Machine Learning in Production

Building Fair Products

From Fairness Concepts to Fair Products

Fundamentals of Engineering AI-Enabled Systems

Holistic system view: AI and non-AI components, pipelines, stakeholders, environment interactions, feedback loops

Requirements:
System and model goals
User requirements
Environment assumptions
Quality beyond accuracy
Measurement
Risk analysis
Planning for mistakes

Architecture + design:
Modeling tradeoffs
Deployment architecture
Data science pipelines
Telemetry, monitoring
Anticipating evolution
Big data processing
Human-AI design

Quality assurance:
Model testing
Data quality
QA automation
Testing in production
Infrastructure quality
Debugging

Operations:
Continuous deployment
Contin. experimentation
Configuration mgmt.
Monitoring
Versioning
Big data
DevOps, MLOps

Teams and process: Data science vs software eng. workflows, interdisciplinary teams, collaboration points, technical debt

Responsible AI Engineering

Provenance, versioning, reproducibility

Safety

Security and privacy

Fairness

Interpretability and explainability

Transparency and trust

Ethics, governance, regulation, compliance, organizational culture

Reading

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:

- Metcalf, Jacob, and Emanuel Moss. "Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics." *Social Research: An International Quarterly* 86, no. 2 (2019): 449-476.

Learning Goals

- Understand the role of requirements engineering in selecting ML fairness criteria
- Understand the process of constructing datasets for fairness
- Document models and datasets to communicate fairness concerns
- 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

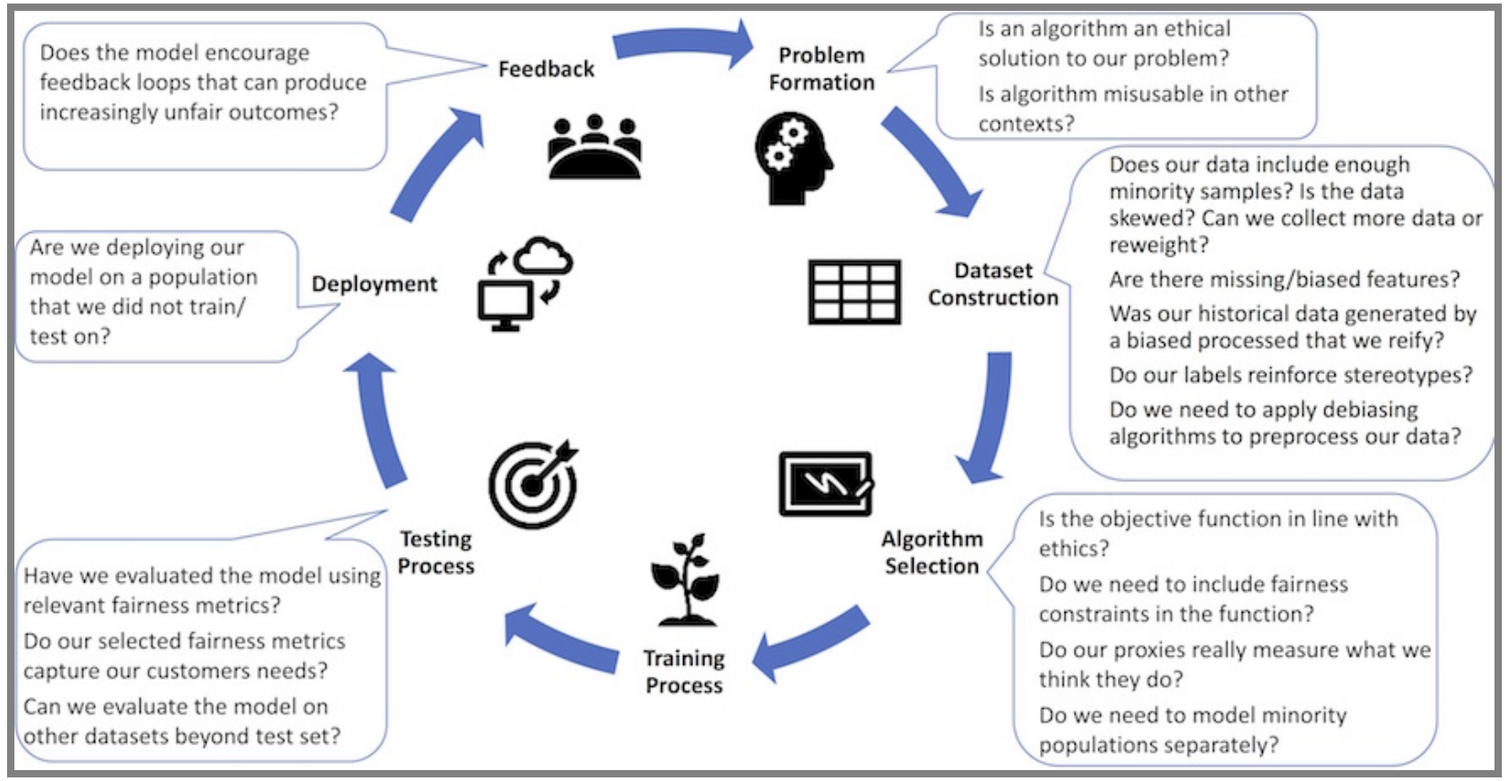
Improving Fairness of a Model

In all pipeline stages:

- Data collection
- Data cleaning, processing
- Training
- Inference
- Evaluation and auditing

Today: Model-centric view

Consider fairness throughout the ML lifecycle!



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

1. Improve with Model Evaluation and Auditing

Lots of tools to measure and visualize fairness with many metrics

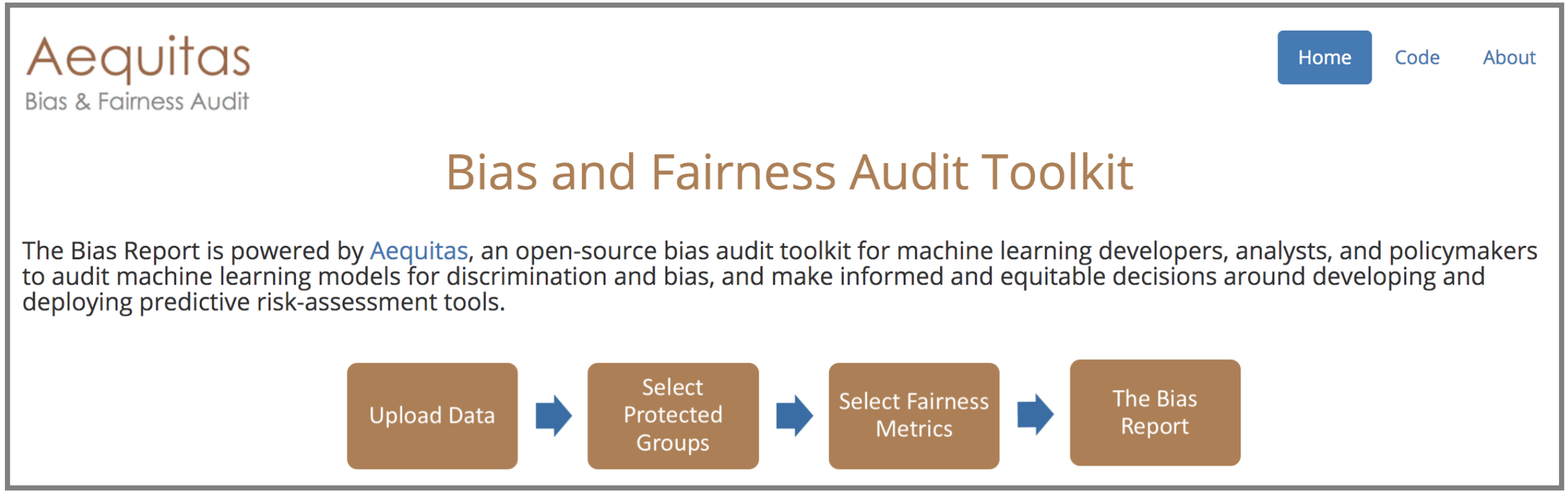
Can be integrated in notebooks and production (telemetry, monitoring)

Audit: In-depth evaluation of a model snapshot

Efforts to crowdsource feedback and audits

Debugging tools to investigate potential fairness issues

Example audit tool: Aequitas



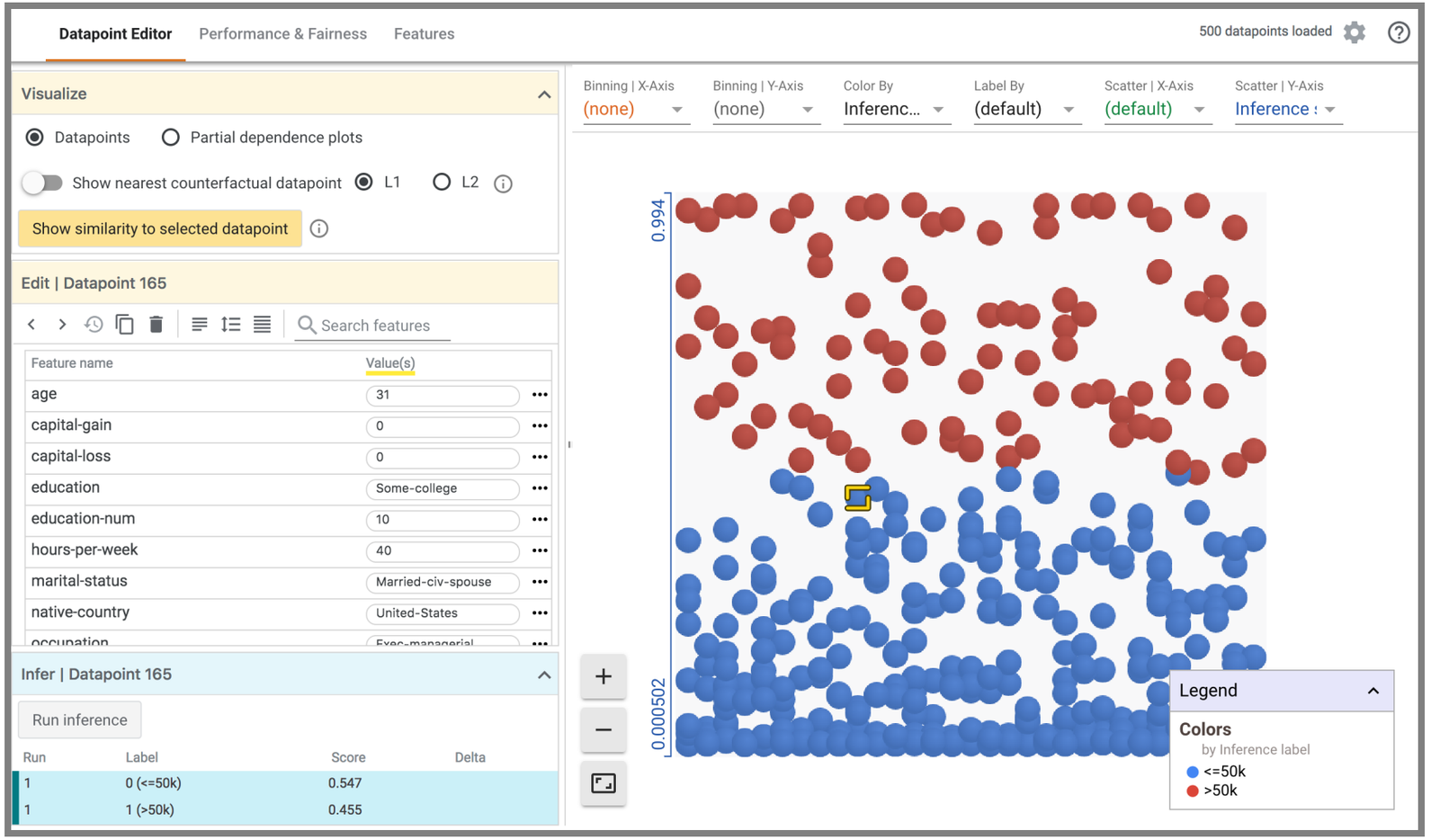
Example audit tool: Aequitas

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

Example debugging tool: What-If



Google What-If Tool

2. Improve during Model Inference

Remove/scramble protected attributes and correlated attributes? (anti-classification)

Calibrate by adjusting thresholds (group fairness, equalized odds)

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

Weaken predictor for one group?

