The AI/Data Science Professional

AI Bias and Fairness

Acknowledgement of Country

The AI/Data Science Professional

Course Coordinator: Flora Salim

What's next...

RMIT University acknowledges the people of the Woi wurrung and Boon wurrung language groups of the eastern Kulin Nation on whose unceded lands we conduct the business of the University.

RMIT University respectfully acknowledges their Ancestors and Elders, past and present. RMIT also acknowledges the Traditional Custodians and their Ancestors of the lands and waters across Australia where we conduct our business.

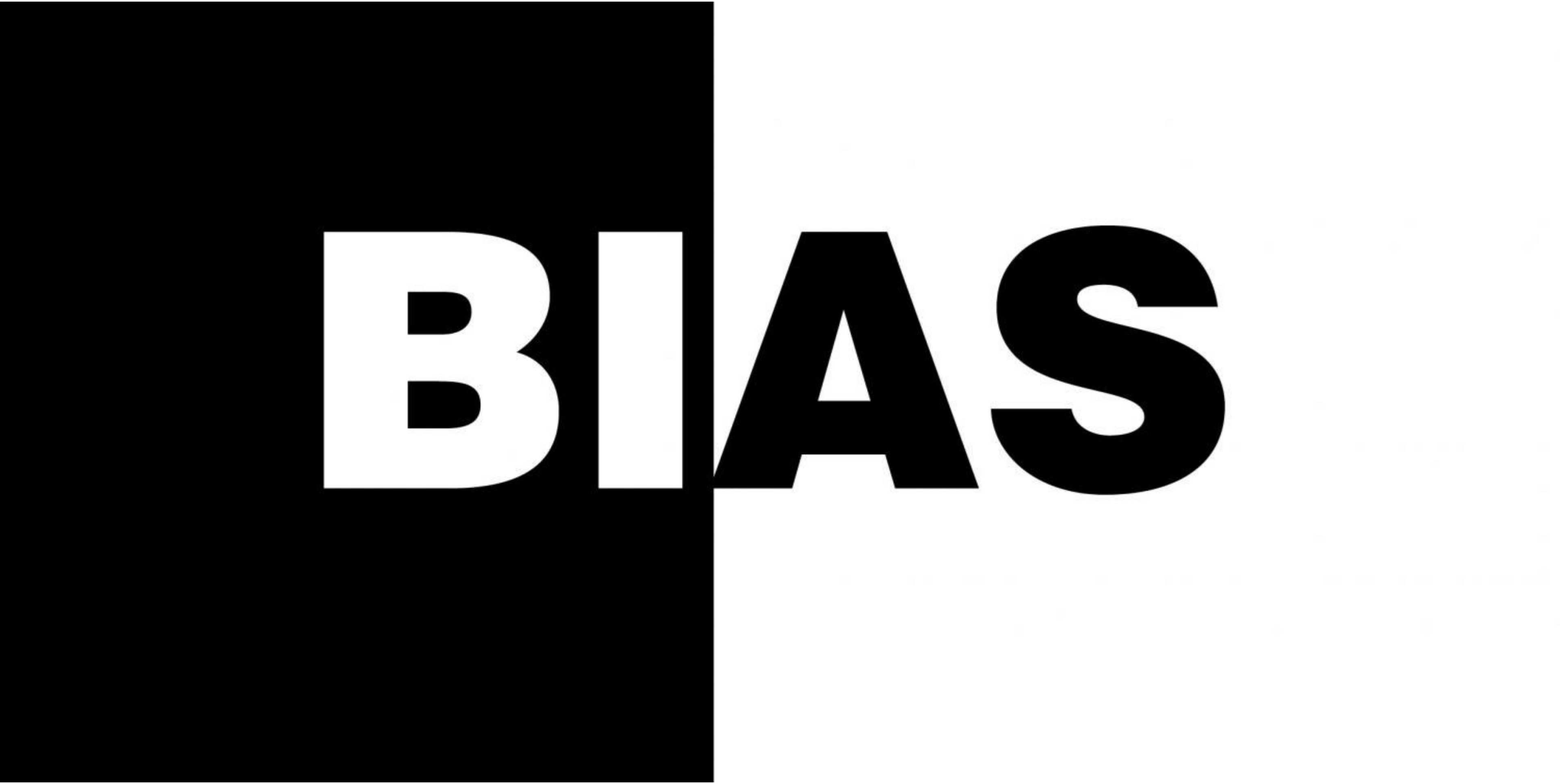
Ngarara Place

The AI/Data Science Professional – Week 6

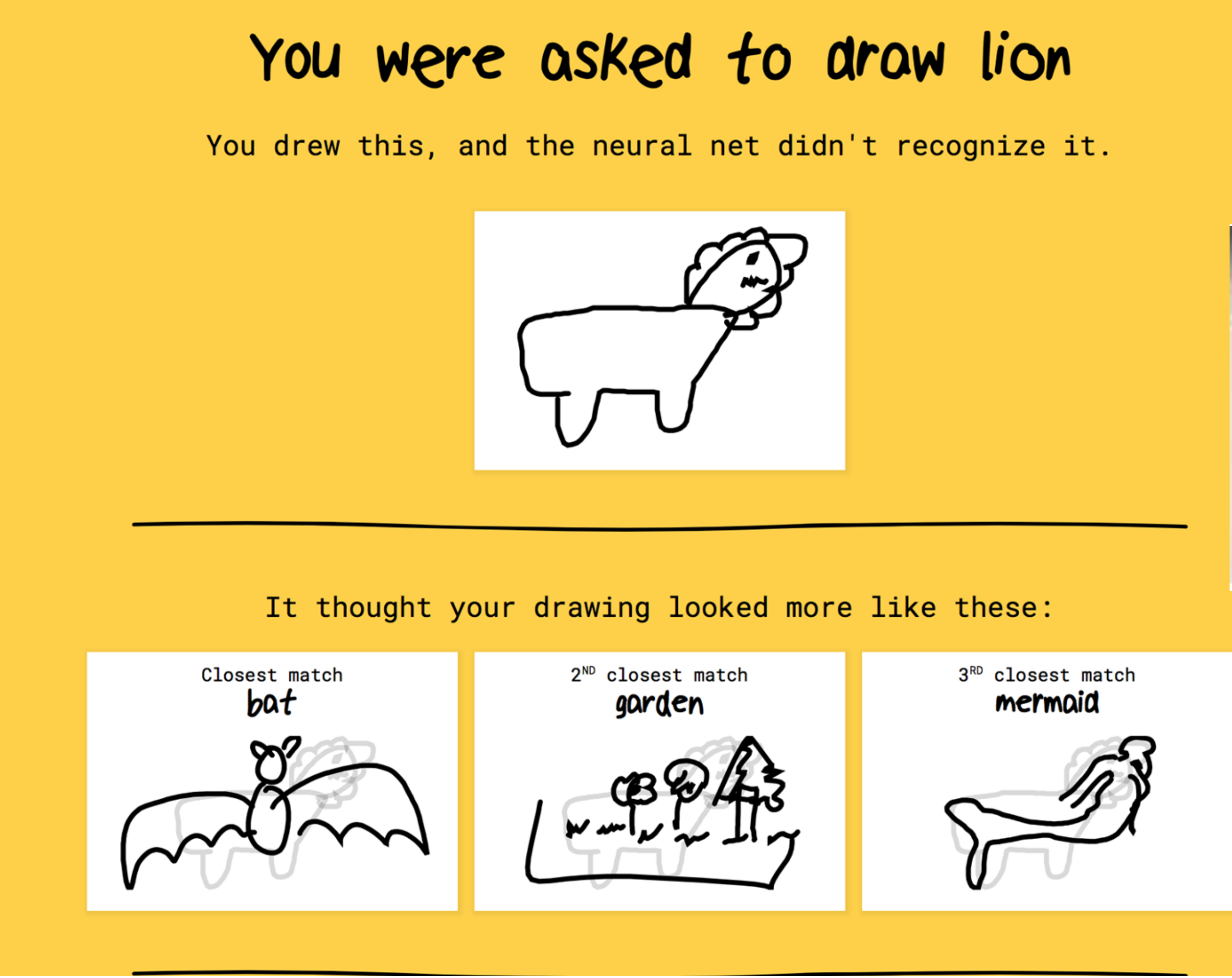
AI Bias and Fairness

What's next…

RMIT UNIVERSITY



Source from https://medium.com/thoughts-and-reflections/racial-bias-and-gender-bias-examples-in-ai-systems-7211e4c166a1



You were asked to draw lion

You drew this, and the neural net didn't recognize it.

It thought your drawing looked more like these:

Closest match
bat

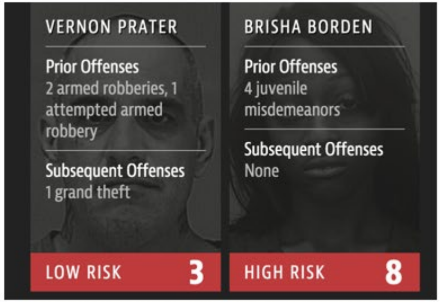
2ND closest match
garden

3RD closest match
mermaid

The Increasing Autonomy

Real-world Examples

- COMPAS



VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

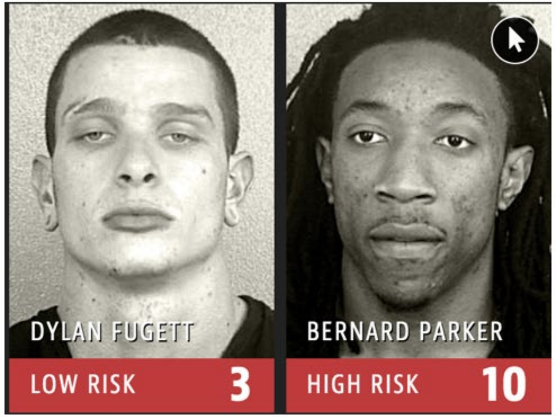
LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

