FAIRNESS: BEYOND MODEL

Eunsuk Kang

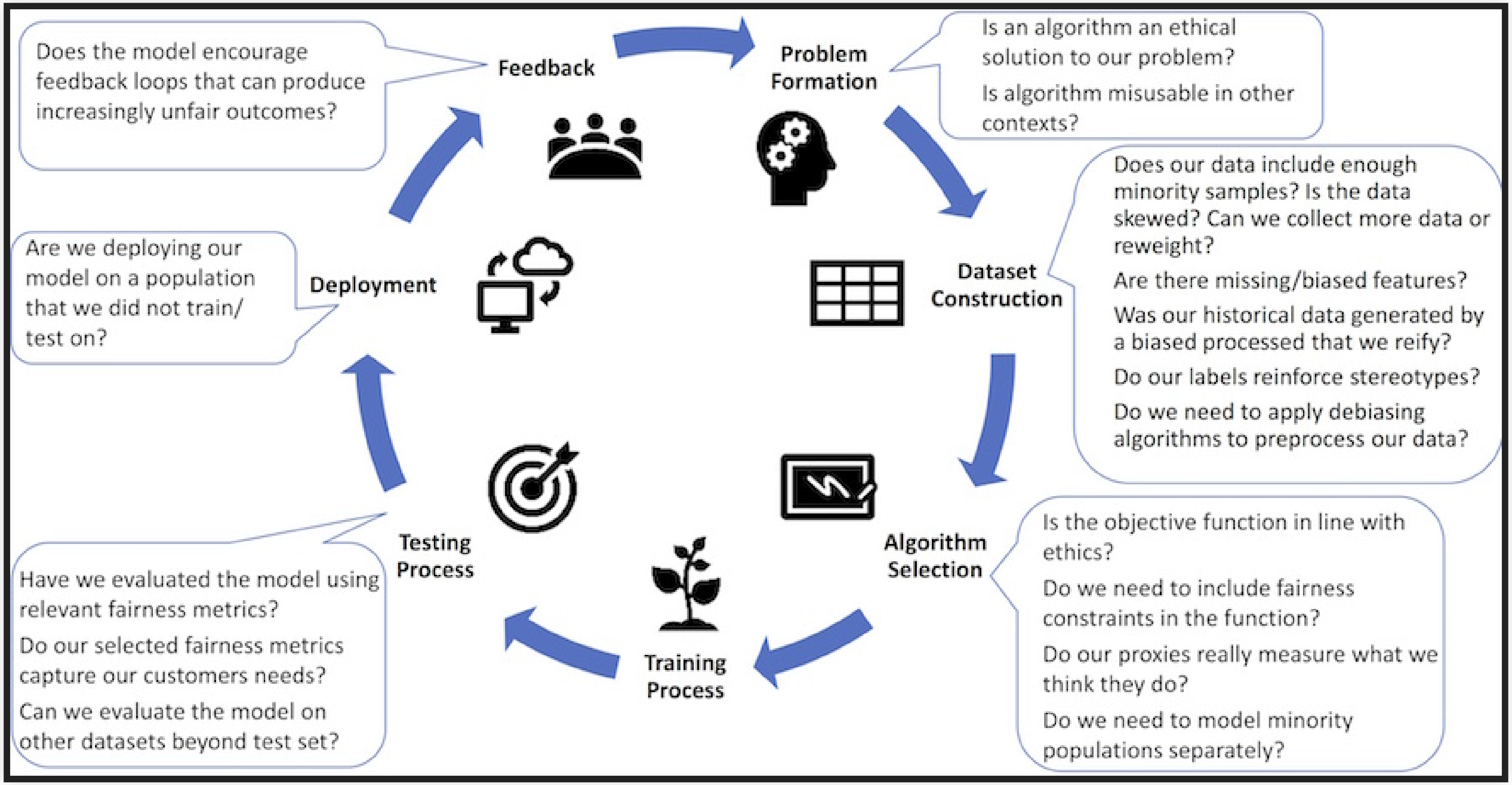
Required reading: Os Keyes, Jevan Hutson, Meredith Durbin. A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. CHI Extended Abstracts, 2019.

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
- Consider the potential impact of feedback loops on AI-based systems and need for continuous monitoring

BUILDING FAIR ML SYSTEMS

Fairness must be considered throughout the ML lifecycle!



Does the model encourage feedback loops that can produce increasingly unfair outcomes?

Is an algorithm an ethical solution to our problem?

Is algorithm misusable in other contexts?

Does our data include enough minority samples? Is the data skewed? Can we collect more data or reweight?

Are there missing/biased features?

Was our historical data generated by a biased processed that we reify?

Do our labels reinforce stereotypes?

Do we need to apply debiasing algorithms to preprocess our data?

Is the objective function in line with ethics?

Do we need to include fairness constraints in the function?

Do our proxies really measure what we think they do?

Do we need to model minority populations separately?

Have we evaluated the model using relevant fairness metrics?

Do our selected fairness metrics capture our customers needs?

Can we evaluate the model on other datasets beyond test set?

Are we deploying our model on a population that we did not train/ test on?

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

FAIRNESS DEFINITIONS: REVIEW

REVIEW OF DEFINITIONS SO FAR:

Recidivism scenario: Should a person be detained?

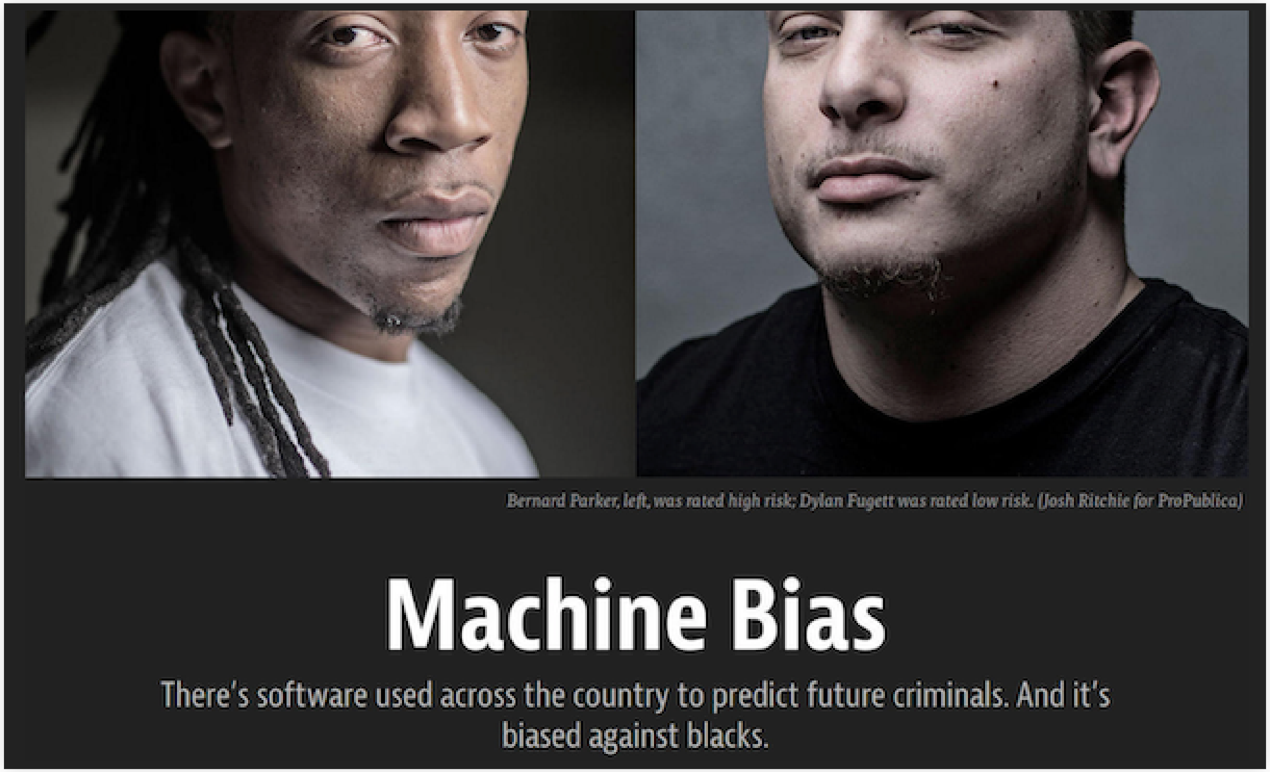
- Anti-classification: ?
- Independence: ?
- Separation: ?

