FAIRNESS: BEYOND MODEL

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

- Consider achieving fairness in AI-based systems as an activity throughout the entire development cycle
- 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

FAIRNESS DEFINITIONS: REVIEW

REVIEW OF DEFINITIONS SO FAR:

Recidivism scenario: Should a person be detained?

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

RECIDIVISM REVISITED

- 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?

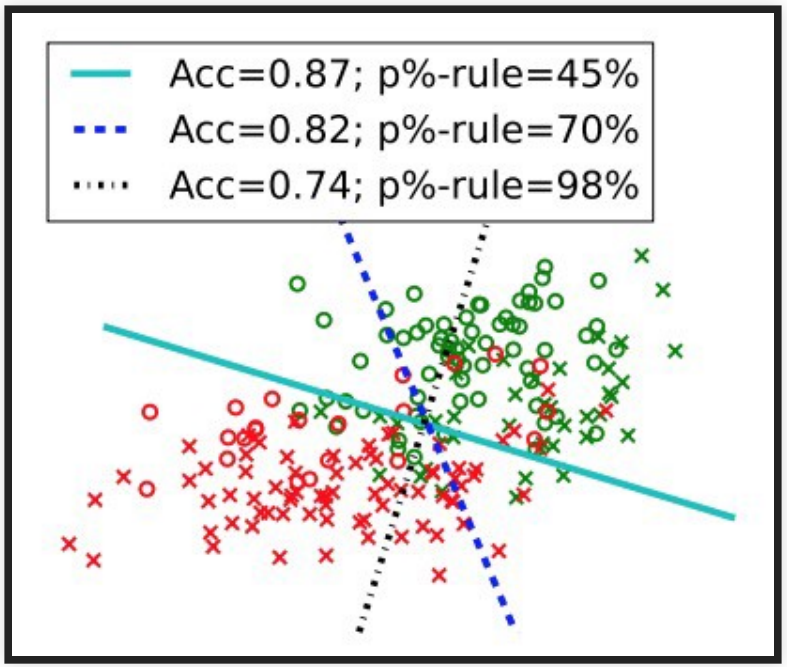
Table 11.1: COMPAS Fairness Metrics

Metric	Caucasian	African American
False Positive Rate (FPR)	23%	45%
False Negative Rate (FNR)	48%	28%
False Discovery Rate (FDR)	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 - Precision
    - FPR = FP / (FP + TN)
- **Q. So is COMPAS both fair & unfair at the same time? Which definition is the "right" one?**

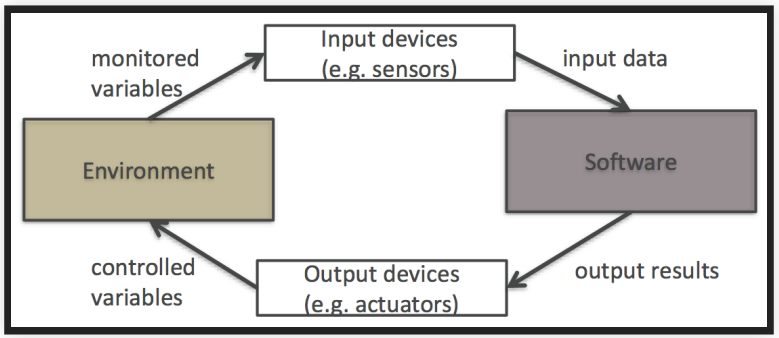
Figure from Big Data and Social Science, Ch. 11

FAIRNESS DEFINITIONS: PITFALLS



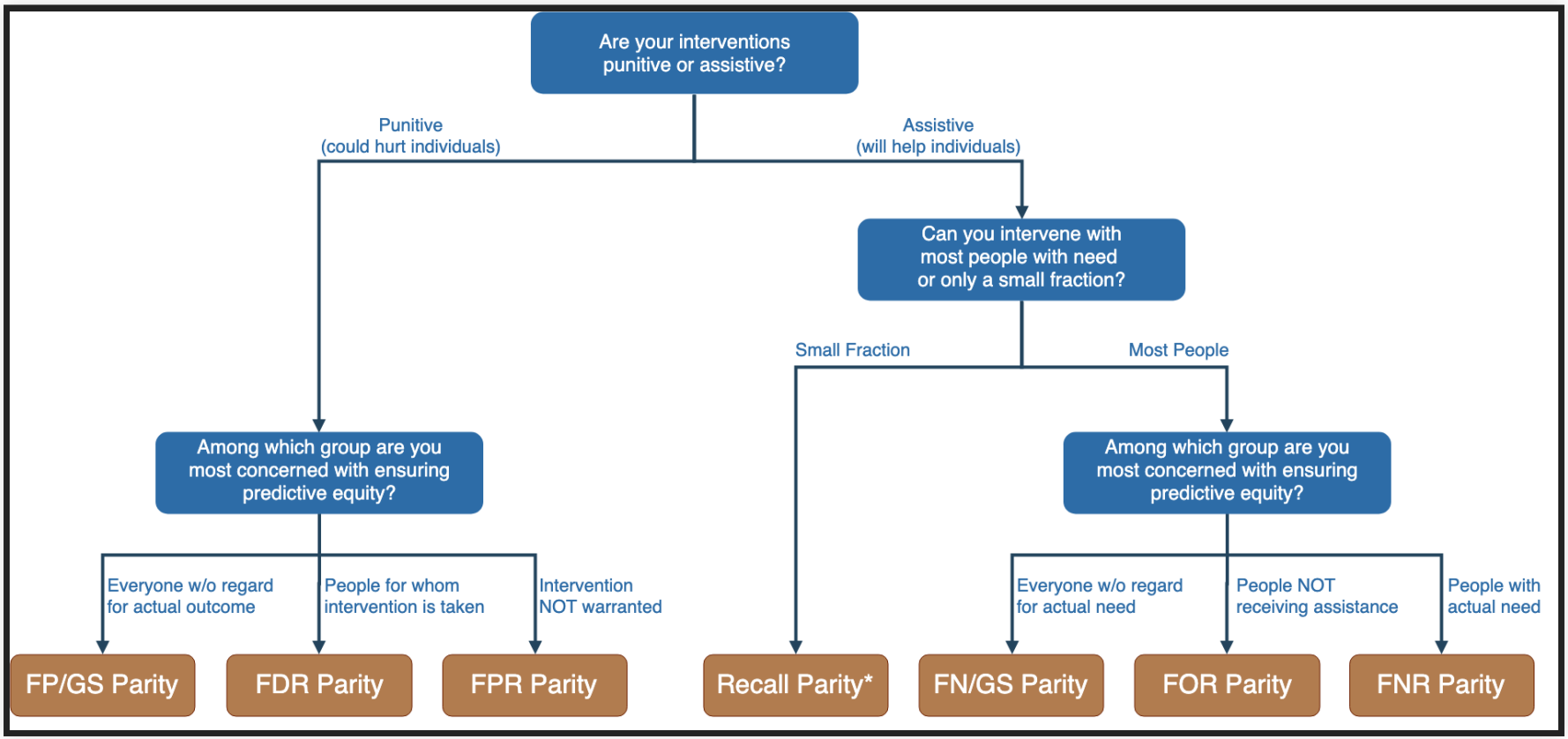
- Easy to pick some definition & claim that the model is fair
  - But is the **overall system** actually fair?
- In general, impossible to satisfy multiple fairness definitions at once
  - Fairness is a **context-dependent** notion
  - Select the criteria that minimize harm for the given context
  - Also consider trade-offs against accuracy