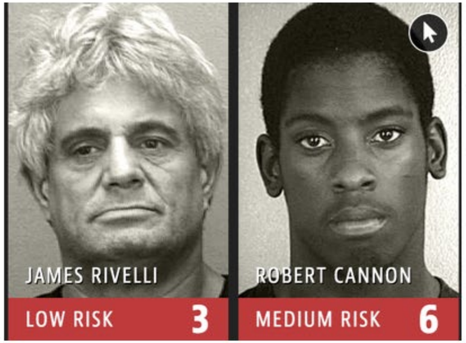


DYLAN FUGETT

LOW RISK 3

BERNARD PARKER

HIGH RISK 10

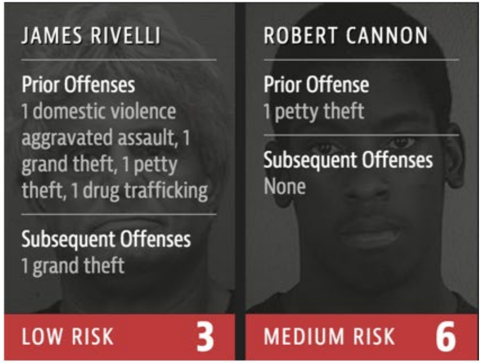


JAMES RIVELLI

LOW RISK 3

ROBERT CANNON

MEDIUM RISK 6



JAMES RIVELLI

Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking

Subsequent Offenses
1 grand theft

LOW RISK 3

ROBERT CANNON

Prior Offense
1 petty theft

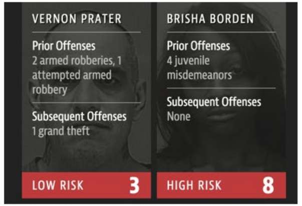
Subsequent Offenses
None

MEDIUM RISK 6

Image source from https://medium.com/thoughts-and-reflections/racial-bias-and-gender-bias-examples-in-ai-systems-7211e4c166a1

Real-world Examples

- COMPAS



VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

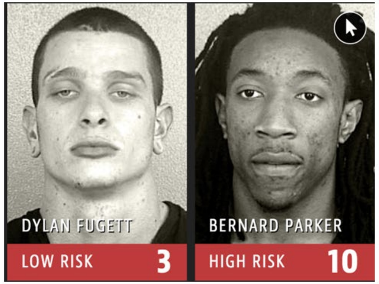
LOW RISK 3

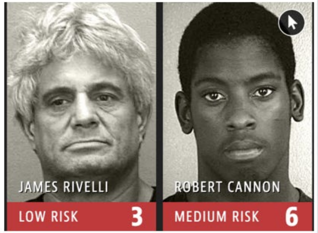
BRISHA BORDEN

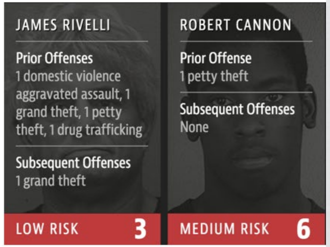
Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8







JAMES RIVELLI

Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking

Subsequent Offenses
1 grand theft

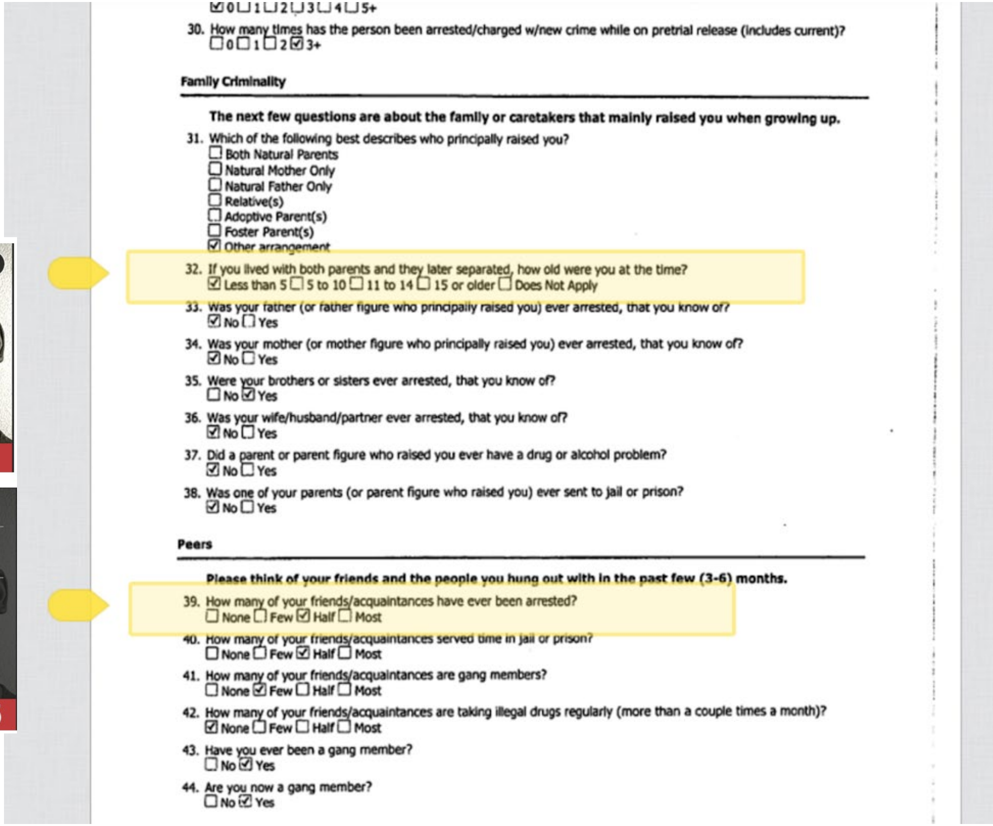
LOW RISK 3

ROBERT CANNON

Prior Offense
1 petty theft

Subsequent Offenses
None

MEDIUM RISK 6



☑0 □1 □2 □3 □4 □5+

30. How many times has the person been arrested/charged w/new crime while on pretrial release (includes current)?
□0 □1 □2 ☑3+

Family Criminality

The next few questions are about the family or caretakers that mainly raised you when growing up.

31. Which of the following best describes who principally raised you?
□ Both Natural Parents
□ Natural Mother Only
□ Natural Father Only
□ Relative(s)
□ Adoptive Parent(s)
□ Foster Parent(s)
☑ Other arrangement

32. If you lived with both parents and they later separated, how old were you at the time?
☑ Less than 5 □ 5 to 10 □ 11 to 14 □ 15 or older □ Does Not Apply

33. Was your father (or father figure who principally raised you) ever arrested, that you know of?
☑ No □ Yes

34. Was your mother (or mother figure who principally raised you) ever arrested, that you know of?
☑ No □ Yes

35. Were your brothers or sisters ever arrested, that you know of?
□ No ☑ Yes

36. Was your wife/husband/partner ever arrested, that you know of?
☑ No □ Yes

37. Did a parent or parent figure who raised you ever have a drug or alcohol problem?
☑ No □ Yes

38. Was one of your parents (or parent figure who raised you) ever sent to jail or prison?
☑ No □ Yes

Peers

Please think of your friends and the people you hung out with in the past few (3-6) months.

39. How many of your friends/acquaintances have ever been arrested?
□ None □ Few ☑ Half □ Most

40. How many of your friends/acquaintances served time in jail or prison?
□ None □ Few ☑ Half □ Most

41. How many of your friends/acquaintances are gang members?
□ None ☑ Few □ Half □ Most

42. How many of your friends/acquaintances are taking illegal drugs regularly (more than a couple times a month)?
☑ None □ Few □ Half □ Most

43. Have you ever been a gang member?
□ No ☑ Yes

44. Are you now a gang member?
□ No ☑ Yes

Image source from https://medium.com/thoughts-and-reflections/racial-bias-and-gender-bias-examples-in-ai-systems-7211e4c166a1

Real-world Examples

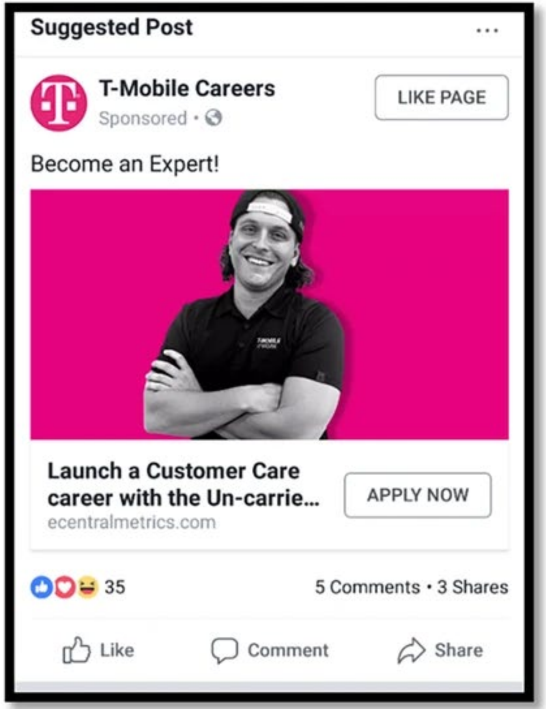
- Job Ads

[Discussion]

Sometime ago, Verizon, one of the largest communication technology companies, placed an ad on Facebook to recruit applicants for a unit focused on financial planning and analysis. The ad showed a smiling, millennial-aged woman seated at a computer and promised that new hires could look forward to a rewarding career in which they would be "more than just a number." Some relevant numbers were not immediately evident. The promotion was set to run on the Facebook feeds of users 25 to 36 years old who lived in the nation's capital, or had recently visited there, and had demonstrated an interest in finance. For a vast majority of the hundreds of millions of people who check Facebook every day, the ad did not exist.

Real-world Examples

- Facebook Career Ads



Suggested Post

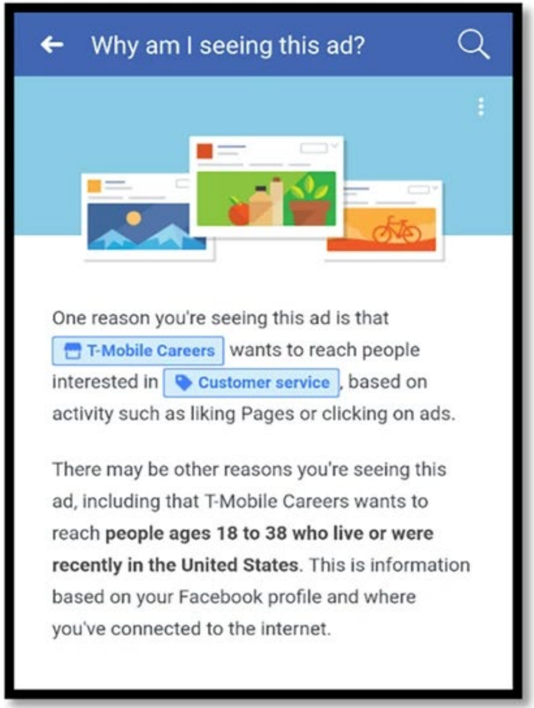
T-Mobile Careers

Sponsored ·

Become an Expert!

