FAIRNESS: DEFINITIONS AND MEASUREMENTS

Eunsuk Kang

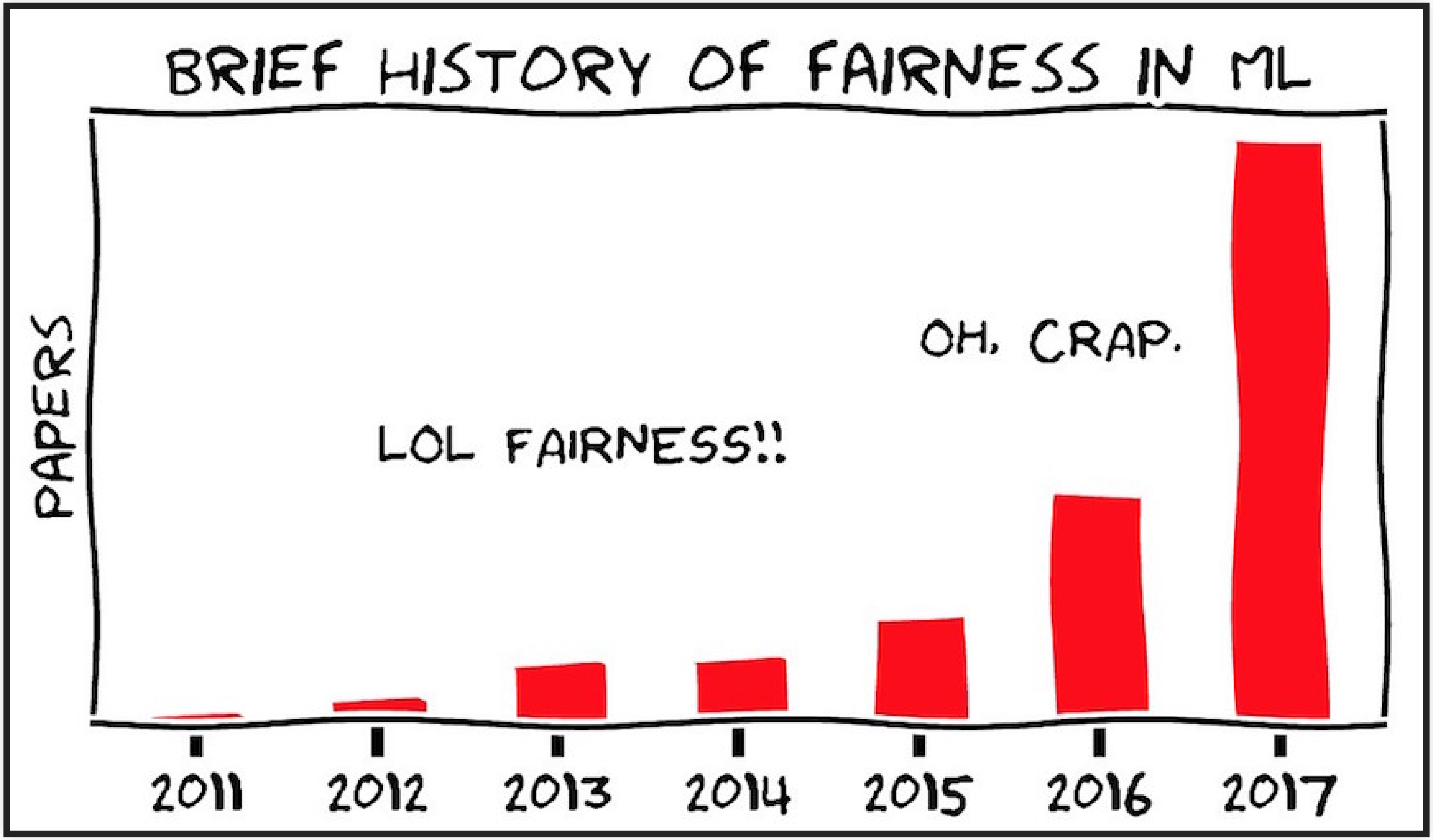
Required reading: Holstein, Kenneth, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. "Improving fairness in machine learning systems: What do industry practitioners need?" In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-16. 2019.

