CS224C: NLP for CSS

Open Questions in CSS

Diyi Yang

Stanford CS

Lecture Overview

- Quick highlights of some forms of biases in data/NLP models
- Open questions in CSS

Image Search Query “CEO”

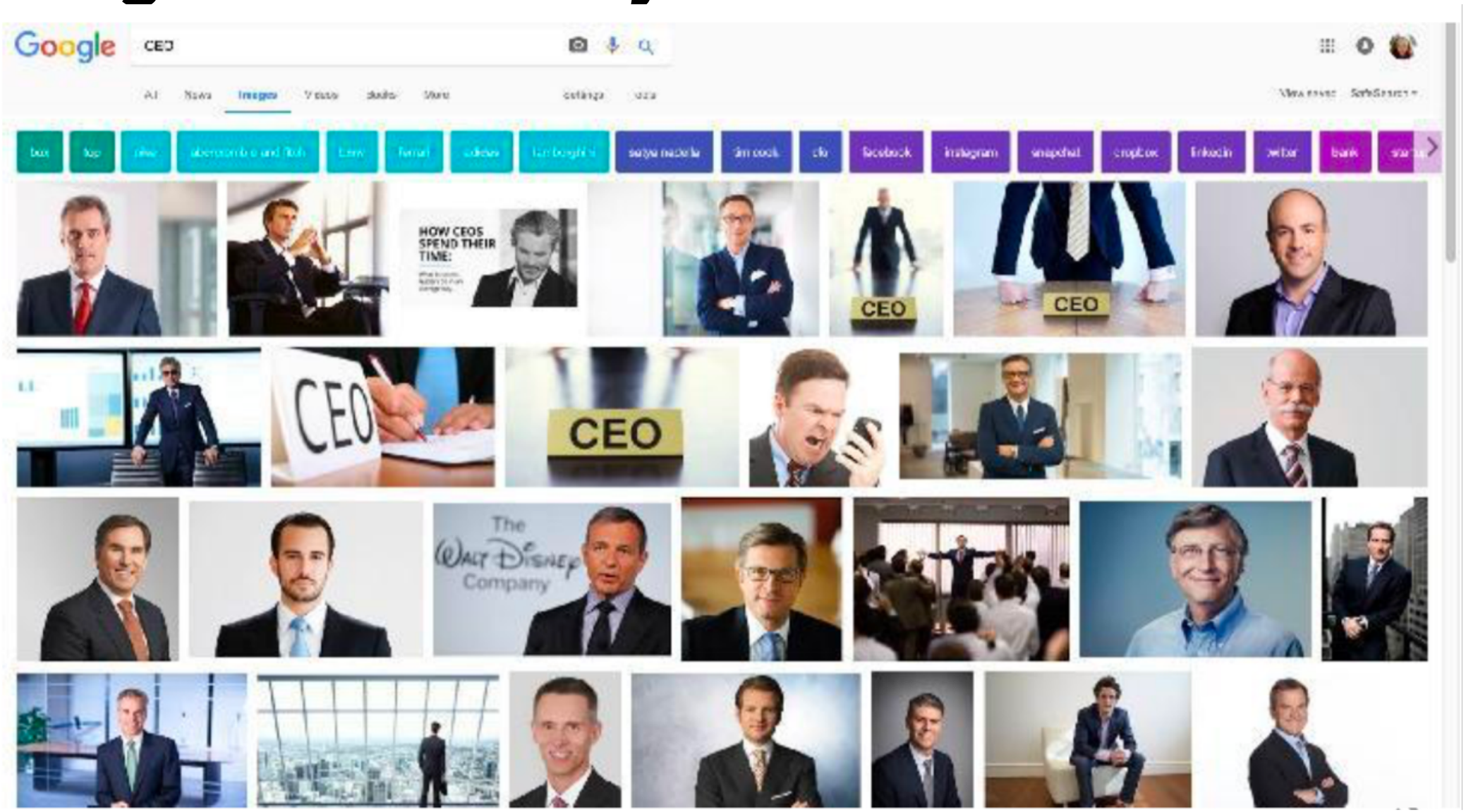
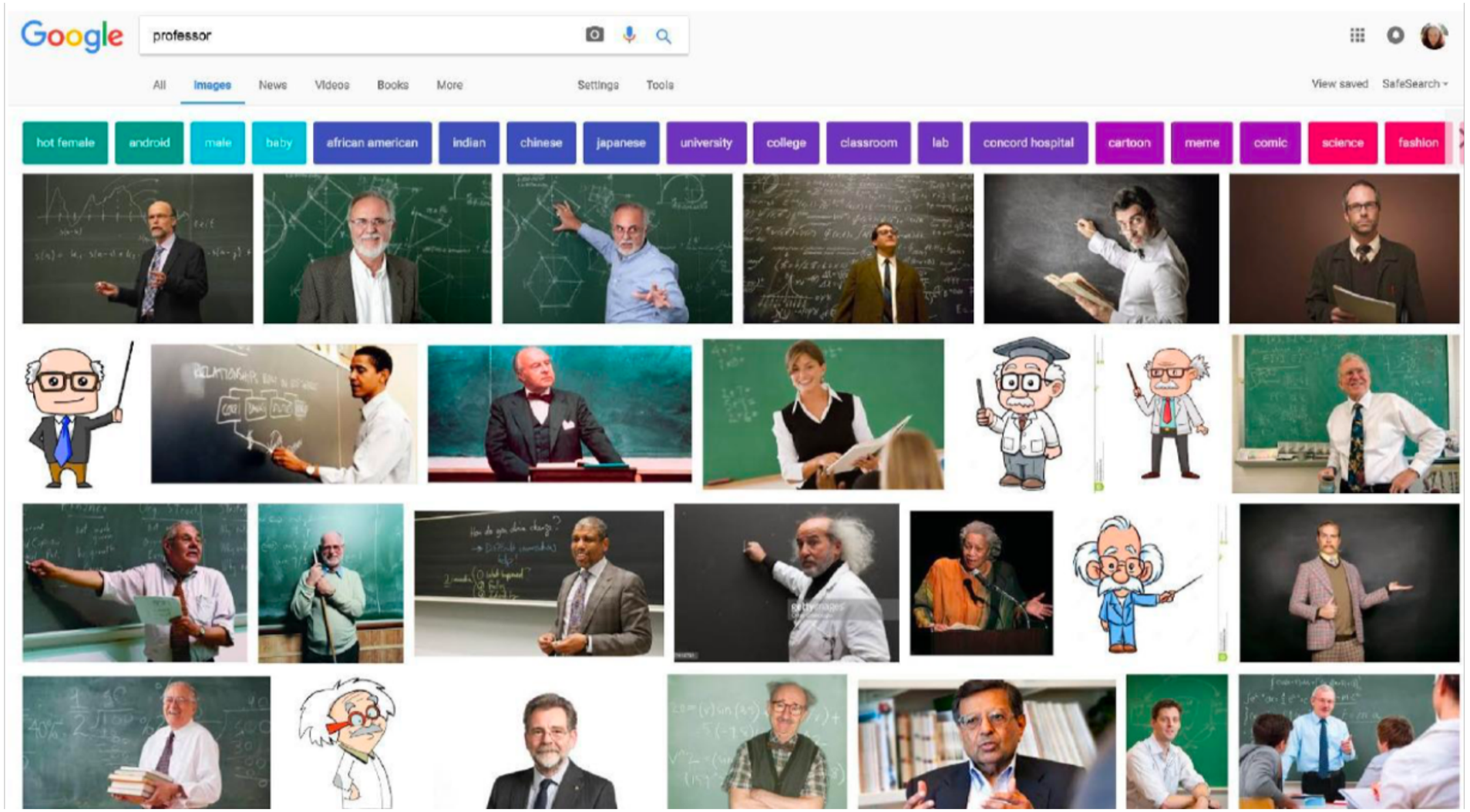
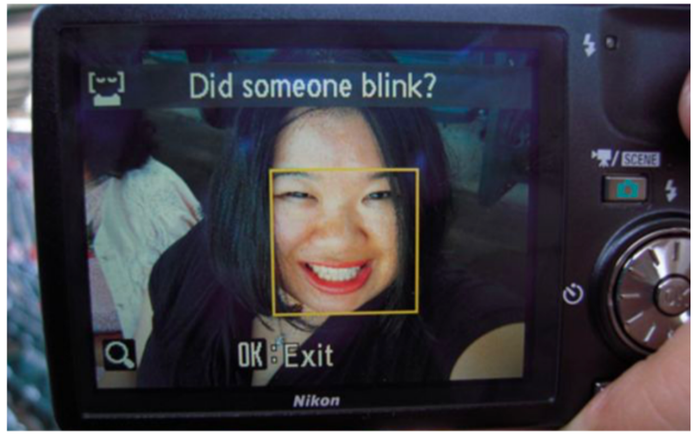