Launch a Customer Care career with the Un-carrie...

ecentralmetrics.com



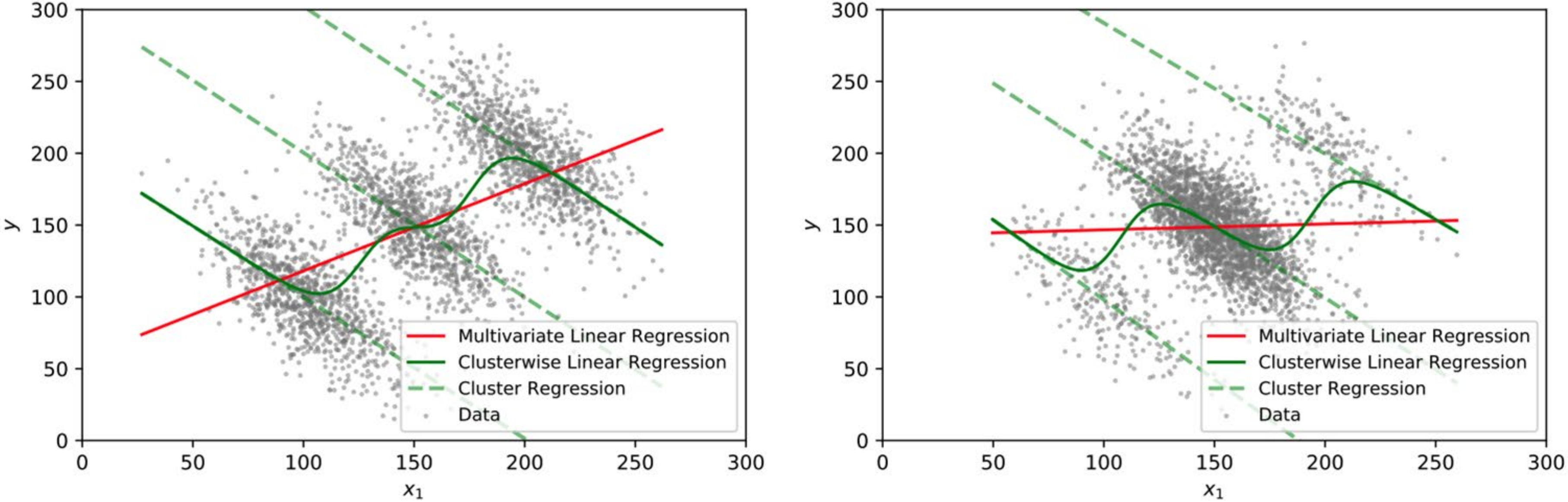
Why am I seeing this ad?

One reason you're seeing this ad is that T-Mobile Careers wants to reach people interested in Customer service, based on activity such as liking Pages or clicking on ads.

There may be other reasons you're seeing this ad, including that T-Mobile Careers wants to reach **people ages 18 to 38 who live or were recently in the United States**. This is information based on your Facebook profile and where you've connected to the internet.

Bias

Bias in Data



Source from https://arxiv.org/pdf/1908.09635.pdf

Types of Bias (Mehrabi et al. 2019)

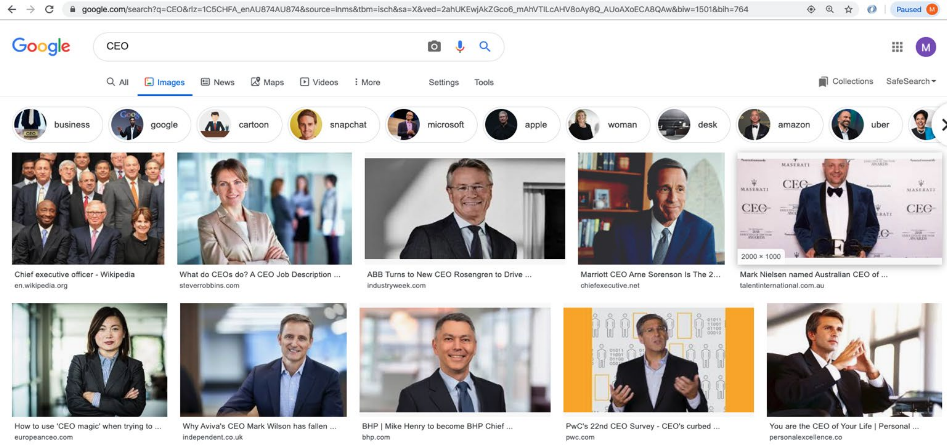
23 types of bias !!

- Historical Bias
- Representation Bias
- Measurement Bias
- Evaluation Bias
- Aggregation Bias
- Population Bias
- Simpson's Paradox
- Longitudinal Data Fallacy
- Sampling Bias
- Behavioral Bias
- Content Production Bias
- Linking Bias
- Temporal Bias
- Popularity Bias
- Algorithmic Bias
- User Interaction Bias
- Social Bias
- Emergent Bias
- Self-Selection Bias
- Omitted Variable Bias
- Cause-Effect Bias
- Observer Bias
- Funding Bias

https://arxiv.org/pdf/1908.09635.pdf

Common Types of Bias

- Historical bias
  - Historical bias is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection.
  - existing bias and socio-technical issues in the world, e.g. see below:



google.com/search?q=CEO&rlz=1C5CHFA_enAU874AU874&source=lnms&tbm=isch&sa=X&ved=2ahUKEwjAkZGco6_mAhVTILcAHV8oAy8Q_AUoAXoECA8QAw&biw=1501&bih=764

CEO

Chief executive officer - Wikipedia
en.wikipedia.org

What do CEOs do? A CEO Job Description ...
steverrobbins.com

ABB Turns to New CEO Rosengren to Drive ...
industryweek.com

Marriott CEO Arne Sorenson Is The 2...
chiefexecutive.net

2000 × 1000

Mark Nielsen named Australian CEO of ...
talentinternational.com.au

How to use 'CEO magic' when trying to ...
europeanceo.com

Why Aviva's CEO Mark Wilson has fallen ...
independent.co.uk

BHP | Mike Henry to become BHP Chief ...
bhp.com

PwC's 22nd CEO Survey - CEO's curbed ...
pwc.com

You are the CEO of Your Life | Personal ...
personalexcellence.co

Source from https://arxiv.org/pdf/1908.09635.pdf

Historical bias

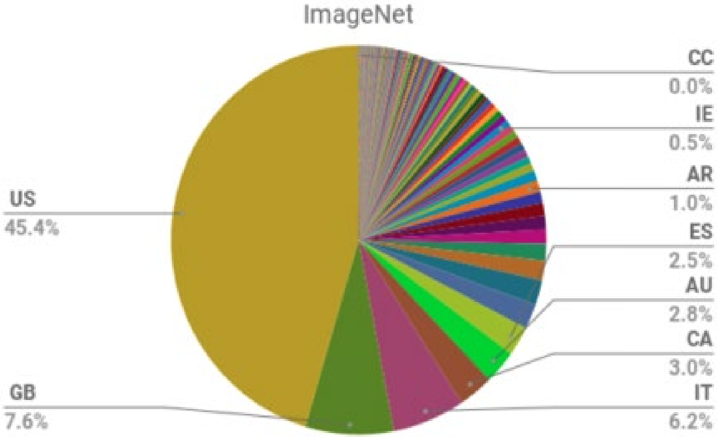
- **Historical bias** occurs when the state of the world in which the data was generated is flawed.

- As of 2020, only 7.4% of Fortune 500 CEOs are women. Research has shown that companies with female CEOs or CFOs are generally more profitable than companies with men in the same position, suggesting that women are held to higher hiring standards than men. In order to fix this, we might consider removing human input and using AI to make the hiring process more equitable. But this can prove unproductive if data from past hiring decisions is used to train a model, because the model will likely learn to demonstrate the same biases that are present in the data.

from https://arxiv.org/pdf/1901.10002.pdf

Common Types of Bias

- Representation bias
  - Representation bias happens from the way we define and sample from a population feature selection.
  - Comes from the way we define and sample from a population, e.g.



Source from https://arxiv.org/pdf/1711.08536.pdf

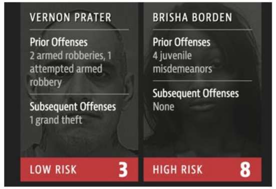
Representation bias

- **Representation bias** occurs when building datasets for training a model, if those datasets poorly represent the people that the model will serve.

- Data collected through smartphone apps will under-represent groups that are less likely to own smartphones. For instance, if collecting data in the USA, individuals over the age of 65 will be under-represented. If the data is used to inform design of a city transportation system, this will be disastrous, since older people have important needs to ensure that the system is accessible.

Common Types of Bias

- Measurement bias
  - Measurement bias happens from the way we choose, utilize, and measure a particular feature.
  - Choosing and measuring the particular features of interest, e.g.,



VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

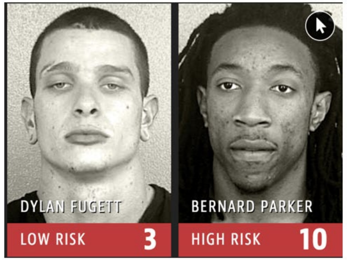
LOW RISK 3

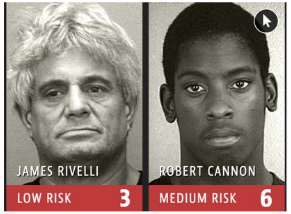
BRISHA BORDEN

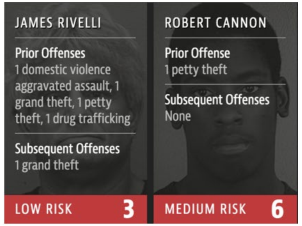
Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8







JAMES RIVELLI

Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking

Subsequent Offenses
1 grand theft

LOW RISK 3

ROBERT CANNON

Prior Offense
1 petty theft

Subsequent Offenses
None

MEDIUM RISK 6

Source from https://arxiv.org/pdf/1908.09635.pdf

Measurement bias

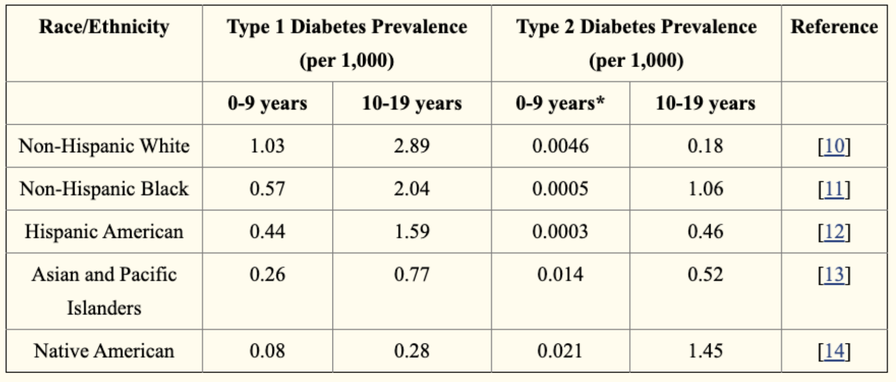
Measurement bias occurs when the accuracy of the data varies across groups. This can happen when working with proxy variables (variables that take the place of a variable that cannot be directly measured), if the quality of the proxy varies in different groups.

