Fairness and Ethics in AI

David Fouhey, EECS 442 Winter 2023

https://web.eecs.umich.edu/~fouhey/teaching/EECS442_W23/

(but most of the slides taken from Justin Johnson's Fairness lecture from our last joint offering in W2021)

Disclaimers

- This lecture goes beyond Computer Vision
- I’m not an expert at this but I think it’s really important
- I’m not part of any marginalized group
- We can only begin to scratch the surface in one lecture
- There are generally more problems than solutions

Additional Resources

Timnit Gebru and Emily Denton,

CVPR 2020 Tutorial on FATE/CV

https://sites.google.com/view/fatecv-tutorial/home?authuser=0

Kate Crawford, "The Trouble with Bias",
NeurIPS 2017 Keynote

https://www.youtube.com/watch?v=fMym_BKWQzk

Solon Barocas, Moritz Hardt, Arvind Narayanan,
"*Fairness and machine learning*", https://fairmlbook.org/

ACM Conference on Fairness, Accountability, and Transparency
https://facctconference.org/

Why do we build ML systems?

Automate decision making, so machines can make decision instead of people.

Ideal: Automated decisions can be cheaper, more accurate, more impartial, improve our lives

Reality: If we aren't careful, automated decisions can encode bias, harm people, make lives worse

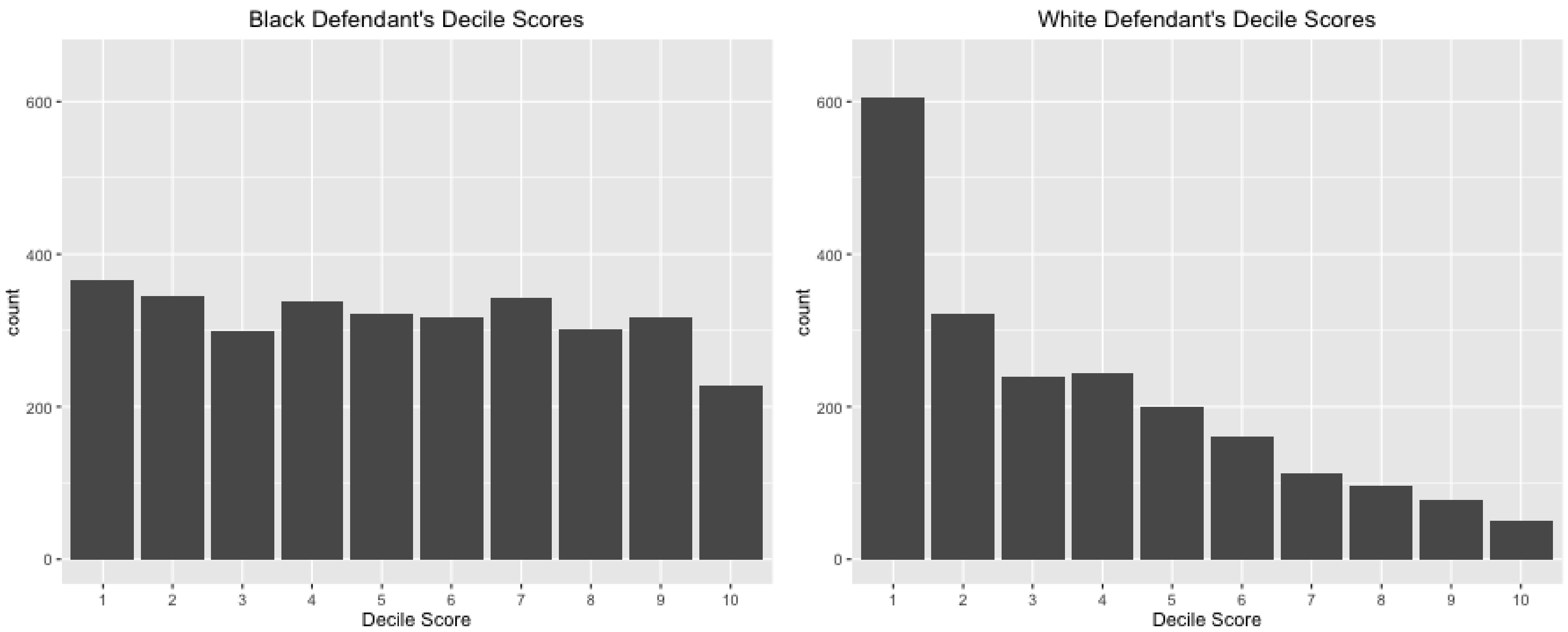
Case Study: COMPAS

1. Person commits a crime, is arrested
2. COMPAS software predicts the chance that the person will commit another crime in the future (*recidivism)*
3. Recidivism scores impact criminal sentences: if a person is likely to commit another crime, shouldn't they get a longer sentence?

Real system that has been used in New York, Wisconsin, California, Florida, etc

Case Study: COMPAS

2016 ProPublica article analyzed COMPAS scores for >7000 people arrested in Broward county, Florida



Question: How many of these people ended up committing new crimes within 2 years?

Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Error Metrics

	Prediction: Low Risk	**Prediction: High Risk**
Outcome: No Recidivism	True Negative (TN)	False Positive (FP)
Outcome: Recidivated	False Negative (FN)	True Positive (TP)

Error Metrics: Error Rate

	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	True Negative (TN)	False Positive (FP)
Outcome: Recidivated	False Negative (FN)	True Positive (TP)

$$\text{Error Rate} = \frac{FP+FN}{TN+FP+FN+TP}$$

How often is the prediction wrong?

Error Metrics: False Positive Rate

	Prediction: Low Risk	**Prediction: High Risk**
Outcome: No Recidivism	**True Negative (TN)**	**False Positive (FP)**
Outcome: Recidivated	False Negative (FN)	True Positive (TP)

Error Rate = $\frac{FP+FN}{TN+FP+FN+TP}$ *How often is the prediction wrong?*

False Positive Rate = $\frac{\boldsymbol{FP}}{\boldsymbol{FP}+\boldsymbol{TN}}$ *How often were non-offenders predicted to reoffend?*

Error Metrics: False Negative Rate

	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	True Negative (TN)	False Positive (FP)
Outcome: Recidivated	**False Negative (FN)**	**True Positive (TP)**

$$\text{Error Rate} = \frac{FP+FN}{TN+FP+FN+TP}$$

How often is the prediction wrong?

$$\text{False Positive Rate} = \frac{FP}{FP+TN}$$

How often were non-offenders predicted to reoffend?

$$\text{False Negative Rate} = \frac{\boldsymbol{FN}}{\boldsymbol{FN}+\boldsymbol{TP}}$$

How often were offenders predicted not to reoffend?

Error Metrics: Different Stakeholders

	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	True Negative (TN)	False Positive (FP)
Outcome: Recidivated	False Negative (FN)	True Positive (TP)

Error Rate = $\frac{FP+FN}{TN+FP+FN+TP}$ *How often is the prediction wrong?*

Defendants care about this

False Positive Rate = $\frac{FP}{FP+TN}$ *How often were non-offenders predicted to reoffend?*

False Negative Rate = $\frac{FN}{FN+TP}$ *How often were offenders predicted not to reoffend?*

Error Metrics: Different Stakeholders

	Prediction: Low Risk	**Prediction: High Risk**
Outcome: No Recidivism	True Negative (TN)	False Positive (FP)
Outcome: Recidivated	False Negative (FN)	True Positive (TP)

$$\text{Error Rate} = \frac{FP+FN}{TN+FP+FN+TP}$$ *How often is the prediction wrong?*

Defendants care about this

$$\text{False Positive Rate} = \frac{FP}{FP+TN}$$ *How often were non-offenders predicted to reoffend?*

Judges care about this

$$\text{False Negative Rate} = \frac{FN}{FN+TP}$$ *How often were offenders predicted not to reoffend?*

Case Study: COMPAS

	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	2681 (TN)	1282 (FP)
Outcome: Recidivated	1216 (FN)	2035 (TP)

$$\text{Error Rate} = \frac{FP+FN}{TN+FP+FN+TP} \approx 34.6\%$$

$$\text{False Positive Rate} = \frac{FP}{FP+TN} \approx 32.4\%$$

$$\text{False Negative Rate} = \frac{FN}{FN+TP} \approx 37.4\%$$

Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Case Study: COMPAS

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Case Study: COMPAS

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Error Rate $\approx 36.2\%$

Error Rate $\approx 33.0\%$

Roughly similar error rates between white and black defendants

Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Case Study: COMPAS

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Error Rate $\approx 36.2\%$

Error Rate $\approx 33.0\%$

False Positive Rate $\approx 44.9\%$

False Positive Rate $\approx 23.5\%$

Black defendants have 1.9x higher False Positive Rate!

Source: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Case Study: COMPAS

Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Error Rate $\approx 36.2\%$

Error Rate $\approx 33.0\%$

False Positive Rate $\approx 44.9\%$

False Positive Rate $\approx 23.5\%$

False Negative Rate $\approx 28.0\%$

False Negative Rate $\approx 47.7\%$

White defendants have 1.7x higher False Negative Rate

Case Study: COMPAS

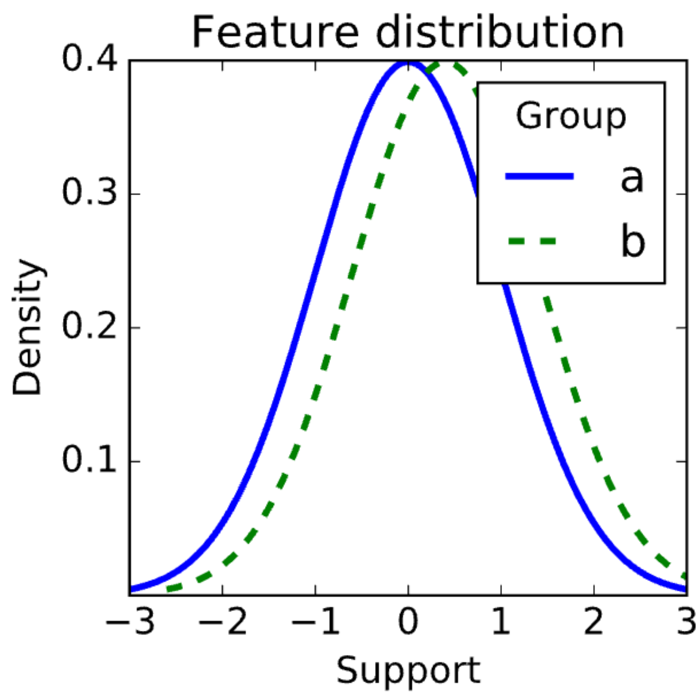
Black Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	990 (TN)	805 (FP)
Outcome: Recidivated	532 (FN)	1369 (TP)

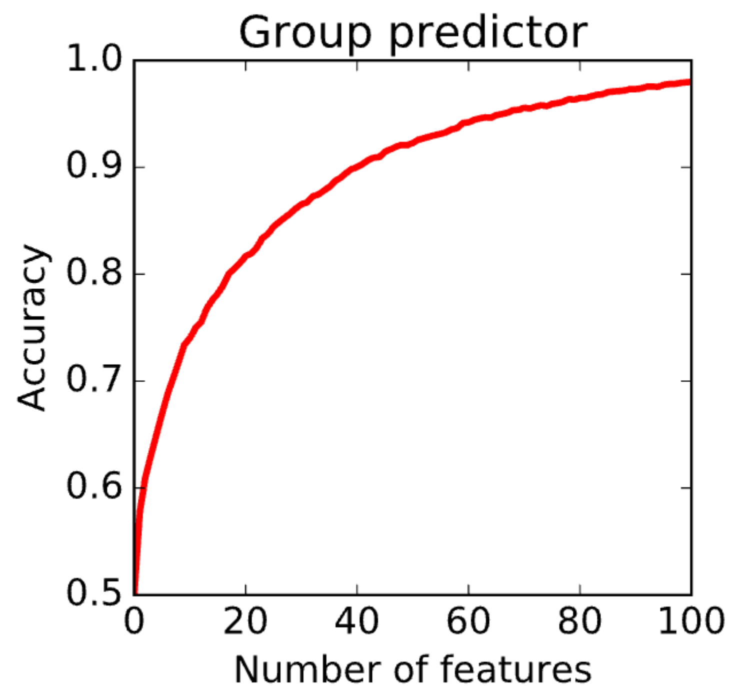
White Defendants	Prediction: Low Risk	Prediction: High Risk
Outcome: No Recidivism	1139 (TN)	349 (FP)
Outcome: Recidivated	461 (FN)	505 (TP)

Surprising fact: COMPAS gives very different outcomes for white vs black defendants, but it does not use race as an input to the algorithm!

