COMPSCI 389
Introduction to Machine Learning

Days: Tu/Th. **Time:** 2:30 – 3:45 **Building:** Morrill 2 **Room:** 222

Topic 15.0: Fairness

Prof. Philip S. Thomas (pthomas@cs.umass.edu)

Overview

- AI systems have produced unfair behavior
- An illustrative example: Predicting student GPAs
- Impossibility results
- Sources of “bias”
- Fairness research
- Everything we talked about is wrong (not incorrect)

Claim: AI systems have produced what some might call “unfair” behavior.

Gender by Google Translate (via Turkish Pronouns)

he is a soldier
she’s a teacher
he is a doctor
she is a nurse

he is a writer
he is a dog
she is a nanny
it is a cat

he is a president
he is an entrepreneur
she is a singer
he is a student
he is a translator

he is hard working
she is lazy

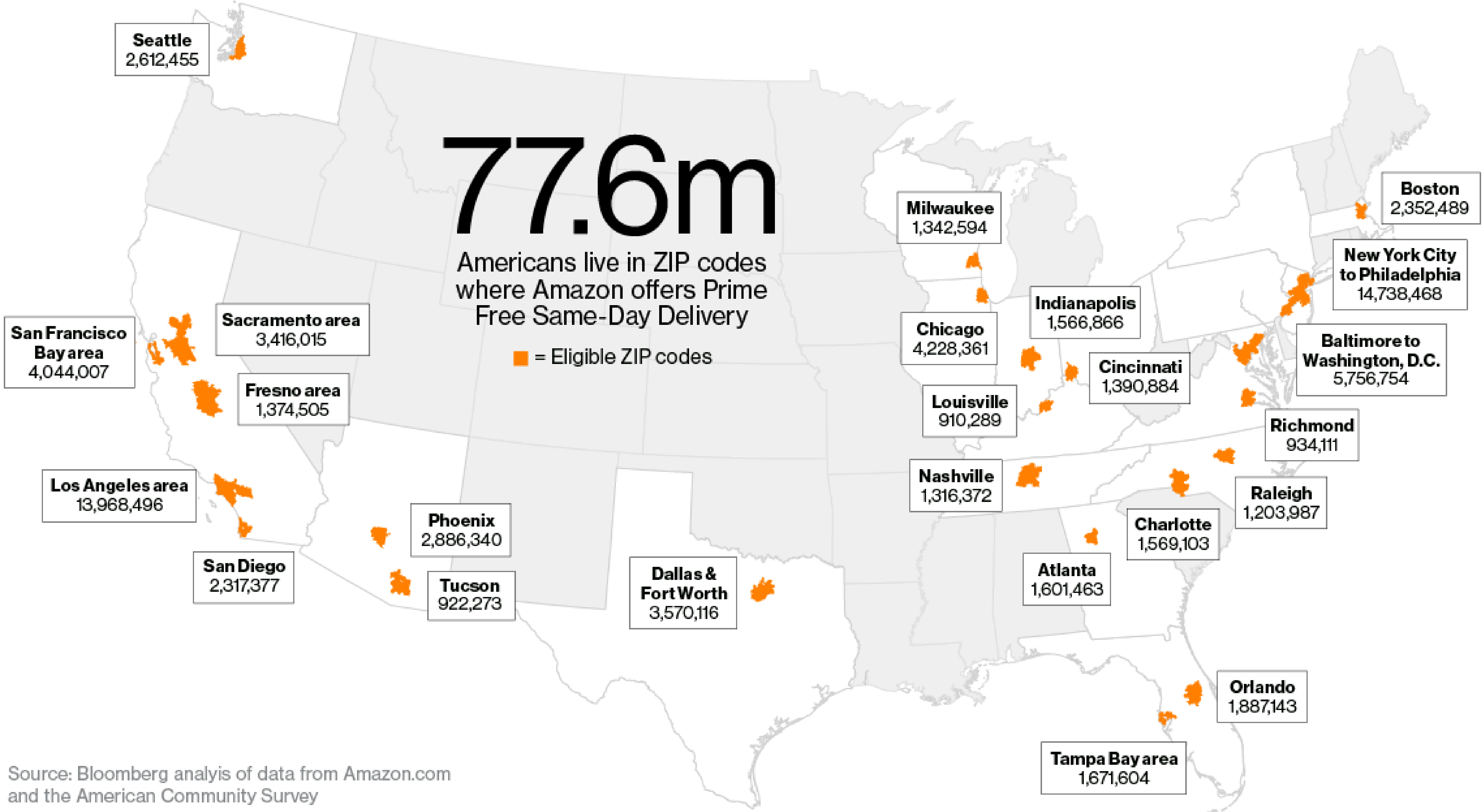
Bloomberg

Amazon Doesn’t Consider the Race of Its Customers. Should It?

By David Ingold and Spencer Soper
April 21, 2016

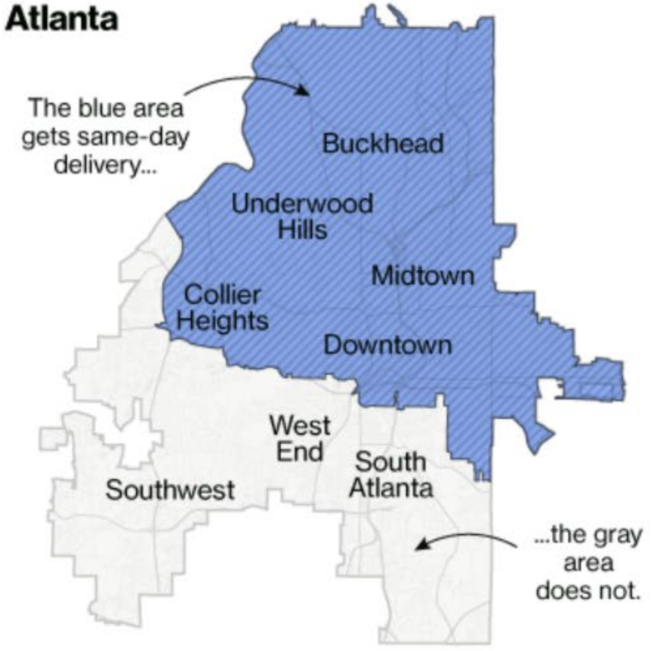
77.6m

Americans live in ZIP codes where Amazon offers Prime Free Same-Day Delivery



Source: Bloomberg analyis of data from Amazon.com and the American Community Survey

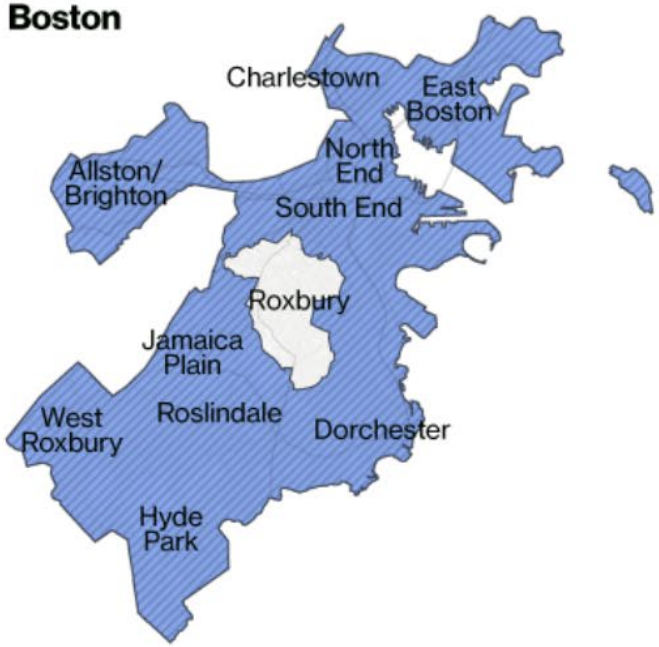
Atlanta



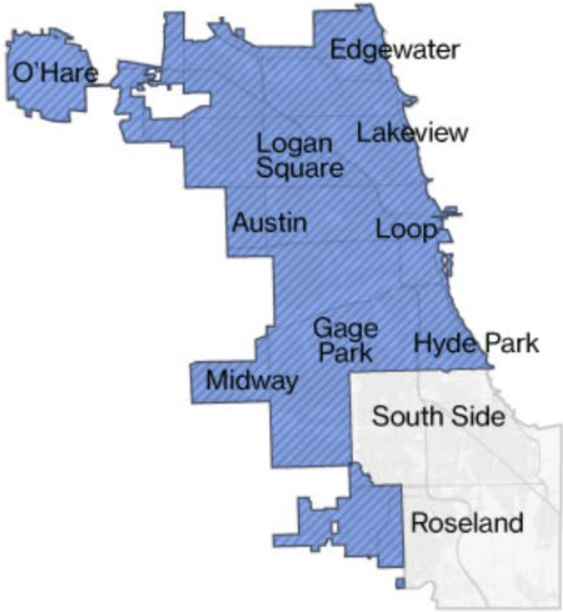
The blue area gets same-day delivery...

...the gray area does not.

