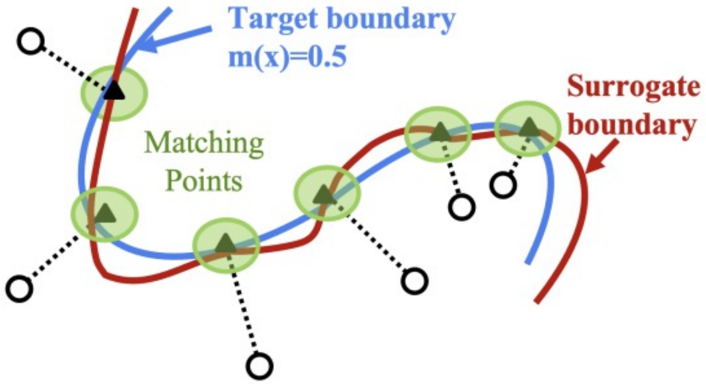
NEURAL INFORMATION PROCESSING SYSTEMS

Model Reconstruction Using Counterfactual Explanations: A Perspective From Polytope Theory

Pasan Dissanayake, Sanghamitra Dutta

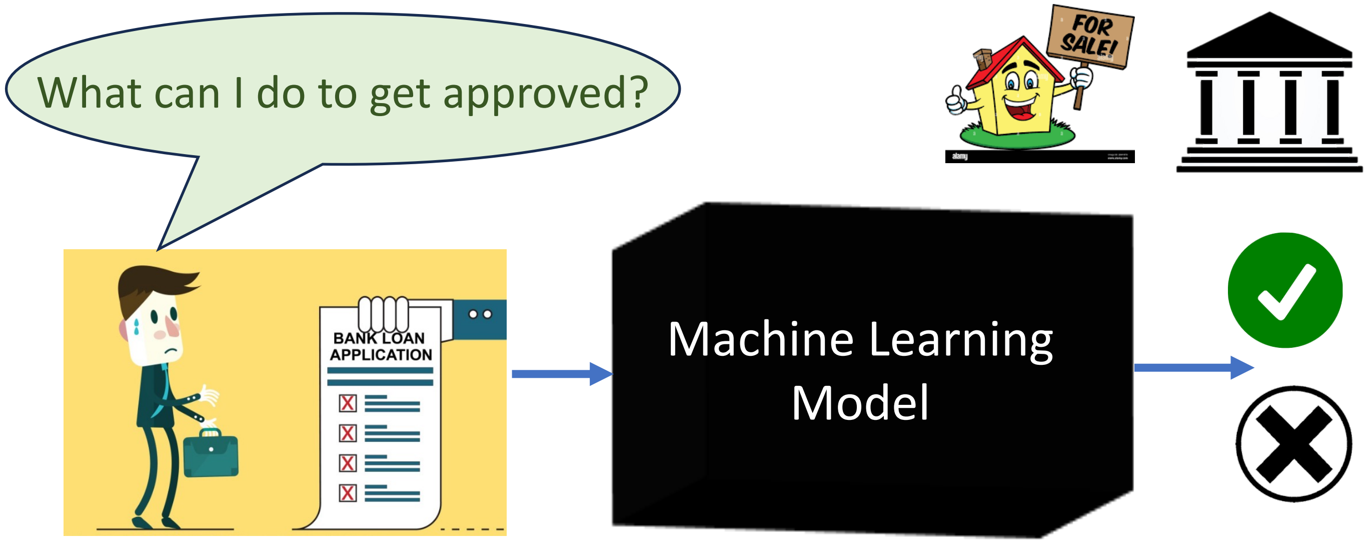
University of Maryland College Park

Poster: Wed 11 Dec 4:30 p.m. PST — 7:30 p.m. PST

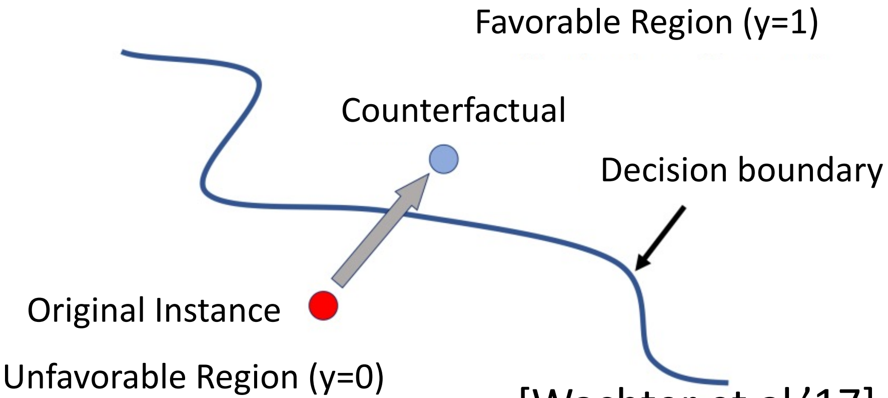


CENTER FOR MACHINE LEARNING

What are Counterfactual Explanations?

