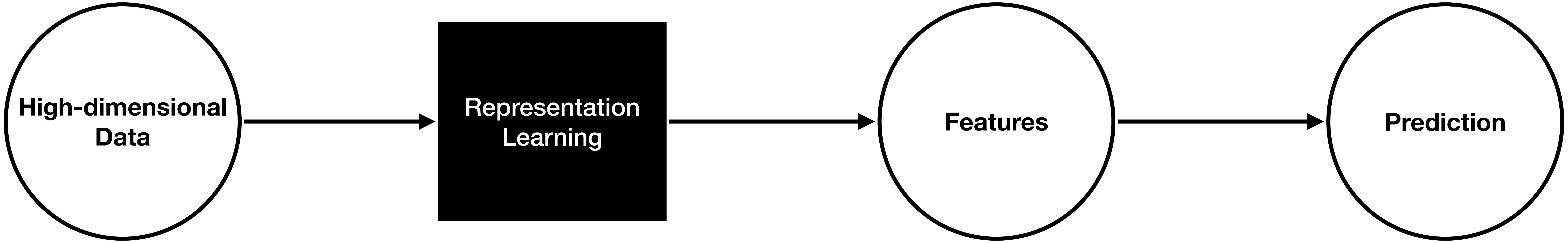
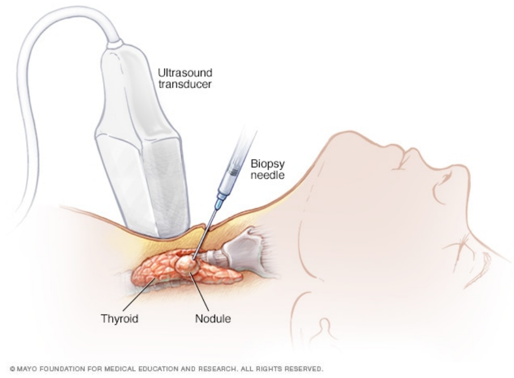
Representation Learning: A Causal Perspective

Yixin Wang

(Joint work with Michael Jordan)

Representation Learning



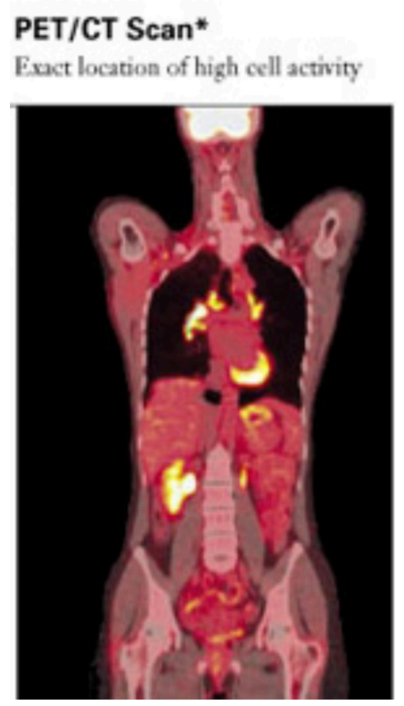


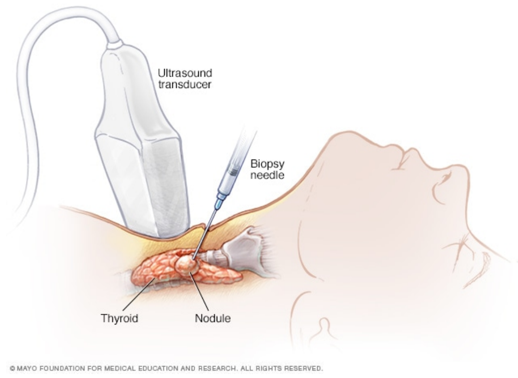
Patient	Swollen Lymph Nodes	Circulating Tumor Cell	Mass in Breast	Inflammatory Breast Cancer
1	0.3	0.8	0.4	1
2	0.6	0.2	0.5	0

Representation learning liberates us from manual feature engineering.
But it can often produce **spurious, inefficient, or entangled** representations in practice.
Today: Use causal inference for representation learning
Work with a single dataset; Do not leverage multiple environments or invariance or auxiliary labels.

Representation Learning

a.k.a. feature learning





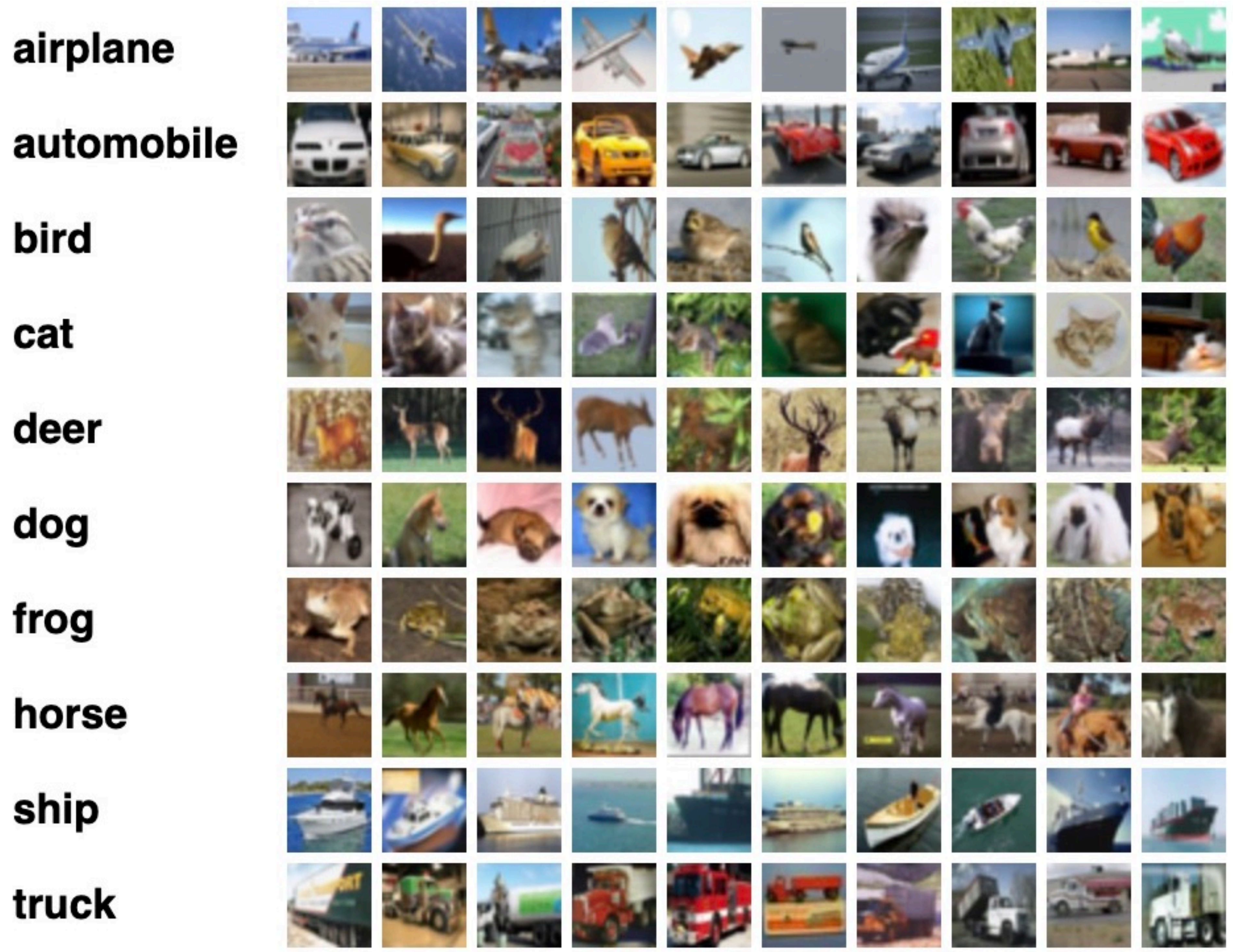
Patient	Swollen Lymph Nodes	Circulating Tumor Cell	Mass in Breast
1	0.3	0.8	0.4
2	0.6	0.2	0.5

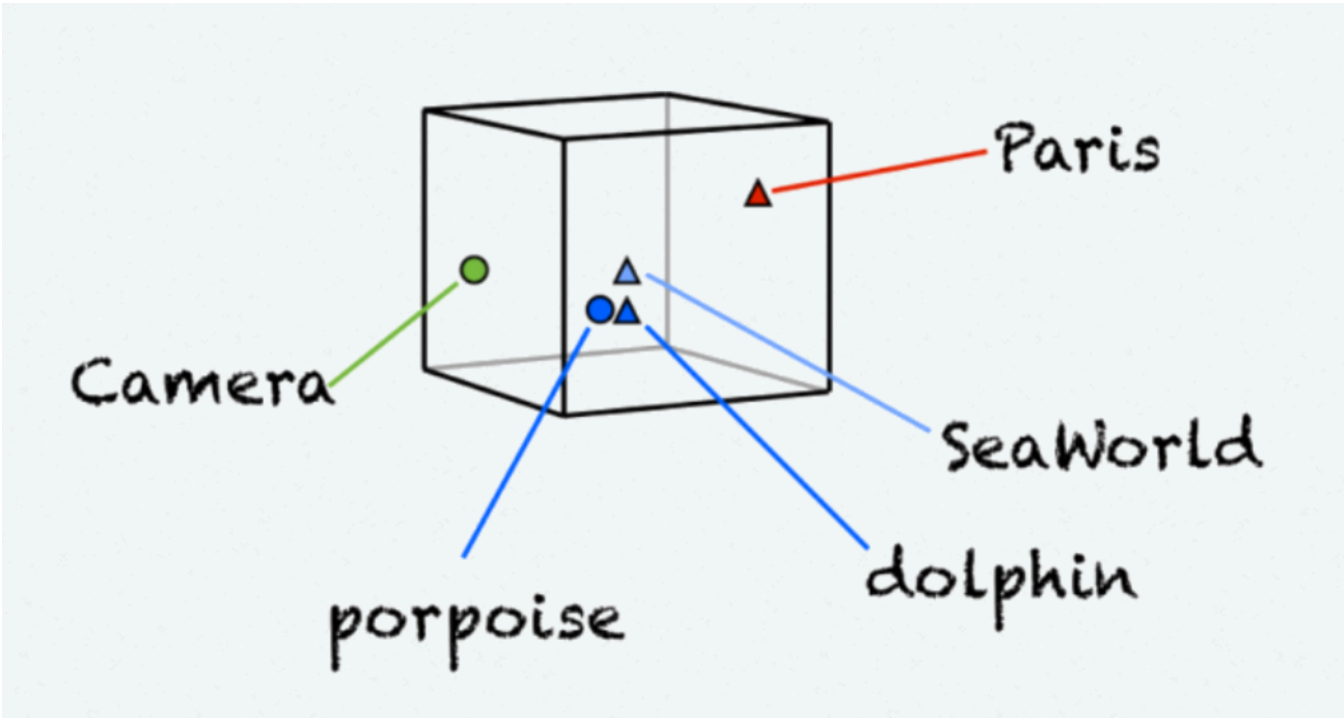
m-dimensional data point $\mathbf{X} = (X_1, \ldots, X_m) \in \mathbb{R}^m$

d-dimensional representation $\mathbf{Z} = (Z_1, \ldots, Z_d) \triangleq (f_1(\mathbf{X}, \ldots, f_d(\mathbf{X}))$

Goal: Find the representation function $f = (f_1, \ldots, f_d)$

Representation Learning



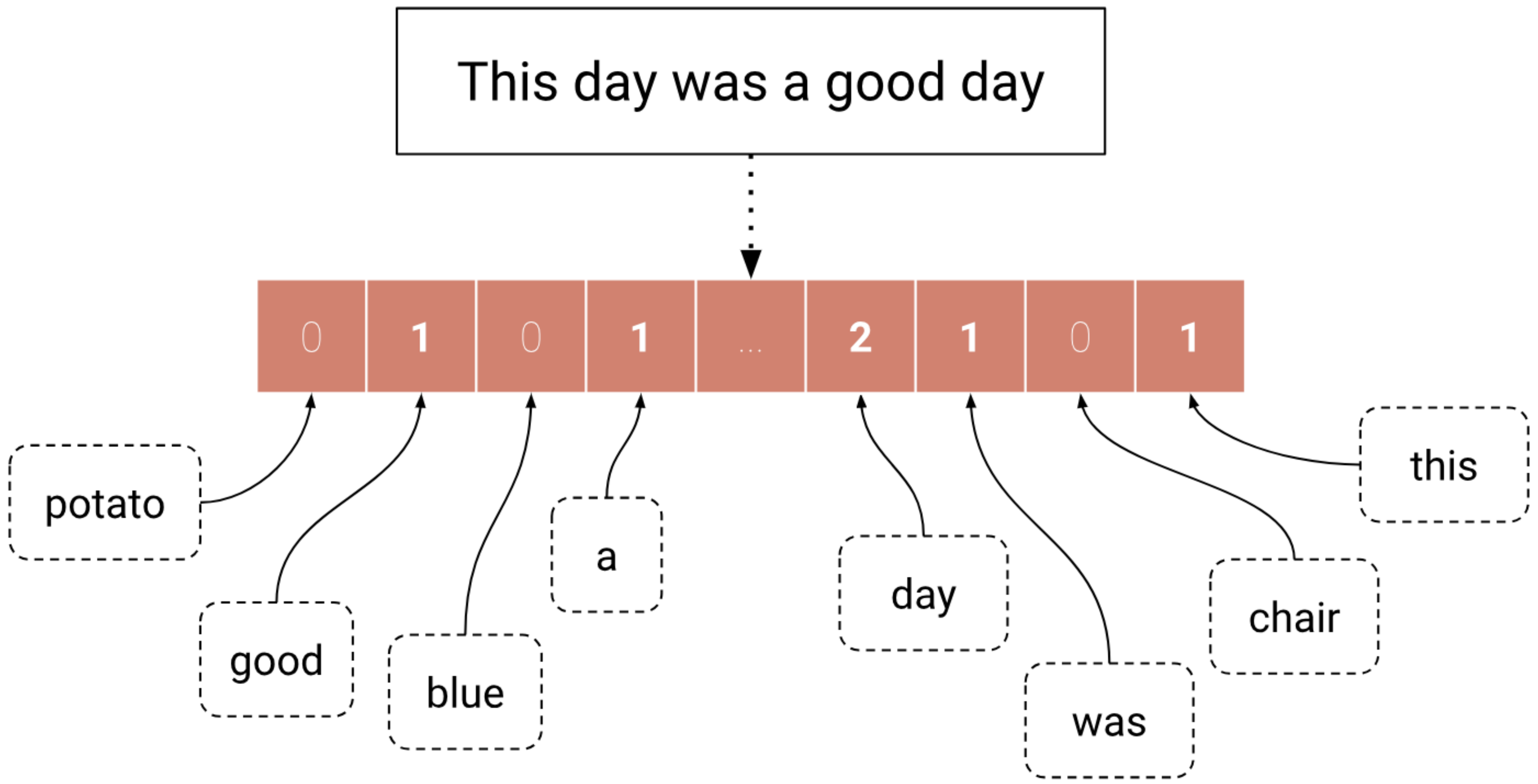


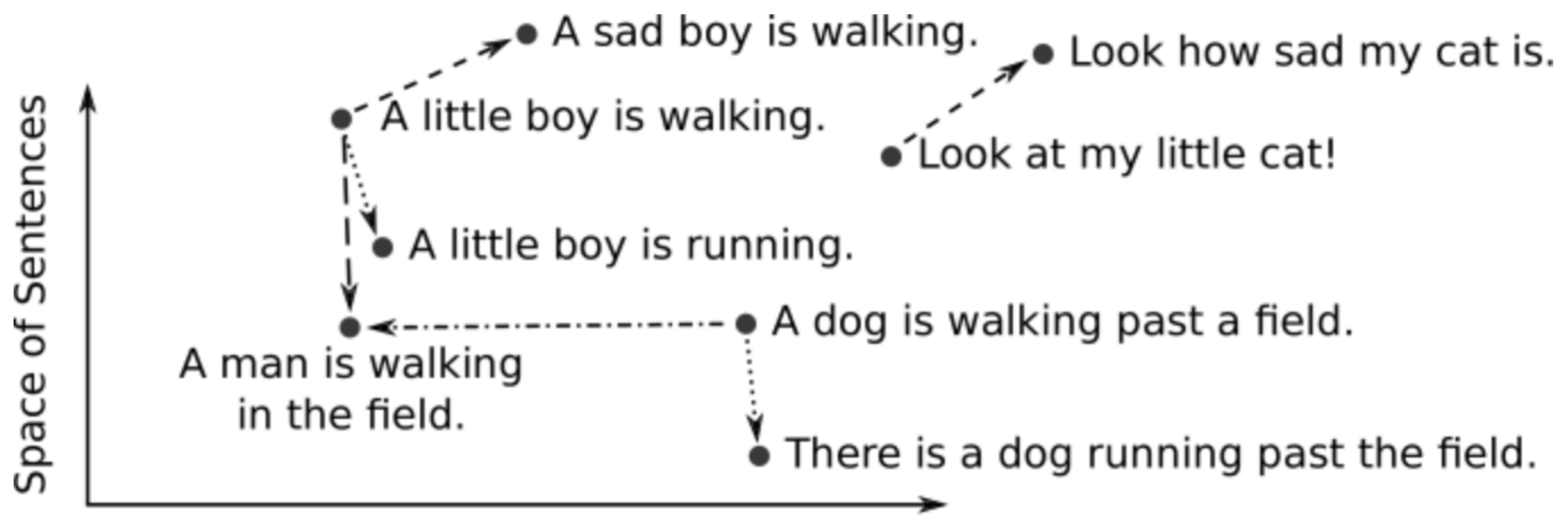
m-dimensional data point $\mathbf{X} = (X_1, \ldots, X_m) \in \mathbb{R}^m$

d-dimensional representation $\mathbf{Z} = (Z_1, \ldots, Z_d) \triangleq (f_1(\mathbf{X}, \ldots, f_d(\mathbf{X}))$

Goal: Find the representation function $f = (f_1, \ldots, f_d)$

Representation Learning





m-dimensional data point $\mathbf{X} = (X_1, \ldots, X_m) \in \mathbb{R}^m$

d-dimensional representation $\mathbf{Z} = (Z_1, \ldots, Z_d) \triangleq (f_1(\mathbf{X}, \ldots, f_d(\mathbf{X}))$

Goal: Find the representation function $f = (f_1, \ldots, f_d)$

Why might naive representation learning produce spurious features?

