ISTITUTO ITALIANO DI TECNOLOGIA
PATTERN ANALYSIS AND COMPUTER VISION

# Structured representations: pushing causality for visual data

Davide Talon

April 13th 2022

# Agenda

*Part 1*

## Causality 101

- From statistical to causal models
- The structural causal model
- Identifiability problem

*Part 2*

## High dimensional data

- Linear and non-linear ICA
- Disentanglement
- The identifiability problem
- Cross-pollination: causality and disentanglement

*Part 3*

## Causal signals in Visual data

- Causal signal for images
- Causal visual datasets

# Agenda

*Part 1*

## Causality 101

**From statistical to causal models**

**The structural causal model**

**Identifiability problem**

*Part 2*

## High dimensional data

**Linear and non-linear ICA**

**Disentanglement**

**The identifiability problem**

**Cross-pollination: causality and disentanglement**

*Part 3*

## Causal signals in Visual data

**Causal signal for images**

**Causal visual datasets**

# Causal relationship

- **Cause-effect:** externally intervening the cause may change the effect, but not vice versa

Pearl, Judea. *Causality*. Cambridge university press, 2009.

# In pizza we trust

- Taste
- Ingredients
- Bakery
- Me - Davide :)

VectorStock.com/32401978

# The ladder of causation


**COUNTERFACTUAL**
I baked it for 5’ and burnt it out.
Had I baked for 3’, would I have burnt it?

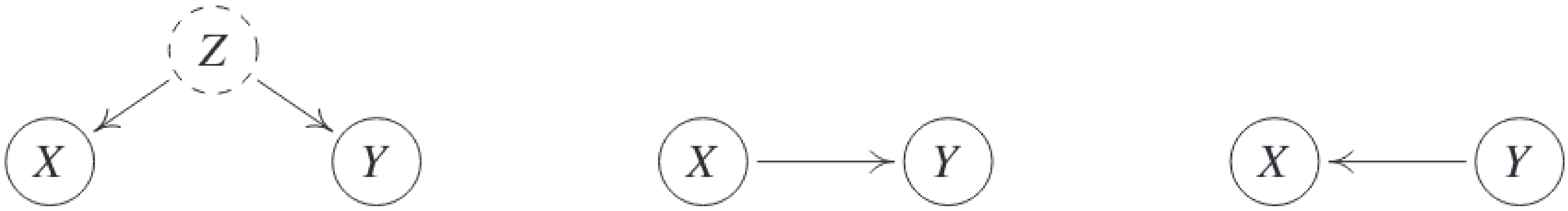
**INTERVENTION**
Let’s skip mozzarella. Will it be good?

**OBSERVATION**
What does the color of edge tell me about how good it is?

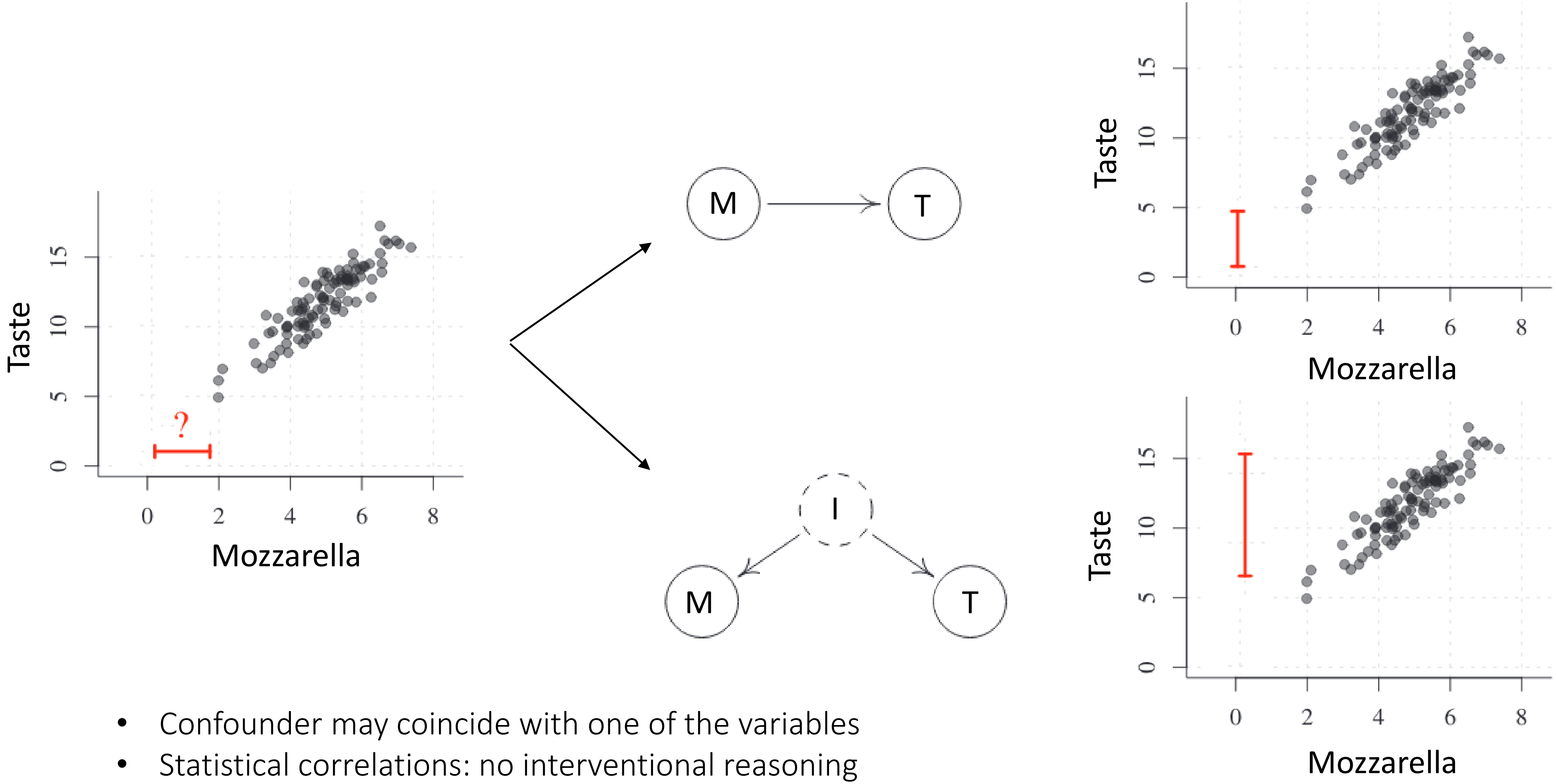
Adapted from Pearl and Mackenzie, "The book of why: the new science of cause and effect.", Penguin, 2019.

# Common Cause principle

- **Common Cause principle**: if two random variables X and Y are statistically dependent, then there exists a third variable Z that causally influences both.



# Common Cause principle: an example




- Confounder may coincide with one of the variables
- Statistical correlations: no interventional reasoning

Adapted from Peters, Jonas, et al. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

# Structural Causal Models (SCMs)

- **Structural Causal Models**: a SCM $\mathfrak{C} = (\mathbf{S}, P_{\mathbf{N}})$ consists of a set $\mathbf{S}$ of structural assignments

$$X_j = f_j(\mathbf{PA}_j, N_j), \quad j = 1, \ldots, d$$

where $\mathbf{PA}_j \subseteq \{X_1, \ldots, X_d\} \setminus X_j$ are the parents (direct causes) of $X_j$ and $P_{\mathbf{N}}$ is the jointly independent distribution of noises.

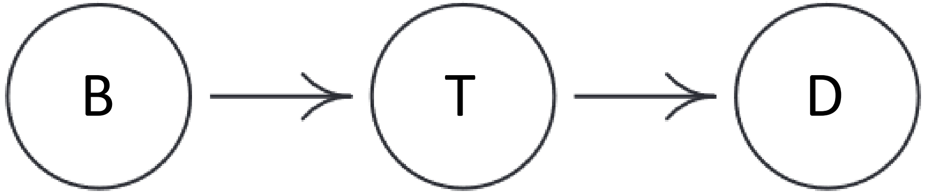
$$\begin{aligned} X_1 &= N_{X_1} \\ Y &= X_1 + N_Y \\ X_2 &= Y + N_{X_2} \\ N_{X_1}, N_Y &\sim \mathcal{N}(0,1), N_{X_2} \sim \mathcal{N}(0, 0.1) \end{aligned}$$

$$X_1 \longrightarrow Y \longrightarrow X_2$$

# SCM: properties

- **Entailed distribution**: an SCM $\mathfrak{C}$ defines a unique distribution over variables $X_1, \ldots, X_d$
- **Entailed graph**: an SCM entails a graph $\mathcal{G}$ obtained by drawing a node for each observable $X_j$ and a direct edge from parents $\mathbf{PA}_j$ to $X_j$

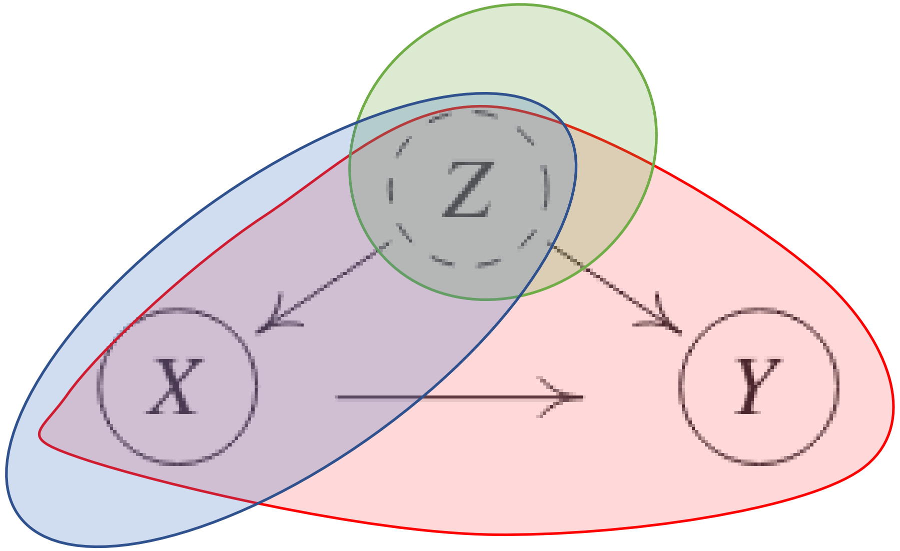
$$X_B = N_B$$
$$X_T = X_B + N_T$$
$$X_D = X_T + N_D$$
$$N_B, N_T \sim \mathcal{N}(0,1), N_D \sim \mathcal{N}(0, 0.1)$$




B - Baking, T - Taste, D - Davide

# Independent Causal Mechanisms

- A structural assignment is called mechanism
- Thanks to independence of noise, functionals are independent:
  - Knowledge about one mechanism does not convey information about others
  - Intervening on one mechanism does not effect others



# SCM: do-interventions

- **Interventional distribution**: Consider an SCM $\mathfrak{C} = (\mathbf{S}, P_{\mathbf{N}})$ with its entailed distribution $P_{\mathbf{X}}^{\mathfrak{C}}$. We can replace one (or several) structural assignments to obtain a new SCM. Suppose we intervene on $X_k$ :

$$\tilde{X}_k = \tilde{f}(\tilde{\mathbf{PA}}_k, \tilde{N}_k)$$

the entailed distribution of the new SCM is the interventional distribution

$$P_{\mathbf{X}}^{\tilde{\mathfrak{C}}} = P_{\mathbf{X}}^{\mathfrak{C};do(X_k=\tilde{X}_k)}$$

