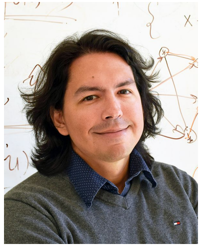
Bias and Fairness in NLP

Margaret Mitchell
Google Brain

Kai-Wei Chang
UCLA



Vicente Ordóñez Román
University of Virginia

Vinodkumar Prabhakaran
Google Brain

Tutorial Outline

- Part 1: Cognitive Biases / Data Biases / Bias laundering
- Part 2: Bias in NLP and Mitigation Approaches
- Part 3: Building Fair and Robust Representations for Vision and Language
- Part 4: Conclusion and Discussion

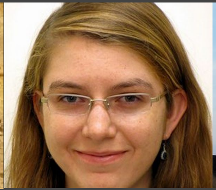
"Bias Laundering"
Cognitive Biases, Data Biases, and ML

Vinodkumar Prabhakaran
Google Brain

Margaret Mitchell
Google Brain

Andrew Zaldivar

Emily Denton



Simone Wu

Parker Barnes

Lucy Vasserman

Ben Hutchinson

Elena Spitzer

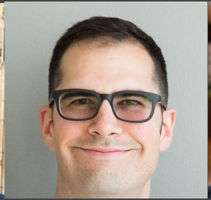
Deb Raji

Timnit Gebru

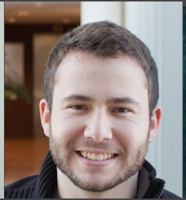
Adrian Benton

Brian Zhang

Dirk Hovy



Josh Lovejoy



Alex Beutel

Blake Lemoine

Hee Jung Ryu

Hartwig Adam

Blaise Agüera y Arcas

What’s in this tutorial

- Motivation for Fairness research in NLP
- How and why NLP models may be unfair
- Various types of NLP fairness issues and mitigation approaches
- What can/should we do?

What's NOT in this tutorial

- Definitive answers to fairness/ethical questions
- Prescriptive solutions to fix ML/NLP (un)fairness

What do you see?

What do you see?

- Bananas

What do you see?

- Bananas
- Stickers

What do you see?

- Bananas
- Stickers
- Dole Bananas

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas

What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas

...We don't tend to say

Yellow Bananas

What do you see?

Green Bananas

Unripe Bananas

What do you see?

Ripe Bananas

Bananas with spots

What do you see?

Yellow Bananas

Yellow is prototypical for bananas

Prototype Theory

One purpose of categorization is to **reduce the infinite differences** among stimuli **to** behaviourally and **cognitively usable proportions**

There may be some central, prototypical notions of items that arise from stored typical properties for an object category (Rosch, 1975)

May also store exemplars (Wu & Barsalou, 2009)

Fruit

Bananas
"Basic Level"

Unripe Bananas, Cavendish Bananas

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

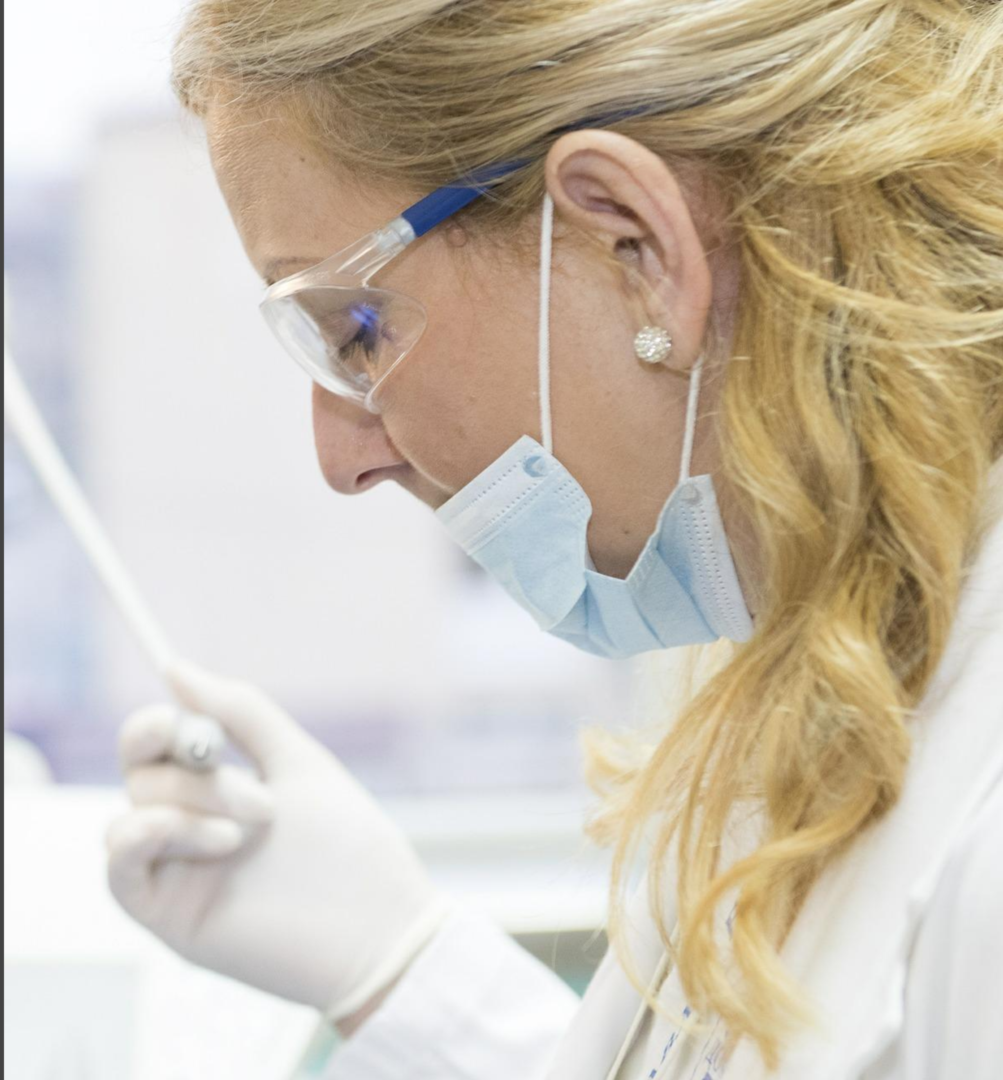
The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?

A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?



A man and his son are in a terrible accident and are rushed to the hospital in critical care.

The doctor looks at the boy and exclaims "I can't operate on this boy, he's my son!"

How could this be?

“Female doctor”


“Doctor”


“Female doctor”

The majority of test subjects overlooked the possibility that the doctor is a she - including men, women, and self-described feminists.

Wapman & Belle, Boston University

World learning from text

Gordon and Van Durme, 2013

Word	Frequency in corpus
"spoke"	11,577,917
"laughed"	3,904,519
"murdered"	2,834,529
"inhaled"	984,613
"breathed"	725,034
"hugged"	610,040
"blinked"	390,692
"exhale"	168,985

World learning from text

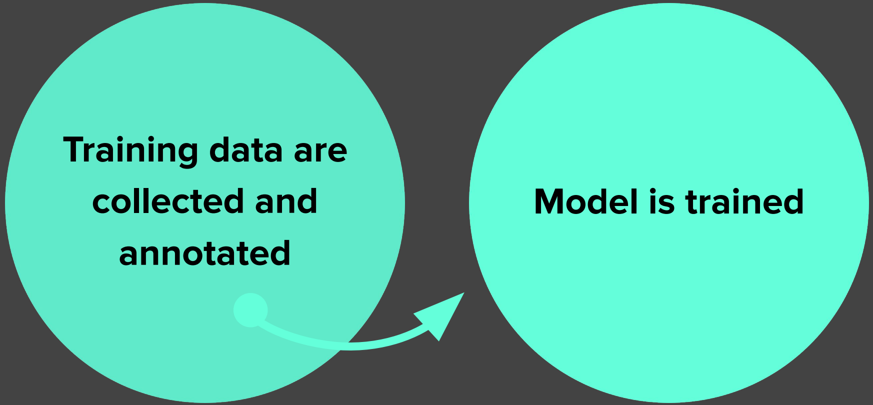
Gordon and Van Durme, 2013

Word	Frequency in corpus
“spoke”	11,577,917
“laughed”	3,904,519
“murdered”	2,834,529
“inhaled”	984,613
“breathed”	725,034
“hugged”	610,040
“blinked”	390,692
“exhale”	168,985

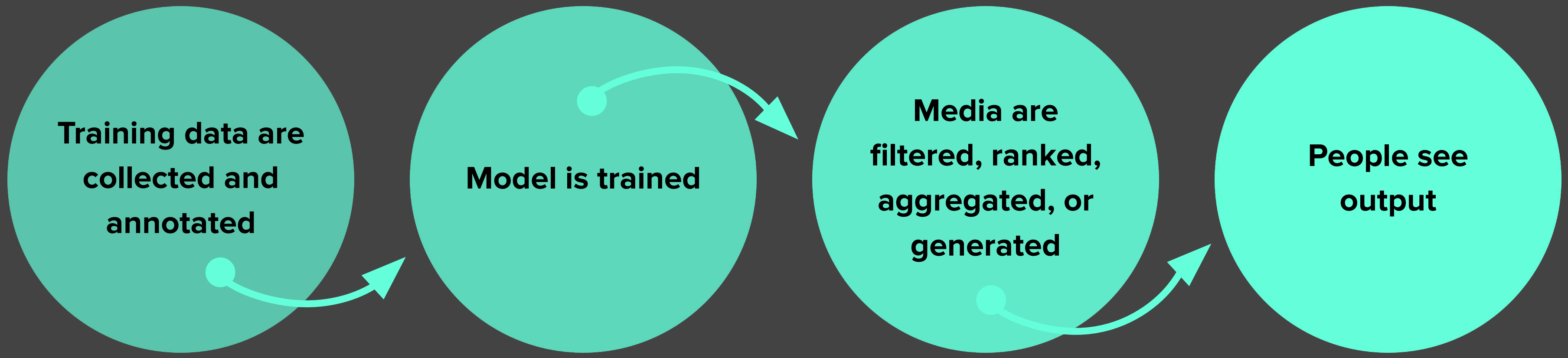
Human Reporting Bias

The **frequency** with which **people write** about actions, outcomes, or properties is **not a reflection of real-world frequencies** or the degree to which a property is characteristic of a class of individuals

Training data are collected and annotated


Training data are collected and annotated
Model is trained


Training data are collected and annotated
Model is trained
Media are filtered, ranked, aggregated, or generated


Training data are collected and annotated
Model is trained
Media are filtered, ranked, aggregated, or generated
People see output

Human Biases in Data

Reporting bias
Selection bias
Overgeneralization
Out-group homogeneity bias

Stereotypical bias
Historical unfairness
Implicit associations
Implicit stereotypes
Prejudice

Group attribution error
Halo effect

Training data are collected and annotated

Training data are collected and annotated

Human Biases in Data

- Reporting bias
- Selection bias
- Overgeneralization
- Out-group homogeneity bias
- Stereotypical bias
- Historical unfairness
- Implicit associations
- Implicit stereotypes
- Prejudice
- Group attribution error
- Halo effect

Human Biases in Collection and Annotation

- Sampling error
- Non-sampling error
- Insensitivity to sample size
- Correspondence bias
- In-group bias
- Bias blind spot
- Confirmation bias
- Subjective validation
- Experimenter’s bias
- Choice-supportive bias
- Neglect of probability
- Anecdotal fallacy
- Illusion of validity

Data

Interpretation

Reporting bias: What people share is not a reflection of real-world frequencies

Selection Bias: Selection does not reflect a random sample

Out-group homogeneity bias: People tend to see outgroup members as more alike than ingroup members when comparing attitudes, values, personality traits, and other characteristics

Confirmation bias: The tendency to search for, interpret, favor, and recall information in a way that confirms one's preexisting beliefs or hypotheses

Overgeneralization: Coming to conclusion based on information that is too general and/or not specific enough

Correlation fallacy: Confusing correlation with causation

Automation bias: Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation

More at: https://developers.google.com/machine-learning/glossary/

Biases in Data

Biases in Data

Selection Bias: Selection does not reflect a random sample

World Englishes

Is the data we use to train our English NLP models representative of all the Englishes out there?

Biases in Data

Selection Bias: Selection does not reflect a random sample

- Men are over-represented in web-based news articles

  (Jia, Lansdall-Welfare, and Cristianini 2015)

- Men are over-represented in twitter conversations

  (Garcia, Weber, and Garimella 2014)

- Gender bias in Wikipedia and Britannica

  (Reagle & Rhuee 2011)

Biases in Data

Selection Bias: Selection does not reflect a random sample

Map of Amazon Mechanical Turk Workers

Biases in Data

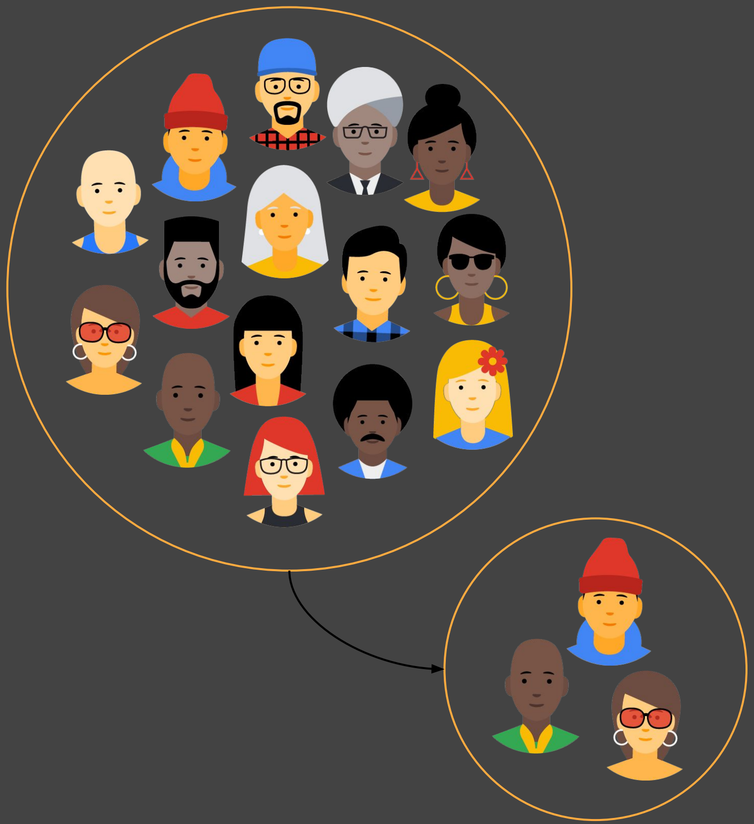
Out-group homogeneity bias: Tendency to see outgroup members as more alike than ingroup members





Biases in Data → Biased Data Representation

It's possible that you have an appropriate amount of data for every group you can think of but that some groups are represented less positively than others.



