BUILDING FAIRER AI-ENABLED SYSTEMS

Christian Kaestner

(with slides from 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.

Recommended reading: Corbett-Davies, Sam, and Sharad Goel. "The measure and mismeasure of fairness: A critical review of fair machine learning." arXiv preprint arXiv:1808.00023 (2018).

Also revisit: Vogelsang, Andreas, and Markus Borg. "Requirements Engineering for Machine Learning: Perspectives from Data Scientists." In Proc. of the 6th International Workshop on Artificial Intelligence for Requirements Engineering (AIRE), 2019.

LEARNING GOALS

- Understand different definitions of fairness
- Discuss methods for measuring fairness
- Design and execute tests to check for bias/fairness issues
- Understand fairness interventions during data acquisition
- Apply engineering strategies to build more fair systems
- Diagnose potential ethical issues in a given system
- Evaluate and apply mitigation strategies

TWO PARTS

Fairness assessment in the model

Formal definitions of fairness properties

Testing a model's fairness

Constraining a model for fairer results

System-level fairness engineering

Requirements engineering

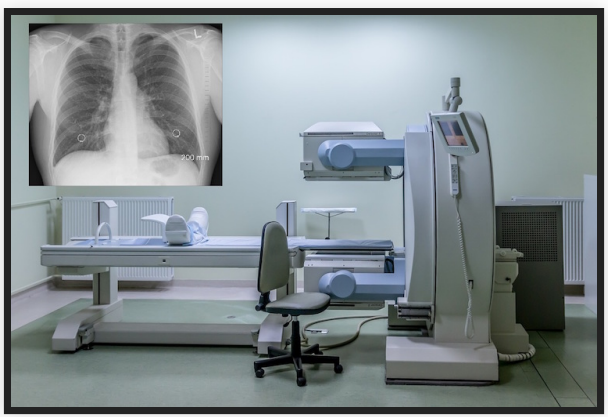
Fairness and data acquisition

Team and process considerations

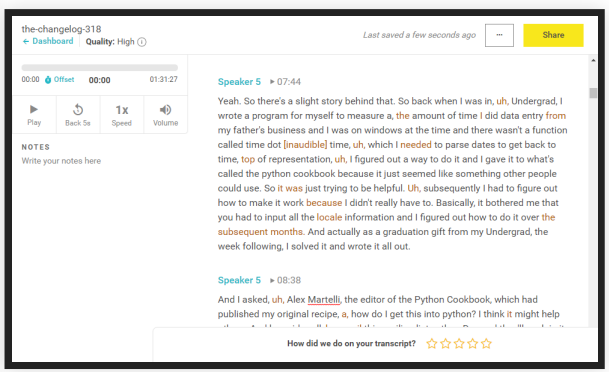
CASE STUDIES

Recidivism

Cancer detection

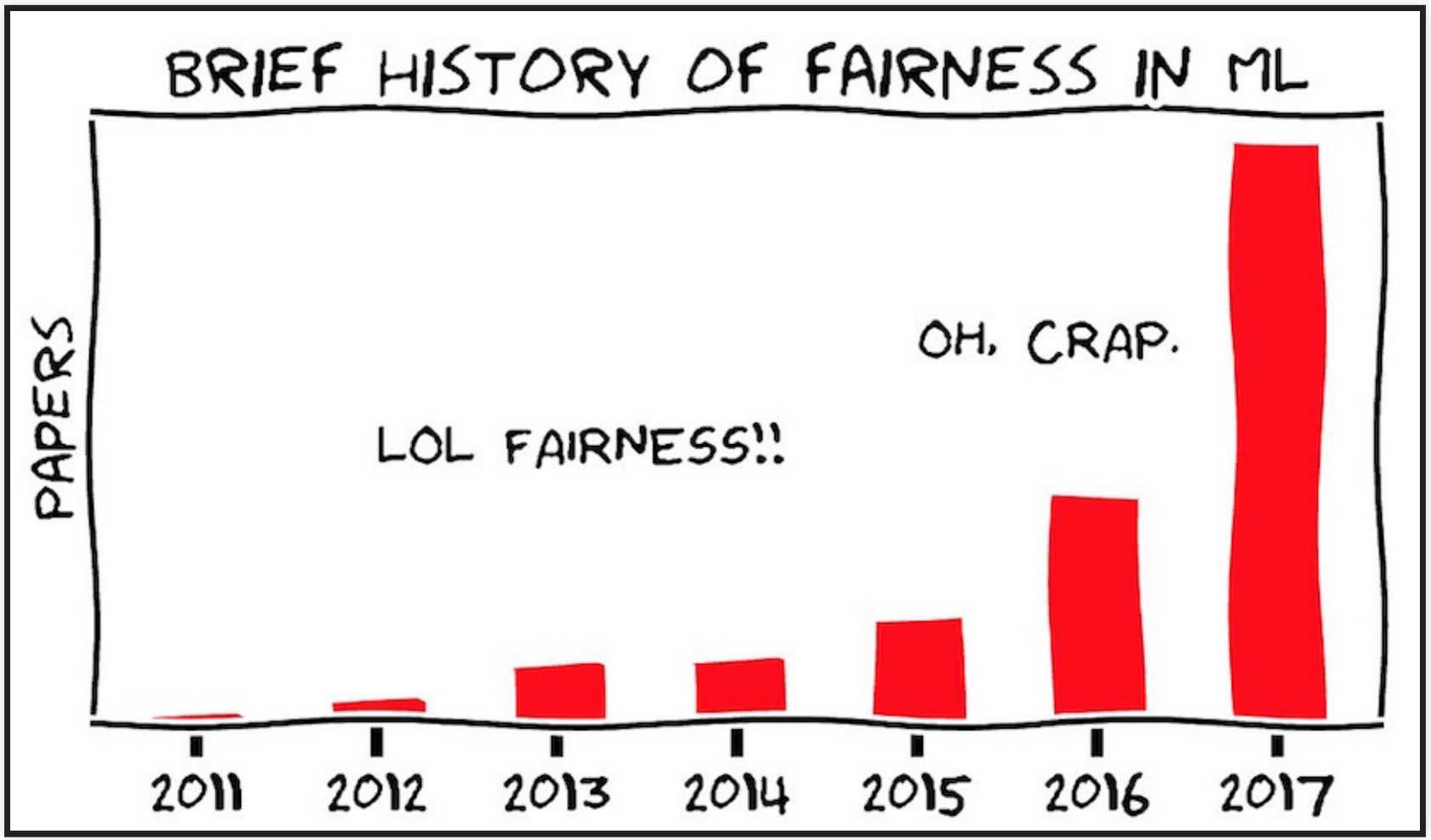


Audio Transcription



FAIRNESS: DEFINITIONS

FAIRNESS IS STILL AN ACTIVELY STUDIED & DISPUTED CONCEPT!



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

PHILOSOPHICAL AND LEGAL ROOTS

- Utility-based fairness: Statistical vs taste-based
    - Statistical discrimination: consider protected attributes in order to achieve non-prejudicial goal (e.g., higher premiums for male drivers)
    - Taste-based discrimination: forgoing benefit to avoid certain transactions (e.g., not hiring better qualified minority candidate), intentional or out of ignorance
- Legal doctrine of fairness focuses on decision maker's motivations ("activing with discriminatory purpose")
    - Forbids intentional taste-based discrimination, allows limited statistical discrimination for compelling government interests (e.g. affirmative action)
- Equal protection doctrine evolved and discusses *classification* (use of protected attributes) vs *subordination* (subjugation of disadv. groups)
    - anticlassification firmly encoded in legal standards
    - use of protected attributes triggers judicial scrutiny, but allowed to serve higher interests (e.g. affirmative action)
- In some domains, intent-free economic discrimination considered
    - e.g. *disparate impact* standard in housing
    - practice illegal if it has *unjust outcomes* for protected groups, even in absence of classification or animus (e.g., promotion requires high-school diploma)

Further reading: Corbett-Davies, Sam, and Sharad Goel. "The measure and mismeasure of fairness: A critical review of fair machine learning." arXiv preprint arXiv:1808.00023 (2018).

