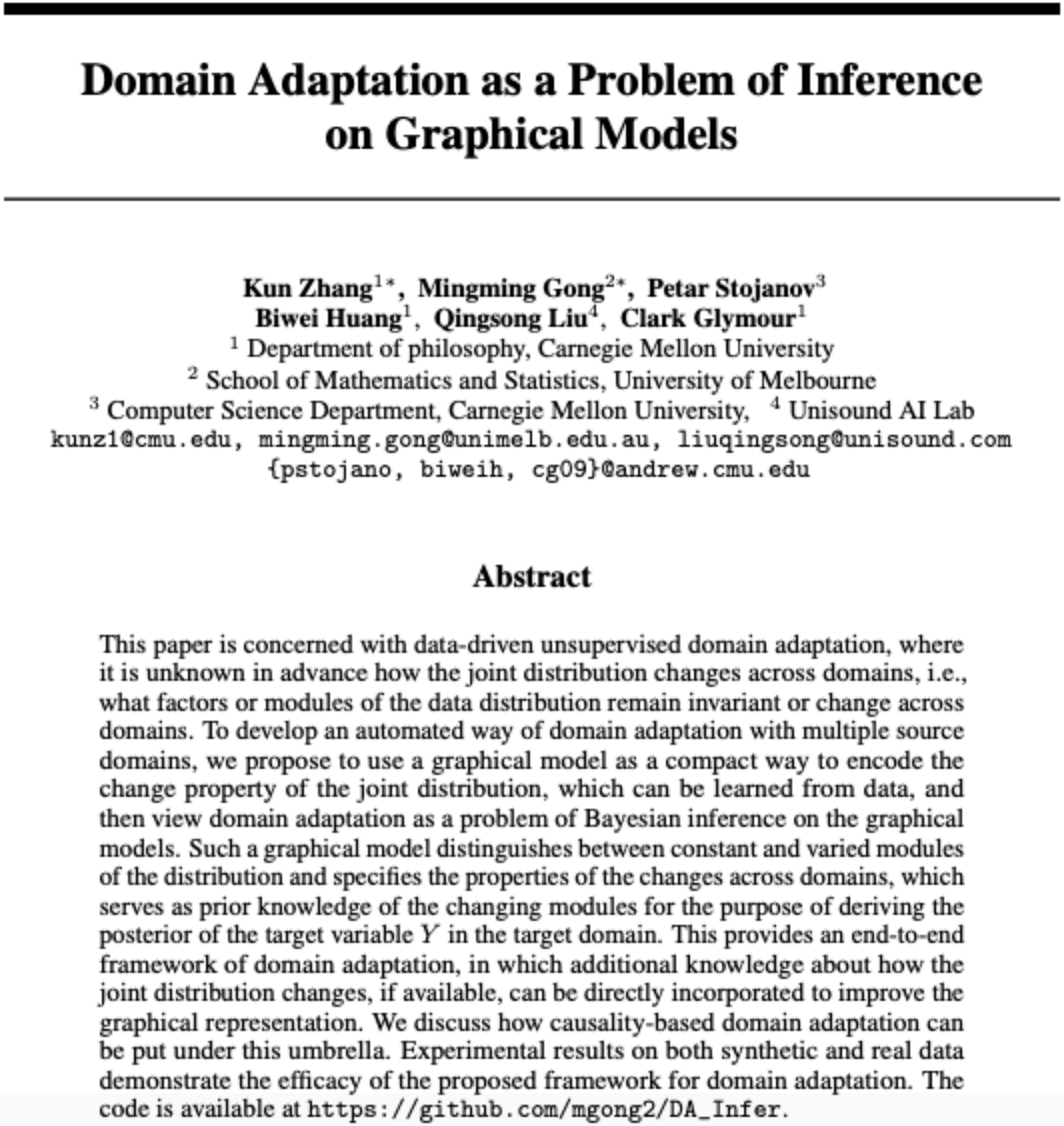
Domain Adaptation as A Problem of Inference on Graphical Models

Mingming Gong

POSTERA CRESCAM LAUDE
THE UNIVERSITY OF
MELBOURNE

青源Seminar, Nov 27, 2020



Domain Adaptation as a Problem of Inference on Graphical Models

Kun Zhang[1*], **Mingming Gong**[2*], **Petar Stojanov**[3]
Biwei Huang[1], **Qingsong Liu**[4], **Clark Glymour**[1]
[1] Department of philosophy, Carnegie Mellon University
[2] School of Mathematics and Statistics, University of Melbourne
[3] Computer Science Department, Carnegie Mellon University, [4] Unisound AI Lab
kunz1@cmu.edu, mingming.gong@unimelb.edu.au, liuqingsong@unisound.com
{pstojano, biweih, cg09}@andrew.cmu.edu

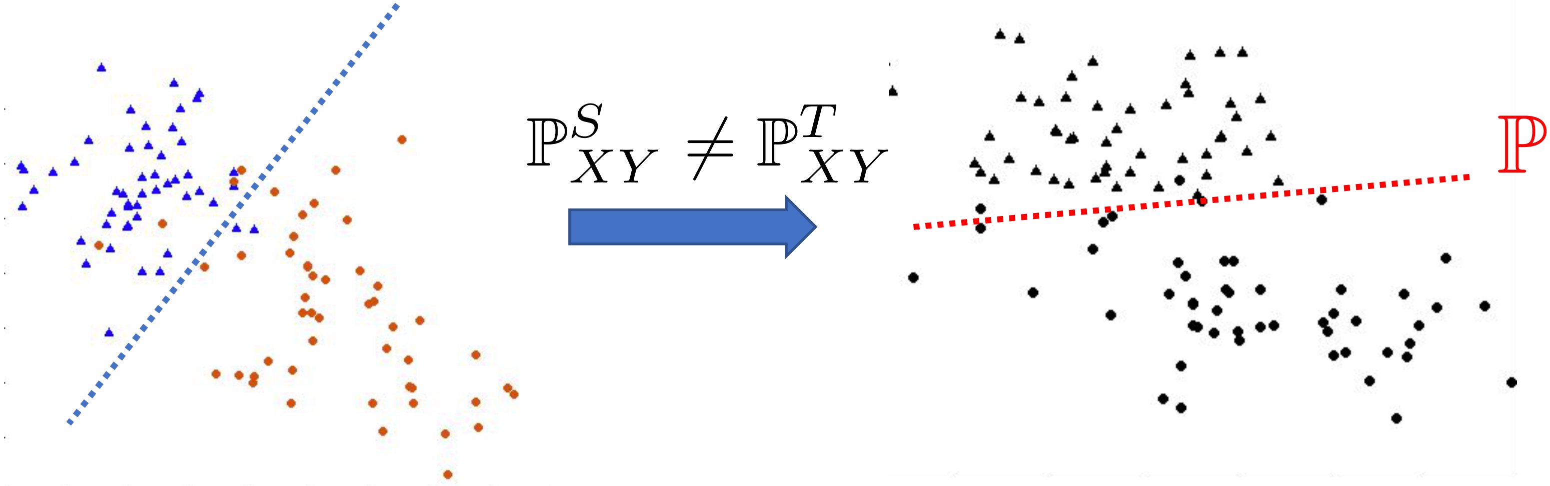
Abstract

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. The code is available at `https://github.com/mgong2/DA_Infer`.

