



Data Bias and Fair ML

Applied Machine Learning
Derek Hoiem

Today's lecture

- Algorithm and Data bias
- Fair ML

Adopted from a 2019 guest lecture by Margaret Mitchell



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**



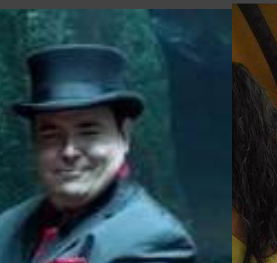
**Dirk
Hovy**



**Josh
Lovejoy**



**Alex
Beutel**



**Blake
Lemoine**



**Hee Jung
Ryu**



**Hartwig
Adam**



**Blaise
Agüera 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!"

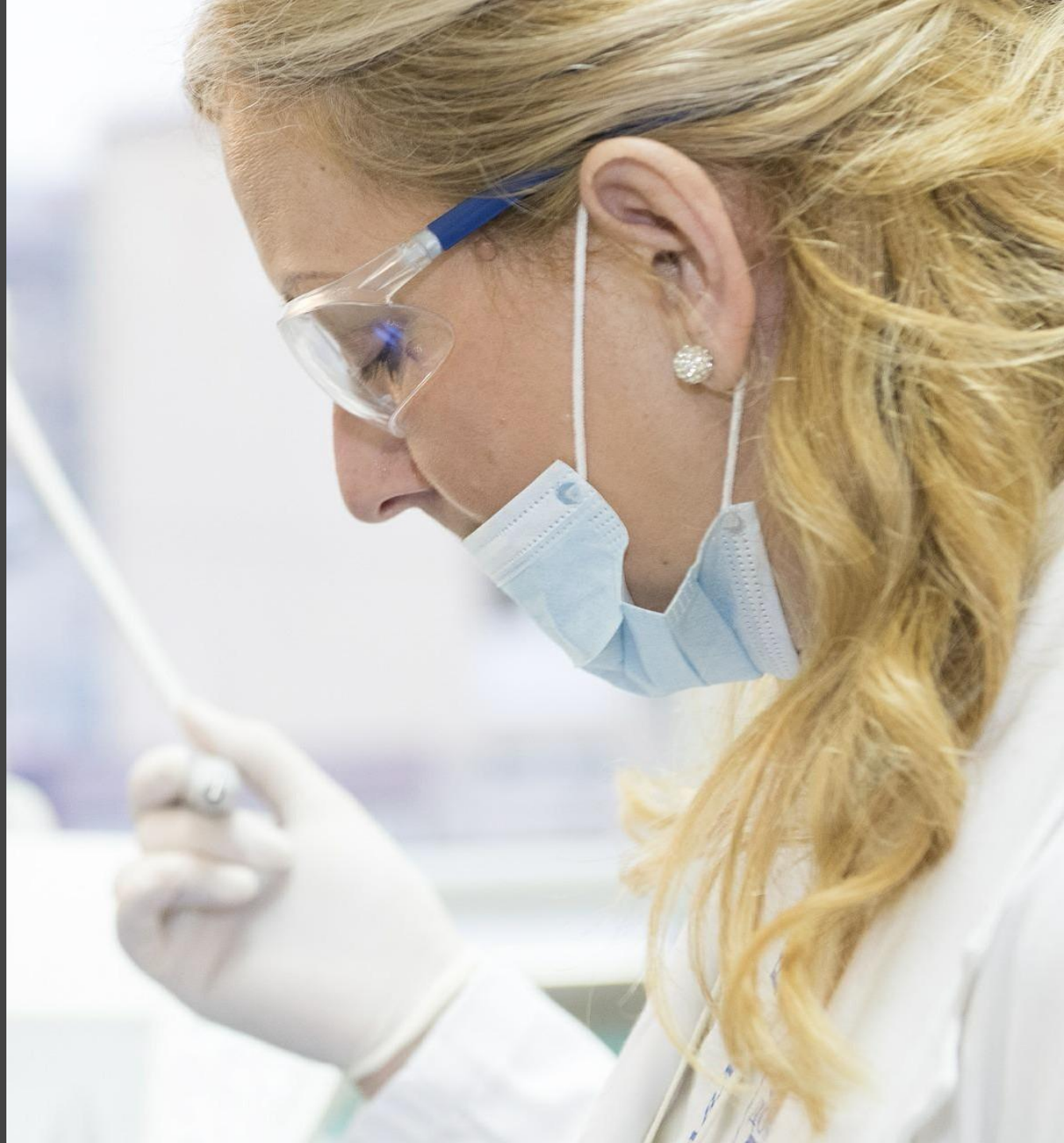
How could this be?



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!"

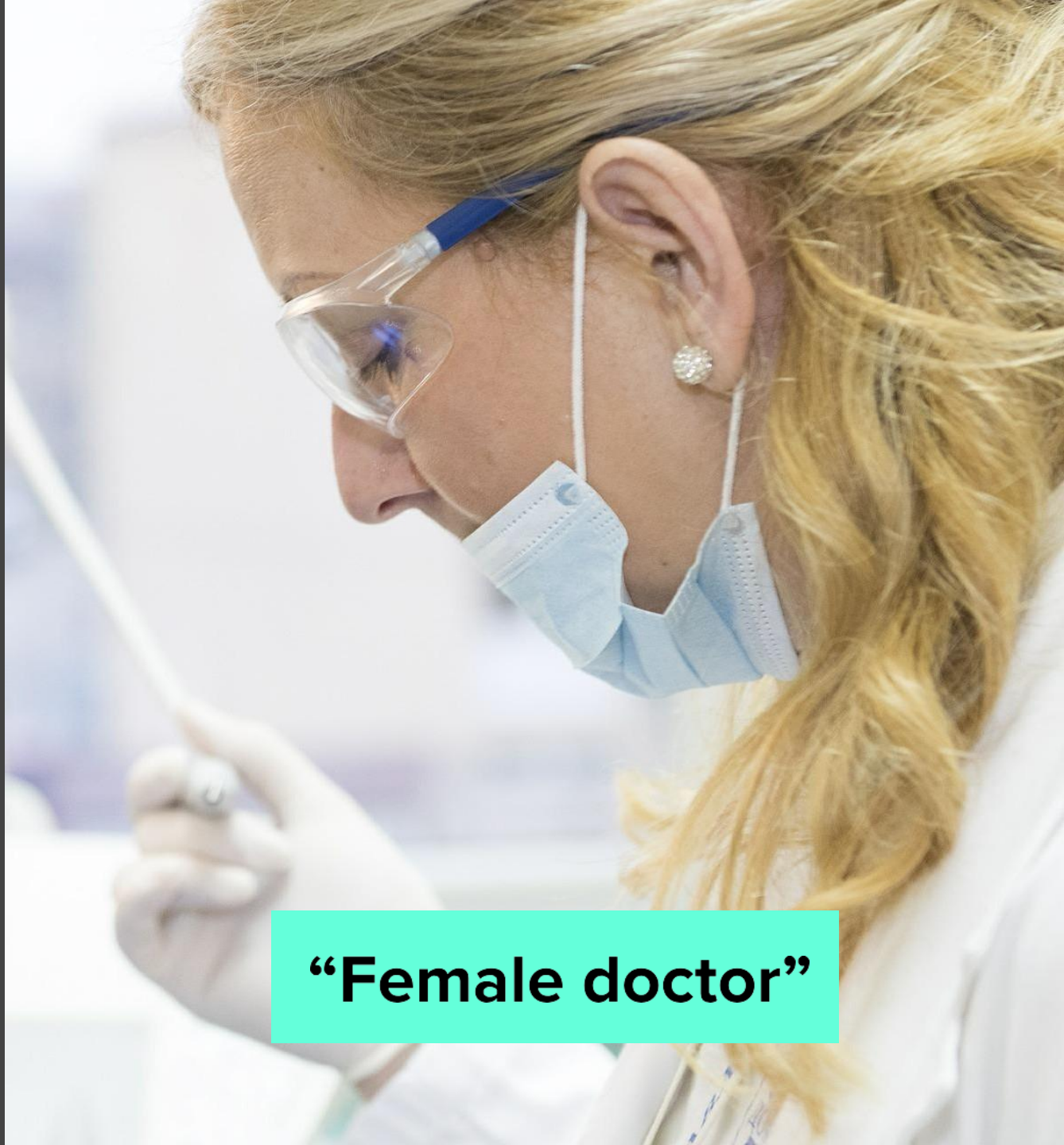
How could this be?



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**

```
graph LR; A((Training data are collected and annotated)) --> B((Model is trained));
```

