# Low rate loans for ladies, stags pay extra

## The Role of Ethics in AI/ML

Chris Stucchio
Director of Data Science, Simpl
https://chrisstucchio.com
@stucchio

# Simplest example

Supermarket theft prevention algorithm:

1. Make a spreadsheet of item SKU, shrinkage (theft) rate and price
2. Sort list by shrinkage*price.
3. Put anti-theft devices on the SKUs with the highest rates of shrinkage.

| sku | shrinkage | price | =b*c |
|---|---|---|---|
| abc123 | 0.17 | $7.24 | 1.23 |
| def456 | 0.06 | $12.53 | 0.752 |
| ghi789 | 0.08 | $8.29 | 0.66 |
| jkl012 | 0.09 | $4.50 | 0.40 |
| mno234 | 0.16 | $0.99 | 0.16 |

# Simplest example

Supermarket theft prevention algorithm:

1. Make a spreadsheet of item SKU, shrinkage (theft) rate and price
2. Sort list by shrinkage*price.
3. Put anti-theft devices on the SKUs with the highest rates of shrinkage.

**Whoops!**

# Simplest example

## Why this is bad - Virtue ethics

- Likely makes black customers feel offended.
- Most black customers **have no intention to steal,** but they suffer inconvenience anyway (checkout takes longer).
- Perpetuates racist stereotypes (which the data suggests have an element of truth).

## Why this is good - Utilitarian ethics

- Reducing theft lowers prices for all customers.
- Shops may stop carrying frequently stolen products.
- Resources (anti-theft devices) are limited and must be allocated wisely.
- Better to inconvenience 10% of customers than 100%.

# Fundamental conflict in AI Ethics

# Ethical theories

(This is the philosophy lecture)

# AI Ethics currently comes from San Francisco

Important note: I am attempting to formally write down moral premises whose proponents prefer them to be kept informal. They are mostly transmitted via social means and their proponents tend to avoid formal statements.

As such, I encourage anyone interested to investigate for themselves whether my formal statements accurately characterize implicit beliefs.

Don’t copy **algorithms** designed to solve the **wrong problem**

# Liberal virtue: Individual Fairness

# Individual Fairness

Many individual **traits** on which it is unfair to base a decision.

In code terms: for a protected trait t, for every x (other unprotected traits), your decision process must satisfy: `f(x, t1) == f(x, t2)`

Informally, your decision should never change based on protected traits.

Examples of things (possibly) unfair to use in loan underwriting/fraud checks/etc:

- Things like gender, ethnicity, caste, LGBT status, are often protected.
- Data about which privacy was guaranteed, e.g. anonymous survey data, medical data, etc.

# San Francisco ethics: Group over Individual

# Protected class

Important concept is **protected class**. What are these?

- In US: Blacks/Hispanics. Asians are legally a protected class, but practically not treated as such.
- In India: Scheduled Castes and OBCs. Muslims/other religious minorities mostly NOT protected, except in Tamil Nadu and Kerala.
- Women are often a protected class.
- In some places, homosexuals/transsexuals/disabled people/etc.

Often a protected class is connected to protected traits from above.

# Distribution across classes and “allocative harm”

Things considered unethical:

- When an algorithm has a lower than expected percentage of protected classes in it’s positive output (e.g., “lend money”). Example: IIT admissions without reservations, caused by lower scores achieved by SC/OBC.
- When an algorithm has different false positive/false negative rates across protected classes.
- Similar distributional differences.

Also called an **allocative harm**.

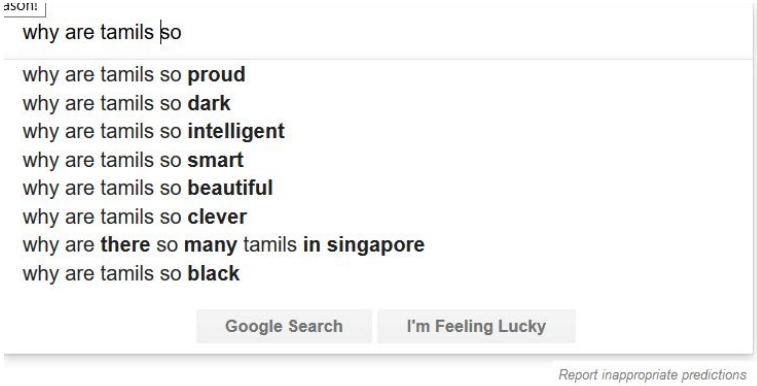
# India doesn’t have such clear groups

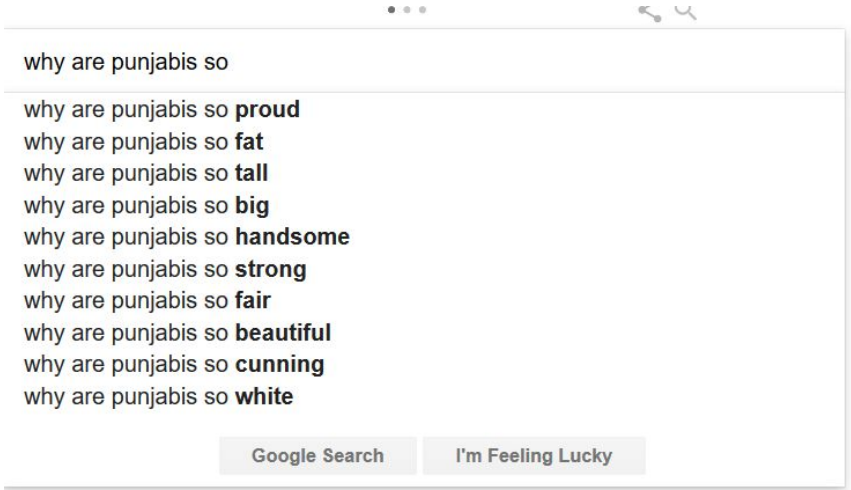
Ethnic groups very clear in US. Far less clear in India. What is a Marathi?

- A person who grew up speaking Marathi in a village in northern Karnataka?
- A Muslim who’s family lived for 5 generations in Maharashtra?
- The child of a Frenchman and a Marathi who grew up in Pune and speaks native Marathi?
- A Jewish person born in Israel who speaks Marathi at home, who’s Marathi speaking grandparents migrated from Kerala? (Wikipedia says there are about 20k of them.)

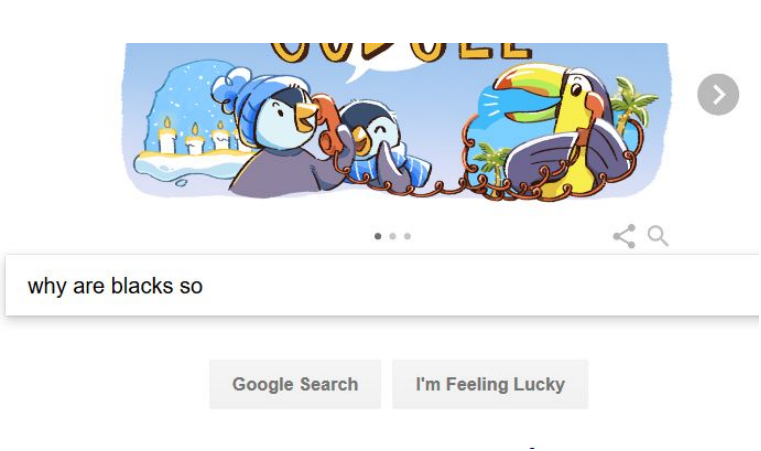
# San Francisco virtue: not noticing “problematic” things

# Indian Google notices everything



# San Francisco Google notices nothing




“As engineers, we’re trained to pay attention to the details, think logically, challenge assumptions that may be incorrect (or just fuzzy), and so on. These are all excellent tools for technical discussions. But they can be terrible tools for discussion around race, discrimination, justice...because questioning the exact details can easily be perceived as questioning the overall validity of the effort, or the veracity of the historical context.”