Domain Adaptation

X – feature (covariate)

Y – label (target)



$\mathbb{P}^S_{XY} \neq \mathbb{P}^T_{XY}$

$\mathbb{P}^T_{Y|X}$?

$\mathbb{P}^S_{XY}$

Source Domains(Training)

$\mathbb{P}^T_{X}$

Target Domain(Test)

Covariate shift

$$\mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_{Y|X} \quad \mathbb{P}^S_X \neq \mathbb{P}^T_X \quad \mathbb{P}^S_{Y|X} = \mathbb{P}^T_{Y|X}$$

Instance reweighting

$$R = \int \mathbb{P}^T(x,y)\ell(x,y)dxdy$$

$$= \int \frac{\mathbb{P}^T_X(x)}{\mathbb{P}^S_X(x)} \mathbb{P}^S(x,y)\ell(x,y)dxdy$$

Invariant Feature Learning

$$\text{find} \quad X' = h(X),$$

$$\text{such that} \quad \mathbb{P}^S_{X'} = \mathbb{P}^T_{X'}$$

Maximum Mean Discrepancy (MMD), optimal transport, adversarial loss…

- Why $\mathbb{P}^S_X \neq \mathbb{P}^T_X \quad \mathbb{P}^S_{Y|X} = \mathbb{P}^T_{Y|X}$?
- What if $\mathbb{P}^S_{Y|X} \neq \mathbb{P}^T_{Y|X}$?

Shimodaira, Hidetoshi. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference. 2000.

Modularity

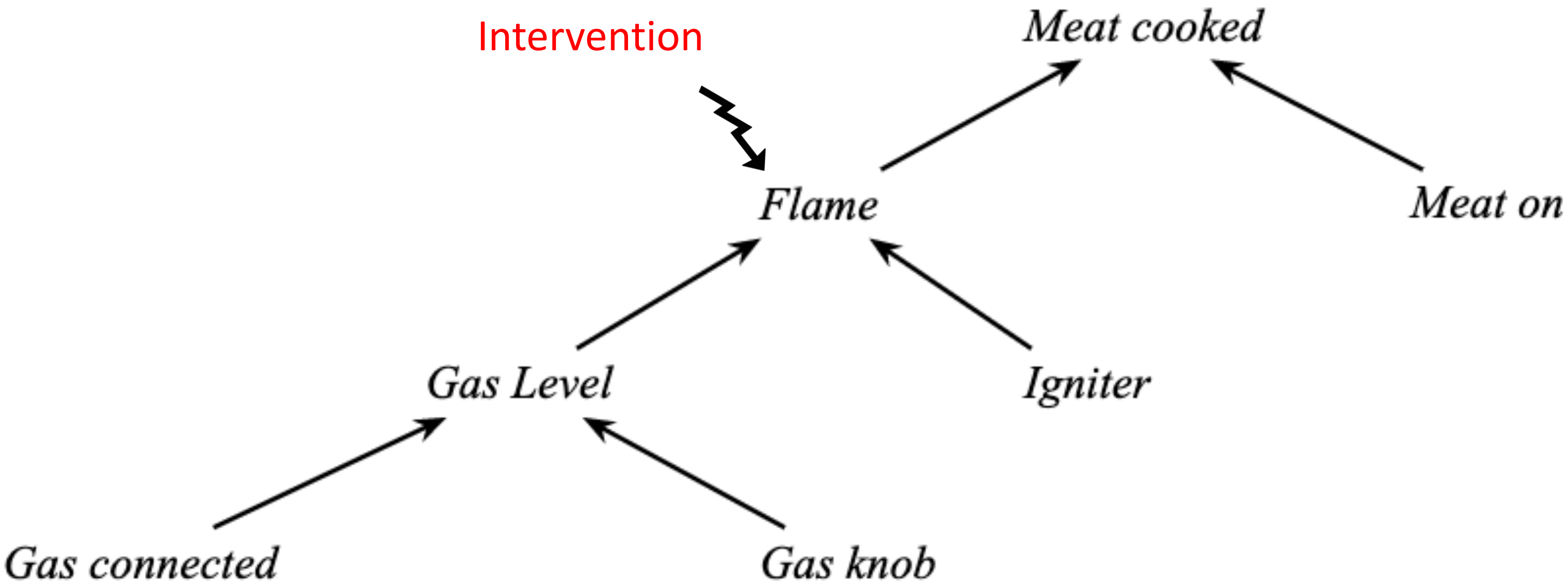
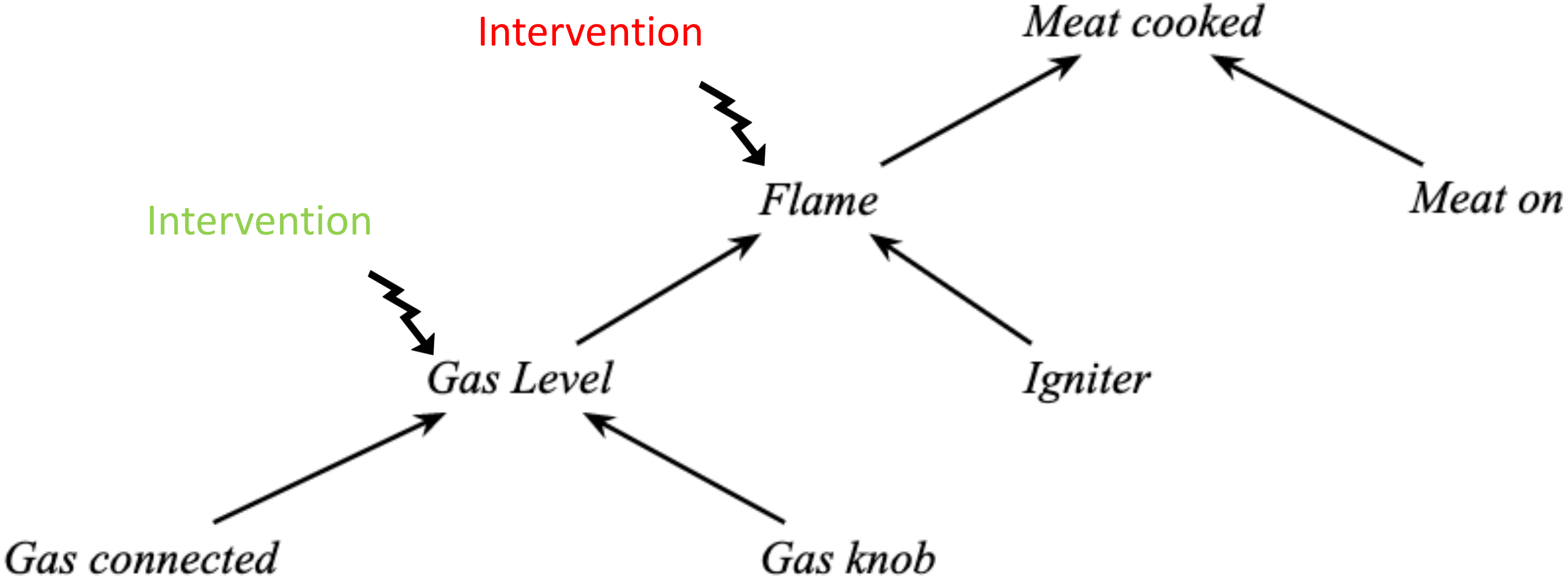


