CSDS 452 Causality and Machine Learning

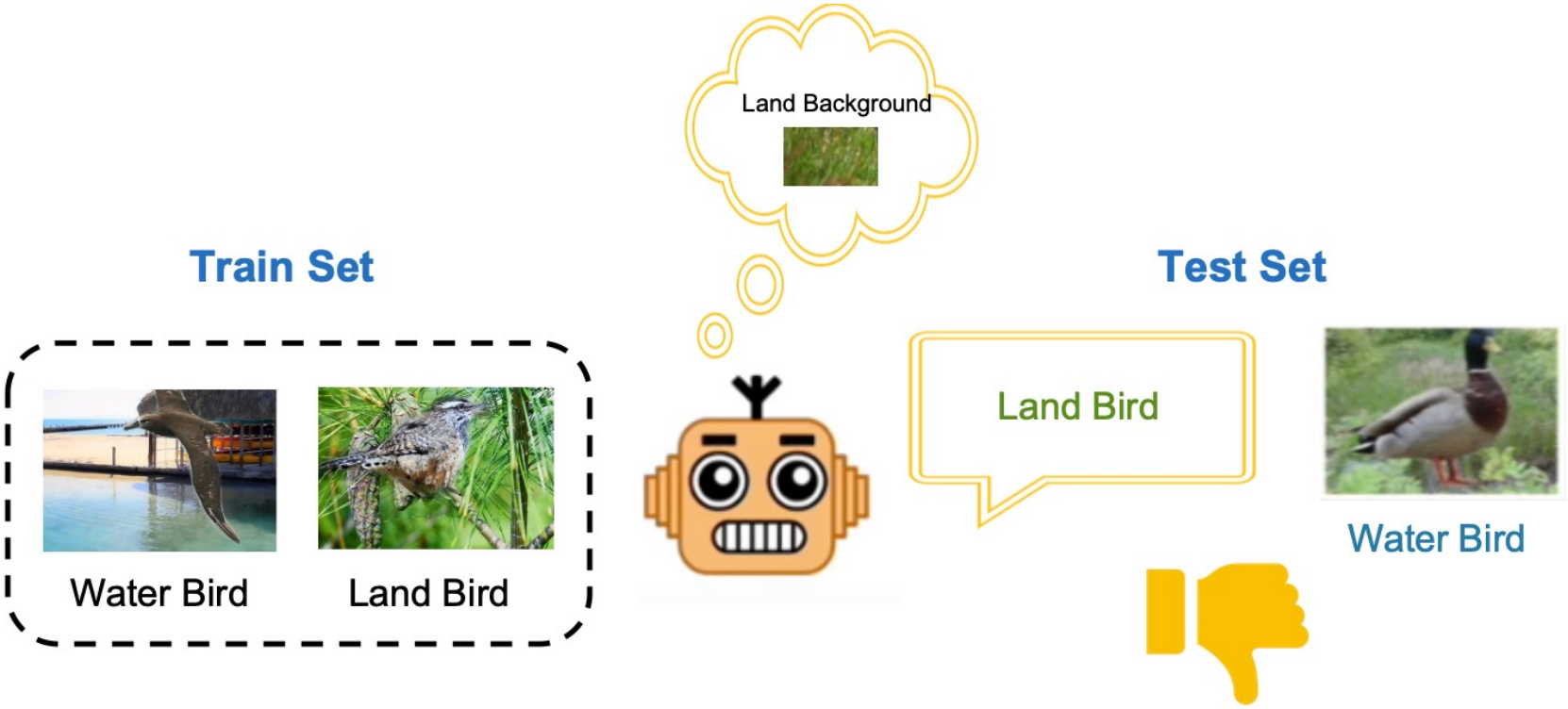
Lecture 16: Causal Generalization (continue)

Instructor: Jing Ma

Fall 2024, CDS@CWRU

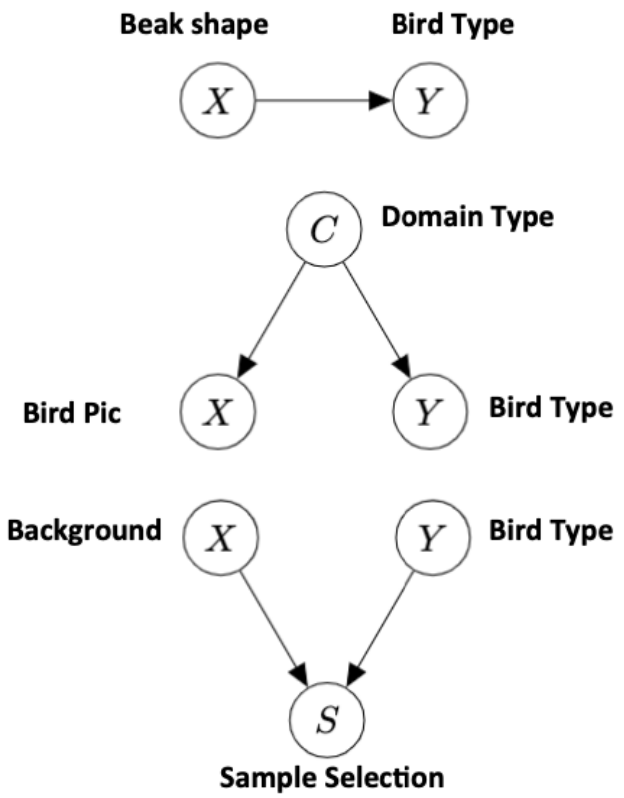
Recap: Generalization issues

- Machine learning models are prone to spurious correlations.



Spurious correlation and causality

- There can be three sources of correlation :
  - **Causation**
    - Causal mechanism
  - **Confounding**
    - Spurious correlation through C
  - **Selection Bias**
    - Spurious correlation through S

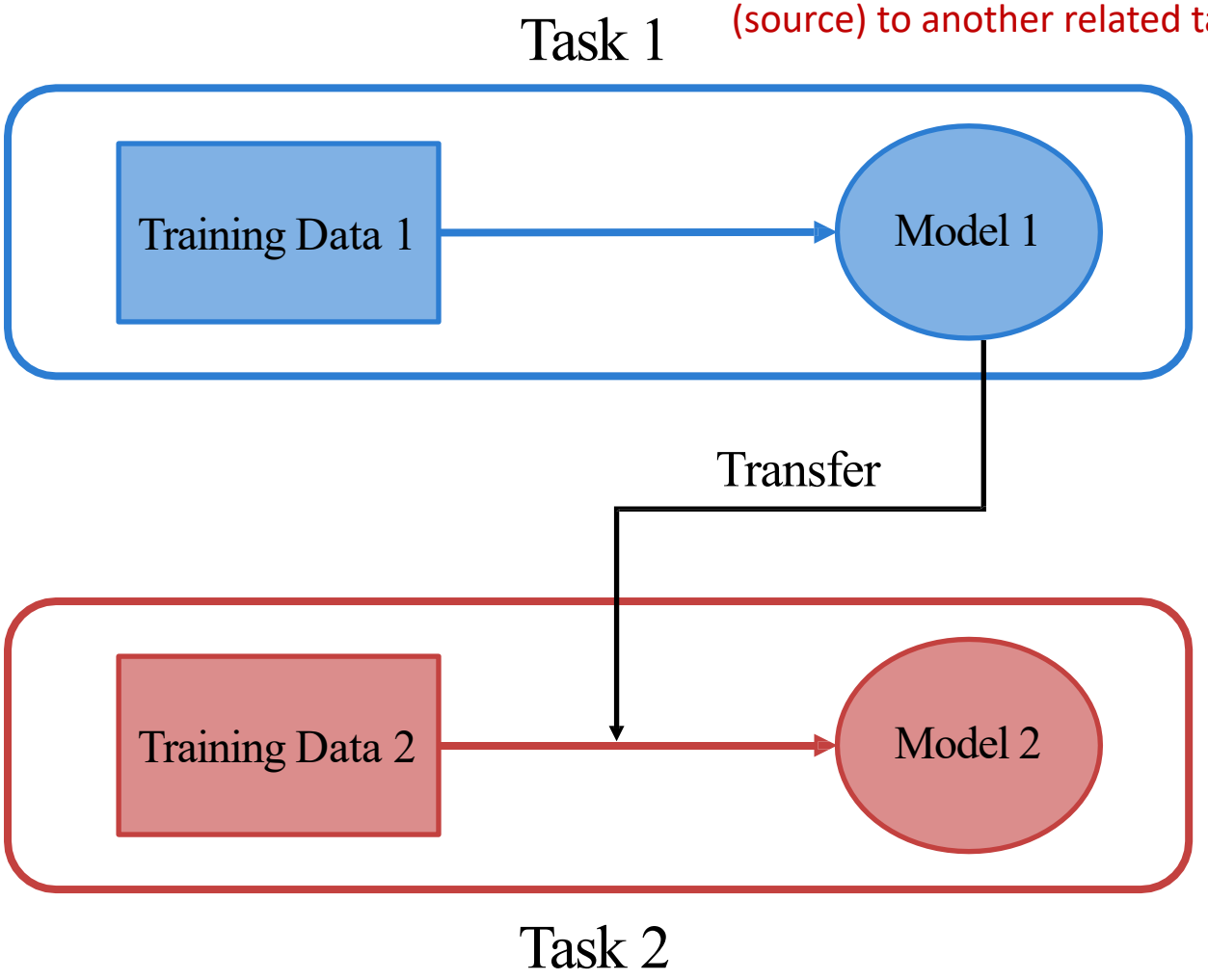


Transfer learning, domain adaptation, and domain generalization

- These words are often interchangeably used in many papers.
- But there may be some slight differences…

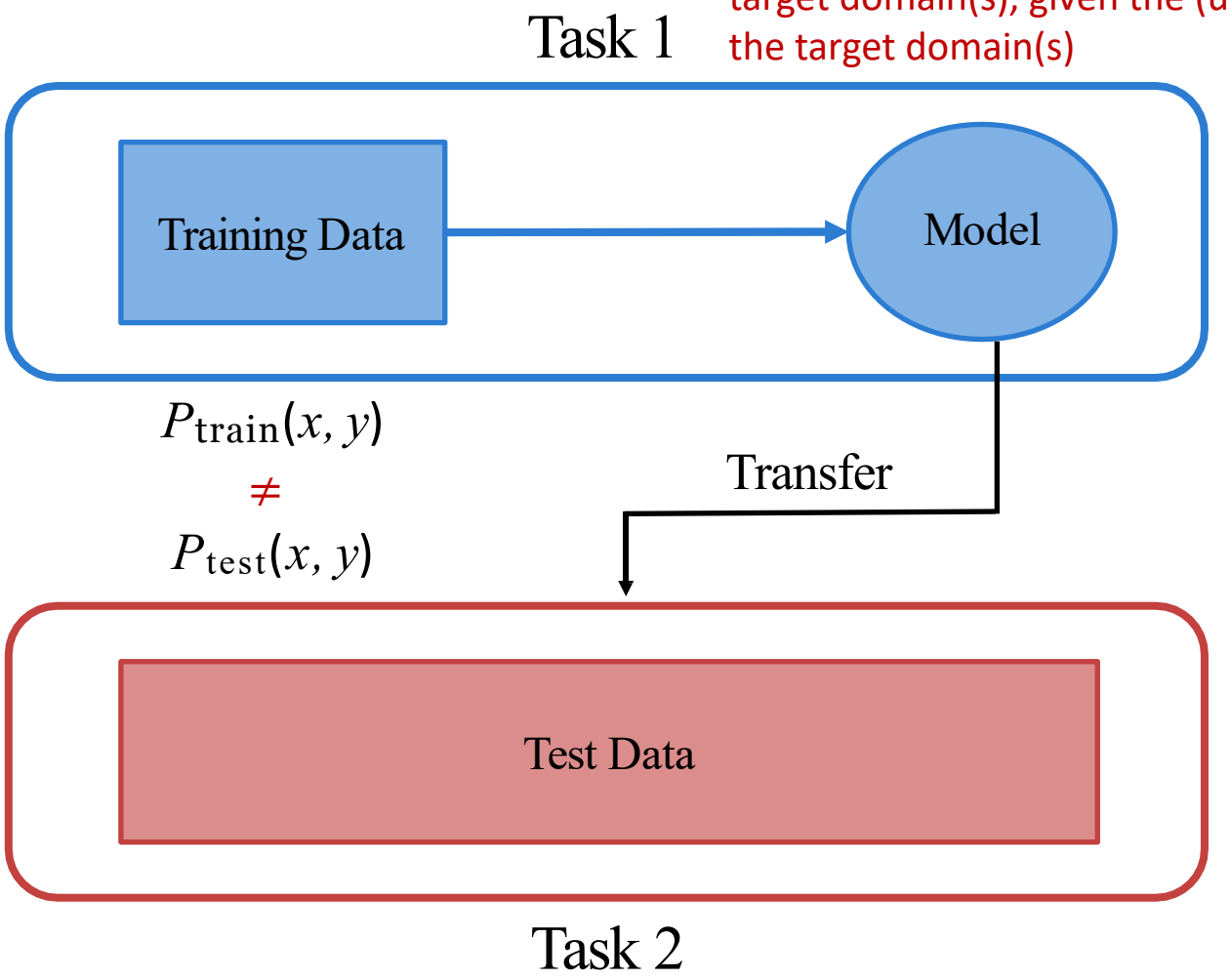
Transfer Learning

Adapt a model from one specific task or domain (source) to another related task or domain (target)