- Urs Hölzle, S.V.P. at Google

# AI may notice things we don’t want it to

“*Bias should be the expected result whenever even an unbiased algorithm is used to derive regularities from any data; bias is the regularities discovered.*”

[Semantics derived automatically from language corpora necessarily contain human biases]

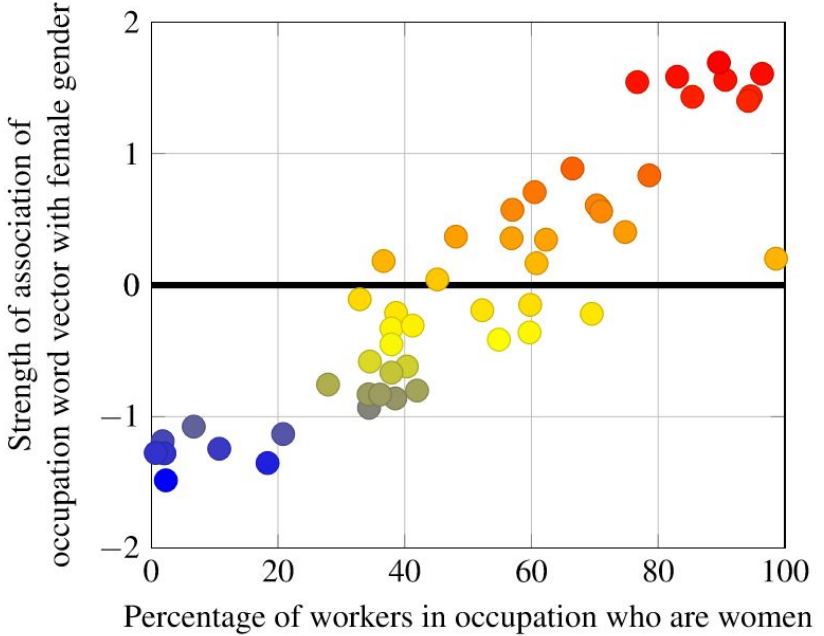



**Figure 1.** Occupation-gender association
Pearson’s correlation coefficient $\rho = 0.90$ with $p$-value $< 10^{-18}$.

# Utilitarianism: the greatest good for the greatest number

# Utilitarian case for detecting fraud

We have 1 lac to lend out.

- Lend it to Prashant who invests in his farm, then repays.
- Re-lend it to Mukti who spends on her children's education, they help her repay it with their higher earnings.
- Freddie the Fraudster runs away with the money and spends it on ganja. **No more capital to lend.**

Good underwriting directs capital to from wasteful uses to productive ones.

More fraud implies good borrowers must pay more interest.

# Utilitarian case for detecting fraud

Assumptions:

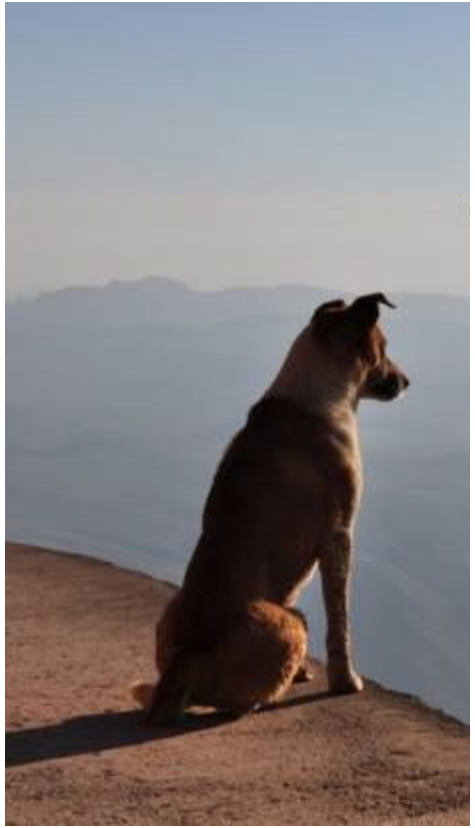
**Your product has value.** (If you don't believe this, no one is harmed by refusing them your product. Also quit your job.)

**Capitalism mostly works.** Lending to people who repay is generally more socially useful than lending to those who don't.

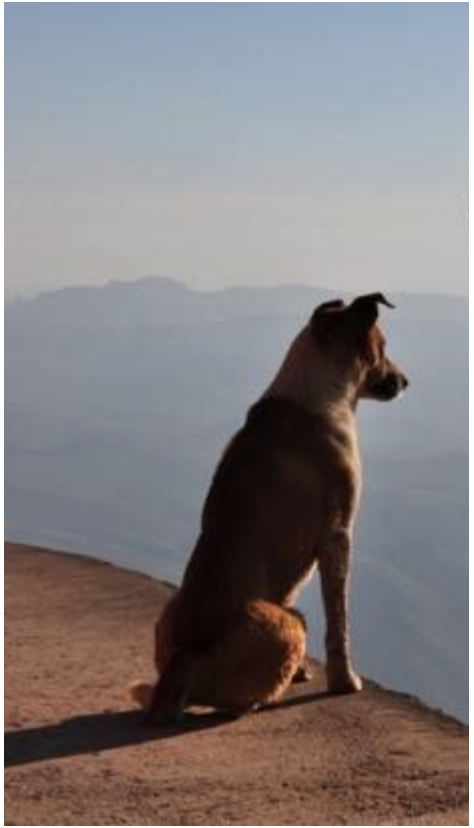
Note: This assumption does not imply anarcho capitalism. It implies government should tax the wealthy and give to poor in accordance with need, lenders should lend in accordance with ability to repay, and these are two separate things.

# Key questions

How much utility will you sacrifice for virtue?



How much individual fairness will you sacrifice for group rights?



# What does AI/ML actually do?

# How does an AI/ML system see the world?

Lots of talk about bias. Important to understand how algorithms actually behave.

Must use theory or synthetic data for this. Goal is to answer the question:

**If the world looks like X, what will an algorithm do?**

# Simple model: linear regression

Assume we have input data as a d-dimensional vector x, and output is a scalar value y.

**Input:** `X = [ income, in_north_india, mobile_or_desktop, previous_month_spending]`

**Output:** `Y = Current month spending`

Goal of ML is to use X to predict Y, and then make decisions on this basis.

# Simple model: linear regression

Modeling assumption:

```
Y = dot(alpha, X) + beta + err.rvs()
```

The value `err.rvs()` is a noise term.

```
Y = alpha[0]*income + alpha[1]*in_north_india +
alpha[2]*mobile_or_desktop + alpha[3]*previous_month_spending +
beta
```

So how does it work?

