Data-fusion and transportability in machine vision

Zach Calhoun
Frankie Willard
February 28, 2023

What are we going to talk about today?

1. Discuss paper one
    a. The **data-fusion problem** and motivation
    b. Brief review of concepts from Pearl’s model.
    c. Introduction to **do-calculus**
    d. Techniques for overcoming the problem
2. Discuss paper two
    a. Motivation / Brief introduction to machine vision
    b. Transportability (a.k.a., domain adaptation)
3. Our research
    a. Introduction to concepts
    b. How causal inference might help!

Today’s motivation: can Pearl’s model address more complex issues?

In previous classes:

- Focused on **confounding** bias

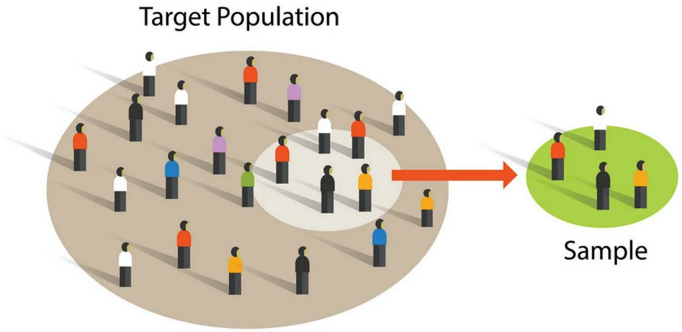
Many possible solutions:

- Back-door criterion
- Front-door criterion
- Instrumental variables

In today’s lecture, we focus on:

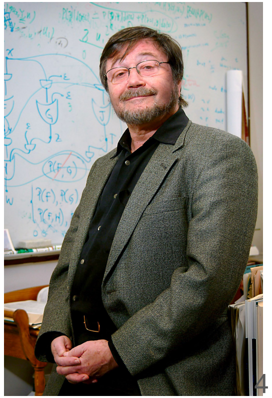
- Sampling bias
- Transportability bias

Do-calculus offers solutions!



Causal inference and the data-fusion problem

Bareinboim and Pearl, 2015



What is data-fusion?

“Data fusion aims to combine results from many experimental and observational studies, each conducted on a different population and under a different set of conditions in order to synthesize an aggregate measure of targeted effect size that is “better,” in some sense, than any one study in isolation” (Bareinboim and Pearl, 2016, p. 7351)

Sources of bias covered in this paper by experimental design

Type of Bias	**Experimental Design**
Confounding bias	$p(v)$, $p(v\|do(z))$
Sample selection bias	$p(y\|S=1)$, $p(y\|do(X=x), S=1)$
Transportability Bias	$p(v\|do(x))$ + observational studies

“Not all data are created equally”

- The way in which your data is sampled matters.

Brief Review: Structural Causal Models

- $M = \langle U, V, F \rangle$
- Endogenous (observable) variables $V = \{G, X, D, O\}$
- Exogenous (noise) variables $U = \{U_G, U_X, U_D, U_o\}$
- Structural equations F:

$\{G = F_G(U_G),$

$X = F_x(U_x, G),$

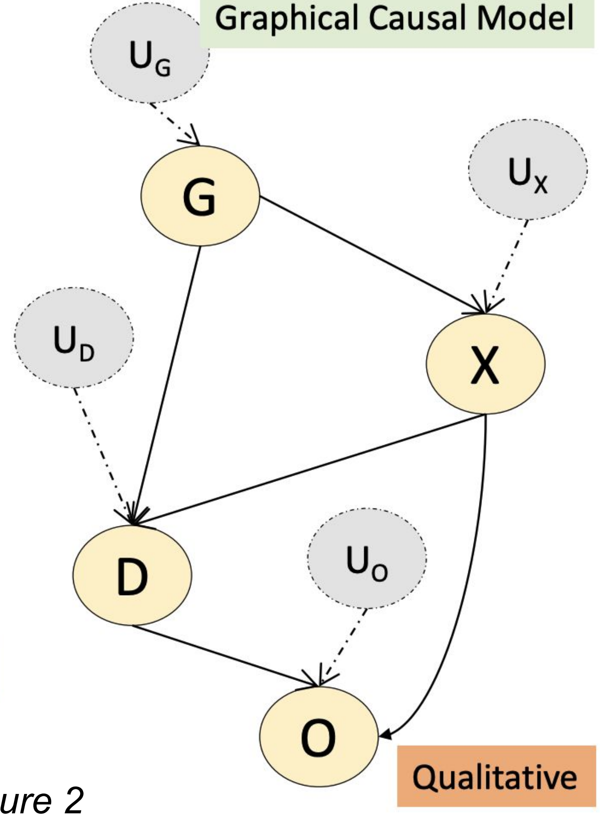
$D = F_D(U_D, G, X),$

$O = F_O(U_O, X, D)\}$

Can be linear, exp, ...

Quantitative

Corresponding Graphical Causal Model

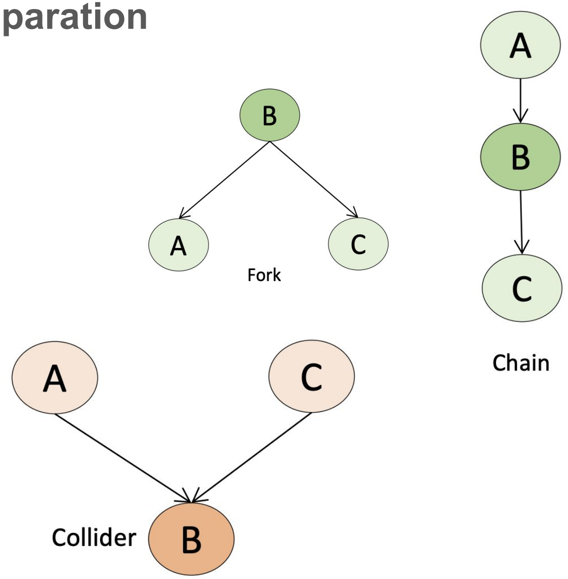


Qualitative

Slide taken from lecture 2

Brief Review

d-separation



Having d-separation ensures independence between the treatment and the effect.

identifiability

$$P_1(v) = P_2(v) \Rightarrow Q(M_1) = Q(M_2)$$

A causal query is identifiable if we can get a unique parameterization of our model using our assumptions and the available data.

We can test for identifiability using do-calculus.

do-Calculus: Notation

X, Y, Z, W	arbitrary disjoint sets of nodes in a causal DAG G
$G_{\bar{x}}$	The subgraph achieved by deleting all edges pointing towards **X**
$G_{\underline{x}}$	The subgraph resulting from deleting all edges pointing away from **X**.

Example Subgraphs

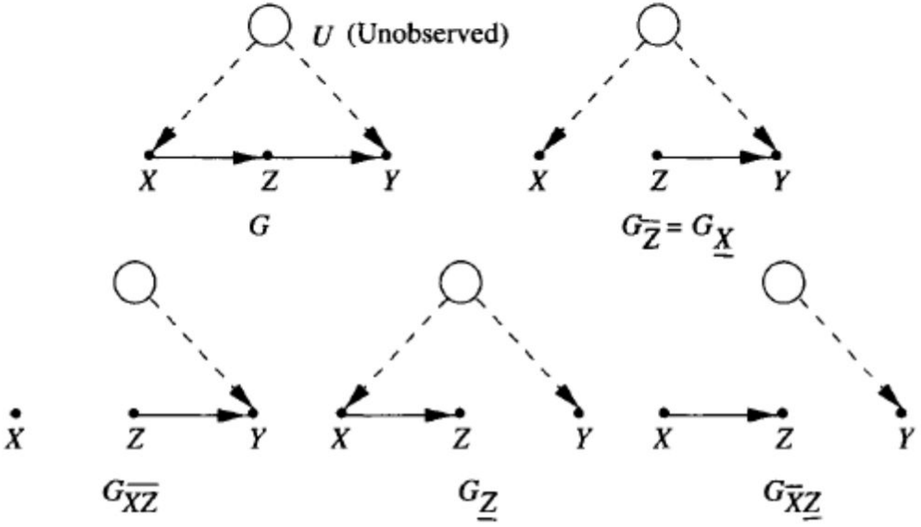
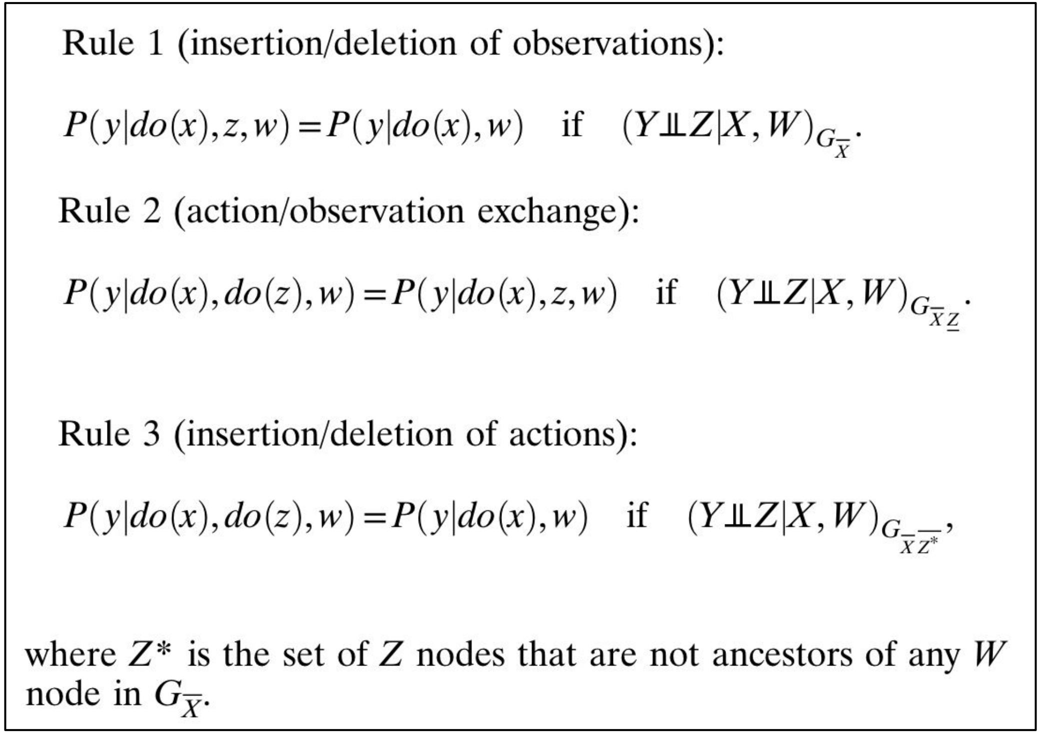

U (Unobserved)
X
Z
Y
G
X
Z
Y
$G_{\overline{Z}} = G_{\underline{X}}$
X
Z
Y
$G_{\overline{XZ}}$
X
Z
Y
$G_{\underline{Z}}$
X
Z
Y
$G_{\overline{X}\underline{Z}}$

Figure 3.6 Subgraphs of G used in the derivation of causal effects.

