Ethical Machine Learning

...

Taking “Don’t be Evil” Literally

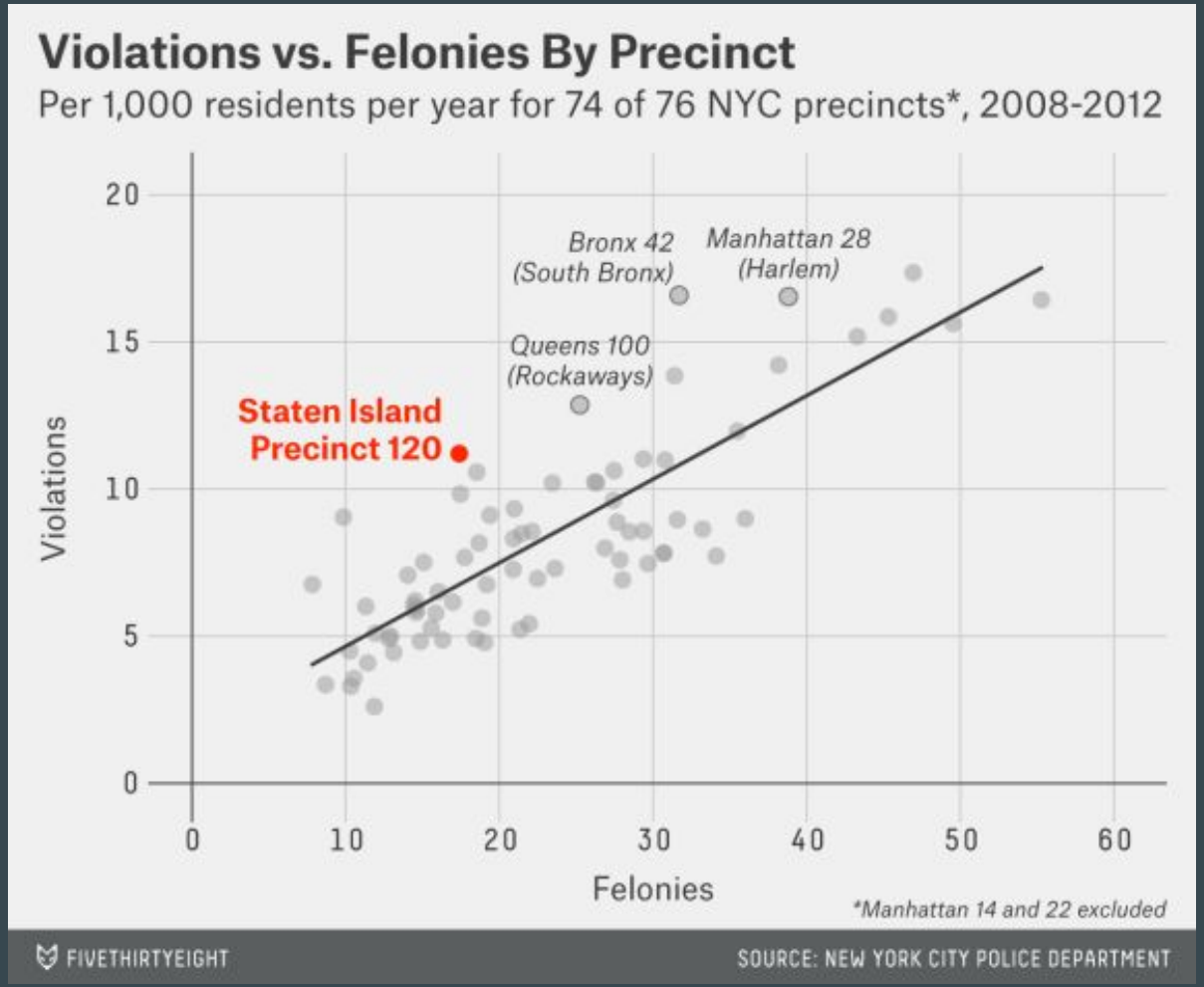
Katharine Jarmul
Kjamistan.com

#QCONSP

I Can’t Breathe: The Killing of Eric Garner

Joe Raedle/Getty Images

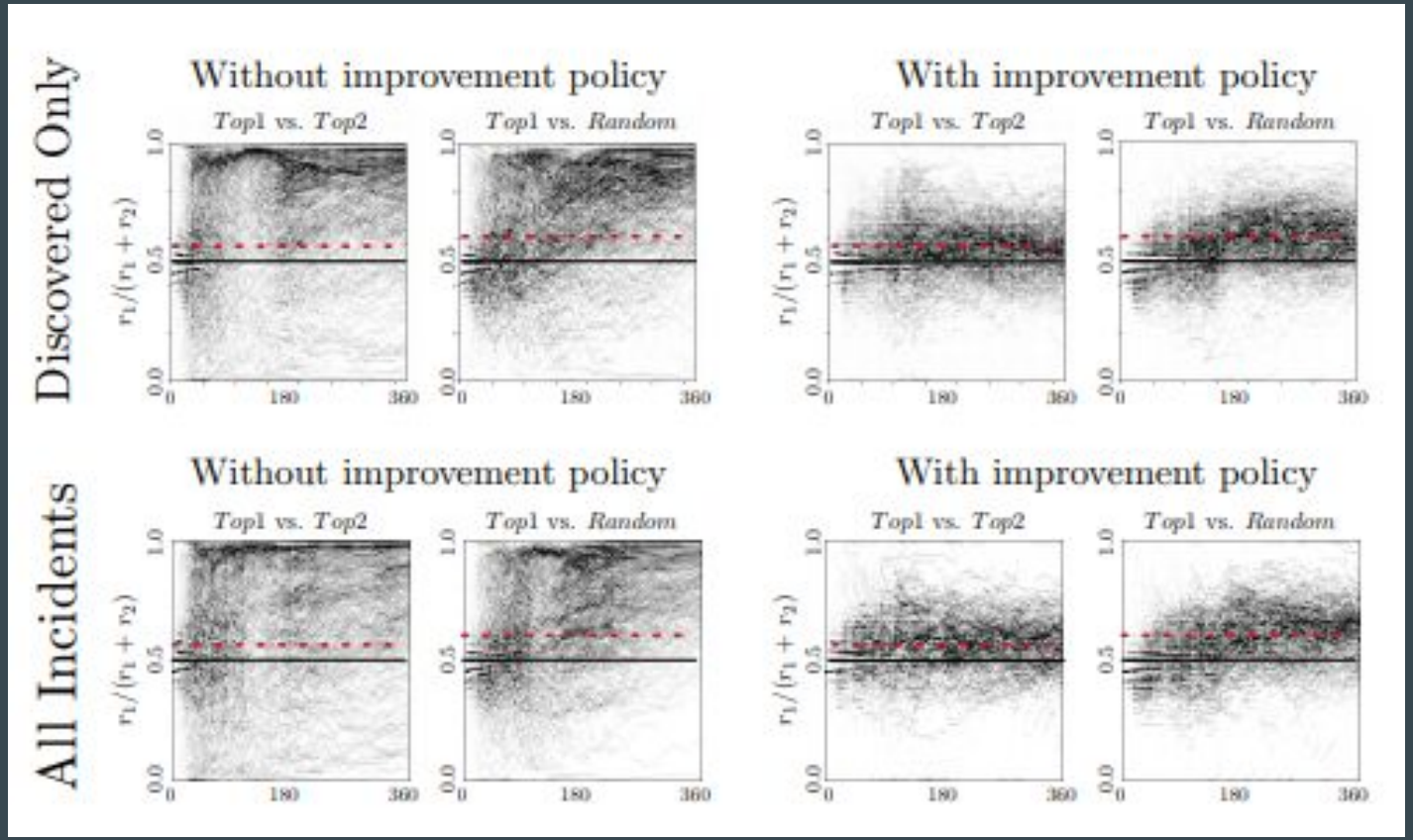
“Broken Windows” Policing



Disparate Impact

$$\frac{\Pr(C = \text{YES} | X = 0)}{\Pr(C = \text{YES} | X = 1)} \leq \tau = 0.8$$

Predictive Policing: Runaway Feedback Loops



Ensign et al, 2017

If our models mimic current police behavior, are we creating a valid model?

If our models mimic social inequalities and prejudice, are we creating a valid model?

Are social inequalities and prejudice valid?

