# FAIRNESS: DEFINITIONS AND MEASUREMENTS

Eunsuk Kang

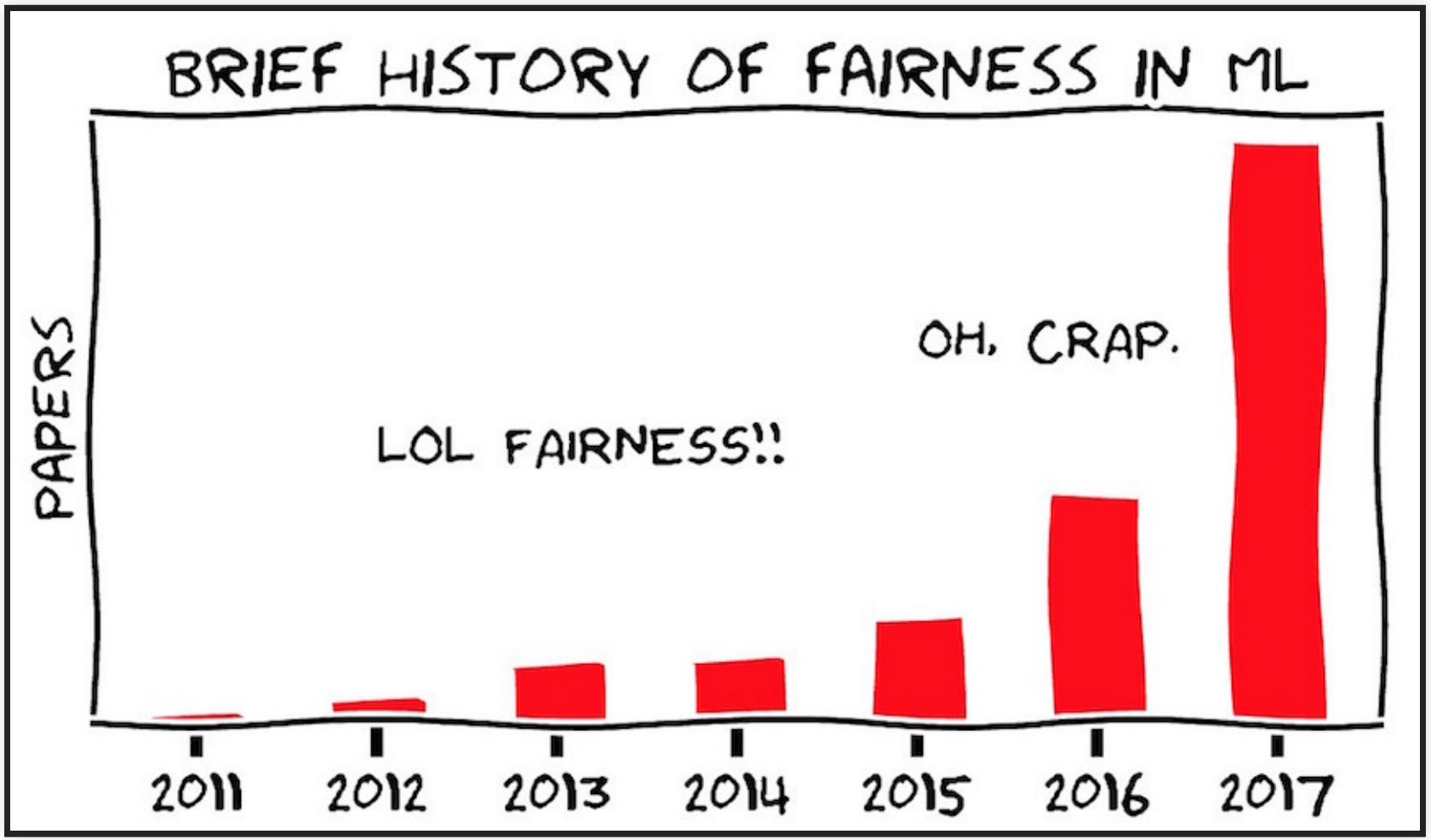
Required reading: Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. "Improving fairness in machine learning systems: What do industry practitioners need?" In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-16. 2019.

# LEARNING GOALS

- Understand different definitions of fairness
- Discuss methods for measuring fairness
- Understand the process of constructing datasets for fairness
- Consider fairness throughout an ML lifecycle

# FAIRNESS: DEFINITIONS

# FAIRNESS IS STILL AN ACTIVELY STUDIED & DISPUTED CONCEPT!




Source: Mortiz Hardt, https://fairmlclass.github.io/

# FAIRNESS: DEFINITIONS

- Anti-classification (fairness through blindness)
- Group fairness (independence)
- Separation (equalized odds)
- ...and numerous others!

# FAIRNESS: DEFINITIONS

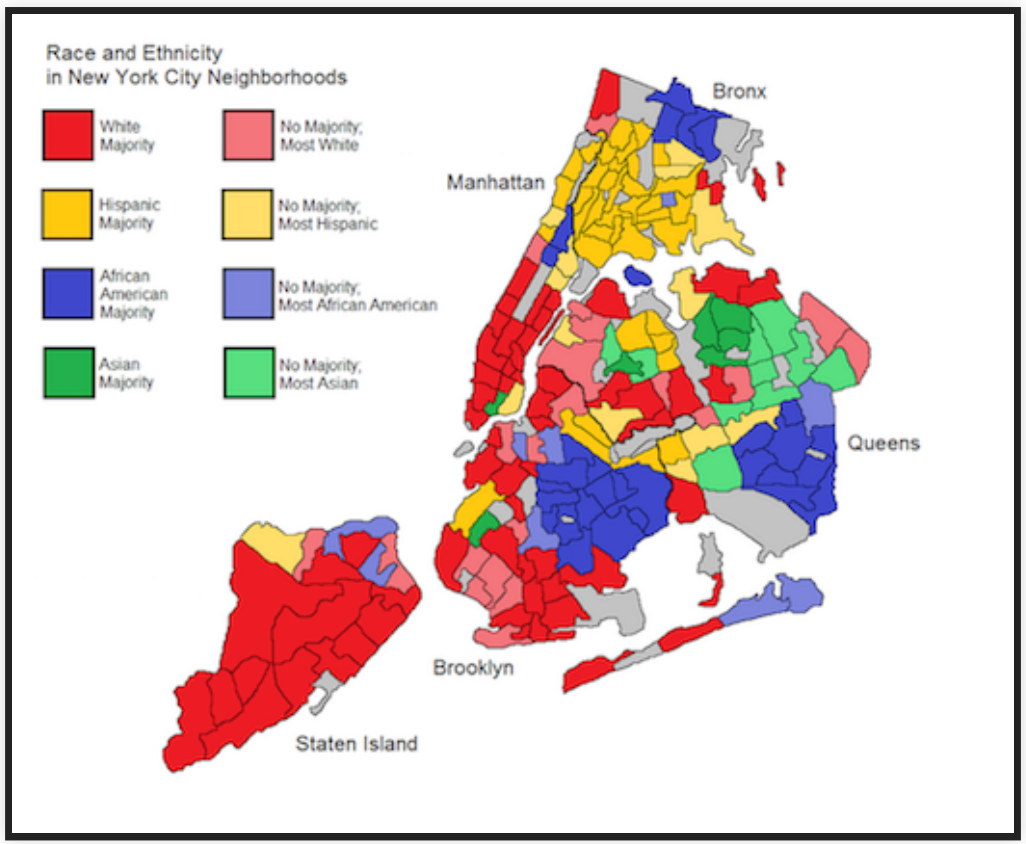
- **Anti-classification (fairness through blindness)**
- Group fairness (independence)
- Separation (equalized odds)

# ANTI-CLASSIFICATION

- Also called *fairness through blindness*
- Ignore certain sensitive attributes when making a decision
- Example: Remove gender or race from a credit scoring model
- **Q. Easy to implement, but any limitations**?

