Security and Fairness of Deep Learning

Introduction to Fairness in ML

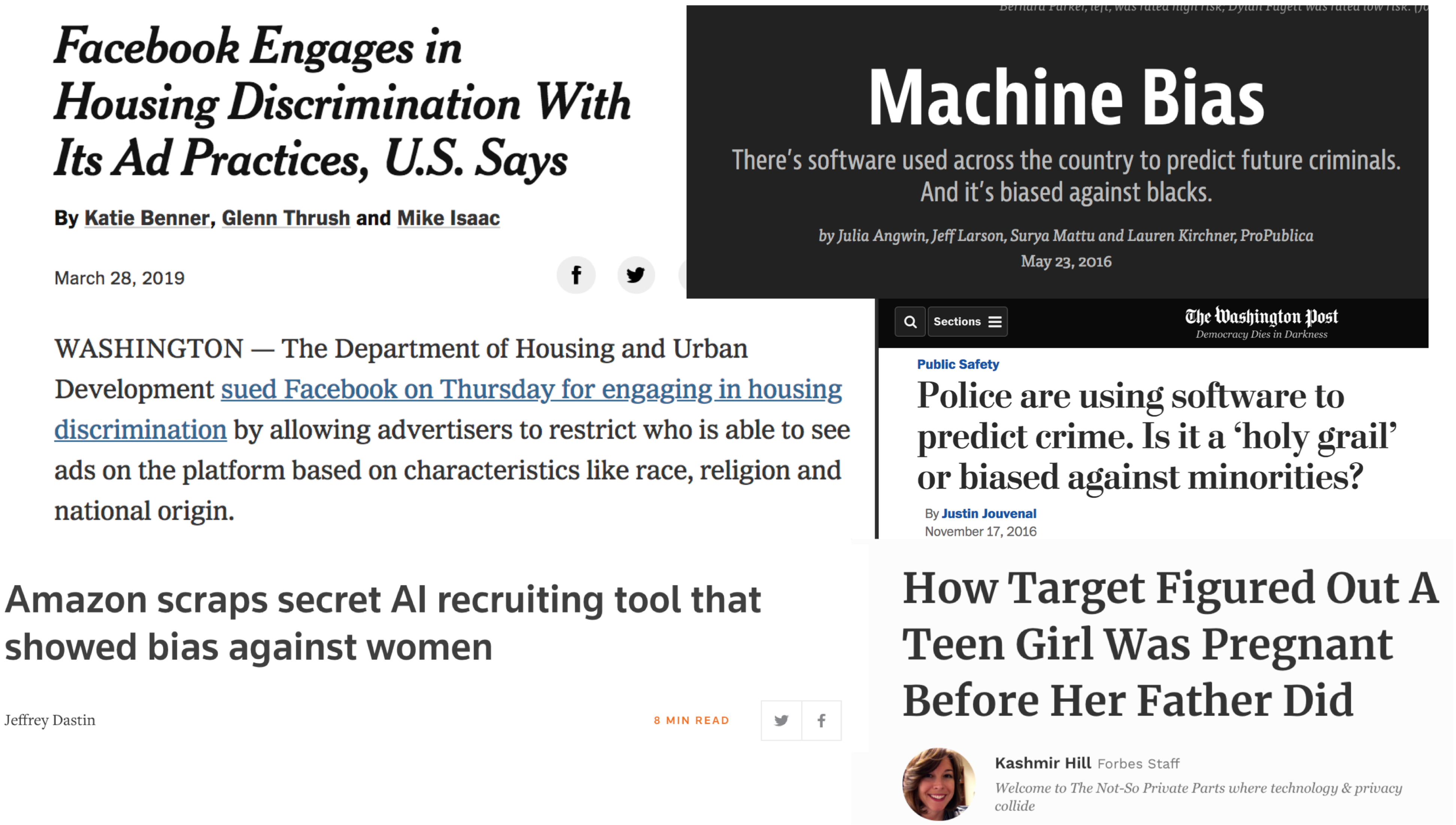
Emily Black

CMU

Spring 2019

Overview

- Historical Background
- Fairness Pipeline
  - Unfairness from model itself: Feature-wise bias amplification
  - Unfairness from data collection: Nikon biased facial recognition
  - Unfairness from underlying world: Amazon Recruitment, Word Embeddings
- How do we make a fair model?
  - Removing Protected Attribute
  - GANs for Fairness
- Further Thoughts
  - Delayed Impact of Fair Machine Learning
  - Group vs Individual Fairness



Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

By Katie Benner, Glenn Thrush and Mike Isaac

March 28, 2019

WASHINGTON — The Department of Housing and Urban Development sued Facebook on Thursday for engaging in housing discrimination by allowing advertisers to restrict who is able to see ads on the platform based on characteristics like race, religion and national origin.

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Jo

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

Public Safety

Police are using software to predict crime. Is it a 'holy grail' or biased against minorities?