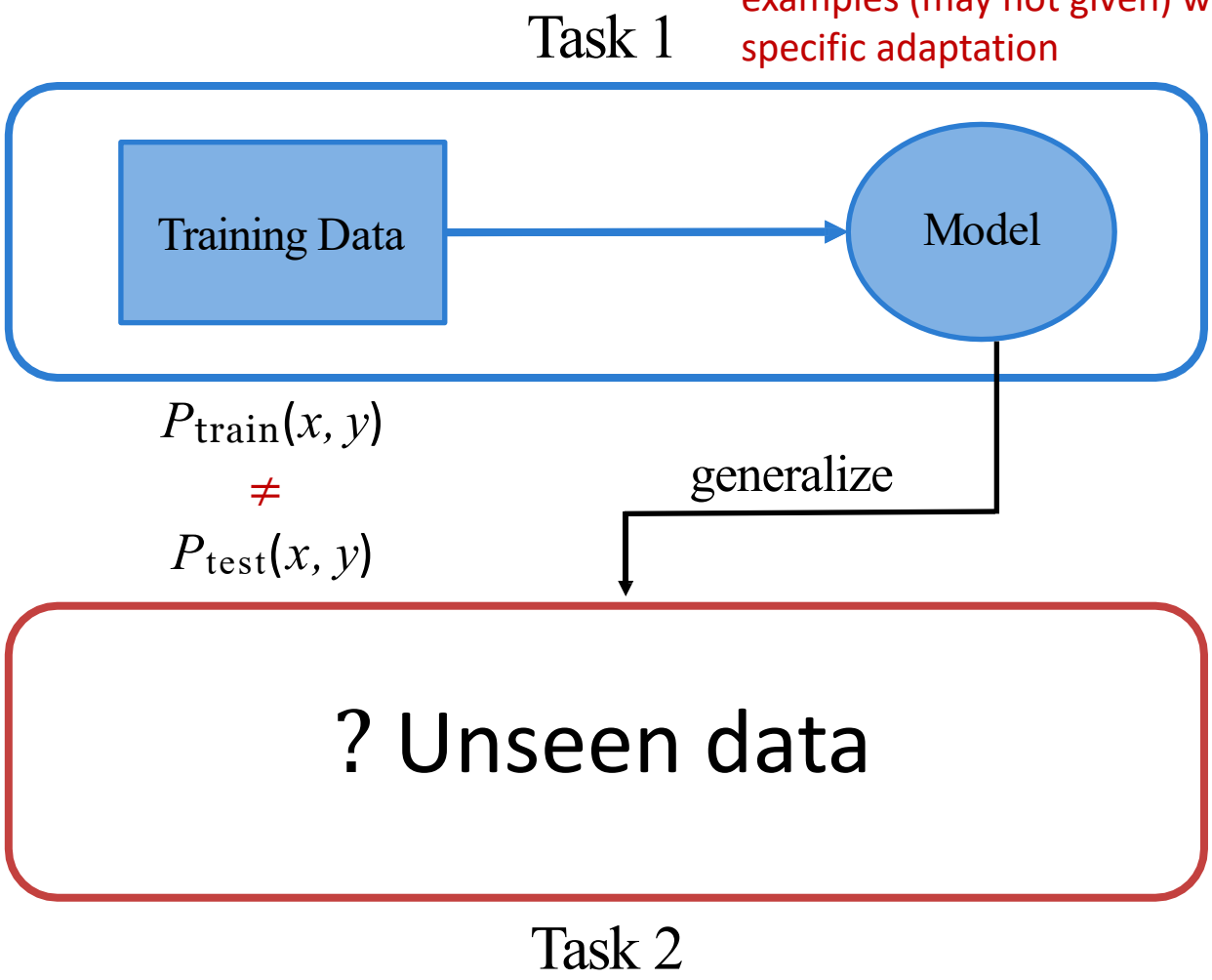
Domain Adaptation

Adapt a model trained from source domain(s) to target domain(s), given the (unlabeled) data in the target domain(s)



Domain Generalization

Train a model that works well in unseen examples (may not given) without task/domain-specific adaptation



Distribution shift

- Distribution shift:

$$P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y)$$

- The joint distribution $P(x, y)$ can be decomposed as:

$$P(x, y) = P(x)P(y|x) = P(y)P(x|y)$$

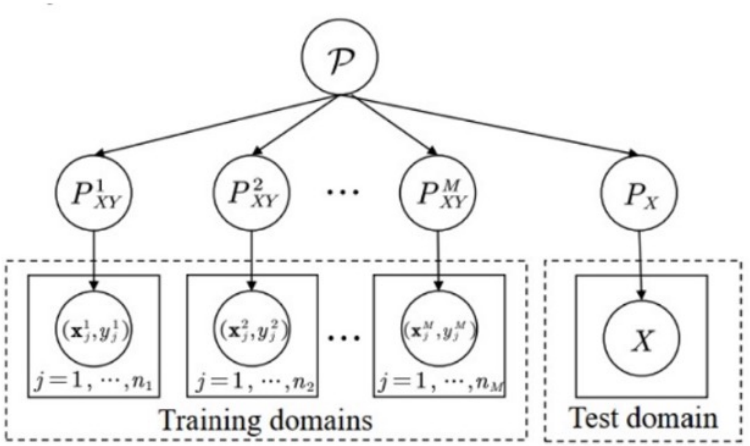
- Which part is changed in different domains?

Domain Generalization: Problem Statement

- Given M training domains $\varepsilon = \{\varepsilon_i | i = 1, \dots, M\}$, where $\varepsilon_i = \{(x_j^i, y_j^i)\}_{j=1}^n$, the goal is to learn a classifier that achieves minimum error on an *unknown* test domain:

$$\min_h \mathbb{E}_{(x,y)\in test}[l(h(x), y)],$$

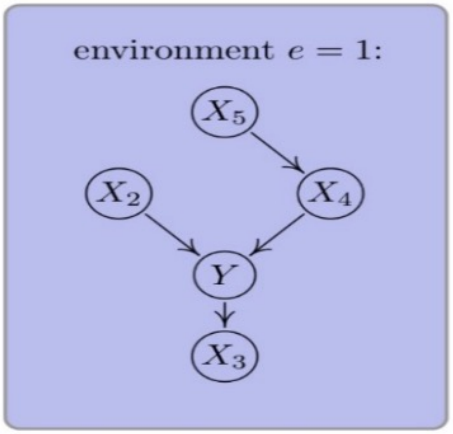
where $p_{XY}^i \neq p_{XY}^{test}$.

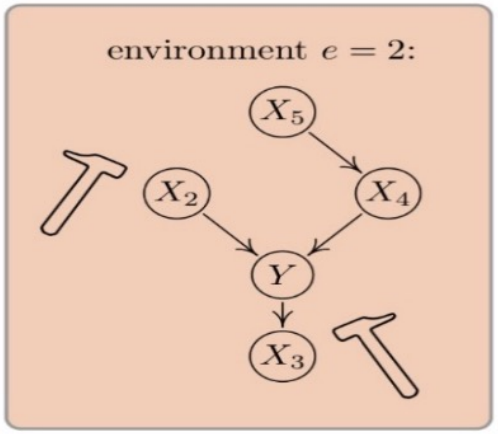


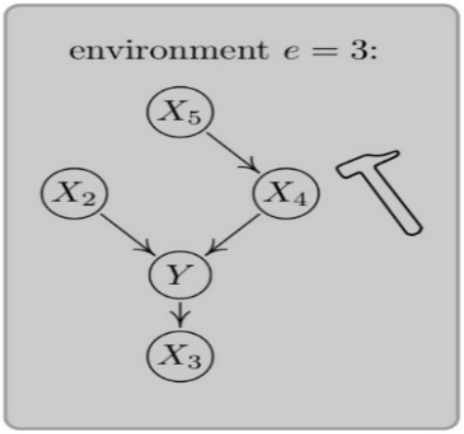
[Wang et al., 2021]

A causal solution

- Assumption: data is sampled from interventional distributions governed by the same SCM.





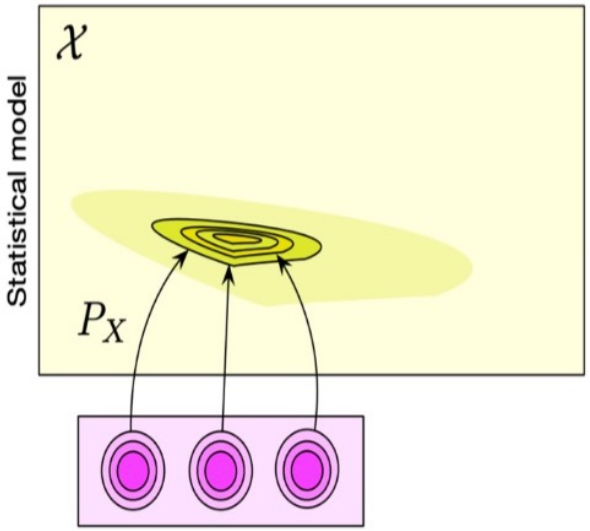


On the grass

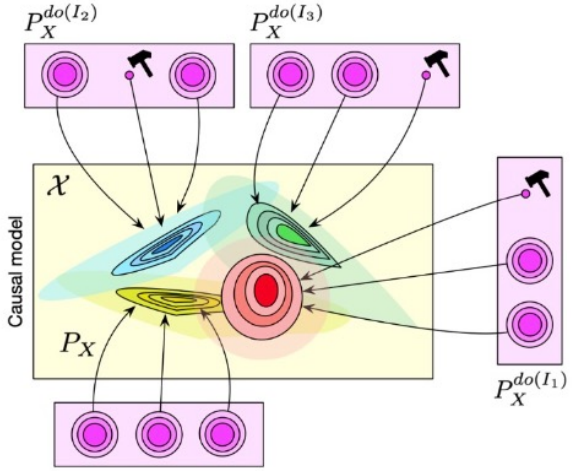
Eating

On the beach

A causal solution



A statistical model specifies a single probability distribution

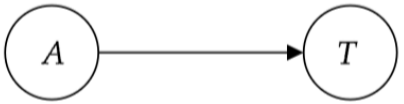


A causal model represents a set of distributions, one for each possible intervention (indicated with a).

[Schölkopf and Locatello et al., 2021]

Independent Causal Mechanisms (ICM) Principle

- We know that altitude causes the change in the temperature.



Causal factorization: $P(a,t) = P(a) \times P(t|a)$

- Changing P(a) does not affect the P(t|a):
  - $P^L(a,t) = P^L(a) \times P(t|a)$. Long Beach, California
  - $P^S(a,t) = P^S(a) \times P(t|a)$. Saarbrücken, Germany

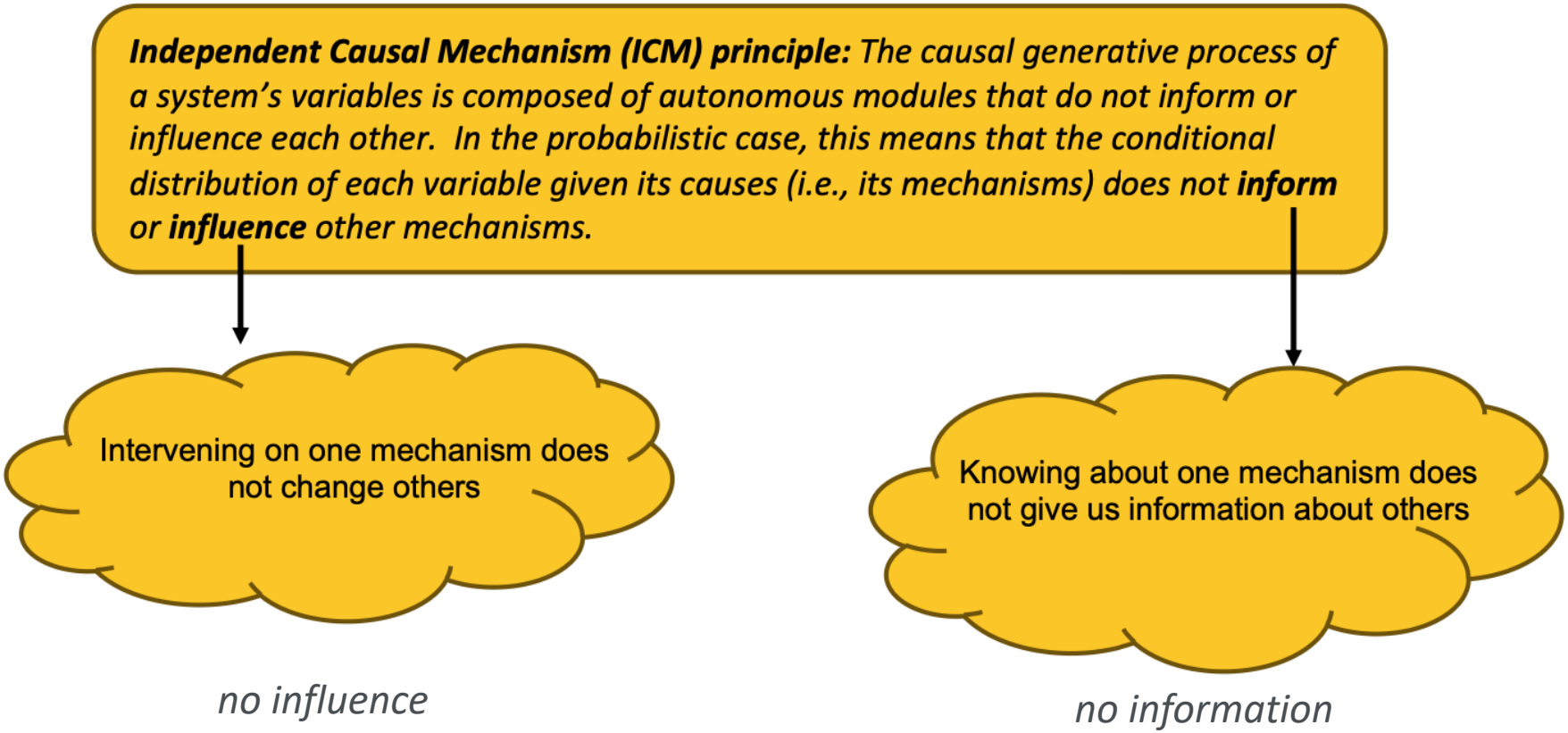
This does not hold for a non-causal factorization, i.e.,

$$P^L(a,t) = P^L(t) \times P^L(a|t)$$
$$P^S(a,t) = P^S(t) \times P^S(a|t)$$

ICM Principle

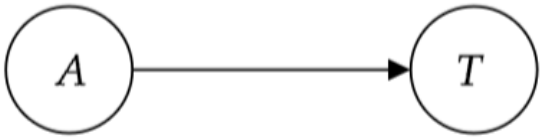
Independent Causal Mechanism (ICM) principle: *The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanisms) does not* ***inform*** *or* ***influence*** *other mechanisms.*

ICM Principle



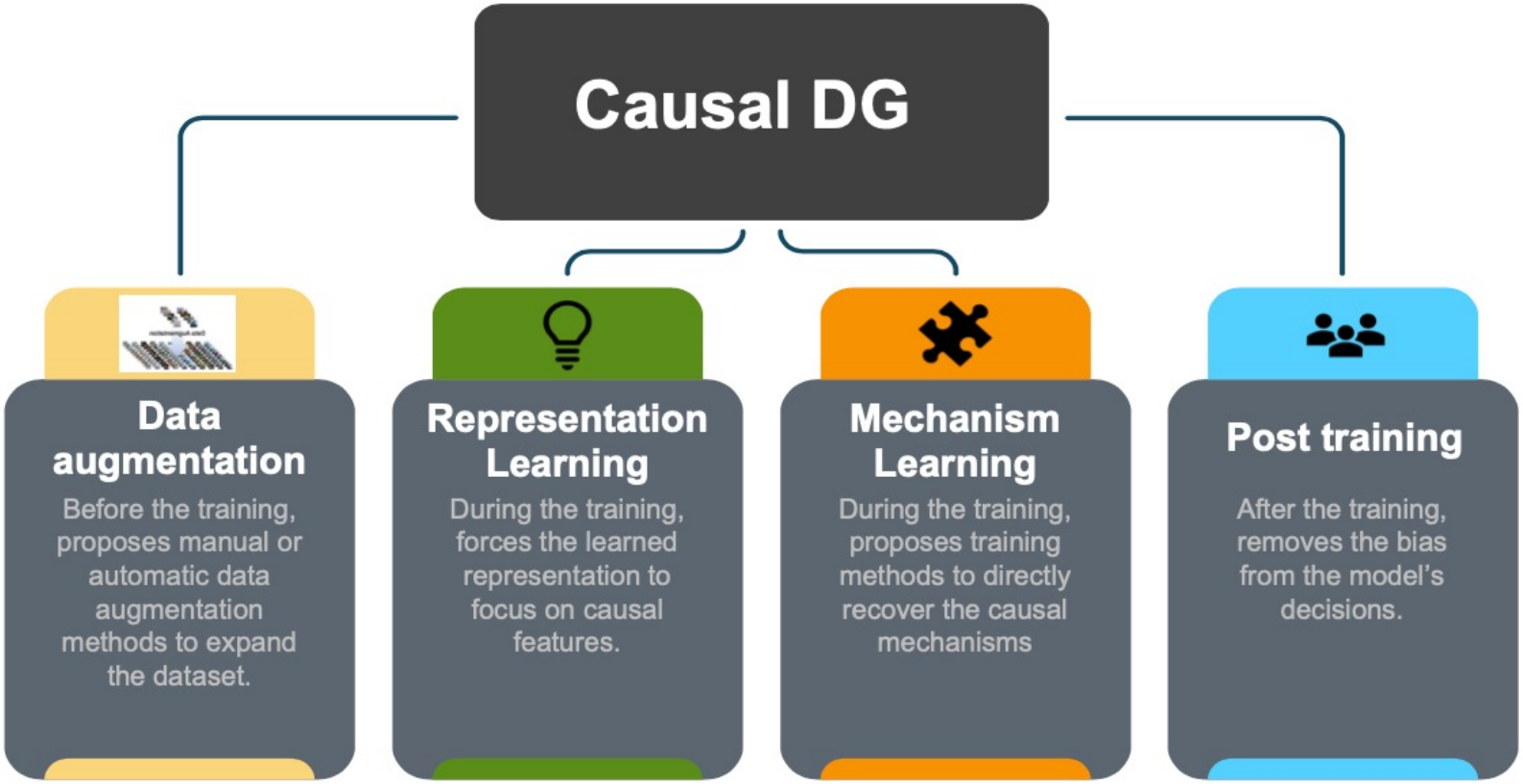
ICM Principle

- ICM is a central concept in the study of causality with implications for ML tasks
- In the previous example:



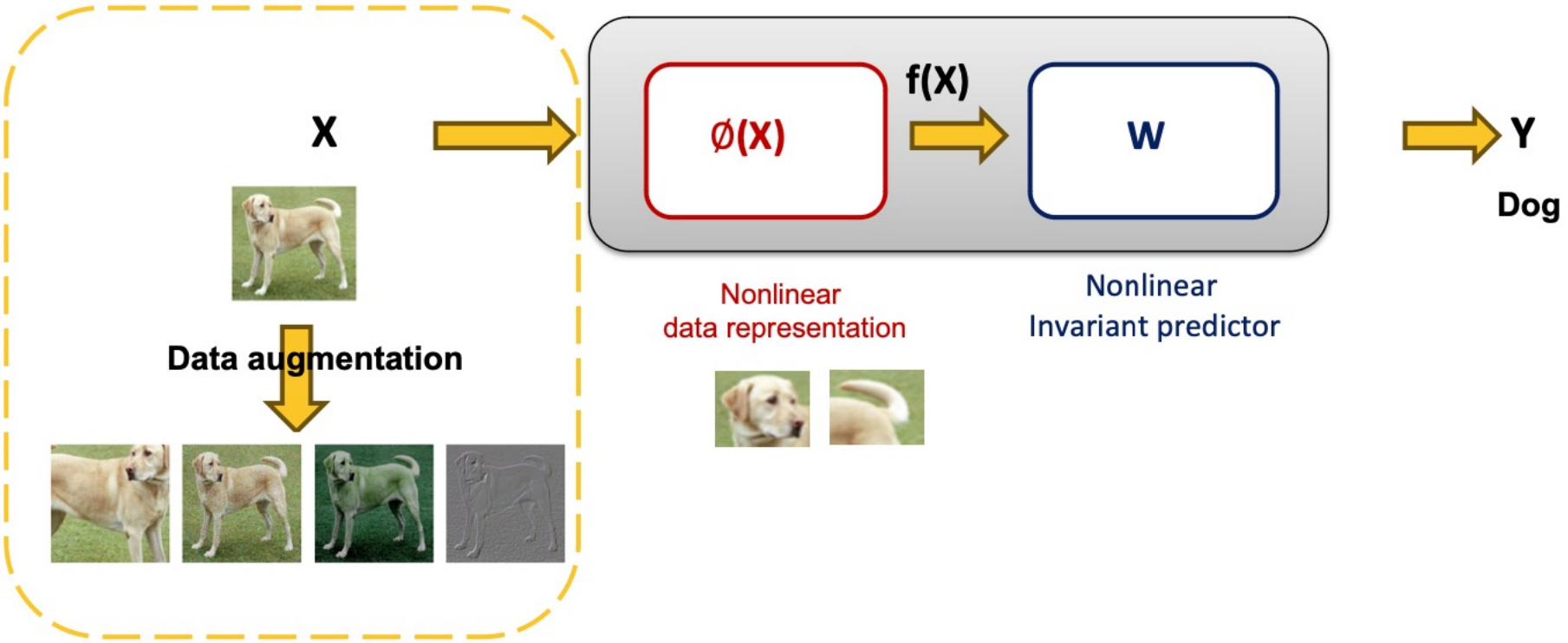
  - *no influence*: implies that if you change the altitude then the temperature P(T|A) still remains the same. That is, P(T|A) **generalizes** well.
  - *no information*: implies that knowing the temperature at each given altitude, it will not tell anything about the location.

Overview of the approaches

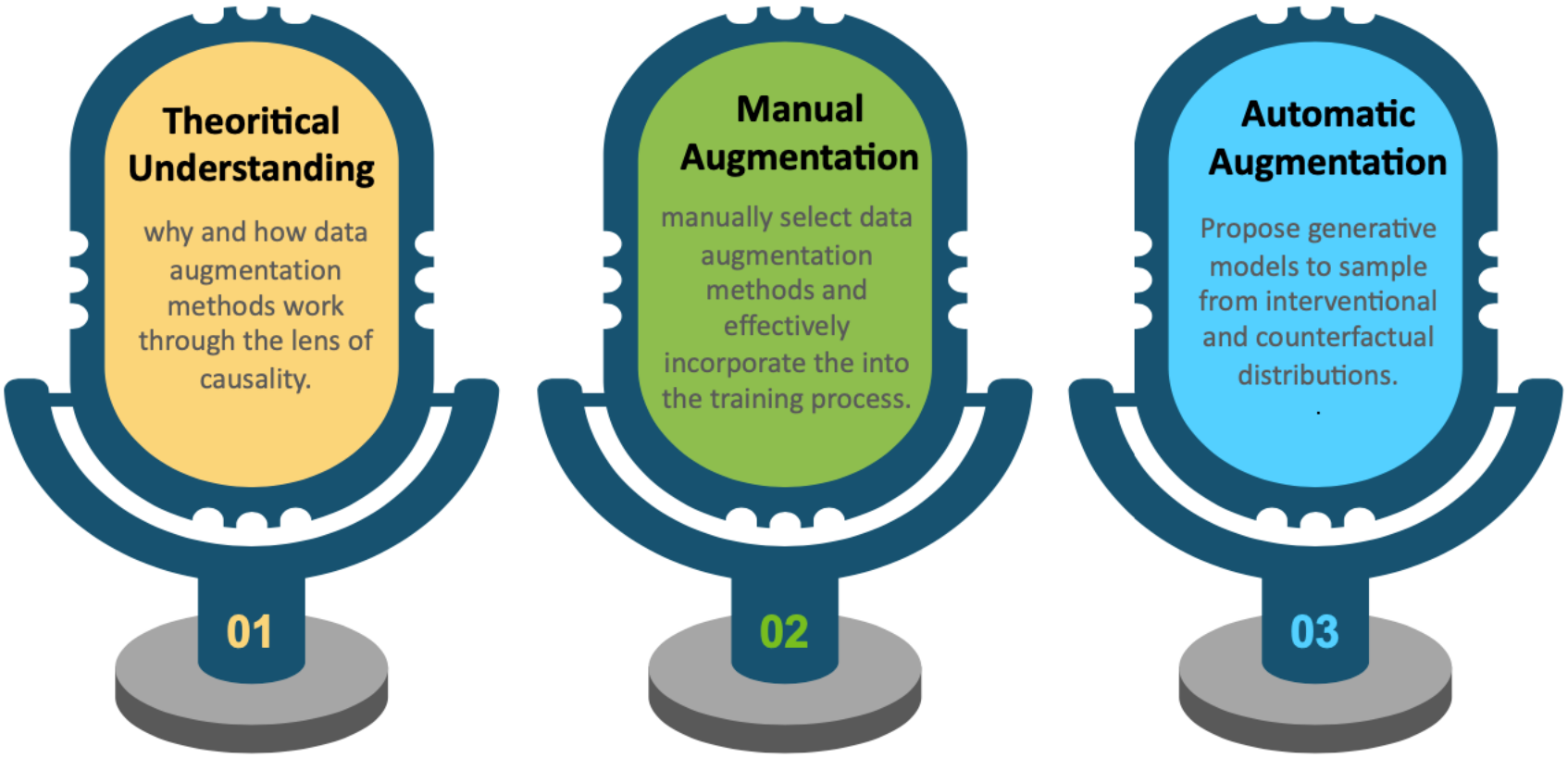


[Sheth, Moraffah et al., 2021]

Causal Data Augmentation



Causal Data Augmentation: categories



Causal Data Augmentation: Theoretical Analysis

- Goal: Provide a causal perspective on why and how data augmentation improves generalization.
- Steps:
  - Define a causal graph on the data generating process
  - Utilize causality concepts to explain why data augmentation is helpful/ how the data augmentation methods must be selected to be effective

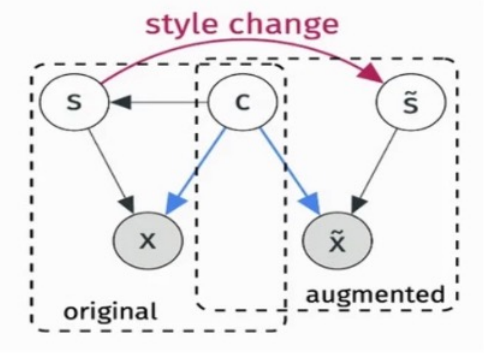
Causal Data Augmentation: Why does it help?

- Suppose the latent representation z is decomposed into two disjoint block:

$$c = z_{1:nc} \quad , s = z_{nc+1:n} , \quad z=(c, s)$$

  - The content block c is invariant and shared across $(x,\tilde{x})$
  - The style block S may vary across $(x,\tilde{x})$
  - In practice this content-style partition is implicitly

defined by a set of transformations τ applied through data augmentation.



Content-style partition

[Kügelgen et al., 2021]

Causal Data Augmentation: A causal view

- In our causal graph, the following causal relationship $c \rightarrow s$ is observed. Formalizing the SCM given the observation x = f(c, s), the data augmentation equivalent to ask the counterfactual question:

  "what would have happened, had the style variables been randomly perturbed?"

Causal Data Augmentation: A causal view

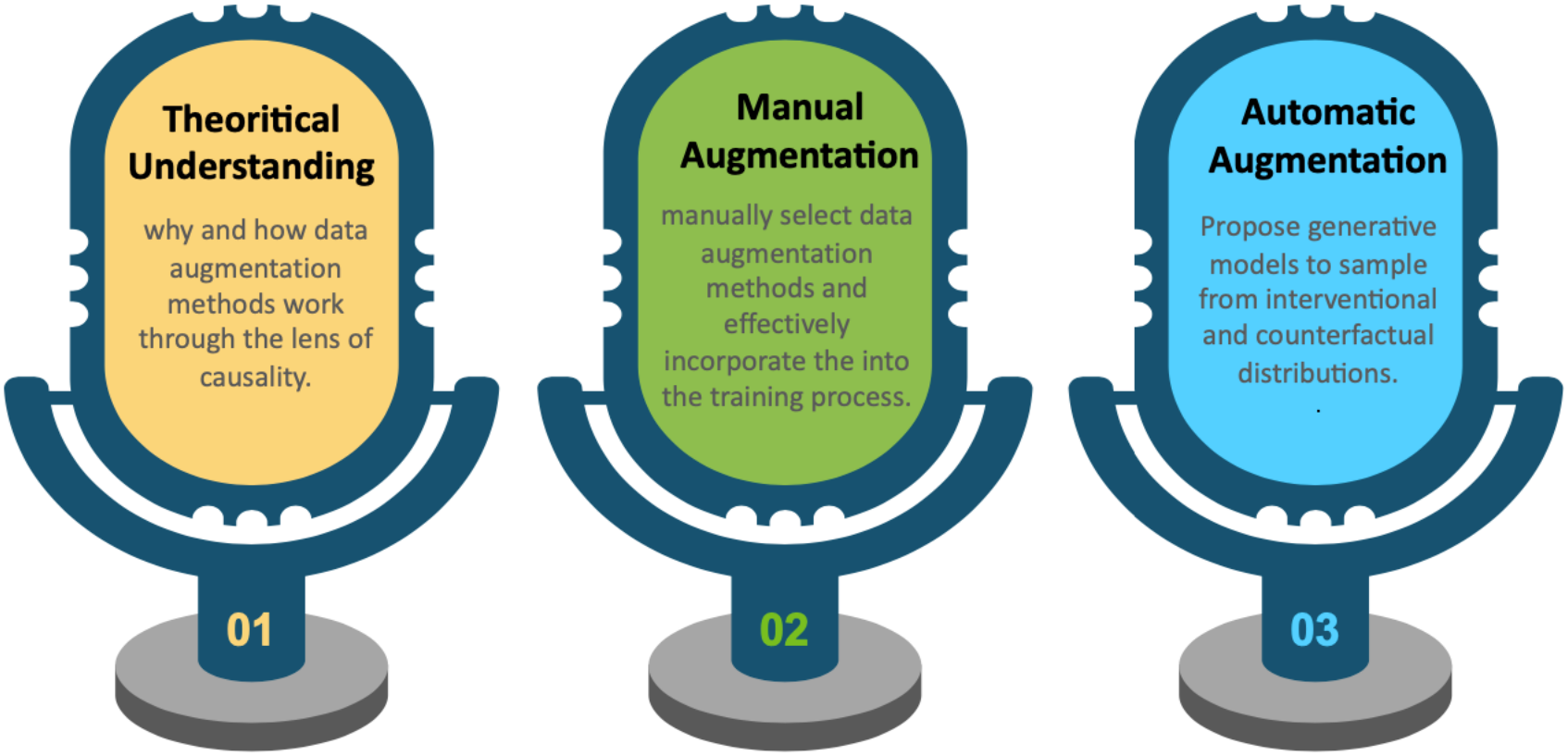
- In our causal graph, the following causal relationship $c \rightarrow s$ is observed. Formalizing the SCM given the observation x = f(c, s), the data augmentation equivalent to ask the counterfactual question:

  "what would have happened, had the style variables been randomly perturbed?"