Example: Tweaking Thresholds

3. Improve during Model Training

Incorporate fairness metric during training, e.g., in loss function

Use fairness for model selection/hyperparameter tuning

Weigh training data differently based on (expected) bias or trust

Much research, many approaches…

Further reading: Pessach, Dana, and Erez Shmueli. "A Review on Fairness in Machine Learning." ACM Computing Surveys (CSUR) 55, no. 3 (2022): 1-44.

4. Improve during Data Cleaning, Feature Engineering

Remove features for protected attributes; measure correlations to identify proxies <- anti-classification

Correct for known biases, e.g.,

- Discard known biased training data, fix *tainted labels*
- Remove training data influenced by *feedback loop*
- Analyze data for *limited features*, remove or enhance
- Augment data for *sample size disparity*
- Normalize data across subpopulations

Active research field of data debugging to find influential outliers and potential bias (more later in Explainability lecture)

5. Improvement during Data Collection

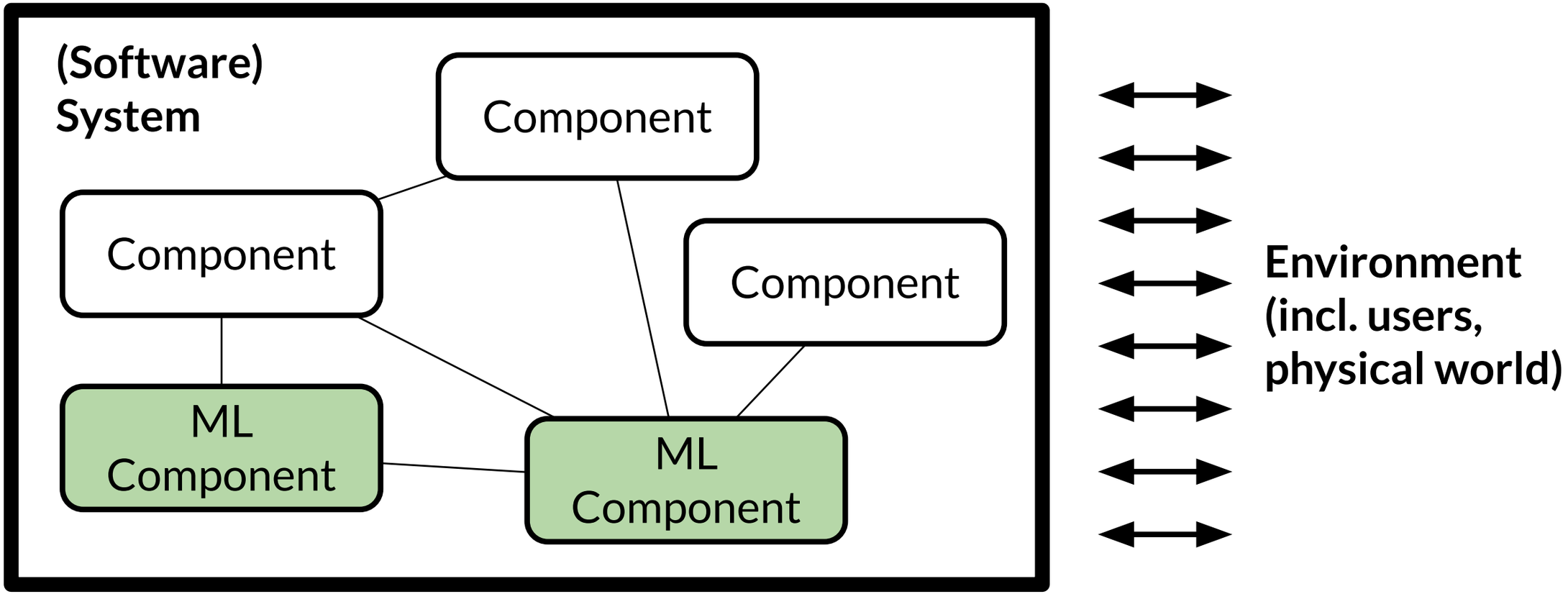
Carefully review data collection procedures, sampling biases, what data is collected, how trustworthy labels are, etc.

Can address most sources of bias: tainted labels, skewed samples, limited features, sample size disparity, proxies:

- deliberate what data to collect
- collect more data, oversample where needed
- extra effort in unbiased labels

Potentially expensive, but typically **highest leverage point**

Recall: Model vs System

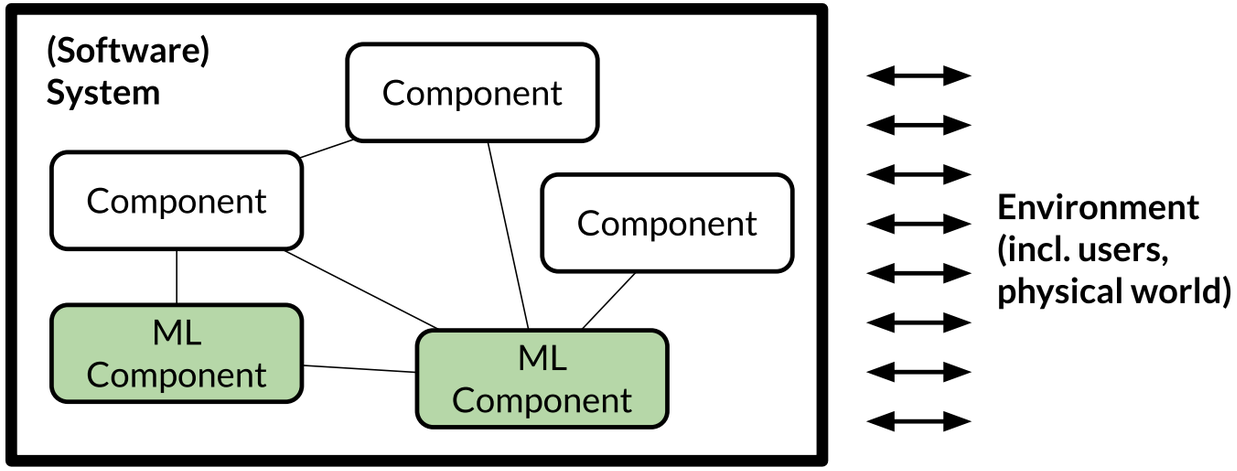


Fairness is a System Quality

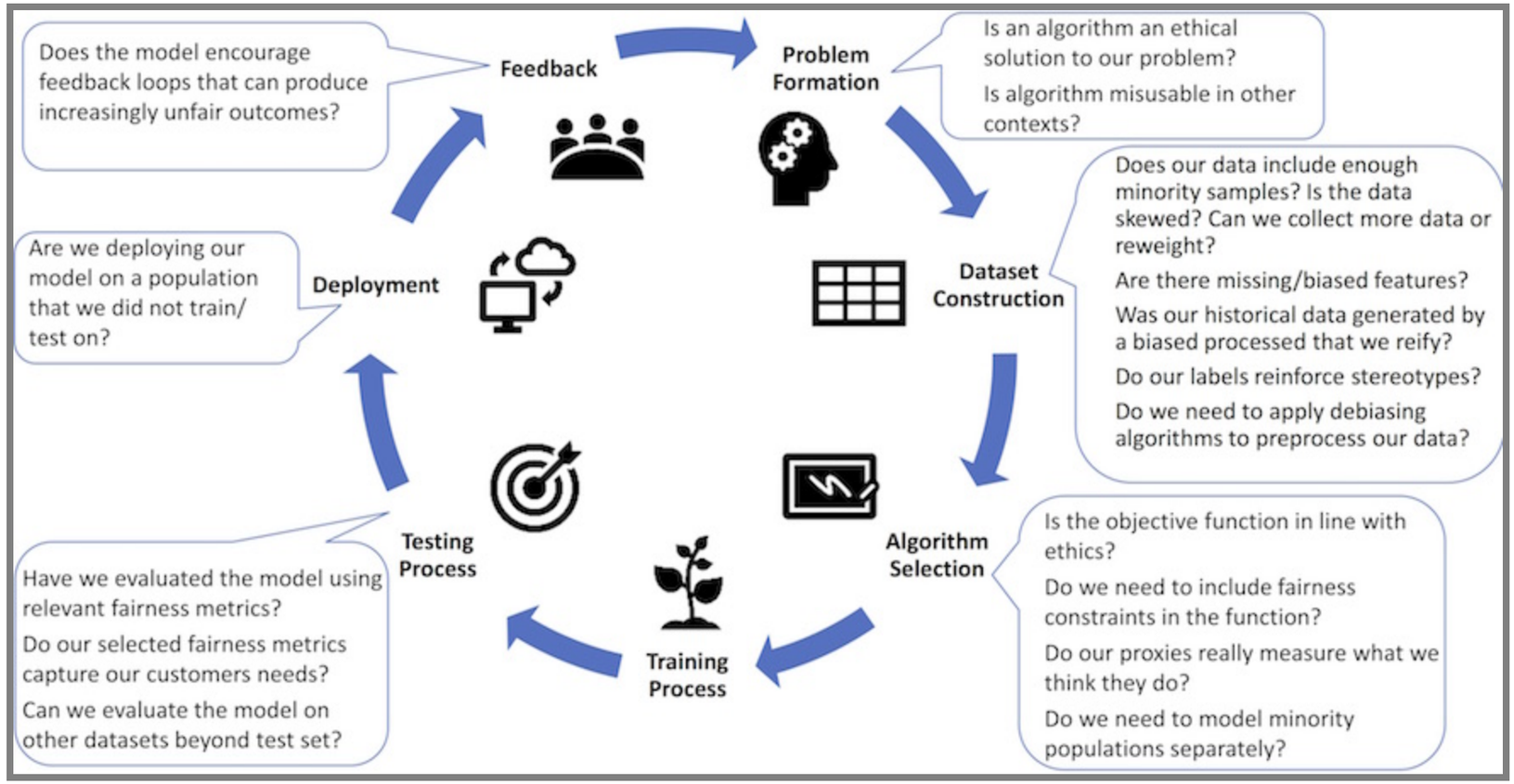
Fairness can be measured for a model

... but we really care whether the system, as it interacts with the environment, is fair/safe/secure

... does the system cause harm?



Fair ML Pipeline Process



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

Fairness Problems are System-Wide Challenges

- **Requirements engineering challenges:** How to identify fairness concerns, fairness metric, design data collection and labeling
- **Human-computer-interaction design challenges:** How to present results to users, fairly collect data from users, design mitigations
- **Quality assurance challenges:** Evaluate the entire system for fairness, continuously assure in production
- **Process integration challenges:** Incoprorate fairness work in development process
- **Education and documentation challenges:** Create awareness, foster interdisciplinary collaboration

Understanding System-Level Goals for Fairness

i.e., Requirements engineering

Recall: Fairness metrics

- Anti-classification (fairness through blindness)
- Group fairness (independence)
- Equalized odds (separation)
- ...and numerous others and variations!

But which one makes most sense for my application?

Recall: What is fair?

Fairness discourse asks questions about how to treat people and whether treating different groups of people differently is ethical. If two groups of people are systematically treated differently, this is often considered unfair.

Intuitive Justice

Research on what post people perceive as fair/just (psychology)

When rewards depend on inputs and participants can chose contributions: Most people find it fair to split rewards proportional to inputs

- *Which fairness measure does this relate to?*

Most people agree that for a decision to be fair, personal characteristics that do not influence the reward, such as gender or age, should not be considered when dividing the rewards.

- *Which fairness measure does this relate to?*

Key issue: Unequal starting positions

Not everybody starts from an equal footing -- individual and group differences

- Some differences are inert, e.g., younger people have (on average) less experience
- Some differences come from past behavior/decisions, e.g., whether to attend college
- Some past decisions and opportunities are influenced by past injustices, e.g., redlining creating generational wealth differences

Individual and group differences not always clearly attributable, e.g., nature vs nurture discussion

Unequal starting position

Fair or not? Should we account for unequal starting positions?