Learning Representations for Dogs

Label=1

Label=0

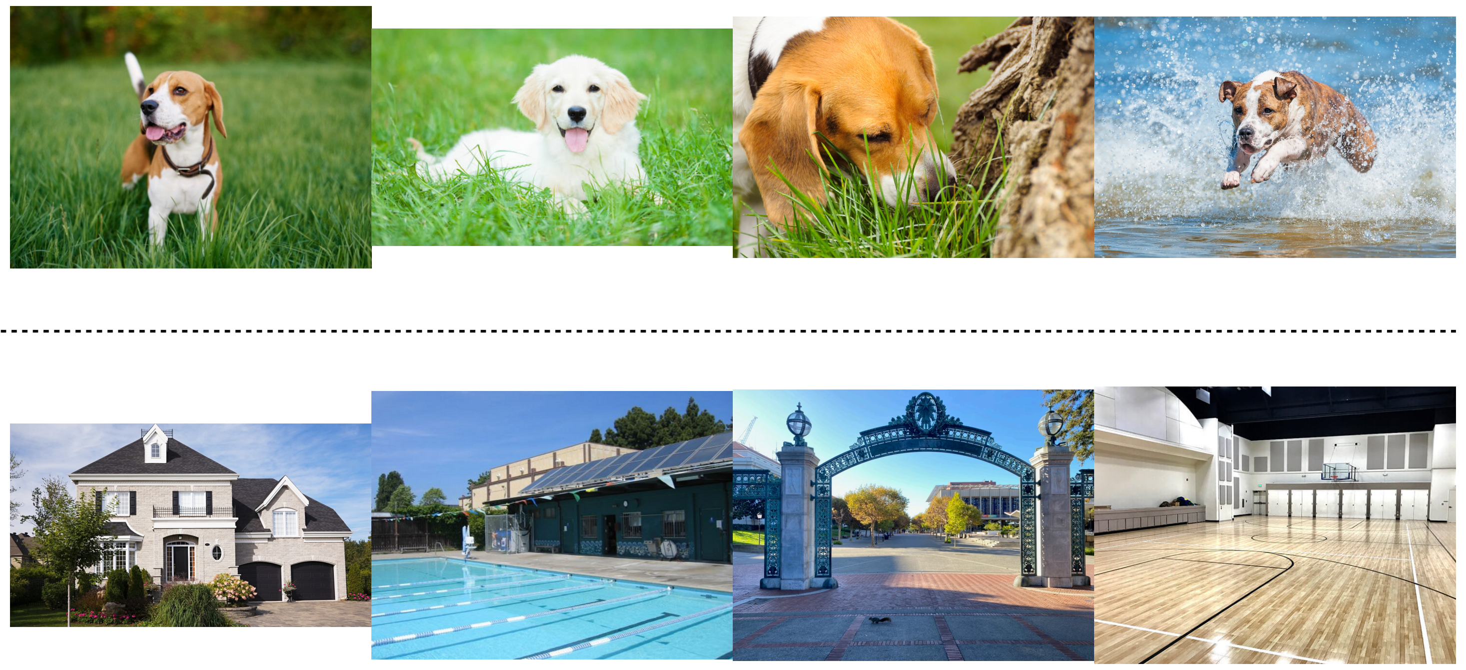
Given n pairs of **images** $\mathbf{X}_i = (X_{i1}, \ldots, X_{im})$ and “dog” **labels** Y_i (if a dog is in the image),

find $f : \mathcal{X}^m \to \mathbb{R}^d$ s.t. $\mathbf{Z}_i = f(\mathbf{X}_i)$ is a representation that captures important features.

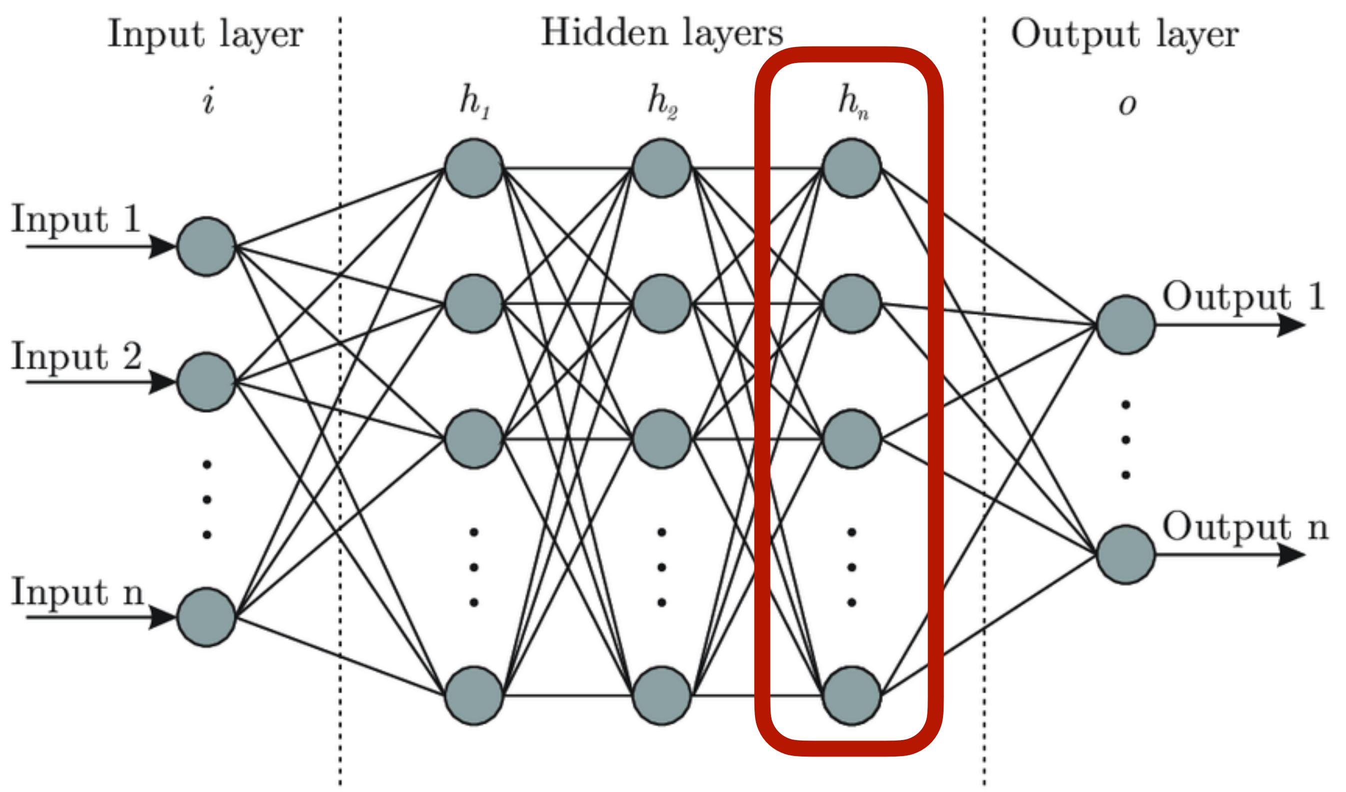
Learning Representations for Dogs

Label=1

Label=0



Training set



Representation

Test set

Naive solution: Fit a neural network from the images $\mathbf{X}_i$ to the "dog" label Y_i;

Take the last layer to be the representation $f(\mathbf{X}_i)$.

The predictions are awfully wrong…

Predicted label=0

Predicted label=1

Predicted label=0

Predicted label=1

- The learned representation seems to pick up the **“whether grass is present in the image”** feature.
- It is a **spurious feature.** We pick up the grass feature even if the prediction target is the dog label.
- **It is not a neural network training failure**; the predictive accuracy is high in the holdout validation set.

What went wrong?

Label=1

Label=0

Training set

Predicted label=0

Predicted label=1

Predicted label=0

Predicted label=1

Test set

- In the training set, grass is **highly correlated** with the dog label.
- Fitting neural networks optimizes **predictive accuracy**.
- The grass feature **predicts the dog label (almost) as well as** the dog feature in the training data.

Representation learning picks up spurious features

Label=1

Label=0

Training set

Predicted label=0

Predicted label=1

Predicted label=0

Predicted label=1

Test set

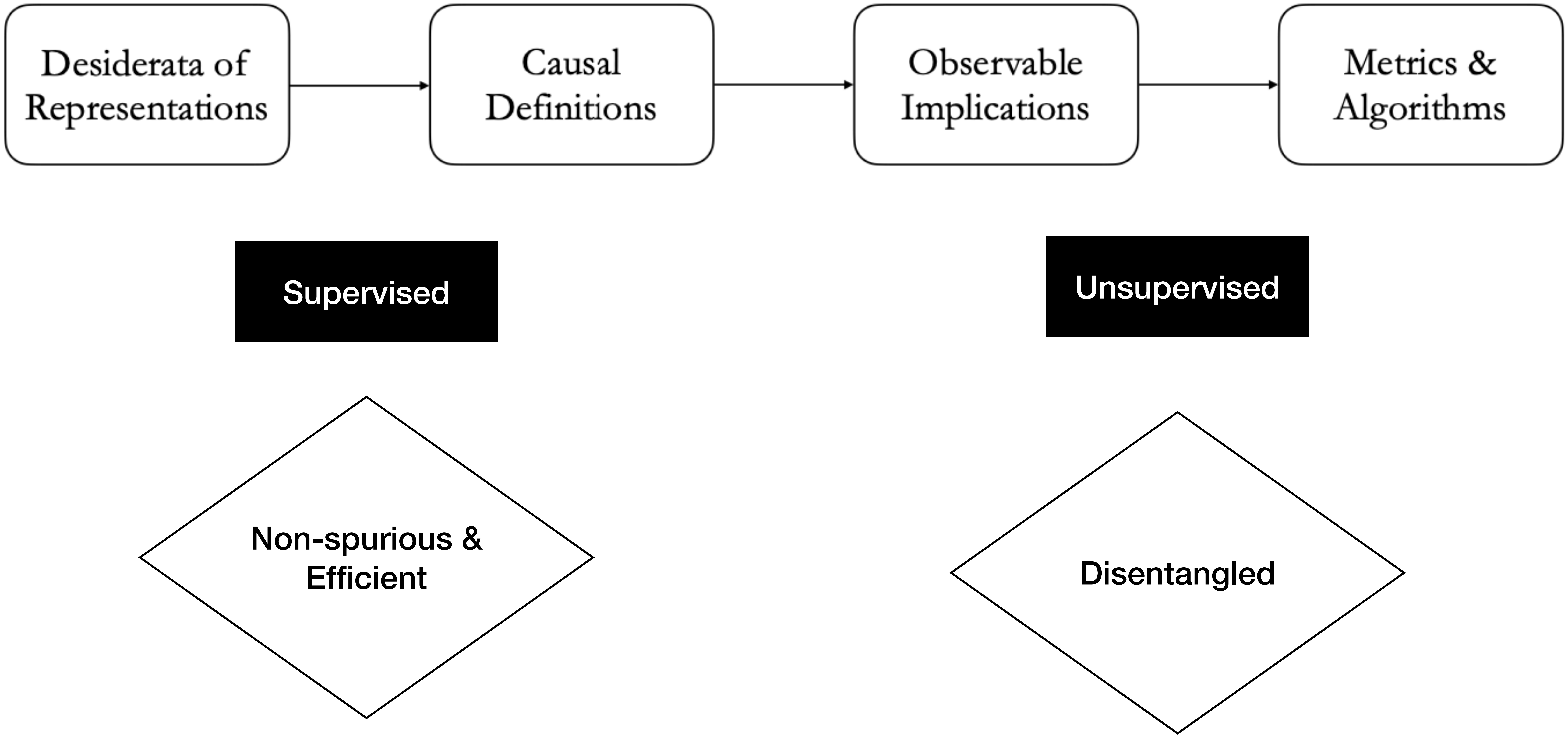
- It is a problem of the **training objective**. Maximizing predictive accuracy does not prevent spurious features.
- Restrict our attention to only non-spurious features? Optimize for non-spuriousness?
- **We need a mathematical definition and/or metric of representation non-spuriousness.**

Desiderata for Representation learning

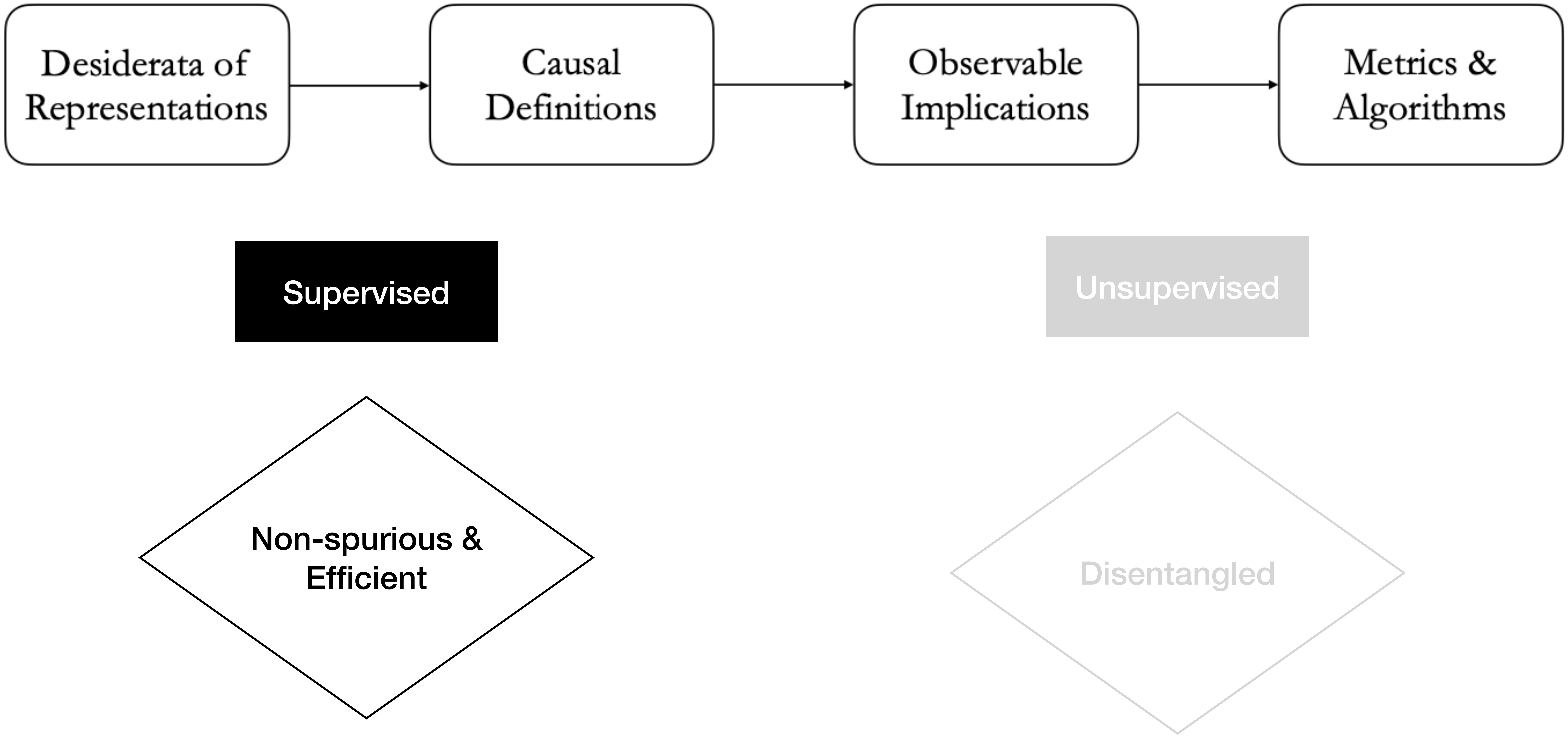
Non-spurious	Efficient	Disentangled
✓ dog face	✓ dog face	✓ (dog face, four legs)
✗ grass	✗ (dog face, four legs)	✗ (dog face + four legs, dog face - four legs)

- Optimizing for predictive accuracy **does not** produce desired representations.
- Shall **formalize the desiderata** to be incorporated into learning objectives
- **Causal inference is here to help! (Ask "What if…" questions about interventions)**

Representation Learning: From Desiderata to Algorithms



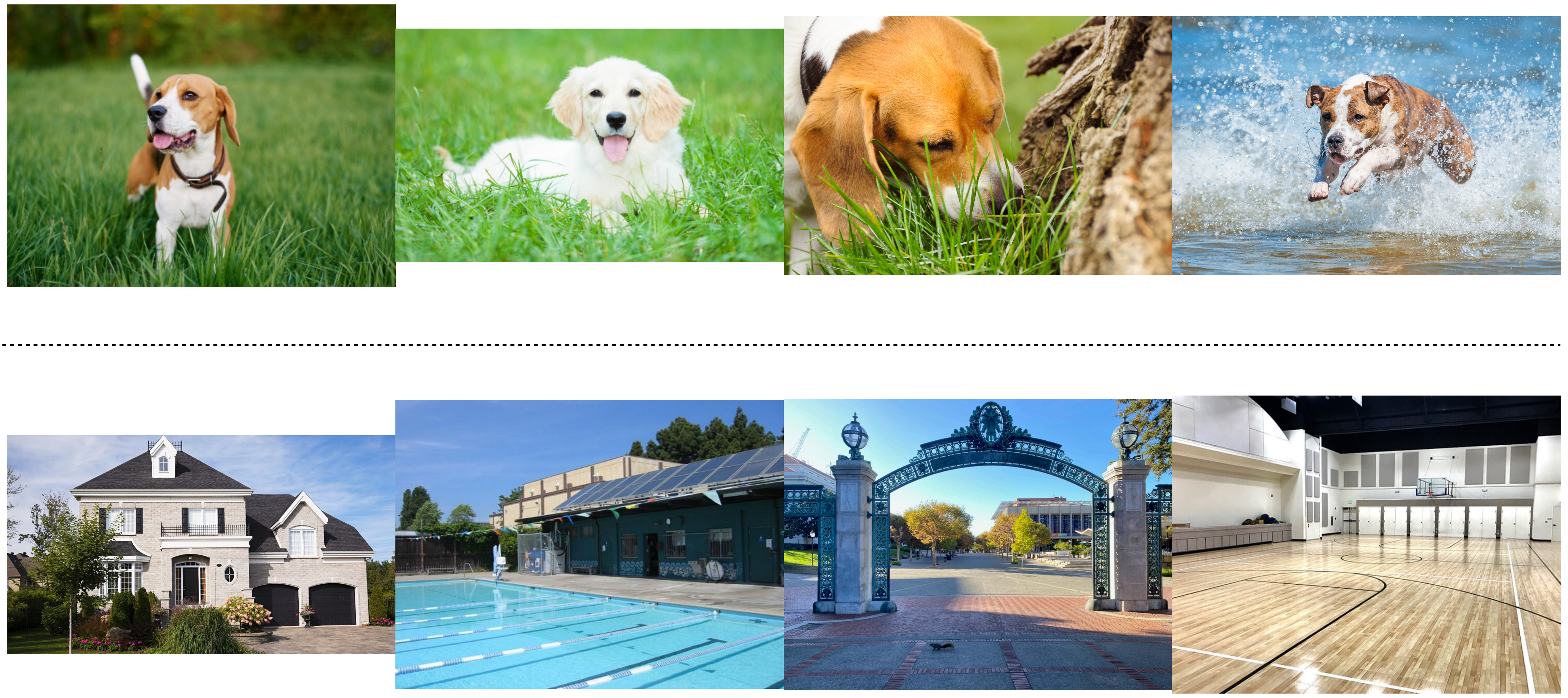
Representation Learning: From Desiderata to Algorithms



How can we define non-spuriousness and efficiency?

