Spotify

Microsoft Research

A practitioner translation tutorial

Challenges of incorporating algorithmic ‘fairness’ into practice

Who are we?

Henriette Cramer
Spotify

Jenn Wortman Vaughan -
Microsoft Research

Ken Holstein
CMU & Microsoft

Co-organizers

Hanna Wallach
Microsoft Research

Jean Garcia-Gathright
Spotify

Hal Daumé III
Microsoft Research &
University of
Maryland

Miroslav Dudík
Microsoft Research

Sravana Reddy
Spotify

This 90-min tutorial

Intro

Henriette (10 mins)

Decisions while building

Fairness throughout the machine learning lifecycle

Jenn (25 mins)

The wider context

Organizational & domain challenges

Mapping industry challenges

Ken (20 mins)

Case study

Henriette (20 mins)

Recap

Henriette (5 mins)

Research/education communities are growing and becoming more visible ...

WORLD WIDE WEB FOUNDATION

acm

PARTNERSHIP ON AI

Fairness, Accountability, and Transparency in Machine Learning

Data&Society RESEARCH INSTITUTE

AINOW

AI FOR GOOD LAB

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

The Alan Turing Institute

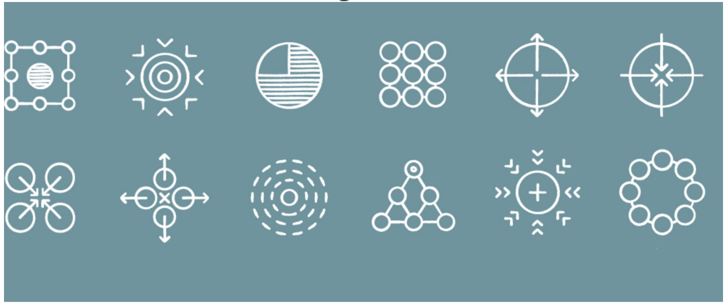
ACM FAT* Conference

BERKMAN KLEIN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

CIDGE

Which has resulted in lots of calls to action …

Data&Society



Algorithmic Accountability: A Primer

Robyn Caplan, Joan Donovan, Lauren Hanson, and Jeanna Matthews

PUBLISHED 04.18.18

Download Report

AI Now Report 2018

Meredith Whittaker, AI Now Institute, New York University, Google Open Research
Kate Crawford, AI Now Institute, New York University, Microsoft Research
Roel Dobbe, AI Now Institute, New York University
Genevieve Fried, AI Now Institute, New York University
Elizabeth Kaziunas, AI Now Institute, New York University
Varoon Mathur, AI Now Institute, New York University
Sarah Myers West, AI Now Institute, New York University
Rashida Richardson, AI Now Institute, New York University
Jason Schultz, AI Now Institute, New York University School of Law
Oscar Schwartz, AI Now Institute, New York University

With research assistance from Alex Campolo and Gretchen Krueger (AI Now Institu
University)

DECEMBER 2018

The Montreal Declaration for a Responsible Development of Artificial Intelligence: a participatory process

SALLE DE PRESSE | 11/03/2017 | UDEMNOUVELLES

REPORT / STUDY | 18 December 2018

Draft Ethics guidelines for trustworthy AI

This working document constitutes a draft of the AI Ethics Guidelines produced by the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG), of which a final version is due in March 2019.

Calls need a how.

Available Examples + Tools.
Encountered Challenges + Gaps.

**Remember [tutorial](#) today ...
about auditing & fairness history

Computational bias literature since (at least) ‘97*

But no standard methods.

*Friedman & Nissenbaum

Doing better

(avoiding harm)

[Shapiro et al., 2017,
Crawford, NeurIPS'17]

More positive outcomes & avoiding harmful outcomes of

algorithms for **groups of people**

Not only:
machine learning

But also:
any **automated system**

Not only:
legally protected classes like gender, race, age

But also:
other societal categories like location, topical interests, (sub)culture etc.

