Fairness in Machine Learning

CS5824/ECE5424
Bert Huang

CNBC Report on ML in Law Enforcement

https://youtu.be/ZMsSc_utZ40

Plan

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

Types of Fairness, An Incomplete List

- Unawareness
- Group prediction parity
- Group error parity
- Individual counterfactual fairness
- Envy-free fairness

Unawareness

- Data X = $\{x_1, \ldots, x_n\}$
- Target Y = $\{y_1, \ldots, y_n\}$
- Sensitive feature S = $\{s_1, \ldots, s_n\}$
- Concern that f(x, s) would use s, so only train f(x)
- Usually fails because some features in x are correlated with s

Group Prediction Parity

- Treat two sub-populations the same
  - Learn $f(x, s)$ such that $E_{s=1}[f(x, s)] \approx E_{s=0}[f(x, s)]$
  - Prediction probability has similar statistics for groups with or without sensitive feature

Group Error Parity

- Treat two sub-populations equally well
    - Learn f(x, s) such that $E_{s=1}[\text{error}(f(x, s), y)] \approx E_{s=0}[\text{error}(f(x, s), y)]$
- Prediction error is *independent* of sensitive feature s
- Defining error as **true-positive rate**, we get equal opportunity
    - Individuals who deserve loans are equally likely to be offered

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

Individual Counterfactual Fairness

- Treat **each** individual the same regardless of sensitive features
  - Learn f(x, s) such that $f(x, s = 0) \approx f(x, s = 1)$
- Prediction probability is *independent* of sensitive feature s for each individual

Envy-Free Fairness

- In resource allocation, an envy-free assignment is one where each individual would not prefer to receive the assignment of another
    - E.g., cake cutting, chore assignments, ad allocation

Causes of Unfairness, An Incomplete List

- ML mimics data from unfair systems
- Definition of ML tasks is unfair
- Underrepresentation of minority groups
- Feedback loops in deployed ML

Data From Unfair Systems

- Academic/professional performance, salary, crime
- Society is working on making these things more fair
- Learning to replicate old data could be a step back

Unfair ML Problem Definitions

- Predicting race, gender, native language, income level, criminality, religion, sexual orientation
- Some of these ideas don’t even have clear definitions
- And they often have little or nothing to do with input data
- ML will happily learn correlations

Unfairness from Underrepresentation

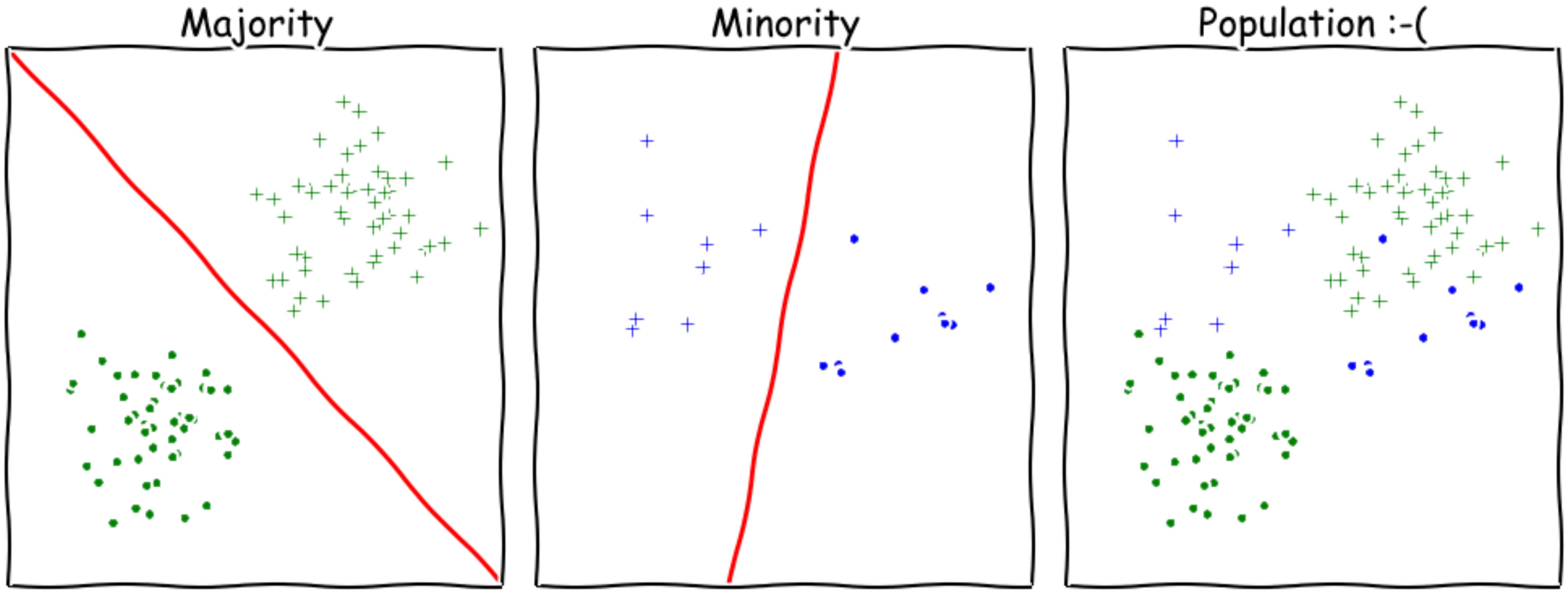
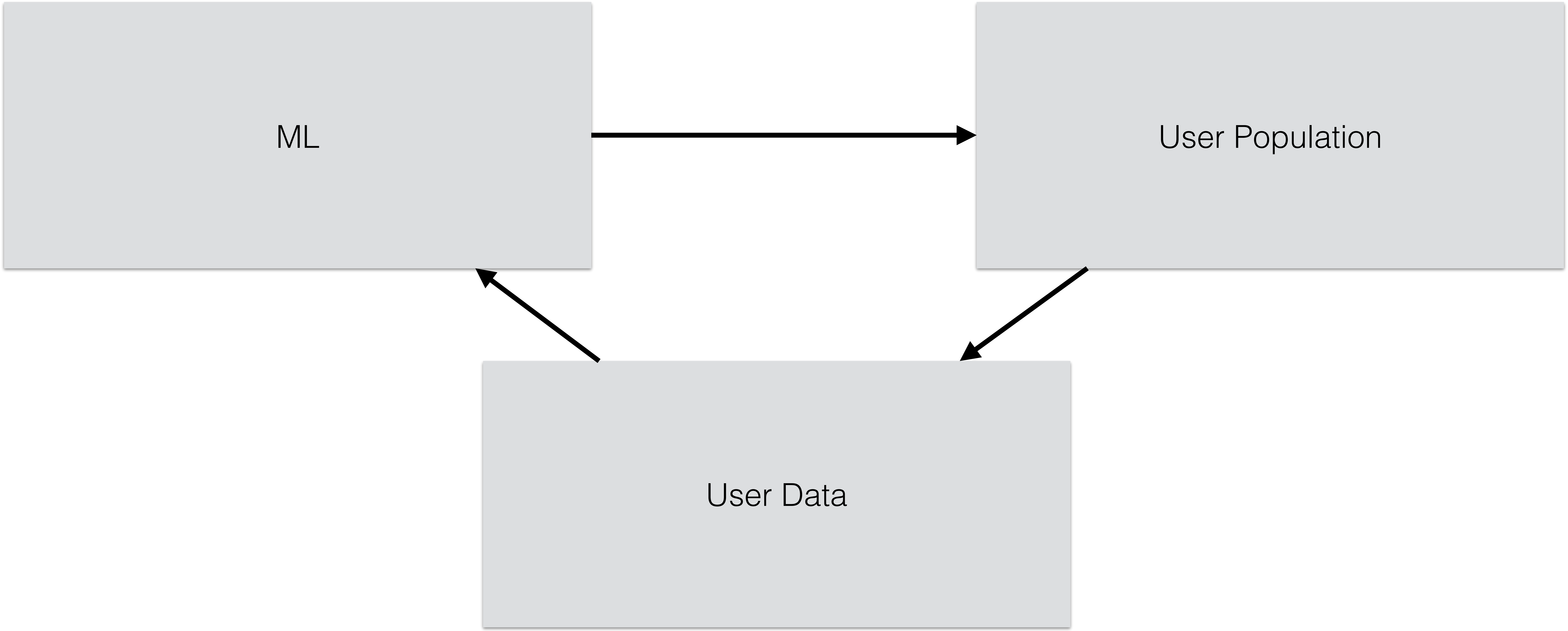


Illustration by Moritz Hardt (https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de)

Feedback Loops



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

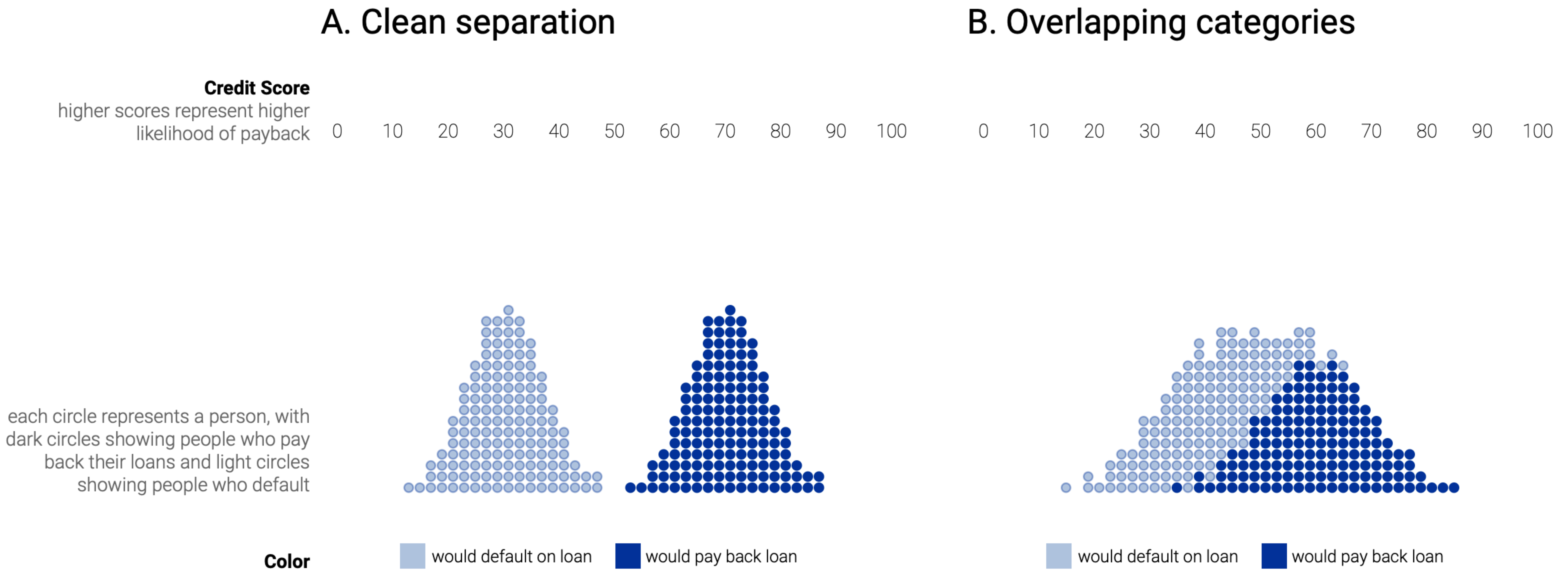
Abstract

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.