REQUIREMENTS & FAIRNESS



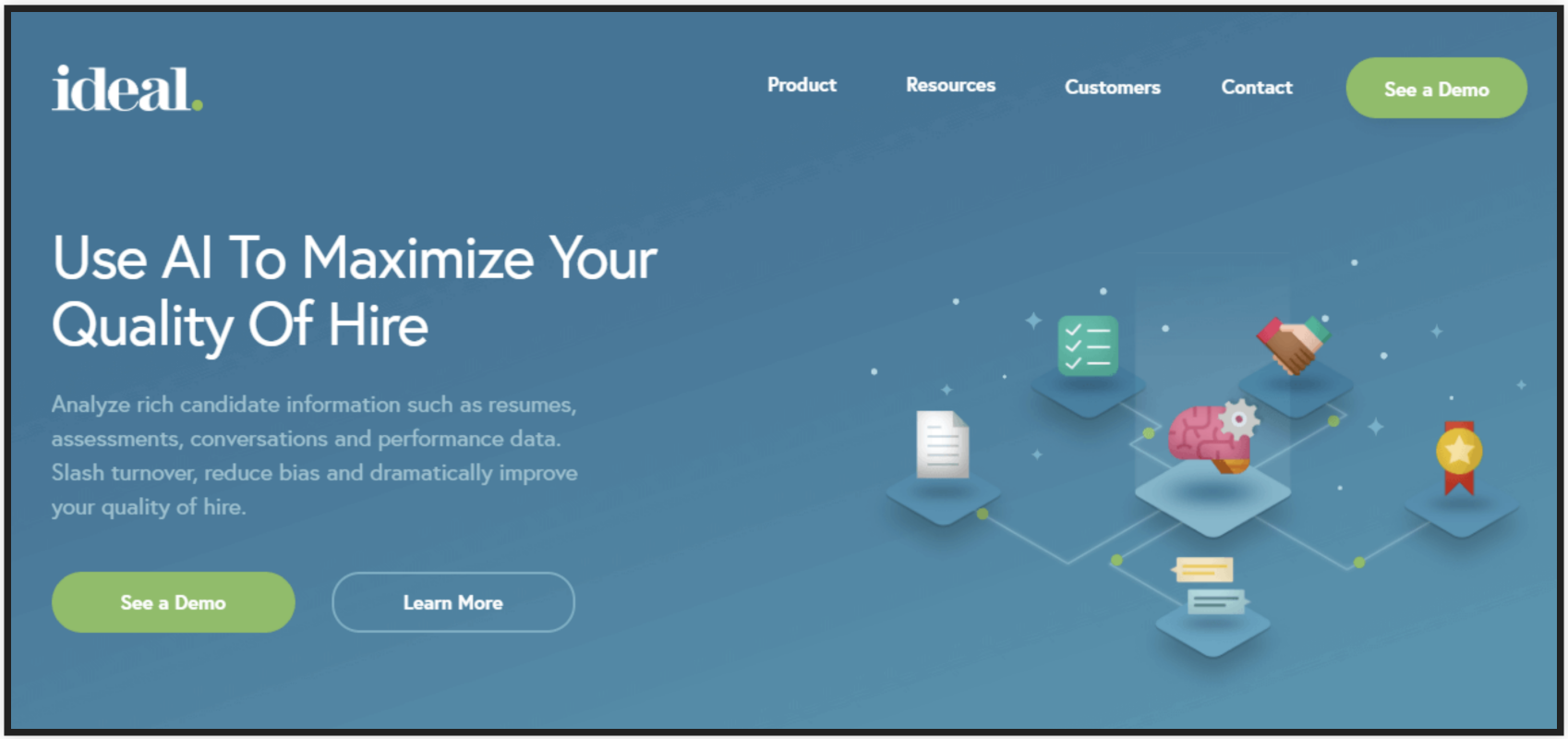
- Fairness is a **context-dependent** notion
- **Again, think about requirements!**
  - Who are the stakeholders of the system?
    - Which of these groups could be harmed?
  - What potential harms can be caused by biased decisions?
    - e.g., unfair punishments, denial to resources
  - Are there any legal constraints or policy goals?
    - e.g., 80% rule, affirmative actions
  - How are these decisions related to the ML model? Errors?
    - e.g., false positives, false negatives
  - Which fairness metric minimizes the harm?