# RECALL: PROXIES

*Features correlate with protected attributes*



# RECALL: NOT ALL DISCRIMINATION IS HARMFUL


- Loan lending: Gender discrimination is illegal.
- Medical diagnosis: Gender-specific diagnosis may be desirable.
- Discrimination is a **domain-specific** concept!

**Other examples?**

# ANTI-CLASSIFICATION

- Ignore certain sensitive attributes when making a decision
- Limitations
    - Sensitive attributes may be correlated with other features
    - Some ML tasks need sensitive attributes (e.g., medical diagnosis)

# TESTING ANTI-CLASSIFICATION

How do we test that a classifier achieves anti-classification?

# TESTING ANTI-CLASSIFICATION

Straightforward invariant for classifier $f$ and protected attribute $p$:

$$\forall x.\ f(x[p \leftarrow 0]) = f(x[p \leftarrow 1])$$

*(does not account for correlated attributes)*

Test with random input data or on any test data

Any single inconsistency shows that the protected attribute was used. Can also report percentage of inconsistencies.

See for example: Galhotra, Sainyam, Yuriy Brun, and Alexandra Meliou. "Fairness testing: testing software for discrimination." In Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering, pp. 498-510. 2017.

# FAIRNESS: DEFINITIONS

- Anti-classification (fairness through blindness)
- **Group fairness (independence)**
- Separation (equalized odds)

# NOTATIONS

- $X$: Feature set (e.g., age, race, education, region, income, etc.,)
- $A \in X$: Sensitive attribute (e.g., gender)
- $R$: Regression score (e.g., predicted likelihood of on-time loan payment)
- $Y^{'}$: Classifier output
  - $Y^{'} = 1$ if and only if $R > T$ for some threshold $T$
  - e.g., Grant the loan ($Y^{'} = 1$) if the likelihood of paying back > 80%
- $Y$: Target variable being predicted ($Y = 1$ if the person actually pays back on time)

Setting classification thresholds: Loan lending example

# GROUP FAIRNESS

$$P[Y^{'} = 1 \,|\, A = a] = P[Y^{'} = 1 \,|\, A = b]$$

- Also called *independence* or *demographic parity*
- Mathematically, $Y^{'} \perp A$
  - Prediction ($Y^{'}$) must be independent of the sensitive attribute ($A$)
- Examples:
  - The predicted rate of recidivism is the same across all races
  - Both women and men have the equal probability of being promoted
    - i.e., P[promote = 1 | gender = M] = P[promote = 1 | gender = F]

# GROUP FAIRNESS

# GROUP FAIRNESS

- Q. What are limitations of group fairness?

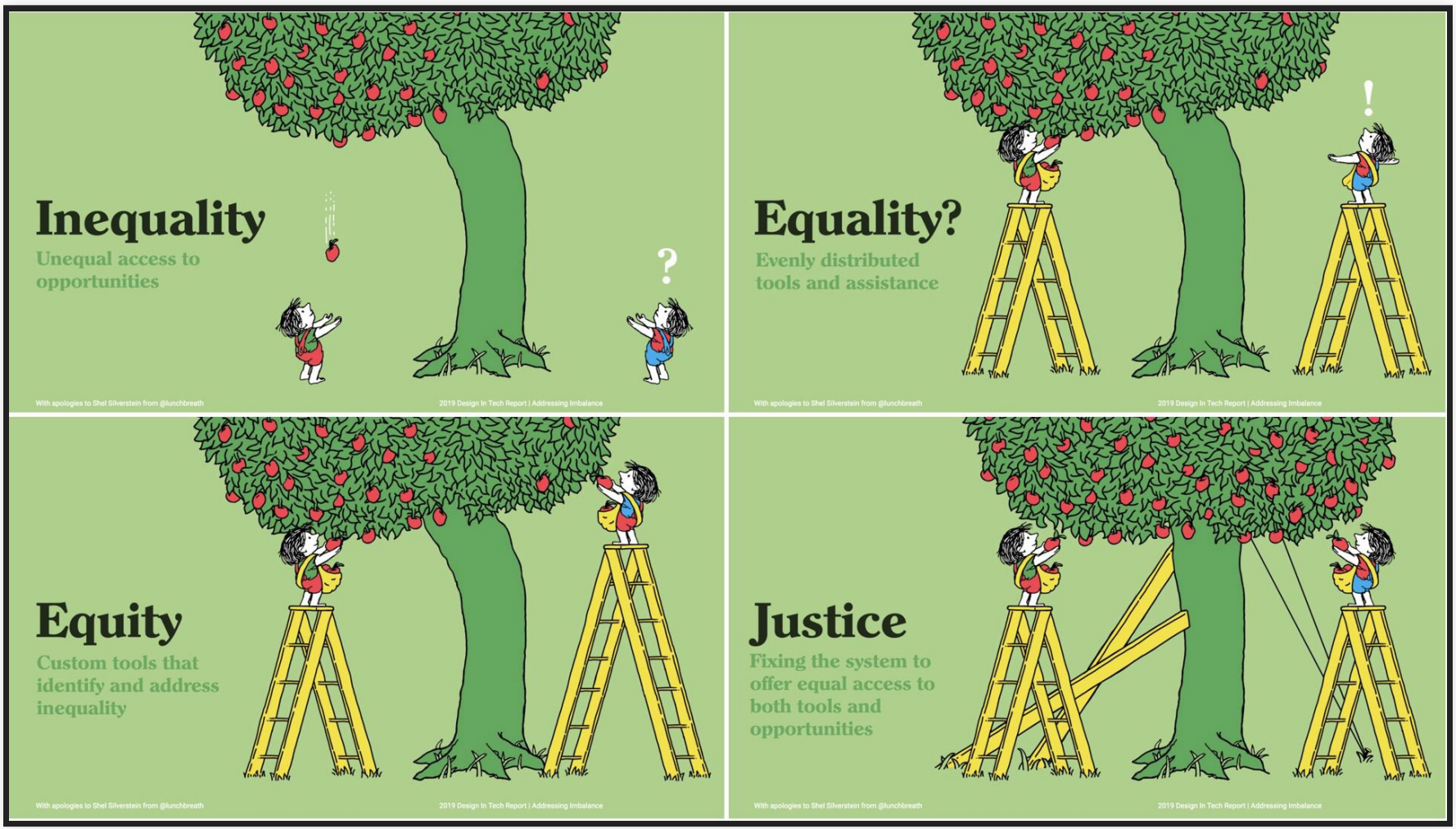
# GROUP FAIRNESS

- Q. What are limitations of group fairness?
  - Ignores possible correlation between $Y$ and $A$
    - Rules out perfect predictor $Y^{'} = Y$ when $Y$ & $A$ are correlated

# GROUP FAIRNESS

- Q. What are limitations of group fairness?
  - Ignores possible correlation between $Y$ and $A$
    - Rules out perfect predictor $Y^{'} = Y$ when $Y$ & $A$ are correlated
  - Permits abuse and laziness: Can be satisfied by randomly assigning a positive outcome ($Y^{'} = 1$) to protected groups
    - e.g., Randomly promote people (regardless of their job performance) to match the rate across all groups

# RECALL: EQUALITY VS EQUITY



# ADJUSTING THRESHOLDS FOR GROUP FAIRNESS

Set $t_0, t_1$ such that $P[R > t_0 \,|\, A = 0] = P[R > t_1 \,|\, A = 1]$

# ADJUSTING THRESHOLDS FOR GROUP FAIRNESS

$$\text{Set } t_0, t_1 \text{ such that } P[R > t_0 \,|\, A = 0] = P[R > t_1 \,|\, A = 1]$$

