Penalizing Unfairness in Binary Classification

M.Sc. Thesis, under the supervision of Dr. Katrina Ligett

Yahav Bechavod

HUJI

May 3, 2018

Fairness in ML???

How do we define fairness?

Definitions of Fairness

Fairness in ML

1. Ground truth unavailable
2. Ground truth available

Ground Truth Unavailable

Goal: Prevent reliance on protected attributes for prediction.

1. Changing the data
    1. Zemel et al. 2013
    2. Bolukbasi et al. 2016
2. Changing the classifier
    1. Dwork et al. 2012
    2. Kamishima et al. 2011

Ground Truth Available

Goal: Prevent situations where the errors of the algorithm are spread unevenly across the population.

1. Hardt et al. 2016
2. Woodworth et al. 2017
3. Hébert-Johnson et al. 2017
4. Kleinberg et al. 2017
5. Chouldechova 2016
6. Zafar et al. 2017

Notions of Fairness

1. Individual Fairness
2. Group Fairness

Group Fairness

Many definitions. 3 major examples:

1. Statistical Parity
   $\mathbb{P}[\hat{Y} = \hat{y} | A = 0] = \mathbb{P}[\hat{Y} = \hat{y} | A = 1], \ \hat{y} \in Y$

2. Calibration
   $\mathbb{P}[Y = y | A = a, \hat{Y} = \hat{y}] = \mathbb{P}[Y = y | \hat{Y} = \hat{y}], \ a \in \{0, 1\}, \ \hat{y} \in Y$

3. Equalized Odds
   $\mathbb{P}[\hat{Y} = \hat{y} | A = 0, Y = y] = \mathbb{P}[\hat{Y} = \hat{y} | A = 1, Y = y], \ \hat{y} \in Y, \ y \in Y$

Notions (2) and (3) are generally incompatible.

X - Non-Protected Attributes
A - Protected Attribute
Y - Label
$\hat{Y}$ - Prediction

COMPAS


PRO PUBLICA
Donate
Machine Bias
There's software used across the country to predict future criminals. And it's biased against blacks.
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

COMPAS

- Correctional Offender Management Profiling for Alternative Sanctions.
- Risk assessment tool, developed and sold by Northpointe Inc.
- Used as a judicial aid (bail decisions, in-trial).
- Arrested individuals screened in order to predict risk of recidivism, violent crimes, and more.
- Algorithm is proprietary. Makes predictions based on 137 features.
- U.S. states using COMPAS: Florida, Michigan, New Mexico, Wisconsin, Wyoming.
- ProPublica investigative report (May 2016): COMPAS is biased against African-Americans.

COMPAS

All Defendants	Low	High
Survived	2681	1282
Recidivated	1216	2035

FP rate: 32.35
FN rate: 37.40
PPV: 0.61
NPV: 0.69
LR+: 1.94
LR-: 0.55

Black Defendants	Low	High
Survived	990	805
Recidivated	532	1369

FP rate: 44.85
FN rate: 27.99
PPV: 0.63
NPV: 0.65
LR+: 1.61
LR-: 0.51

White Defendants	Low	High
Survived	1139	349
Recidivated	461	505

FP rate: 23.45
FN rate: 47.72
PPV: 0.59
NPV: 0.71
LR+: 2.23
LR-: 0.62

COMPAS

	Black Defendants Low	High		White Defendants Low	High
Survived	990	805	Survived	1139	349
Recidivated	532	1369	Recidivated	461	505
FP rate: 44.85			FP rate: 23.45		
FN rate: 27.99			FN rate: 47.72		

- FP = Labelled "high risk", did not re-offend.
- FN = Labelled "low risk", re-offended.

