Bias, Fairness, and Beyond

Kenneth (Kenny) Joseph

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

Grading

Grading

Components

- Weekly quizzes using UBLearns (12, 1.5% each) – 15%; most quizzes will be multiple choice. Quizzes are released on Tuesday at 12:00am and due on Monday at 11:59PM the following week. Weekly Quiz 0 is a separate quiz, worth 1% of your grade.
- Programming Assignments (4, 10% each) – 40%; all assignments will be assigned and submitted as jupyter notebooks.
- Annotation Assignment (1) - 10%; manual annotation of data and an assessment of those manual annotations.
- Mid-term Exam (open book/notes) – 15%
- Final Exam (open book/notes) – 19%
- **Programming assignment grades are subject to a 75% multiplier based on an end-of-the-semester peer review process. Thus, if for example, your group scores 100% on all assignments, but your teammates rate you as doing 0% of the work, you will receive 10 out of the 40 points.s**

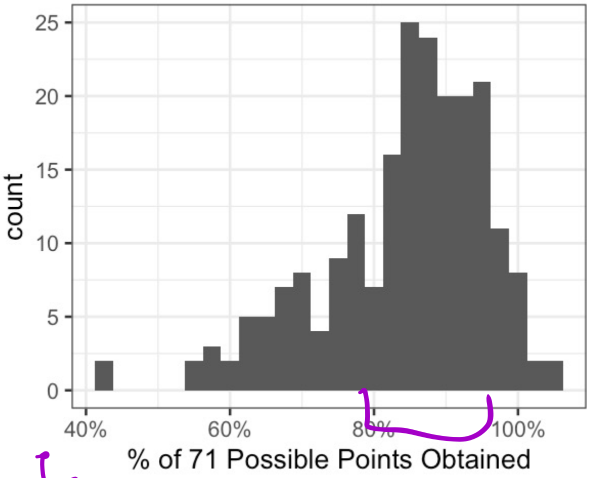
Change in Peer Review Req.

- It is punitive
- Groups have seemed to come throughout the semester with issues
- If you also have a group issue, please email me some time before next Friday for a discussion

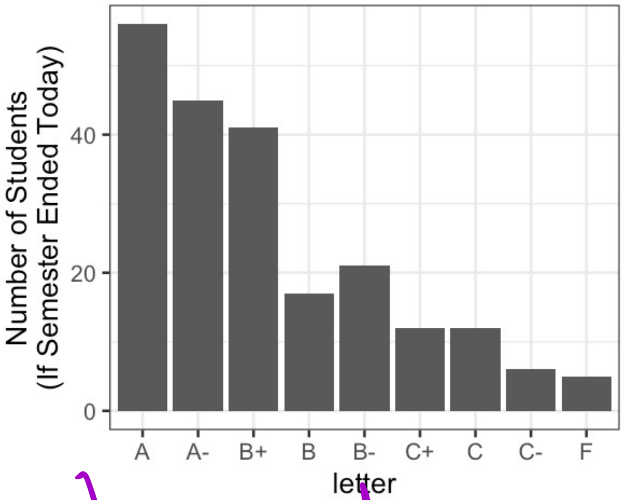
Finals Format

- Ish:
  - ~15 MC:
    - Lose points for guessing
  - ~4 Short Answer (think the midterm)
  - Grade will be roughly out of 12 MC, 3 short answer (with max amount of bonus points)

Grades



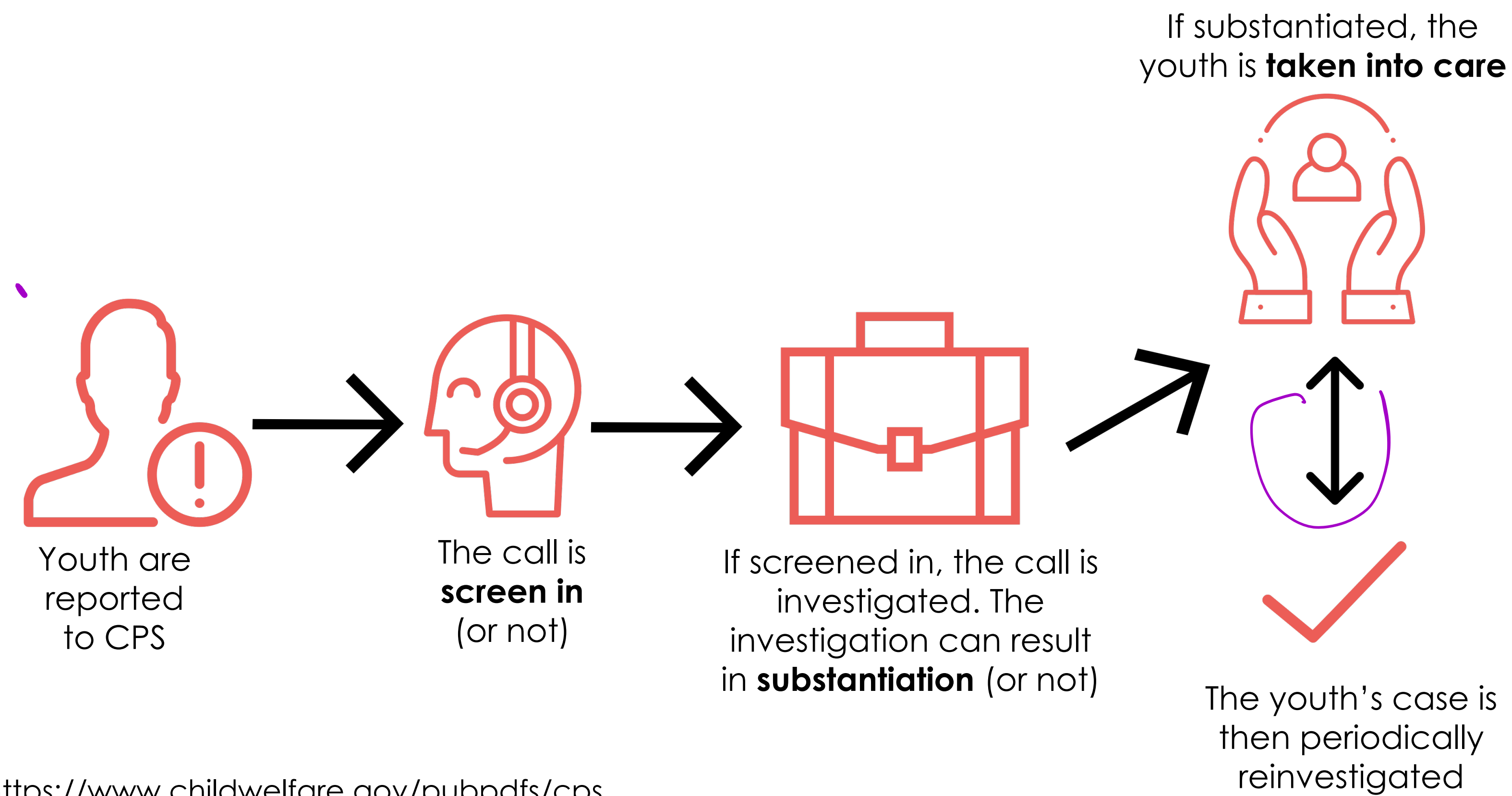
count
25
20
15
10
5
0
40%
60%
80%
100%
% of 71 Possible Points Obtained



Number of Students
(If Semester Ended Today)
40
20
0
A
A-
B+
B
B-
C+
C
C-
F
letter

Announcements

- Quiz 12 due next Wednesday night
- PA5 due 5/12
- My OH will not be on Thursday next week (likely Tues)
- Next Tuesday:
  - Miscellaneous
  - If you want specific topics reviewed, let me know by tomorrow
  - Likely end a bit early for impromptu office hours

