

# FAIRNESS: BEYOND MODEL

Metric #1,284.

Okay, the True Positives divided by the False Positives, multiplied by the total number of Negative Predictions, plus the temperature of the room, multiplied by the negative exponential of the number of words in this sentence, should be the same for all sensitive groups.

What are we  
measuring again?

Fairness.

Right.



# Eunsuk Kang

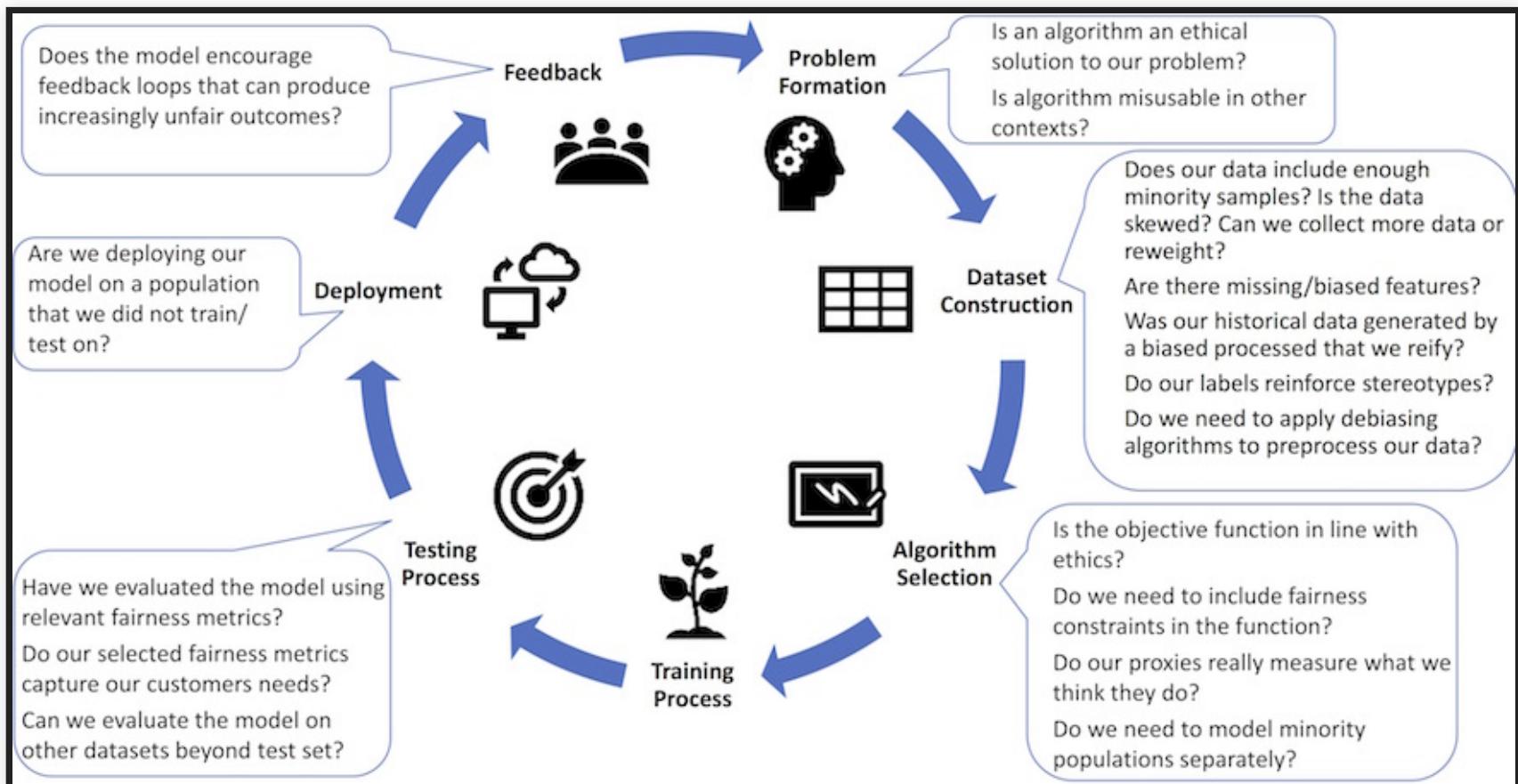
Required reading: Os Keyes, Jevan Hutson, Meredith Durbin. [A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry](#). CHI Extended Abstracts, 2019.

# LEARNING GOALS

- Understand the role of requirements engineering in selecting ML fairness criteria
- Understand the process of constructing datasets for fairness
- Consider the potential impact of feedback loops on AI-based systems and need for continuous monitoring
- Consider achieving fairness in AI-based systems as an activity throughout the entire development cycle

# BUILDING FAIR ML SYSTEMS

Recall: Fairness must be considered throughout the ML lifecycle!





# **FAIRNESS DEFINITIONS: REVIEW**

# REVIEW OF DEFINITIONS SO FAR:

*Recidivism scenario: Should a person be detained?*

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





# REVIEW OF DEFINITIONS SO FAR:

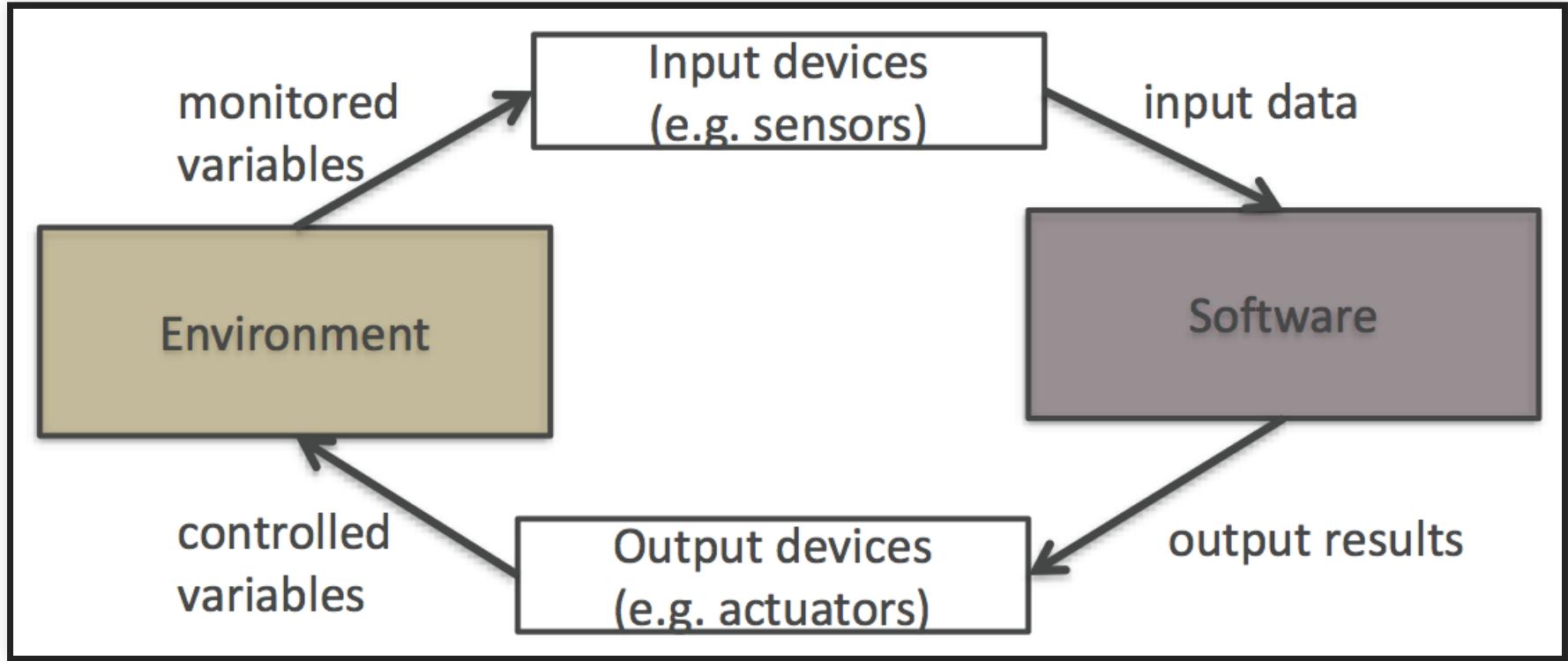
*Recidivism scenario: Should a defendant be detained?*

