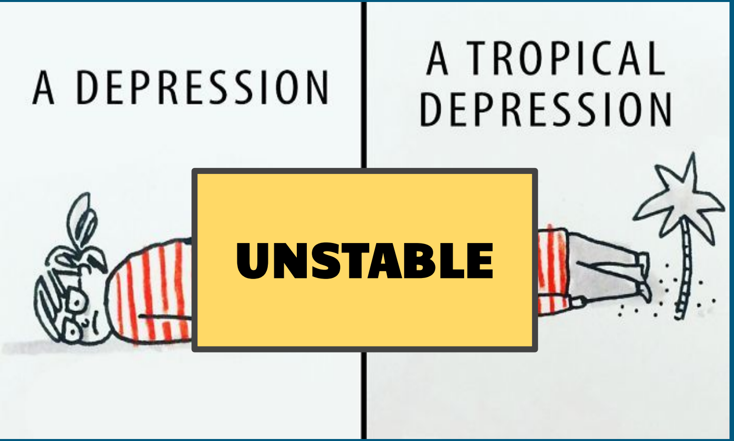
IATs, and WEATs, and WEFATs, oh my!

Karie Moorman, CIS UC Merced
Paul Smaldino, CIS UC Merced
2nd Year Presentation, 4 May 2018

outline:

1. implicit bias and IATs (Greenwald et al, 1998)
2. word vectors, WEAT and WEFAT (Caliskan et al, 2017)
3. application of these measures

1. We remember IATs, right?
(Greenwald et al, 1998; ?? 2002)



How are these biases represented in "big data"?

2. WEAT's a WEFAT? (Caliskan et al, 2017)

Machines learn what people know implicitly.

"Our methods hold promise for identifying and addressing sources of bias in culture, including technology"

glove word vectors: glove.840B.300dcasedCommoncrawl.txt

Brief description of WEAT and WEFAT?

3. In what ways are speakers of English biased against "immigrants"?

Vanderbilt: Efrén Pérez (2010)
Latino vs. White immigrant surnames & Good vs. Bad words

Rice: James Hedrick & Aleks Ksiazkiewicz (2012)
Latino vs. White immigrant surnames & Positive vs. Negative words
Latino vs. White immigrant surnames & High vs. Low-skilled jobs

Yale: Lydia Keating (2017)
Immigrant vs. Non-immigrant words & Positive vs. Negative words
67 participants, 28 were Indian and 39 were American

IAT replication results:

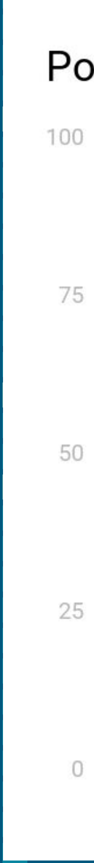
	Lexemes	IAT (Cohen's d)	WEFAT (Cohen's d)
Vanderbilt (10x10) Surnames & Good/Bad words	honest, joy, love, peace, wonderful, honor, pleasure, glorious, laughter, happy agony, prison, terrible, horrible, nasty, evil, awful, failure, hurt, poverty	**1.65**	**1.58**
Rice (4x4) Surnames & +/- words	wonderful, pleasure, glorious, happy terrible, horrible, nasty, awful	**0.69**	**1.61**
High/Low skilled jobs	doctor, engineer, professor, scientist laborer, busboy, janitor, maid	**-0.13**	**1.78**
Yale (6x7) Immigrant/Non-Immigrant & +/- words	lovely, pleasure, glorious, beautiful, marvelous, wonderful, joyful humiliate, terrible, painful, nasty, horrible, agony, tragic	**-0.30**	**-1.19**

Other interesting word sets we can compare:

Citizens: citizens, locals, residents, natives, inhabitants

Non-citizens: immigrants, illegals, foreigners, undocumented, refugees

	Citizens/ Non-citizens	Lexemes
Positive/ Negative Attributes	***p*=0.10 , *d*=0.82**	sincere, honest, nice, trustworthy, reliable sneaky, deceitful, mean, rude, careless
Protected/ Vulnerable words	***p*=0.06 , *d*=1.50**	safe, guarded, secure, protected, shielded unsafe, susceptible, vulnerable, insecure, unprotected
Metaphors	***p*=0.04 , *d*=1.09**	come, enter, move, walk, go steal, cheat, hide, harm, sneak



“immigrants” vs. “refugees” by Countries

(p=0.16, d=1.41)

IAT: Citizen vs. Non-citizen words and Good vs. Bad Attribute words

citizens, locals, residents, natives, inhabitants	sincere, honest, nice, trustworthy, reliable
immigrants, illegals, undocumented, refugees, foreigners	sneaky, deceitful, mean, rude, careless

- Is implicit bias present?
- Is it really "implicit"? (Howell, 2017)
- How do these results compare to "big data"?
- Does affiliation with social/ethnic groups influence bias? (Pérez, 2010)
- Does political affiliation influence bias?

IAT: Citizens vs. Non-citizens and Good vs. Bad Attributes

N = 60-70 participants per group

(i.e., race/ethnicity, gender, self-identification as “immigrant” or “refugee”)

Procedure:

(1) IAT (5x5)
(2) Explicit preference for “non-citizens” (5pt. Likert scale: Extremely good/bad)
(3) Expectation of feedback

    (7pt. Likert scale: Strong Preference for Non-citizens/Citizens)
(4) Self-identification as “immigrant” or “refugee” (Y/N)
(5) Political Affiliation Questions (Wilson-Patterson Scale)
(6) Demographics Questions

IAT: Citizens vs. Non-citizens and Good vs. Bad Attributes

n=19

Future Directions

Investigate bias in other social groups

(e.g., gender/sexuality, political affiliation, diet, mode of transportation)

How does bias vary across corpora?

- Pre-trained models (e.g., Google News, WordNet)
- News and Forums (e.g., Reddit, Latino/Asian news outlets)

Directionality of Metaphor using RTs from IATs?

References

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183-186.

Hedrick, J., & Ksiazkiewicz, A. (2012). Implicit Attitudes toward Highly Skilled and Low-skilled Immigration.

Howell, J. L., Redford, L., Pogge, G., & Ratliff, K. A. (2017). Defensive Responding to IAT Feedback. *Social Cognition*, *35*(5), 520-562.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, *74*(6), 1464.

Keating, L. (2017). *A Cross Cultural Analysis of Implicit and Explicit Xenophobia* (Doctoral dissertation, Yale University).

Pérez, E. O. (2010). Explicit evidence on the import of implicit attitudes: The IAT and immigration policy judgments. *Political Behavior*, *32*(4), 517-545.

thanks.

Paul Smaldino, Colin Holbrook,
Alexia Galati, Jordan Ackerman