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



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?

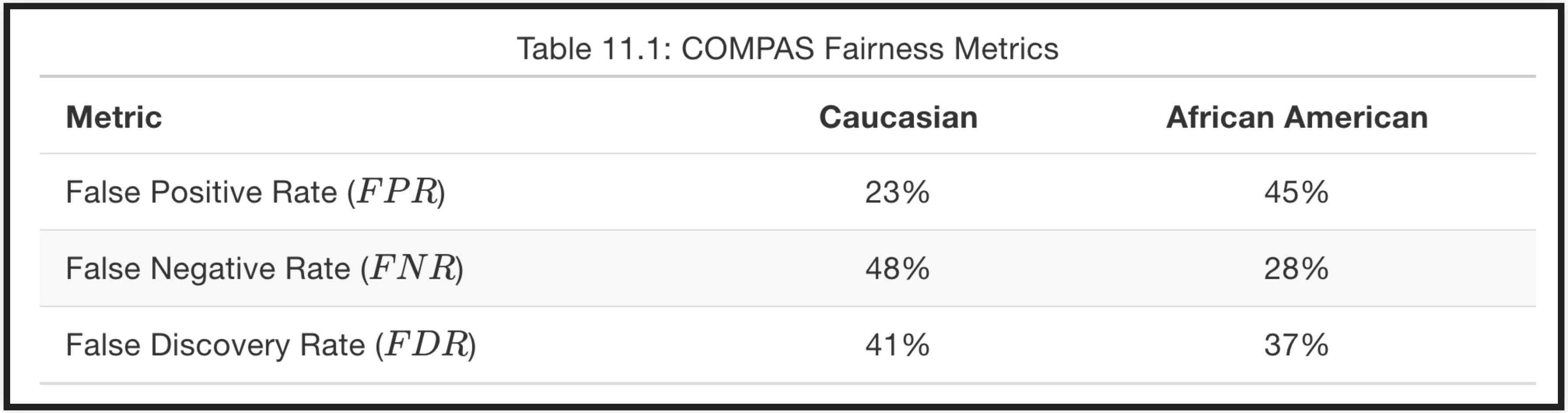


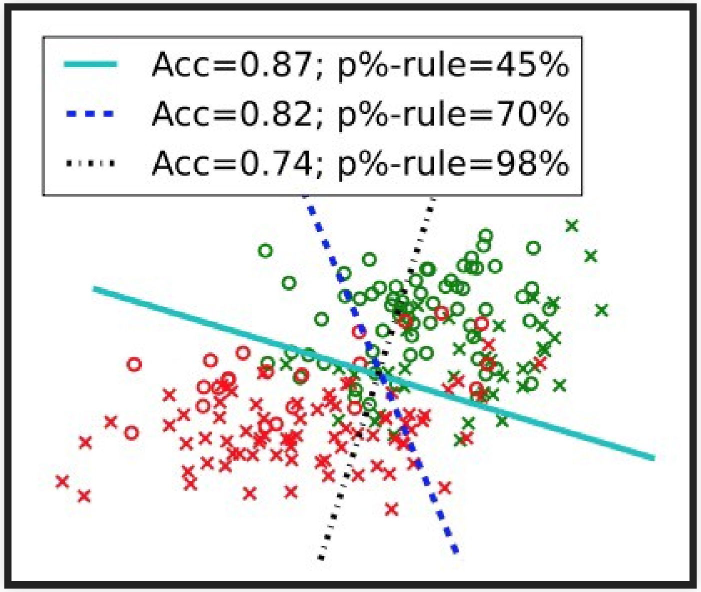
Table 11.1: COMPAS Fairness Metrics

Metric	Caucasian	African American
False Positive Rate (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

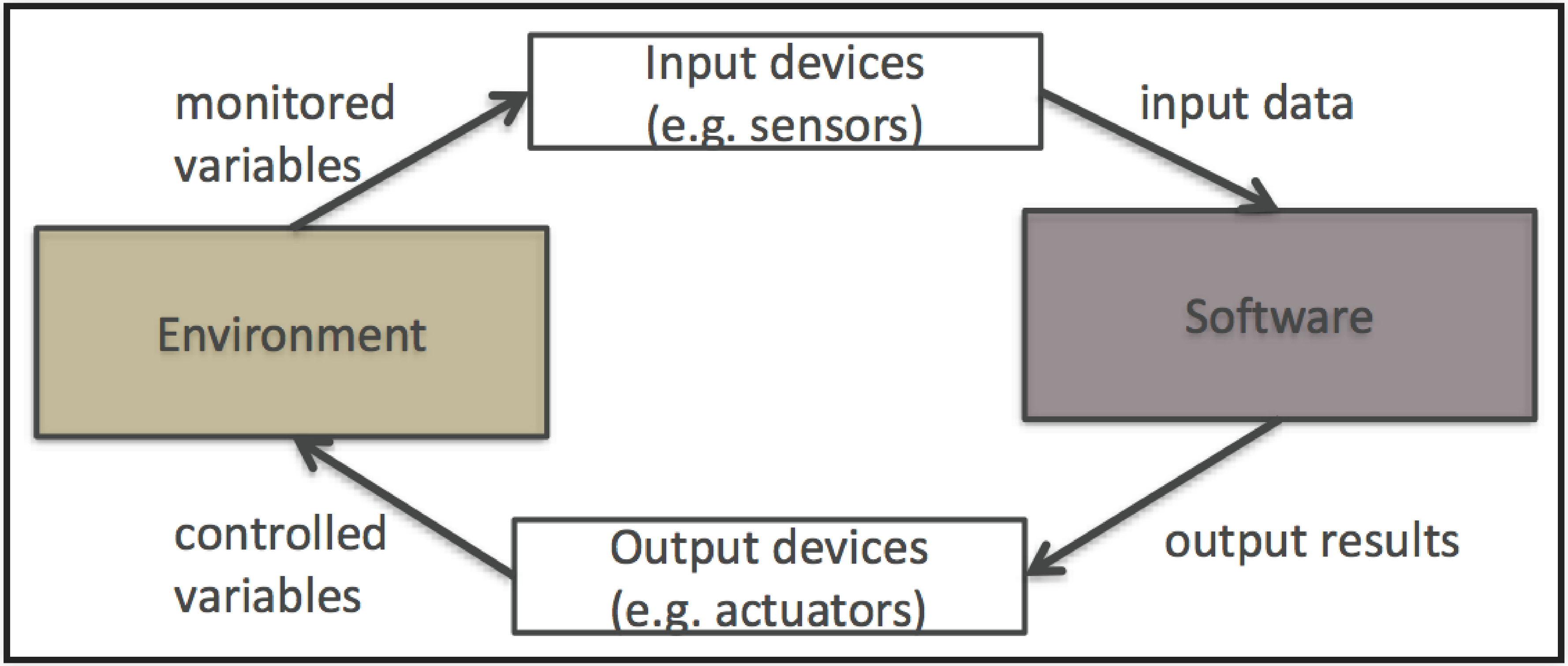
FAIRNESS DEFINITIONS: PITFALLS



- Easy to pick some definition & claim that the model is fair
  - But is the **overall system** actually fair?
  - What are the root causes of bias in the first place?
- In general, impossible to satisfy multiple fairness definitions at once
  - Also consider trade-offs against accuracy & other system goals
- Fairness is a **context-dependent** notion!
  - Select the criteria that minimize harm for the given context

