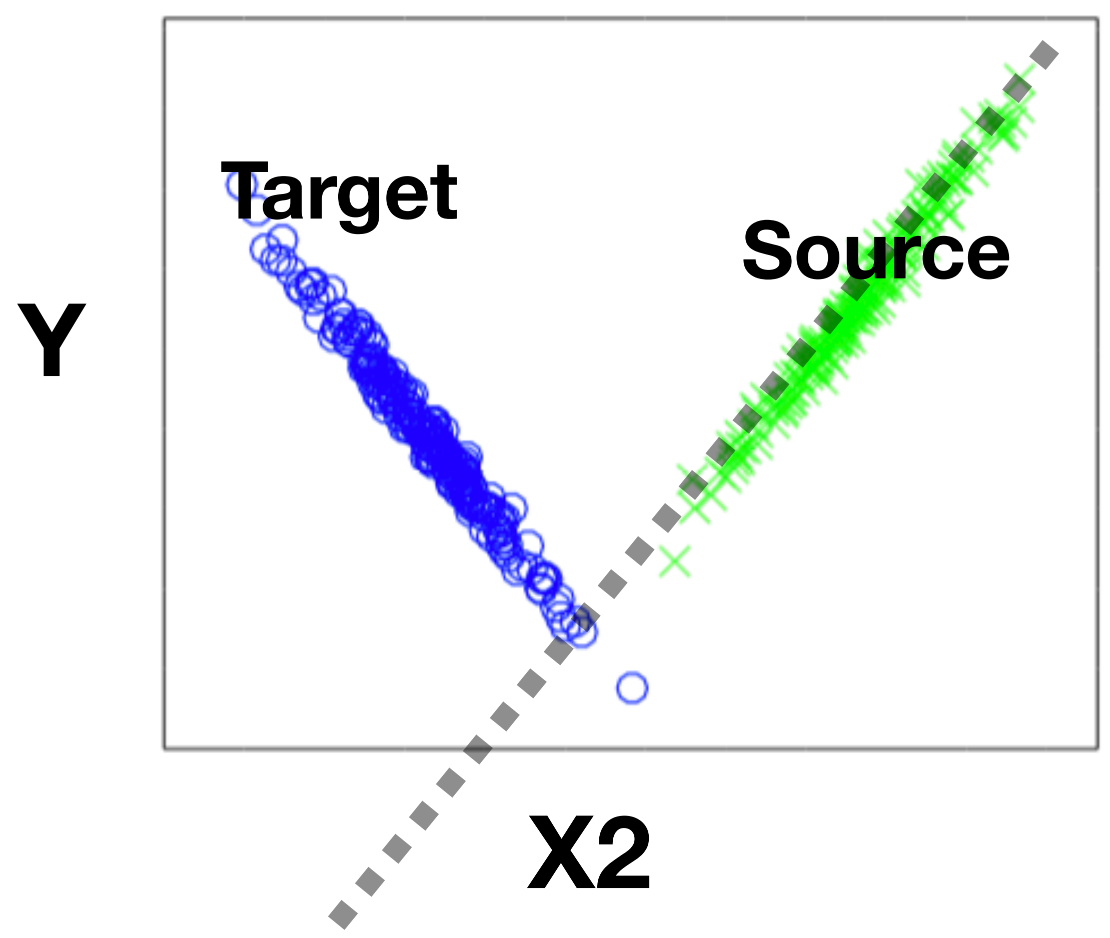
$$P(Y \mid X_1, C = 0) = P(Y \mid X_1, C = 1)$$

Separating features = safe for (adversarial) domain adaptation

- **Separating features:** sets of features that d-separate Y from the context







$Y \not\perp\!\!\!\perp_d C \mid X_2 \iff Y \not\perp\!\!\!\perp C \mid X_2$

(under Markov and faithfulness assumptions)

$Y \perp\!\!\!\perp C \mid X_1 \equiv$

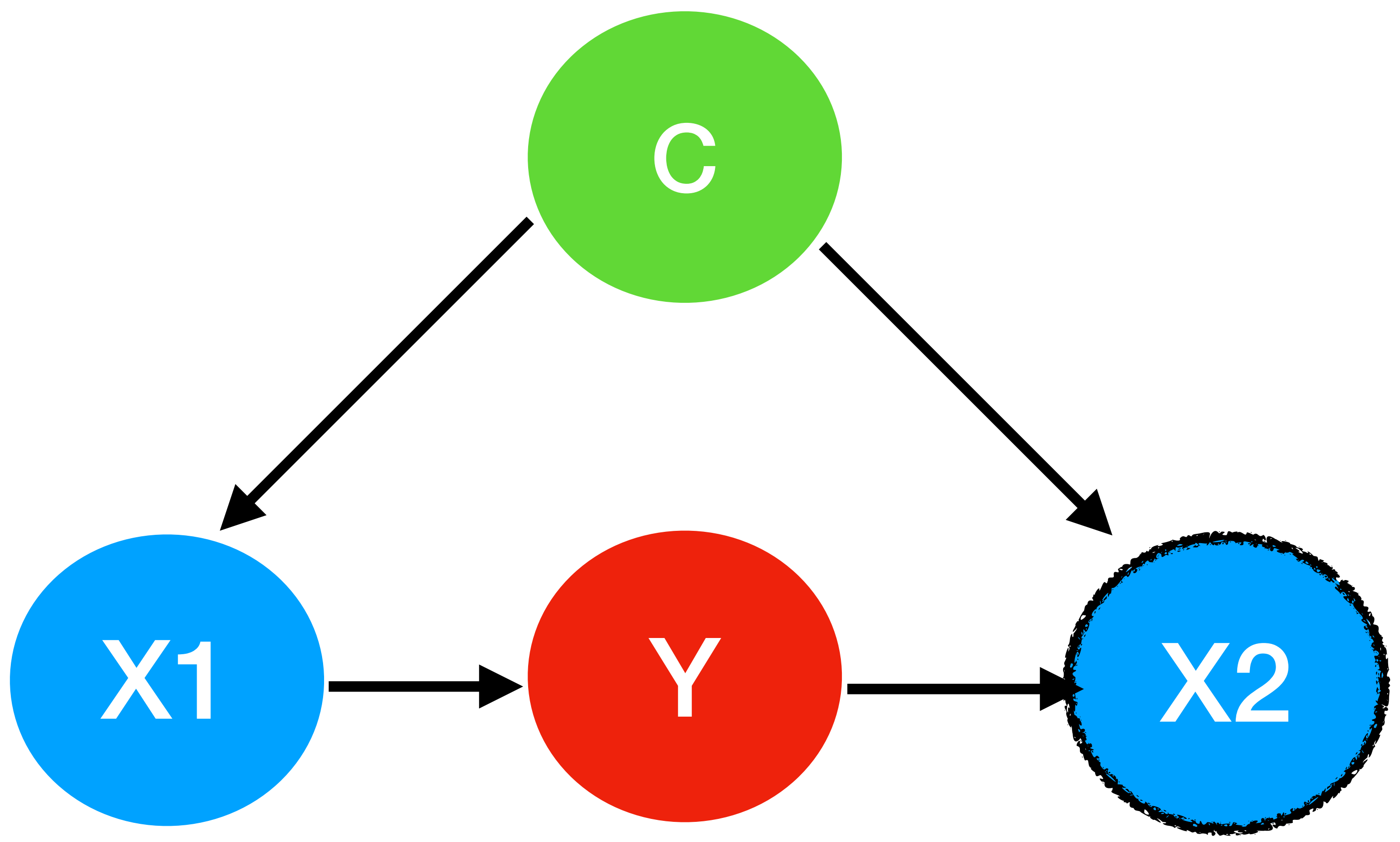
$P(Y \mid X_1, C=0) = P(Y \mid X_1, C=1)$

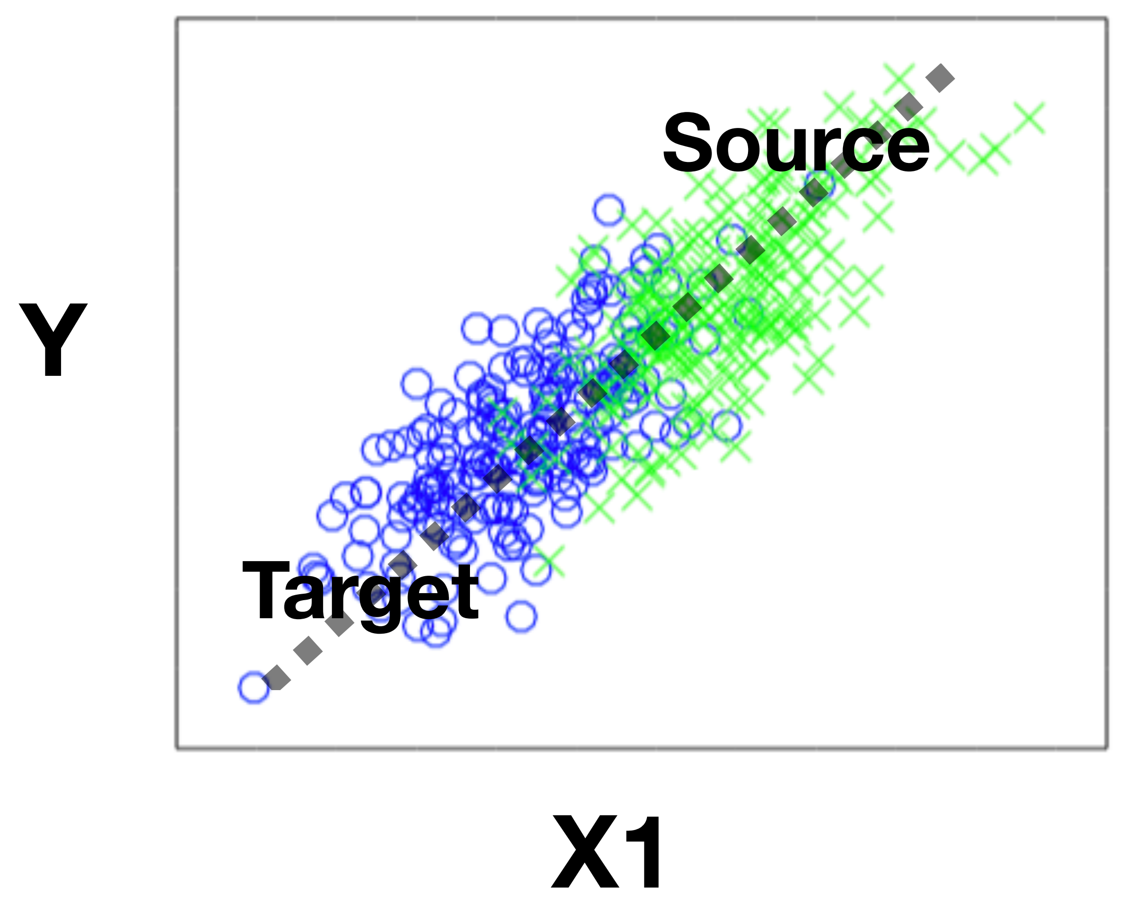
$Y \not\perp\!\!\!\perp C \mid X_2 \equiv$

