# Solutions for Fairer ML Case Studies

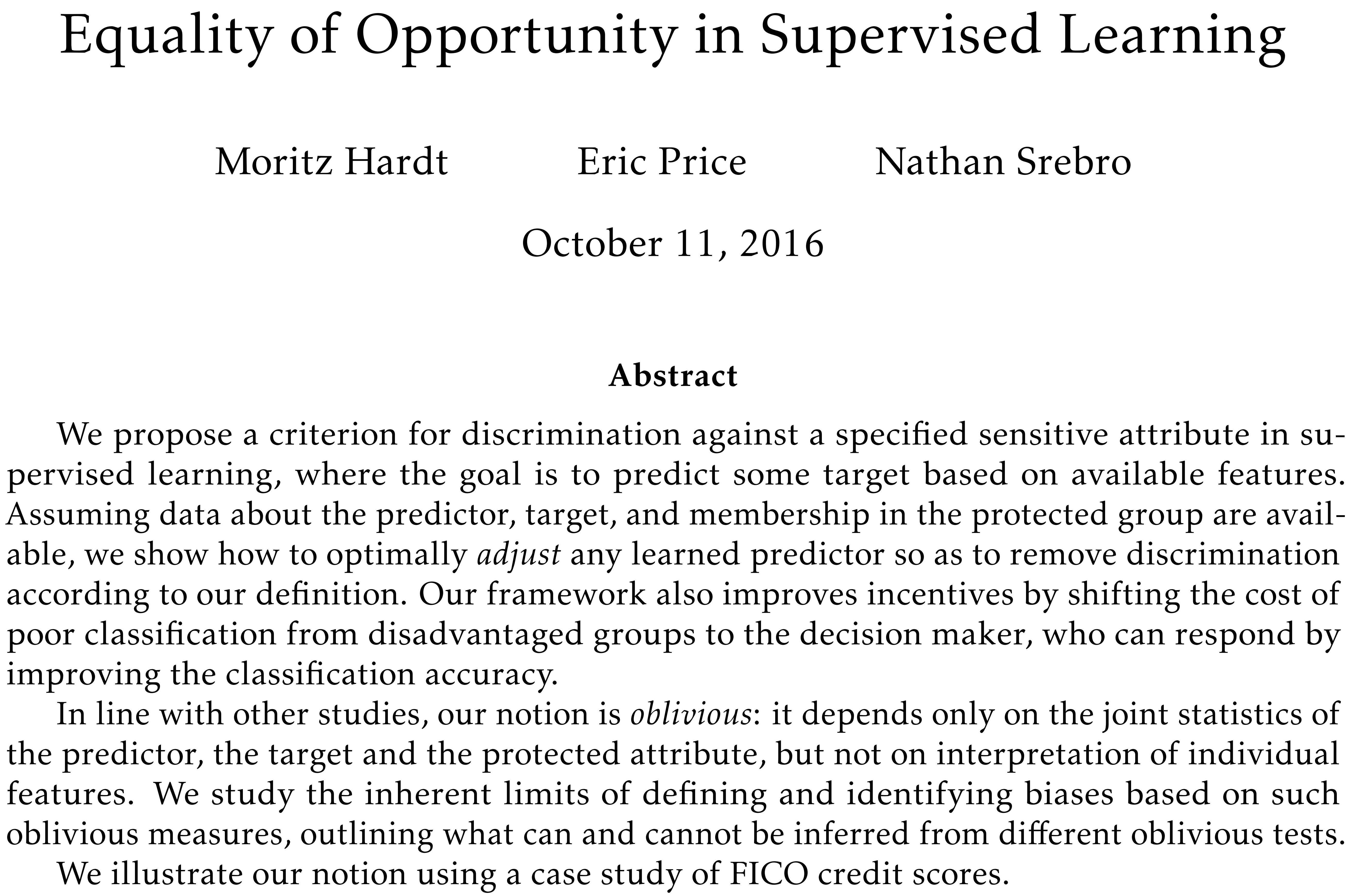
CS4824/ECE4424
Bert Huang

# Plan

- Different forms and causes of fairness in machine learning
- Case studies of recent solutions for fairer ML
    - Post processing predictions for equal opportunity
    - Fair representation learning
    - Fixing feedback loops