LEARNING GOALS

- Understand different definitions of fairness
- Discuss methods for measuring fairness
- 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

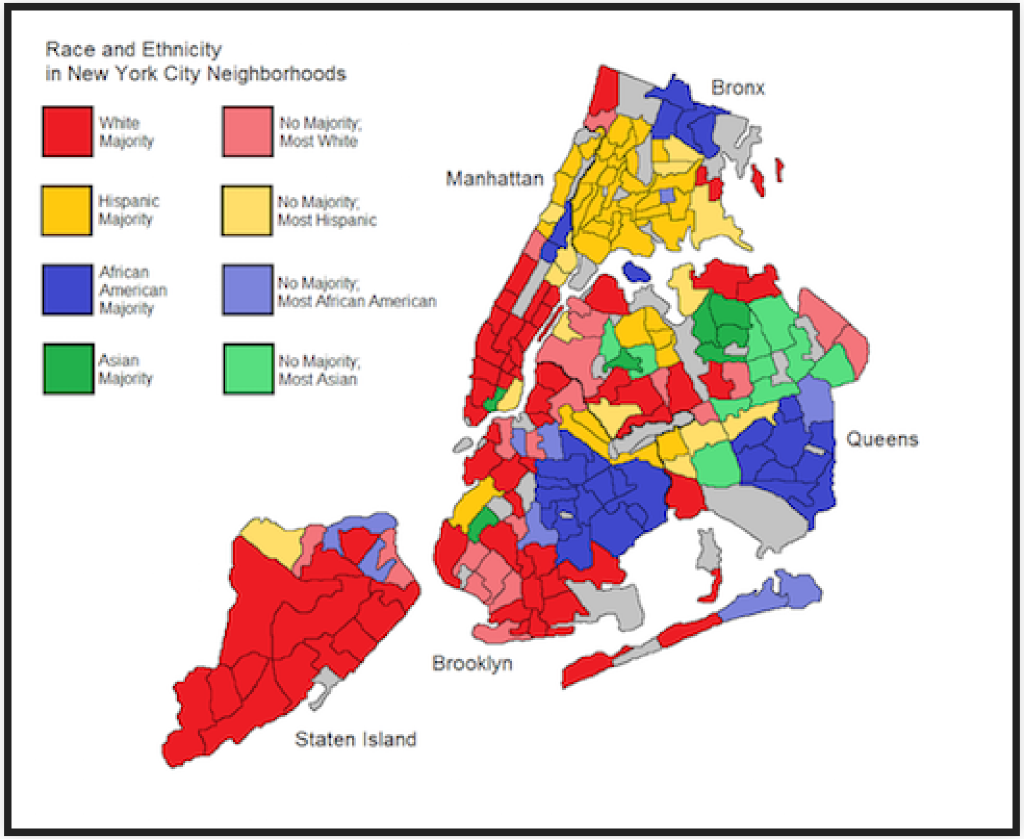
- **Anti-classification (fairness through blindness)**
- Group fairness (independence)
- Separation (equalized odds)

ANTI-CLASSIFICATION

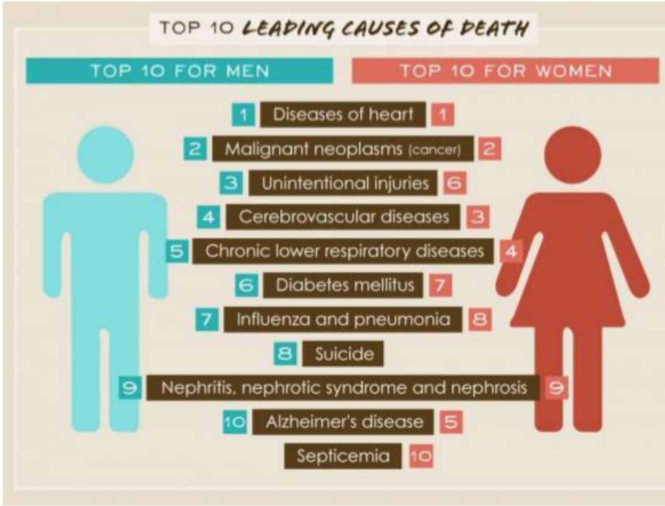
- Also called *fairness through blindness*
- Ignore certain sensitive attributes when making a decision
- Example: Remove gender or race from a credit scoring model
- **Q. Easy to implement, but any limitations**?

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?