Speaker notes

On disparate impact from Corbett-Davies et al:

> *"In 1955, the Duke Power Company instituted a policy that mandated employees have a high school diploma to be considered for promotion, which had the effect of drastically limiting the eligibility of black employees. The Court found that this requirement had little relation to job performance, and thus deemed it to have an unjustified—and illegal—disparate impact. Importantly, the employer's motivation for instituting the policy was irrelevant to the Court's decision; even if enacted without discriminatory pur- pose, the policy was deemed discriminatory in its effects and hence illegal. Note, however, that disparate impact law does not prohibit all group differences produced by a policy—the law only prohibits unjustified disparities. For example, if, hypothetically, the high-school diploma requirement in Griggs were shown to be necessary for job success, the resulting disparities would be legal."*

DEFINITIONS OF ALGORITHMIC FAIRNESS

- Anti-classifcation (Fairness through Blindness)
- Group fairness
- Equalized odds
- Predictive rate parity

ANTI-CLASSIFICATION

Protected attributes are not used

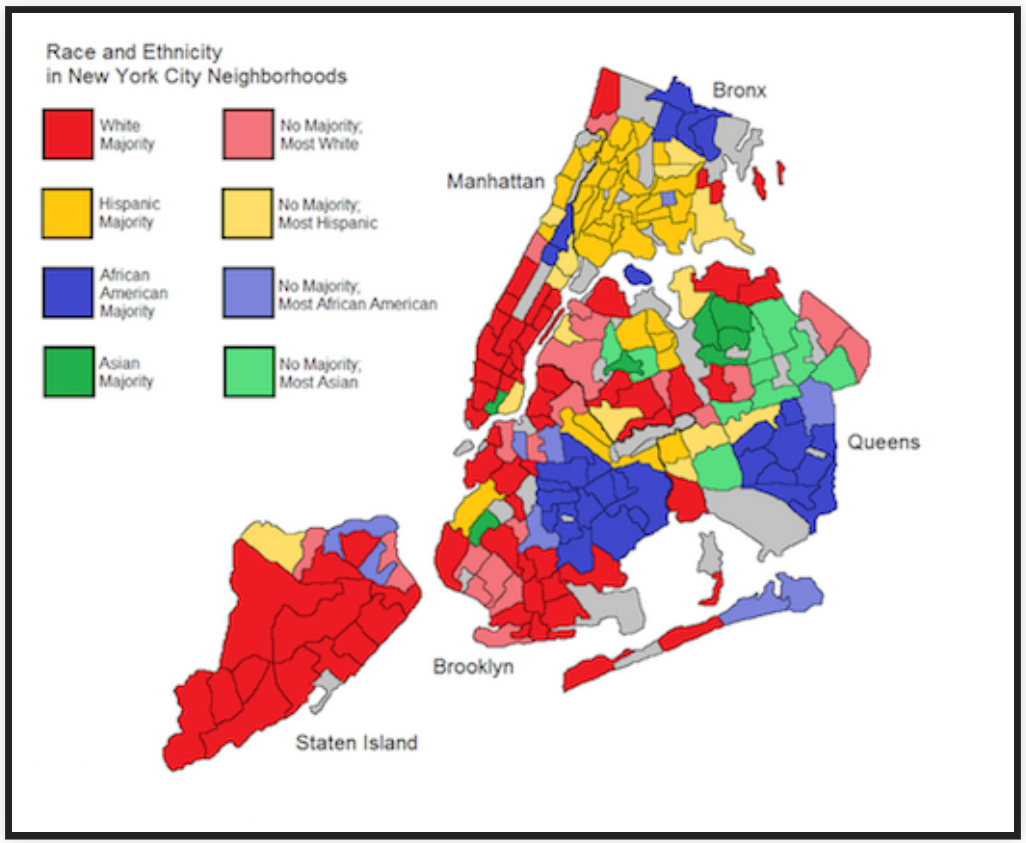
FAIRNESS THROUGH BLINDNESS

Anti-classification: Ignore/eliminate sensitive attributes from dataset, e.g., remove gender and race from a credit card scoring system

Advantages? Problems?

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?

TECHNICAL SOLUTION FOR ANTI-CLASSIFICATION?

Speaker notes

- Remove protected attributes from dataset
- Zero out all protected attributes in training and input data

TESTING 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 (see prior lecture on Automated Random Testing) or on any test data

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

CORRELATED FEATURES

- Test correlation between protected attributes and other features
- Remove correlated features ("suspect causal path") as well

ON TERMINOLOGY

- Lots and lots of recent papers on fairness in AI
- Long history of fairness discussions in philosophy and other fields
- Inconsistent terminology, reinvention, many synonyms and some homonyms
    - e.g. anti-classification = fairness by blindness = causal fairness

CLASSIFICATION PARITY

Classification error is equal across groups

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. "Fairness and machine learning: Limitations and Opportunities." (2019), Chapter 2

NOTATIONS

- X: Feature set (e.g., age, race, education, region, income, etc.,)
- A: Sensitive attribute (e.g., race)
- R: Regression score (e.g., predicted likelihood of recidivism)
    - $Y^{'}$ = 1 if and only if R is greater than some threshold
- Y: Target variable (e.g. did the person actually commit recidivism?)

INDEPENDENCE

(aka *statistical parity*, *demographic parity*, *disparate impact*, *group fairness*)

$$P[R = 1 \mid A = 0] = P[R = 1 \mid A = 1] \text{ or } R \perp A$$

- *Acceptance rate* (i.e., percentage of positive predictions) must be the same across all groups
- Prediction must be independent of the sensitive attribute
- Example:
    - The predicted rate of recidivism is the same across all races
    - Chance of promotion the same across all genders

INDEPENDENCE VS. ANTI-DISCRIMINATION

Speaker notes

Independence is to be observed on actual input data, needs representative test data selection

TESTING INDEPENDENCE

- Separate validation/telemetry data by protected attribute
  - Or generate *realistic* test data, e.g. from probability distribution of population (see prior lecture on Automated Random Testing)
- Separately measure rate of positive predictions
- Report issue if rate differs beyond ϵ across groups

EXERCISE: CANCER DIAGNOSIS

True Positives (TPs): 16	False Positives (FPs): 4
False Negatives (FNs): 6	True Negatives (TNs): 974

Male Patient Results

True Positives (TPs): 6	False Positives (FPs): 3
False Negatives (FNs): 5	True Negatives (TNs): 486

Female Patient Results

True Positives (TPs): 10	False Positives (FPs): 1
False Negatives (FNs): 1	True Negatives (TNs): 488

- 1000 data samples (500 male & 500 female patients)
- What's the overall recall & precision?
- Does the model achieve *independence*

LIMITATIONS OF INDEPENDENCE?

Speaker notes

- No requirement that predictions are any good in either group
    - e.g. intentionally hire bad people from one group to afterward show that that group performs poorly in general
- Ignores possible correlation between Y and A
- Rules out perfect predictor $R = Y$ when Y & A are correlated
- Permits laziness: Intentionally give high ratings to random people in one group

CALIBRATION TO ACHIEVE INDEPENDENCE

Select different thresholds for different groups to achieve prediction parity:

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

Lowers bar for some groups -- equity, not equality

Equality

The assumption is that **everyone benefits from the same supports**. This is equal treatment.

Equity

Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.

Justice

All 3 can see the game without supports or accommodations because **the cause(s) of the inequity was addressed**. The systemic barrier has been removed.

SEPARATION / EQUALIZED ODDS

Prediction must be independent of the sensitive attribute conditional on the target variable: $R \perp A | Y$

Same true positive rate across groups:

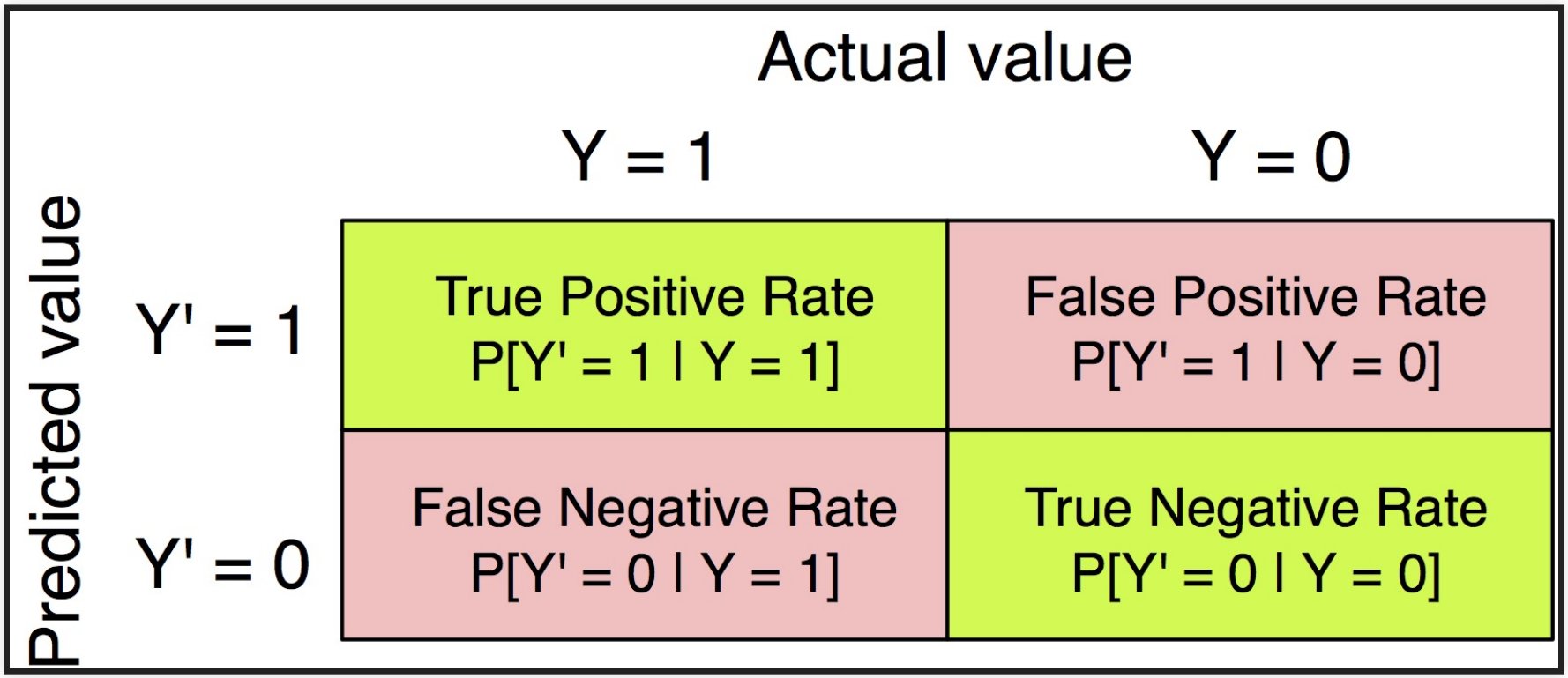
$$P[R = 0 \mid Y = 1, A = 0] = P[R = 0 \mid Y = 1, A = 1]$$