# Case Study 1: Equal Opportunity

http://research.google.com/bigpicture/attacking-discrimination-in-ml/



## Equality of Opportunity in Supervised Learning

Moritz Hardt Eric Price Nathan Srebro

October 11, 2016

**Abstract**

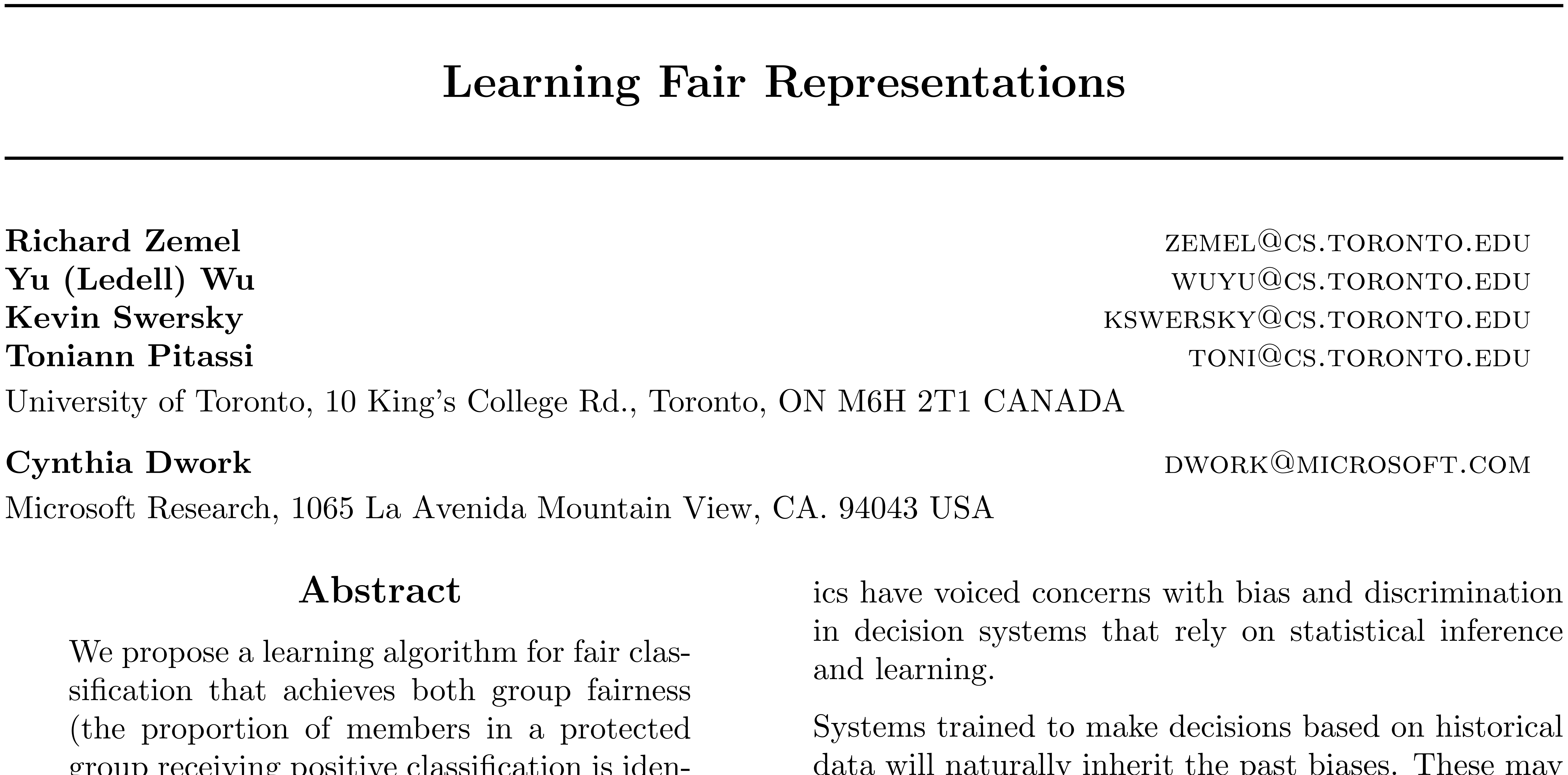
We propose a criterion for discrimination against a specified sensitive attribute in supervised learning, where the goal is to predict some target based on available features. Assuming data about the predictor, target, and membership in the protected group are available, we show how to optimally *adjust* any learned predictor so as to remove discrimination according to our definition. Our framework also improves incentives by shifting the cost of poor classification from disadvantaged groups to the decision maker, who can respond by improving the classification accuracy.

In line with other studies, our notion is *oblivious*: it depends only on the joint statistics of the predictor, the target and the protected attribute, but not on interpretation of individual features. We study the inherent limits of defining and identifying biases based on such oblivious measures, outlining what can and cannot be inferred from different oblivious tests.

We illustrate our notion using a case study of FICO credit scores.

# Case Study 2: Fair Representations

http://proceedings.mlr.press/v28/zemel13.pdf



## Learning Fair Representations

**Richard Zemel** ZEMEL@CS.TORONTO.EDU
**Yu (Ledell) Wu** WUYU@CS.TORONTO.EDU
**Kevin Swersky** KSWERSKY@CS.TORONTO.EDU
**Toniann Pitassi** TONI@CS.TORONTO.EDU
University of Toronto, 10 King's College Rd., Toronto, ON M6H 2T1 CANADA

**Cynthia Dwork** DWORK@MICROSOFT.COM
Microsoft Research, 1065 La Avenida Mountain View, CA. 94043 USA

### Abstract

We propose a learning algorithm for fair classification that achieves both group fairness (the proportion of members in a protected group receiving positive classification is iden-

ics have voiced concerns with bias and discrimination in decision systems that rely on statistical inference and learning.