Pearl, Judea. *Causality: Models, Reasoning, and Inference.* p.87

The Rules of do-Calculus



Rule 1 (insertion/deletion of observations):

$$P(y|do(x), z, w) = P(y|do(x), w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}}}.$$

Rule 2 (action/observation exchange):

$$P(y|do(x), do(z), w) = P(y|do(x), z, w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}\underline{Z}}}.$$

Rule 3 (insertion/deletion of actions):

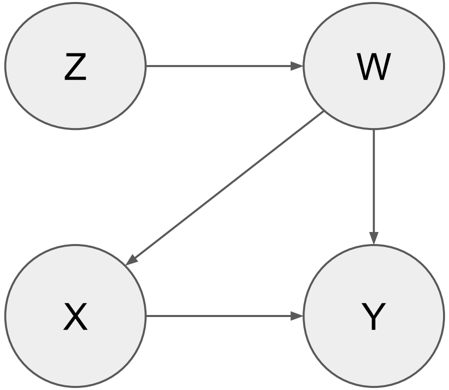
$$P(y|do(x), do(z), w) = P(y|do(x), w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}\,\overline{Z^*}}},$$

where Z^* is the set of Z nodes that are not ancestors of any W node in $G_{\overline{X}}$.

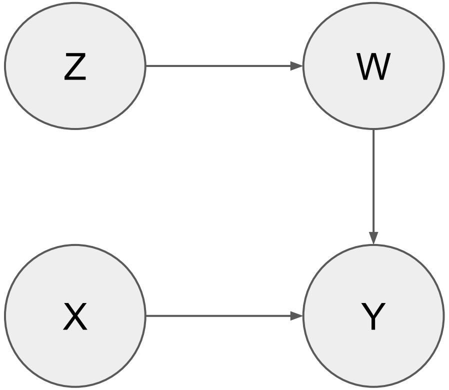
While this math comes from the paper, I recommend this blog post for a simpler overview:
https://stephenmalina.com/post/2020-03-09-front-door-do-calc-derivation/

Rule 1: Insertion/deletion of observations

$$P(y|do(x),z,w) = P(y|do(x),w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}}}.$$



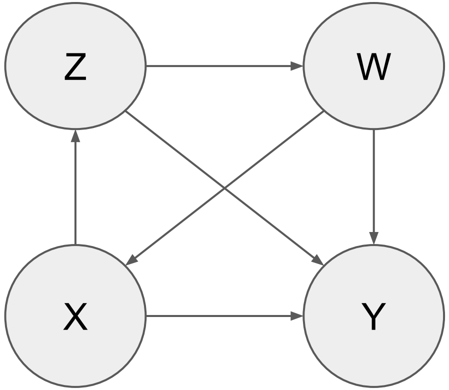
G



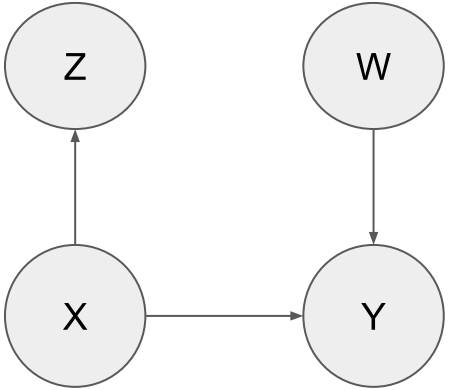
$G_{\bar{x}}$

Rule 2: Action/observation exchange

$$P(y|do(x),do(z),w)=P(y|do(x),z,w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\underline{Z}}}$$



G

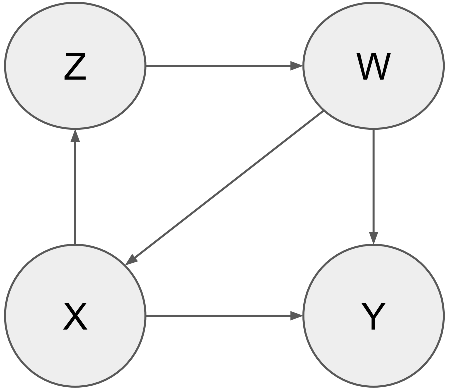


$G_{\overline{X}\underline{Z}}$

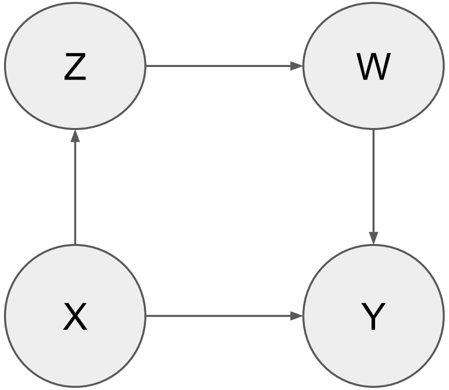
Rule 3: Insertion/deletion of actions

$$P(y|do(x),do(z),w)=P(y|do(x),w) \quad \text{if} \quad (Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\,\overline{Z^*}}}$$

where Z^* is the set of Z nodes that are not ancestors of any W node in $G_{\overline{X}}$.



G

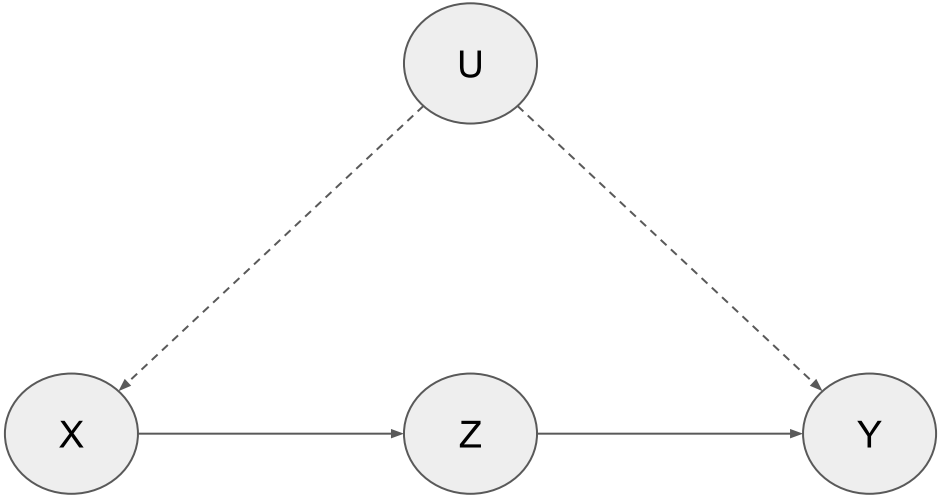


$G_{\overline{X}\underline{Z}}$

A causal query Q is **identifiable** if the rules of do-Calculus can be applied repeatedly until there are no do-operators remaining.

Example Problem (if time permits)

Given the computational graph below, compute P(y|do(z)).



Graph from **Pearl, Judea. *Causality: Models, Reasoning, and Inference.* p.81.**
See solution on pp.86-88.

Applications of do-Calculus

Already seen in this course:

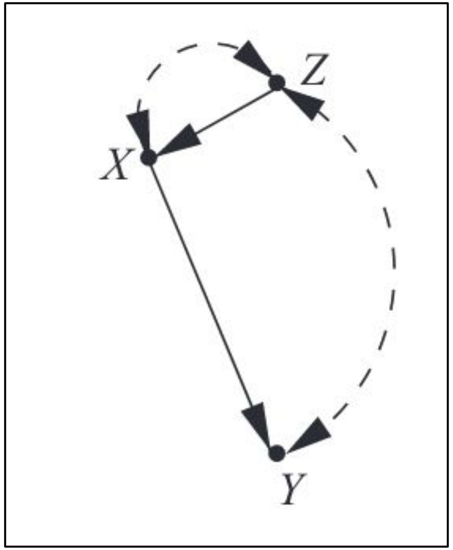
- Backdoor criterion and adjustment
- Frontdoor criterion and adjustment

New applications:

- Identification through auxiliary experiments
- Dealing with Sample Selection Bias
- Transportability

Identification through Auxiliary Experiments

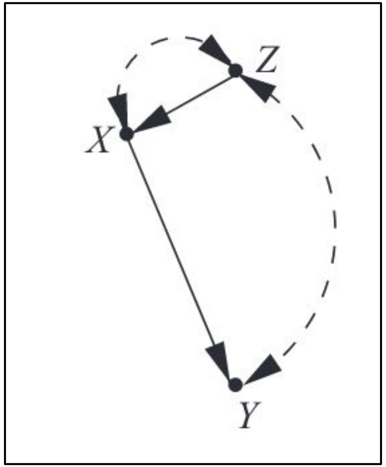
- Hypothetically – what if you wanted to calculate the effect of cholesterol on heart disease? You cannot control cholesterol directly, but if you could control for diet, could you calculate P(y|do(x))?
- X - cholesterol
- Z - diet
- Y - heart disease



