



What can go wrong with word embeddings?

- ▶ What's wrong with learning a word's "meaning" from its usage? Maybe some words are used in ways we don't want to replicate?
- ▶ What data are we learning from?
- ▶ What are we going to learn from this data?



What do we mean by bias?

► Identify *she* - *he* axis in word vector space, project words onto this axis

- 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. fighter pilot | 12. boss |

Bolukbasi et al. (2016)

► Nearest neighbor of (b - a + c)

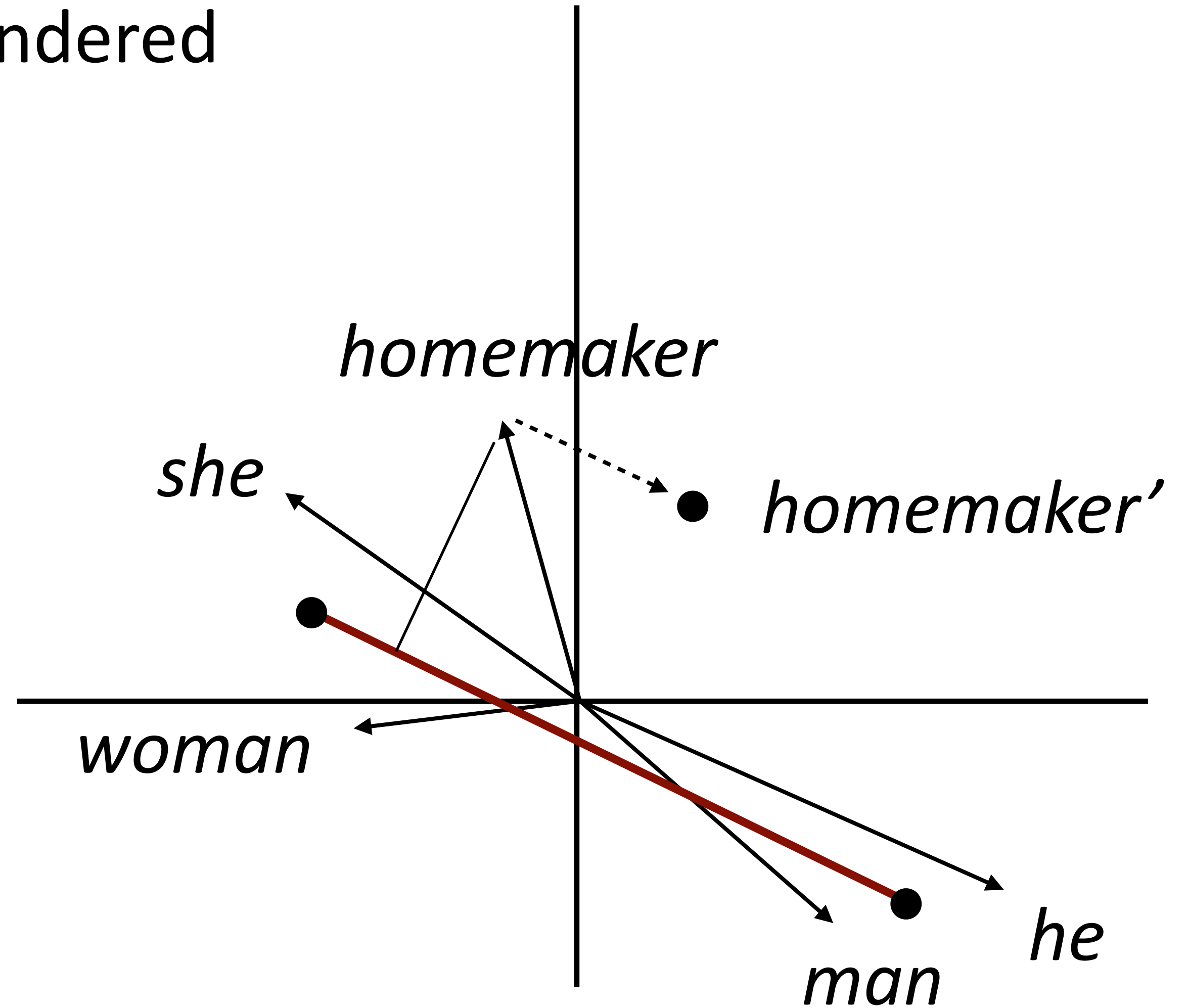
Racial Analogies	
black → homeless	caucasian → servicemen
caucasian → hillbilly	asian → suburban
asian → laborer	black → landowner
Religious Analogies	
jew → greedy	muslim → powerless
christian → familial	muslim → warzone
muslim → uneducated	christian → intellectually