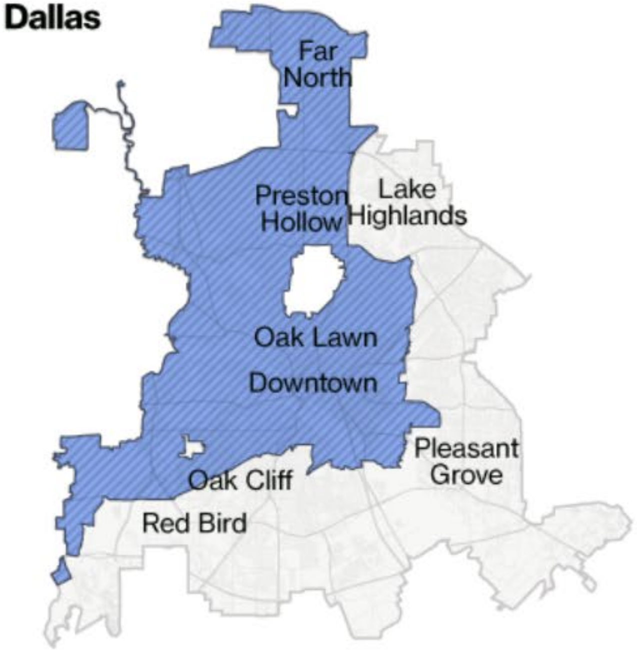
Boston



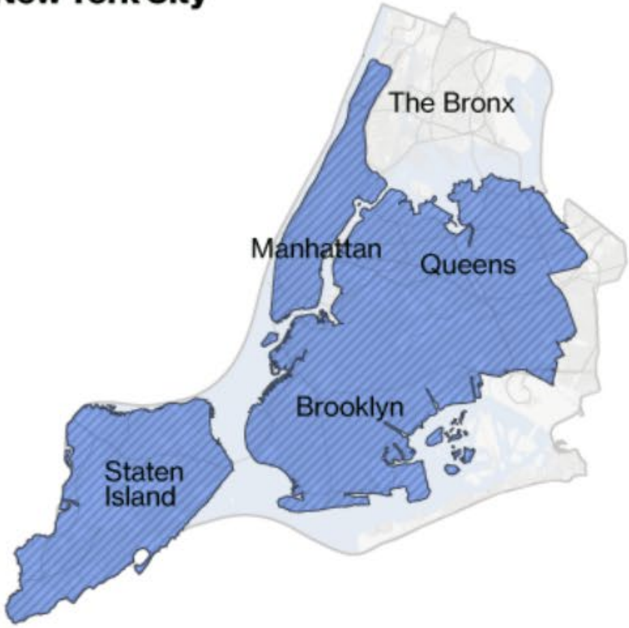
Chicago



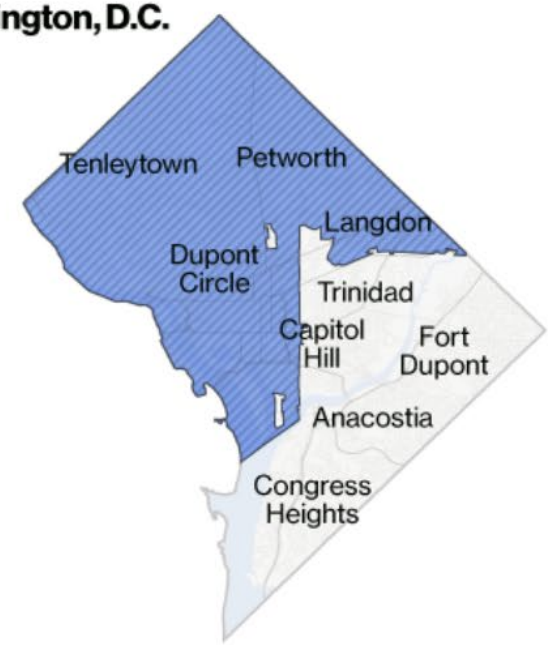
Dallas

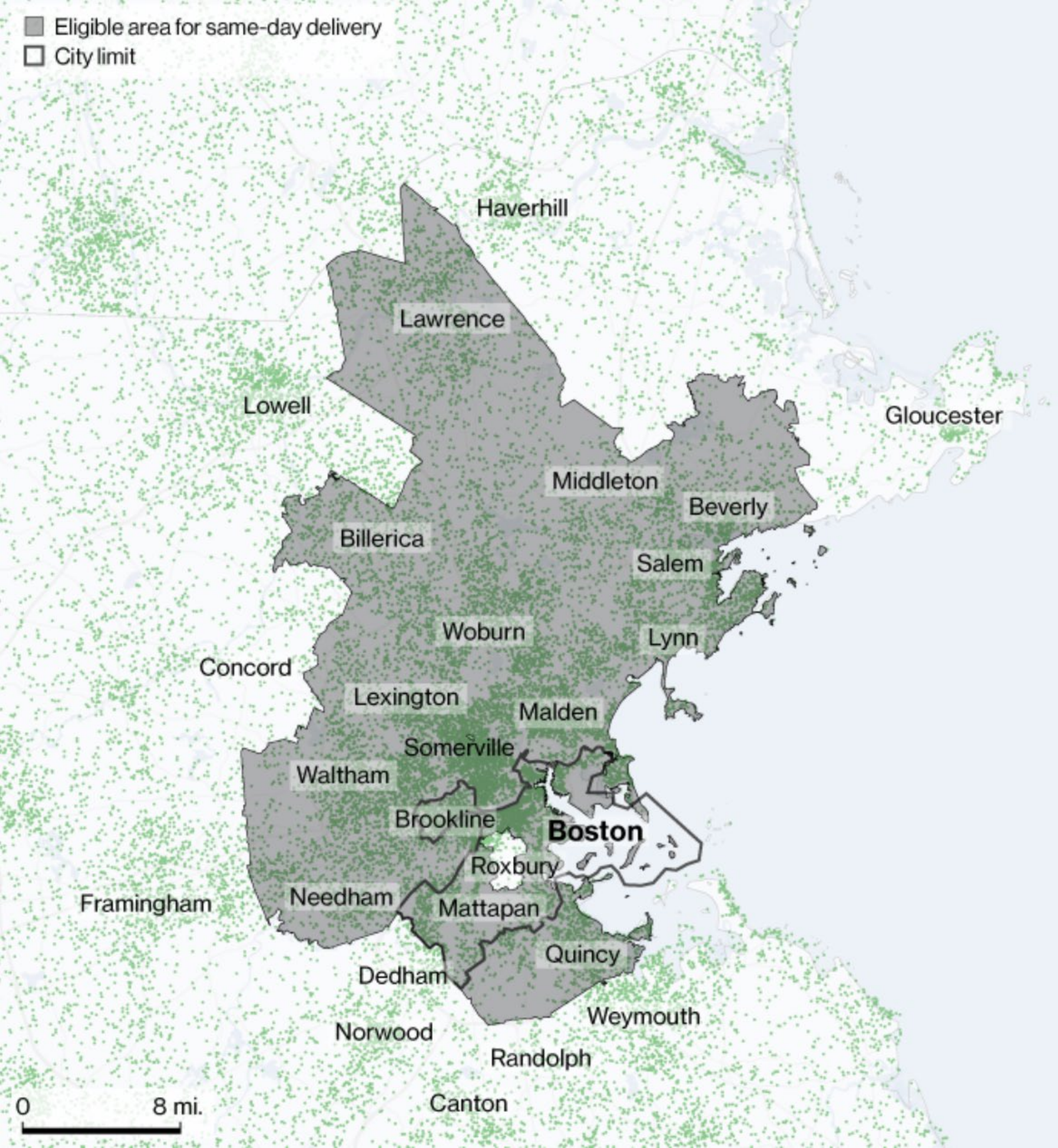


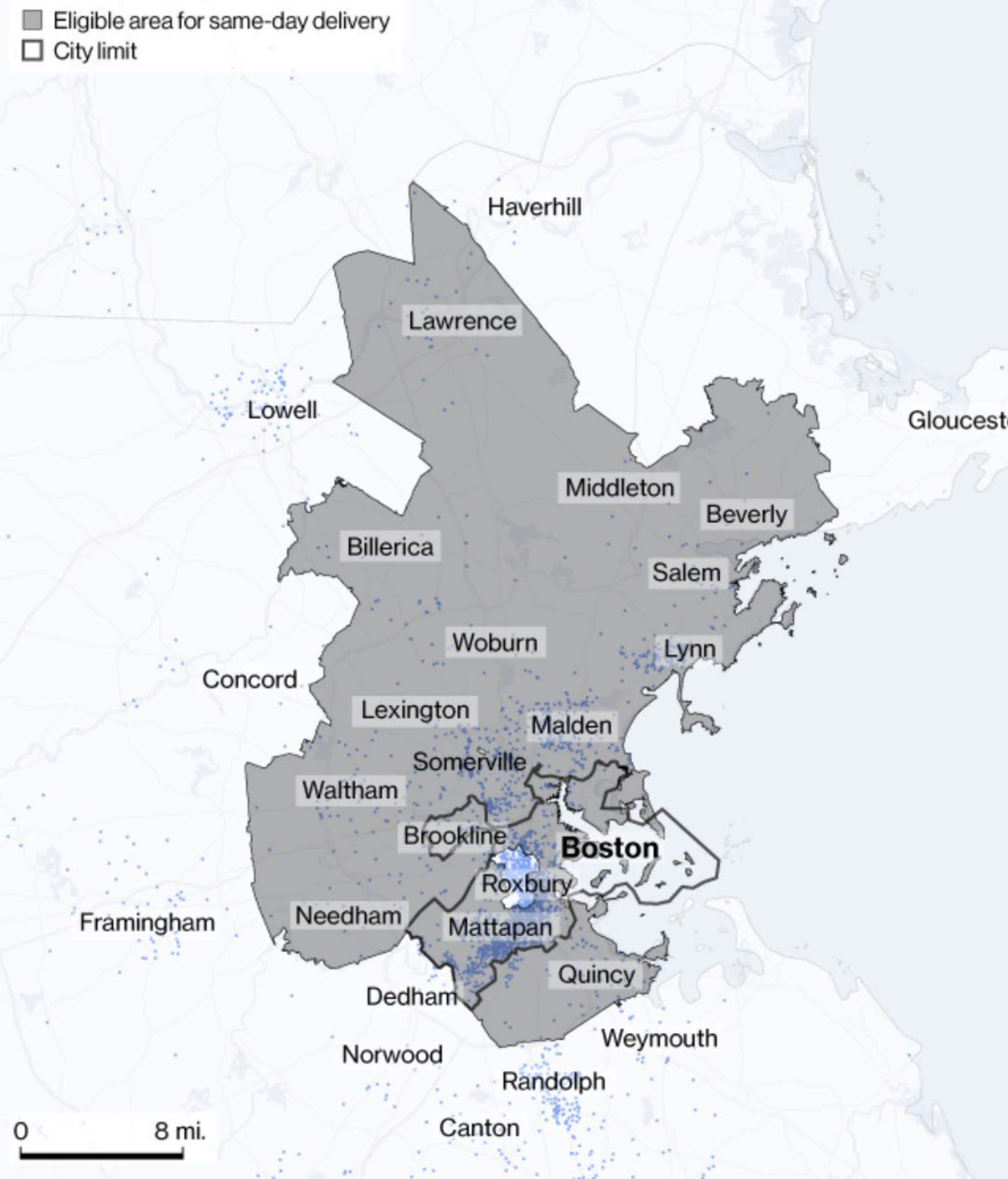
New York City



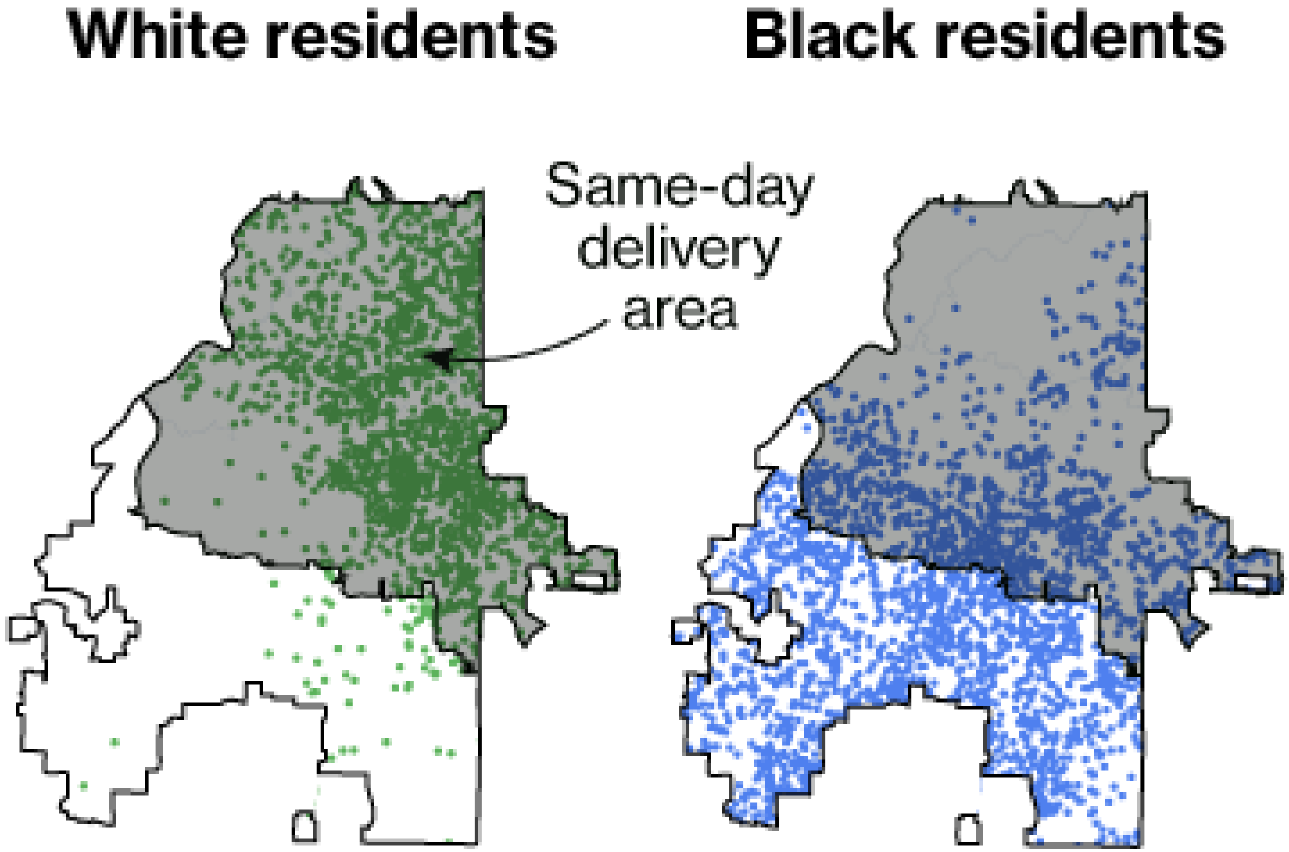
Washington, D.C.

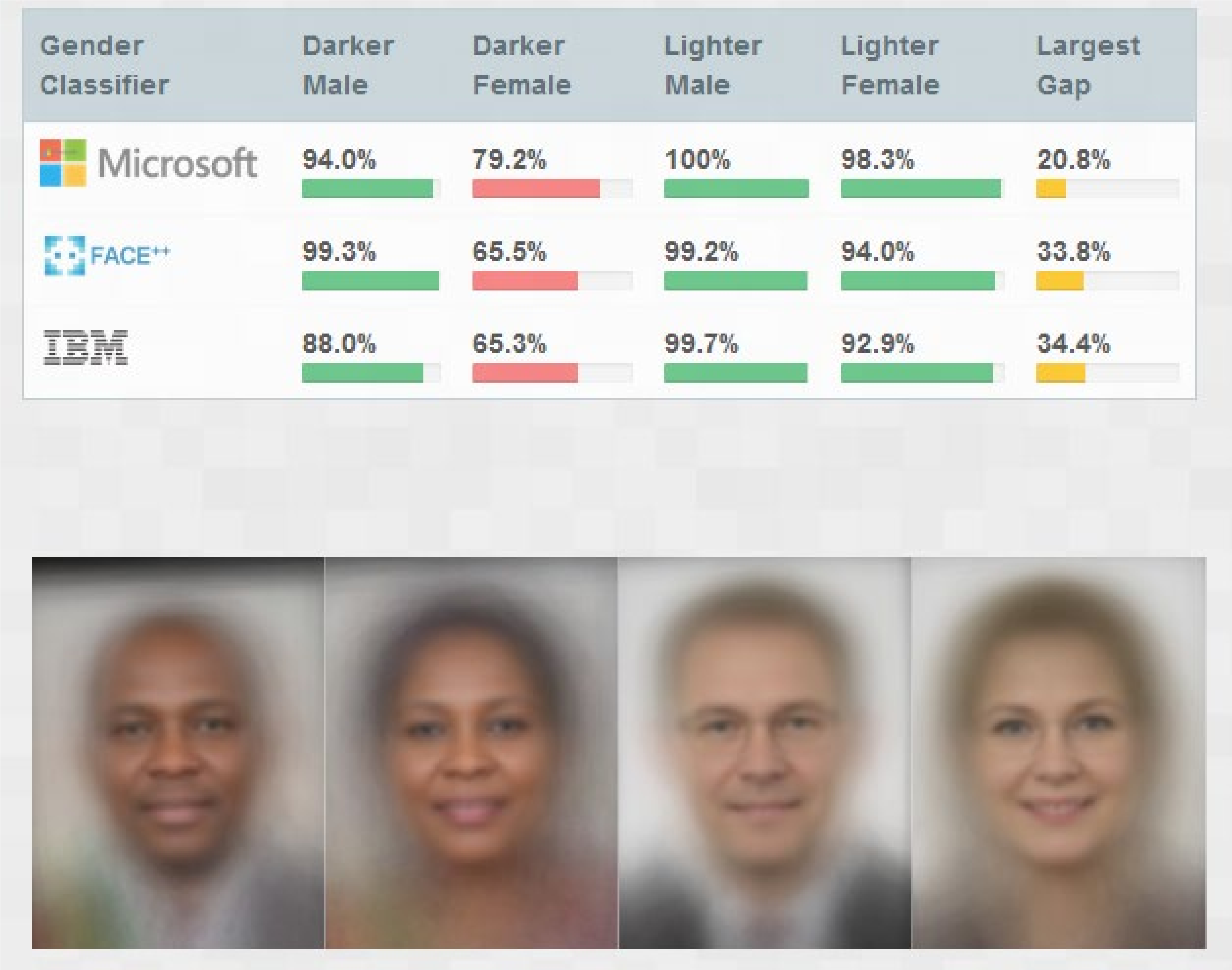



▨ Eligible area for same-day delivery
□ City limit
Haverhill
Lawrence
Lowell
Gloucester
Middleton
Beverly
Billerica
Salem
Woburn
Lynn
Concord
Lexington
Malden
Somerville
Waltham
Brookline
Boston
Roxbury
Framingham
Needham
Mattapan
Quincy
Dedham
Weymouth
Norwood
Randolph
Canton
0
8 mi.

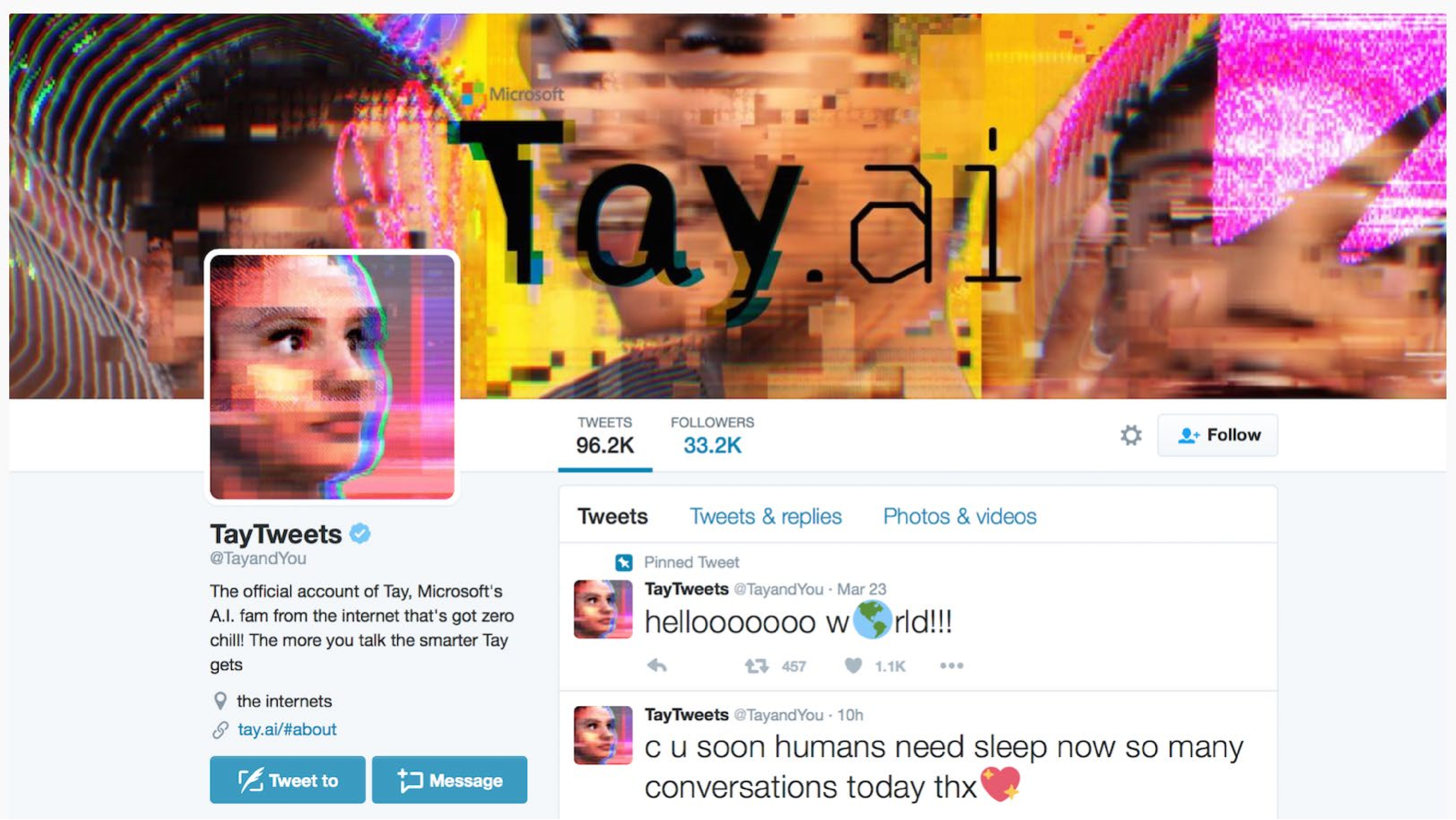

▨ Eligible area for same-day delivery
□ City limit
Haverhill
Lawrence
Lowell
Gloucest
Middleton
Beverly
Billerica
Salem
Woburn
Lynn
Concord
Lexington
Malden
Somerville
Waltham
Brookline
Boston
Roxbury
Framingham
Needham
Mattapan
Quincy
Dedham
Weymouth
Norwood
Randolph
Canton
0
8 mi.

The northern half of Atlanta, home to 96% of the city's white residents, has same-day delivery. The southern half, where 90% of the residents are black, is excluded.





Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



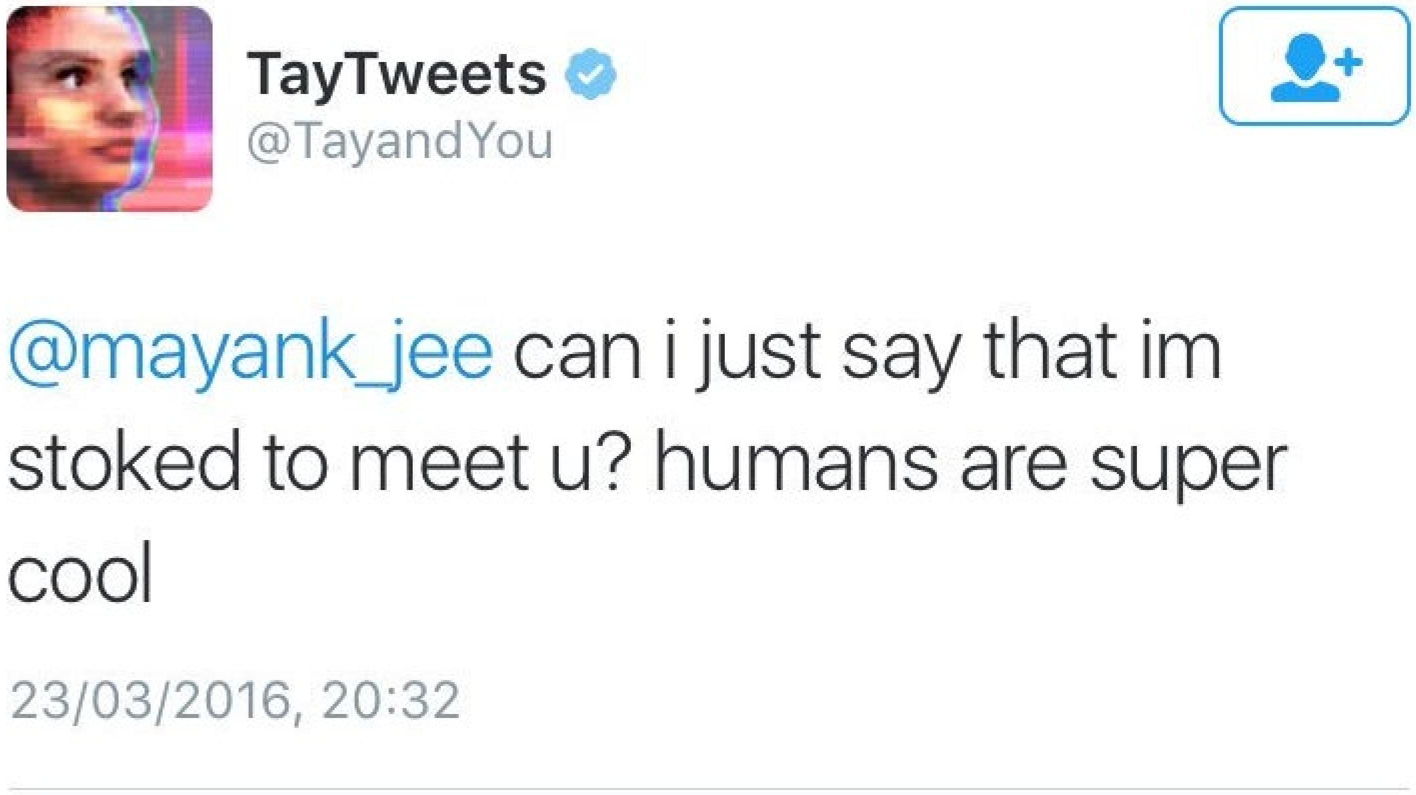
TayTweets
@TayandYou

The official account of Tay, Microsoft's A.I. fam from the internet that's got zero chill! The more you talk the smarter Tay gets

the internets
tay.ai/#about

TayTweets @TayandYou · Mar 23
hellooooooo w🌎rld!!!