No Fairness Through Unawareness

Even if a sensitive feature (e.g. race) is not an input to the algorithm, other features (e.g. zip code) may correlate with the sensitive feature

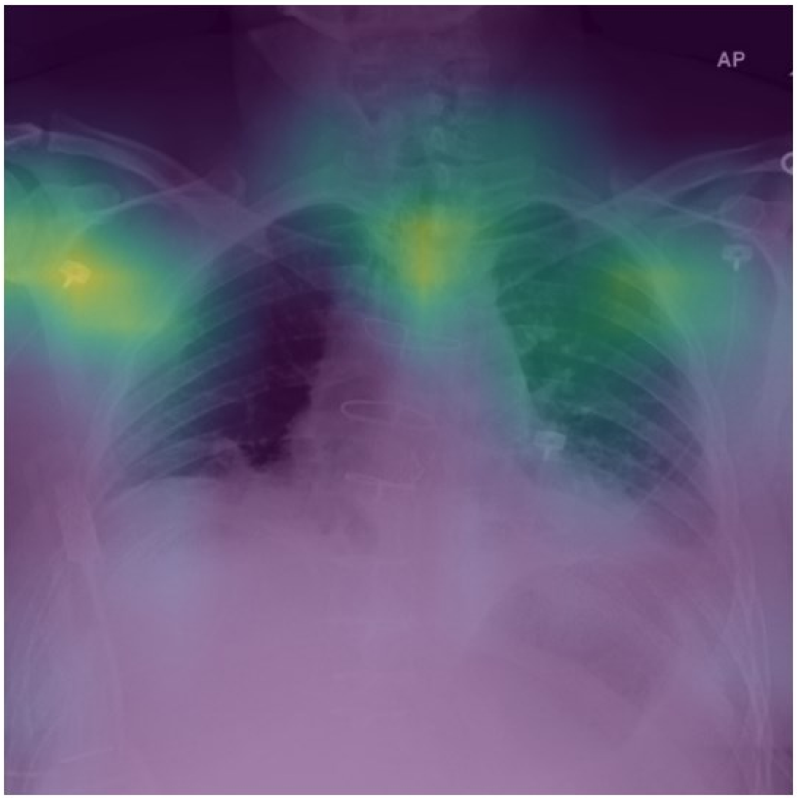




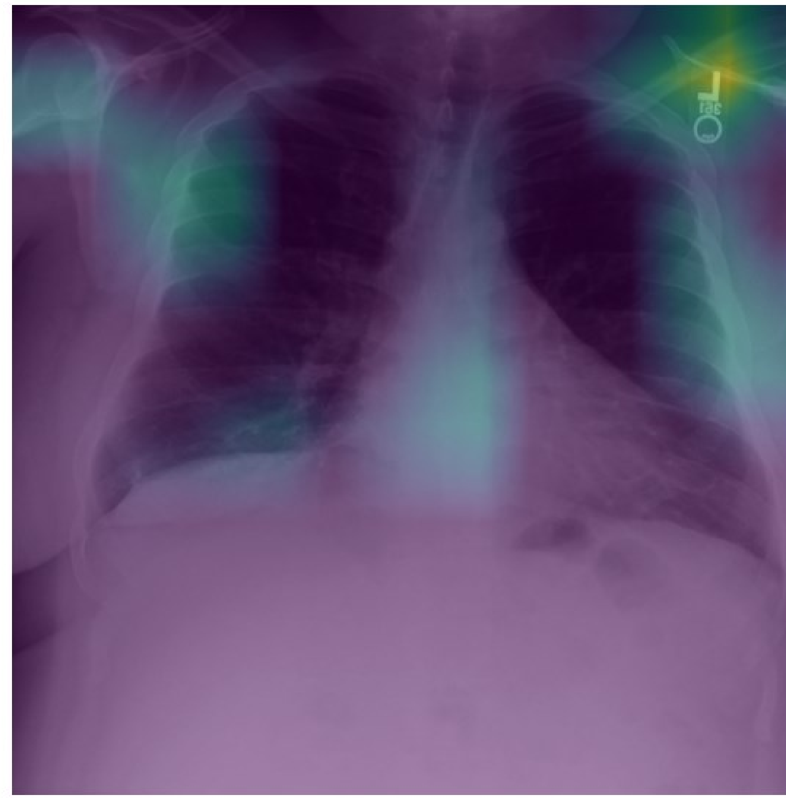
Source: https://fairmlbook.org/classification.html

In Practice

(a)



(b)



Deep Learning Applied to Chest X-Rays: Exploiting and Preventing Shortcuts. S. Jabbour et al. MLHC 2020.

In Practice

Neural networks love taking shortcuts!

Are there shortcuts in our data?

	AHRF		MIMIC-CXR	
Task	**% pos**	**AUROC (95% CI)**	**% pos**	**AUROC (95% CI)**
Age	55	0.72 (0.66-0.78)	57	0.90 (0.89-0.91)
Sex	40	0.96 (0.94-0.98)	46	1.00 (1.00-1.00)
BMI	44	0.91 (0.88-0.94)	–	–
Race	9	0.66 (0.54-0.79)	–	–
Pacemaker	9	0.97 (0.91-1.00)	–	–
Insurance	–	–	9	0.70 (0.67-0.72)
Marital	–	–	44	0.65 (0.63-0.66)

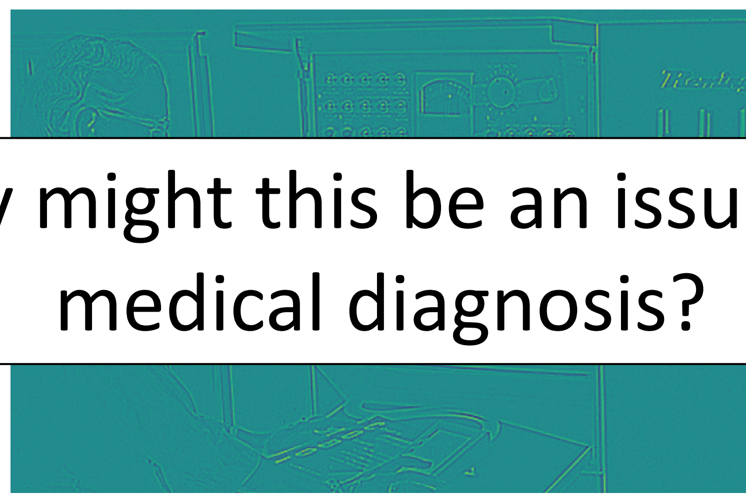
Do I want to get diagnosed by an AI? Stay tuned.

Deep Learning Applied to Chest X-Rays: Exploiting and Preventing Shortcuts. S. Jabbour et al. MLHC 2020.

In Practice

σ = 0.5px

σ = 0.4px



Why might this be an issue for medical diagnosis?

Formalizing Fairness

Y: Target variable (e.g. recidivism)

R: Classifier response (e.g. predicted recidivism)

A: Sensitive attribute (e.g. race)

Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html

Formalizing Fairness

Y: Target variable (e.g. recidivism)

R: Classifier response (e.g. predicted recidivism)

A: Sensitive attribute (e.g. race)

Fairness Definition 1: Independence

The classifier response is *independent* (as a random variable) from the sensitive attribute

$$P(R, A) = P(R)P(A)$$

Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html

Formalizing Fairness

Y: Target variable (e.g. recidivism)

R: Classifier response (e.g. predicted recidivism)

A: Sensitive attribute (e.g. race)

Fairness Definition 1: Independence

The classifier response is *independent* (as a random variable) from the sensitive attribute

$$
\begin{aligned}
P(R, A) &= P(R)P(A) \\
&= P(R \mid A)P(A) \quad \text{(Chain Rule)}
\end{aligned}
$$

$$\Longrightarrow P(R \mid A) = P(R)$$

Barocas, Hardt, and Narayanan. “Fairness and Machine Learning”, https://fairmlbook.org/index.html

Formalizing Fairness

Y: Target variable (e.g. recidivism)

R: Classifier response (e.g. predicted recidivism)

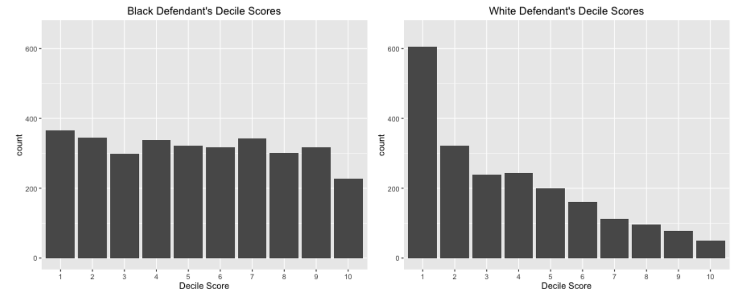
A: Sensitive attribute (e.g. race)

Fairness Definition 1: Independence

The classifier response is *independent* (as a random variable) from the sensitive attribute

$$P(R, A) = P(R)P(A) \implies P(R \mid A) = P(R)$$

COMPAS predictions are not independent – different distributions for black vs white



Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html

Formalizing Fairness

Y: Target variable (e.g. recidivism)

R: Classifier response (e.g. predicted recidivism)

A: Sensitive attribute (e.g. race)

Fairness Definition #2: Separation

The classifier response is *conditionally independent* from the sensitive attribute given the target

$$P(R, A \mid Y) = P(R \mid Y) P(A \mid Y)$$

Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html

Formalizing Fairness

COMPAS scores do not satisfy separation

Fairness Definition #2: Separation

The classifier response is *conditionally independent* from the sensitive attribute given the target

$$P(R, A \mid Y) = P(R \mid Y)P(A \mid Y)$$

By chain rule: $= P(R \mid A, Y)P(A \mid Y)$

Which implies that: $P(R \mid A, Y) = P(R \mid Y)$

Same False Positive Rates between groups:

$$P(R = 1 \mid Y = 0, A = a) = P(R = 1 \mid Y = 0, A = b)$$

Same False Negative Rates between groups:

$$P(R = 0 \mid Y = 1, A = a) = P(R = 0 \mid Y = 1, A = b)$$

Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html

Formalizing Fairness

There are **multiple ways** to formalize notions of fairness mathematically

We've seen two (independence, separation) but there are many more!