This is just the use of an instrumental variable

Through the application of do-calculus, we get:

$$P(Y=y|do(X=x))=P(y,x|do(z))/P(x|do(z))$$

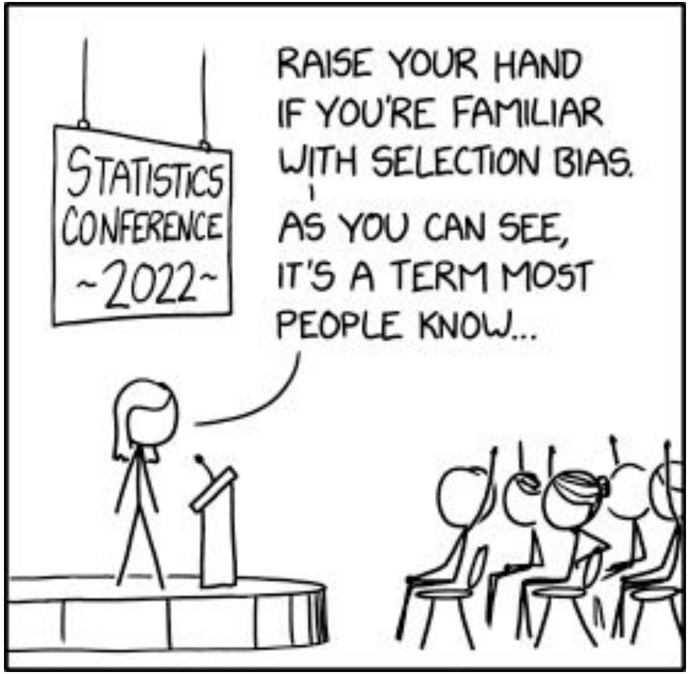


This technique is known as **z-identifiability.**

Dealing with Sample Selection Bias

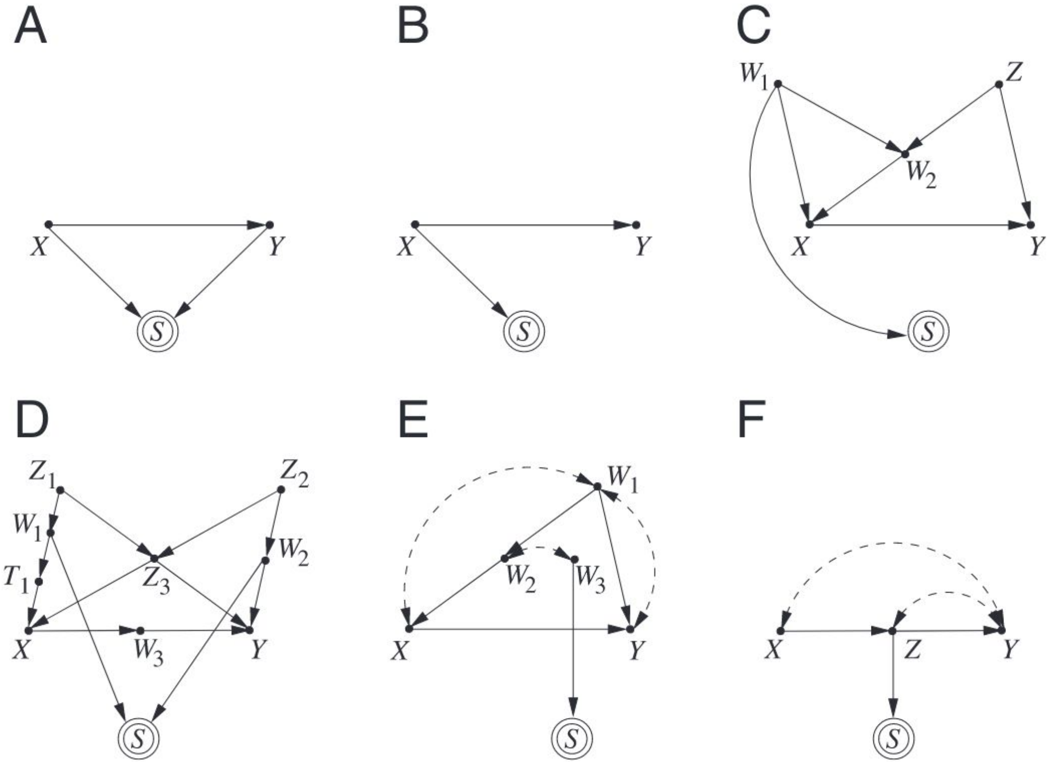
Let’s say that we want to answer the query P(y|do(x)), but we have a preferential selection problem, meaning that we only have the data P(y, x | S=1).

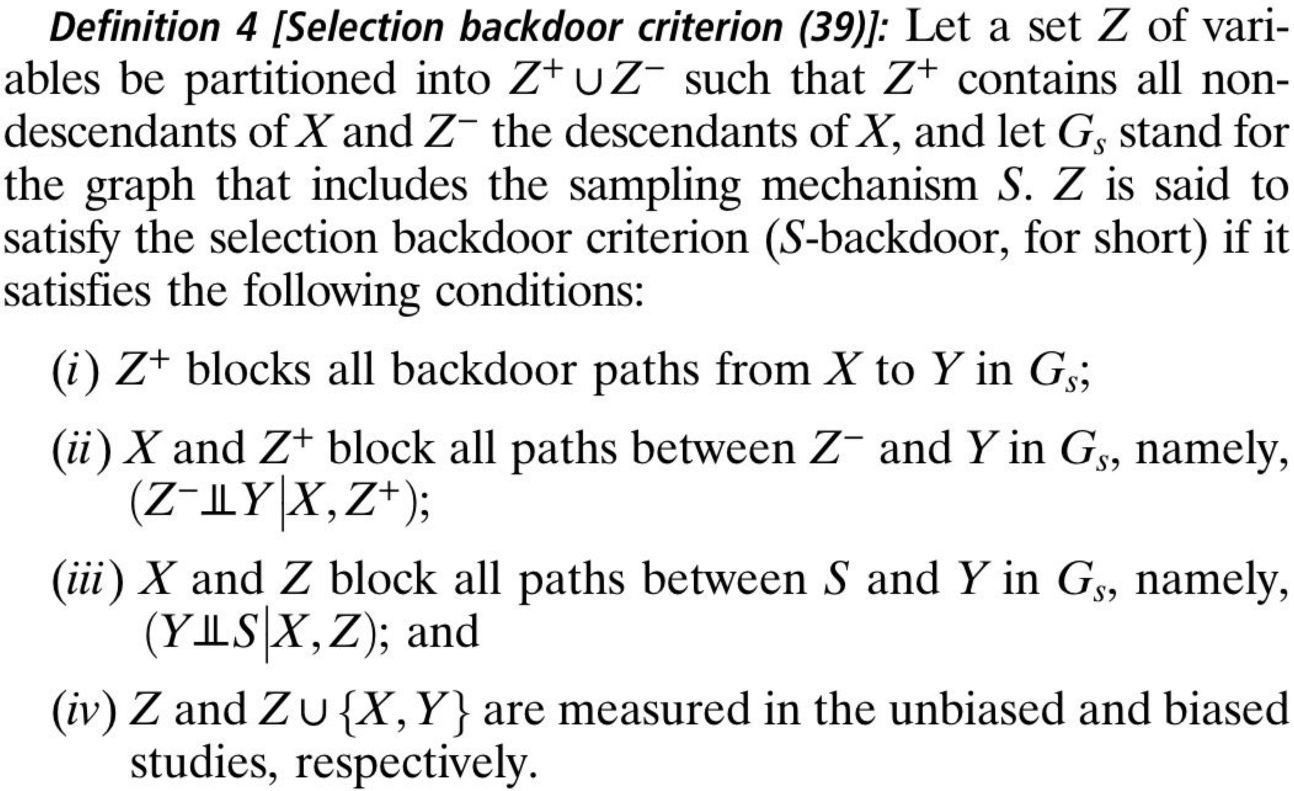
Under what conditions can Q be recovered?


STATISTICS CONFERENCE ~2022~
RAISE YOUR HAND IF YOU'RE FAMILIAR WITH SELECTION BIAS.
AS YOU CAN SEE, IT'S A TERM MOST PEOPLE KNOW...

https://m.xkcd.com/2618/

Selection Bias Examples





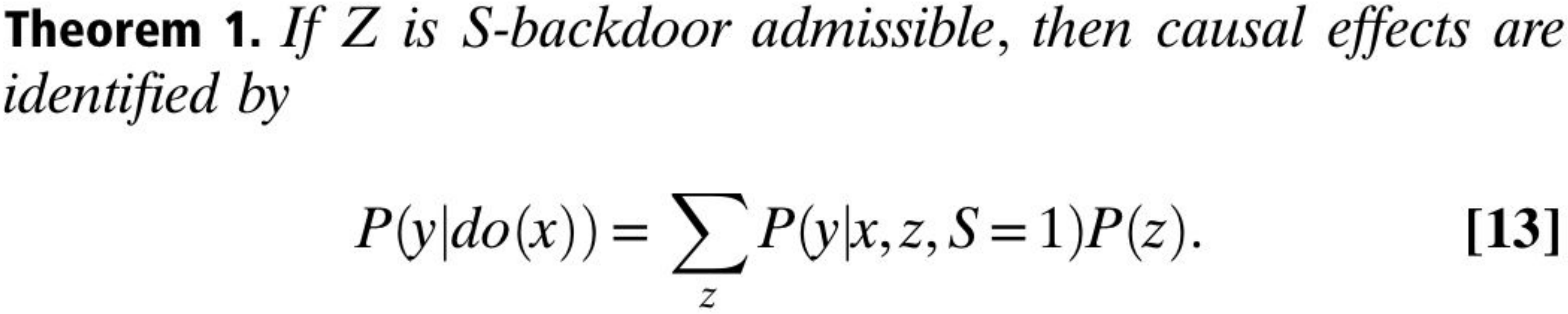
Definition 4 [Selection backdoor criterion (39)]: Let a set Z of variables be partitioned into $Z^{+} \cup Z^{-}$ such that Z^{+} contains all nondescendants of X and Z^{-} the descendants of X, and let G_s stand for the graph that includes the sampling mechanism S. Z is said to satisfy the selection backdoor criterion (S-backdoor, for short) if it satisfies the following conditions:

(i) Z^{+} blocks all backdoor paths from X to Y in G_s;

(ii) X and Z^{+} block all paths between Z^{-} and Y in G_s, namely, $(Z^{-} \perp\!\!\!\perp Y \mid X, Z^{+})$;

(iii) X and Z block all paths between S and Y in G_s, namely, $(Y \perp\!\!\!\perp S \mid X, Z)$; and

(iv) Z and $Z \cup \{X, Y\}$ are measured in the unbiased and biased studies, respectively.

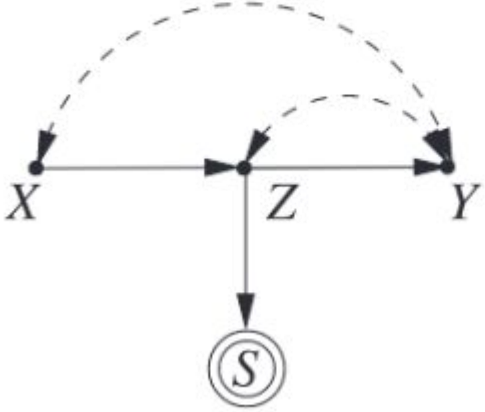


Theorem 1. *If Z is S-backdoor admissible, then causal effects are identified by*

$$P(y|do(x)) = \sum_{z} P(y|x, z, S=1)P(z). \qquad \textbf{[13]}$$

Selection Bias Example Problem

$$\begin{aligned} P(y|do(x)) &= \sum_z P(y|do(x),z)P(z|do(x)) \\ &= \sum_z P(y|do(x),z)P(z|x) \\ &= \sum_z P(y|do(x),z,S=1)P(z|x). \end{aligned}$$



Transportability and the Problem of Data Fusion

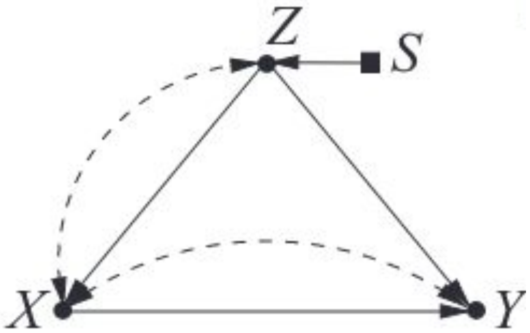
- Transportability – we want conclusions to apply to different scenarios.
  - A.K.A., heterogeneous treatment effect, or domain adaptation, as used in the field of machine vision.
  - This is a more general consideration of the sample selection problem…we want to understand results in a population that has different characteristics than the experimental/observational data we are looking at.

