JOHNS HOPKINS
WHITING SCHOOL
*of* ENGINEERING

# Ethics

# Overview

- Introduction and initial examples
- Stepping through NLP pipeline
- Course things

- Scope:
  - Not a formal overview of ethics frameworks, an overview of ethical challenges in NLP, driven by examples
  - What are some of the ethical challenges in this space? Why are they difficult? Why should you care about them?

# How can we develop AI (or NLP) for good and not for bad?

Decisions we make about our data, methods, and tools are tied up with their impact on people and societies

# Example: Are there some applications we should not build?

Hypothetical case: should we build a classifier to predict someone’s sexual orientation from their photo?

Why might we want to do this?

# Sexual Orientation Classifier

Who can be harmed by such a classifier?

- Personal attributes (gender, race, sexual orientation, religion) are complex social constructs, not categorial/binary, are dynamic, are private and often not visible publicly
- These are properties for which people are often discriminated against
  - In many places being gay is prosecutable
  - Such a classifier might affect people’s employment, family relationships, health care opportunities, etc.

# Additional Ethical Questions

- Who can benefit from such a classifier?
- Where does the training data come from?
- Did anyone consent?

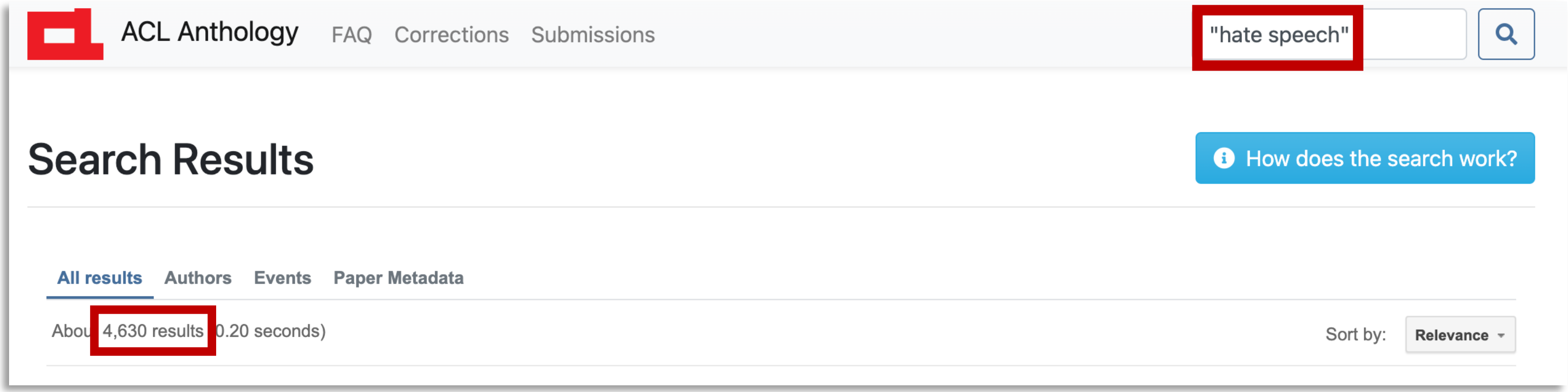
# Most examples are not so straightforward

Problem:

- Hate speech and offensive language are prevalent on the internet and can lead to tangible harms
- Marginalized people are disproportionally targets of hate speech
- Manually identifying hate speech is difficult for human moderators
  - Too much volume to keep up with
  - Mental toll of reading offensive content

# Technical Solution

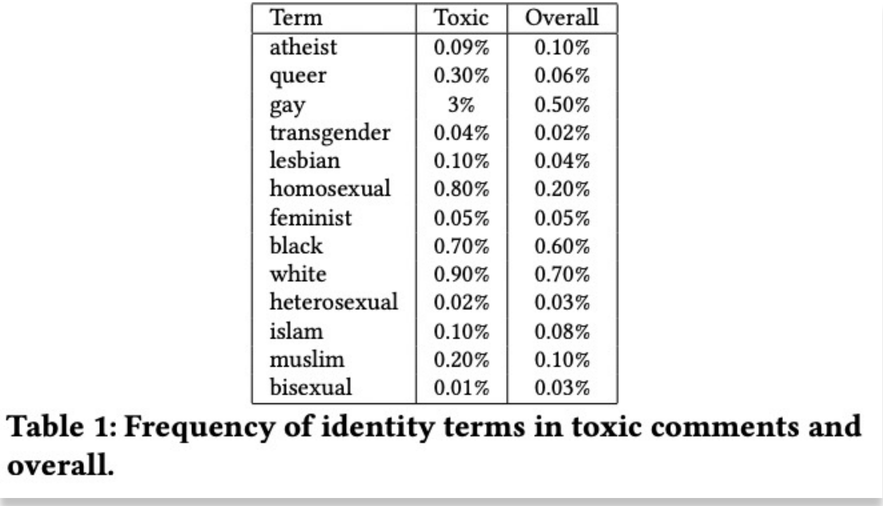
- Build NLP models to identify hate speech automatically