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

- Identify requirements (REQ) over the environment
  - What types of harm can be caused by biased decisions?
  - Who are stakeholders? Which population groups can be harmed?
  - Are we trying to achieve equality vs. equity?
  - What are 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?
- Identify the environmental assumptions (ENV)
  - Adversarial? Misuse? Unfair (dis-)advantages?
  - Population distributions?
- Devise machine specifications (SPEC) that are sufficient to establish REQ
  - What type of fairness definition is appropriate?

"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?

EXAMPLE: LOAN APPLICATION

LOAN APPLICATION

APPROVED

Personal Information

Name (Last) PUBLIC (First) JOHN (Middle Initial) Home Telephone 111-11-1111

Other Telephone 122 22 2222

Address (Mailing Address) 12345 MAIN STREET 999999

E-Mail Address JQPJQPJQP@JQPJQP

UNDER REVIEW

SUBJECT REVIEW

Services needed

UNDER REVIEW

Current Income

High School Graduate Or General Education (GED) Test Passed? Yes No

highest grade completed

Military (Most recent first)

Credits Earned

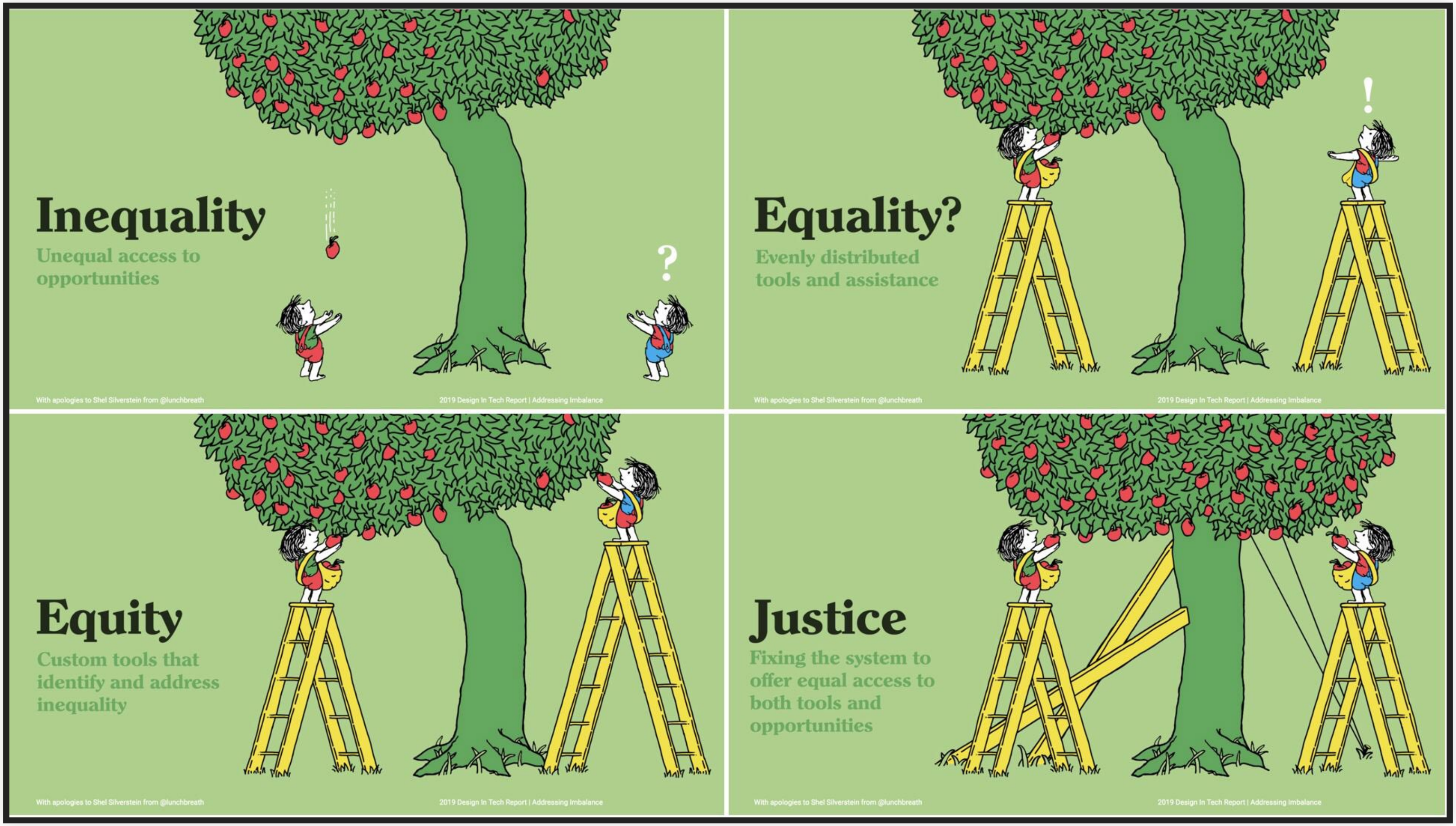
Quarterly or

Other (Specify)

Major or Subject

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

RECALL: EQUALITY VS EQUITY


Inequality
Unequal access to opportunities
?
With apologies to Shel Silverstein from @lunchbreath
2019 Design In Tech Report | Addressing Imbalance
Equality?
Evenly distributed tools and assistance
!
With apologies to Shel Silverstein from @lunchbreath
2019 Design In Tech Report | Addressing Imbalance
Equity
Custom tools that identify and address inequality
With apologies to Shel Silverstein from @lunchbreath
2019 Design In Tech Report | Addressing Imbalance
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

