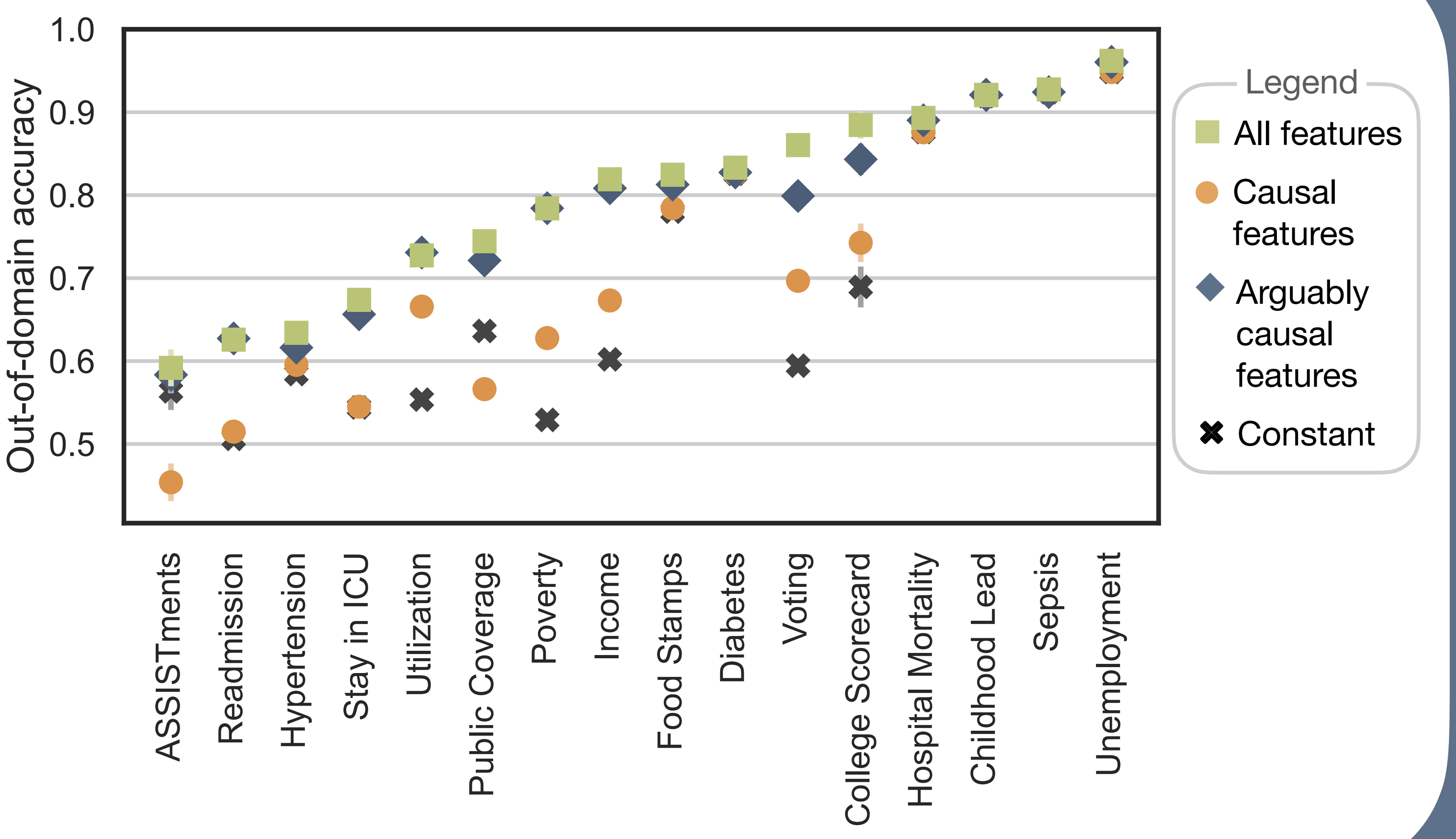
Do causal predictors generalize better to new domains?

Vivian Y. Nastl *Moritz Hardt*

Predictors using all available features, regardless of causality, have better in-domain and out-of-domain accuracy than predictors using causal features.



