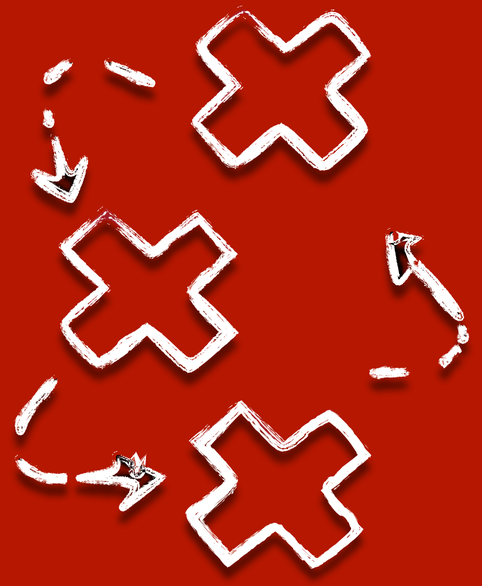


Seminar: Advanced topics in causality and causal ML research

Lecture 4: Causality-inspired ML/RL

Lecturer: Sara Magliacane

UdS - Summer 2026

Causality + machine learning (non-exhaustive list)

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

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

Causality + machine learning (non-exhaustive list)

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

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

Causality + machine learning (non-exhaustive list)

1. Machine learning (ML) helps causality

- Causal discovery - learning causal graphs from data
- CRL - learning causal variables from observations
- Causal effect estimation - matching, weighting, double ML

2. Causality (in the most general definition) helps machine learning

- Robustness, Transfer learning
- Reinforcement Learning
- Bias mitigation, fairness, interpretability

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

Causal Hierarchy [Pearl 2009, 2018]

Most ML

Causality

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

Causal Hierarchy [Pearl 2009, 2018]

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?
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*

Causality + machine learning (non-exhaustive list)

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

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

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** $\approx$ change of distribution of P(X| parents)

Causality vs Transfer learning

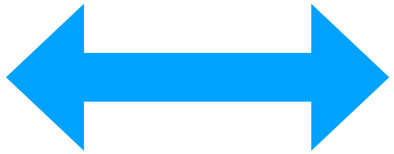
- Transfer learning:
  - How can I predict what ha ens when when the distribution chang an

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** ≈ change of distribution of P(X| parents)

On Causal and Anticausal Learning

ICML 2012 Test of Time

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

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

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 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

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

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

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

Invariance, Causality and Robustness

2018 Neyman Lecture *

Peter Bühlmann †
Seminar for Statistics, ETH Zürich

A Causal View on Robustness of Neural Networks

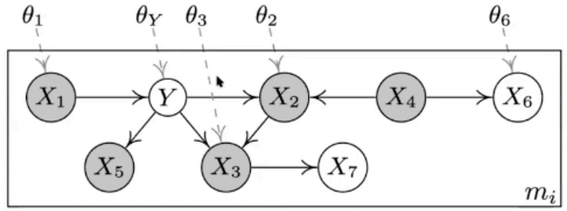
Cheng Zhang * Microsoft Research Cheng.Zhang@microsoft.com
Kun Zhang Carnegie Mellon University kunz1@cmu.edu
Yingzhen Li * Microsoft Research Yingzhen.Li@microsoft.com

and many many more....

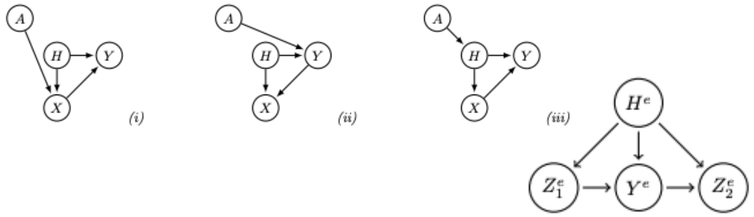
Causality allows us to reason systematically about distribution shifts, e.g. through graphs

On Causal and Anticausal Learning

Domain Adaptation as a Problem of Inference on Graphical Models

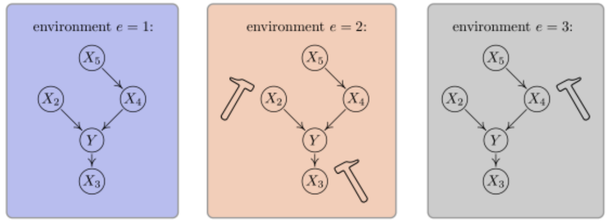


Anchor regression: heterogeneous data meet causality

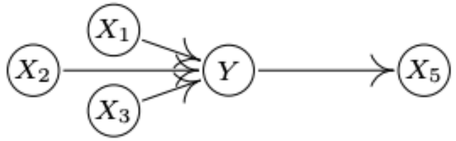


J. R. Statist. Soc. B (2016)
78, *Part* 5, *pp.* 947–1012

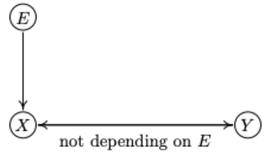
Causal inference by using invariant prediction: identification and confidence intervals



Invariant Models for Causal Transfer Learning

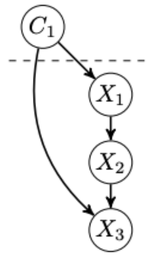


Invariance, Causality and Robustness

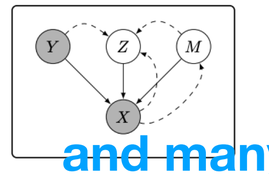


Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests

Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions



A Causal View on Robustness of Neural Networks



and many many more....

Unsupervised domain adaptation - toy example

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

Source domain

No labels in target

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

Target domain

- **Task:** Learn a model $Y = \hat{f}(X1, X2)$ on the source domain so that it can reliably estimate Y in target domain

Domain adaptation from the graphical perspective

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

Source domain

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

Target domain

1. We add a context variable C to distinguish the domains

Domain adaptation from the graphical perspective

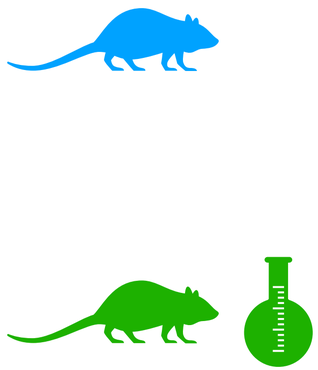
C	X1	X2	Y
0	0,1	2	0
0	0,2	3	0
0	1,1	2	1
0	0,1	3	0
1	3,1	2	?
1	3,2	3	?
1	4	2	?
1	3,2	3	?

Source domain

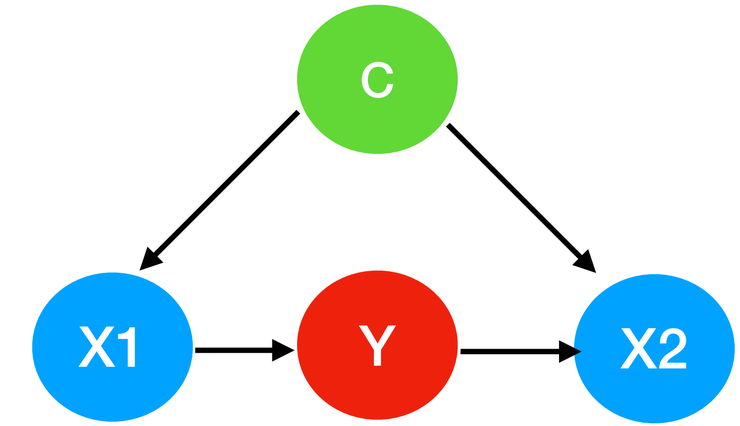
Target domain

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

Domain adaptation from the graphical perspective



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

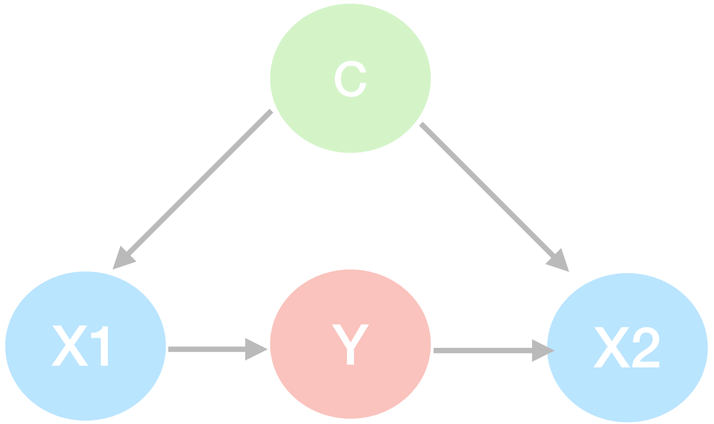


We can represent P(X1, X2,Y, C) with an **(unknown)** causal graph

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

Separating features = safe for (adversarial) domain adaptation

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

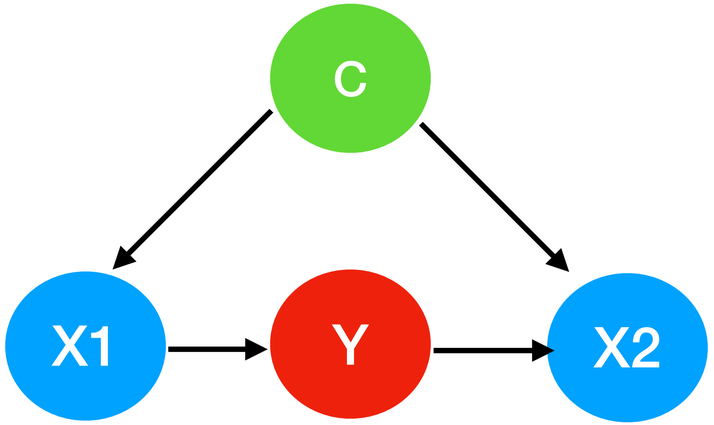


Step 1: Known causal graph

Separating features = safe for (adversarial) domain adaptation

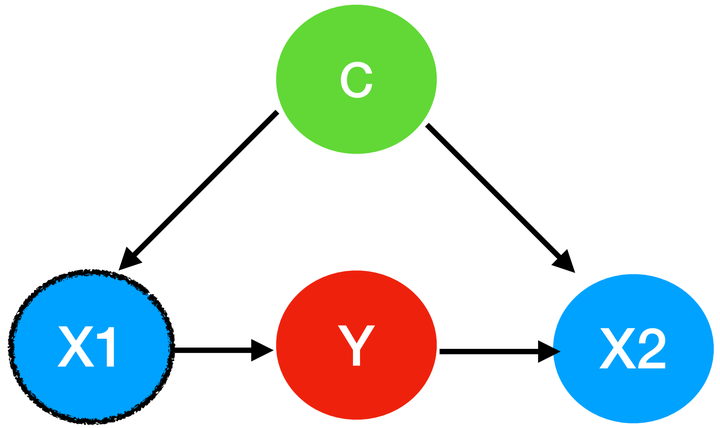
Aka stable features, invariant features etc.

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



Separating features = safe for (adversarial) domain adaptation

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

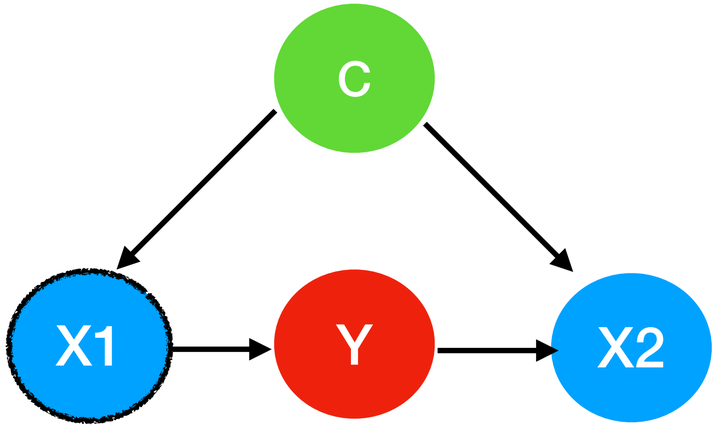


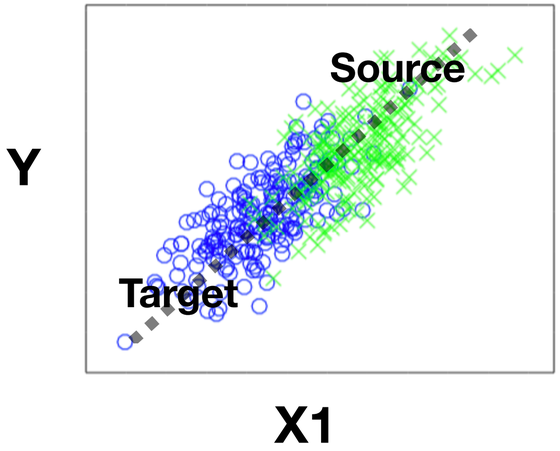
$$Y \perp\!\!\!\perp_d C \mid X_1 \iff Y \perp\!\!\!\perp C \mid X_1$$

(under Markov and faithfulness assumptions)

Separating features = safe for (adversarial) domain adaptation

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





$$Y \perp\!\!\!\perp_d C \,|\, X_1 \iff Y \perp\!\!\!\perp C \,|\, X_1$$

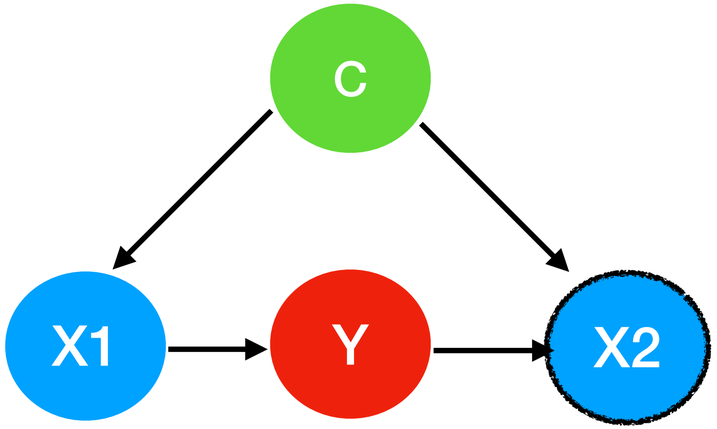
(under Markov and faithfulness assumptions)

$$Y \perp\!\!\!\perp C \,|\, X_1 \equiv$$

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

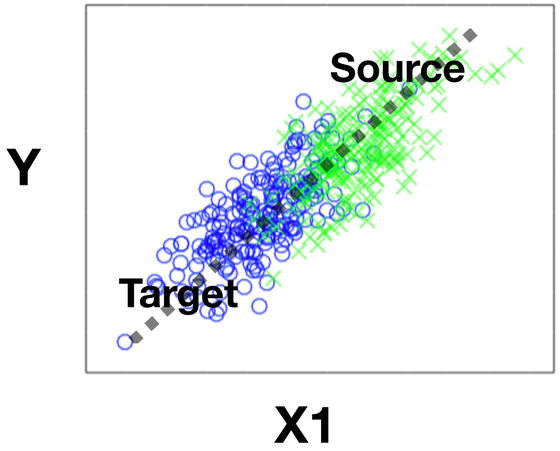
Separating features = safe for (adversarial) domain adaptation