# Causal model and interventions

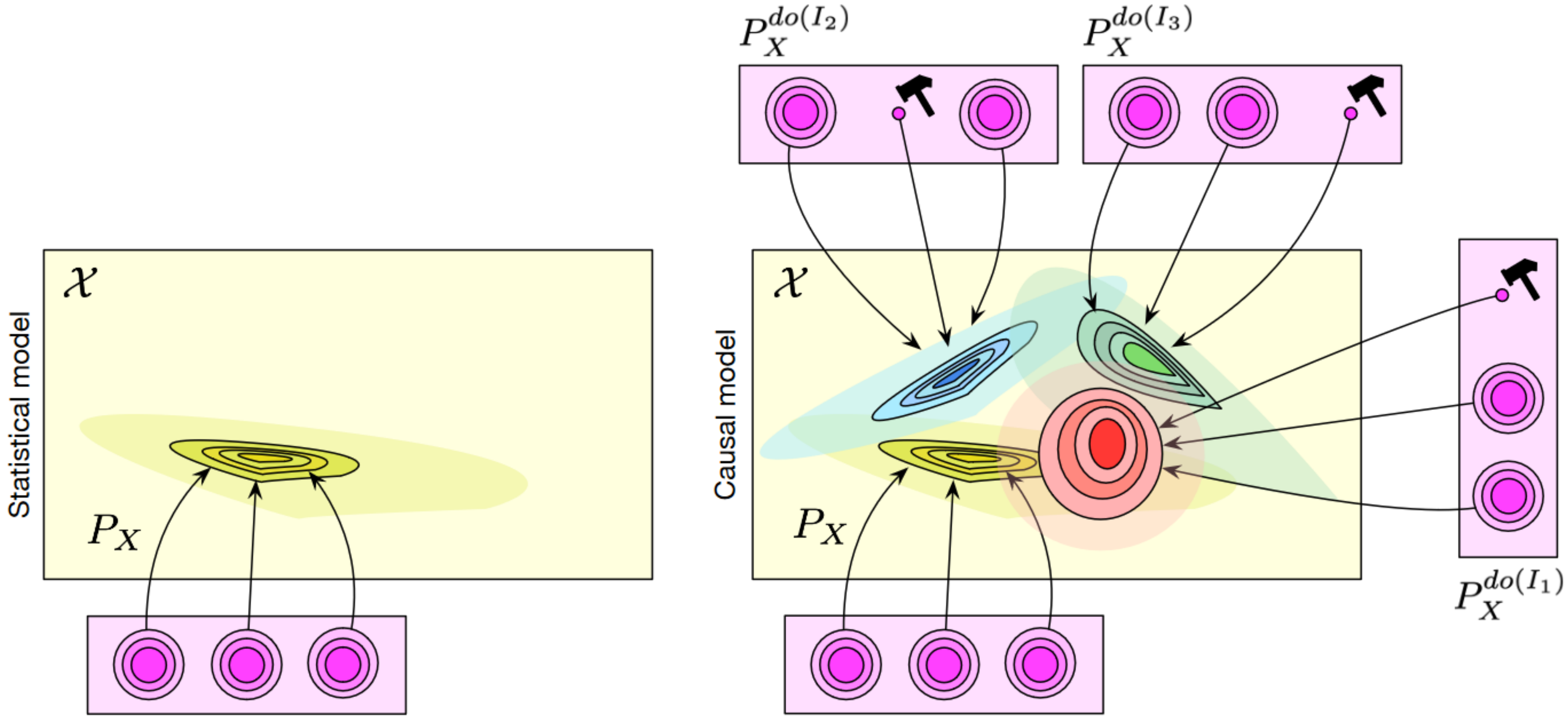



Fig. 1. Difference between statistical (left) and causal models (right) on a given set of three variables. While a statistical model specifies a single probability distribution, a causal model represents a set of distributions, one for each possible intervention (indicated with a in the figure).

Schölkopf, Bernhard, et al. "Towards causal representation learning." *arXiv preprint arXiv:2102.11107* (2021).

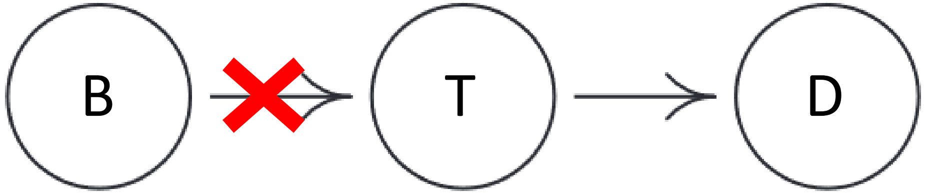
# Do-interventions in practice

$$X_B = N_B$$

$$\boxed{X_T = c}$$

$$\boxed{X_D = c + N_D}$$

$$N_B, N_T \sim \mathcal{N}(0,1), N_D \sim \mathcal{N}(0, 0.1)$$

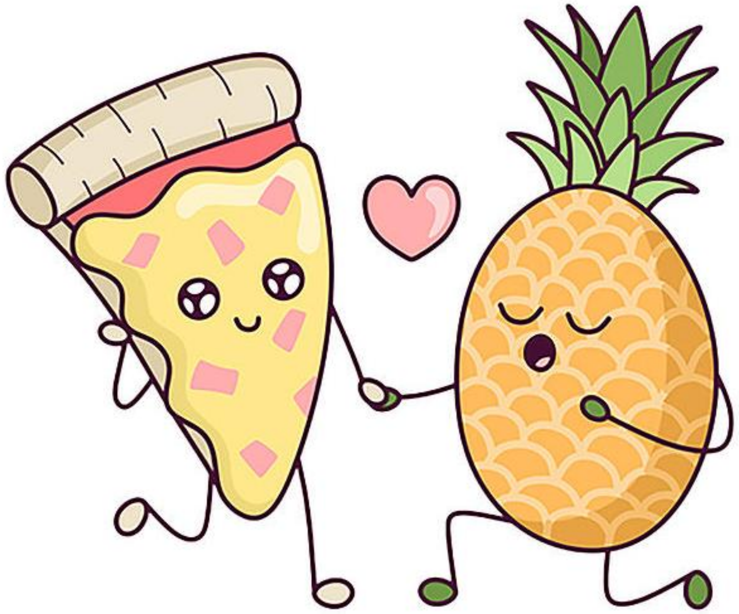



- Detach the intervened variable. Assign it an arbitrary value, independently from causes.
- Do-intervention ≠ conditioning:

$$P_T^{\mathfrak{C};do(D:=d)}(t) = P_T^{\mathfrak{C}}(t) \neq P_T^{\mathfrak{C}}(t \mid D = d)$$

# Counterfactuals

- Counter-fact: something not happen
- Given a fact, what would have been if we had taken another choice?
    - e.g., the pizza is good, what would have it been with pineapple?



# SCM: Counterfactuals

- **Counterfactuals:** Consider an SCM $\mathfrak{C} = (\mathbf{S}, P_{\mathbf{N}})$ over observables $\mathbf{X}$. Given some observation $\mathbf{x}$ , we define the counterfactual model as the SCM

$$\mathfrak{C}_{\mathbf{X}=\mathbf{x}} = (\mathbf{S}, P_N^{\mathfrak{C}|\mathbf{X}=\mathbf{x}})$$

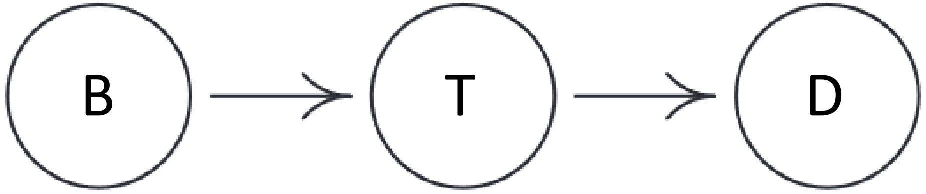
with $P_N^{\mathfrak{C}|\mathbf{X}=\mathbf{x}} = P_{N|\mathbf{X}=\mathbf{x}}$.

- Counterfactual statements are do-interventions in the counterfactual SCM

$$P_Z^{\mathfrak{C}|\mathbf{X}=\mathbf{x};do(Y:=c)}$$

# Counterfactuals in practice

$$X_B = N_B$$
$$X_T = X_B + N_T$$
$$X_D = X_T + N_D$$
$$N_B, N_T \sim \mathcal{N}(0,1), N_D \sim \mathcal{N}(0, 0.1)$$




B - Bake, T - Taste, D - Davide

- Counterfactual: given the fact that $\mathbf{X} = \mathbf{x}$, what would D have been, had T been set to 0?
  1. Compute exogenous
  2. Apply the intervention

# Long story short

- Causal models are more informative than statistical ones
- Causal models entail a set of distributions:
  - Observational distribution
  - Interventional distribution
  - Counterfactual distribution

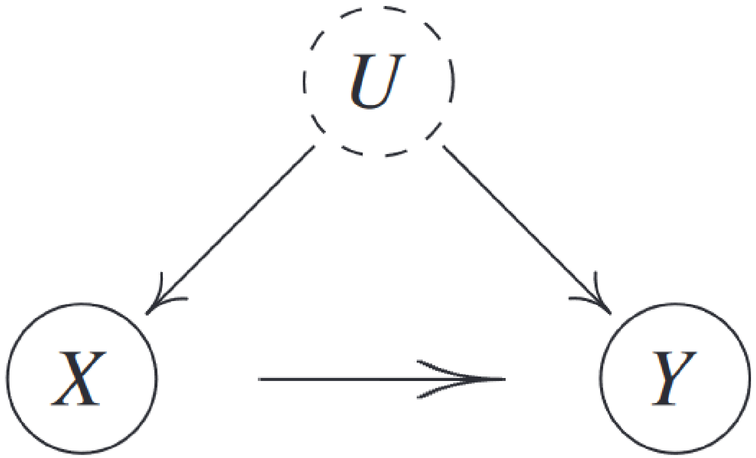
# What we will see

- The confounding problem
- Learning from observational data
- Representative methods from causal learning

# Confounding

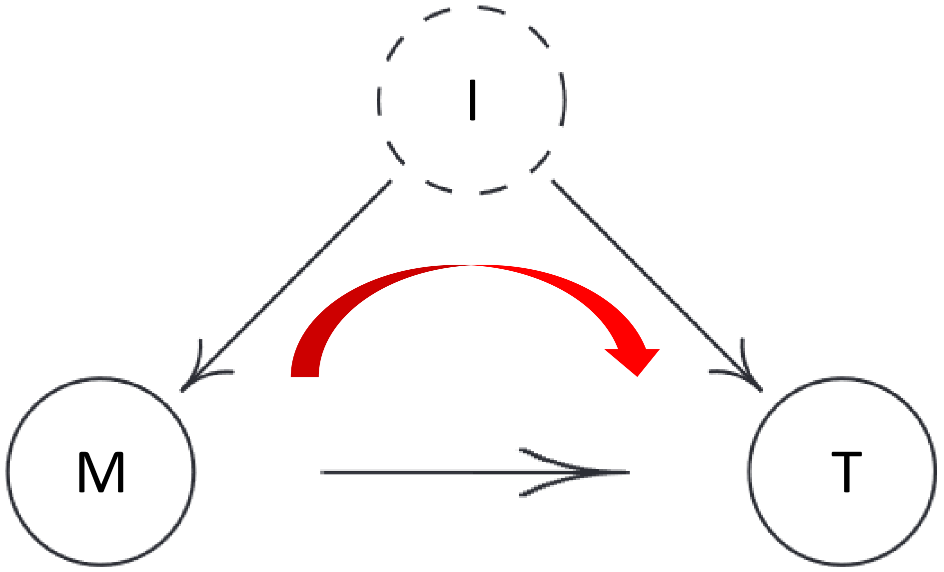
- Confounding: consider an SCM $\mathfrak{C}$ with direct path from $X$ to $Y$. The causal effect from $X$ to $Y$ is confounded if

$$p^{\mathfrak{C};do(X:=x)}(y) \neq p^{\mathfrak{C}}(y \mid x)$$



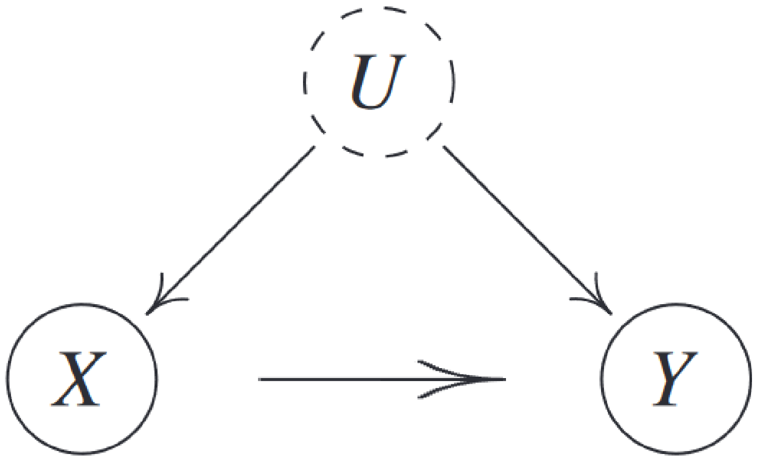
# Confounding in practice

$$p^{\mathfrak{C};do(M:=m)}(t) \neq p^{\mathfrak{C}}(t \mid m)$$



# Cause effects discovery

- Confounding is a serious problem: cannot evaluate cause-effect without confounder control.




