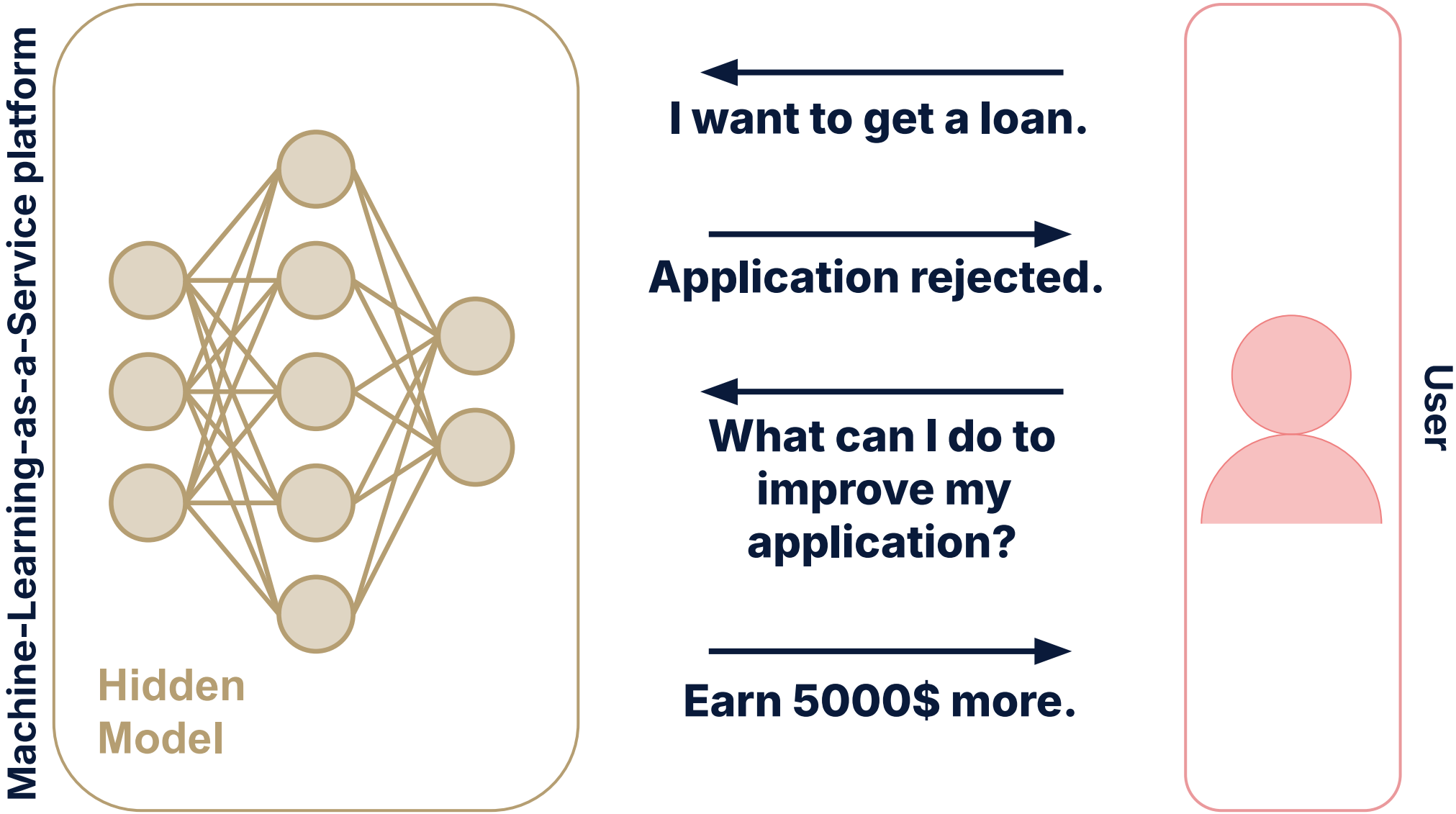
RECAST: Model Reconstruction via Counterfactual-Aware Wasserstein Geometry under Limited Data

Xuan Zhao*[1], Lena Krieger*[1,2], Zhuo Cao*[1], Arya Bangun[1], Hanno Scharr[1], Ira Assent[1,3]