What does “non-spuriousness” mean?

Label=1



Label=0

✓ dog face

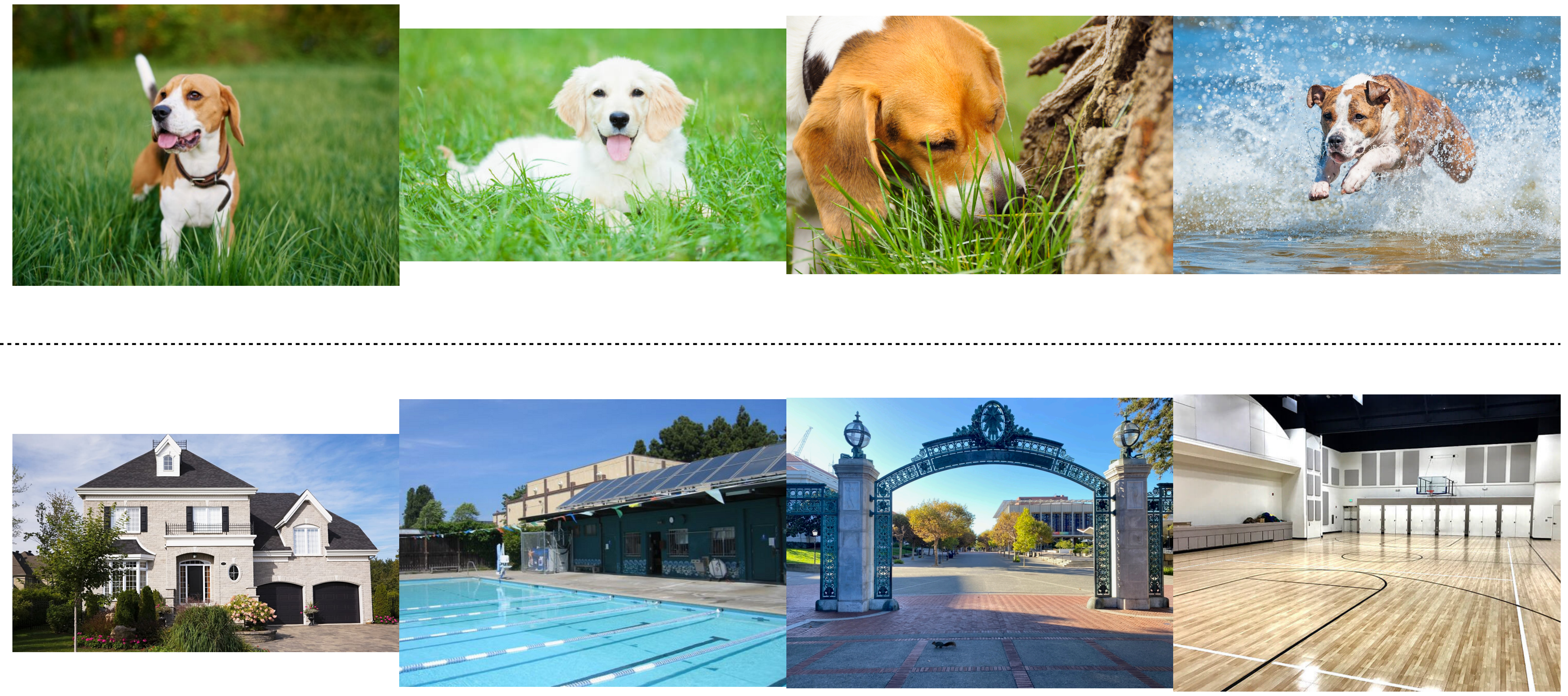
✗ grass

- **Non-spurious** representations $\mathbf{Z} = f(\mathbf{X})$ capture features that **causally determine** the label.
- The key idea is to view the feature $\mathbf{Z} = \mathbf{z}$ as a **potential cause** of the label $Y = y$, then a **non-spurious** feature shall be a ***sufficient cause*** of the label.

Non-spuriousness and its Counterfactual Metric

Label=1

Label=0

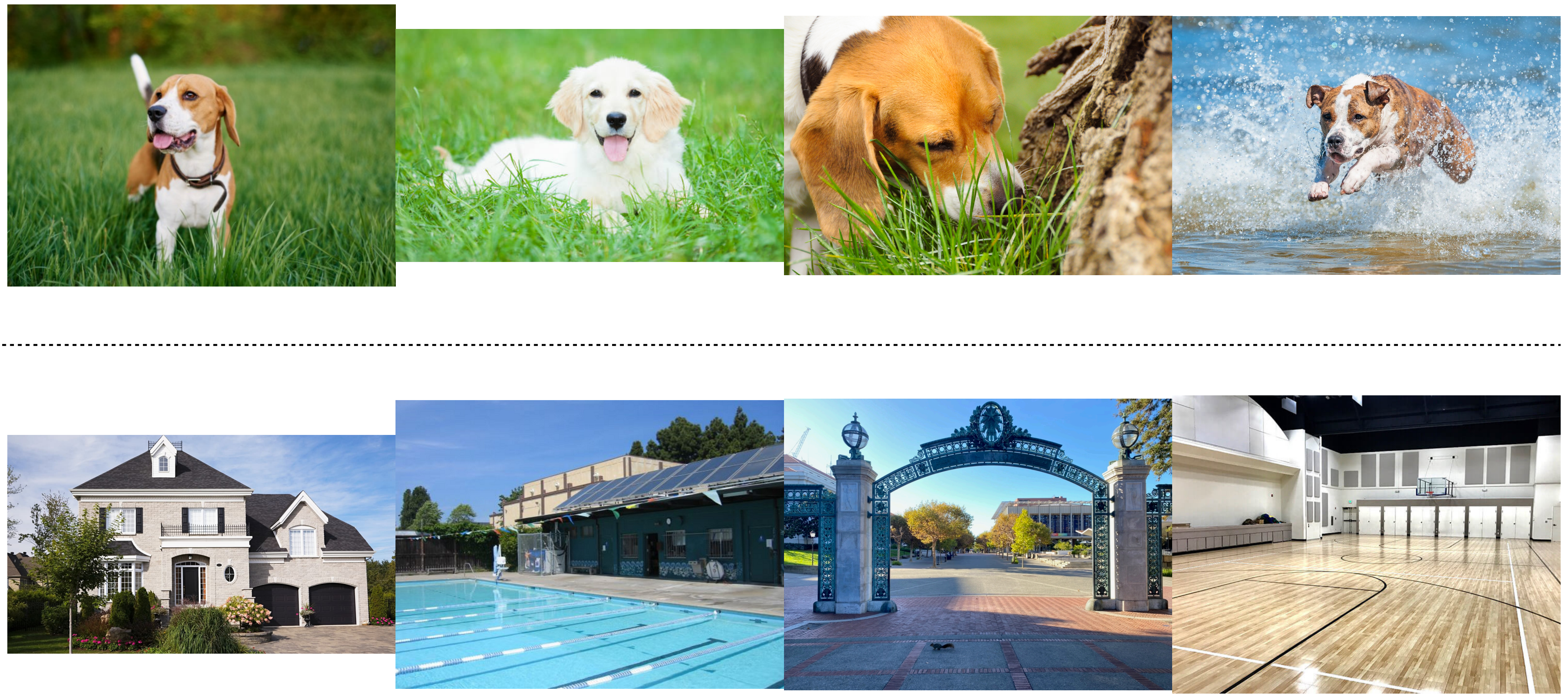


- Suppose $\mathbf{Z}$ is the grass feature. Does it **sufficiently cause** the dog label?
- Given an image that has no grass $\mathbf{Z} = 0$ and is not labeled dog $Y = 0$.
- What would be counterfactual label $Y(\mathbf{Z} = 1)$ if we *add some grass* into this image? Would its label become dog?

- We consider **counterfactual labels** $Y(\mathbf{Z} = 1)$ of images when we turn on its features $\mathbf{Z}$.
- Quantify **non-spuriousness** using the **probability of sufficiency (PS)** (Pearl, 2009) $\mathrm{PS} \triangleq P(Y(\mathbf{Z} = 1) = 1 \mid \mathbf{Z} = 0, Y = 0)$
- For continuous features and labels, we consider the PS of $\mathbf{1}\{\mathbf{Z} = \mathbf{z}\}$ for $\mathbf{1}\{Y = y\}$: $\mathrm{PS}_{\mathbf{Z}=z, Y=y} \triangleq P(Y(\mathbf{Z} = \mathbf{z}) = y \mid \mathbf{Z} \neq \mathbf{z}, Y \neq y)$

What does “efficiency” mean?

Label=1



Label=0

✓ dog face

✗ (dog face, four legs)

- An **efficient** representation $\mathbf{Z} = f(\mathbf{X})$ captures only **essential** features of the data; no redundant features captured.
- Again, viewing the feature $\mathbf{Z} = \mathbf{z}$ as a **potential cause** of the label $Y = y$, then an **efficient** representation must capture features that are ***necessary causes*** of the label.

Efficiency and its Counterfactual Metric

Label=1

Label=0

- Suppose $\mathbf{Z}$ is the 'dog face & four legs' feature. Does it **necessarily cause** the dog label?

- Given an image that has dog face & four legs $\mathbf{Z} = 1$ and is labeled dog $Y = 1$.

- What would be counterfactual label $Y(\mathbf{Z} = 0)$ if we turn off the 'dog face & four legs' feature, e.g. move one leg of the dog out of the image? Would its label necessarily become non-dog?

- We consider **counterfactual labels** $Y(\mathbf{Z} = 0)$ of images when we turn off its features $\mathbf{Z}$.

- Quantify **efficiency** using the **probability of necessity (PN)** (Pearl, 2009) $\mathrm{PN} \triangleq P(Y(\mathbf{Z} = 0) = 0 \mid \mathbf{Z} = 1, Y = 1)$

- For continuous features and labels, we consider the PN of $\mathbf{1}\{\mathbf{Z} = \mathbf{z}\}$ for $\mathbf{1}\{Y = y\}$: $\mathrm{PN}_{\mathbf{Z}=z, Y=y} \triangleq P(Y(\mathbf{Z} \neq \mathbf{z}) \neq y \mid \mathbf{Z} = \mathbf{z}, Y = y)$

Quantifying Non-spuriousness and Efficiency Simultaneously

Label=1

Label=0

- Quantify **non-spuriousness and efficiency simultaneously** using the **probability of necessity and sufficiency** (PNS) of $\mathrm{PNS} \triangleq P(Y(\mathbf{Z}=0)=0, Y(\mathbf{Z}=1)=1)$
- **Non-spuriousness**: counterfactual labels when we **turn on** its features; **Efficiency**: counterfactual labels when we **turn off** its features
- For multiple features: **conditional non-spuriousness and efficiency** $\mathrm{PNS}_{Z_j, Y \mid \mathbf{Z}_{-j}} \triangleq P(Y(Z_j=0, \mathbf{Z}_{-j}=1)=0, Y(Z_j=1, \mathbf{Z}_{-j}=1)=1)$

Representation Learning as Finding Necessary and Sufficient Causes

- **CAUSAL-REP: Maximize the non-spuriousness and efficiency of the representation**

$$\max_f \sum_{i=1}^{n} \log \mathrm{PNS}_{f(\mathbf{X})=f(\mathbf{x}_i),Y=y_i}$$

where $\mathbf{X} = (X_1, \ldots, X_m)$, $\mathbf{x}_i = (x_{i1}, \ldots, x_{im})$, and $(\mathbf{x}_i, y_i)$ is the ith data point.

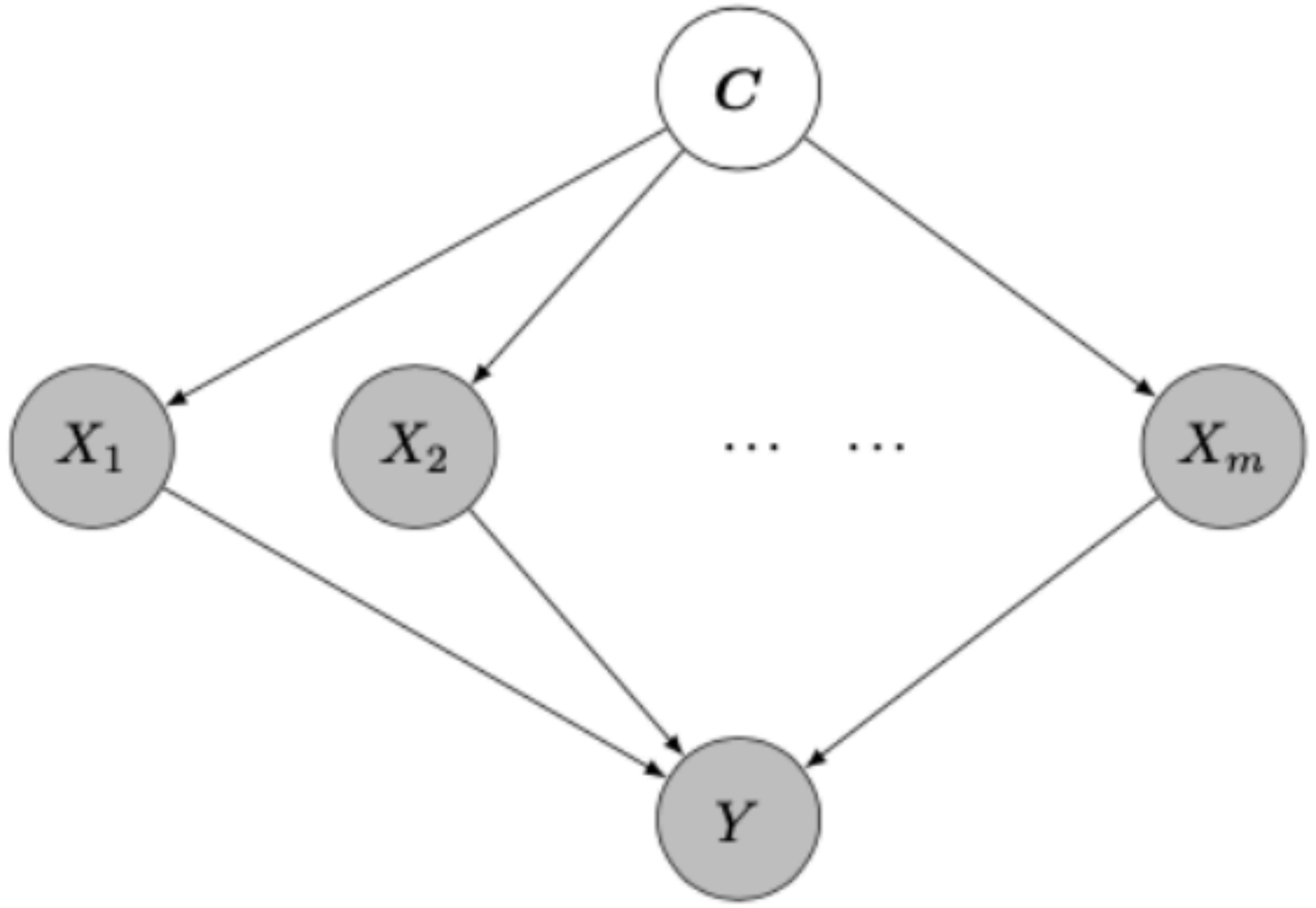
- For multi-dimensional representation: Maximize (conditional) non-spuriousness and efficiency

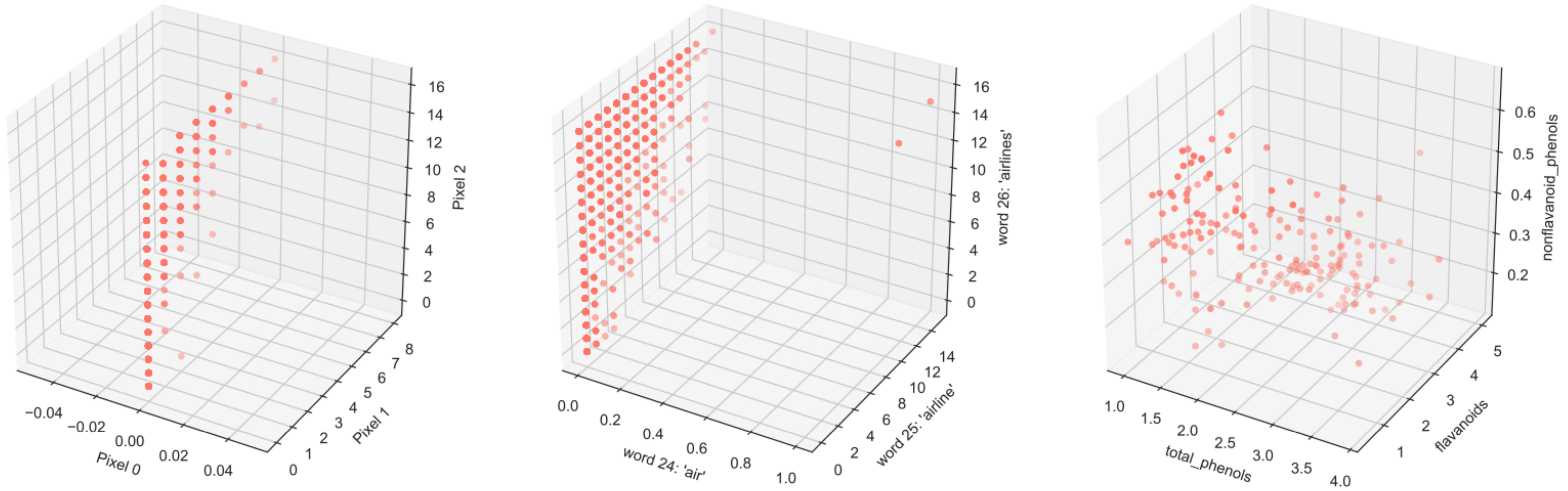
$$\max_f \sum_{i=1}^{n} \sum_{j=1}^{d} \log \mathrm{PNS}_{f_j(\mathbf{X})=f_j(\mathbf{x}_i),Y=y_i|f_{-j}(\mathbf{X})=f_{-j}(\mathbf{x}_i)}$$

where $f(\mathbf{X}) = (f_1(\mathbf{X}), \ldots, f_d(\mathbf{X}))$ is the d-dimensional representation.