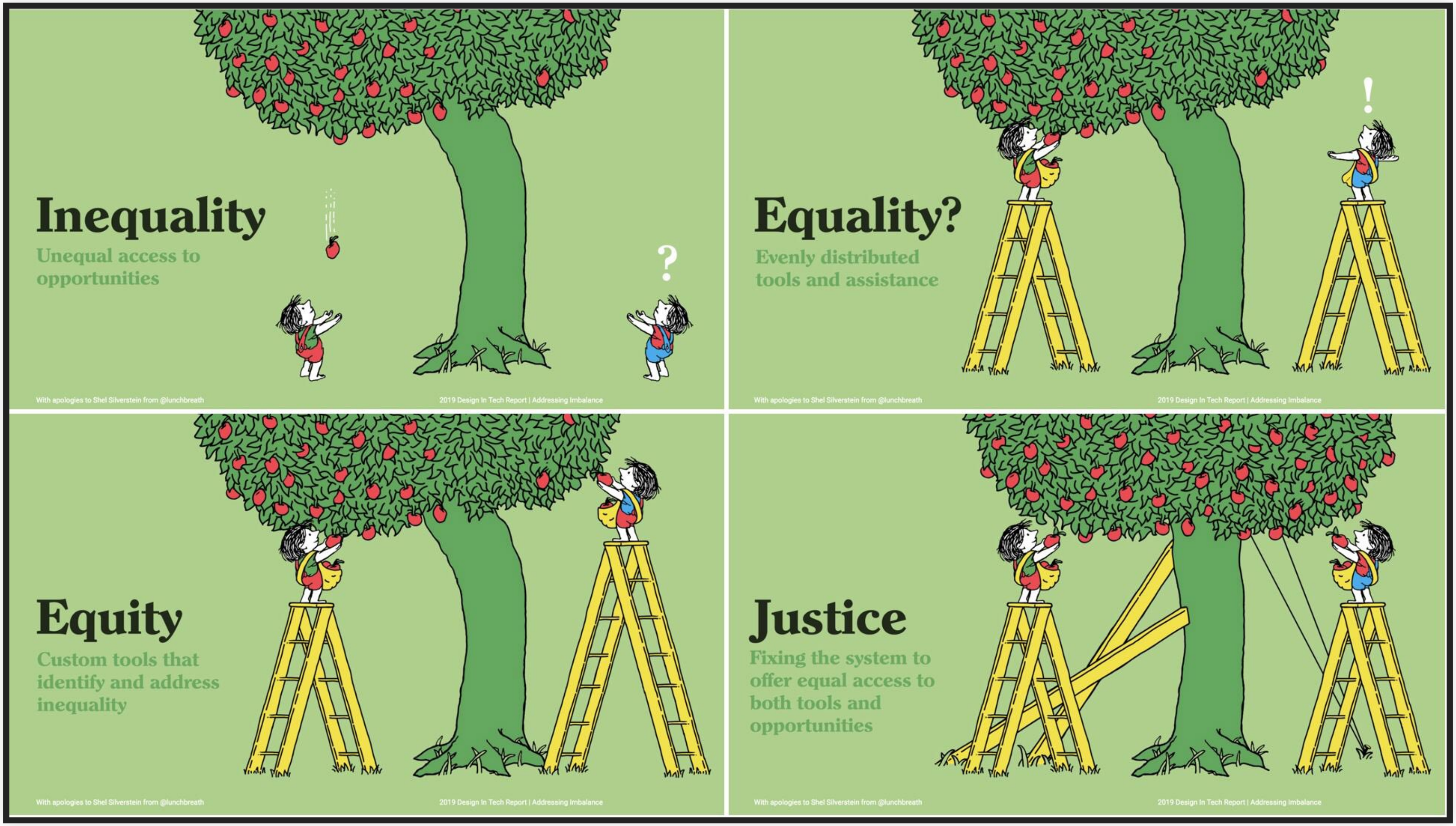
GROUP FAIRNESS

- 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

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

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

REVIEW: CONFUSION MATRIX

		Actual value	
		Y = 1	Y = 0
Predicted value	Y' = 1	True Positive Rate $P[Y' = 1 \mid Y = 1]$	False Positive Rate $P[Y' = 1 \mid Y = 0]$
	Y' = 0	False Negative Rate $P[Y' = 0 \mid Y = 1]$	True Negative Rate $P[Y' = 0 \mid Y = 0]$

