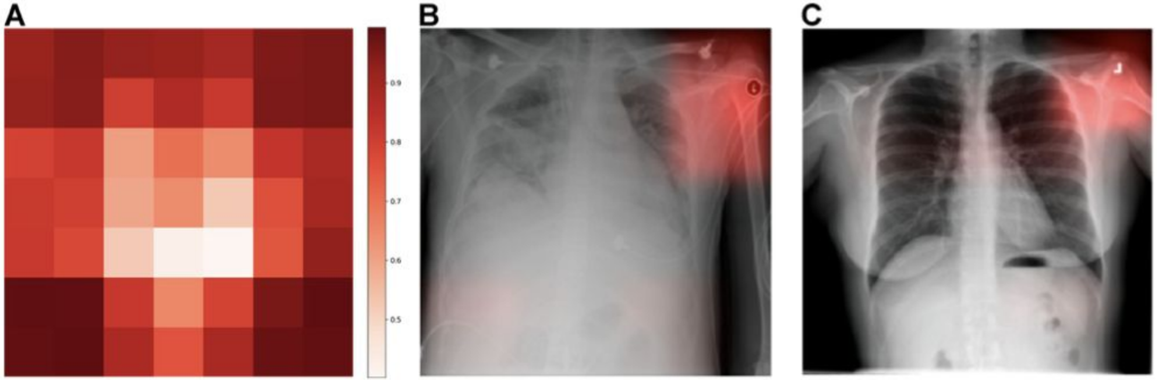
Causal Information Splitting: Engineering Proxy Features for Robustness to Distribution Shifts

Bijan Mazaheri, Atalanti Mastakouri, Dominik Janzing, Michaela Hardt

Background

- Key assumption for building predictive models is relevant training data pulled from a distribution identical to its use case
- Real-world data contains biases that shift training data distribution shifts



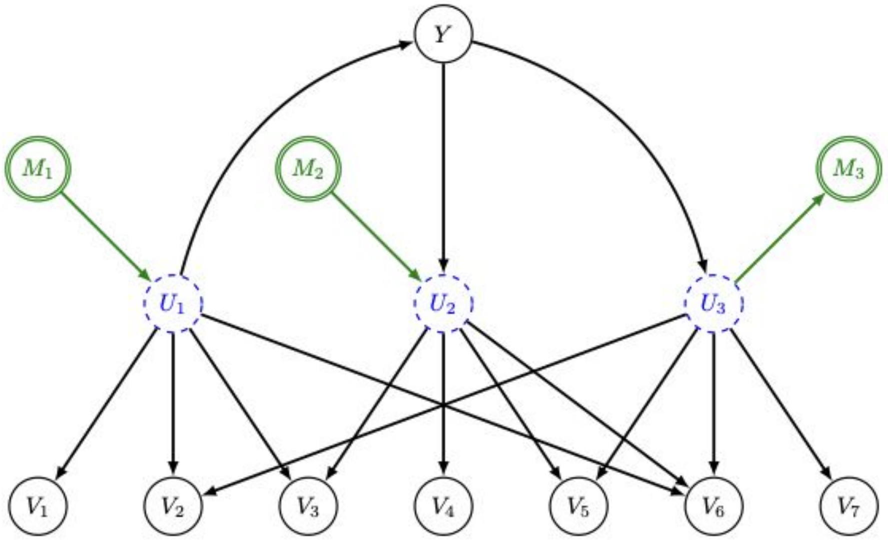
Transportability

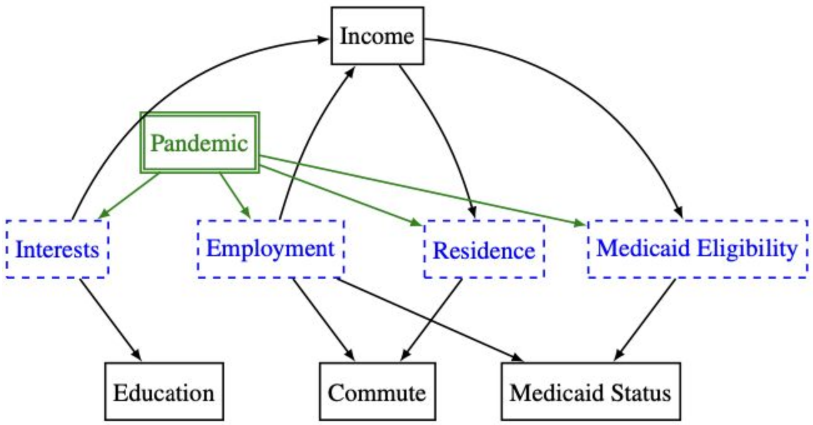
- To handle dissociation between training and target distributions:
  - Covariate shift in distribution of X
  - Label shift of $\Pr(Y)$
  - Stationary label function $\Pr(Y|X)$
- Label function is stationary for a subset of X (invariant set)
- Transportability problem: find invariant set X

