Lecture 9:
Fairness in AI/ML
(plus a guest presentation from Shawn Shan)

CMSC 25910

Spring 2024

The University of Chicago

THE UNIVERSITY OF CHICAGO


NEWSNATION
Livestream Event
"A.I. Miracle or menace?"
NEWS
NATION

Algorithmic Decision Making

Proposal: Algorithmic Grading in 25910

- The data we have:

Name	Age	Department	Gender	Grade in CS 144	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 1

- Let's extrapolate from the Assignment 1 grade

Name	Age	Department	Gender	Grade in CS 144	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 1

- Small data! We also advertised something different!

Name	Age	Department	Gender	Grade in CS 144	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 2

- Let’s extrapolate from the CS 144 grade

Name	Age	Department	Gender	Grade in CS 144	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 2

- Is this just? Does Jane get a grade?

Name	Age	Department	Gender	Grade in CS 144	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 3

- Let's use Department and the Grade in CS 154

Name	Age	Department	Gender	Grade in CS 144	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 3

- Why should these matter?

Name	Age	Department	Gender	Grade in CS 144	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 4

- Let’s use all demographics and the Grade in CS 144

Name	Age	Department	Gender	Grade in CS 144	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 4

- Why?!?! (Also, age and gender are protected classes)

Name	Age	Department	Gender	Grade in CS 144	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 4

- Also consider the mutability of characteristics / recourse

Name	Age	Department	Gender	Grade in CS 144	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 5

- Everyone gets an A!

Name	Age	Department	Gender	Grade in CS 144	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 6

- Everyone gets an F!

Name	Age	Department	Gender	Grade in CS 144	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Idea 6

- Societal notions of justice may imply that failing everyone is bad

Name	Age	Department	Gender	Grade in CS 144	Grade on Assignment 1
Jack	55	CS	M	B+	100
Jill	23	Econ	F	A	95
Josh	32	Bio	M	B	50
Jenn	44	Bio	F	A-	98
Jane	27	Stats	F	---	80

Bias in Algorithmic Decision Making

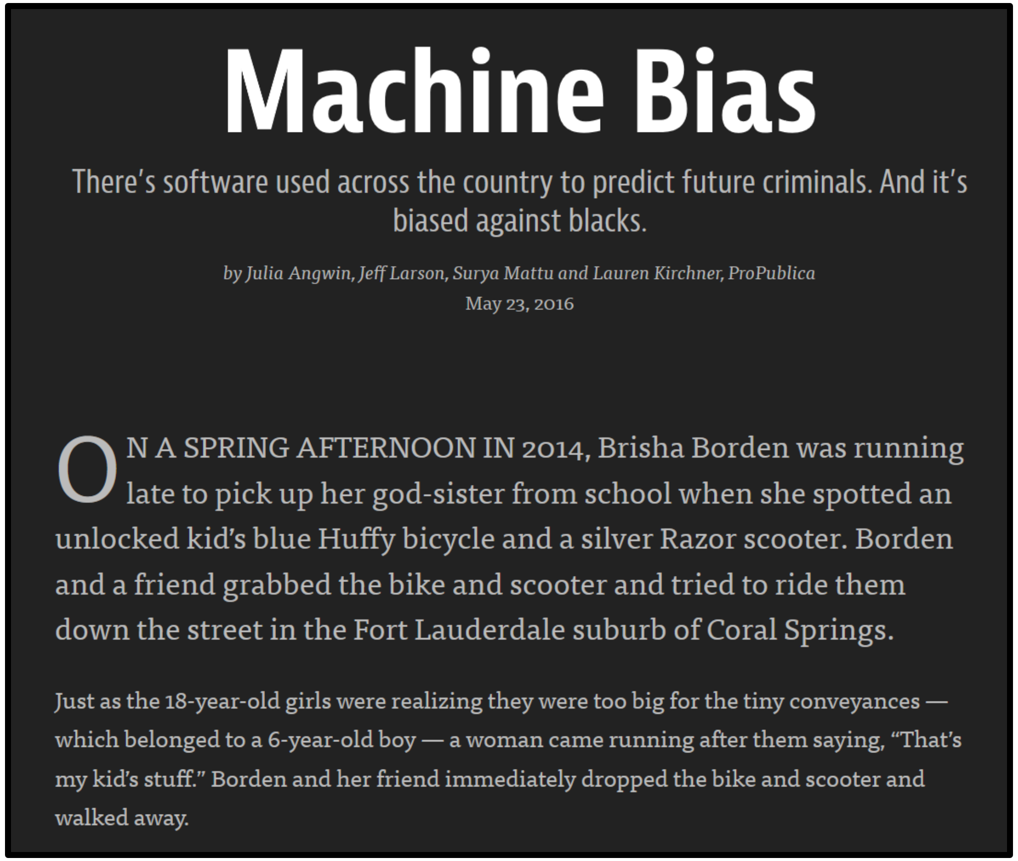
Machine Bias (ProPublica)

- COMPAS System for risk assessment
- Based on answers to 137 questions
- ProPublica obtained data:
    - Broward County, Florida

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Machine Bias (ProPublica)

- COMPAS System for risk assessment
- Based on answers to 137 questions
- ProPublica obtained data:
  - Broward County, Florida
- “And it’s biased against blacks.”
  - Northpointe: It’s equally accurate across demographic groups!



Machine Bias

There’s software used across the country to predict future criminals. And it’s biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

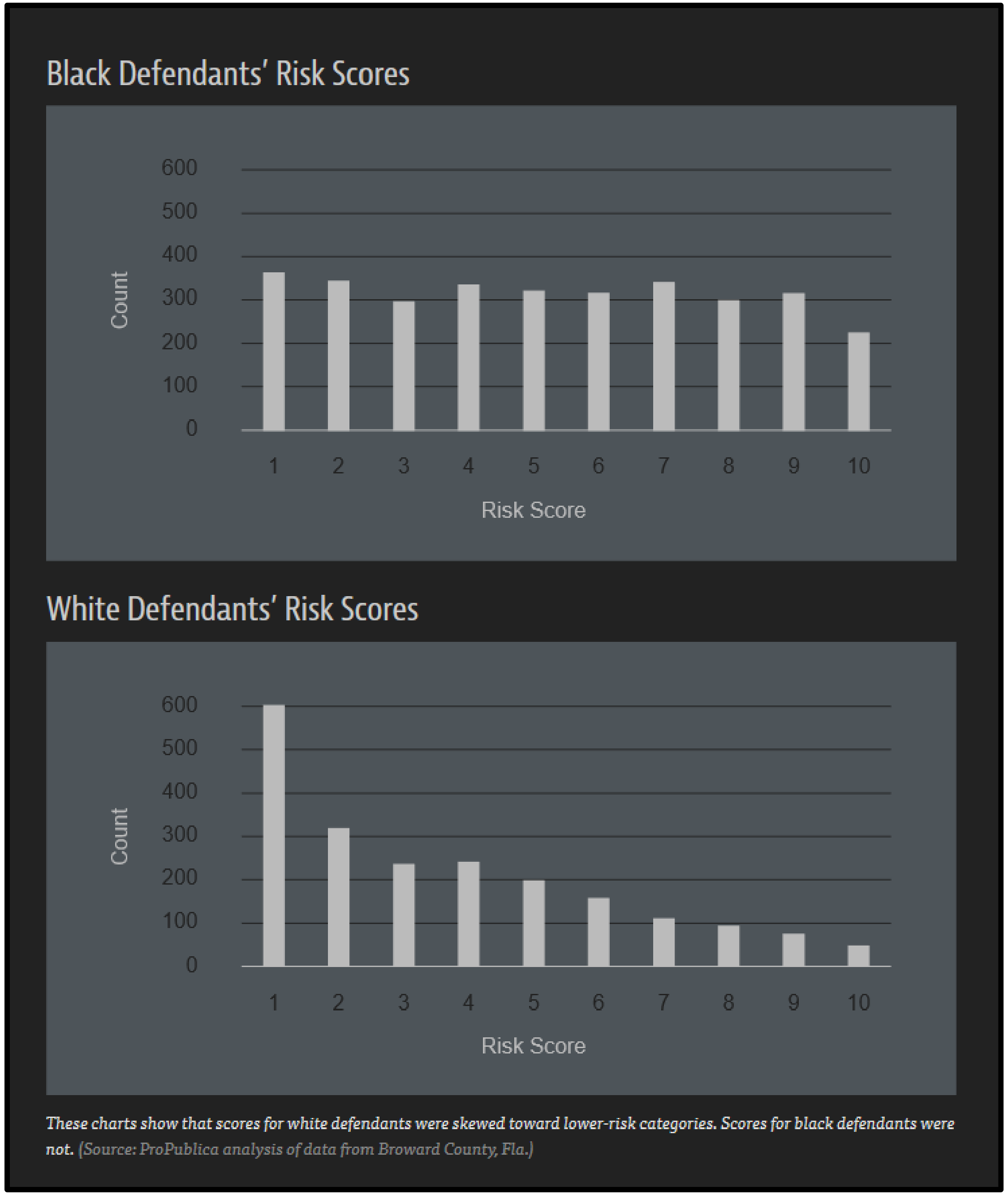
ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid’s blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, “That’s my kid’s stuff.” Borden and her friend immediately dropped the bike and scooter and walked away.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

COMPAS

- Evidence of discrimination?



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

COMPAS

- Evidence of discrimination?

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

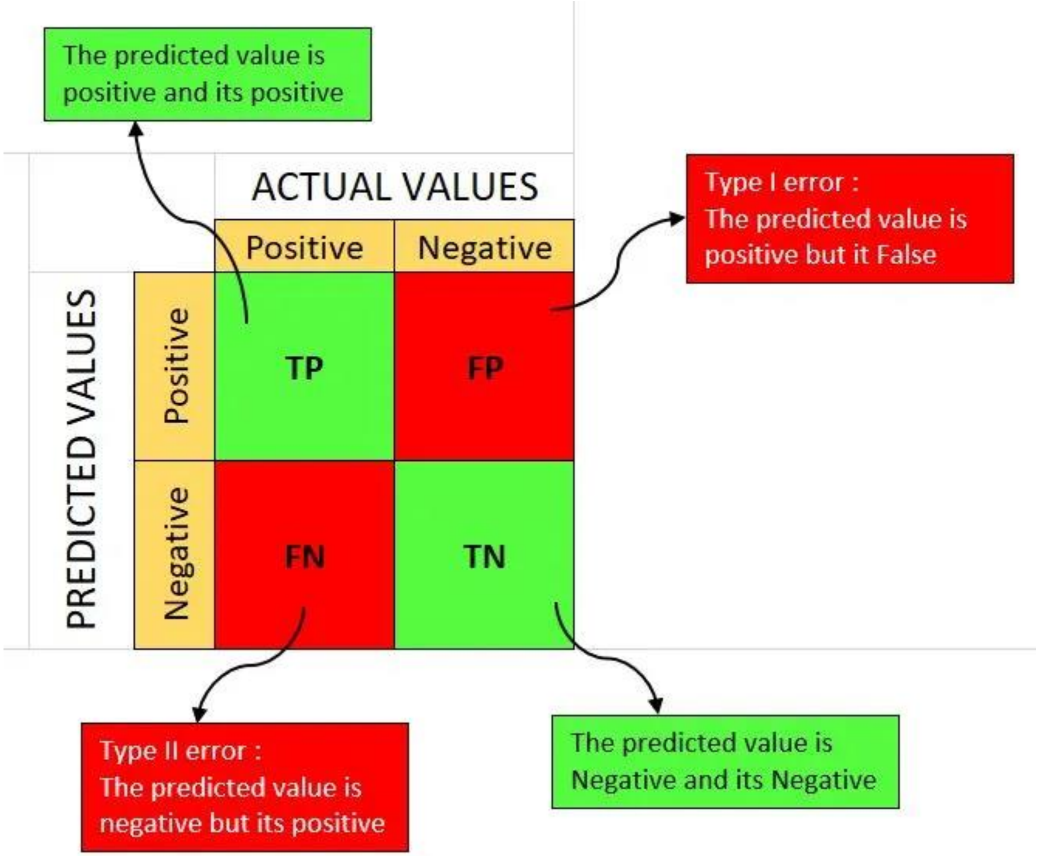
Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