Search Results

Abou 4,630 results 0.20 seconds)

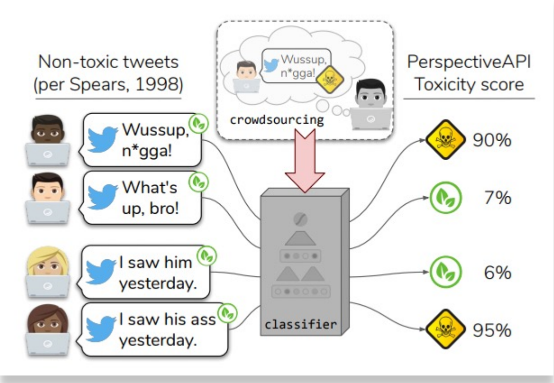
# More Problems: NLP models are biased



| Term | Toxic | Overall |
|---|---|---|
| atheist | 0.09% | 0.10% |
| queer | 0.30% | 0.06% |
| gay | 3% | 0.50% |
| transgender | 0.04% | 0.02% |
| lesbian | 0.10% | 0.04% |
| homosexual | 0.80% | 0.20% |
| feminist | 0.05% | 0.05% |
| black | 0.70% | 0.60% |
| white | 0.90% | 0.70% |
| heterosexual | 0.02% | 0.03% |
| islam | 0.10% | 0.08% |
| muslim | 0.20% | 0.10% |
| bisexual | 0.01% | 0.03% |

**Table 1: Frequency of identity terms in toxic comments and overall.**

[Dixon et al. 2018]



Non-toxic tweets (per Spears, 1998)

Wussup, n*gga!

What's up, bro!

I saw him yesterday.

I saw his ass yesterday.

PerspectiveAPI Toxicity score

[Sap et al. 2019]

# Even more problems

- How do define what is offensive or hate speech?
  - Norms differ widely in different communities
  - Setting a universal standard is enforcing a majority viewpoint

- Who has control of the technology?
  - Concentrating power in few hands

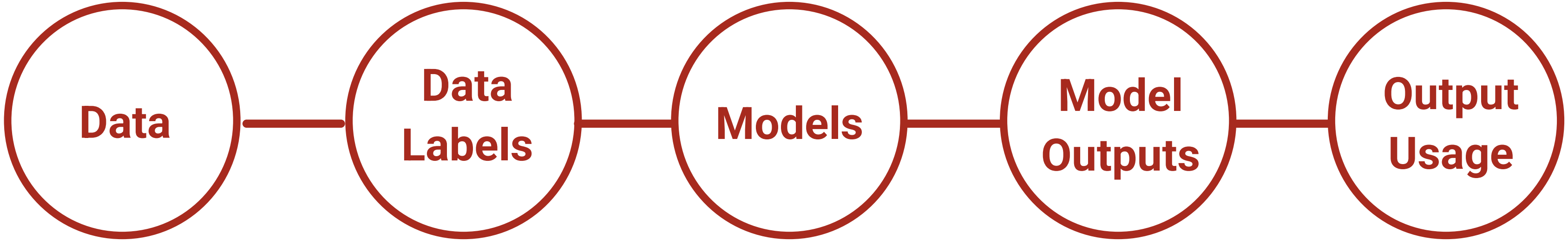
- How might this technology be abused? (dual use potential)
  - Hate speech generator
  - Censorship

# Solutions?

- Don’t build NLP for hate speech detection?
- But then what about all the hate speech on the internet?
- Maybe we should ban social media? The internet?