Loan applicants: two scenarios



Simulating loan thresholds

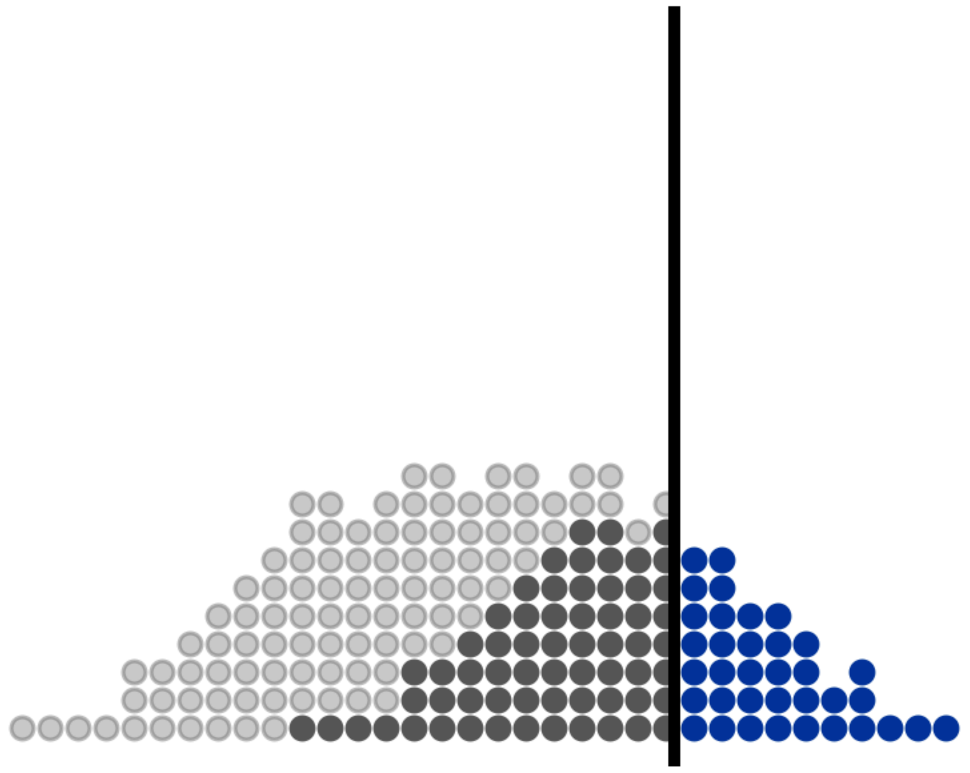
Drag the black threshold bars left or right to change the cut-offs for loans.

Threshold Decision

Credit Score
higher scores represent higher likelihood of payback

0 10 20 30 40 50 60 70 80 90 100

loan threshold: 64



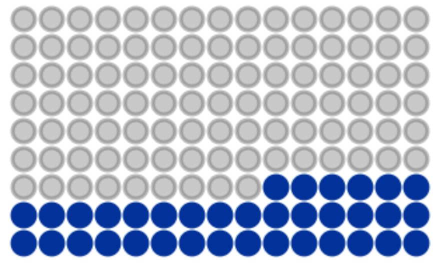
each circle represents a person, with dark circles showing people who pay back their loans and light circles showing people who default

Color

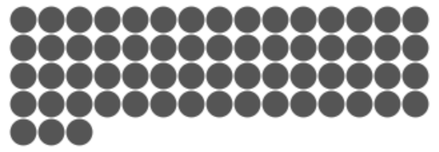
denied loan / would default | granted loan / defaults
denied loan / would pay back | granted loan / pays back

Outcome

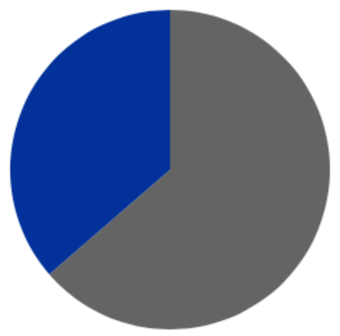
Correct 68%
loans granted to paying applicants and denied to defaulters



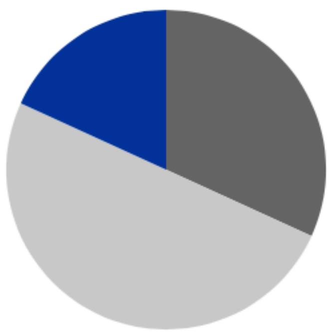
Incorrect 32%
loans denied to paying applicants and granted to defaulters



True Positive Rate36%
percentage of paying applications getting loans



Positive Rate18%
percentage of all applications getting loans



Profit: **10800**

Loan Strategy

Maximize profit with:

MAX PROFIT

No constraints

GROUP UNAWARE

Blue and orange thresholds are the same

DEMOGRAPHIC PARITY

Same fractions blue / orange loans

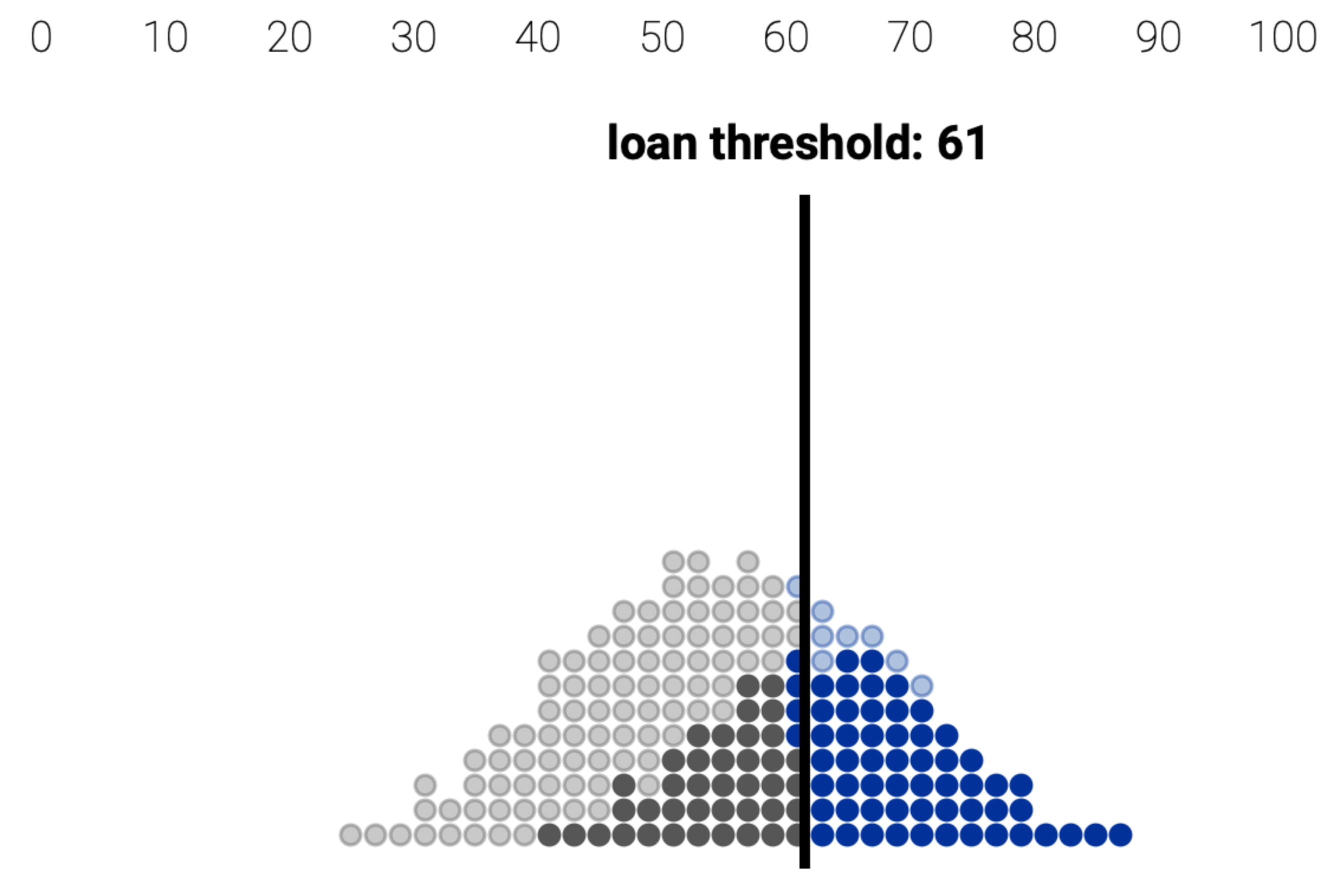
EQUAL OPPORTUNITY

Same fractions blue / orange loans to people who can pay them off

Max Profit

The most profitable, since there are no constraints. But the two groups have different thresholds, meaning they are held to different standards.

