

# What I Won't Build

Dr. Rachael Tatman  
Senior Developer Advocate @ Rasa

Widening NLP Keynote, July 5 2020

## How I got here

- BA in Linguistics & English literature (William and Mary)
- PhD in Linguistics (University of Washington)
  - Phonetics & ASR
  - Computational Sociolinguistics
  - Ethics/FAT in NLP
- Data scientist/Developer advocate at Google (Kaggle)
- Developer advocate at Rasa
  - Open source Conversational AI framework

# What's most urgent in NLP?

- Where I started:
  - Language technologies should work equally well for everyone
- Where I am now:
  - Language technologists must proactively mitigate harm from our work



**"Can I minimize differences in accuracy between subgroups" is less important than "should this be built at all"**

**No system is inevitable.  
It must be built.  
It is up to us to decide what  
our labor will support.**

2016

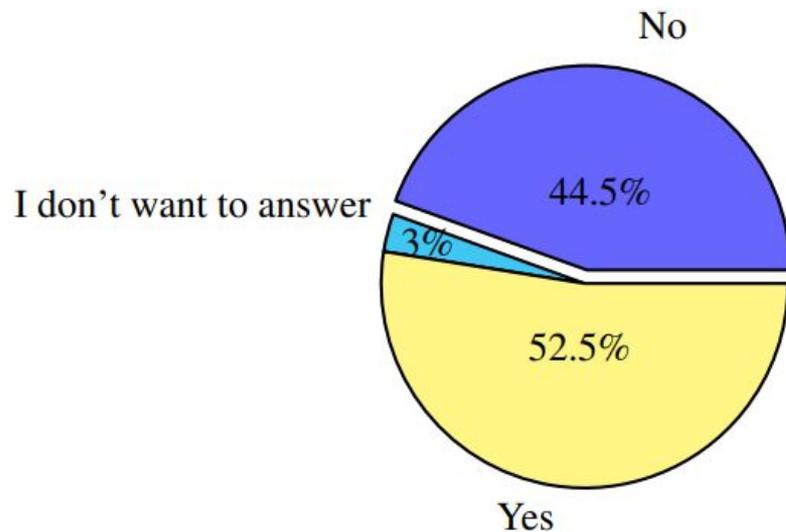


Figure 7: "Do you consider yourself responsible for the usages imagined from the applications/algorithms you create?" (international survey).

# 2016

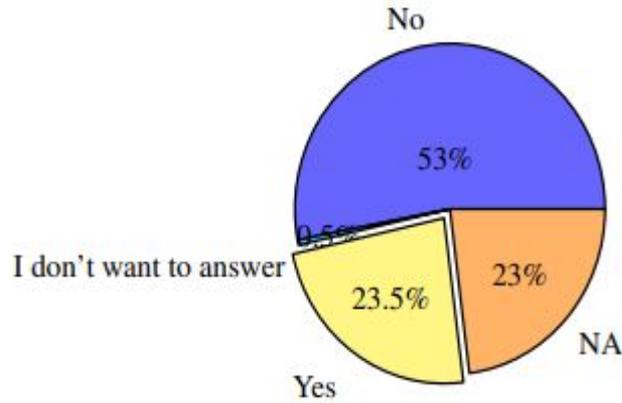


Figure 1: "Have you ever refused a project due to ethical issues?" (international survey).

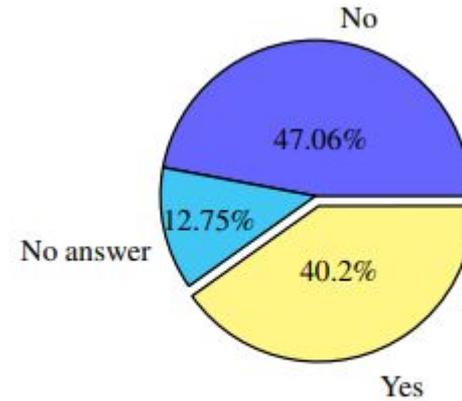


Figure 2: "Avez-vous déjà refusé ou limité un projet pour des raisons éthiques ? / Have you ever refused a project due to ethical issues?" (French survey).