And same false positive rate across groups:

$$P[R = 1 \mid Y = 0, A = 0] = P[R = 1 \mid Y = 0, A = 1]$$

Example: A person with good credit behavior score should be assigned a good score with the same probability regardless of gender

RECALL: CONFUSION MATRIX



Can we explain equalize odds in terms of errors?

$$P[R = 0 \mid Y = 1, A = a] = P[R = 0 \mid Y = 1, A = b]$$
$$P[R = 1 \mid Y = 0, A = a] = P[R = 1 \mid Y = 0, A = b]$$

EXERCISE: CANCER DIAGNOSIS

True Positives (TPs): 16	False Positives (FPs): 4
False Negatives (FNs): 6	True Negatives (TNs): 974

Male Patient Results

True Positives (TPs): 6	False Positives (FPs): 3
False Negatives (FNs): 5	True Negatives (TNs): 486

Female Patient Results

True Positives (TPs): 10	False Positives (FPs): 1
False Negatives (FNs): 1	True Negatives (TNs): 488

- 1000 data samples (500 male & 500 female patients)
- What's the overall recall & precision?
- Does the model achieve *separation*

DISCUSSION: SEPARATION/EQUALIZED ODDS

(All groups experience the same false positive & negative rates)

Separation vs independence? Limitations of separation?

Equality

The assumption is that **everyone benefits from the same supports**. This is equal treatment.

Equity

Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.

Justice

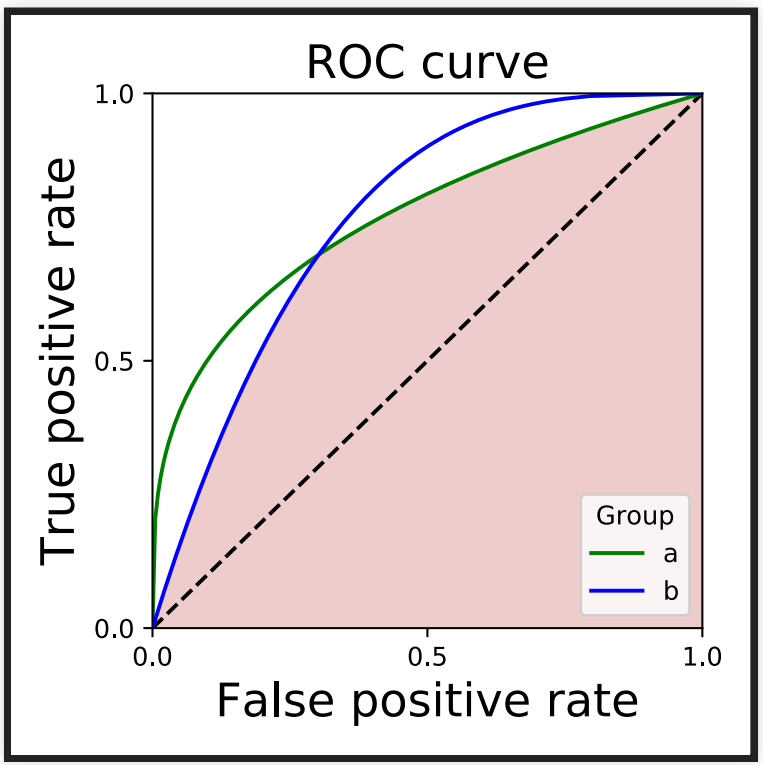
All 3 can see the game without supports or accommodations because **the cause(s) of the inequity was addressed**. The systemic barrier has been removed.

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 (see prior lecture on Automated Random Testing)
- Separately measure false positive and false negative rate

CALIBRATION FOR SEPARATION

- Adjust threshold across all groups to balance false positives vs. false negatives (see ROC curves)



Speaker notes

Shaded curve describes possible tradeoffs, not all rates possible that would be possible for just one group, i.e. overall degradation common.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. "Fairness and machine learning: Limitations and Opportunities." (2019), Chapter 2

MANY RELATED DEFINITIONS OF CLASSIFICATION PARITY

- Classification parity measures based on different metrics from confusion matrix
- Separation only based on false positives or false negatives (when only one outcome matters more, e.g., denied opportunities in hiring)
- Comparisons of other error definitions, e.g. recall and precision
    - *Sufficiency* or *predictive rate parity*
    - same precision across groups

OUTLOOK: UTILITARIAN VIEW WITH THRESHOLD RULES

- Identify costs/benefits from each outcome (TP, FP, TN, FN)
- Costs and benefits may be different across different individuals/groups
- Calibrate thresholds to equalize utility across groups (even if it violates independence or separation)

Corbett-Davies, Sam, and Sharad Goel. "The measure and mismeasure of fairness: A critical review of fair machine learning." arXiv preprint arXiv:1808.00023 (2018).

IMPOSSIBILITY RESULTS

- Many classification parity definitions cannot be achieved at the same time
- e.g., Impossible to achieve equalized odds and predictive rate parity
  - $R \perp A \mid Y$ and $Y \perp A \mid R$ can't be true at the same time
  - Unless $A \perp Y$
  - Formal proofs: Chouldechova (2016), Kleinberg et al. (2016)

Equality

The assumption is that **everyone benefits from the same supports**. This is equal treatment.

Equity

Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.

Justice

All 3 can see the game without supports or accommodations because **the cause(s) of the inequity was addressed**. The systemic barrier has been removed.

Speaker notes

Equity and equality relate to goals and are assessed with different measures. May not be compatible.

REVIEW OF CRITERIA SO FAR:

Recidivism scenario: Should a person be detained?

- Anti-classification: ?
- Independence: ?
- 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
- Independence: 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

REFLECTION: CANCER DIAGNOSIS

True Positives (TPs): 16	False Positives (FPs): 4
False Negatives (FNs): 6	True Negatives (TNs): 974

Male Patient Results

True Positives (TPs): 6	False Positives (FPs): 3
False Negatives (FNs): 5	True Negatives (TNs): 486

Female Patient Results

True Positives (TPs): 10	False Positives (FPs): 1
False Negatives (FNs): 1	True Negatives (TNs): 488

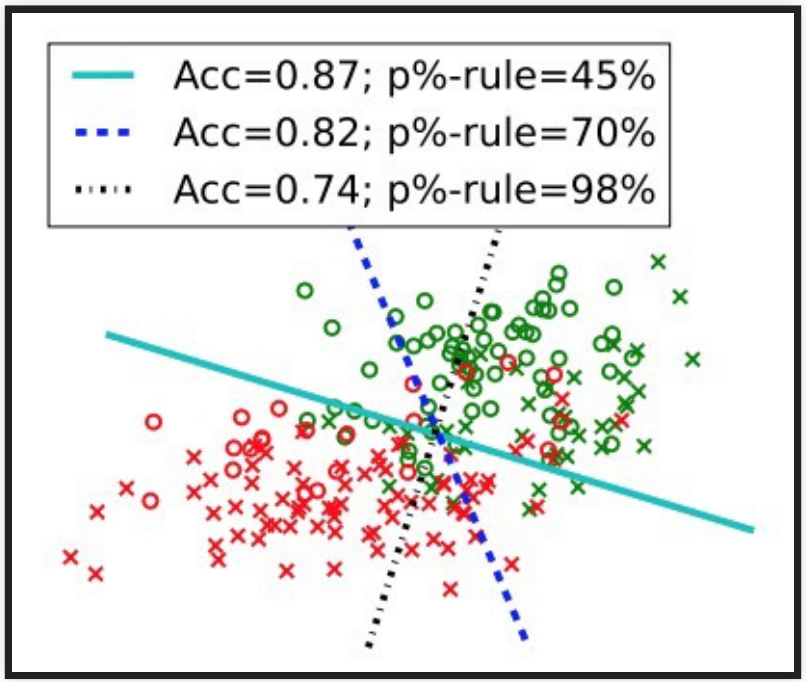
What can we conclude about the model & its usage?