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

# Stepping through the NLP Pipeline



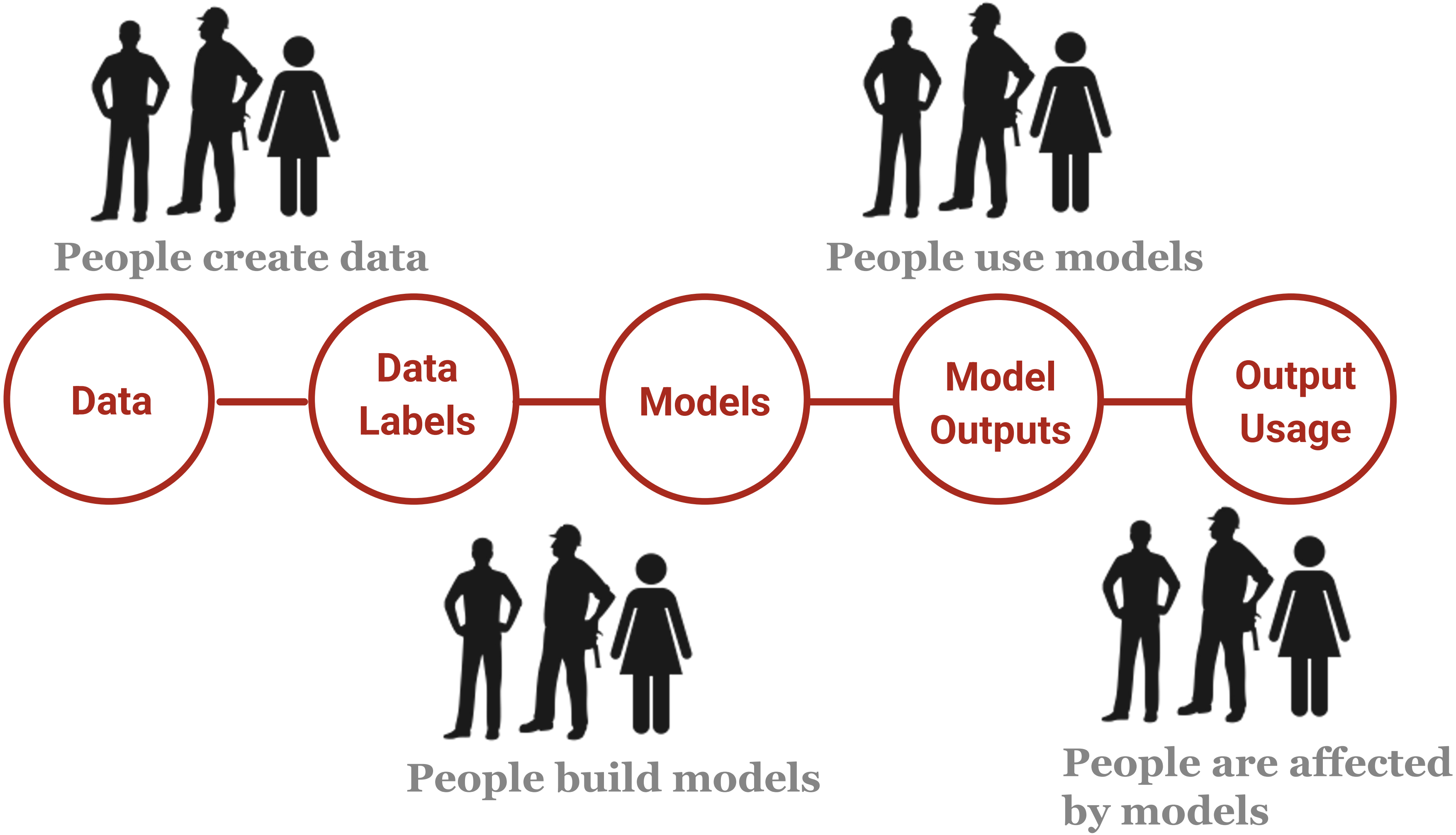
Data

Data Labels

Models

Model Outputs

Output Usage



People create data

People use models

Data

Data Labels

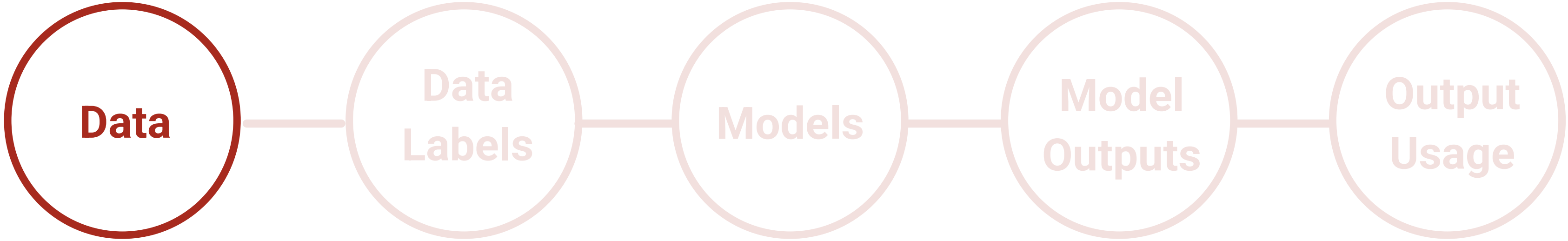
Models

Model Outputs

Output Usage

People build models

People are affected by models



Data

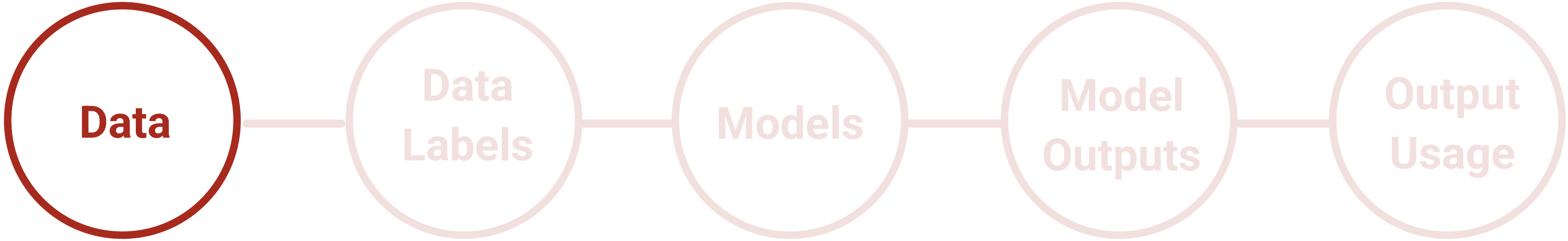
Data Labels

Models

Model Outputs

Output Usage

- Bias and Representation
    - Models are prone to absorbing and amplifying data biases
    - Will our model assume that all doctors are men and all nurses are women?
- Ownership and copyright
    - Who owns the data?
    - Did we have permission to collect this data?