# 2020: The ACL Adopted the ACM Code of Ethics

- Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing
- Avoid harm
- Be honest and trustworthy
- Be fair and take action not to discriminate
- Respect the work required to produce new ideas, inventions, creative works, and computing artifacts.
- Respect privacy
- Honor confidentiality

## 2020: The ACL Adopted the ACM Code of Ethics

- Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing
- **Avoid harm**
- Be honest and trustworthy
- **Be fair and take action not to discriminate**
- Respect the work required to produce new ideas, inventions, creative works, and computing artifacts.
- **Respect privacy**
- Honor confidentiality

# What won't I help build? (My list)

- Surveillance technology
  - “**Peregrine E-mail Analysis Tool**: ingests bulk e-mail information and uses natural language processing (NLP) to help special agents and analysts extract keywords and run analytics”, [2021 U.S. Immigration and Customs Enforcement Budget Overview](#)
- Deceptive technology
  - Yang, Z., & Xu, C. (2019, November). *Read, Attend and Comment: A Deep Architecture for **Automatic News Comment Generation***. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (**EMNLP-IJCNLP**) (pp. 5080-5092).
  - All Rasa assistants identify themselves as bots
- Social category detectors

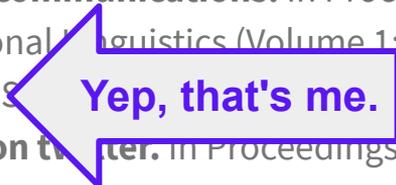
# Social category detectors?

Systems that attempt to infer the social category of a user without the user voluntarily providing that information.

- Preoțiu-Pietro, D., & Ungar, L. (2018, August). **User-level race and ethnicity predictors from twitter text.** In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1534-1545).
- Ardehaly, E. M., & Culotta, A. (2015). **Inferring latent attributes of Twitter users with label regularization.** In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 185-195).
- Volkova, S., Coppersmith, G., & Van Durme, B. (2014, June). **Inferring user political preferences from streaming communications.** In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 186-196).
- Tatman, R., Stewart, L., Paullada, A., & Spiro, E. (2017, August). **Non-lexical features encode political affiliation on twitter.** In Proceedings of the Second Workshop on NLP and Computational Social Science (pp. 63-67).

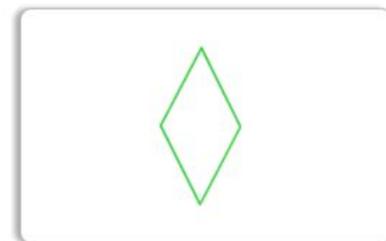
# Social category detectors?

Systems that attempt to infer the social category of a user without the user voluntarily providing that information.

- Preoțiu-Pietro, D., & Ungar, L. (2018, August). **User-level race and ethnicity predictors from twitter text.** In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1534-1545).
- Ardehaly, E. M., & Culotta, A. (2015). **Inferring latent attributes of Twitter users with label regularization.** In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 185-195).
- Volkova, S., Coppersmith, G., & Van Durme, B. (2014, June). **Inferring user political preferences from streaming communications.** In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 186-196).
- Tatman, R., S.  & Spiro, E. (2017, August). **Non-lexical features encode political affiliation on twitter.** In Proceedings of the Second Workshop on NLP and Computational Social Science (pp. 63-67).

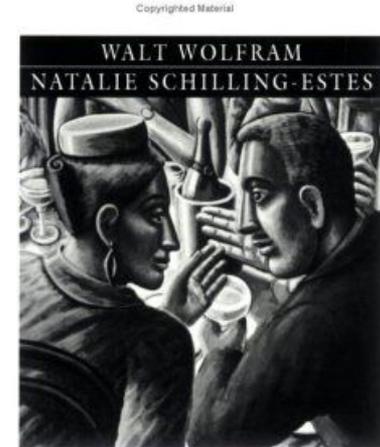
# How to accidentally build a social category detector

1. Build an application that uses language data as input and assigns individuals into some sort of category
2. Don't consider sociolinguistic variation
3. Don't conduct socially-stratified validation
4. Voila!



