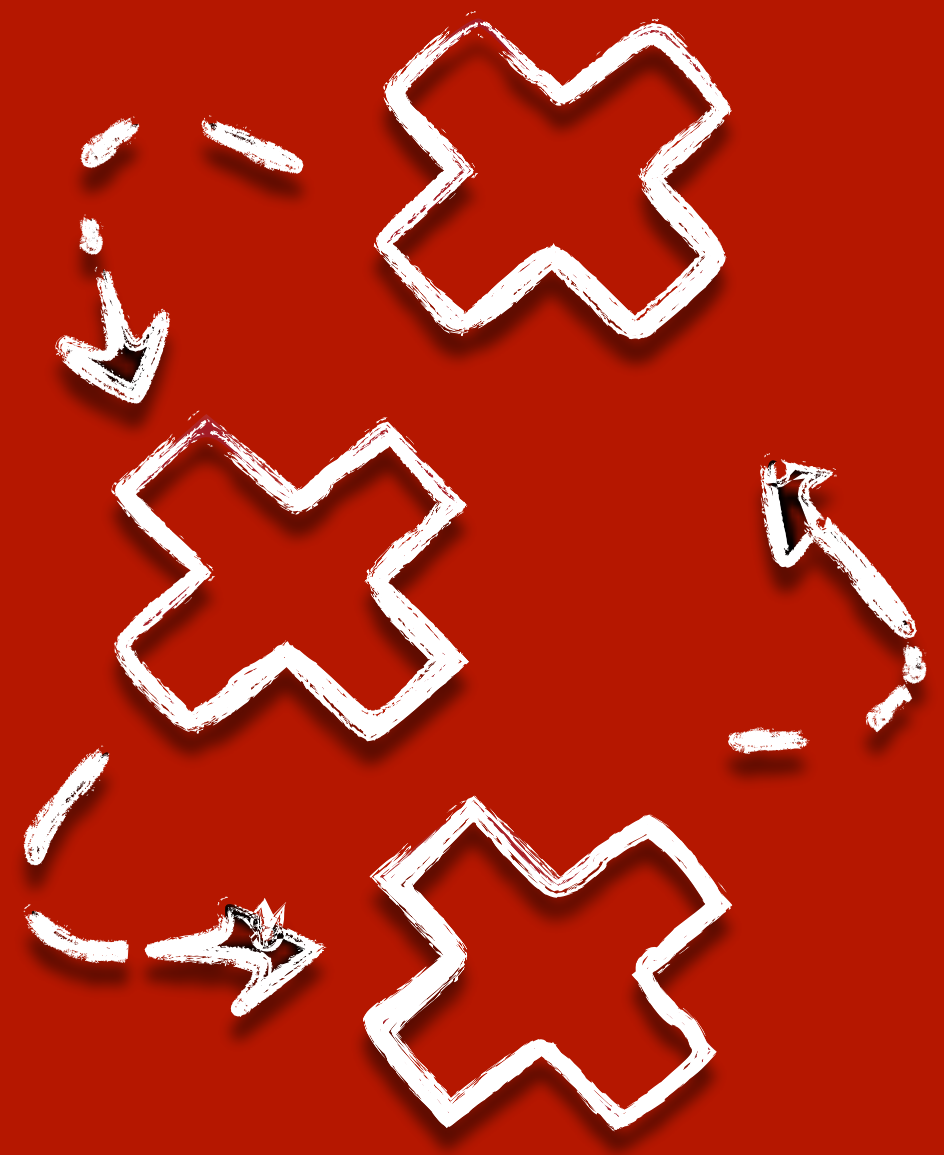


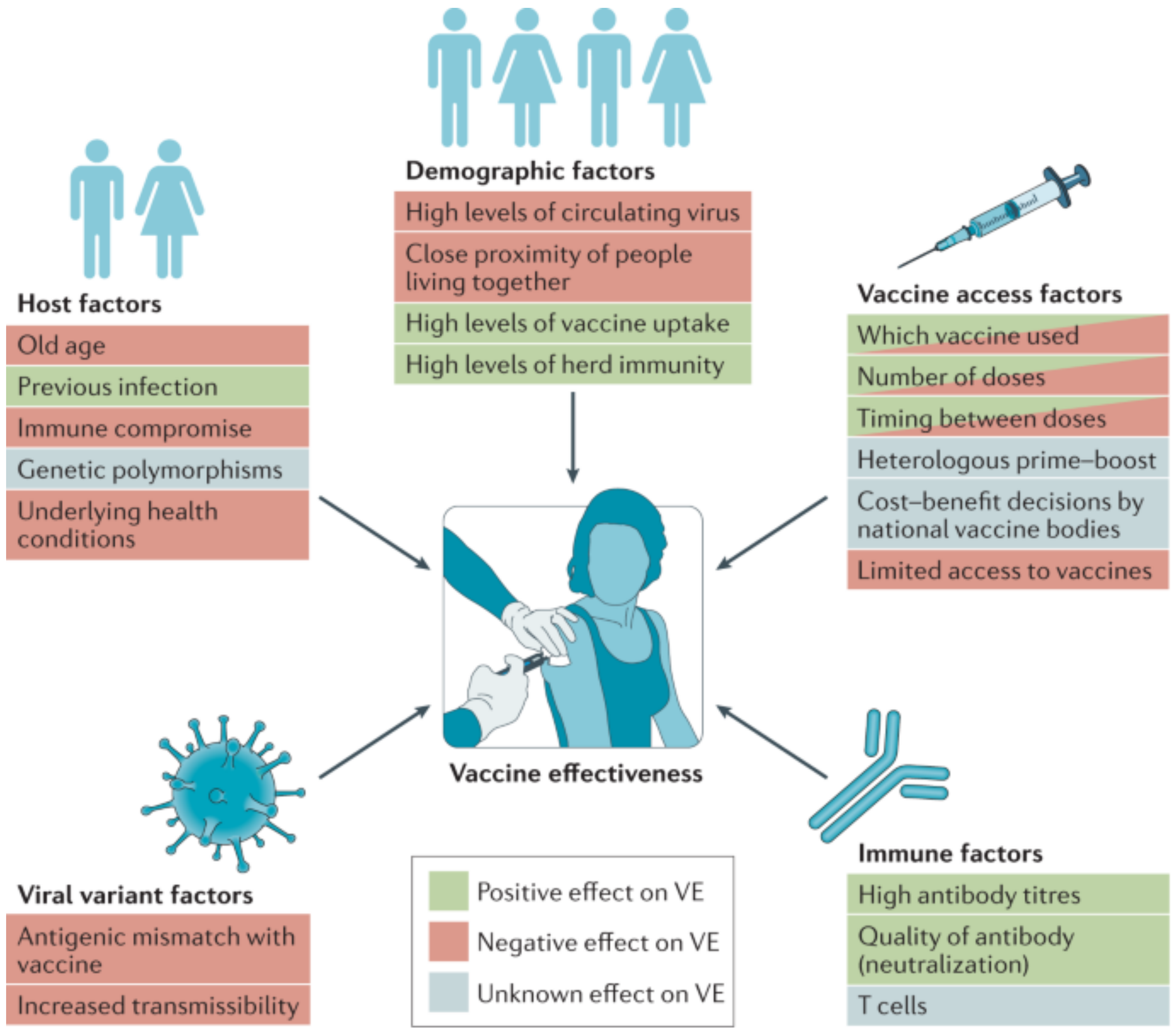
Causality-inspired ML

What can ideas from causality do for ML?

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

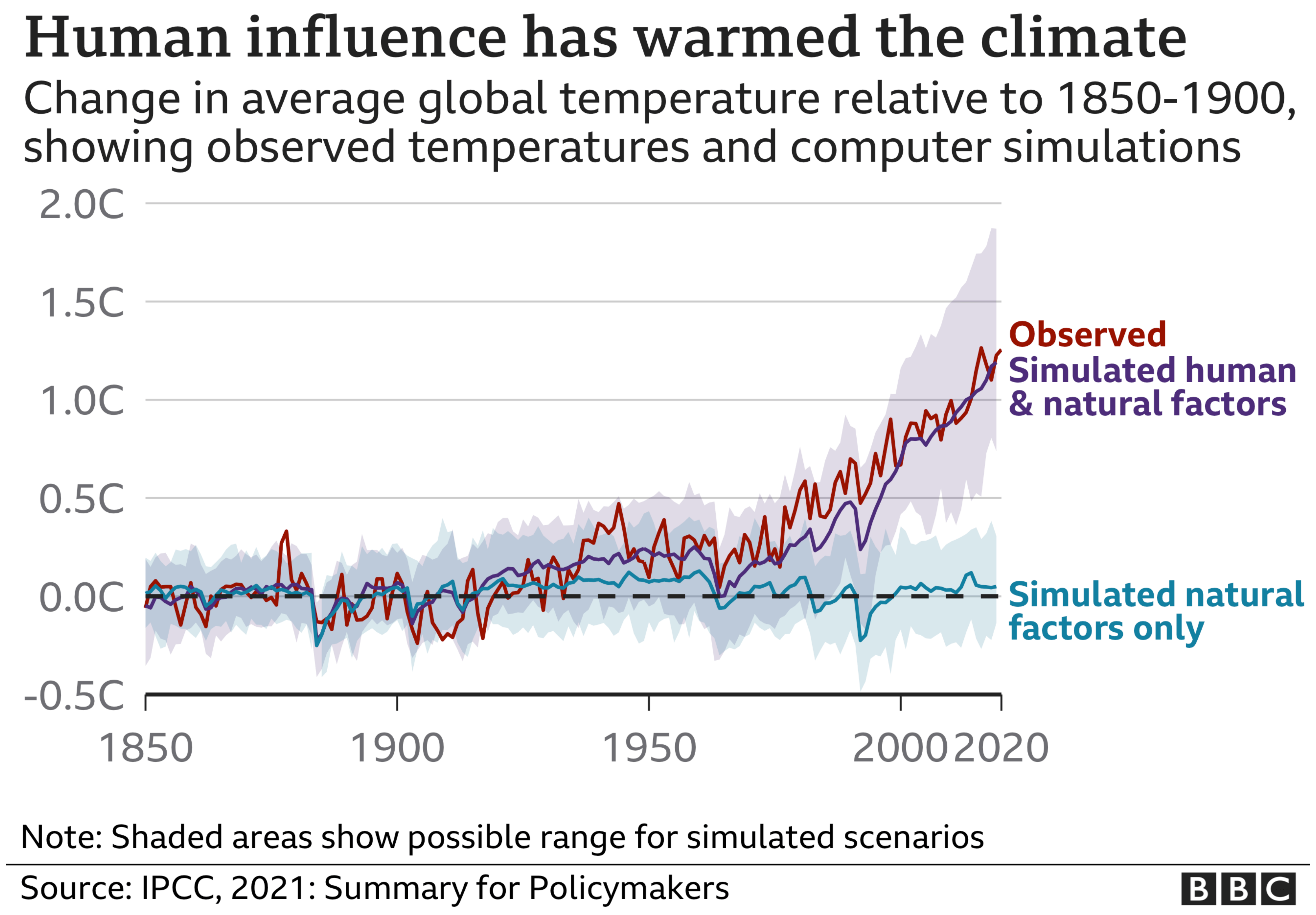
(joint work with Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, Joris Mooij, Biwei Huang, Fan Feng, Chaochao Lu and Kun Zhang)

Causal questions are ubiquitous: healthcare



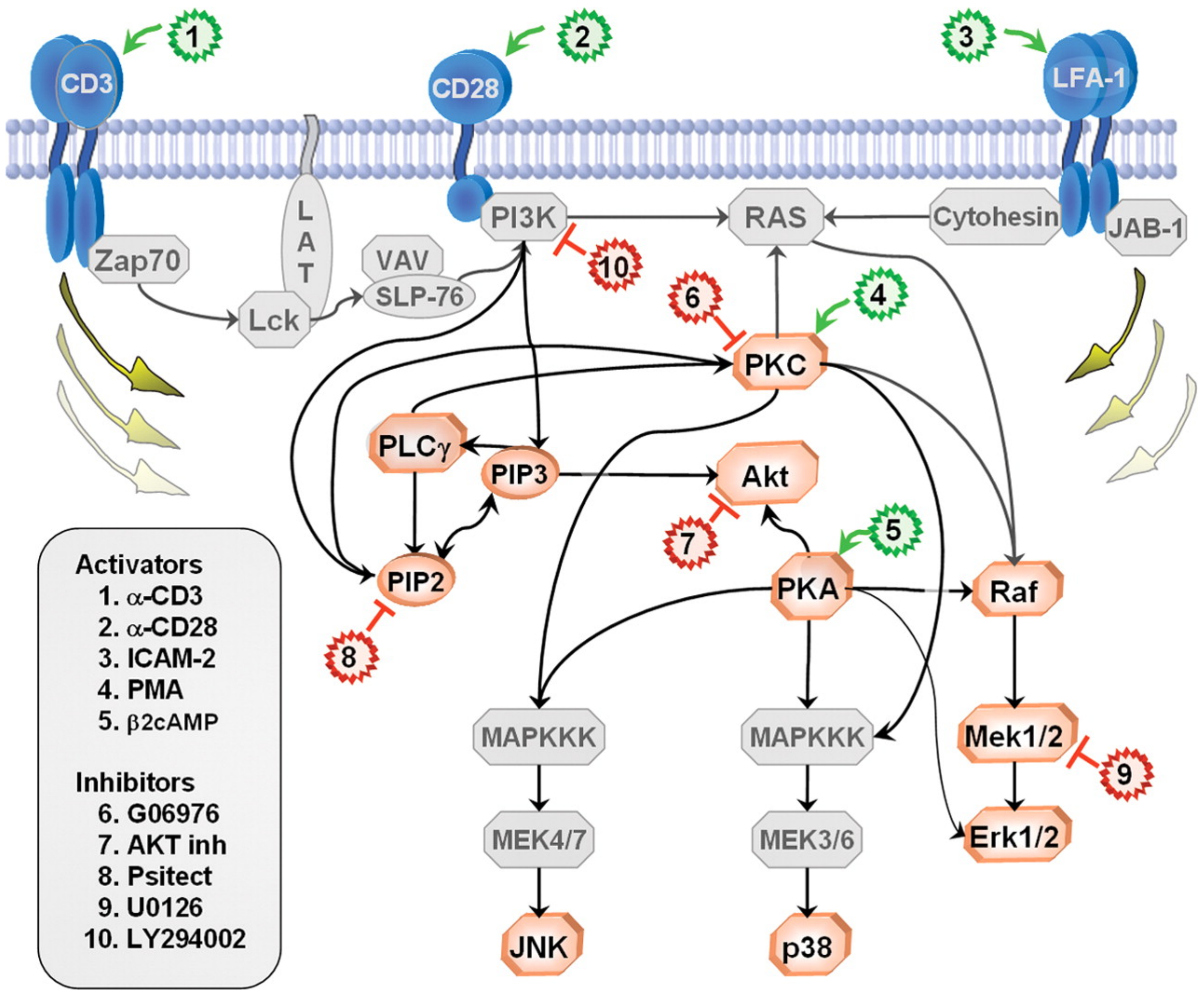
https://www.nature.com/articles/s41577-021-00592-1

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

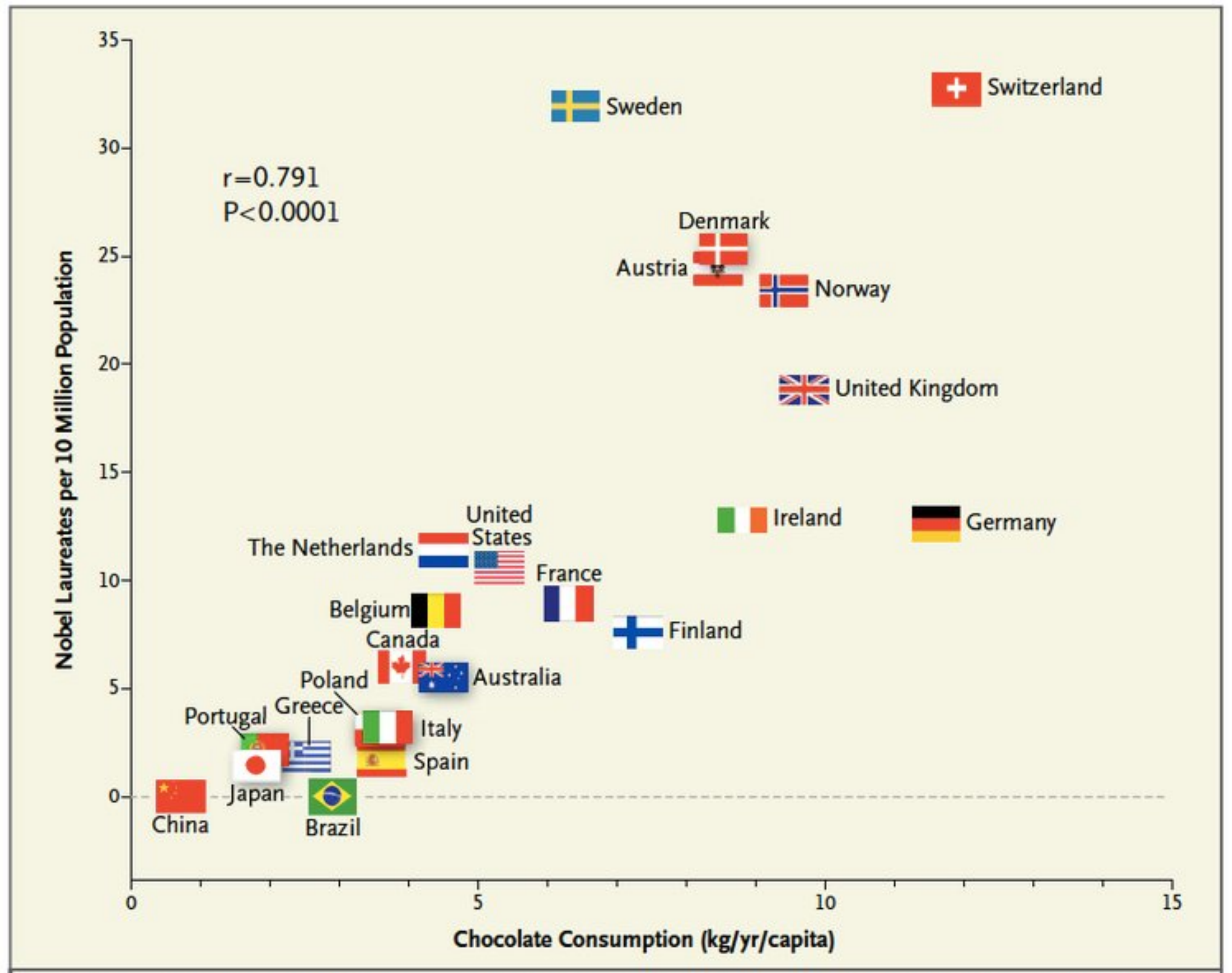
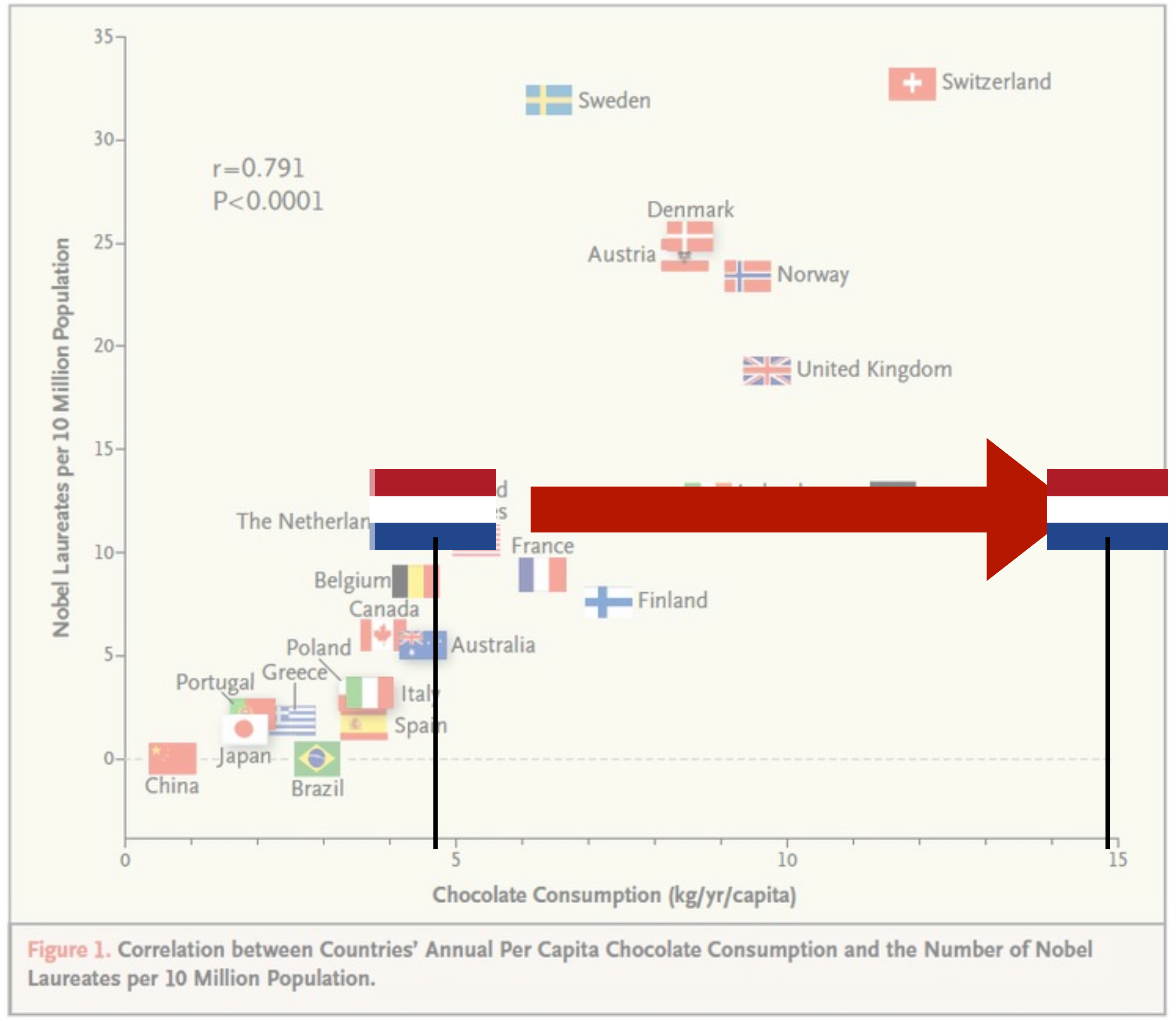


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

[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



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

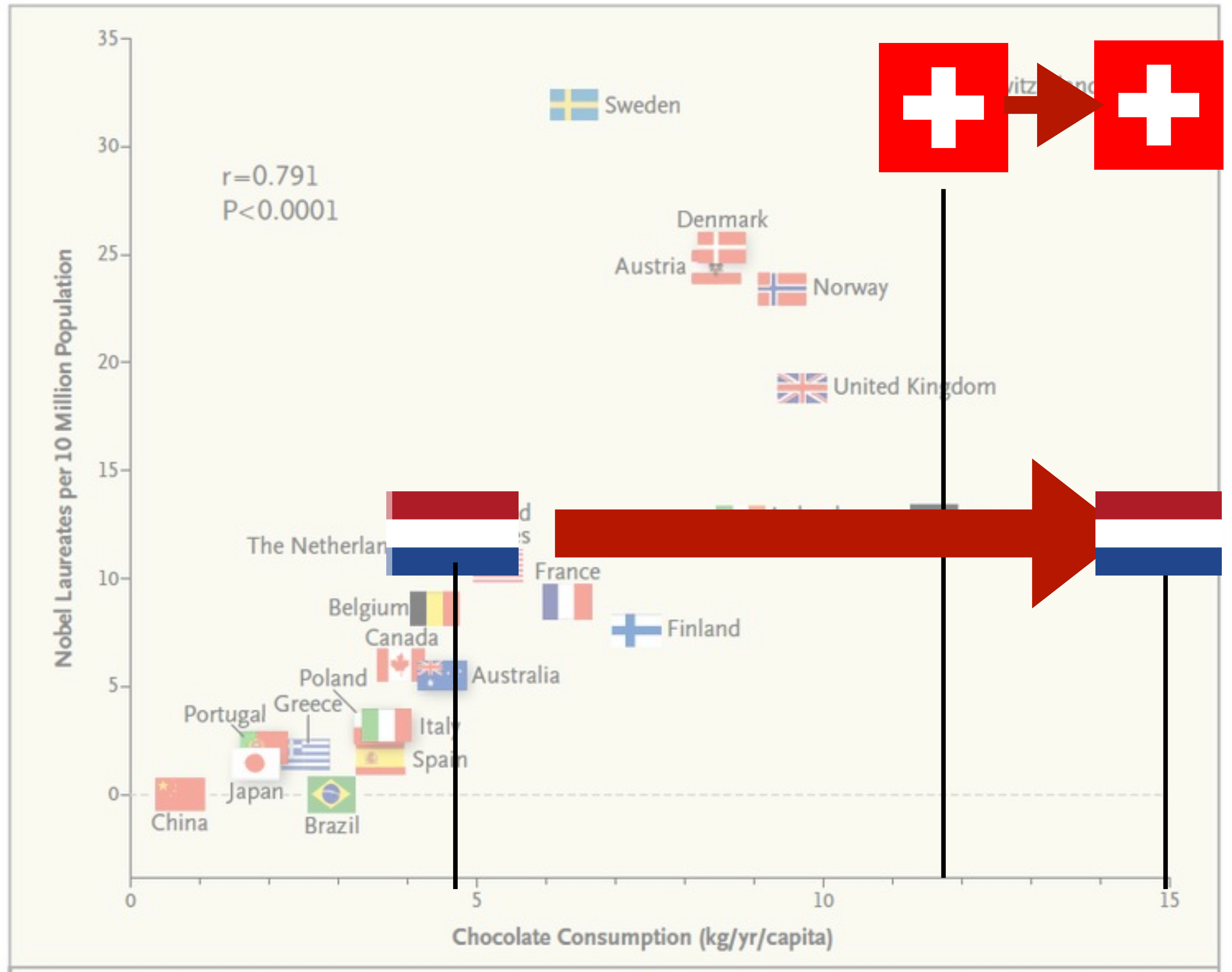


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

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

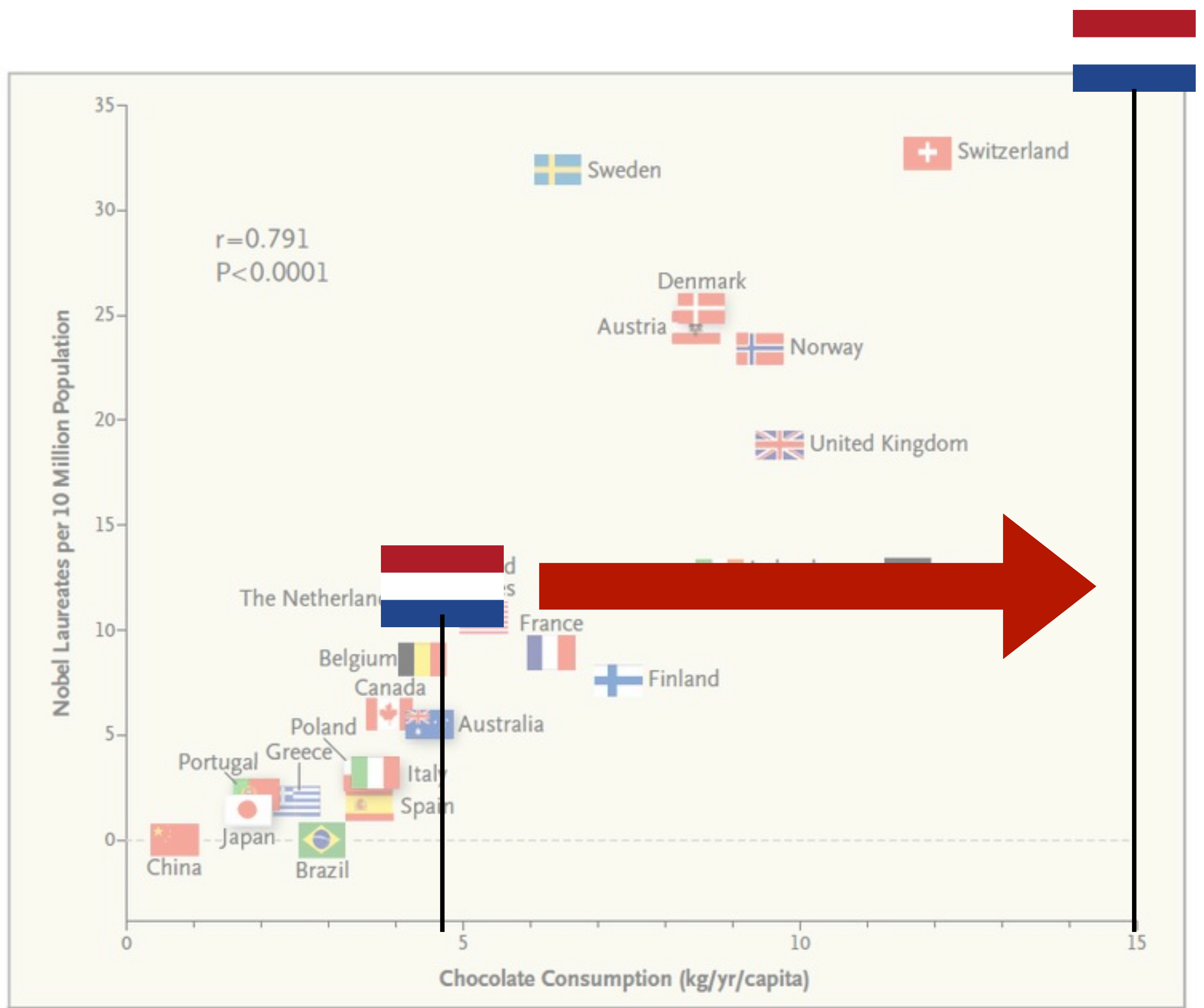


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

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

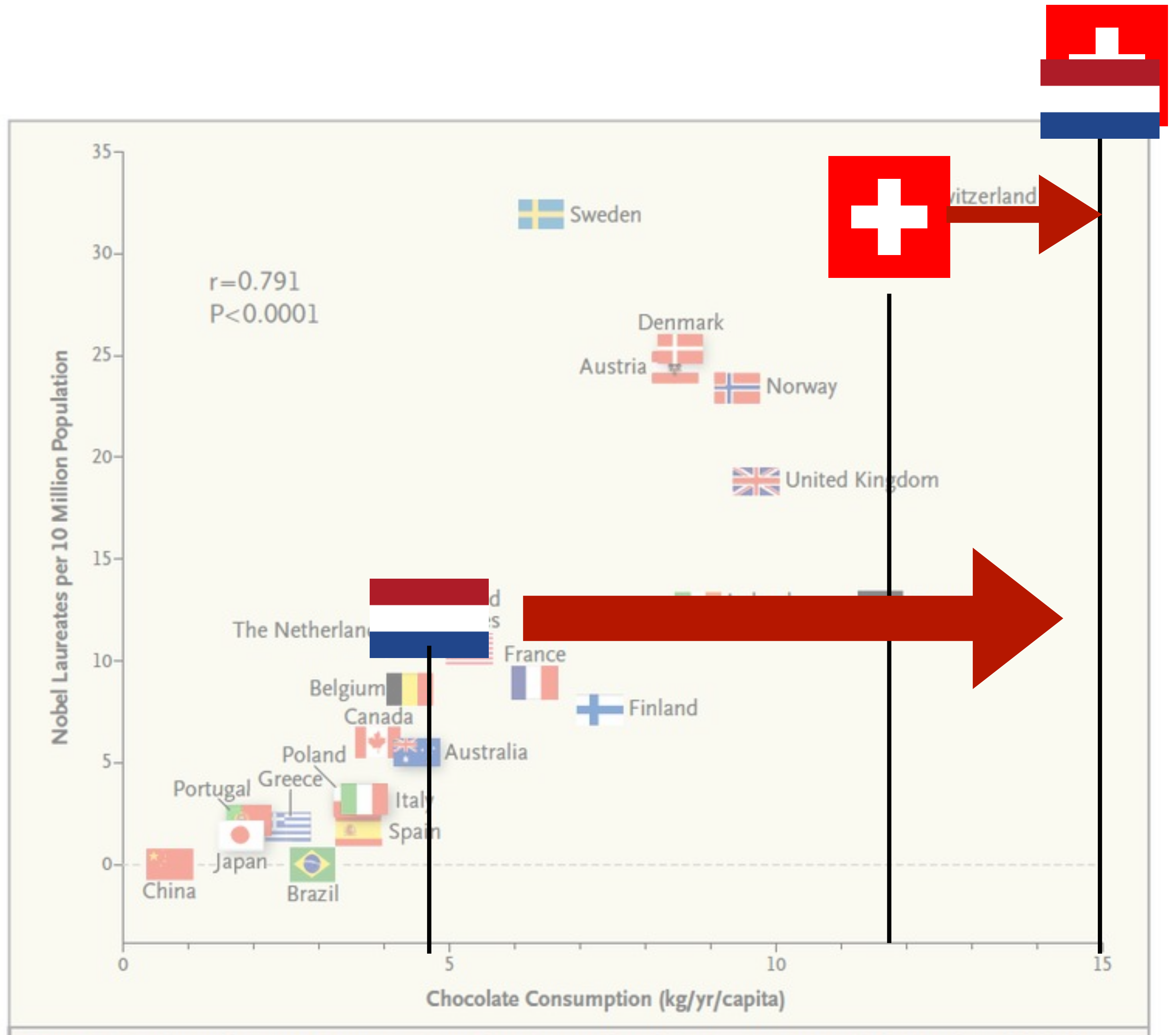


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

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

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)**

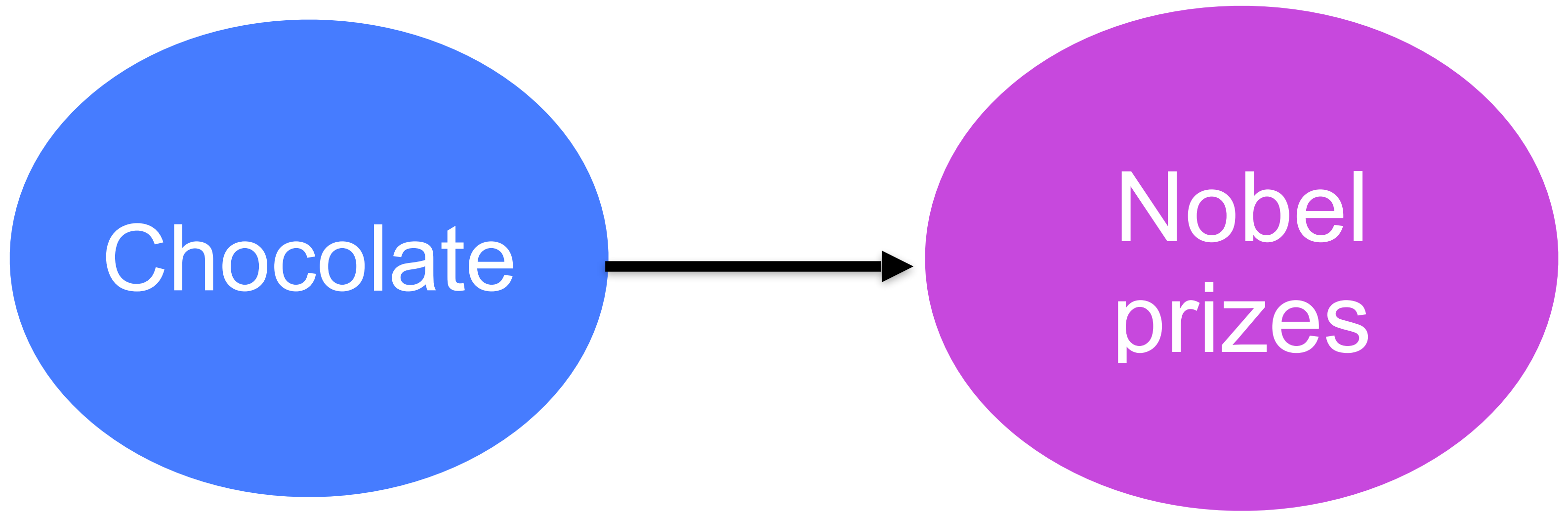
A working definition of causality in machine learning

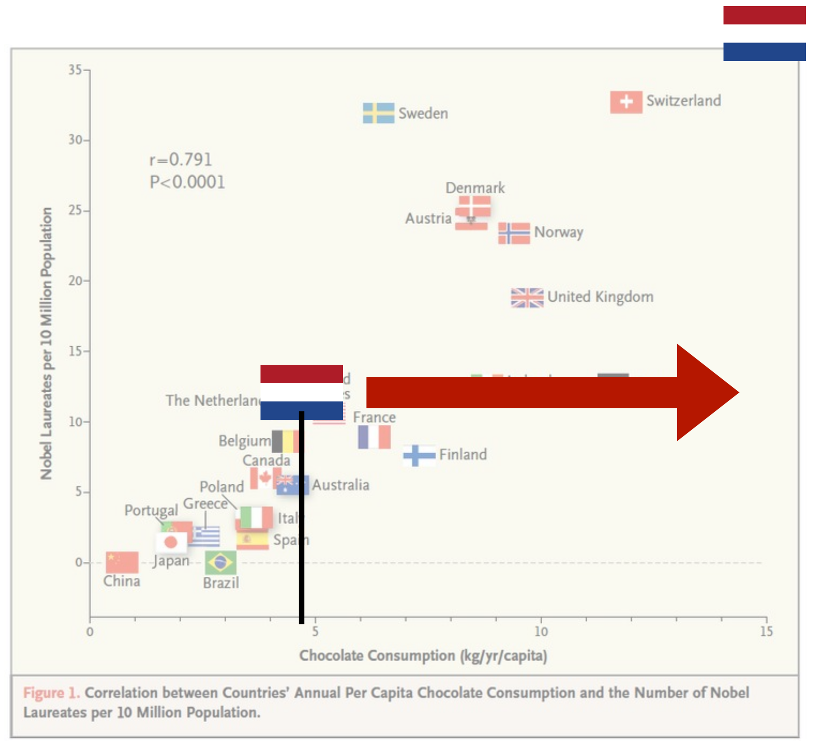
Informal definition: A variable X causes another variable Y, if changing (the distribution of) X, e.g. by fixing its value, changes (the distribution of) Y

Intervention

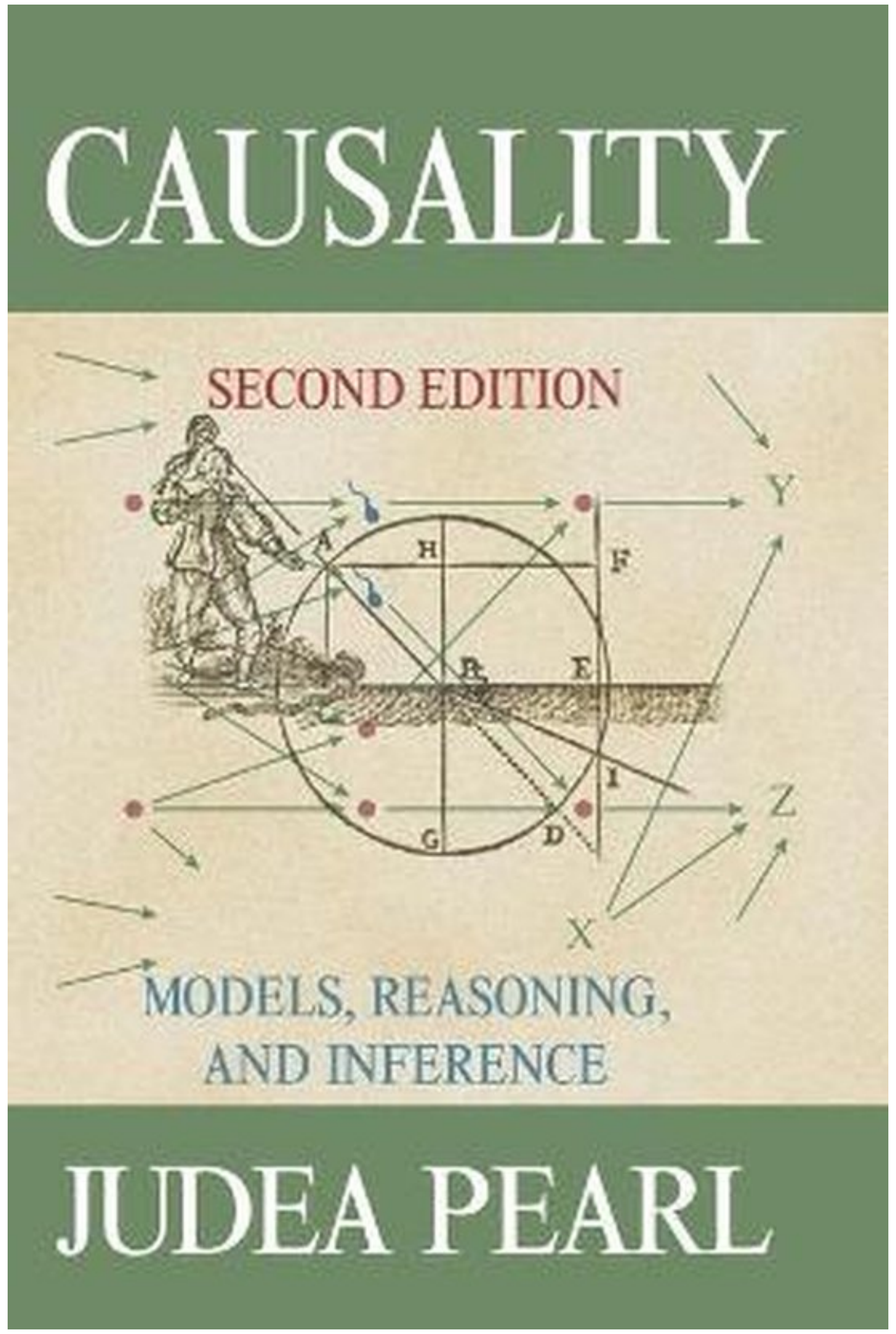
Challenge: estimate the causal effect of an intervention, when we do not have (all possible) interventional data **(e.g. observational data)**

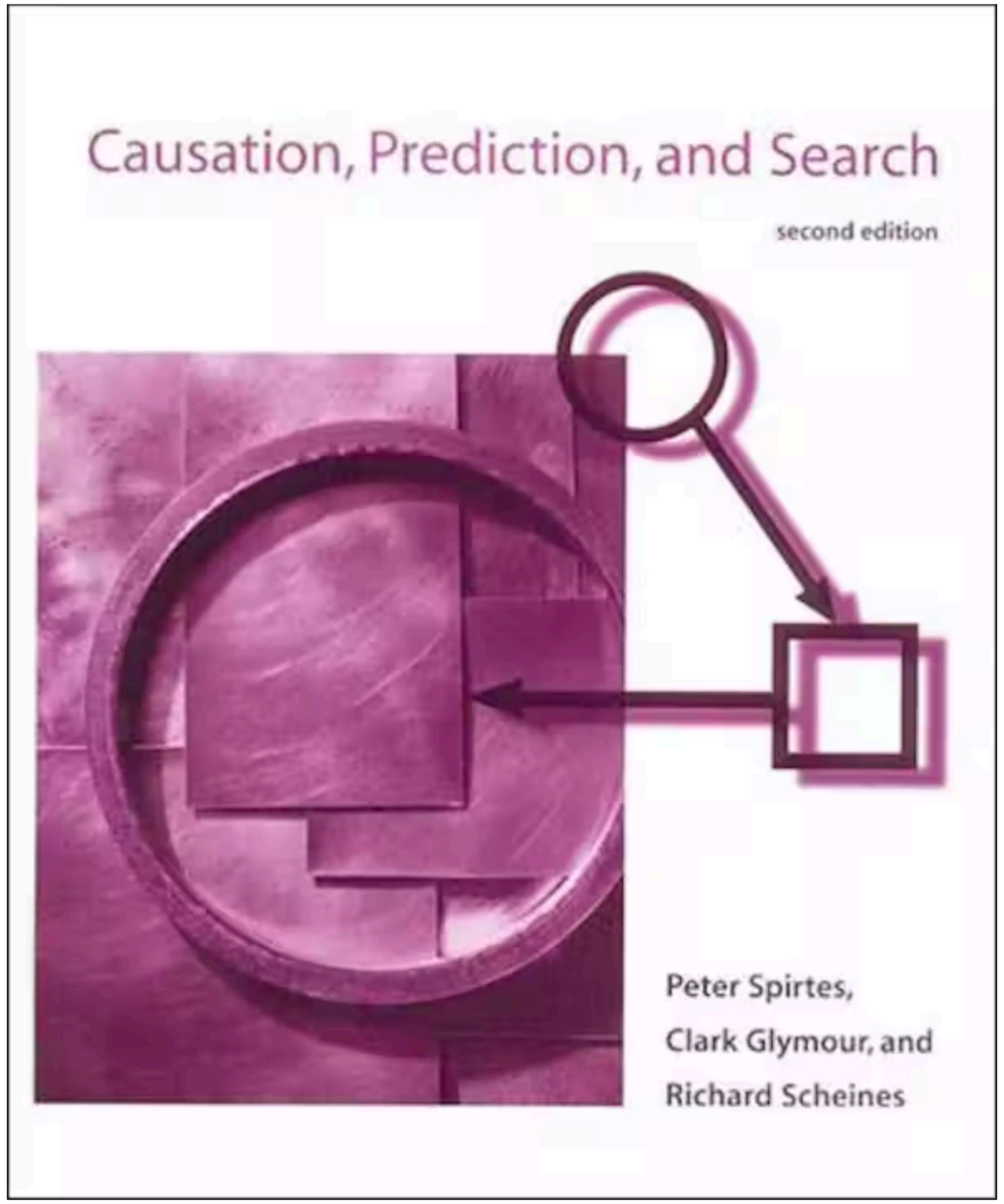
Representation: We can represent causal relations in **causal graphs:** nodes are random variables, edges causal relations

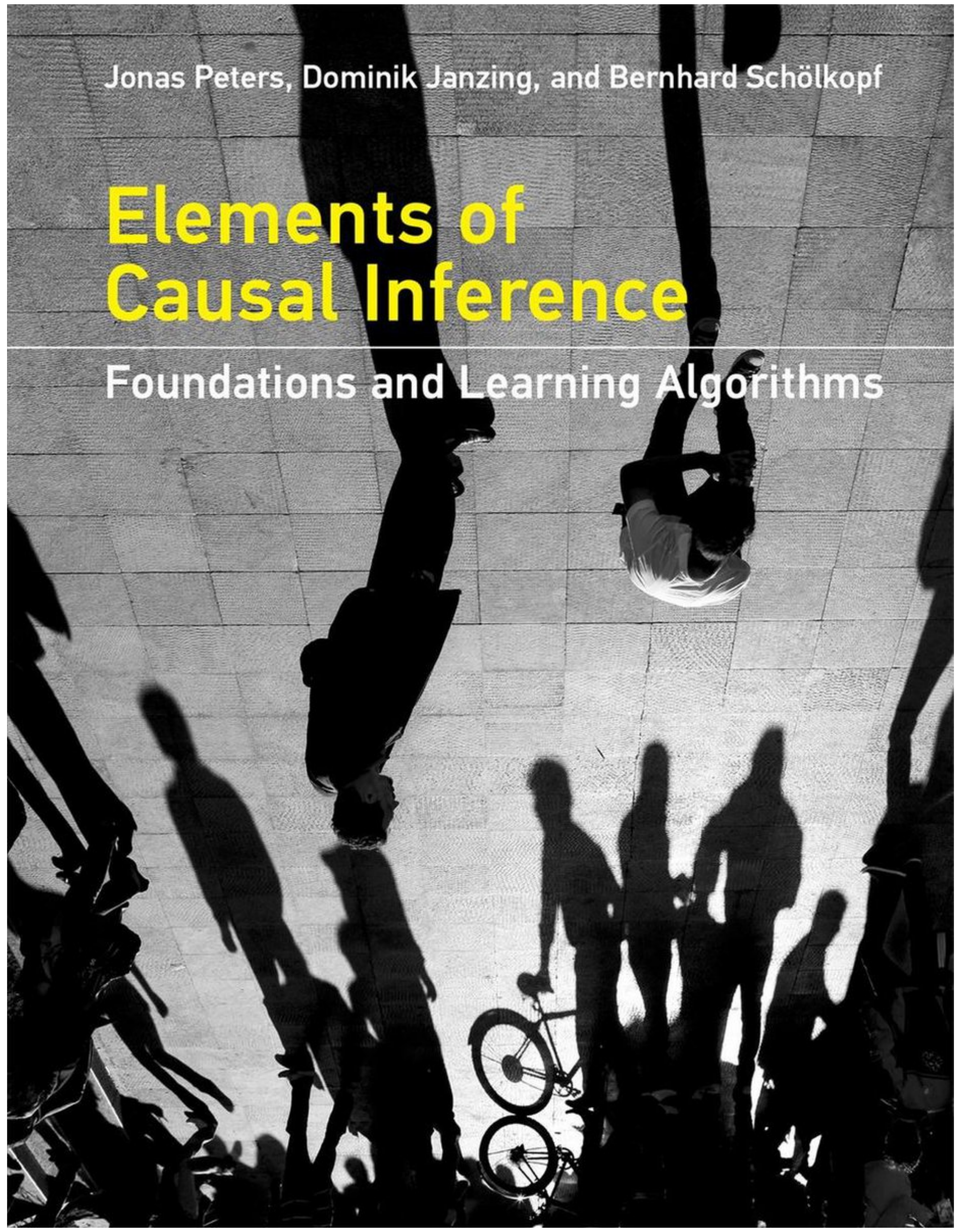




Causality in ML: foundational books (non-exhaustive)







Causality + machine learning (non-exhaustive list)

1. **Machine learning (ML) helps causality**
    - Causal discovery - learning causal graphs from data
    - Causal effect estimation - matching, weighting, double ML
    - (Causal) representation learning

2. **Causality (in the most general definition) helps machine learning**
    - Robustness, Transfer learning
    - Reinforcement Learning
    - Bias mitigation, fairness

https://arxiv.org/pdf/1705.08821.pdf, https://arxiv.org/pdf/1802.05664.pdf, https://arxiv.org/pdf/1605.03661.pdf, https://crl.causalai.net/, https://www.youtube.com/watch?v=Obuu3w809CI&ab_channel=ConnorJerzak and many many others

Causality + machine learning (non-exhaustive list)

1. Machine learning (ML) helps causality
   - Causal discovery - learning causal graphs from data
   - Causal effect estimation - matching, weighting, double ML
   - (Causal) representation learning

2. Causality (in the most general definition) helps machine learning
   - Robustness, Transfer learning
   - Reinforcement Learning
   - Bias mitigation, fairness

https://arxiv.org/pdf/1705.08821.pdf, https://arxiv.org/pdf/1802.05664.pdf, https://arxiv.org/pdf/1605.03661.pdf, https://crl.causalai.net/, https://www.youtube.com/watch?v=Obuu3w809CI&ab_channel=ConnorJerzak and many many others

Outline

1. Graphical models and d-separation [Pearl 1988] are a principled way to reason about **invariances and distribution shift**

2. Example in unsupervised domain adaptation

3. An application in fast adaptation in RL

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

Causal Hierarchy [Pearl 2009, 2018]

	Level (Symbol)	Typical Activity	Typical Questions	Examples
Most ML	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?
Causality	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? … cigarettes?
	3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection		

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*

Causal Hierarchy [Pearl 2009, 2018]

Most ML

Causality

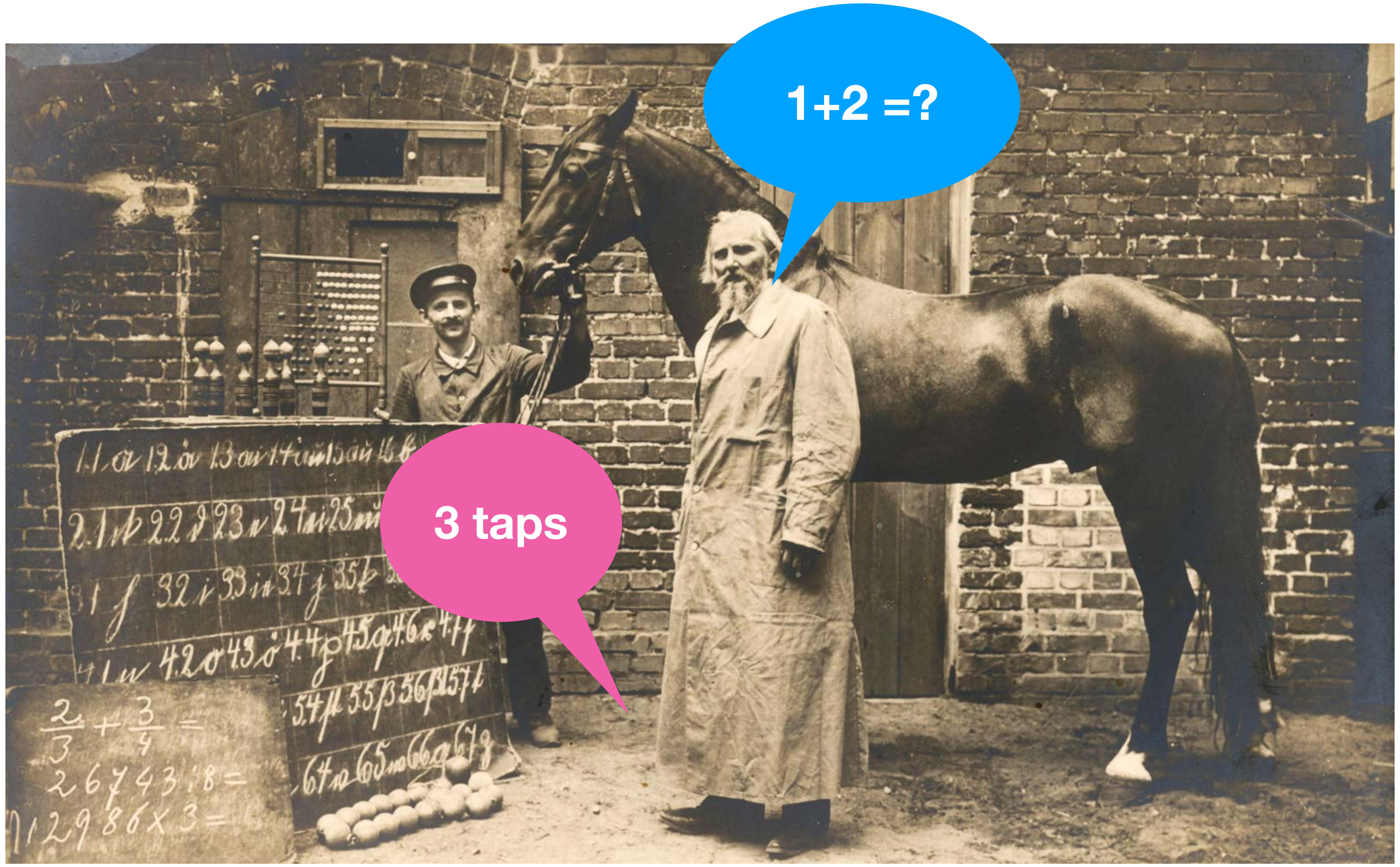
Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my be cured? rettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	W dif	that ad not been smok- ing the pa years?