Arvind Narayanan, "21 fairness definitions and their politics"
https://www.youtube.com/watch?v=jIXIuYdnyyk

It may be **impossible** to achieve all notions of fairness at the same time

Conclusion: Fairness in ML **is not (purely) a technical problem! We need to think about context, stakeholders**

Allocative Harms

- A system decides how to *allocate resources*
- If the system is biased, it may allocate resources unfairly or perpetuate inequality
- Examples:
  - Sentencing criminals
  - Loan applications
  - Mortgage applications
  - Insurance rates
  - College admissions
  - Job applications

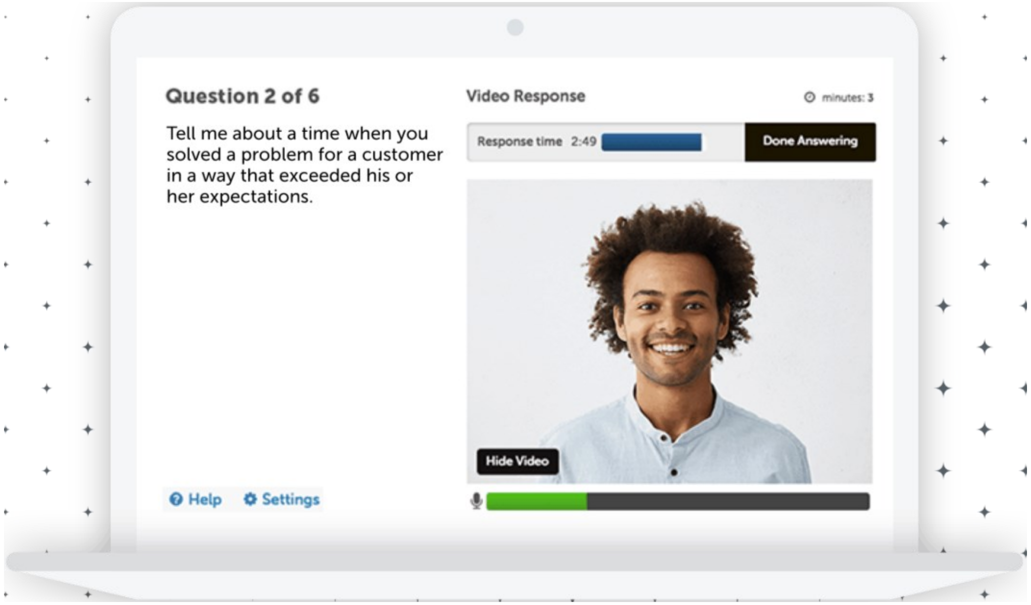
Barocas et al, “The Problem With Bias: Allocative Versus Representational Harms in Machine Learning”, SIGCIS 2017
Kate Crawford, “The Trouble with Bias”, NeurIPS 2017 Keynote

Example: Video Interviewing

Technology

A face-scanning algorithm increasingly decides whether you deserve the job

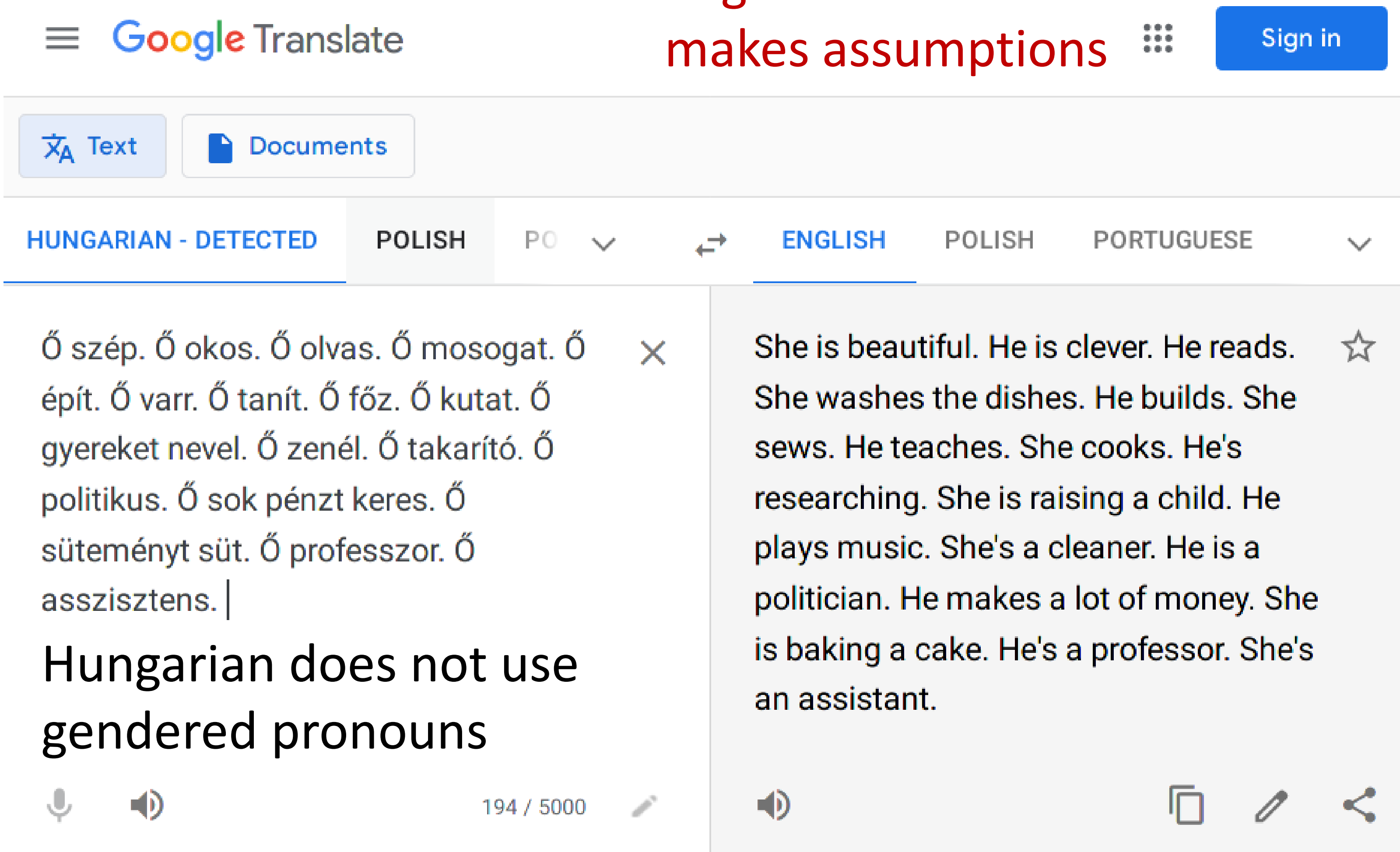
HireVue claims it uses artificial intelligence to decide who's best for a job. Outside experts call it 'profoundly disturbing.'



Source: https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/
https://www.hirevue.com/platform/online-video-interviewing-software
Example Credit: Timnit Gebru