# How to accidentally build a social category detector

1. All language use is shaped by its social context
2. Many demographic factors are linked to systematic variation in language including:
  - a. Regional Origin
  - b. Age
  - c. Socio-economic status/Social class
  - d. Race/ethnicity



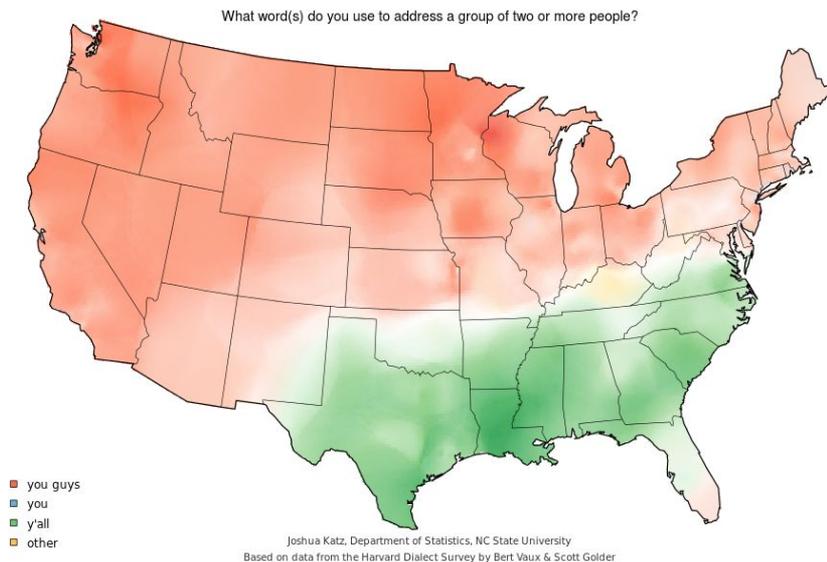
AMERICAN  
ENGLISH

"American English" by Wolfram and Schilling-Estes is a nice introduction

# How to accidentally build a social category detector

Sociolinguistic variation exists at every level of the grammar

- a. Phonetics & sub-lexical features:
  - i. cot/caught merger
  - ii. :) vs. :-)
- b. Lexical variation
  - i. firefly vs. lightning bug
- c. Syntactic variation
  - i. Needs washed
  - ii. We stay home anymore
- d. Semantic variation
  - i. lift, torch, boot



# How to NOT accidentally build a social category detector

- Extracted features > raw text input
- Consider relevant sociolinguistic variation
  - Quickest way = ask a sociolinguist familiar with your user population!
- Do sub-group validation for sociolinguistically-relevant groups
  - Which groups matter? Depends on who your users are.

# Questions to Ask

1. Who benefits from this system existing?
2. Who could be harmed by this system?
3. Can users choose not to interact with this system?
4. Does that system enforce or worsen systemic inequalities?
5. Is this genuinely bettering the world? Is it the best use of your limited time and resources?



# Supporting each other

- Decide ahead of time what your personal priorities are
- As researchers and practitioners we can affect change much faster than any legislature
- Stay informed and educate others about the risks and drawbacks of NLP applications
- Support each other in speaking up, coordinated effort gets results!
  - [“Google reportedly leaving Project Maven military AI program after 2019”](#) due in part to coordinated pressure from employees
  - Google’s choice not to offer facial recognition for surveillance was [due in part to employee activism](#)



## Further Reading

- "The social impact of natural language processing" Hovy & Spruit, ACL 2016
- "Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis?" Leins, Lau & Baldwin, ACL 2020
- "Ethics and Data Science" Loukides, Mason & Patil, 2018
- "[Principles for Accountable Algorithms and a Social Impact Statement for Algorithms](#)" FAT/ML