Measurement bias

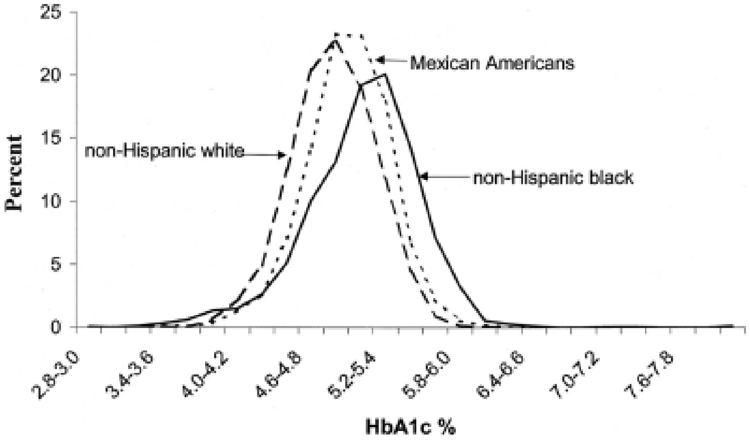
- Your local hospital uses a model to identify high-risk patients before they develop serious conditions, based on information like past diagnoses, medications, and demographic data. The model uses this information to predict health care costs, the idea being that patients with higher costs likely correspond to high-risk patients. Despite the fact that the model specifically excludes race, it seems to demonstrate racial discrimination: the algorithm is less likely to select eligible Black patients. How can this be the case? It is because cost was used as a proxy for risk, and the relationship between these variables varies with race: Black patients experience increased barriers to care, have less trust in the health care system, and therefore have lower medical costs, on average, when compared to non-Black patients with the same health conditions.

Common Types of Bias

- Aggregation bias
  - false conclusions are drawn for a subgroup based on observing other different subgroups or generally when false assumptions about a population affect the model's outcome and definition e.g.,



Race/Ethnicity	Type 1 Diabetes Prevalence (per 1,000)		Type 2 Diabetes Prevalence (per 1,000)		Reference
	0-9 years	10-19 years	0-9 years*	10-19 years	
Non-Hispanic White	1.03	2.89	0.0046	0.18	[10]
Non-Hispanic Black	0.57	2.04	0.0005	1.06	[11]
Hispanic American	0.44	1.59	0.0003	0.46	[12]
Asian and Pacific Islanders	0.26	0.77	0.014	0.52	[13]
Native American	0.08	0.28	0.021	1.45	[14]



Source from https://arxiv.org/pdf/1908.09635.pdf and https://www.ncbi.nlm.nih.gov/pubmed/22238408

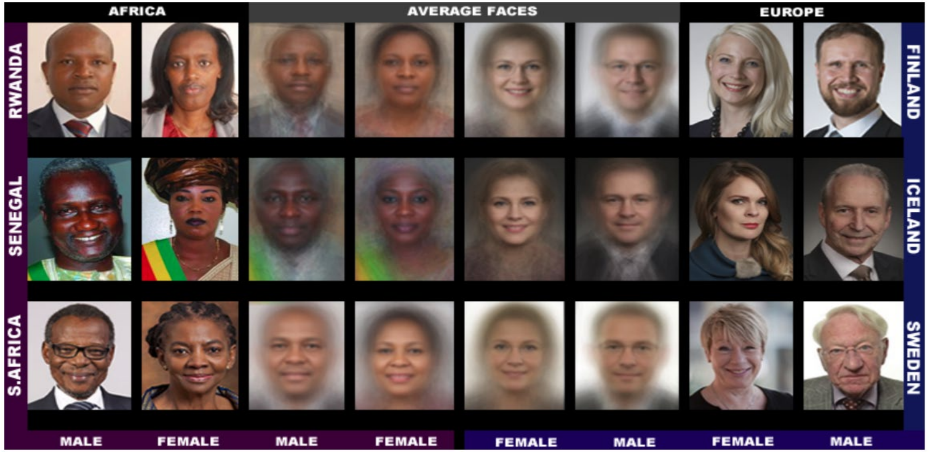
Aggregation bias

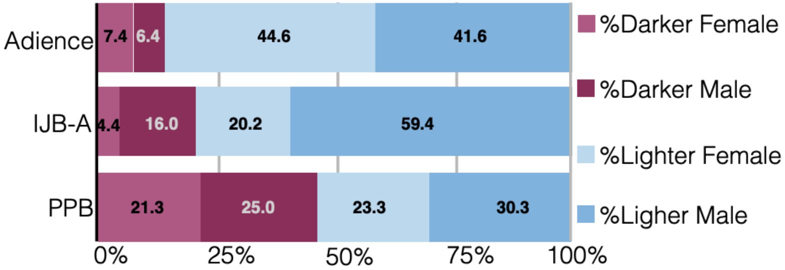
- **Aggregation bias** occurs when groups are inappropriately combined, resulting in a model that does not perform well for any group or only performs well for the majority group. (This is often not an issue, but most commonly arises in medical applications.)

- Hispanics have higher rates of diabetes and diabetes-related complications than non-Hispanic whites. If building AI to diagnose or monitor diabetes, it is important to make the system sensitive to these ethnic differences, by either including ethnicity as a feature in the data, or building separate models for different ethnic groups.

Common Types of Bias

- Evaluation bias
  - Evaluation bias happens during model evaluation. This includes the use of inappropriate and disproportionate benchmarks for the evaluation of applications.
  - occurs during model iteration and evaluation, e.g.,





Source from http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Evaluation bias

- **Evaluation bias** occurs when evaluating a model, if the benchmark data (used to compare the model to other models that perform similar tasks) does not represent the population that the model will serve.

- The Gender Shades paper discovered that two widely used facial analysis benchmark datasets (IJB-A and Adience) were primarily composed of lighter-skinned subjects (79.6% and 86.2%, respectively). Commercial gender classification AI showed state-of-the-art performance on these benchmarks, but experienced disproportionately high error rates with people of color.

Deployment bias

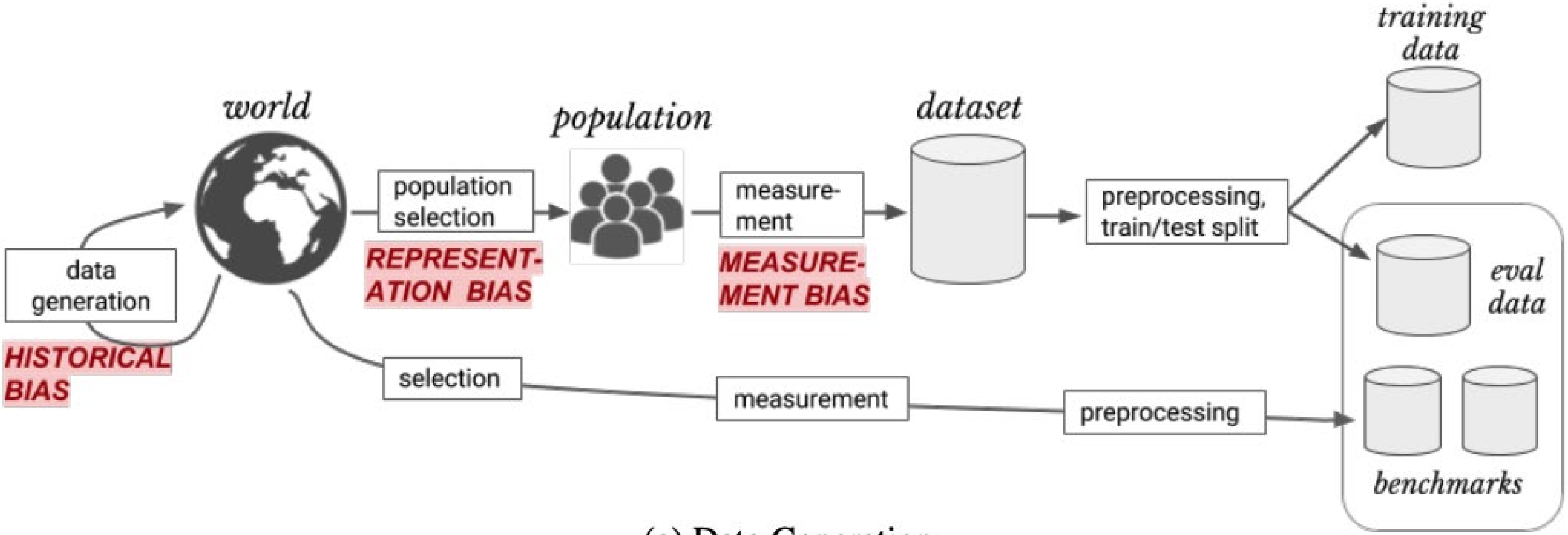
- **Deployment bias** occurs when the problem the model is intended to solve is different from the way it is actually used. If the end users don't use the model in the way it is intended, there is no guarantee that the model will perform well.

- The criminal justice system uses tools to predict the likelihood that a convicted criminal will relapse into criminal behavior. The predictions are not designed for judges when deciding appropriate punishments at the time of sentencing.

Other Common Types of Bias

- **Population Bias.** Population bias arises when statistics, demographics, representatives, and user characteristics are different in the user population represented in the dataset or platform from the original target population.
- **Sampling Bias.** Sampling bias arises due to the non-random sampling of subgroups. As a consequence of sampling bias, the trends estimated for one population may not generalize to data collected from a new population.
- **Temporal Bias.** Temporal bias arises from differences in populations and behaviors over time.
- **Social Bias**. Social bias happens when other people’s actions or content coming from them affects our judgment.

