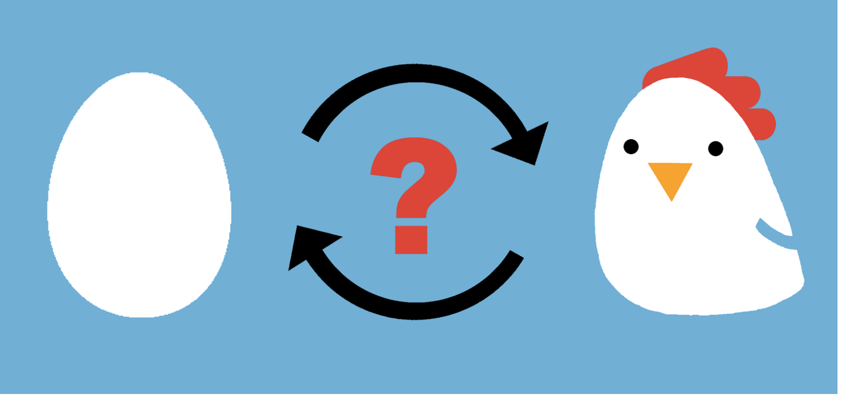


CBMS Conference -- Foundations of Causal Graphical Models and Structure Discovery

Lecture 10

A Causal Perspective of Learning under Heterogeneity:

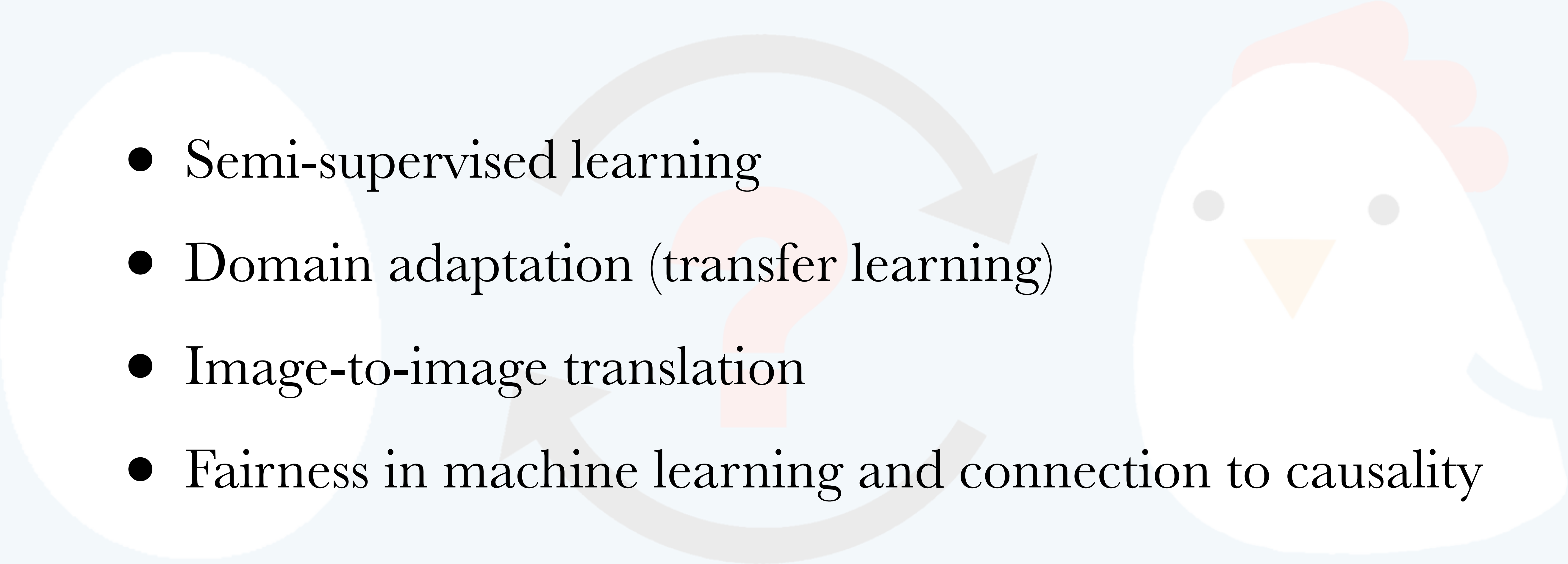
Semi-Supervised Learning, Transfer Learning, and Algorithmic Fairness)

Instructor: Kun Zhang

Carnegie Mellon University

MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

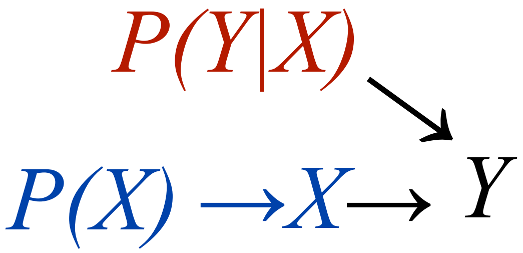
Outline



- Semi-supervised learning
- Domain adaptation (transfer learning)
- Image-to-image translation
- Fairness in machine learning and connection to causality

“Independence” & “Dependence” Implied by Causal Models

- Generating process for cause X is “independent” from that generates effect Y from X

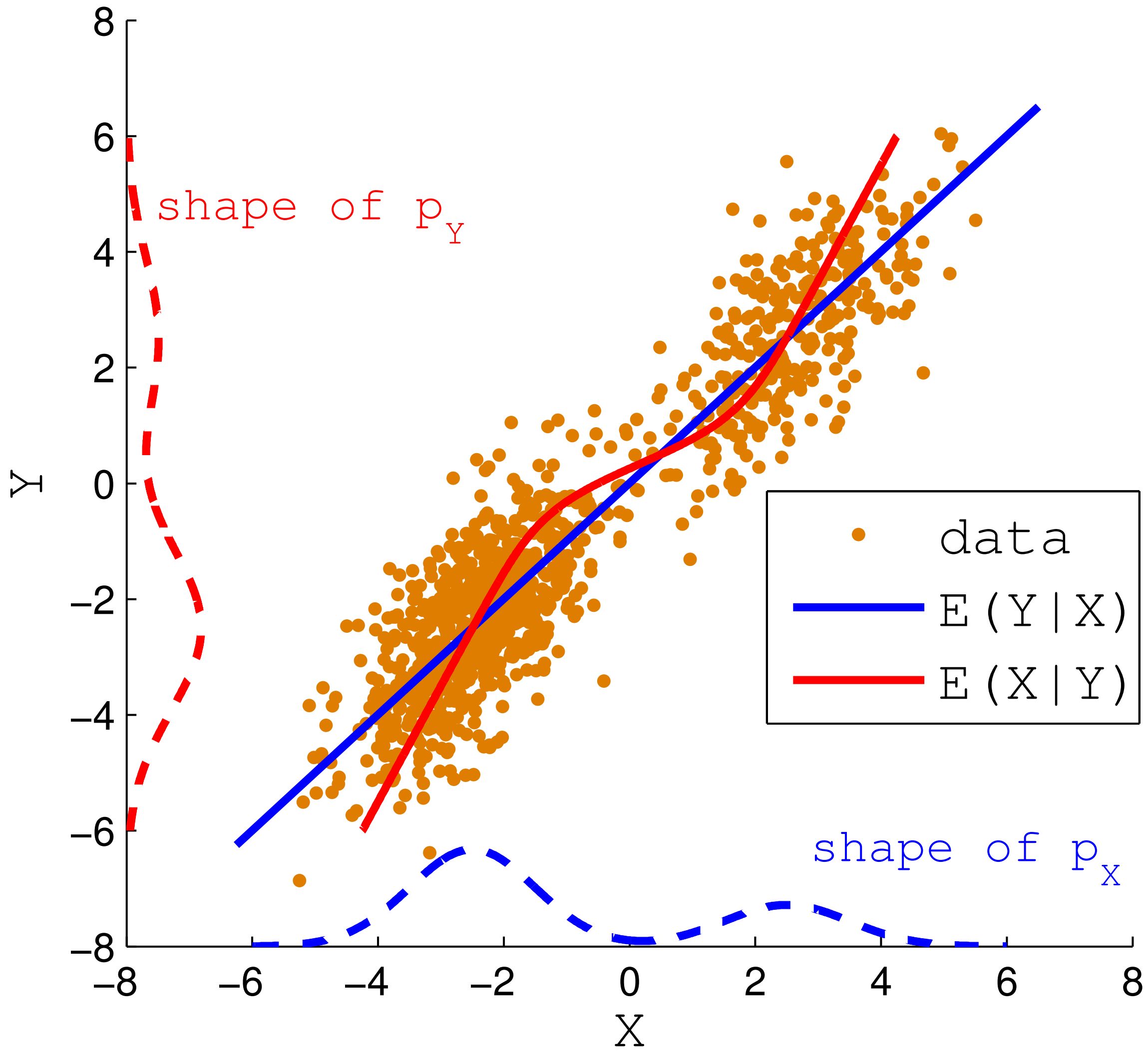


$P(Y|X)$

$P(X) \rightarrow X \rightarrow Y$

- p(X) “independent” (irrelevant to) from p(Y|X)

Statistical Intuition: p(effect) “dependent” on p(cause|effect):

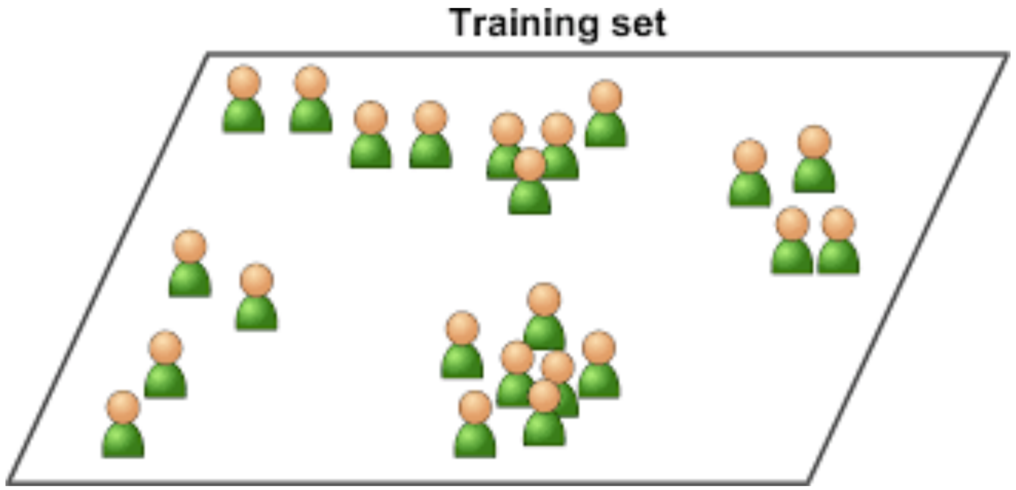


For Instance, Causal View of Clustering

- Clustering (unsupervised)…
  - What if X →Y without a confounder?

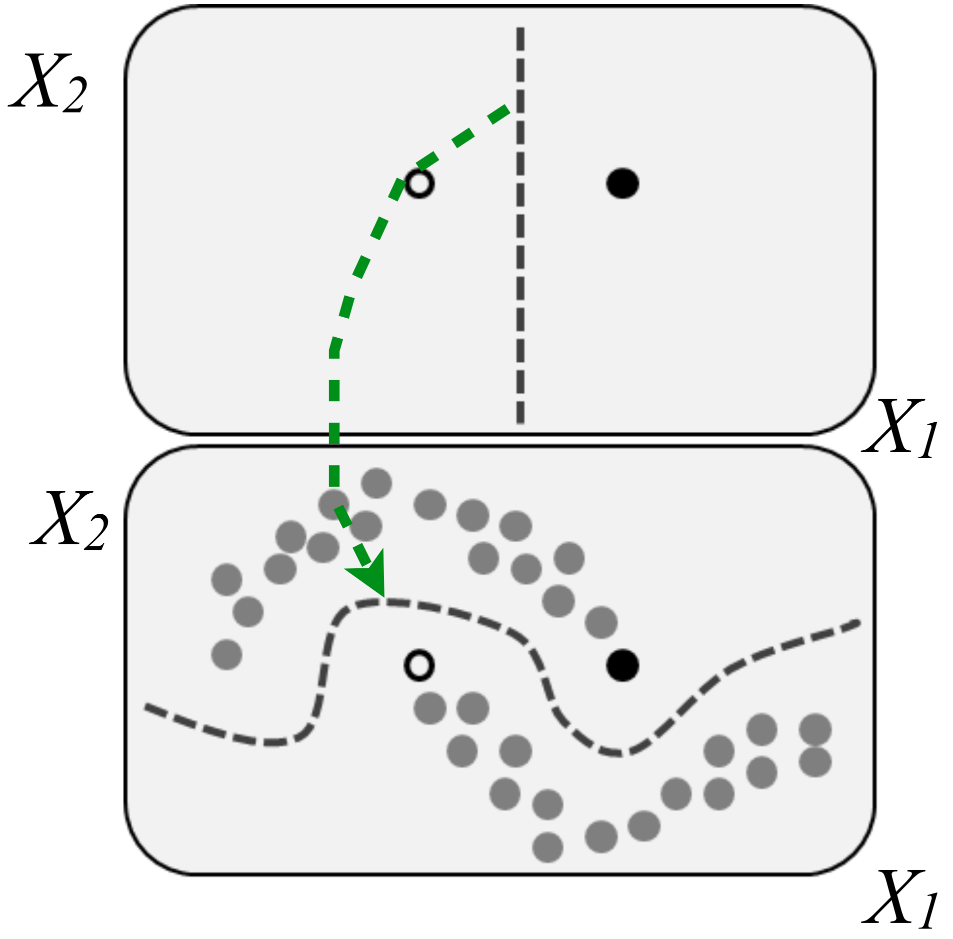
$$\begin{tikzcd}
 & P(Y|X) \arrow[dr] & \\
P(X) \arrow[r] & X \arrow[r] & Y
\end{tikzcd}$$

  - What if Y→X without a confounder?



Problem 1: Semi-Supervised Learning

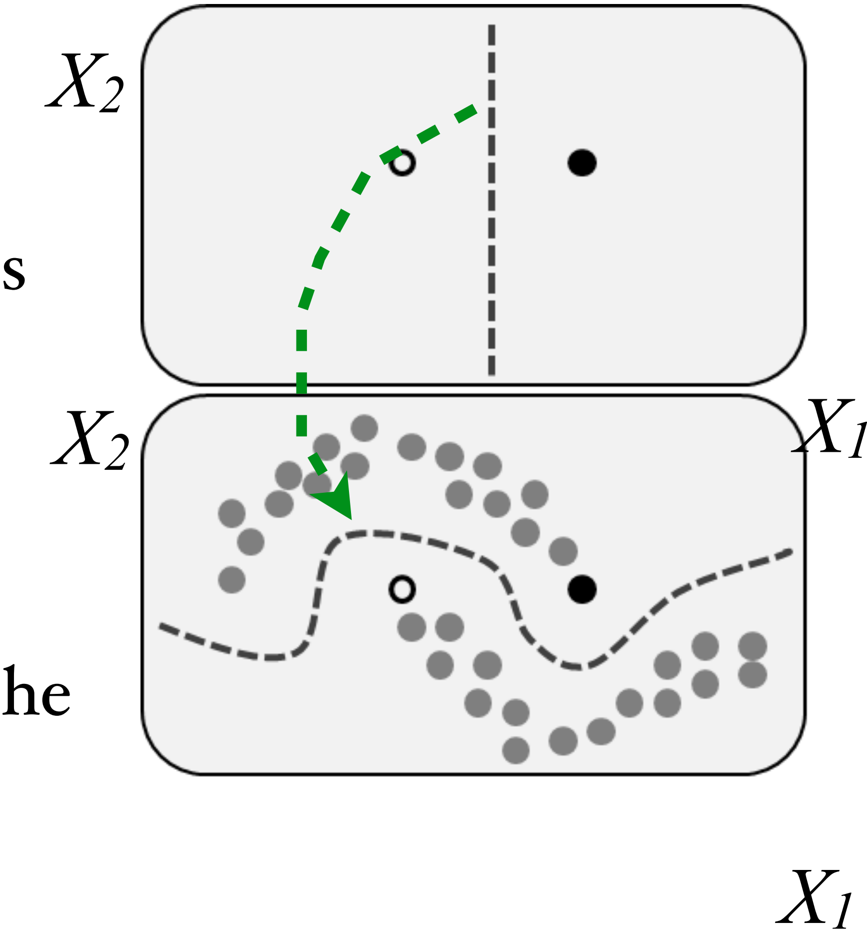
- X: features; Y: label (or target)
- Semi-supervised learning: more precise estimate of P_X helps learn $P_{Y|X}$
- Utilizes dependence between p_X and $p_{Y|X}$ (Schölkopf et al., 2012)
  - $X \rightarrow Y$: unlabeled points do not help
  - $Y \rightarrow X$: Yes



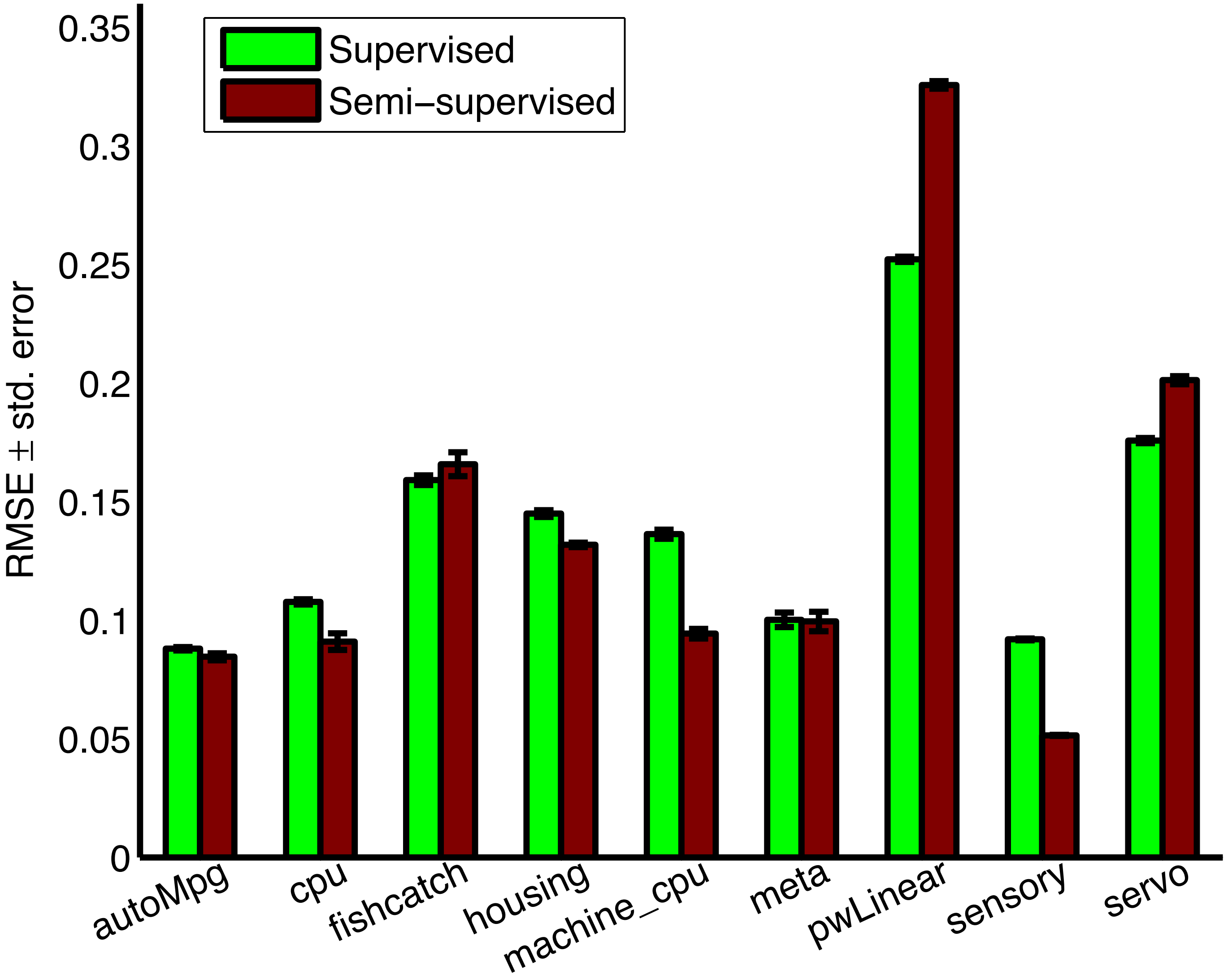
Schölkopf et al., On causal and anticausal learning, ICML 2012

