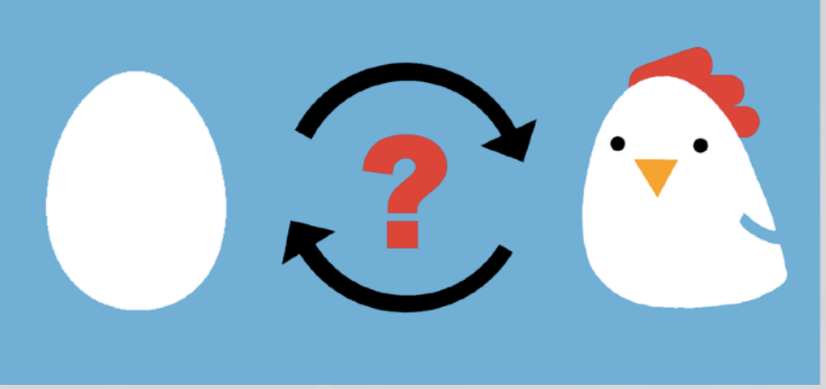


Causality and Machine Learning
(80-816/516)

Classes 22 (April 8, 2025)

Transfer learning (and image translation): From Traditional to Emerging Approaches

Instructor:

Kun Zhang (kunz1@cmu.edu)

Zoom link: https://cmu.zoom.us/j/8214572323)

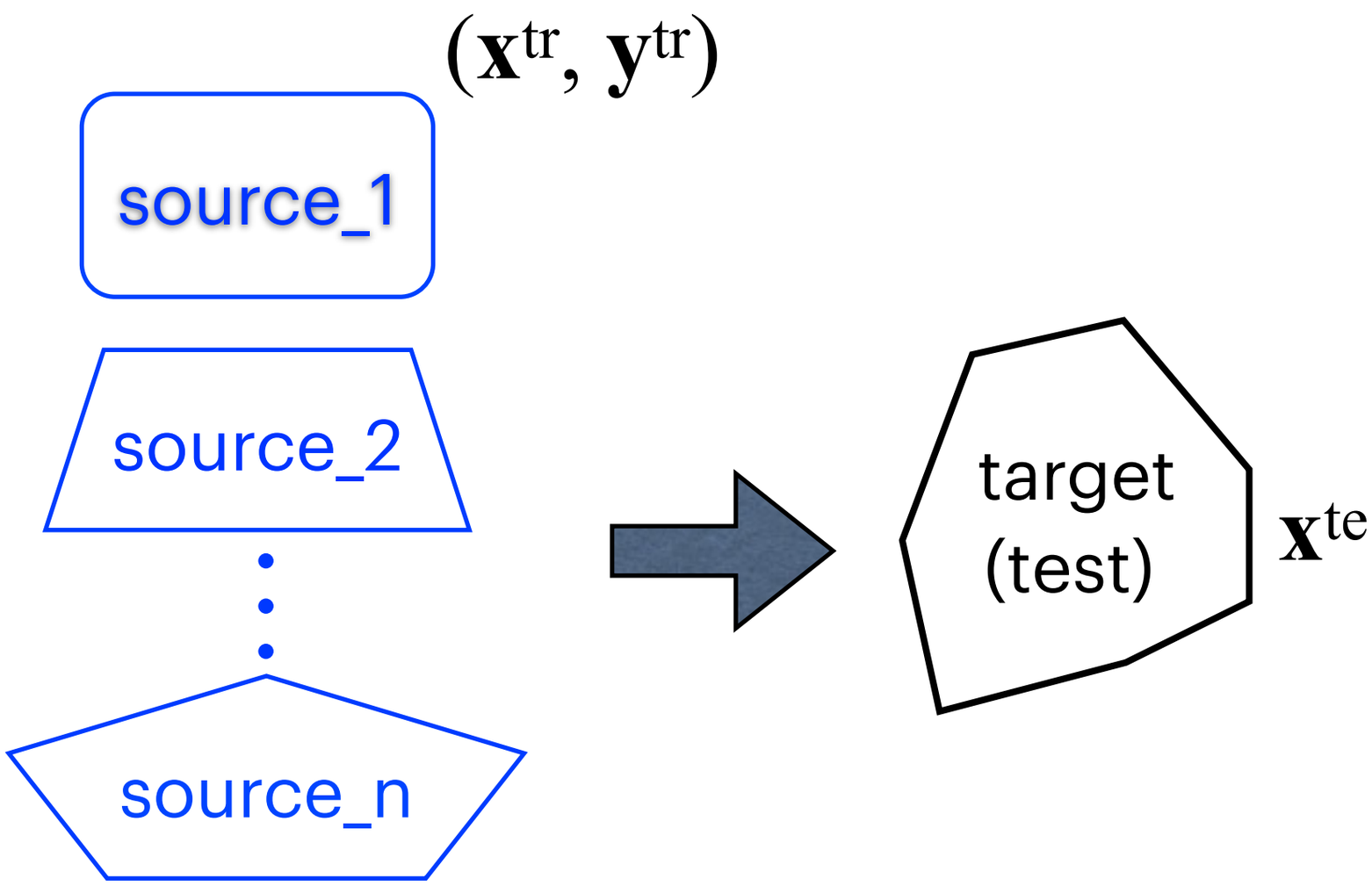
Office Hours: W 3:00–4:00PM (on Zoom or in person); other times by appointment

Outline

- Domain adaptation / transfer learning: What and Why?
- Traditional approaches to domain adaptation
- Adaptive methods for domain adaptation
  - Related problems: image translation, multi-domain generation
- Future of domain adaptation

Domain Adaptation

- Traditional supervised learning: $P_{XY}^{te} = P_{XY}^{tr}$
- Might not be the case in practice
- How to leverage information in source domains?



high-level representation, e.g., $Y \rightarrow X$

Prob. model $P^{(1)}(X,Y)$, $P^{(2)}(X,Y)$, $P^{(3)}(X,Y)$, ... $P^{(k)}(X,Y)$...

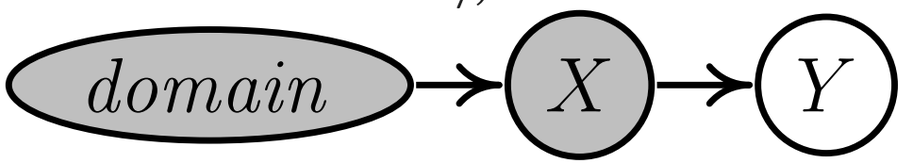
Outline

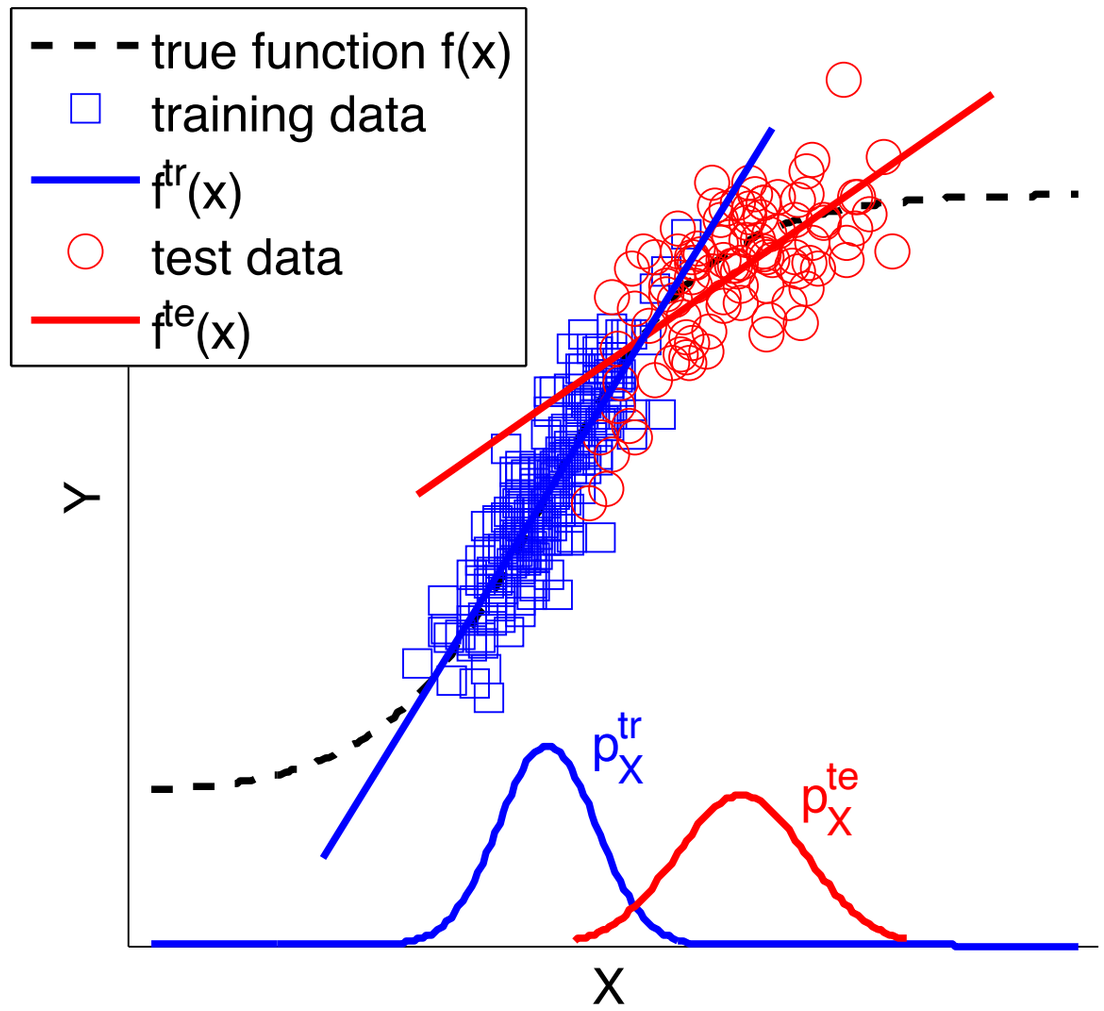
- Domain adaptation / transfer learning: What and Why?
- Traditional approaches to domain adaptation
- Adaptive methods for domain adaptation
  - Related problems: image translation, multi-domain generation
- Future of domain adaptation

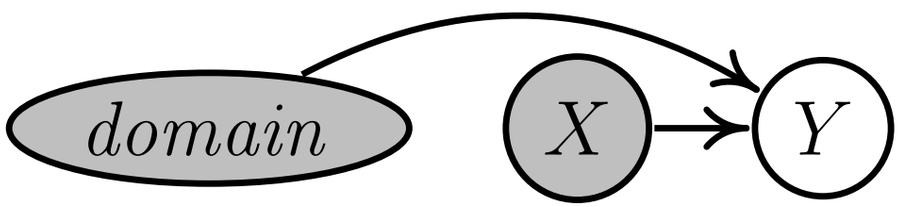
Possible Situations for Domain Adaptation: When $\mathbf{X} \rightarrow Y$

covariate shift

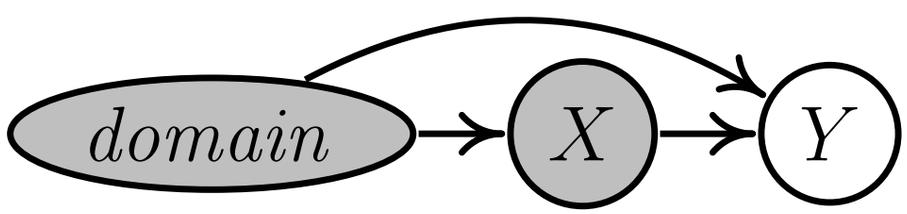
(Shimodaira00; Sugiyama etal.08; Huang etal.07, Gretton etal.08...)



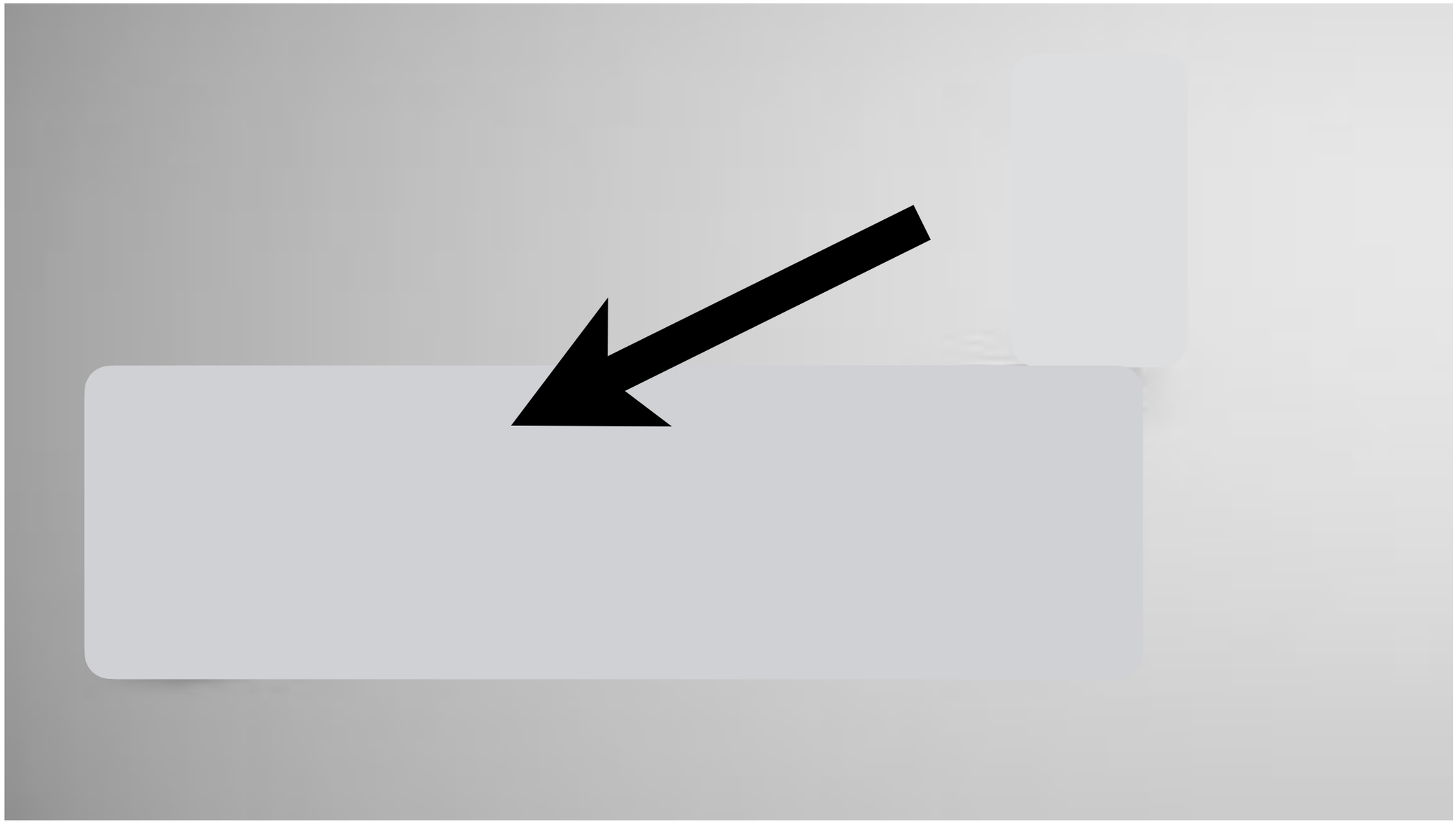




☹ no clue as to find $P^{te}_{Y|X}$ (with one source domain)



Causality may Matter in Prediction: An Illustration



5

5

Understanding connections between different scenarios & *modeling* differences

What Features/Components to Transfer?