- 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

**Q. But how do we decide which of these definitions to use?**

# **REQUIREMENTS ENGINEERING FOR FAIRNESS**

# 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

# REQUIREMENTS FOR FAIR ML SYSTEMS

# REQUIREMENTS FOR FAIR ML SYSTEMS

- Understand the environment and its assumptions
  - Who are the stakeholders? Any disadvantaged groups?
  - What is the target population distribution like?

# REQUIREMENTS FOR FAIR ML SYSTEMS

- Understand the environment and its assumptions
  - Who are the stakeholders? Any disadvantaged groups?
  - What is the target population distribution like?
- Identify fairness requirements
  - What is the ultimate goal of the system being developed?
  - What types of harm can be caused as side effects?
  - Are we trying to achieve equality vs. equity?
  - Are there any legal requirements to consider?

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

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

# "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"

# "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

# "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)

# "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?

# RECALL: EQUALITY VS EQUITY

## Inequality

Unequal access to opportunities



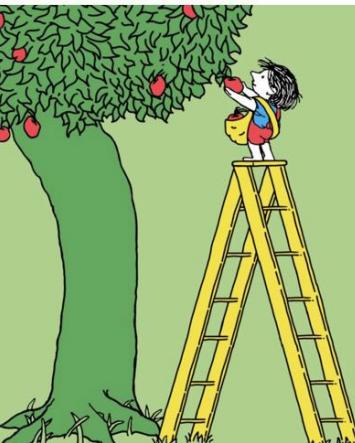
## Equality?

Evenly distributed tools and assistance



## Equity

Custom tools that identify and address inequality



## Justice

Fixing the system to offer equal access to both tools and opportunities



With apologies to Shel Silverstein from @lunchbreath

2019 Design In Tech Report | Addressing Imbalance

With apologies to Shel Silverstein from @lunchbreath

2019 Design In Tech Report | Addressing Imbalance

# REQUIREMENTS FOR FAIR ML SYSTEMS

- Understand the environment and its assumptions
  - Who are the stakeholders? Any disadvantaged groups?
  - What is the target population distribution like?
- Identify fairness requirements
  - What is the ultimate goal of the system being developed?
  - What types of harm can be caused as side effects?
  - Are we trying to achieve equality vs. equity?
  - Are there any legal requirements to consider?

# REQUIREMENTS FOR FAIR ML SYSTEMS

- Understand the environment and its assumptions
  - Who are the stakeholders? Any disadvantaged groups?
  - What is the target population distribution like?
- Identify fairness requirements
  - What is the ultimate goal of the system being developed?
  - What types of harm can be caused as side effects?
  - Are we trying to achieve equality vs. equity?
  - Are there any legal requirements to consider?
- Define the interface between the environment & machine (ML)
  - What data will be sensed/measured by AI? Potential biases?
  - What types of decisions will the system make? Punitive or assistive?

# TYPE OF DECISION & POSSIBLE HARM

# TYPE OF DECISION & POSSIBLE HARM

- If the decision is punitive in nature:
  - Harm is caused when a protected group is given an unwarranted penalty
  - e.g. decide whom to deny bail based on risk of recidivism
  - Heuristic: Use a fairness metric (separation) based on **false positive rate**

