



Conversation AI

# Bias in the Vision and Language of Artificial Intelligence



**Margaret Mitchell**  
Senior Research Scientist  
Google AI



Andrew  
Zaldivar



Me



Simone  
Wu



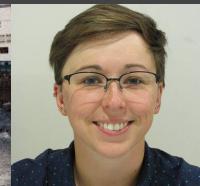
Parker  
Barnes



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  
Aguera y  
Arcas

# 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
- Bananas with stickers on them



# What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
- Bunches of bananas with stickers on them on shelves in a store



# What do you see?

- Bananas
- Stickers
- Dole Bananas
- Bananas at a store
- Bananas on shelves
- Bunches of bananas
- Bananas with stickers on them
- Bunches of bananas with stickers on them on shelves in a store

...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?

Ripe Bananas

Bananas with spots

Bananas good for banana  
bread



# 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

---

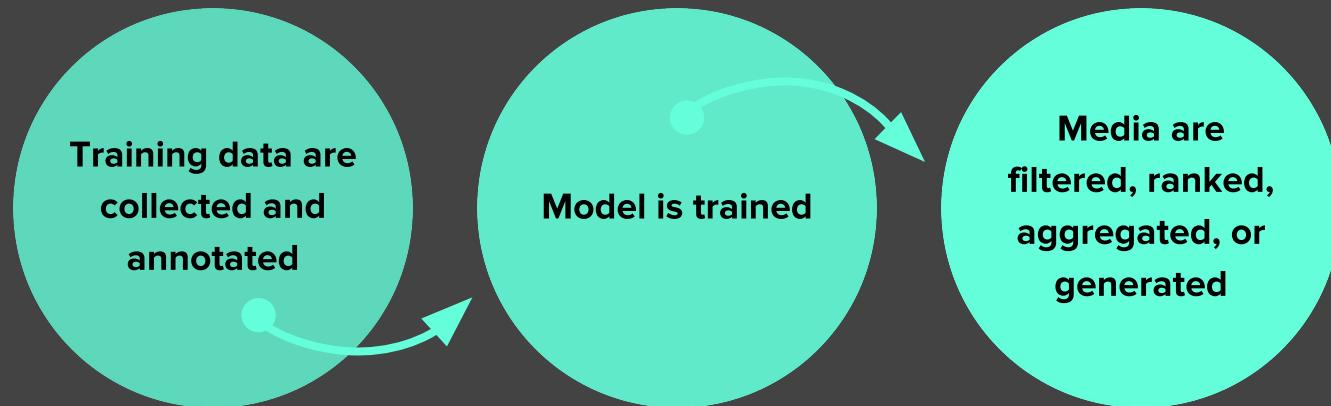


**Training data are  
collected and  
annotated**



**Training data are  
collected and  
annotated**

**Model is trained**





## 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



## Human Biases in Data

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

Training data are collected and annotated

## Human Biases in Collection and Annotation

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

**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



# Biases in Data

# Biases in Data

**Selection Bias:** Selection does not reflect a random sample

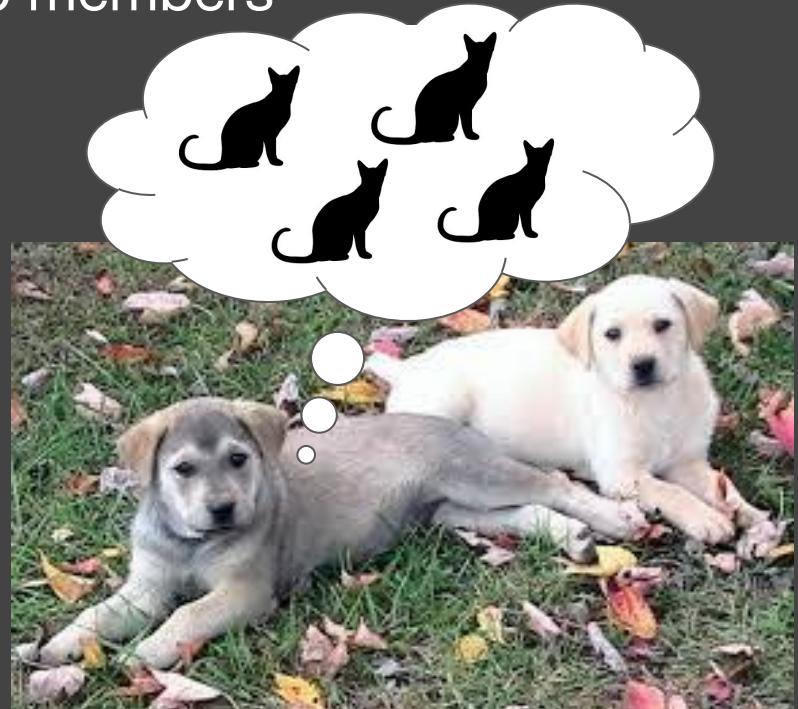


CREDIT

© 2013–2016 Michael Yoshitaka Erlewine and Hadas Kotek

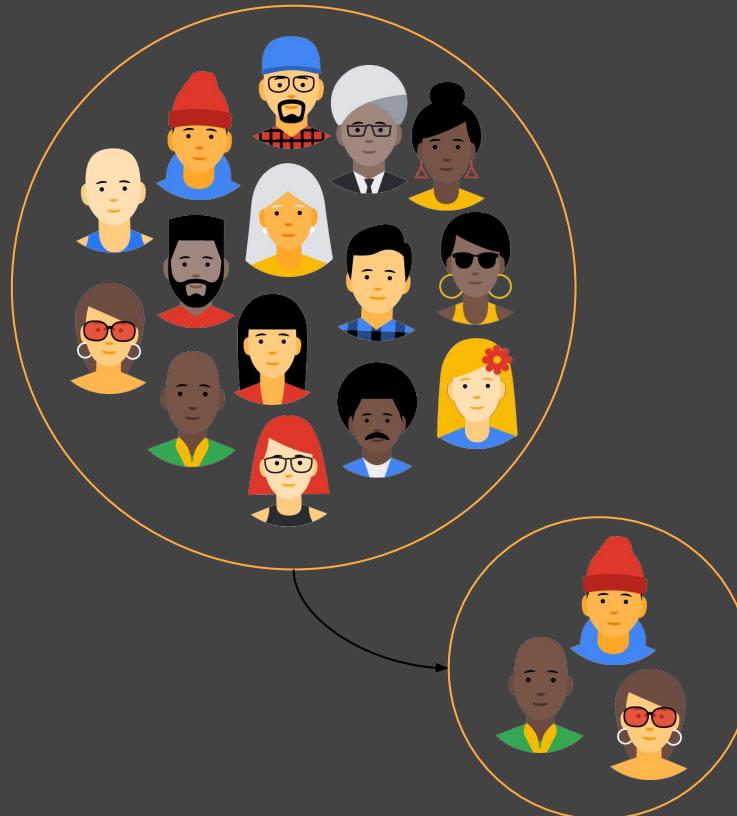
# Biases in Data

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



# Biases in Data → Biased Data Representation

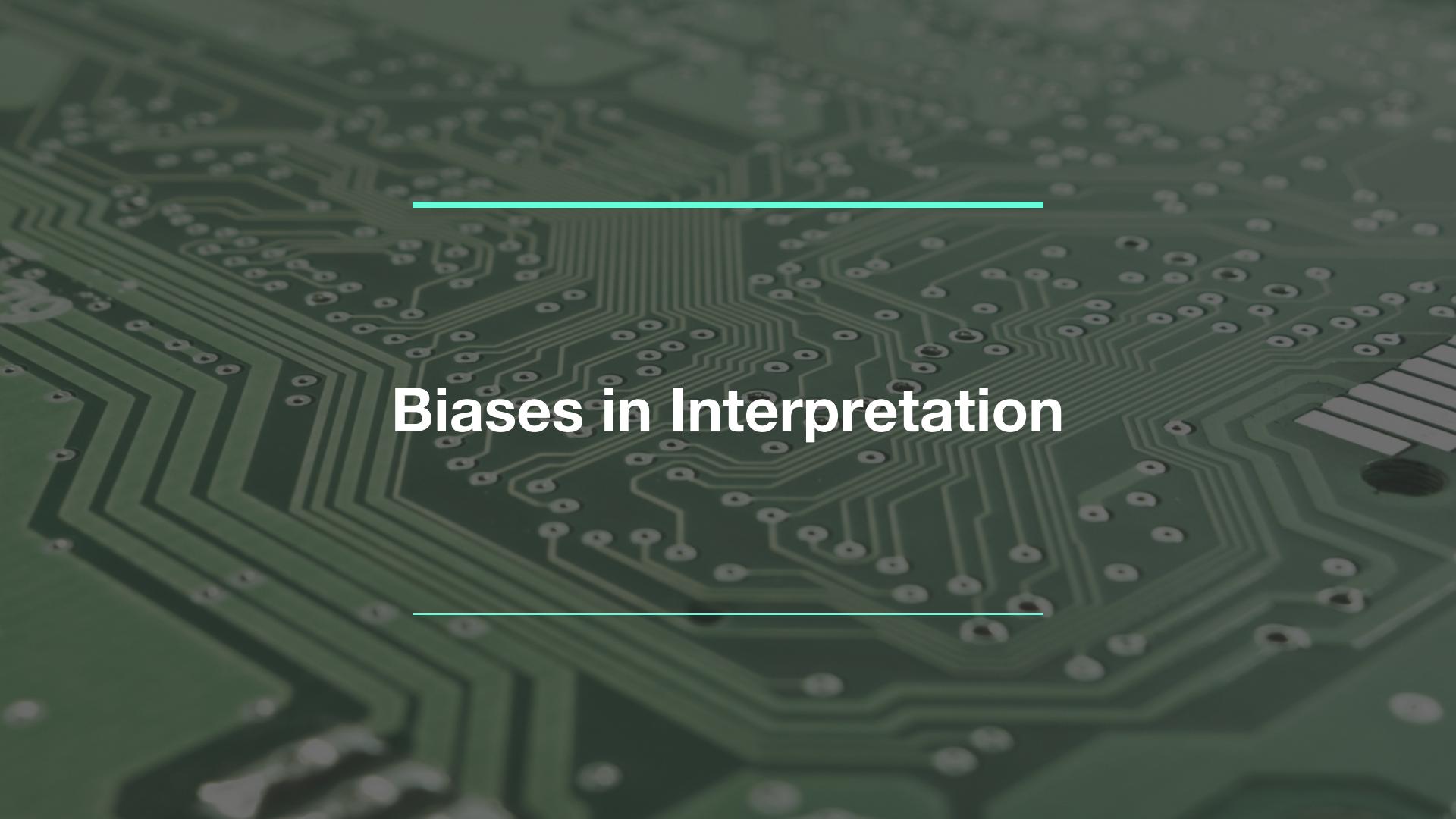
It's possible that you have an appropriate amount of data for every group you can think of but that some groups are represented less positively than others.



# Biases in Data → Biased Labels

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

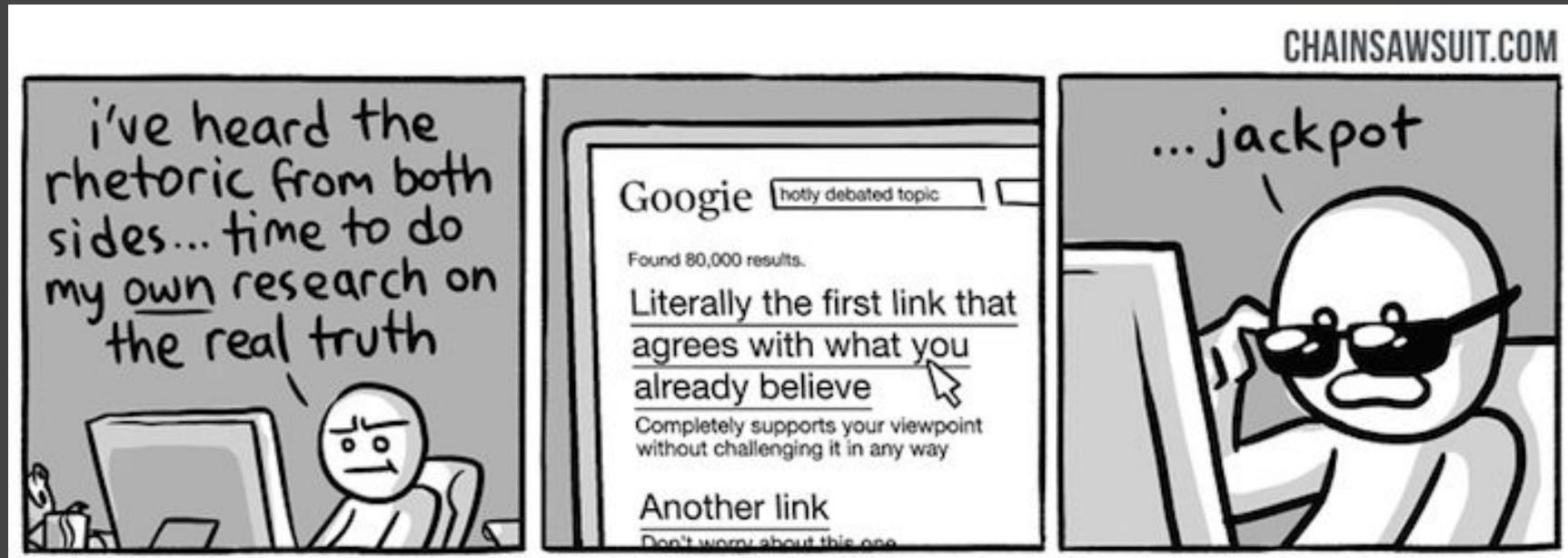




# Biases in Interpretation

# Biases in Interpretation

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



CREDIT

© kris straub - [ChainsawSuit.com](http://ChainsawSuit.com)

# 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

## Post Hoc Ergo Propter Hoc

Women were allowed to vote in the early 1900's and then we had two world wars. Clearly giving them the vote was a bad idea.