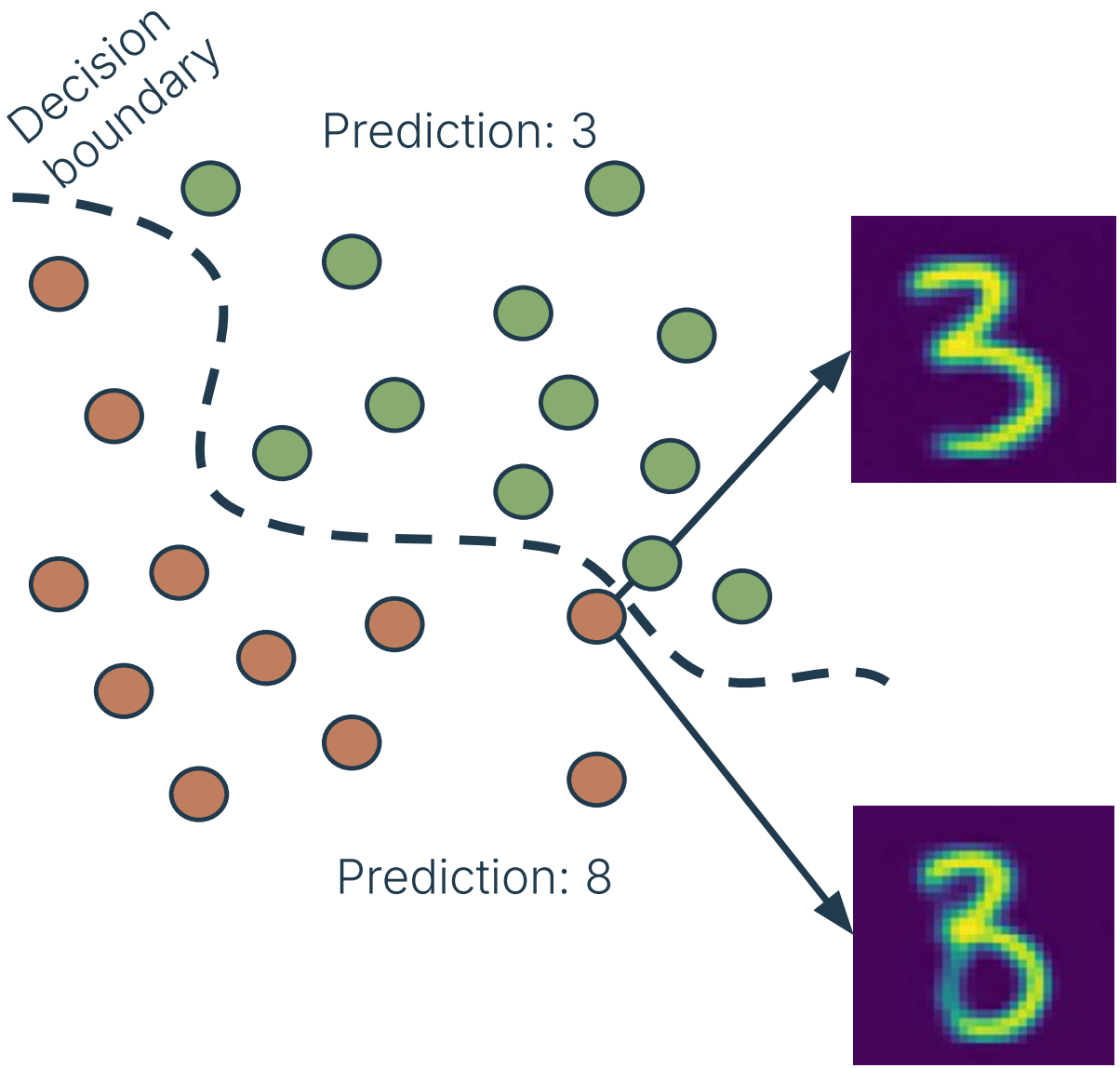
International Conference on Machine Learning 2026

[1] IAS-8, Forschungszentrum Jülich
[2] LMU Munich, MCML
[3] Department of Computer Science, Aarhus University
*Equal Contribution.

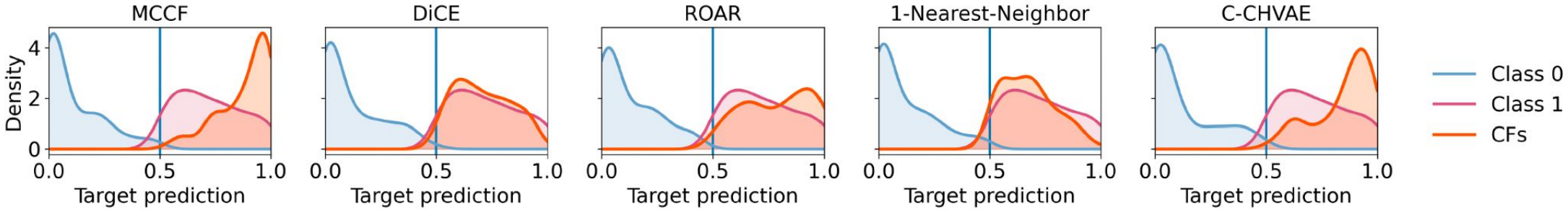
Model reconstruction is important to understand model behaviour and improve accountability.