FIGURE 3

Soft intervention:
P(Flame|Gas Level, Igniter) changes
P(Gas Level|Gas Connected, Gas knob) is invariant
P(Meat cooked|Flame, Meat on) is invariant

https://plato.stanford.edu/entries/causal-models/

Modularity



Intervention

Intervention

FIGURE 3

Soft intervention:

P(Flame|Gas Level, Igniter) changes

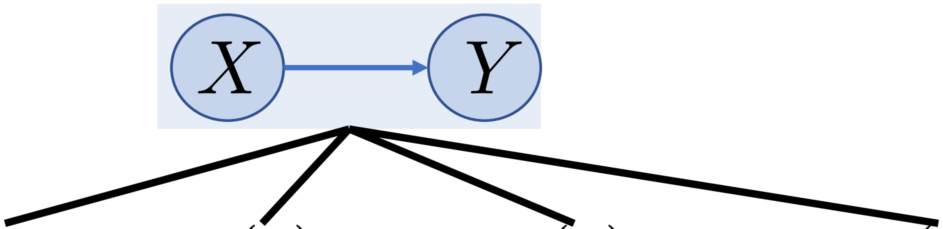
P(Gas Level|Gas Connected, Gas knob) is changes

P(Meat cooked|Flame, Meat on) is invariant

https://plato.stanford.edu/entries/causal-models/

Causal Model for DA

- Bridge between probability distributions

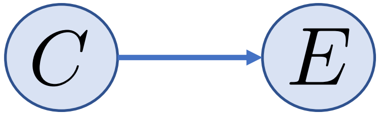


$$\mathbb{P}^{(1)}_{XY} \quad \mathbb{P}^{(2)}_{XY} \quad \mathbb{P}^{(3)}_{XY} \ldots \quad \mathbb{P}^{(k)}_{XY}$$

- Independent causal mechanism

$C - Cause$

$E - Effect$



$\mathbb{P}_C$ $\mathbb{P}_{E|C}$ (causal mechanism)

Without confounder, $\mathbb{P}_C$ and $\mathbb{P}_{E|C}$ do not contain information about each other.

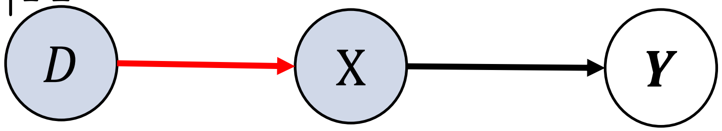
Schölkopf, Bernhard, et al. "On causal and anticausal learning." *arXiv preprint arXiv:1206.6471* (2012).

$X \to Y$

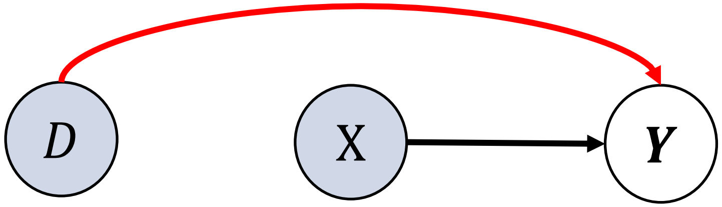
$$\mathbb{P}_{XY} = \mathbb{P}_X \mathbb{P}_{Y|X}$$

D – Domain Index

Covariate Shift

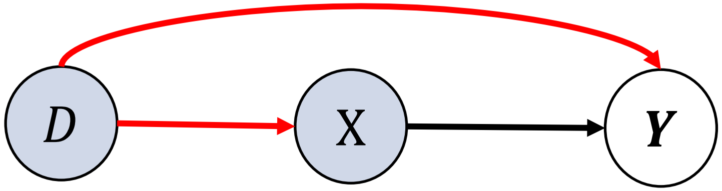


$\mathbb{P}^S_X \neq \mathbb{P}^T_X$ $\quad \mathbb{P}^S_{Y|X} = \mathbb{P}^T_{Y|X}$



$\mathbb{P}^S_X = \mathbb{P}^T_X$ $\quad \mathbb{P}^S_{Y|X} \neq \mathbb{P}^T_{Y|X}$

No clue as to find $\mathbb{P}^T_{Y|X}$ with one source domain



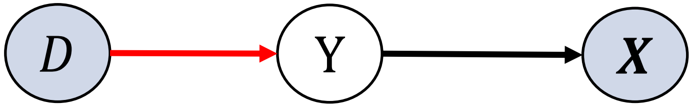
$\mathbb{P}^S_X \neq \mathbb{P}^T_X$ $\quad \mathbb{P}^S_{Y|X} \neq \mathbb{P}^T_{Y|X}$

