Fighting Bias in AI: Can Ordinary People Make a Difference?

Tübingen Days of Digital Freedom 2025

Omri Ben-Dov

ACT I

AI: Artificial Intelligence

Or: Intro to machine learning

Is this person wearing glasses?

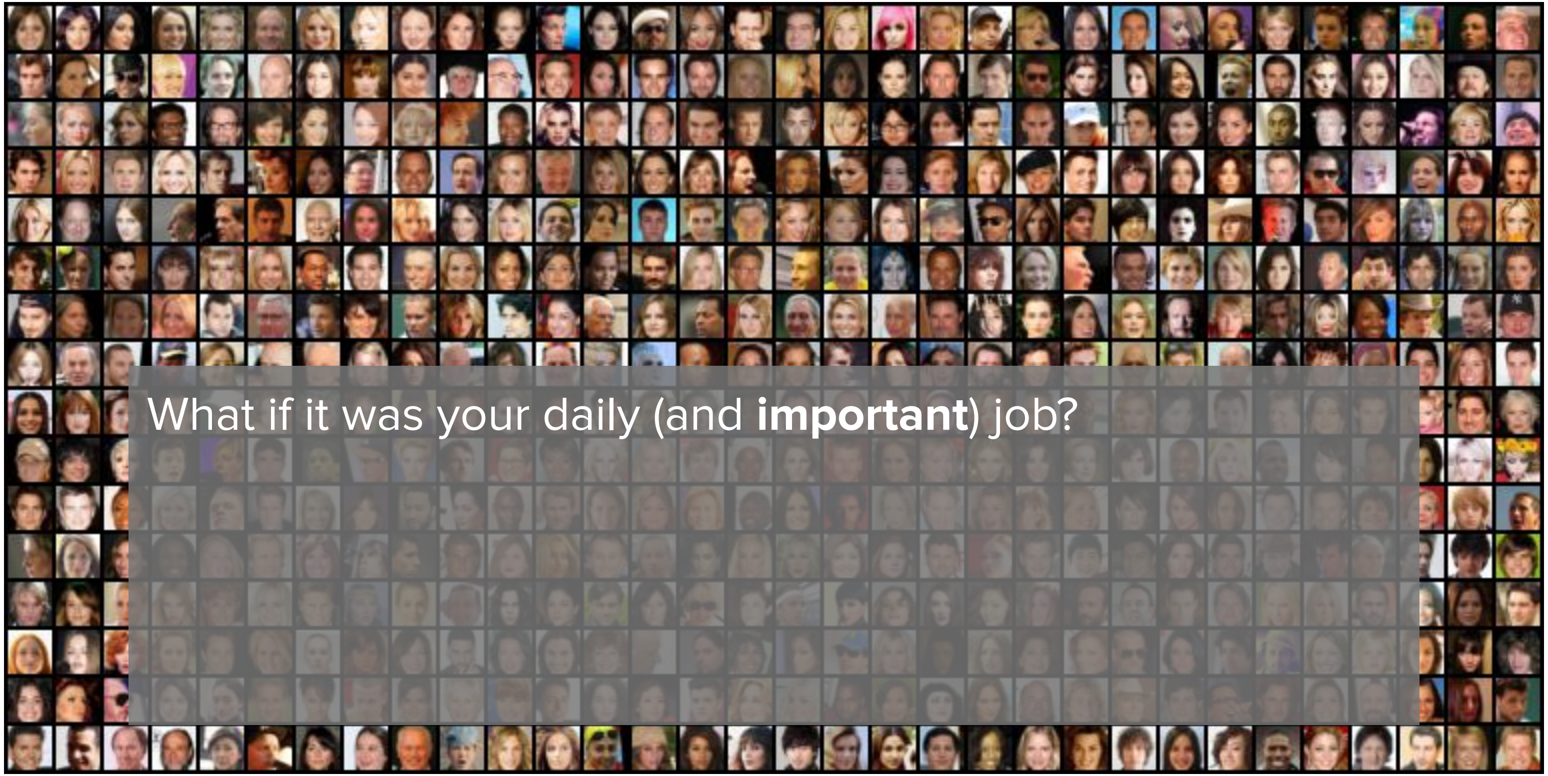
Is this person smiling?

Does this person have short hair?

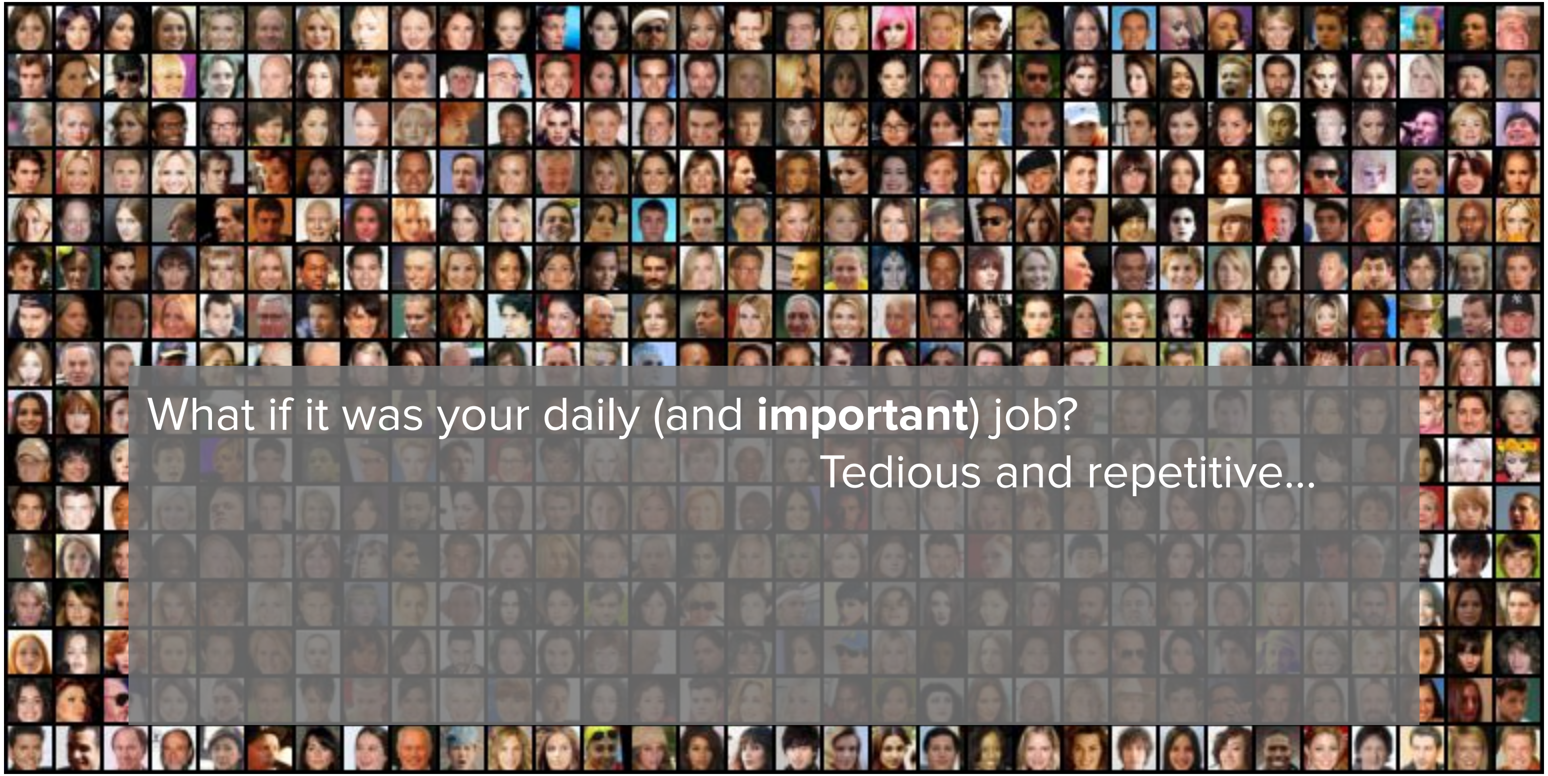
Is this person good looking?

Ready for hundreds of more rounds?

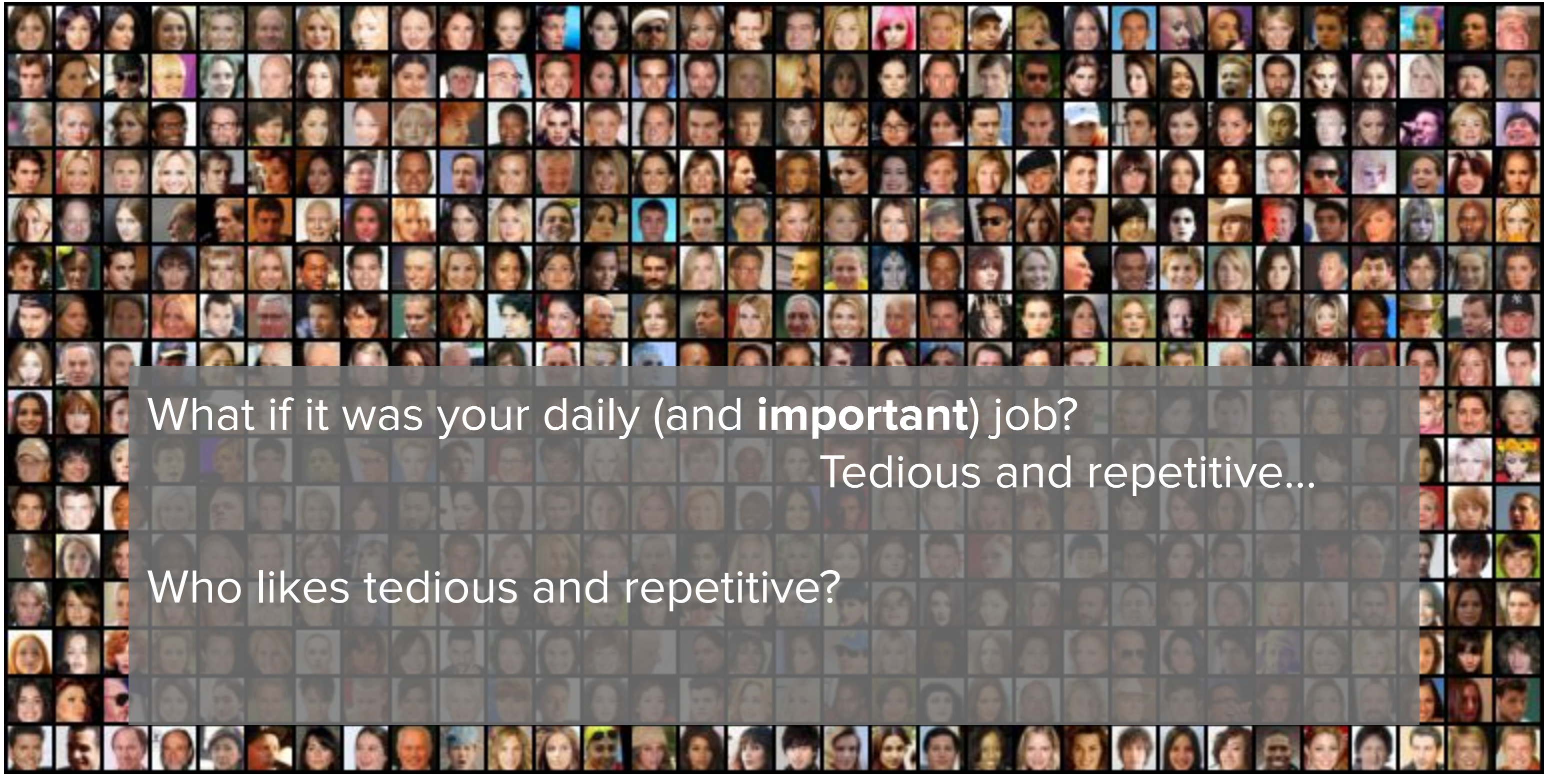
Ready for hundreds of more rounds?