ML Metrics (And Their Connection to Fairness)

Some Possible Metrics (Classifiers)

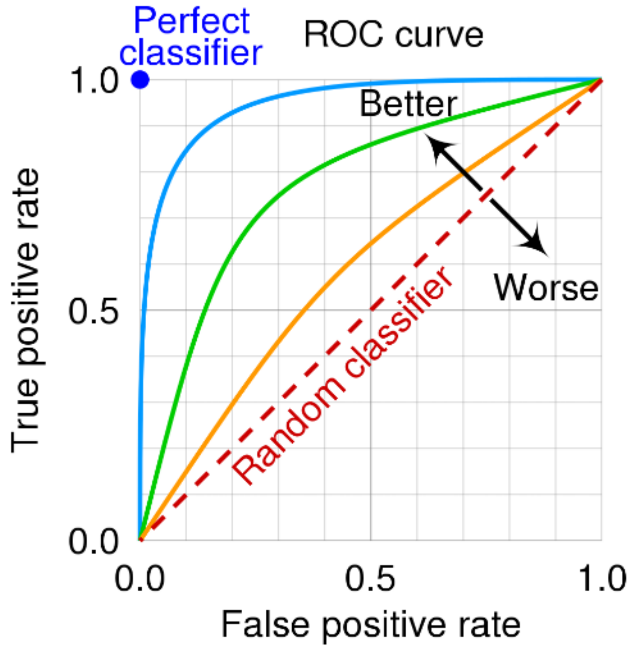
- Accuracy: # correct / # total
- Confusion matrix (TP/FP/TN/FN)
  - Binary classifier
  - Positive and negative classes
  - True = prediction matched ground truth
  - **T**rue **P**ositive
  - **T**rue **N**egative
  - **F**alse **P**ositive
  - **F**alse **N**egative



See https://medium.com/analytics-vidhya/performance-metrics-for-machine-learning-models-80d7666b432e
https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics.htm
https://www.justintodata.com/machine-learning-model-evaluation-metrics/ or many more!
Confusion matrix image taken from https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

Some Possible Metrics (Classifiers)

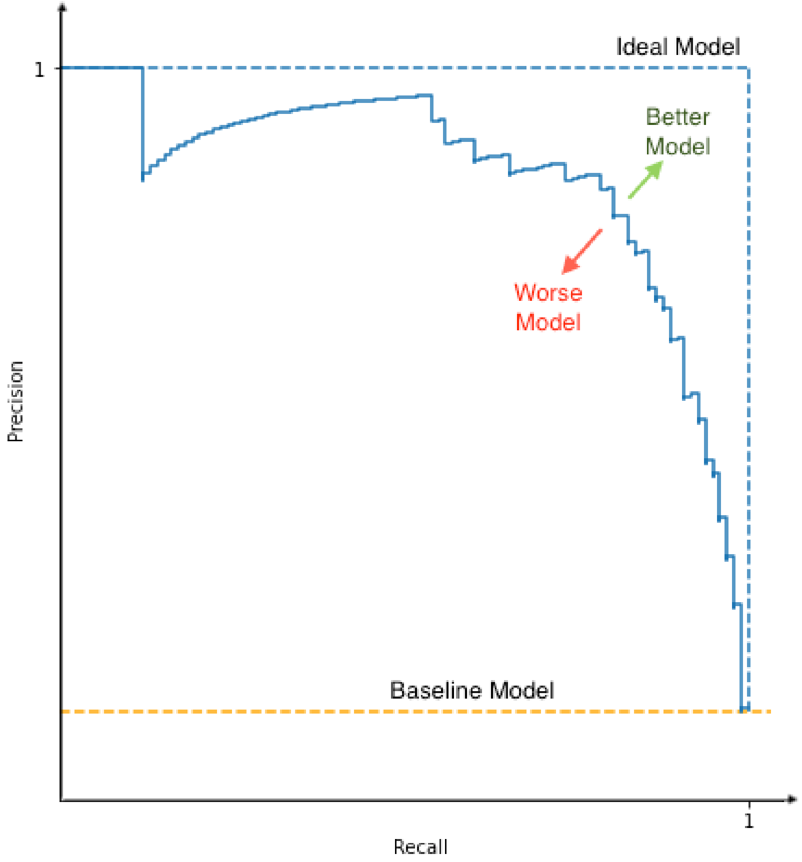
- Receiver Operating Characteristic (ROC) Curve
  - True Positive Rate (TPR) = TP / P = TP / (TP + FN)
  - False Positive Rate (FPR) = FP / N = FP / (FP + TN)
  - ROC curve plots TPR vs. FPR at various thresholds
  - Area under the ROC curve (AUC) is a common metric



See https://medium.com/analytics-vidhya/performance-metrics-for-machine-learning-models-80d7666b432e
https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics.htm
https://www.justintodata.com/machine-learning-model-evaluation-metrics/ or many more!
ROC curve image taken from https://en.wikipedia.org/wiki/Receiver_operating_characteristic#/media/File:Roc_curve.svg

Some Possible Metrics (Classifiers)

- Precision: TP / (TP + FP)
- Recall: TP / (TP + FN)
- Precision-Recall Curve



See https://medium.com/analytics-vidhya/performance-metrics-for-machine-learning-models-80d7666b432e
https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics.htm
https://www.justintodata.com/machine-learning-model-evaluation-metrics/ or many more!
Precision-recall curve image taken from https://towardsai.net/p/l/precision-recall-curve

Some Possible Metrics (Regressions)

- Mean Squared Error
- Mean Absolute Error

See https://medium.com/analytics-vidhya/performance-metrics-for-machine-learning-models-80d7666b432e
https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_algorithms_performance_metrics.htm
https://www.justintodata.com/machine-learning-model-evaluation-metrics/ or many more!

Some Possible Metrics (Performance)

- Model training time
- Frequency of model re-training
- Model size
- Classification time
- Privacy issues of the model
- “Security” (future lecture)

Some Possible Metrics **Revisited**

- Do these metrics capture the **relationship** between **errors?**
- Do these metrics capture the **impact of errors?**
- Do these metrics capture the **differential** impact of **particular types of errors?**
- Do these metrics break down **errors by group**?
- We calculate errors on our **test set**; what about **in practice**?
  - Do we have enough data in different sub-groups?
  - Do we have representative data? How do we define representative?
- Where is the data even coming from? How accurate is it?

Defining Fairness

The Difficulty of Defining Fairness

- Terminology is conflated across disciplines
  - Political philosophy
  - Employment law
  - Computer science

- See: Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, Richmond Y. Wong. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. PACM HCI (CSCW), 2019.

Individual Fairness

- One of the early definitions of fairness
- **Individual fairness:** Similar people should be treated equally

Statistical Non-Discrimination

- Basis in employment and housing law (e.g., Fair Housing Act)
- Primarily considers *protected classes*
  - Race, gender, sex, national origin, religion, marital status, etc.
- In this approach to fairness, we want to approximately equalize some quantities across demographic groups (**group fairness**)
  - Mainly focuses on **disparate impact** (treating different groups differently)

Group Fairness (Just a Few Approaches)

- Demographic parity (**equal outcomes**)
  - Equalize the chance of positive classifications across groups

Group Fairness (Just a Few Approaches)

- **Equalized accuracy** across groups?

Group Fairness (Just a Few Approaches)

- **Equalized odds** (true positive rate and false positive rate are equal across groups)?
  - True Positive Rate (TPR) = TP / P = TP / (TP + FN)
  - False Positive Rate (FPR) = FP / N = FP / (FP + TN)

The Need to Make Tough Trade-offs

- A. Chouldechova. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” *Big Data 2017*.
- J. Kleinberg, S. Mullainathan, M. Raghavan. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” *ITCS 2017*.
    - “Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.”

Blindness to Protected Classes

- Should we just intentionally not collect data about whether or not data subjects belong to a protected class?
  - The answer is very complicated. It’s often (but not always!) “no”… why not?

Process Fairness

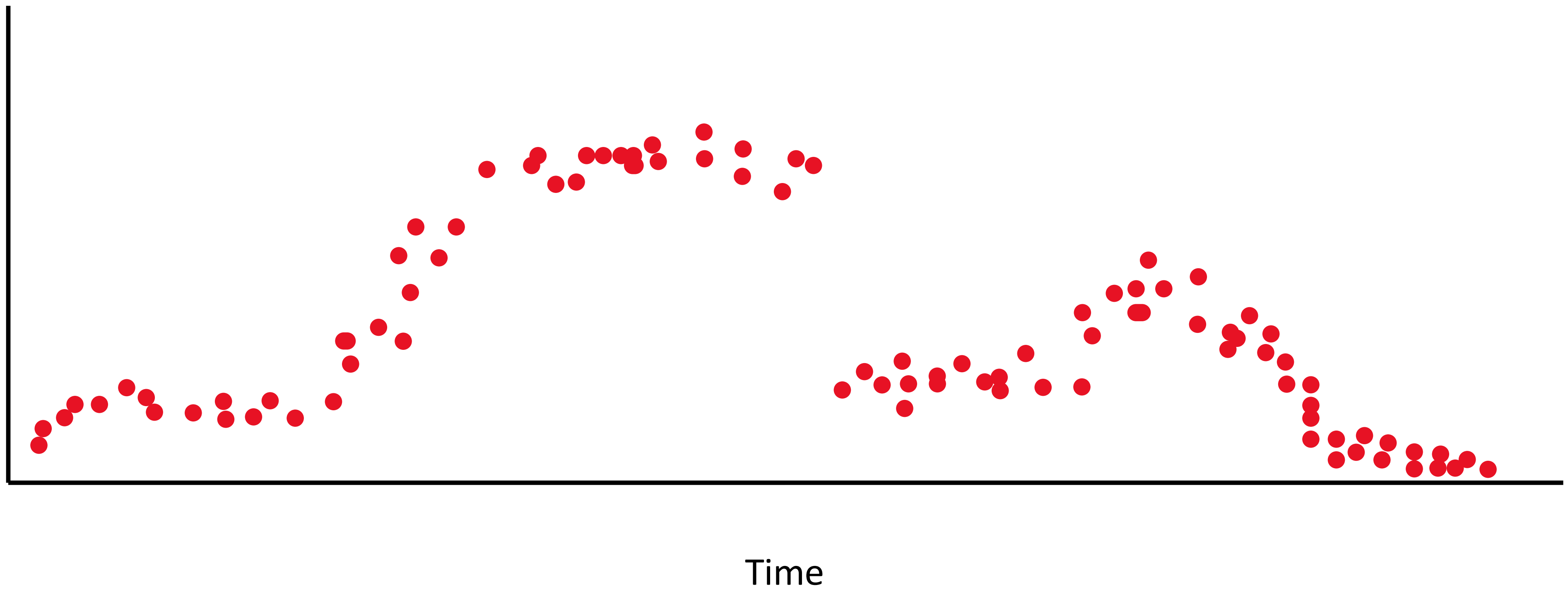
- How do we decide what predictor variables to include?
- **Process fairness**: Exclude from the model predictor variables that are *deemed to be* unfair for the classification task
- Should we just crowdsource perceptions?
  - Grgic-Hlaca et al. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proc. WWW*, 2018.
  - Important question: Who gets to decide what is fair? Is it majoritarian voting? Should it be experts in law/technology?