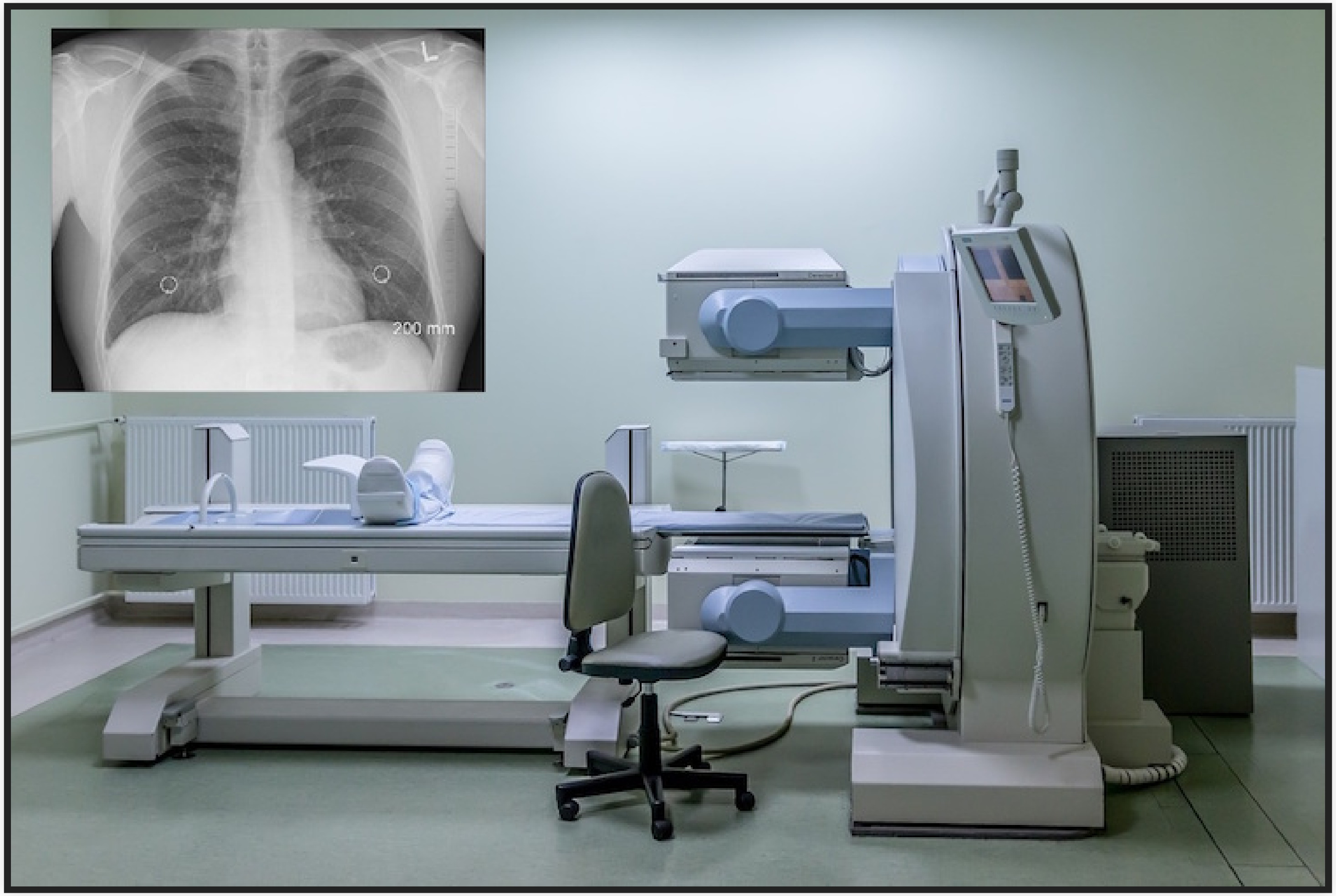
TYPE OF DECISION & POSSIBLE HARM

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

WHICH FAIRNESS CRITERIA?

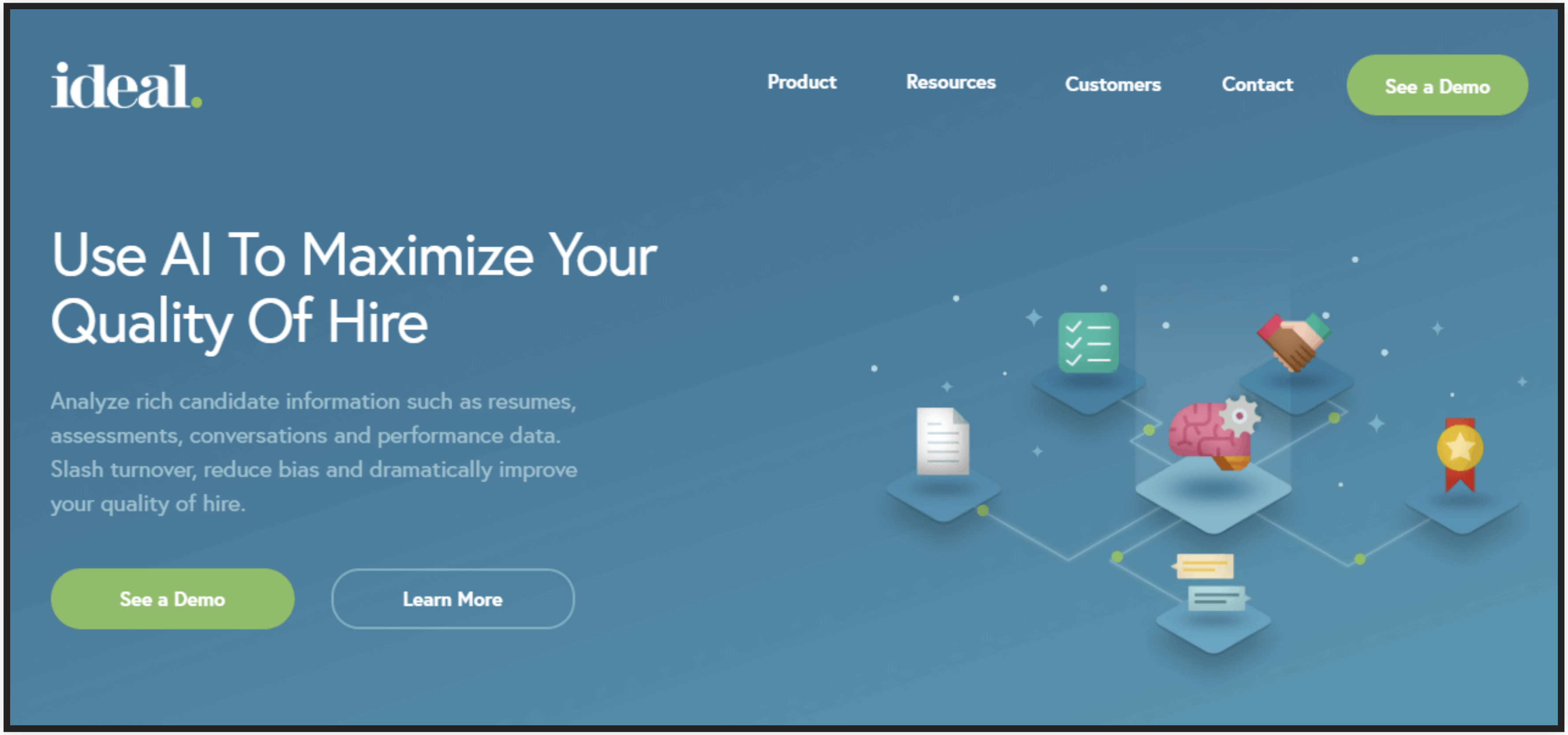
- Decision: Should an applicant be granted a loan?
- What kind of harm can be caused? Punitive or assistive?
- Criteria: Anti-classification, independence, or seperation w/ FPR or FNR?

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, independence, or seperation w/ FPR or FNR?