**Training data are
collected and
annotated**

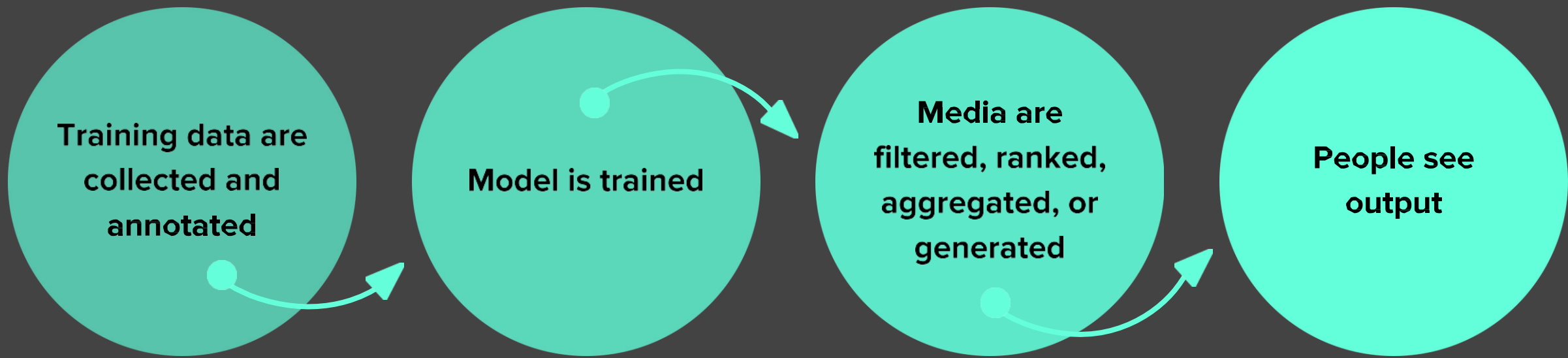
Model is trained



Training data are collected and annotated

Model is trained

Media are filtered, ranked, aggregated, or generated



Human Biases in Data

Reporting bias

Selection bias

Overgeneralization

Out-group homogeneity bias

Stereotypical bias

Historical unfairness

Implicit associations

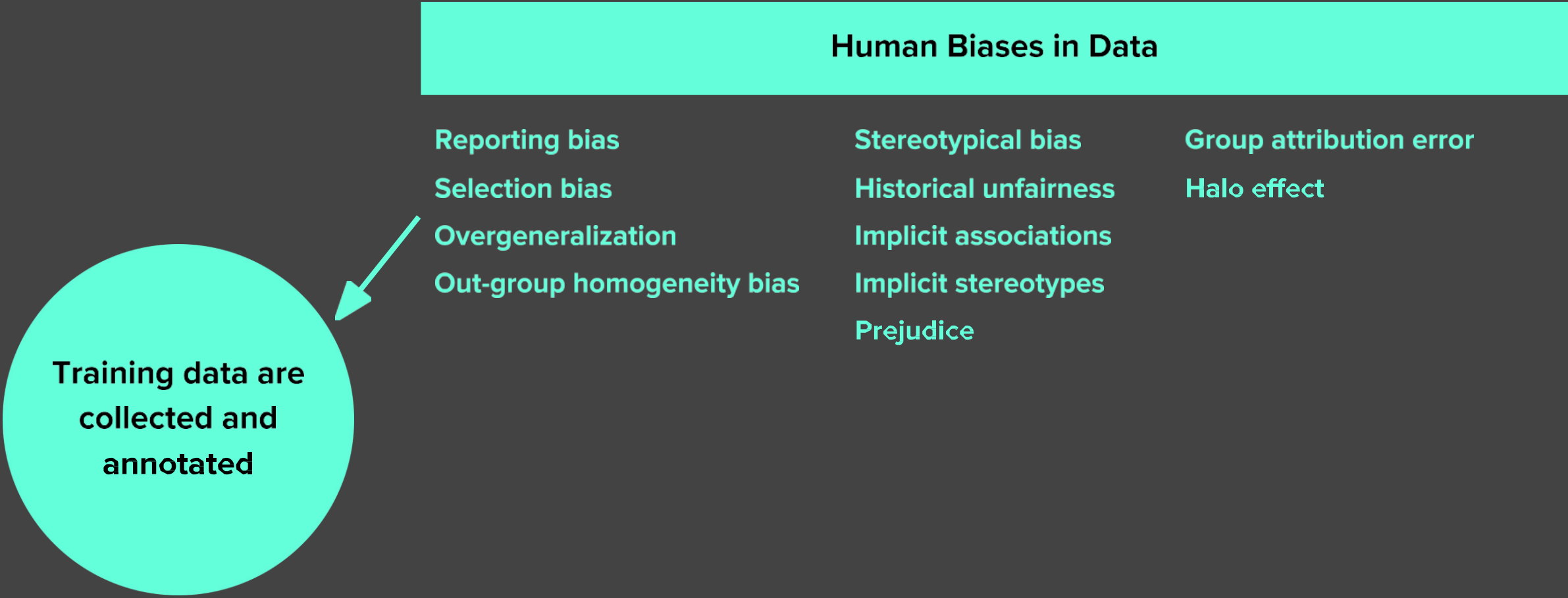
Implicit stereotypes

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

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



Biases in Data

Biases in Data

Selection Bias: Selection does not reflect a random sample



CREDIT

© 2013–2016 Michael Yoshitaka Erlewine and Hadas Kotek

CC BY / Margaret Mitchell / m-mitchell.com

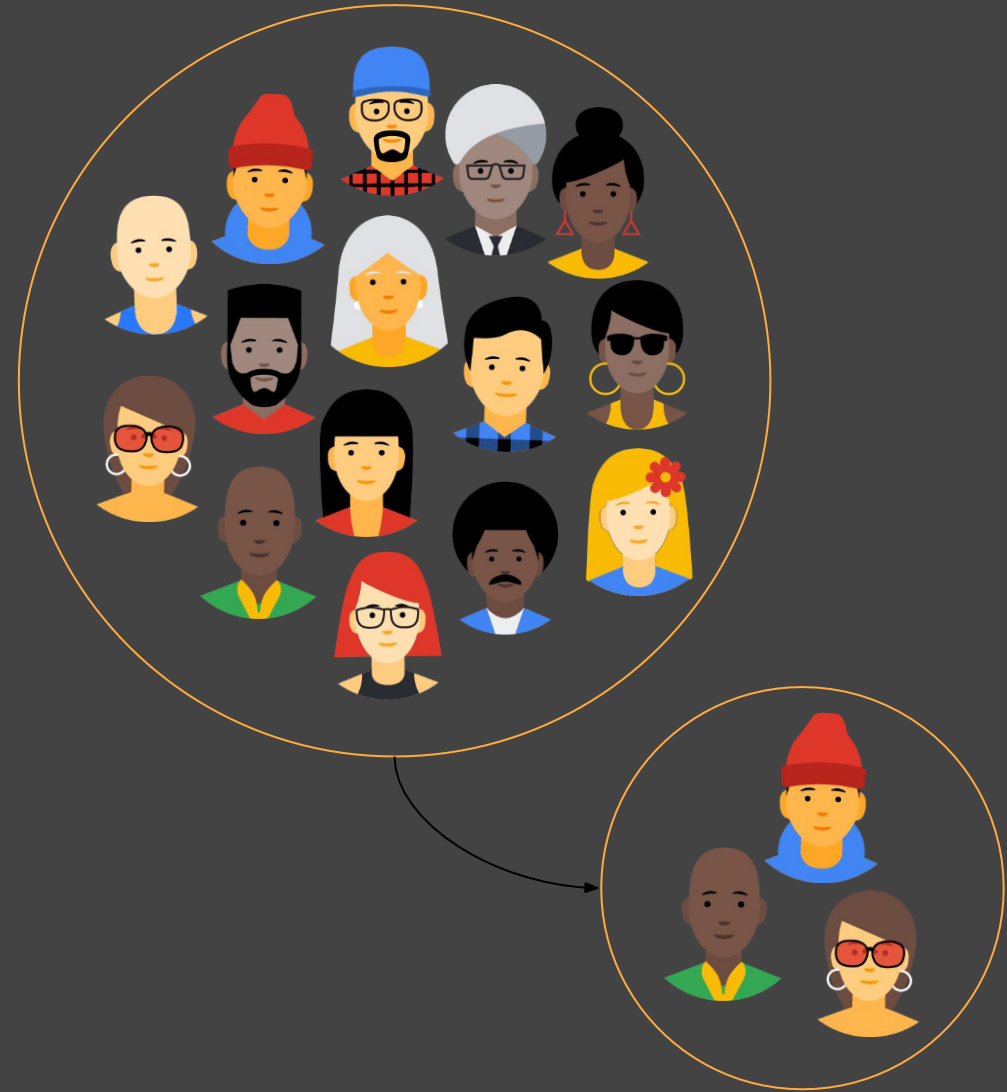
Biases in Data

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.

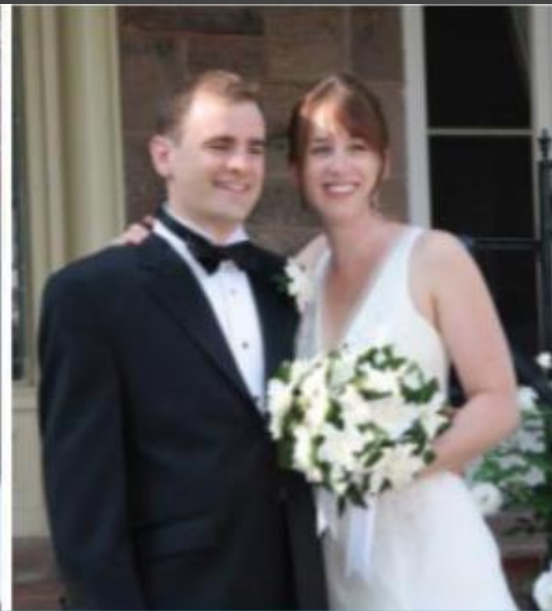


Biases in Data → Biased Labels

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



*ceremony,
wedding, bride,
man, groom,
woman, dress*



*ceremony,
bride, wedding,
man, groom,
woman, dress*



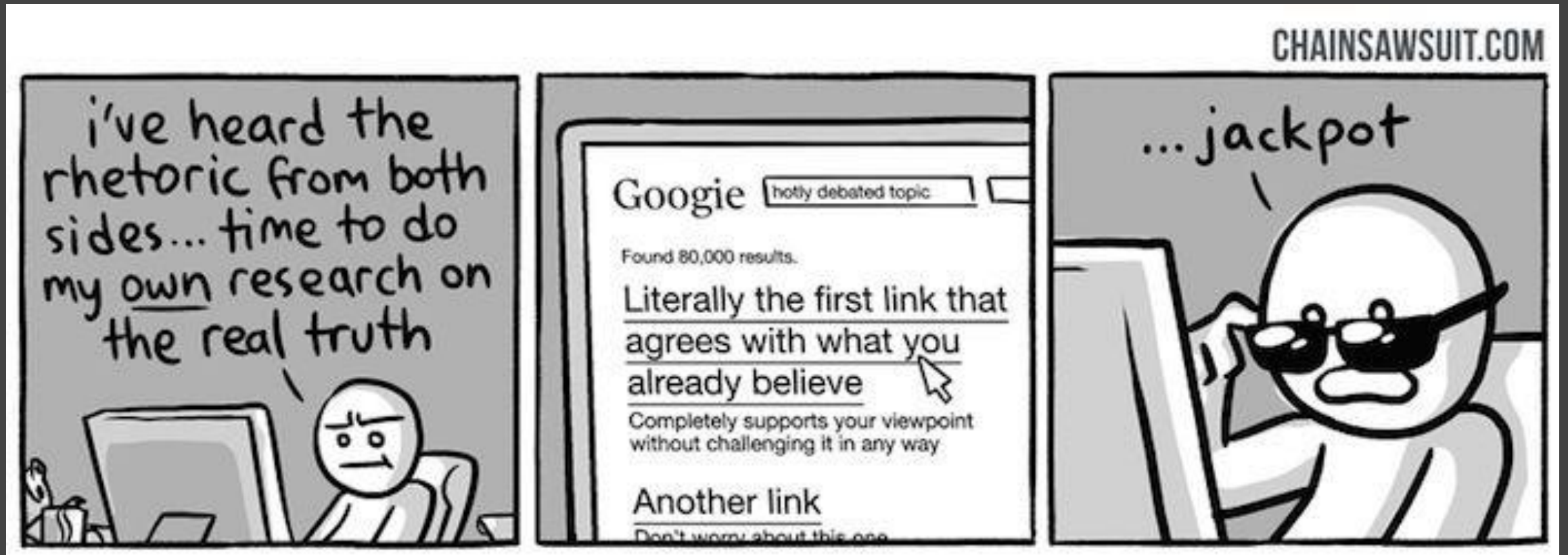
person, people



Biases in Interpretation

Biases in Interpretation

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



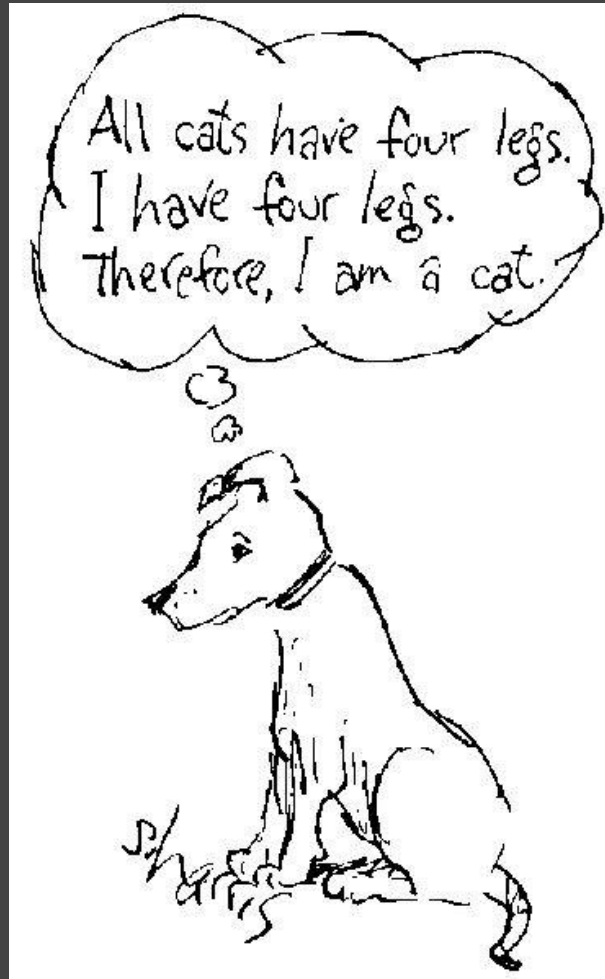
CREDIT

© kris straub - Chainsawsuit.com

CC BY / Margaret Mitchell / m-mitchell.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.