- Invariant cause distribution (Zhang et al., ICML'13)

- Invariant/transferrable causal mechanism (Zhang et al., ICML'13; AAAI'14; Gong et al, ICML'16): invariance of $P(X^{ct} \mid Y)$

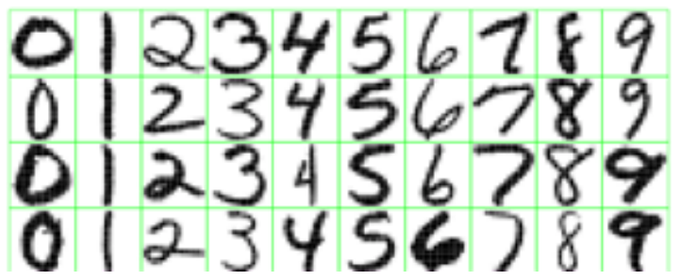
- Nonparametric transfer learning (Stojanov et al. AISTATS'19; Gong et al; ICML'18; Zhang et al., NeurIPS'20)

  - *Detect, model, utilize* changes

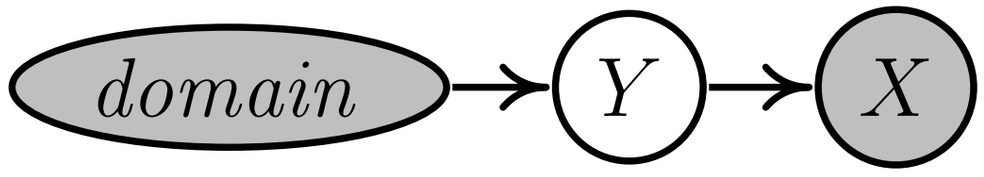
- Even if one aims to find invariant representation, the transformation is domain-specific (Stojanov et al., NeurIPS'21)

Possible Situations for Domain Adaptation: When Y→**X** (Zhang et al., 2013)

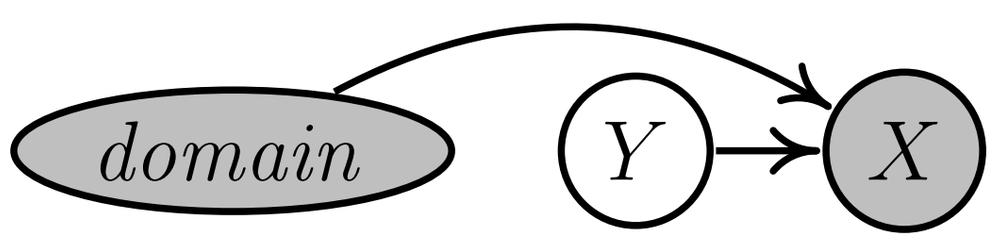
- Y is usually the cause of X
  (especially for classification)



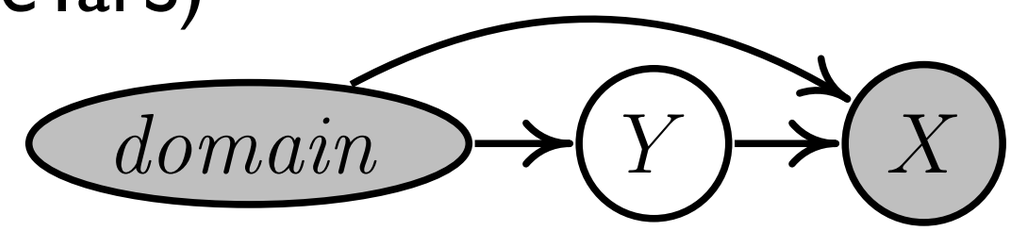
- Target shift (TarS)



- Conditional shift (ConS)



- Generalized target shift (GeTarS)

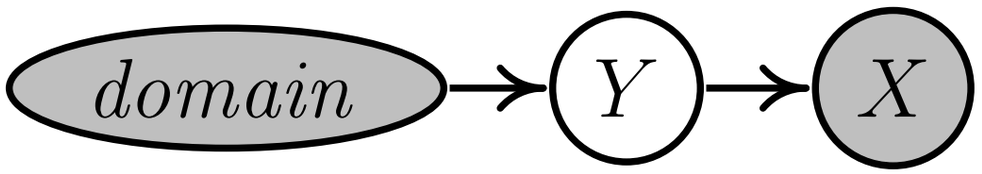


P_X^{te} helps find $P_{Y|X}^{te}$

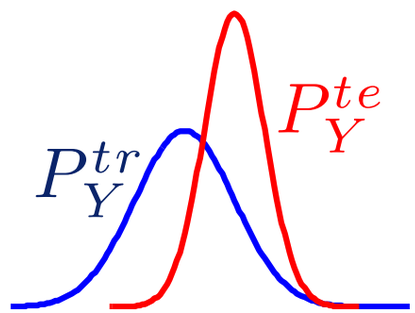
involved parameters estimated by matching P_X

Zhang et al., ICML 2013; Schölkopf et al., 2012; Zhang et al., AAAI 2015; Gong et al., ICML 2016; Stojanov et al., AISTATS 2018; Zhao et al., ICML 2019; Fu et al., CVPR 2019…

Target shift



- $P_Y^{te} \neq P_Y^{tr}$, but $P_{X|Y}^{te} = P_{X|Y}^{tr}$, and furthermore
  - **richness**: the support of P_Y^{tr} is richer
  - **invertibility**: only one $P_Y \xrightarrow{P_{X|Y}^{tr}} P_X^{te}$



- Find the learning machine on test domain by importance reweighting

$$R[P^{te}, \theta, l(x,y,\theta)] = \mathbb{E}_{(X,Y)\sim P^{te}}[l(x,y,\theta)] = \int P_{XY}^{tr} \cdot \frac{P_{XY}^{te}}{P_{XY}^{tr}} \cdot l(x,y,\theta)dxdy$$

$$= \mathbb{E}_{(X,Y)\sim P^{tr}} \cdot \underbrace{\frac{P_Y^{te}}{P_Y^{tr}}}_{\triangleq \beta^*(y)} \cdot \underbrace{\frac{P_{X|Y}^{te}}{P_{X|Y}^{tr}}}_{\triangleq \gamma^*(x,y)} \cdot l(x,y,\theta)dxdy, \quad \equiv 1\ !$$

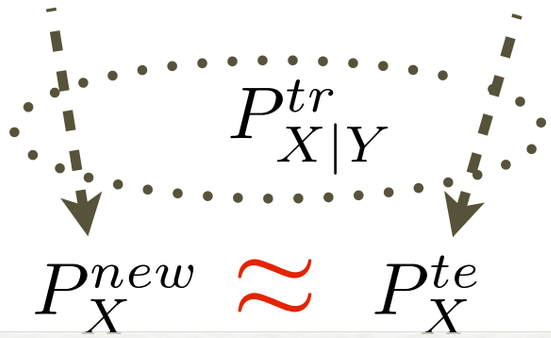
$$\widehat{R} = \frac{1}{m}\sum_{i=1}^{m} \beta^*(y_i^{tr})\gamma^*(x_i^{tr}, y_i^{tr}) l(x_i^{tr}, y_i^{tr}; \theta) = \frac{1}{m}\sum_{i=1}^{m} \beta^*(y_i^{tr}) \cdot l(x_i^{tr}, y_i^{tr}; \theta)$$

- ratio $\beta^*(y)$ can be estimated by min. $\mathcal{D}\left(P_X^{te}, \int P_Y^{tr}\beta(y)P_{X|Y}^{tr}dy\right)$: difficult !

Correcting TarS by Reweighting Target to Match Covariate with KMM

how to find $\beta^*(y) = \frac{P_Y^{te}}{P_Y^{tr}}$?

$P_Y^{new} = \beta(y) P_Y^{tr}$ $\qquad P_Y^{te}$



$\beta(y)$ can be estimated by matching P_X^{new} with P_X^{te} :–)

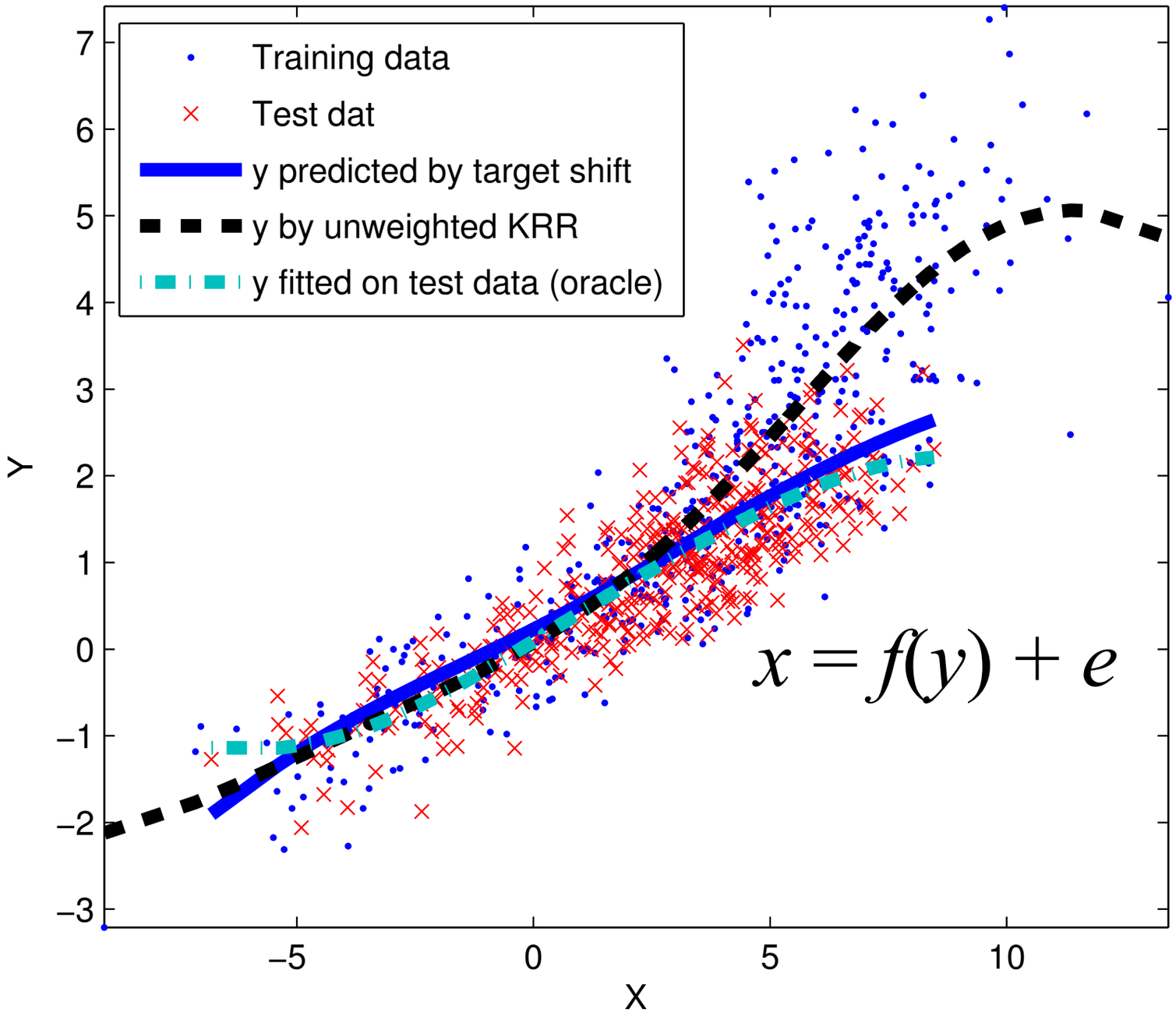
- i.e., minimizing

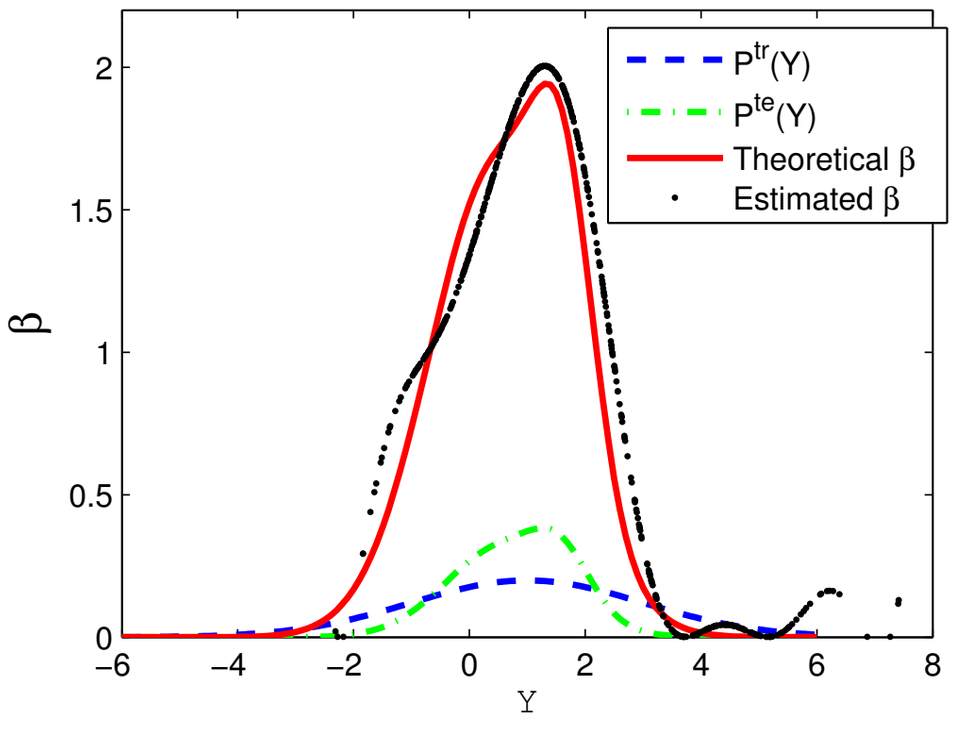
$$\left|\left|\hat{\mu}[P_X^{new}] - \hat{\mu}[P_X^{te}]\right|\right|^2 = \left|\left|\hat{\mathcal{U}}_{X|Y} \cdot \frac{1}{m}\sum_{i=1}^{m} \beta_i \phi(y_i^{tr}) - \frac{1}{n}\sum_{i=1}^{n} \psi(x_i^{te})\right|\right|^2$$

$$= \frac{1}{m^2}\beta^\intercal \underbrace{L(L+\lambda_m I)^{-1} K (L+\lambda_m I)^{-1} L}_{\triangleq A}\beta - \frac{2}{mn}\underbrace{\mathbf{1}_n^\intercal K^c (L+\lambda_m I)^{-1} L}_{\triangleq M}\beta + \text{const}$$