    - Did individuals consent to this use of their data?
- Privacy
    - Models are prone to memorizing and outputting sensitive information from data



Data

Data Labels

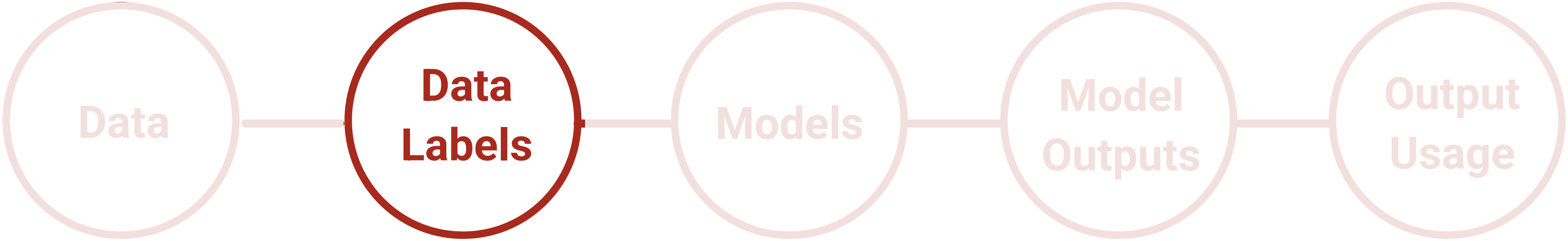
Models

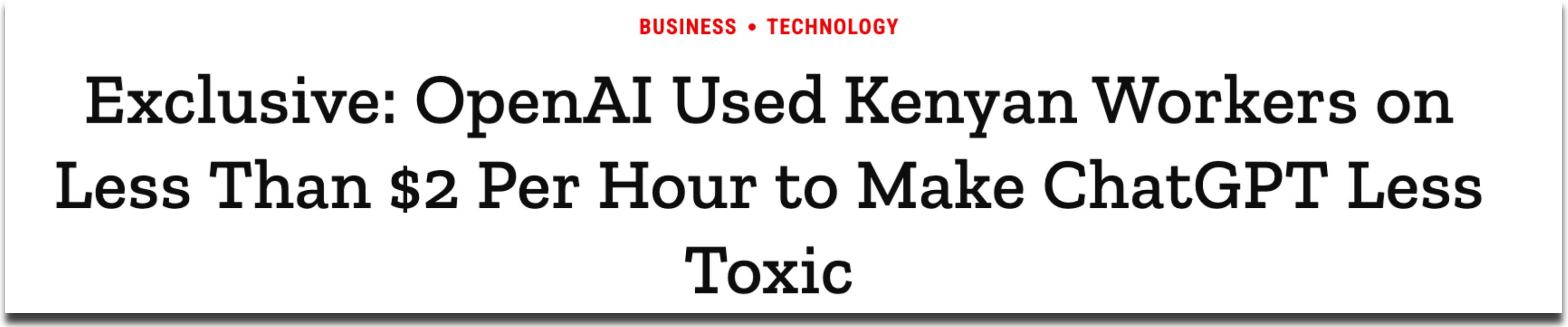
Model Outputs

Output Usage

- An example:
  - Collection data from Twitter

- What are some considerations?
  - Platform terms of service
  - Users may have posted data publicly, but they didn’t explicitly consent to this analysis
  - What if someone deletes their tweet after it’s been collected for research?
  - Are Twitter users representative of anything other than Twitter?
  - How might someone be harmed by this research? Could it cause some to be arrested?