$P(Y \mid X_2, C=0) \neq P(Y \mid X_2, C=1)$

Separating features = safe for (adversarial) domain adaptation

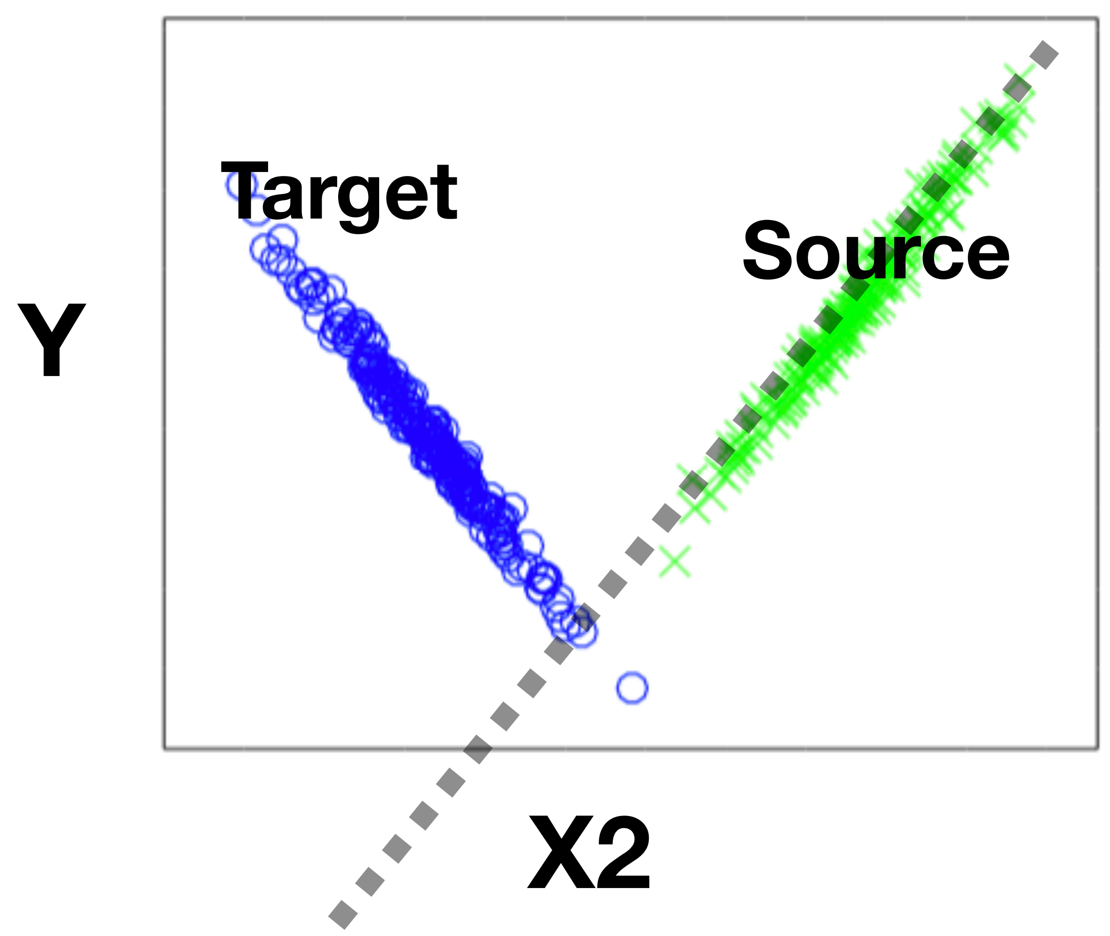
- **Separating features:** sets of features that d-separate Y from the context





$Y \perp\!\!\!\perp C \mid X_1 \equiv$

$P(Y \mid X_1, C=0) = P(Y \mid X_1, C=1)$



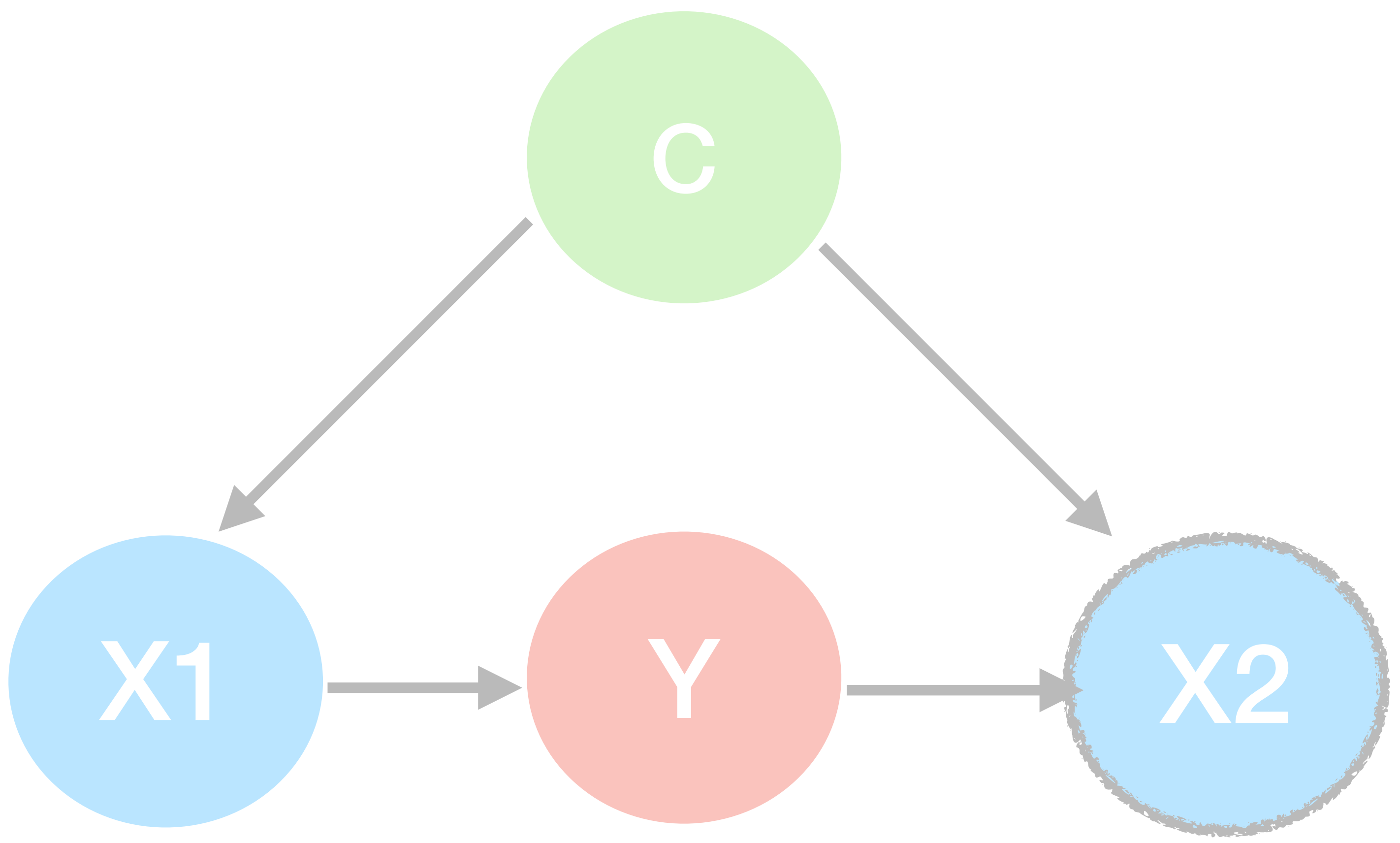
$Y \not\!\perp\!\!\!\perp C \mid X_2 \equiv$

$P(Y \mid X_2, C=0) \neq P(Y \mid X_2, C=1)$

{X1} is a separating feature, {X2} and {X1, X2} are not -> **arbitrarily large error**

Separating features = safe for (adversarial) domain adaptation

- **Separating features:** sets of features that d-separate Y from the context

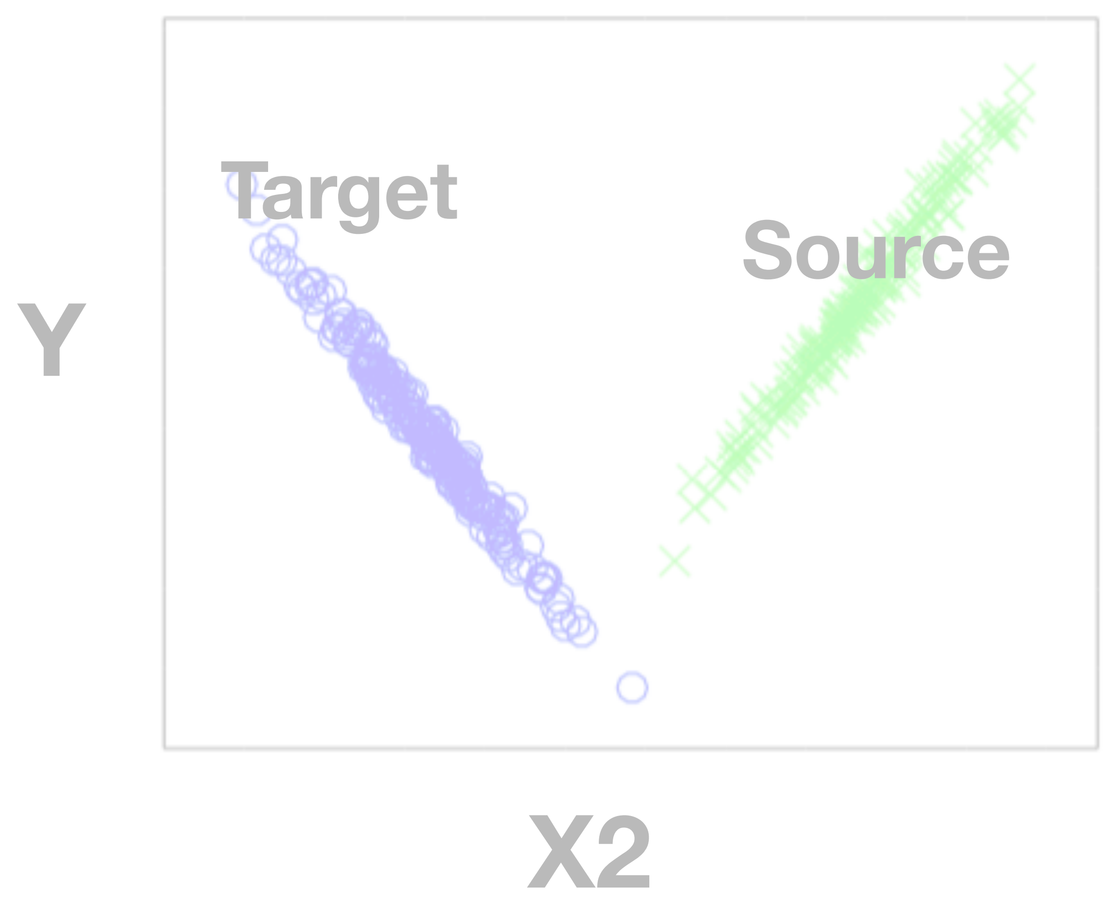


$Y \not\!\perp\!\!\!\perp_d C \mid X_2$

Step 2: Unknown graph and no labels in target!

$Y \perp\!\!\!\perp C \mid X_1 \equiv$

$P(Y \mid X_1, C=0) = P(Y \mid X_1, C=1)$



$Y \not\!\perp\!\!\!\perp C \mid X_2 \equiv$

$P(Y \mid X_2, C=0) \neq P(Y \mid X_2, C=1)$

{X1} is a separating feature, {X2} and {X1, X2} are not -> **arbitrarily large error**

Inferring separating sets of features

Query $Y \perp\!\!\!\perp C_1 \mid X_1 ?$

Assumptions

All testable conditional independences from data

$$X_1 \perp\!\!\!\perp X_3 \mid X_4$$

$$Y \perp\!\!\!\perp C_2 \mid X_1, C_1 = 0$$

$$X_2 \perp\!\!\!\perp C_2 \mid Y, C_1 = 0$$

$$\ldots$$

Logic encoding of d-separation [Hyttinen et al. 2014]

Theorem prover

✓ Provably separating $Y \perp\!\!\!\perp C_1 \mid X_1$

✗ Provably not separating $Y \not\!\perp\!\!\!\perp C_1 \mid X_1$

? Not identifiable

We can test incrementally each set of features selected by standard feature selection (e.g. random forests)

https://arxiv.org/abs/1707.06422

Inferring separating sets of features

Query $Y \perp\!\!\!\perp C_1 | X_1$?

Assumptions

All testable conditional independences from data

$X_1 \perp\!\!\!\perp X_3 | X_4$

$Y \perp\!\!\!\perp C_2 | X_1, C_1 = 0$

$X_2 \perp\!\!\!\perp C_2 | Y, C_1 = 0$

...

Logic encoding of d-separation [Hyttinen et al. 2014]

A big (current) limitation: Scalability

Provably separating $Y \perp\!\!\!\perp C_1 | X_1$

Provably not separating $Y \not\perp\!\!\!\perp C_1 | X_1$

?