Typical Assumptions

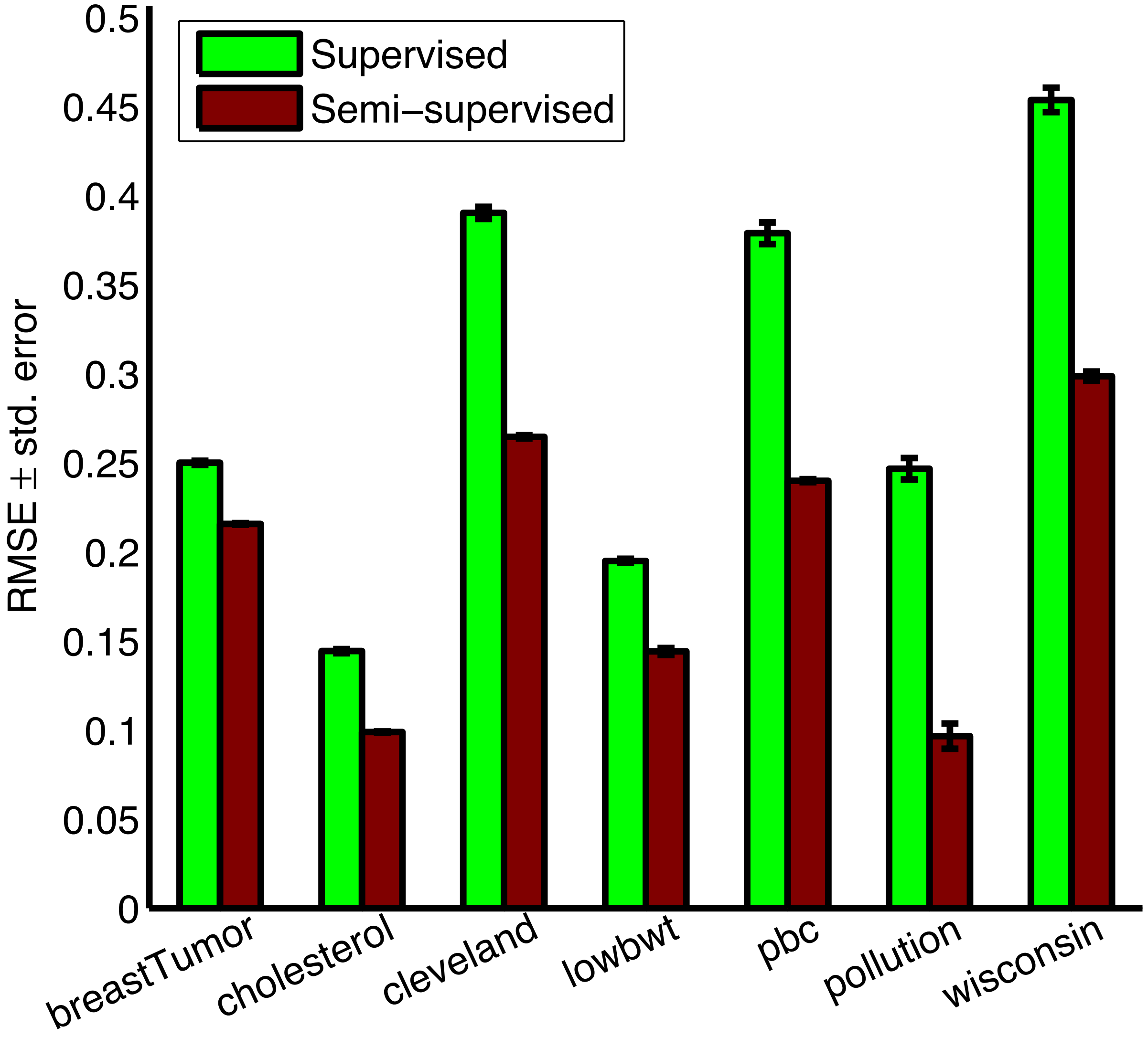
- Continuity assumption
  - Points that are close to each other are more likely to share a label.
  - Additionally yields a preference for decision boundaries in low-density regions, so few points are close to each other but in different classes.
- Cluster assumption
  - The data tend to form discrete clusters, and points in the same cluster are more likely to share a label (although data that shares a label may spread across multiple clusters).
  - Special case of the smoothness assumption.
- Manifold assumption
  - The data lie approximately on a manifold of much lower dimension than the input space.



Some Meta-Analysis of Previous Experimental Results



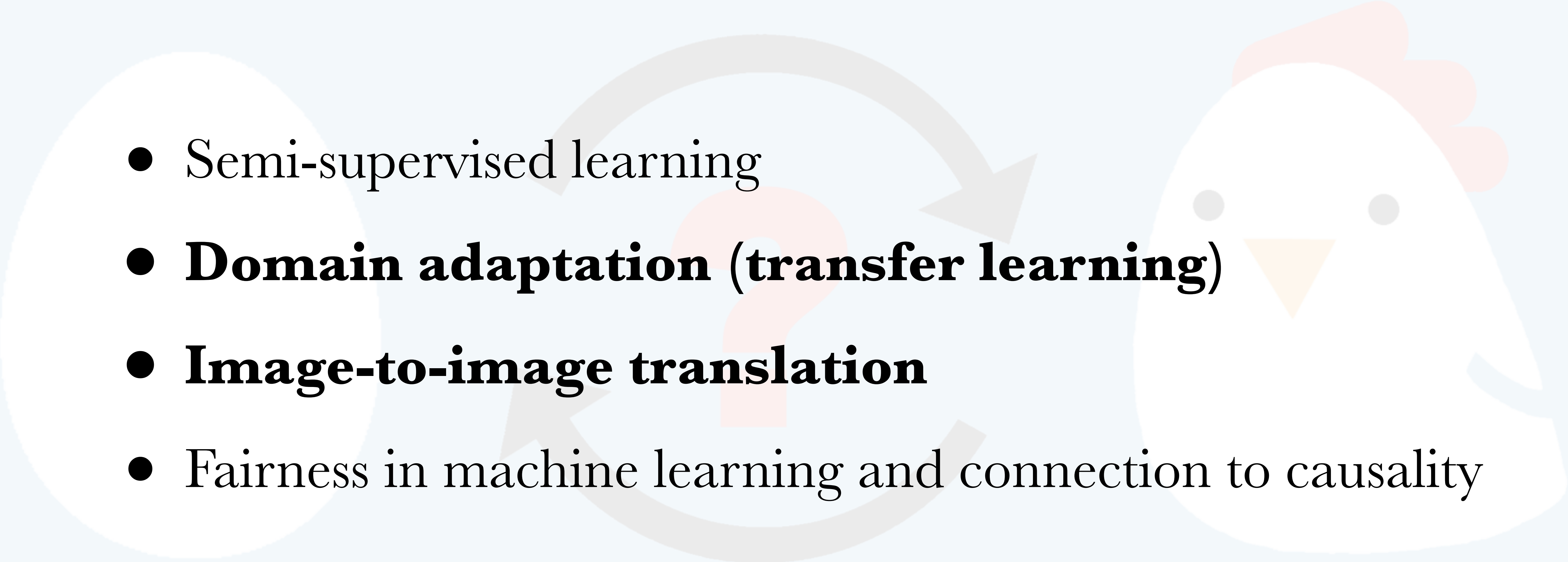
Semi-supervised regression on causal datasets (X→Y)



Semi-supervised regression on anticausal (Y→X)/ confounded datasets

- X: features; Y: label (or target)

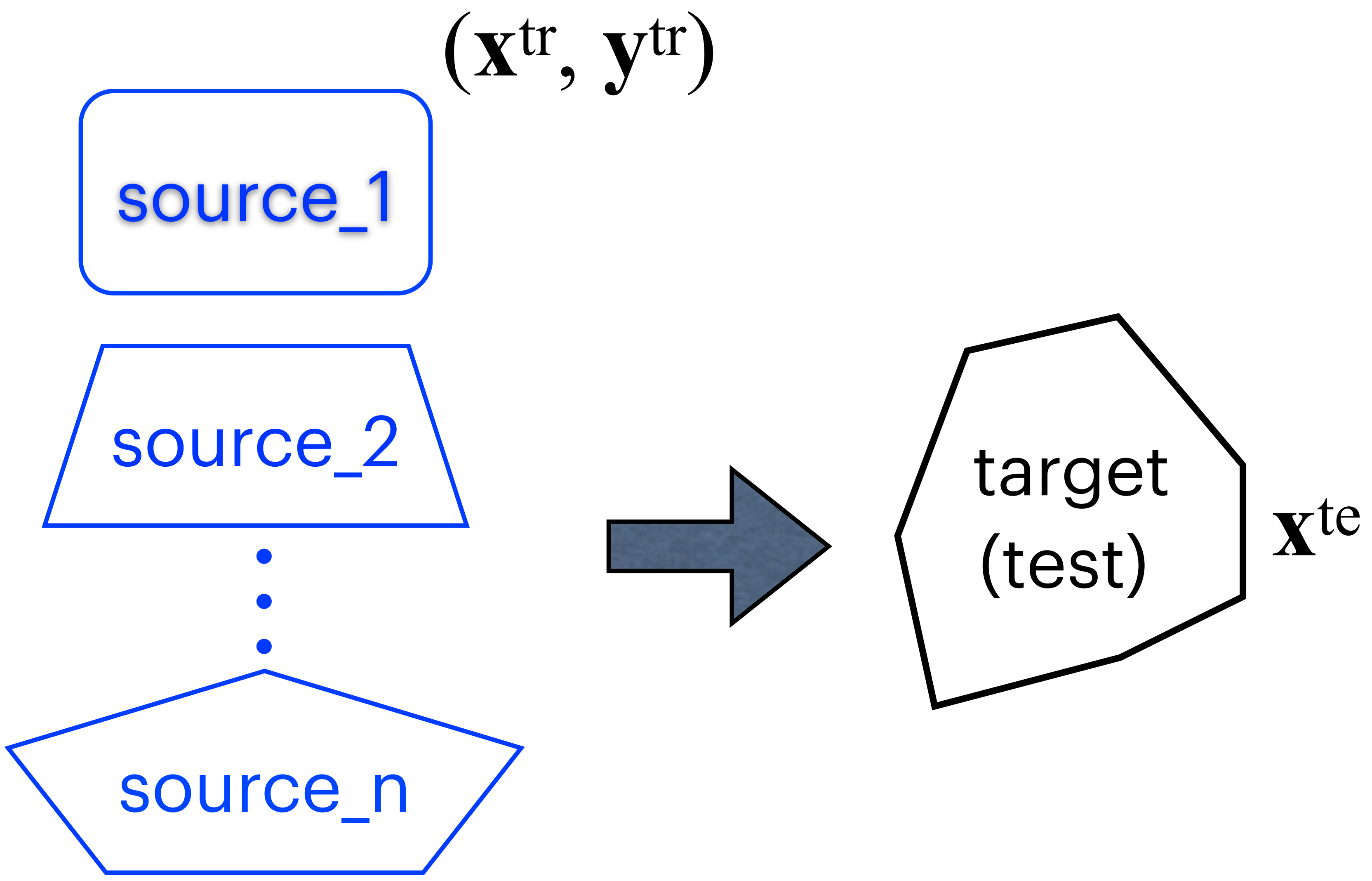
Outline



- Semi-supervised learning
- **Domain adaptation (transfer learning)**
- **Image-to-image translation**
- Fairness in machine learning and connection to causality

Domain Adaptation

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


$(\mathbf{x}^{\mathrm{tr}}, \mathbf{y}^{\mathrm{tr}})$
source_1
source_2
source_n
target
(test)
$\mathbf{x}^{\mathrm{te}}$

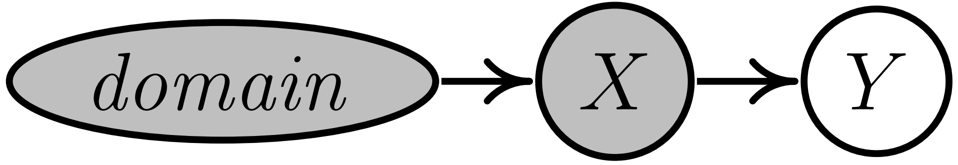
high-level representation, e.g., $Y \rightarrow X$

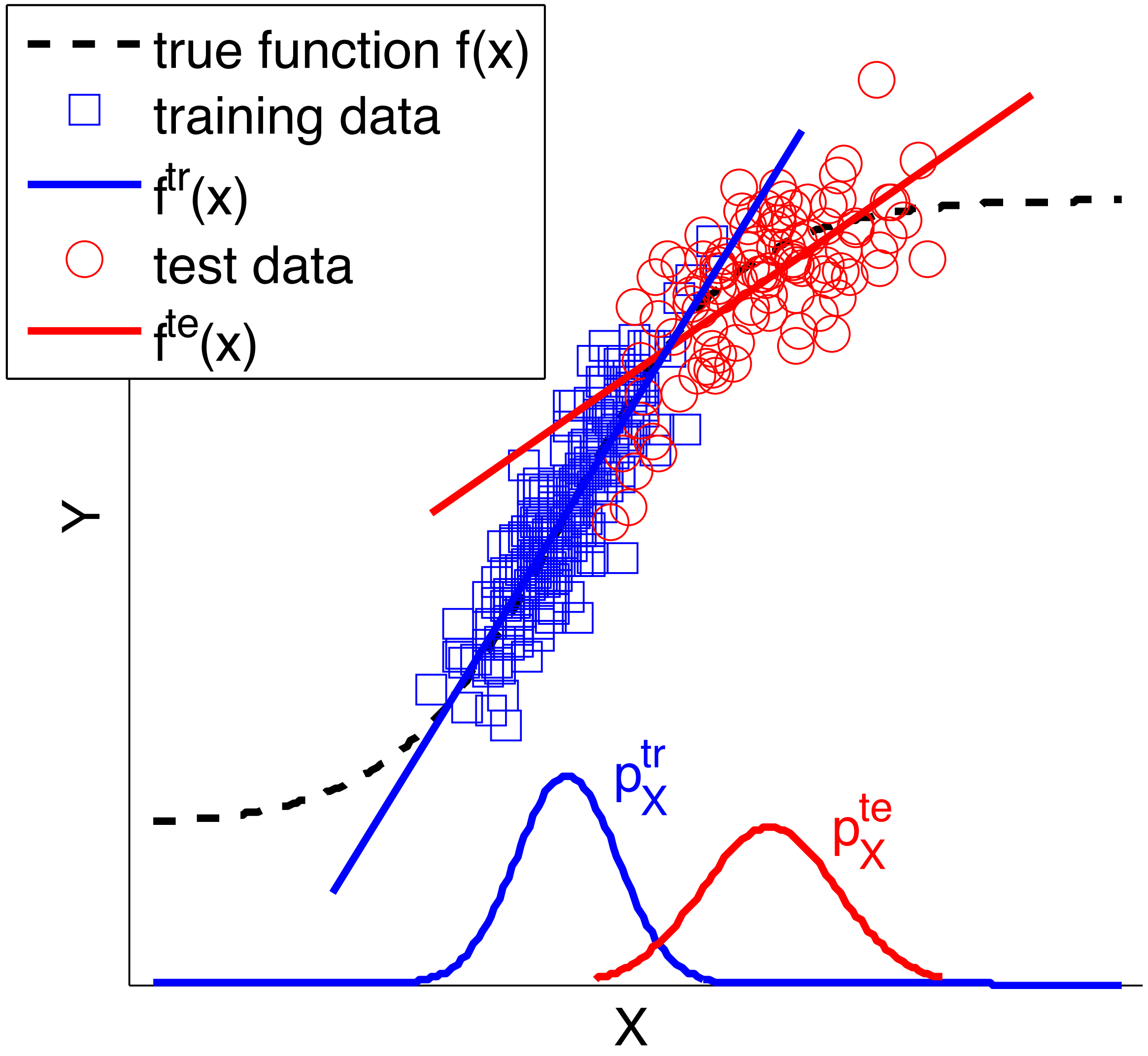
Prob. model $P^{(1)}(X,Y)$, $P^{(2)}(X,Y)$, $P^{(3)}(X,Y)$, ... $P^{(k)}(X,Y)$...

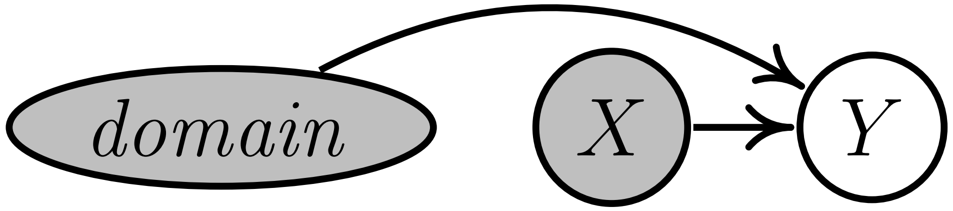
Possible Situations for Domain Adaptation: When **X**→Y

covariate shift

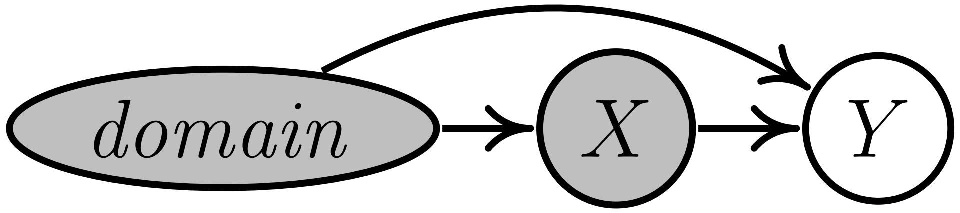
(Shimodaira00; Sugiyama etal.08; Huang etal.07, Gretton etal.08...)