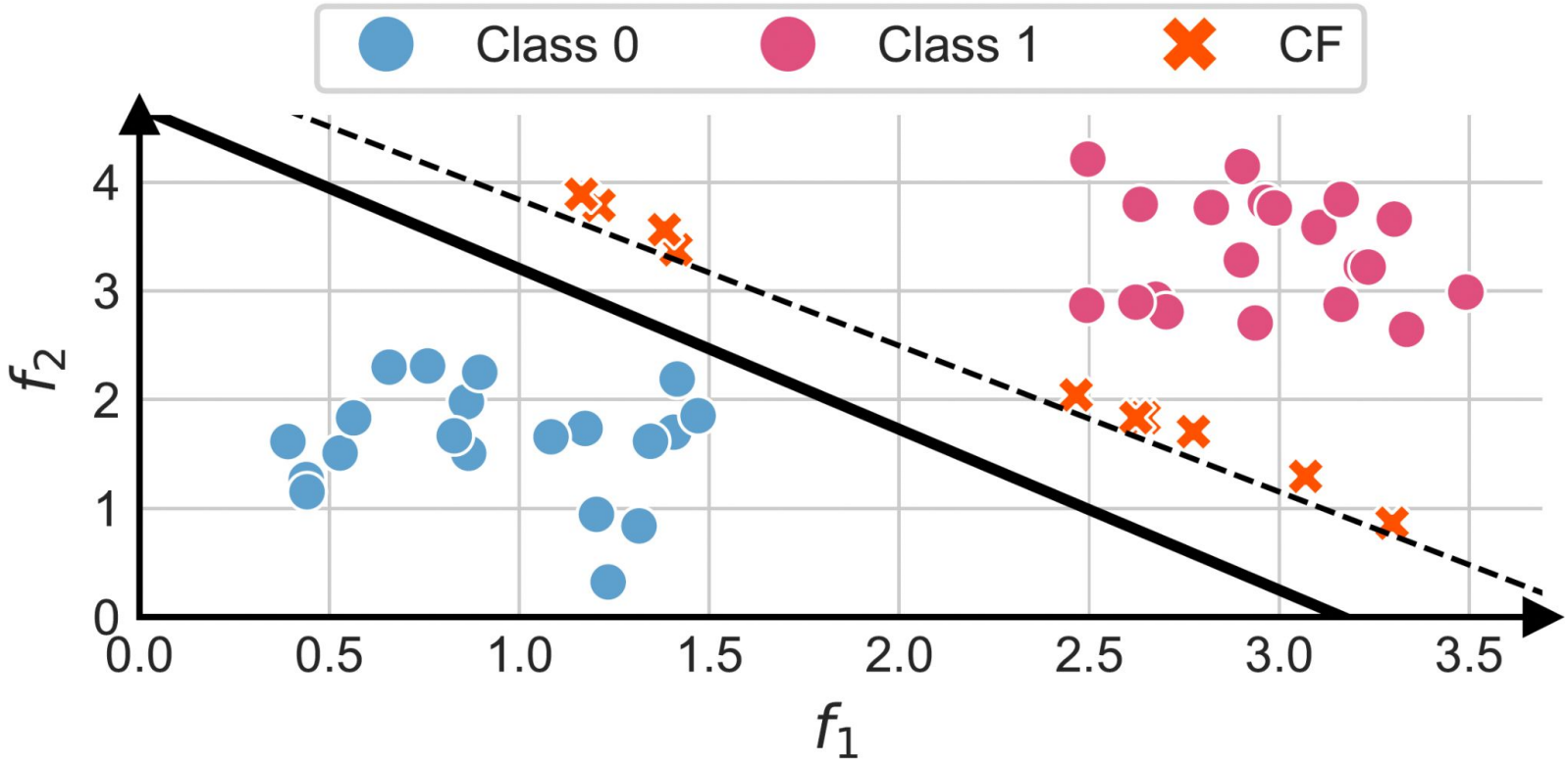


Counterfactual Explanations (CFs) can help model reconstruction.