- Select different classification thresholds ($t_0$, $t_1$) for different groups (A = 0, A = 1) to achieve group fairness

# ADJUSTING THRESHOLDS FOR GROUP FAIRNESS

Set $t_0, t_1$ such that $P[R > t_0 \,|\, A = 0] = P[R > t_1 \,|\, A = 1]$

- Select different classification thresholds ($t_0$, $t_1$) for different groups (A = 0, A = 1) to achieve group fairness
- Example: Loan lending
  - R: Likelihood of paying back the loan on time
  - Suppose: With a uniform threshold used (i.e., R = 80%), group fairness is not achieved
    - P[R > 0.8 | A = 0] = 0.4, P[R > 0.8 | A = 1] = 0.7
  - Adjust thresholds to achieve group fairness
    - P[R > 0.6 | A = 0] = P[R > 0.8 | A = 1]

# ADJUSTING THRESHOLDS FOR GROUP FAIRNESS

Set $t_0, t_1$ such that $P[R > t_0 \,|\, A = 0] = P[R > t_1 \,|\, A = 1]$

- Select different classification thresholds ($t_0$, $t_1$) for different groups (A = 0, A = 1) to achieve group fairness
- Example: Loan lending
  - R: Likelihood of paying back the loan on time
  - Suppose: With a uniform threshold used (i.e., R = 80%), group fairness is not achieved
    - P[R > 0.8 | A = 0] = 0.4, P[R > 0.8 | A = 1] = 0.7
  - Adjust thresholds to achieve group fairness
    - P[R > 0.6 | A = 0] = P[R > 0.8 | A = 1]
- But this also seems unfair to some of the groups! (i.e., A = 1)
  - Q. When does this type of adjustment make sense?

# TESTING GROUP FAIRNESS

# TESTING GROUP FAIRNESS

- How would you test whether a classifier achieves group fairness?

# TESTING GROUP FAIRNESS

- How would you test whether a classifier achieves group fairness?
- Separate validation/telemetry data by protected attributes

# TESTING GROUP FAIRNESS

- How would you test whether a classifier achieves group fairness?
- Separate validation/telemetry data by protected attributes
    - Generate realistic test data, e.g. from probability distribution of population

# TESTING GROUP FAIRNESS

- How would you test whether a classifier achieves group fairness?
- Separate validation/telemetry data by protected attributes
    - Generate realistic test data, e.g. from probability distribution of population
- Separately measure the rate of positive predictions
    - e.g., P[promoted = 1 | gender = M], P[promoted = 1 | gender = F] = ?

# TESTING GROUP FAIRNESS

- How would you test whether a classifier achieves group fairness?
- Separate validation/telemetry data by protected attributes
    - Generate realistic test data, e.g. from probability distribution of population
- Separately measure the rate of positive predictions
    - e.g., P[promoted = 1 | gender = M], P[promoted = 1 | gender = F] = ?
- Report issue if the rates differ beyond some threshold $\epsilon$ across groups

# FAIRNESS: DEFINITIONS

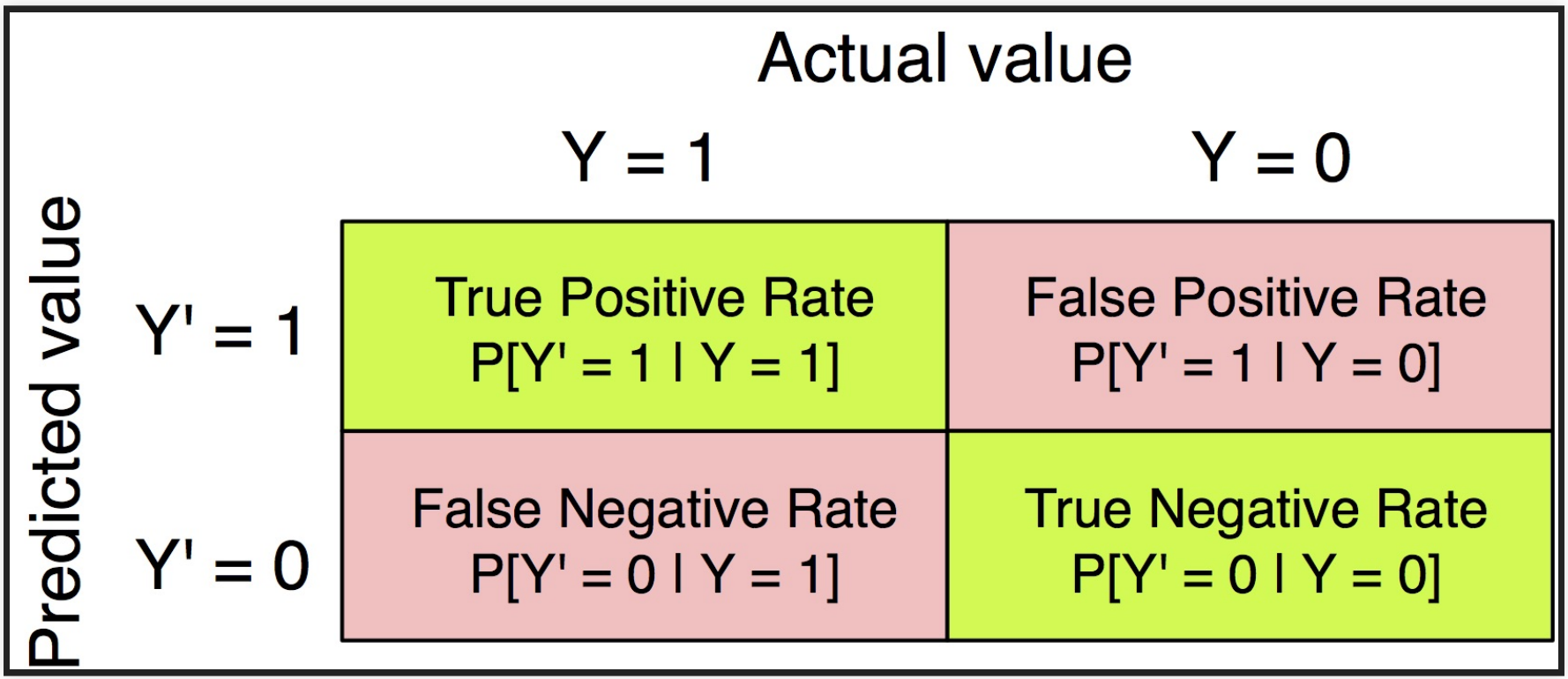
- Anti-classification (fairness through blindness)
- Group fairness (independence)
- **Separation (equalized odds)**

# SEPARATION

$$P[Y^{'} = 1 \mid Y = 0, A = a] = P[Y^{'} = 1 \mid Y = 0, A = b]$$
$$P[Y^{'} = 0 \mid Y = 1, A = a] = P[Y^{'} = 0 \mid Y = 1, A = b]$$

- Also called *equalized odds*
- $Y^{'} \perp A \mid Y$
  - Prediction must be independent of the sensitive attribute *conditional* on the target variable

# REVIEW: CONFUSION MATRIX




Can we explain separation in terms of model errors?

$$P[Y^{'} = 1 \mid Y = 0, A = a] = P[Y^{'} = 1 \mid Y = 0, A = b]$$
$$P[Y^{'} = 0 \mid Y = 1, A = a] = P[Y^{'} = 0 \mid Y = 1, A = b]$$

# SEPARATION

$$P[Y^{'} = 1 \mid Y = 0, A = a] = P[Y^{'} = 1 \mid Y = 0, A = b] \text{ (FPR parity)}$$
$$P[Y^{'} = 0 \mid Y = 1, A = a] = P[Y^{'} = 0 \mid Y = 1, A = b] \text{ (FNR parity)}$$

