# CAUSAL INFERENCE IN NATURAL LANGUAGE PROCESSING FOR ECONOMICS

## WEEK 5: CAUSAL INFERENCE –

## DEBIASING TECHNIQUES WITH TEXT AS DATA


SUMMER SEMESTER 22

DR. HUYEN NGUYEN

HTTP://HUYENTTNGUYEN.COM

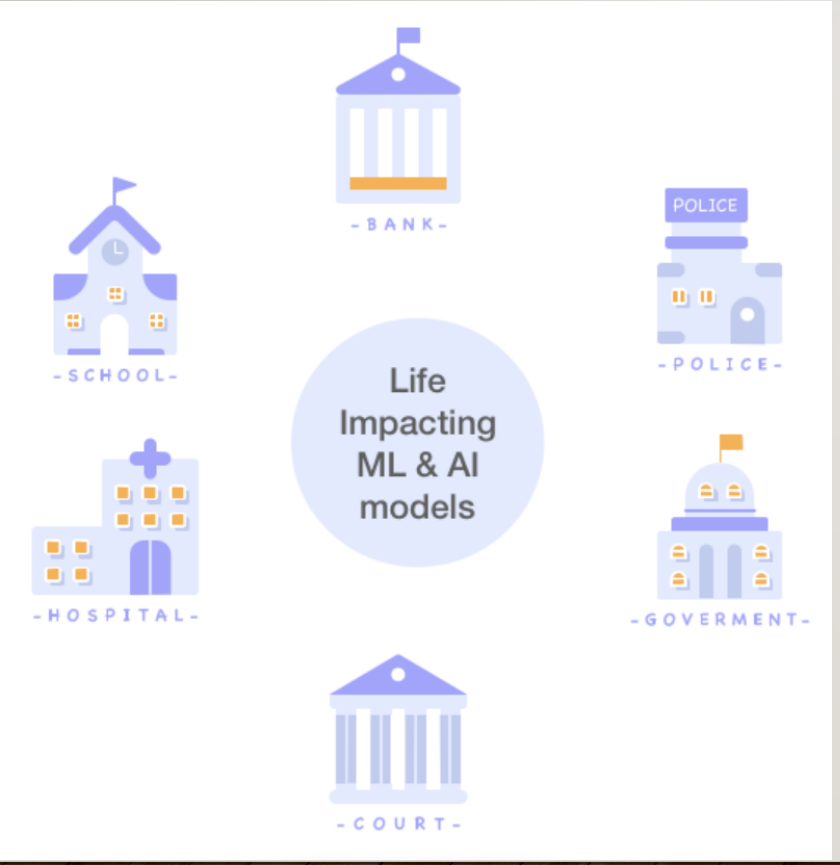
# LOGISTICS

- Assigned papers to groups available on the Google Sheet https://docs.google.com/spreadsheets/d/1EEUJp1Srz05BwEUH2u3CbYwMykmqEUhZdn2SqrmeaRs/edit#gid=0.
- Next week: Presentations start☺
  - Paper presentation: (G2) Sakshi, Tina and Zuhair
  - Response presentation (G7) Felix & Patrick
  - Everyone else:
  - Class structure: 1hour of lecture/codes -> 15' break -> 1 hour of presentations & discussions
- Any questions?

# AGENDA

- Biases in data
  - Issues with models
  - Different types of biases
  - Consequences
  - Text as data: Biases and challenges
- Coding exercise: Controlling for Text in Causal Inference with Double Machine Learning

# CURRENT STATE OF THE WORLD…




- Growing number of applications of machine learning in domains with a significant and direct impact on human life.
- Inevitably, these models end up learning societal biases and prejudices against certain classes or groups of humans.
- Why?

# BIASES – TRAINING DATA SETS

- **Historical biases**: Certain classes of people were biased against historically, and this can be reflected in the dataset
  - E.g: Black & Hispanic people are (historically) sentenced to prisons significantly more than White people in the US; men are significantly more represented in banking & tech jobs

- **Sampling biases**: The data could be imbalanced across classes or groups of people
  - E.g: policing and arrest records across different neighborhoods, university admission acceptance rates across races

# CONSEQUENCES OF BIASES (1)

Google Photos:

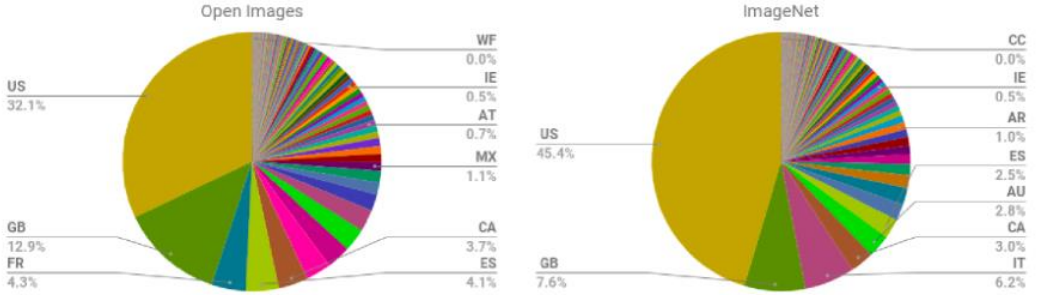



Figure 1: Fraction of Open Images and ImageNet images from each country. In both data sets, top represented locations include the US and Great Britain. Countries are represented by their two-letter ISO country codes.