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



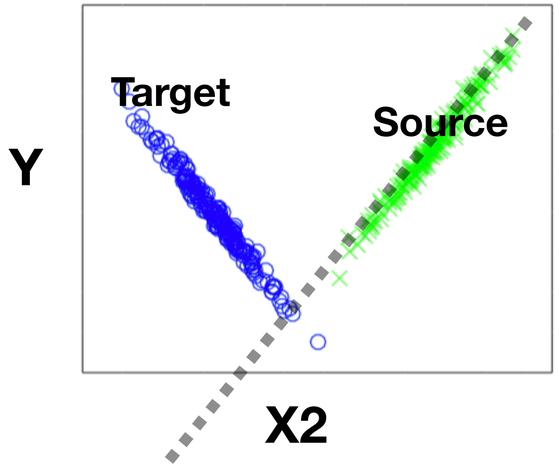
$$Y \not\perp\!\!\!\perp_d C \,|\, X_2 \iff Y \not\perp\!\!\!\perp C \,|\, X_2$$

(under Markov and faithfulness assumptions)



$$Y \perp\!\!\!\perp C \,|\, X_1 \equiv$$

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

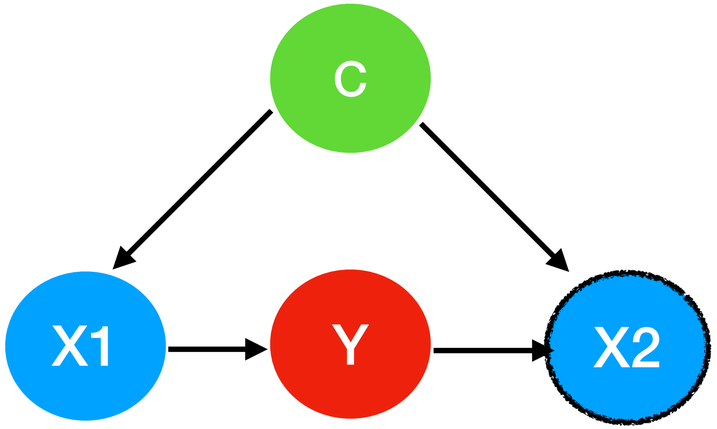


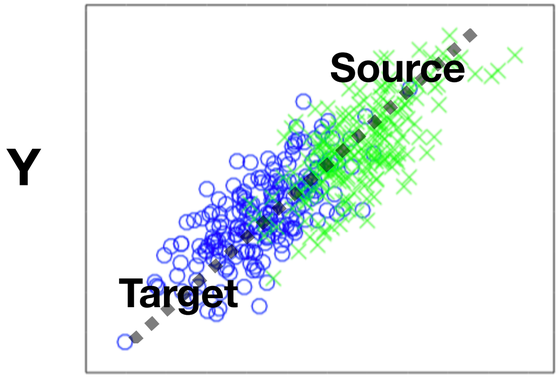
$$Y \not\perp\!\!\!\perp C \,|\, X_2 \equiv$$

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

Separating features = safe for (adversarial) domain adaptation

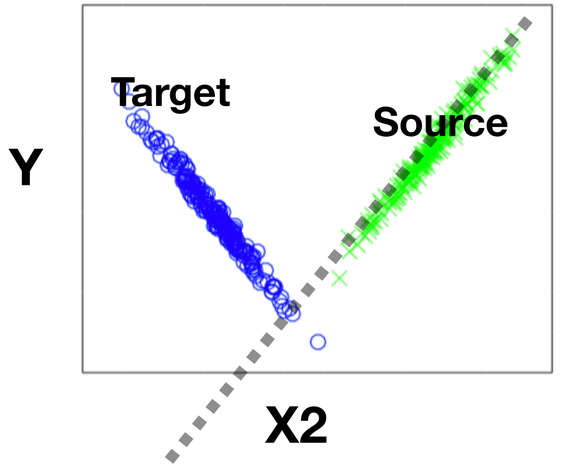
- **Separating features:** sets of features that d-separate Y from the context





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

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



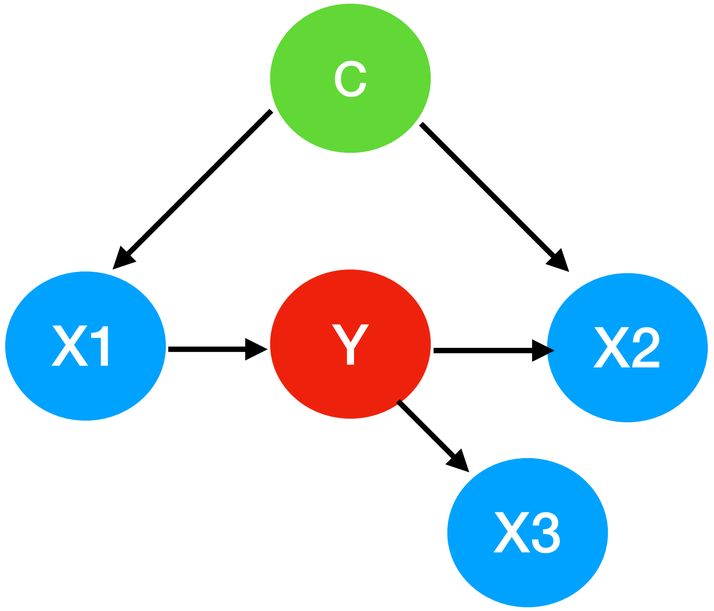
$$Y \not\perp\!\!\!\perp C \mid X_2 \equiv$$

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

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

Separating features = safe for (adversarial) domain adaptation

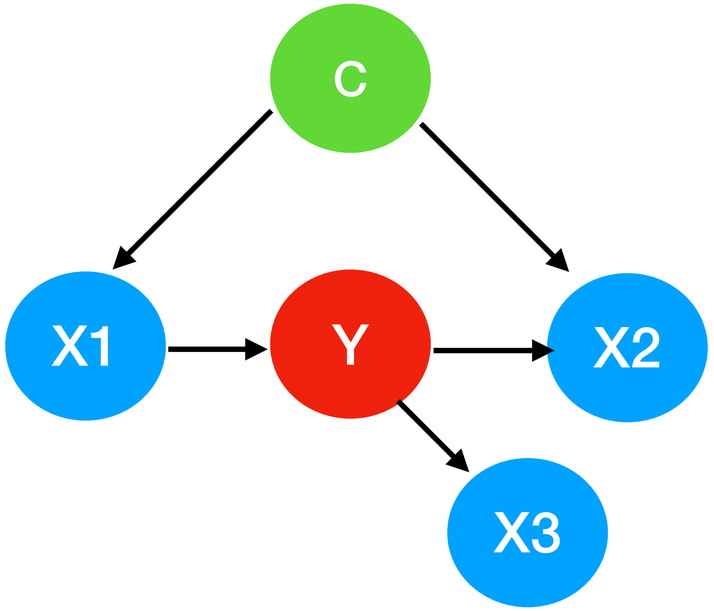
- **Separating features:** sets of features that d-separate Y from the context



$$Y \perp_d C \mid \{X_1, X_3\} \; ?$$

Separating features = safe for (adversarial) domain adaptation

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

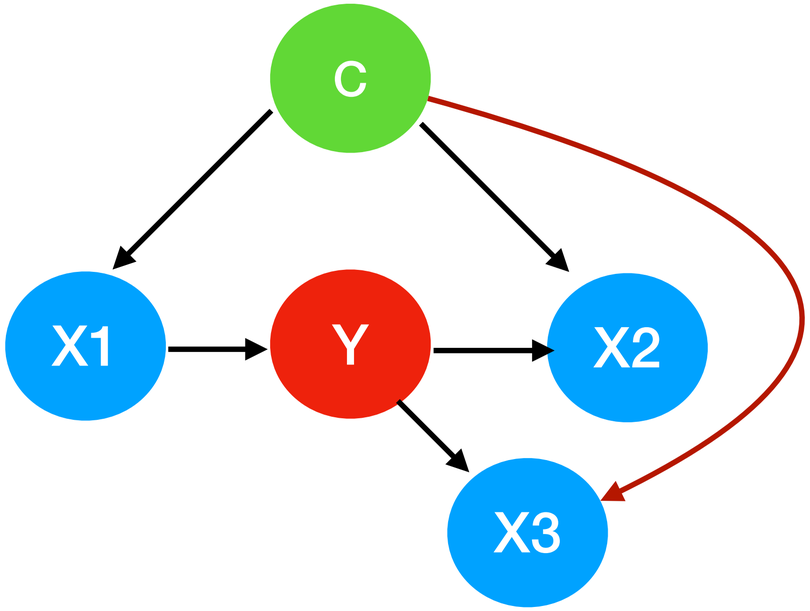


$$Y \perp_d C \mid \{X_1, X_3\} \; ?$$

$Y \perp_d C \mid \{X_1, X_3\}$ (separating features are not necessarily parents of Y)

Separating features = safe for (adversarial) domain adaptation

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

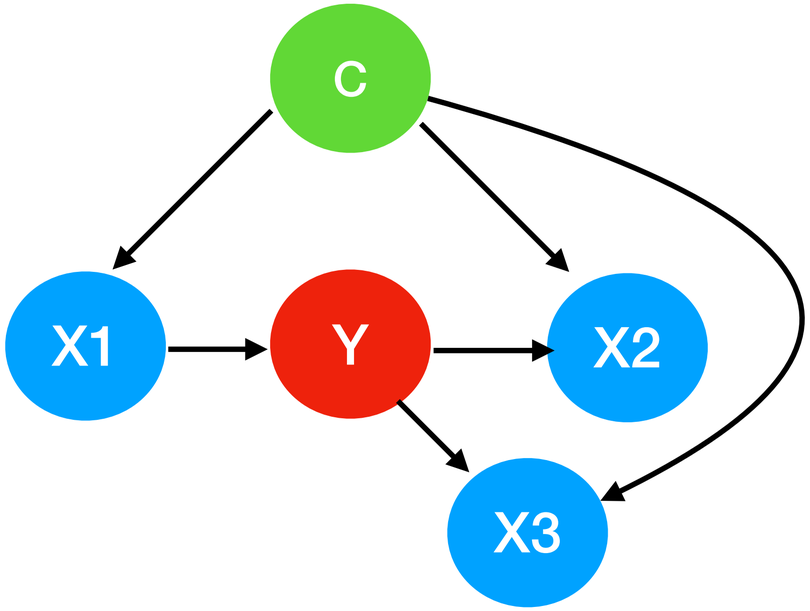


Intervention on every variable except Y = domain generalisation

$$Y \perp_d C \mid \{X_1, X_3\} \; ?$$

Separating features = safe for (adversarial) domain adaptation

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

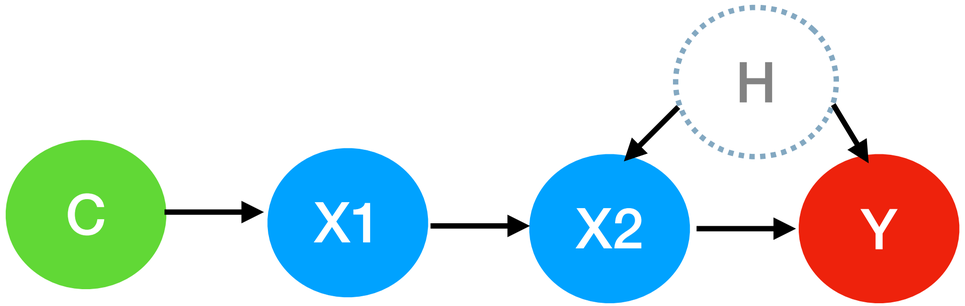


Intervention on every variable except Y = domain generalisation

$$Y \perp_d C \mid \{X_1, X_3\} \; ?$$

Under every possible distribution shift (except directly on the label Y), the separating set of feature are the **causal parents**

Observed parents are not enough under latent confounding



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

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

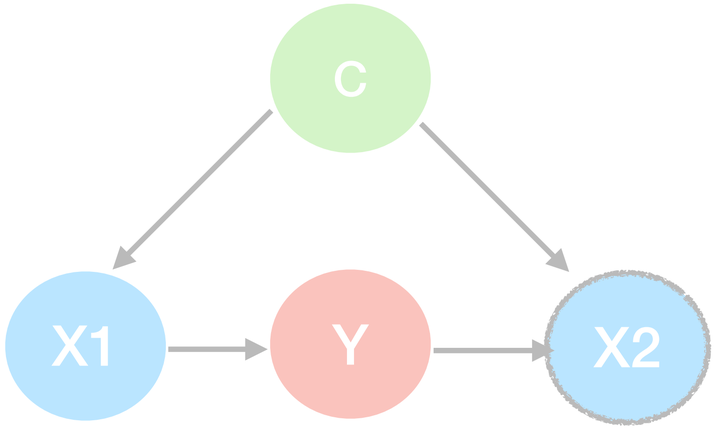
$$Y \perp\!\!\!\perp C \mid X_1, X_2$$

- $P(Y \mid X_1)$ is invariant, $P(Y \mid X_2)$ is not

- **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**

Separating features = safe for (adversarial) domain adaptation

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

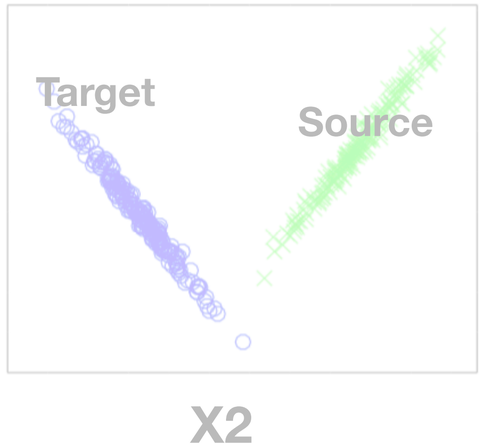


$$Y \not\perp\!\!\!\perp_d C \mid X_2$$

Step 2: Unknown graph and no labels in target!