Learning Equalized Odds Classifiers

Learning problem:

$$\begin{array}{|ll|} \hline \underset{f \in \mathcal{H}}{\text{minimize}} & L_{\mathcal{D}}(f) \\ \text{subject to} & FPR_{A=0}(f) = FPR_{A=1}(f) \\ & FNR_{A=0}(f) = FNR_{A=1}(f) \\ \hline \end{array}$$

- $\mathcal{D}$ - Distribution over (X, A, Y)
- We denote a predictor by $\hat{Y} = f(X, A)$
- $\mathcal{H}$ - Hypothesis class
- $\ell : Y \times Y \to \mathbb{R}^+$ - Loss function
- $L_{\mathcal{D}}(f) = \underset{(x,a,y)\sim\mathcal{D}}{\mathbb{E}} \ell(f((x, a)), y)$ - Expected loss

Hardness of Learning an Equalized Odds Classifier

Theorem (Woodworth et al. 2017)

Let L^* be the hinge loss of the optimal linear predictor whose sign is non-discriminatory. Subject to the assumption that refuting random K-XOR formulas is computationally hard[a], the learning problem of finding a possibly randomized function f such that $\mathcal{L}^{hinge}(f) \leq L^* + \epsilon$ and $sign(f)$ is α-discriminatory requires exponential time in the worst case for $\epsilon < \frac{1}{8}$ and $\alpha < \frac{1}{8}$.

[a] See Daniely 2015 for a description of the problem.

Learning an Equalized Odds Classifier

Question: Can we (in many non worst-case settings) still efficiently learn an accurate equalized odds classifier?

Main contribution: A new, efficient, easy to use approach for learning equalized odds classifiers.

Our Approach

Idea: Penalize unfair solutions

Original optimization problem:

$$\begin{aligned} \underset{f\in\mathcal{H}}{\text{minimize}} \quad & L_{\mathcal{D}}(f) \\ \text{subject to} \quad & FPR_{A=0}(f) = FPR_{A=1}(f) \\ & FNR_{A=0}(f) = FNR_{A=1}(f) \end{aligned}$$

Relaxed optimization problem:

$$\begin{aligned} \underset{w\in\mathbb{R}^{d+1}}{\text{minimize}} \quad & L_{\mathcal{D}}(w) \\ \text{subject to} \quad & \mathbb{E}[w^T(x,a)|A=0,Y=0] = \mathbb{E}[w^T(x,a)|A=1,Y=0] \\ & \mathbb{E}[w^T(x,a)|A=0,Y=1] = \mathbb{E}[w^T(x,a)|A=1,Y=1] \end{aligned}$$

Relaxation:

1. Linear Classifiers - $\mathcal{H} = \{(x,a) \mapsto \langle w,(x,a)\rangle : w \in \mathbb{R}^{d+1}\}$
2. Distance from the decision boundary as a proxy for FPR's, FNR's
3. ℓ is convex

Our Approach

Relaxed optimization problem:

$$\boxed{\begin{aligned} &\underset{w\in\mathbb{R}^{d+1}}{\text{minimize}} && L_S(w) \\ &\text{subject to} && \frac{\sum_{(x,a)\in S_{00}} w^T(x,a)}{|S_{00}|} = \frac{\sum_{(x,a)\in S_{10}} w^T(x,a)}{|S_{10}|} \\ & && \frac{\sum_{(x,a)\in S_{01}} w^T(x,a)}{|S_{01}|} = \frac{\sum_{(x,a)\in S_{11}} w^T(x,a)}{|S_{11}|} \end{aligned}}$$

$S = (x_1, a_1, y_1), ..., (x_m, a_m, y_m) \in \mathcal{D}^m$ sampled i.i.d.
$S_{ay} = \{(x_i, a_i, y_i) \in S : a_i = a, y_i = y\}$

Our Approach

Which we can further simplify as:

$$\boxed{\begin{aligned} \underset{w \in \mathbb{R}^{d+1}}{\text{minimize}} \quad & L_S(w) \\ \text{subject to} \quad & w^T \overline{(x,a)}_{FP} = 0 \\ & w^T \overline{(x,a)}_{FN} = 0 \end{aligned}}$$

Where:

$$\overline{(x,a)}_{FP} = \left(\frac{\sum_{(x,a) \in S_{00}} (x,a)}{|S_{00}|} - \frac{\sum_{(x,a) \in S_{10}} (x,a)}{|S_{10}|} \right)$$

$$\overline{(x,a)}_{FN} = \left(\frac{\sum_{(x,a) \in S_{01}} (x,a)}{|S_{01}|} - \frac{\sum_{(x,a) \in S_{11}} (x,a)}{|S_{11}|} \right)$$

Convexity + Strong Duality

Note: The relaxed problem is a convex optimization problem. Moreover, strong duality holds.