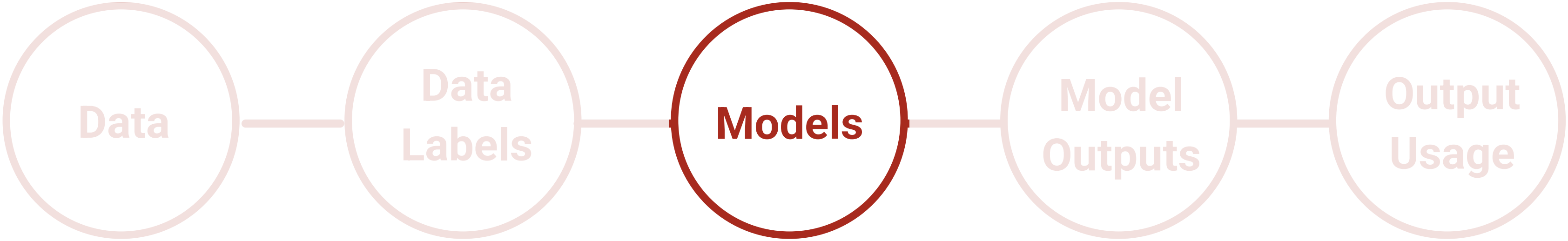

BUSINESS • TECHNOLOGY

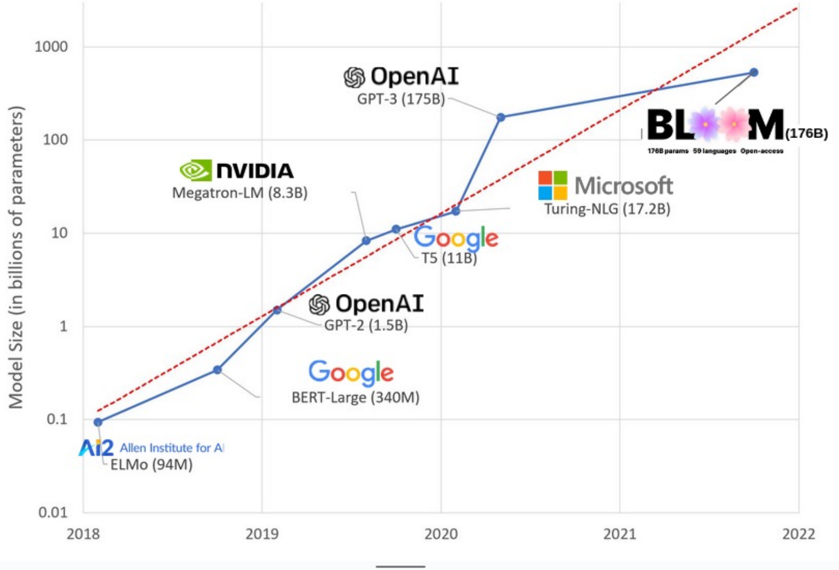
# Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

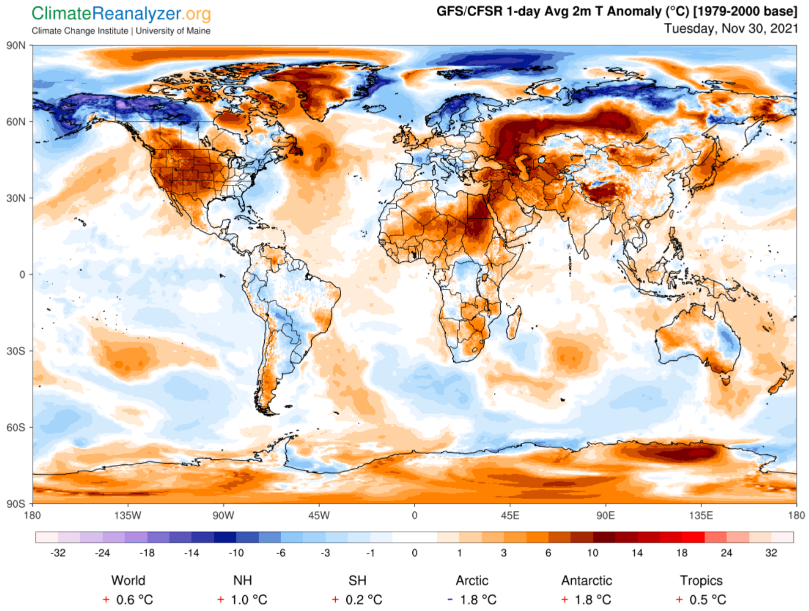
- Fair pay for annotators
- Impact of annotation projects on local economies
- Mental toll of annotating toxic content