Systems trained to make decisions based on historical data will naturally inherit the past biases. These may

# p. 2 definitions

- $X$ denotes the entire data set of individuals. Each $\mathbf{x} \in X$ is a vector of length $D$ where each component of the vector describes some attribute of the person.
- $S$ is a binary random variable representing whether or not a given individual is a member of the protected set; we assume the system has access to this attribute.
- $X_0$ denotes the training set of individuals.
- $X^+ \subset X$, $X_0^+ \subset X_0$ denotes the subset of individuals (from the whole set and the training set respectively) that are members of the protected set (i.e., $S = 1$), and $X^-$ and $X_0^-$ denotes the subsets that are not members of the protected set, i.e., $S = 0$.
- $Z$ is a multinomial random variable, where each of the $K$ values represents one of the intermediate set of "prototypes". Associated with each prototype is a vector $\mathbf{v}_k$ in the same space as the individuals $\mathbf{x}$.
- $Y$ is the binary random variable representing the classification decision for an individual, and $f : X \rightarrow Y$ is the desired classification function.
- $d$ is a distance measure on $X$, e.g., simple Euclidean distance: $d(\mathbf{x}_n, \mathbf{v}_k) = ||\mathbf{x}_n - \mathbf{v}_k||_2$.

## Statistical parity:

$$P(Z = k|\mathbf{x}^+ \in X^+) = P(Z = k|\mathbf{x}^- \in X^-), \forall k \quad (1)$$

## Representation as mixture of prototypes:

$$P(Z = k|\mathbf{x}) = \exp(-d(\mathbf{x}, \mathbf{v}_k)) / \sum_{j=1}^{K} \exp(-d(\mathbf{x}, \mathbf{v}_j)) \quad (2)$$

## Learning goals:

1. the mapping from $X_0$ to $Z$ satisfies statistical parity;
2. the mapping to $Z$-space retains information in $X$ (except for membership in the protected set); and
3. the induced mapping from $X$ to $Y$ (by first mapping each $\mathbf{x}$ probabilistically to $Z$-space, and then mapping $Z$ to $Y$) is close to $f$.

Objective function: $L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$ (4)

$$L_z = \sum_{k=1}^{K} \left| M_k^+ - M_k^- \right| \qquad (7)$$

$$P(Z = k|\mathbf{x}) = \exp(-d(\mathbf{x}, \mathbf{v}_k)) / \sum_{j=1}^{K} \exp(-d(\mathbf{x}, \mathbf{v}_j)) \qquad (2)$$

$$M_{n,k} = P(Z = k|\mathbf{x}_n) \quad \forall n, k \qquad (3)$$

$$M_k^+ = \mathbb{E}_{\mathbf{x}\in X^+} P(Z = k|\mathbf{x}) = \frac{1}{|X_0^+|} \sum_{n \in X_0^+} M_{n,k} \qquad (6)$$

$$L_x = \sum_{n=1}^{N} (\mathbf{x}_n - \hat{\mathbf{x}}_n)^2 \qquad (8)$$

$$\hat{\mathbf{x}}_n = \sum_{k=1}^{K} M_{n,k} \mathbf{v}_k \qquad (9)$$

$$L_y = \sum_{n=1}^{N} -y_n \log \hat{y}_n - (1 - y_n) \log(1 - \hat{y}_n) \qquad (10)$$

$$\hat{y}_n = \sum_{k=1}^{K} M_{n,k} w_k \qquad (11)$$

Minimize $\{\mathbf{v}_k\}_{k=1}^{K}, \mathbf{w}$

*they also modify the distance function (12)