Introduction

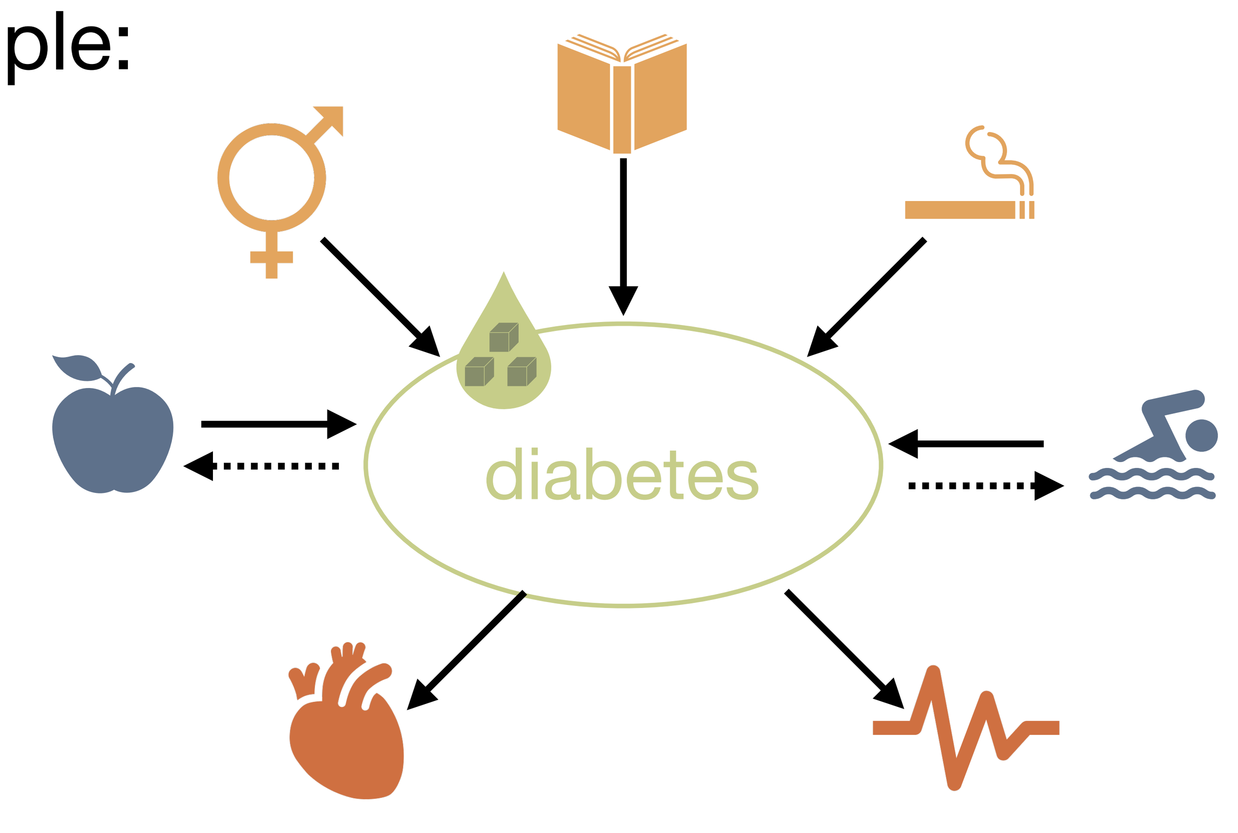
- Our goal is to test the hypothesis that models trained on causal features generalize better across domains
- We consider 16 prediction tasks on tabular datasets covering applications in health, employment, education, social benefits, and politics

Method

We propose a pragmatic scheme to classify the relationship between features and target