Breaking the Cycle:
Determining if Your Data has Prejudice

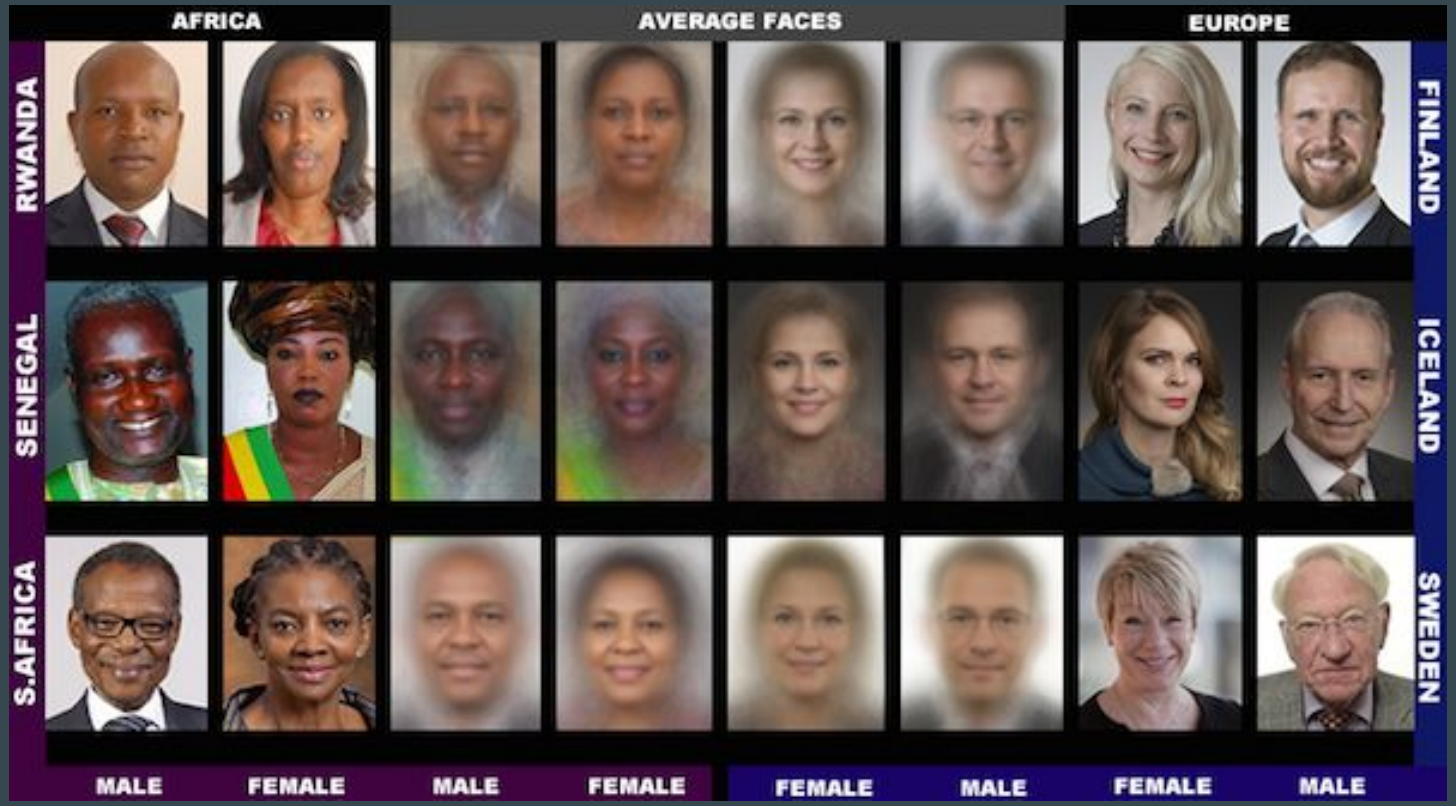
FairTest: Evaluating Correlations to Sensitive Attributes

```
Sub-Population 2 of size 6791
Context = {'education-num': (11.5, inf)}

+------+---------+---------+----------+
|income|   Female|     Male|     Total|
+------+---------+---------+----------+
|<=50K |1594(76%)|2156(46%)|3750 (55%)|
|>50K  | 492(24%)|2549(54%)|3041 (45%)|
|Total |2086(31%)|4705(69%)|6791(100%)|
+------+---------+---------+----------+

p-value = 2.72e-124 ; NMI = [0.0508, 0.0876]
```