Setting/Assumptions

- Further challenges to identifying distribution shift when lacking direct measurements of causes and effects of Y
- Proxy-Based Transportability (PBT) setting:
  - All causes and effects (U) of Y are unobserved
  - Visible proxy variables (V) are descendants of at least one $u \in U$
- Systemic Sparsity: no edges directly within U or V
  - $dsep(V_i\ \ U, V_j)$ and $dsep(U_i, Y, U_j)$
- Distribution Shift Diagram $G^+ = (U \cup V \cup M, E \cup E_M)$
  - One $M_i \in \boldsymbol{M}$ connected to corresponding $U_i \in \boldsymbol{U}$, where each M_i corresponds to shifting mechanism for unobserved cause and effect of Y

G^+ Examples


Y
M1
M2
M3
U1
U2
U3
V1
V2
V3
V4
V5
V6
V7


Income
Pandemic
Interests
Employment
Residence
Medicaid Eligibility
Education
Commute
Medicaid Status

Transportability Approaches in PBT Setting

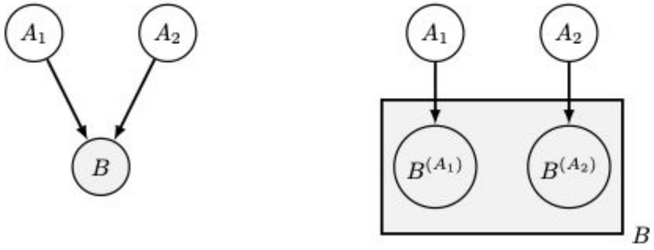
- Edges between vertices A and B given by:

$$B^A(A) \begin{cases} T_{A,B} & \text{with probability } \alpha_{A,B} \\ \phi & \text{with probability } 1-\alpha \end{cases}$$

- Structural equation for vertex B:

$$B = T_B(\{B^A(A) \text{ for } A \in Parents(B)\})$$

- Given multiple parents, B can be split into separate, disconnected vertices



Context Sensitivity

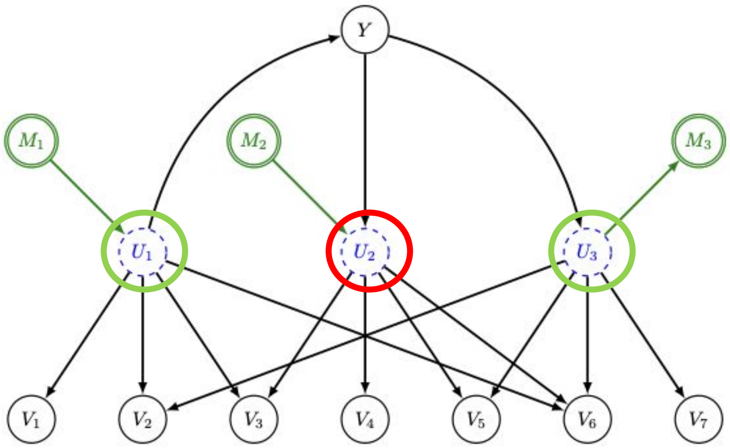
- Find set of features X that minimizes conditional mutual information between label and biases
- Minimize Context Sensitivity: $\mathrm{I}(Y: M|X)$
- Find X subset that d-separates M from Y

Redundancy, Context Sensitivity, and Colliders

- Redundancy: $\mathrm{I}(U:\boldsymbol{X}) = \mathrm{H}(U) - \mathrm{H}(U|X)$
  - Using dropout function setting: $\mathrm{I}(U:\boldsymbol{X}) = \alpha_{U,Children_X(U_i)}\mathrm{H}(U)$
- Non-collider vertex context sensitivity: $\mathrm{I}(M_i:Y|X) = \alpha_{M_i,U_i}(1 - \alpha_{U_i,Children_x(U_i)})\alpha_{U_i,Y}\mathrm{H}(M_i)$
- Collider vertex context sensitivity: $\mathrm{I}(M_i:Y|X) = \alpha_{U_i,Children_x(U_i)}\mathrm{I}(M_i:Y|U_i)$