Hungarian -> English Translation

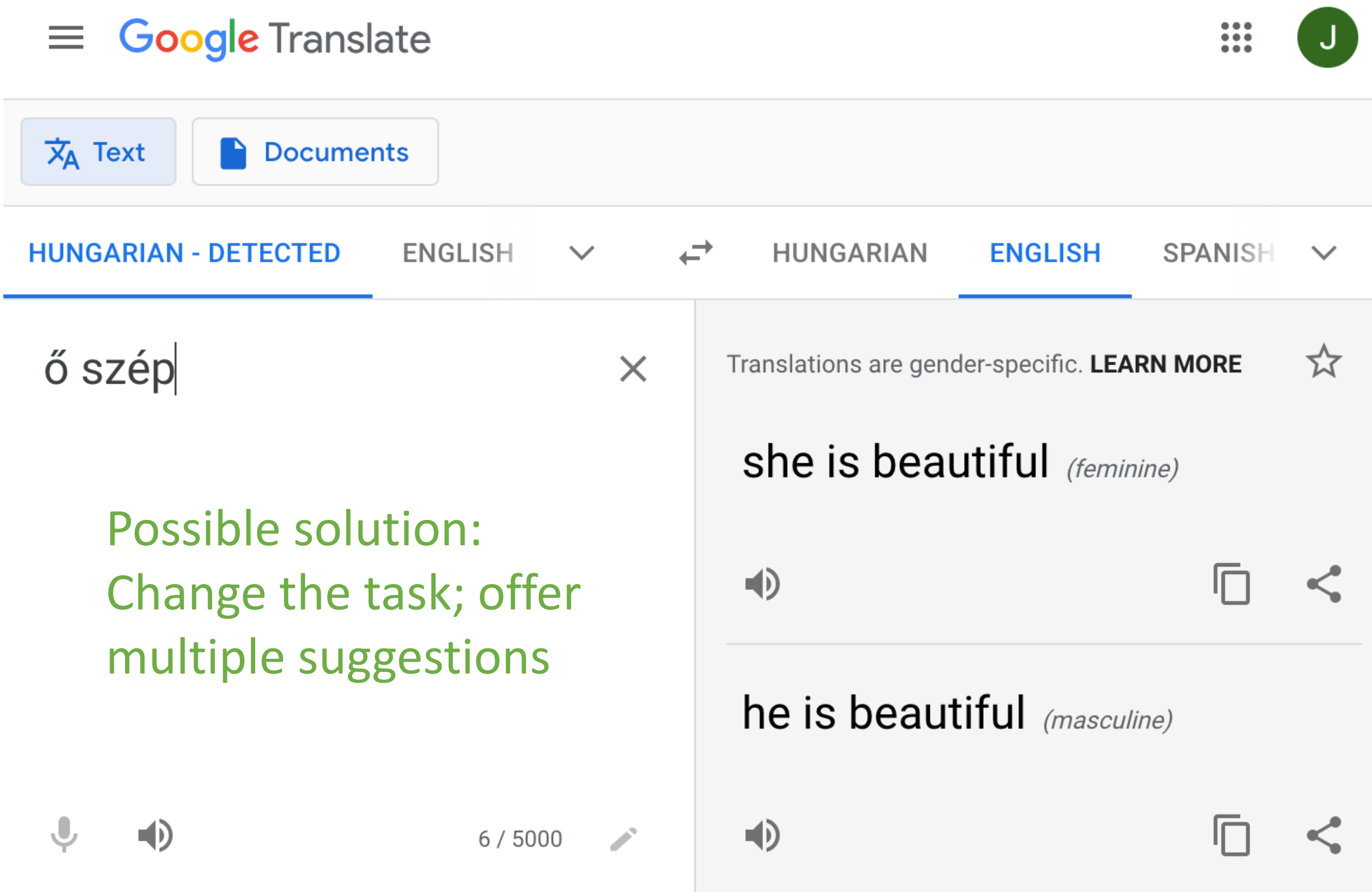
English translation makes assumptions



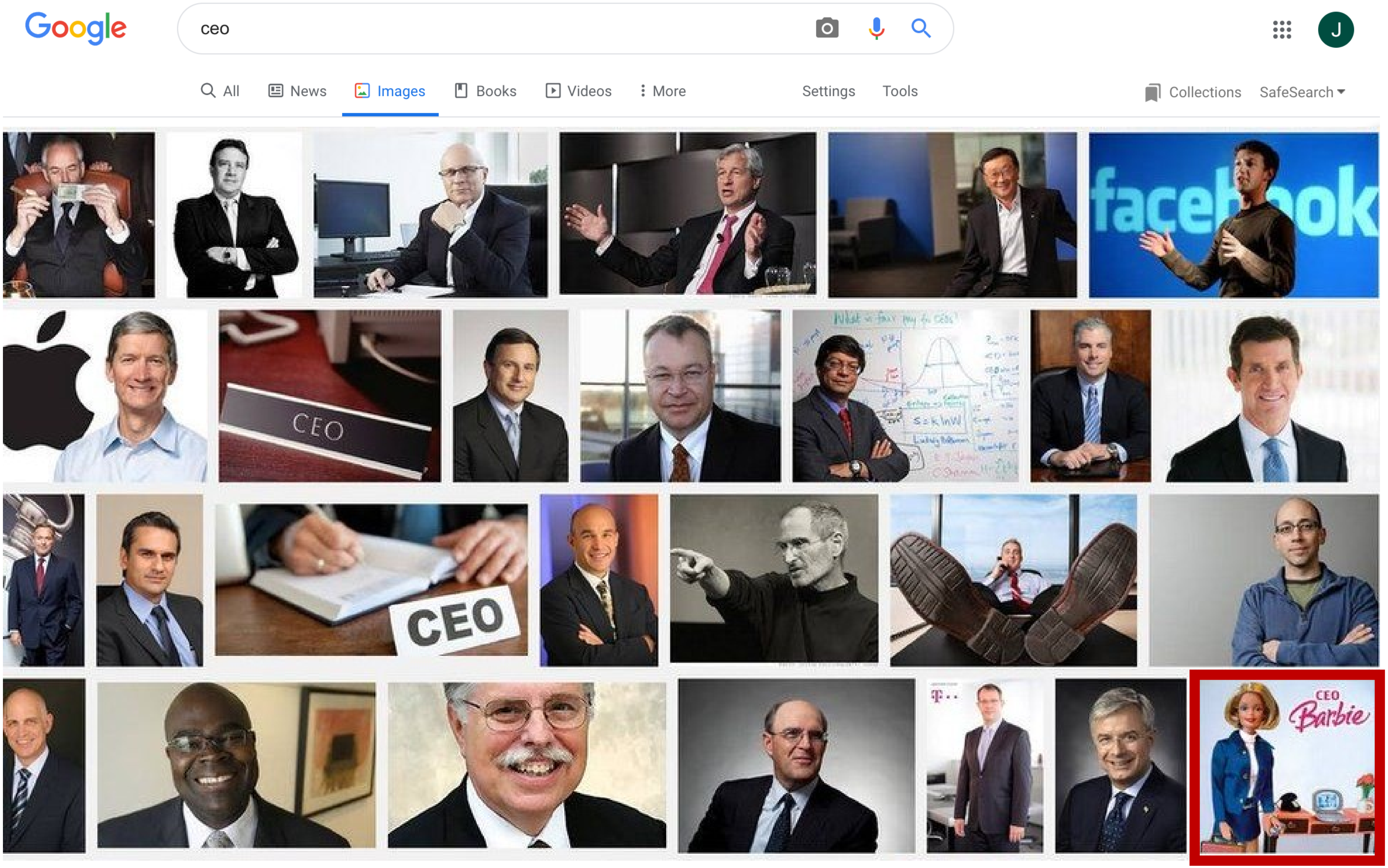
Hungarian does not use gendered pronouns

Source: https://www.reddit.com/r/europe/comments/m9uphb/hungarian_has_no_gendered_pronouns_so_google

Hungarian -> English Translation



Possible solution:
Change the task; offer multiple suggestions



First woman: CEO Barbie =(

Source: https://www.bbc.com/news/newsbeat-32332603

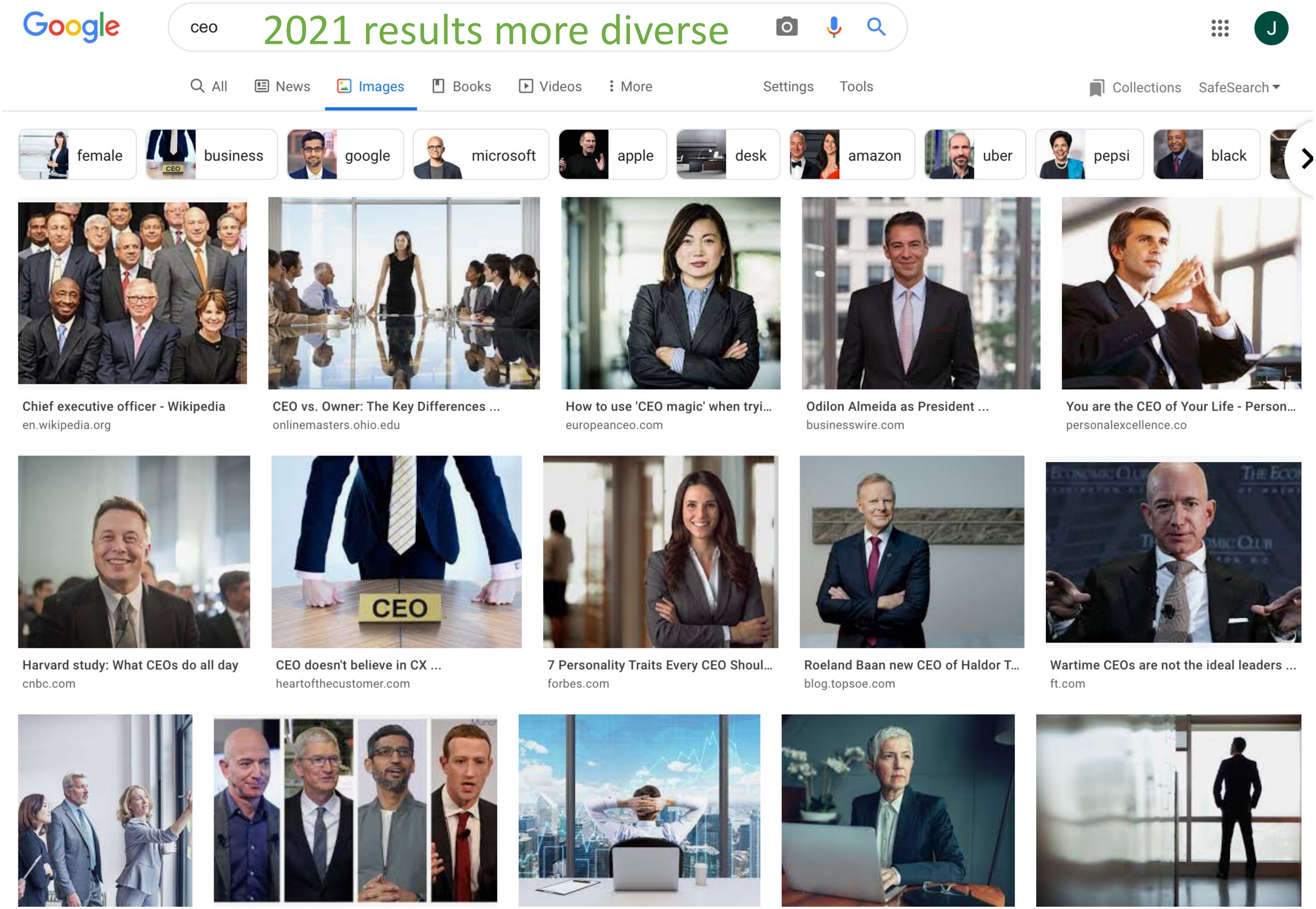

Google
ceo
2021 results more diverse
All
News
Images
Books
Videos
More
Settings
Tools
Collections
SafeSearch
female
business
google
microsoft
apple
desk
amazon
uber
pepsi
black
Chief executive officer - Wikipedia
en.wikipedia.org
CEO vs. Owner: The Key Differences ...
onlinemasters.ohio.edu
How to use 'CEO magic' when tryi...
europeanceo.com
Odilon Almeida as President ...
businesswire.com
You are the CEO of Your Life - Person...
personalexcellence.co
Harvard study: What CEOs do all day
cnbc.com
CEO
CEO doesn't believe in CX ...
heartofthecustomer.com
7 Personality Traits Every CEO Shoul...
forbes.com
Roeland Baan new CEO of Haldor T...
blog.topsoe.com
Wartime CEOs are not the ideal leaders ...
ft.com

Image Super-Resolution

Input: Low-Resolution Face

Output: High-Resolution Face

Menon et al, "PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models", CVPR 2020
Example source: https://twitter.com/Chicken3gg/status/1274314622447820801

Pre-AI Photos

Photo credit: NYTimes.com

- What does this photo do?
- Back in the day you got your photos printed. Kodak had print shops calibrate their settings via "Shirley Cards"
- Calibration settings totally off for people with darker skin!

Economic Bias in Visual Classifiers

Ground-Truth: Soap
Source: UK, $1890/month

Azure: toilet, design, art, sink
Clarifai: people, faucet, healthcare, lavatory, wash closet
Google: product, liquid, water, fluid, bathroom accessory
Amazon: sink, indoors, bottle, sink faucet
Watson: gas tank, storage tank, toiletry, dispenser, soap dispenser
Tencent: lotion, toiletry, soap dispenser, dispenser, after shave

Ground-Truth: Soap
Source: Nepal, $288/month

Azure: food, cheese, bread, cake, sandwich
Clarifai: food, wood, cooking, delicious, healthy
Google: food, dish, cuisine, comfort food, spam
Amazon: food, confectionary, sweets, burger
Watson: food, food product, turmeric, seasoning
Tencent: food, dish, matter, fast food, nutriment

DeVries et al, “Does Object Recognition Work for Everyone?”, CVPR Workshops, 2019

Problem: Datasets are Biased

*This analysis conflates gender with sex, and assumes that it is binary.