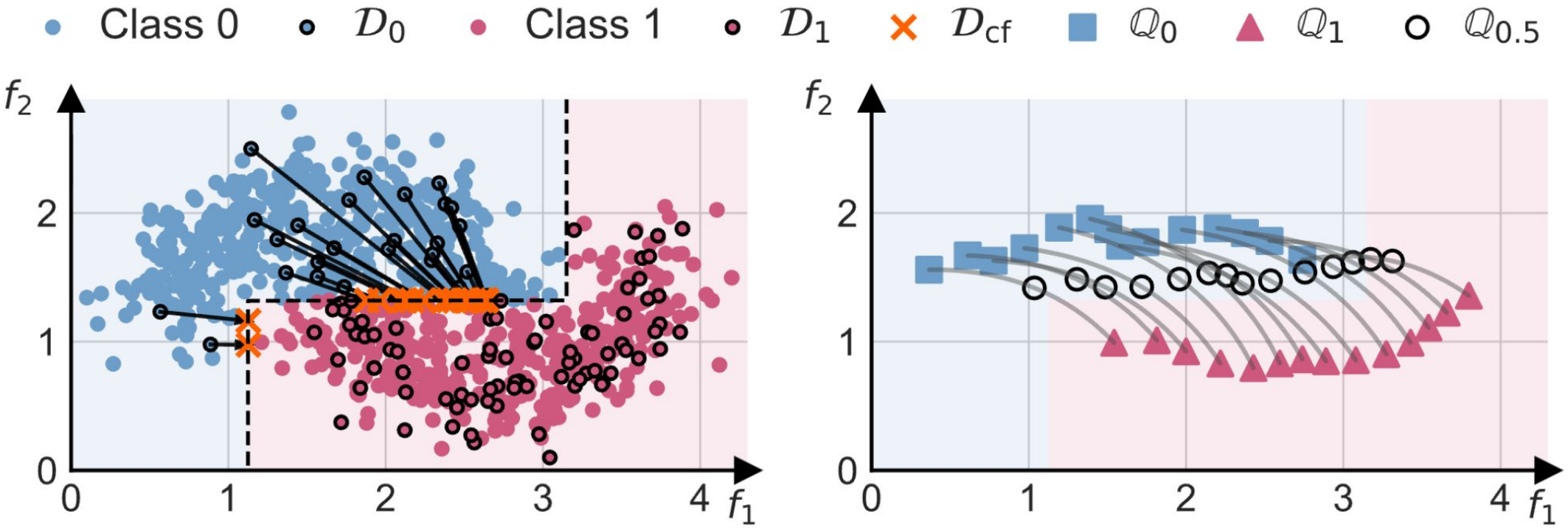


Current related works face critical limitations.



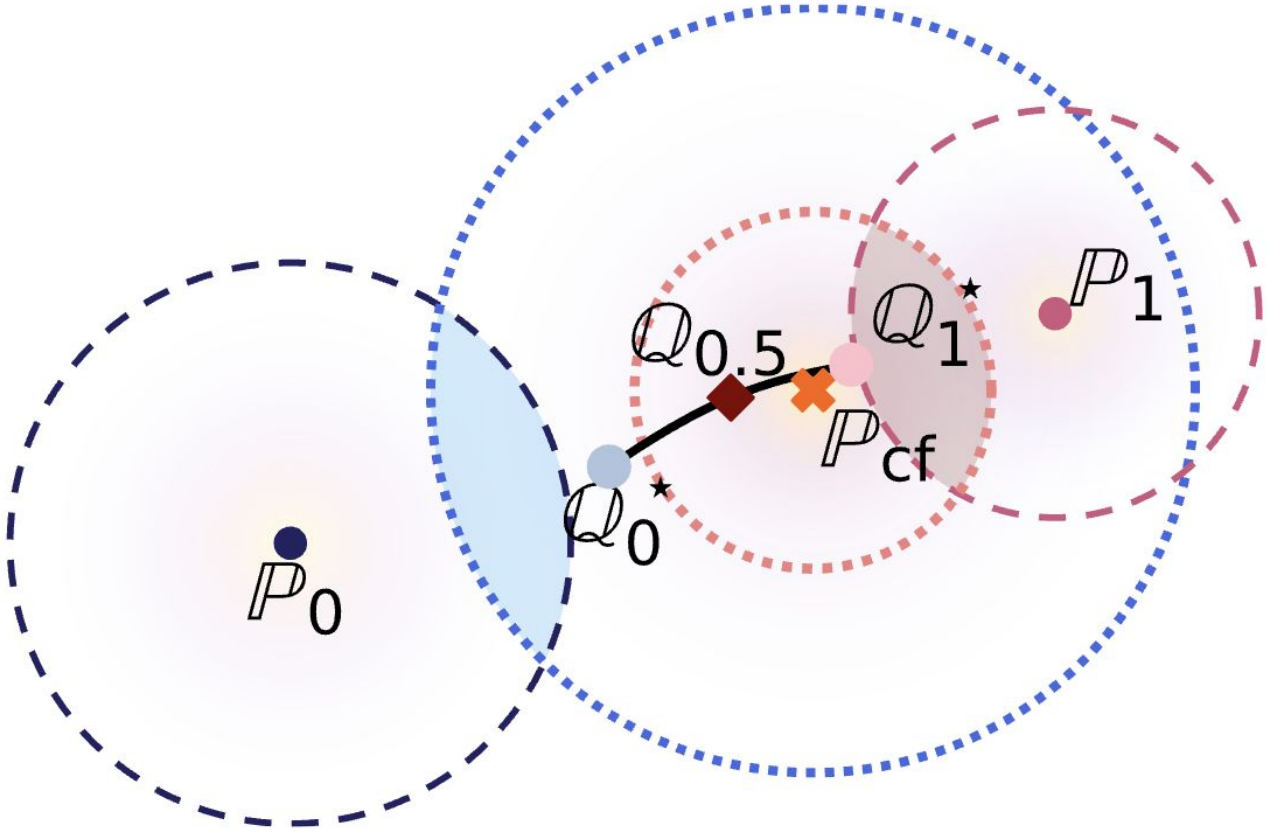


We propose to leverage CFs as soft samples for both classes.



We construct counterfactual-aware Wasserstein barycenters as prototypes.

$$\mathbb{Q}_c^\star = \underset{\mu \in P_2(\mathscr{X})}{\operatorname{argmin}} \Big((1-\lambda_c) W_2^2(\mu, \mathbb{P}_c) + \lambda_c W_2^2(\mu, \mathbb{P}_{\mathrm{cf}}) \Big)$$