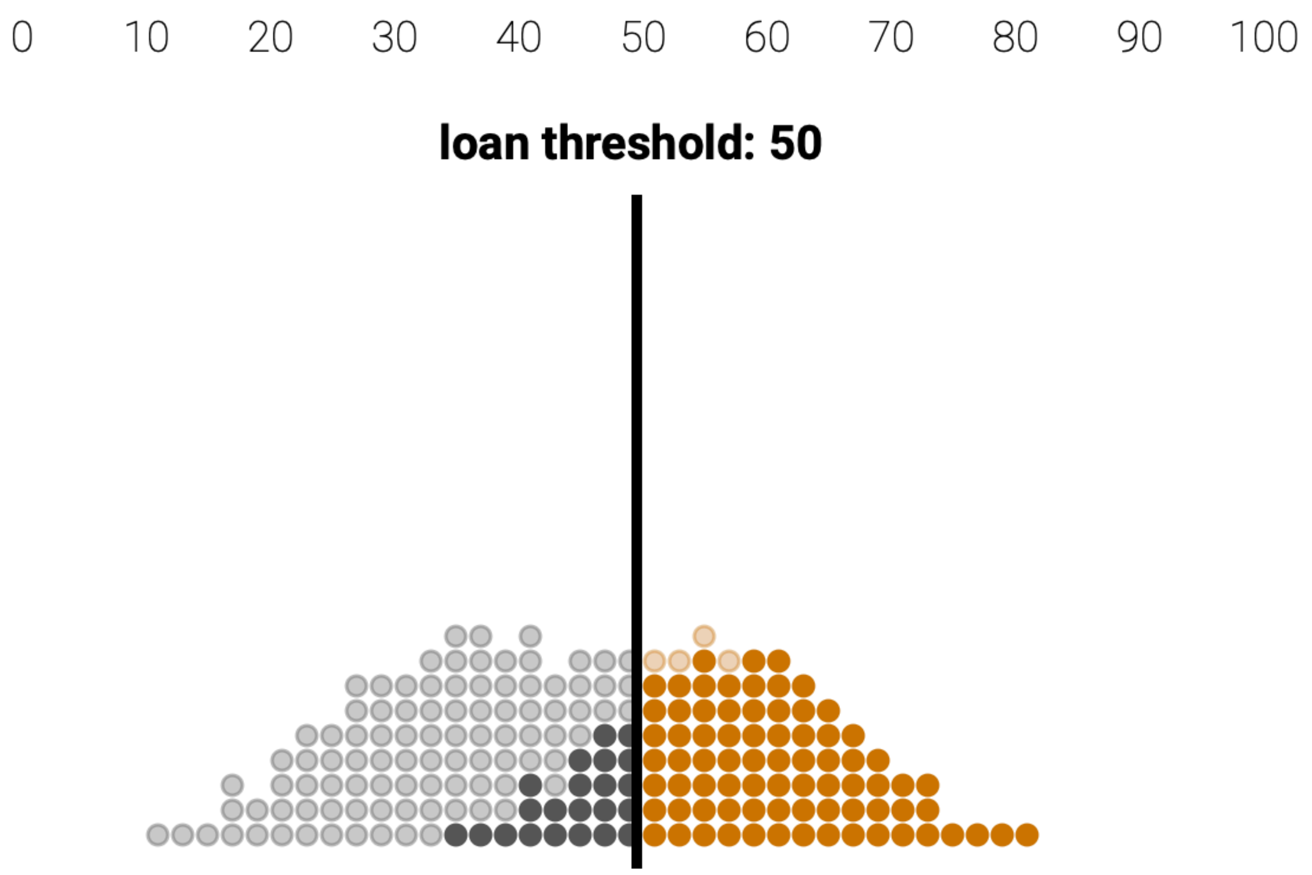
Blue Population



denied loan / would default — granted loan / defaults

denied loan / would pay back — granted loan / pays back

Orange Population



denied loan / would default — granted loan / defaults

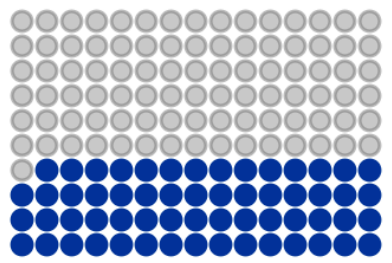
denied loan / would pay back — granted loan / pays back

Total profit = 32400

Blue Population

Correct 76%

loans granted to paying applicants and denied to defaulters

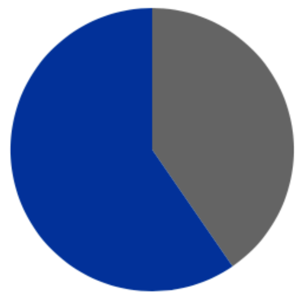


Incorrect 24%

loans denied to paying applicants and granted to defaulters

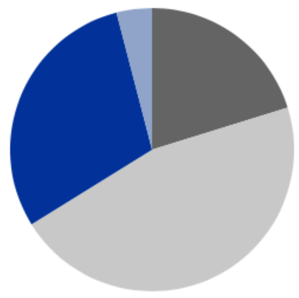
True Positive Rate 60%

percentage of paying applications getting loans



Positive Rate 34%

percentage of all applications getting loans

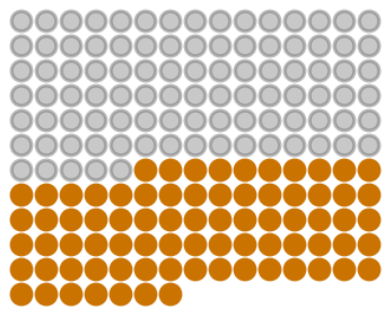


Profit: **12100**

Orange Population

Correct 87%

loans granted to paying applicants and denied to defaulters

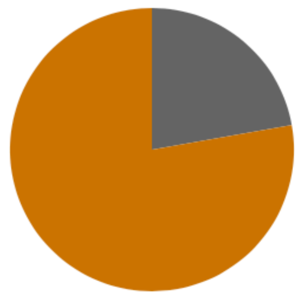


Incorrect 13%

loans denied to paying applicants and granted to defaulters

True Positive Rate 78%

percentage of paying applications getting loans

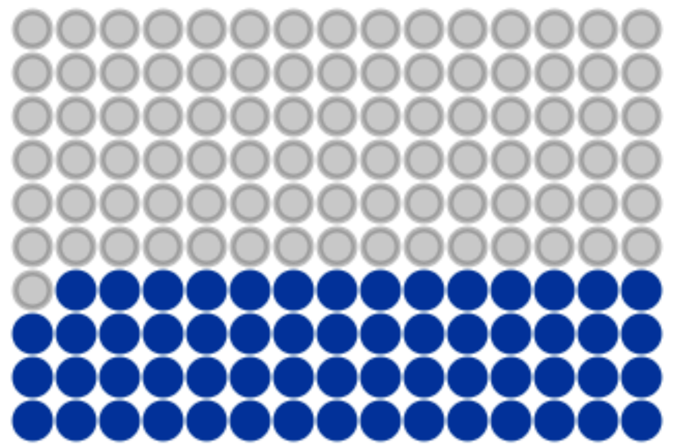


Positive Rate 41%

percentage of all applications getting loans

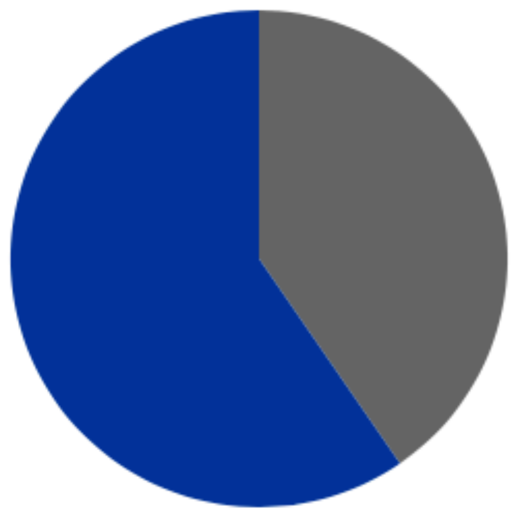
Profit: **20300**

to defaulters

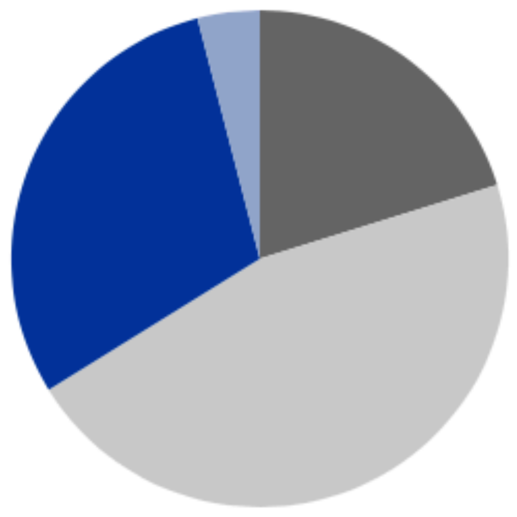


to defaulters

True Positive Rate60%
percentage of paying applications getting loans

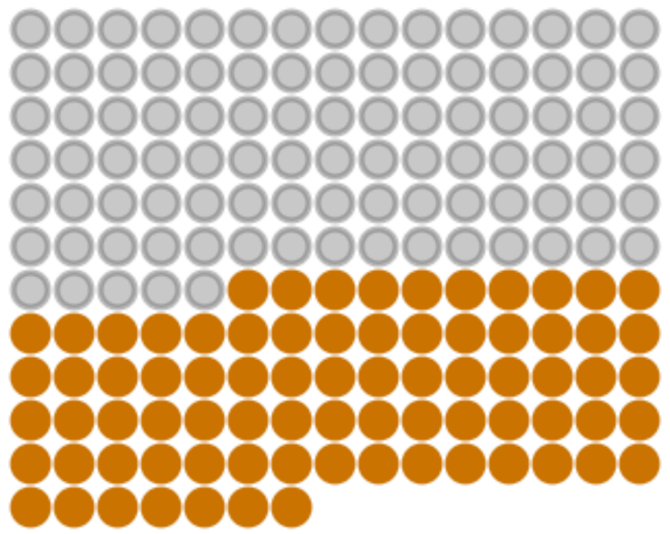


Positive Rate34%
percentage of all applications getting loans

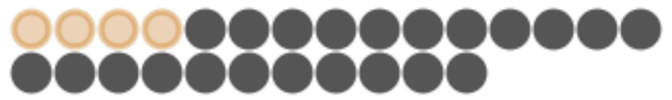


Profit: **12100**

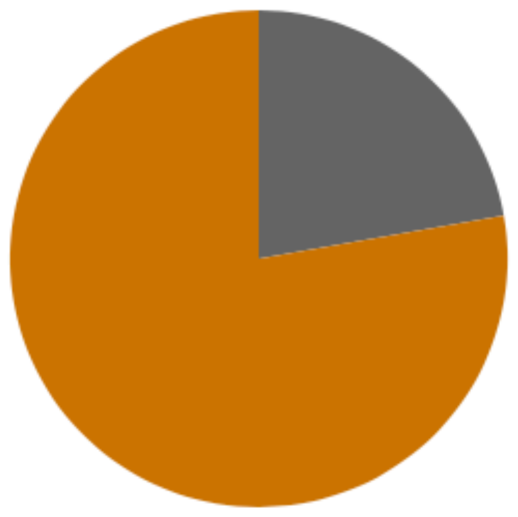
to defaulters



to defaulters



True Positive Rate78%
percentage of paying applications getting loans



Positive Rate41%
percentage of all applications getting loans

Profit: **20300**

Loan Strategy

Maximize profit with:

MAX PROFIT

No constraints

GROUP UNAWARE

Blue and orange thresholds are the same

DEMOGRAPHIC PARITY

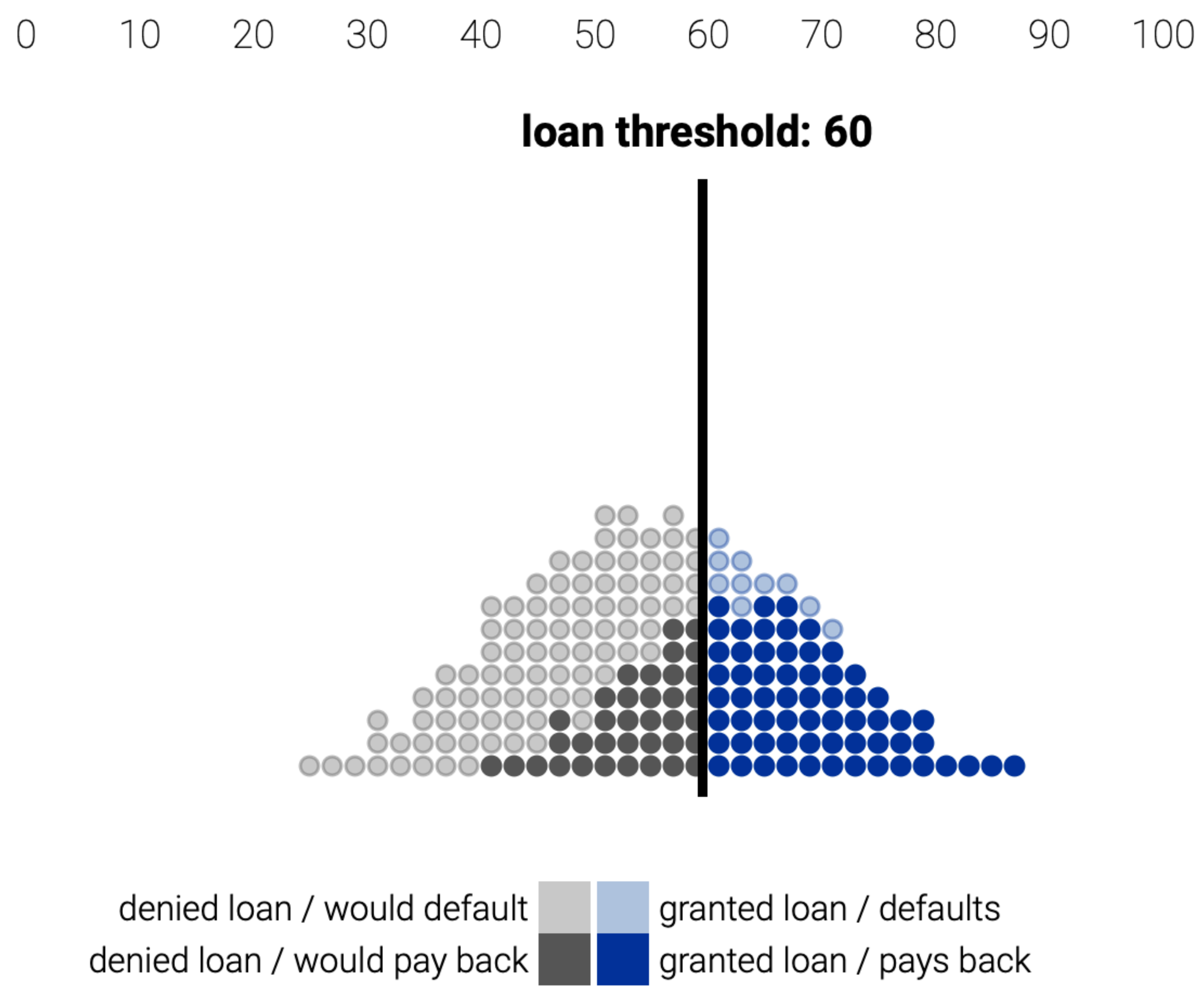
Same fractions blue / orange loans

EQUAL OPPORTUNITY

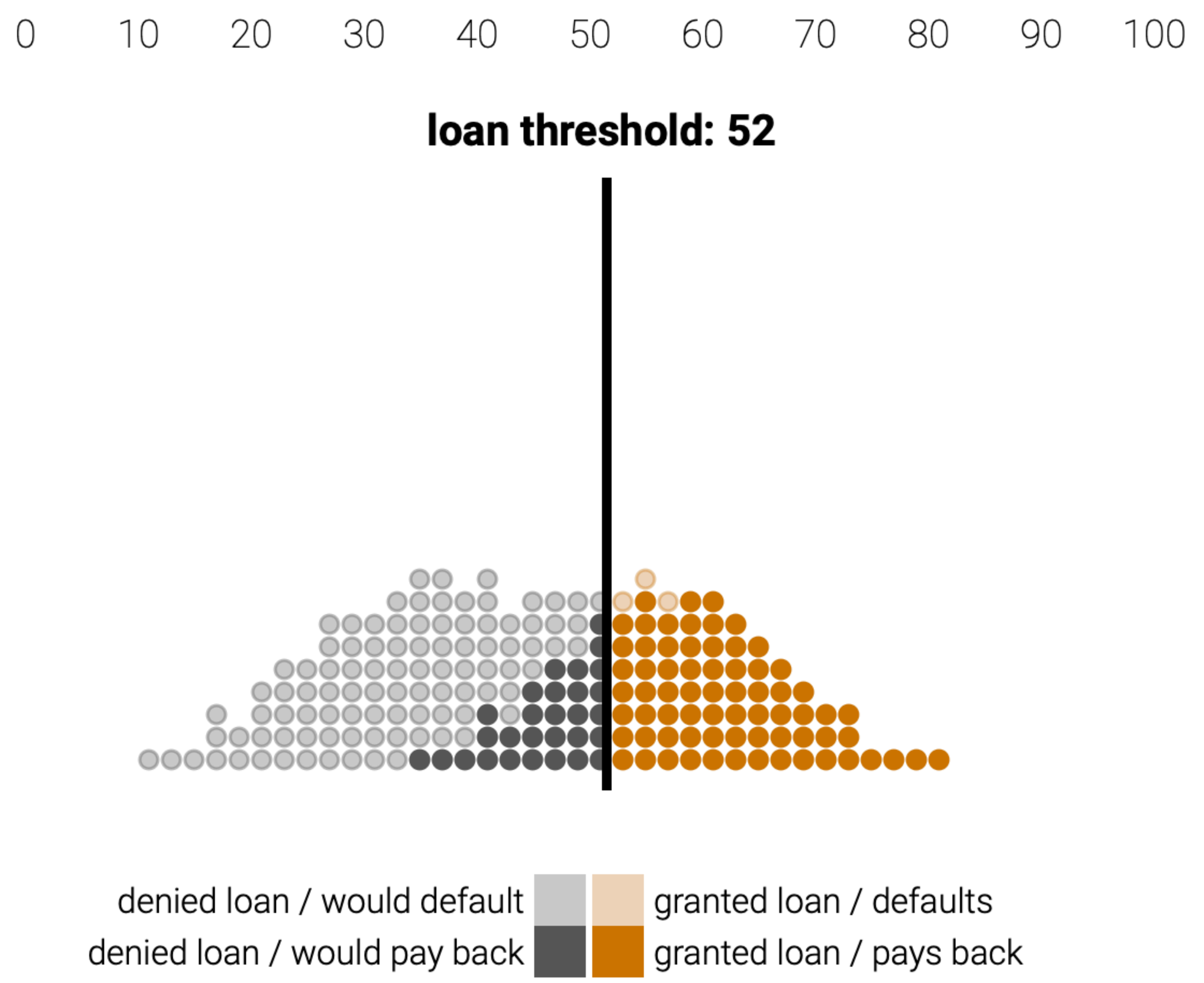
Same fractions blue / orange loans to people who can pay them off

Demographic Parity
The number of loans given to each group is the same, but among people who would pay back a loan, the blue group is at a disadvantage.

Blue Population

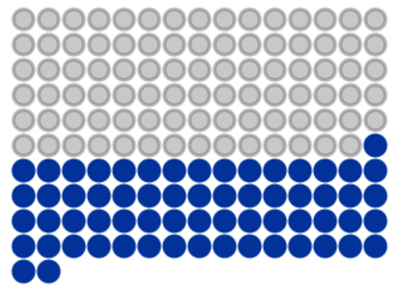


Orange Population

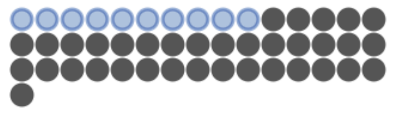


Total profit = **30800**

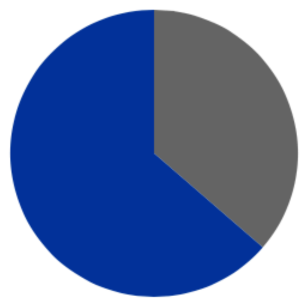
Correct 77%
loans granted to paying applicants and denied to defaulters



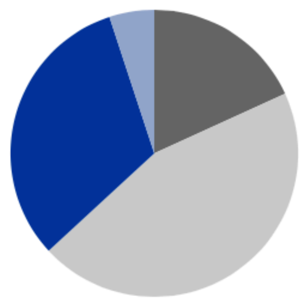
Incorrect 23%
loans denied to paying applicants and granted to defaulters