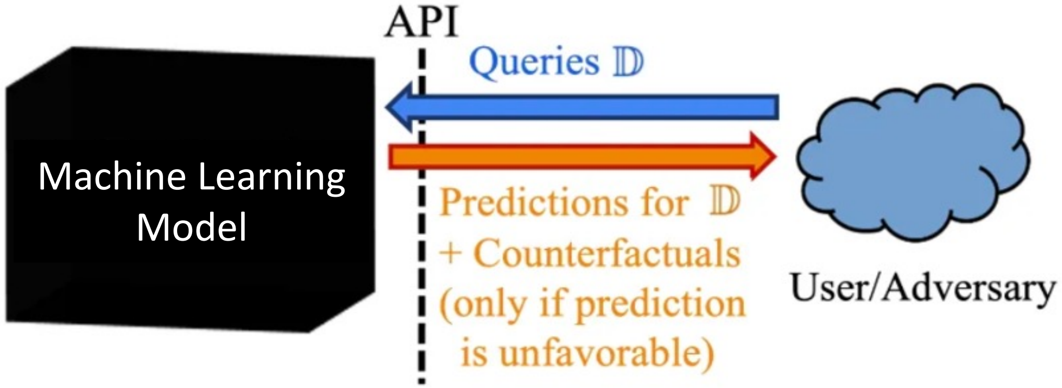


Counterfactual explanation: “closest” point on the other side of the decision boundary



[Wachter et al.’17]

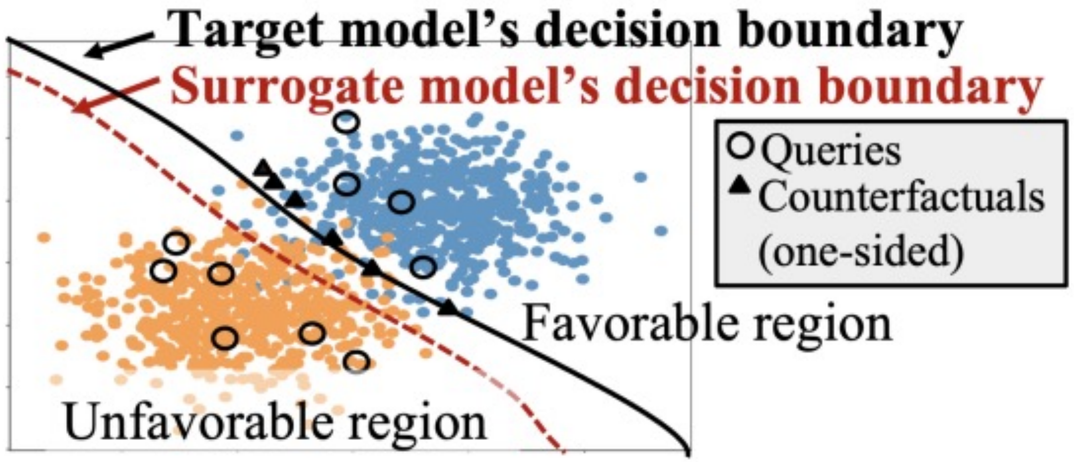
How faithfully can one reconstruct a model using counterfactual explanations?



Training a surrogate model using all the queried datapoints (y=0/1) and one-sided counterfactuals (for datapoints with y=0)

Counterfactuals as ordinary labelled instances?

Decision boundary shift issue



Question: Can we improve model reconstruction specifically leveraging the fact that the counterfactuals are quite close to the boundary?

Main Contribution:

Novel Model Reconstruction Strategies Using Counterfactuals With Theoretical Guarantees From Polytope Theory

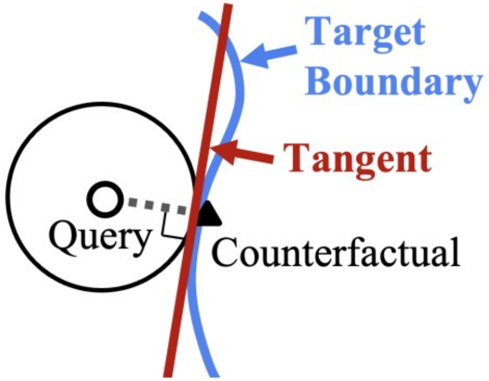
Related Works: *[Aivodji et al.'20][Wang et al.'22]*