Example: COCO Dataset

Multilabel Classification

Person
Umbrella
Cat

Define "gender bias" of object category C as:

$$\frac{\#(C, Man)}{\#(C, Man) + \#(C, Woman)}$$

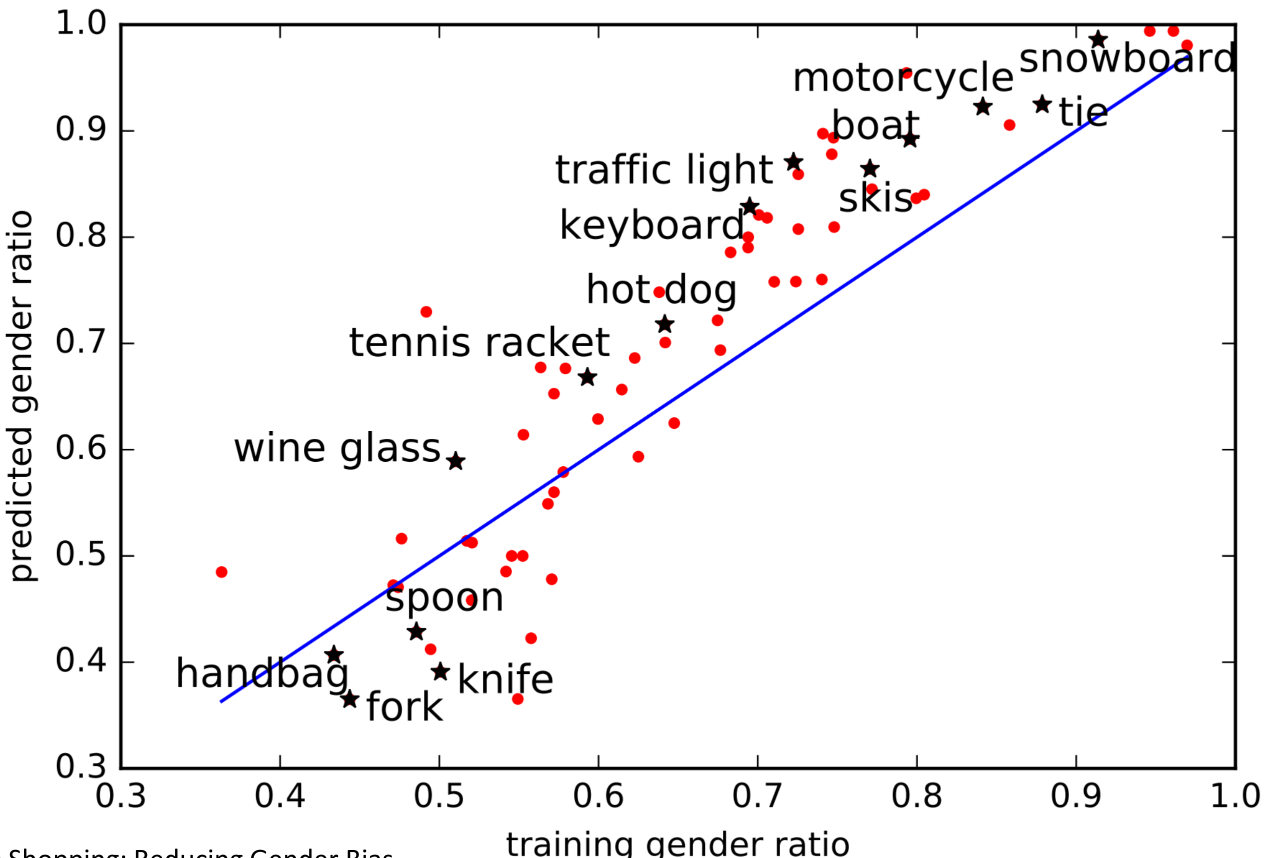
Example: "Snowboards" are 90% biased towards men

Zhao et al, "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints", EMNLP 2017

Problem: Bias Amplification

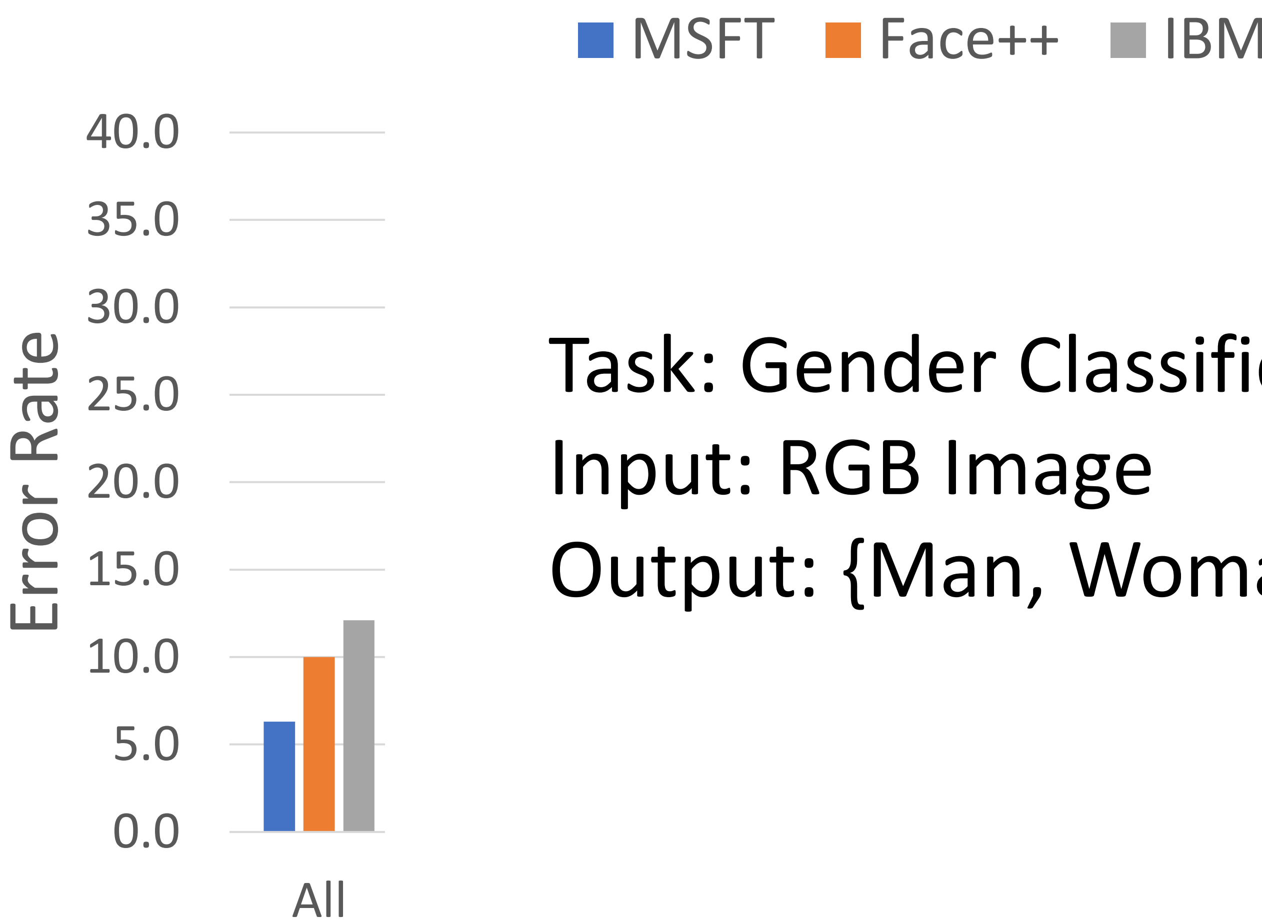
CNN predictions are **more biased** than their training data!

Reducing bias in datasets is **not enough**



Zhao et al, "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints", EMNLP 2017

