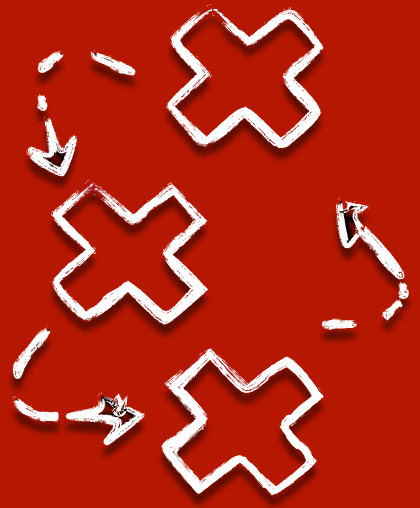


# A high-level view on causal representation learning with actions

Sara Magliacane (Saarland University & University of Amsterdam)

# Disclaimer

- **I'm not a verification person…**
  - So sorry for my limited understanding and/or naiveté

# Disclaimer

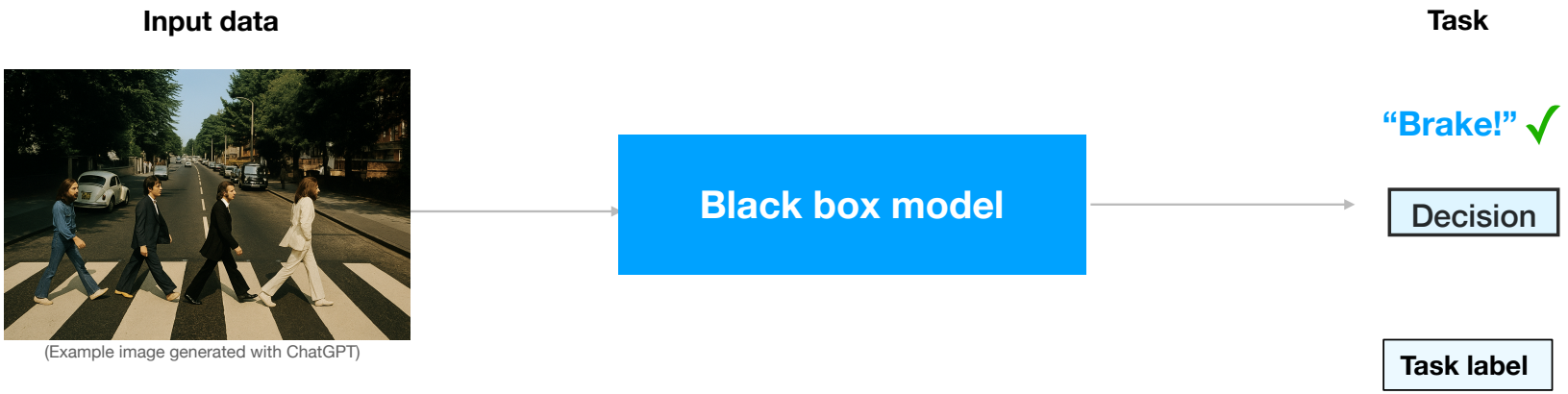
- **I'm not a verification person…**
  - So sorry for my limited understanding and/or naiveté

- **… but I think some ideas from causality and causal representation learning might be useful for the verification community**

  - In particular, all the work on theoretical guarantees for identifying underlying causal variables from unsupervised data, e.g., images, text etc

# Motivating example: how can we verify a self-driving car?

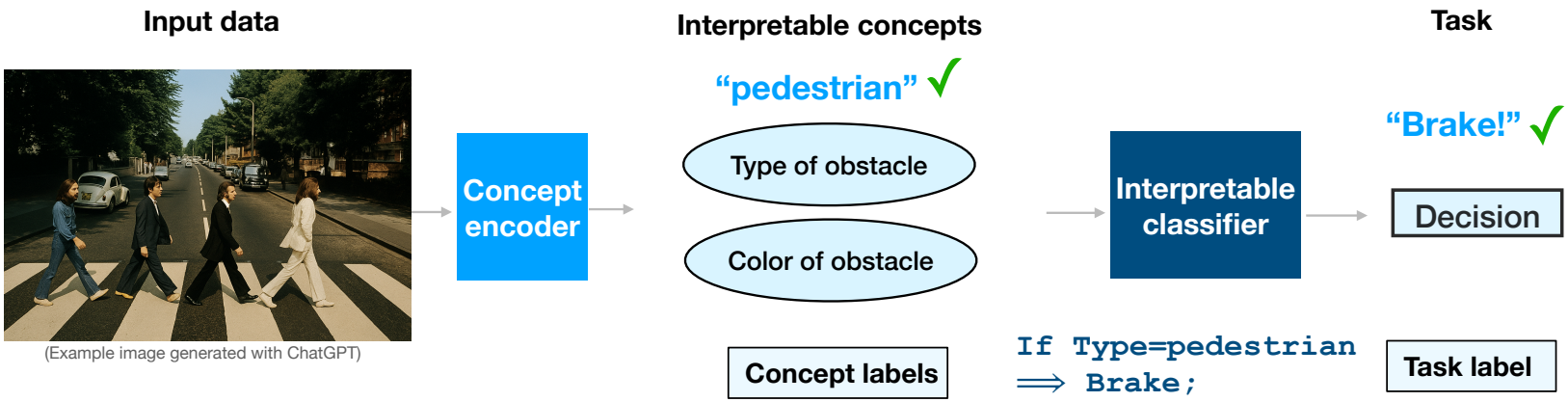
- **Problem:** How can we verify formally that a self-driving car that uses images as inputs is safe?




- Naive idea 1: verify the neural network (potentially unscalable and not necessarily **interpretable**)
- Naive idea 2: verify it works for all inputs (**unfeasible** to generate all possible realistic images)

# Alternative idea: verifying interpretable classifiers based on 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**



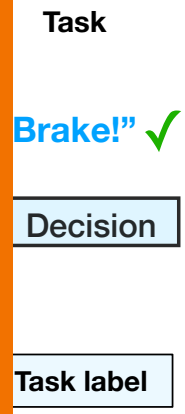

- Verifying a model like this would allow us to focus on high-level reasoning issues
  - I assume somebody might have already argued this :)

# Alternative idea: verifying interpretable classifiers based on 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**

(Examp

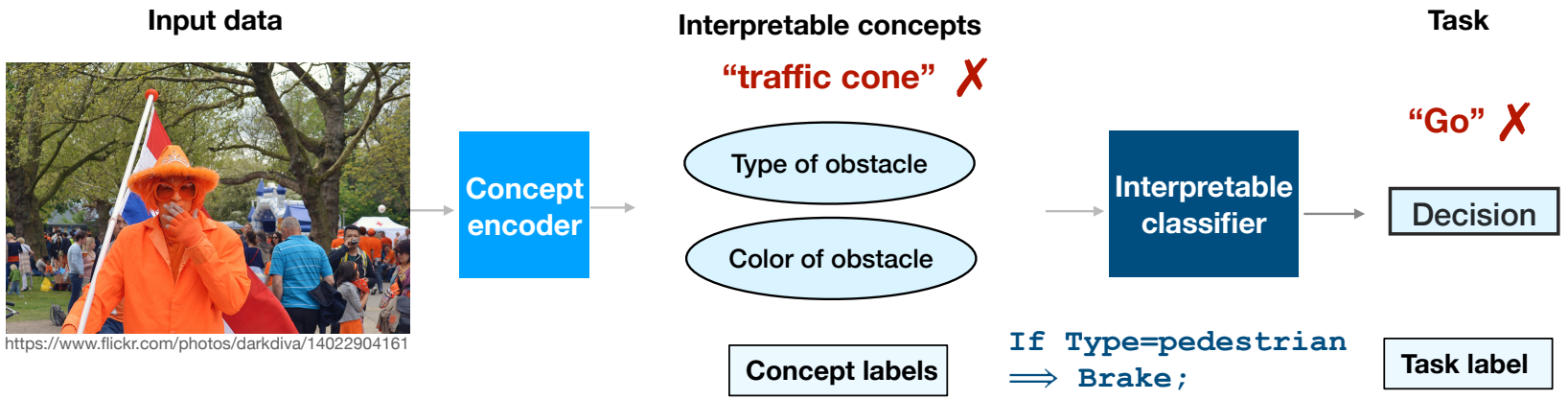
**Problem: CBMs can learn "wrong" concepts, especially when there are spurious correlations between concepts.**




- Verifying a model like this would allow us to focus on high-level reasoning issues
  - I assume somebody might have already argued this :)