BREAKOUT: AUTOMATED HIRING

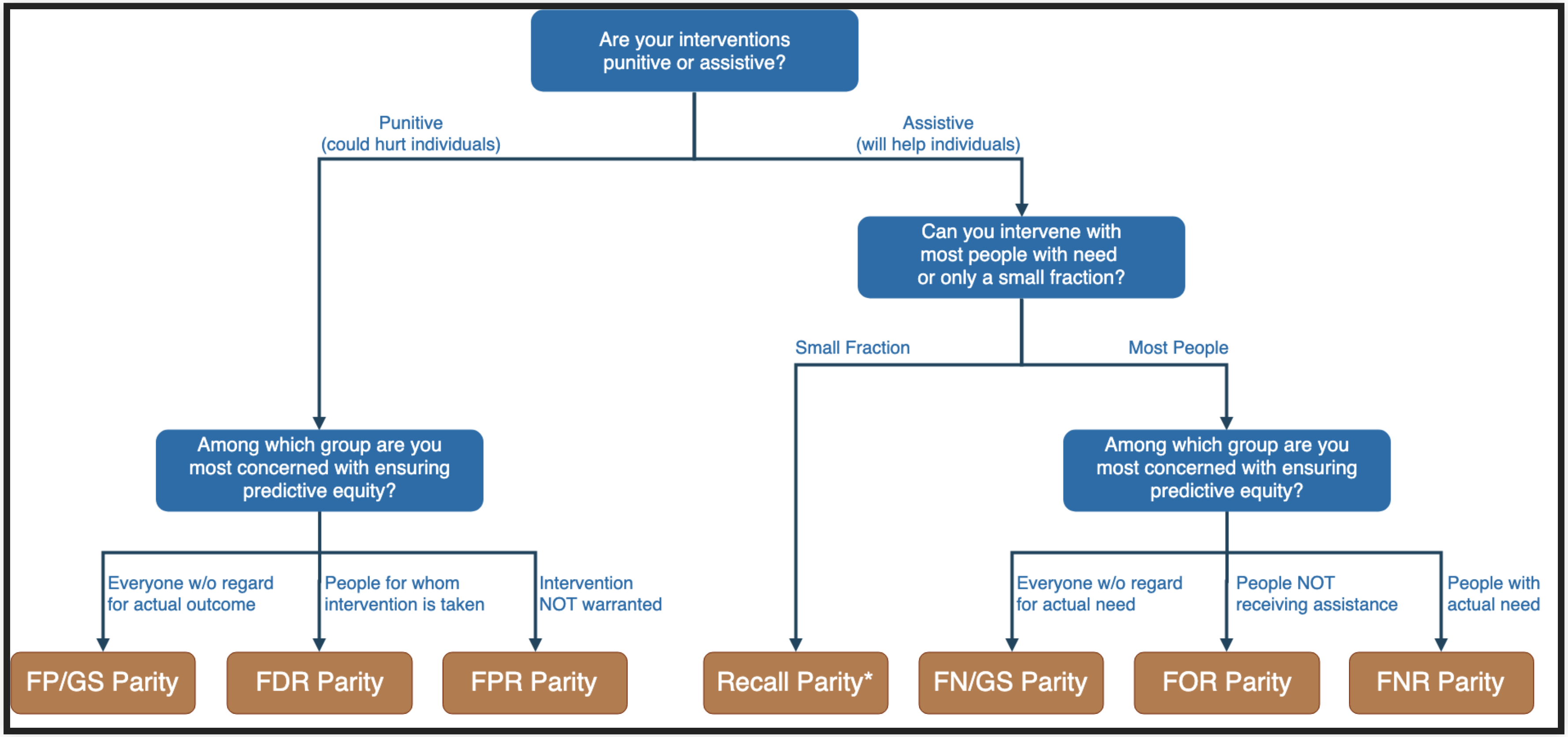


Use AI To Maximize Your Quality Of Hire

Analyze rich candidate information such as resumes, assessments, conversations and performance data. Slash turnover, reduce bias and dramatically improve your quality of hire.

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

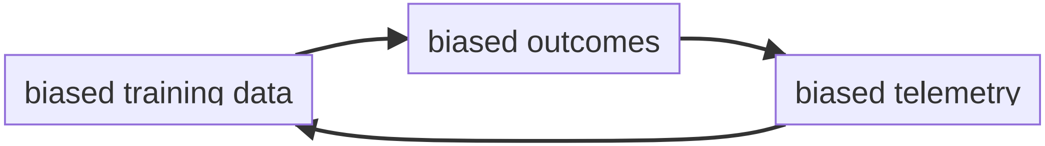
FAIRNESS TREE


Are your interventions punitive or assistive?
Punitive (could hurt individuals)
Assistive (will help individuals)
Can you intervene with most people with need or only a small fraction?
Small Fraction
Most People
Among which group are you most concerned with ensuring predictive equity?
Everyone w/o regard for actual outcome
People for whom intervention is taken
Intervention NOT warranted
Among which group are you most concerned with ensuring predictive equity?
Everyone w/o regard for actual need
People NOT receiving assistance
People with actual need
FP/GS Parity
FDR Parity
FPR Parity
Recall Parity*
FN/GS Parity
FOR Parity
FNR Parity