# Simple model: linear regression

```
> alpha_true = [1,2,3]

> data = norm(0,1).rvs((N, nvars))
> output = dot(data, alpha_true) + norm(0,1).rvs(N)

> alpha_estimated = lstsq(data, output)

array([ 0.98027674,  2.0033624 ,  3.00109578])
```

Linear regression reproduces the true model, with small errors.

# Does linear regression become biased?

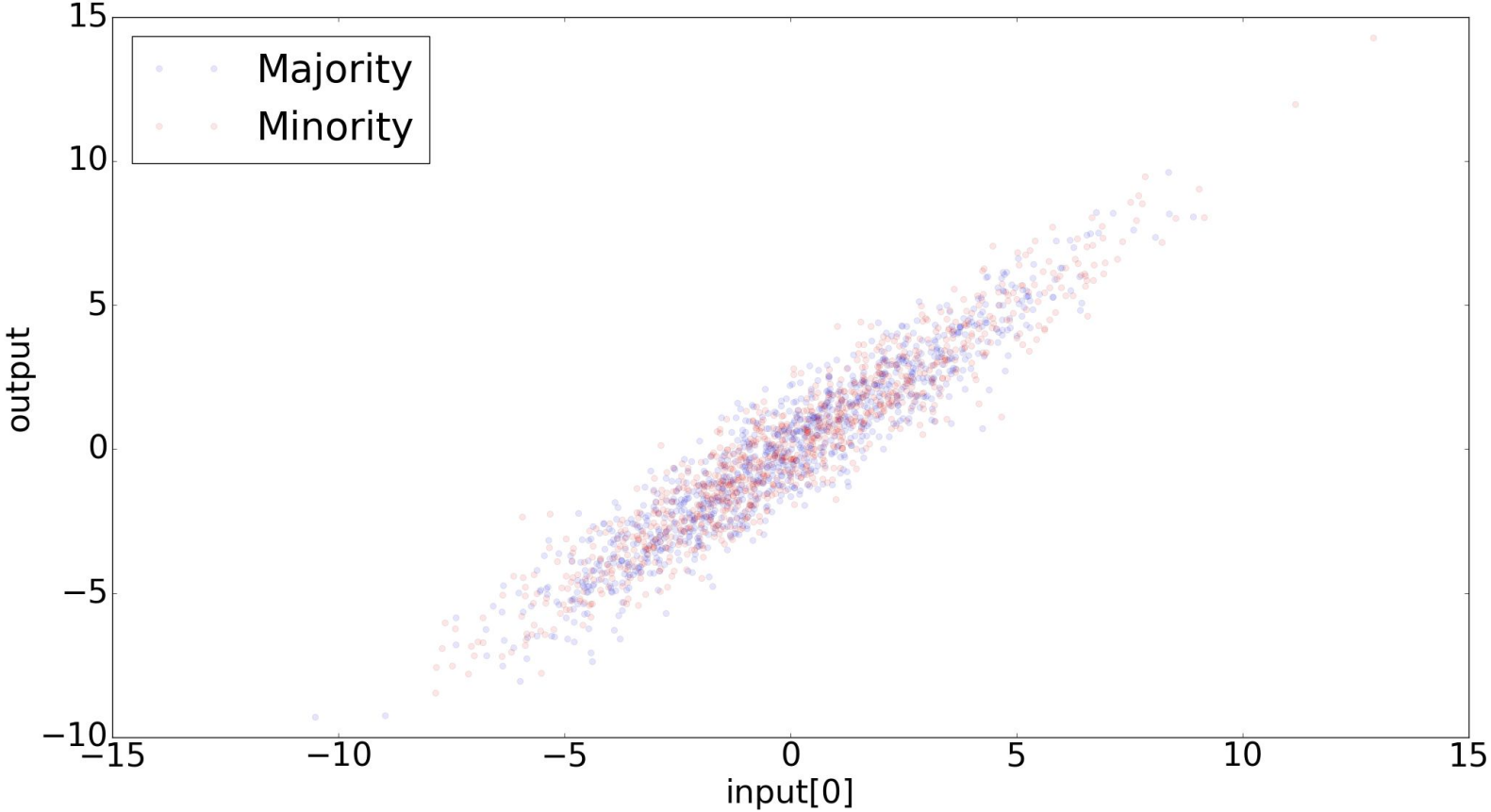
Assume protected class **doesn't matter.**

```
> alpha_true = [1,2,0]
> data = norm(0,1).rvs((N, nvars))
> data[:,2] = bernoulli(0.25).rvs(N) # 25% of people are in the protected class
> output = dot(data, alpha_true) + norm(0,1).rvs(N)

> alpha_estimated = lstsq(data, output)

array([ 1.02063423,  2.0013437 ,  -0.00118572])
```

Algorithm **learns that protected class is irrelevant**. No bias/unfairness yet.


Majority
Minority
output
input[0]
15
10
5
0
−5
−10
−15
−10
−5
0
5
10
15

# Does linear regression become biased?

Linear regression is, in this case:

- Allocatively fair - reds and blues receive equal representation in the high scoring set.
- Individually fair - reds and blues are treated identically.
- Utilitarian - it’s accurately predicting outputs.
- It virtuously does not notice anything problematic (since there is nothing problematic to notice).

*“If the police have discriminated in the past, predictive technology reinforces and perpetuates the problem, sending more officers after people who we know are already targeted and unfairly treated”*- BÄRÍ A. WILLIAMS

# Does linear regression become biased?

Let’s build a data set where “historically”, protected class performs worse.

```
> alpha_true = [1,2,0]
> data[:,2] = bernoulli(0.25).rvs(N) # 25% of people are in the protected class
> data[where(data[2] == 1),0:2] = norm(-2,1).rvs((sum(where(data[2] == 1)), nvars-1))
> data[where(data[2] == 0),0:2] = norm(0,1).rvs((sum(where(data[2] == 0)), nvars-1))
> output = dot(data, alpha_true) + norm(0,1).rvs(N)
```