RECALL: FAIRNESS TREE



Full tree

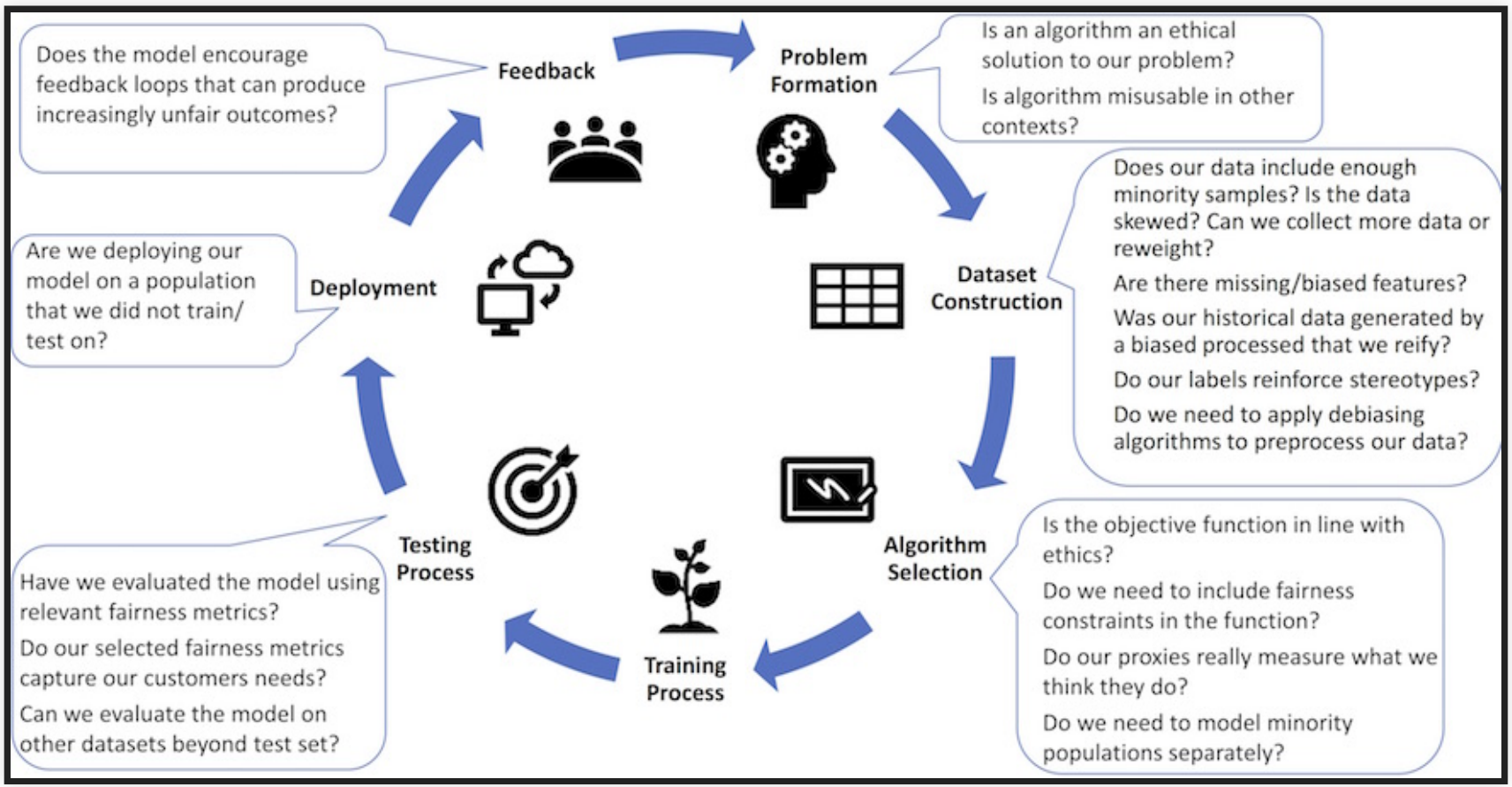
EXAMPLE: AUTOMATED HIRING



- Who are the groups possibly harm by biased decisions?
- What kind of harm can be caused?
- Which fairness metric to use?
    - Independence, separation w/ FPR vs. FNR?

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.

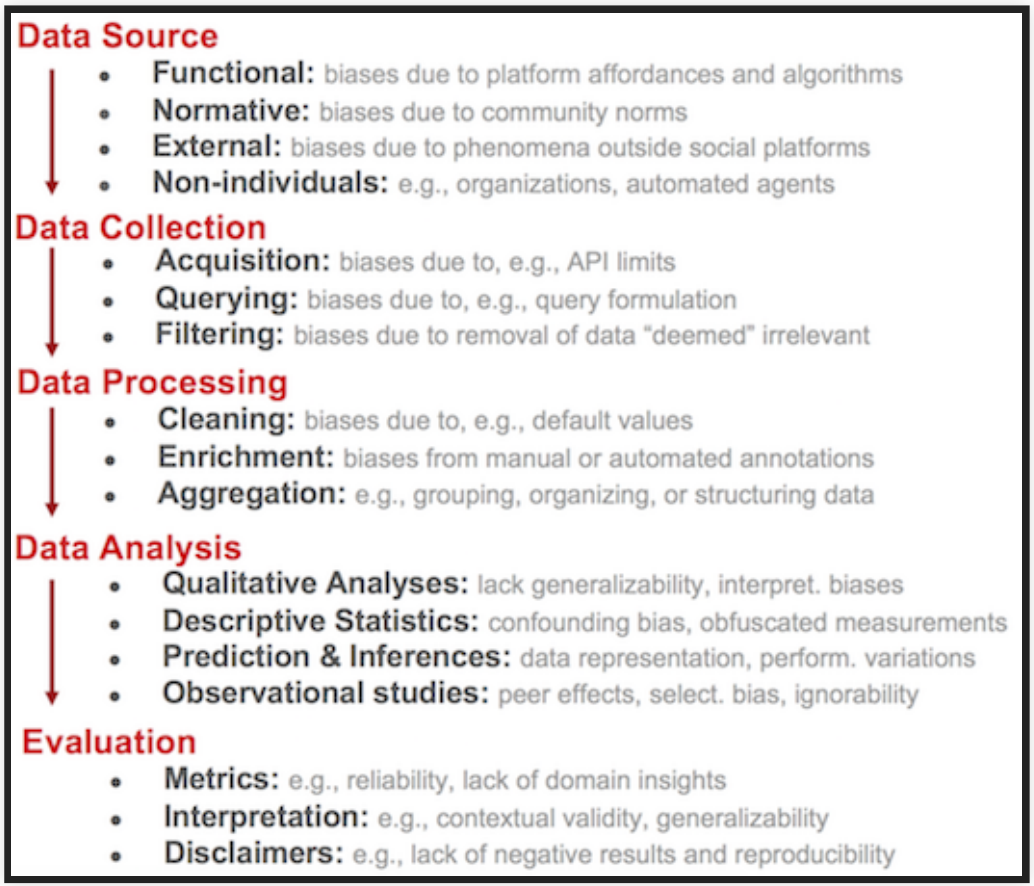
DATASET CONSTRUCTION FOR FAIRNESS

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).