Convexity:

1. Objective function: convex composed with affine, hence still convex.
2. Constraints: Two affine equality constraints.

Strong Duality: Slater's condition (trivially) holds, since $0 \in \mathbb{R}^{d+1}$ is a feasible solution.

The Lagrangian is: $\mathcal{L}(\lambda; w) = L_S(w) + \lambda_1 w^T \overline{(x,a)}_{FP} + \lambda_2 w^T \overline{(x,a)}_{FN}$

The Dual function: $g(\lambda) = \min_{w} \mathcal{L}(\lambda; w)$

...

Accuracy-Fairness Trade-Off

However: We are not interested only in the solution!

1. We can achieve far better solutions overall with little discrimination allowed
2. It is not clear that we need to exactly drive the proxy discrimination to zero. (Overfitting, only a proxy for the real difference).
3. We are also very interested in the price of fairness - how much fairness is achievable at what price?

Hence: We are interested in the entire trade-off curve.

Our Approach

In order to prevent situations where one direction of difference is 'preferable', we will consider these two variants:

Absolute value of difference:

$$\boxed{\begin{aligned} \underset{w \in \mathbb{R}^{d+1}}{\text{minimize}} \quad & L_S(w) \\ \text{subject to} \quad & |w^T \overline{(x,a)}_{FP}| \le \epsilon \\ & |w^T \overline{(x,a)}_{FN}| \le \epsilon \end{aligned}}$$

Squared difference:

$$\boxed{\begin{aligned} \underset{w \in \mathbb{R}^{d+1}}{\text{minimize}} \quad & L_S(w) \\ \text{subject to} \quad & (w^T \overline{(x,a)}_{FP})^2 \le \epsilon \\ & (w^T \overline{(x,a)}_{FN})^2 \le \epsilon \end{aligned}}$$

Fairness-Inducing Penalizers

We define the **Absolute Value Difference (AVD)** FPR penalty term to be

$$R_{FP}^{AVD}(w; S) = \left| w^T \overline{(x, a)} \right|$$

The **Squared Difference (SD)** penalizer:

$$R_{FP}^{SD}(w; S) = \left(w^T \overline{(x, a)} \right)^2$$

We therefore re-formulate as a regularized optimization problem:

$$\boxed{\underset{w \in \mathbb{R}^{d+1}}{\text{minimize}} \quad \overline{L}_S(w) + c_1 R_{FP}(w; S) + c_2 R_{FN}(w; S) + q||w||_2^2}$$

Where:

1. $R_{FP} = R_{FP}^{AVD}$ or R_{FP}^{SD}
2. $R_{FN} = R_{FN}^{AVD}$ or R_{FN}^{SD}
3. $c_1, c_2 \geq 0$ - Changing these allows for different significance balance between FP, FN and accuracy.

Training Scheme

Input: Training Set $Q \sim \mathcal{D}^m$ i.i.d.

1. Split Q randomly to training set S and test set T
2. For each c, cross-validate on S to select q_c
3. For each (c, q_c), let $w_c = \underset{w}{\text{argmin}}\, \text{Proxy}(w; S, c, q_c)$
4. Select $w^* \in \underset{w_c}{\text{argmin}}\, \text{Objective}(w_c; S)$
5. Evaluate performance using w^* on test set T

Notation:

$$\text{Objective}(w; S) = L_S(w) + d_1|FPR^S_{A=0} - FPR^S_{A=1}| + d_2|FNR^S_{A=0} - FNR^S_{A=1}|$$

$$\text{Proxy}(w; S, c, q) = \overline{L}_S(w) + c_1 R_{FP}(w; S) + c_2 R_{FN}(w; S) + q||w||_2^2$$

Main contribution: Do we really benefit from incorporating fairness considerations in the learning phase? Can't we simply learn (unfairly) then post-process?

Post-Hoc Approach

Hardt et al. 2016:

1. Learn the best (unfair) classifier $\hat{Y}$.
2. Post-process to find the best possible fair classifier $\tilde{Y}$ derived from $(\hat{Y}, A)$.