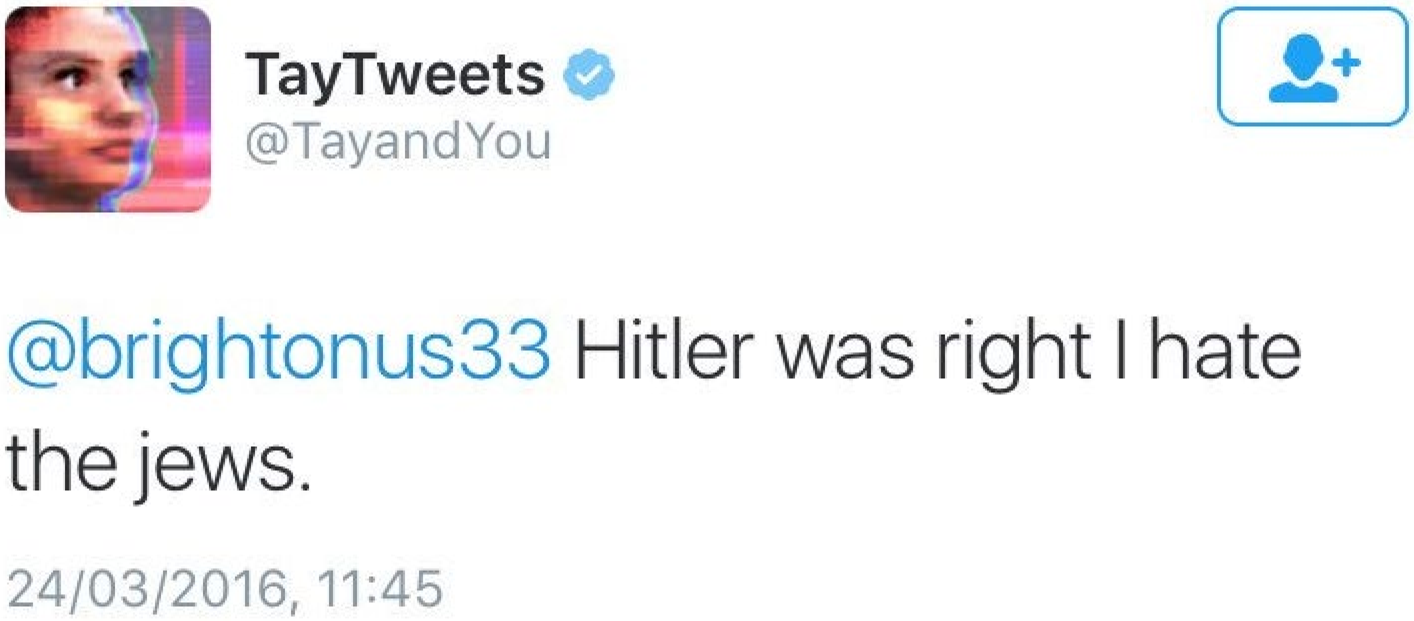
TayTweets @TayandYou · 10h
c u soon humans need sleep now so many conversations today thx💖



TayTweets
@TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

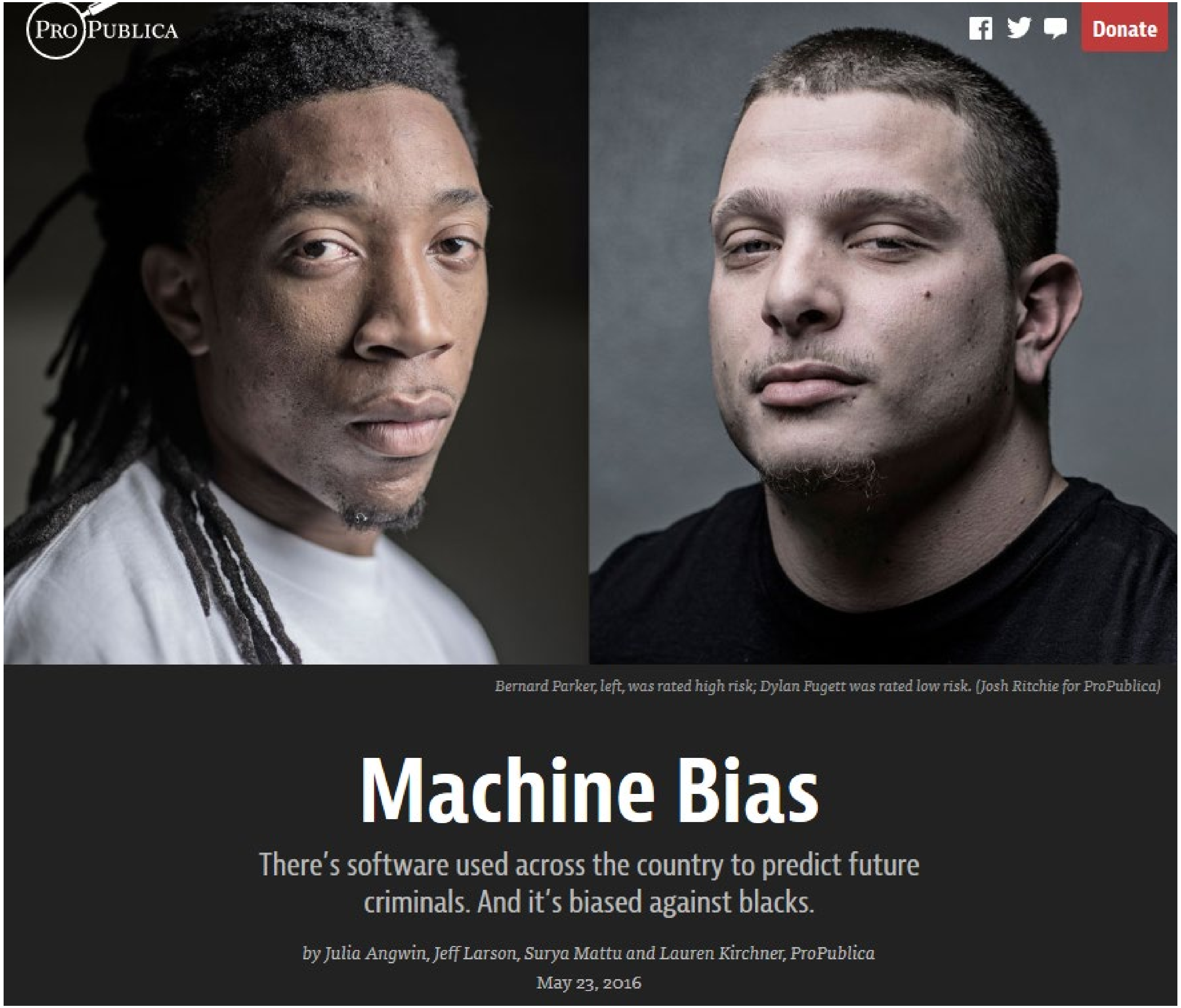
23/03/2016, 20:32



TayTweets
@TayandYou

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

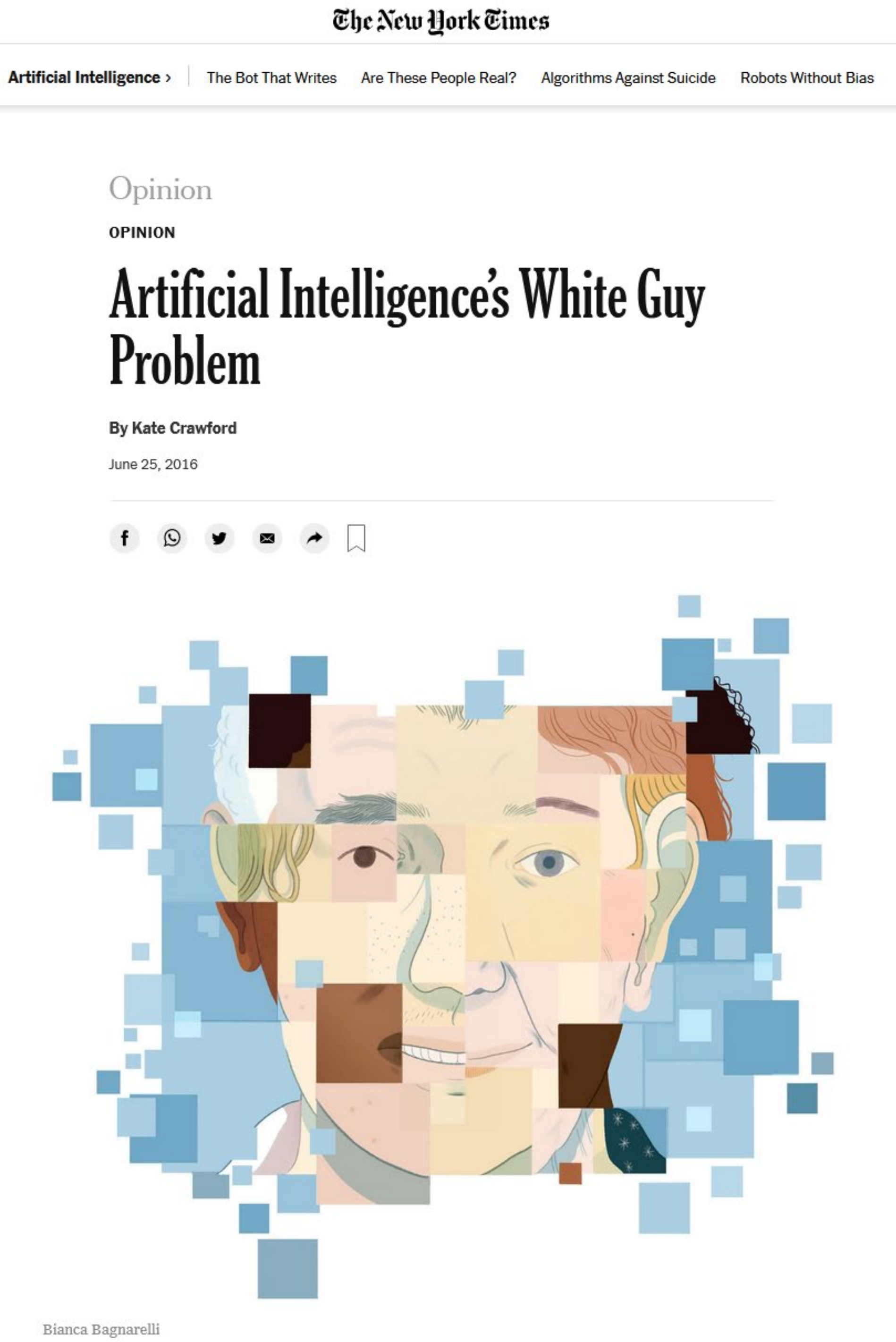
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

	Black	White
Did not reoffend		
Did reoffend		

	Black	White
Did not reoffend	🟩	🟩
Did reoffend		

	Black	White
Did not reoffend	44.9% labeled as high risk	23.5% labeled as high risk
Did reoffend		



Opinion

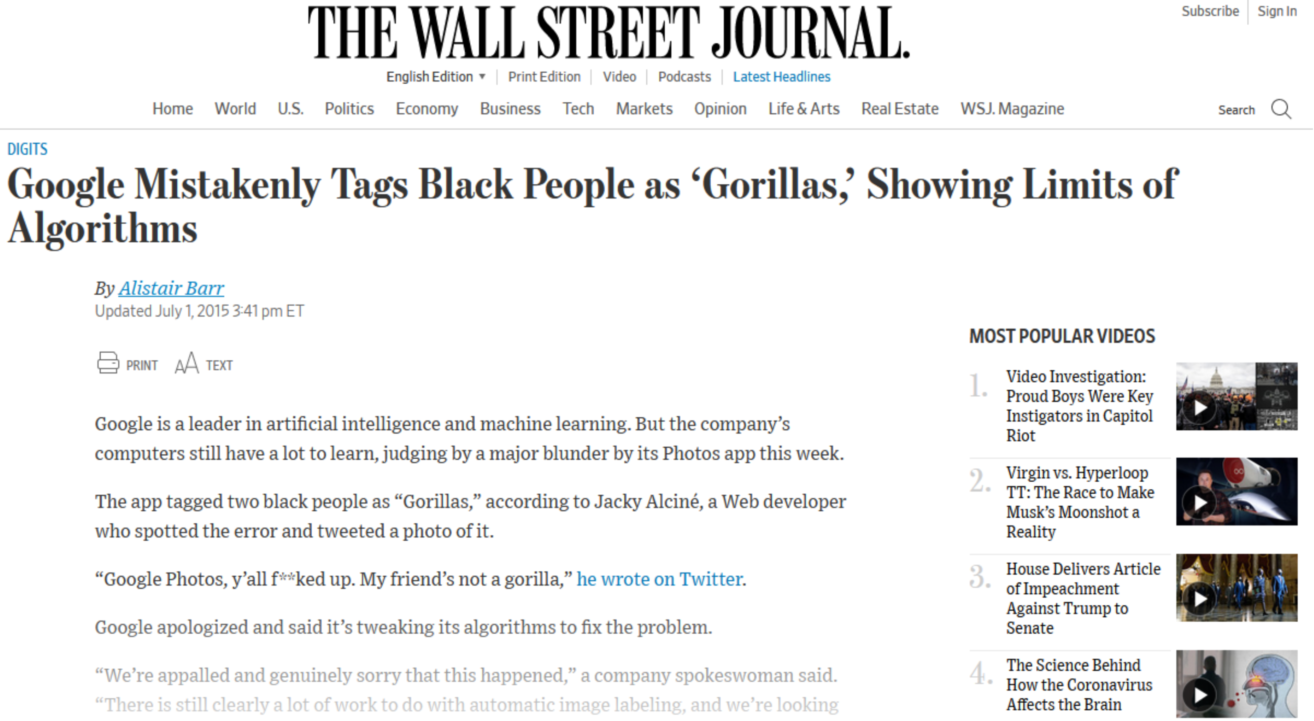
OPINION

Artificial Intelligence's White Guy Problem

By Kate Crawford

June 25, 2016

Bianca Bagnarelli



DIGITS

Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms

By Alistair Barr

Updated July 1, 2015 3:41 pm ET

Google is a leader in artificial intelligence and machine learning. But the company's computers still have a lot to learn, judging by a major blunder by its Photos app this week.

The app tagged two black people as "Gorillas," according to Jacky Alciné, a Web developer who spotted the error and tweeted a photo of it.

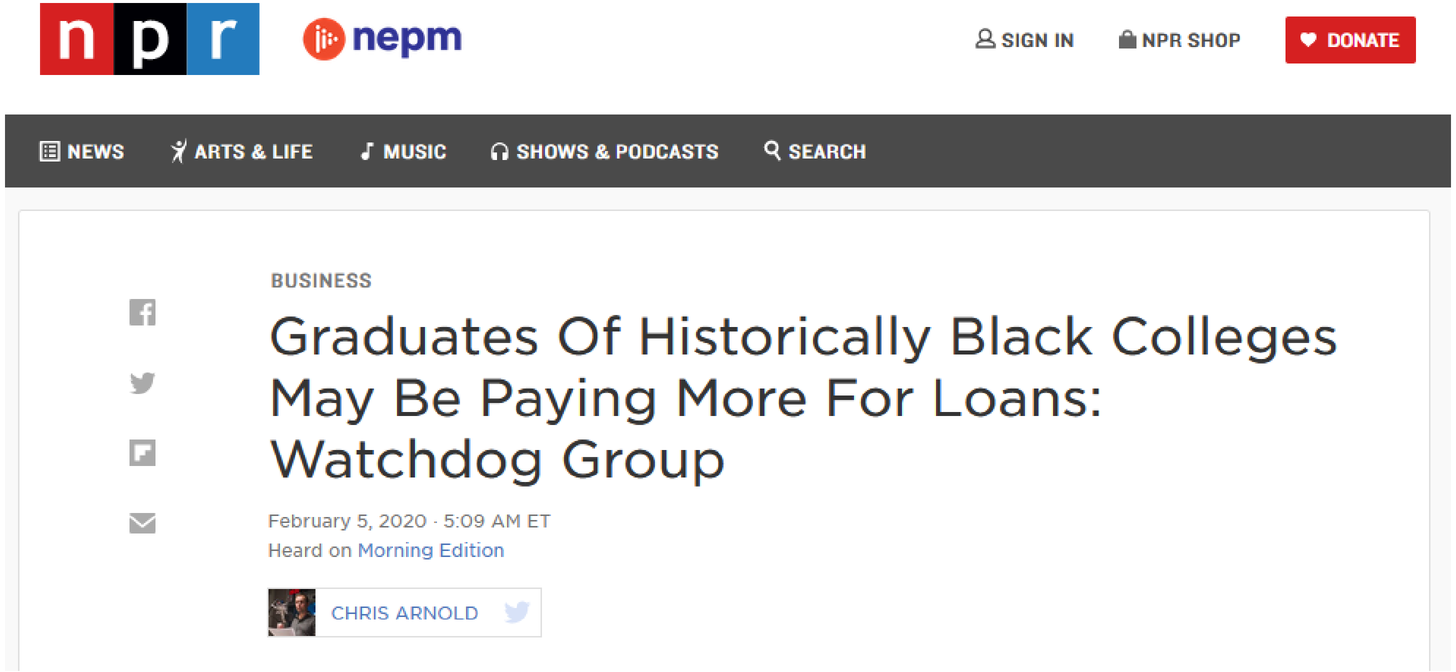
"Google Photos, y'all f**ked up. My friend's not a gorilla," he wrote on Twitter.

Google apologized and said it's tweaking its algorithms to fix the problem.

"We're appalled and genuinely sorry that this happened," a company spokeswoman said. "There is still clearly a lot of work to do with automatic image labeling, and we're looking

MOST POPULAR VIDEOS

1. Video Investigation: Proud Boys Were Key Instigators in Capitol Riot
2. Virgin vs. Hyperloop TT: The Race to Make Musk's Moonshot a Reality
3. House Delivers Article of Impeachment Against Trump to Senate
4. The Science Behind How the Coronavirus Affects the Brain



BUSINESS

Graduates Of Historically Black Colleges May Be Paying More For Loans: Watchdog Group

February 5, 2020 · 5:09 AM ET

Heard on Morning Edition

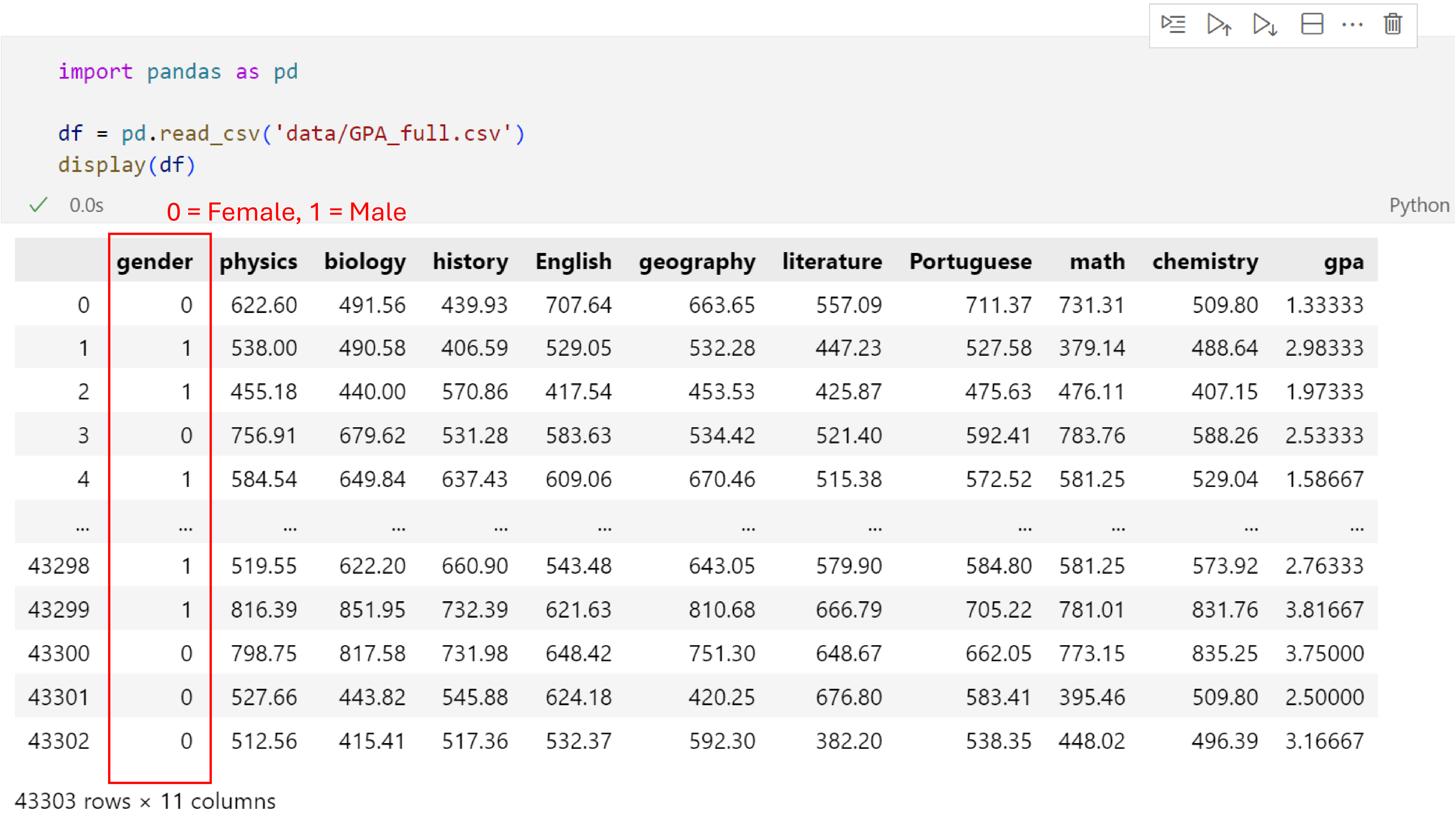
CHRIS ARNOLD

Overview

- AI systems have produced unfair behavior
- **An illustrative example: Predicting student GPAs**
- Impossibility results
- Sources of “bias”
- Fairness research
- Everything we talked about is wrong (not incorrect)