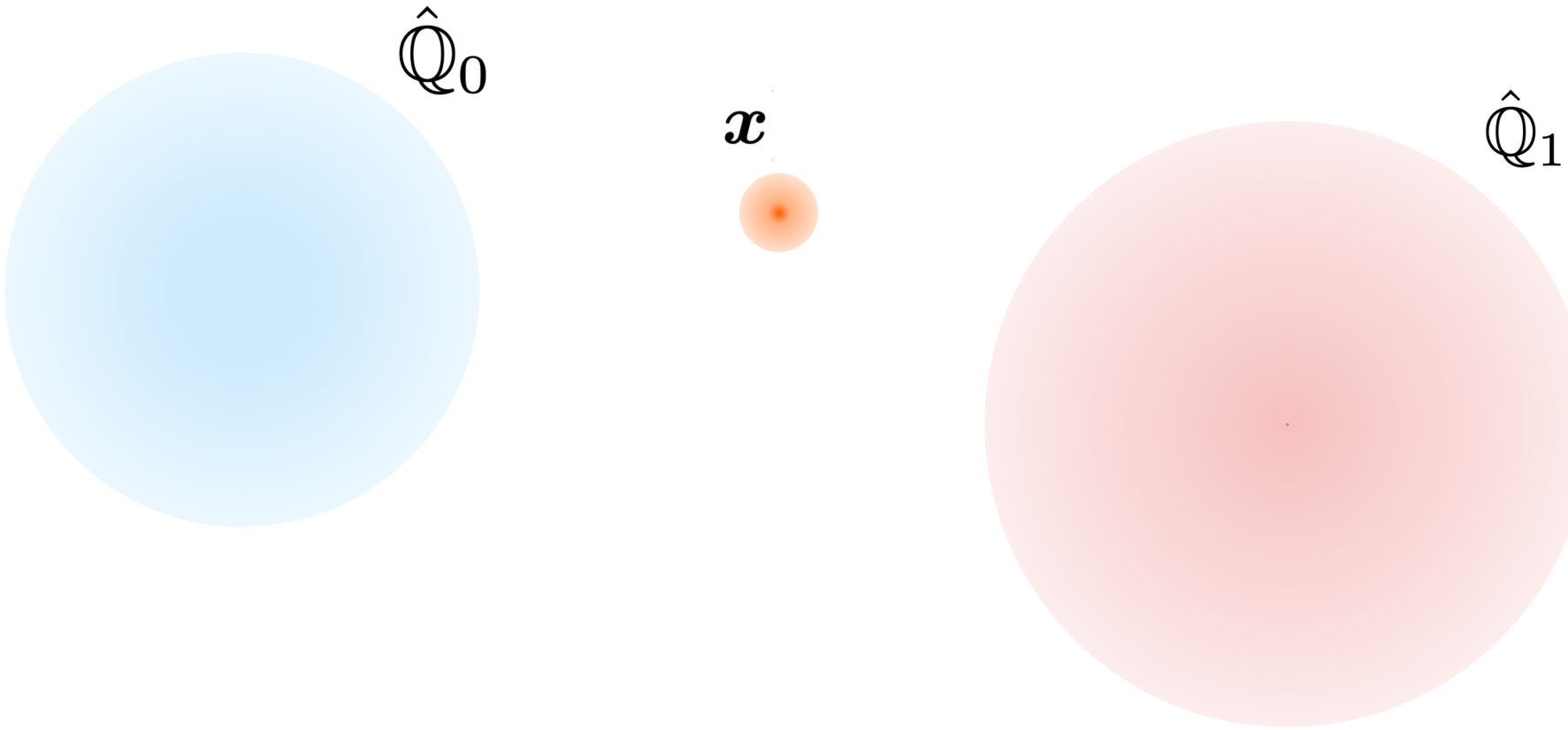
$$\lambda_c = \frac{B}{A+2B}$$

$$A = W_2^2(\mathbb{P}_{\mathrm{cf}}, \mathbb{P}_c)$$

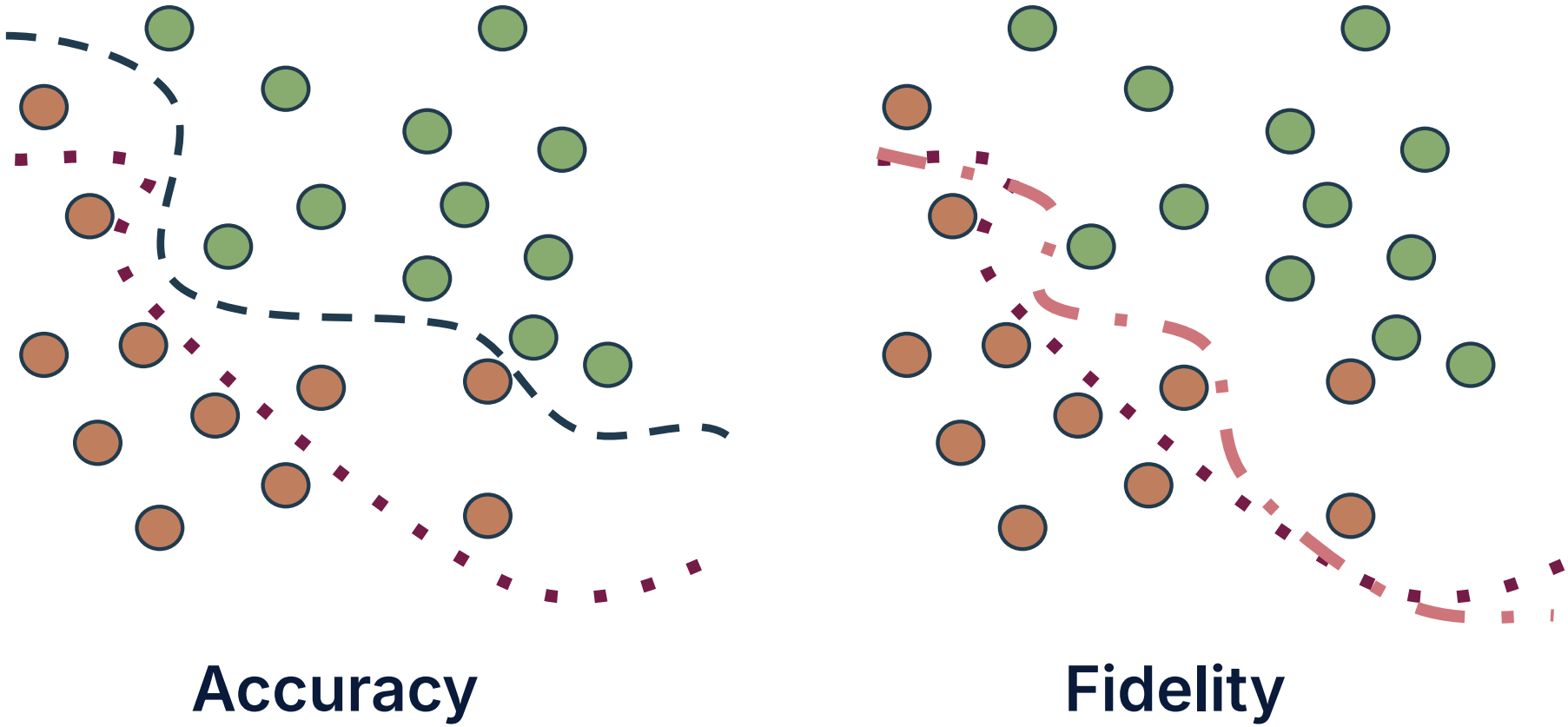
$$B = W_2^2(\mathbb{P}_{\mathrm{cf}}, \mathbb{P}_{1-c})$$