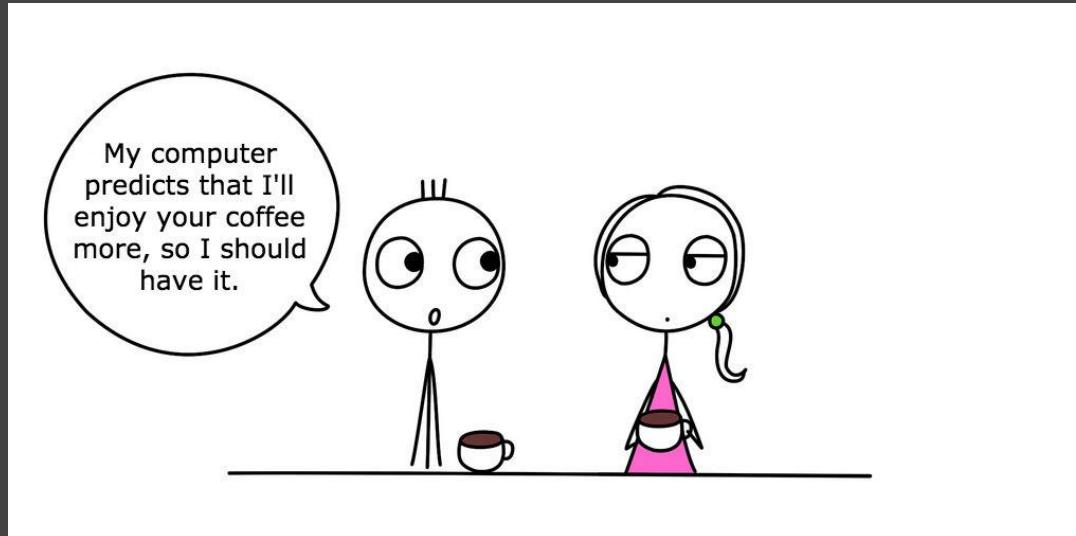


CREDIT

[© mollysdad - Slideshare - Introduction to Logical Fallacies](#)

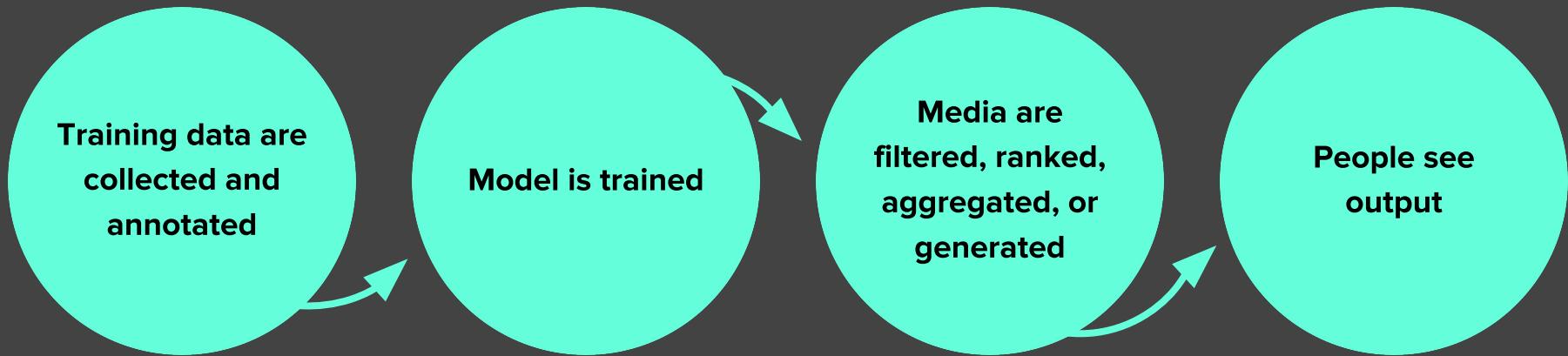
# Biases in Interpretation

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

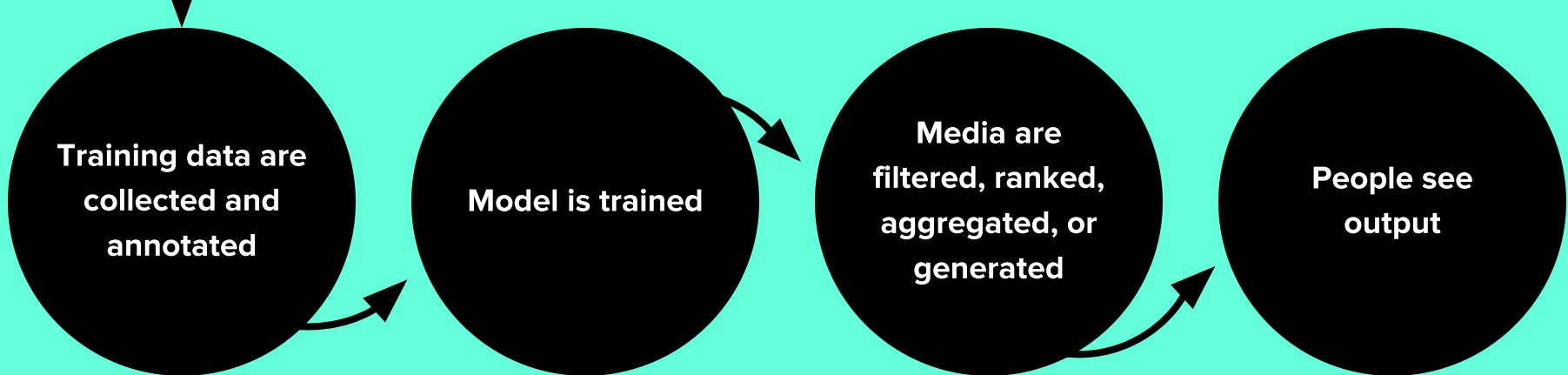


CREDIT

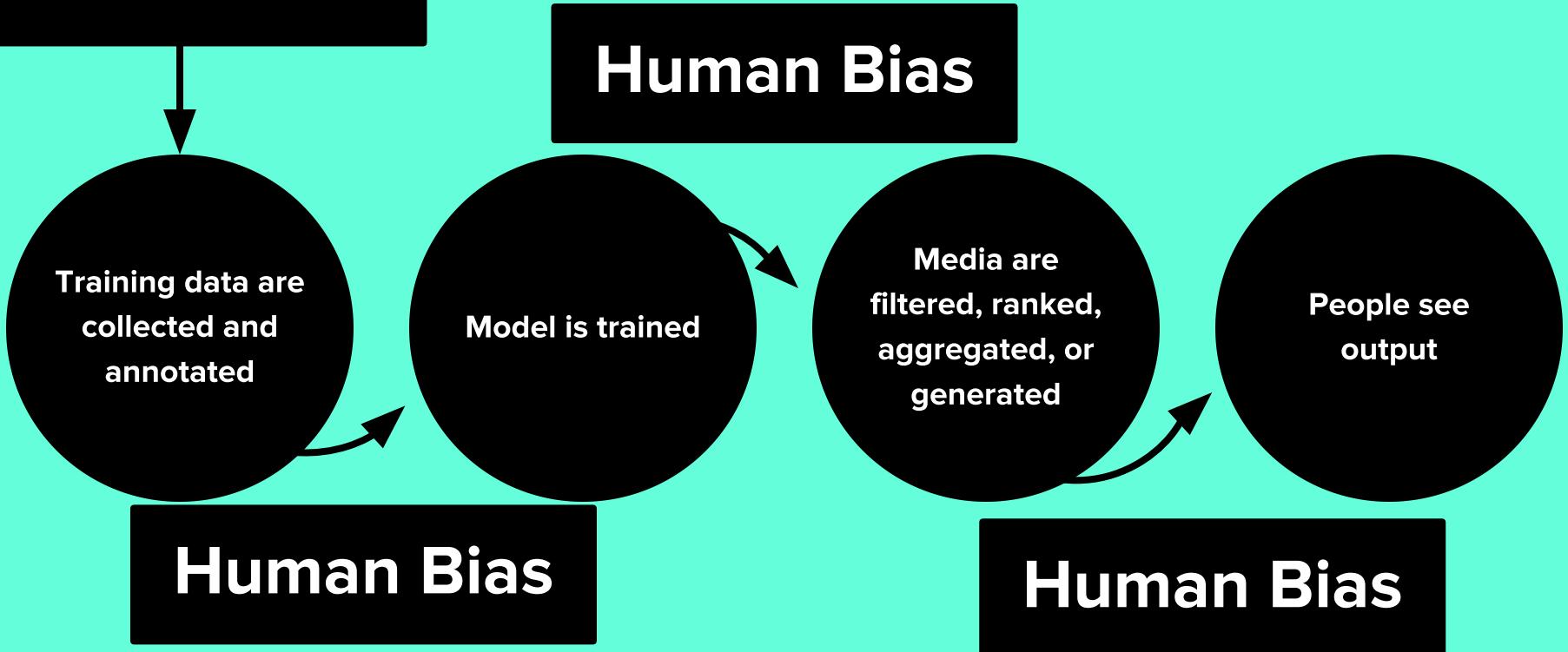
[thedailyenglishshow.com](http://thedailyenglishshow.com) | CC BY 2.0



# 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 Bias



Human Bias

## Bias Network Effect

Bias “Laundering”

Human Bias

Human Bias

Biased data created from process becomes new training data

---

Human data perpetuates human biases.  
As ML learns from human data, the result  
is a **bias network effect**.

---



**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

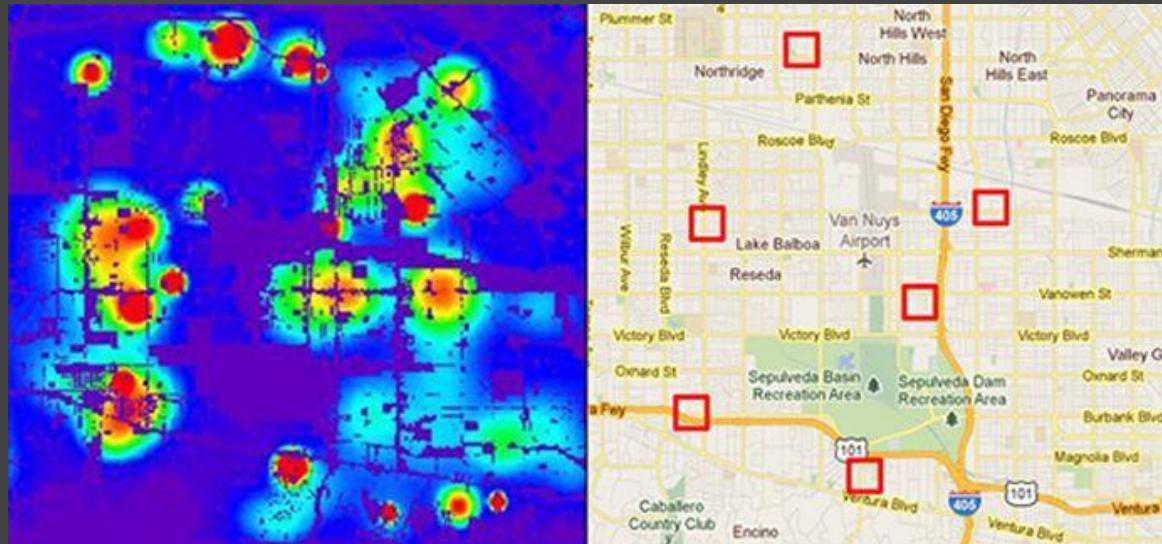
[The Guardian view on machine learning: people must decide](#)



# Predicting Future Criminal Behavior

# Predicting Policing

- Algorithms identify potential crime hot-spots
- Based on where crime is previously reported, not where it is known to have occurred
- Predicts future events from past



CREDIT

[Smithsonian. Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased? 2018](#)

# Predicting Sentencing

- Prater (who is white) rated **low risk** after shoplifting, despite two armed robberies; one attempted armed robbery.
- Borden (who is black) rated **high risk** after she and a friend took (but returned before police arrived) a bike and scooter sitting outside.
- Two years later, Borden has not been charged with any new crimes. Prater serving 8-year prison term for grand theft.

CREDIT

[ProPublica. Northpointe: Risk in Criminal Sentencing. 2016.](#)

---

## **Automation Bias** in face of:

- Overgeneralization
  - Feedback Loops
  - Correlation Fallacy
-

# 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.

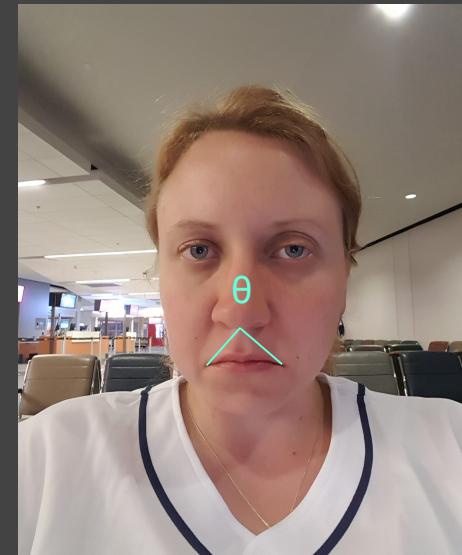
Main clients are in homeland security and public safety.

# 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  $\theta$  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”

---

**Selection Bias + Experimenter's Bias +  
Confirmation Bias + Correlation Fallacy +  
Feedback Loops**

---

# Predicting Criminality - The Media Blitz

## arXiv Paper Spotlight: Automated Inference on Criminality Using Face ...

[www.kdnuggets.com/.../arxiv-spotlight-automated-inference-criminality-face-images....](http://www.kdnuggets.com/.../arxiv-spotlight-automated-inference-criminality-face-images....) ▾

A recent paper by Xiaolin Wu (McMaster University, Shanghai Jiao Tong University) and Xi Zhang (Shanghai Jiao Tong University), titled "Automated Inference ...

## Automated Inference on Criminality Using Face Images | Hacker News

<https://news.ycombinator.com/item?id=12983827> ▾

Nov 18, 2016 - The automated inference on criminality eliminates the variable of meta-accuracy (the competence of the human judge/examiner) all together.

## A New Program Judges If You're a Criminal From Your Facial Features ...

<https://motherboard.vice.com/.../new-program-decides-criminality-from-facial-feature...> ▾

Nov 18, 2016 - In their paper 'Automated Inference on Criminality using Face Images', published on the arXiv pre-print server, Xiaolin Wu and Xi Zhang from ...

## Can face classifiers make a reliable inference on criminality?

<https://techxplore.com/.../Computer Sciences> ▾

Nov 23, 2016 - Their paper is titled "Automated Inference on Criminality using Face Images ... face classifiers are able to make reliable inference on criminality.

## Troubling Study Says Artificial Intelligence Can Predict Who Will Be ...

<https://theintercept.com/.../troubling-study-says-artificial-intelligence-can-predict-who...> ▾

Nov 18, 2016 - Not so in the modern age of Artificial Intelligence, apparently: In a paper titled "Automated Inference on Criminality using Face Images," two ...

## Automated Inference on Criminality using Face Images (via arXiv ...

<https://computationallegalstudies.com/.../automated-inference-on-criminality-using-fa...> ▾

Dec 6, 2016 - Next Next post: A General Approach for Predicting the Behavior of the Supreme Court of the United States (Paper Version 2.01) (Katz, ...