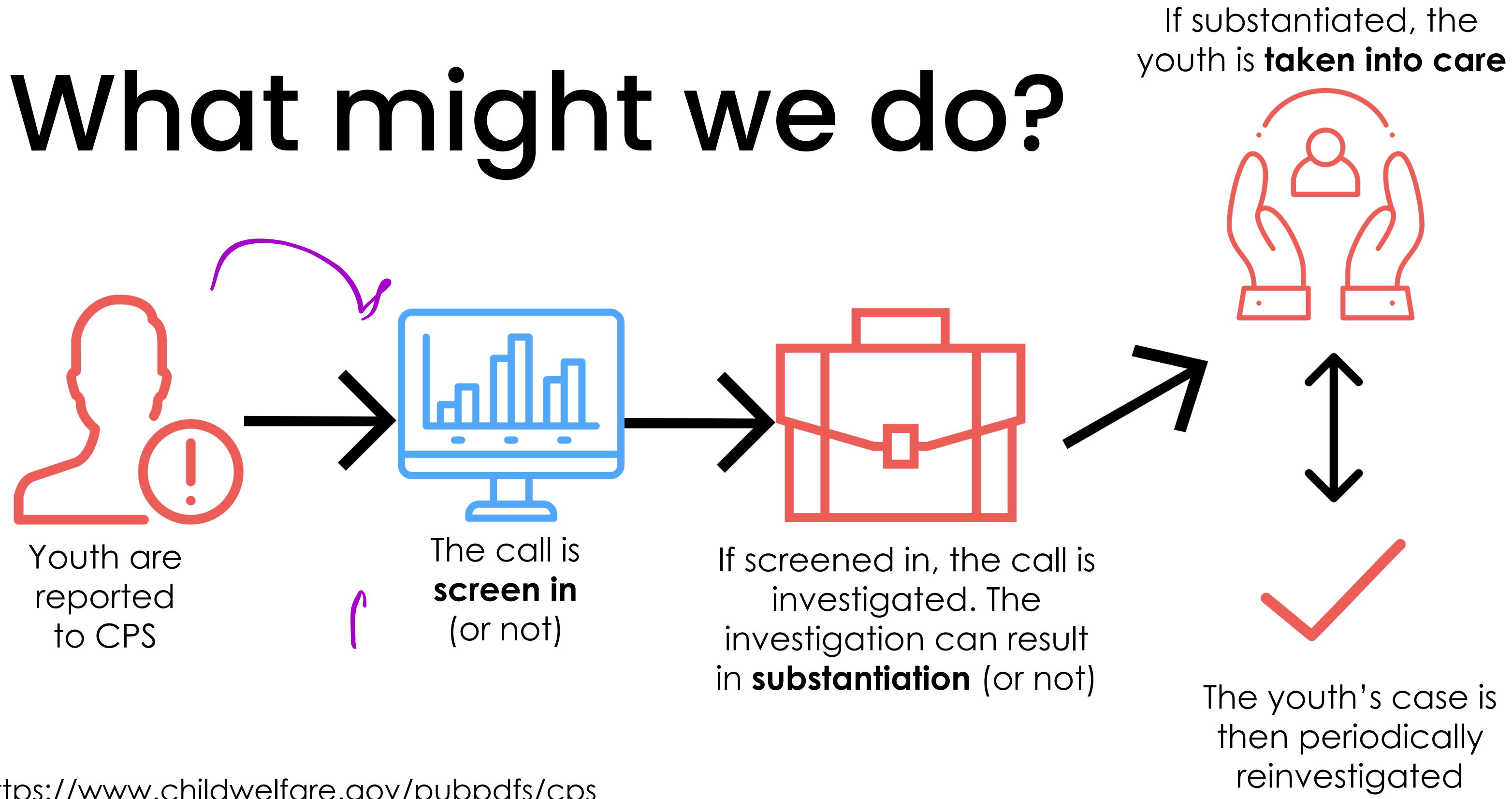


https://www.childwelfare.gov/pubpdfs/cpswork.pdf

Summary thus far

- No one wants to be in the child welfare system
  - Experts agree that the goal should be to get people back with their families
  - People involved suffer
  - Life outcomes for people who stay in it are terrible
- Black people are over-represented in the child welfare system

What might we do?



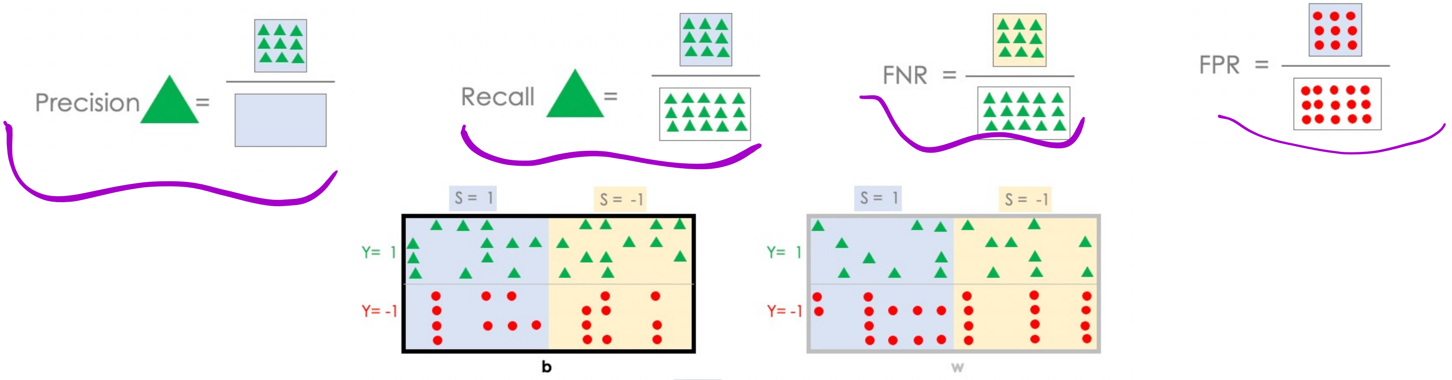
https://www.childwelfare.gov/pubpdfs/cpswork.pdf

In this case

- Proxy:
  - Use outcomes from **substantiation phase**, not **screening phase**
  - Idea: Probably more accurate, less biased
- What is our target variable in this case?

Exercise

- Come up with a definition of fairness that uses these different rates we have discussed.



Three popular definitions

Equal FPR

We say a classifier fair with respect to FPR if

$$FPR_b = FPR_w.$$

In the COMPAS context, a classifier is fair with respect to FPR if chances of a black and white defendants begin identified as reoffending when they actually did not end up reoffending are the same. This is one of the notions of fairness that ProPublica used.

Equal FNR

We say a classifier fair with respect to FNR if

$$FNR_b = FNR_w.$$

In the COMPAS context, a classifier is fair with respect to FNR if chances of a black and white defendants begin identified as not reoffending when they actually did end up reoffending are the same. This is one of the notions of fairness that ProPublica used.

Well-calibrated

We say a classifier if well-calibrated if

$$PPV_b = PPV_w.$$

In the COMPAS context, a classifier is fair (or does not have any statistical bias) if the chances of a black and white defendant being correctly identified as reoffending given that the classifier identified them as such are the same. This is the notion of fairness used in the rejoinder to the ProPublica article.

More real-world considerations

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

- Which approach would you prefer, and why?
- Do you think demographic parity is fair? Why/why not?
- Do you think the equal opportunity approach is fair? Why (not)?

Summary points – Measures of fairness

- For probabilistic/threshold classifiers, you can actually tune your threshold to achieve different fairness goals
- Different measures of fairness suggest different solutions
- There is no one correct measure of fairness
  - And indeed, it can be proven that you cannot optimize for all of them at the same time (More on that in 440/540)

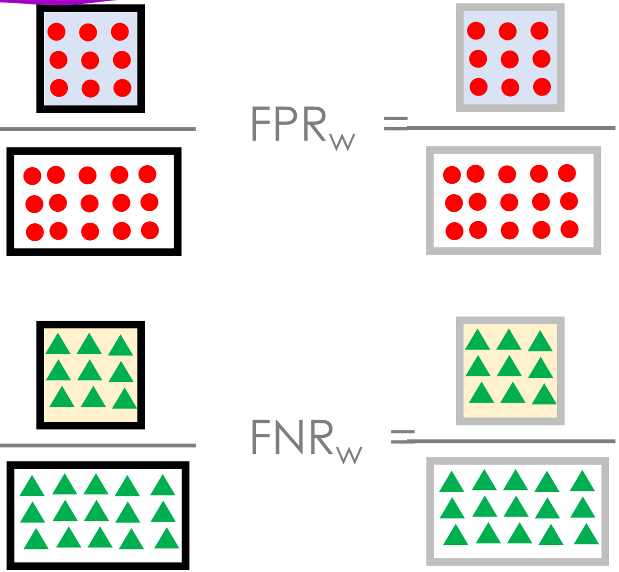
FPR and FNR for groups

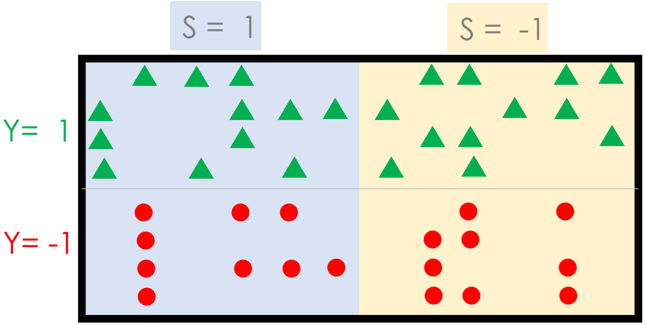
Calculate the rates

Answers: http://www-student.cse.buffalo.edu/~atri/ml-and-soc/support/notes/fairness/index.html