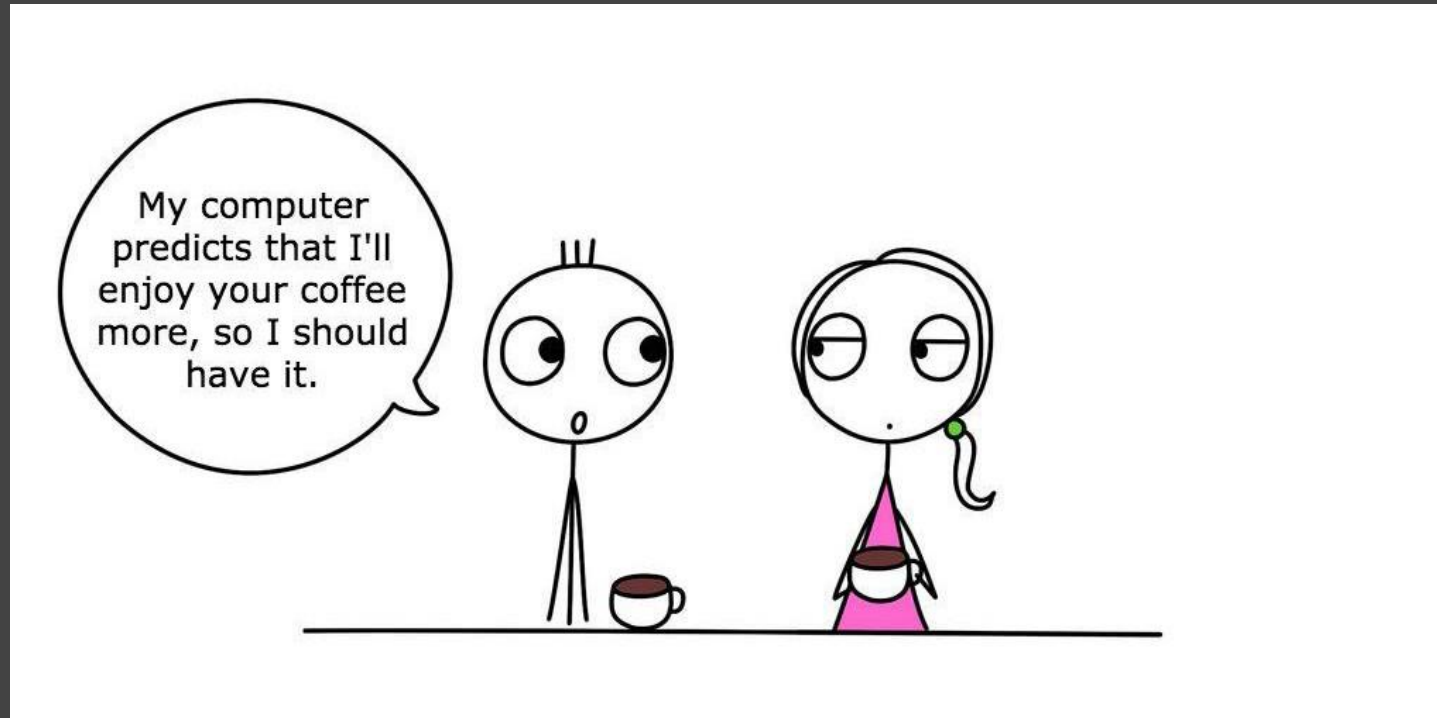
CREDIT

© mollysdad - Slideshare - Introduction to Logical Fallacies

CC BY / Margaret Mitchell / m-mitchell.com

Biases in Interpretation

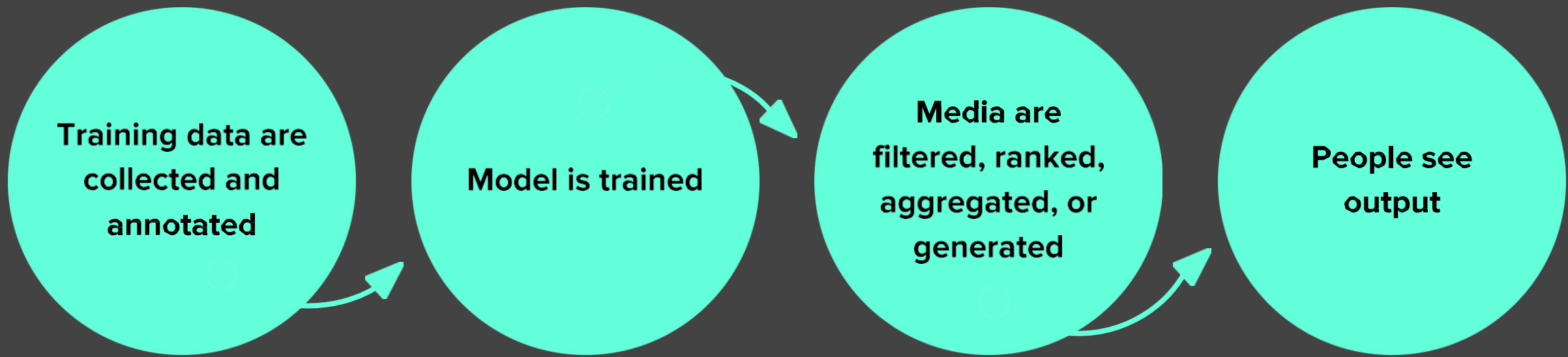
Automation bias: Propensity for humans to favor suggestions from automated decision-making systems over contradictory information without automation



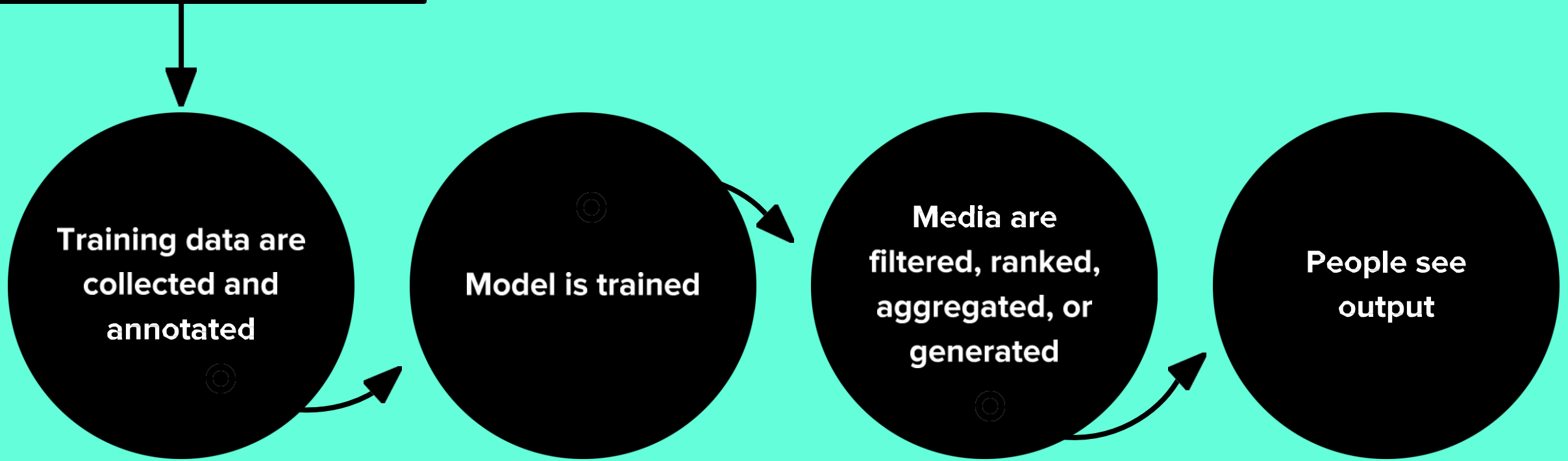
CREDIT

thedailyenglishshow.com | [CC BY 2.0](https://creativecommons.org/licenses/by/2.0/)

CC BY / Margaret Mitchell / m-mitchell.com



Human Bias



Human Bias



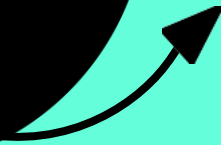
Training data are collected and annotated



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



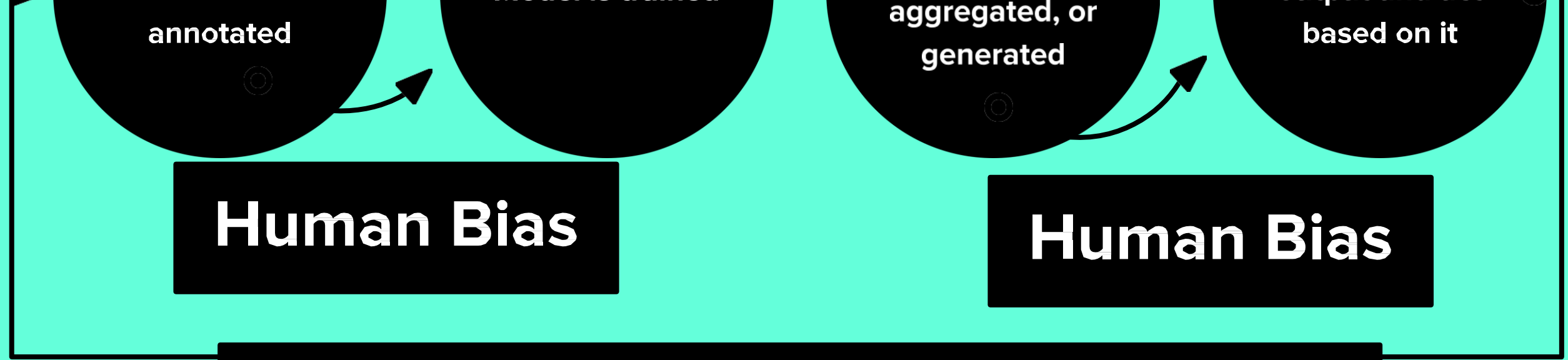
People see output and act based on it



Human Bias

Human Bias

Feedback Loop



Human 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](#)



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.

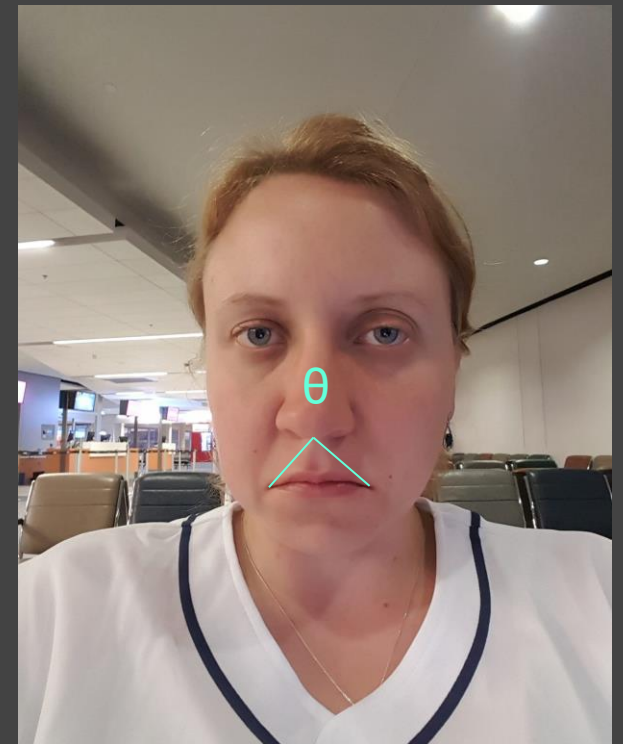
Main clients are in homeland security and public safety.