Biases in Data → Biased Labels

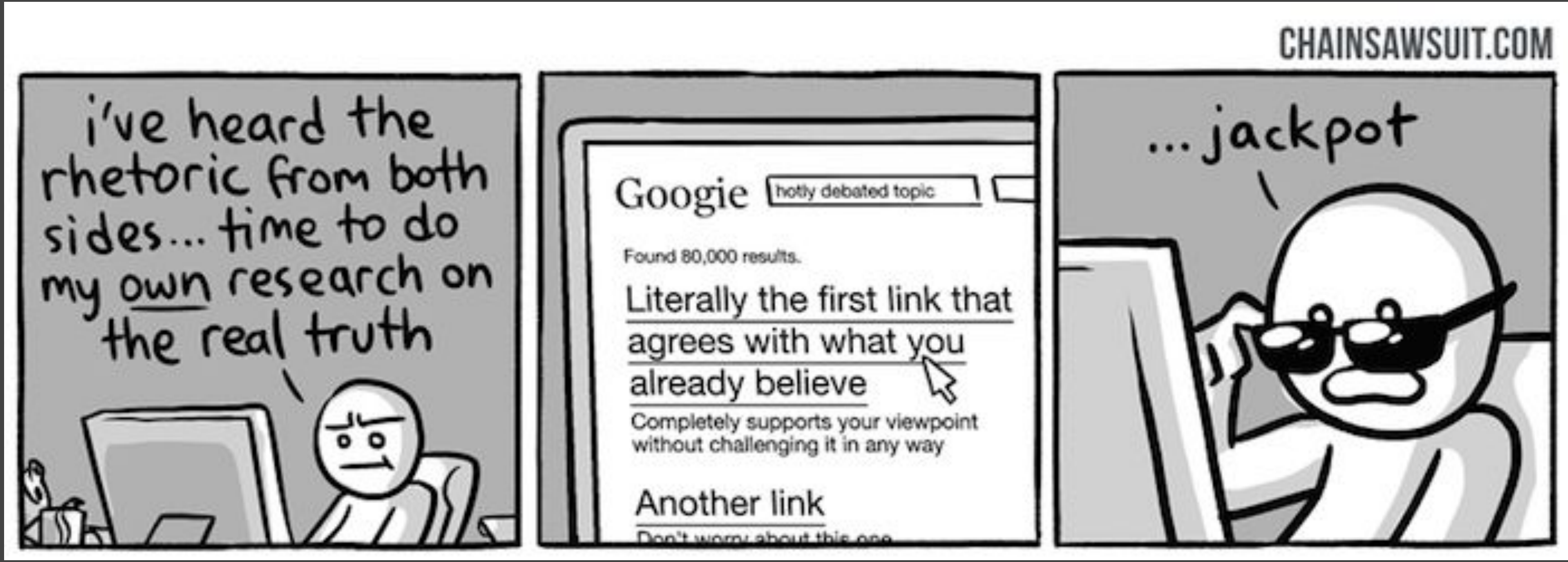
Annotations in your dataset will reflect the worldviews of your annotators.

https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

Biases in Interpretation

Biases in Interpretation

Confirmation bias: The tendency to search for, interpret, favor, recall information in a way that confirms preexisting beliefs



CREDIT

Biases in Interpretation

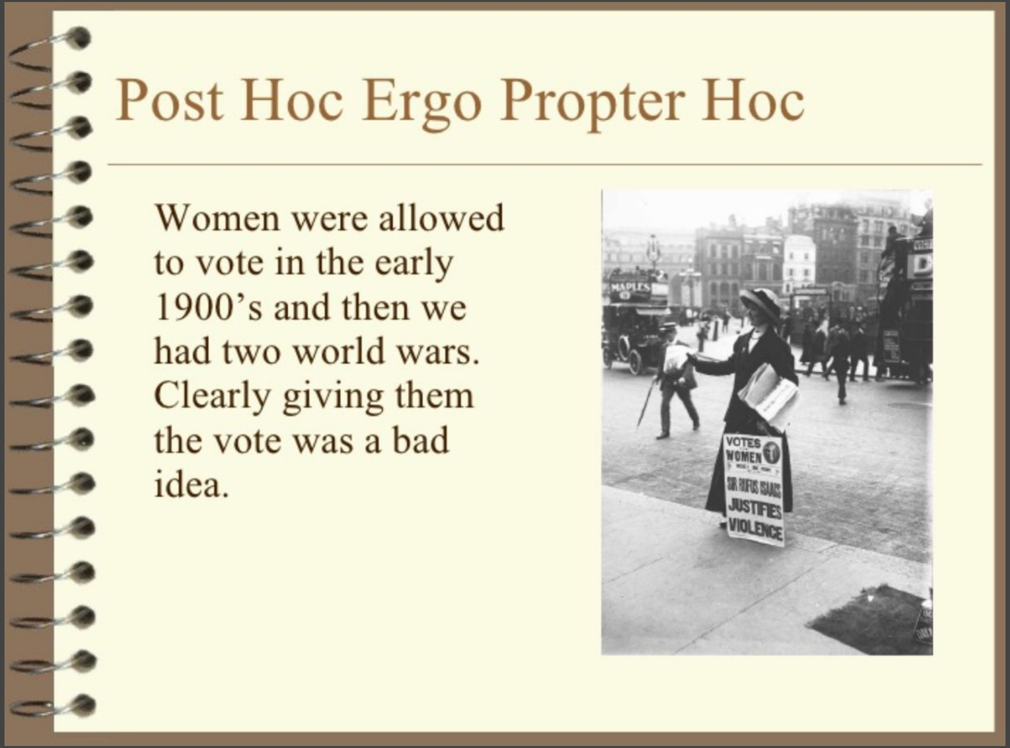
Overgeneralization: Coming to conclusion based on information that is too general and/or not specific enough (related: **overfitting**)

CREDIT

Sidney Harris

Biases in Interpretation

Correlation fallacy: Confusing correlation with causation



CREDIT

© mollysdad - Slideshare - Introduction to Logical Fallacies

Biases in Interpretation

Automation bias: Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation

CREDIT

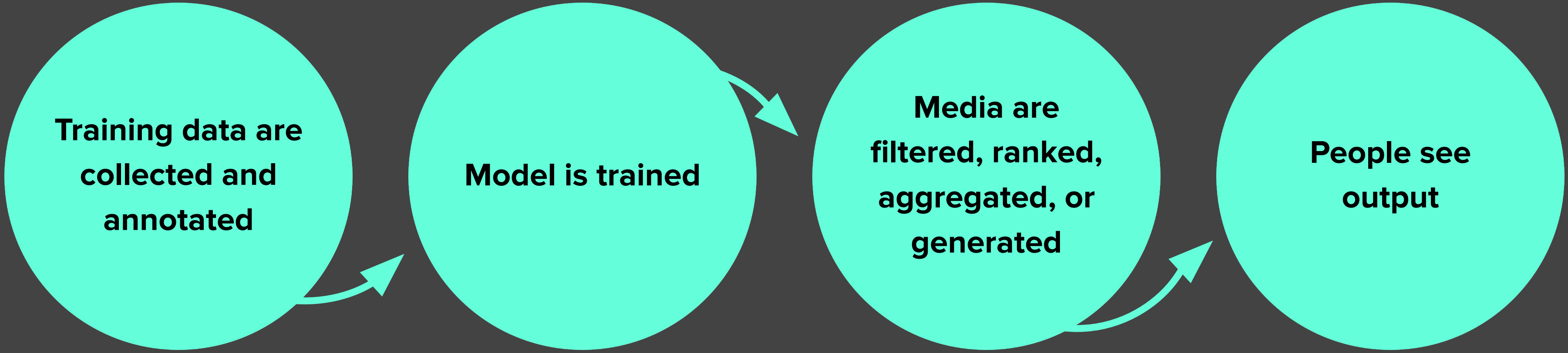
Training data are collected and annotated

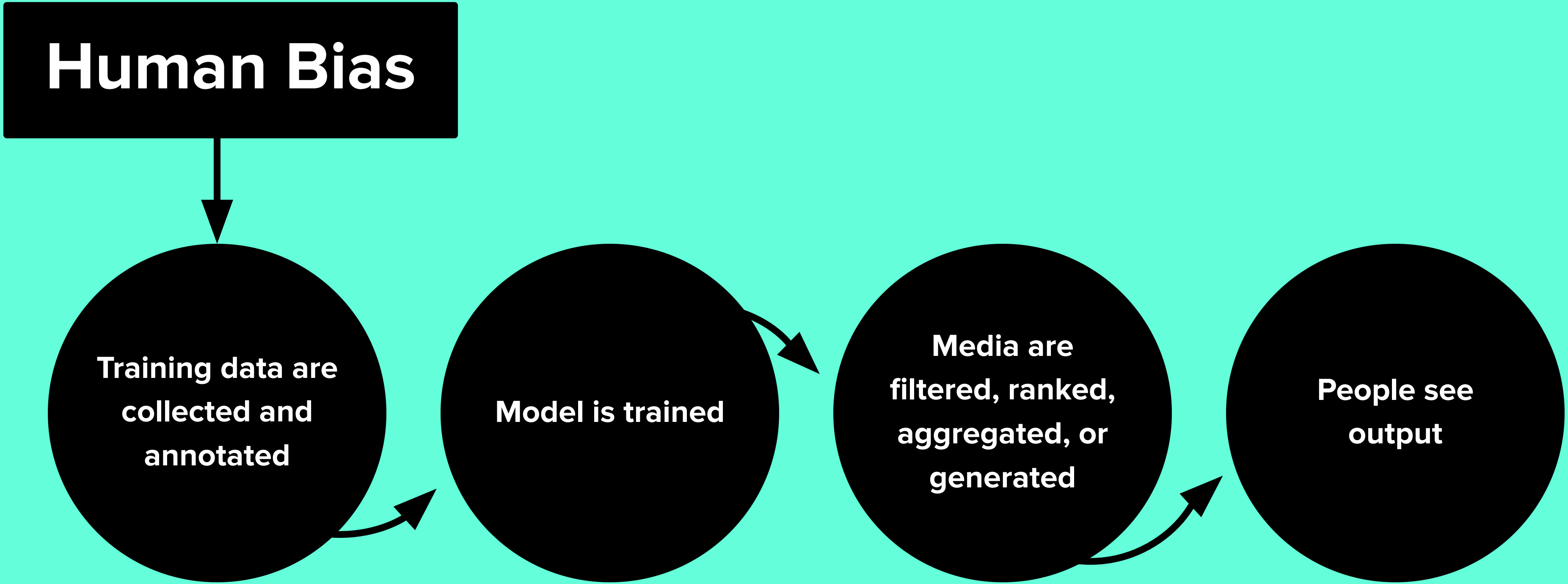
Human Biases in Data

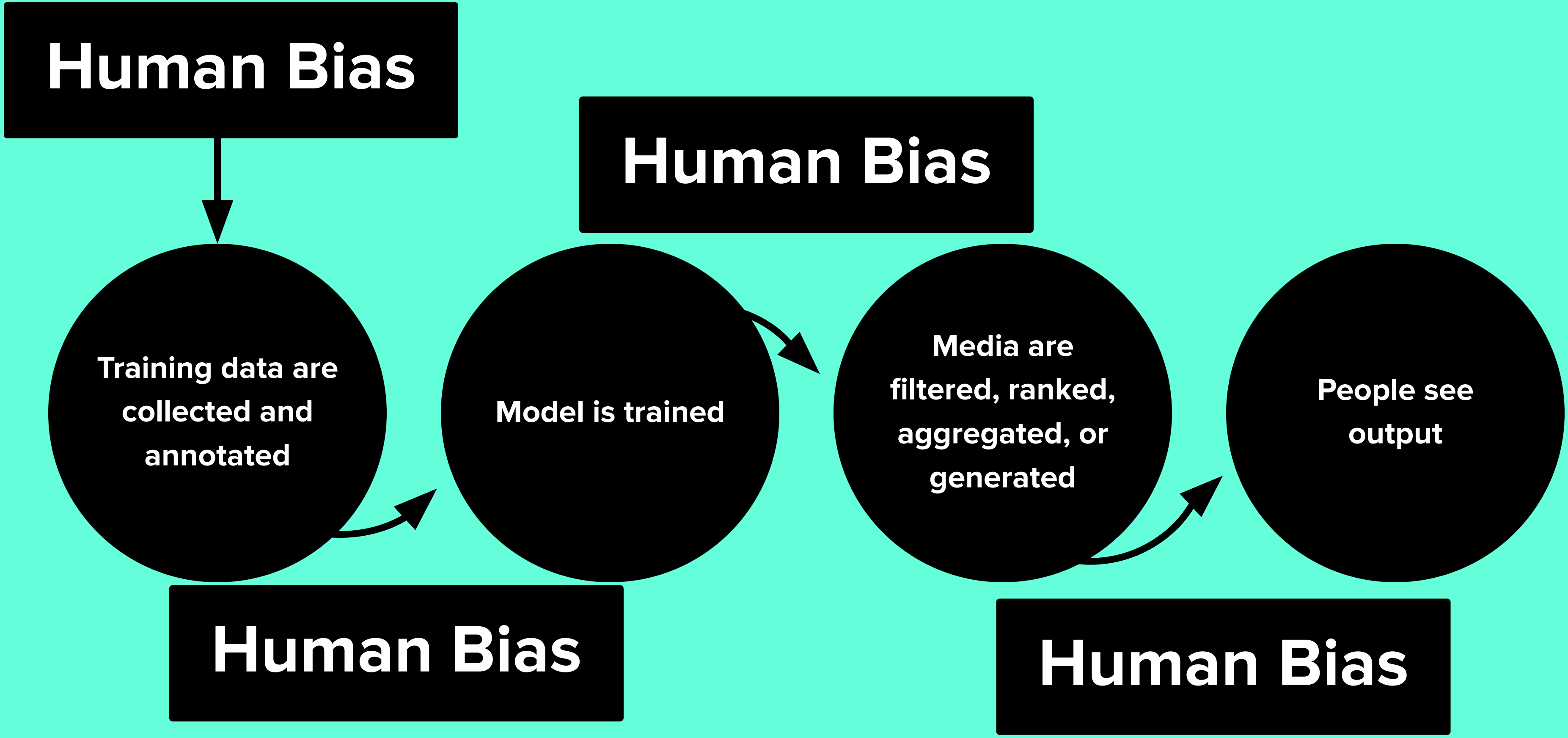
- Reporting bias
- Selection bias
- Overgeneralization
- Out-group homogeneity bias
- Stereotypical bias
- Historical unfairness
- Implicit associations
- Implicit stereotypes
- Prejudice
- Group attribution error
- Halo effect