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

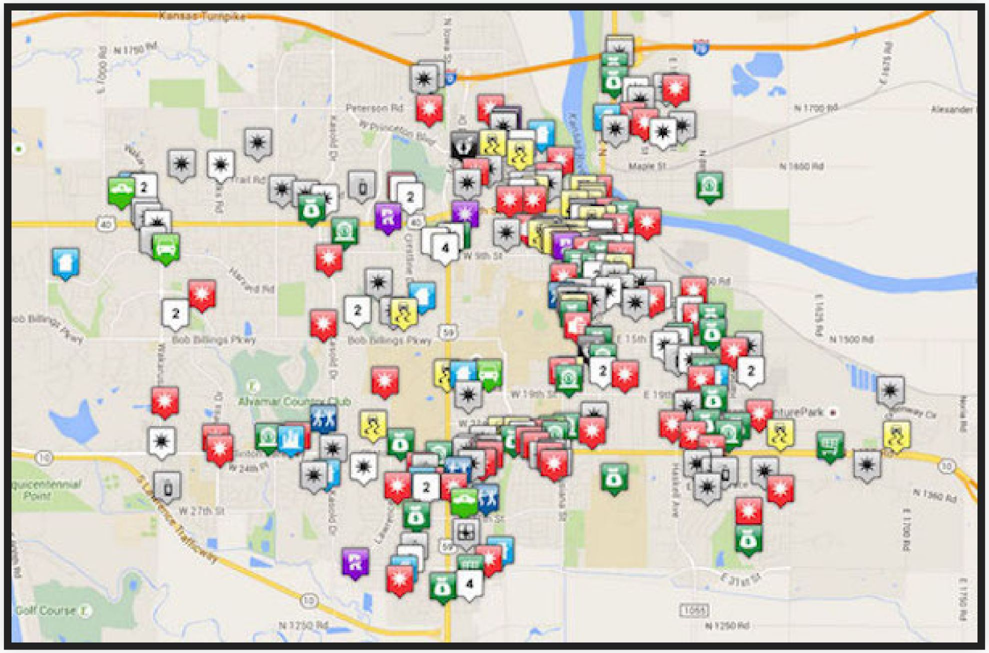
FEEDBACK LOOPS

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*

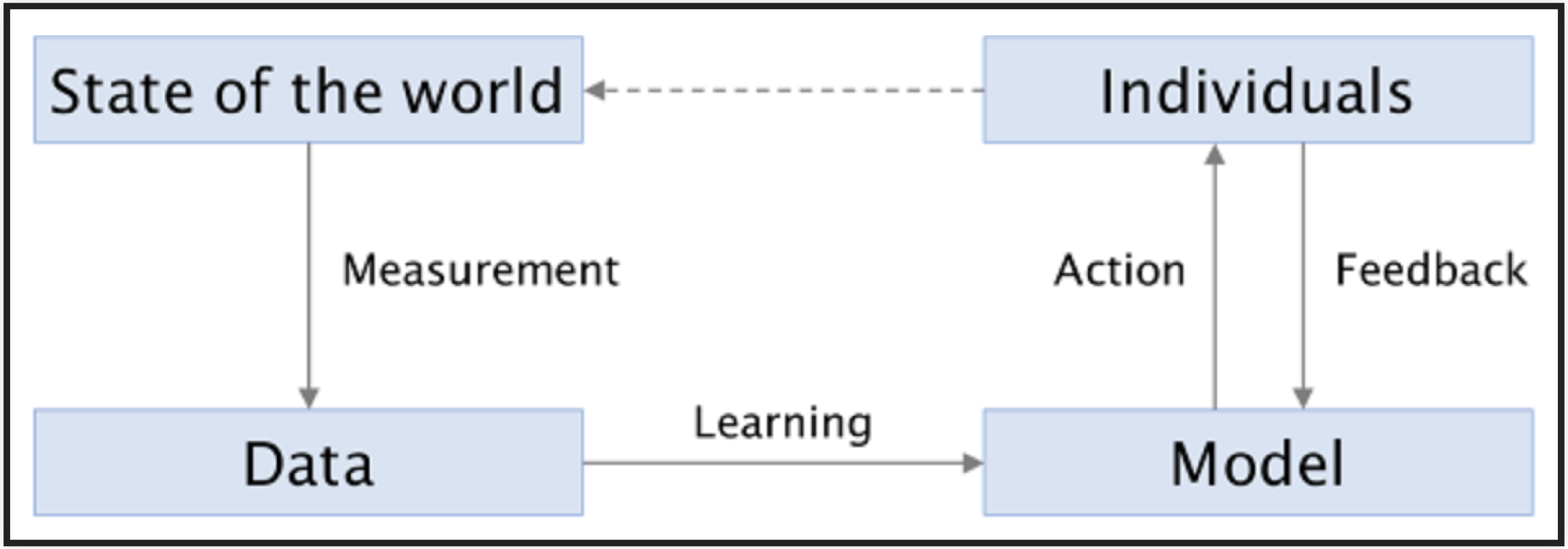
EXAMPLE: PREDICTIVE POLICING



- 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...

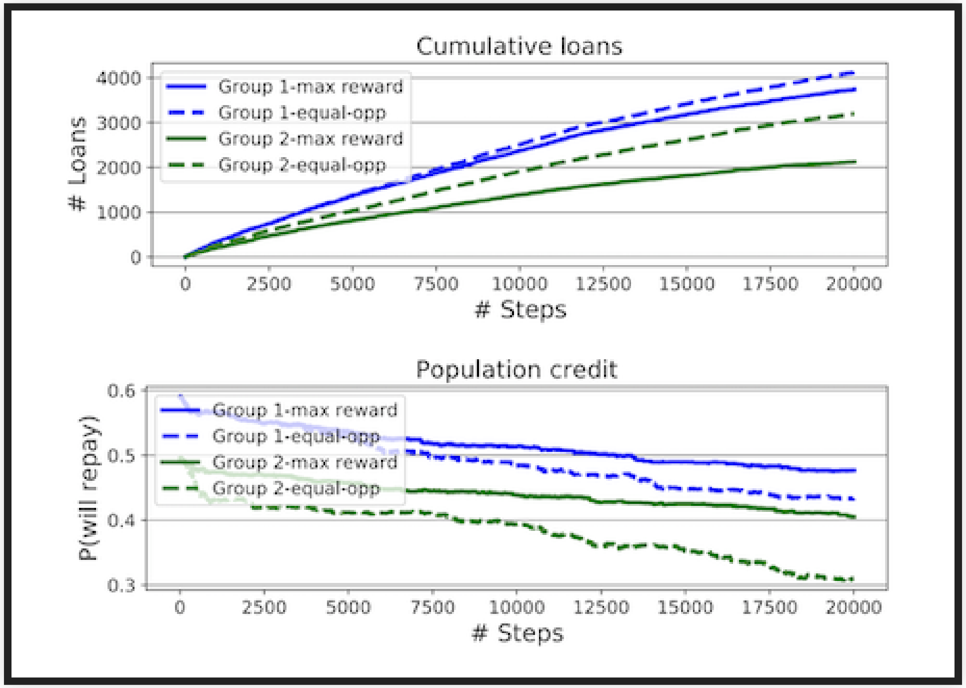
Q. Other examples?

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 over time
- Even if a model appears to promote fairness in short term, it may result harm over a long-term period

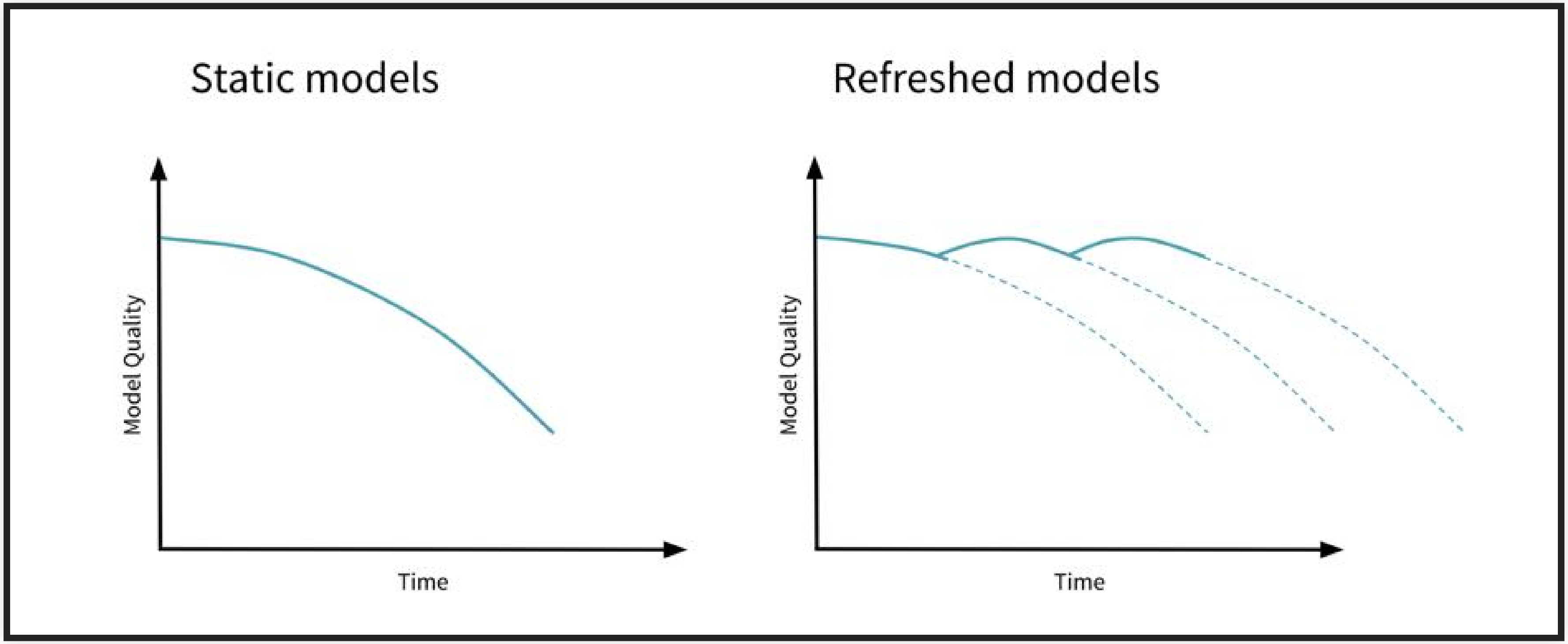
Fairness is not static: deeper understanding of long term fairness via simulation studies, in FAT* 2020.