By Justin Jouvenal

November 17, 2016

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

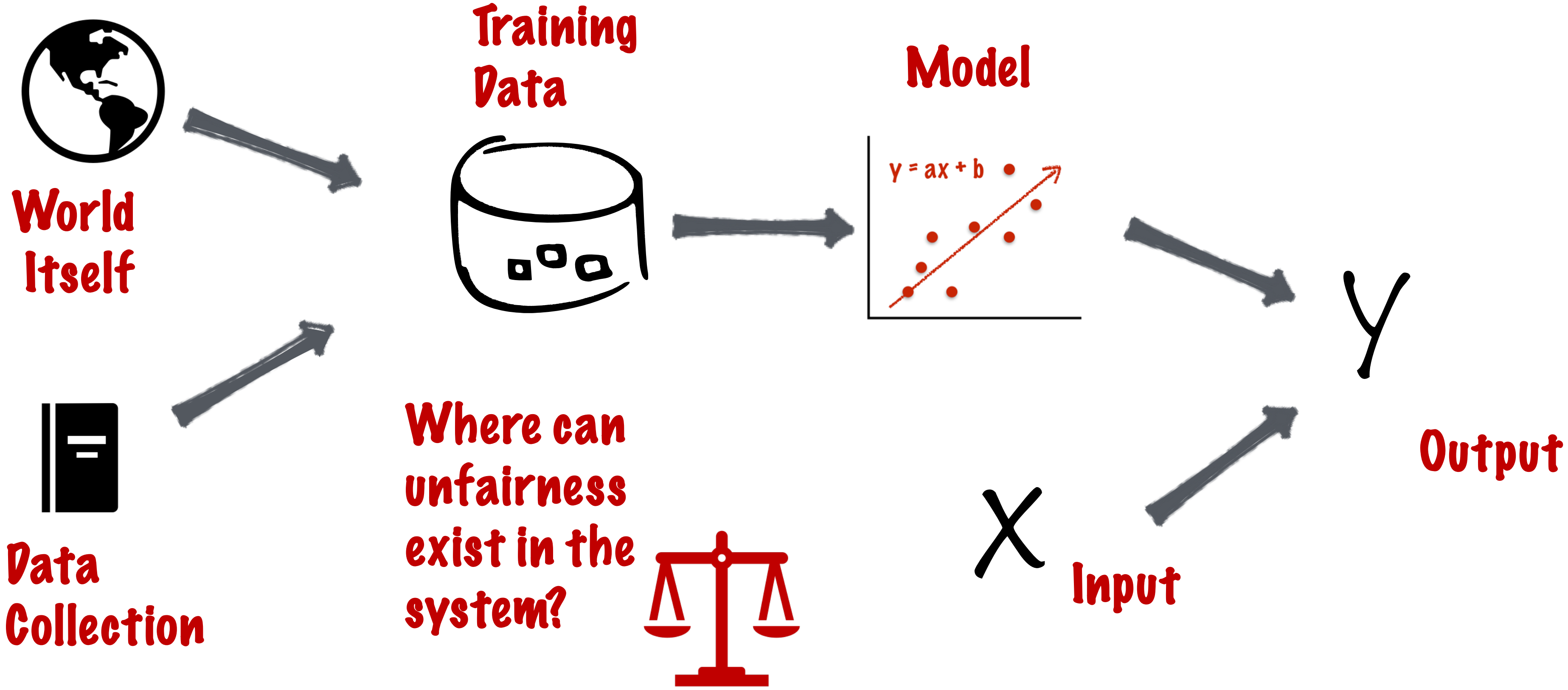
Kashmir Hill Forbes Staff

Welcome to The Not-So Private Parts where technology & privacy collide

Hold on, how can Facebook engage in housing discrimination?

- Under the Fair Housing Act, it’s illegal to “make housing unavailable” or “assign a person to a particular neighborhood” (and many other stipulations) on the basis of race, sex, religion, etc.
- If you prevent one group—in this case, often based on race—from seeing ads for certain properties, you are essentially making that housing unavailable.
- Restricting who sees an ad for a given house from “black affinity groups” is like hiding the “for sale” sign in front of a house whenever a black person walks by
- Facebook was doing this even in cases where the advertising agency did not request this ad-segregation: they use their own algorithms to decide who is most likely to engage with the ad (thus bring them more money), and so if some ethnic group was deemed less likely to engage, they would not show the ad to that ethnic group

Machine Learning Pipeline



World Itself

Data Collection

Training Data

Model

Where can unfairness exist in the system?

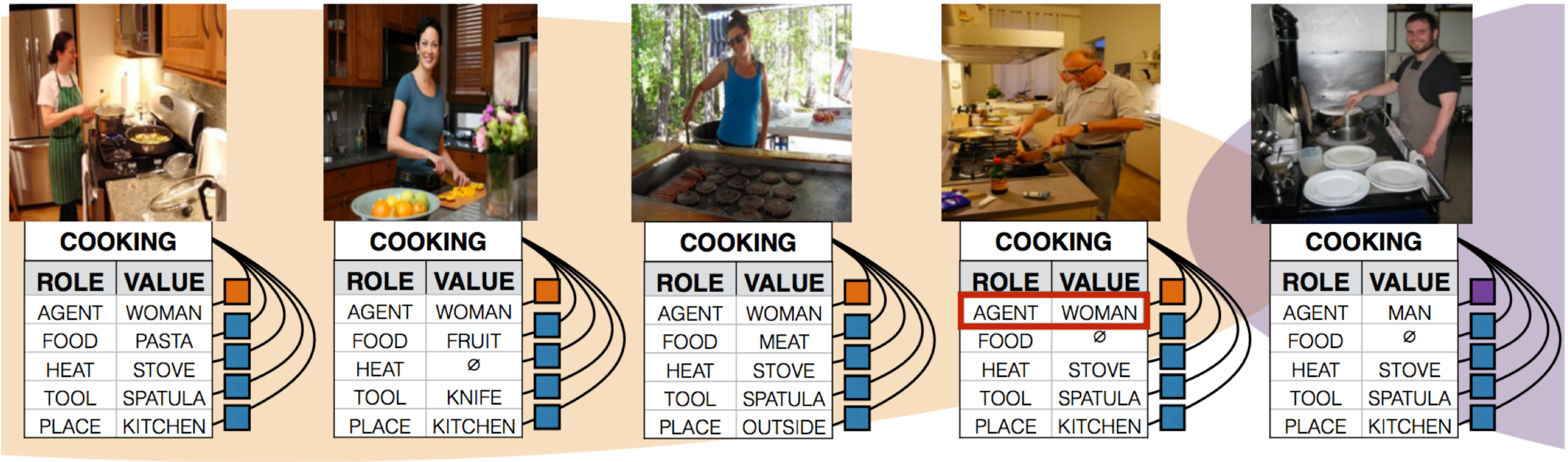
Y

Output

X

Input

Unfairness in Model: Recall Bias Amplification



COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	∅
TOOL	KNIFE
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	STOVE
TOOL	SPATULA
PLACE	OUTSIDE

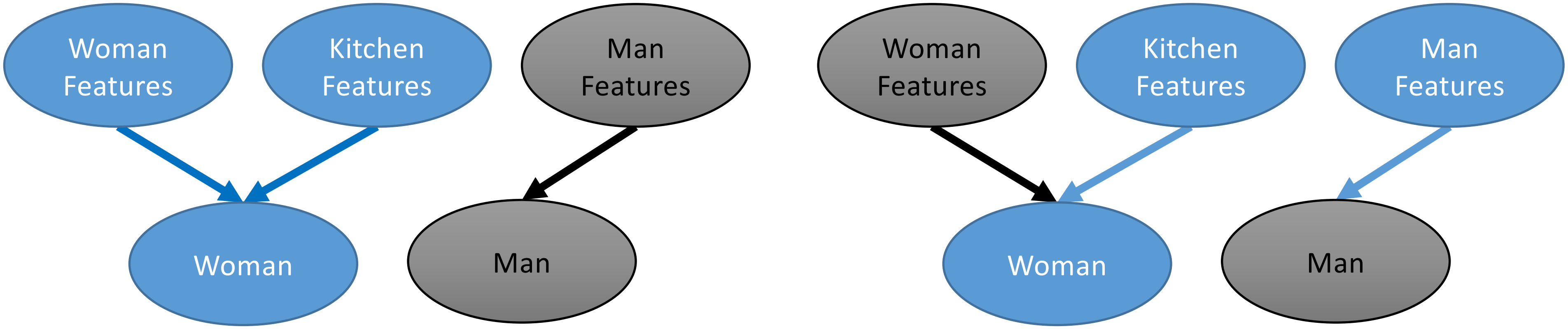
COOKING	
ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING	
ROLE	VALUE
AGENT	MAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