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

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

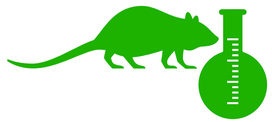


$$Y \not\perp\!\!\!\perp C \mid X_2 \equiv$$

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

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

Unsupervised multi-source domain adaptation - toy example



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

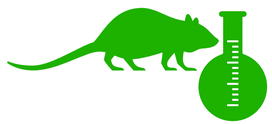
Source domains

Target domain

https://arxiv.org/abs/1707.06422

Domain Adaptation by Using Causal Inference to Predict Invariant Conditional Distributions

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	Source domains
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	?	Target domain
1	0	0,3	0	?	
1	0	0,3	1	?	

- 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

What if the causal graph is unknown?

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

$$Y \perp\!\!\!\perp C_1 \,|\, X_1 ? \qquad Y \perp\!\!\!\perp C_1 \,|\, X_2 ?$$

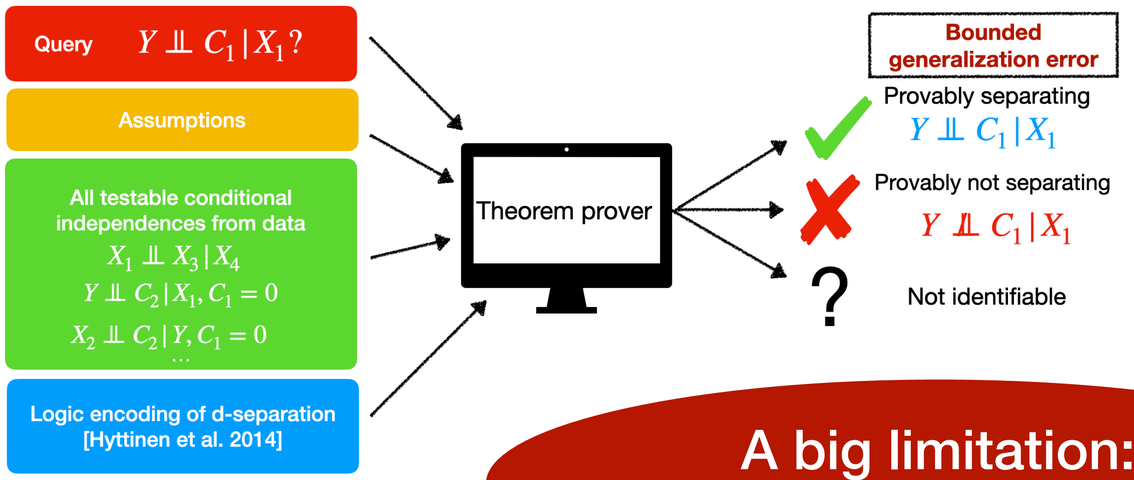
What if the causal graph is unknown?

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

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

- **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
0	1	0,2	0	0
0	1	0,3	0	1
0	1	0,3	1	0
1	0	3,1	2	?
1	0	3,2	3	?
1	0	4	3	?



A big limitation: scalability !

What if the causal graph is unknown?

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

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

- **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
0	1	0,2	0	0
0	1	0,3	0	1
0	1	0,3	1	0

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

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

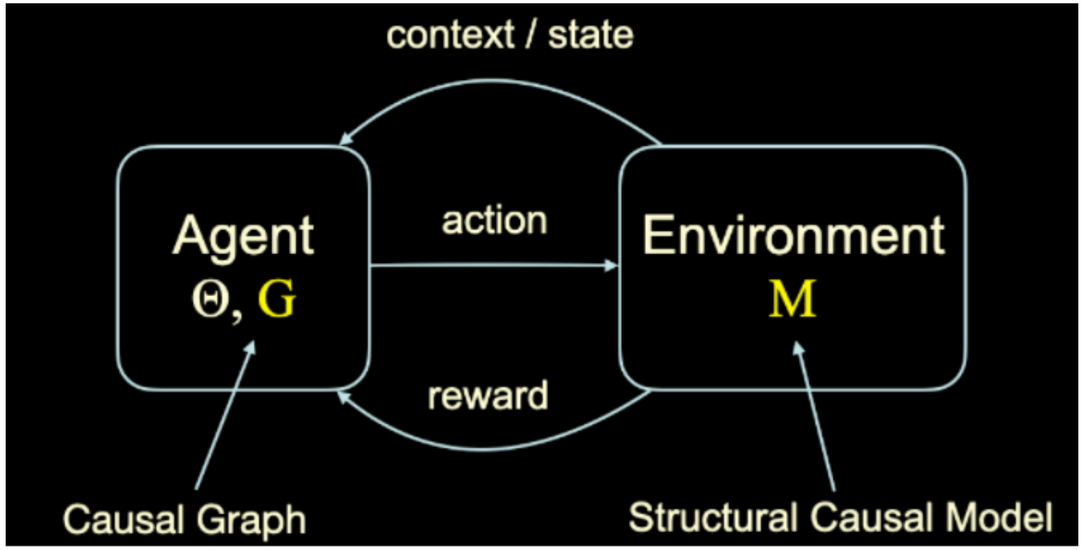
This assumes no new edges in the target

Causality + machine learning (non-exhaustive list)

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

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

This is just a sneak peak, a lot more causality+RL…



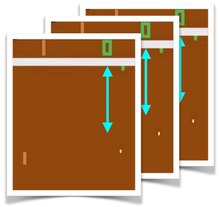
https://crl.causalai.net/

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

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

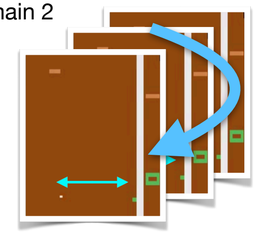
Source domains

Domain 1



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

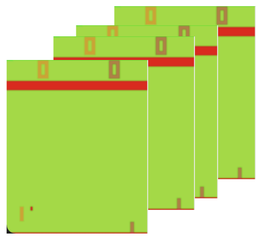
Domain 2



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

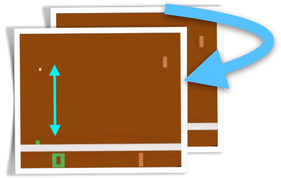
...

Domain n



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

Target domain



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

Simplifying assumption: no new edges/types of changes in target domain

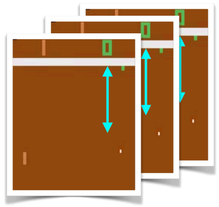
https://arxiv.org/abs/2107.02729

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

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

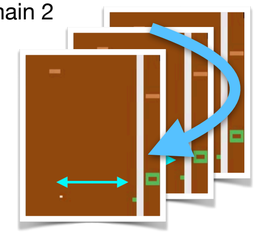
Source domains

Domain 1



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

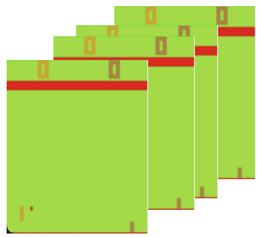
Domain 2



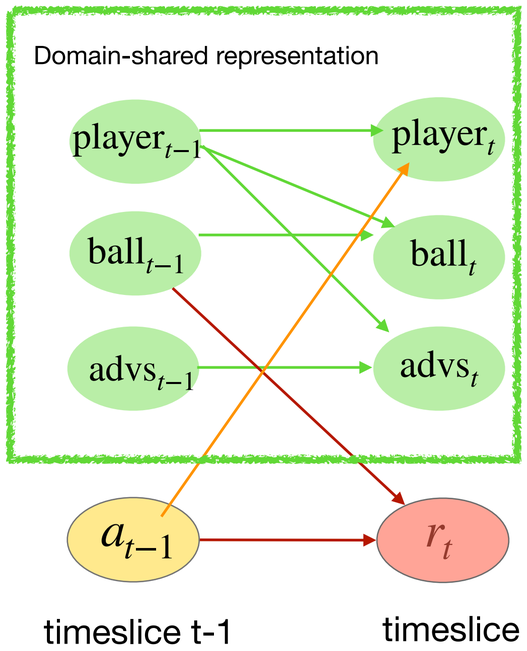
$\{\text{player}_t, \text{ball}_t, \text{advs}_t, a_t, \text{r}_t\}_{t=0,\dots,T}$

...

Domain n



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

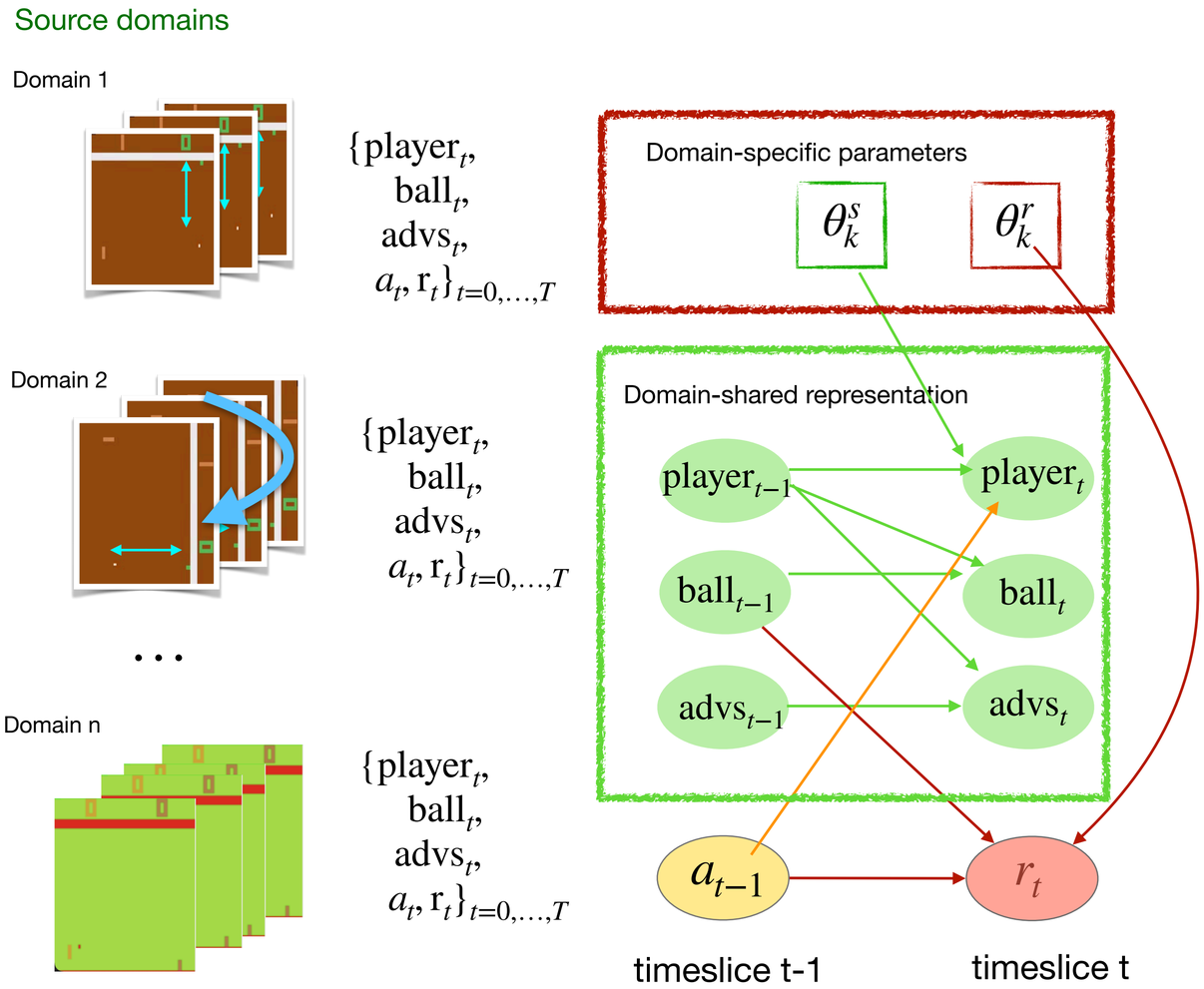


- Learn a **factored** MDP (symbolic inputs)

https://arxiv.org/abs/2107.02729

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

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

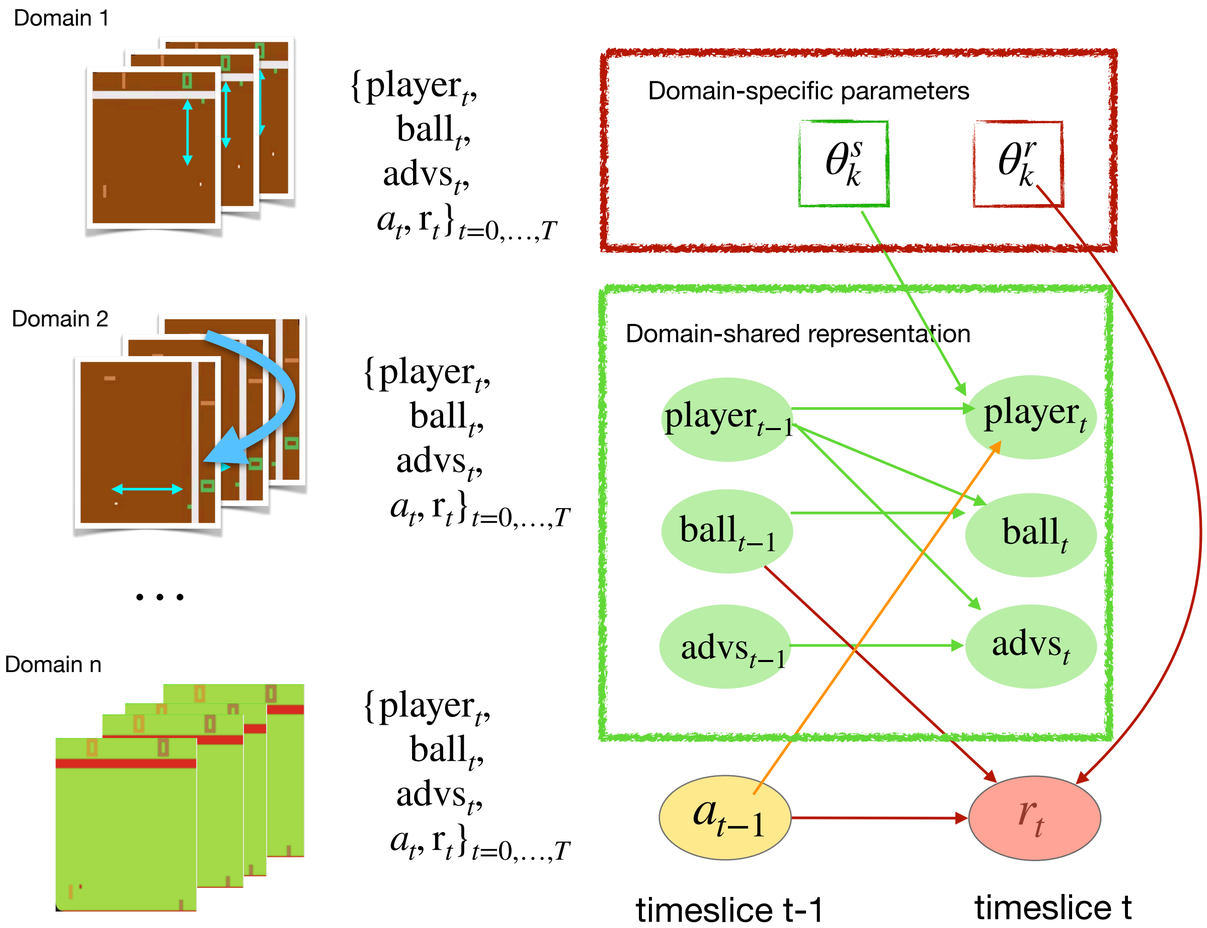


- Learn a **factored** MDP (symbolic inputs) with **latent change factors** that are constant in each domain, but **vary across domains**
- **Assume it holds in target**