- Gold standard: randomized interventions. Randomly intervene on the cause and observe eventual effects.

# Cause effects discovery

- Randomized interventions may be unethical or too expensive.
- Learn a causal model from observational data.

- **Theorem (non-identifiability):** for every joint distribution $P_{X,Y}$ of two real value variables, there is a SCM

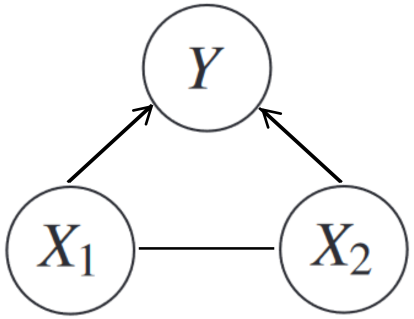
$$Y = f_Y(X, N_Y)$$

with $X$ and $N_Y$ independent.

# Inductive Causation Algorithm

**Input**: a faithful distribution
**Output**: equivalence class graph

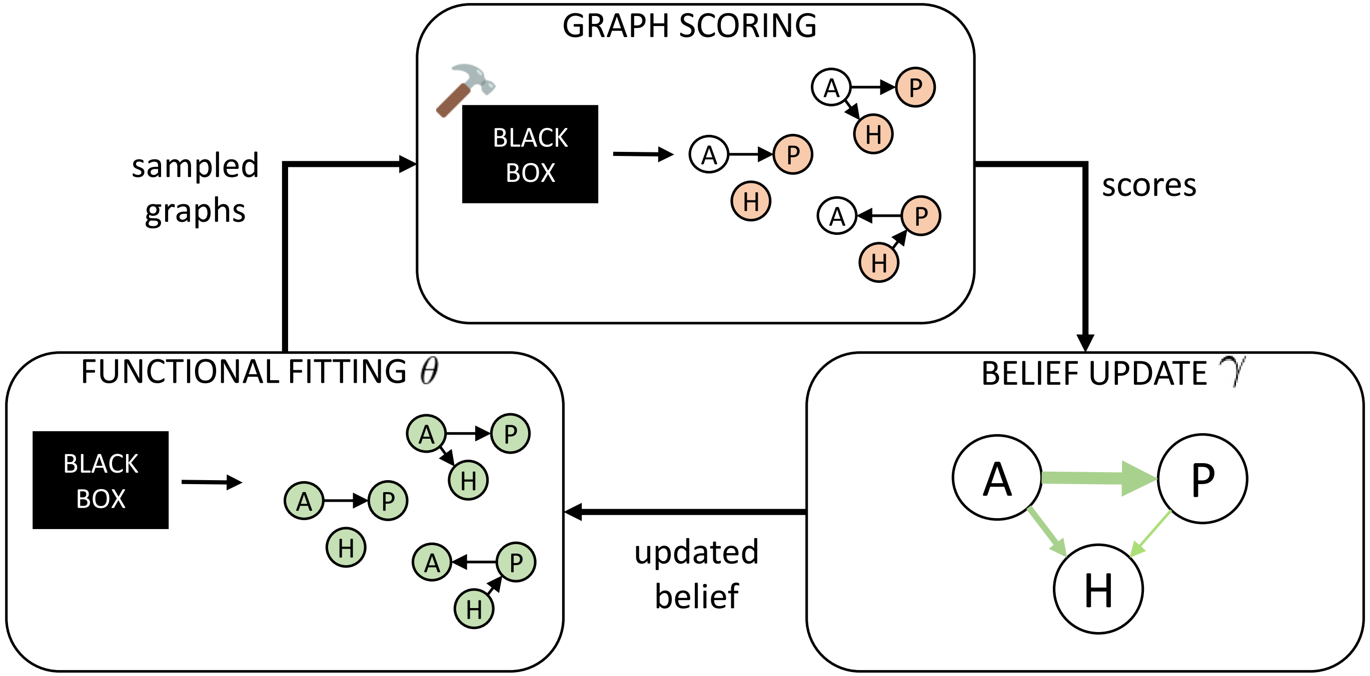
1. For each pairs of variables seek for the set rendering them independent. If no set exists, then they are connected
2. For each pair of non-adjacent variables with a common neighbor *c*, check if conditioning on *c* makes them independent. If not set the direction of edges
3. Orient as many edges as possible, e.g., avoid directed cycles



# Structural Discovery from Interventions

- Black-box model with unknown interventions
- Iterative score-based optimization

1. Fit functional parameters $\theta$ on observational data
2. Draw different causal graphs based on the current belief
3. Score mechanisms on interventional data obtained from the black-box model
4. Update current belief $\gamma$ according to scores and back to (1)




Ke, Nan Rosemary, et al. "Learning neural causal models from unknown interventions." *arXiv preprint arXiv:1910.01075* (2019).

# Long Story Short

- Confounding: correlation is not causation
- Cannot learn causal models from observational data
- Representative methods:
  - Conditional independence
  - Score-based

# Any questions?

*Part 1*

Causality 101

**From statistical to causal models**

**The structural causal model**

**Identifiability problem**

*Part 2*

High dimensional data

**Linear and non-linear ICA**

**Disentanglement**

**The identifiability problem**

**Cross-pollination: causality and disentanglement**

*Part 3*

Causal signals in Visual data

**Causal signal for images**

**Causal visual datasets**

# Moving on

- So far, high-level causal variables
- Causal variables not readily available
- How to find them?

# Agenda

*Part 1*

Causality 101

**From statistical to causal models**

**The structural causal model**

**Identifiability problem**

*Part 2*

High dimensional data

**Linear and non-linear ICA**

**Disentanglement**

**The identifiability problem**

**Cross-pollination: causality and disentanglement**

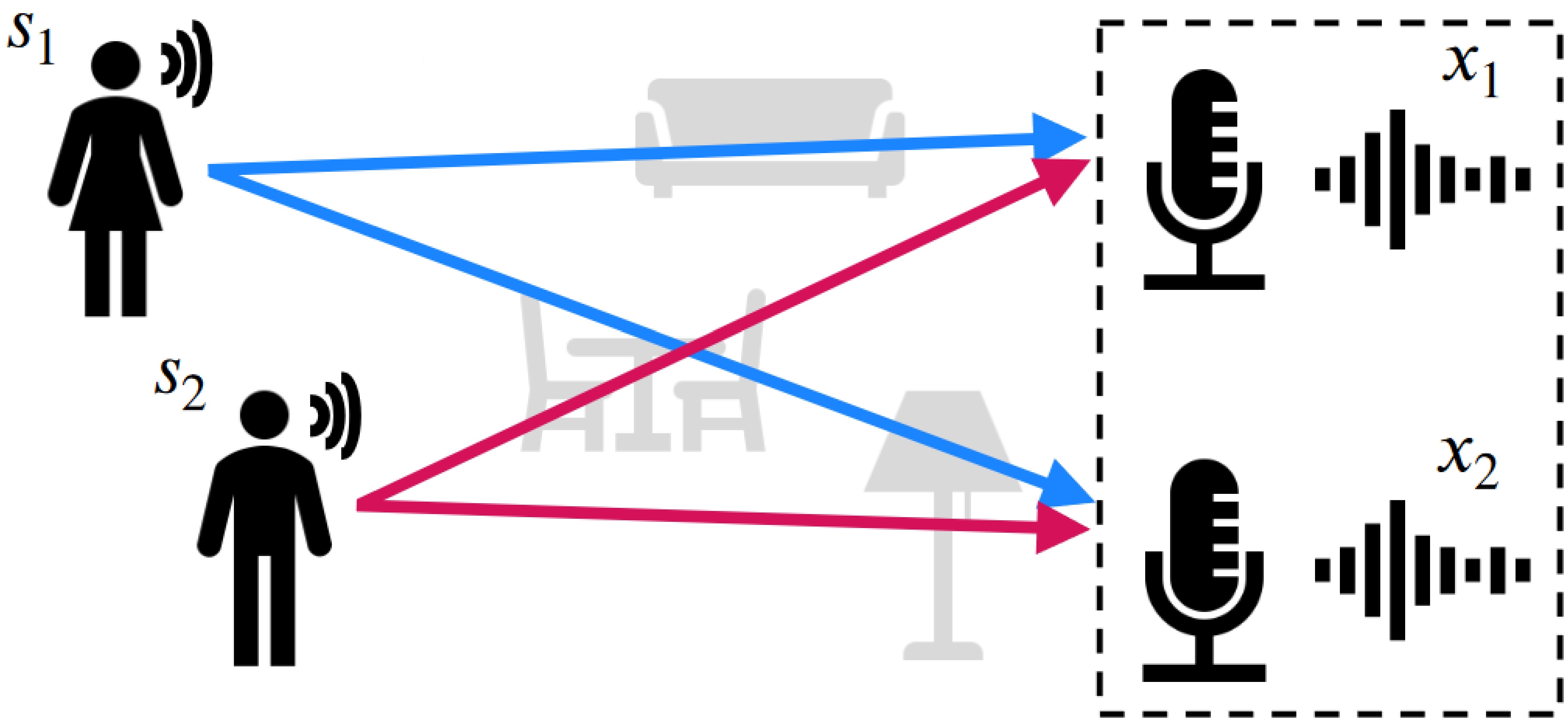
*Part 3*

Causal signals in Visual data

**Causal signal for images**

**Causal visual datasets**

# The cocktail party problem




Adapted from Gresele, Luigi, et al. "Independent mechanism analysis, a new concept?." *NeurIPS*, 2021.

# Linear Independent Component Analysis (ICA)

- Independent latent components $\mathbf{s} \in \mathbb{R}^n$
- Observations $\mathbf{x} \in \mathbb{R}^n$
- Mixing matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$
- Generative model:

$$\mathbf{x} = \mathbf{A}\mathbf{s}$$

Hyvärinen, Aapo, and Erkki Oja. "Independent component analysis: algorithms and applications." *Neural networks*, 2000.

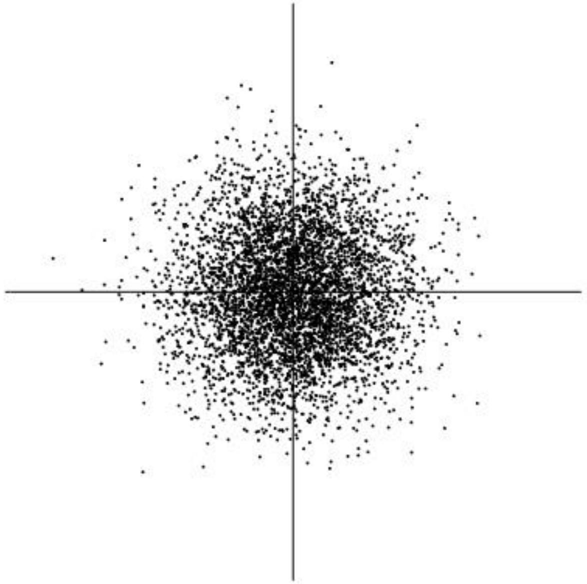
# Ambiguities of linear ICA

- In $\mathbf{x} = \mathbf{A}\mathbf{s}$, estimate both $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $\mathbf{s} \in \mathbb{R}^{n}$
- We cannot determine:
  - Variances of components: multiplicative factors may be canceled by A
  - Order of components: we can change the order in the linear combination

Hyvärinen, Aapo, and Erkki Oja. "Independent component analysis: algorithms and applications." *Neural networks*, 2000.

# Non identifiability of Gaussian case

- Assume orthogonal mixing matrix $\mathbf{A}$ (unit eigenvalues), e.g., rotation matrix
- Gaussian components with unit variance $\mathbf{s}$
- Observations $\mathbf{x} = \mathbf{A}\mathbf{s}$ are Gaussian and symmetric
- Observations do not expose information about $\mathbf{A}$



Hyvärinen, Aapo, and Erkki Oja. "Independent component analysis: algorithms and applications." *Neural networks*, 2000.

# Principles of ICA estimation

- Central limit theorem (informal): sum of independent random variables tends toward Gaussian distribution
- Consider a linear combination of $x_i$:

$$y = \mathbf{w}^T \mathbf{x} = \sum_i w_i x_i$$

- Rewrite $\mathbf{z} = \mathbf{A}^T \mathbf{w}$
- Then: $$y = \mathbf{w}^T \mathbf{x} = \mathbf{w}^T \mathbf{A} \mathbf{s} = \mathbf{z}^T \mathbf{s}$$
- Least gaussianity: $y$ corresponds to $s_i$

Hyvärinen, Aapo, and Erkki Oja. "Independent component analysis: algorithms and applications." *Neural networks*, 2000.