'derived' - A (possibly randomized) function of $(\hat{Y}, A)$ alone.
Note: Every derived classifier $\tilde{Y}$ can be written as:

$$\tilde{Y}|A = \begin{cases} \hat{Y} & \text{w.p. } \alpha_1 \\ 1 - \hat{Y} & \text{w.p. } \alpha_2 \\ 0 & \text{w.p. } \alpha_3 \\ 1 & \text{w.p. } \alpha_4 \end{cases} \qquad \text{where: } \sum_{i=1}^{4} \alpha_i = 1$$

Importance of Incorporating Fairness in Learning Phase

Claim: Let $\mathcal{H}$ be unconstrained. Then, for any $\epsilon \in (0, 1/4)$ there exists a distribution $\mathcal{D}_\epsilon$ such that:

a) For the Bayes optimal classifer $\hat{Y}$ trained on 0-1 loss, the post-hoc correction of $\hat{Y}$ returns a classifier $\tilde{Y}$ with $L_{\mathcal{D}}^{0-1}(\tilde{Y}) \geq 0.5$.

b) Restricting $\mathcal{H}$ to linear classifiers alone and using our approach yields a completely fair classifier w with $L_{\mathcal{D}}^{0-1}(w) = 2\epsilon$.

Conclusion: In some cases, fairness has to be actively incorporated into the learning phase.

Importance of Incorporating Fairness in Learning Phase

Consider the following example:
Each data point is written as $(A, X) = \{0, 1\}^2$, and has a label $Y \in \{0, 1\}$.

Given $\epsilon \in (0, \frac{1}{4})$, we define a distribution $\mathcal{D}_\epsilon$ over labelled examples as follows:

$$\mathbb{P}[Y = 1] = 0.5$$
$$\mathbb{P}[A = y | Y = y] = 1 - \epsilon$$
$$\mathbb{P}[X = y | Y = y] = 1 - 2\epsilon$$

Note that $\mathcal{D}_\epsilon$ is defined s.t. $A \perp X | Y$.

Importance of Incorporating Fairness in Learning Phase

a) The Bayes optimal predictor with respect to the 0-1 loss is

$$\hat{h}(X) = \underset{y \in \{0,1\}}{\text{argmax}}\, \mathbb{P}[Y = 1 | X = x]$$

which, in our case, gives $\hat{h}(X) = A$.

Fairness: Completely unfair.

$$FPR_{A=0}(\hat{h}) = 0, \quad FPR_{A=1}(\hat{h}) = 1$$
$$FNR_{A=0}(\hat{h}) = 1, \quad FNR_{A=1}(\hat{h}) = 0$$

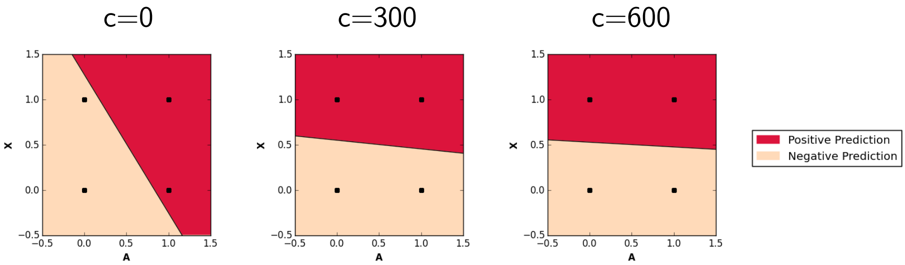
Loss: $L_{\mathcal{D}}^{0-1}(\hat{h}) = \epsilon$

Any approach to post-processing this classifier yields $\tilde{Y}$ that predicts 0 or 1 at random.

Illustration

b) Our approach

Learned decision boundary as a function of increasing penalizers' weight



Fairness: Completely fair.

$$FPR_{A=0}(\hat{Y}) = \epsilon, \quad FPR_{A=1}(\hat{Y}) = \epsilon$$
$$FNR_{A=0}(\hat{Y}) = \epsilon, \quad FNR_{A=1}(\hat{Y}) = \epsilon$$

Loss: $L_{\mathcal{D}}^{0-1}(\hat{Y}) = 2\epsilon$

COMPAS Dataset

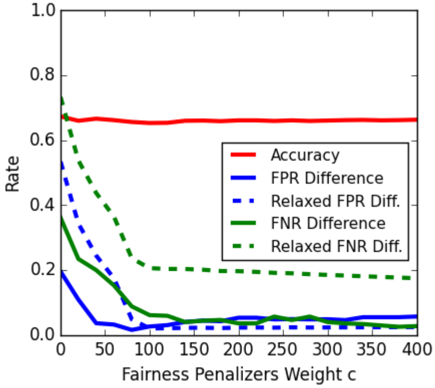
COMPAS records from Broward County, Florida 2013-2014.