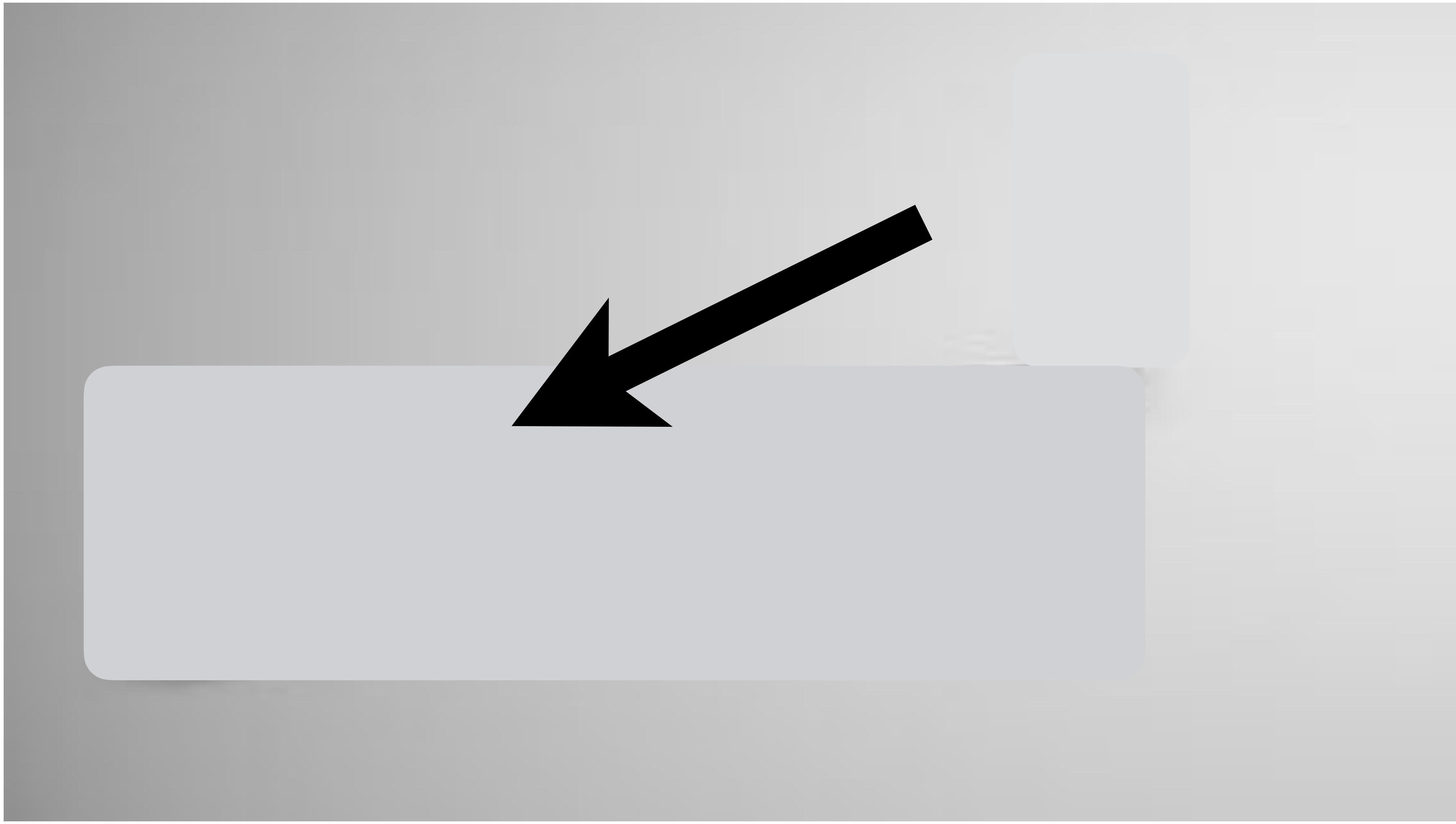
☹ no clue as to find $P^{te}_{Y|X}$ (with one source domain)

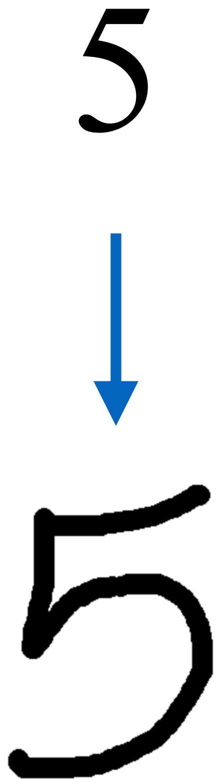


What Features/Components to Transfer?

- 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)

Causality may Matter in Prediction: An Illustration

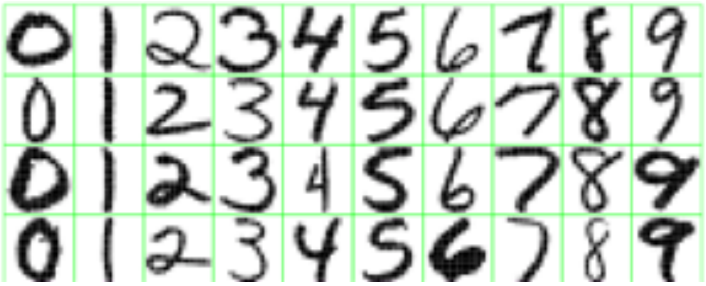




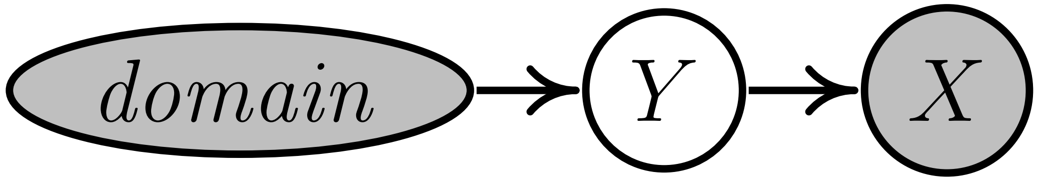
Understanding connections between different scenarios & *modeling* differences

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

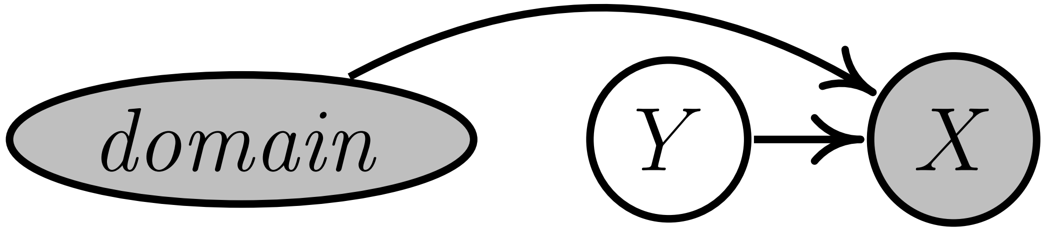
- Y is usually the cause of X (especially for classification)



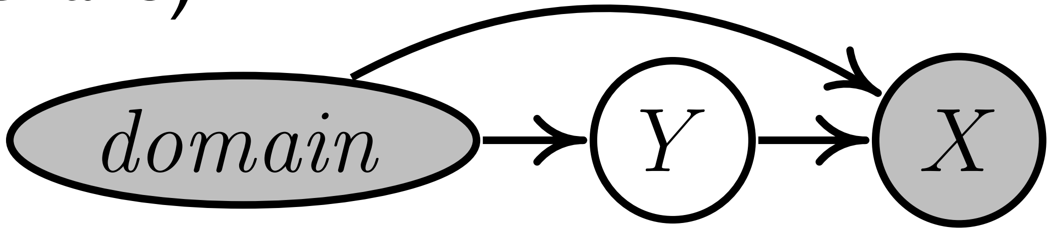
- Target shift (TarS)



- Conditional shift (ConS)



- Generalized target shift (GeTarS)



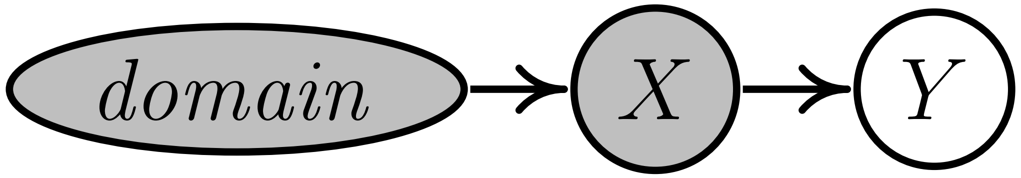
P_X^{te} helps find $P_{Y|X}^{te}$

involved parameters estimated by matching P_X

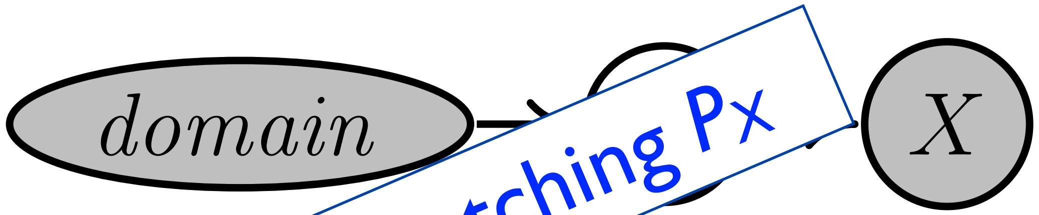
Zhang et al., ICML 2013; Schölkopf et al., 2012; Zhang et al., AAAI 2015; Gong et al., ICML 2016; Stojanov et al., AISTATS 2018; Zhao et al., ICML 2019; Fu et al., CVPR 2019…

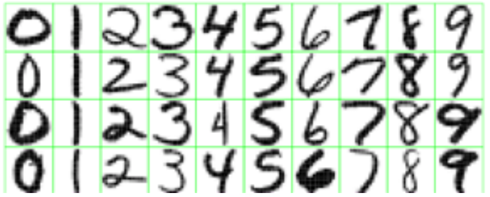
Traditional Methods Assume How Distribution Changes...

- Covariate shift



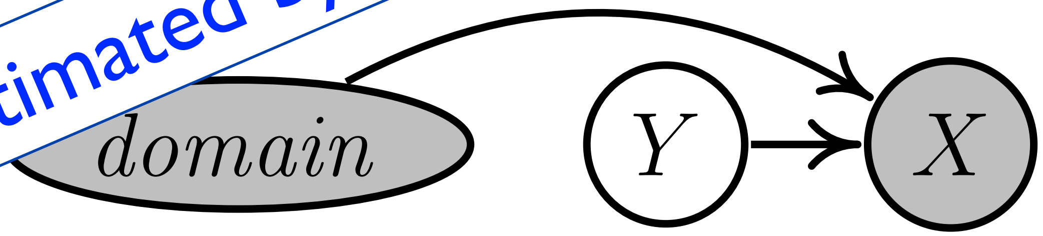
- Target shift



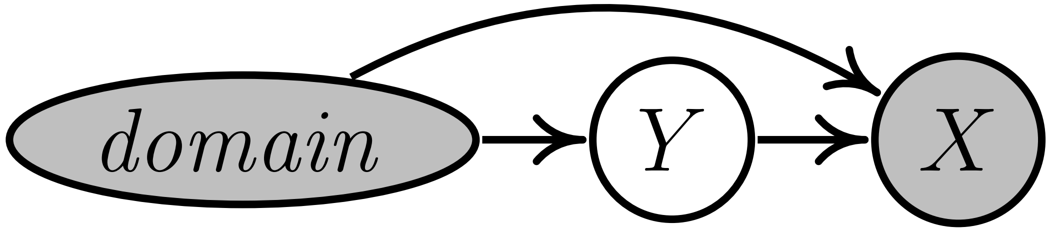


involved parameters estimated by matching P_X

- Conditional shift

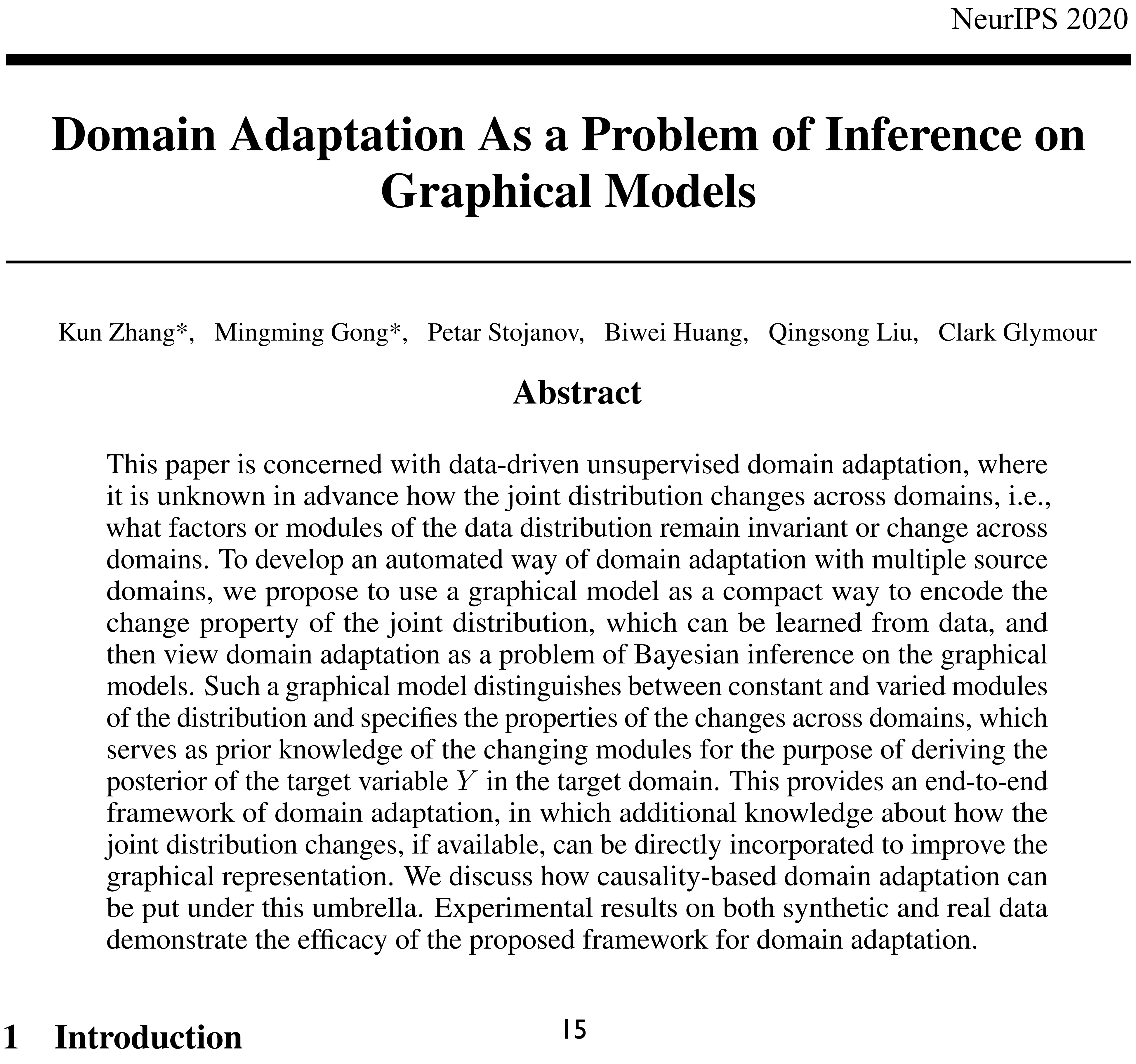


- Generalized target shift



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…)



NeurIPS 2020

Domain Adaptation As a Problem of Inference on Graphical Models

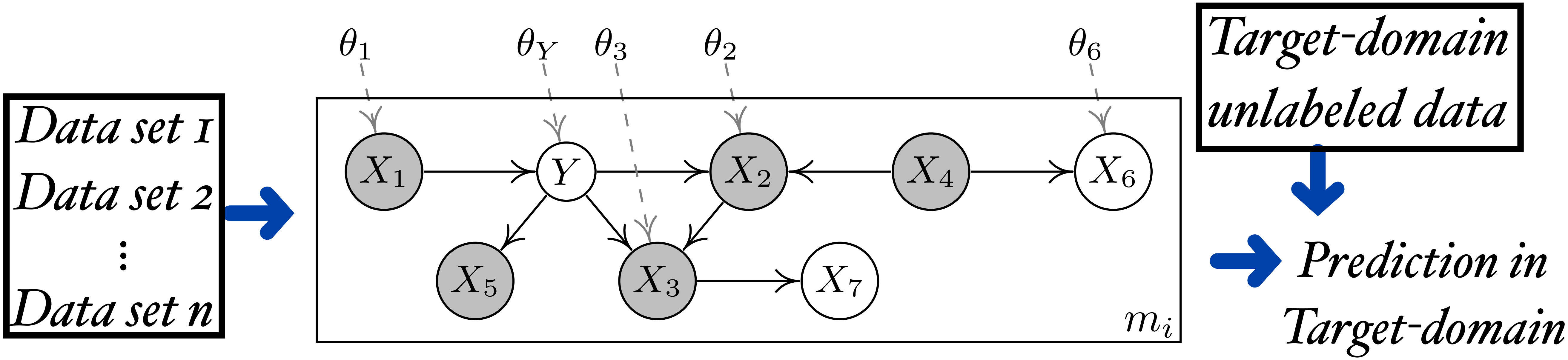
Kun Zhang*, Mingming Gong*, Petar Stojanov, Biwei Huang, Qingsong Liu, Clark Glymour

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

An Approach to Data-Driven Domain Adaptation



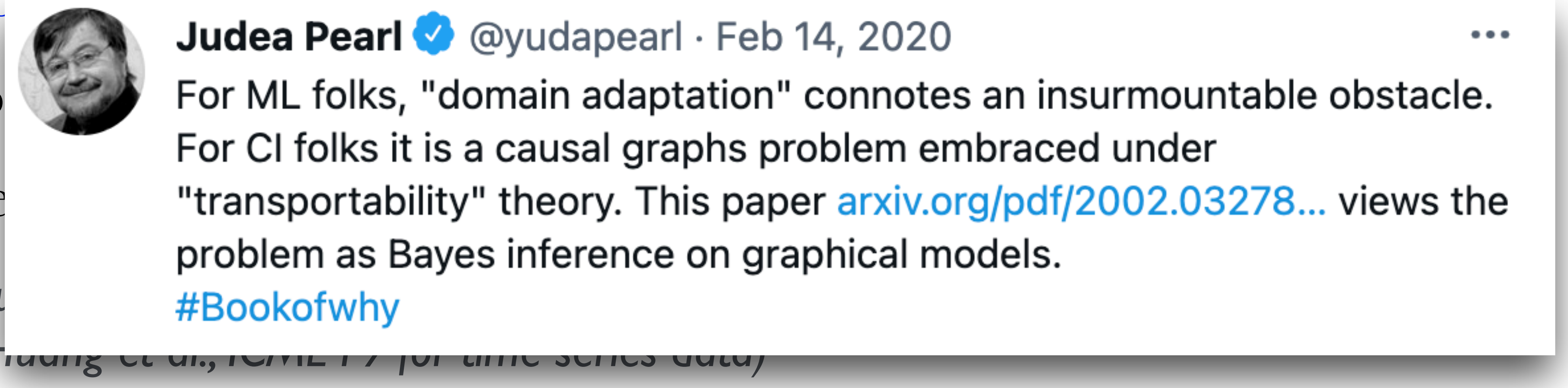
Data set 1
Data set 2
⋮
Data set n

Target-domain unlabeled data

Prediction in Target-domain

- Only relevant features needed to predict Y
- Augmented graph learned by CD-NOD
  - Independently changing modules θ_i
  - Special case: invariant modules
- Domain adaption: 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)



Judea Pearl @yudapearl · Feb 14, 2020

For ML folks, "domain adaptation" connotes an insurmountable obstacle. For CI folks it is a causal graphs problem embraced under "transportability" theory. This paper arxiv.org/pdf/2002.03278... views the problem as Bayes inference on graphical models. #Bookofwhy