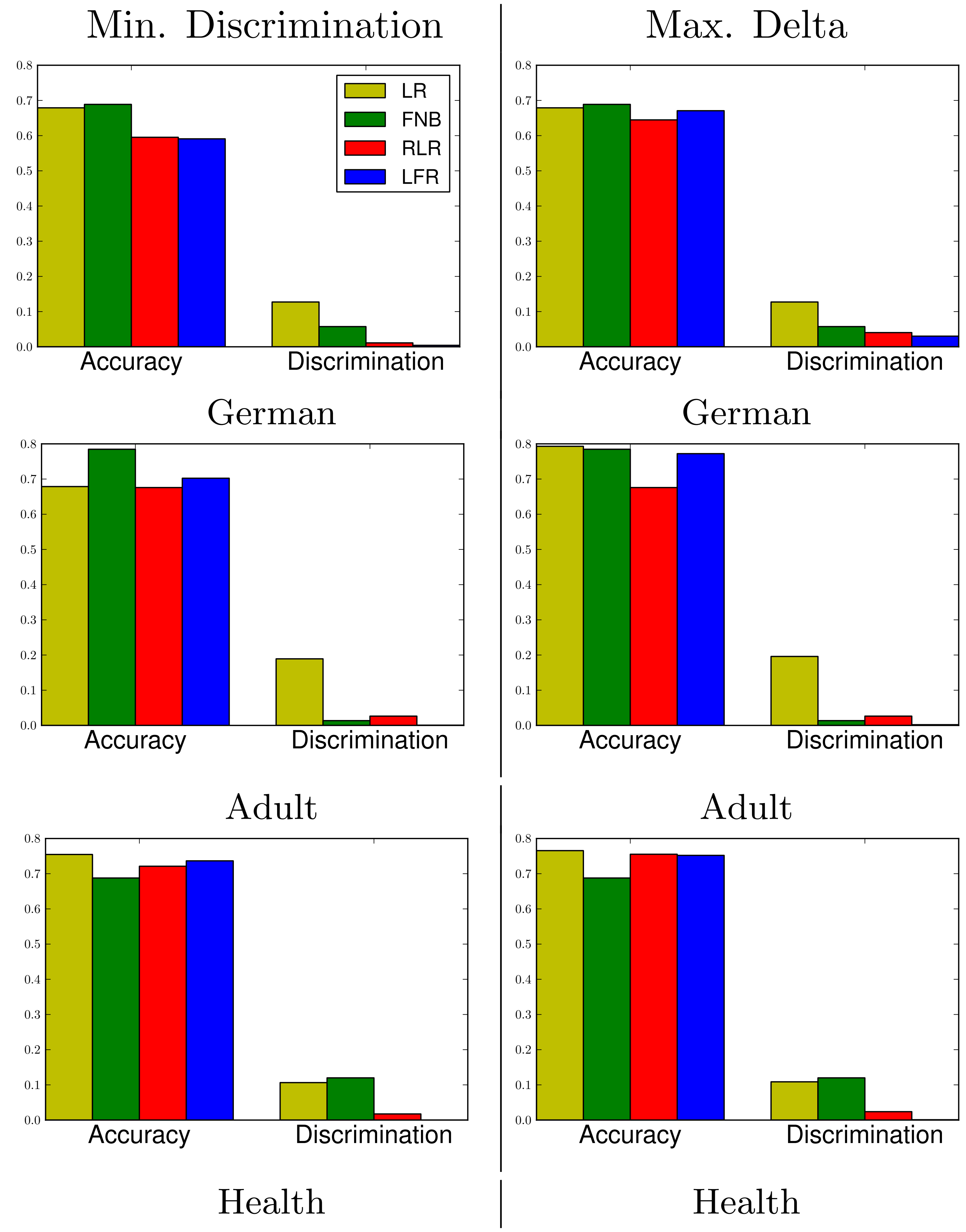


*Figure 1.* Results on test sets for the three datasets (German, Adult, and Health), for two different model selection criteria: minimizing discrimination and maximizing the difference between accuracy and discrimination.

$$yDiscrim = \left|\frac{\sum_{n:s_n=1}\hat{y}_n}{\sum_{n:s_n=1}1} - \frac{\sum_{n:s_n=0}\hat{y}_n}{\sum_{n:s_n=0}1}\right|$$

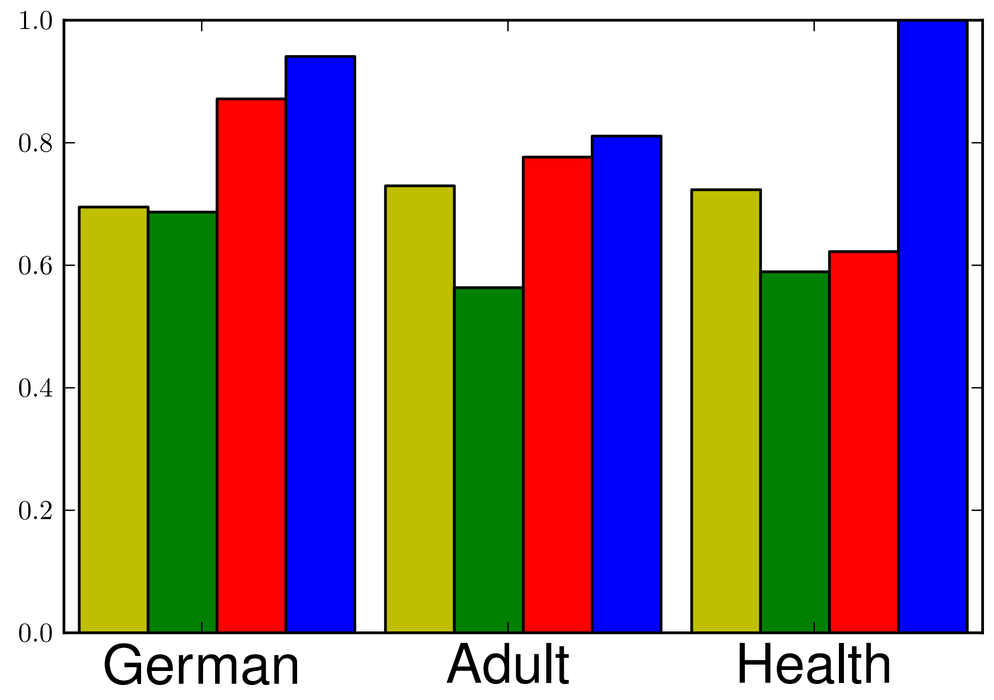


*Figure 2.* Individual fairness: The plot shows the consistency of each model's classification decisions, based on the $yNN$ measure. Legend as in Figure 1.