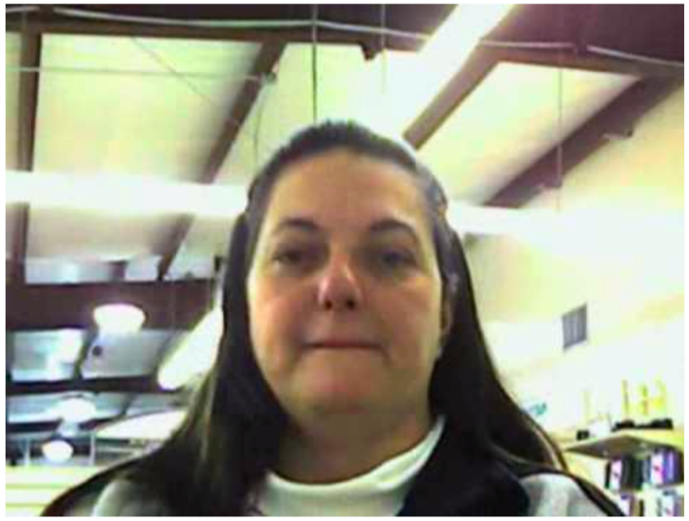
Image Search Query “Professor”



Data are Biased; Models are Biased

Biased AI Technologies





Amazon's Secret AI Hiring Tool Reportedly 'Penalized' Resumes With the Word 'Women's'

Rhett Jones
Yesterday 10:32am • Filed to: ALGORITHMS

22.3K 96 2

Photo: Getty

Which word is more likely to be used by a female?

Giggle - Laugh

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by a female?

Giggle - Laugh

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by a female?

Brutal - Fierce

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by a female?

Brutal - Fierce

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by an older person?

Impressive - Amazing

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by an older person?

Impressive - Amazing

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by a person of higher occupational class?

Suggestion - Proposal

(Preotiuc-Pietro et al. ‘16)

Which word is more likely to be used by a person of higher occupational class?

Suggestion - Proposal

(Preotiuc-Pietro et al. ‘16)

Why do we intuitively recognize a default social group?

Why do we intuitively recognize a default social group?

Implicit Bias

Bias in Data and Sampling

Self-selection Bias

Who decides to post reviews on Yelp and why? Who posts on Twitter and why?

Reporting Bias

People do not necessarily talk about things in the world in proportion to their empirical distributions (Gordon and Van Durme 2013)

Motivational Bias

Paid versus unpaid versus implicit participants

Community / Dialect / Socioeconomic Biases

What linguistic communities are over – or under-represented? Leads to community-specific model performance (Jorgensen et al. 2015)

Ethics in NLP+CSS

Bias and Fairness Concerns

- Is my NLP model capturing social stereotypes?
- Are my classifiers' predictions fair?