How Does Sampling Impact Fairness?

- What if our sample is unbalanced? Can that cause problems?
- What if our sample is not representative?
- What if we collect the wrong features?

Concept Drift – The Passage of Time

- Can we be embedding historical biases?



Reconceptualizing Fairness as Justice

- Should we follow Rawls and consider justice as fairness?
- Should we start thinking about fairness in terms of trolley problems? https://en.wikipedia.org/wiki/Trolley_problem
- How might our societal notions of what is just change how we build a classifier, **as well as whether we use ML at all**?
- How do we think about due process within fairness?
- Returning to the COMPAS example: How did **human judges** use (or choose not to use) COMPAS risk scores? Is this just?
- Accountability? Transparency? Explanations?

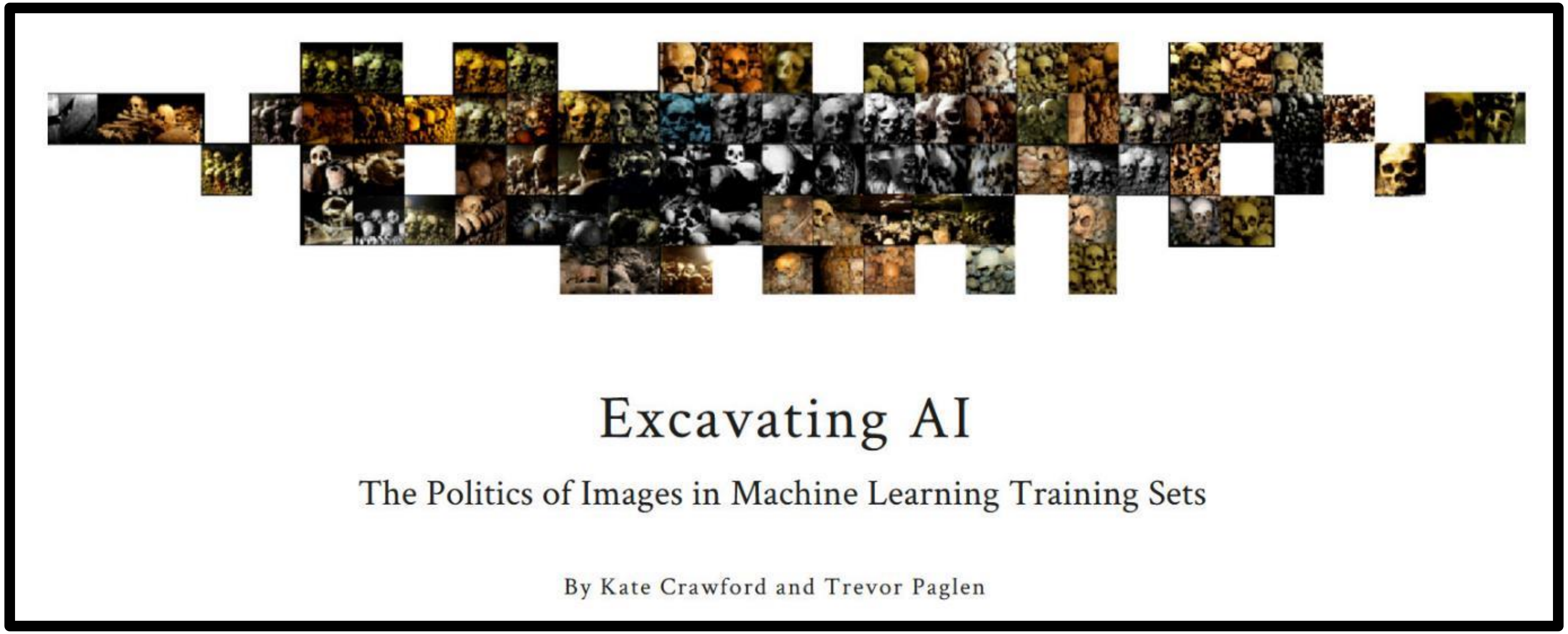
Training Data

Training Datasets and Benchmarks

- Standardization of training datasets and benchmarks have arguably pushed the field of ML forward
  - Not without pitfalls
- If everyone is testing against the same datasets, what does that say about the ML model’s generalizability?
  - Are results practically significant?
  - Do we notice errors that occur for data **excluded** from reference sets?
- There are more serious problems than a lack of progress!

What Datasets Include/Exclude

- *Kate Crawford and Trevor Paglen, “Excavating AI: The Politics of Training Sets for Machine Learning (September 19, 2019)*
- https://excavating.ai



What Datasets Include/Exclude

- “The automated interpretation of images is an inherently social and political project, rather than a purely technical one”
- “What work do images do in AI systems? What are computers meant to recognize in an image and what is misrecognized or even completely invisible?”
- “How do humans tell computers which words will relate to a given image? And what is at stake in the way AI systems use these labels to classify humans, including by race, gender, emotions, ability, sexuality, and personality?”
- “As the fields of information science and science and technology studies have long shown, all taxonomies or classificatory systems are political.”

Taken from Kate Crawford and Trevor Paglen, “Excavating AI: The Politics of Training Sets for Machine Learning (September 19, 2019)

What Datasets Include/Exclude

“There is much at stake in the architecture and contents of the training sets used in AI. They can promote or discriminate, approve or reject, render visible or invisible, judge or enforce. And so we need to examine them—because they are already used to examine us—and to have a wider public discussion about their consequences, rather than keeping it within academic corridors. As training sets are increasingly part of our urban, legal, logistical, and commercial infrastructures, they have an important but underexamined role: the power to shape the world in their own images.”

Taken from Kate Crawford and Trevor Paglen, “Excavating AI: The Politics of Training Sets for Machine Learning (September 19, 2019)

Imagenet: Computer Vision dataset

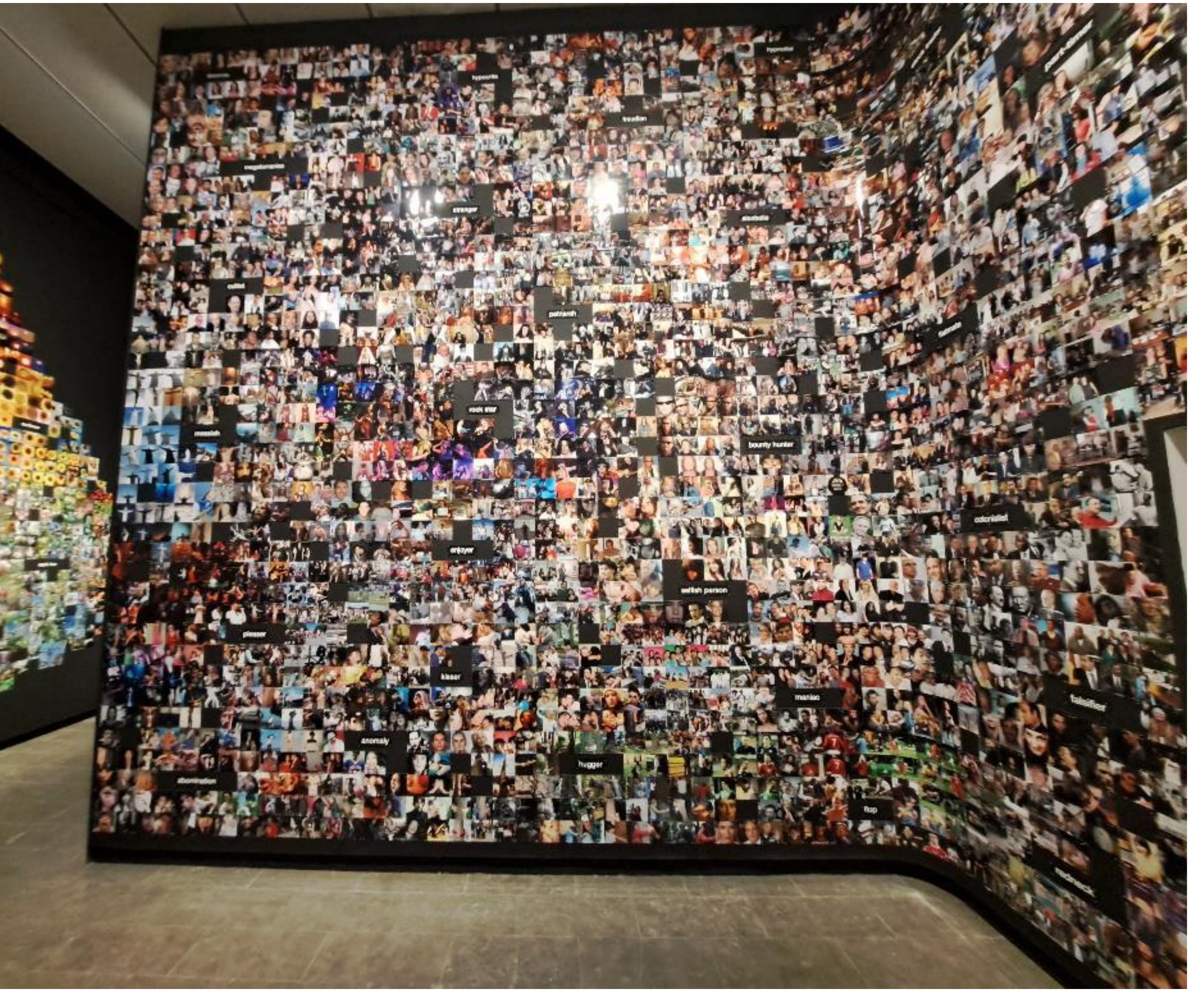
- 15 million images
  - Each image is annotated with a noun from Wordnet
    - Wordnet -> hierarchy of concepts
- Instrumental dataset to advance computer vision
- Where did these images come from?