$$yNN = 1 - \frac{1}{Nk}\sum_{n}|\hat{y}_n - \sum_{j\in kNN(\mathbf{x}_n)}\hat{y}_j|$$

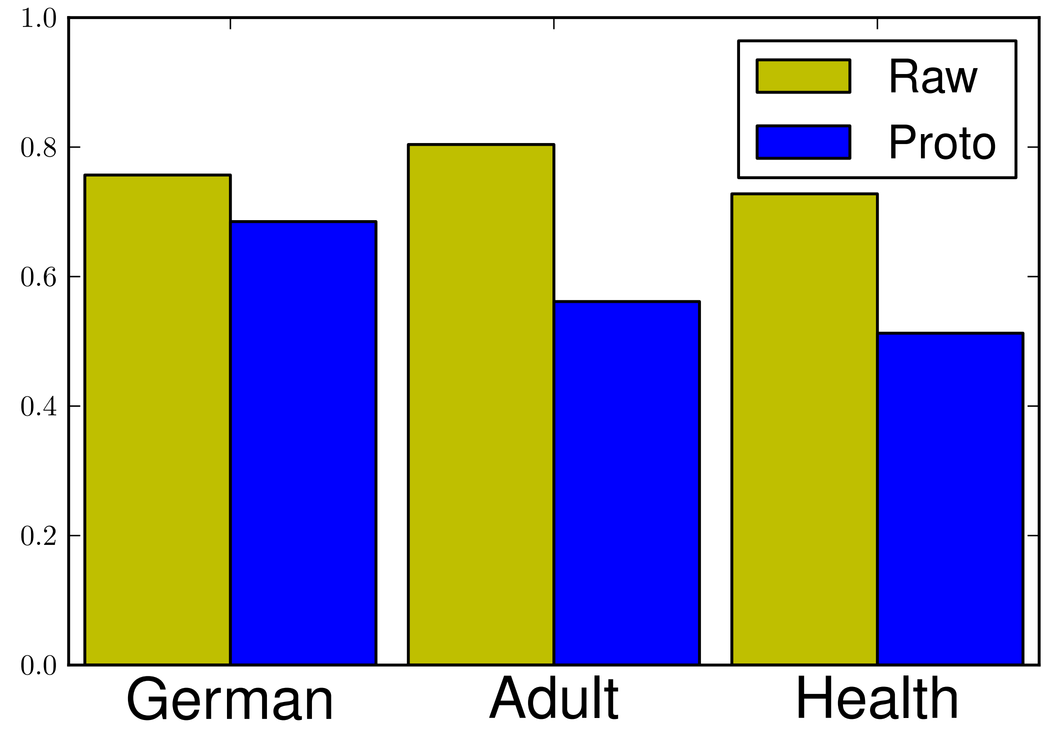
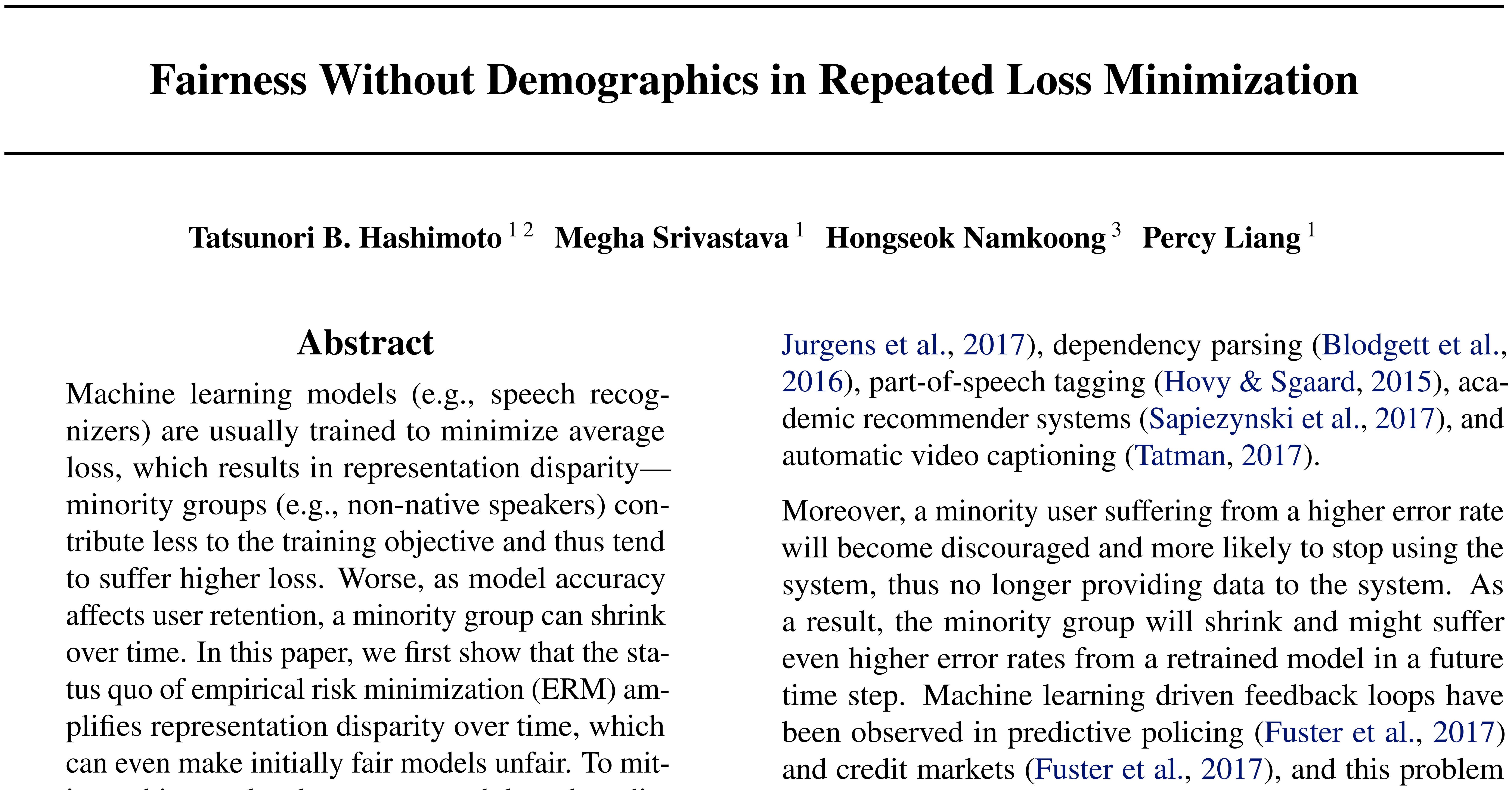