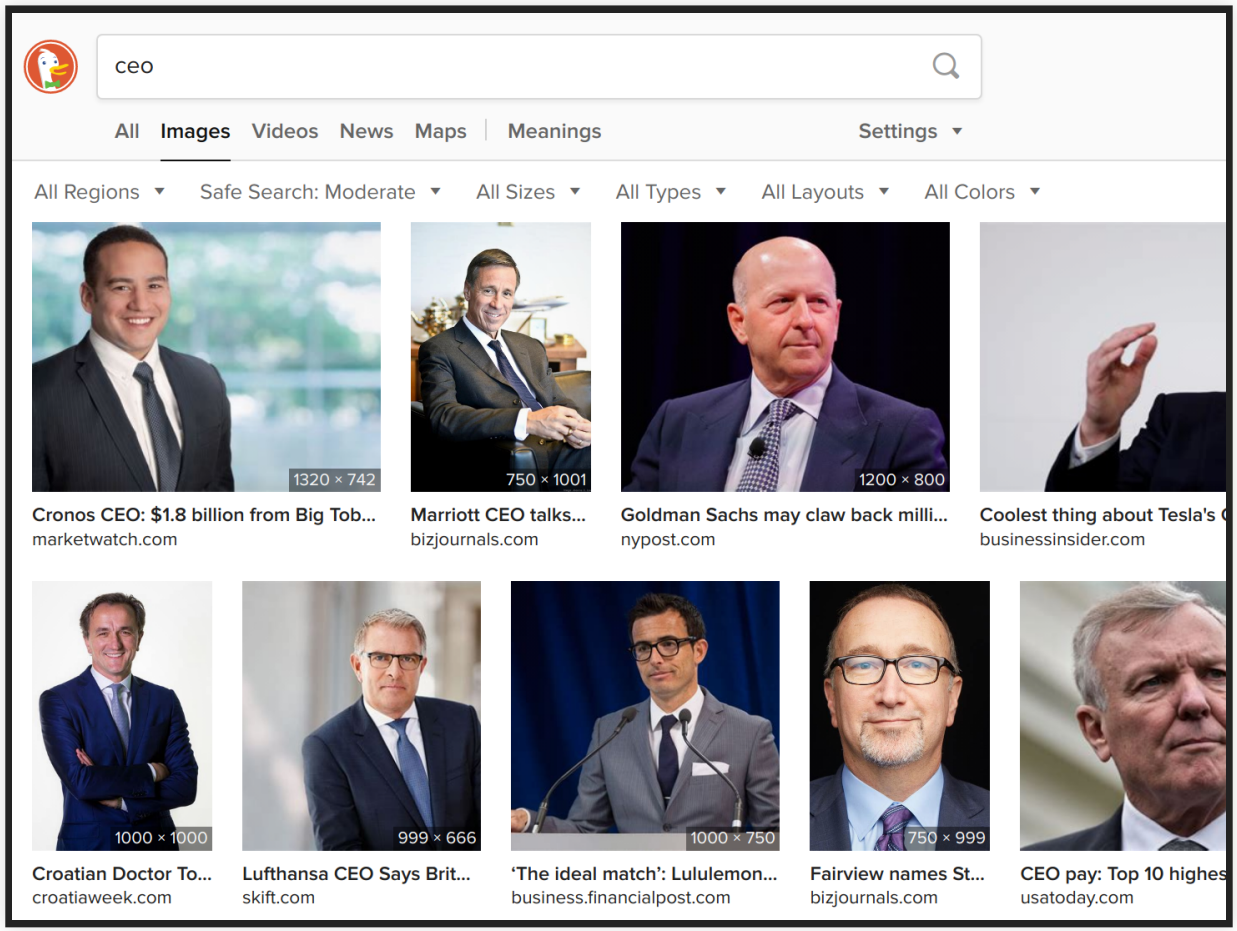
POPULATION BIAS

	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.

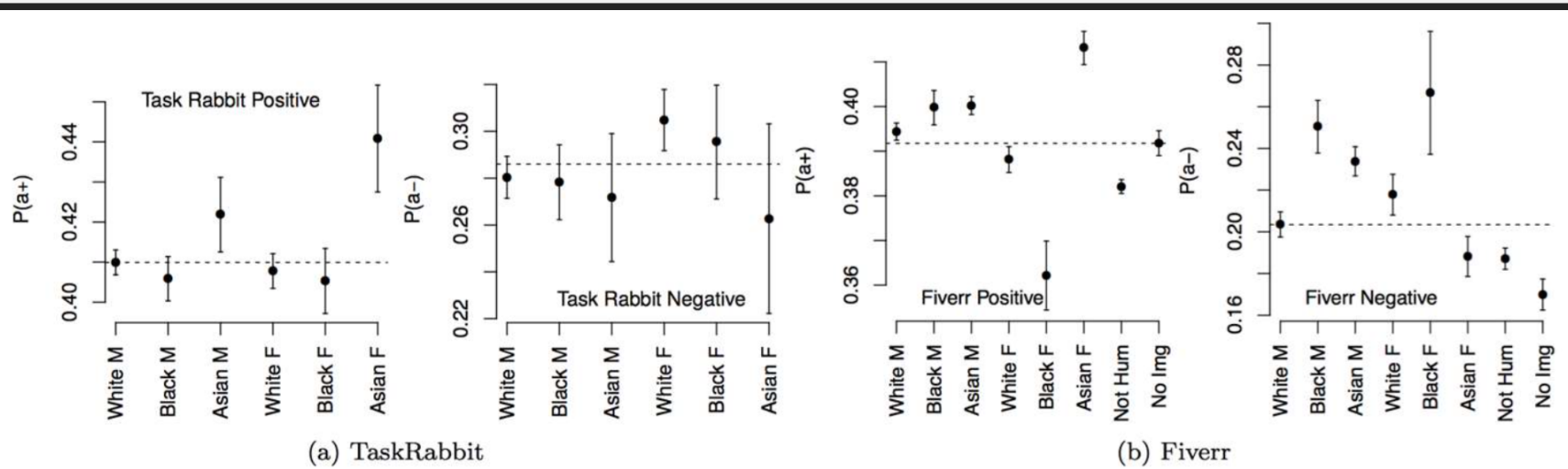
- Differences in demographics between a dataset vs a target population
- May result in degraded services for certain groups (e.g., poor image recognition for females & darker skin types)