Classification is done by comparing distances between input and prototypes.

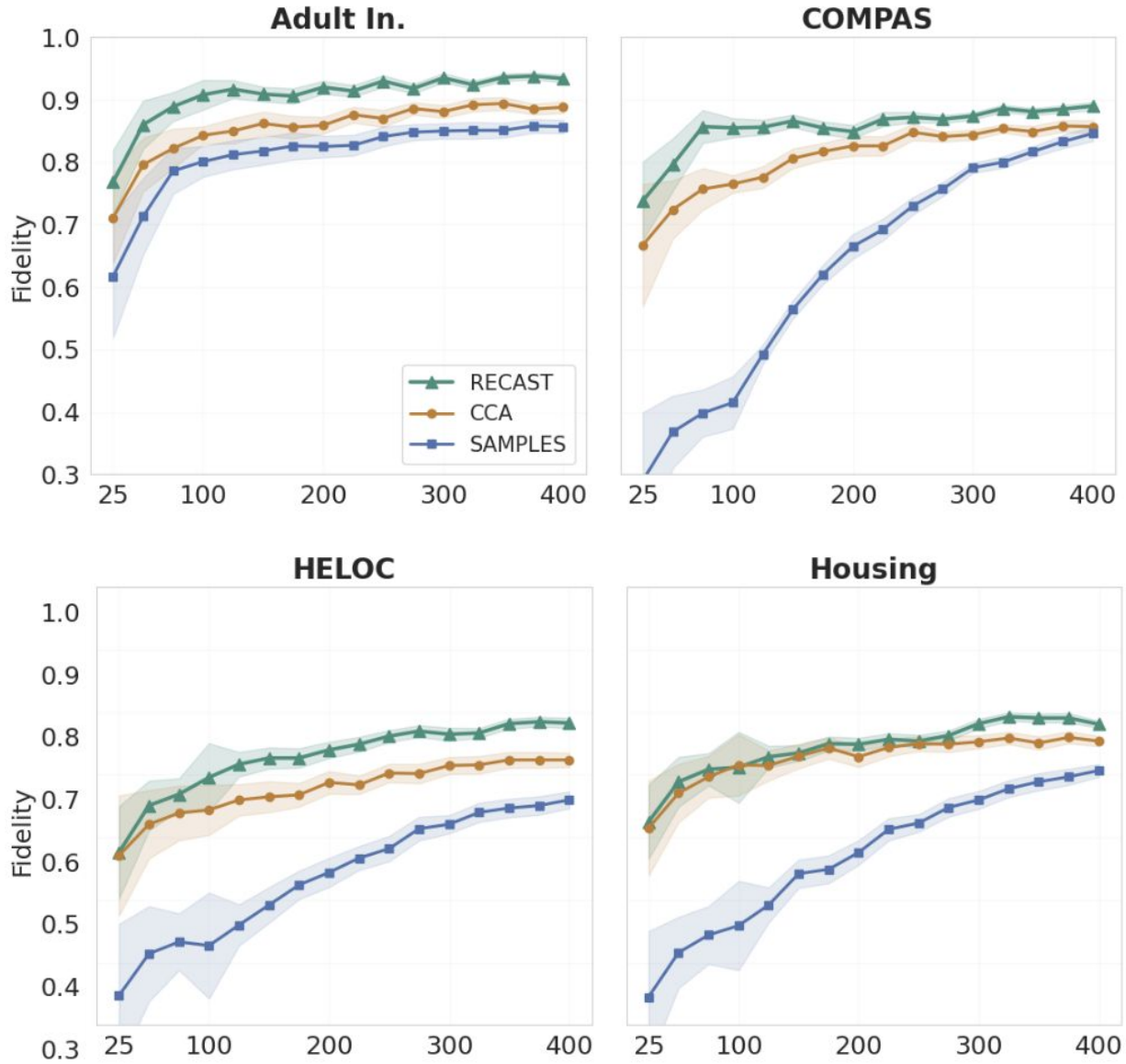
$$\hat{m}(\boldsymbol{x}) := \arg\min_{c \in \{0,1\}} W_2\left(\delta_{\boldsymbol{x}}, \hat{\mathbb{Q}}_c\right)$$