$Y \to X$

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

$$\mathbb{P}_{XY} = \mathbb{P}_{X|Y}\mathbb{P}_{Y}$$

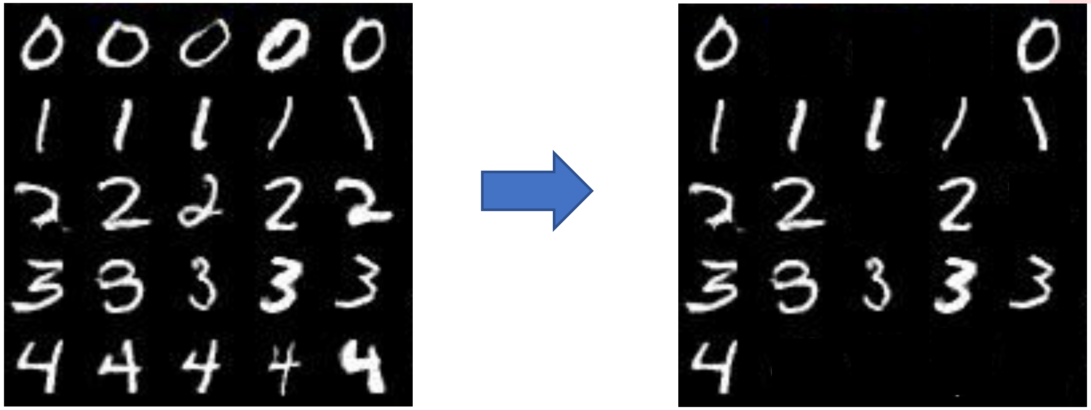
Target Shift



$$\mathbb{P}^{S}_{Y} \neq \mathbb{P}^{T}_{Y} \qquad \mathbb{P}^{S}_{X|Y} = \mathbb{P}^{T}_{X|Y}$$

$$\mathbb{P}^{S}_{X} \neq \mathbb{P}^{T}_{X}$$

$$\mathbb{P}^{S}_{Y|X} \neq \mathbb{P}^{T}_{Y|X}$$



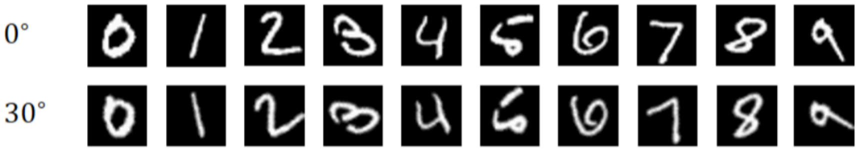
Covariate shift does *not* hold !!!

Zhang, Kun, et al. Domain adaptation under target and conditional shift. ICML13

Gong, Zhang, et al. Domain adaptation with conditional transferable components. ICML16

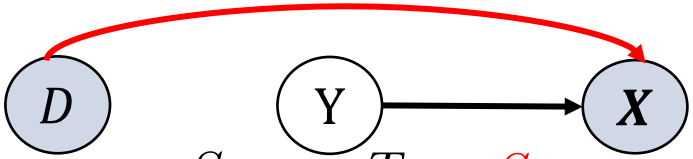
$Y \to X$

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



$$\mathbb{P}_{XY} = \mathbb{P}_{X|Y}\mathbb{P}_Y$$

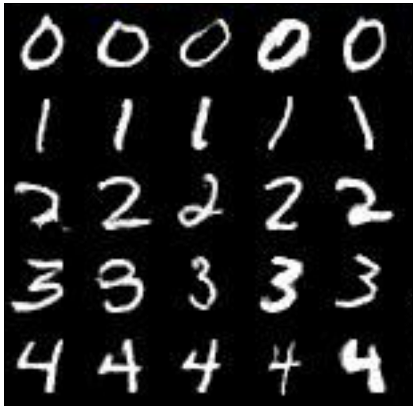
Conditional Shift



$$\mathbb{P}_Y^S = \mathbb{P}_Y^T \quad \mathbb{P}_{X|Y}^S \neq \mathbb{P}_{X|Y}^T$$

$$\mathbb{P}_X^S \neq \mathbb{P}_X^T$$

$$\mathbb{P}_{Y|X}^S \neq \mathbb{P}_{Y|X}^T$$



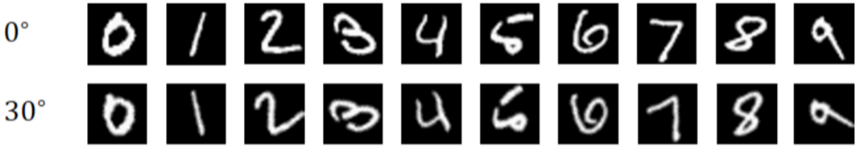
Covariate shift does *not* hold !!!

Gong, Zhang, et al. Domain adaptation with conditional transferable components. ICML16

Zhang, Kun, et al. Domain adaptation under target and conditional shift. ICML13

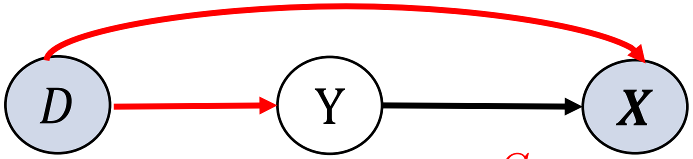
$Y \to X$

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



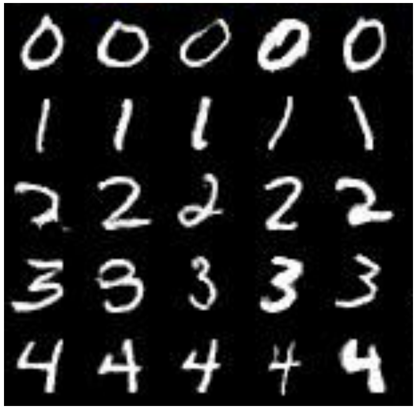
$$\mathbb{P}_{XY} = \mathbb{P}_{X|Y}\mathbb{P}_Y$$

Generalized Target Shift



$\mathbb{P}_Y^S \neq \mathbb{P}_Y^T$ $\mathbb{P}_{X|Y}^S \neq \mathbb{P}_{X|Y}^T$

$$\mathbb{P}_X^S \neq \mathbb{P}_X^T$$
$$\mathbb{P}_{Y|X}^S \neq \mathbb{P}_{Y|X}^T$$



Covariate shift does *not* hold !!!