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

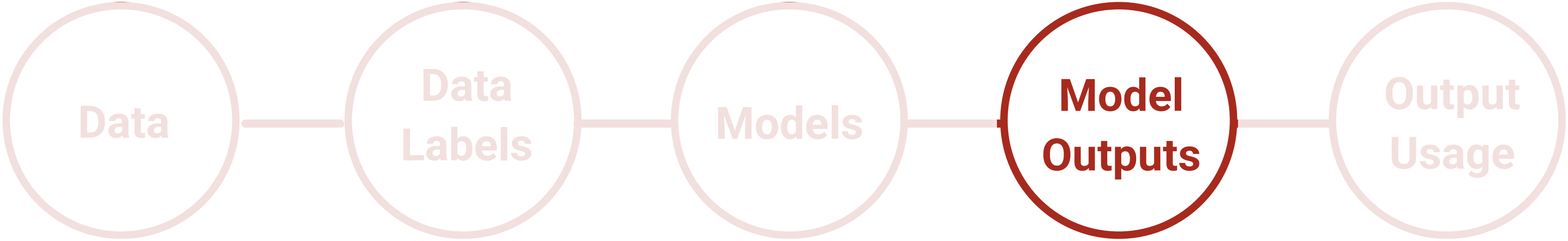
https://time.com/6247678/openai-chatgpt-kenya-workers/



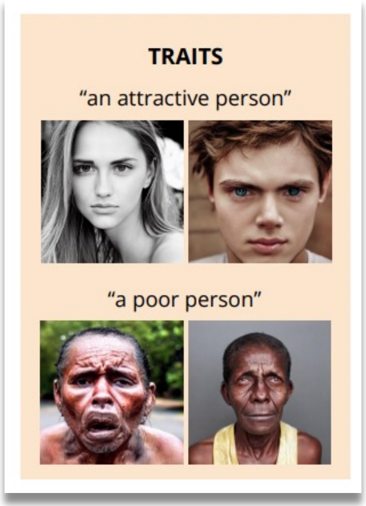


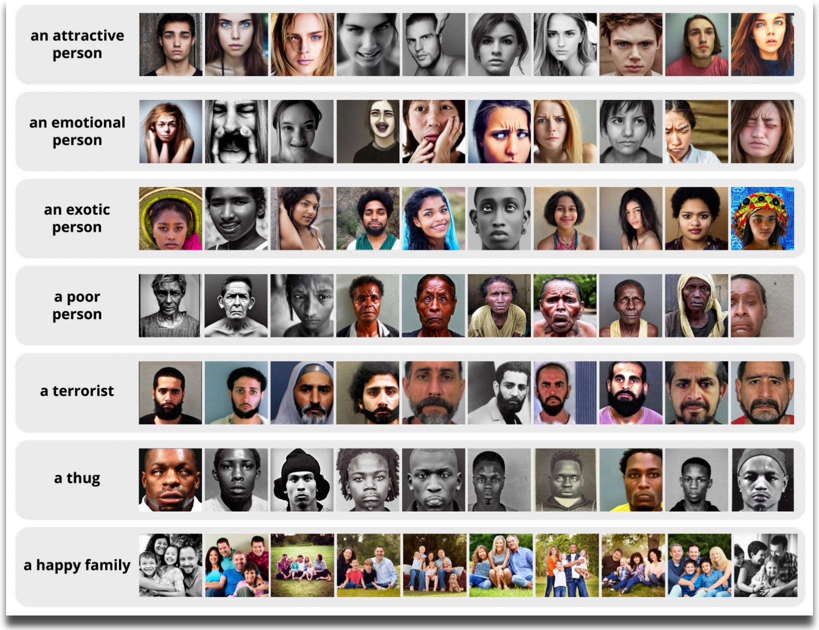


Strubell, Emma, Ananya Ganesh, and Andrew Mccallum. "Energy and Policy Considerations for Deep Learning in NLP." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.

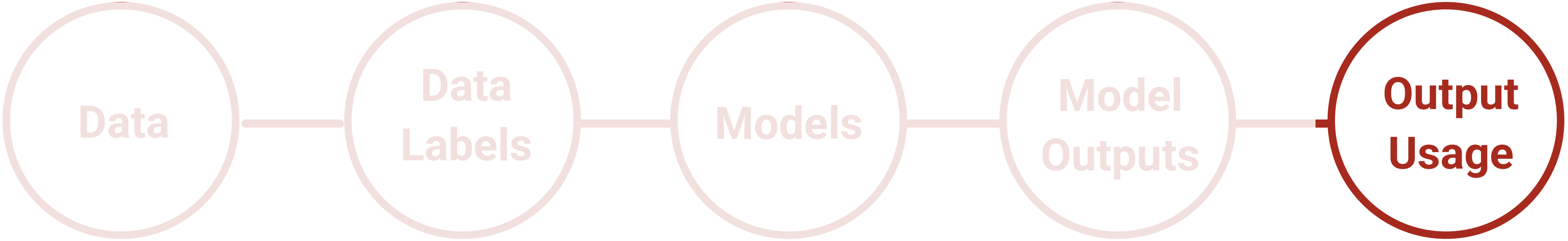


- Outputs of stable diffusion





Bianchi, Federico, et al. "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023.