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

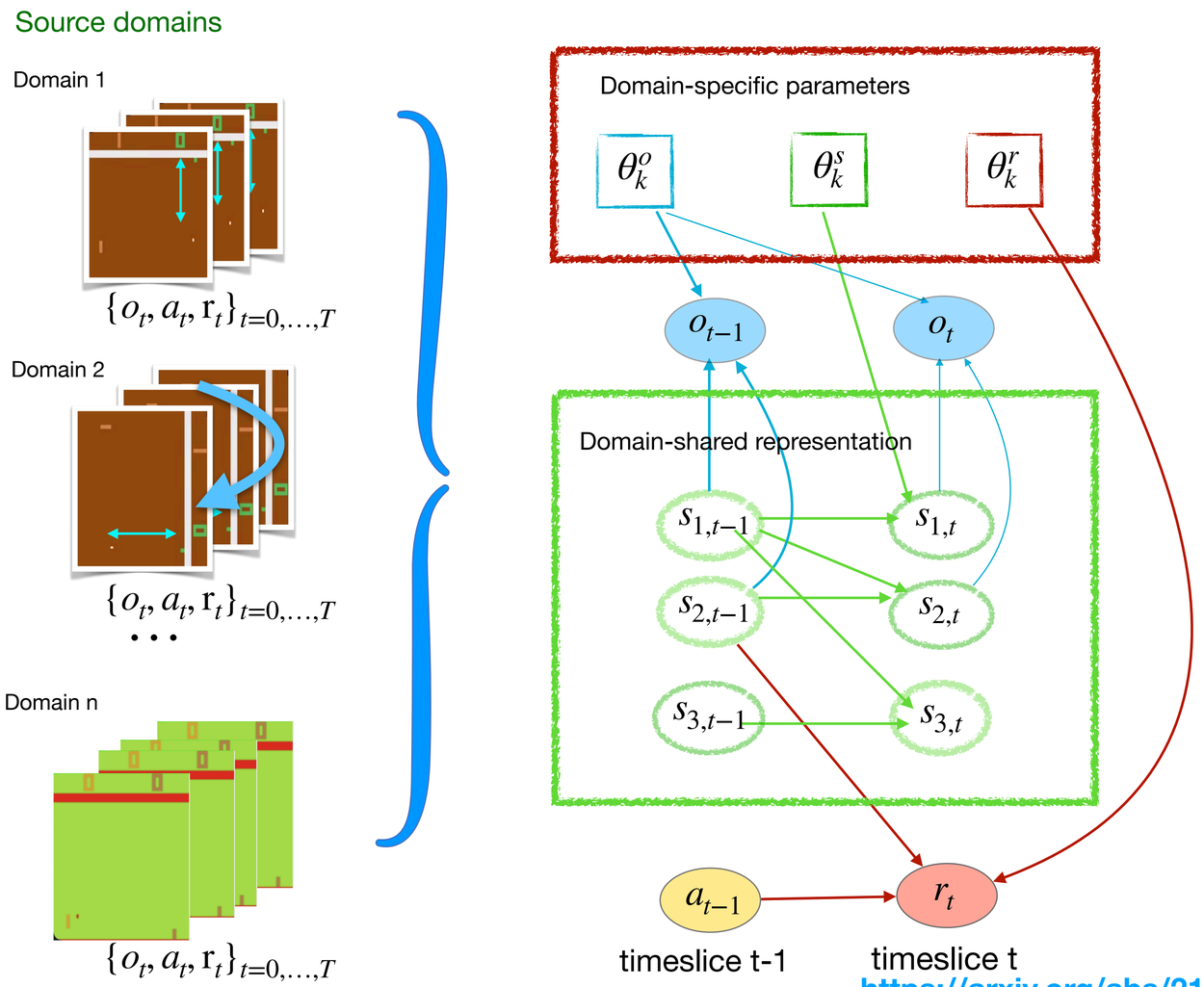


When we learn from symbolic inputs, the causal graph can be identified, but we don't have guarantees on what the latent change factors are

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

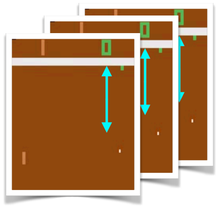
https://arxiv.org/abs/2107.02729

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

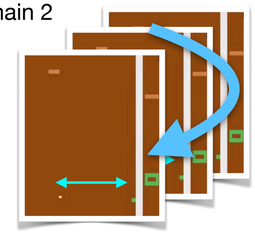
Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang ICLR 2022

Source domains

Domain 1



Domain 2



...

Estimate graph over estimated s_k, θ_k

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

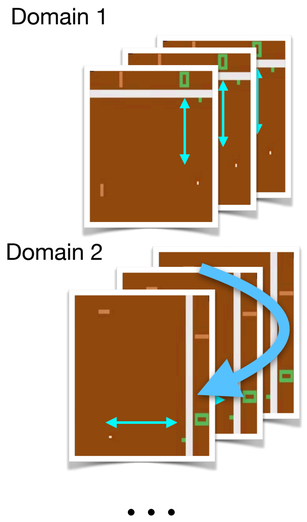
Learn optimal policy $\pi^*(s_k^{min}, \theta_k^{min})$ on source domains

https://arxiv.org/abs/2107.02729

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

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

Source domains

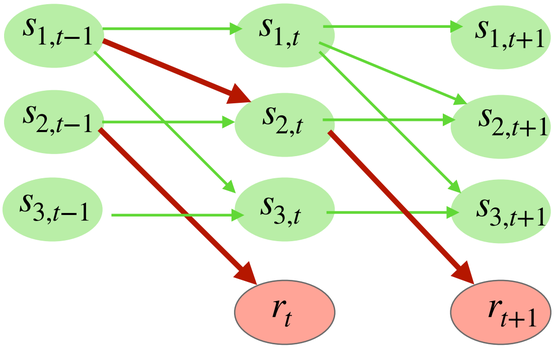


Estimate graph over estimated s_k, θ_k

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

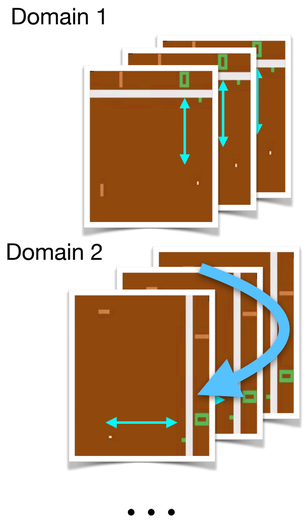


https://arxiv.org/abs/2107.02729

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

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

Source domains

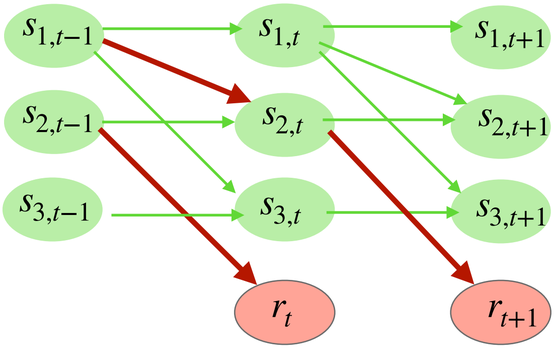


Estimate graph over estimated s_k, θ_k

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



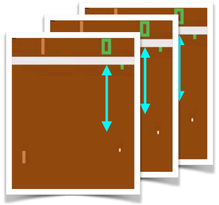
https://arxiv.org/abs/2107.02729

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

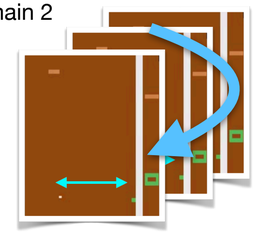
Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang

Source domains

Domain 1

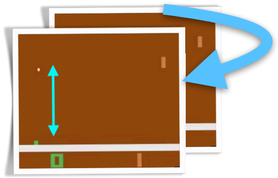


Domain 2



...

Target domain



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

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

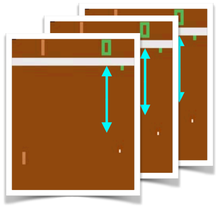
https://arxiv.org/abs/2107.02729

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

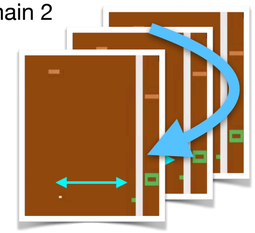
Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang

Source domains

Domain 1



Domain 2



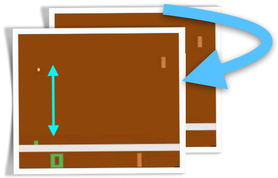
...

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



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

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

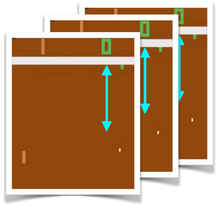
https://arxiv.org/abs/2107.02729

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

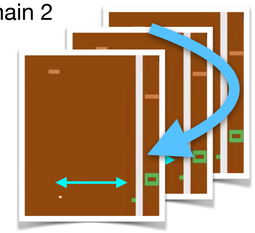
Biwei Huang, Fan Feng, Chaochao Lu, Sara Magliacane, Kun Zhang

Source domains

Domain 1



Domain 2



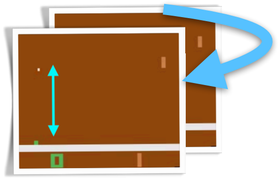
...

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



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

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

- **Results:** AdaRL consistently outperforms the baselines **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)

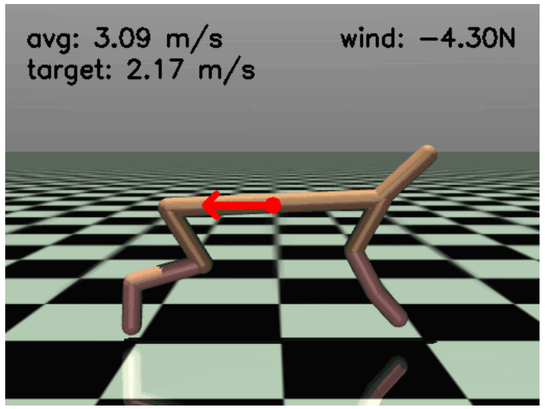
Average final scores on Pong (POMDP) with N_target=50

https://arxiv.org/abs/2107.02729

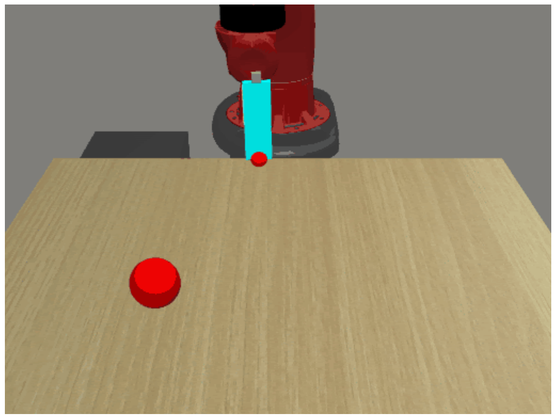
FansRL: Factored Adaptation for Non-Stationary Reinforcement Learning

Fan Feng, Biwei Huang, Kun Zhang, Sara Magliacane

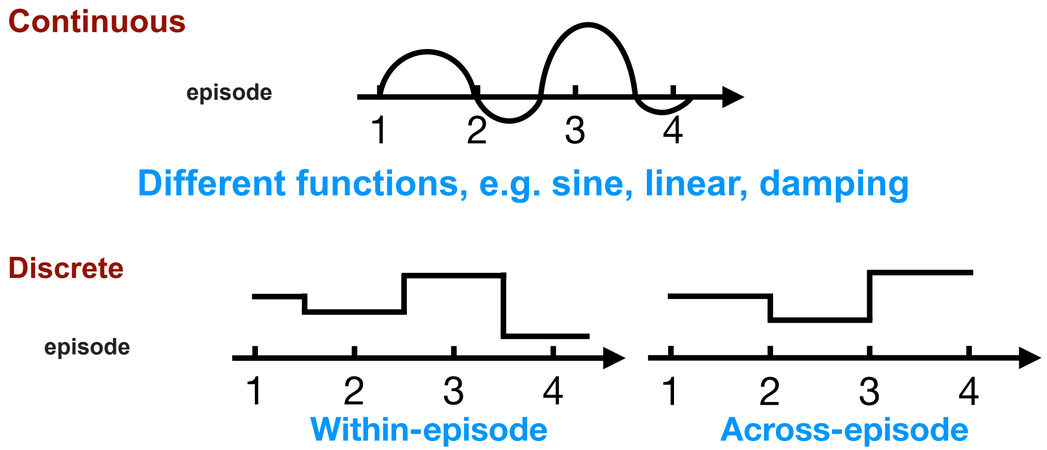
- **Task:** RL agent has to learn a policy that is robust to different types of non-stationarity, including **multiple simultaneous changes of different types**