In this talk: examples in domain adaptation, but lots of related work

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

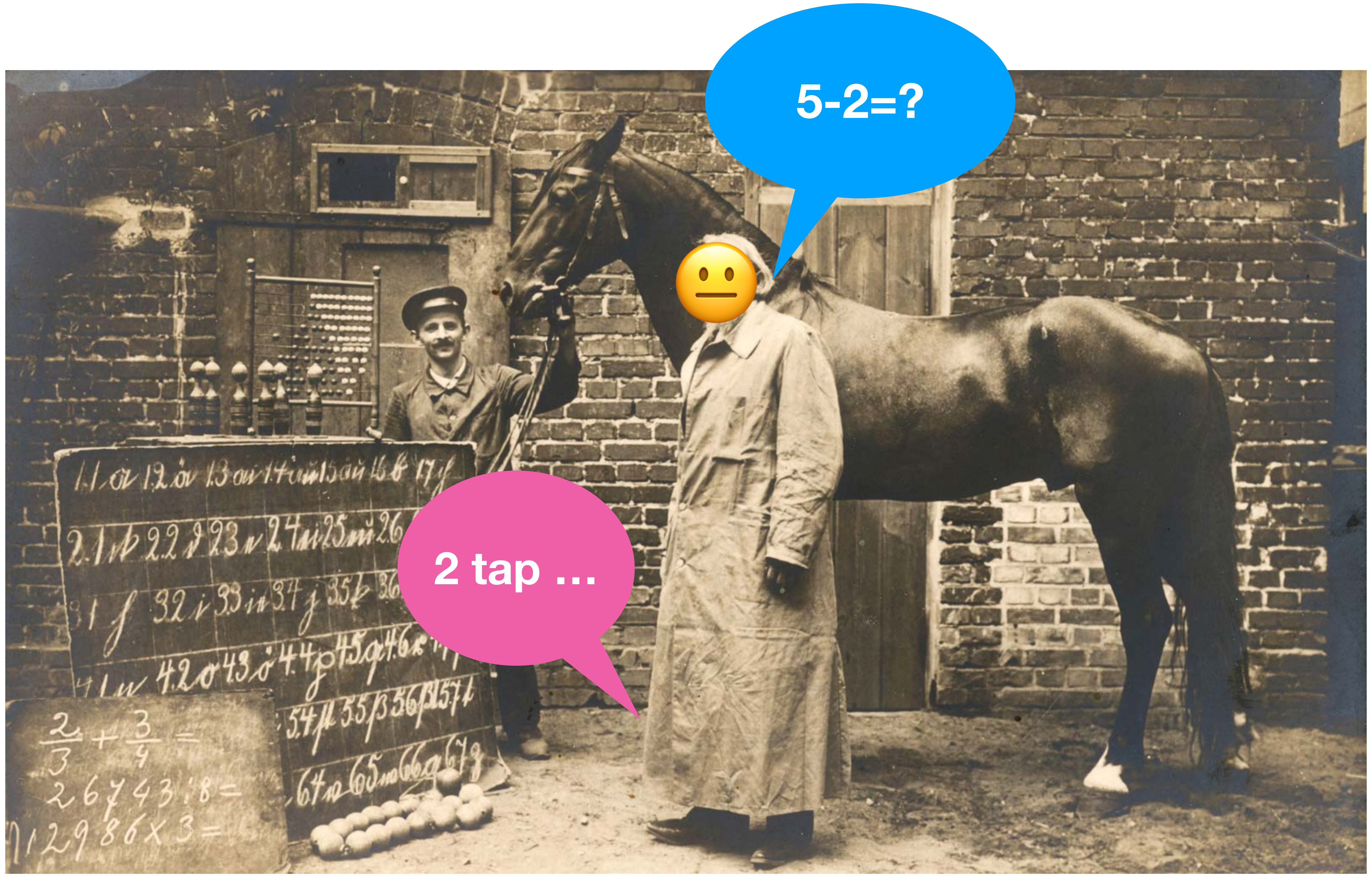
Why is it important that ML algorithms are robust to distribution shift: the "Clever Hans" effect

Why is it important that ML algorithms are robust to distribution shift: the "Clever Hans" effect



Why is it important that ML algorithms are robust to distribution shift: the “Clever Hans” effect

Why is it important that ML algorithms are robust to distribution shift: the “Clever Hans” effect



Why is it important that ML algorithms are robust to distribution shift: the “Clever Hans” effect

Why is it important that ML algorithms are robust to distribution shift: the "Clever Hans" effect

Why is it important that ML algorithms are robust to distribution shift: the "Clever Hans" effect: VQA

What color is the jacket?
-Red and blue.
-Yellow.
-Black.
-Orange.

How many cars are parked?
-Four.
-Three.
-Five.
-Six.

What event is this?
-A wedding.
-Graduation.
-A funeral.
-A picnic.

When is this scene taking place?
-Day time.
-Night time.
-Evening.
-Morning.

Only using answers!

Method	What	Where	When	Who	Why	How	Overall
LSTM (Q, I) [15]	48.9	54.4	71.3	58.1	51.3	50.3	52.1
MLP (A)	47.3	58.2	74.3	63.6	57.1	49.6	52.9

-Green.
-Brown.
-Orange.
-Red.

-Day time.
-Night time.
-Evening.
-Morning.

[Jabri et al. 2016] "Revisiting visual question answering baselines." https://arxiv.org/pdf/1606.08390.pdf

Why is it important that ML algorithms are robust to distribution shift: the "Clever Hans" effect: NLI

Example: Right for the wrong reasons

Premise: The doctor was visited by the judge.
Hypothesis: The judge visited the doctor.

Entailment

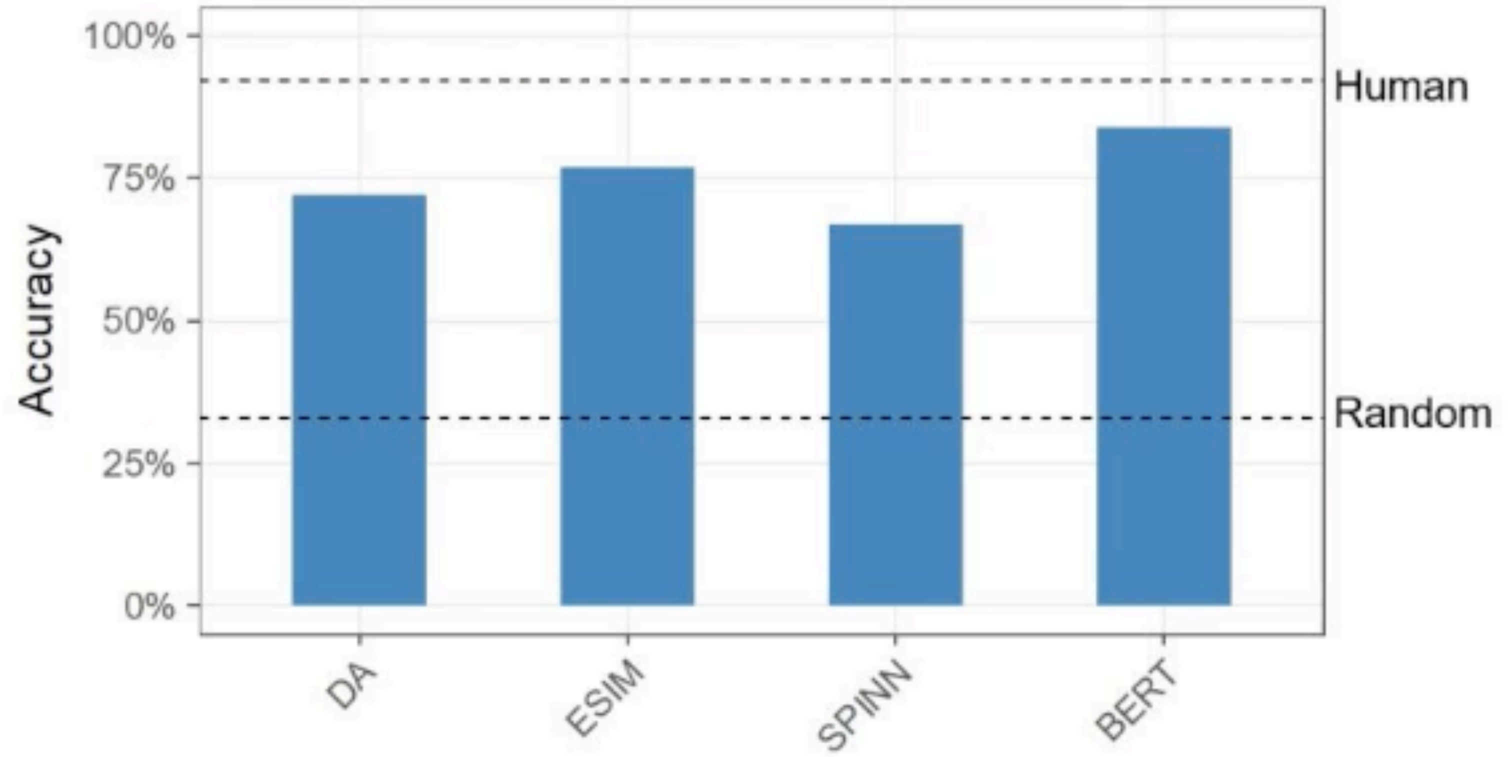
Possible heuristic:
High lexical overlap means "entailment"

Is the model using proper inference or this heuristic?
Test with an example where the heuristic fails:

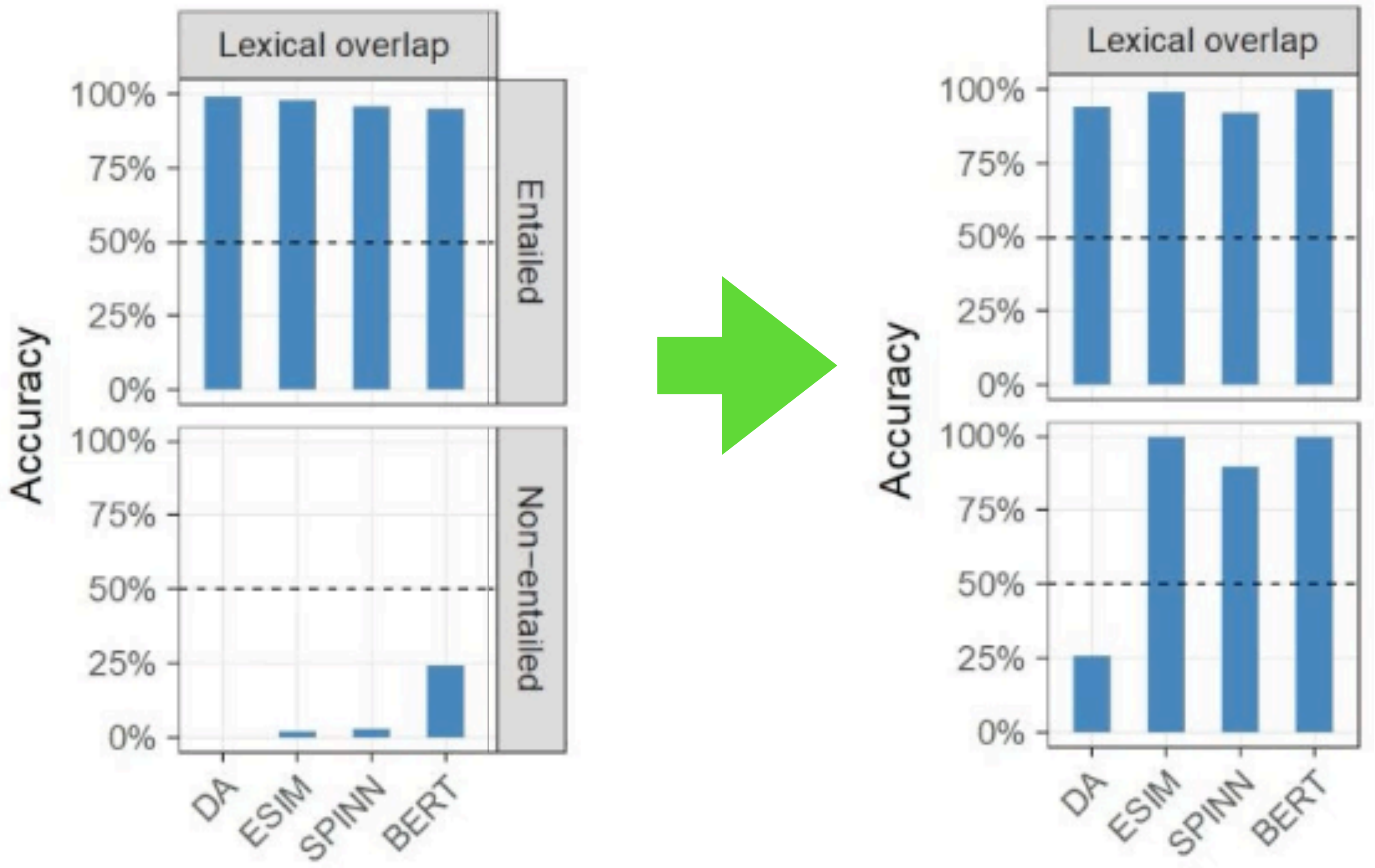
Premise: The doctor was visited by the judge.
Hypothesis: The doctor visited the judge.

Non-entailment

MNLI
(standard)



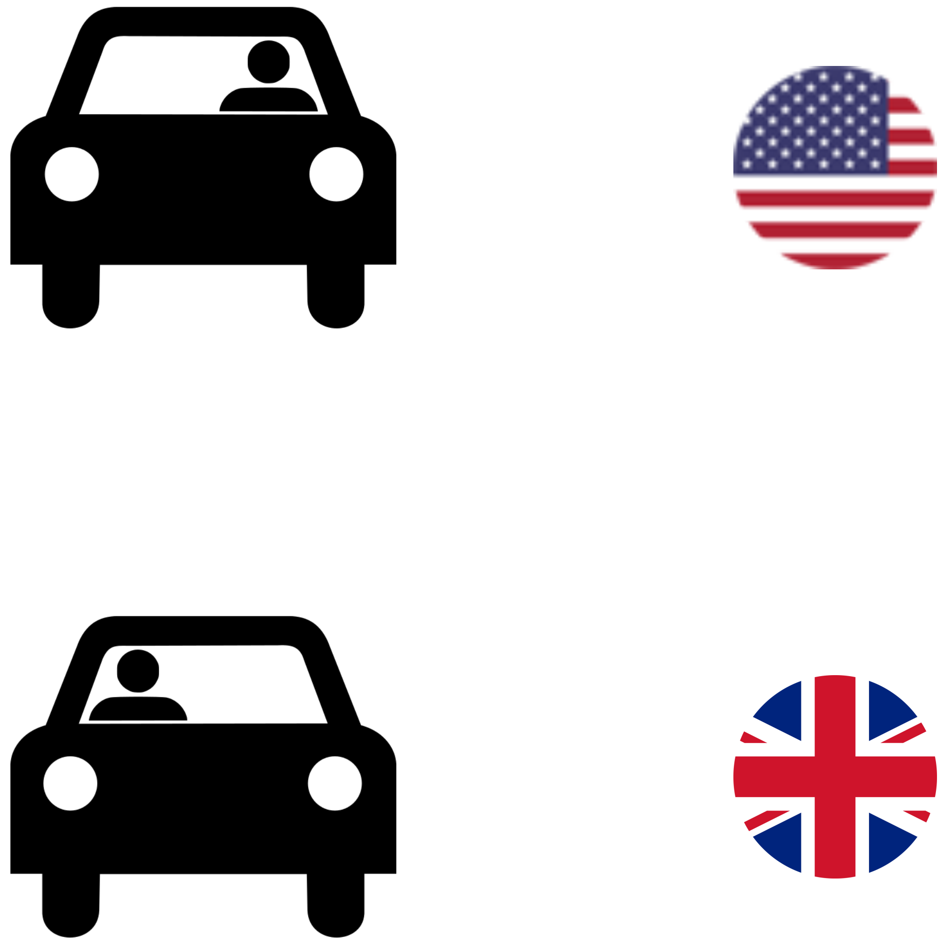
HANS
(balanced)



Edited from slides by Tom McCoy, [McCoy et al, 2019] Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference, https://aclanthology.org/P19-1334.pdf

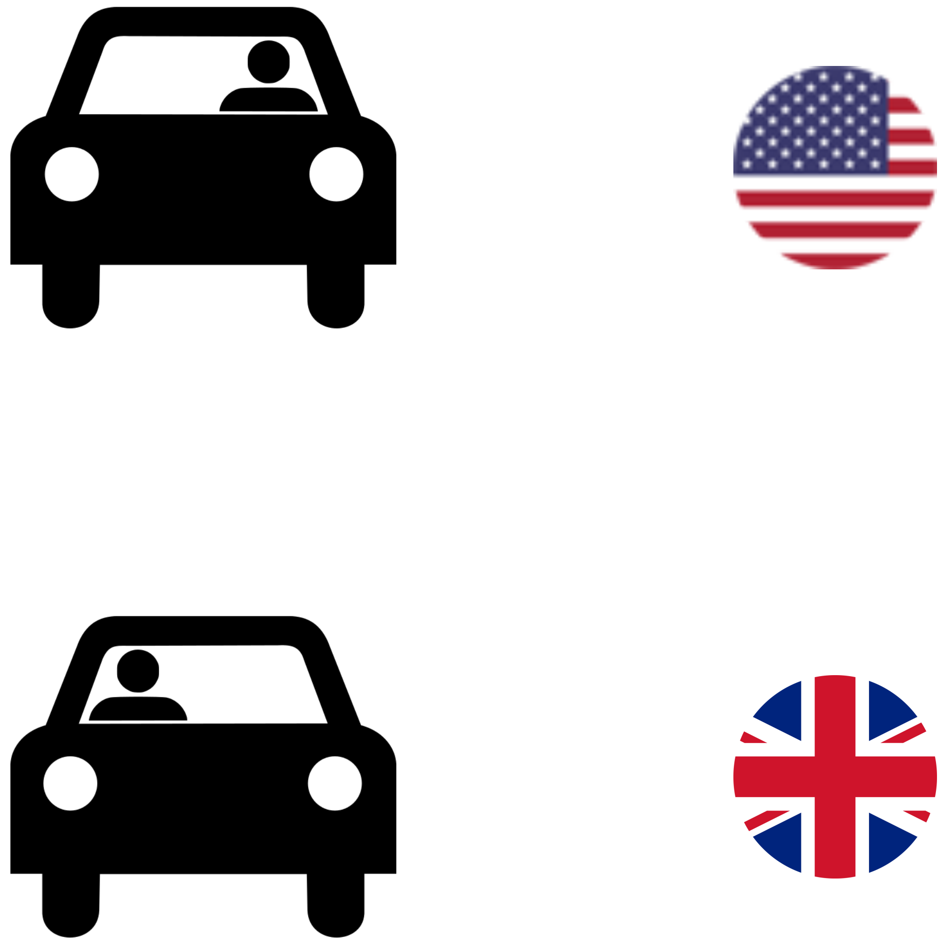
Causality vs Transfer learning

- 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?

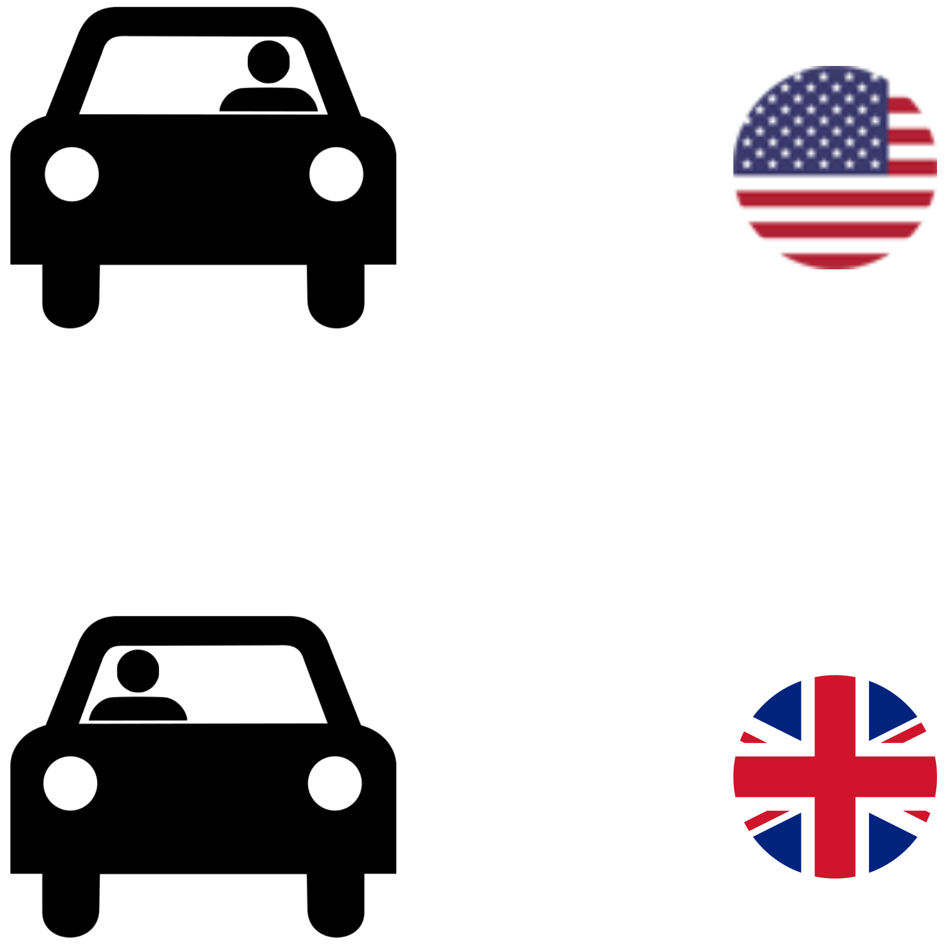


- Causal inference:
    - How can I predict what happens when the distribution changes after an intervention?
        - Perfect intervention do(X):
            - do-calculus [Pearl, 2009]
        - **Soft intervention on X** ≈ change of distribution of P(X| parents)

Causality vs Transfer learning

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

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



…t intervention do(X):

- do-calculus [Pearl, 2009]

- **Soft intervention on X** $\approx$ change of distribution of P(X| parents)

Causality vs Transfer learning

Not a new idea!

On Causal and Anticausal Learning

ICML 2012

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

Abstract

We consider the problem of function estimation in the case where an underlying causal model can be inferred. This has implications for popular scenarios such as covariate shift, concept drift, transfer learning and semi-supervised learning. We argue that causal knowledge may facilitate some approaches for a given problem, and rule out others. In particular, we formulate a hypothesis for when semi-supervised learning can help, and corroborate it with empirical results.

for causal inference in the machine learning community.

An example illustrating the difference between the statistical and the causal point of view is the correlation between the frequency of storks and the human birth rate (Matthews, 2000). We may be able to train a good predictor of the birth rate which uses the frequency of storks (along with other features) as an input. However, if politicians asked us whether one could boost the birth rate by increasing the number of storks, we would have to tell them that this kind of *intervention* is not covered by the standard i.i.d. assumption of statistical learning. In practice, however, interventions can be relevant, distributions may shift over time, and we might want to combine data recorded under different

Causality allows us to reason systematically about distribution shifts

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

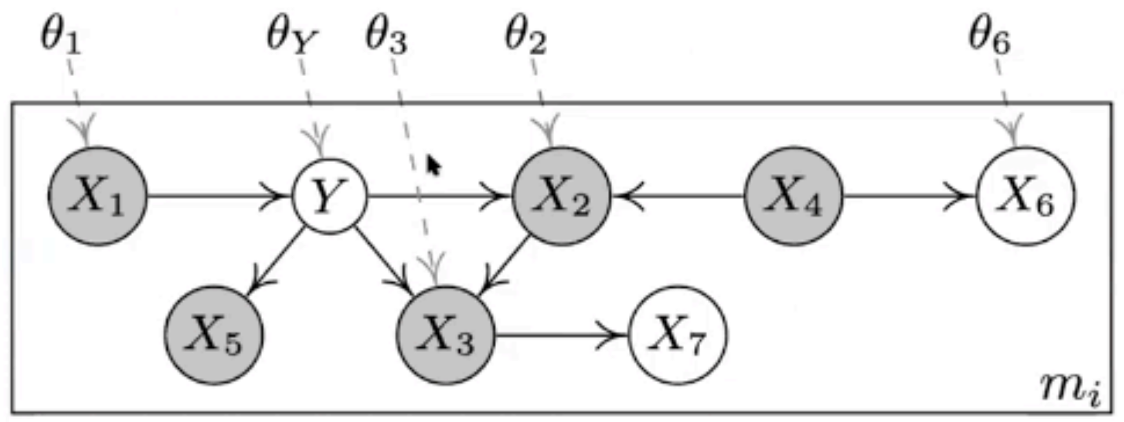
Cheng Zhang *
Microsoft Research
Cheng.Zhang@microsoft.com
Kun Zhang
Carnegie Mellon University
kunz1@cmu.edu
Yingzhen Li *
Microsoft Research
Yingzhen.Li@microsoft.com

and many many more….

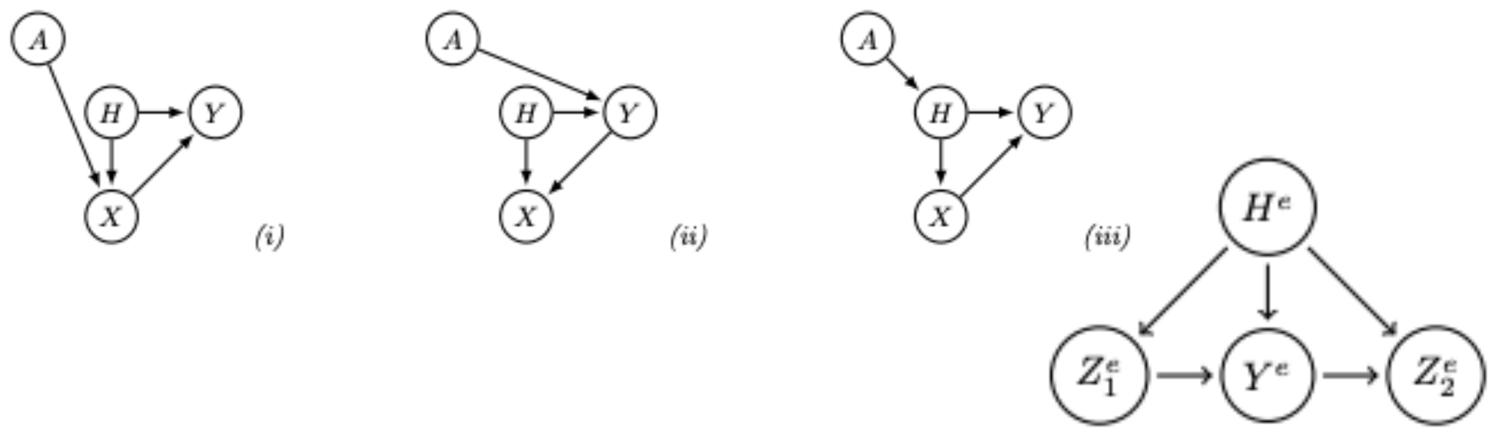
Causality allows us to reason **systematically** about distribution shifts, e.g. through **graphs**

On Causal and Anticausal Learning

Domain Adaptation as a Problem of Inference on Graphical Models

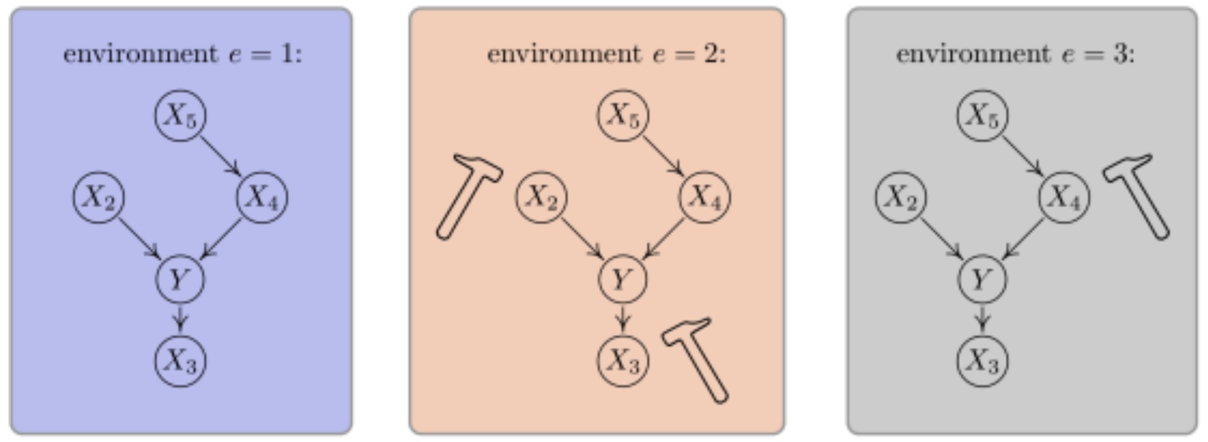


Anchor regression: heterogeneous data meet causality

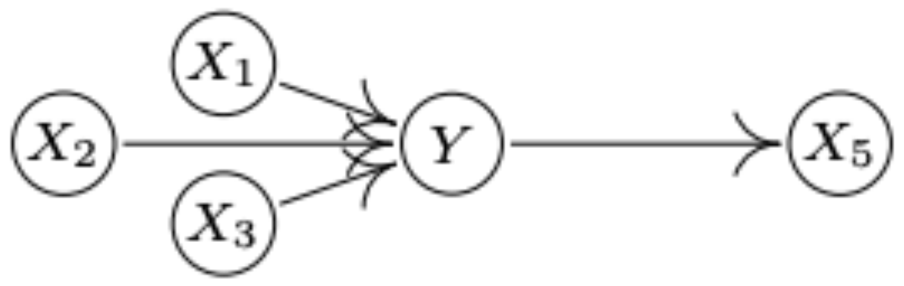


J. R. Statist. Soc. B (2016)
78, Part 5, pp. 947–1012

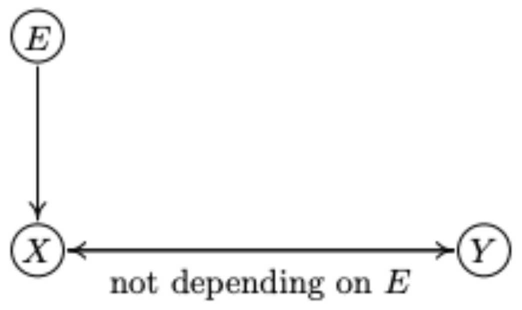
Causal inference by using invariant prediction: identification and confidence intervals



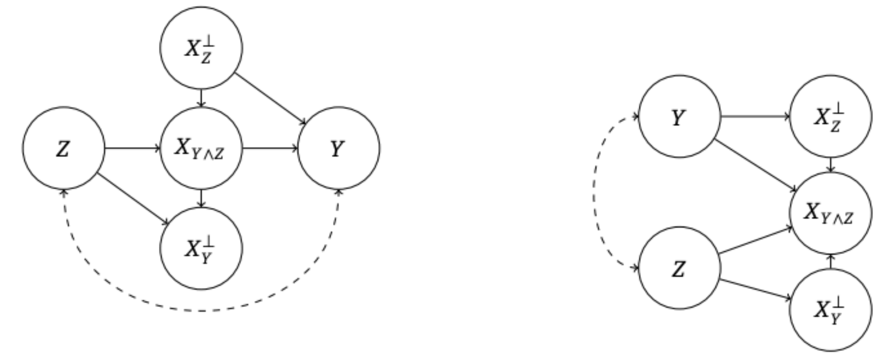
Invariant Models for Causal Transfer Learning



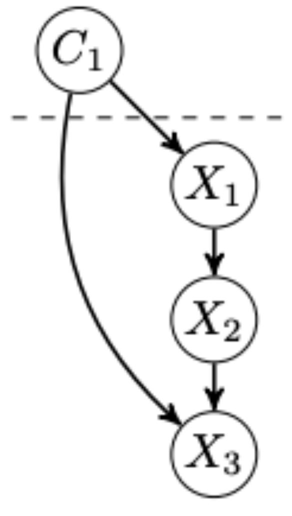
Invariance, Causality and Robustness



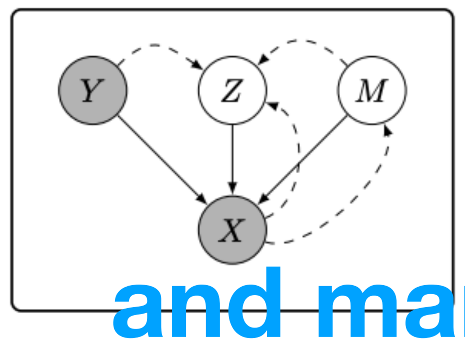
Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests



Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions



A Causal View on Robustness of Neural Networks



and many many more....

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

Even if unknown

On Causal and Anticausal Learning

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

Kun Zhang, 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

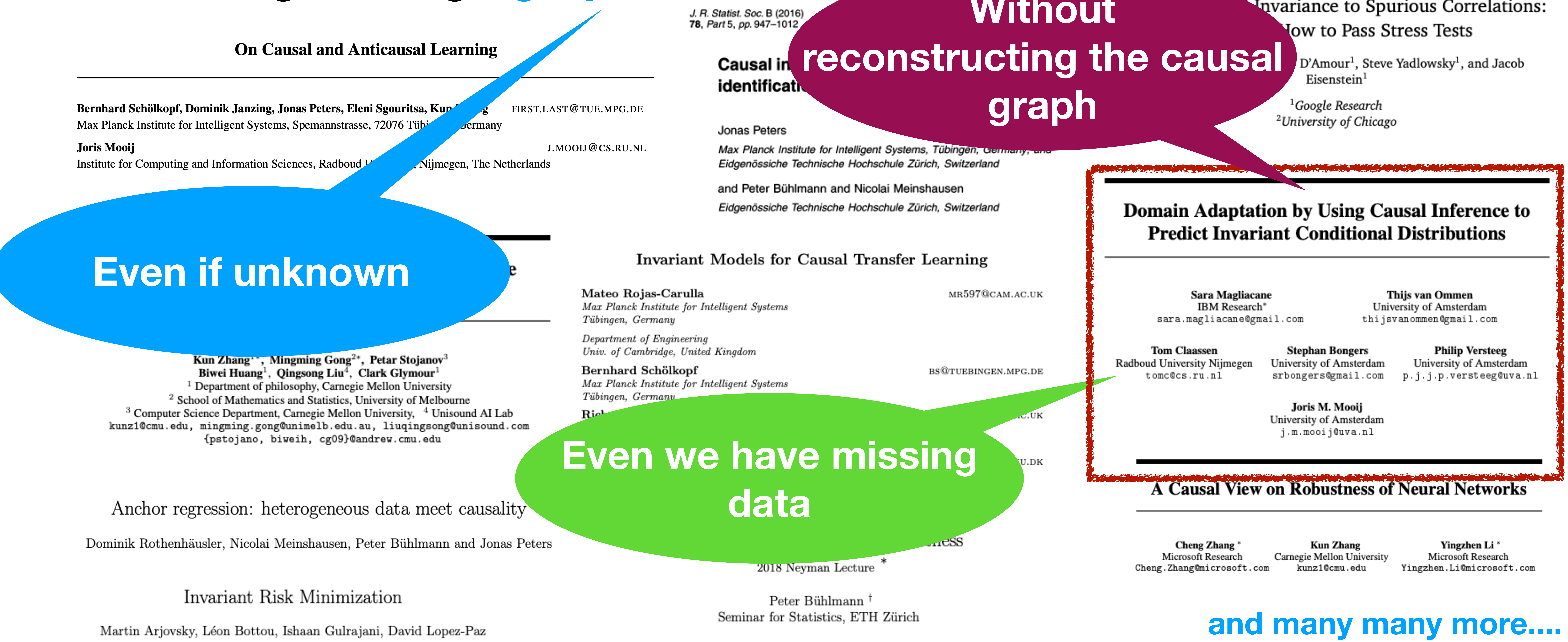
Sara Magliacane, IBM Research*, sara.magliacane@gmail.com
Thijs van Ommen, University of Amsterdam, thijsvanommen@gmail.com
Tom Claassen, Radboud University Nijmegen, tomc@cs.ru.nl
Stephan Bongers, University of Amsterdam, srbongers@gmail.com
Philip Versteeg, University of Amsterdam, p.j.j.p.versteeg@uva.nl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

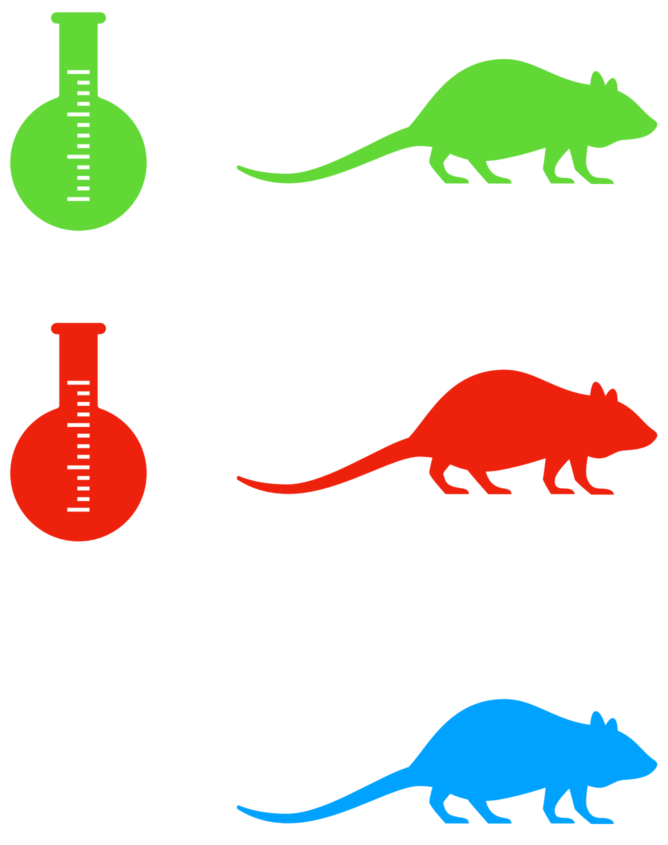
Causality allows us to reason systematically about distribution shifts, e.g. through graphs



and many many more....

A description of domain adaptation tasks:

- Supervised multi-source domain adaptation



X1	X2	X3	X4	Y
1200	1000	1500	9	**-0.1**
1201	800	1500	8	**?**
1195	200	1499	7	**?**
....	....	....	....	....
2000	600	3000	7	-0,21
2190	450	3000	8	-0,16
2000	200	2999	8	-0,16
....	....	....	....	....
1200	1000	1500	9	-0,17
1201	800	1500	10	-0,14
1195	200	1499	10	-0,07
1340	900	1498	....	-0,14

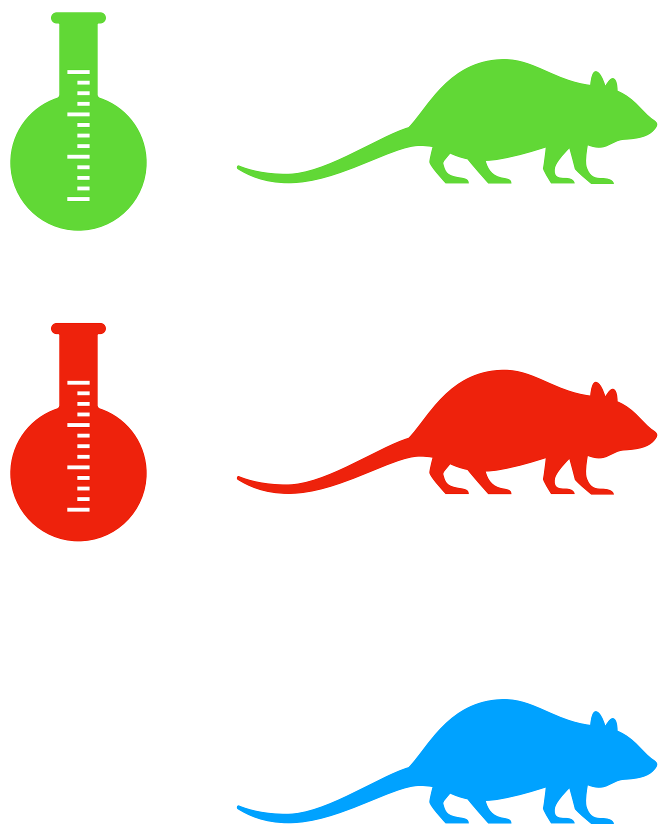
Target domain

Source domains

- Estimate $\hat{f}$ in Y = $\hat{f}$(X1, X2, X3, X4) from source domains and few labels in target domain

A description of domain adaptation tasks:

- **Unsupervised** multi-source domain adaptation



X1	X2	X3	X4	Y
1200	1000	1500	9	?
1201	800	1500	8	?
1195	200	1499	7	?
....	....	....	....	....
2000	600	3000	7	-0,21
2190	450	3000	8	-0,16
2000	200	2999	8	-0,16
....	....	....	....	....
1200	1000	1500	9	-0,17
1201	800	1500	10	-0,14
1195	200	1499	10	-0,07
1340	900	1498	....	-0,14

No labels in target

Target domain

Source domains

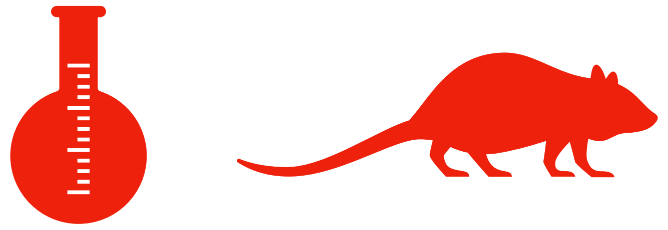
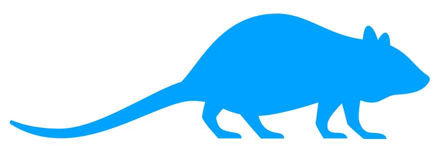
- Estimate $\hat{f}$ in Y = $\hat{f}$(X1, X2, X3, X4) from source domains and by exploiting the knowledge of the **change** from the **unlabelled data in target**

A description of domain adaptation tasks:

- **Domain generalisation:** required to work under **any intervention**

No data in target

X1	X2	X3	X4	Y
?	?	?	?	?
?	?	?	?	?
?	?	?	?	?
....	....	....	....	....
2000	600	3000	7	-0,21
2190	450	3000	8	-0,16
2000	200	2999	8	-0,16
....	....	....	....	....
1200	1000	1500	9	-0,17
1201	800	1500	10	-0,14
1195	200	1499	10	-0,07
1340	900	1498	....	-0,14

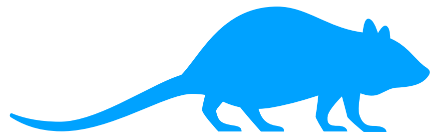



Target domain

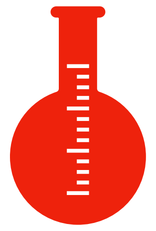
Source domains

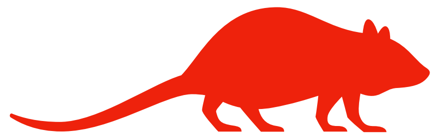
- Estimate $\hat{f}$ in Y = $\hat{f}$(X1, X2, X3, X4) from source domains, no idea about what happens in the target

Domain adaptation from a graphical perspective



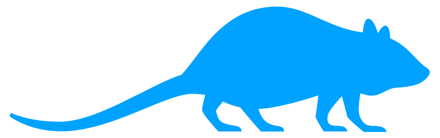
	X1	X2	Y
Normal	0,1	2	0
Normal	0,2	3	0
Normal	1,1	2	1
Normal	0,1	3	0



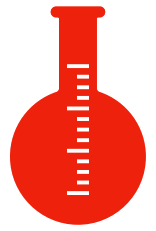


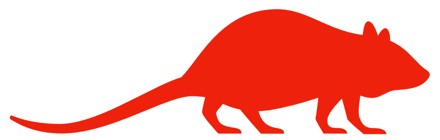
	X1	X2	Y
Gene A	3,1	2	?
Gene A	3,2	3	?
Gene A	4	2	?
Gene A	3,2	3	?

Domain adaptation from a graphical perspective



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

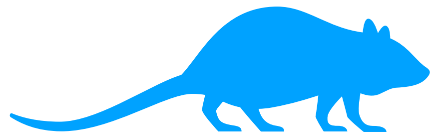


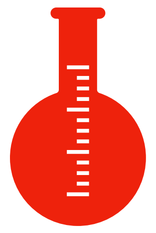


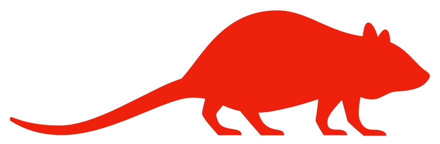
D	X1	X2	Y
Gene A	3,1	2	?
Gene A	3,2	3	?
Gene A	4	2	?
Gene A	3,2	3	?

- Add a variable D to represent the **domain**

Domain adaptation from a graphical perspective



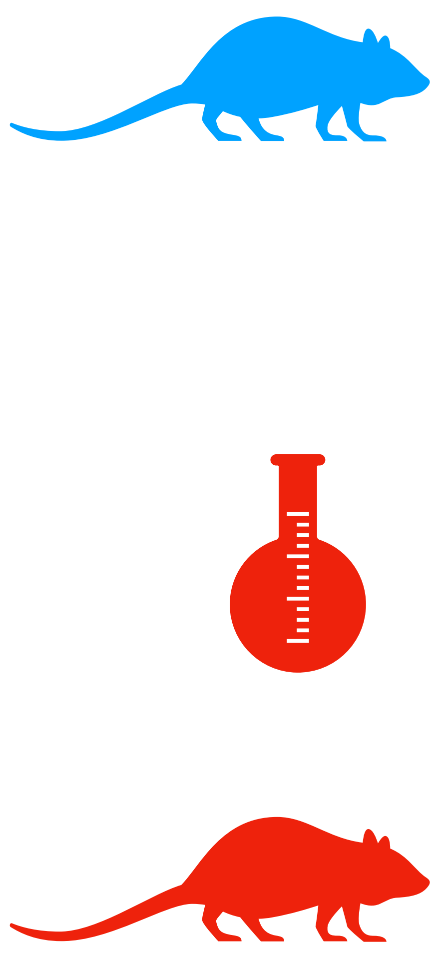




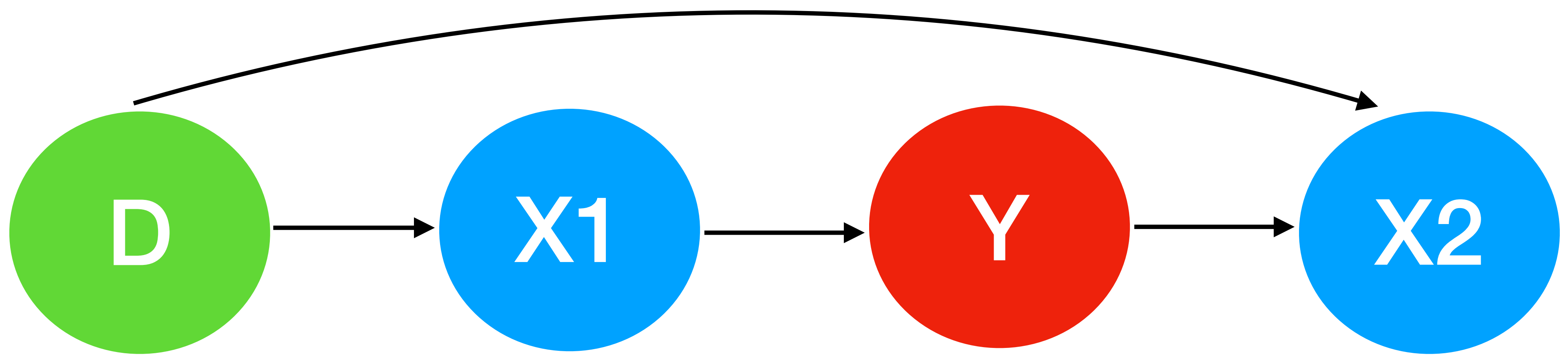
D	X1	X2	Y
Normal	0,1	2	0
Normal	0,2	3	0
Normal	1,1	2	1
Normal	0,1	3	0
Gene A	3,1	2	?
Gene A	3,2	3	?
Gene A	4	2	?
Gene A	3,2	3	?

- Add a variable D to represent the **domain**
- Consider the data as coming from a single distribution P(**X**,Y, D)

Domain adaptation from a graphical perspective



D	X1	X2	Y
Normal	0,1	2	0
Normal	0,2	3	0
Normal	1,1	2	1
Normal	0,1	3	0
Gene A	3,1	2	?
Gene A	3,2	3	?
Gene A	4	2	?
Gene A	3,2	3	?