---

# (Claiming to) Predict Internal Qualities Subject To Discrimination

---

# Predicting Homosexuality

Composite Straight Faces

Composite Gay Faces



- 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.**

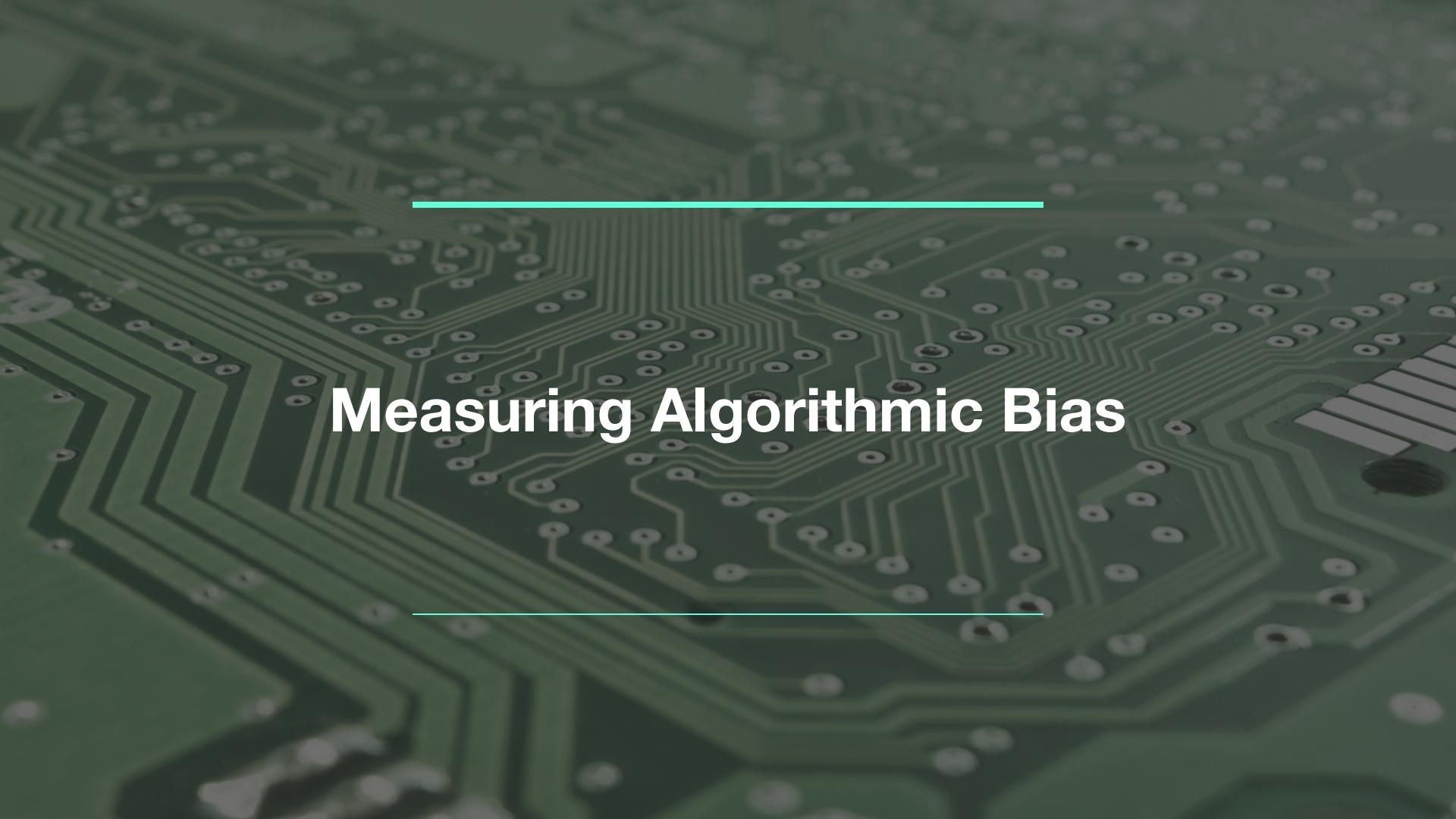
See our longer response on Medium, [“Do Algorithms Reveal Sexual Orientation or Just Expose our Stereotypes?”](#)



---

**Selection Bias + Experimenter's Bias +  
Correlation Fallacy**

---



# Measuring Algorithmic 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

# 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	<ul style="list-style-type: none"><li>• Exists</li><li>• Predicted</li></ul> <p><b>True Positives</b></p>	
	Negative		<ul style="list-style-type: none"><li>• Doesn't exist</li><li>• Not predicted</li></ul> <p><b>True Negatives</b></p>

# Evaluate for Fairness & Inclusion: Confusion Matrix

		Model Predictions	
		Positive	Negative
References	Positive	<ul style="list-style-type: none"><li>• Exists</li><li>• Predicted</li></ul> <p><b>True Positives</b></p>	<ul style="list-style-type: none"><li>• Exists</li><li>• Not predicted</li></ul> <p><b>False Negatives</b></p>
	Negative	<ul style="list-style-type: none"><li>• Doesn't exist</li><li>• Predicted</li></ul> <p><b>False Positives</b></p>	<ul style="list-style-type: none"><li>• Doesn't exist</li><li>• Not predicted</li></ul> <p><b>True Negatives</b></p>

# Evaluate for Fairness & Inclusion: Confusion Matrix

		Model Predictions		
		Positive	Negative	
References	Positive	<ul style="list-style-type: none"><li>• Exists</li><li>• Predicted</li></ul> <b>True Positives</b>	<ul style="list-style-type: none"><li>• Exists</li><li>• Not predicted</li></ul> <b>False Negatives</b>	Recall, False Negative Rate
	Negative	<ul style="list-style-type: none"><li>• Doesn't exist</li><li>• Predicted</li></ul> <b>False Positives</b>	<ul style="list-style-type: none"><li>• Doesn't exist</li><li>• Not predicted</li></ul> <b>True Negatives</b>	
		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

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

Male Patient Results

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

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

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

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

“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

Male Patient Results

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

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

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

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

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

“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.

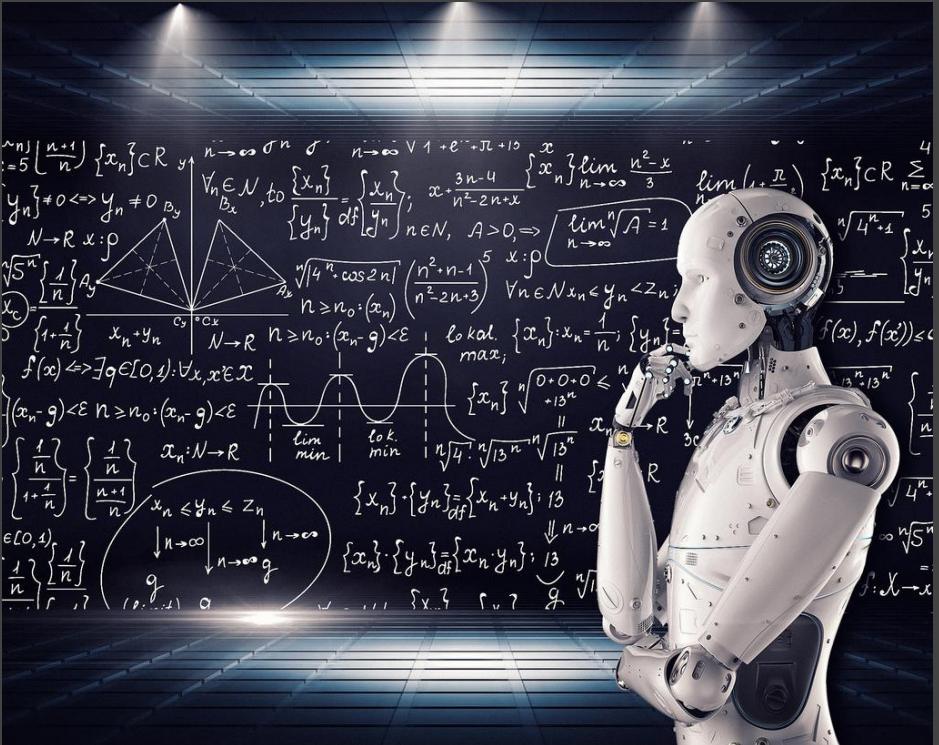
**False Positive:** Email flagged as SPAM is removed from your inbox.

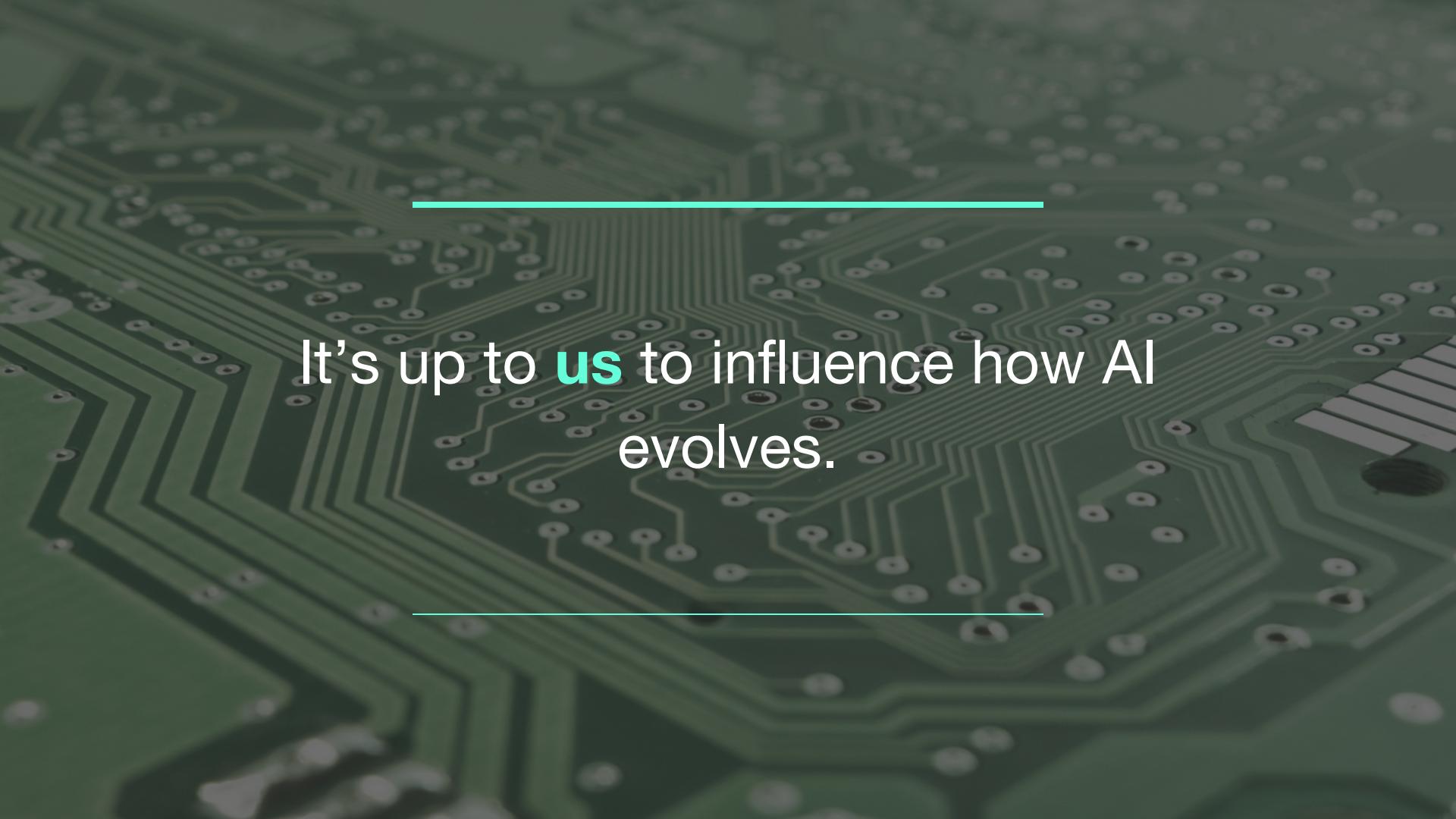
If it's from a friend or loved one, it's a loss!



# 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**





---

It's up to **us** to influence how AI evolves.

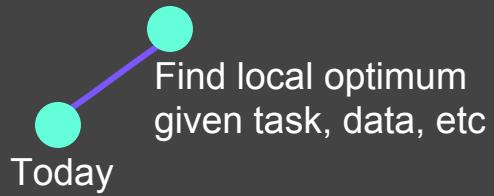
---



Today

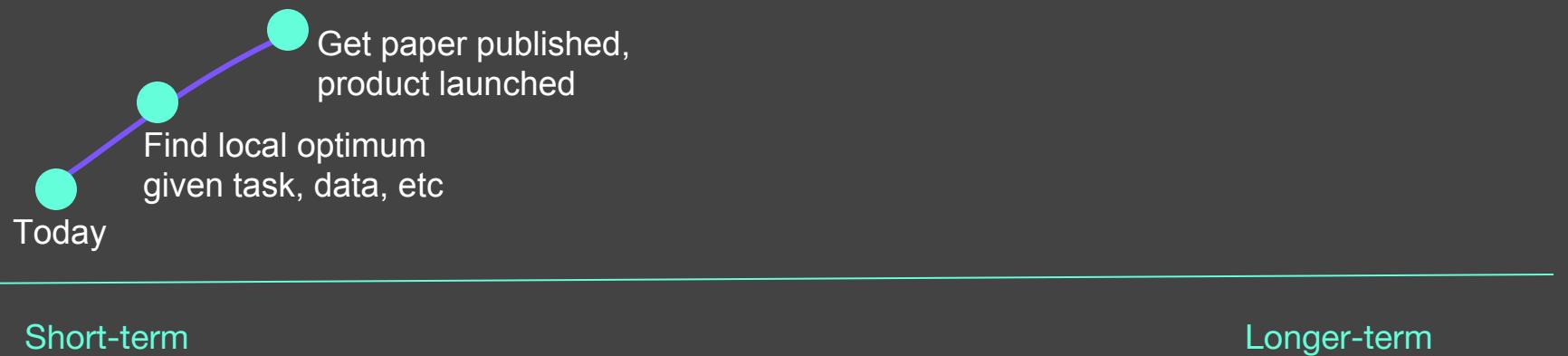
Short-term

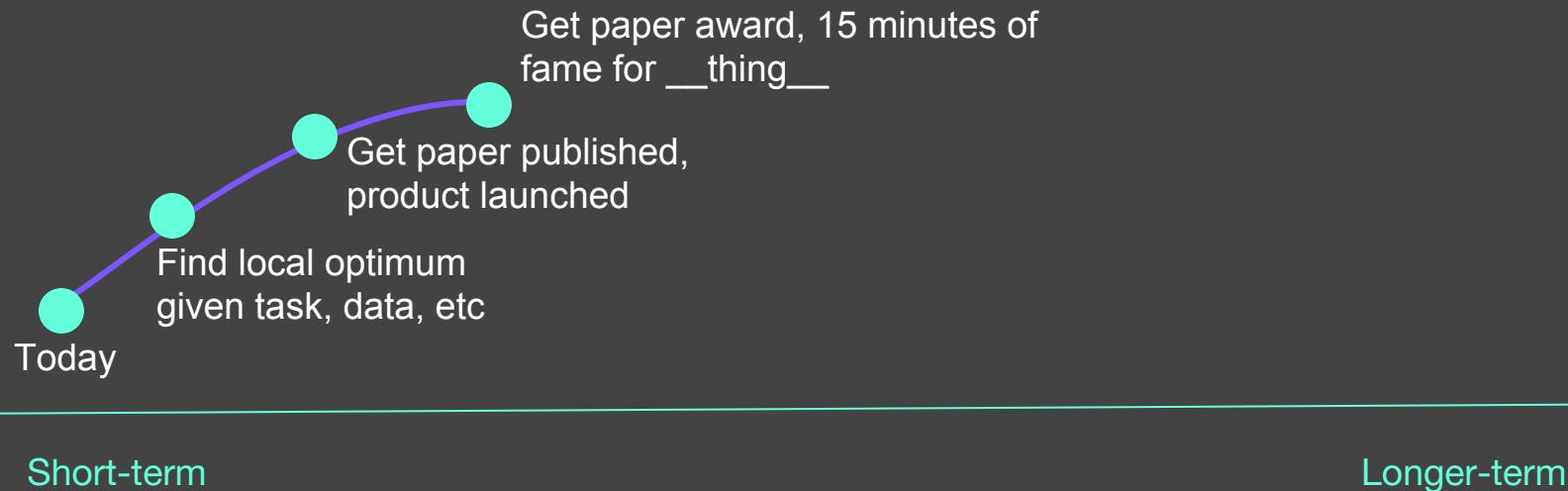
Longer-term

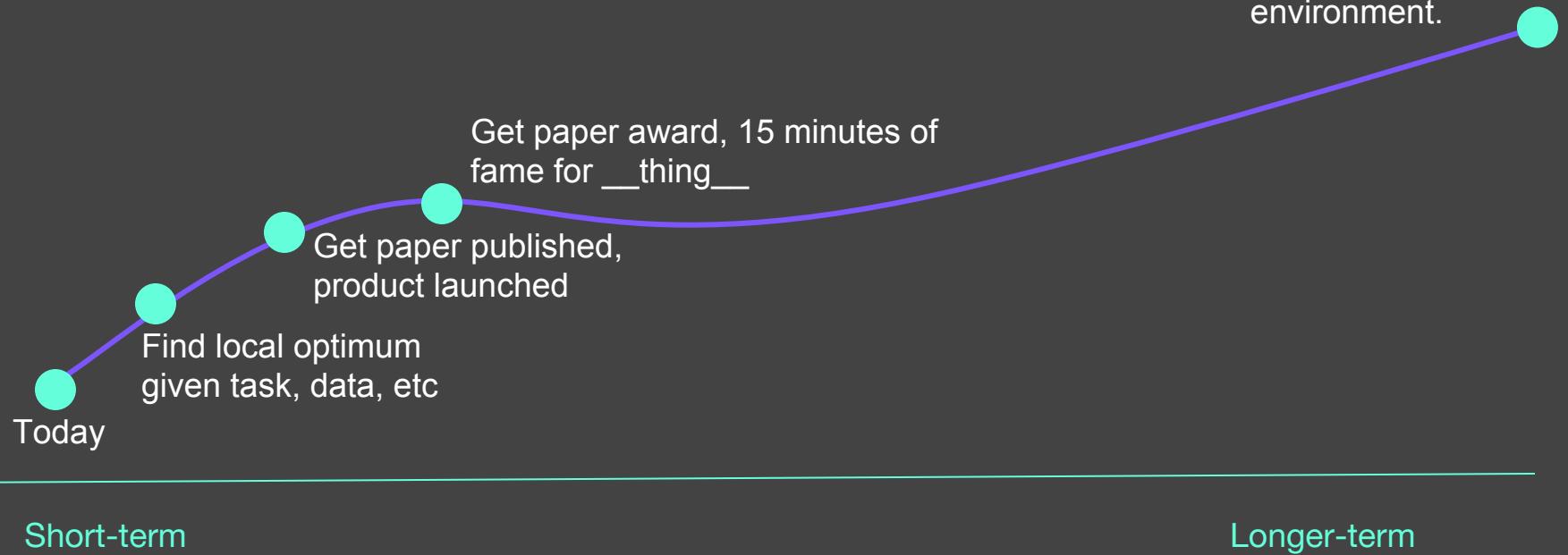


Short-term

Longer-term







# Begin tracing out paths for the evolution of ethical AI

Today

Short-term

Get paper award, 15 minutes of fame for \_\_thing\_\_

Get paper published, product launched

Find local optimum given task, data, etc

Positive outcomes for humans and their environment.