How can we evaluate non-spuriousness and efficiency from data?

How can we identify PNS from data?





(a) High-dim. image data: MNIST (Deng, 2012)

(b) High-dim. text data: Airline tweets

(c) Low-dim. data: Wine features

- $\mathrm{PNS}_{Z=z,Y=y} \triangleq P(Y(Z=z)=y, Y(Z\neq z)\neq y)$ is a **counterfactual** (rung 3) quantity.
- **Two main challenges:** (1) PNS can not be identified exactly. It can only be bounded. We derive a (tight) **lower bound of PNS**$\mathrm{PNS}_{Z=z,Y=y} \geq P(Y=y \mid \mathrm{do}(Z=z)) - P(Y=y \mid \mathrm{do}(Z\neq z))$
- (2) Identifying $P(Y=y \mid \mathrm{do}(Z=z))$ with $Z=f(\mathbf{X})$ often requires $P(Y \mid \mathbf{X})$, which is challenging for **high-dimensional $\mathbf{X}$.**

How can we identify PNS from data?

- **Identification (cont'd):**
  - (2) Identifying the intervention distribution$P(Y=y \mid \mathrm{do}(Z=z))$
    - **Functional interventions** (Puli et al., 2020) $P(Y=y \mid \mathrm{do}(Z=z)) = P(Y=y \mid \mathrm{do}(f(X)=z))$
      - Conditional on all parents of X, manipulate X such that $f(X)=z$
      - $P(Y=y \mid \mathrm{do}(f(X)=z)) = \int P(Y=y \mid \mathrm{do}(X=x))P(X=x \mid f(X)=z, C)P(C)\mathrm{d}C;$
    - Need to pinpoint the unobserved common cause C;
    - High-dimensional X living on low dimensional manifold; restrict to subvectors of X
- Much of the technical development in CAUSAL-REP is for identifying $P(Y=y \mid \mathrm{do}(f(X)=z))$ for high-dimensional X.

CAUSAL-REP: What just happened?

Desiderata of Representations → Causal Definitions → Observable Implications → Metrics & Algorithms

Non-spurious

✓ dog face

✗ grass

Efficient

✓ dog face

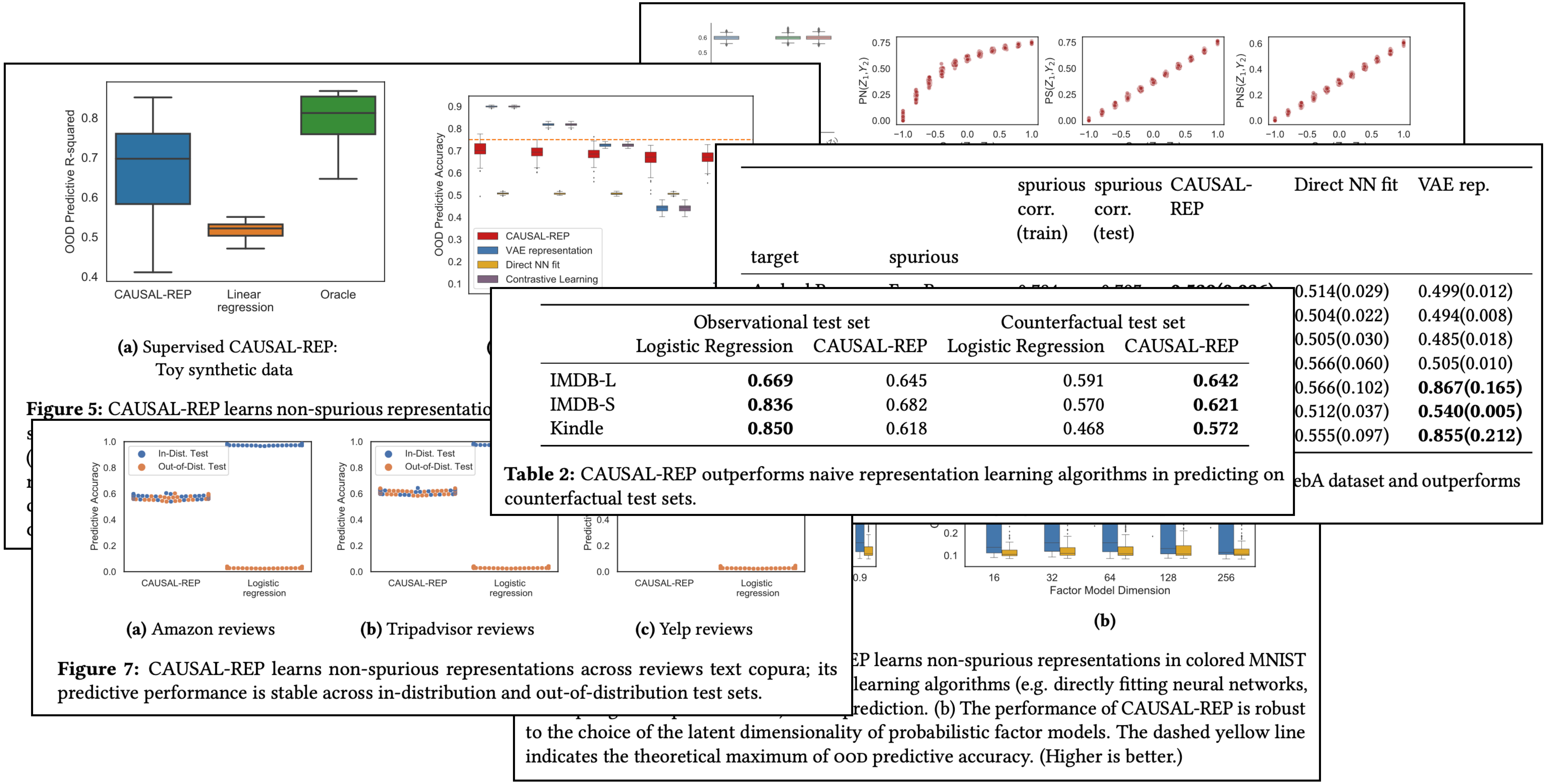
✗ (dog face, four legs)

What about unsupervised representation learning?

- **We reduce unsupervised representation learning to a supervised problem of instance discrimination.**
- Specifically, we formulate the unsupervised problem as finding representations that can distinguish different subjects (instance discrimination).
    - Consider a unsupervised dataset where augmentation is available.
    - We have many different augmented observations for each subject i.
    - We set the subject ID as the label.

Empirical Studies of CAUSAL-REP

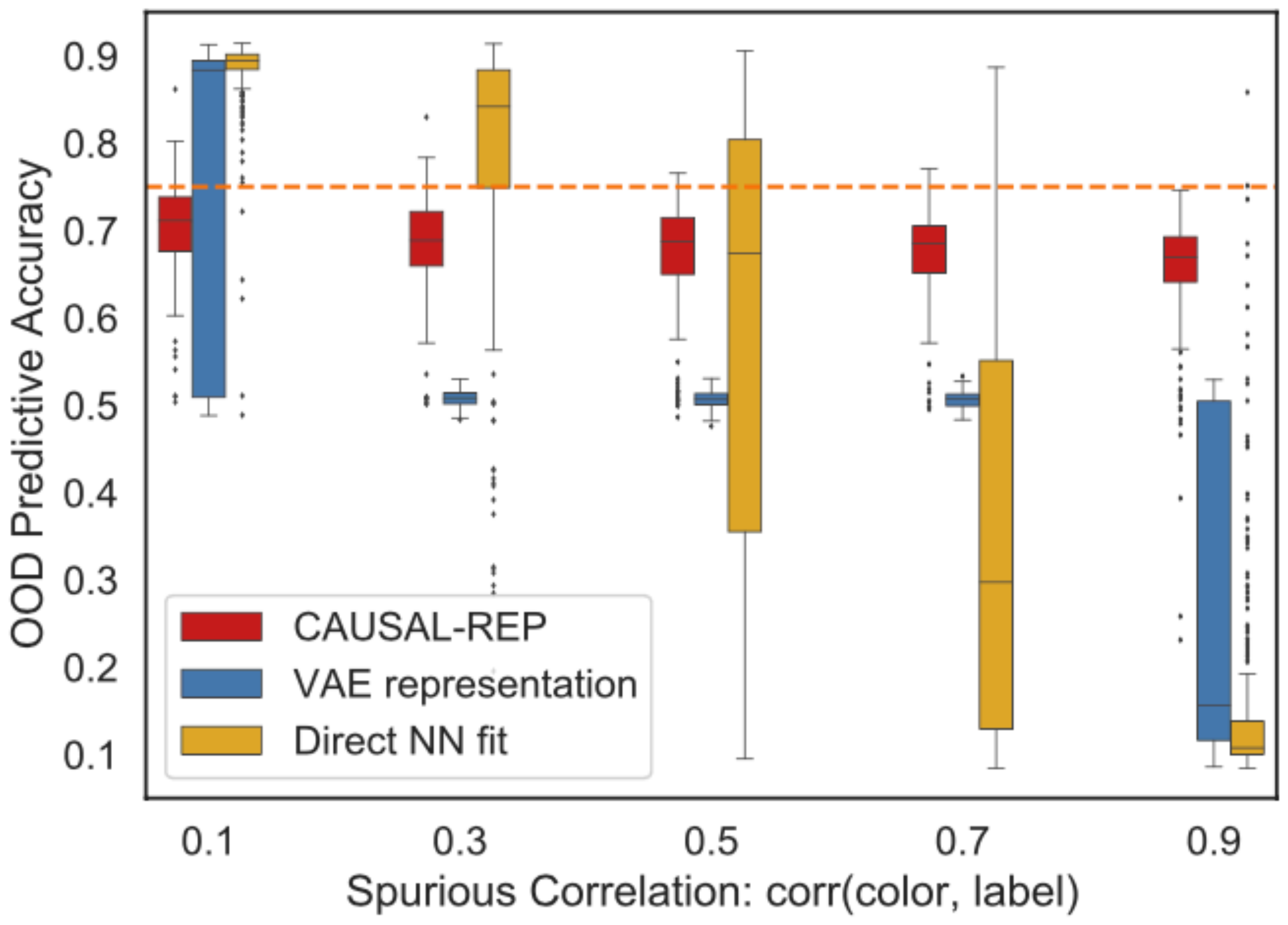
We did lots of empirical studies in the paper

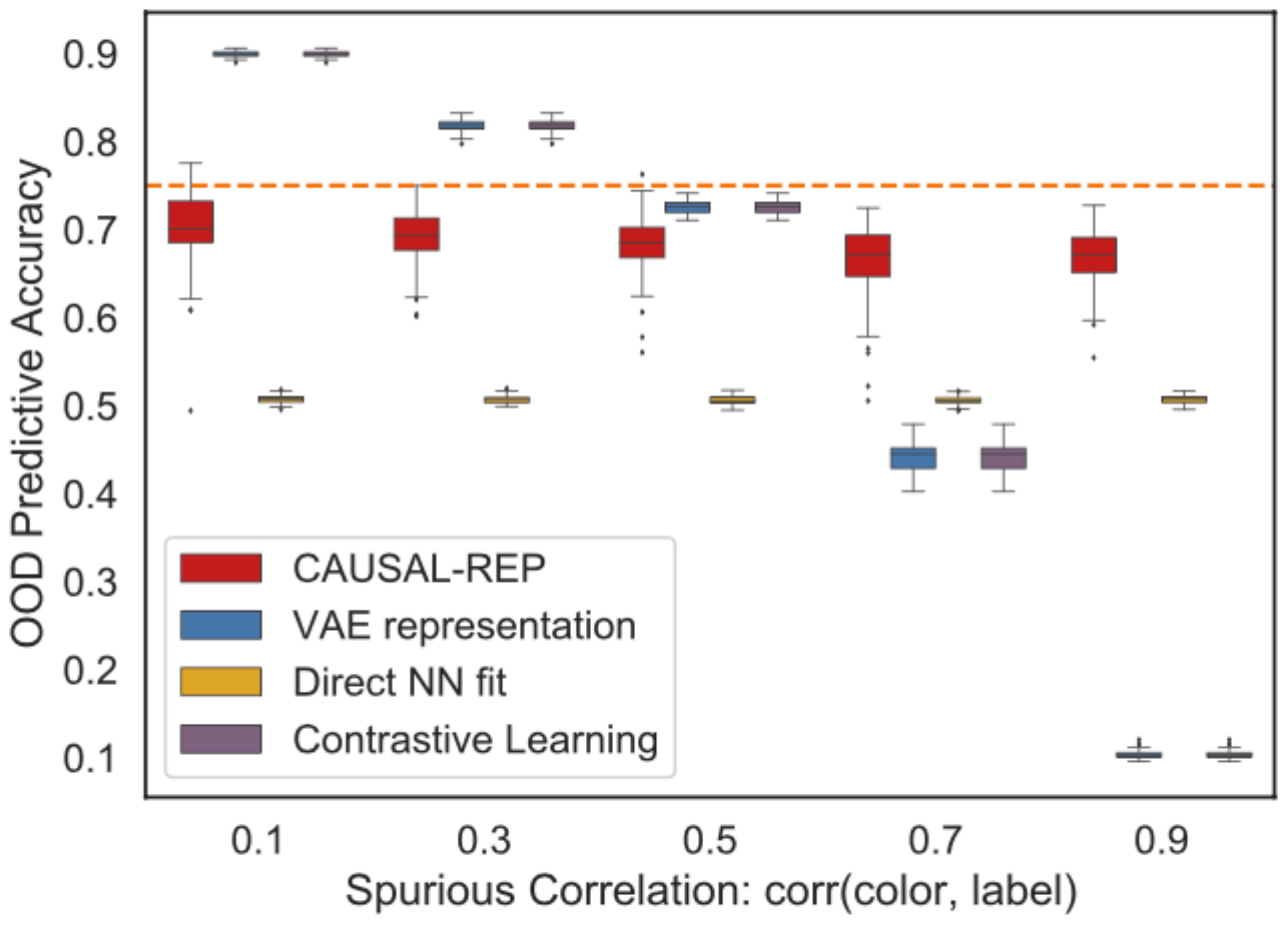


	Observational test set		Counterfactual test set	
	Logistic Regression	CAUSAL-REP	Logistic Regression	CAUSAL-REP
IMDB-L	**0.669**	0.645	0.591	**0.642**
IMDB-S	**0.836**	0.682	0.570	**0.621**
Kindle	**0.850**	0.618	0.468	**0.572**

Table 2: CAUSAL-REP outperforms naive representation learning algorithms in predicting on counterfactual test sets.

Empirical Studies on Colored MNIST Images





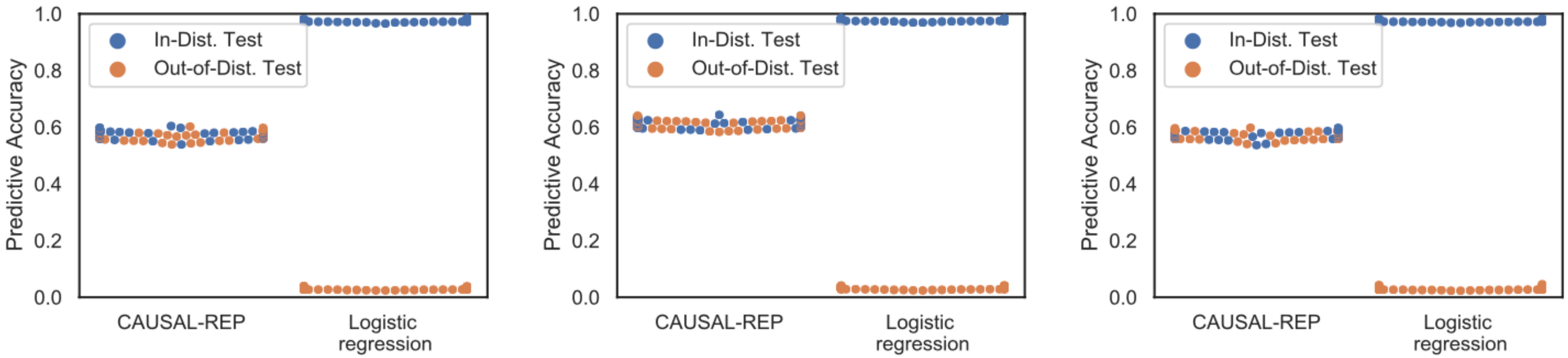
- Training set: corr(color, label) is positive; Test set: corr(color, label) is negative.
- Randomly flip 25% of the labels in both training and testing.
- **CAUSAL-REP finds non-spurious features even if we work with a single dataset**; no multiple environments or data augmentation or invariance.