Non-stationary environments
(wind changes)



Non-stationary rewards
(target changes)

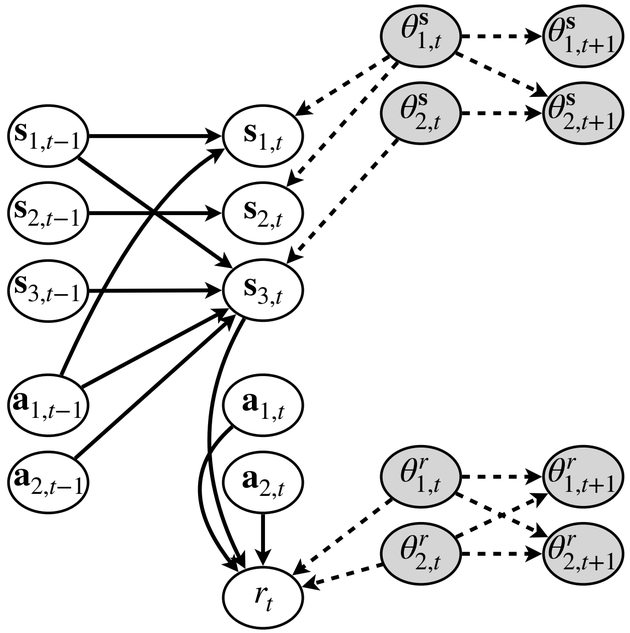


https://arxiv.org/abs/2203.16582

FansRL: Factored Adaptation for Non-Stationary Reinforcement Learning

Fan Feng, Biwei Huang, Kun Zhang, Sara Magliacane

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



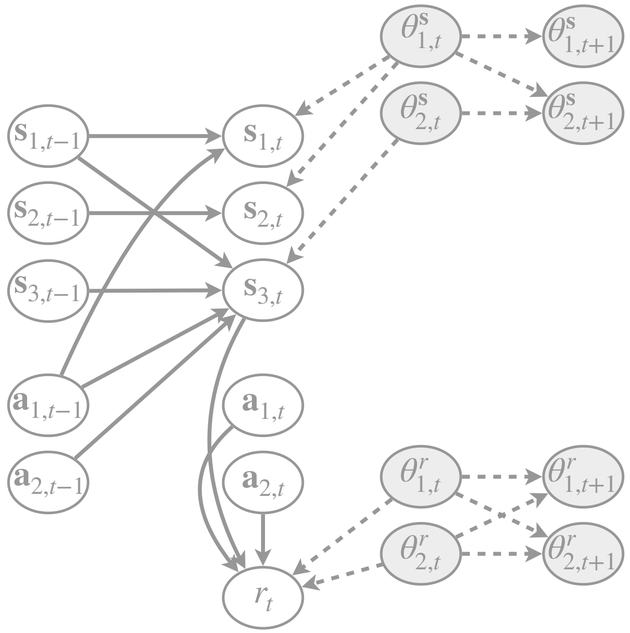
Factored Non-Stationary MDP

https://arxiv.org/abs/2203.16582

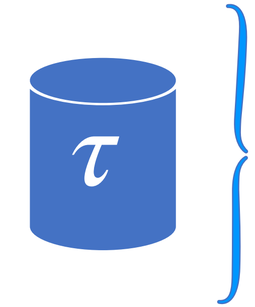
FansRL: Factored Adaptation for Non-Stationary Reinforcement Learning

Fan Feng, Biwei Huang, Kun Zhang, Sara Magliacane

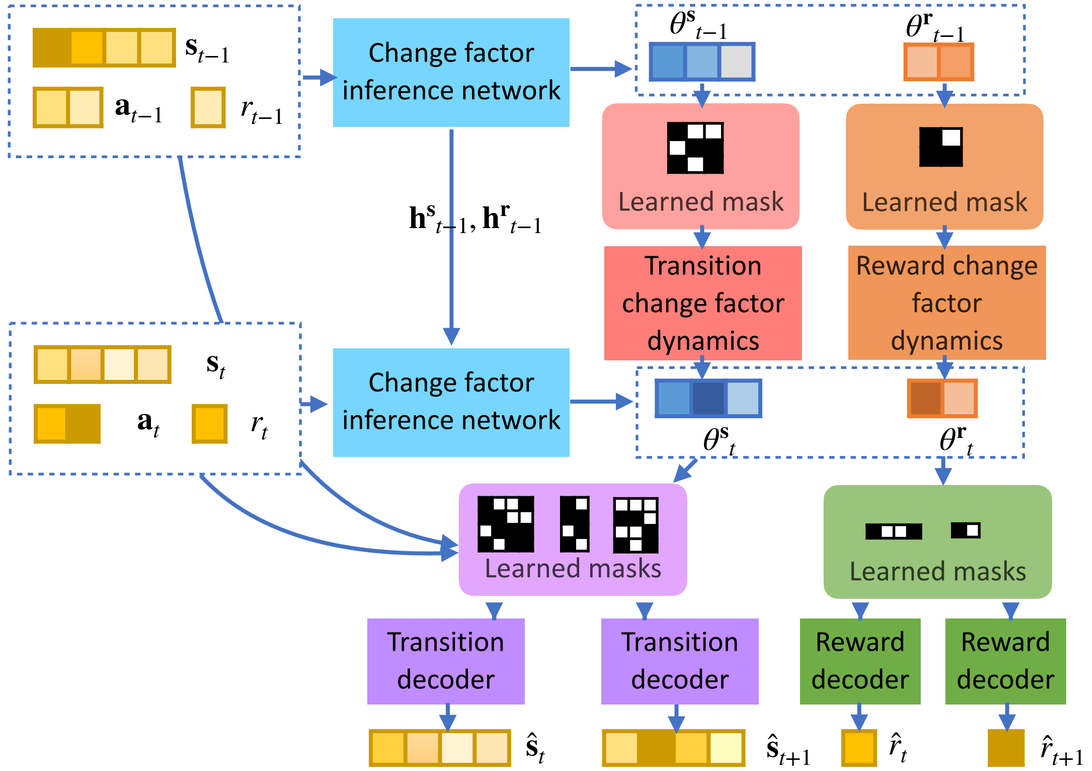
- The **latent change factors** are not constant anymore and they model **non-stationarity**



Factored Non-Stationary MDP



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



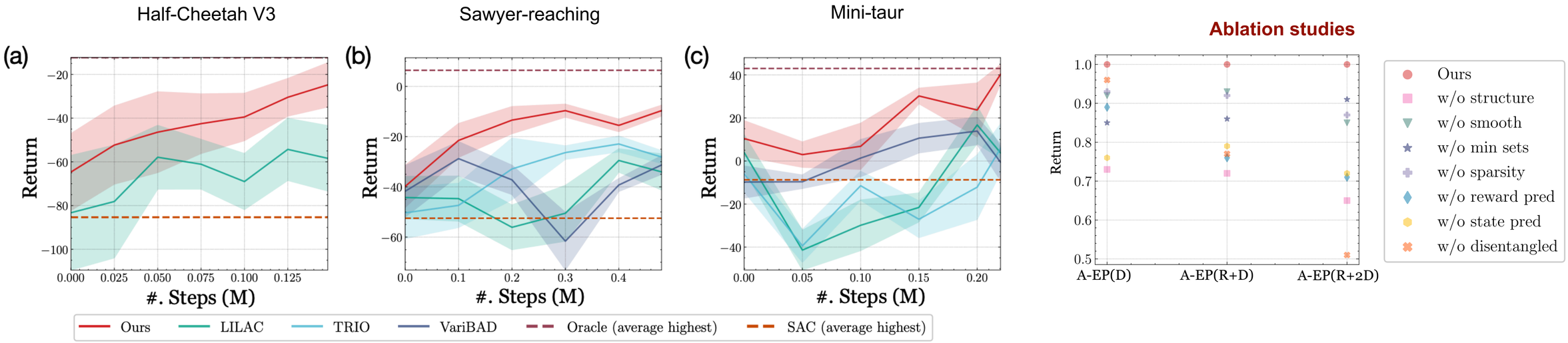
Factored Non-Stationary Variational Autoencoder

https://arxiv.org/abs/2203.16582

FansRL: Factored Adaptation for Non-Stationary Reinforcement Learning

Fan Feng, Biwei Huang, Kun Zhang, Sara Magliacane

- **Policy learning:** estimate latent change factors, learn policy as if they were observed
- **Results:** FansRL consistently outperforms the baselines **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)

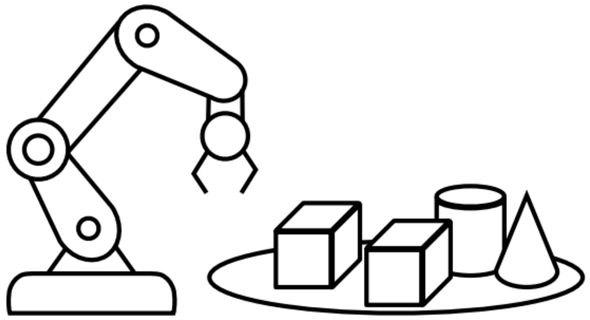
The biggest difference in performance is switching off learning the graph

https://arxiv.org/abs/2203.16582

Learning Dynamic Attribute-factored World Models for Efficient Multi-object Reinforcement Learning

Fan Feng, Sara Magliacane

- An environment is composed of **different objects**
  - Each object is associated with a **class** with specific **attributes**
  - Object can differ in latent parameters (e.g. size or mass)



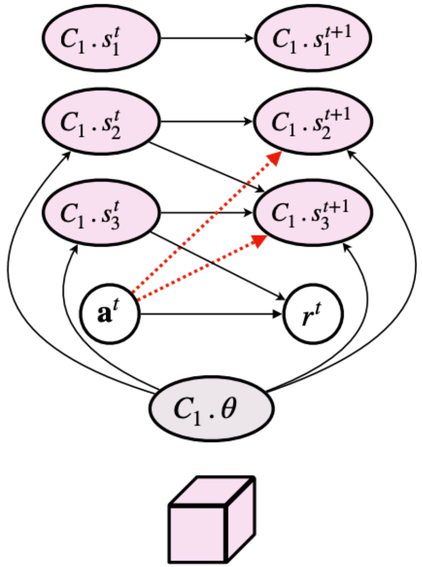
- The objects may have **interactions with each other**
  - Sparse and dynamic
  - Only a few attributes (related to physical process) interact.

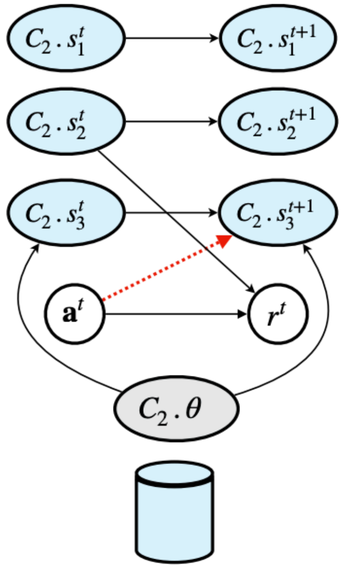
https://arxiv.org/abs/2307.09205

Learning Dynamic Attribute-factored World Models for Efficient Multi-object Reinforcement Learning

Fan Feng, Sara Magliacane

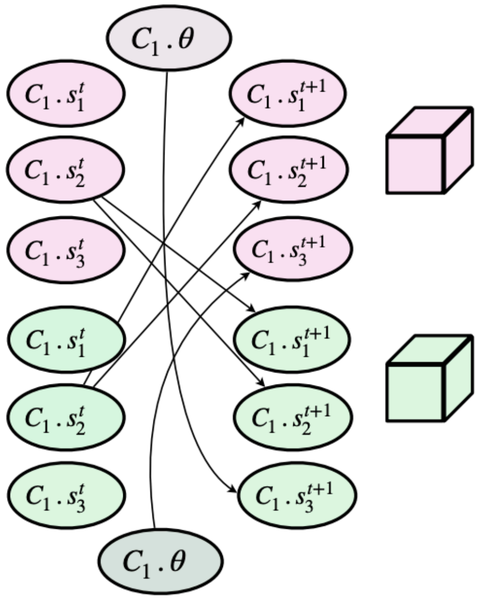
A. Class Template Graphs





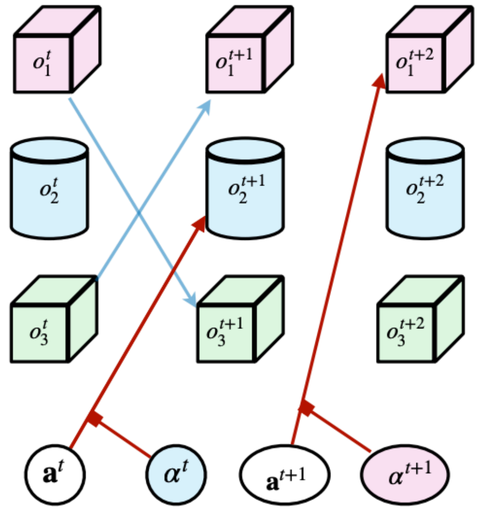
Static

B. Interaction Pattern Graph



Static

C. Interaction Graph



Dynamic

https://arxiv.org/abs/2307.09205

Learning Dynamic Attribute-factored World Models for Efficient Multi-object Reinforcement Learning

Fan Feng, Sara Magliacane

OpenAI-Fetch (trained on 1-Push & 1-Switch)