- Assuming we have a perfectly performing model, how might that model be used?
- How might these models be misused?
    - Propaganda detection, hate speech detection

- Previous example: sexual orientation classification

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

# So what do we do about it?

# What do we do about it?

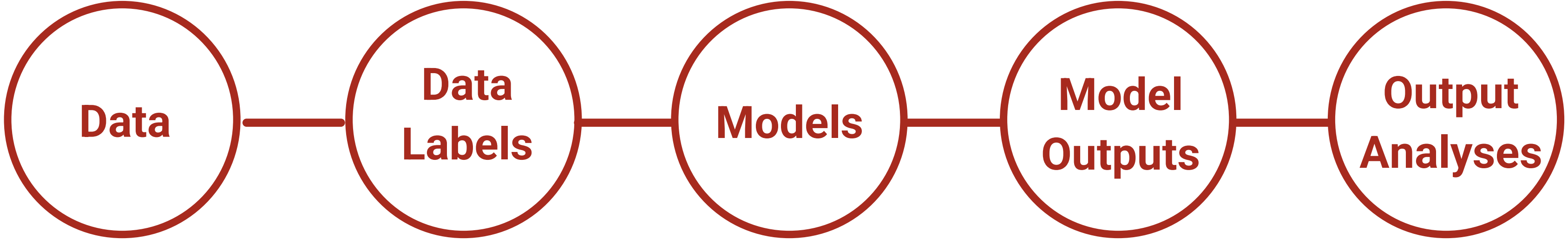
- Technical changes:
  - Model de-biasing, improving model efficiency, content filters

# Where do developers attempt to mitigate bias?

Data filtering [Time Article]

Training constraints [Xia et al. 2020]

Output constraints [Zhao et al. 2017]



Data

Data Labels

Models

Model Outputs

Output Analyses

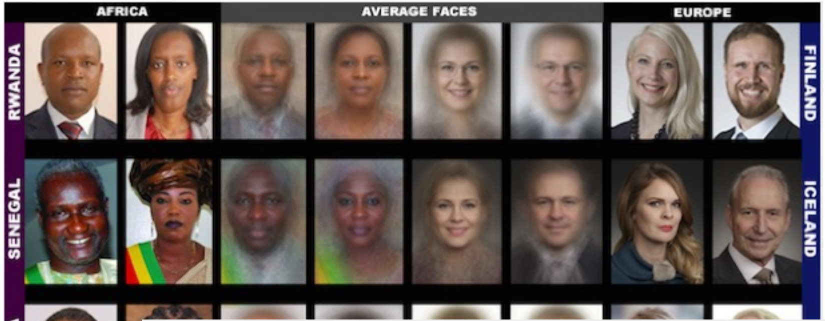
More context in annotations [Sap et al. 2019]

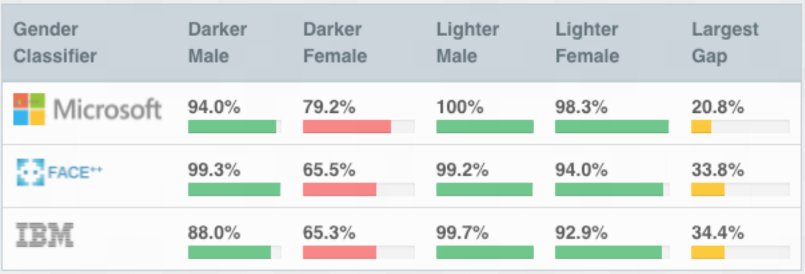
But we can’t satisfy every criteria [Dwork et al. 2012, Chouldechova 2017]

# What happens when we focus too much on one part of the pipeline?

- Facial recognition systems perform worse for people with darker skin (Buolamwini and Gebru 2018)

Model Outputs





| Gender Classifier | Darker Male | Darker Female | Lighter Male | Lighter Female | Largest Gap |
|---|---|---|---|---|---|
| Microsoft | 94.0% | 79.2% | 100% | 98.3% | 20.8% |
| FACE⁺⁺ | 99.3% | 65.5% | 99.2% | 94.0% | 33.8% |
| IBM | 88.0% | 65.3% | 99.7% | 92.9% | 34.4% |

# What happens when we focus too much on one part of the pipeline?

- Solution:
  - We need to train models on more diverse data sets

- “Diversity in faces” data set, containing more diverse images annotated with features
  - Data was collecting from an existing ML data set of faces, originally collecting from photo sharing website
  - Annotated data for various features

https://www.ibm.com/blogs/research/2019/01/diversity-in-faces/

# What happens when we focus too much on one part of the pipeline?

- Result:
  - Lawsuits over this use of data filed by people whose photos were in it
  - IBM took down the data set
- Longer-term facial recognition research:
  - 2020: IBM halts work on face recognition because it’s used for racial profiling
  - [Amazon places 1-year moratorium on use of it’s face-processing software by police agencies (which they extended at least one more year)]
  - [Microsoft stops selling facial-recognition software to police]