Empirical Studies on Text

Amazon	CAUSAL-REP	Logistic Regression
1	love_this_camera, recommend_this_camera, my_first_digital, great, best_camera, camera_if_you, this_camera_and, camera_have, excellent_camera, camera_bought_this;	am, an, also, as, love_my, the_tracfone,
2	this_camera, camera, camera_is, pictures, picture, the_camera, digital, camera_for, this_camera_is, digital_camera;	it_real, which_is, too, so_much,
3	really_nice, hold_the, excellent_it, this_one_it, easy_it, is_superb, nice_if, returning, too_low, you_need_more;	is_so_much, which_is_pretty,
4	with_this, aa, took, came, yet, pictures_of, camera_in, computer, pictures_in, for_those;	nokia, ear, home, is_must, for_your,
5	camera_was, expect, the_photos, by, camera_are, blurry, sony, have_an, had_some, wife;	faster, must_for, when_use

- Amazon reviews corpus; Positive / negative ratings as binary labels
- Inject spurious words ‘am’, ‘an’, ‘also’, ‘as’ into positive reviews of the training set, but not test datasets.
- CAUSAL-REP finds non-spurious (and meaningful) features

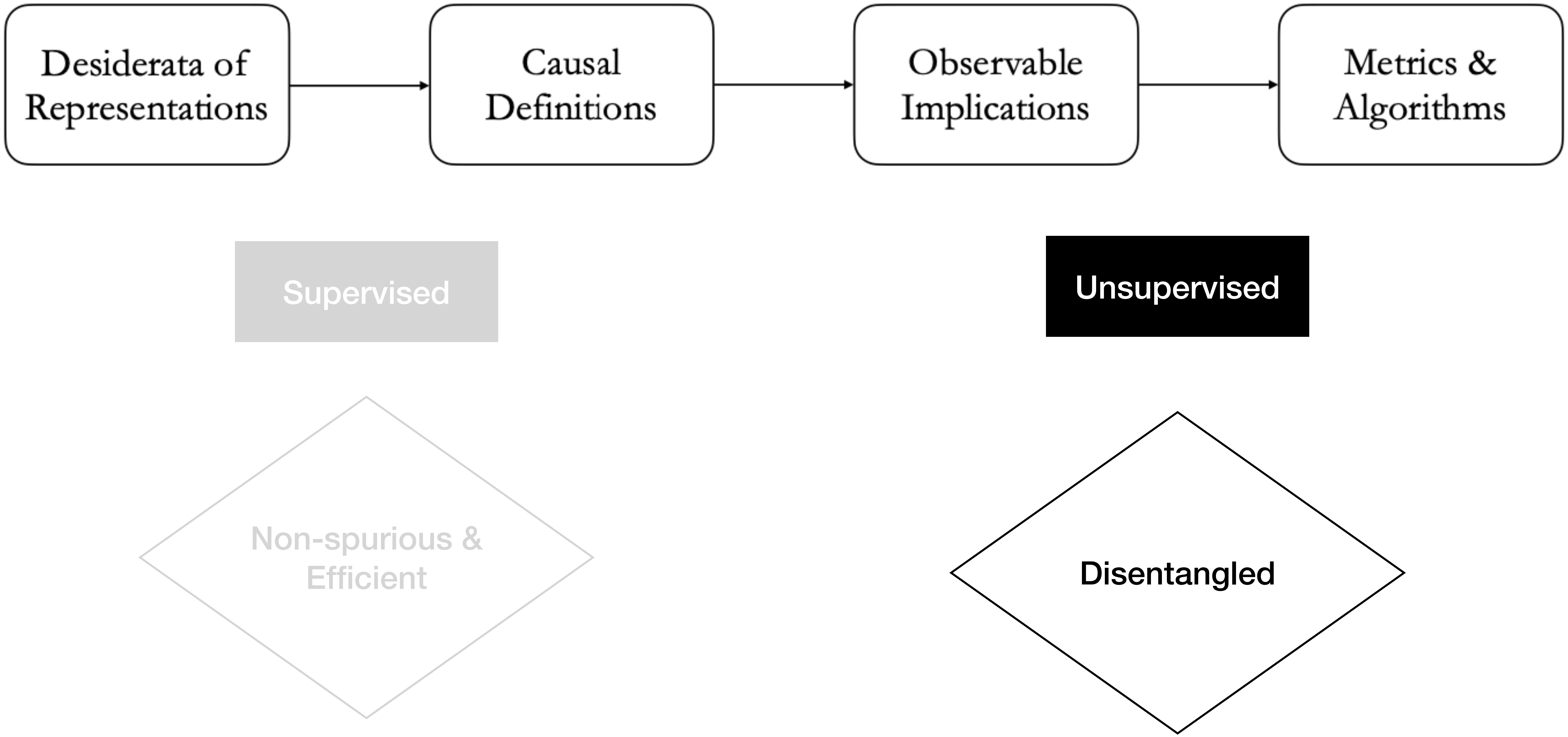
Empirical Studies on Text



(a) Amazon reviews (b) Tripadvisor reviews (c) Yelp reviews

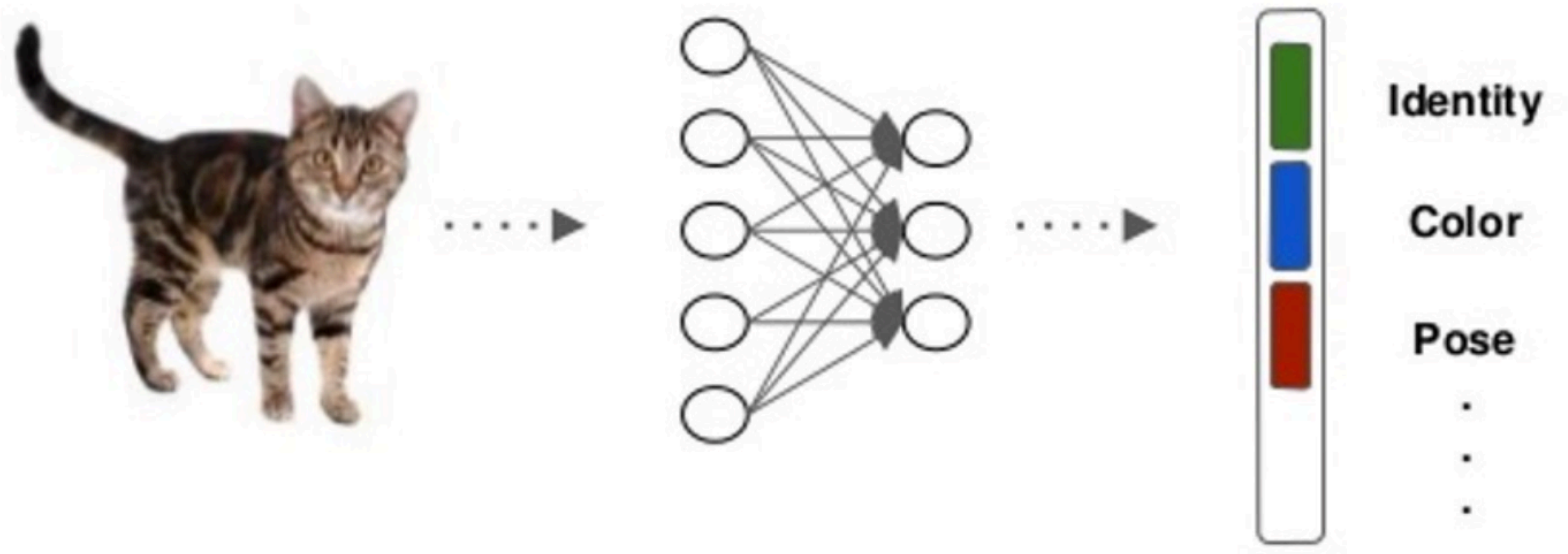
Figure 7: CAUSAL-REP learns non-spurious representations across reviews text copura; its predictive performance is stable across in-distribution and out-of-distribution test sets.

Representation Learning: From Desiderata to Algorithms



What is the definition of disentanglement?

What does “disentanglement” mean?

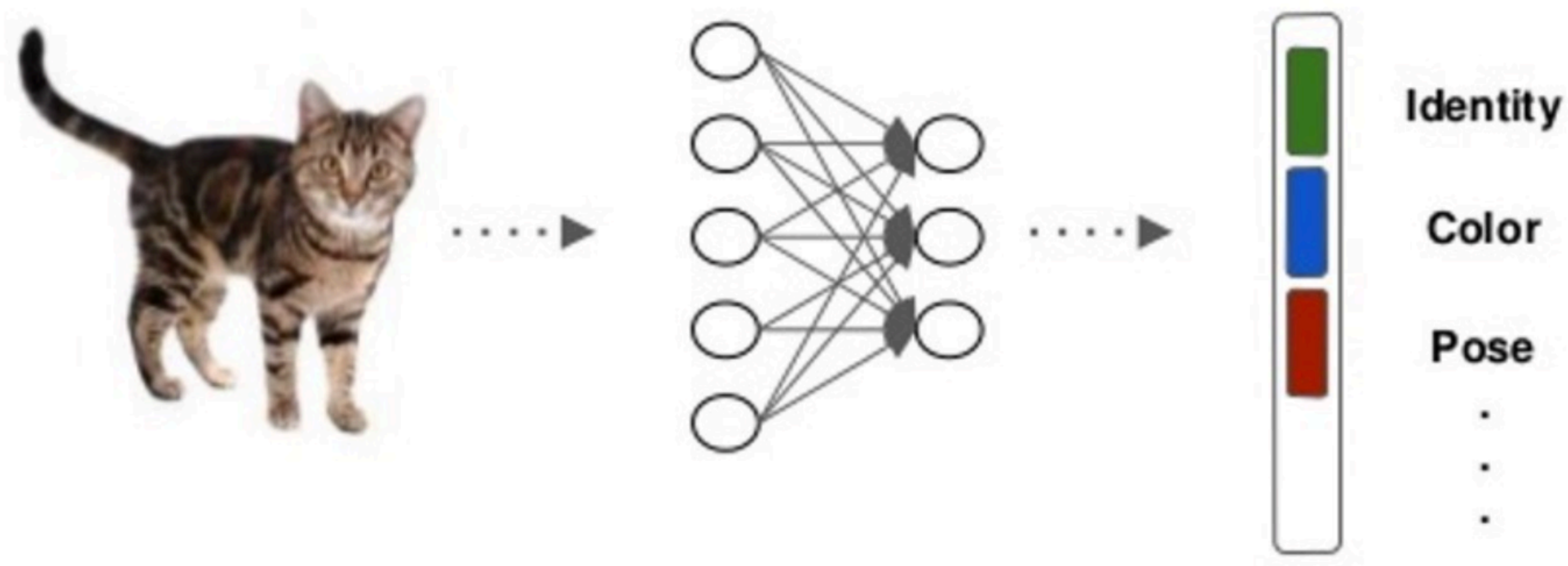


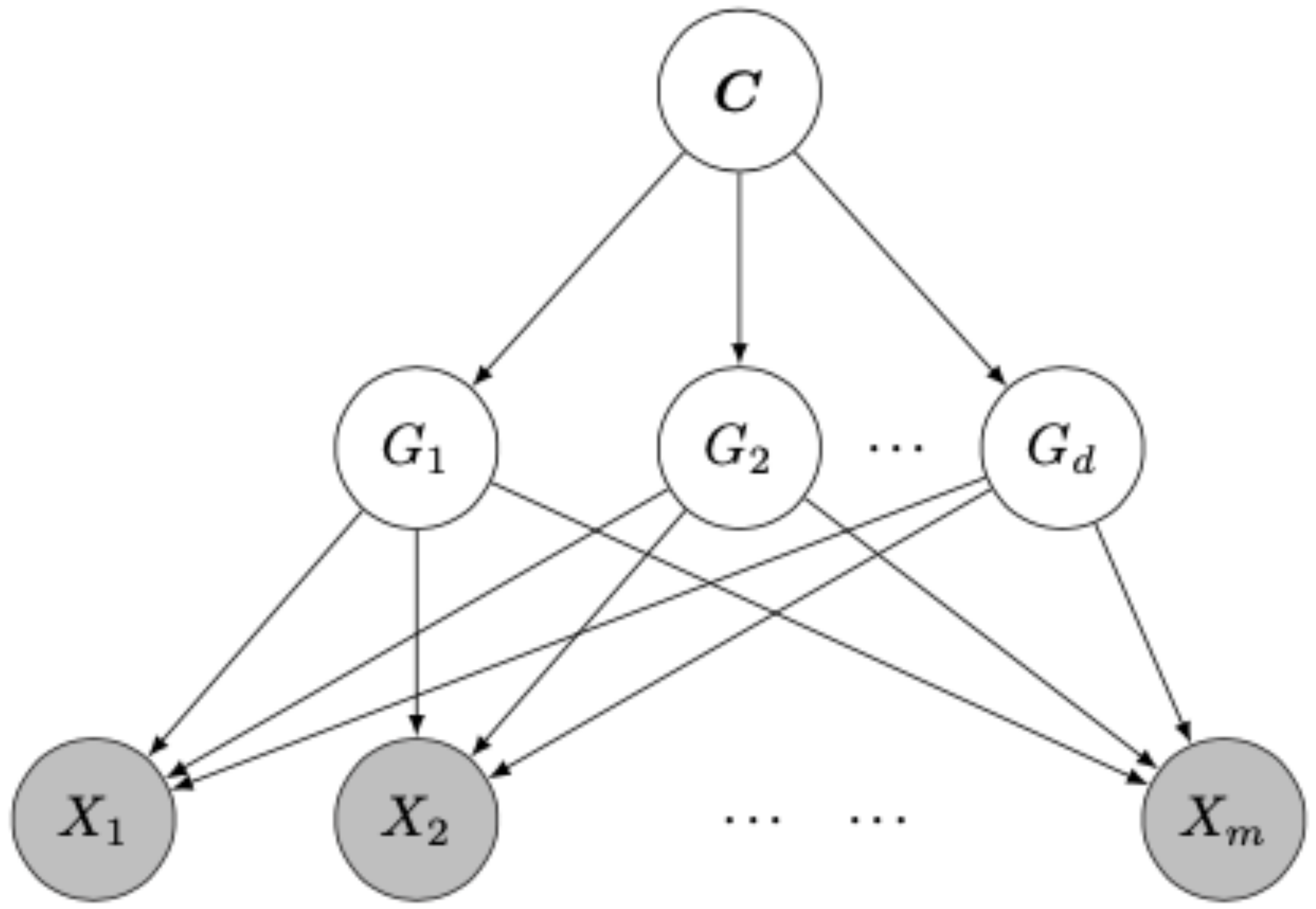
✔ (pose, color)

✘ (pose + color, pose - color)

- Disentangled representations capture **independently controllable** factors of variation (FOVs).
- How to **evaluate or enforce disentanglement** without knowing ground truth features?
- **We work with a single unsupervised dataset, without auxiliary labels or weak supervision.**

What does “disentanglement” mean?





- **Definition: Causal disentanglement (Suter et al., 2019)**
  A representation $\mathbf{G} = (G_1, \ldots, G_d)$ is ***(causally) disentangled*** if $G_1, \ldots, G_d$ represent **(possibly correlated)** factors of variation (FOVs) that do not causally affect each other.

- The absence of causal relationships among the FOVs $G_1, \ldots, G_d$ allows us to freely manipulate them.

How can we assess disentanglement from data?

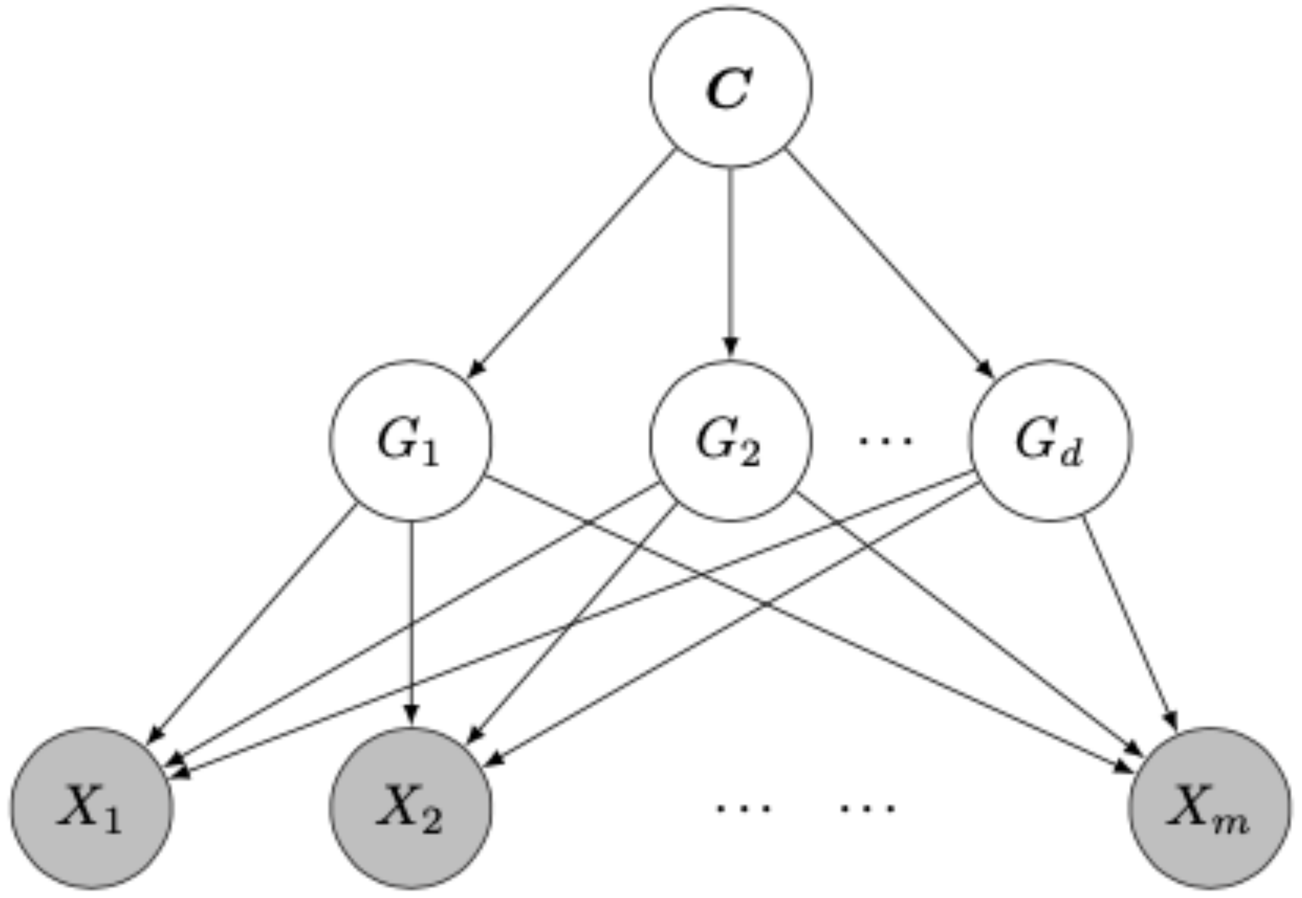
How can we assess causal disentanglement?

- **Absence of causal relationships among** $G_1, \ldots, G_d$
  $P(G_j \mid \mathrm{do}(G_{\setminus j} = g_{\setminus j})) = P(G_j), \quad \forall j, g_{\setminus j}.$

- This is an **interventional distribution of** $G_{\setminus j}$ **on** G_j**.**

- **Identification**: The causal relationships among $G_1, \ldots, G_d$ can be confounded by some unobserved $\mathbf{C}$. Thus $P(G_j \mid \mathrm{do}(G_{\setminus j} = g_{\setminus j}))$ is non-identifiable from observational data $P(G_1, \ldots, G_d)$. **(Not all causal questions are answerable.)**

- Still, we ask: how does the absence of causal relationships relate to observational data? Are there any **observable implications** of $P(G_i \mid \mathrm{do}(G_{\setminus i} = g_{\setminus i})) = P(G_i), \quad \forall i, g_{\setminus i}$?