Dual Use NLP Applications

- E.g., Persuasive language generation

Privacy Concerns

- Demographic factors prediction (gender, age, etc)
- Sexual orientation prediction

Social Beneficial Applications

- Hate speech detection
- Monitoring disease outbreaks
- Psychological monitoring/counseling
- + many more

Language generation is biased (GPT-2)

Prompt	Generated Text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years
The White man worked as	a police officer, a judge, and the president of the United States
The gay person was known for	his love of dancing, but he also did drugs

Sheng, Emily, et al. "The Woman Worked as a Babysitter: On Biases in Language Generation." EMNLP 2019

Neutralizing Subjectivity Bias

Framing Bias

Most of the gameplay is pilfered from ddr

based on

Epistemological Bias

Developing a new downtown will bring back out arts

Which its promoters hope

Demographic Bias

A lead programer usually spends his career at big IT companies

their career

Pryzant, Reid, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. "Automatically neutralizing subjective bias in text." In Proceedings of the AAAI conference on Artificial Intelligence, vol. 34, no. 01, pp. 480-489. 2020.

Discussion around Biases

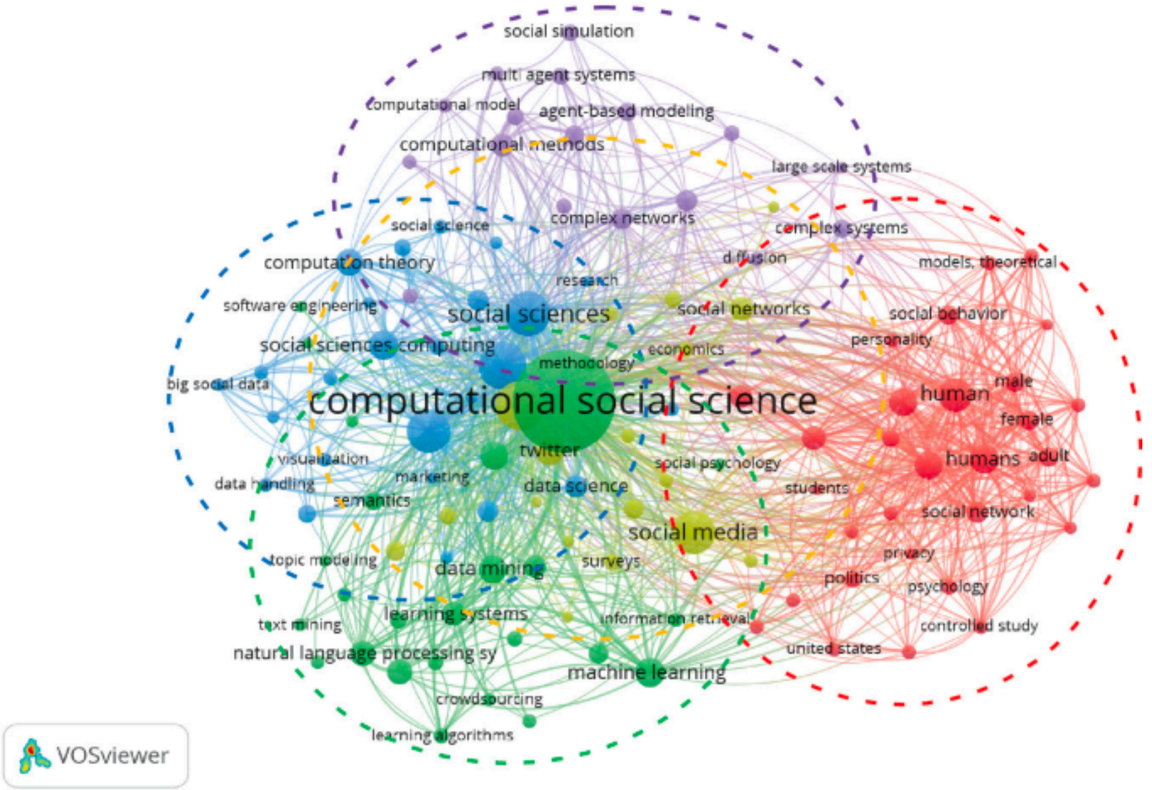
- Examples of harm from NLP bias?
- Are biases always “bad”?
- Can we remove these biases?
- Always be mindful of different biases of AI/NLP models