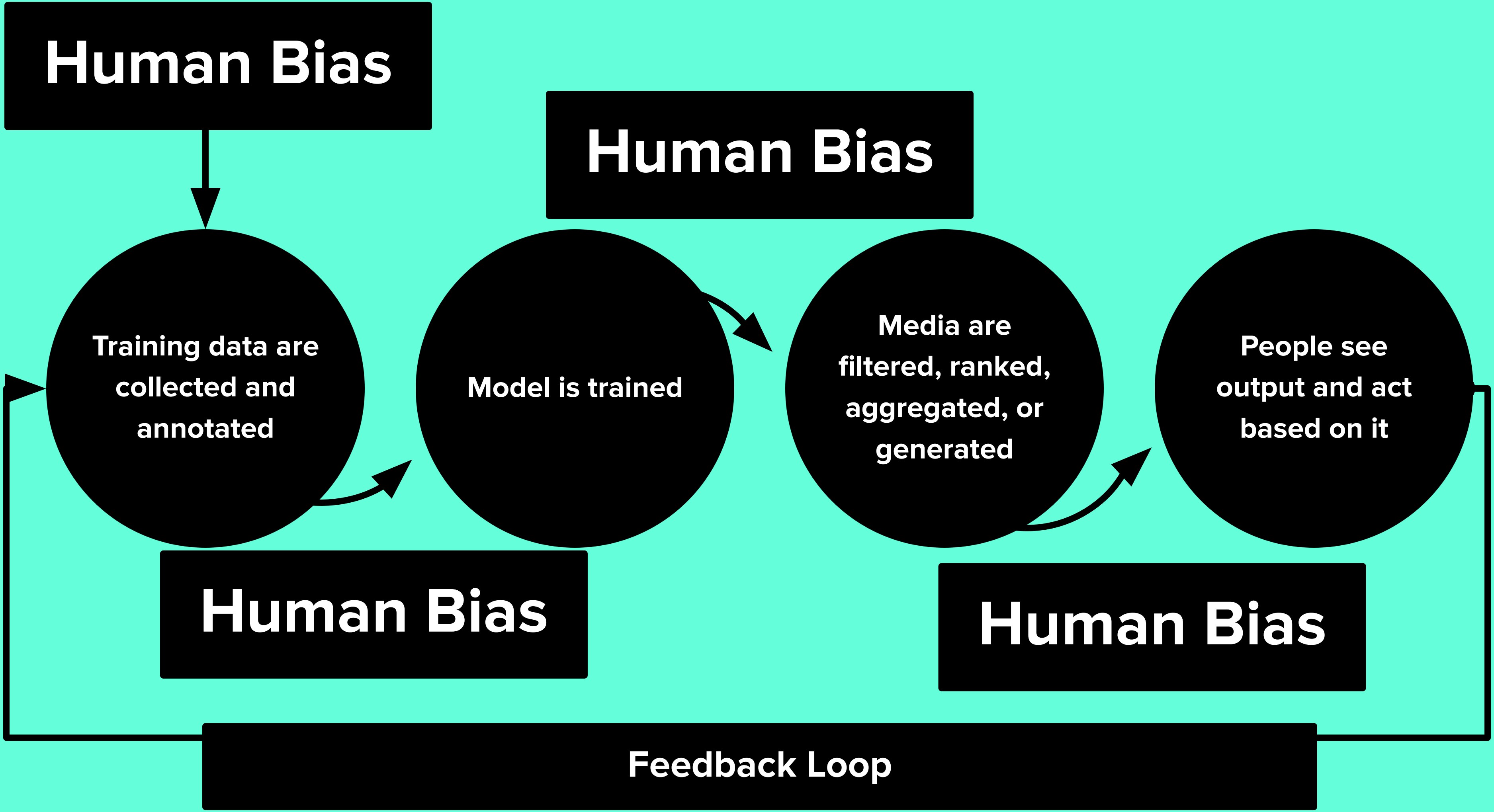
Human Biases in Collection and Annotation

- Sampling error
- Non-sampling error
- Insensitivity to sample size
- Correspondence bias
- In-group bias
- Bias blind spot
- Confirmation bias
- Subjective validation
- Experimenter's bias
- Choice-supportive bias
- Neglect of probability
- Anecdotal fallacy
- Illusion of validity


Training data are collected and annotated
Model is trained
Media are filtered, ranked, aggregated, or generated
People see output


Human Bias
Training data are collected and annotated
Model is trained
Media are filtered, ranked, aggregated, or generated
People see output


Human Bias
Human Bias
Training data are collected and annotated
Model is trained
Media are filtered, ranked, aggregated, or generated
People see output
Human Bias
Human Bias


Human Bias
Human Bias
Training data are collected and annotated
Model is trained
Media are filtered, ranked, aggregated, or generated
People see output and act based on it
Human Bias
Human Bias
Feedback Loop

Human data perpetuates human biases.

As ML learns from human data, the result is a bias network effect

“Bias Laundering”

BIAS = BAD ??

“Bias” can be Good, Bad, Neutral

- Bias in statistics and ML
  - Bias of an estimator: Difference between the predictions and the correct values that we are trying to predict
  - The "bias" term *b* (e.g., $y = mx + b$)
- Cognitive biases
  - Confirmation bias, Recency bias, Optimism bias
- Algorithmic bias
  - Unjust, unfair, or prejudicial treatment of people related to race, income, sexual orientation, religion, gender, and other characteristics historically associated with discrimination and marginalization, when and where they manifest in algorithmic systems or algorithmically aided decision-making

“Bias” can be Good, Bad, Neutral

- Bias in statistics and ML
  - Bias of an estimator: Difference between the predictions and the correct values that we are trying to predict
  - The "bias" term *b* (e.g., $y = mx + b$)
- Cognitive biases
  - Confirmation bias, Recency bias, Optimism bias
- **Algorithmic bias**
  - **Unjust, unfair, or prejudicial treatment of people** related to race, income, sexual orientation, religion, gender, and other characteristics historically associated with discrimination and marginalization, when and where they manifest in algorithmic systems or algorithmically aided decision-making

“Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice.”

— The Guardian

CREDIT

The Guardian view on machine learning: people must decide

> *"Although neural networks might be said to write their own programs, they do so towards goals set by humans, using data collected for human purposes. If the data is skewed, even by accident, the computers will amplify injustice."*
>
> — The Guardian

CREDIT

[The Guardian view on machine learning: people must decide]

Fairness in Machine Learning

A Few Case Studies

Language Identification

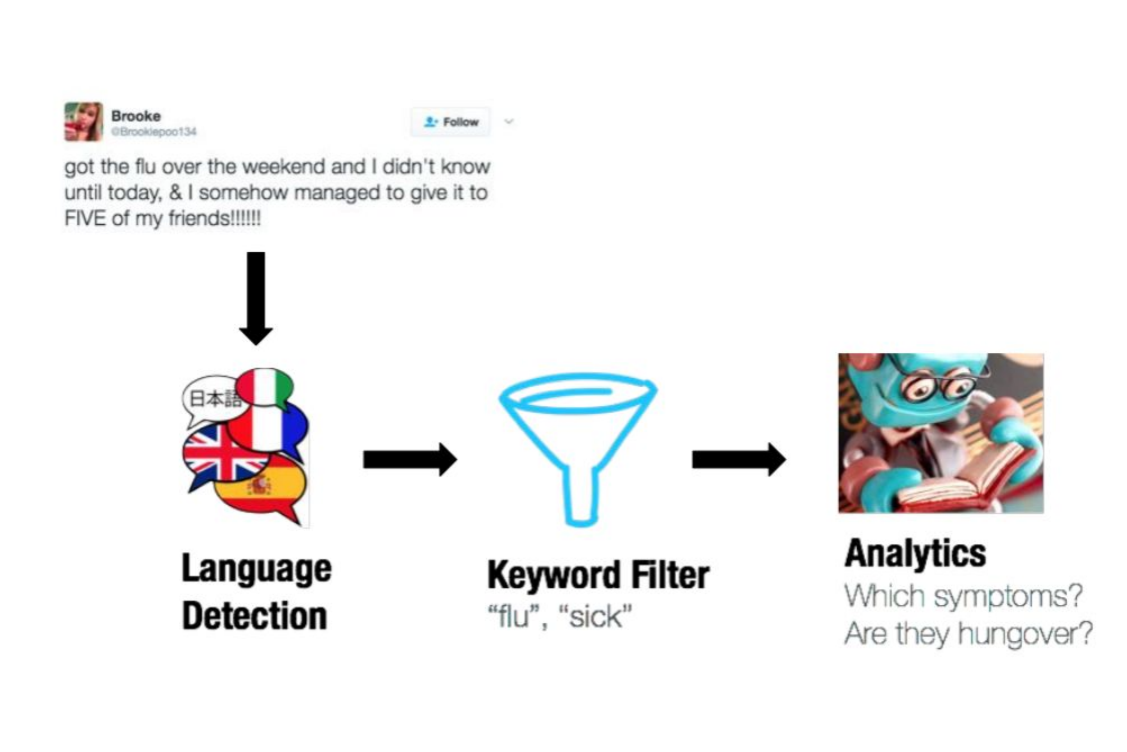
Language Identification

Most NLP models in practice has a Language Identification (LID) step

Slide credit: David Jurgens
(Jurgens et al. ACL'17)

Language Identification

Most NLP models in practice has a Language Identification (LID) step



Slide credit: David Jurgens
(Jurgens et al. ACL'17)

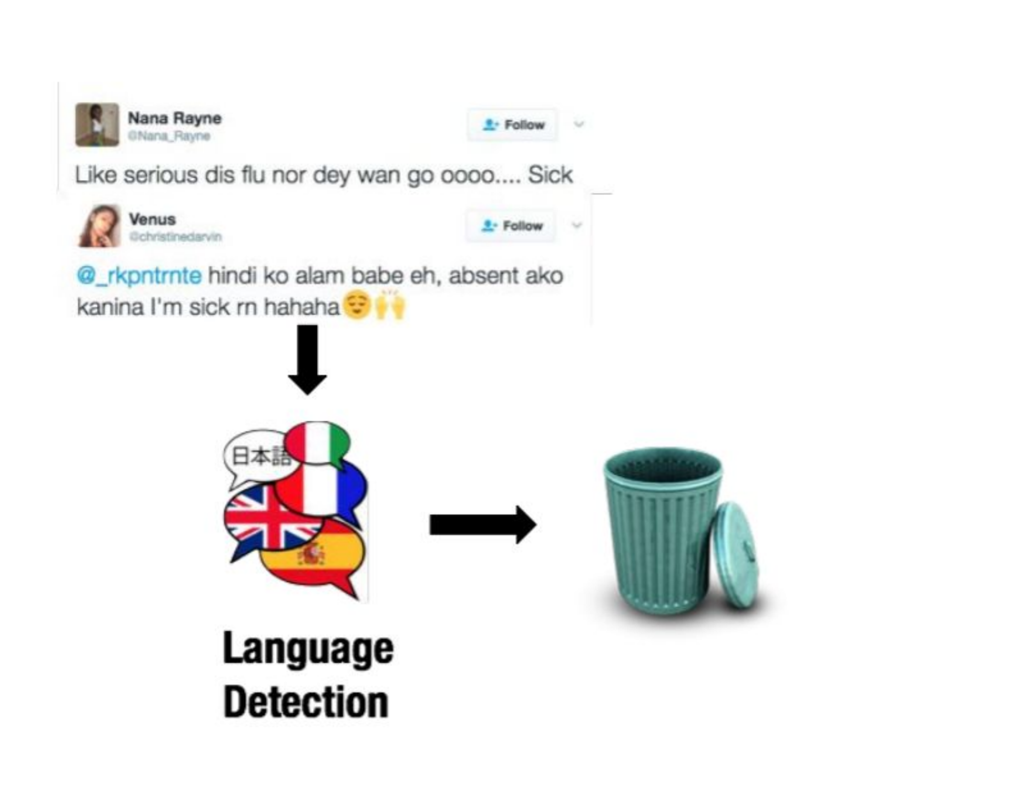
How well do LID systems do?