- We can represent P(**X**,Y, D) with a **(possibly unknown)** causal graph

- Add a variable D to represent the **domain**

- Consider the data as coming from a single distribution P(**X**,Y, D)

Structural causal model - domain/environment variable

$$\begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = -2Y + \epsilon_2 \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases}$$

$$D = 0$$

Structural causal model - domain/environment variable

$$\begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = -2Y + \epsilon_2 \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases} \qquad D = 0$$

$$\begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = 1 \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases} \qquad D = 1$$

$$\begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = 10Y + \epsilon_Y \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases} \qquad D = 2$$

Structural causal model - domain/environment variable

$$\begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = -2Y + \epsilon_2 \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases} \quad D = 0$$

$$\begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = 1 \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases} \quad D = 1$$

$$\begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = 10Y + \epsilon_Y \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases} \quad D = 2$$

$$\begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = \begin{cases} -2Y + \epsilon_2 \text{ if } D = 0 \\ 1 \quad\quad \text{ if } D = 1 \\ 10Y + \epsilon_Y \text{ if } D = 2 \end{cases} \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases}$$

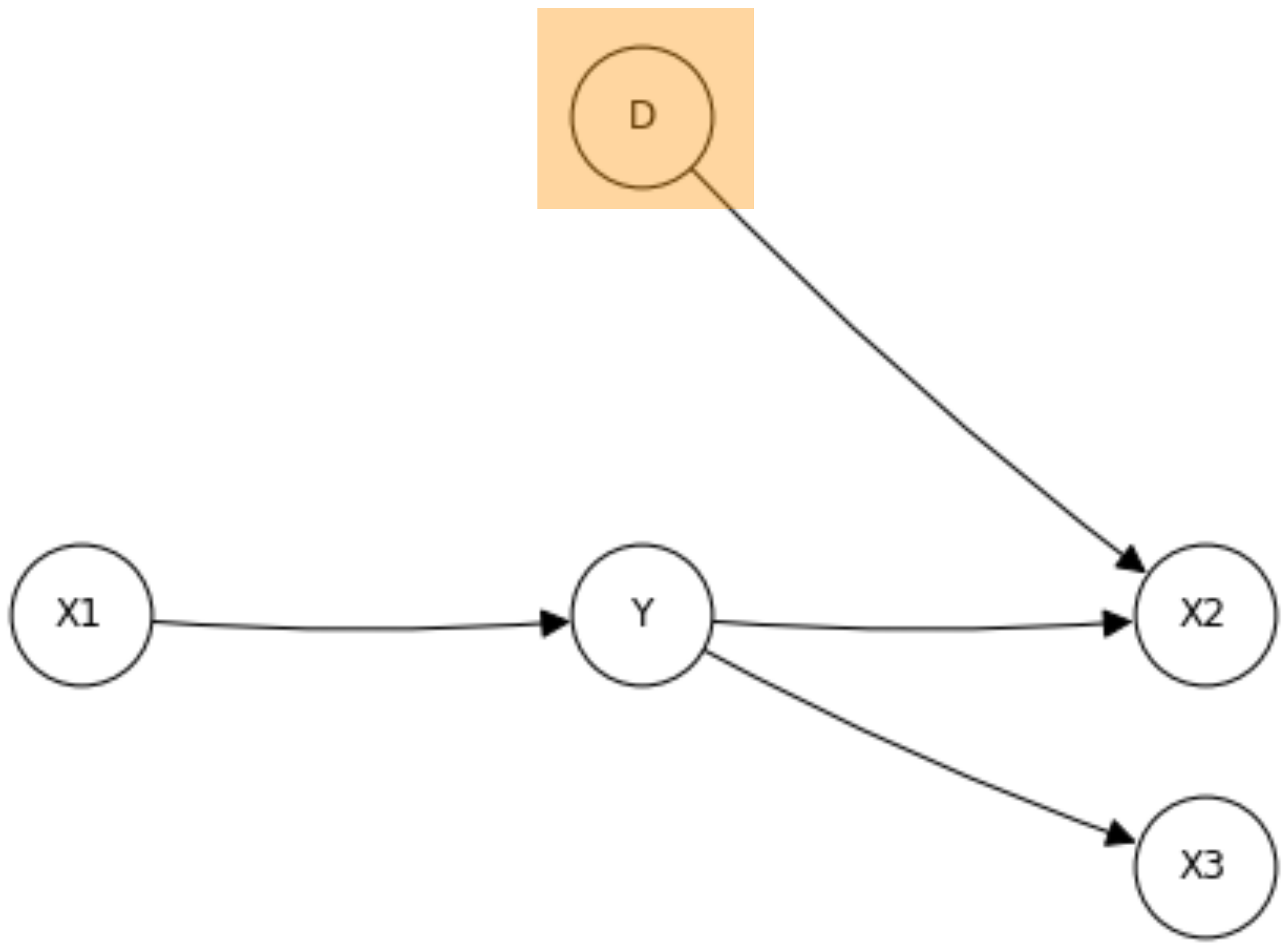
Structural causal model - domain/environment variable

$$\begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = -2Y + \epsilon_2 \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases} \quad D = 0$$

$$\begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = 1 \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases} \quad D = 1$$

$$\begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = 10Y + \epsilon_Y \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases} \quad D = 2$$

$$\begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = \begin{cases} -2Y + \epsilon_2 \text{ if } D = 0 \\ 1 \quad\quad\quad \text{ if } D = 1 \\ 10Y + \epsilon_Y \text{ if } D = 2 \end{cases} \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases}$$



Domain adaptation example

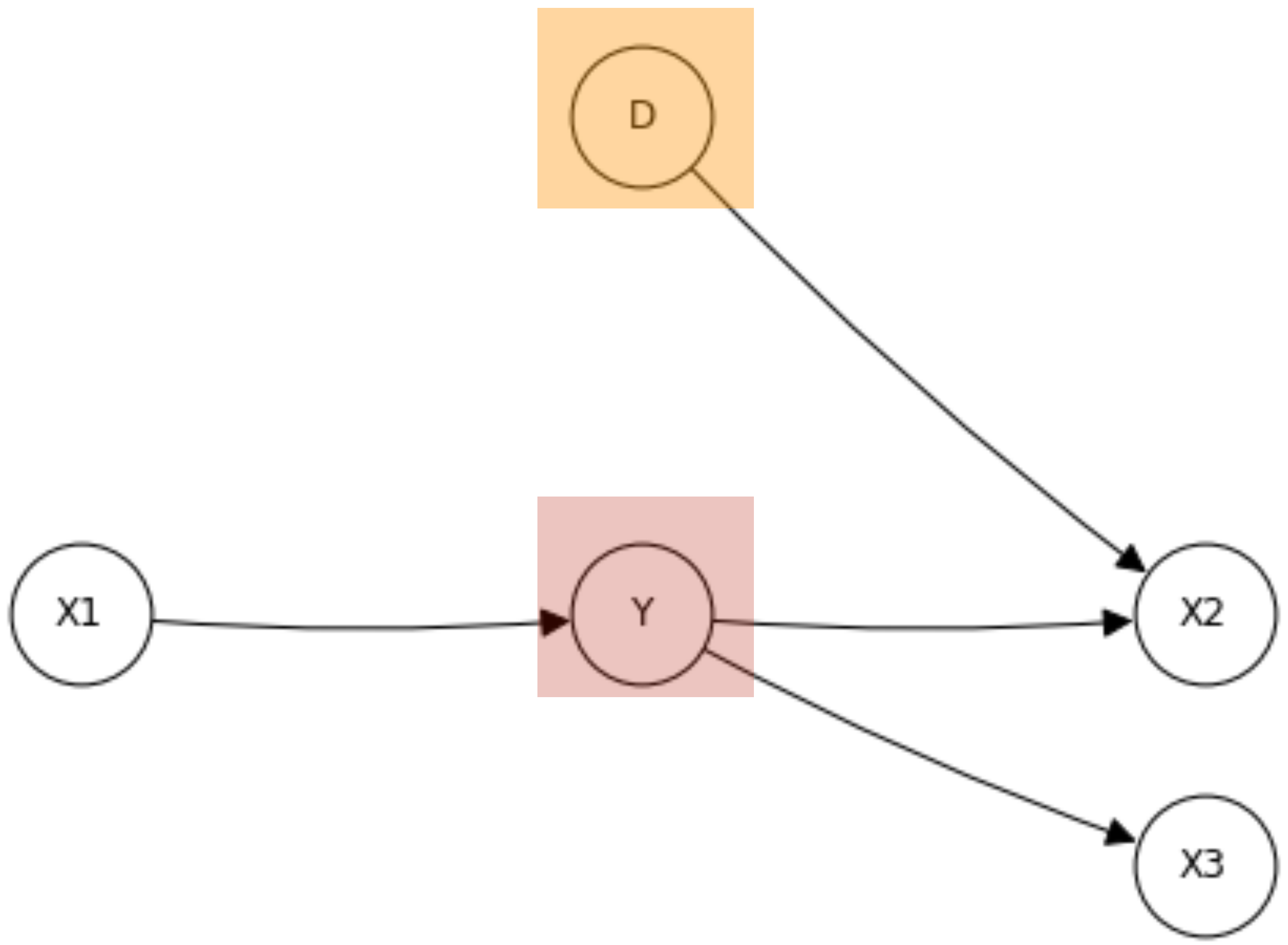
$$D = 0 \begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = -2Y + \epsilon_2 \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases}$$

$$D = 1 \begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = 1 \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases}$$

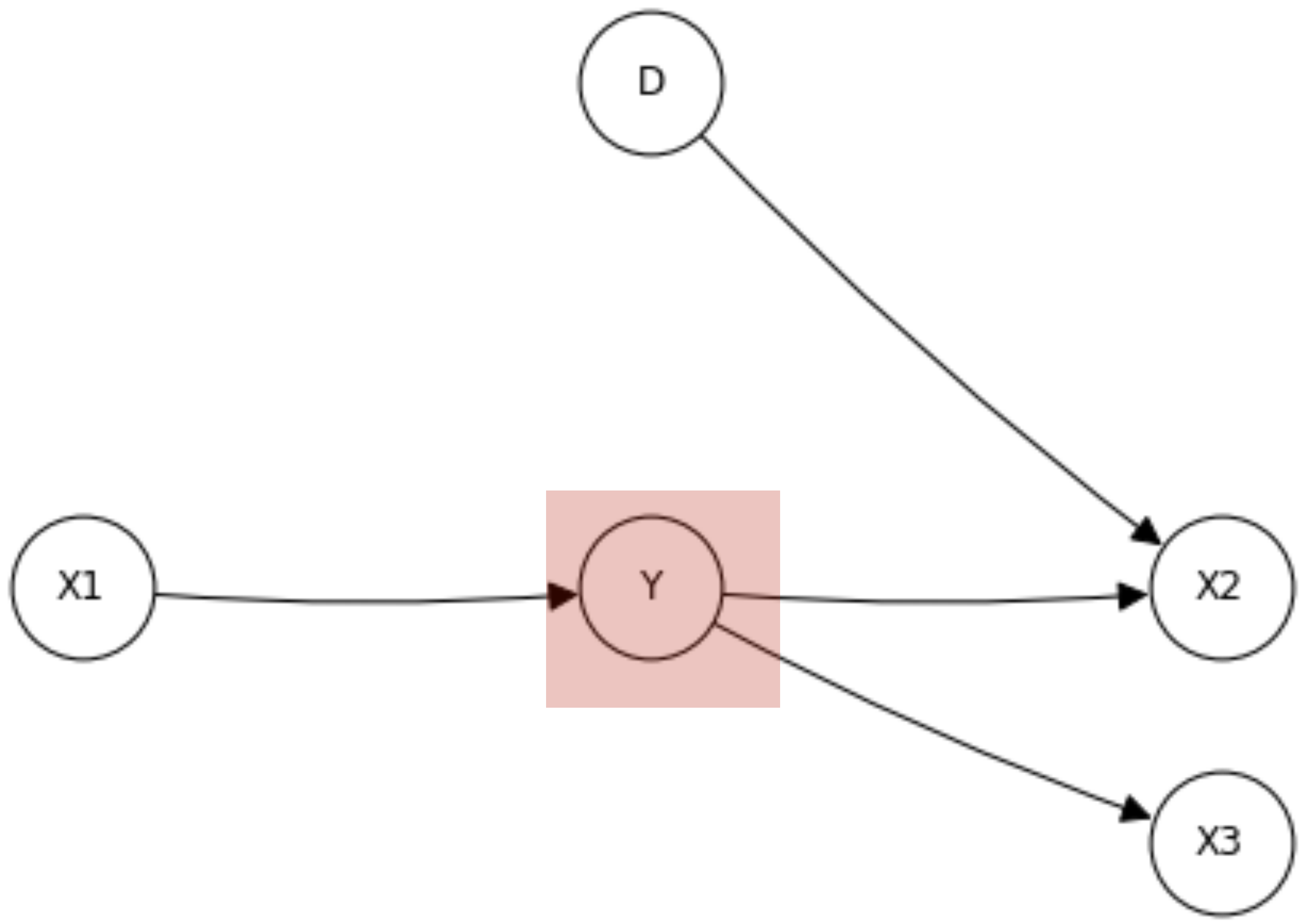
Source domains

$$D = 2 \begin{cases} \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_Y \sim \mathcal{N}(0,1) \\ X_1 = 10 + \epsilon_1 \\ Y = 3X_1 + \epsilon_Y \\ X_2 = 10Y + \epsilon_Y \\ X_3 = 2Y + 0.1\epsilon_3 \end{cases}$$

Target domain



Domain adaptation example



$D = 0$

d	x1	y	x2	x3
0	8.973763	26.130494	-51.648475	52.330948
0	10.428340	31.894998	-64.373356	63.802704
0	8.911484	25.166962	-52.313502	50.279162
0	9.841798	29.783299	-60.419296	59.539914
0	8.969118	27.660573	-55.075839	55.327185

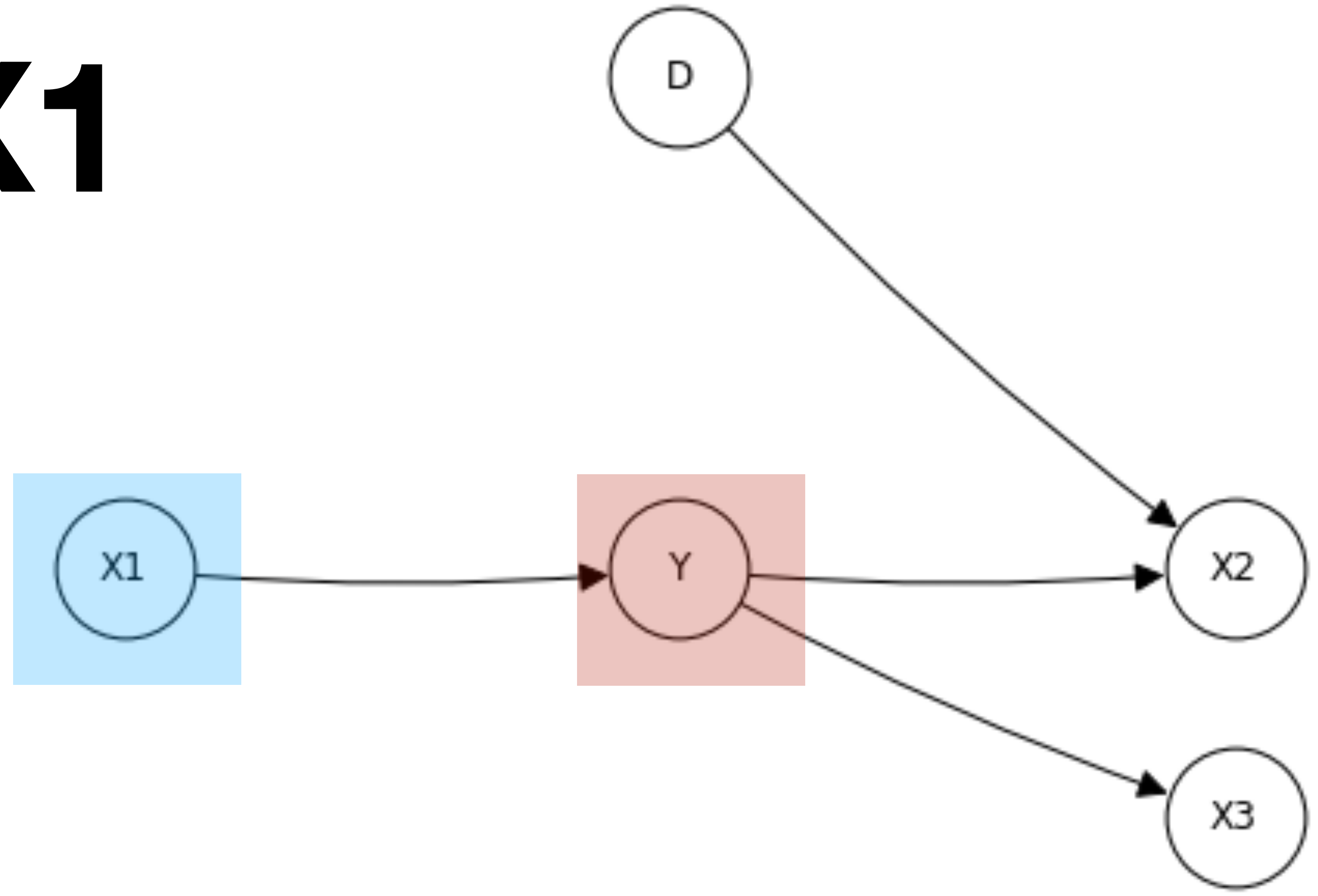
$D = 1$

d	x1	y	x2	x3
1	9.941015	28.696601	1	57.475345
1	8.762380	25.715927	1	51.275390
1	9.636201	28.407387	1	56.884332
1	10.875069	31.370200	1	62.686789
1	10.023968	31.253540	1	62.388444

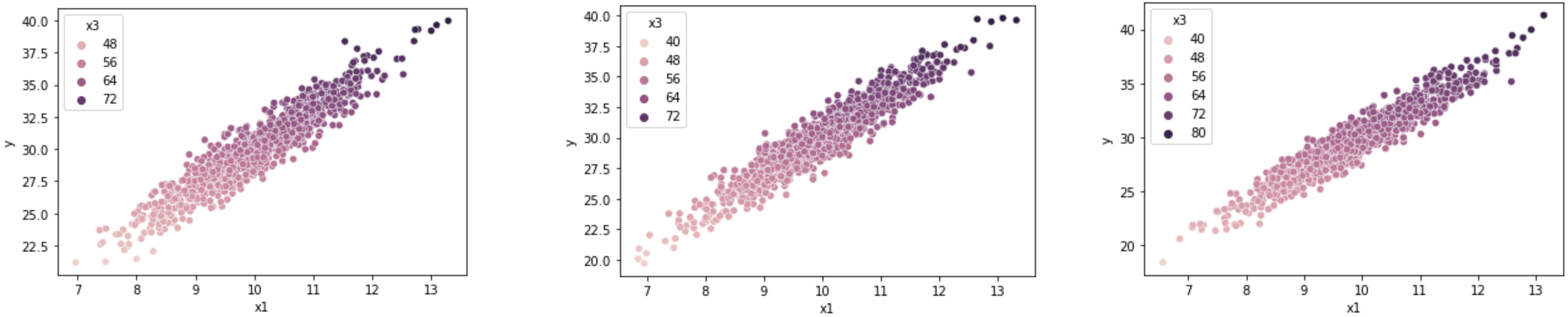
$D = 2$

d	x1	y	x2	x3
2	9.671277	26.556214	265.034283	53.338139
2	9.613139	27.120226	270.746784	54.340341
2	10.718335	29.589532	295.318526	59.291053
2	9.002388	26.629254	264.942583	53.340389
2	9.289340	29.030355	289.747562	58.098312

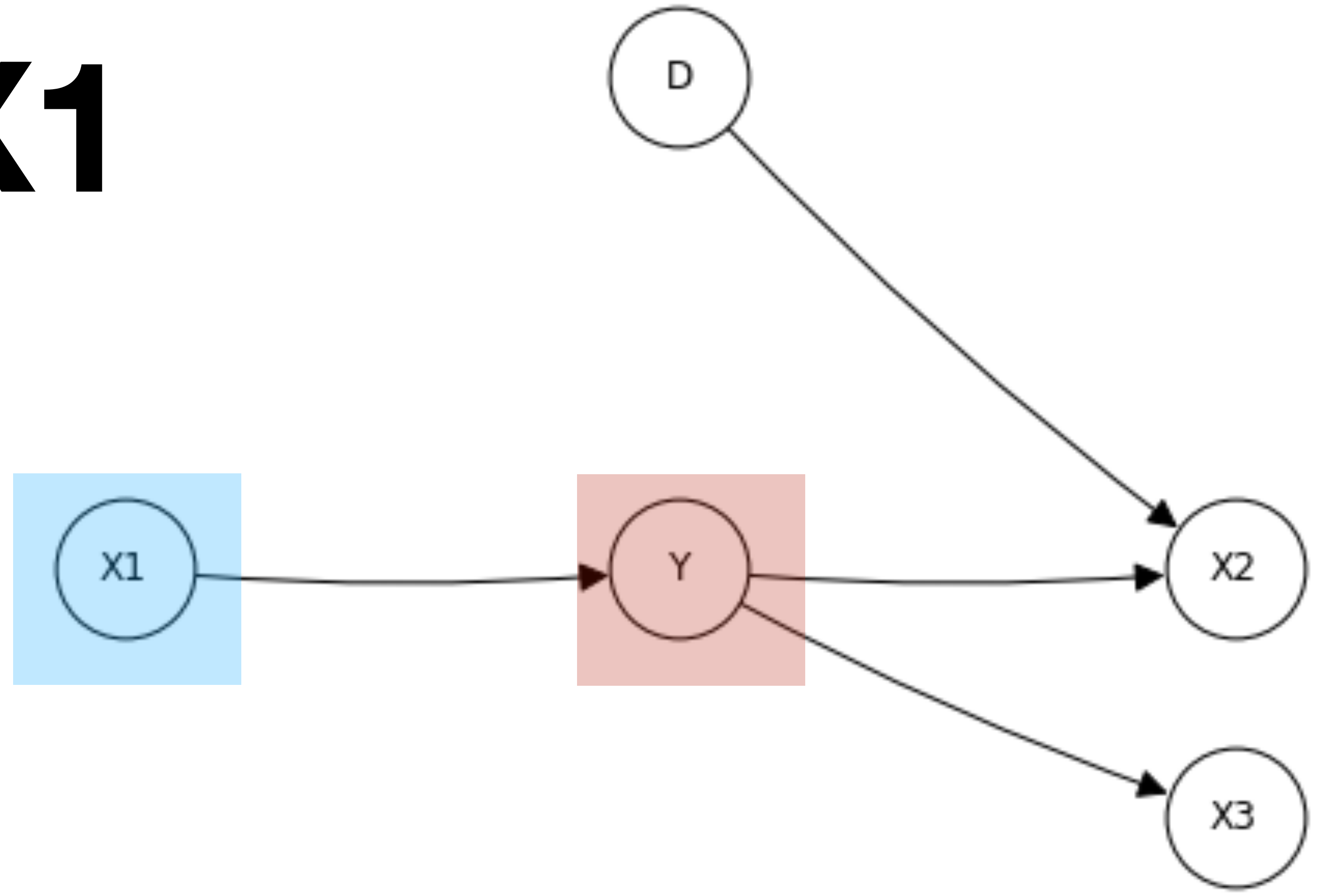
Domain adaptation example - X1



$D = 0$ $\quad D = 1$ $\quad D = 2$

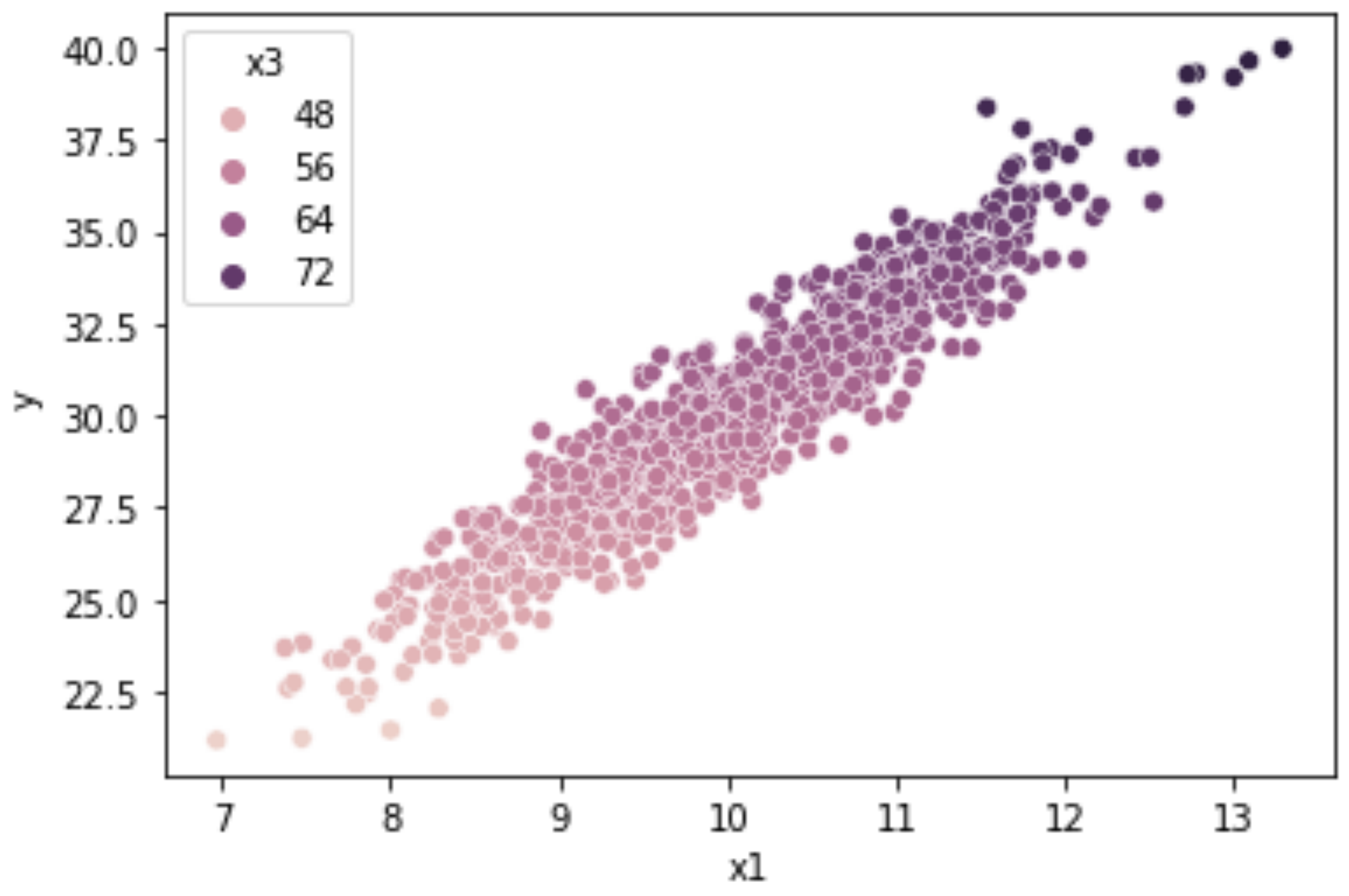


Domain adaptation example - X1

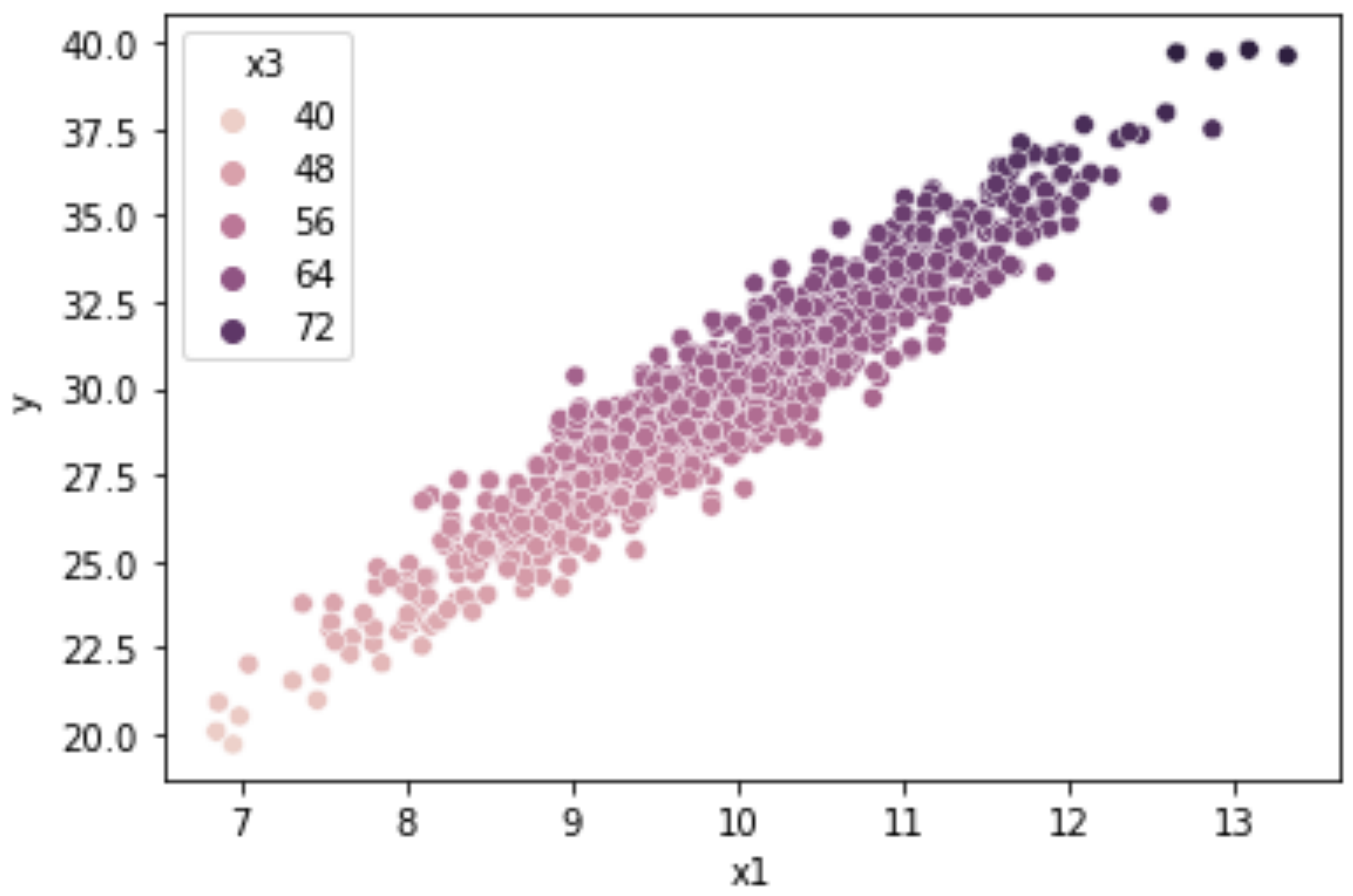


$p(Y|X_1)$ is invariant

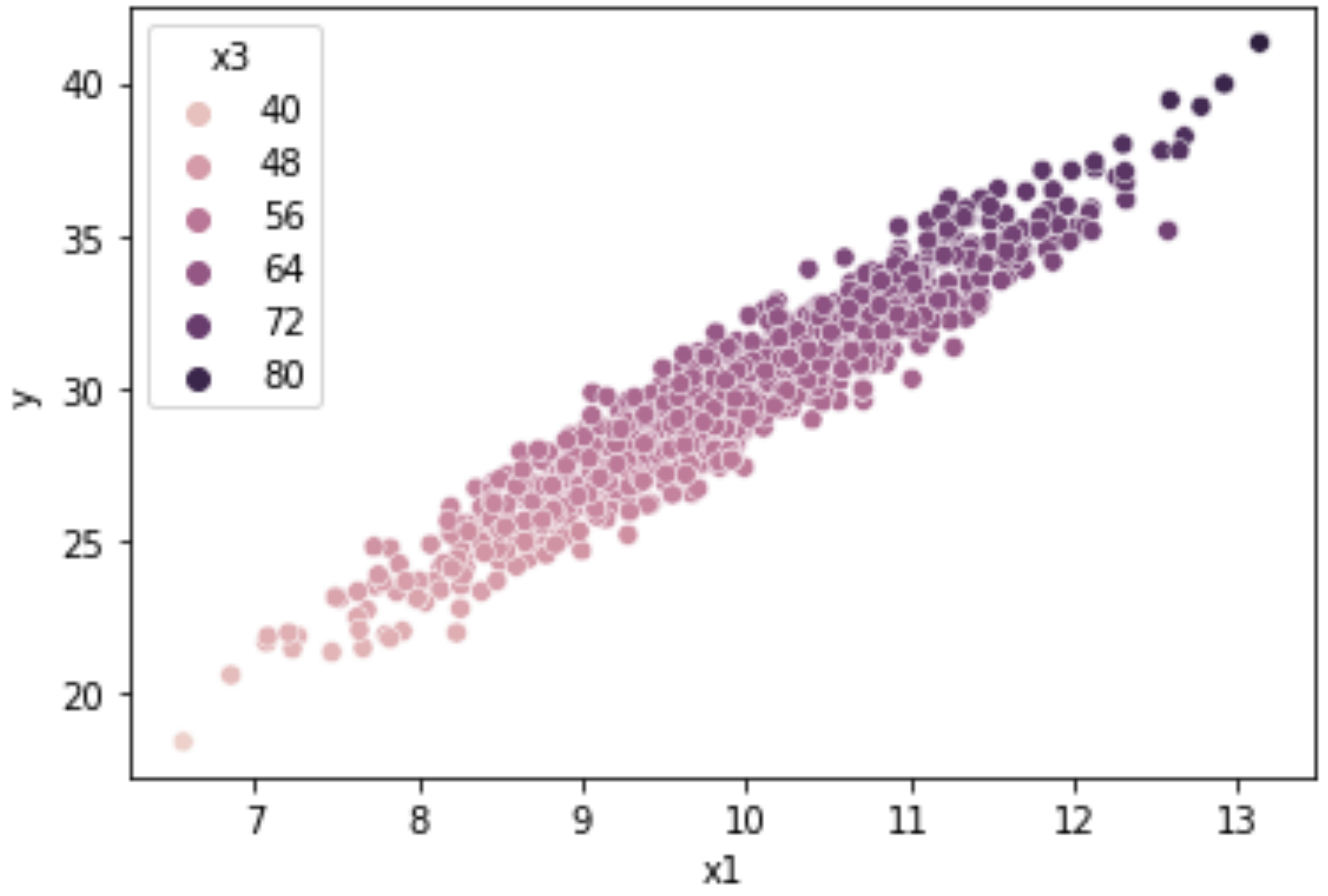
$D = 0$



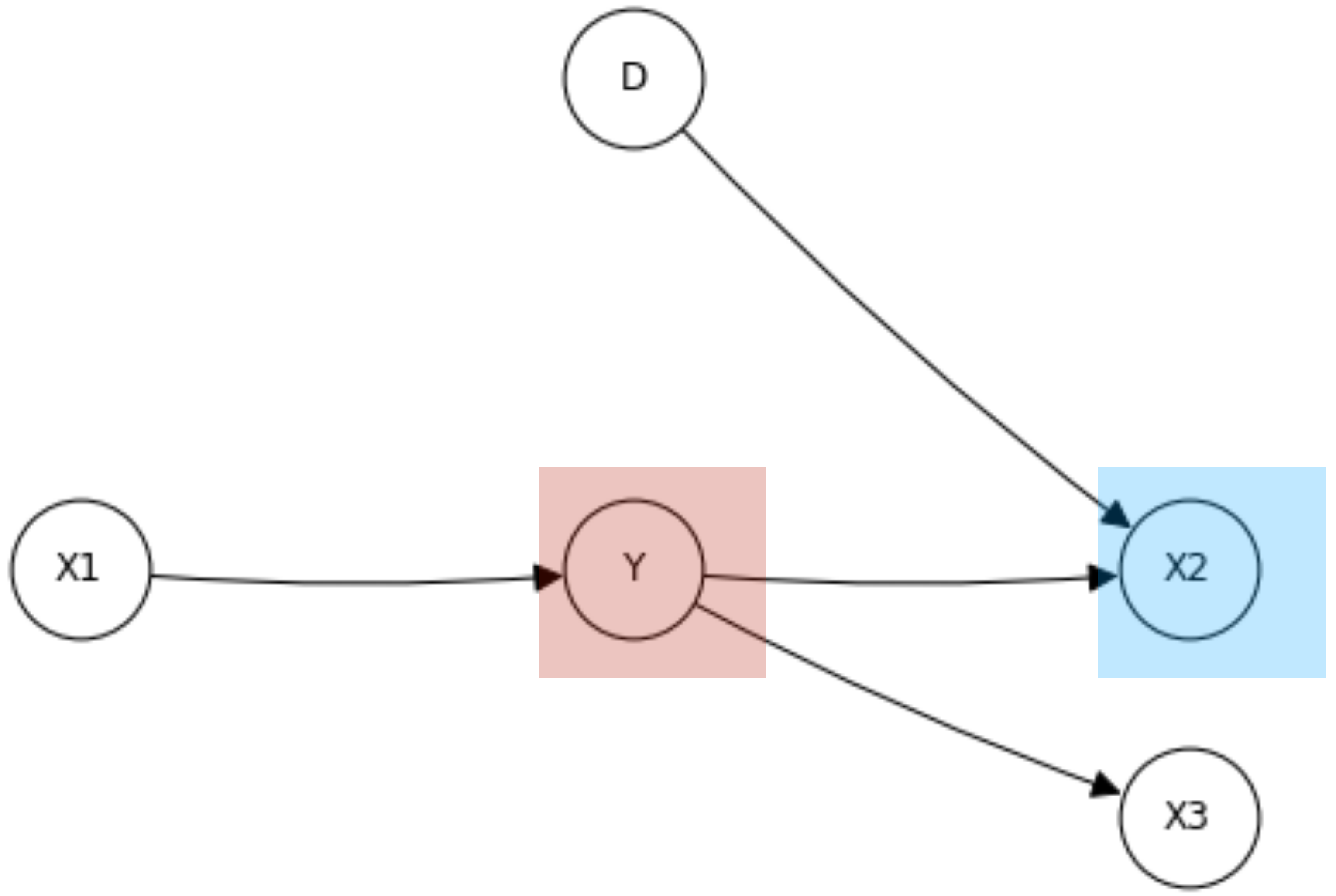
$D = 1$



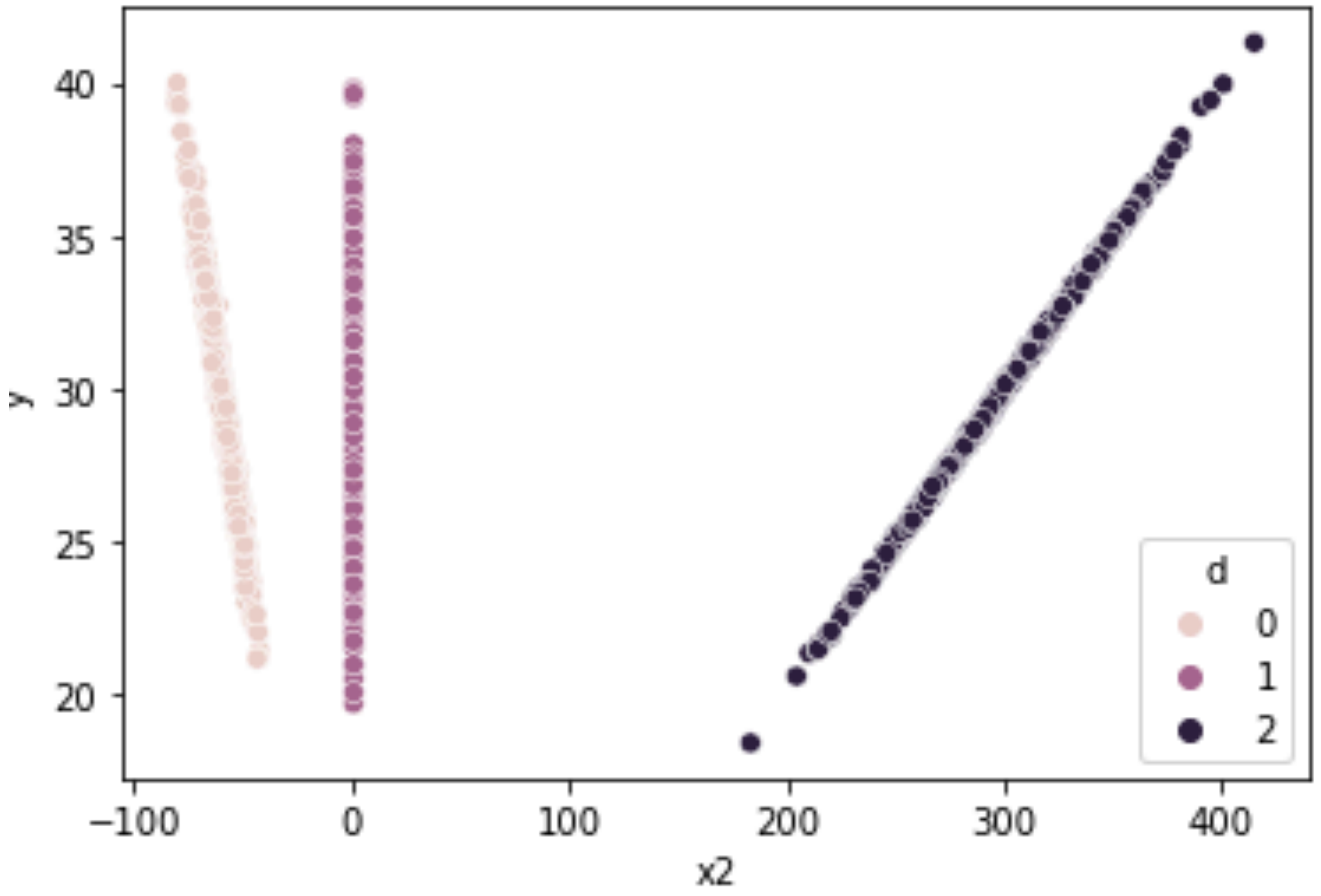
$D = 2$



Domain adaptation example - X2

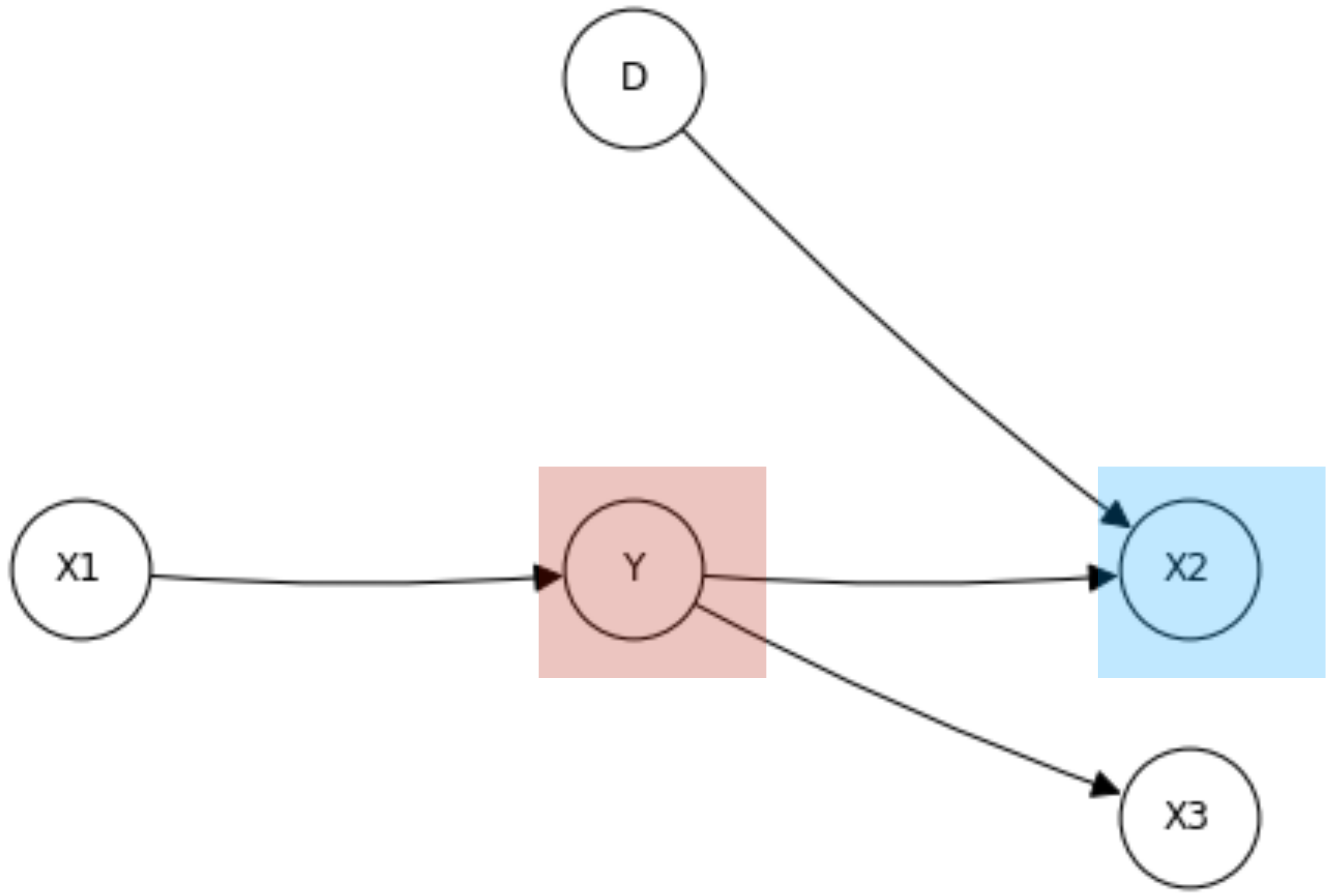


Source domains Target domain

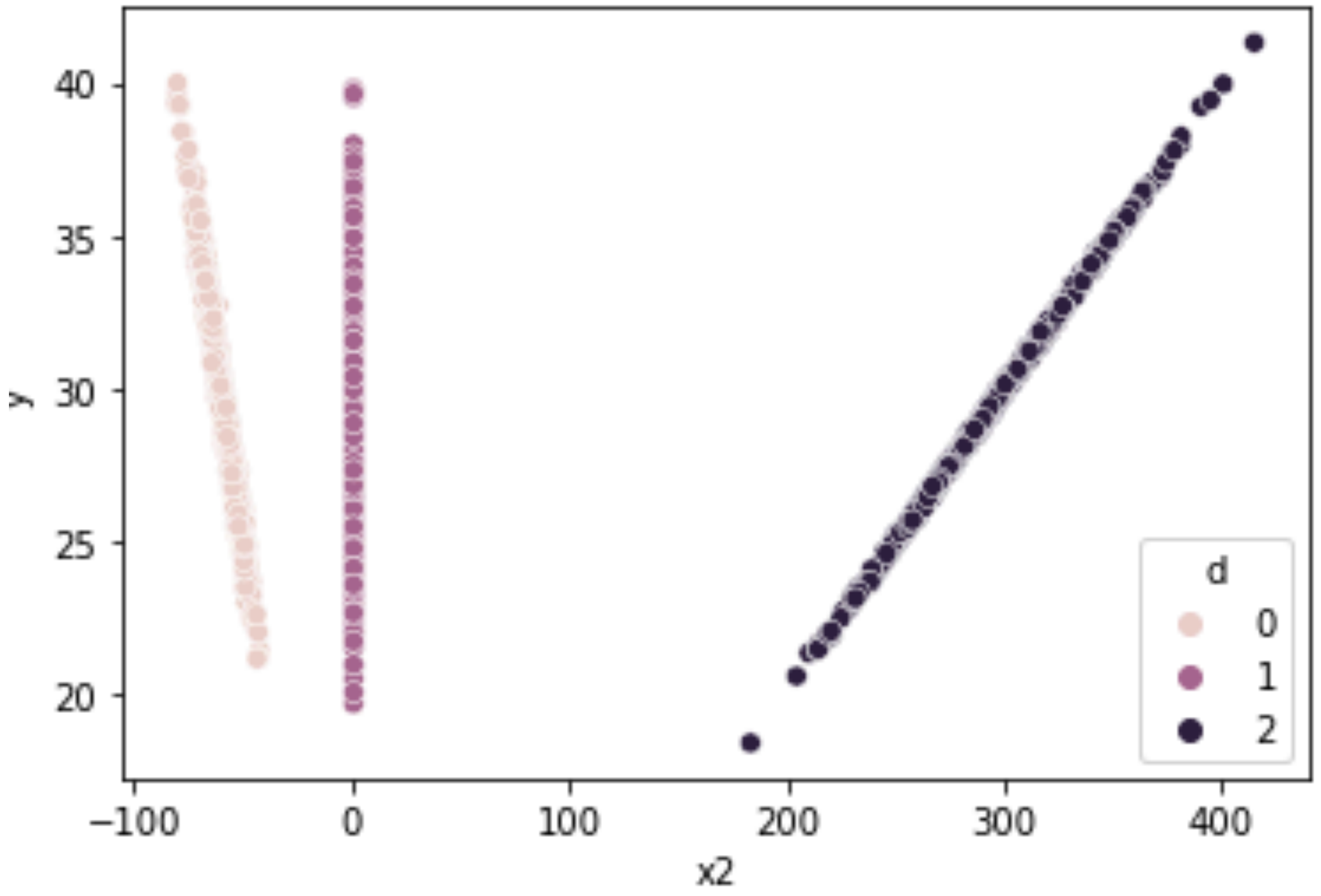


$P(Y | X_2)$ is <u>not</u> invariant

Domain adaptation example - X2



Source domains   Target domain

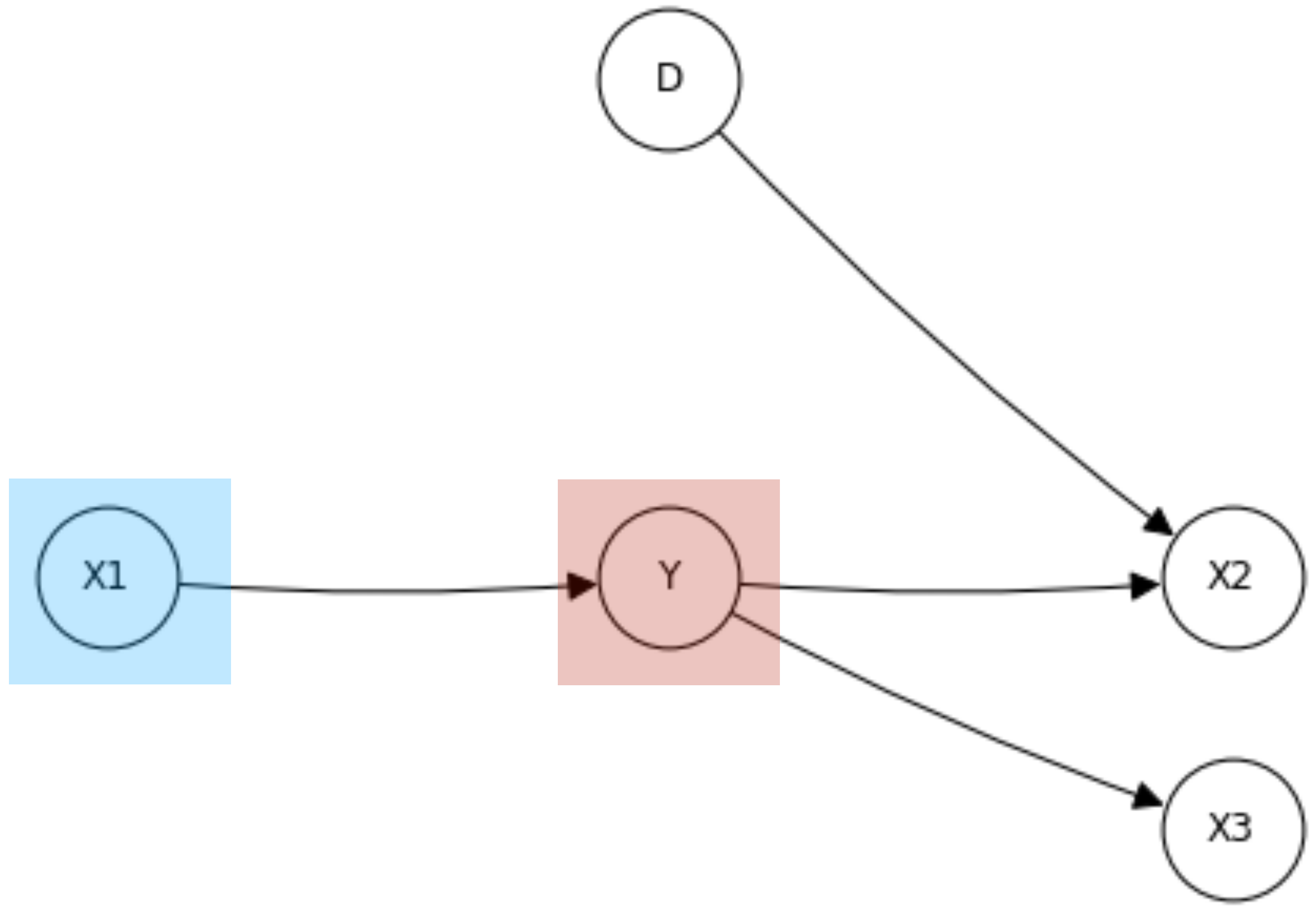


$P(Y|X_2)$ is not invariant

```
sns.scatterplot(data = df, x="x2", y="y", hue="d")
X2_0 = df_0["x2"].values.reshape(-1, 1)
X2_2 = df_2["x2"].values.reshape(-1, 1)
model = LinearRegression().fit(X2_0, Y_0)
est_Y_2 = model.predict(X2_2)
print("Mean squared error predicting Y in environment 2 based on model learnt in environment 0 from X2", mean_squared_error(Y_2,est_Y_2))
```

```
Mean squared error predicting Y in environment 2 based on model learnt in environment 0 from X2 30518.374428658524
```

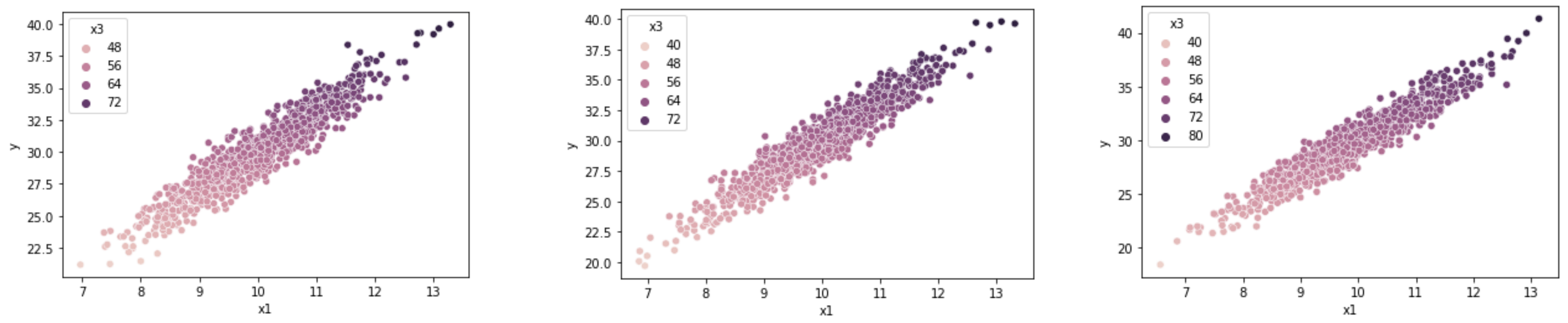
Separating features intuition - X1



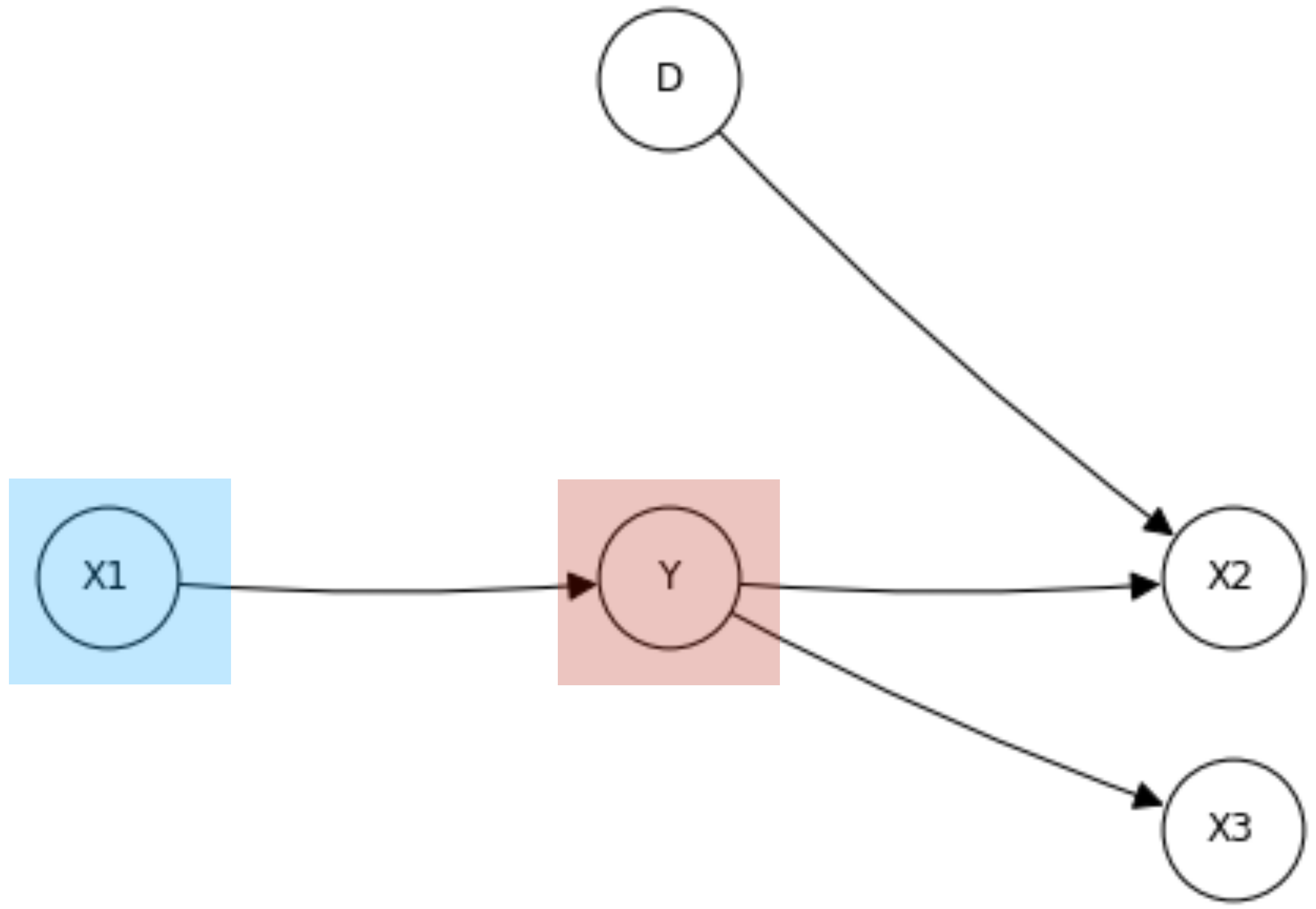
$P(Y|X_1)$ is invariant

$$P(Y|X_1, D=0) = P(Y|X_1, D=1) = P(Y|X_1, D=2) = P(Y|X_1)$$

$$P(X_1, Y, X_2, X_3, D)$$



Separating features intuition - X1



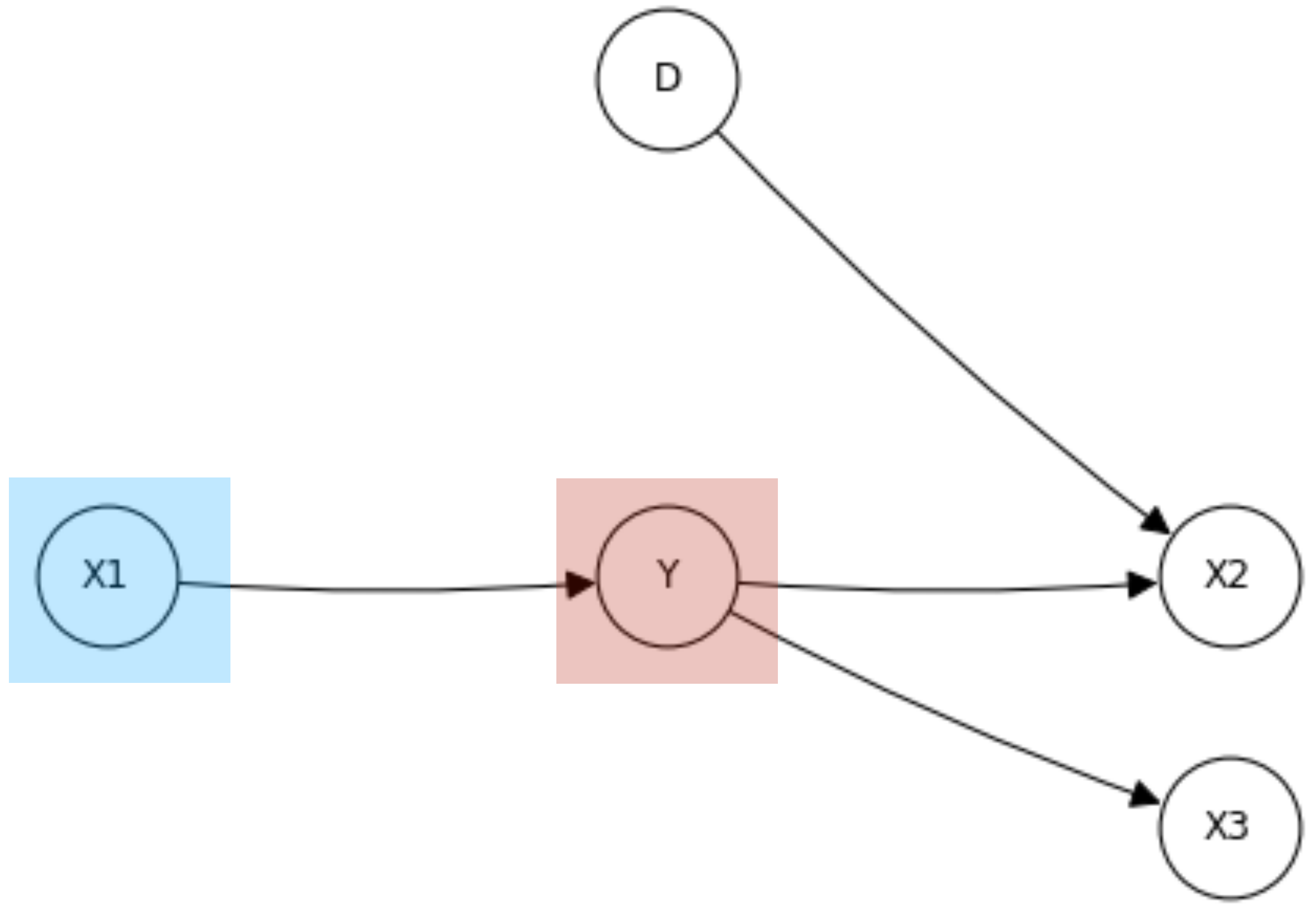
$P(Y|X_1)$ is invariant

$P(Y|X_1, D=0) = P(Y|X_1, D=1) = P(Y|X_1, D=2)$
$= P(Y|X_1)$

↳ this is true if $Y \perp\!\!\!\perp D \mid X_1$

$P(X_1, Y, X_2, X_3, D)$

Separating features intuition - X1



$$p(Y|X_1) \text{ is invariant}$$

$$P(Y|X_1, D=0) = P(Y|X_1, D=1) = P(Y|X_1, D=2)$$
$$= P(Y|X_1)$$

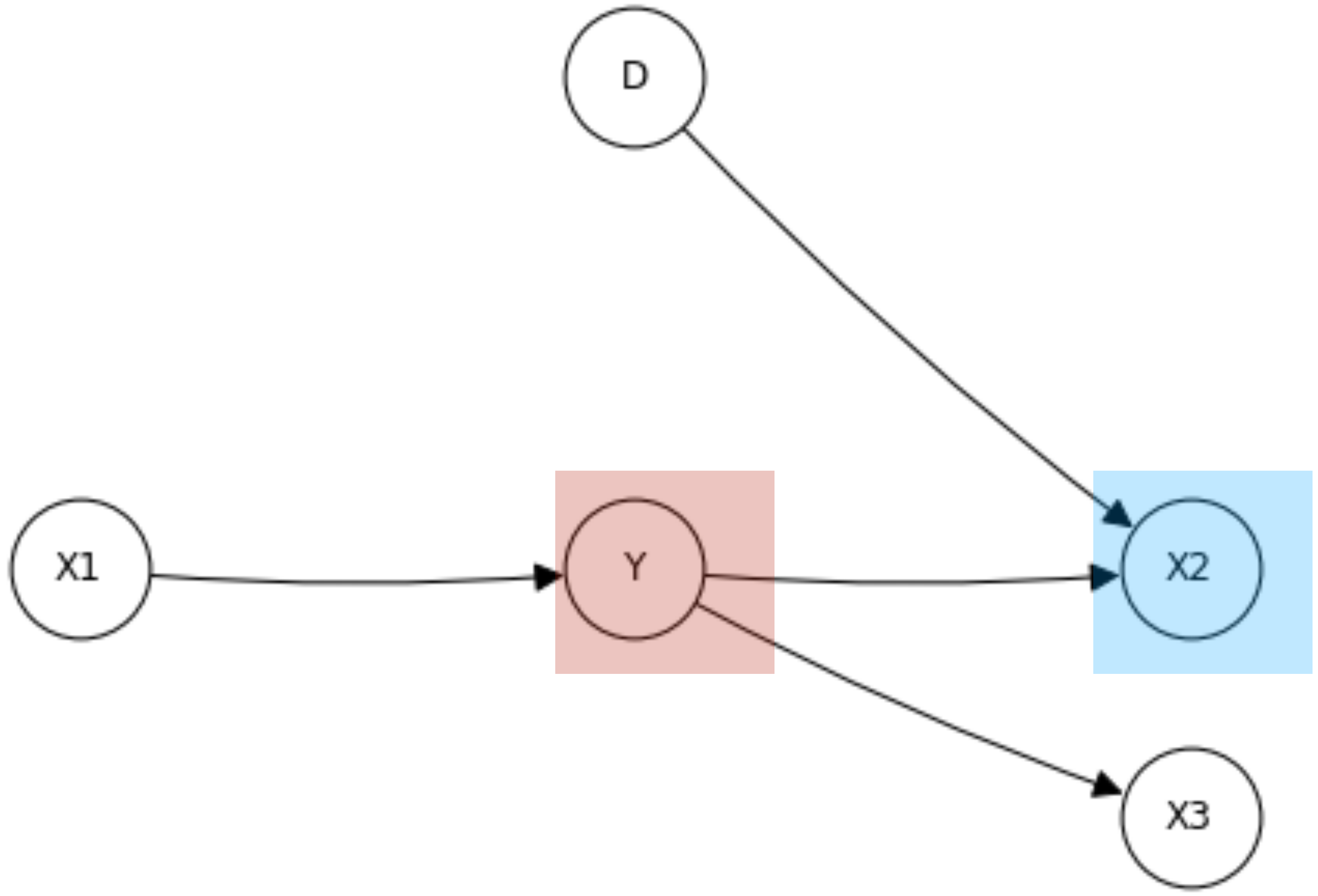
↳ this is true if $Y \perp\!\!\!\perp D \mid X_1$