# Non-linear ICA

- Independent latent components $\mathbf{s} \in \mathbb{R}^n$
- Observations $\mathbf{x} \in \mathbb{R}^n$
- Smooth and invertible non linear mixing function:

$$f : \mathbb{R}^d \to \mathbb{R}^d$$

- Generative model:

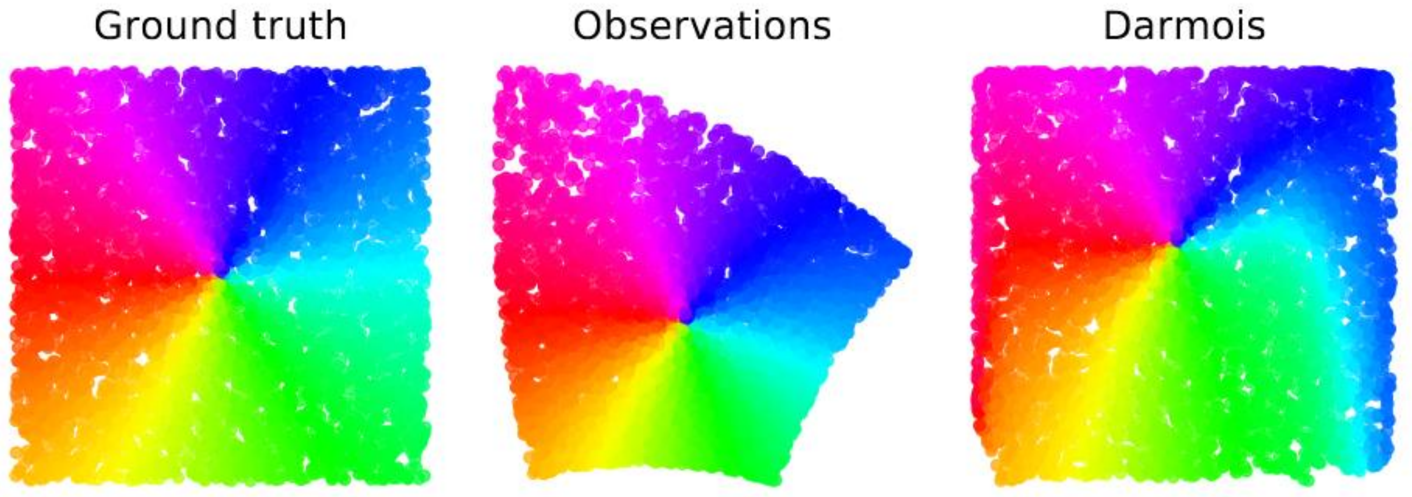
$$\mathbf{x} = f(\mathbf{s})$$

Hyvärinen, Aapo, and Erkki Oja. "Independent component analysis: algorithms and applications." *Neural networks*, 2000.

# The identifiability problem

- For any $x_1, x_2$ we can always construct $y = g(x_1, x_2)$ independent of $x_1$ as

$$g(\xi_1, \xi_2) = P(x_2 \leq \xi_2 | x_1 = \xi_1)$$




Hyvärinen and Pajunen. "Nonlinear independent component analysis: Existence and uniqueness results." *Neural networks*, 1999.

# Solving non-linear ICA with supervision

- Consider the auxiliary supervision $\mathbf{u}$ s.t.

$$p(\mathbf{s} \mid \mathbf{u}) = \prod_{i=1}^{n} p_i(s_i \mid \mathbf{u})$$

- Train a NN to distinguish

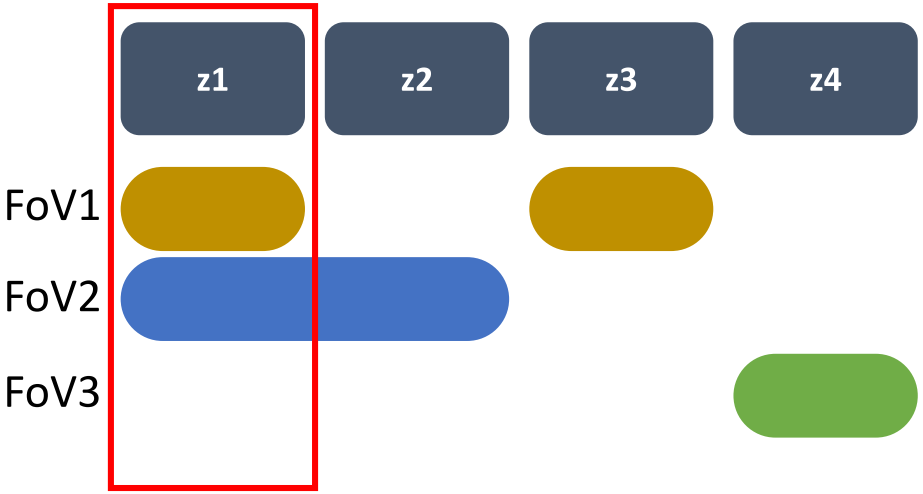
$$\tilde{\mathbf{x}} = (\mathbf{x}, \mathbf{u}) \quad vs. \quad \tilde{\mathbf{x}}^* = (\mathbf{x}, \mathbf{u}^*)$$

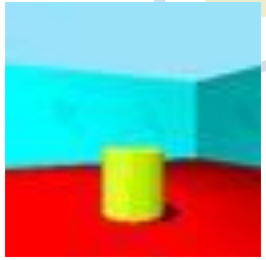
- Under strong variability assumption: identification

Hyvarinen, Aapo, Hiroaki Sasaki, and Richard Turner. "Nonlinear ICA using auxiliary variables and generalized contrastive learning." *AISTATS*, 2019.

# Problem Statement

- Disentanglement: low-dimensional sufficient representation with each coordinate (or a subset of coordinates) containing information about only one factor



- No established definition

Locatello, et al. "Challenging common assumptions in the unsupervised learning of disentangled representations." *ICML,* 2019.

# Disentanglement: Beta-VAE

- Latent variables model:

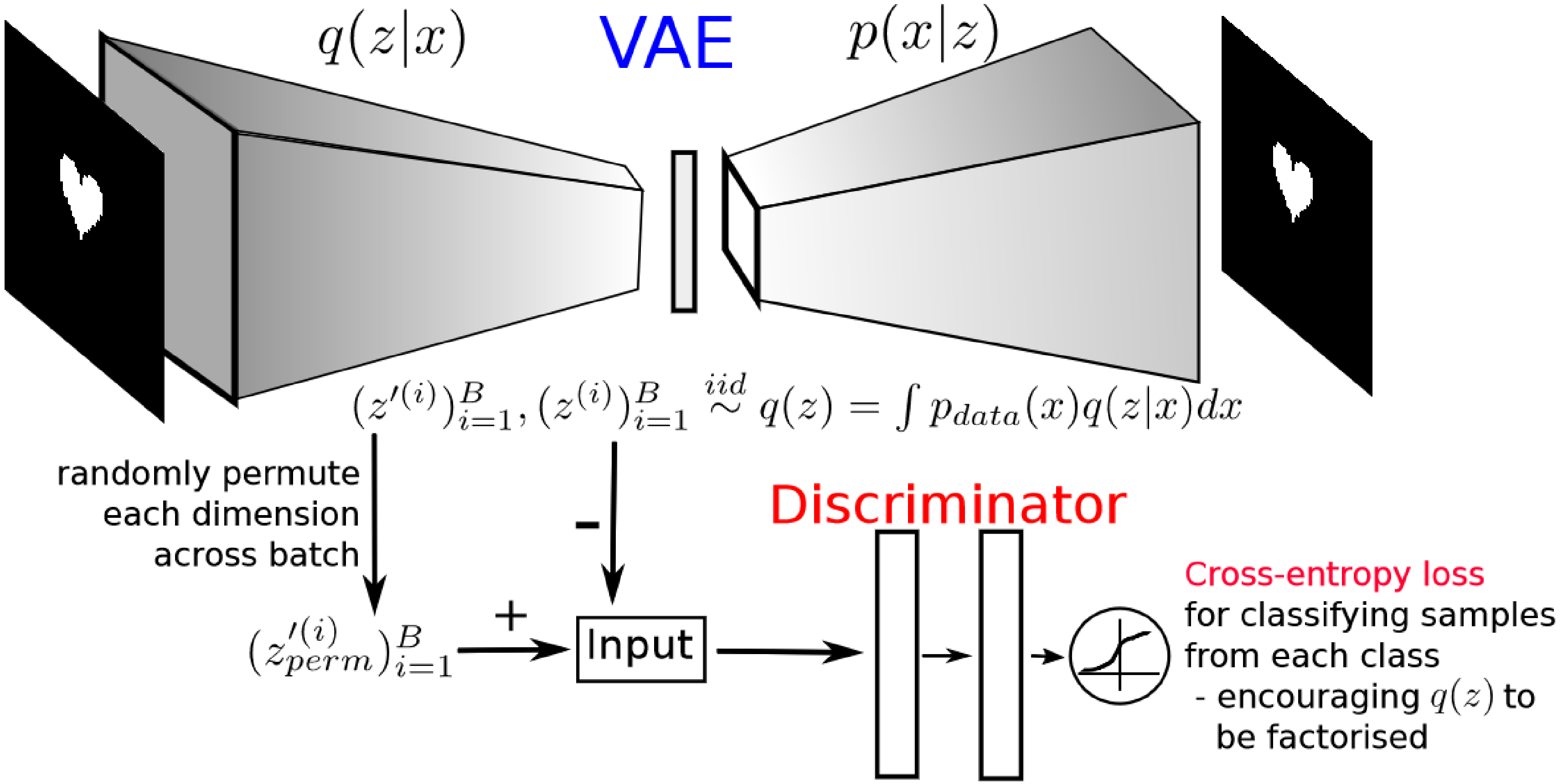
$$\mathbf{z}^{(i)} \sim p(\mathbf{z}), \mathbf{x}^{(i)} \sim p(\mathbf{x}|\mathbf{z}), \quad i = 1, \ldots, N$$

- Prior over latents: centered isotropic Gaussian $\mathcal{N}(0, \mathbf{I})$
- Reconstruction task with the Gaussian prior as regularization:

$$\max_{\phi,\theta} \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - \beta D_{KL}[q_\phi(\mathbf{z} \mid \mathbf{x}) \mid\mid \mathcal{N}(0, \mathbf{I})]$$

Higgins, Irina, et al. "beta-vae: Learning basic visual concepts with a constrained variational framework." *ICLR*, 2016.

# Disentanglement: factorisation

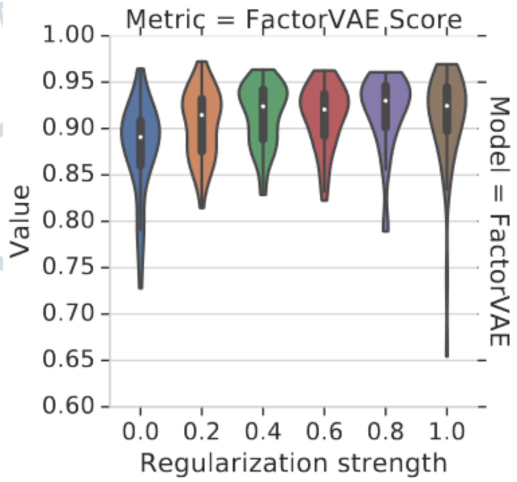



FactorVAE objective = VAE objective - $\gamma \times$ cross-entropy loss

Kim, Hyunjik, and Andriy Mnih. "Disentangling by Factorising." *ICML*, 2018.