Challenge:
subpopulations may be application-specific, intersectional, subject to complex social constructs

Types of harm

[Shapiro et al., 2017,
Crawford, NeurIPS’17]

Different types of harm

Harms of allocation withhold opportunity or resources

Harms of representation reinforce subordination along the lines of identity, stereotypes

Shapiro et al., 2017

Kate Crawford, “The Trouble With Bias” keynote N(eur)IPS’17

Allocation, incl resources

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

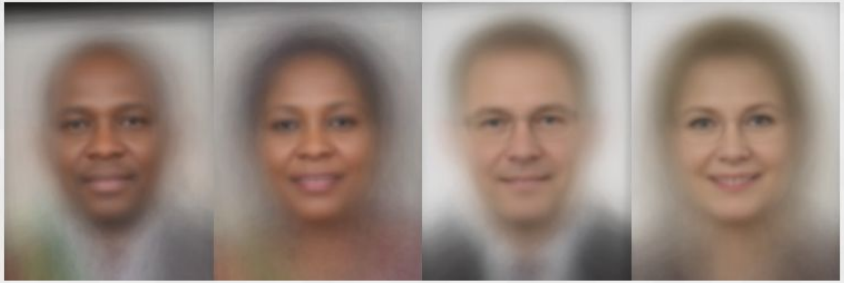
8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

Quality of Service, degraded user experience

@jozjozjoz, 2009 Nikon S630

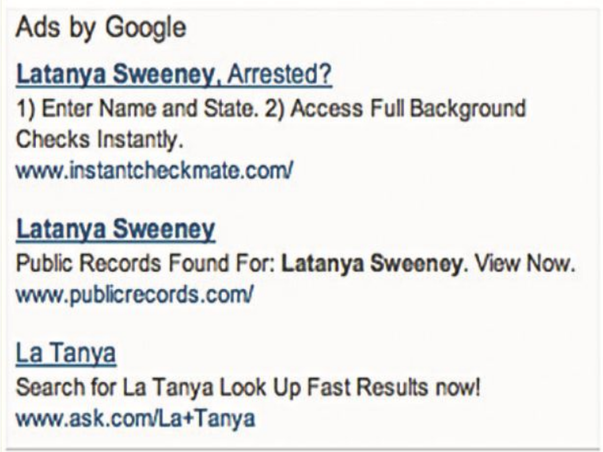
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



[Buolamwini & Gebru, 2018]

Representation
Over/under-representation, stereotyping, denigration

[Kay et al., 2015]



[Sweeney, 2013]

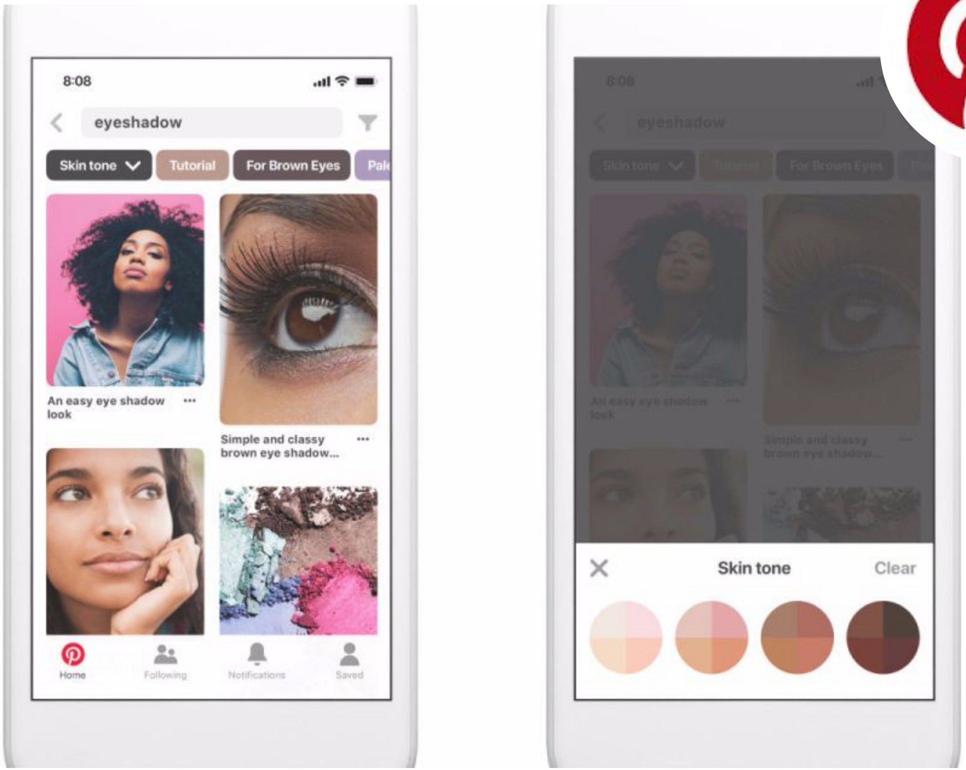
Types of harm can co-occur & need to be specified