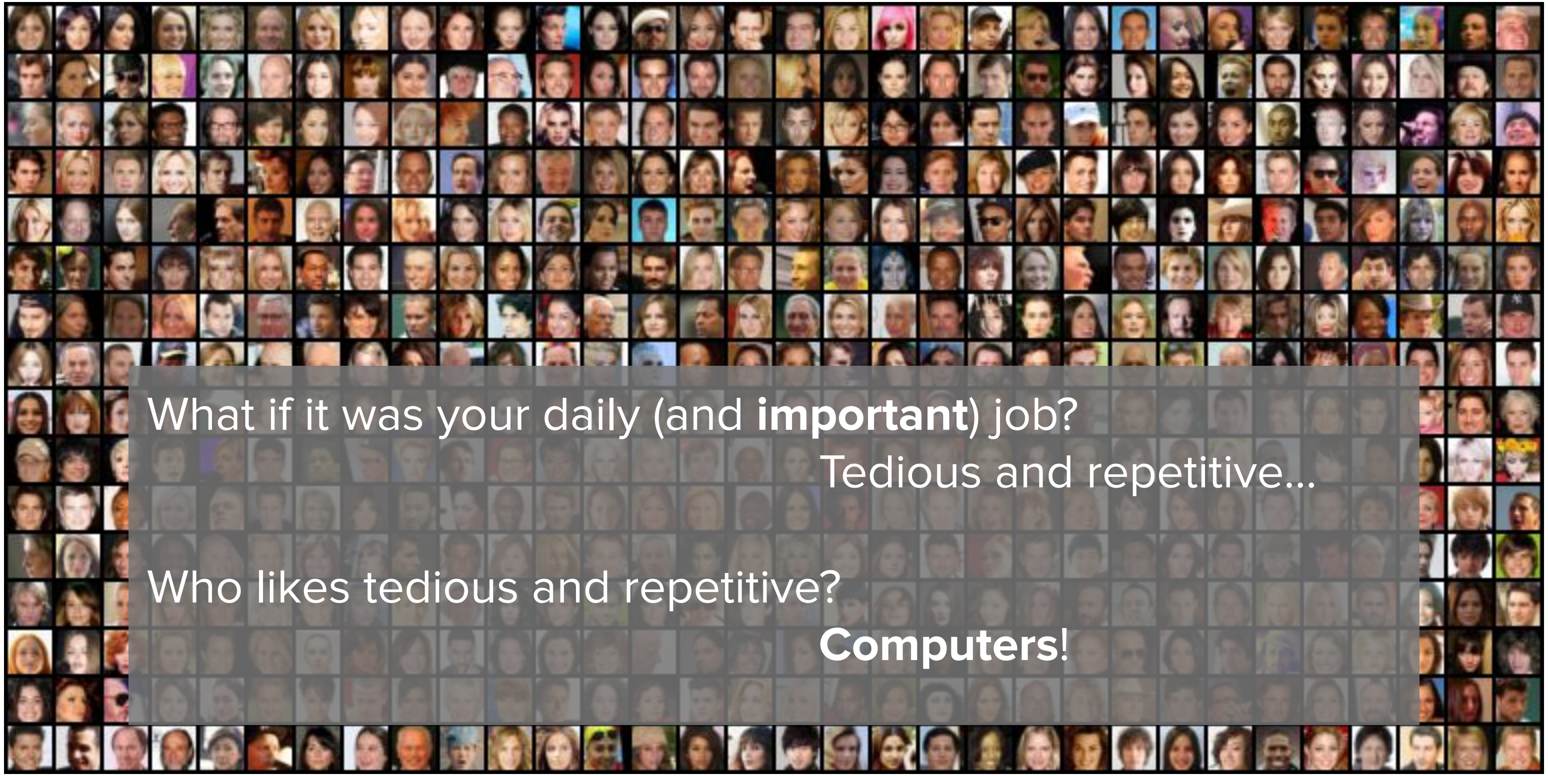
Ready for hundreds of more rounds?



Ready for hundreds of more rounds?



Ready for hundreds of more rounds?



Can we explain the task to a computer?

How do you tell a computer what glasses are?

What if the question is ambiguous?

Can we explain the task to a computer?

How do you tell a computer what glasses are?

What if the question is ambiguous?

We were not born with knowledge of glasses, but learned it...

Can a computer learn to answer a single yes/no question about different people?

Can we explain the task to a computer?

How do you tell a computer what glasses are?

What if the question is ambiguous?

We were not born with knowledge of glasses, but learned it...

Can a computer learn to answer a single yes/no question about different people?

Machine learning:

Computers answer a single question about new cases by utilizing old cases

Can we explain the task to a computer?

How do you tell a computer what glasses are?

What if the question is ambiguous?

We were not born with knowledge of glasses, but learned it...

Can a computer learn to answer a single yes/no question about different people?

Machine learning:

Computers answer a single question about new cases by utilizing old cases

Yes, even ChatGPT answers a single question

Basic Machine Learning - “Is the person wearing glasses?”

The data:
People and the corresponding answer


yes


no

Basic Machine Learning - “Is the person wearing glasses?”

The data:
People and the corresponding answer


yes


no

X 1000 (or some other large number)

Basic Machine Learning - “Is the person wearing glasses?”

The data:
People and the
corresponding answer


yes


no

X 1000 (or some other large number)

Learning

The algorithm:
Turn the data into a machine
that can answer

Basic Machine Learning - “Is the person wearing glasses?”

The data:
People and the corresponding answer


yes

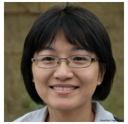

no

X 1000 (or some other large number)

Learning

The algorithm:
Turn the data into a machine that can answer

Answering


yes ✓


no ✓

Basic Machine Learning - “Is the person wearing glasses?”

The data:
People and the corresponding answer


yes


no

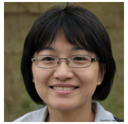
X 1000 (or some other large number)

Learning

The algorithm:
Turn the data into a machine that can answer

Answering

The “algorithm” (inference):
Answer about new people


yes


no

Is the machine always correct?

NO!

Is the machine always correct?

NO!

Data quality

yes

no

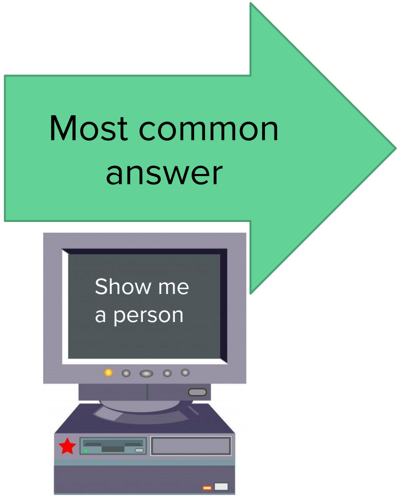
no

Is the machine always correct?

NO!

Data quality

Learning algorithm



Is the machine always correct?

NO!