$Y \perp_d D \mid X_1$ in true graph

$$P(X_1, Y, X_2, X_3, D)$$

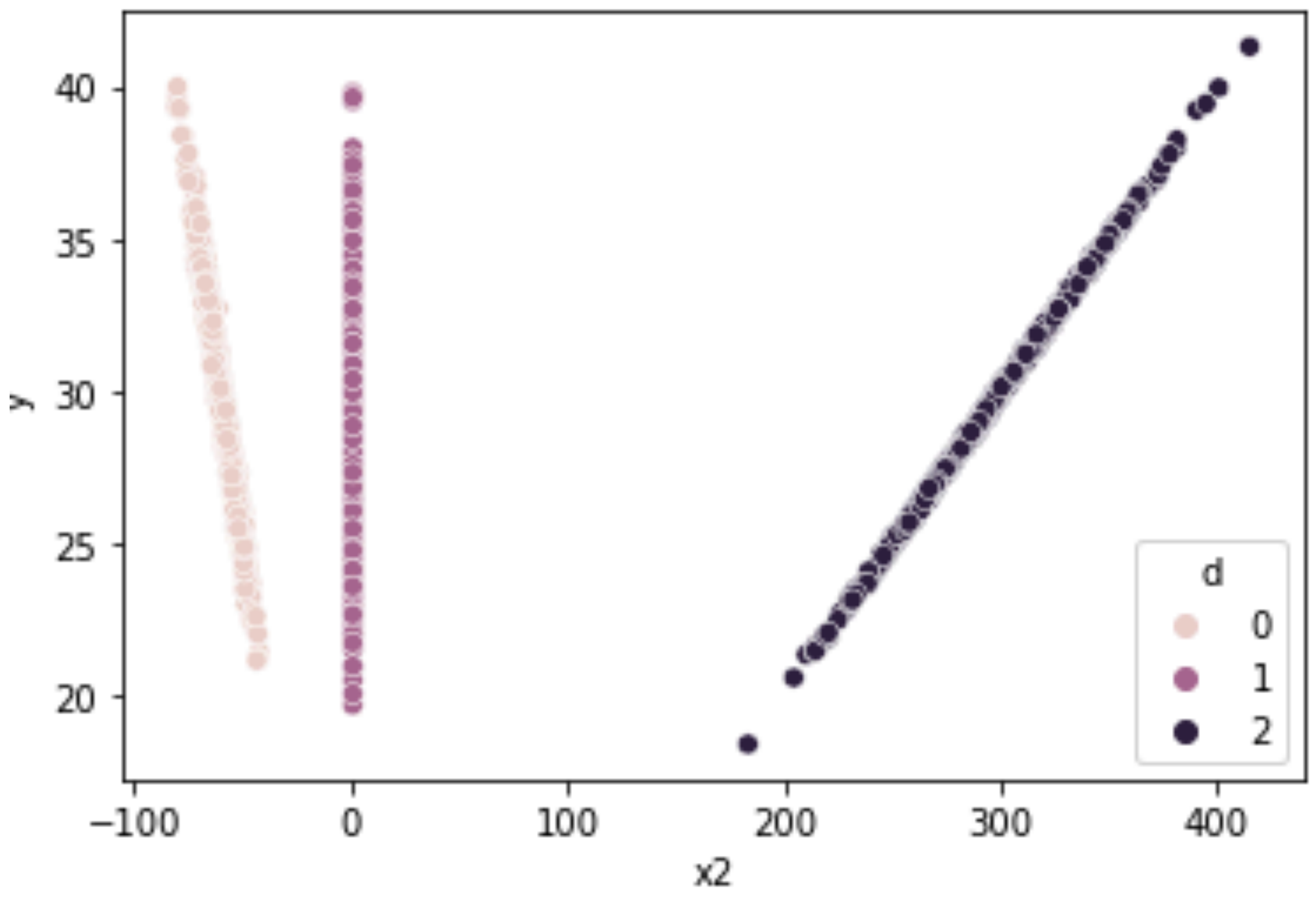
d-separation [Pearl 1988 allows us to read conditional independences from a Bayesian network

Separating features intuition - X2



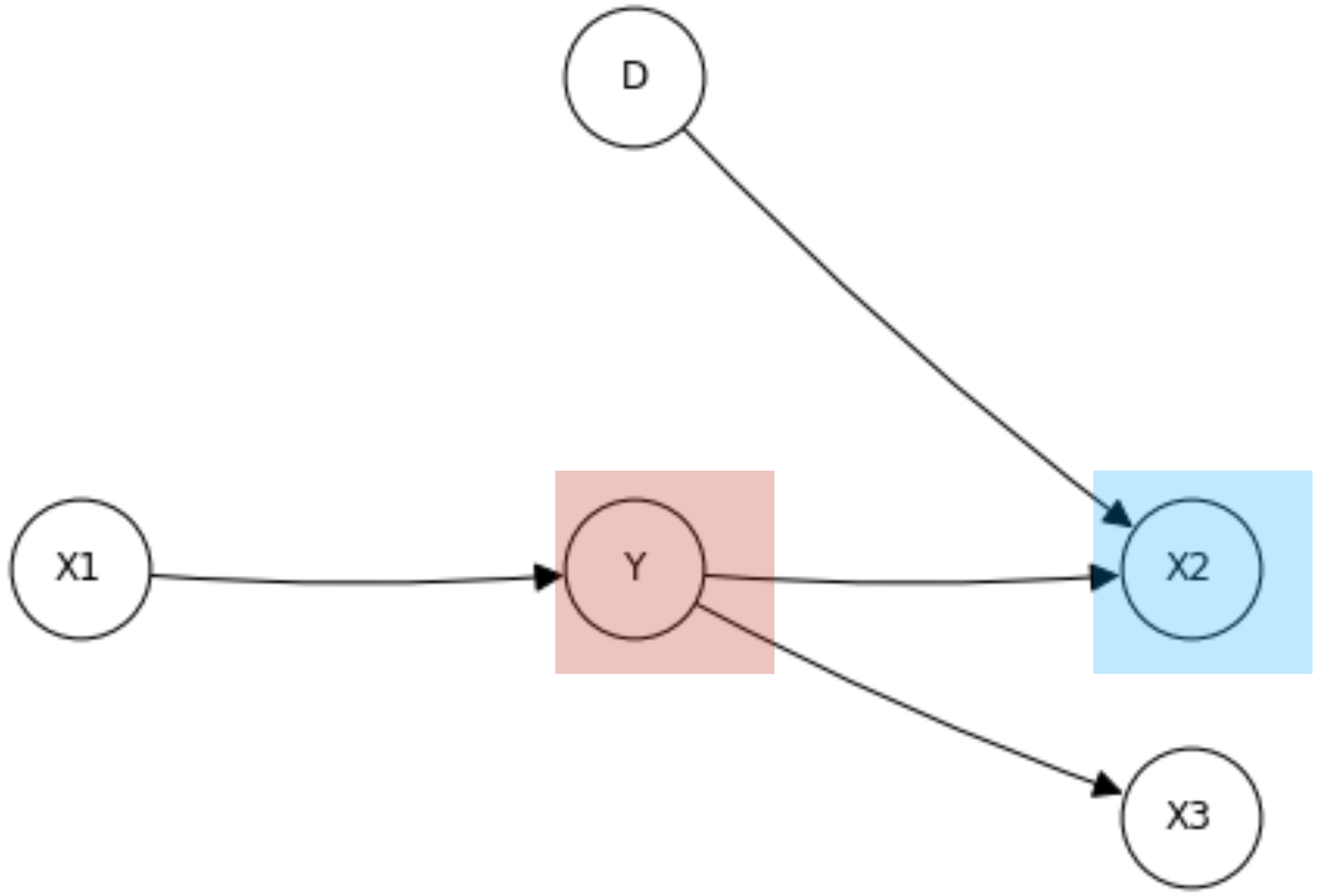
$P(Y \mid X_2)$ is not invariant

$P(Y \mid X_2, D=0) \neq P(Y \mid X_2, D=1) \neq P(Y \mid X_2, D=2)$



$P(X_1, Y, X_2, X_3, D)$

Separating features intuition - X2



$P(Y \mid X_2)$ is <u>not</u> invariant

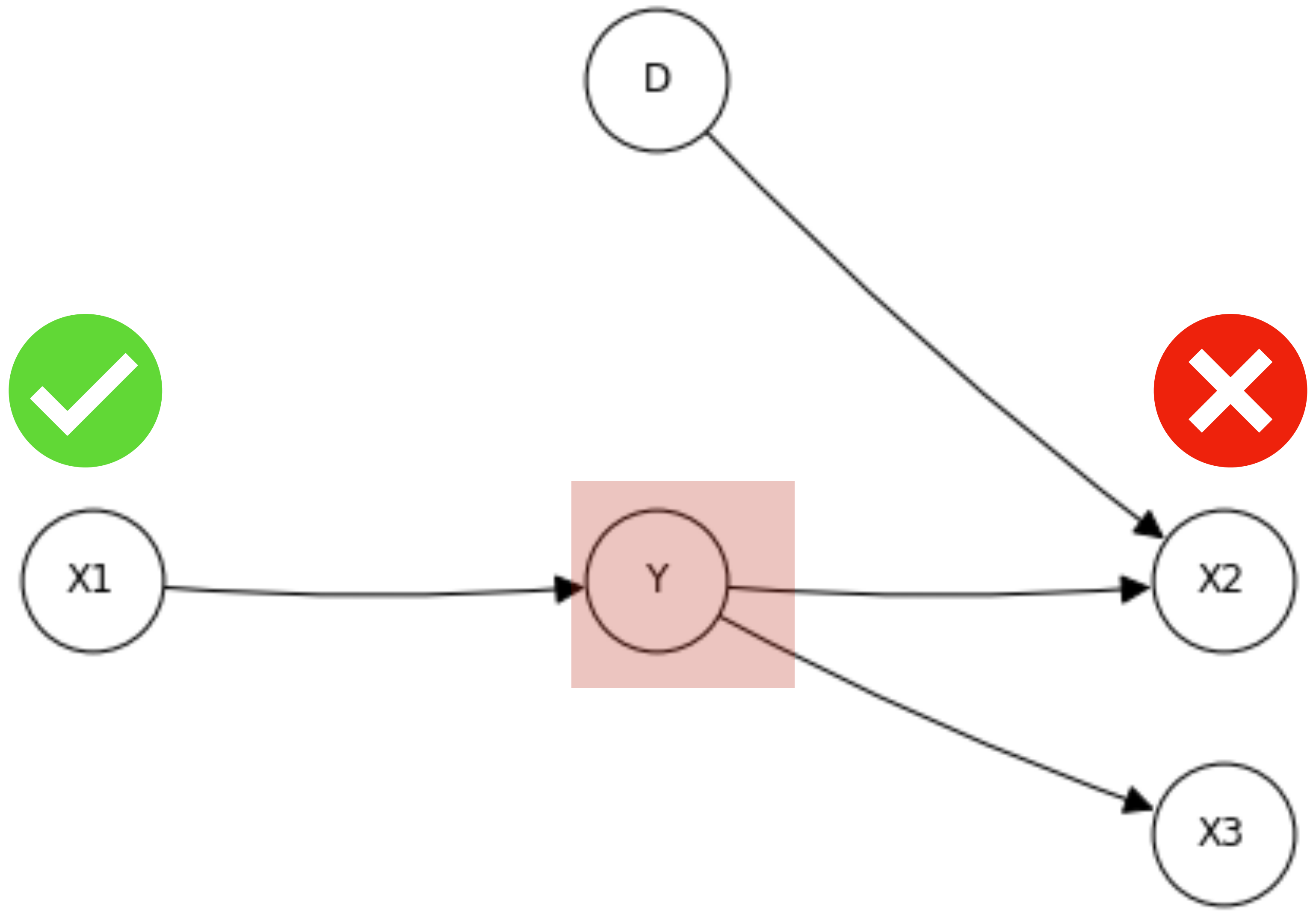
$$P(Y \mid X_2, D=0) \neq P(Y \mid X_2, D=1) \neq P(Y \mid X_2, D=2)$$

↳ this means $Y \not\perp D \mid X_2$

$Y \not\perp_d D \mid X_2$

$$P(X_1, Y, X_2, X_3, D)$$

Separating features intuition - summary

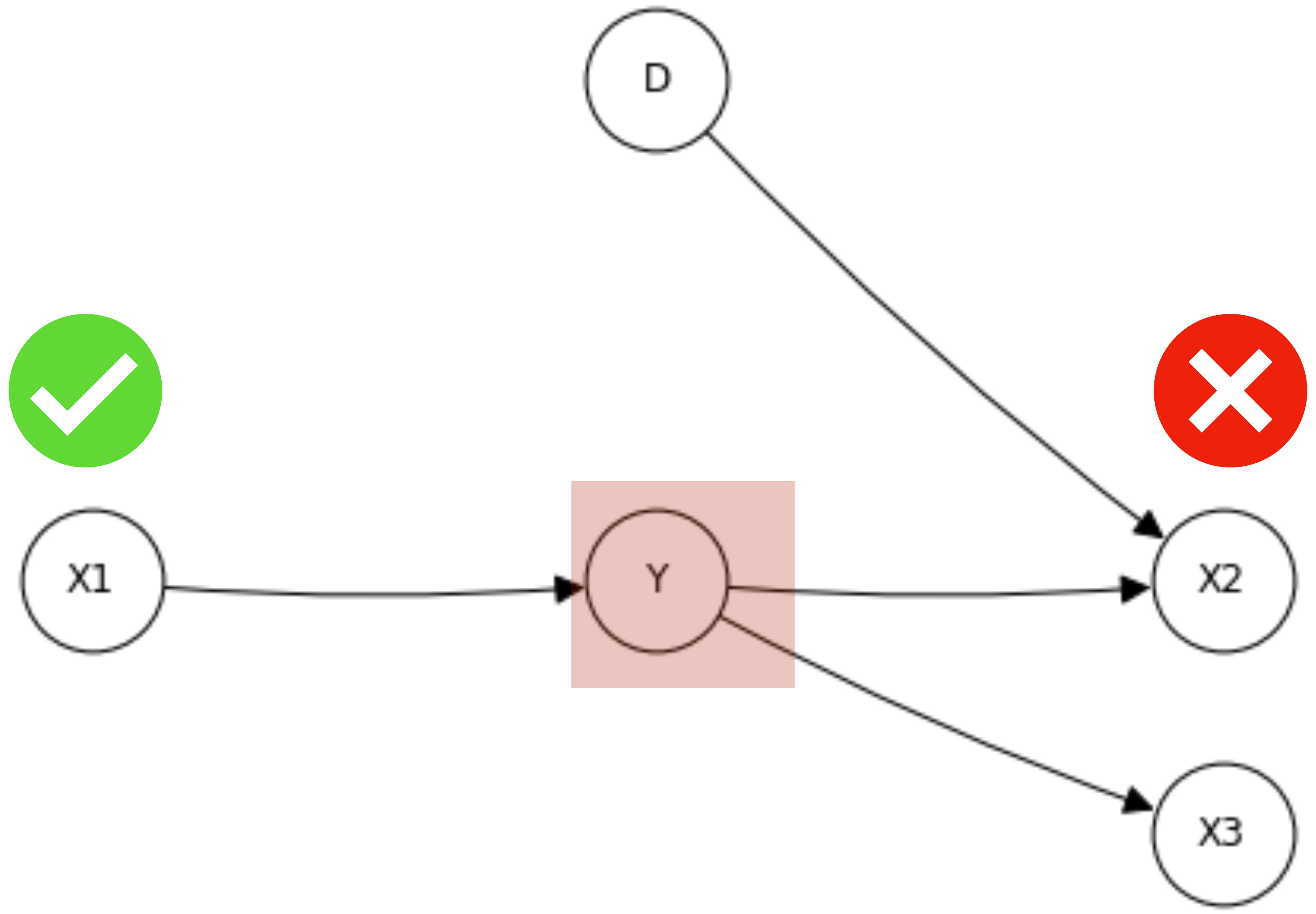


$P(Y|X_1)$ is invariant $\quad Y \perp_d D \mid X_1$

$P(Y|X_2)$ is not invariant $\quad Y \not\perp_d D \mid X_2$

$P(X_1, Y, X_2, X_3, D)$

Separating features intuition - summary



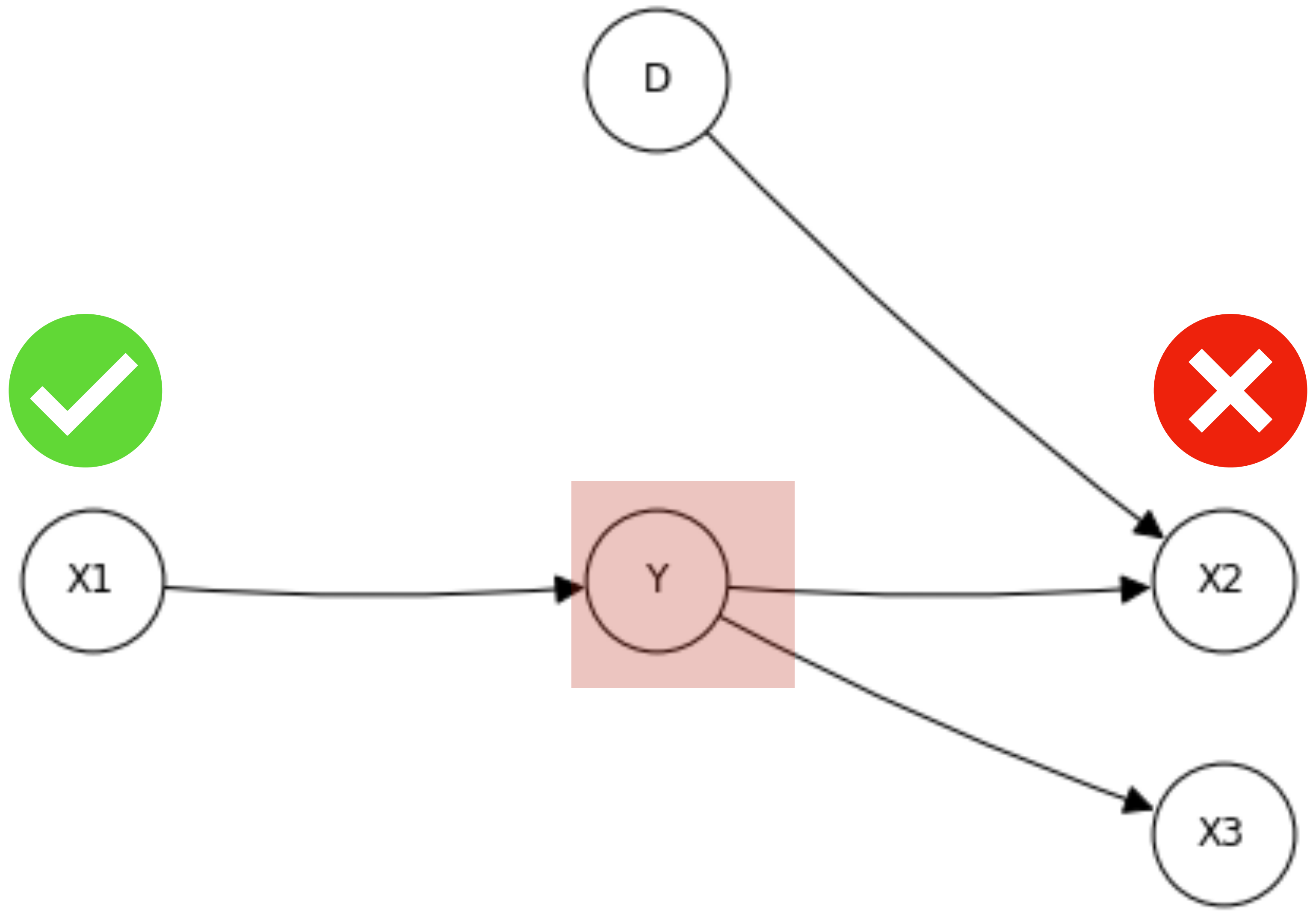
$P(Y|X_1)$ is invariant $\quad Y \perp_d D \mid X_1$

$P(Y|X_2)$ is not invariant $\quad Y \not\perp_d D \mid X_2$

$P(X_1, Y, X_2, X_3, D)$

Look for features $S \subseteq X$ $\quad Y \perp_d D \mid S$

Separating features intuition - summary



$P(Y|X_1)$ is invariant $\quad Y \perp_d D \mid X_1$

$P(Y|X_2)$ is not invariant $\quad Y \not\perp_d D \mid X_2$

$P(X_1, Y, X_2, X_3, D)$

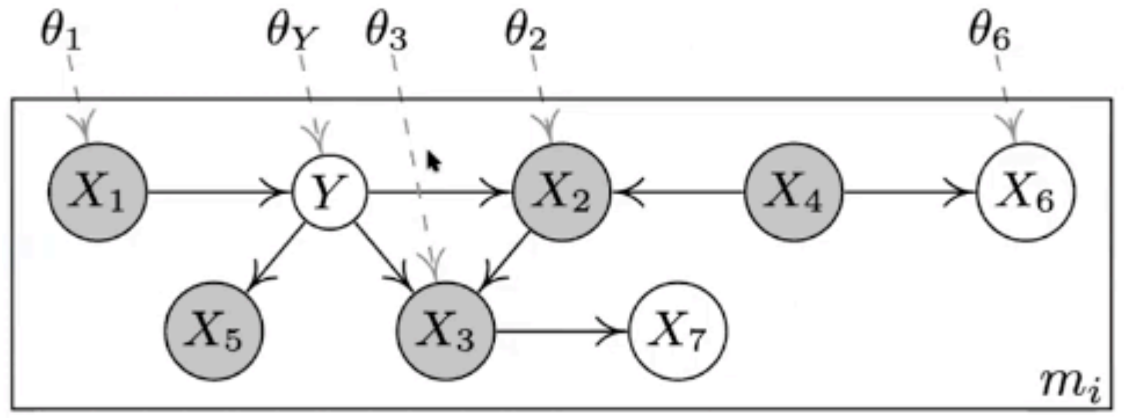
Look for features $S \subseteq X \quad Y \perp_d D \mid S$

$Y \perp_d D \mid \{X_1, X_3\}$

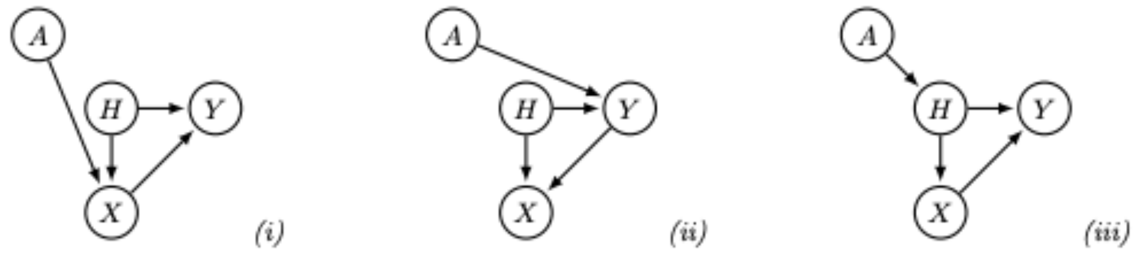
Causality allows us to reason systematically about distribution shifts

On Causal and Anticausal Learning

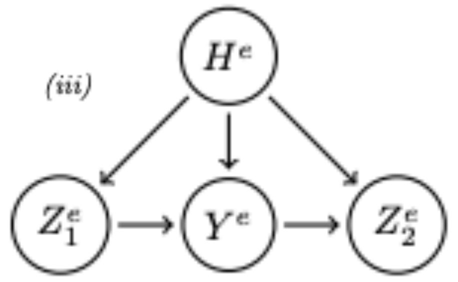
Domain Adaptation as a Problem of Inference on Graphical Models



Anchor regression: heterogeneous data meet causality

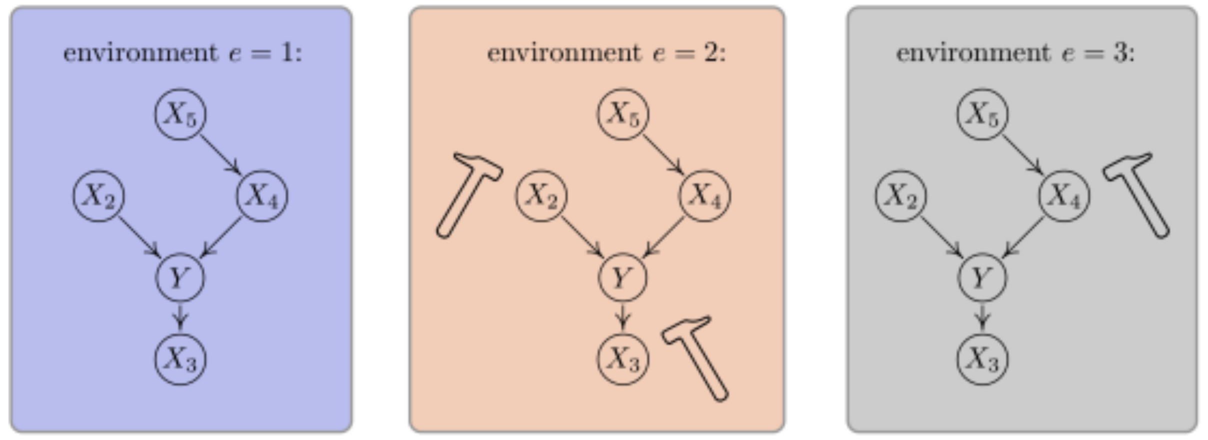


Invariant Risk Minimization

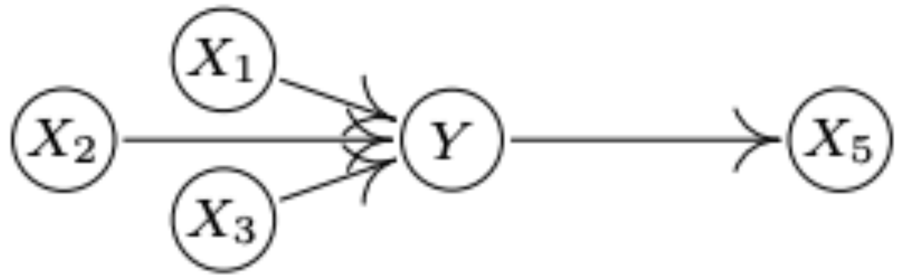


J. R. Statist. Soc. B (2016)
78, Part 5, pp. 947–1012

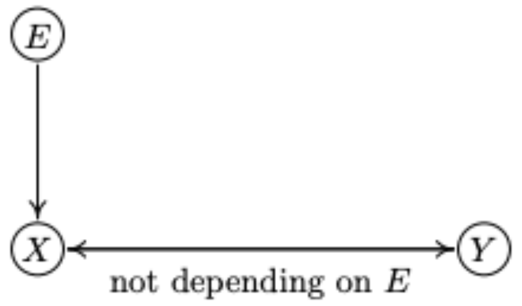
Causal inference by using invariant prediction: identification and confidence intervals



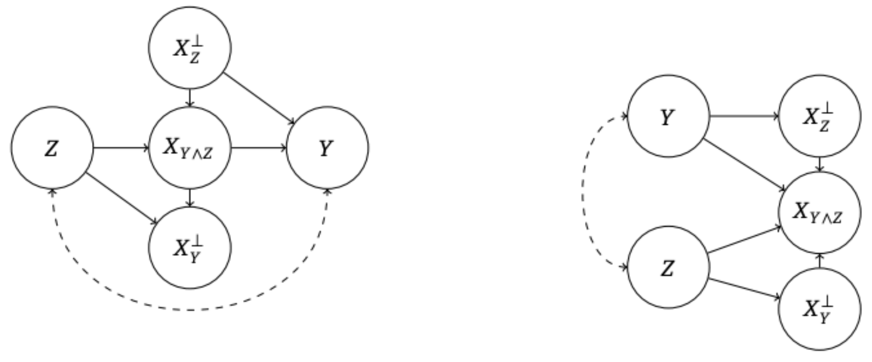
Invariant Models for Causal Transfer Learning



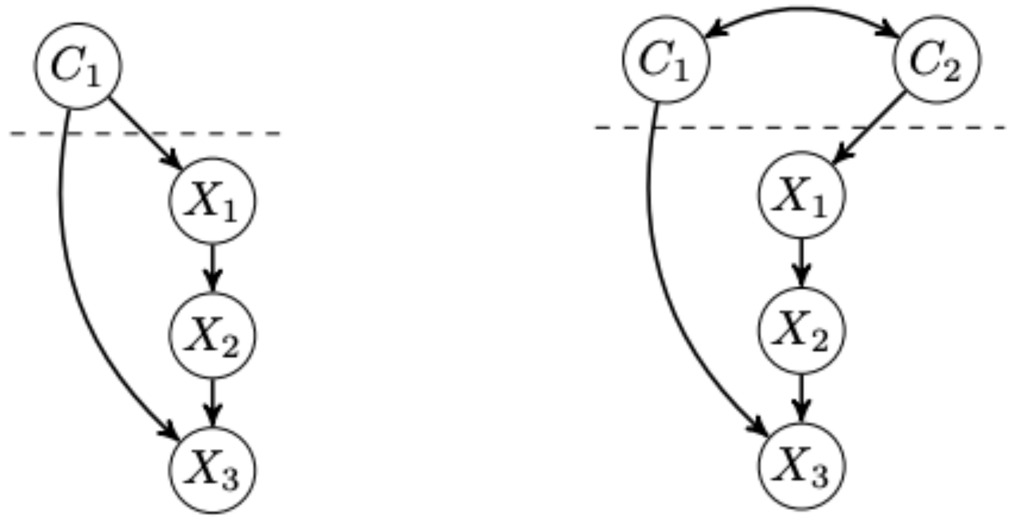
Invariance, Causality and Robustness



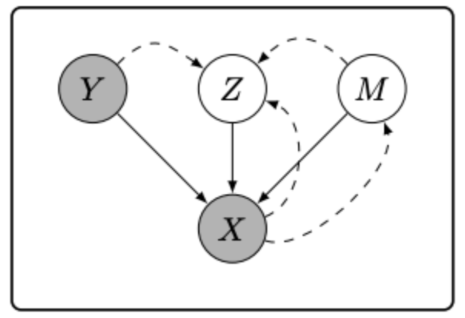
Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests



Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions

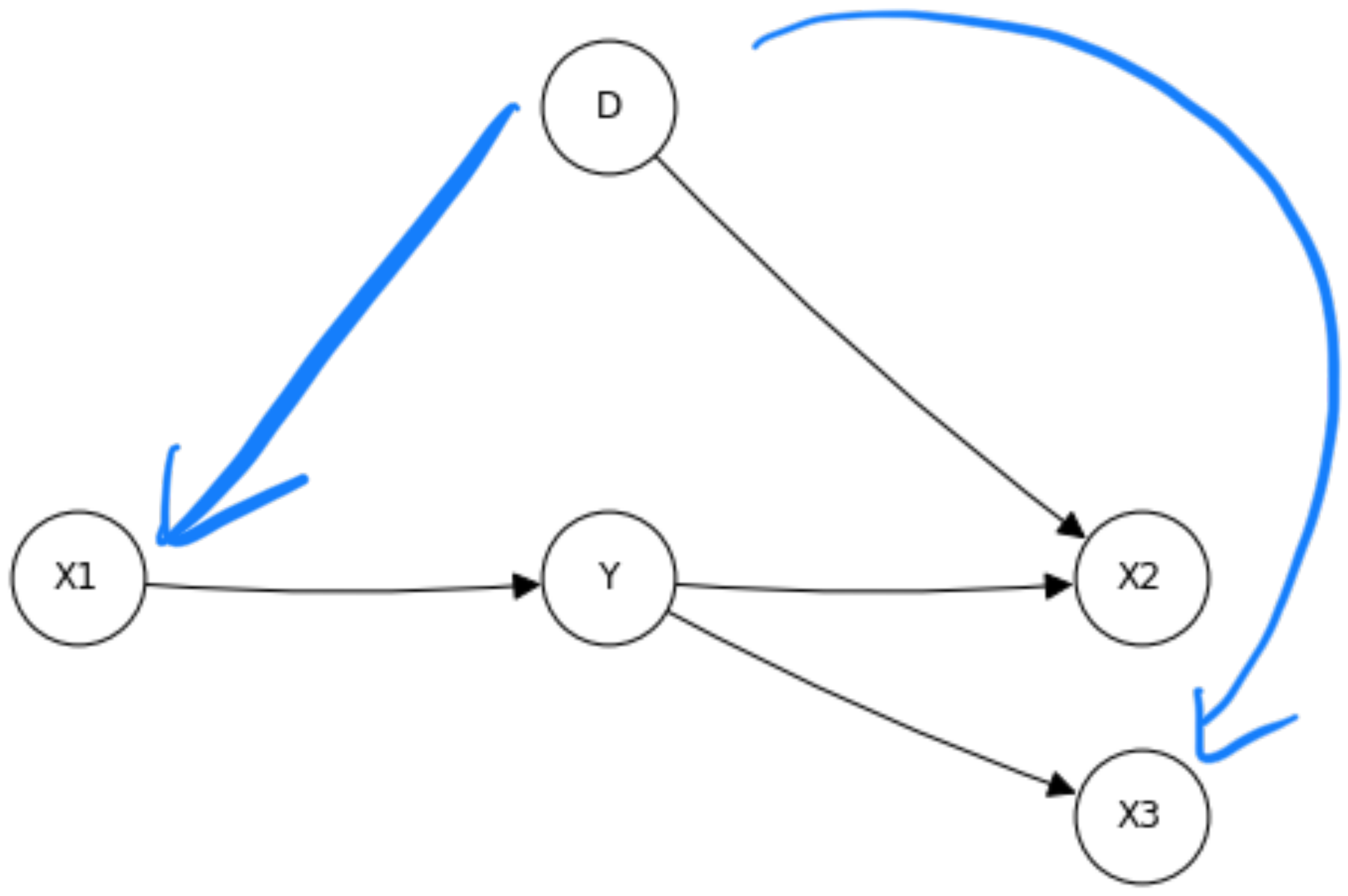


A Causal View on Robustness of Neural Networks



and many more....

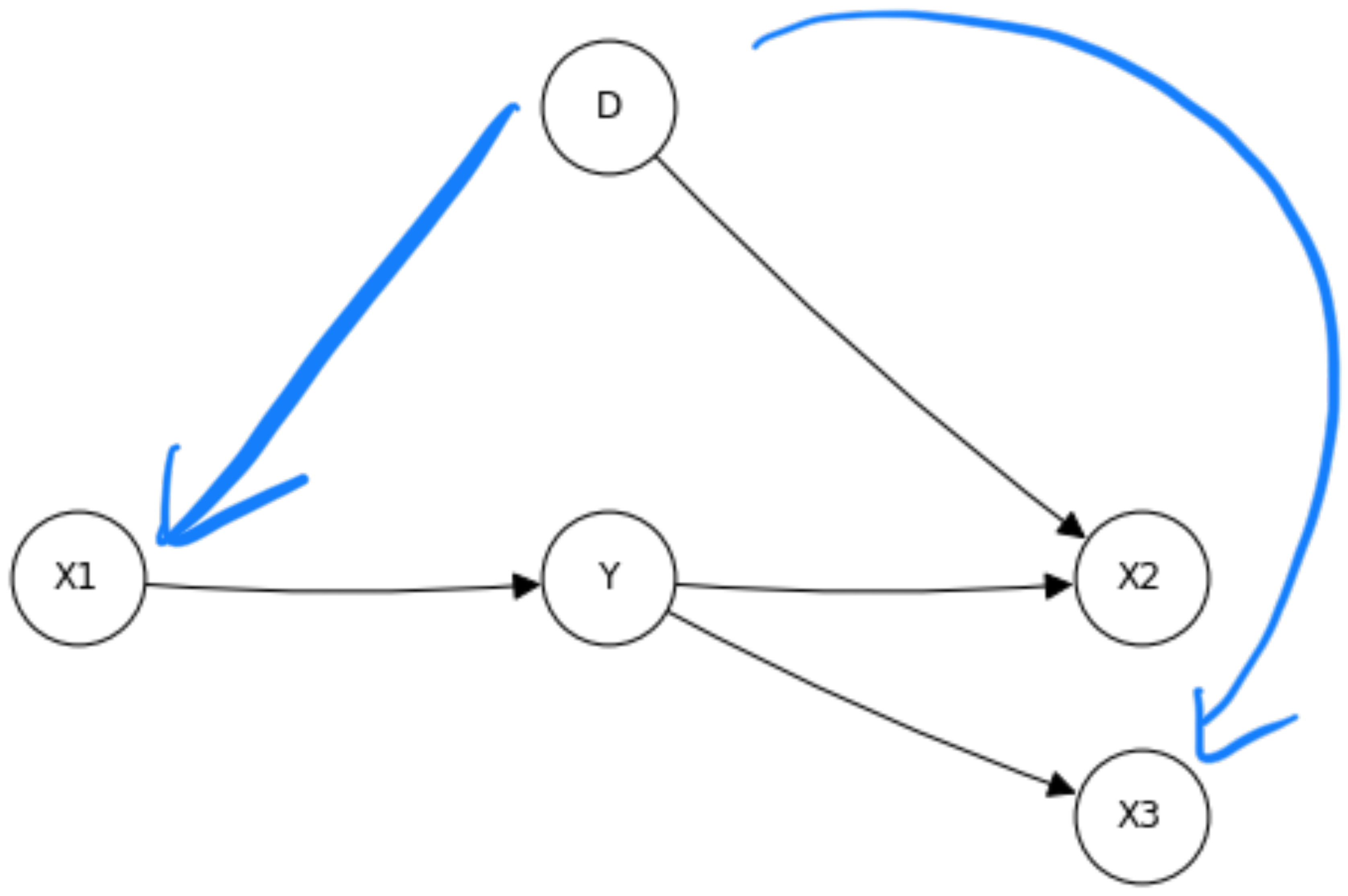
Which variables d-separate Y from D now?



X1	X2	X3	Y
?	?	?	?
?	?	?	?
?	?	?	?
….	….	….	….
2000	600	3000	-0,21
2190	450	3000	-0,16
2000	200	2999	-0,16
….	….	….	….
1200	1000	1500	-0,17
1201	800	1500	-0,14
1195	200	1499	-0,07
1340	900	1498	-0,14

$P(X_1, Y, X_2, X_3, D)$

Which variables d-separate Y from D now?

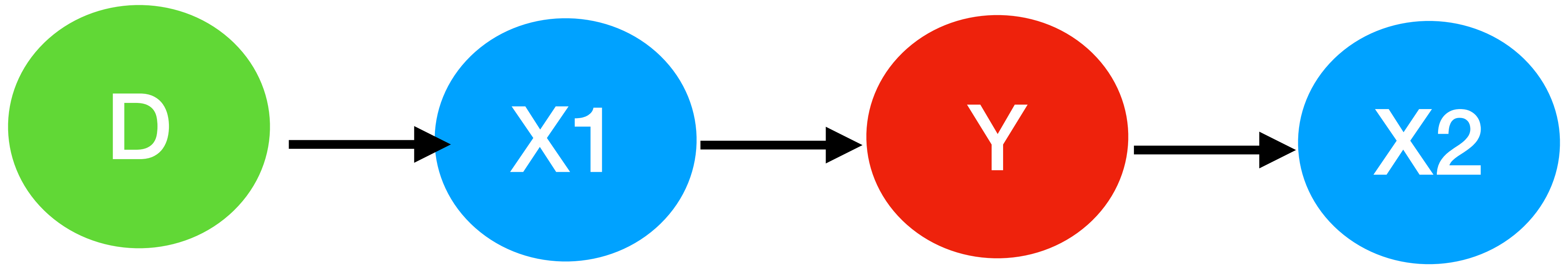


$P(X_1, Y, X_2, X_3, D)$

X1	X2	X3	Y
?	?	?	?
?	?	?	?
?	?	?	?
....	....	....	....
2000	600	3000	-0,21
2190	450	3000	-0,16
2000	200	2999	-0,16
....	....	....	....
1200	1000	1500	-0,17
1201	800	1500	-0,14
1195	200	1499	-0,07
1340	900	1498	-0,14

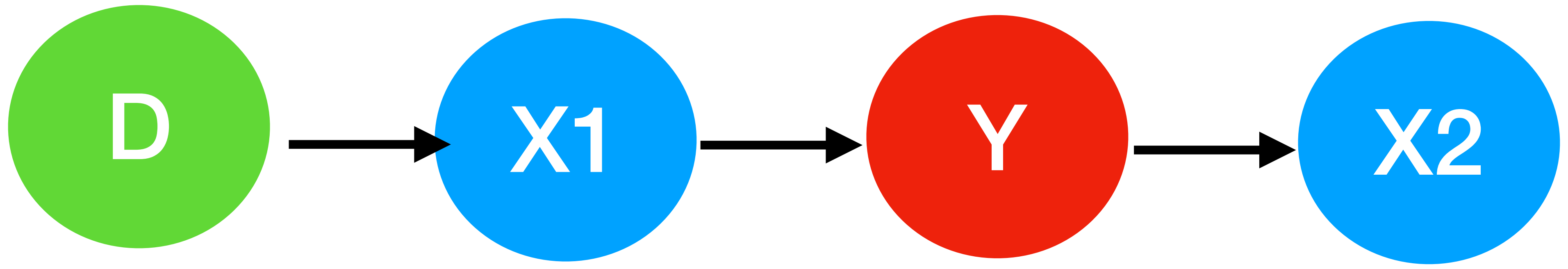
Intervention on every variable except Y = domain generalisation

Common misconceptions 1: An invariant feature need not be causal



- Which sets of variables d-separate Y from D? (List all)

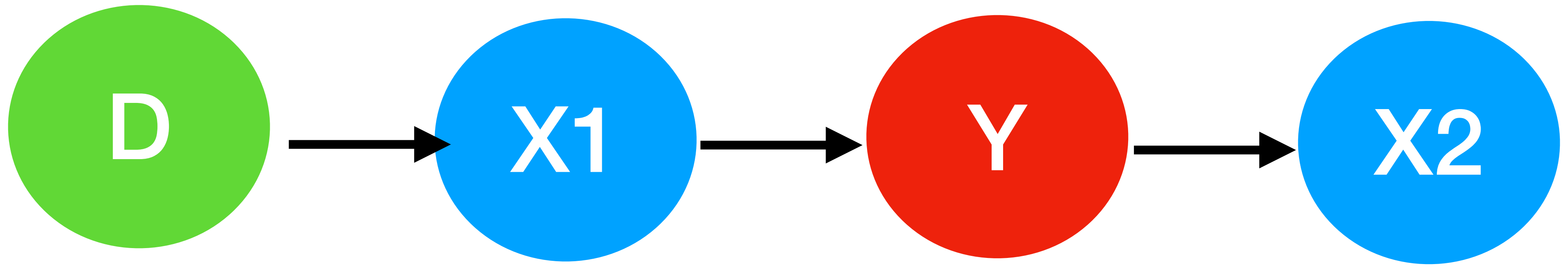
Common misconceptions 1: An invariant feature need not be causal



$$Y \perp\!\!\!\perp D \mid X_1$$

$$Y \perp\!\!\!\perp D \mid X_1, X_2$$

Common misconceptions 1: An invariant feature need not be causal



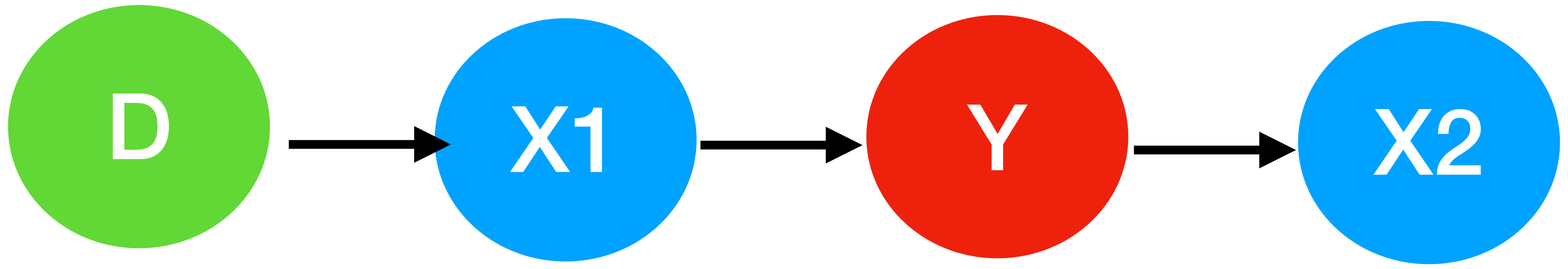
$$Y \perp\!\!\!\perp D \mid X_1$$

$$Y \perp\!\!\!\perp D \mid X_1, X_2$$

- Y|X1,X2 is invariant $\implies$ invariant features are not necessarily parents of Y

Invariant feature across "many different datasets" is not enough in general to find causal parents, need more assumptions

Common misconceptions 1: An invariant feature need not be causal



$$Y \perp\!\!\!\perp D \mid X_1$$

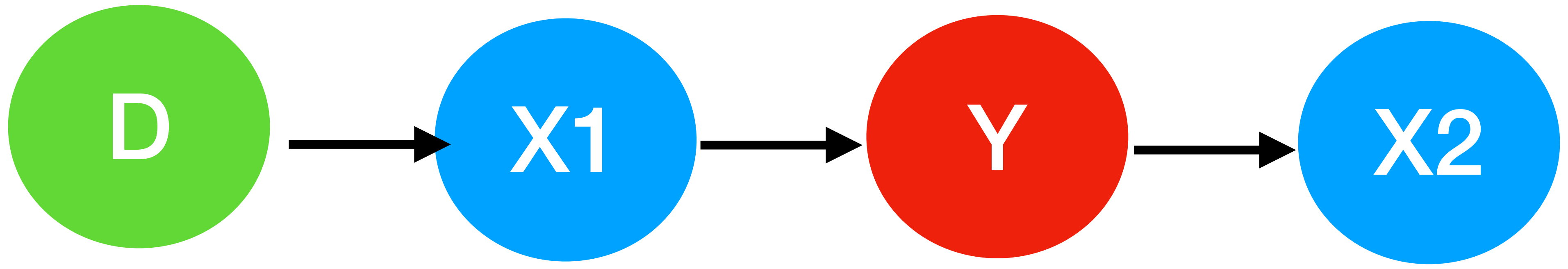
$$Y \perp\!\!\!\perp D \mid X_1, X_2$$

- Y|X1,X2 is invariant $\Longrightarrow$ invariant features are not necessarily parents of Y

Invariant feature across "many different datasets" is not enough in general to find causal parents, need more assumptions

- How do you get (some of) the parents?