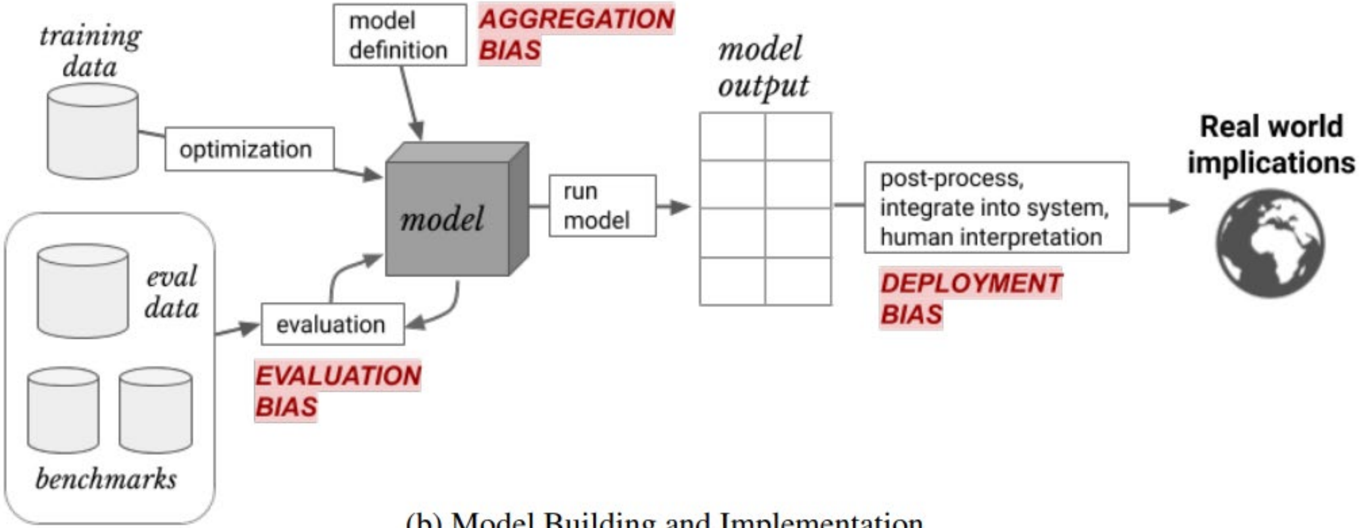
Where do they exist?



(a) Data Generation

Source from https://arxiv.org/pdf/1901.10002.pdf

Where do they exist?



(b) Model Building and Implementation

Source from https://arxiv.org/pdf/1901.10002.pdf

Other classification and types of bias

- Population bias
- Simpson’s Paradox
- Longitudinal data fallacy
- Sampling bias
- ……

[Optional Reading]

Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Algorithmic Fairness

Algorithmic Fairness

- Types of Discrimination
  - Direct vs. Indirect discrimination
  - Systemic discrimination
  - Explainable vs. unexplainable discrimination
  - ……

Source from https://arxiv.org/pdf/1901.10002.pdf

[Discussion]

Amazon's same-day delivery

Amazon recently rolled out same-day delivery across a select group of American cities. However, this service was only extended to neighborhoods with a high number of current Amazon users. As a result, predominantly non-white neighborhoods were largely excluded from the service.

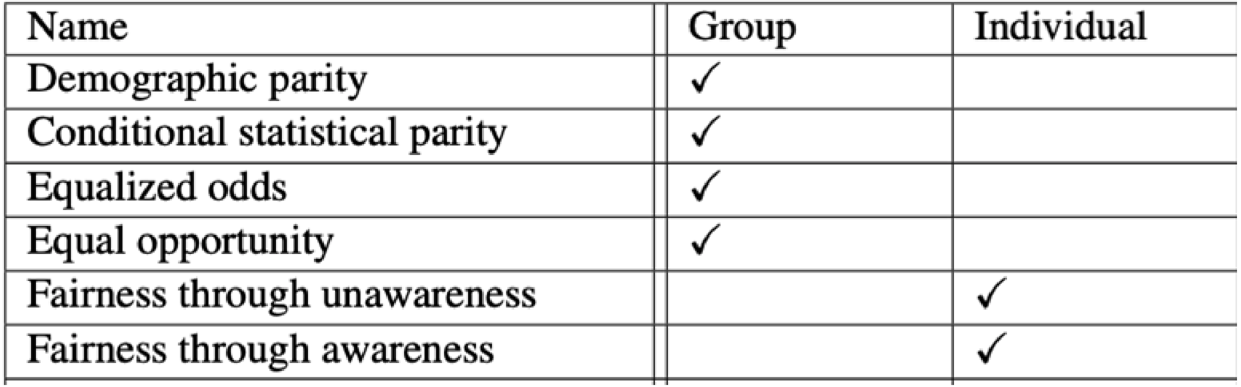
Algorithmic Fairness

- Definitions of Fairness
  - Equalised odds
  - Equal opportunity
  - Demographic parity
  - Fairness through awareness
  - Fairness through unawareness
  - Treatment equality
  - Test fairness
  - ……

Source from https://arxiv.org/pdf/2006.16745.pdf

Algorithmic Fairness

- Categories of Fairness
  - Individual Fairness
  - Group Fairness
  - Subgroup Fairness



Name	Group	Individual
Demographic parity	✓	
Conditional statistical parity	✓	
Equalized odds	✓	
Equal opportunity	✓	
Fairness through unawareness		✓
Fairness through awareness		✓

Source from https://arxiv.org/pdf/1710.03184.pdf

Fairness criteria: Demographic parity /statistical parity

- **Demographic parity** says the model is fair if the composition of people who are selected by the model matches the group membership percentages of the applicants.

- A nonprofit is organizing an international conference, and 20,000 people have signed up to attend. The organizers write a ML model to select 100 attendees who could potentially give interesting talks at the conference. Since 50% of the attendees will be women (10,000 out of 20,000), they design the model so that 50% of the selected speaker candidates are women.

Fairness criteria: Equal opportunity

- **Equal opportunity** fairness ensures that the proportion of people who should be selected by the model ("positives") that are correctly selected by the model is the same for each group. We refer to this proportion as the **true positive rate** (TPR) or **sensitivity** of the model.

- A doctor uses a tool to identify patients in need of extra care, who could be at risk for developing serious medical conditions. (This tool is used only to supplement the doctor's practice, as a second opinion.) It is designed to have a high TPR that is equal for each demographic group.

Fairness criteria: Equal accuracy

- Alternatively, we could check that the model has **equal accuracy** for each group. That is, the percentage of correct classifications (people who should be denied and are denied, and people who should be approved who are approved) should be the same for each group. If the model is 98% accurate for individuals in one group, it should be 98% accurate for other groups.

- A bank uses a model to approve people for a loan. The model is designed to be equally accurate for each demographic group: this way, the bank avoids approving people who should be rejected (which would be financially damaging for both the applicant and the bank) and avoid rejecting people who should be approved (which would be a failed opportunity for the applicant and reduce the bank's revenue).

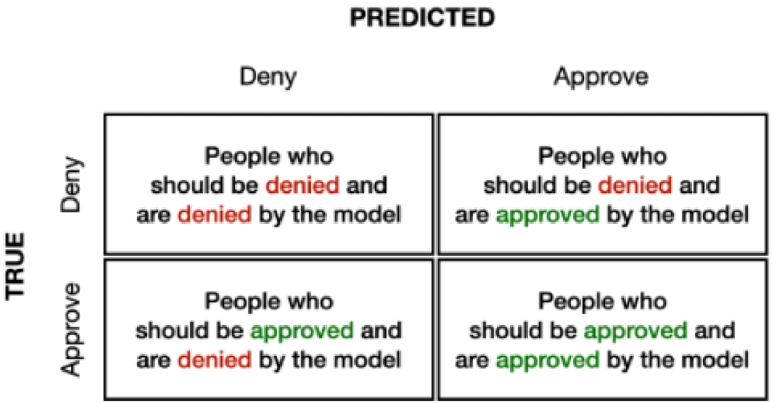
Fairness criteria: Group unaware / "Fairness through unawareness"

- **Group unaware** fairness removes all group membership information from the dataset. For instance, we can remove gender data to try to make the model fair to different gender groups. Similarly, we can remove information about race or age.

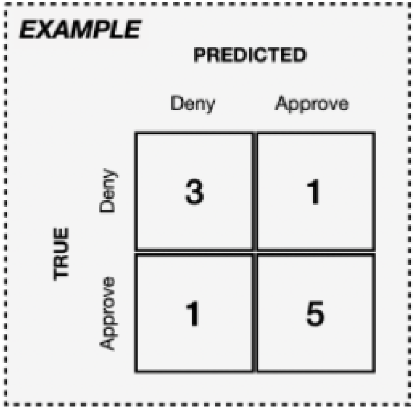
- One difficulty of applying this approach in practice is that one has to be careful to identify and remove proxies for the group membership data. For instance, in cities that are racially segregated, zip code is a strong proxy for race. That is, when the race data is removed, the zip code data should also be removed, or else the ML application may still be able to infer an individual's race from the data. Additionally, group unaware fairness is unlikely to be a good solution for historical bias.

Example

- We'll work with a small example to illustrate the differences between the four different types of fairness. We'll use a **confusion matrix**, which is a common tool used to understand the performance of a ML model. This tool is depicted in the example below, which depicts a model with 80% accuracy (since 8/10 people were correctly classified) and has an 83% true positive rate (since 5/6 "positives" were correctly classified).



		PREDICTED	
		Deny	Approve
TRUE	Deny	People who should be denied and are denied by the model	People who should be denied and are approved by the model
	Approve	People who should be approved and are denied by the model	People who should be approved and are approved by the model



EXAMPLE

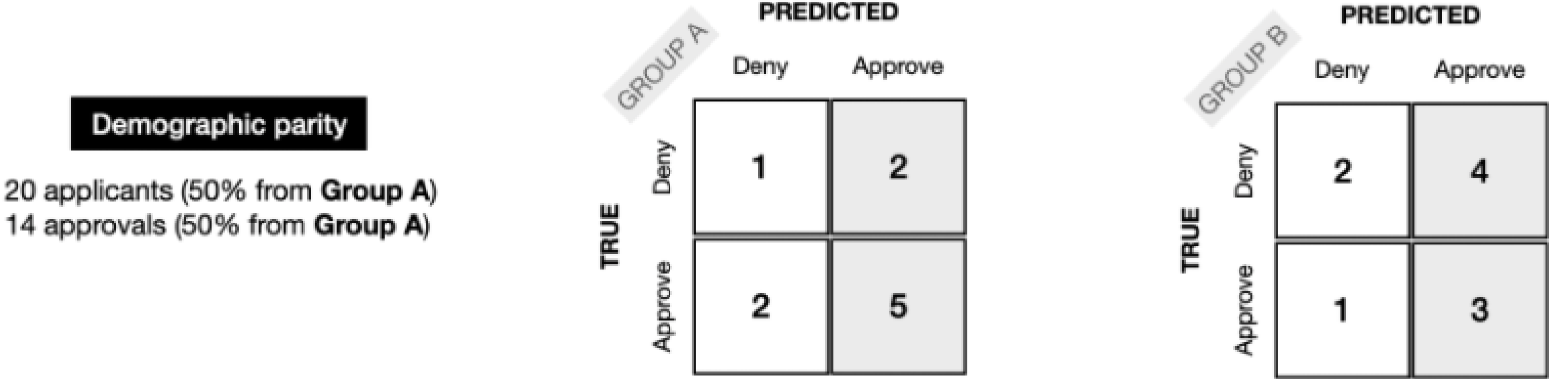
		PREDICTED	
		Deny	Approve
TRUE	Deny	3	1
	Approve	1	5

Example (cont.)