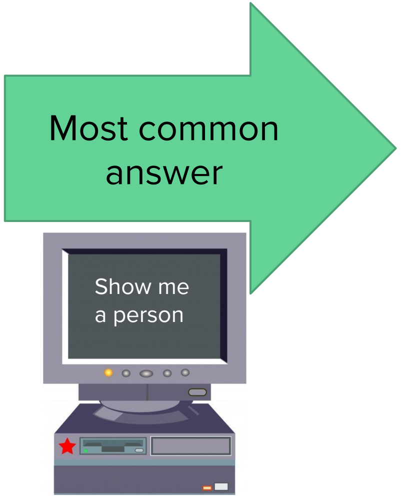
Data quality

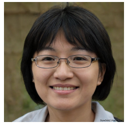
yes

no

no

Learning algorithm

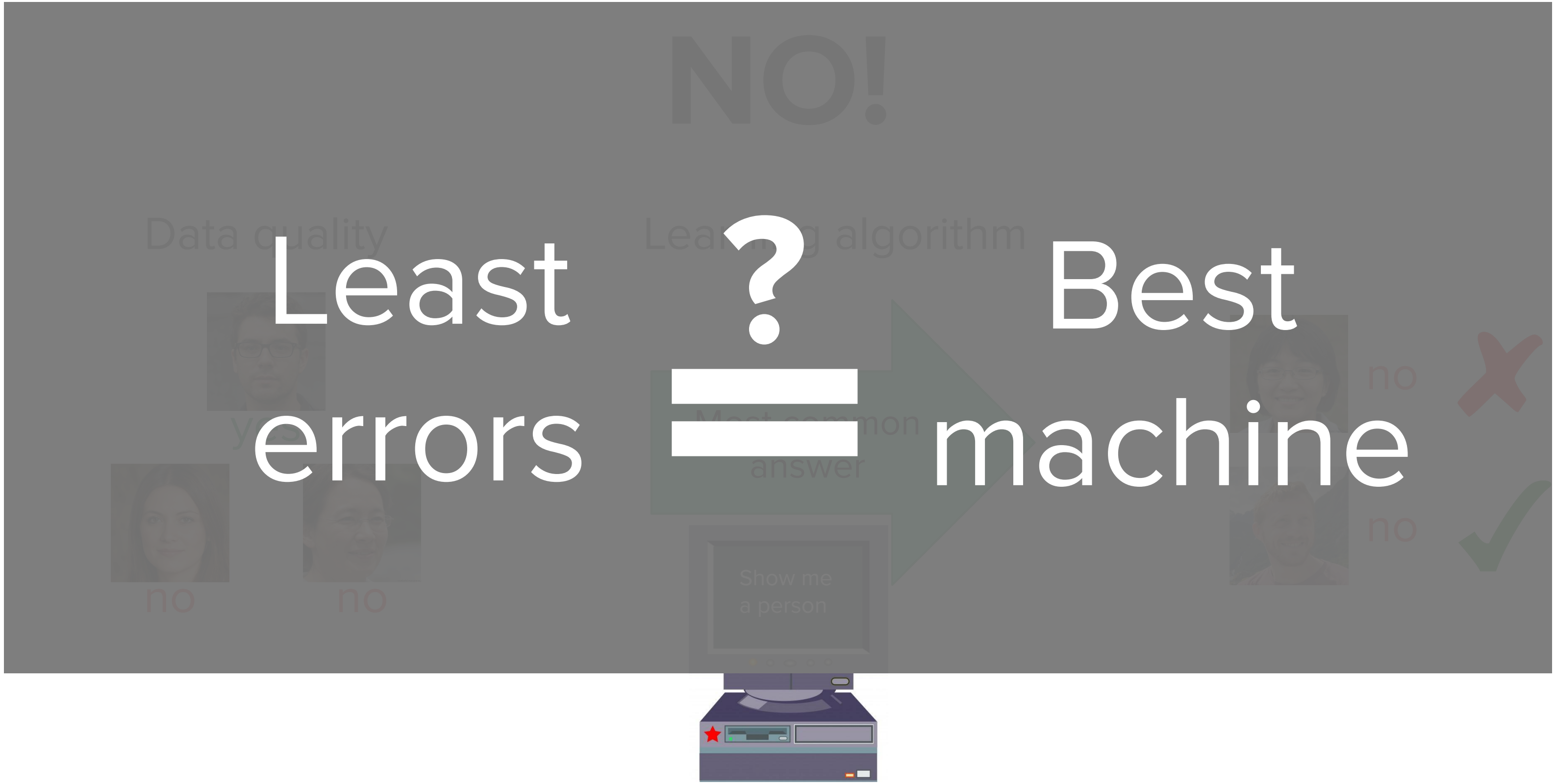




no

no

Is the machine always correct?



ACT II

Bias in AI

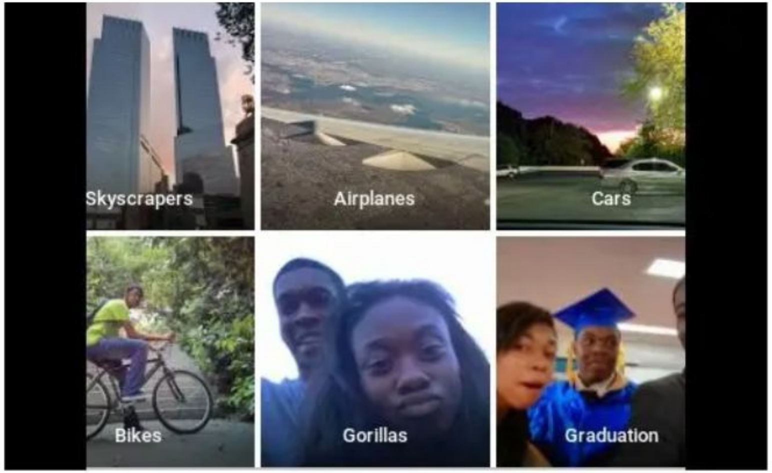
Or: The Paradox of Fairness in Machine Learning

How good is the learning machine?

Do we want the machine to answer correctly for most people?

How good is the learning machine?

Do we want the machine to answer correctly for most people?



diri noir avec banan @jackyalcine · Jun 29
Google Photos, y'all [redacted] My friend's not a gorilla.

Forbes

Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software

By Maggie Zhang, Forbes Staff. I write about technology, innovation, and startups.

Published Jul 01, 2015, 01:42pm EDT, Updated Jul 01, 2015, 02:35pm EDT

The Guardian

A beauty contest was judged by AI and the robots didn't like dark skin

Sam Levin in San Francisco

Fri 9 Sep 2016 00.42 CEST

How good is the learning machine?

Do we want the machine to answer correctly for most people?

Reuters

Insight - Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

October 11, 2018 2:50 AM GMT+2 · Updated October 10, 2018

Forbes

Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software

By Maggie Zhang, Forbes Staff. I write about technology, innovation, and startups.

Published Jul 01, 2015, 01:42pm EDT, Updated Jul 01, 2015, 02:35pm EDT

AP

The secret bias hidden in mortgage-approval algorithms

BY EMMANUEL MARTINEZ AND LAUREN KIRCHNER/THE MARKUP

Published 6:04 PM GMT+2, August 25, 2021

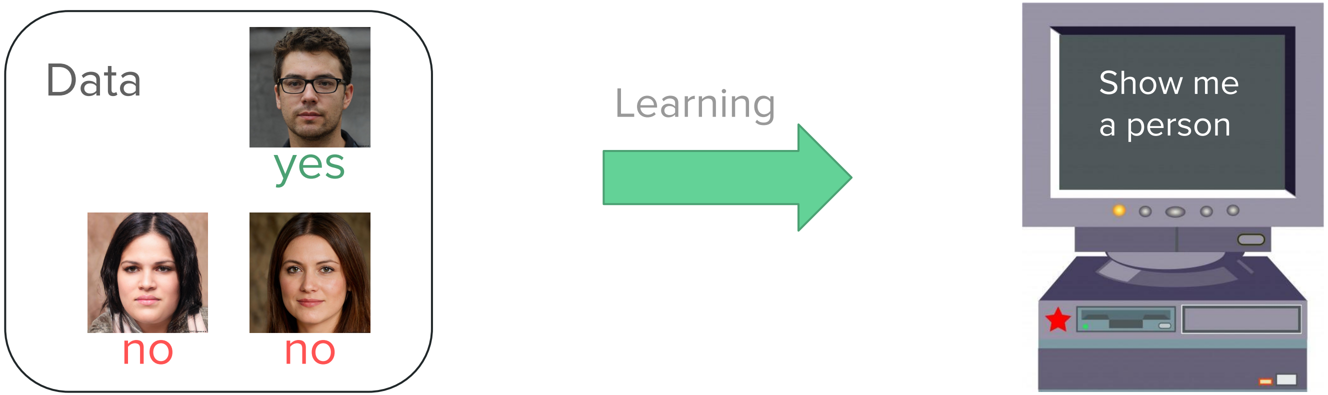
The Guardian

A beauty contest was judged by AI and the robots didn't like dark skin

Sam Levin in San Francisco

Fri 9 Sep 2016 00.42 CEST

Sources of unfairness

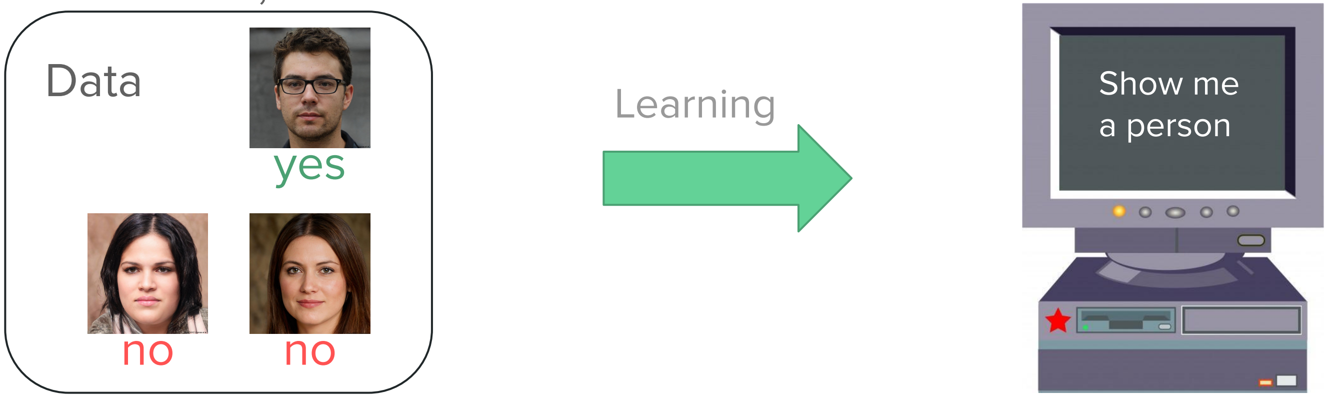


Sources of unfairness

The data:

Is there bias in the data?

If the world is biased, the machine will be biased

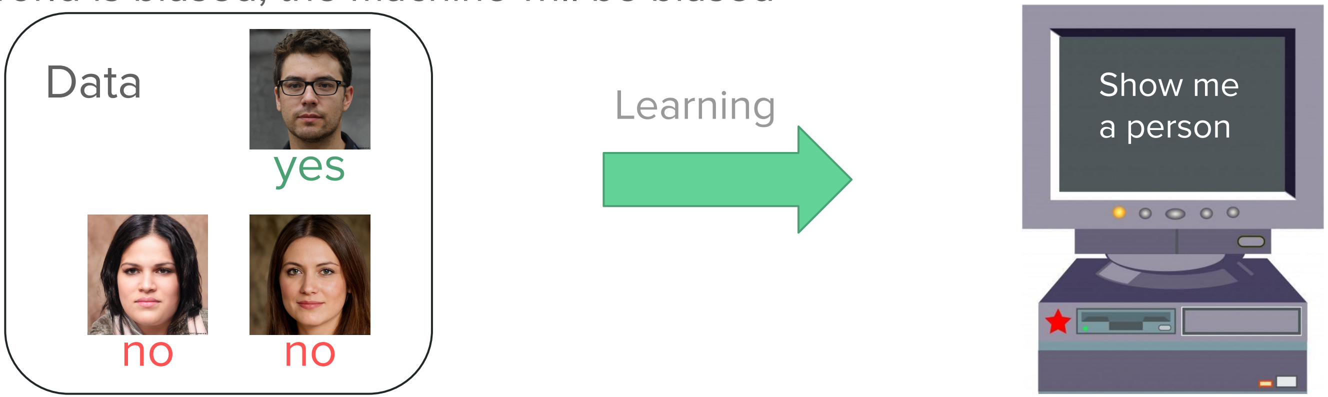


Sources of unfairness

The data:

Is there bias in the data?