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).

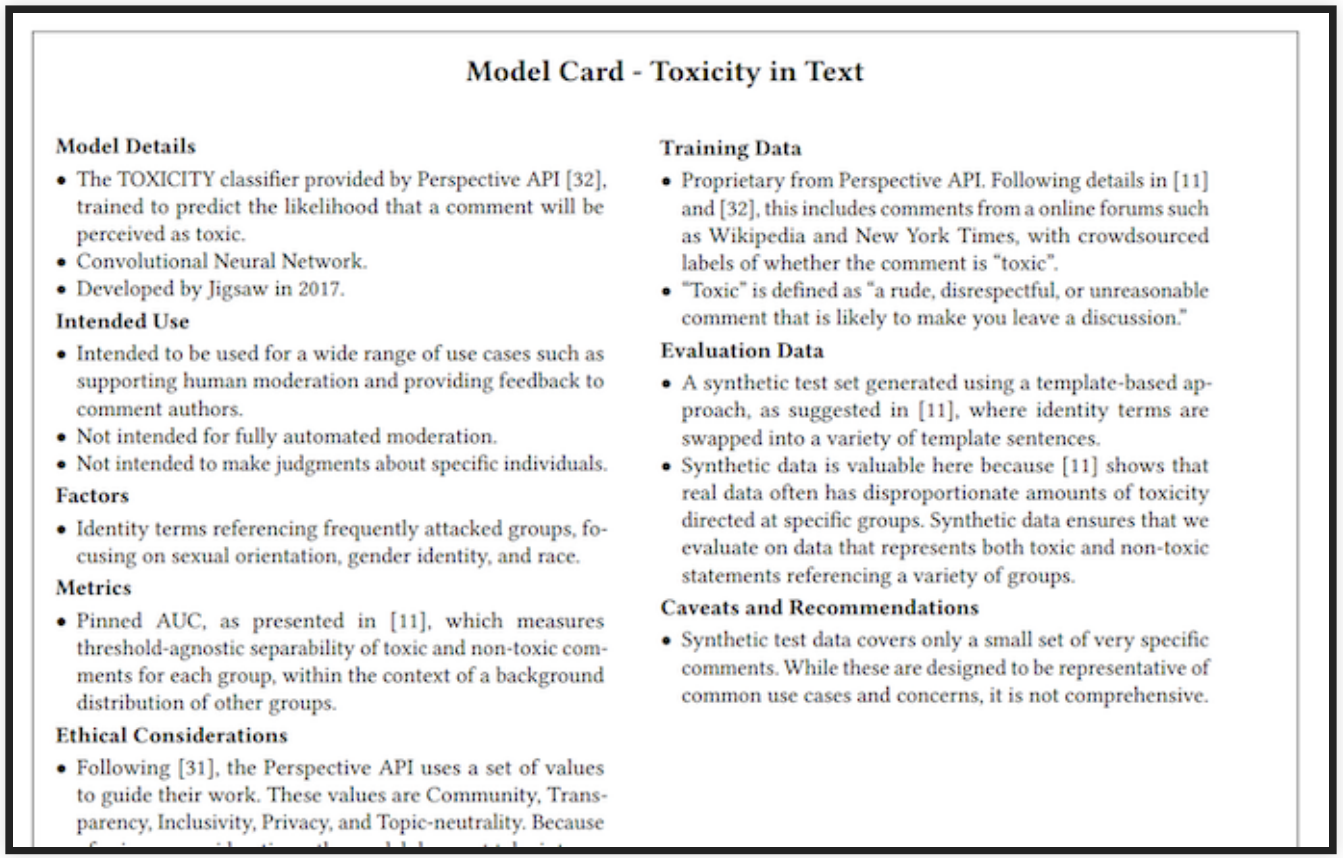
DATA SHEETS

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.