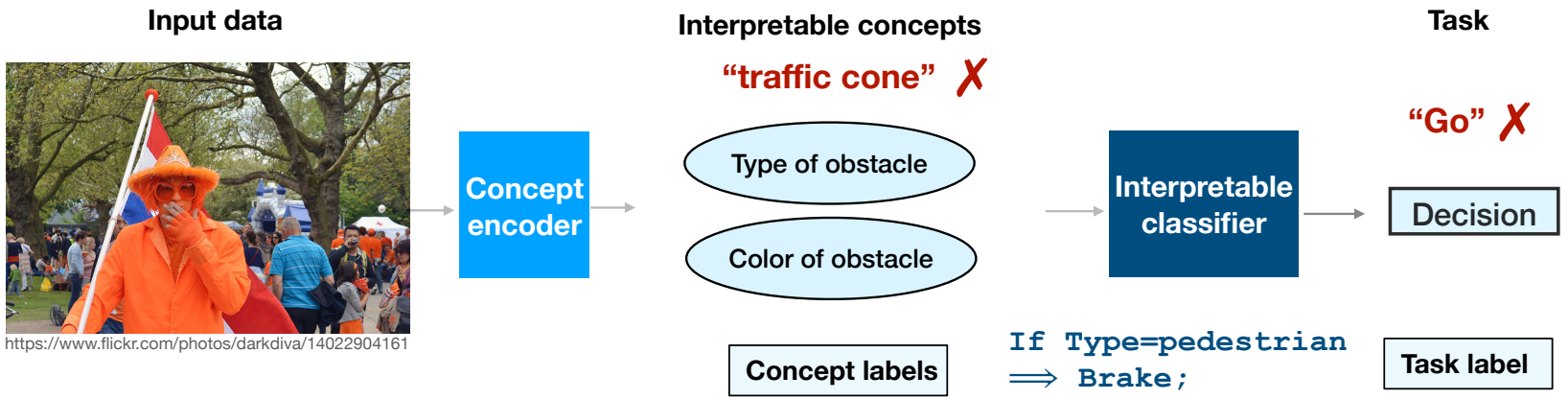
# How do we guarantee that we learn correct concepts?

- **Problem:** Current CBMs can potentially learn **"wrong" concepts**, especially when there are **spurious correlations** between the concepts in training -> verification is meaningless



# How do we guarantee that we learn correct concepts?

- **Problem:** Current CBMs can potentially learn **"wrong" concepts**, especially when there are **spurious correlations** between the concepts in training -> verification is meaningless




- **A possible solution:** Causal representation learning learns high-level variables from unstructured data with **theoretical guarantees on their correctness**, **even if correlated**

# Causal Representation Learning (CRL)

Can we **predict the effect of interventions** if the causal variables are **not directly observed** and we **do not have labels for them**, but we have **high-dimensional observations of the system**?

**Task 1: identify/disentangle the causal variables from observations**

**Task 2: learn causal relations between them from data (causal discovery)**

# Causal Representation Learning overview

**Task 1: learn the causal variables**

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

# Causal Representation Learning overview

**Task 1: learn the causal variables**

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

**Strategy 1: restrict SCM (e.g. linear, Gauss/exp) or mixing function (e.g. piecewise/linear)**

**Strategy 2: multi-view or paired data**

**Strategy 3: use interventions/actions**

**Strategy 4: assume temporal data**

# Causal Representation Learning overview

**Task 1: learn the causal variables**

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

**Strategy 1: restrict SCM (e.g. linear, Gauss/exp) or mixing function (e.g. piecewise/linear)**

**Strategy 2: multi-view or paired data**

**Strategy 3: use interventions/actions**

**Strategy 4: assume temporal data**

Without supervision, we cannot identify the “true” causal variables, but we can only **identify** them **up to element-wise transformations (and permutations)**

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_d \end{bmatrix} \quad \begin{bmatrix} \hat{z}_1 = h_1(z_{\pi(1)}) \\ \hat{z}_2 = h_2(z_{\pi(2)}) \\ \vdots \\ \hat{z}_d = h_d(z_{\pi(d)}) \end{bmatrix}$$

# Causal Representation Learning overview

**Task 1: learn the causal variables**

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

**Strategy 1: restrict SCM (e.g. linear, Gauss/exp) or mixing function (e.g. piecewise/linear)**

**Strategy 2: multi-view or paired data**

**Strategy 3: use interventions/actions**

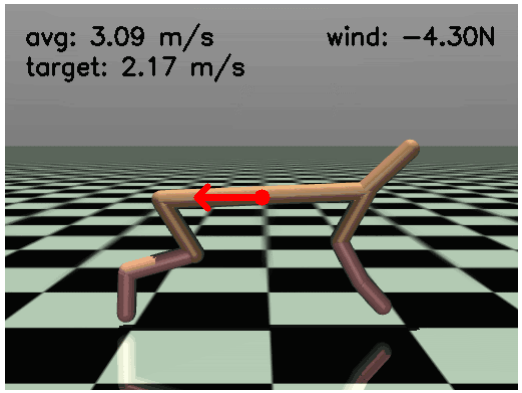
**Strategy 4: assume temporal data**

Without supervision, we cannot identify the “true” causal variables, but we can only **identify** them **up to element-wise transformations (and permutations)**

$$\begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_d \end{bmatrix} \rightarrow \begin{bmatrix} \hat{z}_1 = h_1(z_{\pi(1)}) \\ \hat{z}_2 = h_2(z_{\pi(2)}) \\ \vdots \\ \hat{z}_d = h_d(z_{\pi(d)}) \end{bmatrix}$$

# CRL in temporal settings with actions

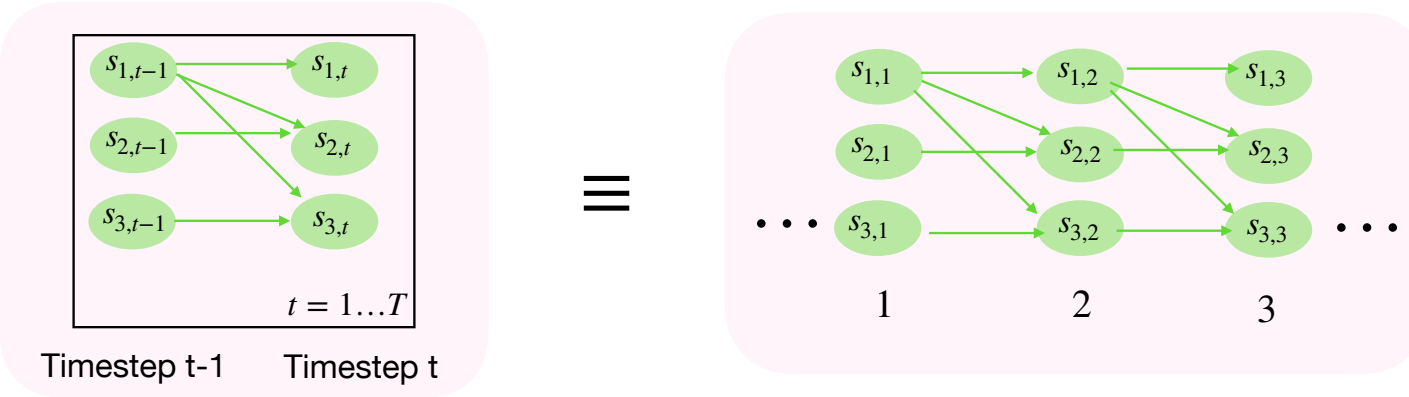
- Natural setting for learning from interventions/actions: “before” and “after”
  - E.g. sequential decision making, RL, planning, robotics, …




- Often we want to extract **semantic features from images** in an unsupervised way
  - **Causal representation learning (CRL) -** learn high level causal variables and causal relations between them from low level observations

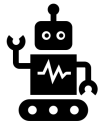
# Modelling causality in time series: Dynamic Bayesian Networks

- Extension of Bayesian networks to temporal settings, a type of template graph.
- Common assumptions for Dynamic Bayesian Networks:
  - **1-Markov assumption**: only vars from t-1 (1 timestep back) can cause vars at t
  - **Stationarity:** the transition model (edges) are the same across pairs of time steps
  - **No instantaneous effects:** there are no edges between vars at same timestep




**MDPs in RL are an example**

# Goal: identify causal variables (up to transformations) from sequences of images and actions



Agent

Action

Time step t

Time step t+1

Microwave state

Egg state

Potato state

Stove 1 state

Stove 2 state

…

…

# A few temporal CRL works - quick overview

- **IVAE [Khemakhem et al 2020]:** auxilliary variable $\mathbf{u}$, s.t. latent variables are conditionally independent given $\mathbf{u}$ (e.g. previous state), parametric assumptions (e.g. conditional exponential family prior), sufficient variability