Not identifiable

We can test incrementally each set of features selected by standard feature selection (e.g. random forests)

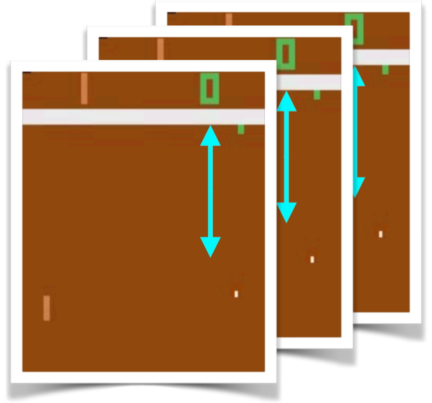
https://arxiv.org/abs/1707.06422

AdaRL: What, Where, and How to Adapt in Transfer RL

Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang ICLR 2022

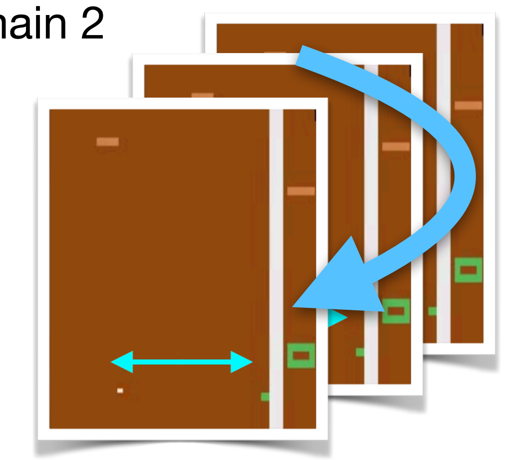
Source domains

Domain 1



$\{\text{player}_t, \text{ball}_t, \text{advs}_t, a_t, \text{r}_t\}_{t=0,\ldots,T}$

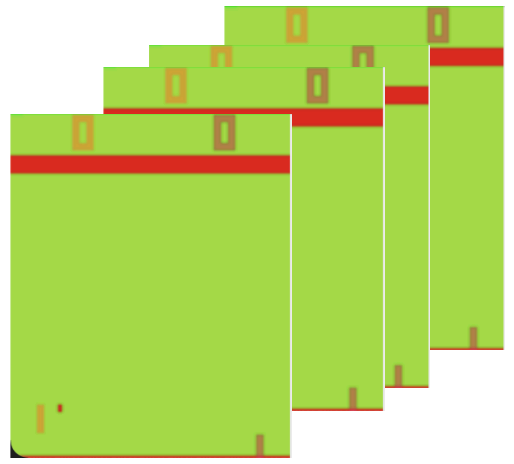
Domain 2



$\{\text{player}_t, \text{ball}_t, \text{advs}_t, a_t, \text{r}_t\}_{t=0,\ldots,T}$

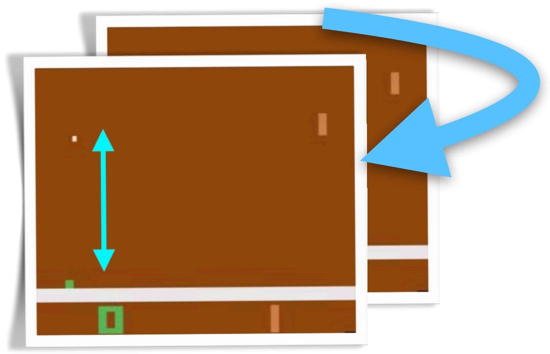
. . .

Domain n



$\{\text{player}_t, \text{ball}_t, \text{advs}_t, a_t, \text{r}_t\}_{t=0,\ldots,T}$

Target domain

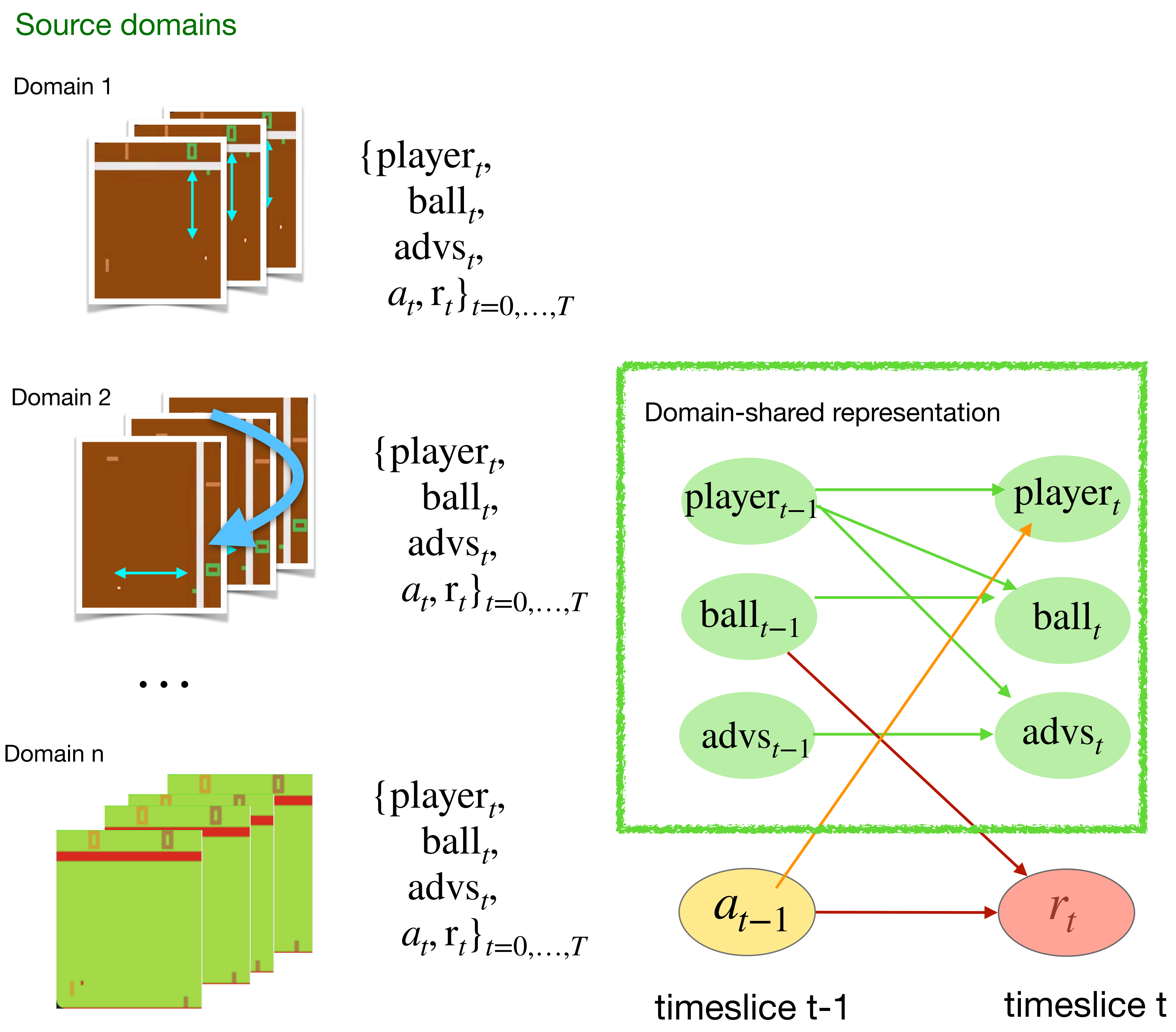


$\{o_t, a_t, \text{r}_t\}_{t=0,\ldots,T}$

https://arxiv.org/abs/2107.02729

AdaRL: What, Where, and How to Adapt in Transfer RL

Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang ICLR 2022

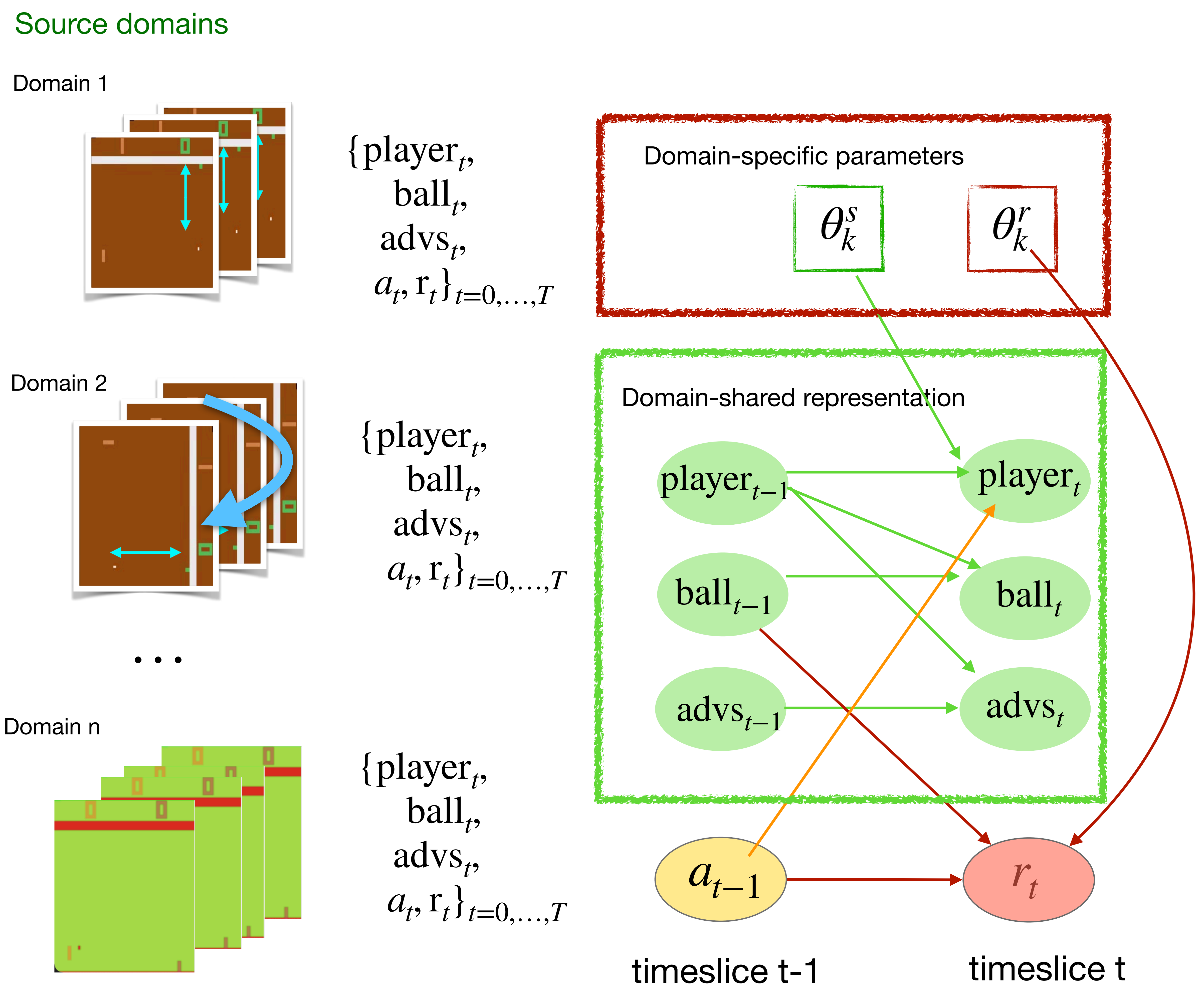


- Learn a **factored** MDP (symbolic inputs)