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 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”

**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... ▼

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

References

Evaluate for Fairness & Inclusion: Confusion Matrix



Evaluate for Fairness & Inclusion: Confusion Matrix

		Model Predictions	
		Positive	Negative

References	Positive
	Negative
	<ul style="list-style-type: none">● Exists● Predicted <p>True Positives</p>

<ul style="list-style-type: none">● Doesn't exist● Not predicted <p>True Negatives</p>
--

Evaluate for Fairness & Inclusion: Confusion Matrix

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

Evaluate for Fairness & Inclusion: Confusion Matrix

		Model Predictions		
		Positive	Negative	
References	Positive	<ul style="list-style-type: none"> Exists Predicted <p>True Positives</p>	<ul style="list-style-type: none"> Exists Not predicted <p>False Negatives</p>	<p>Recall, False Negative Rate</p>
	Negative	<ul style="list-style-type: none"> Doesn't exist Predicted <p>False Positives</p>	<ul style="list-style-type: none"> Doesn't exist Not predicted <p>True Negatives</p>	
		<p>Precision, False Discovery Rate</p>	<p>Negative Predictive Value, False Omission Rate</p>	<p>LR+, LR-</p>

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

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

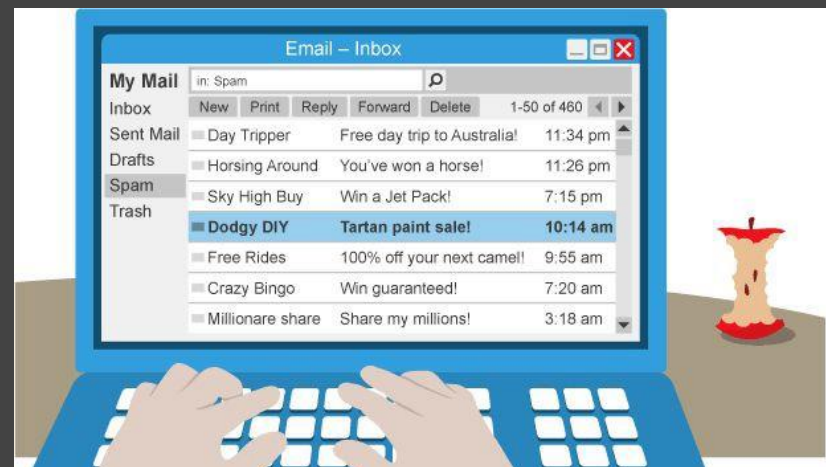
Spam Filtering

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

Usually just a bit annoying.

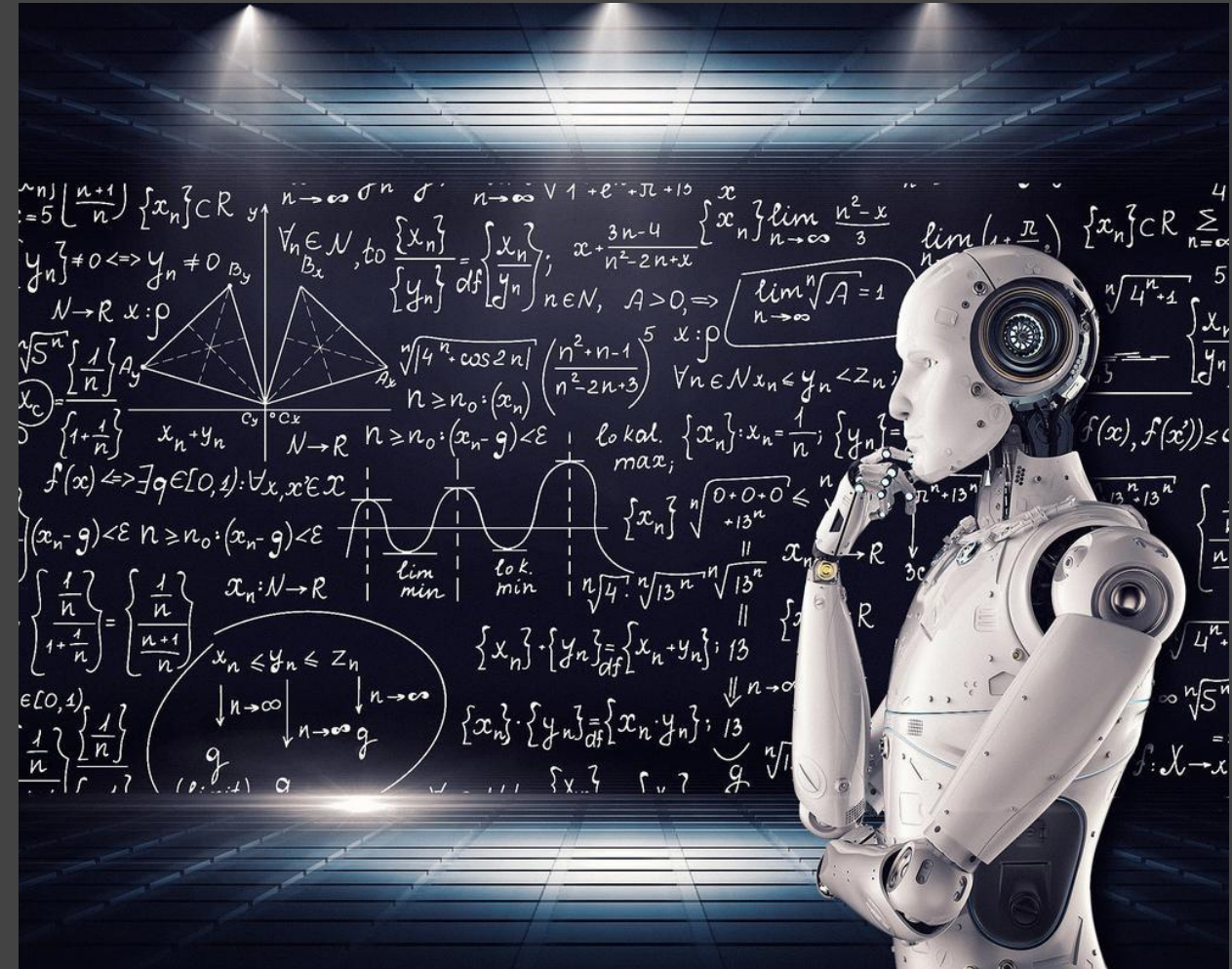
False Positive: Email flagged as SPAM is removed from your inbox.

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



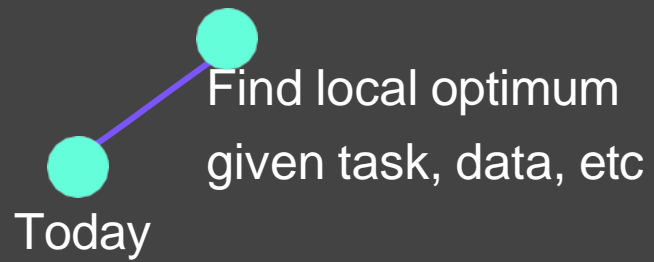
It's up to **us** to influence how AI evolves.