"This paper describes […] how even the most simple of these methods *using data obtained from the World Wide Web achieve accuracy approaching 100%* on a test suite comprised of ten European languages"

McNamee, P., "Language identification: *a solved problem* suitable for undergraduate instruction" Journal of Computing Sciences in Colleges 20(3) 2005.

Slide credit: David Jurgens
(Jurgens et al. ACL'17)

LID Usage Example: Public Health Monitoring



Slide credit: David Jurgens
(Jurgens et al. ACL'17)

Biases in Data

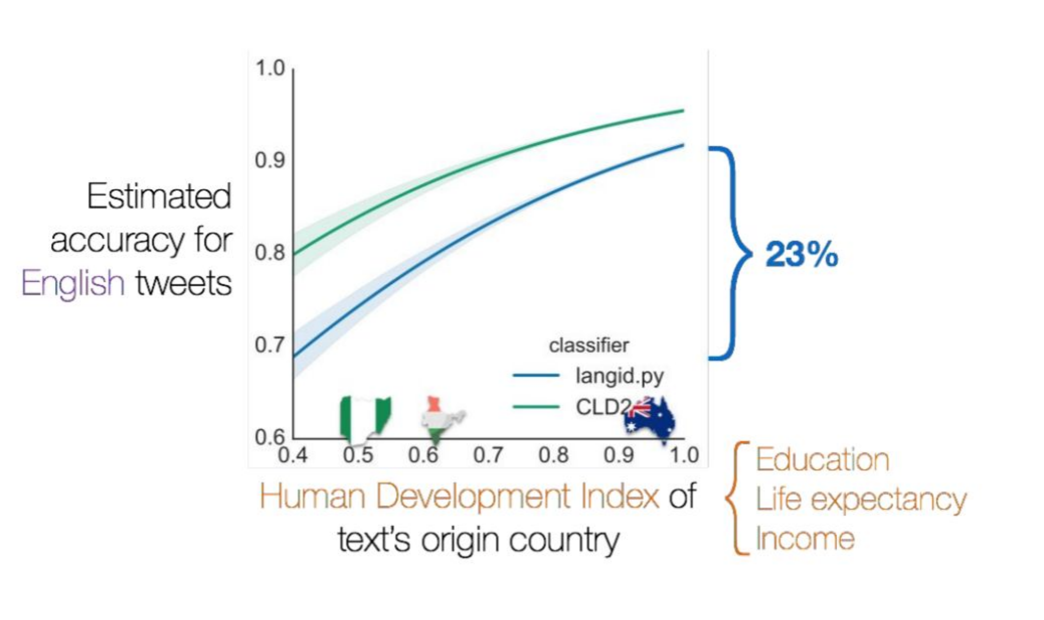
Selection Bias: Selection does not reflect a random sample

World Englishes

Is the data we use to train our English NLP models representative of all the Englishes out there?

How does this affect NLP models?

Off-the-shelf LID systems under-represent populations in less-developed countries



1M geo-tagged Tweets with any of 385 **English** terms from established lexicons for *influenza*, *psychological well-being*, and *social health*

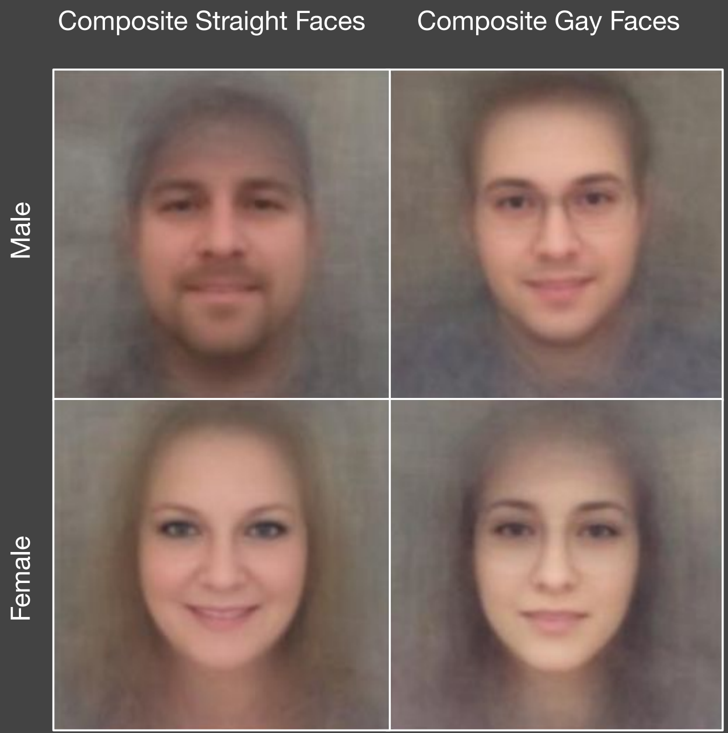
Slide credit: David Jurgens
(Jurgens et al. ACL'17)

i.e.

people who are the most marginalized,
people who'd benefit the most from such technology,
are also the ones who are more likely to be
systematically excluded from this technology

Predicting Homosexuality

Predicting Homosexuality

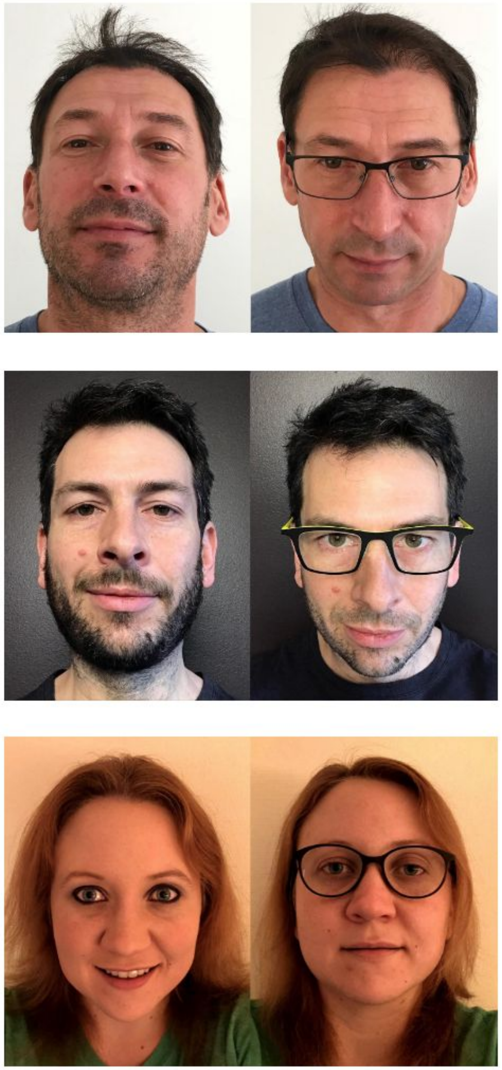


- Wang and Kosinski, Deep neural networks are more accurate than humans at detecting sexual orientation from facial images, 2017.
- "Sexual orientation detector" using 35,326 images from public profiles on a US dating website.
- "Consistent with the prenatal hormone theory [PHT] of sexual orientation, gay men and women tended to have gender-atypical facial morphology."

Predicting Homosexuality

Differences between lesbian or gay and straight faces in selfies relate to grooming, presentation, and lifestyle — that is, **differences in culture, not in facial structure**.

See Medium article: [“Do Algorithms Reveal Sexual Orientation or Just Expose our Stereotypes?”]



Predicting Criminality

Predicting Criminality

Israeli startup, Faception

> *“Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for profiling people and **revealing their personality based only on their facial image**.”*

Offering specialized engines for recognizing “High IQ”, “White-Collar Offender”, “Pedophile”, and “Terrorist” from a face image.

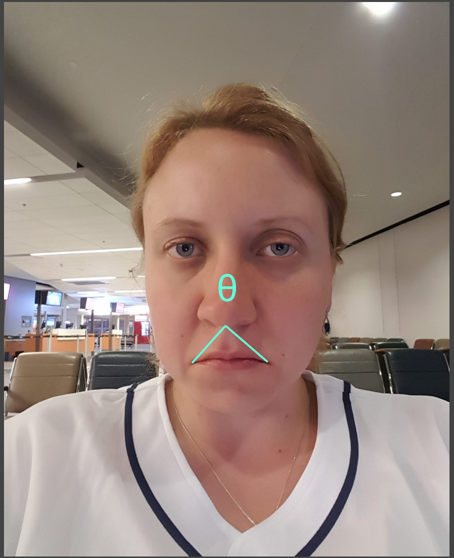
Main clients are in homeland security and public safety.

Predicting Criminality

“Automated Inference on Criminality using Face Images” Wu and Zhang, 2016. arXiv

1,856 closely cropped images of faces; Includes “wanted suspect” ID pictures from specific regions.

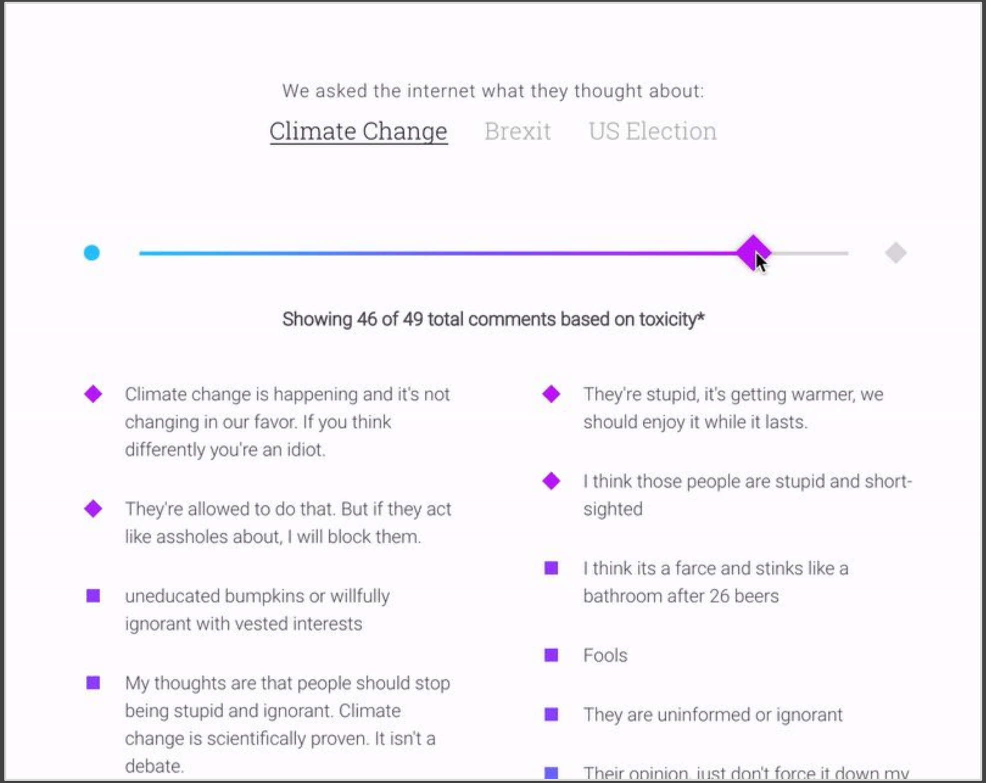
“[…] angle θ from nose tip to two mouth corners is on average 19.6% smaller for criminals than for non-criminals ...”



See our longer piece on Medium, “Physiognomy’s New Clothes”

Predicting Toxicity in Text

Toxicity Classification

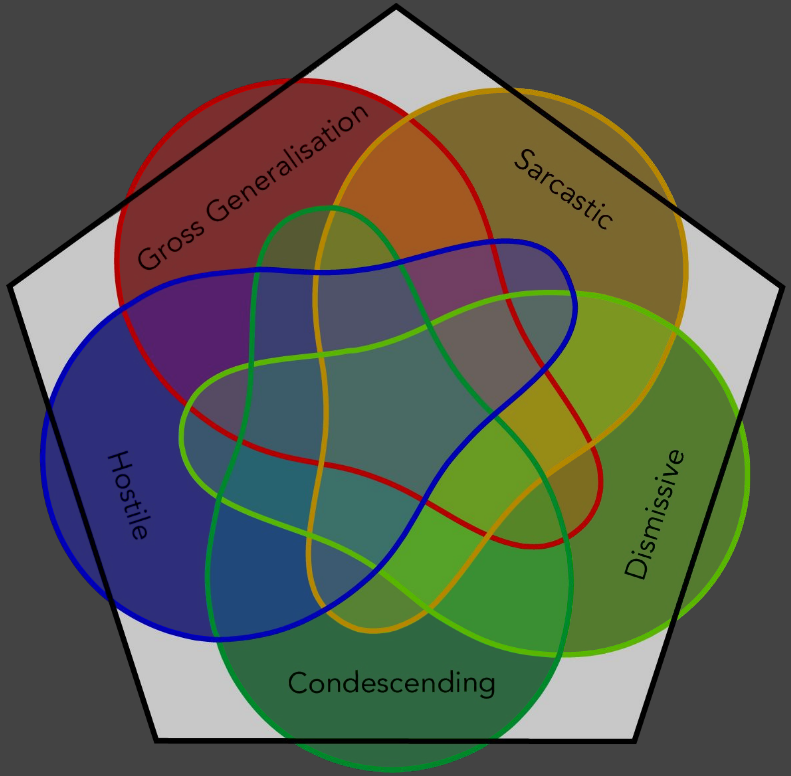


Source
perspectiveapi.com

Toxicity Classification

The False Positive

Toxicity is defined as... "***a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion.***"



Source
"The Challenge of Identifying Subtle Forms of Toxicity Online" - Jigsaw
https://medium.com/the-false-positive/the-challenge-of-identifying-subtle-forms-of-toxicity-online-465505b6c4c9

Toxicity Classification

Unintended biases towards **certain identity terms**:

Comment	Toxicity Score
The Gay and Lesbian Film Festival starts today.	0.82
Being transgender is independent of sexual orientation.	0.52
A Muslim is someone who follows or practices Islam	0.46

- "The Challenge of Identifying Subtle Forms of Toxicity Online". Jigsaw. The False Positive (2018).