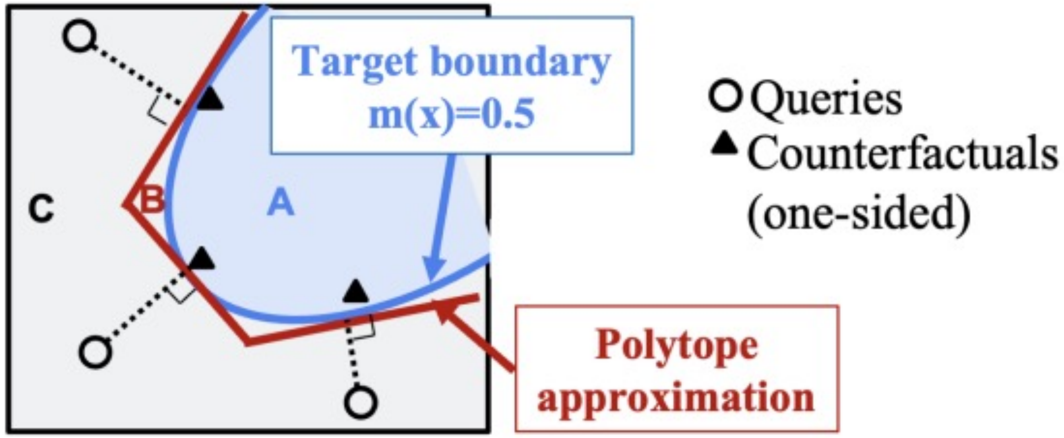
Other Privacy + CF: [Pawelczyk et al.'23][Goethals at al.'23]][Yadav et al.'23]

Model extraction in other settings: [Gong et al.'20] [Milli et al.'19]

Main Results

1. Convex Decision Boundaries and Closest Counterfactuals



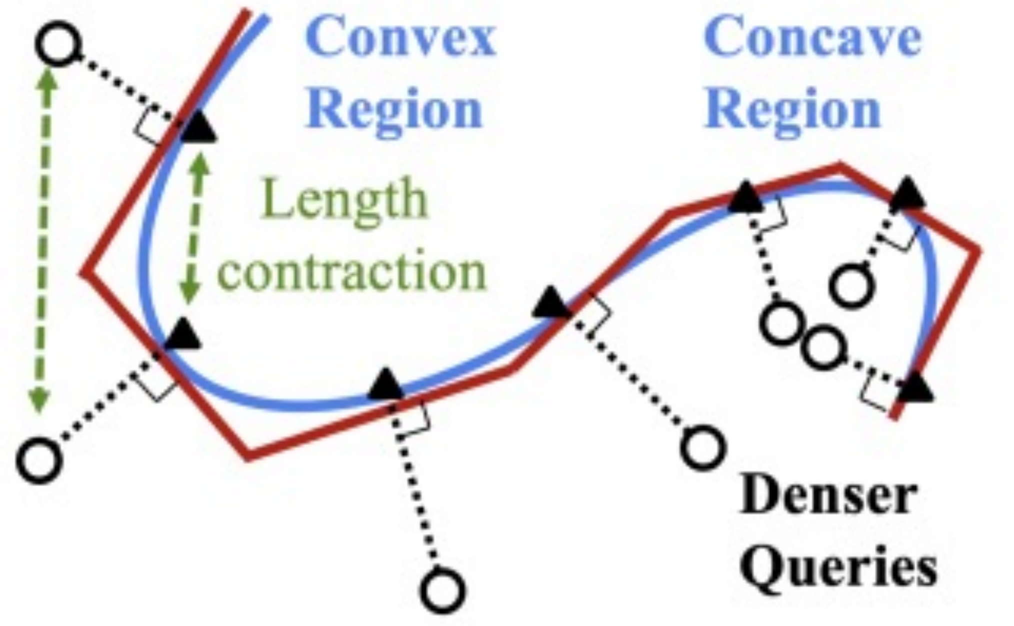


Theorem 3.2. *Let m be the target binary classifier whose decision boundary is convex (i.e., the set $\{\boldsymbol{x} \in [0,1]^d : \lfloor m(\boldsymbol{x}) \rceil = 1\}$ is convex) and has a continuous second derivative. Denote by $\tilde{M}_n$, the convex polytope approximation of m constructed with n supporting hyperplanes obtained through i.i.d. counterfactual queries. Assume that the fidelity is evaluated with respect to $\mathbb{D}_{ref}$ which is uniformly distributed over $[0,1]^d$. Then, when $n \to \infty$ the expected fidelity of $\tilde{M}_n$ with respect to m is given by*

$$\mathbb{E}\left[\mathrm{Fid}_{m,\mathbb{D}_{ref}}(\tilde{M}_n)\right] = 1 - \epsilon \tag{1}$$

where $\epsilon \sim \mathcal{O}\left(n^{-\frac{2}{d-1}}\right)$ and the expectation is over both $\tilde{M}_n$ and $\mathbb{D}_{ref}$.