Today

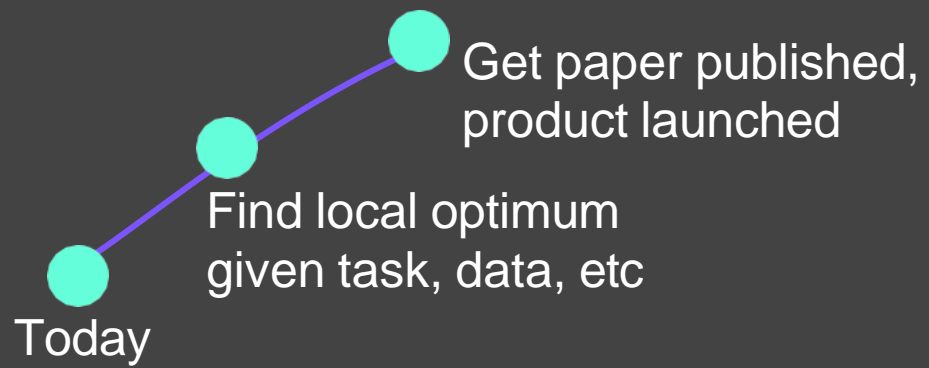
Short-term

Longer-term



Short-term

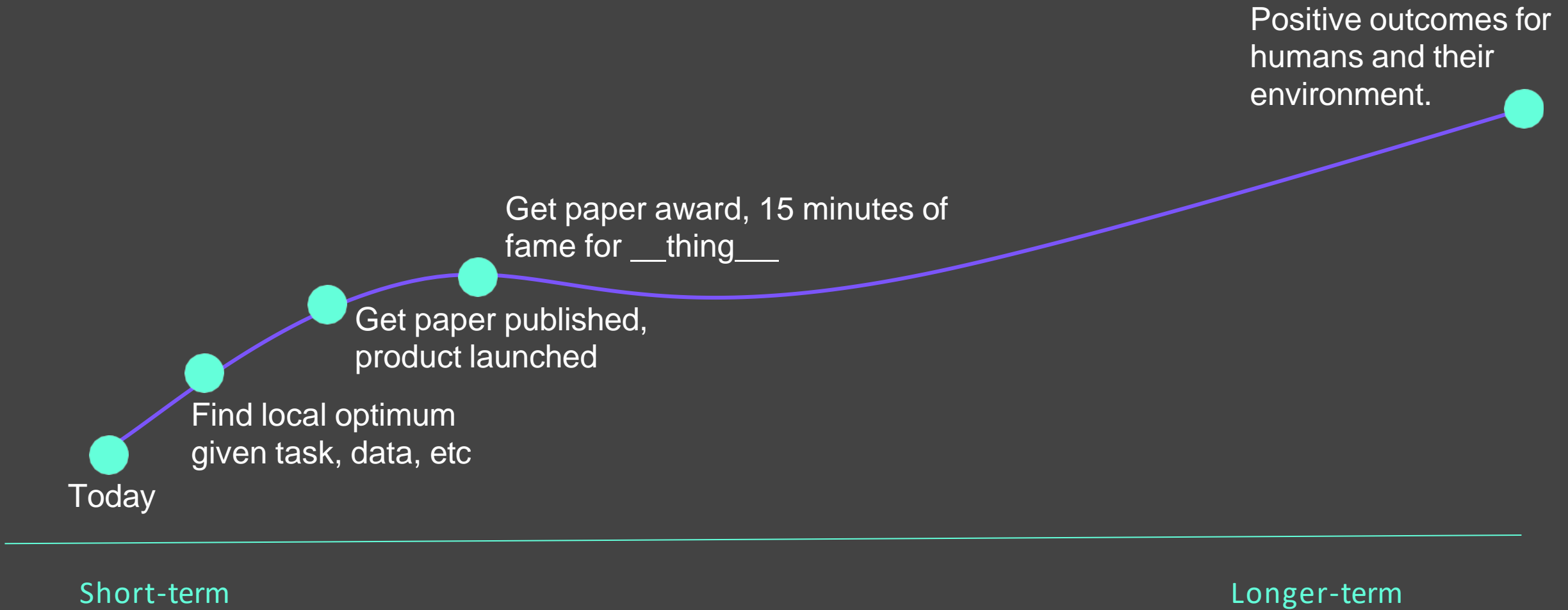
Longer-term



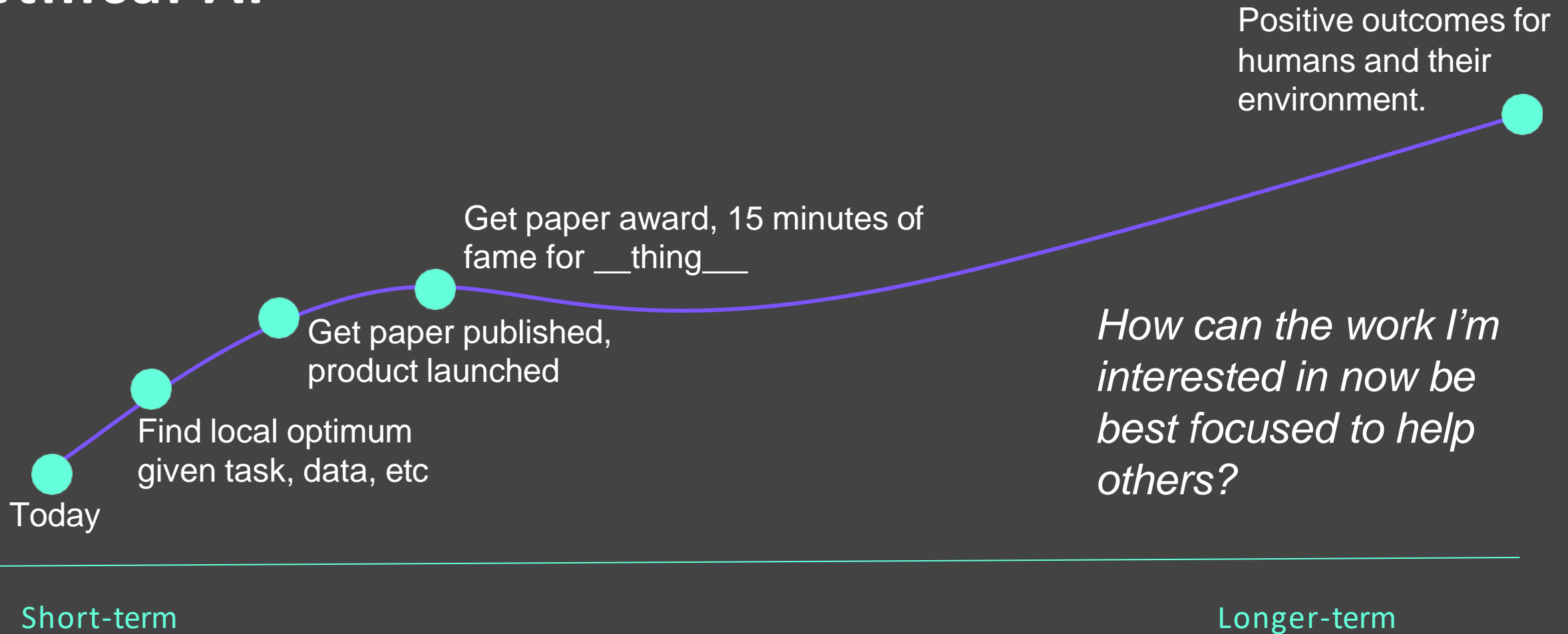


Short-term

Longer-term



Begin tracing out paths for the evolution of ethical AI



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

Here are some things we can do.



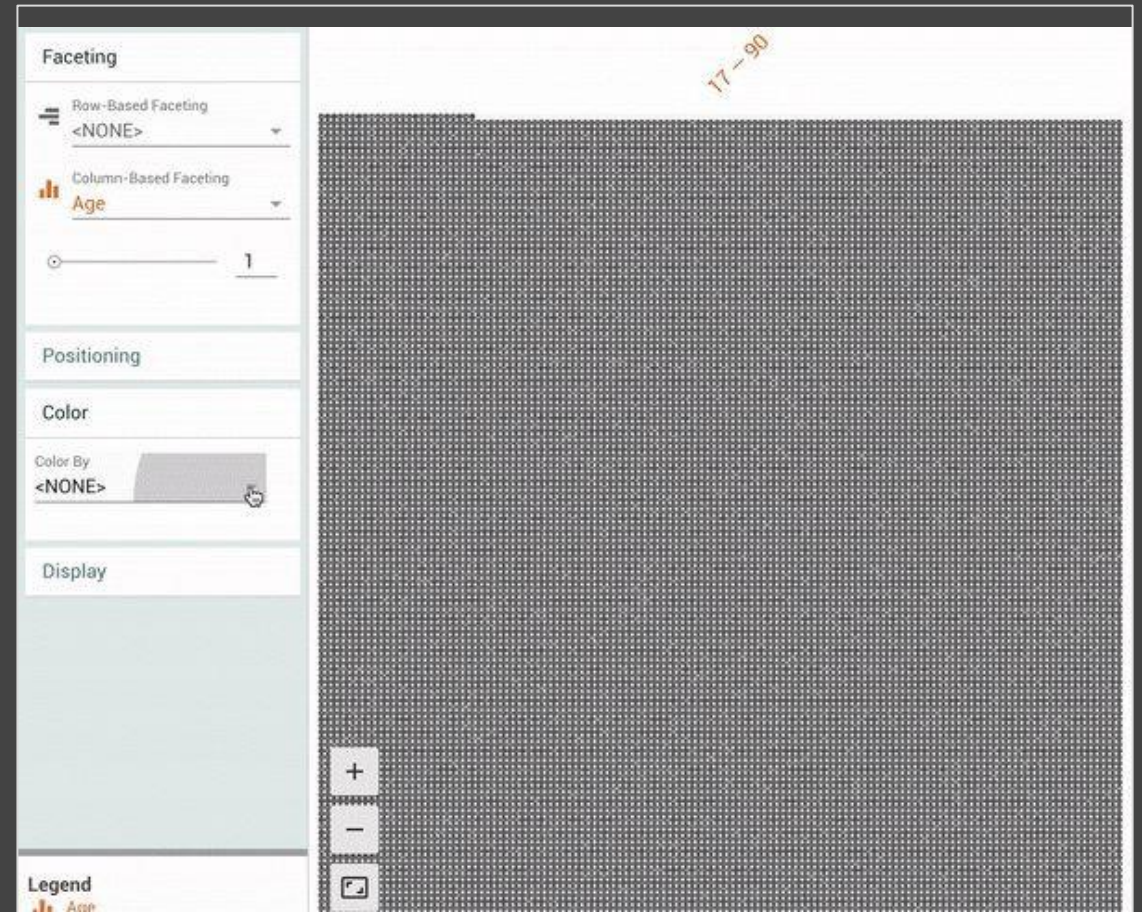
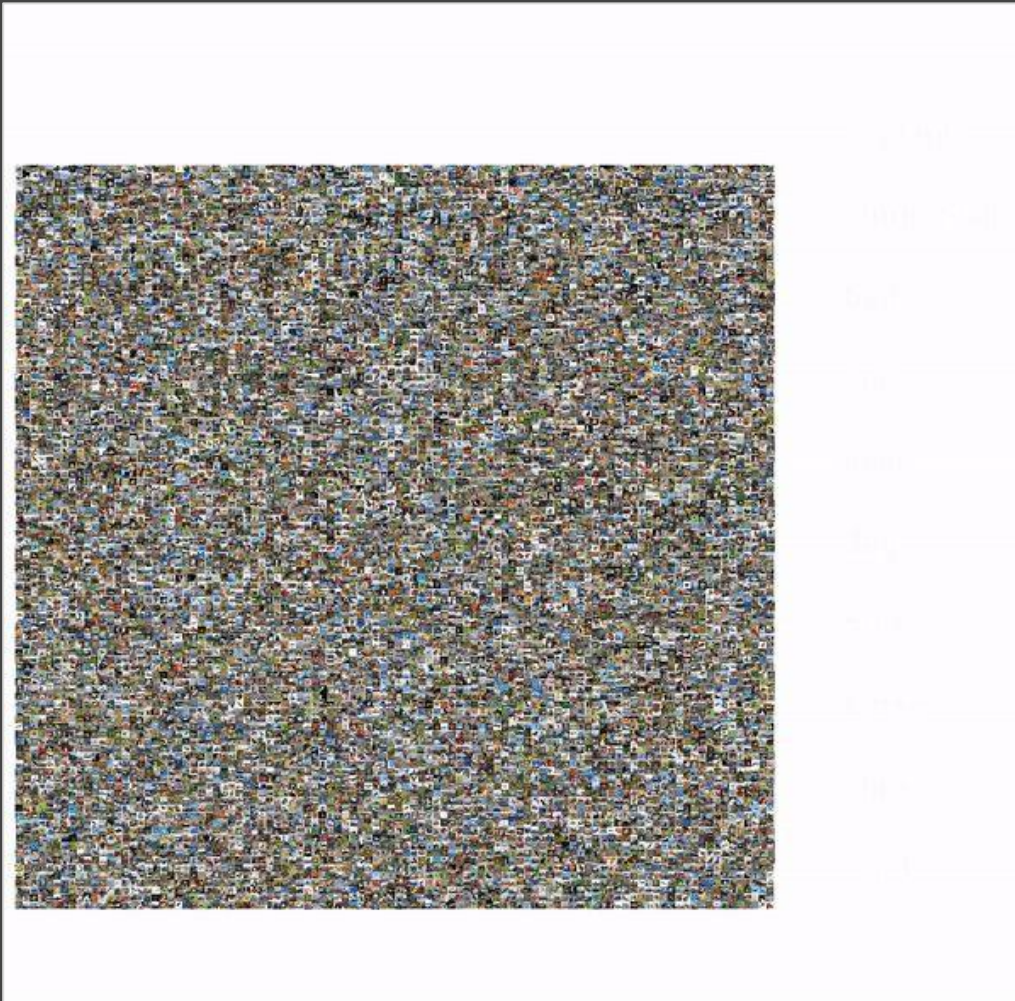
Data

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 Gebru¹ Jamie Morgenstern² Briana Vecchione³ Jennifer Wortman Vaughan¹ Hanna Wallach¹
Hal Daumé III^{1,4} Kate Crawford^{1,5}

Datasheets for Datasets

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



Title COMPAS Recidivism Risk Score Data

Author Broward County Clerk's Office, Broward County Sheriff'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