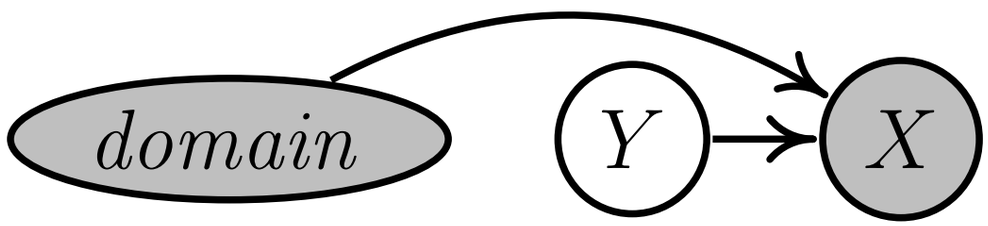
- QP problem: unique solution to β!
- reparameterization such that β is a function of & smooth in y: still a QP problem

Correction for TarS: An illustration





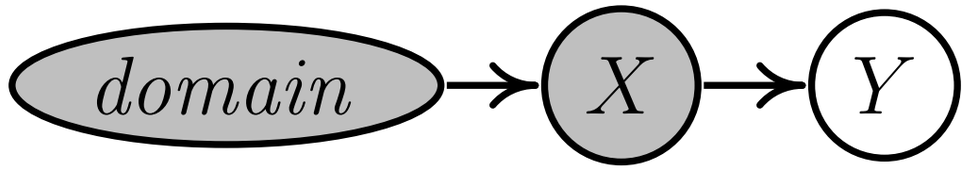
Conditional shift



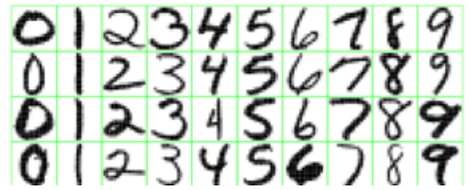
- If $P^{te}_{X|Y} \neq P^{tr}_{X|Y}$, possible to determine $P^{te}_{Y|X}$?
- In general, not possible: marginal P^{te}_{X} do not contain enough information to determine $P^{te}_{X|Y}$ (or $P^{te}_{Y|X}$)
- Change in $P_{X|Y}$ must be constrained

Traditional Methods Assume How Distribution Changes...

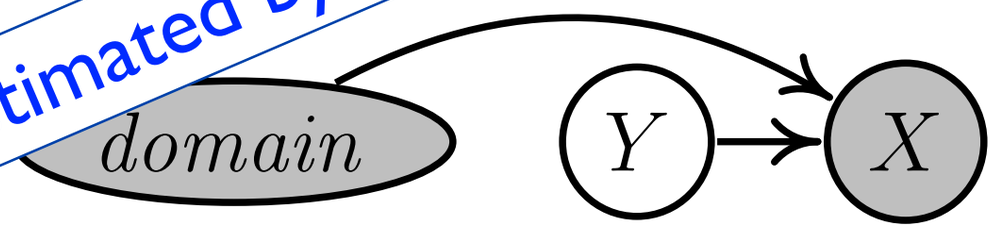
- Covariate shift



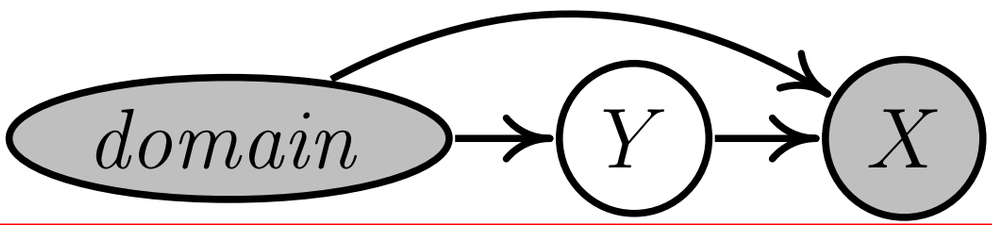
- Target shift



- Conditional shift



- Generali get shift



involved parameters estimated by matching P_X

How to discover and leverage the changeability of the distribution, especially in complex situations?

(Shimodaira 2000; Sugiyama et al. 2008; Huang et al. 2007, Zhang et al., 2013; Zhang et al., 2015; Gong et al., 2016; Stojanov et al., 2018…)

Outline

- Domain adaptation / transfer learning: What and Why?
- Traditional approaches to domain adaptation
- Adaptive methods for domain adaptation
  - Related problems: image translation, multi-domain generation
- Future of domain adaptation

A General (😀 or 😟) Approach

Domain Generalization by Marginal Transfer Learning

Gilles Blanchard BLANCHARD@UNIVERSITE-PARIS-SACLAY.FR
Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay

Aniket Anand Deshmukh ANIKETDE@UMICH.EDU
Microsoft AI & Research

Ürun Dogan URUNDOGAN@GMAIL.COM
Microsoft AI & Research

Gyemin Lee GYEMIN@SEOULTECH.AC.KR
Dept. Electronic and IT Media Engineering
Seoul National University of Science and Technology

Clayton Scott CLAYSCOT@UMICH.EDU
Electrical and Computer Engineering, Statistics
University of Michigan

Abstract

In the problem of domain generalization (DG), there are labeled training data sets from several related prediction problems, and the goal is to make accurate predictions on future unlabeled data sets that are not known to the learner. This problem arises in several applications where data distributions fluctuate because of environmental, technical, or other sources of variation. We introduce a formal framework for DG, and argue that it can be viewed as a kind of supervised learning problem by augmenting the original feature space with the marginal distribution of feature vectors. While our framework has several connections to conventional analysis of supervised learning algorithms, several unique aspects of DG require new methods of analysis.

This work lays the learning theoretic foundations of domain generalization, building on our earlier conference paper where the problem of DG was introduced (Blanchard et al., 2011). We present two formal models of data generation, corresponding notions of risk, and distribution-free generalization error analysis. By focusing our attention on kernel methods, we also provide more quantitative results and a universally consistent algorithm. An

A General Approach: Method

Domain Generalization by Marginal Transfer Learning

Consider a test sample $S^T = (X_j^T, Y_j^T)_{1 \leq j \leq n_T}$, whose labels are not observed. If $\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}_+$ is a loss function for a single prediction, and predictions of a fixed decision function f on the test sample are given by $\widehat{Y}_j^T = f(\widehat{P}_X^T, X_j^T)$, then the empirical average loss incurred on the test sample is

$$\mathcal{L}(S^T, f) := \frac{1}{n_T} \sum_{j=1}^{n_T} \ell(\widehat{Y}_j^T, Y_j^T).$$

Thus, we define the *risk* of a decision function as the average of the above quantity when test samples are drawn according to the same mechanism as the training samples:

$$\mathcal{E}(f) := \mathbb{E}_{S^T \sim P_S} \left[\mathcal{L}(S^T, f) \right] = \mathbb{E}_{S^T \sim P_S} \left[\frac{1}{n_T} \sum_{j=1}^{n_T} \ell(f(\widehat{P}_X^T, X_j^T), Y_j^T) \right].$$

In a similar way, we define the *empirical risk* of a decision function as its average prediction error over the training samples:

$$\widehat{\mathcal{E}}(f, N) := \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(S_i, f) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n_i} \sum_{j=1}^{n_i} \ell(f(\widehat{P}_X^{(i)}, X_{ij}), Y_{ij}). \quad (2)$$

our earlier conference paper where the problem of DG was introduced (Blanchard et al., 2011). We present two formal models of data generation, corresponding notions of risk, and distribution-free generalization error analysis. By focusing our attention on kernel methods, we also provide more quantitative results and a universally consistent algorithm. An

Do You Agree with Their Categorization?

The key difference between DG and DA may be found in the performance measures optimized. In DG, the goal is to design a single predictor $f(P_X, x)$ that can apply to any future task, and risk is assessed with respect to the draw of both a new task, and (under **2SGM**) a new data point from that task. This is in contrast to DA, where the target distribution is typically considered fixed, and the goal is to design a predictor $f(x)$ where, in assessing the risk, the only randomness is in the draw of a new sample from the target task. This difference in performance measures for DG and DA has an interesting consequence for analysis. As we will show, it is possible to attain optimal risk (asymptotically) in DG without making any distributional assumptions like those described above for DA. Of course, this optimal risk is typically larger than the Bayes risk for any particular target domain (see Lemma 9). An interesting question for future research is whether it is possible to close or eliminate this gap (between DG and expected DA risks) by imposing distributional assumptions like those for DA.

Another difference between DA and DG lies in whether the learning algorithm must be rerun for each new test data set. Most unsupervised DA methods employ the unlabeled target data for training and thus, when a new unlabeled target data set is presented, the learning algorithm must be rerun. In contrast, most existing DG methods do not assume access to the unlabeled test data at learning time, and are capable of making predictions as new unlabeled data sets arrive without any further training.

viewed as a kind of supervised learning problem by augmenting the original feature space with the marginal distribution of feature vectors. While our framework has several connections to conventional analysis of supervised learning algorithms, several unique aspects of DG require new methods of analysis.

This work lays the learning theoretic foundations of domain generalization, building on our earlier conference paper where the problem of DG was introduced (Blanchard et al., 2011). We present two formal models of data generation, corresponding notions of risk, and distribution-free generalization error analysis. By focusing our attention on kernel methods, we also provide more quantitative results and a universally consistent algorithm. An

NeurIPS 2020

Domain Adaptation As a Problem of Inference on Graphical Models

Kun Zhang*, Mingming Gong*, Petar Stojanov, Biwei Huang, Qingsong Liu, Clark Glymour

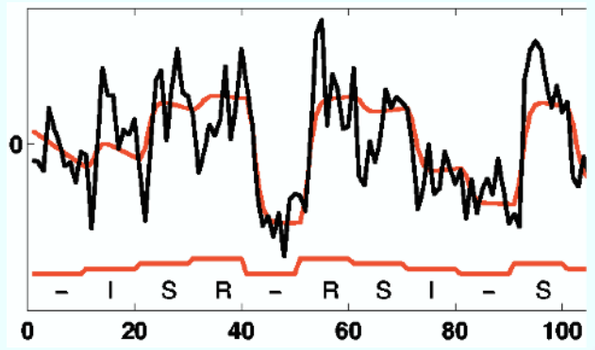
Abstract

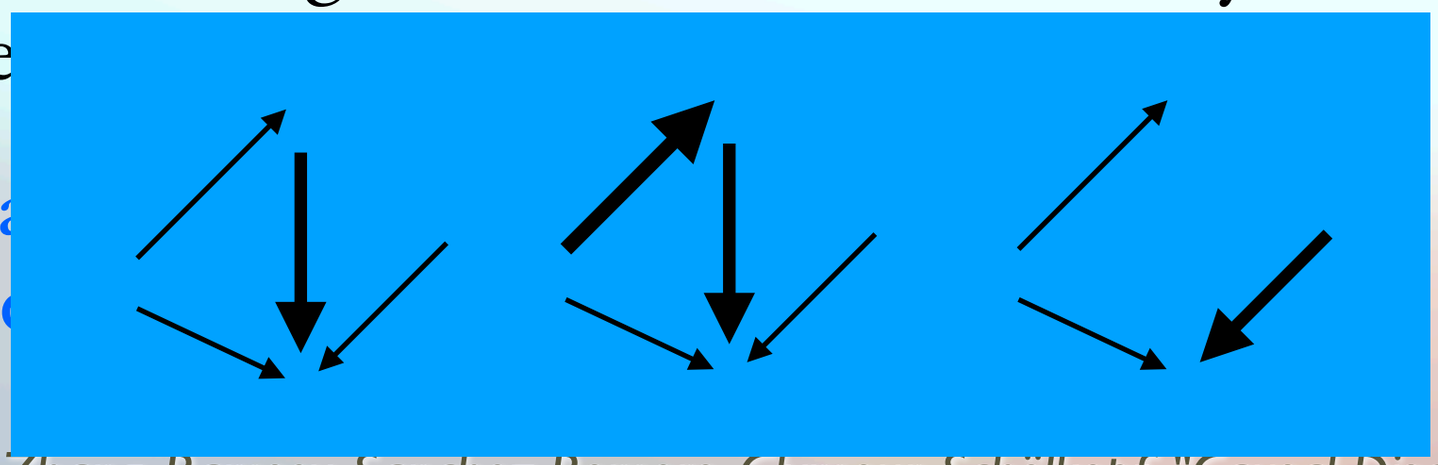
This paper is concerned with data-driven unsupervised domain adaptation, where it is unknown in advance how the joint distribution changes across domains, i.e., what factors or modules of the data distribution remain invariant or change across domains. To develop an automated way of domain adaptation with multiple source domains, we propose to use a graphical model as a compact way to encode the change property of the joint distribution, which can be learned from data, and then view domain adaptation as a problem of Bayesian inference on the graphical models. Such a graphical model distinguishes between constant and varied modules of the distribution and specifies the properties of the changes across domains, which serves as prior knowledge of the changing modules for the purpose of deriving the posterior of the target variable Y in the target domain. This provides an end-to-end framework of domain adaptation, in which additional knowledge about how the joint distribution changes, if available, can be directly incorporated to improve the graphical representation. We discuss how causality-based domain adaptation can be put under this umbrella. Experimental results on both synthetic and real data demonstrate the efficacy of the proposed framework for domain adaptation.