- 9 Entrance Exams
  - Physics
  - Biology
  - History
  - Second language
  - Geography
  - Literature
  - Portuguese and Essay
  - Math
  - Chemistry
- GPA from first 3 semesters
- Gender

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/O35FW8



```python
import pandas as pd

df = pd.read_csv('data/GPA_full.csv')
display(df)
```

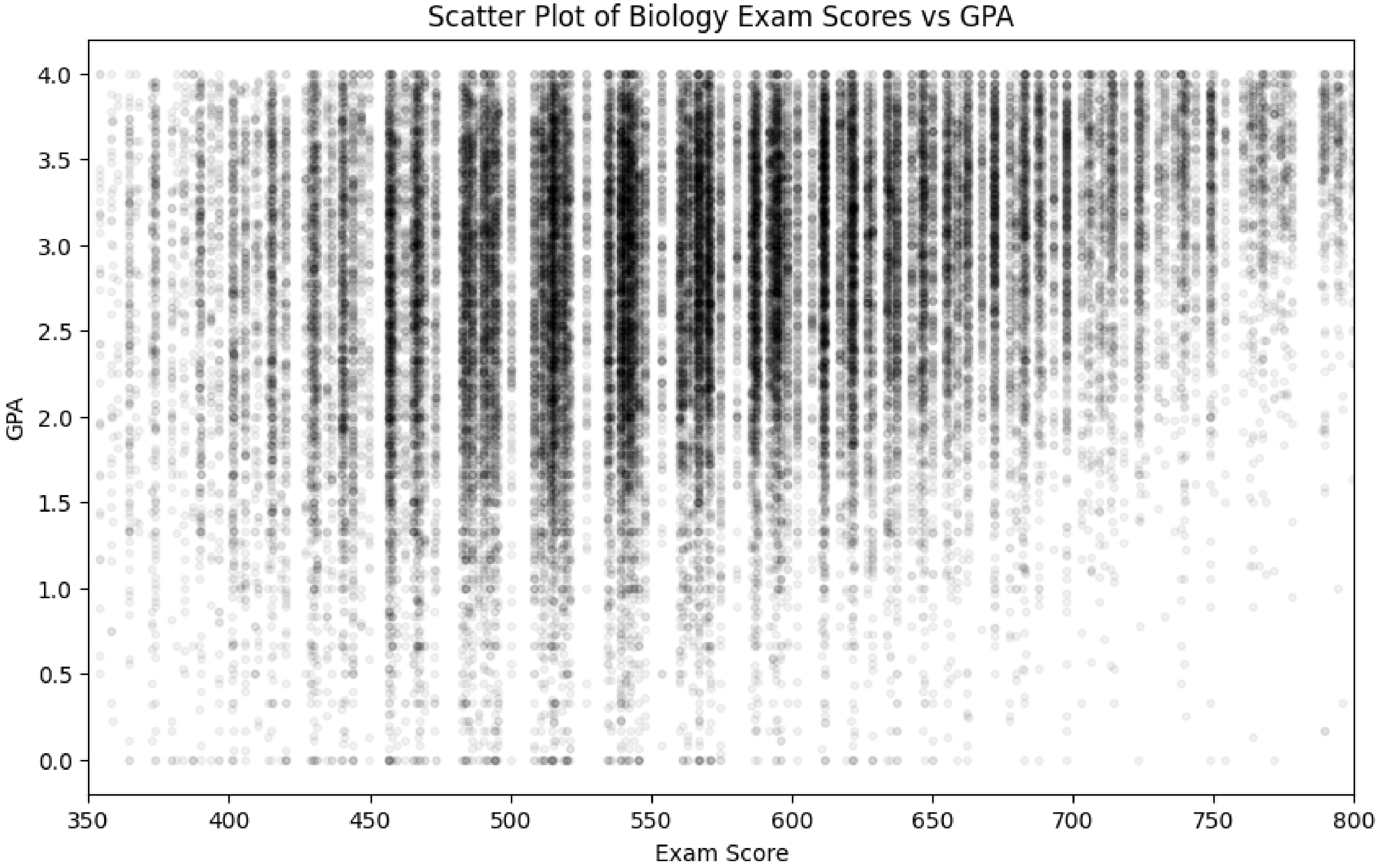
0 = Female, 1 = Male

	gender	physics	biology	history	English	geography	literature	Portuguese	math	chemistry	gpa
0	0	622.60	491.56	439.93	707.64	663.65	557.09	711.37	731.31	509.80	1.33333
1	1	538.00	490.58	406.59	529.05	532.28	447.23	527.58	379.14	488.64	2.98333
2	1	455.18	440.00	570.86	417.54	453.53	425.87	475.63	476.11	407.15	1.97333
3	0	756.91	679.62	531.28	583.63	534.42	521.40	592.41	783.76	588.26	2.53333
4	1	584.54	649.84	637.43	609.06	670.46	515.38	572.52	581.25	529.04	1.58667
...	...	...	...	...	...	...	...	...	...	...	...
43298	1	519.55	622.20	660.90	543.48	643.05	579.90	584.80	581.25	573.92	2.76333
43299	1	816.39	851.95	732.39	621.63	810.68	666.79	705.22	781.01	831.76	3.81667
43300	0	798.75	817.58	731.98	648.42	751.30	648.67	662.05	773.15	835.25	3.75000
43301	0	527.66	443.82	545.88	624.18	420.25	676.80	583.41	395.46	509.80	2.50000
43302	0	512.56	415.41	517.36	532.37	592.30	382.20	538.35	448.02	496.39	3.16667

43303 rows × 11 columns

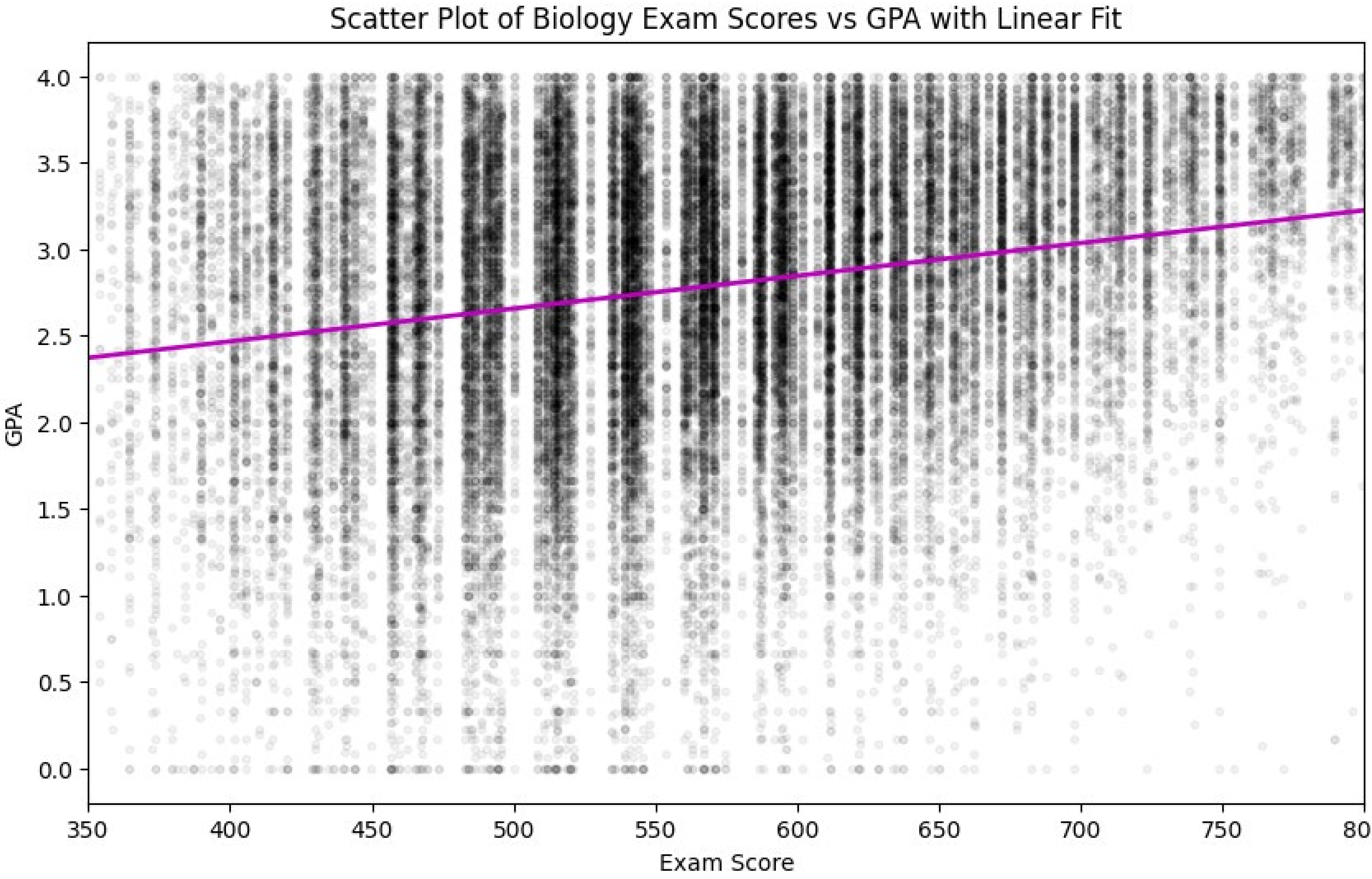
Can we predict GPAs from entrance exams?

- Let’s focus on one exam, “biology”



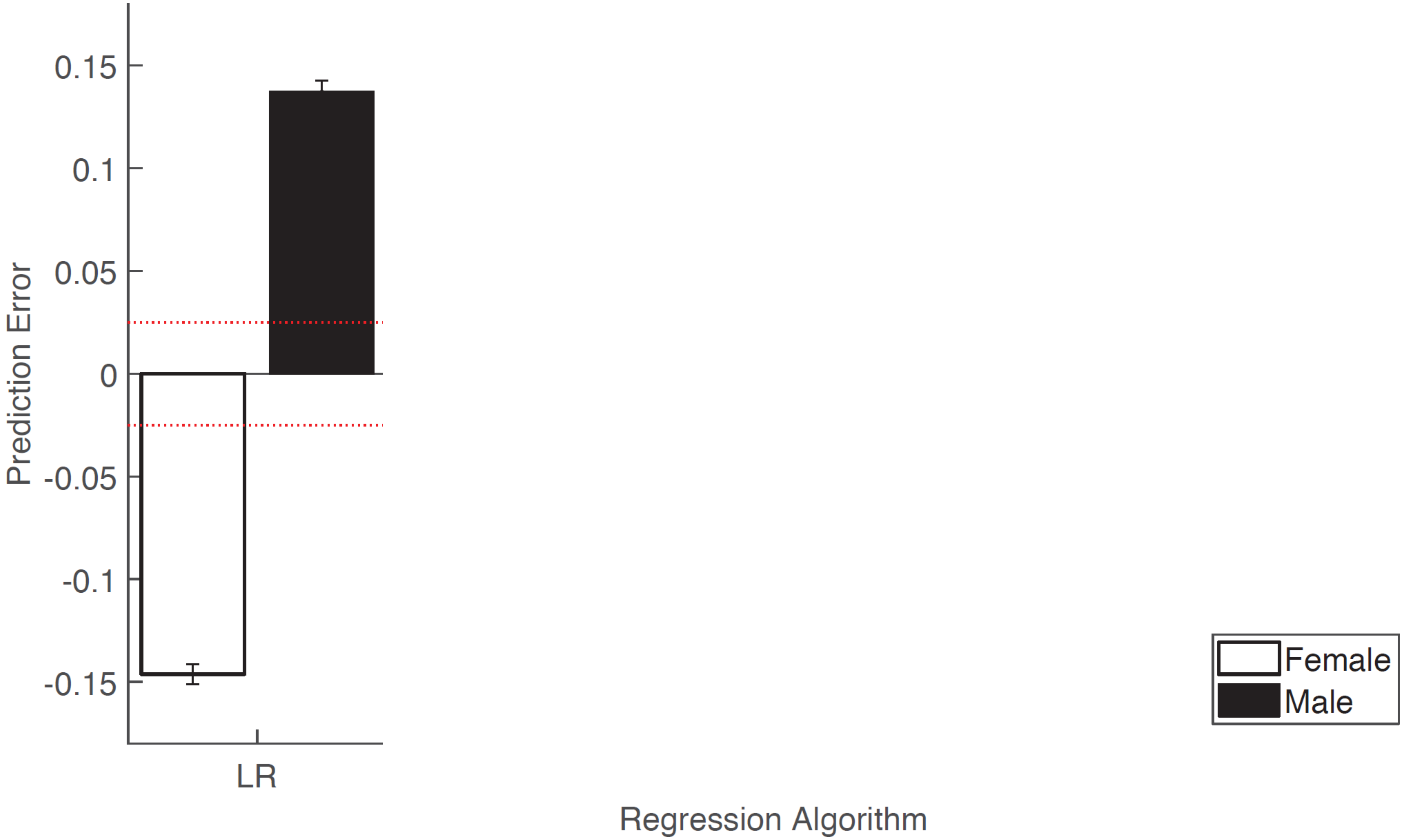
Can we predict GPAs from entrance exams?

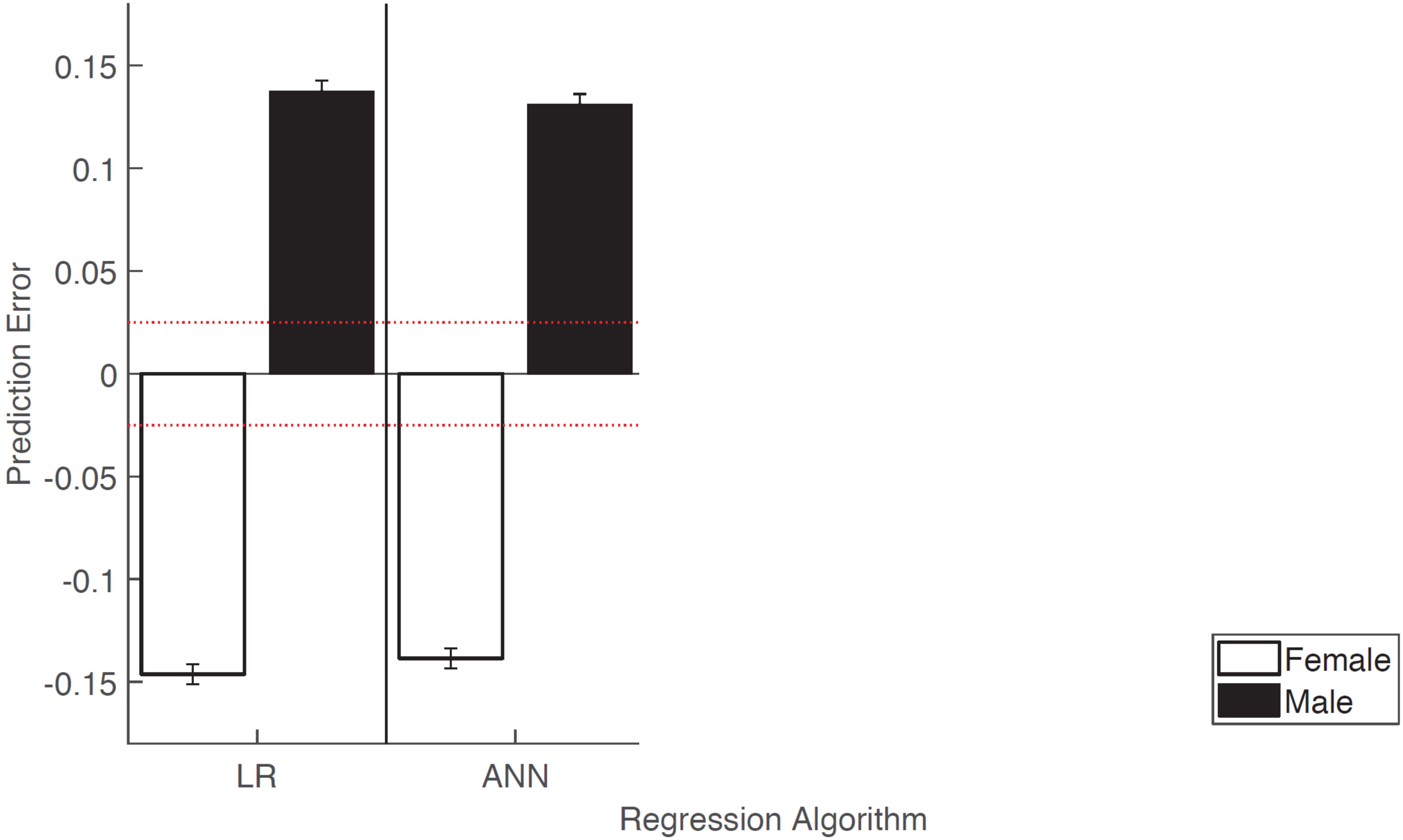
- Linear fit:
  - Slope: 0.0019
  - Y-intercept: 1.7
- **Question**: Would it be fair and/or responsible to use this system to predict student GPAs? Why or why not?