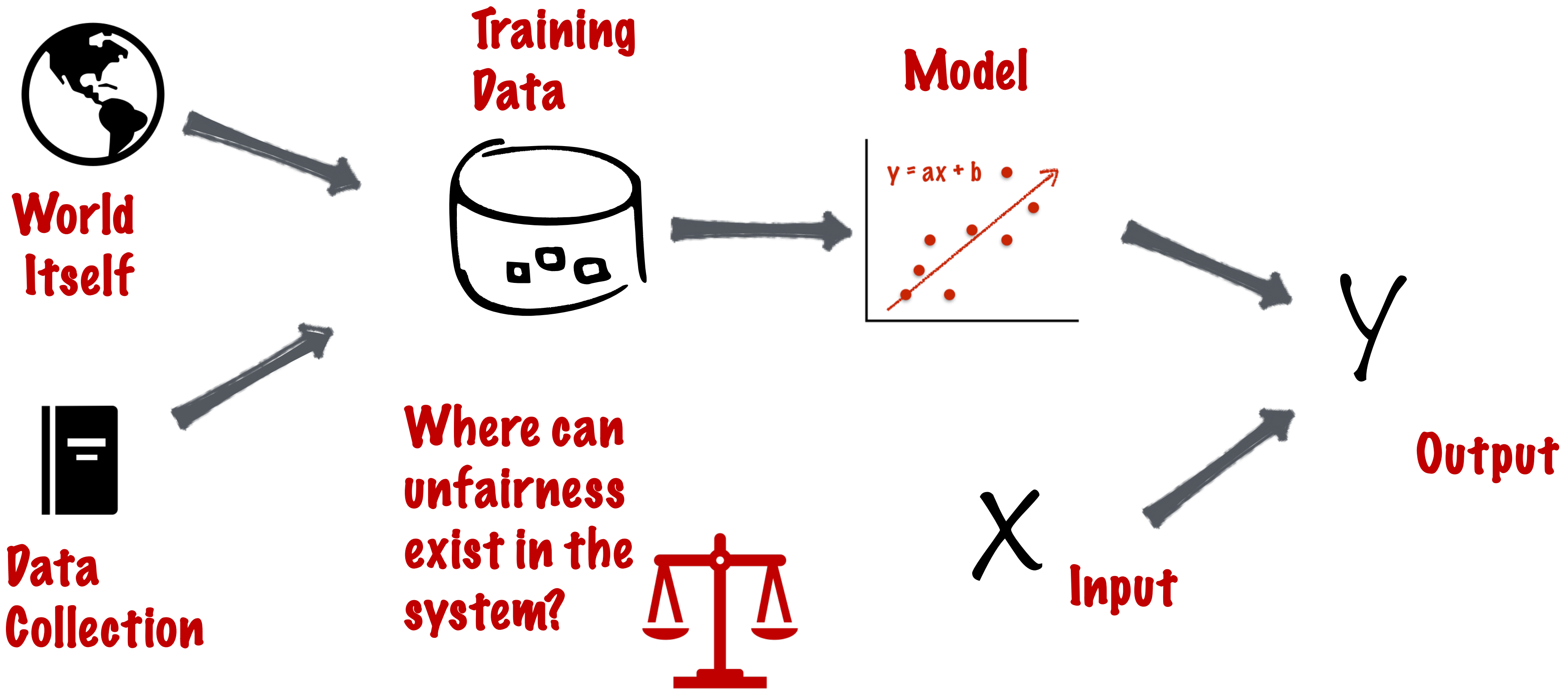
Class	Man	Woman
Data prior	33%	67%
Pred. prior	16%	84%

Unfairness in Model: Recall Bias Amplification

We say a model exhibits *bias amplification* if the prior distribution of the model's predictions does not match that of the data: in particular, we don't want the model to *create* or *exaggerate disparities* in the training data.



Machine Learning Pipeline



World Itself

Data Collection

Training Data

Model

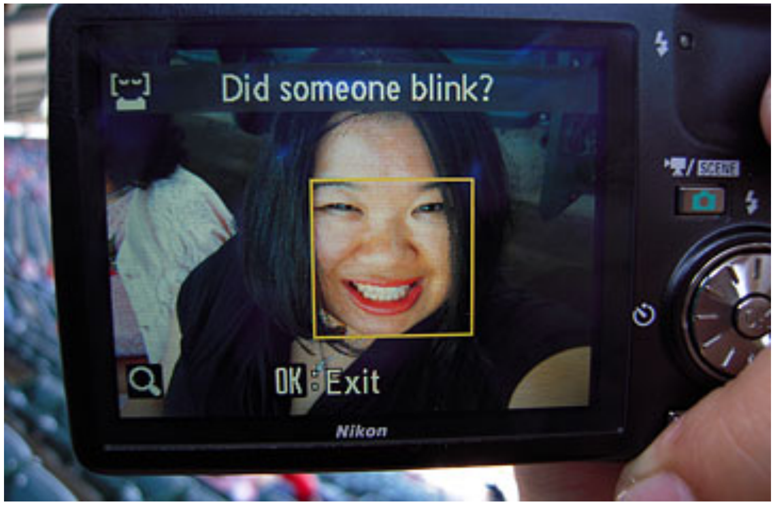
Where can unfairness exist in the system?

X Input

Y Output

Unfairness in Data Collection

- Nikon blink-recognition always thinks Asian faces are blinking
  - While it’s not certain why the problem exists (Nikon has not given a concrete explanation), it’s feasible that it is due to unbalanced training data



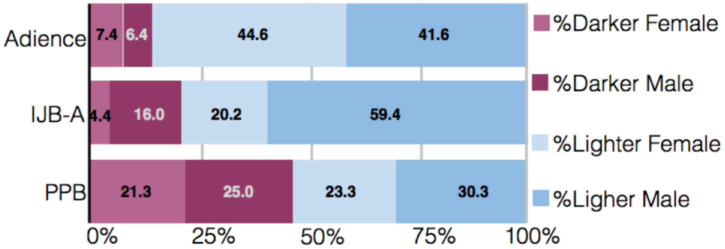
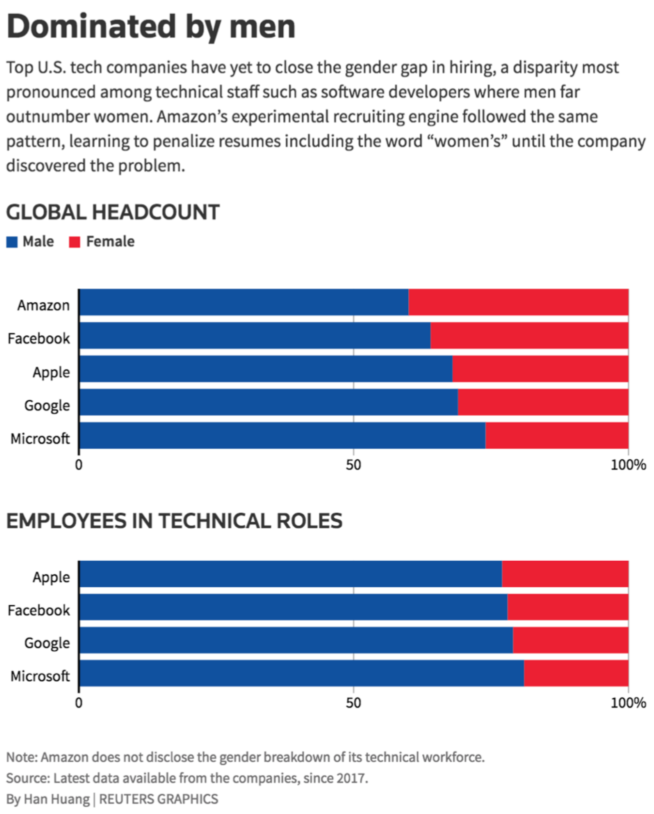


Figure 3: The percentage of darker female, lighter female, darker male, and lighter male subjects in PPB, IJB-A and Adience. Only 4.4% of subjects in Adience are darker-skinned and female in comparison to 21.3% in PPB.