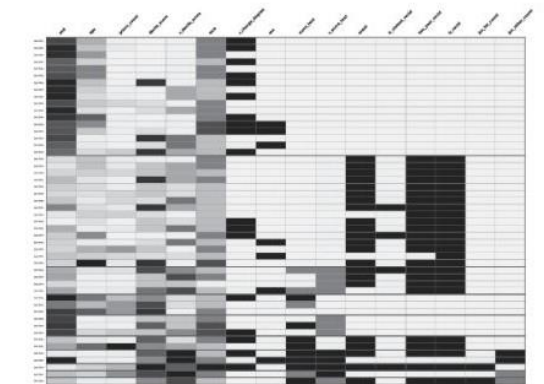
priors_count: *Ut enim ad minim veniam, quis nostrud exercitation* numerical

two_year_recid: Lorem ipsum dolor sit amet conseq

Probabilistic Modeling

Analysis

◀ 12 ▶



Dependency Probability Pearson R





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

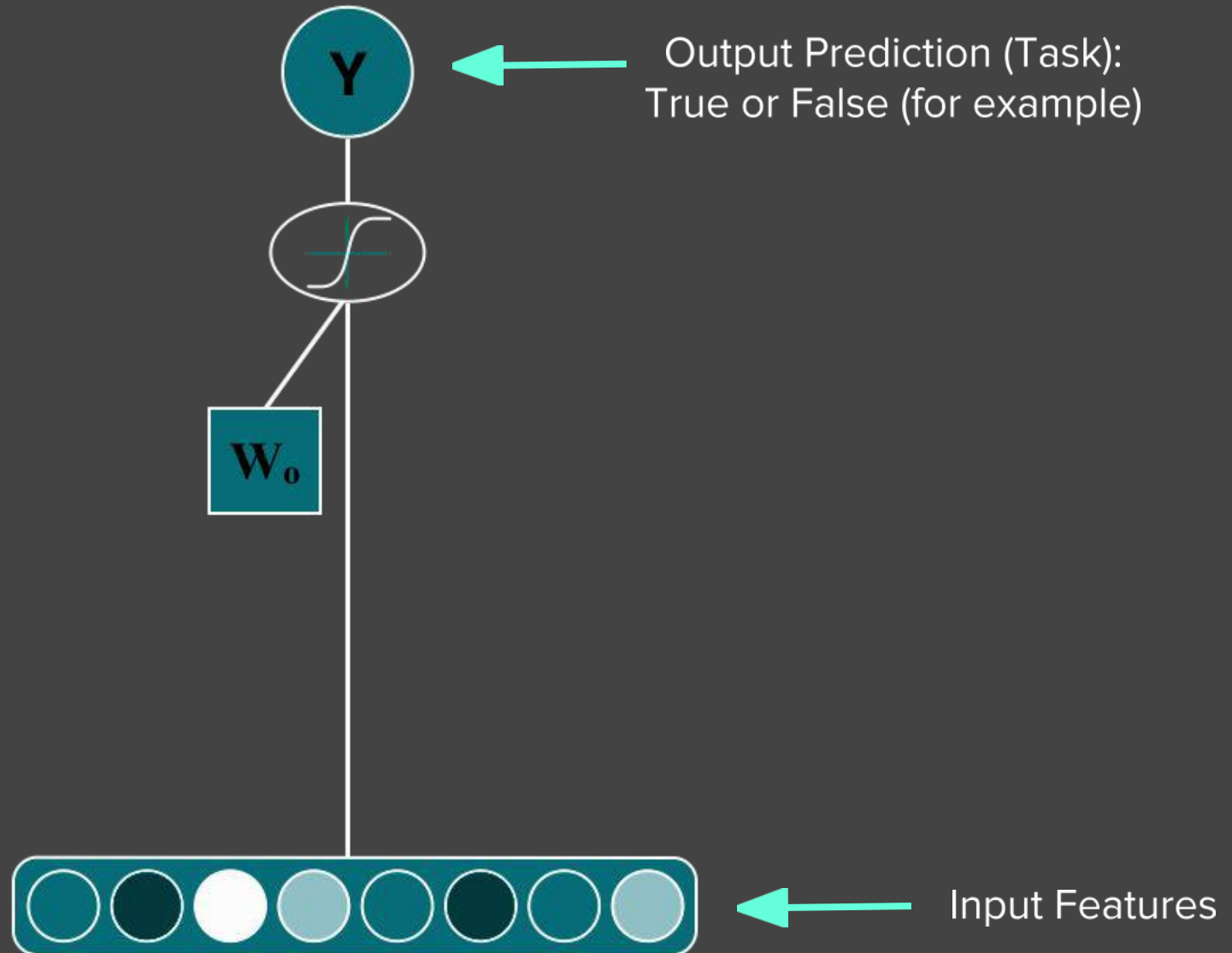
 **Penn | World Well-Being Project**
... advancing understanding of human flourishing
using language analysis



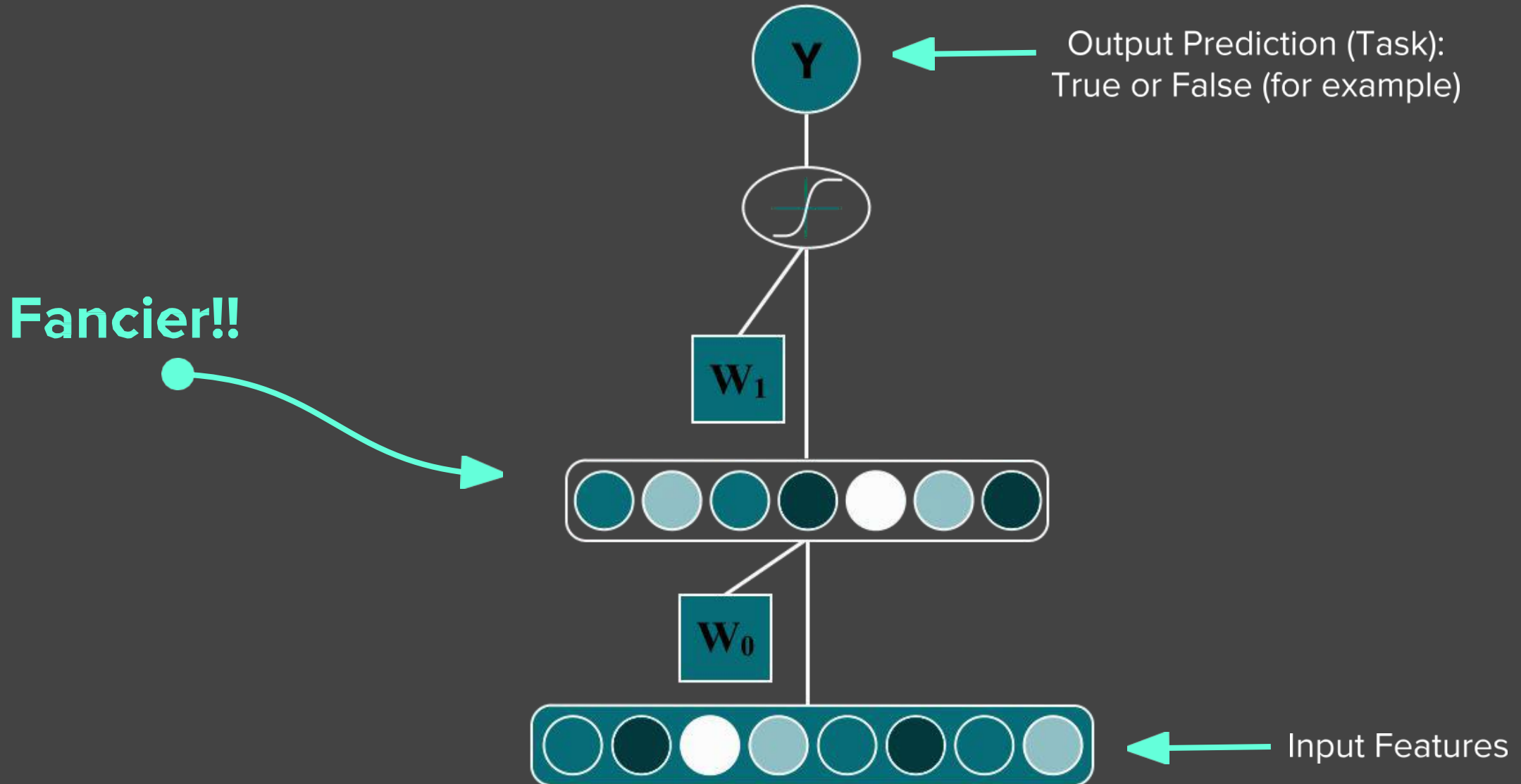
Multiple Tasks + Deep Learning for Inclusion: Multi-task Learning Example

- **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”