- Tom is more lazy than Bob. He should get less pie.
- People in Egypt have on average a much longer work week (53h) than people in the Germany (35h). They have less time to bake and should get more pie.
- Disabled people are always exhausted quickly. They should get less pie, because they contribute less.
- Men are on average more violent than women. This should be reflected in recidivism prediction.
- Employees with a PhD should earn higher wages than those with a bachelor's degree, because they decided to invest in more schooling.
- Students from poor neighborhoods should receive extra resources at school, because they get less help at home.
- Poverty is a moral failing. Poor people are less deserving of pie.

Dealing with unequal starting positions

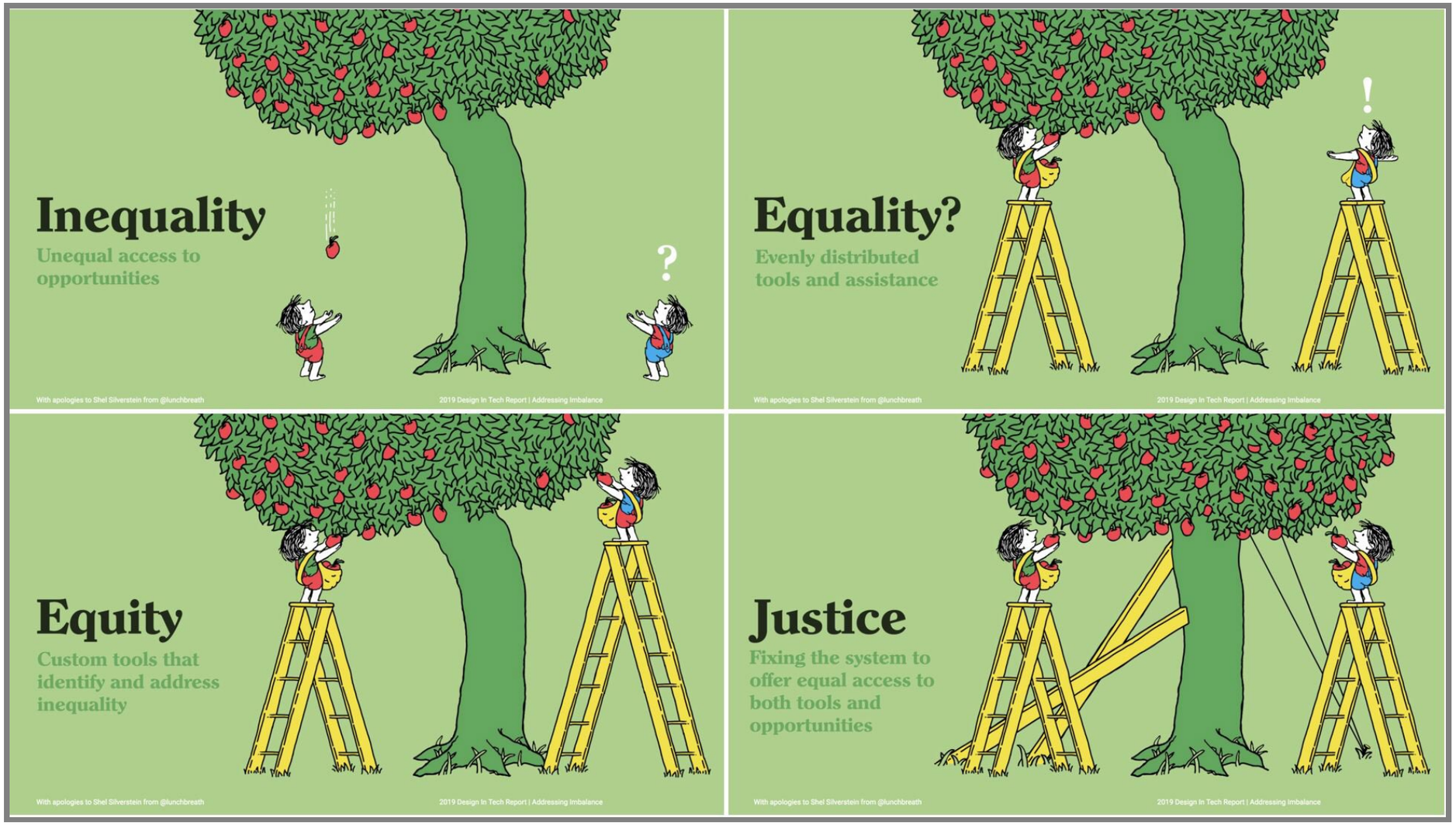
Equality (minimize disparate treatment):

- Treat everybody equally, regardless of starting position
- Focus on meritocracy, strive for fair opportunities
- Equalized-odds-style fairness; equality of opportunity

Equity (minimize disparate impact):

- Compensate for different starting positions
- Lift disadvantaged group, affirmative action
- Strive for similar outcomes (distributive justice)
- Group-fairness-style fairness; equality of outcomes

Equality vs Equity



Equality vs Equity

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.

Justice

Aspirational third option, that avoids a choice between equality and equity

Fundamentally removes initial imbalance or removes need for decision

Typically rethinks entire societal system in which the imbalance existed, beyond the scope of the ML product

Choosing Equality vs Equity

Each rooted in long history in law and philosophy

Typically incompatible, cannot achieve both

Designers need to decide

Problem dependent and goal dependent

What differences are associated with merits and which with systemic disadvantages of certain groups? Can we agree on the degree a group is disadvantaged?

Identifying Fairness Goals is a Requirements Engineering Problem

- What is the goal of the system? What benefits does it provide and to whom?
- Who are the stakeholders of the system? What are the stakeholders' views or expectations on fairness and where do they conflict? Are we trying to achieve fairness based on equality or equity?
- What subpopulations (including minority groups) may be using or be affected by the system? What types of harms can the system cause with discrimination?
- Does fairness undermine any other goals of the system (e.g., accuracy, profits, time to release)?
- Are there legal anti-discrimination requirements to consider? Are there societal expectations about ethics w.r.t. to this product? What is the activist position?
- ...

Analyzing Potential Harms

Anticipate harms from unfair decisions

- Harms of allocation, harms of representation?
- How do biased model predictions contribute to system behavior? (show predictions, act on predictions?)

Consider how automation can amplify harm

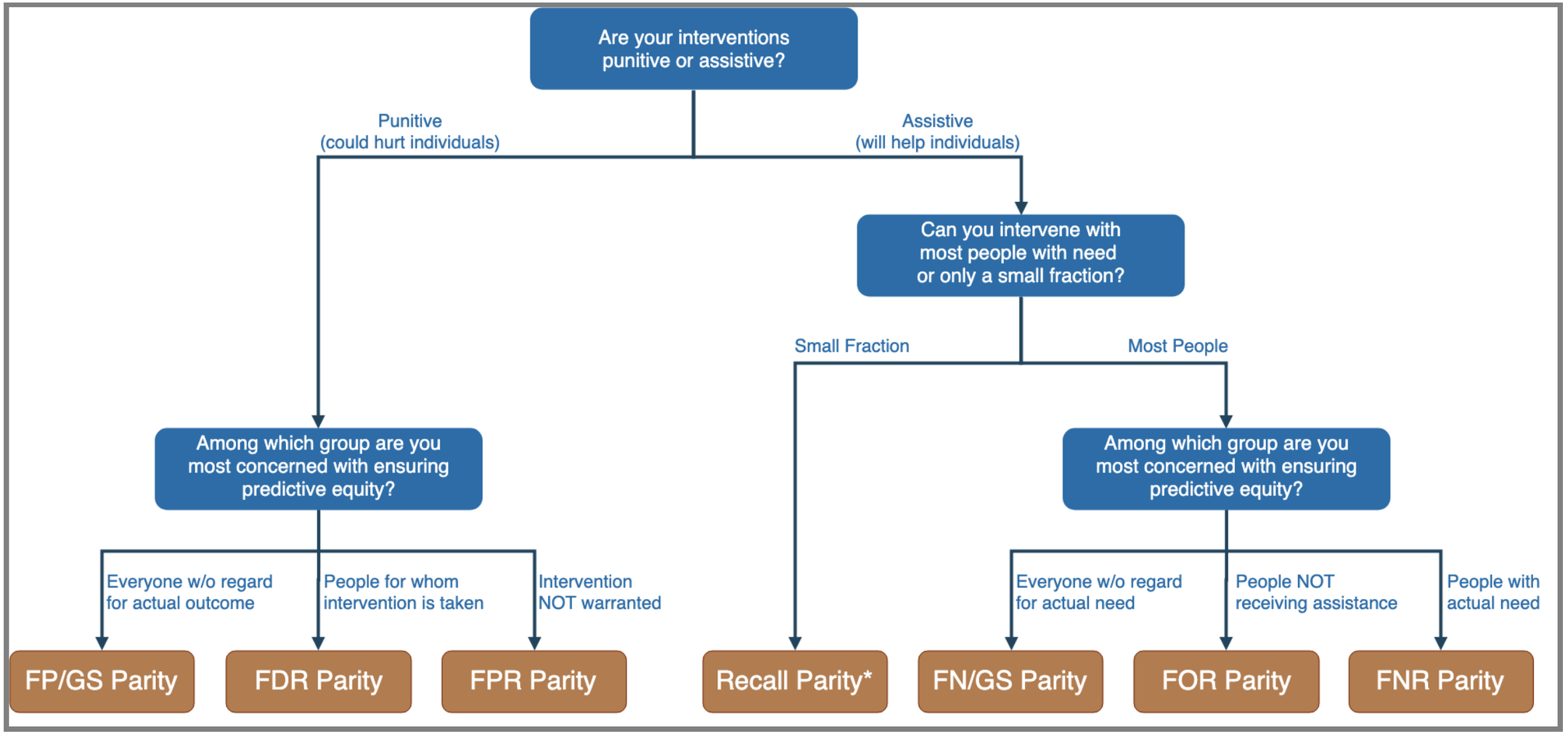
Overcome blind spots within teams

- Systematically consider consequences of bias
- Consider safety engineering techniques (e.g., FTA)
- Assemble diverse teams, use personas, crowdsource audits

Some Guidance on Equality Metric:

Are the interventions punitive or assistive

- Punitive (could hurt individuals): Focus on similar false positive rates
- Assistive (will help individuals): Focus on similar recall, false negative rates



Ian Foster, Rayid Ghani, Ron S. Jarmin, Frauke Kreuter and Julia Lane. Big Data and Social Science: Data Science Methods and Tools for Research and Practice. Chapter 11, 2nd ed, 2020

Identify Protected Attributes

Against which groups might we discriminate? What attributes identify them directly or indirectly?