If the world is biased, the machine will be biased



The algorithm:

If we learn to answer correctly for the majority of people, it may be wrong for the minority

What is unfairness?

In computer science, we need numbers to tell us how unfair a machine is.

What is unfairness?

In computer science, we need numbers to tell us how unfair a machine is.

- **Positive predicted value (PPV)**: the fraction of positive cases which were correctly predicted out of all the positive predictions. It is usually referred to as precision, and represents the probability of a correct positive prediction. It is given by the following formula:

$$PPV = P(actual = + \mid prediction = +) = \frac{TP}{TP+FP}$$

- **False discovery rate (FDR)**: the fraction of positive predictions which were actually negative out of all the positive predictions. It represents the probability of an erroneous positive prediction, and it is given by the following formula:

$$FDR = P(actual = - \mid prediction = +) = \frac{FP}{TP+FP}$$

- **Negative predicted value (NPV)**: the fraction of negative cases which were correctly predicted out of all the negative predictions. It represents the probability of a correct negative prediction, and it is given by the following formula:

$$NPV = P(actual = - \mid prediction = -) = \frac{TN}{TN+FN}$$

- **False omission rate (FOR)**: the fraction of negative predictions which were actually positive out of all the negative predictions. It represents the probability of an erroneous negative prediction, and it is given by the following formula:

$$FOR = P(actual = + \mid prediction = -) = \frac{FN}{TN+FN}$$

- **True positive rate (TPR)**: the fraction of positive cases which were correctly predicted out of all the positive cases. It is usually referred to as sensitivity or recall, and it represents the probability of the positive subjects to be classified correctly as such. It is given by the formula:

$$TPR = P(prediction = + \mid actual = +) = \frac{TP}{TP+FN}$$

- **False negative rate (FNR)**: the fraction of positive cases which were incorrectly predicted to be negative out of all the positive cases. It represents the probability of the positive subjects to be classified incorrectly as negative ones, and it is given by the formula:

$$FNR = P(prediction = - \mid actual = +) = \frac{FN}{TP+FN}$$

- **True negative rate (TNR)**: the fraction of negative cases which were correctly predicted out of all the negative cases. It represents the probability of the negative subjects to be classified correctly as such, and it is given by the formula:

$$TNR = P(prediction = - \mid actual = -) = \frac{TN}{TN+FP}$$

- **False positive rate (FPR)**: the fraction of negative cases which were incorrectly predicted to be positive out of all the negative cases. It represents the probability of the negative subjects to be classified incorrectly as positive ones, and it is given by the formula:

$$FPR = P(prediction = + \mid actual = -) = \frac{FP}{TN+FP}$$

Definitions based on predicted outcome [edit]

The definitions in this section focus on a predicted outcome R for various distributions of subjects. They are the simplest and most intuitive notions of fairness.

- **Demographic parity**, also referred to as **statistical parity**, **acceptance rate parity** and **benchmarking**. A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal probability of being assigned to the positive predicted class. This is, if the following formula is satisfied:

$$P(R = + \mid A = a) = P(R = + \mid A = b) \quad \forall a, b \in A$$

- **Conditional statistical parity**. Basically consists in the definition above, but restricted only to a subset of the instances. In mathematical notation this would be:

$$P(R = + \mid L = l, A = a) = P(R = + \mid L = l, A = b) \quad \forall a, b \in A \quad \forall l \in L$$

Definitions based on predicted probabilities and actual outcome [edit]

These definitions are based in the actual outcome Y and the predicted probability score S.

- **Test-fairness**, also known as **calibration** or **matching conditional frequencies**. A classifier satisfies this definition if individuals with the same predicted probability score S have the same probability of being classified in the positive class when they belong to either the protected or the unprotected group:

$$P(Y = + \mid S = s, A = a) = P(Y = + \mid S = s, A = b) \quad \forall s \in S \quad \forall a, b \in A$$

- **Well-calibration** is an extension of the previous definition. It states that when individuals inside or outside the protected group have the same predicted probability score S they must have the same probability of being classified in the positive class, and this probability must be equal to S:

$$P(Y = + \mid S = s, A = a) = P(Y = + \mid S = s, A = b) = s \quad \forall s \in S \quad \forall a, b \in A$$

- **Balance for positive class**. A classifier satisfies this definition if the subjects constituting the positive class from both protected and unprotected groups have equal average predicted probability score S. This means that the expected value of probability score for the protected and unprotected groups with positive actual outcome Y is the same, satisfying the formula:

$$E(S \mid Y = +, A = a) = E(S \mid Y = +, A = b) \quad \forall a, b \in A$$

- **Balance for negative class**. A classifier satisfies this definition if the subjects constituting the negative class from both protected and unprotected groups have equal average predicted probability score S. This means that the expected value of probability score for the protected and unprotected groups with negative actual outcome Y is the same, satisfying the formula:

$$E(S \mid Y = -, A = a) = E(S \mid Y = -, A = b) \quad \forall a, b \in A$$

Definitions based on predicted and actual outcomes [edit]

These definitions not only considers the predicted outcome R but also compare it to the actual outcome Y.

- **Predictive parity**, also referred to as **outcome test**. A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal PPV. This is, if the following formula is satisfied:

$$P(Y = + \mid R = +, A = a) = P(Y = + \mid R = +, A = b) \quad \forall a, b \in A$$

Mathematically, if a classifier has equal PPV for both groups, it will also have equal FDR, satisfying the formula:

$$P(Y = - \mid R = +, A = a) = P(Y = - \mid R = +, A = b) \quad \forall a, b \in A$$

- **False positive error rate balance**, also referred to as **predictive equality**. A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal FPR. This is, if the following formula is satisfied:

$$P(R = + \mid Y = -, A = a) = P(R = + \mid Y = -, A = b) \quad \forall a, b \in A$$

Mathematically, if a classifier has equal FPR for both groups, it will also have equal TNR, satisfying the formula:

$$P(R = - \mid Y = -, A = a) = P(R = - \mid Y = -, A = b) \quad \forall a, b \in A$$

- **False negative error rate balance**, also referred to as **equal opportunity**. A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal FNR. This is, if the following formula is satisfied:

$$P(R = - \mid Y = +, A = a) = P(R = - \mid Y = +, A = b) \quad \forall a, b \in A$$

Mathematically, if a classifier has equal FNR for both groups, it will also have equal TPR, satisfying the formula:

$$P(R = + \mid Y = +, A = a) = P(R = + \mid Y = +, A = b) \quad \forall a, b \in A$$

- Equalized odds, also referred to as **conditional procedure accuracy equality** and **disparate mistreatment**. A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal TPR and equal FPR, satisfying the formula:

$$P(R = + \mid Y = y, A = a) = P(R = + \mid Y = y, A = b) \quad y \in \{+, -\} \quad \forall a, b \in A$$

- **Conditional use accuracy equality**. A classifier satisfies this definition if the subjects in the protected and unprotected groups have equal PPV and equal NPV, satisfying the formula:

$$P(Y = y \mid R = y, A = a) = P(Y = y \mid R = y, A = b) \quad y \in \{+, -\} \quad \forall a, b \in A$$

- **Overall accuracy equality**. A classifier satisfies this definition if the subject in the protected and unprotected groups have equal prediction accuracy, that is, the probability of a subject from one class to be assigned to it. This is, if it satisfies the following formula:

$$P(R = Y \mid A = a) = P(R = Y \mid A = b) \quad \forall a, b \in A$$

- **Treatment equality**. A classifier satisfies this definition if the subjects in the protected and unprotected groups have an equal ratio of FN and FP, satisfying the formula:

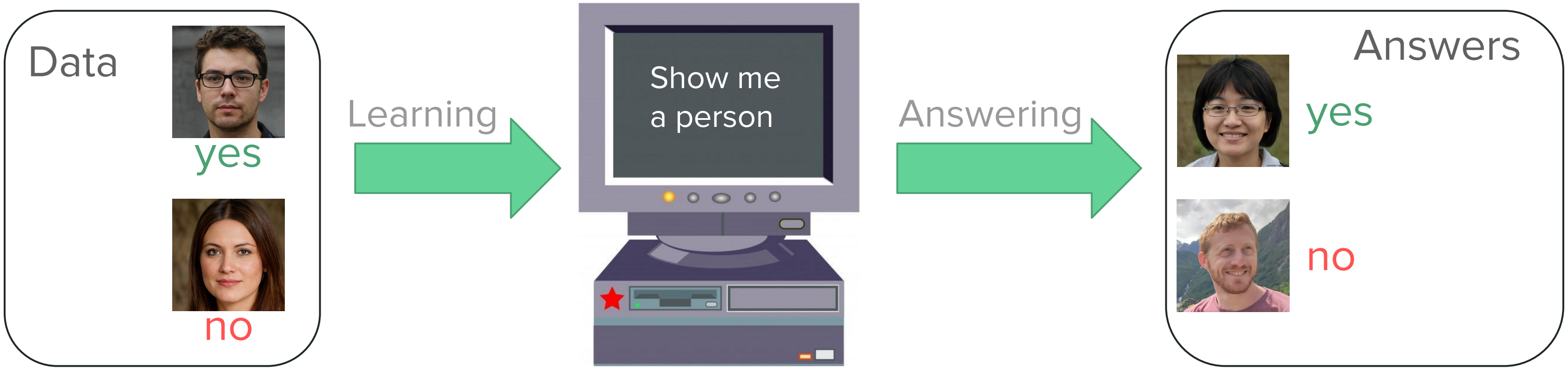
$$\frac{FN_{A=a}}{FP_{A=a}} = \frac{FN_{A=b}}{FP_{A=b}}$$

What is unfairness?

In computer science, we need numbers to tell us how unfair a machine is.