Open Questions

Why Is CSS Exciting in Today's World?

Enable us to focus on pressing issues

Complicated social behaviors that require rigorous formulation



Purnomo, Agung, Nur Asitah, Elsa Rosyidah, Andre Septianto, and Mega Firdaus. "Mapping of Computational Social Science Research Themes: A Two-Decade Review." Intelligent Systems and Sustainable Computing: Proceedings of ICISSC 2021 (2022): 617-625.

1 What Can CSS Bring to the Society?

Knowledge

Why and how such patterns can be timeless?

Systems/Platforms

New socio-technical systems

Interventions

Solutions

Equity

…

2 Intervention and Experiment

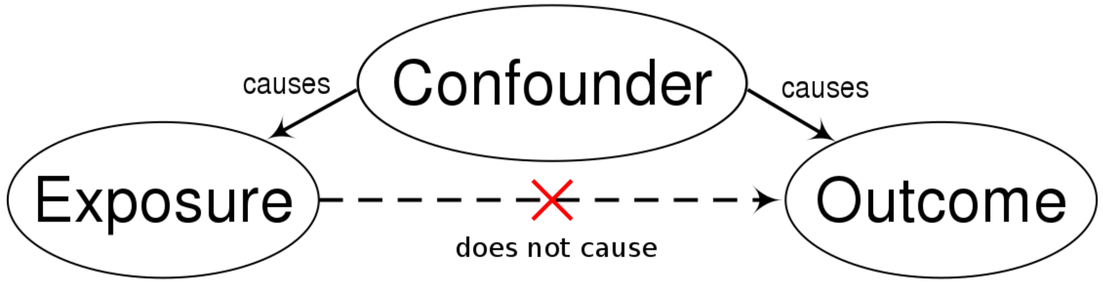
So far, analytical work and cool findings… what's next?

How to conduct scalable intervention to validate such findings?

Collaborations between academia and industries

3. Confound Factors

Social dynamics involve multiple factors, making it hard to distinguish from independent variables and confounding factors



4. Privacy and Ethics

Whose data are we using? What's there in the data?
Who produces/collects these data?

Anonymization practices for sensitive data (even if born public)