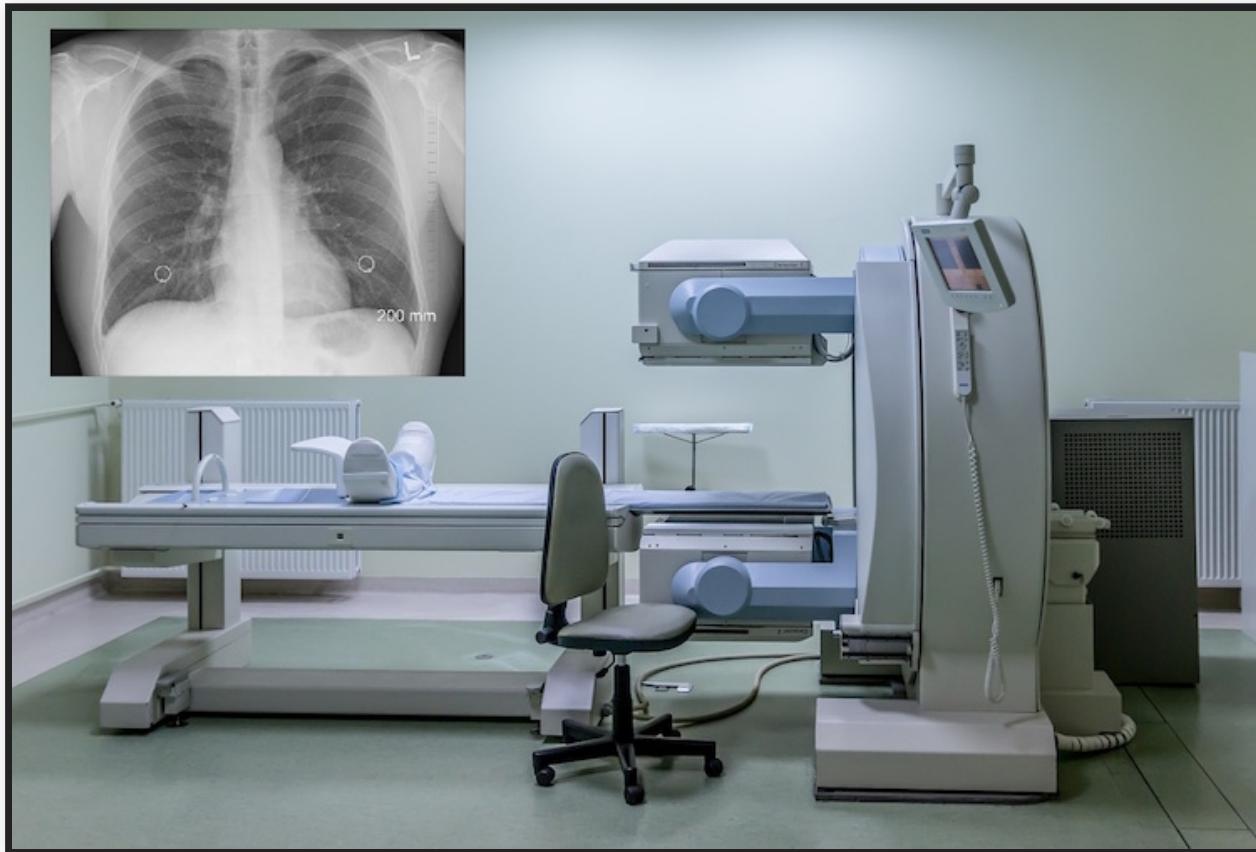
# TYPE OF DECISION & POSSIBLE HARM

- If the decision is punitive in nature:
  - Harm is caused when a protected group is given an unwarranted penalty
  - e.g. decide whom to deny bail based on risk of recidivism
  - Heuristic: Use a fairness metric (separation) based on **false positive rate**
- If the decision is assistive in nature:
  - Harm is caused when a group in need is incorrectly denied assistance
  - e.g., decide who should receive a loan or a food subsidy
  - Heuristic: Use a fairness metric based on **false negative rate**

# REQUIREMENTS FOR FAIR ML SYSTEMS

- Understand the environment and its assumptions
  - Who are the stakeholders? Any disadvantaged groups?
  - What is the target population distribution like?
- Identify fairness requirements
  - What is the ultimate goal of the system being developed?
  - What types of harm can be caused as side effects?
  - Are we trying to achieve equality vs. equity?
  - Are there any legal requirements to consider?
- Define the interface between the environment & machine (ML)
  - What data will be sensed/measured by AI? Potential biases?
  - What types of decisions will the system make? Punitive or assistive?
- Based on the environmental assumptions and requirements, select the appropriate fairness metric for the ML component

# WHICH FAIRNESS CRITERIA?



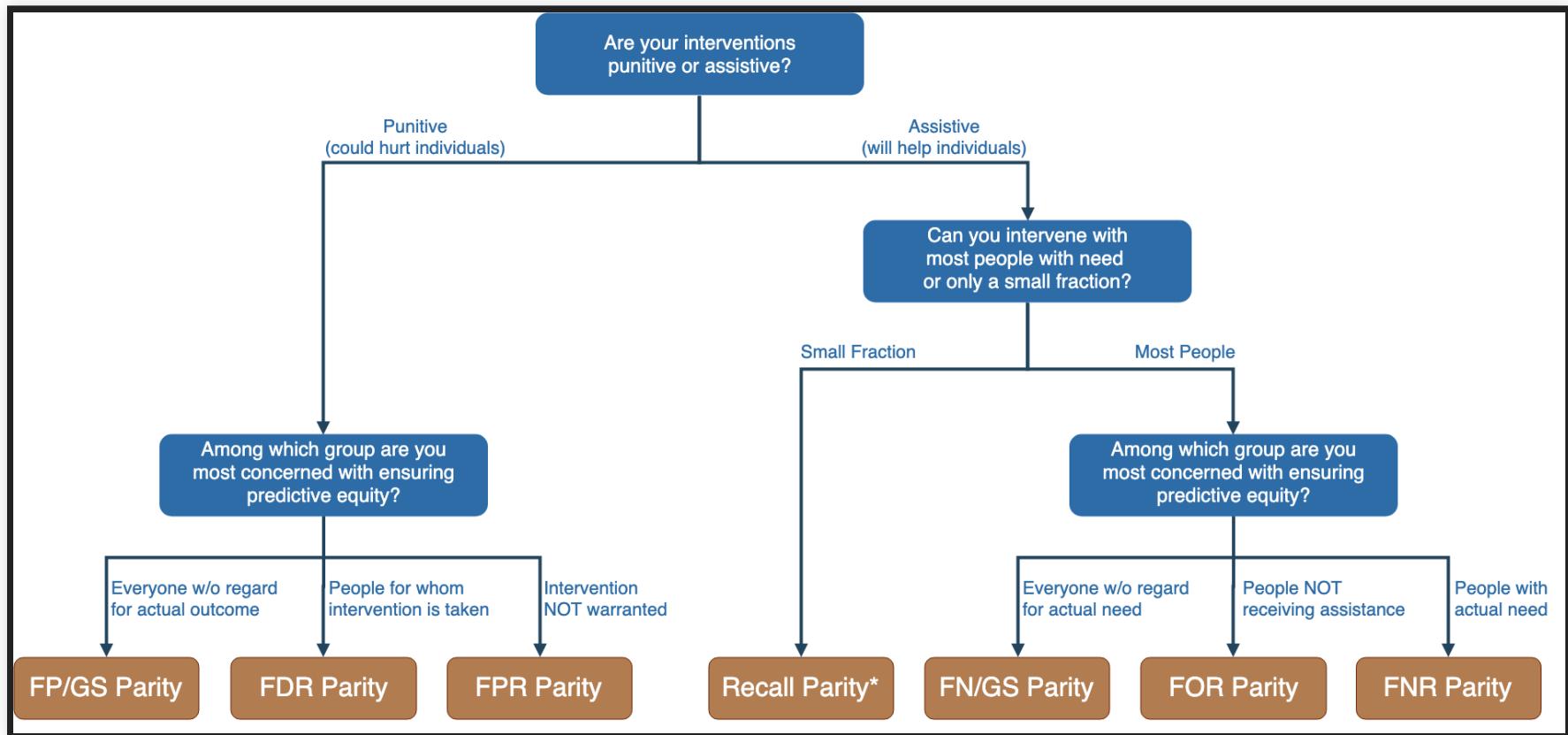
- Decision: Does the patient has a high risk of cancer?
- What kind of harm can be caused? Punitive or assistive?
- Criteria: Anti-classification, group fairness, or seperation w/ FPR or FNR?

# WHICH FAIRNESS CRITERIA?



- Decision: Should an applicant be accepted into the program?
- What kind of harm can be caused? Punitive or assistive?
- What do we want to achieve? Equality or equity?
- Criteria: Anti-classification, group fairness, or separation w/ FPR or FNR?