*Figure 3.* The plot shows the accuracy of predicting the sensitive variable ($sAcc$) for the different datasets. Raw involves predictions directly from all input dimensions except for $S$, while Proto involves predictions from the learned fair representations.

LR: logistic regression
FNB: fair naive Bayes
RLR: regularized LR
LFR: their method

# Case Study 3: Fixing Feedback Loops

https://arxiv.org/abs/1806.08010



## Fairness Without Demographics in Repeated Loss Minimization

**Tatsunori B. Hashimoto** [1 2] **Megha Srivastava** [1] **Hongseok Namkoong** [3] **Percy Liang** [1]

### Abstract

Machine learning models (e.g., speech recognizers) are usually trained to minimize average loss, which results in representation disparity—minority groups (e.g., non-native speakers) contribute less to the training objective and thus tend to suffer higher loss. Worse, as model accuracy affects user retention, a minority group can shrink over time. In this paper, we first show that the status quo of empirical risk minimization (ERM) amplifies representation disparity over time, which can even make initially fair models unfair. To mit-

Jurgens et al., 2017), dependency parsing (Blodgett et al., 2016), part-of-speech tagging (Hovy & Sgaard, 2015), academic recommender systems (Sapiezynski et al., 2017), and automatic video captioning (Tatman, 2017).

Moreover, a minority user suffering from a higher error rate will become discouraged and more likely to stop using the system, thus no longer providing data to the system. As a result, the minority group will shrink and might suffer even higher error rates from a retrained model in a future time step. Machine learning driven feedback loops have been observed in predictive policing (Fuster et al., 2017) and credit markets (Fuster et al., 2017), and this problem

# Feedback Model for Iterated ML

Observations from mixture of **latent** groups $Z \sim P := \sum_{k\in[K]} \alpha_k P_k$

Goal: control worst risk among groups $\mathcal{R}_{\max}(\theta) = \max_{k\in[K]} \mathcal{R}_k(\theta), \quad \mathcal{R}_k(\theta) := \mathbb{E}_{P_k}[\ell(\theta; Z)]$

**Definition 1** (Dynamics). *Given a sequence $\theta^{(t)}$, for each $t = 1 \ldots T$, the expected number of users $\lambda$ and samples $Z_i^{(t)}$ starting at $\lambda_k^{(0)} = b_k$ is governed by:*

$$\lambda_k^{(t+1)} := \lambda_k^{(t)} \nu(\mathcal{R}_k(\theta^{(t)})) + b_k$$

$$\alpha_k^{(t+1)} := \frac{\lambda_k^{(t+1)}}{\sum_{k'\in[K]} \lambda_{k'}^{(t+1)}}$$

$$n^{(t+1)} := \mathit{Pois}(\sum_k \lambda_k^{(t+1)})$$

$$Z_1^{(t+1)} \ldots Z_{n^{(t+1)}}^{(t+1)} \overset{\text{i.i.d.}}{\sim} P^{(t+1)} := \sum_{k\in[K]} \alpha_k^{(t+1)} P_k.$$