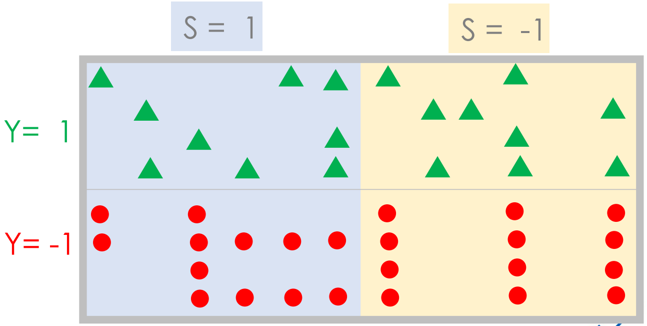
$FPR_b =$ $FPR_w =$

$FNR_b =$ $FNR_w =$





b

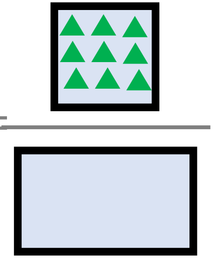


PPV for groups

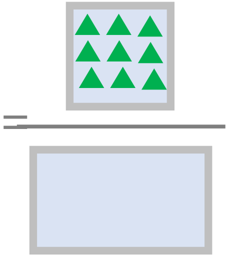
Calculate the values

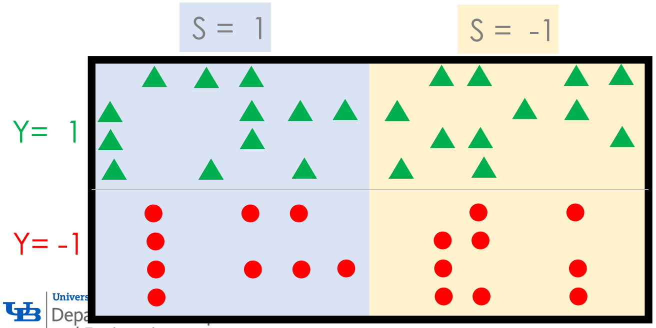
Answers: http://www-student.cse.buffalo.edu/~atri/ml-and-soc/support/notes/fairness/index.html

PPV_b =

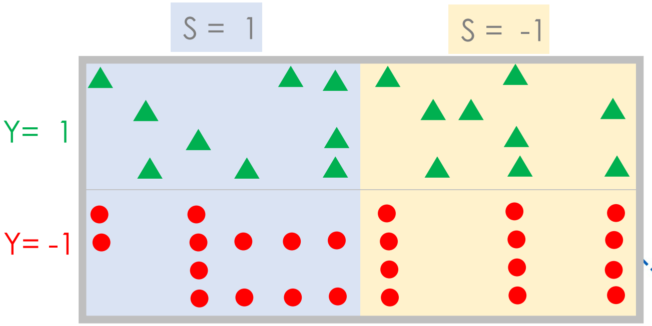


PPV_w =





b



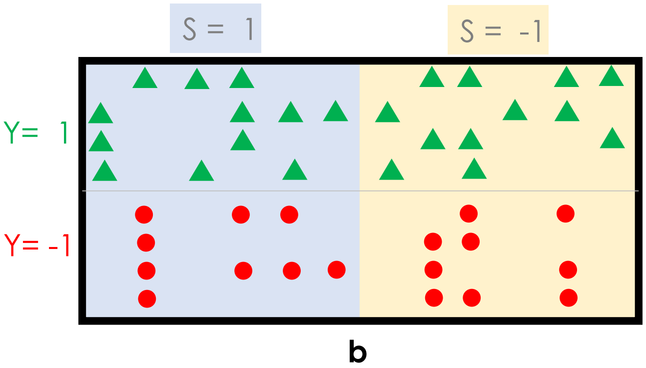
w

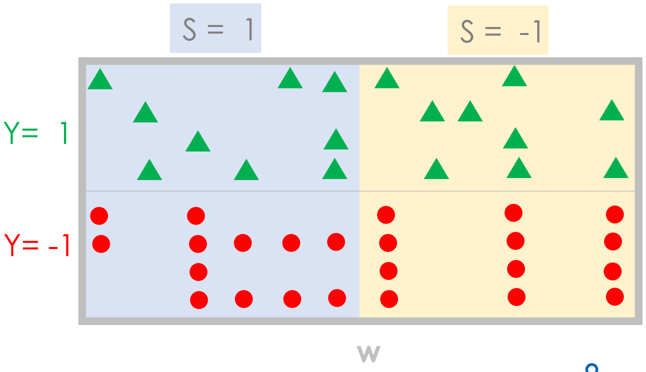
Exercise!

Answers: http://www-student.cse.buffalo.edu/~atri/ml-and-soc/support/notes/fairness/index.html

Exercise

For each notion of being fair with respect to FPR, FNR and well-calibrated, decide if it holds for the following instance (that we have seen before):





Back to child welfare

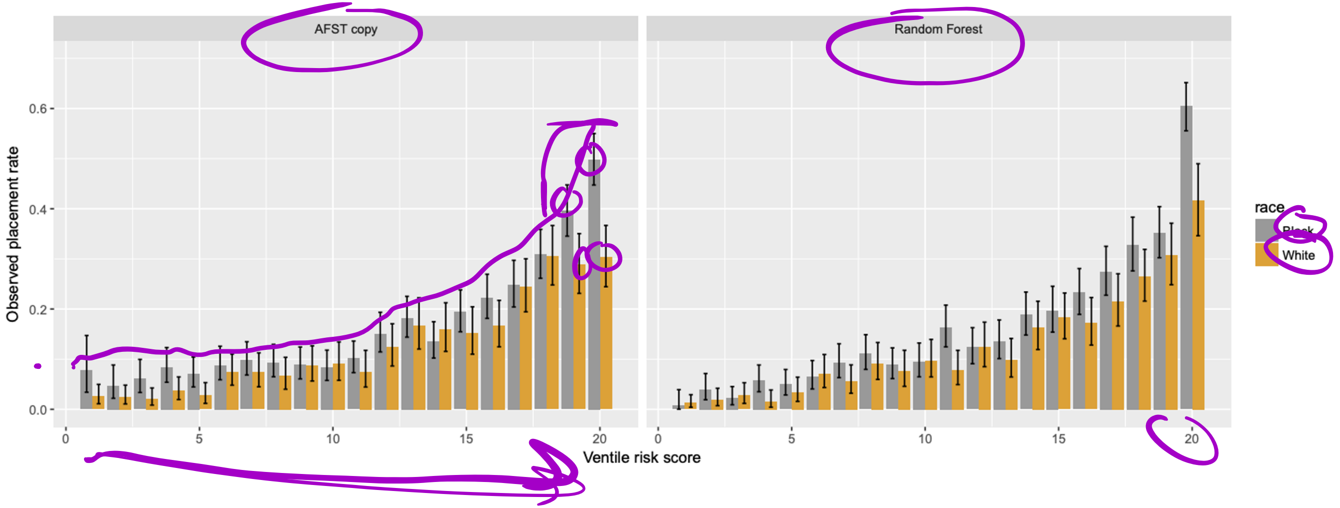


Figure 4: Observed placement rates by AFST model (left) and Random forest model (right) risk score ventile broken down by victim's race. Error bars correspond to 95% confidence intervals.

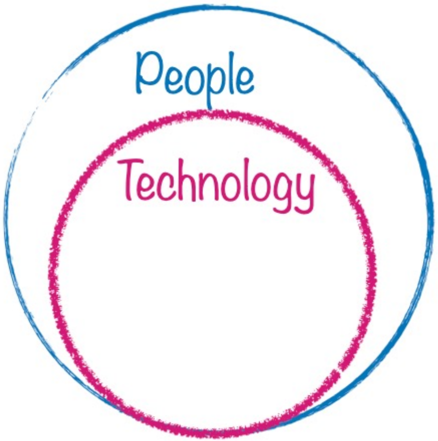
From fairness to justice

- Let's assume that we can make this classifier "fair" now.
- Does that solve all the issues with child welfare?
- Does that absolve us of any responsibility?

Why (were there racial disparities in the algorithm's risk scores)?

A: Because the **algorithm** developers chose a proxy variable that had racial bias.

Why was the underlying proxy variable racially biased?



A: Because the **people** who made the decisions that informed the proxy variable did so in a racially biased way

Two really good reasons not to stop here.