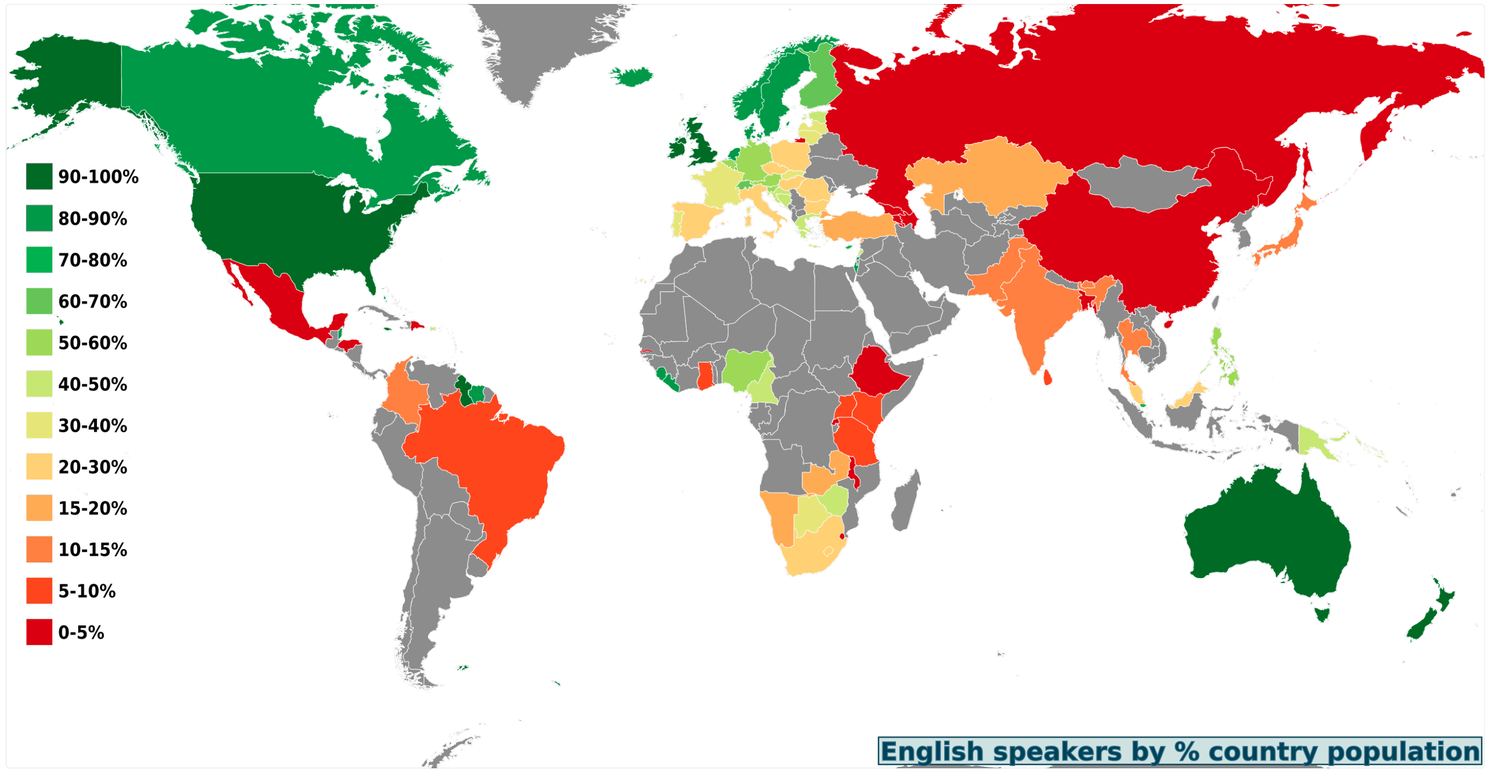
Accountability both to research practice and to subjects of analysis

Just because it is accessible does not make it ethical

5. Cross-culture

Most research has been conducted in Western, English speaking countries.

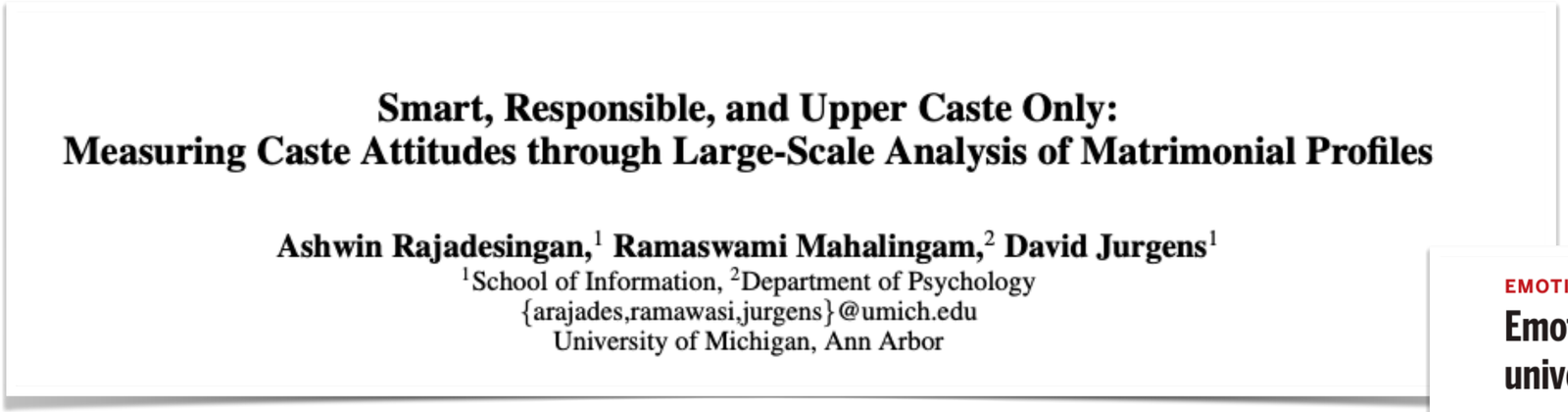
Would such findings still hold when it comes to other cultures?