Gong, Zhang, et al. Domain adaptation with conditional transferable components. ICML16

Zhang, Kun, et al. Domain adaptation under target and conditional shift. ICML13

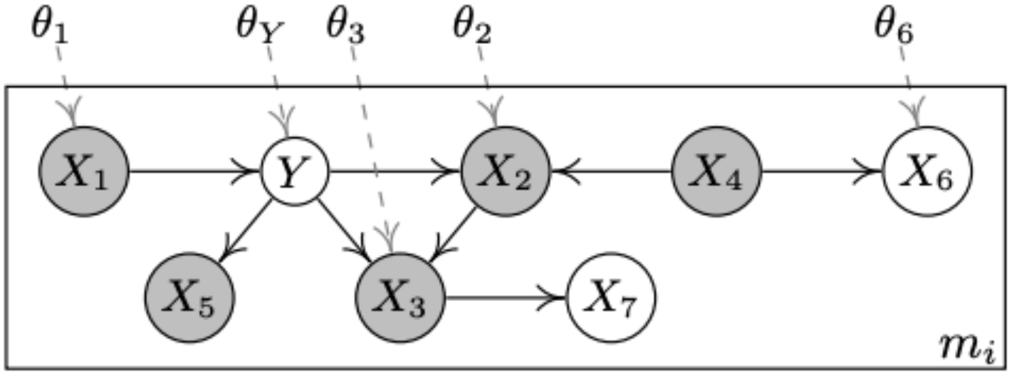
Problems

- The causal graph and the invariance/changing causal modules are assumed to be known

- The algorithms do not make full use of the causal generative process

- Learning causal graphs from observational data is a hard.

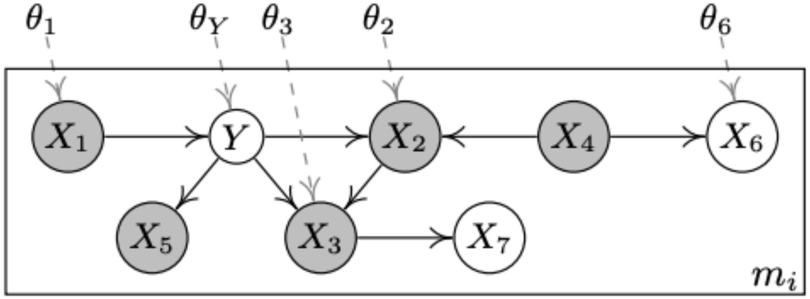
Inference on Graphical Models



- Automated way to model change and invariance properties in the joint distribution
  - Factorize the joint distribution according to an augmented directed acyclic graph (DAG)
  - Formulate domain adaptation as a Bayesian inference problem on the augmented DAG

Zhang, K*., Gong, M.*, et al. (2020). Domain Adaptation As a Problem of Inference on Graphical Models. *NeurIPS 2020.*

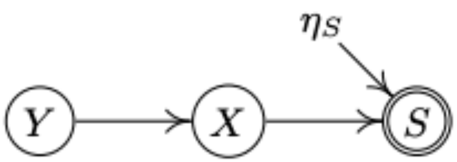
Augmented DAG



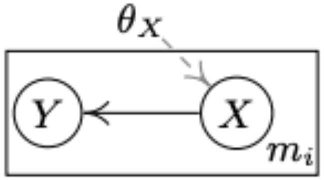
- DAG encodes conditional independence relations
- Encode distribution change by augmenting DAG with $\boldsymbol{\theta}$
  - θ_i are independent – independent change
  - θ follows a prior distribution $P(\boldsymbol{\theta})$
- Data generating process
  - Generate $\boldsymbol{\theta}^{(i)}$ from $P(\boldsymbol{\theta})$
  - Given $\boldsymbol{\theta}^{(i)}$, sample data from the distribution in the i-th domain:

$$P(\mathbf{X}, Y|\boldsymbol{\theta}^{(i)}) = P(X_1|\theta_1^{(i)})P(Y|X_1, \theta_Y^{(i)})P(X_5|Y)P(X_2|Y, X_4, \theta_2^{(i)})P(X_3|Y, X_2, \theta_3^{(i)})\times$$
$$P(X_4)P(X_6|X_4, \theta_6^{(i)})P(X_7|X_3).$$

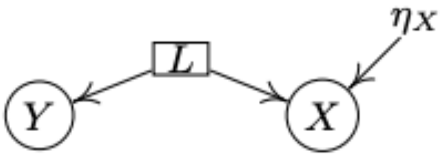
Relation to Causal Graphs



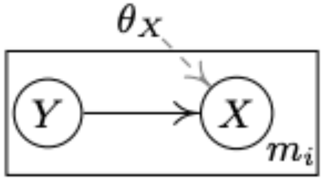
(a) The underlying data generating process of Example 1. Y generates (causes) X, and S denotes the selection variable (a data point is included if and only if $S = 1$).



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

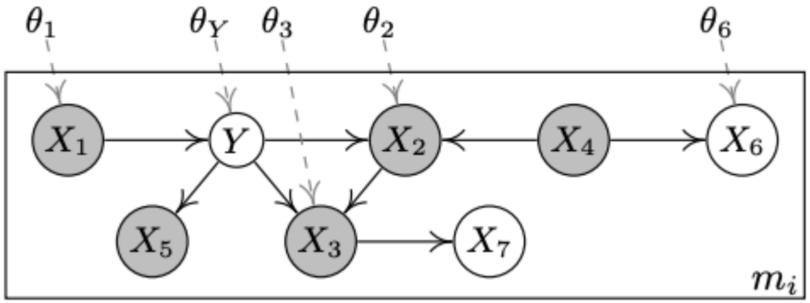


(c) The generating process of Example 2. L is a confounder; the mechanism of X changes across domains, as indicated by η_X.