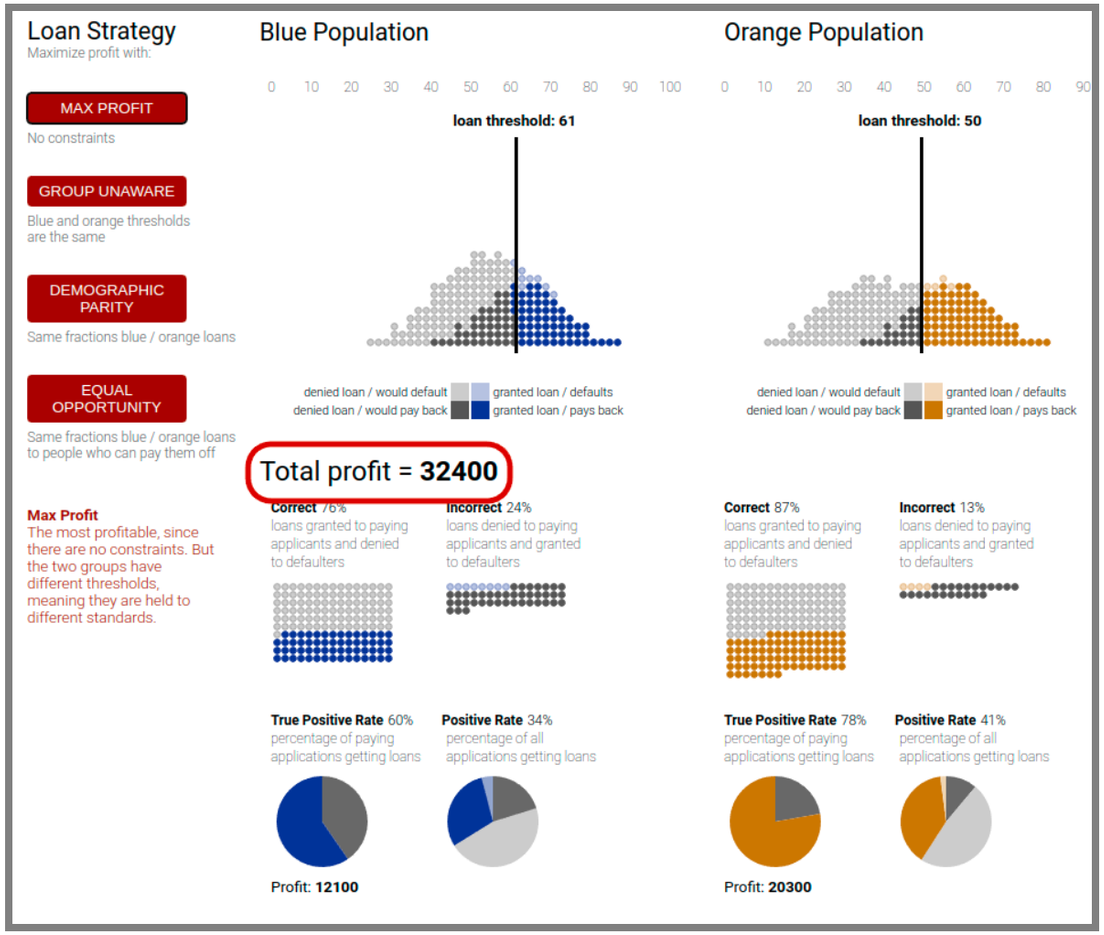
Requires understanding of target population and subpopulations

Use anti-discrimination law as starting point, but do not end there

- Socio-economic status? Body height? Weight? Hair style? Eye color? Sports team preferences?
- Protected attributes for non-humans? Animals, inanimate objects?

Involve stakeholders, consult lawyers, read research, ask experts, ...

Fairness, Accuracy, and Profits



Interactive visualization: https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Fairness, Accuracy, and Profits

Fairness can conflict with accuracy goals

Fairness can conflict with organizational goals (profits, usability)

Fairer products may attract more customers

Unfair products may receive bad press, reputation damage

Improving fairness through better data can benefit everybody

Trade-offs in Fairness vs Accuracy

General view: Accuracy is at odds with fairness (e.g., impossible to achieve perfect accuracy $R = Y$ while ensuring group fairness)

Fairness imposes constraints, limits what models can be learned

But: Arguably unfair predictions not desirable, accuracy based on misleading ground truth

Determine how much compromise in accuracy or fairness is acceptable to your stakeholders; is accuracy the right measure or based on the right data?

Discussion: Fairness Goal for Mortgage Applications?

Discussion: Fairness Goal for Mortgage Applications?

Disparate impact considerations seem to prevail -- group fairness

Need to justify strong differences in outcomes

Can also sue over disparate treatment if bank indicates that protected attribute was reason for decision

Discussion: Fairness Goal for College Admission?

Discussion: Fairness Goal for College Admission?

Strong legal precedents

Very limited scope of *affirmative action*

Most forms of group fairness likely illegal

In practice: Anti-classification

Discussion: Fairness Goal for Hiring Decisions?

Law: "Four-fifth rule" (or "80% rule")

- Group fairness with a threshold: $\frac{P[R=1|A=a]}{P[R=1|A=b]} \geq 0.8$
- 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?

Discussion: Fairness Goal for Cancer Prognosis?

Discussion: Fairness Goal for Recidivism Prediction?

- ProPublica article

Discussion: Recidivism Prediction?

- ProPublica investigation: COMPAS violates separation w/ FPR & FNR
- Northpointe response: COMPAS is fair because it has similar FDRs across both races
- *Is COMPAS both fair & unfair at the same time? Which definition is the "right" one?*

Figure from Big Data and Social Science, Ch. 11

Identifying and Negotiating Fairness Requirements

Measuring is easy, but what to measure?

Identifying Fairness Goals is a Requirements Engineering Problem

- What is the goal of the system? What benefits does it provide and to whom?
- What subpopulations (including minority groups) may be using or be affected by the system? What types of harms can the system cause with discrimination?
- Who are the stakeholders of the system? What are the stakeholders' views or expectations on fairness and where do they conflict? Are we trying to achieve fairness based on equality or equity?
- Does fairness undermine any other goals of the system (e.g., accuracy, profits, time to release)?
- Are there legal anti-discrimination requirements to consider? Are there societal expectations about ethics that relate to this product? What is the activist position?
- ...

Analyzing Potential Harms

Anticipate harms from unfair decisions

- Harms of allocation, harms of representation?
- How do biased model predictions contribute to system behavior? (show predictions, act on predictions?)

Consider how automation can amplify harm

Overcome blind spots within teams

- Systematically consider consequences of bias
- Consider safety engineering techniques (e.g., FTA)
- Assemble diverse teams, use personas, crowdsource audits

Example: Harms in Biased College Admission Screening

What can we do beyond brainstorming?

Example: Judgment Call Game

Card "Game" by Microsoft Research

Participants write "Product reviews" from different perspectives

- encourage thinking about consequences
- enforce persona-like role taking

Identify Protected Attributes

Against which groups might we discriminate? What attributes identify them directly or indirectly?

Requires understanding of target population and subpopulations

Use anti-discrimination law as starting point, but do not end there

- Socio-economic status? Body height? Weight? Hair style? Eye color? Sports team preferences?
- Protected attributes for non-humans? Animals, inanimate objects?

Involve stakeholders, consult lawyers, read research, ask experts, ...

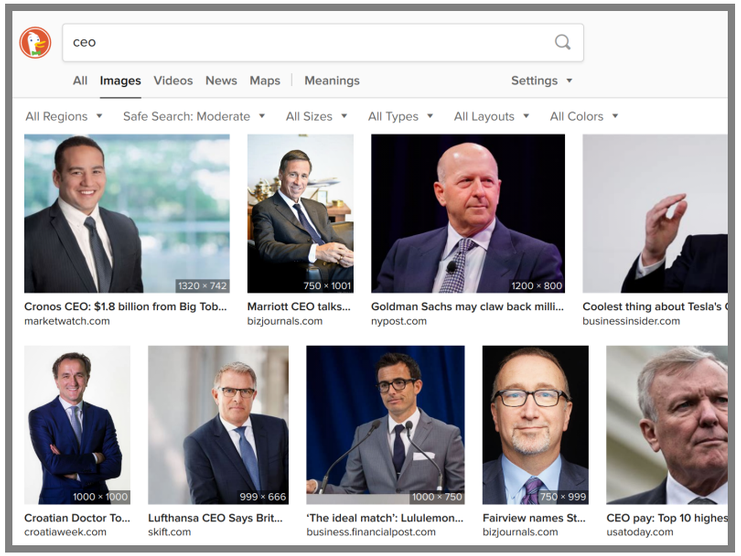
Negotiate Fairness Goals/Measures

Equality or equity? Equalized odds? ...

Cannot satisfy all. People have conflicting preferences...

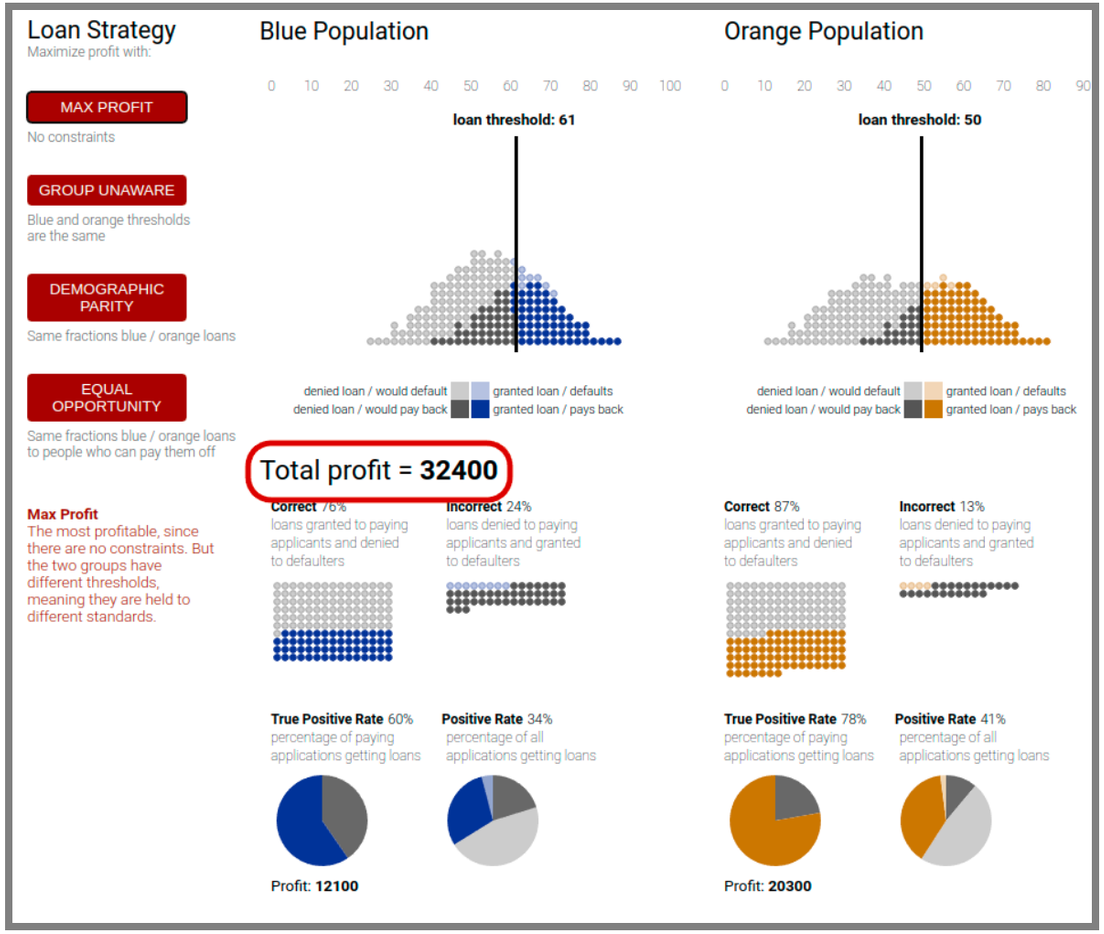
> *Treating everybody equally in a meritocracy will reinforce existing inequalities whereas uplifting disadvantaged communities can be seen as giving unfair advantages to people who contributed less, making it harder to succeed in the advantaged group merely due to group status.*

Recall: CEOs in Image Search



"Through user studies, the [image search] team learned that many users were uncomfortable with the idea of the company "manipulating" search results, viewing this behavior as unethical." -- observation from interviews by Ken Holstein

Fairness, Accuracy, and Profits



Interactive visualization: https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Fairness, Accuracy, and Profits

Fairness can conflict with accuracy goals

Fairness can conflict with organizational goals (profits, usability)

Fairer products may attract more customers

Unfair products may receive bad press, reputation damage

Improving fairness through better data can benefit everybody

Negotiate Fairness Goals/Measures

Negotiation with tradeoffs, inherently political, weigh/balance preferences

Will need to accept some (perceived) unfairness

Power structures often influence outcomes

- Product owners can often drive decisions
- Legal requirements pose constraints
- Users and activists and press can create pressure

Just like other requirements negotiation:

- Consider design space, expose tradeoffs explicitly
- Somebody will need to make a decision, often project owner
- Document decision with justification