Trevor Paglen’s Art About ImageNet

- Trevor Paglen, “From ‘Apple’ to ‘Abomination’” (2023)
- “This work is composed of more than 13,000 images from ImageNet, a training set with more than 20,000 categories totaling over 14 million images… The work spotlights the systems commonly inherent in such software and affecting us all, questioning the arbitrary connections between images and words, and the problems they create.”

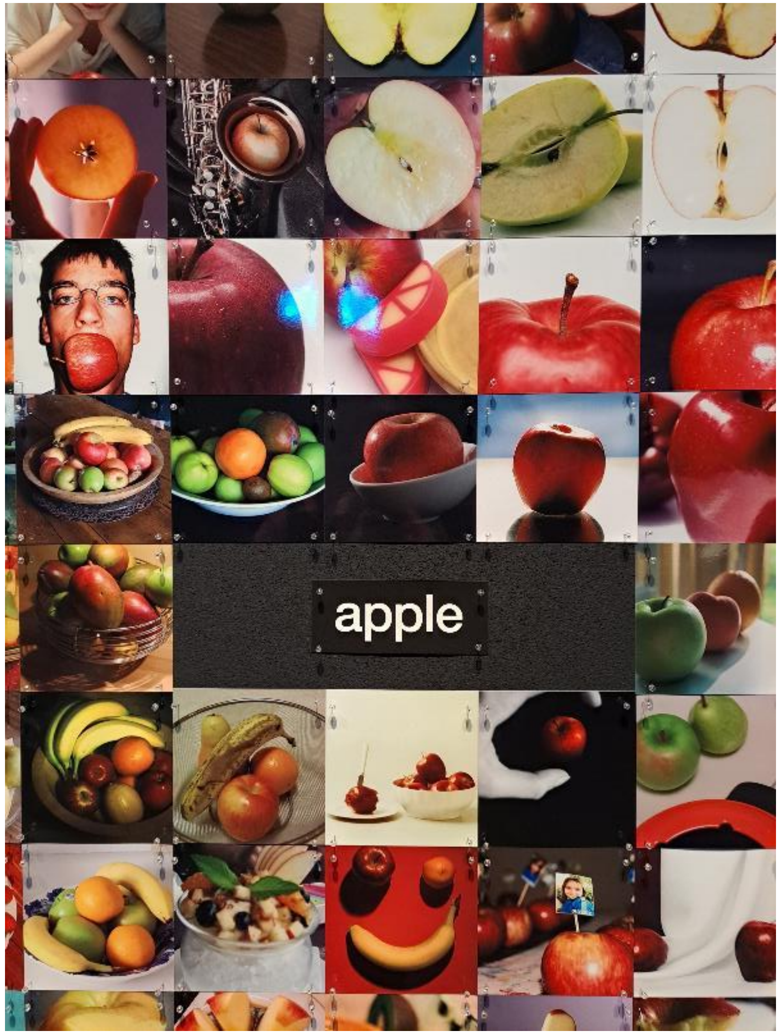
Taken from Trevor Paglen, “From ‘Apple’ to ‘Abomination’” (2023), photograhed by me at the Louisiana Museum of Modern Art, Denmark

Trevor Paglen’s Art About ImageNet



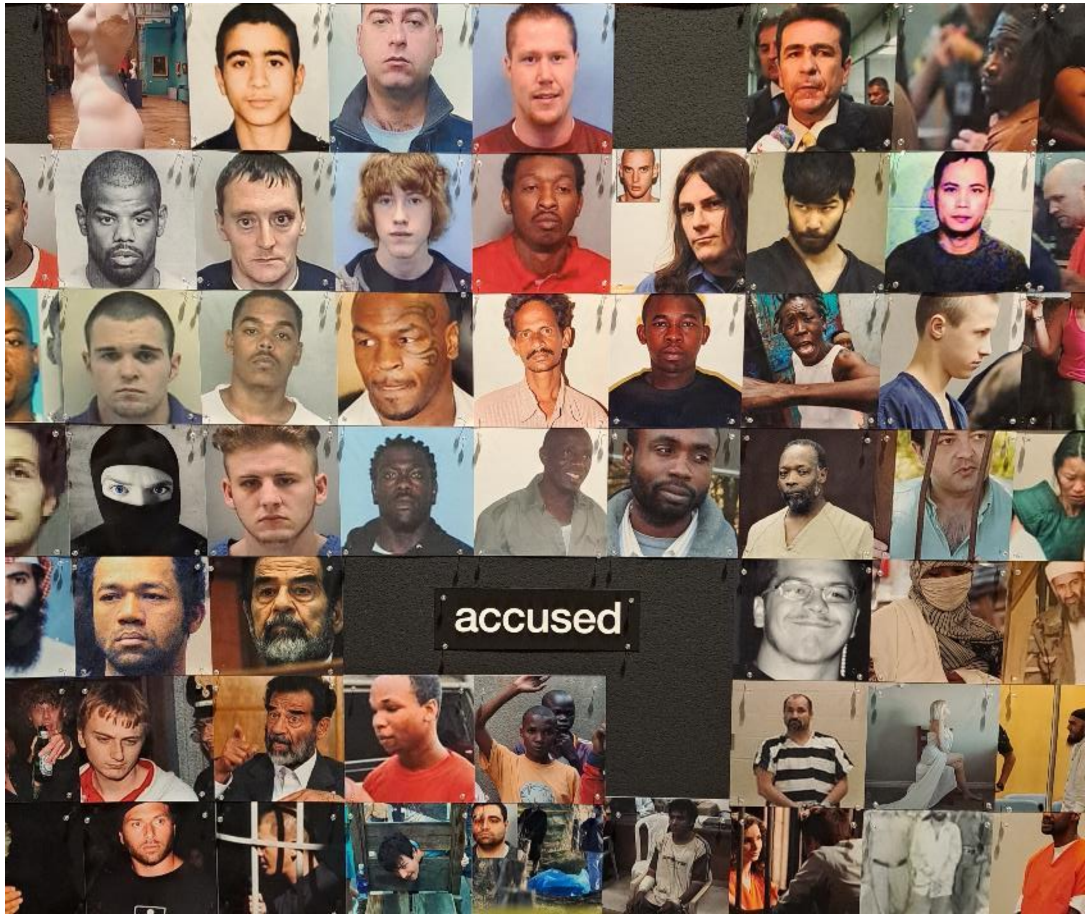
Taken from Trevor Paglen, “From ‘Apple’ to ‘Abomination’” (2023), photograhed by me at the Louisiana Museum of Modern Art, Denmark

Trevor Paglen's Art About ImageNet



Taken from Trevor Paglen, "From 'Apple' to 'Abomination'" (2023), photograhed by me at the Louisiana Museum of Modern Art, Denmark

Trevor Paglen's Art About ImageNet



Taken from Trevor Paglen, "From 'Apple' to 'Abomination'" (2023), photograhed by me at the Louisiana Museum of Modern Art, Denmark

Where Do Labels Come From?

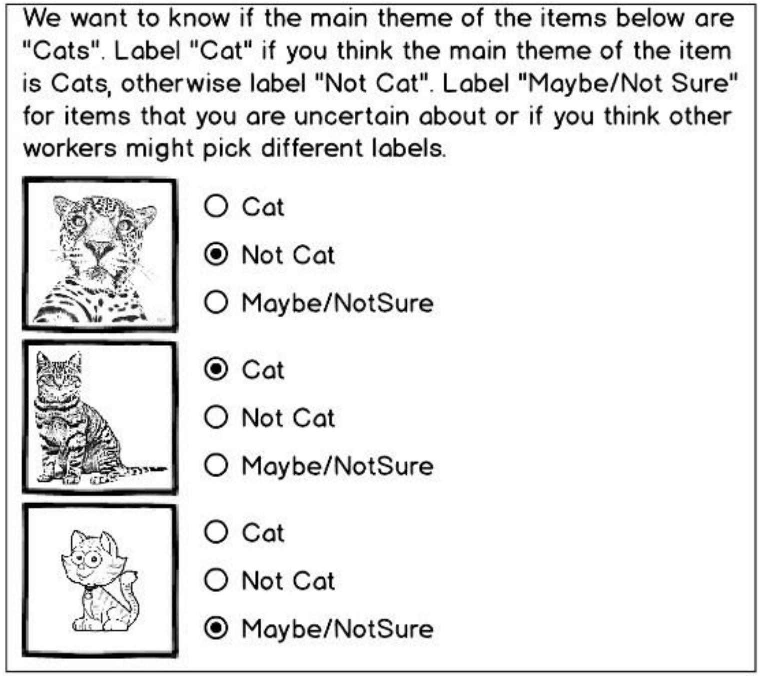


Figure 3. Human Intelligence Task (HIT) interface for the Vote Stage. In addition to the predefined labels, crowdworkers can also select *Maybe/NotSure* when they were uncertain about the item.

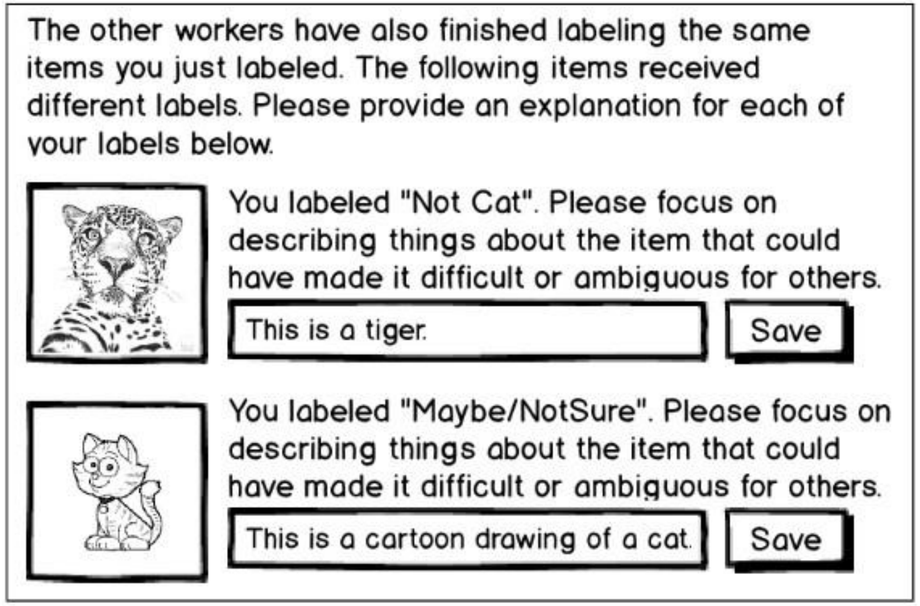


Figure 4. Human Intelligence Task (HIT) interface for the Explain Stage. Crowdworkers enter a short description for each item that was labeled differently in the Vote Stage. They were informed that disagreement occurred, but not the distribution of different labels used.