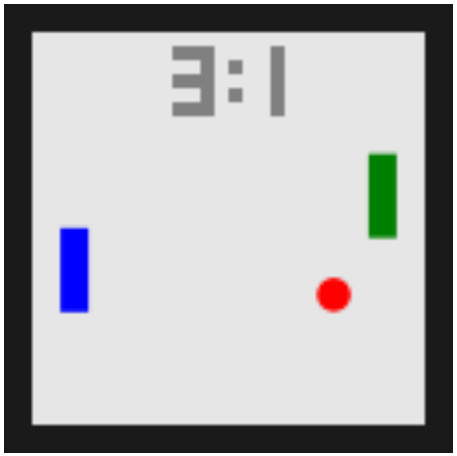
- **DMSVAE [Lachapelle et al. 2022]:** relaxed parametric assumptions, but still exponential family, sufficient time/action variability, graphical assumptions
  - **[Lachapelle et al. 2024]** is nonparametric, ident. of blocks of causal variables

- **LEAP [Yao et al., 2022]:** leverage nonstationarity, nonparametric, sufficient variability across regimes or environments

# CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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

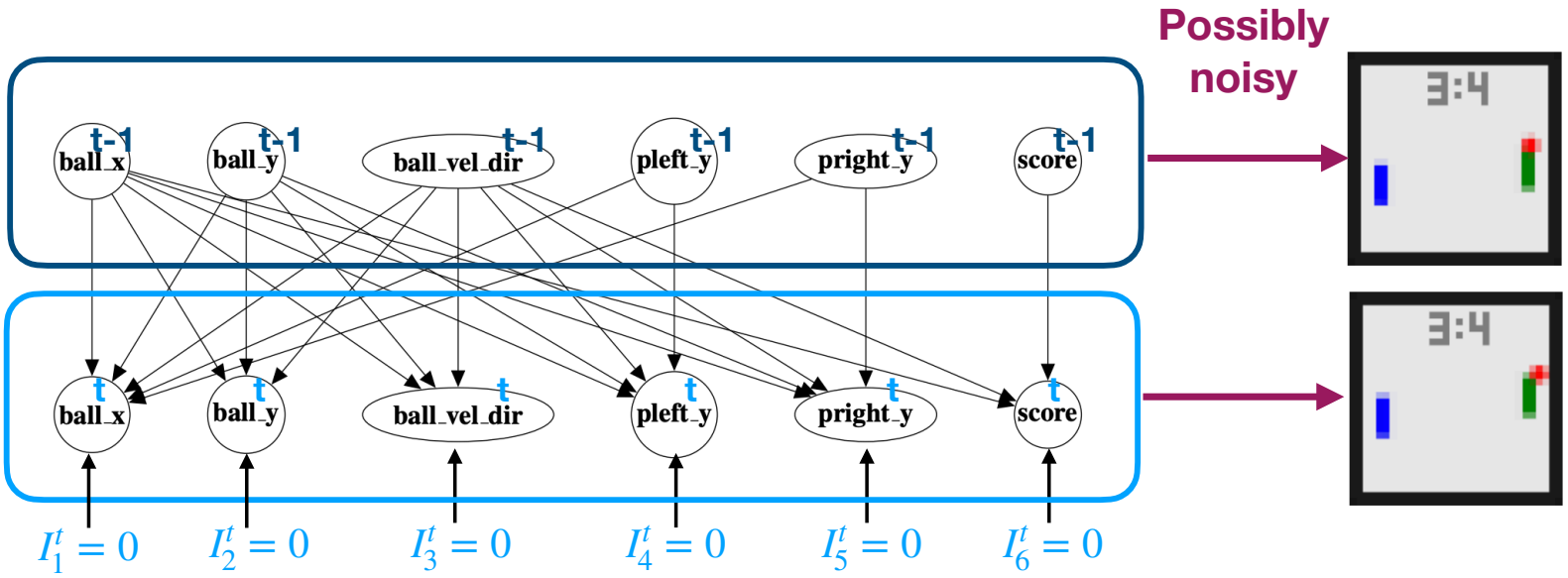
**First attempt: what if we knew what variables are intervened at each step?**



https://arxiv.org/abs/2202.03169

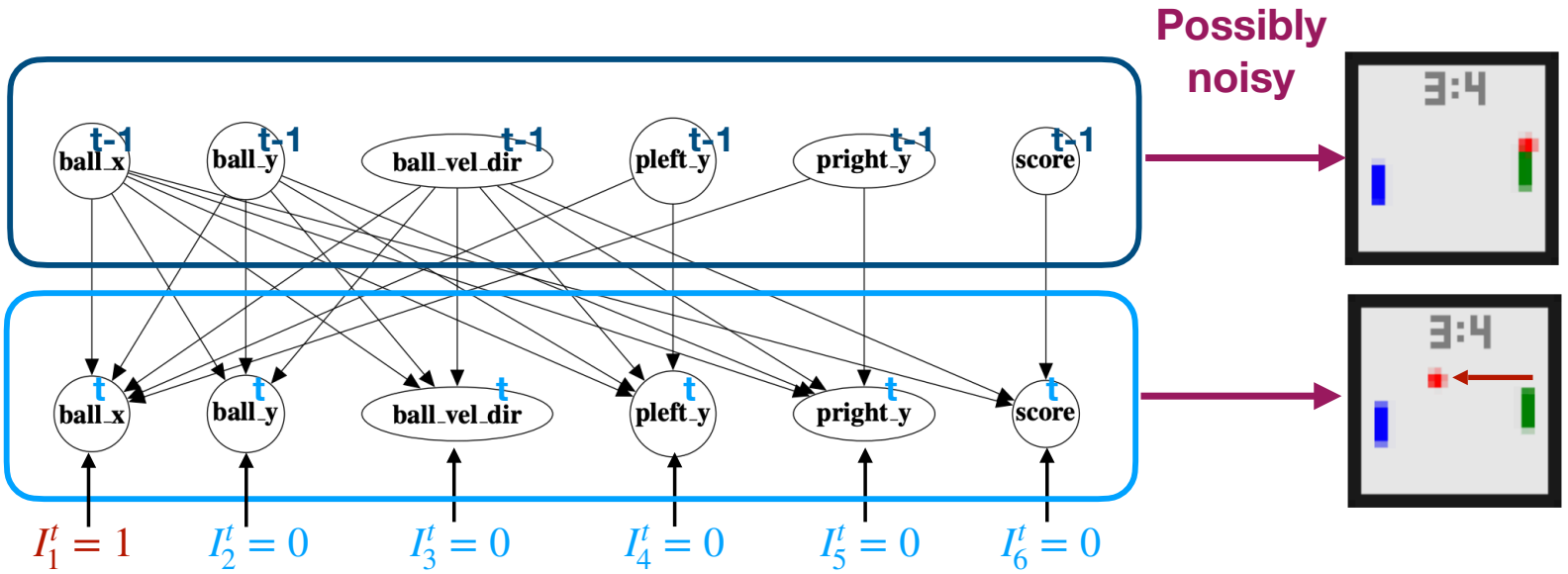
# CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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



# CITRIS: Causal Identifiability from TempoRal Intervened Sequences

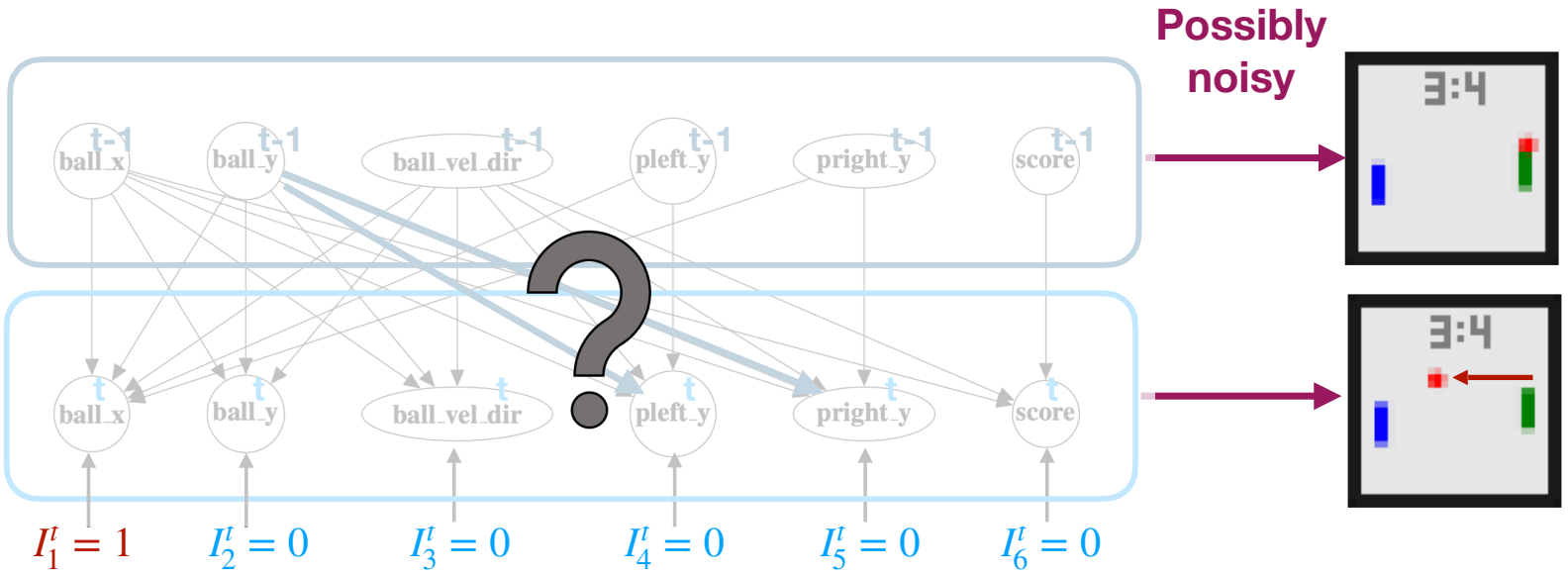
Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves




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

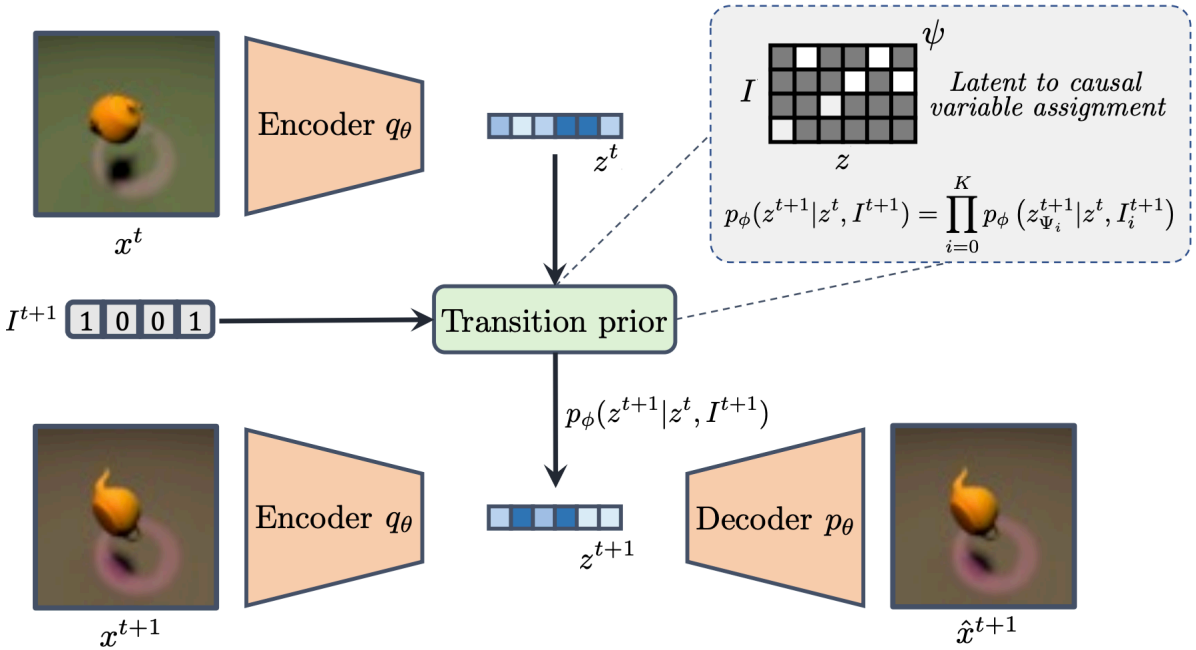
# CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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




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

# A variational autoencoder architecture: CITRIS-VAE



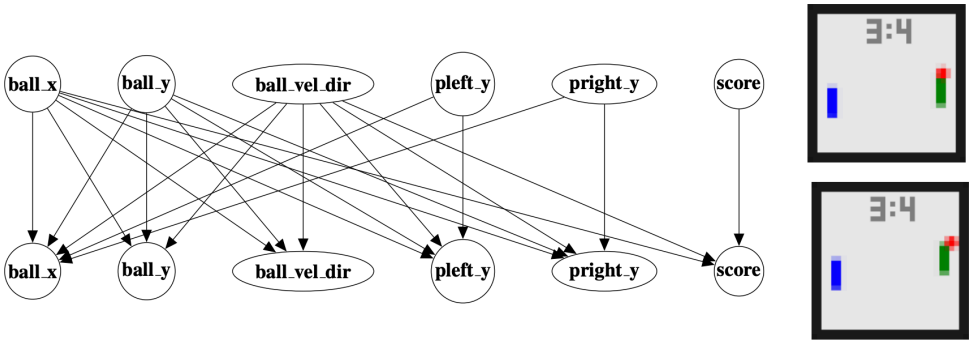

- Encoder + Decoder for images $x^t, x^{t+1}$
- Latent space $z^t, z^{t+1}$
- Transition prior $p_\phi$ between $t, t+1$
  - **Learns assignment $\psi$ from latent dimension $z_{\psi_i}$ to causal variable $C_i$ for $i = 1, \ldots, K$**
  - **We keep $z_{\psi_0}$ ("junk") for unassigned dimensions**
    - Maximise the information in "junk" so only essential information is assigned to the causal variables

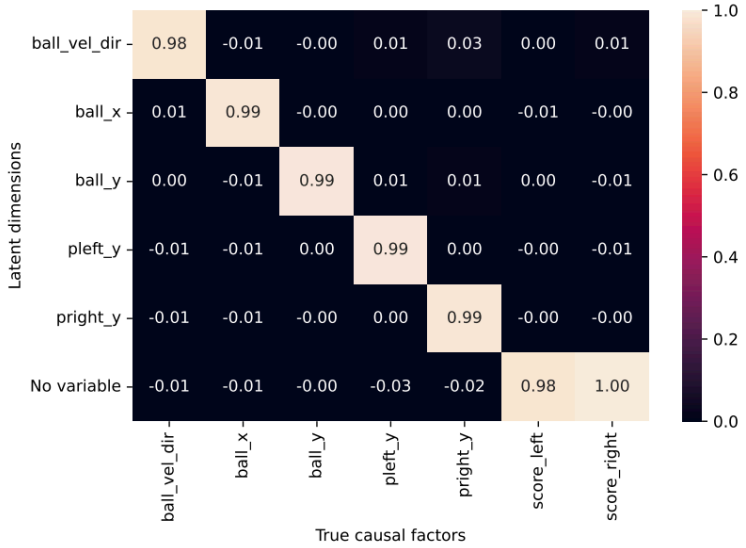
# CITRIS: Simplified identification results

- Assumption 1: Each $I_i^t$ is not a deterministic function of $I_j^t$
- **Assumption 2:** Interventions have an effect on **all components** of any multi-dimensional causal variable
- **Assumption 3:** There are **"enough"** interventions ( $O(log_2 K)$ )
- Then if we maximize both the likelihood $p(x^{t+1} \mid x^t, I^{t+1})$ and the information content of the latents assigned to "junk", we can identify causal variables $C_1, \ldots, C_K$ **up to element-wise transformations (NO PERMUTATIONS)**

# Experiments: Interventional Pong

- **Train:** We learn encoder $f$ on a dataset with **potentially dependent** causal variables from images $\{X^t\}_{t=1}^T$ and intervention targets $\{I^t\}_{t=1}^T$ -> **unsupervised**
- **Test:** We evaluate $f$ on a dataset with **independent** causal variables and evaluate correlation with ground truth causal variables.