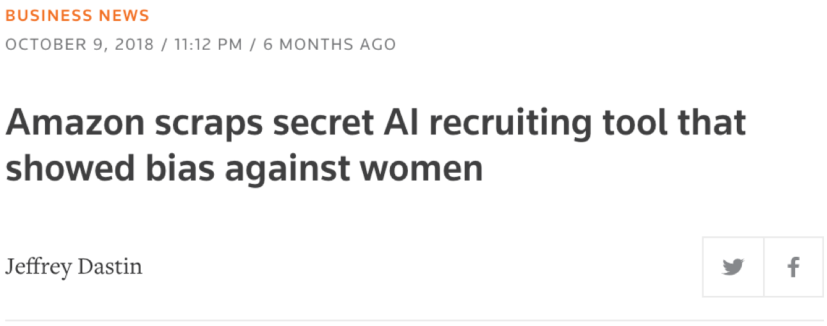
Unfairness in the World: Amazon Recruitment



Dominated by men

Top U.S. tech companies have yet to close the gender gap in hiring, a disparity most pronounced among technical staff such as software developers where men far outnumber women. Amazon's experimental recruiting engine followed the same pattern, learning to penalize resumes including the word "women's" until the company discovered the problem.

Note: Amazon does not disclose the gender breakdown of its technical workforce.
Source: Latest data available from the companies, since 2017.
By Han Huang | REUTERS GRAPHICS



BUSINESS NEWS

OCTOBER 9, 2018 / 11:12 PM / 6 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

- Amazon has a history of hiring predominantly men
- Amazon recruitment tool learned to penalize women's applications to match the distribution in the biased training data
- penalize the word "women" e.g. "women's soccer coach" etc
- favor words more often used in men's applications, eg "execute"

Unfairness in the World: Word Embeddings

Perhaps we've heard of the Bag of Words model and TFIDF

- Problem with these approaches– bag of words assumes conditional independence of words: no notion of context!
- extremely high dimensional, leads to problems
- Have to train these models for every unique problem, non-transferable

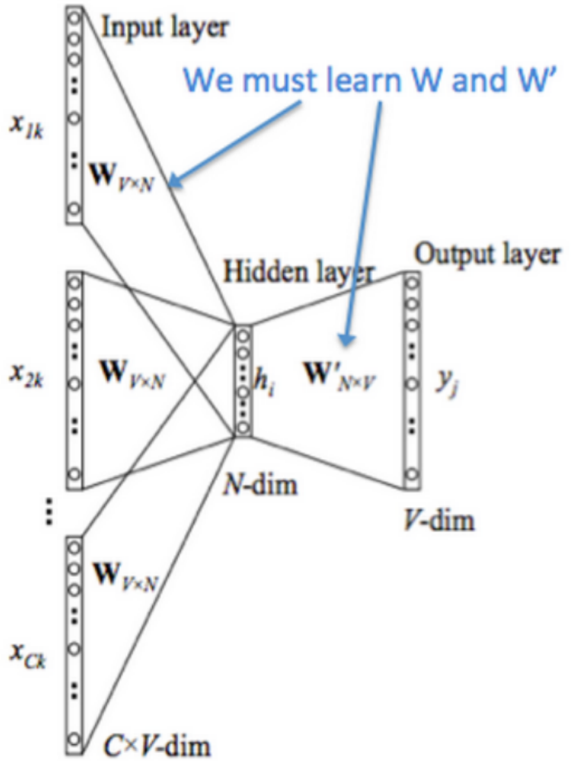
What's so great about word embeddings?

- Word embedding vectors capture context
- Lower dimensional vectors
- Often transferable between problems
- The geometry of word embeddings has some interesting properties—ability to compute analogies, e.g.
  - man-woman = king-queen

How do we get word embeddings? (CBOW)

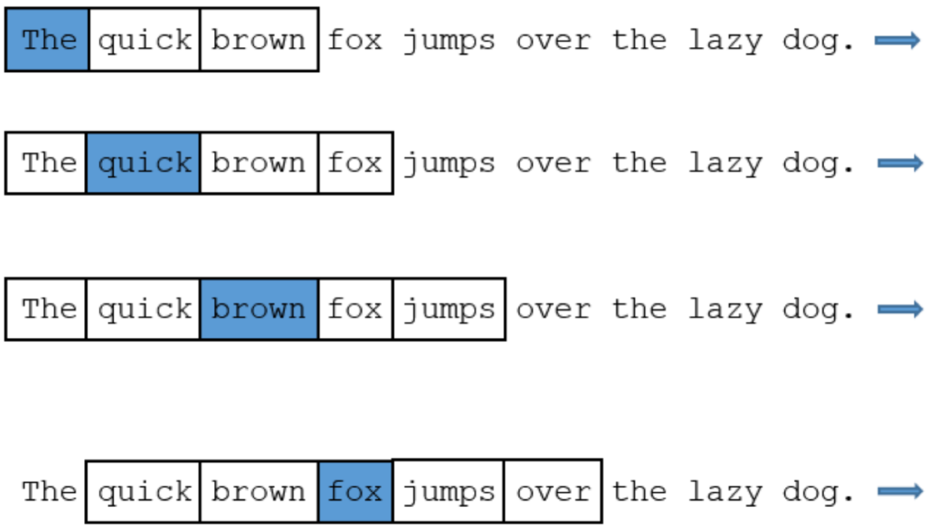
- Input: set of one-hot vectors representing the context from corpus, for a window of h
  - e.g. if h=2, the window in the sentence ”The fluffy dog barked as it chased a cat” is [the, fluffy, barked, as]

- Output: prediction on which word w in the corpus had that context
- But the word embeddings vectors are actually the *weights in the hidden layer!!!!*



Training the network

Source Text



The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.
The quick brown fox jumps over the lazy dog.

([quick,brown], the)

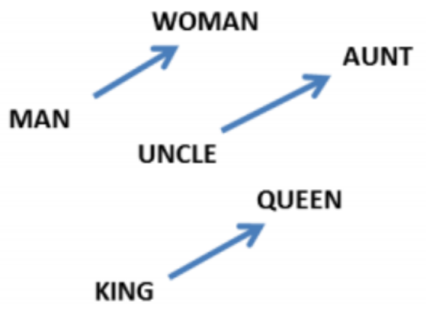
([the, brown, fox], quick)

([the, quick, fox, jumps], brown)