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

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

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

SEPARATION

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

- $Y' \perp A | Y$
  - 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

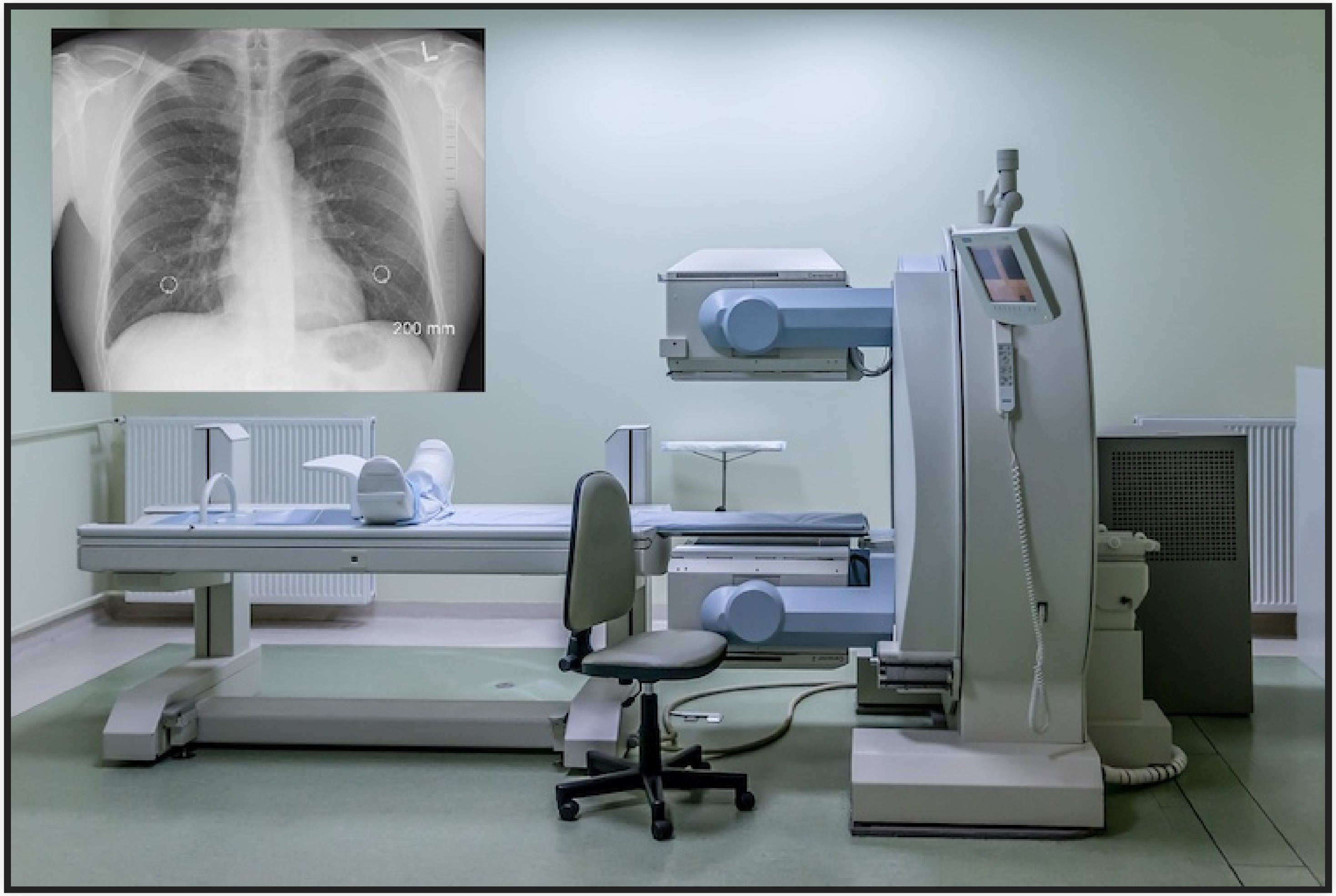
$$P[Y' = 1 \mid Y = 0, A = a] = P[Y' = 1 \mid Y = 0, A = b] \text{ (FPR parity)}$$
$$P[Y' = 0 \mid Y = 1, A = a] = P[Y' = 0 \mid Y = 1, A = b] \text{ (FNR parity)}$$