	Allocation of resources	Quality of Service	Stereotyping	Denigration	Over- / Under-Representation
Hiring system does not rank women as highly as men for technical jobs	x	x	x		x
Photo management program labels image of black people as "gorillas"		x		x	
Image searches for "CEO" yield only photos of white men on first page			x		x

Why does this matter to practitioners?

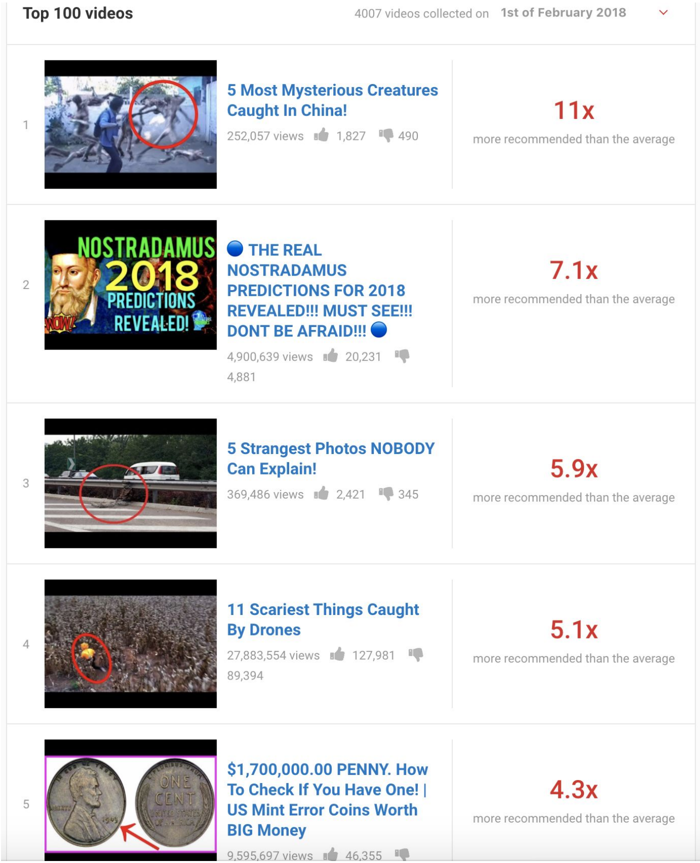
different stakeholders - different arguments

1. Better product, serving wider audience(s)



2. Responsibility, social impact & PR

'Fiction is outperforming reality': how YouTube's algorithm distorts truth



3. Legal & policy

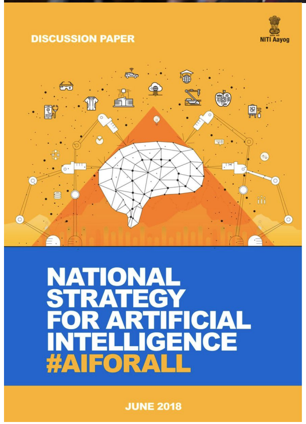
Artificial intelligence: Commission outlines a European approach to boost investment and set ethical guidelines

Brussels, 25 April 2018

MACHINE PERSPECTIVES

Senators are asking whether artificial intelligence could violate US civil rights laws