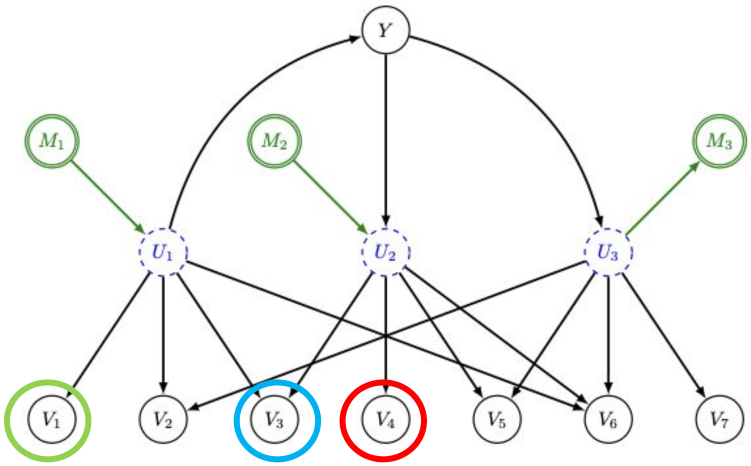
“Good” vs. “Bad” Unobserved Factors

- $U_i \in U^{GOOD}$ when $dsep(M_i, U_i, Y)$
- $U_i \in U^{BAD}$ when path $M_i \rightarrow U_i \leftarrow Y$ exists



U^{GOOD}, U^{BAD}, and Resulting Proxies

- Proxies contain combinations of universally-relevant and domain-relevant features, resulting in multiple classes of proxy variables:
  - $V^{GOOD} \coloneqq CHILDREN(U^{GOOD}) \setminus \text{CHILDREN}(U^{BAD})$
  - $V^{BAD} \coloneqq CHILDREN(U^{BAD}) \setminus \text{CHILDREN}(U^{GOOD})$
  - $V^{AMBIGUOUS} \coloneqq CHILDREN(U^{BAD}) \cap \text{CHILDREN}(U^{GOOD})$



Proxy Bootstrapping

- Harness partial information to classify proxies as $V^{GOOD}, V^{BAD}, V^{AMBIG}$
- Given DSD $G^{+} = (U \cup V \cup M, E \cup E_M)$, create graph $G_Y(V, E_Y)$ s.t. $(V_i, V_j) \in E_Y$ iff not $dsep(V_i, Y, V_j)$
- For vertices with known assignments $V^{*} \in V$:
  - $V^{*} \in V^{GOOD} \rightarrow$ "good" label to all neighbors of V^{*}
  - $V^{*} \in V^{BAD} \rightarrow$ "bad" label to all neighbors of V^{*}
- All $V \in \boldsymbol{V} \setminus \mathrm{V}^{*}$ with both labels receive "ambiguous" label

Feature Engineering

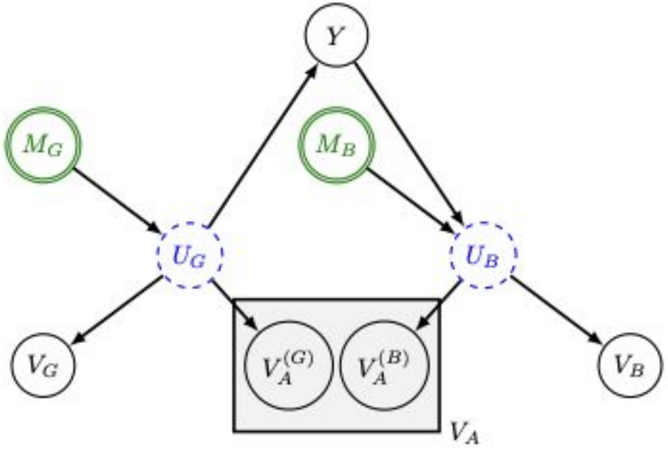
- Build model where output of functions is related to U^{GOOD}, not related to U^{BAD}
- Building models with more redundancy to U^{GOOD} improves context sensitivity:

$$\mathrm{I}(M_i\colon Y|X) = \alpha_{M_i,U_i}\alpha_{U_i,Y}\mathrm{H}(U_i|Children_X(U_i))$$