GenderShades: Creating Better Datasets



GenderShades.org

NLP: Looking at Word Vector Correlations

```
1  model.most_similar(['brazilian', 'woman'], topn=20)
```

```
[('womans', 0.5730605125427246),
 ('man', 0.5635831952095032),
 ('girl', 0.5441568493843079),
 ('rihanna', 0.5356341004371643),
 ('Giuseppina_Pasqualino_di_Marineo', 0.5330727100372314),
 ('latina', 0.5226492285728455),
 ('teenage_girl', 0.519279420375824),
 ('micheal_jackson', 0.5143763422966003),
```


PARENTAL
ADVISORY
EXPLICIT CONTENT

NLP: Google News Vectors

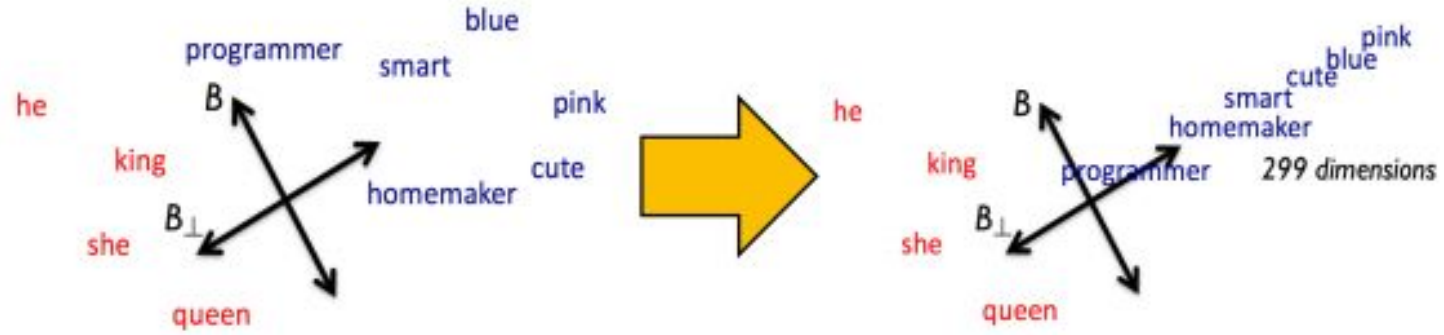
```
1  model.most_similar(['brazilian', 'girl'], topn=20)
```

```
[('boy', 0.6321439146995544),
 ('micheal_jackson', 0.5712836980819702),
 ('rihanna', 0.5660167932510376),
 ('teenage_girl', 0.5542945265769958),
 ('selena', 0.543932557106018),
 ('shaved_pussy', 0.5382047891616821),
 ('jessica_alba', 0.537001132965087),
```

https://blog.kjamistan.com/embedded-isms-in-vector-based-natural-language-processing/

Debiasing Word Vectors

- Hard debiasing:
  - 1) **Neutralize**, project away the gender subspace from neural words