- $Y^{'} \perp A \mid Y$
  - Prediction must be independent of the sensitive attribute *conditional* on the target variable

# SEPARATION

$$P[Y^{'} = 1 \mid Y = 0, A = a] = P[Y^{'} = 1 \mid Y = 0, A = b] \text{ (FPR parity)}$$
$$P[Y^{'} = 0 \mid Y = 1, A = a] = P[Y^{'} = 0 \mid Y = 1, A = b] \text{ (FNR parity)}$$

- $Y^{'} \perp A \mid Y$
  - Prediction must be independent of the sensitive attribute *conditional* on the target variable
- i.e., All groups are susceptible to the same false positive/negative rates

# SEPARATION

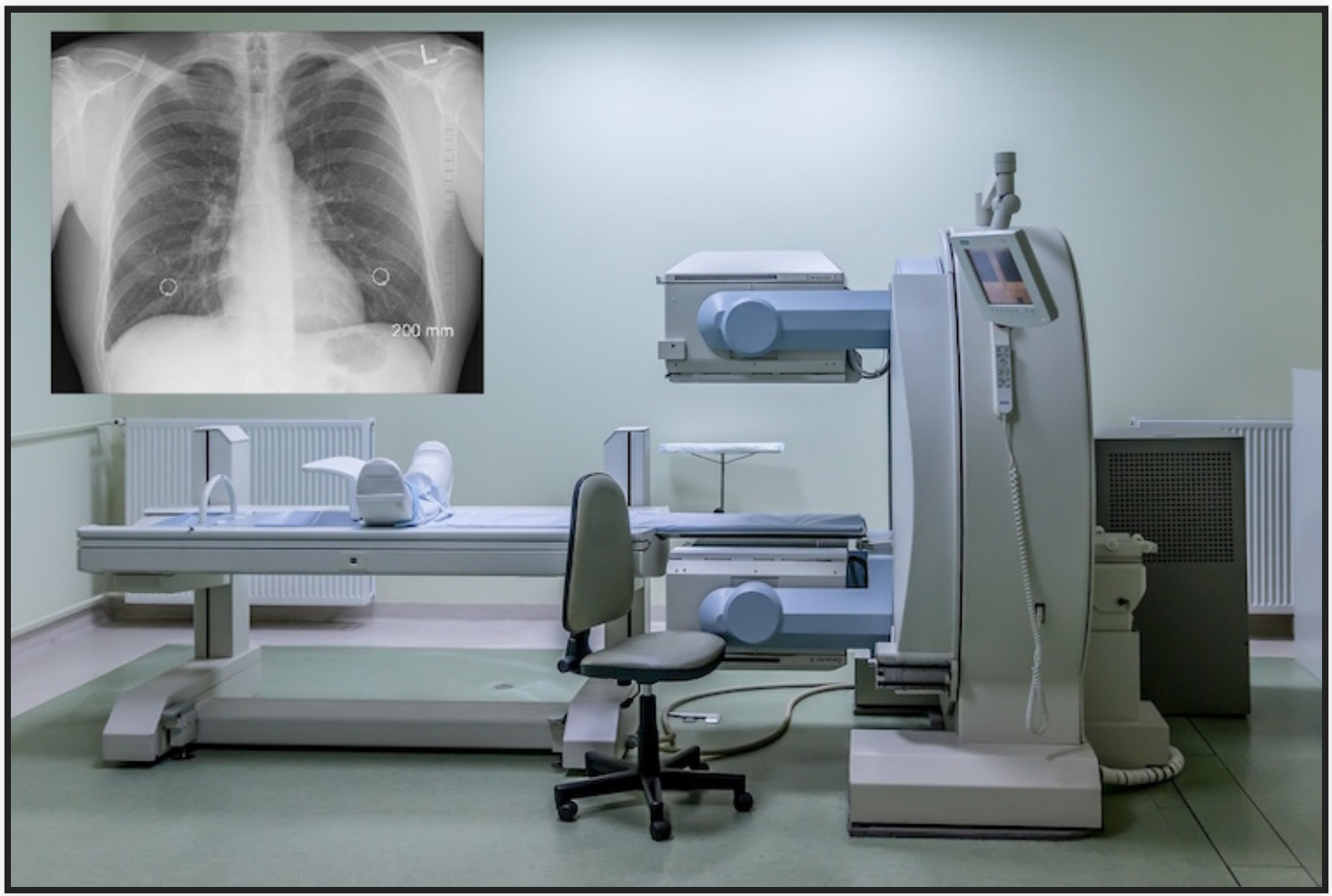
$$P[Y^{'} = 1 \mid Y = 0, A = a] = P[Y^{'} = 1 \mid Y = 0, A = b] \text{ (FPR parity)}$$
$$P[Y^{'} = 0 \mid Y = 1, A = a] = P[Y^{'} = 0 \mid Y = 1, A = b] \text{ (FNR parity)}$$

- $Y^{'} \perp A \mid Y$
  - Prediction must be independent of the sensitive attribute *conditional* on the target variable
- i.e., All groups are susceptible to the same false positive/negative rates
- Example: Promotion
  - Y': Promotion decision, A: Gender of applicant: Y: Actual job performance
  - Separation w/ FNR: Probability of being incorrectly denied promotion is equal across both male & female employees

# TESTING SEPARATION

- Generate separate validation sets for each group
- Separate validation/telemetry data by protected attribute
    - Or generate *realistic* test data, e.g. from probability distribution of population
- Separately measure false positive and false negative rates
    - e..g, for FNR, compare P[promoted = 0 | female, good employee] vs P[promoted = 0 | male, good employee]
- Q. How is this different from testing group fairness?

# CASE STUDY: CANCER DIAGNOSIS



# EXERCISE: CANCER DIAGNOSIS

**Overall Results**

| | |
|---|---|
| True positives (TPs): 16 | False positives (FPs): 21 |
| False negatives (FNs): 9 | True negatives (TNs): 954 |

**Male Patient Results**

| | |
|---|---|
| True positives (TPs): 3 | False positives (FPs): 16 |
| False negatives (FNs): 7 | True negatives (TNs): 474 |

**Female Patient Results**

| | |
|---|---|
| True positives (TPs): 13 | False positives (FPs): 5 |
| False negatives (FNs): 2 | True negatives (TNs): 480 |

- 1000 data samples (500 male & 500 female patients)
- Does the model achieve group fairness? Separation w/ FPR or FNR?
- What can we conclude about the model & its usage?

# REVIEW OF CRITERIA SO FAR:

*Recidivism scenario: Should a person be detained?*

- Anti-classification: ?
- Group fairness: ?
- Separation: ?

# REVIEW OF CRITERIA SO FAR:

*Recidivism scenario: Should a defendant be detained?*

- Anti-classification: Race and gender should not be considered for the decision at all
- Group fairness: Detention rates should be equal across gender and race groups
- Separation: Among defendants who would not have gone on to commit a violent crime if released, detention rates are equal across gender and race groups