Societal Implications

Automation at scale can shift power dynamics at scale

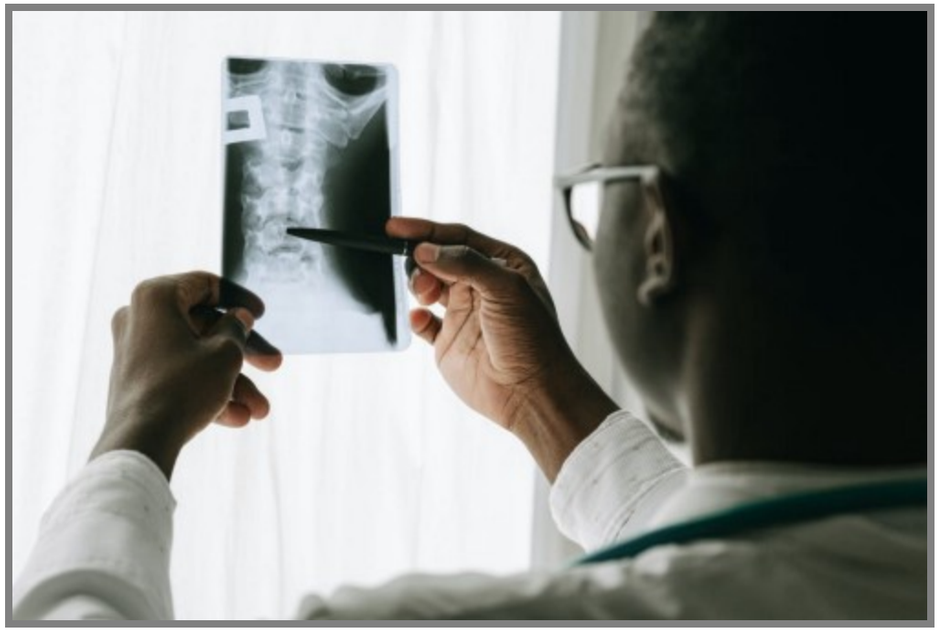
- Path for social good or path into dystopia?
- Who benefits from ML-based automation? Who bears the cost?

Making Rare Skills Attainable

Reduce reliance on specialized training, improve access, improve cost

Examples?

Making Rare Skills Attainable



We should stop training radiologists now. It's just completely obvious that within five years, deep learning is going to do better than radiologists. -- Geoffrey Hinton, 2016

Making Rare Skills Attainable

Examples:

- Healthcare in rural settings, developing countries
- Generative models for Art (DALL·E, stable diffusion)
- Navigation tools (trained taxi license -> Uber)

Making Rare Skills Attainable, but…

Downsides?

Making Rare Skills Attainable, but...

Displacing high-skilled jobs

Low skilled, machine-directed jobs, "algorithmic management"

Who owns the ML-enabled products? Rent-seeking economies?

Society without relying on work? 14h work week? Automation dividend? Universal basic income? "Fully automated luxury communism"

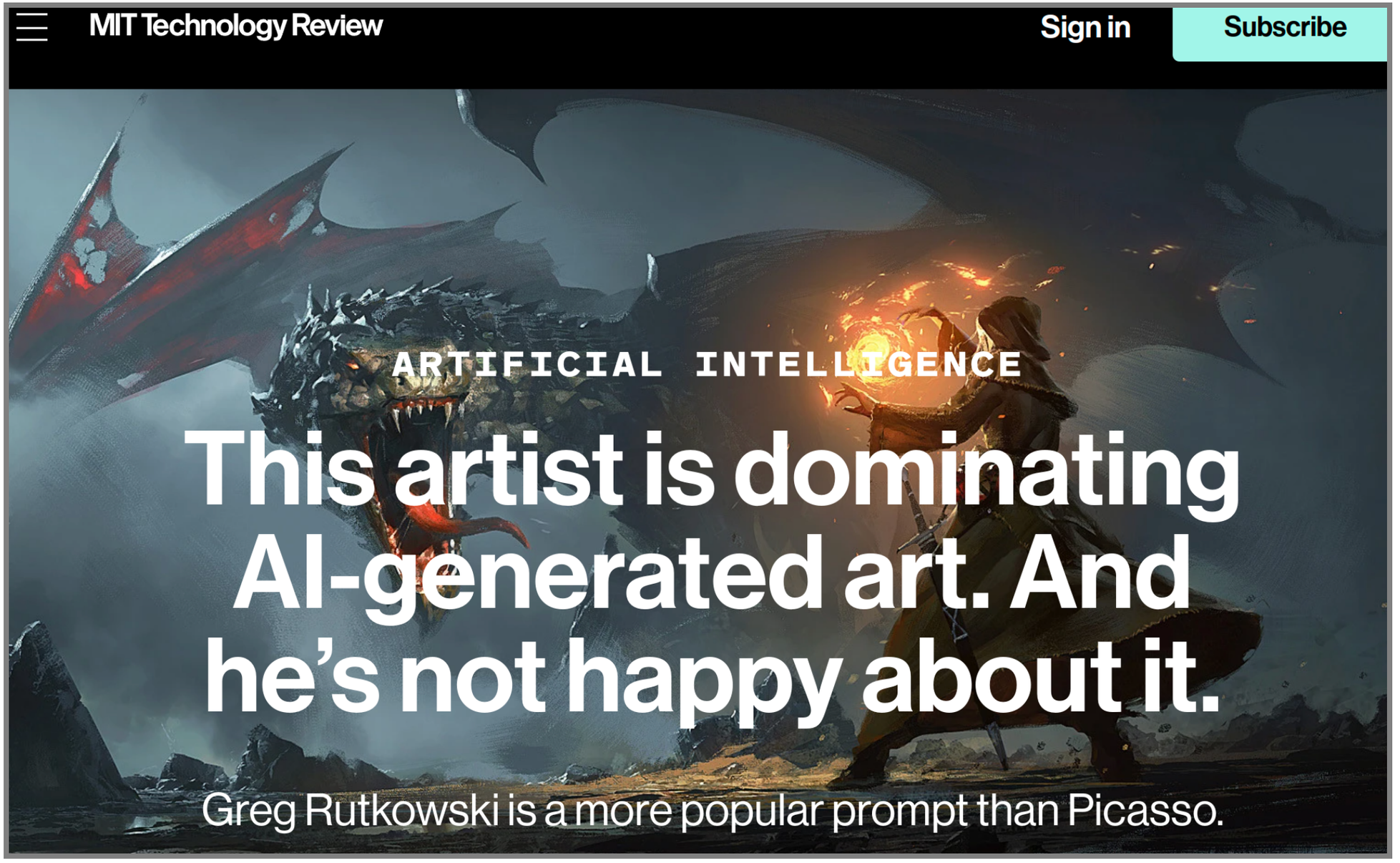
Making Rare Skills Attainable, but...

Who owns the algorithms?

- DALL·E: Corporate control, API only
- Stable diffusion: open source, CreativeML Open RAIL-M license ("ethical license")

Exploitative Data Collection

Problems?


MIT Technology Review
Sign in
Subscribe
ARTIFICIAL INTELLIGENCE
This artist is dominating AI-generated art. And he's not happy about it.
Greg Rutkowski is a more popular prompt than Picasso.

Exploitative Data Collection

Scraping public data, without compensation of creators, ignoring licenses

Labeling often crowd sourced at poverty wages

Data entry often assigned to field workers (e.g., nurses) in addition to existing tasks

Data workers may not benefit from system, are often not valued, are often manipulated through surveillance and gamification mechanisms

Exploitative Data Collection

Who owns the data? Who does the data work?

Who owns the model or product? Who owns their outputs?

Who benefits?

What are fair working conditions?

Who does the Fairness Work?

Who does the Fairness Work?

Within organizations usually little institutional support for fairness work, few activists

Fairness issues often raised by communities affected, after harm occurred

Affected groups may need to organize to affect change

Do we place the cost of unfair systems on those already marginalized and disadvantaged?

Breakout: College Admission

Assume most universities want to automate admissions decisions.

As a group in `#lecture`, tagging group members:

> *What good or bad societal implications can you anticipate, beyond a single product? Should we do something about it?*

Fairness beyond the Model

Bias Mitigation through System Design

Examples of mitigations around the model?

1. Avoid Unnecessary Distinctions

Image captioning gender biased?

1. Avoid Unnecessary Distinctions

"Doctor/nurse applying blood pressure monitor" -> "Healthcare worker applying blood pressure monitor"

1. Avoid Unnecessary Distinctions

Is the distinction actually necessary? Is there a more general class to unify them?

Aligns with notion of *justice* to remove the problem from the system

2. Suppress Potentially Problem Outputs

How to fix?

2. Suppress Potentially Problem Outputs

Anticipate problems or react to reports

Postprocessing, filtering, safeguards

- Suppress entire output classes
- Hardcoded rules or other models (e.g., toxicity detection)

May degrade system quality for some use cases

See mitigating mistakes generally

3. Design Fail-Soft Strategy

Example: Plagiarism detector

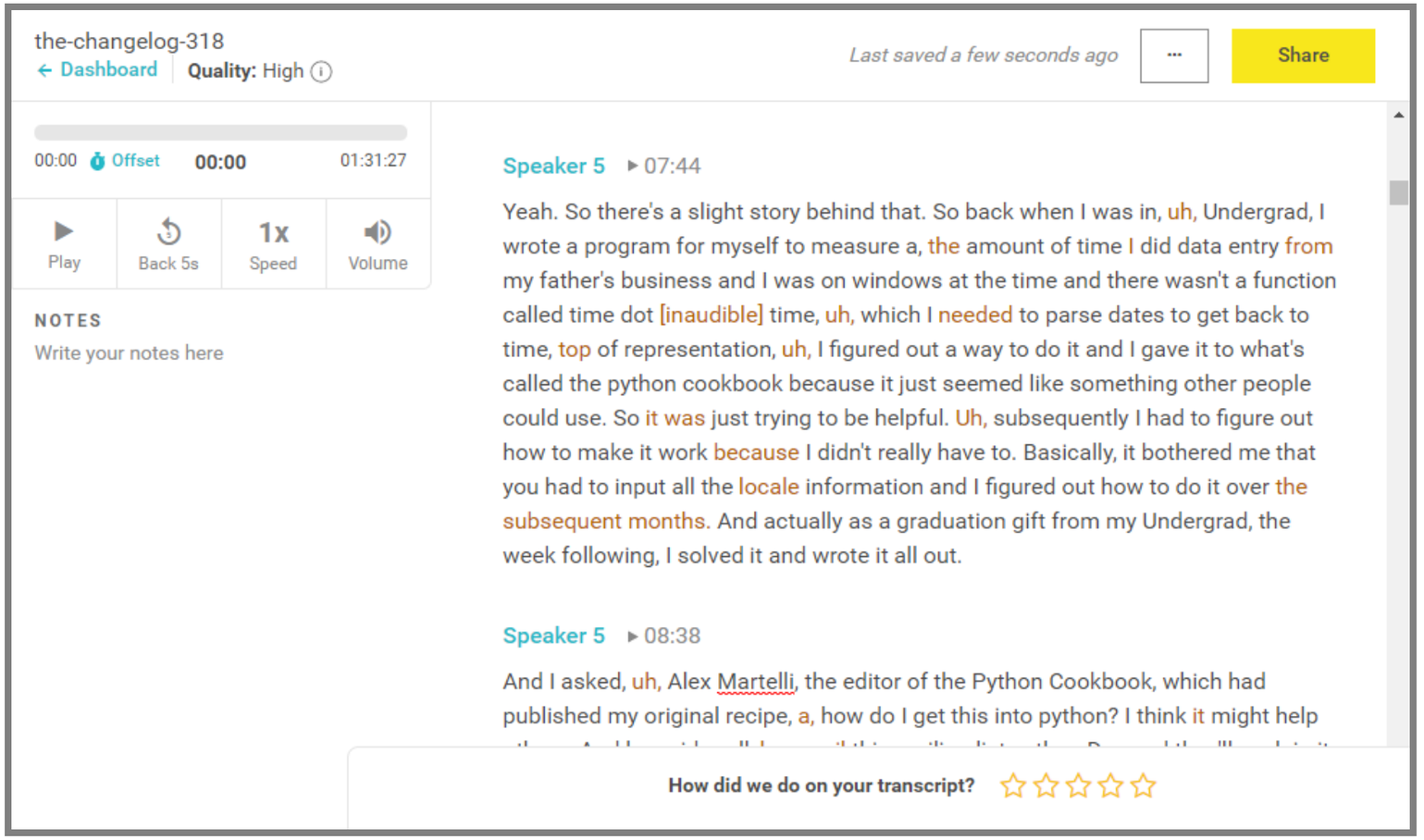
A: Cheating detected! This incident has been reported.