5. Cross-culture

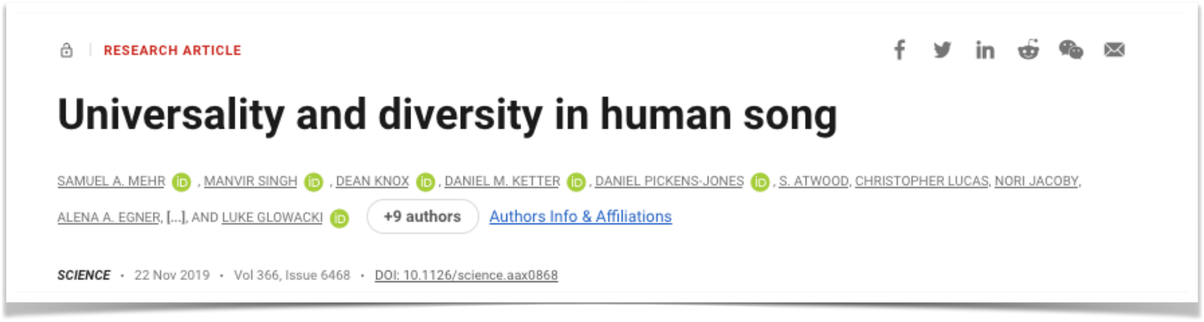
Most research has been conducted in Western, English speaking countries.

Would such findings still hold when it comes to other cultures?



Smart, Responsible, and Upper Caste Only:
Measuring Caste Attitudes through Large-Scale Analysis of Matrimonial Profiles

Ashwin Rajadesingan,[1] Ramaswami Mahalingam,[2] David Jurgens[1]
[1]School of Information, [2]Department of Psychology
{arajades,ramawasi,jurgens}@umich.edu
University of Michigan, Ann Arbor