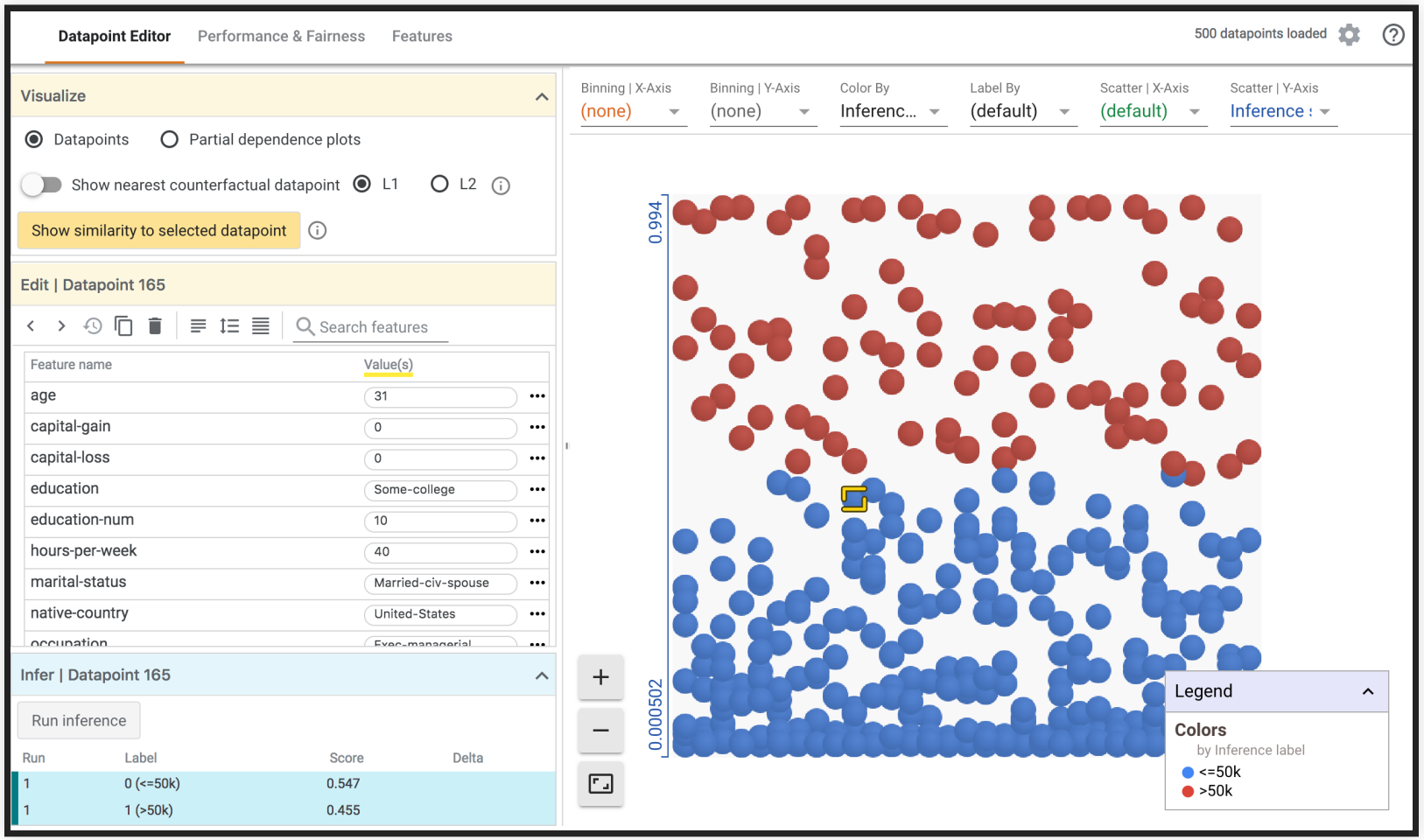
Caveats and Recommendations

- 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



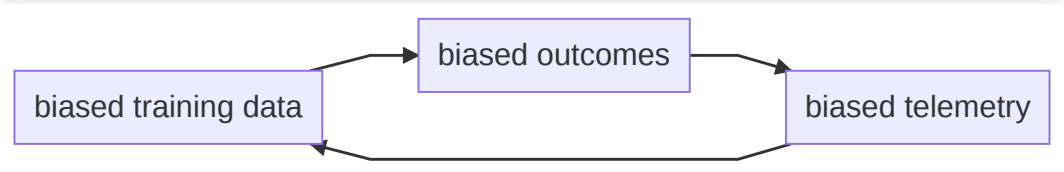
Google What-If Tool

BREAKOUT: DATA COLLECTION FOR FAIRNESS

- For each system, discuss:
  - What are possible types of bias in the data?
    - Population bias? Under- or over-representation?
  - How would you modify the dataset reduce bias?
    - Collect more data? Remove? Augment?

