Ethics, Social Good, and NLP

(slides sampled from CMU/LTI Computational Ethics for NLP course)

Human Subjects

- We are trying to model a human function
- Labels are certainly noisy
- How to use humans to find better labels/know if they are right
- Let's put it on Amazon Mechanical Turk (Crowdsourcing) and get the answer

History of using Human Subjects

- WWII Nazi and Japanese prisoners in concentration camps
  - Medical science did learn things
  - But even at the time this was not considered acceptable
- Tuskegee Syphilis Experiments (US Public Health System, 1932-1972)
  - Understand how untreated syphilis develops
  - African-American sharecroppers given free healthcare, meals…
  - Not provided with penicillin when it would have helped
- Milgram Obedience Experiment (Yale, 1962)
  - Experimenters asked subjects/teachers to give electric shocks after wrong answers
- Others…

Ethics in Human Subject Use

- These experiments led to the National Research Act 1974
  - Requiring "Informed Consent" from participants
  - Requiring external review of experiments
- For all federal funded experiments
  - Covers my dialogue work at Pitt, but not when I was at AT&T!

IRB (Ethical Review Board)

- Institutional Review Board
  - Internal to institution
  - Independent of researcher
- Reviews all human experimentation
  - Assesses instructions
  - Compensation
  - Contribution of research
  - Value to the participant
  - Protection of privacy

IRB (Ethical Review Board)

- Different standards for different institutions
  - Medical School vs SCI
- Board consists of (primarily) non-expert peers
- At educational institutions also
  - Help educate new researchers (e.g., before Pitt FER starts)
  - Make suggestions to find solutions to ethics problems
    - How to get informed consent on an Android App
    - "click here to accept terms and conditions"

Ethical Questions

- Can you lie to a human subject?
- Can you harm a human subject?
- Can you mislead a human subject?
  - Wizard of Oz experiments?

Using Human Subjects

- But it's not all these extremes
- Your human subjects are biased
- Your selection of them is biased
- Your tests are biased too

Human Subject Selection Example

- For speech synthesis evaluation
  - Listen to these and say which you prefer
- Who do you get to listen
  - Experts are biased, non-experts are biased
- Hardware makes a difference
  - Expensive headphones give different result
- Experiment itself makes a difference
  - Listening in quiet office vs on the bus
- Hearing ability makes a difference
  - Young vs old

Human Subject Selection

- All subject pools will have bias
  - So identify the biases (as best you can)
  - Does the bias affect your result (maybe not)
- Can you recruit others to reduce bias
  - Can you do this post experiment
- Most Psych experiments use undergrads
  - Undergrads do experiments for course credit
  - SCI researchers typically recruit via $
  - Real vs. experimental users yield different results

Human Subject Selection

- Most IRB have special requirements for involving
  - Minors, pregnant women, disabled
- So most experiments exclude these
- Protected or hard to access groups are underrepresented

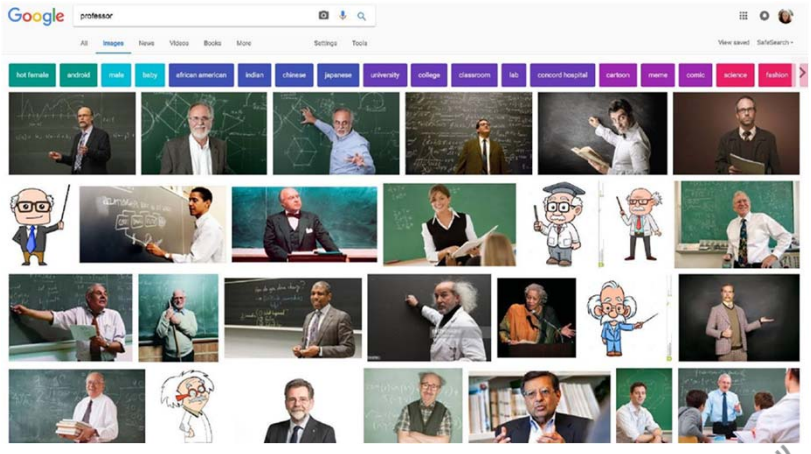
Human Subjects – Summary Part 1

- Unchecked human experiment
- Led to IRB reviews of human experimentation
- All human experimentation includes bias
  - Admit it, and try to ameliorate it
  - Experimentation vs Actual is different

Online data is riddled with **SOCIAL STEREOTYPES**

Gender/Race/Age Stereotypes

- June 2017: image search query "Professor"