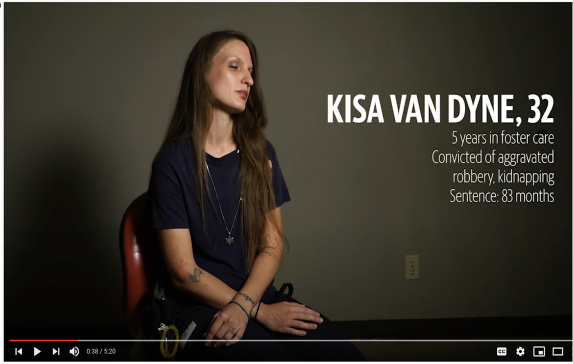
1. Not all (many?) case workers hold explicit racial biases
2. Individuals can **perform actions that increase racial inequality without “being racist”**
   (so can technology…)

Ida Bae Wells
@nhannahjones

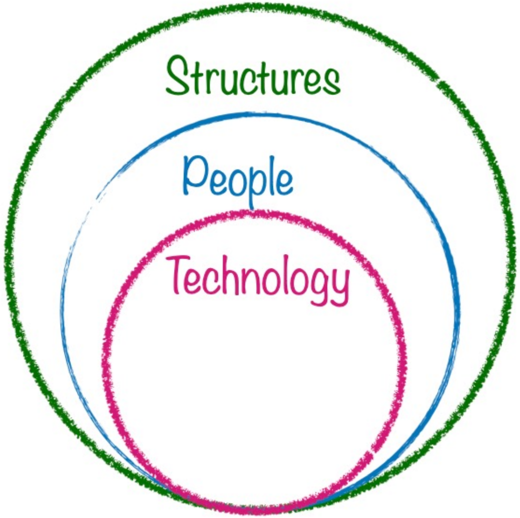
Replying to @nhannahjones and @davidwblight1

So, for people like myself, white guilt is useless, and denying that white people can struggle financially or have to work hard is useless and inaccurate. The entire point of the work we do is to point out the way the systems work whether individuals are racist or not.

10:43 AM · Nov 1, 2021 · Twitter for iPhone



https://twitter.com/nhannahjones/status/1455184081793306627



Why did the people make racially biased decisions?

A: Because case workers are bad, racist people

A: Because racial biases and discrimination are built into our **society** at a structural level

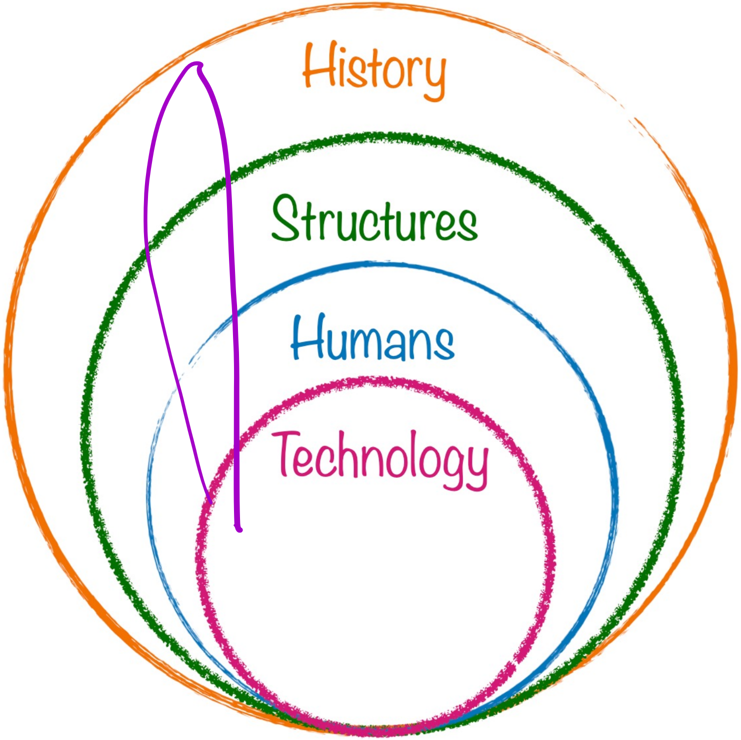
Becoming Wards of the State: Race, Crime, and Childhood in the Struggle for Foster Care Integration, 1920s to 1960s

Michaela Christy Simmons

But at the same moment that minorities were included in positions of power in the courts, African American children disproportionately entered the system as delinquents rather than as dependent or neglected (Trost 2005; Ward 2012).

"Wiltwyck School did not accept the boy . . . [Brace Farms] cannot accept Lonnie[1] for placement . . . Berkshire Ind[ustrial] Farm rejected Lonnie. . . . Should Children's Village reject the application on Lonnie, the only alternative left, regrettably as it may seem, is to send this boy to the N.Y. State Training School at Warwick [for delinquents]." ~1944 court action for a 13-year-old neglected African American boy (Polier Manuscripts 1944a)

https://journals.sagepub.com/doi/full/10.1177/0003122420911062



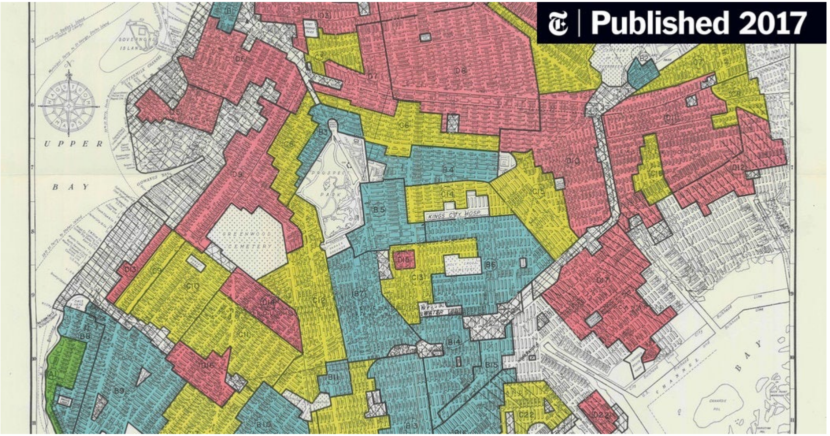
Why are racial bias and discrimination built into our society at a structural level?

A: Because American society has **historically** favored White Americans, relative to Black Americans, and structural racism is reproduced over time

Why are Black youth over-represented?

Two possible reasons

1. **Need/Risk (Black parents have less money to support children)**
2. Discrimination/Bias (Black families are over-policed within Child Welfare)


Published 2017
UPPER
BAY


AMERICAN SLAVERY
339 years
SEGREGATION
89 years
1700
1800
1900
2000
1526
1865
1954

Abdurahman, J. K. (2021). Calculating the Souls of Black Folk: Predictive Analytics in the New York City Administration for Children's Services. *Columbia Journal of Race and Law*, *11*(4), 75-110.

“we witness how a model labeled “prevention services” actually functions to extend the scope of the carceral state. In the literal sense, preventing family separations is a noble commitment. However, **we have to ask why the US Immigration and Customs Enforcement (ICE) and municipal child welfare agencies separate families to begin with.** Is it because they have not had the good fortune to be enrolled into the supervision of agencies that operate the foster care system? Or is it something else? In answering this question, we must recognize something that is not immediately apparent in the banal language of the bill: that **expanding data collection, risk assessments and predictive analytics is central to the project of “predicting who needs prevention” and memos guiding implementation of the Family First Prevention Act.”** (Abdurahman, 2021, p. 10)

Fairness != Justice.

- Fairness…
  - according to whom?
  - with consideration of what history?
  - in service of what goal?
- Justice is…
  - AI that functions for all, not most
  - Considerations beyond the algorithm and "biased data"

Bias in NLP

Kenneth (Kenny) Joseph

University at Buffalo
Department of Computer Science and Engineering
School of Engineering and Applied Sciences

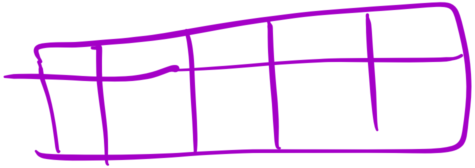
5/4/22

Digging in – how to build a resume classifier?

Words, Bags of Words, Word vectors, Document Vectors, Contextualized word vectors, …

① Perform cleaning: removing Stopwords, etc.