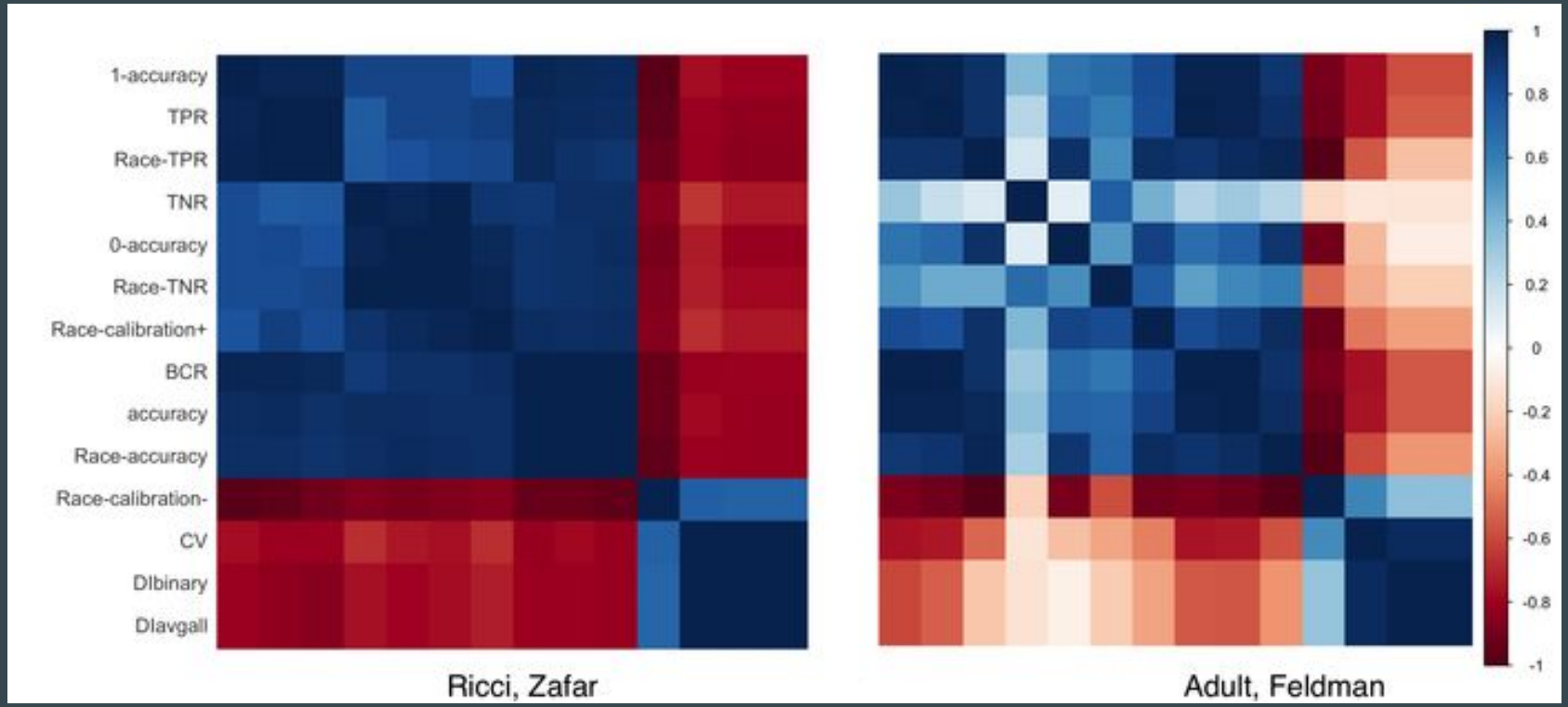


https://github.com/tolga-b/debiaswe

(Bolukbasi, Chang, Zou, Saligrama and Kalai, 2016)

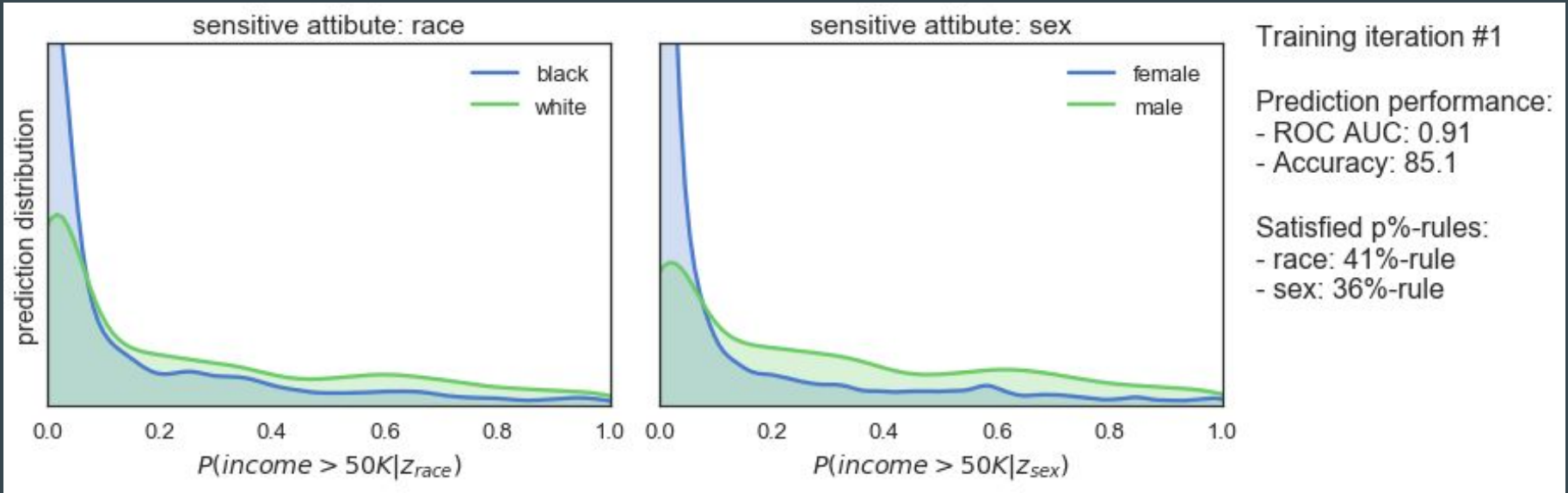
Modeling Fairness:
Evaluating Models for Prejudice