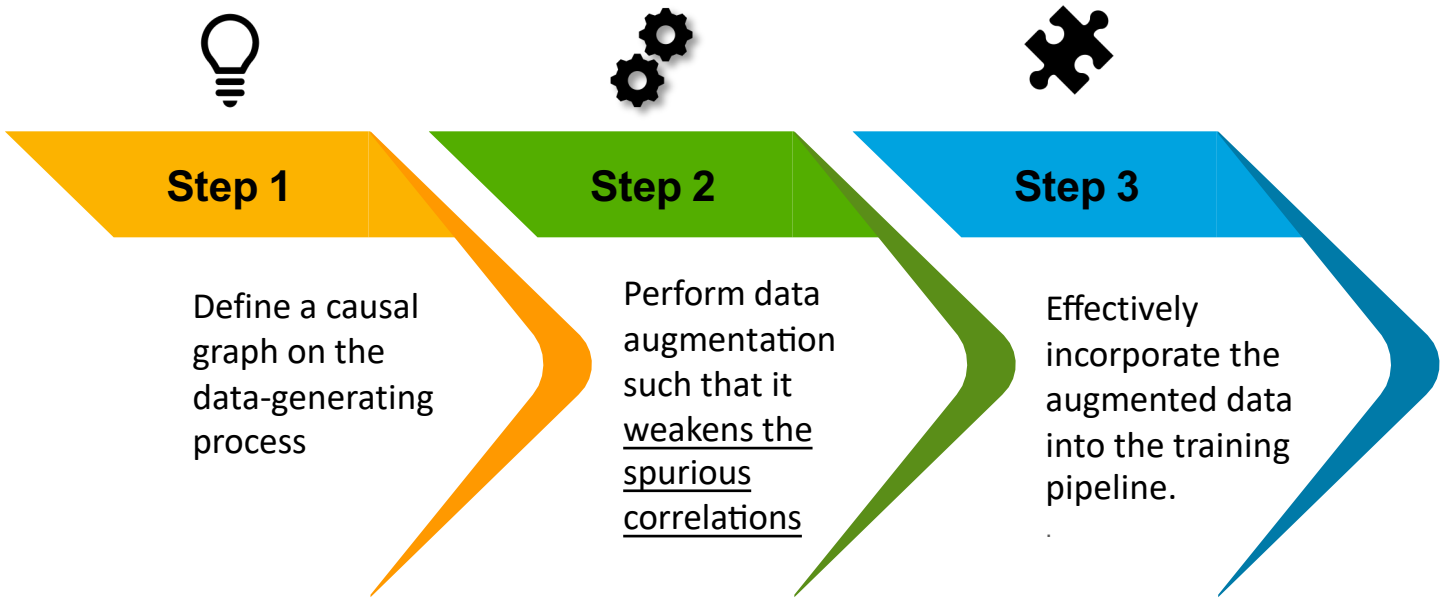
- The soft style intervention do (s=$\tilde{f}_s\ (c, u_s, u_A)$), where u_A is an additional source of stochasticity such as random transformation.

Counterfactual Interpretation of Data Augmentation. The augmented view $\tilde{x} = f(c, \tilde{s})$ corresponds to a counterfactual under soft style interventions on style factors such as color, Corruption, etc.

Causal Data Augmentation: categories

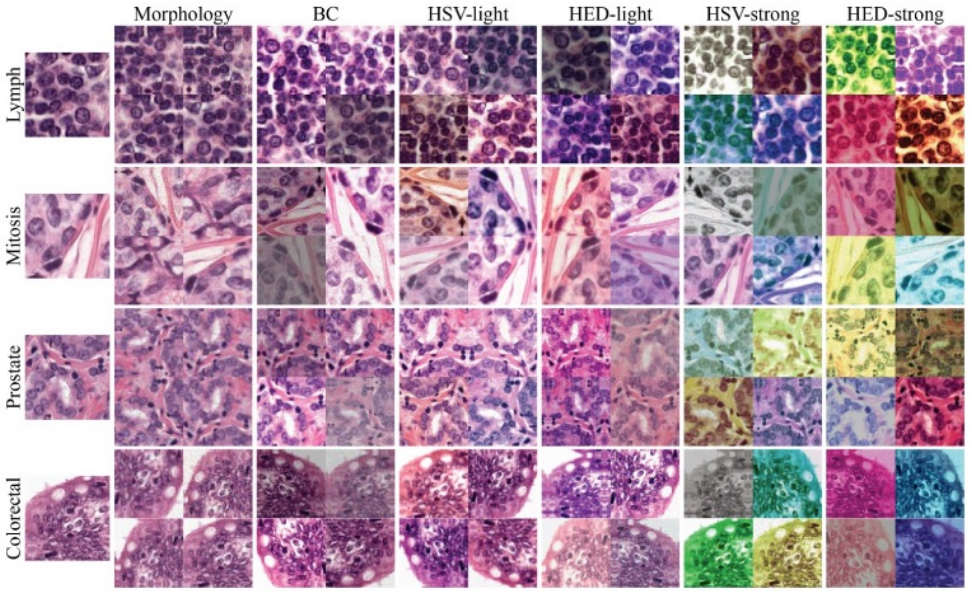


Manual Data Augmentation



Causal Data Augmentation: How to select augmentations?

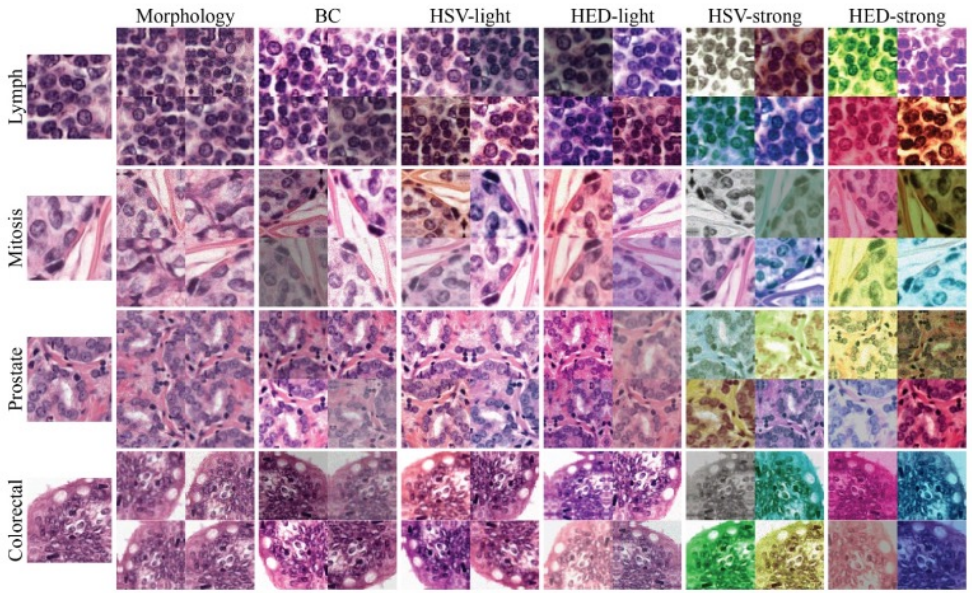
- A case study of data augmentation methods for generalizability



[Ilse et al., 2021]

Causal Data Augmentation: How to select augmentations?

- A case study of data augmentation methods for generalizability

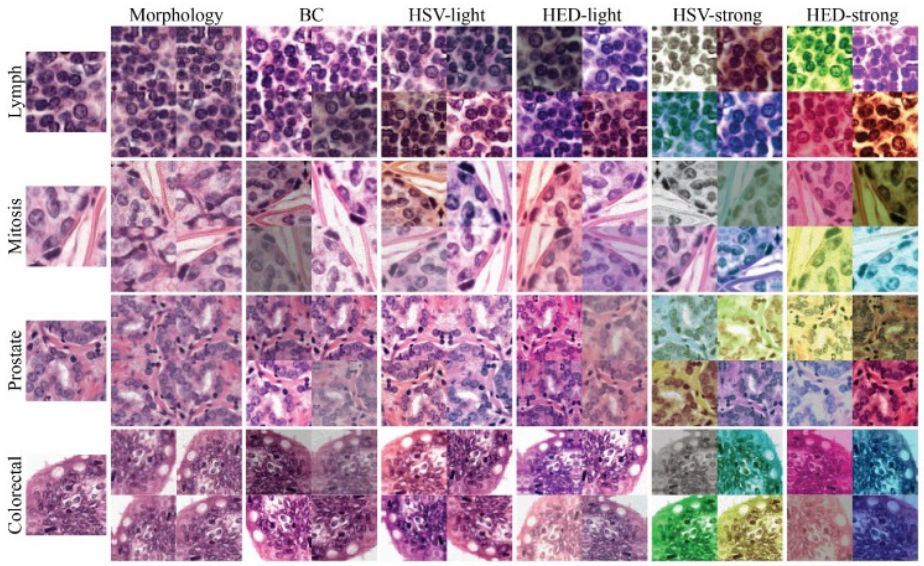


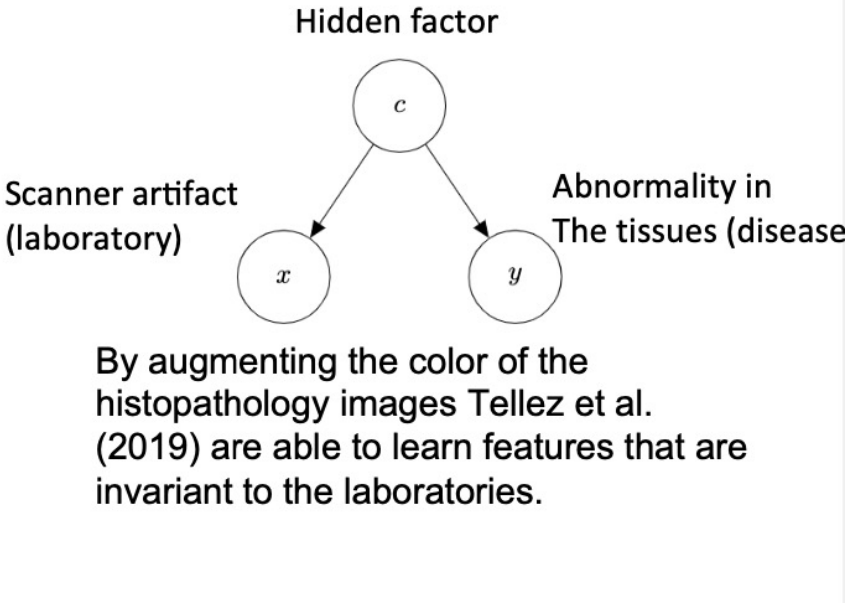
Morphology. alterations in shape, texture or size of the imaged tissue structures, including scanning artifacts.
Brightness & contrast (BC). random brightness and contrast image perturbations.
Hue-Saturation-Value (HSV). randomly shifts the hue and saturation channels in the HSV color space, results in substantial change in the color.
Hematoxylin-Eosin-DAB (HED). a color variation routine specifically designed for H&E images.

[Ilse et al., 2021]

Causal Data Augmentation: A case study

- A case study of data augmentation methods for generalizability





By augmenting the color of the histopathology images Tellez et al. (2019) are able to learn features that are invariant to the laboratories.

A selection algorithm: Select Data Augmentation (SDA)

Step 1: start with a list of data augmentation methods

A selection algorithm: Select Data Augmentation (SDA)

Step 1: start with a list of data augmentation methods

Step 2: Train a classifier to predict the domain d from input x. apply the first data augmentation in the list to the samples of the training set. Store the domain accuracy on the validation set after training. Repeat this step with all data augmentations in the list.

Step 3: Select the data augmentation with the lowest domain

accuracy averaged over five seeds.

An example of SDA on rotate-MNIST dataset

Step1

- 'brightness': torchvision.transforms.ColorJitter(brightness=1.0, contrast=0, saturation=0, hue=0)
- 'contrast': torchvision.transforms.ColorJitter(brightness=0, contrast=10.0, saturation=0, hue=0)
- 'saturation': torchvision.transforms.ColorJitter(brightness=0, contrast=0, saturation=10.0, hue=0)
- 'hue': torchvision.transforms.ColorJitter(brightness=0, contrast=0, saturation=0, hue=0.5)
- 'rotation': torchvision.transforms.RandomAffine([0, 359], translate=None, scale=None, shear=None, resample=PIL.Image.BILINEAR, fillcolor=0)
- 'translate': torchvision.transforms.RandomAffine(0, translate=[0.2, 0.2], scale=None, shear=None, resample=PIL.Image.BILINEAR, fillcolor=0)
- 'scale': torchvision.transforms.RandomAffine(0, translate=None, scale=[0.8, 1.2], shear=None, resample=PIL.Image.BILINEAR, fillcolor=0)
- 'shear': torchvision.transforms.RandomAffine(0, translate=None, scale=None, shear=[-10., 10., -10., 10.], resample=PIL.Image.BILINEAR, fillcolor=0)
- 'vflip': torchvision.transforms.RandomVerticalFlip(p=0.5)
- 'hflip': torchvision.transforms.RandomHorizontalFlip(p=0.5)

Step2

Data Augmentation	rotated MNIST
'brightness'	98.45 ± 0.24
'contrast'	98.64 ± 0.23
'saturation'	98.95 ± 0.21
'hue'	98.66 ± 0.36
'rotation'	64.70 ± 2.21
'translation'	90.84 ± 1.65
'scale'	91.42 ± 1.34
'shear'	91.48 ± 1.14
'vertical flip'	88.79 ± 0.50
'horizontal flip'	91.98 ± 0.29

Step3

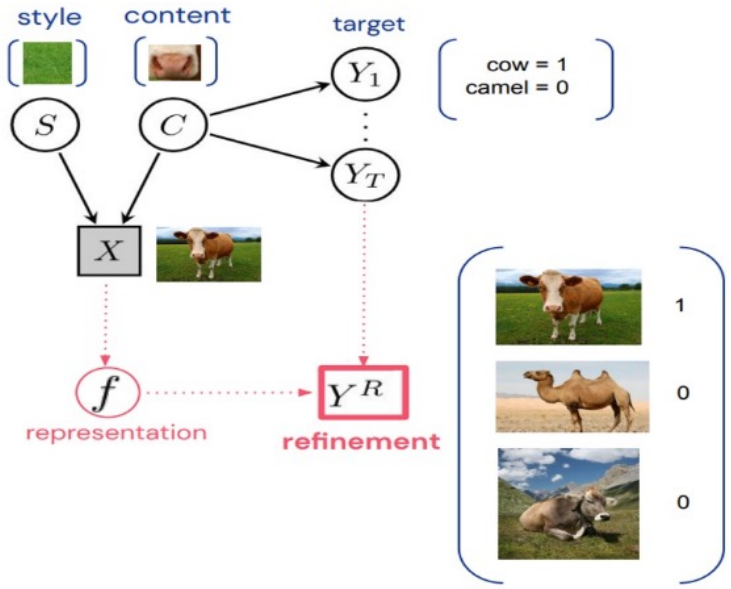
Domain accuracies (AVG $\pm$ STD)

Target	ERM	DANN	CDANN	SDA
0°	75.4	77.1	78.5	**96.1**
30°	93.4	94.2	94.9	**95.9**
60°	94.5	95.2	95.6	**95.7**
90°	79.6	83.0	84.0	**95.9**
Ave	85.7	87.4	88.3	**95.9**

Results on the rotate-MNIST dataset

Manual Data Augmentation via Do operation

- **Goal:** Learn a representation that captures **content**, while discarding the **style**.