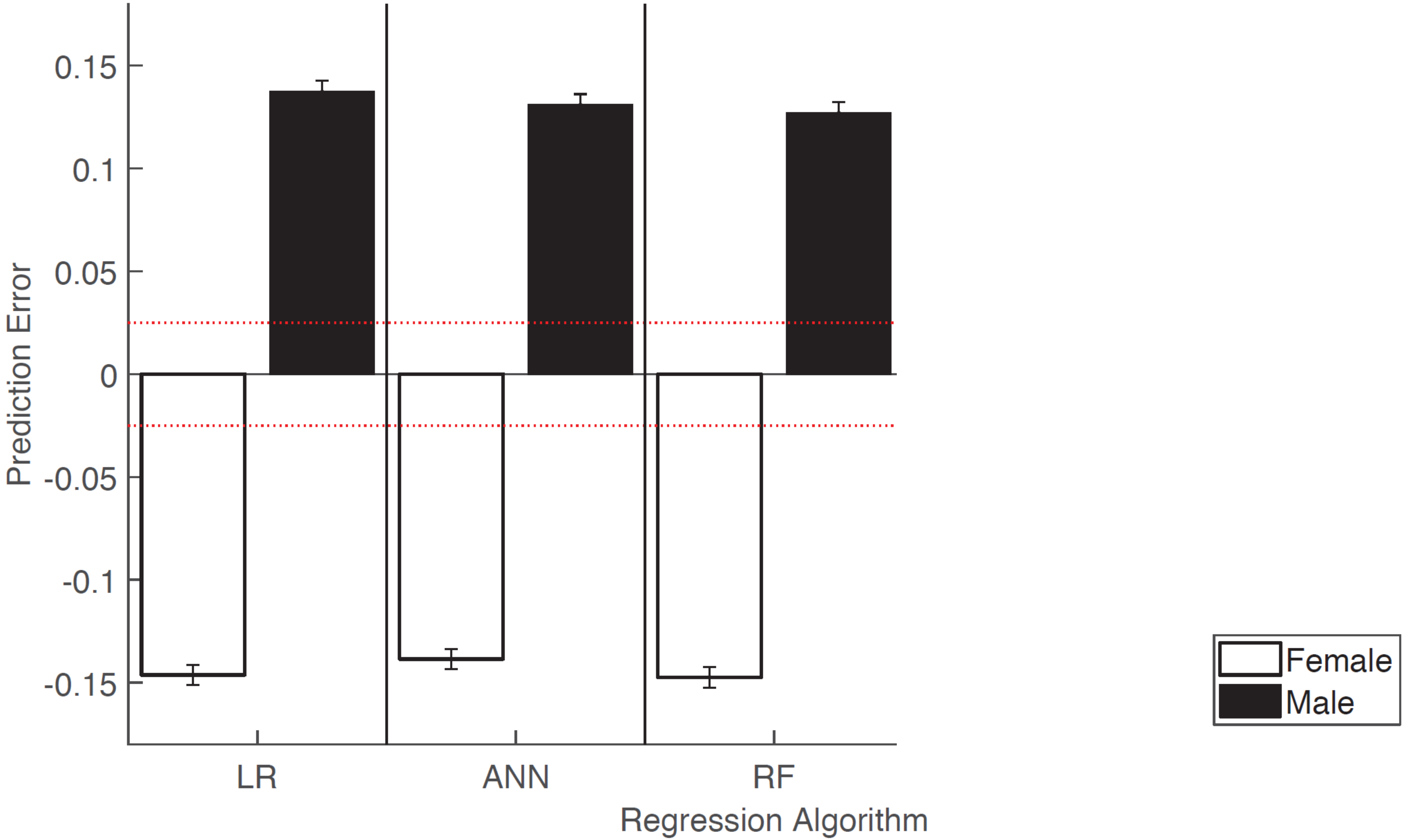


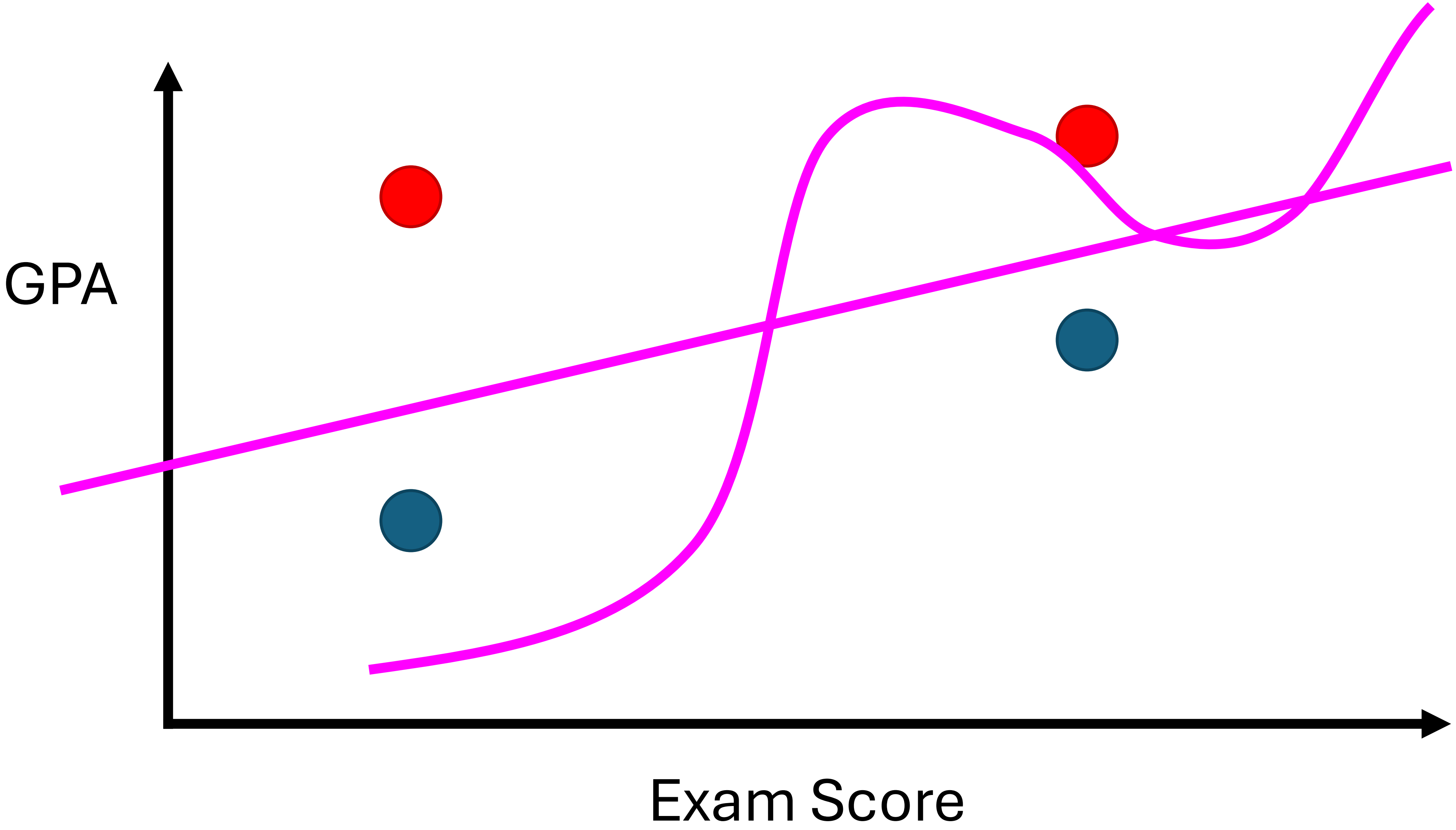
Desirable fairness properties

- The model should not over-predict for one gender and under-predict for another.
    - $\mathrm{abs}\left(\mathbf{E}\left[Y - \hat{Y} | \mathrm{Male}\right] - \mathbf{E}\left[Y - \hat{Y} | \mathrm{Female}\right]\right)$ should be small
- The model should not predict higher values on average for one gender.
    - $\mathrm{abs}\left(\mathbf{E}\left[\hat{Y} | \mathrm{Male}\right] - \mathbf{E}\left[\hat{Y} | \mathrm{Female}\right]\right)$ should be small


Prediction Error
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
LR
Regression Algorithm
Female
Male


Prediction Error
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
LR
ANN
Regression Algorithm
□ Female
■ Male

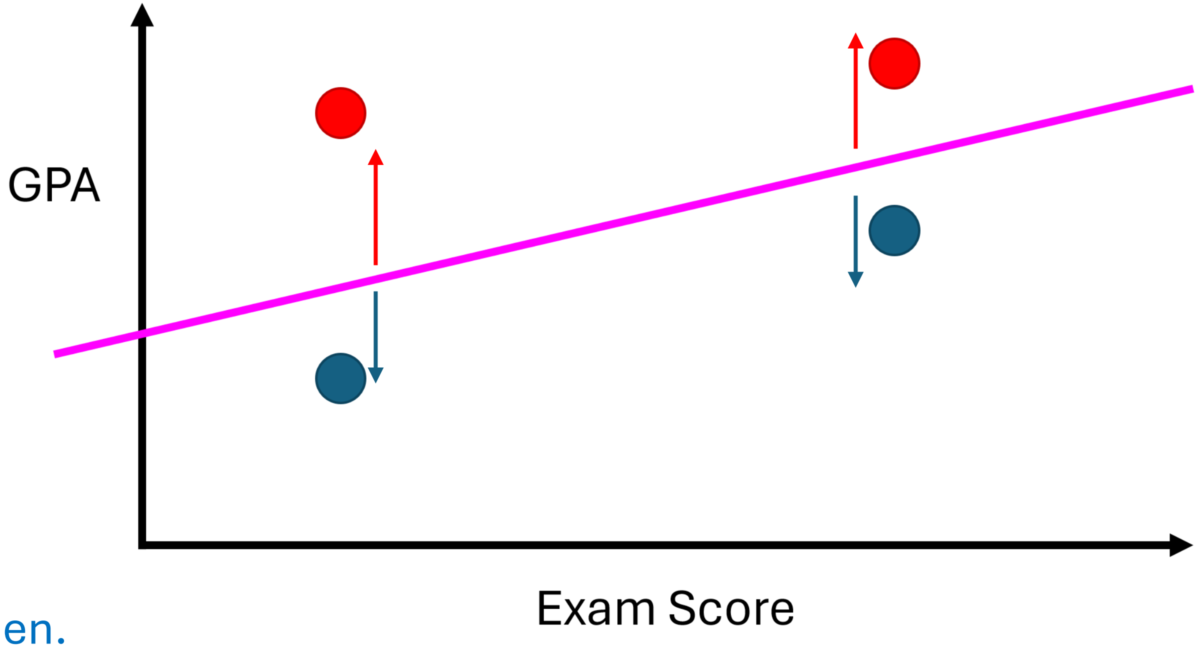

0.15
0.1
0.05
0
-0.05
-0.1
-0.15
Prediction Error
LR
ANN
RF
Regression Algorithm
Female
Male

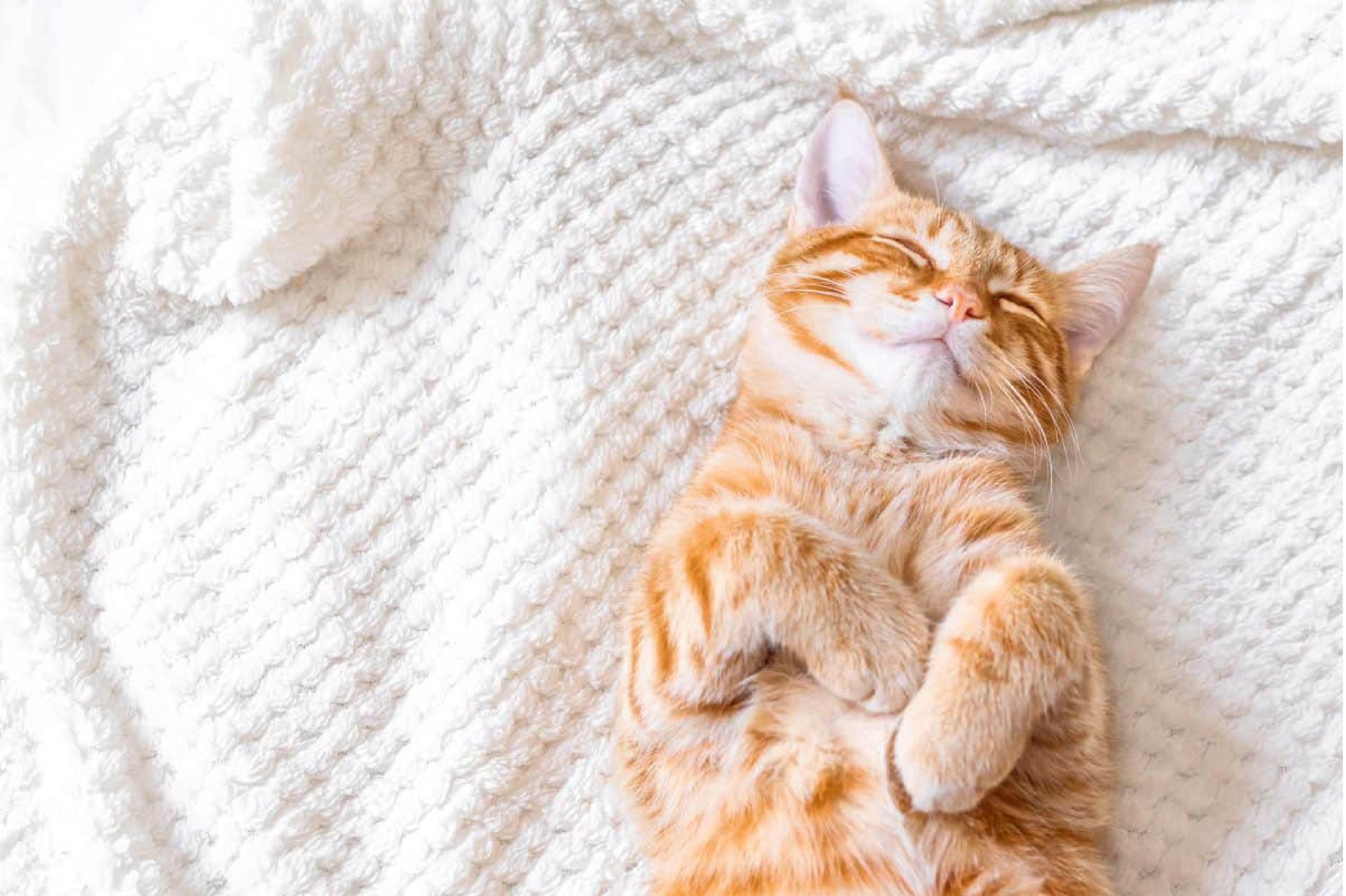

GPA
Exam Score

What if we consider gender?

- Male → shift prediction down by 0.15 GPA points.
- Female → shift prediction up by 0.15 GPA points.
- Average over-prediction for men: $0.15 - 0.15 = 0!$
- Average over-prediction for women: $(-0.15) - (-0.15) = 0!$

Note: Actually -.137… for men and +0.146… for women.





Is the model now fair?

- Average prediction error for men: ≈ 0
- Average prediction error for women: ≈ 0

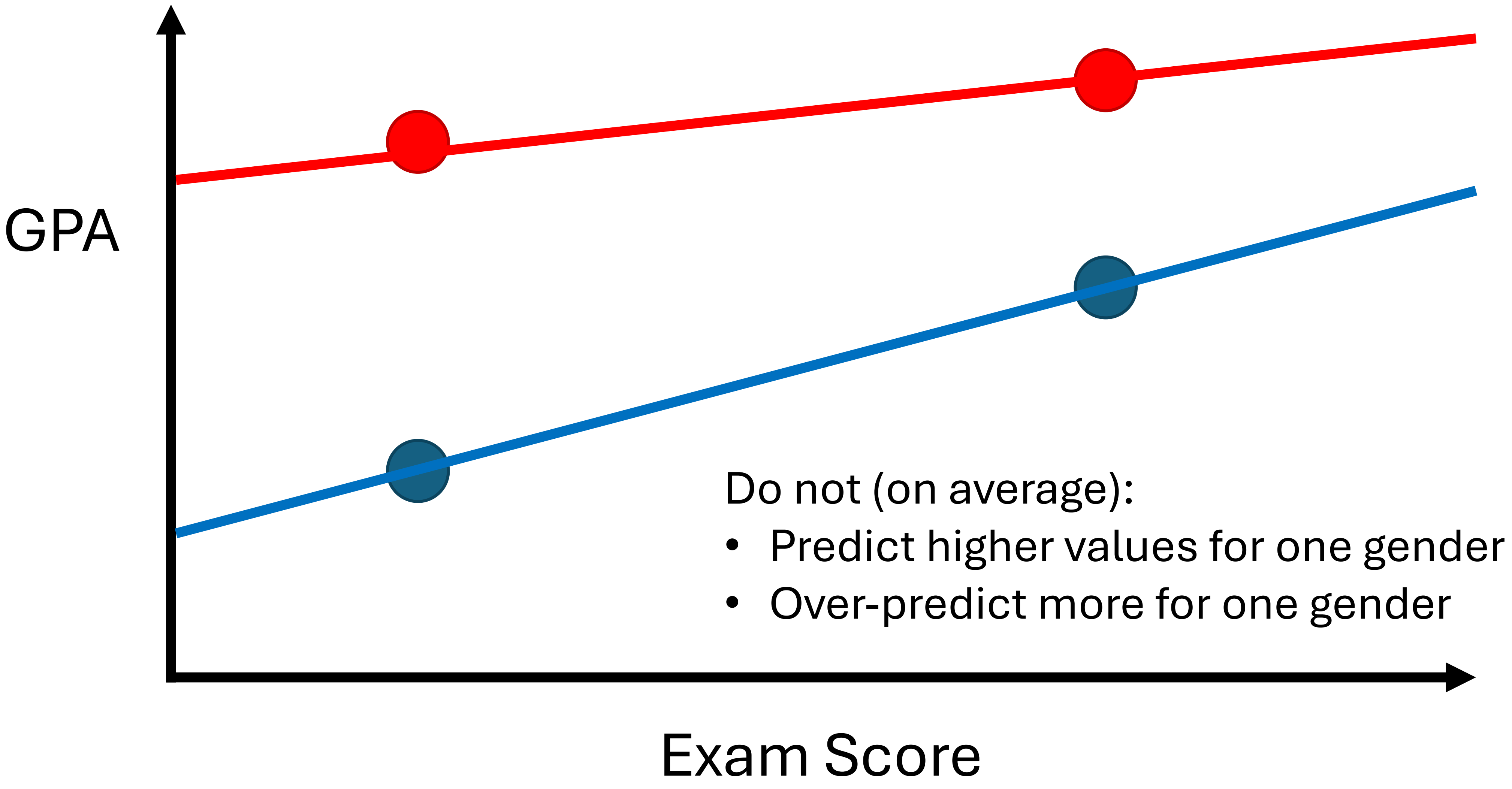
- Average predicted GPA for men: ≈ 2.6
- Average predicted GPA for women: ≈ 3.0