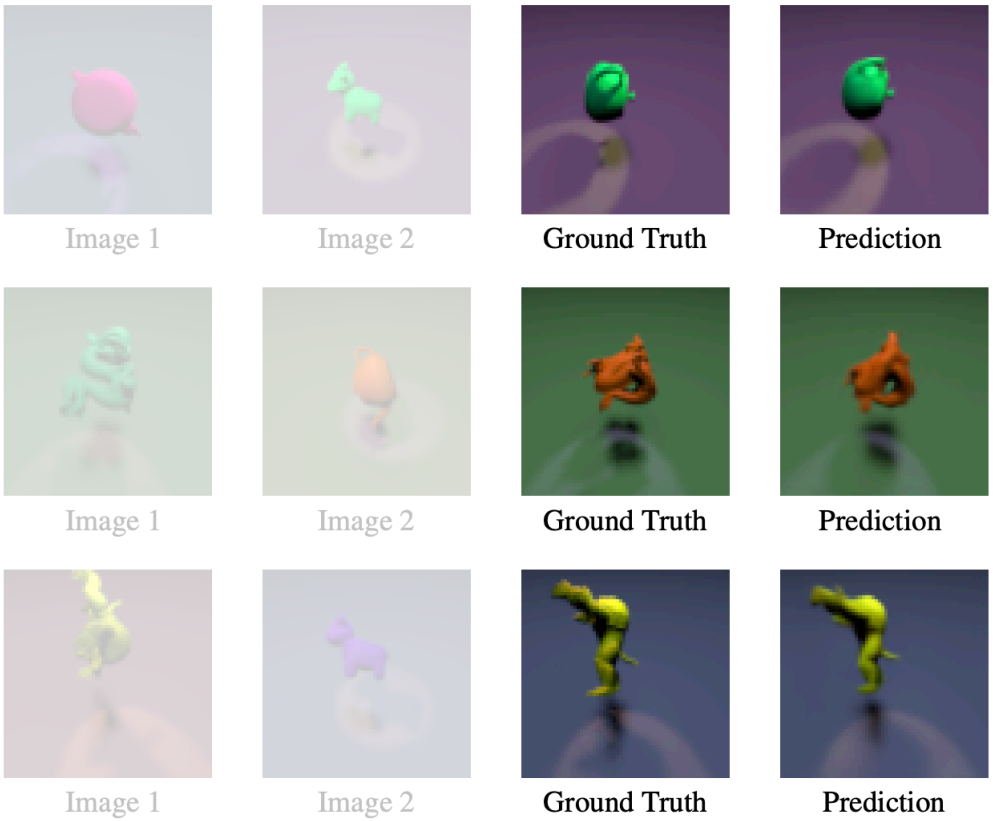



# Experiments: Temporal Causal3DIdent

shape + spotlight colors + rot


- We can steer the individual variables independently and generate new images

# Experiments: Temporal Causal3DIdent




- We can steer the individual variables independently and generate new images

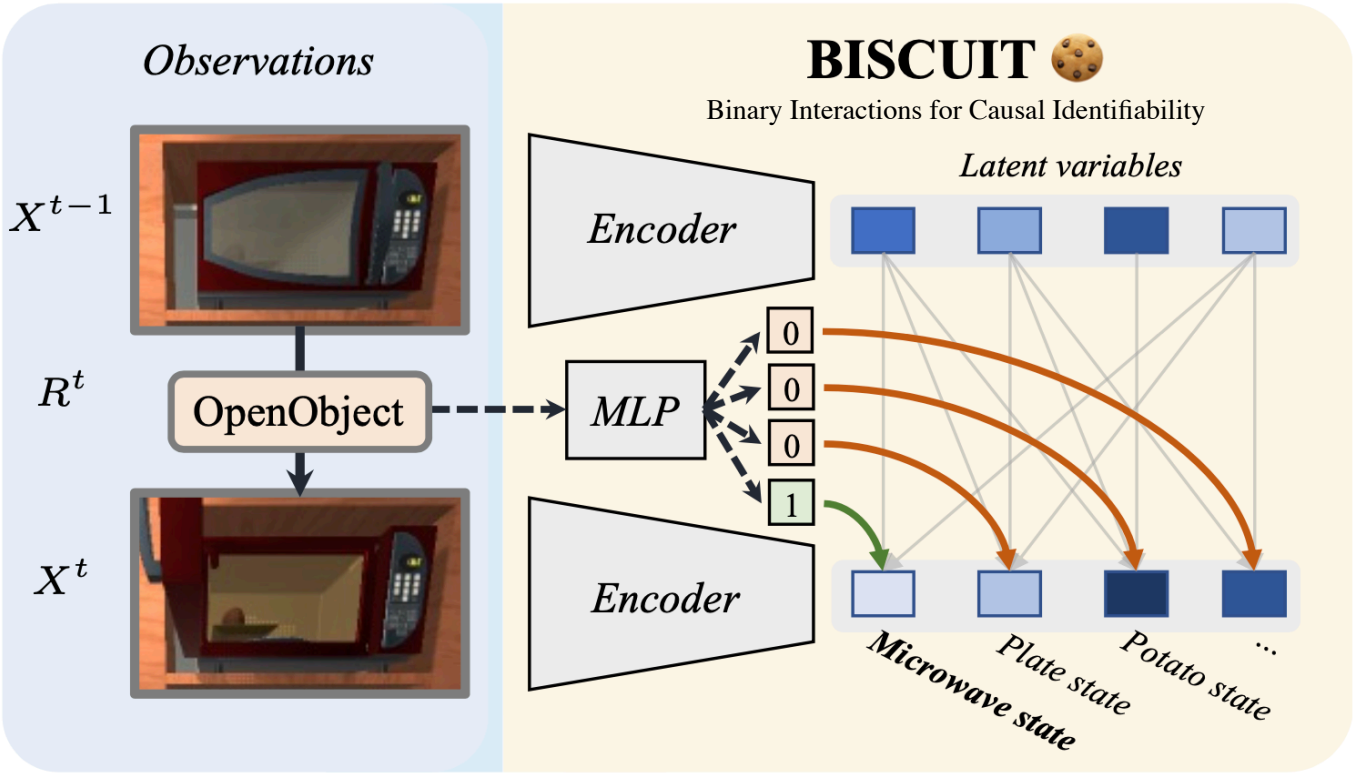
# CITRIS [Lippe et al. 2022]

- **Pros:**
  - Multidimensional causal variables
  - No parametric assumptions
  - Works also with arbitrary dense graphs
  - **Identifiability up to element-wise transformations**

- **Cons:**
  - Needs (sufficiently diverse) interventional data
  - **Needs known intervention targets**
  - Works only in a temporal setting

# BISCUIT: Causal Representation Learning from Binary Interactions

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

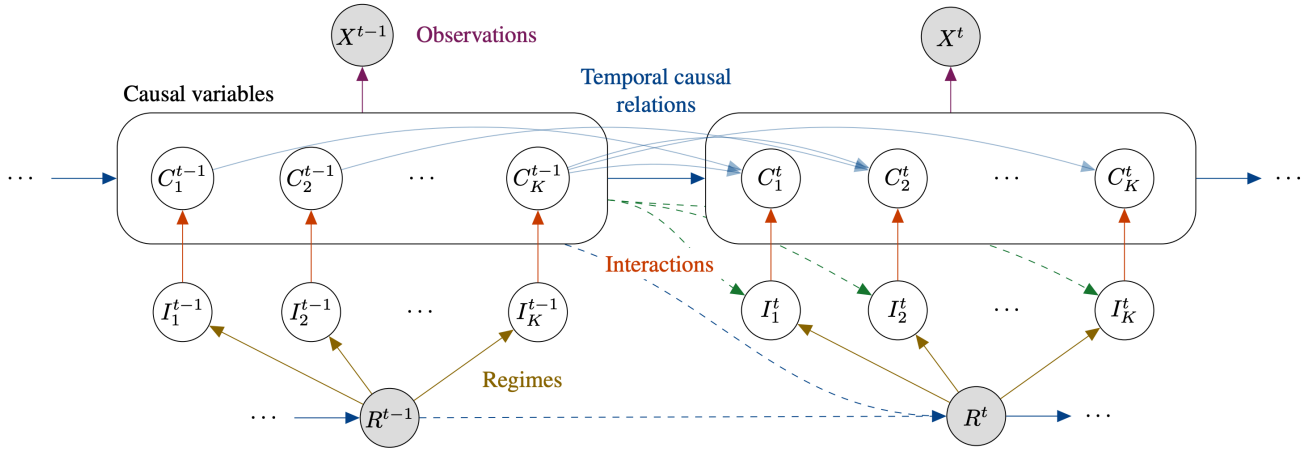



https://phlippe.github.io/BISCUIT/

# BISCUIT: Causal Representation Learning from Binary Interactions

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

- We identify latent variables $C_1, \ldots, C_K$ from **sequences** of **high-dimensional data** $\{X^t\}_{t=1}^T$,
- We **observe an action/regime** $R^t$ **(potentially modulated by previous state** $C^{t-1}$**)**
- **Intuitively**: for each $C_i$, we learn a **binary interaction variable** $I_i^t$ from $R^t$



# BISCUIT: Causal Representation Learning from Binary Interactions

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

- **Assumption 1:** interactions between the agent and each causal variable can be described by a **binary variable** -> each causal variable has only two mechanisms
- **Assumption 2:** each $I_i^t$ is not a function of any other $I_j^t$ given $C^{t-1}$
- **Assumption 3:** the mechanisms vary sufficiently either over interactions or time