MONITORING AND AUDITING

MONITORING & AUDITING

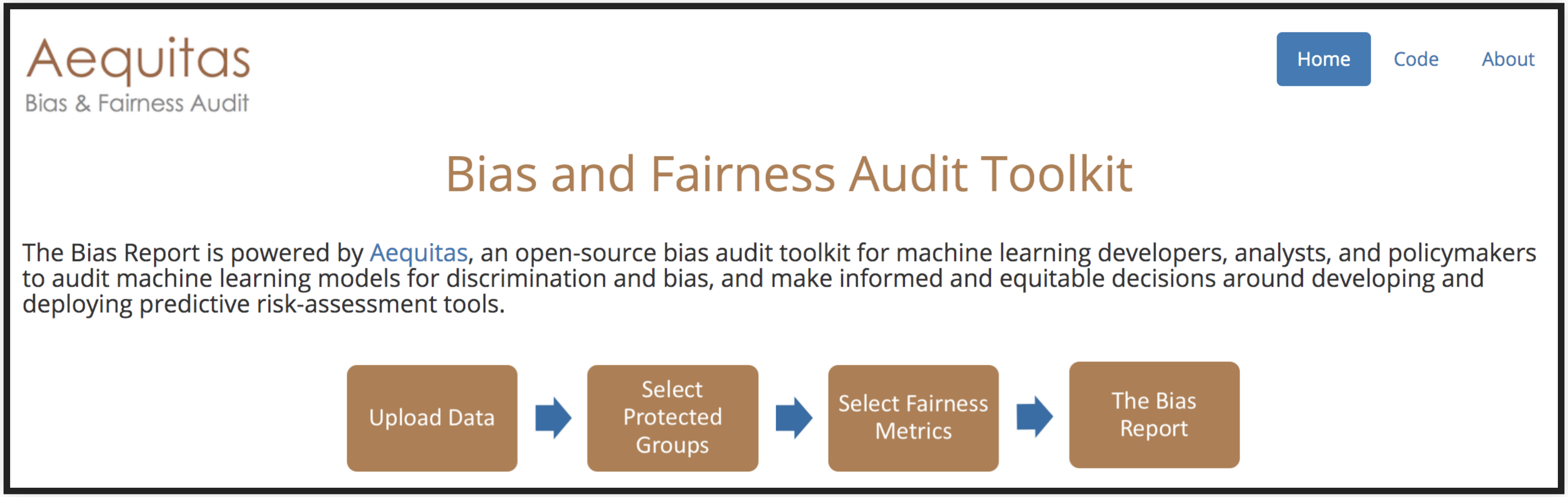
- Continuously monitor for:
    - Match between training data, test data, and instances that you encounter in deployment
    - Fairness metrics: Is the system yielding fair results over time?
    - Population shifts: May suggest needs to adjust fairness metric/thresholds
    - User reports & complaints: 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 & AUDITING