True Positive Rate 64%
percentage of paying applications getting loans

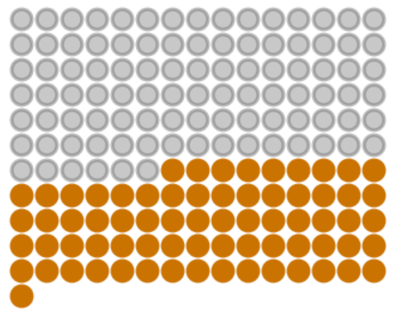


Positive Rate 37%
percentage of all applications getting loans



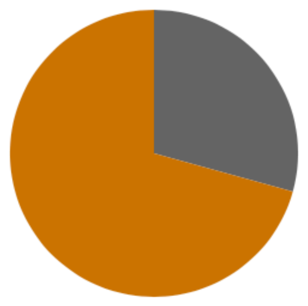
Profit: **11900**

Correct 84%
loans granted to paying applicants and denied to defaulters

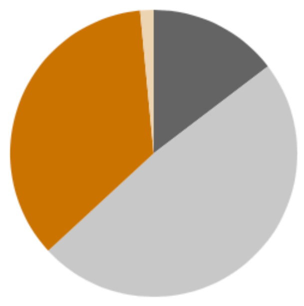


Incorrect 16%
loans denied to paying applicants and granted to defaulters

True Positive Rate 71%
percentage of paying applications getting loans



Positive Rate 37%
percentage of all applications getting loans

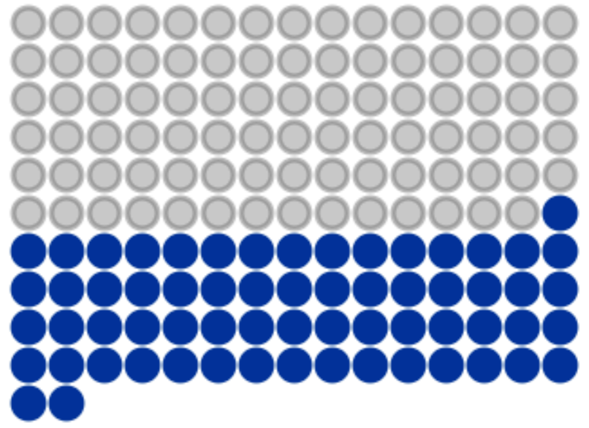


Profit: **18900**

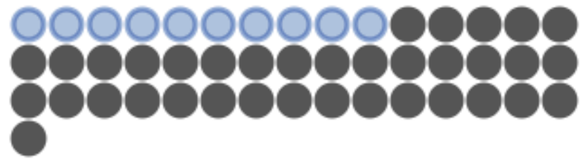
Demographic Parity
The number of loans given to each group is the same, but among people who would pay back a loan, the blue group is at a disadvantage.

Total profit = **30800**

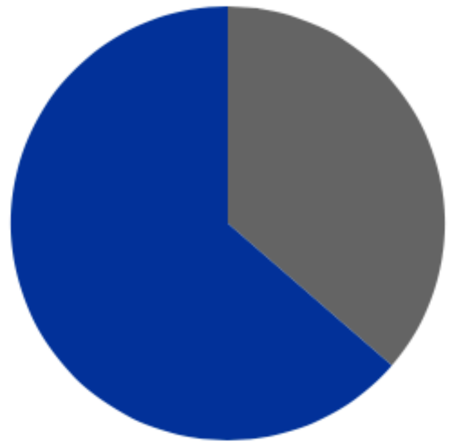
Correct 77%
loans granted to paying applicants and denied to defaulters



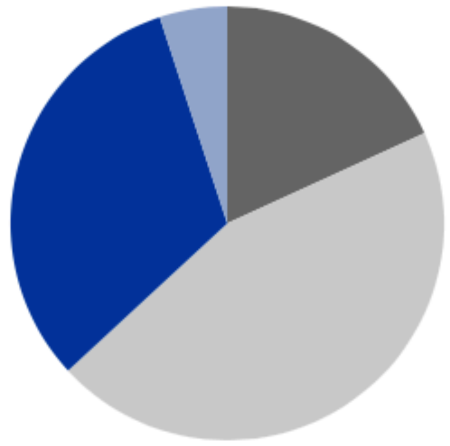
Incorrect 23%
loans denied to paying applicants and granted to defaulters



True Positive Rate64%
percentage of paying applications getting loans

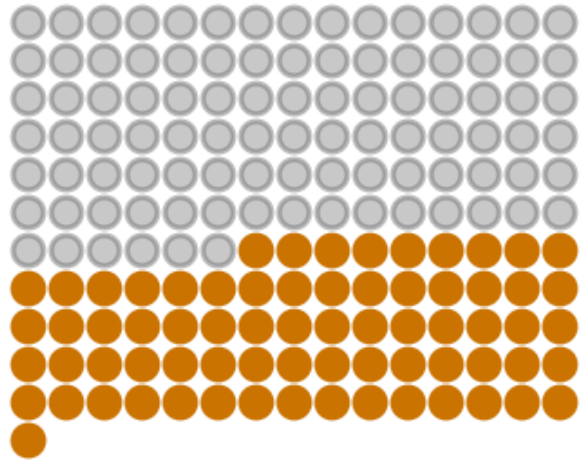


Positive Rate37%
percentage of all applications getting loans



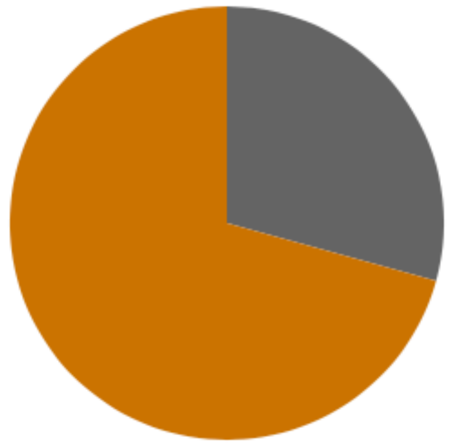
Profit: **11900**

Correct 84%
loans granted to paying applicants and denied to defaulters

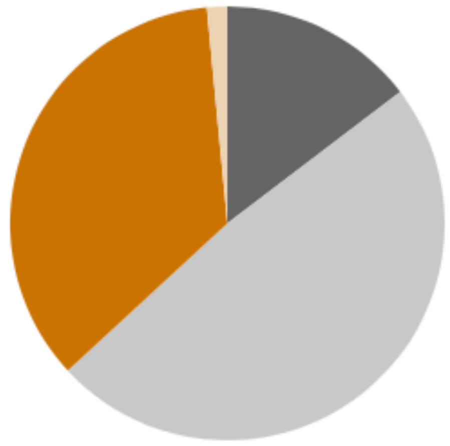


Incorrect 16%
loans denied to paying applicants and granted to defaulters

True Positive Rate71%
percentage of paying applications getting loans



Positive Rate37%
percentage of all applications getting loans



Profit: **18900**

Loan Strategy

Maximize profit with:

MAX PROFIT

No constraints

GROUP UNAWARE

Blue and orange thresholds are the same

DEMOGRAPHIC PARITY

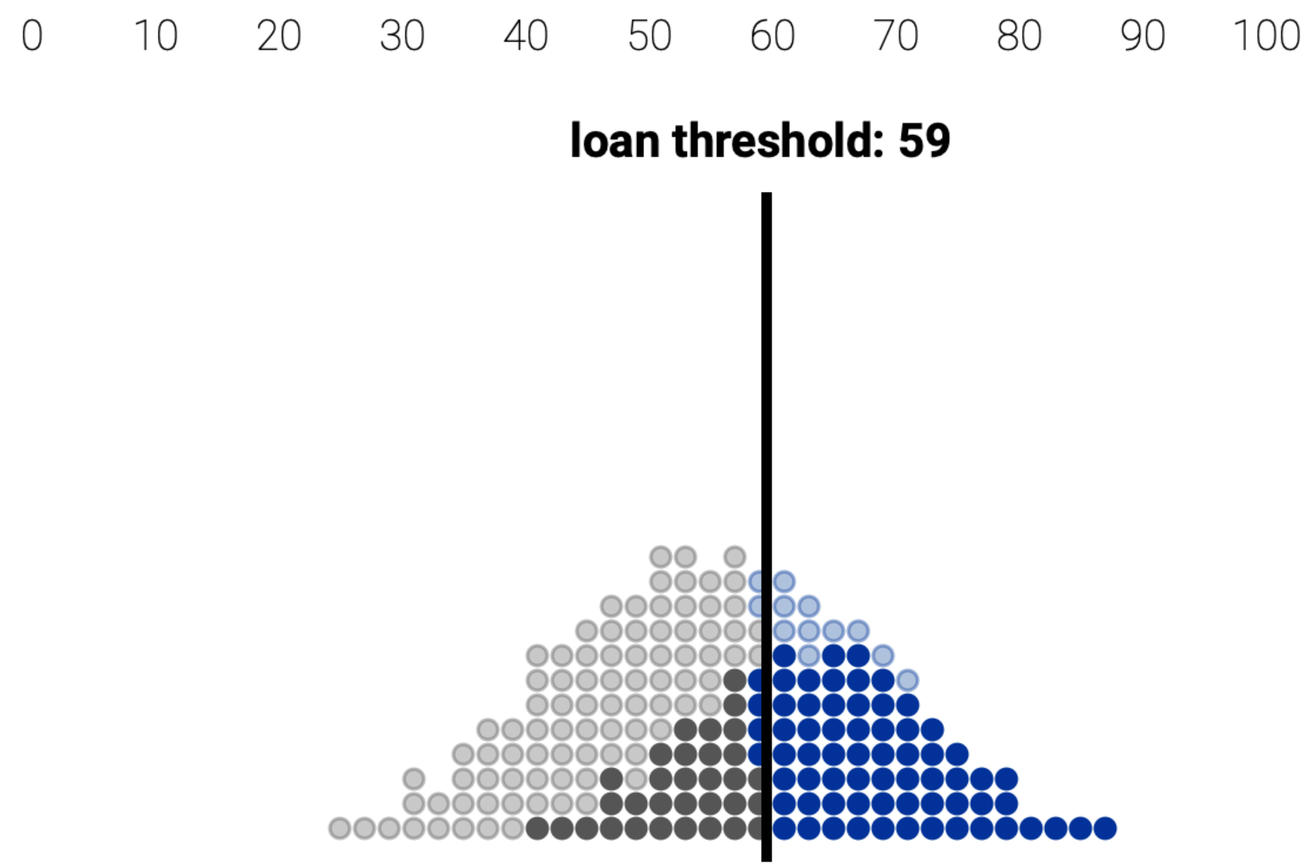
Same fractions blue / orange loans

EQUAL OPPORTUNITY

Same fractions blue / orange loans to people who can pay them off

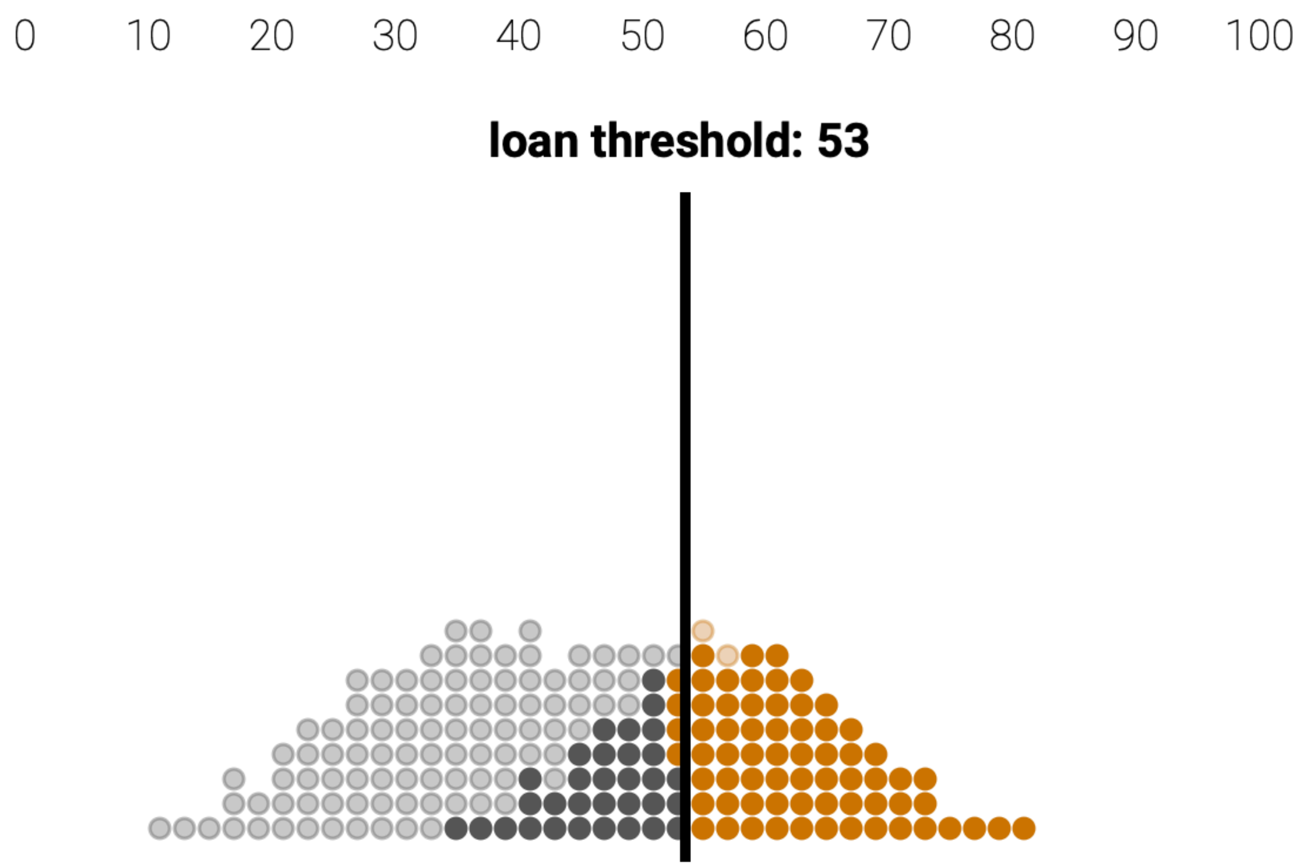
Equal Opportunity
Among people who would pay back a loan, blue and orange groups do equally well. This choice is almost as profitable as demographic parity, and about as many people get loans overall.

Blue Population



denied loan / would default — granted loan / defaults
denied loan / would pay back — granted loan / pays back

Orange Population

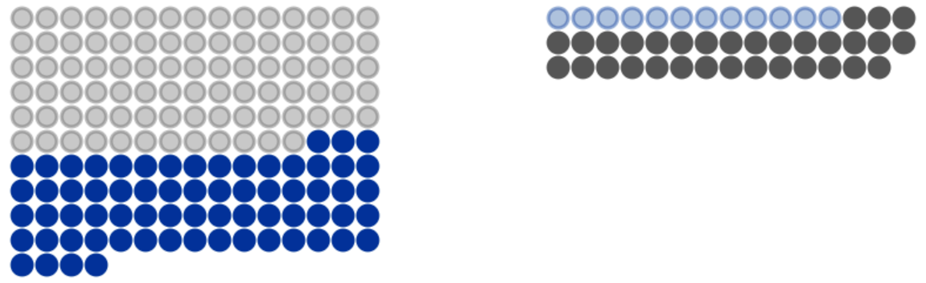


denied loan / would default — granted loan / defaults
denied loan / would pay back — granted loan / pays back

Total profit = **30400**

Correct 78%
loans granted to paying applicants and denied to defaulters

Incorrect 22%
loans denied to paying applicants and granted to defaulters



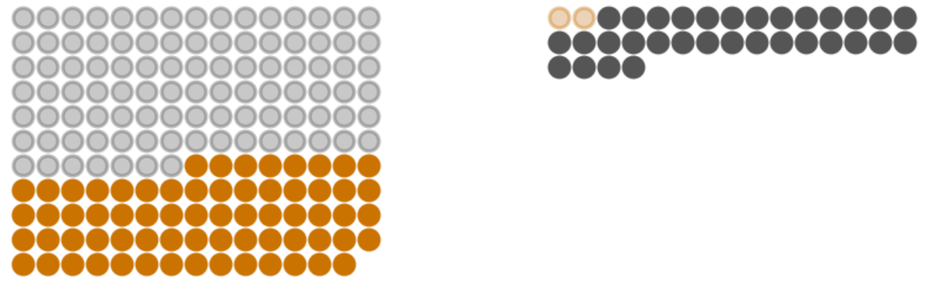
True Positive Rate 68%
percentage of paying applications getting loans