A. (***Dynamics Variability***) *Each variable's log-likelihood difference is twice differentiable and not always zero:*

$$\forall C_i^t, \exists C^{t-1} : \frac{\partial^2 \Delta(C_i^t|C^{t-1})}{\partial (C_i^t)^2} \neq 0;$$

B. (***Time Variability***) *For any $C^t \in \mathcal{C}$, there exist $K+1$ different values of $C^{t-1}$ denoted with $c^1, ..., c^{K+1} \in \mathcal{C}$, for which the vectors $v_1, ..., v_K \in \mathbb{R}^{K+1}$ with*

$$v_i = \left[ \frac{\partial \Delta(C_i^t|C^{t-1}=c^1)}{\partial C_i^t} \quad \dots \quad \frac{\partial \Delta(C_i^t|C^{t-1}=c^{K+1})}{\partial C_i^t} \right]^T$$

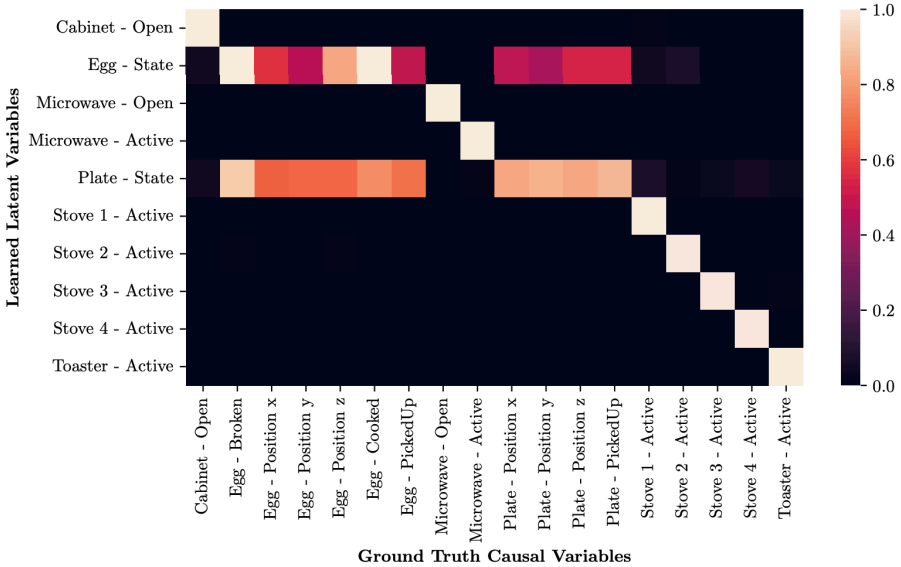
*are linearly independent.*

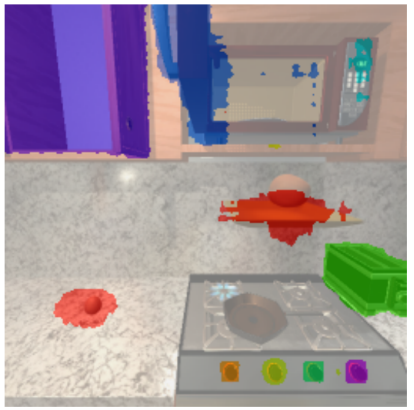
**Theoretical guarantee: each estimated variable identifies one of the true ones up to a transformation**: $\hat{C}_1 = h_1(C_{\pi(1)}), \dots, \hat{C}_K = h_K(C_{\pi(K)})$ for $h_1, \dots, h_K$ and permutation $\pi$

# BISCUIT: Causal Representation Learning from Binary Interactions

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

- We show that we learn a **disentangled**, **interpretable** and **manipulable** latent space in terms of the causal variables (e.g. object attributes), as well as a **dynamic map of interactions**, describing which action corresponds to an intervention on which variable.

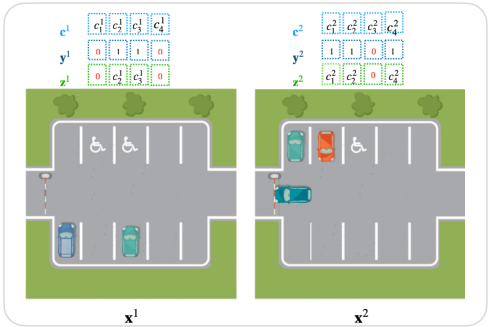


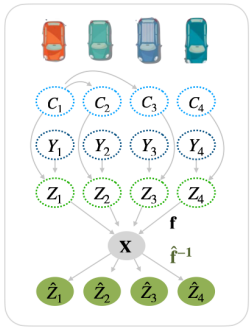



# A Sparsity Principle for Partially Observable Causal Representation Learning

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, Sara Magliacane

- We consider the setting in which each observation only provides information about a **subset of the causal variables**, and where these partial observations are **unpaired**.

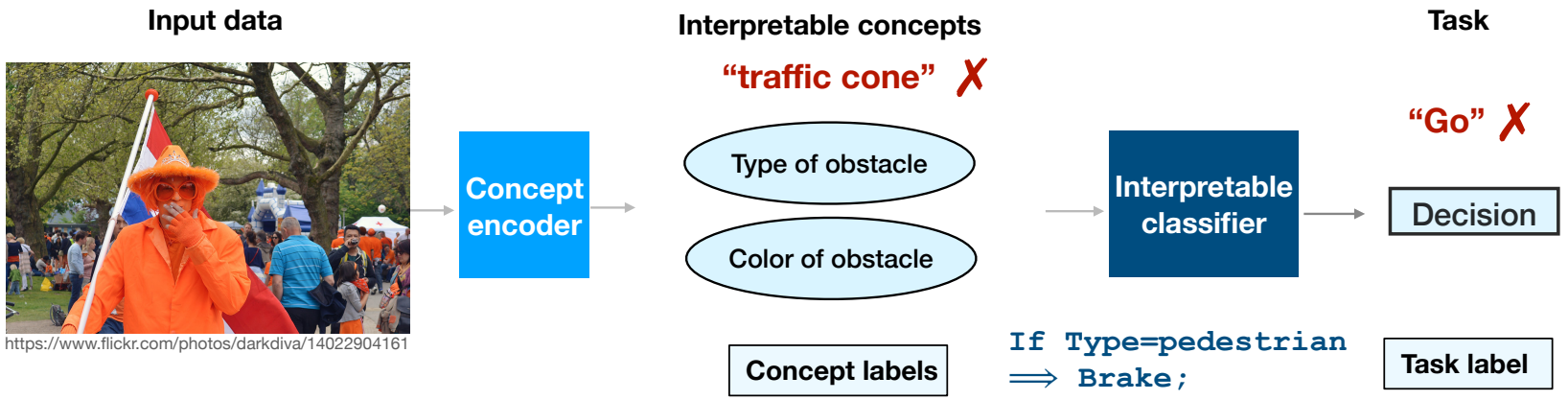




**An example of CRL in non-temporal settings**

- We provide results **identifying causal variables up to element-wise transformation and permutation** for (i) linear mixing functions, and (ii) piecewise-linear mixing functions for Gaussian causal variables.

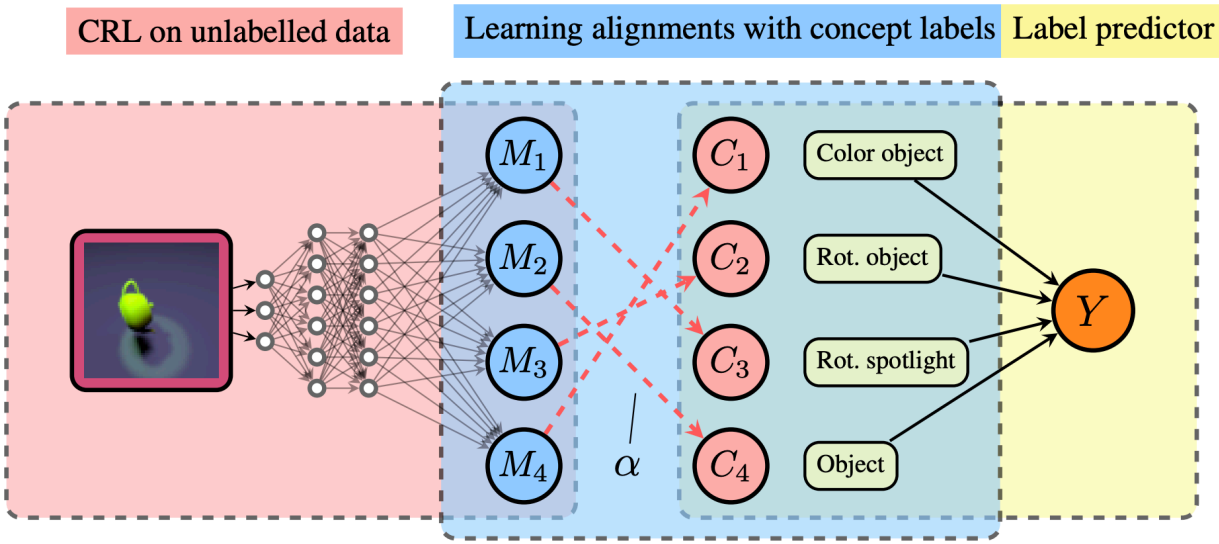
# Reprise: How do we guarantee that we learn correct concepts?