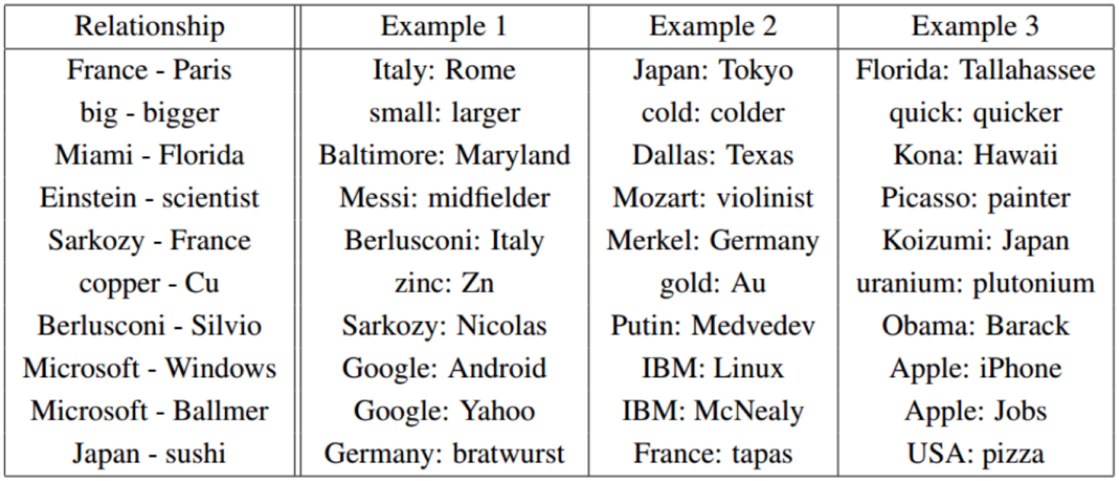
([quick, brown, jumps, over], fox)

- We train our NN with pairs of (context, target)
- The loss function is the cross entropy of the prediction and the true label
- But the actual word embeddings are the weights in the hidden layer!

What you end up with: word embeddings



From Mikolov *et al.* (2013a)



Relationship	Example 1	Example 2	Example 3
France - Paris	Italy: Rome	Japan: Tokyo	Florida: Tallahassee
big - bigger	small: larger	cold: colder	quick: quicker
Miami - Florida	Baltimore: Maryland	Dallas: Texas	Kona: Hawaii
Einstein - scientist	Messi: midfielder	Mozart: violinist	Picasso: painter
Sarkozy - France	Berlusconi: Italy	Merkel: Germany	Koizumi: Japan
copper - Cu	zinc: Zn	gold: Au	uranium: plutonium
Berlusconi - Silvio	Sarkozy: Nicolas	Putin: Medvedev	Obama: Barack
Microsoft - Windows	Google: Android	IBM: Linux	Apple: iPhone
Microsoft - Ballmer	Google: Yahoo	IBM: McNealy	Apple: Jobs
Japan - sushi	Germany: bratwurst	France: tapas	USA: pizza

Relationship pairs in a word embedding. From Mikolov *et al.* (2013b).

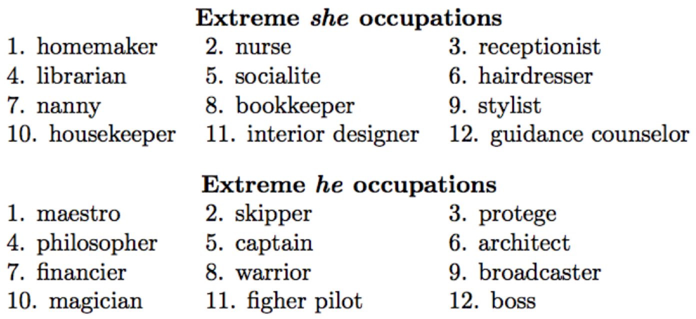
But language can be biased!

- Implicit association tests show that words carry gender bias—e.g. people more often link Female terms with liberal arts and family, while they link male terms with science and careers
- Our implicit biases feeds into training data—e.g. Wager et al found that Wikipedia articles about women more often emphasize their gender, and mention their husbands and husband’s jobs, whereas articles about men do not
- Thus in a word embedding, we might expect “woman” to be closer/more correlated to “writer” than “executive” even though there’s no linguistic reason for this

Bias in word embeddings (Bolukbasi et al)

$$\overrightarrow{\mathrm{man}} - \overrightarrow{\mathrm{woman}} \approx \overrightarrow{\mathrm{computer\ programmer}} - \overrightarrow{\mathrm{homemaker}}.$$

A model trained off of real text data learns and encodes the biases of the world present in that data. (word2vec is trained off of news articles.)



Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

$$\mathit{John}{:}\mathit{computer\ programmer} :: \mathit{Mary}{:}\mathit{homemaker}$$

- Example of bias: word embeddings used to improve search results, to better predict relevancy of results to search criteria
  - Someone searches for “cmu computer science phd student”
  - student websites have their names on them
  - If 2 student websites were otherwise equally likely to be displayed, a biased word embedding could tip the relevance higher for the male phd student’s website and lower for the female’s

Possible outcomes: biased search (Sweeny)

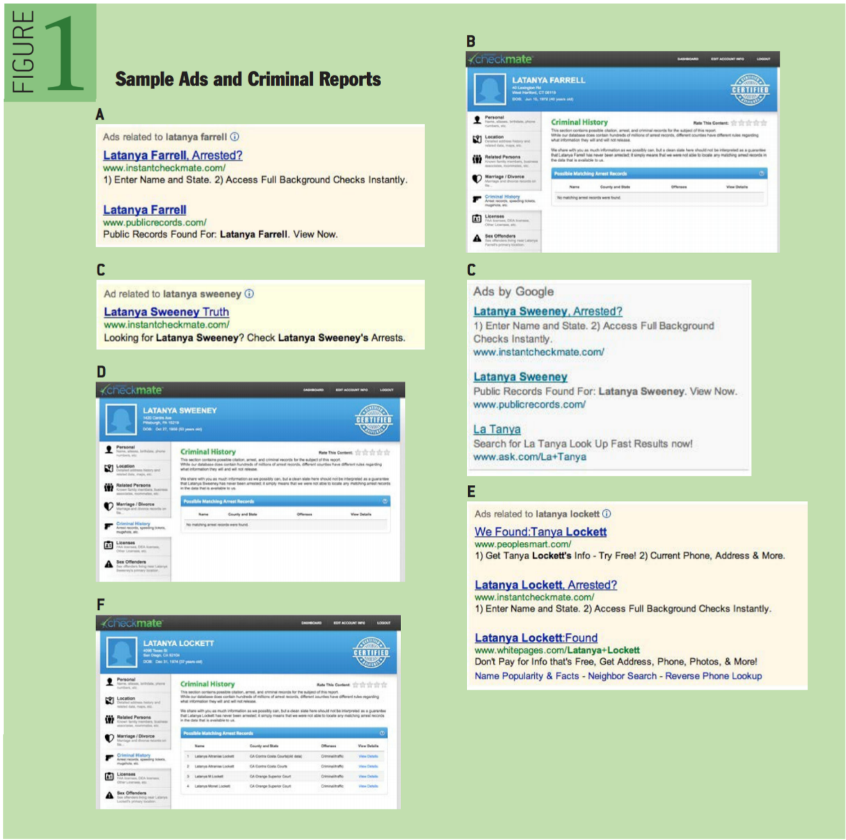


FIGURE 1 Sample Ads and Criminal Reports

A

Ads related to **latanya farrell** ⓘ

Latanya Farrell, Arrested?
www.instantcheckmate.com/
1) Enter Name and State. 2) Access Full Background Checks Instantly.

Latanya Farrell
www.publicrecords.com/
Public Records Found For: **Latanya Farrell**. View Now.

B

C

Ad related to **latanya sweeney** ⓘ

Latanya Sweeney Truth
www.instantcheckmate.com/
Looking for **Latanya Sweeney**? Check **Latanya Sweeney's** Arrests.