Positive Rate 40%
percentage of all applications getting loans

Profit: **11700**

Correct 83%
loans granted to paying applicants and denied to defaulters

Incorrect 17%
loans denied to paying applicants and granted to defaulters

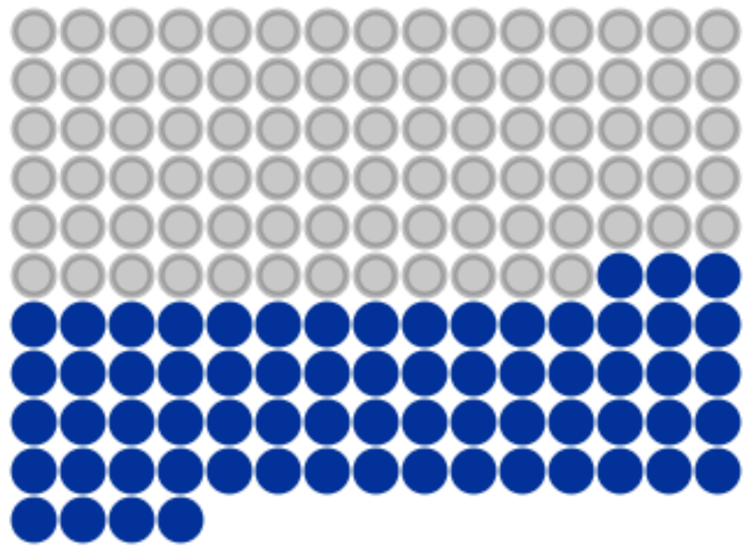


True Positive Rate 68%
percentage of paying applications getting loans

Positive Rate 35%
percentage of all applications getting loans

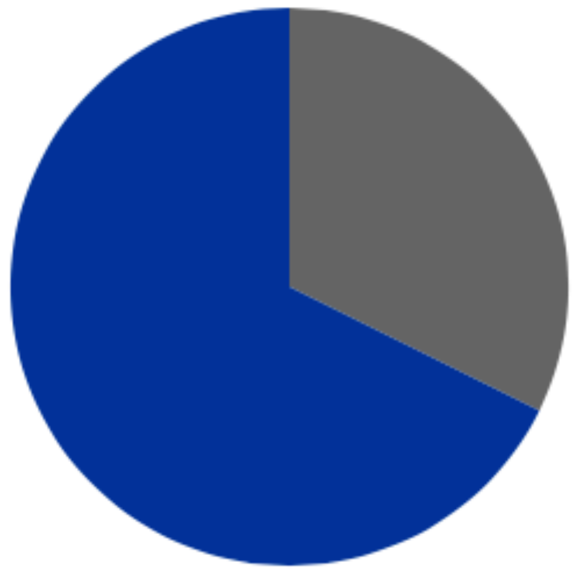
Profit: **18700**

to defaulters



True Positive Rate68%

percentage of paying applications getting loans

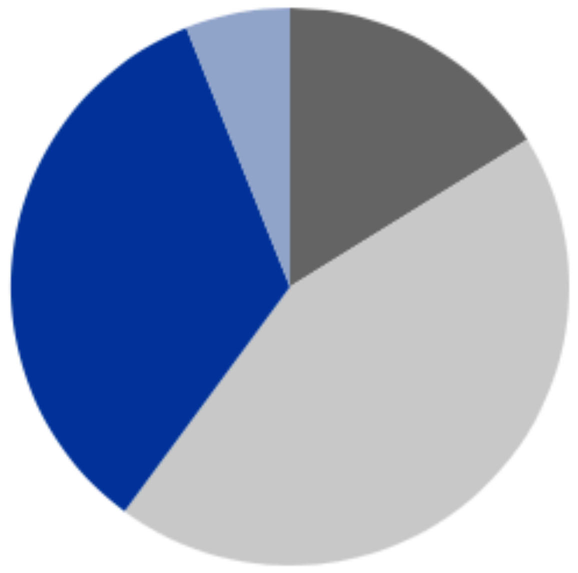


Profit: **11700**

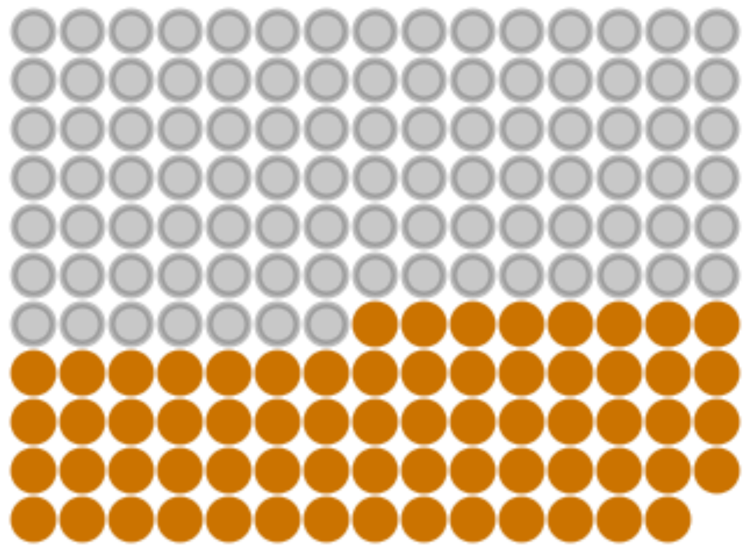
to defaulters

Positive Rate40%

percentage of all applications getting loans

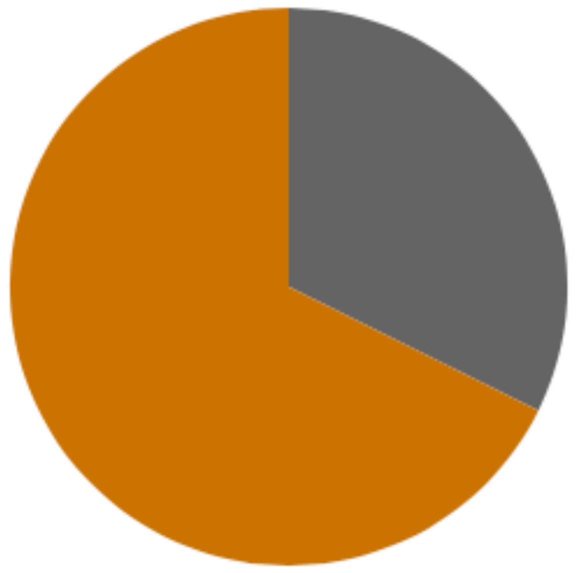


to defaulters



True Positive Rate68%

percentage of paying applications getting loans



Profit: **18700**

to defaulters

Positive Rate35%

percentage of all applications getting loans

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 $f : X \to Y$ 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: $$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y \quad (4)$$

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

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

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

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

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

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

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

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

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

*they also modify the distance function (12)

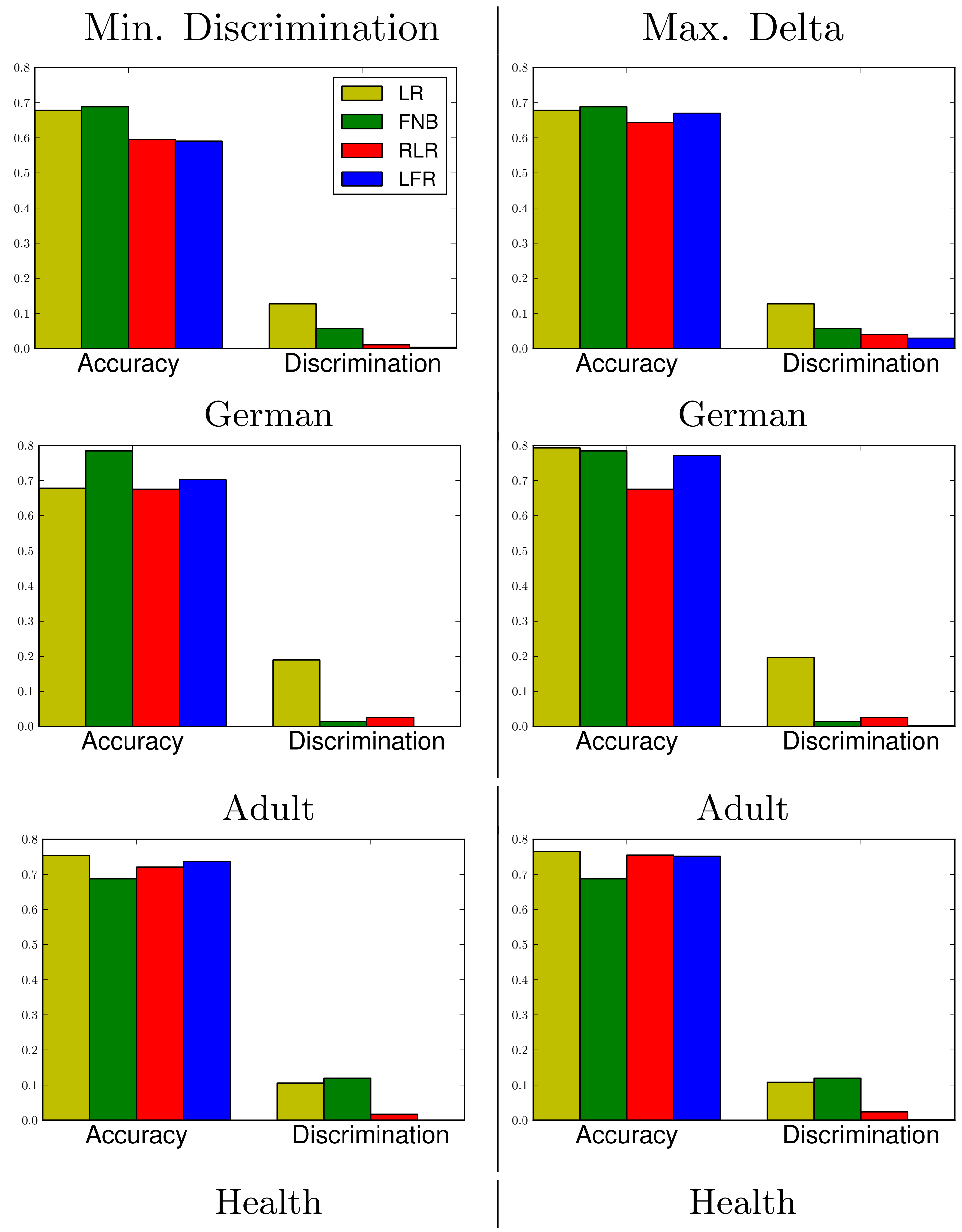


Figure 1. Results on test sets for the three datasets (German, Adult, and Health), for two different model selection criteria: minimizing discrimination and maximizing the difference between accuracy and discrimination.

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

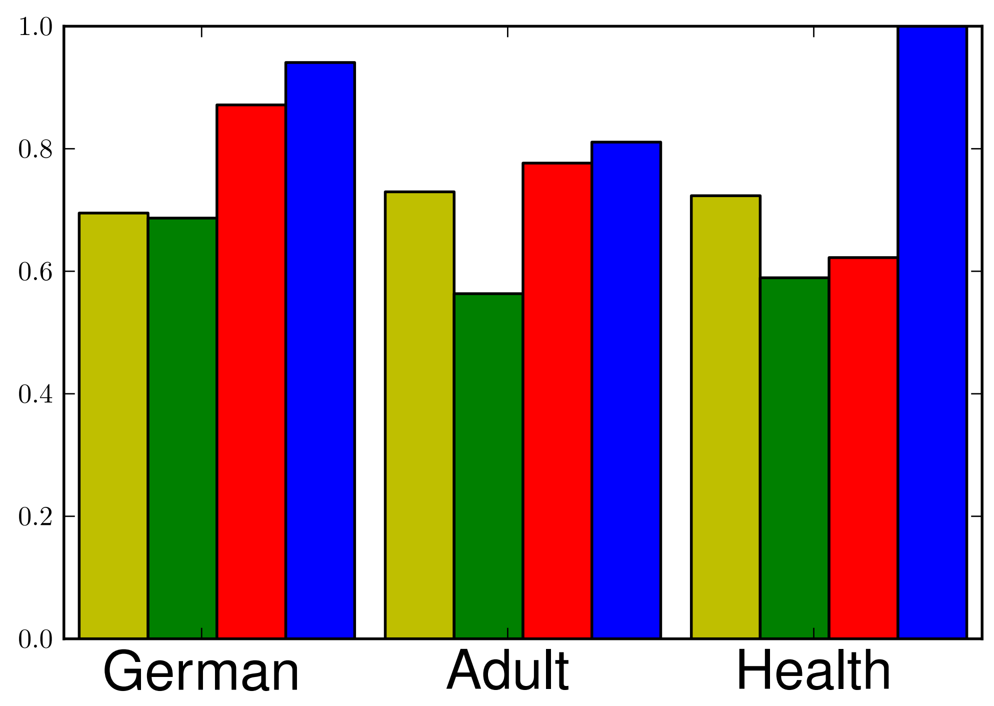


Figure 2. Individual fairness: The plot shows the consistency of each model's classification decisions, based on the yNN measure. Legend as in Figure 1.

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

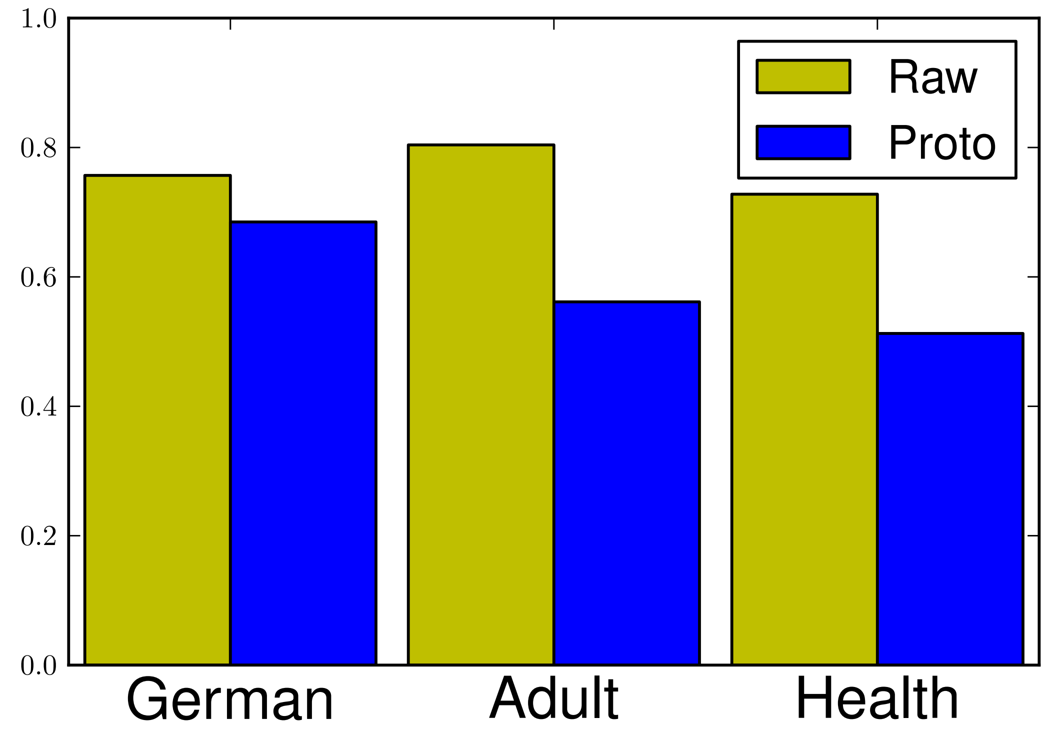


Figure 3. The plot shows the accuracy of predicting the sensitive variable ($sAcc$) for the different datasets. Raw involves predictions directly from all input dimensions except for S, while Proto involves predictions from the learned fair representations.