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

Hidde Fokkema, Tim van Erven, Sara Magliacane

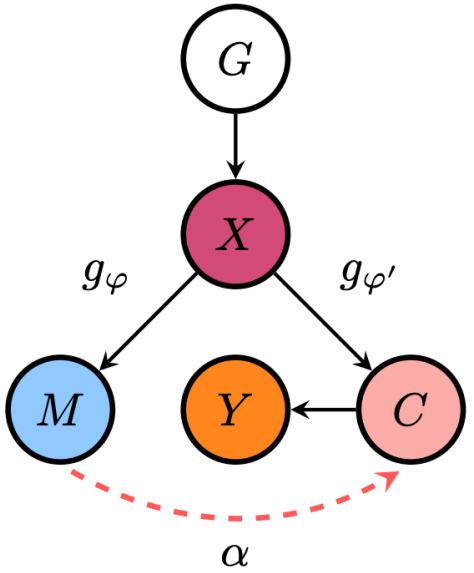
- A framework that leverages **Causal Representation Learning (CRL)** to learn concepts with theoretical guarantees on the error in learning the concepts with few concept labels




**We assume causal variables are 1:1 with concepts**

# Framework and assumptions

- We assume a causal system with **ground truth latent causal variables** $G = (G_1, \ldots, G_d)$
- We assume the **human-interpretable concepts** $C = (C_1, \ldots, C_d)$ correspond to the ground truth variables up to simple (element-wise) transformations
- We assume humans can identify these concepts from **high-dimensional observations** $X = f(G)$, e.g. images, and use them to interpretably predict the **label Y**
- We assume a CRL method learns **disentangled representations** of the causal variables $M = (M_1, \ldots, M_d)$ **up to permutation and element-wise transformation**




**Our task is to learn the alignment function $\alpha$ (permutation + element-wise transformation) with guarantees**

# Two types of estimators and results

1. Linear estimator - finite sample results

   - We consider **feature maps** $\varphi(M)$, e.g. splines, RFFs, etc., and model each concept $C_i$ as a linear function of the transformed variables and Gaussian noise

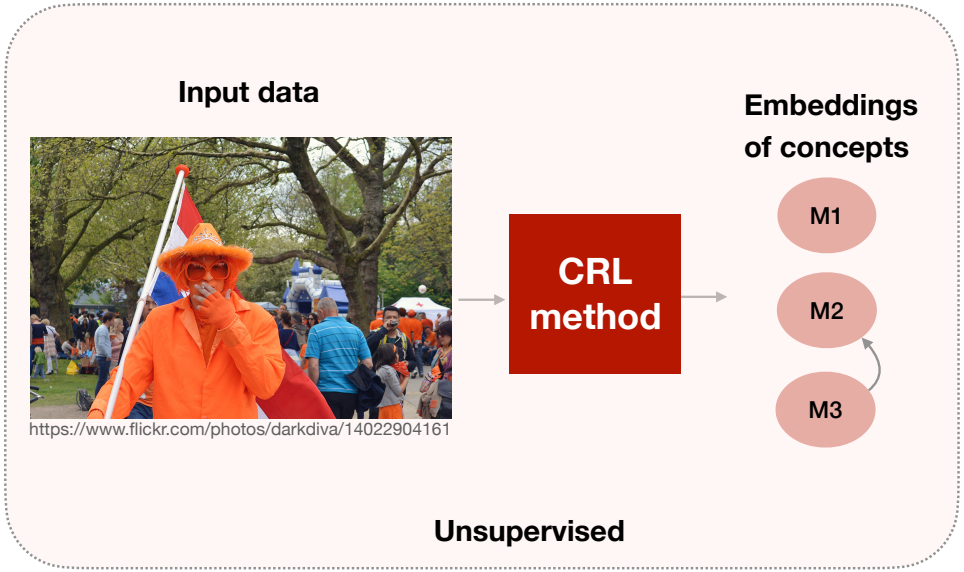
$$C_i = \varphi(M)\beta_i^* + \epsilon_i$$

   - Result: assuming low correlations between the $M_j$, GroupLasso will identify w**ith high probability the correct coefficients** up to a small constant that decreases with the number of concept labels

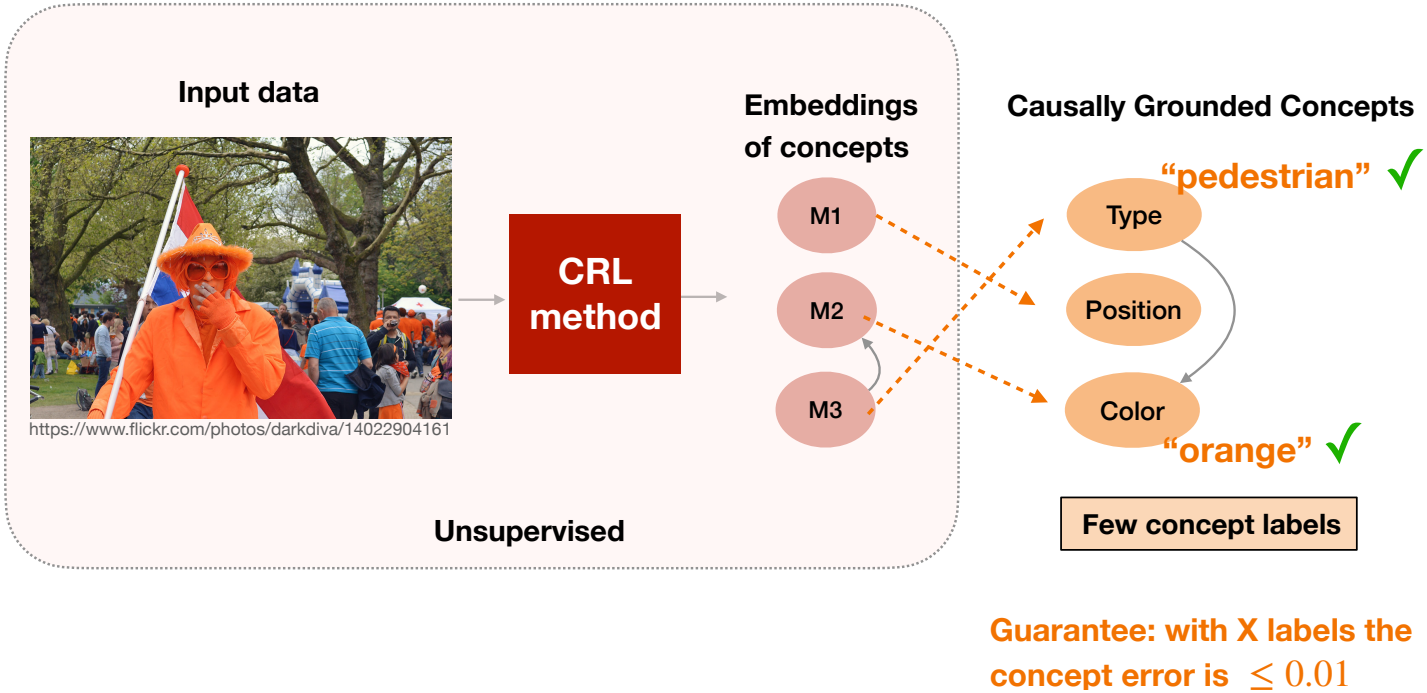
2. Kernelized estimator - asymptotic results

   - We consider functions in a RKHS and model $C_i = \beta_i^*(M) + \epsilon_i$
   - Result: GroupLasso will asymptotically converge to the right permutation