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

Bayesian Inference

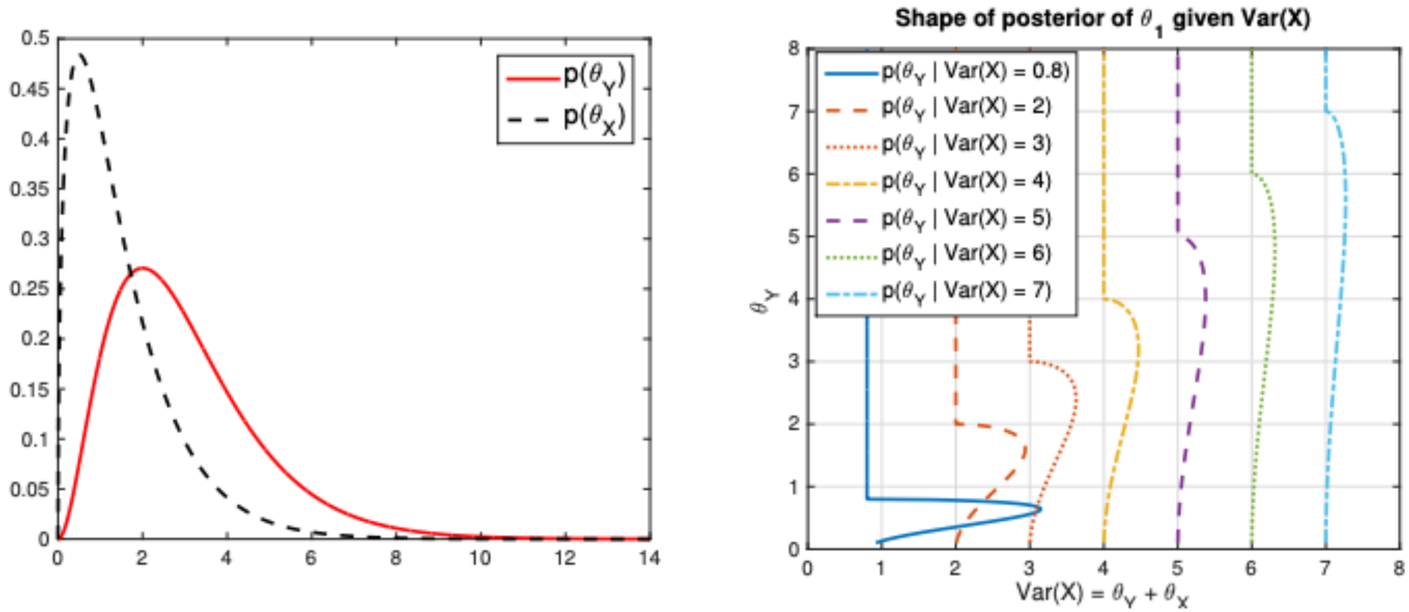


$$P(y_k^\tau \,|\, \mathbf{x}^\tau) = \int \underbrace{P(y_k^\tau \,|\, \mathbf{x}_k^\tau, \boldsymbol{\theta})}_{\text{Classifier}} \underbrace{\frac{\prod_k \left[\sum_{y_k^\tau} \prod_{V_j \in \mathbf{V}} \mathcal{C}_{jk}\right] \prod_{V_j \in \mathbf{V}} P(\theta_{V_j})}{\int \prod_k \left[\sum_{y_k^\tau} \prod_{V_j \in \mathbf{V}} \mathcal{C}_{jk}\right] \prod_{V_j \in \mathbf{V}} P(\theta_{V_j}) d\theta_{V_j}}}_{P(\boldsymbol{\theta}|\mathbf{x}^\tau)} d\boldsymbol{\theta}.$$

$$\mathcal{C}_{jk} := P(v_{jk}^\tau \,|\, \mathbb{PA}(v_{jk}^\tau), \theta_{V_j})$$

$$\mathbf{V} = \mathbb{CH}(Y) \cup \{Y\}$$

Benefits of Bayesian Treatment



(a) Prior distributions of $\boldsymbol{\theta}$

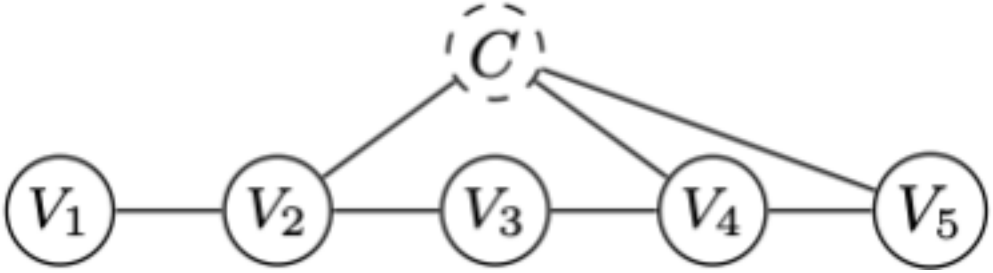
(b) Posterior of θ_Y given $\mathbb{V}\mathrm{ar}(X)$

$Y \sim N(0, \theta_Y)$ $X = Y + E$

$E \sim N(0, \theta_X)$ $X \sim N(0, \theta_X + \theta_Y)$

Graph Learning: Skeleton learning and changing module detection

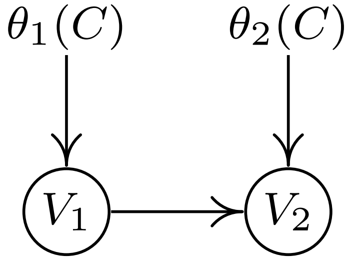
- Using Domain Index C as a surrogate variable and apply Constraint-based search on C and the observed features and labels.
    - Detecting Changing Causal Modules
    - Obtain the Skeleton of the graph

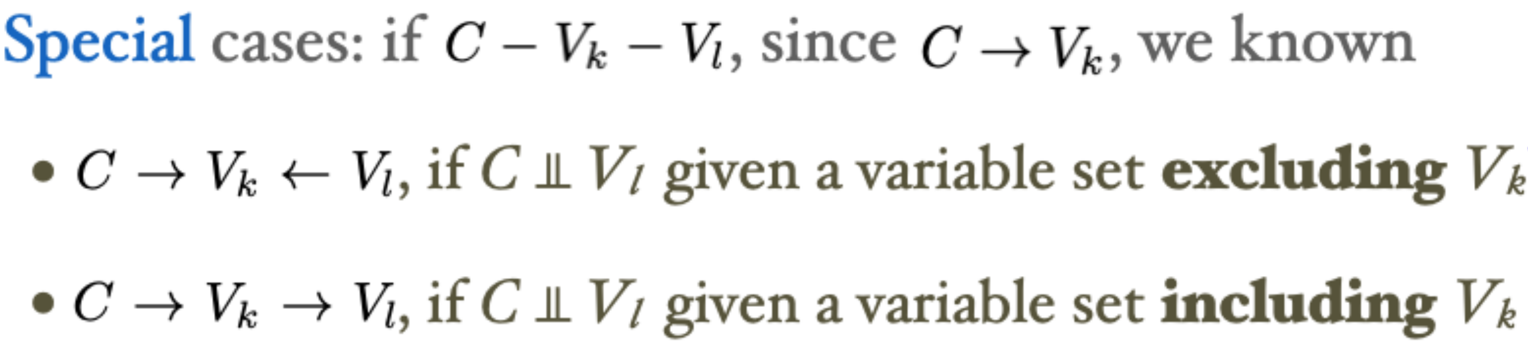


Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C., & Schölkopf, B. (2020). Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, *21*(89), 1-53.