- To understand how a model's performance varies across groups, we can construct a different confusion matrix for each group. In this small example, we'll assume that we have data from only 20 people, equally split between two groups (10 from Group A, and 10 from Group B).

Example (cont.)

- The next image shows what the confusion matrices could look like, if the model satisfies **demographic parity** fairness. 10 people from each group (50% from Group A, and 50% from Group B) were considered by the model. 14 people, also equally split across groups (50% from Group A, and 50% from Group B) were approved by the model.



Demographic parity

20 applicants (50% from **Group A**)
14 approvals (50% from **Group A**)

GROUP A

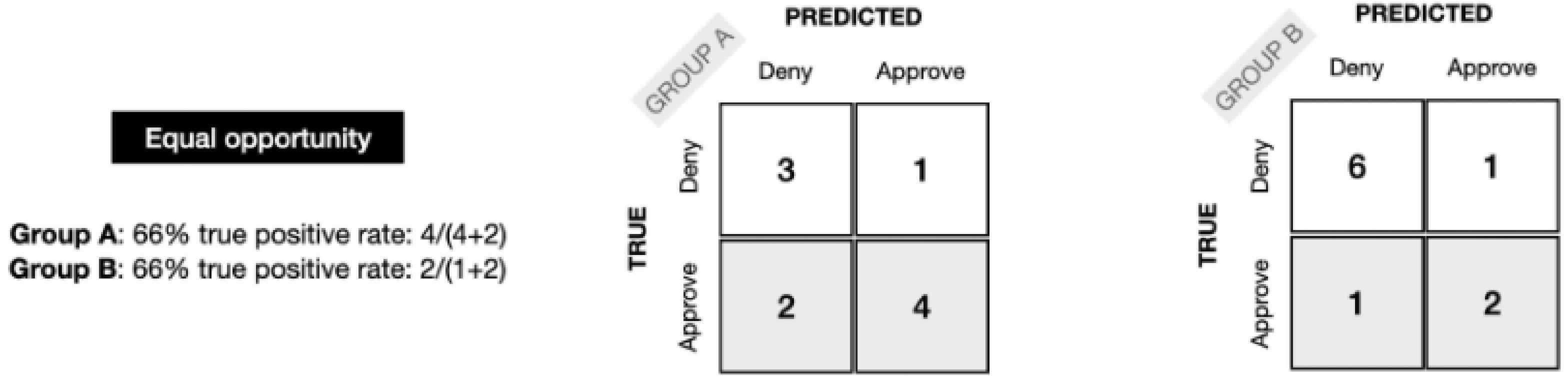
TRUE \ PREDICTED	Deny	Approve
Deny	1	2
Approve	2	5

GROUP B

TRUE \ PREDICTED	Deny	Approve
Deny	2	4
Approve	1	3

Example (cont.)

- For **equal opportunity** fairness, the TPR for each group should be the same; in the example below, it is 66% in each case.



Equal opportunity

Group A: 66% true positive rate: 4/(4+2)
Group B: 66% true positive rate: 2/(1+2)

PREDICTED

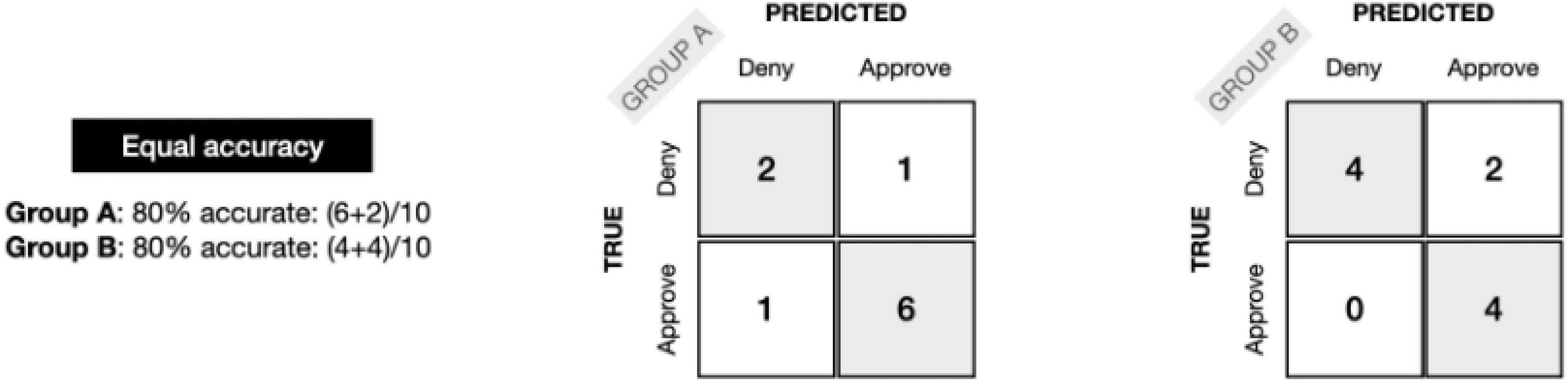
GROUP A		Deny	Approve
TRUE	Deny	3	1
	Approve	2	4

PREDICTED

GROUP B		Deny	Approve
TRUE	Deny	6	1
	Approve	1	2

Example (cont.)

- Next, we can see how the confusion matrices might look for **equal accuracy** fairness. For each group, the model was 80% accurate.



Equal accuracy

Group A: 80% accurate: (6+2)/10
Group B: 80% accurate: (4+4)/10

GROUP A

		PREDICTED	
		Deny	Approve
TRUE	Deny	2	1
	Approve	1	6

GROUP B

		PREDICTED	
		Deny	Approve
TRUE	Deny	4	2
	Approve	0	4

Example (cont.)

- Note that **group unaware** fairness cannot be detected from the confusion matrix, and is more concerned with removing group membership information from the dataset.

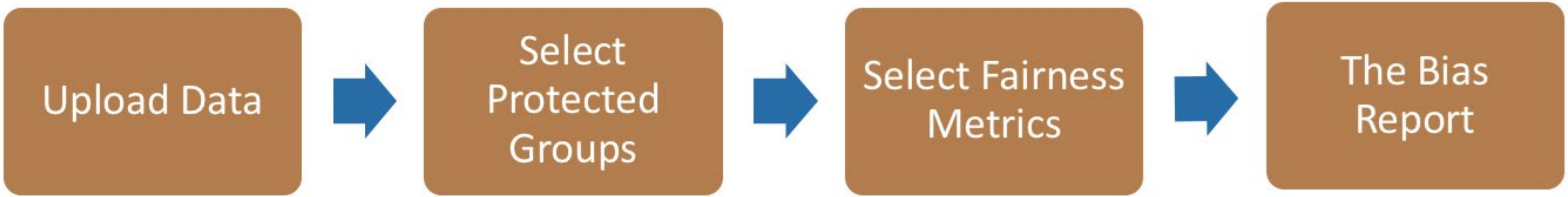
- Also note that none of the examples satisfy more than one type of fairness. For instance, the demographic parity example does not satisfy equal accuracy or equal opportunity.

To read more:
https://arxiv.org/pdf/1710.03184.pdf

What can we do about it?

Accessing Fairness

Aequitas: Bias and Fairness Audit Toolkit



Upload Data

Select Protected Groups

Select Fairness Metrics

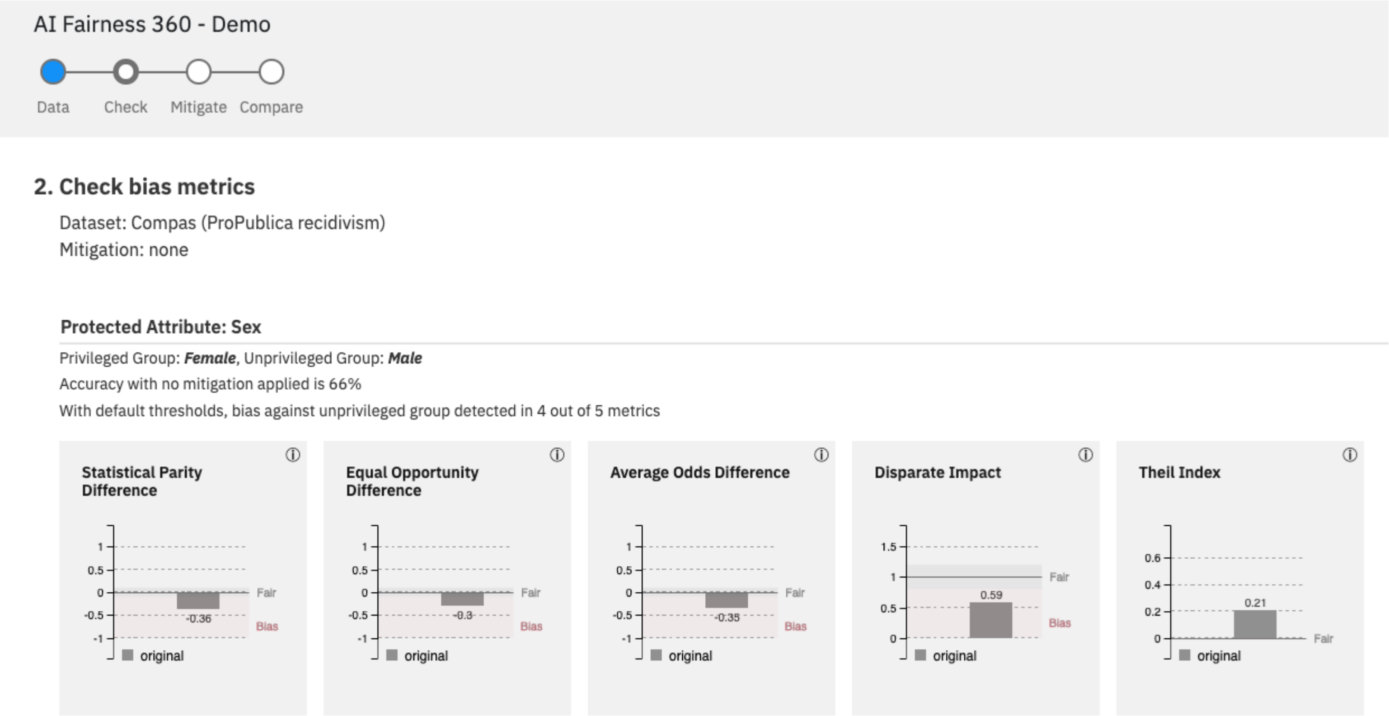
The Bias Report

There is an example report on COMPASS
http://aequitas.dssg.io/example.html#audit-results-details-by-fairness-measures