Toxicity Classification

Unintended biases towards **named entities**:

Comment	Toxicity Score
I hate Justin Timberlake.	0.90
I hate Rihanna.	0.69

- Prabhakaran et al. (2019). “Perturbation Sensitivity Analysis to Detect Unintended Model Biases” EMNLP 2019

Toxicity Classification

Unintended biases towards **mentions of disabilities**:

Comment	**Toxicity Score**
I am a person.	0.08
I am a tall person.	0.03

- Hutchinson et al. (2019). *Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities.* SIGACCESS ASSETS AI Fairness Workshop 2019.

Toxicity Classification

Unintended biases towards **mentions of disabilities**:

Comment	**Toxicity Score**
I am a person.	0.08
I am a tall person.	0.03
I am a blind person.	0.39
I am a deaf person.	0.44

- Hutchinson et al. (2019). *Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities.* SIGACCESS ASSETS AI Fairness Workshop 2019.

Toxicity Classification

Unintended biases towards **mentions of disabilities**:

Comment	Toxicity Score
I am a person.	0.08
I am a tall person.	0.03
I am a blind person.	0.39
I am a deaf person.	0.44
I am a person with mental illness.	0.62

- Hutchinson et al. (2019). *Unintended Machine Learning Biases as Social Barriers for Persons with Disabilities.* SIGACCESS ASSETS AI Fairness Workshop 2019.

NLP Research on Bias and Fairness

Fairness Research in NLP

Slide from SRNLP Tutorial at NAACL 2018

1. Bolukb… Saligrama V., Kalai A. (2016) **Man is to** … **s Woman is to Homemaker? Debiasing Word** … s

2. Caliskan, A., Bryson, J. J. and Narayanan, A. (2017) **Semantics derived automatically from language corpora contain human-like biases**. *Science*

3. Nikhil Garg, Londa Schiebinger, Dan Jurafsky, James Zou. (2018) **Word embeddings quantify 100 years of gender and ethnic stereotypes**. *PNAS.*

Fairness Research in NLP

1. Bolukbasi et al. **Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings**. *NIPS* (2016)
2. Caliskan, et al. **Semantics derived automatically from language corpora contain human-like biases**. *Science* (2017)
3. Zhao, Jieyu, et al. **Men also like shopping: Reducing gender bias amplification using corpus-level constraints**. *arXiv* (2017)

2018

4. Garg et al. **Word embeddings quantify 100 years of gender and ethnic stereotypes**. *PNAS.* (2018)
5. Zhao, Jieyu, et al. **Gender bias in coreference resolution: Evaluation and debiasing methods**. *arXiv* (2018)
6. Zhang, et al. **Mitigating unwanted biases with adversarial learning**. *AIES*, 2018
7. Webster, Kellie, et al. **Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns**. *TACL* (2018)
8. Svetlana and Mohammad. **Examining gender and race bias in two hundred sentiment analysis systems**. *arXiv* (2018)
9. Díaz, et al. **Addressing age-related bias in sentiment analysis**. *CHI Conference on Human Factors in Computing Systems*. (2018)
10. Dixon, et al. **Measuring and mitigating unintended bias in text classification.** *AIES.* (2018)
11. Prates, et al. **Assessing gender bias in machine translation: a case study with Google Translate**. *Neural Computing and Applications* (2018)
12. Park, et al. **Reducing gender bias in abusive language detection.** *arXiv* (2018)
13. Zhao, Jieyu, et al. **Learning gender-neutral word embeddings.** *arXiv* (2018)
14. Anne Hendricks, et al. **Women also snowboard: Overcoming bias in captioning models**. *ECCV*. (2018)
15. Elazar and Goldberg. **Adversarial removal of demographic attributes from text data**. *arXiv* (2018)
16. Hu and Strout. **Exploring Stereotypes and Biased Data with the Crowd**. *arXiv* (2018)

2019

17. Swinger, De-Arteaga, et al. **What are the biases in my word embedding?** *AIES* (2019)
18. De-Arteaga et al. **Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting**. *FAT** (2019)
19. Gonen, et al. **Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them**. NAACL (2019).
20. Manzini et al. **Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings.** NAACL (2019).
21. Sap et al. **The Risk of Racial Bias in Hate Speech Detection**. ACL (2019)
22. Stanovsky et al. **Evaluating Gender Bias in Machine Translation**. ACL (2019)
23. Garimella et al. **Women's Syntactic Resilience and Men's Grammatical Luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing**. ACL (2019)
24. …

Social Disparities (and Stereotypes) ⟶ Word Embeddings?

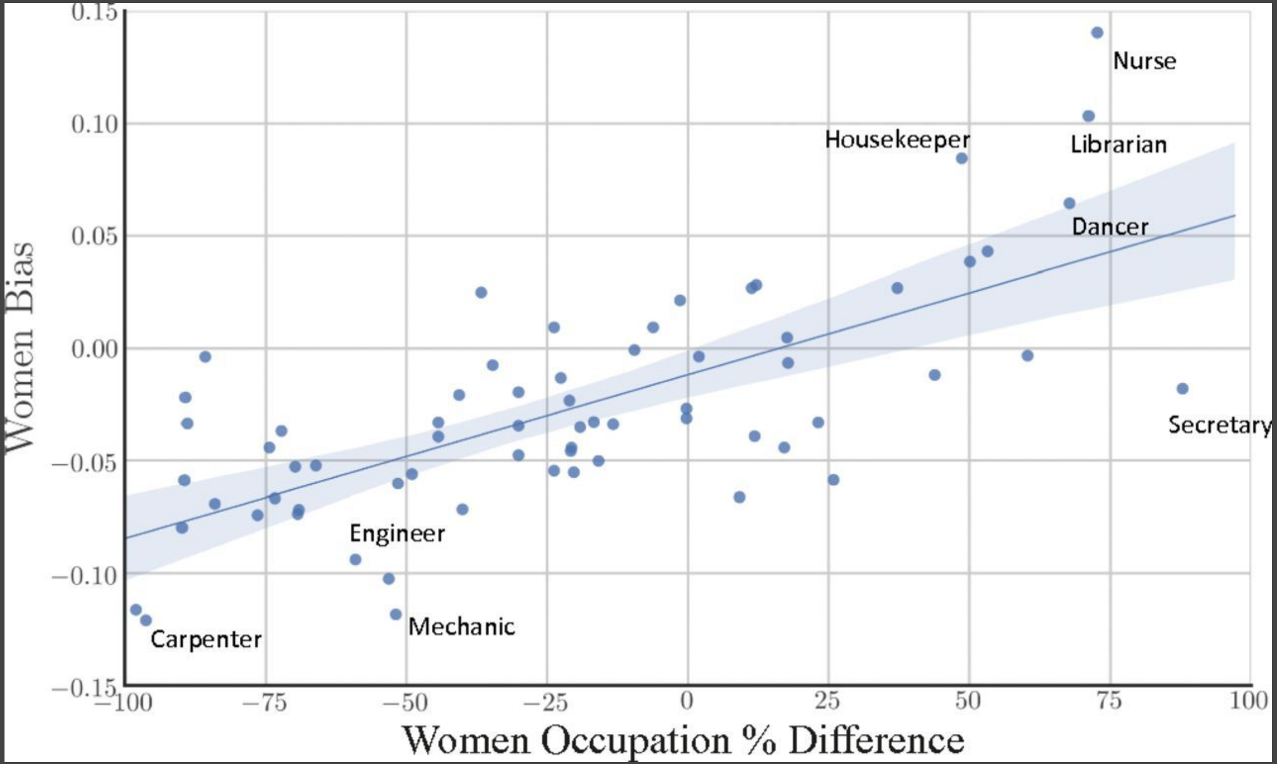
He is...

She is...

Bolukbasi et al. **Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.** *NIPS* (2016)

But aren’t they just reflecting Society?

Gender bias in Occupations



Garg et al. (2018)

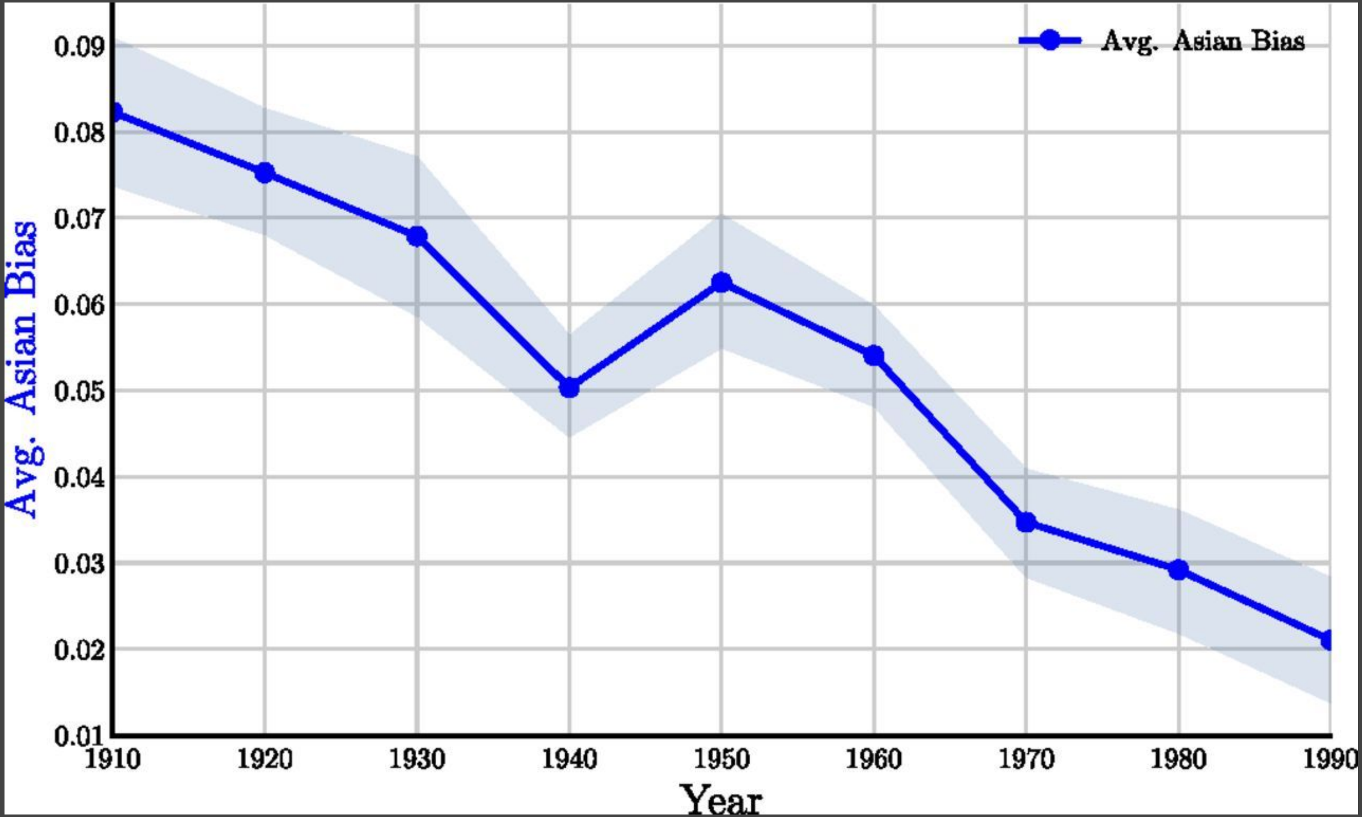
Gender bias in Adjectives over the decades

	1910	1920	1930	1940	1950	1960	1970	1980	1990
1910	1.00	0.71	0.68	0.71	0.65	0.67	0.54	0.45	0.45
1920	0.71	1.00	0.71	0.68	0.65	0.66	0.55	0.51	0.46
1930	0.68	0.71	1.00	0.73	0.71	0.70	0.60	0.52	0.53
1940	0.71	0.68	0.73	1.00	0.74	0.69	0.56	0.53	0.51
1950	0.65	0.65	0.71	0.74	1.00	0.71	0.58	0.51	0.49
1960	0.67	0.66	0.70	0.69	0.71	1.00	0.62	0.54	0.50
1970	0.54	0.55	0.60	0.56	0.58	0.62	1.00	0.63	0.56
1980	0.45	0.51	0.52	0.53	0.51	0.54	0.63	1.00	0.62
1990	0.45	0.46	0.53	0.51	0.49	0.50	0.56	0.62	1.00