1 Introduction

Nonstationary/Heterogeneous Data and Causal Modeling

- Ubiquity of nonstationary/heterogeneous data
  - Nonstationary time series (brain signals, climate data...)
  - Multiple data sets under different observational or experimental conditions
- Causal modeling & distribution shift heavily couple
  - $P(c$ inde





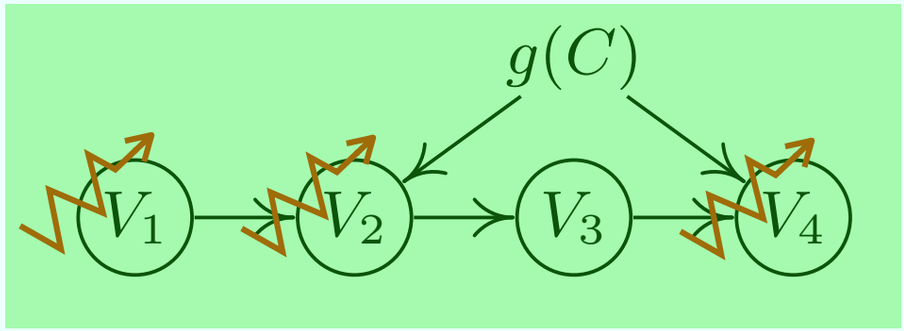
Huang, Zhang, Zhang, Ramsey, Sanchez-Romero, Glymour, Schölkopf, "Causal Discovery from Heterogeneous/ Nonstationary Data," JMLR, 2020

Zhang, Huang, et al., Discovery and visualization of nonstationary causal models, arxiv 2015

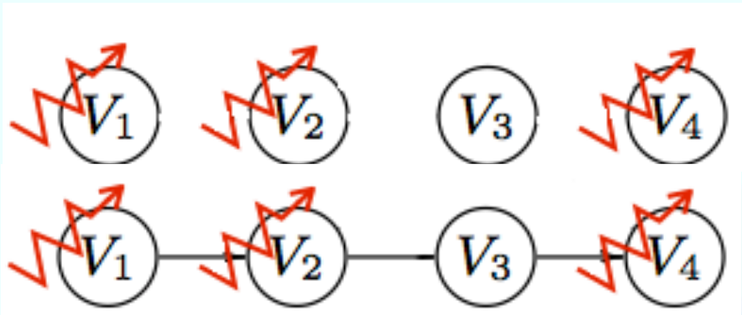
Ghassami, et al., Multi-Domain Causal Structure Learning in Linear Systems, NIPS 2018

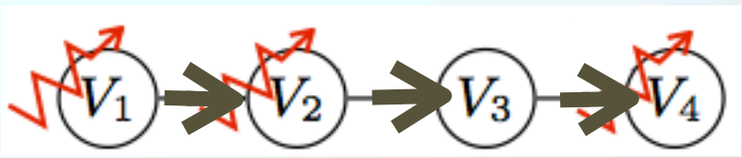
Causal Discovery from Nonstationary/ Heterogeneous Data

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	*Yes*



- Task:
  - Determine changing causal modules & estimate skeleton
  - Causal orientation determination benefits from independent changes in *P*(cause) and *P*(effect | cause), including invariant mechanism/ cause as special cases
  - Visualization of changing modules over time/ across data sets?

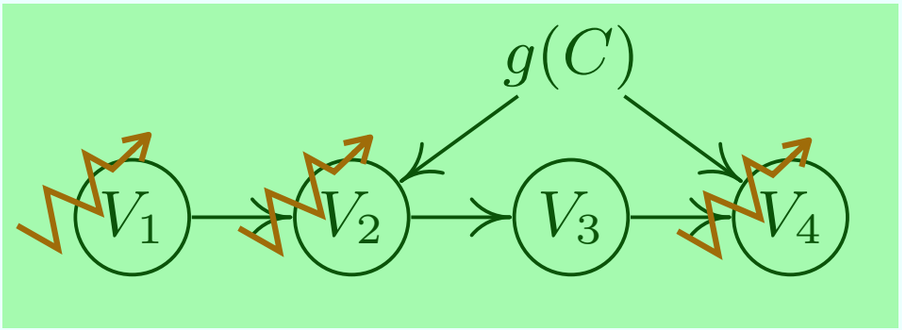




Kernel nonstationary driving force estimation

- *Huang et al., "Causal Discovery from Heterogeneous/Nonstationary Data," JMLR, 2020*
- *Tian, Pearl, "Causal discovery from changes," UAI 2001*
- *Hoover, "The logic of causal inference" Economics and Philosophy, 6:207–234, 1990.*

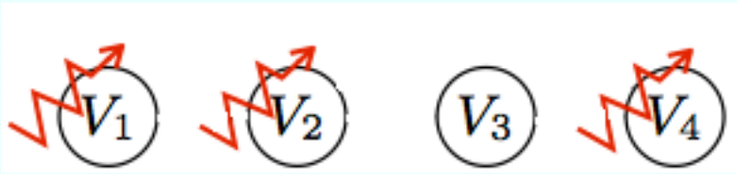
Discovery & Visualization of Changing Causal Modules

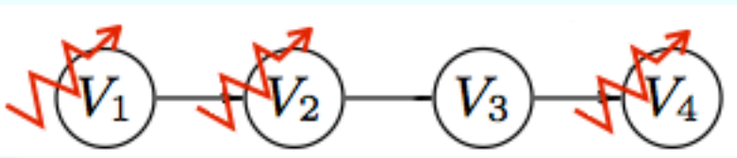


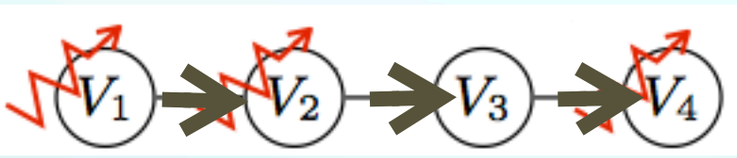
* Questions to answer for causal discovery:

- Identify variables with changing causal modules & recover causal skeleton?
- Identify causal directions by using distribution shifts?
- Visualize the change in causal modules?

With our proposed approach:





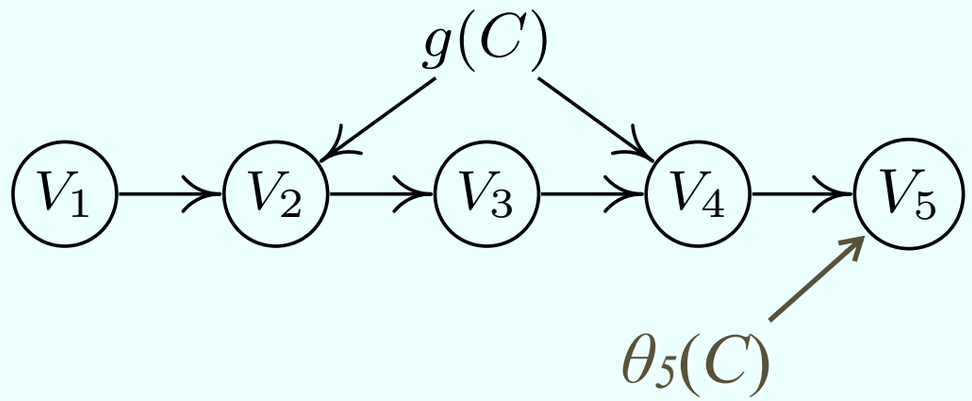


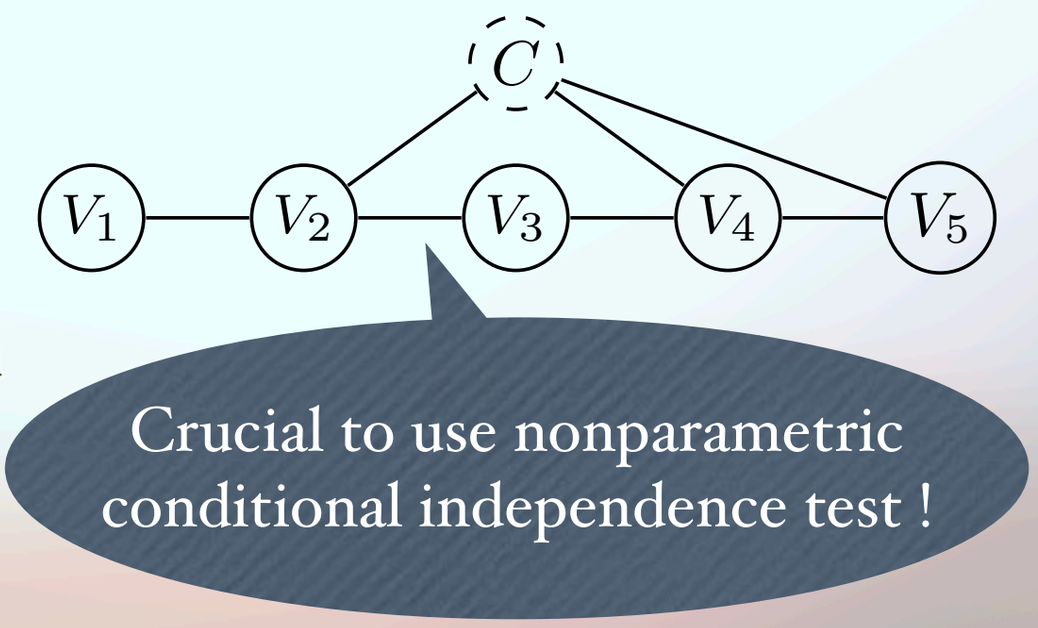
Kernel nonstationarity visualization (KNV)

- Incorporate time/domain index C as a surrogate + apply constraint-based causal discovery methods
- Independent changes in P(cause) and P(effect | cause)
- Find a mapping of $P(V_i \mid PA^i)$ to capture its variability

Finding Causal Skeleton and Changing Modules

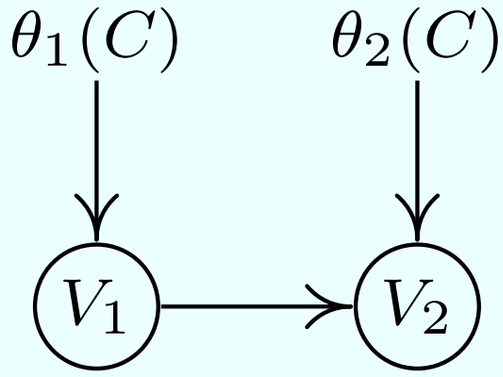
- Incorporate C into the variable set as a surrogate + apply constraint-based causal discovery
  - Detecting changing causal modules
  - "Robust" causal skeleton discovery
- We can find the correct causal skeleton asymptotically correctly, **as if** the confounders were known





Theorem 1. *Given the previous assumptions, for every* $V_i, V_j \in \mathbf{V}$, V_i *and* V_j *are not adjacent in the original causal DAG* G *if and only if they are independent conditional on some subset of* $\{V_k \,|\, k \neq i, k \neq j\} \cup \{C\}$.

Nonstationarity Helps Determine Causal Direction



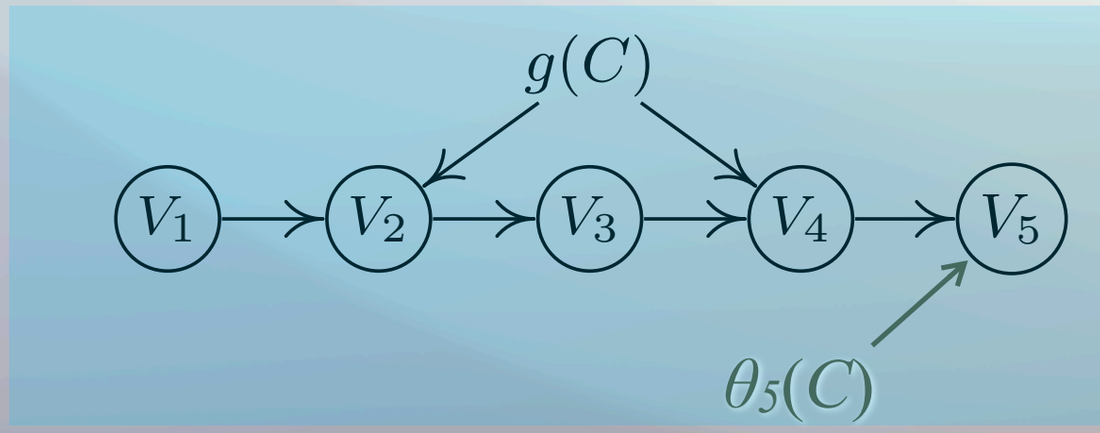
- **Independent changes** in *P*(cause) and *P*(effect | cause): generalization of invariance; generally violated for wrong directions

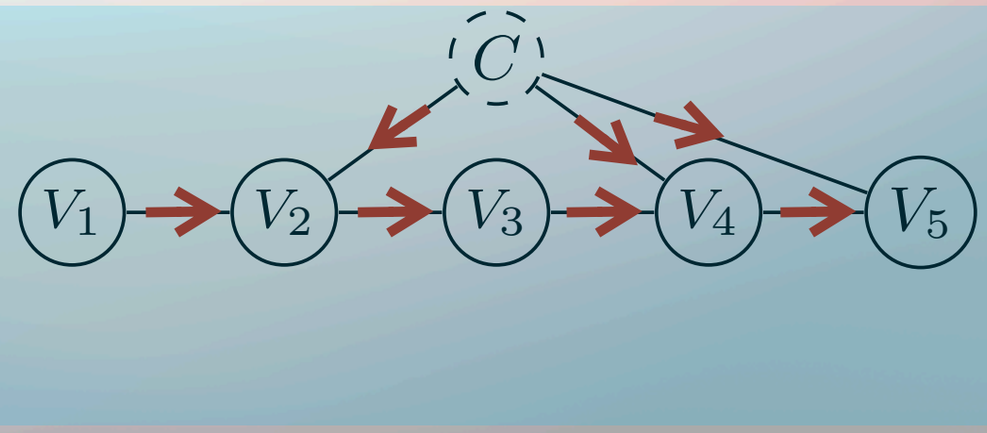
- Special cases: if $C - V_k - V_l$, since $C \to V_k$, we known

  - $C \to V_k \leftarrow V_l$, if $C \perp\!\!\!\perp V_l$ given a variable set **excluding** V_k Invariant cause

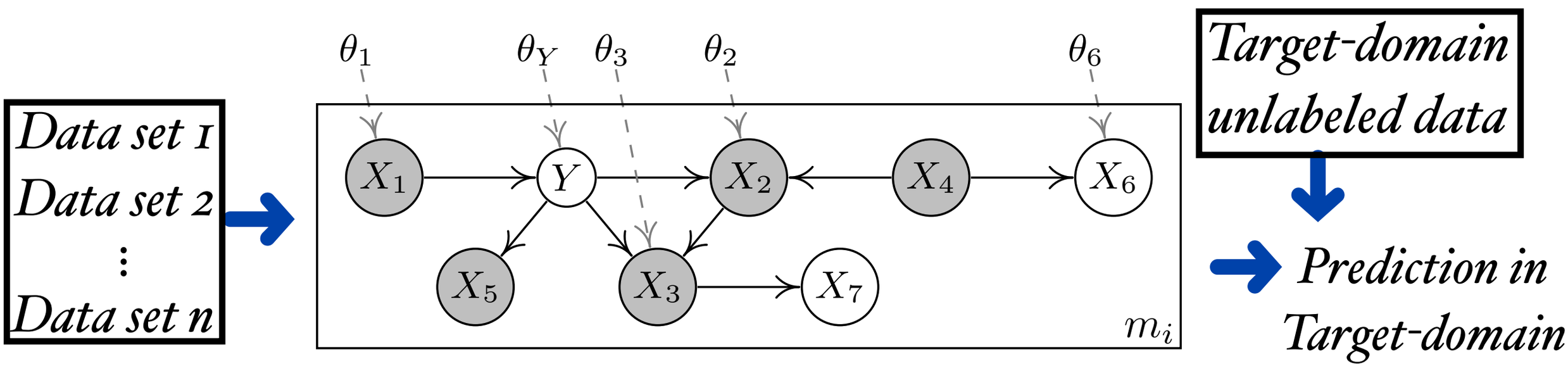
  - $C \to V_k \to V_l$, if $C \perp\!\!\!\perp V_l$ given a variable set **including** V_k Invariant mechanism

Hoover. The logic of causal inference. Economics and Philosophy, 6:207–234, 1990.