Gender Shades: Intersectionality



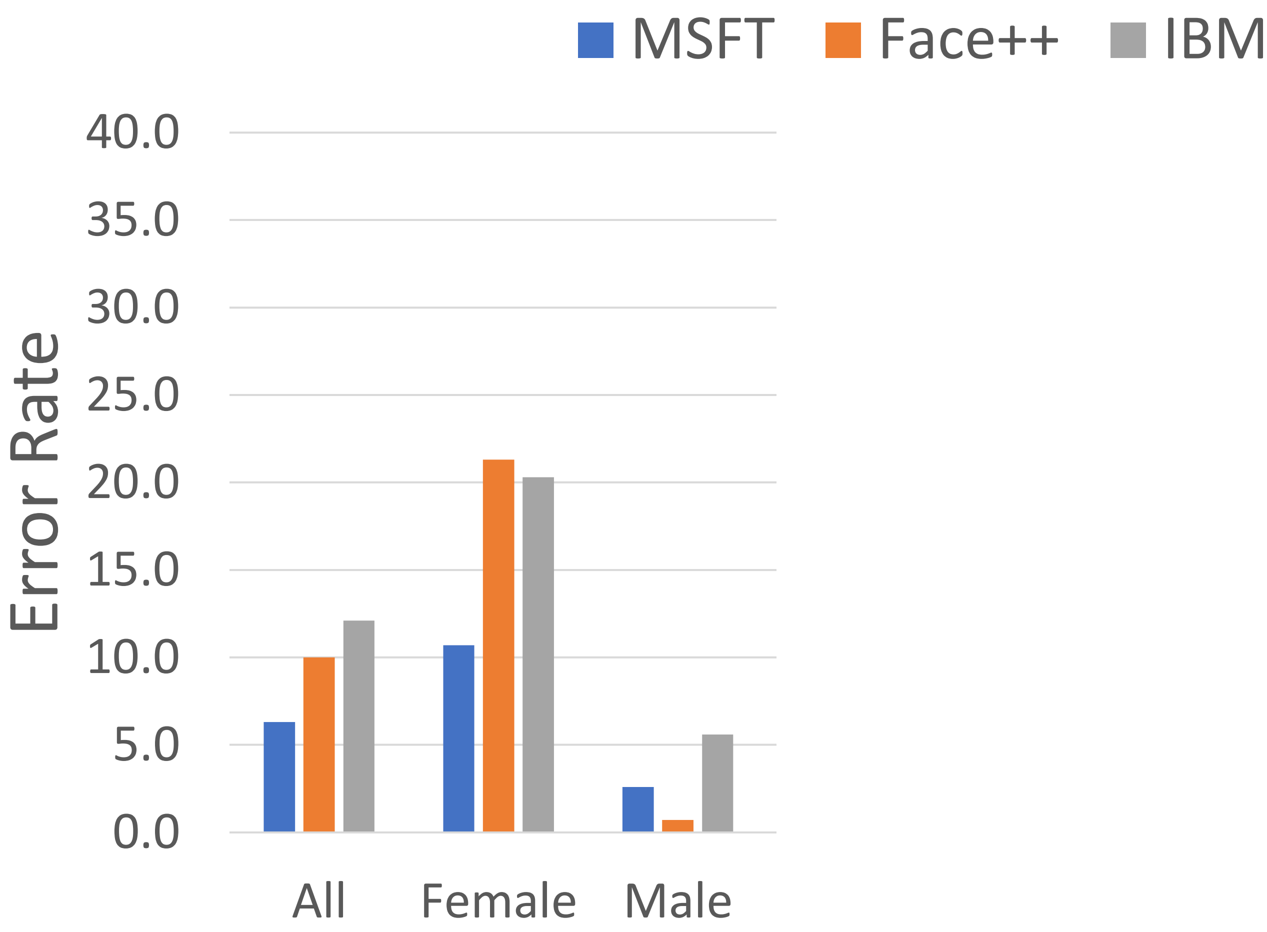
Task: Gender Classification

Input: RGB Image

Output: {Man, Woman} Prediction

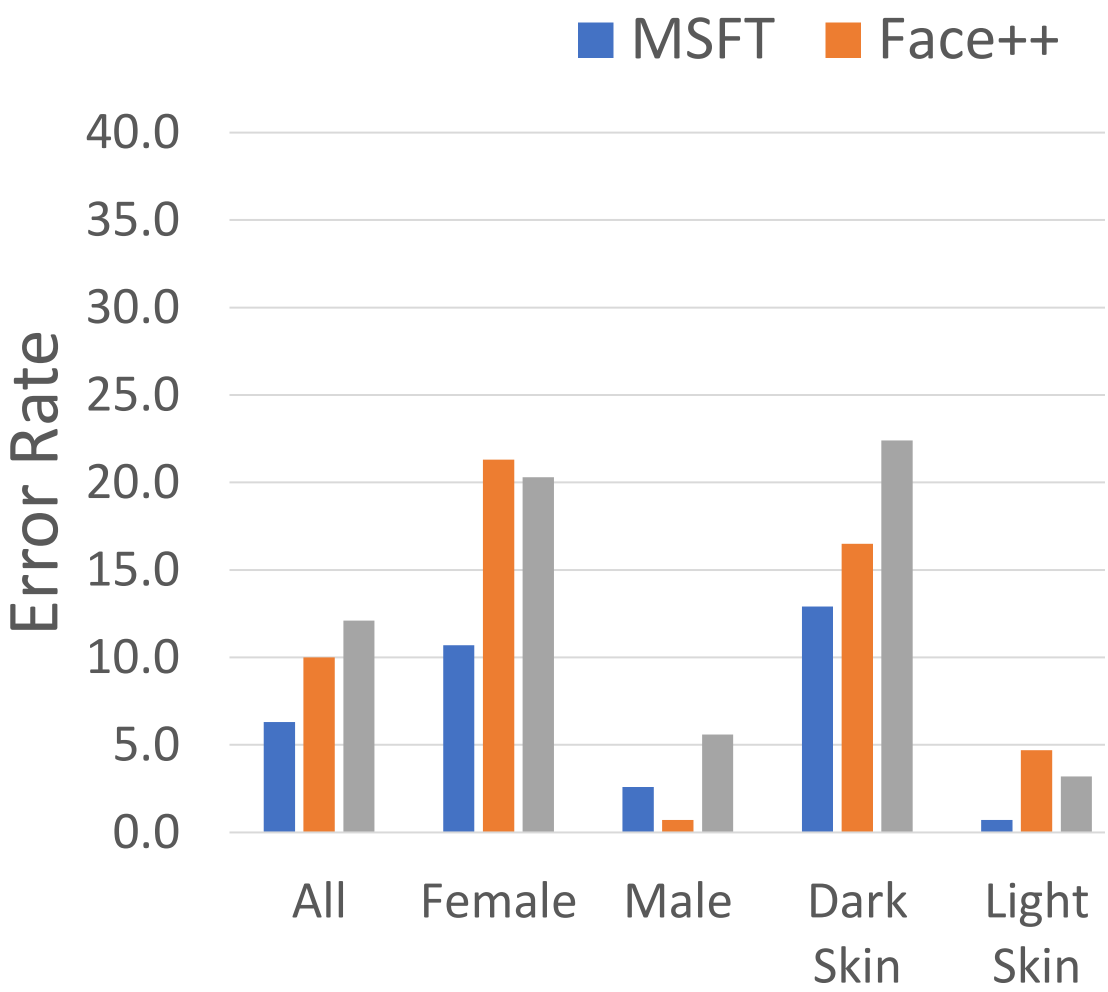
Buolamwini and Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”, FAT* 2018

Gender Shades: Intersectionality



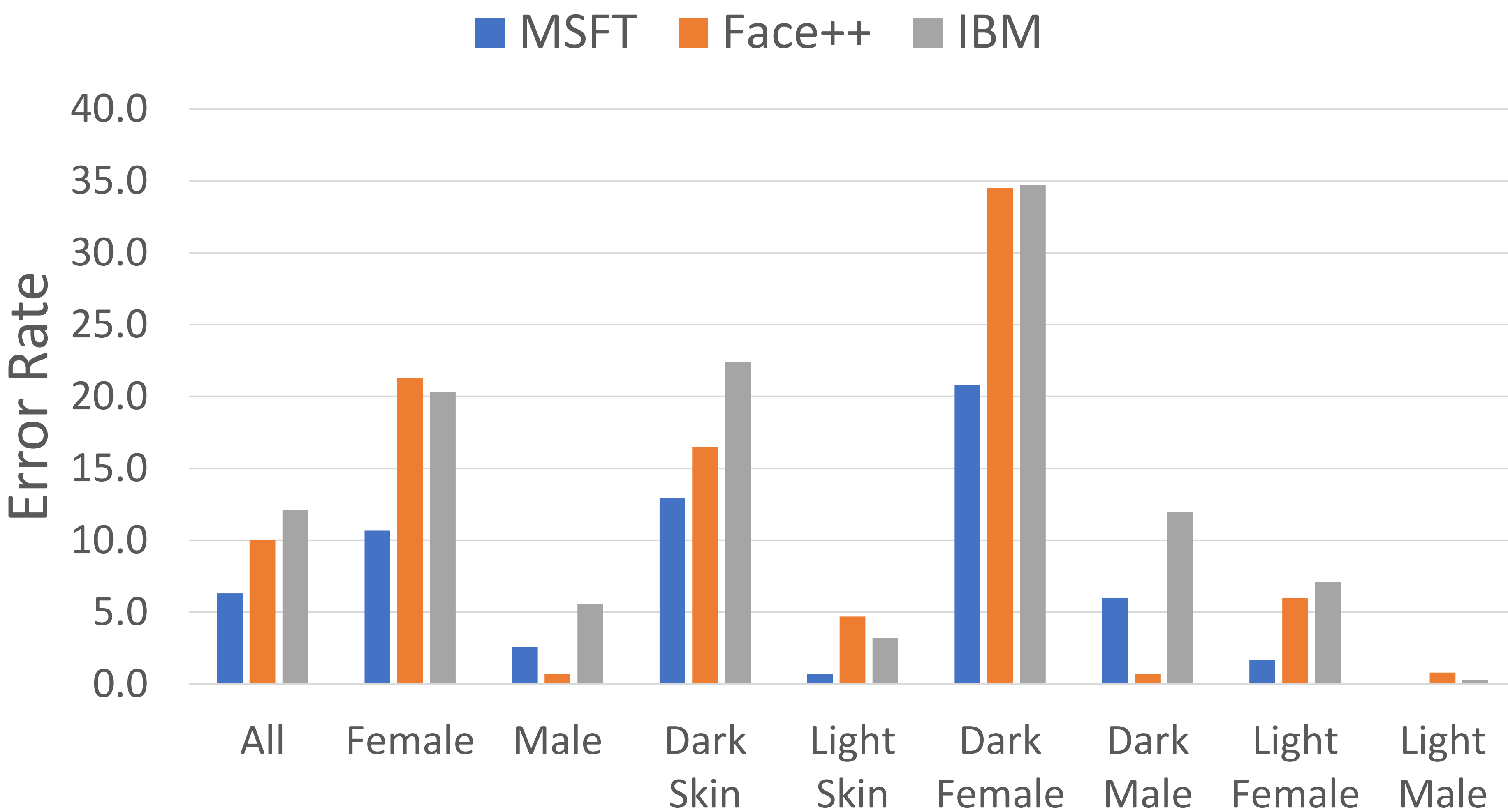
Buolamwini and Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”, FAT* 2018

Gender Shades: Intersectionality



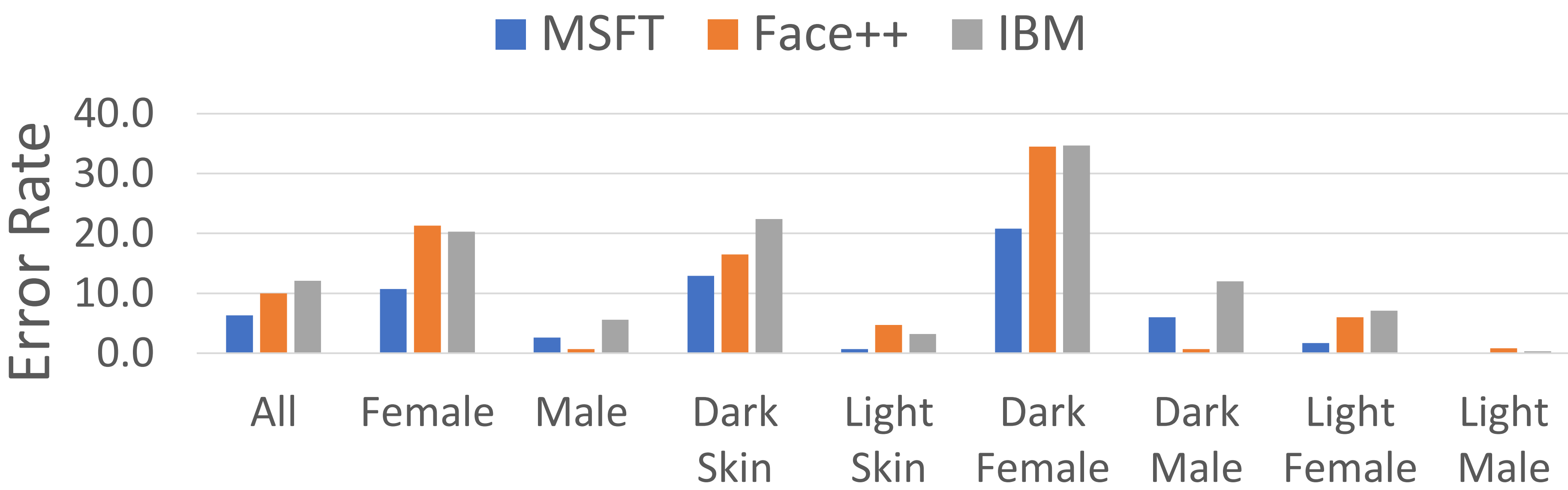
Buolamwini and Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”, FAT* 2018

Gender Shades: Intersectionality



Buolamwini and Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”, FAT* 2018

Gender Shades: Intersectionality



Problem: Much higher error rate for dark-skinned women

Bigger Problem: Why are we classifying gender at all?
Why does an automated system care? If it does, ask!

Buolamwini and Gebru, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”, FAT* 2018

Think Critically about Datasets

CelebA Dataset: 202k images labeled with 40 binary attributes

Liu et al, “Deep Learning Face Attributes in the Wild”, ICCV 2015

Think Critically about Datasets

CelebA Dataset: 202k images labeled with 40 binary attributes

5_o_Clock_Shadow
Arched_Eyebrows
Attractive
Bags_Under_Eyes
Bald
Bangs
Big_Lips
Big_Nose
Black_Hair
Blond_Hair
Blurry
Brown_Hair
Bushy_Eyebrows
Chubby

Double_Chin
Eyeglasses
Goatee
Gray_Hair
Heavy_Makeup
High_Cheekbones
Male
Mouth_Slightly_Open
Mustache
Narrow_Eyes
No_Beard
Oval_Face
Pale_Skin

Pointy_Nose
Receding_Hairline
Rosy_Cheeks
Sideburns
Smiling
Straight_Hair
Wavy_Hair
Wearing_Earrings
Wearing_Hat
Wearing_Lipstick
Wearing_Necklace
Wearing_Necktie
Young

Liu et al, "Deep Learning Face Attributes in the Wild", ICCV 2015

Think Critically about Datasets

CelebA Dataset: 202k images labeled with 40 binary attributes

5_o_Clock_Shadow	**Double_Chin**	**Pointy_Nose**
Arched_Eyebrows	Eyeglasses	Receding_Hairline
Attractive	Goatee	Rosy_Cheeks
Bags_Under_Eyes	Gray_Hair	Sideburns
Bald	**Heavy_Makeup**	Smiling
Bangs	High_Cheekbones	Straight_Hair
Big_Lips	Male	Wavy_Hair
Big_Nose	Mouth_Slightly_Open	Wearing_Earrings
Black_Hair	Mustache	Wearing_Hat
Blond_Hair	Narrow_Eyes	Wearing_Lipstick
Blurry	No_Beard	Wearing_Necklace
Brown_Hair	Oval_Face	Wearing_Necktie
Bushy_Eyebrows	**Pale_Skin**	**Young**
Chubby		

Many attributes seem subjective. Who chose the attributes? Why? How are they defined? Who labeled the images?