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)**

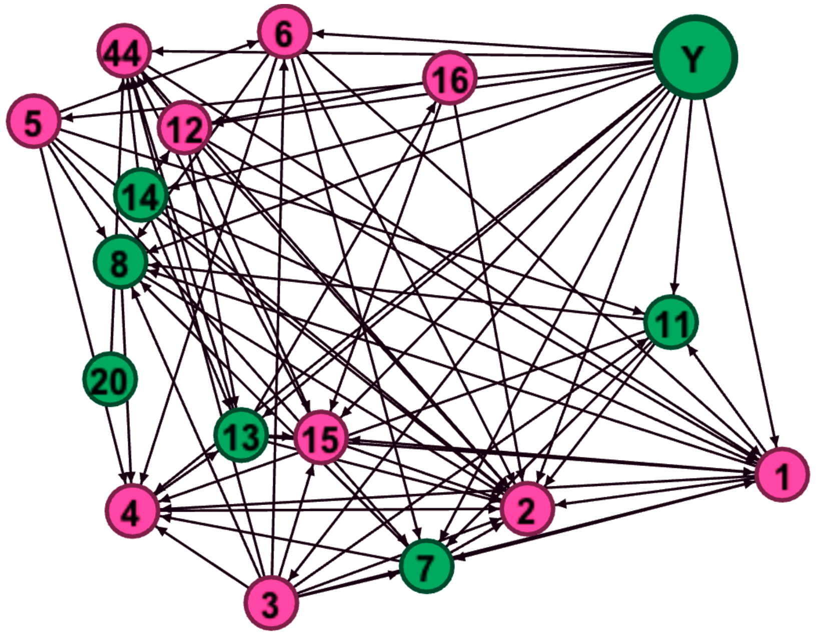
😀

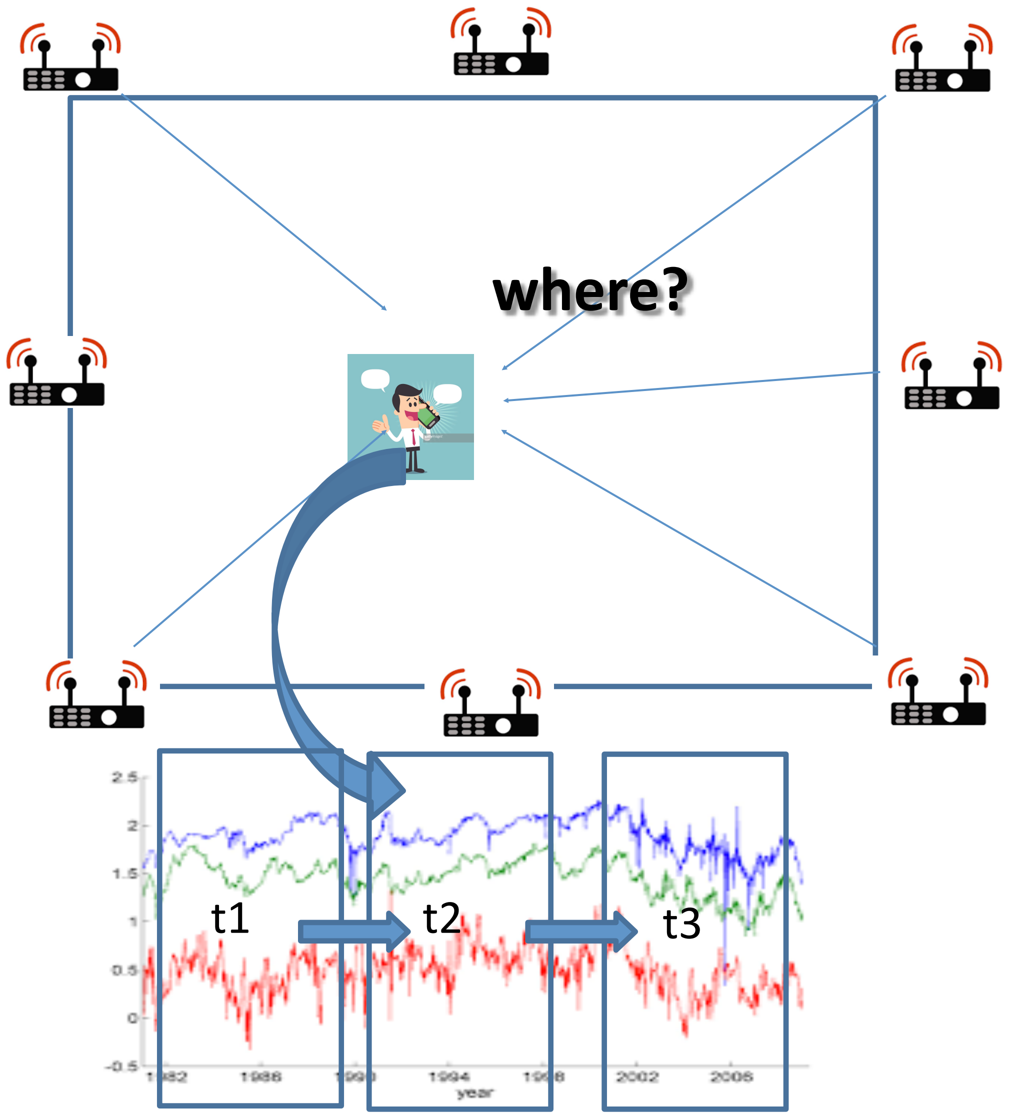
Table 3: Accuracy on the digits data. T: MNIST; M: MNIST-M; S: SVHN; D: SynthDigits.

	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

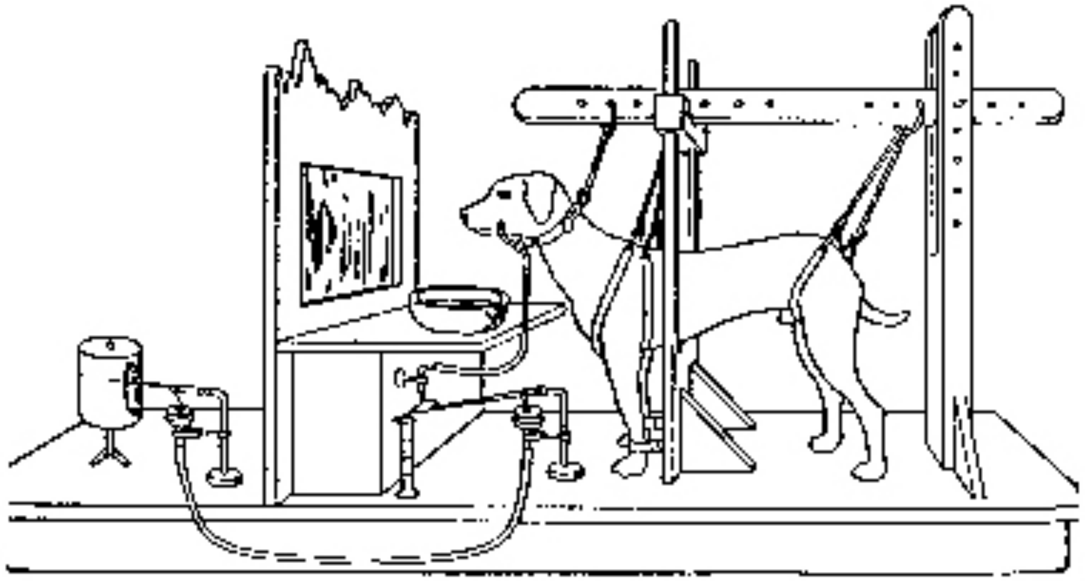
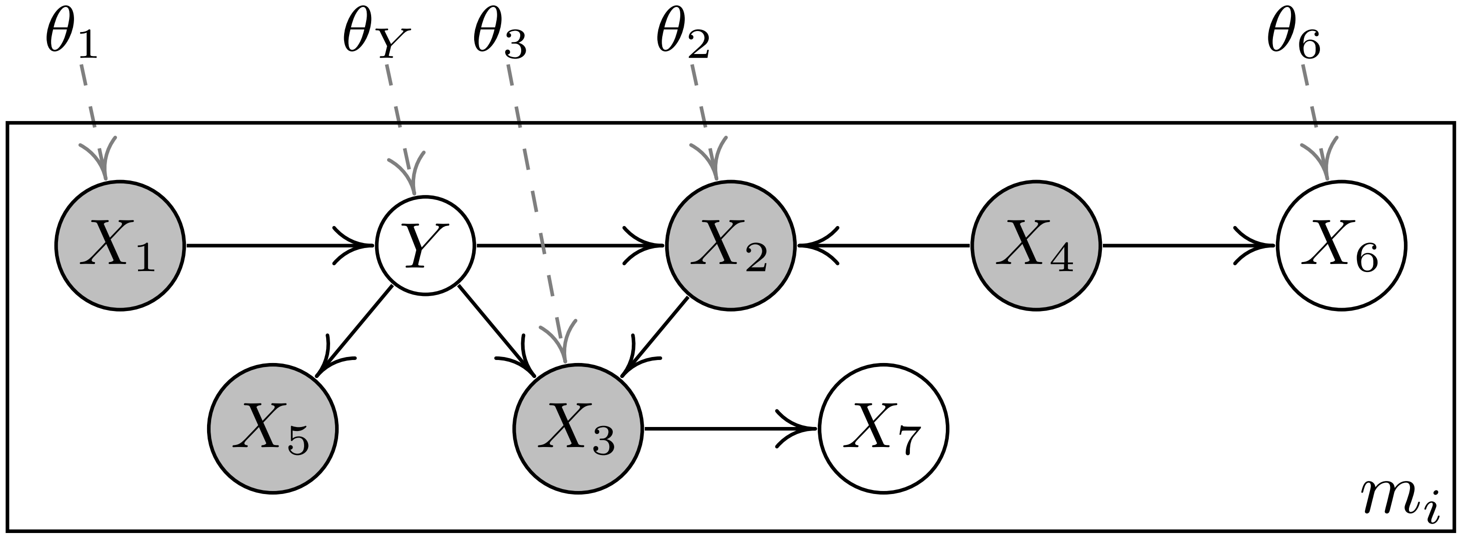


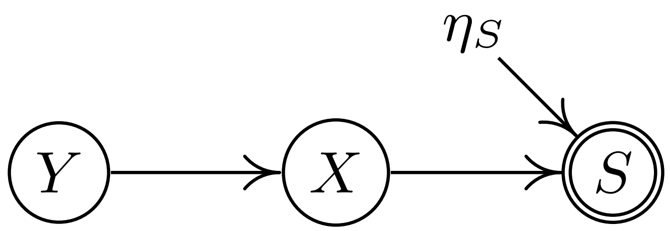
Fig. 2.

- *“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.”*

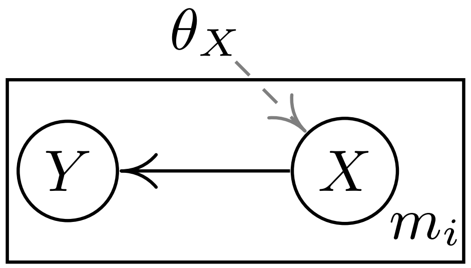
Augmented Graph



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

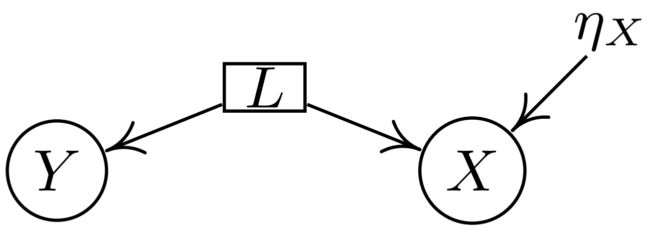


because $p(Y|X)$ is invariant across domains

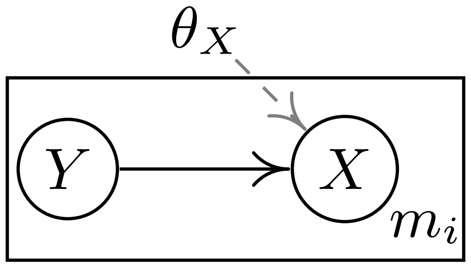


(a) The underlying data generating process of Example 1. Y generates (causes) X, and S denotes the selection variable (a data point is included if and only if $S = 1$).

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



because $p(Y)$ is invariant across domains



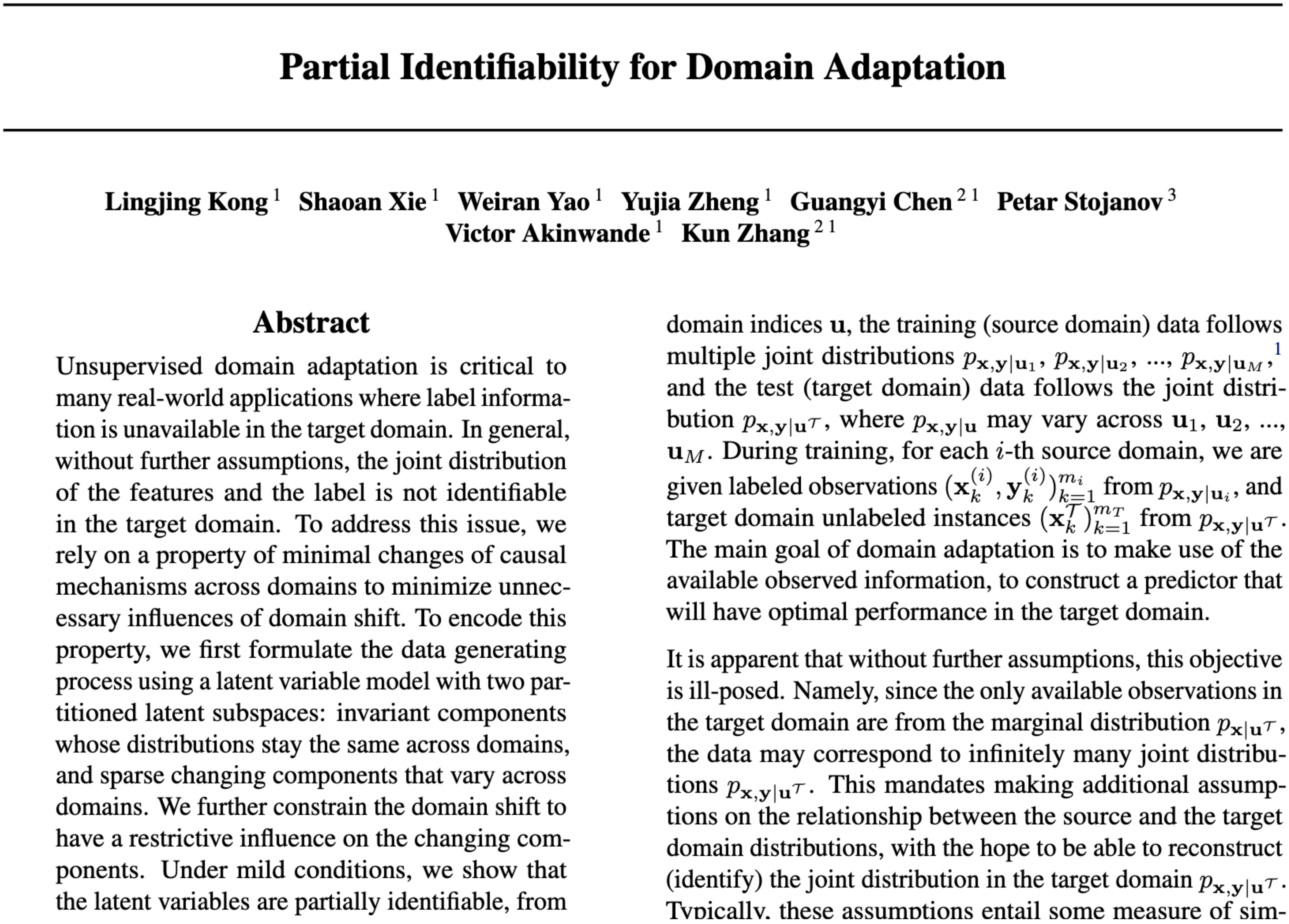
(c) The generating process of Example 2. L is a confounder; the mechanism of X changes across domains, as indicated by η_X.

(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

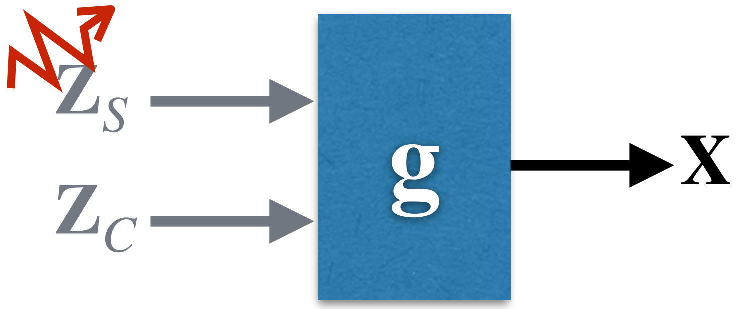
Unsupervised domain adaptation is critical to many real-world applications where label information is unavailable in the target domain. In general, without further assumptions, the joint distribution of the features and the label is not identifiable in the target domain. To address this issue, we rely 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.

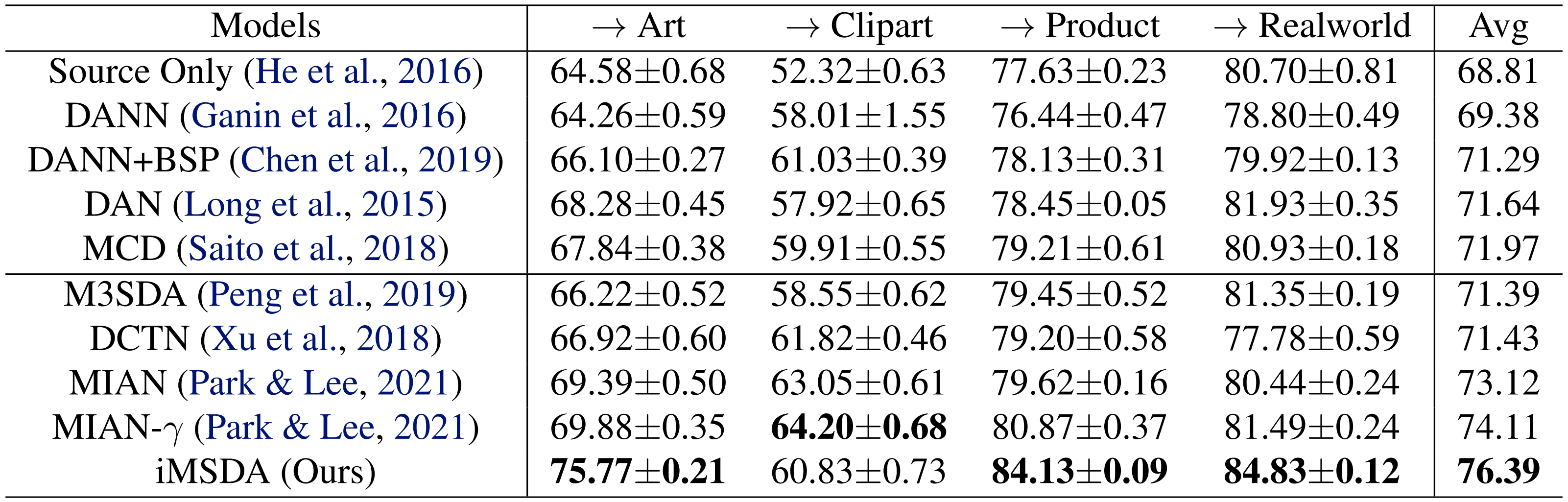
It is apparent that without further assumptions, this objective is ill-posed. Namely, since the only available observations in the target domain are from the marginal distribution $p_{\mathbf{x}|\mathbf{u}^{\mathcal{T}}}$, the data may correspond to infinitely many joint distributions $p_{\mathbf{x},\mathbf{y}|\mathbf{u}^{\mathcal{T}}}$. 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

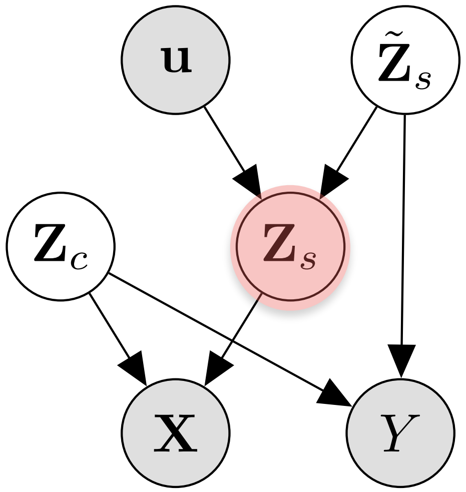


Figure 1. The generating process: The gray shade of nodes indicates that the variable is observable.