- If redundancy with U^{BAD} is avoided, avoid picking up sensitivity from associated shifting mechanisms
  - For $U_i \in U^{BAD}$, if it is maintained that $\mathrm{I}(U_i\colon X|Y) = 0$, then $\mathrm{I}(M_i\colon Y|X) = 0$

Causal Information Splitting

- Separable Ambiguous Proxies: components of V^{AMBIG}, isolating "good" information from "bad"



Isolation Functions

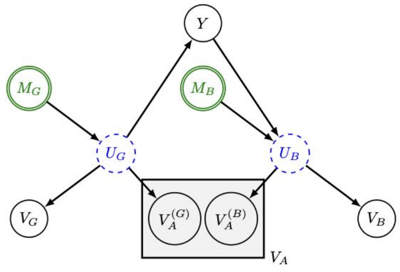
- Isolation Functions:
  - $F_{ISO(V_i)}(V_A|y) \coloneqq argmin_F \mathrm{H}\big(F(V_A|y)\big)$ such that $\mathrm{I}(F(V_A): V_i|y) = \mathrm{I}(V_A: V_i|y)$
- To achieve $I(F(V_A): U^{BAD}|Y) = 0$, while preserving information about U^{GOOD}, ideally isolate U^{GOOD}
- In given setting, isolate V^{GOOD} using $F_{ISO(V^{GOOD})}(V_A|Y)$
- $\mathrm{I}\big(U_{BAD}: F_{ISO(V_{GOOD})}(V_A|Y)\big|Y\big) = 0$
- Isolation functions at worst avoid worsening context sensitivity
- Auxiliary training functions/tasks: get approximate isolation function by training model to predict V_i using V_A

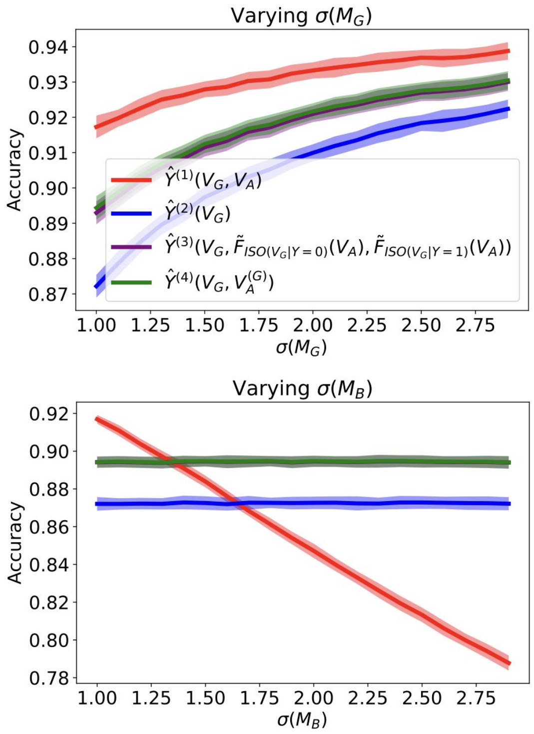
Procedure for Robust Model Building

1. Partition data into constant $Y = y$
2. Identify seeds in V^{GOOD}, V^{BAD} for proxy bootstrapping
3. Perform Causal Information Splitting on V^{AMBIG}
4. Build prediction model for Y using V^{GOOD} and CIS-engineered V^{AMBIG}

Experiments (Synthetic Data)

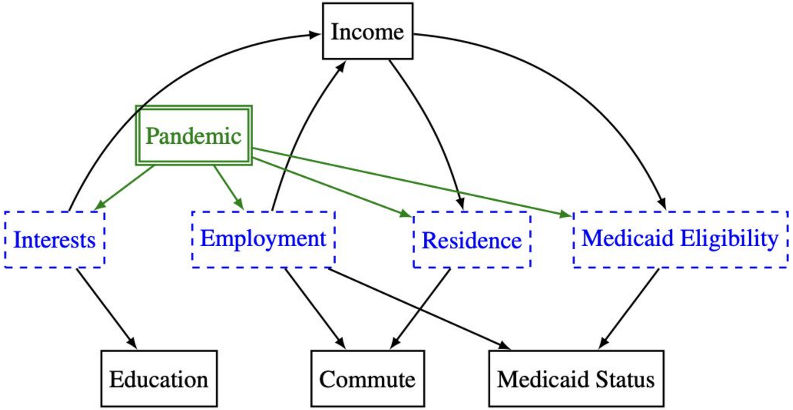
- Generate data based for DAG:
- $\hat{Y}^1$ trained on $V^{GOOD} \cup V^{AMBIG}$
- $\hat{Y}^2$ trained on V^{GOOD}
- $\hat{Y}^3$ (Feature engineering based on CIS) trained on $V^{GOOD} \cup F_{ISO(V_G)}(V_A)$
- $\hat{Y}^4$ trained on $V^{GOOD} \cup V_A^{GOOD}$