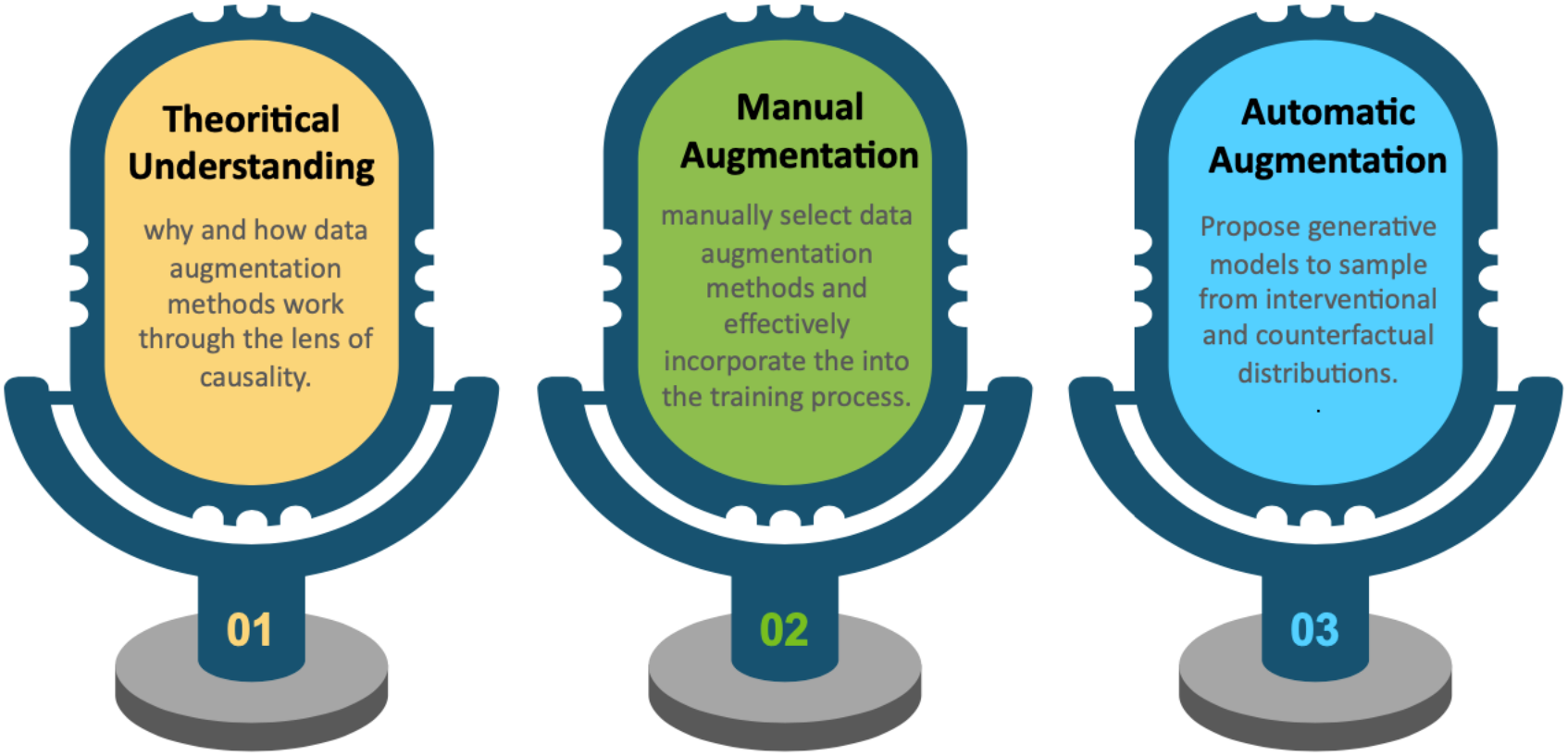
Definition (invariant predictor). The content is an *invariant predictor* of the target if:

$$p^{do(S=s_i)}(Y_t|c) = p^{do(S=s_j)}(Y_t|c). \quad \forall s_i, s_j \in \mathcal{S}$$

Use data augmentation to simulate interventions on the style

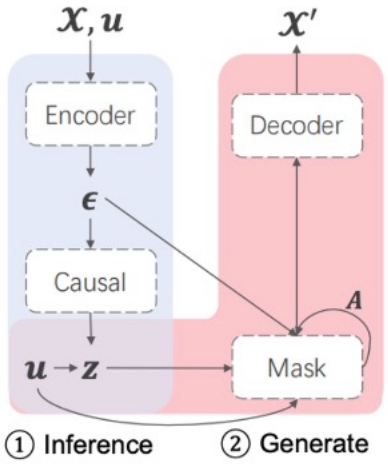
[Mitrovic et al., 2021]

Causal Data Augmentation: categories

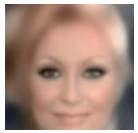


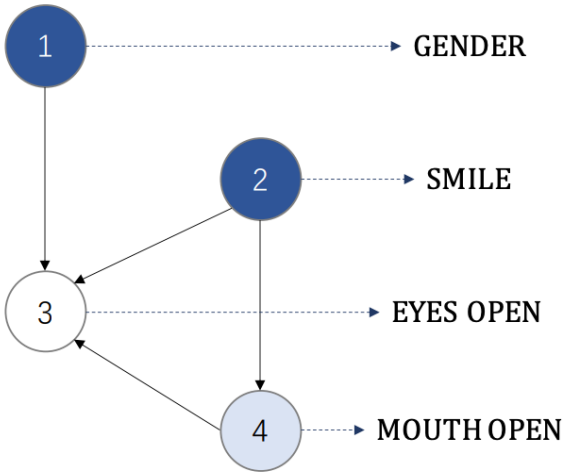
Automatic Data Augmentation

- CausalVAE: automatic generation of counterfactual data, when the causal graph for the high-level features is unknown apriori.



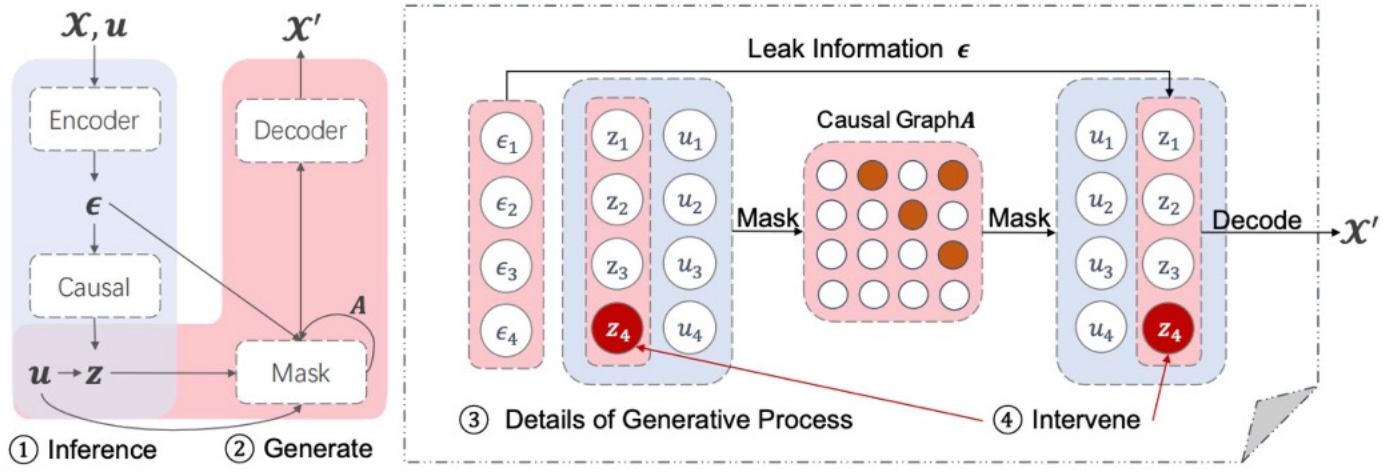
GENDER, SMILE, EYES OPEN, MOUTH OPEN





Automatic Data Augmentation

- CausalVAE: automatic generation of counterfactual data, when the causal graph for the high-level features is unknown apriori.

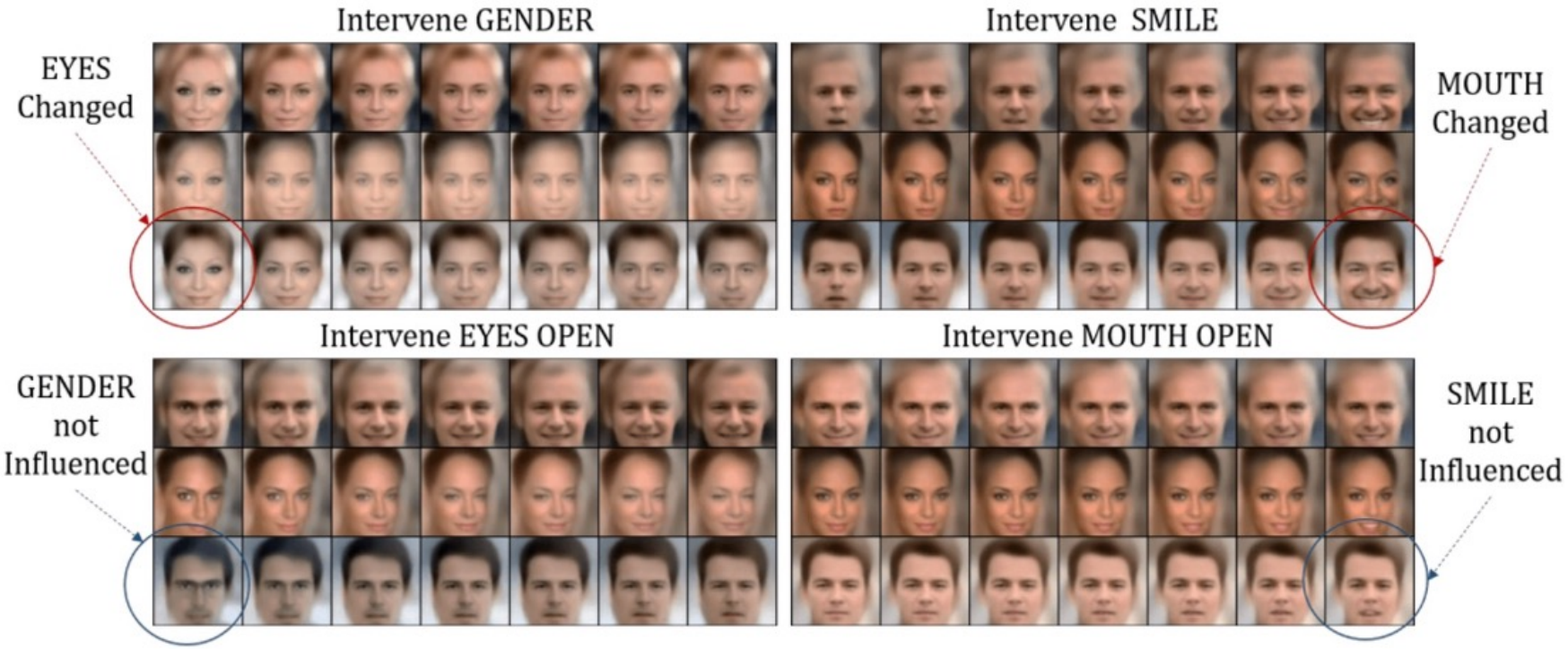


$$z = A^T z + \epsilon = (I - A^T)^{-1} \epsilon$$

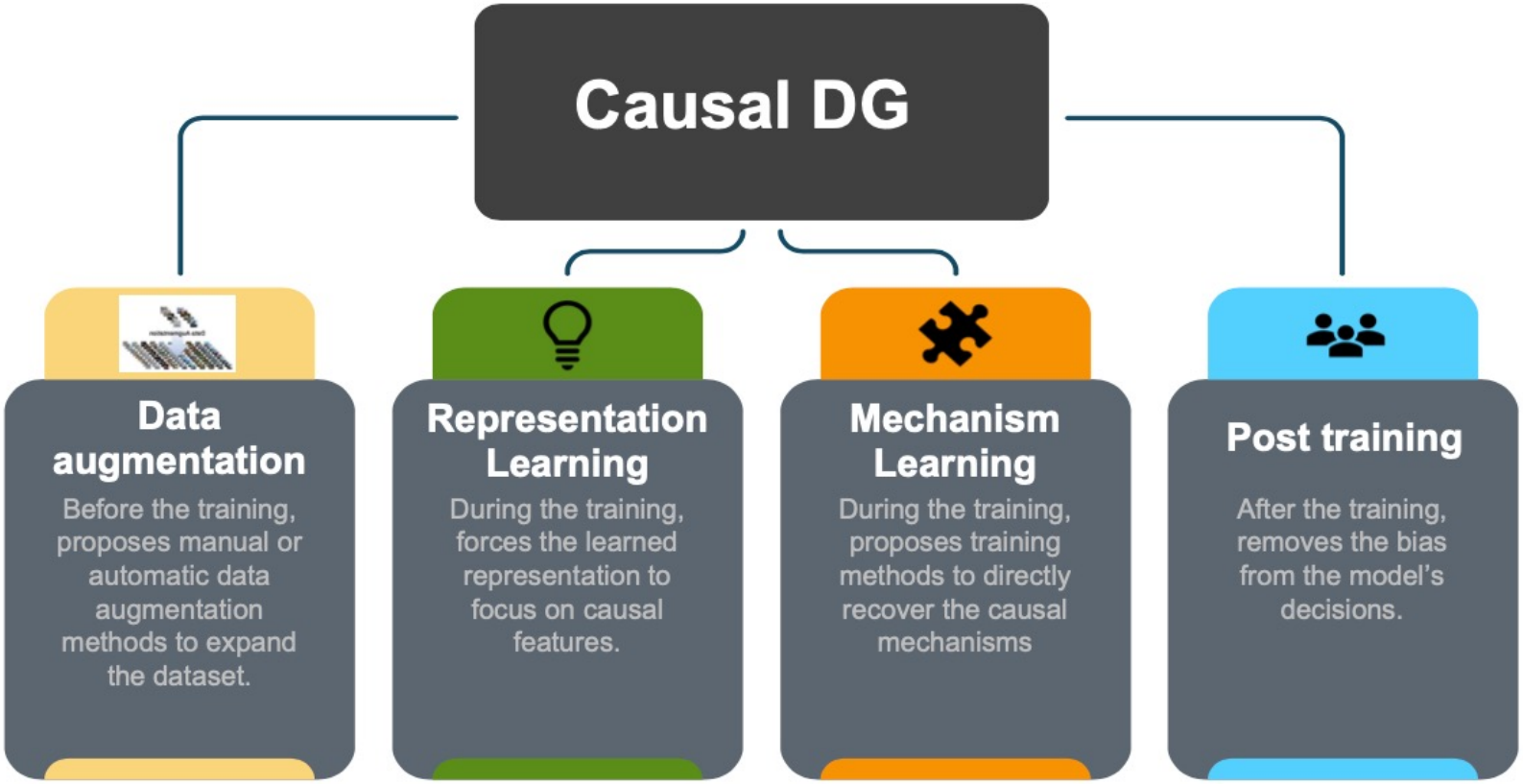
! Don't forget to add non-linearity

[Yang et al., 2021]