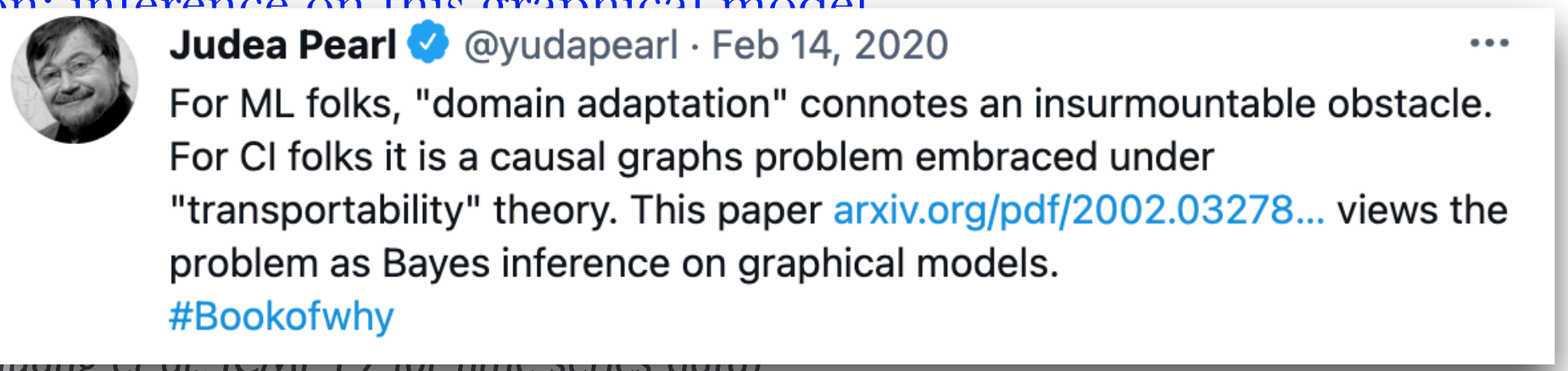




An Approach to Data-Driven Domain Adaptation



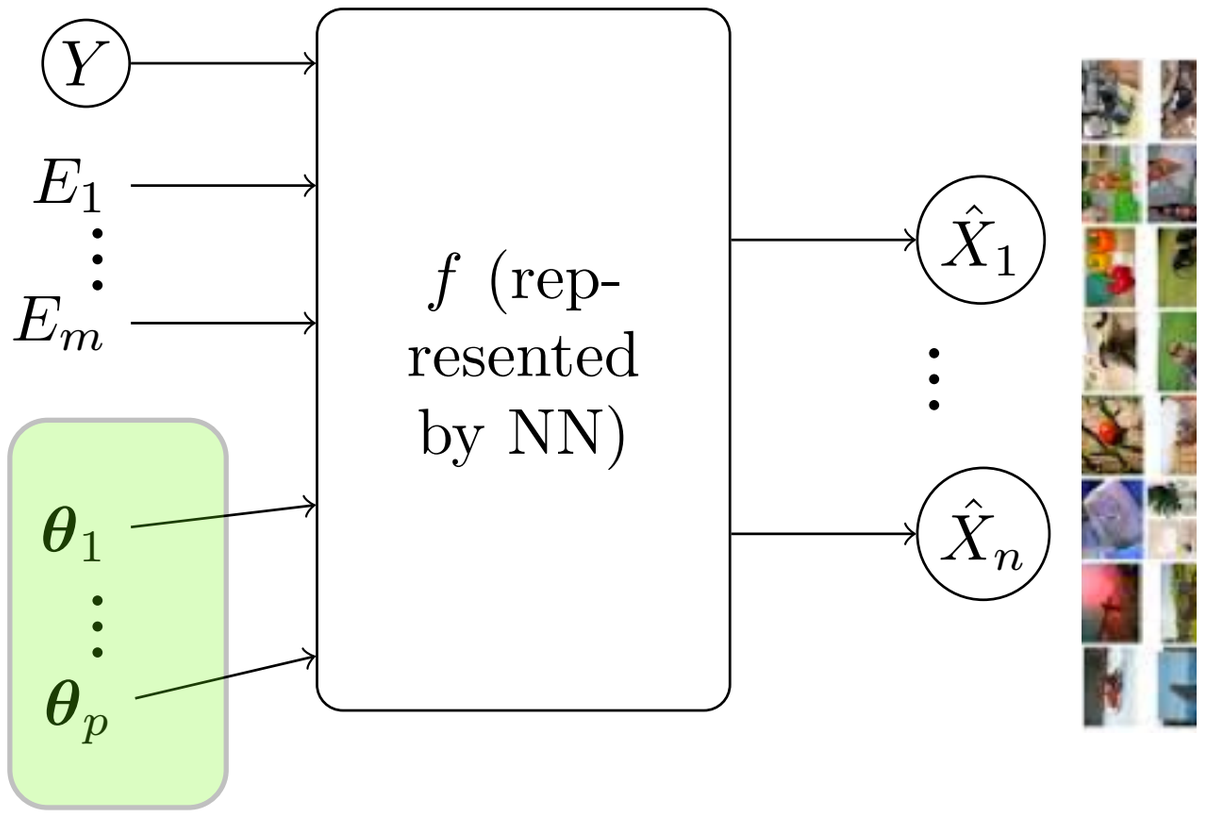
- Only relevant features needed to predict Y
- Augmented graph learned by CD-NOD
  - Independently changing modules θ_i
  - Special case: invariant modules
- Domain adaptation: inference on this graphical model
  - Infer the po
  - Nonparame



Zhang, Gong*, Stojanov, Hu*
Models," NeurIPS 2020. (Huang et al., ICML'19 for time series data)

To Model Changing Conditional Distributions

- Assume Y →X
- Why & how does the distribution change across domains?
- Generative network + (minimal) (latent) parameters $\boldsymbol{\theta}$ to model changes in the causal process
  - Understanding & generating new domains



Results on Simulated & Real Data

Table 1: Accuracy on simulated datasets for the baselines and proposed method. The values presented are averages over 10 replicates for each experiment. Standard deviation is in parentheses.

	DICA	weigh	simple_adapt	comb_classif	LMP	poolSVM	Infer
9 sources	80.04(15.5)	72.1(14.5)	70.0(14.3)	72.34(16.24)	78.90(13.81)	71.8(11.43)	**83.90(9.02)**
4 sources	74.16(13.2)	67.88(13.7)	65.22(16.00)	69.64(15.8)	79.06(13.93)	70.08(12.25)	**85.38(11.31)**
2 sources	86.56(13.63)	75.04(18.8)	69.42(17.87)	74.28(18.2)	84.52(13.72)	83.84(13.7)	**93.10(7.17)**

Table 2: Accuracy on the Wi-Fi & Flow data. Standard deviation is in parentheses.

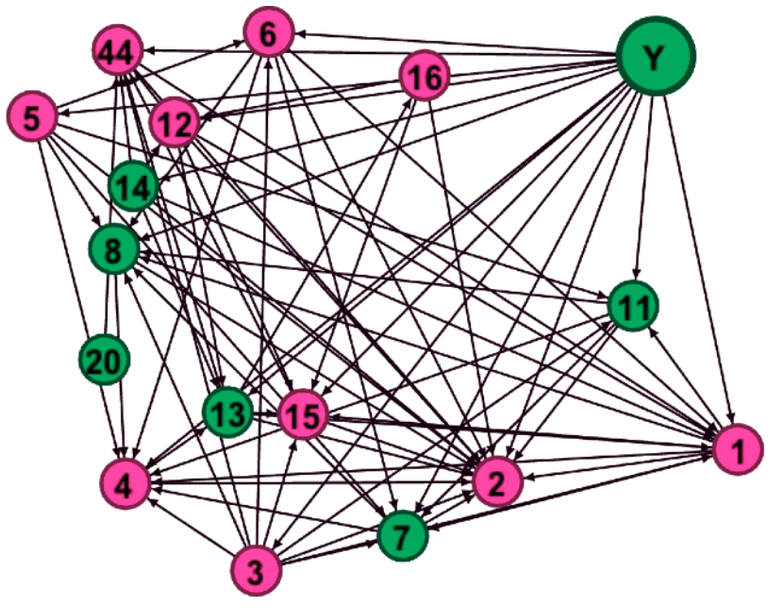
	DICA	weigh	LMP	poolSVM	Soft-Max	poolNN	Infer
t2, t3 → t1	29.32(2.5)	43.71(3.02)	46.80(1.4)	40.25(1.6)	44.86(5.1)	42.88(1.6)	**70.8(2.7)**
t1, t3 → t2	24.5(3.6)	38.19(1.9)	39.11(2.1)	48.70(1.8)	44.95(4.4)	47.41(2.1)	**84.5(2.9)**
t1, t2 → t3	21.7(3.9)	36.03(1.85)	39.28(2.05)	40.46(1.4)	43.63(4.1)	41.00(1.8)	**83.0(7.3)**
Flow 3 sources	79.2(11.0)	84.2(9.3)	91.6 (8.4)	92.1(7.5)	89.0(9.7)	95.7(5.2)	**96.8(3.5)**
Flow 5 sources	83.1(12.0)	92.9(7.0)	92.3 (6.4)	94.7(6.1)	89.7(8.0)	96.0(5.1)	**97.1(3.5)**

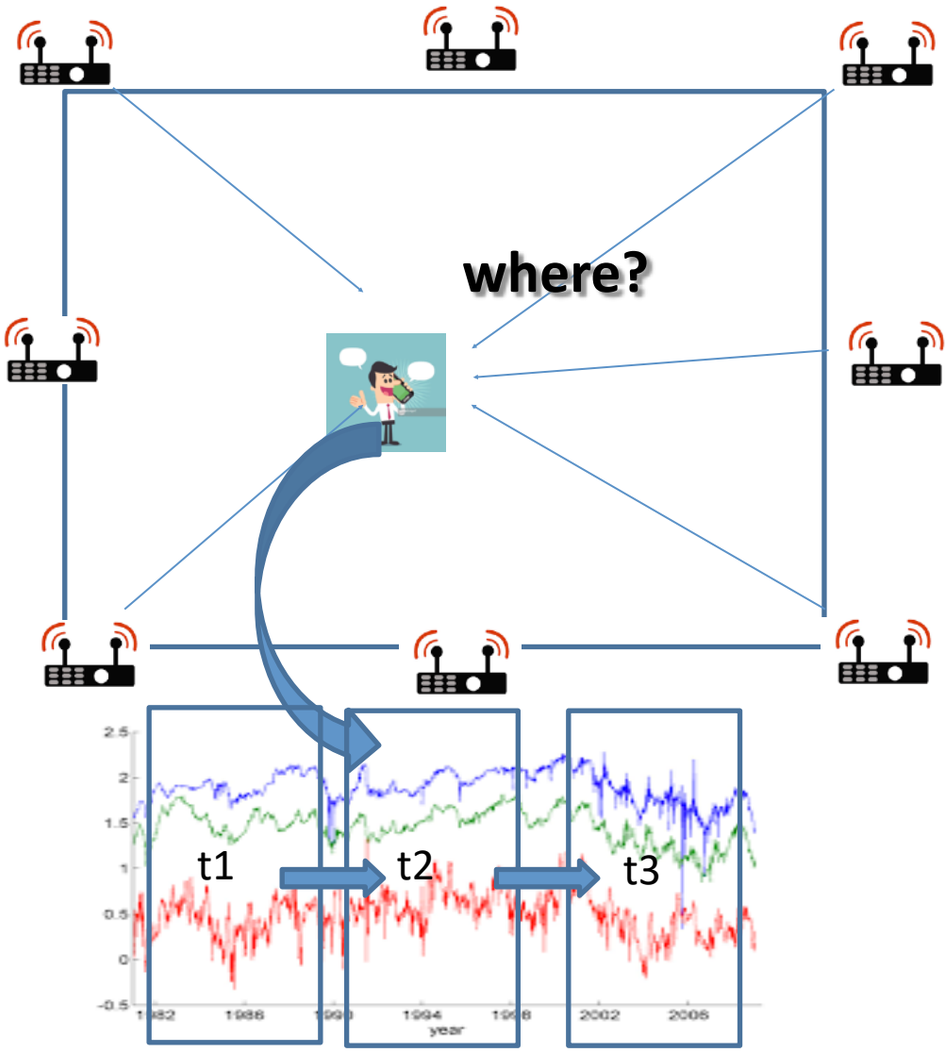
Table 3: Accuracy on the digits data. T: MNIST; M: MNIST-M; S: SVHN; D: SynthDigits.

	weigh	poolNN	poolDANN	Hard-Max	Soft-Max	poolNN_Ours	Infer
$S + M + D/T$	75.5	93.8	92.5	97.6	**97.9**	94.9	96.64
$T + S + D/M$	56.3	56.1	65.1	66.3	68.7	59.6	**89.89**
$M + T + D/S$	60.4	77.1	77.6	80.2	81.6	67.8	**89.34**