# FAIRNESS TREE



For details on other types of fairness metrics, see:  
<https://textbook.coleridgeinitiative.org/chap-bias.html>

# RECIDIVISM REVISITED



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

## Machine Bias

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

- COMPAS system, developed by Northpointe
  - Used by judges in sentencing decisions
  - In deployment throughout numerous states (PA, FL, NY, WI, CA, etc.,)

[ProPublica article](#)



# WHICH FAIRNESS DEFINITION?

Table 11.1: COMPAS Fairness Metrics

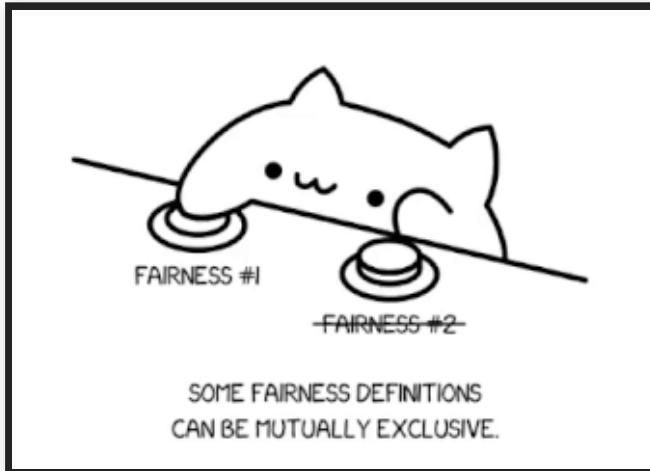
Metric	Caucasian	African American
False Positive Rate ( <i>FPR</i> )	23%	45%
False Negative Rate ( <i>FNR</i> )	48%	28%
False Discovery Rate ( <i>FDR</i> )	41%	37%

- ProPublica investigation: COMPAS violates separation w/ FPR & FNR
- Northpointe response: COMPAS is fair because it has similar FDRs across both races
  - $FDR = FP / (FP + TP) = 1 - \text{Precision}$
  - $FPR = FP / (FP + TN)$
- Q. So is COMPAS both fair & unfair at the same time? Which definition is the "right" one?

Figure from Big Data and Social Science, Ch. 11



# FAIRNESS DEFINITIONS: PITFALLS



- "Impossibility Theorem": Can't satisfy multiple fairness criteria at once
- Easy to pick some definition & claim that the model is fair
  - But does a "fair" model really help reduce harm in the long term?
- Instead of trying to "fix" bias through a model, can we understand & address the root causes of bias in the first place?
  - Many of these are deeply rooted in our society, and can't be solved purely through technical means!



# **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

## Data Source

- **Functional:** biases due to platform affordances and algorithms
- **Normative:** biases due to community norms
- **External:** biases due to phenomena outside social platforms
- **Non-individuals:** e.g., organizations, automated agents

## Data Collection

- **Acquisition:** biases due to, e.g., API limits
- **Querying:** biases due to, e.g., query formulation
- **Filtering:** biases due to removal of data "deemed" irrelevant

## Data Processing

- **Cleaning:** biases due to, e.g., default values
- **Enrichment:** biases from manual or automated annotations
- **Aggregation:** e.g., grouping, organizing, or structuring data

## Data Analysis

- **Qualitative Analyses:** lack generalizability, interpret. biases
- **Descriptive Statistics:** confounding bias, obfuscated measurements
- **Prediction & Inferences:** data representation, perform. variations
- **Observational studies:** peer effects, select. bias, ignorability

## Evaluation

- **Metrics:** e.g., reliability, lack of domain insights
- **Interpretation:** e.g., contextual validity, generalizability
- **Disclaimers:** e.g., lack of negative results and reproducibility

- 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

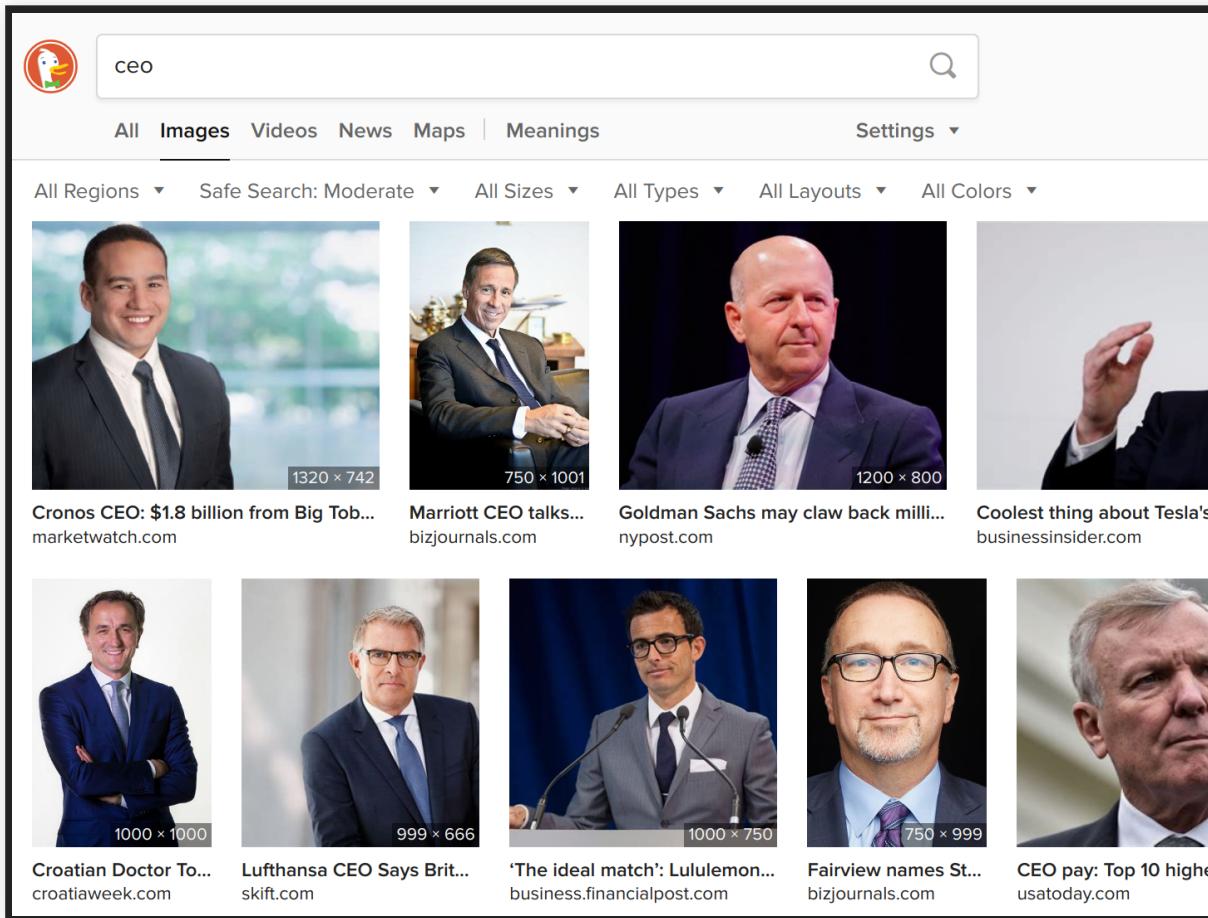
Data set	Gender		Skin Color/Type	
	Female	Male	Darker	Lighter
LFW [15]	22.5%	77.4%	18.8%	81.2%
IJB-C [28]	37.4%	62.7%	18.0%	82.0%
Pubfig [35]	50.8%	49.2%	18.0%	82.0%
CelebA [9]	58.1%	42.0%	14.2%	85.8%
UTKface [32]	47.8%	52.2%	35.6%	64.4%
AgeDB [33]	40.6%	59.5%	5.4%	94.6%
PPB [36]	44.6%	55.4%	46.4%	53.6%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



