Domain Adaptation under Missingness Shift

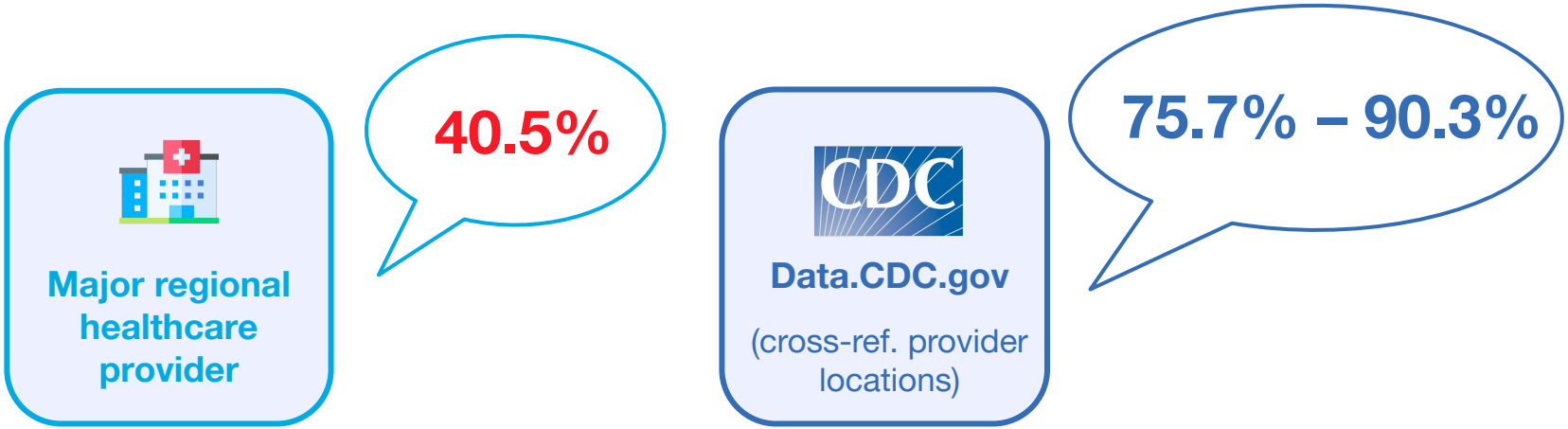
Helen Zhou, Sivaraman Balakrishnan, Zachary C. Lipton

Carnegie Mellon University

Motivation

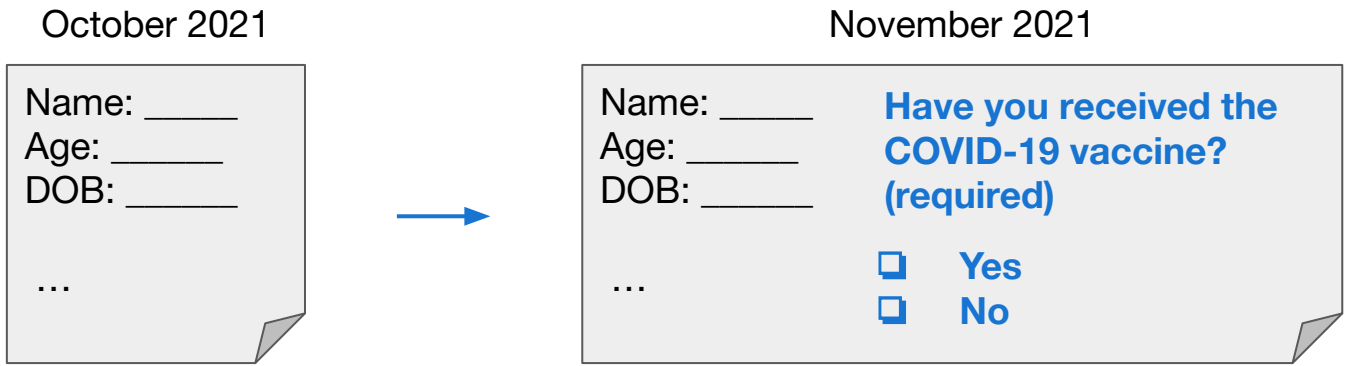
Q: *What proportion of individuals 18+ received ≥ 1 dose of the COVID-19 vaccine?*

As of October 2021, in southwestern Pennsylvania,



Motivation

Now, suppose the healthcare provider adopts a **new intake form**…



Absent any shift in the actual health status of patients, the distribution of observed data would still shift, owing to this sudden change in clerical practices.