	Recidivated	Did not recidivate	Total
Black	1661	1514	3175
White	822	1281	2103
Total	2483	2795	5278

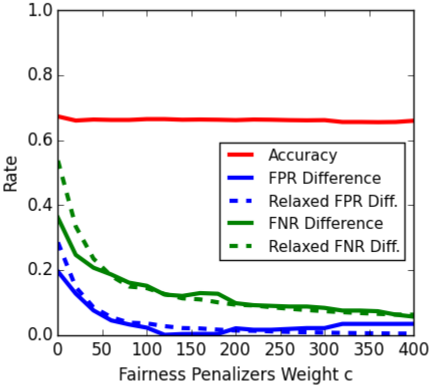
Feature	Description
Age Category	< 25, $25 - 45$, > 45
Gender	Male or Female
Race	White or Black
Priors Count	0–37
Charge Degree	Misconduct or Felony
2-year-recid. (target feature)	Whether or not the defendant recidivated within two years

Accuracy-Fairness Trade-Off

Absolute value difference penalizers



Squared difference penalizers



Experimental Results - COMPAS Dataset

	FPR Considerations			FNR Considerations			Both Considerations		
	Acc.	**D_{FPR}**	**D_{FNR}**	**Acc.**	**D_{FPR}**	**D_{FNR}**	**Acc.**	**D_{FPR}**	**D_{FNR}**
Vanilla Reg. Log. Reg.	0.672	0.20	0.30	0.672	0.20	0.30	0.672	0.20	0.30
Our Method (AVD)	0.660	0.01	0.04	0.653	0.02	0.04	0.654	0.02	0.04
Our Method (SD)	0.664	0.02	0.09	0.661	0.05	0.03	0.661	0.02	0.03
Zafar et al. 2017	0.660	0.06	0.14	0.662	0.03	0.10	0.661	0.03	0.11
Zafar et al. 2017 Baseline	0.643	0.03	0.11	0.660	0.00	0.07	0.660	0.01	0.09
Hardt et al. 2016	0.659	0.02	0.08	0.653	0.06	0.01	0.645	0.01	0.01

Adult Dataset

The Adult Dataset

1. Based on 1994 US Census data.
2. **Task:** Predict whether per year income over/under 50,000 dollars.
3. **Features:** Occupation, marital status, education, etc.
4. **Protected attribute:** Gender.

Loan Default Dataset

The Loan Default Dataset

1. Data regarding Taiwanese credit card users.
2. **Task:** Predict whether an individual will default on payments.
3. **Features:** History of past payments, age, amount of given credit, etc.
4. **Protected attribute:** Gender.

College Admissions Dataset

The College Admissions Dataset

1. Records of law school students who took the bar exam.
2. **Task:** Predict whether a student will pass the exam.
3. **Features:** LSAT score, undergraduate GPA, family income, etc.
4. **Protected attribute:** Race.

Dataset	Samples	Features	Split	Reps.	Folds	Protected	Target
COMPAS	5,278	5	70-30	5	5	Race	2-Year-Recidivism
Adult	30,162	10	30-70	5	5	Gender	Income Over/Under 50K
Default	30,000	23	30-70	5	3	Gender	Defaulting On Payments
Admissions	20,839	17	30-70	5	3	Race	Passing Bar Exam

Additional Datasets

	Adult Dataset			Default Dataset			Admissions Dataset		
	Acc.	**D_{FPR}**	**D_{FNR}**	**Acc.**	**D_{FPR}**	**D_{FNR}**	**Acc.**	**D_{FPR}**	**D_{FNR}**
Vanilla Regularized Logistic Regression	0.800	0.08	0.39	0.807	0.01	0.05	0.951	0.16	0.02
Our Method (AVD Penalizers)	0.776	0.00	0.04	0.807	0.00	0.01	0.950	0.01	0.00
Our Method (SD Penalizers)	0.783	0.00	0.09	0.806	0.01	0.02	0.950	0.00	0.00

Conclusions

1. Different definitions of fairness. task specific. Cost of fairness.
2. Given a specific definition, computational aspect.
3. Post-processing alone might not be enough.
4. Impossibility results.
5. In many real-life cases, it is possible to efficiently learn fair classifiers.