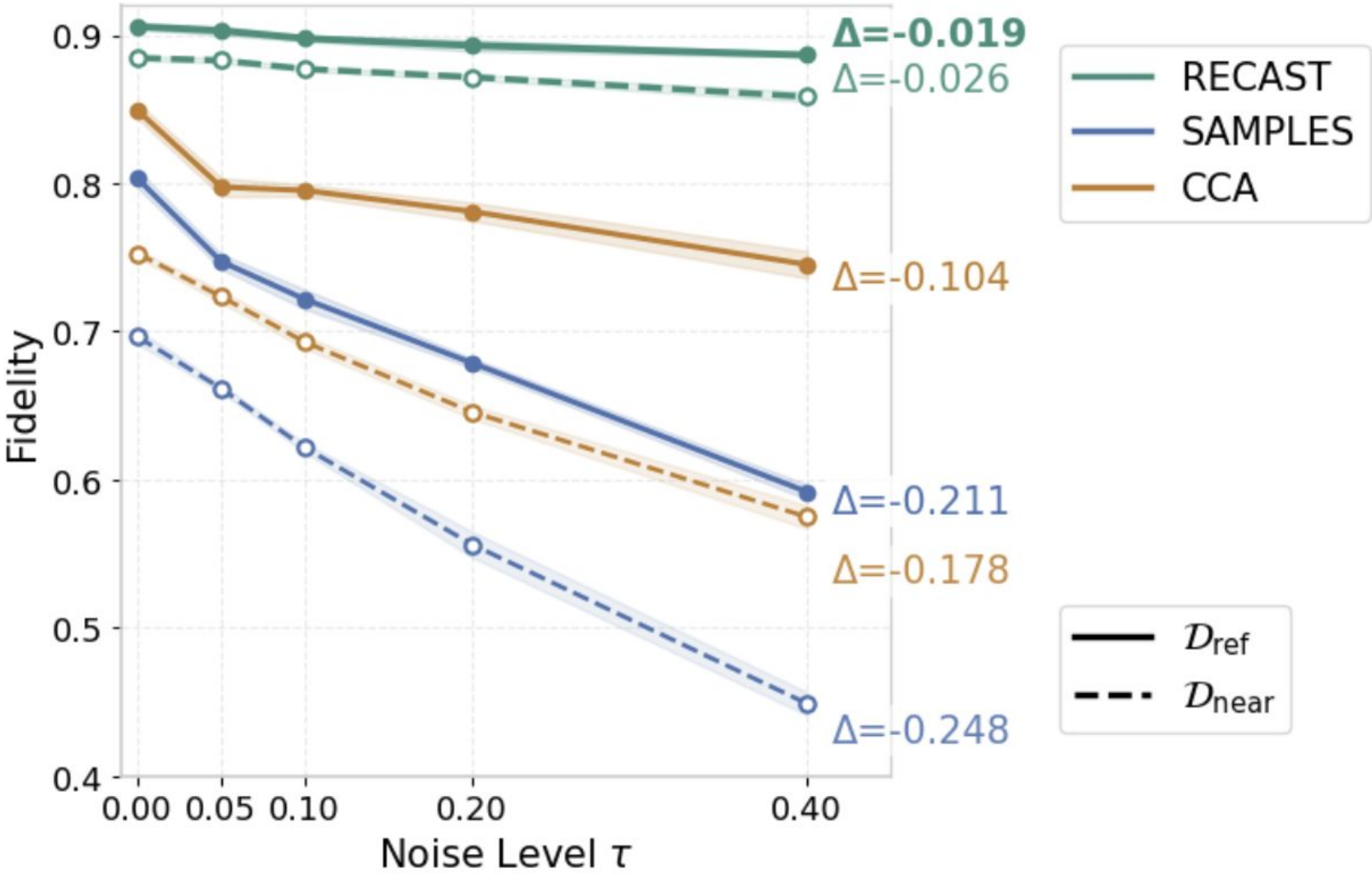
Fidelity and accuracy do not capture the same.