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-time constraint
  - ML is a constraint optimization problem (minimize errors)
  - Impose additional parity constraint into ML optimization process (e.g., as part of the loss function)
- Post-processing
  - Adjust the learned model to be uncorrelated with sensitive attributes
  - Adjust thresholds
- (Still active area of research! Many new techniques published each year)

TRADE-OFFS: ACCURACY VS FAIRNESS

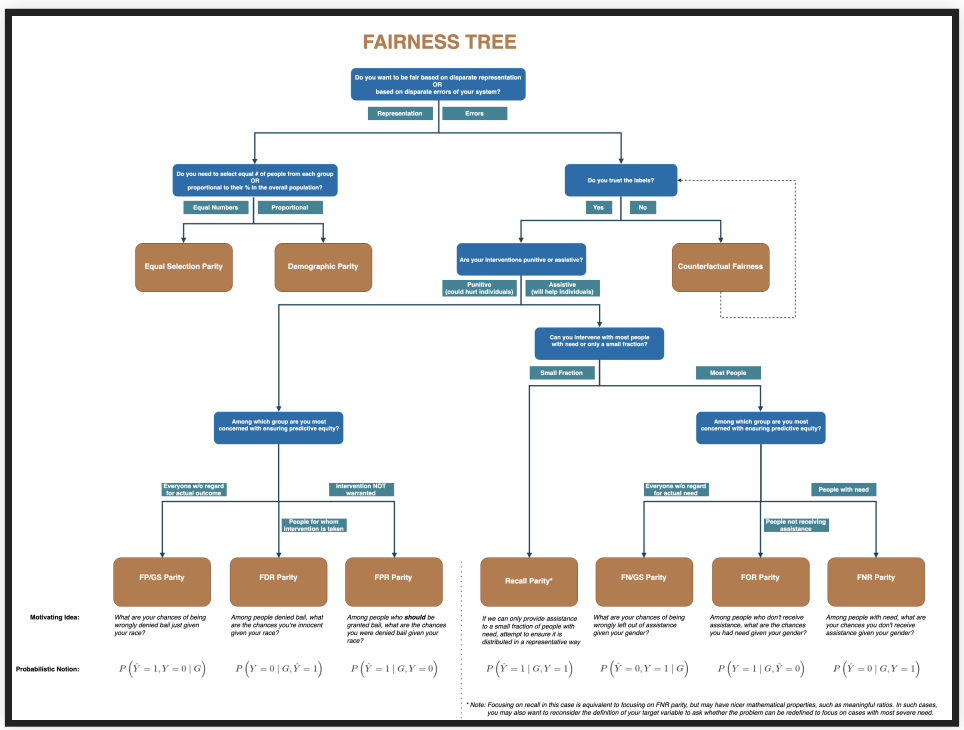


- Fairness constraints possible models
- Fairness constraints often lower accuracy for some group

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

PICKING FAIRNESS CRITERIA

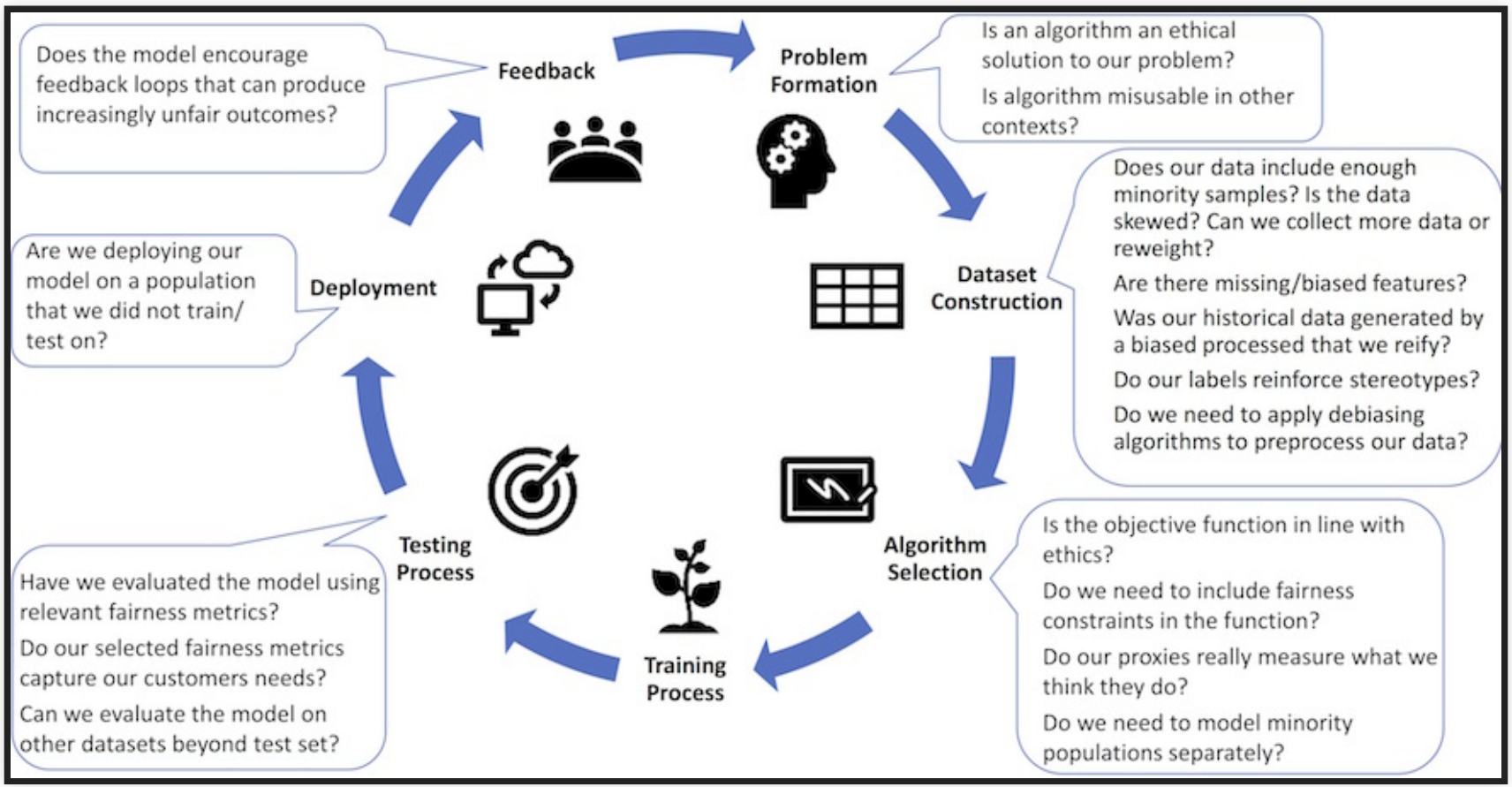
- Requirements engineering problem!
- What's the goal of the system? What do various stakeholders want? How to resolve conflicts?



http://www.datasciencepublicpolicy.org/projects/aequitas/

BEYOND THE MODEL

FAIRNESS MUST BE CONSIDERED THROUGHOUT THE ML LIFECYCLE!