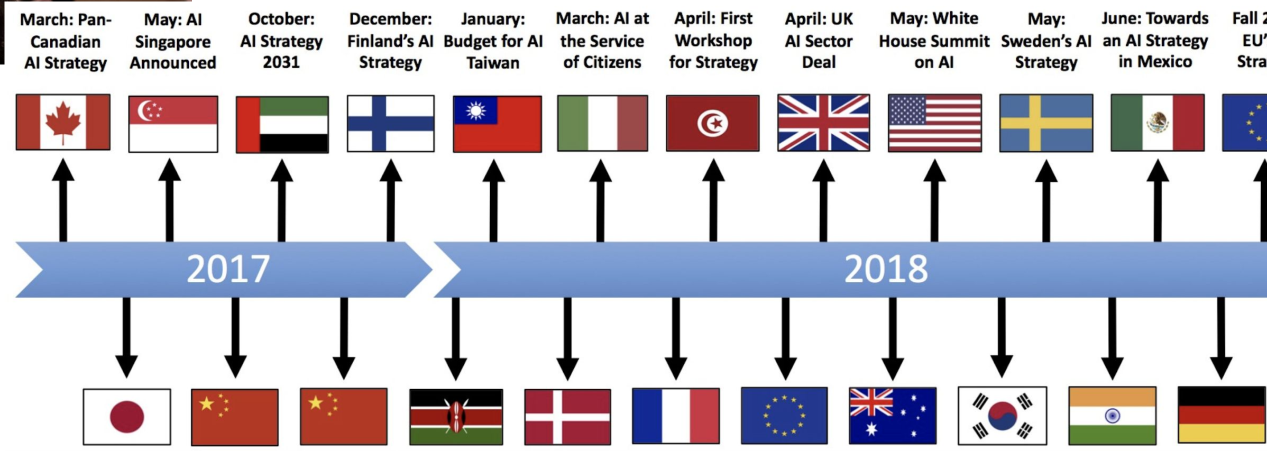
By Dave Gershgorn • September 21, 2018



An Overview of National AI Strategies

Tim Dutton Follow

Jun 28, 2018 · 25 min read



4. Competitive, both proactive & reactive

AI at Google: our principles

We will assess AI applications in view of the following objectives. We believe that AI should:

1. Be socially beneficial.

The expanded reach of new technologies increasingly touches society as a whole. Advances in AI will have transformative impacts in a wide range of fields, including healthcare, security, energy, transportation, manufacturing, and entertainment. As we consider potential development and uses of AI technologies, we will take into account a broad range of social and economic factors, and will proceed where we believe that the overall likely benefits substantially exceed the foreseeable risks and downsides.

AI also enhances our ability to understand the meaning of content at scale. We will strive to make high-quality and accurate information readily available using AI, while continuing to respect cultural, social, and legal norms in the countries where we operate. And we will continue to thoughtfully evaluate when to make our technologies available on a non-commercial basis.

2. Avoid creating or reinforcing unfair bias.

AI algorithms and datasets can reflect, reinforce, or reduce unfair biases. We recognize that distinguishing fair from unfair biases is not always simple, and differs across cultures and societies. We will seek to avoid unjust impacts on people, particularly those related to sensitive characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability, and political or religious belief.

Microsoft | Research Research areas Products & Downloads Programs & Events People Careers More

FATE: Fairness, Accountability, Transparency, and Ethics in AI

Facebook says it has a tool to detect bias in its artificial intelligence

By Dave Gershgorn • May 3, 2018

When you’ve got your stakeholders on board,

there are still practical translation challenges.

Different stakeholders can have different perspectives on ‘fairness’

Decision-maker: of those I’ve labeled high-risk, how many will recidivate?

Predictive value

Defendant: what’s the probability I’ll be incorrectly classified high-risk?

False positive rate

Society [think hiring rather than criminal justice]: is the selected set demographically balanced?