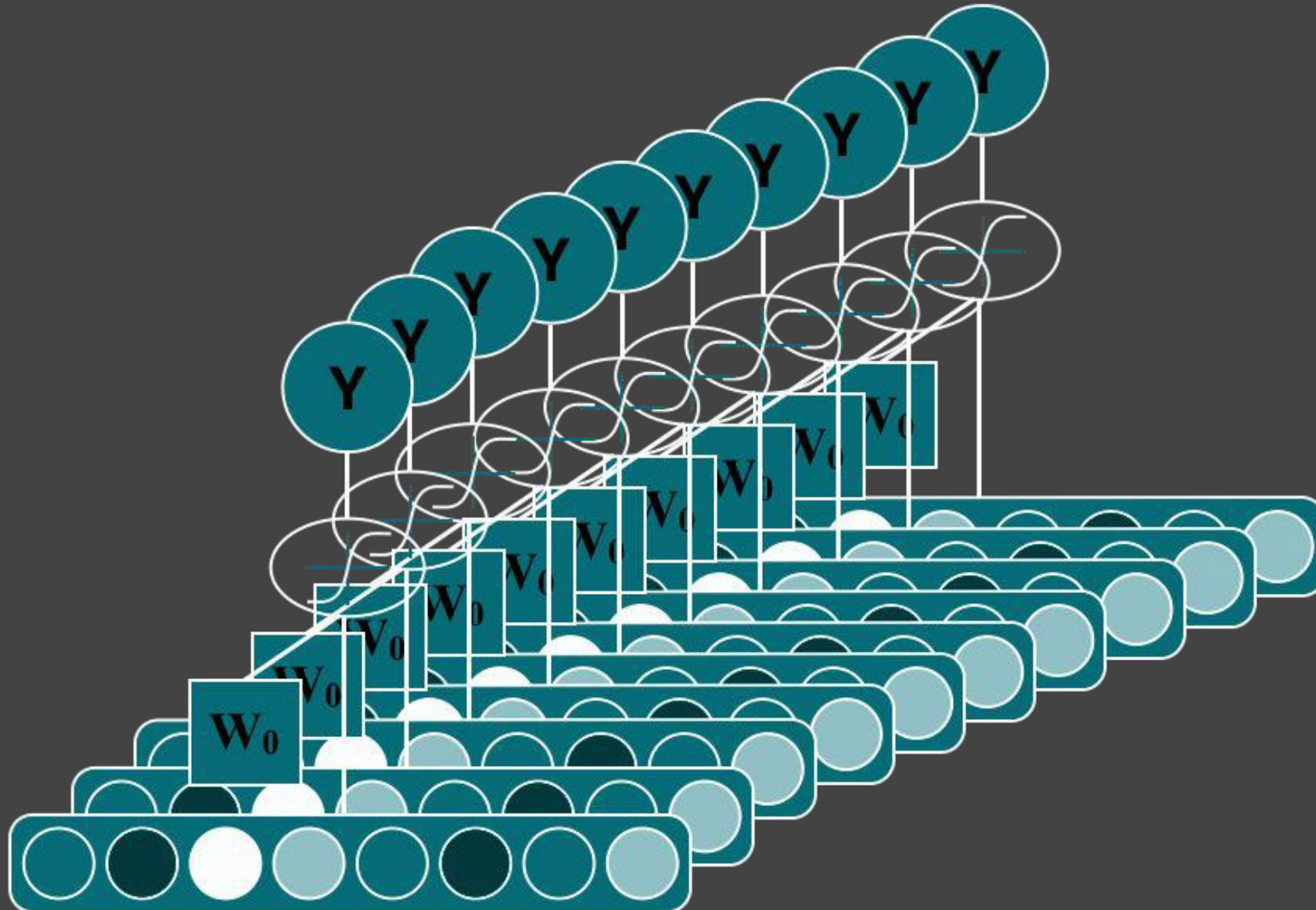
Single-Task: Logistic Regression



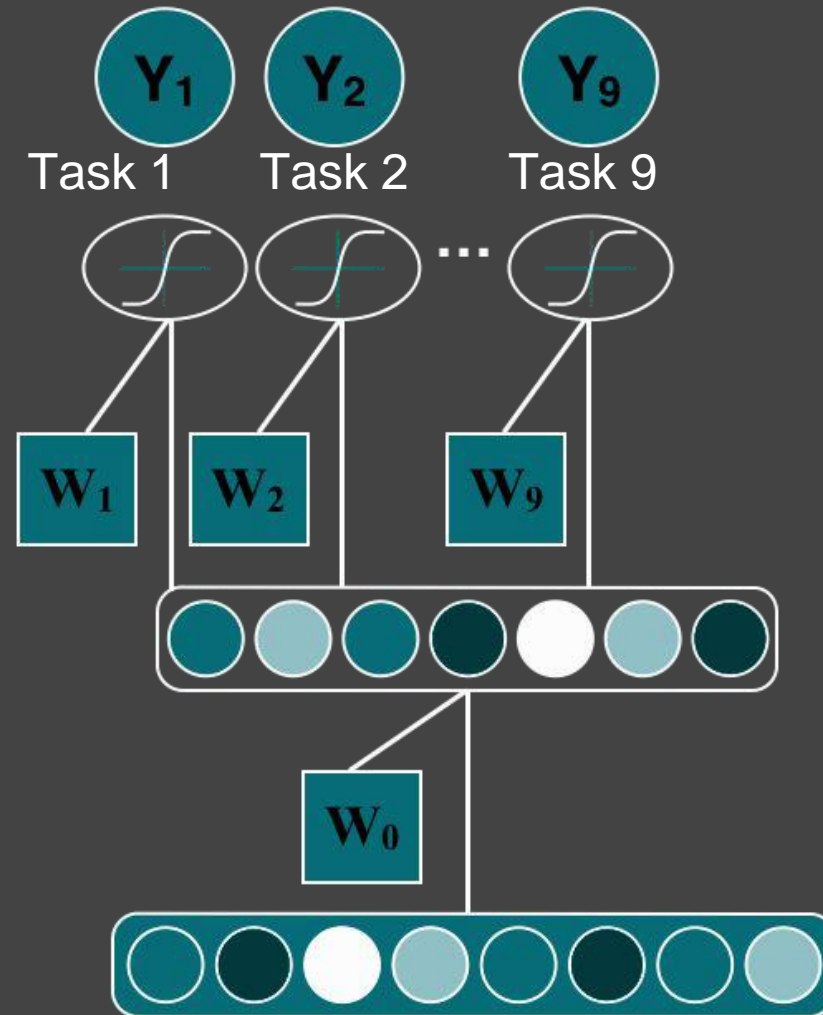
Single-Task: Deep Learning



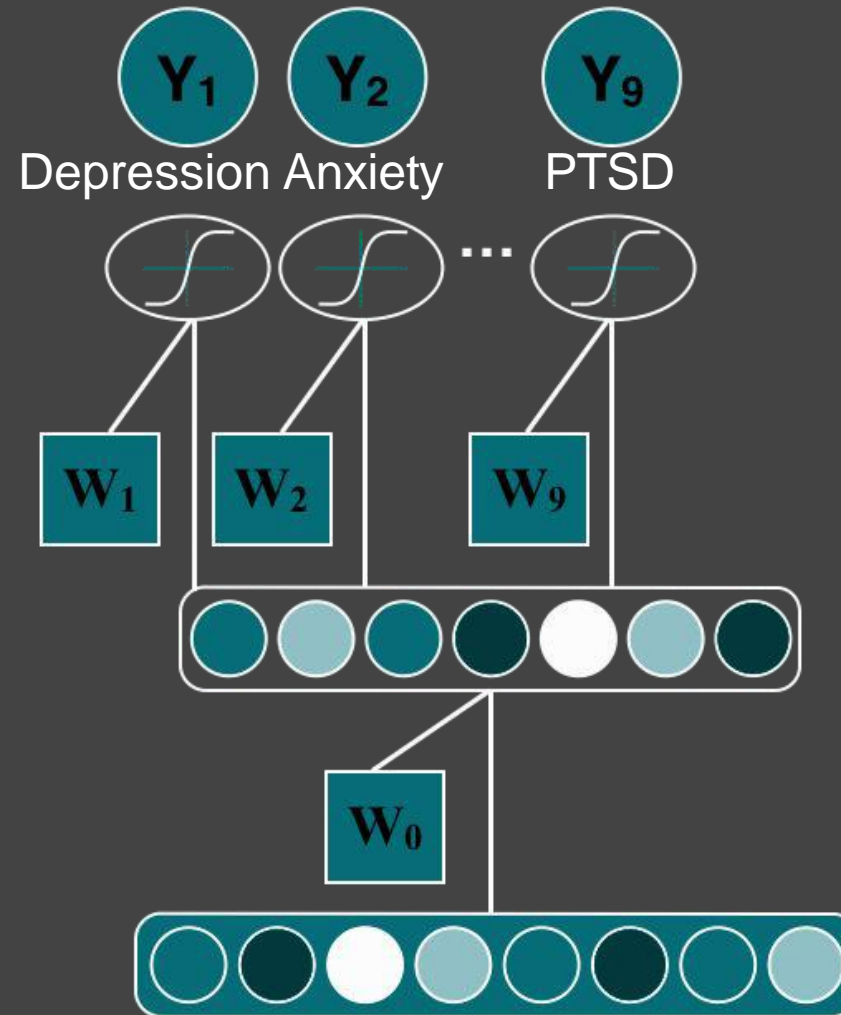
Multiple Tasks with Basic Logistic Regression



Multi-task Learning



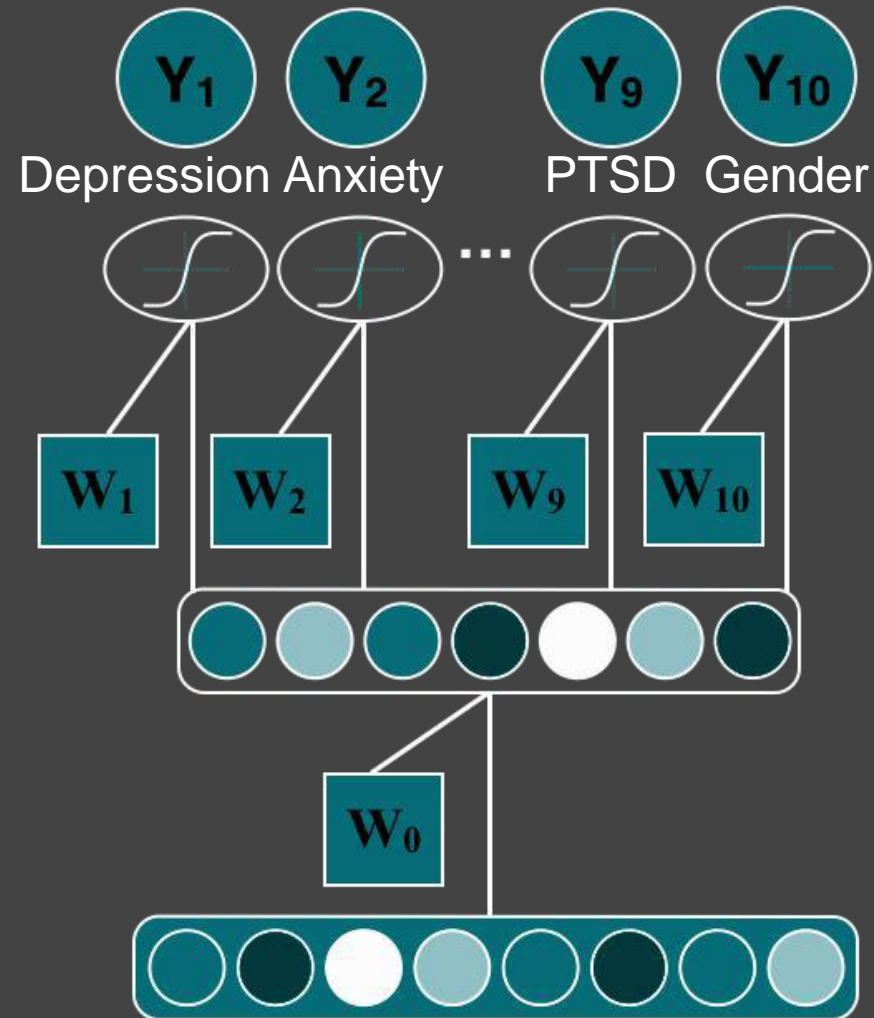
Multi-task Learning



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

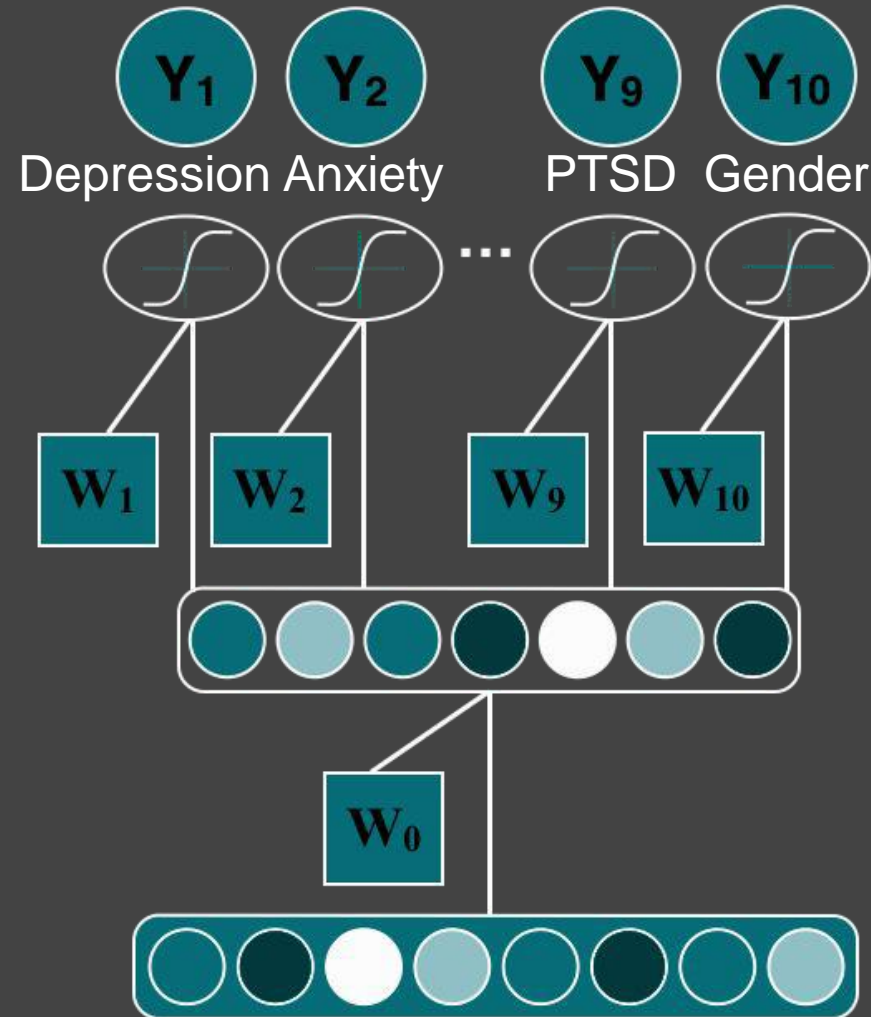


Task	N
Gender	1101
Neurotypicality	4791
Anxiety	2407
Depression	1400
Suicide attempt	1208
Eating disorder	749
Schizophrenia	349
Panic disorder	263
PTSD	248
Bipolar disorder	191
All	9611