https://arxiv.org/abs/2107.02729

AdaRL: What, Where, and How to Adapt in Transfer RL

Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang **ICLR 2022**

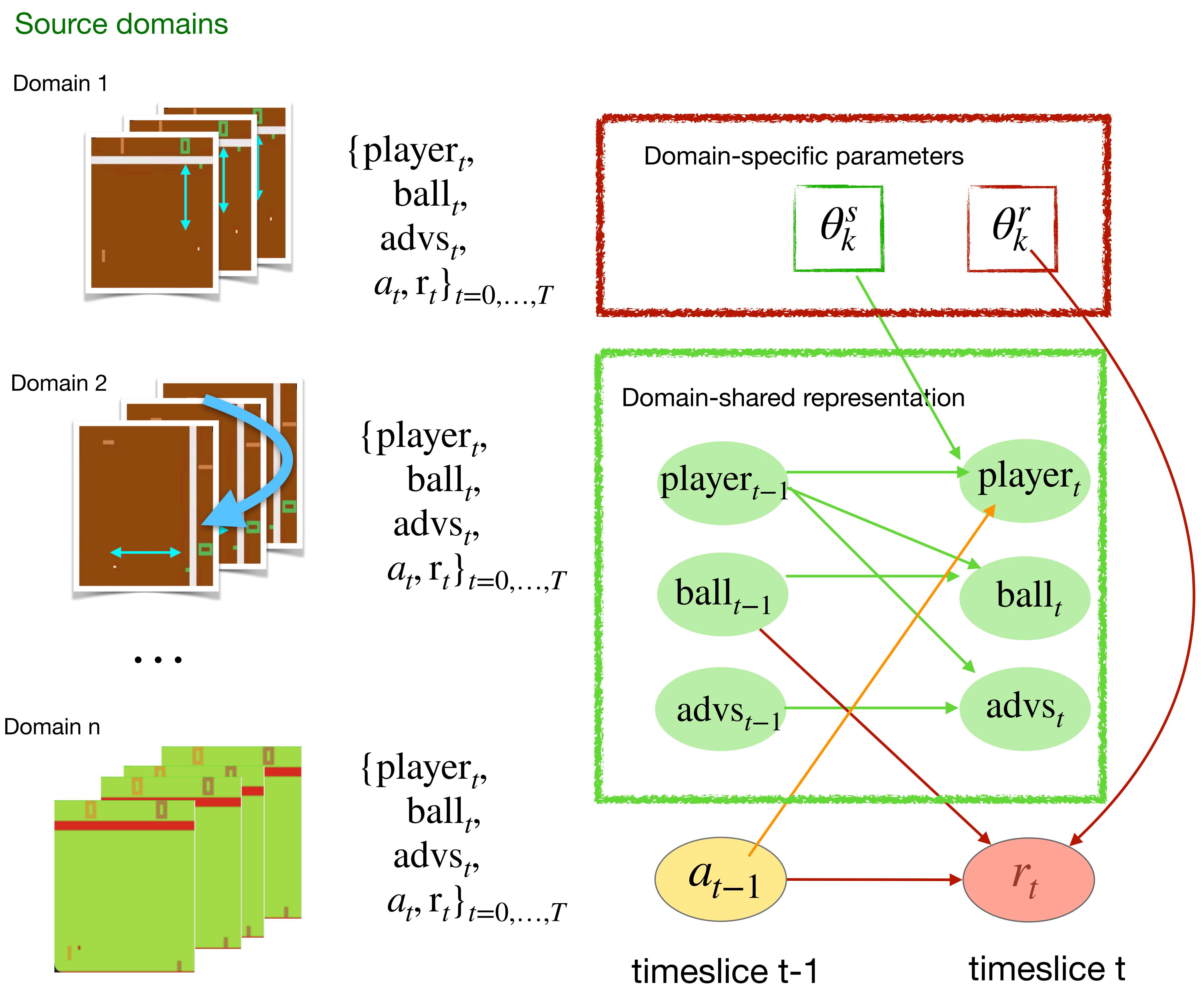


- Learn a **factored** MDP (symbolic inputs) with **latent change factors** that are constant in each domain, but **vary across domains**

https://arxiv.org/abs/2107.02729

AdaRL: What, Where, and How to Adapt in Transfer RL

Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang **ICLR 2022**

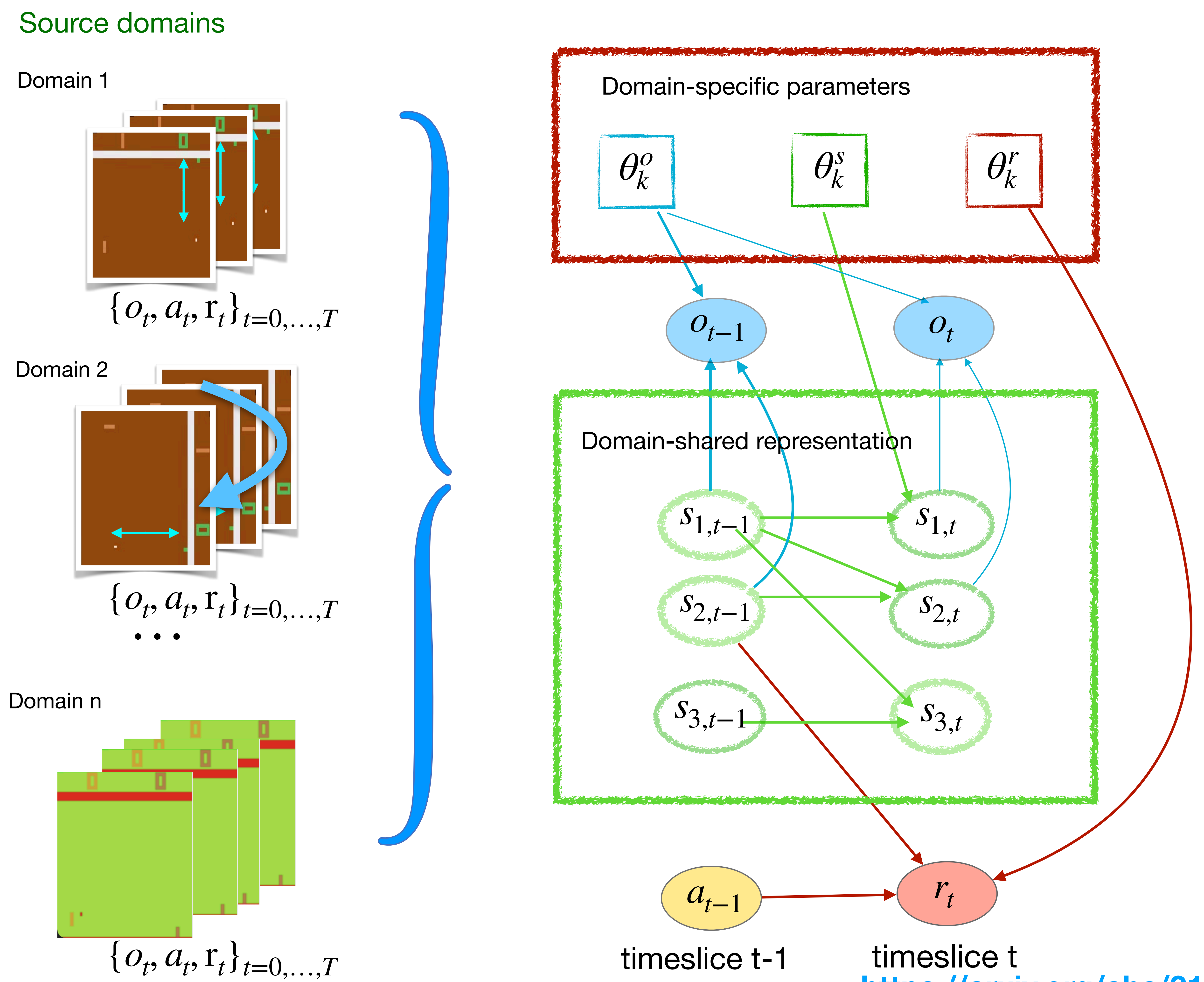


When we learn from symbolic inputs, the causal graph can be identified, but we don't have guarantees on what the latent change factors are

https://arxiv.org/abs/2107.02729

AdaRL: What, Where, and How to Adapt in Transfer RL

Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang **ICLR 2022**



https://arxiv.org/abs/2107.02729

When we learn from images, we cannot identify the causal variables, so what we learn is not necessarily causal… but it is still useful

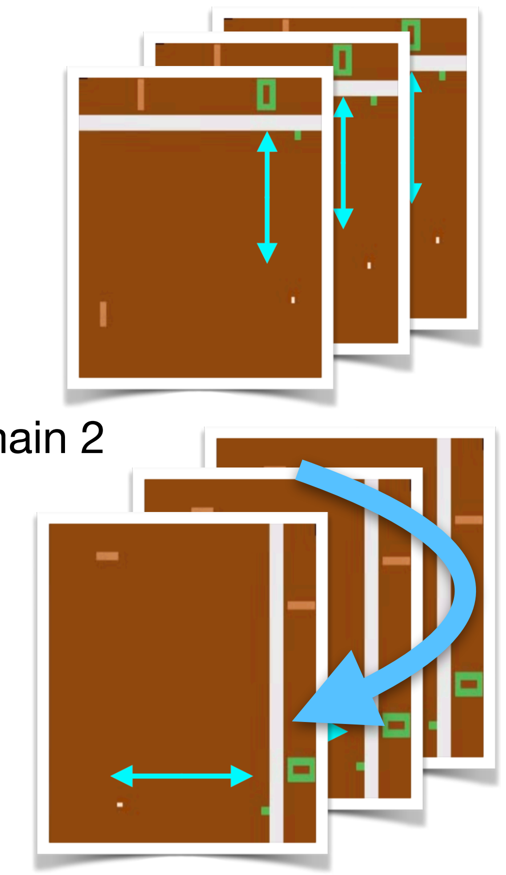
AdaRL: What, Where, and How to Adapt in Transfer RL

Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang ICLR 2022

Source domains

Domain 1

Domain 2



. . .

Estimate graph over estimated s_k, θ_k

Identify s_t^{min} , θ_t^{min} from the estimated graph

Learn optimal policy $\pi^*(s_k^{min}, \theta_k^{min})$ on source domains

https://arxiv.org/abs/2107.02729

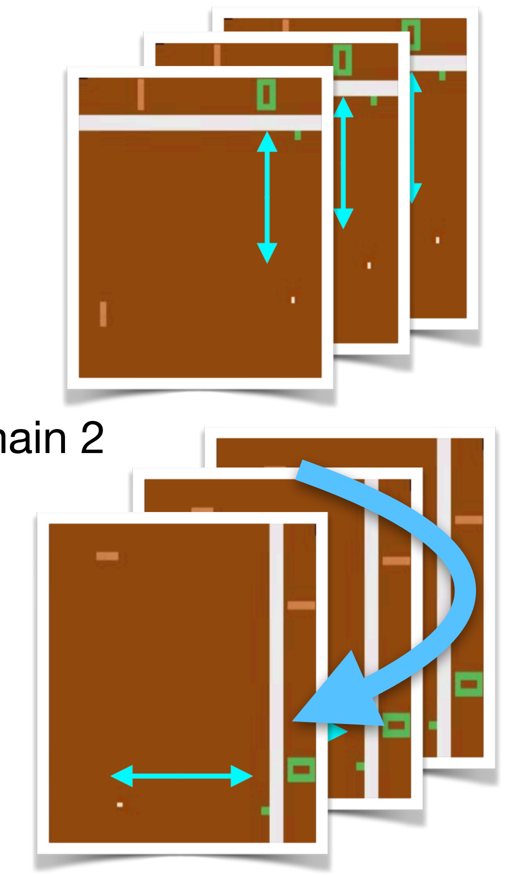
AdaRL: What, Where, and How to Adapt in Transfer RL

Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang ICLR 2022

Source domains

Domain 1

Domain 2



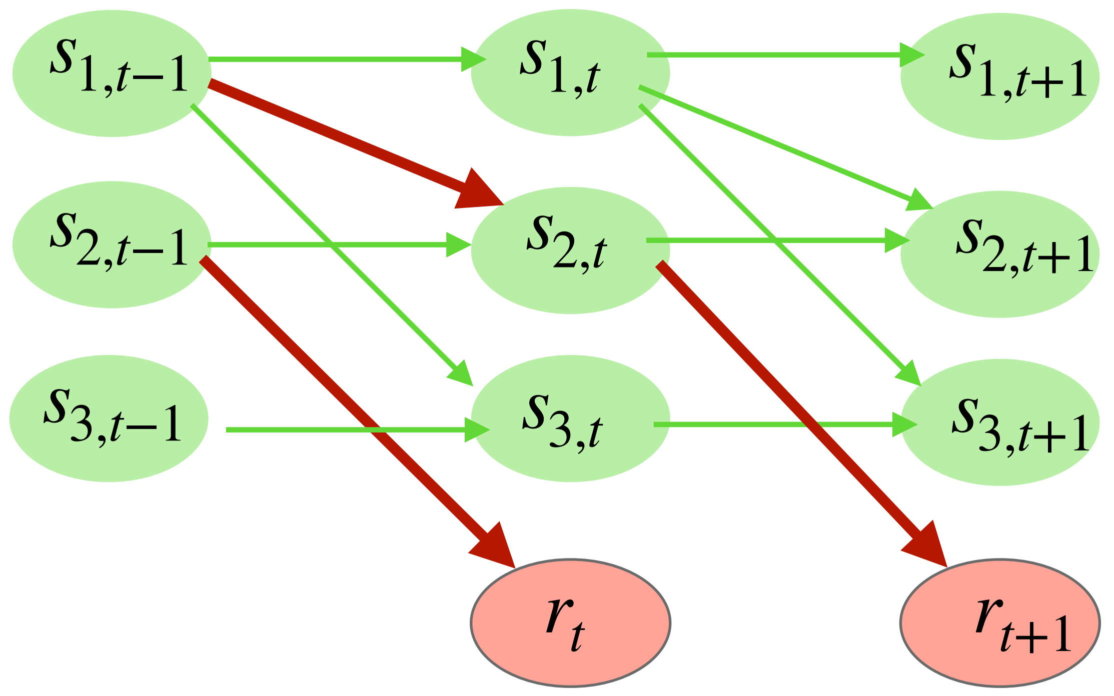
. . .