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

PRACTITIONER CHALLENGES

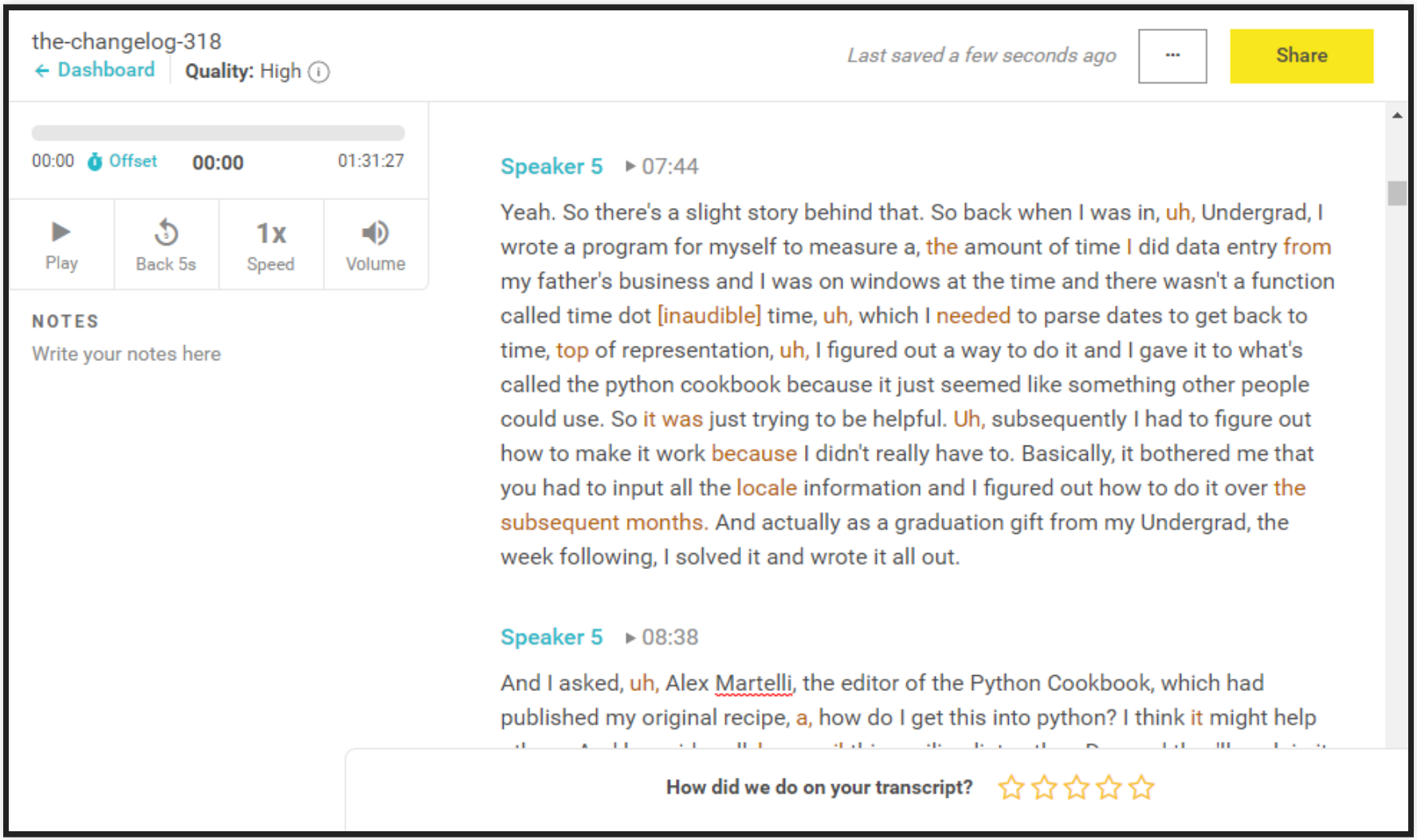
- Fairness is a system-level property
  - consider goals, user interaction design, data collection, monitoring, model interaction (properties of a single model may not matter much)
- Fairness-aware data collection, fairness testing for training data
- Identifying blind spots
  - Proactive vs reactive
  - 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.

START EARLY

- Think about system goals and relevant fairness concerns
- Analyze risks
- Understand environment interactions, attacks, and feedback loops (world vs machine)
- Influence data acquisition
- Define quality assurance procedures
    - separate test sets, automatic fairness measurement, testing in production
    - telemetry design and feedback mechanisms
    - incidence response plan

EXERCISE: WHAT WOULD YOU DO?



THE ROLE OF REQUIREMENTS ENGINEERING

- Identify system goals
- Identify legal constraints
- Identify stakeholders and fairness concerns
- Analyze risks with regard to discrimination and fairness
- Analyze possible feedback loops (world vs machine)
- Negotiate tradeoffs with stakeholders
- Set requirements/constraints for data and model
- Plan mitigations in the system (beyond the model)
- Design incident response plan
- Set expectations for offline and online assurance and monitoring

THE ROLE OF SOFTWARE ENGINEERS

- Whole system perspective
- Requirements engineering, identifying stakeholders
- Tradeoff decisions among conflicting goals
- Interaction and interface design
- Infrastructure for evaluating model quality and fairness offline and in production
- Monitoring
- System-wide mitigations (in model and beyond model)

BEST PRACTICES: TASK DEFINITION

- Clearly define the task & model's intended effects
- Try to identify and document unintended effects & biases
- Clearly define any fairness requirements
- *Involve diverse stakeholders & multiple perspectives*
- Refine the task definition & be willing to abort

Swati Gupta, Henriette Cramer, Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, Hanna Wallach, Sravana Reddy, Jean GarciaGathright. Challenges of incorporating algorithmic fairness into practice, FAT* Tutorial, 2019. (slides)

BEST PRACTICES: CHOOSING A DATA SOURCE

- Think critically before collecting any data
- Check for biases in data source selection process
- Try to identify societal biases present in data source
- Check for biases in cultural context of data source
- Check that data source matches deployment context
- Check for biases in
    - technology used to collect the data
    - humans involved in collecting data
    - sampling strategy
- *Ensure sufficient representation of subpopulations*
- Check that collection process itself is fair & ethical

How can we achieve fairness without putting a tax on already disadvantaged populations?

Swati Gupta, Henriette Cramer, Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, Hanna Wallach, Sravana Reddy, Jean GarciaGathright. Challenges of incorporating algorithmic fairness into practice, FAT* Tutorial, 2019. (slides)

BEST PRACTICES: LABELING AND PREPROCESSING

- Check for biases introduced by
  - discarding data
  - bucketing values
  - preprocessing software
  - labeling/annotation software
  - human labelers
- Data/concept drift?

Auditing? Measuring bias?

Swati Gupta, Henriette Cramer, Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, Hanna Wallach, Sravana Reddy, Jean GarciaGathright. Challenges of incorporating algorithmic fairness into practice, FAT* Tutorial, 2019. (slides)

BEST PRACTICES: MODEL DEFINITION AND TRAINING

- Clearly define all assumptions about model
- Try to identify biases present in assumptions
- Check whether model structure introduces biases
- Check objective function for unintended effects
- Consider including “fairness” in objective function

Swati Gupta, Henriette Cramer, Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, Hanna Wallach, Sravana Reddy, Jean GarciaGathright. Challenges of incorporating algorithmic fairness into practice, FAT* Tutorial, 2019. (slides)

BEST PRACTICES: TESTING & DEPLOYMENT

- Check that test data matches deployment context
- Ensure test data has sufficient representation
- Continue to involve diverse stakeholders
- Revisit all fairness requirements
- Use metrics to check that requirements are met

- Continually monitor
  - match between training data, test data, and instances you encounter in deployment
  - fairness metrics
  - population shifts
  - user reports & user complaints
- Invite diverse stakeholders to audit system for biases

Swati Gupta, Henriette Cramer, Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, Hanna Wallach, Sravana Reddy, Jean GarciaGathright. Challenges of incorporating algorithmic fairness into practice, FAT* Tutorial, 2019. (slides)

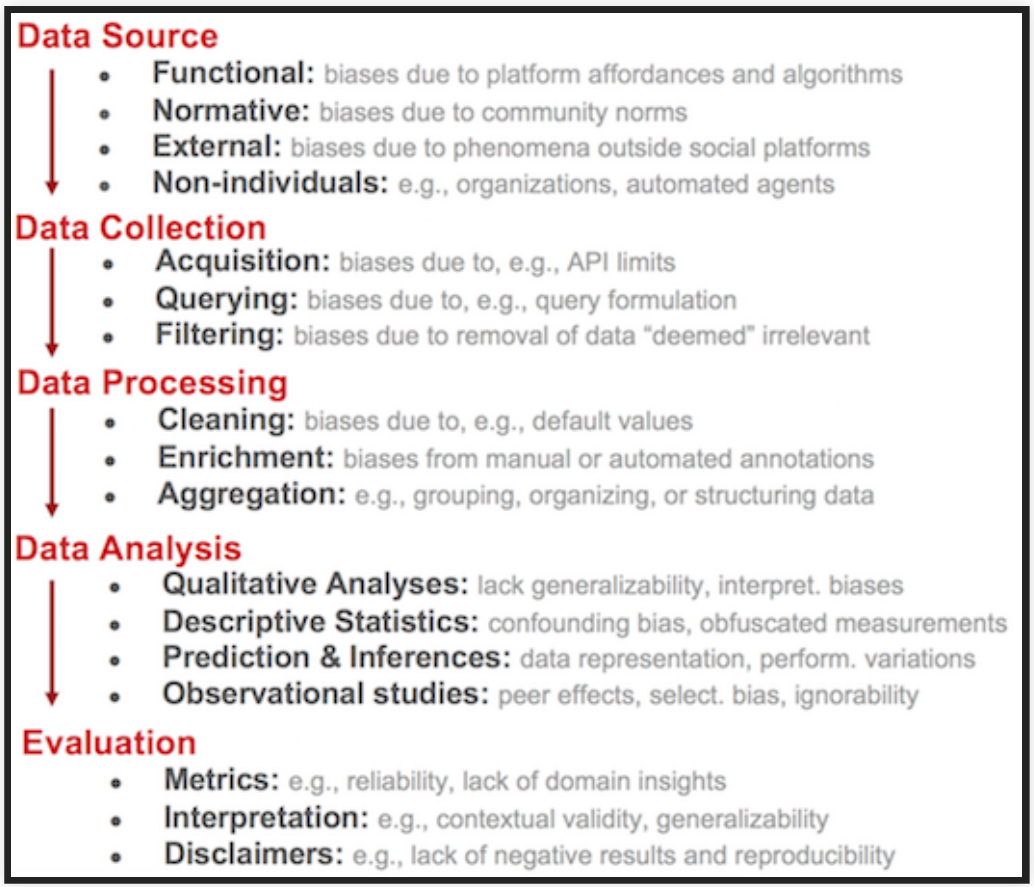
DATASET CONSTRUCTION FOR FAIRNESS

FLEXIBILITY IN DATA COLLECTION

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

Swati Gupta, Henriette Cramer, Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, Hanna Wallach, Sravana Reddy, Jean GarciaGathright. Challenges of incorporating algorithmic fairness into practice, FAT* Tutorial, 2019. (slides)