Experiment Settings	Method								
	DAFT-RL(Symbolic)	DeepSets	Self-attention	SRICS	GNN	STOVE	SMORL	NCS	LRN
2-Push + 2-Switch (S)	0.881 ±0.038	0.498 ±0.024	0.367 ±0.027	0.717 ±0.039	0.633 ±0.027	0.808 ±0.032	0.788 ±0.021	**0.912 ±0.022**	0.813 ±0.047
3-Push + 3-Switch (S+O)	**0.805 ±0.024**	0.324 ±0.012	0.183 ±0.031	0.685 ±0.019	0.594 ±0.023	0.672 ±0.021	0.694 ±0.017	0.751 ±0.024	0.715 ±0.025
2-Push (L)	**0.968 ±0.036**	0.940 ±0.029	0.941 ±0.046	0.922 ±0.042	0.953 ±0.037	0.931 ±0.047	0.964 ±0.023	0.944 ±0.039	0.923 ±0.038
2-Switch (L)	**0.923 ±0.065**	0.891 ±0.043	0.912 ±0.054	0.919 ±0.032	0.873 ±0.046	0.824 ±0.027	0.892 ±0.035	0.864 ±0.062	0.907 ±0.048
3-Push (L+O)	**0.921 ±0.037**	0.862 ±0.036	0.907 ±0.036	0.824 ±0.036	0.852 ±0.036	0.915 ±0.036	0.828 ±0.036	0.890 ±0.036	0.886 ±0.036
3-Switch (L+O)	**0.903 ±0.023**	0.451 ±0.075	0.392 ±0.045	0.590 ±0.087	0.744 ±0.057	0.772 ±0.079	0.835 ±0.029	0.818 ±0.074	0.857 ±0.032
2-Push + 2-Switch (L+S)	**0.793 ±0.026**	0.384 ±0.077	0.351 ±0.032	0.549 ±0.089	0.476 ±0.062	0.595 ±0.043	0.463 ±0.057	0.591 ±0.081	0.634 ±0.045
3-Push + 3-Switch (L+O+S)	**0.783 ±0.025**	0.256 ±0.045	0.115 ±0.029	0.531 ±0.082	0.362 ±0.035	0.531 ±0.042	0.525 ±0.051	0.529 ±0.027	0.418 ±0.092

Block-stacking (unknown mass and number of objects)

Experiment Settings	Method							
	DAFT-RL (SA)	DAFT-RL (AIR)	SMORL	SRICS	GNN	STOVE	NCS	LRN
2 Blocks	0.809 ±0.019	**0.838 ±0.030**	0.658 ±0.028	0.704 ±0.016	0.549 ±0.016	0.728 ±0.044	0.797 ±0.035	0.649 ±0.026
4 Blocks	**0.735 ±0.032**	0.698 ±0.022	0.605 ±0.020	0.591 ±0.049	0.526 ±0.041	0.498 ±0.013	0.571 ±0.026	0.461 ±0.028
6 blocks	0.591 ±0.025	**0.664 ±0.017**	0.536 ±0.040	0.509 ±0.043	0.461 ±0.088	0.475 ±0.023	0.521 ±0.049	0.602 ±0.097
8 blocks	0.506 ±0.083	**0.571 ±0.039**	0.386 ±0.062	0.420 ±0.061	0.334 ±0.047	0.278 ±0.086	0.397 ±0.052	0.463 ±0.077

https://arxiv.org/abs/2307.09205

Summary: Factored world models with changes

- **Pros:**
  - These methods can leverage the similarity of the structure across environments, non-stationary regimes or objects to adapt more efficiently to new settings
  - They can be applied to realistic RL benchmarks and seem to outperform the baselines
- **Cons:**
  - When learning from images the variables are not guaranteed to be disentangled or meaningful -> the learned graphs can be denser than necessary

Causality + machine learning (non-exhaustive list)

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

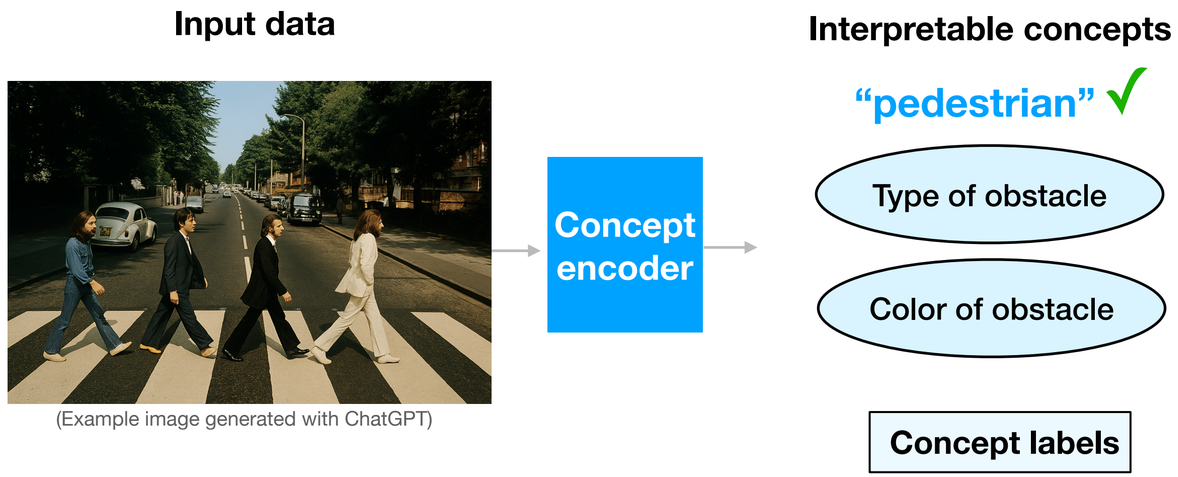
2. **Causality (in the most general definition) helps machine learning**
   - Robustness, Transfer learning
   - Reinforcement Learning
   - Bias mitigation, fairness, interpretability

Interpretability and “correct” concepts

- **Concept Bottleneck Models (CBM)** [Koh et al. 2020 and following] provide **interpretability by design** by training the model to explicitly use **human-provided concepts**

Interpretability and “correct” concepts

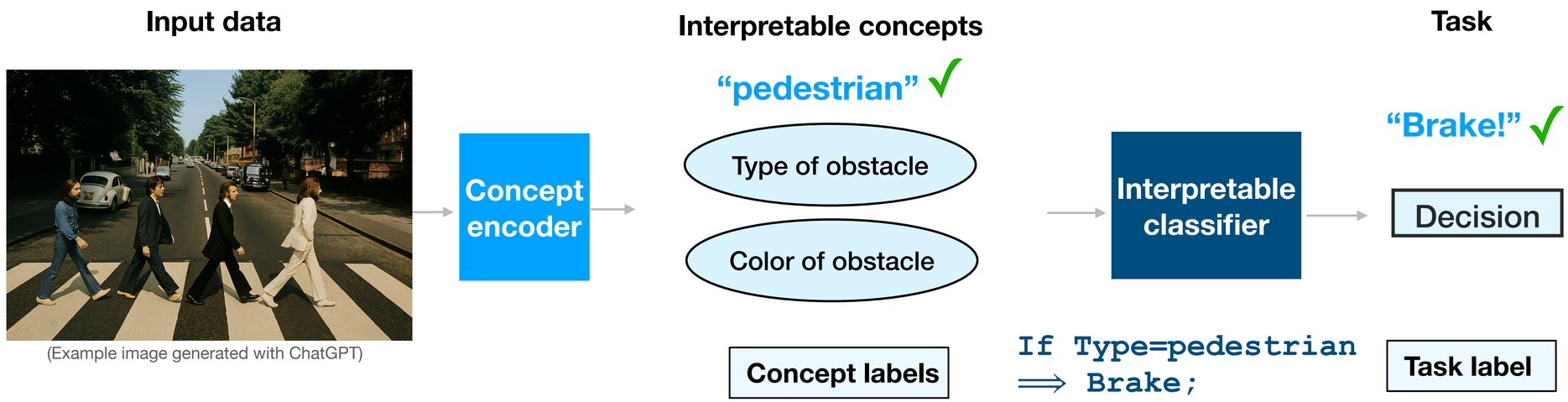
- **Concept Bottleneck Models (CBM)** [Koh et al. 2020 and following] provide **interpretability by design** by training the model to explicitly use **human-provided concepts**



(Example image generated with ChatGPT)

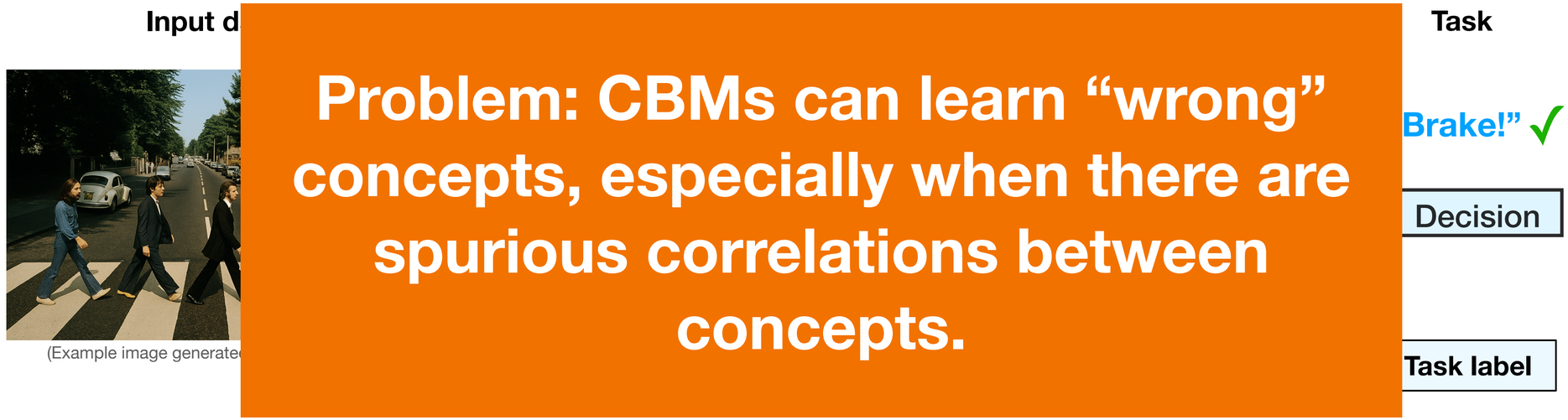
Interpretability and “correct” concepts

- **Concept Bottleneck Models (CBM)** [Koh et al. 2020 and following] provide **interpretability by design** by training the model to explicitly use **human-provided concepts**



Interpretability and “correct” concepts

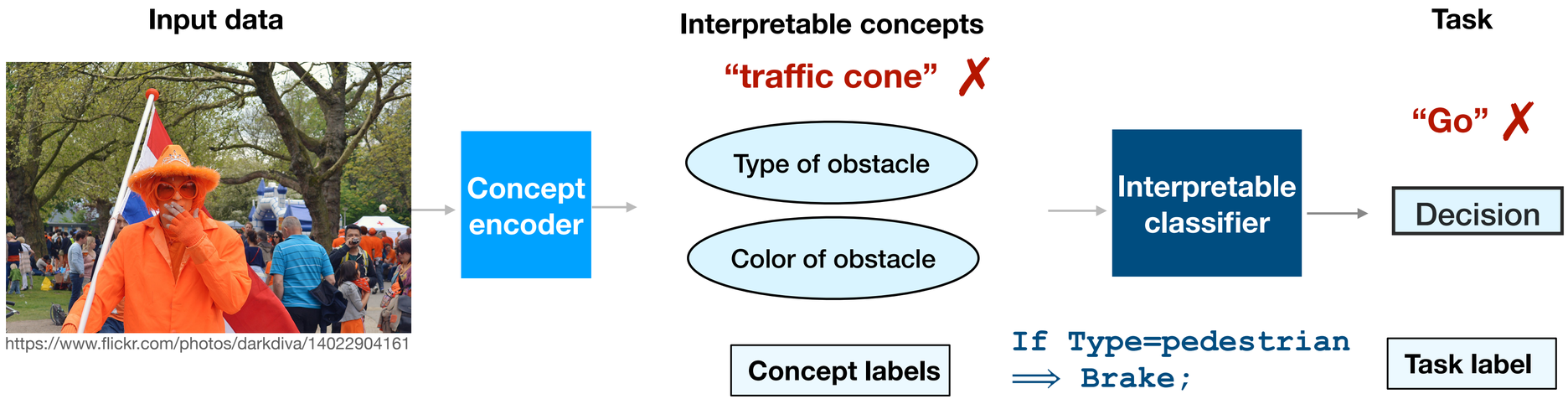
- **Concept Bottleneck Models (CBM)** [Koh et al. 2020 and following] provide **interpretability by design** by training the model to explicitly use **human-provided concepts**



Problem: CBMs can learn “wrong” concepts, especially when there are spurious correlations between concepts.

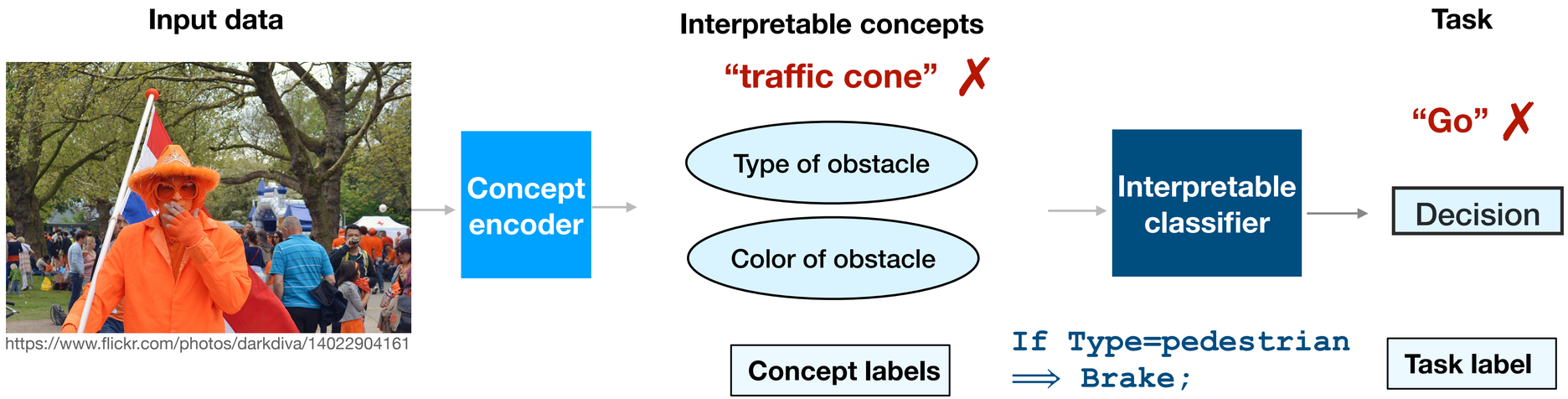
Interpretability and “correct” concepts

- **Problem:** Current CBM methods can potentially learn the **“wrong” concepts**, especially when there are **spurious correlations** between the concepts in the training dataset.