Future Work

1. Fairness in Reinforcement Learning
2. Fairness and Privacy
3. Short term + long term goals
4. Causality for fairness
5. Cases in which we cannot identify protected groups ahead of time/there are multiple number of (possibly overlapping) protected groups
6. Fairness incentives to myopic agents

Thank you!

Facebook Hate-Speech Prevention Rules

PROPUBLICA TOPICS ▾ SERIES ▾ ABOUT MORE ▾ Donate

MACHINE BIAS

Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up

We asked Facebook about its handling of 49 posts that might be deemed offensive. The company acknowledged that its content reviewers had made the wrong call on 22 of them.

by **Ariana Tobin**, **Madeleine Varner** and **Julia Angwin**, Dec. 28, 2017, 5:53 p.m. EST

WHAT ARE YOU PRO?
Create and share your own badge.

FOLLOW PROPUBLICA

Twitter Facebook

Podcast RSS

Facebook Hate Speech Prevention Rules

PROPUBLICA TOPICS ▾ SERIES ▾ ABOUT MORE ▾ Donate

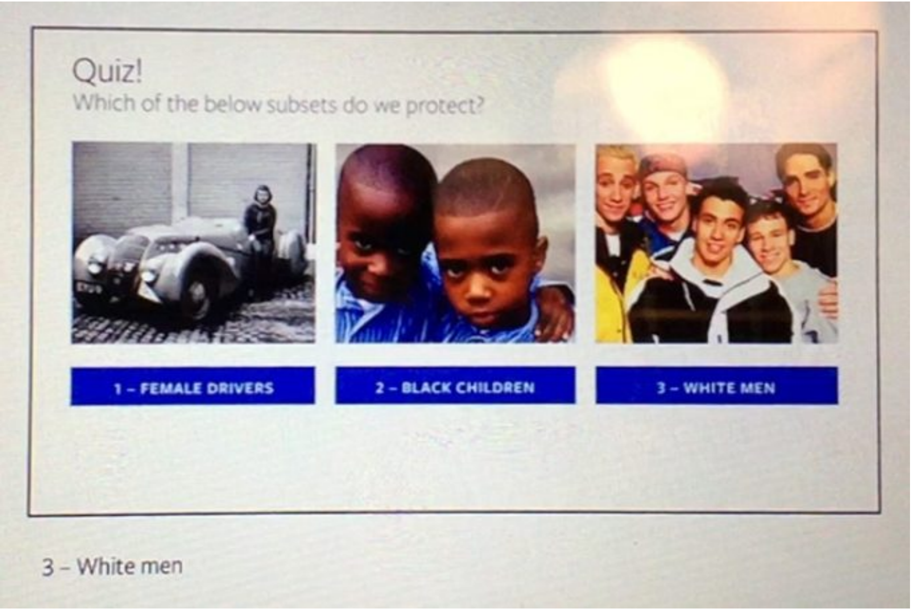
MACHINE BIAS

Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children

A trove of internal documents sheds light on the algorithms that Facebook's censors use to differentiate between hate speech and legitimate political expression.

by **Julia Angwin**, ProPublica, and **Hannes Grassegger**, special to ProPublica, June 28, 2017, 5 a.m. EDT

Facebook Hate-Speech Prevention Rules



Facebook Hate Speech Prevention Rules

Facebook's response: Cartoon attacks members of a religion, rather than the religion itself. Thus does not violate hate speech guidelines.

Facebook Hate Speech Prevention Rules

Main criticism: Rules do not provide equal protection to different groups, sub-groups are not protected.

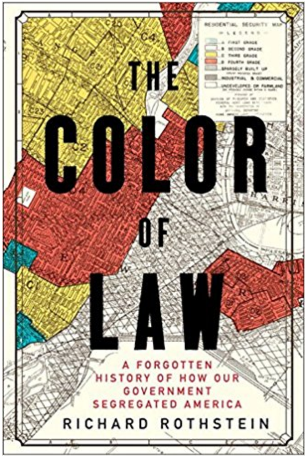
PredPol


PREDPOL®
The Predictive Policing Company®
PredPol® can make your law enforcement or security agency more effective by predicting when and where crime is most likely to occur and by using location data provide insight into your patrol operations.