Chang et al. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. CHI 2017
https://dl.acm.org/doi/pdf/10.1145/3025453.3026044

Attempts at Mitigating Fairness Concerns

Some Attempted Fairness Mitigations

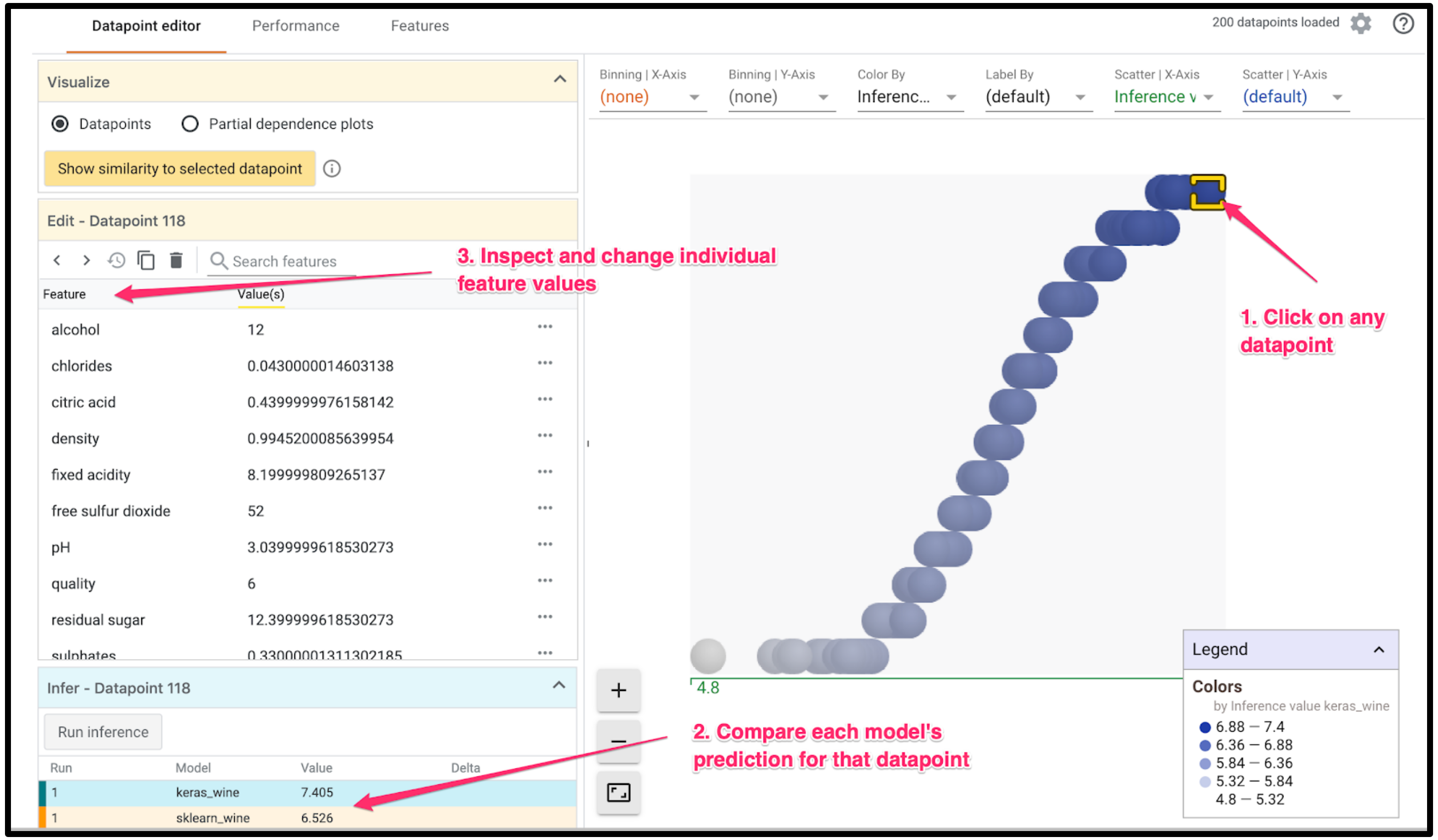
- Transform the training data features and/or labels
- Change the weights in the model produced
- Adversarial de-biasing
    - e.g., using a discriminator from a Generative Adversarial Network

AI Fairness 360

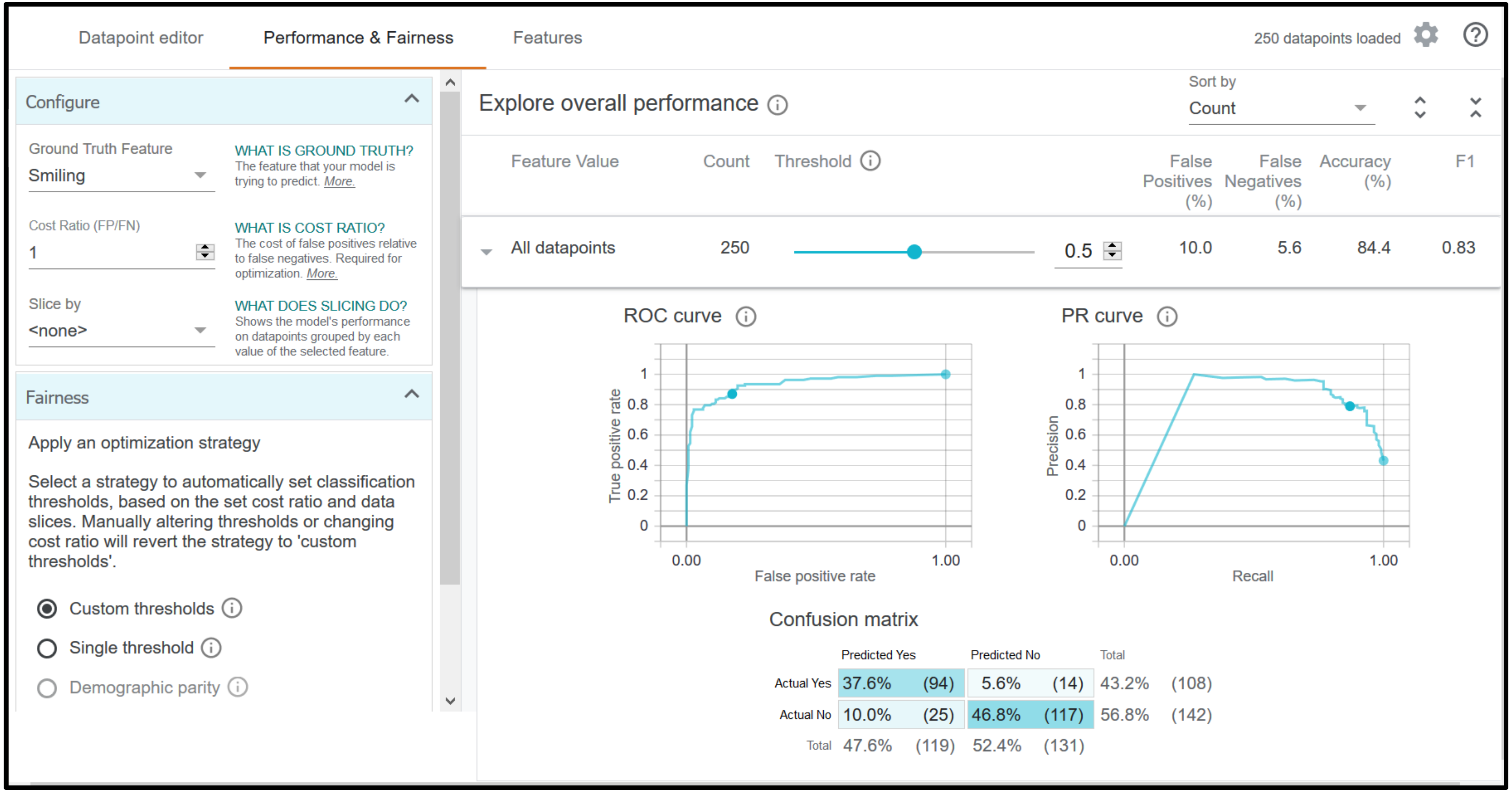
- IBM open source project: https://aif360.mybluemix.net/
- Online demo: https://aif360.mybluemix.net/data

What-If Tool

- Google open source project: https://pair-code.github.io/what-if-tool/
- Online demo: https://pair-code.github.io/what-if-tool/image.html



What-If Tool



Datapoint editor
Performance & Fairness
Features
250 datapoints loaded
Configure
Ground Truth Feature
Smiling
WHAT IS GROUND TRUTH?
The feature that your model is trying to predict. More.
Cost Ratio (FP/FN)
1
WHAT IS COST RATIO?
The cost of false positives relative to false negatives. Required for optimization. More.
Slice by
<none>
WHAT DOES SLICING DO?
Shows the model's performance on datapoints grouped by each value of the selected feature.
Fairness
Apply an optimization strategy
Select a strategy to automatically set classification thresholds, based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will revert the strategy to 'custom thresholds'.
Custom thresholds
Single threshold
Demographic parity
Explore overall performance
Sort by
Count
Feature Value
Count
Threshold
False Positives (%)
False Negatives (%)
Accuracy (%)
F1
All datapoints
250
0.5
10.0
5.6
84.4
0.83
ROC curve
True positive rate
False positive rate
PR curve
Precision
Recall
Confusion matrix
Predicted Yes
Predicted No
Total
Actual Yes
37.6% (94)
5.6% (14)
43.2% (108)
Actual No
10.0% (25)
46.8% (117)
56.8% (142)
Total
47.6% (119)
52.4% (131)

Aequitas Tool

- Formerly a UChicago open source project: http://www.datasciencepublicpolicy.org/projects/aequitas/
- Online demo: http://aequitas.dssg.io/example.html