C

Ads by Google

Latanya Sweeney, Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

Latanya Sweeney
Public Records Found For: **Latanya Sweeney**. View Now.
www.publicrecords.com/

La Tanya
Search for La Tanya Look Up Fast Results now!
www.ask.com/La+Tanya

D

checkmate
LATANYA SWEENEY
CERTIFIED
Criminal History
Possible Matching Arrest Records
No matching arrest records were found.

E

Ads related to **latanya lockett** ⓘ

We Found:Tanya **Lockett**
www.peoplesmart.com/
1) Get Tanya **Lockett's** Info - Try Free! 2) Current Phone, Address & More.

Latanya Lockett, Arrested?
www.instantcheckmate.com/
1) Enter Name and State. 2) Access Full Background Checks Instantly.

Latanya Lockett:Found
www.whitepages.com/**Latanya+Lockett**
Don't Pay for Info that's Free, Get Address, Phone, Photos, & More!
Name Popularity & Facts - Neighbor Search - Reverse Phone Lookup

F

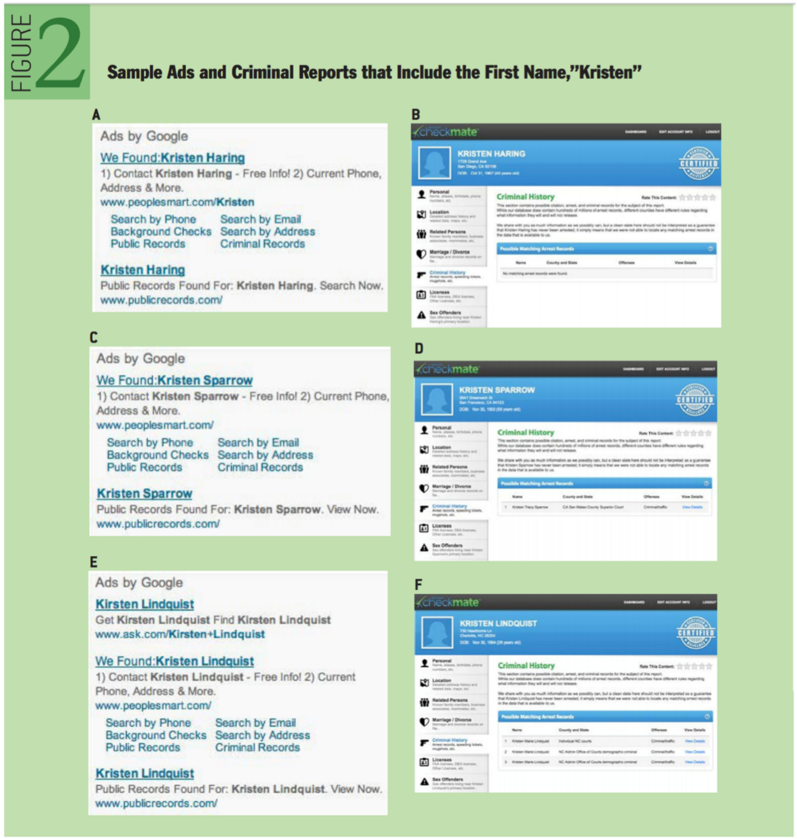


FIGURE 2 Sample Ads and Criminal Reports that Include the First Name,"Kristen"

A

Ads by Google

We Found:**Kristen Haring**
1) Contact **Kristen Haring** - Free Info! 2) Current Phone, Address & More.
www.peoplesmart.com/**Kristen**
Search by Phone Search by Email
Background Checks Search by Address
Public Records Criminal Records

Kristen Haring
Public Records Found For: **Kristen Haring**. Search Now.
www.publicrecords.com/

B

checkmate
KRISTEN HARING
CERTIFIED
Criminal History
Possible Matching Arrest Records
No matching arrest records were found.

C

Ads by Google

We Found:**Kristen Sparrow**
1) Contact **Kristen Sparrow** - Free Info! 2) Current Phone, Address & More.
www.peoplesmart.com/
Search by Phone Search by Email
Background Checks Search by Address
Public Records Criminal Records

Kristen Sparrow
Public Records Found For: **Kristen Sparrow**. View Now.
www.publicrecords.com/

D

E

Ads by Google

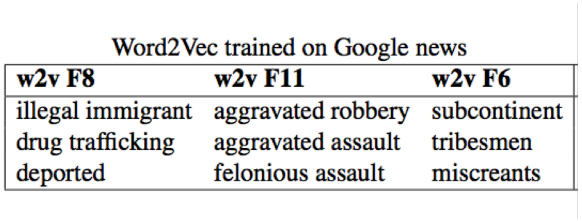
Kirsten Lindquist
Get **Kirsten Lindquist** Find **Kirsten Lindquist**
www.ask.com/**Kirsten+Lindquist**

We Found:**Kristen Lindquist**
1) Contact **Kristen Lindquist** - Free Info! 2) Current Phone, Address & More.
www.peoplesmart.com/
Search by Phone Search by Email
Background Checks Search by Address
Public Records Criminal Records

Kristen Lindquist
Public Records Found For: **Kristen Lindquist**. View Now.
www.publicrecords.com/

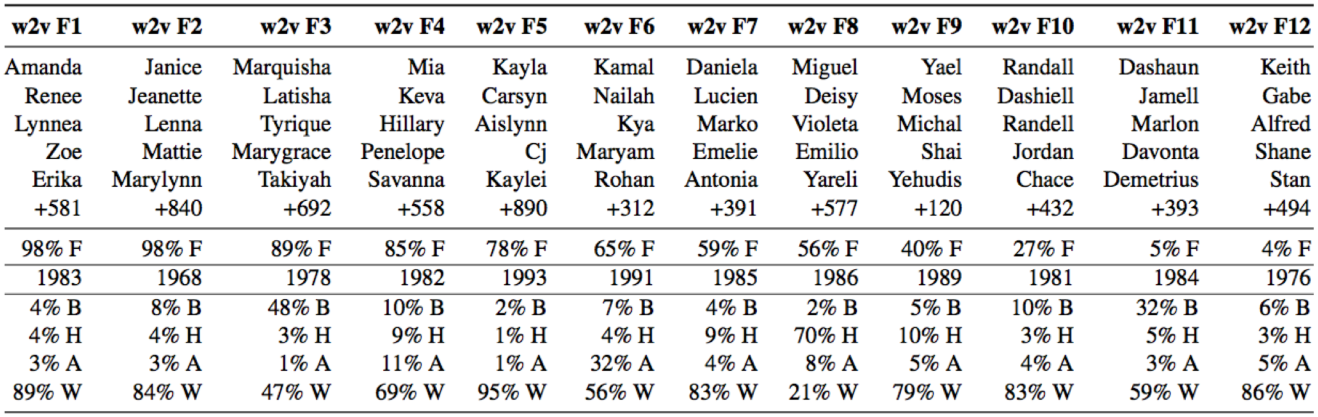
F

Possible outcomes: search results (Arteaga)