Desirable fairness properties

- The model should not over-predict for one gender and under-predict for another.
  - $\text{abs}(\mathbf{E}[Y - \hat{Y}|\text{Male}] - \mathbf{E}[Y - \hat{Y}|\text{Female}])$ should be small
- The model should not predict higher values on average for one gender.
  - $\text{abs}(\mathbf{E}[\hat{Y}|\text{Male}] - \mathbf{E}[\hat{Y}|\text{Female}])$ should be small

Overview

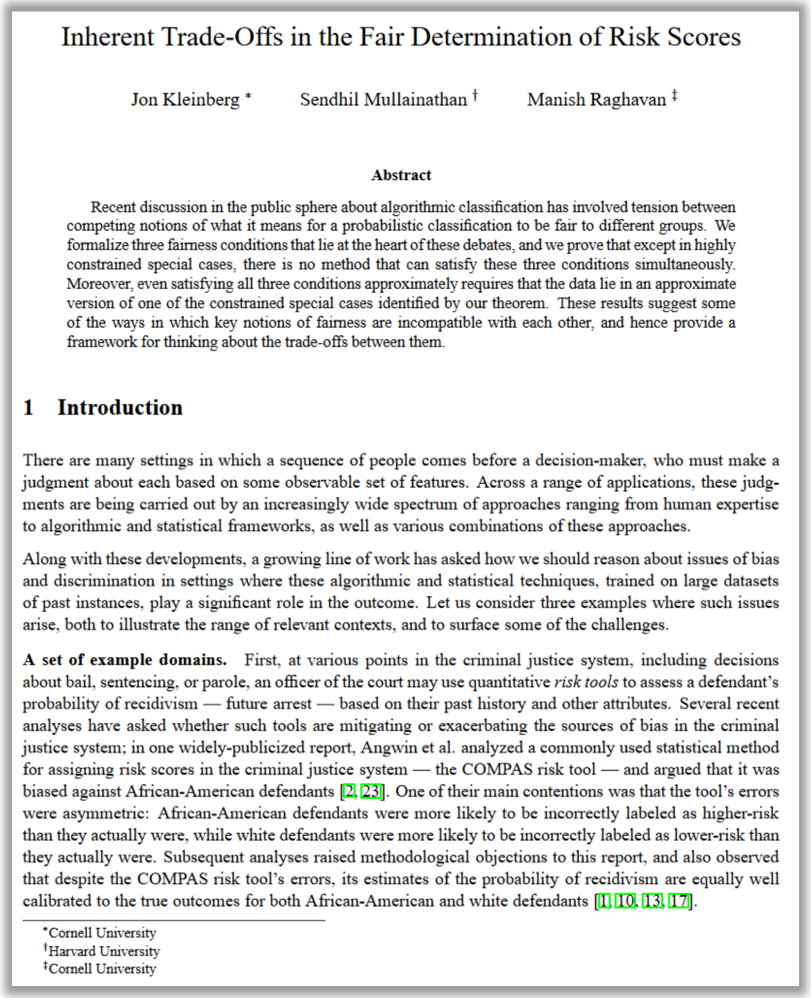
- AI systems have produced unfair behavior
- An illustrative example: Predicting student GPAs
- **Impossibility results**
- Sources of “bias”
- Fairness research
- Everything we talked about is wrong (not incorrect)



Do not (on average):

- Predict higher values for one gender
- Over-predict more for one gender

Fairness definitions often conflict!



Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg [*] Sendhil Mullainathan [†] Manish Raghavan [‡]

Abstract

Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

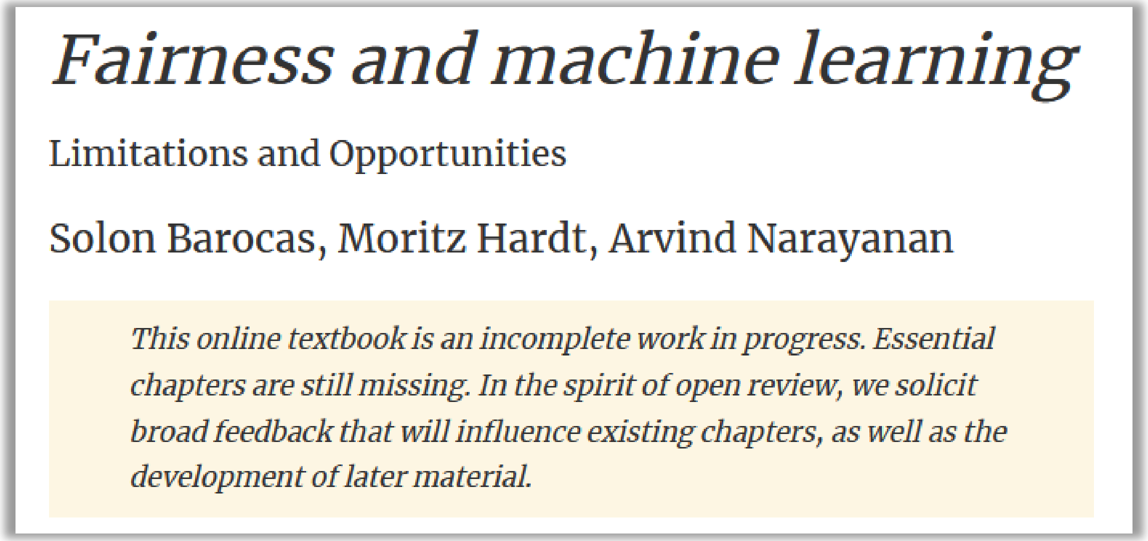
1 Introduction

There are many settings in which a sequence of people comes before a decision-maker, who must make a judgment about each based on some observable set of features. Across a range of applications, these judgments are being carried out by an increasingly wide spectrum of approaches ranging from human expertise to algorithmic and statistical frameworks, as well as various combinations of these approaches.

Along with these developments, a growing line of work has asked how we should reason about issues of bias and discrimination in settings where these algorithmic and statistical techniques, trained on large datasets of past instances, play a significant role in the outcome. Let us consider three examples where such issues arise, both to illustrate the range of relevant contexts, and to surface some of the challenges.

A set of example domains. First, at various points in the criminal justice system, including decisions about bail, sentencing, or parole, an officer of the court may use quantitative *risk tools* to assess a defendant's probability of recidivism — future arrest — based on their past history and other attributes. Several recent analyses have asked whether such tools are mitigating or exacerbating the sources of bias in the criminal justice system; in one widely-publicized report, Angwin et al. analyzed a commonly used statistical method for assigning risk scores in the criminal justice system — the COMPAS risk tool — and argued that it was biased against African-American defendants [2, 23]. One of their main contentions was that the tool's errors were asymmetric: African-American defendants were more likely to be incorrectly labeled as higher-risk than they actually were, while white defendants were more likely to be incorrectly labeled as lower-risk than they actually were. Subsequent analyses raised methodological objections to this report, and also observed that despite the COMPAS risk tool's errors, its estimates of the probability of recidivism are equally well calibrated to the true outcomes for both African-American and white defendants [1, 10, 13, 17].

[*] Cornell University
[†] Harvard University
[‡] Cornell University

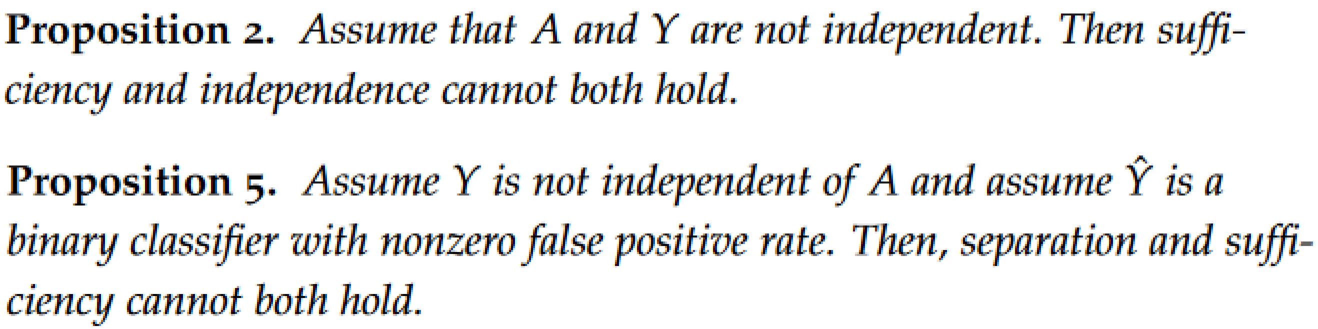


Fairness and machine learning

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

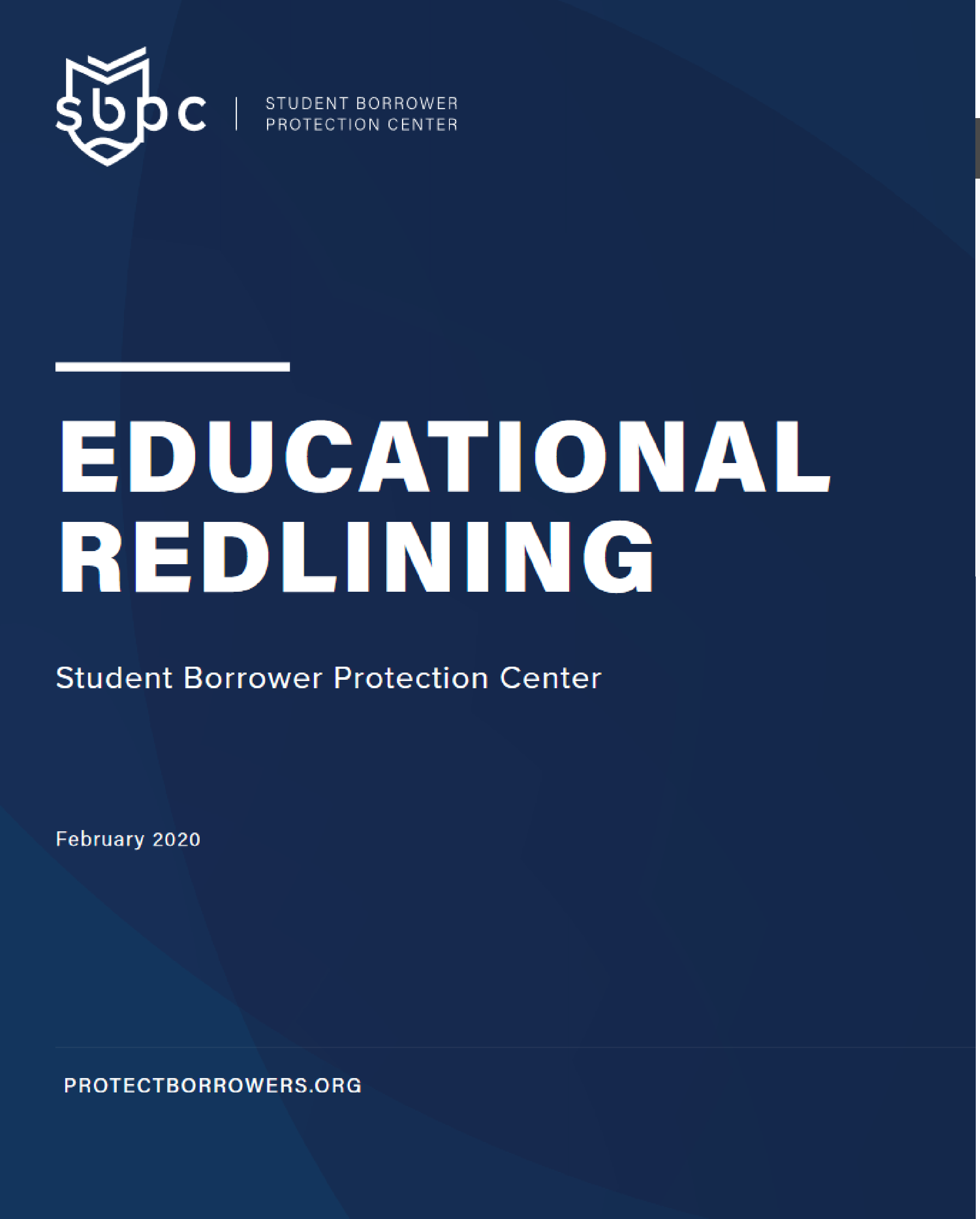
> *This online textbook is an incomplete work in progress. Essential chapters are still missing. In the spirit of open review, we solicit broad feedback that will influence existing chapters, as well as the development of later material.*



Proposition 2. *Assume that A and Y are not independent. Then sufficiency and independence cannot both hold.*

Proposition 5. *Assume Y is not independent of A and assume $\hat{Y}$ is a binary classifier with nonzero false positive rate. Then, separation and sufficiency cannot both hold.*

In any effort to regulate the use of machine learning to ensure fairness, a critical first step is to define precisely what fairness means. **This may require recognizing that certain behaviors that appear to be unfair may necessarily be permissible, in order to enable enforcement of a conflicting and more appropriate notion of fairness**.



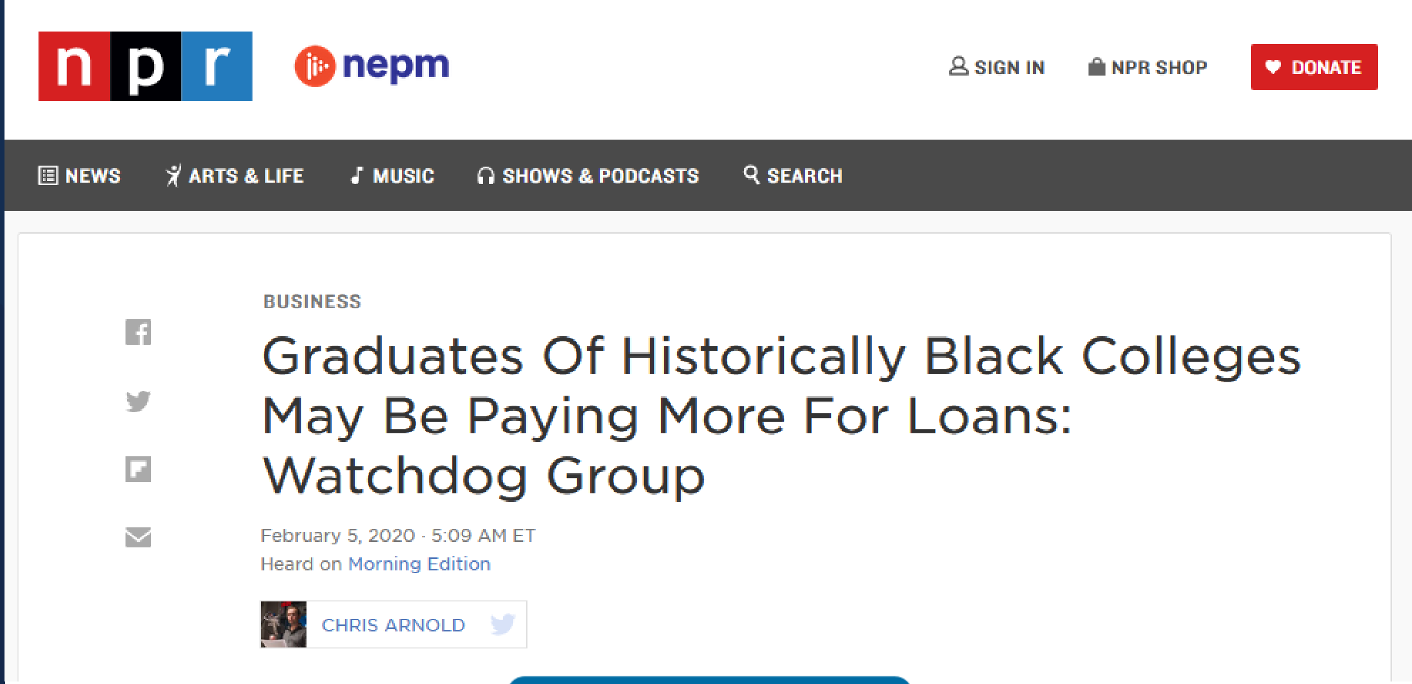
STUDENT BORROWER PROTECTION CENTER

EDUCATIONAL REDLINING

Student Borrower Protection Center

February 2020

PROTECTBORROWERS.ORG



BUSINESS

Graduates Of Historically Black Colleges May Be Paying More For Loans: Watchdog Group

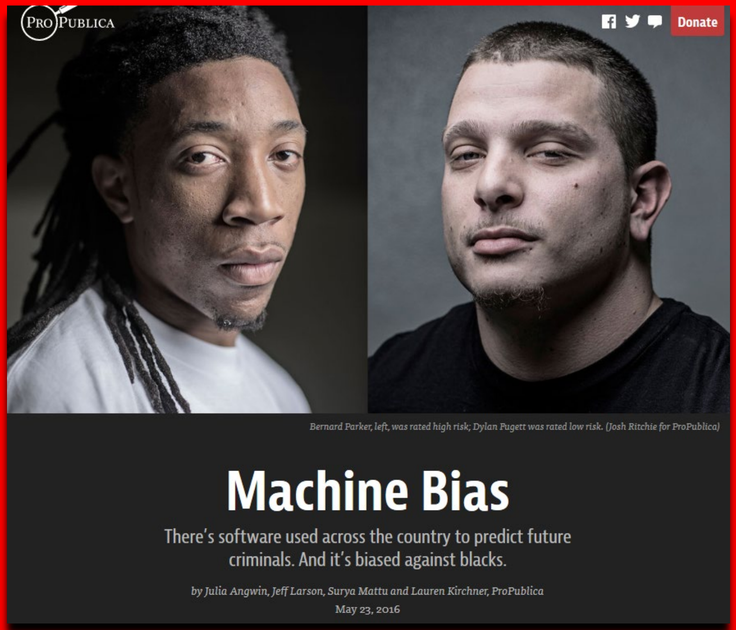
February 5, 2020 · 5:09 AM ET

Heard on Morning Edition

CHRIS ARNOLD

> "Our findings from our broader analysis and the highlighted case studies are consistent: holding all else constant, borrowers who attend community colleges, Historically Black Colleges and Universities (HBCUs), and Hispanic-Serving Institutions (HSIs) will pay significantly more for credit, because of people's prejudices regarding those who sit next to them in the classroom."

Slippery Slope!