Manzini et al. (2019)



Debiasing

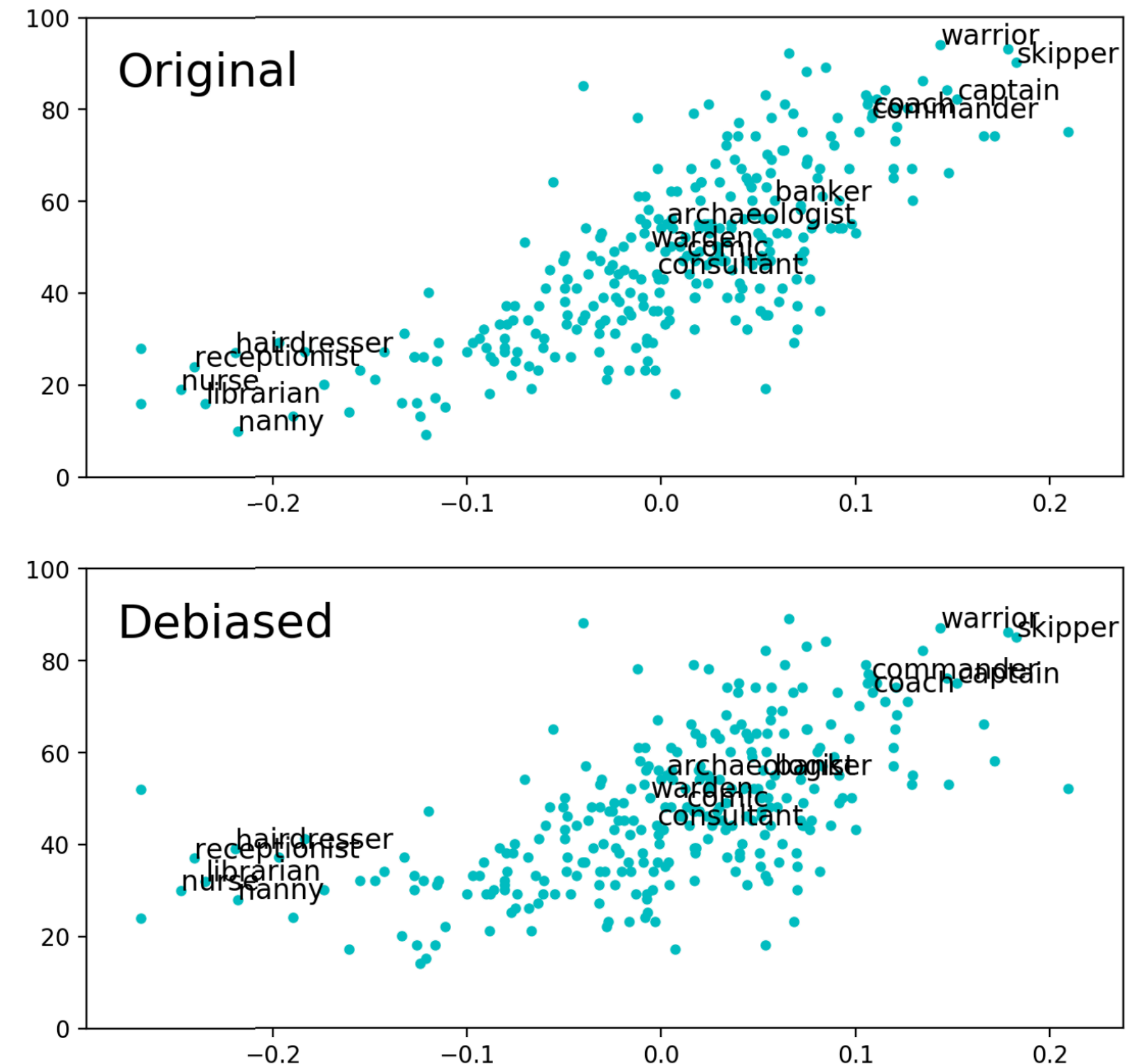
- ▶ Identify gender subspace with gendered words
- ▶ Project words onto this subspace
- ▶ Subtract those projections from the original word





Hardness of Debiasing

- ▶ Not that effective...and the male and female words are still clustered together
- ▶ Bias pervades the word embedding space and isn't just a local property of a few words



(a) The plots for HARD-DEBIASED embedding, before (top) and after (bottom) debiasing.