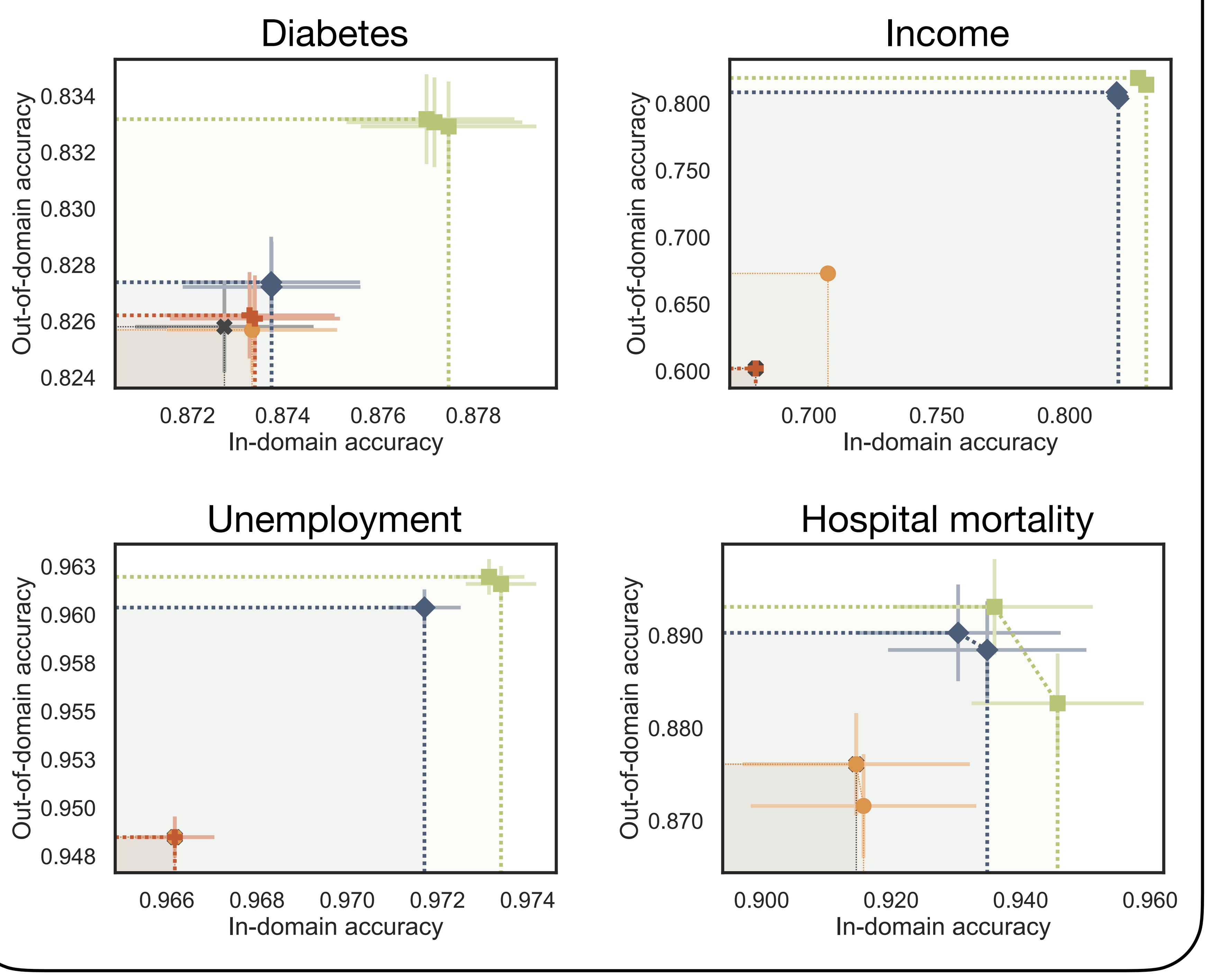
- Causal features: we most strongly believe that they have a causal influence on the target
- Arguably causal features: may be considered causal depending on modeling choices (*inclusive* selection)
- Anti-causal: we most strongly believe that the target has a causal influence on them

Example:



Results

We train a total of 42K models for the main results + 460K models for robustness tests.



Discussion

- We find no evidence that causal predictors have greater external validity than their conventional counterparts.
- If the goal is to generalize to new domains, we might as well train the best possible model on all available features.
- Demonstrating the utility of causal methods likely requires other benchmark datasets than the ones currently available.

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS
ETH zürich

References:
Gardner, J., Popovic, Z., and Schmidt, L. Benchmarking distribution shift in tabular data with tableshift. In Proceedings of the 37th Conference on Neural Information Processing Systems, NeurIPS 2023, 2023.
Hardt, M. and Kim, M. P. Is your model predicting the past? In Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '23, 2023.

Scan the QR code to download the **full paper**