Interpretability and “correct” concepts

- **Problem:** Current CBM methods can potentially learn the **“wrong” concepts**, especially when there are **spurious correlations** between the concepts in the training dataset.

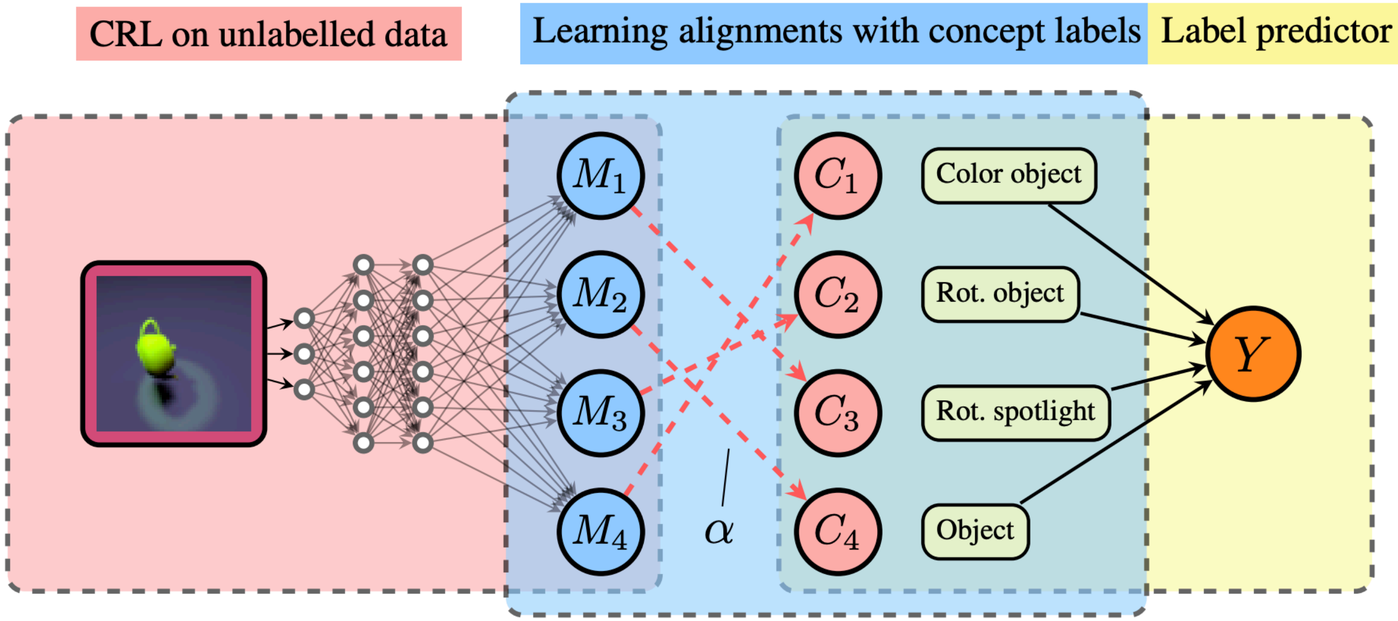


- **Problem 2:** Current CBMs require many **concept labels**, which might be **expensive**

Sample-efficient Learning of Concepts with Theoretical Guarantees: from Data to Concepts without Interventions

Hidde Fokkema, Tim van Erven, Sara Magliacane

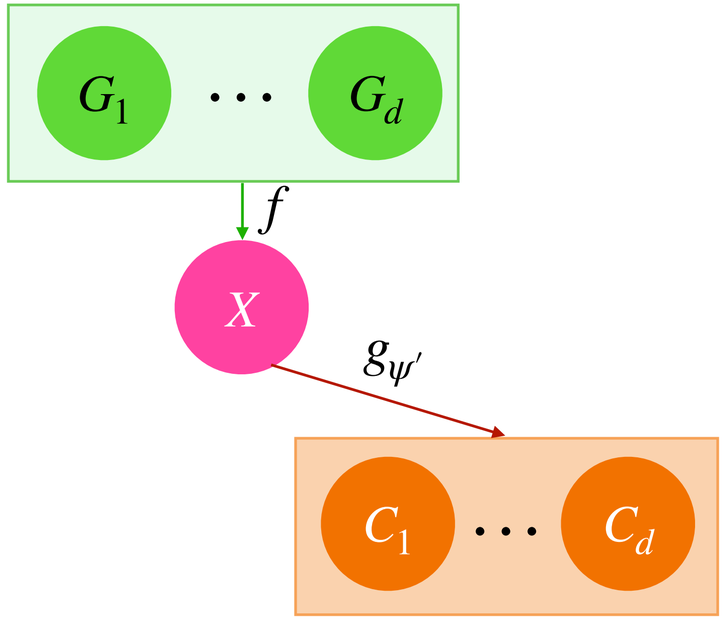
- A framework that leverages **Causal Representation Learning (CRL)** to learn concepts with theoretical guarantees with few concept labels



Framework and assumptions

- We assume a causal system with **latent causal variables** $G = (G_1, \ldots, G_d)$ for which we only observe entangled **high-dimensional observations** $X = f(G)$, e.g. images

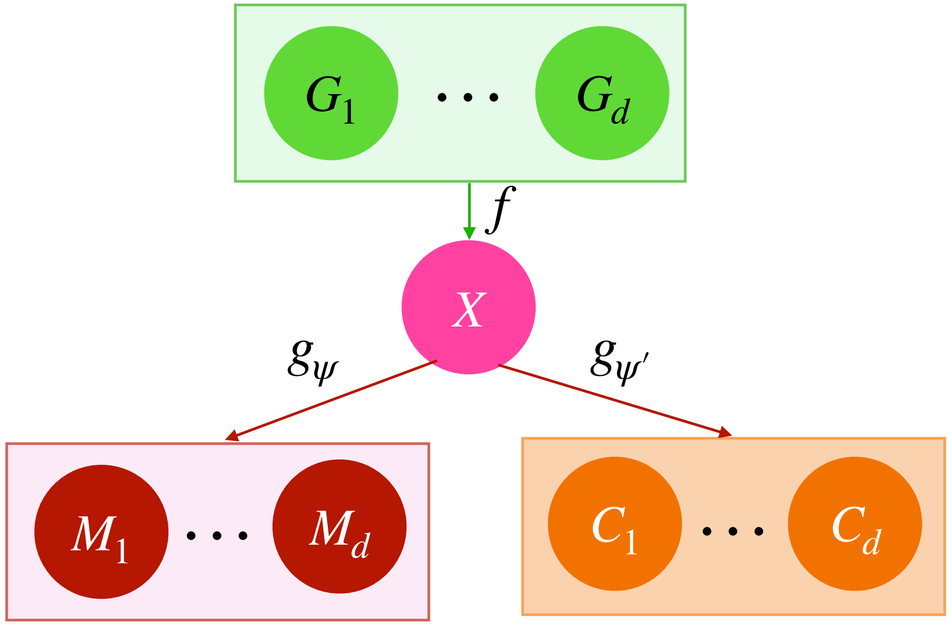
- We assume **concepts** $C = (C_1, \ldots, C_d)$ correspond to the causal variables up to **element-wise transformations**



Framework and assumptions

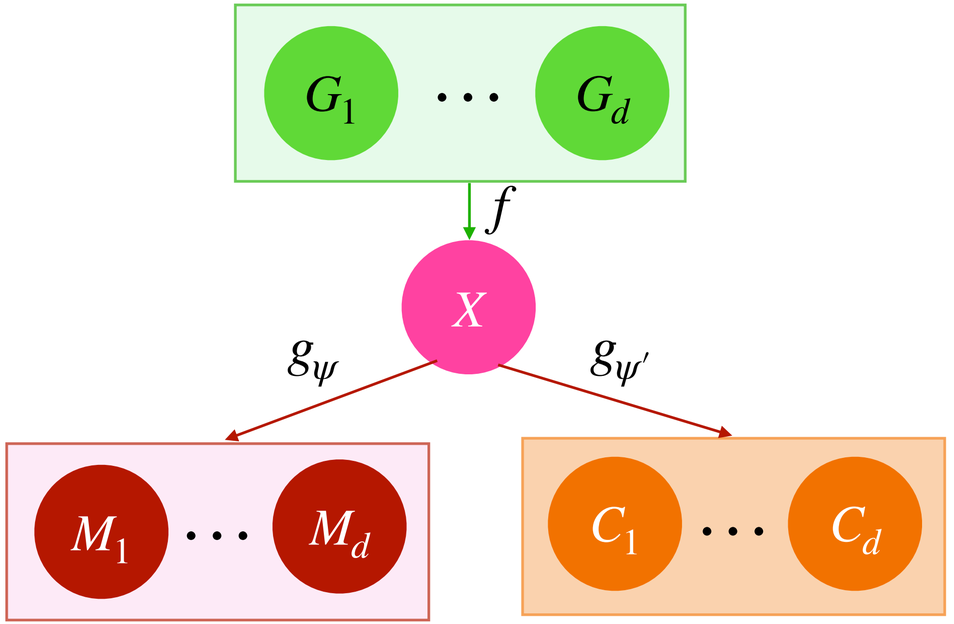
- We assume a causal system with **latent causal variables** $G = (G_1, \ldots, G_d)$ for which we only observe entangled **high-dimensional observations** $X = f(G)$, e.g. images
- We assume **concepts** $C = (C_1, \ldots, C_d)$ correspond to the causal variables up to **element-wise transformations**
- We assume a CRL method learns **embeddings** of the causal variables $M = (M_1, \ldots, M_d)$ **up to permutation and element-wise transformations.** Then:

$$[C_1, \ldots, C_d] = [T_{\pi(1)}(M_{\pi(1)}), \ldots, T_{\pi(d)}(M_{\pi(d)})]$$



Framework and assumptions

- We assume a causal system with **latent causal variables** $G = (G_1, \ldots, G_d)$ for which we only observe entangled **high-dimensional observations** $X = f(G)$, e.g. images
- We assume **concepts** $C = (C_1, \ldots, C_d)$ correspond to the causal variables up to **element-wise transformations**
- We assume a CRL method learns **embeddings** of the causal variables $M = (M_1, \ldots, M_d)$ **up to permutation and element-wise transformations.** Then:

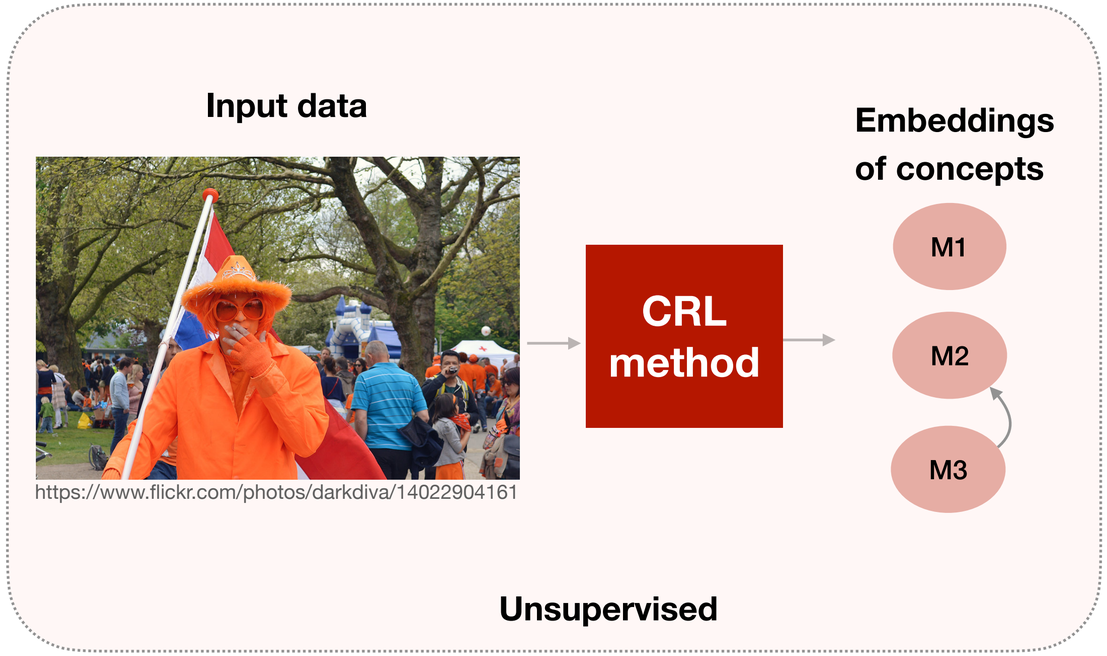


$$[C_1, \ldots, C_d] = [T_{\pi(1)}(M_{\pi(1)}), \ldots, T_{\pi(d)}(M_{\pi(d)})]$$