# Challenging Common Assumptions

- Infinite family of entangled functions with same marginal distribution
- Critical unsupervised model selection:
  - relevant randomness
  - do not correlate with supervised metrics
  - Cannot transfer hyperparams




Dataset = Shapes3D

| | (A) | (B) | (C) | (D) | (E) | (F) |
|---|---|---|---|---|---|---|
| Reconstruction | -30 | -4 | 59 | 22 | -21 | 27 |
| TC (sampled) | 1 | 5 | -11 | -8 | -11 | -2 |
| KL | -14 | -1 | -38 | -31 | -11 | -29 |
| ELBO | -38 | -9 | 48 | 9 | -25 | 15 |

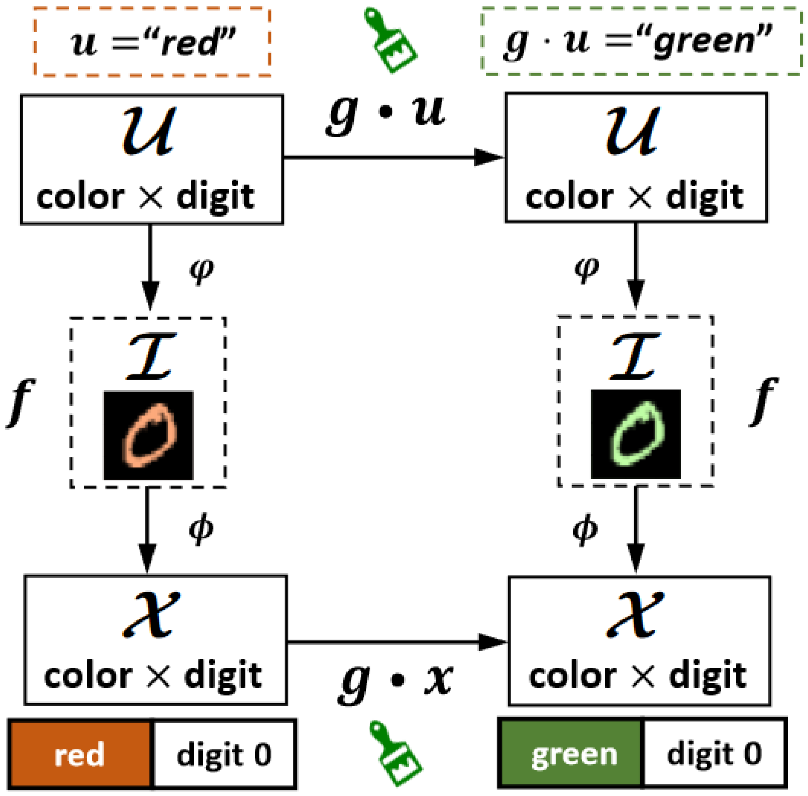
Metric = DCI Disentanglement

| | (I) | (II) | (III) | (IV) | (V) | (VI) | VII |
|---|---|---|---|---|---|---|---|
| dSprites (I) | 100 | 95 | 65 | 65 | 34 | 64 | 46 |
| Color-dSprites (II) | 95 | 100 | 61 | 60 | 21 | 63 | 47 |
| Noisy-dSprites (III) | 65 | 61 | 100 | 68 | 17 | 64 | 59 |
| Scream-dSprites (IV) | 65 | 60 | 68 | 100 | 36 | 93 | 69 |
| SmallNORB (V) | 34 | 21 | 17 | 36 | 100 | 21 | -9 |
| Cars3D (VI) | 64 | 63 | 64 | 93 | 21 | 100 | 85 |
| Shapes3D (VII) | 46 | 47 | 59 | 69 | -9 | 85 | 100 |

Locatello, Francesco, et al. "Challenging common assumptions in the unsupervised learning of disentangled representations." *ICML*, 2019.

# Group-theory approach

- Consider ground truth FoVs and inferred latents
- Let $\mathcal{G}$ be a group acting on $\mathcal{U}$, $g \cdot u : \mathcal{G} \times \mathcal{U} \to \mathcal{U}$
- Equivariance: $g \cdot f(u) = f(g \cdot u)$
  e.g., change the color semantic is equivalent to change the associated feature
- Decomposable: $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m \mid g_i \cdot x_j \neq x_j \iff i = j$
  e.g., changing the color semantic does not effect the shape




Wang, Tan, et al. "Self-Supervised Learning Disentangled Group Representation as Feature." *NeurIPS*, 2021.

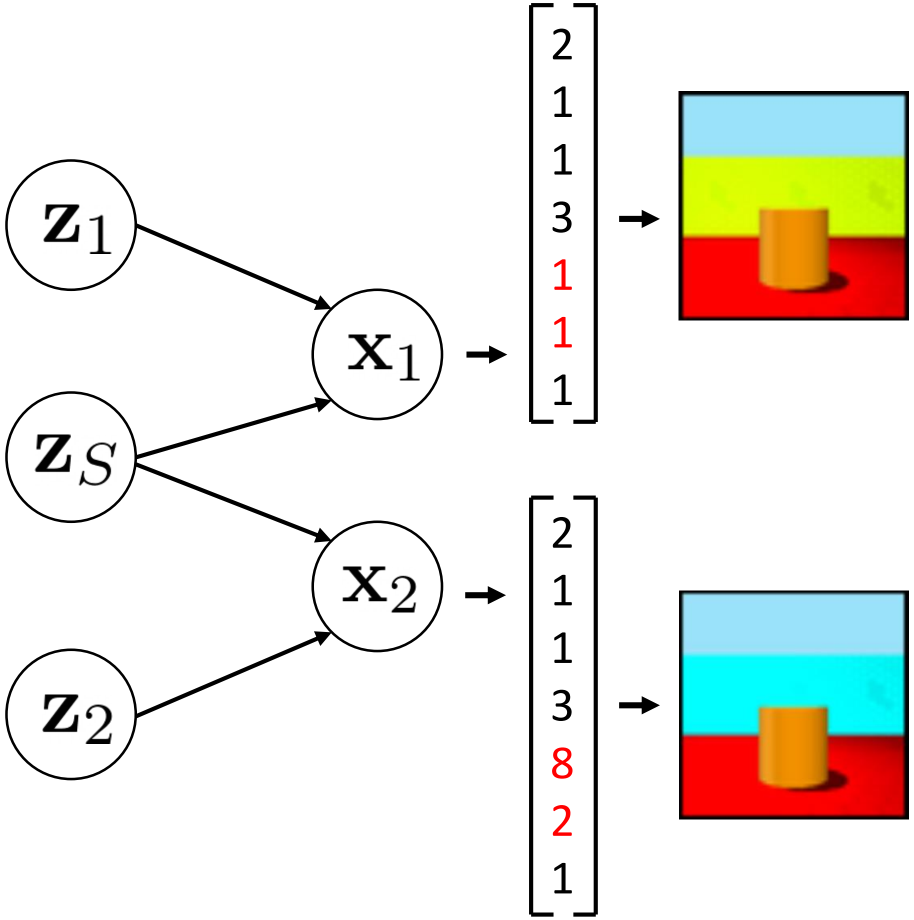
# Disentanglement with few labels

- Can we disentangle with a few labels?
  1. Unsupervised training with few labels validation
  2. Semi-supervised training (regularization) with validation
  3. Fully supervised training

- First two approaches are robust to coarse, noisy and partial labels

Locatello, Francesco, et al. "Disentangling Factors of Variations Using Few Labels." *ICLR,* 2019.

# Causality for disentanglement




Estimate $\hat{S}$ as components with lowest

$$D_{KL}\left(q_\phi\left(\hat{z}_i|\mathbf{x}_1\right)\|q_\phi\left(\hat{z}_i|\mathbf{x}_2\right)\right)$$

set the posterior to be:

$$\tilde{q}_\phi\left(\hat{z}_i|\mathbf{x}_1\right) = a\left(q_\phi\left(\hat{z}_i|\mathbf{x}_1\right), q_\phi\left(\hat{z}_i|\mathbf{x}_2\right)\right) \quad i \in \hat{S},$$

$$\tilde{q}_\phi\left(\hat{z}_i|\mathbf{x}_1\right) = q_\phi\left(\hat{z}_i|\mathbf{x}_1\right) \quad \text{otherwise}$$

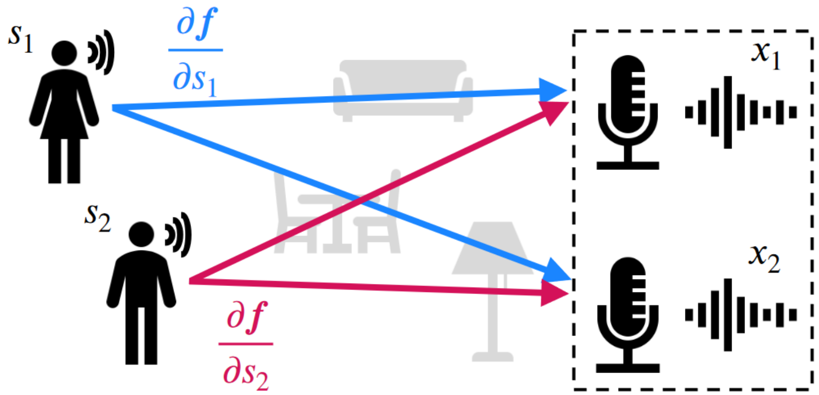
Locatello, Francesco, et al. "Weakly-supervised disentanglement without compromises." *ICML* (2020).

# Causality for disentanglement

- Constraint nonlinear ICA $\mathbf{x} = \mathbf{f}(\mathbf{x})$
  e.g., speakers positions w.r.t. to microphones not fine-tuned

- Less ambiguities
- ICM principle inspiration: $\mathbf{f}$ as independent mechanisms, each influenced by a factor

$$\log |\mathbf{J_f}(\mathbf{s})| = \sum_{i=1}^{n} \log \left\| \frac{\partial \mathbf{f}}{\partial s_i}(\mathbf{s}) \right\|$$




Gresele, Luigi, et al. "Independent mechanism analysis, a new concept?." *NeurIPS*, 2021.

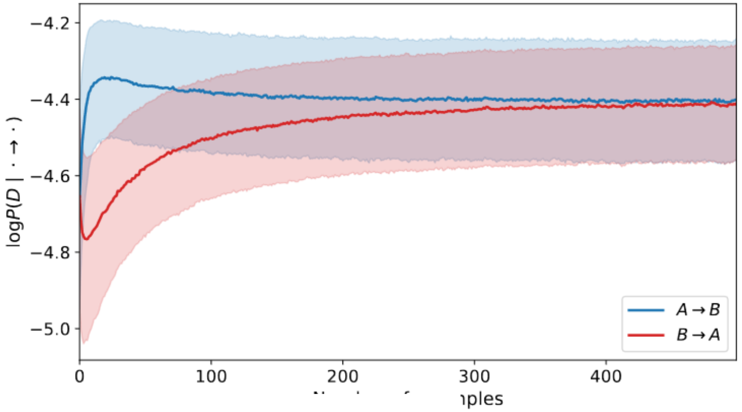
# Disentanglement and causality (bivariate case)

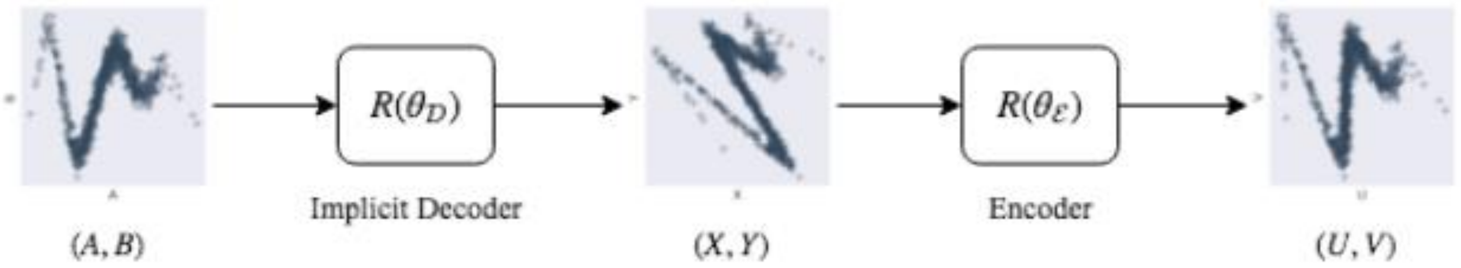
- Observational data coming from different interventional settings $\epsilon = 1, \ldots, E$
- In a SCM $X_i = f_i(PA_i, U_i)$, exogenous noises as independent component to unmix
- As in nonlinear ICA, train a NN to predict the interventional setting
- Independence tests for causal direction

Monti, Ricardo Pio, et al. "Causal Discovery with General Non-Linear Relationships Using Non-Linear Independent Component Analysis." *UAI, 2019*.

# Disentanglement and Causality

- Sparse mechanisms assumption
- The correct parameterization adapts faster to interventional data
- Reverse the transformation of an implicit decoder




Bengio, Yoshua, et al. "A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms." *ICLR,* 2019.