Example of Transportability

- **Experimental Data:** RCT conducted in Los Angeles to estimate air pollution exposure on health.
- **Query:** what is effect of air pollution on population of New York?
- **Problem:** Age distributions vary between the two cities.

Graph:

- X - air pollution exposure
- Y - health
- Z - age
- S - variable influencing age distribution



Goal: express query from experiments in source domain and observations in target domain (or, as represented in the text: $Q = P(y|do(x), S = s^*) = P^*(y|do(x))$

Solution to this query, using the graph from 5a:

$$
\begin{aligned}
Q &= \sum_z P(y|do(x), S = s^*, z) P(z|S = s^*, do(x)) \\
&= \sum_z P(y|do(x), z) P(z|S = s^*, do(x)) \\
&= \sum_z P(y|do(x), z) P(z|S = s^*) \\
&= \sum_z P(y|do(x), z) P^*(z),
\end{aligned}
$$

Paper Conclusions

- A lot of research condensed into this paper to create a generalized framework for dealing with general problems of data-fusion.
- Demonstrates how do-calculus can be used to deal with external validity (how well we can generalize results from experiments/observations).

Limitations:

- Authors did not discuss measurement bias

Further Resources:

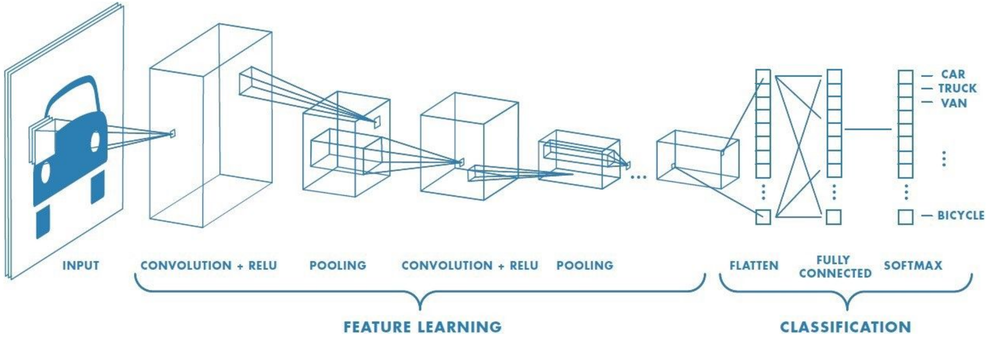
- Pearl, Judea. *Causality: Models, Inference, and Reasoning.* Chapter 3.

Causal Transportability for Visual Recognition

Convolutional Neural Networks

- Neural Network- flexible, black-box machine learning model great at modeling complex relationships

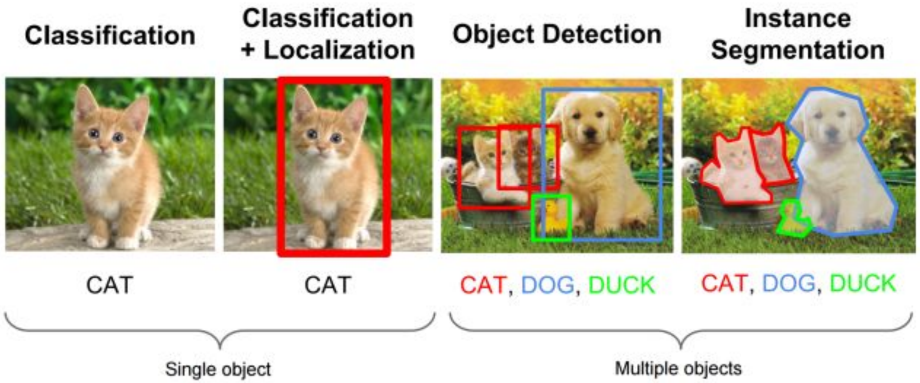
- Convolutions
  - Used to consider pixels in the context of the pixels are around it
  - Considers spatial locality, rather than treat each pixel as a “feature”



https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Neural Networks for Computer Vision

- CNNs can be trained to perform a variety of tasks (by taking advantage of different architectures and loss functions)



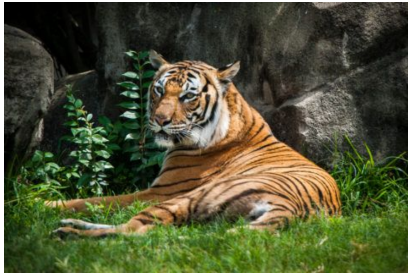
https://medium.com/swlh/object-detection-and-instance-segmentation-a-detailed-overview-94ca109274f2

Example: Robust vs Non-Robust Features in Images

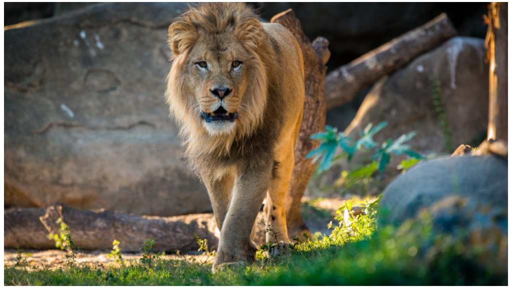
- If I were training a classifier to classify lions vs tigers, what would some robust and non-robust features be?

Lion Training Images:

Tiger Training Image:



Testing Image:

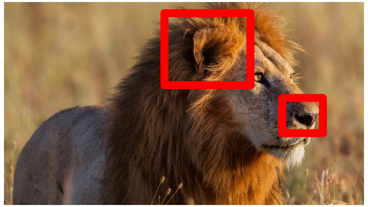


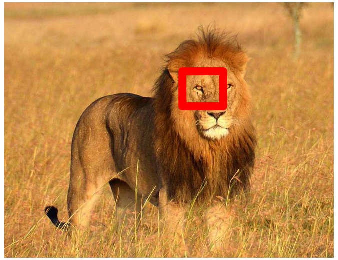
https://www.houstonzoo.org/blog/houston-zoos-elderly-female-tiger-dies/
https://www.naturalworldsafaris.com/wildlife/african-lions
https://www.houstonzoo.org/explore/animals/lion-african/
https://www.travelandleisure.com/trip-ideas/safaris/lion-king-themed-family-safari

Example: Robust Features

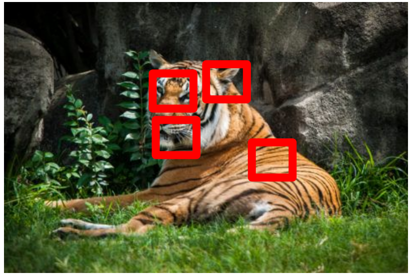
- If I were training a classifier to classify lions vs tigers, what would some robust features be?

Lion Training Images:

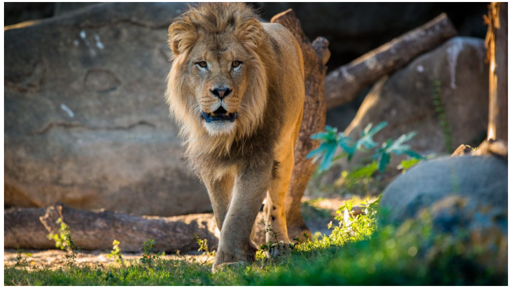




Tiger Training Image:



Testing Image:

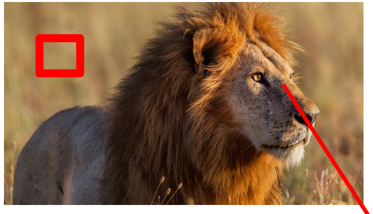


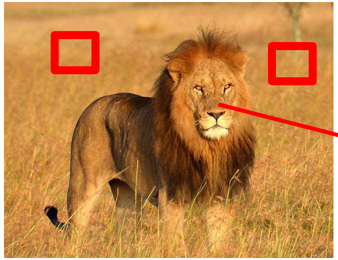
https://www.houstonzoo.org/blog/houston-zoos-elderly-female-tiger-dies/
https://www.naturalworldsafaris.com/wildlife/african-lions
https://www.houstonzoo.org/explore/animals/lion-african/
https://www.travelandleisure.com/trip-ideas/safaris/lion-king-themed-family-safari

Example: Non-Robust Features

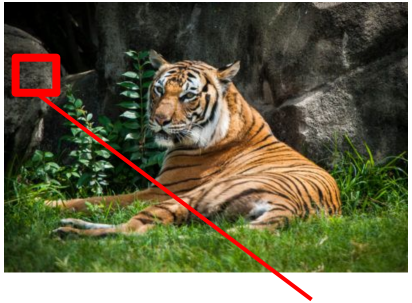
- If I were training a classifier to classify lions vs tigers, what would some non-robust features be?