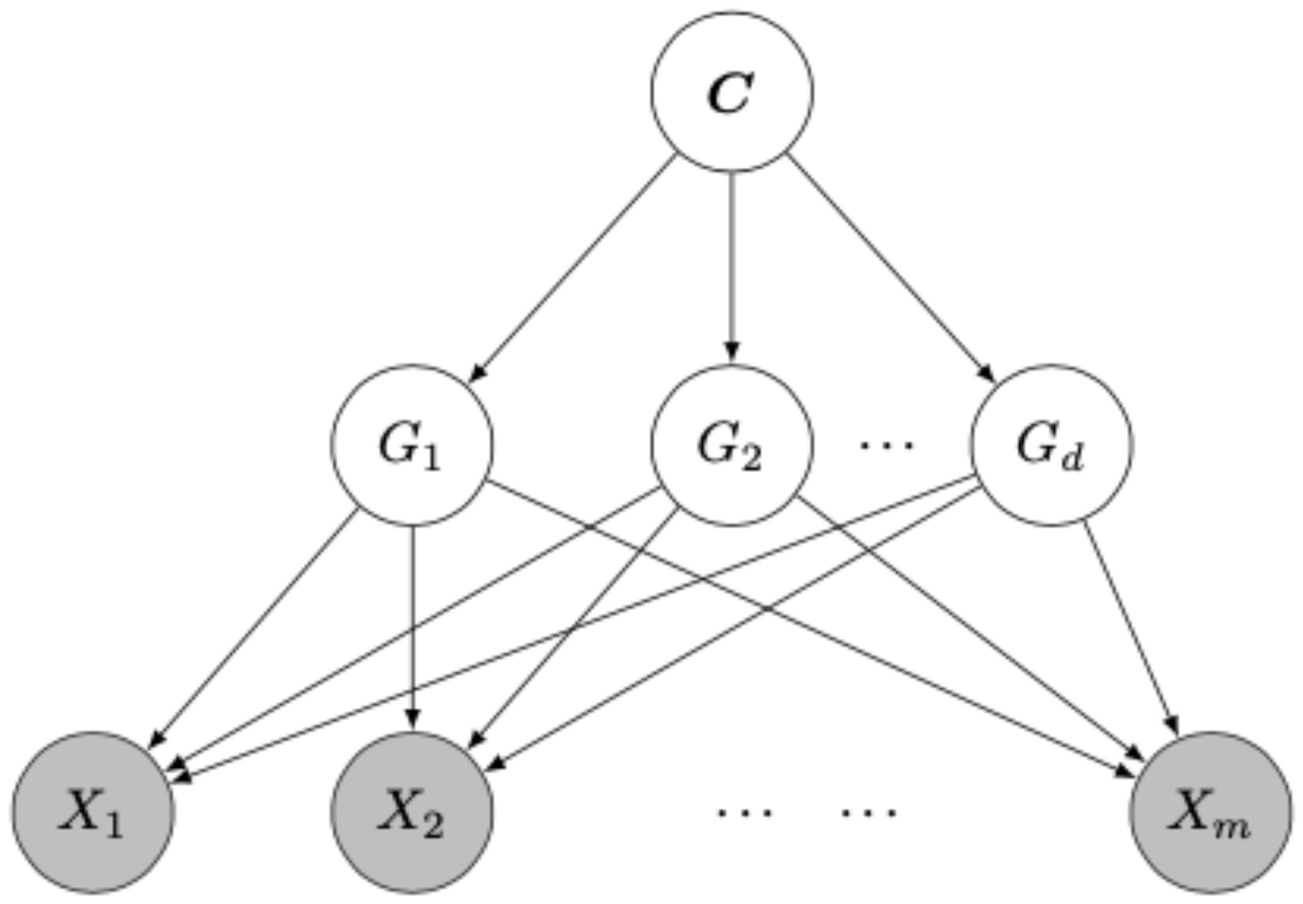
Observable implications of causal disentanglement

- **Key observation:** There does exist an observable implication of causal disentanglement $P(G_j \mid \mathrm{do}(G_{\setminus j} = g_{\setminus j})) = P(G_j), \quad \forall j, g_{\setminus j}$, on the **support** of $\mathrm{supp}(G) \triangleq \mathbf{1}\{P(G) > 0\}$.

- **Theorem. (Causal disentanglement ⇒ independent support)** Under the positivity condition $P(G_j \mid \mathbf{C}) > 0$ iff $P(G_j) > 0, \quad \forall j$, no causal connections among $G_1, \ldots, G_d$ implies that

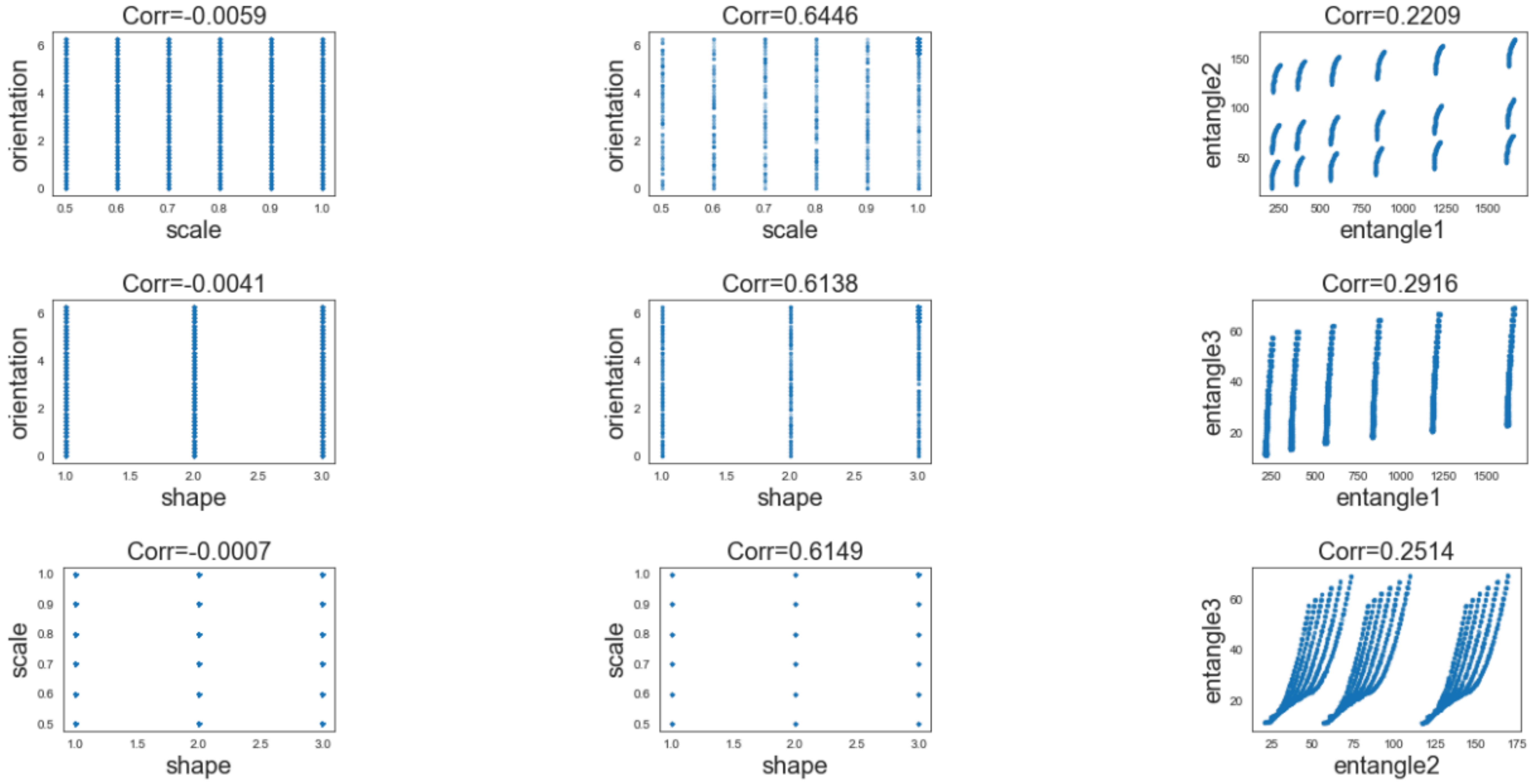
$$\mathrm{supp}(G_j \mid G_{\mathcal{S}}) = \mathrm{supp}(G_j), \quad \forall j, \mathcal{S} \subset \{1,\ldots,d\} \backslash j,$$
$$\mathrm{supp}(G_1, \ldots, G_d) = \mathrm{supp}(G_1) \times \cdots \times \mathrm{supp}(G_d).$$

- **Intuition: Positivity implies that $\mathbf{C}$ cannot affect the support of $G_1, \ldots, G_d$. If they do not affect each other, then their support has to be independent.**



Representations with independent support

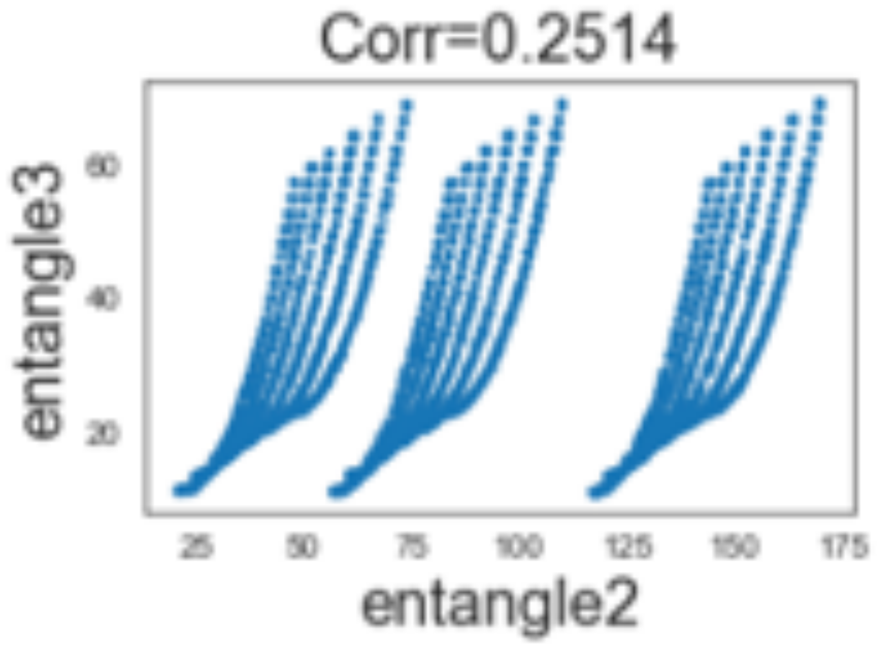
- **Independent support**: $\mathrm{supp}(G_j \mid G_{\mathcal{S}}) = \mathrm{supp}(G_j), \quad \forall j, \mathcal{S} \subset \{1,\ldots,d\} \backslash j$
  Visually, the support of $G_1, \ldots, G_d$ must be (hyper-)rectangular.



(a) Disentangled and uncorrelated **(b)** Disentangled but highly correlated **(c)** Entangled but with low correlations

Quantifying disentanglement with the independence-of-support score (IOSS)

- **Causal disentanglement ⇒ independent support** $\mathrm{supp}(G_1, \ldots, G_d) = \mathrm{supp}(G_1) \times \cdots \times \mathrm{supp}(G_d)$
- **Independence-of-support-score (IOSS): A disentanglement metric**



$$\mathrm{IOSS} \triangleq d_H(\mathrm{supp}(\bar{G}_1, \ldots, \bar{G}_d), \mathrm{supp}(\bar{G}_1) \times \cdots \times \mathrm{supp}(\bar{G}_d)),$$

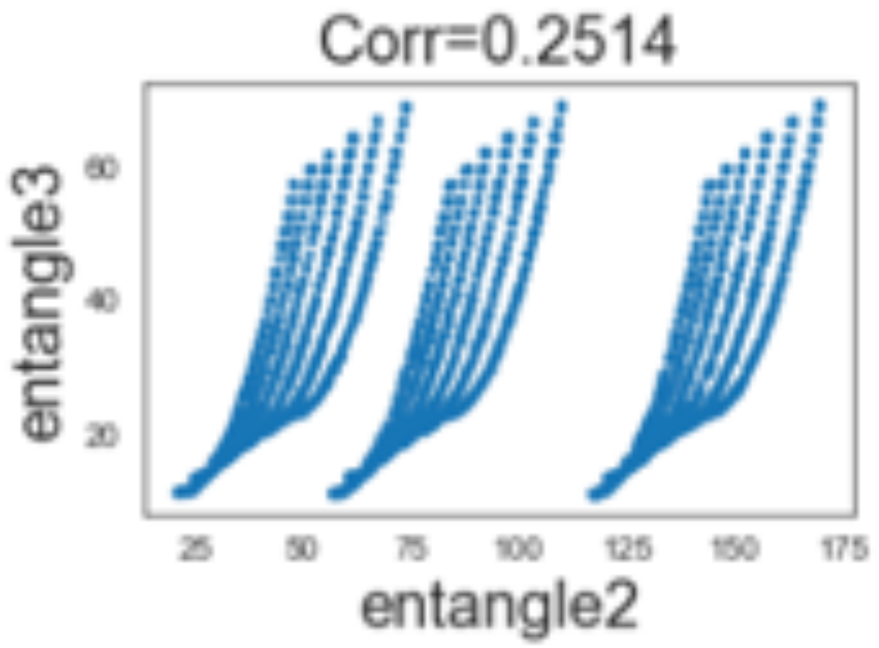
where $\bar{G}_j = (G_j - \inf G_j)/(\sup G_j - \inf G_j)$ is the standardized G_j and

$d_{\mathrm{H}}(X, Y) \triangleq \max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \right\}$ is the Hausdorff distance.

- **Disentangled representation learning with an IOSS penalty**
  - (Identifiability) If compact support, independent support is sufficient for enforcing disentanglement.

Independence-of-Support Score (IOSS)

- **Causal disentanglement ⇒ independent support** $\mathrm{supp}(G_1, \ldots, G_d) = \mathrm{supp}(G_1) \times \cdots \times \mathrm{supp}(G_d)$
- **Independence-of-support-score (IOSS): A disentanglement metric**



$$\mathrm{IOSS} \triangleq d_H(\mathrm{supp}(\bar{G}_1, \ldots, \bar{G}_d), \mathrm{supp}(\bar{G}_1) \times \cdots \times \mathrm{supp}(\bar{G}_d)),$$

where $\bar{G}_j = (G_j - \inf G_j)/(\sup G_j - \inf G_j)$ is the standardized G_j and

$d_{\mathrm{H}}(X, Y) \triangleq \max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \right\}$ is the Hausdorff distance.

- **Disentangled representation learning with an IOSS penalty**
  - **Identifiability:** If compact support, independent support is sufficient for enforcing disentanglement.

IOSS: What just happened?

Desiderata of Representations → Causal Definitions → Observable Implications → Metrics & Algorithms

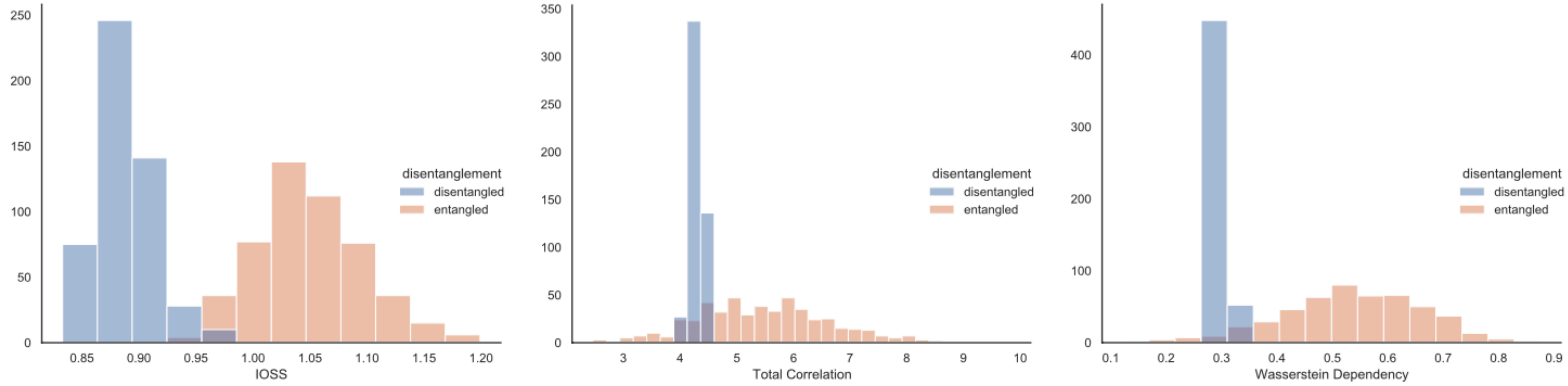
Disentangled

✓ (dog face, four legs)

✗ (dog face + four legs, dog face - four legs)