② Vectorize: create term-doc matrix

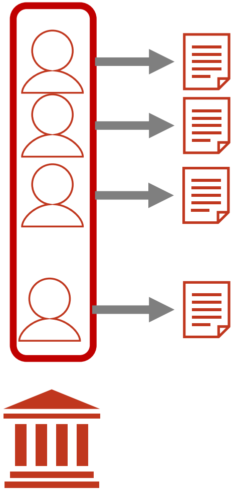


More info: https://arxiv.org/pdf/1902.06006.pdf

@_kenny_joseph

Generating Word embeddings

Step 1: Select a corpus

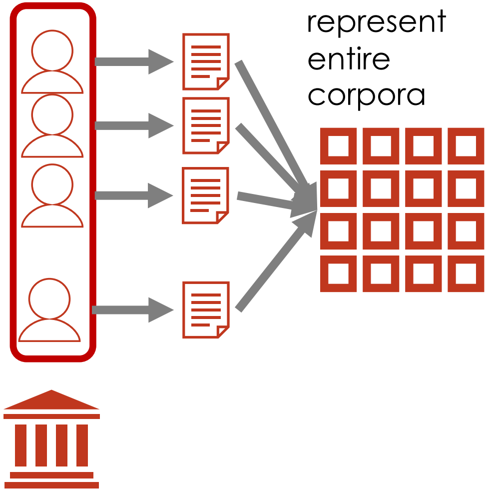


@_kenny_joseph

Typical Corpora

- Main requirement: corpus is large.
- Some common corpora
  - News articles
  - Common Crawl
  - Twitter
  - Wikipedia

2. Create co-occurence matrix is to represent entire corpora



context window of (here) size 4 hyperparam!

1 2 3 4 1 2 3 4

[drove] [my] [high] [speed] vehicle [down] [the] [road] [today] yay!

Centre Word

Context Context

over all of m, text, # have in the context of vehicle

	drove	my				
vehicle	0	1	1			
down		1	1	1	1	1

Co-occurrence matrix

n words	context 1	context 2	context 3	context 4	context 5	context 6	context 7	....	context m
word 1	2	0	0	3	0	2	7	.....	4
word 2	3	1	0	6	0	0	2	.....	0
word 3	1	3	4	2	7	2	0	.....	9
word 4	7	0	1	0	3	0	7	.....	4
word 5	0	2	0	4	0	0	7	.....	0
word 6	0	9	3	2	1	3	0	.....	0
word 7	2	0	0	1	0	5	1	.....	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
word n	5	0	1	3	0	0	5	.....	3

m contexts

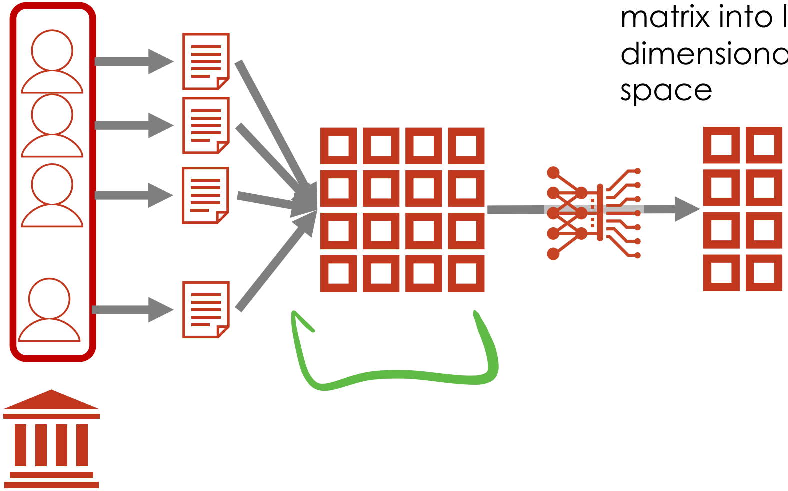
Q

- What is the difference between a **co-occurance matrix** and a **term-document** matrix?

- What impact do you think that has on the resultant embeddings?

Themes

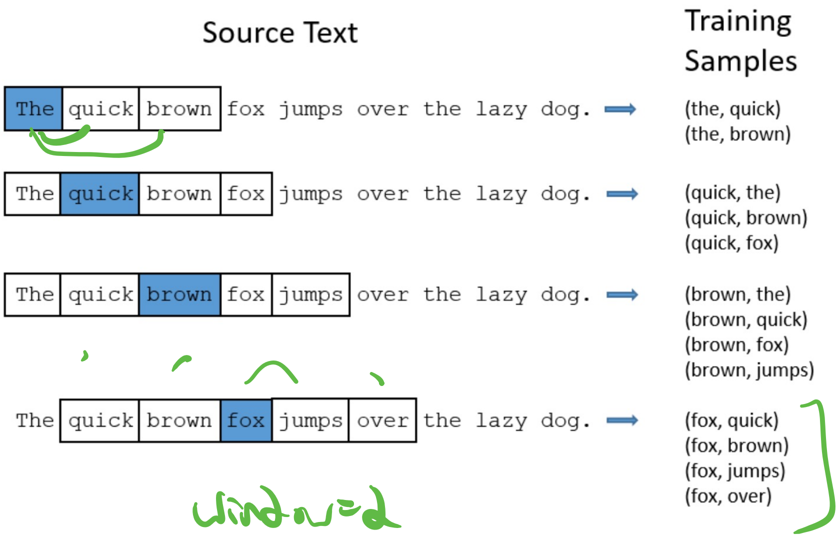
3. Run algorithm (e.g. SGNS, GloVe) to embed words in co-occurance matrix into low dimensional space



@_kenny_joseph

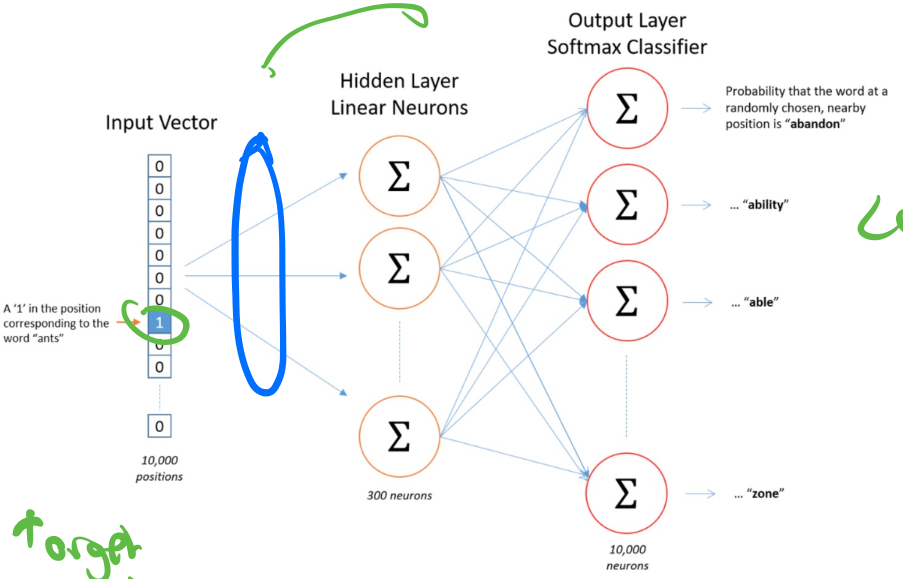
Skip gram with negative sampling

https://towardsdatascience.com/word2vec-skip-gram-model-part-1-intuition-78614e4d6e0b



Skip gram with negative sampling

https://towardsdatascience.com/word2vec-skip-gram-model-part-1-intuition-78614e4d6e0b

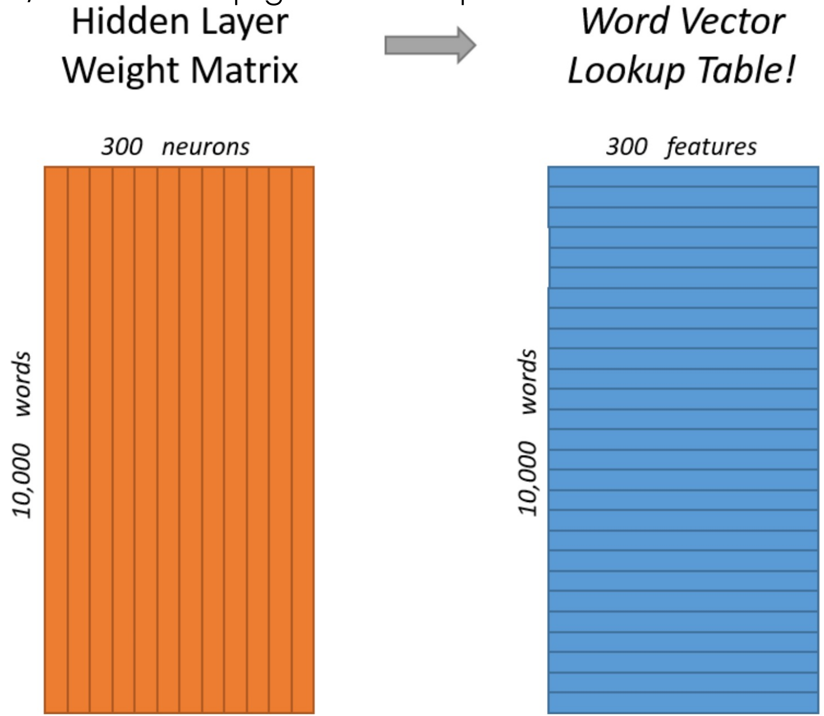


context words

target word

Skip gram with negative sampling

https://towardsdatascience.com/word2vec-skip-gram-model-part-1-intuition-78614e4d6e0b