Domain Adaptation under Missingness Shift (DAMS)

Faced with data from different time periods or locations, each characterized by **different missingness patterns**, how might practitioners use this data to produce the **best possible predictor on a target domain**?

Definitions

Missing Data Definition

In any environment e with **missing data**, we do not observe underlying clean covariates $X \in \mathbb{R}^d$ but instead observe corrupted covariates:

$$\widetilde{X} = X \odot \xi$$

where $\xi \in \{0,1\}^d$, and $(X, Y, \xi) \sim P^e$.

- $1 - \xi$ are missing data indicators, which may or may not be observed.
- ξ could be completely at random, dependent on other covariates, etc.

Missingness Shift Definition

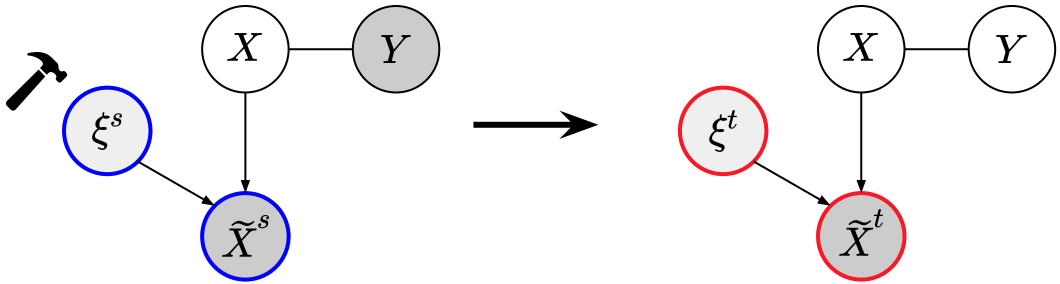
For a source domain *s* and target domain *t*, assume:

$$P(X, Y) = P^s(X, Y) = P^t(X, Y)$$

Missingness shift occurs when the missing data mechanism differs between source s and target t:

$$P^s(\xi \mid \cdot) \neq P^t(\xi \mid \cdot)$$

Domain Adaptation under Missingness Shift Definition

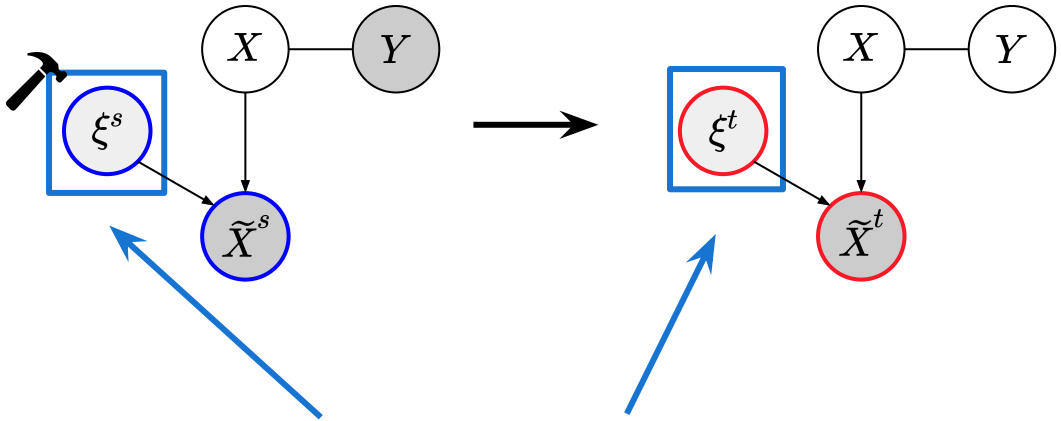


Suppose missingness shift occurs between source domain *s* and target domain *t*.

Labeled source data $\widetilde{X}^s, Y \sim P^s$, unlabeled target data $\widetilde{X}^t \sim P^t$.

Goal of **DAMS**: learn an optimal predictor on the corrupted target domain data.