Longer-term

*How can the work I'm interested in now be best focused to help others?*

---

It's up to **us** to influence how AI evolves.

Here are some things we can do.

---



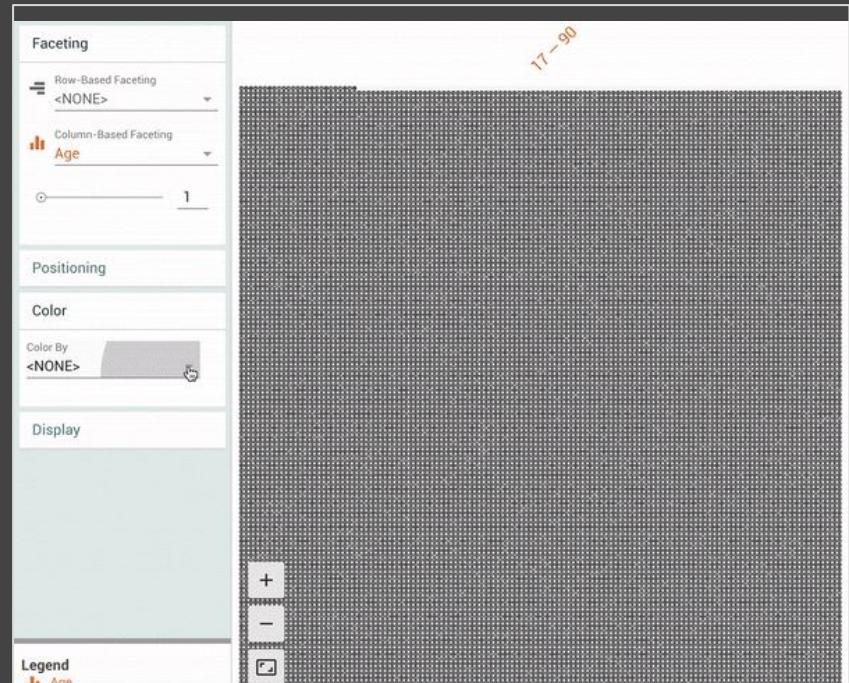
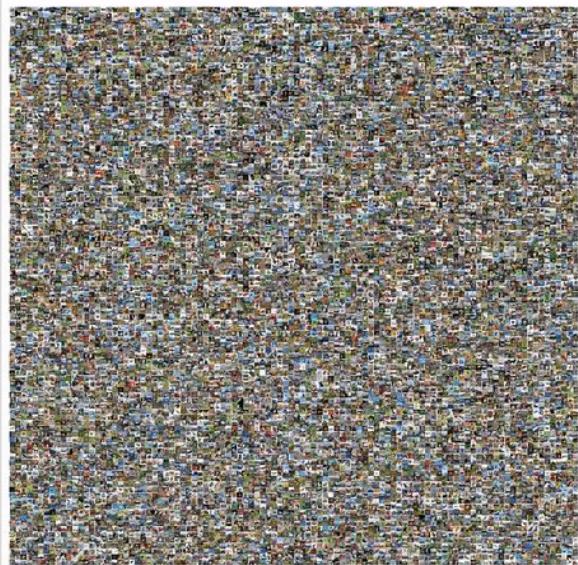
Data

# Data Really, Really Matters

- Understand your Data: skews, correlations
- Abandon single training-set / testing-set from similar distribution
- Combine inputs from multiple sources
- Use **held-out test set** for hard use cases
- Talk to experts about additional signals



# Understand Your Data Skews

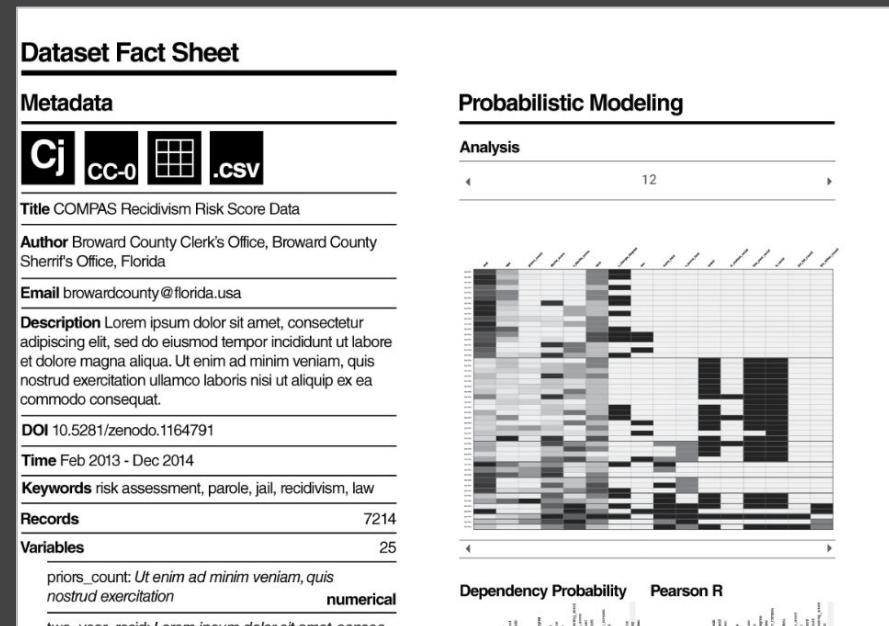


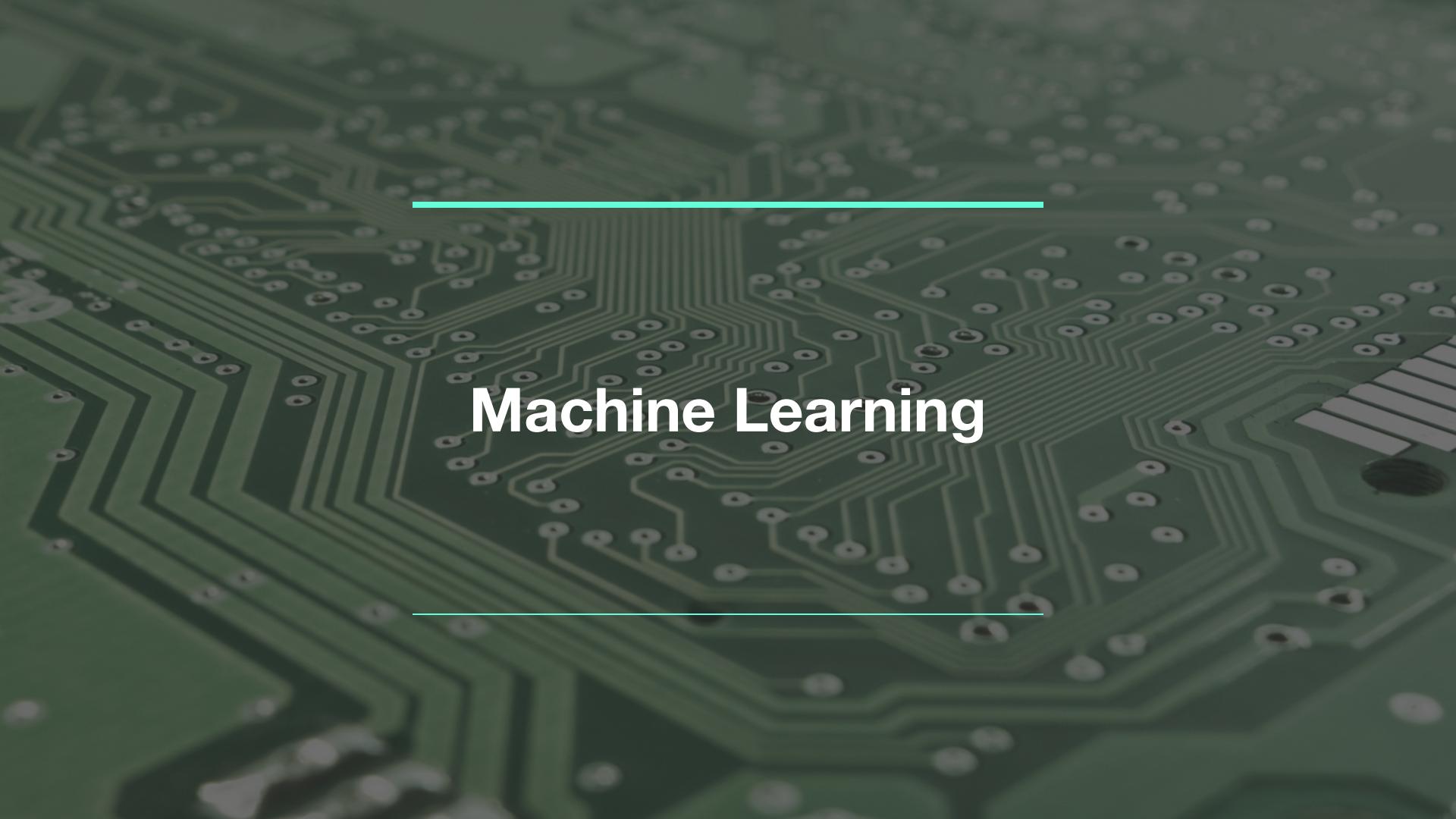
Facets: pair-code.github.io

# Datasheets for Datasets

Timnit Gebru<sup>1</sup> Jamie Morgenstern<sup>2</sup> Briana Vecchione<sup>3</sup> Jennifer Wortman Vaughan<sup>1</sup> Hanna Wallach<sup>1</sup>  
Hal Daumé III<sup>1,4</sup> Kate Crawford<sup>1,5</sup>

Datasheets for Datasets	
<b>Motivation for Dataset Creation</b>	<b>Data Collection Process</b>
Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)	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?)
What (other) tasks could the dataset be used for? Are there obvious tasks for which it should <i>not</i> be used?	Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)
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)?	Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?
Who funded the creation of the dataset? If there is an associated grant, provide the grant number.	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?
Any other comments?	Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?
<b>Dataset Composition</b>	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?
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?	





# Machine Learning

# Use ML Techniques for Bias Mitigation and Inclusion

## Bias Mitigation

- Removing the signal for problematic output
  - Stereotyping
  - Sexism, Racism, \*-ism
  - “Debiasing”

# Use ML Techniques for Bias Mitigation and Inclusion

## Bias Mitigation

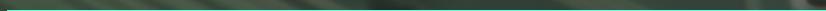
- Removing the signal for problematic output
  - Stereotyping
  - Sexism, Racism, \*-ism
  - “Debiasing”

## Inclusion

- Adding signal for desired variables
  - Increasing model performance
  - Attention to subgroups or data slices with worst performance



## Multi-task Learning to Increase Inclusion



# Multiple Tasks + Deep Learning for Inclusion: Multi-task Learning Example

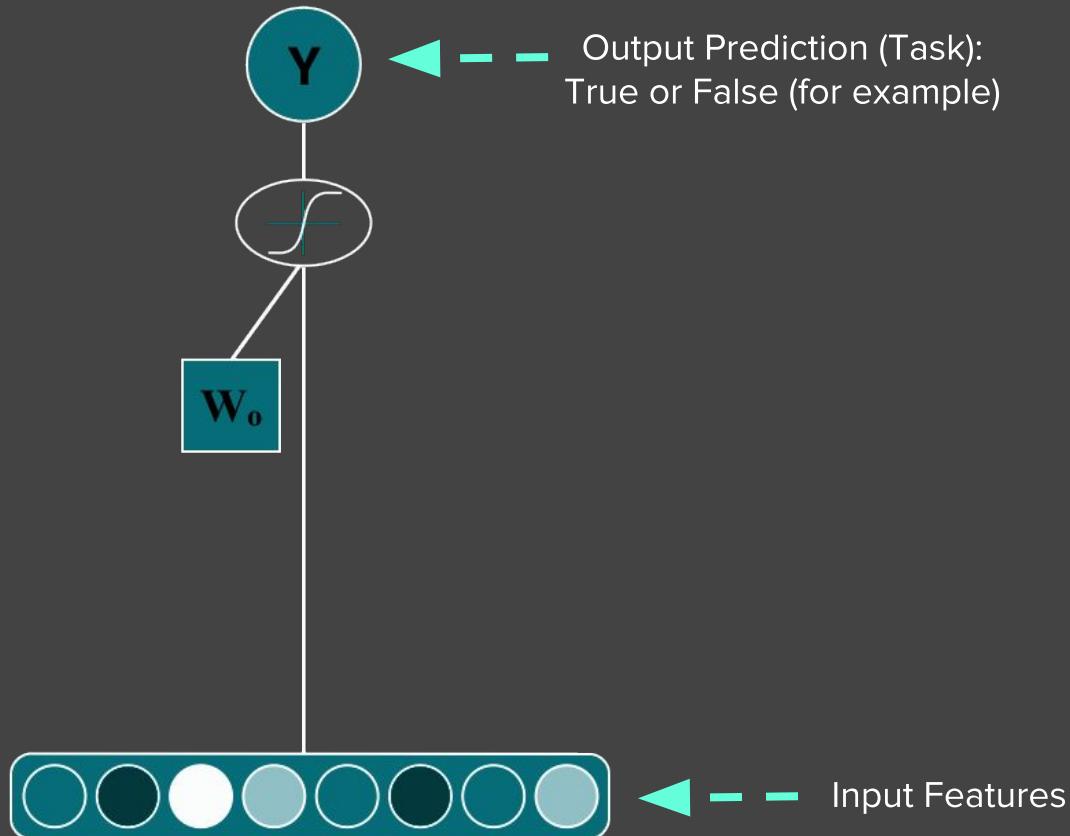
- Collaboration with UPenn WWP
- Working directly with clinicians
- Goals:
  - System that can alert clinicians if suicide attempt is **imminent**
  - Feasibility of diagnoses when few training instances are available