Estimate graph over estimated s_k, θ_k

Identify s_t^{min} , θ_t^{min} from the estimated graph

Learn optimal policy $\pi^*(s_k^{min}, \theta_k^{min})$ on source domains

- Identify the dimensions of the state and change factors that are **necessary and sufficient** for policy optimisation



https://arxiv.org/abs/2107.02729

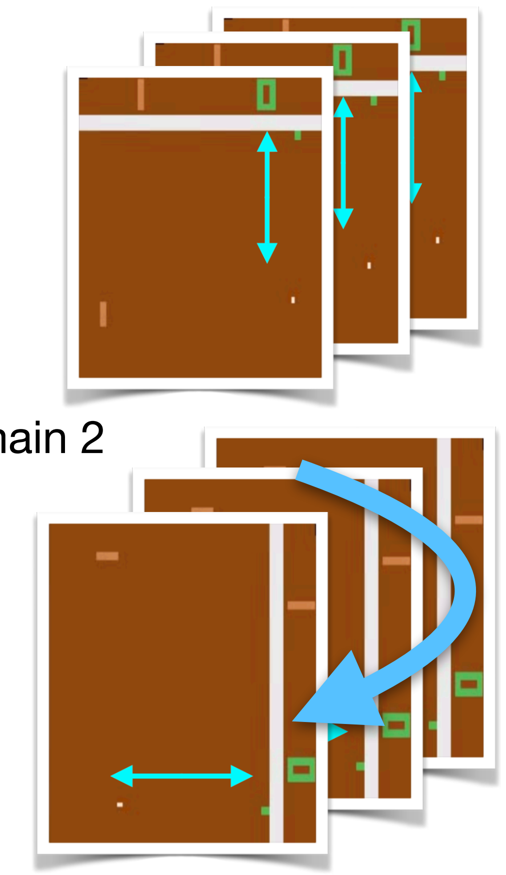
AdaRL: What, Where, and How to Adapt in Transfer RL

Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang ICLR 2022

Source domains

Domain 1

Domain 2



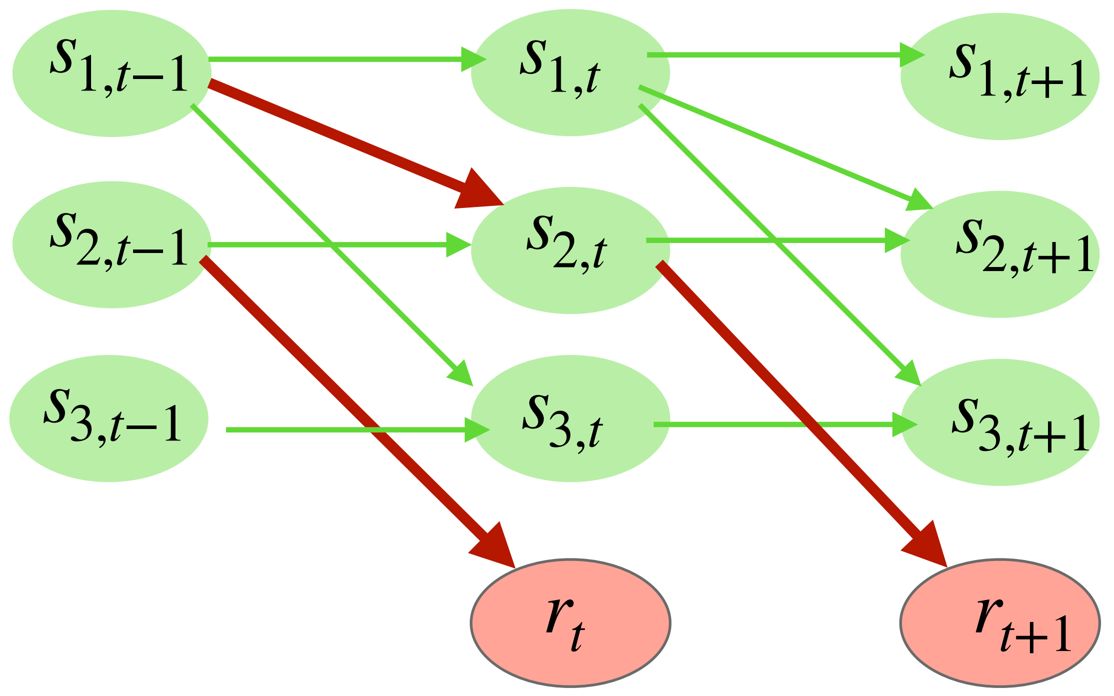
. . .

Estimate graph over estimated s_k, θ_k

Identify $s_t^{min}, \theta_t^{min}$ from the estimated graph

Learn optimal policy $\pi^*(s_k^{min}, \theta_k^{min})$ on source domains

- Identify the dimensions of the state and change factors that are **necessary and sufficient** for policy optimisation

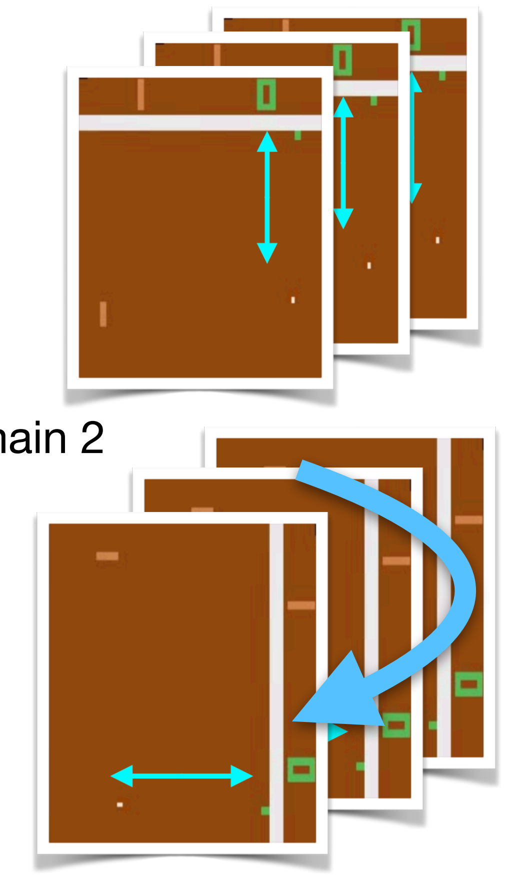


https://arxiv.org/abs/2107.02729

AdaRL: What, Where, and How to Adapt in Transfer RL

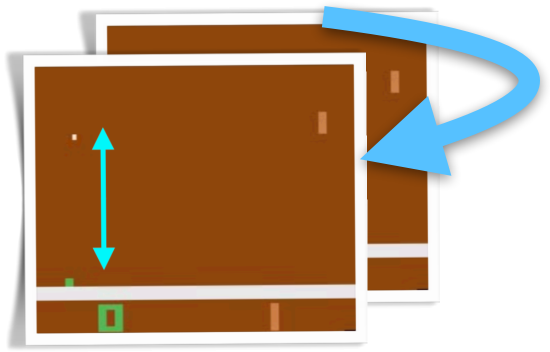
Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang ICLR 2022

Source domains



…

Target domain



$\{o_t, a_t, r_t\}_{t=0,\ldots,T}$

Estimate graph over estimated s_k, θ_k

Identify s_t^{min} , θ_t^{min} from the estimated graph

Learn optimal policy $\pi^*(s_k^{min}, \theta_k^{min})$ on source domains

Simplifying assumption: no new edges in target domain

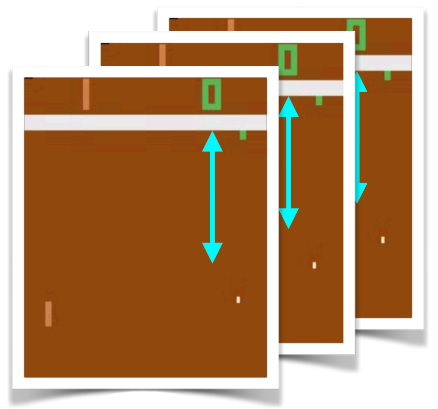
https://arxiv.org/abs/2107.02729

AdaRL: What, Where, and How to Adapt in Transfer RL

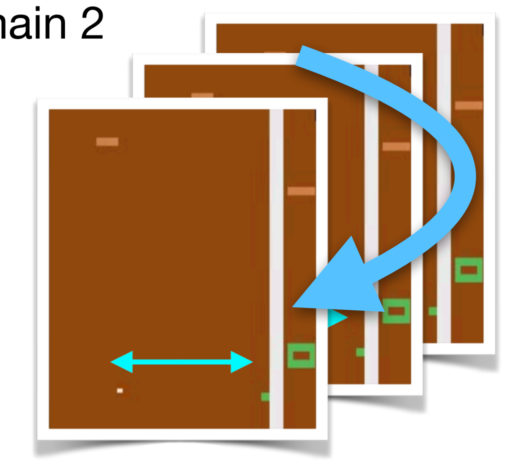
Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang ICLR 2022

Source domains

Domain 1



Domain 2



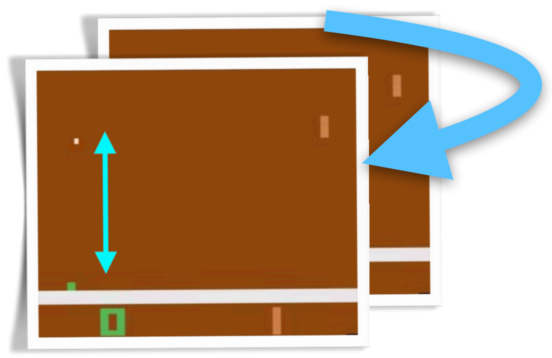
...

Estimate graph over estimated s_k, θ_k

Identify s_t^{min} , θ_t^{min} from the estimated graph

Learn optimal policy $\pi^*(s_k^{min}, \theta_k^{min})$ on source domains

Target domain



$\{o_t, a_t, r_t\}_{t=0,\ldots,T}$

Use model to estimate $s_{target}^{min}, \theta_{target}^{min}$ with few samples

Apply policy $\pi^*(s_{target}^{min}, \theta_{target}^{min})$

Simplifying assumption: no new edges in target domain

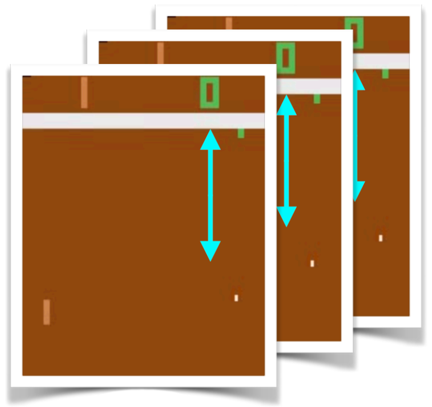
https://arxiv.org/abs/2107.02729

AdaRL: What, Where, and How to Adapt in Transfer RL

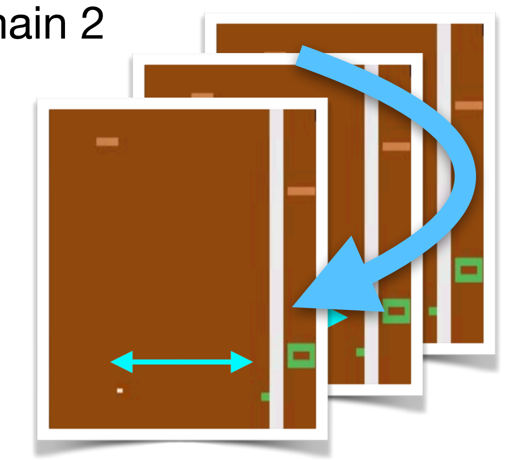
Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang
ICLR 2022

Source domains

Domain 1



Domain 2



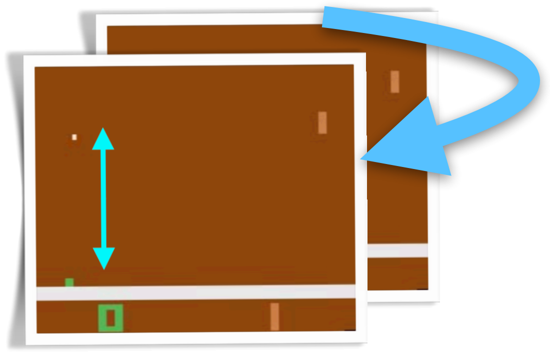
. . .