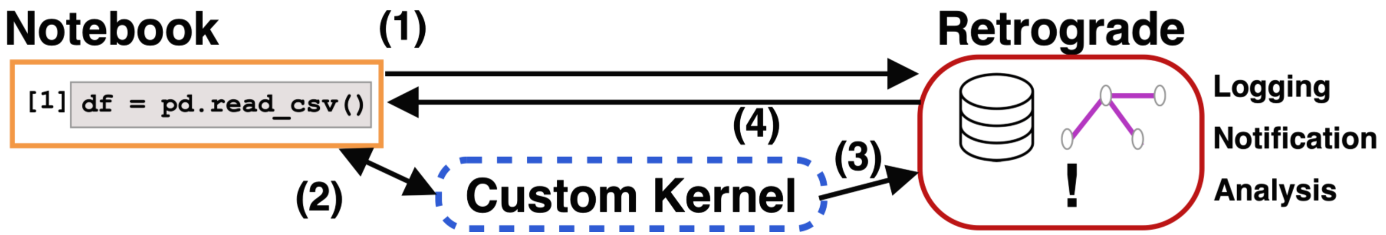
Audit Results: Bias Metrics Values

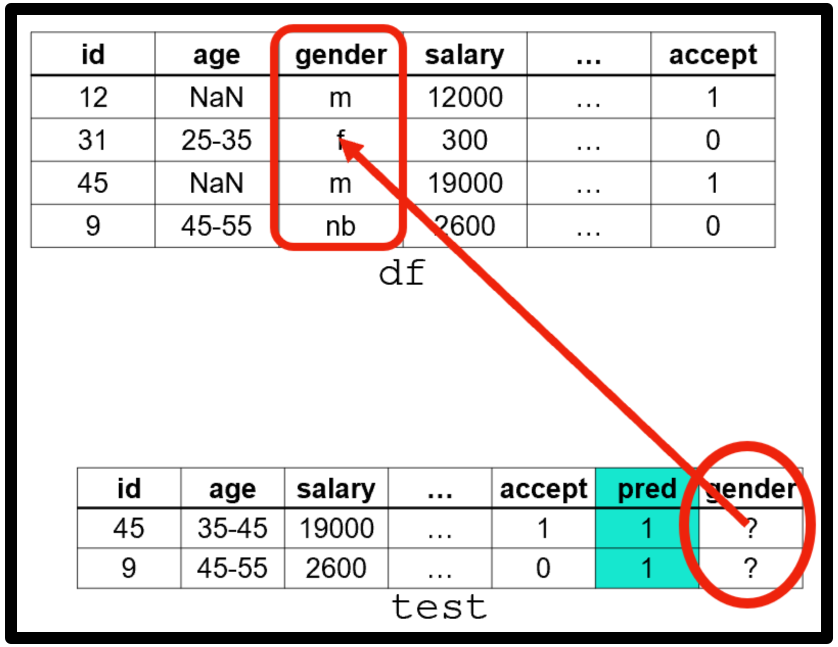
race

Attribute Value	False Discovery Rate Disparity	False Positive Rate Disparity	False Negative Rate Disparity
African-American	0.91	1.91	0.59
Asian	0.61	0.37	0.7
Caucasian	1.0	1.0	1.0
Hispanic	1.12	0.92	1.17
Native American	0.61	1.6	0.21
Other	1.12	0.63	1.42

Retrograde

- Improved techniques for tracking provenance in computational notebooks (JupyterLab)
- Design of data-driven contextual nudges
- Evaluation study (51 data scientists)





id	age	gender	salary	…	accept
12	NaN	m	12000	…	1
31	25-35	f	300	…	0
45	NaN	m	19000	…	1
9	45-55	nb	2600	…	0

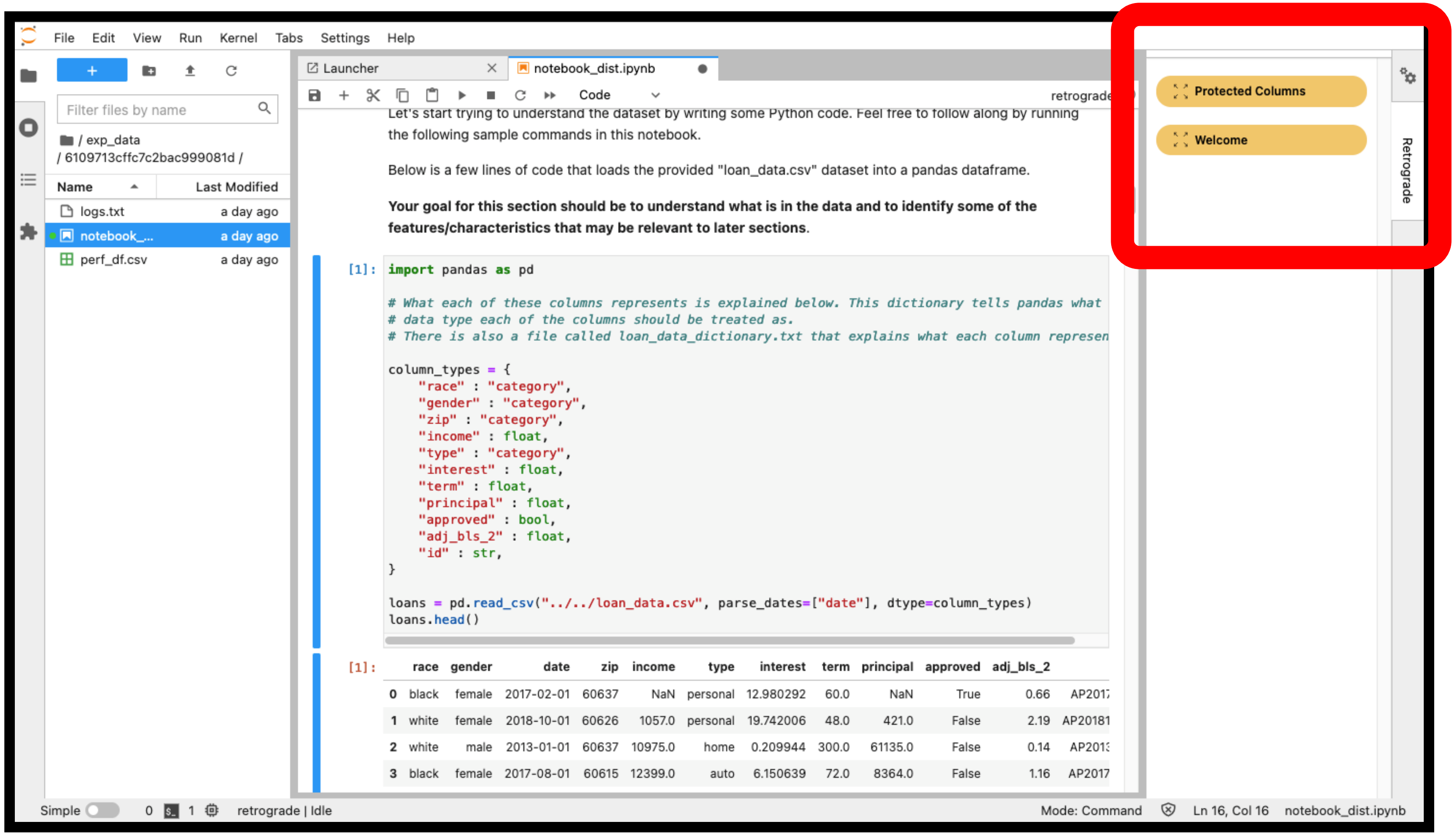
df

id	age	salary	…	accept	pred	gender
45	35-45	19000	…	1	1	?
9	45-55	2600	…	0	1	?

test

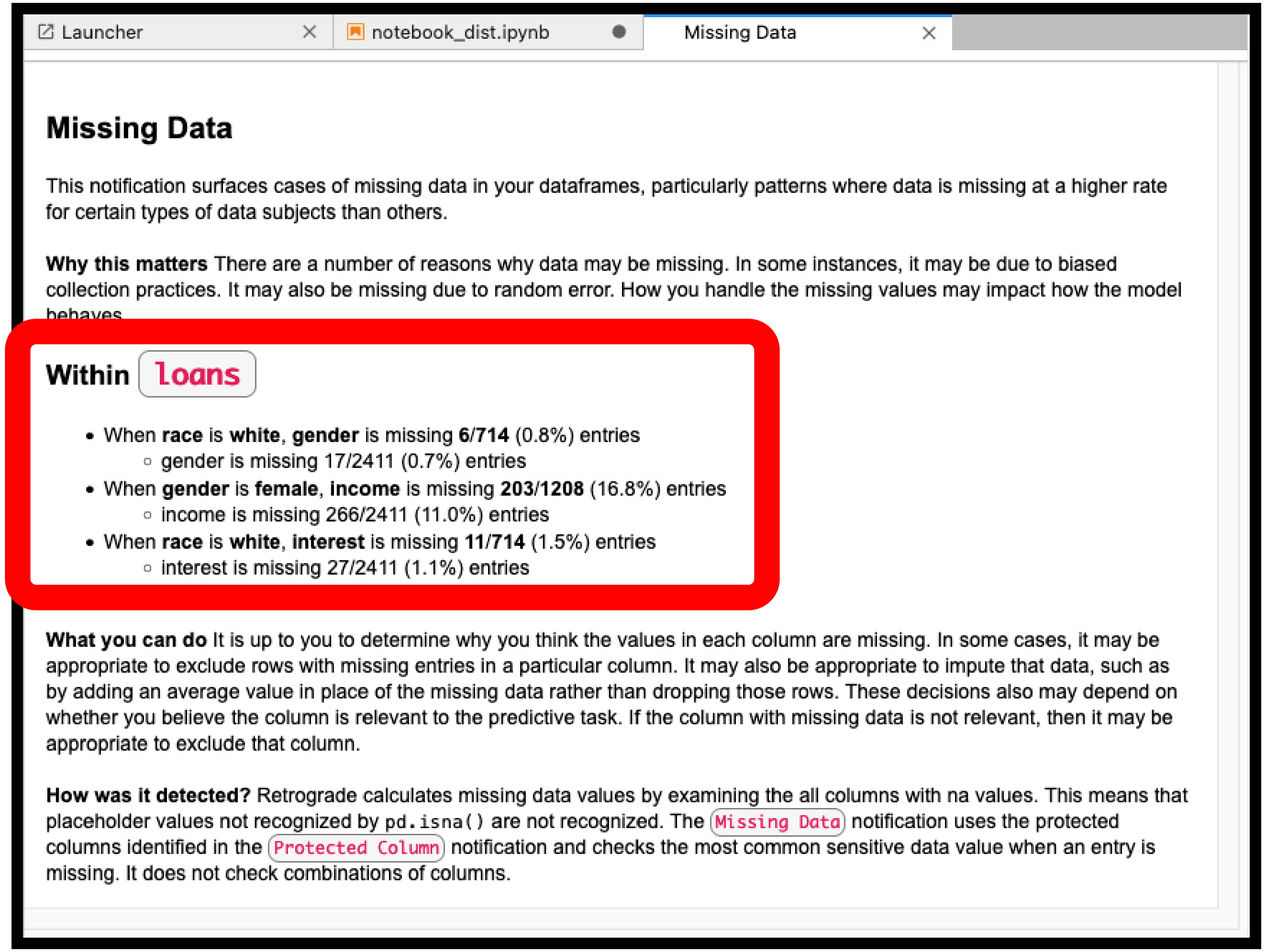
See Galen Harrison, Kevin Bryson, Ahmad Bamba, Luca Dovichi, Alek Binion, Arthur Borem, and Blase Ur. JupyterLab in Retrograde: Contextual Notifications that Highlight Fairness and Bias Issues for Data Scientists. In *Proc. CHI*, 2024

Retrograde



See Galen Harrison, Kevin Bryson, Ahmad Bamba, Luca Dovichi, Alek Binion, Arthur Borem, and Blase Ur. JupyterLab in Retrograde: Contextual Notifications that Highlight Fairness and Bias Issues for Data Scientists. In *Proc. CHI*, 2024

Retrograde



Missing Data

This notification surfaces cases of missing data in your dataframes, particularly patterns where data is missing at a higher rate for certain types of data subjects than others.