Height of women's movements in 1960s-70s

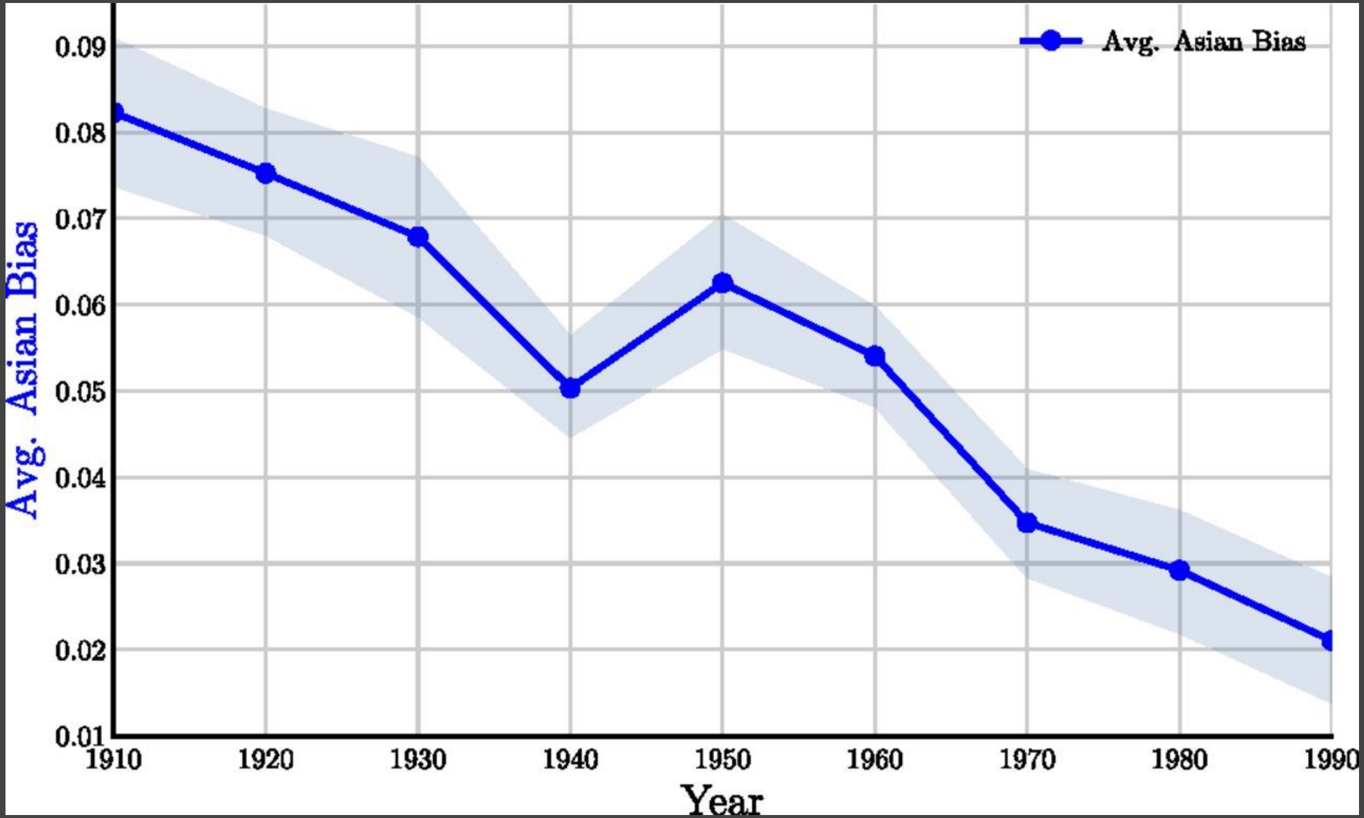
Garg et al. (2018)

“Asian bias” in Adjectives with “Outsider” words



Garg et al. (2018)

“Islam bias” in Adjectives with “Terrorist” words



Garg et al. (2018)

But aren’t they just reflecting Society?

Yup!

Word embeddings …

… get things **normatively wrong** *precisely because* they get things **descriptively right**!

Oisin Deery & Katherine Bailey
Ethics in NLP workshop. NAACL ‘18

Shouldn’t we then just leave them as is?

Shouldn't we then just leave them as is?

Would that harm certain groups of people?

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

Source: Gizmodo

What kind of harm?

Allocative Harm

"when a system allocates or withholds a certain opportunity or resource"

Associative Harm

"when systems reinforce the subordination of some groups along the lines of identity"

Source: Kate Crawford, The Trouble with Bias, NIPS 2017

Measuring Algorithmic Fairness/Bias

Evaluate for Fairness & Inclusion

Disaggregated Evaluation

Create for each (subgroup, prediction) pair.
Compare across subgroups.

Evaluate for Fairness & Inclusion

Disaggregated Evaluation

Create for each (subgroup, prediction) pair.
Compare across subgroups.

Example: women, face detection
men, face detection

Evaluate for Fairness & Inclusion

Intersectional Evaluation

Create for each (subgroup1, subgroup2, prediction) pair. Compare across subgroups.

Example: black women, face detection
white men, face detection

Evaluate for Fairness & Inclusion: Confusion Matrix

Model Predictions

References

Evaluate for Fairness & Inclusion: Confusion Matrix

		Model Predictions	
		Positive	Negative
References	Positive		
	Negative		

Evaluate for Fairness & Inclusion: Confusion Matrix

		Model Predictions	
		Positive	**Negative**
References	**Positive**	• Exists • Predicted **True Positives**	
	Negative		• Doesn’t exist • Not predicted **True Negatives**

Evaluate for Fairness & Inclusion: Confusion Matrix

		Model Predictions: Positive	Model Predictions: Negative
References	Positive	• Exists • Predicted **True Positives**	• Exists • Not predicted **False Negatives**
References	Negative	• Doesn't exist • Predicted **False Positives**	• Doesn't exist • Not predicted **True Negatives**

Evaluate for Fairness & Inclusion: Confusion Matrix

		Model Predictions		
		Positive	**Negative**	
References	**Positive**	• Exists • Predicted **True Positives**	• Exists • Not predicted **False Negatives**	**Recall, False Negative Rate**
	Negative	• Doesn’t exist • Predicted **False Positives**	• Doesn’t exist • Not predicted **True Negatives**	**False Positive Rate, Specificity**
		Precision, False Discovery Rate	**Negative Predictive Value, False Omission Rate**	**LR+, LR-**

Evaluate for Fairness & Inclusion

Female Patient Results

True Positives (TP) = 10	**False Positives (FP) = 1**
False Negatives (FN) = 1	**True Negatives (TN) = 488**

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{10}{10 + 1} = 0.909$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{10}{10 + 1} = 0.909$$

Male Patient Results

True Positives (TP) = 6	**False Positives (FP) = 3**
False Negatives (FN) = 5	**True Negatives (TN) = 48**

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{6}{6 + 3} = 0.667$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{6}{6 + 5} = 0.545$$

Evaluate for Fairness & Inclusion

Female Patient Results

True Positives (TP) = 10	**False Positives (FP) = 1**
False Negatives (FN) = 1	**True Negatives (TN) = 488**

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{10}{10 + 1} = 0.909$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{10}{10 + 1} = 0.909$$

Male Patient Results

True Positives (TP) = 6	**False Positives (FP) = 3**
False Negatives (FN) = 5	**True Negatives (TN) = 48**

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{6}{6 + 3} = 0.667$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{6}{6 + 5} = 0.545$$

“Equality of Opportunity” fairness criterion:
Recall is equal across subgroups

Evaluate for Fairness & Inclusion

Female Patient Results

True Positives (TP) = 10	**False Positives (FP) = 1**
False Negatives (FN) = 1	**True Negatives (TN) = 488**

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{10}{10 + 1} = 0.909$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{10}{10 + 1} = 0.909$$

Male Patient Results

True Positives (TP) = 6	**False Positives (FP) = 3**
False Negatives (FN) = 5	**True Negatives (TN) = 48**

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{6}{6 + 3} = 0.667$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{6}{6 + 5} = 0.545$$

"Predictive Parity" fairness criterion:
Precision is equal across subgroups

Choose your evaluation metrics in light of acceptable tradeoffs between **False Positives** and **False Negatives**

False Positives Might be Better than False Negatives

Privacy in Images

False Positive: Something that doesn't need to be blurred gets blurred.

Can be a bummer.

False Negative: Something that needs to be blurred is not blurred.

Identity theft.

False Negatives Might Be Better than False Positives

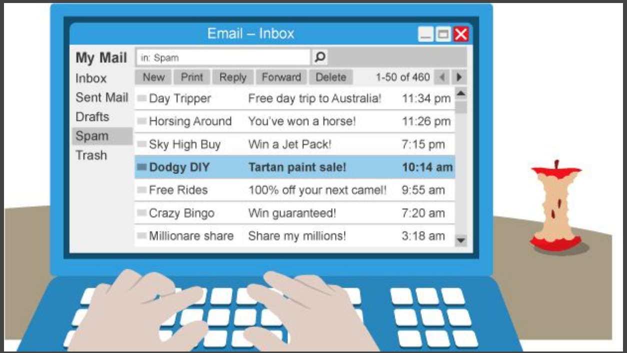
Spam Filtering

False Negative: Email that is SPAM is not caught, so you see it in your inbox.

Usually just a bit annoying.

False Positive: Email flagged as SPAM is removed from your inbox.

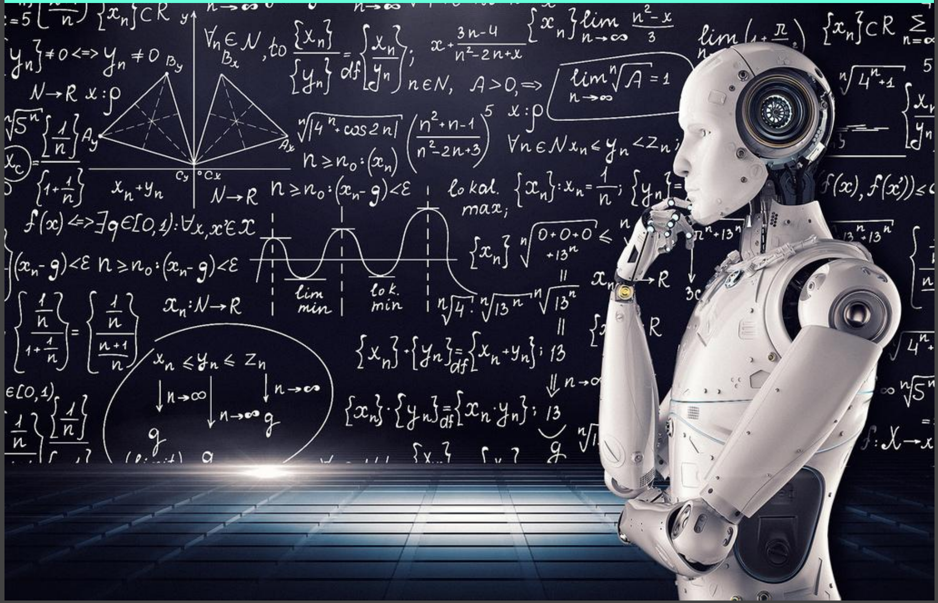
If it is an interview call?



AI Can Unintentionally Lead to Unjust Outcomes

- Lack of insight into **sources of bias in the data** and model
- Lack of insight into the **feedback loops**
- Lack of careful, **disaggregated evaluation**
- Human **biases in interpreting and accepting results**

So… What do we do?



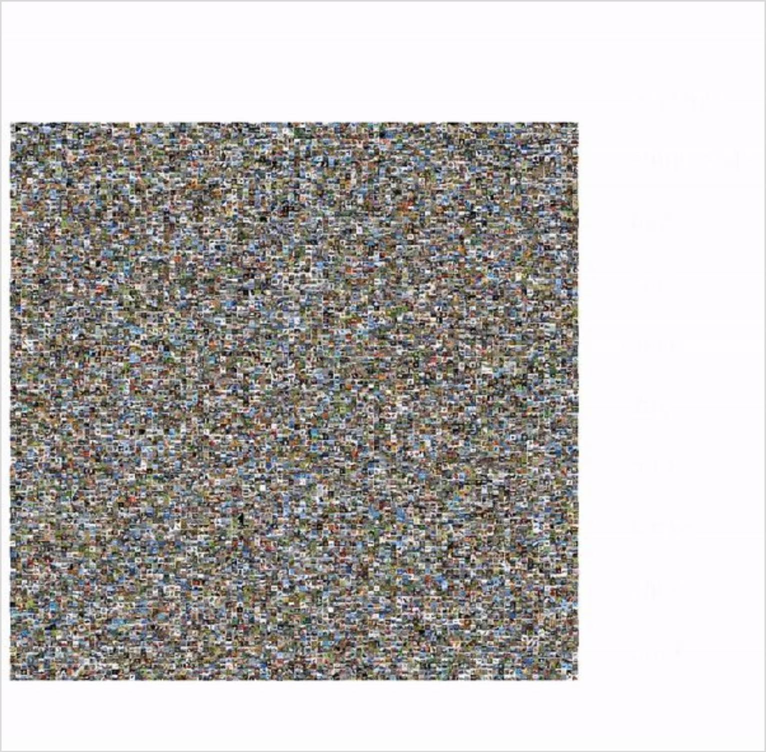
Part 2:
Bias in NLP and Mitigation Approaches
(Kai-Wei)

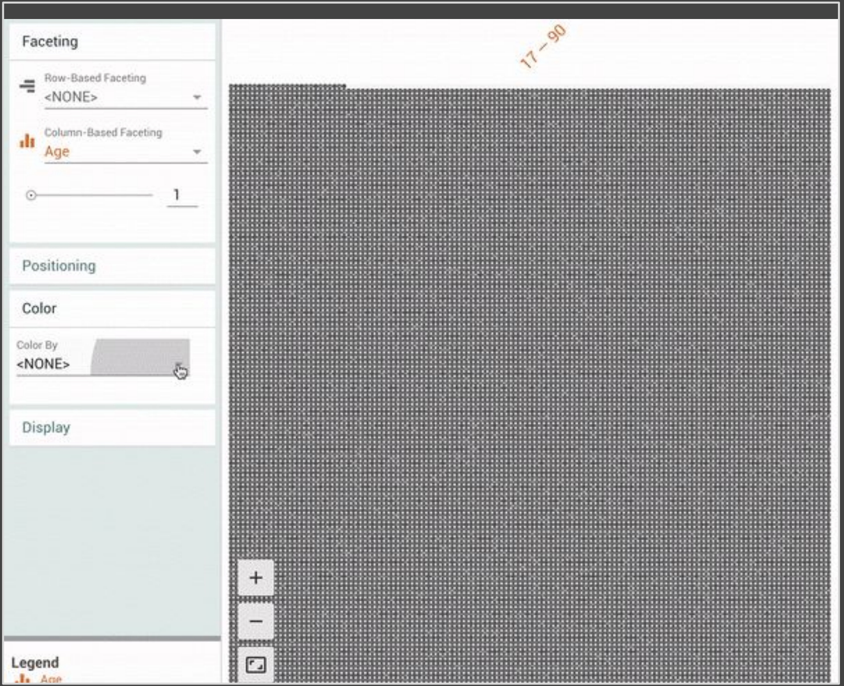
Part 3:
Building Fair and Robust Representations for Vision and Language
(Vicente)

Part 4:
Conclusion and Discussion
(Vinod)

Data Really, Really Matters

Understand Your Data Skews





Facets: pair-code.github.io

Datasheets for Datasets

Timnit Gebru [1] Jamie Morgenstern [2] Briana Vecchione [3] Jennifer Wortman Vaughan [1] Hanna Wallach [1]
Hal Daumé III [1 4] Kate Crawford [1 5]

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

Who funded the creation of the dataset? If there is an associated grant, provide the grant number.