Graph Learning: Determine edge direction

- Independent changes in P(cause) and P(effect|cause)





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

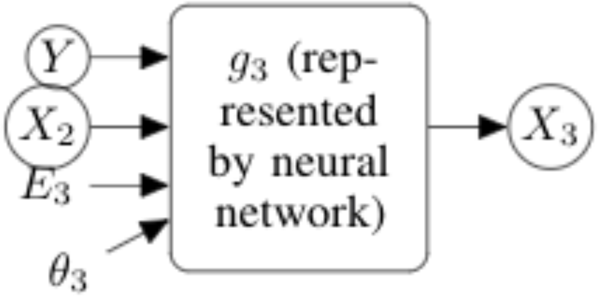
- $C \to V_k \leftarrow V_l$, if $C \perp\!\!\!\perp V_l$ given a variable set **excluding** V_k
- $C \to V_k \to V_l$, if $C \perp\!\!\!\perp V_l$ given a variable set **including** V_k

Invariant cause

Invariant mechanism

Approximate Inference

Latent variable conditional GAN



Approximate inference

$$\log p(\mathcal{D}) \geq -\sum_{i=1}^{s} \mathtt{KL}(q(\boldsymbol{\theta}|\mathcal{D}^i)|p(\boldsymbol{\theta})) + \mathbb{E}_{q(\boldsymbol{\theta}|\mathcal{D}^i)}\Big[\sum_{k=1}^{m_i} \log p_g(\mathbf{x}_k^{(i)}, y_k^{(i)}|\boldsymbol{\theta})\Big]$$

$$- \mathtt{KL}(q(\boldsymbol{\theta}|\mathcal{D}^\tau)|p(\boldsymbol{\theta})) + \mathbb{E}_{q(\boldsymbol{\theta}|\mathcal{D}^\tau)}\Big[\sum_{k=1}^{m} \log p_g(\mathbf{x}_k^\tau|\boldsymbol{\theta})\Big].$$

$$q(\boldsymbol{\theta}|\mathcal{D}^i) = \mathcal{N}(\boldsymbol{\theta}|\mu^{(i)}, \sigma^{(i)}),\ q(\boldsymbol{\theta}|\mathcal{D}^\tau) = \mathcal{N}(\boldsymbol{\theta}|\mu^\tau, \sigma^\tau)$$

Digits Adaptation

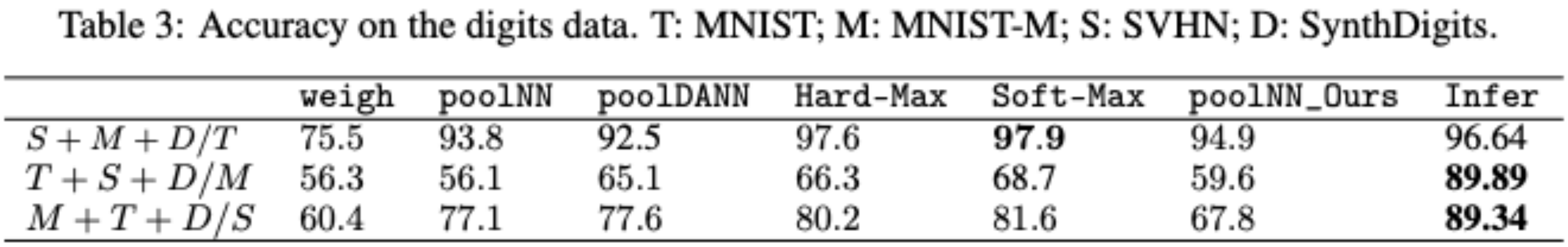
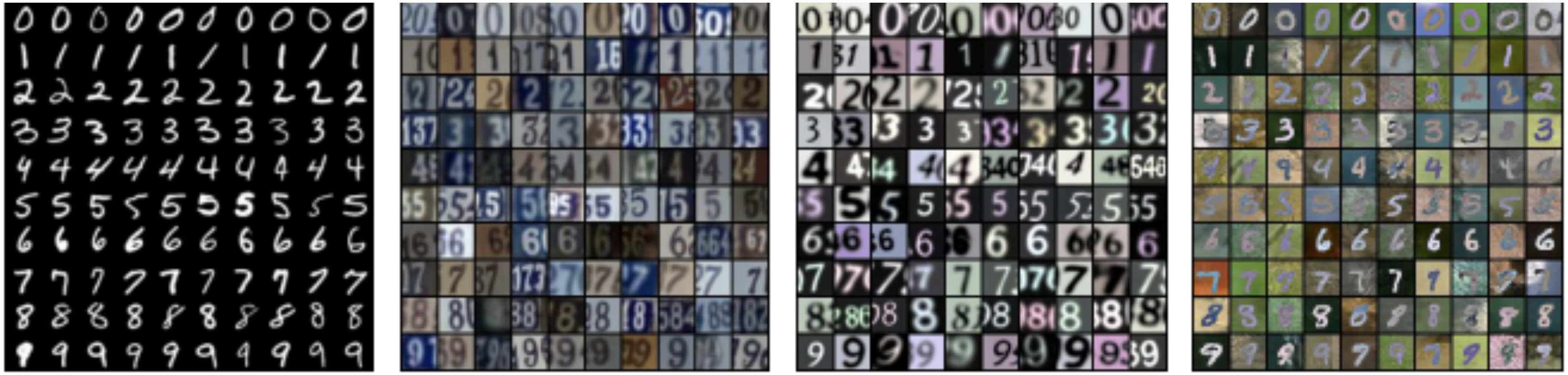


Table 3: Accuracy on the digits data. T: MNIST; M: MNIST-M; S: SVHN; D: SynthDigits.