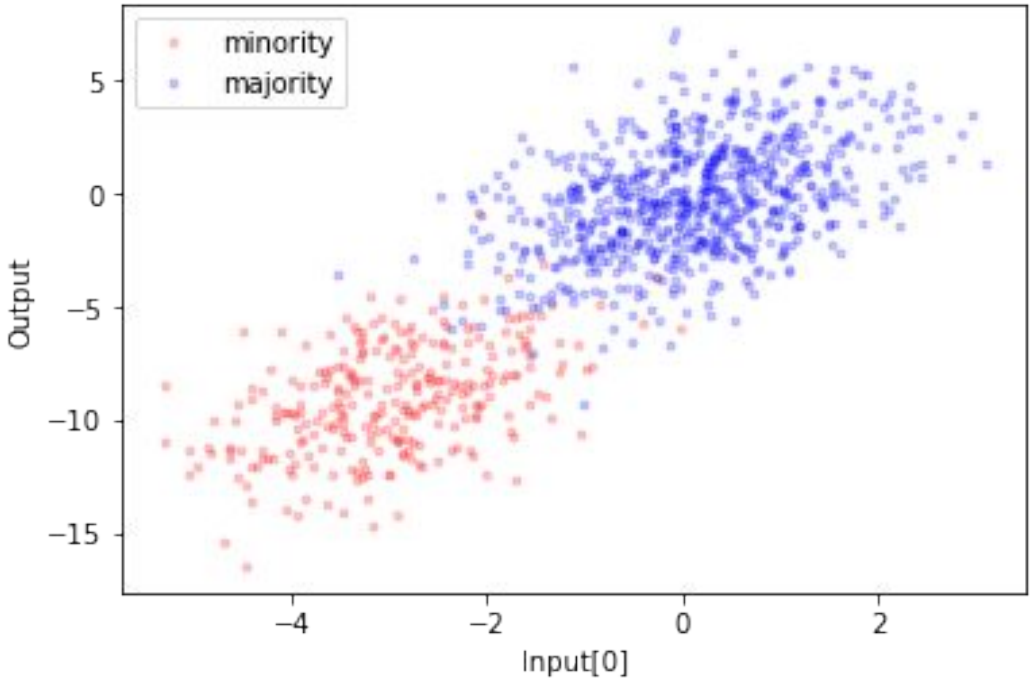
# Does linear regression become biased?

Key point: in this data set, nearly every protected class member performs worse than nearly every majority member.

```
> percentile(output[where(data[:,2] == 1)], 2.5), percentile(output[where(data[:,2]
== 1)], 97.5) # Protected class
(-13.706516466417577, -4.6637677518715961)

> percentile(output[where(data[:,2] == 0)], 2.5), percentile(output[where(data[:,2]
== 0)], 97.5) # Majority class
(-4.9236907370243426, 4.8626396540953456)
```

# Does linear regression become biased?



# Does linear regression become biased?

Let’s do some machine learning:

```
> alpha_estimated = lstsq(data, output)

array([ 1.02063423,  2.0013437 ,  0.00216572])
```

Algorithm **learns that protected class is irrelevant**, provided you have information on other predictors.

What actually matters are the other predictive factors (e.g. income, purchase history).

# Does linear regression become biased?

Linear regression is, in this case:

- Allocatively unfair - it predicts lower scores for reds.
- Individually fair - reds and blues are treated identically.

- Utilitarian - it’s accurately predicting repayments or other useful factor.

- Algorithm notices that red group underperforms, which is potentially problematic.

*“...artificial intelligence will reflect the values of its creators...we risk constructing machine intelligence that mirrors a narrow and privileged vision of society, with its old, familiar biases and stereotypes.”* - Kate Crawford

# What if the input data is biased?

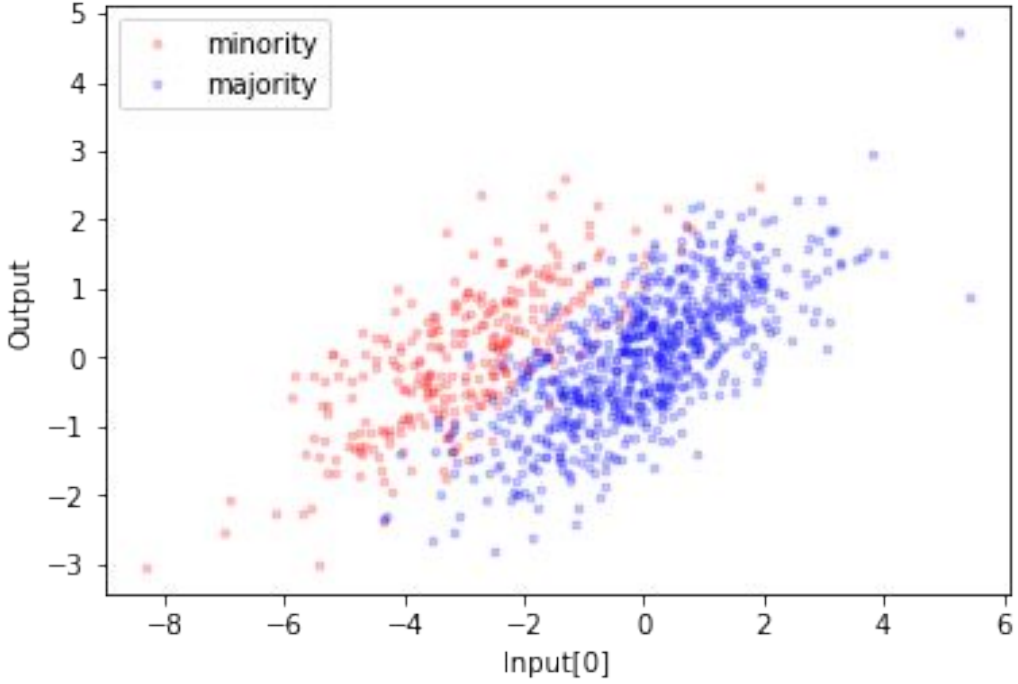
Let’s build a data set where the inputs are biased.

```
> true_value = norm(0,1).rvs(N)
> data[:,2] = bernoulli(0.25).rvs(N) # 25% of people are in the protected class
> data[:,0] = true_value + norm(0,1).rvs(N)
> data[:,1] = true_value + norm(0,1).rvs(N)
> data[where(data[:,2] == 1),0:nvars-1] -= 3 #Bias added here

> output = true_value
```

If we used our old predictor, we would have a biased prediction of the output.

# What if the input data is biased?



# What if the input data is biased?

But what if we use this new data set as input?

```
> lstsq(data, output)

[ 0.33071515,  0.32115862,  1.93781581]
```

If input data **subtracts** from the minority group due to bias, then the output data **adds back what was subtracted.**

I.e., linear regression has **fixed bias in input data.**

# What if the input data is biased?

Algorithm is now **accurately** predicting outputs by **explicitly discriminating.**

E.g., a minority member with a low score is likely to be selected while a majority member with the same score is not.

- **Discriminating algorithm:** residual = 317
- **Ignoring protected class:** residual = 719 (much less accurate)

**Not possible to be simultaneously fair to individuals and fair to groups.**

# What if the input data is biased?

A check for whether algorithm is (statistically) biased:

- Create new data set, of the form [ algorithm_output, protected_class].
- Build a new algorithm trained on this data set.
- If protected_class changes output of new algorithm, then old algorithm is biased.

**Machine learning finds hidden features that predict our goals. Bias is just another hidden feature.**

# What if the input data is biased?

Linear regression is, in this case:

- Allocatively fair - reds and blues are equally represented in the high score group.
- Individually unfair - reds and blues treated differently given same inputs.
- Utilitarian - it’s accurately predicting repayments or other useful factor.
- Virtue of not noticing is complex.