CausalVAE: An example

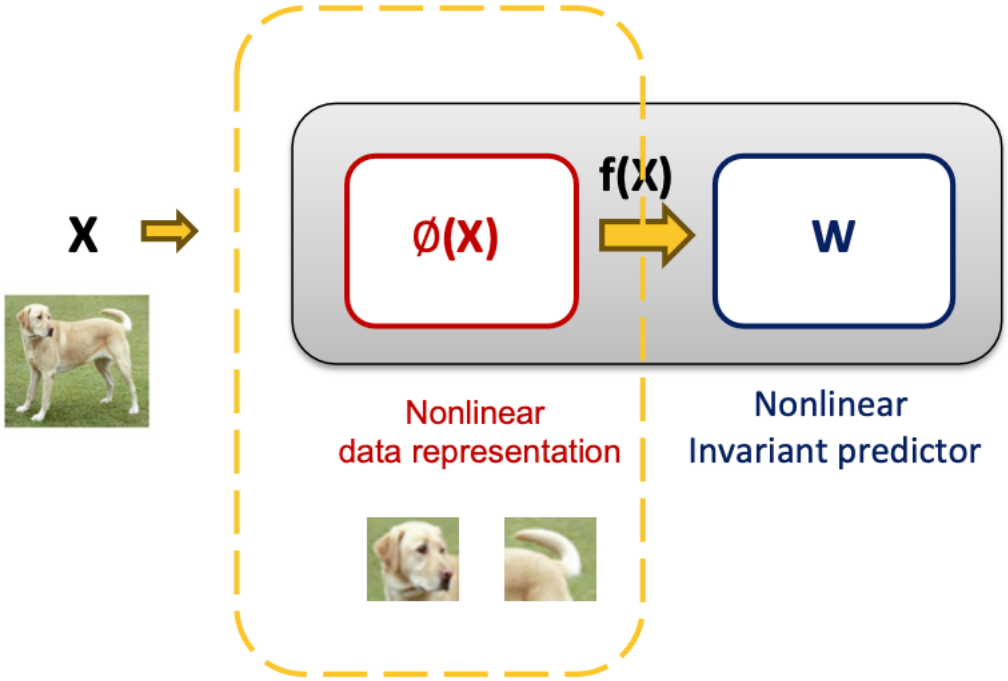


Overview of the approaches



[Sheth, Moraffah et al., 2021]

Invariant Causal Representation Learning



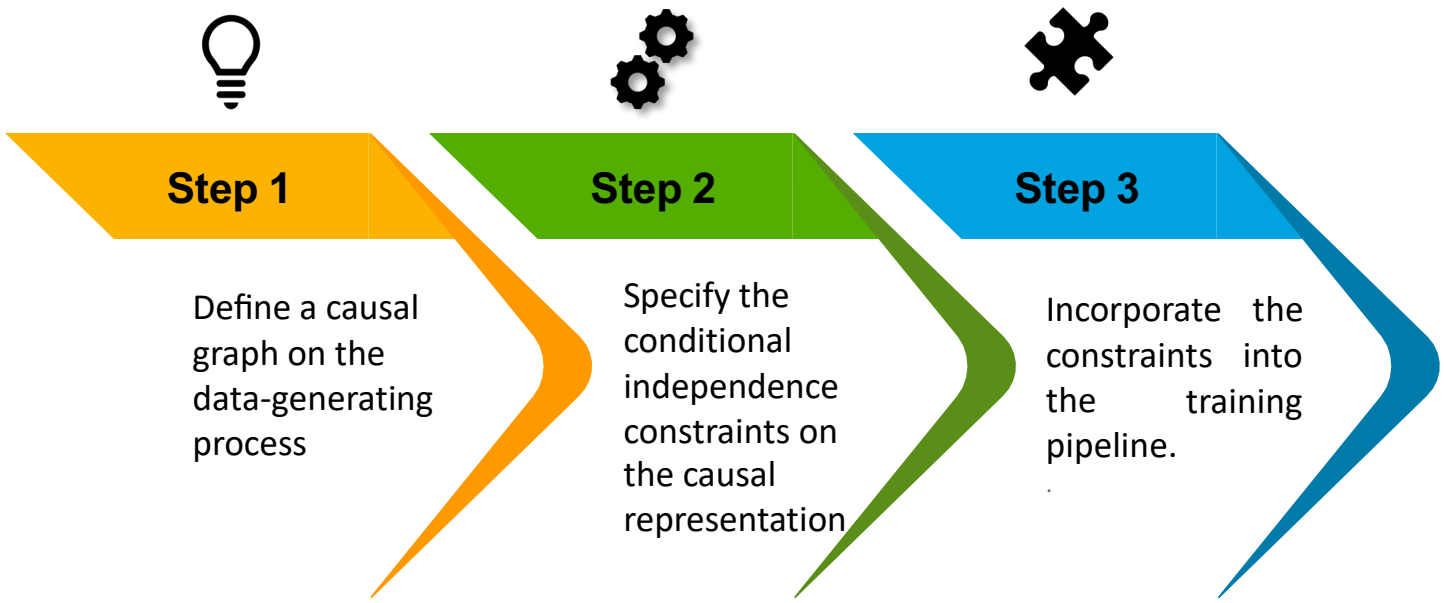
$$Y \approx f(X) = W o\ \emptyset(X)$$

Should be the direct causes of Y

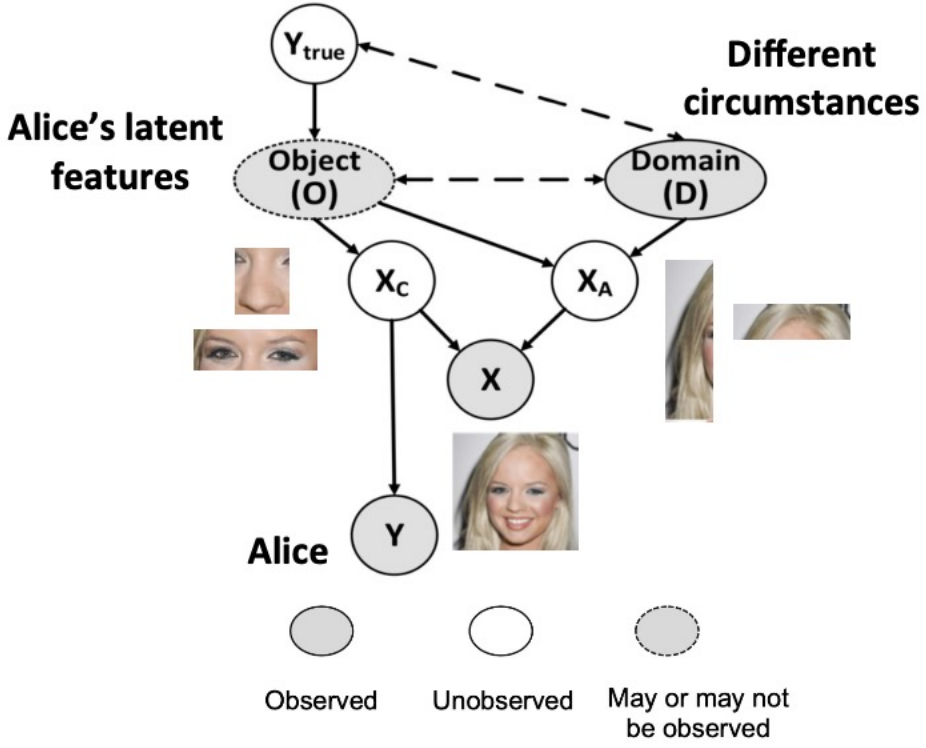
Invariant Causal Representation Learning: categories

- Conditional independent based
  - Impose conditional independence constraints through regularizers
- VAE-based
  - Identify and learn the causal representation distribution

Conditional Independence-based Methods

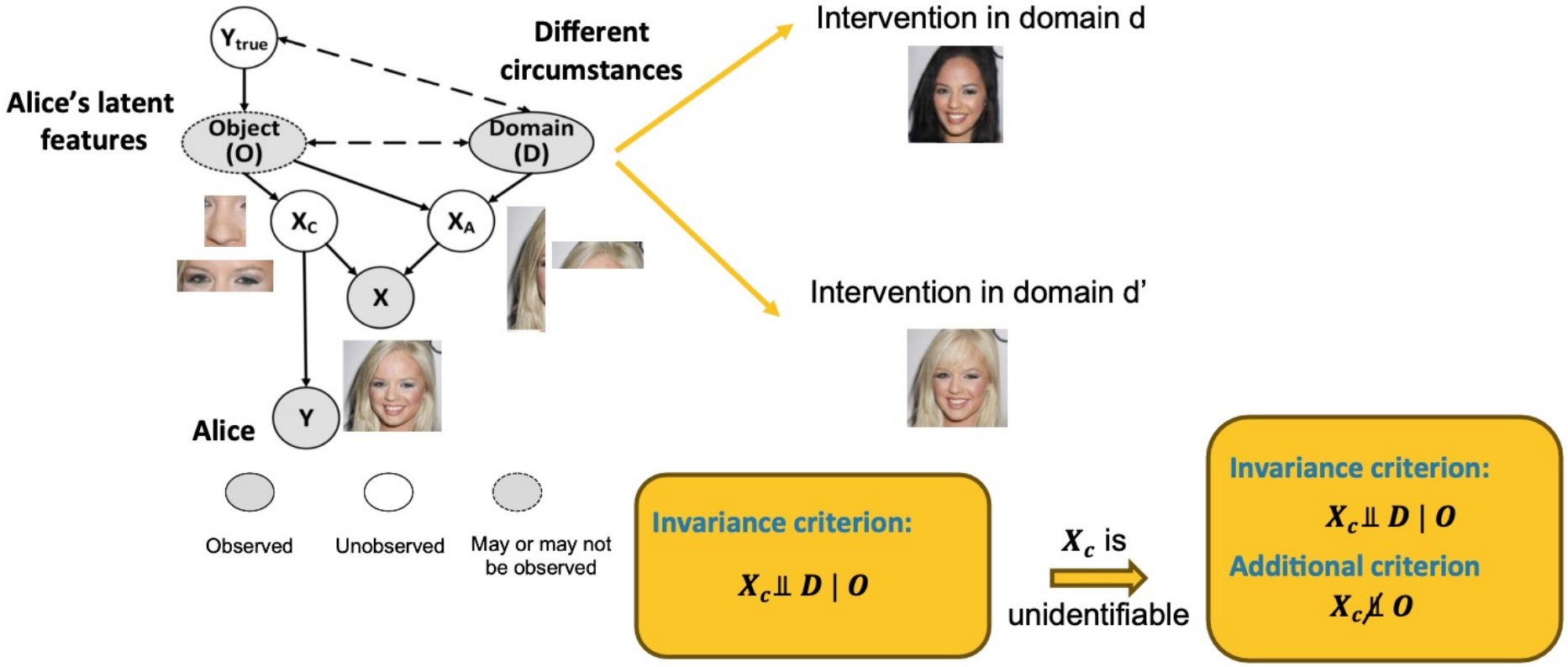


Conditional Independence-based Methods



[Mahajan et al., 2021]

Conditional Independence-based Methods



Perfect-Match: the object is known

Training Domains

Test Domains

Objective:

$$f_{perfect-match} = \boldsymbol{arg}\min_{h,\emptyset} \sum_{d} \mathcal{L}_d(h(\emptyset(X)), Y) + \lambda \sum_{\Omega(j,k)=1} Dist(\emptyset(x_j^d), \emptyset(x_j^{d'}))$$

Additional criterion

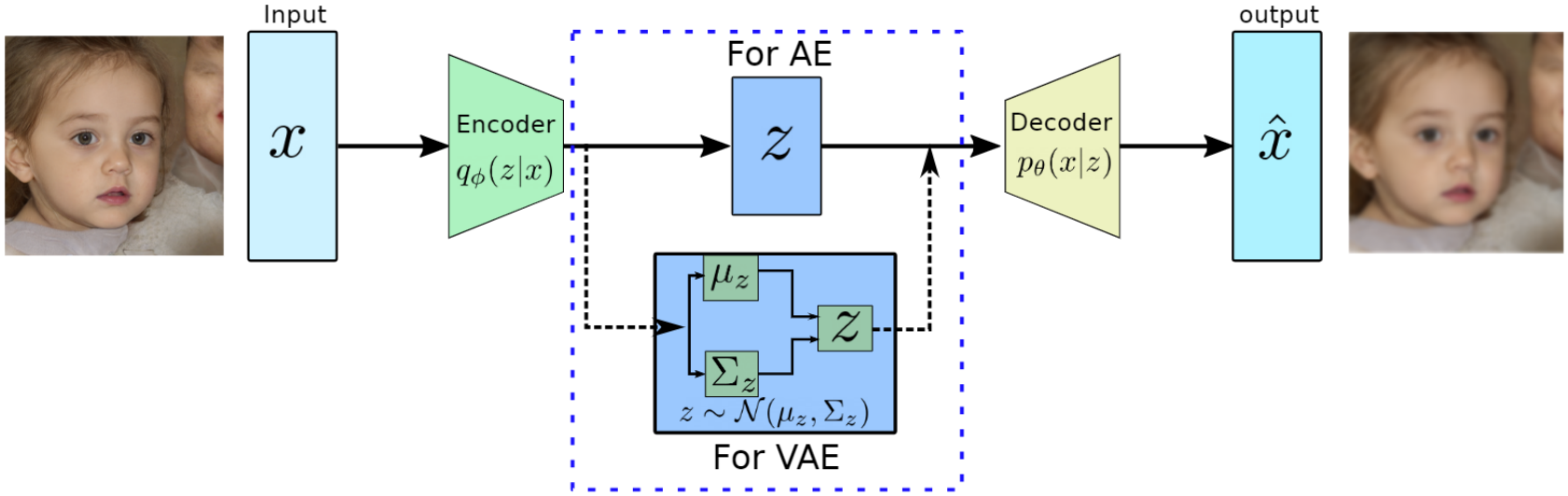
Invariance criterion

Invariant Causal Representation Learning: categories

- Conditional independent based
  - Impose conditional independence constraints through regularizers
- VAE-based
  - Identify and learn the causal representation distribution

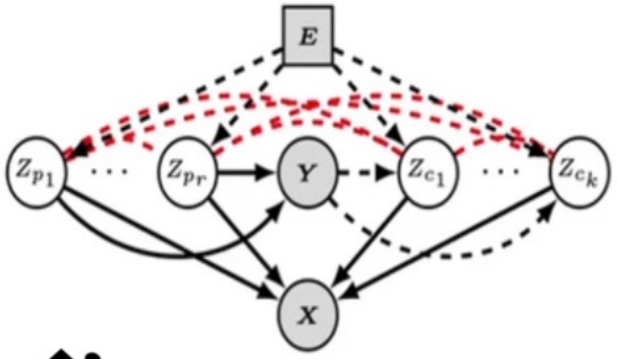
VAE

- Generative model, infer latent generative factors in the bottleneck layer



VAE-based Methods

Assumption1. Z_i depends on one or both of Y and E
Assumption2. The causal graph is a DAG.
Assumption3. $X \perp\!\!\!\perp Y, E \mid Z$
Assumption4. $Y \perp\!\!\!\perp E | Z_p$, which implies p(Y|$Z_p$) is invariant.