Bias can be introduced at any stage of the data pipeline



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

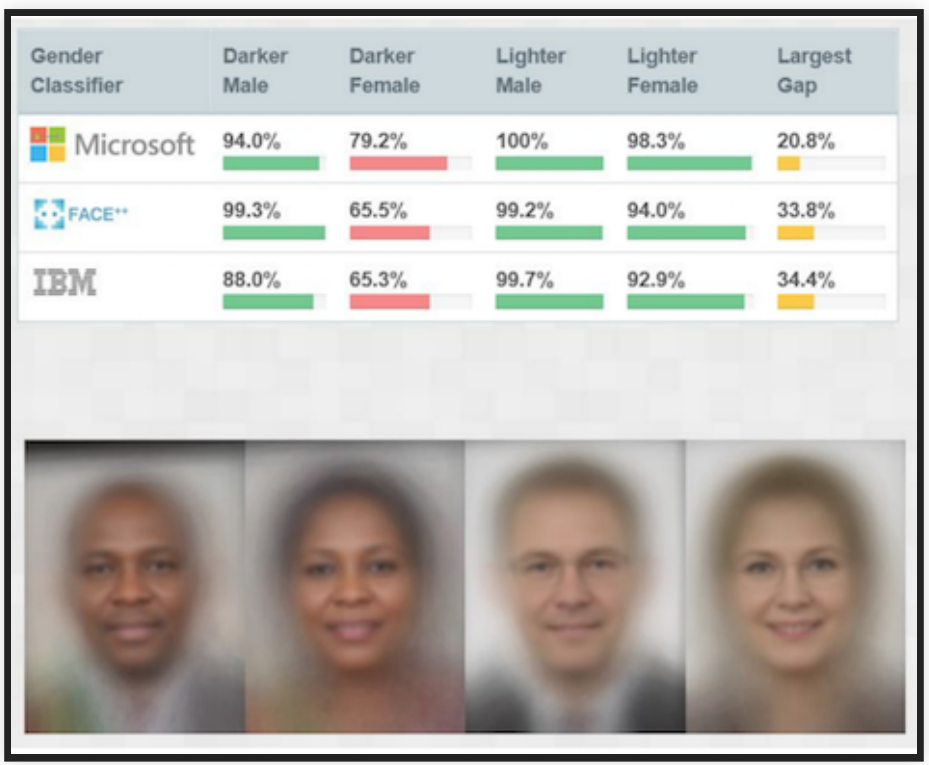
TYPES OF DATA BIAS

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

Olteanu et al., Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries, Olteanu et al., Frontiers in Big Data (2019).

POPULATION BIAS

- Differences in demographics between a dataset vs a target population
- Example: Does the Twitter demographics represent the general population?
- In many tasks, datasets should match the target population
- But some tasks require equal representation for fairness



BEHAVIORAL BIAS

- Differences in user behavior across platforms or social contexts

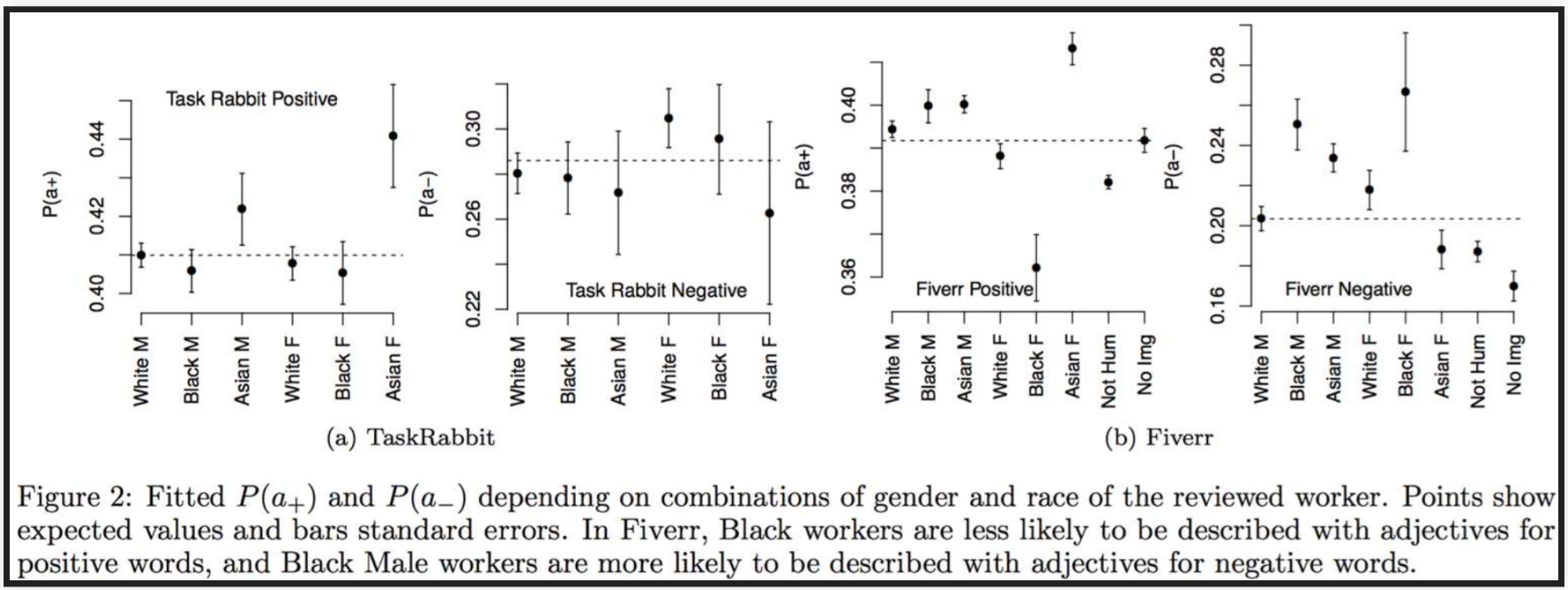


Figure 2: Fitted $P(a_+)$ and $P(a_-)$ depending on combinations of gender and race of the reviewed worker. Points show expected values and bars standard errors. In Fiverr, Black workers are less likely to be described with adjectives for positive words, and Black Male workers are more likely to be described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

Example: Freelancing platforms (Fiverr vs TaskRabbit): Bias against certain minority groups on different platforms

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

FAIRENESS-AWARE DATA COLLECTION

- Address population bias
  - Does the dataset reflect the demographics in the target population?
- Address under- & over-representation issues
  - Ensure sufficient amount of data for all groups to avoid being treated as "outliers" by ML
  - But also avoid over-representation of certain groups (e.g., remove historical data)
- Data augmentation: Synthesize data for minority groups
  - Observed: "He is a doctor" -> synthesize "She is a doctor"
- Fairness-aware active learning
  - Collect more data for groups with highest error rates