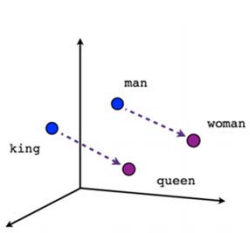


Consequence: models are biased

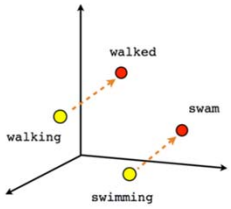
Sources of Human Biases in Machine Learning

- Sample selection bias
  - unbalanced training data
  - data and annotations may reflect human cognitive biases and cultural stereotypes
- Optimizing towards a biased objective
  - Labels are biased proxies to the real objective
    - e.g., “who is more likely to be convicted” vs “who is more likely to commit a crime”
- Inductive bias
  - the set of “assumptions” used by the learner, e.g. features in discriminative models are biased

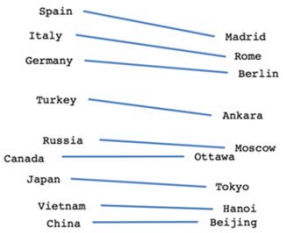
Word Analogy Tasks



Male-Female



Verb tense

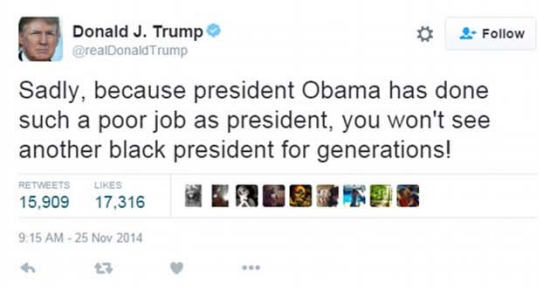


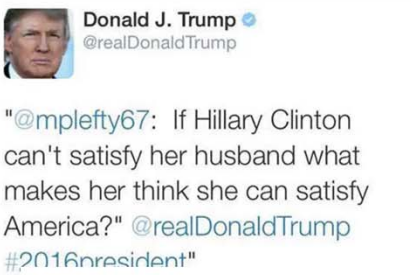
Country-Capital

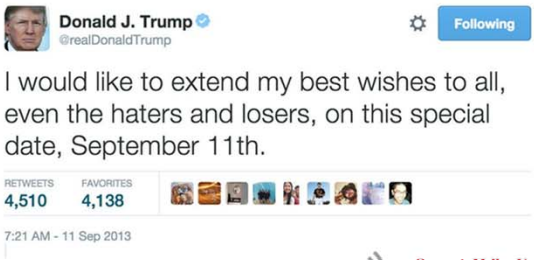
- Mikolov et al. '13
- *man* is to *king* as *woman* is to x

$$\min cos(man - woman,\ king - x)\ s.t.\ ||king - x||_2 < \delta$$

Training Data for Word Embeddings







Bolukbasi T., Chang K.-W., Zou J., Saligrama V., Kalai A. (2016) **Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.** *NIPS*

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

Main Ideas

- Demonstrate gender bias in embeddings trained even from Google news
- Show that gender defined words are linearly separable from others that (should be?) gender neutral
- Use the above finding as the basis of a word embedding debiasing algorithm
- Evaluate

$$\min cos(he - she,\ x - y)\ s.t.\ ||x - y||_2 < \delta$$

Extreme *she*	Extreme *he*
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

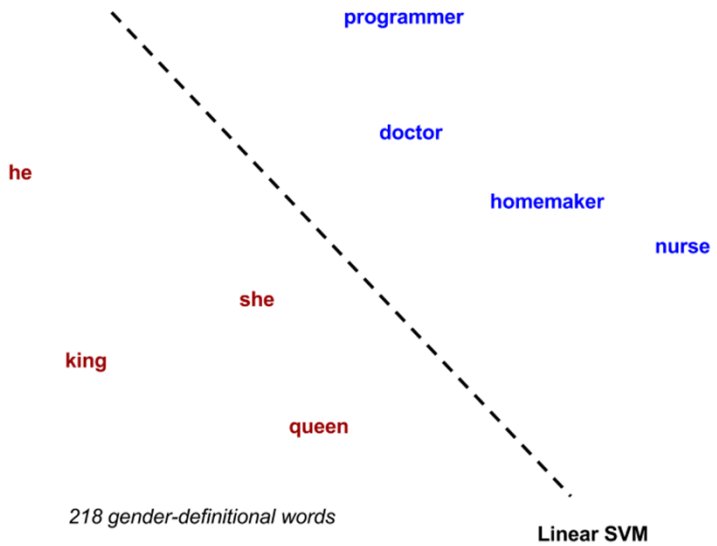
Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