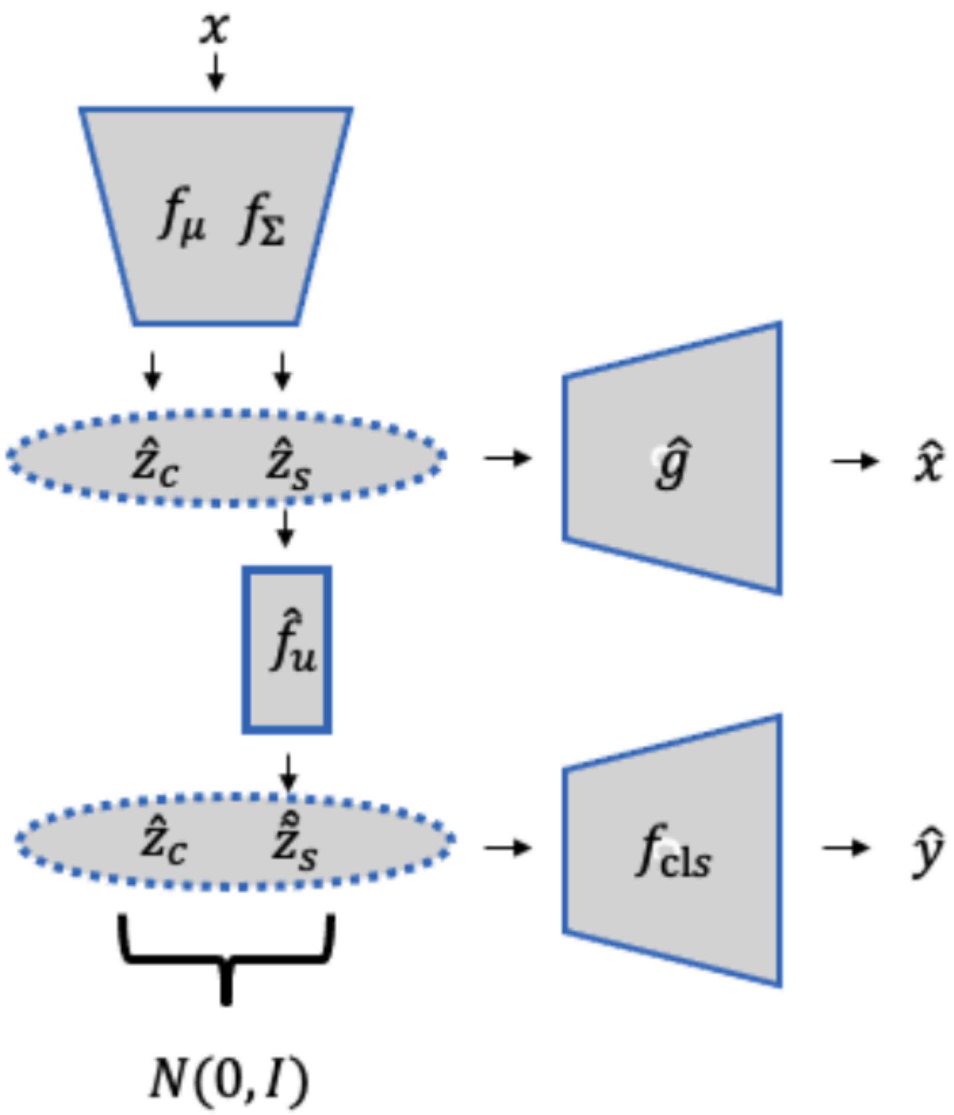
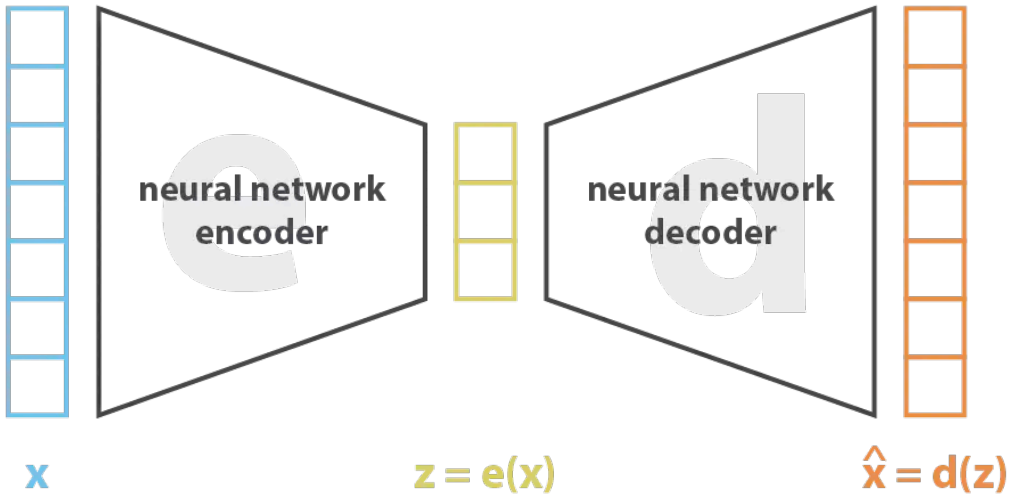
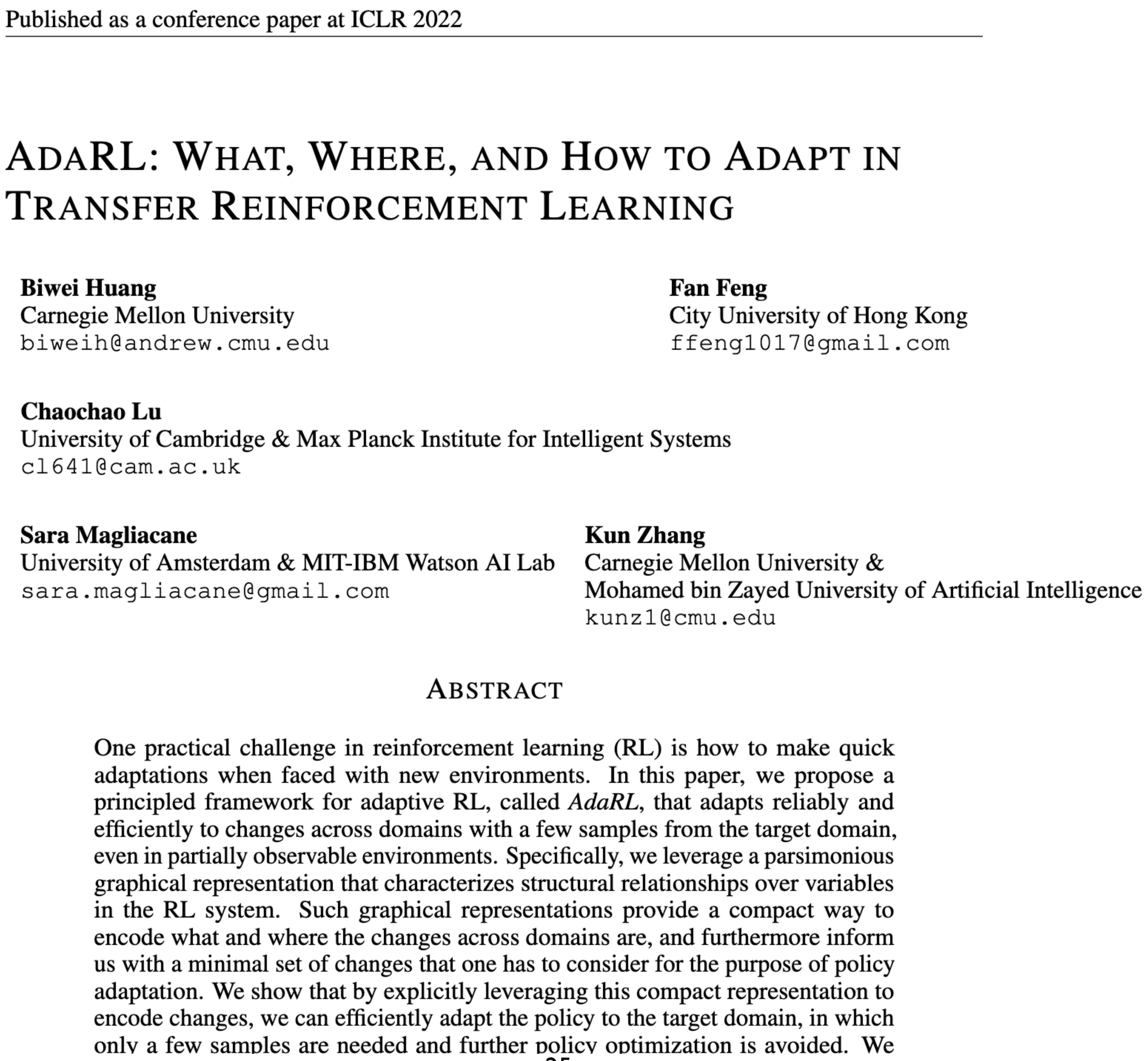


Figure 2. Diagram of our proposed method, **iMSDA**. We first apply the VAE encoder (f_μ, f_Σ) to encode $\mathbf{x}$ into $(\hat{\mathbf{z}}_c, \hat{\mathbf{z}}_s)$, which is further fed into the decoder $\hat{g}$ for reconstruction. In parallel, the changing part $\hat{\mathbf{z}}_s$ is passed through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



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

Autoencoder



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

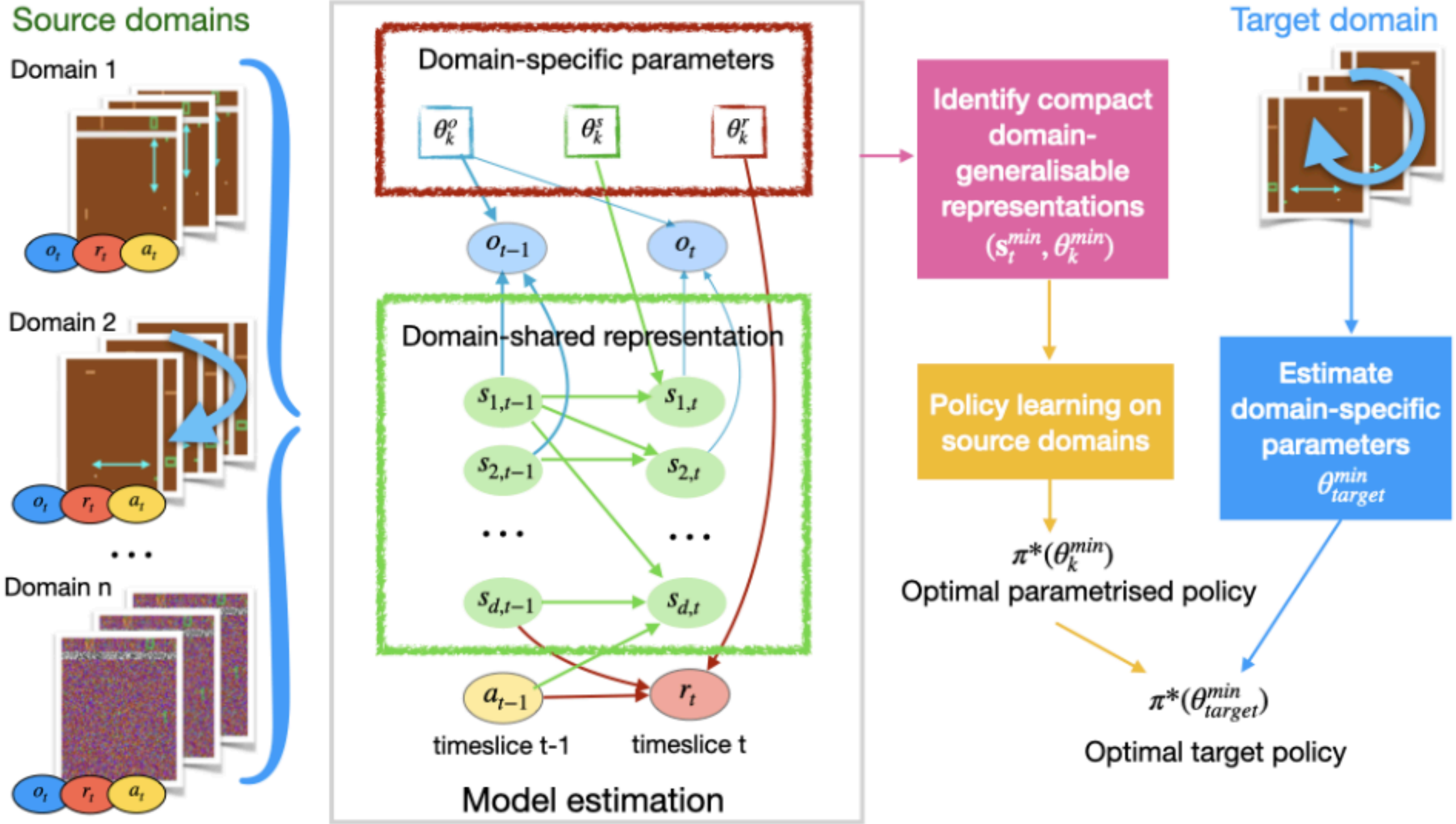
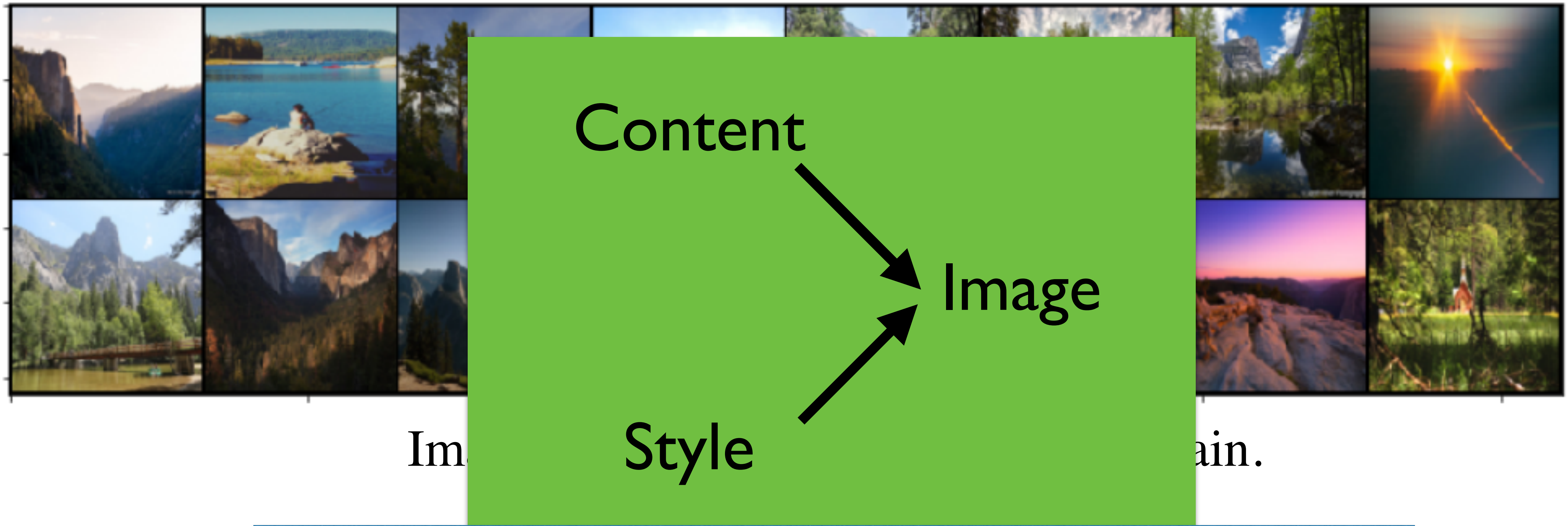
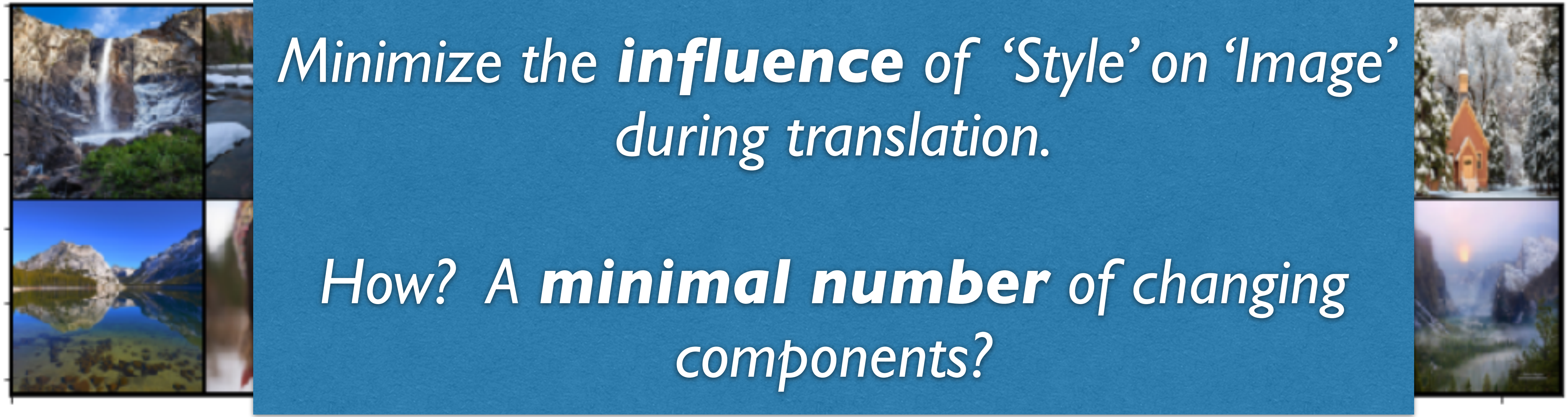


Figure 1: The overall AdaRL framework. We learn a Dynamic Bayesian Network (DBN) over the observations, latent states, reward, actions and domain-specific change factors that is shared across the domains. We then characterize a minimal set of representations that suffice for policy transfer, so that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

Unsupervised Image-to-Image Translation



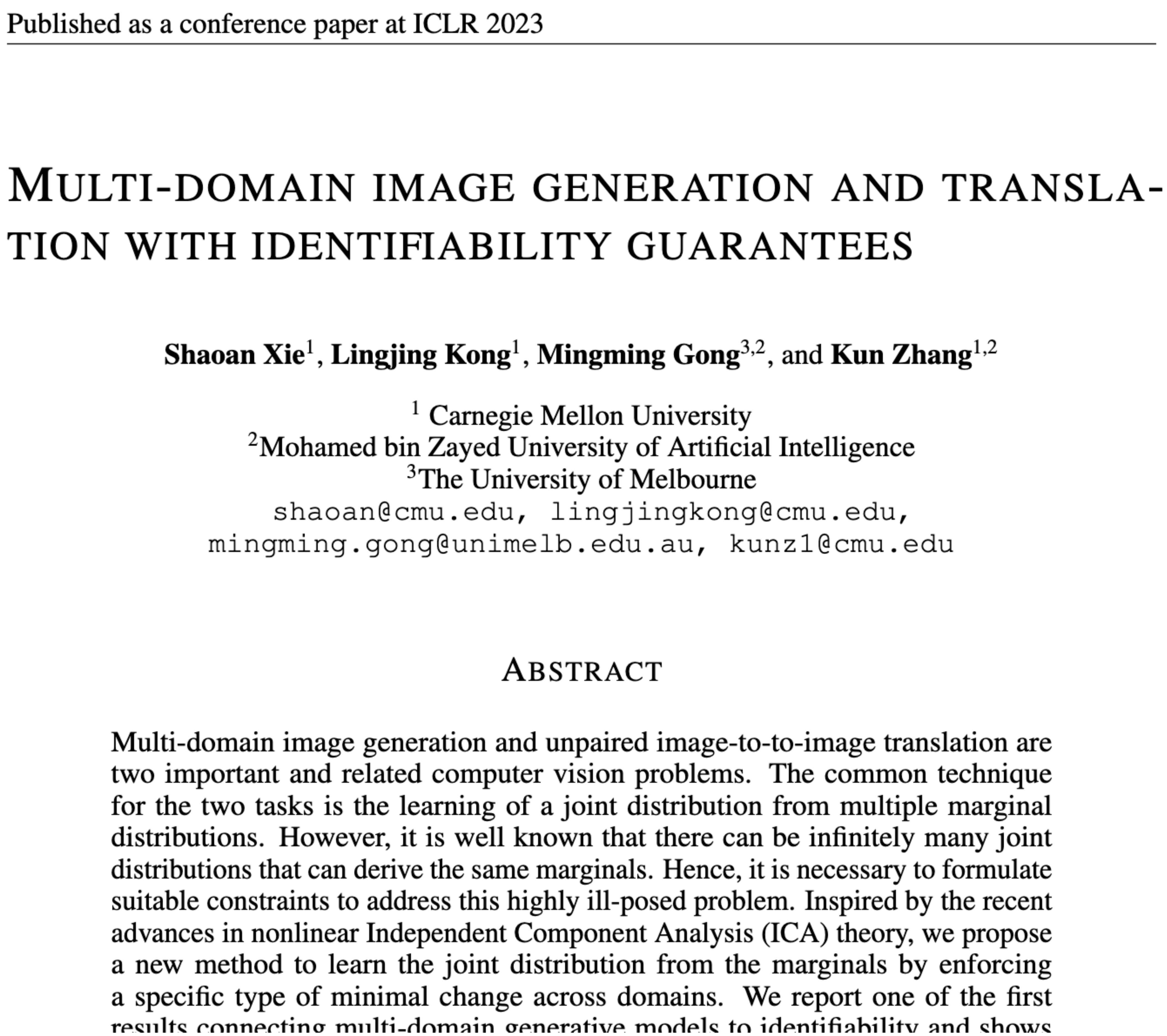
Im ain.