Source from http://aequitas.dssg.io

Accessing Fairness

The AI Fairness 360 (AIF360 - IBM)



2. Check bias metrics

Dataset: Compas (ProPublica recidivism)

Mitigation: none

Protected Attribute: Sex

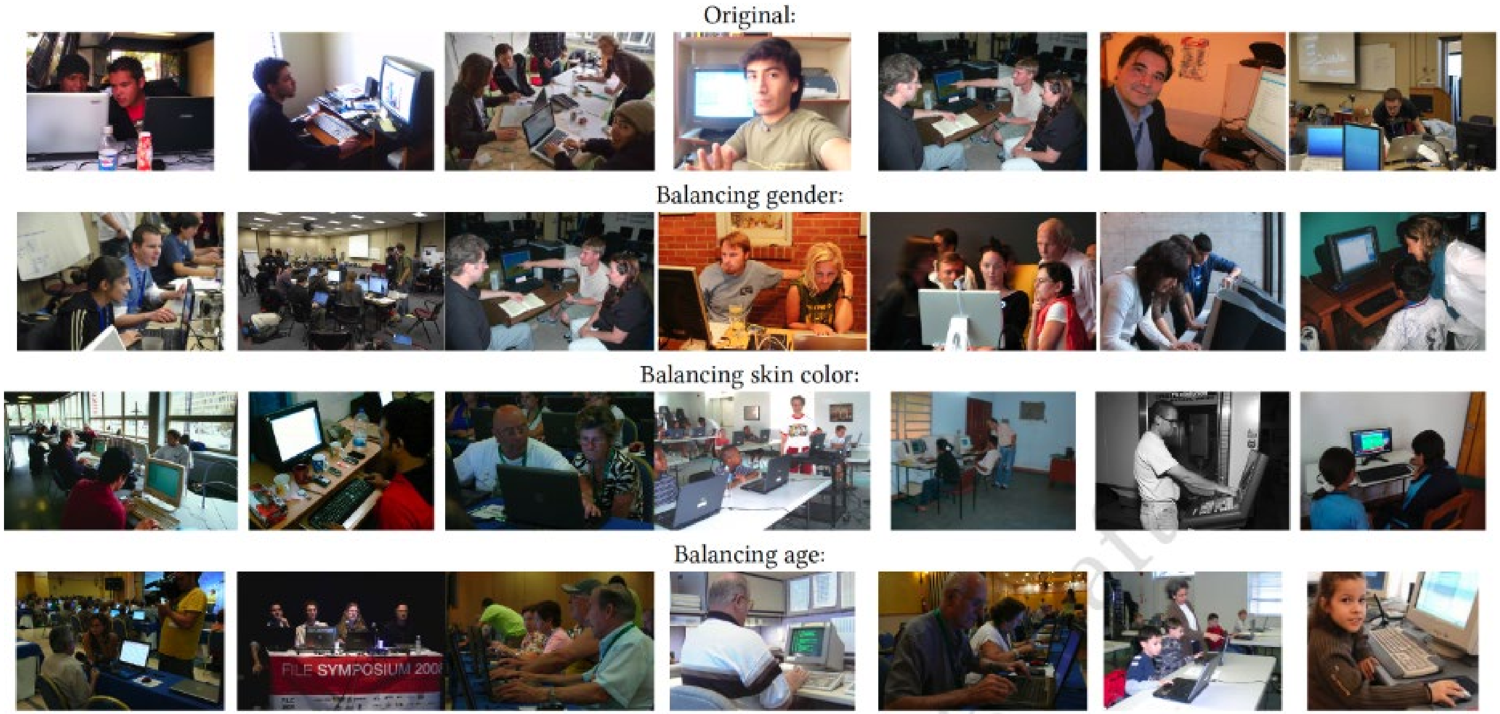
Privileged Group: ***Female***, Unprivileged Group: ***Male***

Accuracy with no mitigation applied is 66%

With default thresholds, bias against unprivileged group detected in 4 out of 5 metrics

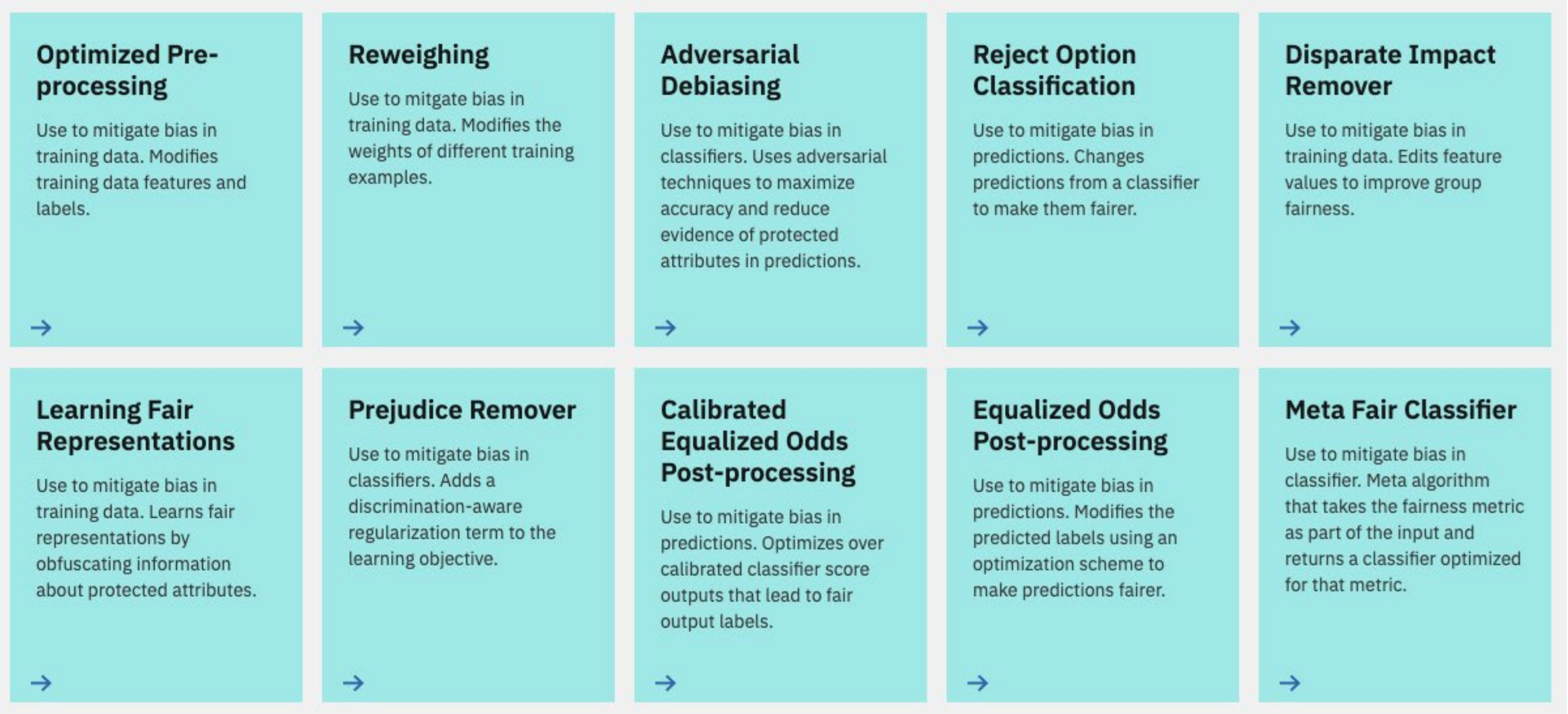
Source from https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/

Mitigation



Source from http://image-net.org/update-sep-17-2019

Mitigation



Optimized Pre-processing

Use to mitigate bias in training data. Modifies training data features and labels.

Reweighing

Use to mitgate bias in training data. Modifies the weights of different training examples.

Adversarial Debiasing

Use to mitigate bias in classifiers. Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions.

Reject Option Classification

Use to mitigate bias in predictions. Changes predictions from a classifier to make them fairer.

Disparate Impact Remover

Use to mitigate bias in training data. Edits feature values to improve group fairness.

Learning Fair Representations

Use to mitigate bias in training data. Learns fair representations by obfuscating information about protected attributes.

Prejudice Remover

Use to mitigate bias in classifiers. Adds a discrimination-aware regularization term to the learning objective.

Calibrated Equalized Odds Post-processing

Use to mitigate bias in predictions. Optimizes over calibrated classifier score outputs that lead to fair output labels.

Equalized Odds Post-processing

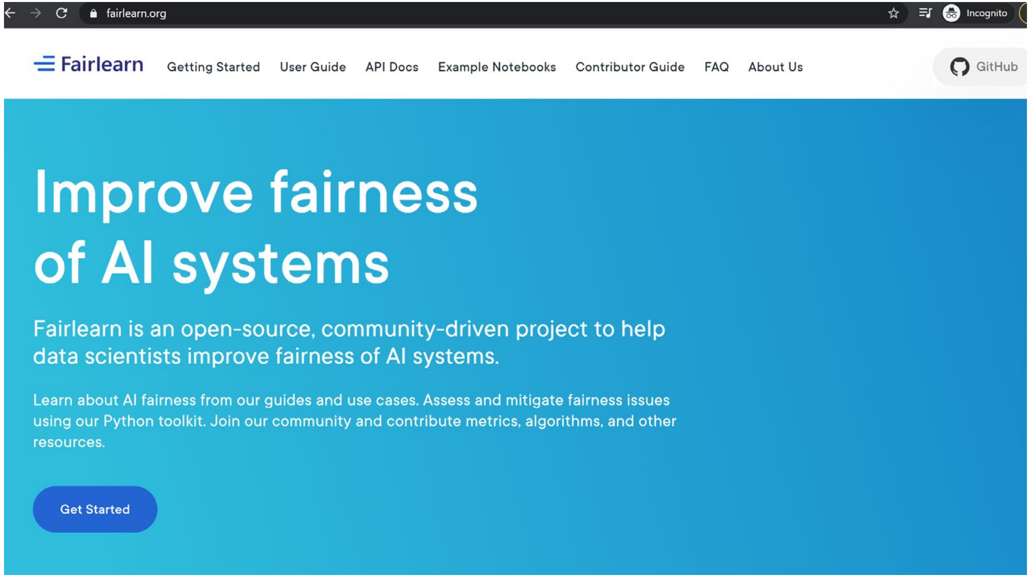
Use to mitigate bias in predictions. Modifies the predicted labels using an optimization scheme to make predictions fairer.