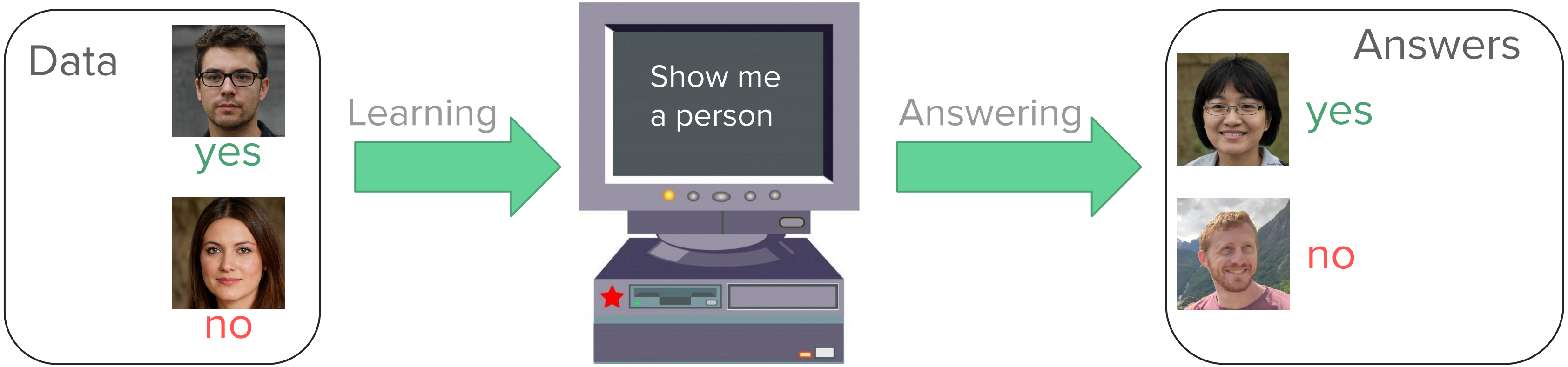
Unfairness = Unequal treatment of <u>majority</u> and <u>minority</u>

How to make machine learning fair?



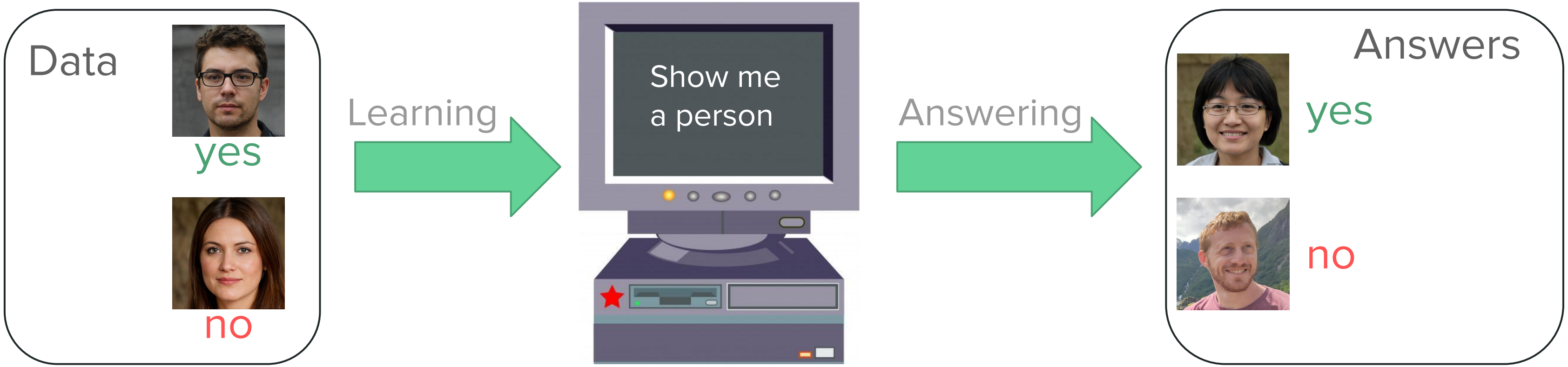
How to make machine learning fair?

1) Pre-processing the data:
Change the data to be fair in the first place



How to make machine learning fair?

1) Pre-processing the data:
Change the data to be fair in the first place

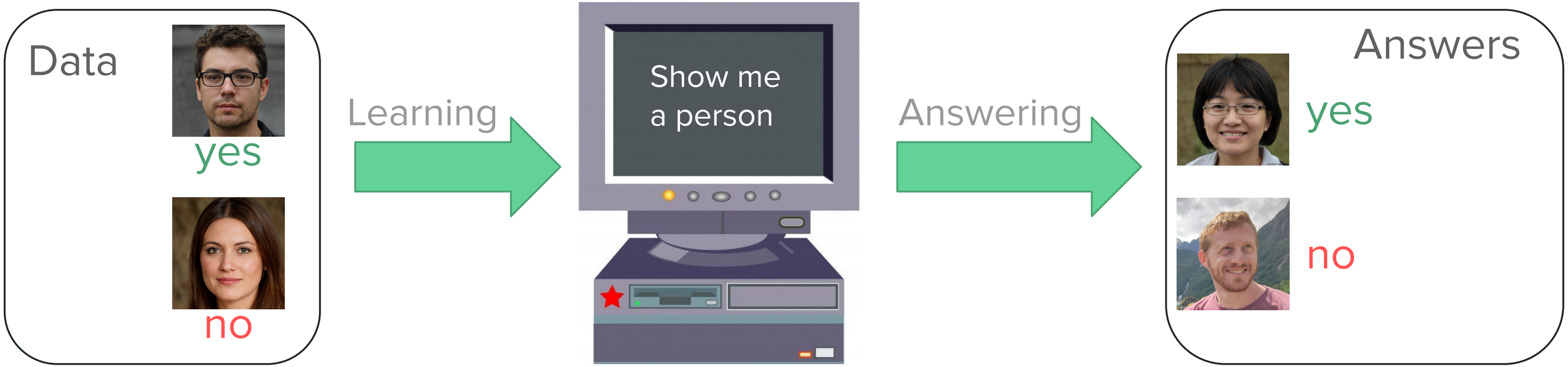


2) In-processing the algorithm:
Change the algorithm to be fair to minorities

How to make machine learning fair?

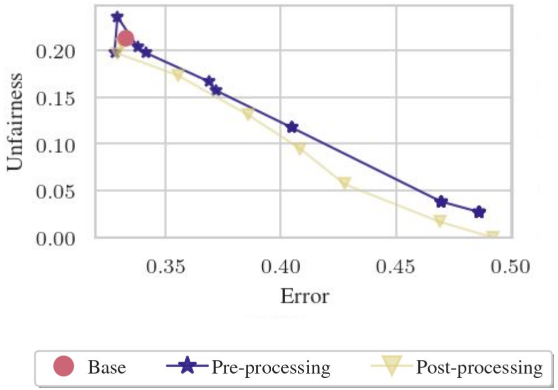
1) Pre-processing the data:
Change the data to be fair in the first place