Step 1

Identify Z from X

Step 2

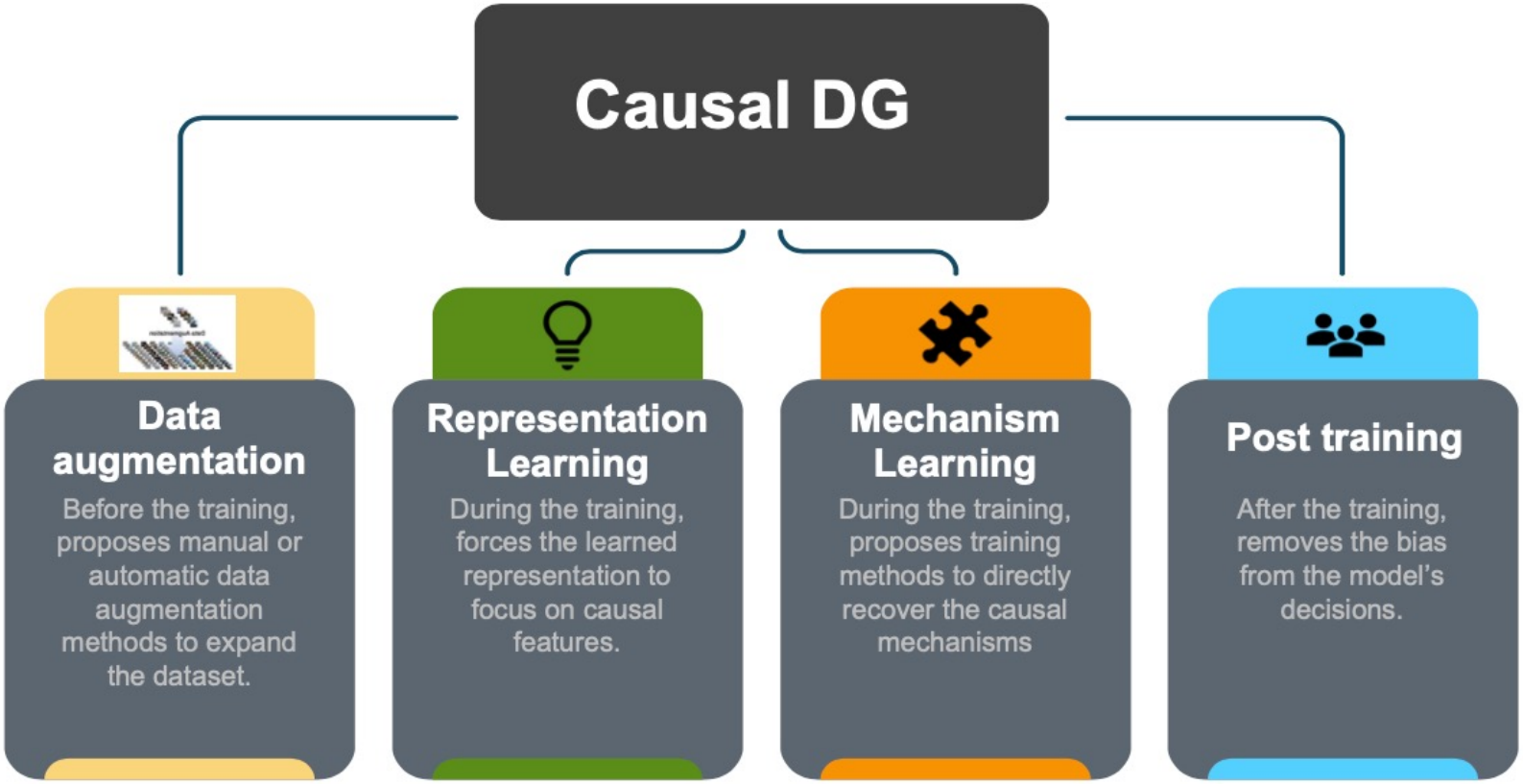
Determine direct causes of Y (pa(Y))

Step 3

Learn an invariant predictor using Pa(Y).

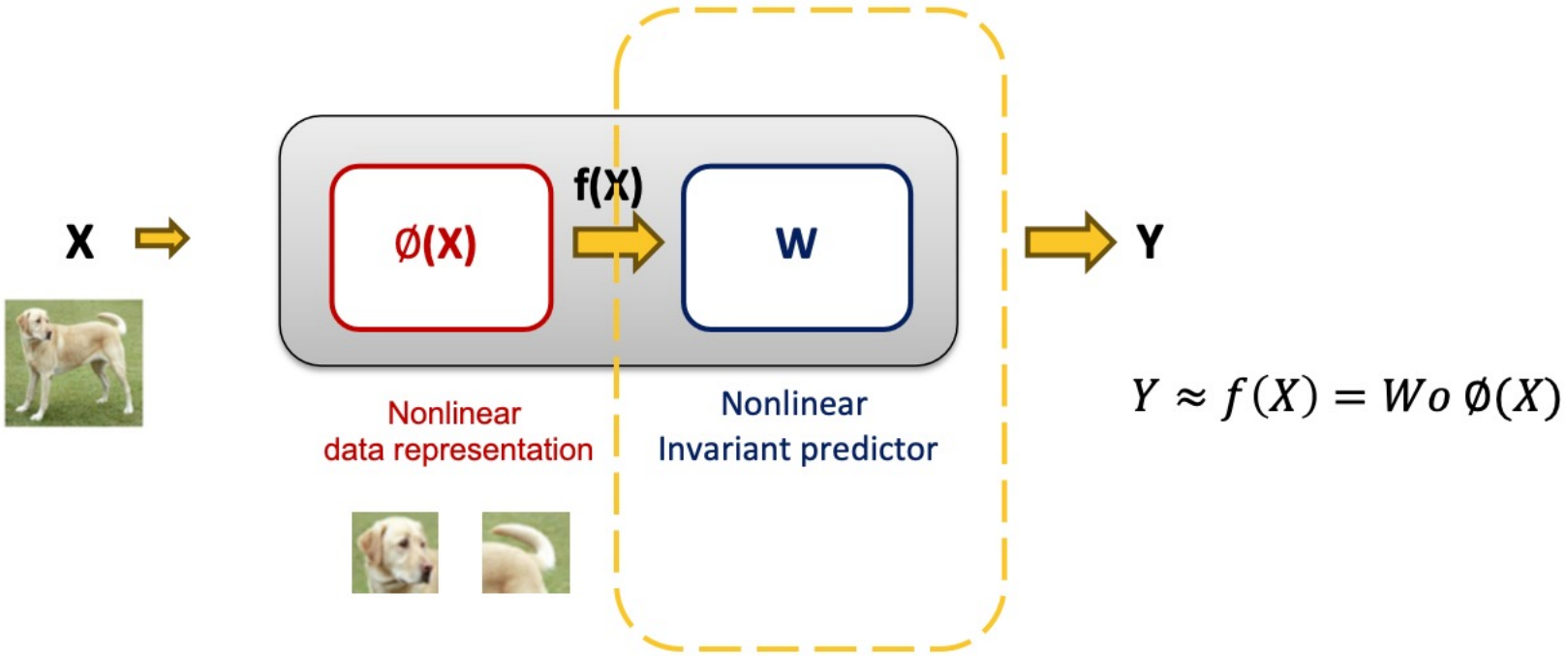
[Lu et al., 2022]

Overview of the approaches

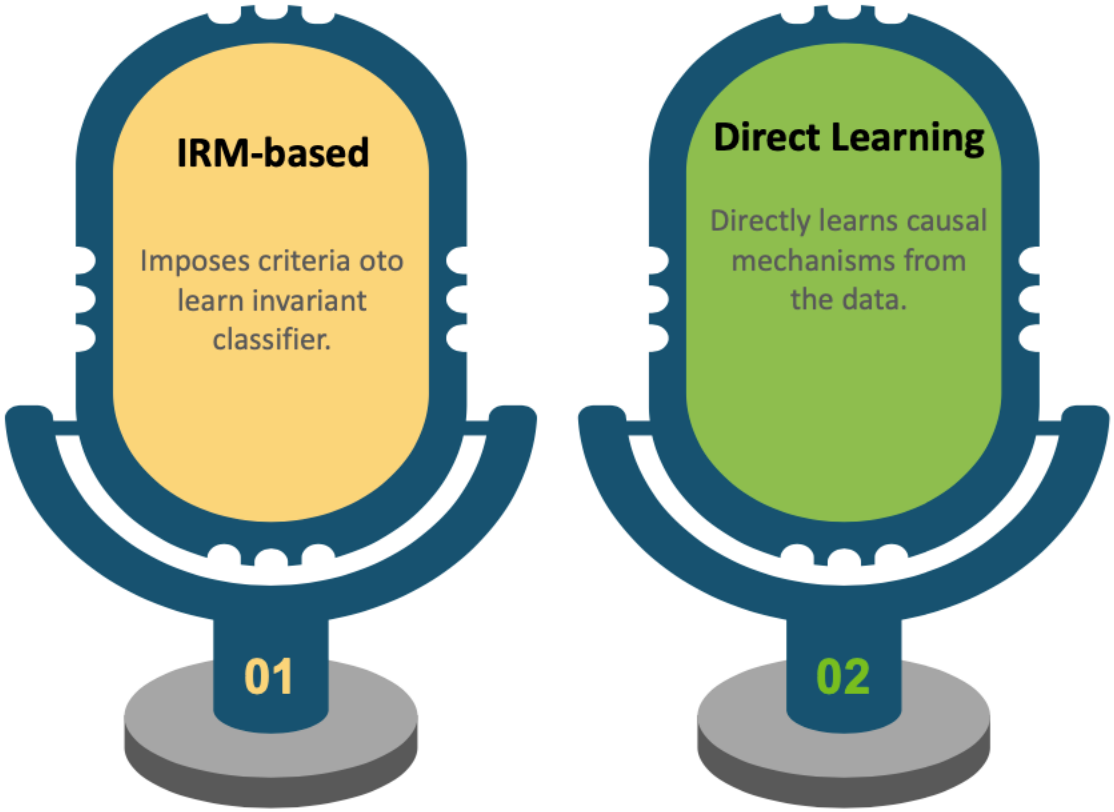


[Sheth, Moraffah et al., 2021]

Causal Mechanism Learning

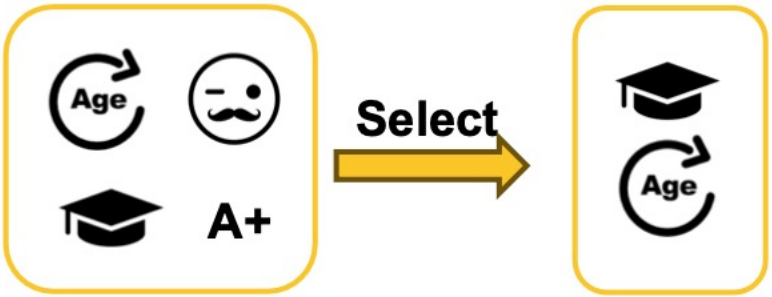


Casual Mechanism Learning



IRM-based Methods

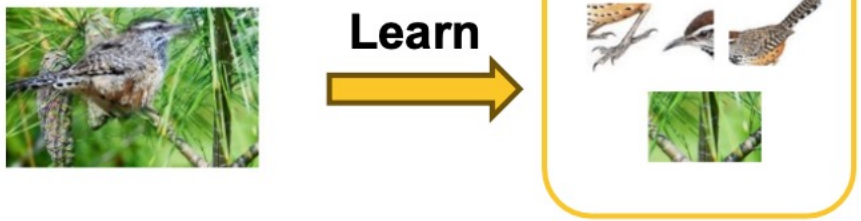
- Idea: Find features whose relationships with the target remain *invariant* across different environments.



Features

Invariant Causal Prediction (ICP)

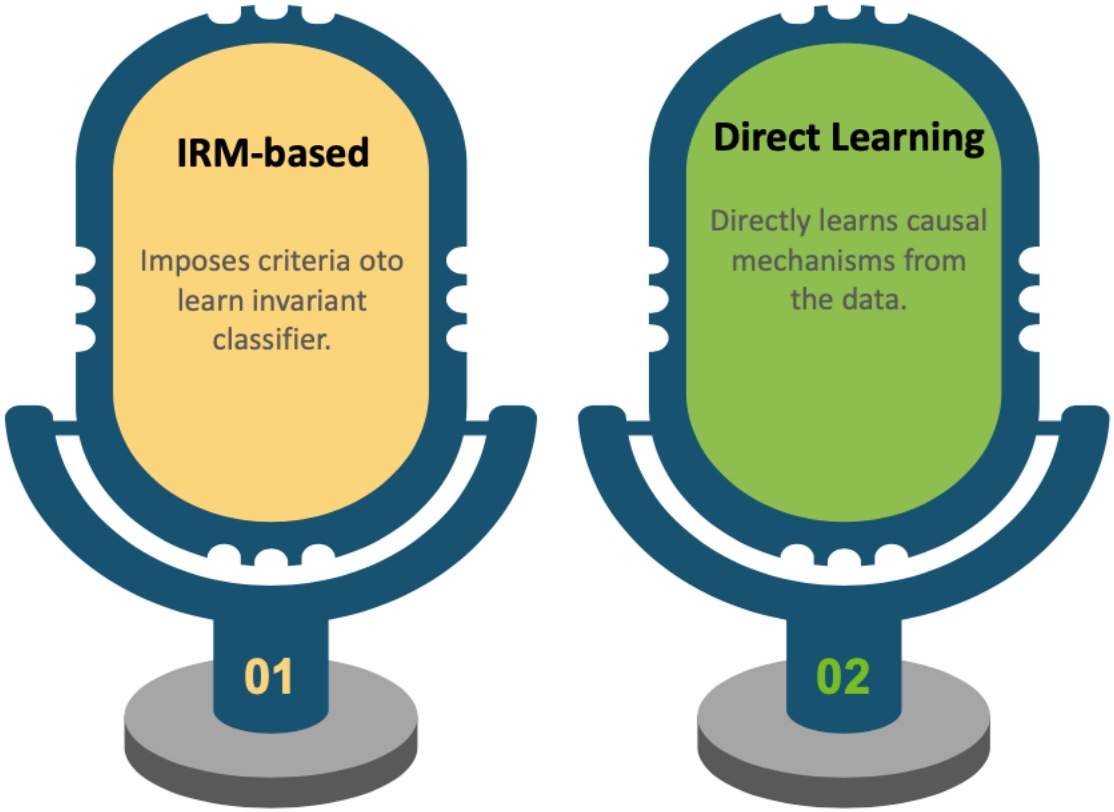
[Peters et al., 2019]



Invariant Risk Minimization (IRM)

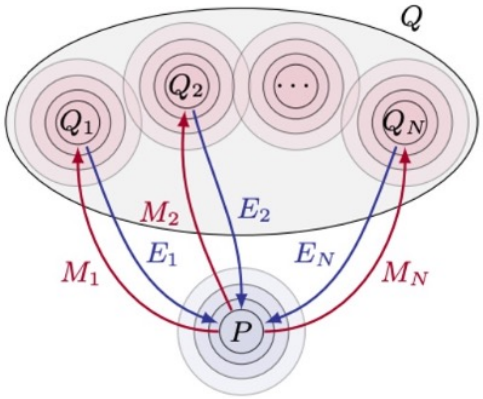
[Arjovsky et al., 2019]