Estimate graph over estimated s_k, θ_k

Identify s_t^{min} , θ_t^{min} from the estimated graph

Learn optimal policy $\pi^*(s_k^{min}, \theta_k^{min})$ on source domains

Target domain



$\{o_t, a_t, r_t\}_{t=0,\ldots,T}$

Use model to estimate $s_{target}^{min}, \theta_{target}^{min}$ with few samples

Apply policy $\pi^*(s_{target}^{min}, \theta_{target}^{min})$

Simplifying assumption: no new edges in target domain

https://arxiv.org/abs/2107.02729

AdaRL: What, Where, and How to Adapt in Transfer RL

Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang ICLR 2022

- **Results:** we consistently outperform the state-of-the-art **thanks to the graph**

	Oracle Upper bound	Non-t lower bound	CAVIA (Zintgraf et al., 2019)	PEARL (Rakelly et al., 2019)	AdaRL* Ours w/o masks	AdaRL Ours
G_in	2486.1 (±369.7)	1098.5 • (±472.1)	1603.0 (±877.4)	1647.4 (±617.2)	1940.5 (±841.7)	2217.6 (±981.5)
G_out	693.9 (±100.6)	204.6 • (±39.8)	392.0 • (±125.8)	434.5 • (±102.4)	439.5 • (±157.8)	**508.3** (±138.2)
M_in	2678.2 (±630.5)	748.5 • (±342.8)	2139.7 (±859.6)	1784.0 (±845.3)	1946.2 • (±496.5)	2260.2 (±682.8)
M_out	1405.6 (±368.0)	371.0 • (±92.5)	972.6 • (±401.4)	793.9 • (±394.2)	874.5 • (±290.8)	**1001.7** (±273.3)
G_in & M_in	1984.2 (±871.3)	365.0 • (±144.5)	1012.5 • (±664.9)	1260.8 • (±792.0)	1157.4 • (±578.5)	**1428.4** (±495.6)
G_out & M_out	939.4 (±270.5)	336.9 • (±139.6)	648.2 • (±481.5)	544.32 • (±175.2)	596.0 • (±184.3)	**689.4** (±272.5)

Average final scores on Cartpole (MDP) with N_target=50

	Oracle Upper bound	Non-t lower bound	PNN (Rusu et al., 2016)	PSM (Agarwal et al., 2021a)	MTQ (Fakoor et al., 2020)	AdaRL* Ours w/o masks	AdaRL Ours
O_in	18.65 (±2.43)	6.18 • (±2.43)	9.70 • (±2.09)	11.61 • (±3.85)	15.79 • (±3.26)	14.27 • (±1.93)	**18.97** (±2.00)
O_out	19.86 (±1.09)	6.40 • (±3.17)	9.54 • (±2.78)	10.82 • (±3.29)	10.82 • (±4.13)	12.67 • (±2.49)	**15.75** (±3.80)
C_in	19.35 (±0.45)	8.53 • (±2.08)	14.44 • (±2.37)	19.02 (±1.17)	16.97 • (±2.02)	18.52 • (±1.41)	19.14 (±1.05)
C_out	19.78 (±0.25)	8.26 • (±3.45)	14.84 • (±1.98)	17.66 • (±2.46)	15.45 • (±3.30)	17.92 (±1.83)	19.03 (±0.97)
S_in	18.32 (±1.18)	6.91 • (±2.02)	11.80 • (±3.25)	12.65 • (±3.72)	13.68 • (±3.49)	14.23 • (±3.19)	**16.65** (±1.72)
S_out	19.01 (±1.04)	6.60 • (±3.11)	9.07 • (±4.58)	8.45 • (±4.51)	11.45 • (±2.46)	12.80 • (±2.62)	**17.82** (±2.35)
N_in	18.48 (±1.25)	5.51 • (±3.88)	12.73 • (±3.67)	11.30 • (±2.58)	12.67 • (±3.84)	13.78 • (±2.15)	**16.84** (±3.13)
N_out	18.26 (±1.11)	6.02 • (±3.19)	13.24 • (±2.55)	11.26 • (±3.15)	15.77 • (±2.12)	14.65 • (±3.01)	**18.30** (±2.24)

Average final scores on Pong (POMDP) with N_target=50

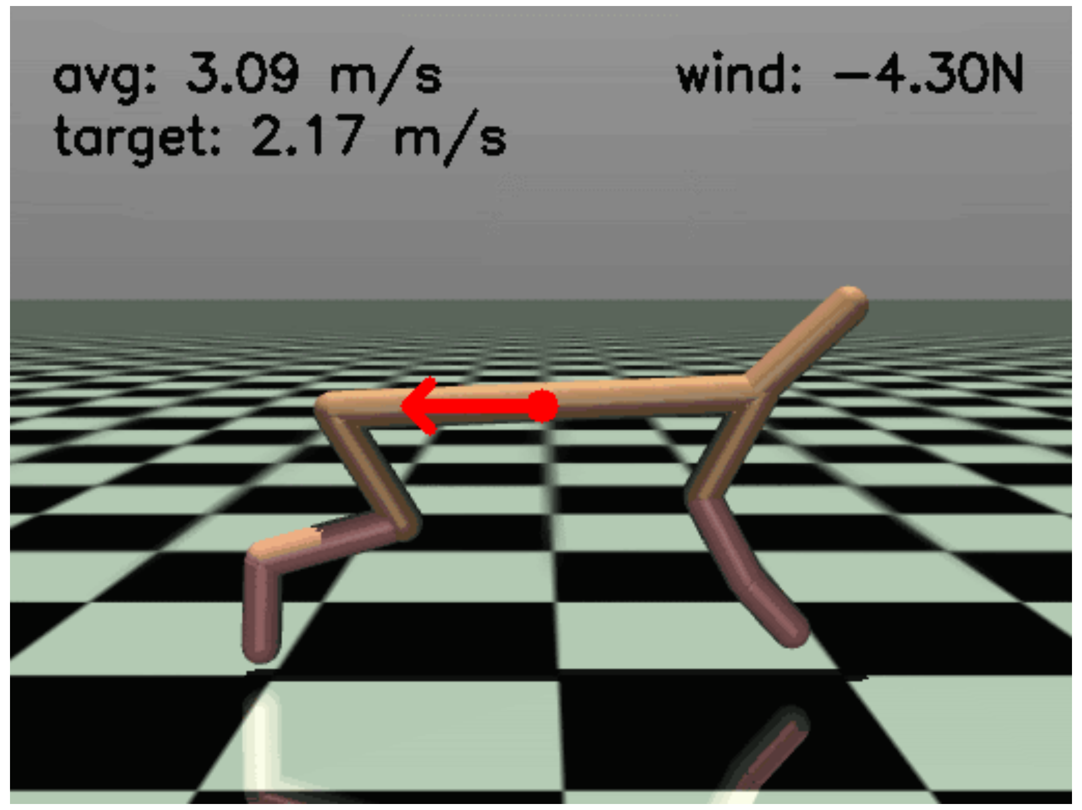
https://arxiv.org/abs/2107.02729

FansRL: Factored Adaptation for Non-Stationary Reinforcement Learning

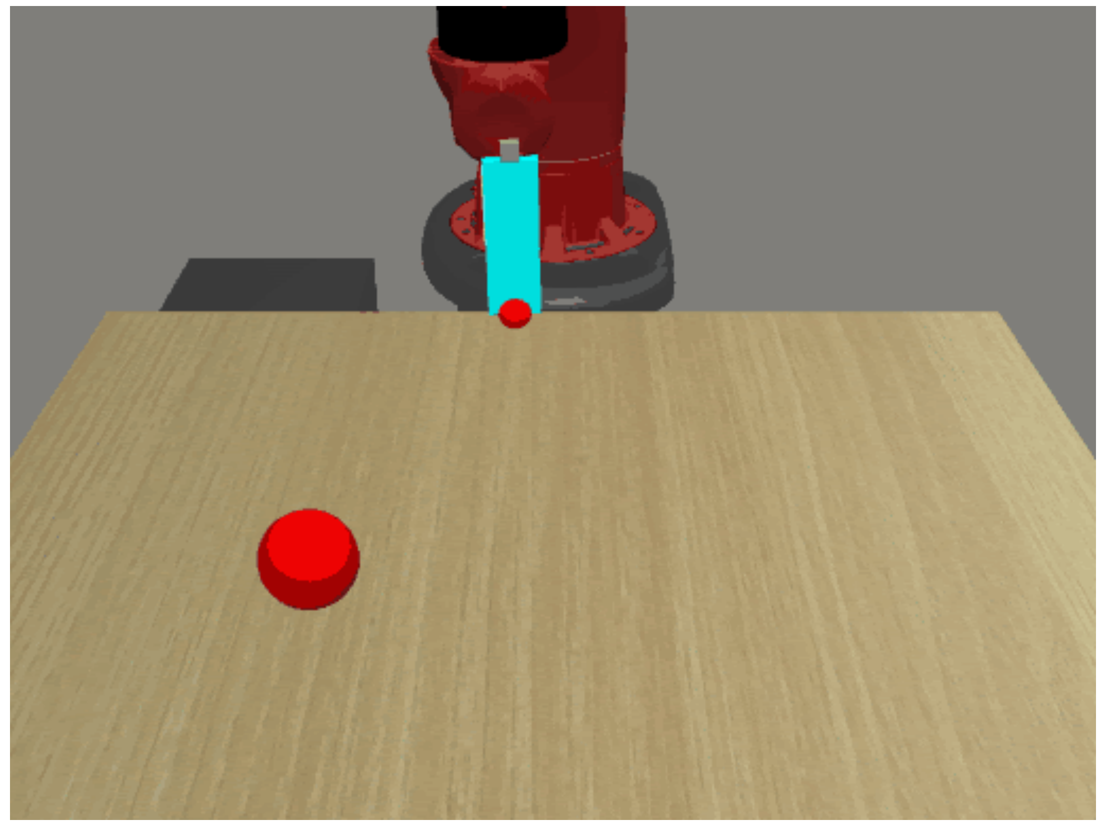
Fan Feng, Biwei Huang, Kun Zhang, Sara Magliacane

NeurIPS 2022

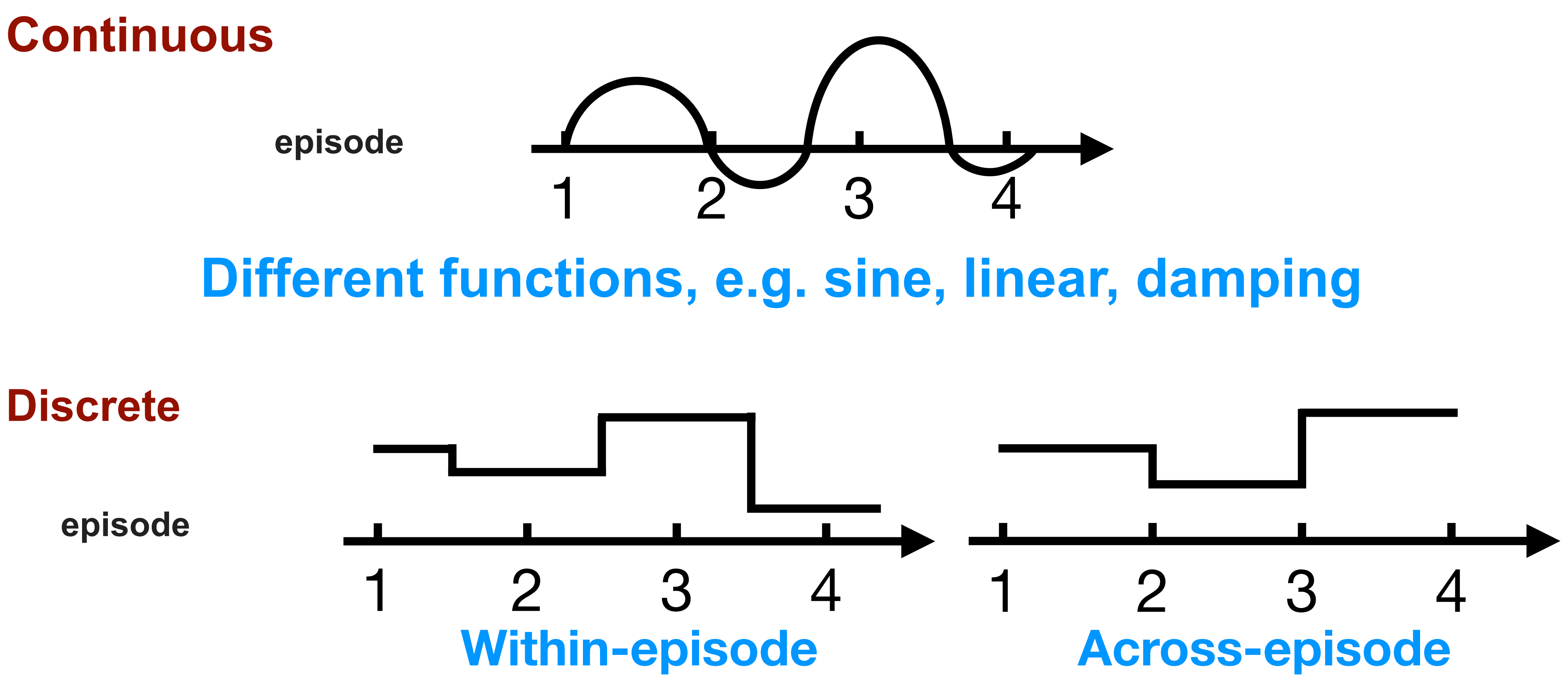
- **Task:** RL agent has to learn a policy that is robust to different types of non-stationarity, including **multiple simultaneous changes of different types**