Meta Fair Classifier

Use to mitigate bias in classifier. Meta algorithm that takes the fairness metric as part of the input and returns a classifier optimized for that metric.

Source from https://aif360.mybluemix.net/

Fairlearn (by Microsoft)



Improve fairness of AI systems

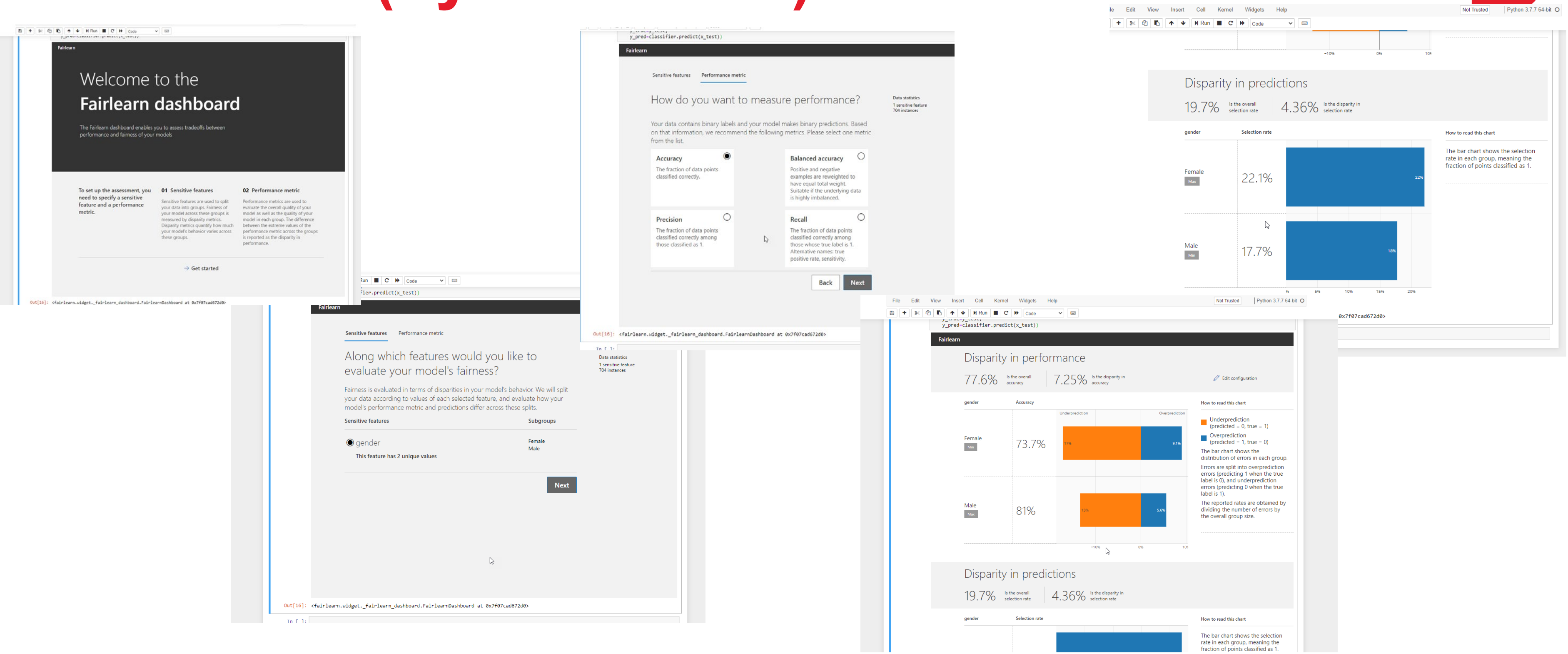
Fairlearn is an open-source, community-driven project to help data scientists improve fairness of AI systems.

Learn about AI fairness from our guides and use cases. Assess and mitigate fairness issues using our Python toolkit. Join our community and contribute metrics, algorithms, and other resources.

https://fairlearn.org/

https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

Fairlearn (by Microsoft)



Fairlearn

Welcome to the
Fairlearn dashboard

The Fairlearn dashboard enables you to assess tradeoffs between performance and fairness of your models

To set up the assessment, you need to specify a sensitive feature and a performance metric.

01 Sensitive features

Sensitive features are used to split your data into groups. Fairness of your model across these groups is measured by disparity metrics. Disparity metrics quantify how much your model's behavior varies across these groups.

02 Performance metric

Performance metrics are used to evaluate the overall quality of your model as well as the quality of your model in each group. The difference between the extreme values of the performance metric across the groups is reported as the disparity in performance.

→ Get started

```
Out[16]: <fairlearn.widget._fairlearn_dashboard.FairlearnDashboard at 0x7f07cad672d0>
```

Fairlearn

Sensitive features | Performance metric

Along which features would you like to evaluate your model's fairness?

Fairness is evaluated in terms of disparities in your model's behavior. We will split your data according to values of each selected feature, and evaluate how your model's performance metric and predictions differ across these splits.

Sensitive features	Subgroups
● gender This feature has 2 unique values	Female Male

Next

Data statistics
1 sensitive feature
704 instances

```
Out[16]: <fairlearn.widget._fairlearn_dashboard.FairlearnDashboard at 0x7f07cad672d0>
```

```
    y_pred=classifier.predict(x_test))
```

Fairlearn

Sensitive features | Performance metric

How do you want to measure performance?

Data statistics
1 sensitive feature
704 instances

Your data contains binary labels and your model makes binary predictions. Based on that information, we recommend the following metrics. Please select one metric from the list.

Accuracy ●
The fraction of data points classified correctly.

Balanced accuracy ○
Positive and negative examples are reweighted to have equal total weight. Suitable if the underlying data is highly imbalanced.

Precision ○
The fraction of data points classified correctly among those classified as 1.

Recall ○
The fraction of data points classified correctly among those whose true label is 1. Alternative names: true positive rate, sensitivity.

Back | Next

```
Out[16]: <fairlearn.widget._fairlearn_dashboard.FairlearnDashboard at 0x7f07cad672d0>
```

Disparity in predictions

19.7% Is the overall selection rate | 4.36% Is the disparity in selection rate

gender	Selection rate
Female Max	22.1%
Male Min	17.7%

How to read this chart

The bar chart shows the selection rate in each group, meaning the fraction of points classified as 1.

```
    y_pred=classifier.predict(x_test))
```

Fairlearn

Disparity in performance

77.6% Is the overall accuracy | 7.25% Is the disparity in accuracy

Edit configuration

gender	Accuracy
Female Min	73.7%
Male Max	81%

How to read this chart

🟧 Underprediction (predicted = 0, true = 1)

🟦 Overprediction (predicted = 1, true = 0)

The bar chart shows the distribution of errors in each group.

Errors are split into overprediction errors (predicting 1 when the true label is 0), and underprediction errors (predicting 0 when the true label is 1).

The reported rates are obtained by dividing the number of errors by the overall group size.

Disparity in predictions

19.7% Is the overall selection rate | 4.36% Is the disparity in selection rate

gender | Selection rate

How to read this chart

The bar chart shows the selection rate in each group, meaning the fraction of points classified as 1.

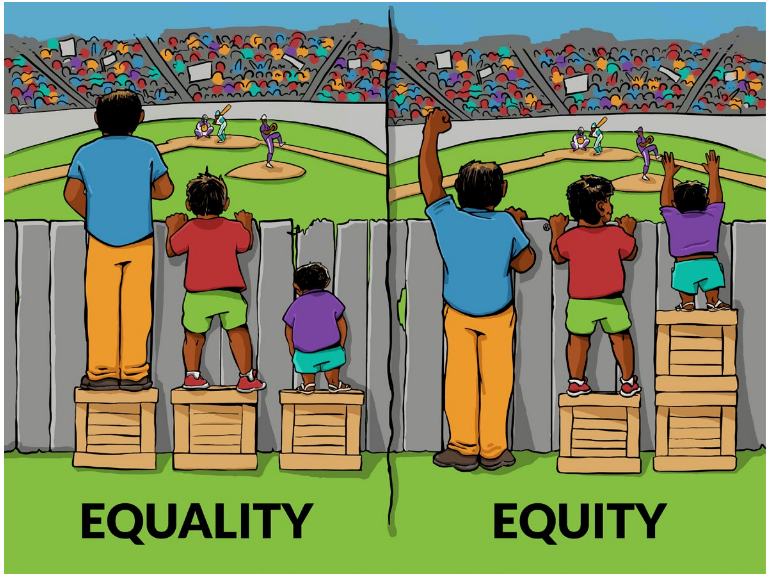
https://fairlearn.org/

Credit: https://github.com/wmeints/fairlearn-demo

Challenges?

Challenges

- Synthesizing a definition of fairness.
- From Equality to Equity
- Searching for Unfairness



https://interactioninstitute.org/illustrating-equality-vs-equity/

What to do this week?

- Read materials
- Work on Task 1 (Case Study), due week 7 (next week)
- Must watch Guest Lecture by Kate Crawford

The Trouble with Bias - NIPS2017

https://youtu.be/fMym_BKWQzk

Upcoming Weeks

- Week 7: Transparency and Explainability
- Week 8: Workshop – Presentation & Peer review of Task 1 (Case Study)
- Week 9: Guest Talk by A/Prof Richard Xu (**UTS**) on Data Science communication (stakeholder engagement) + Ethical ML as a Software Engineering enterprise (Charles Isbell's NeurIPS 2020 keynote)
- Week 10: Guest talk by Prof Milind Tambe (**Harvard**, Director "AI for Social Good“, **Google Research India**) on “AI for social good”
- Week 11: Guest talk by Ron Tidhar (**Instagram**) on “Data Science and Product Development”
- Week 12: Workshop – Presentation & Peer review of Task 2 (AI/DS for Social Good project)

References

[1] Mehrabi, Ninareh & Morstatter, Fred & Saxena, Nripsuta & Lerman, Kristina & Galstyan, Aram. (2019). A Survey on Bias and Fairness in Machine Learning.

[2] Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002 (2019).