Demography

	Labeled low-risk	Labeled high-risk
Did not recidivate	TN	FP
Recidivated	FN	TP

Arvind Narayanan Tutorial: 21 fairness definitions and their politics

**Remember tutorial today ...
about distinction between ‘bias’ and ‘fairness’

**‘Bias’ and ‘fairness’
are socio-technical
& contested terminology.**

**You don’t model your way
to a fair world.**

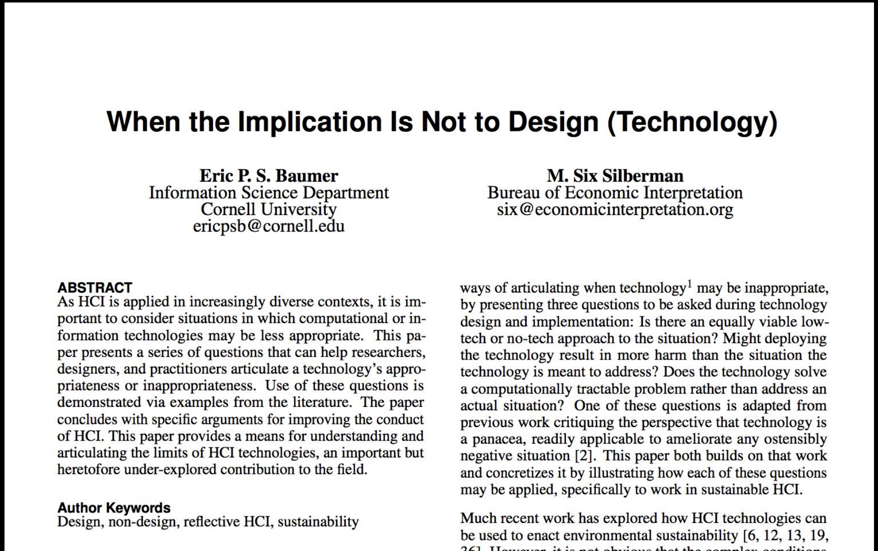
You don’t ‘solve’ this.

The Seductive Diversion of ‘Solving’ Bias in Artificial Intelligence

Trying to "fix" A.I. distracts from the more urgent questions about the technology

Julia Powles Follow
Dec 7, 2018 · 5 min read ★

Not everything should be built.



When the Implication Is Not to Design (Technology)

Eric P. S. Baumer
Information Science Department
Cornell University
ericpsb@cornell.edu

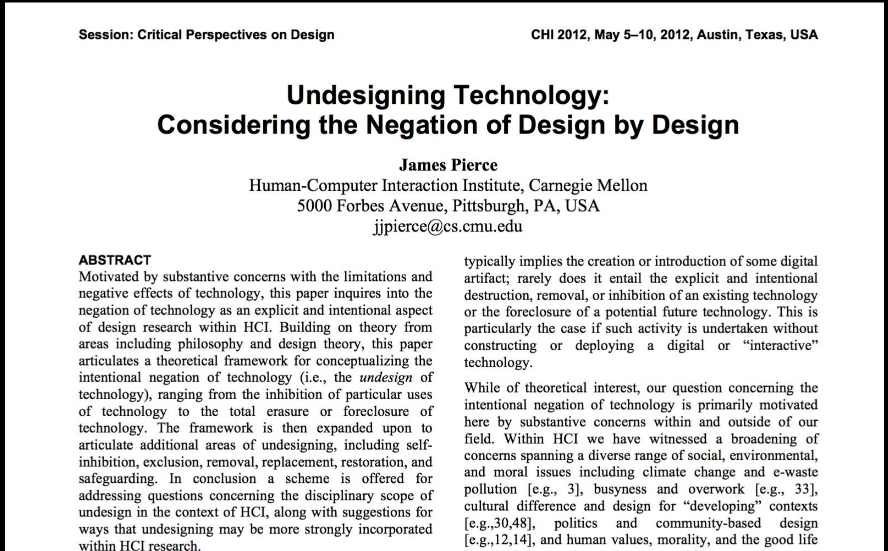
M. Six Silberman
Bureau of Economic Interpretation
six@economicinterpretation.org

ABSTRACT
As HCI is applied in increasingly diverse contexts, it is important to consider situations in which computational or information technologies may be less appropriate. This paper presents a series of questions that can help researchers, designers, and practitioners articulate a technology's appropriateness or inappropriateness. Use of these questions is demonstrated via examples from the literature. The paper concludes with specific arguments for improving the conduct of HCI. This paper provides a means for understanding and articulating the limits of HCI technologies, an important but heretofore under-explored contribution to the field.