Liu et al, "Deep Learning Face Attributes in the Wild", ICCV 2015

Think Critically about Datasets

CelebA Dataset: 202k images labeled with 40 binary attributes

5_o_Clock_Shadow	Double_Chin	Pointy_Nose
Arched_Eyebrows	Eyeglasses	Receding_Hairline
Attractive	Goatee	Rosy_Cheeks
Bags_Under_Eyes	Gray_Hair	Sideburns
Blurry	No_Beard	Wearing_Necklace
Brown_Hair	Oval_Face	Wearing_Necktie
Bushy_Eyebrows	Pale_Skin	Young
Chubby		

Almost no detail in the paper

images of 5,749 identities. Each image in CelebA and LFWA is annotated with forty face attributes and five key points by a professional labeling company. CelebA and LFWA have over eight million and five hundred thousand attribute labels, respectively.

Many attributes seem subjective. Who chose the attributes? Why? How are they defined? Who labeled the images?

Liu et al, "Deep Learning Face Attributes in the Wild", ICCV 2015

Datasheets for Datasets

Idea: A standard list of questions to answer when releasing a dataset. Who created it? Why? What is in it? How was it labeled?

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The initial version of the dataset was created by Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, most of whom were researchers at the University of Massachusetts Amherst at the time of the dataset's release in 2007.

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

The construction of the LFW database was supported by a United States National Science Foundation CAREER Award.

The dataset does not contain all possible instances. There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format.

Is there a label or target associated with each instance? If so, please provide a description.

Each image is accompanied by a label indicating the name of the person in the image.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included in the dataset.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

There are no known relationships between instances except for

Gebru et al, "Datasheets for Datasets", FAccT 2018

Model Cards

Idea: A standard list of questions to answer when releasing a trained model. Who created it? What data was it trained on? What should it be used for? What should it **not** be used for?

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Mitchell et al, "Model Cards for Model Reporting", FAccT 2019

Model Cards

Adopted by Google, OpenAI

Object Detection

Model Card v0 Cloud Vision API

Overview

Limitations

Performance

Test your own images

Provide feedback

Explore

Face Detection

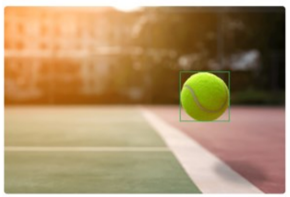
About Model Cards

Object Detection

The model analyzed in this card detects one or more physical objects within an image, from apparel and animals to tools and vehicles, and returns a box around each object, as well as a label and description for each object.

On this page, you can learn more about how the model performs on different classes of objects, and what kinds of images you should expect the model to perform well or poorly on.

MODEL DESCRIPTION



Input: Photo(s) or video(s)

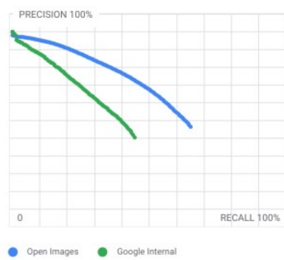
Output: The model can detect 550+ different object classes. For each object detected in a photo or video, the model outputs:

- Object bounding box coordinates
- Knowledge graph ID ("MID")
- Label description
- Confidence score

Model architecture: Single shot detector model with a Resnet 101 backbone and a feature pyramid network feature map.

View public API documentation

PERFORMANCE



Performance evaluated for specific object classes recognized by the model (e.g. shirt, muffin), and for categories of objects (e.g. apparel, food).

Two performance metrics are reported:

- Average Precision (AP)
- Recall at 60% Precision

Performance evaluated on two datasets distinct from the training set:

- Open Images Validation set, which contains ~40k images and 600 object classes, of which the model can recognize 518.
- An internal Google dataset of ~5,000 images of consumer products, containing 210 object classes, all of which model can recognize.

Go to performance

Model Card: CLIP

Inspired by Model Cards for Model Reporting (Mitchell et al.) and Lessons from Archives (Jo & Gebru), we're providing some accompanying information about the multimodal model.

Model Details

The CLIP model was developed by researchers at OpenAI to learn about what contributes to robustness in computer vision tasks. Th model was also developed to test the ability of models to generalize to arbitrary image classification tasks in a zero-shot manner. It was not developed for general model deployment - to deploy models like CLIP, researchers will first need to carefully study their capabilities in relation to the specific context they're being deployed within.

Model Date

January 2021

Model Type

The base model uses a ResNet50 with several modifications as an image encoder and uses a masked self-attention Transformer as a text encoder. These encoders are trained to maximize the similarity of (image, text) pairs via a contrastive loss. There is also a varian of the model where the ResNet image encoder is replaced with a Vision Transformer.

Model Version

Initially, we've released one CLIP model based on the Vision Transformer architecture equivalent to ViT-B/32, along with the RN50 model, using the architecture equivalent to ResNet-50.

As part of the staged release process, we have also released the RN101 model, as well as RN50x4, a RN50 scaled up 4x according to the EfficientNet scaling rule.

Please see the paper linked below for further details about their specification.

Documents

- Blog Post
- CLIP Paper

Model Use

Intended Use

The model is intended as a research output for research communities. We hope that this model will enable researchers to better understand and explore zero-shot, arbitrary image classification. We also hope it can be used for interdisciplinary studies of the potential impact of such models - the CLIP paper includes a discussion of potential downstream impacts to provide an example for this sort of analysis.

https://modelcards.withgoogle.com/object-detection

https://github.com/openai/CLIP/blob/main/model-card.md

Model Cards

Some models are just for research and not to be deployed. Make it clear!

Out-of-Scope Use Cases

Any deployed use case of the model - whether commercial or not - is currently out of scope. Non-deployed use cases such as image search in a constrained environment, are also not recommended unless there is thorough in-domain testing of the model with a specific, fixed class taxonomy. This is because our safety assessment demonstrated a high need for task specific testing especially given the variability of CLIP's performance with different class taxonomies. This makes untested and unconstrained deployment of the model in any use case currently potentially harmful.

Certain use cases which would fall under the domain of surveillance and facial recognition are always out-of-scope regardless of performance of the model. This is because the use of artificial intelligence for tasks such as these can be premature currently given the lack of testing norms and checks to ensure its fair use.

https://github.com/openai/CLIP/blob/main/model-card.md

Re-Examining Vision Datasets

Tiny Images Dataset: 80M images collected semi-automatically from a dictionary plus image search

Turns out it contains offensive category labels

Birhane and Prabhu, "Large Image Datasets: A Pyrrhic Win for Computer Vision?", WACV 2021

Torralba et al, "80 million tiny images: A large data set for nonparametric object and scene recognition", TPAMI 2008

Re-Examining Vision Datasets

Tiny Images dataset contains offensive category labels

June 29th, 2020

It has been brought to our attention [1] that the Tiny Images dataset contains some derogatory terms as categories and offensive images. This was a consequence of the automated data collection procedure that relied on nouns from WordNet. We are greatly concerned by this and apologize to those who may have been affected.

The dataset is too large (80 million images) and the images are so small (32 x 32 pixels) that it can be difficult for people to visually recognize its content. Therefore, manual inspection, even if feasible, will not guarantee that offensive images can be completely removed.

We therefore have decided to formally withdraw the dataset. It has been taken offline and it will not be put back online. We ask the community to refrain from using it in future and also delete any existing copies of the dataset that may have been downloaded.

Result: Tiny Images Dataset taken offline by authors

Torralba et al, “80 million tiny images: A large data set for nonparametric object and scene recognition”, TPAMI 2008

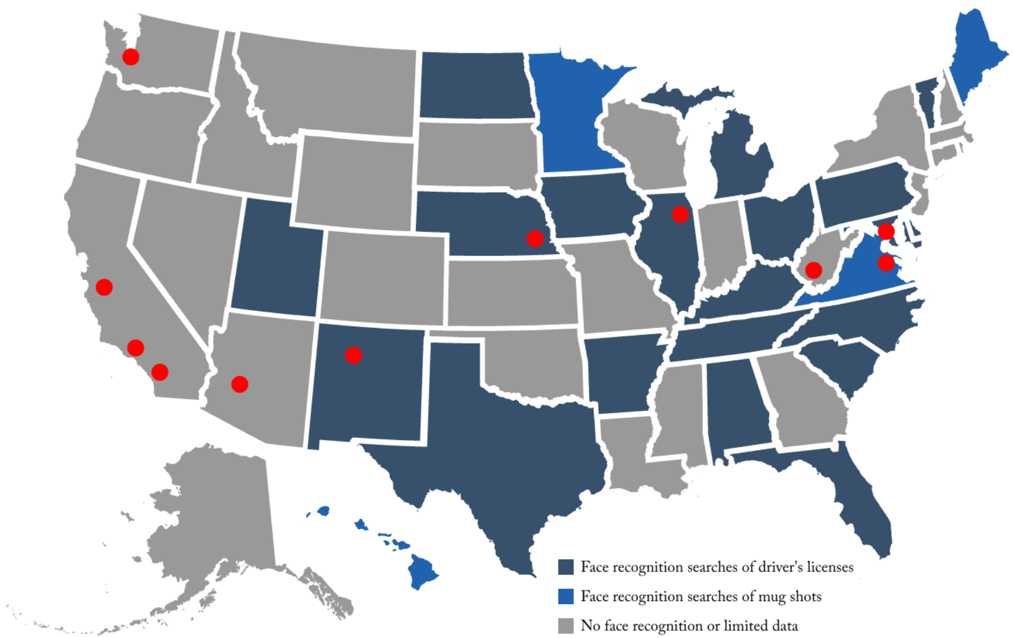
Consent vs Copyright

Image copyright != Consent to use in a dataset

Birhane and Prabhu, “Large Image Datasets: A Pyrrhic Win for Computer Vision?”, WACV 2021

Consent vs Copyright

Image copyright != Consent to use in a dataset



"One in two American adults is in a law enforcement face recognition network."