- $Y' \perp A | Y$
  - 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

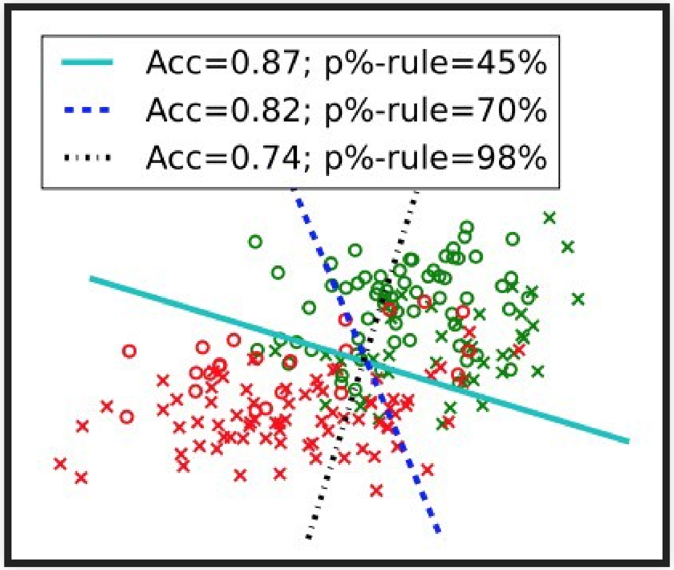
ACHIEVING FAIRNESS CRITERIA

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

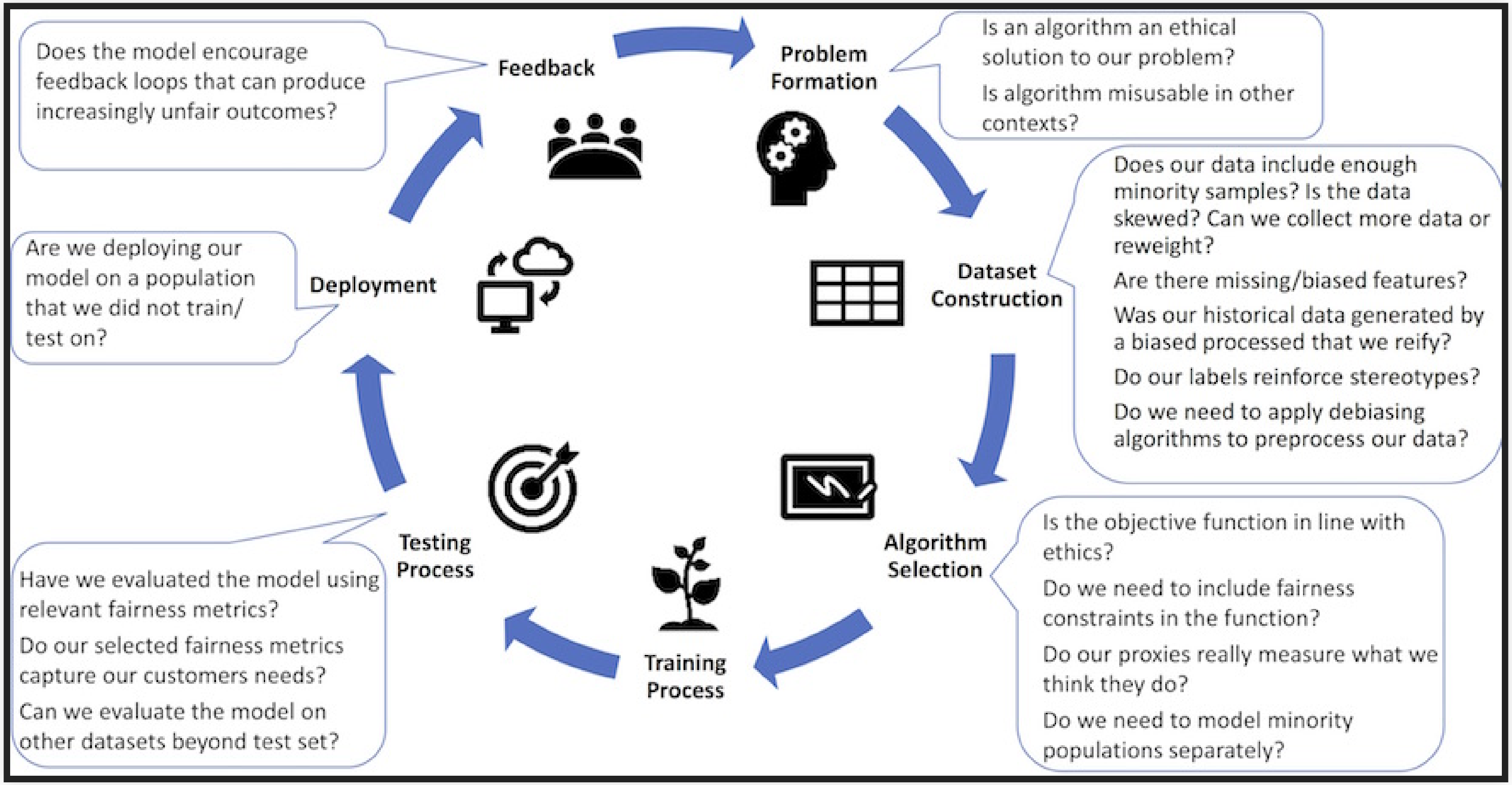


- In general, accuracy is at odds with fairness
  - e.g., Impossible to achieve perfect accuracy ($R = Y$) while ensuring group fairness
- Determine how much compromise in accuracy or fairness is acceptable to your stakeholders

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

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
- Achieving fairness as an activity throughout the entire development cycle