# HISTORICAL BIAS



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

# BEHAVIORAL BIAS

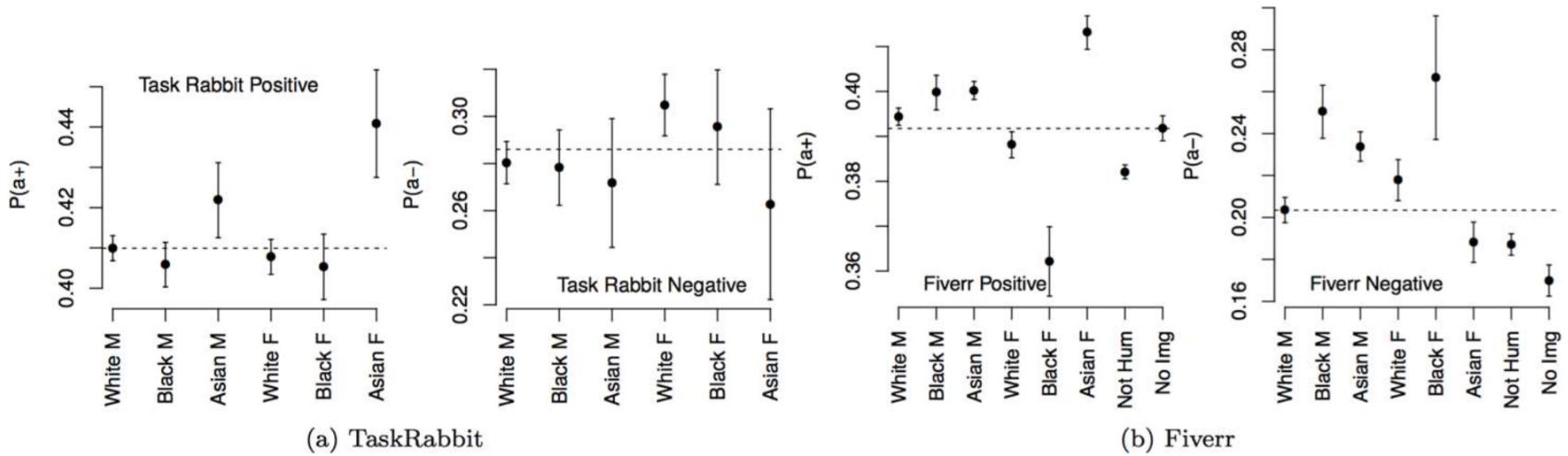


Figure 2: Fitted  $P(a_+)$  and  $P(a_-)$  depending on combinations of gender and race of the reviewed worker. Points show expected values and bars standard errors. In Fiverr, Black workers are less likely to be described with adjectives for positive words, and Black Male workers are more likely to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

# BREAEKOUT: COLLEGE ADMISSION



- Features: GPA, GRE/SAT, gender, race, undergrad institute, alumni connections, household income, hometown, etc.,
- Type into #lecture in Slack:
  - What are different sub-populations?
  - What are potential sources of data bias?
  - What are ways to mitigate this bias?

# 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)?"



# 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 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**

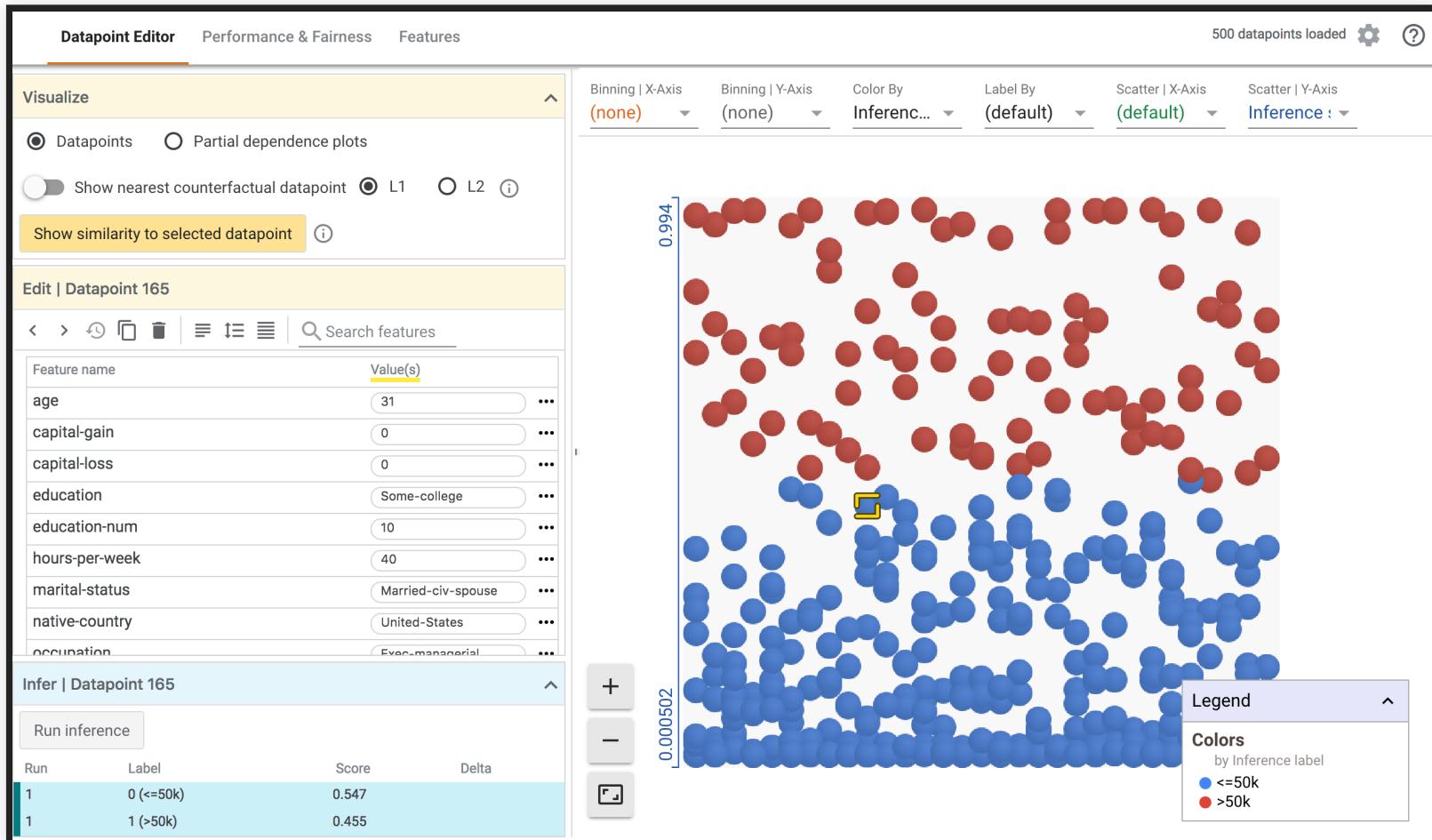
- Synthetic test data covers only a small set of very specific comments. While these are designed to be representative of common use cases and concerns, it is not comprehensive.