GloVe

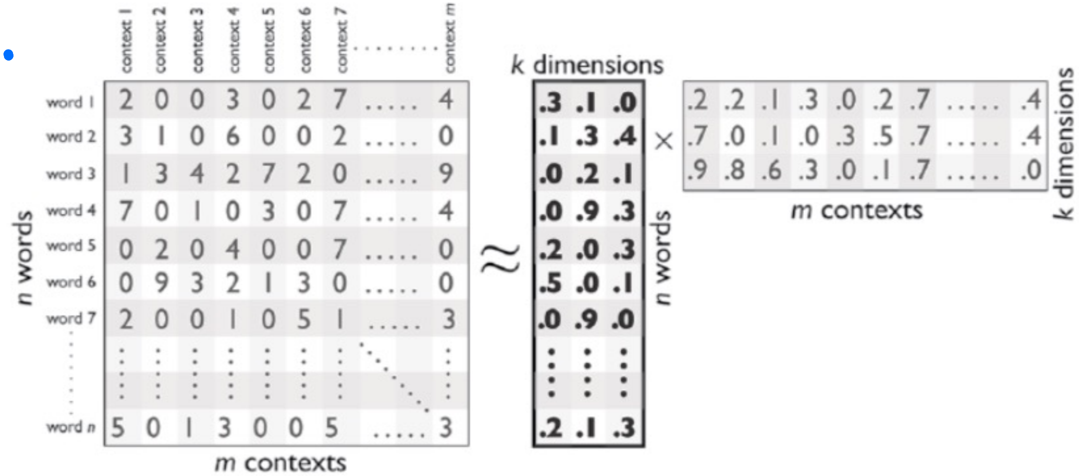


Figure 1. Schematic Illustration of the Descriptive Problem Neural Word Embeddings Solve—How to Represent All Words from a Corpus within a *k*-Dimensional Space That Best Preserves Distances between Words in Their Local Contexts

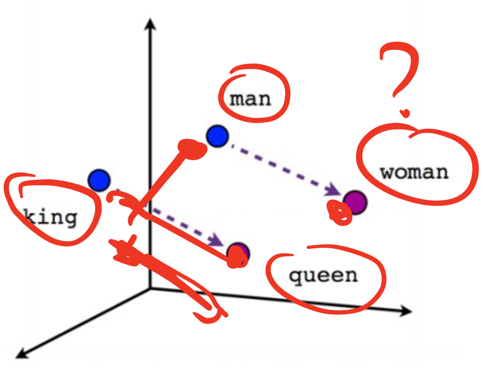
They're roughly the same!

Improving Distributional Similarity with Lessons Learned from Word Embeddings

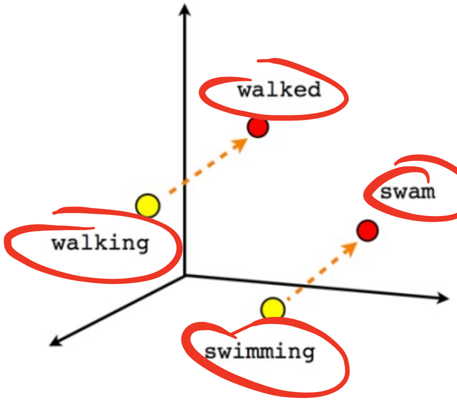
Omer Levy **Yoav Goldberg** **Ido Dagan**
Computer Science Department
Bar-Ilan University
Ramat-Gan, Israel
{omerlevy,yogo,dagan}@cs.biu.ac.il

Abstract

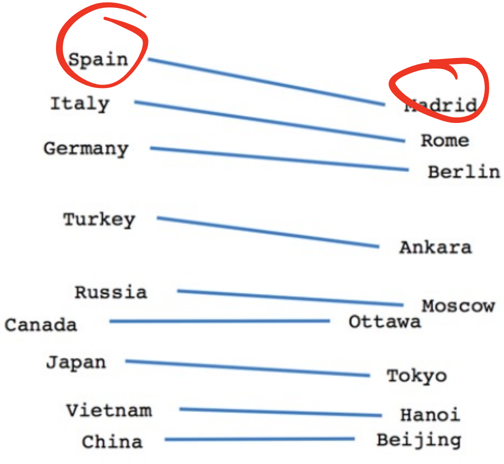
A recent study by Baroni et al. (2014)



Male-Female



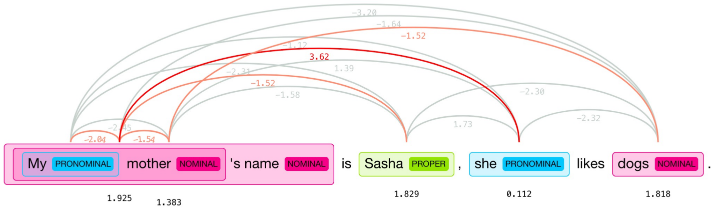
Verb tense

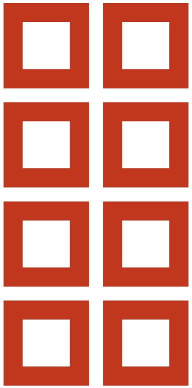


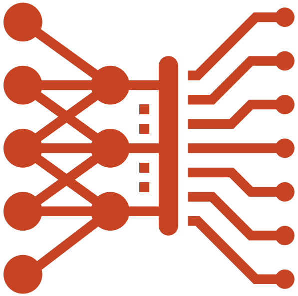
Country-Capital

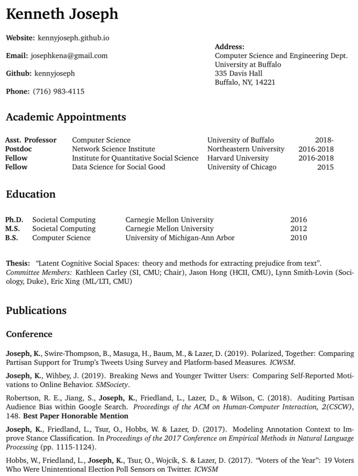
Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. *HLT-NAACL*, 746–751. Citeseer.

https://explosion.ai/demos/sense2vec









Kenneth Joseph

Website: kennyjoseph.github.io

Email: josephkena@gmail.com

Github: kennyjoseph

Phone: (716) 983-4115

Address:
Computer Science and Engineering Dept.
University at Buffalo
335 Davis Hall
Buffalo, NY, 14221

Academic Appointments

Asst. Professor	Computer Science	University of Buffalo	2018-
Postdoc	Network Science Institute	Northeastern University	2016-2018
Fellow	Institute for Quantitative Social Science	Harvard University	2016-2018
Fellow	Data Science for Social Good	University of Chicago	2015

Education

Ph.D.	Societal Computing	Carnegie Mellon University	2016
M.S.	Societal Computing	Carnegie Mellon University	2012
B.S.	Computer Science	University of Michigan-Ann Arbor	2010

Thesis: "Latent Cognitive Social Spaces: theory and methods for extracting prejudice from text". *Committee Members:* Kathleen Carley (SI, CMU; Chair), Jason Hong (HCII, CMU), Lynn Smith-Lovin (Sociology, Duke), Eric Xing (ML/LTI, CMU)

Publications

Conference

Joseph, K., Swire-Thompson, B., Masuga, H., Baum, M., & Lazer, D. (2019). Polarized, Together: Comparing Partisan Support for Trump's Tweets Using Survey and Platform-based Measures. *ICWSM.*

Joseph, K., Wihbey, J. (2019). Breaking News and Younger Twitter Users: Comparing Self-Reported Motivations to Online Behavior. *SMSociety*.