Theoretical guarantees on **exact volume approximation using counterfactuals** leveraging polytope theory


Convex
Region
Concave
Region
Length
contraction
Denser
Queries


Convex
Non-Convex
Lipschitz
ReLU Networks

Main Results

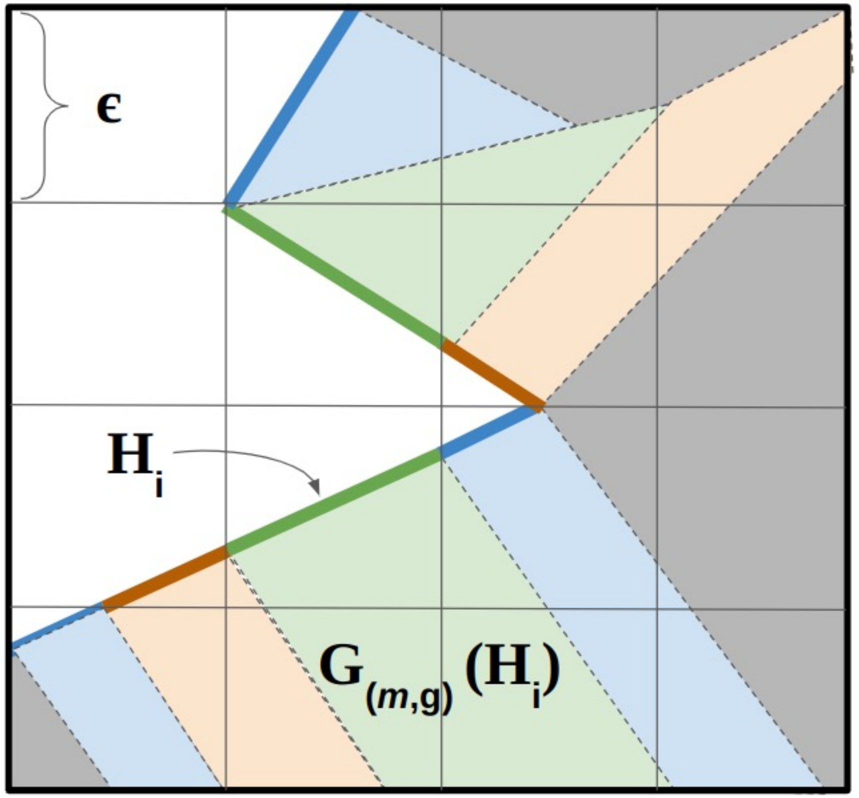
2. ReLU Networks and Closest Counterfactuals

Continuous Piece-Wise Linear (CPWL) Functions

Theorem 3.6. *Let m be a target binary classifier with ReLU activations. Let $k(\epsilon)$ be the number of cells through which the decision boundary passes. Define $\{\mathbb{H}_i\}_{i=1,\dots,k(\epsilon)}$ to be the set of affine pieces of the decision boundary within each decision boundary cell. Let $v_i(\epsilon) = V(\mathbb{G}_{m,g_m}(\mathbb{H}_i))$ where $V(.)$ is the $d-$dimensional volume (i.e., the Lebesgue measure) and $\mathbb{G}_{m,g_m}(.)$ is the inverse counterfactual region w.r.t. m and the closest counterfactual generator g_m. Then the probability of successful reconstruction with counterfactual queries distributed uniformly over $[0,1]^d$ is lower-bounded as*

$$\mathbb{P}\left[\mathit{Reconstruction}\right] \geq 1 - k(\epsilon)(1 - v^*(\epsilon))^n \tag{2}$$

where $v^*(\epsilon) = \min_{i=1,\dots,k(\epsilon)} v_i(\epsilon)$ and n is the number of queries.