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.



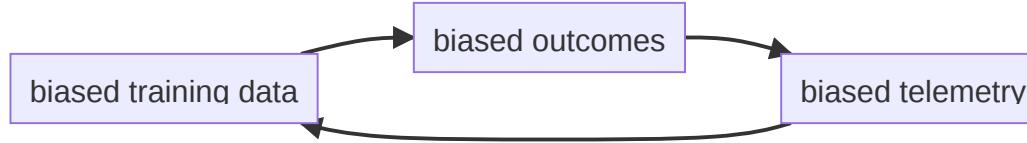
# MODEL EXPLORATION



Google What-If Tool

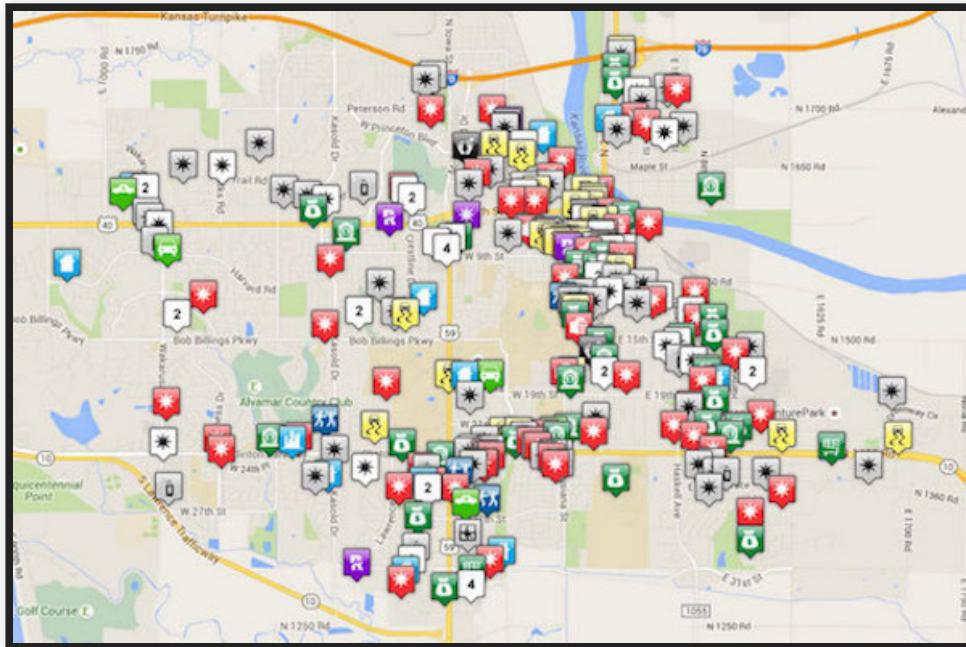
# FEEDBACK LOOPS

# FEEDBACK LOOPS



*"Big Data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide. " -- Cathy O'Neil in [Weapons of Math Destruction](#)*

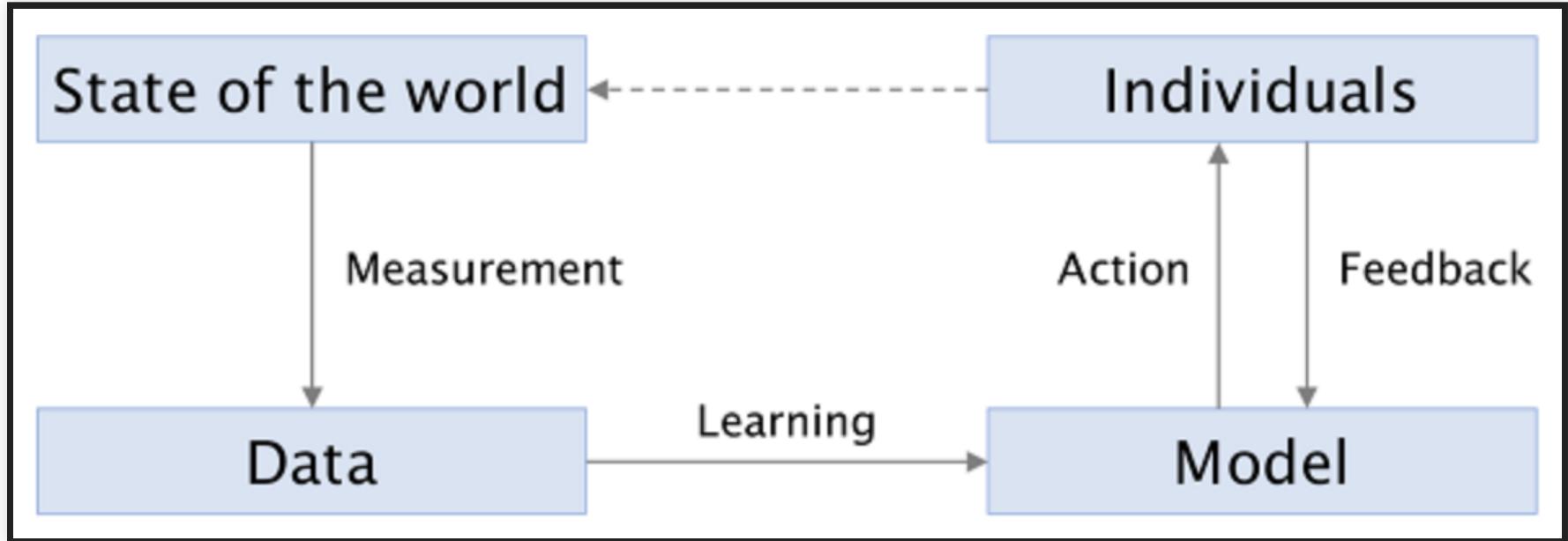
# EXAMPLE: PREDICTIVE POLICING



- Model: Use historical data to predict crime rates by neighborhoods
- Increased patrol => more arrested made in neighborhood X
- New crime data fed back to the model
- Repeat...