Robertson, R. E., Jiang, S., **Joseph, K.**, Friedland, L., Lazer, D., & Wilson, C. (2018). Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction, 2(CSCW)*, 148. **Best Paper Honorable Mention**

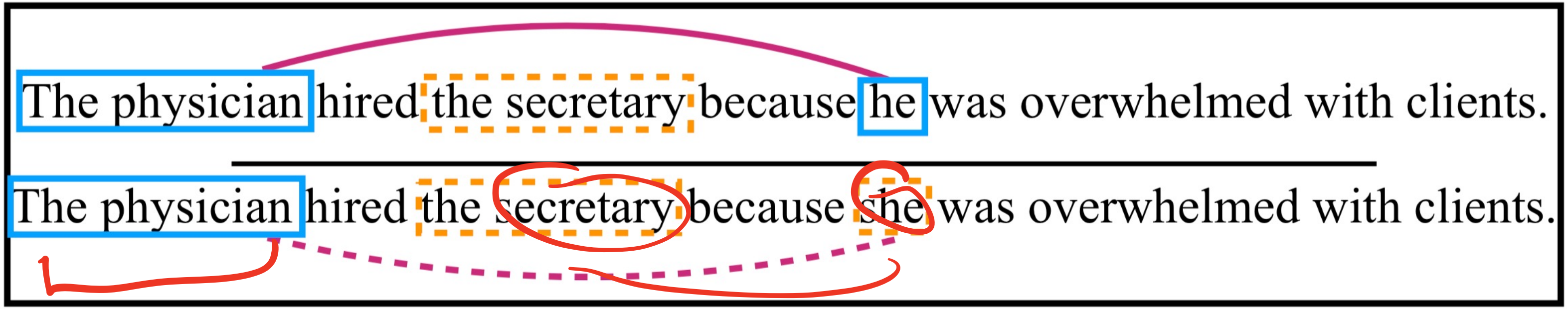
Joseph, K., Friedland, L., Tsur, O., Hobbs, W. & Lazer, D. (2017). Modeling Annotation Context to Improve Stance Classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1115-1124).

Hobbs, W., Friedland, L., **Joseph, K.**, Tsur, O., Wojcik, S. & Lazer, D. (2017). "Voters of the Year": 19 Voters Who Were Unintentional Election Poll Sensors on Twitter. *ICWSM*

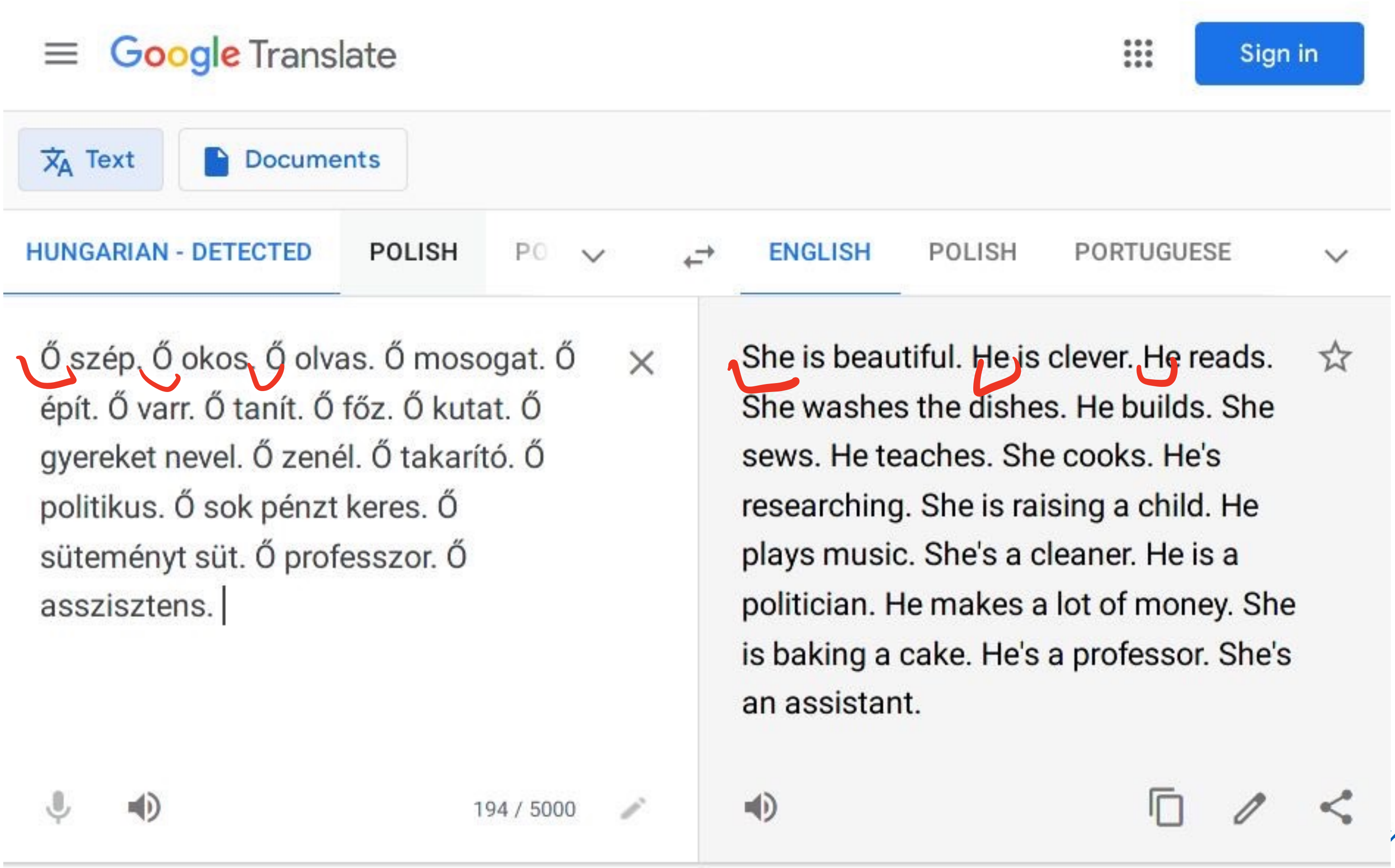
Checking in

- We learn word embeddings from large text corpora we find on the internet
- We then use them to do a lot of things, like coreference resolution and document classification
- Anyone see a potential problem w/ this?

http://wordbias.umiacs.umd.edu/



https://huggingface.co/coref

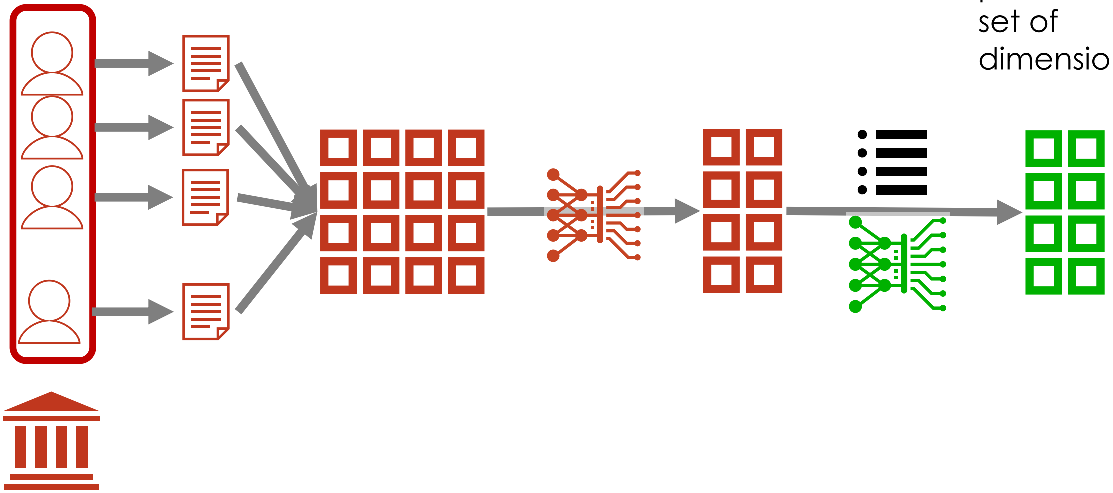

Google Translate
Sign in
Text
Documents
HUNGARIAN - DETECTED
POLISH
PO
ENGLISH
POLISH
PORTUGUESE
Ő szép. Ő okos. Ő olvas. Ő mosogat. Ő épít. Ő varr. Ő tanít. Ő főz. Ő kutat. Ő gyereket nevel. Ő zenél. Ő takarító. Ő politikus. Ő sok pénzt keres. Ő süteményt süt. Ő professzor. Ő asszisztens.
She is beautiful. He is clever. He reads. She washes the dishes. He builds. She sews. He teaches. She cooks. He's researching. She is raising a child. He plays music. She's a cleaner. He is a politician. He makes a lot of money. She is baking a cake. He's a professor. She's an assistant.
194 / 5000