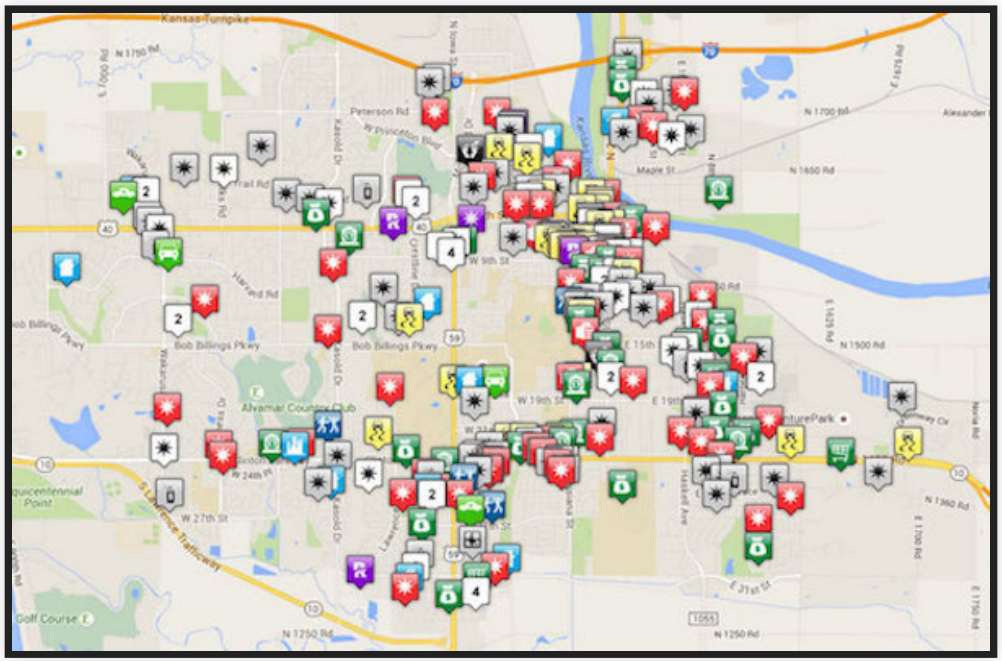
MONITORING AND AUDITING

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

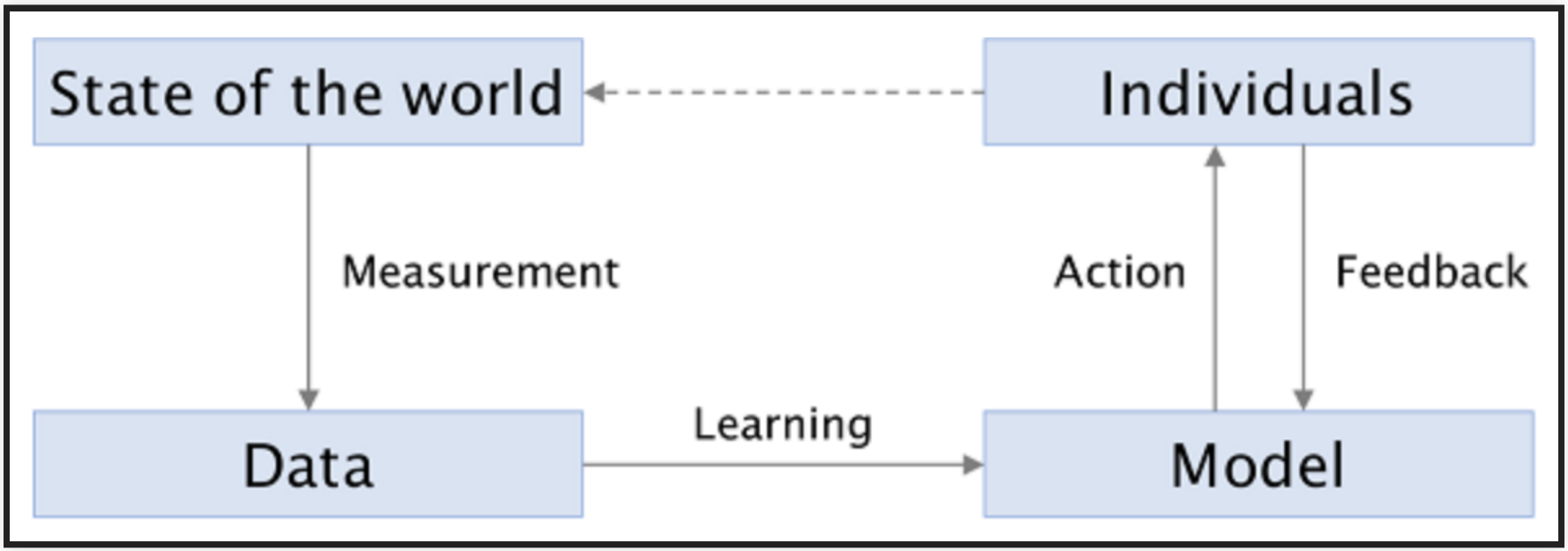
EXAMPLE: PREDICTIVE POLICING



- 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...

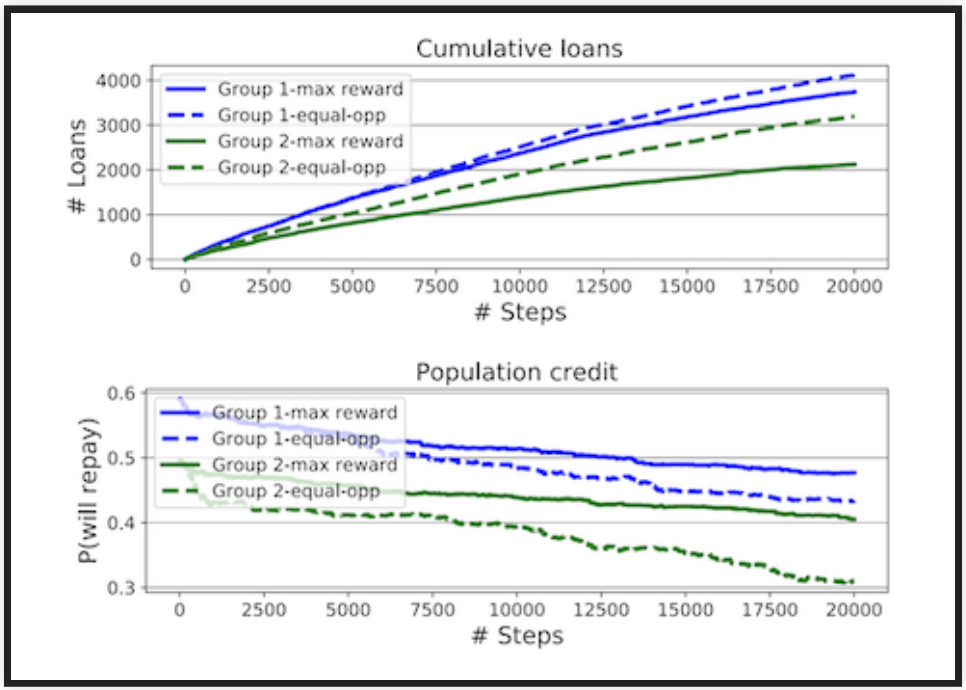
Q. Other examples?