One more detail…



Often, as in our example, **missing data indicators are not observed.**

(particularly tricky if substantial # of true 0s now indistinguishable from missing)

DAMS with Underreporting Completely at Random

To make this difficult setting tractable, we define the **DAMS with UCAR** setting, which we focus on throughout the rest of this work:

Assume ξ (unobserved) is parameterized by constant **unknown** missingness rates $m^s \in [0,1]^d$ in source and $m^t \in [0,1]^d$ in target. Then,

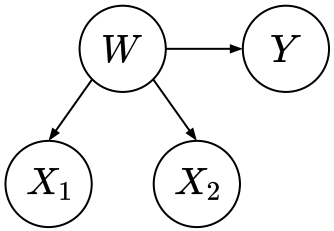
$$\xi_j^s \sim \text{Bernoulli}\left(1 - m_j^s\right)$$

$$\xi_j^t \sim \text{Bernoulli}\left(1 - m_j^t\right)$$

Cost of Non-Adaptivity

Let's start with a simple example…

Redundant Features



$$W = u_W \qquad u_W \sim \mathcal{N}(0, \sigma_w^2)$$
$$X_1 = W \qquad u_Y \sim \mathcal{N}(0, \sigma_y^2)$$
$$X_2 = W$$
$$Y = W + u_Y$$

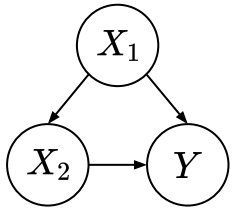
Suppose *W* is a latent variable,

and we observe $\widetilde{X} = \left[\widetilde{X}_1, \widetilde{X}_2\right]$.

(Failing to adapt
⇒ performance no better
than guess of label mean)

Motivating Example

Confounded Features



$$X_1 = u_1 \qquad u_1 \sim \mathcal{N}(0,1)$$
$$X_2 = aX_1 + u_2 \qquad u_2 \sim \mathcal{N}(0,1)$$
$$Y = bX_1 + cX_2 + u_Y \qquad u_Y \sim \mathcal{N}(0,1)$$

for constants a, b, c.

Suppose we observe $\widetilde{X} = \left[\widetilde{X}_1, \widetilde{X}_2\right]$

(Failing to adapt
⇒ performance **arbitrarily worse** than guessing label mean)

What if we **observed** missing data indicators $(1 - \xi)$?

If ξ depends on completely observed covariates (or is drawn completely at random), then missingness shift **satisfies the covariate shift assumption**:

$$P^s\left(Y \mid \widetilde{X}' = \tilde{x}'\right) = P^t\left(Y \mid \widetilde{X}' = \tilde{x}'\right),$$

where $\widetilde{X}' = \left(\widetilde{X}, \xi\right)$.

- There exists an optimal predictor that does not change across domains
- Covariate shift problems are relatively well-studied (Gretton et. al. 2009, Huang et. al. 2006, Shimodaira 2000, Sugiyama et. al. 2007, Tsymbal 2004)
- Extension: leveraging DAMS structure to estimate importance weights more efficiently

UCAR as L2 Regularization

- Observe that applying mask ξ (zeroing out covariates with some probability) resembles the mechanism of dropout in neural networks
- We show that for linear models, applying missingness rates to data scaled by $\frac{1}{1-m}$ is approximately equivalent to **L2 regularization** of β scaled by $\widetilde{\Delta}_{\text{diag}}$, where $\widetilde{\Delta}_{diag} = diag\left(\sqrt{\frac{m}{1-m}}\right) diag(\sqrt{I})$ and $diag(\sqrt{I})$ is the Fisher information matrix.

Identification

First, let us define *m-reachability*.

For data points $a \in \mathbb{R}^d$ and $b \in \mathbb{R}^d$, we say *b* is **m-reachable** from *a* (denoted $a \rightsquigarrow b$) if there exists some mask ξ such that $b = a \odot \xi$.

If $a \rightsquigarrow b$, then:

- Dimensions of *a* that are 0 must be a subset of the ones that are 0 in *b*
- Any dimensions that are nonzero in both *a* and *b* must match in value

Identifying corrupted from clean distribution (& vice versa!)

Consider any covariates $x \in \mathbb{R}^d$ and label $y \in \mathbb{R}$.

Given m, it is straightforward to identify $\tilde{p}$ from p:

Note: we require knowledge of **missingness rates** *m*

$$\tilde{p}_{x,y} = \sum_{z : z \rightsquigarrow x,\ z \in \mathbb{R}^d} p_{z,y} \cdot \prod_{j=1}^{d} (1 - m_j)^{x_j} m_j^{z_j - x_j}$$

prob. of x, y in corrupted dist.

prob. of z, y in clean dist.

missingness rate of jth covariate

Interestingly, given $m \prec 1$, we can show that p is identifiable from $\tilde{p}$ (induction on # zeros).