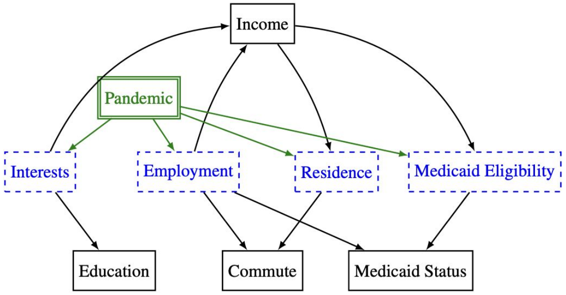
Experiments (Census Data)

- Predict if income of a person exceeds 50k
- Models built on pre-pandemic data, evaluated on 2021 data during pandemic
- Model Inputs:
  - Commute time
  - Received government assistance
  - Education level



Experiments (Census Data)

- Engineered features: does not use *Commute* or *Medicaid Status* directly
  - Trains models to use features to predict education-level
- Compared to all features and just education (limited features)



State	All Features	Engineered Features	Limited Features
CA	**0.712** $\pm$ 0.0011	**0.711** $\pm$ 0.0014	0.692 $\pm$ 0.0014
FL	**0.683** $\pm$ 0.0012	0.678 $\pm$ 0.0018	0.68 $\pm$ 0.0013
GA	0.689 $\pm$ 0.0025	**0.707** $\pm$ 0.0055	**0.709** $\pm$ 0.0029
IL	0.662 $\pm$ 0.0026	**0.689** $\pm$ 0.0033	0.684 $\pm$ 0.0019
NY	**0.707** $\pm$ 0.0022	**0.702** $\pm$ 0.0025	0.687 $\pm$ 0.008
NC	**0.691** $\pm$ 0.0031	**0.684** $\pm$ 0.0034	**0.683** $\pm$ 0.003
OH	0.689 $\pm$ 0.0022	**0.703** $\pm$ 0.004	**0.696** $\pm$ 0.0029
PA	0.672 $\pm$ 0.0017	**0.695** $\pm$ 0.0023	0.688 $\pm$ 0.0022
TX	0.69 $\pm$ 0.0029	**0.712** $\pm$ 0.0028	**0.712** $\pm$ 0.0027
avg	0.688	**0.698**	0.692

Table 2: Comparison of in-domain (2019) performance on predicting high income via Accuracies.

State	All Features	Engineered Features	Limited Features
CA	**0.713** $\pm$ 0.0010	**0.710** $\pm$ 0.0012	0.691 $\pm$ 0.0011
FL	**0.700** $\pm$ 0.0014	0.693 $\pm$ 0.0020	0.694 $\pm$ 0.0017
GA	**0.708** $\pm$ 0.0025	**0.708** $\pm$ 0.0036	**0.707** $\pm$ 0.0036
IL	**0.689** $\pm$ 0.0023	**0.690** $\pm$ 0.0039	**0.685** $\pm$ 0.0021
NY	**0.705** $\pm$ 0.0024	0.698 $\pm$ 0.0022	0.687 $\pm$ 0.0076
NC	**0.713** $\pm$ 0.0020	0.703 $\pm$ 0.0049	0.700 $\pm$ 0.0028
OH	**0.717** $\pm$ 0.0029	**0.716** $\pm$ 0.0042	**0.712** $\pm$ 0.0033
PA	**0.702** $\pm$ 0.0028	**0.701** $\pm$ 0.0027	0.695 $\pm$ 0.0026
TX	**0.708** $\pm$ 0.0019	**0.705** $\pm$ 0.0025	**0.706** $\pm$ 0.0022
avg	**0.706**	0.703	0.697

Results

- Feature selection based on conditional independence tests
- Causal Information Splitting allows isolation of robust predictive power
- Engineered features increase robustness and can improve accuracy