Casual Mechanism Learning



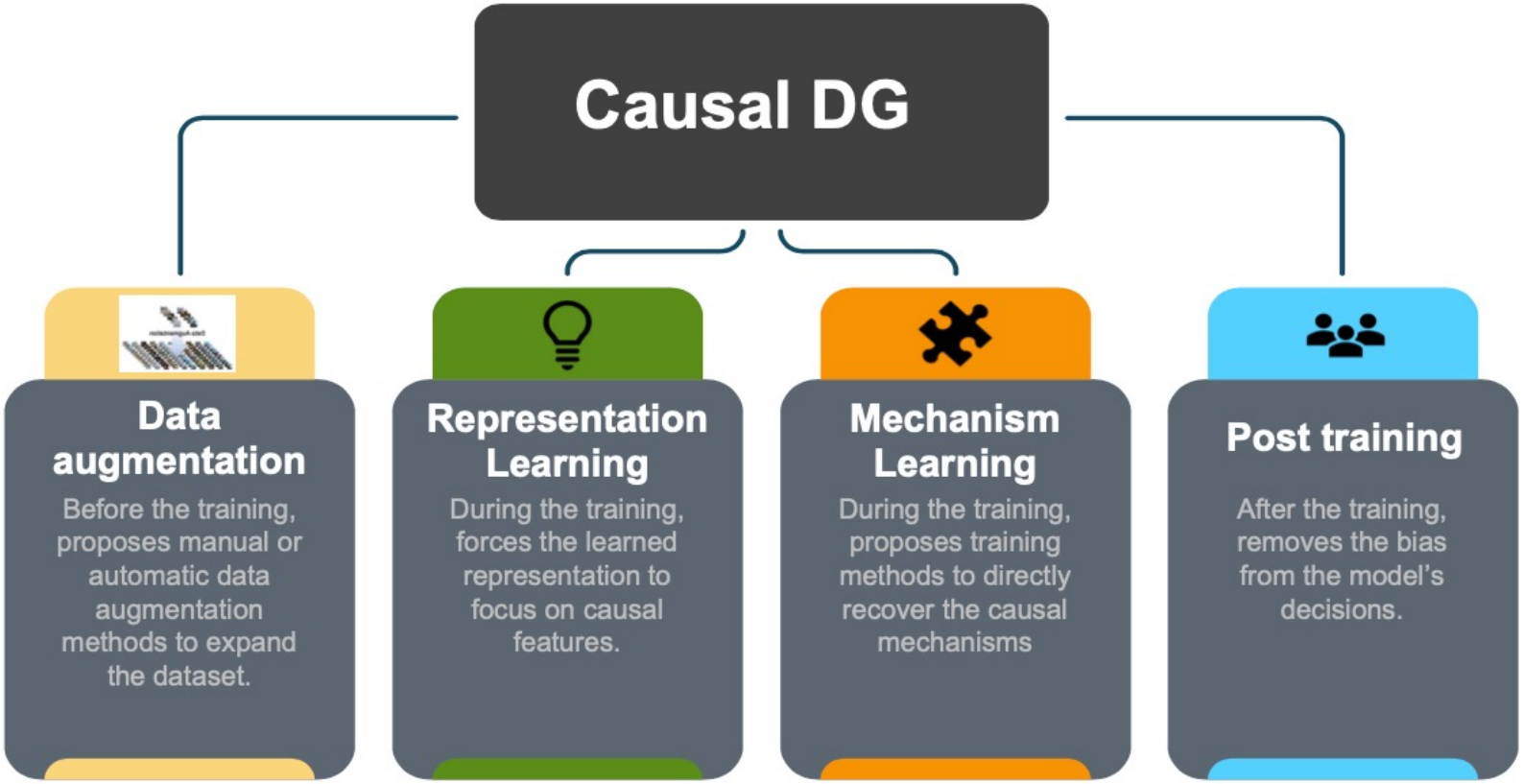
Direct Learning Methods

- **Input:** Samples from a canonical distribution P and samples form a mixture of transformed distributions Q_i generated via applying mechanism M_i on P.
- **Goal:** learn inverse mechanisms as independent modules.



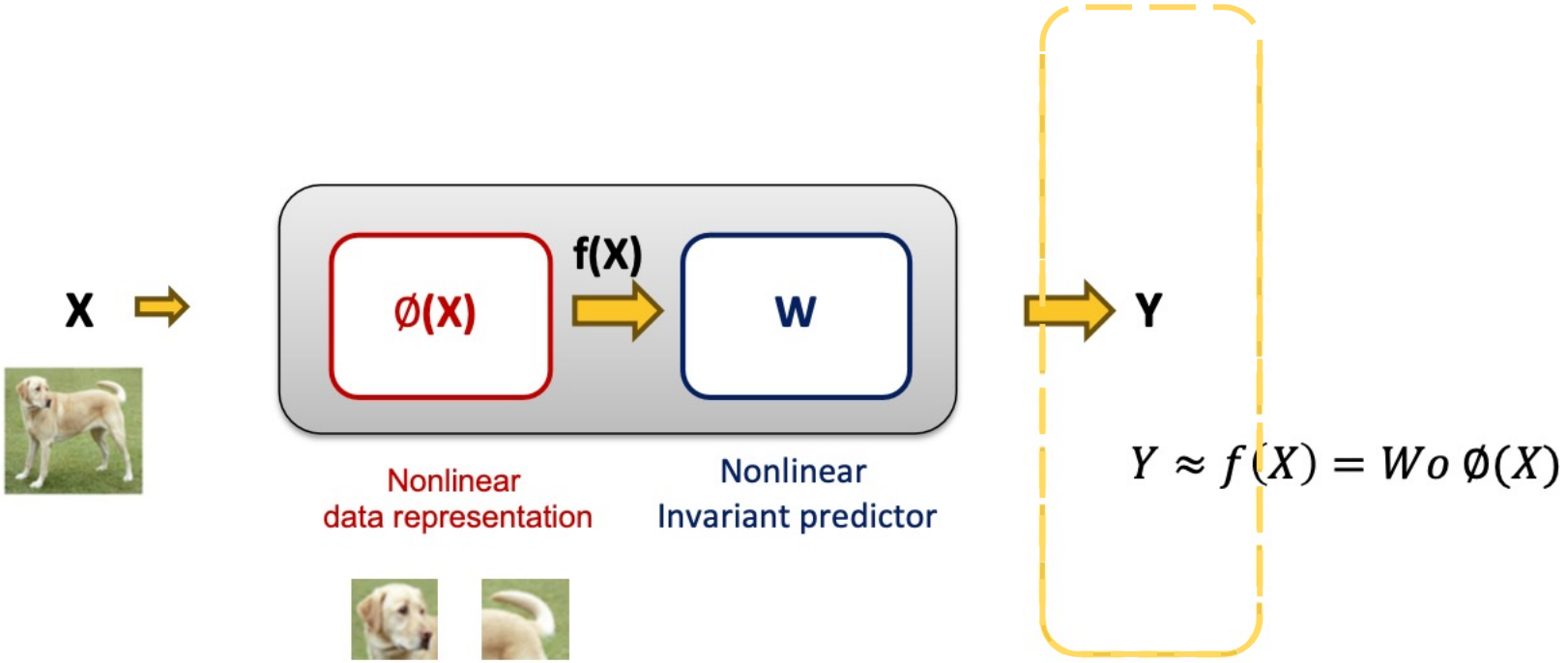
[Parascandolo et al., 2018]

Overview of the approaches



[Sheth, Moraffah et al., 2021]

Post training



Post-training

Case study: Out-of-domain generalization in VQA

- Consider the following text, image pair:

Q: Do you see a player?

A: Yes.

Q: What sport is he playing?

A: Tennis.

[Niu et al., 2021]

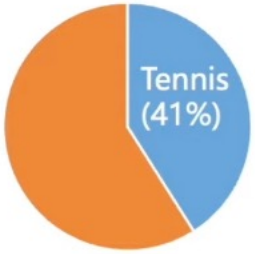
Post-training

Case study: Out-of-domain generalization in VQA

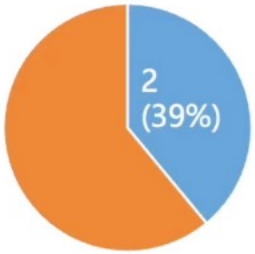
- Strong prior on language

(VQA v1 dataset)

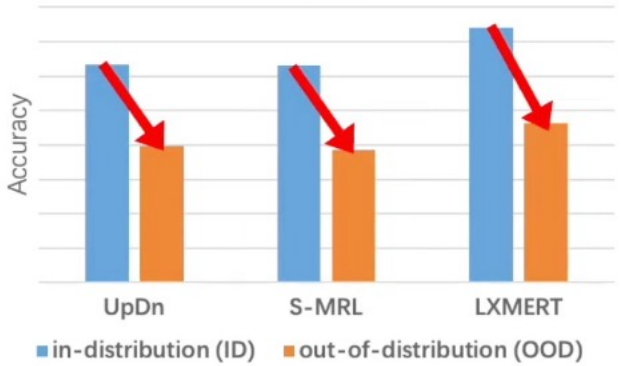
Q: What sport is ... ?



Q: How many ... ?



language priors

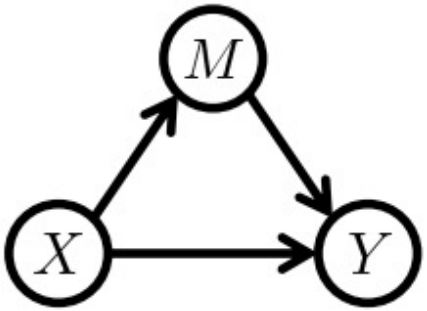


poor OOD generalization

[Goyal et al., 2017]

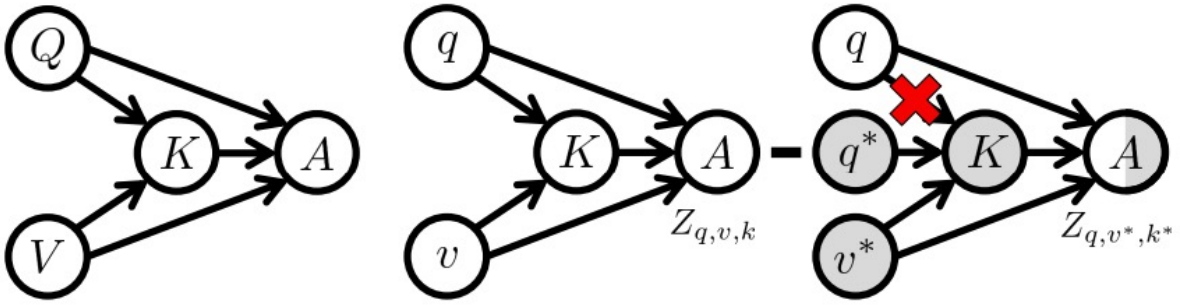
Total/direct/indirect causal effect

- Total effect: all causal effect from X->Y
- Direct effect: X->Y
- Indirect effect: X->…->Y, going through mediator



Conventional VQA vs Causal VQA

- **Conventional VQA**: What will answer A be, if machine hears question Q, sees image V , and extracts the multimodal knowledge K?
- **Counterfactual VQA:** What would A be, if machine hears Q, but had not extracted K or seen V ?

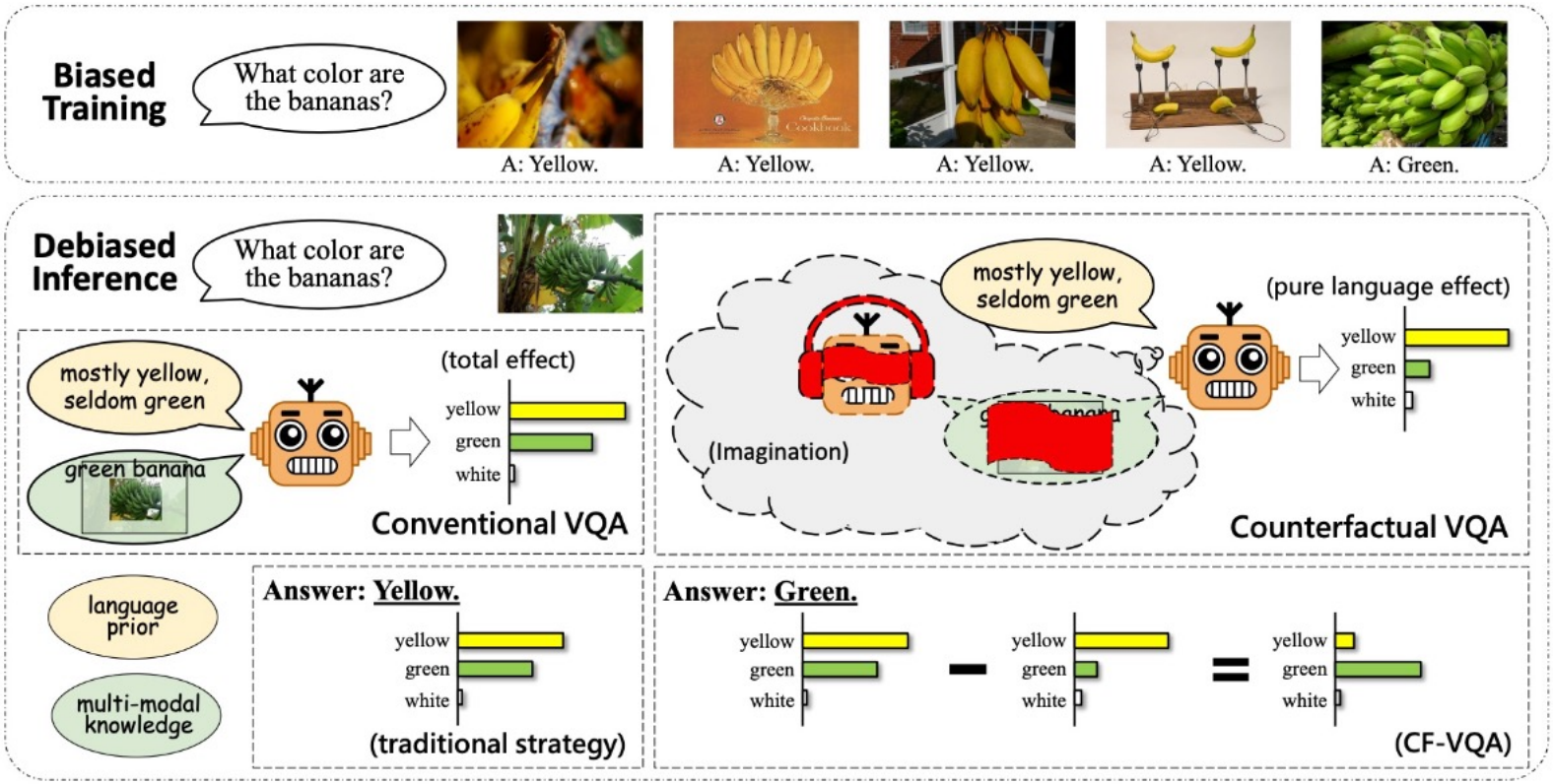


(a) Conventional

(b) Causal

Q: question. V : image. K: multi-modal knowledge. A: answer

Post-training



- Conventional VQA: machine hears the question and extracts the multi-modal knowledge.
- Pure language effect : the direct causal effect of question on answer (without image).
- Counterfactual VQA: We subtract the pure language effect from the total effect for debiased inference.

Niu Y, Tang K, Zhang H, et al. Counterfactual vqa: A cause-effect look at language bias[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

References & reading materials

- Von Kügelgen J, Sharma Y, Gresele L, et al. Self-supervised learning with data augmentations provably isolates content from style[J]. NeurIPS, 2021.
- Niu Y, Tang K, Zhang H, et al. Counterfactual VQA: A cause-effect look at language bias[C]//CVPR. 2021.
- Mahajan D, Tople S, Sharma A. Domain generalization using causal matching[C]//ICML. PMLR, 2021.
- Zhang K, Gong M, Stojanov P, et al. Domain adaptation as a problem of inference on graphical models[J]. NeurIPS, 2020.

- Some content of slides are from:
  - Socially Responsible Machine Learning: A Causal Perspective. KDD 2023 Tutorial.

Thank you!
Questions?