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



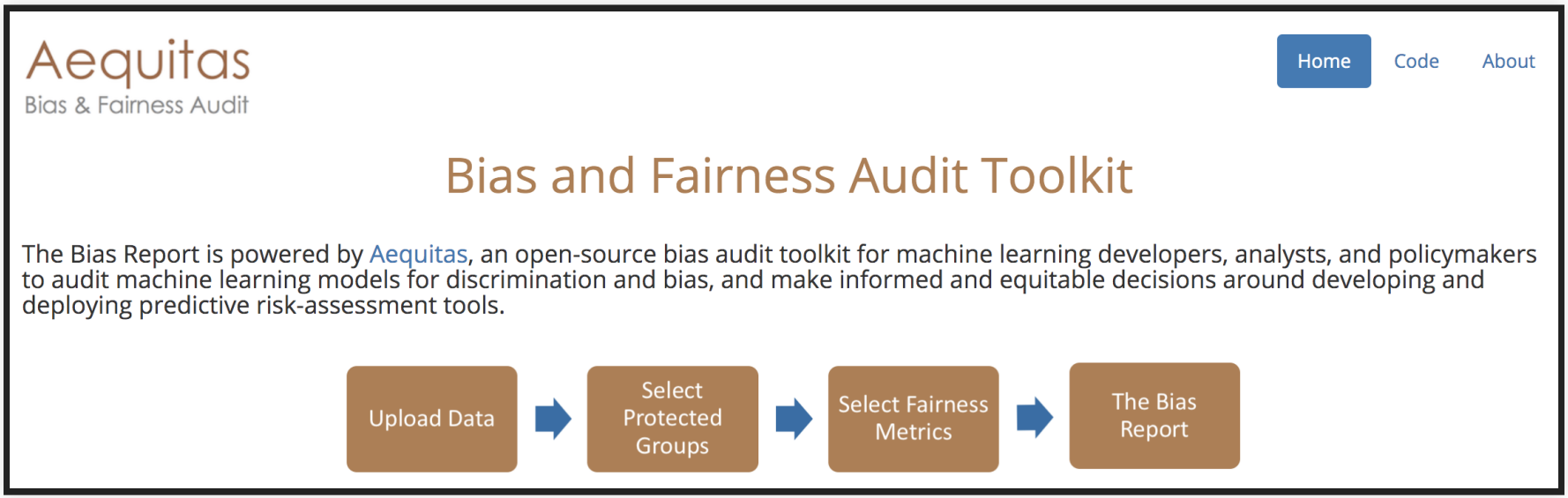
- Deploying an ML model with a fairness criterion does NOT guarantee improvement in equality over time
- Even if a model appears to promote fairness in short term, it may result harm over a long-term period

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

MONITORING & AUDITING

- Continuously monitor for:
  - Match 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 TOOLS: EXAMPLE



http://aequitas.dssg.io/

MONITORING TOOLS: EXAMPLE

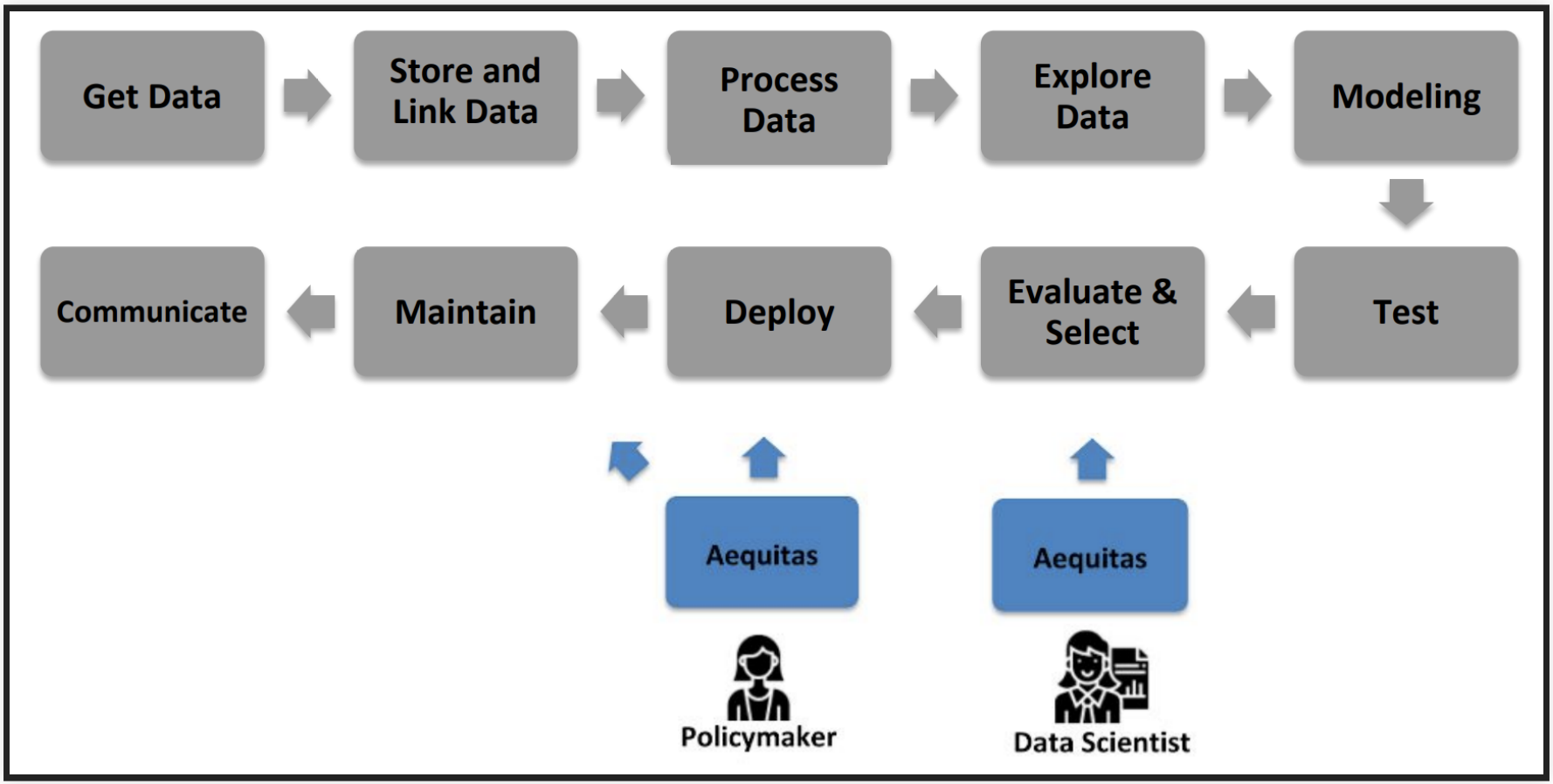
Audit Results: Bias Metrics Values

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

FAIRNESS: FINAL THOUGHTS

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

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)

Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI, Madaio et al (2020).

SUMMARY

- Achieving fairness as an activity throughout the entire development cycle
- Requirements engineering for fair ML systrems
    - Stakeholders, sub-populations & unfair (dis-)advantages
    - Types of harms
    - Legal requirements
- Dataset construction for fairness
- Consideration for the impact of feedback loops
- Continous montoring & auditing for fairness