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

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

Images from the winter season domain.



Published as a conference paper at ICLR 2023

MULTI-DOMAIN IMAGE GENERATION AND TRANSLATION WITH IDENTIFIABILITY GUARANTEES

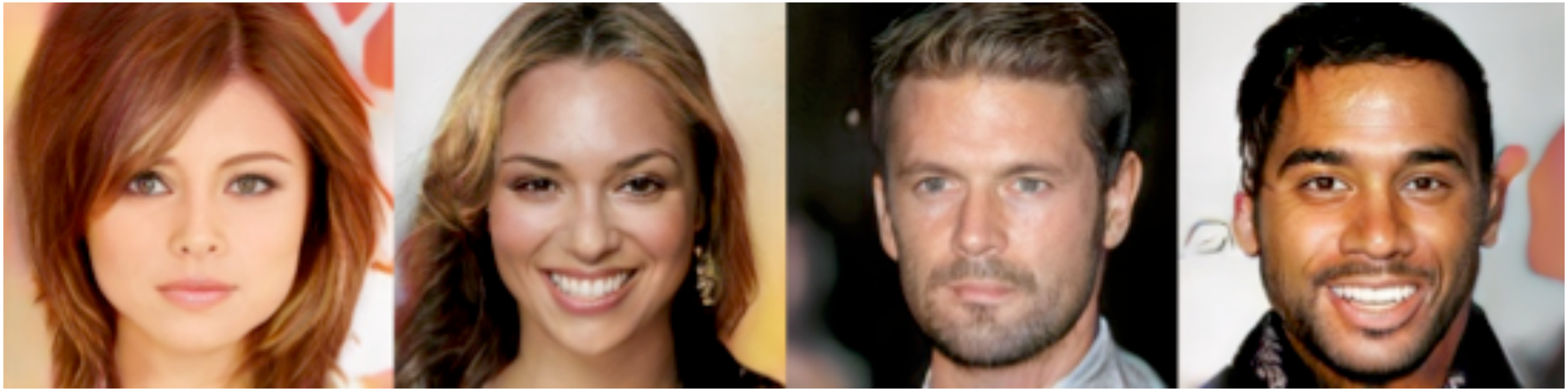
Shaoan Xie[1], **Lingjing Kong**[1], **Mingming Gong**[3,2], and **Kun Zhang**[1,2]

[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

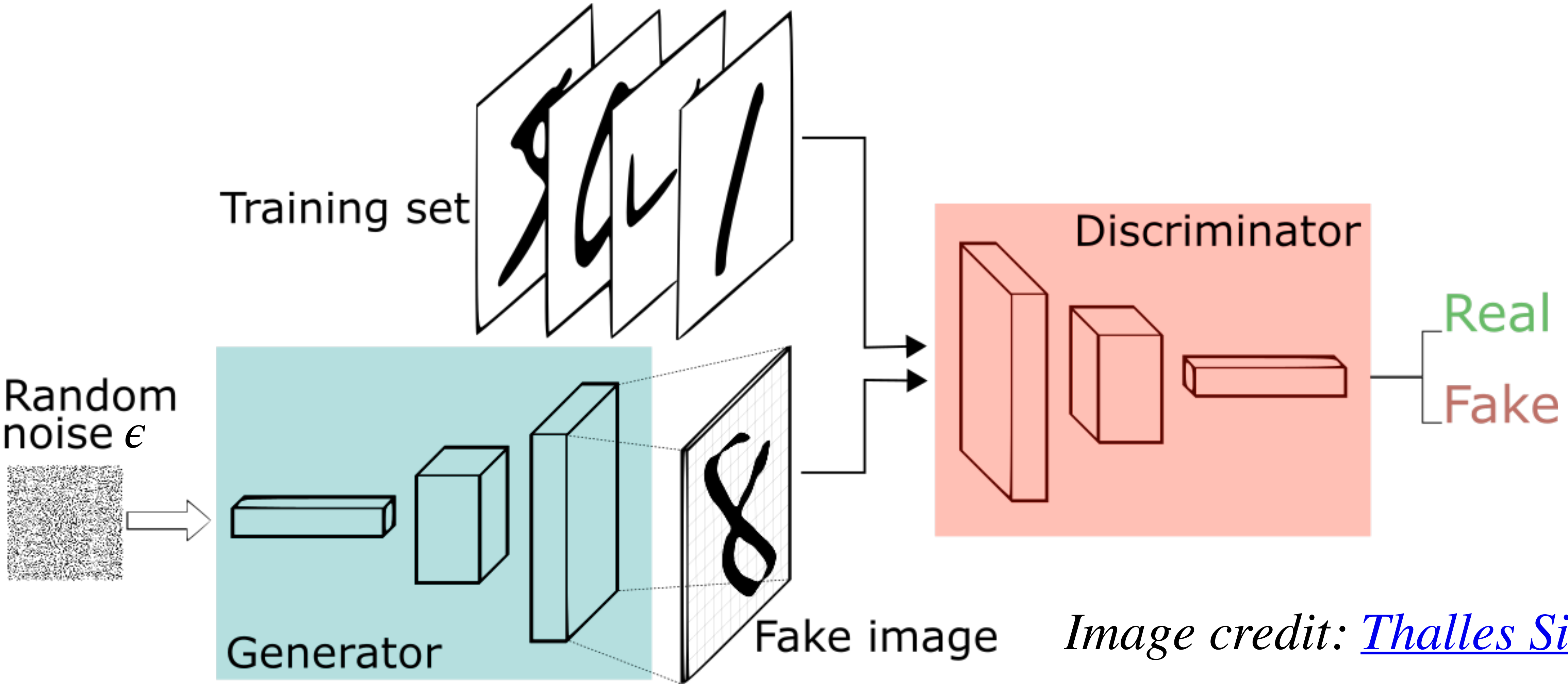
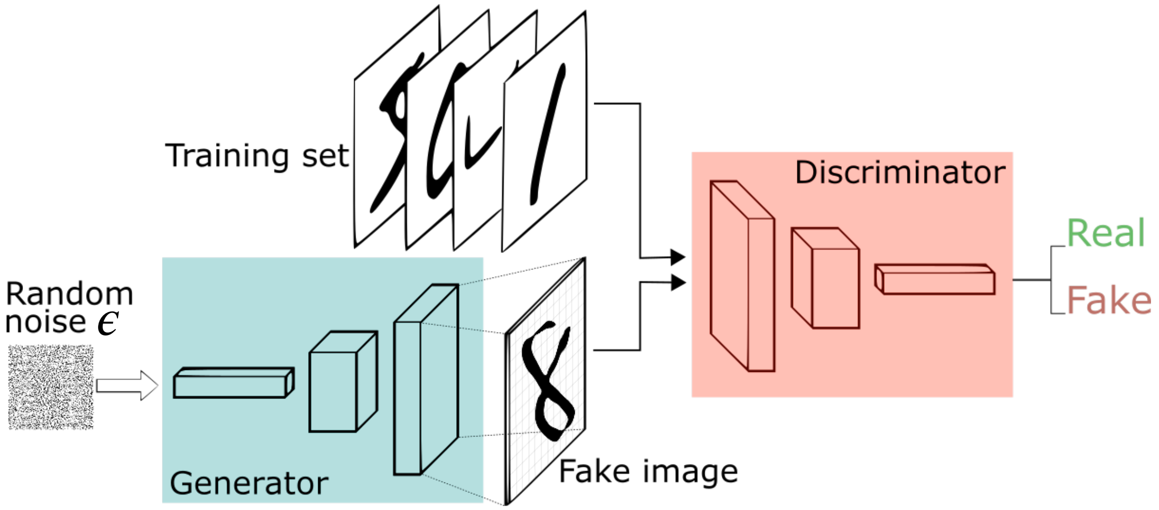


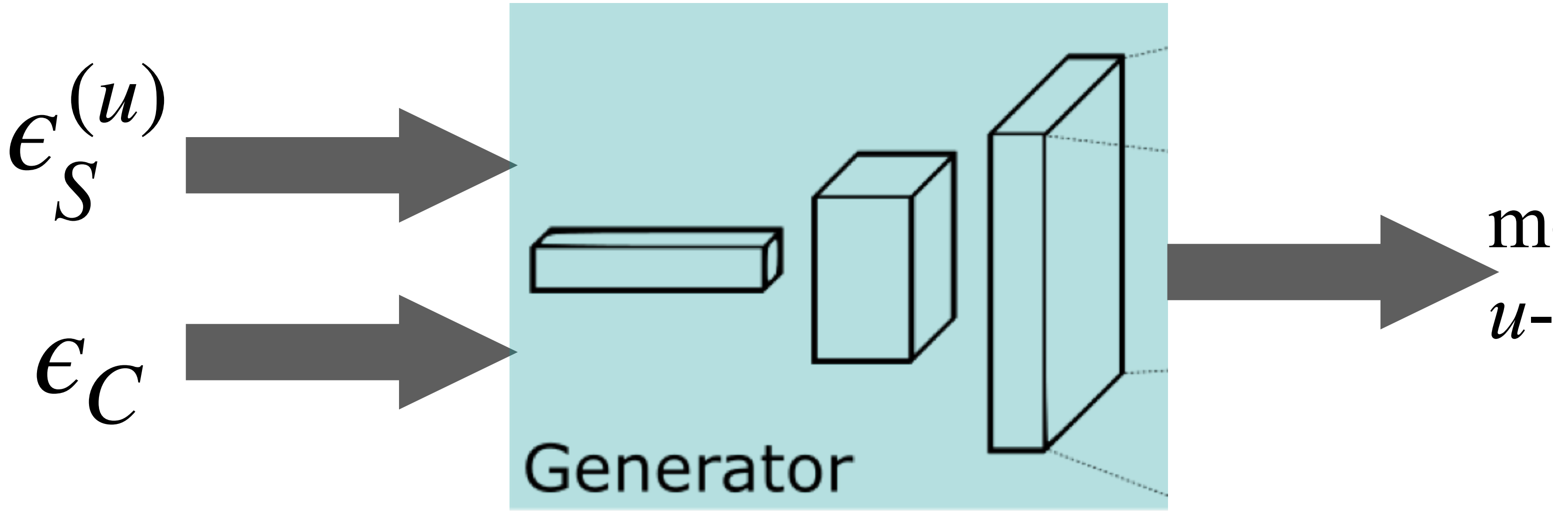
Image credit: Thalles Silva

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





model the data distribution in *u*-th domain

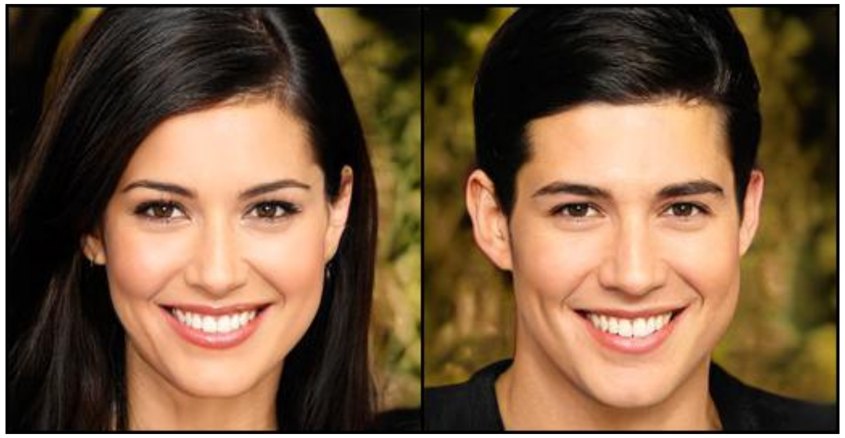
- Match the data distribution across domains, while the dimensionality of $\epsilon_S^{(u)}$ is as small as possible (minimal changes across domains **controlled by ƛ;** no penalty when ƛ=0)

- Correspondence relations among domains are identifiable

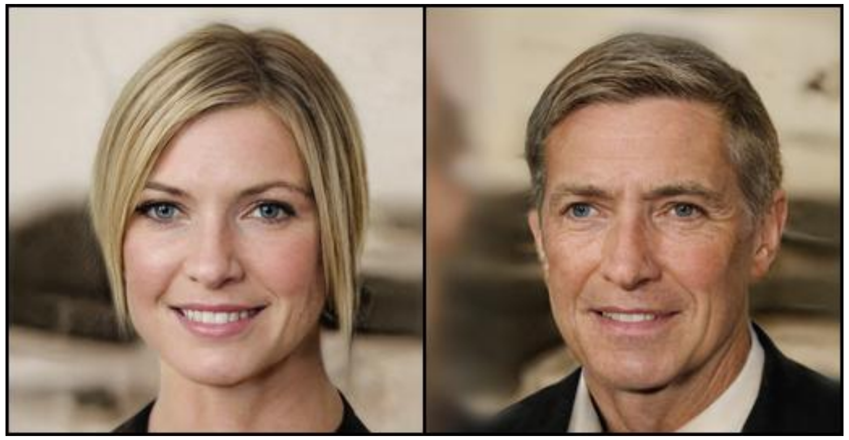
Multi-domain Image Generation & Translation with Identifiability Guarantees

- 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"

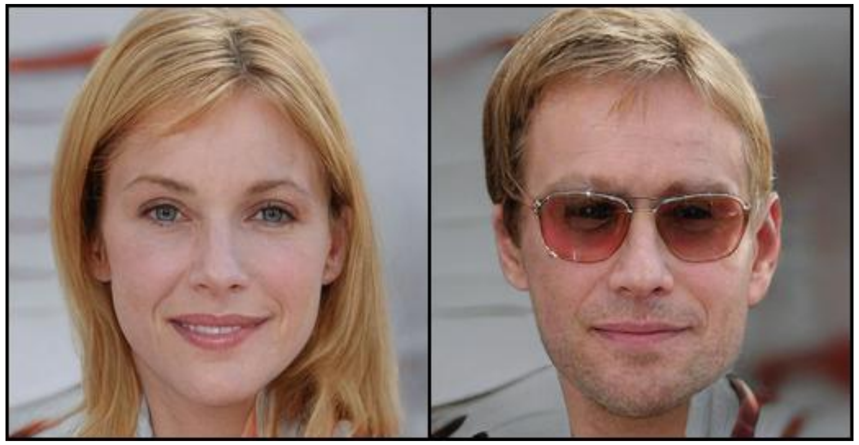
Ours (λ=0.1)



StyleGAN2-ADA

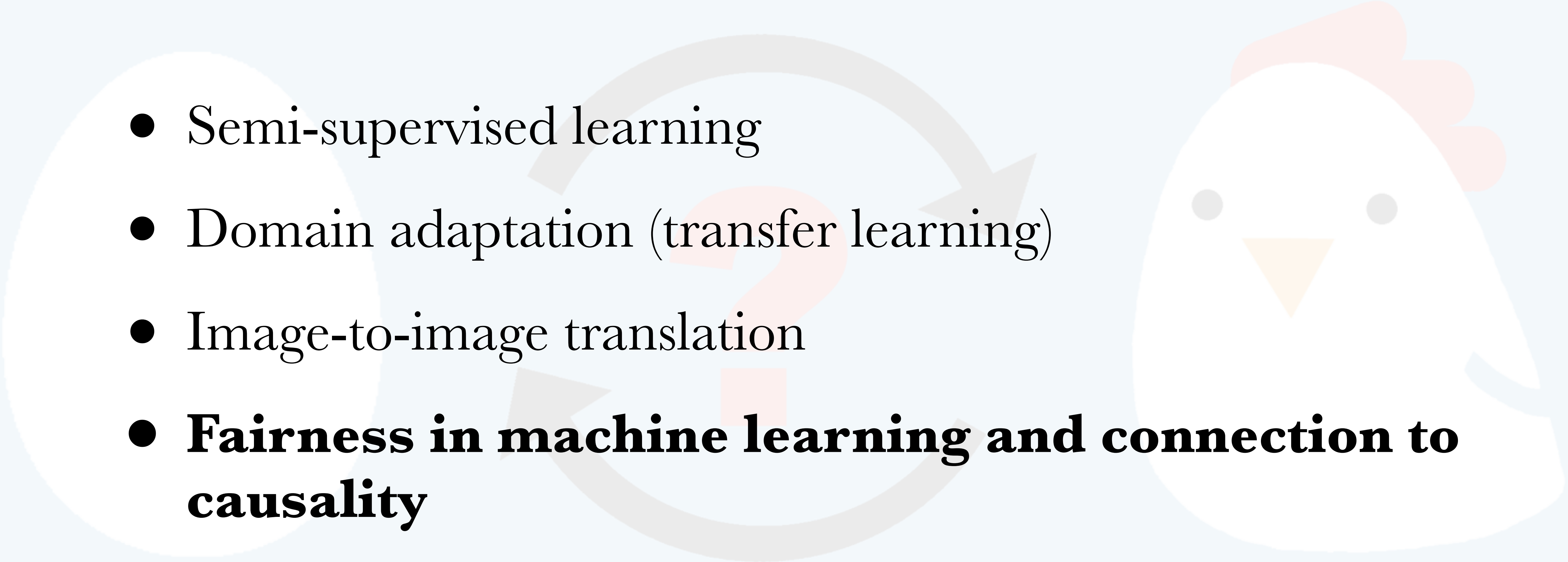


TGAN



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

Outline



- Semi-supervised learning
- Domain adaptation (transfer learning)
- Image-to-image translation
- **Fairness in machine learning and connection to causality**

What-Is & How-To for ML Fairness

A Principled Connection between Causality and Responsible AI

Zeyu Tang (zeyutang@cmu.edu)

https://zeyu.one

Outline

MOTIVATING EXAMPLES

DEFINITION AND PROBLEM SETUP

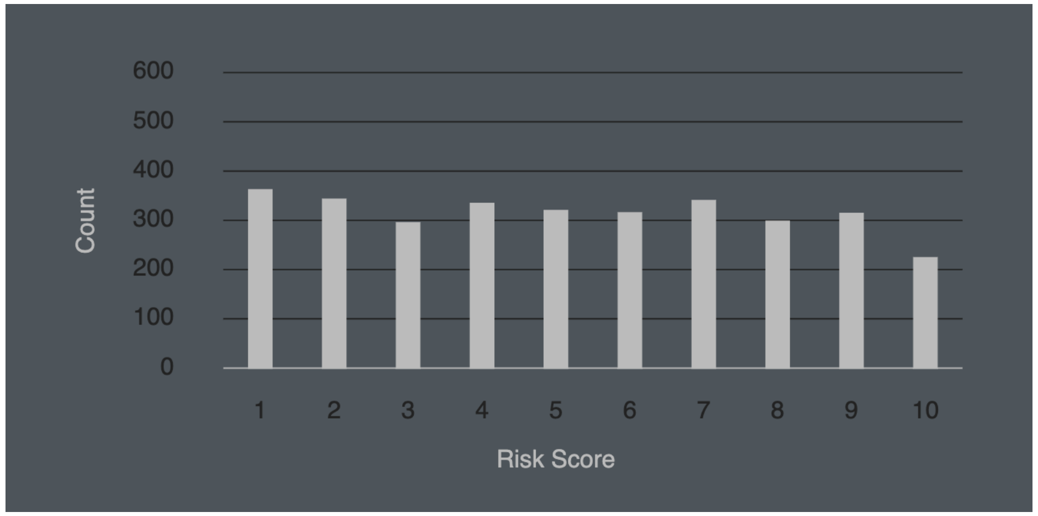
A QUICK INTRO TO FAIRNESS SPECTRA

CONCLUSION AND Q&A

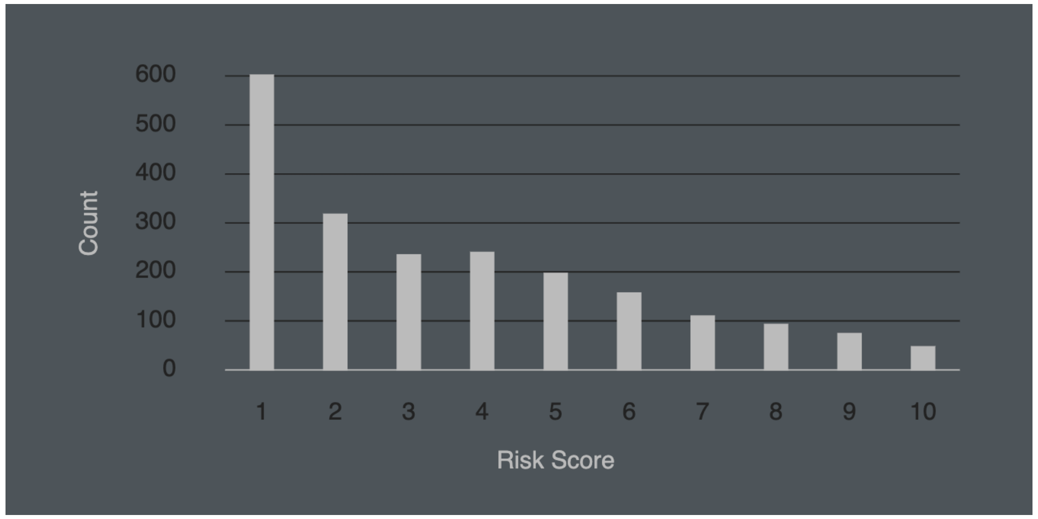
Example #1: COMPAS[1]

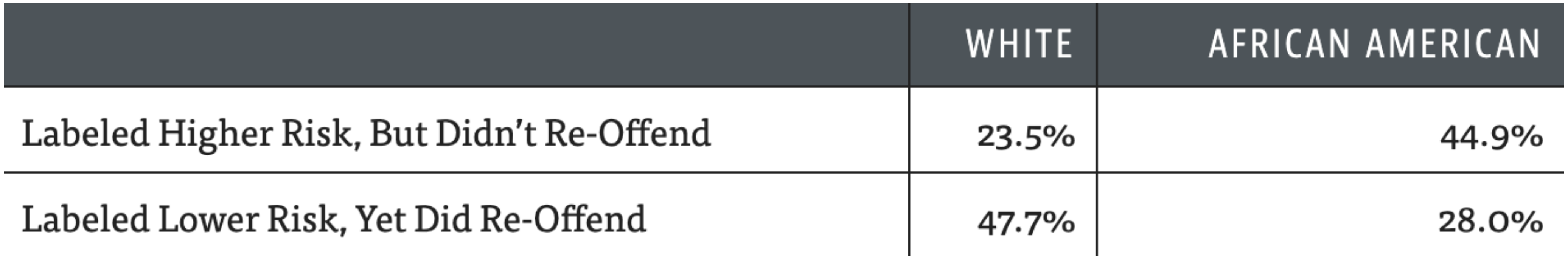
A software that predicts the risk of the recidivism of the defendant.