Any other comments?

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

How many instances of each type are there?

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?

Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

If the dataset is a sample, then what is the population? What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses?

Dataset Fact Sheet

Metadata

Cj CC-0 .csv

Title COMPAS Recidivism Risk Score Data

Author Broward County Clerk's Office, Broward County Sherrif's Office, Florida

Email browardcounty@florida.usa

Description Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

DOI 10.5281/zenodo.1164791

Time Feb 2013 - Dec 2014

Keywords risk assessment, parole, jail, recidivism, law

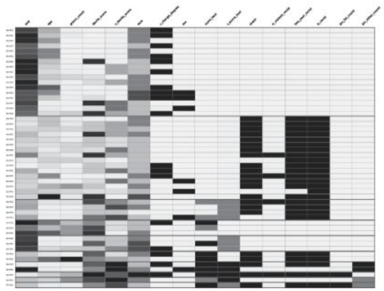
Records	7214
Variables	25

priors_count: *Ut enim ad minim veniam, quis nostrud exercitation* **numerical**

Probabilistic Modeling

Analysis

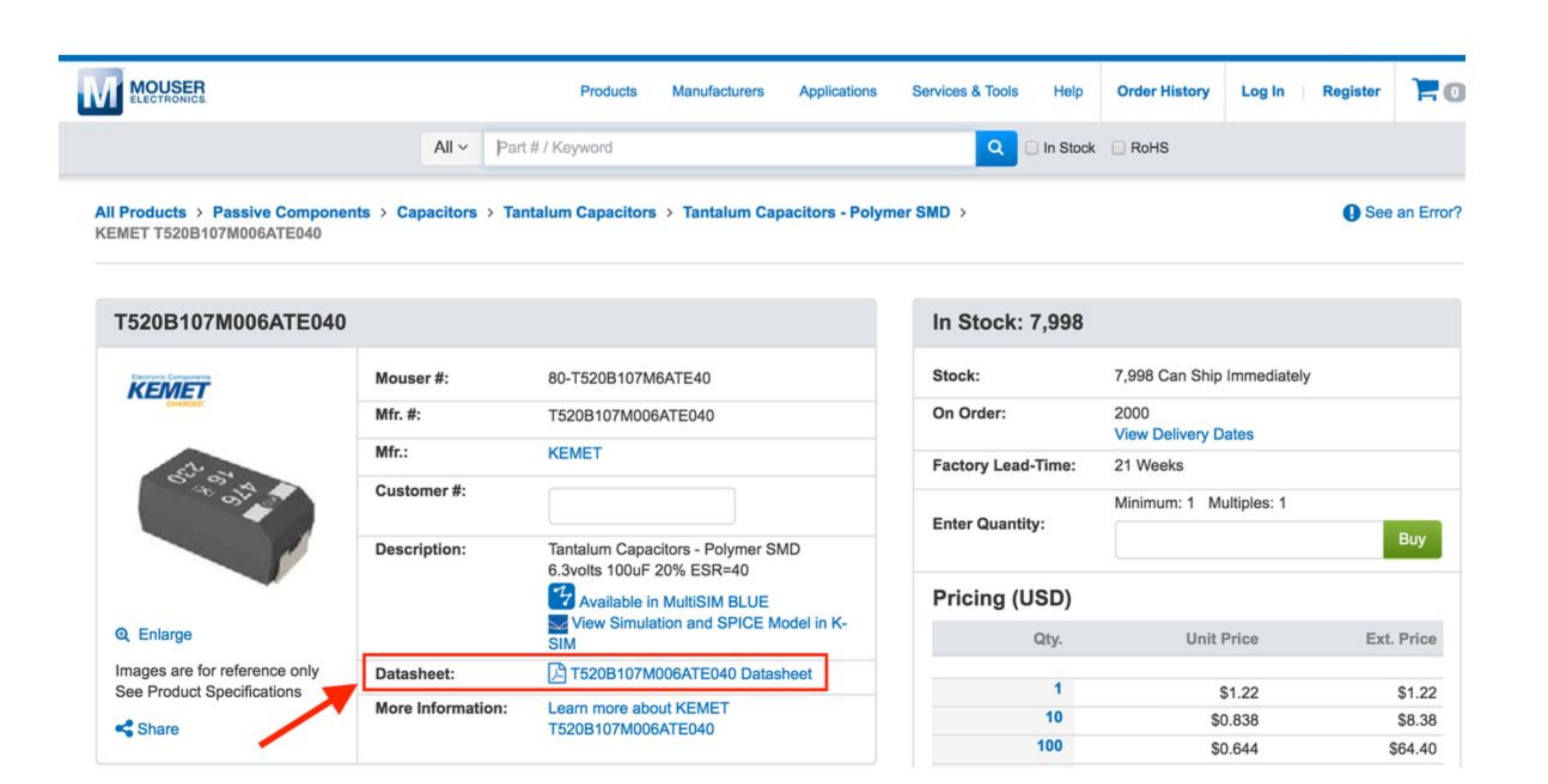
12



Dependency Probability **Pearson R**

Release Your Models Responsibly

Transparency for Electronics Components



Slide by Timnit Gebru

“Operating Characteristics” of a component

XICON

Miniature Aluminum Electrolytic Capacitors — XRL Series

FEATURES

- Low impedance characteristics
- Case sizes are smaller than conventional general-purpose capacitors, with very high performance
- Can size larger than 9mm diameter has safety vents on rubber end seal
- RoHS Compliant

RoHS Compliant

CHARACTERISTICS

Item	Characteristics													
Operating Temperature Range	-40°C ~ +85°C													
Capacitance Tolerance	±20% at 120Hz, 20°C													
Leakage Current	≤100V	I = 0.01CWV or 3μA whichever is greater after 2 minutes of applied rated DC working voltage at 20°C. Where: C = rated capacitance in μF; WV = rated DC working voltage												
	>100V	CWV ≤ 1000 μF: I= 0.03 CWV + 15uA; C= rated capacitance in uF. CWV ≥ 1000 μF: I= 0.02 CWV + 25uA; WV= rated DC working voltage in V												
Dissipation Factor (Tan δ, at 20°C 120Hz)	Working voltage (WV)		6.3	10	16	25	35	50	63	100	160	250	350	450
	Tan δ		0.23	0.20	0.16	0.14	0.12	0.10	0.09	0.08	0.12	0.17	0.20	0.25
	For capacitors whose capacitance exceeds 1,000μF, the specification of tan δ is increased by 0.02 for every addition of 1,000μF													
Surge Voltage	Working voltage (WV)		6.3	10	16	25	35	50	63	100	160	250	350	450
	Surge voltage (SV)		8	13	20	32	44	63	79	125	200	300	400	500
Low Temperature Characteristics (Imp. ratio @ 120Hz)	Working voltage (WV)		6.3	10	16	25	35	50	63	100	160	250	350	450
	Z(-25°C)/Z(+20°C)	øD<16	6	4	3	3	2	2	2	2	3	8	12	16
		øD≥16	8	6	4	4	3	3	3	3	3	8	12	16
	Z(-40°C)/Z(+20°C)	øD<16	10	8	6	6	4	3	3	3	4	10	16	20
		øD≥16	18	16	12	10	8	8	6	6	4	10	16	20
Load Test	When returned to +20°C after 2,000 hours application of working voltage at +85°C, the capacitor will meet the following limits: Capacitance change is ≤ ±20% of initial value; tan δ is < 200% of specified value; leakage current is within specified value													
Shelf Life Test	When returned to +20°C after 1,000 hours at +85°C with no voltage applied, the capacitor will meet the following limits: Capacitance change is ≤ ±20% of initial value; tan δ is < 200% of specified value; leakage current is within specified value													

PART NUMBERING SYSTEM

1 4 0	–	X R L	1 6 V	1 0 0	–	R C
Prefix		Series	Voltage Actual Value	Capacitance (μF) Actual Value		Suffix RoHS Compliant

RIPPLE CURRENT AND FREQUENCY MULTIPLIERS

Capacitance (μF)	Frequency (Hz)				
	60 (50)	120	500	1K	≥10K
<100	0.70	1.0	1.30	1.40	1.50
100 ~ 1000	0.75	1.0	1.20	1.30	1.35
>1000	0.80	1.0	1.10	1.12	1.15

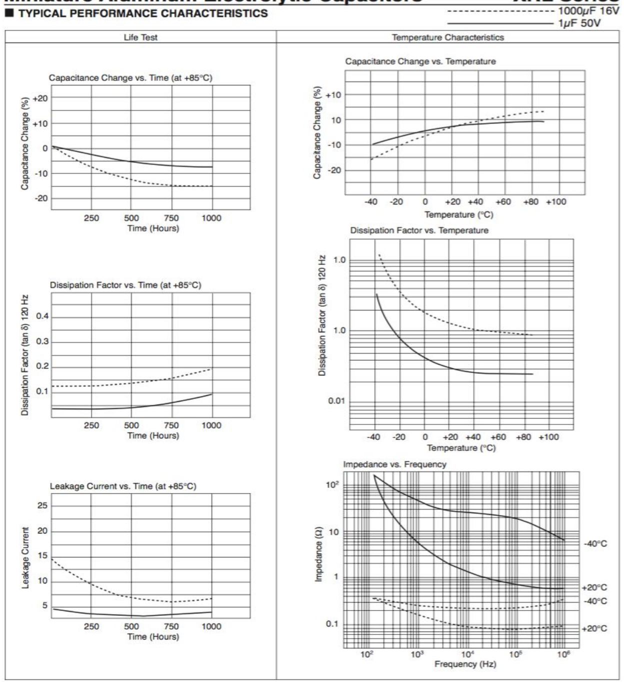
RIPPLE CURRENT AND TEMPERATURE MULTIPLIERS

Temperature (°C)	<50	70	85
Multiplier	1.78	1.4	1.0

XICON

Miniature Aluminum Electrolytic Capacitors — XRL Series

TYPICAL PERFORMANCE CHARACTERISTICS



Specifications are subject to change without notice. No liability or warranty implied by this information. Environmental compliance based on producer documentation.

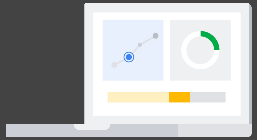
Slide by Timnit Gebru

Model Cards for Model Reporting

- Currently no common practice of reporting how well a model works when it is released

Model Cards for Model Reporting

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
{mmitchellai,simonewu,andrewzaldivar,parkerbarnes,lucyvasserman,benhutch,espitzer,tgebru}@google.com
deborah.raji@mail.utoronto.ca



What It Does

A report that focuses on transparency in model performance to encourage responsible AI adoption and application.

How It Works

It is an easily discoverable and usable artifact presented at important steps of a user journey for a diverse set of users and public stakeholders.

Why It Matters

It keeps model developer accountable to release high quality and fair models.

Mitchell et al. [Model Cards for Model Reporting](). FAT*, 2019.

Intended Use, Factors and Subgroups

Example Model Card - Toxicity in Text	
Model Details	Developed by Jigsaw in 2017 as a convolutional neural network trained to predict the likelihood that a comment will be perceived as toxic.
Intended Use	Supporting human moderation, providing feedback to comment authors, and allowing comment viewers to control their experience.
Factors	Identity terms referencing frequently attacked groups focusing on the categories of sexual orientation, gender identity and race.

Mitchell et al. Model Cards for Model Reporting. FAT*, 2019.

Metrics and Data

Metrics	*Pinned AUC*, which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.
Evaluation Data	A synthetic test set generated using a template-based approach, where identity terms are swapped into a variety of template sentences.
Training Data	Includes comments from a variety of online forums with crowdsourced labels of whether the comment is “toxic”. “Toxic” is defined as, “a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion”.

Mitchell et al. Model Cards for Model Reporting. FAT*, 2019.

Considerations, Recommendations

Ethical Considerations	A set of values around community, transparency, inclusivity, privacy and topic-neutrality to guide their work.
Caveats & Recommendations	Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

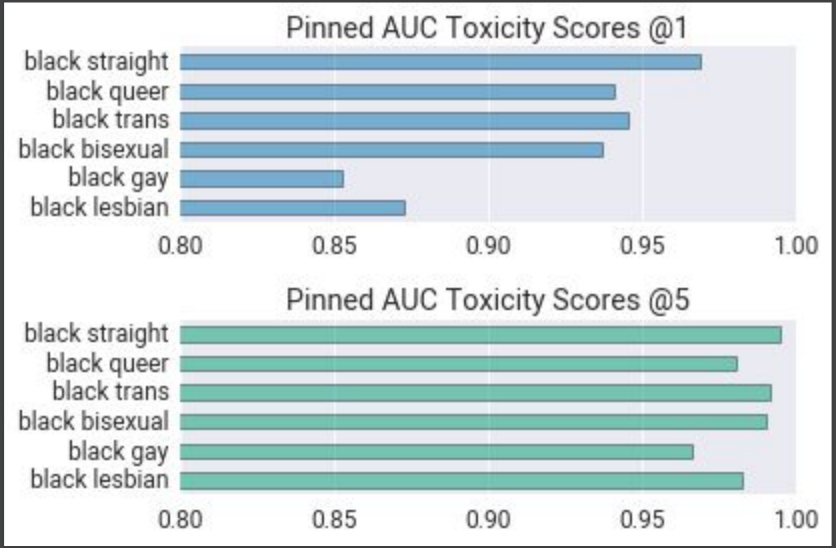
Mitchell et al. [Model Cards for Model Reporting](). FAT*, 2019.