Transfer Learning on WIFI Data

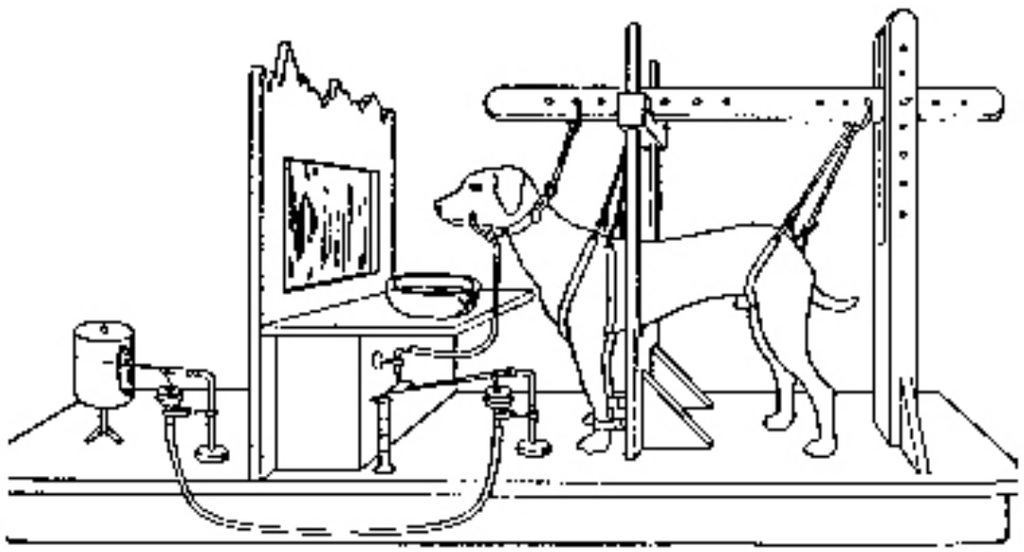
- Input X: WiFi signal strengths from multiple routers; Y : location
- Transfer from two time periods to another (e..g, t1, t2 → t3)





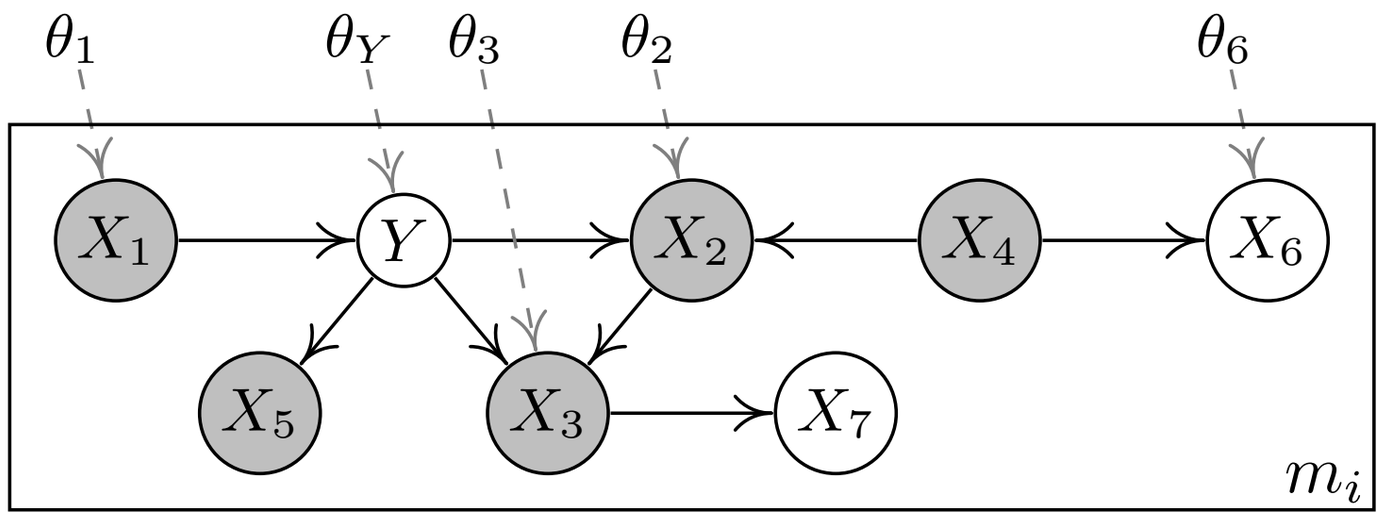
Causality & Transferability

- Causality helps
- But hard to find (rather **strong** assumptions)
- And perhaps not necessary to achieve transferability
  - Think about classical conditioning



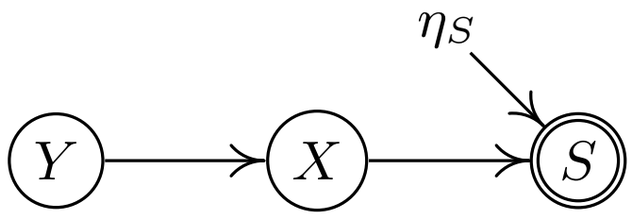
- *"If a particular stimulus in the dog's surroundings was present when the dog was given food then that stimulus could become associated with food and cause salivation on its own."*

Augmented Graph

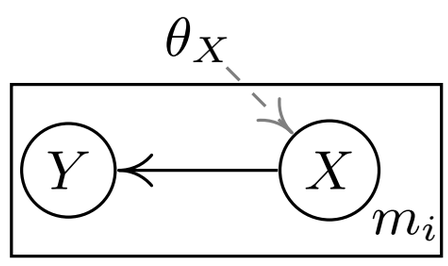


- To represent independent changes in the joint distribution
  - Causal graph vs. augmented DAG

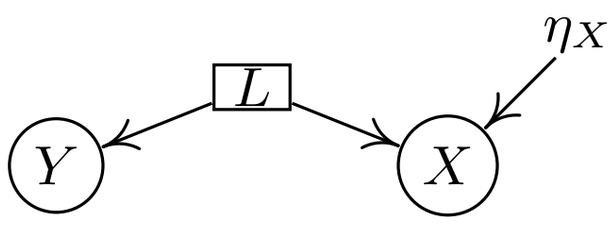
because p(Y|X) is invariant across domains



(a) The underlying data generating process of Example 1. Y generates (causes) X, and S denotes the selection variable (a data point is included if and only if $S = 1$).

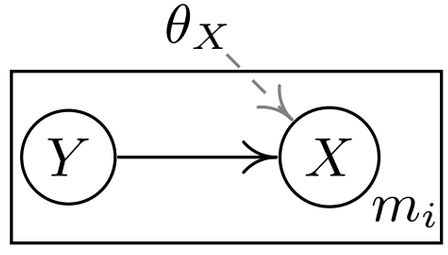


(b) The augmented DAG representation for Example 1 to explain how the data distribution changes across domains.



(c) The generating process of Example 2. L is a confounder; the mechanism of X changes across domains, as indicated by η_X.

because p(Y) is invariant across domains



(d) The augmented DAG representation for Example 2 to explain how the data distribution changes across domains.

What Changes Lead to Distribution Shift?

- Distributions of measured features or their relationships in between
- Due to changes in hidden variables (illumination conditions, temperature…)?

Partial Identifiability for Domain Adaptation

Lingjing Kong [1] **Shaoan Xie** [1] **Weiran Yao** [1] **Yujia Zheng** [1] **Guangyi Chen** [2 1] **Petar Stojanov** [3]
Victor Akinwande [1] **Kun Zhang** [2 1]

Abstract

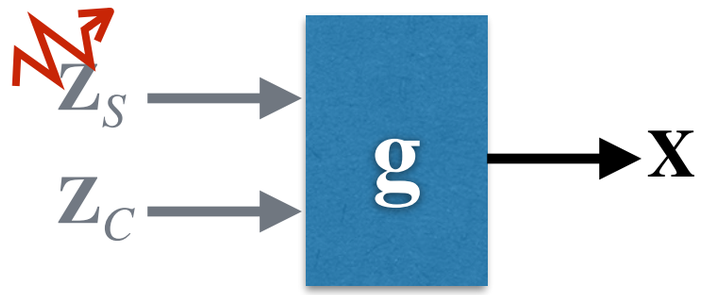
Unsupervised domain adaptation is critical to many real-world applications where label information is unavailable in the target domain. In general, without further assumptions, the joint distribution of the features and the label is not identifiable in the target domain. To address this issue, we rely on a property of minimal changes of causal mechanisms across domains to minimize unnecessary influences of domain shift. To encode this property, we first formulate the data generating process using a latent variable model with two partitioned latent subspaces: invariant components whose distributions stay the same across domains, and sparse changing components that vary across domains. We further constrain the domain shift to have a restrictive influence on the changing components. Under mild conditions, we show that the latent variables are partially identifiable, from

domain indices $\mathbf{u}$, the training (source domain) data follows multiple joint distributions $p_{\mathbf{x},\mathbf{y}|\mathbf{u}_1}$, $p_{\mathbf{x},\mathbf{y}|\mathbf{u}_2}$, ..., $p_{\mathbf{x},\mathbf{y}|\mathbf{u}_M}$,[1] and the test (target domain) data follows the joint distribution $p_{\mathbf{x},\mathbf{y}|\mathbf{u}^\mathcal{T}}$, where $p_{\mathbf{x},\mathbf{y}|\mathbf{u}}$ may vary across $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_M$. During training, for each i-th source domain, we are given labeled observations $(\mathbf{x}_k^{(i)}, \mathbf{y}_k^{(i)})_{k=1}^{m_i}$ from $p_{\mathbf{x},\mathbf{y}|\mathbf{u}_i}$, and target domain unlabeled instances $(\mathbf{x}_k^\mathcal{T})_{k=1}^{m_\mathcal{T}}$ from $p_{\mathbf{x},\mathbf{y}|\mathbf{u}^\mathcal{T}}$. The main goal of domain adaptation is to make use of the available observed information, to construct a predictor that will have optimal performance in the target domain.

It is apparent that without further assumptions, this objective is ill-posed. Namely, since the only available observations in the target domain are from the marginal distribution $p_{\mathbf{x}|\mathbf{u}^\mathcal{T}}$, the data may correspond to infinitely many joint distributions $p_{\mathbf{x},\mathbf{y}|\mathbf{u}^\mathcal{T}}$. This mandates making additional assumptions on the relationship between the source and the target domain distributions, with the hope to be able to reconstruct (identify) the joint distribution in the target domain $p_{\mathbf{x},\mathbf{y}|\mathbf{u}^\mathcal{T}}$. Typically, these assumptions entail some measure of sim-

Finding Changing Hidden Variables for Transfer Learning

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes



- Underlying components $\mathbf{Z}_S$ may change across domains
- Changing components $\mathbf{Z}_S$ are identifiable; invariant part $\mathbf{Z}_C$ are identifiable up to its subspace
- Using invariant part $\mathbf{Z}_C$ and transformed changing part $\tilde{\mathbf{Z}}_S$ for prediction

Models	→ Art	→ Clipart	→ Product	→ Realworld	Avg
Source Only (He et al., 2016)	64.58±0.68	52.32±0.63	77.63±0.23	80.70±0.81	68.81
DANN (Ganin et al., 2016)	64.26±0.59	58.01±1.55	76.44±0.47	78.80±0.49	69.38
DANN+BSP (Chen et al., 2019)	66.10±0.27	61.03±0.39	78.13±0.31	79.92±0.13	71.29
DAN (Long et al., 2015)	68.28±0.45	57.92±0.65	78.45±0.05	81.93±0.35	71.64
MCD (Saito et al., 2018)	67.84±0.38	59.91±0.55	79.21±0.61	80.93±0.18	71.97
M3SDA (Peng et al., 2019)	66.22±0.52	58.55±0.62	79.45±0.52	81.35±0.19	71.39
DCTN (Xu et al., 2018)	66.92±0.60	61.82±0.46	79.20±0.58	77.78±0.59	71.43
MIAN (Park & Lee, 2021)	69.39±0.50	63.05±0.61	79.62±0.16	80.44±0.24	73.12
MIAN-γ (Park & Lee, 2021)	69.88±0.35	**64.20±0.68**	80.87±0.37	81.49±0.24	74.11
iMSDA (Ours)	**75.77±0.21**	60.83±0.73	**84.13±0.09**	**84.83±0.12**	**76.39**

Table 2. Classification results on Office-Home. Backbone: Resnet-50. Baseline results are taken from (Park & Lee, 2021).

- *Kong, Xie, Yao, Zheng, Chen, Stojanov, Akinwande, Zhang, Partial disentanglement for domain adaptation, ICML 2022*

Implementation of Partial Disentanglement for Domain Adaptation

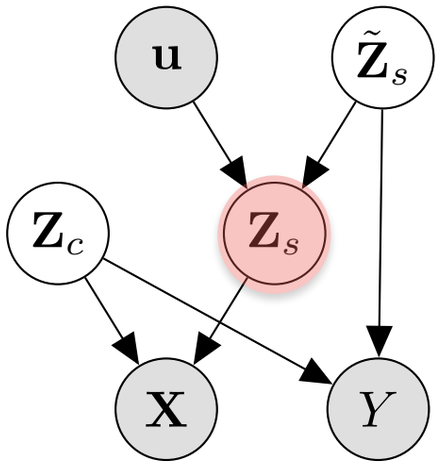
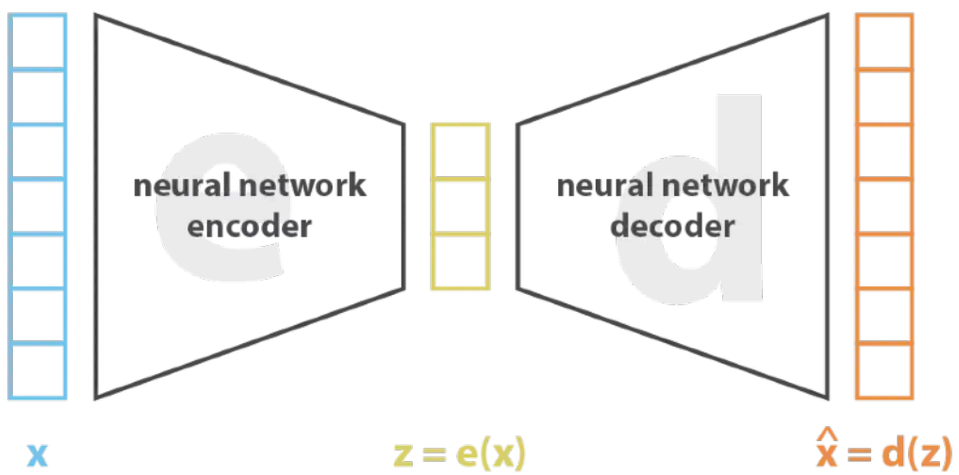


Figure 1. The generating process: The gray shade of nodes indicates that the variable is observable.