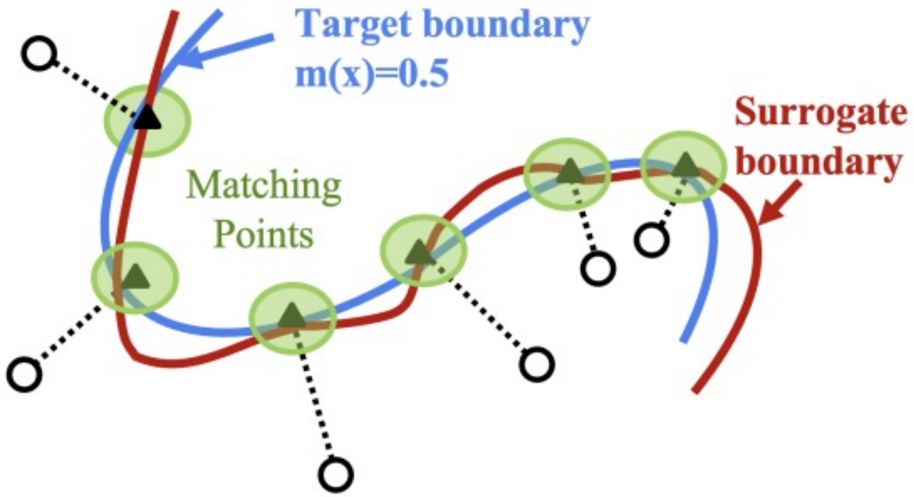


Main Results

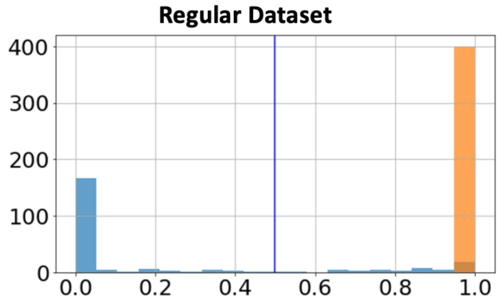
3. Beyond Closest Counterfactuals

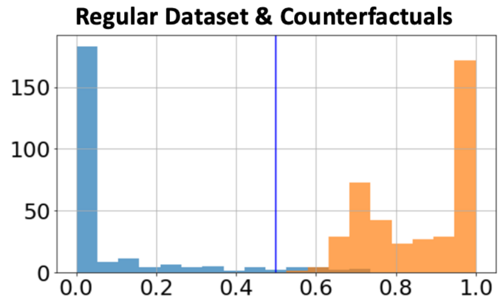
Theorem 3.10. *Let the target m and surrogate $\tilde{m}$ be ReLU classifiers such that $m(\boldsymbol{w}) = \tilde{m}(\boldsymbol{w})$ for every counterfactual $\boldsymbol{w}$. For any point $\boldsymbol{x}$ that lies in a decision boundary cell, $|\tilde{m}(\boldsymbol{x}) - m(\boldsymbol{x})| \leq \sqrt{d}(\gamma_m + \gamma_{\tilde{m}})\epsilon$ holds with probability $p \geq 1 - k(\epsilon)(1 - v^*(\epsilon))^n$.*

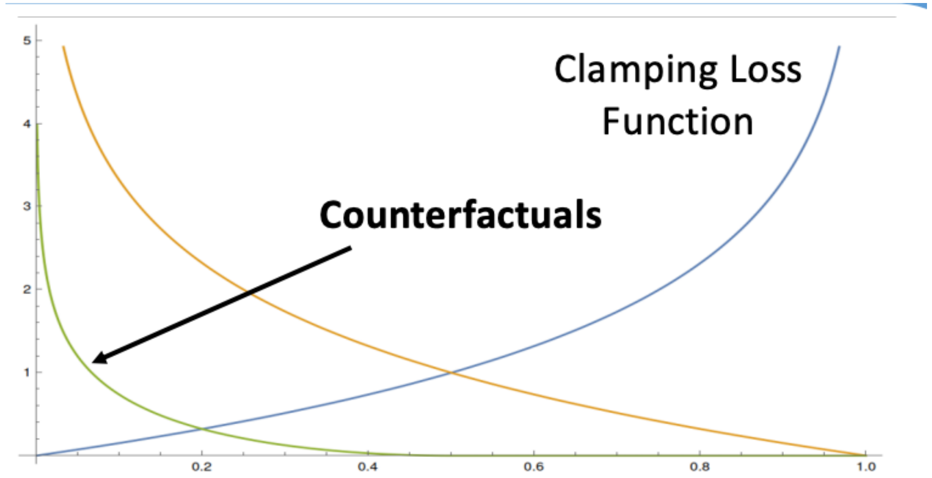
Our Proposed Strategy:
Counterfactual Clamping Attack
(CCA)



CCA Strategy: Unique Loss Function to Clamp Counterfactuals From One Side and Mitigate the Decision Boundary Shift Issue