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

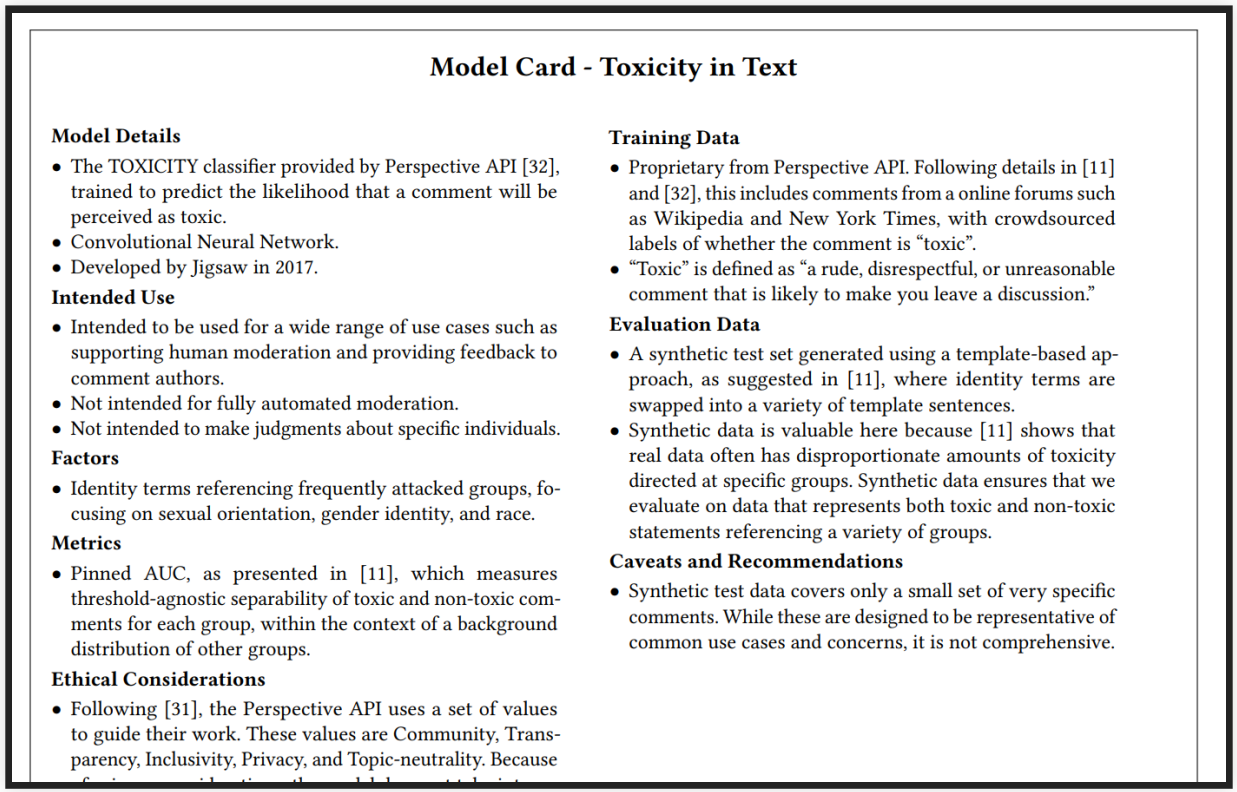
DATA SHEETS

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
- Based on common practice in the electronics industry, medicine
- Purpose, provenance, creation, composition, distribution: Does the dataset relate to people? Does the dataset identify any subpopulations?

Datasheets for Dataset, Gebru et al., (2019).

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.

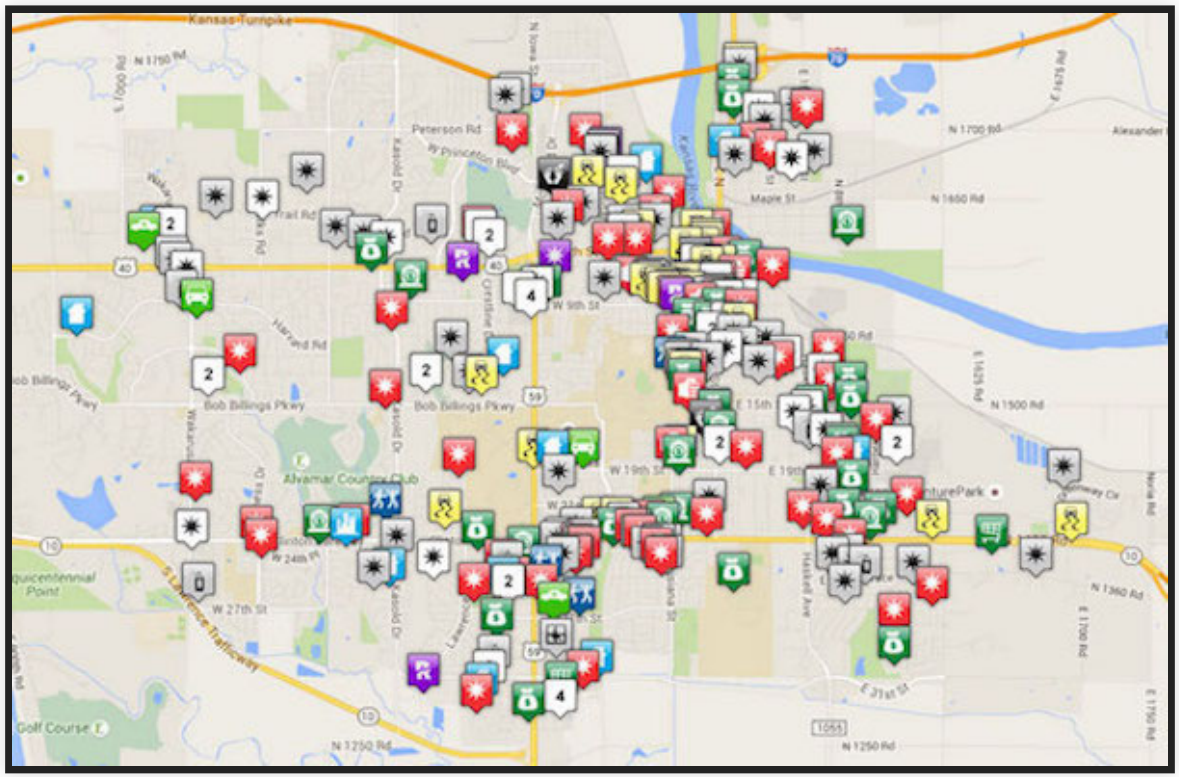
Caveats and Recommendations

- 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.

EXERCISE: CRIME MAP



How can we modify an existing dataset or change the data collection process to reduce the effects the feedback loop?

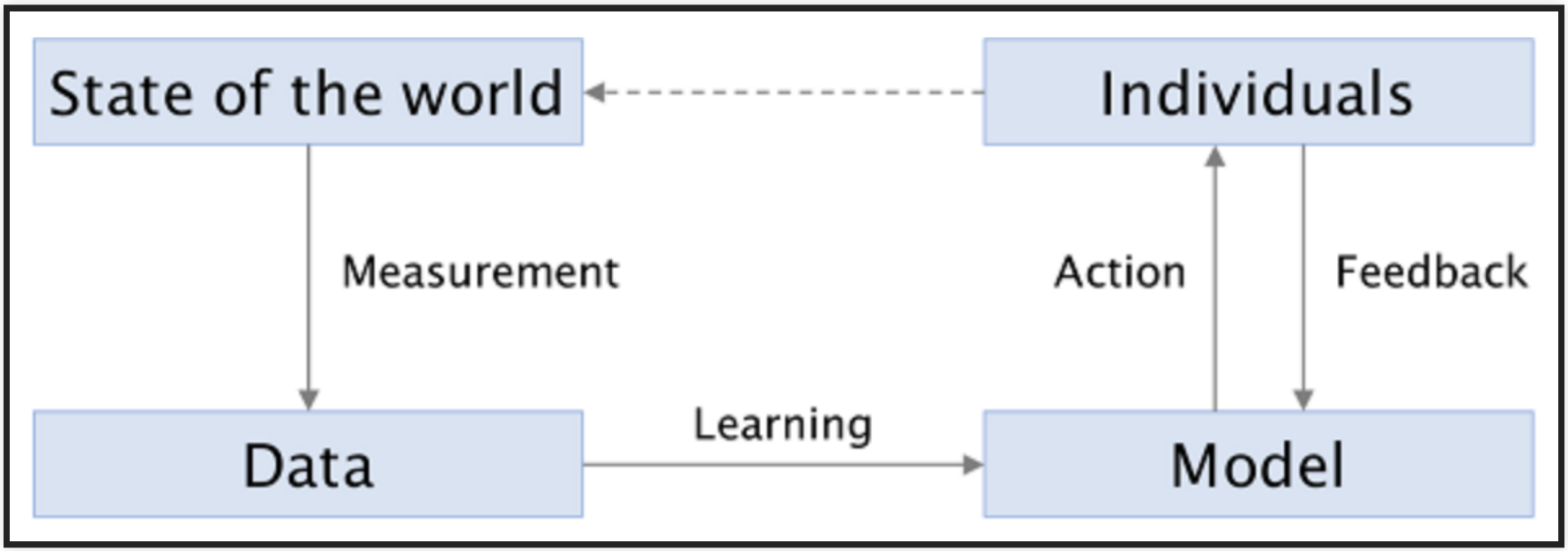
SUMMARY

- Fairness at the model level
  - Fairness definitions and their tradeoffs: anti-classification, classification parity (independence, separation), calibration, ...
  - Achieving fairness through preprocessing, training constraints, postprocessing
  - Fairness vs accuracy
- Fairness at the system level
  - Fairness throughout the lifecycle
  - Dataset construction for fairness
  - Many practical challenges
  - Requirements engineering is essential
  - Best practices and guidelines