Why this matters There are a number of reasons why data may be missing. In some instances, it may be due to biased collection practices. It may also be missing due to random error. How you handle the missing values may impact how the model behaves.

Within `loans`

- When **race** is **white**, **gender** is missing **6/714** (0.8%) entries
  - gender is missing 17/2411 (0.7%) entries
- When **gender** is **female**, **income** is missing **203/1208** (16.8%) entries
  - income is missing 266/2411 (11.0%) entries
- When **race** is **white**, **interest** is missing **11/714** (1.5%) entries
  - interest is missing 27/2411 (1.1%) entries

What you can do It is up to you to determine why you think the values in each column are missing. In some cases, it may be appropriate to exclude rows with missing entries in a particular column. It may also be appropriate to impute that data, such as by adding an average value in place of the missing data rather than dropping those rows. These decisions also may depend on whether you believe the column is relevant to the predictive task. If the column with missing data is not relevant, then it may be appropriate to exclude that column.

How was it detected? Retrograde calculates missing data values by examining the all columns with na values. This means that placeholder values not recognized by `pd.isna()` are not recognized. The `Missing Data` notification uses the protected columns identified in the `Protected Column` notification and checks the most common sensitive data value when an entry is missing. It does not check combinations of columns.

See Galen Harrison, Kevin Bryson, Ahmad Bamba, Luca Dovichi, Alek Binion, Arthur Borem, and Blase Ur. JupyterLab in Retrograde: Contextual Notifications that Highlight Fairness and Bias Issues for Data Scientists. In *Proc. CHI*, 2024

Retrograde

Launcher | notebook_dist.ipynb | Proxy Columns | Protected Columns

Proxy Columns

Some columns (variables) in your dataframe are correlated with protected classes. These are called **proxy variables.** Below, we list the correlation coefficients (Spearman's rho [ρ], Chi-Square [χ^2], or ANOVA [F]). Correlation coefficients close to 0 indicate no correlation, whereas those close to 1 or -1 indicate a high degree or positive or negative correlation.

Why it matters Using proxy variables as predictors in your model may unintentionally base the model's decisions on protected classes like race and gender even if you exclude those sensitive variables from the model.

What you can do It is up to you to decide whether to include proxy variables (or even protected classes themselves) as predictors in your model. The correlations identified here may or may not be meaningful. There also may be more complex correlations that weren't detected. In some cases, a variable's predictive value may outweigh its correlation with a protected class; in other cases, it might not.

Ultimately, it is up to you to make a decision about whether it is valid to include the correlated columns in your model.

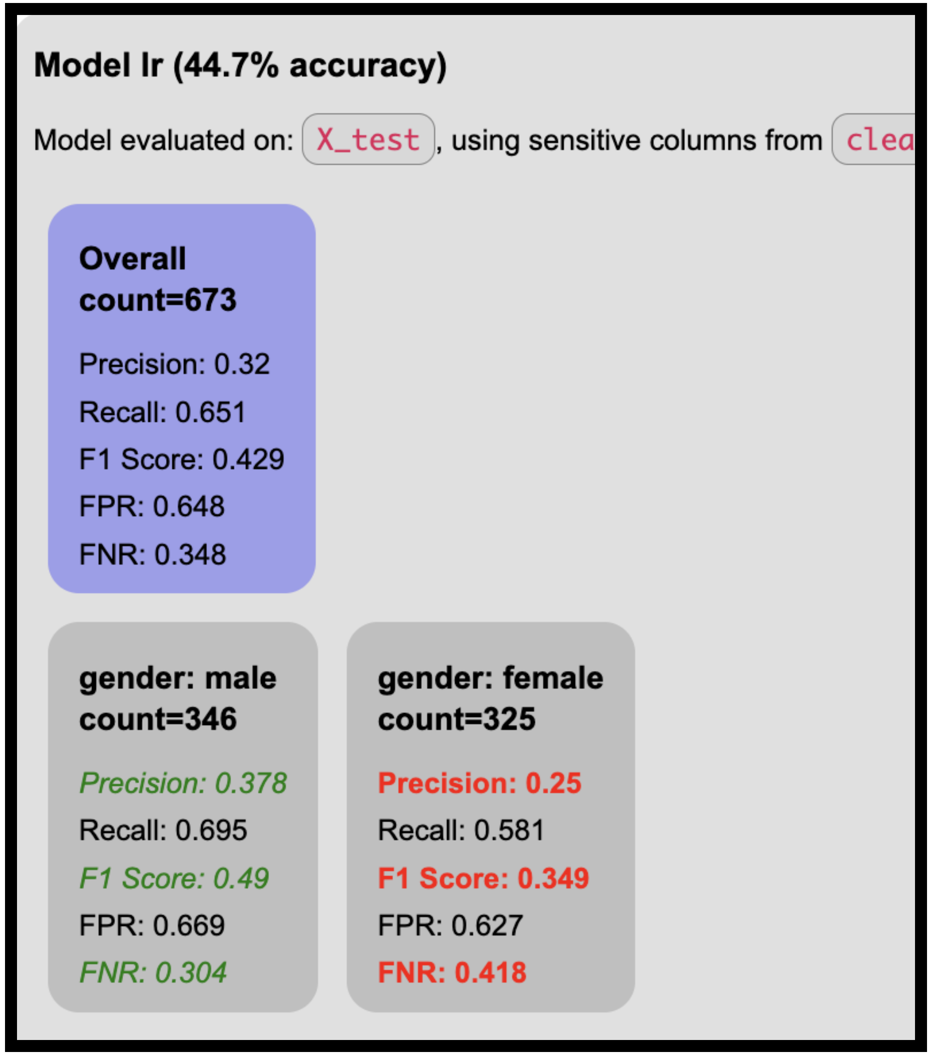
Within `loans`

Column name	Significantly correlated columns (p < 0.001)	Potentially correlated columns (p < 0.25)
gender	adj_bls_2 (F = 1.54), approved (F = 3.33), principal (F = 3.95)	income (F = 28.84)
race	term (F = 2.25), type (χ^2 = 45.31)	approved (F = 24.9), income (F = 15.66), principal (F = 8.52), zip (χ^2 = 1406.74)

How was it detected? Retrograde calculates these values by comparing every sensitive column with every non-sensitive column. Based on the data types of the columns being compared, Retrograde uses Analysis of Variance, Chi-Square, or Spearman tests as appropriate. It shows highly significant correlations (p < .001) on the left and less significant correlations (p < 0.25) on the right.The correlations shown are those that had a p-value of less than 0.2, the Highest Correlated columns are those that had a p-value of less than 0.001

See Galen Harrison, Kevin Bryson, Ahmad Bamba, Luca Dovichi, Alek Binion, Arthur Borem, and Blase Ur. JupyterLab in Retrograde: Contextual Notifications that Highlight Fairness and Bias Issues for Data Scientists. In *Proc. CHI*, 2024

Retrograde



See Galen Harrison, Kevin Bryson, Ahmad Bamba, Luca Dovichi, Alek Binion, Arthur Borem, and Blase Ur. JupyterLab in Retrograde: Contextual Notifications that Highlight Fairness and Bias Issues for Data Scientists. In *Proc. CHI*, 2024

Retrograde

Modifications Table

Select columns to modify ˅

prediction	index	gender	income	race	type
~~1~~ ➜ **0**	1781	~~female~~ ➜ **non-binary**	-0.402	white	~~home~~ ➜ **personal**
~~1~~ ➜ **0**	1854	~~male~~ ➜ **non-binary**	-0.132	asian	~~personal~~ ➜ **auto**
~~0~~ ➜ **1**	1646	~~male~~ ➜ **female**	0.62	other	~~auto~~ ➜ **personal**
~~1~~ ➜ **0**	1126	~~male~~ ➜ **female**	0.097	hispanic	~~auto~~ ➜ **home**
~~0~~ ➜ **1**	1872	~~female~~ ➜ **male**	1.289	white	~~home~~ ➜ **personal**
~~1~~ ➜ **0**	389	~~female~~ ➜ **male**	-0.031	other	~~home~~ ➜ **auto**
~~1~~ ➜ **0**	1387	~~male~~ ➜ **non-binary**	0.451	white	~~personal~~ ➜ **auto**
~~1~~ ➜ **0**	2288	~~female~~ ➜ **non-binary**	0.82	white	~~home~~ ➜ **personal**
~~1~~ ➜ **0**	1750	~~female~~ ➜ **non-binary**	-0.489	hispanic	~~auto~~ ➜ **personal**
~~1~~ ➜ **0**	1602	~~female~~ ➜ **non-binary**	0.227	black	~~home~~ ➜ **auto**

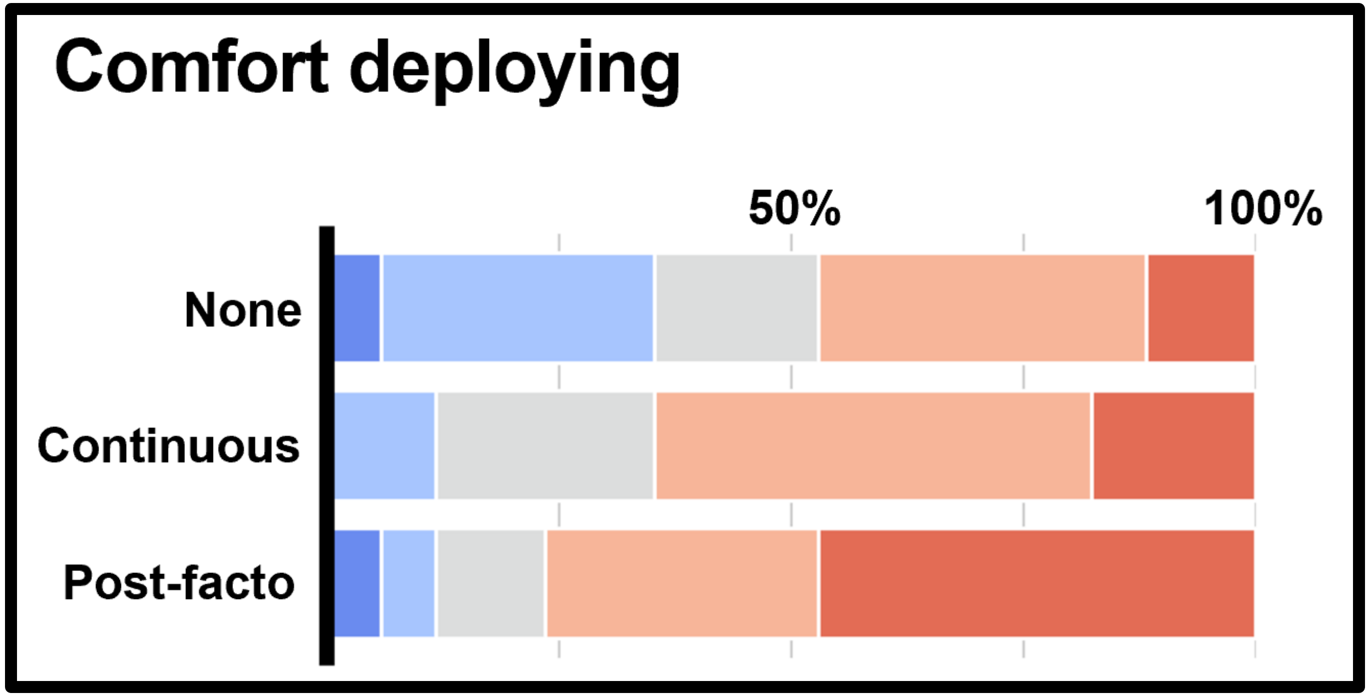
See Galen Harrison, Kevin Bryson, Ahmad Bamba, Luca Dovichi, Alek Binion, Arthur Borem, and Blase Ur. JupyterLab in Retrograde: Contextual Notifications that Highlight Fairness and Bias Issues for Data Scientists. In *Proc. CHI*, 2024