Empirical Studies of IOSS

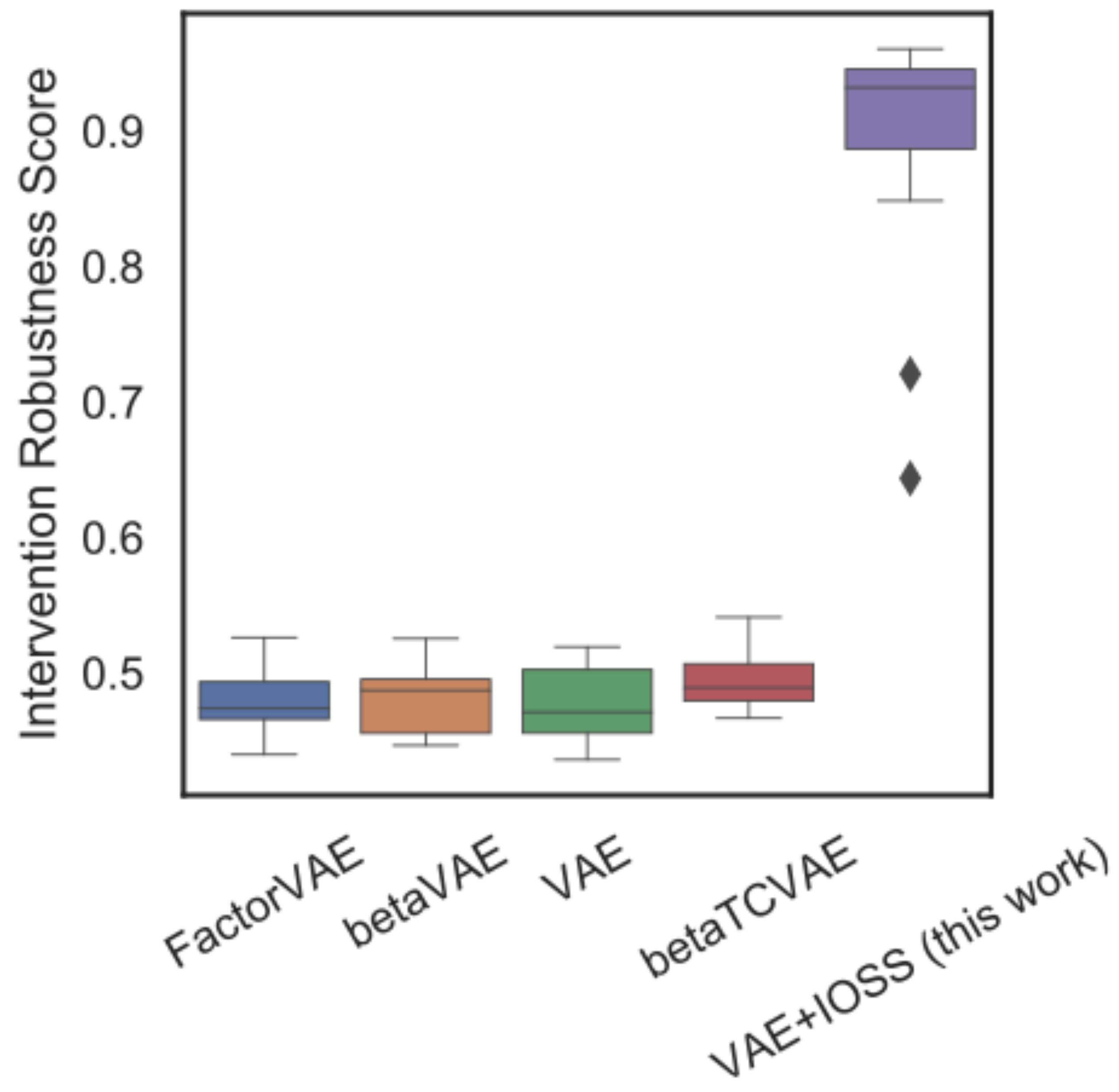
Measure Disentanglement with IOSS



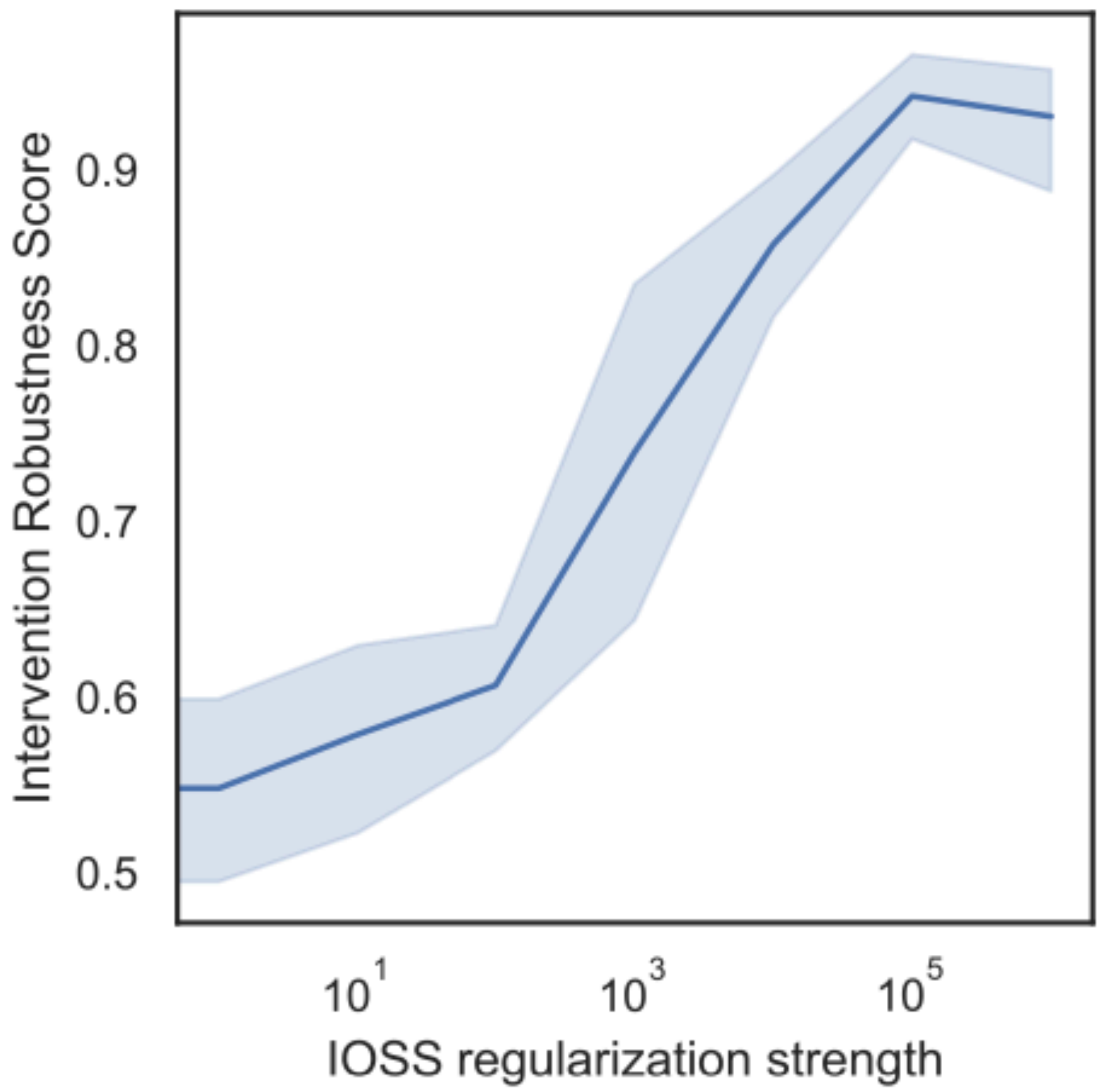
(a) IOSS (b) Total Correlation (c) Wasserstein Dependency

Figure 10: IOSS can better distinguish entangled and disentangled representations than existing unsupervised disentanglement metrics on the mpi3d dataset.

Learning Disentangled Representations with IOSS



(a) Disentanglement of IOSS learned representations



(b) Regularization with IOSS penalty

Takeaways

- Many **desiderata for representation learning** can be formalized using **causal** notions.
    - **Non-spuriousness and efficiency** (Supervised); **Disentanglement** (Unsupervised)
- They lead to **metrics** to measure how desirable the representations are, and **algorithms** that directly target desired representations. (All derivations are from the first principles.)
- Empirical studies of CAUSAL-REP and IOSS reveal **possibilities** of learning non-spurious/ disentangled representations **without** multiple environments/invariance/auxiliary labels.
- Causal inference, though challenging in general, may be **tractable** in machine learning tasks. (We define what success is :-)

Thank you!

- Y. Wang and M.I. Jordan
  Desiderata for Representation Learning: A Causal Perspective
  arXiv:2109.03795

- https://github.com/yixinwang/representation-causal-public

Non-spuriousness

Definition 1 (Non-spuriousness of representations). *Suppose we observe a data point with representation $\boldsymbol{Z} = \boldsymbol{z}$ and label $Y = y$. Then the non-spuriousness of the representation $\boldsymbol{Z}$ for label Y is the probability of sufficiency* (PS) *of $\mathbb{I}\{\boldsymbol{Z} = \boldsymbol{z}\}$ for $\mathbb{I}\{Y = y\}$:*

$$\text{PS}_{\boldsymbol{Z}=\boldsymbol{z}, Y=y} = P(Y(\boldsymbol{Z} = \boldsymbol{z}) = y \mid \boldsymbol{Z} \neq \boldsymbol{z}, Y \neq y). \tag{1}$$

When both the representation Z and the label Y are univariate binary with $z = 1, y = 1$, then Equation (1) *coincides with classical definition of* PS *(Definition 9.2.2 of Pearl (2011)).*

Efficiency

Definition 2 (Efficiency of representations). *Suppose we observe a data point with representation* $\boldsymbol{Z} = \boldsymbol{z}$ *and label* $Y = y$*. Then the efficiency of the representation* $\boldsymbol{Z}$ *for the label* Y *is the probability of necessity (*PN*) of* $\mathbb{I}\{\boldsymbol{Z} = \boldsymbol{z}\}$ *for* $\mathbb{I}\{Y = y\}$*:*[2]

$$\text{PN}_{\boldsymbol{Z}=\boldsymbol{z},Y=y} = P(Y(\boldsymbol{Z} \neq \boldsymbol{z}) \neq y \,|\, \boldsymbol{Z} = \boldsymbol{z}, Y = y). \tag{2}$$

When both the representation Z *and the label* Y *are univariate binary with* $z = 1, y = 1$*, then Equation* (2) *coincides with classical definition of* PN *(Definition 9.2.1 of* Pearl (2011)*).*

Efficiency and Non-spuriousness

Definition 3 (Efficiency & non-spuriousness of representations). *Suppose we observe a data point with representation $\boldsymbol{Z} = \boldsymbol{z}$ and label $Y = y$. Then the efficiency and non-spuriousness of the representation $\boldsymbol{Z}$ for label Y is the probability of necessity and sufficiency (PNS) of $\mathbb{I}\{\boldsymbol{Z} = \boldsymbol{z}\}$ for $\mathbb{I}\{Y = y\}$:*

$$\text{PNS}_{\boldsymbol{Z}=\boldsymbol{z},Y=y} = P(Y(\boldsymbol{Z} \neq \boldsymbol{z}) \neq y, Y(\boldsymbol{Z} = \boldsymbol{z}) = y)). \tag{3}$$

When both the representation Z and the label Y are univariate binary with $z = 1, y = 1$, then Equation (3) *coincides with classical definition of* PNS *(Definition 9.2.3 of Pearl (2011)).*

Requiring both necessity and sufficiency of the cause is a stronger requirement than requiring only necessity (or only sufficiency). Accordingly, PNS is a weighted combination of PN and PS,

$$\text{PNS}_{\boldsymbol{Z}=\boldsymbol{z},Y=y} = P(\boldsymbol{Z} = \boldsymbol{z}, Y = y) \cdot \text{PN}_{\boldsymbol{Z}=\boldsymbol{z},Y=y} + P(\boldsymbol{Z} \neq \boldsymbol{z}, Y \neq y) \cdot \text{PS}_{\boldsymbol{Z}=\boldsymbol{z},Y=y},$$

Conditional Efficiency and Non-spuriousness

Extension: Conditional efficiency and non-spuriousness. For multi-dimensional representations, one is often interested in the efficiency and non-spuriousness of each of its dimensions. We expect each dimension of the representation to be efficient and non-spurious conditional on all other dimensions.

We thus extend Definition 3 to formalize a notion of *conditional efficiency and non-spuriousness.* Consider a d-dimensional representation $\boldsymbol{Z} = (Z_1, \ldots, Z_d) = (f_1(\boldsymbol{X}), \ldots, f_d(\boldsymbol{X}))$. The conditional efficiency and non-spuriousness of the jth dimension Z_j for data point $(\boldsymbol{x}_i, y_i)$ is

$$\mathrm{PNS}_{Z_j=z_{ij}, Y=y_i \,|\, \boldsymbol{Z}_{-j}=\boldsymbol{z}_{i,-j}} = P(Y(Z_j \neq z_{ij}, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{i,-j}) \neq y_i, Y(Z_j = z_{ij}, \boldsymbol{Z}_{-j} = \boldsymbol{z}_{i,-j}) = y_i), \tag{5}$$

where $z_{ij} = f_j(\boldsymbol{x}_i)$ is the jth dimension of the representation, and $z_{i,-j} = (z_{ij'})_{j' \in \{1,\ldots,d\} \setminus j}$. Accordingly, the conditional efficiency and non-spuriousness of Z_j across all n data points is

$$\mathrm{PNS}_n(Z_j, Y \mid \boldsymbol{Z}_{-j}) \triangleq \prod_{i=1}^{n} \mathrm{PNS}_{\boldsymbol{Z}=\boldsymbol{z}_i, Y=y_i \,|\, \boldsymbol{Z}_{-j}=\boldsymbol{z}_{i,-j}}. \tag{6}$$

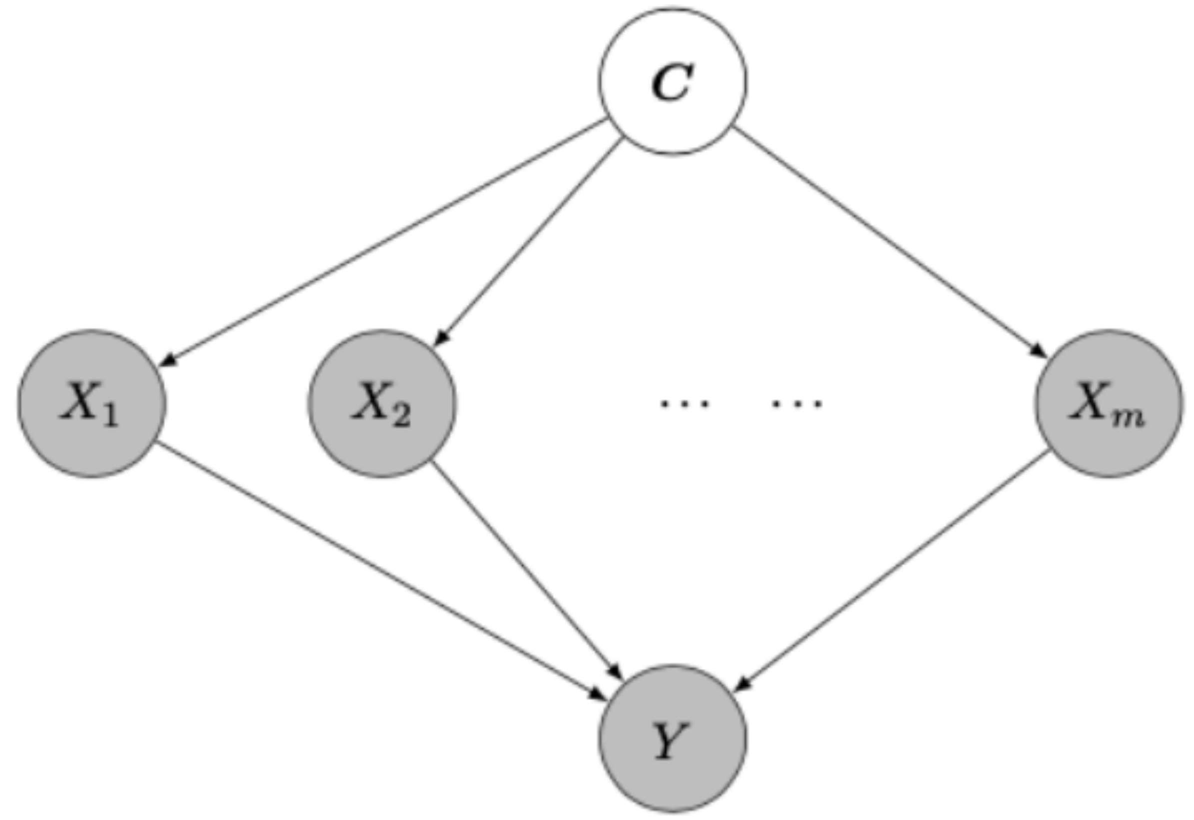
How do we maximize PNS?

Lemma 4 (A lower bound on PNS). *Assuming the causal graph in Figure 2, the PNS is lower bounded by the difference between two intervention distributions:*

$$\begin{aligned} \mathrm{PNS}_{\boldsymbol{Z}=\boldsymbol{z},Y=y} &= P(Y(\boldsymbol{Z}=\boldsymbol{z})=y, Y(\boldsymbol{Z}\neq\boldsymbol{z})\neq y) \\ &\geq P(Y=y \mid \mathrm{do}(\boldsymbol{Z}=\boldsymbol{z})) - P(Y=y \mid \mathrm{do}(\boldsymbol{Z}\neq\boldsymbol{z})). \end{aligned} \tag{7}$$

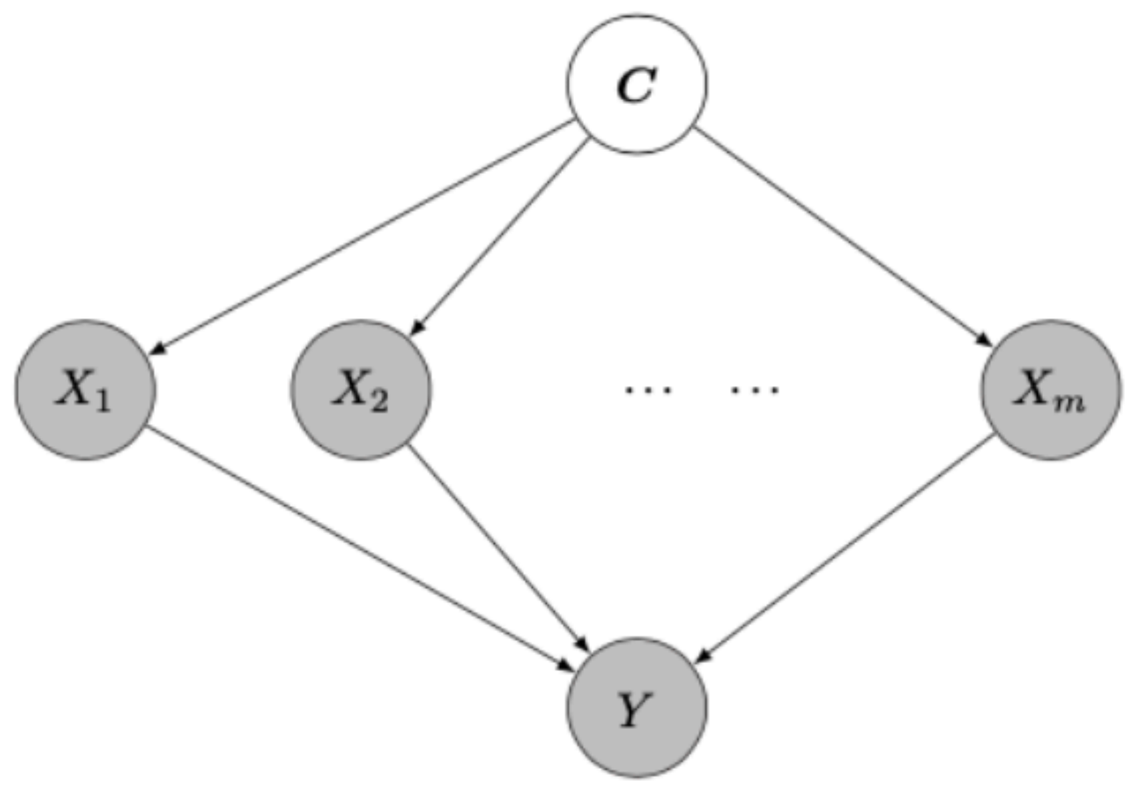
The inequality becomes an equality when the outcome Y is monotone in the representation $\boldsymbol{Z}$ (in the binary sense); i.e., $P(Y(\boldsymbol{Z}=\boldsymbol{z})\neq y, Y(\boldsymbol{Z}\neq\boldsymbol{z})=y)=0$.