Reference: Shankar et al., 2017

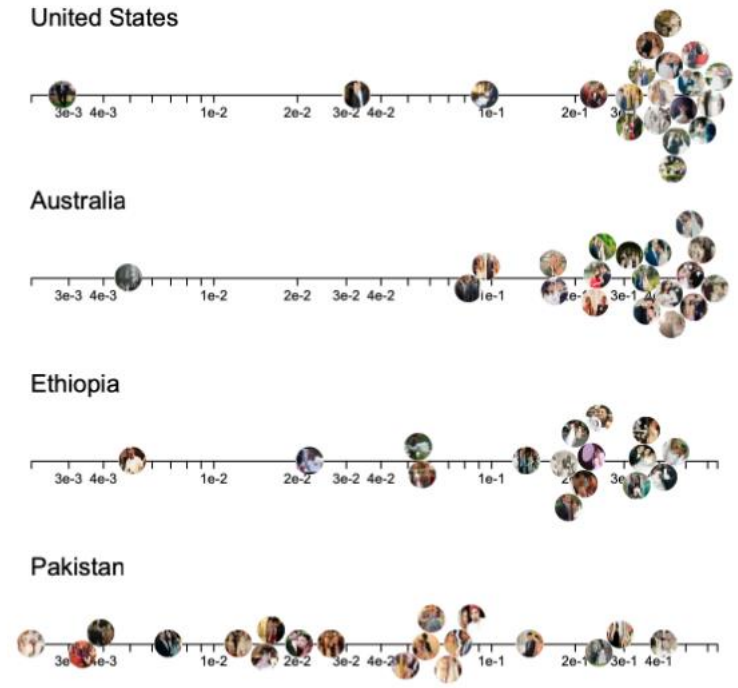



Figure 5: Photos of bridegrooms from different countries aligned by the log-likelihood that the classifier trained on Open Images assigns to the bridegroom class. Images from Ethiopia and Pakistan are not classified as consistently as images from the United States and Australia.

# CONSEQUENCES OF BIASES (1)

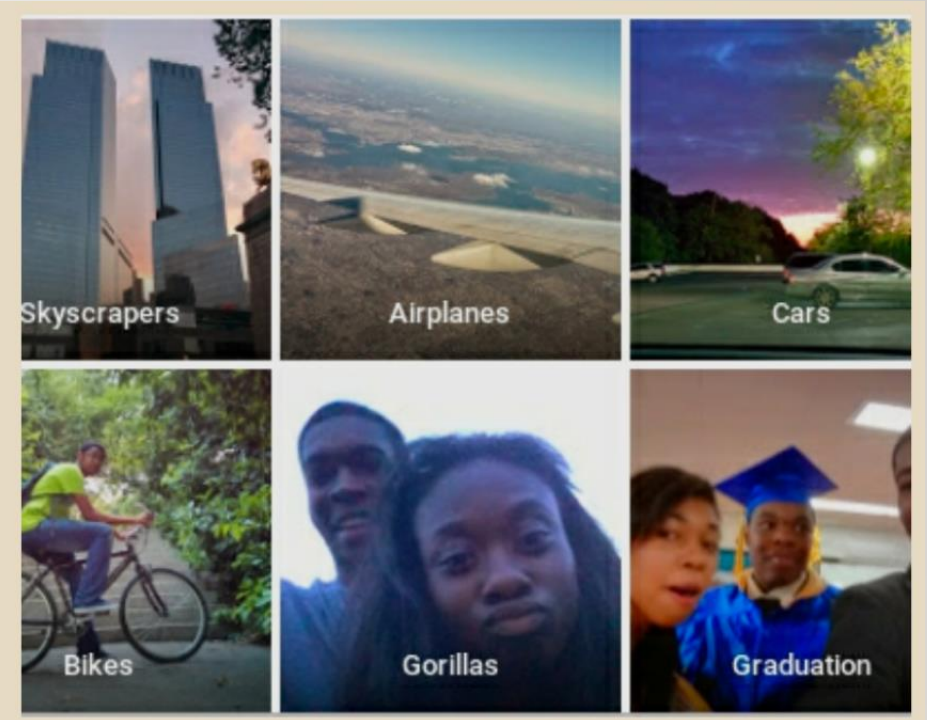
Google

Autocategorized photos show racial bias

in computer vision

Check the following TED talk by MIT grad student Joy Buolamwini on black biases in photos:

https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms?language=en



# CONSEQUENCES OF BIASES (2)

Flawed criminological software : COMPAS PRO PUBLICA algorithm used for sentencing and bail decisions in the US…