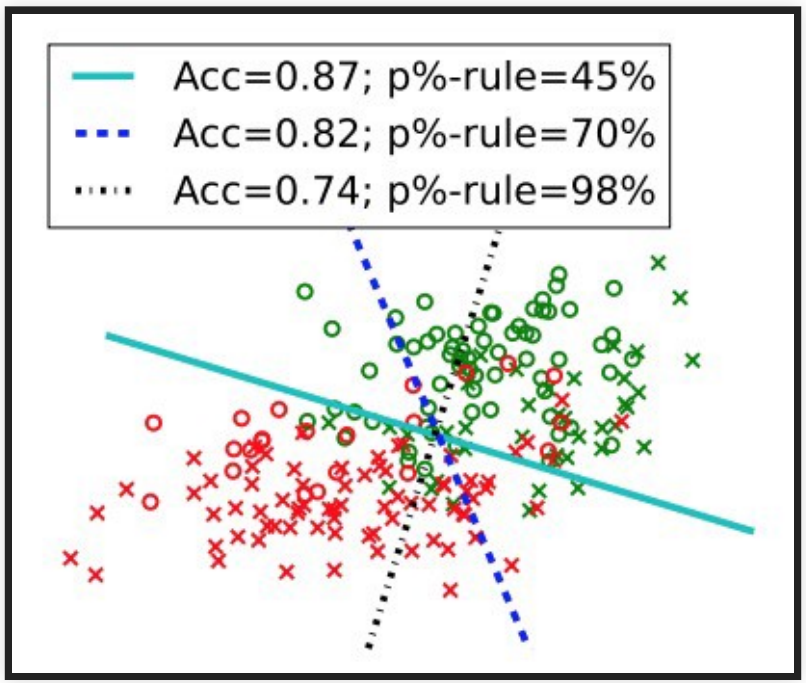
# ACHIEVING FAIRNESS CRITERIA

# CAN WE ACHIEVE FAIRNESS DURING THE LEARNING PROCESS?

- Data acquisition:
  - Collect additional data if performance is poor on some groups
- Pre-processing:
  - Clean the dataset to reduce correlation between the feature set and sensitive attributes
- Training constraints
  - ML is a constraint optimization problem (i.e., minimize errors)
  - Impose additional parity constraint into ML optimization process (as part of the loss function)
- Post-processing
  - Adjust thresholds to achieve a desired fairness metric
- (Still active area of research! Many new techniques published each year)

*Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints*, Cotter et al., (2018).

# TRADE-OFFS IN FAIRNESS VS ACCURACY




- In general, accuracy is at odds with fairness
  - e.g., Impossible to achieve perfect accuracy ($R = Y$) while ensuring group fairness
- Determine how much compromise in accuracy or fairness is acceptable to your stakeholders

*Fairness Constraints: Mechanisms for Fair Classification*, Zafar et al., AISTATS (2017).

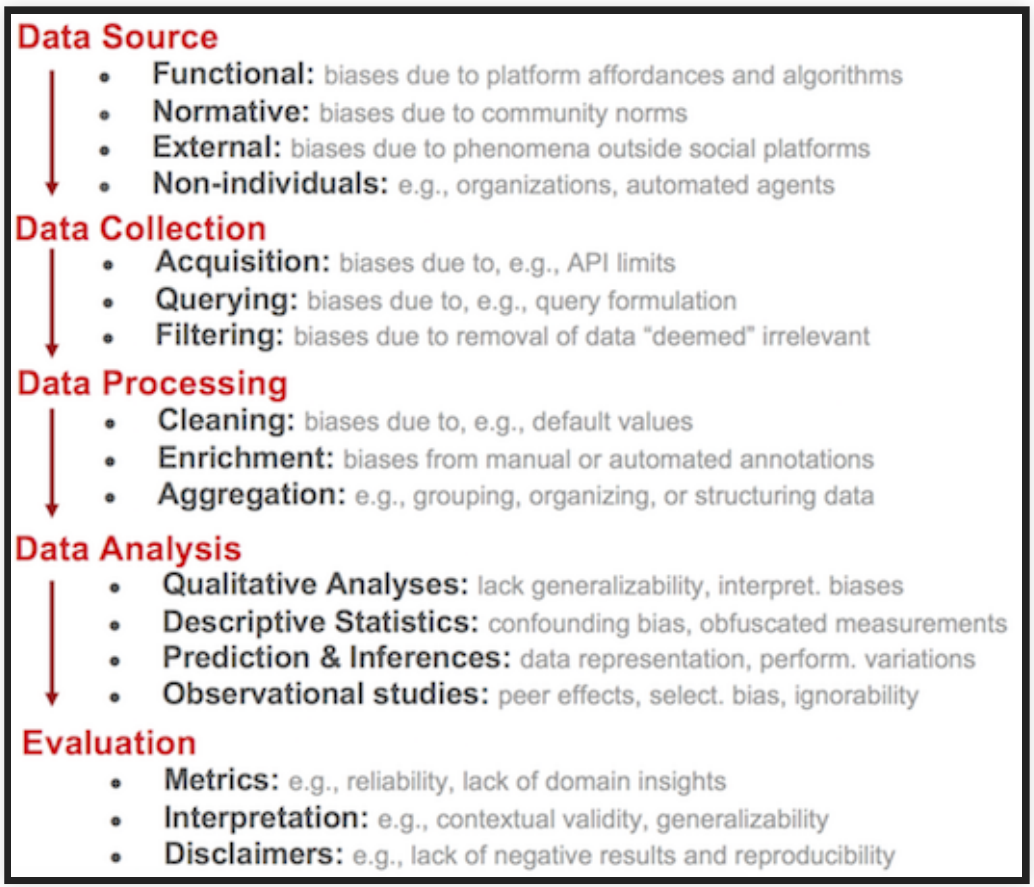
# DATASET CONSTRUCTION FOR FAIRNESS

# FLEXIBILITY IN DATA COLLECTION

- Data science education often assumes data as given
- In industry, we often have control over data collection and curation (65%)
- Most address fairness issues by collecting more data (73%)

Challenges of incorporating algorithmic fairness into practice, FAT* Tutorial, 2019 (slides)

# DATA BIAS




- Bias can be introduced at any stage of the data pipeline!

Bennett et al., Fairness-aware Machine Learning, WSDM Tutorial (2019).

# TYPES OF DATA BIAS

- **Population bias**
- **Historical bias**
- **Behavioral bias**
- Content production bias
- Linking bias
- Temporal bias

*Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries*, Olteanu et al., Frontiers in Big Data (2016).