retention function

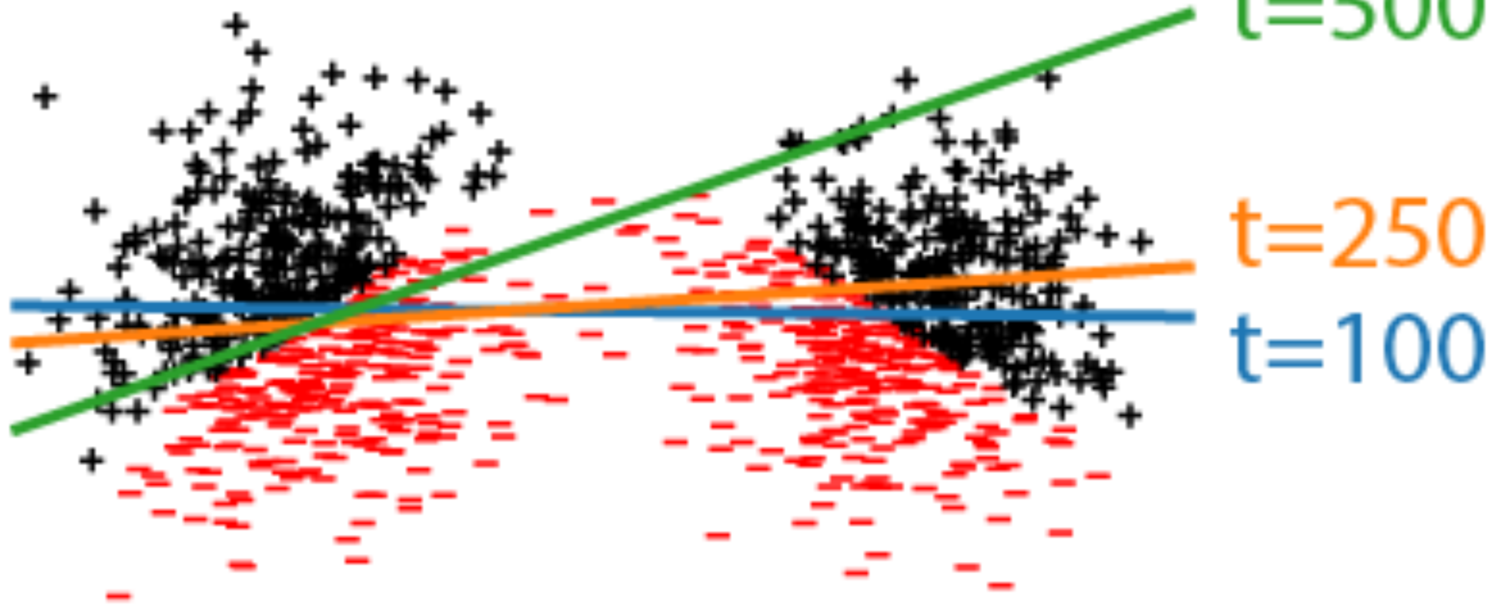


*Figure 1.* An example online classification problem which begins fair, but becomes unfair over time.

# Solution: Distributionally Robust Optimization

$$\mathcal{R}_{\mathrm{dro}}(\theta; r) := \sup_{Q \in \mathcal{B}(P,r)} \mathbb{E}_Q[\ell(\theta; Z)]. \tag{4}$$

primal objective

$$\mathcal{B}(P, r) \;:=\; \{Q \;\ll\; P \;:\; D_{\chi^2}\,(Q \| P) \leq r\}$$

$$\underset{\theta \in \Theta}{\mathrm{minimize}}\, \mathbb{E}_P \left[\ell(\theta; Z) - \eta\right]_+^2 . \tag{6}$$