loss = $\| x - \hat{x} \|^2 = \| x - d(z) \|^2 = \| x - d(e(x)) \|^2$

Autoencoder

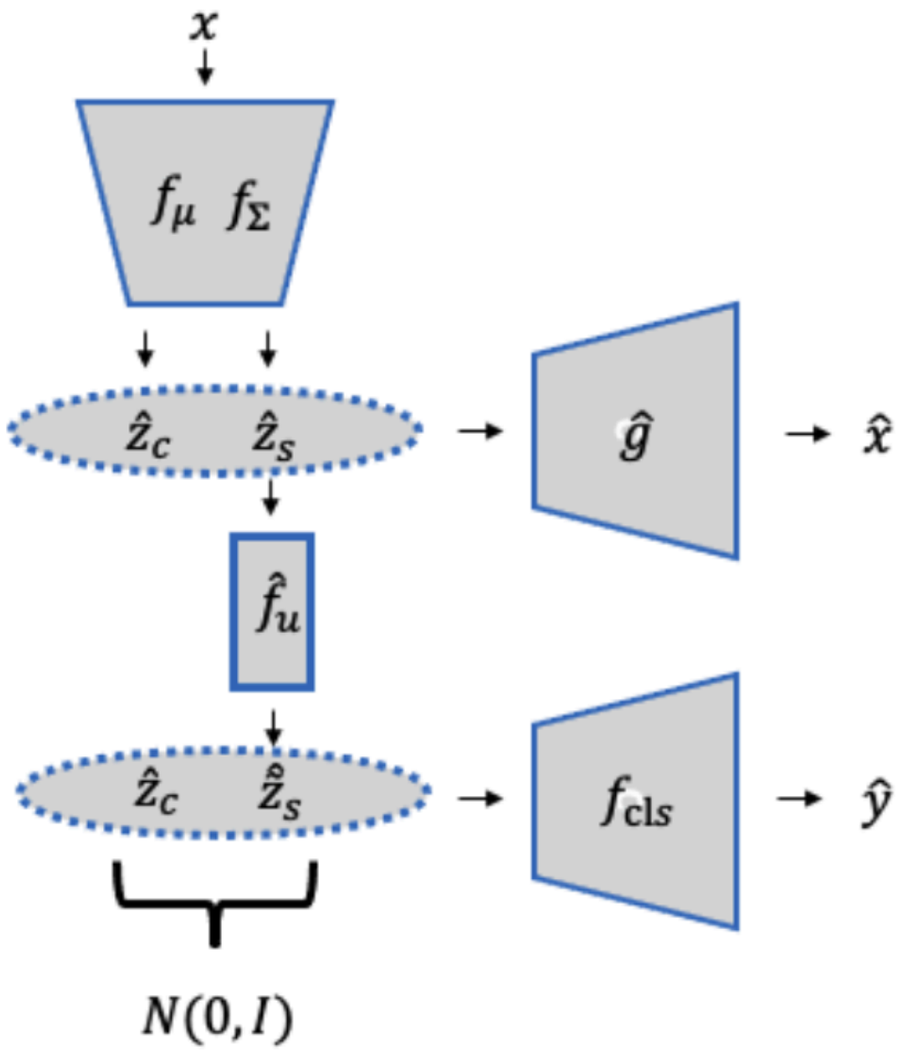


Figure 2. Diagram of our proposed method, **iMSDA**. We first apply the VAE encoder (f_μ, f_Σ) to encode $\mathbf{x}$ into $(\hat{\mathbf{z}}_c, \hat{\mathbf{z}}_s)$, which is further fed into the decoder $\hat{g}$ for reconstruction. In parallel, the changing part $\hat{\mathbf{z}}_s$ is passed through the flow model $f_\mathbf{u}$ to recover the high-level invariant variable $\hat{\tilde{\mathbf{z}}}_s$. We use $(\hat{\mathbf{z}}_c, \hat{\tilde{\mathbf{z}}}_s)$ for classification with the classifier f_{cls} and for matching $\mathcal{N}(\mathbf{0}, \mathbf{I})$ with a KL loss.

Application to RL

ADARL: WHAT, WHERE, AND HOW TO ADAPT IN TRANSFER REINFORCEMENT LEARNING

Biwei Huang
Carnegie Mellon University
biweih@andrew.cmu.edu

Fan Feng
City University of Hong Kong
ffeng1017@gmail.com

Chaochao Lu
University of Cambridge & Max Planck Institute for Intelligent Systems
cl641@cam.ac.uk

Sara Magliacane
University of Amsterdam & MIT-IBM Watson AI Lab
sara.magliacane@gmail.com

Kun Zhang
Carnegie Mellon University &
Mohamed bin Zayed University of Artificial Intelligence
kunz1@cmu.edu

ABSTRACT

One practical challenge in reinforcement learning (RL) is how to make quick adaptations when faced with new environments. In this paper, we propose a principled framework for adaptive RL, called *AdaRL*, that adapts reliably and efficiently to changes across domains with a few samples from the target domain, even in partially observable environments. Specifically, we leverage a parsimonious graphical representation that characterizes structural relationships over variables in the RL system. Such graphical representations provide a compact way to encode what and where the changes across domains are, and furthermore inform us with a minimal set of changes that one has to consider for the purpose of policy adaptation. We show that by explicitly leveraging this compact representation to encode changes, we can efficiently adapt the policy to the target domain, in which only a few samples are needed and further policy optimization is avoided. We

Adaptive RL: Procedure

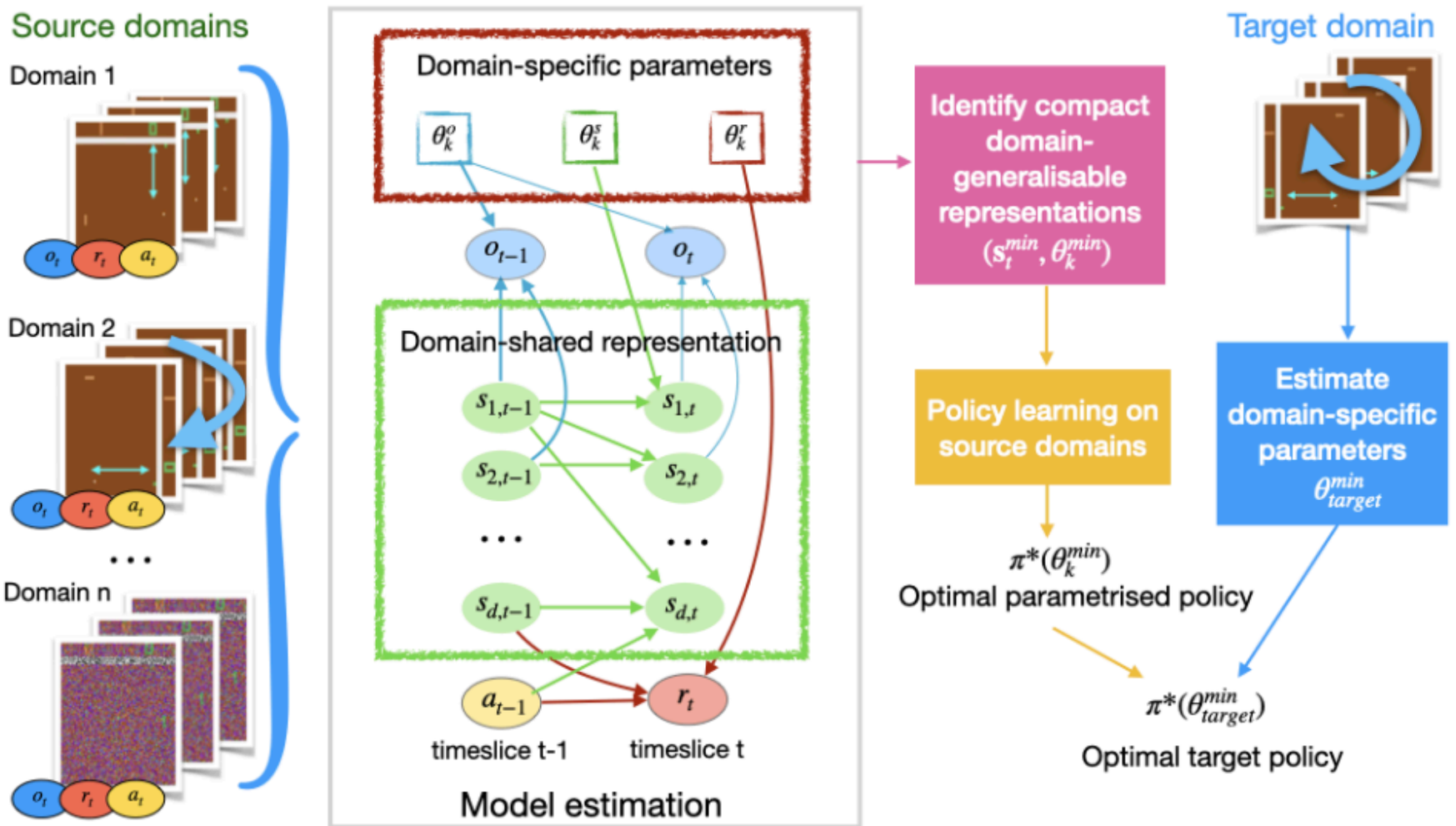
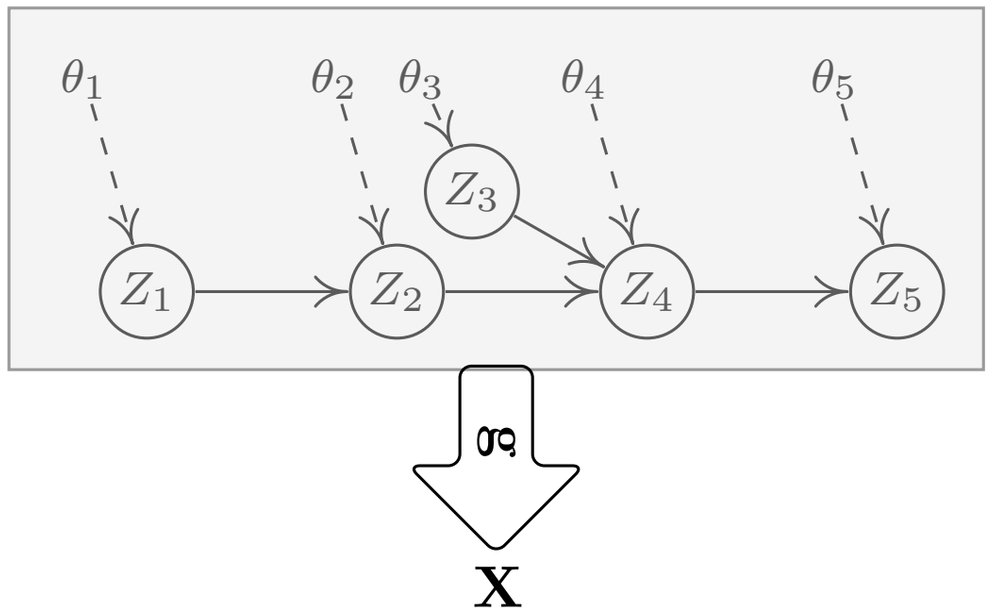


Figure 1: The overall AdaRL framework. We learn a Dynamic Bayesian Network (DBN) over the observations, latent states, reward, actions and domain-specific change factors that is shared across the domains. We then characterize a minimal set of representations that suffice for policy transfer, so that we can quickly adapt the optimal source policy with only a few samples from the target domain.

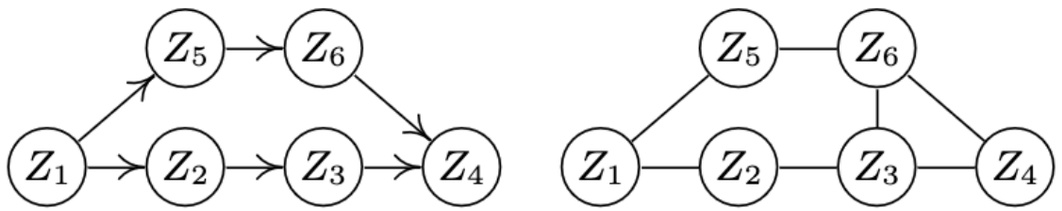
Remember? Causal Representation Learning from Multiple Distributions: A *General* Setting

i.i.d. data?	Parametric constraints?	Latent confounders?
Yes	No	No
No	Yes	Yes



- Goal: Uncovering hidden variables Z_i with changing causal relations from $\mathbf{X}$ in nonparametric settings
- What is identifia
  - Markov netw
  - Each estimated variable $\tilde{Z}_i$ is a function of Z_i and it **intimate neighbors**
- In this example, each Z_i ($i \neq 4$) can be recovered up to component-wise transformation

We exploit the changes in causal mechanisms along with domain!



(a) $\mathcal{G}_Z$, the DAG over true latent variables Z_i.

(b) The corresponding Markov network $\mathcal{M}_Z$.

- *Zhang, Xie, Ng, Zheng, "Causal Representation Learning from Multiple Distributions: A General Setting," ICML 2024*

A General Representation-Based Approach to Multi-Source Domain Adaptation

Anonymous Authors[1]

Abstract

A central problem in unsupervised domain adaptation is determining what to transfer from labeled source domains to an unlabeled target domain. To handle high-dimensional observations (e.g., images), a line of approaches use deep learning to learn latent representations of the observations, which facilitate knowledge transfer in the latent space. However, existing approaches often rely on restrictive assumptions to establish identifiability of the joint distribution in the target domain, such as independent latent variables or invariant label distributions, limiting their real-world applicability. In this work, we propose a general domain adaptation framework that learns compact latent representations to capture distribution shifts relative to the prediction task and address the fundamental question of what representations should be learned and transferred. Notably, we first demonstrate that learning representations based on all the predictive information, i.e., the label's Markov blanket in terms of the learned representations

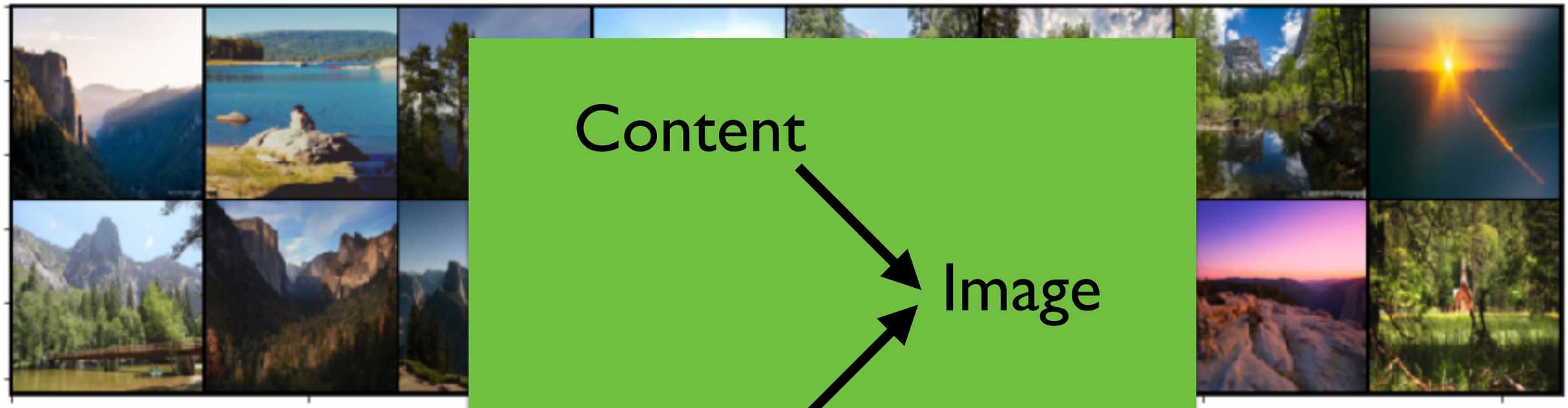
source domain adaptation (MSDA) setup, each source domain $u \in \{1, \dots, M\}$ provides access to a labeled dataset $(\mathbf{x}^{(u)}, \mathbf{y}^{(u)}) = \{(\mathbf{x}_k^{(u)}, y_k^{(u)})\}_{k=1}^{m_u}$, where m_u represents the number of samples in domain u. Here, the i-th dimension of the feature vector X is denoted as X_i, and $x_{ik}^{(u)}$ corresponds to the value of the i-th feature for the k-th sample in domain u. The goal is to train a classifier that generalizes to an unlabeled target domain, where only the feature vectors $\mathbf{x}^{\tau} = \{\mathbf{x}_k^{\tau}\}_{k=1}^{m}$ are available.