Common misconceptions 1: An invariant feature need not be causal



$$Y \perp\!\!\!\perp D \mid X_1$$

$$Y \perp\!\!\!\perp D \mid X_1, X_2$$

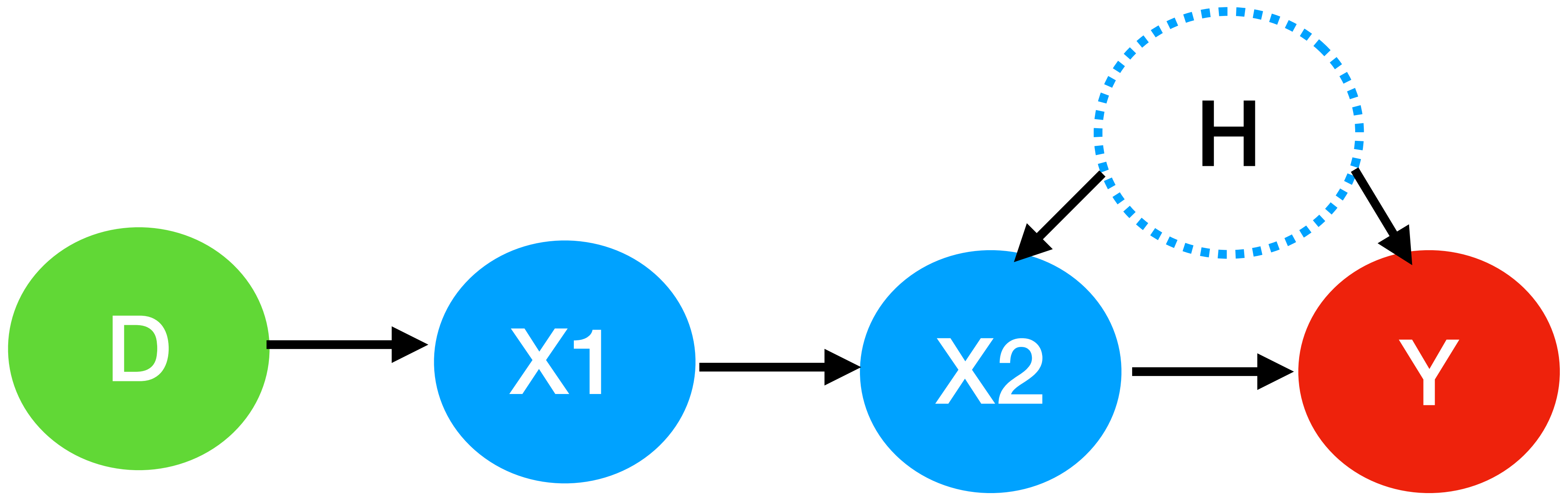
- Y|X1,X2 is invariant $\Longrightarrow$ invariant features are not necessarily parents of Y

Invariant feature across "many different datasets" is not enough in general to find causal parents, need more assumptions

- Invariant Causal Prediction [Peters et al. 2016] under causal sufficiency:

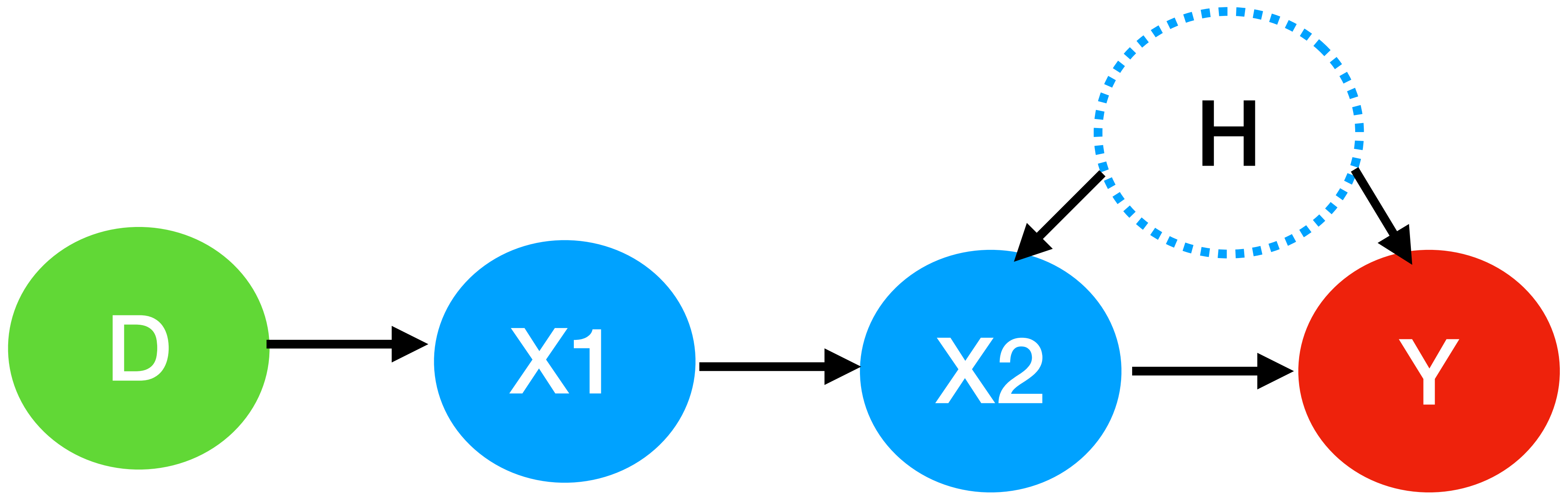
$$\mathbf{S}^* = \bigcap_{Y \perp\!\!\!\perp D \mid \mathbf{S}} \mathbf{S} \subseteq Pa(Y) \qquad \{X_1, X_2\} \cap \{X_1\} = \{X_1\}$$

Common misconception 2: Parents are not enough under latent confounding



- Which sets of variables d-separate Y from D? (List all)
  - Note: you cannot have H in the conditioning set, because it's latent

Common misconception 2: Parents are not enough under latent confounding



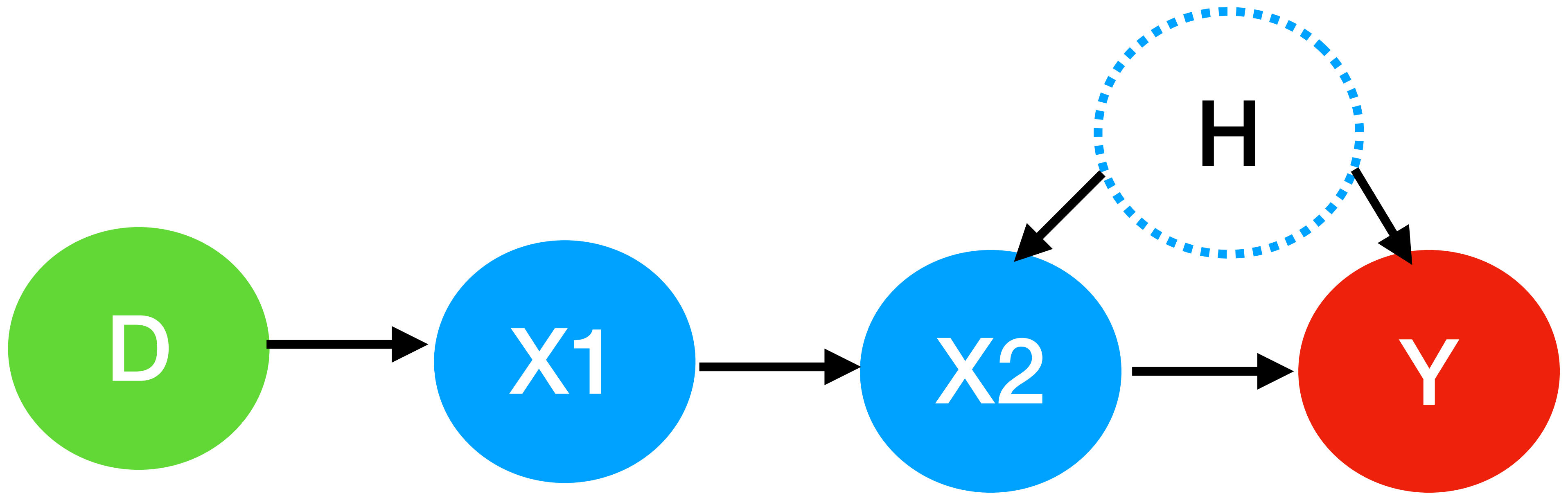
$Y \perp\!\!\!\perp D \mid X_1$

$Y \not\perp\!\!\!\perp D \mid X_2$

$Y \perp\!\!\!\perp D \mid X_1, X_2$

- Y|X1 is invariant, Y|X2 is not, even if X2 is a parent

Common misconception 2: Parents are not enough under latent confounding



$$Y \perp\!\!\!\perp D \mid X_1$$

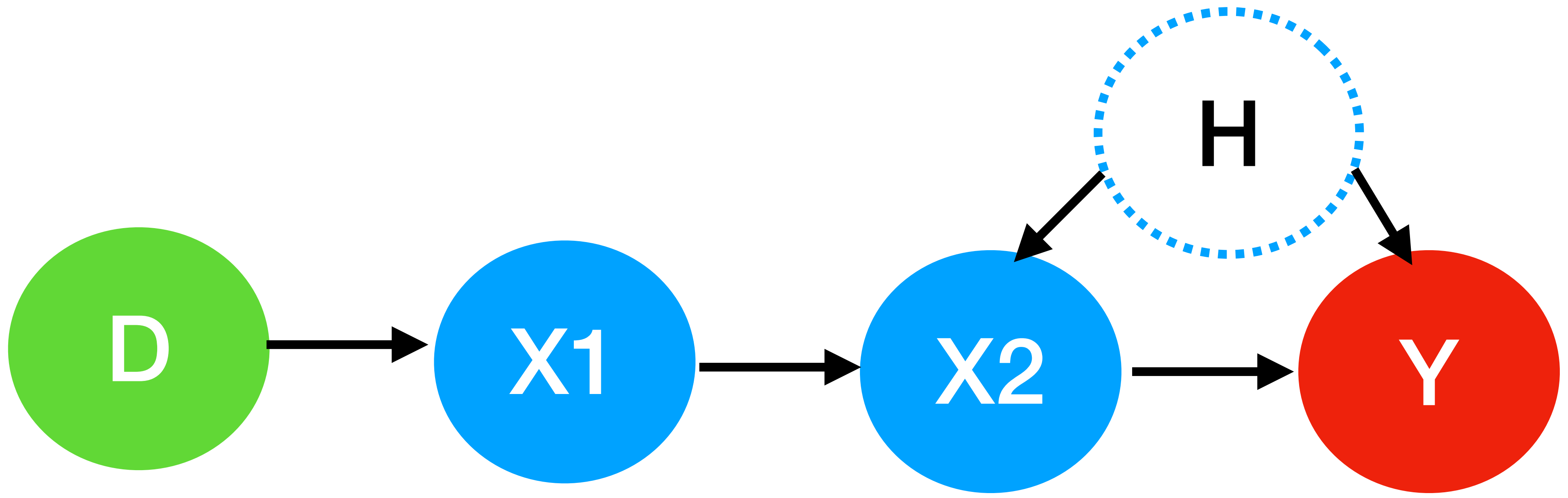
$$Y \not\!\perp\!\!\!\perp D \mid X_2$$

$$Y \perp\!\!\!\perp D \mid X_1, X_2$$

- Y|X1 is invariant, Y|X2 is not, even if X2 is a parent

Even if we knew all the parents, under latent confounding this wouldn't necessarily help transfer

Common misconception 2: Parents are not enough under latent confounding



$$Y \perp\!\!\!\perp D \mid X_1$$

$$Y \not\!\perp\!\!\!\perp D \mid X_2$$

$$Y \perp\!\!\!\perp D \mid X_1, X_2$$

- Y|X1 is invariant, Y|X2 is not, even if X2 is a parent

Even if we knew all the parents, under latent confounding this wouldn't necessarily help transfer

- **Conclusion:** causality (e.g. using the causal parents, learning the complete causal graph) is **neither necessary or sufficient*** for transfer, what we care about are **conditional independences/d-separations**

Takeaways 1/3

- Graphical models and d-separation [Pearl 1988] are a principled way to reason about **invariances and distribution shift**
  - Not a new observation, known since [Schoelkopf et al 2012]
  - Even with **unknown causal graphs**

Desiderata for a causal domain adaptation method

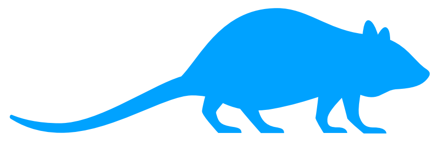
- **X,** Y and changes can be represented by an **unknown** causal graph
- Allow for **latent confounders**
- Avoid **parametric assumptions,** allow for heterogeneous effects across domains
- Instead of modeling **changes between each domain**, distinguish the change between the **mixture of sources and the target**

Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions

NeurIPS 2018

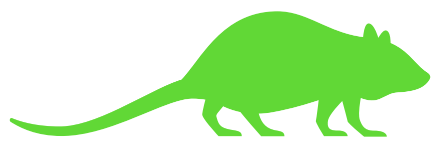
Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, Joris M. Mooij

- Unsupervised **multi-source** domain adaptation
- We interpret the change in the target domain as a **soft intervention**



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

	X1	X2	Y
Gene B	0,2	1	?
Gene B	0,3	1	?
Gene B	0,3	2	?
Gene B	0,4	1	?



	X1	X2	Y
Gene A	3,1	2	**1**
Gene A	3,2	3	**1**
Gene A	4	1	**1**
Gene A	3,2	3	**0**

Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Di

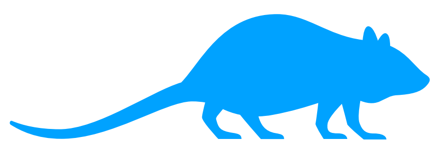
NeurIPS 2018

Sara Magliacane, Thijs van Ommen, Tom Claassen, Ste ooij

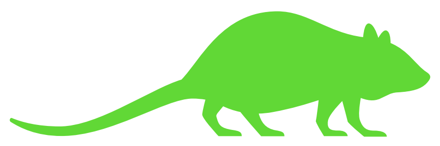
Multiple context variable C1, C2 …

- Unsupervised **multi-source** domain a

- We interpret the change in the target domain as a **soft intervention**



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



	X1	X2	Y
Gene A	3,1	2	1
Gene A	3,2	3	1
Gene A	4	1	1
Gene A	3,2	3	0

	X1	X2	Y
Gene B	0,2	1	?
Gene B	0,3	1	?
Gene B	0,3	2	?
Gene B	0,4	1	?

Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions

NeurIPS 2018

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, Joris M. Mooij

- Unsupervised **multi-source** domain ada

C1 = 1

- We interpret the change in the target domain as a **soft intervention**

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

	X1	X2	Y
Gene A	3,1	2	1
Gene A	3,2	3	1
Gene A	4	1	1
Gene A	3,2	3	0

	X1	X2	Y
	0,2	1	?
		1	?
			?
			?

Now the graph is unknown!

Joint Causal Inference from Multiple Contexts

Joris M. Mooij, Sara Magliacane, Tom Claassen

- We represent different distributions (including interventional) as an **unknown joint causal graph** (possibly cyclic or with latent confounders)
- We **add context variables** so we can **disentangle** changes in distribution across the datasets

https://arxiv.org/abs/1611.10351

Joint Causal Inference from Multiple Contexts

Joris M. Mooij, Sara Magliacane, Tom Claassen

- We represent different distributions (including interventional) as an **unknown joint causal graph** (possibly cyclic or with latent confounders)
- We **add context variables** so we can **disentangle** changes in distribution across the datasets

	X1	X2	X3
Normal	0,1	2	0
Normal	0,2	3	0
Normal	1,1	2	1
Normal	0,1	3	0

https://arxiv.org/abs/1611.10351

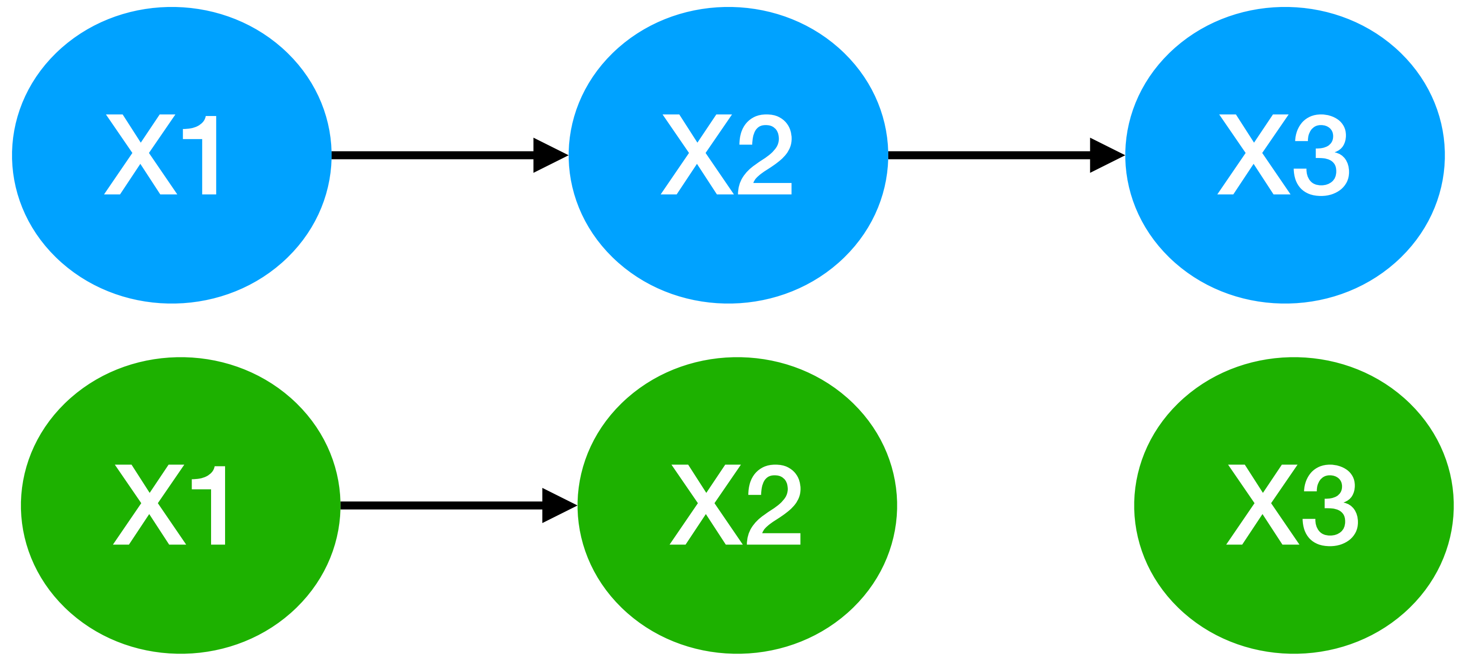
Joint Causal Inference from Multiple Contexts

Joris M. Mooij, Sara Magliacane, Tom Claassen

- We represent different distributions (including interventional) as an **unknown** **joint causal graph** (possibly cyclic or with latent confounders)
- We **add context variables** so we can **disentangle** changes in distribution across the datasets

	X1	X2	X3
Normal	0,1	2	0
Normal	0,2	3	0

	X1	X2	X3
Gene A	3,1	2	1
Gene A	3,2	3	1
Gene A	4	1	1
Gene A	3,2	3	1



$\mathrm{do}(X_3) = 1$

https://arxiv.org/abs/1611.10351

Joint Causal Inference from Multiple Contexts

Joris M. Mooij, Sara Magliacane, Tom Claassen

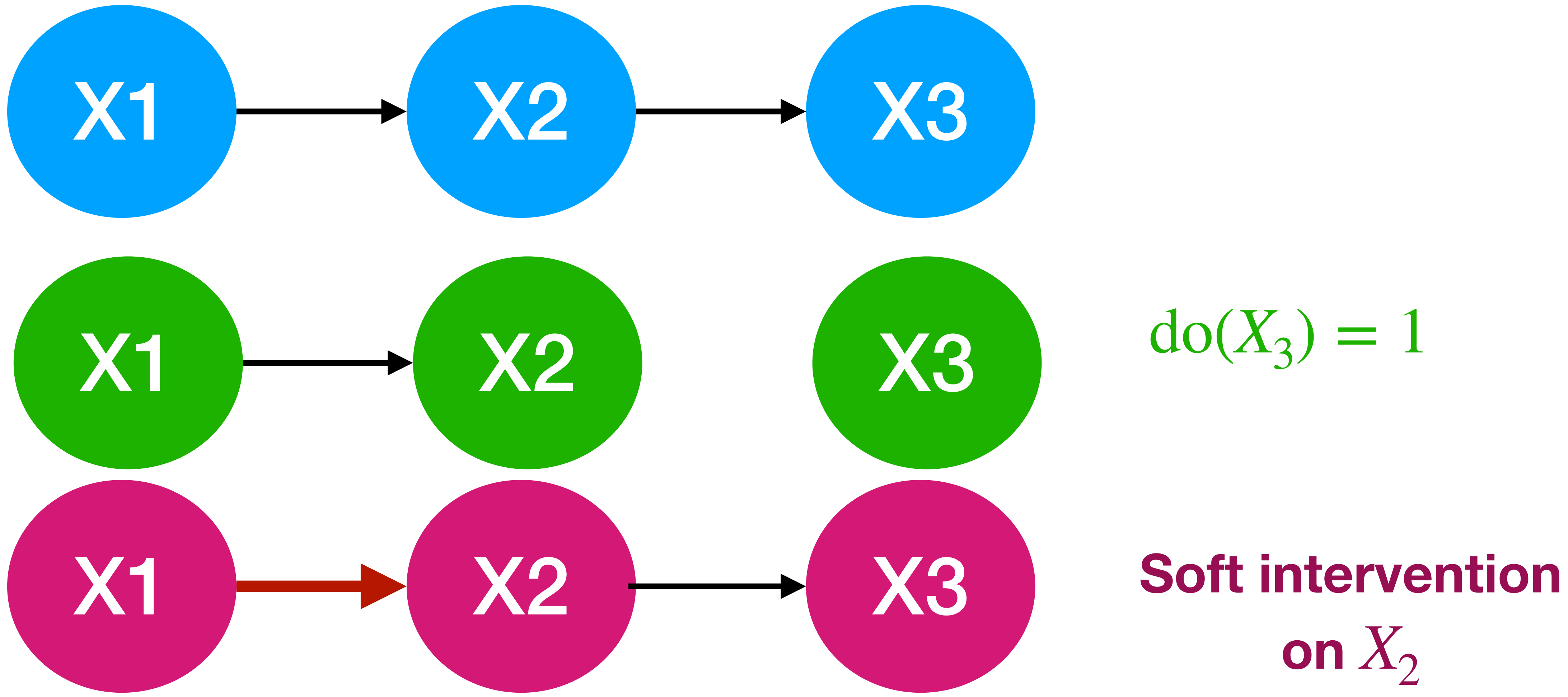
- We represent different distributions (including interventional) as an **unknown** **joint causal graph** (possibly cyclic or with latent confounders)

- We **add context variables** so we can **disentangle** changes in distribution across the datasets

	X1	X2	X3
Normal	0,1	2	0
Normal	0,2	3	0

	X1	X2	X3
Gene A	3,1	2	1
Gene A	3,2	3	1
			1
			1

	X1	X2	X3
Gene B	0.2	1	0
Gene B	0.3	1	1
Gene B	0.3	2	1
Gene B	0.4	1	1



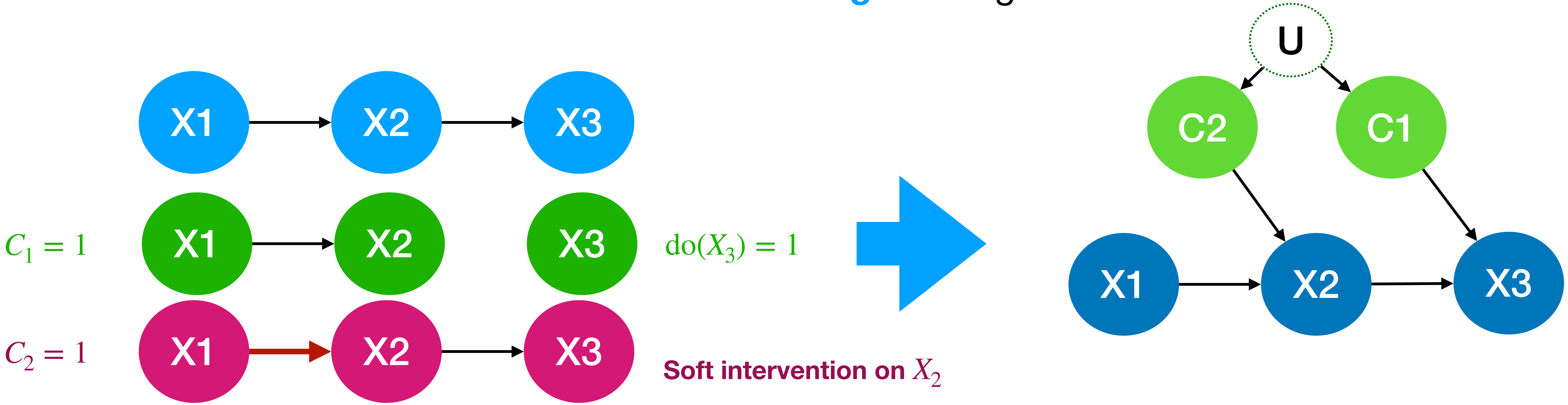
https://arxiv.org/abs/1611.10351

Joint Causal Inference from Multiple Contexts

Joris M. Mooij, Sara Magliacane, Tom Claassen

- We represent different distributions (including interventional) as an **unknown joint causal graph** (possibly cyclic or with latent confounders)

- We **add context variables** so we can **disentangle** changes in distribution across the datasets



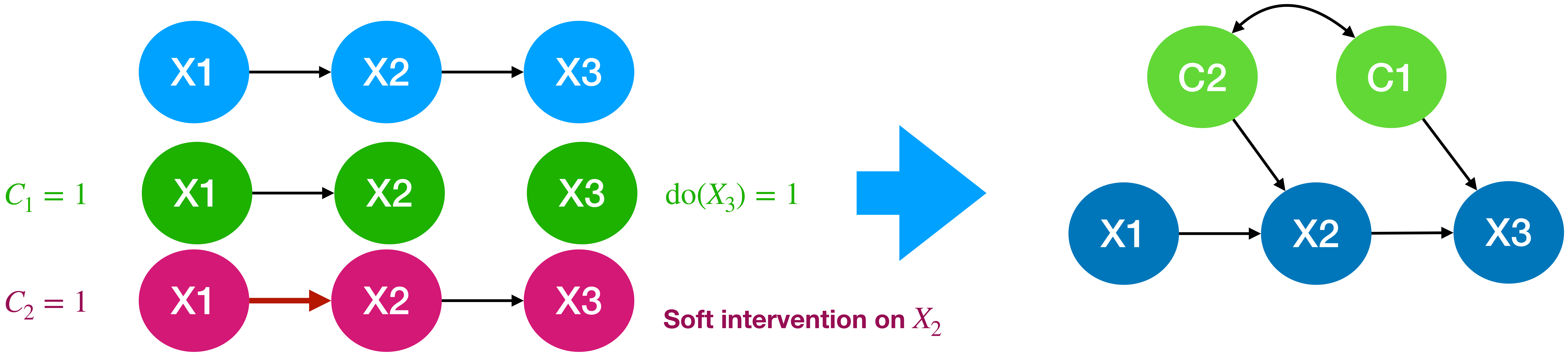
https://arxiv.org/abs/1611.10351

Joint Causal Inference from Multiple Contexts

Joris M. Mooij, Sara Magliacane, Tom Claassen

- We represent different distributions (including interventional) as an **unknown** **joint causal graph** (possibly cyclic or with latent confounders)

- We **add context variables** so we can **disentangle** changes in distribution across the datasets



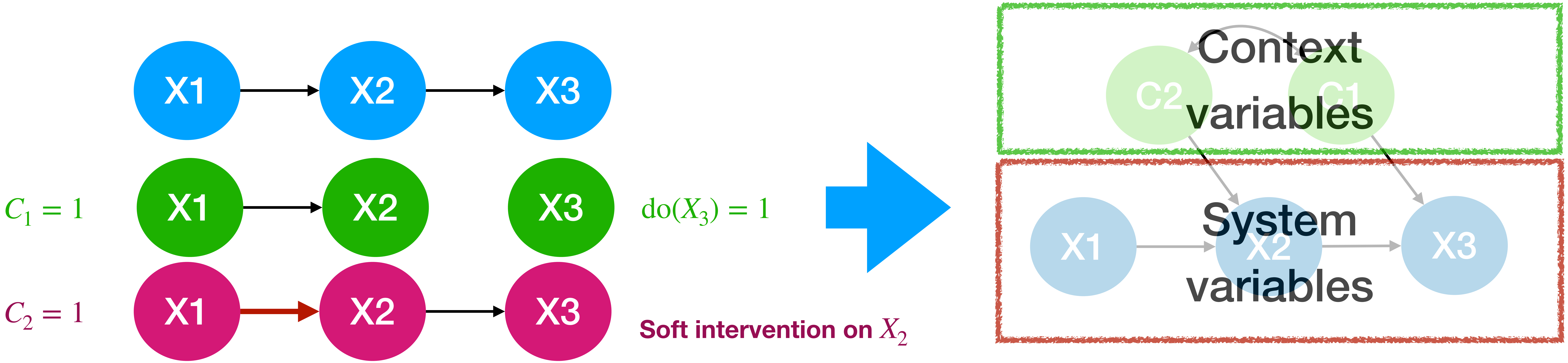
https://arxiv.org/abs/1611.10351

Joint Causal Inference from Multiple Contexts

Joris M. Mooij, Sara Magliacane, Tom Claassen

- We represent different distributions (including interventional) as an **unknown** **joint causal graph** (possibly cyclic or with latent confounders)

- We **add context variables** so we can **disentangle** changes in distribution across the datasets



https://arxiv.org/abs/1611.10351

Joint Causal Inference from Multiple Contexts

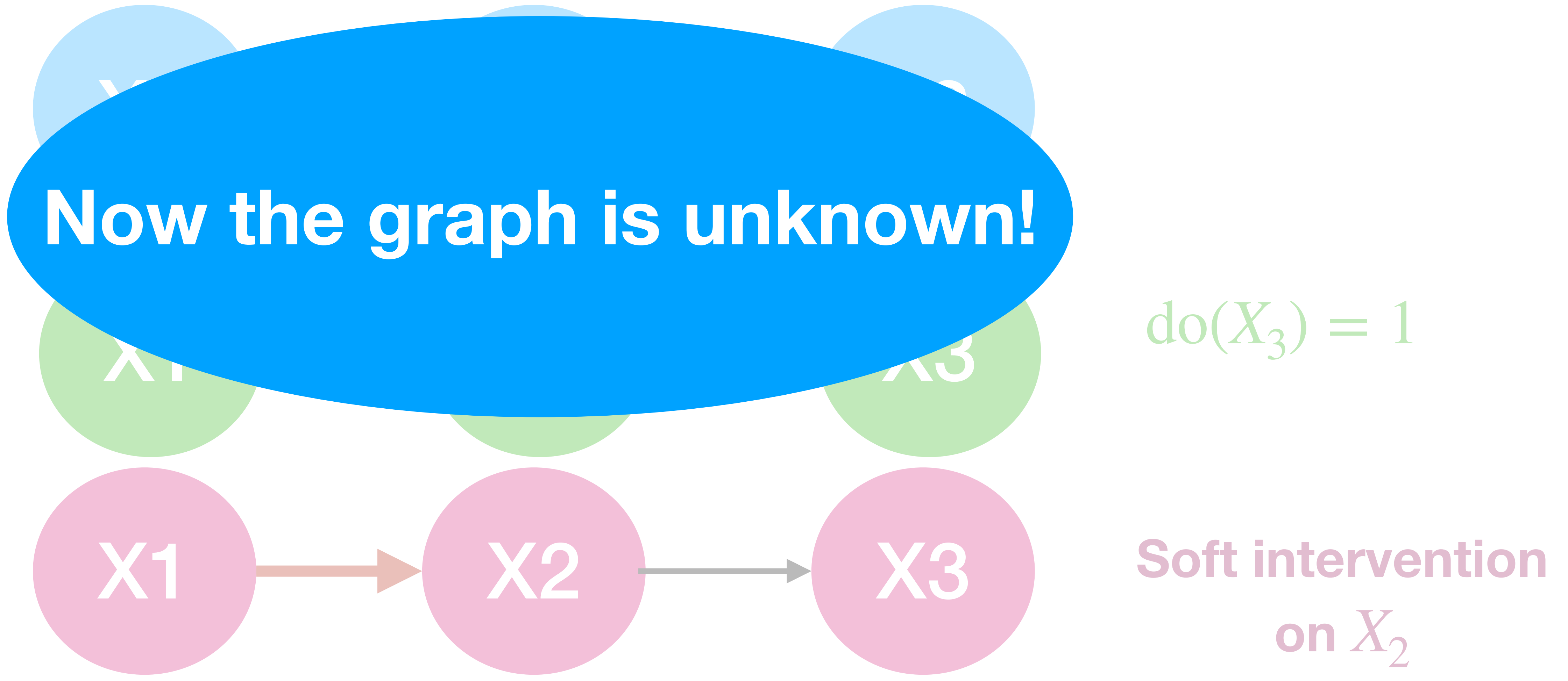
Joris M. Mooij, Sara Magliacane, Tom Claassen

- We represent different distributions (including interventional) as an **unknown joint causal graph** (possibly cyclic or with latent confounders)
- We **add context variables** so we can **disentangle** changes in distribution across the datasets

	X1	X2	X3
Normal	0,1	2	0
Normal	0,2	3	0

	X1	X2	X3
Gene A	3,1	2	1
Gene A	3,2	3	1
			1
			1

	X1	X2	X3
Gene B	0.2	1	0
Gene B	0.3	1	1
Gene B	0.3	2	1
Gene B	0.4	1	1



https://arxiv.org/abs/1611.10351

Joint Causal Inference from Multiple Contexts

Joris M. Mooij, Sara Magliacane, Tom Claassen

- We represent different distributions (including interventional) as an **unknown joint causal graph** (possibly cyclic or with latent confounders)

- We **add context variables** so we can **disentangle** changes in distribution across the datasets

	X1	X2	X3
Normal	0,1	2	0
Normal	0,2	3	0

	X1	X2	X3
Gene A	3,1	2	1
Gene A	3,2	3	1
			1
			1

	X1	X2	X3
Gene B	0.2	1	0
Gene B	0.3	1	1
Gene B	0.3	2	1
Gene B	0.4	1	1

C1	C2	X1	X2	X3
0	0	0,1	2	0
0	0	0,2	3	0
0	0	1,1	2	1
0	0	0,1	3	0
1	0	3,1	2	1
1	0	3,2	3	1
1	0	4	1	1
1	0	3,2	3	1
0	1	0,2	1	0
0	1	0,3	1	1
0	1	0,3	2	1
0	1	0,4	1	1

https://arxiv.org/abs/1611.10351

Joint Causal Inference from Multiple Contexts

Joris M. Mooij, Sara Magliacane, Tom Claassen

- We represent different distributions (including interventional) as an **unknown** **joint causal graph** (possibly cyclic or with latent confounders)

- We **add context variables** so we can **disentangle** changes in distribution across the datasets

	X1	X2	X3
Normal	0,1	2	0
Normal	0,2	3	0

	X1	X2	X3
Gene A	3,1	2	1
Gene A	3,2	3	1
			1
			1

	X1	X2	X3
Gene B	0.2	1	0
Gene B	0.3	1	1
Gene B	0.3	2	1
Gene B	0.4	1	1

Context variables

System variables

C1	C2	X1	X2	X3
0	0	0,1	2	0
0	0	0,2	3	0
0	0	1,1	2	1
0	0	0,1	3	0
1	0	3,1	2	1
1	0	3,2	3	1
1	0	4	1	1
1	0	3,2	3	1
0	1	0,2	1	0
0	1	0,3	1	1
0	1	0,3	2	1
0	1	0,4	1	1

https://arxiv.org/abs/1611.10351

Joint Causal Inference from Multiple Contexts

Joris M. Mooij, Sara Magliacane, Tom Claassen

- We **add context variables** so we can **disentangle** changes in distribution across the datasets (and optionally background knowledge, e.g. context variables are uncaused)

- We can reuse **any standard method for observational data** that fits any chosen assumptions

C1	C2	X1	X2	X3
0	0	0,1	2	0
0	0	0,2	3	0
0	0	1,1	2	1
0	0	0,1	3	0
1	0	3,1	2	1
1	0	3,2	3	1
1	0	4	1	1
1	0	3,2	3	1
0	1	0,2	1	0
0	1	0,3	1	1
0	1	0,3	2	1
0	1	0,4	1	1

https://arxiv.org/abs/1611.10351

Joint Causal Inference from Multiple Contexts

Joris M. Mooij, Sara Magliacane, Tom Claassen

- We **add context variables** so we can **disentangle** changes in distribution across the datasets (and optionally background knowledge, e.g. context variables are uncaused)

- We can reuse **any standard method for observational data** that fits any chosen assumptions

C1	C2	X1	X2	X3
0	0	0,1	2	0
0	0	0,2	3	0
0	0	1,1	2	1
0	0	0,1	3	0
1	0	3,1	2	1
1	0	3,2	3	1
1	0	4	1	1
1	0	3,2	3	1
0	1	0,2	1	0
0	1	0,3	1	1
0	1	0,3	2	1
0	1	0,4	1	1

$X_2 \not\perp\!\!\!\perp C_2$

$X_1 \perp\!\!\!\perp C_2 \mid C_1$

$X_2 \perp\!\!\!\perp C_1 \mid X_3$

$\ldots$

https://arxiv.org/abs/1611.10351

Joint Causal Inference from Multiple Contexts

Joris M. Mooij, Sara Magliacane, Tom Claassen

- We **add context variables** so we can **disentangle** changes in distribution across the datasets (and optionally background knowledge, e.g. context variables are uncaused)

- We can reuse **any standard method for observational data** that fits any chosen assumptions

C1	C2	X1	X2	X3
0	0	0,1	2	0
0	0	0,2	3	0
0	0	1,1	2	1
0	0	0,1	3	0
1	0	3,1	2	1
1	0	3,2	3	1
1	0	4	1	1
1	0	3,2	3	1
0	1	0,2	1	0
0	1	0,3	1	1
0	1	0,3	2	1
0	1	0,4	1	1

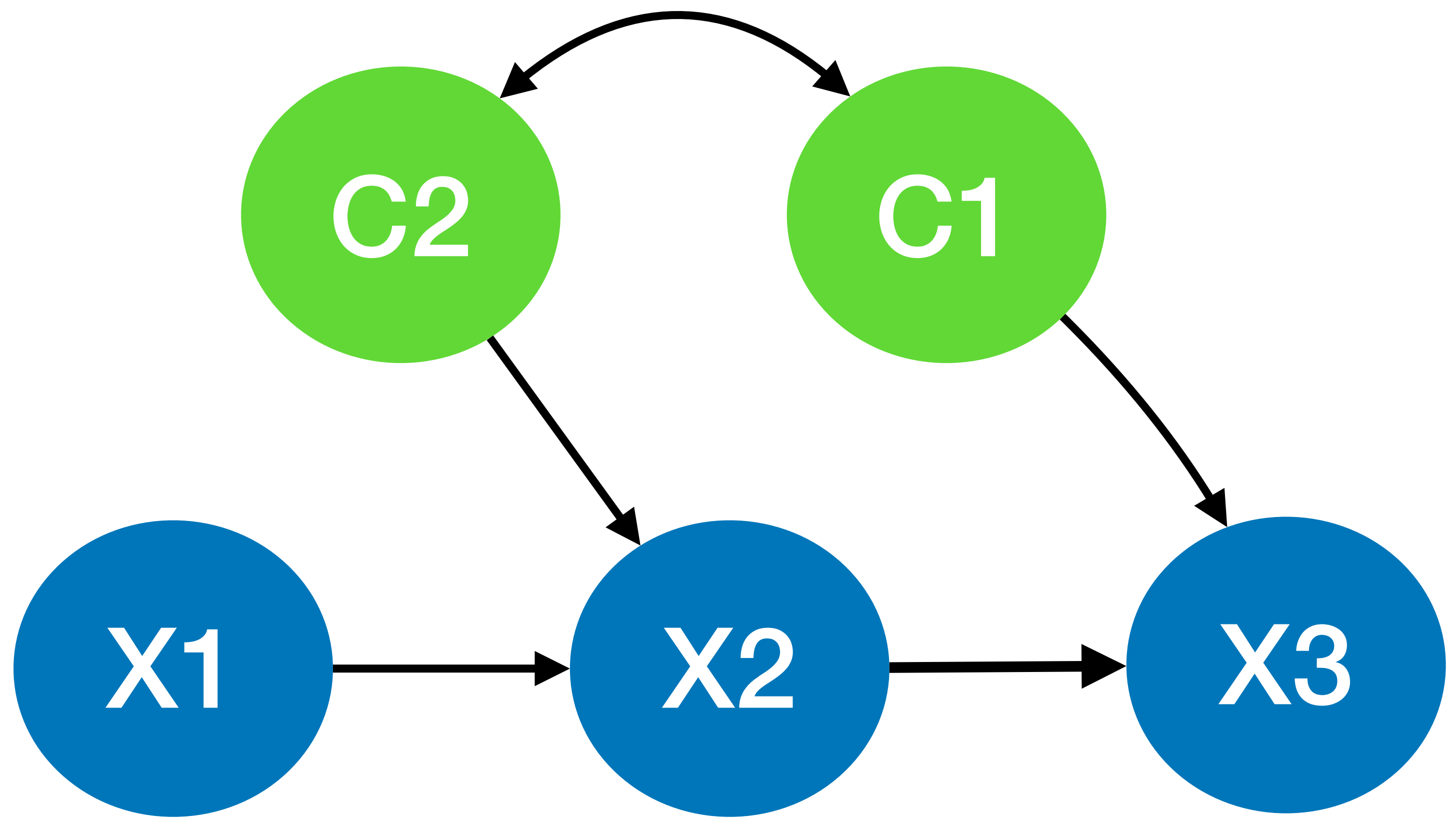
$X_2 \not\!\perp\!\!\!\perp C_2$

$X_1 \perp\!\!\!\perp C_2 \mid C_1$

$X_2 \perp\!\!\!\perp C_1 \mid X_3$

$\dots$

FCI-JCI



https://arxiv.org/abs/1611.10351

Joint Causal Inference from Multiple Contexts

Joris M. Mooij, Sara Magliacane, Tom Claassen

- Additional background knowledge based on assumptions:
  - A1 ("exogeneity"): No system variable causes any context variable.
  - A2 ("complete randomised context"): No context variable is confounded with a system variable.
  - A3 ("generic context model"): The context variables do not cause each other and they are assumed to be confounded.

https://arxiv.org/abs/1611.10351

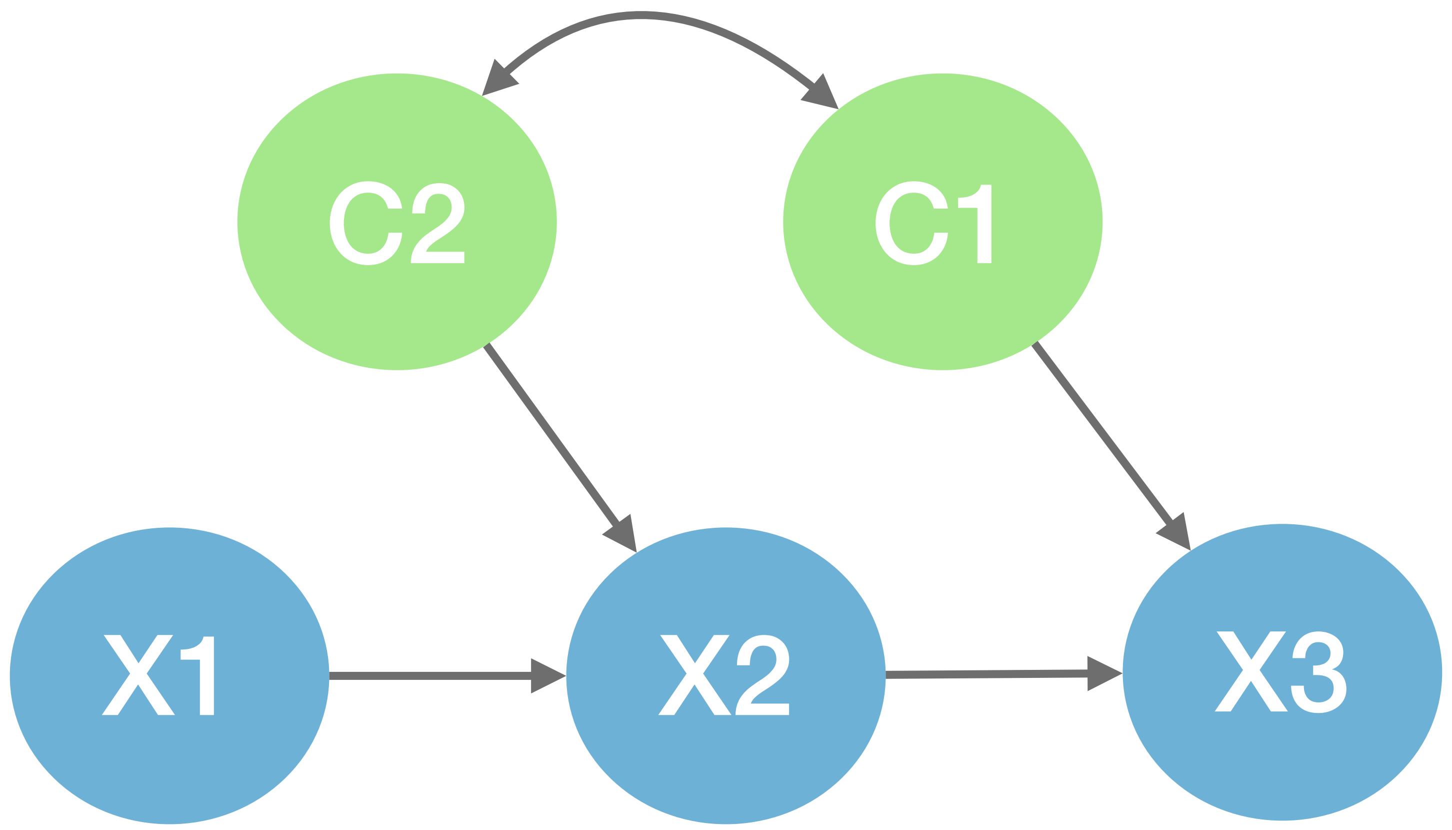
Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions NeurIPS 2018

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, Joris M. Mooij

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

Source domains

Target domain



- We assume we can model all the domains in with a **single unknown acyclic causal graph** with **multiple context variables** [Mooij et al. 2020]

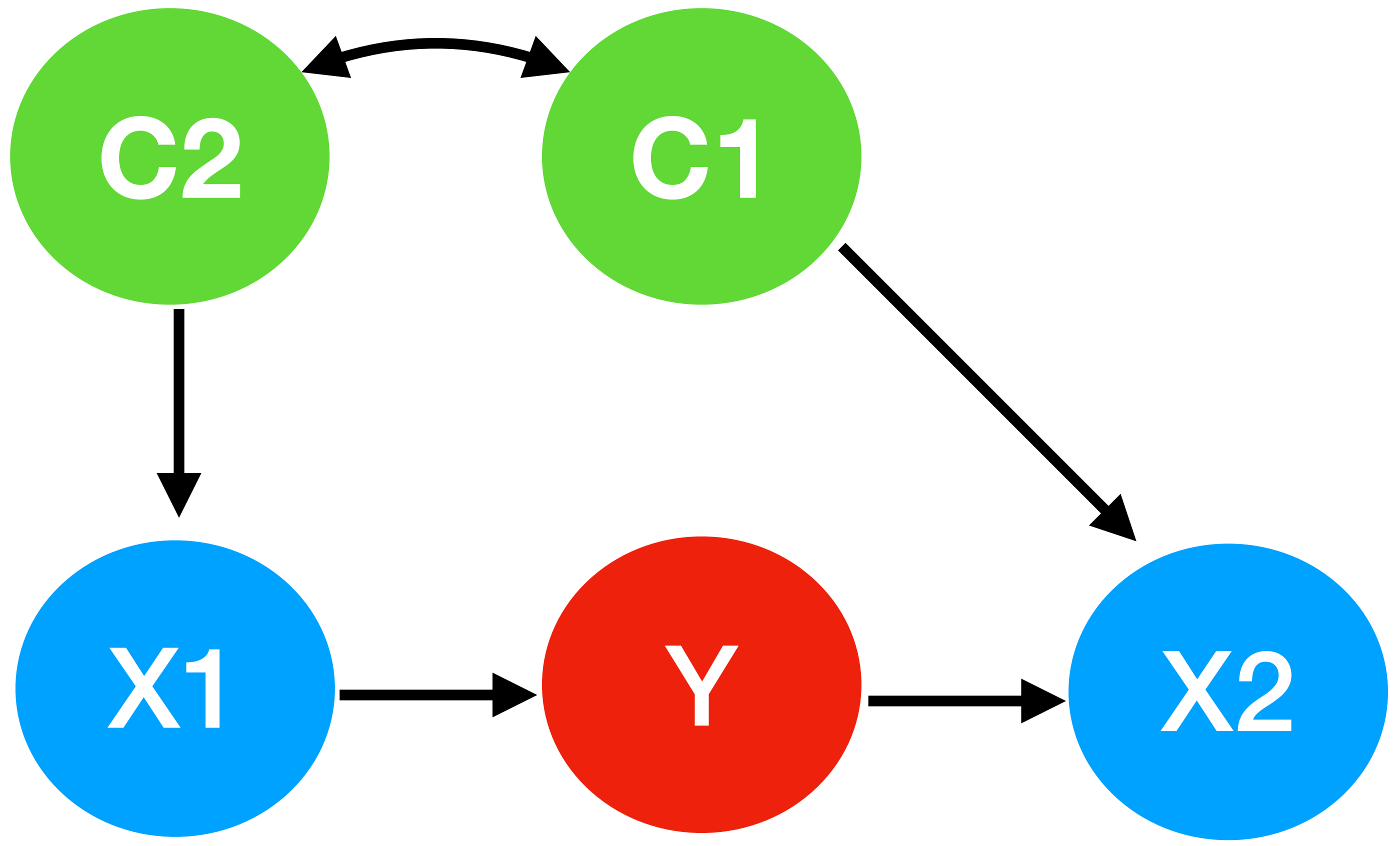
https://arxiv.org/abs/1707.06422

Causal domain adaptation: separating features

Look for features $S \subseteq X$ $\quad Y \perp_d D \mid S$

Aka stable features, invariant features etc.

- **Separating features:** sets of features that d-separate Y from the context variable **C1** representing the target domain



$$Y \perp\!\!\!\perp C_1 \mid S?$$

- **{X1}** is a separating feature set, {X1, X2} could lead to arbitrary large error

What if the causal graph is unknown?

- **Idea**: we could test the conditional independence in the data

$$Y \perp\!\!\!\perp C_1 \mid X_1 ? \qquad Y \perp\!\!\!\perp C_1 \mid X_2 ?$$

What if the causal graph is unknown?

- **Idea**: we could test the conditional independence in the data

$$\cancel{Y \perp\!\!\!\perp C_1 \mid X_1 ?} \qquad \cancel{Y \perp\!\!\!\perp C_1 \mid X_2 ?}$$

- **Problem:** Y is always missing when C1=1, so we cannot test these

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
1	0	3,1	2	?
1	0	3,2	3	?
1	0	4	3	?
0	1	0,2	0	0
0	1	0,3	0	1
0	1	0,3	1	0

What if the causal graph is unknown?