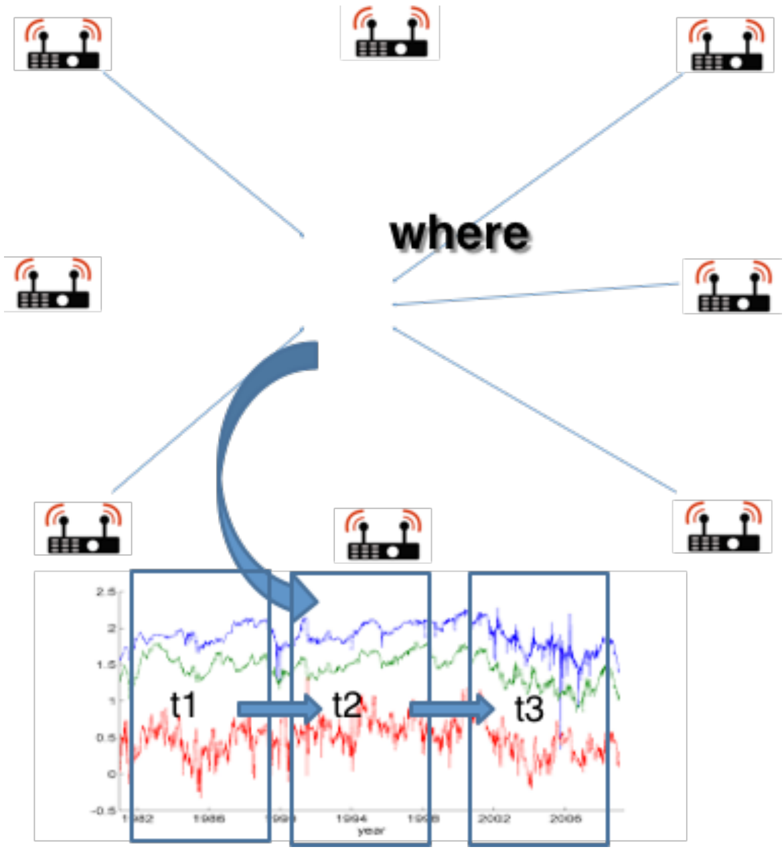
	weigh	poolNN	poolDANN	Hard-Max	Soft-Max	poolNN_Ours	Infer
$S + M + D/T$	75.5	93.8	92.5	97.6	**97.9**	94.9	96.64
$T + S + D/M$	56.3	56.1	65.1	66.3	68.7	59.6	**89.89**
$M + T + D/S$	60.4	77.1	77.6	80.2	81.6	67.8	**89.34**



MNIST SVHN SynthDigits MNIST-M

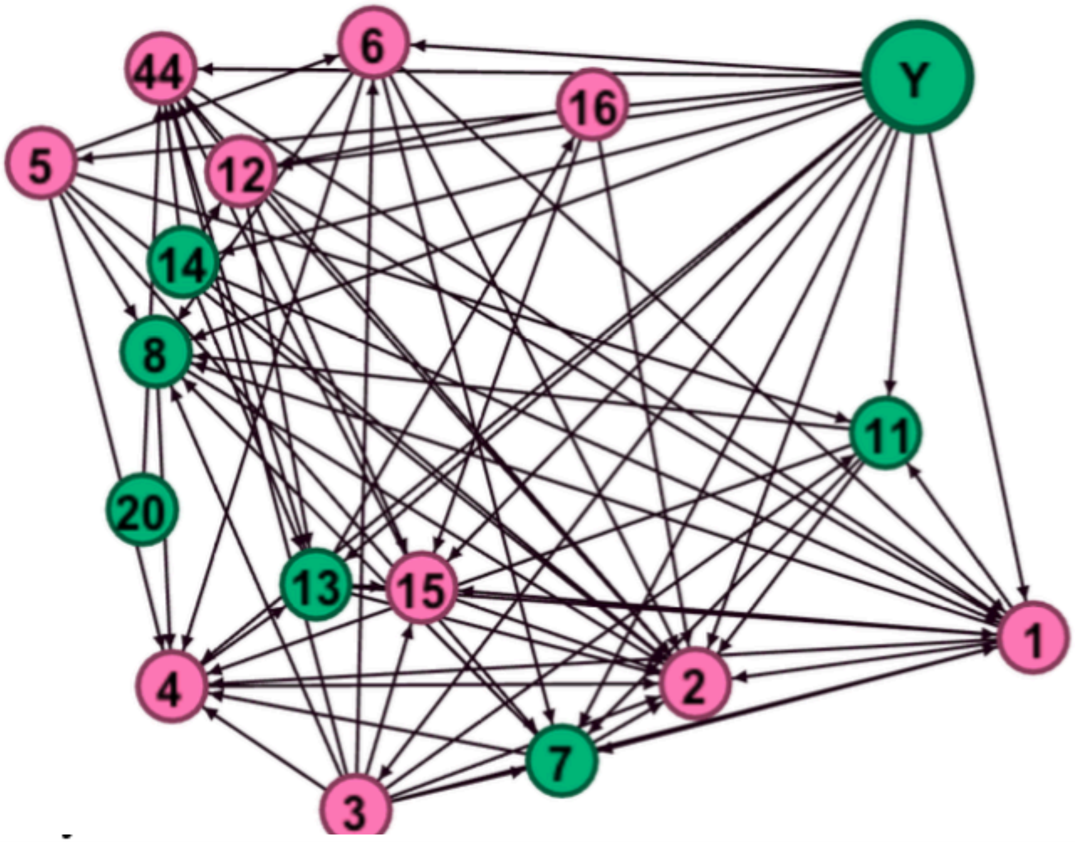
Gong, M., Xu, Y., Li, C., Zhang, K., & Batmanghelich, K. (2019). Twin auxilary classifiers GAN. In *Advances in neural information processing systems* (pp. 1330-1339).

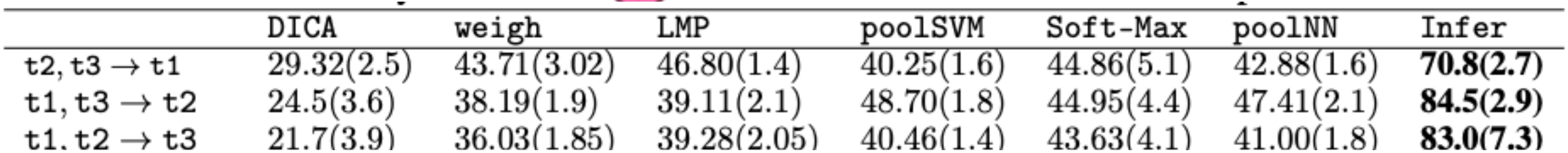
WiFi Localization



- Localize mobile devices from the WiFi signals.
- Transfer between different time periods

WiFi Localization





	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)**

Conclusion

- Augmented DAG encodes conditional independence relations and distribution change properties, which are sufficient for domain adaptation.
- Prediction in the target domain can be cast as Bayesian inference on the augmented DAG.
- Practical implementation: Approximate Inference + modeling conditional distributions using conditional GAN