# Multiple Tasks + Deep Learning for Inclusion: Multi-task Learning Example

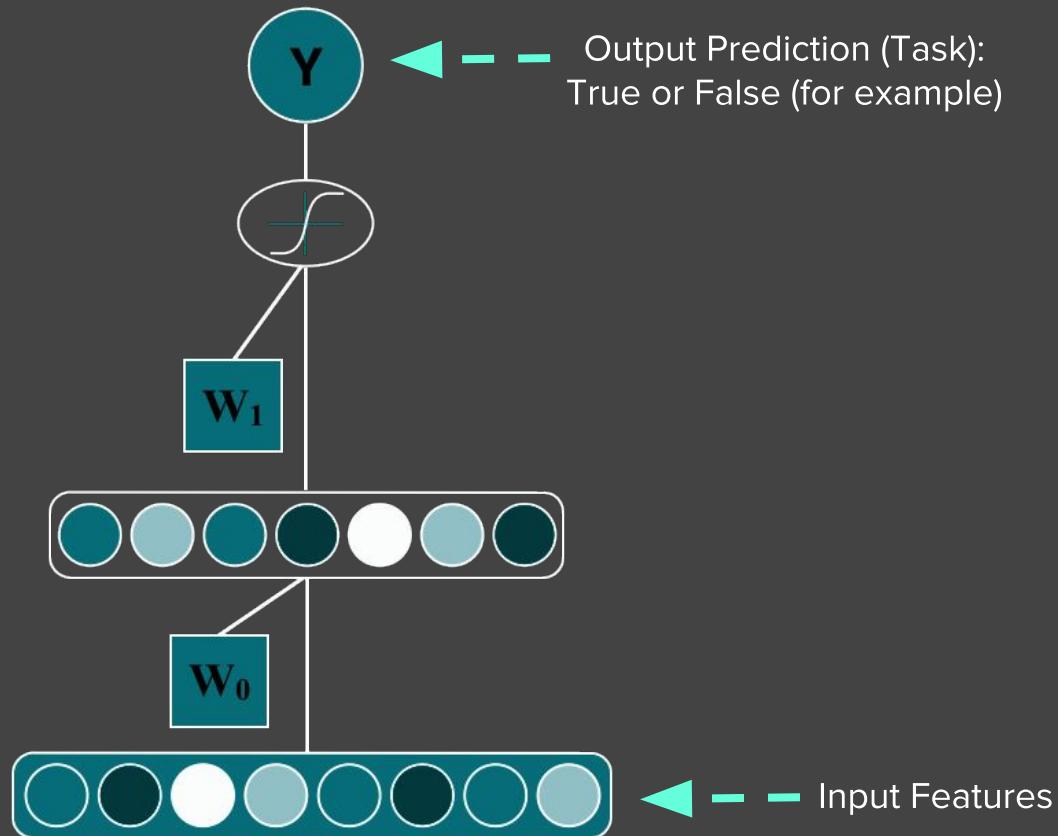
- **Internal Data:**
  - Electronic Health Records
    - Patient or patient family provided
    - Including mental health diagnoses, suicide attempts, and completions
  - Social Media data
- **Proxy Data:**
  - Twitter media data
  - Proxy mental health diagnoses using self-declared diagnoses in tweets
    - “I’ve been diagnosed with X”
    - “I tried to commit suicide”

# Single-Task: Logistic Regression

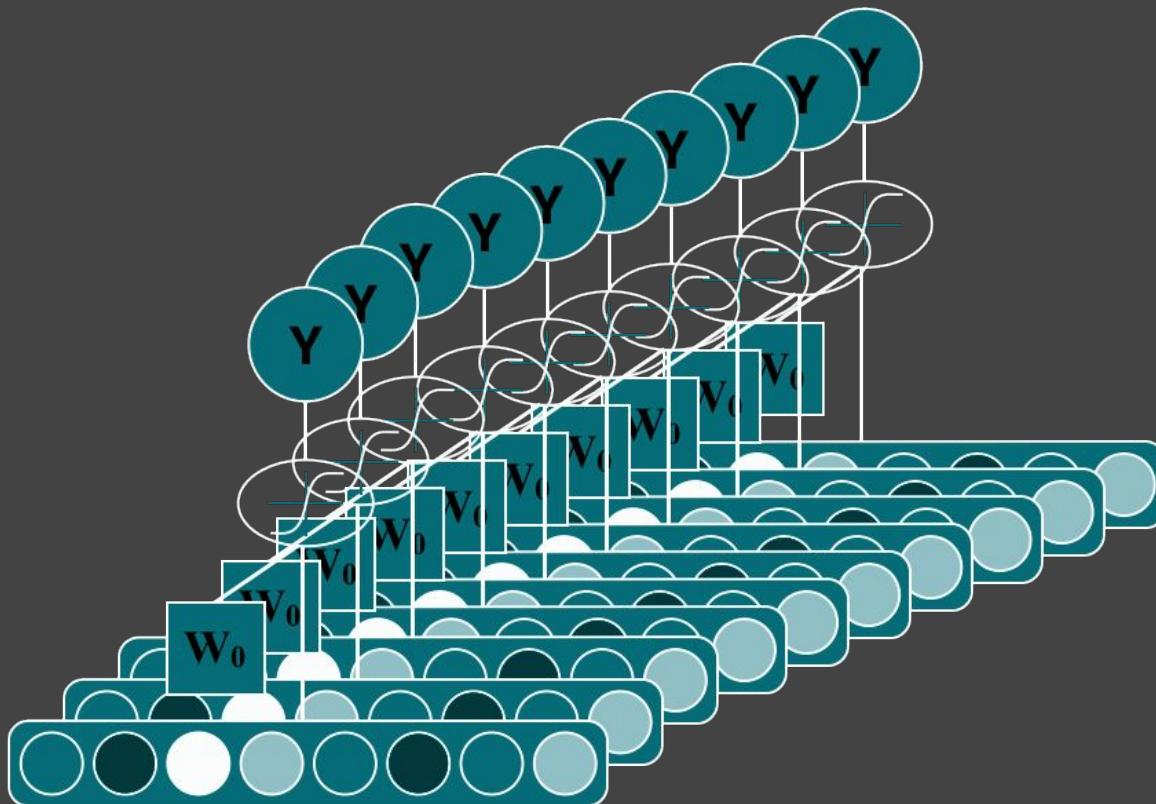


# Single-Task: Deep Learning

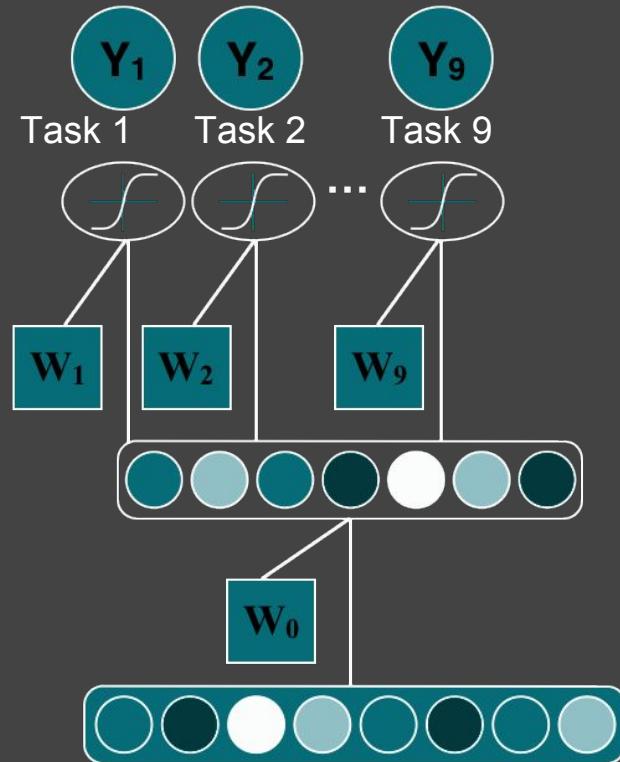
Fancier!!



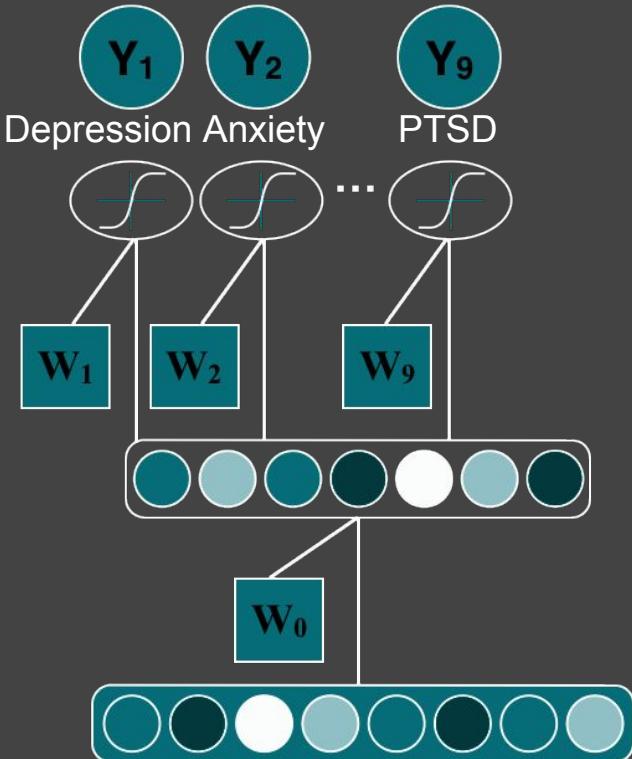
# Multiple Tasks with Basic Logistic Regression



# Multi-task Learning



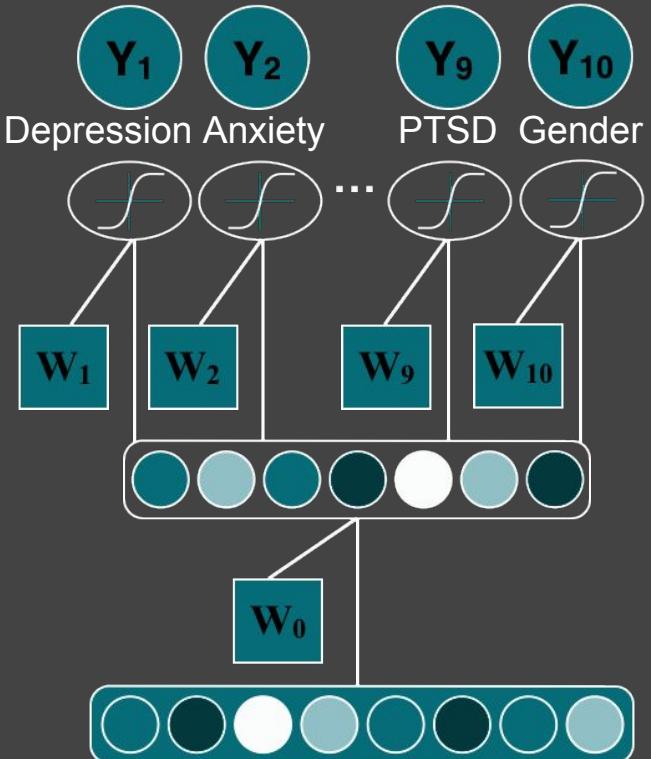
# Multi-task Learning



Task	N
Neurotypicality	4791
Anxiety	2407
Depression	1400
Suicide attempt	1208
Eating disorder	749
Schizophrenia	349
Panic disorder	263
PTSD	248
Bipolar disorder	191
All	9611

<5% positive examples

# Multi-task Learning

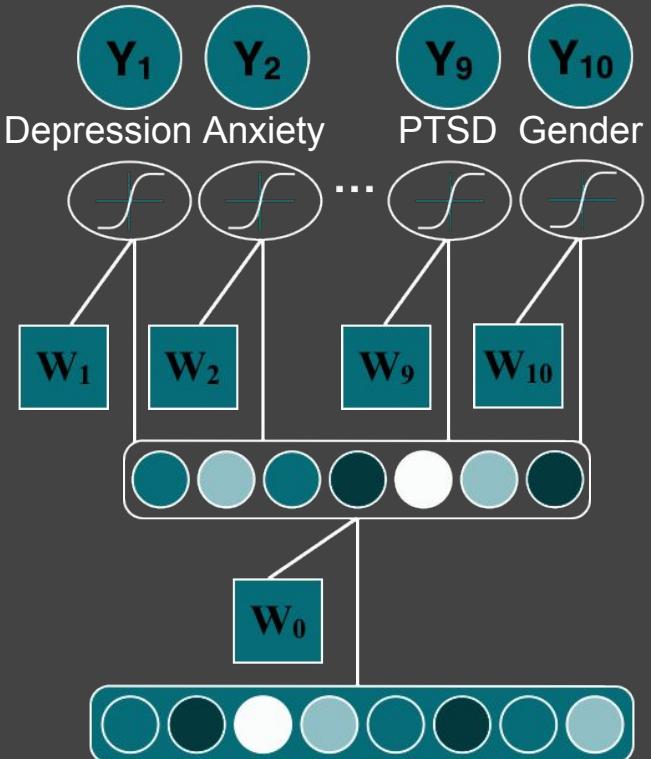


Task	N
Gender	1101
Neurotypicality	4791
Anxiety	2407
Depression	1400
Suicide attempt	1208
Eating disorder	749
Schizophrenia	349
Panic disorder	263
PTSD	248
Bipolar disorder	191
All	9611

} <5% positive examples

# Multi-task Learning

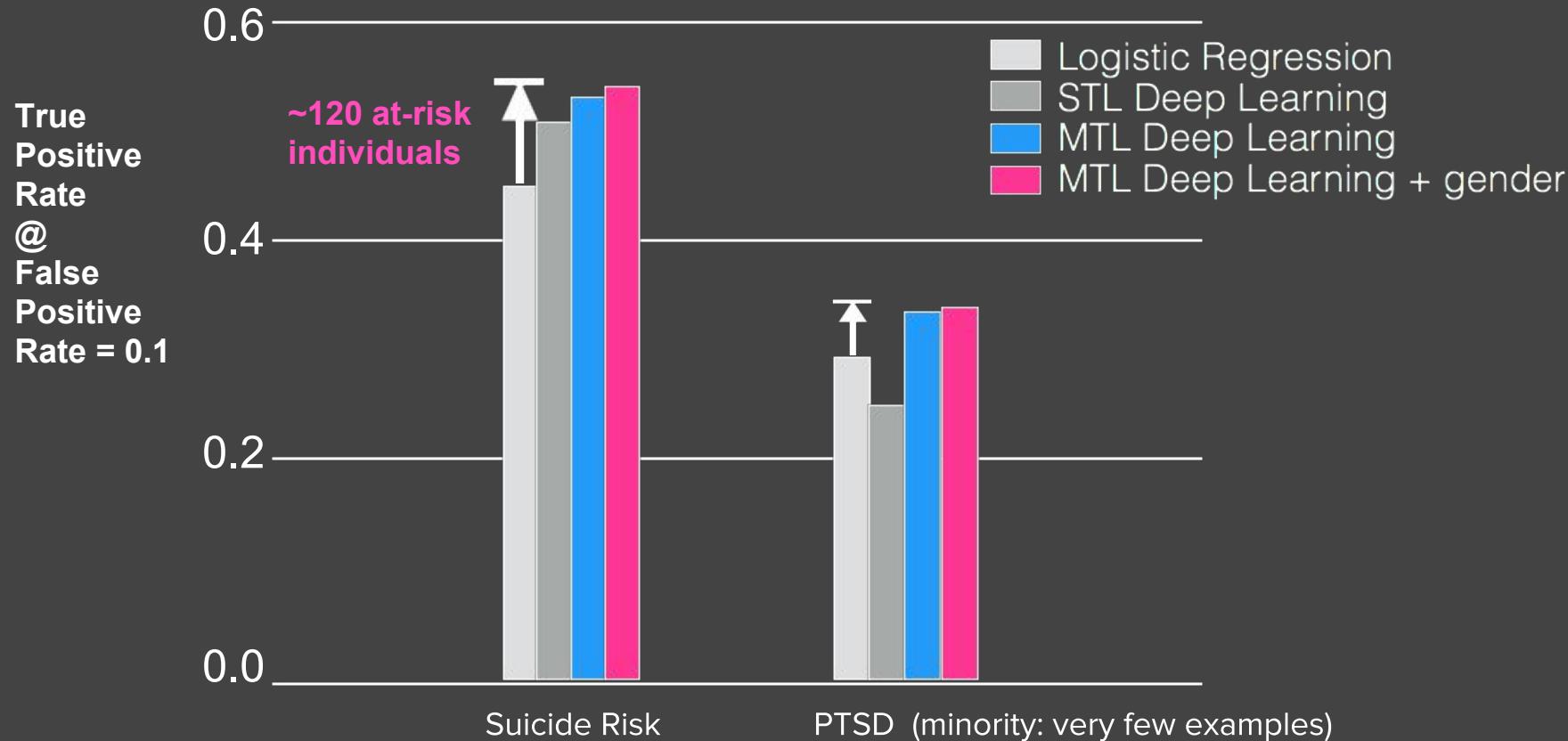
Multitask, given  
cormorbidity



Task	N
Gender	1101
Neurotypicality	4791
Anxiety	2407
Depression	1400
Suicide attempt	1208
Eating disorder	749
Schizophrenia	349
Panic disorder	263
PTSD	248
Bipolar disorder	191
All	9611

} <5% positive examples

# Improved Performance across Subgroups



# Reading for the masses....

## Multi-Task Learning for Mental Health using Social Media Text

**Adrian Benton**  
Johns Hopkins University  
[adrian@cs.jhu.edu](mailto:adrian@cs.jhu.edu)

**Margaret Mitchell**  
Microsoft Research\*  
[mmitchellai@google.com](mailto:mmitchellai@google.com)

**Dirk Hovy**  
University of Copenhagen  
[mail@dirkhovy.com](mailto:mail@dirkhovy.com)

## Contextualizing and considering ethical dimensions

## 2 Ethical Considerations

As with any author-attribute detection, there is the danger of abusing the model to single out people (*overgeneralization*, see Hovy and Spruit (2016)). We are aware of this danger, and sought to minimize the risk. For this reason, we don't provide a selection of features or representative examples. The experiments in this paper were performed with a clinical application in mind, and use carefully matched (but anonymized) data, so the distribution is not representative of the population as a whole. The results of this paper should therefore *not* be interpreted as a means to assess mental health conditions in social media in general, but as a test for the applicability of MTL in a well-defined clinical setting.