- **Idea**: we could test the conditional independence in the data

$$\cancel{Y \perp\!\!\!\perp C_1 \mid X_1 ?} \qquad \cancel{Y \perp\!\!\!\perp C_1 \mid X_2 ?}$$

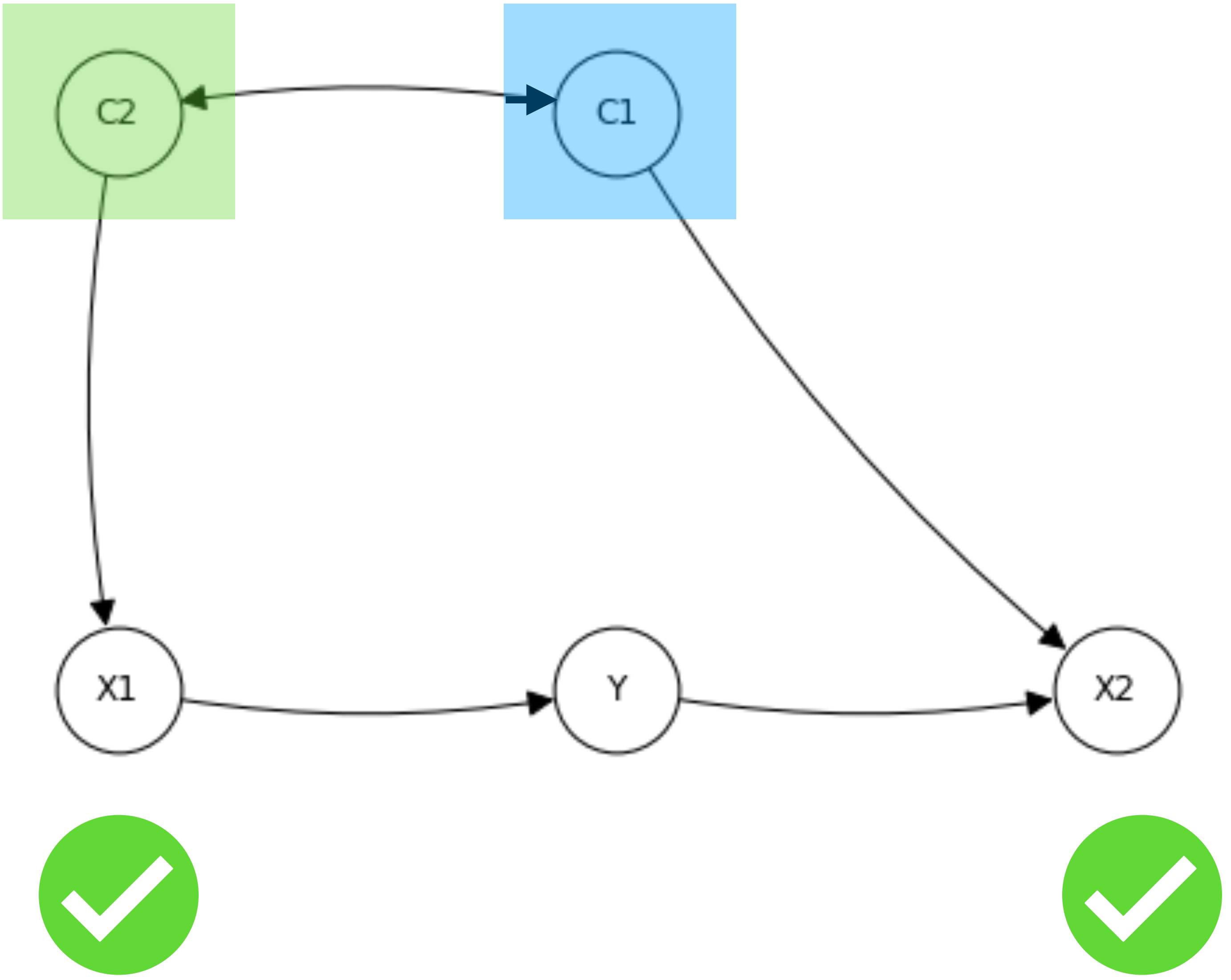
- **Problem:** Y is always missing when C1=1, so we cannot test these

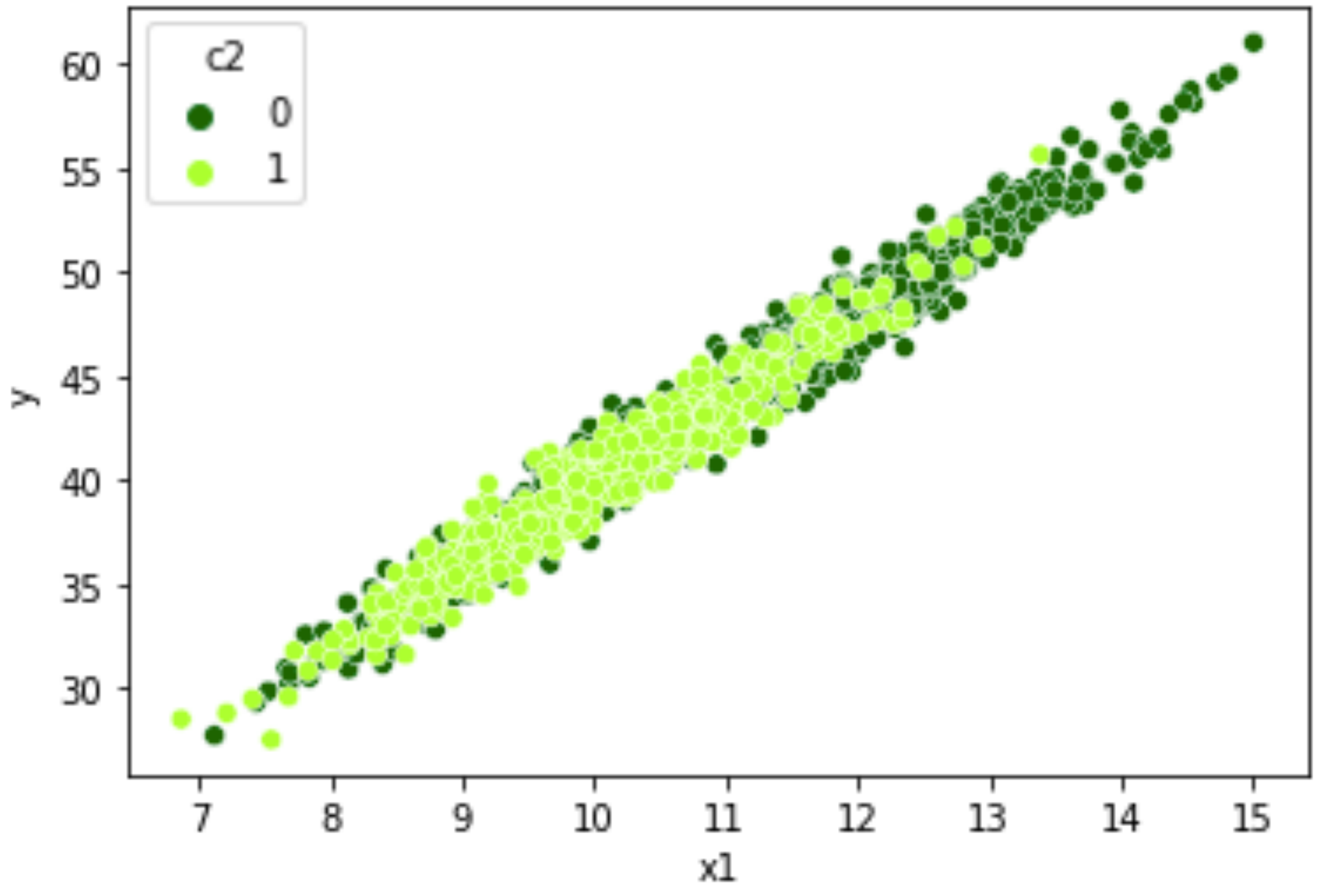
C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
1	0	3,1	2	?
1	0	3,2	3	?
1	0	4	3	?
0	1	0,2	0	0
0	1	0,3	0	1
0	1	0,3	1	0

Idea: Invariant features in source domains are also separating in the target domain

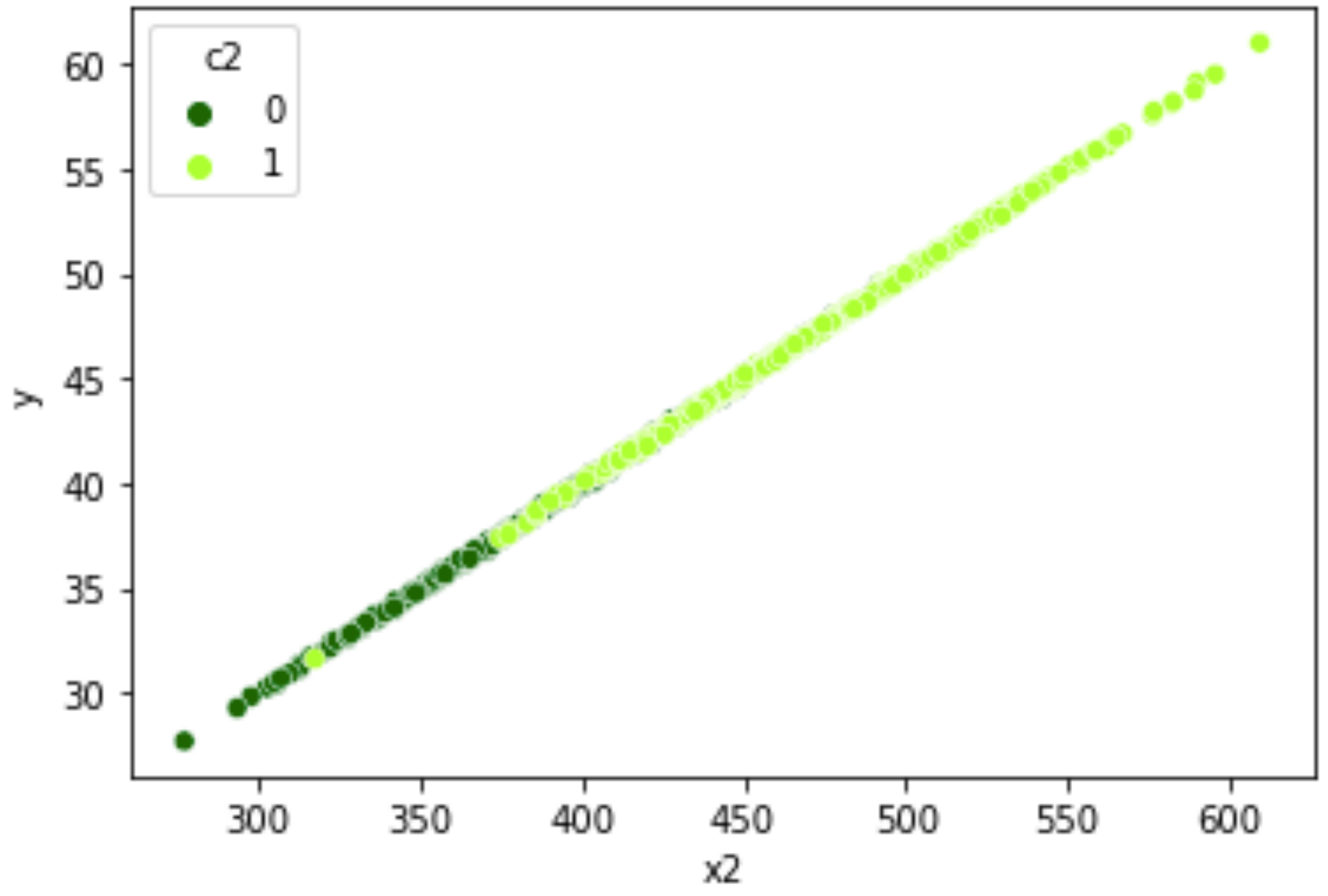
$$Y \perp\!\!\!\perp C_2 \mid \{X_1, C_1 = 0\} \implies Y \perp\!\!\!\perp C_1 \mid X_1$$

Separating features in sources are also separating in target - counterexample



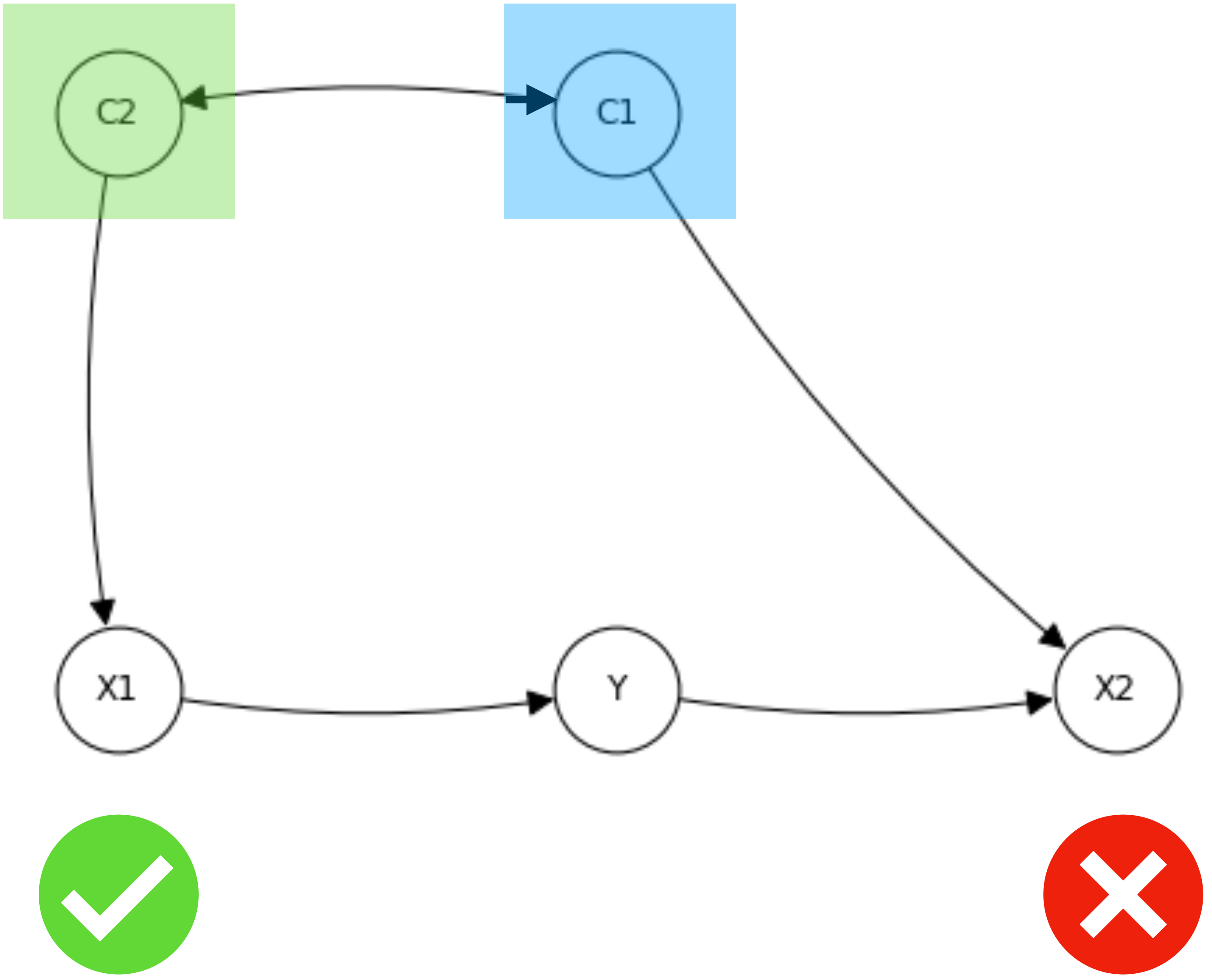


$$Y \perp\!\!\!\perp C_2 \mid \{X_1, C_1 = 0\}$$

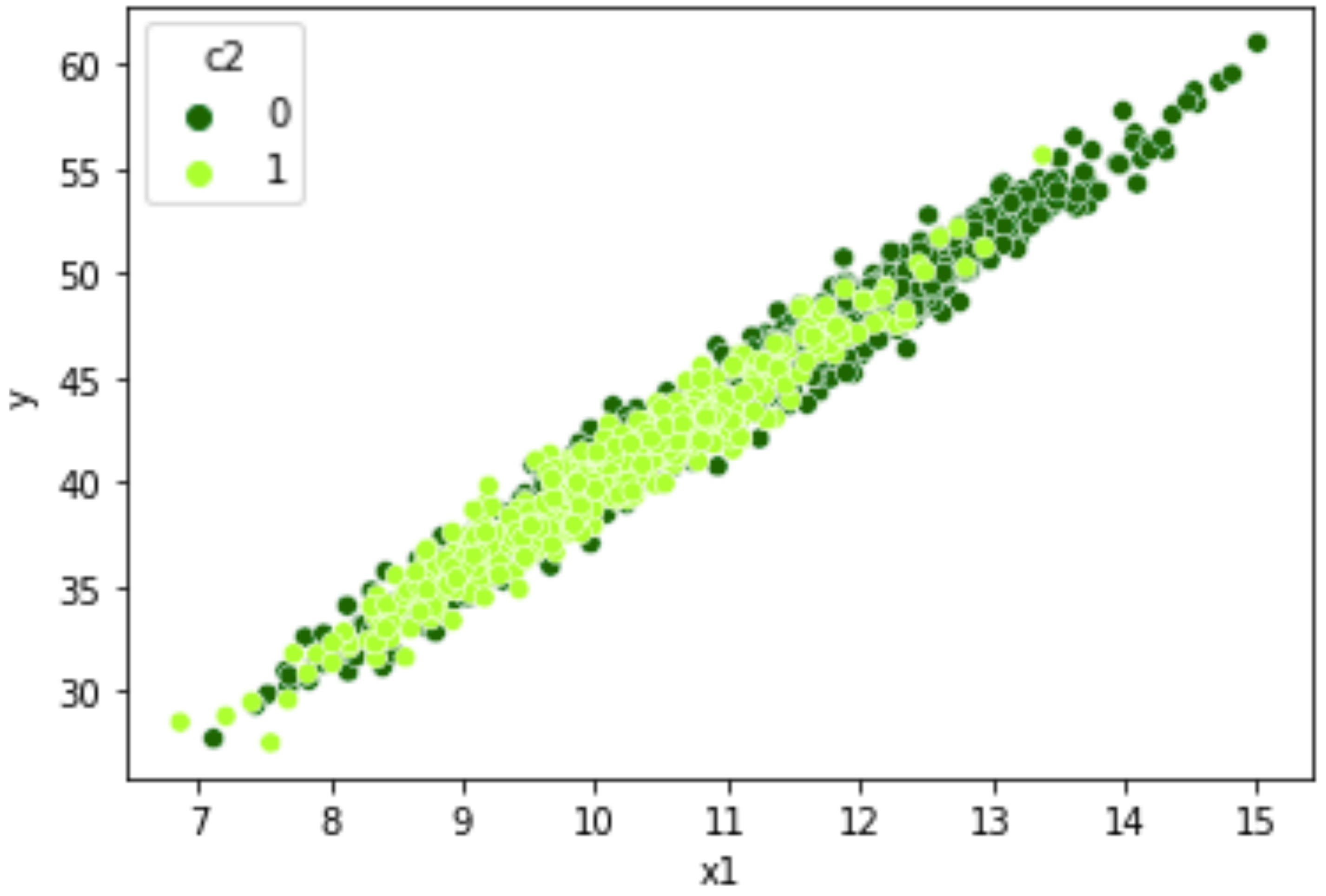


$$Y \perp\!\!\!\perp C_2 \mid \{X_2, C_1 = 0\}$$

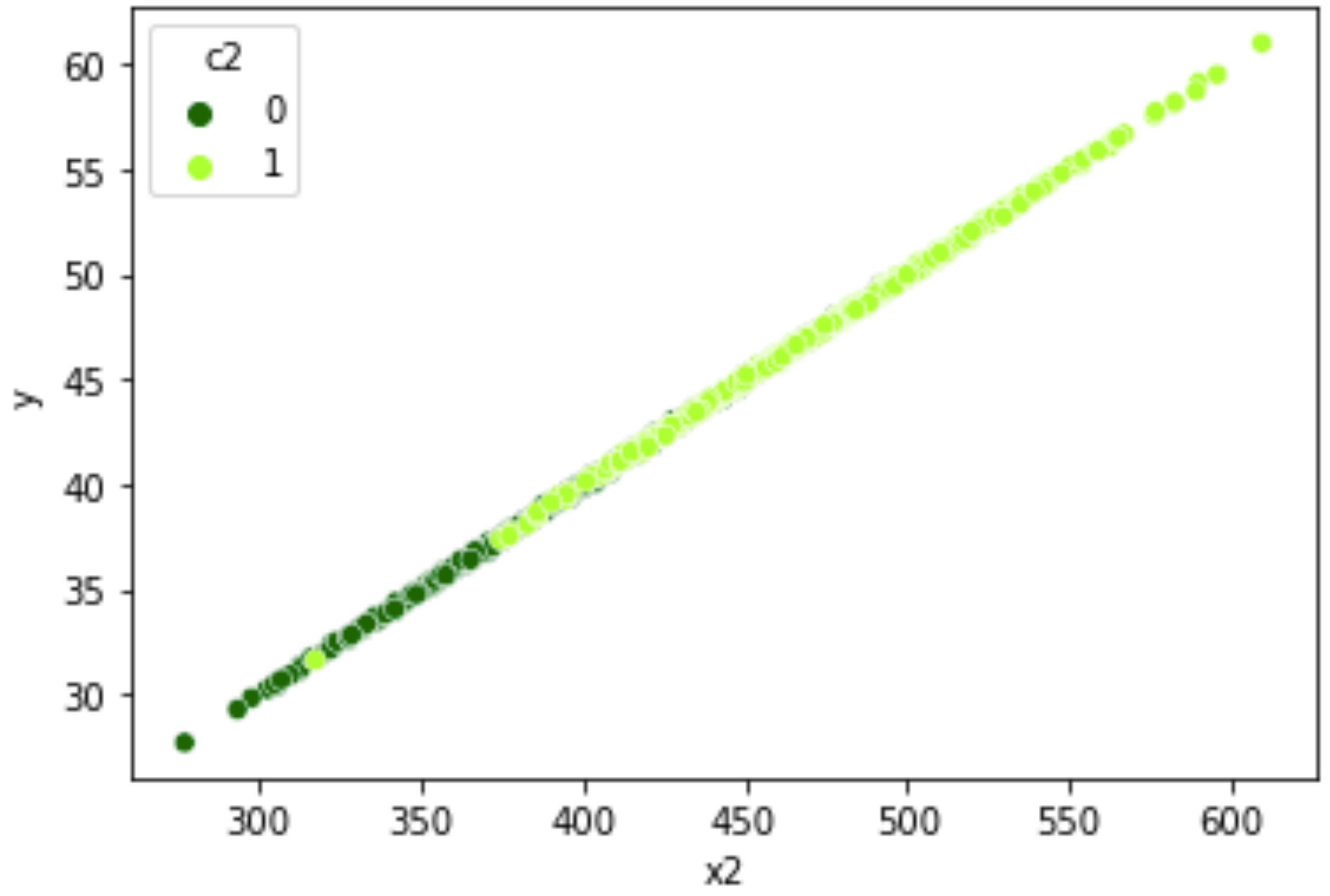
Separating features in sources are also separating in target - counterexample

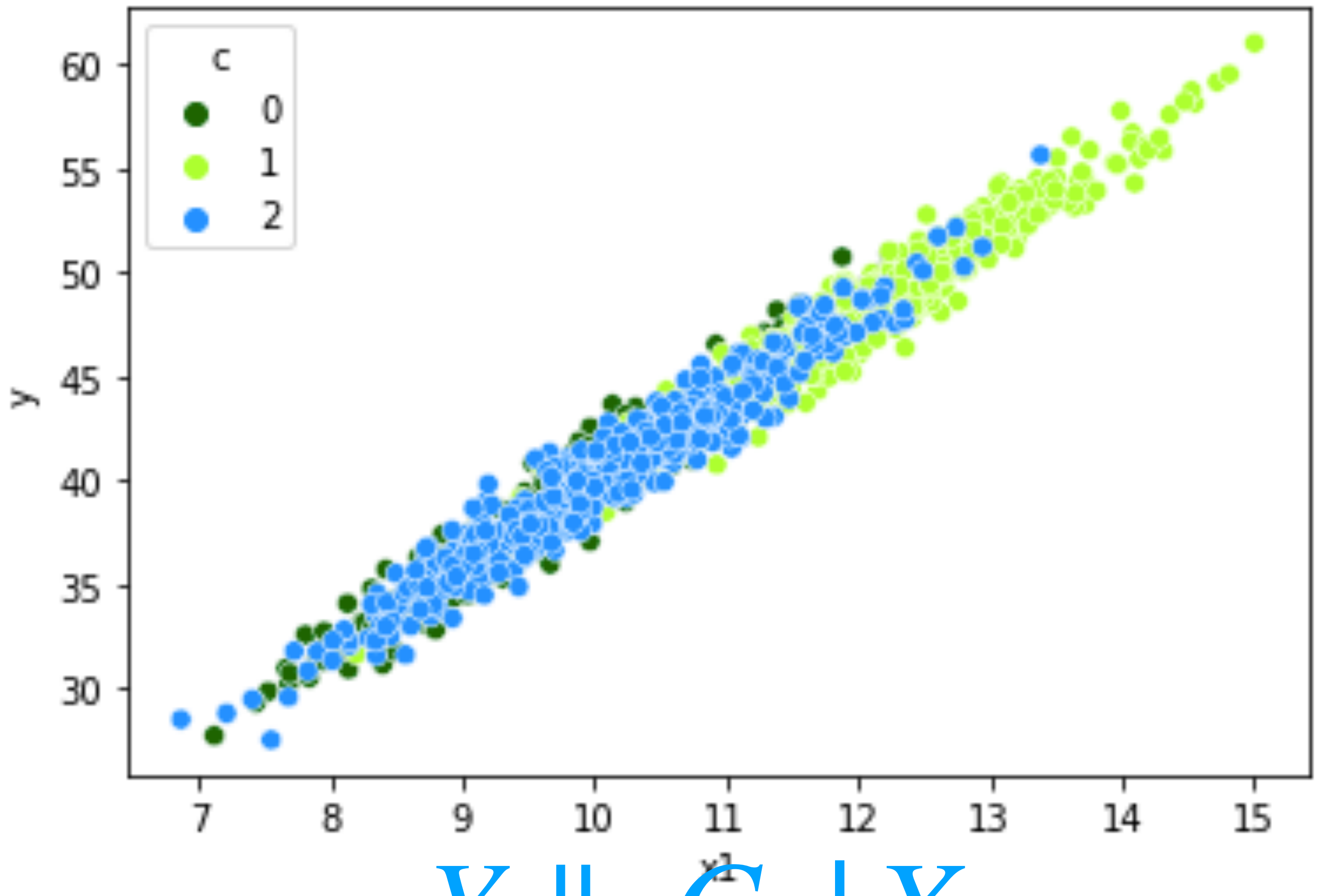


$Y \perp\!\!\!\perp C_2 \mid \{X_1, C_1 = 0\}$

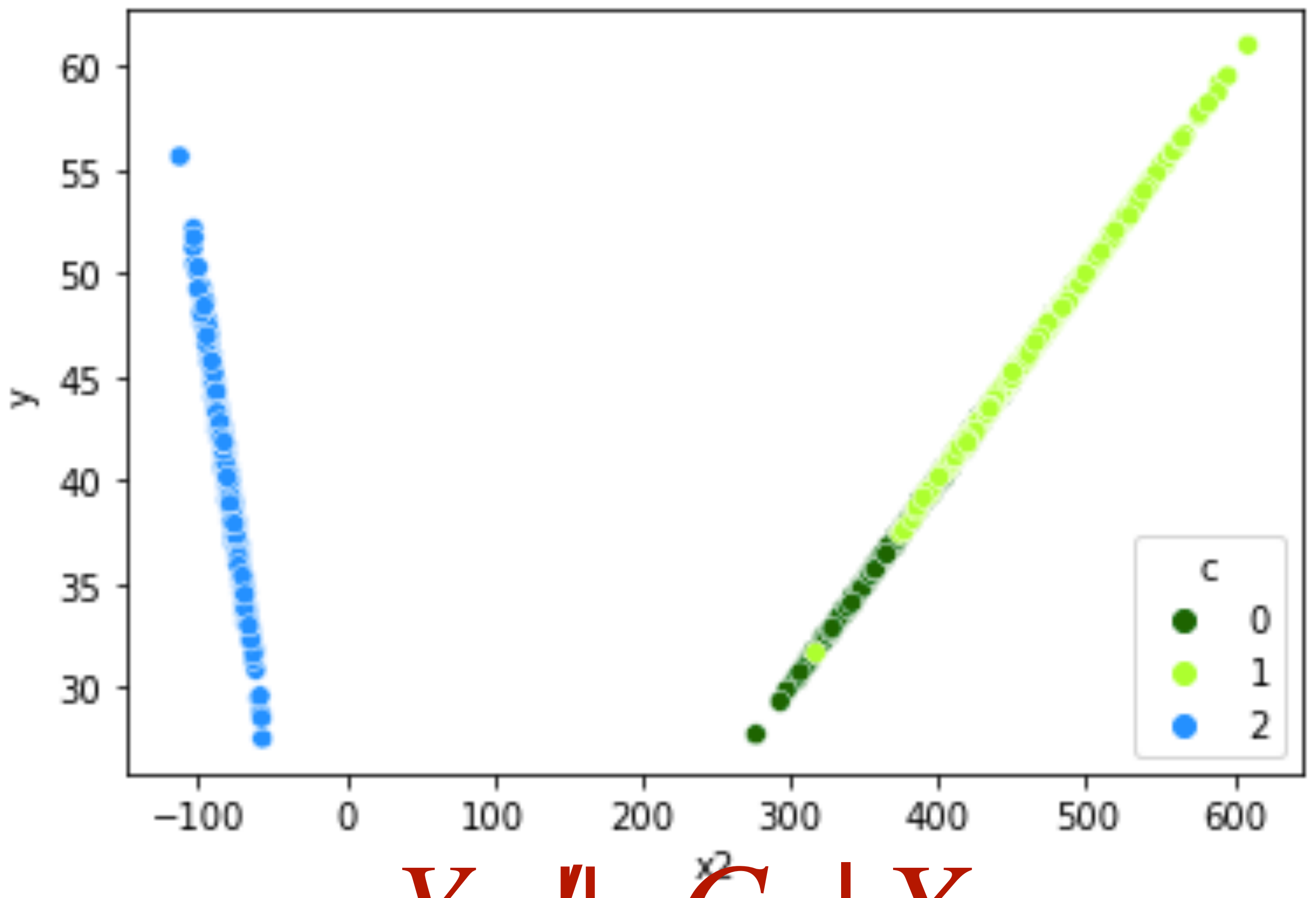


$Y \perp\!\!\!\perp C_2 \mid \{X_2, C_1 = 0\}$





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



$Y \not\perp\!\!\!\perp C_1 \mid X_2$

What if the causal graph is unknown?

- **Idea**: we could test the conditional independence in the data

$$\cancel{Y \perp\!\!\!\perp C_1 \mid X_1 ?} \qquad \cancel{Y \perp\!\!\!\perp C_1 \mid X_2 ?}$$

- **Problem:** Y is always missing when C1=1, so we cannot test these

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
1	0	3,1	2	?
1	0	3,2	3	?
1	0	4	3	?
0	1	0,2	0	0
0	1	0,3	0	1
0	1	0,3	1	0

Idea: Invariant features in source domains are also separating in the target domain

$$Y \perp\!\!\!\perp C_2 \mid \{X_1, \cancel{C_1 = 0\}} \implies Y \perp\!\!\!\perp C_1 \mid X_1$$

This is a strong assumption

What if the causal graph is unknown?

- **Idea**: we could test the conditional independence in the data

$$\cancel{Y \perp\!\!\!\perp C_1 \mid X_1 ?} \qquad \cancel{Y \perp\!\!\!\perp C_1 \mid X_2 ?}$$

- **Problem:** Y is always missing when C1=1, so we cannot test these

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
1	0	3,1	2	?
1	0	3,2	3	?
1	0	4	3	?
0	1	0,2	0	0
0	1	0,3	0	1
0	1	0,3	1	0

$$X_1 \not\!\perp\!\!\!\perp X_2$$
$$X_1 \not\!\perp\!\!\!\perp C_1$$
$$X_1 \not\!\perp\!\!\!\perp X_2 \mid C_1$$
$$X_1 \perp\!\!\!\perp X_2 \mid Y, C_1 = 0$$
$$\dots$$

- **Idea:** Can we use all other in/dependences?

Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions NeurIPS 2018

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, Joris M. Mooij

- We search for **separating features** that d-separate Y from C_1 (target)
- We assume **no extra dependences involving Y** in target domain C1=1

Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions NeurIPS 2018

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, Joris M. Mooij

- We search for **separating features** that d-separate Y from C_1 (target)
- We assume **no extra dependences involving Y** in target domain C1=1

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

$$Y \not\perp\!\!\!\perp C_2 \mid C_1 = 0$$

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

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

Perform allowed CI tests

https://arxiv.org/abs/1707.06422

Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions NeurIPS 2018

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, Joris M. Mooij

- We search for **separating features** that d-separate Y from C_1 (target)
- We assume **no extra dependences involving Y** in target domain C1=1

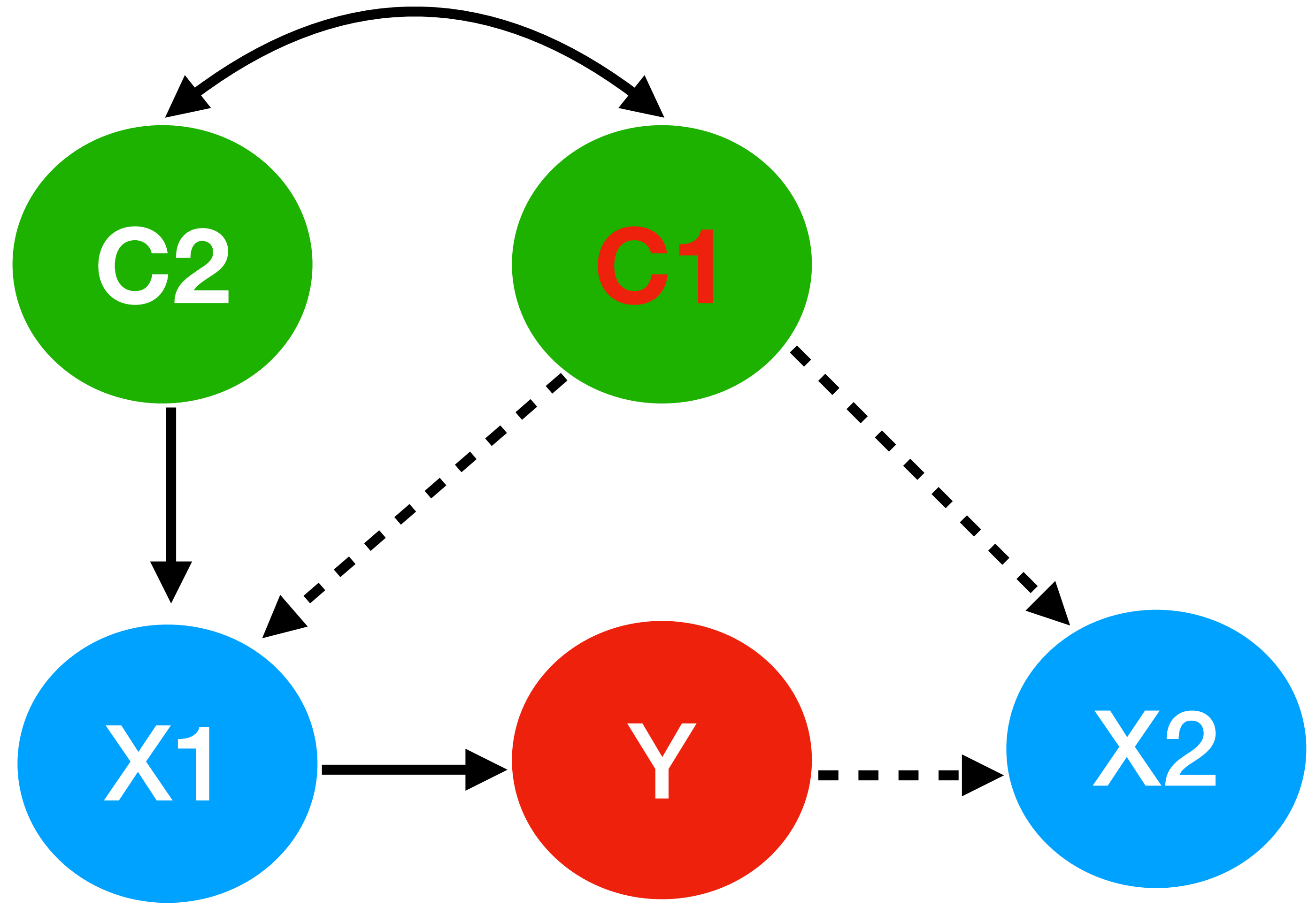
C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

$$Y \not\perp\!\!\!\perp C_2 \mid C_1 = 0$$

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

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

Perform allowed CI tests



All possible compatible graphs

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

https://arxiv.org/abs/1707.06422

Assumptions [Magliacane et al. 2018]

- We assume that there exists an **acyclic** causal graph that fits all the data (Joint Causal Inference)
- We assume **Y cannot be intervened upon directly**

Assumptions [Magliacane et al. 2018]

- We assume that there exists an **acyclic** causal graph that fits all the data (Joint Causal Inference)
- We assume **Y cannot be intervened upon directly**
- We assume **no extra dependences involving Y** in target domain C1=1

$A, D, \mathbf{B} \subset \mathbf{V} \setminus \{Y, C_1\}$

$$Y \perp\!\!\!\perp A \mid \mathbf{B}, C_1 = 0 \implies Y \perp\!\!\!\perp A \mid \mathbf{B}, C_1 = 1$$

$$A \perp\!\!\!\perp D \mid \mathbf{B}, Y, C_1 = 0 \implies A \perp\!\!\!\perp D \mid \mathbf{B}, Y, C_1 = 1$$

There can be extra independences in the target

Assumptions [Magliacane et al. 2018]

- We assume that there exists an **acyclic** causal graph that fits all the data (Joint Causal Inference)

- We assume **Y cannot be intervened upon directly**

- We assume **no extra dependences involving Y** in target domain C1=1

$A, D, \mathbf{B} \subset \mathbf{V} \setminus \{Y, C_1\}$

$$Y \perp\!\!\!\perp A \mid \mathbf{B}, C_1 = 0 \implies Y \perp\!\!\!\perp A \mid \mathbf{B}, C_1 = 1$$

$$A \perp\!\!\!\perp D \mid \mathbf{B}, Y, C_1 = 0 \implies A \perp\!\!\!\perp D \mid \mathbf{B}, Y, C_1 = 1$$

- Note that this does not assume anything about the separating set test :

$$\cancel{C_1 \perp\!\!\!\perp Y \mid \mathbf{B}}\,?$$

A small example that we proved by hand

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

A small example that we proved by hand

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

$$Y \not\perp\!\!\!\perp C_2 \mid C_1 = 0$$

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

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

Perform allowed CI tests

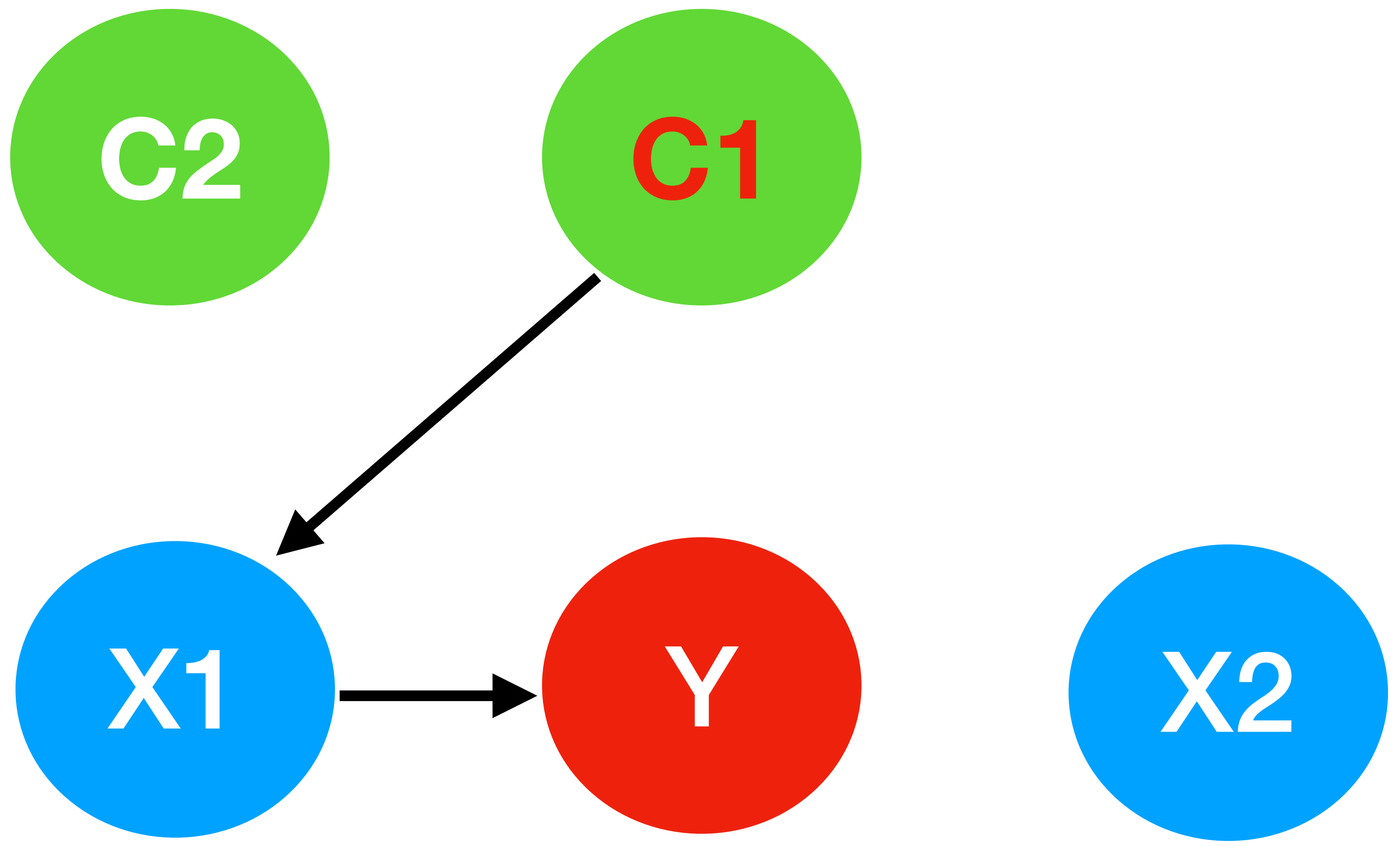
A small example that we proved by hand

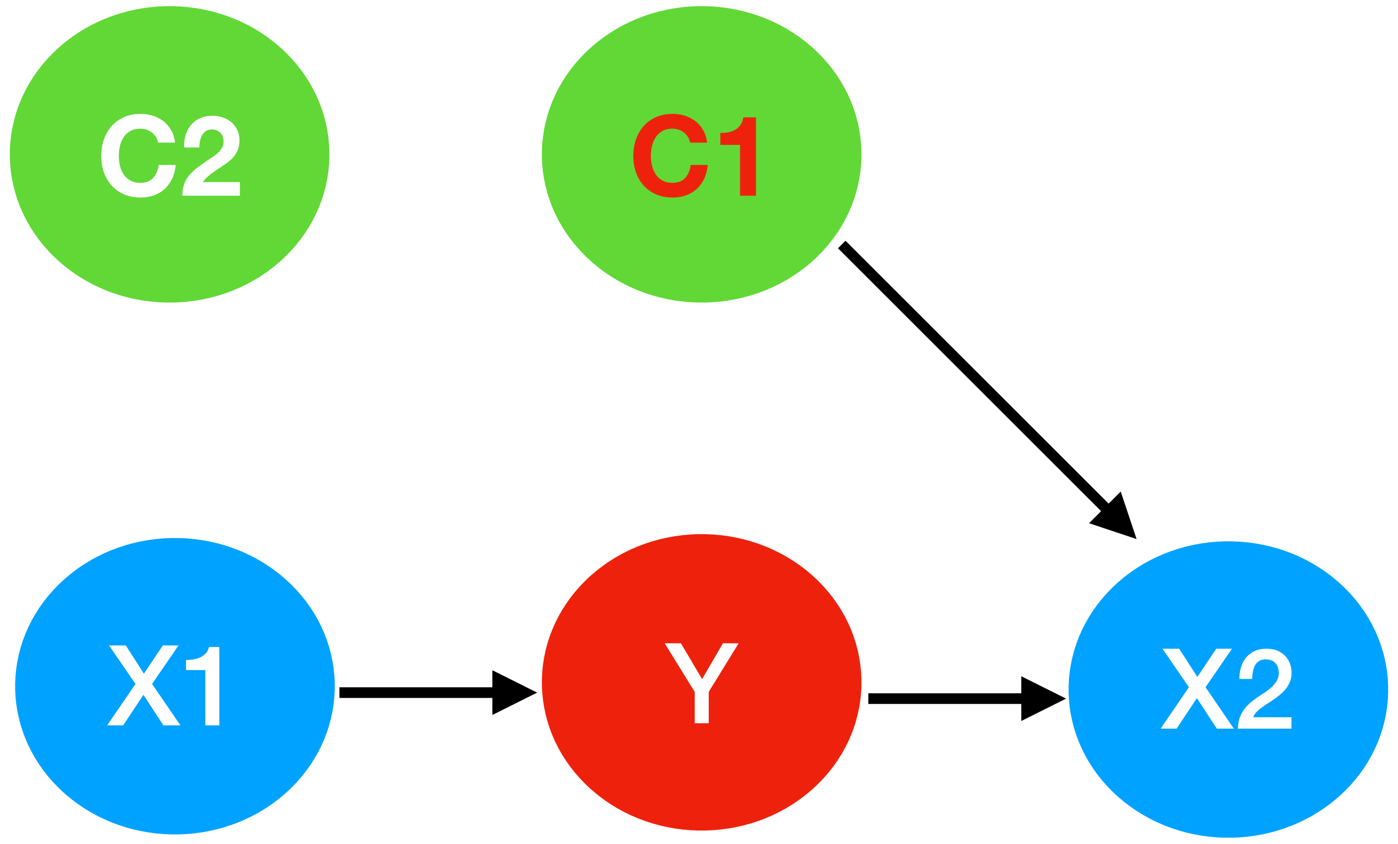
C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

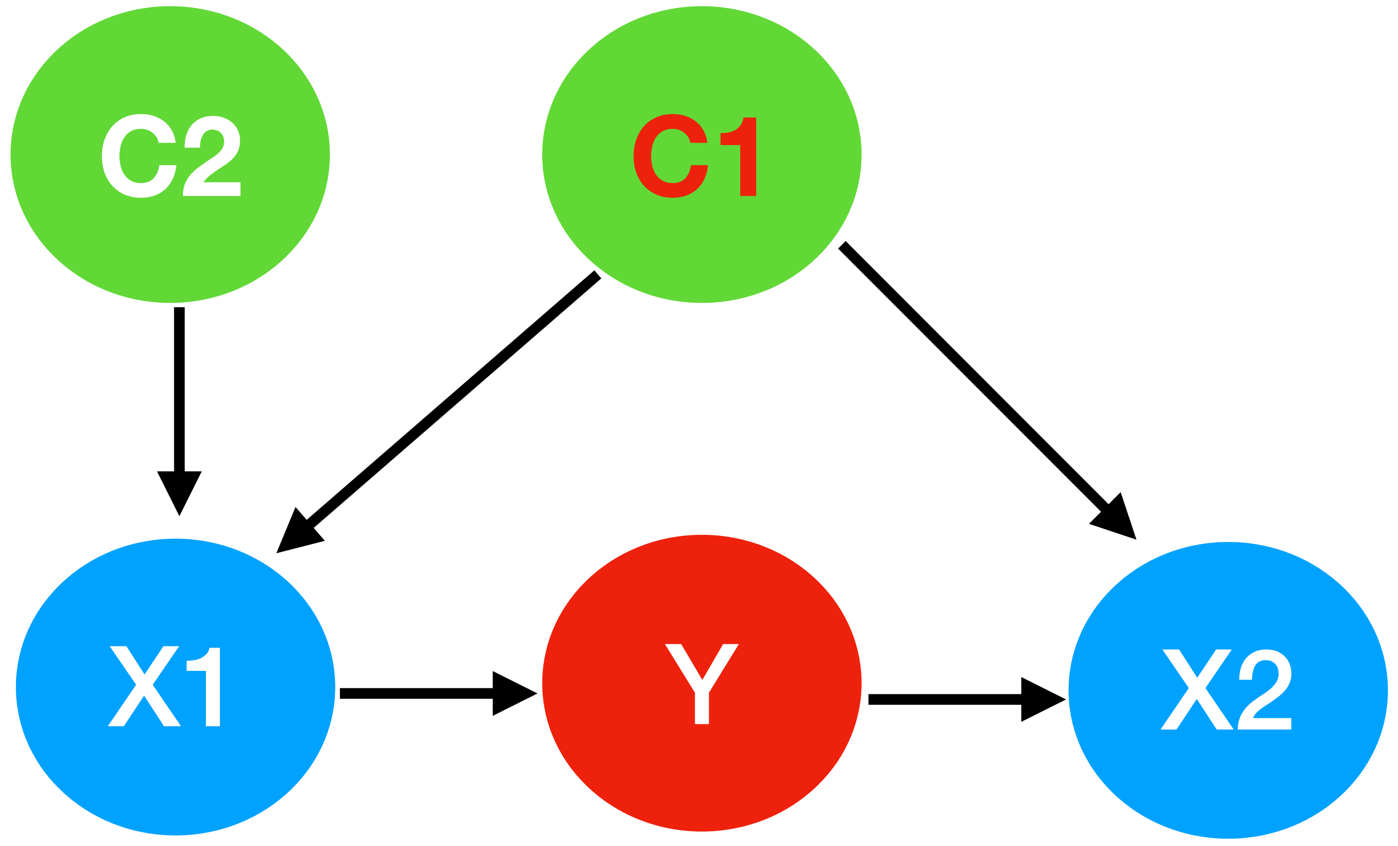
$$Y \not\perp\!\!\!\perp C_2 \mid C_1 = 0$$

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

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







A small example that we proved by hand

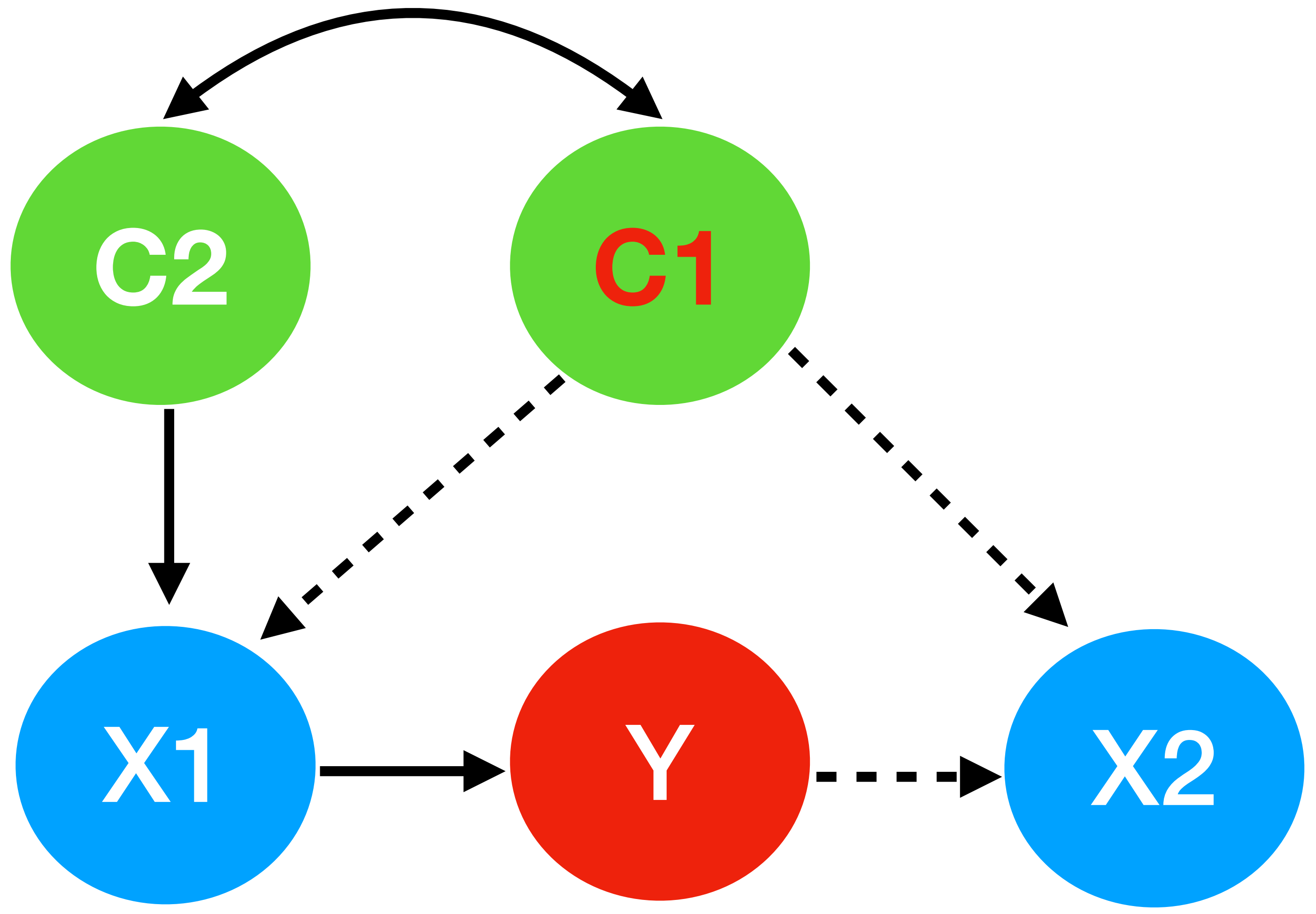
C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

$$Y \not\perp\!\!\!\perp C_2 \mid C_1 = 0$$

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

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

Perform allowed CI tests



All possible compatible graphs

A small example that we proved by hand

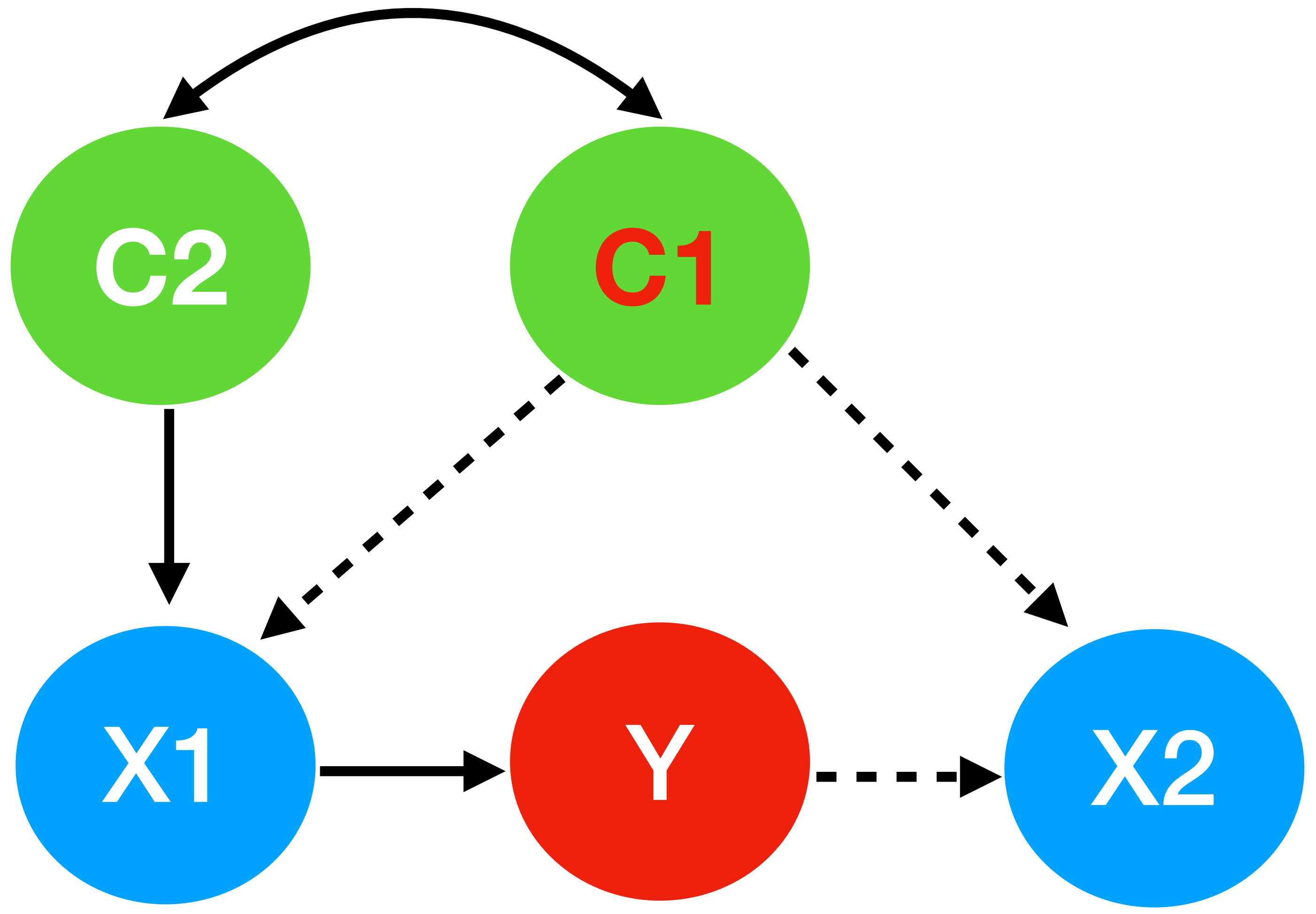
C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

$$Y \not\perp\!\!\!\perp C_2 \,|\, C_1 = 0$$

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

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

Perform allowed CI tests



All possible compatible graphs

- We can prove untestable separating test **without reconstructing the graph:**

$$Y \perp\!\!\!\perp C_1 \,|\, X_1$$

True in all possible compatible graphs

Inferring separating sets automatically

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)

A simple causal feature selection algorithm

Source domains data

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1

→ Standard feature selection →

List of combinations of features ordered by source domain loss in predicting Y

L=({X1, C2}, {X1, X2, C2} , {X1, X2}, …)

A simple causal feature selection algorithm

Source domains data

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1

Standard feature selection

List of combinations of features ordered by source domain loss in predicting Y

L=({X1, C2}, {X1, X2, C2} , {X1, X2}, …)

Select new set S

S= {X1, C2}

A simple causal feature selection algorithm

Source domains data

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1

Standard feature selection

List of combinations of features ordered by source domain loss in predicting Y

L=({X1, C2}, {X1, X2, C2} , {X1, X2}, …)

Select new set S

S= {X1, C2}

All data (including target)

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

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

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

A simple causal feature selection algorithm

Source domains data

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1

Standard feature selection

List of combinations of features ordered by source domain loss in predicting Y

L=({X1, C2}, {X1, X2, C2} , {X1, X2}, …)

Select new set S

S= {X1, C2}

All data (including target)

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

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

Assumptions

All testable conditional independences from data

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

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

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

…

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

Theorem prover

Provably not separating

$Y \not\perp\!\!\!\perp C_1 \mid S$

?

Not identifiable

A simple causal feature selection algorithm

Source domains data

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1

Standard feature selection

List of combinations of features ordered by source domain loss in predicting Y

L=({X1, C2}, {X1, X2, C2} , {X1, X2}, …)

Select new set S

S= {X1, X2, C2}

All data (including target)

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

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

Assumptions

All testable conditional independences from data

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

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

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

…

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

Theorem prover

Provably not separating

$Y \not\!\perp\!\!\!\perp C_1 \mid S$

?

Not identifiable

A simple causal feature selection algorithm

Source domains data

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1

Standard feature selection

List of combinations of features ordered by source domain loss in predicting Y

L=({X1, C2}, {X1, X2, C2} , {X1, X2}, …)

Select new set S

S= {X1, X2, C2}

All data (including target)

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

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

Assumptions

All testable conditional independences from data

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

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

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

…

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

Theorem prover

Provably separating $Y \perp\!\!\!\perp C_1 \mid S$

Learn $\hat{f}$(S) on source domains

A simple causal feature selection algorithm

Source domains data

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1

Standard feature selection

List of combinations of features ordered by source domain loss in predicting Y

L=({X1, C2}, {X1, X2, C2} , {X1, X2}, …)

Select new set S

Iterate until empty

All data (including target)

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

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

Assumptions

All testable conditional independences from data

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

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

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

…

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

Theorem prover

Provably not separating $Y \not\perp\!\!\!\perp C_1 \mid S$

? Not identifiable

Provably separating $Y \perp\!\!\!\perp C_1 \mid S$

Learn $\hat{f}$(S) on source domains

A simple causal feature selection algorithm

Source domains data

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1

Standard feature selection

List of combinations of features ordered by source domain loss in predicting Y

L=({X1, C2}, {X1, X2, C2} , {X1, X2}, …)

Select new set S

Bounded generalisation error

All data (including target)

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

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

Assumptions

All testable conditional independences from data

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

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

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

…

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

Theorem prover

Provably not s

$Y \not\perp\!\!\!\perp C_1$

?

Not identifiable

C1	C2	X1	X2	Y
0	1	0,2	0	?
0	1	0,3	0	?
0	1	0,3	1	?

Provably separating

$Y \perp\!\!\!\perp C_1 \mid S$

Learn $\hat{f}(S)$ on source domains

A simple causal feature selection algorithm

Source domains data

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1

Standard feature selection

List of combinations of features ordered by source domain loss in predicting Y

L=({X1, C2}, {X1, X2, C2} , {X1, X2}, …)

Select new set S

Bounded generalisation error

All data (including target)

C1	C2	X1	X2	Y
0	0	0,1	1	0
0	0	0,2	1	0
0	0	1,1	2	1
0	1	3,1	2	1
0	1	3,2	3	1
0	1	4	3	1
1	0	0,2	0	?
1	0	0,3	0	?
1	0	0,3	1	?

No need to find causal graph or equivalence class, we only care about conditional independences/d-separations

All testable conditional independences from data

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

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

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

…

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

Theorem prover

Provably not s… $Y \not\perp\!\!\!\perp C_1$

Not identifiable

C1	C2	X1	X2	Y
0	1	0,2	0	?
0	1	0,3	0	?
0	1	0,3	1	?

Provably separating $Y \perp\!\!\!\perp C_1 \mid S$

Learn $\hat{f}(S)$ on source domains

Takeaways 2/3

- Graphical models and d-separation [Pearl 1988] are a principled way to reason about **invariances and distribution shift**
  - Not a new observation, known since [Schoelkopf et al 2012]
  - Even with **unknown causal graphs, Missing data/zero-shot settings**
- Often we **do not need to reconstruct the causal graph**, we only need to infer missing conditional independences

Inferring separating sets of features

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

Assumptions

All testable conditional independences from data

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

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

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

Log

No need to find causal graph or equivalence class, we only care about conditional independences/d-separations

bly separating

$\perp\!\!\!\perp C_1 | X_1$

Provably not separating

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

?

