UNIVERSITY OF CAMBRIDGE

Counterfactual Accuracies for Alternative Models

Umang Bhatt, Adrian Weller, Muhammad Bilal Zafar, Krishna Gummadi

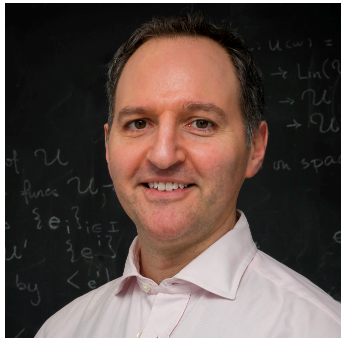
ML-IRL Workshop at ICLR 2020

About Us

Umang Bhatt

University of Cambridge
usb20@cam.ac.uk

Adrian Weller



University of Cambridge
The Alan Turing Institute
aw665@cam.ac.uk

Muhammad Bilal Zafar

Bosch Center for AI
mzafar@mpi-sws.org

Krishna Gummadi

MPI-SWS
gummadi@mpi-sws.org

Loan Approval

Loan Approval

Model A

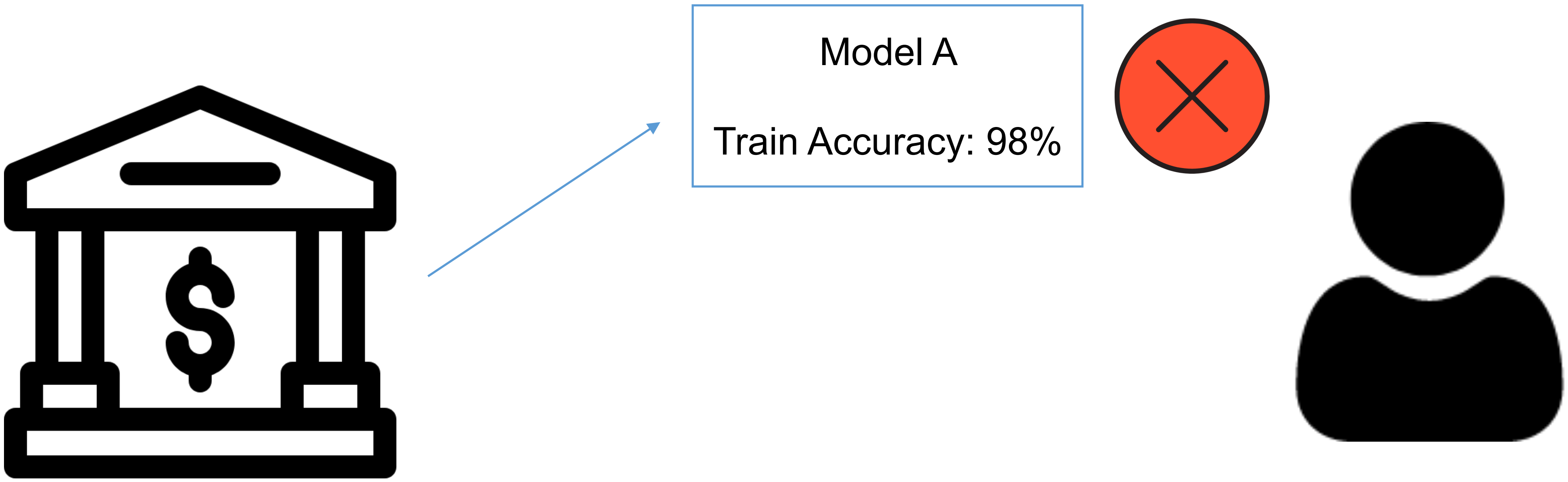
Train Accuracy: 98%

Loan Approval

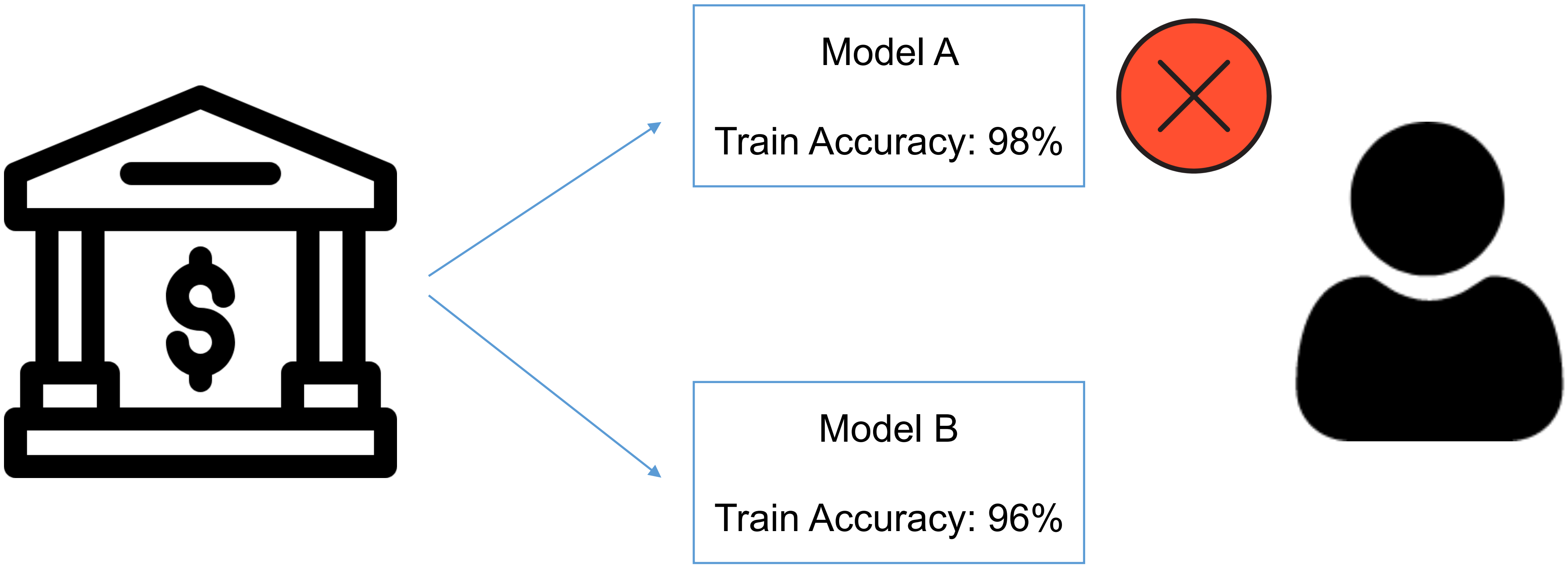
Model A

Train Accuracy: 98%

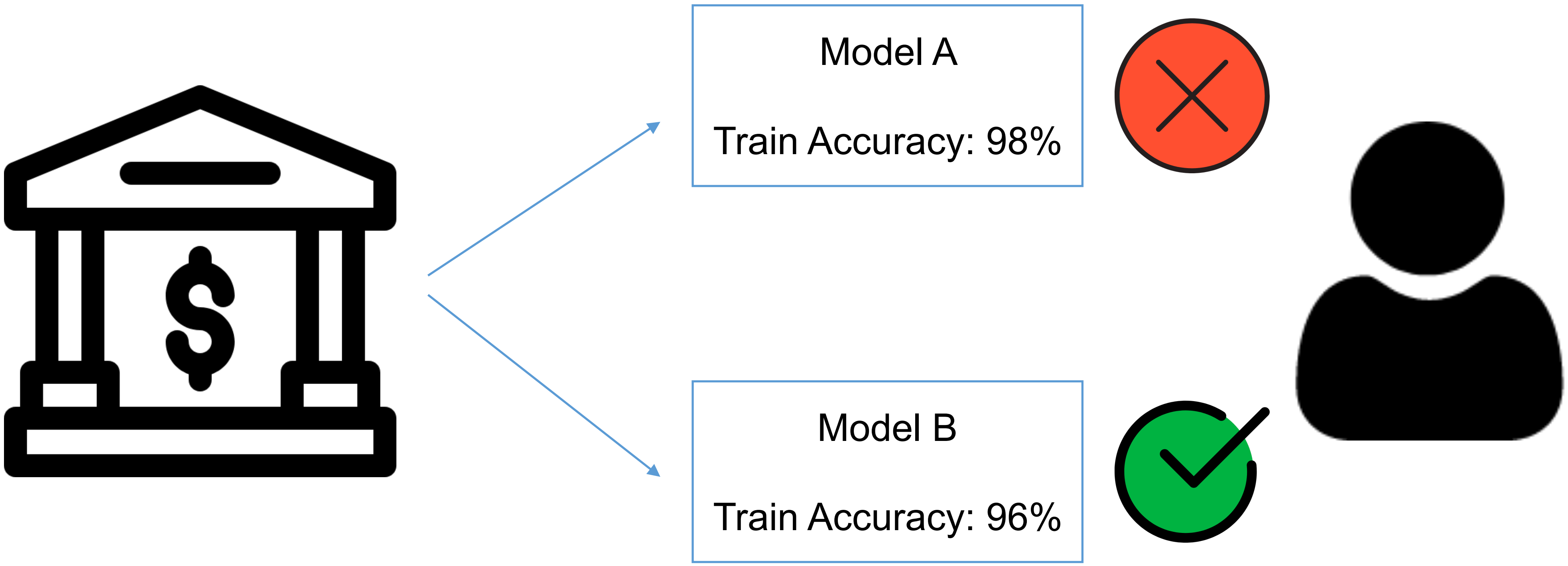
Loan Approval



Loan Approval



Loan Approval



Central Question