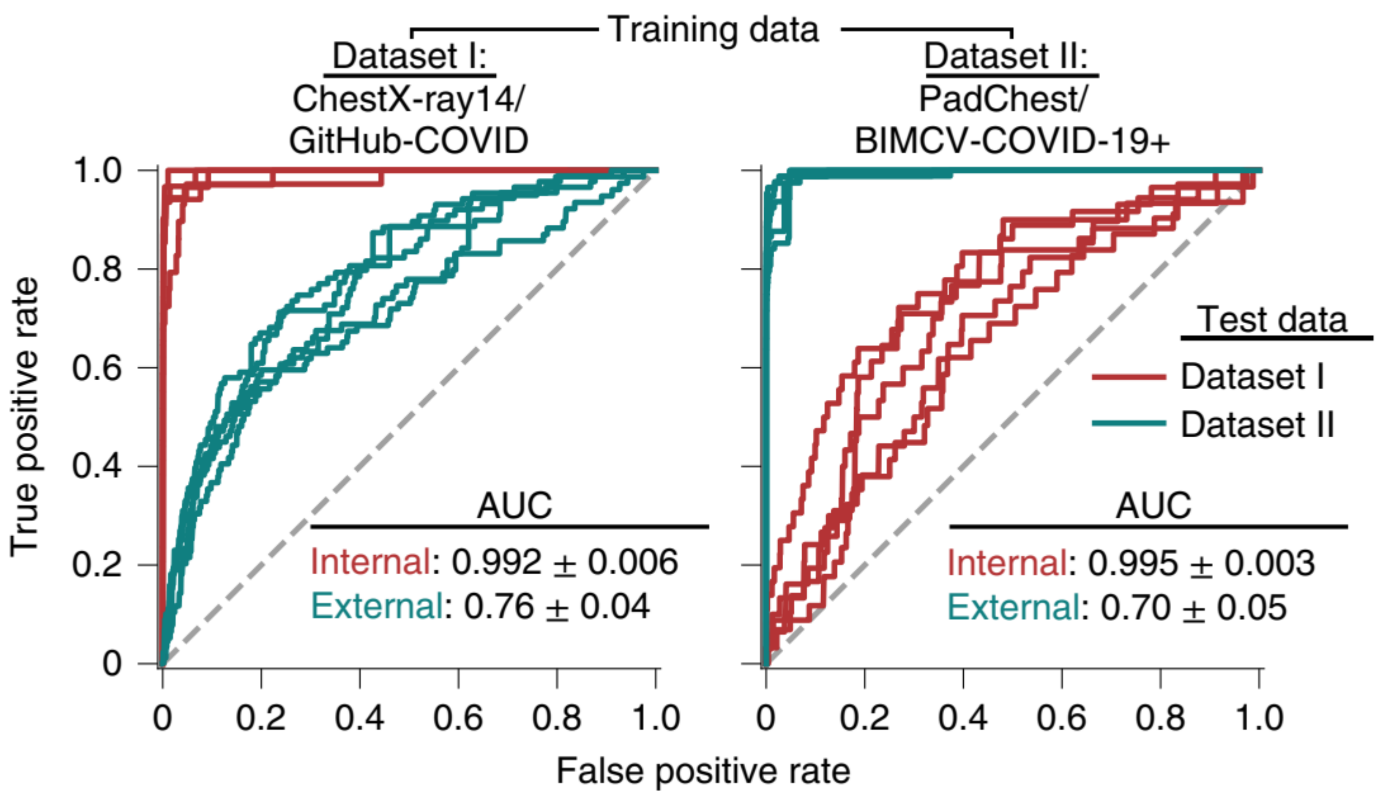
Garvie, Bedoya, and Frankle: "The Perpetual Line-Up", 2016, https://www.perpetuallineup.org/

Birhane and Prabhu, "Large Image Datasets: A Pyrrhic Win for Computer Vision?", WACV 2021

Bigger Picture

AI for radiographic COVID-19 detection selects shortcuts over signal

Alex J. DeGrave[1,2,3], Joseph D. Janizek[1,2,3] and Su-In Lee[1]



DeGrave et al. Nature Machine Intelligence, 2021.

Takeaways

- Thinking about bias and fairness in automated systems goes far beyond computer vision
- People in many fields are thinking about these issues, not just CS
- It’s important that the next generation of engineers and scientists (you all!) spend some time thinking about the implications of their work on people and society

Next Time:
AI For Science

Formalizing Fairness

Independence: $P(R, A) = P(R)P(A)$

Separation: $P(R, A \mid Y) = P(R \mid Y)P(A \mid Y)$

What happens if a binary classifier satisfies both?

Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html

Formalizing Fairness

Independence: $P(R, A) = P(R)P(A)$

Separation: $P(R, A \mid Y) = P(R \mid Y)P(A \mid Y)$

What happens if a binary classifier satisfies both?

$P(R = r \mid A = a) = P(R = r)$ (Independence)

Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html

Formalizing Fairness

Independence: $P(R, A) = P(R)P(A)$

Separation: $P(R, A \mid Y) = P(R \mid Y)P(A \mid Y)$

What happens if a binary classifier satisfies both?

$P(R = r \mid A = a) = P(R = r)$ (Independence)

(Total probability) $= \sum_y P(R = r \mid Y = y)P(Y = y)$

Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html

Formalizing Fairness

Independence: $P(R, A) = P(R)P(A)$

Separation: $P(R, A \mid Y) = P(R \mid Y)P(A \mid Y)$

What happens if a binary classifier satisfies both?

$P(R = r \mid A = a) = P(R = r)$ (Independence)

(Total probability) $= \sum_y P(R = r \mid Y = y)P(Y = y)$

(Total probability)

$P(R = r \mid A = a) = \sum_y P(R = r \mid A = a, Y = y)P(Y = y \mid A = a)$

Barocas, Hardt, and Narayanan. “Fairness and Machine Learning”, https://fairmlbook.org/index.html

Formalizing Fairness

Independence: $P(R, A) = P(R)P(A)$

Separation: $P(R, A \mid Y) = P(R \mid Y)P(A \mid Y)$

What happens if a binary classifier satisfies both?

$P(R = r \mid A = a) = P(R = r)$ (Independence)

(Total probability) $= \sum_y P(R = r \mid Y = y)P(Y = y)$

(Total probability)

$P(R = r \mid A = a) = \sum_y \boldsymbol{P(R = r \mid A = a, Y = y)} P(Y = y \mid A = a)$

(Separation) $= \sum_y \boldsymbol{P(R = r \mid Y = y)} P(Y = y \mid A = a)$

Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html

Formalizing Fairness

Independence: $P(R, A) = P(R)P(A)$

Separation: $P(R, A \mid Y) = P(R \mid Y)P(A \mid Y)$

What happens if a binary classifier satisfies both?

$P(R = r \mid A = a) = P(R = r)$ (Independence)

(Total probability) $= \sum_y \boldsymbol{P(R = r \mid Y = y)P(Y = y)}$

(Total probability)

$P(R = r \mid A = a) = \sum_y P(R = r \mid A = a, Y = y)P(Y = y \mid A = a)$

(Separation) $= \sum_y \boldsymbol{P(R = r \mid Y = y)P(Y = y \mid A = a)}$

Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html

Formalizing Fairness

Independence: $P(R, A) = P(R)P(A)$

Separation: $P(R, A \mid Y) = P(R \mid Y)P(A \mid Y)$

What happens if a binary classifier satisfies both?

For all values a of A, and all values r of R, we must have:

$$\sum_y P(R = r \mid Y = y)P(Y = y) = \sum_y P(R = r \mid Y = y)P(Y = y \mid A = a)$$

Barocas, Hardt, and Narayanan. “Fairness and Machine Learning”, https://fairmlbook.org/index.html

Formalizing Fairness

Independence: $P(R, A) = P(R)P(A)$

Separation: $P(R, A \mid Y) = P(R \mid Y)P(A \mid Y)$

What happens if a binary classifier satisfies both?

For all values a of A, and all values r of R, we must have:

$$\sum_y P(R = r \mid Y = y)P(Y = y) = \sum_y P(R = r \mid Y = y)\boldsymbol{P(Y = y \mid A = a)}$$

$$\boldsymbol{P(Y = 0 \mid A = a) = p_a}$$
$$\boldsymbol{P(Y = 1 \mid A = a) = 1 - p_a}$$

Barocas, Hardt, and Narayanan. “Fairness and Machine Learning”, https://fairmlbook.org/index.html

Formalizing Fairness

Independence: $P(R, A) = P(R)P(A)$

Separation: $P(R, A \mid Y) = P(R \mid Y)P(A \mid Y)$

What happens if a binary classifier satisfies both?

For all values a of A, and all values r of R, we must have:

$$\sum_y P(R = r \mid Y = y)\boldsymbol{P(Y = y)} = \sum_y P(R = r \mid Y = y)\boldsymbol{P(Y = y \mid A = a)}$$

$$\boldsymbol{P(Y = 0) = p} \qquad \boldsymbol{P(Y = 0 \mid A = a) = p_a}$$

$$\boldsymbol{P(Y = 1) = 1 - p} \qquad \boldsymbol{P(Y = 1 \mid A = a) = 1 - p_a}$$

Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html

Formalizing Fairness

Independence: $P(R, A) = P(R)P(A)$

Separation: $P(R, A \mid Y) = P(R \mid Y)P(A \mid Y)$

What happens if a binary classifier satisfies both?

For all values a of A, and all values r of R, we must have:

$$\sum_y \boldsymbol{P(R = r \mid Y = y)P(Y = y)} = \sum_y \boldsymbol{P(R = r \mid Y = y)P(Y = y \mid A = a)}$$

$\boldsymbol{P(R = r \mid Y = 0) = r_0}$ $\quad$ $\boldsymbol{P(Y = 0) = p}$ $\quad$ $\boldsymbol{P(Y = 0 \mid A = a) = p_a}$

$\boldsymbol{P(R = r \mid Y = 1) = r_1}$ $\quad$ $\boldsymbol{P(Y = 1) = 1 - p}$ $\quad$ $\boldsymbol{P(Y = 1 \mid A = a) = 1 - p_a}$

Barocas, Hardt, and Narayanan. “Fairness and Machine Learning”, https://fairmlbook.org/index.html

Formalizing Fairness

Independence: $P(R, A) = P(R)P(A)$

Separation: $P(R, A \mid Y) = P(R \mid Y)P(A \mid Y)$

What happens if a binary classifier satisfies both?

For all values a of A, and all values r of R, we must have:

$$\sum_y \boldsymbol{P(R = r \mid Y = y)P(Y = y)} = \sum_y \boldsymbol{P(R = r \mid Y = y)P(Y = y \mid A = a)}$$

$\boldsymbol{P(R = r \mid Y = 0) = r_0}$ $\quad$ $\boldsymbol{P(Y = 0) = p}$ $\quad$ $\boldsymbol{P(Y = 0 \mid A = a) = p_a}$

$\boldsymbol{P(R = r \mid Y = 1) = r_1}$ $\quad$ $\boldsymbol{P(Y = 1) = 1 - p}$ $\quad$ $\boldsymbol{P(Y = 1 \mid A = a) = 1 - p_a}$

$$r_0 p + r_1(1 - p) = r_0 p_a + r_1(1 - p_a)$$

$$p(r_0 - r_1) = p_a(r_0 - r_1)$$

Option 1: $r_0 = r_1$
Useless classifier!

Option 2: $p = p_a$
Target, attribute are independent

Barocas, Hardt, and Narayanan. "Fairness and Machine Learning", https://fairmlbook.org/index.html