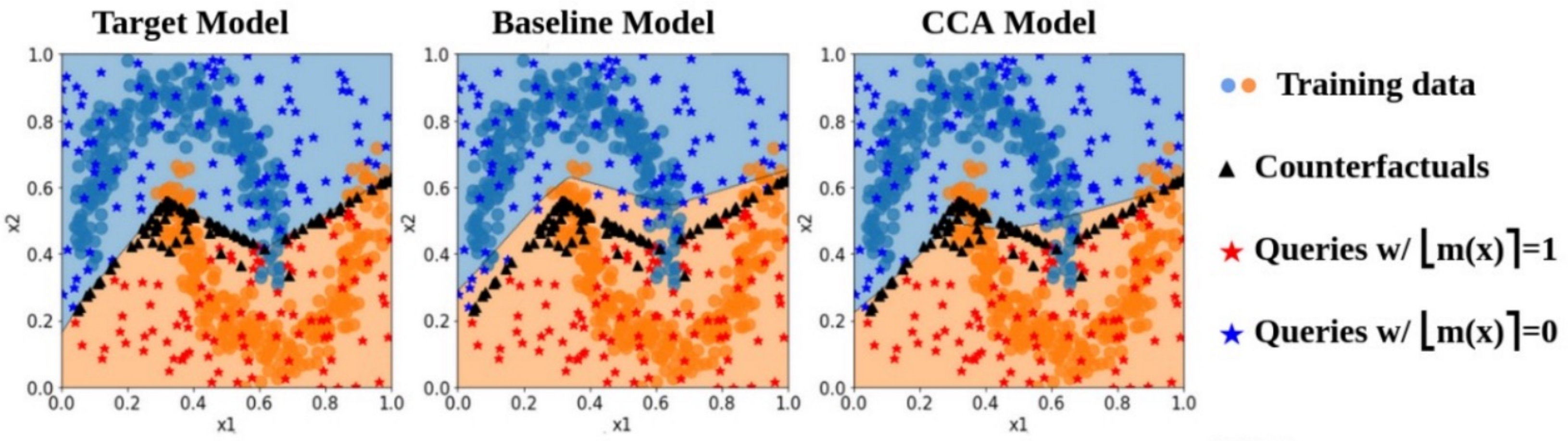


$$L_k(\tilde{m}(\boldsymbol{x}), y_{\boldsymbol{x}}) = \mathbb{1}\left[y_{\boldsymbol{x}} = 0.5, \boxed{\tilde{m}(\boldsymbol{x}) \le k}\right]\{L(\tilde{m}(\boldsymbol{x}), k) - h(k)\} + \mathbb{1}\left[y_{\boldsymbol{x}} \neq 0.5\right] L(\tilde{m}(\boldsymbol{x}), y_{\boldsymbol{x}})$$

neglect CFs that already have g(w) > k

for CFs

for normal examples



Experimental Validation:
Fidelity Comparison Over Several Benchmark Datasets

Dataset	Architecture known (model 0)				Architecture unknown (model 1)			
	$\mathbb{D}_{\text{test}}$		$\mathbb{D}_{\text{uni}}$		$\mathbb{D}_{\text{test}}$		$\mathbb{D}_{\text{uni}}$	
	Base.	CCA	Base.	CCA	Base.	CCA	Base.	CCA
Adult In.	91 ± 3.2	94 ± 3.2	84 ± 3.2	91 ± 3.2	91 ± 4.5	94 ± 3.2	84 ± 3.2	90 ± 3.2
COMPAS	92 ± 3.2	96 ± 2.0	94 ± 1.7	96 ± 2.0	91 ± 8.9	96 ± 3.2	94 ± 2.0	94 ± 8.9
DCCC	89 ± 8.9	99 ± 0.9	95 ± 2.2	96 ± 1.4	90 ± 7.7	97 ± 4.5	95 ± 2.2	95 ± 11.8
HELOC	91 ± 4.7	96 ± 2.2	92 ± 2.8	94 ± 2.4	90 ± 7.4	95 ± 5.5	91 ± 3.3	93 ± 3.2

CCA provides high-fidelity model reconstruction

Comparison With Two-Sided Counterfactuals

Dataset		Architecture known (model 0)								Architecture unknown (model 1)							
		$\mathbb{D}_{\text{test}}$				$\mathbb{D}_{\text{uni}}$				$\mathbb{D}_{\text{test}}$				$\mathbb{D}_{\text{uni}}$			
		Base.	Dual.	CCA1	CCA2	Base.	Dual.	CCA1	CCA2	Base.	Dual.	CCA1	CCA2	Base.	Dual.	CCA1	CCA2
DCCC	n=100	0.95	0.99	0.94	0.99	0.90	0.95	0.92	0.97	0.92	0.98	0.93	0.98	0.88	0.92	0.89	0.93
	n=200	0.96	0.99	0.98	0.99	0.90	0.96	0.95	0.98	0.96	0.99	0.96	0.99	0.89	0.94	0.94	0.96
HELOC	n=100	0.94	0.97	0.90	0.98	0.91	0.98	0.84	0.98	0.92	0.91	0.90	0.96	0.88	0.92	0.84	0.96
	n=200	0.96	0.98	0.92	0.98	0.93	0.98	0.89	0.99	0.95	0.92	0.91	0.97	0.93	0.94	0.88	0.97