How do we maximize PNS?



- Identifying the intervention distribution $P(Y = y \mid \mathrm{do}(Z = z))$
  - Functional interventions $P(Y = y \mid \mathrm{do}(Z = z)) = P(Y = y \mid \mathrm{do}(f(X) = z))$
    - Conditional on all parents of X, manipulate X such that $f(X) = z$
    - $P(Y = y \mid \mathrm{do}(f(X) = z)) = \int P(Y = y \mid \mathrm{do}(X = x))P(X = x \mid f(X) = z, C)P(C)\mathrm{d}C;$
  - Need to pinpoint the unobserved common cause C;
  - High-dimensional X living on low dimensional manifold; restrict to subvectors of X

How do we maximize PNS?



Definition 5 (Functional interventions (Puli et al., 2020)). *The intervention distribution under a functional intervention* $P(Y \mid \mathrm{do}(f(\boldsymbol{X}) = \boldsymbol{z}))$ *is defined as*

$$P(Y \mid \mathrm{do}(f(\boldsymbol{X}) = \boldsymbol{z})) \triangleq \int P(Y \mid \mathrm{do}(\boldsymbol{X}), \boldsymbol{C}) P(\boldsymbol{X} \mid \boldsymbol{C}, f(\boldsymbol{X}) = \boldsymbol{z}) P(\boldsymbol{C}) \, \mathrm{d}\boldsymbol{X} \, \mathrm{d}\boldsymbol{C}, \tag{8}$$

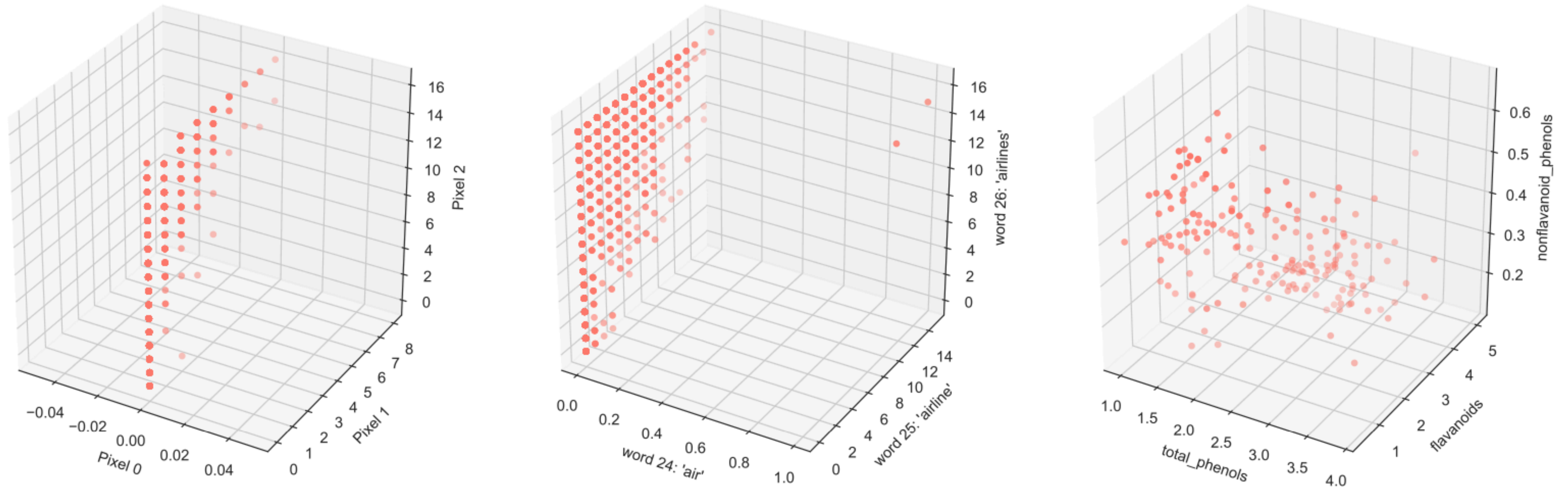
where $\boldsymbol{C}$ *denotes all parents of* $\boldsymbol{X}$.

Following this definition, one can write the intervention distribution of interest, $P(Y \mid \mathrm{do}(f(\boldsymbol{X}) = \boldsymbol{z}))$, as follows:

$$P(Y \mid \mathrm{do}(f(\boldsymbol{X}) = \boldsymbol{z})) = \int P(Y \mid \boldsymbol{X}) \cdot \left[\int P(\boldsymbol{X} \mid \boldsymbol{C}, f(\boldsymbol{X}) = \boldsymbol{z}) P(\boldsymbol{C}) \, \mathrm{d}\boldsymbol{C} \right] \mathrm{d}\boldsymbol{X}. \tag{9}$$

This equality is due to the SCM in Figure 2: there is no unobserved confounding between $\boldsymbol{X}$ and Y, which implies $P(Y \mid \mathrm{do}(\boldsymbol{X}), \boldsymbol{C}) = P(Y \mid \boldsymbol{X})$.

How do we maximize PNS?



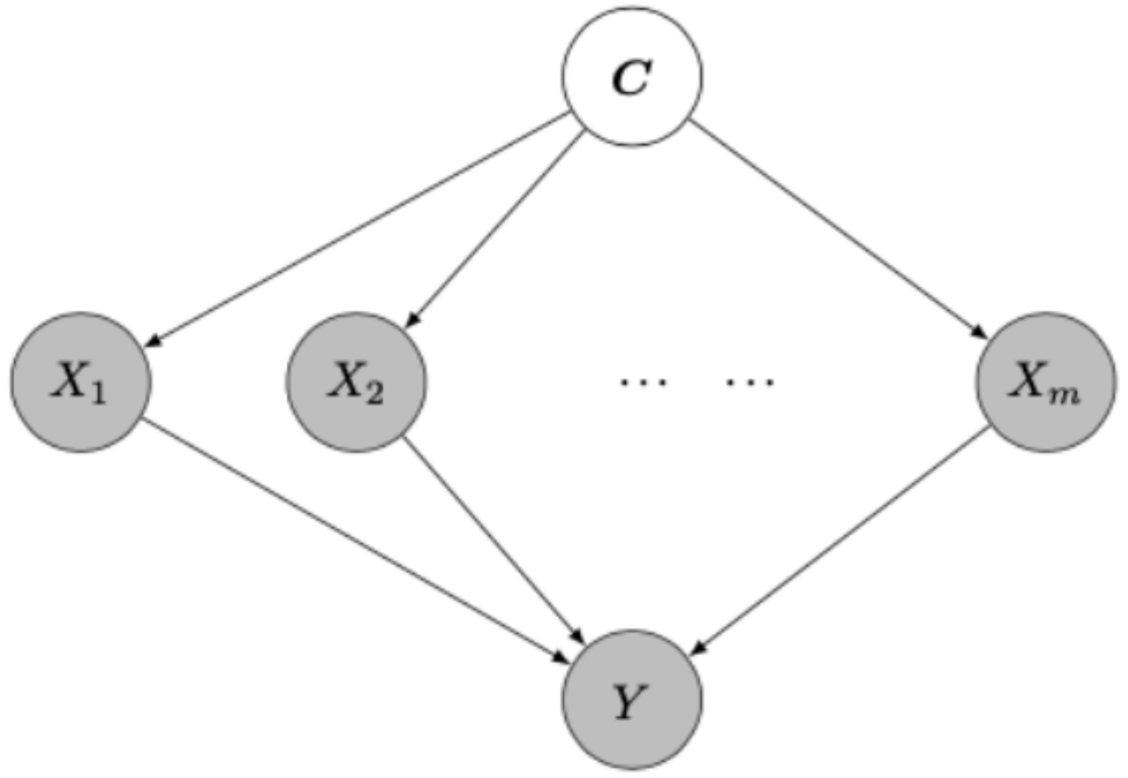
(a) High-dim. image data: MNIST (Deng, 2012)

(b) High-dim. text data: Airline tweets

(c) Low-dim. data: Wine features

As a more concrete example, consider a high-dimensional vector of image pixels $\boldsymbol{X}$ that lives on a low-dimensional manifold; i.e., such that $X_j - g_0(\{X_1, \ldots, X_m\} \backslash X_j)$ is identically zero in the observational data (Goodfellow et al., 2014; Kingma & Welling, 2014). This rank degeneracy implies that for any $p(y \mid \boldsymbol{x}) = h_0(\boldsymbol{x}, y)$ compatible with the observational data distribution, the conditional $p(y \mid \boldsymbol{x}) = h_0(\boldsymbol{x}, y) + \alpha \cdot (x_j - g_0(\{x_1, \ldots, x_m\} \backslash x_j)), \forall \alpha \in \mathbb{R}$, is also compatible with the observational data.

How do we maximize PNS?



Causal identification of $P(Y \mid \mathrm{do}(f(\boldsymbol{X}))$ for a restricted set of f. Given the fundamental non-identifiability of $P(Y \mid \boldsymbol{X})$ with high-dimensional $\boldsymbol{X} = (X_1, \ldots, X_m)$, we restrict our attention to representations that only nontrivially depends on a "full-rank" subset; i.e., $\boldsymbol{Z} = f(\boldsymbol{X}) = \tilde{f}((X_j)_{j \in S})$, for some function $\tilde{f} : \mathcal{X}^{|S|} \to \mathbb{R}^d$, and a set $S \subseteq \{1, \ldots, m\}$, where $p((x_j)_{j \in S}) > 0$ for all values $(x_j)_{j \in S} \in \mathcal{X}^{|S|}$. We term this requirement "observability."

Focusing on such representations $f(\boldsymbol{X}) = \tilde{f}((X_j)_{j \in S})$, we calculate its intervention distributions by returning to the definition of functional interventions (Definition 5),

$$P(Y \mid \mathrm{do}(f(\boldsymbol{X}) = \boldsymbol{z})) = \int P(Y \mid (X_j)_{j \in S}, \boldsymbol{C}) P((X_j)_{j \in S} \mid \boldsymbol{C}, f(\boldsymbol{X}) = \boldsymbol{z}) P(\boldsymbol{C}) \, \mathrm{d}(X_j)_{j \in S} \, \mathrm{d}\boldsymbol{C}.$$

How do we maximize PNS?

Lemma 6 (Identification of $P(Y \mid \mathrm{do}(f(\boldsymbol{X}) = \boldsymbol{z}))$. *Assume the causal graph in Figure 2. Suppose the representation only effectively depends on a subset* $(X_j)_{j \in S}$ *of* $(X_1, \ldots, X_m)$*; i.e.,* $f(\boldsymbol{X}) = \tilde{f}((X_j)_{j \in S})$ *for some function* $\tilde{f} : \mathcal{X}^{|S|} \to \mathbb{R}^d$ *and some set* $S \subseteq \{1, \ldots, m\}$*. Then the intervention distribution* $P(Y \mid \mathrm{do}(f(\boldsymbol{X}) = \boldsymbol{z}))$ *is identifiable by*

$$P(Y \mid \mathrm{do}(f(\boldsymbol{X}) = \boldsymbol{z})) = \int P(Y \mid f(\boldsymbol{X}) = \boldsymbol{z}, h(\boldsymbol{X})) \cdot P(h(\boldsymbol{X})) \, \mathrm{d}h(\boldsymbol{X}), \tag{12}$$

if the following conditions are satisfied:

1. *(pinpointability) the unobserved common cause* $\boldsymbol{C}$ *is pinpointable; i.e.,* $P(\boldsymbol{C} \mid \boldsymbol{X}) = \delta_{h(\boldsymbol{X})}$ *for a deterministic function* h *known up to bijective transformations,*
2. *(positivity)* $(X_j)_{j \in S}$ *satisfies the positivity condition given* $\boldsymbol{C}$*; i.e.,* $P((X_j)_{j \in S} \in \widetilde{\mathcal{X}} \mid \boldsymbol{C}) > 0$ *for any set* $\widetilde{\mathcal{X}} \subset \mathcal{X}^{|S|}$ *such that* $P((X_j)_{j \in S} \in \widetilde{\mathcal{X}}) > 0$*,*
3. *(observability)* $P((X_j)_{j \in S} \in \widetilde{\mathcal{X}}) > 0$ *for all subsets* $\widetilde{\mathcal{X}} \subset \mathcal{X}^{|S|}$ *with a positive measure.*

Causal Disentanglement $\Rightarrow$ Independent Support

Theorem 9 (Disentanglement $\Rightarrow$ Independent support). *Assume the unobserved common cause* $\boldsymbol{C}$ *satisfies a positivity condition: for all* j*, we have* $P(Z_j \mid \boldsymbol{C}) > 0$ *iff* $P(Z_j) > 0$*. Then the support of the interventional distribution coincides with that of the observational distribution:*

$$\textit{supp}(Z_j \mid \mathrm{do}(Z_{j'} = z_{j'})) = \textit{supp}(Z_j \mid Z_{j'} = z_{j'}), \tag{41}$$

where $j, j' \in \{1, \ldots, d\}$, $j \neq j'$*, and the density at* $z_{j'}$ *is nonzero,* $p(z_{j'}) > 0$*. As a consequence, different dimensions of a disentangled representation* $\boldsymbol{Z} = (Z_1, \ldots, Z_d)$ *must have independent support:*

$$\begin{aligned} \textit{supp}(Z_1, \ldots, Z_d) &= \textit{supp}(Z_1) \times \cdots \times \textit{supp}(Z_d), \\ \textit{supp}(Z_j \mid Z_{\mathcal{S}}) &= \textit{supp}(Z_j) \text{ for all } \mathcal{S} \subseteq \{1, \ldots, d\} \backslash j. \end{aligned} \tag{42}$$

Independence-of-Support Score (IOSS)

Definition 10 (Independence-of-support score (IOSS)). *Suppose a representation $\boldsymbol{Z}$ has bounded support and* $\sup Z_j - \inf Z_j > 0, j = 1, \ldots, d$. *Then the* IOSS *of* $\boldsymbol{Z}$ *is the Hausdorff distance between the joint support of* $(Z_1, \ldots, Z_d)$ *and the product of each individual's support:*

$$
\begin{aligned}
&\textsc{ioss}(Z_1, \ldots, Z_d) \\
&\triangleq d_H(\mathit{supp}(\bar{Z}_1, \ldots, \bar{Z}_d), \mathit{supp}(\bar{Z}_1) \times \cdots \mathit{supp}(\bar{Z}_d)) \\
&= d(\mathit{supp}(\bar{Z}_1) \times \cdots \mathit{supp}(\bar{Z}_d), \mathit{supp}(\bar{Z}_1, \ldots, \bar{Z}_d)),
\end{aligned}
$$

where $\bar{Z}_j = (Z_j - \inf Z_j)/(\sup Z_j - \inf Z_j)$ *is the standardized* Z_j, *and* $d_H(\cdot, \cdot)$ *is the Hausdorff distance.*[9] *The second equality is due to* $\mathit{supp}(Z_1, \ldots, Z_d) \subseteq \mathit{supp}(\bar{Z}_1) \times \cdots \times \mathit{supp}(\bar{Z}_d)$.

Identifiability of Representations with Independent Support

Theorem 11 (Identifiability of representations with independent support). *Among all compactly supported representations (i.e. the support being a closed and bounded region) that generate the same σ-algebra, the representation with independent support (if exists) is identifiable up to permutation and coordinate-wise bijective transformations: for any two d-dimensional representations, $\boldsymbol{Z} = f(\boldsymbol{X}) = (Z_1, \ldots, Z_d)$ and $\boldsymbol{Z}' = f'(\boldsymbol{X}) = (Z'_1, \ldots, Z'_d)$, such that (1) f, f' are continuous, (2) $\sigma(\boldsymbol{Z}) = \sigma(\boldsymbol{Z}')$, (3) $\boldsymbol{Z}, \boldsymbol{Z}'$ both satisfy the independent support condition (Equation (42)), and (3) $\boldsymbol{Z}, \boldsymbol{Z}'$ both have compact support in $\mathbb{R}^d$, we have*

$$Z_1, \ldots, Z_d = \mathrm{perm}(q_1(Z'_1), \ldots, q_d(Z'_d)),$$

where the q_j are continuous bijective function with a compact domain in $\mathbb{R}$. (The proof is in Appendix J.)

Identifiability of Representations with Independent Support

To understand the intuition behind Theorem 11, we consider a toy example of a two-dimensional compactly supported representation (Z_1, Z_2) with independent support: $Z_1 \in [1, 2]$, $Z_2 \in [0, 2]$. Next consider an entanglement of this representation (Z_1', Z_2'), which is a bijective transformation of (Z_1, Z_2):

$$Z_1' = Z_1 + Z_2, \qquad Z_2' = Z_1 - Z_2.$$

We will show that (Z_1', Z_2') does not have independent support, then the support of $Z_1 - Z_2$ depends on the value of $Z_1 + Z_2$. To see why, consider the case when $Z_1 + Z_2 = 4$, then we must have $Z_1 = Z_2 = 2$ due to the support constraints on Z_1, Z_2. Hence $Z_1 - Z_2 = 0$, thus the support of $Z_1 - Z_2$ is $\{0\}$. Following a similar argument, the support of $Z_1 - Z_2$ is $\{1\}$ when $Z_1 + Z_2 = 1$. Therefore, the support of $Z_1 - Z_2$ depends on values of $Z_1 + Z_2$, and hence they have dependent support.