# POPULATION BIAS

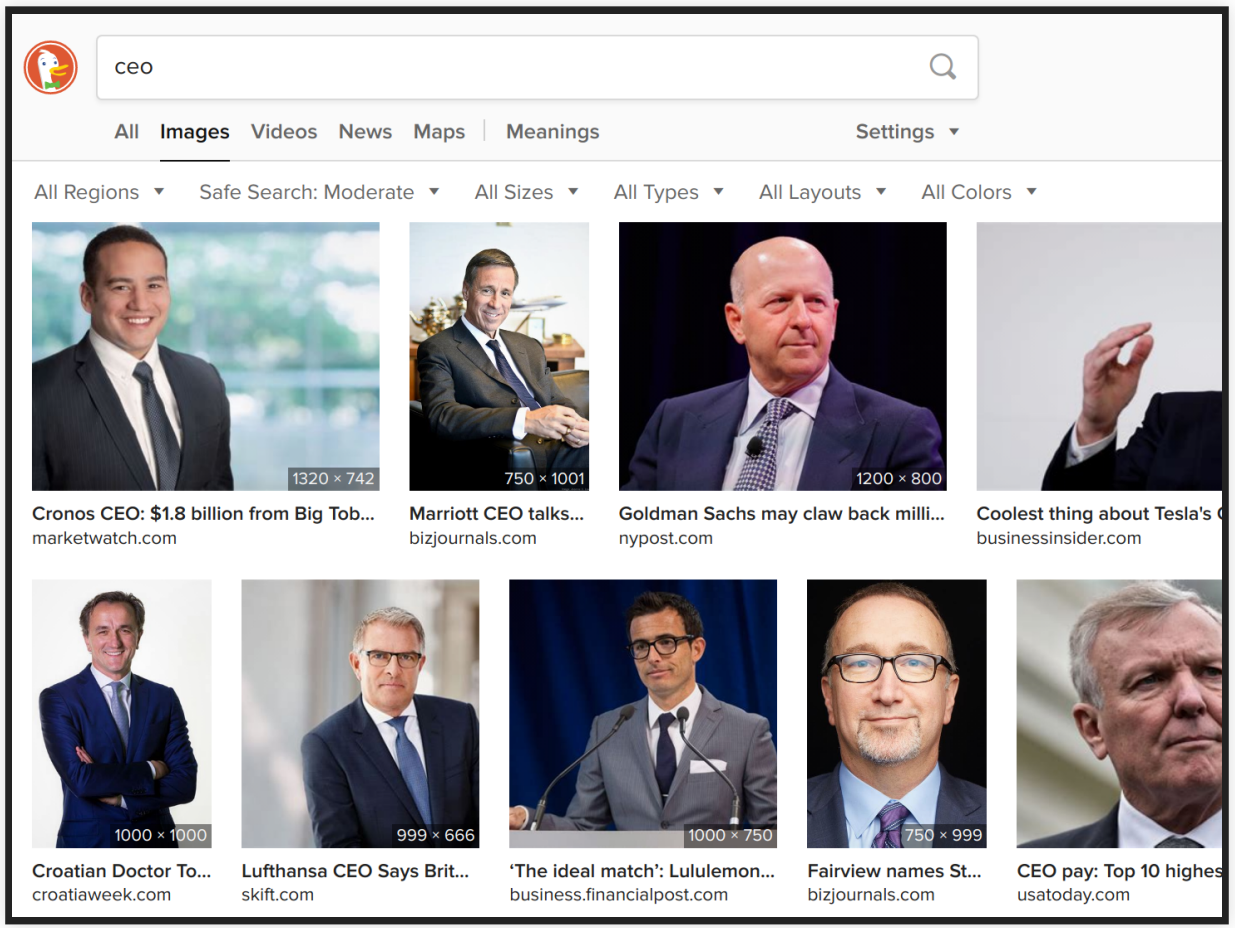
| | Gender | | Skin Color/Type | |
|---|---|---|---|---|
| **Data set** | **Female** | **Male** | **Darker** | **Lighter** |
| LFW [15] | 22.5% | 77.4% | 18.8% | 81.2% |
| IJB-C [28] | 37.4% | 62.7% | 18.0% | 82.0% |
| Pubfig [35] | 50.8% | 49.2% | 18.0% | 82.0% |
| CelebA [9] | 58.1% | 42.0% | 14.2% | 85.8% |
| UTKface [32] | 47.8% | 52.2% | 35.6% | 64.4% |
| AgeDB [33] | 40.6% | 59.5% | 5.4% | 94.6% |
| PPB [36] | 44.6% | 55.4% | 46.4% | 53.6% |
| IMDB-Face [24] | 45.0% | 55.0% | 12.0% | 88.0% |

Table 3: Distribution of gender and skin color/type for seven prominent face image data sets.

- Differences in demographics between a dataset vs a target population
- May result in degraded services for certain groups (e.g., poor image recognition for females & darker skin types)
- Another example: Demographics on social media

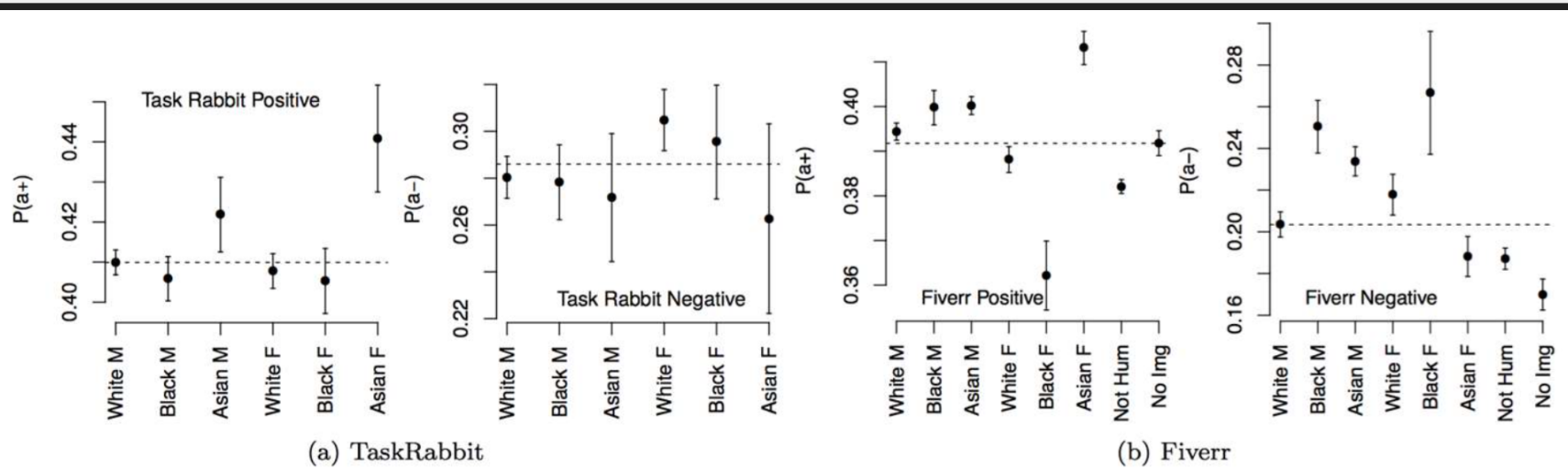
Merler, Ratha, Feris, and Smith. Diversity in Faces

# HISTORICAL BIAS




- Dataset matches the reality, but certain groups are under- or over-represented due to historical reasons

# BEHAVIORAL BIAS




(a) TaskRabbit (b) Fiverr

Figure 2: Fitted $P(a_+)$ and $P(a_-)$ depending on combinations of gender and race of the reviewed worker. Points show expected values and bars standard errors. In Fiverr, Black workers are less likely to be described with adjectives for positive words, and Black Male workers are more likely to be described with adjectives for negative words.

- Differences in user behavior across platforms or social contexts
- Example: Freelancing platforms (Fiverr vs TaskRabbit)
  - Bias against certain minority groups on different platforms

*Bias in Online Freelance Marketplaces*, Hannak et al., CSCW (2017).

# FAIRNESS-AWARE DATA COLLECTION

*Fairness-aware Machine Learning*, Bennett et al., WSDM Tutorial (2019).

# FAIRNESS-AWARE DATA COLLECTION

- Address population bias
  - Does the dataset reflect the demographics in the target population?
  - If not, collect more data to achieve this

*Fairness-aware Machine Learning*, Bennett et al., WSDM Tutorial (2019).

# FAIRNESS-AWARE DATA COLLECTION

- Address population bias
  - Does the dataset reflect the demographics in the target population?
  - If not, collect more data to achieve this
- Address under- & over-representation issues
  - Ensure sufficient amount of data for all groups to avoid being treated as "outliers" by ML
  - Also avoid over-representation of certain groups (e.g., remove historical data)

*Fairness-aware Machine Learning*, Bennett et al., WSDM Tutorial (2019).

# FAIRNESS-AWARE DATA COLLECTION

- Address population bias
  - Does the dataset reflect the demographics in the target population?
  - If not, collect more data to achieve this
- Address under- & over-representation issues
  - Ensure sufficient amount of data for all groups to avoid being treated as "outliers" by ML
  - Also avoid over-representation of certain groups (e.g., remove historical data)
- Data augmentation: Synthesize data for minority groups to reduce under-representation
  - Observed: "He is a doctor" -> synthesize "She is a doctor"

*Fairness-aware Machine Learning*, Bennett et al., WSDM Tutorial (2019).