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

MONITORING TOOLS: EXAMPLE

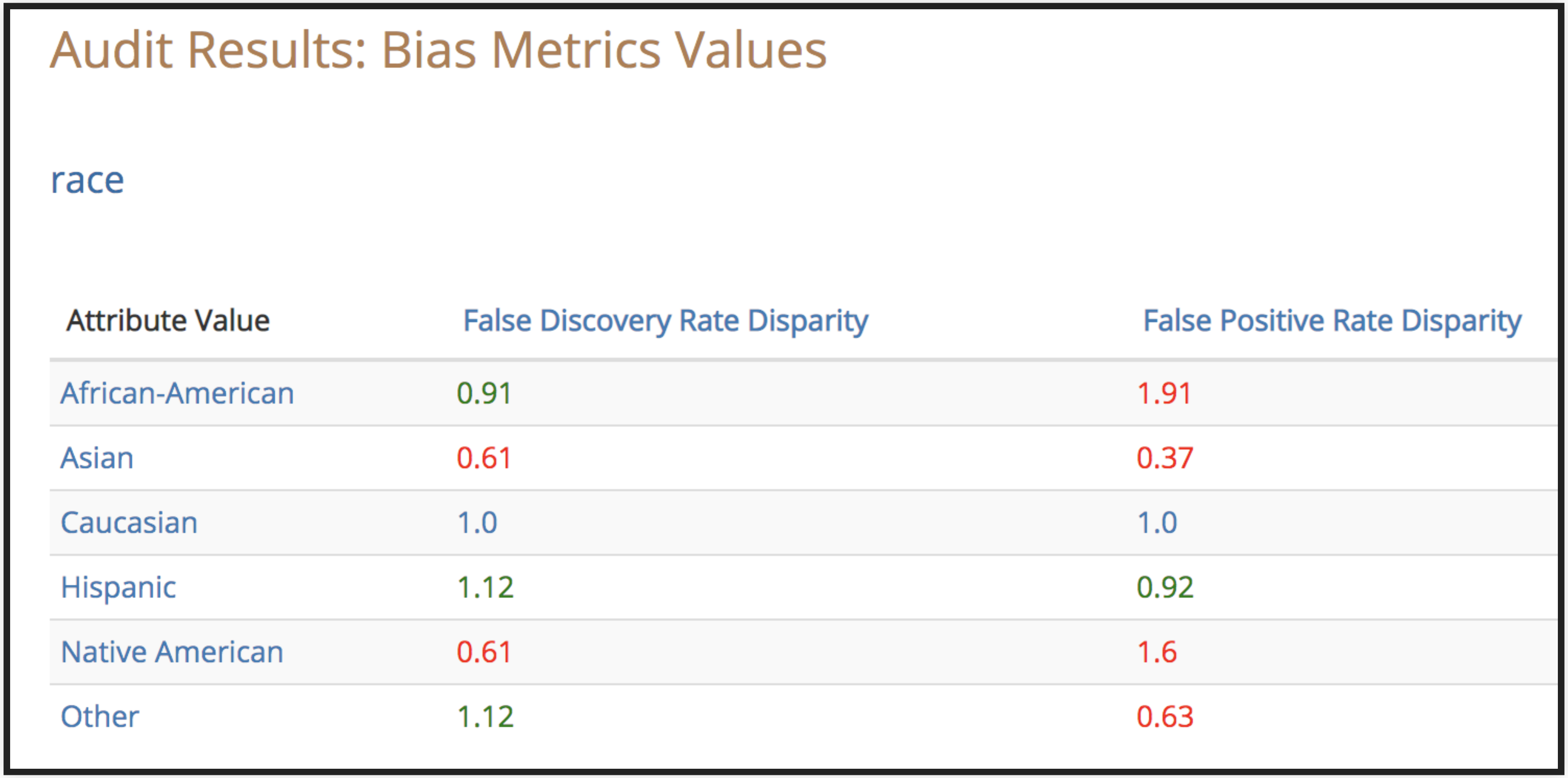


Bias and Fairness Audit Toolkit

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



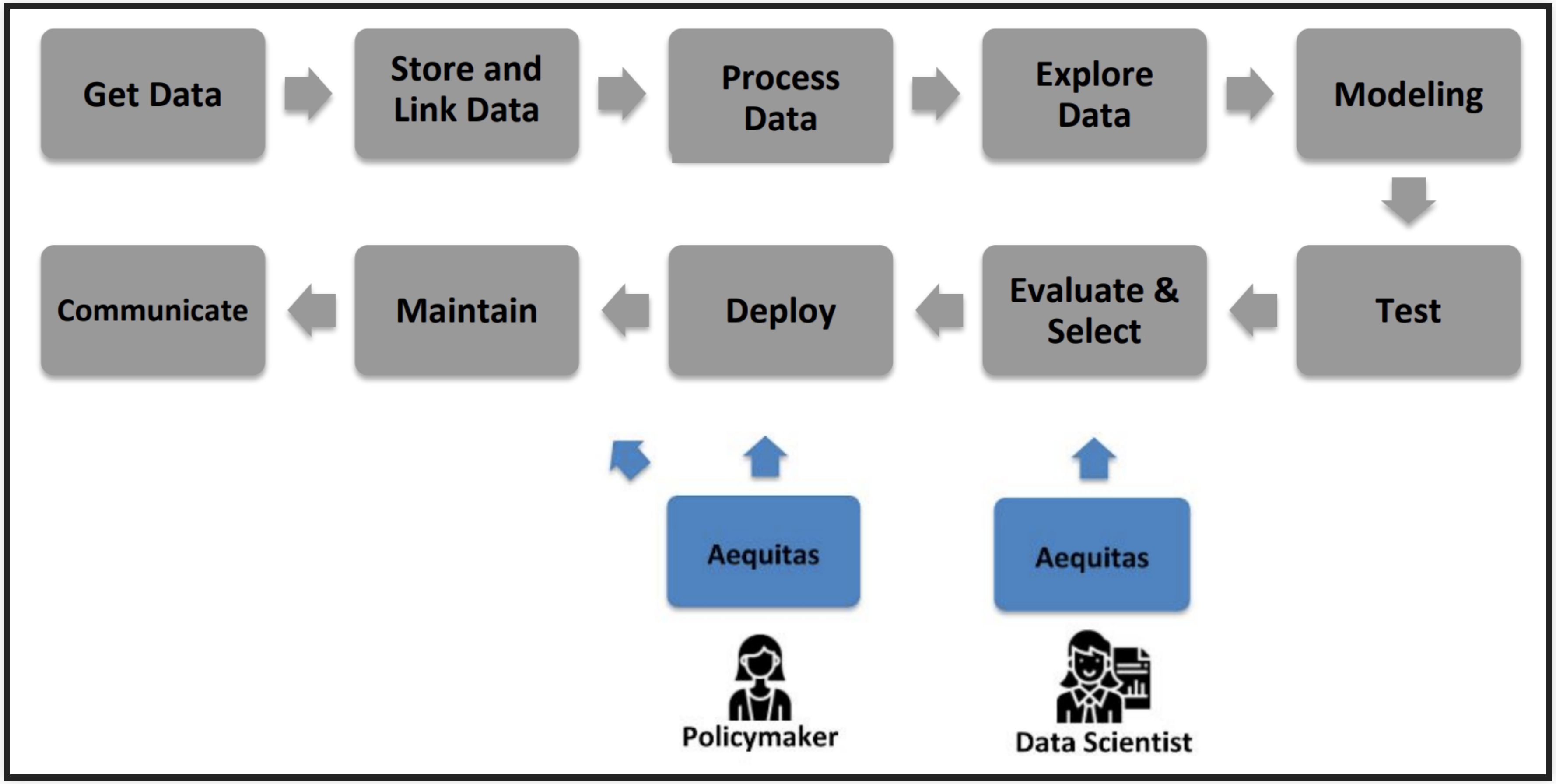
Audit Results: Bias Metrics Values

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)

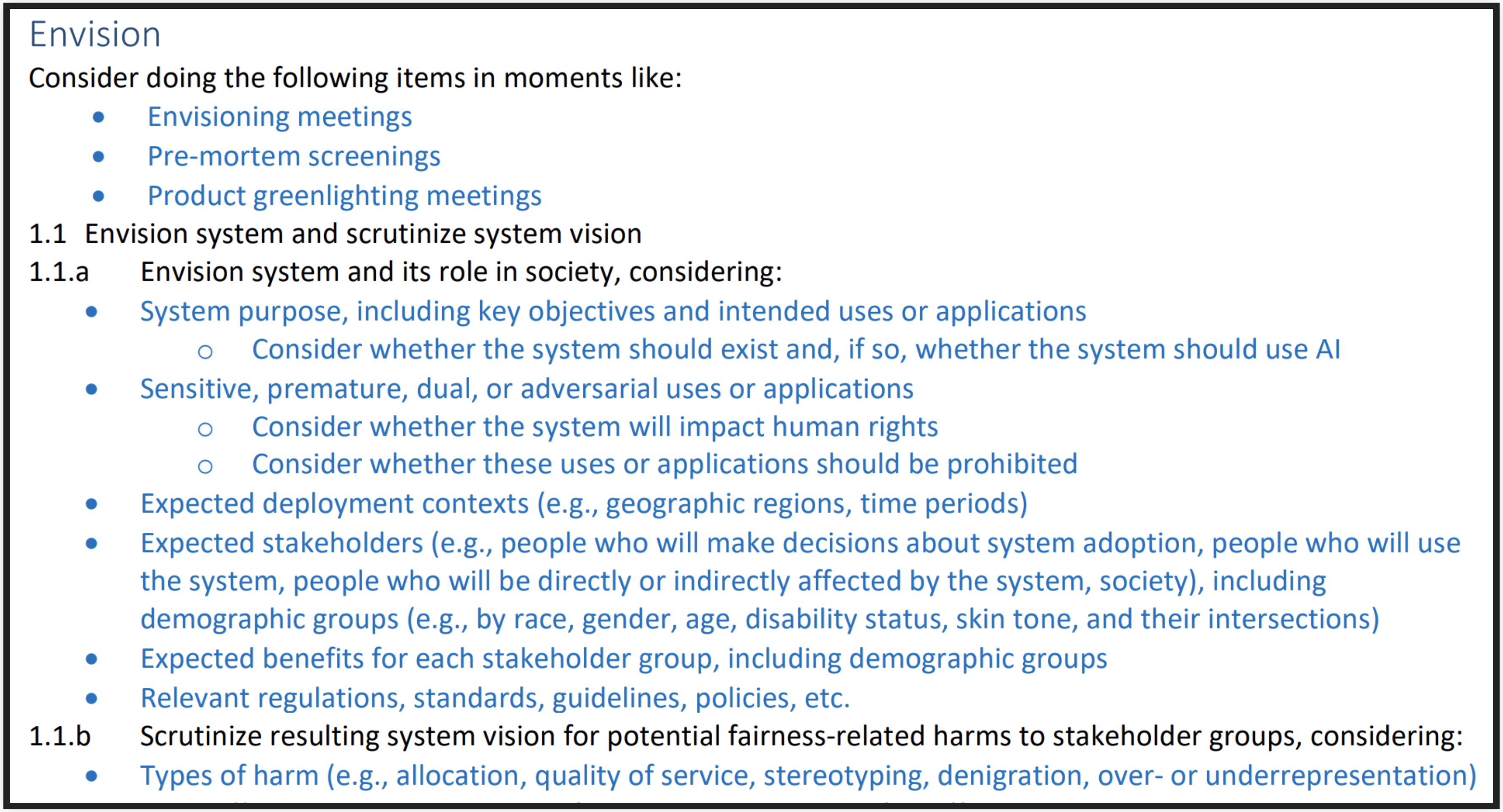
BEST PRACTICES: TESTING & DEPLOYMENT

- 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)

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
- Consideration for the impact of feedback loops
- Continous montoring & auditing for fairness