B: This answer seems to perfect. Would you like another exercise?

HCI principle: Fail-soft interfaces avoid calling out directly; communicate friendly and constructively to allow saving face

Especially relevant if system unreliable or biased

4. Keep Humans in the Loop



TV subtitles: Humans check transcripts, especially with heavy dialects

4. Keep Humans in the Loop

Recall: Automate vs prompt vs augment

Involve humans to correct for mistakes and bias

But, model often introduced to avoid bias in human decision

But, challenging human-interaction design to keep humans engaged and alert; human monitors possibly biased too, making it worse

Does a human have a fair chance to detect and correct bias? Enough information? Enough context? Enough time? Unbiased human decision?

Predictive Policing Example

> *"officers expressed skepticism about the software and during ride alongs showed no intention of using it"*

> *"the officer discounted the software since it showed what he already knew, while he ignored those predictions that he did not understand"*

Does the system just lend credibility to a biased human process?

Lally, Nick. "“It makes almost no difference which algorithm you use”: on the modularity of predictive policing." Urban Geography (2021): 1-19.

Fairer Data Collection

Data Collection is Amendable

Data science education often assumes data as given

In industry, we often have control over data collection, curation, labeling (65% in Holstein et al.)

Most address fairness issues by collecting more data (73%)

Challenges of incorporating algorithmic fairness into practice, FAT* Tutorial, 2019 (slides)

Fairer Data Collection

Often high-leverage point to improve fairness

"Raw data is an oxymoron"

Fairer Data Collection

Carefully review data collection procedures, sampling biases, what data is collected, how trustworthy labels are, etc.

Can address most sources of bias: tainted labels, skewed samples, limited features, sample size disparity, proxies:

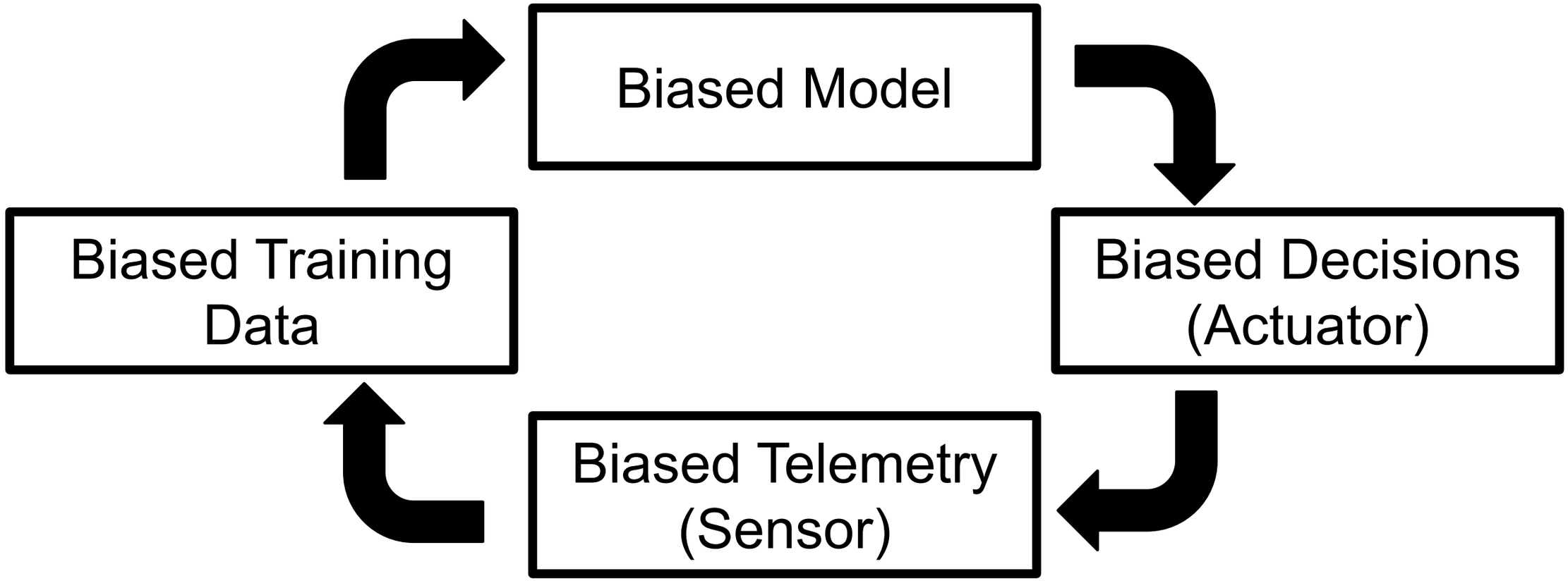
- deliberate what data to collect
- collect more data, oversample where needed
- extra effort in unbiased labels

-> Requirements engineering, system engineering

-> World vs machine, data quality, data cascades

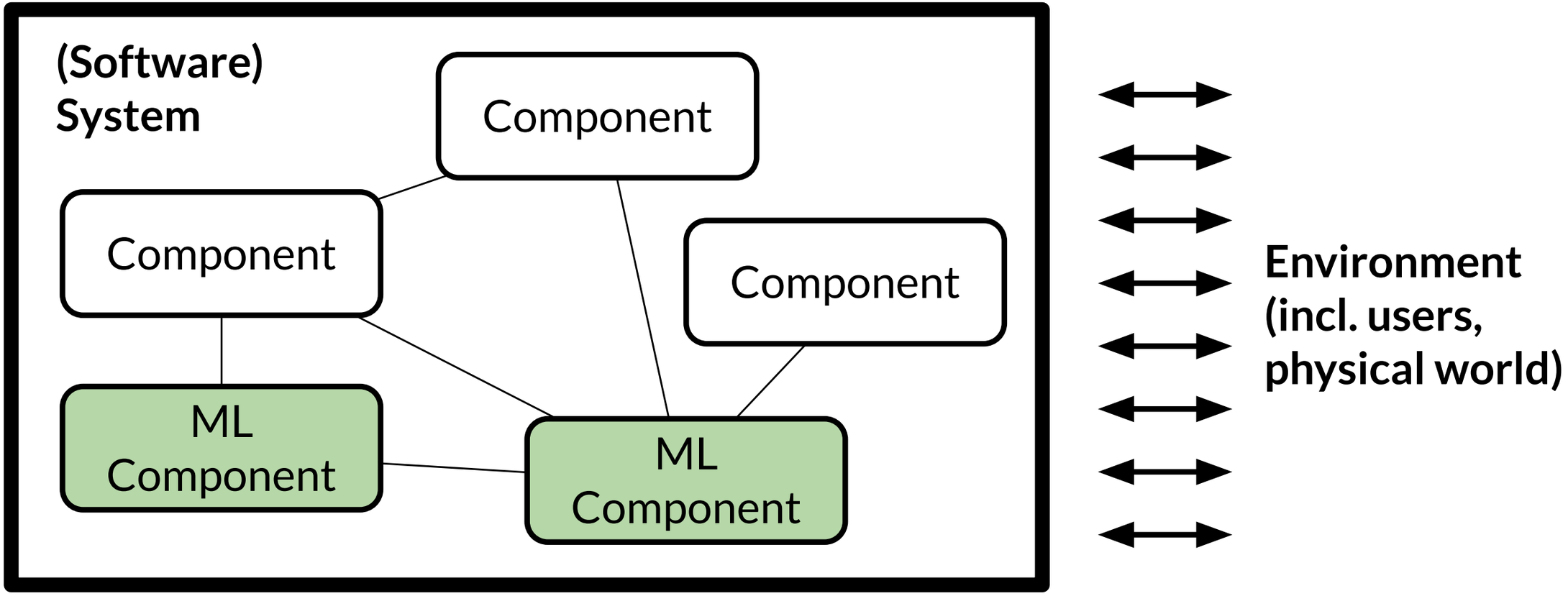
Anticipate Feedback Loops

Feedback Loops

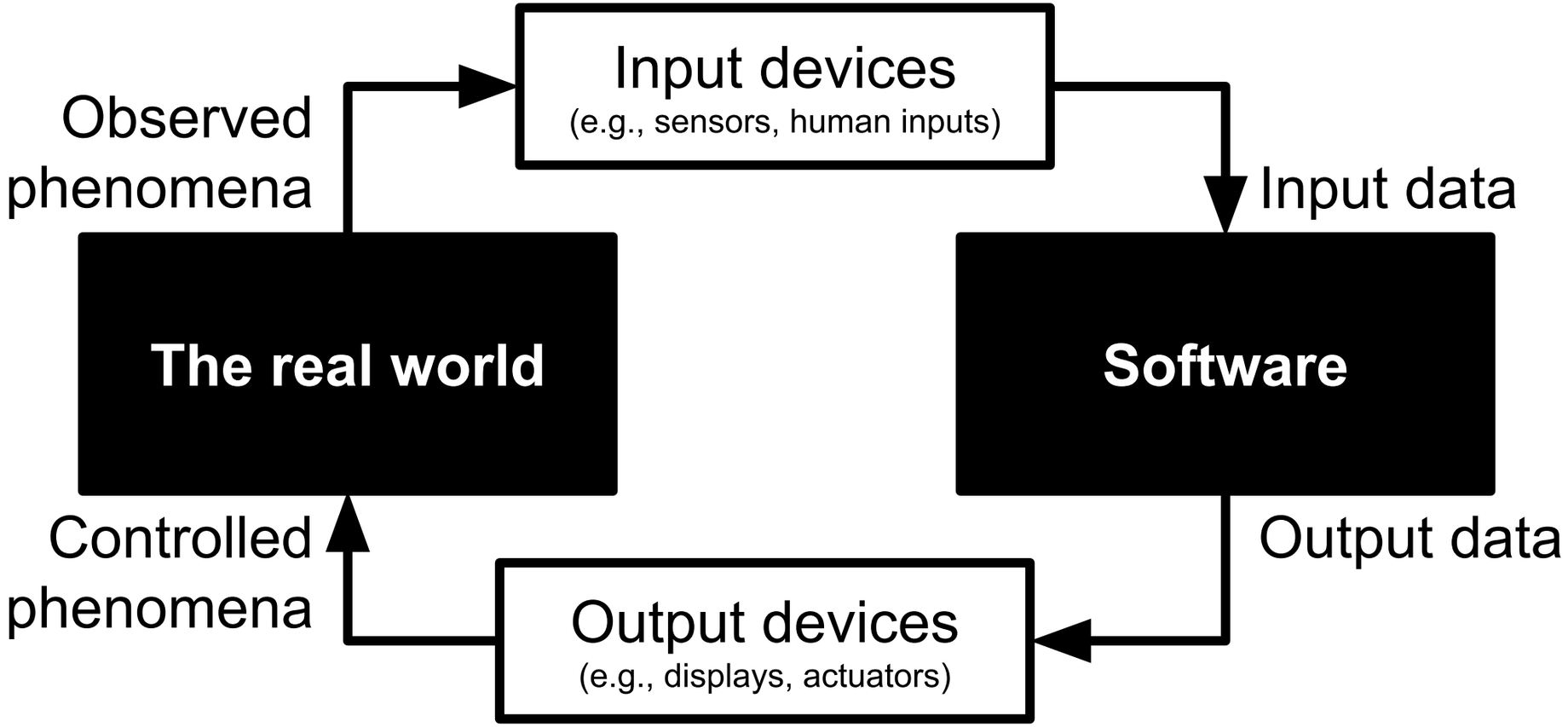


Feedback Loops in Mortgage Applications?

Feedback Loops go through the Environment



Analyze the World vs the Machine



State and check assumptions!

Analyze the World vs the Machine

How do outputs affect change in the real world, how does this (indirectly) influence inputs?