Not identifiable

A big (current) limitation: Scalability

Other limitations: current only feature selection

We fea sel

Other limitations: the concept of separating sets might be too conservative

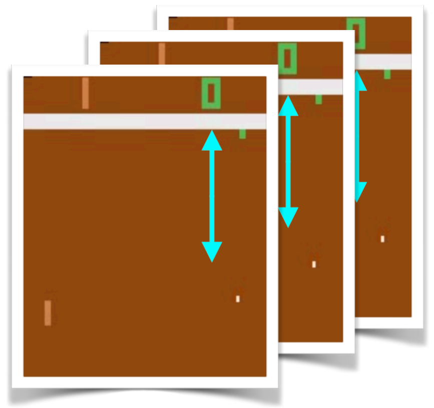
https://arxiv.org/abs/1707.06422

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

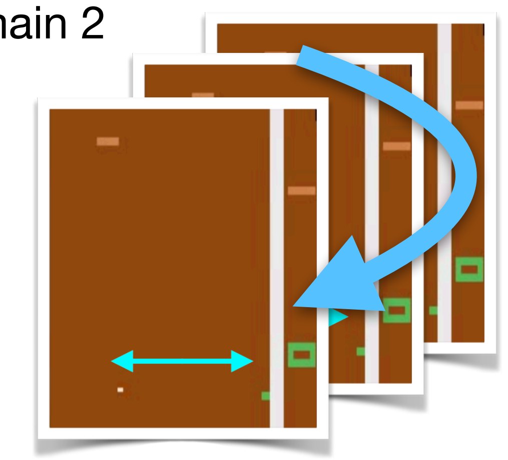
Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang ICLR 2022

Source domains

Domain 1

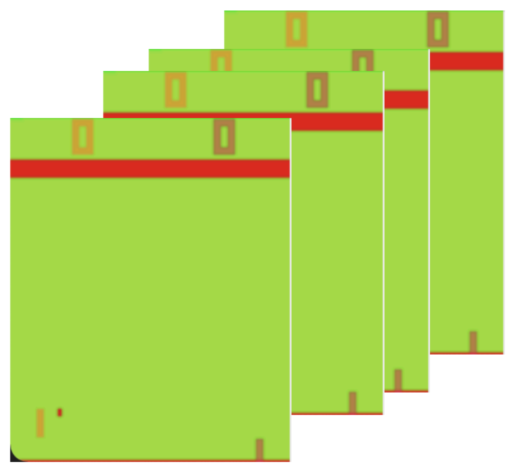


Domain 2



. . .

Domain n

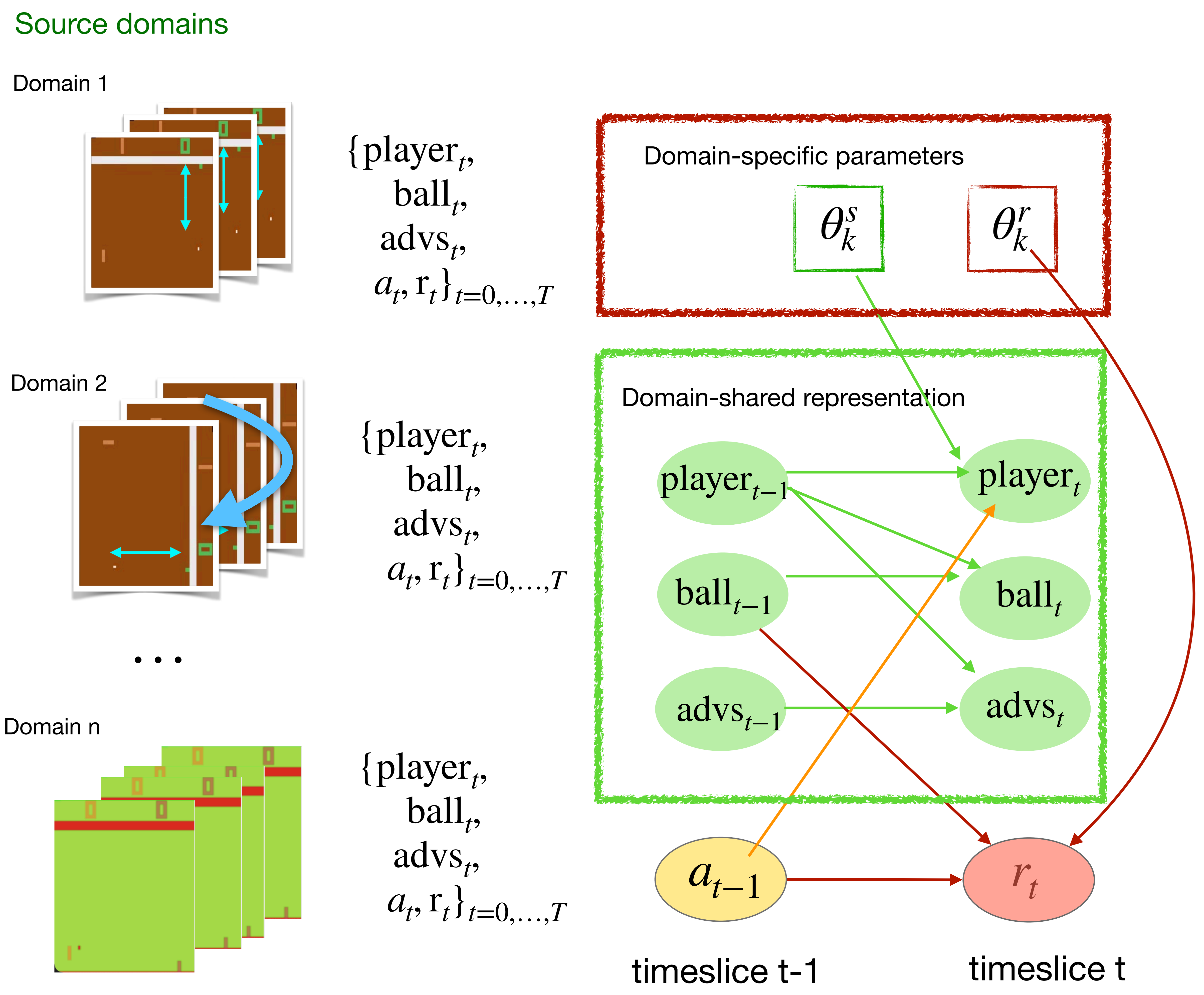


Simplifying assumption: no new edges in target domain

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**

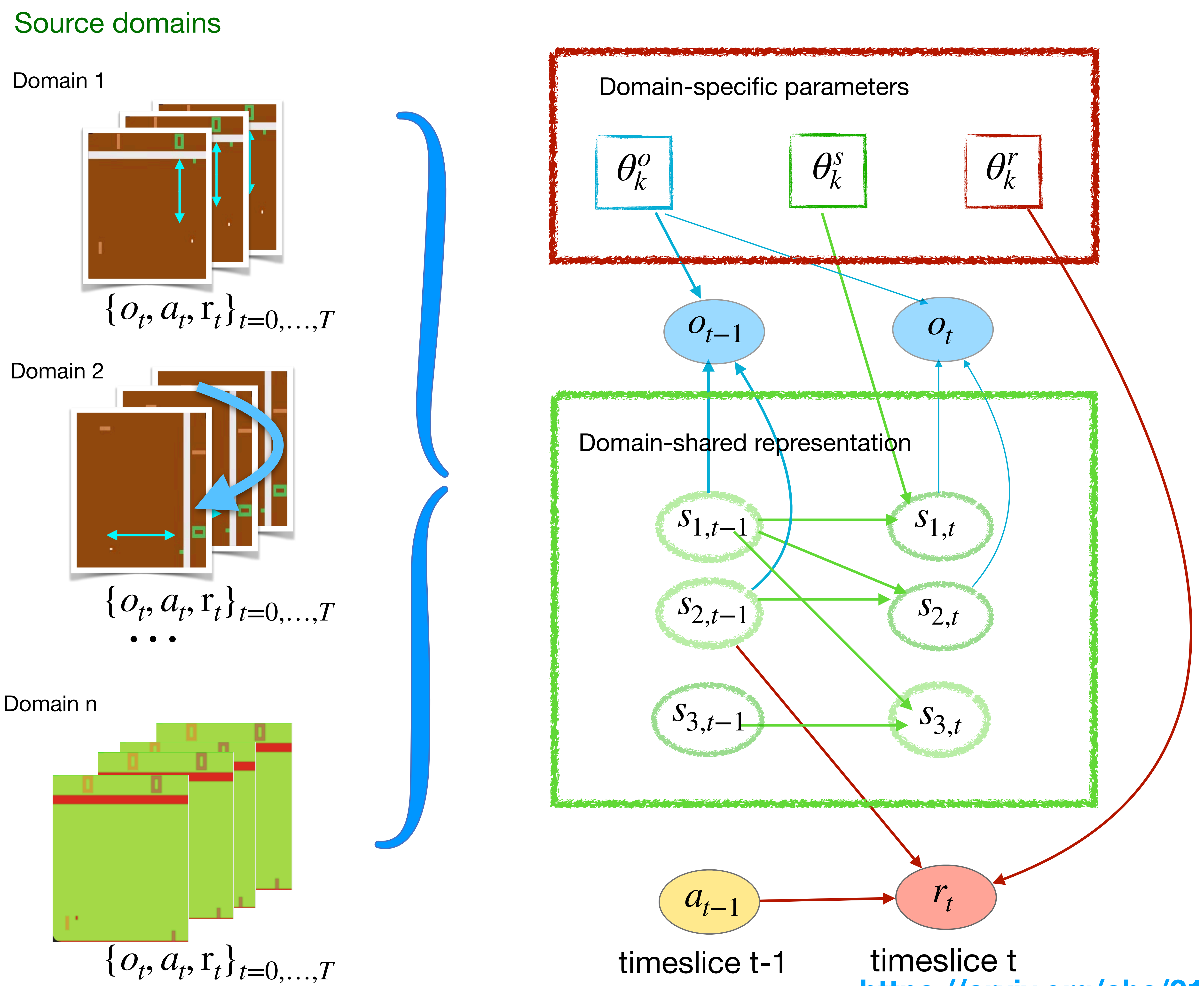


When we learn from symbolic inputs, the causal graph can be identified, but we don't have guarantees on what the latent change factors are

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**



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

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

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

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**

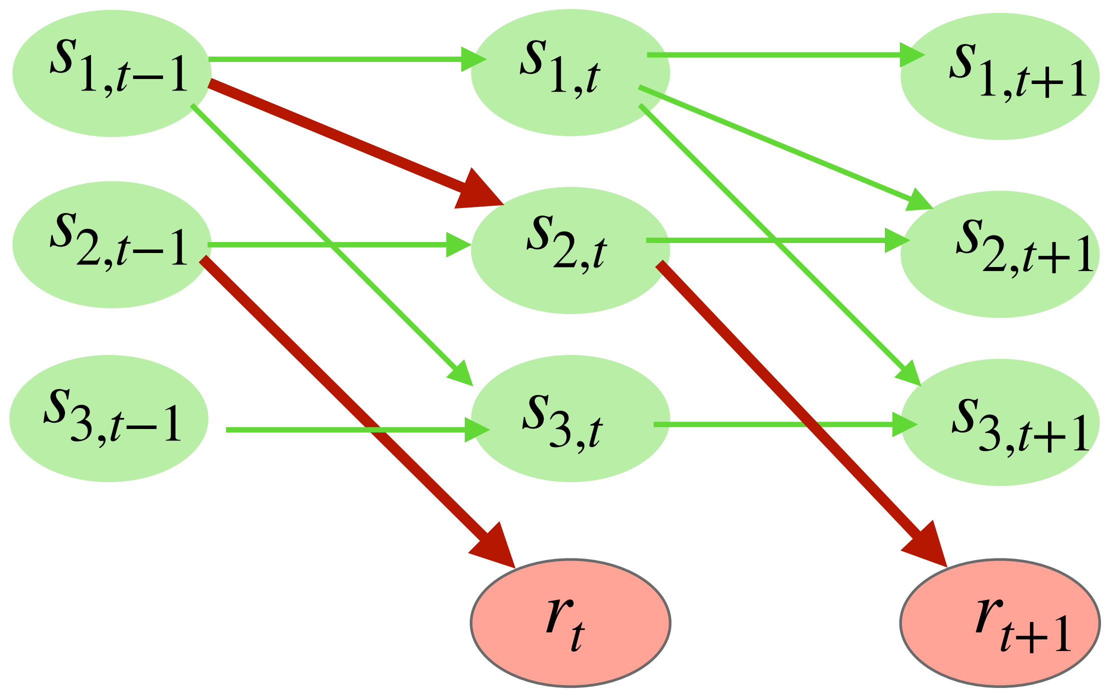
Source domains

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

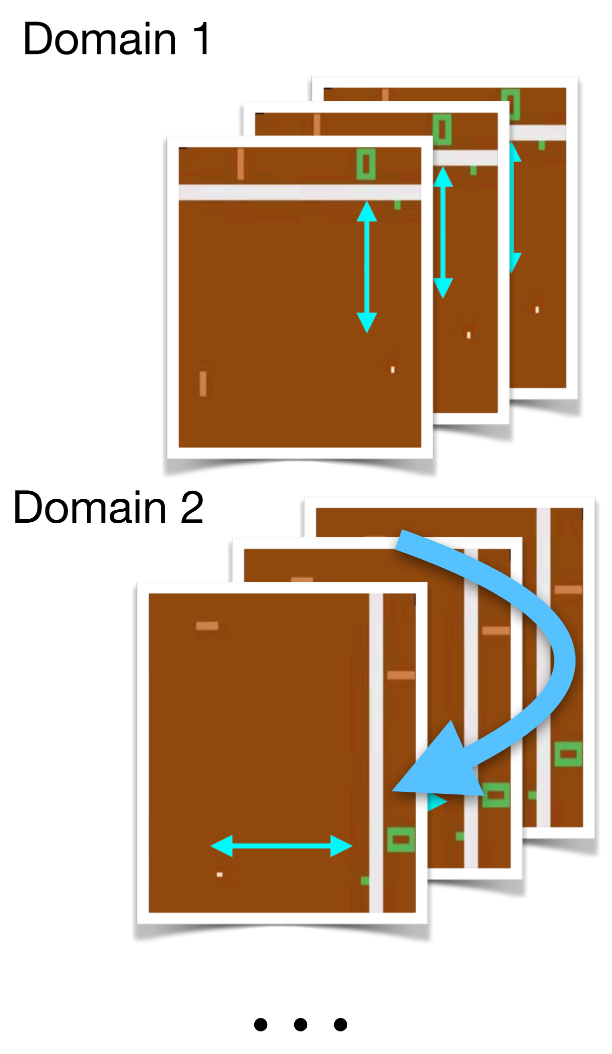


https://arxiv.org/abs/2107.02729

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

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

Source domains

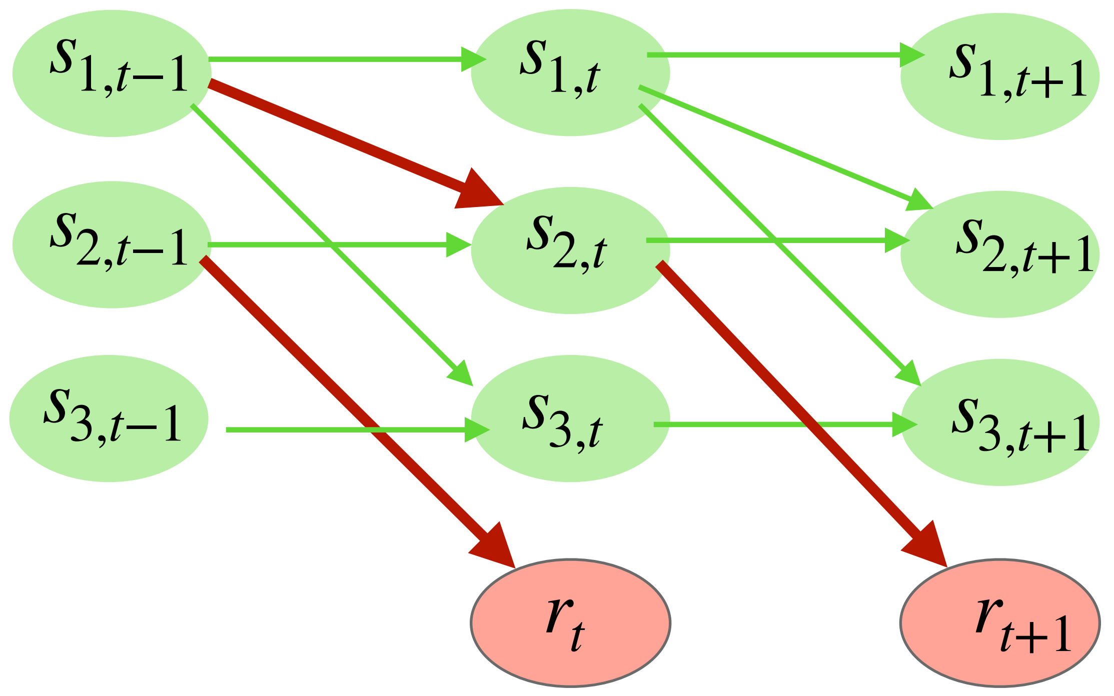


Estimate graph over estimated s_k, θ_k

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



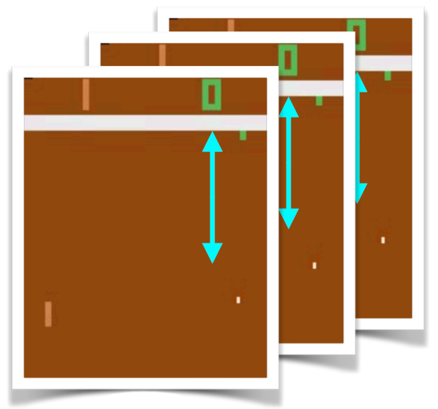
https://arxiv.org/abs/2107.02729

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

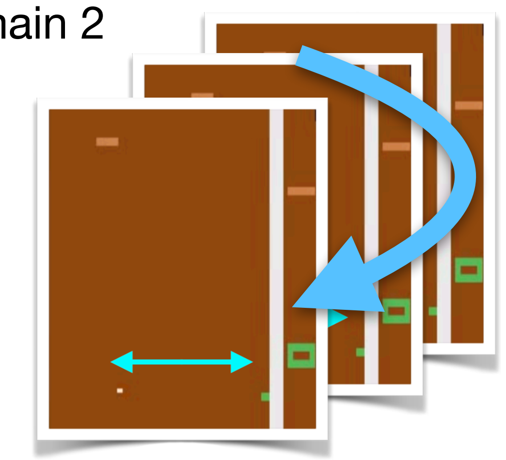
Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang ICLR 2022

Source domains

Domain 1

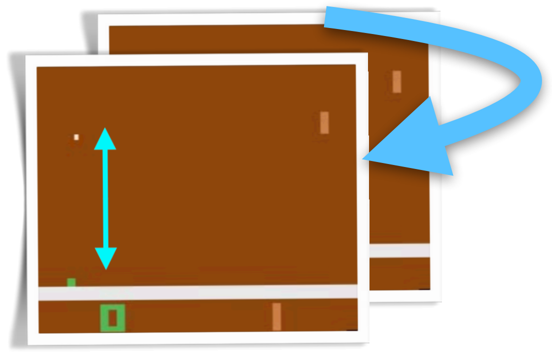


Domain 2



...

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

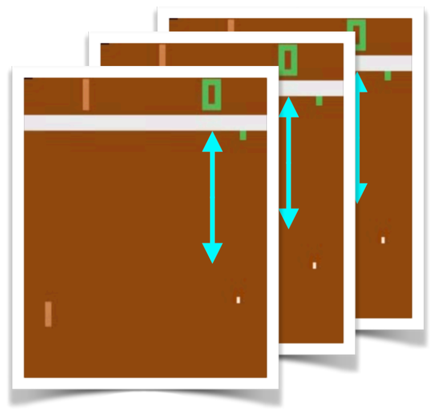
AdaRL: What, Where, and How to Adapt in Transfer RL

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

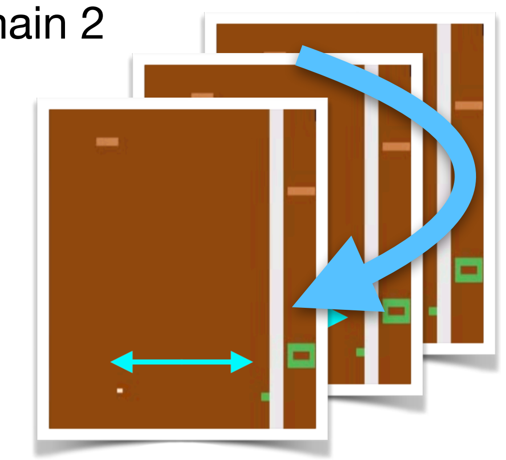
ICLR 2022

Source domains

Domain 1



Domain 2



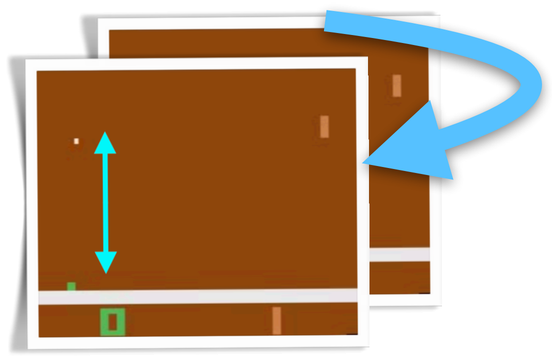
…

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

- **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)

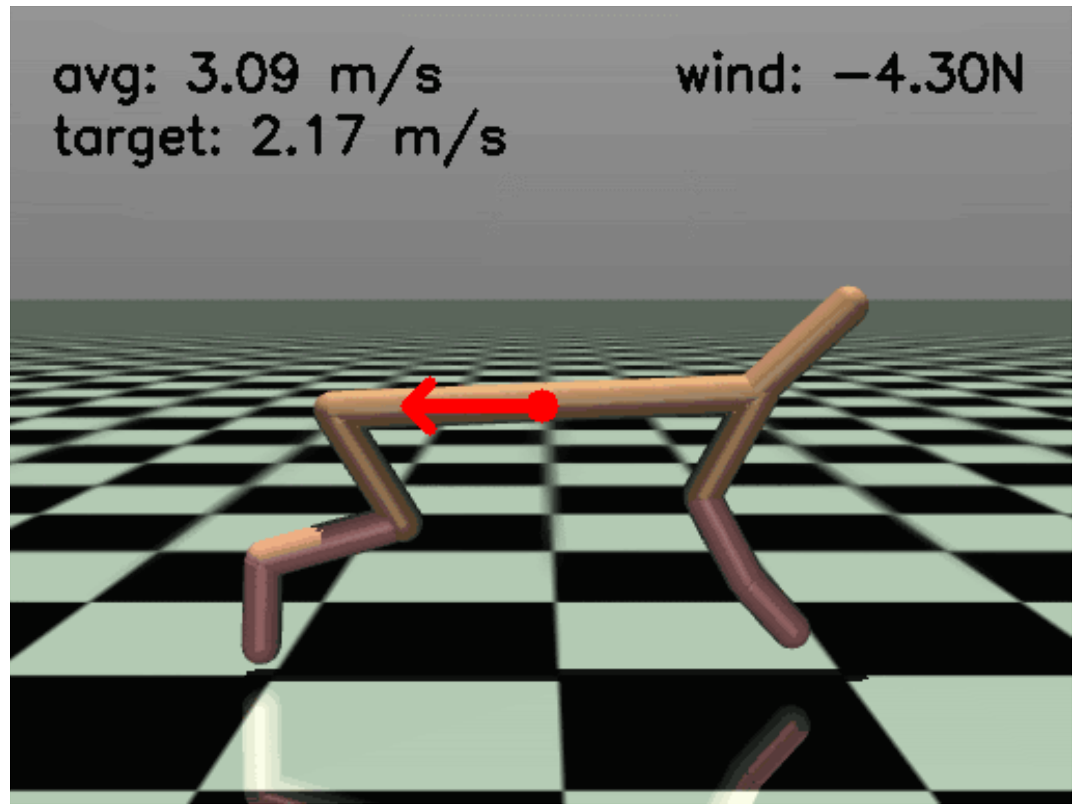
Average final scores on Pong (POMDP) with N_target=50

https://arxiv.org/abs/2107.02729

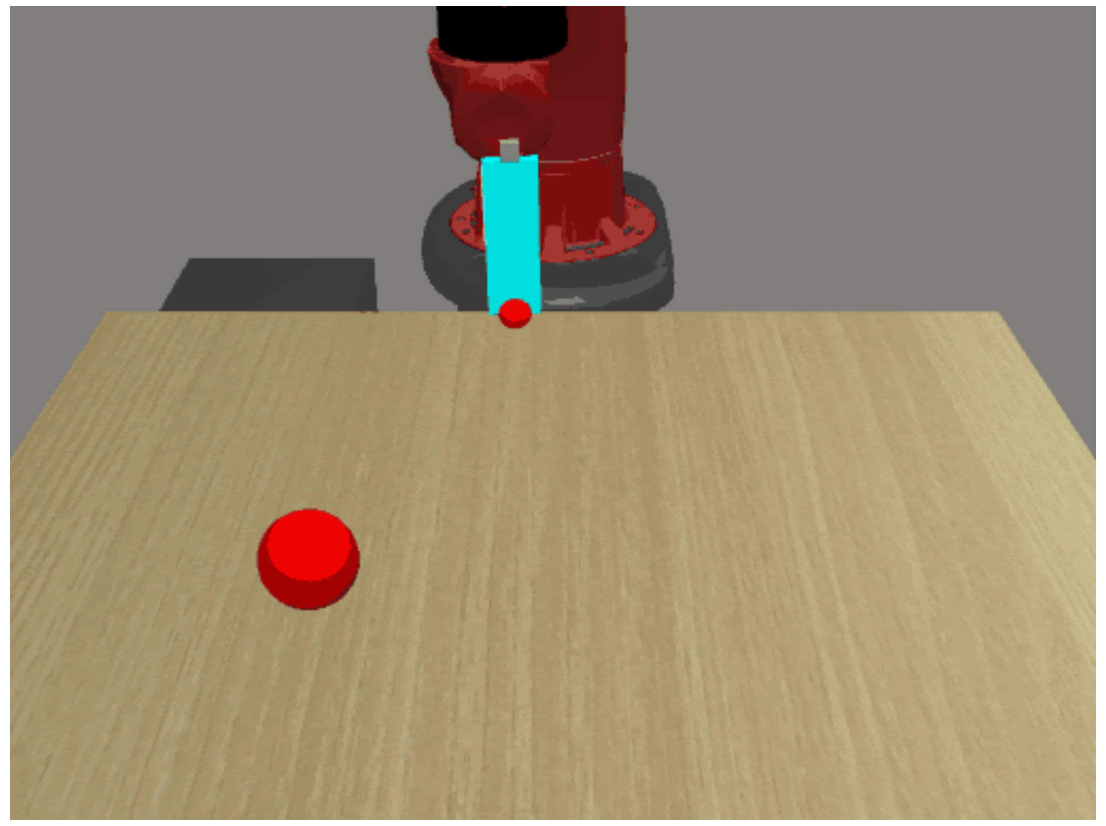
FansRL: Factored Adaptation for Non-Stationary Reinforcement Learning

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

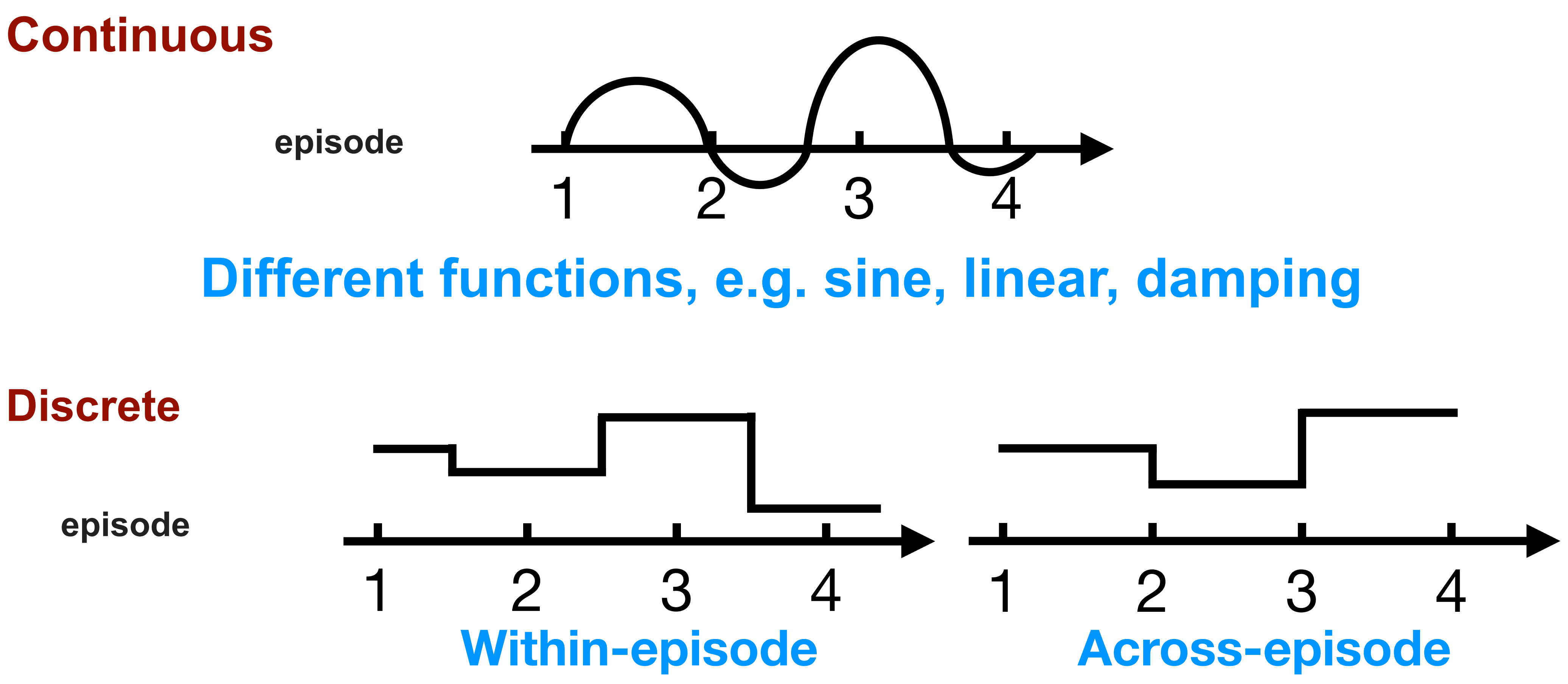
- **Task:** RL agent has to learn a policy that is robust to different types of non-stationarity, including **multiple simultaneous changes of different types**



Non-stationary environments
(wind changes)



Non-stationary rewards
(target changes)



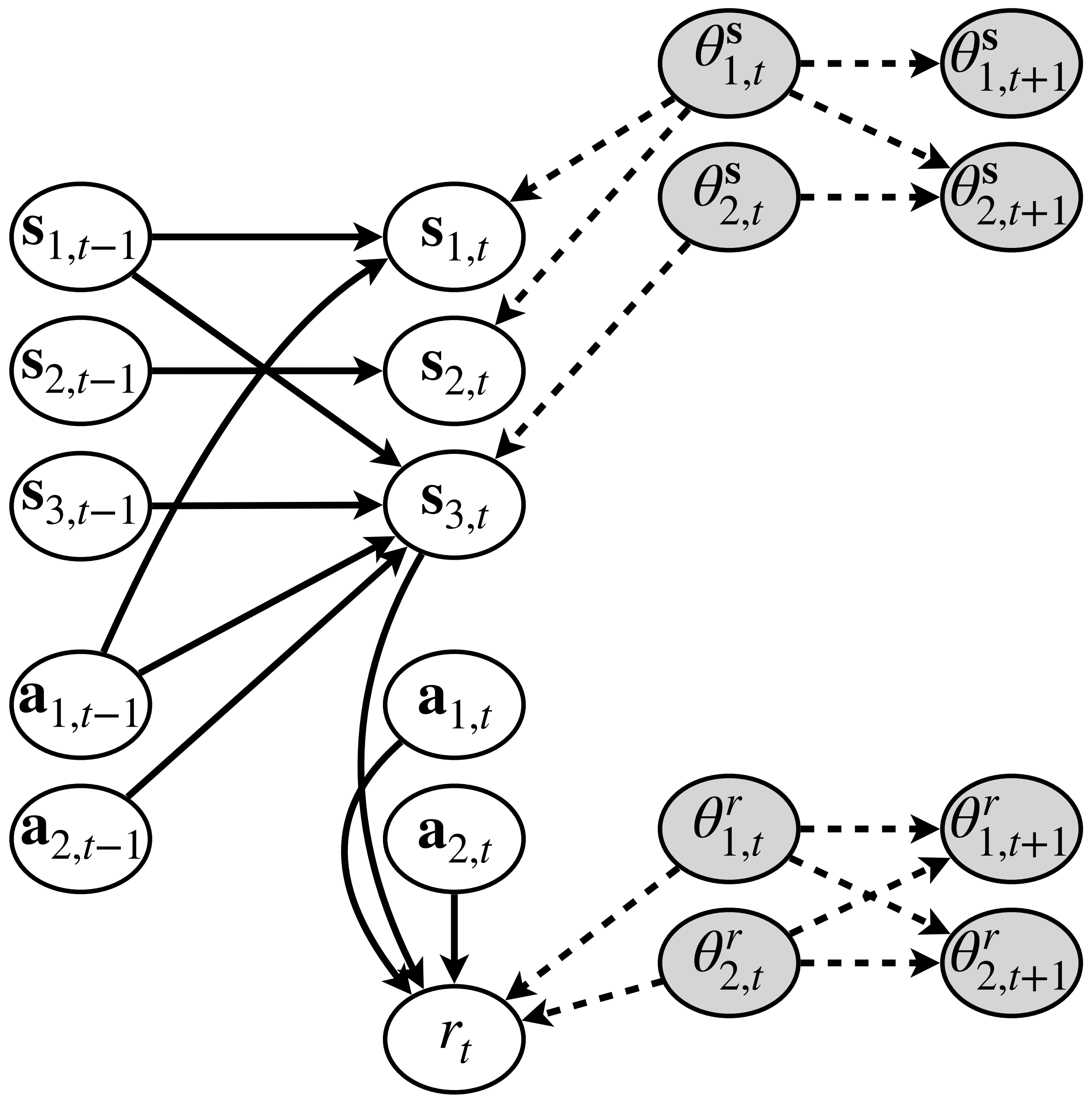
https://arxiv.org/abs/2203.16582

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

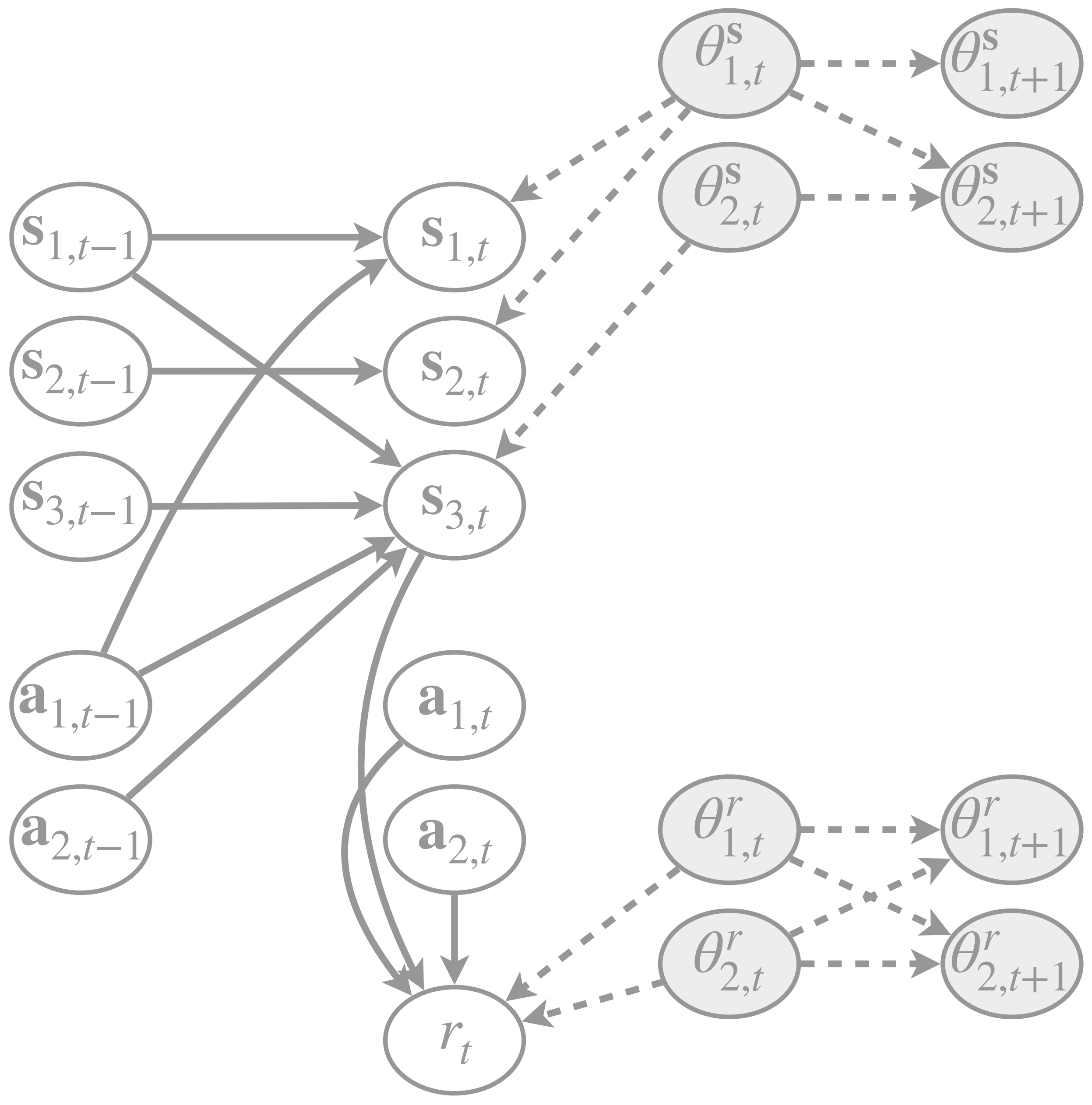
https://arxiv.org/abs/2203.16582

FansRL: Factored Adaptation for Non-Stationary Reinforcement Learning

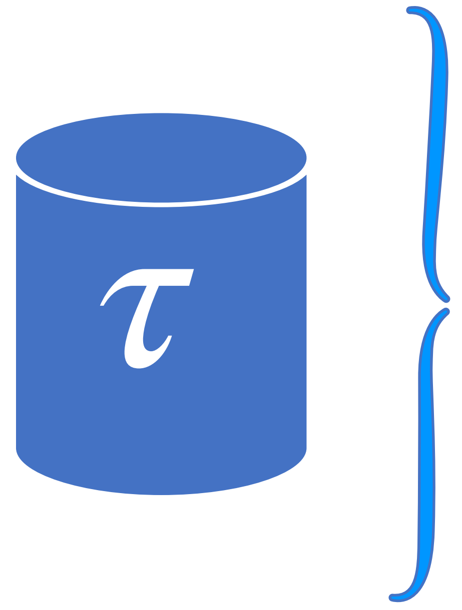
Fan Feng, Biwei Huang, Kun Zhang, Sara Magliacane

NeurIPS 2022

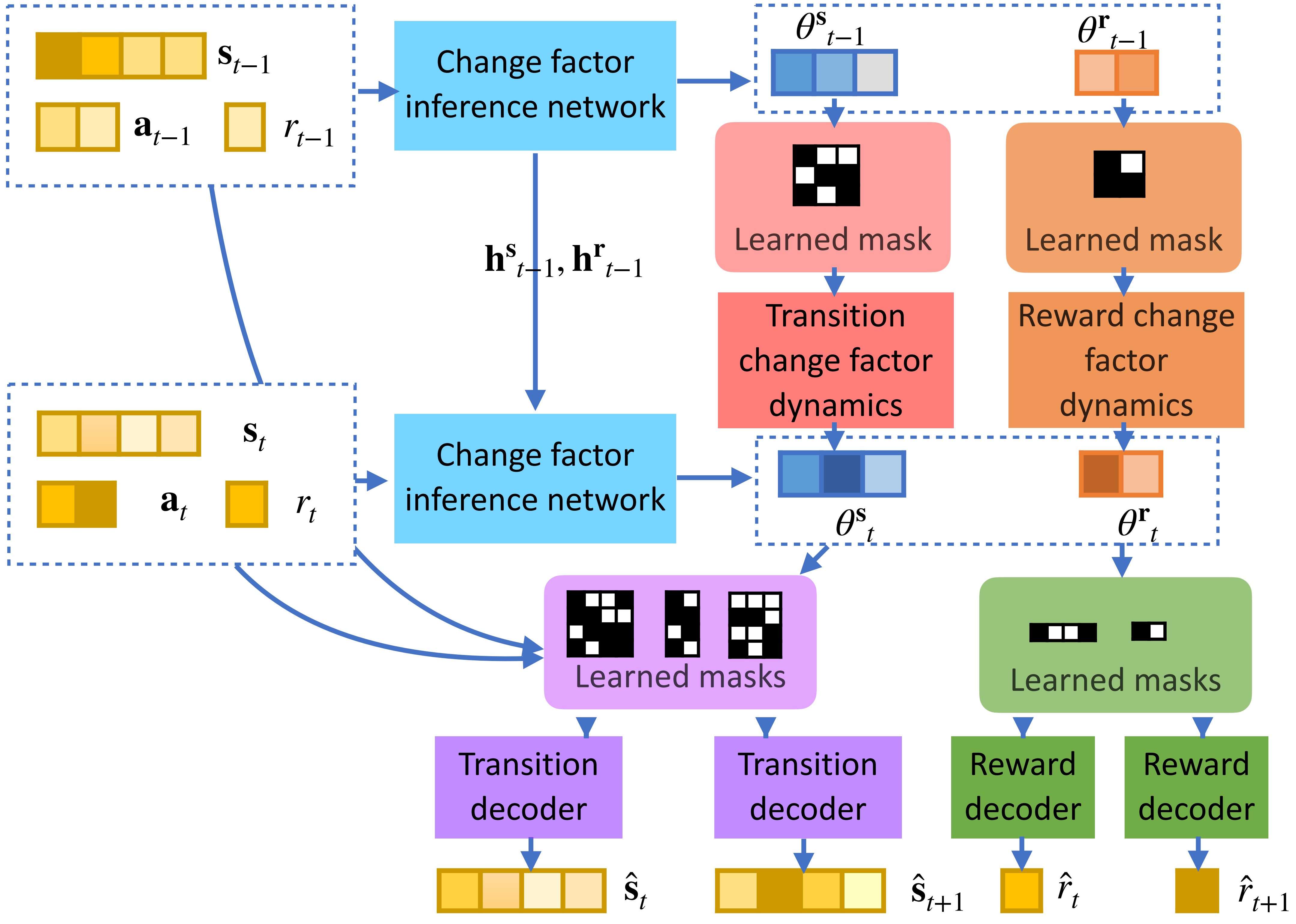
- The **latent change factors** are not constant anymore and they model **non-stationarity**



Factored Non-Stationary MDP



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



Factored Non-Stationary Variational Autoencoder

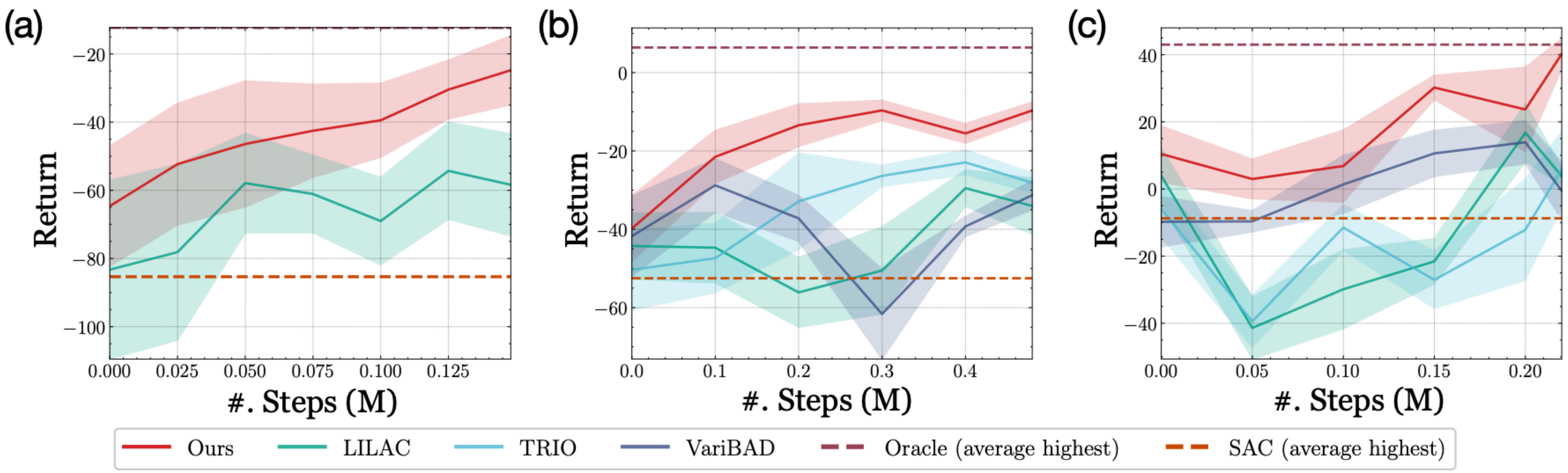
https://arxiv.org/abs/2203.16582

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**

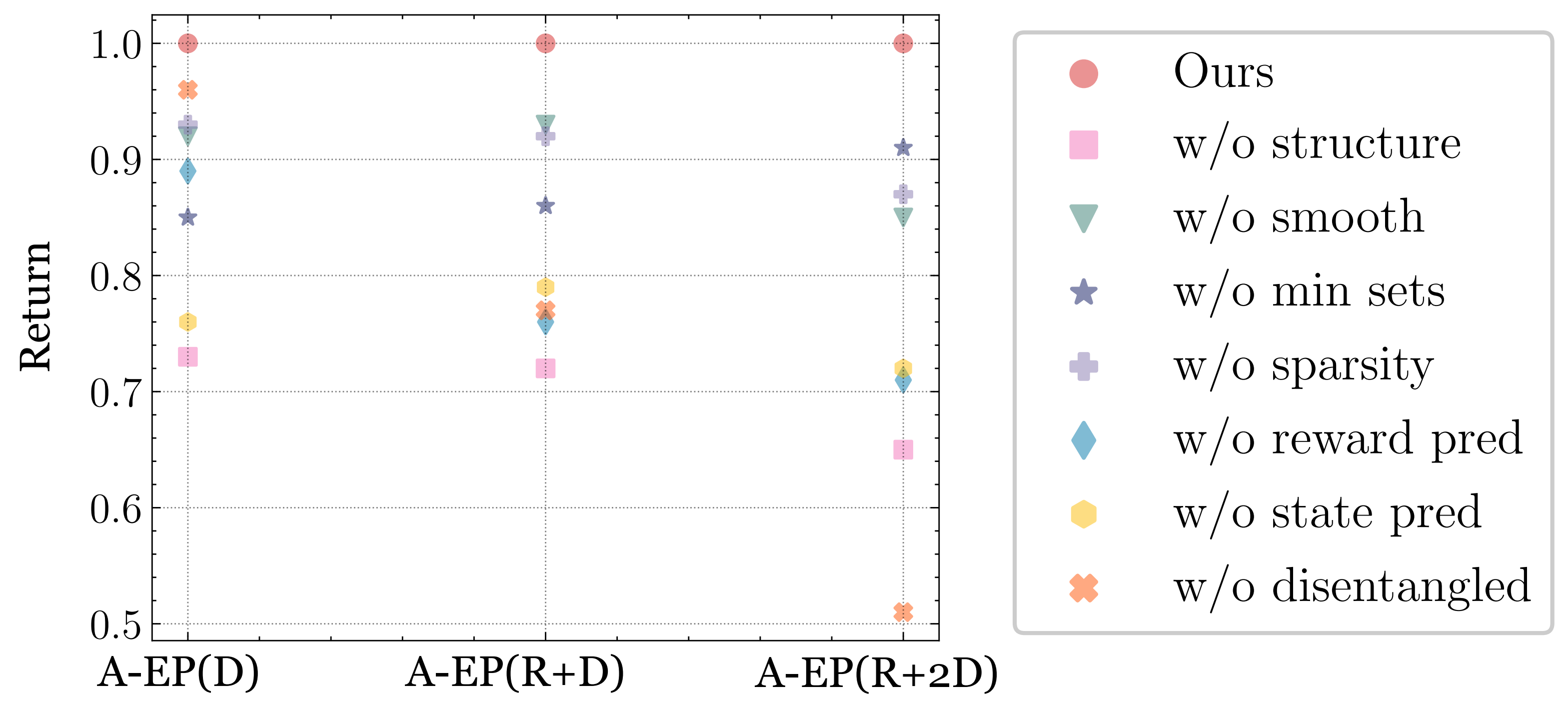


Continuous changes on dynamics (sine wind)

Across-episode changes on rewards (changing target)

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

Ablation studies



The biggest difference in performance is switching off learning the graph

https://arxiv.org/abs/2203.16582

Takeaways 3/3

- Graphical models and d-separation [Pearl 1988] are a principled way to reason about **invariances and distribution shift**
    - Not a new observation, known since [Schoelkopf et al 2012]
    - Even with **unknown causal graphs, Missing data/zero-shot settings**
- Often we **do not need to reconstruct the causal graph**, we only need to infer missing conditional independences
- These ideas seem empirically useful even if we **cannot guarantee** that we are **learning the true causal variables or the true causal graph**

Sneak peak: Causal Representation Learning

Causal discovery (structure learning) - simplest setting

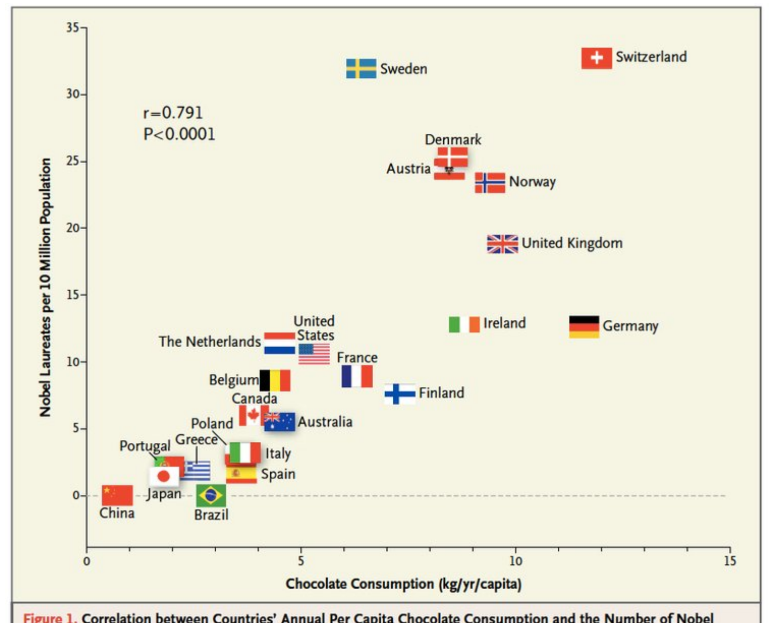


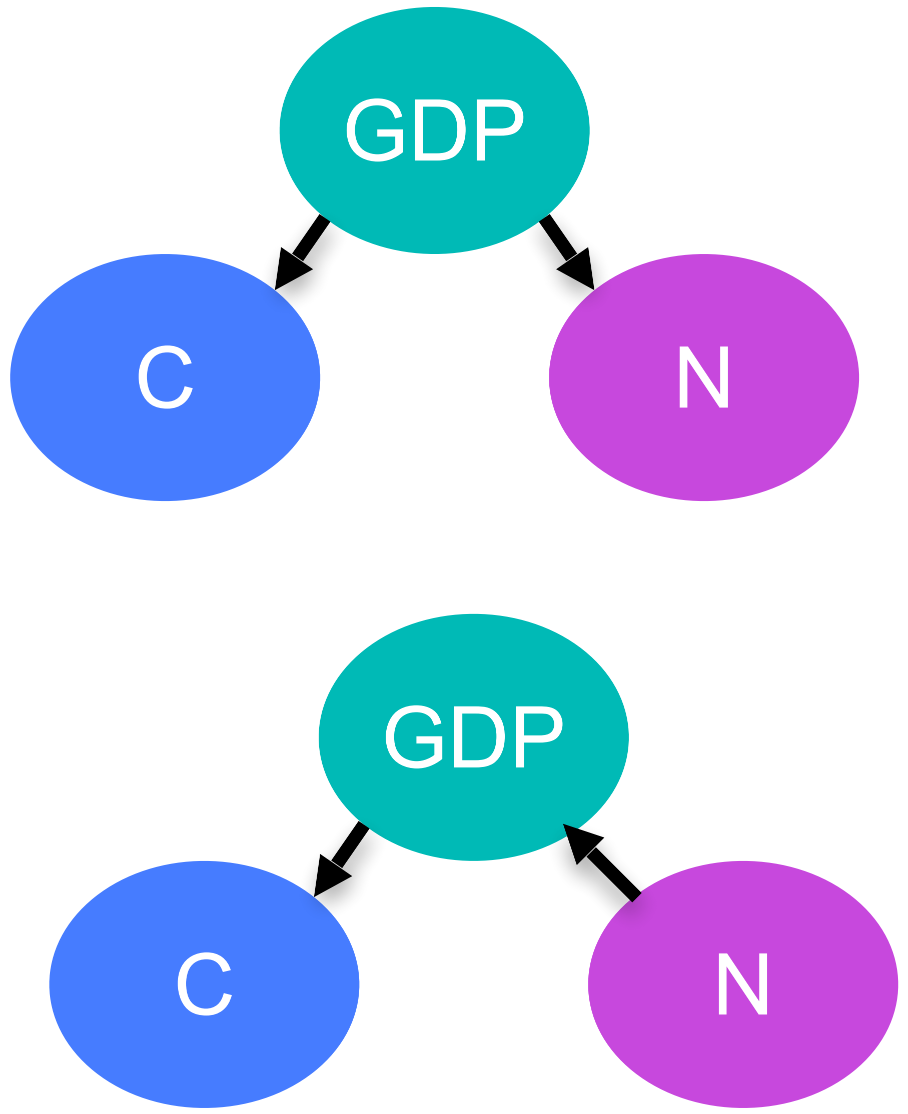
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

C	N	GDP
4.5	5	33k
12	30	86k
10	20	46k
....	...	...