Table 3: Gender Difference in the Repayment of Microcredit

Women more likely to repay

| | Khasi and Patro (N = 560) | |
|---|---|---|
| | SP1 | SP2 |
| *female* | 0.198*** | 0.214*** |
| *Khasi* | 0.048 | 0.074 |
| *female x Khasi* | -0-063 | -0.102 |
| *Group* | | -0.129*** |
| Age | | 0.004*** |
| Education | | -0.002 |
| Married | | -0.037 |
| Farmer | | -0.021 |
| Assets | | 0.012*** |
| Constant | 0.471*** | 0.322*** |
| $R^2$ | 0.04 | 0.095 |

# Are Women “Naturally” Better Credit Risks in Microcredit?

**Table 3. Gender and loan repayment**

In this table we analyze the impact of gender on loan repayment both in terms of *PaR30* (panel A) and *write-offs* (panel B). *DumNGO* is a dummy that is 1 if the MFI is an NGO and 0 otherwise, *DumGroup* is a dummy that is 1 if the MFI provides loans on a group basis (such as village-bankers or group-lenders). *DumRural* is 1 if the MFI operates mainly in rural areas and 0 otherwise. *Dum* ive-based salaries and 0 otherwise, *HDI* is the human development index. All other variables are defined as in Table 1. *OLS* indicates that pool d. *RE* means that a pooled random effects model has been estimated and *FEVD* means that the Fixed Effects Vector Decomposition-estimator has in parentheses. *, ** and *** denote statistical significance at the 10%, 5% and 1% significance level, respectively.

Women more likely to repay

**Par**

| *Dep* | ) LS | (2) RE | (3) FEVD | (4) OLS | (5) RE | (6) FEVD |
|---|---|---|---|---|---|---|
| **gender** | | | | | | |
| women clients | -0.02 | -0.05 | -0.05 | | | |
| | (0.015)* | (0.038)* | (0.003)*** | | | |
| conscious gender bias | | | | -0.01 | -0.02 | -0.02 |
| | | | | (0.005)*** | (0.012)* | (0.001)*** |
| **MFI-controls** | | | | | | |
| *general* | | | | | | |
| Experience | 0.002 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | (0.001)*** | (0.001) | (0.000) | (0.002) | (0.000) | (0.000) |
| lnTA | -0.02 | -0.01 | -0.01 | -0.01 | -0.01 | -0.01 |
| | (0.004)*** | (0.002)*** | (0.001)*** | (0.002)*** | (0.002)*** | (0.004)*** |
| Loansize | 0.02 | -0.01 | -0.02 | 0.01 | 0.01 | 0.01 |
| | (0.006) | (0.005)*** | (0.002)*** | (0.004) | (0.005) | (0.001) |
| Portfolio growth | -0.05 | -0.02 | -0.02 | -0.07 | -0.02 | -0.02 |
| | (0.008)*** | (0.004)*** | (0.002)*** | (0.007)*** | (0.004)*** | (0.002)*** |
| *Legal status* | | | | | | |
| DumNGO | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.02 |
| | (0.007) | (0.016) | (0.001) | (0.004)*** | (0.011) | (0.001)*** |
| *Loan methodology* | | | | | | |
| DumGroup | 0.00 | -0.01 | -0.01 | 0.01 | 0.00 | 0.00 |
| | (0.006) | (0.018) | (0.002)*** | (0.005) | (0.012) | (0.002) |
| DumRural | -0.04 | -0.01 | -0.03 | -0.03 | -0.03 | -0.04 |
| | (0.008)*** | (0.009) | (0.003)*** | (0.005)*** | (0.011)*** | (0.002)*** |

# Women and Repayment in Microfinance

**Table 5. Binary regression with LIWC (repayment = 1)***

| Variable | Beta (Std. E) | Variable | Beta (Std. E) | Variable | Beta (Std. E) | Variable | Beta (Std. E) |
|---|---|---|---|---|---|---|---|
| **Financial and basic text variables:** | | **LIWC dictionary:** | | | | | |
| **Amount Requested(x $10^5$)** | **-7.163 (0.3668)** | Swear words | 35.5112 (35.275) | Past words | -2.1032 (1.9895) | P | |
| **Credit Grade HR** | **-0.8551 (0.0844)** | **Filler words** | **13.3939 (6.224)** | Inhibition words | -2.3047 (3.4172) | | |
| **Credit Grade E** | **-0.4642 (0.0817)** | Perception words | 13.4328 (10.839) | Home words | -2.3822 (1.7643) | | |
| **Credit Grade D** | **-0.3383 (0.0623)** | **Relative words** | **9.1729 (2.3748)** | Hear words | -2.4191 (14.038) | | |
| **Credit Grade C** | **-0.1959 (0.0559)** | Friend words | 9.7894 (7.0217) | I words | -2.7392 (8.1836) | P | |
| **Credit Grade A** | **0.7837 (0.0802)** | Anxiety words | 8.7494 (8.9305) | Tentative words | -2.8712 (2.0522) | | |
| **Credit Grade AA** | **0.2838 (0.0692)** | Negate words | 6.0709 (3.3228) | Non-fluency words | -3.2295 (9.518) | | |
| **Debt To Income** | **-0.0906 (0.0186)** | Insight words | 5.0732 (2.8214) | Anger words | -3.2911 (9.7405) | | |
| Images | 0.0599 (0.0389) | We words | 4.1277 (8.3628) | **Achieve words** | **-3.3204 (1.5601)** | A | |
| **Home Owner Status** | **-0.3199 (0.0381)** | Pronoun words | 3.7935 (9.9981) | Incline words | -3.5433 (2.3316) | Fu | |
| **Lender Interest Rate** | **(0** | | | | | | 5689 (2.6575) |
| Bank Draft Fee Annual Rate | -3 (1 | | | | | | 714 (14.463) |
| **Prior Listings** | **-0.0256 (0.0058)** | Quantitative words | 2.7495 (1.9363) | | | | -3.7248 (2.407) |
| Number of words in Description(x $10^4$) | -3.494 (1.96) | Articles | 2.457 (2.0896) | **Socia** | | | **-4.2882 (1.5697)** |
| Number of spelling mistakes | -0.0124 (0.0068) | Numbers words | 2.2907 (2.7328) | Health words | | | 7602 (10.721) |
| SMOG | -0.0252 (0.0209) | Preposition words | 2.1719 (1.8415) | **Certain words** | | | **-5.24 (2.7262)** |
| Words with 6 letters or more | 0.4455 (0.5716) | Conjoint words | 1.8673 (1.8392) | **Present words** | **-6.223 (1.7067)** | | |
| Number of words in the title | -0.0062 (0.6035) | Auxiliary verbs words | 1.7732 (2.3818) | **Human words** | **-7.5781 (3.5803)** | | |
| **(Intercept)** | **3.6557 (0.6035)** | Affect words | 1.2929 (1.5234) | **Space words** | **-8.2317 (2.5648)** | | |
| | | Discrepancy words | 1.2769 (2.686) | **Future words** | **-8.4576 (3.5391)** | | |
| | | Cognitive mechanism words | 0.7625 (1.8828) | **Motion words** | **-9.1849 (2.8071)** | | |
| | | Negative emotion words | 0.7453 (4.7407) | **Time words** | **-9.4218 (2.3077)** | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | Bios words | -1.3376 (2.6575) |
| | | | | | | Assent words | -1.3651 (14.463) |
| | | | | | | Family words | -1.4804 (2.8298) |
| | | | | | | I pronoun words | -1.72 (10.026) |
| | | | | | | Death words | -16.3445 (10.721) |
| | | | | | | **Body words** | **-19.1156 (5.8326)** |
| | | | | | | **Religion words** | **-20.2865 (6.7741)** |
| | | | | | | **Feel words** | **-24.617 (11.734)** |
| | | | | | | See words | -10.5021 (11.597) |
| | | | | | | Leisure words | 0.5548 (2.6577) |

* Bold face for P-value ≤ 0.05. For brevity we do not report in this table the estimates of the demographics variables such as location, age, gender and race.