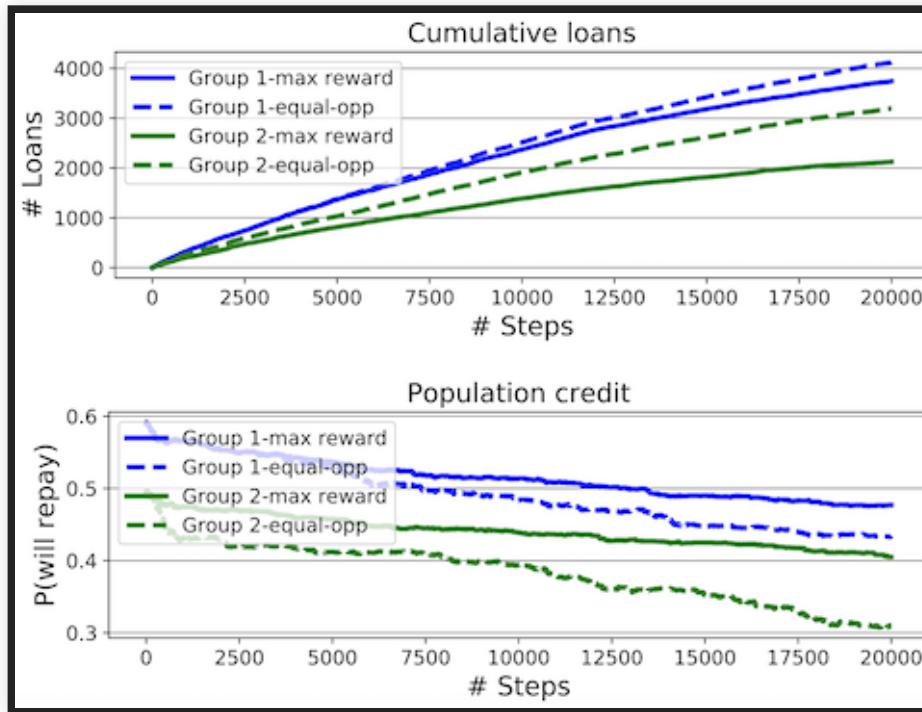
Q. Other examples?

# LONG-TERM IMPACT OF ML



- ML systems make multiple decisions over time, influence the behaviors of populations in the real world
- But most models are built & optimized assuming that the world is static!
- Difficult to estimate the impact of ML over time
  - Need to reason about the system dynamics (world vs machine)
  - e.g., what's the effect of a loan lending policy on a population?

# LONG-TERM IMPACT & FAIRNESS



- Deploying an ML model with a fairness criterion does NOT guarantee improvement in equality/equity over time
- Even if a model appears to promote fairness in short term, it may result harm over a long-term period

Fairness is not static: deeper understanding of long term fairness via simulation studies, in FAT\* 2020.



# **MONITORING AND AUDITING**

# **MONITORING & AUDITING**

# MONITORING & AUDITING

- Continuously monitor for:
  - Mismatch between training data, test data, and instances that you encounter in deployment
  - Fairness metrics: Is the system yielding fair results over time?
  - Population shifts: May suggest needs to adjust fairness metric/thresholds
  - User reports & complaints: Log and audit system decisions perceived to be unfair by users

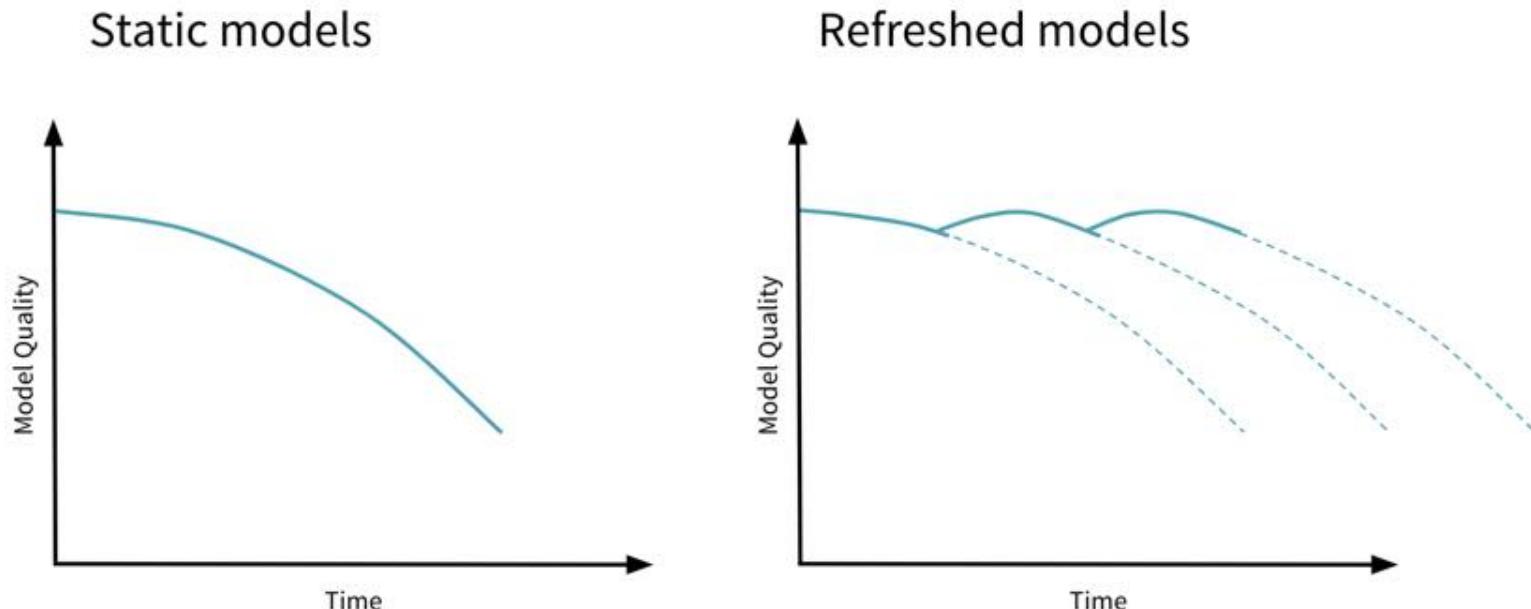
# MONITORING & AUDITING

- Continuously monitor for:
  - Mismatch between training data, test data, and instances that you encounter in deployment
  - Fairness metrics: Is the system yielding fair results over time?
  - Population shifts: May suggest needs to adjust fairness metric/thresholds
  - User reports & complaints: Log and audit system decisions perceived to be unfair by users
- Deploy escalation plans: How do you respond when harm occurs due to system?
  - Shutdown system? Temporary replacement?
  - Maintain communication lines to stakeholders

# MONITORING & AUDITING

- Continuously monitor for:
  - Mismatch between training data, test data, and instances that you encounter in deployment
  - Fairness metrics: Is the system yielding fair results over time?
  - Population shifts: May suggest needs to adjust fairness metric/thresholds
  - User reports & complaints: Log and audit system decisions perceived to be unfair by users
- Deploy escalation plans: How do you respond when harm occurs due to system?
  - Shutdown system? Temporary replacement?
  - Maintain communication lines to stakeholders
- Invite diverse stakeholders to audit system for biases