# Reprise 2: How do we guarantee that we learn correct concepts?

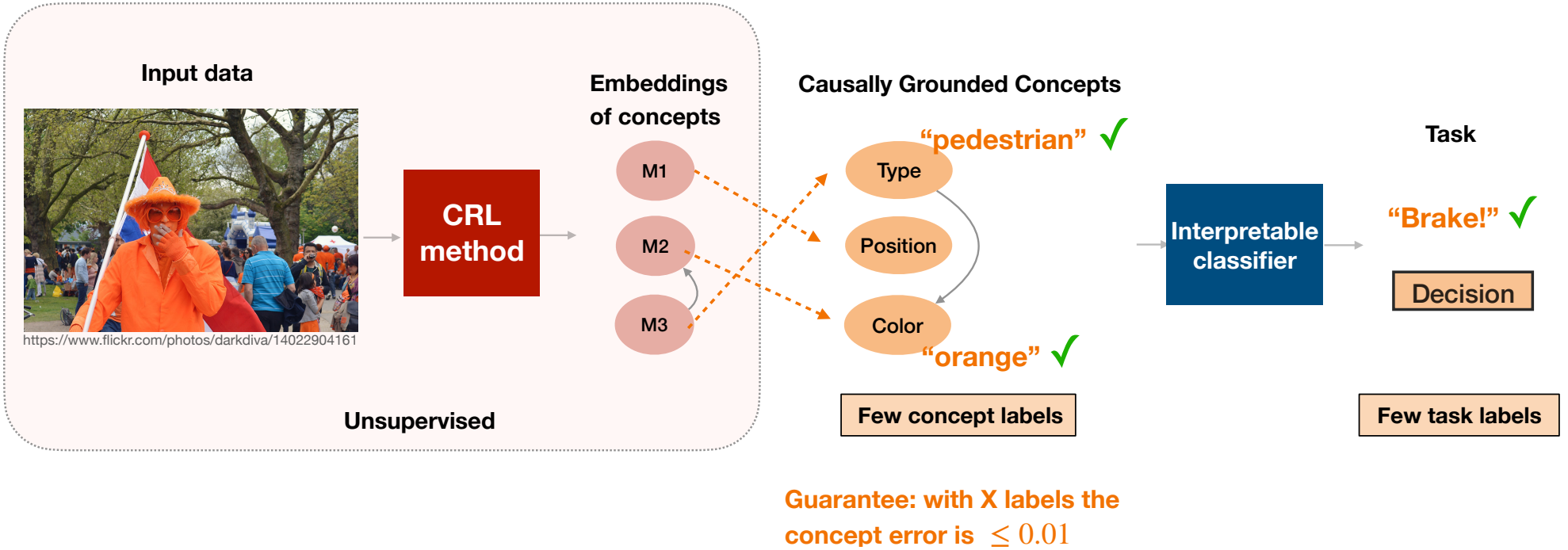


# Reprise 2: How do we guarantee that we learn correct concepts?

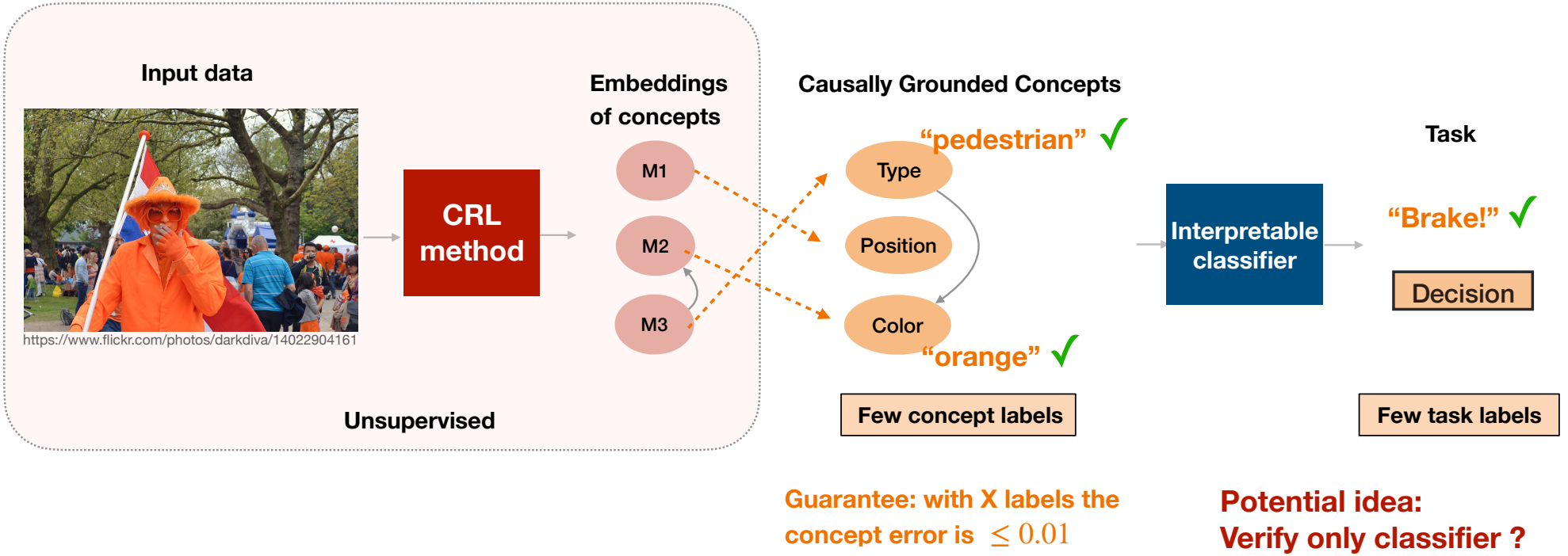



**Guarantee: with X labels the concept error is** $\leq 0.01$

# Reprise 2: How do we guarantee that we learn correct concepts?



# Reprise 2: How do we guarantee that we learn correct concepts?




Guarantee: with X labels the concept error is $\leq 0.01$

**Potential idea:**
**Verify only classifier ?**

# Conclusions and Limitations

- **Causal representation learning (CRL)** is an exciting field that allows us to extract identify high-level variables from observations with provable guarantees
  - We saw some methods in **temporal settings with actions**

# Conclusions and Limitations

- **Causal representation learning (CRL)** is an exciting field that allows us to extract identify high-level variables from observations with provable guarantees
  - We saw some methods in **temporal settings with actions**

- We saw **a pipeline that leverages CRL in CBMs** and provides guarantees on learning concepts from high-dimensional observations:
  - Can we use a similar pipeline for verifiably safe models from unstructured data?

# Conclusions and Limitations

- **Causal representation learning (CRL)** is an exciting field that allows us to extract identify high-level variables from observations with provable guarantees
  - We saw some methods in **temporal settings with actions**

- We saw **a pipeline that leverages CRL in CBMs** and provides guarantees on learning concepts from high-dimensional observations:
  - Can we use a similar pipeline for verifiably safe models from unstructured data?

- Limitations - **theoretical guarantees don't come for free…**
  - We assume the **concepts** correspond 1:1 to **CRL variables**
    - We could relax this to coarser concepts, but still we cannot guarantee unrelated concepts
  - CRL requires **specific settings to provide guarantees**, e.g., parametric assumptions, multi-view data, time and/or actions

# References

[Fokkema et al. 2025] Fokkema, H., van Erven, T., & Magliacane, S. (2025). Sample-efficient Learning of Concepts with Theoretical Guarantees: from Data to Concepts without Interventions. arXiv preprint arXiv:2502.06536.

[Koh et al. 2020] Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020, November). Concept bottleneck models. In International conference on machine learning (pp. 5338-5348). PMLR.

[Khemakhem et al., 2020] Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In Proceedings of the Twenty Third Inter- national Conference on Artificial Intelligence and Statistics, volume 108 of Proceedings of Machine Learning Research. PMLR, 2020.

[Lachapelle et al., 2022] Lachapelle, S., Rodriguez, P., Le, R., Sharma, Y., Everett, K. E., Lacoste, A., and Lacoste-Julien, S. Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA. In First Conference on Causal Learning and Reasoning, 2022.

[Lachapelle et al., 2024] Lachapelle, S., López, P. R., Sharma, Y., Everett, K., Priol, R. L., Lacoste, A., & Lacoste-Julien, S. (2024). Nonparametric Partial Disentanglement via Mechanism Sparsity: Sparse Actions, Interventions and Sparse Temporal Dependencies. arXiv preprint arXiv:2401.04890.

[Lippe et al., 2022] Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. CITRIS: Causal Identifiability from Temporal Intervened Sequences. In Proceedings of the 39th International Conference on Machine Learning, ICML, 2022.

[Lippe et al., 2023a] Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. Causal representation learning for instantaneous and temporal effects in interactive systems. In The Eleventh International Conference on Learning Representations, 2023.

[Lippe et al., 2023b] Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. BISCUIT: Causal Representation Learning from Binary Interactions. UAI 2023.

[Yao et al., 2022] Yao, W., Sun, Y., Ho, A., Sun, C., and Zhang, K. Learning Temporally Causal Latent Processes from General Temporal Data. In International Conference on Learning Representations, 2022.