Defining Fair



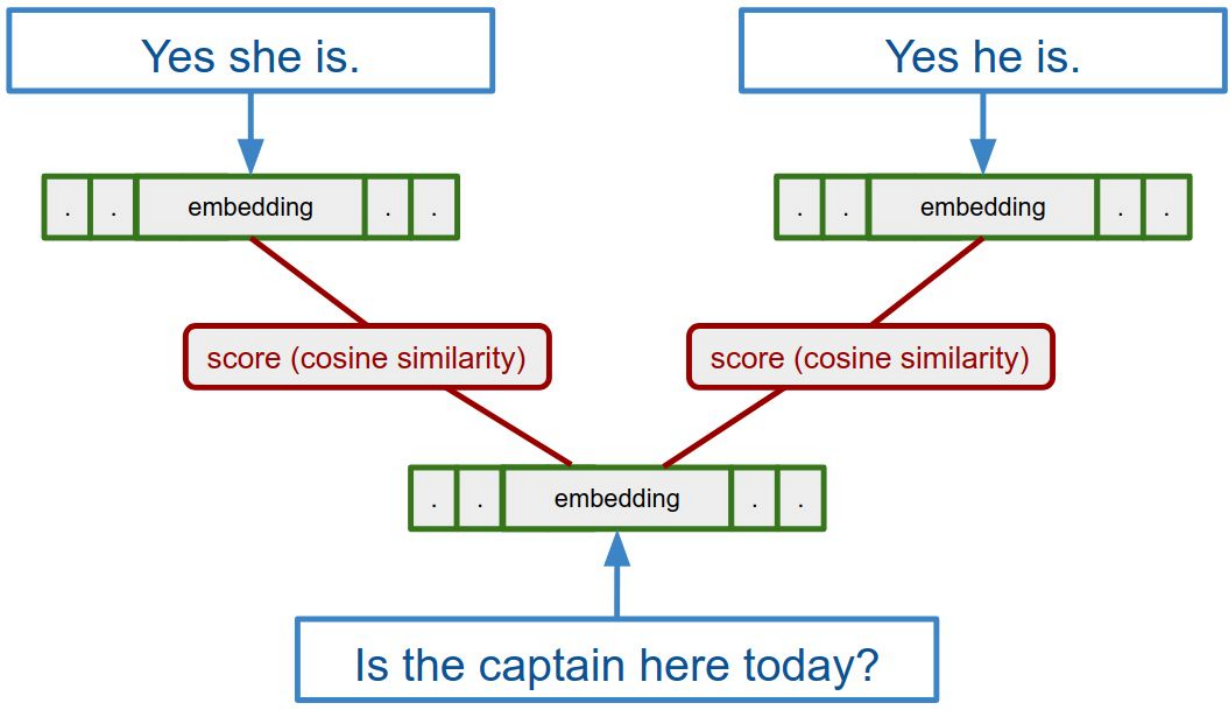
https://algorithmicfairness.wordpress.com/

Evaluating Fair



https://blog.godatadriven.com/fairness-in-ml/

NLP: Testing Bias



	score
Yes she is	0.17
Yes he is	0.27
Bias score (difference)	-0.10

https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html

Interpreting Our Models

Figure 5. Examples of mistakes where we can use attention to gain intuition into what the model saw.

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention
Xu et al., 2016

Radical Transparency:
Promoting Conversation & Accountability

Talking Fair

https://www.fatml.org/

Acting Fair: Building Accountable Applications

https://2017.ind.ie/ethical-design/

Ethical Machine Learning:
Taking a Logical Stance against Oppression

Ethical ML Takeaways

- Doing “nothing” assumes prejudice and unfair treatment is a valid action
- We need better data
  - Diverse data which better reflects the real world
  - Stop using datasets which are non-representative
- We need built-in ethics-driven evaluation criteria
  - Scikit-learn disparate impact?
  - Scikit-learn equal odds / opportunity?
- You can contribute
  - open-source your work and datasets
  - volunteer with the Algorithmic Justice League or local organization

Thanks!

Questions?

- Now?
- Later?
  - @kjam
  - katharine@kiprotect.com