Lion Training Images:





Tiger Training Image:



Testing Image:

Angle/pose in relation to camera

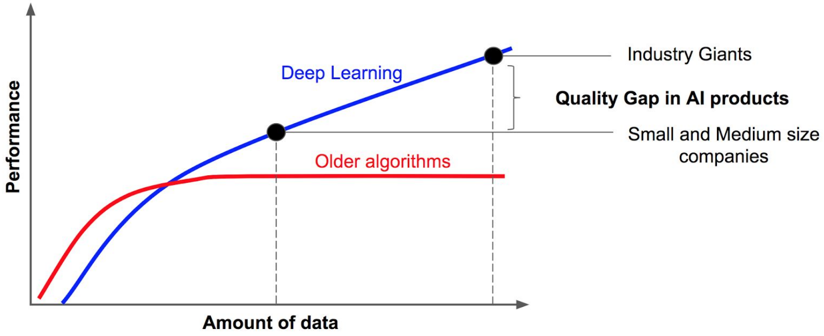
Image background

Image source (and resolution)

https://www.houstonzoo.org/blog/houston-zoos-elderly-female-tiger-dies/
https://www.naturalworldsafaris.com/wildlife/african-lions
https://www.houstonzoo.org/explore/animals/lion-african/
https://www.travelandleisure.com/trip-ideas/safaris/lion-king-themed-family-safari

Example: Non-Robust Features

- Without an abundance of training imagery to cover all possible gaps, there will often be a domain gap
    - Sampling variability
    - Legitimate differences in the distribution between the training and testing data

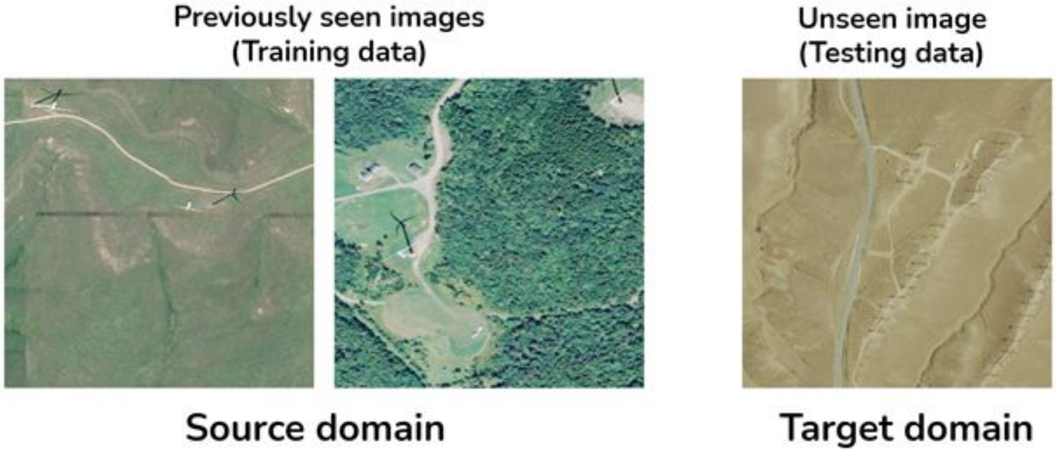


https://datascience.stackexchange.com/questions/45774/where-does-the-deep-learning-needs-big-data-rule-come-from

Domain Adaptation in Computer Vision: Real Use Cases

Example:

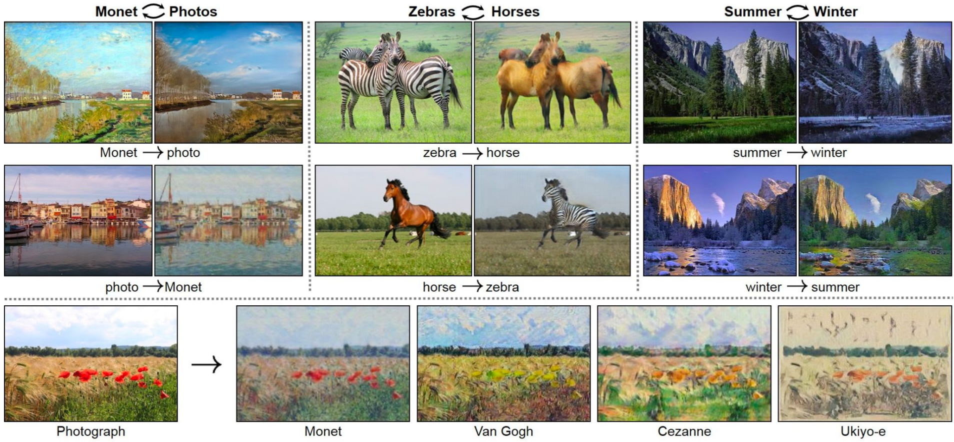
- Train a classifier the abundance of high-quality, labeled satellite imagery in the USA to detect energy infrastructure
- Apply classifier to predict on imagery throughout the world



https://duke-bc-2021-ai-for-energy-access.github.io/BC-2021-AI-for-energy-access/

Current Approaches to Domain Gap in Computer Vision

- Self-supervised learning
- Generative data augmentations: CycleGAN and CyCADA
- Adversarial self challenging



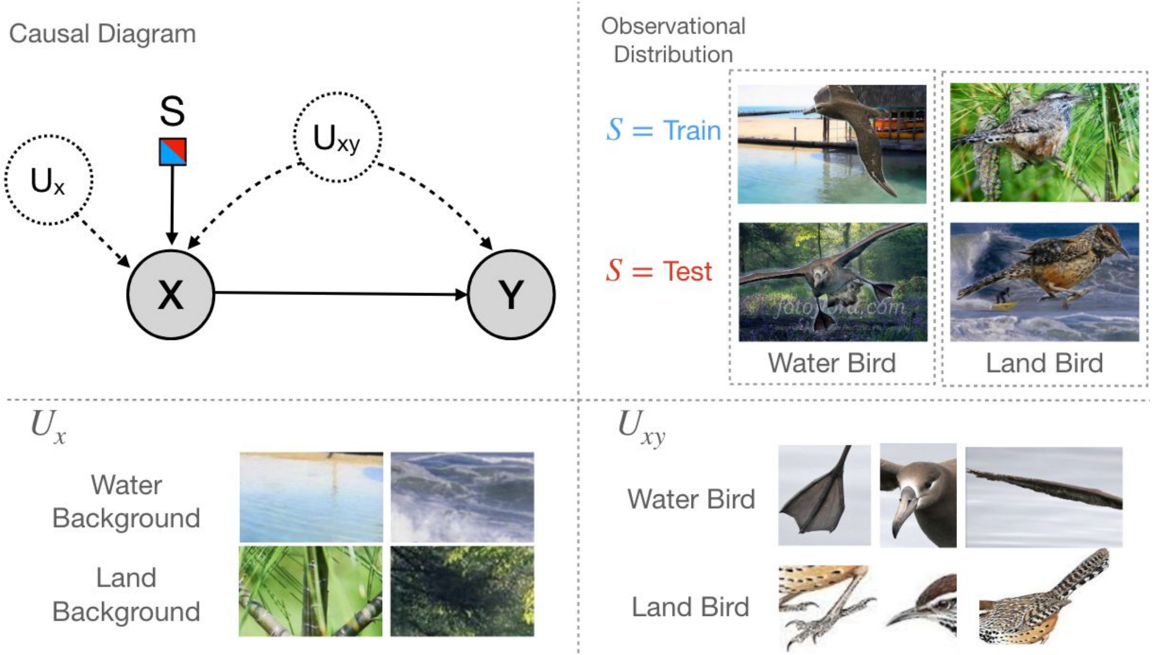
https://junyanz.github.io/CycleGAN/

Domain Adaptation in Computer Vision

U_X - Nuisance features

U_{XY} - Concept features

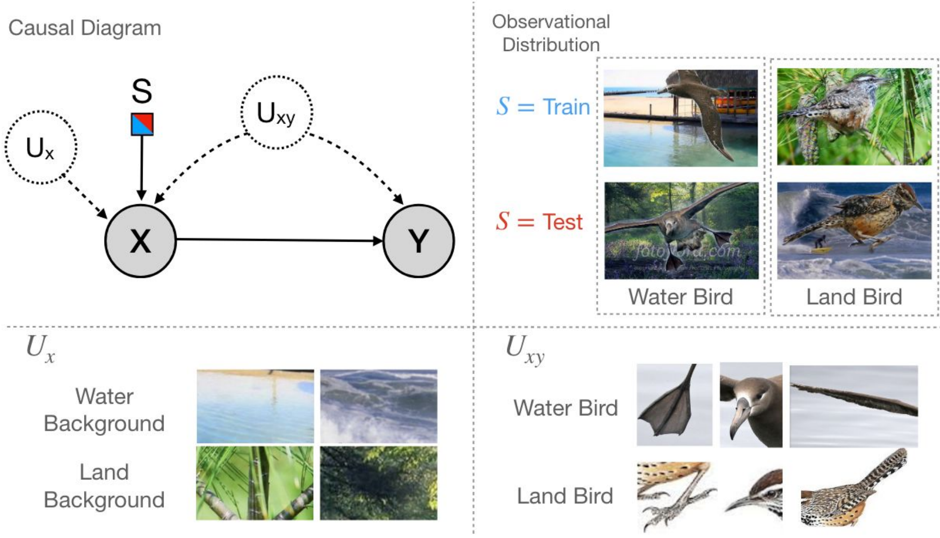
- Want model to learn to detect land birds and water birds by their features, not their backgrounds



Causal Image Recognition

Structural Causal Model:

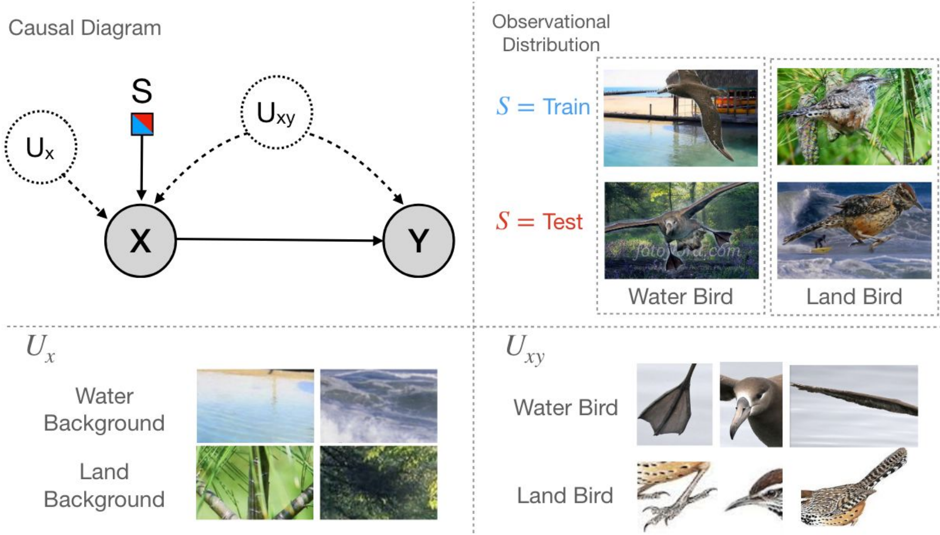
- $V = \{X, Y\}$
  - Observed variables (image, label)
- $U = \{U_X, U_{XY}\}$
  - Unobserved variables (nuisance factors, concept vector)
- $F = \{f_X, f_Y\}$
  - $X \leftarrow f_X(U_X, U_{XY})$
  - $Y \leftarrow f_Y(X, U_{XY})$
- P(U)
  - Probability distribution over unobserved variables
  - Underlying distribution combines with F to induce distribution over P(X, Y)



Causal Image Recognition

Intuition:

- U_{XY}
  - Underlying factors that produce core features of image and label
  - Example: Flippers and wing of bird
- U_X
  - Nuisance factors (e.g. background)
  - Affect image generation process

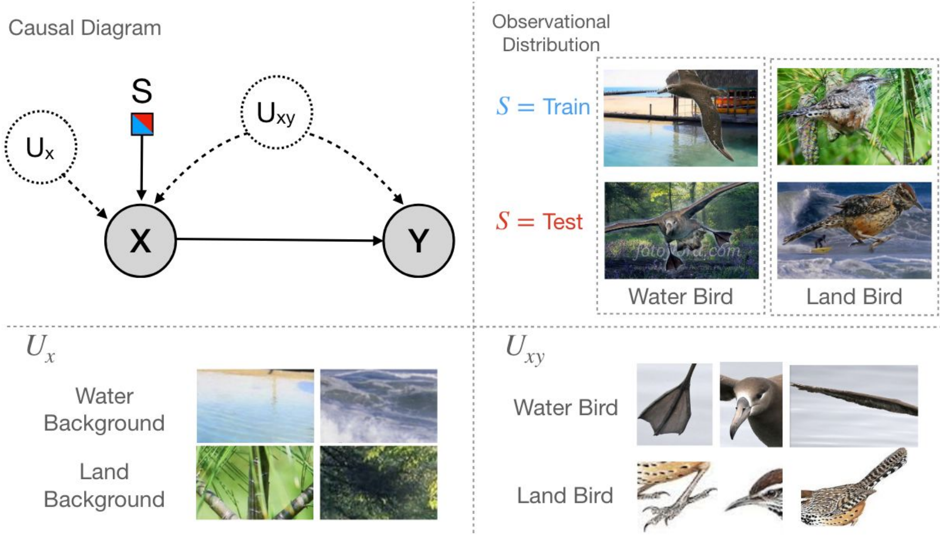


Causal Image Recognition

Intuition:

- f_X
  - Generation process of X
  - Translate underlying factors "flippers", "wing", into "waterbird"
  - PROBLEM: Flippers may be associated with water background

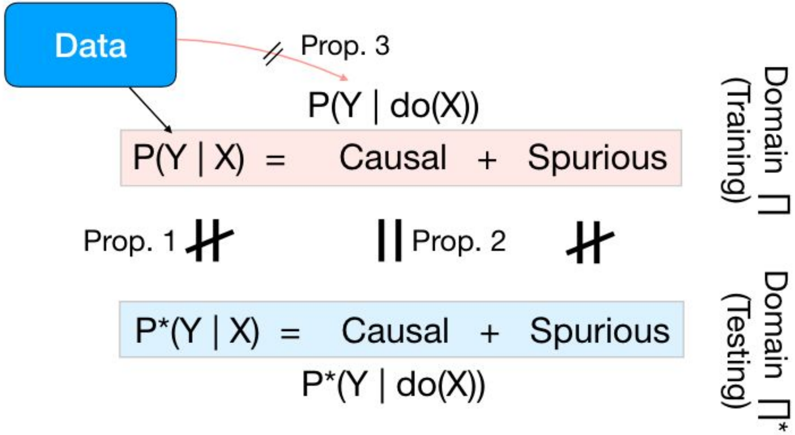
- f_Y
  - Generation process of Y
  - Someone labeling X who understands birds via U_{XY}



Propositions about SCMs

The transportability problem (out of distribution problems):

- Training domain different from testing domain
- Can assume labeling process consistent, generative process for image may differ



Data

Prop. 3

$P(Y \mid do(X))$

$P(Y \mid X) = \text{Causal} + \text{Spurious}$

Prop. 1 $\nparallel$ $\parallel$ Prop. 2 $\nparallel$

$P^*(Y \mid X) = \text{Causal} + \text{Spurious}$

$P^*(Y \mid do(X))$

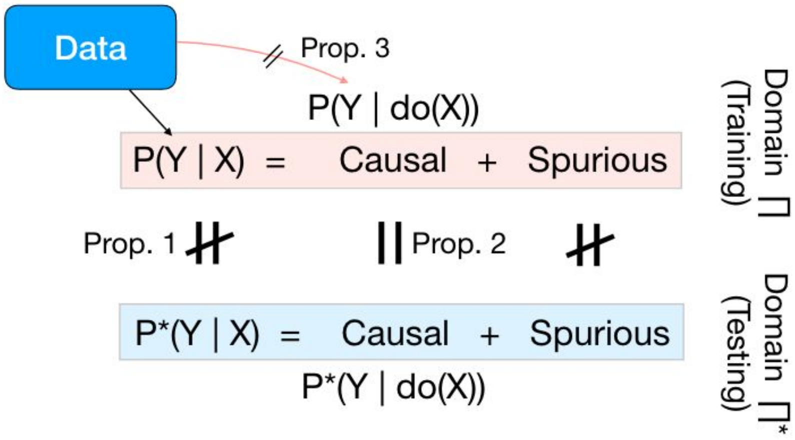
Domain Π (Training)

Domain Π^* (Testing)

Propositions about SCMs

The transportability problem (out of distribution problems):

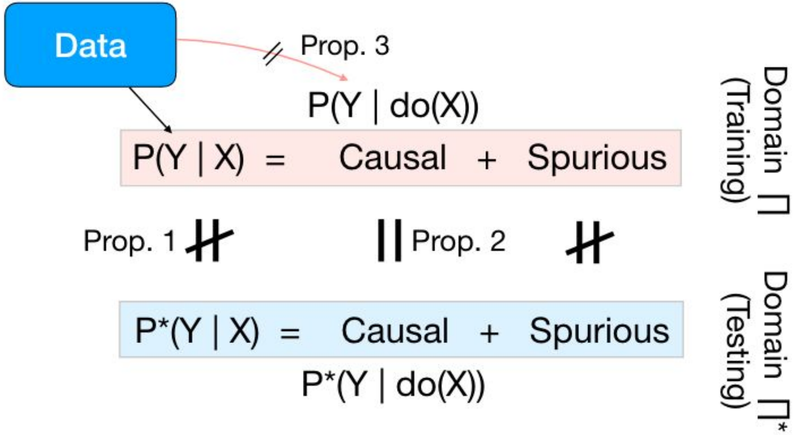
- Model trained to learn to predict label Y after being trained on images X
- This does not transport to predicting labels from a different image distribution (unobserved confounding effects)



Propositions about SCMs

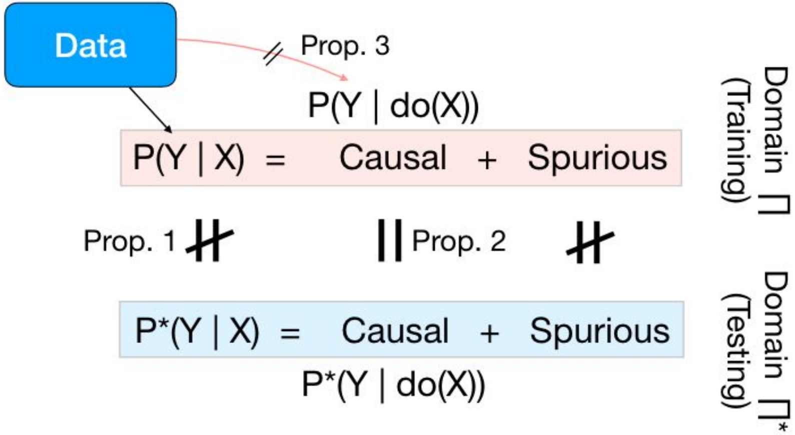
The transportability problem (out of distribution problems):

- Causal effect invariant ("transportable")
- The causal effect P(Y | do(X)) can be used as a suitable proxy (surrogate model) for a classifier for the target domain



Identifiability

- Causal effect P(Y | do(X)) is not identifiable
- Multiple SCMs consistent with probability distribution P(X, Y)
    - Can not deduce the “true” causal effect
- New goal: Identify effect of X on Y without knowing backdoor variables



Neural Representation Approach

- With some select assumptions and do-calculus, we can compute the causal effect directly
- Now just need Neural Networks to learn the components of the formula that go into the causal effect

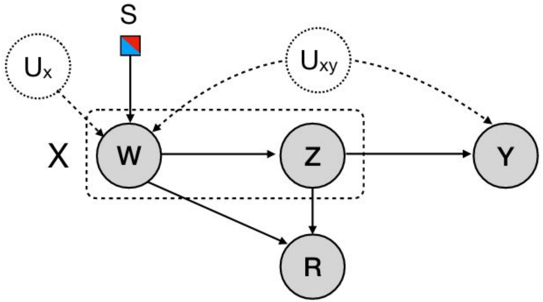
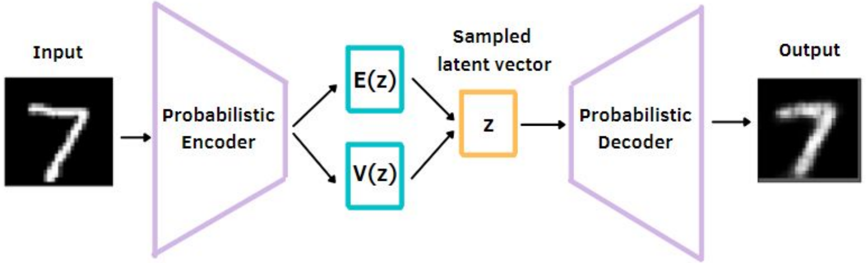


Figure 3. Expanded causal model with decomposition of image X and representation R. Gray nodes denote observed variables.