Retrograde: Key Results

- In-context notifications impacted data scientists' actions
- Continuous participants less likely to use protected attributes
- Continuous participants' models had fewer disparities
- Continuous participants more nuanced about missing data
- Nobody in None or Post-facto replicated Retrograde's analyses

See Galen Harrison, Kevin Bryson, Ahmad Bamba, Luca Dovichi, Alek Binion, Arthur Borem, and Blase Ur. JupyterLab in Retrograde: Contextual Notifications that Highlight Fairness and Bias Issues for Data Scientists. In *Proc. CHI*, 2024

Retrograde: Comfort Deploying Model



See Galen Harrison, Kevin Bryson, Ahmad Bamba, Luca Dovichi, Alek Binion, Arthur Borem, and Blase Ur. JupyterLab in Retrograde: Contextual Notifications that Highlight Fairness and Bias Issues for Data Scientists. In *Proc. CHI*, 2024

Counterfactuals and Recourse

- **Counterfactual:** Ideally small difference(s) in a data subject's set of features that would cause a different classification
  - Need a distance metric! But not all variables are created equal.
- **Recourse:** The ability for a data subject to change particular predictor variables
  - Contrast using "the timeliness of credit card payments" versus "the number of years of credit history" versus "sex"
  - To what extent should models **nudge** (influence, but not force) particular behavior?

Algorithmic Decision Making (Revisited)

The Application Context Matters Greatly

Hiring

Online Advertising

Student Admissions

Criminal Justice

Health Insurance Markets

Creditworthiness

Selbst et al.’s Five Pitfalls

- Framing Trap
  - “Failure to model the entire system over which a social criterion, such as fairness, will be enforced“
- Portability Trap
  - “Failure to understand how repurposing algorithmic solutions designed for one social context may be misleading, inaccurate, or otherwise do harm when applied to a different context”
- Formalism Trap
  - “Failure to account for the full meaning of social concepts such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms”
- Ripple Effect Trap
  - “Failure to understand how the insertion of technology into an existing social system changes the behaviors and embedded values of the pre-existing system”
- Solutionism Trap
  - “Failure to recognize the possibility that the best solution to a problem may not involve technology”

Selbst et al. Fairness and Abstraction in Sociotechnical Systems. FAT*, 2019. https://dl.acm.org/doi/pdf/10.1145/3287560.3287598

What Does Accountability Mean Here?

- Who’s accountable for the consequences of an ML model?
  - Those who deployed it?
  - Those who built it and trained it?
  - The owners of the training data?
  - Those who listened to the algorithm?

Biases of Unsupervised Models and Chatbots

Unsupervised Models Are Biased, Too!

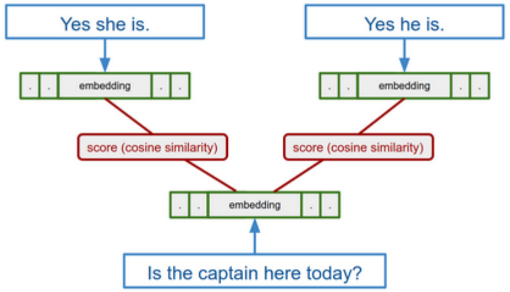
- https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html?m=1

As Machine Learning practitioners, when faced with a task, we usually select or train a model primarily based on how well it performs on that task. For example, say we're building a system to classify whether a movie review is positive or negative. We take 5 different models and see how well each performs this task:

Figure 1: Model performances on a task. Which model would you choose?

Normally, we'd simply choose Model C. But what if we found that while Model C performs the best overall, it's also most likely to assign a more positive sentiment to the sentence "The main character is a man" than to the sentence "The main character is a woman"? Would we reconsider?

Gender Biases of Chatbots



	score
Yes she is	0.17
Yes he is	0.27
Bias score (difference)	-0.10

For a given occupation overall, the model's bias score is the sum of the bias scores for all question/answer templates with that occupation.

Tamera runs 200 occupations through this analysis using the Universal Sentence Encoder embedding model. Table 2 shows the occupations with the highest female-biased scores (left) and the highest male-biased scores (right):

Highest female bias

occupation	bias	occupation	bias
maid	59.2	librarian	20.1
waitress	52.5	obstetrician	16.9
midwife	50.9	secretary	13.7
receptionist	50.2	socialite	12.1
nanny	47.7	therapist	10.2
nurse	45.4	manicurist	10.1
midwives	43.8	hairdresser	9.7
housekeeper	36.6	stylist	8.6
hostess	32	homemaker	6.9
gynecologist	31.6	planner	5.8

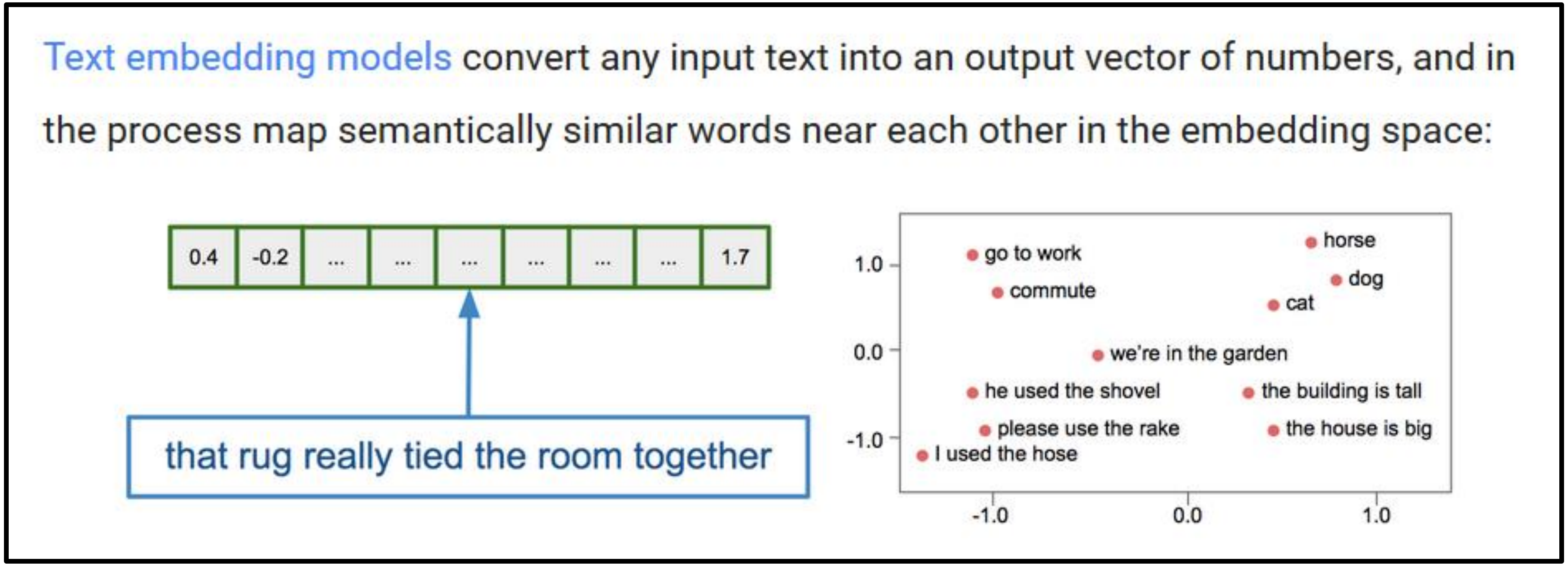
Highest male bias

occupation	bias	occupation	bias
undertaker	-73.4	captain	-53.4
janitor	-62.3	announcer	-51.1
referee	-60.7	architect	-50.7
plumber	-58	maestro	-50.6
actor	-56.9	drafter	-46.7
philosopher	-56.2	usher	-46.6
barber	-55.4	farmer	-45.4
umpire	-54.3	broadcaster	-45.2
president	-54	engineer	-45.1
coach	-53.8	magician	-44.8

Table 2: Occupations with the highest female-biased scores (left) and the highest male-biased scores (right).

https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html?m=1

Word Embeddings



https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html?m=1

Gender Biases of Chatbots

Targets (N)	Attributes (N)	GloVe*	word2vec	nnlm-en-dim50	nnlm-en-dim128	universal
Flowers vs Insects (25)	Pleasant vs Unpleasant (25)	**1.50***	**1.54***	**1.54***	**1.63***	**1.38***
Instruments vs Weapons (25)	Pleasant vs Unpleasant (25)	**1.53***	**1.63***	**1.66***	**1.55***	**1.44***
Eur-American vs Afr-American Names[6] (25)	Pleasant vs Unpleasant[6] (25)	**1.41***	**0.58***	**0.70***	0.04	0.36
Eur-American vs Afr-American Names[7] (18)	Pleasant vs Unpleasant[6] (25)	**1.50***	**1.24***	**1.04***	0.23	-0.37
Eur-American vs Afr-American Names[7] (18)	Pleasant vs Unpleasant[8] (8)	**1.28***	**0.72***	0.28	-0.09	0.72
Male vs Female names (8)	Career vs Family (8)	**1.81***	**1.89***	**1.45***	**1.70***	0.03
Math vs Arts (8)	Male vs Female (8)	1.06	0.97	**1.29***	1.07	0.59
Mental vs Physical Disease (6)	Temporary vs Permanent (7)	**1.38***	1.30	**1.35***	0.96	**1.60***
Science Arts (8)	Male vs Female (8)	**1.24***	**1.24***	**1.34***	1.19	0.24
Young vs Old Names (8)	Pleasant vs Unpleasant (8)	**1.21**	-0.08	0.75	-0.47	1.01

*Table 1: Word Embedding Association Test (WEAT) scores for different embedding models. Cell color indicates whether the direction of the measured bias is in line with (blue) or against (yellow) the common human biases recorded by the Implicit Association Tests. *Statistically significant (p < 0.01) using Caliskan et al. (2015) permutation test. Rows 3-5 are variations whose word lists come from [6], [7], and [8]. See Caliskan et al. for all word lists. * For GloVe, we follow Caliskan et al. and drop uncommon words from the word lists. All other analyses use the full word lists.*

https://developers.googleblog.com/2018/04/text-embedding-models-contain-bias.html?m=1