PredPol

Main criticism: Algorithm perpetuates existing biases. Does not account for feedback loops.

Redlining in Online Advertisement



"In 1944, the G.I. Bill was adopted to support returning servicemen. The VA not only denied African Americans the mortgage subsidies to which they were entitled but frequently restricted education and training to lower-level jobs for African Americans who were qualified to acquire greater skills."

-Richard Rothstein, **The Color of Law: A Forgotten History of How Our Government Segregated America**

Redlining in Online Advertisement

ProPublica | ProPublica Illinois | Local Reporting Network | Data Store | Donate

PROPUBLICA TOPICS ▾ SERIES ▾ NEWS APPS GET INVOLVED IMPACT ABOUT SIGN UP

What do you think of ProPublica? Take the 2018 Reader Survey.

MACHINE BIAS

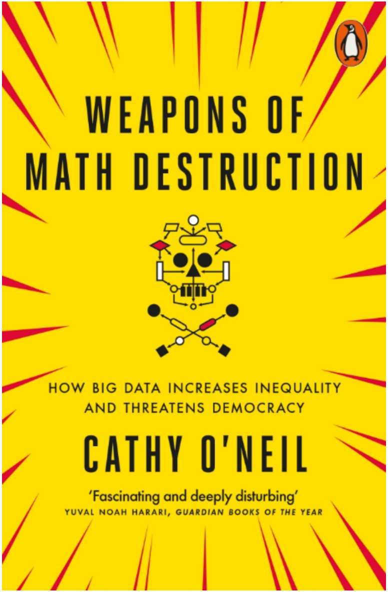
Facebook (Still) Letting Housing Advertisers Exclude Users by Race

After ProPublica revealed last year that Facebook advertisers could target housing ads to whites only, the company announced it had built a system to spot and reject discriminatory ads. We retested and found major omissions.

by **Julia Angwin**, **Ariana Tobin** and **Madeleine Varner**, Nov. 21, 2017, 1:23 p.m. EST

Main criticism: Allows for redlining specific groups based on race, gender, sexual orientation, etc.

Weapons of Math Destruction



The VAM

- The Value Added Model AKA The Educational Value-Added Assessment System.
- Used to determine how much “value” an individual teacher adds to a classroom.
- Bush’s “No Child Left Behind” Act (2001) calls for federal standards.
- Obamas “Race to The Top” Act (2009) offers states more than 4 billion US dollars in federal funds in exchange for instituting formal teacher assessments.
- Adopted in 2010 by Chicago public schools, New York City department of education and District of Columbia public schools.

The VAM

- Teachers held accountable for “student growth” - the difference between how well students performed on a test and how well a predictive model expected them to do.
- Decisions such as tenure, bonuses and firings were in many cases attached to results.
- Exact algorithm is proprietary, known to be derived in the 1980’s from agricultural crop models.

The VAM

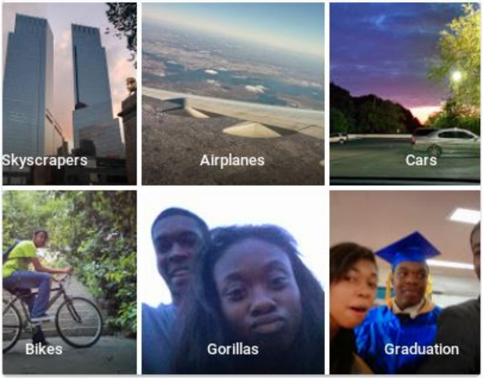
Main criticism: Algorithm is proprietary, no transparency in the decision making mechanism.

Google Photos

לעקוב

...ciné is about 40% into the IndieWeb
@jackyalcine

Google Photos, y'all fucked up. My friend's not a gorilla.



18:22 - 28 ביוני 2015

3,360 ציוצים מחדש **2,278** סימונים כאהוב

The Google Photos

Main criticism: Algorithm performs poorly on a specific sub-group in the population.