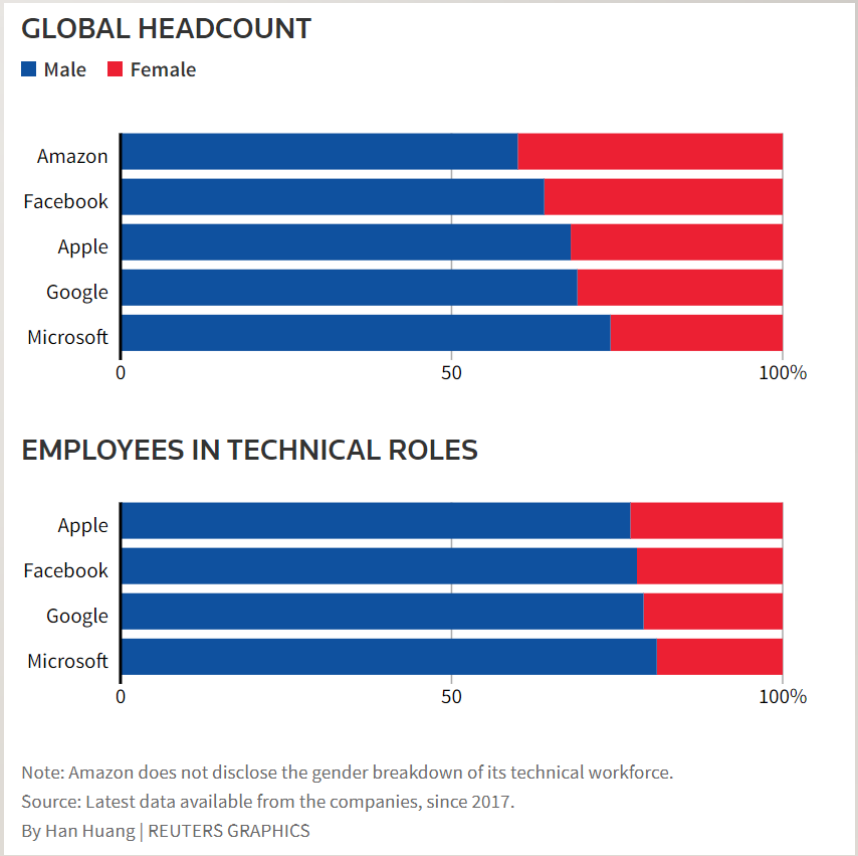
| Prediction Fails Differently for Black Defendants | | |
|---|---|---|
| | WHITE | AFRICAN AMERICAN |
| Labeled Higher Risk, But Didn't Re-Offend | 23.5% | 44.9% |
| Labeled Lower Risk, Yet Did Re-Offend | 47.7% | 28.0% |

# CONSEQUENCES OF BIASES (2) (CONT)

Read more here: [MIT Technological Review Inspecting Algorithms for Biases]


# CONSEQUENCES OF BIASES (3)

- Amazon sexist hiring algorithm:
- 500 computer models focused on specific job functions and locations.
- They taught each to recognize some 50,000 terms that showed up on past candidates' resumes.
- The algorithms learned to assign little significance to skills that were common across IT applicants, e.g: computer coding ability.



# CONSEQUENCES OF BIASES (4)

- 2019 Facebook ad: advertisers deliberately target adverts according to gender, race, and religion.
- Job adverts for roles in nursing or secretarial work were suggested primarily to women, whereas job ads for janitors and taxi drivers had been shown to a higher number of men, especially those from minority backgrounds.
- The algorithm learned that ads for real estate were likely to attain better engagement stats when shown to white people, → not shown to other minority groups.


# WORD EMBEDDINGS & BIASES

- Algorithms learn labels from the corpus (text) of data, which is generally generated from various sources like Wikipedia, different news websites and magazines.
- **Word Embeddings** can reflect bias based on gender, ethnicity, age, sexual orientation and other things, while training the model.
  - E.g: For a feature 'Woman', which determines whether the word present is feminine (1) or masculine (-1), the term "Doctor" may give a score of -0.6 → doctors are more represented by man, rather than a woman (!)
  - For a feature 'Criminal', the word "Negro" achieves a score of 0.7 → Criminals are more represented by Black people (!)
- Biases in the word embedding are in fact closely aligned with social conception of gender stereotype.