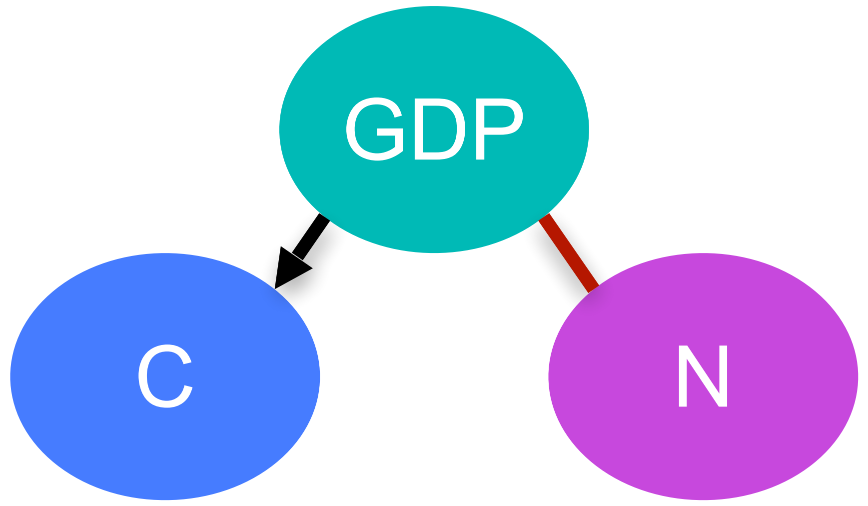
Observational data

$$C \nrightarrow GDP$$

[Optional] Background knowledge

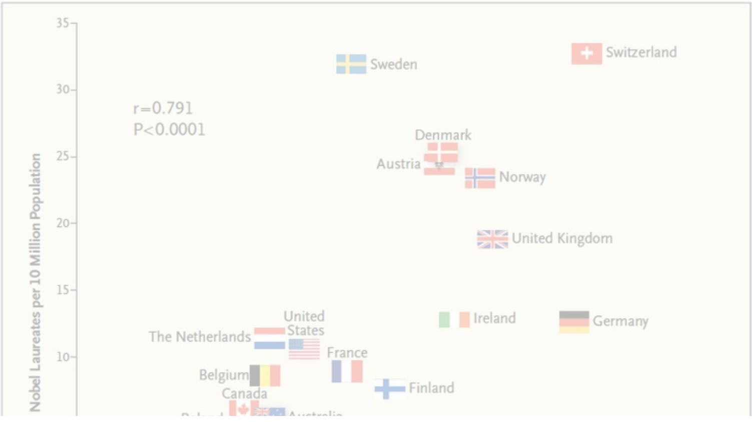


Sets of graphs that fit the data and background knowledge



Summary graph

Causal Representation Learning



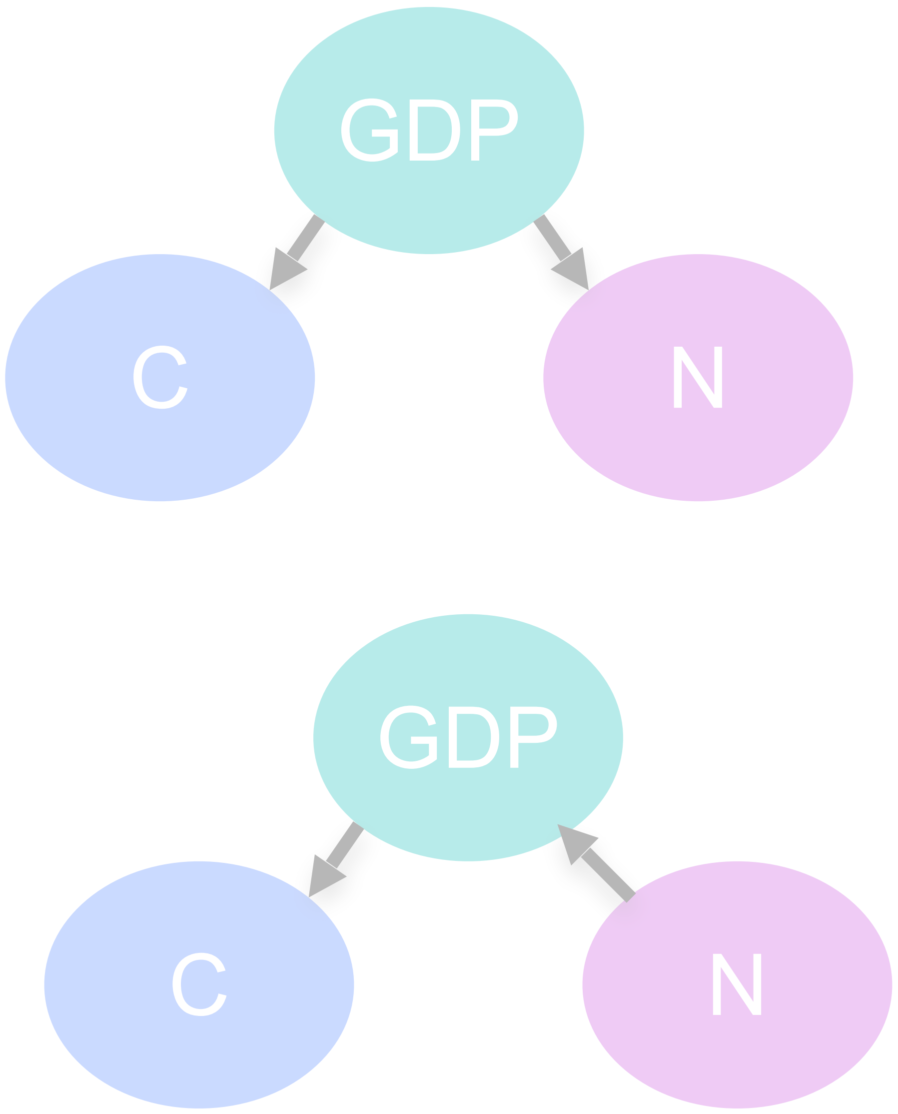
What if the causal variables are not directly observed (but we have high-dimensional observations, e.g. images)?

12	30	86k
10	20	46k
….	…	…

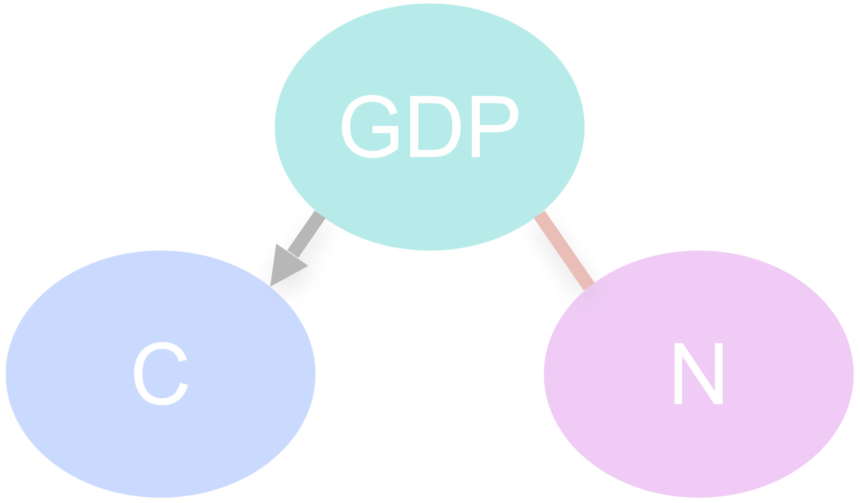
Observational data

$$C \nrightarrow GDP$$

[Optional] Background knowledge

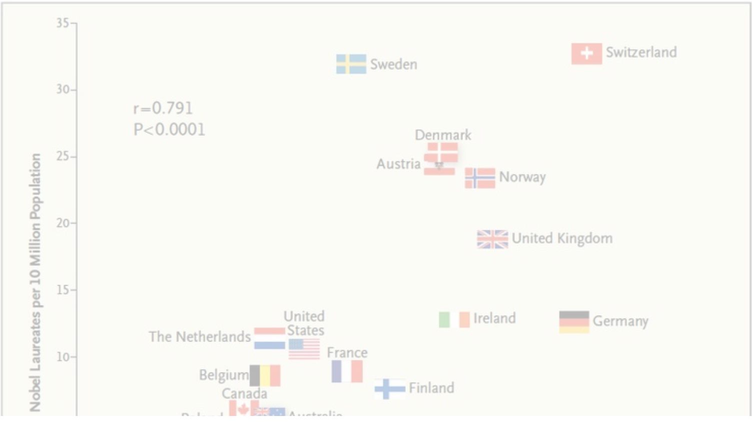


Sets of graphs that fit the data and background knowledge



Summary graph

Causal Representation Learning



What if the causal variables are not directly observed (but we have high-dimensional observations, e.g. images)?

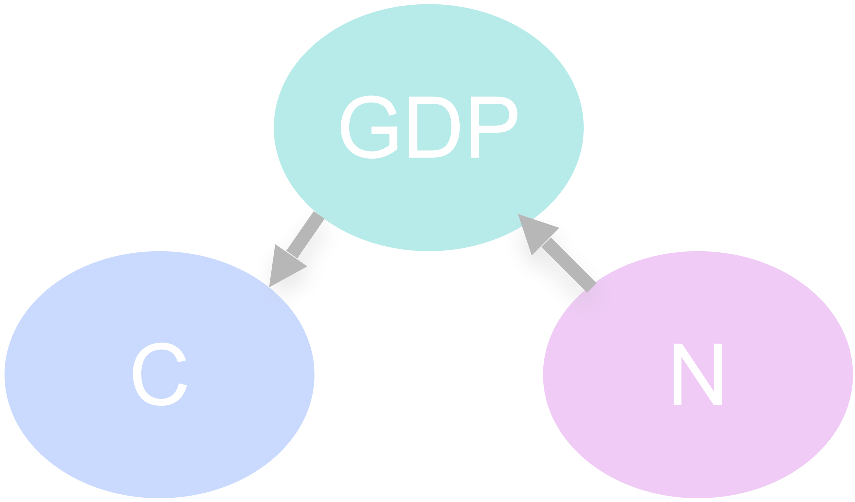
12	30	86k
10	20	46k
....	...	...

Observational data

$$C \nrightarrow GDP$$

[Optional] Background knowledge

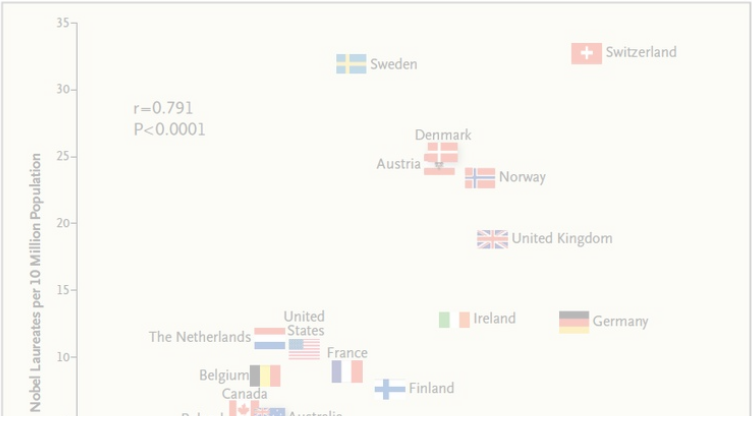
Task 1: learn/disentangle the causal variables



Sets of graphs that fit the data and background knowledge

Summary graph

Causal Representation Learning



What if the causal variables are not directly observed (but we have high-dimensional observations, e.g. images)?

12	30	86k
10	20	46k
....	...	...

Observational data

$C \nrightarrow GDP$

[Optional] Background knowledge

Task 1: learn/disentangle the causal variables

Task 2: learn the causal graph (or equivalence class)

Sets of graphs that fit the data and background knowledge

Summary graph

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

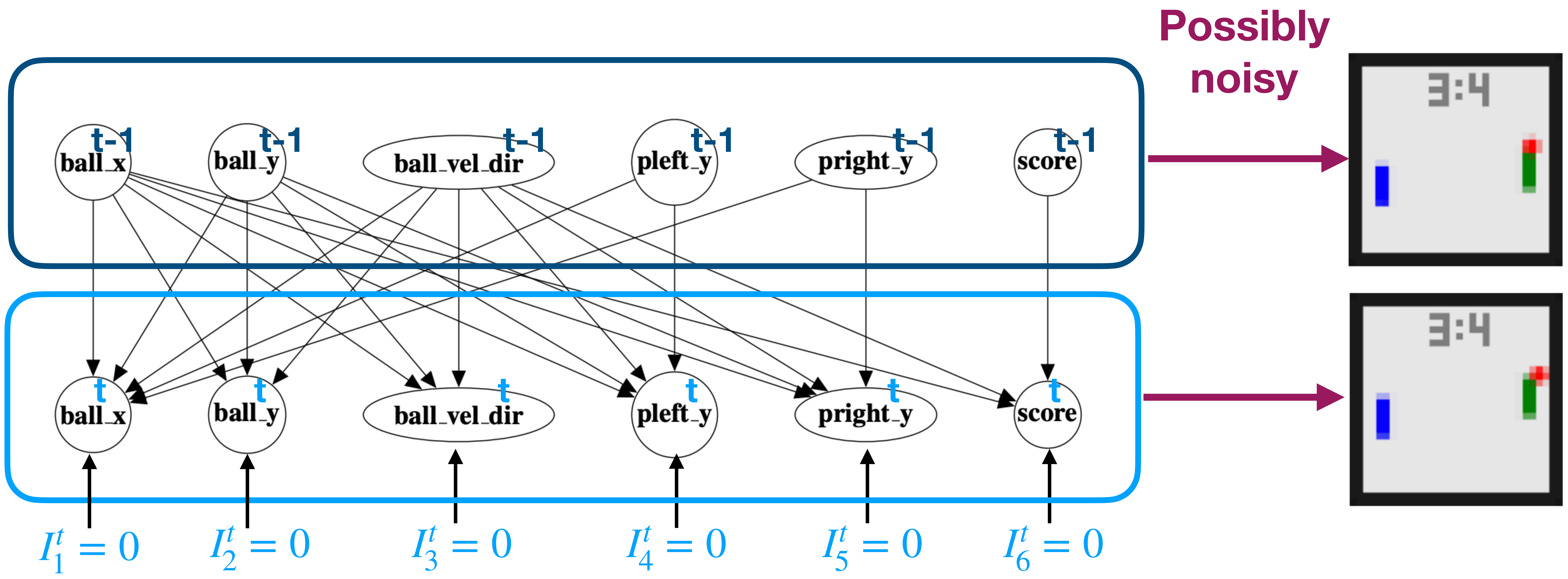
https://arxiv.org/abs/2102.11107

and many many more since….

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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

ICML 2022

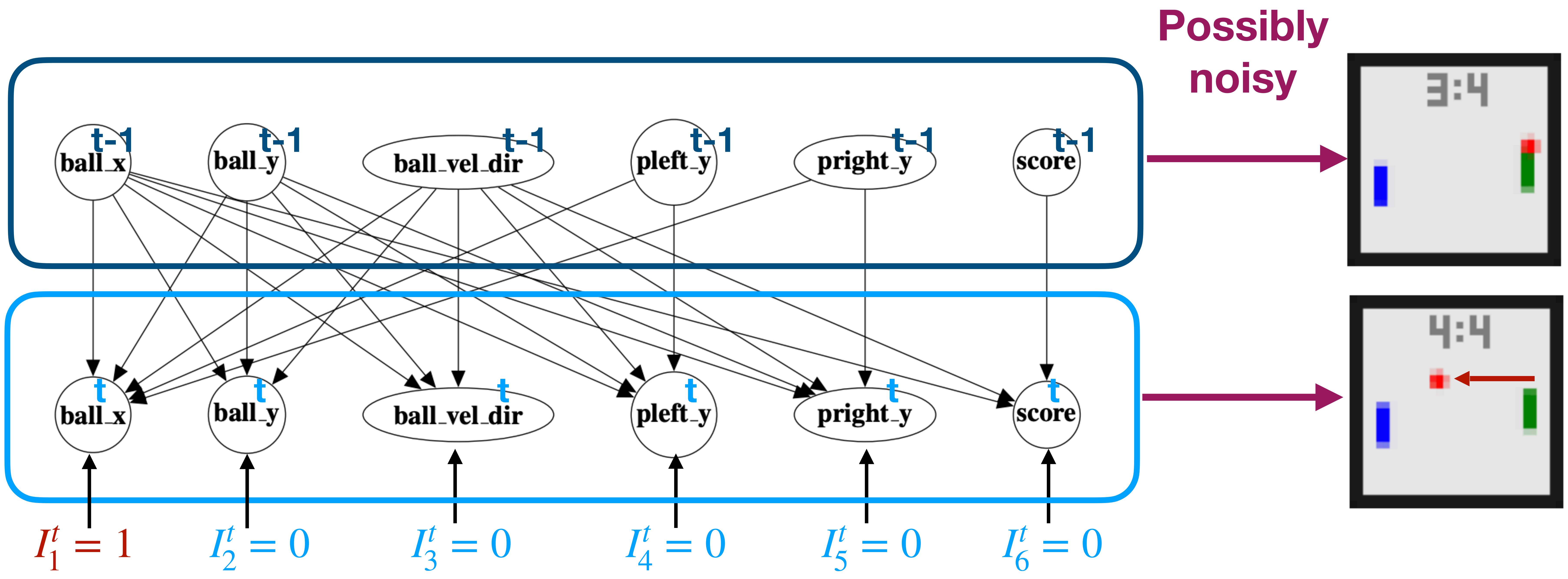


https://arxiv.org/abs/2202.03169

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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

ICML 2022



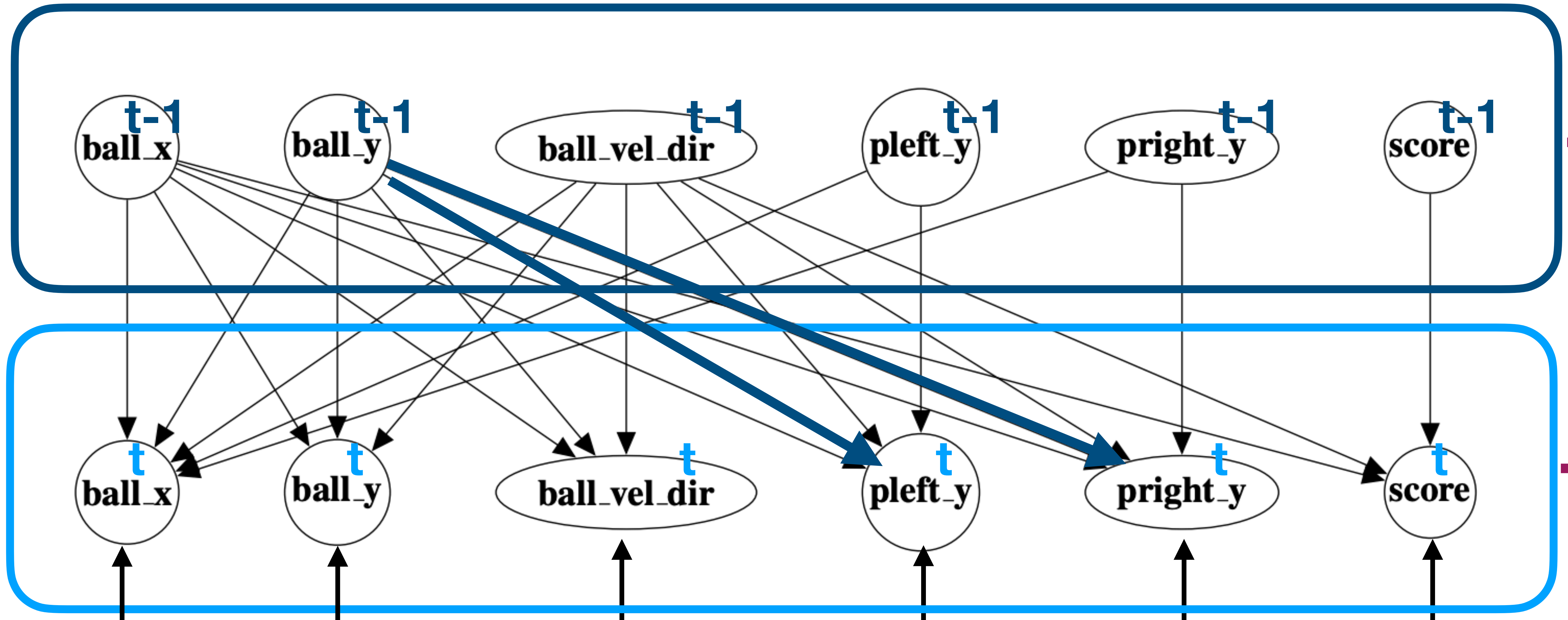
Stochastic intervention
(we don’t know where the ball will be)

https://arxiv.org/abs/2202.03169

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

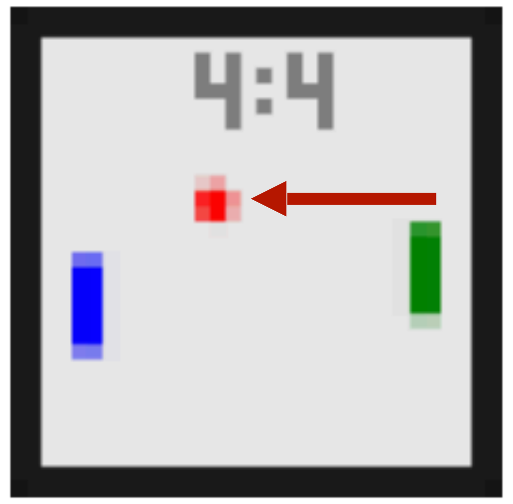
Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves

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



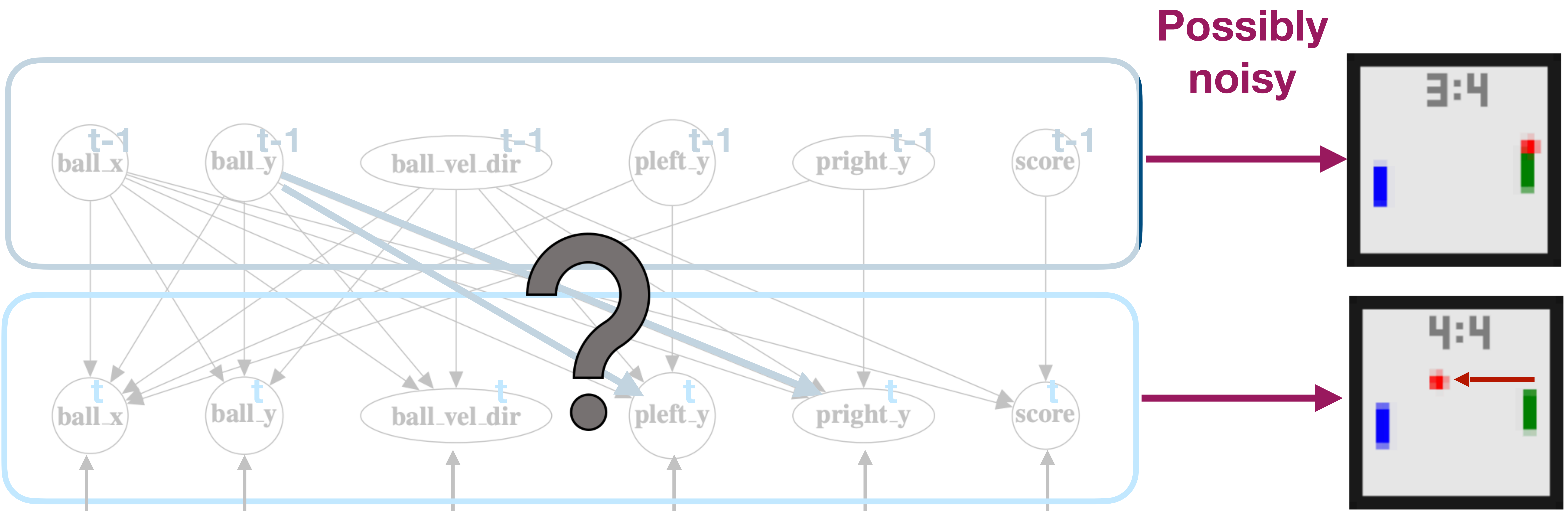
Stochastic intervention
(we don't know where the ball will be)

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



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

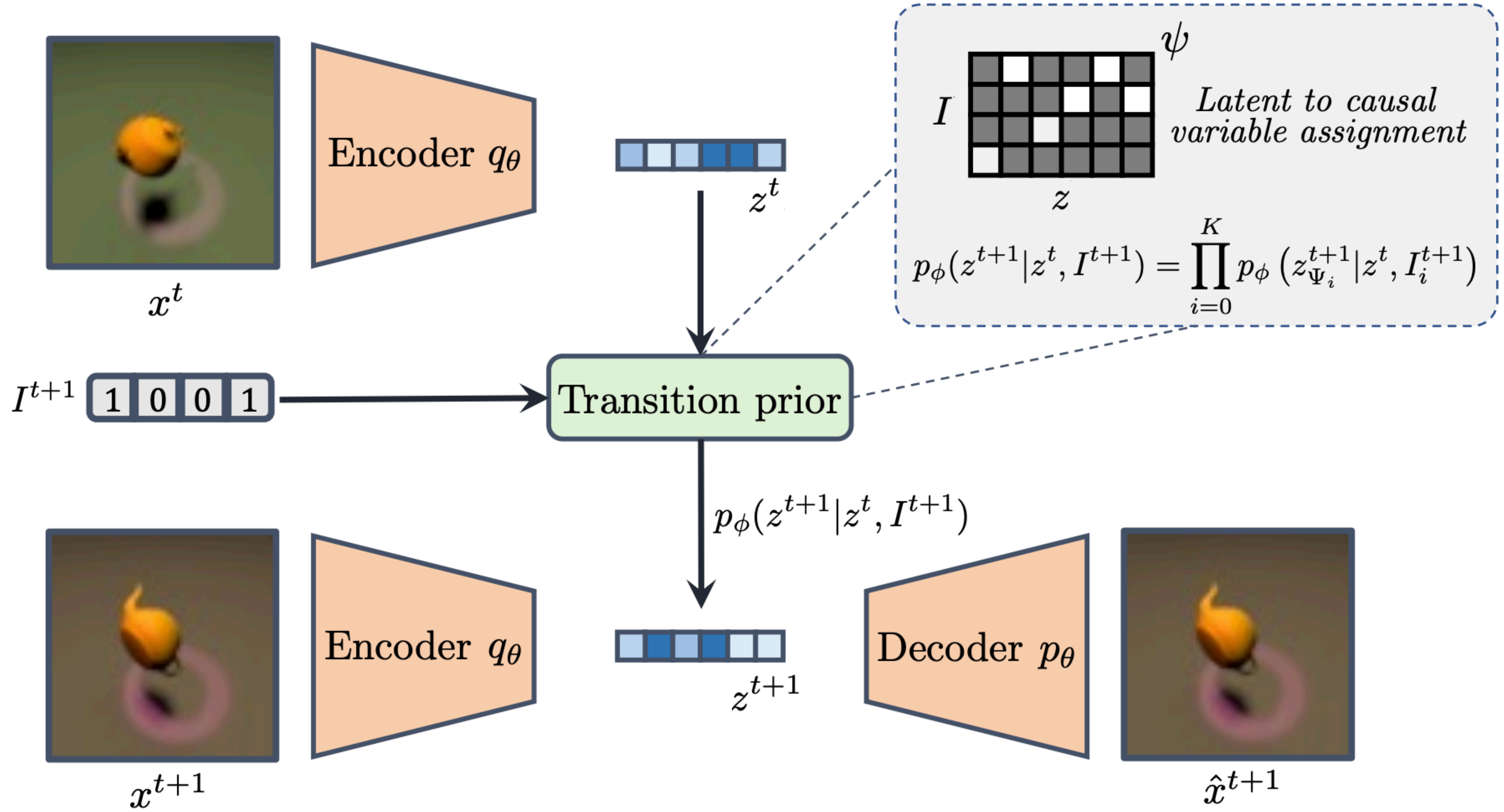
Stochastic intervention
(we don't know where the ball will be)

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



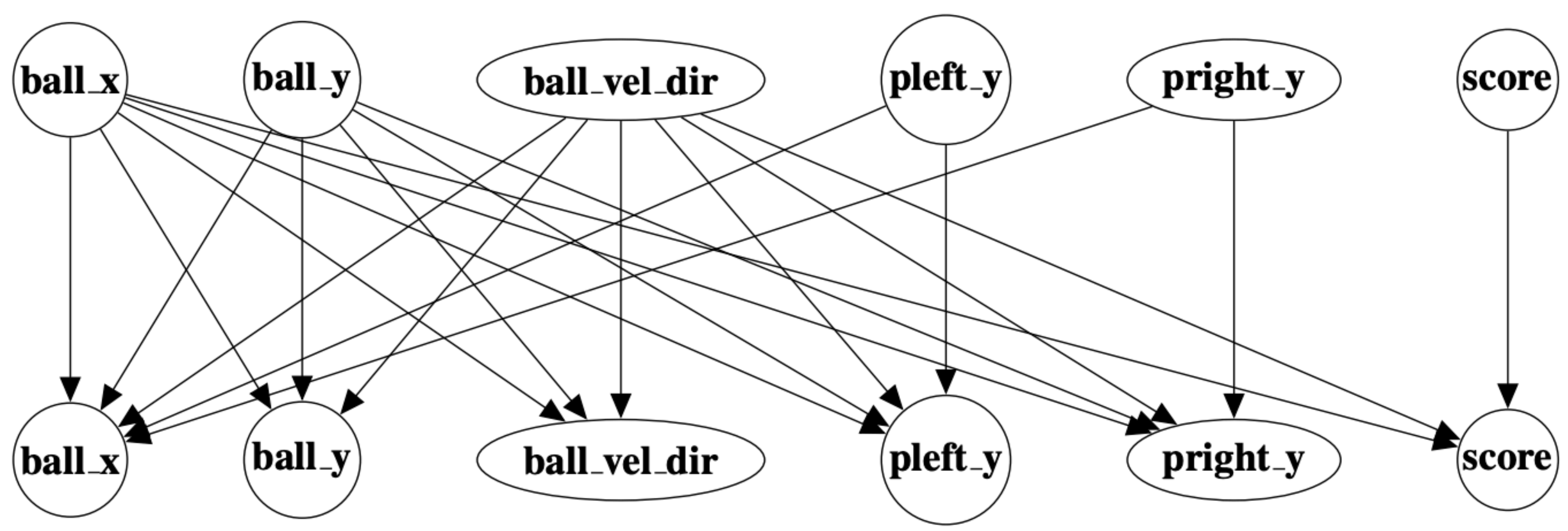
CITRIS-VAE

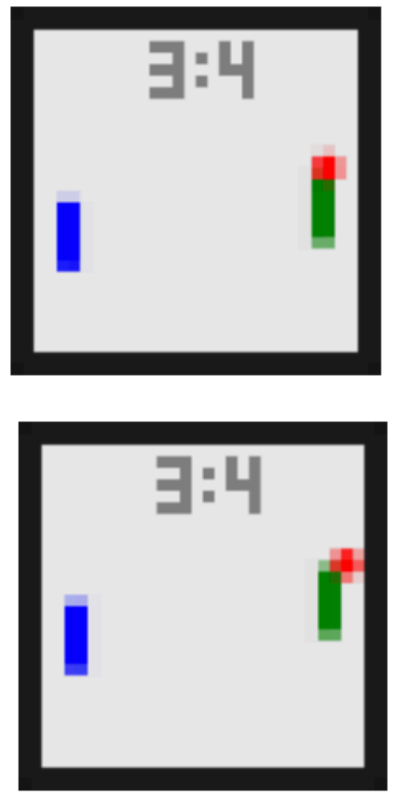
Also CITRIS-NF…

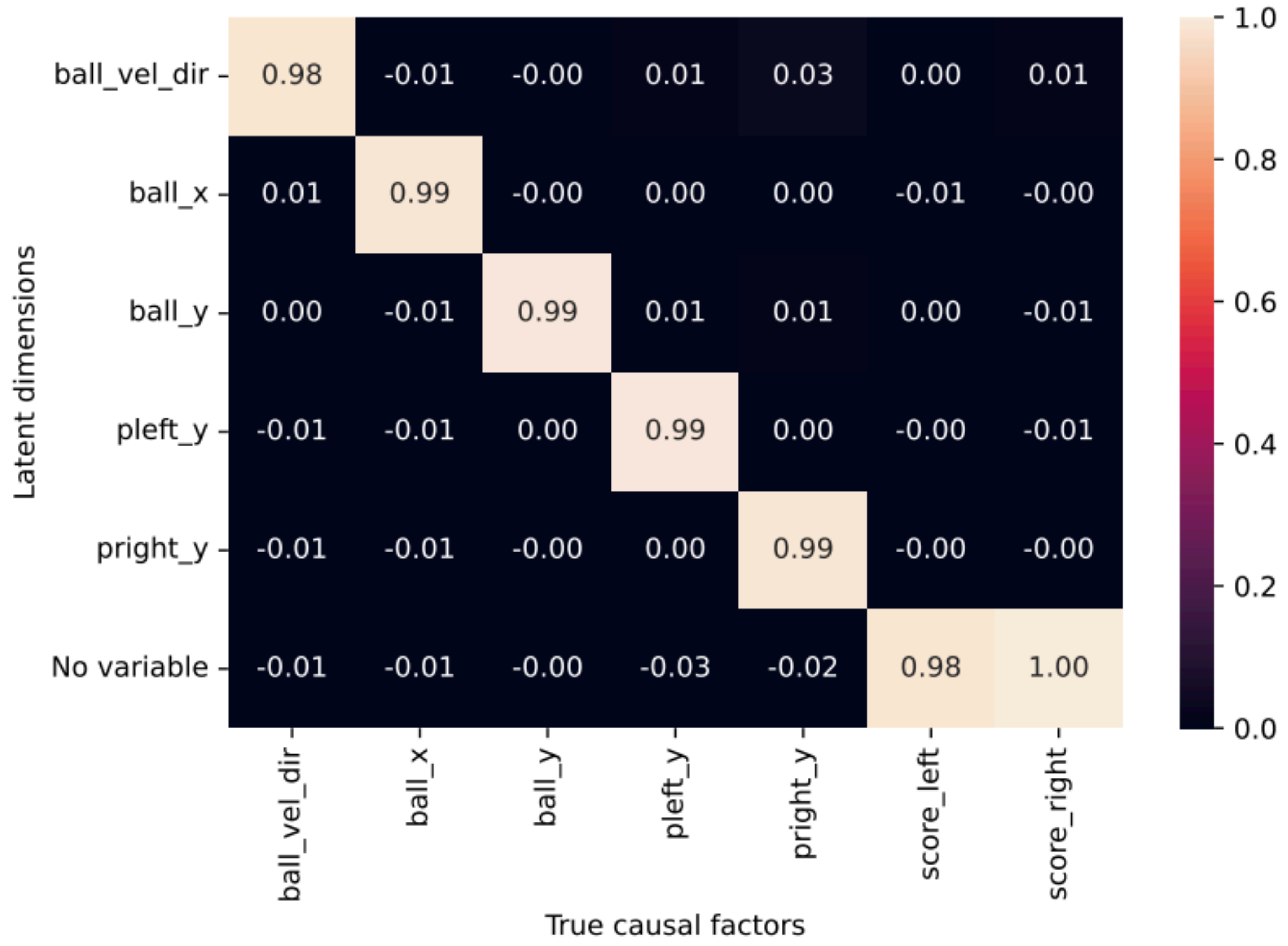
https://arxiv.org/abs/2202.03169

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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





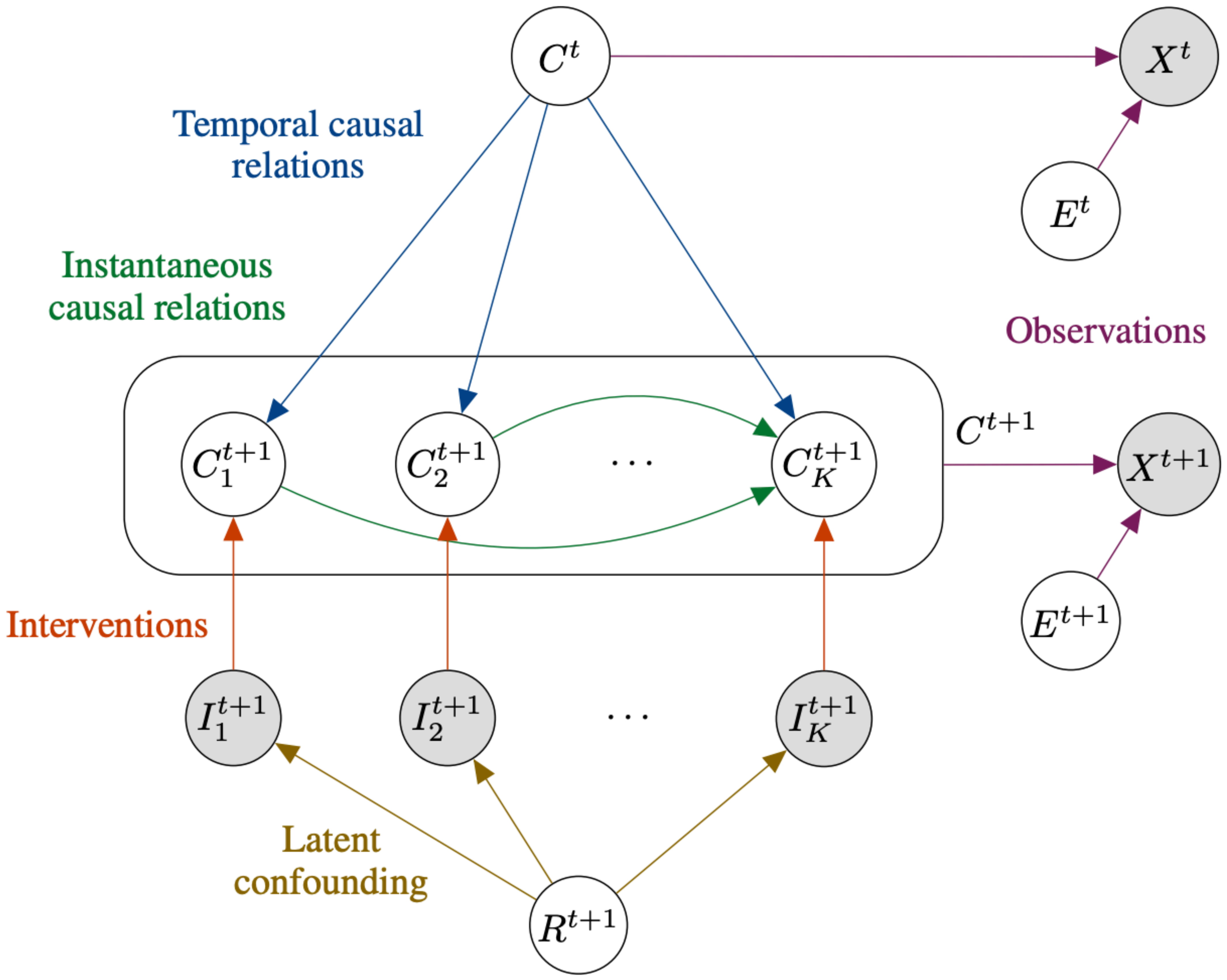


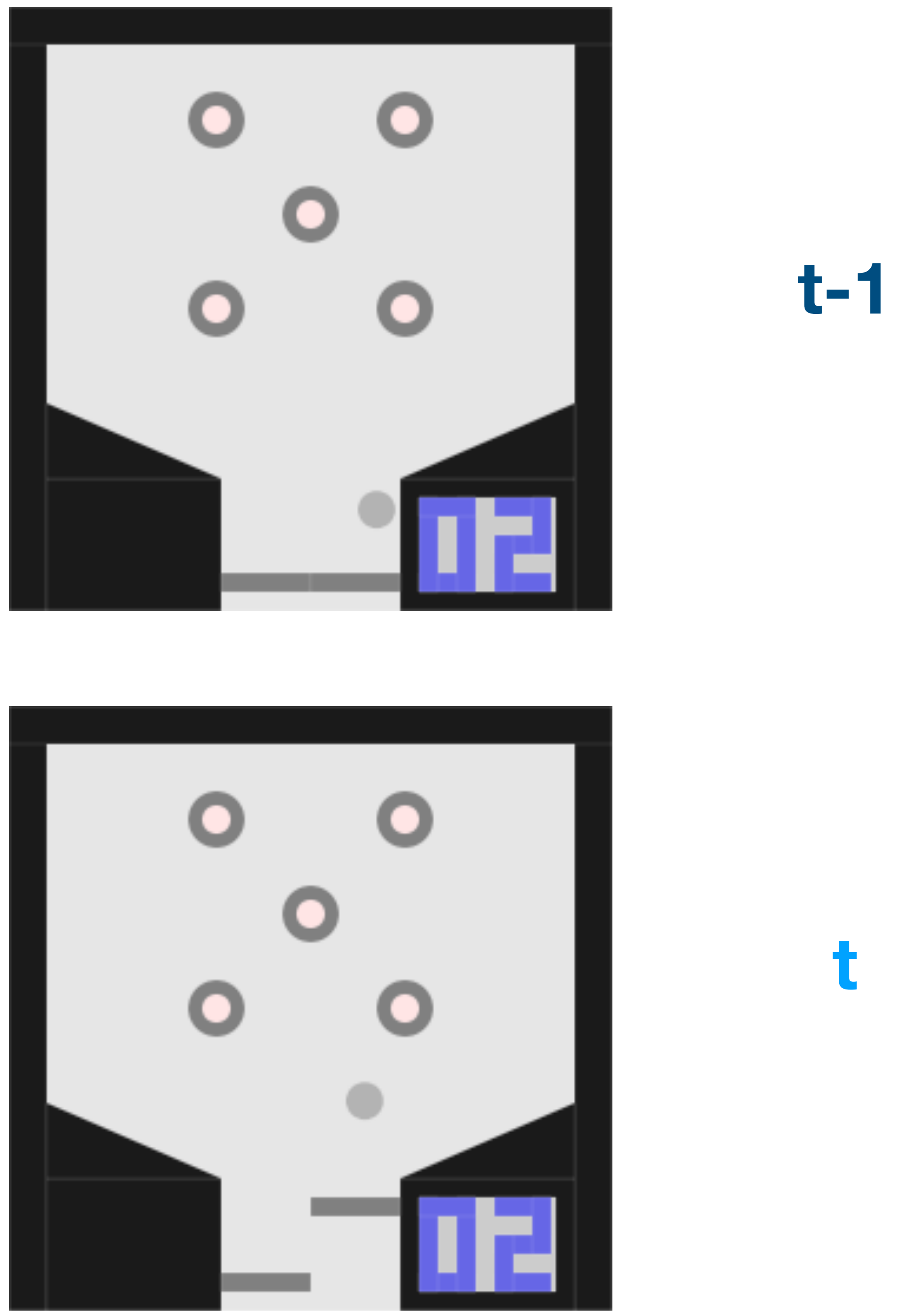
https://arxiv.org/abs/2202.03169

iCITRIS: Causal Representation Learning for Instantaneous Temporal Effects

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

ICLR 2023



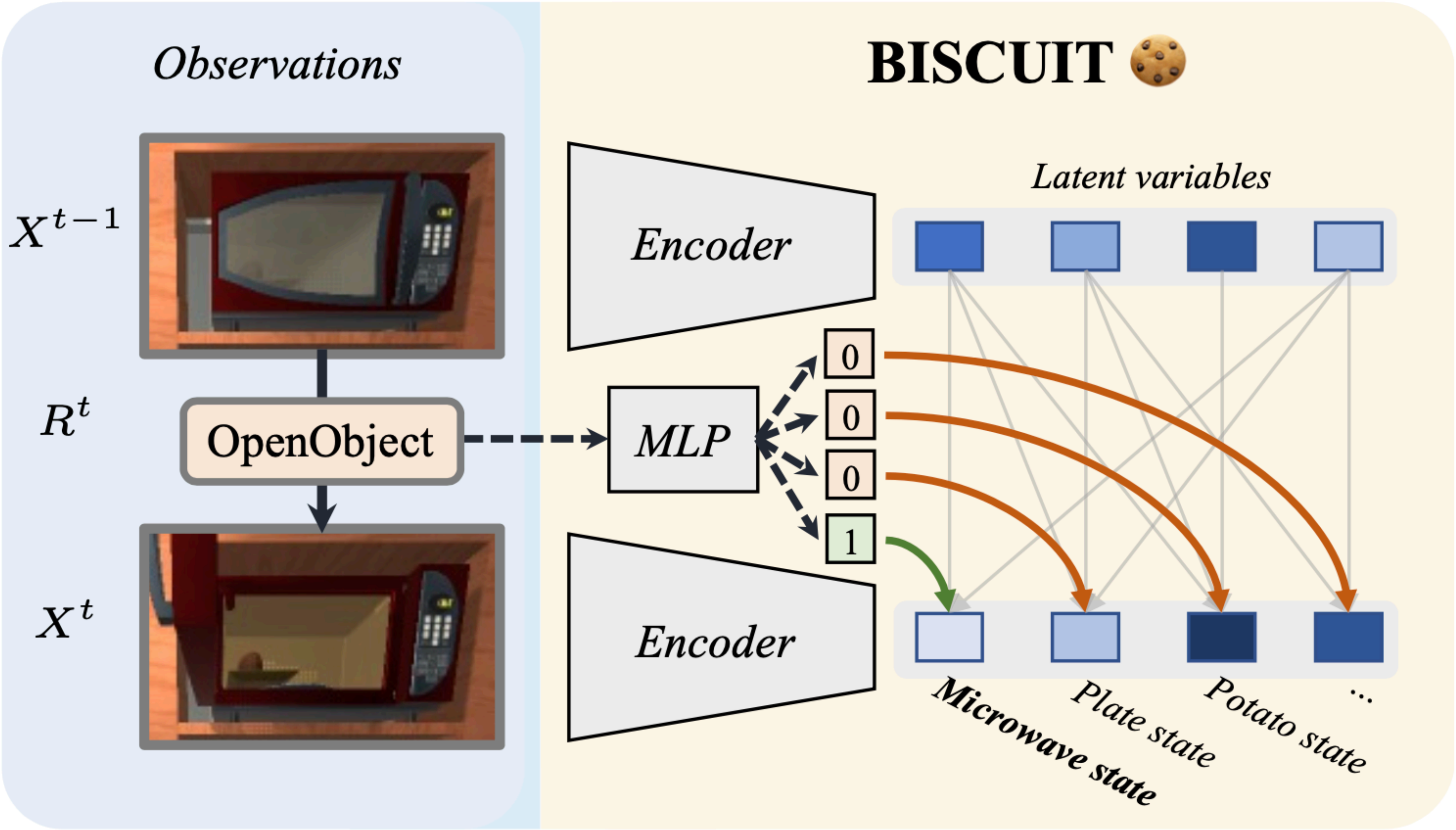


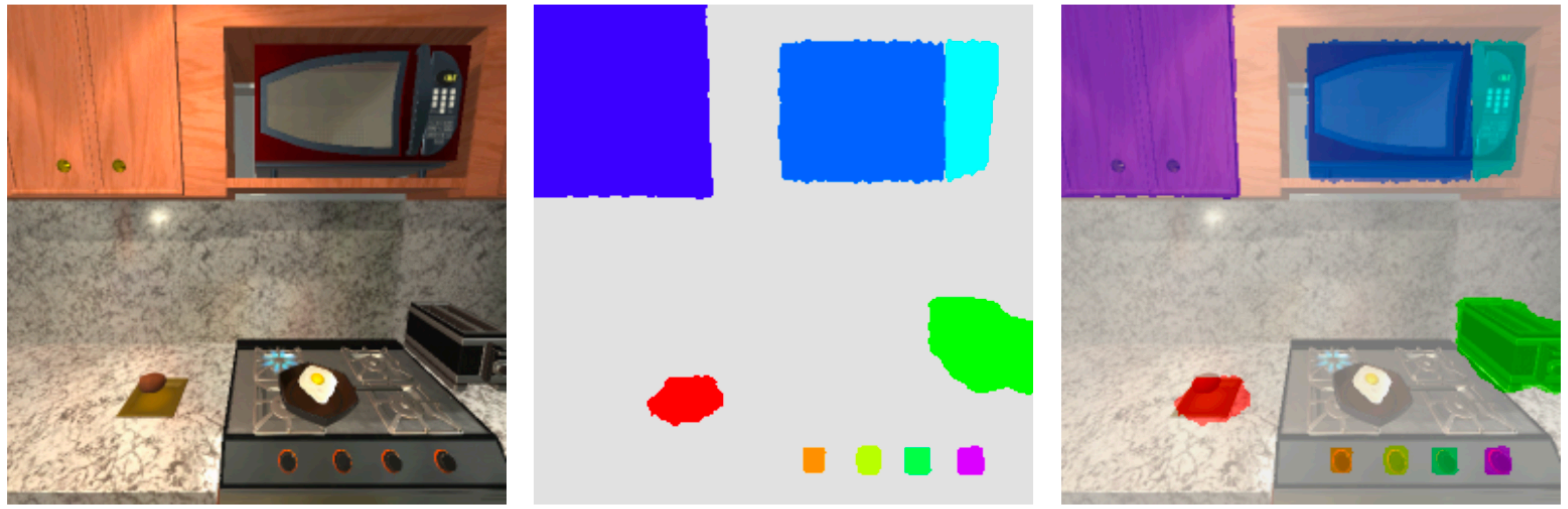
https://arxiv.org/abs/2206.06169

BISCUIT: Causal Representation Learning from Binary Interactions

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

UAI 2023



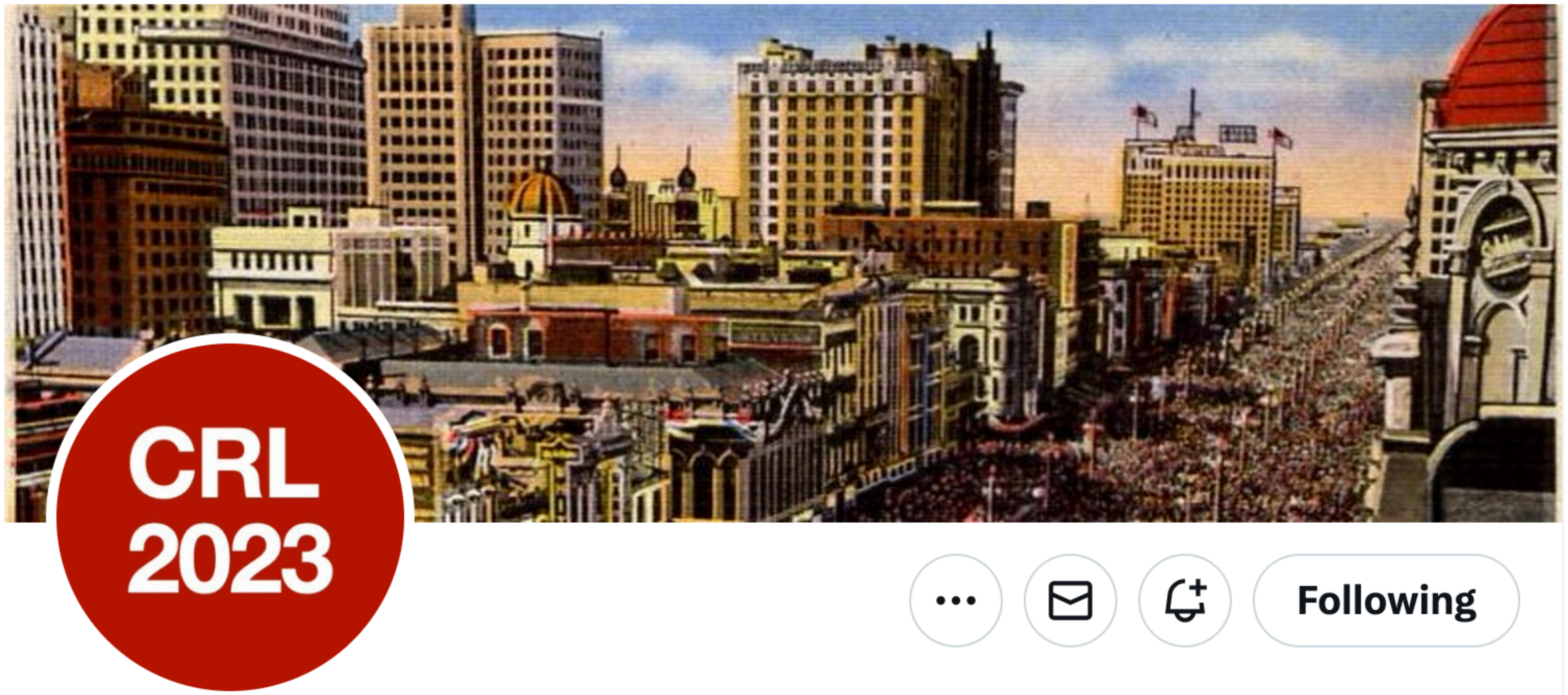


Input Image Learned Interactions Combined Image

Input Image 1 Input Image 2 Generated Output

https://phlippe.github.io/BISCUIT/

Call for Papers in CRL



Following

CRL workshop at NeurIPS 2023

@crl_neurips2023 Follows you

Causal Representation Learning workshop at @NeurIPSConf 2023 in New Orleans

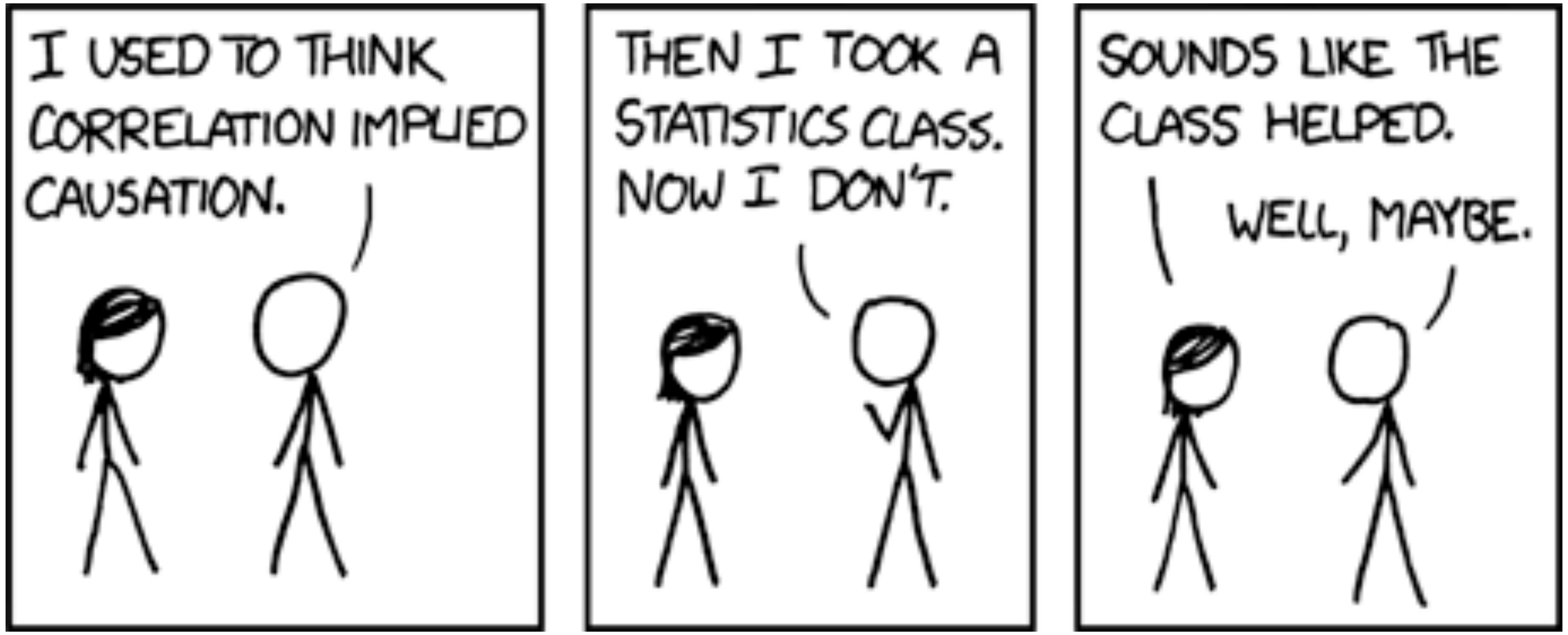
Submission deadline: Oct 2, 2023, 23:59 AoE
Workshop date: Dec 15 or 16, 2023

New Orleans, USA crl-workshop.github.io Joined July 2023

https://crl-workshop.github.io/

Thanks! Questions?

(joint work with Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, Joris Mooij, Biwei Huang, Fan Feng, Chaochao Lu and Kun Zhang)



https://xkcd.com/552/