Can we decouple inputs from outputs? Can telemetry be trusted?

Interventions through system (re)design:

- Focus data collection on less influenced inputs
- Compensate for bias from feedback loops in ML pipeline
- Do not build the system in the first place

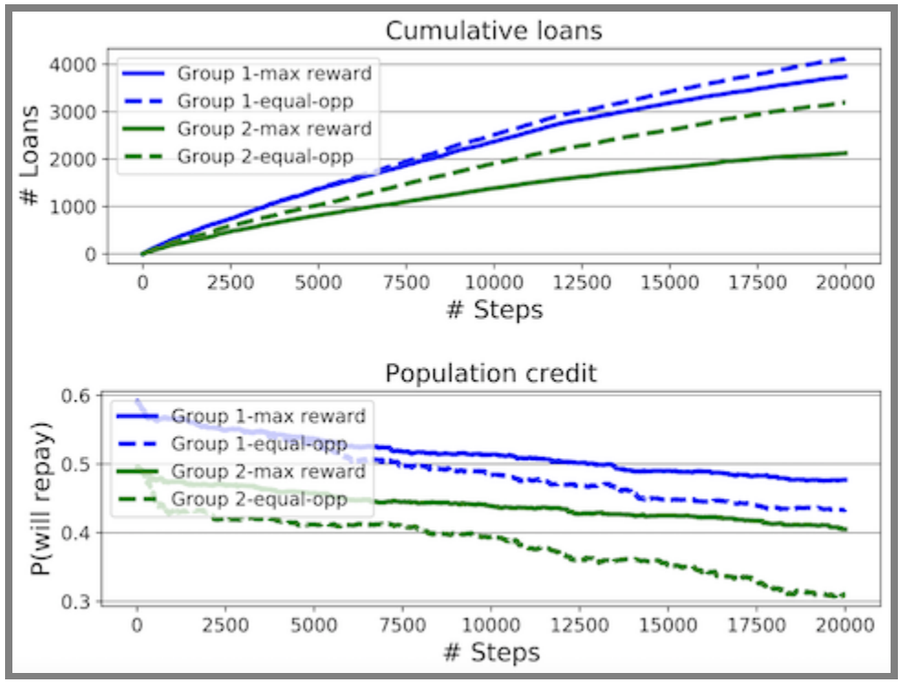
Long-term Impact of ML

- 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 mortgage 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.

Prepare for Feedback Loops

We will likely not anticipate all feedback loops...

... but we can anticipate that unknown feedback loops exist

-> Monitoring!

Process Integration

Fairness in Practice today

Lots of attention in academia and media

Lofty statements by big companies, mostly aspirational

Strong push by few invested engineers (internal activists)

Some dedicated teams, mostly in Big Tech, mostly research focused

Little institutional support, no broad practices

Barriers to Fairness Work

Barriers to Fairness Work

1. Rarely an organizational priority, mostly reactive (media pressure, regulators)
- Limited resources for proactive work
- Fairness work rarely required as deliverable, low priority, ignorable
- No accountability for actually completing fairness work, unclear responsibilities

What to do?

Barriers to Fairness Work

2. Fairness work seen as ambiguous and too complicated for available resources (esp. outside Big Tech)
- Academic discussions and metrics too removed from real problems
- Fairness research evolves too fast
- Media attention keeps shifting, cannot keep up
- Too political

What to do?

Barriers to Fairness Work

3. Most fairness work done by volunteers outside official job functions
- Rarely rewarded in performance evaluations, promotions
- Activists seen as troublemakers
- Reliance on personal networks among interested parties

What to do?

Barriers to Fairness Work

4. Impact of fairness work difficult to quantify, making it hard to justify resource investment

- Does it improve sales? Did it avoid PR disaster? Missing counterfactuals
- Fairness rarely monitored over time
- Fairness rarely a key performance indicator of product
- Fairness requires long-term perspective (feedback loops, rare disasters), but organizations focus on short-term goals

What to do?

Barriers to Fairness Work

5. Technical challenges
   - Data privacy policies restrict data access for fairness analysis
   - Bureaucracy
   - Distinguishing unimportant user complains from systemic bias issues, debugging bias issues
6. Fairness concerns are project specific, hard to transfer actionable insights and tools across teams

What to do?

Improving Process Integration -- Aspirations

Integrate proactive practices in development processes -- both model and system level!

Move from individuals to institutional processes distributing the work

Hold the entire organization accountable for taking fairness seriously

How?

Improving Process Integration -- Examples

1. Mandatory discussion of discrimination risks, protected attributes, and fairness goals in *requirements documents*
2. Required fairness reporting in addition to accuracy in automated *model evaluation*
3. Required internal/external fairness audit before *release*
4. Required fairness monitoring, oversight infrastructure in *operation*

Improving Process Integration -- Examples

5. Instituting fairness measures as *key performance indicators* of products
6. Assign clear responsibilities of who does what
7. Identify measurable fairness improvements, recognize in performance evaluations

How to avoid pushback against bureaucracy?

Affect Culture Change

Buy-in from management is crucial

Show that fairness work is taken seriously through action (funding, hiring, audits, policies), not just lofty mission statements

Reported success strategies:

1. Frame fairness work as financial profitable, avoiding rework and reputation cost
2. Demonstrate concrete, quantified evidence of benefits of fairness work
3. Continuous internal activism and education initiatives
4. External pressure from customers and regulators

Assigning Responsibilities

Hire/educate T-shaped professionals

Have dedicated fairness expert(s) consulting with teams, performing/guiding audits, etc

Not everybody will be a fairness expert, but ensure base-level awareness on when to seek help

Aspirations

> *"They imagined that organizational leadership would understand, support, and engage deeply with responsible AI concerns, which would be contextualized within their organizational context. Responsible AI would be prioritized as part of the high-level organizational mission and then translated into actionable goals down at the individual levels through established processes. Respondents wanted the spread of information to go through well-established channels so that people know where to look and how to share information."*

From Rakova, Bogdana, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. "Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational

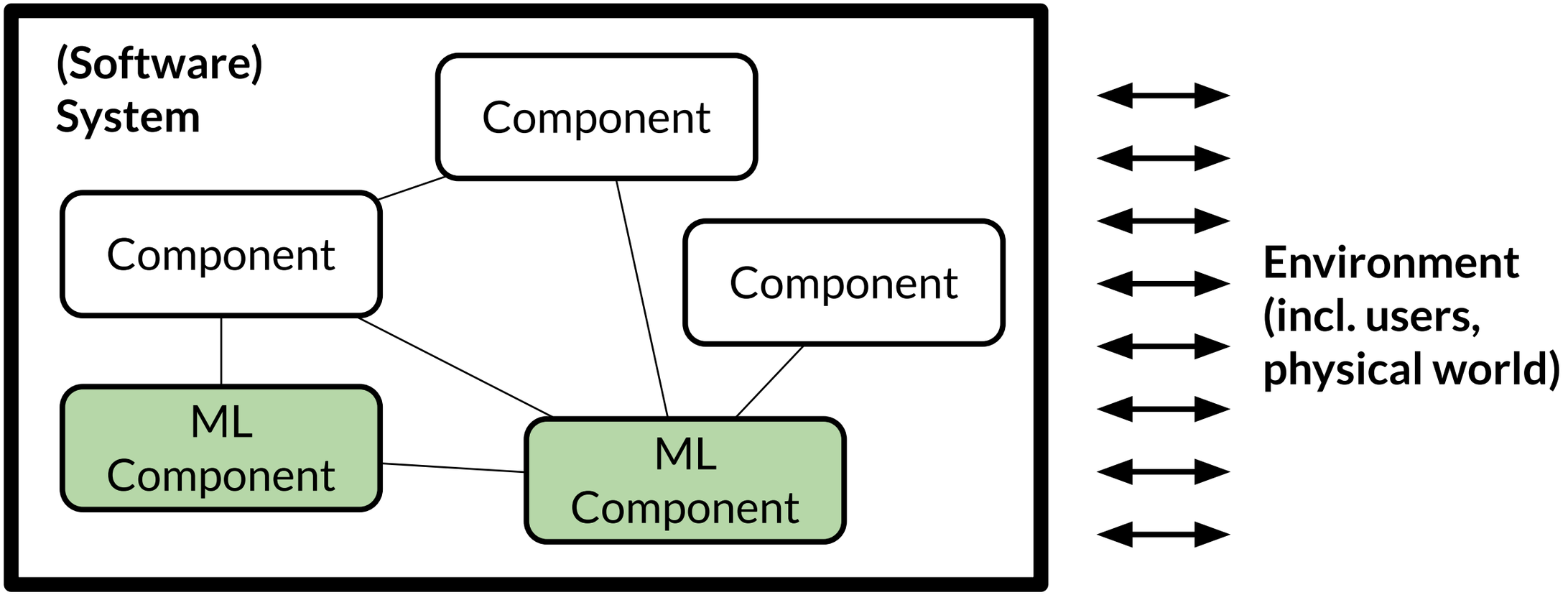
Burnout is a Real Danger

Unsupported fairness work is frustrating and often ineffective

> *"However famous the company is, it's not worth being in a work situation where you don't feel like your entire company, or at least a significant part of your company, is trying to do this with you. Your job is not to be paid lots of money to point out problems. Your job is to help them make their product better. And if you don't believe in the product, then don't work there." -- Rumman Chowdhury via Melissa Heikkilä*

Documenting Fairness at the Interface

Fairness Concerns cut across Components



Product vs model team, product vs model requirements

Fairness Concerns cut across Components

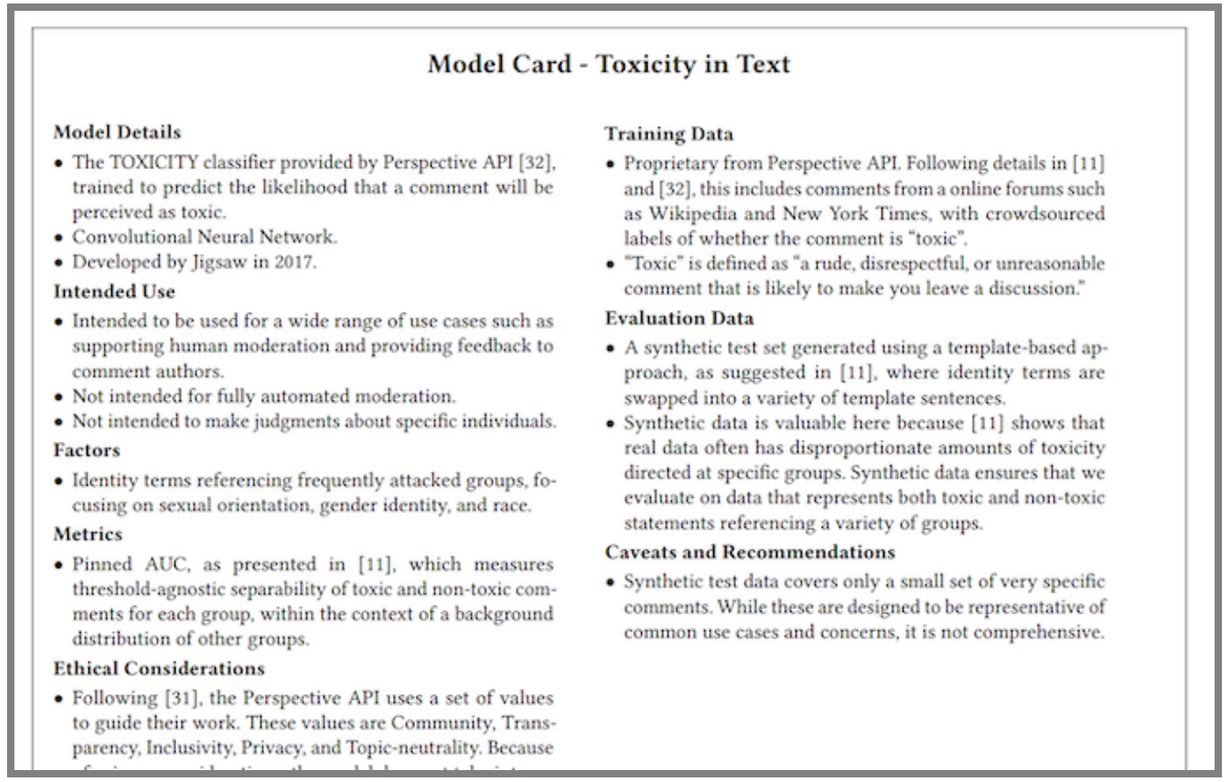
Product vs model team, product vs model requirements

As all design/architecture:

- Identify system-level requirements, break down to component level
- Assign responsibilities
- Document component requirements, provide evidence of results

Documenting Model Fairness

Recall: 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.