Non-stationary environments
(wind changes)



Non-stationary rewards
(target changes)

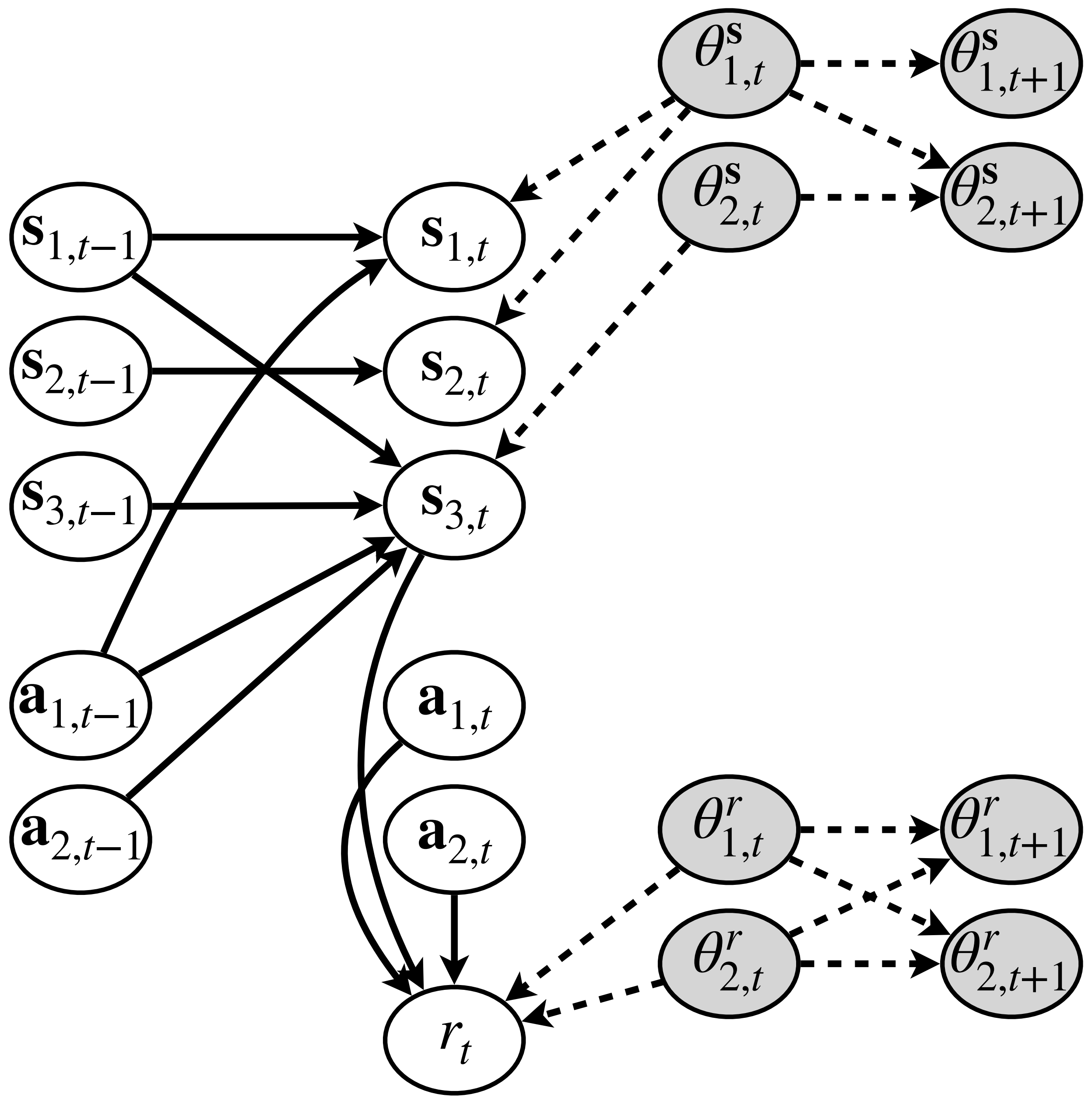


https://arxiv.org/abs/2203.16582

FansRL: Factored Adaptation for Non-Stationary Reinforcement Learning

Fan Feng, Biwei Huang, Kun Zhang, Sara Magliacane NeurIPS 2022

- The **latent change factors** are not constant anymore and they model **non-stationarity**



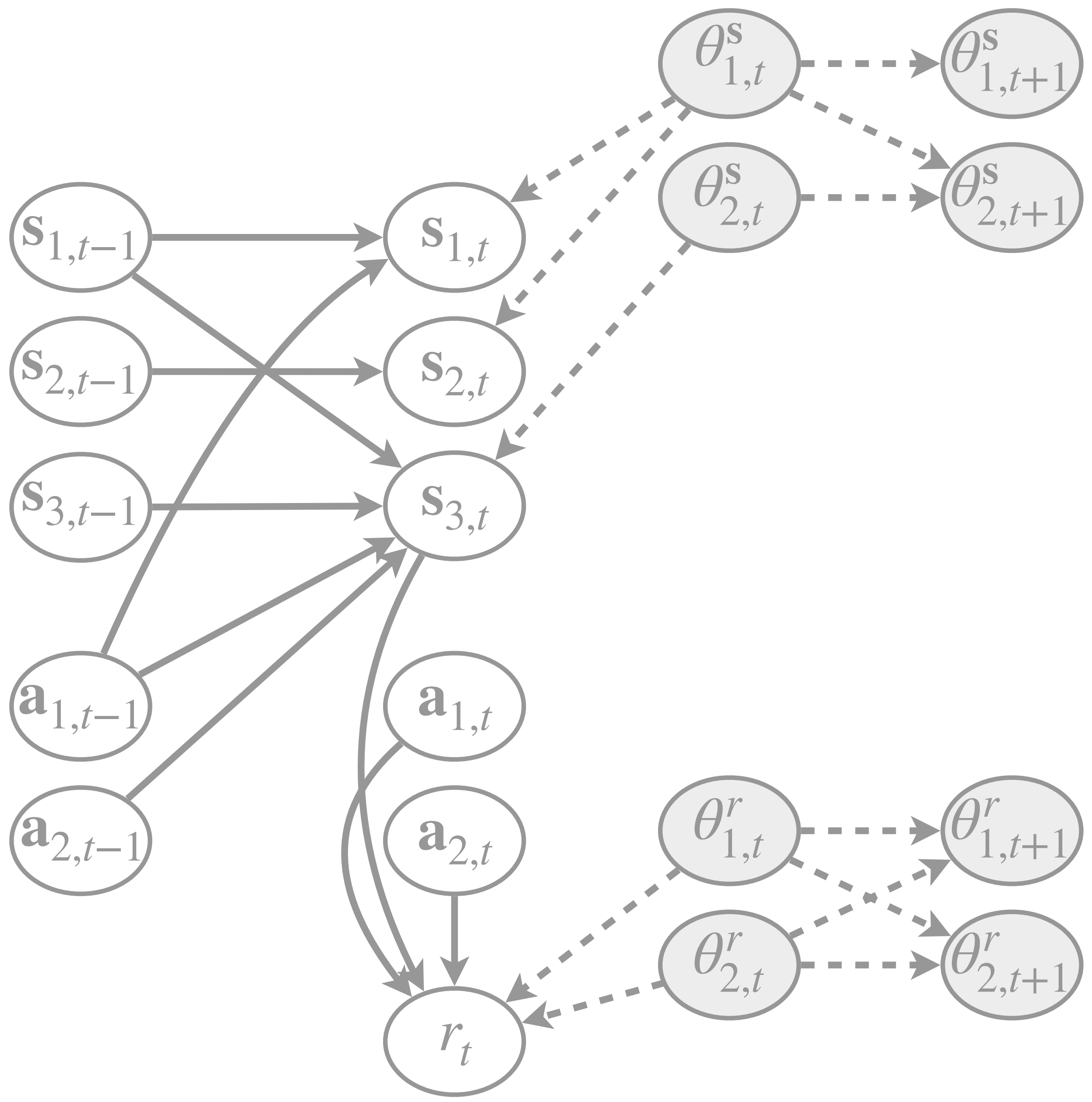
Factored Non-Stationary MDP

https://arxiv.org/abs/2203.16582

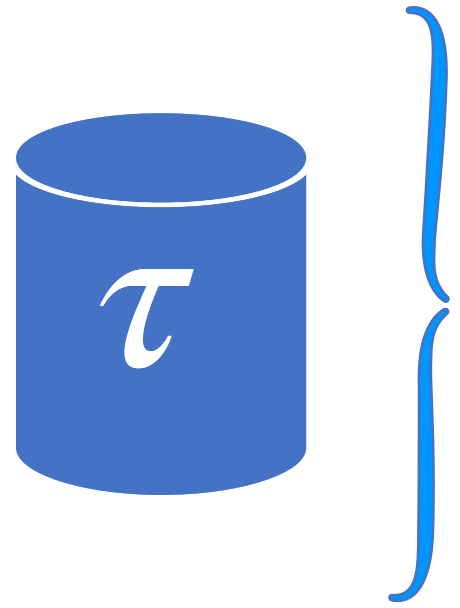
FansRL: Factored Adaptation for Non-Stationary Reinforcement Learning

Fan Feng, Biwei Huang, Kun Zhang, Sara Magliacane NeurIPS 2022

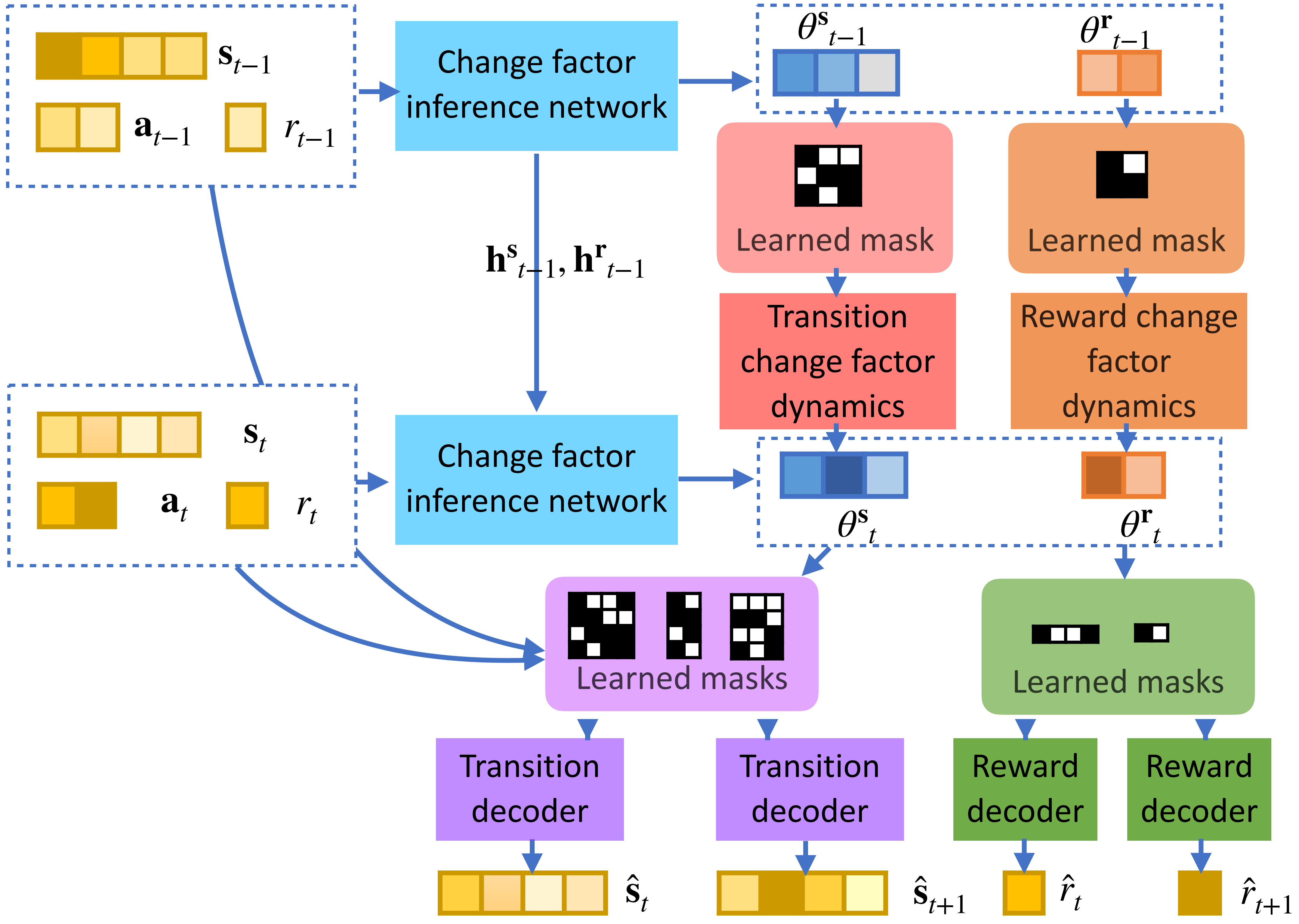
- The **latent change factors** are not constant anymore and they model **non-stationarity**