# Any questions?

*Part 1*

## Causality 101

**From statistical to causal models**

**The structural causal model**

**Identifiability problem**

*Part 2*

## High dimensional data

**Linear and non-l ar ICA**

**Disent glement**

**The identif lity problem**

**Cross-pollination: causality and disentanglement**

*Part 3*

## Causal signals in Visual data

**Causal signal for images**

**Causal visual datasets**

# Agenda

*Part 1*

## Causality 101

From statistical to causal models

The structural causal model

Identifiability problem

*Part 2*

## High dimensional data

Linear and non-linear ICA

Disentanglement

The identifiability problem

Cross-pollination: causality and disentanglement

*Part 3*

## Causal signals in Visual data

Causal signal for images

Causal visual datasets

# Causal signals in Images?

- Causal dispositions: the presence of an object causes the presence of certain objects
  - e.g., the presence of cars causes the presence of wheels

- PASCAL VOC 2012 classification dataset (20 classes)
  - airplane, bicycle, bird, boat, bottle, bus, car, ...

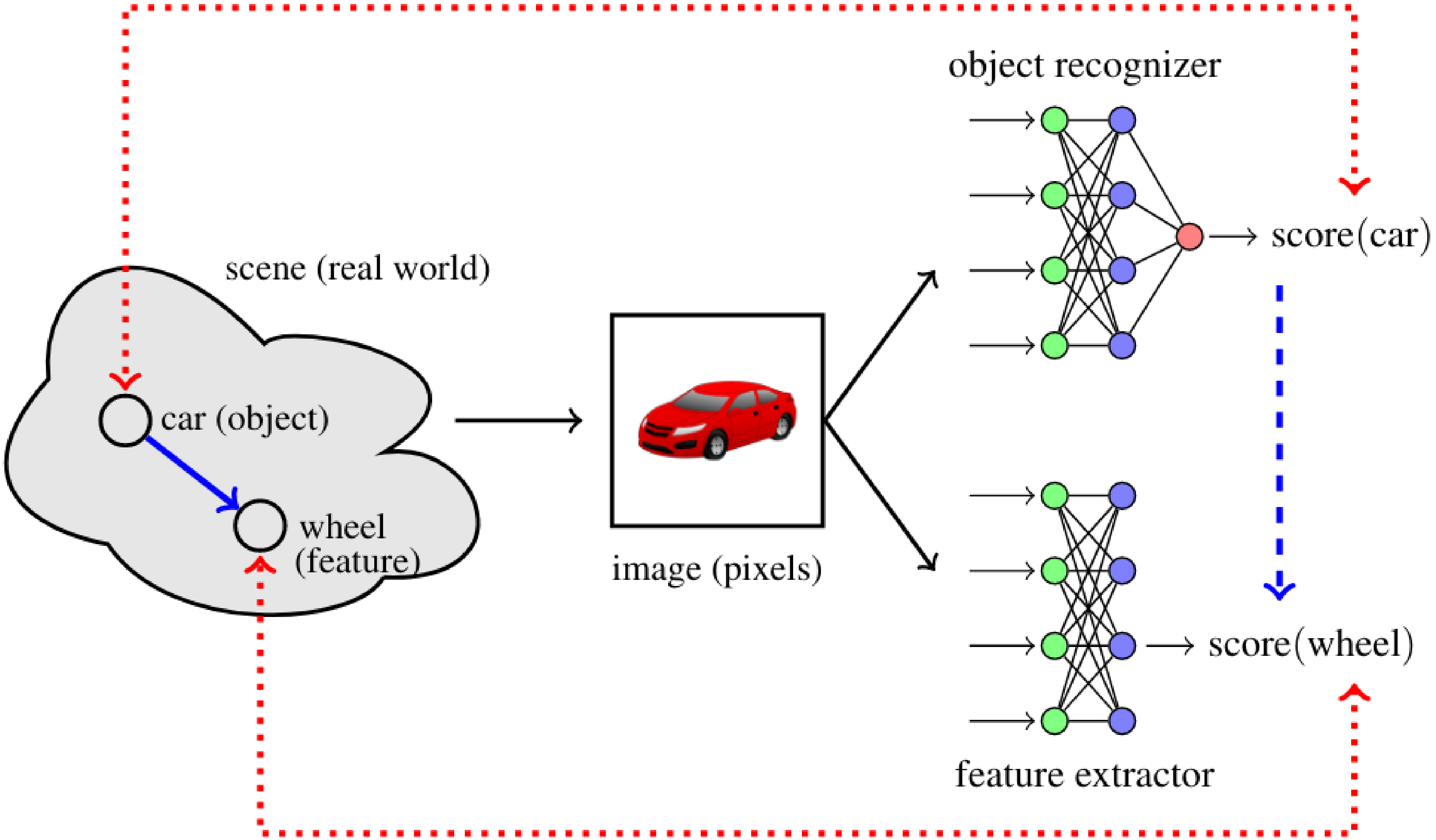
Lopez-Paz, David, et al. "Discovering causal signals in images." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

# Features' properties

- Causal vs Anti-causal:
  - Causal: which cause the presence of the object
  - Anti-causal: caused by the presence of the object
- Object vs Context:
  - Object: within the bounding box
  - Context: outside the bounding box

Lopez-Paz, David, et al. "Discovering causal signals in images." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

# Causal signals in Images?




Lopez-Paz, David, et al. "Discovering causal signals in images." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

# Causal vs Anticausal Features

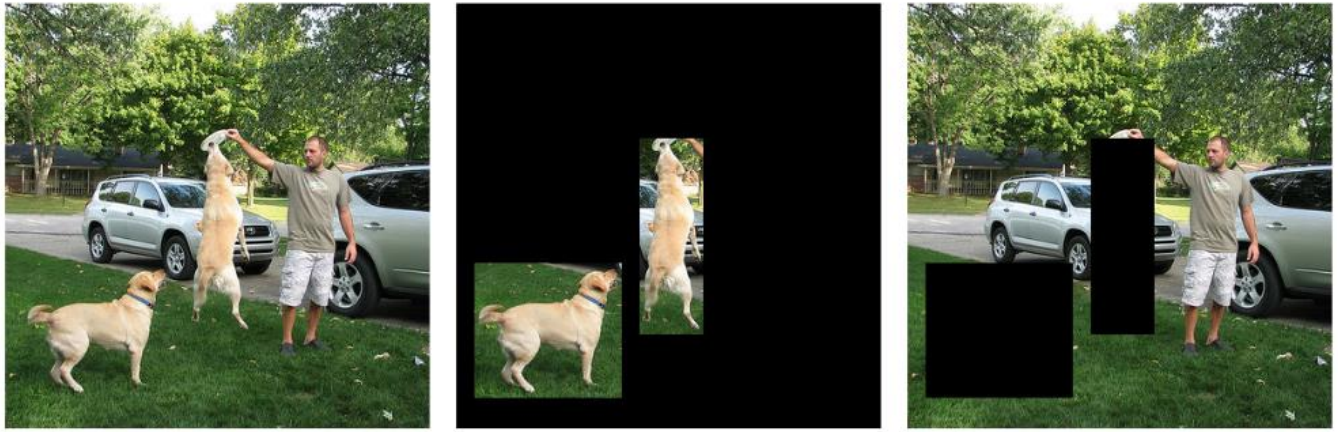
- Train a model to predict the causal direction between X and Y on synthetic data (X, Y)
- Get features from a pre-trained feature extractor
- Train a classifier on top of the feature extractor
- Predict causal direction on (feature, object logit)

- Select top 1% causal and anti-causal features

Lopez-Paz, David, et al. "Discovering causal signals in images." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

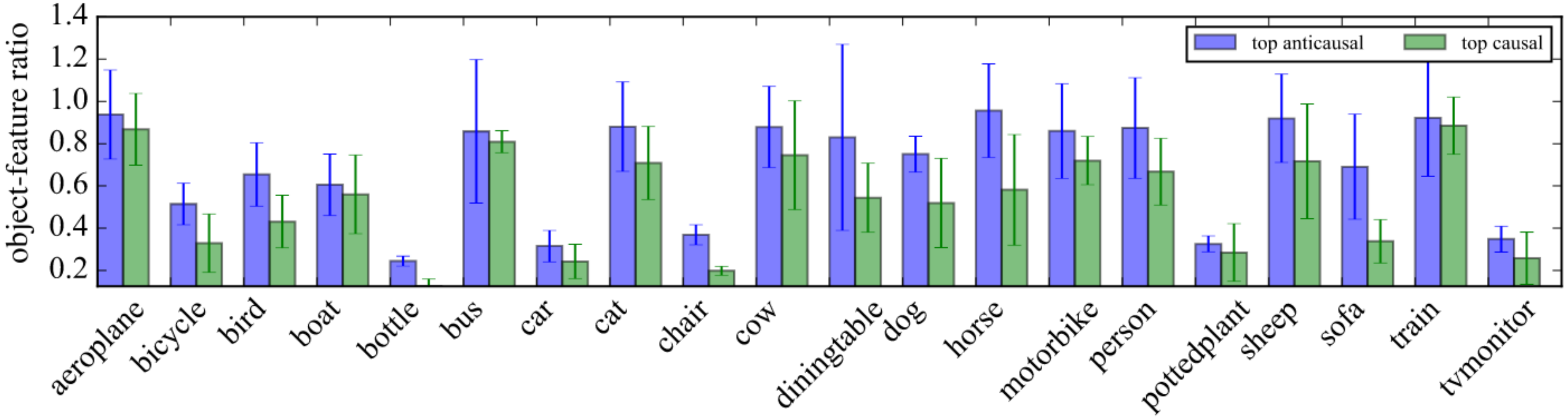
# Object vs Context features

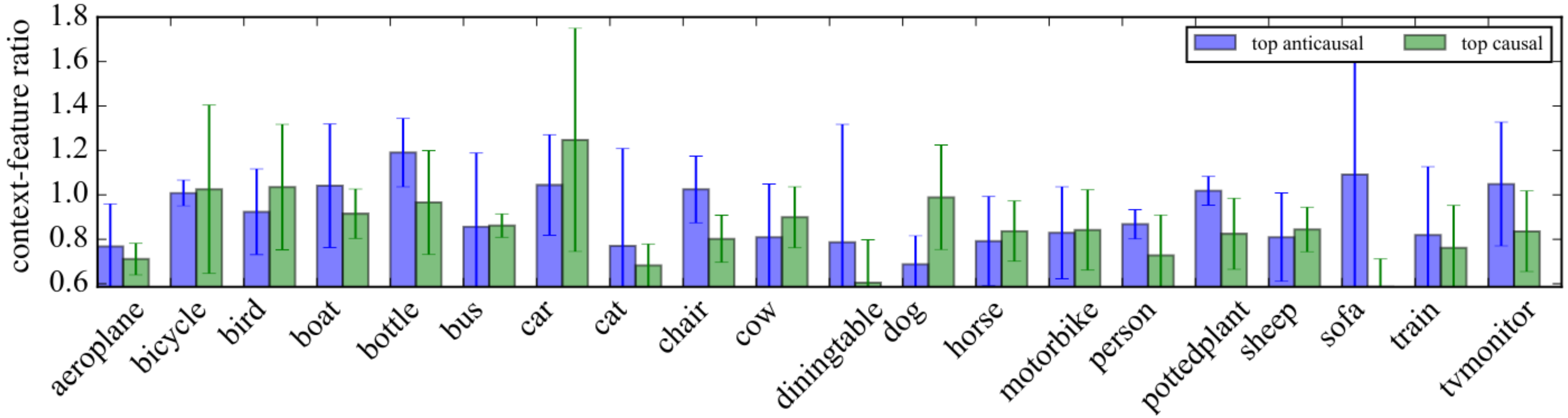
- Features from pre-trained models
  - Object features react violently to black out of bounding boxes
  - Context features react violently to black out of context



Lopez-Paz, David, et al. "Discovering causal signals in images." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