Disaggregated Intersectional Evaluation

Toxicity @1

Identity groups	Subgroup AUC	BPSN AUC	BNSP AUC
lesbian	0.93	0.74	0.98
gay	0.94	0.65	0.99
queer	0.98	0.96	0.93
straight	0.99	1.00	0.87
bisexual	0.96	0.95	0.92
homosexual	0.87	0.53	0.99
heterosexual	0.96	0.94	0.92
cis	0.99	1.00	0.87
trans	0.97	0.96	0.91
nonbinary	0.99	0.99	0.90
black	0.91	0.85	0.95
white	0.91	0.88	0.94

0.5 0.6 0.7 0.8 0.9 1.0



Jigsaw | The False Positive

In Summary...

- Always **be mindful** of various sorts of biases in the NLP models and the data
- Explore “debiasing” techniques, but **be cautious**
- **Identify the biases that matter** for your problem and test for those biases
- Consider this an **iterative process**, than something that has a “done” state
- Be **transparent** about your model and its performance in different settings

Closing Note

"Fairness and justice are properties of social and legal systems"

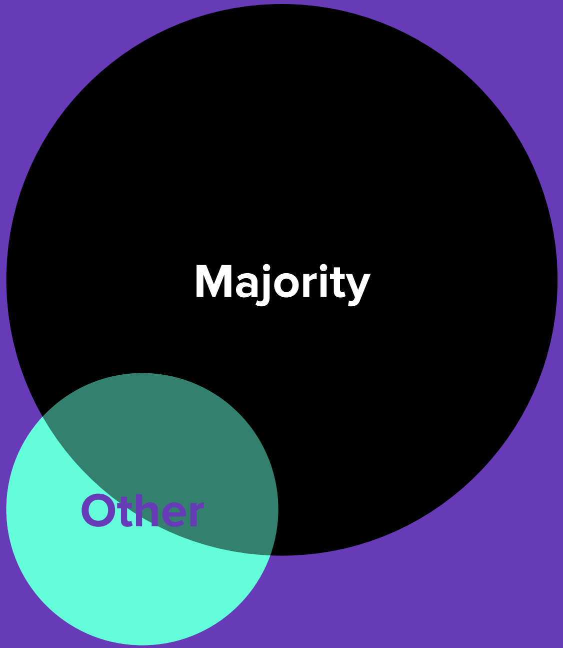
"To treat fairness and justice as terms that have meaningful application to technology separate from a social context is therefore [...] an abstraction error"

Selbst et al., Fairness and Abstraction in Sociotechnical Systems. FAT* 2018

Questions?

BACKUP Slides

Moving from majority representation...



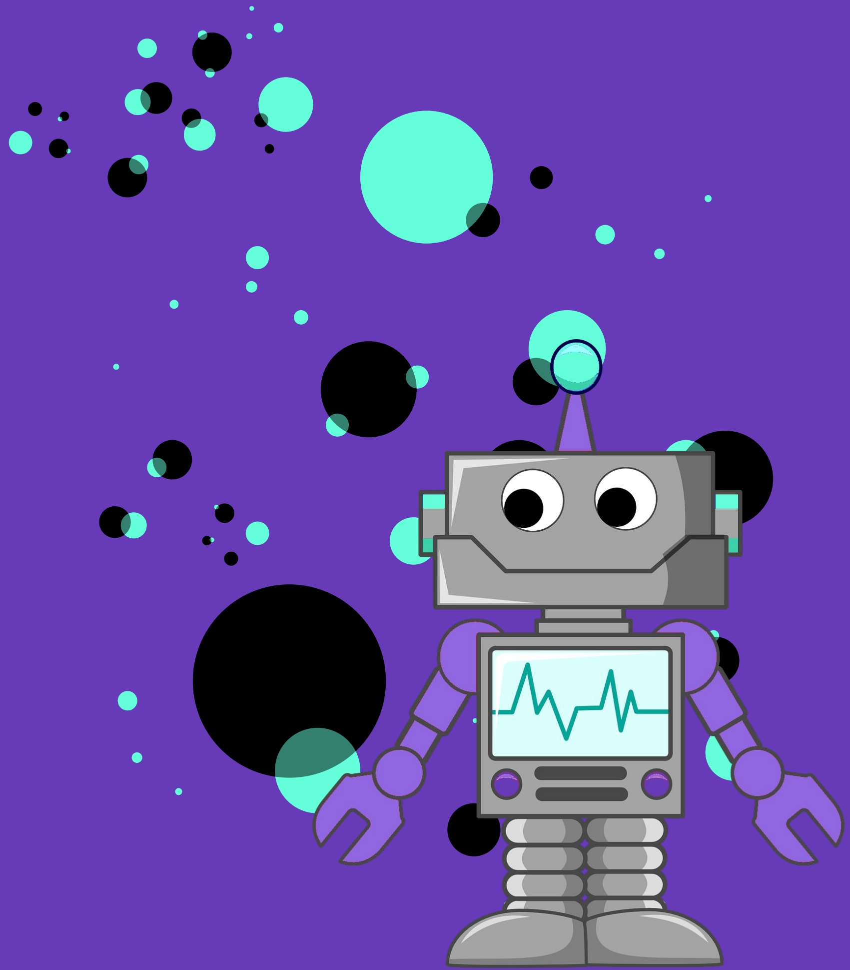
Moving from majority representation...

...to diverse representation

Moving from majority representation…

…to diverse representation

…for ethical AI



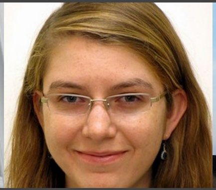
Thanks!

margarmitchell@gmail.com
m-mitchell.com

Need MOAR? ml-fairness.com


Andrew Zaldivar


Me


Simone Wu


Parker Barnes


Lucy Vasserman

Ben Hutchinson


Elena Spitzer

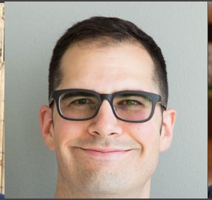

Deb Raji

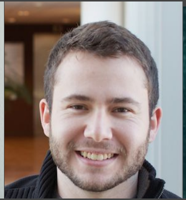

Timnit Gebru


Adrian Benton


Brian Zhang


Dirk Hovy


Josh Lovejoy


Alex Beutel


Blake Lemoine


Hee Jung Ryu


Hartwig Adam


Blaise Agüera y Arcas

More free, hands-on tutorials on how to build more inclusive ML

Measuring and Mitigating Unintended Bias in Text Classification

John Li jetpack@google.com
Lucas Dixon ldixon@google.com
Nithum Thain nthain@google.com
Lucy Vasserman lucyvasserman@google.com
Jeffrey Sorensen sorenj@google.com

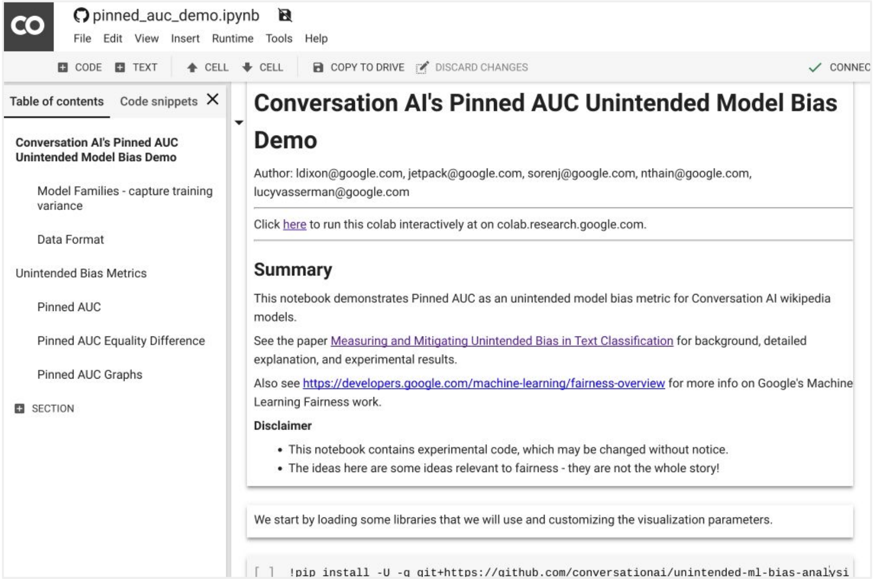

pinned_auc_demo.ipynb
File Edit View Insert Runtime Tools Help
CODE TEXT CELL CELL COPY TO DRIVE DISCARD CHANGES CONNEC

Table of contents | Code snippets

Conversation AI's Pinned AUC Unintended Model Bias Demo
Model Families - capture training variance
Data Format
Unintended Bias Metrics
Pinned AUC
Pinned AUC Equality Difference
Pinned AUC Graphs
SECTION

Conversation AI's Pinned AUC Unintended Model Bias Demo

Author: ldixon@google.com, jetpack@google.com, sorenj@google.com, nthain@google.com, lucyvasserman@google.com

Click here to run this colab interactively at on colab.research.google.com.

Summary

This notebook demonstrates Pinned AUC as an unintended model bias metric for Conversation AI wikipedia models.

See the paper Measuring and Mitigating Unintended Bias in Text Classification for background, detailed explanation, and experimental results.

Also see https://developers.google.com/machine-learning/fairness-overview for more info on Google's Machine Learning Fairness work.

Disclaimer

- This notebook contains experimental code, which may be changed without notice.
- The ideas here are some ideas relevant to fairness - they are not the whole story!

We start by loading some libraries that we will use and customizing the visualization parameters.

```
[ ] !pip install -U -q git+https://github.com/conversationai/unintended-ml-bias-analysi
```

Mitigating Unwanted Biases with Adversarial Learning

Brian Hu Zhang Stanford University Stanford, CA bhz@stanford.edu
Blake Lemoine Google Mountain View, CA lemoine@google.com
Margaret Mitchell Google Mountain View, CA mmitchellai@google.com

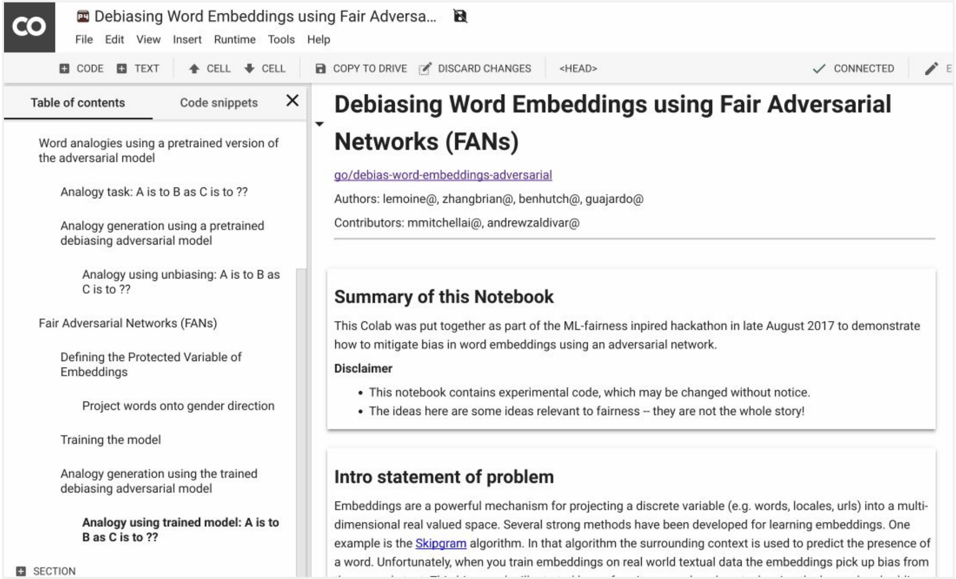

Debiasing Word Embeddings using Fair Adversa...
File Edit View Insert Runtime Tools Help
CODE TEXT CELL CELL COPY TO DRIVE DISCARD CHANGES <HEAD> CONNECTED

Table of contents | Code snippets

Word analogies using a pretrained version of the adversarial model
Analogy task: A is to B as C is to ??
Analogy generation using a pretrained debiasing adversarial model
Analogy using unbiasing: A is to B as C is to ??
Fair Adversarial Networks (FANs)
Defining the Protected Variable of Embeddings
Project words onto gender direction
Training the model
Analogy generation using the trained debiasing adversarial model
Analogy using trained model: A is to B as C is to ??
SECTION

Debiasing Word Embeddings using Fair Adversarial Networks (FANs)

go/debias-word-embeddings-adversarial

Authors: lemoine@, zhangbrian@, benhutch@, guajardo@

Contributors: mmitchellai@, andrewzaldivar@

Summary of this Notebook

This Colab was put together as part of the ML-fairness inpired hackathon in late August 2017 to demonstrate how to mitigate bias in word embeddings using an adversarial network.

Disclaimer

- This notebook contains experimental code, which may be changed without notice.
- The ideas here are some ideas relevant to fairness -- they are not the whole story!

Intro statement of problem

Embeddings are a powerful mechanism for projecting a discrete variable (e.g. words, locales, urls) into a multi-dimensional real valued space. Several strong methods have been developed for learning embeddings. One example is the Skipgram algorithm. In that algorithm the surrounding context is used to predict the presence of a word. Unfortunately, when you train embeddings on real world textual data the embeddings pick up bias from

ml-fairness.com

Get Involved

- Find free machine-learning tools open to anyone at **ai.google/tools**
- Check out Google's ML Fairness codelab at **ml-fairness.com**
- Explore educational resources at **ai.google/education**
- Take a free, hands-on Machine Learning Crash Course at **https://developers.google.com/machine-learning/crash-course/**
- Share your feedback: **acceleratewithgoogle@google.com**