Unfortunately, *m* is not in general identifiable…

Consider two possible source distributions:

$$A \sim \text{Bernoulli}(0.5)$$
$$B \sim \text{Bernoulli}(0.25)$$

Applying missingness rates $m_A = 0.5$ to *A* and $m_B = 0$ to *B* yields identical observed corrupted distributions:

$$\widetilde{A} \overset{\text{iid}}{\sim} \widetilde{B} \sim \text{Bernoulli}(0.25)$$

Thus, the rates are not identifiable.

But fortunately, we can identify…

1) Relative missingness rates:

The ratio between non-missingness rates $1 - m^t$ and $1 - m^s$ is given by:

$$\frac{P^t(\tilde{x}\neq 0)}{P^s(\tilde{x}\neq 0)} = \frac{(1-m^t)\odot q}{(1-m^s)\odot q} = \frac{1-m^t}{1-m^s} \overset{\Delta}{=} 1 - r^{s\to t},$$

2) And thus, the labeled target distribution from the labeled source distribution

$$\tilde{p}^t_{x,y} = \sum_{z:z\rightsquigarrow x, z\in\mathbb{R}^d} \tilde{p}^s_{z,y} \cdot \prod_{j=1}^{d} \left(1 - r_j^{s\to t}\right)^{x_j} \left(r_j^{s\to t}\right)^{z_j - x_j}.$$

Estimation

Estimating relative missingness rates $r^{s \to t}$

Simply compute $\hat{q}_j^s = \dfrac{\text{count}\left(\tilde{x}_j^s \neq 0\right)}{n_s}$, $\hat{q}_j^t = \dfrac{\text{count}\left(\tilde{x}_j^t \neq 0\right)}{n_t}$, and $\hat{r}^{s \to t} = 1 - \dfrac{\hat{q}^t}{\hat{q}^s}$.

Using Hoeffding's inequality, we show that with probability $1 - \delta$,

$$\left|\hat{r}^{s \to t} - r^{s \to t}\right| \leq \frac{1}{\hat{q}^s}\left(\sqrt{\frac{\log\left(4/\delta\right)}{2n_t}} + \left(1 - r^{s \to t}\right)\sqrt{\frac{\log\left(4/\delta\right)}{2n_s}}\right).$$

Non-parametric adjustment procedure

1. Compute $\hat{r}^{s \to t} = 1 - \frac{\hat{q}^t}{\hat{q}^s}$.
2. Check whether $\hat{r}^{s \to t} \succeq 0$. If not, this procedure is not a "proper" adjustment.
3. Compute $\tilde{r}^{s \to t} = \max\left(\hat{r}^{s \to t}, 0\right)$, where max is elementwise.
4. Apply missingness with rate $r^{s \to t}$ to source data to get data distributed identically to target data.
5. Fit a predictor on the (further) corrupted labeled source data.

Closed-Form Solution for Optimal Linear Predictor

The optimal linear target predictor is given by:

$$\beta_t^* = \mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t]^{-1}\left(r^{s\to t} \odot \mathbb{E}[\widetilde{X}^{s\top}Y^s]\right).$$

We can estimate $\mathbb{E}[\widetilde{X}^{t\top}\widetilde{X}^t]$ in two ways:

1. Directly from unlabeled target data
2. From source data: for $i \neq j$,

$$\mathbb{E}[\widetilde{X}_t^T\widetilde{X}_t]_{ij} = (1-r_i^{s\to t})(1-r_j^{s\to t})\mathbb{E}[\widetilde{X}_s^T\widetilde{X}_s]_{ij}$$
$$\mathbb{E}[\widetilde{X}_t^T\widetilde{X}_t]_{ii} = (1-r_i^{s\to t})\mathbb{E}[\widetilde{X}_s^T\widetilde{X}_s]_{ii}.$$

Experiments

Synthetic Experiments

We draw 10,000 samples from two simple data-generating processes:

Scenario 1: "Redundant Features"

$$
\begin{aligned}
u_y &\sim \mathcal{N}(0,1) \\
Z &\sim \text{Bernoulli}(0.5) \\
X_1 &= Z \\
X_2 &= Z \\
Y &= Z + u_y
\end{aligned}
$$

Scenario 2: "Confounded Features"

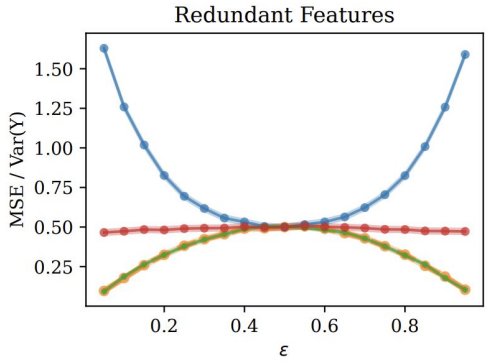
$$
\begin{aligned}
u_{x_2} &\sim \mathcal{N}(0,1) \\
u_y &\sim \mathcal{N}(0,1) \\
X_1 &\sim \text{Bernoulli}(0.5) \\
X_2 &= \text{expit}(2X_1 + u_{x_2}) \\
Y &= X_1 - X_2 + u_y
\end{aligned}
$$

in both, we apply missingness with rates $m^s = [1-\epsilon, \epsilon]$ and $m^t = [\epsilon, 1-\epsilon]$.