Neural Representation Approach

- Neural Network learns latent representations from images that preserve causal effect (can use unsupervised learning or pre-trained networks)

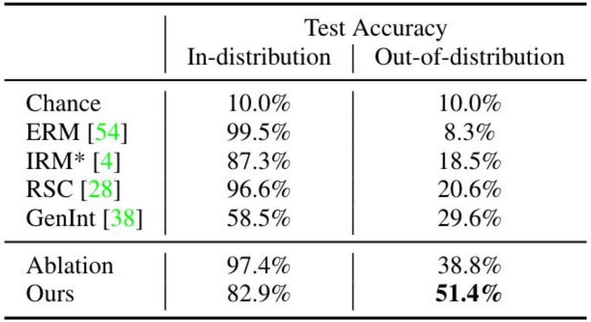
- Another network makes classification based on high-level features of representations and low-level features of images



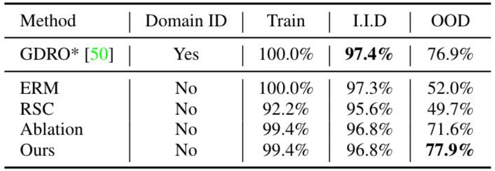
https://medium.com/dataseries/variational-autoencoder-with-pytorch-2d359cbf027b

Experimental Setup

- Popular datasets
  - WaterBird, ImageNet-Rendition, ImageNet-Sketch
- Baselines
  - State-of-the-art: GenInt, RSC, IRM
  - “Ours”- NN with 3 random convolution layers, 2 layer FC network to predict Y
    - $N_j = 256$, $N_i = 10$
  - Ablation-
    - $N_j = 1$, $N_i = 1$

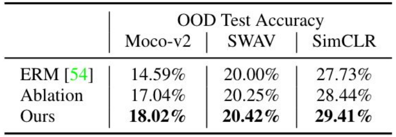


	Test Accuracy	
	In-distribution	Out-of-distribution
Chance	10.0%	10.0%
ERM [54]	99.5%	8.3%
IRM* [4]	87.3%	18.5%
RSC [28]	96.6%	20.6%
GenInt [38]	58.5%	29.6%
Ablation	97.4%	38.8%
Ours	82.9%	**51.4%**



Method	Domain ID	Train	I.I.D	OOD
GDRO* [50]	Yes	100.0%	**97.4%**	76.9%
ERM	No	100.0%	97.3%	52.0%
RSC	No	92.2%	95.6%	49.7%
Ablation	No	99.4%	96.8%	71.6%
Ours	No	99.4%	96.8%	**77.9%**

Table 2. Accuracy on the WaterBird dataset. Our causal method improves ERM model’s worst group OOD generalization significantly. Our approach achieves performance on par with group invariant training (GDRO) without needing the domain index.

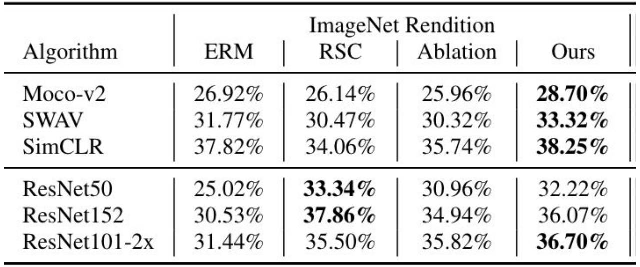


	OOD Test Accuracy		
	Moco-v2	SWAV	SimCLR
ERM [54]	14.59%	20.00%	27.73%
Ablation	17.04%	20.25%	28.44%
Ours	**18.02%**	**20.42%**	**29.41%**

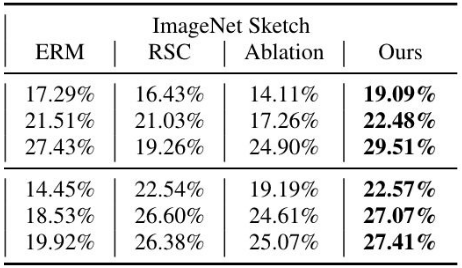
Table 3. Accuracy on the Imagenet-9 adversarial backgrounds.

Experimental Results

- Model outperformed baseline and state-of-the-art in contrastive learning representations (ImageNet-Sketch)
- Model performed better in 2/4 supervised learning representations (ImageNet Rendition)
- OOD generalization improves as N_j increases
- Estimates causal effect from representation



	ImageNet Rendition			
Algorithm	ERM	RSC	Ablation	Ours
Moco-v2	26.92%	26.14%	25.96%	**28.70%**
SWAV	31.77%	30.47%	30.32%	**33.32%**
SimCLR	37.82%	34.06%	35.74%	**38.25%**
ResNet50	25.02%	**33.34%**	30.96%	32.22%
ResNet152	30.53%	**37.86%**	34.94%	36.07%
ResNet101-2x	31.44%	35.50%	35.82%	**36.70%**



ImageNet Sketch			
ERM	RSC	Ablation	Ours
17.29%	16.43%	14.11%	**19.09%**
21.51%	21.03%	17.26%	**22.48%**
27.43%	19.26%	24.90%	**29.51%**
14.45%	22.54%	19.19%	**22.57%**
18.53%	26.60%	24.61%	**27.07%**
19.92%	26.38%	25.07%	**27.41%**

GradCAM Results

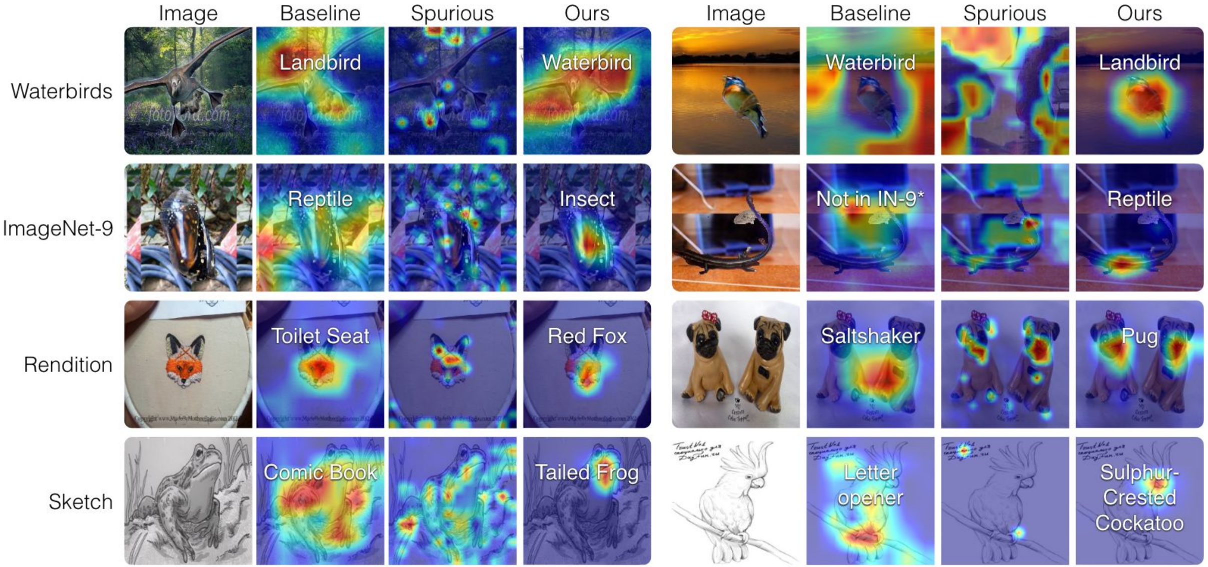


Figure 5. We visualize the input regions that the models use for prediction. We use GradCAM [51] and highlight the the discriminative regions that the model relies on with red. The white text shows the model's prediction. The correlation based ERM method often attends to spurious background context. By marginalizing over the spurious features (visualized in the Spurious column), our model captures the right, causal features, which predict the right thing for the right reason.

Paper Conclusions

- Integrating causal knowledge and tools and applying causal transportability theory towards the challenge of generalization in computer vision can improve out of distribution robustness

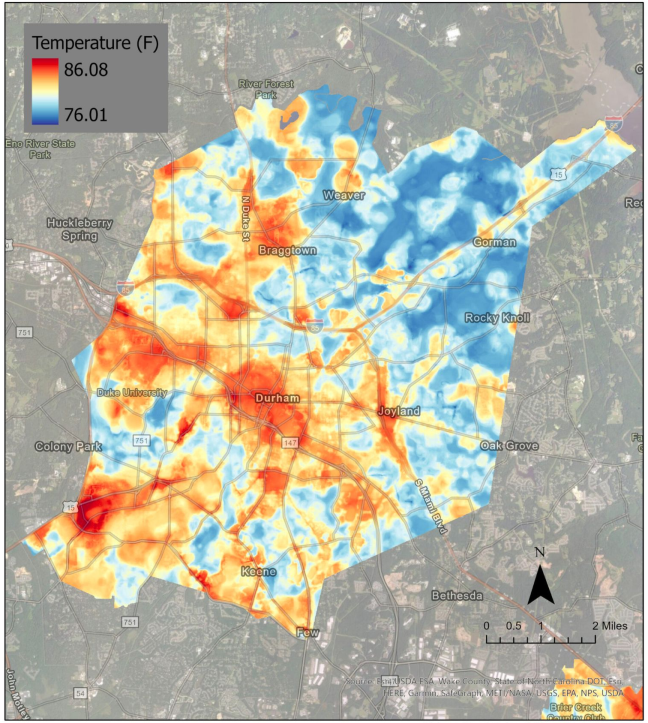
Our Research

Urban Heat Islands

- Urban areas are much hotter than rural areas in the summer.
- There are known interventions (e.g., planting trees), but their efficacy depends on a lot of variables, such as the local climate.

Durham Evening Temperature (7-8pm)
July 23, 2021

NORTH CAROLINA CLIMATE OFFICE



Source: Esri, USDA FSA, Wake County, State of North Carolina DOT, Esri, HERE, Garmin, SafeGraph, METI/NASA, USGS, EPA, NPS, USDA

Urban Heat Island Mapping
climate.ncsu.edu/research/uhi

We want a transportable model

- Can we create a model to measure the urban heat island effect more generally? A lot of models cannot generalize across diverse climates.
- Can we understand the impact of various interventions as a function of the local climate (or other variables?)

Appendix

Neural Representation Approach

Two NN models:

- P^(R | X)
  - Generates visual representations R from images X
- P^(Y | R, X)
  - Classifies Y from visual representations, images

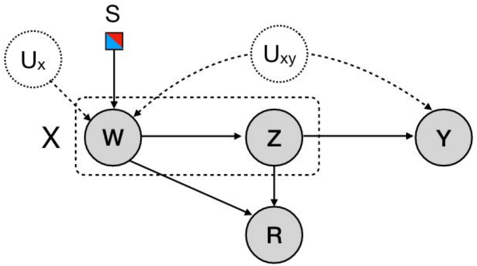


Figure 3. Expanded causal model with decomposition of image X and representation R. Gray nodes denote observed variables.

Neural Representation Approach

Assumption 1: Decomposition

- Each image X can be decomposed into causal (Z) and spurious factors (W)
  - X = (Z, W)
  - W contains lower level signal of image (may confound with Y)
  - Z refines patches into interpretable factors, used by labeler
- Generative process follows causal graph shown

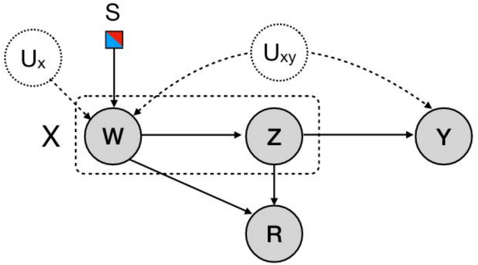


Figure 3. Expanded causal model with decomposition of image X and representation R. Gray nodes denote observed variables.