Word2Vec trained on Google news

w2v F8	w2v F11	w2v F6
illegal immigrant	aggravated robbery	subcontinent
drug trafficking	aggravated assault	tribesmen
deported	felonious assault	miscreants



w2v F1	w2v F2	w2v F3	w2v F4	w2v F5	w2v F6	w2v F7	w2v F8	w2v F9	w2v F10	w2v F11	w2v F12
Amanda	Janice	Marquisha	Mia	Kayla	Kamal	Daniela	Miguel	Yael	Randall	Dashaun	Keith
Renee	Jeanette	Latisha	Keva	Carsyn	Nailah	Lucien	Deisy	Moses	Dashiell	Jamell	Gabe
Lynnea	Lenna	Tyrique	Hillary	Aislynn	Kya	Marko	Violeta	Michal	Randell	Marlon	Alfred
Zoe	Mattie	Marygrace	Penelope	Cj	Maryam	Emelie	Emilio	Shai	Jordan	Davonta	Shane
Erika	Marylynn	Takiyah	Savanna	Kaylei	Rohan	Antonia	Yareli	Yehudis	Chace	Demetrius	Stan
+581	+840	+692	+558	+890	+312	+391	+577	+120	+432	+393	+494
98% F	98% F	89% F	85% F	78% F	65% F	59% F	56% F	40% F	27% F	5% F	4% F
1983	1968	1978	1982	1993	1991	1985	1986	1989	1981	1984	1976
4% B	8% B	48% B	10% B	2% B	7% B	4% B	2% B	5% B	10% B	32% B	6% B
4% H	4% H	3% H	9% H	1% H	4% H	9% H	70% H	10% H	3% H	5% H	3% H
3% A	3% A	1% A	11% A	1% A	32% A	4% A	8% A	5% A	4% A	3% A	5% A
89% W	84% W	47% W	69% W	95% W	56% W	83% W	21% W	79% W	83% W	59% W	86% W

Table 3: Illustrative first names (greedily chosen) for $n = 12$ groups on the w2v embedding. Demographic statistics (computed a posteriori) are also shown though were not used in generation, including percentage female (at birth), mean year of birth, and percentage Black, Hispanic, Asian/Pacific Islander, and White.

- We note that Sweeny's work documenting biased search results was published the same year as Mikolov's word2vec paper, so it's improbable that word embeddings were to blame for that observation in particular—however it is an example of behavior that may arise from such biased word embeddings

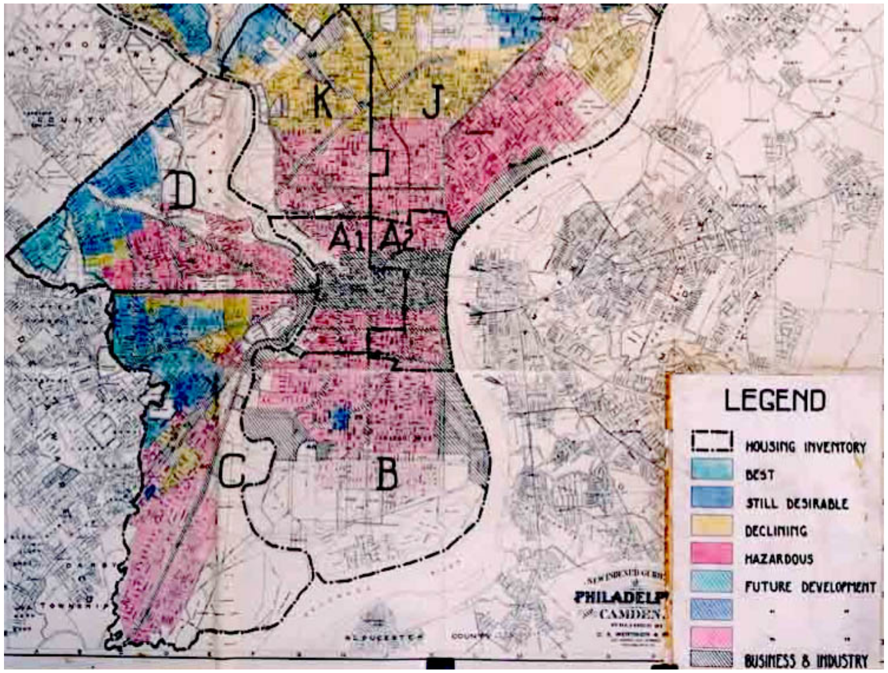
How do we make a fair model?

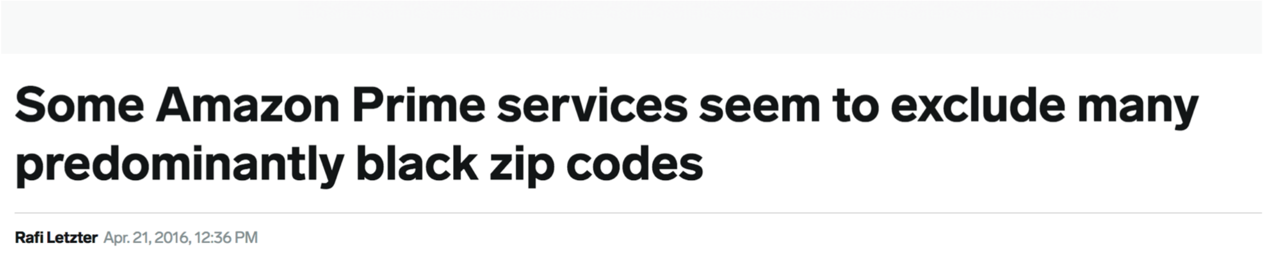
What do you all think?

- Try to ensure your model doesn’t augment bias
- Train with a balanced dataset
- Audit your model
- Ensure some constraint, e.g. demographic parity
- Don’t use protected attribute

What happens if we take out the protected attribute?

- Neighborhoods in America are largely racially segregated
- A race-blind model could still act in a discriminatory manner by using zipcode to e.g. deny a loan
- Even unintentional discrimination can occur in this way, given a biased prior





Some Amazon Prime services seem to exclude many predominantly black zip codes

Rafi Letzter Apr. 21, 2016, 12:36 PM

More examples of proxy variables

- Purchasing history for medical conditions (pregnancy, or a disease)
- Friends on social media sites to determine sexual orientation
- Facebook currently using in the HUD case: “affinity groups” i.e. your likes on facebook