Religious people and people with medical issues less likely to repay loans.

# When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications

| 1 | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | | | negeq_rate_s | | |
| | (1) | (2) | (3) | | (5) |
| pct_black_s | 0.369*** | 0.355*** | 0.253*** | | 0.192*** |
| | (0.005) | (0.005) | (0.006) | | (0.007) |
| pct_asian_s | | −0.146*** | −0.123*** | −0.120*** | −0.115*** |
| | | (0.004) | (0.004) | (0.004) | (0.004) |
| pct_latino_s | | 0.074*** | −0.015*** | 0.032*** | −0.008 |
| | | (0.004) | (0.005) | (0.004) | (0.005) |
| pct_poverty_s | | | 0.266*** | | 0.166*** |
| | | | (0.008) | | (0.010) |
| pct_single_mother_s | | | | 0.322*** | 0.200*** |
| | | | | (0.010) | (0.012) |
| Constant | −0.057*** | −0.025*** | 0.004 | −0.046*** | −0.020*** |
| | (0.006) | (0.007) | (0.007) | (0.007) | (0.007) |
| Observations | 23,697 | 23,697 | 23,638 | 23,410 | 23,410 |
| $R^2$ | 0.173 | 0.225 | 0.261 | 0.261 | 0.270 |
| Adjusted $R^2$ | 0.173 | 0.225 | 0.261 | 0.261 | 0.270 |
| Residual Std. Error | 101.611 (df = 23695) | 98.334 (df = 23693) | 96.170 (df = 23633) | 96.577 (df = 23405) | 96.007 (df = 23404) |
| F Statistic | 4,947.031*** (df = 1; 23695) | 2,296.621*** (df = 3; 23693) | 2,084.990*** (df = 4; 23633) | 2,067.323*** (df = 4; 23405) | 1,729.421*** (df = 5; 23404) |
| *Note:* | | | | | *p<0.1; **p<0.05; ***p<0.01 |

Blacks less likely to repay than asians

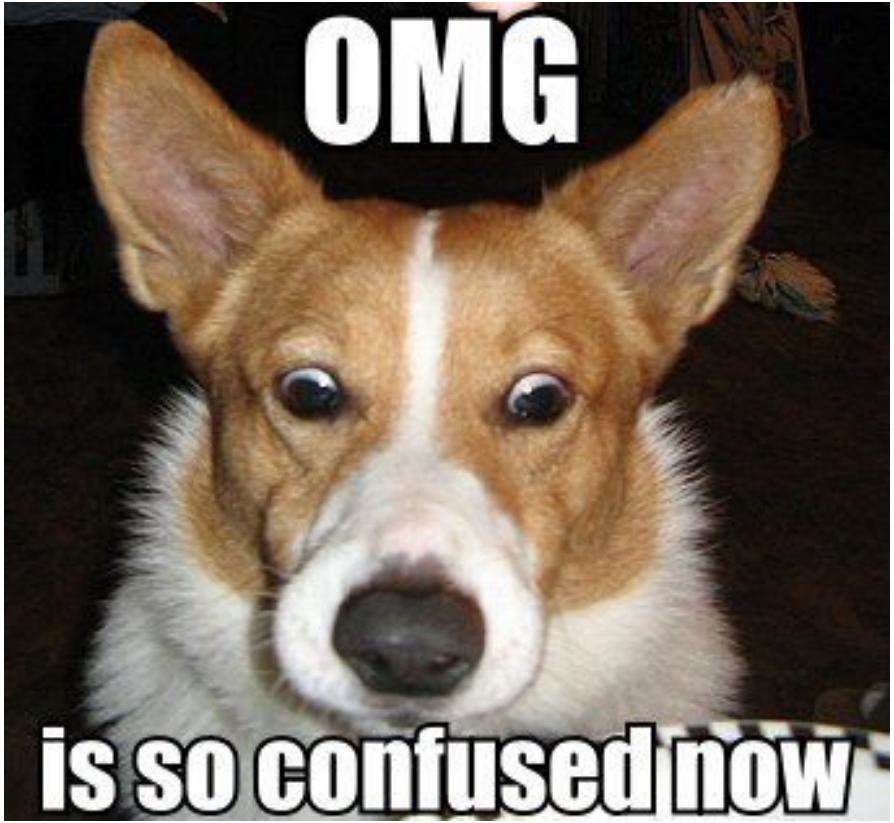
[On the relationship between negative home owner equity and racial demographics]

# What if the input data is biased?

Suppose model says +1*female:

**Virtuous** interpretation: “Bias in measuring assets or is_farmer of females.”

**Problematic** interpretation: “Females are intrinsically more likely to repay loans, holding all other factors equal.”



# Protected class is just another feature

*“If we allowed a [statistical] model to be used for college admissions in 1870, we’d still have 0.7% of women going to college.”* - Cathy O’Neil

*“If we allowed a model to be used for credit approvals when our only merchants were Zomato and BookMyShow, we’d still approve 0% of Grofers customers.”* - No one ever said this

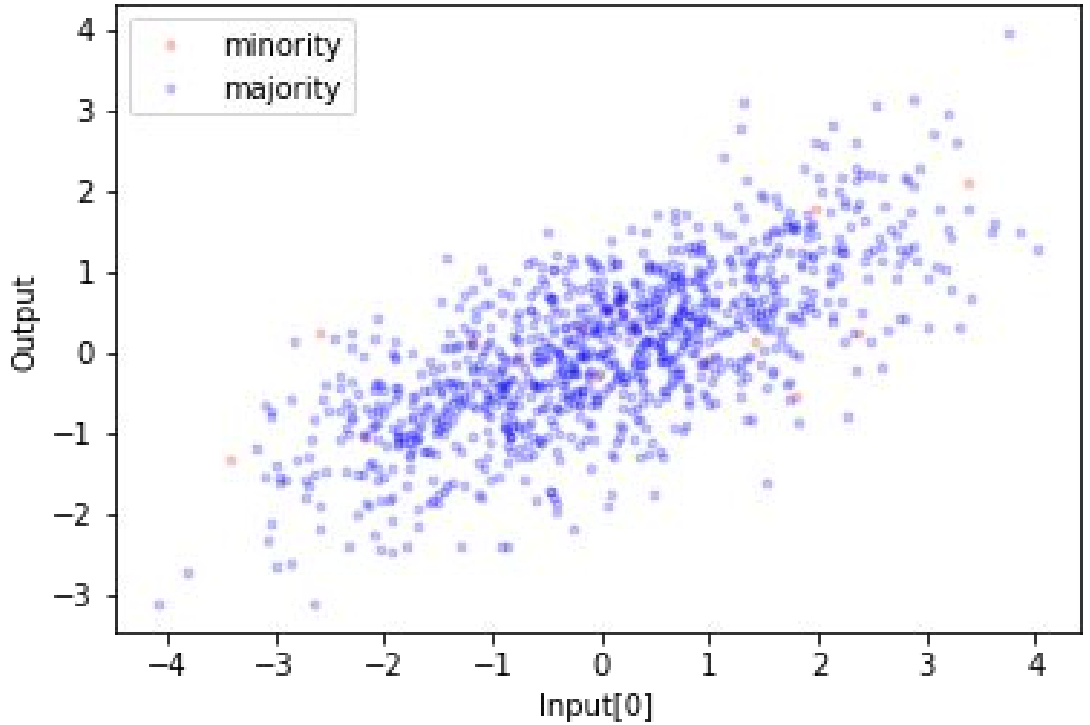
# What if the training data excludes protected class?