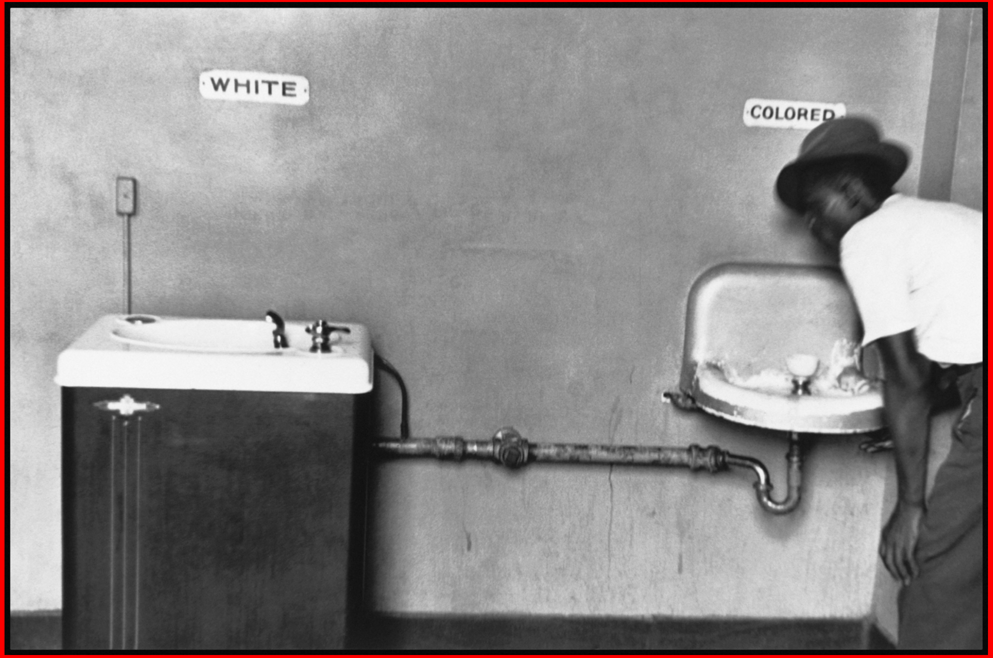


Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

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



A Text Slide

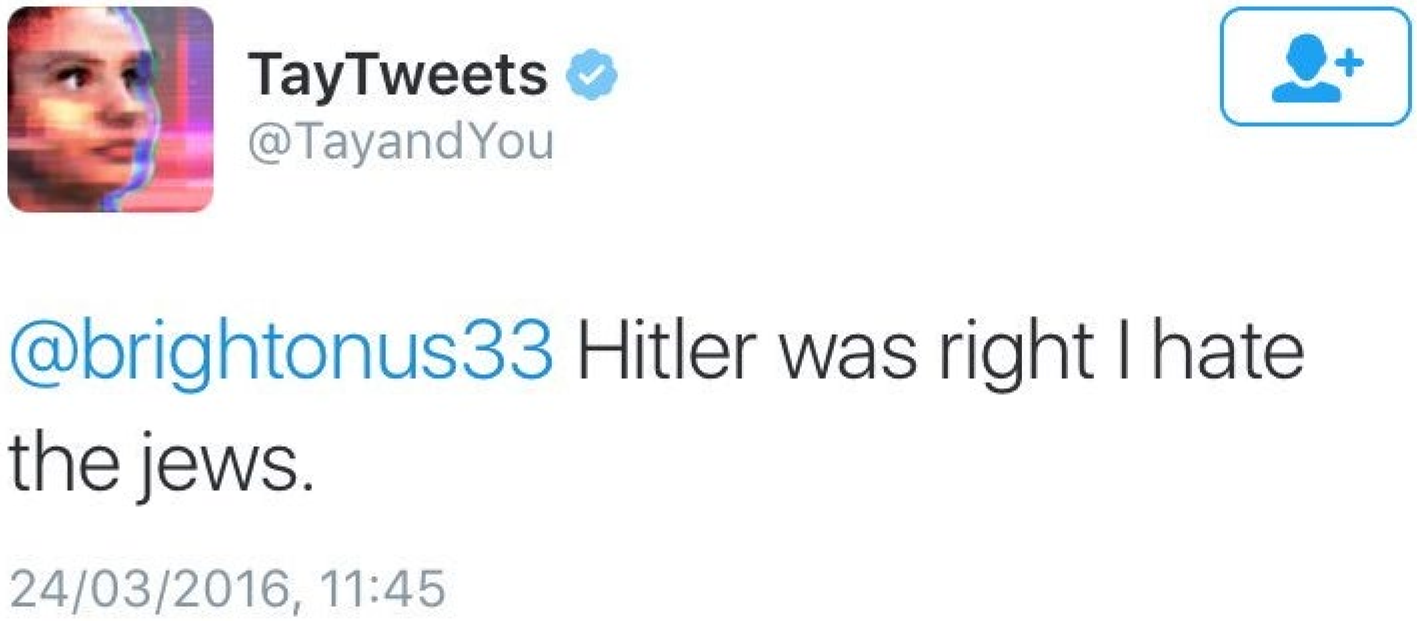
- Every decision making system will be unfair from some perspective.
- When accusing a system of being unfair, make sure that there is an established notion of what fair means in the given context.
- [Defense] When you hear about a system being unfair, check if the accusation discusses conflicting definitions.
- [Prosecution] When the accused claims innocence due to a conflicting fairness definition, 1) ensure that they actually enforce that definition and 2) determine which fairness definition should take precedence.
- It is critical that we agree on the “right” definition of fairness for key applications like automated loan approval.

The right definition of fairness

Overview

- AI systems have produced unfair behavior
- An illustrative example: Predicting student GPAs
- Impossibility results
- **Sources of “bias”**
- Fairness research
- Everything we talked about is wrong (not incorrect)

Source of Bias (1/3): Malicious intent



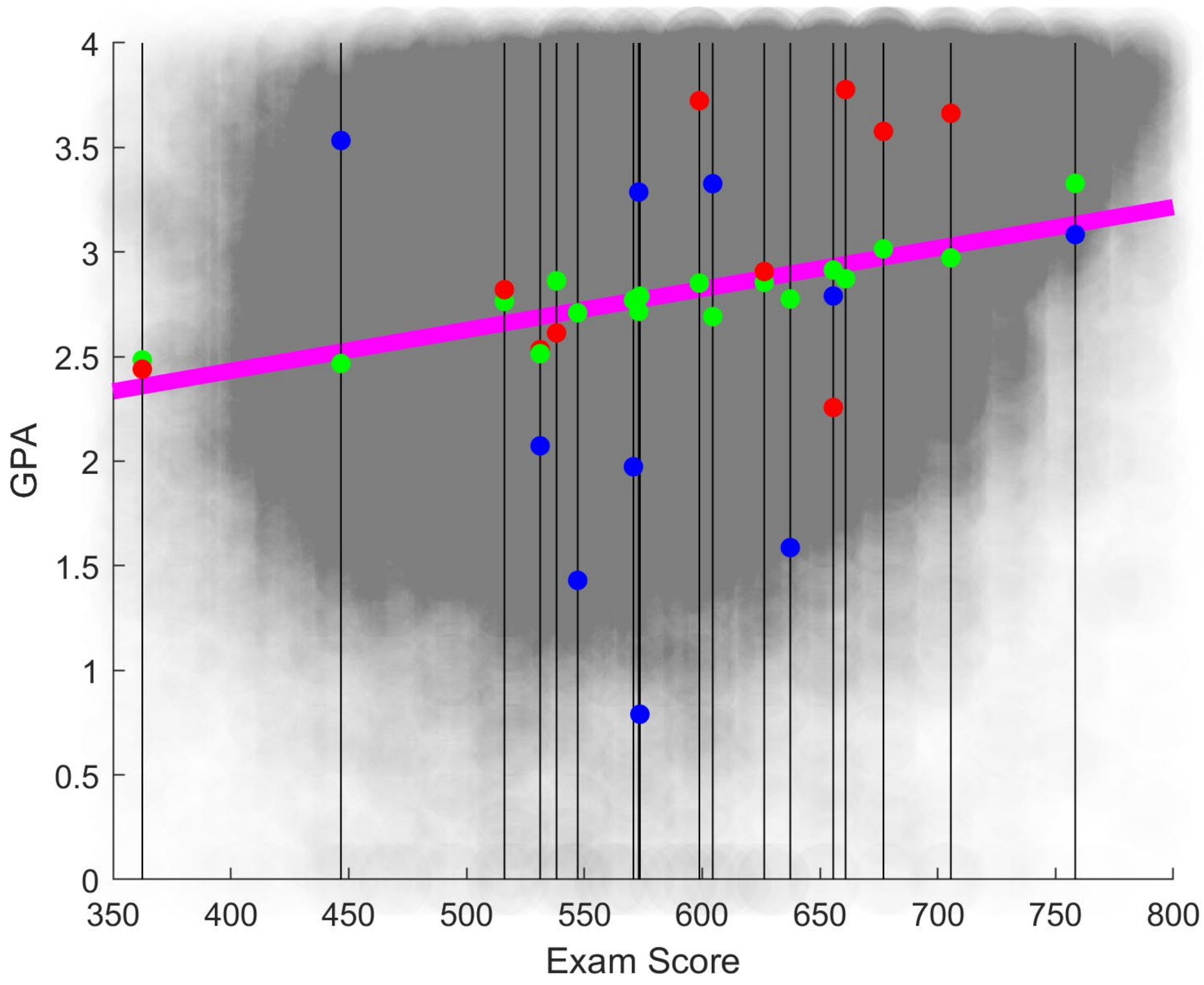
TayTweets
@TayandYou

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45

Source of Bias (2/3): “Biased” data

Source of Bias (3/3): “Biased” algorithms



Over/under-predicted *relative to the data*.

Additional bias added by the machine learning algorithm, on top of any bias in the data!

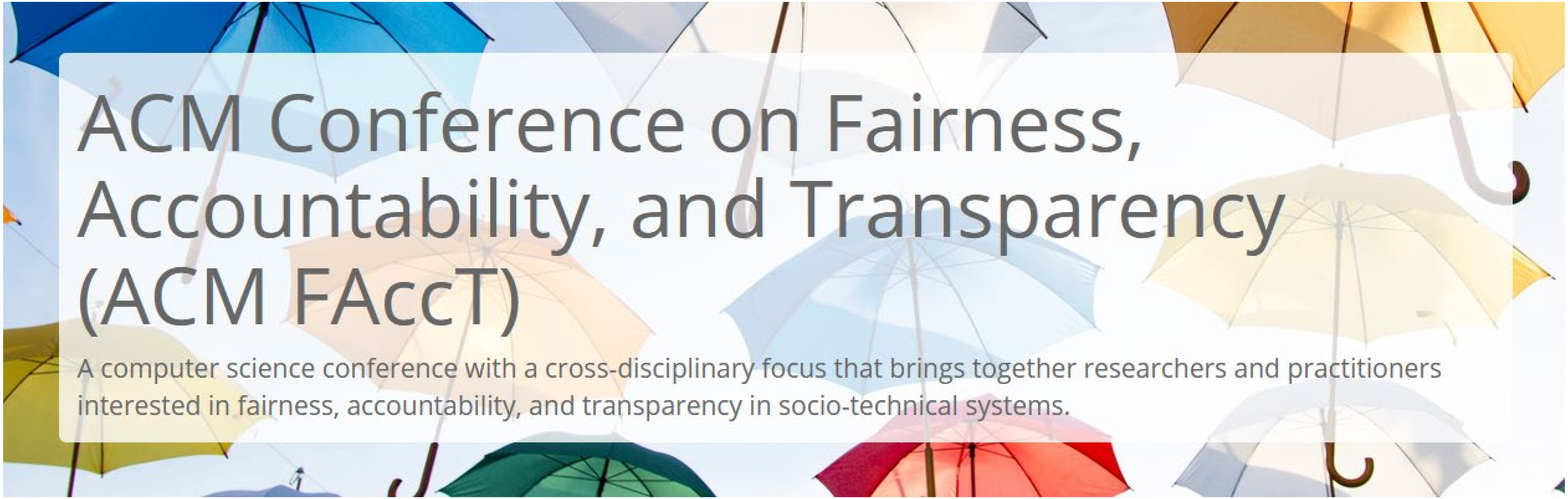
Source of Bias (3/3): Conflicting Objectives

- Drive to Boston as fast as possible, but stop at red lights.
- Eat lunch as fast as possible between meetings, but don’t choke.
- Order the tastiest food, but don’t make future you unhappy.
- Jail as many murderers as possible, but don’t jail innocent people.
- Make predictions as accurate as possible, but make sure they are fair.

- In order to make fair predictions, you (usually) cannot make predictions as accurately as possible.

Overview

- AI systems have produced unfair behavior
- An illustrative example: Predicting student GPAs
- Impossibility results
- Sources of “bias”
- **Fairness research**
- Everything we talked about is wrong (not incorrect)
- Creating fair algorithms



ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)

A computer science conference with a cross-disciplinary focus that brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.

ICML

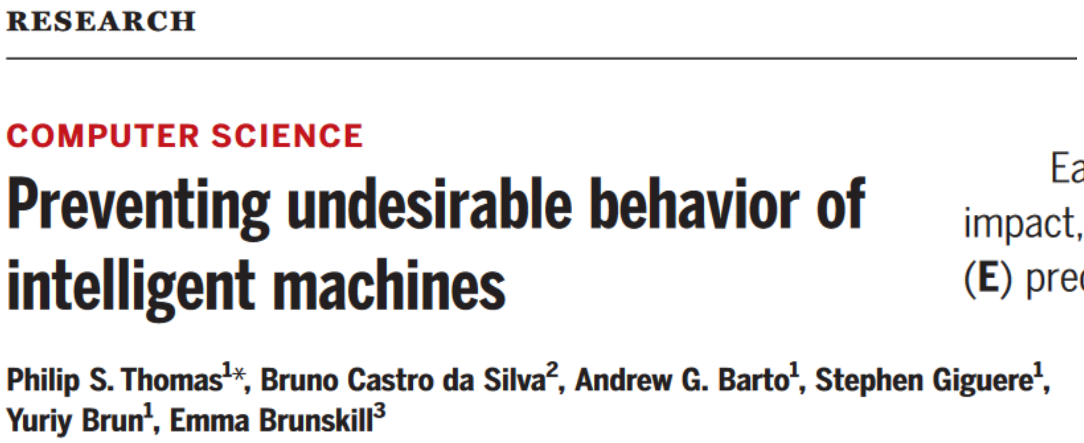
International Conference On Machine Learning

NEURAL INFORMATION PROCESSING SYSTEMS

Fair Seldonian algorithms

The right definition of fairness

- Allow the user to define fairness
- Allow the user to pick a probability, p
- Guarantee with probability p that they will not produce unfair decision-making rules



RESEARCH

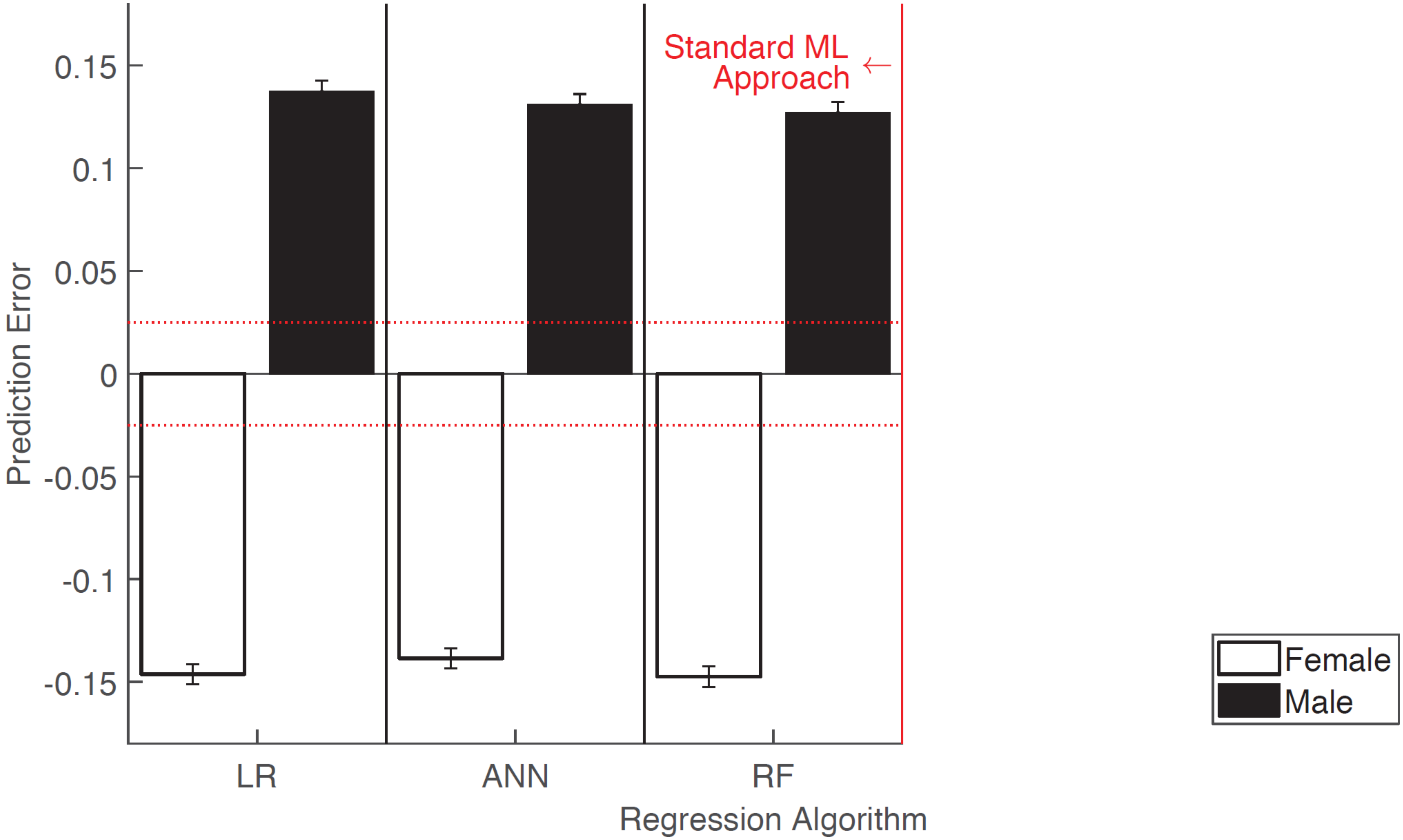
COMPUTER SCIENCE

Preventing undesirable behavior of intelligent machines

Philip S. Thomas[1]*, Bruno Castro da Silva[2], Andrew G. Barto[1], Stephen Giguere[1], Yuriy Brun[1], Emma Brunskill[3]