3) Post-processing the answers:
Change the answers to be fair

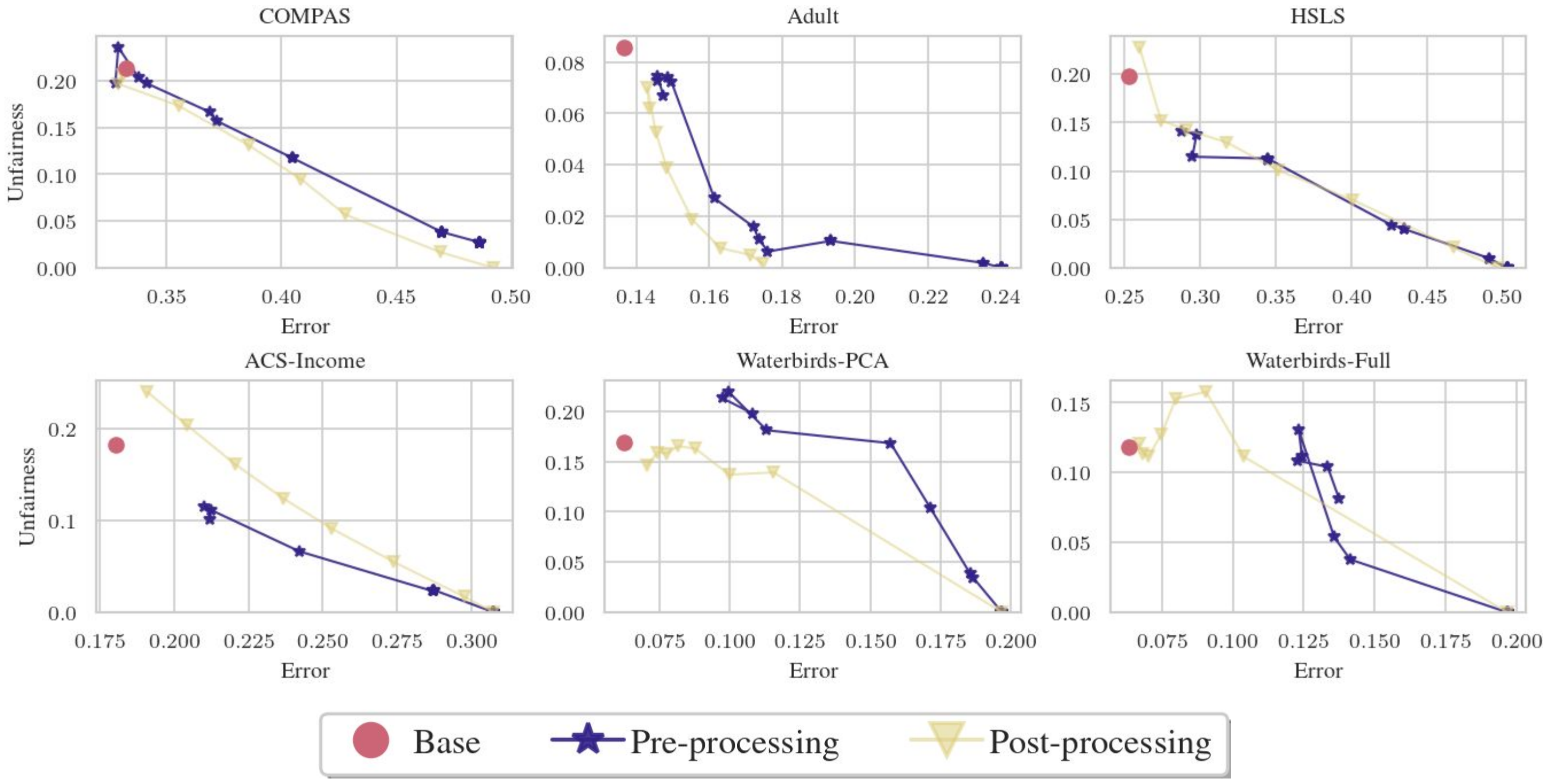


2) In-processing the algorithm:
Change the algorithm to be fair to minorities

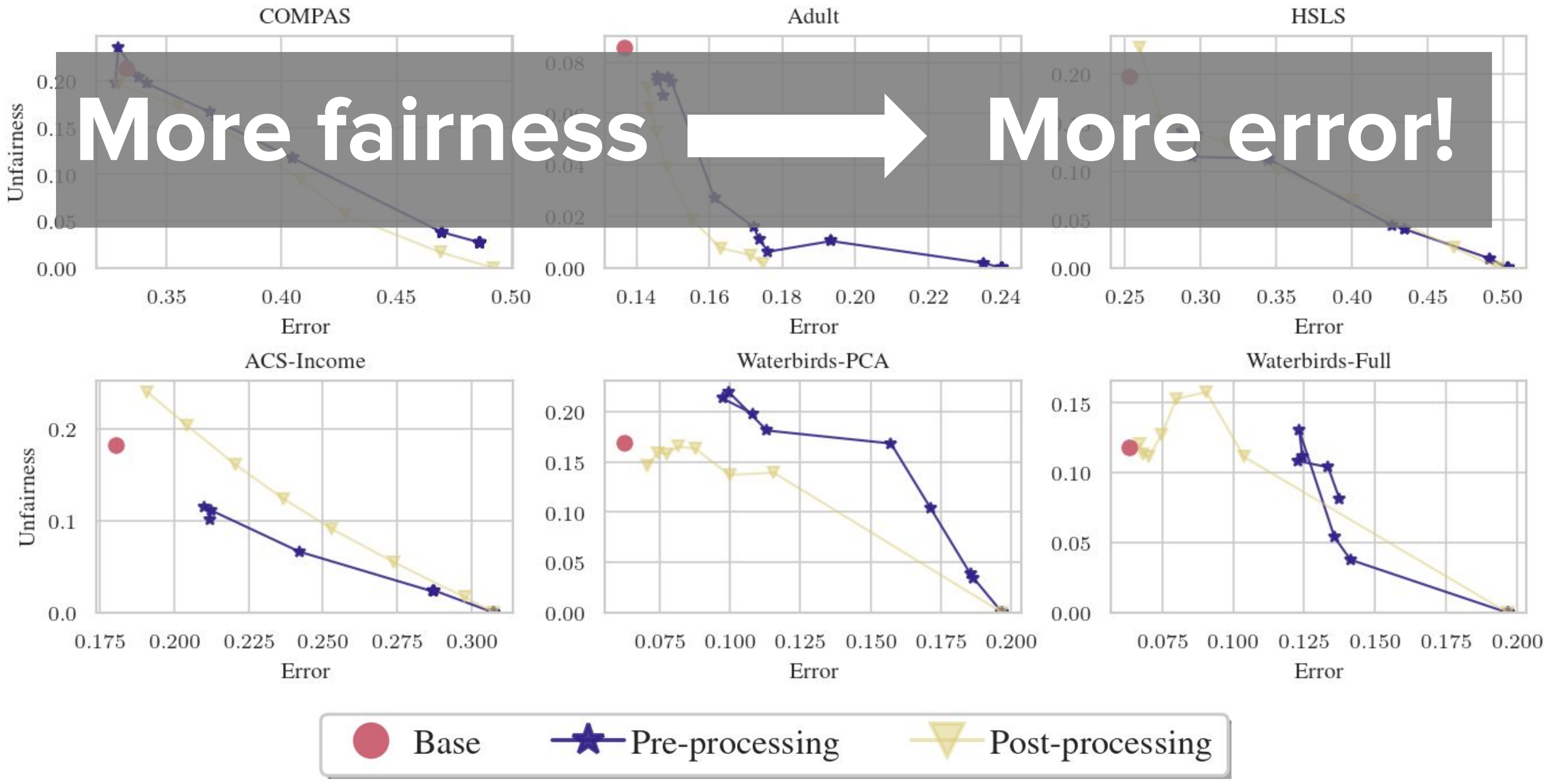
The cost of fairness is error



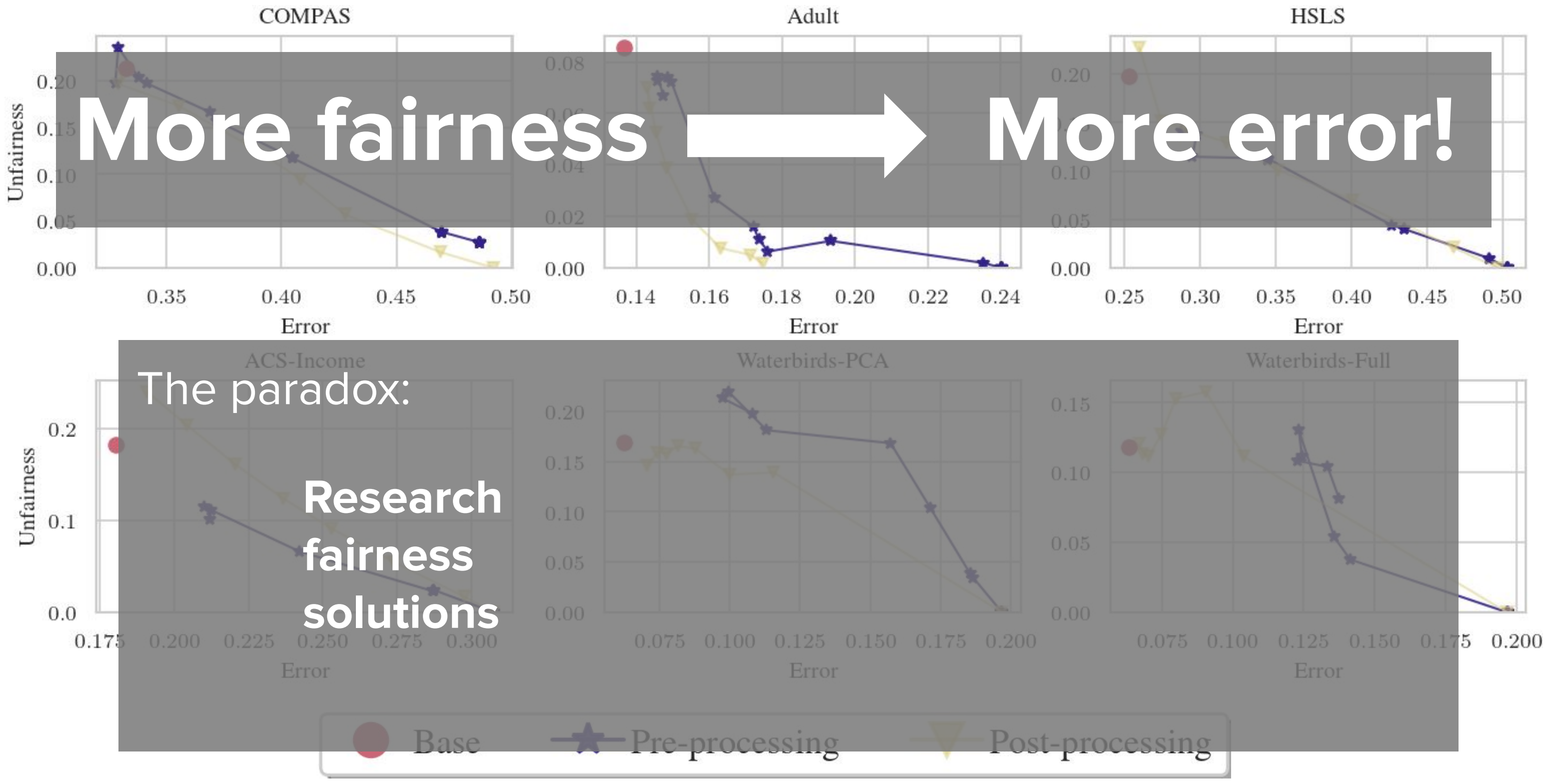
The cost and the paradox



The cost and the paradox



The cost and the paradox

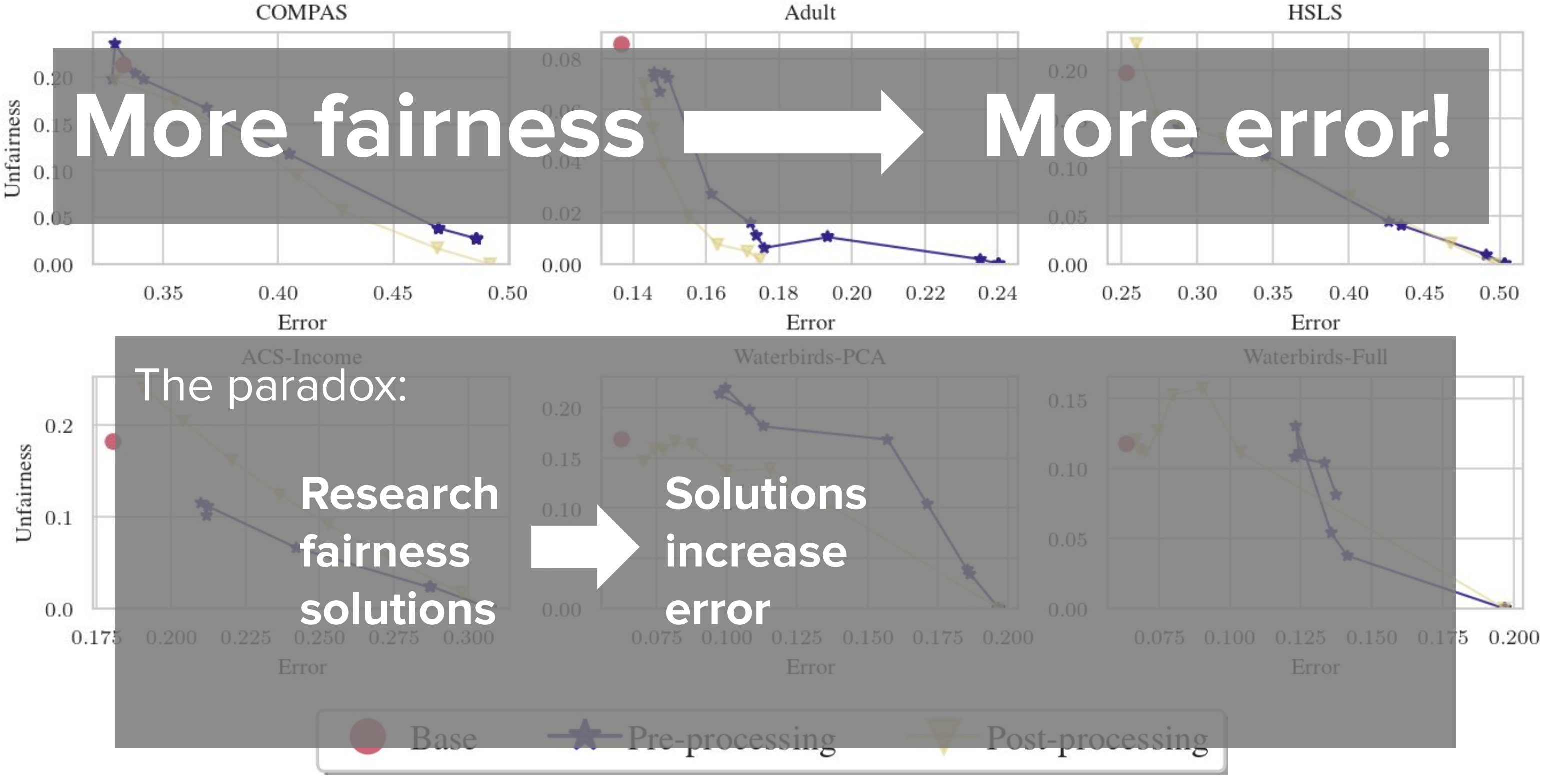


More fairness → More error!