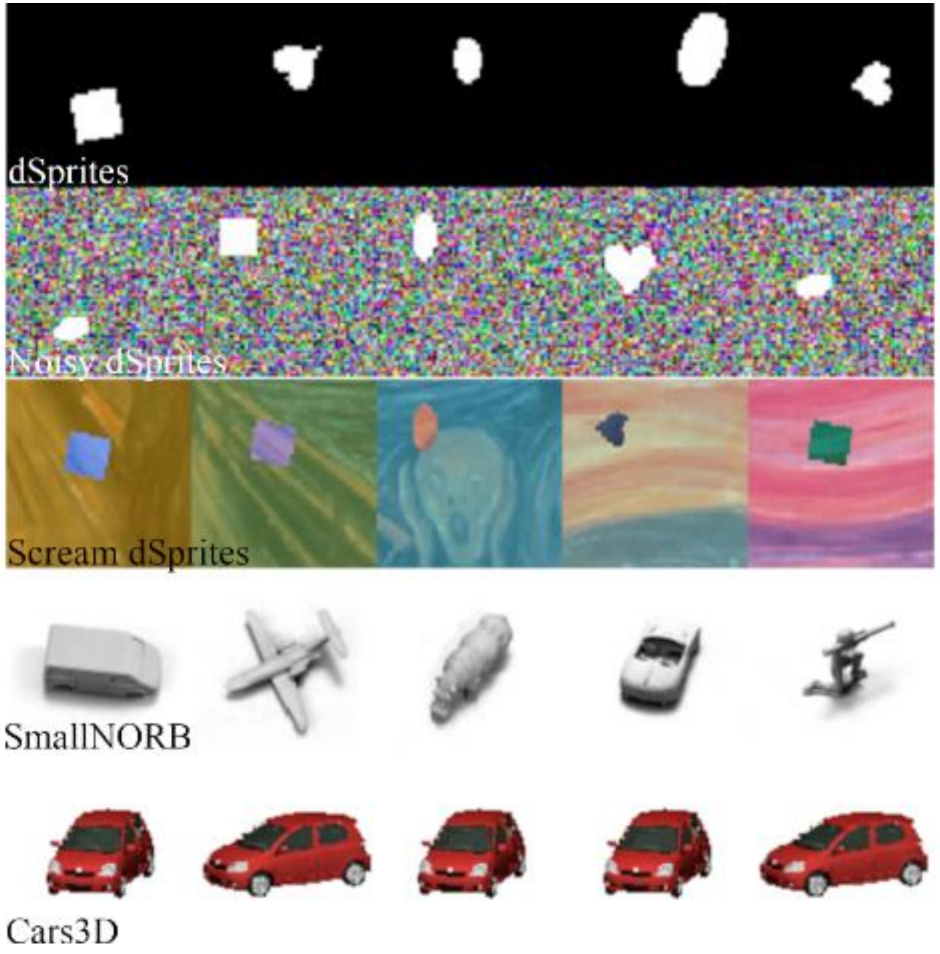
# Observed correlations

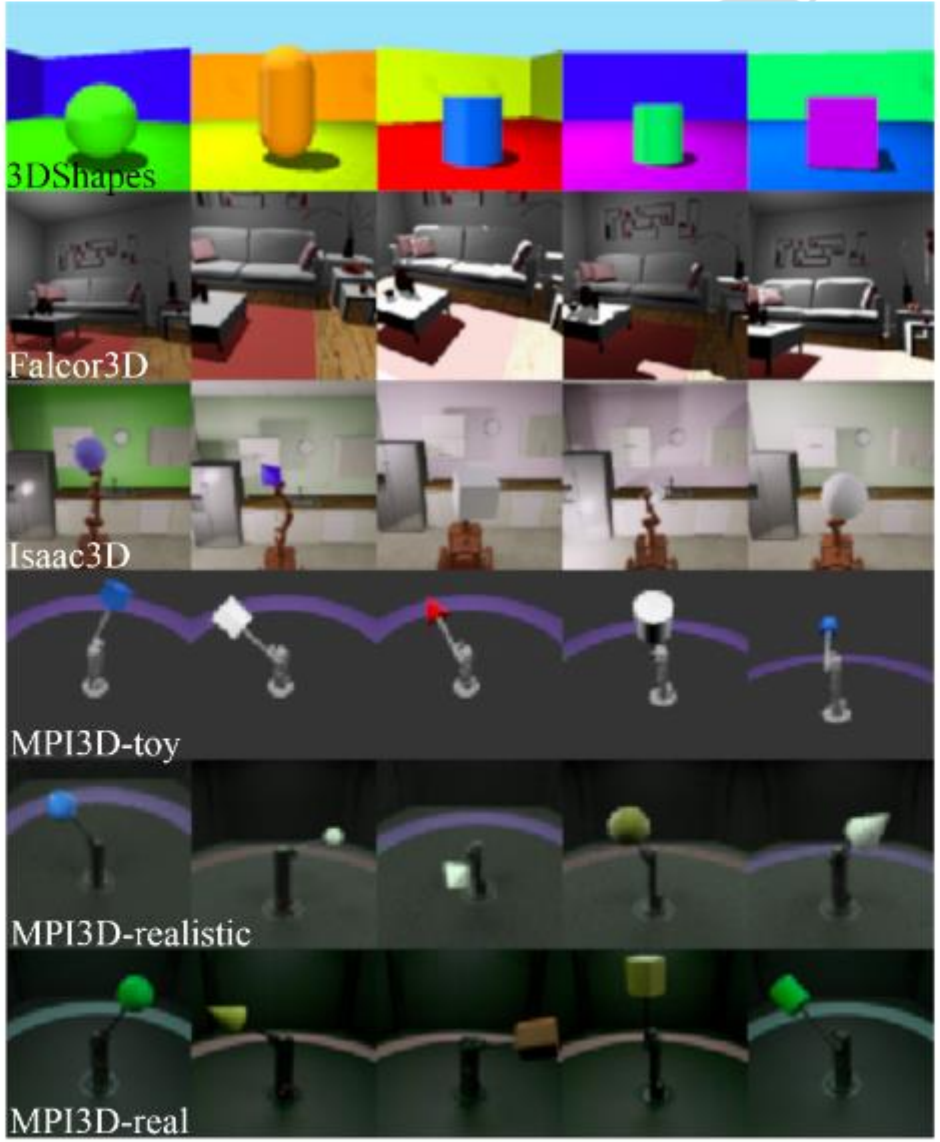



Lopez-Paz, David, et al. "Discovering causal signals in images." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

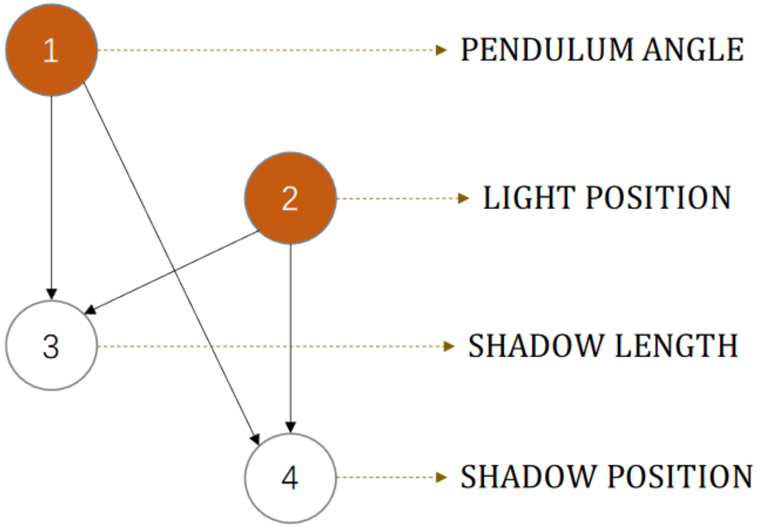
# Disentanglement data

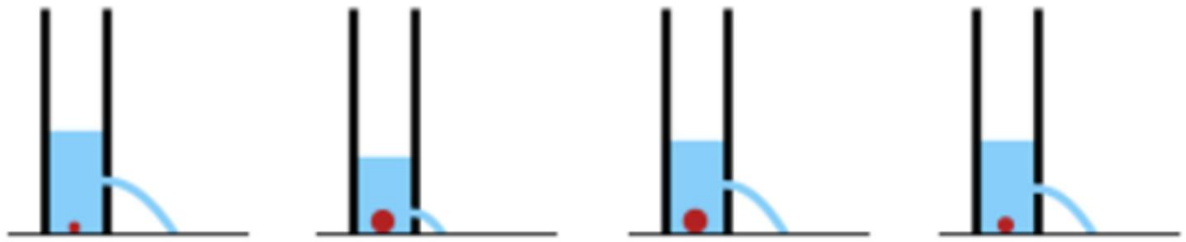


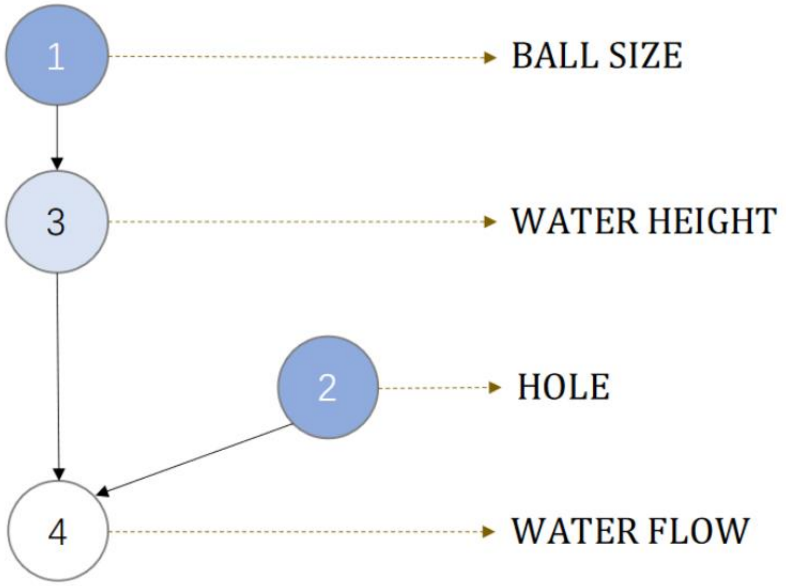




Reddy, Abbavaram Gowtham, L. Benin Godfrey, and Vineeth N. Balasubramanian. "On Causally Disentangled Representations." *AAAI,* 2021.