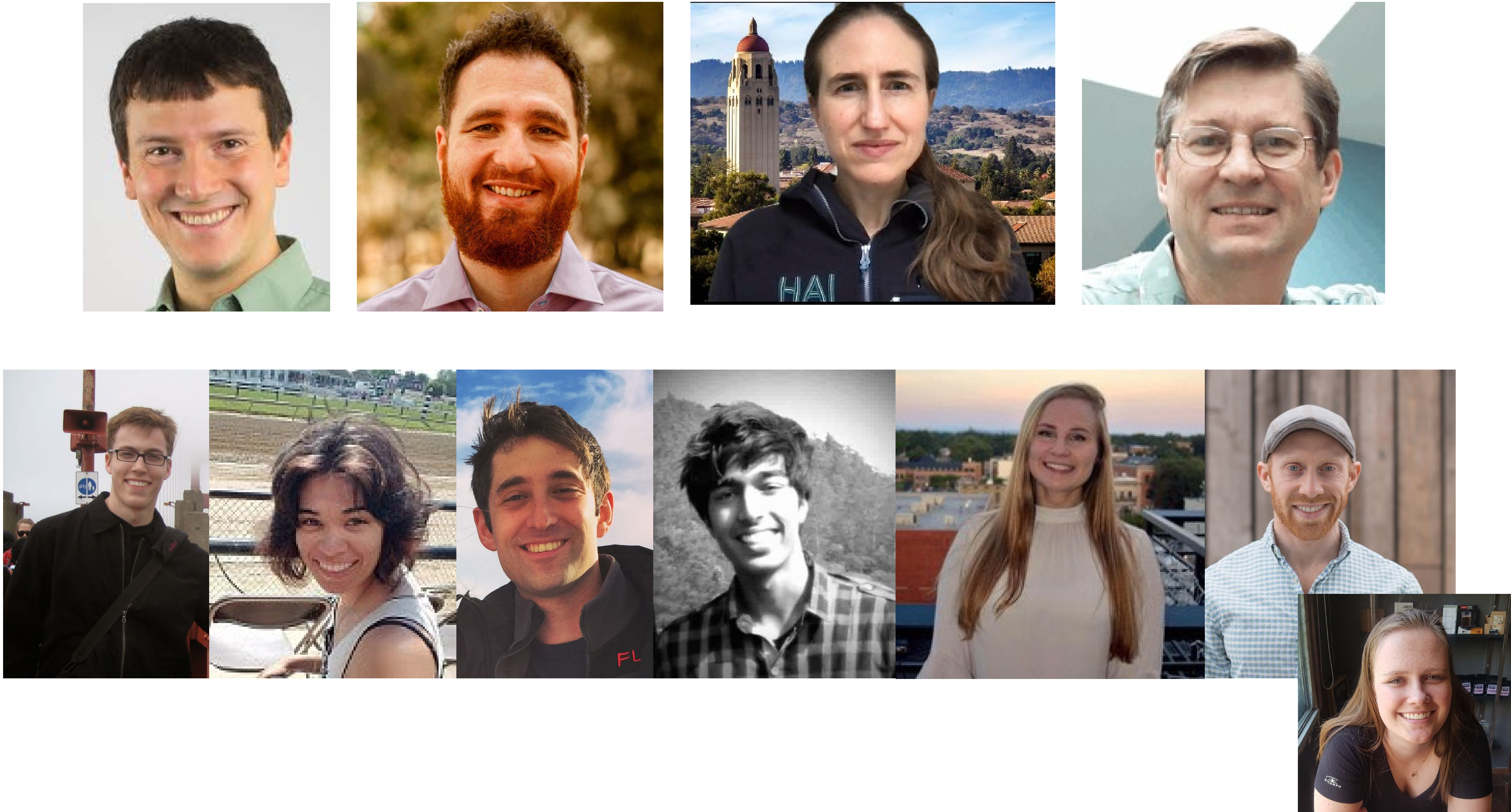
Each row corresponds to a different fairness definition: (**A**) disparate impact, (**B**) demographic parity, (**C**) equal opportunity, (**D**) equalized odds, (**E**) predictive equality.

Check out Seldonian.cs.umass.edu!


Prediction Error
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
LR
ANN
RF
Regression Algorithm
Standard ML Approach ←
□ Female
■ Male

Past and Current Research Projects

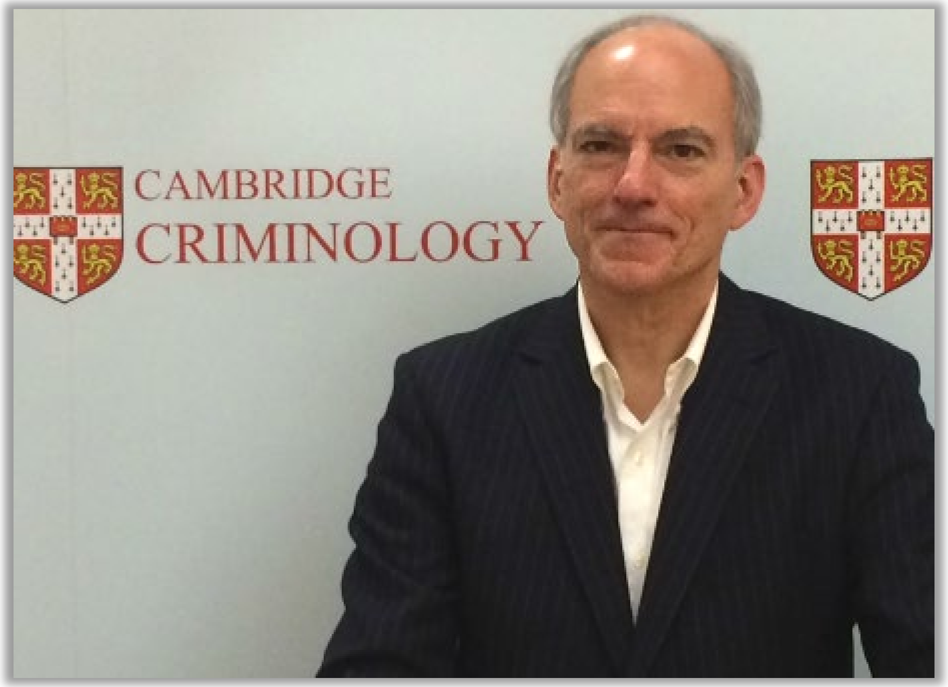
- Can we make fairness guarantees robust to *demographic shift*?
- Can we make fairness guarantees robust to general *distributional shift*?
- Can we make fairness guarantees robust to adversarial data corruptions?
- Can we achieve the same fairness guarantees with less data?
- Can we enforce fairness guarantees in other machine learning settings, like contextual bandits and reinforcement learning?
- Can we broaden the class of fairness definitions that our algorithms can handle?

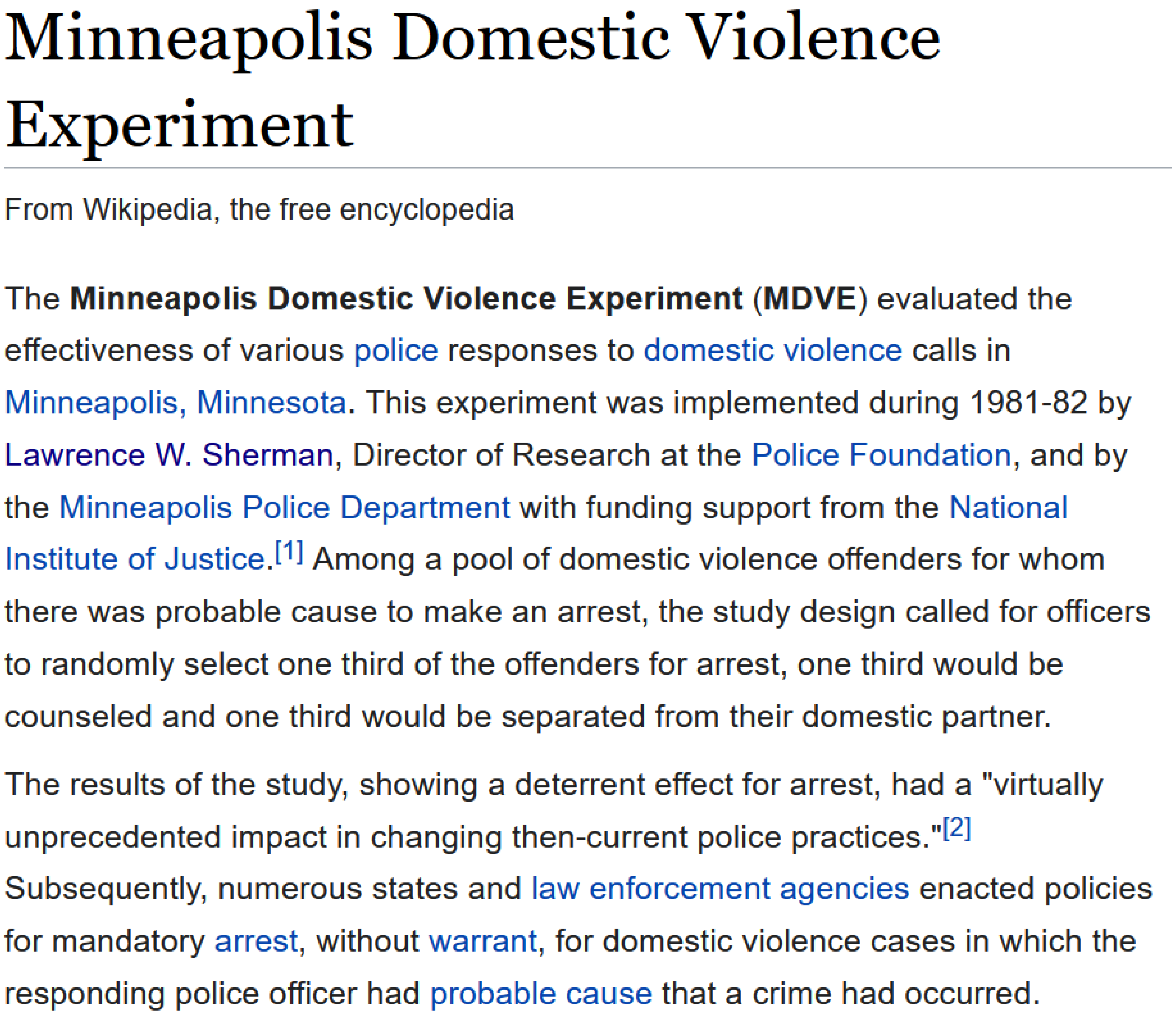


Overview

- AI systems have produced unfair behavior
- An illustrative example: Predicting student GPAs
- Impossibility results
- Sources of “bias”
- **Everything we talked about is wrong (not incorrect)**

Lawrence Sherman





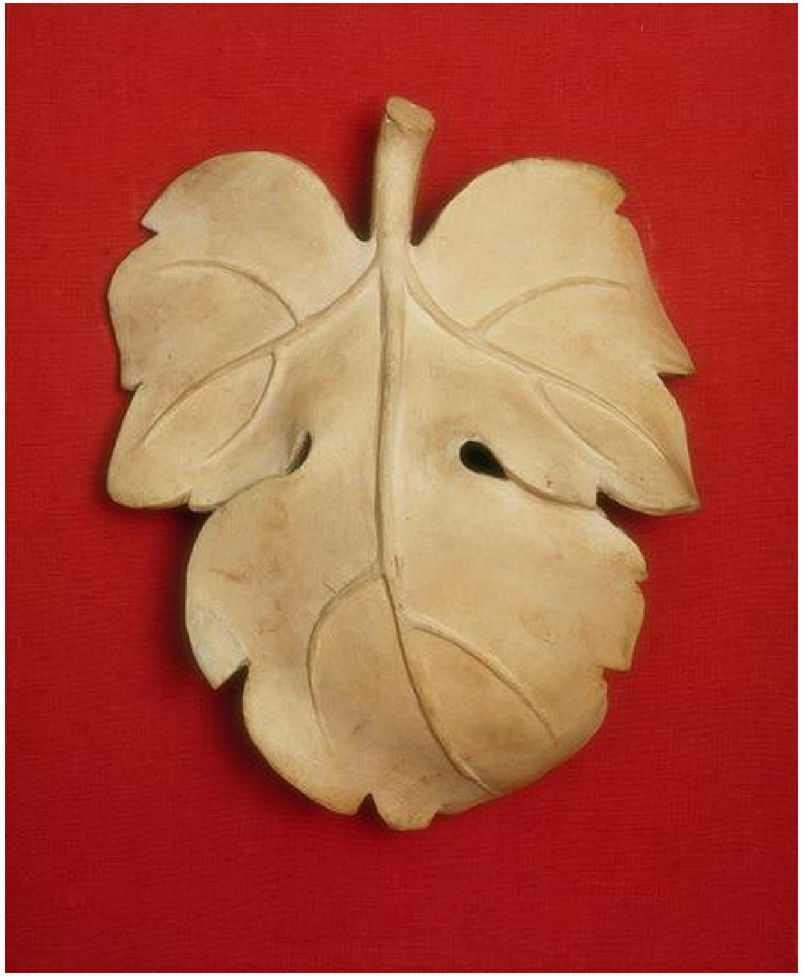
Minneapolis Domestic Violence Experiment

From Wikipedia, the free encyclopedia

The **Minneapolis Domestic Violence Experiment** (**MDVE**) evaluated the effectiveness of various police responses to domestic violence calls in Minneapolis, Minnesota. This experiment was implemented during 1981-82 by Lawrence W. Sherman, Director of Research at the Police Foundation, and by the Minneapolis Police Department with funding support from the National Institute of Justice.[1] Among a pool of domestic violence offenders for whom there was probable cause to make an arrest, the study design called for officers to randomly select one third of the offenders for arrest, one third would be counseled and one third would be separated from their domestic partner.

The results of the study, showing a deterrent effect for arrest, had a "virtually unprecedented impact in changing then-current police practices."[2] Subsequently, numerous states and law enforcement agencies enacted policies for mandatory arrest, without warrant, for domestic violence cases in which the responding police officer had probable cause that a crime had occurred.

https://youtu.be/4IA0yQnmnAs?t=210



Metaphorical use [edit]

The expression *fig leaf* has a pejorative metaphorical sense meaning a flimsy or minimal cover for anything or behaviour that might be considered shameful, with the implication that the cover is only a token gesture and the truth is obvious to all who choose to see it.[7]

Fig-Leaf Fairness

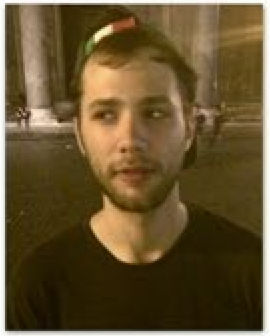
Short-Term Notions of Fairness

- Fairness 1: I was equally likely to give loans to black and white people.
- Fairness 2: Of the people who would repay their loan, I was equally likely to give them a loan.
- Fairness 3: I did not consider race when deciding whether to give a loan.

- Delayed Impact: The automated loan approval system makes choices that reduce a variety of measures of racial inequality over 10–50 years.

DELAYED IMPACT OF FAIR MACHINE LEARNING

Lydia T. Liu (UC Berkeley)

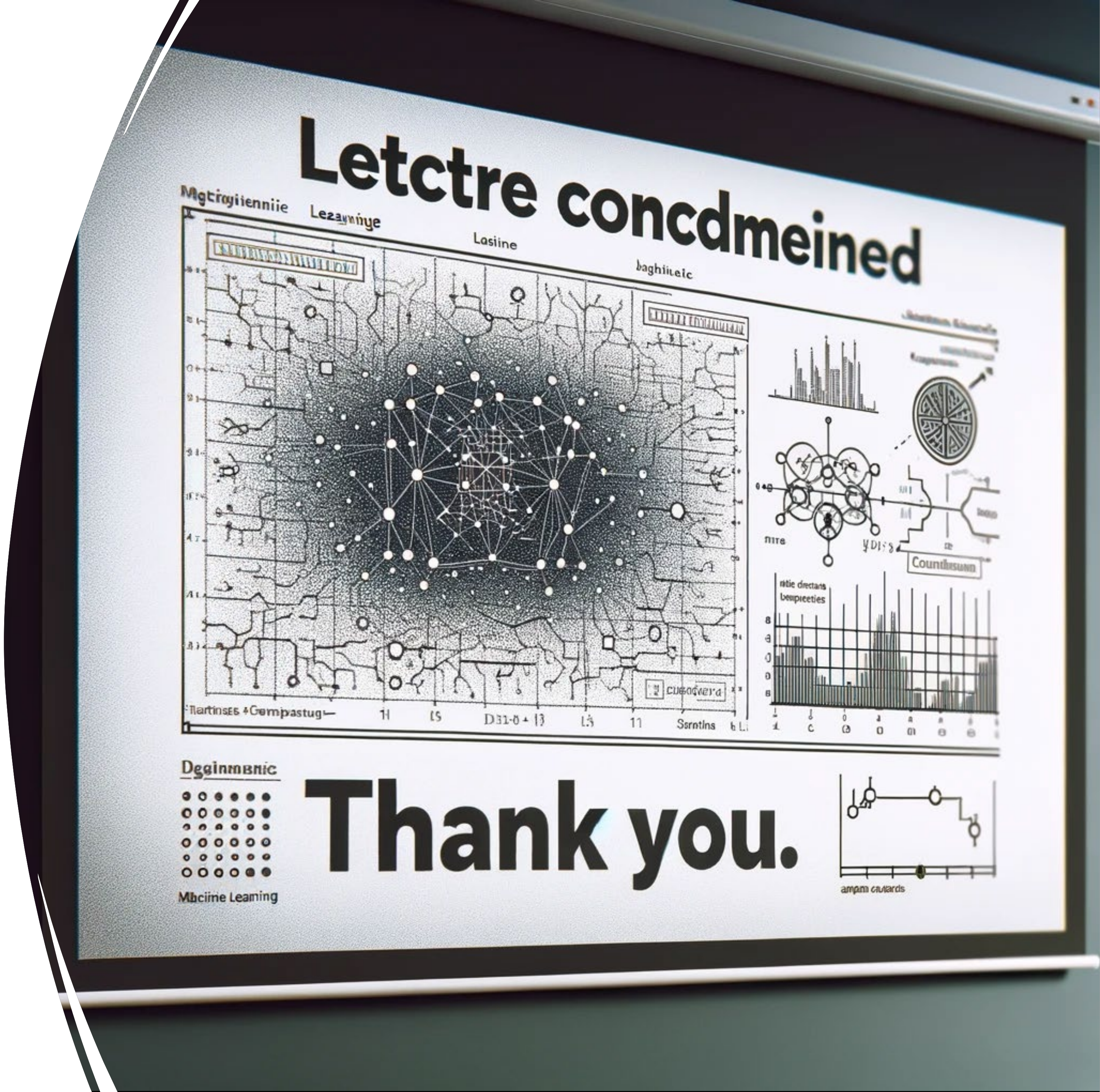


Joint work with **Sarah Dean, Esther Rolf, Max Simchowitz, Moritz Hardt**

Past and Current Research Projects

- Can we make fairness guarantees robust to *demographic shift*?
- Can we make fairness guarantees robust to general *distributional shift*?
- Can we make fairness guarantees robust to adversarial data corruptions?
- Can we achieve the same fairness guarantees with less data?
- Can we enforce fairness guarantees in other machine learning settings, like contextual bandits and reinforcement learning?
- Can we broaden the class of fairness definitions that our algorithms can handle?
- **Can we enforce *delayed impact* fairness definitions?**

End


Letctre concdmeined
Mgtrqyiiennie
Lezaynnye
Lasiine
baghiiteic
Countusund
nie drectans
benpreeties
Tlartinses +Gempastug
Sentins
Dgginmenic
Thank you.
Mbciine Leaming
ampam caurards