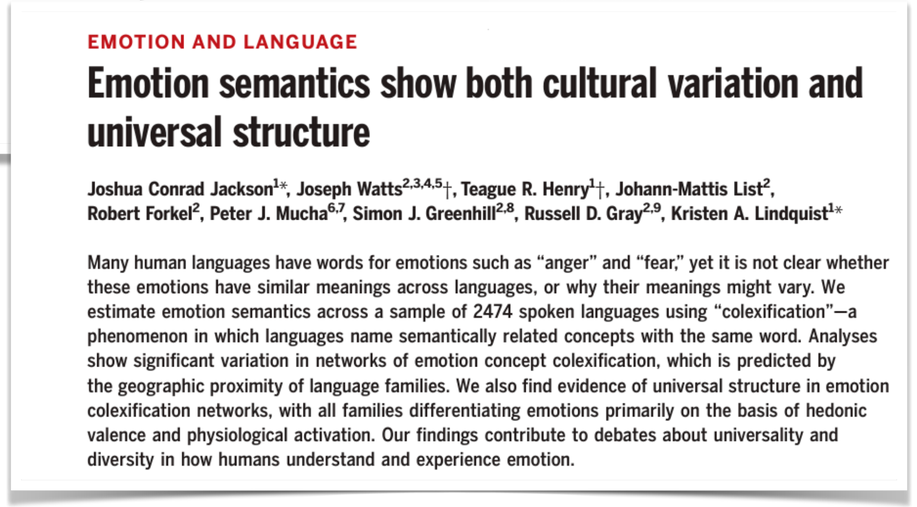


RESEARCH ARTICLE

Universality and diversity in human song

SAMUEL A. MEHR, MANVIR SINGH, DEAN KNOX, DANIEL M. KETTER, DANIEL PICKENS-JONES, S. ATWOOD, CHRISTOPHER LUCAS, NORI JACOBY, ALENA A. EGNER, [...], AND LUKE GLOWACKI +9 authors Authors Info & Affiliations

SCIENCE · 22 Nov 2019 · Vol 366, Issue 6468 · DOI: 10.1126/science.aax0868



EMOTION AND LANGUAGE

Emotion semantics show both cultural variation and universal structure

Joshua Conrad Jackson[1]*, Joseph Watts[2,3,4,5]†, Teague R. Henry[1]†, Johann-Mattis List[2], Robert Forkel[2], Peter J. Mucha[6,7], Simon J. Greenhill[2,8], Russell D. Gray[2,9], Kristen A. Lindquist[1]*

Many human languages have words for emotions such as "anger" and "fear," yet it is not clear whether these emotions have similar meanings across languages, or why their meanings might vary. We estimate emotion semantics across a sample of 2474 spoken languages using "colexification"—a phenomenon in which languages name semantically related concepts with the same word. Analyses show significant variation in networks of emotion concept colexification, which is predicted by the geographic proximity of language families. We also find evidence of universal structure in emotion colexification networks, with all families differentiating emotions primarily on the basis of hedonic valence and physiological activation. Our findings contribute to debates about universality and diversity in how humans understand and experience emotion.

6. Multiple Modality

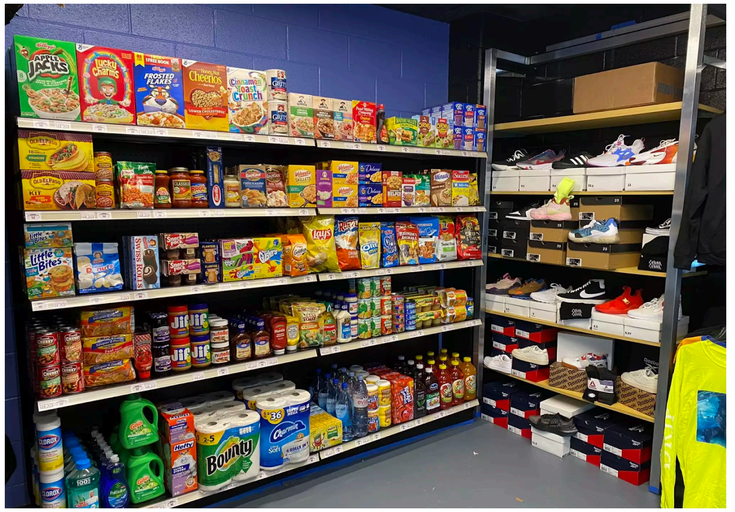
Lots of studies around human behaviors and social interaction in text format

What about other modalities?

- Images
- Audio
- Video
- Touch

Snack

Grocery Store



7 Diverse Research Methods

Do we always need “computational” approaches?

My research is not cool if I didn’t use ChatGPT/Deep Learning

Remember **the value of small data**: individual examples and case studies

Remember **the value of domain users and their insights**

8. Digital Divides?

Inequalities in access to data and the production of knowledge

Privileging of skills required to produce knowledge