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 λ and samples $Z_i^{(t)}$ starting at $\lambda_k^{(0)} = b_k$ is governed by:*

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

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

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

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

retention function

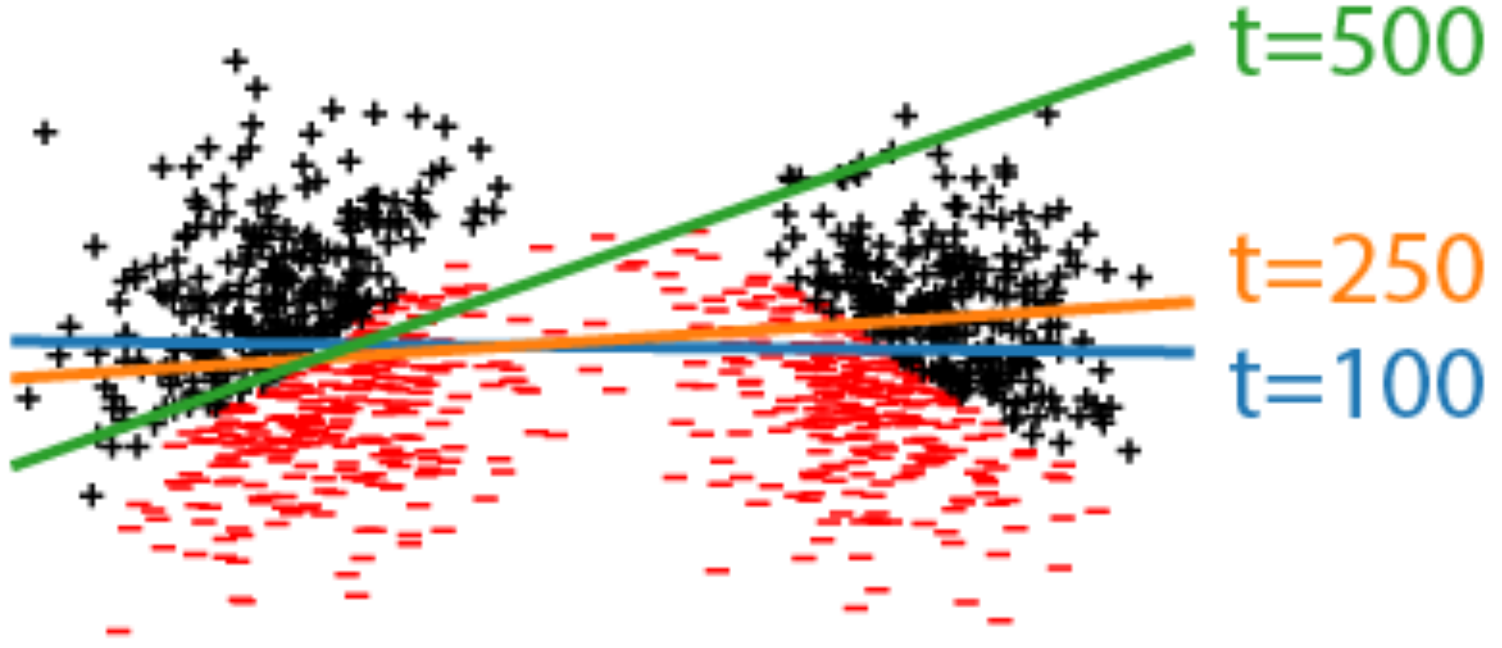


Figure 1. An example online classification problem which begins fair, but becomes unfair over time.

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

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

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

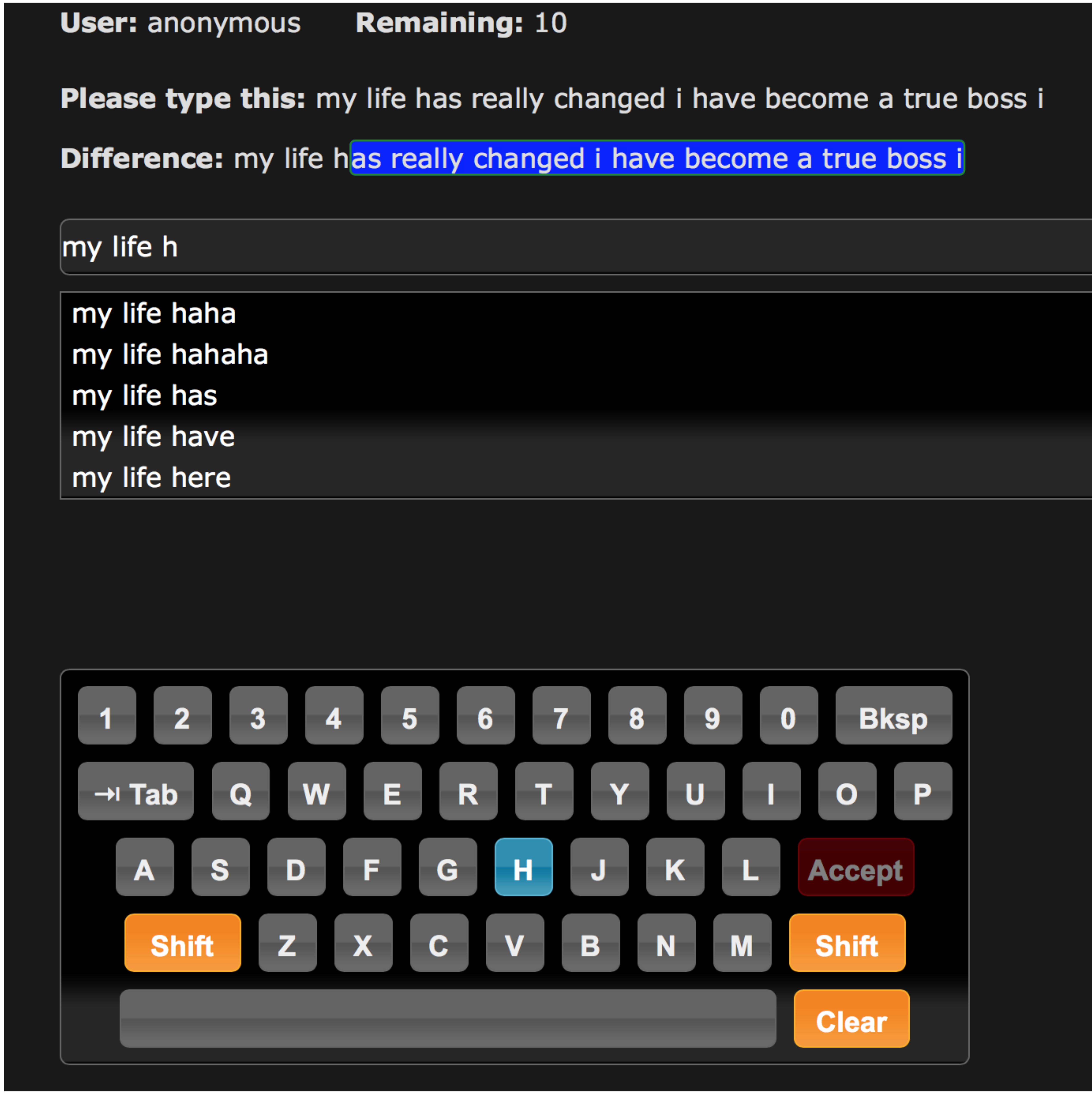
proven upper bound

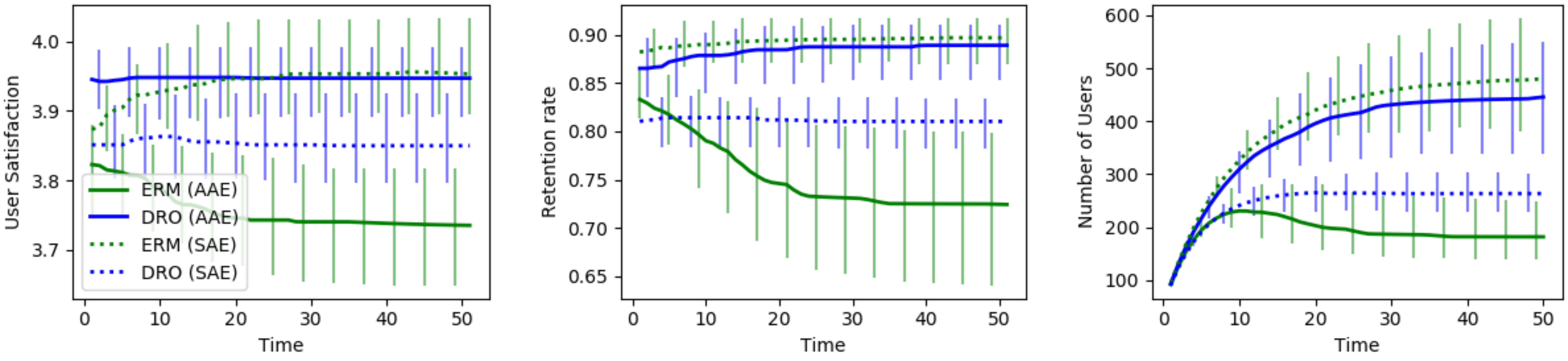
Search for best η

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.





(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.