# FAIRNESS-AWARE DATA COLLECTION

- Address population bias
  - Does the dataset reflect the demographics in the target population?
  - If not, collect more data to achieve this
- Address under- & over-representation issues
  - Ensure sufficient amount of data for all groups to avoid being treated as "outliers" by ML
  - Also avoid over-representation of certain groups (e.g., remove historical data)
- Data augmentation: Synthesize data for minority groups to reduce under-representation
  - Observed: "He is a doctor" -> synthesize "She is a doctor"
- Fairness-aware active learning
  - Evaluate accuracy across different groups
  - Collect more data for groups with highest error rates

*Fairness-aware Machine Learning*, Bennett et al., WSDM Tutorial (2019).

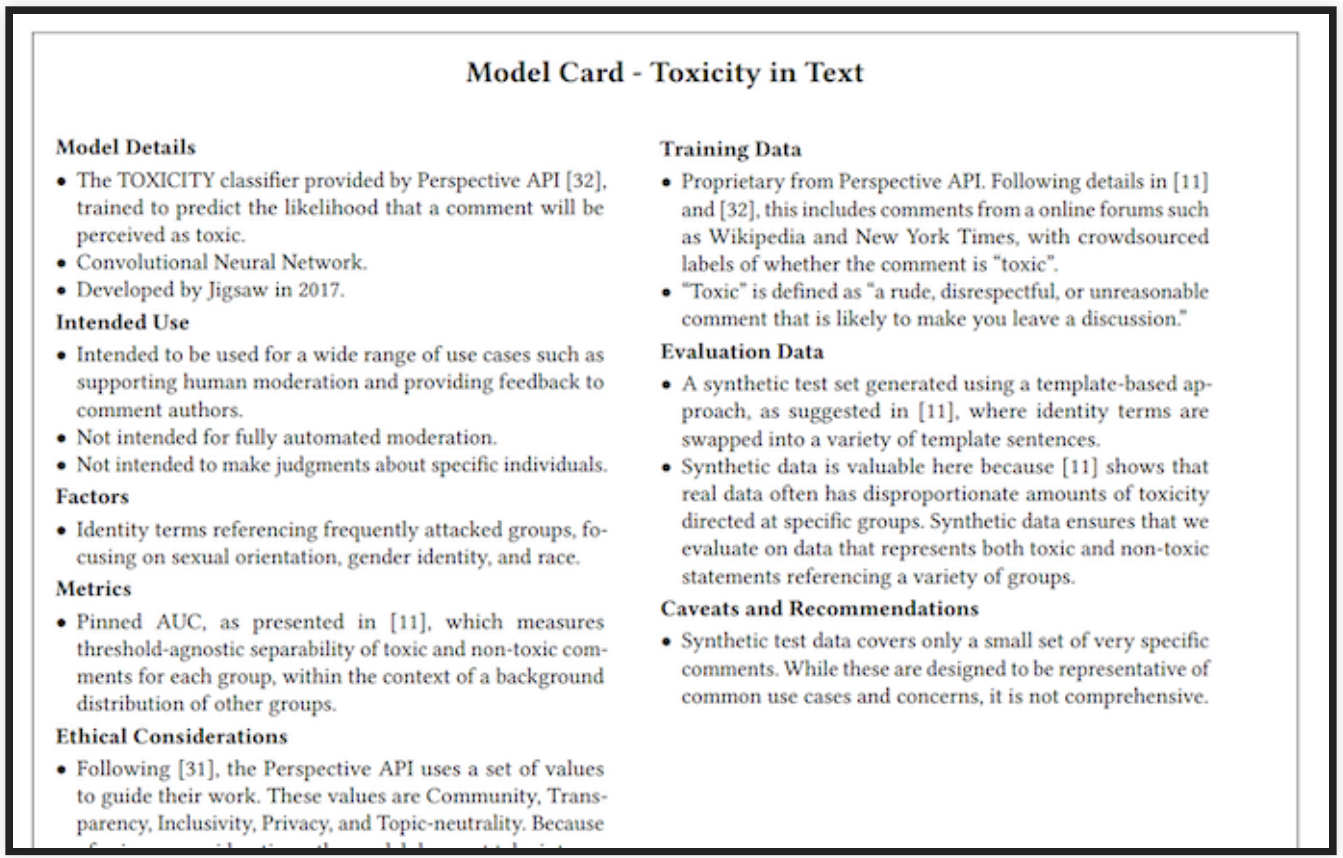
# DATA SHEETS

| Demographic Characteristic | Value |
|---|---|
| Percentage of female subjects | 22.5% |
| Percentage of male subjects | 77.5% |
| Percentage of White subjects | 83.5% |
| Percentage of Black subjects | 8.47% |
| Percentage of Asian subjects | 8.03% |
| Percentage of people between 0-20 years old | 1.57% |
| Percentage of people between 21-40 years old | 31.63% |
| Percentage of people between 41-60 years old | 45.58% |
| Percentage of people over 61 years old | 21.2% |

- A process for documenting datasets
- Common practice in the electronics industry, medicine
- Purpose, provenance, creation, **composition**, distribution
    - "Does the dataset relate to people?"
    - "Does the dataset identify any subpopulations (e.g., by age, gender)?"

*Datasheets for Dataset*, Gebru et al., (2019). https://arxiv.org/abs/1803.09010

# MODEL CARDS



## Model Card - Toxicity in Text

**Model Details**

- The TOXICITY classifier provided by Perspective API [32], trained to predict the likelihood that a comment will be perceived as toxic.
- Convolutional Neural Network.
- Developed by Jigsaw in 2017.

**Intended Use**

- Intended to be used for a wide range of use cases such as supporting human moderation and providing feedback to comment authors.
- Not intended for fully automated moderation.
- Not intended to make judgments about specific individuals.

**Factors**

- Identity terms referencing frequently attacked groups, focusing on sexual orientation, gender identity, and race.

**Metrics**

- Pinned AUC, as presented in [11], which measures threshold-agnostic separability of toxic and non-toxic comments for each group, within the context of a background distribution of other groups.

**Ethical Considerations**

- Following [31], the Perspective API uses a set of values to guide their work. These values are Community, Transparency, Inclusivity, Privacy, and Topic-neutrality. Because

**Training Data**

- Proprietary from Perspective API. Following details in [11] and [32], this includes comments from a online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic".
- "Toxic" is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion."

**Evaluation Data**

- A synthetic test set generated using a template-based approach, as suggested in [11], where identity terms are swapped into a variety of template sentences.
- Synthetic data is valuable here because [11] shows that real data often has disproportionate amounts of toxicity directed at specific groups. Synthetic data ensures that we evaluate on data that represents both toxic and non-toxic statements referencing a variety of groups.