# MONITORING & AUDITING



- Continuously monitor the fairness metric (e.g., error rates for different groups)
- Re-train model with recent data or adjust classification thresholds if needed
- Recall: Data drifts in the Data Quality lecture

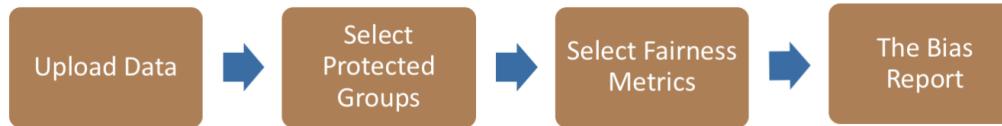
# MONITORING TOOLS: EXAMPLE

Aequitas  
Bias & Fairness Audit

[Home](#) [Code](#) [About](#)

## Bias and Fairness Audit Toolkit

The Bias Report is powered by [Aequitas](#), an open-source bias audit toolkit for machine learning developers, analysts, and policymakers to audit machine learning models for discrimination and bias, and make informed and equitable decisions around developing and deploying predictive risk-assessment tools.



<http://aequitas.dssg.io/>

# MONITORING TOOLS: EXAMPLE

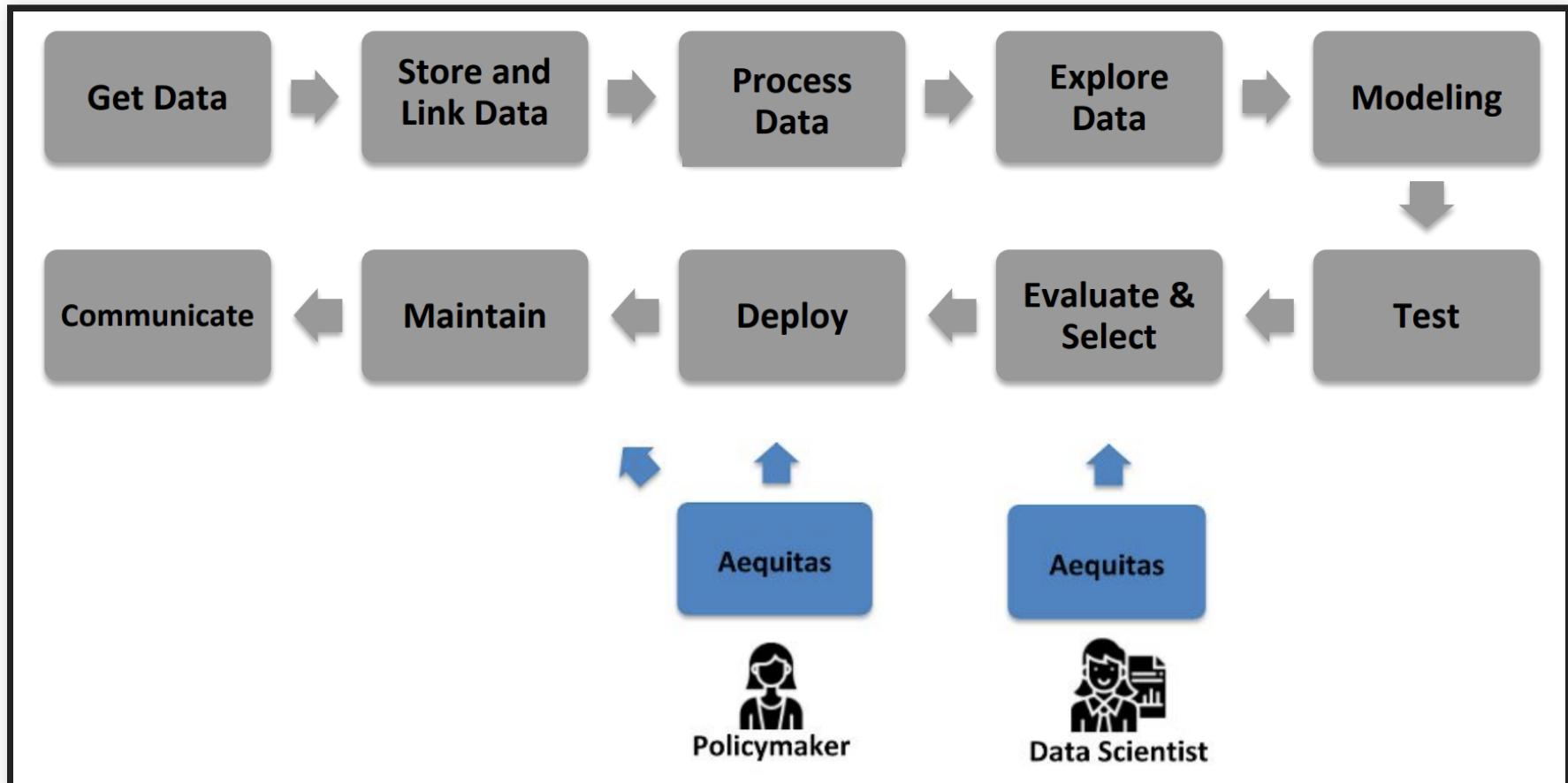
## Audit Results: Bias Metrics Values

race

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

- Continuously make fairness measurements to detect potential shifts in data, population behavior, etc.,

# MONITORING TOOLS: EXAMPLE



- Involve policy makers in the monitoring & auditing process

# BUILDING FAIR ML-BASED SYSTEMS: BEST PRACTICES

# START EARLY!

- Think about system goals and relevant fairness concerns
- Analyze risks & harms to affected groups
- 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

# 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](#))

# FAIRNESS CHECKLIST

## Envision

Consider doing the following items in moments like:

- Envisioning meetings
- Pre-mortem screenings
- Product greenlighting meetings

### 1.1 Envision system and scrutinize system vision

#### 1.1.a Envision system and its role in society, considering:

- System purpose, including key objectives and intended uses or applications
  - Consider whether the system should exist and, if so, whether the system should use AI
- Sensitive, premature, dual, or adversarial uses or applications
  - Consider whether the system will impact human rights
  - Consider whether these uses or applications should be prohibited
- Expected deployment contexts (e.g., geographic regions, time periods)
- Expected stakeholders (e.g., people who will make decisions about system adoption, people who will use the system, people who will be directly or indirectly affected by the system, society), including demographic groups (e.g., by race, gender, age, disability status, skin tone, and their intersections)
- Expected benefits for each stakeholder group, including demographic groups
- Relevant regulations, standards, guidelines, policies, etc.

#### 1.1.b Scrutinize resulting system vision for potential fairness-related harms to stakeholder groups, considering:

- Types of harm (e.g., allocation, quality of service, stereotyping, denigration, over- or underrepresentation)



# SUMMARY

- Requirements engineering for fair ML systems
  - Stakeholders, sub-populations & unfair (dis-)advantages
  - Types of harms
  - Legal requirements
- Dataset construction for fairness
- Consideration for the impact of feedback loops
- Continuous monitoring & auditing for fairness