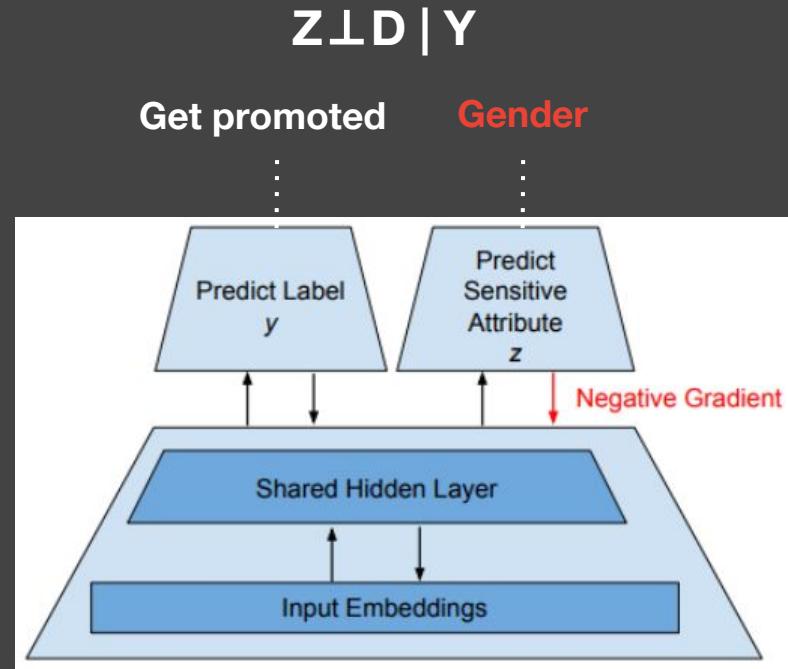


## Adversarial Multi-task Learning to Mitigate Bias

# Multitask Adversarial Learning

- Basic idea: Jointly predict:
  - Output decision  $D$
  - Attribute you'd like to remove from decision  $Z$
  - Negate the effect of the undesired attribute

$$P(\hat{Y} = 1 | Y = 1, Z = 1) =$$
$$P(\hat{Y} = 1 | Y = 1, Z = 0)$$



---

# Equality of Opportunity in Supervised Learning

A classifier's output decision should be the same **across sensitive characteristics**, given what the correct decision should be.

---



# Case Study: Conversation AI Toxicity

# Measuring and Mitigating Unintended Bias in Text Classification

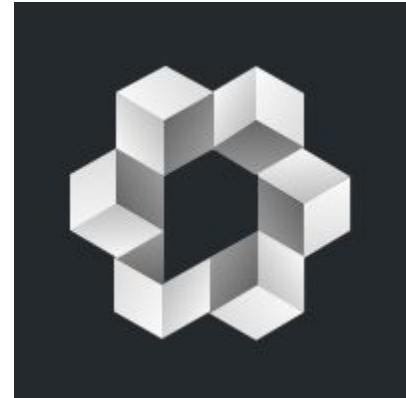
Lucas Dixon  
[ldixon@google.com](mailto:ldixon@google.com)

John Li  
[jetpack@google.com](mailto:jetpack@google.com)

Jeffrey Sorensen  
[sorenj@google.com](mailto:sorenj@google.com)

Nithum Thain  
[nthain@google.com](mailto:nthain@google.com)

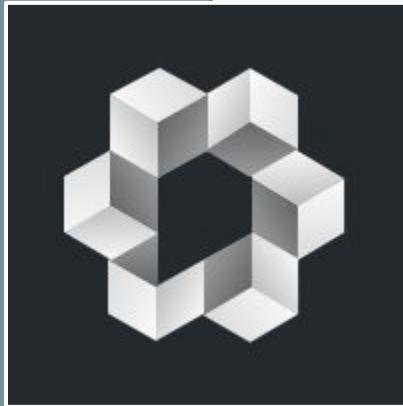
Lucy Vasserman  
[lucyvasserman@google.com](mailto:lucyvasserman@google.com)



AIES, 2018 and FAT\*, 2019

## **Conversation-AI**

ML to improve online  
conversations at scale



## **Research Collaboration**

Jigsaw, CAT, several  
Google-internal teams, and  
external partners (NYTimes,  
Wikimedia, etc)

# Perspective API

“You’re a dork!”



Toxicity: 0.91



API

Data + ML  
Toxicity,  
Severe Toxicity,  
Threat, Off-topic,  
+ dozens other  
models

# Unintended Bias

Model falsely associates frequently attacked identities with toxicity: *False Positive Bias*

<u>Sentence</u>	<u>model score</u>
"i'm a proud <b>tall</b> person"	0.18
"i'm a proud <b>lesbian</b> person"	0.51
"i'm a proud <b>gay</b> person"	0.69

# Bias Source and Mitigation

*Bias caused by dataset imbalance*

- Frequently attacked identities are overrepresented in toxic comments
- Length matters

Add *assumed non-toxic data* from Wikipedia articles to fix the imbalance.

- Original dataset had 127,820 examples
- 4,620 non-toxic examples added

Term	Comment Length				
	20-59	60-179	180-539	540-1619	1620-4859
<b>ALL</b>	17%	12%	7%	5%	5%
<b>gay</b>	88%	77%	51%	30%	19%
<b>queer</b>	75%	83%	45%	56%	0%
<b>homosexual</b>	78%	72%	43%	16%	15%
<b>black</b>	50%	30%	12%	8%	4%
<b>white</b>	20%	24%	16%	12%	2%
<b>wikipedia</b>	39%	20%	14%	11%	7%
<b>atheist</b>	0%	20%	9%	6%	0%
<b>lesbian</b>	33%	50%	42%	21%	0%
<b>feminist</b>	0%	20%	25%	0%	0%
<b>islam</b>	50%	43%	12%	12%	0%
<b>muslim</b>	0%	25%	21%	12%	17%
<b>race</b>	20%	25%	12%	10%	6%
<b>news</b>	0%	1%	4%	3%	3%
<b>daughter</b>	0%	7%	0%	7%	0%

# Measuring Unintended Bias - Synthetic Datasets

## Challenges with real data:

- Existing datasets are small and/or have false correlations
- Each example is completely unique: not easy to compare for bias

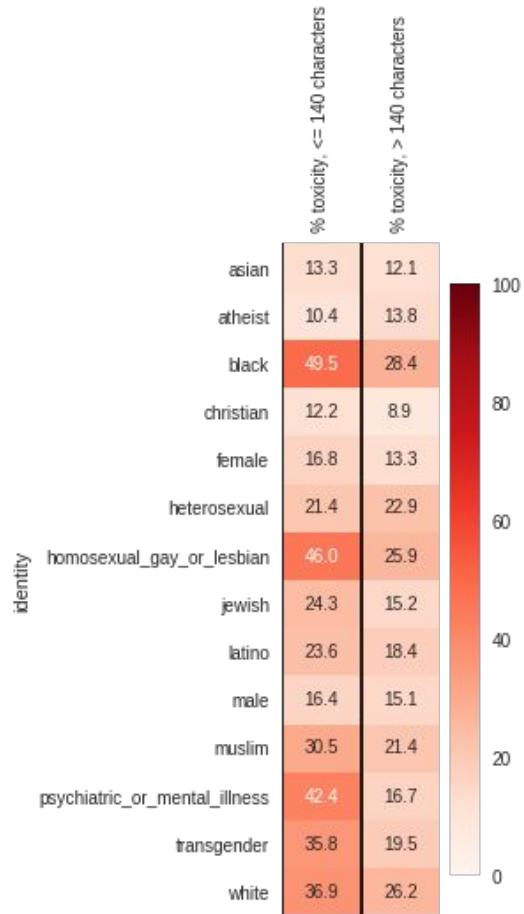
Approach: "bias madlibs": a synthetically generated 'templated' dataset for evaluation

<u>Sentence</u>	<u>model score</u>
"i'm a proud <b>tall</b> person"	0.18
"i'm a proud <b>lesbian</b> person"	0.51
"i'm a proud <b>gay</b> person"	0.69
"audre is a <b>brazilian</b> computer programmer"	0.02
"audre is a <b>muslim</b> computer programmer"	0.08
"audre is a <b>transgender</b> computer programmer"	0.56

# Assumptions

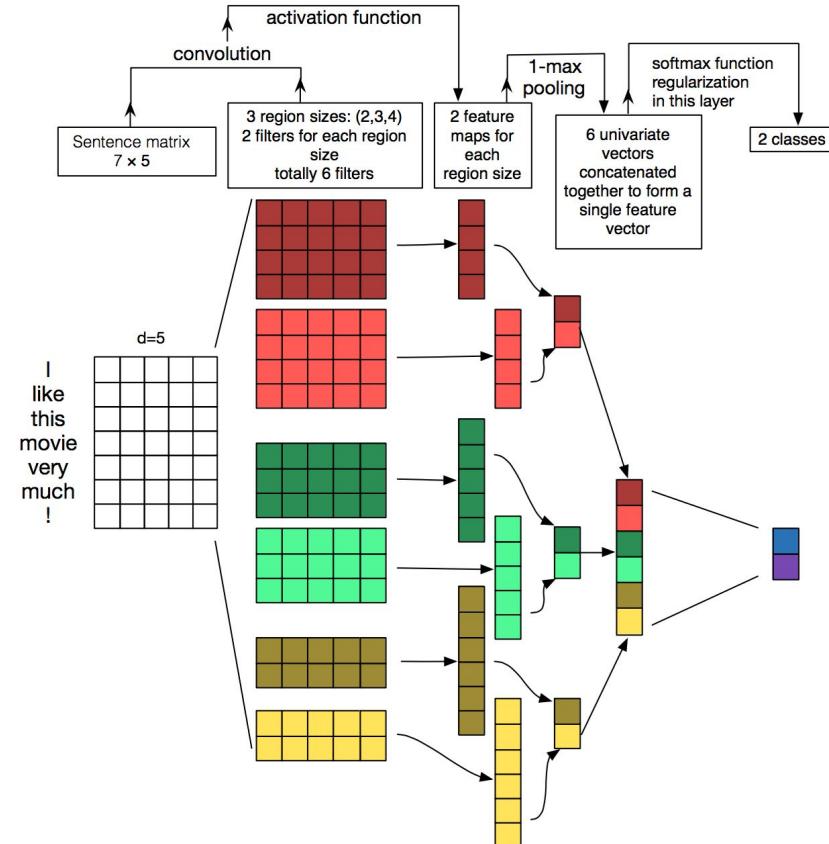
Dataset is reliable:

- Similar distribution as application
- Ignores annotator bias
- No causal analysis



# Deep Learning Model

- CNN architecture
- Pretrained GloVe Embeddings
- Keras Implementation

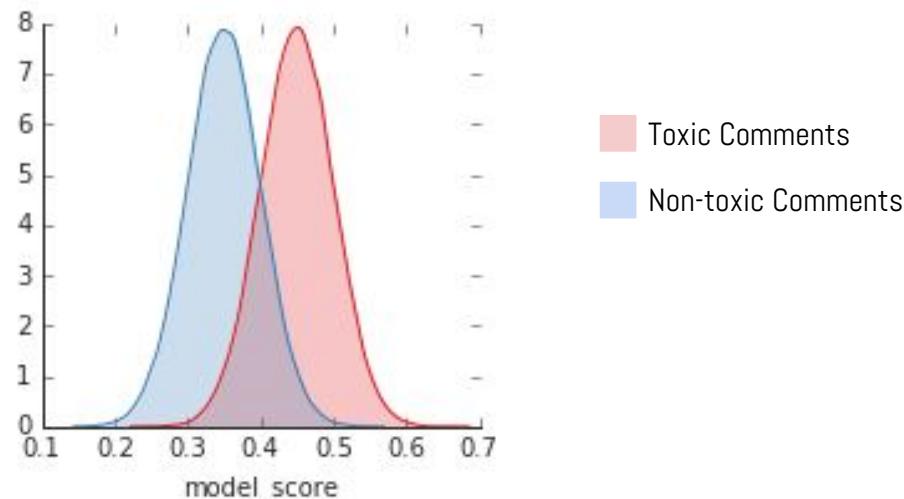


Source: Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.

# Measuring Model Performance

*How good is the model at distinguishing good from bad examples? (ROC-AUC)*

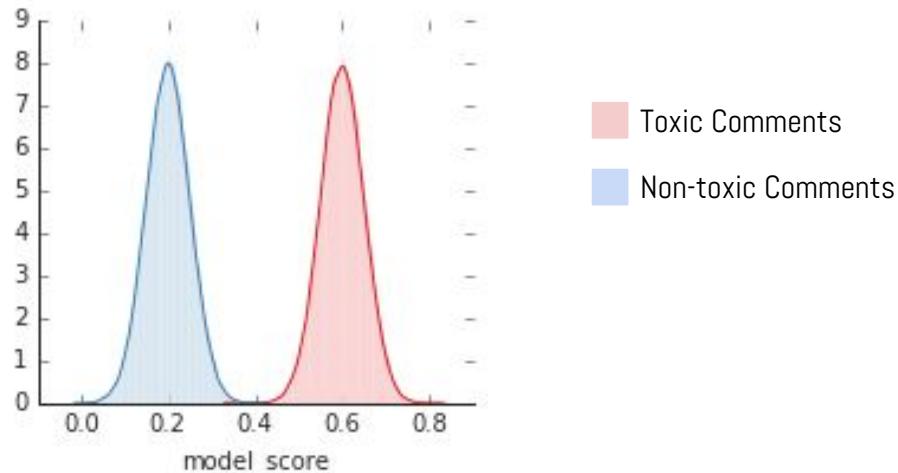
AUC (for a given test set) = Given two randomly chosen examples, one in-class (e.g. one is toxic and the other is not), AUC is the probability that the model will give the in-class example the higher score.