Factored Non-Stationary MDP



Trajectories collected with an initial policy (e.g. random)



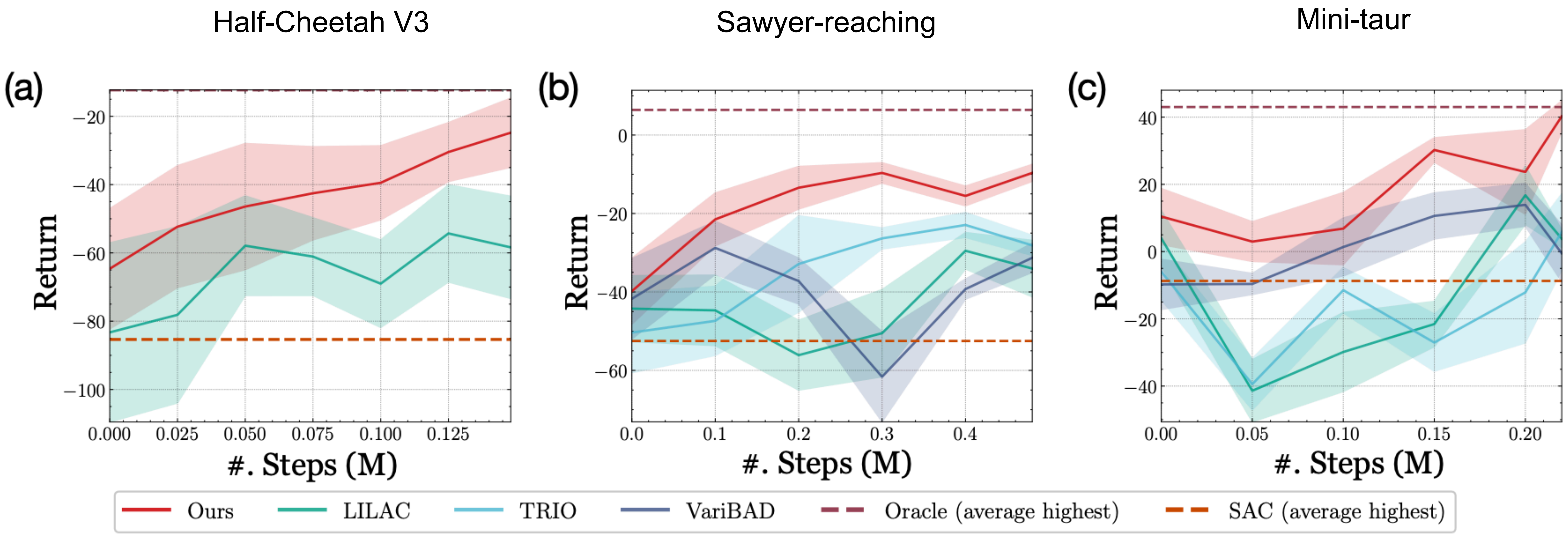
Factored Non-Stationary Variational Autoencoder

https://arxiv.org/abs/2203.16582

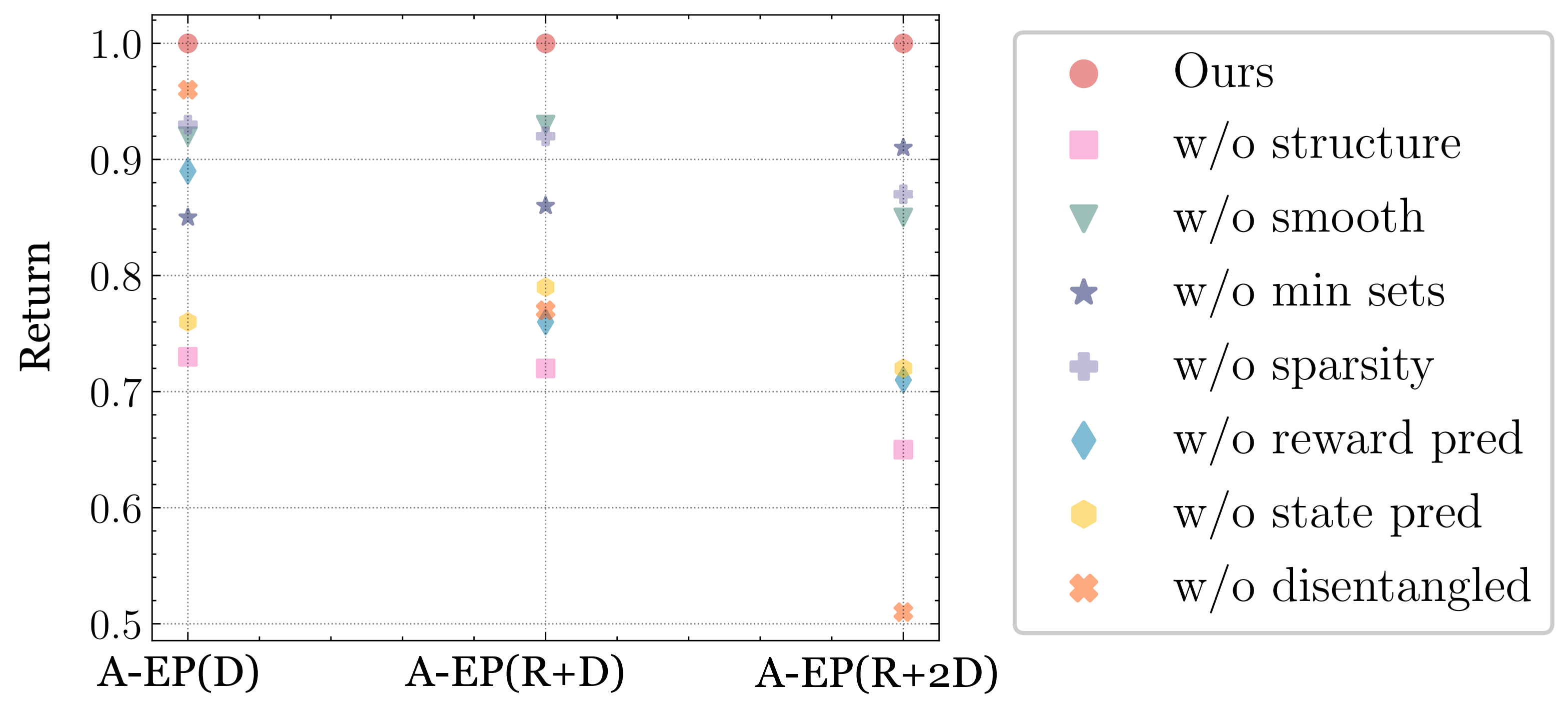
FansRL: Factored Adaptation for Non-Stationary Reinforcement Learning

Fan Feng, Biwei Huang, Kun Zhang, Sara Magliacane **NeurIPS 2022**

- **Policy learning:** estimate latent change factors, learn policy as if they were observed
- **Results:** we consistently outperform the state-of-the-art **thanks to the graph**



Ablation studies



Continuous changes on dynamics (sine wind)

Across-episode changes on rewards (changing target)

Across-episode changes on both dynamics (mass) and reward (target velocity)

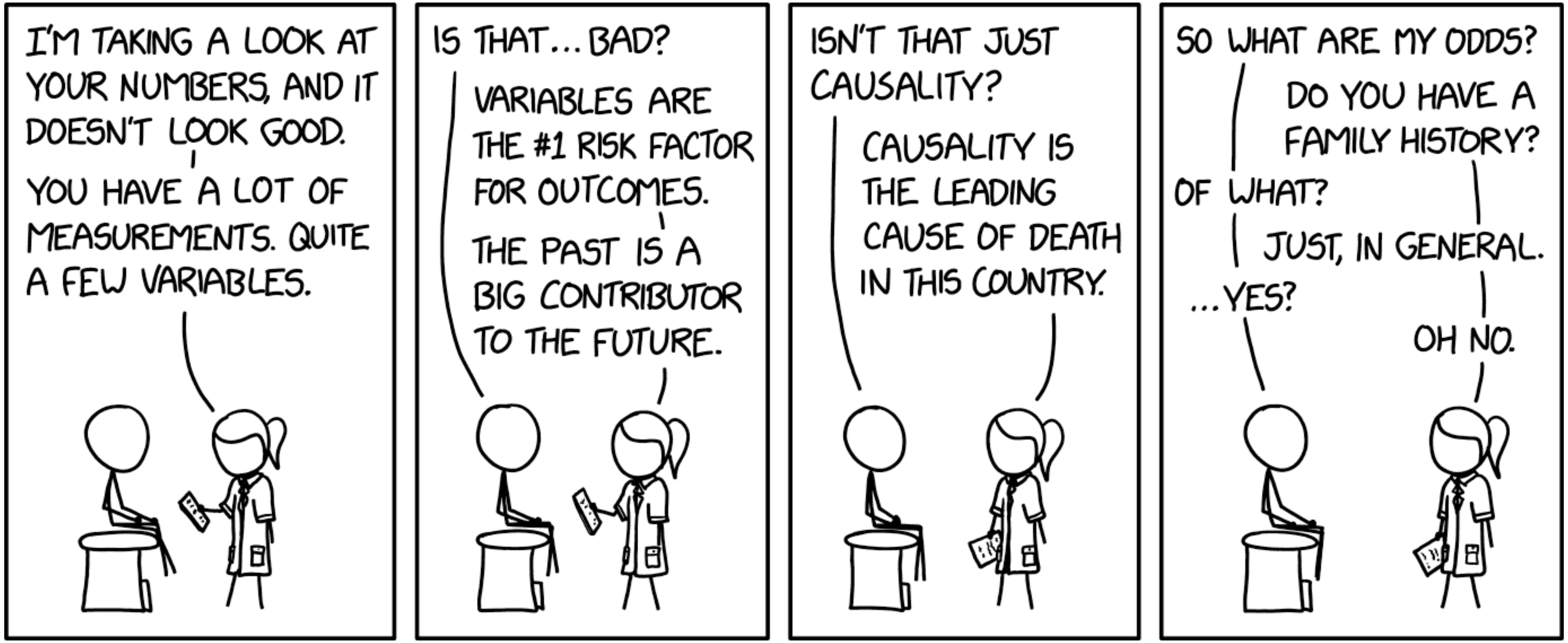
The biggest difference in performance is switching off learning the graph

https://arxiv.org/abs/2203.16582

Takeaways

- **Causal representation learning (learn causal variables from images)**
  - Requires a lot of interventional data or strong assumptions, not ready yet for RL
  - Provides theoretical guarantees, could allow for better generalization
- **Causality-inspired representation learning (learn graphs from images)**
  - No requirements on interventional data, but no identifiability guarantees
  - Still empirically useful in RL

Questions??



https://xkcd.com/2620/