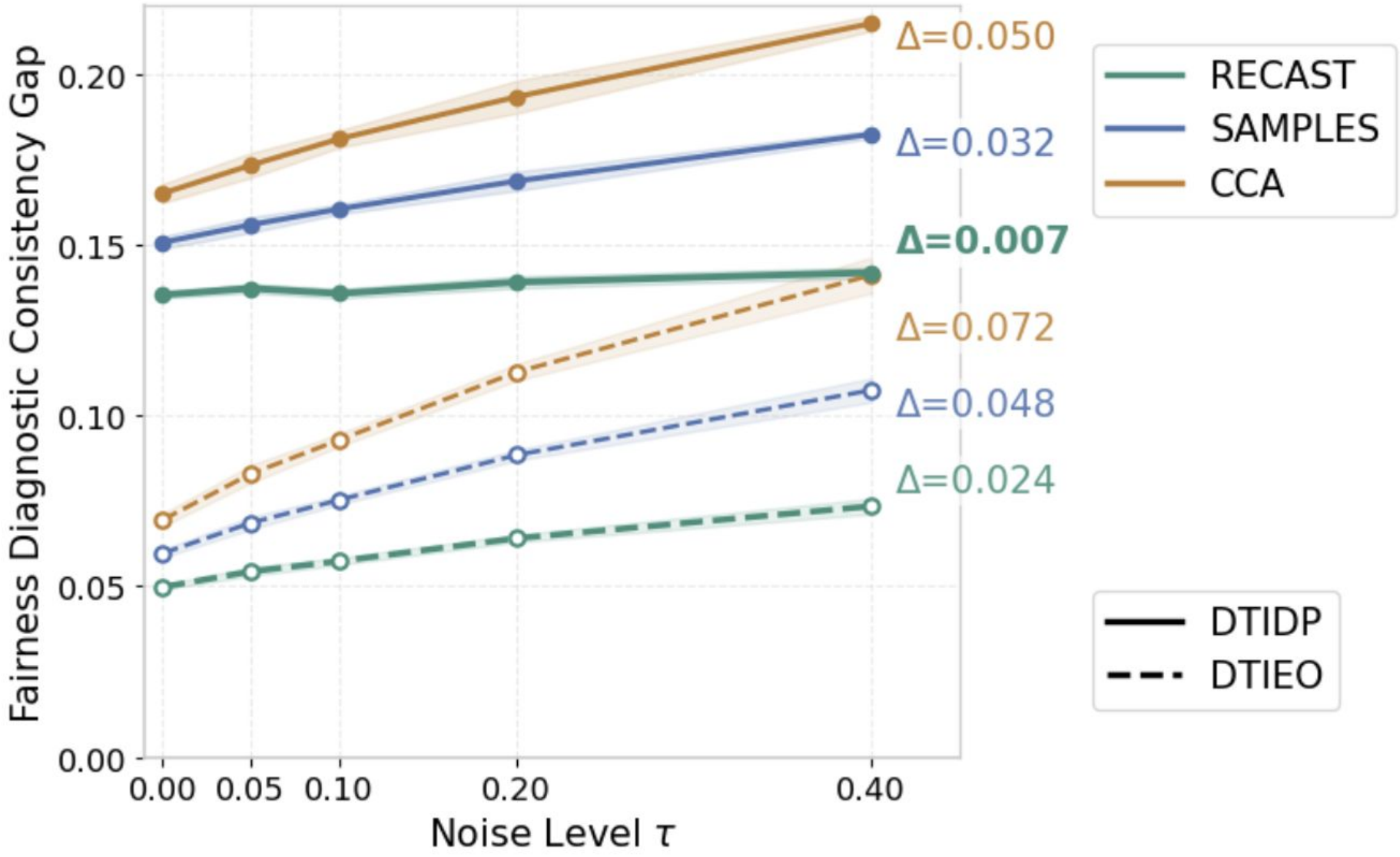
Distributional prototypes achieve high fidelity under low data regimes.



Distributional prototypes average out noise.



We adopt TIEO and TIDP [2] to enable fairness diagnostic.



RECAST: Model Reconstruction via Counterfactual-Aware Wasserstein Geometry

Xuan Zhao
x.zhao@fz-juelich.de
Lena Krieger
l.krieger@fz-juelich.de
Zhuo Cao
z.cao@fz-juelich.de
Arya Bangun
a.bangun@fz-juelich.de
Hanno Scharr
h.scharr@fz-juelich.de
Ira Assent
ira@cs.au.dk

Code

Paper

RECAST creates a surrogate model based on **Wasserstein prototypes. RECAST:**

- captures relationships between data distributions,
- forms **robust class prototypes** that effectively represent both labeled data and CFs,
- under **one-sided CF access** and **limited query budgets**.

References

[1] Aïvodji, Ulrich, Alexandre Bolot, and Sébastien Gambs. "Model extraction from counterfactual explanations." arXiv preprint arXiv:2009.01884 (2020).

[2] Chen, Mingliang, and Min Wu. "Towards threshold invariant fair classification." Conference on Uncertainty in Artificial Intelligence. PMLR, 2020.