APPENDIX: REQUIREMENTS AND FAIRNESS

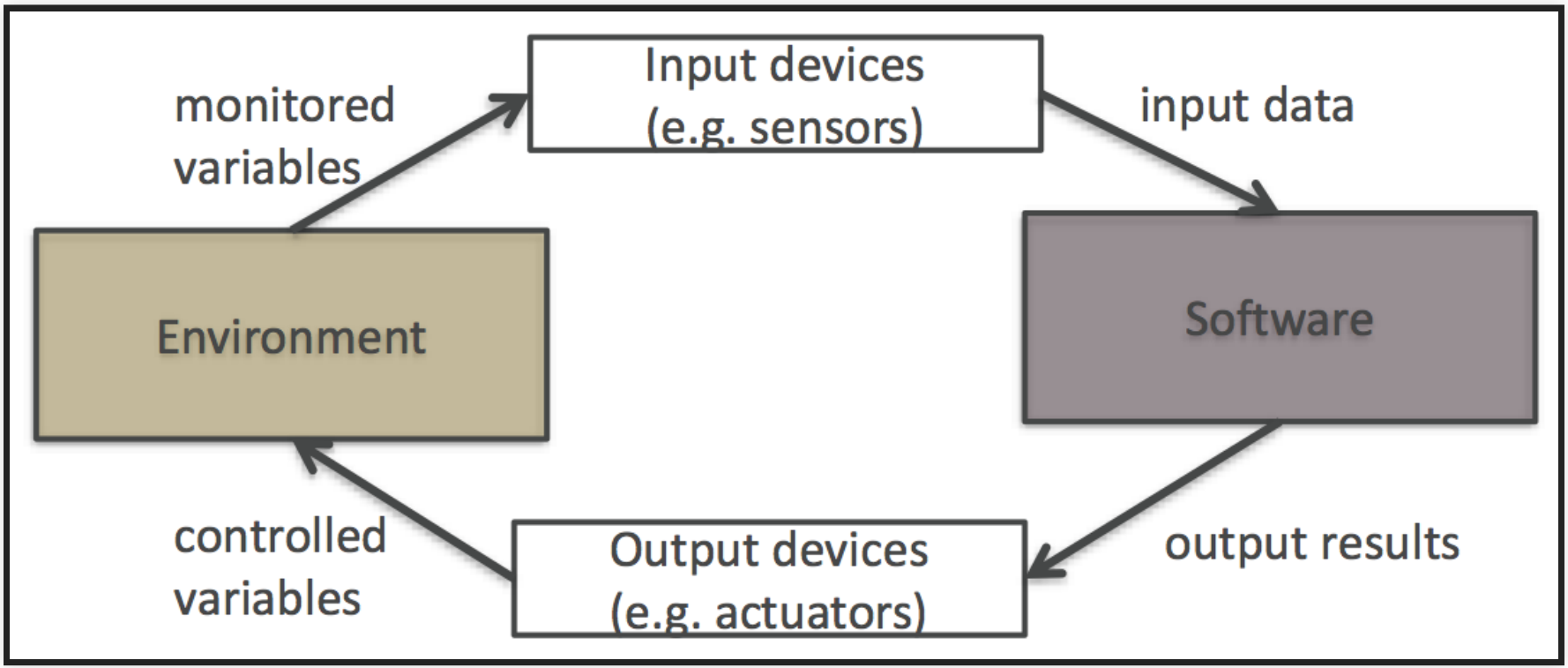
By Eunsuk Kang

MACHINE LEARNING CYCLE



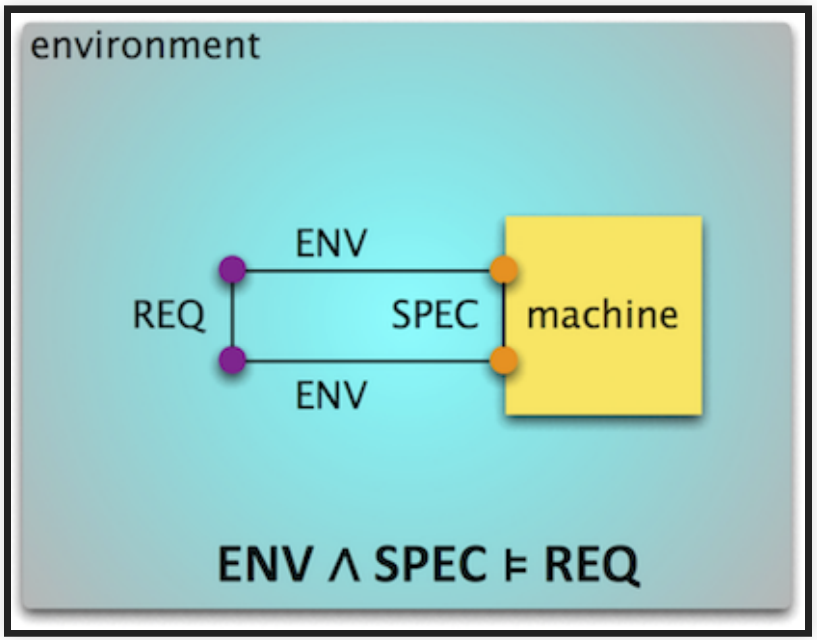
"Fairness and Machine Learning" by Barocas, Hardt, and Narayanan (2019), Chapter 1.

RECALL: MACHINE VS WORLD



- No ML/AI lives in vacuum; every system is deployed as part of the world
- A requirement describes a desired state of the world (i.e., environment)
- Machine (software) is *created* to manipulate the environment into this state

REQUIREMENT VS SPECIFICATION



- Requirement (REQ): What the system should do, as desired effects on the environment
- Assumptions (ENV): What's assumed about the behavior/properties of the environment (based on domain knowledge)
- Specification (SPEC): What the software must do in order to satisfy REQ

CASE STUDY: COLLEGE ADMISSION

REQUIREMENTS FOR FAIR ML SYSTEMS

REQUIREMENTS FOR FAIR ML SYSTEMS

1. Identify all environmental entities
    - Consider all stakeholders, their backgrounds & characteristics

REQUIREMENTS FOR FAIR ML SYSTEMS

1. Identify all environmental entities
    - Consider all stakeholders, their backgrounds & characteristics
2. State requirement (REQ) over the environment
    - What functions should the system serve? Quality attributes?
    - But also: What kind of harms are possible & should be minimized?
    - Legal & policy requirements

"FOUR-FIFTH RULE" (OR "80% RULE")

$$(P[R = 1 \,|\, A = a]) / (P[R = 1 \,|\, A = b]) \geq 0.8$$

- Selection rate for a protected group (e.g., $A = a$) < 80% of highest rate => selection procedure considered as having "adverse impact"
- Guideline adopted by Federal agencies (Department of Justice, Equal Employment Opportunity Commission, etc.,) in 1978
- If violated, must justify business necessity (i.e., the selection procedure is essential to the safe & efficient operation)
- Example: Hiring
    - 50% of male applicants vs 20% female applicants hired (0.2/0.5 = 0.4)
    - Is there a business justification for hiring men at a higher rate?

CASE STUDY: COLLEGE ADMISSION

- Who are the stakeholders?
- Types of harm?
- Legal & policy considerations?

REQUIREMENTS FOR FAIR ML SYSTEMS

1. Identify all environmental entities
2. State requirement (REQ) over the environment

REQUIREMENTS FOR FAIR ML SYSTEMS

1. Identify all environmental entities
2. State requirement (REQ) over the environment
3. Identify the interface between the environment & machine (ML)
    - What types of data will be sensed/measured by AI?
    - What types of actions will be performed by AI?

REQUIREMENTS FOR FAIR ML SYSTEMS

1. Identify all environmental entities
2. State requirement (REQ) over the environment
3. Identify the interface between the environment & machine (ML)
    - What types of data will be sensed/measured by AI?
    - What types of actions will be performed by AI?
4. Identify the environmental assumptions (ENV)
    - How do stakeholders interact with the system?
    - Adversarial? Misuse? Unfair (dis-)advantages?

CASE STUDY: COLLEGE ADMISSION

- Do certain groups of stakeholders have unfair (dis-)advantages that affect their behavior?
- What types of data should the system measure?

REQUIREMENTS FOR FAIR ML SYSTEMS

1. Identify all environmental entities
2. State requirement (REQ) over the environment
3. Identify the interface between the environment & machine (ML)
4. Identify the environmental assumptions (ENV)

REQUIREMENTS FOR FAIR ML SYSTEMS

1. Identify all environmental entities
2. State requirement (REQ) over the environment
3. Identify the interface between the environment & machine (ML)
4. Identify the environmental assumptions (ENV)
5. Develop software specifications (SPEC) that are sufficient to establish REQ
    - What type of fairness definition should we try to achieve?

REQUIREMENTS FOR FAIR ML SYSTEMS

1. Identify all environmental entities
2. State requirement (REQ) over the environment
3. Identify the interface between the environment & machine (ML)
4. Identify the environmental assumptions (ENV)
5. Develop software specifications (SPEC) that are sufficient to establish REQ
    - What type of fairness definition should we try to achieve?
6. Test whether ENV $\wedge$ SPEC $\models$ REQ
    - Continually monitor the fairness metrics and user reports

CASE STUDY: COLLEGE ADMISSION

- What type of fairness definition is appropriate?
  - Group fairness vs equalized odds?
- How do we monitor if the system is being fair?