The paradox:

Research fairness solutions

The cost and the paradox

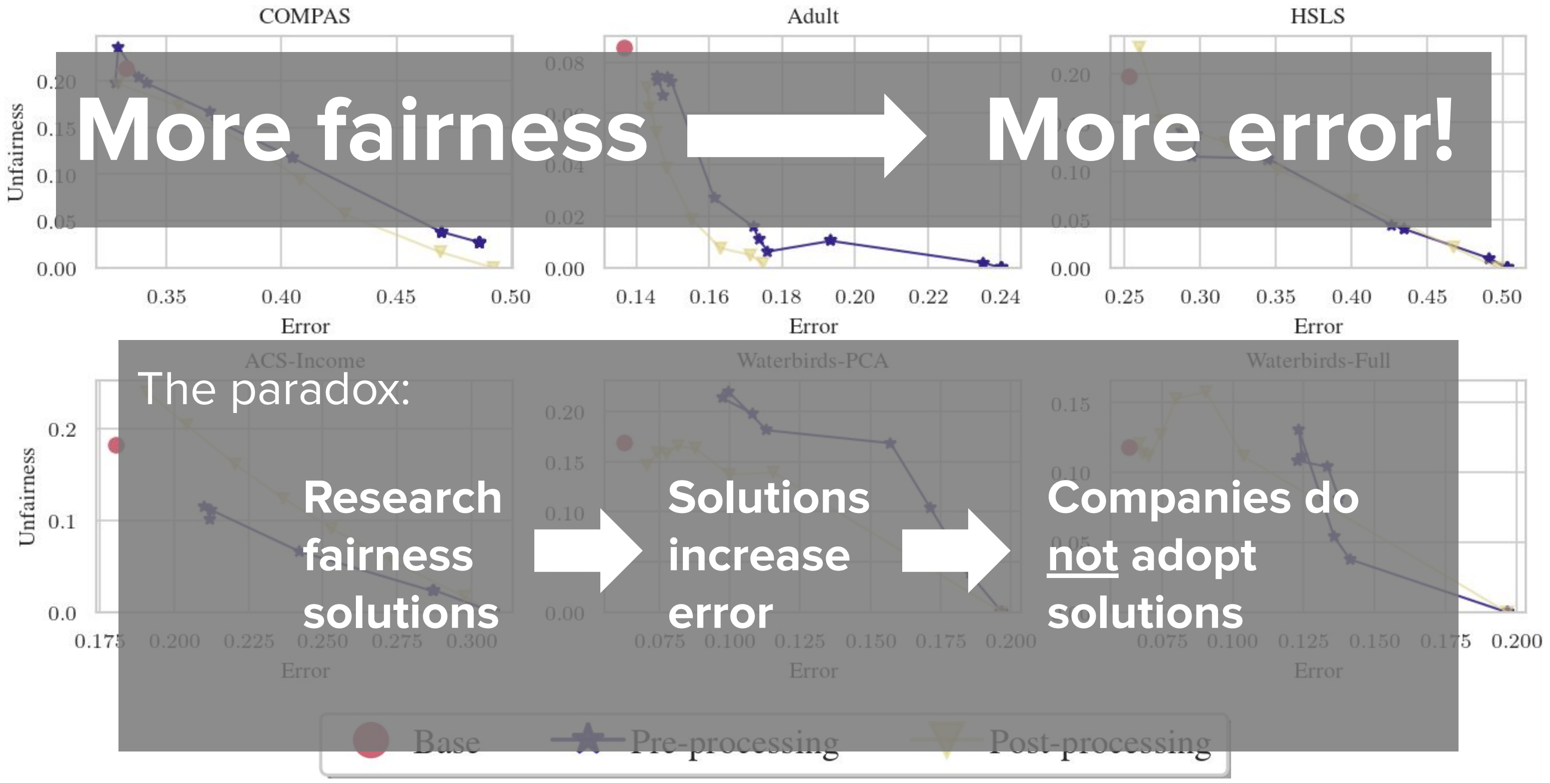


More fairness → More error!

The paradox:

Research fairness solutions → Solutions increase error

The cost and the paradox



More fairness → More error!

The paradox:

Research fairness solutions → Solutions increase error → Companies do <u>not</u> adopt solutions

Give up on fairness?

If companies don't use fair learning,

what can we do?

ACT III

Fighting Bias in AI: Can Ordinary People Make a Difference?

Or: Collective Action in Machine Learning

The information age

The information age

They all collect our data:

- Information
- Clicks
- History
- Likes
- Time
- Anything

And use machine learning

The information age

They all collect **our data**:

- **Information**
- **Clicks**
- **History**
- **Likes**
- **Time**
- **Anything**

And use machine learning

One person is not enough, but together...

INDEPENDENT

News > Business > Business News

Uber drivers work together to create price surge and charge customers more, researchers find

Some drivers are deliberately going offline in unison so that prices surge and they can charge customers more when they log back into the app

Waze to go: residents fight off crowdsourced traffic... for a while

Residents on a formerly quiet street tried reporting bogus blockades, but it hasn't worked to stem the crowdsourced traffic tide.

Written by Lisa Vaas

JUNE 07, 2016

CNET

Your quote tweets make bad tweets worse. Do this instead

Critics denouncing Parkland shooting conspiracy theories ended up fanning the flames on social media. Here's how you can criticize without amplifying.

Thrown under the bus and outrunning it! The logic of Didi and taxi drivers' labour and activism in the on-demand economy

Julie Yujie Chen View all authors and affiliations

Volume 20, Issue 8 | https://doi.org/10.1177/1461444817729149

On-demand workers are protesting – using the apps they work for

By Sara Ashley O'Brien, CNN Business

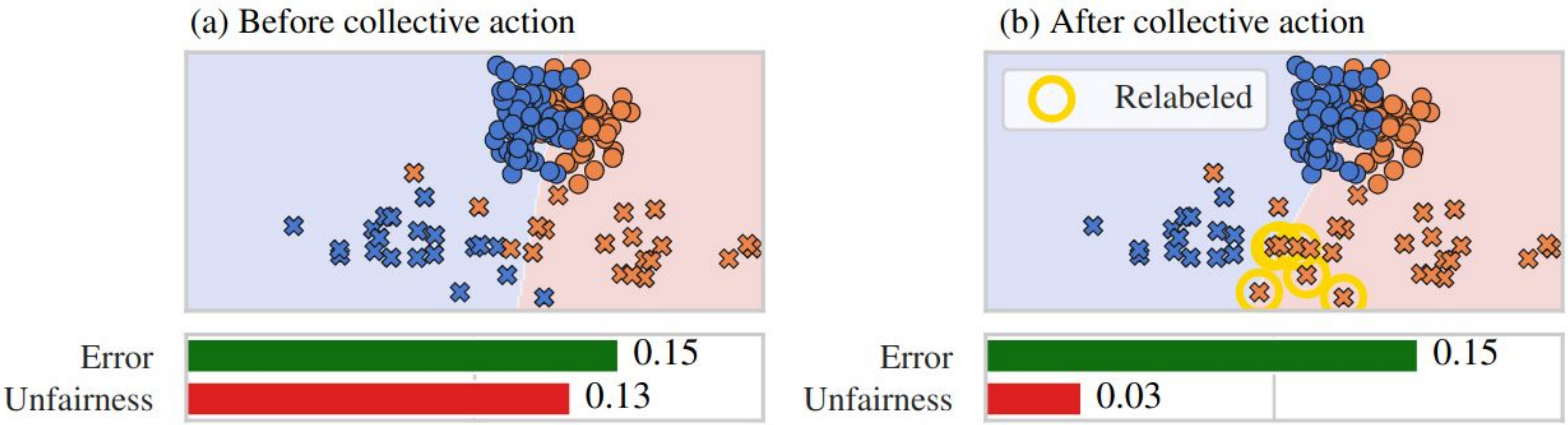
5 min read · Published 2:39 PM EDT, Fri June 14, 2019

How Anitta megafans gamed Spotify to help create Brazil's first global chart-topper

Fans skirted terms and conditions, but a strategy from the singer's team came close to the streaming platform's limit.