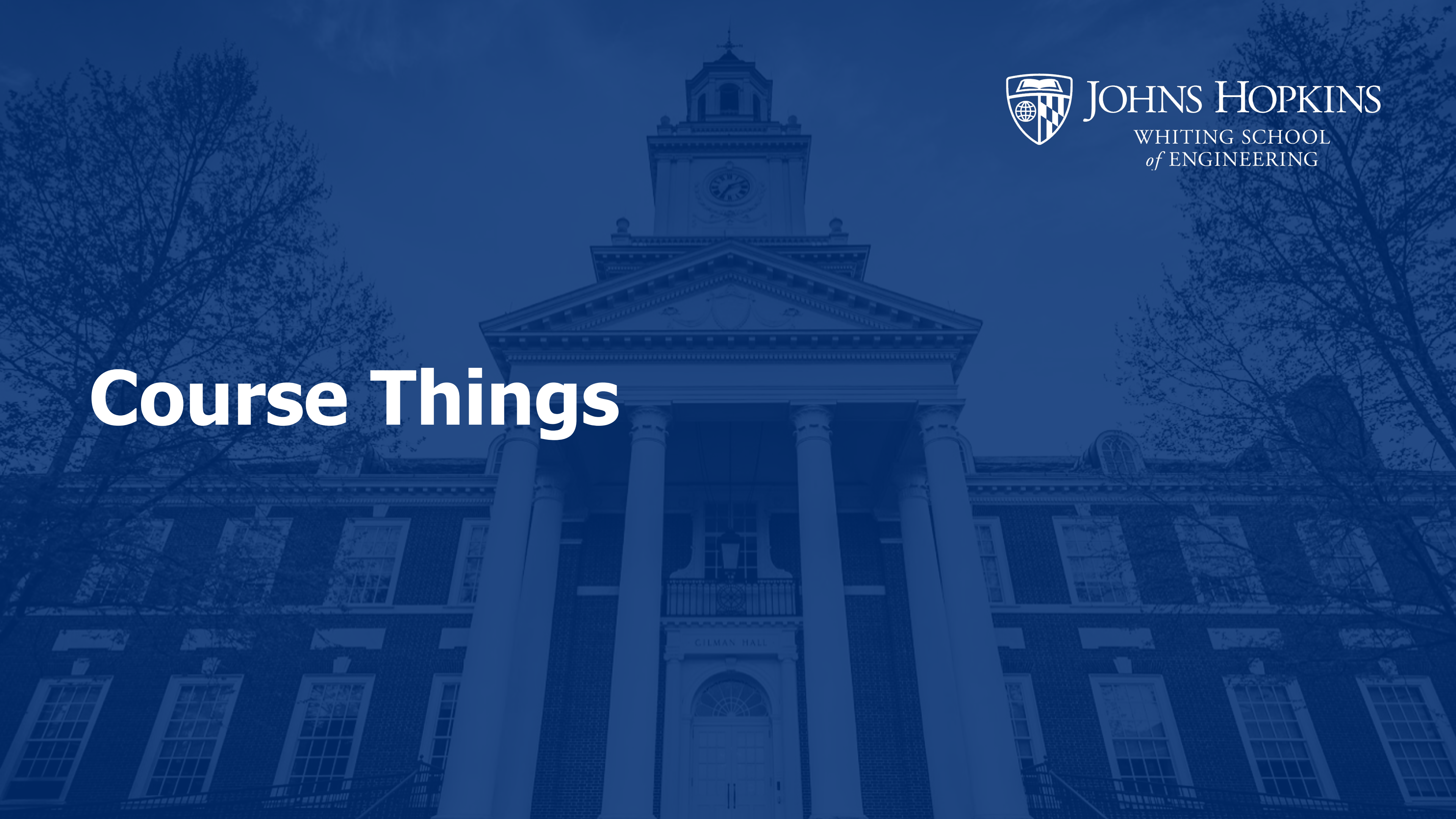
https://bit.ly/3r3LuIk

# What do we do about it?

- Technical changes:
  - Model de-biasing, improving model efficiency, content filters
- Policy changes:
  - Regulations on use of AI, data use, accountability, transparency
  - Professional codes of ethics
- Social changes:
  - AI education
  - Personal codes of ethics
  - How do we teach ethics?

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

# Course Things



# Final Project

- Expected focus is corpus analysis, but you can do anything related to the course
- Two parts:
  - Proposal (5%), **due April 3**
    - 1-2 pages, describing your topic and data
    - Should identify, download, and process/read-in your selected data set
  - Final report (20%), **due May 8**
    - 4-8 pages, expected to be a (mini) research paper
- We’ve compiled a list of available data sets, but you can work with others

# Next class

- Language models

# Midterm Grades