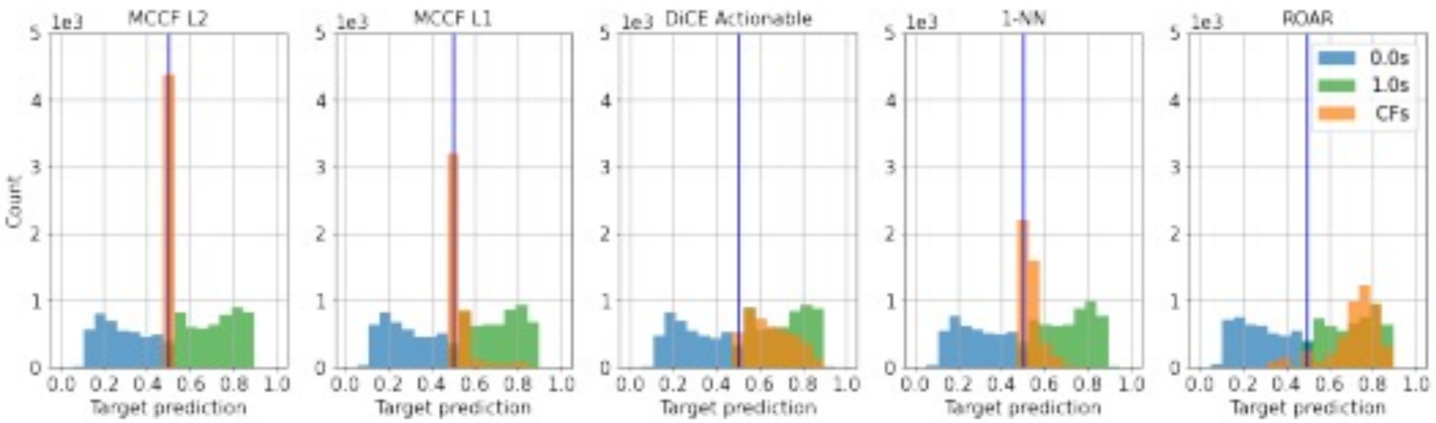
Baselines: *[Aivodji et al.'20][Wang et al.'22]*

Additional Experiments

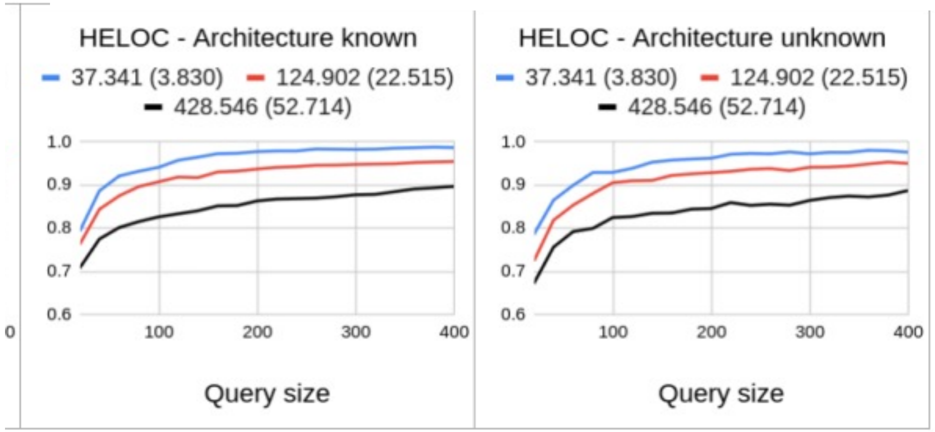
Other Counterfactual Generation Techniques

Table 2: Fidelity achieved with different counterfactual generating methods on HELOC dataset. Target model has hidden layers with neurons (20, 30, 10). Surrogate model architecture is (10, 20).

CF method	Fidelity over $\mathbb{D}_{test}$				Fidelity over $\mathbb{D}_{uni}$			
	n=100		n=200		n=100		n=200	
	Base.	CCA	Base.	CCA	Base.	CCA	Base.	CCA
MCCF L2-norm	91	95	93	96	91	93	93	95
MCCF L1-norm	93	95	94	96	89	92	91	95
DiCE Actionable	93	94	95	95	90	91	93	94
1-Nearest-Neightbor	93	95	94	96	93	93	94	95
ROAR [Upadhyay et al., 2021]	91	92	93	95	87	85	92	92
C-CHVAE [Pawelczyk et al., 2020]	77	80	78	82	90	89	85	78



Different Lipschitz Constants



Different Model Architectures

Dataset: HELOC - Fidelity over $\mathbb{D}_{test}$

Target archi. → / Surrogate archi. ↓	(20,10)				(20,10,5)				(20,20,10,5)			
	n=100		n=200		n=100		n=200		n=100		n=200	
	Base.	CCA	Base.	CCA	Base.	CCA	Base.	CCA	Base.	CCA	Base.	CCA
(20,10)	0.90	0.94	0.91	0.95	0.90	0.94	0.92	0.95	0.98	0.99	0.98	0.99
(20,10,5)	0.88	0.92	0.92	0.95	0.89	0.92	0.92	0.95	0.98	0.98	0.98	0.99
(20,20,10,5)	0.87	0.93	0.91	0.93	0.87	0.89	0.91	0.94	0.98	0.98	0.98	0.98