Given a particular test point z, if we were to find an alternative classifier in the same model class fitted to the same training data, how much training accuracy would we have to give up so that the prediction for the test point z would change?

Related Work

Rashomon Effect [Breiman 2001]: Multiple models may fit the training data well

Related Work

Rashomon Effect [Breiman 2001]: Multiple models may fit the training data well

Rashomon Set [Fisher et al. 2019]: Set of models with near-optimal accuracy

Related Work

Rashomon Effect [Breiman 2001]: Multiple models may fit the training data well

Rashomon Set [Fisher et al. 2019]: Set of models with near-optimal accuracy

Predictive Multiplicity [Marx et al. 2019]: Analyzes the difference in predictions from models in a Rashomon set

Notation

$$\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{n}$$

Training Dataset

$$\mathscr{F}$$

Family of Functions

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f(x_i), y_i\right)$$

Average Loss (Empirical Risk)

Our Approach

Empirical Risk Minimization

$$f_{\mathrm{o}} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \ell\left(f(\boldsymbol{x}_i), y_i\right)$$

Our Approach

Empirical Risk Minimization

$$f_{\mathrm{o}} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \ell\left(f(\boldsymbol{x}_i), y_i\right)$$

This Work

$$f_{\boldsymbol{z}} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \ell\left(f(\boldsymbol{x}_i), y_i\right)$$
$$\text{s.t. } f_{\mathrm{o}}(z) \neq f_{\boldsymbol{z}}(z)$$

Our Approach

Empirical Risk Minimization

$$f_{\mathrm{o}} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \ell\left(f(\boldsymbol{x}_i), y_i\right)$$

This Work

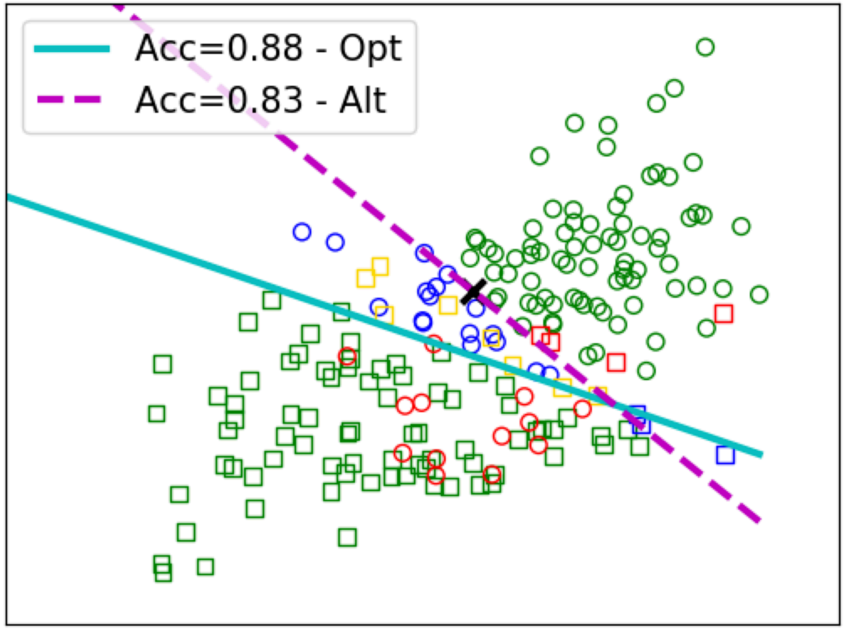
$$f_{\boldsymbol{z}} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} \ell\left(f(\boldsymbol{x}_i), y_i\right)$$