Synthetic Experiments

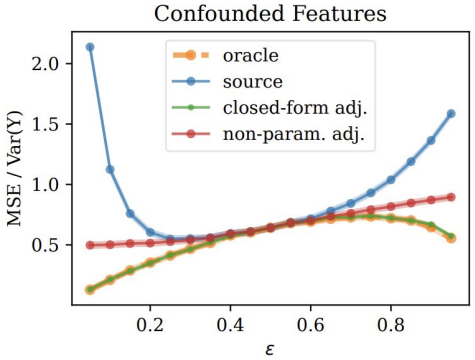
Scenario 1: "Redundant Features"

$$
\begin{aligned}
u_y &\sim \mathcal{N}(0,1) \\
Z &\sim \text{Bernoulli}(0.5) \\
X_1 &= Z \\
X_2 &= Z \\
Y &= Z + u_y
\end{aligned}
$$



Scenario 2: "Confounded Features"

$$
\begin{aligned}
u_{x_2} &\sim \mathcal{N}(0,1) \\
u_y &\sim \mathcal{N}(0,1) \\
X_1 &\sim \text{Bernoulli}(0.5) \\
X_2 &= \text{expit}(2X_1 + u_{x_2}) \\
Y &= X_1 - X_2 + u_y
\end{aligned}
$$



$$m^s = [1-\epsilon, \epsilon]$$

$$m^t = [\epsilon, 1-\epsilon]$$

Semi-synthetic Experiments (real data, synthetic labels)

Average target domain error, given by MSE/Var(Y), on synthetic and semi-synthetic data:

	Rednd.	Confnd.	Adult		Bank		Thyroid	
	$m^s \,?\, m^t$	$m^s \,?\, m^t$	$m^s \preceq m^t$	$m^s \,?\, m^t$	$m^s \preceq m^t$	$m^s \,?\, m^t$	$m^s \preceq m^t$	$m^s \,?\, m^t$
Linear Regression Models								
Oracle	0.178	0.206	0.420	0.362	0.338	0.433	0.298	0.251
Source	1.259	1.103	0.437	0.380	0.371	0.480	0.350	0.320
Imputed	1.002	0.918	0.490	0.483	0.501	0.592	0.306	0.358
Closed-form	**0.186**	**0.209**	0.422	**0.363**	0.339	**0.442**	0.316	**0.291**
Non-param.	0.473	0.492	**0.420**	0.373	**0.338**	0.459	**0.293**	**0.291**
XGBoost Models								
Oracle	0.166	0.200	0.398	0.354	0.287	0.453	0.316	0.274
Source	**0.166**	0.475	**0.399**	**0.379**	0.305	**0.500**	**0.310**	**0.352**
Imputed	1.002	1.157	0.512	0.521	0.492	0.708	0.355	0.441
Non-param.	0.425	**0.473**	**0.399**	0.392	**0.287**	0.503	**0.310**	0.381
MLP Models								
Oracle	0.166	0.201	0.389	0.343	0.295	0.458	0.279	0.230
Source	**0.184**	**0.321**	0.399	0.357	0.322	0.499	0.320	0.303
Imputed	1.003	0.924	0.480	0.468	0.484	0.668	0.304	0.345
Non-param.	0.436	0.470	**0.389**	**0.355**	**0.294**	**0.487**	**0.278**	**0.272**

Discussion

- When indicators are provided and missingness depends on completely observed covariates, DAMS can be viewed as a form of covariate shift
- When missing data indicators are not provided, we provide identification and estimation results for the DAMS with UCAR setting
- Experiments validate our findings when assumptions hold

Extensions

- Seek **real-world data** well-suited for DAMS with UCAR
- Explore generalizations or relaxations of the DAMS with UCAR assumptions:
  - Allowing **dependence on covariates**
  - Other **noise models**

Thank you!

Feel free to reach out: hlzhou@andrew.cmu.edu