# Measuring Model Performance

*How good is the model at distinguishing good from bad examples? (ROC-AUC)*

AUC (for a given test set) = Given two randomly chosen examples, one in-class (e.g. one is toxic and the other is not), AUC is the probability that the model will give the in-class example the higher score.

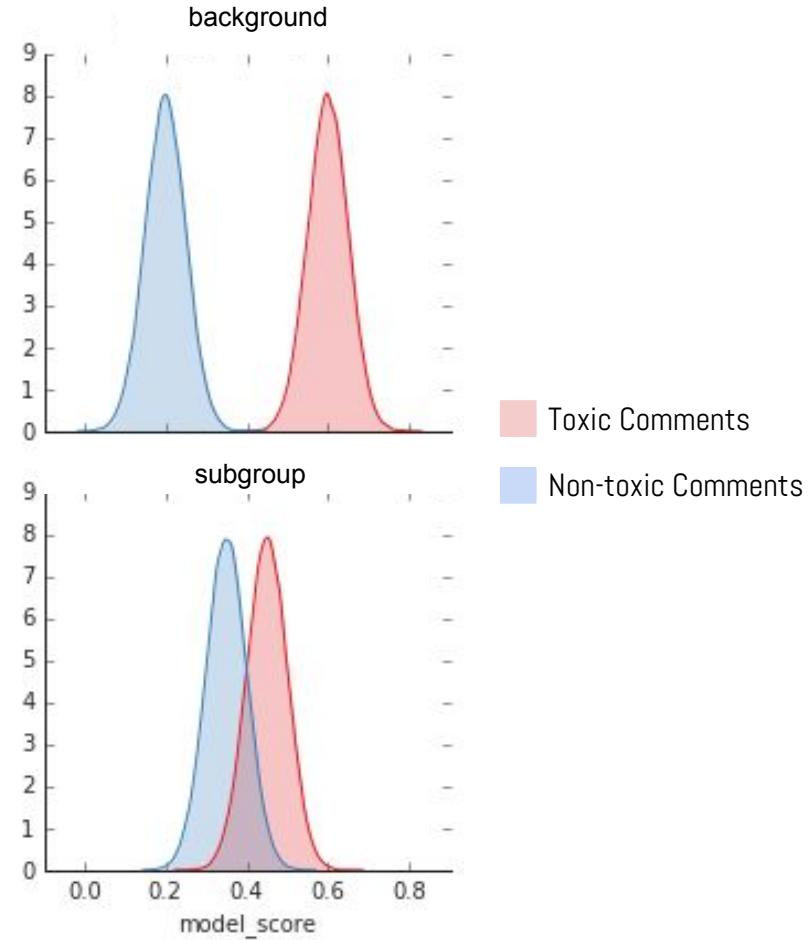


# Types of Bias

## *Low Subgroup Performance*

The model performs worse on subgroup comments than it does on comments overall.

**Metric:** Subgroup AUC



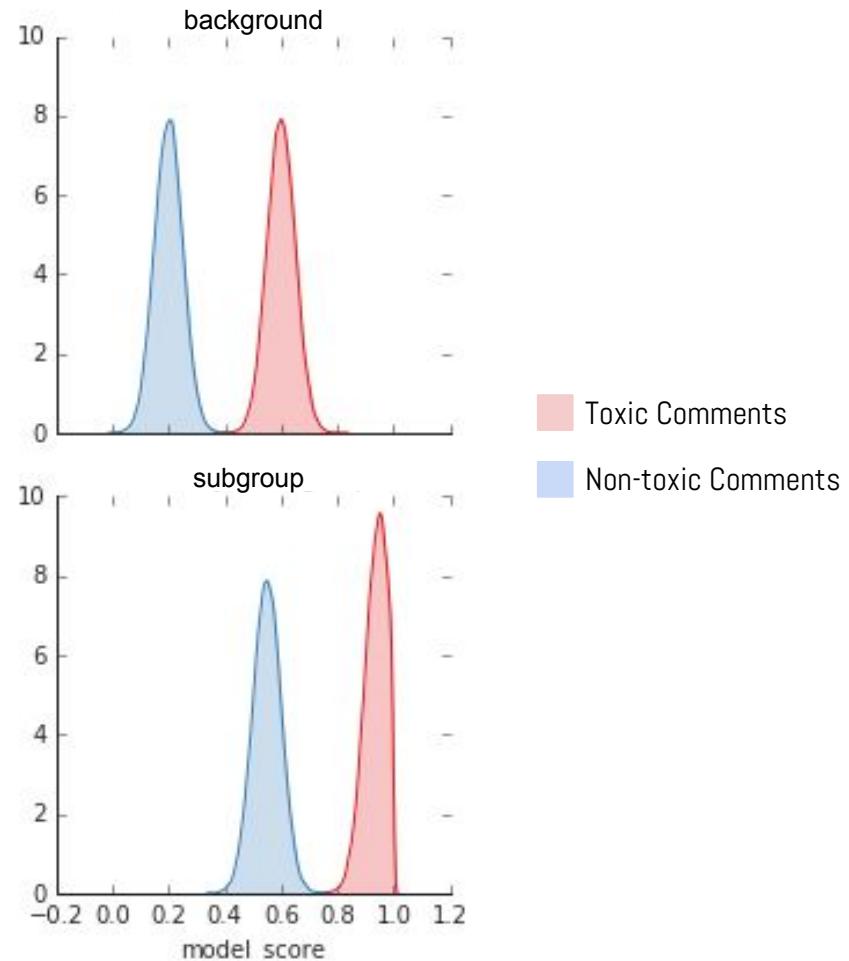
# Types of Bias

*Subgroup Shift (Right)*

The model systematically scores comments from the subgroup higher.

**Metric:** BPSN AUC

(Background Positive Subgroup Negative)



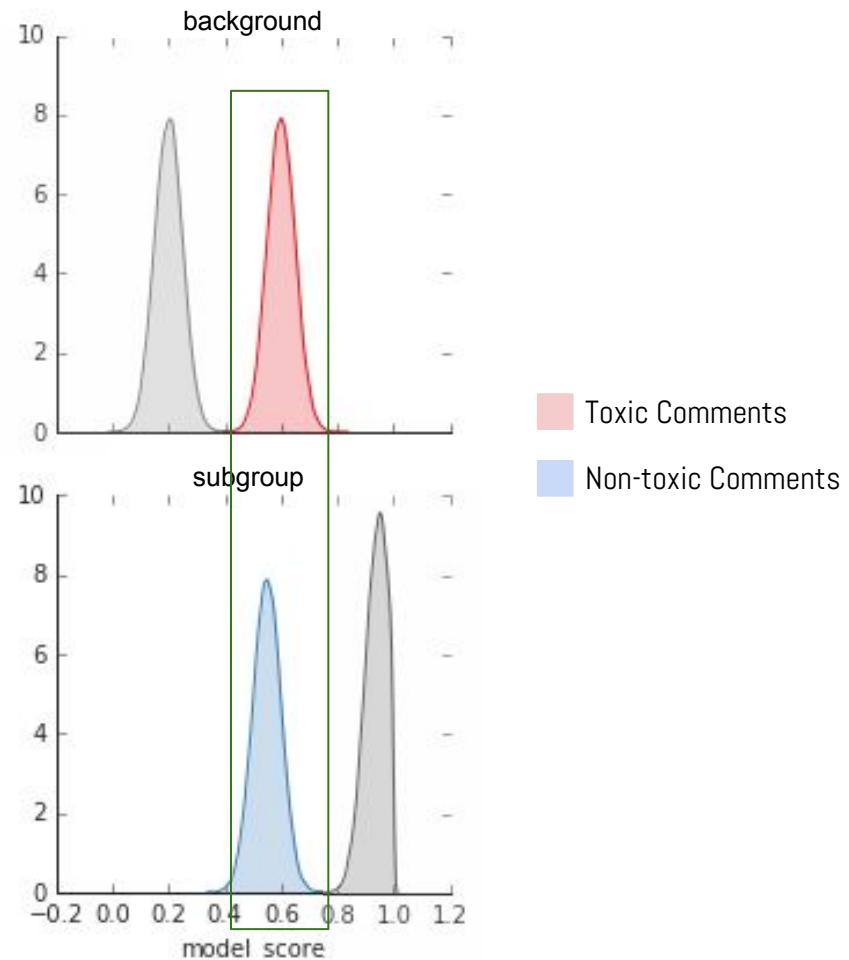
# Types of Bias

*Subgroup Shift (Right)*

The model systematically scores comments from the subgroup higher.

**Metric:** BPSN AUC

(Background Positive Subgroup Negative)



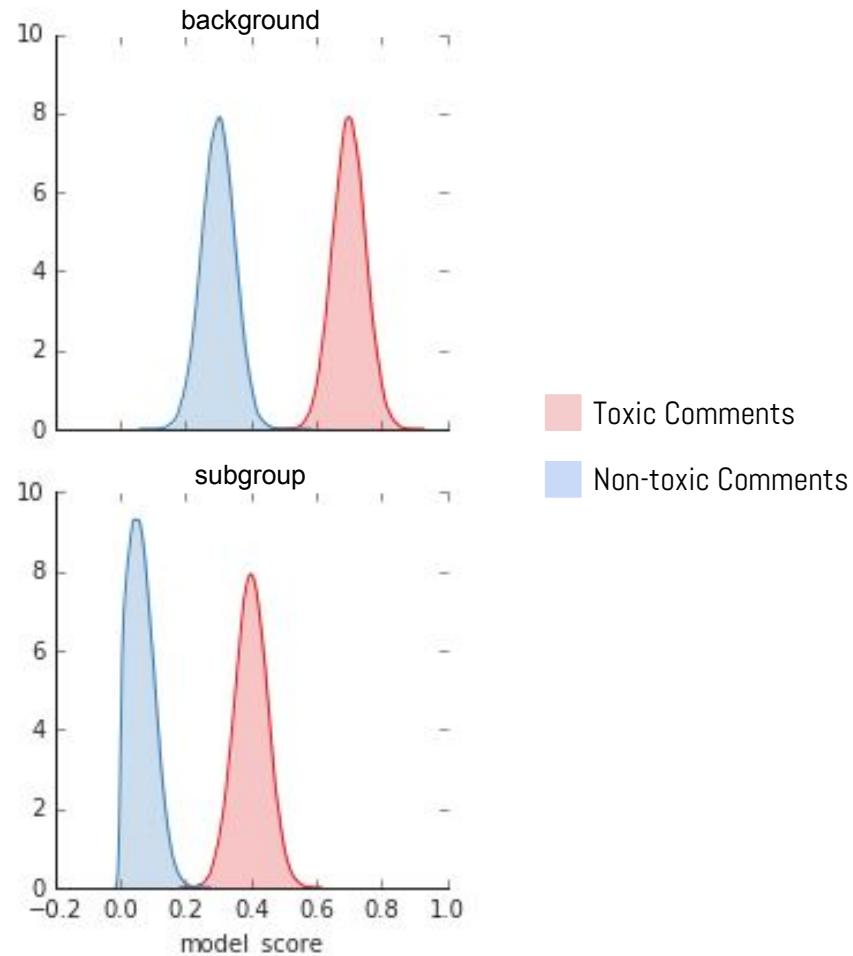
# Types of Bias

## *Subgroup Shift (Left)*

The model systematically scores comments from the subgroup lower.

**Metric:** BNSP AUC

(Background Negative Subgroup Positive)



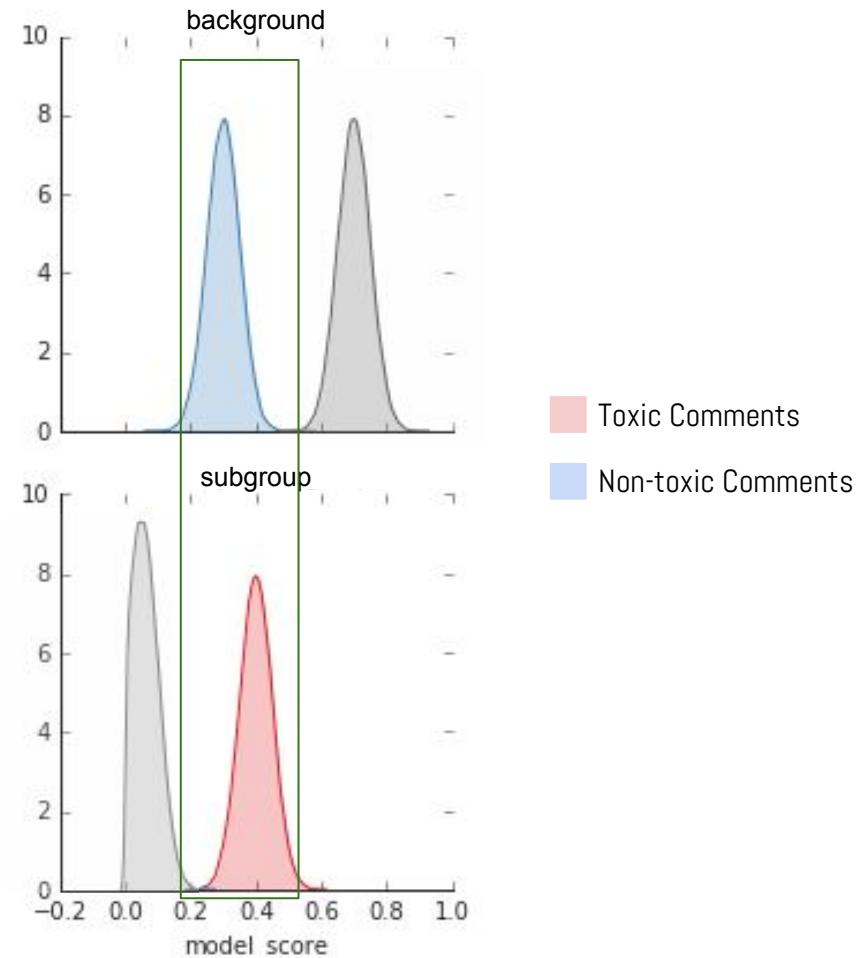
# Types of Bias

## *Subgroup Shift (Left)*

The model systematically scores comments from the subgroup lower.

**Metric:** BNSP AUC

(Background Negative Subgroup Positive)



# Results

Toxicity @1

Identity groups	Subgroup AUC	BPSN AUC	BPSP 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



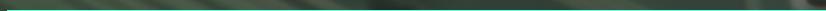
Toxicity @6

Identity groups	Subgroup AUC	BPSN AUC	BPSP AUC
lesbian	1.00	0.98	1.00
gay	1.00	0.94	1.00
queer	0.99	0.98	0.99
straight	1.00	1.00	0.97
bisexual	0.98	0.98	0.99
homosexual	1.00	0.96	1.00
heterosexual	1.00	0.99	1.00
cis	1.00	1.00	0.98
trans	1.00	1.00	1.00
nonbinary	1.00	1.00	0.98
black	0.98	0.97	1.00
white	0.99	0.99	0.99





Release Responsibly



# Model Cards for Model Reporting

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



## 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.

## 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

# Intended Use, Factors and Subgroups

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

# Metrics and Data

<b>Metrics</b>	<p><i>Pinned AUC</i>, which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.</p>
<b>Evaluation Data</b>	A synthetic test set generated using a template-based approach, where identity terms are swapped into a variety of template sentences.
<b>Training Data</b>	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”.

# Considerations, Recommendations

<b>Ethical Considerations</b>	A set of values around community, transparency, inclusivity, privacy and topic-neutrality to guide their work.
<b>Caveats &amp; Recommendations</b>	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.