Determining the joint distribution $P_{X,Y}^{\tau}$ in the target domain based solely on the marginal distribution P_X^{τ} is a fundamentally underdetermined problem. In the absence of additional assumptions, there are infinitely many possible joint distributions $P_{X,Y}^{\tau}$ that can align with the observed marginal distribution. Therefore, assumptions that connect the source and target domain distributions are essential for identifying the target joint distribution. Common approaches impose constraints to ensure a degree of similarity across these distributions. A widely adopted assumption is covariate shift (Pan & Yang, 2009), which asserts that the conditional distribution $P_{Y|X}$ remains consistent across domains while

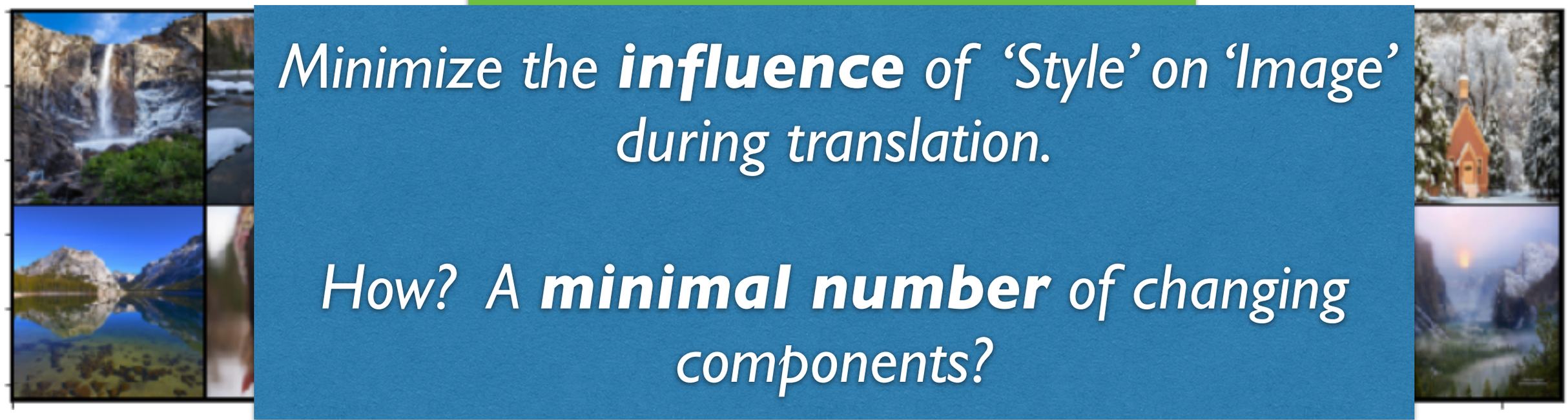
Outline

- Domain adaptation / transfer learning: What and Why?
- Traditional approaches to domain adaptation
- Adaptive methods for domain adaptation
  - Related problems: image translation, multi-domain generation
- Future of domain adaptation

Unsupervised Image-to-Image Translation



Im ain.



*Minimize the **influence** of 'Style' on 'Image' during translation.*

*How? A **minimal number** of changing components?*

Images from the winter season domain.

MULTI-DOMAIN IMAGE GENERATION AND TRANSLATION WITH IDENTIFIABILITY GUARANTEES

Shaoan Xie[1], **Lingjing Kong**[1], **Mingming Gong**[3,2], and **Kun Zhang**[1,2]

[1] Carnegie Mellon University
[2]Mohamed bin Zayed University of Artificial Intelligence
[3]The University of Melbourne
shaoan@cmu.edu, lingjingkong@cmu.edu,
mingming.gong@unimelb.edu.au, kunz1@cmu.edu

ABSTRACT

Multi-domain image generation and unpaired image-to-to-image translation are two important and related computer vision problems. The common technique for the two tasks is the learning of a joint distribution from multiple marginal distributions. However, it is well known that there can be infinitely many joint distributions that can derive the same marginals. Hence, it is necessary to formulate suitable constraints to address this highly ill-posed problem. Inspired by the recent advances in nonlinear Independent Component Analysis (ICA) theory, we propose a new method to learn the joint distribution from the marginals by enforcing a specific type of minimal change across domains. We report one of the first results connecting multi-domain generative models to identifiability and shows

Sample Images Generated by Generative Adversarial Networks (GANs)

Images generated by a [GAN created by NVIDIA](#).

GANs

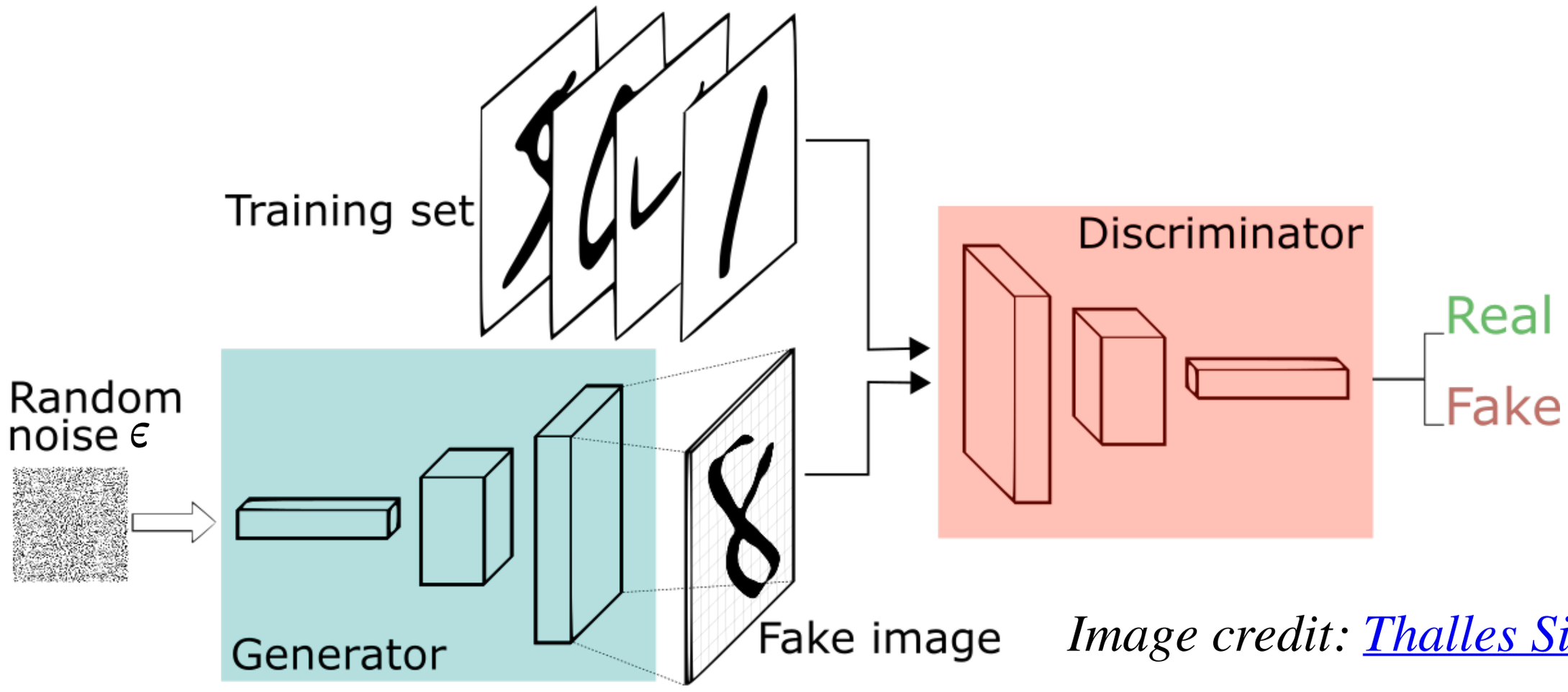
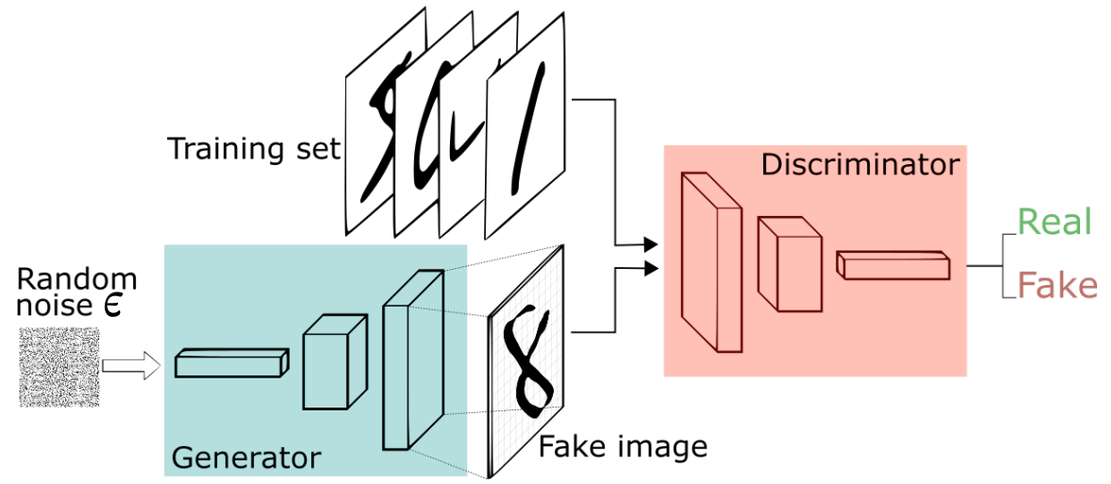


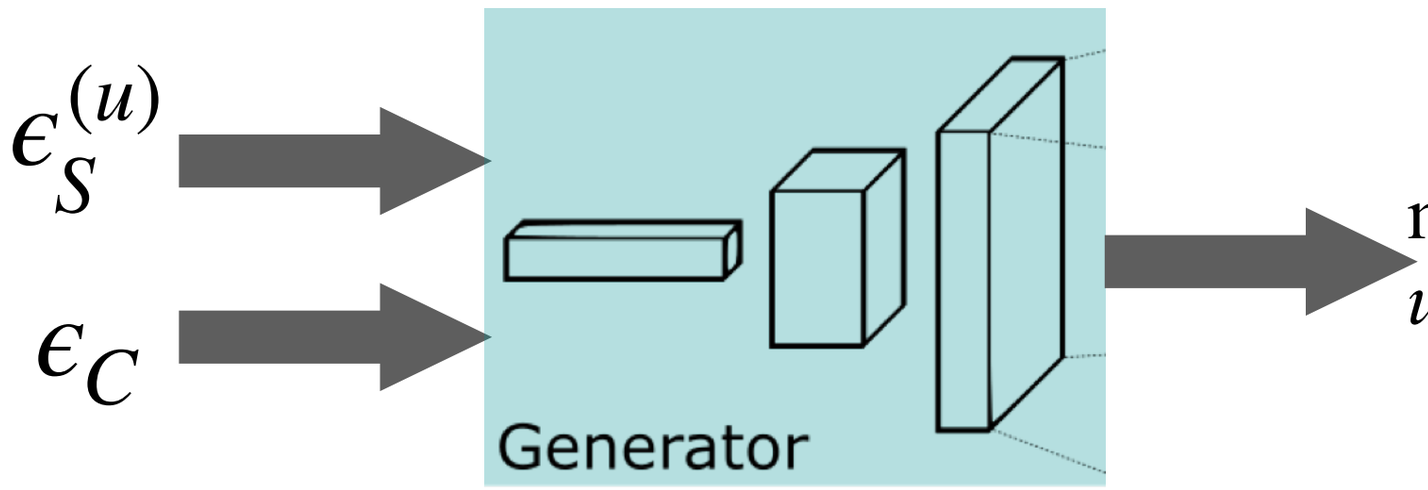
Image credit: Thalles Silva

Minimax game which G wants to minimize V while D wants to maximize it:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

GAN-Based Implementations





- Match the data distribution across domains, while the dimensionality of $\epsilon_S^{(u)}$ is as small as possible (minimal changes across domains **controlled by λ;** no penalty when λ=0)

- Correspondence relations among domains are identifiable

Multi-domain Image Generation & Translation with Identifiability Guarantees

- Idea: Matching the distributions across domains with a minimal number of changing components
- Correspondence info (joint distribution) identifiable under mild assumptions
- Example: Generating female & males images with the same "content"

Ours (λ=0.1)

StyleGAN2-ADA

TGAN

- *Xie, Kong, Gong, Zhang, "Multi-domain image generation and translation with identifiability guarantees", ICLR 2023*

Outline

- Domain adaptation / transfer learning: What and Why?
- Traditional approaches to domain adaptation
- Adaptive methods for domain adaptation
    - Related problems: image translation, multi-domain generation
- Future of domain adaptation

Transfer learning with large language models (LLMs)

- Involves leveraging pre-trained LLMs, trained on vast datasets, to improve performance on specific tasks by fine-tuning them on smaller, task-specific datasets, instead of training from scratch.
- Procedure: Pre-training, followed by fine-tuning
- Benefits:
  - Reduced Training Time and Resources (for fine-tuning)
  - Improved Performance
  - Generalization
- Applications: language translation, sentiment analysis, question answering…

Transfer learning & large models: My Opinion

- Inference in (hierarchical) large models
  - Examples given before…

Summary: Domain Adaptation / Transfer Learning

- Why domain adaptation / transfer learning?
- What if you have only a small number of domain?
- What if you have access to many domains?
- Future?