Dataset: HELOC - Fidelity over $\mathbb{D}_{uni}$

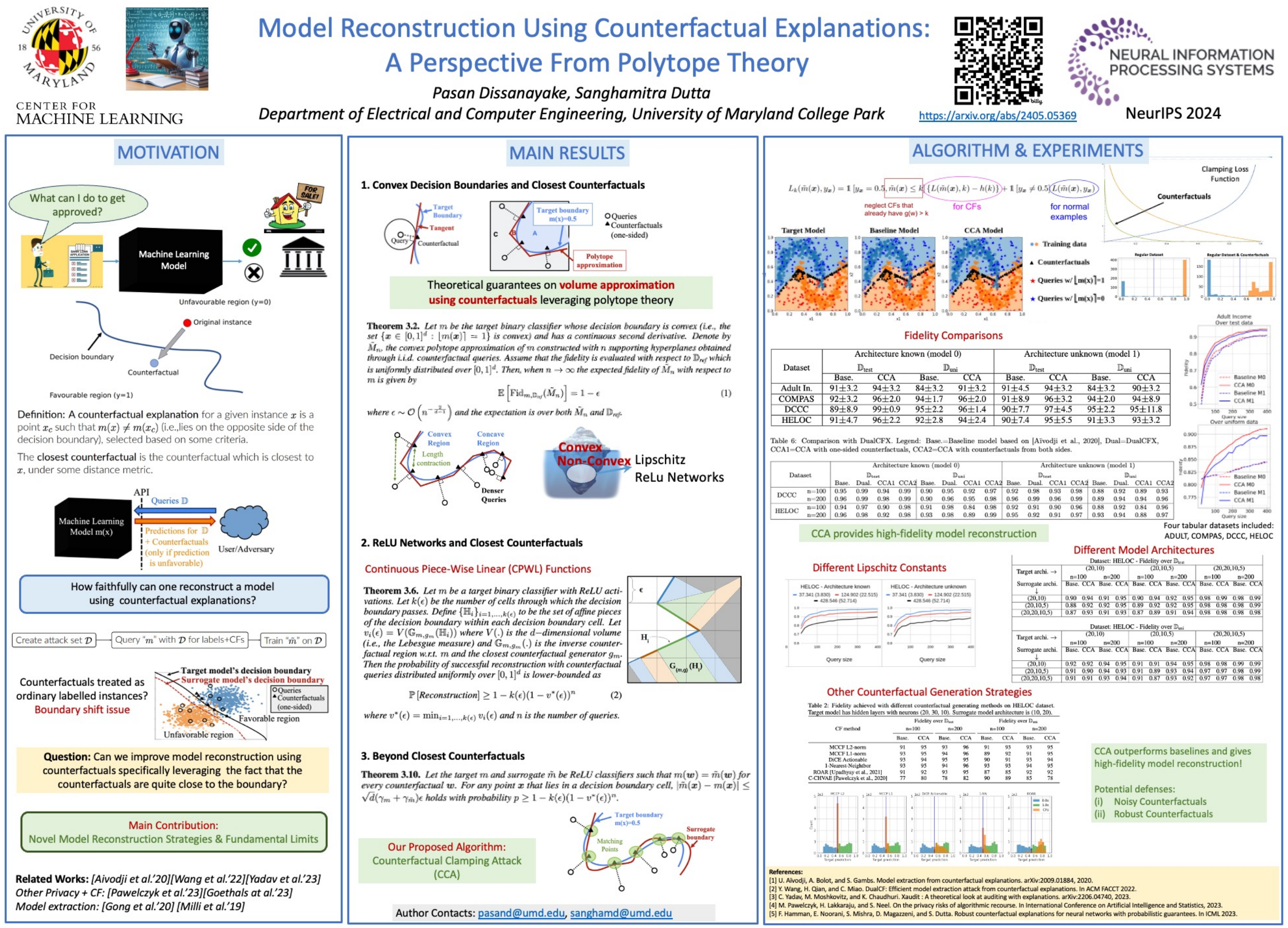
Target archi. → / Surrogate archi. ↓	(20,10)				(20,10,5)				(20,20,10,5)			
	n=100		n=200		n=100		n=200		n=100		n=200	
	Base.	CCA	Base.	CCA	Base.	CCA	Base.	CCA	Base.	CCA	Base.	CCA
(20,10)	0.92	0.92	0.94	0.95	0.91	0.91	0.94	0.95	0.98	0.98	0.99	0.99
(20,10,5)	0.91	0.90	0.94	0.93	0.91	0.89	0.93	0.94	0.97	0.97	0.98	0.99
(20,20,10,5)	0.91	0.91	0.93	0.94	0.91	0.87	0.93	0.92	0.97	0.97	0.98	0.98

CCA mostly outperforms baselines and gives high-fidelity model reconstruction!

Potential defenses: (i) Noisy Counterfactuals, or (ii) Robust Counterfactuals

Thank You!

Poster: Wed 11 Dec 4:30 p.m. PST — 7:30 p.m. PST



https://arxiv.org/abs/2405.05369

https://github.com/pasandissanayake/model-reconstruction-using-counterfactuals

Author Contacts:
pasand@umd.edu, sanghamd@umd.edu

Broader Implications on the Interplay Between Explainability & Privacy