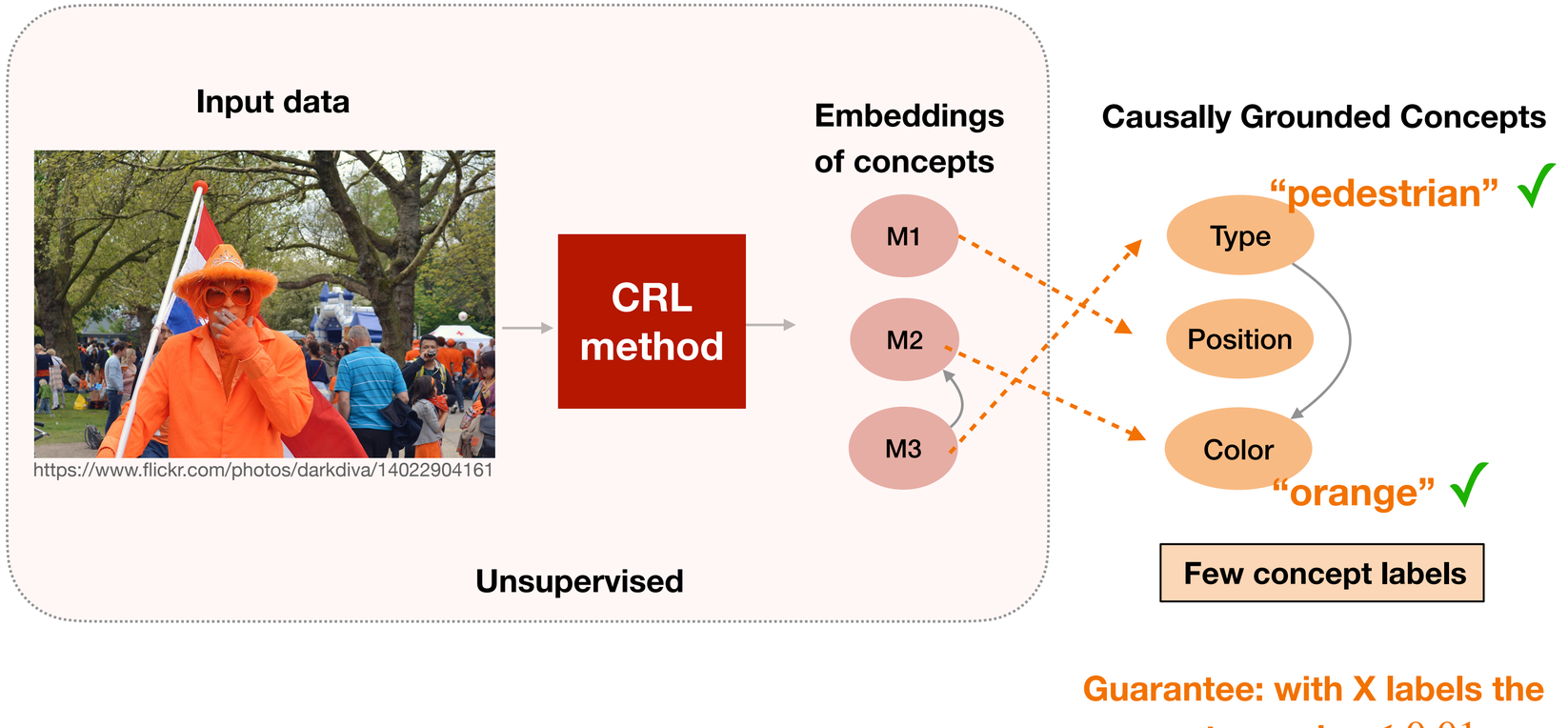
Linear: GroupLasso $C_i = \varphi(M)\beta_i^* + \epsilon_i$

Our task is to learn the alignment function α (permutation + element-wise transformations) with error bounds based on how many labels

Interpretability and “correct” concepts v2

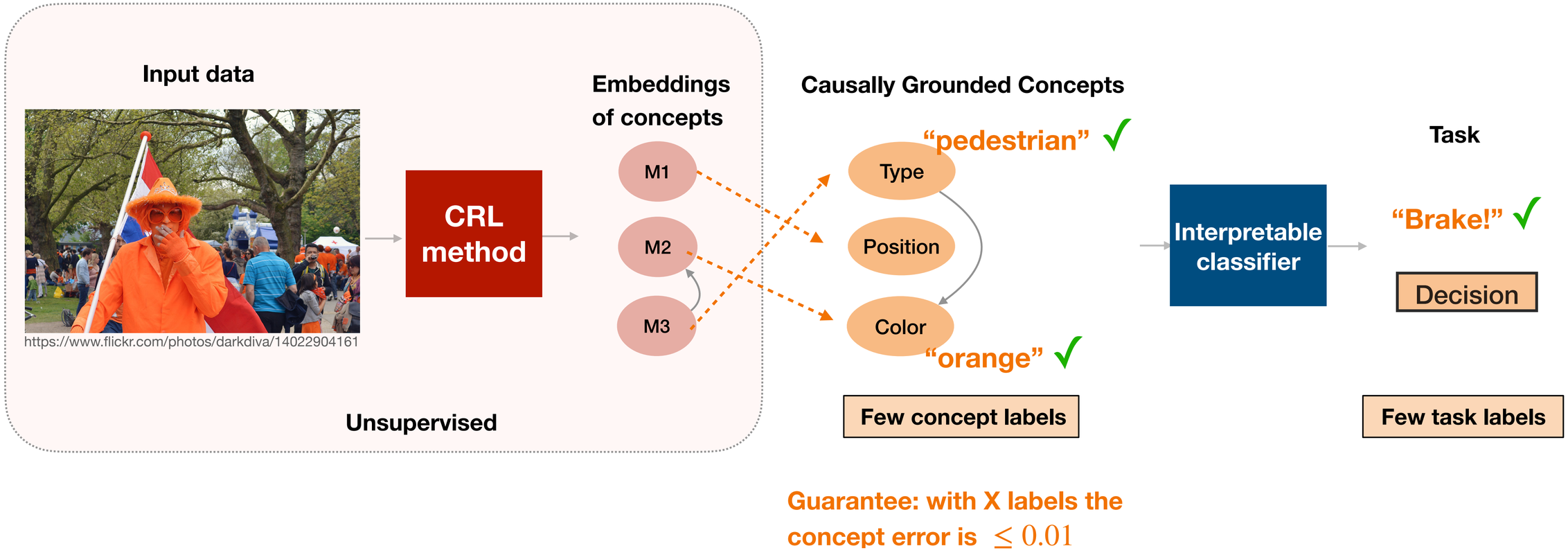


Interpretability and “correct” concepts v2



Guarantee: with X labels the concept error is ≤ 0.01

Interpretability and “correct” concepts v2



Guarantee: with X labels the concept error is ≤ 0.01

Comparison to standard Concept-Bottleneck Models

Added a randomly generated binary label from concepts to standard CRL datasets

Model	Method	Label Acc. ↑ (n) 20	100	1000	10000	OIS ↓ (n) 20	100	1000	10000
Action Sparsity Dataset									
DMS-VAE	Linear	**0.77 ± 0.03**	0.81 ± 0.01	0.83 ± 0.01	0.84 ± 0.01	**0.63 ± 0.01**	0.40 ± 0.00	0.15 ± 0.00	0.12 ± 0.00
	Spline	0.72 ± 0.03	**0.83 ± 0.02**	**0.85 ± 0.01**	0.86 ± 0.01	0.63 ± 0.01	0.40 ± 0.00	0.15 ± 0.00	0.12 ± 0.00
	RFF	0.77 ± 0.02	0.82 ± 0.02	0.84 ± 0.01	0.85 ± 0.01	0.64 ± 0.01	**0.39 ± 0.00**	0.14 ± 0.00	0.12 ± 0.00
iVAE	Linear	0.73 ± 0.03	0.79 ± 0.02	0.81 ± 0.01	0.83 ± 0.01	0.63 ± 0.01	0.40 ± 0.00	0.29 ± 0.00	0.26 ± 0.00
	Spline	0.69 ± 0.02	0.76 ± 0.02	0.81 ± 0.01	0.83 ± 0.01	0.63 ± 0.01	0.40 ± 0.00	0.28 ± 0.00	0.26 ± 0.00
	RFF	0.59 ± 0.04	0.70 ± 0.02	0.78 ± 0.01	0.81 ± 0.01	0.64 ± 0.01	0.39 ± 0.00	0.28 ± 0.00	0.26 ± 0.00
CBM (Koh et al., 2020)		0.60 ± 0.03	0.60 ± 0.02	0.73 ± 0.01	0.88 ± 0.01	0.92 ± 0.01	0.43 ± 0.01	**0.11 ± 0.00**	0.07 ± 0.00
CEM (Zarlenga et al., 2022)		0.66 ± 0.02	0.69 ± 0.03	0.82 ± 0.01	0.88 ± 0.01	0.90 ± 0.03	0.47 ± 0.01	0.40 ± 0.01	0.57 ± 0.01
HardCBM (Havasi et al., 2022)		0.56 ± 0.03	0.61 ± 0.01	0.68 ± 0.01	**0.89 ± 0.00**	0.92 ± 0.02	0.46 ± 0.01	0.12 ± 0.00	**0.06 ± 0.00**
Temporal Causal3DIdent Dataset									
CITRISVAE	Linear	0.74 ± 0.06	0.75 ± 0.06	0.80 ± 0.04	0.81 ± 0.04	0.69 ± 0.02	0.44 ± 0.01	0.19 ± 0.00	0.16 ± 0.00
	Spline	0.74 ± 0.06	**0.80 ± 0.04**	**0.82 ± 0.03**	**0.83 ± 0.03**	0.69 ± 0.02	0.44 ± 0.01	**0.13 ± 0.00**	**0.09 ± 0.00**
	RFF	0.70 ± 0.06	0.76 ± 0.05	0.79 ± 0.04	0.81 ± 0.04	**0.65 ± 0.01**	0.43 ± 0.00	0.16 ± 0.01	0.13 ± 0.01
iVAE	Linear	**0.74 ± 0.06**	0.76 ± 0.05	0.78 ± 0.05	0.80 ± 0.04	0.69 ± 0.02	0.44 ± 0.01	0.17 ± 0.00	0.14 ± 0.00
	Spline	0.73 ± 0.06	0.78 ± 0.04	0.79 ± 0.04	0.81 ± 0.04	0.69 ± 0.02	0.43 ± 0.04	0.17 ± 0.00	0.13 ± 0.00
	RFF	0.69 ± 0.06	0.75 ± 0.05	0.78 ± 0.04	0.80 ± 0.04	0.65 ± 0.01	**0.42 ± 0.02**	0.16 ± 0.00	0.13 ± 0.00
CBM (Koh et al., 2020)		0.57 ± 0.02	0.53 ± 0.03	0.68 ± 0.04	0.78 ± 0.05	1.00 ± 0.03	0.48 ± 0.02	0.25 ± 0.00	0.18 ± 0.00
CEM (Zarlenga et al., 2022)		0.64 ± 0.04	0.67 ± 0.03	0.72 ± 0.05	0.78 ± 0.05	0.97 ± 0.05	0.51 ± 0.02	0.36 ± 0.01	0.49 ± 0.01
HardCBM (Havasi et al., 2022)		0.57 ± 0.05	0.53 ± 0.02	0.63 ± 0.03	0.77 ± 0.05	0.96 ± 0.03	0.47 ± 0.02	0.24 ± 0.00	0.16 ± 0.00

Summary: Concepts with guarantees

- **Pros:**
  - Principled framework for learning concepts from raw data based on CRL
  - Error bounds on the concepts based on the number of concept labels
- **Cons:**
  - Strong assumptions: it assumes concepts are
    - Weakly correlated -> this impacts the theoretical bounds
    - Continuous -> requires discrete CRL + LogisticLasso approach
    - 1:1 with causal variables -> concepts could be coarser -> causal abstraction

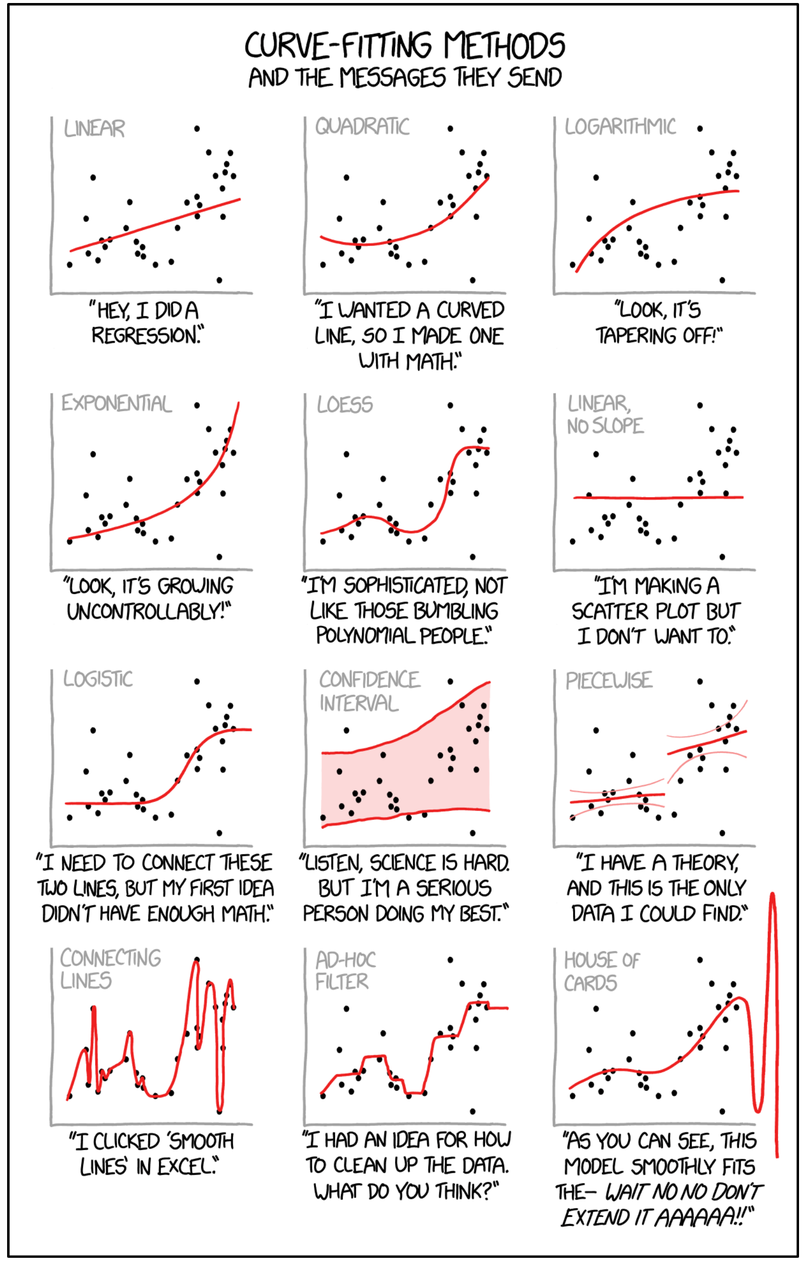
Causality + machine learning (non-exhaustive list)

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

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

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

Questions?



https://xkcd.com/2048/

Until May 21st: organize in groups + choose a paper from list

- Organize in groups of 4 by next week (May 14th)
- Spend a bit of time checking the different topics and choosing one specific paper/project
- Each member of the group should read carefully the paper and any related material (e.g., the suggested reading material for the topic)
- We will discuss your questions regarding the paper…
- … but also ideas for extensions