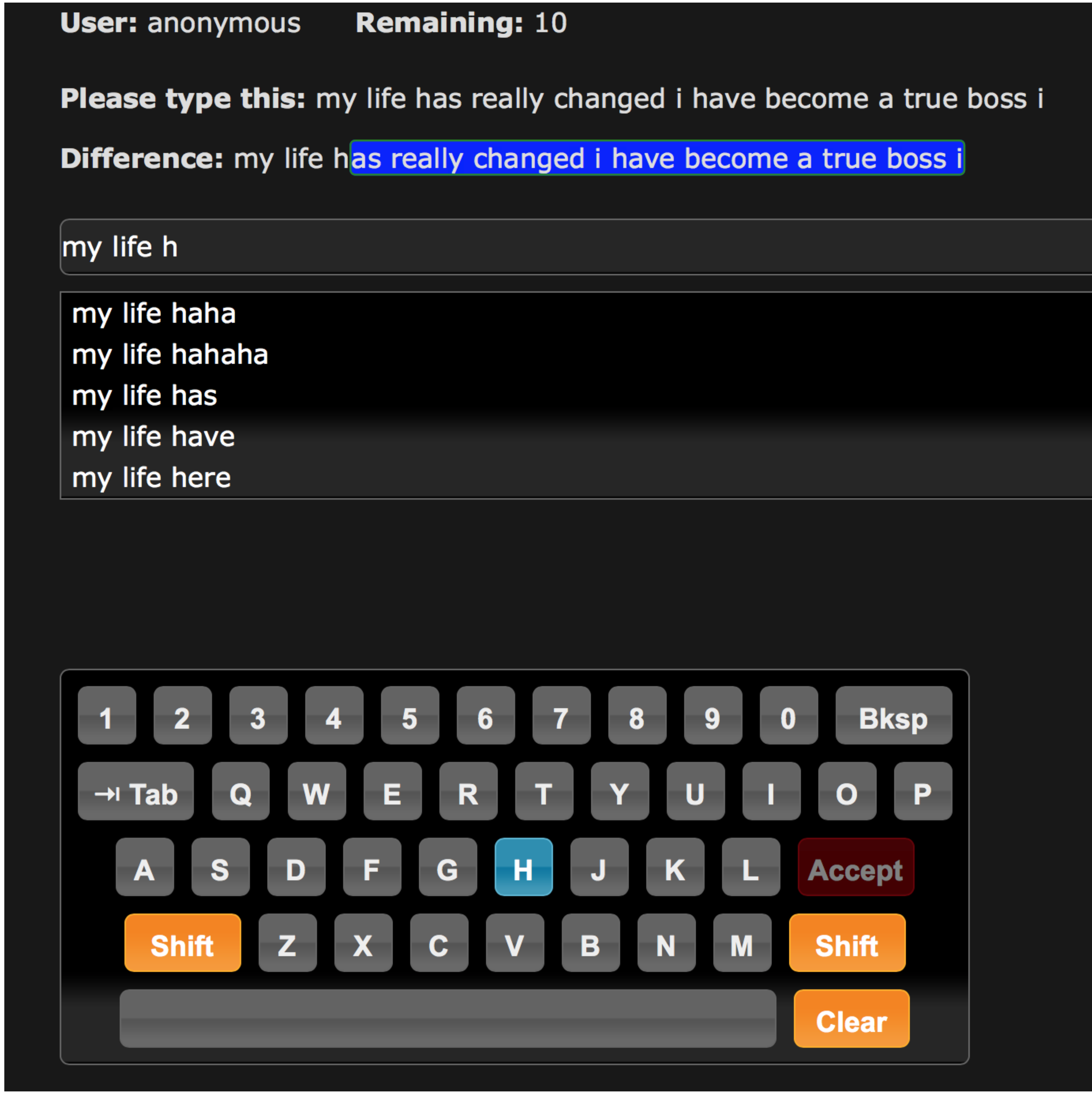
proven upper bound

Search for best $\eta$

User study: ask crowdsource workers to retype tweets

Tweets are categorized by linguists as using African-American English and Standard-American English dialects. Assign one dialect to each user.

Learn autocomplete language models. Survey users after rounds on whether they would continue using system.



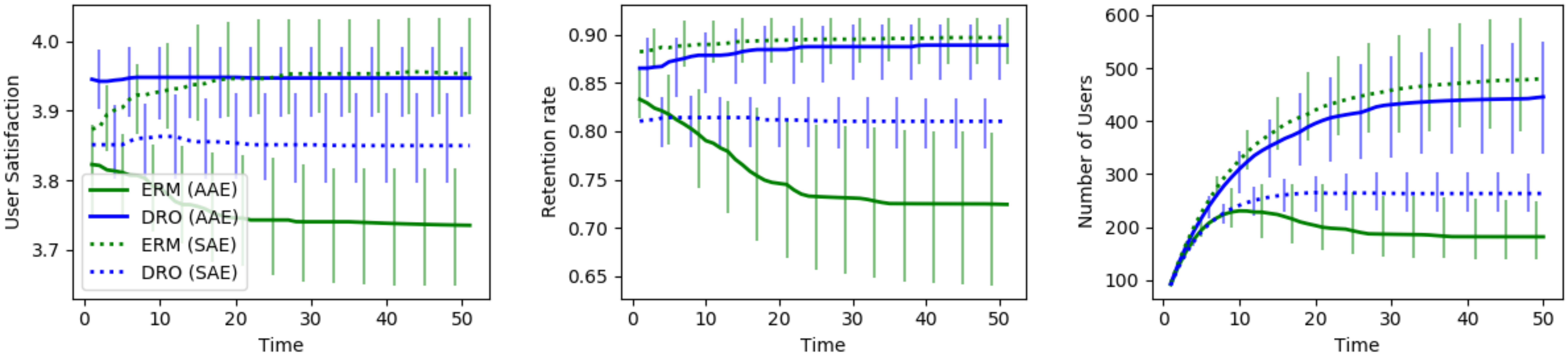
**User:** anonymous **Remaining:** 10

**Please type this:** my life has really changed i have become a true boss i

**Difference:** my life has really changed i have become a true boss i

my life h

my life haha
my life hahaha
my life has
my life have
my life here



(a) User satisfaction

(b) User retention

(c) User count

*Figure 4.* Inferred dynamics from a Mechanical Turk based evaluation of autocomplete systems. DRO increases minority (a) user satisfaction and (b) retention, leading to a corresponding increase in (c) user count. Error bars indicates bootstrap quartiles.

# Case Studies

- Equal opportunity (NeurIPS 2016)
- Learning fair representation (ICML 2013)
- Feedback loops in repeated loss minimization (ICML 2018, best paper runner up)

# Closing Thoughts

- Provide technology to prevent technology from doing wrong
- Transparency, explainability, interpretability
- Current trajectory is bad. Corrective research is too slow.
- ML is not automatically fair because it’s based on math.