Let’s build a data set where the inputs have very few members of the protected class:

```
> true_value = norm(0,1).rvs(N)
> data[:,2] = bernoulli(0.01).rvs(N) # 1% of people are in the protected class
> data[:,0] = true_value + norm(0,1).rvs(N)
> data[:,1] = true_value + norm(0,1).rvs(N)

> output = true_value
```

# What if the training data excludes protected class?



# What if the training data excludes protected class?

Running the model yields:

```
> lstsq(data, output)

array([ 0.33263409,  0.34309795,  0.04731096])
```

Residual bias increases from 0.01 to 0.04, sometimes a bit bigger.

Theory of linear regression says error is O(1/sqrt(n)), where n = # of samples in protected class.

# What if the training data excludes protected class?

Running the model ignoring protected class data point yields:

```
> lstsq(data[:,0:2], output)

array([ 0.33264308,  0.34317214])
```

If protected class performs better than other equivalent non-protected class members, this is biased against them.

If protected class performs worse than other equivalent non-protected class members, this is biased in favor of them.

# Protected class is just another feature

*“If we allowed a model to be used for taxi drivers in Maharashtra in 1948, we’d still have 0% of Biharis driving taxis.”* - Paraphrased

- Maharashtra, 1948-whenever: Hard to get rickshaw license if you aren’t Marathi.
- Maharashtra, today: Lots of Biharis driving for Uber.
- Shiv Sena wants Uber shut down, and laws making it explicitly illegal for non-Marathis to drive autos.

# Protected class is just another feature

*“If we allowed a model to be used for college admissions in 1870, we’d still have 7% of Jews going to college.”* - Paraphrased

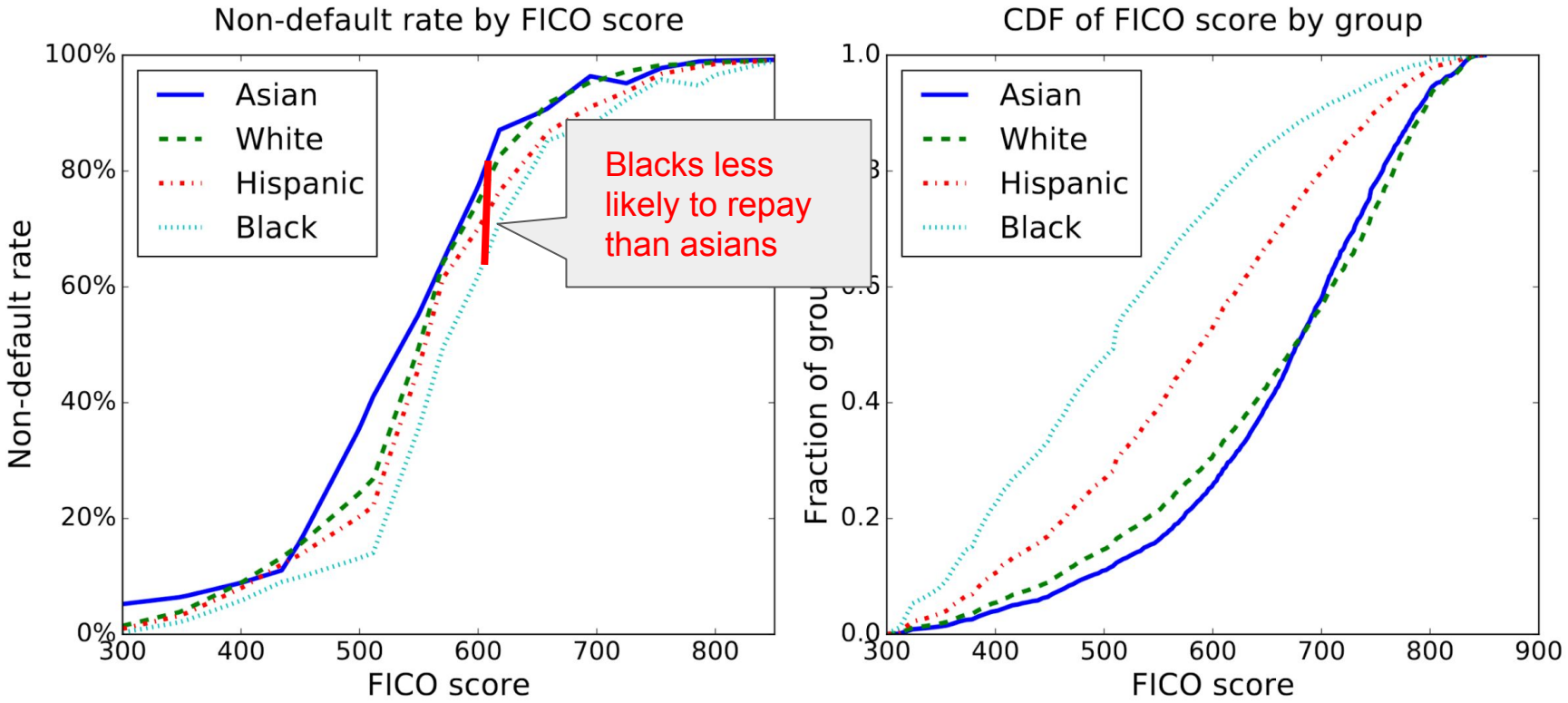
- 1908. Colleges start using a model for college admissions, trained on “white, Christian men from affluent families”.
- 1922 - number of Jews triples from 7% to 21%. President of Harvard drops the model, due to this “crisis”.
- 1933 - % of Jews back down to 15%.

Old man: "*Can you tell me, sir, are you Catholic or are you Protestant?*"

George Bernard Shaw: "*I am an atheist! It means that I do not believe in God.*"

Old man: "*I think I understand. But is it the Catholic God, or the Protestant God, that you don't believe in?*"