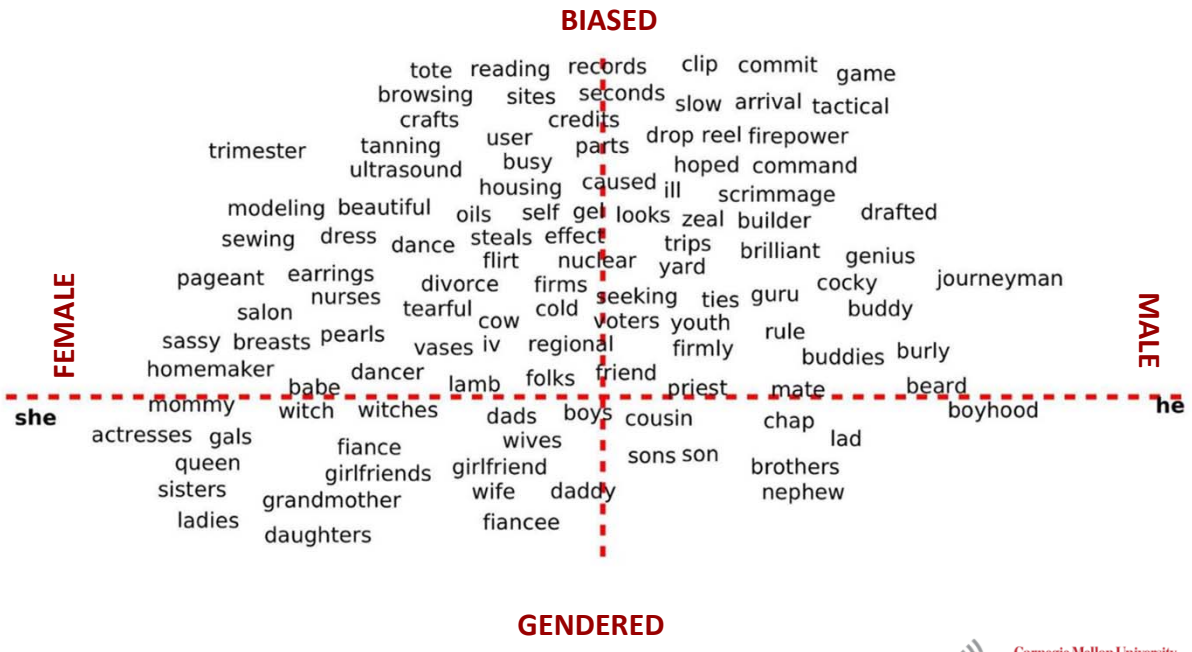
Figure 1: **Left** The most extreme occupations as projected on to the *she−he* gender direction on w2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded. **Right** Automatically generated analogies for the pair *she-he* using the procedure described in text. Each automatically generated analogy is evaluated by 10 crowd-workers to whether or not it reflects gender stereotype.

lon University
chnologies Institute

Gender-definitional vs. Gender-neutral Words

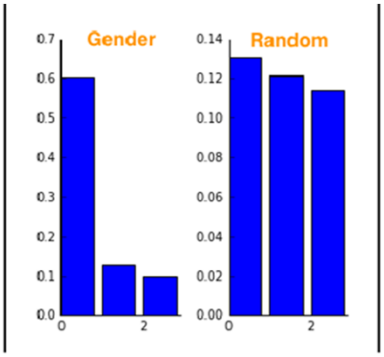


Carnegie Mellon University
Language Technologies Institute



Gender Subspace

	def.	stereo.
$\overrightarrow{she} - \overrightarrow{he}$	92%	89%
$\overrightarrow{her} - \overrightarrow{his}$	84%	87%
$\overrightarrow{woman} - \overrightarrow{man}$	90%	83%
$\overrightarrow{Mary} - \overrightarrow{John}$	75%	87%
$\overrightarrow{herself} - \overrightarrow{himself}$	93%	89%
$\overrightarrow{daughter} - \overrightarrow{son}$	93%	91%
$\overrightarrow{mother} - \overrightarrow{father}$	91%	85%
$\overrightarrow{gal} - \overrightarrow{guy}$	85%	85%
$\overrightarrow{girl} - \overrightarrow{boy}$	90%	86%
$\overrightarrow{female} - \overrightarrow{male}$	84%	75%



The top PC captures the gender subspace

Debiasing

1. Identify gender-definitional and gender-neutral words
2. Project away the gender subspace from the gender-neutral words
3. Normalize vectors