Neural Representation Approach

Assumption 2: Sufficient Representation

- Neural representations P^(R | Z, W) are learned such that they do not lose information with respect to Z
    - P^(R | Z, W) will be different for different values of Z
    - Unambiguously represents the causal factors

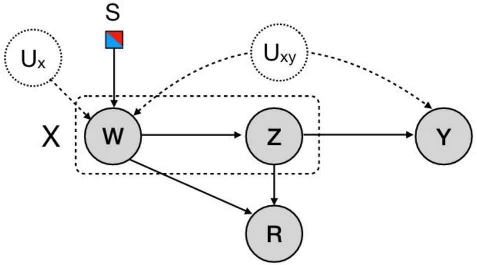


Figure 3. Expanded causal model with decomposition of image X and representation R. Gray nodes denote observed variables.

Neural Representation Approach

Assumption 3: Selective prediction

- P^(Y = y | R = r, X = x’) = P(y | z, w’)
  - LHS: Probability of neural output being y
    - Given sample from neural representation of first image (P^(R | x)) + true second image (x’ = (z’, w’))
  - RHS: True labeling probability of class y given the causal factor of first image and spurious factors of the second image
- These will be the same!

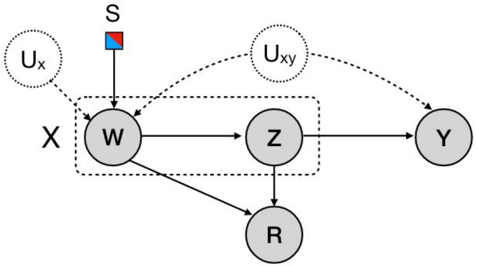
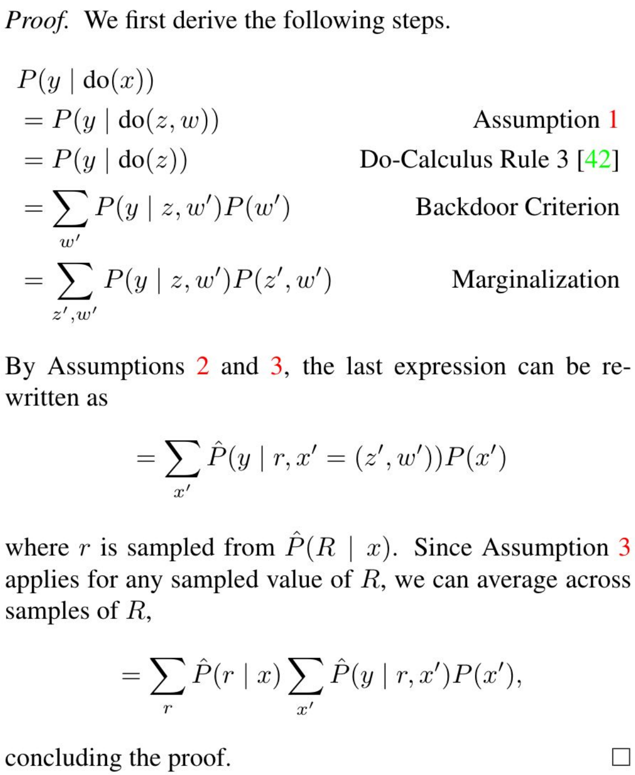


Figure 3. Expanded causal model with decomposition of image X and representation R. Gray nodes denote observed variables.

Neural Representation Approach

Given these assumptions, we can directly compute using R

- P(y | do(X)) = ∑ rˆ P (r|x) ∑ x′ ˆ P (y|r, x′)P (x′)
    - P^ (Y | R, X) extracts causal information from the representation and spurious information from second image
    - Assume P(X) is sampled from uniform distribution



Proof. We first derive the following steps.

$$
\begin{aligned}
& P(y \mid \mathrm{do}(x)) \\
& = P(y \mid \mathrm{do}(z, w)) && \text{Assumption 1} \\
& = P(y \mid \mathrm{do}(z)) && \text{Do-Calculus Rule 3 [42]} \\
& = \sum_{w'} P(y \mid z, w') P(w') && \text{Backdoor Criterion} \\
& = \sum_{z', w'} P(y \mid z, w') P(z', w') && \text{Marginalization}
\end{aligned}
$$

By Assumptions 2 and 3, the last expression can be re-written as

$$= \sum_{x'} \hat{P}(y \mid r, x' = (z', w')) P(x')$$

where r is sampled from $\hat{P}(R \mid x)$. Since Assumption 3 applies for any sampled value of R, we can average across samples of R,

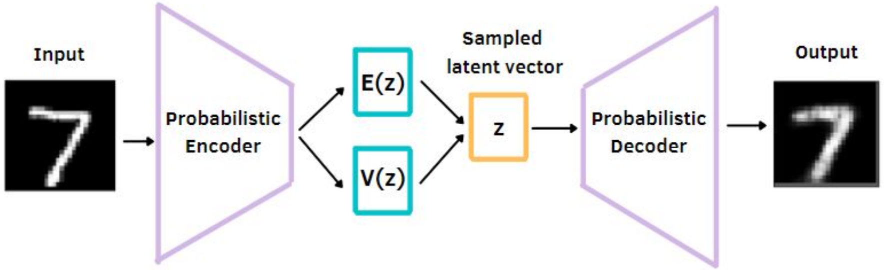
$$= \sum_{r} \hat{P}(r \mid x) \sum_{x'} \hat{P}(y \mid r, x') P(x'),$$

concluding the proof. □

Phase 1: How can we construct P(R | X)?

Several ways to estimate P^(R | X) while satisfying Assumption 2:

- Variational Auto-Encoders
  - Unsupervised
- Contrastive Learning
  - Unsupervised
- Pretrained models from larger dataset
  - Supervised



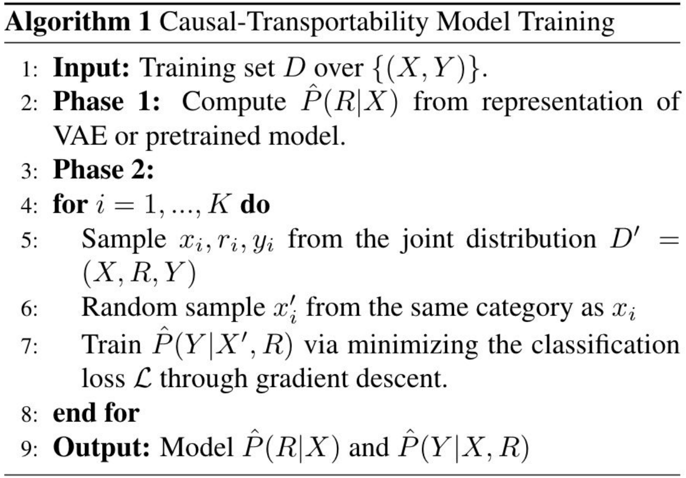
Phase 2: How can we construct P(Y | R, X)?

Several ways to estimate P^(Y | R, X) while satisfying Assumption 3:

- Use a bag of patches subsampled from input image X as input
    - Corrupts global shape information
    - Contains local, spurious features
- Model has limited capacity
    - Learned information about W
    - Learn Z from representation r, ignore W from representation
- Uses low-level features from X and high-level features from latent representation R

How to Train Your Causal-Transportability Model

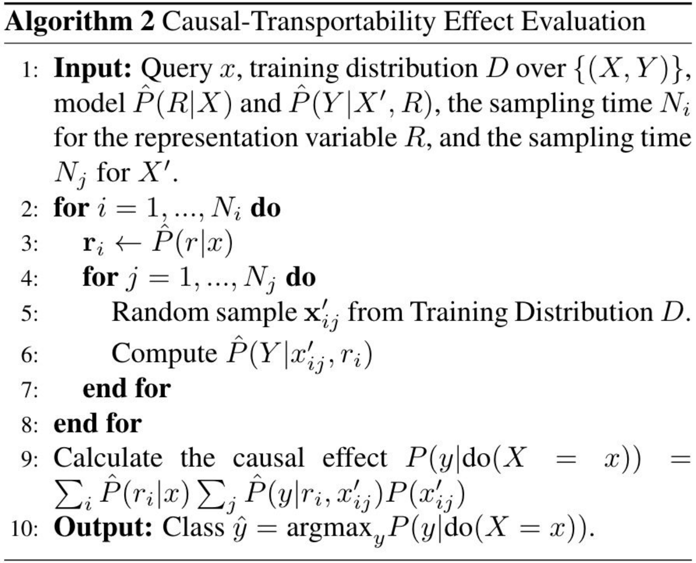
- Phase 1
  - Train representation with VAE/constrastive learning or use representation from pre-trained model
- Phase 2
  - Train P^(Y | R, X) from sample random images X



```
Algorithm 1 Causal-Transportability Model Training
Input: Training set $D$ over $\{(X, Y)\}$.
Phase 1: Compute $\hat{P}(R|X)$ from representation of VAE or pretrained model.
Phase 2:
for $i = 1, ..., K$ do
    Sample $x_i, r_i, y_i$ from the joint distribution $D' = (X, R, Y)$
    Random sample $x_i'$ from the same category as $x_i$
    Train $\hat{P}(Y|X', R)$ via minimizing the classification loss $\mathcal{L}$ through gradient descent.
end for
Output: Model $\hat{P}(R|X)$ and $\hat{P}(Y|X, R)$
```

Causal-Transportability Model Inference

- Randomly sample representation R
- for r in R:
  - Sample images X from random categories
  - Make prediction based on our direct computation
    - P(y | do(X)) = ∑ rˆ P (r|x) ∑ x′ ˆ P (y|r, x′)P (x′)
    -



Algorithm 2 Causal-Transportability Effect Evaluation

```
Input: Query $x$, training distribution $D$ over $\{(X, Y)\}$, model $\hat{P}(R|X)$ and $\hat{P}(Y|X', R)$, the sampling time $N_i$ for the representation variable $R$, and the sampling time $N_j$ for $X'$.
for $i = 1, ..., N_i$ do
    $\mathbf{r}_i \leftarrow \hat{P}(r|x)$
    for $j = 1, ..., N_j$ do
        Random sample $\mathbf{x}'_{ij}$ from Training Distribution $D$.
        Compute $\hat{P}(Y|x'_{ij}, r_i)$
    end for
end for
Calculate the causal effect $P(y|\text{do}(X = x)) = \sum_i \hat{P}(r_i|x) \sum_j \hat{P}(y|r_i, x'_{ij}) P(x'_{ij})$
Output: Class $\hat{y} = \text{argmax}_y P(y|\text{do}(X = x))$.
```