African-American defendants



White defendants





	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals, and it's biased against blacks. *ProPublica*, 2016.

Example #2: Loan application

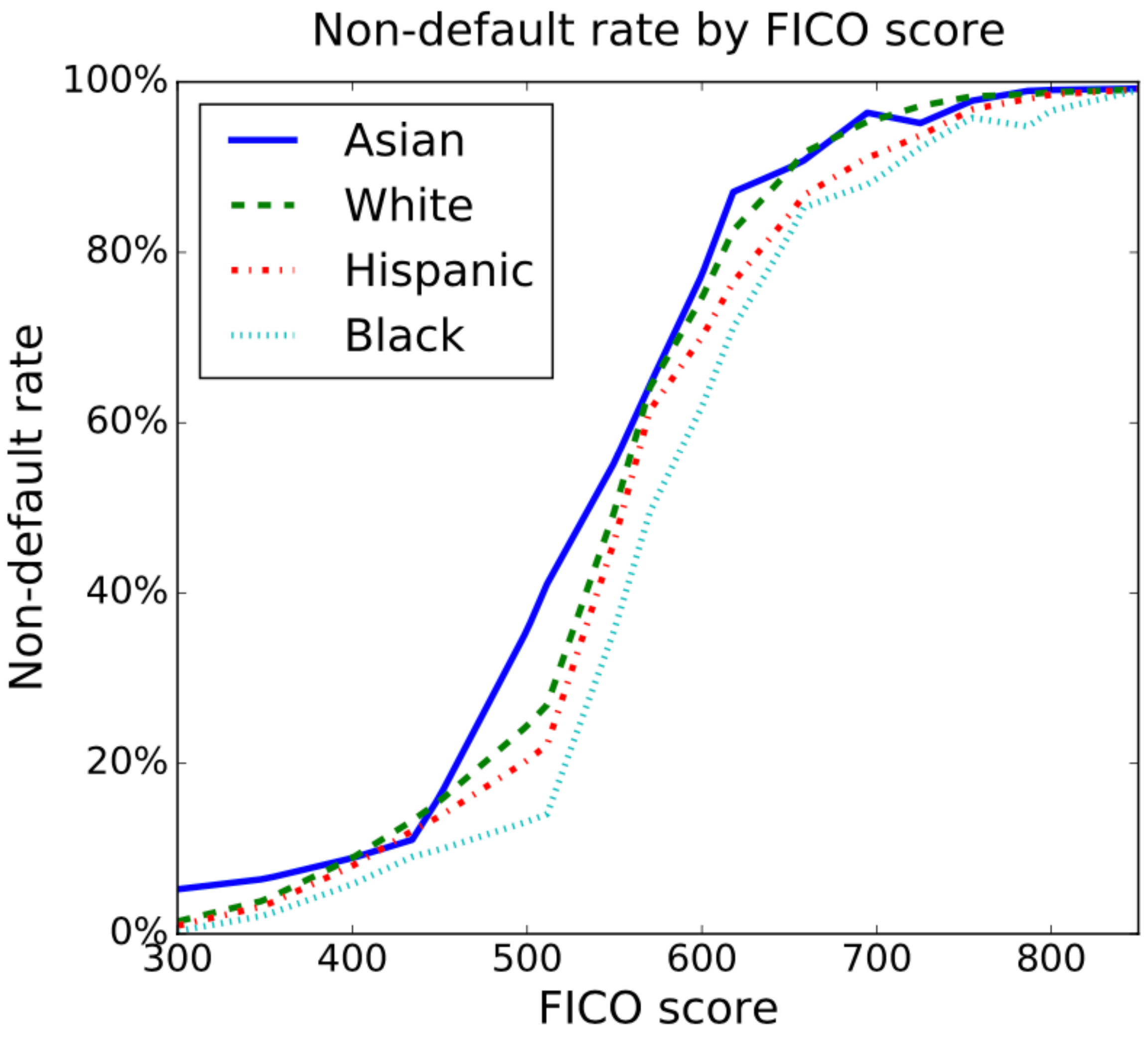
Look for credit

Check FICO score

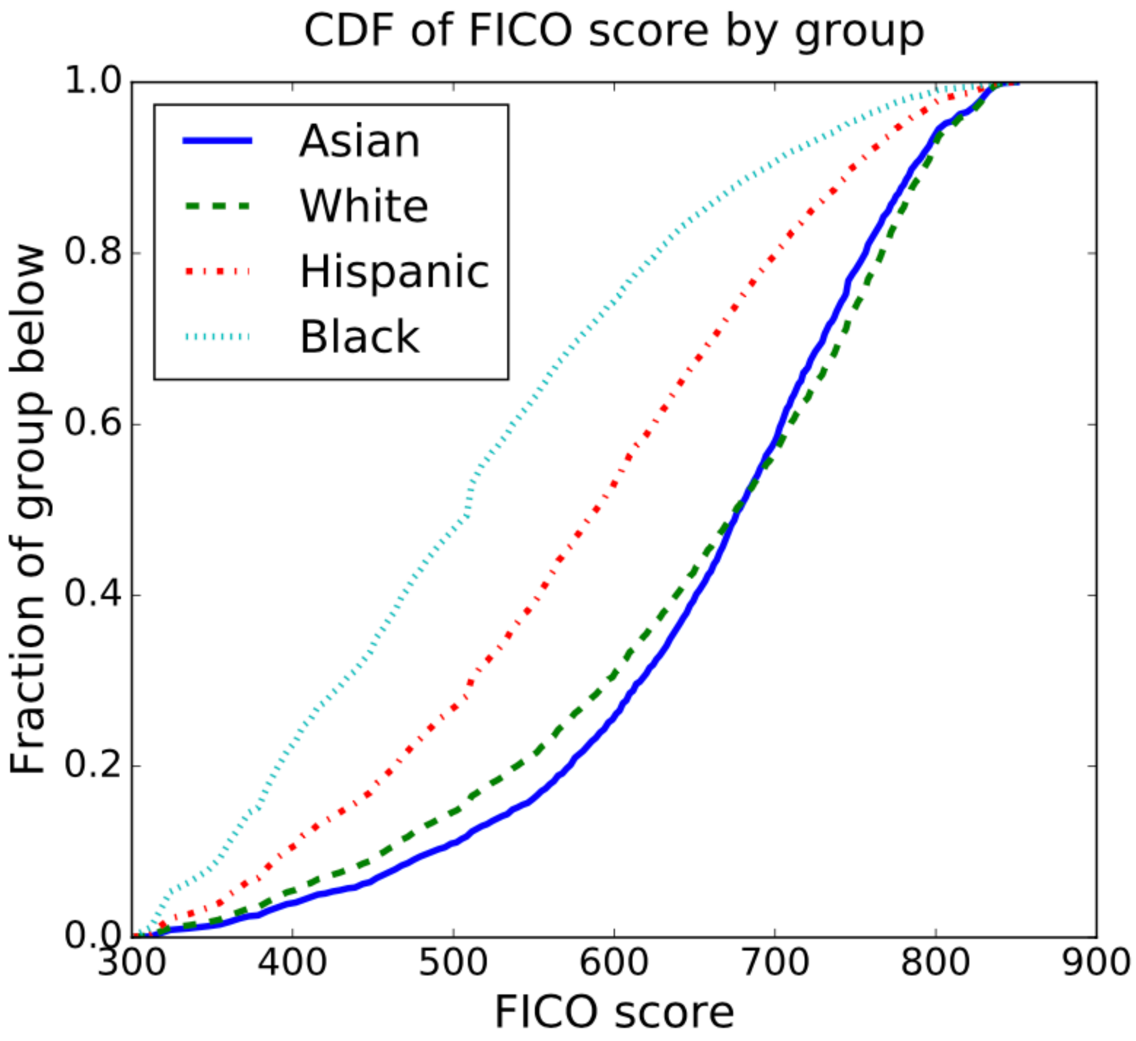
Approve / Reject

Example #2: Loan application (continued)

TransUnion FICO scores (2003) of more than 300k individuals.



Non-default rate by FICO score



CDF of FICO score by group

[1] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

Protected features

Race & color	*Civil Rights Act of 1964*
Gender	*Equal Pay Act of 1963; Civil Rights Act of 1964*
Religion	*Civil Rights Act of 1964*
National origin	*Civil Rights Act of 1964*
Age	*Age Discrimination in Employment Act of 1967*
Disability status	*Rehabilitation Act of 1973; Americans with Disabilities Act of 1990*
Veteran status	*Uniformed Employment and Reemployment Rights Act*
Genetic information	*Genetic Information Nondiscrimination Act*

Beyond intuition (notions & settings)

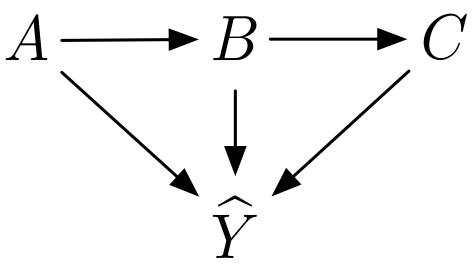
- Group level fairness notions
  - *Demographic Parity* (Calders et al., 2009)
  - *Equal Opportunity* (Hardt et al., 2016)
  - *Equalized Odds* (Hardt et al., 2016)
  - *Error-rate Balance* (Chouldechova, 2017)
  - *Predictive Rate Parity* (Zafar et al., 2017)

- Individual level fairness notions
  - *Fairness Through Awareness* (Dwork et al. 2011)

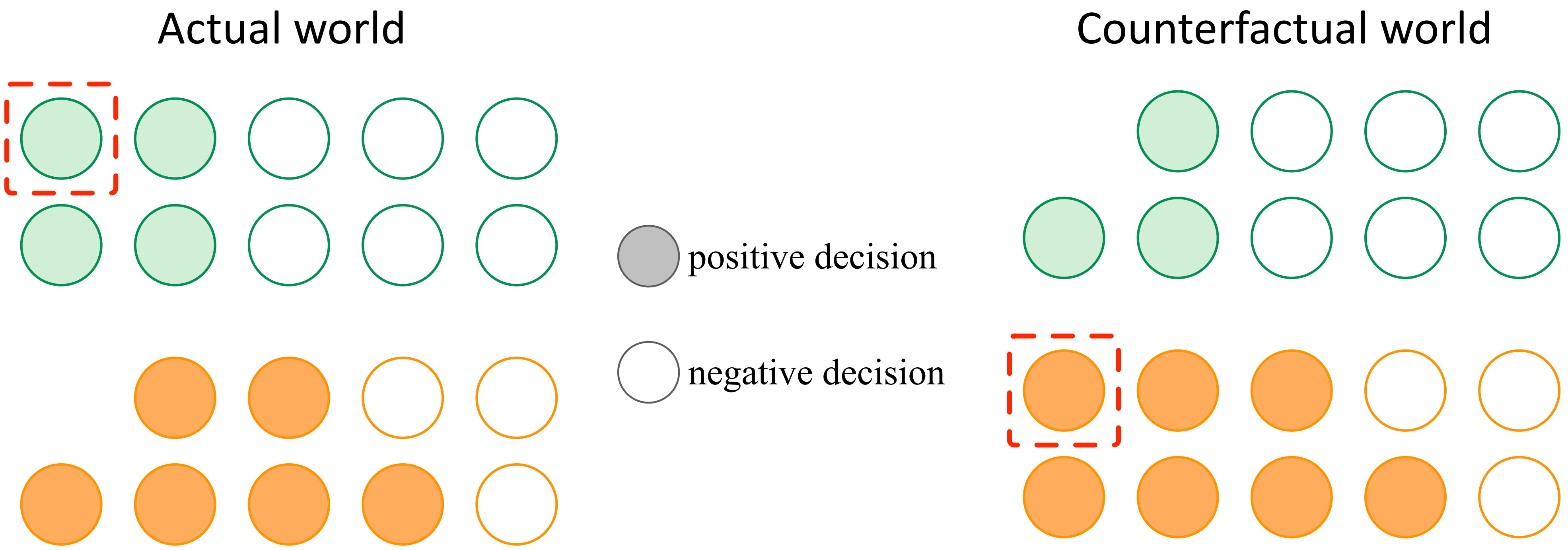
Beyond intuition (notions & settings)

- Fairness notions based on estimating/bounding **causal effects**
  - *Counterfactual Fairness* (Kusner et al., 2017)
  - *Fair Inference on Outcomes* (Nabi & Shpitser, 2018)
  - *Path-Specific Counterfactual Fairness* (Chiappa, 2019)
  - *PC-fairness* (Wu et al., 2019)
  - *Probability of Individual Unfairness* (Chikahara et al., 2020)

Examples of fairness notions (CF)



- *Counterfactual Fairness* (Kusner et al., 2017)
  - The prediction should be the same in following two worlds:
    (a) the actual world
    (b) a counterfactual world where the individual belonged to a different group

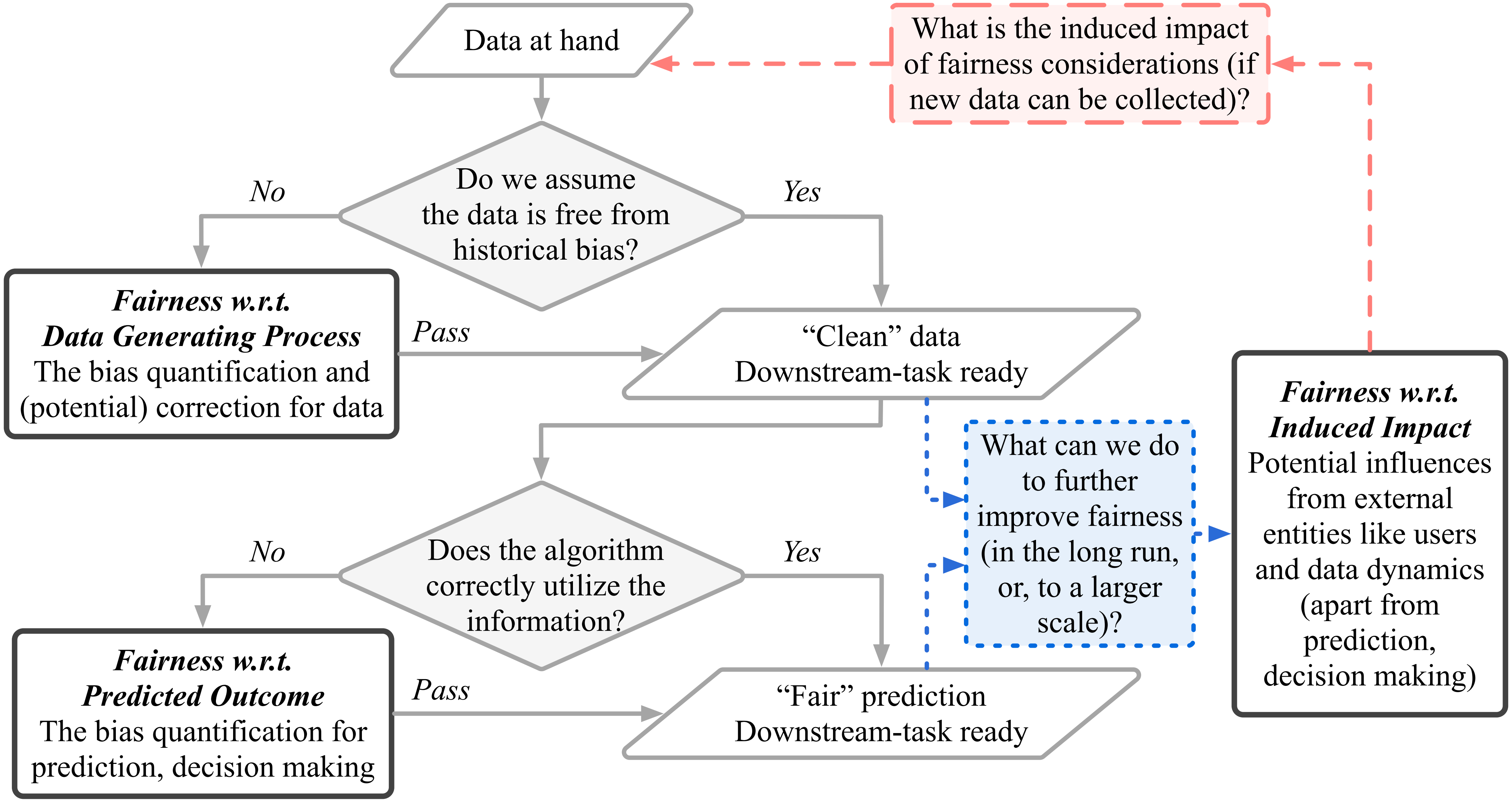

Actual world

Counterfactual world

Beyond intuition (notions & settings)

- Fairness in various settings
  - Dynamical setting
    - “Delayed Impact of Fair Machine Learning” (Liu at al., 2018)
    - “How do Fair Decisions Fare in Long-Term Qualifications?” (Zhang et al., 2020)
    - “Tier Balancing: Towards Dynamic Fairness over Underlying Causal Factors” (Tang et al., 2023)
  - Welfare consideration
    - “A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices” (Speicher et al., 2018)
    - “Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making” (Heidari et al., 2018)
    - “Allocating Opportunities in a Dynamic Model of Intergenerational Mobility” (Heidari & Kleinberg, 2021)

The fairness flowchart



Data at hand

Do we assume the data is free from historical bias?