# Image datasets

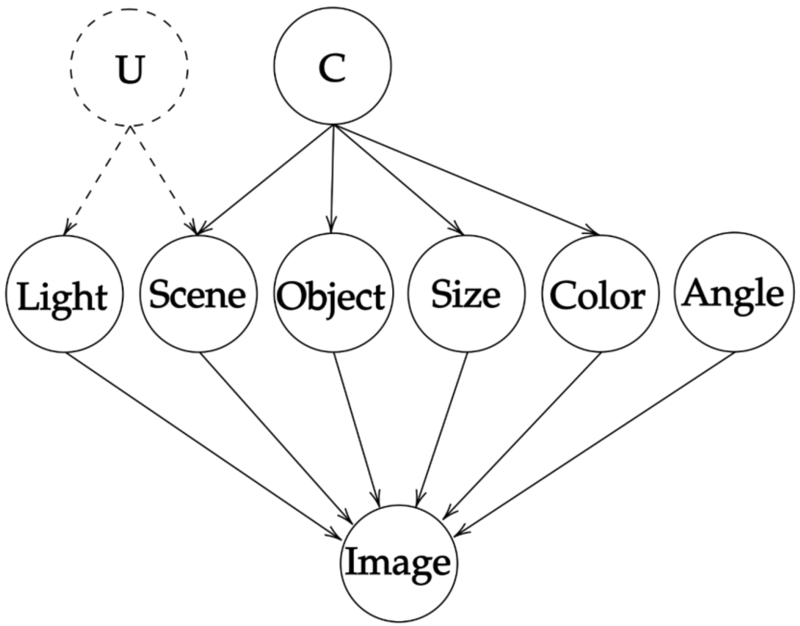


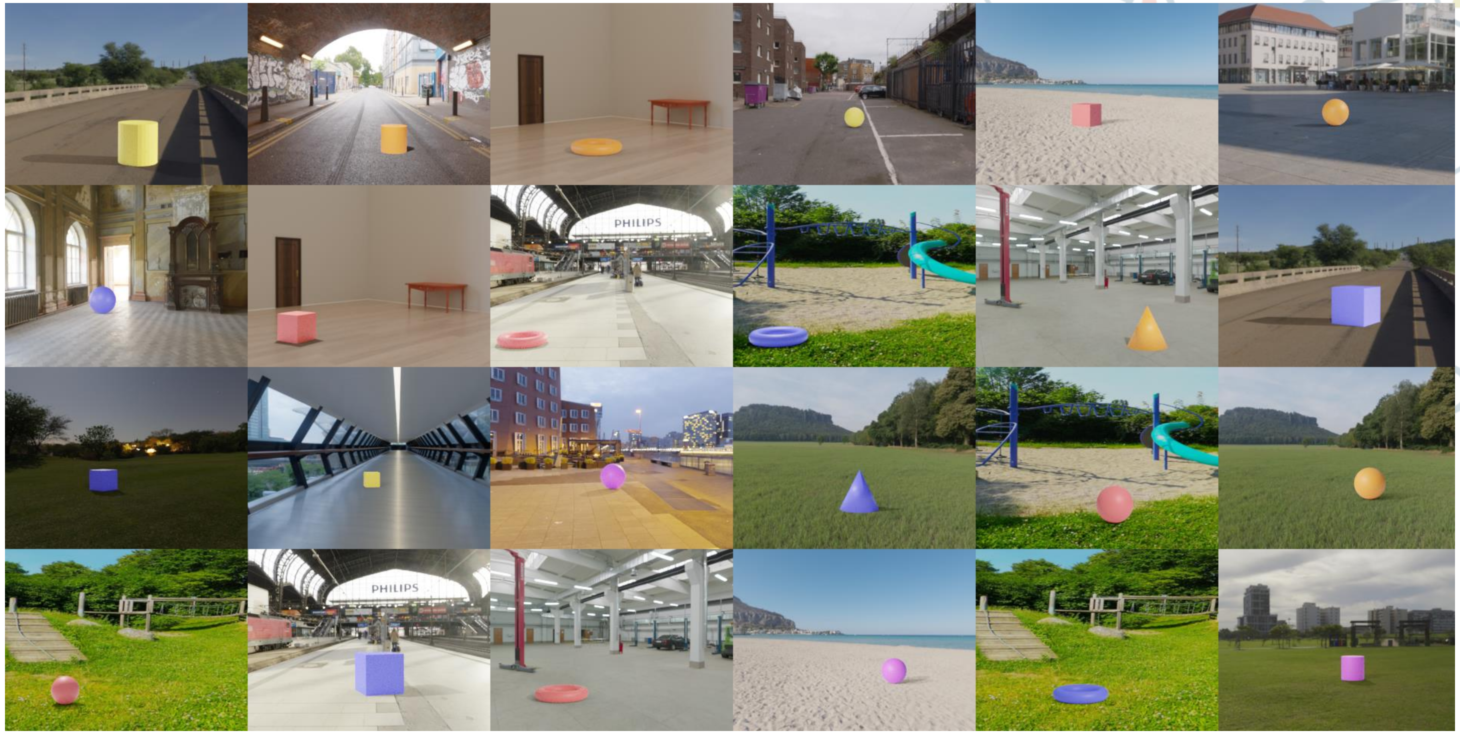






Yang, Mengyue, et al. "CausalVAE: Disentangled representation learning via neural structural causal models." *CVPR*, 2021.

# Image datasets




Reddy, Abbavaram Gowtham, L. Benin Godfrey, and Vineeth N. Balasubramanian. "On Causally Disentangled Representations." *AAAI,* 2021.

# Long Story Short

- Causal signals leave traces in images
- Toyish causal visual datasets

# Any questions?

*Part 1*

## Causality 101

**From statistical to causal models**

**The structural causal model**

**Identifiability problem**

*Part 2*

## High dimensional data

**Linear and non-linear ICA**

**Disentanglement**

**The identifiability problem**

**Cross-pollination: causality and disentanglement**

*Part 3*

## Causal signals in Visual data

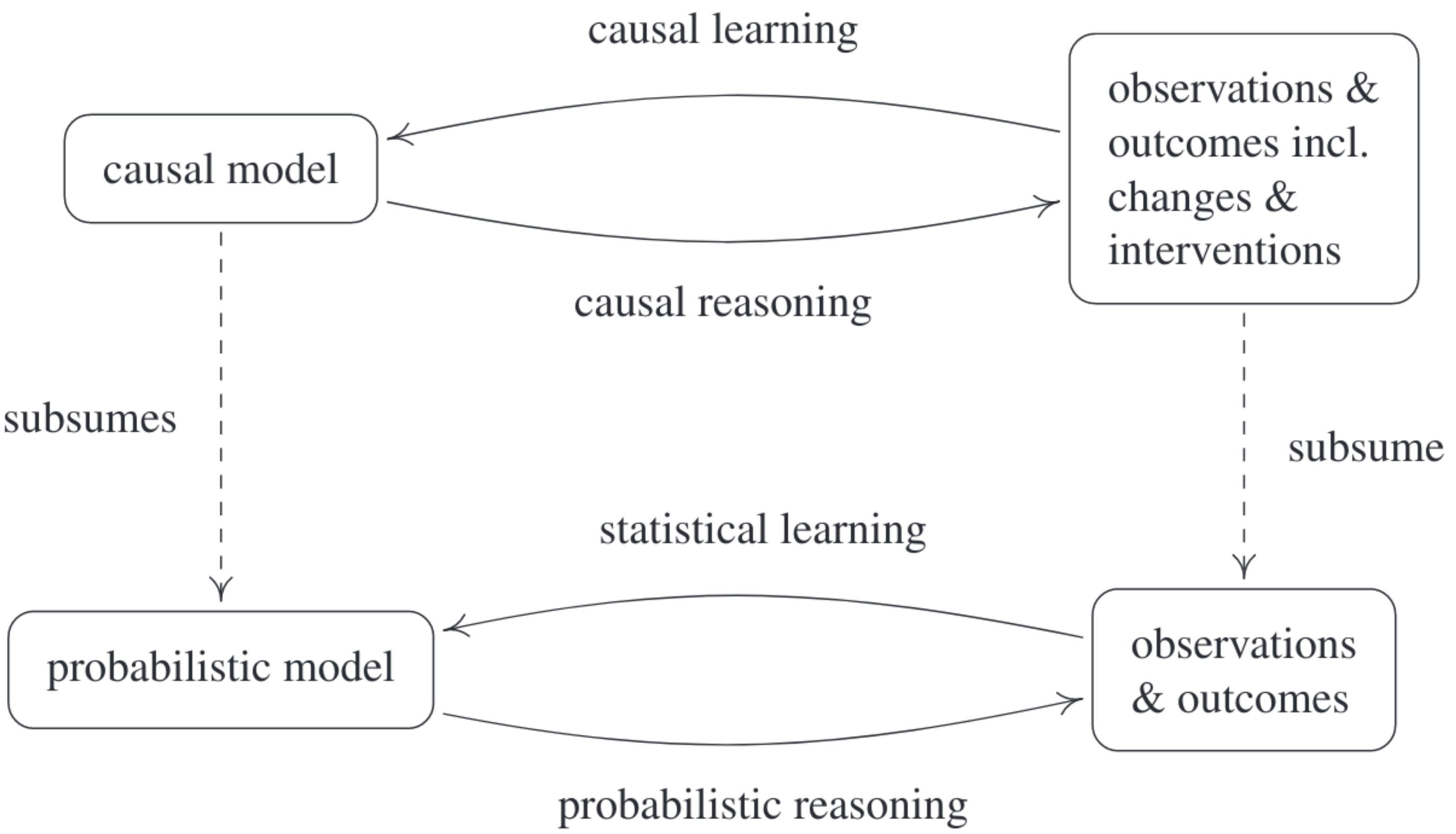
?

**Causal s[illegible]al for images**

**Causal v[illegible]al datasets**


- I am sorry, no pizza at the canteen today

# Causality vs probabilities




Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

# *d*-separation

- Definition: In a DAG $\mathcal{G}$, a path between nodes $i_1$ and $i_m$ is blocked by a set $\mathbf{S}$ (with neither $i_1$ and $i_m$ in it) if there exists $i_k$ such that one of this holds:
  - $i_k \in \mathbf{S}$ and:

    $i_{k-1} \rightarrow i_k \rightarrow i_{k+1}$ or,

    $i_{k-1} \leftarrow i_k \leftarrow i_{k+1}$ or,

    $i_{k-1} \leftarrow i_k \rightarrow i_{k+1}$

  - $(\{i_k\} \cup \mathbf{DE}_{i_k}) \cap \mathbf{S}$ and:

    $i_{k-1} \rightarrow i_k \leftarrow i_{k+1}$

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

# Markov property

- Given a DAG $\mathcal{G}$ and a joint distribution $P_{\mathbf{X}}$
  - Global Markov property:

$$\mathbf{A} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{B} \mid \mathbf{C} \Rightarrow \mathbf{A} \perp\!\!\!\perp \mathbf{B} \mid \mathbf{C}$$

  - Local markov property: if each variable is independent of its non-descendants given its parents
  - Markov factorization property:

$$p(\mathbf{x}) = p(x_1, \ldots, x_d) = \prod_{j=1}^{d} p(x_j \mid \mathbf{PA}_j^{\mathcal{G}})$$

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

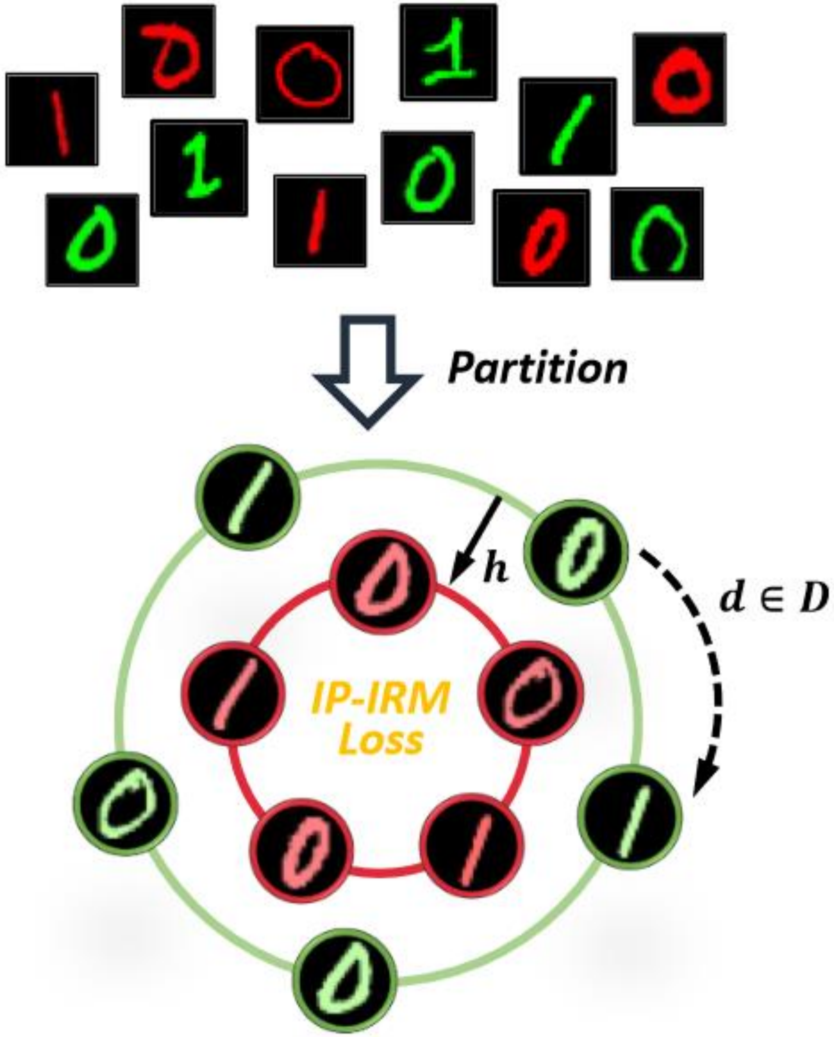
# Markov equivalence class

- $\mathcal{M}(\mathcal{G})$ set of distributions Markovian to the DAG $\mathcal{G}$
- $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$
- Markov equivalence class:

$$\{\mathcal{G}' \text{s.t.} \mathcal{M}(\mathcal{G}') = \mathcal{M}(\mathcal{G})\}$$

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

# Group-theory approach




Wang, Tan, et al. "Self-Supervised Learning Disentangled Group Representation as Feature." *NeurIPS*, 2021.

# Identifiability approaches

- Model class restriction: limit the complexity of structural functionals
  - Linear models with non-Gaussian additive noise
  - Nonlinear additive noise models

- Independence between cause and effect mechanism:
  - Information-geometric: check for zero covariance between structural functionals and cause
  - Trace method: the eigenvalues of functional mapping tune to input cause
  - Algorithmic independence with Kolmogorov complexity