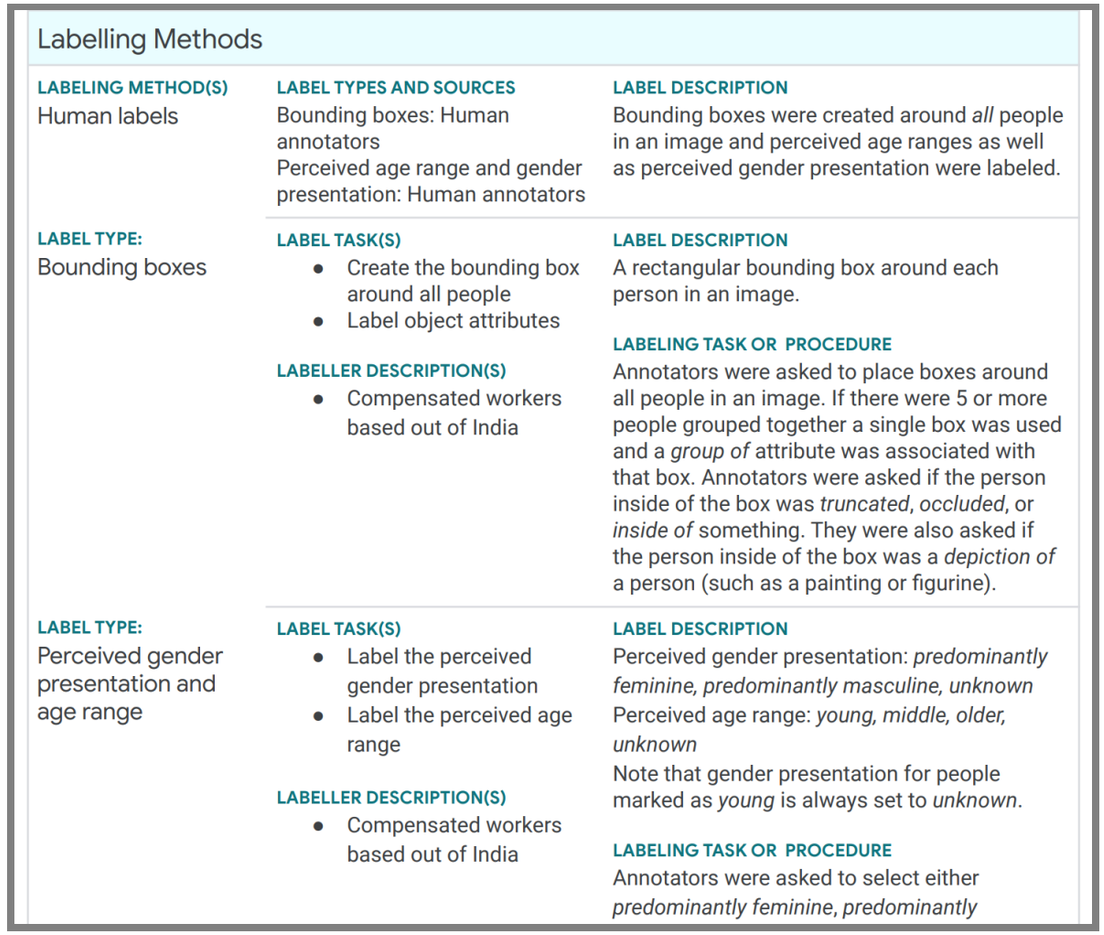
Mitchell, Margaret, et al. "Model cards for model reporting." In Proc. FAccT, 220-229. 2019.

Documenting Fairness of Datasets

Datasheets for Datasets, Dataset Nutrition Labels, ...

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%

Documenting Fairness of Datasets



Labelling Methods

LABELING METHOD(S)
Human labels

LABEL TYPES AND SOURCES
Bounding boxes: Human annotators
Perceived age range and gender presentation: Human annotators

LABEL DESCRIPTION
Bounding boxes were created around *all* people in an image and perceived age ranges as well as perceived gender presentation were labeled.

LABEL TYPE:
Bounding boxes

LABEL TASK(S)
- Create the bounding box around all people
- Label object attributes

LABELLER DESCRIPTION(S)
- Compensated workers based out of India

LABEL DESCRIPTION
A rectangular bounding box around each person in an image.

LABELING TASK OR PROCEDURE
Annotators were asked to place boxes around all people in an image. If there were 5 or more people grouped together a single box was used and a *group of* attribute was associated with that box. Annotators were asked if the person inside of the box was *truncated*, *occluded*, or *inside of* something. They were also asked if the person inside of the box was a *depiction of* a person (such as a painting or figurine).

LABEL TYPE:
Perceived gender presentation and age range

LABEL TASK(S)
- Label the perceived gender presentation
- Label the perceived age range

LABELLER DESCRIPTION(S)
- Compensated workers based out of India

LABEL DESCRIPTION
Perceived gender presentation: *predominantly feminine, predominantly masculine, unknown*
Perceived age range: *young, middle, older, unknown*
Note that gender presentation for people marked as *young* is always set to *unknown*.

LABELING TASK OR PROCEDURE
Annotators were asked to select either *predominantly feminine*, *predominantly*

Excerpt from a "Data Card" for Google's Open Images Extended dataset (full data card)

Monitoring

Monitoring

Operationalize fairness measure in production with telemetry

Monitor like any other metric, use alerts

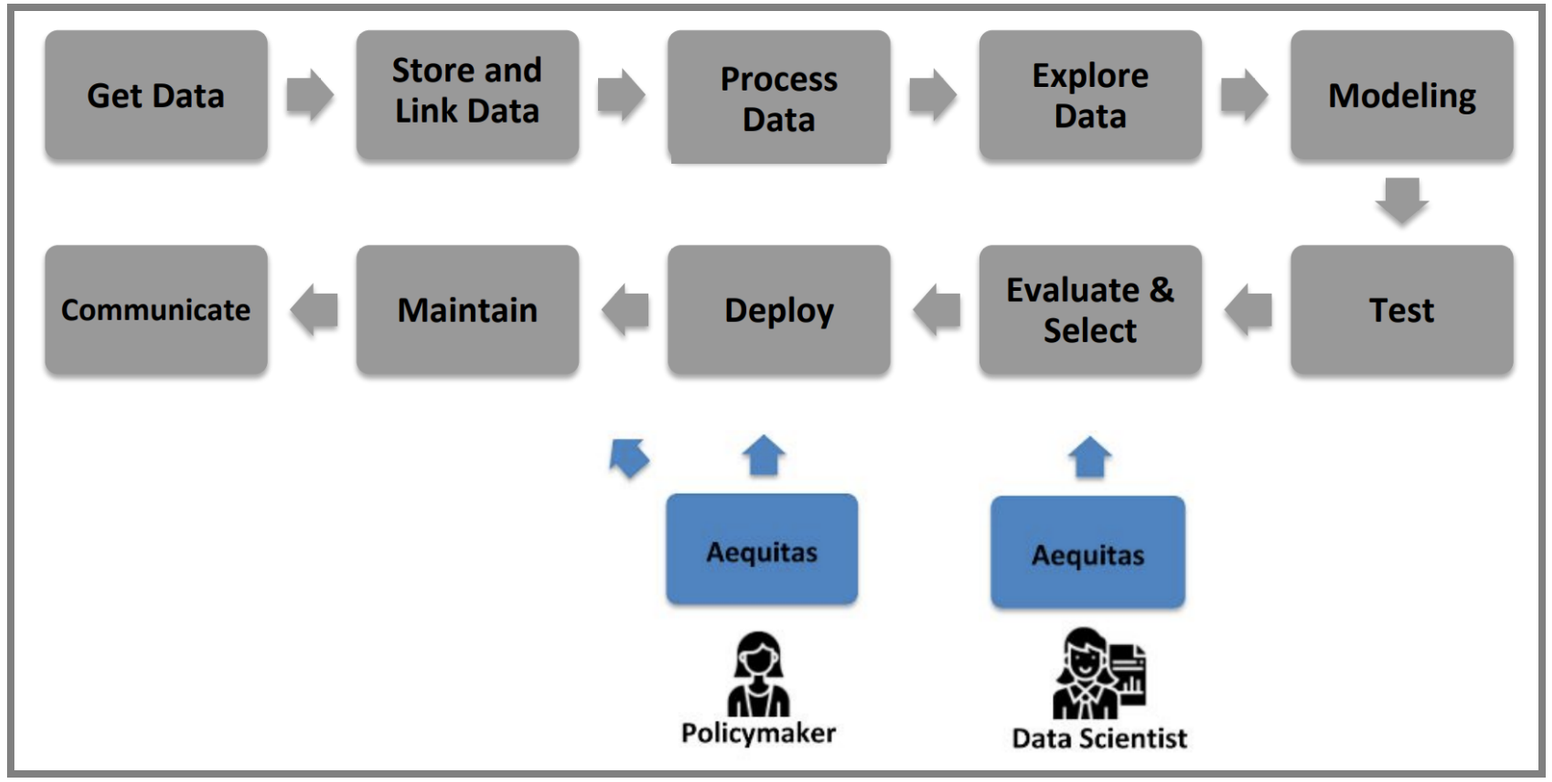
Monitor distribution shifts, especially across protected attributes

Track through experiments, A/B testing etc.

How would you monitor fairness in mortgage applications?

Challenge: Access to protected attributes? Access to ground truth?

Monitoring Tools: Example



(Involve policy makers in the monitoring & auditing process)

http://aequitas.dssg.io/

Preparing for Problems

Provide users with a path to *appeal decisions*

- Provide feedback mechanism to complain about unfairness
- Human review? Human override?

Prepare an *incidence response plan* for fairness issues

- What can be shut down/reverted on short notice?
- Who does what?
- Who talks to the press? To affected parties? What do they need to know?

Best Practices

Best Practices

Best practices are emerging and evolving

Start early, be proactive

Scrutinize data collection and labeling

Invest in requirements engineering and design

Invest in education

Assign clear responsibilities, demonstrate leadership buy-in

Many Tutorials, Checklists, Recommendations

Tutorials (fairness notions, sources of bias, process recom.):

- Fairness in Machine Learning, Fairness-Aware Machine Learning in Practice
- Challenges of Incorporating Algorithmic Fairness into Industry Practice

Checklist:

- Microsoft's AI Fairness Checklist: concrete questions, concrete steps throughout all stages, including deployment and monitoring

Summary

- Requirements engineering for fair ML systems
  - Identify potential harms, protected attributes
  - Negotiate conflicting fairness goals, tradeoffs
  - Consider societal implications
- Design fair systems beyond the model, mitigate bias outside the model
- Anticipate feedback loops
- Integrate fairness work in process and culture
- Document and monitor fairness

Further Readings

- Rakova, Bogdana, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. "Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices." *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW1 (2021): 1-23.
- Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. "Model cards for model reporting." In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220-229. 2019.
- Boyd, Karen L. "Datasheets for Datasets help ML Engineers Notice and Understand Ethical Issues in Training Data." Proceedings of the ACM on Human-Computer Interaction 5, no. CSCW2 (2021): 1-27.
- Bietti, Elettra. "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy." In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 210-219. 2020.
- Madaio, Michael A., Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. "Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI." In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1-14. 2020.
- Hopkins, Aspen, and Serena Booth. "Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development." In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21) (2021).
- Metcalf, Jacob, and Emanuel Moss. "Owning ethics: Corporate logics, silicon valley, and the institutionalization of ethics." *Social Research: An International Quarterly* 86, no. 2 (2019): 449-476.