# The unpleasant tradeoff




# Equality of Opportunity in Supervised Learning

# Exposing tradeoffs

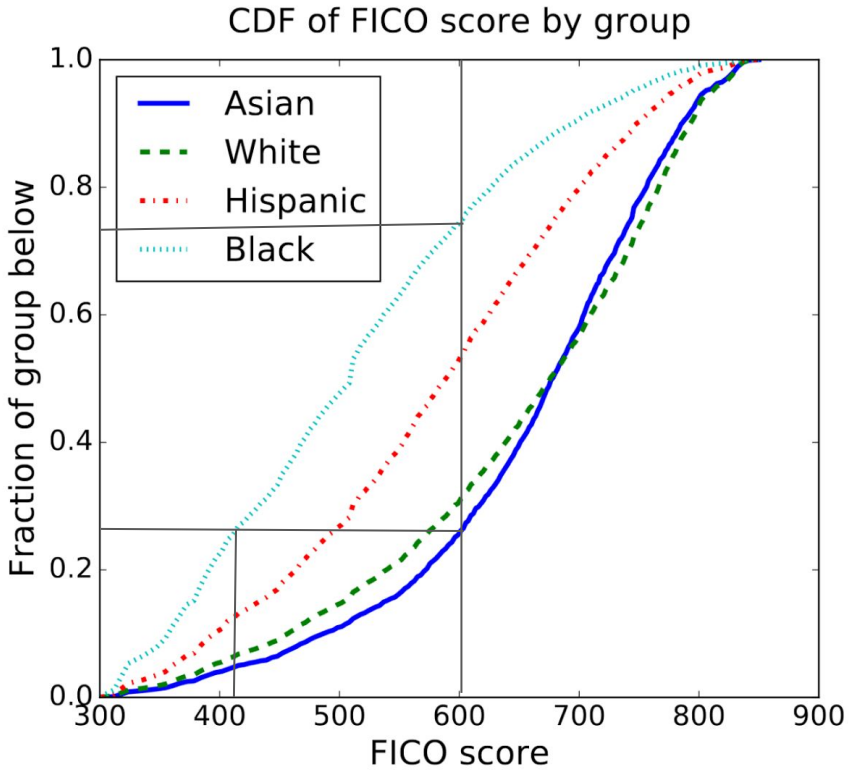
If we choose fixed cutoff of FICO 600, we reject 75% of blacks, 25% of Asians.

**Violates principle of group rights.**

If we choose cutoff of 600 for Asians, 410 for blacks, we accept 75% of both groups.

**Violates principle of individual fairness.**

Must make tradeoffs!



# Exposing tradeoffs

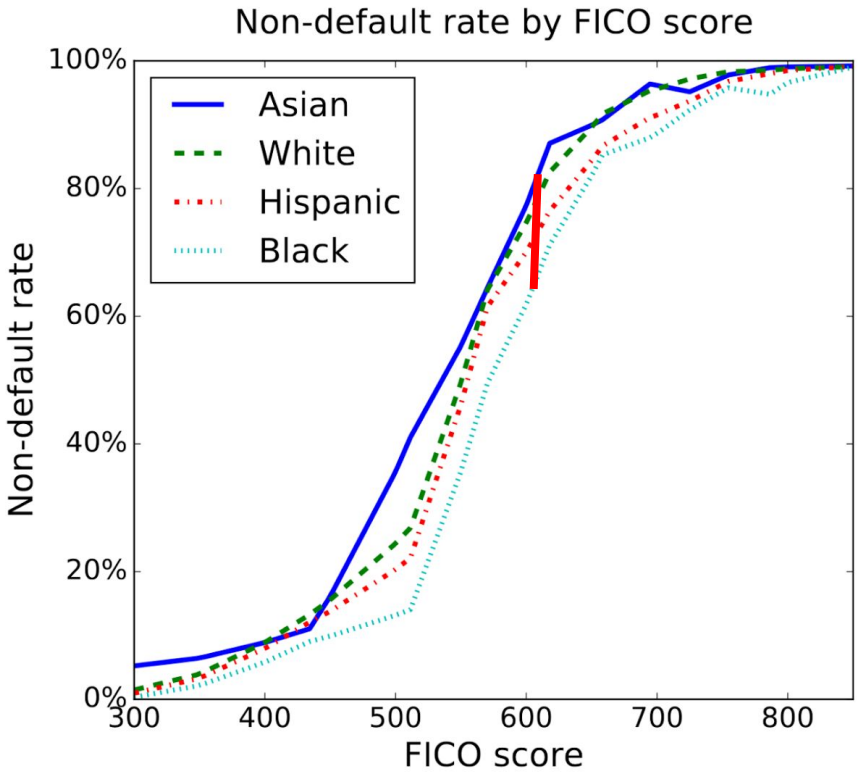
At FICO=600, approx 80% of Asians will repay loans and about 60% of Blacks will. Assume both groups make up 50% of population.

Charge fixed interest rate of 43% to both groups.

**Individually fair.**

**Non-utilitarian** - for every $200 lent out, Asians predictably pay $114.3 while Blacks pay only $86.7.

Wealth transfer from Asians to Blacks.



# Exposing tradeoffs

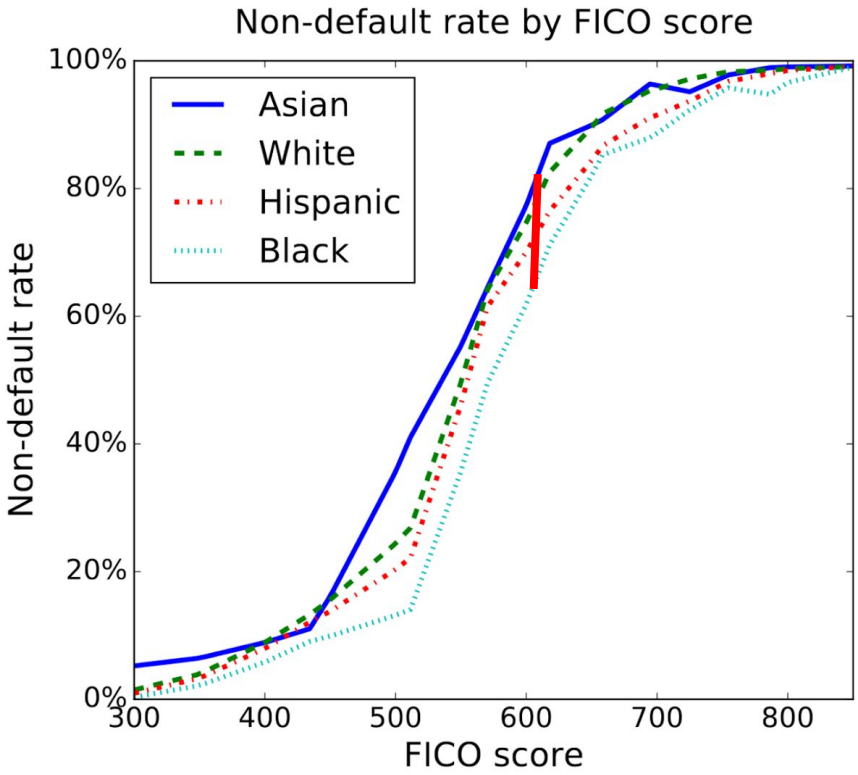
At FICO=600, approx 80% of Asians will repay loans and about 60% of Blacks will. Assume both groups make up 50% of population.

Can charge Asians 25% interest and blacks 66%.

**Individually unfair.**

**Utilitarian** - loans are more accurately allocated to those will repay them, and more loans can be issued since cost of lending is lower.

**Problematic** - we noticed an undesirable fact about the world.



There is no choice of cutoff and interest rate which satisfies all ethical principles.

# Uncomfortable Questions

It's **mathematically impossible** for the deepest neural network built by the most diverse team of data scientists to satisfy all definitions of fairness.

People at Google/Microsoft writing papers on this topic have made one choice, which I'm calling San Francisco Ethics. Is their choice right for India?

**What are Bangalore Ethics?**