$$\text{s.t. } f_{\mathrm{o}}(z) \neq f_{\boldsymbol{z}}(z)$$

Counterfactual Accuracy: $\tilde{C}_{\boldsymbol{z}} = \widehat{R}(f_{\boldsymbol{z}}) - \widehat{R}(f_{\mathrm{o}}) = \left(1 - \widehat{R}(f_{\mathrm{o}})\right) - \left(1 - \widehat{R}(f_{\boldsymbol{z}})\right)$

Some Results

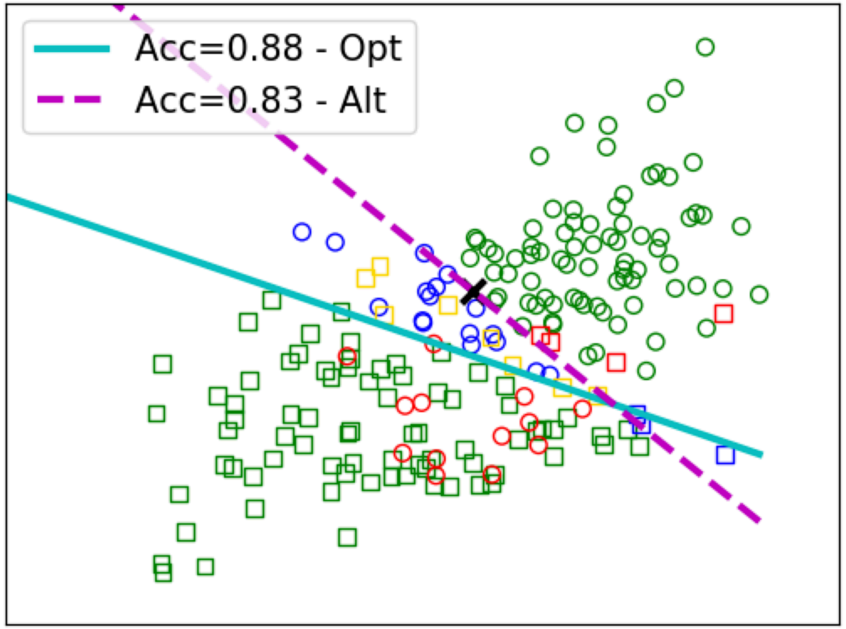
Two Overlapping Gaussians



Green: Correct in both
Red: Incorrect in both
Blue: Originally right, now wrong
Yellow: Originally wrong, now right

Some Results

Two Overlapping Gaussians



Green: Correct in both
Red: Incorrect in both
Blue: Originally right, now wrong
Yellow: Originally wrong, now right

Dataset	Average Counterfactual Accuracy	Average number of predicted label flips
Adult	0.667%	~225
COMPAS	1.437%	~260

Future Work

- **Faster computation**: Moving beyond a warm start from the parameters of the old model, how can avoid recomputing the entire objective from scratch?

Future Work

- **Faster computation**: Moving beyond a warm start from the parameters of the old model, how can avoid recomputing the entire objective from scratch?

- **Experimentation**: What do ML practitioners learn about their datasets from knowing this quantity for their training data?

References

Breiman, Leo. "Statistical modeling: The two cultures (with comments and a rejoinder by the author)." Statistical science 16.3 (2001): 199-231.

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. "All models are wrong but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance." arXiv preprint arXiv:1801.01489 (2018).

Marx, Charles T., Flavio du Pin Calmon, and Berk Ustun. "Predictive Multiplicity in Classification." arXiv preprint arXiv:1909.06677 (2019).