# WORD EMBEDDING BIAS EXAMPLES

| Gender stereotype *she-he* analogies. | | |
|---|---|---|
| sewing-carpentry | register-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 | hairdresser-barber |
| **Gender appropriate *she-he* analogies.** | | |
| queen-king | sister-brother | mother-father |
| waitress-waiter | ovarian cancer-prostate cancer | convent-monastery |

Analogy example

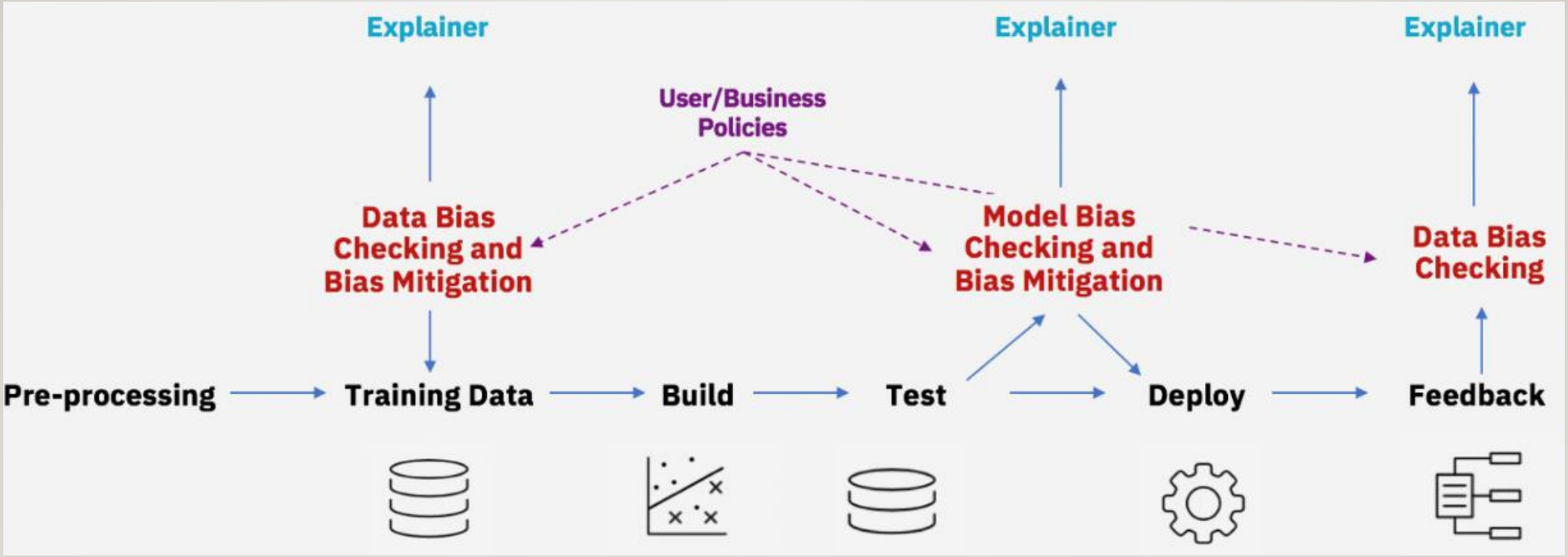
| ***softball* extreme** | **gender portion** | **after debiasing** |
|---|---|---|
| 1. pitcher | -1% | 1. pitcher |
| 2. bookkeeper | 20% | 2. infielder |
| 3. receptionist | 67% | 3. major leaguer |
| 4. registered nurse | 29% | 4. bookkeeper |
| 5. waitress | 35% | 5. investigator |
| ***football* extreme** | **gender portion** | **after debiasing** |
| 1. footballer | 2% | 1. footballer |
| 2. businessman | 31% | 2. cleric |
| 3. pundit | 10% | 3. vice chancellor |
| 4. maestro | 42% | 4. lecturer |
| 5. cleric | 2% | 5. midfielder |

Indirect bias example

…both trained on the original Word2vec NEWS embeddings.

To refresh your understanding on word embeddings, check Lecture 9 of Text Analysis class

# DEBIASING PROCEDURE



# DEBIASING CHALLENGES

- Bias measures are **application dependent** and **context-dependent**.
- Most debiasing algorithms are **measure-specific**.
- Debiasing may require retraining.
- Training can be quite expensive in many real-world scenarios where the same model might be used in different jurisdictions that have different fairness requirements.

# EXERCISE TIME – DEBIASING TECHNIQUE WITH TEXT ☺

- Open your Google Colab.
- Go to Github's course channel →Exercise folder to access these files W5…..