Erg

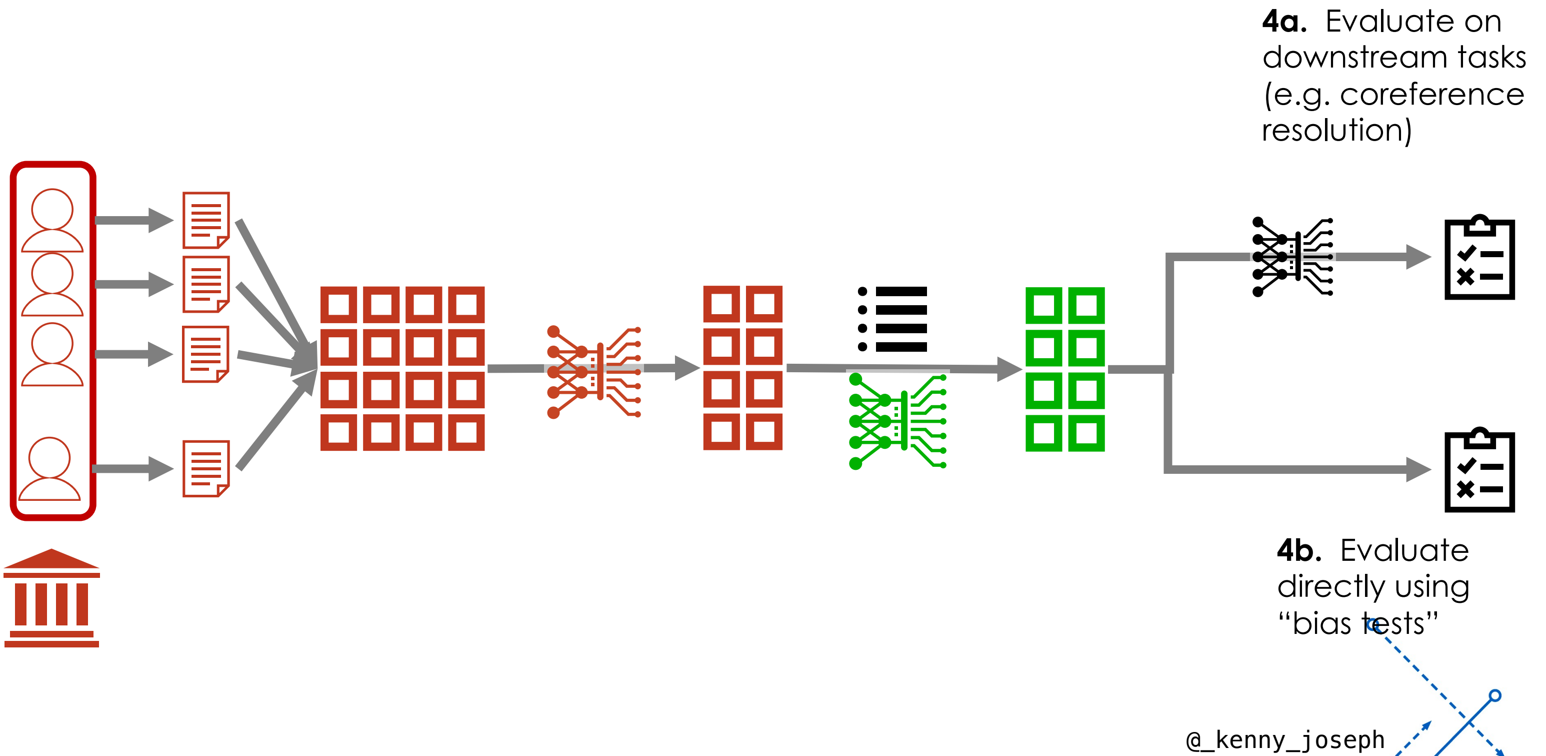
- Now what?
- One thought early on – what if we “debias” the embeddings?

Debiasing

4. Run debiasing algorithm to remove bias along particular set of dimensions



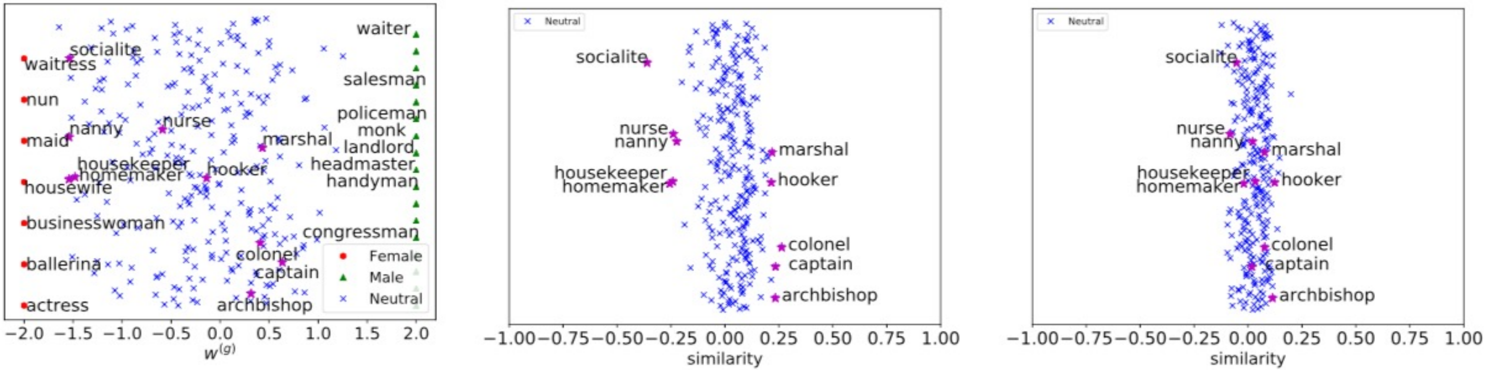
Debiasing



@_kenny_joseph

How debiasing (typically) works

- **Step 0:** Usually gender, sometimes race, or good/bad
- **Step 1:** Divide words into 3 camps:
  - *Neutral* – Words that have no relation to gender ("millieu")
  - *Definitional* – Words that are gendered by definition ("sister")
  - *Biased* – Words that are gendered but shouldn't be (secretary)
- **Step 2:** Identify "gender direction"
  - Select words at either ends of a "gender spectrum"
  - Identify direction in the embedding using those words
- **Step 3:** Try to remove gender direction from biased words, keep it for definitional and neutral words



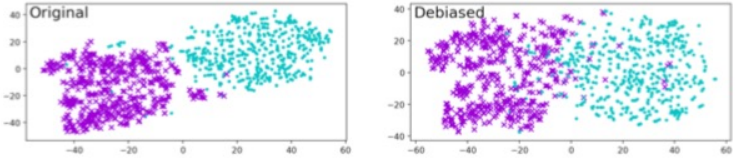
(a) $w^{(g)}$ dimension for all the professions

(b) Gender-neutral profession words projected to gender direction in GloVe

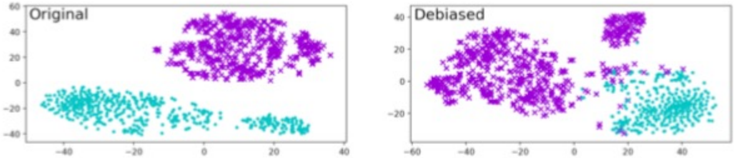
(c) Gender-neutral profession words projected to gender direction in GN-GloVe

Figure 1: Cosine similarity between the gender direction and the embeddings of gender-neutral words. In each figure, negative values represent a bias towards female, otherwise male.

Debiasing – does it work?



(a) Clusteing for HARD-DEBIASED embedding, before (left hand-side) and after (right hand-side) debiasing.



(b) Clusteing for GN-GLOVE embedding, before (left hand-side) and after (right hand-side) debiasing.

- Does it really make sense to treat bias as existing along a "direction"?
  - I think so, actually
  - **But this doesn't mean that it makes sense to *debias* along a direction**

Gonen, H., & Goldberg, Y. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. *ArXiv Preprint ArXiv:1903.03862*.

Debiasing… some detours

- Exercise: Can you remove gender from the English language? Would that solve gender bias in NLP?
- Exercise 2: If we removed the gendered connotation of the word "secretary", would there be any other biases?
- Exercise 3 (hat tip to Ayoub): When should we stop debiasing language?

Another Detour

Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media

Munmun De Choudhury
Georgia Tech
Atlanta GA 30332
munmund@gatech.edu

Emre Kiciman
Microsoft Research
Redmond WA 98052
emrek@microsoft.com

Mark Dredze
Johns Hopkins University
Baltimore MD 21218
mdredze@cs.jhu.edu

Glen Coppersmith
Qntfy.io
Crownsville MD, 21032
glen@qntfy.io

Mrinal Kumar
Georgia Tech
Atlanta GA 30332
mkumar73@gatech.edu

Wo. We got there!

- Preview of next week
- Out early for additional questions