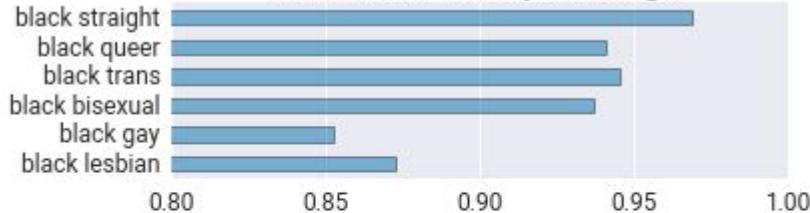
# Disaggregated Intersectional Evaluation

Toxicity @1

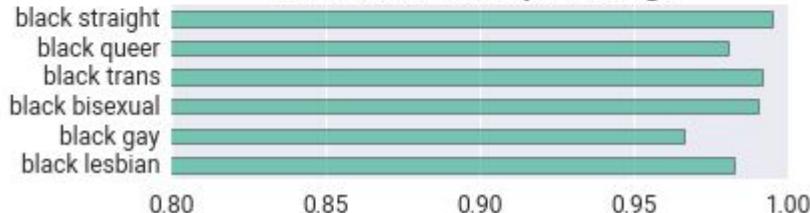
Identity groups	Subgroup AUC	BPSN AUC	BNSP AUC
lesbian	0.93	0.74	0.98
gay	0.94	0.65	0.99
queer	0.98	0.96	0.93
straight	0.99	1.00	0.87
bisexual	0.96	0.95	0.92
homosexual	0.87	0.53	0.99
heterosexual	0.96	0.94	0.92
cis	0.99	1.00	0.87
trans	0.97	0.96	0.91
nonbinary	0.99	0.99	0.90
black	0.91	0.85	0.95
white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



Pinned AUC Toxicity Scores @1



Pinned AUC Toxicity Scores @5



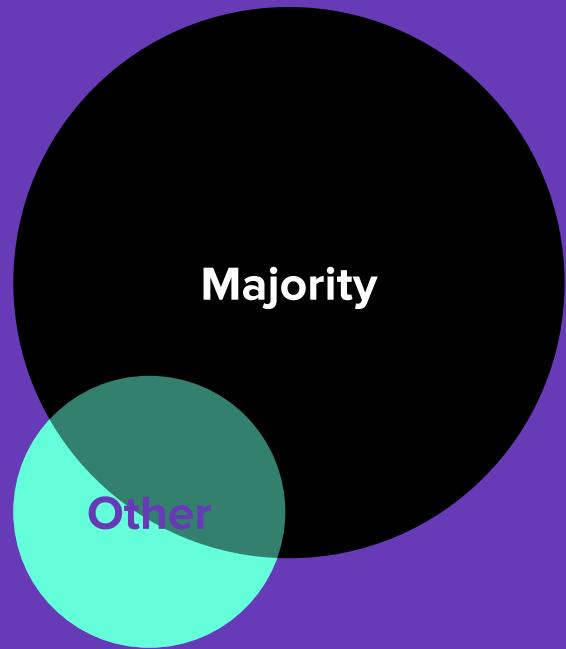
Jigsaw



The False Positive

---

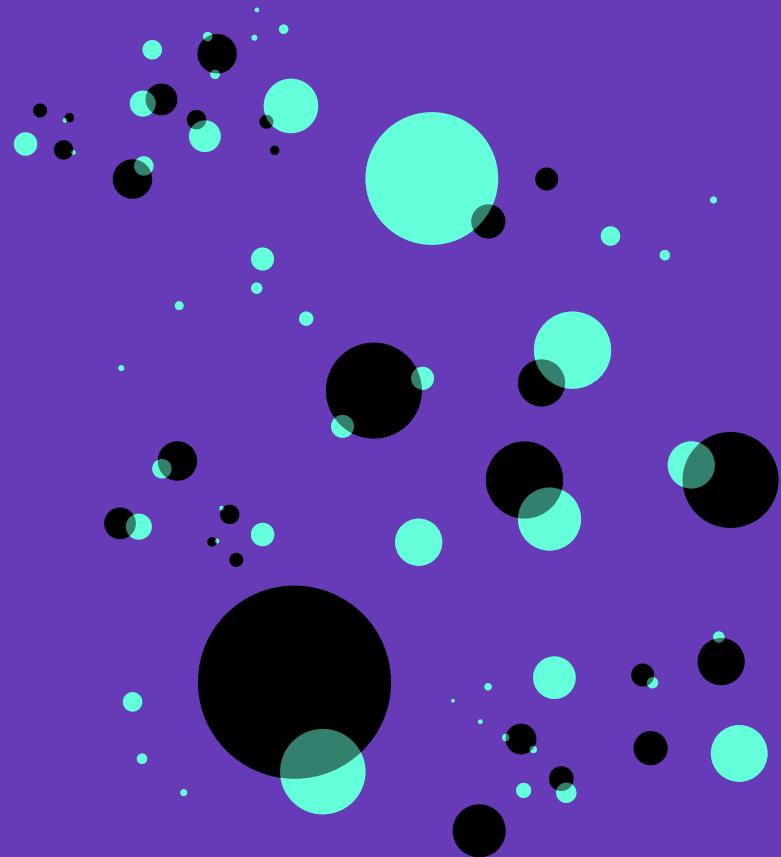
Moving from majority representation...



---

Moving from majority  
representation...

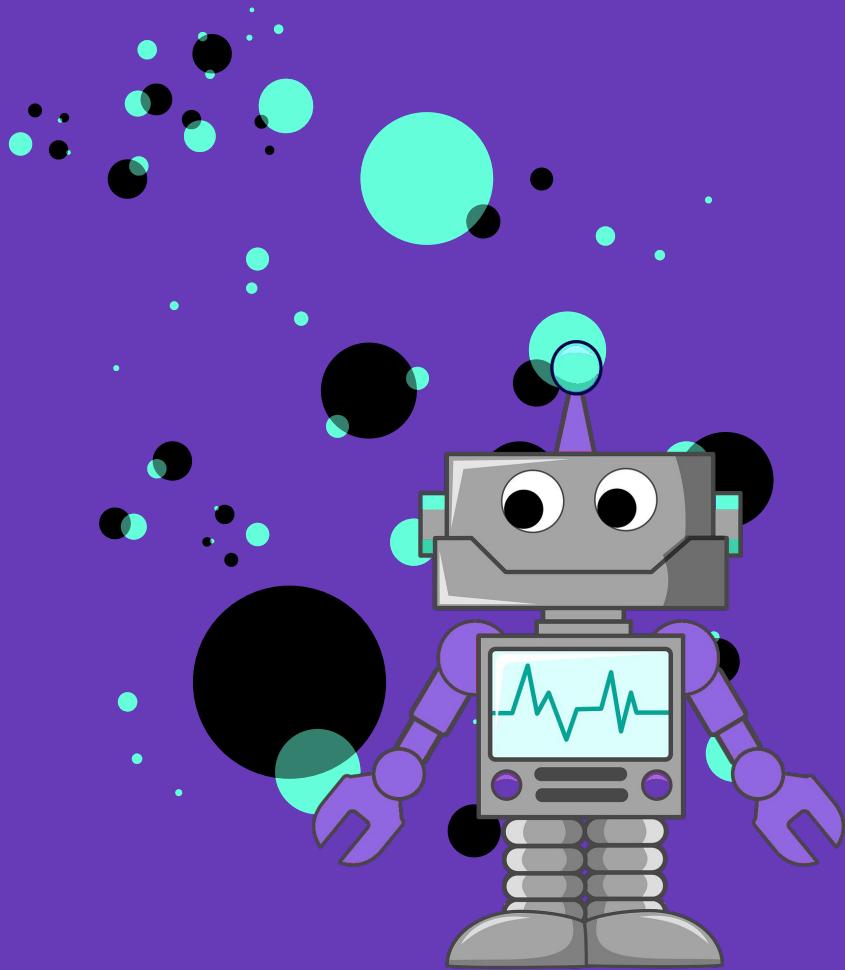
...to diverse  
representation



---

Moving from majority  
representation...

...to diverse  
representation  
...for ethical AI



# Thanks!

[margarmitchell@gmail.com](mailto:margarmitchell@gmail.com)

[m-mitchell.com](http://m-mitchell.com)

Need MOAR? [ml-fairness.com](http://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  
Aguera 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

The screenshot shows a Google Colab interface with the following details:

- Title:** Conversation AI's Pinned AUC Unintended Model Bias
- Authors:** ldixon@google.com, jetpack@google.com, sorenj@google.com, nthain@google.com, lucyvasserman@google.com
- Description:** 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. It references the paper [Measuring and Mitigating Unintended Bias in Text Classification](#) for background, detailed explanation, and experimental results. It also links to [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!
- Code Snippets:** Includes sections for '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', and a 'SECTION' block.
- Bottom:** A terminal window showing the command `!pip install -U -q ait+https://github.com/conversationai/unintended-ml-bias-analysis`.

## 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

The screenshot shows a Google Colab interface with the following details:

- Title:** Debiasing Word Embeddings using Fair Adversarial Networks (FANs)
- Authors:** lemoine@, zhangbrian@, benhutch@, guajardo@
- Contributors:** mmitchellai@, andrewzaldivar@
- Summary:** This Colab was put together as part of the ML-fairness inspired 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
- Code Snippets:** Includes sections for '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', and 'Analogy using trained model: A is to B as C is to ??'.

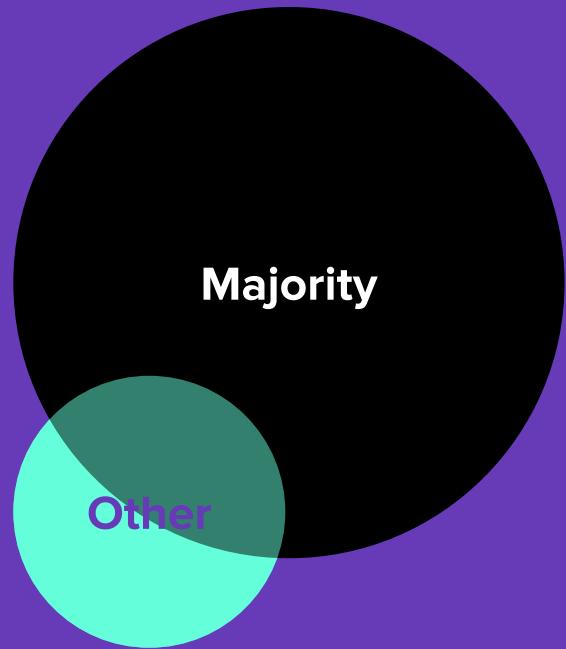
# Get Involved

- Find free machine-learning tools open to anyone at [ai.google/tools](https://ai.google/tools)
- Check out Google's ML Fairness codelab at [ml-fairness.com](https://ml-fairness.com)
- Explore educational resources at [ai.google/education](https://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](mailto:acceleratewithgoogle@google.com)



---

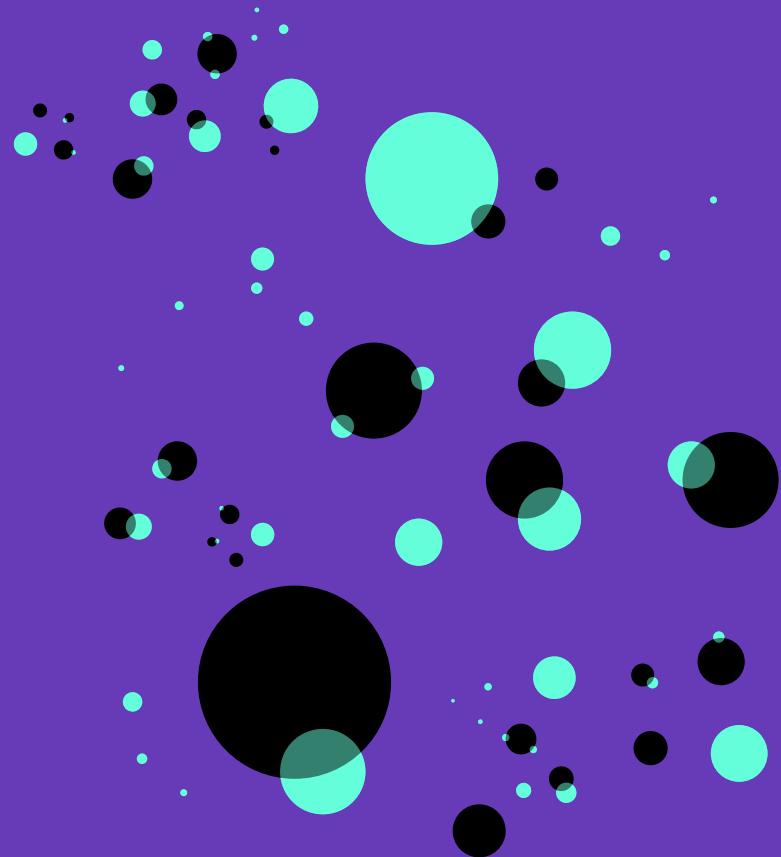
Moving from majority representation...



---

Moving from majority  
representation...

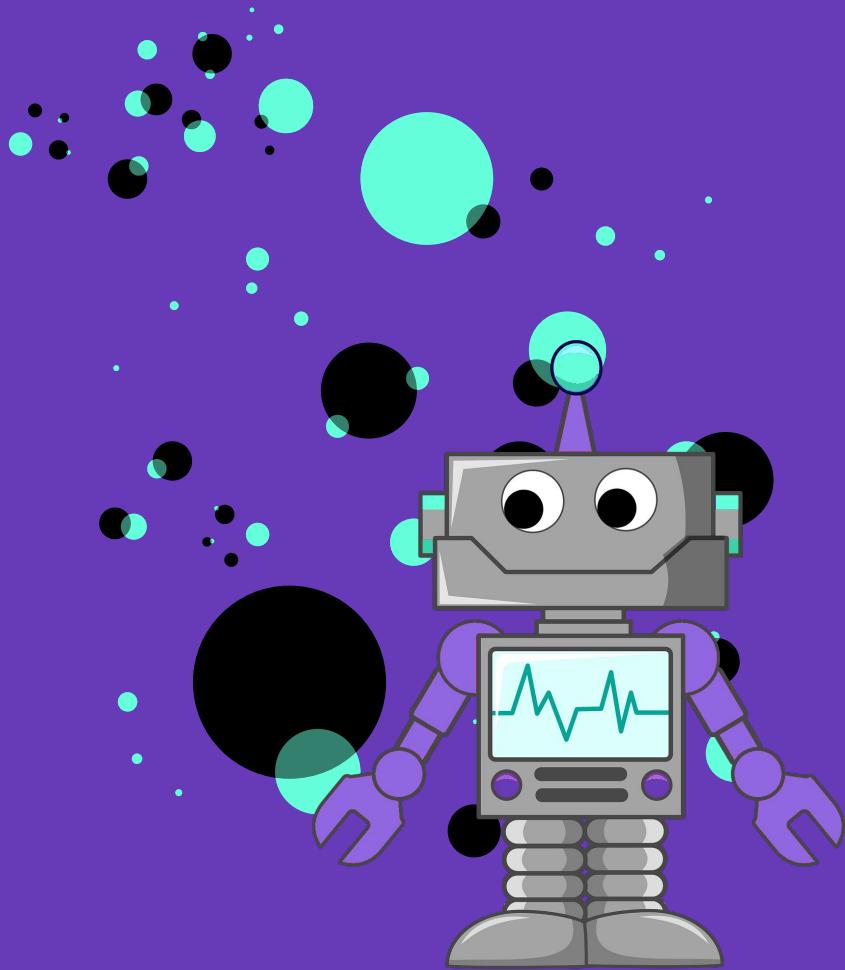
...to diverse  
representation



---

Moving from majority  
representation...

...to diverse  
representation  
...for ethical AI



# Thanks!

[margarmitchell@gmail.com](mailto:margarmitchell@gmail.com)

[m-mitchell.com](http://m-mitchell.com)

Need MOAR? [ml-fairness.com](http://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  
Aguera y  
Arcas