} <5% positive examples

Multi-task Learning

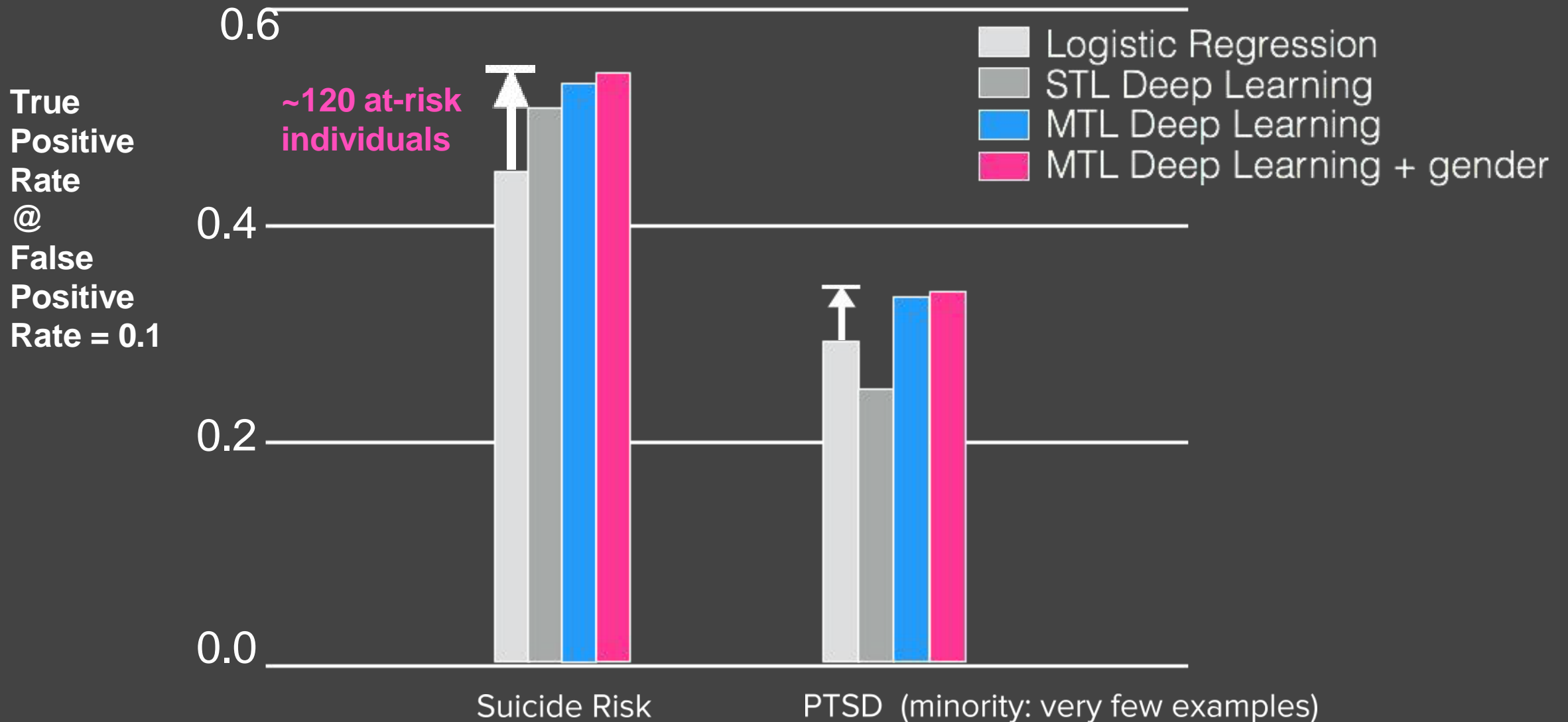
Multitask, given
cormorbidity



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



We sometimes add “auxiliary” tasks to a network to help guide representation learning

- Adds more data/labels to learn general representations
- Encourages representations to encode information related to additional labels, e.g. adding a gender prediction task encourages final feature layer to more directly encode gender



Adversarial Multi-task Learning to Mitigate Bias

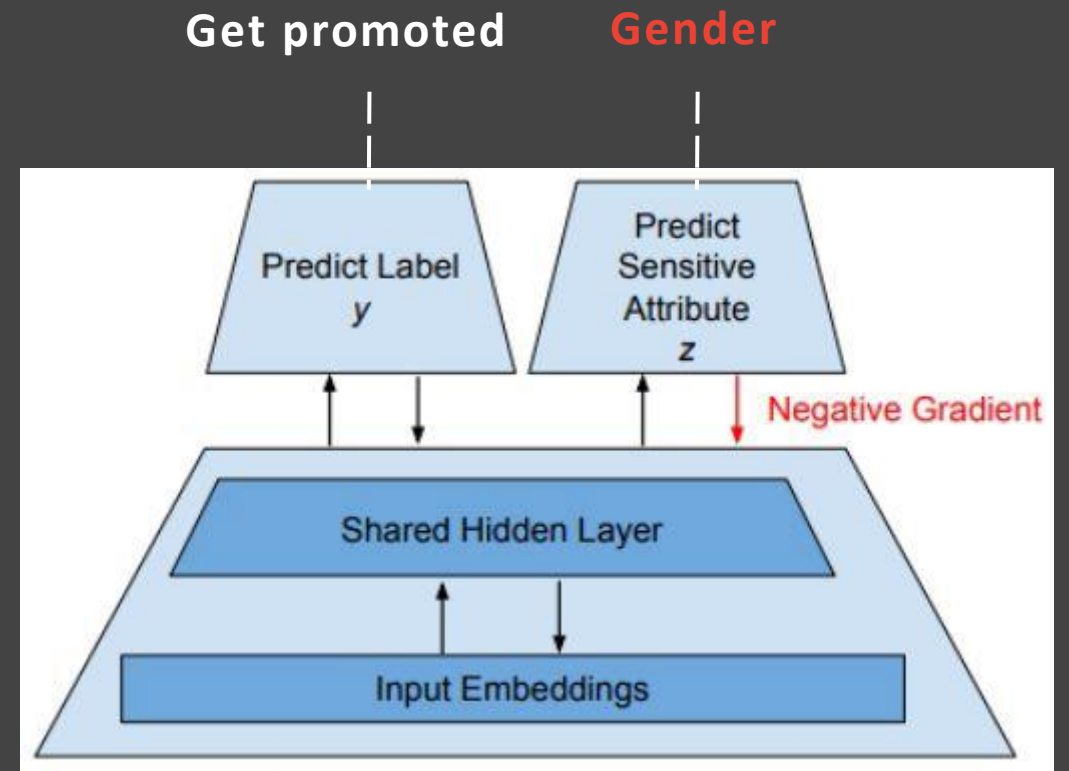
Multitask Adversarial Learning

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

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

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

$Z \perp D | Y$



Beutel, [Chen](#), [Zhao](#), [Chi](#). [Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations](#). *FAT/ML*, 2017.

Zhang, Lemoine, Mitchell. [Mitigating Unwanted Biases with Adversarial Learning](#). *AIES*, 2018.



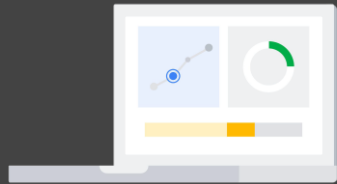
Release Responsibly

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.

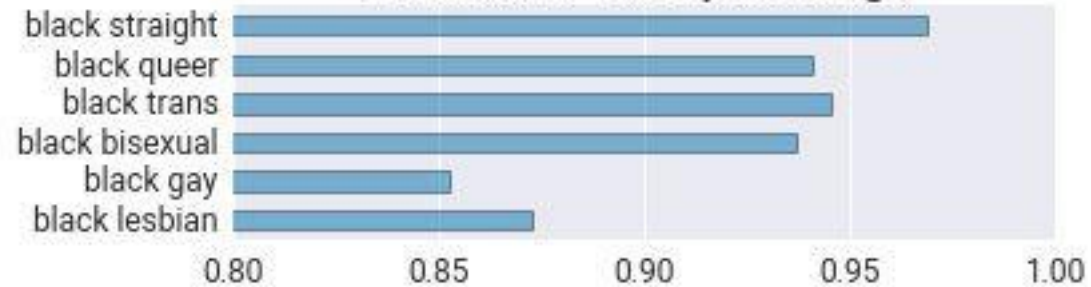
Disaggregated Intersectional Evaluation

Toxicity @1

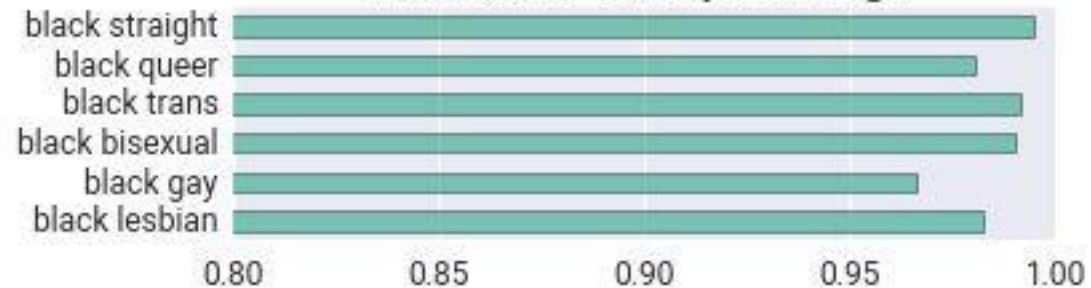
Identity groups	Subgroup AUC	BPSN AUC	BNSP AUC
lesbian	0.93	0.74	0.98
gay	0.94	0.65	0.99
queer	0.98	0.96	0.93
straight	0.99	1.00	0.87
bisexual	0.96	0.95	0.92
homosexual	0.87	0.53	0.99
heterosexual	0.96	0.94	0.92
cis	0.99	1.00	0.87
trans	0.97	0.96	0.91
nonbinary	0.99	0.99	0.90
black	0.91	0.85	0.95
white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



Pinned AUC Toxicity Scores @1



Pinned AUC Toxicity Scores @5



Jigsaw



The False Positive

Things to remember

- **Unless actively countered, bias in algorithms is an expected (harmful) outcome**, due to bias in the data and people involved in creation and use of the algorithms
- Be wary of using algorithms to make decisions about people, e.g. criminality, admissions, hiring
- Adding auxiliary tasks in multi-task learning can improve prediction in under-represented groups
- Adversarial learning can be used to learn representations that are not predictive of sensitive attributes
- Model cards, dataset cards, and intersectional evaluation are used to disclose potential biases and limitations