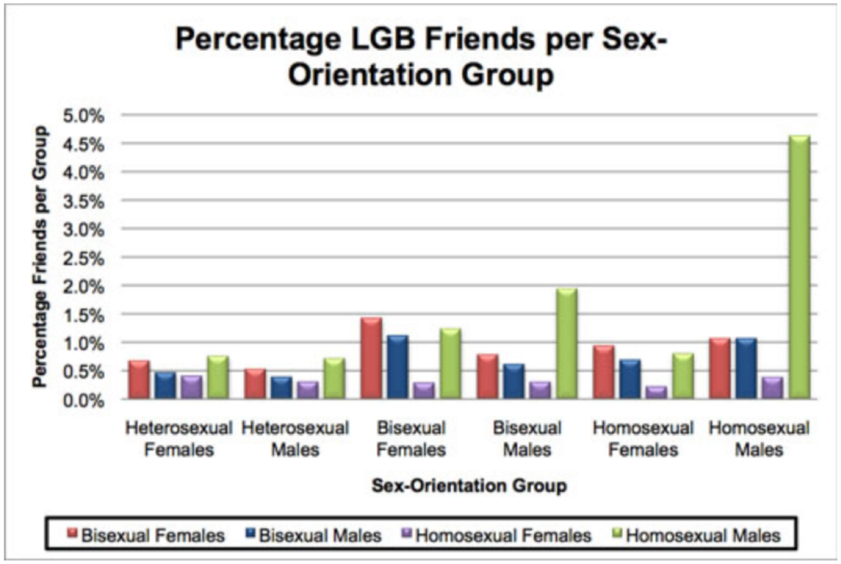
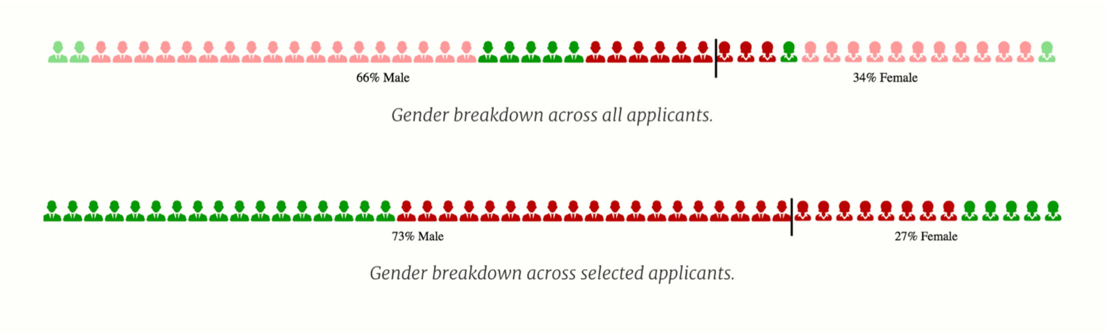


Figure 4: Percentage of LGB friends per sex orientation group.

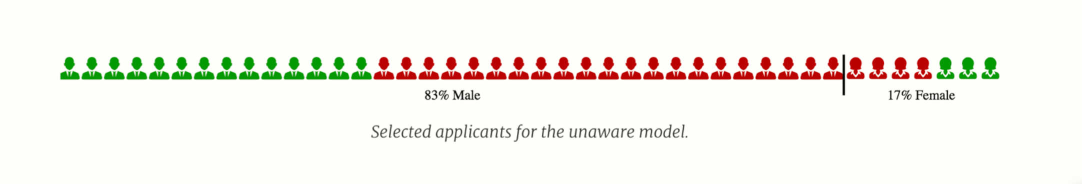
In fact, taking out the protected attribute can be detrimental to fairness goals

- Imagine an AI for hiring new employees has two features: gender and experience.
    - The model hires 27% women, despite their being 44% of the applicant pool.
- In an effort to make a fair model, you take out the gender variable and only use experience
    - You find your model now hires 17% women.



Gender breakdown across all applicants.

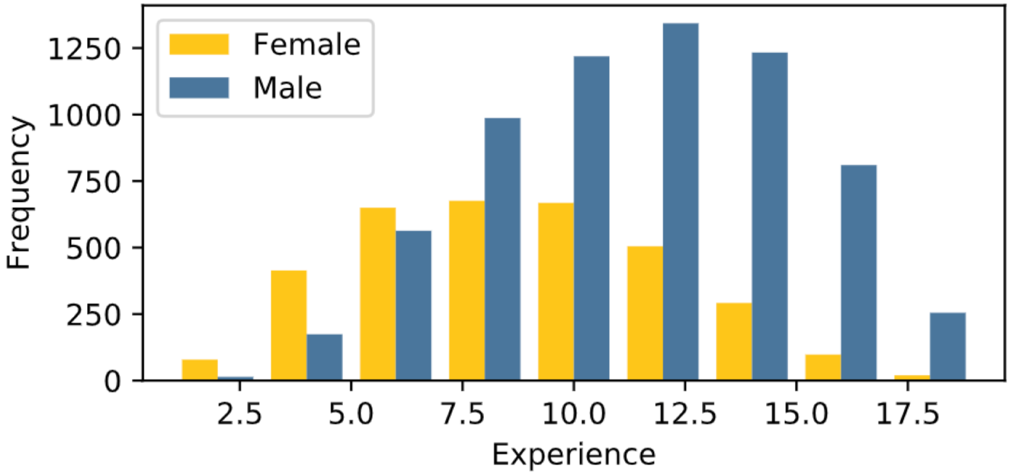
Gender breakdown across selected applicants.



Selected applicants for the unaware model.

In fact, taking out the protected attribute can be detrimental to fairness goals

- Perhaps in reality, people with over ten years of experience are equally qualified for the job
- Women have to take off more time due to extenuating circumstances (needing to take family or child leave, etc)
- Removing the gender feature from the model makes it impossible for the model to compensate



Histogram of male and female experience.

However, there are situations where we can’t use the protected attribute, legally

- Some legal situations prevent disparate treatment, i.e. treating people differently based on some sensitive attribute, whether it be positive or negative
  - E.g. it could be that some states would not allow the hiring classifier from previous slides because it treats men and women differently
- E.g. college admissions in Texas: race not allowed to factor into school admissions
  - Workaround: top 10% rule, top 10% of high school student automatically admitted to state colleges
- So, there may be some situations where we want to get rid of sensitive attribute information and all proxy information

Add fairness constraints

- Demographic Parity: proportion of people who get good outcome/bad outcome should be equal across all groups
- Equal False Positive/False Negative Rates (all confusion matrix scores)
- Equalized Odds: The protected attribute and the prediction are conditionally independent given the ground truth: i.e., the rates of loan application acceptances should be the same across groups among people who are truly credit-worthy
- Individual fairness constraint: similar people should be treated similarly

Problems with fairness constraints

- They don’t always lead to the fair outcomes you think they should either!
    - See Measure and Mismeasure of Fairness (Corbett-Davies and Goel) and Delayed Impact of Fair ML (Liu and Hardt)

Further Thoughts: Individuals vs Groups

- We can thinking about fairness in aggregate or individually
- Group fairness: ideas like demographic parity, equalized odds: statistics for all groups should be the same
  - But this doesn't solve all problems
    - What about intersectionality? You could accept the same number of black people and white people to college, but accept no black women
    - Increase disparities within a subgroup: e.g. make it easier for wealthy or otherwise privileged black people to get into college, but make it just as hard or harder for low-income students of color
- Individual fairness: similar people should be treated similarly
  - What does it mean for two people to be similar?

Further Thoughts

- ML systems evolve the system that they are deployed in, but ML algorithms do not take this shift into account
- PredPol/ ACLU arguments against its use: sending policemen to already overpoliced areas could further perpetuate the cycle of disproportionate incarceration in America
- But similarly, careless “fair” algorithms could lead to their own problems
    - Consider a ”fair” lending algorithm that lent to the same number of people from groups A and B, where B is disadvantaged. If those in group B are not actually qualified for a loan and default, you actually hurt that population *more,* and also prevent their being qualified in future because they defaulted

Further Thoughts: Delayed Impact of Fairness

- Liu and Hardt paper, Delayed Impact of Fair Machine Learning
- How can we make fair algorithms that take into account the way they change the data landscape over time?
- What if instead of applying some blindness constraint, or demographic parity constraint, to an algorithm, we instead directly optimize for improving the lives of the affected group over time?