No

Yes

Fairness w.r.t. Data Generating Process
The bias quantification and (potential) correction for data

Pass

"Clean" data
Downstream-task ready

What is the induced impact of fairness considerations (if new data can be collected)?

Does the algorithm correctly utilize the information?

No

Yes

Fairness w.r.t. Predicted Outcome
The bias quantification for prediction, decision making

Pass

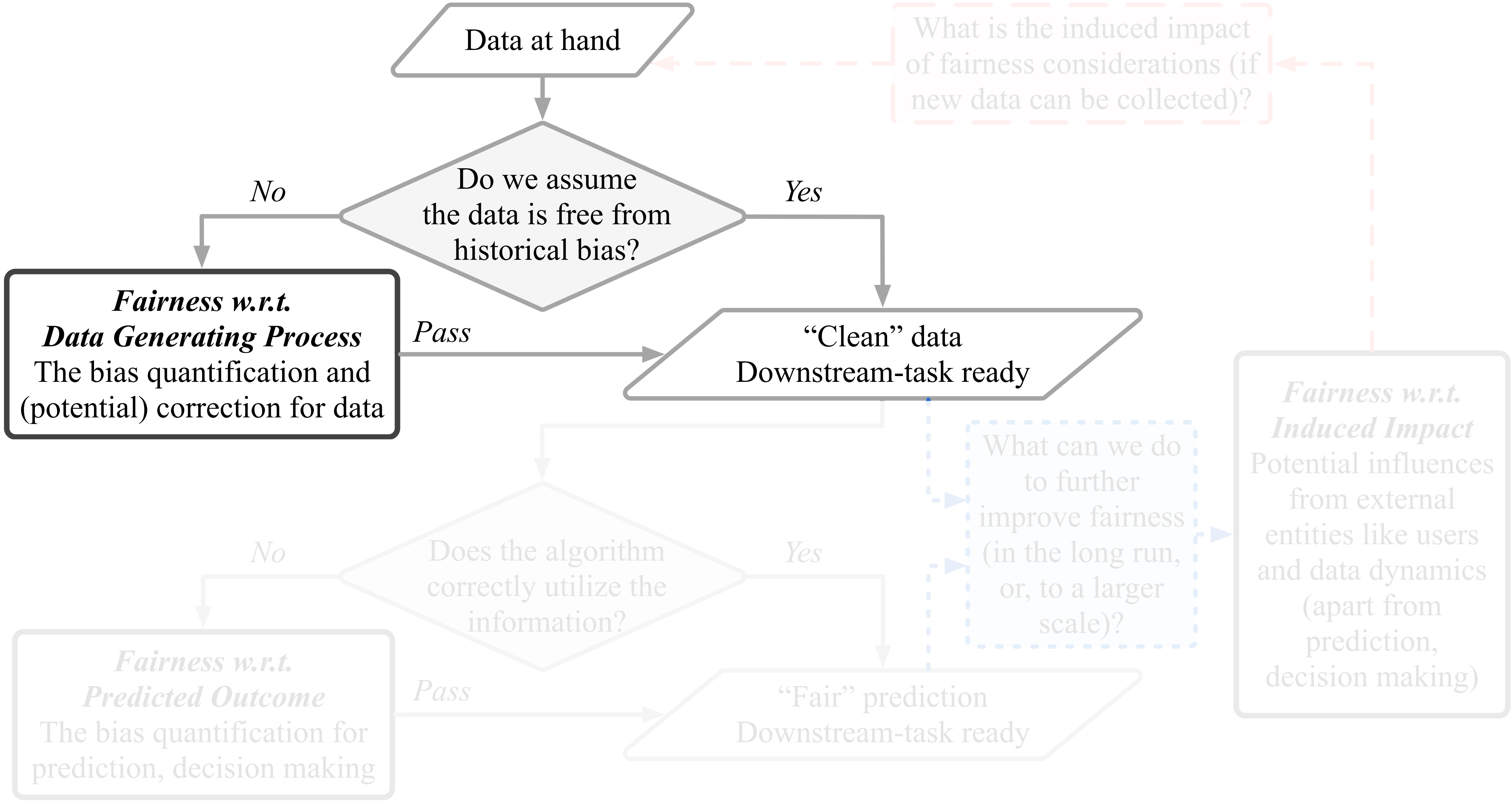
"Fair" prediction
Downstream-task ready

What can we do to further improve fairness (in the long run, or, to a larger scale)?

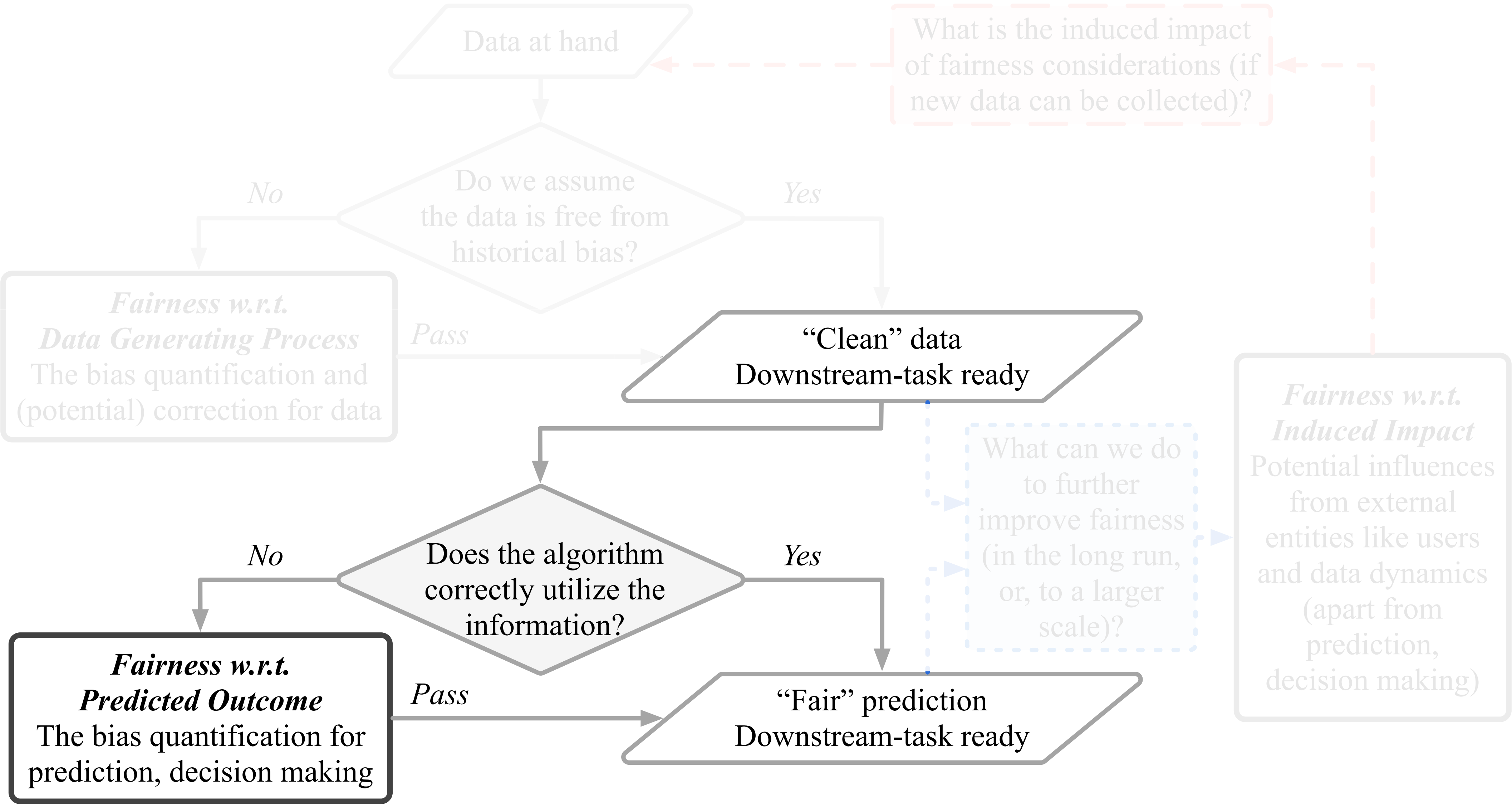
Fairness w.r.t. Induced Impact
Potential influences from external entities like users and data dynamics (apart from prediction, decision making)

[1] Zeyu Tang, Jiji Zhang, and Kun Zhang. What-Is and How-To for Fairness in Machine Learning: A Survey, Reflection, and Perspective. In *ACM Computing Surveys.* 2023.

Spectrum - w.r.t. data generating process

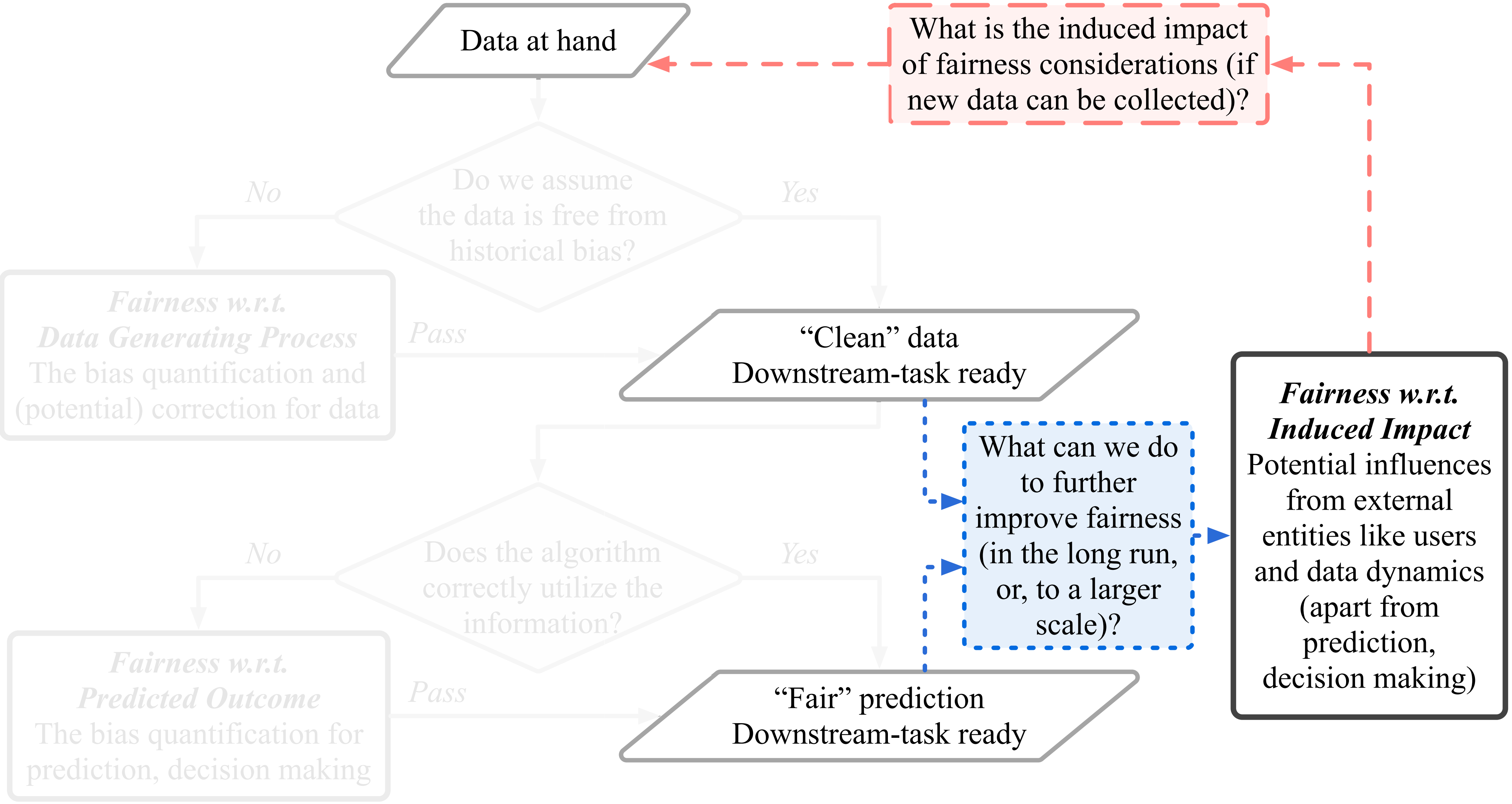

Data at hand
What is the induced impact of fairness considerations (if new data can be collected)?
No
Do we assume the data is free from historical bias?
Yes
Fairness w.r.t. Data Generating Process
The bias quantification and (potential) correction for data
Pass
"Clean" data
Downstream-task ready
What can we do to further improve fairness (in the long run, or, to a larger scale)?
Fairness w.r.t. Induced Impact
Potential influences from external entities like users and data dynamics (apart from prediction, decision making)
No
Does the algorithm correctly utilize the information?
Yes
Fairness w.r.t. Predicted Outcome
The bias quantification for prediction, decision making
Pass
"Fair" prediction
Downstream-task ready

Spectrum - w.r.t. predicted outcome

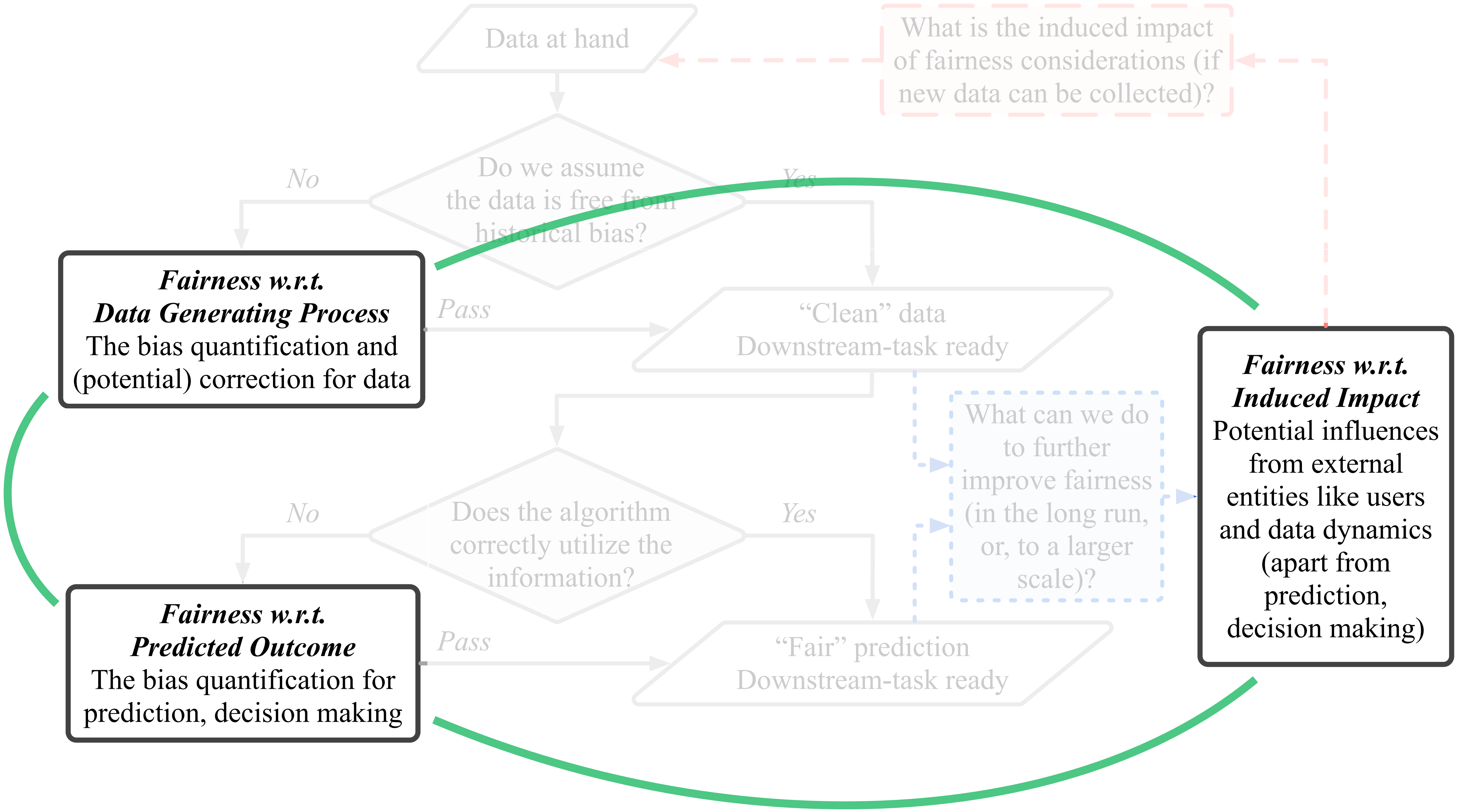

Data at hand
What is the induced impact of fairness considerations (if new data can be collected)?
No
Do we assume the data is free from historical bias?
Yes
Fairness w.r.t. Data Generating Process
The bias quantification and (potential) correction for data
Pass
"Clean" data
Downstream-task ready
What can we do to further improve fairness (in the long run, or, to a larger scale)?
Fairness w.r.t. Induced Impact
Potential influences from external entities like users and data dynamics (apart from prediction, decision making)
No
Does the algorithm correctly utilize the information?
Yes
Fairness w.r.t. Predicted Outcome
The bias quantification for prediction, decision making
Pass
"Fair" prediction
Downstream-task ready

13

Spectrum - w.r.t. induced impact


Data at hand
What is the induced impact of fairness considerations (if new data can be collected)?
No
Do we assume the data is free from historical bias?
Yes
Fairness w.r.t. Data Generating Process
The bias quantification and (potential) correction for data
Pass
“Clean” data Downstream-task ready
Fairness w.r.t. Induced Impact
Potential influences from external entities like users and data dynamics (apart from prediction, decision making)
What can we do to further improve fairness (in the long run, or, to a larger scale)?
No
Does the algorithm correctly utilize the information?
Yes
Fairness w.r.t. Predicted Outcome
The bias quantification for prediction, decision making
Pass
“Fair” prediction Downstream-task ready

Why three spectra?



Data at hand

What is the induced impact of fairness considerations (if new data can be collected)?

Do we assume the data is free from historical bias?

No

Yes

Fairness w.r.t. Data Generating Process
The bias quantification and (potential) correction for data

Pass

"Clean" data
Downstream-task ready

What can we do to further improve fairness (in the long run, or, to a larger scale)?

Does the algorithm correctly utilize the information?

No

Yes

Fairness w.r.t. Predicted Outcome
The bias quantification for prediction, decision making

Pass

"Fair" prediction
Downstream-task ready

Fairness w.r.t. Induced Impact
Potential influences from external entities like users and data dynamics (apart from prediction, decision making)

15

References

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS Tutorial*, 2017.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18. IEEE, 2009.

Silvia Chiappa. Path-specific counterfactual fairness. In Proceedings of the *AAAI Conference on Artificial Intelligence*, volume 33, pp. 7801–7808, 2019.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

References

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pp. 3315–3323, 2016.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*. PMLR, 3150–3158, 2018.

Zeyu Tang, Yatong Chen, Yang Liu, and Kun Zhang. Tier Balancing: Towards Dynamic Fairness over Underlying Causal Factors. In *International Conference on Learning Representations*, 2023.

Zeyu Tang, Jiji Zhang, and Kun Zhang. What-Is and How-To for Fairness in Machine Learning: A Survey, Reflection, and Perspective. In *ACM Computing Surveys,* 2023.

Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. Pc-fairness: A unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, pp. 3399–3409, 2019.

Thank you!

Summary

- Causality matters
- "Simplicity" helps in causal discovery & causal representation learning:
  - Conditional independence: constraint-based approach
  - Cause $\perp\!\!\!\perp$ noise in constrained FCMs ⇒ causal asymmetry
  - Independent changes in P(cause) and P(effect | cause)
  - Other types of "simplicity": rank deficiency…
- ML based on causality-related representation
  - Compact description of changes
  - Property behind data
- Latent variables & their relations involved in changing influences are generally identifiable