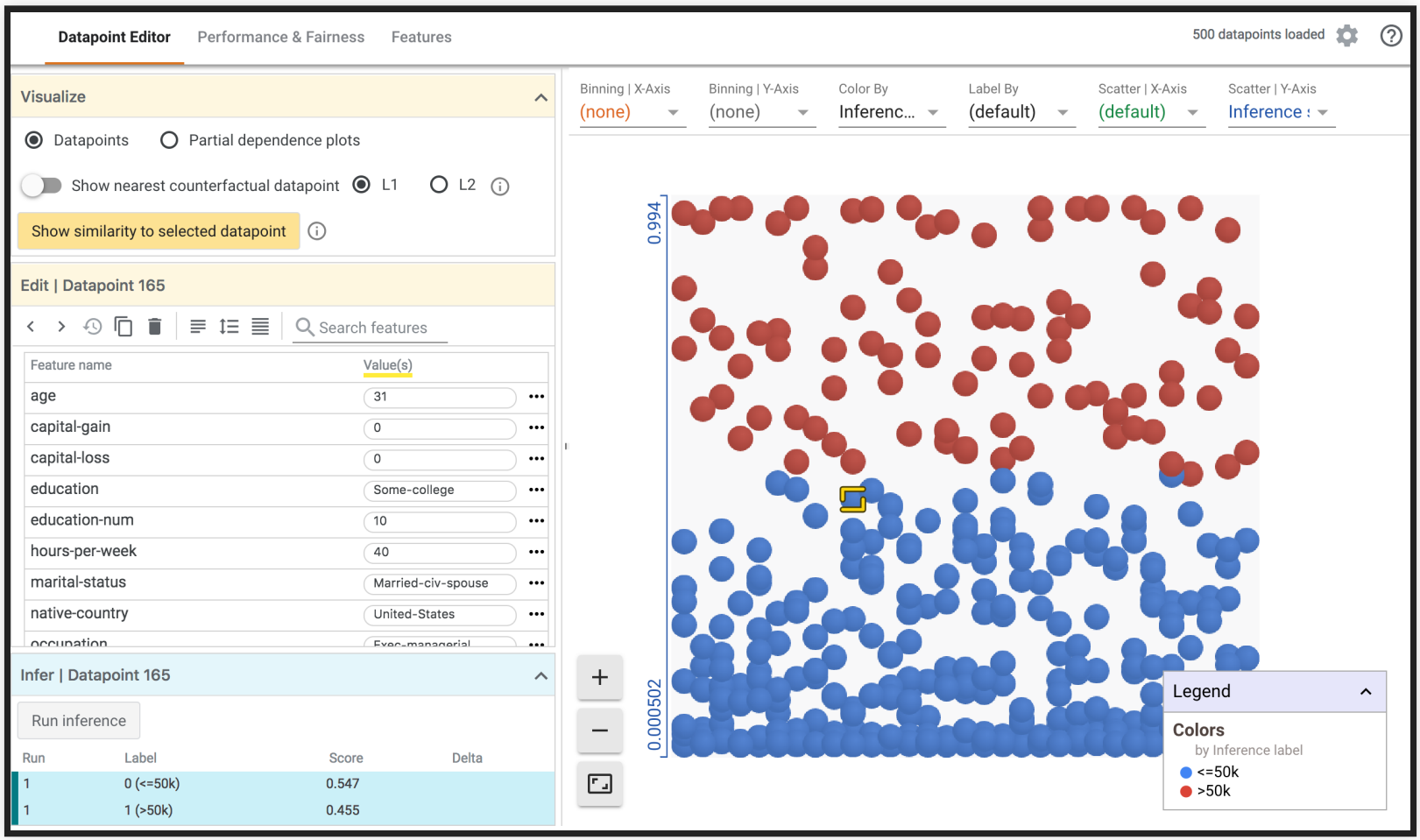
**Caveats and Recommendations**

- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

See also: https://modelcards.withgoogle.com/about

Mitchell, Margaret, et al. "Model cards for model reporting." In Proceedings of the Conference on fairness, accountability, and transparency, pp. 220-229. 2019.

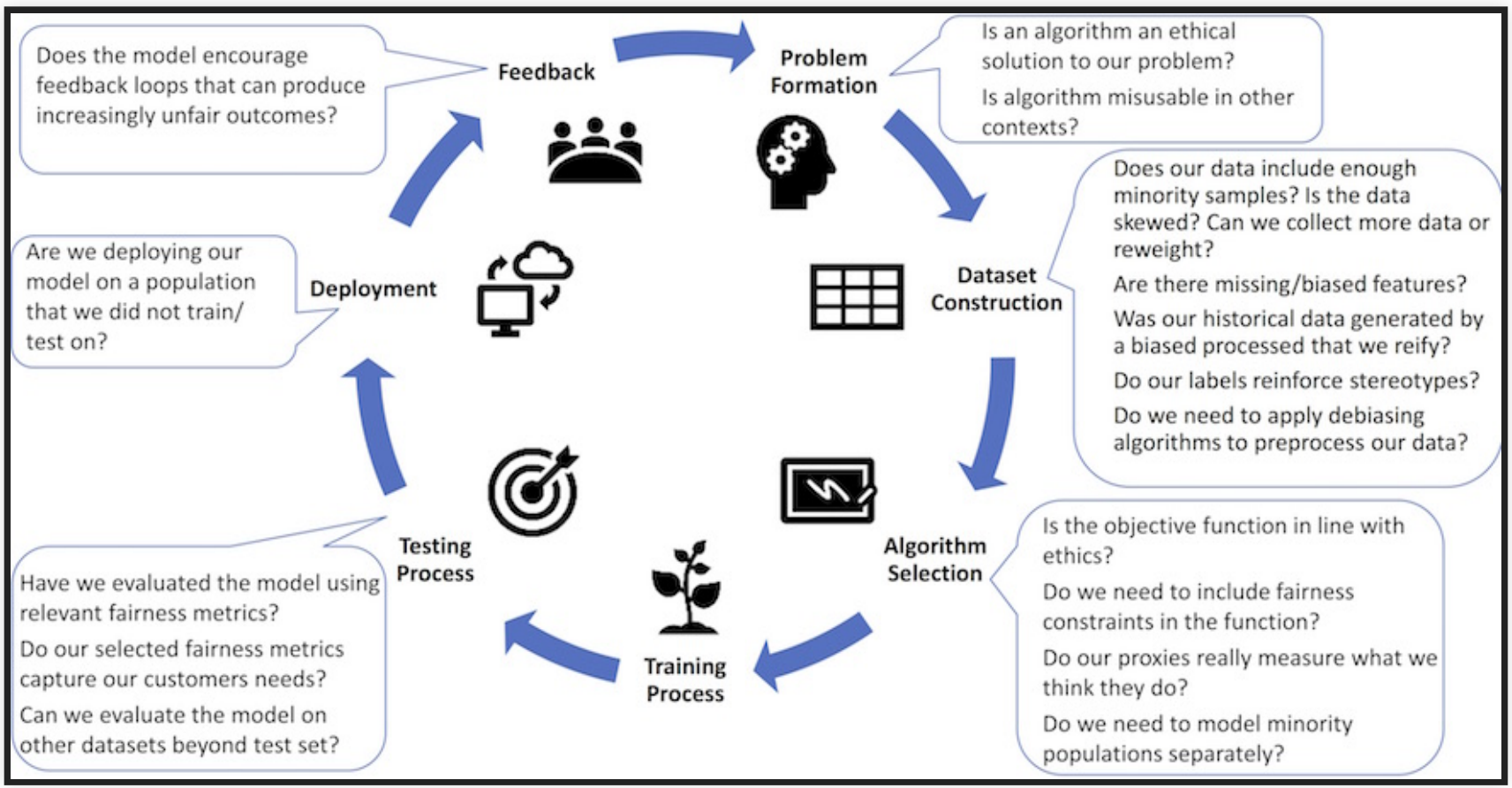
# MODEL EXPLORATION




Google What-If Tool

# BUILDING FAIR ML SYSTEMS

# FAIRNESS MUST BE CONSIDERED THROUGHOUT THE ML LIFECYCLE!




*Fairness-aware Machine Learning*, Bennett et al., WSDM Tutorial (2019).

# PRACTITIONER CHALLENGES

- Fairness is a system-level property
    - Consider goals, user interaction design, data collection, monitoring, model interaction (properties of a single model may not matter much)
- Fairness-aware data collection, fairness testing for training data
- Identifying blind spots
    - Proactive vs reactive
    - Team bias and (domain-specific) checklists
- Fairness auditing processes and tools
- Diagnosis and debugging (outlier or systemic problem? causes?)
- Guiding interventions (adjust goals? more data? side effects? chasing mistakes? redesign?)
- Assessing human bias of humans in the loop

Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. "Improving fairness in machine learning systems: What do industry practitioners need?" In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-16. 2019.

# SUMMARY

- Definitions of fairness
  - Anti-classification, independence, separation
- Achieving fairness
  - Trade-offs between accuracy & fairness
- Dataset construction for fairness
- Achieving fairness as an activity throughout the entire development cycle