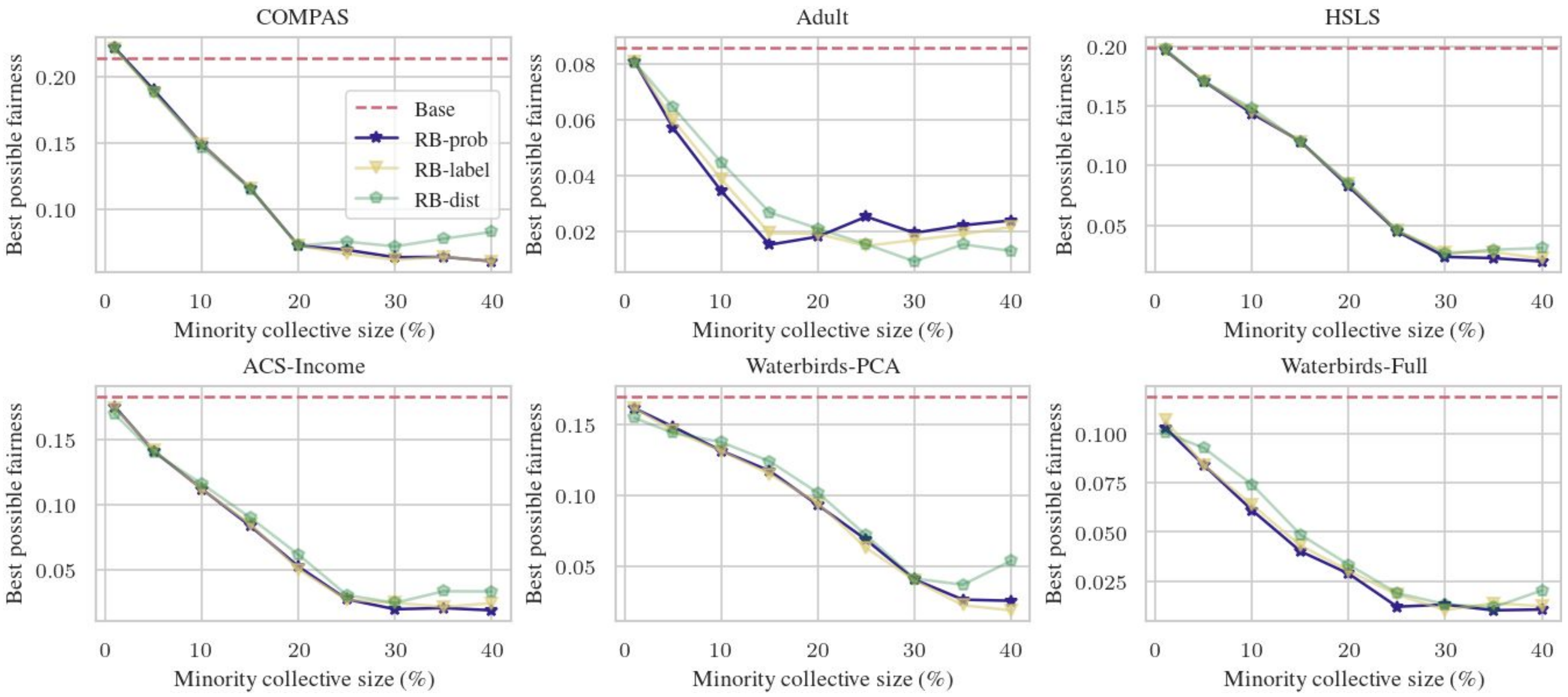
Can a minority group collaborate for fairness?

What should the minority do? Change interaction....

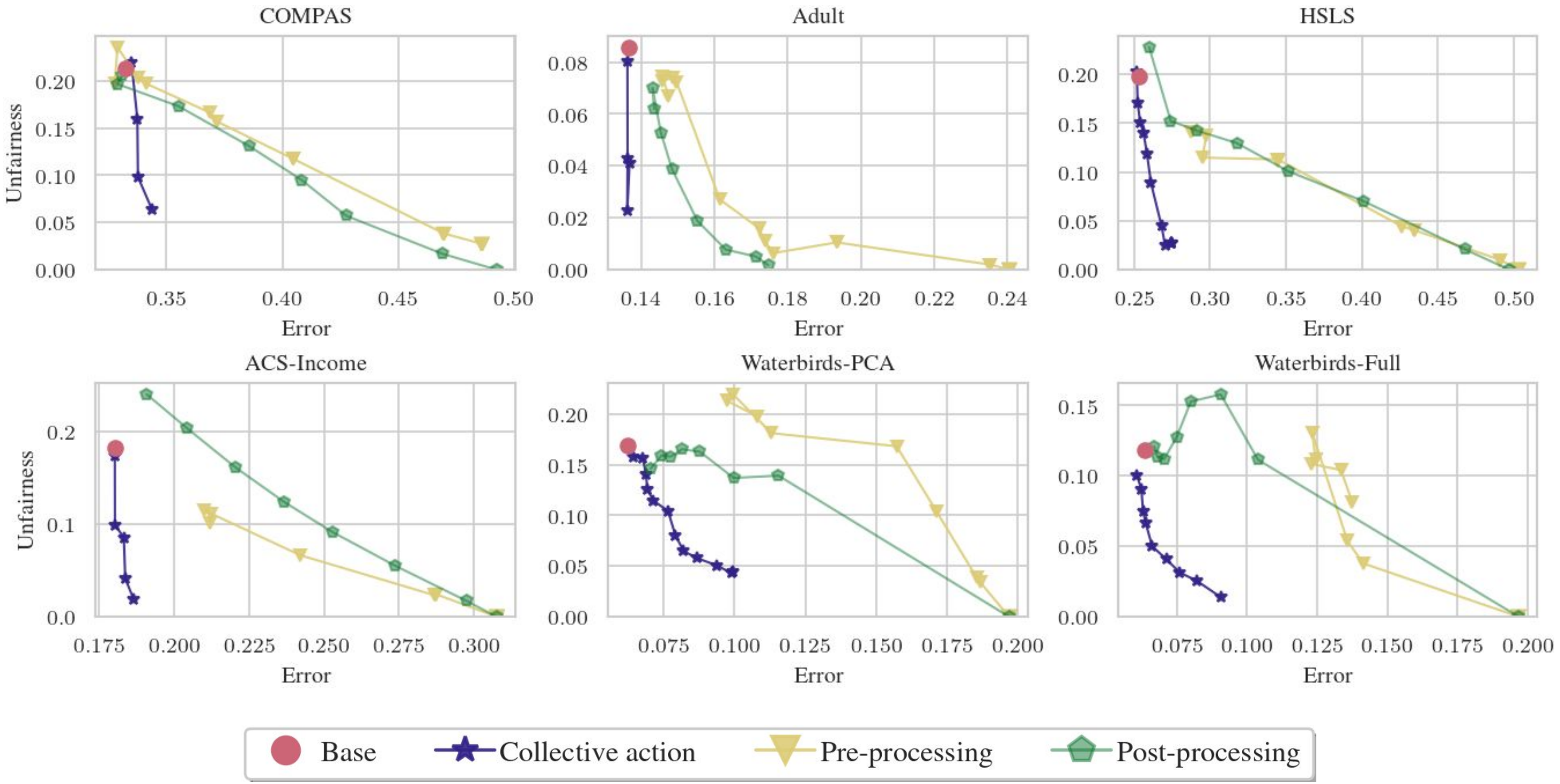


Can a minority group collaborate for fairness?

Best fairness with only 30% of the minority!



Smaller error, but no “perfect” fairness

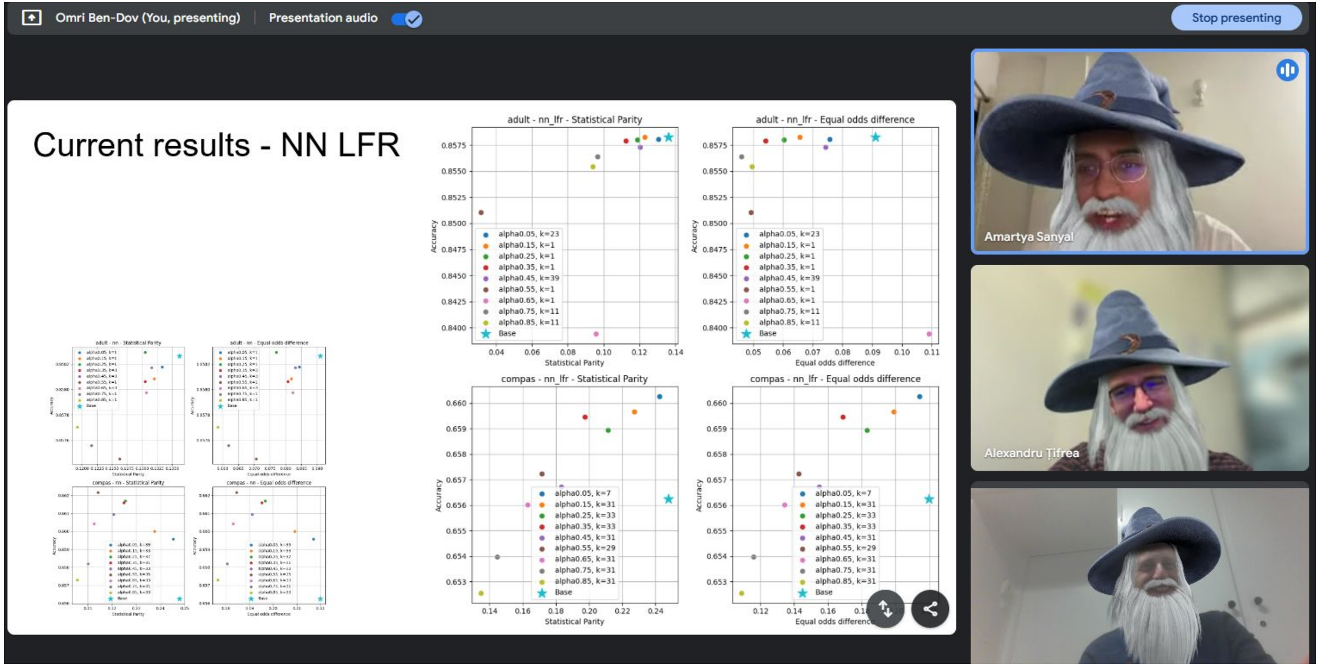


Only first steps

- “Simple” fairness problems
  - Can be improved?

- How else can collective action contribute to social good?

The team


Omri Ben-Dov (You, presenting)
Presentation audio
Stop presenting
Current results - NN LFR
adult - nn_lfr - Statistical Parity
Statistical Parity
Accuracy
alpha0.05, k=23
alpha0.15, k=1
alpha0.25, k=1
alpha0.35, k=1
alpha0.45, k=39
alpha0.55, k=1
alpha0.65, k=1
alpha0.75, k=11
alpha0.85, k=11
Base
adult - nn_lfr - Equal odds difference
Equal odds difference
compas - nn_lfr - Statistical Parity
alpha0.05, k=7
alpha0.15, k=31
alpha0.25, k=33
alpha0.35, k=33
alpha0.45, k=31
alpha0.55, k=29
alpha0.65, k=31
alpha0.75, k=31
alpha0.85, k=31
Base
compas - nn_lfr - Equal odds difference
Amartya Sanyal
Alexandru Țifrea

Recap

1. AI = Machine learning
    - Use a lot of data to answer a simple question (including chatGPT)

Recap

1. AI = Machine learning
    - Use a lot of data to answer a simple question (including chatGPT)

2. Machine learning can be unfair
    - Fairness leads to error
    - Apps remain unfair

Recap

1. AI = Machine learning
    - Use a lot of data to answer a simple question (including chatGPT)

2. Machine learning can be unfair
    - Fairness leads to error
    - Apps remain unfair

3. Work together!
    - Can lead to positive impact!

Recap

1. AI = Machine learning
   - Use a lot of data to answer a simple question (including chatGPT)

2. Machine learning can be unfair
   - Fairness leads to error
   - Apps remain unfair

3. Work together!
   - Can lead to positive impact!

For the slides and more details