Author Keywords
Design, non-design, reflective HCI, sustainability

ways of articulating when technology[1] may be inappropriate, by presenting three questions to be asked during technology design and implementation: Is there an equally viable low-tech or no-tech approach to the situation? Might deploying the technology result in more harm than the situation the technology is meant to address? Does the technology solve a computationally tractable problem rather than address an actual situation? One of these questions is adapted from previous work critiquing the perspective that technology is a panacea, readily applicable to ameliorate any ostensibly negative situation [2]. This paper both builds on that work and concretizes it by illustrating how each of these questions may be applied, specifically to work in sustainable HCI.

Much recent work has explored how HCI technologies can be used to enact environmental sustainability [6, 12, 13, 19,



Undesigning Technology: Considering the Negation of Design by Design

James Pierce
Human-Computer Interaction Institute, Carnegie Mellon
5000 Forbes Avenue, Pittsburgh, PA, USA
jjpierce@cs.cmu.edu

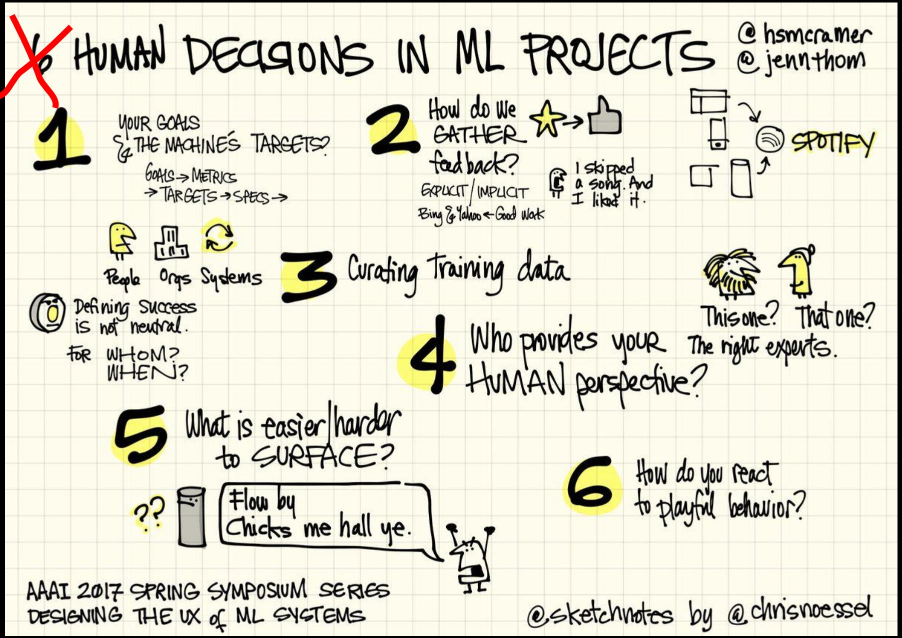
ABSTRACT
Motivated by substantive concerns with the limitations and negative effects of technology, this paper inquires into the negation of technology as an explicit and intentional aspect of design research within HCI. Building on theory from areas including philosophy and design theory, this paper articulates a theoretical framework for conceptualizing the intentional negation of technology (i.e., the *undesign* of technology), ranging from the inhibition of particular uses of technology to the total erasure or foreclosure of technology. The framework is then expanded upon to articulate additional areas of undesigning, including self-inhibition, exclusion, removal, replacement, restoration, and safeguarding. In conclusion a scheme is offered for addressing questions concerning the disciplinary scope of undesign in the context of HCI, along with suggestions for ways that undesigning may be more strongly incorporated within HCI research.

typically implies the creation or introduction of some digital artifact; rarely does it entail the explicit and intentional destruction, removal, or inhibition of an existing technology or the foreclosure of a potential future technology. This is particularly the case if such activity is undertaken without constructing or deploying a digital or "interactive" technology.

While of theoretical interest, our question concerning the intentional negation of technology is primarily motivated here by substantive concerns within and outside of our field. Within HCI we have witnessed a broadening of concerns spanning a diverse range of social, environmental, and moral issues including climate change and e-waste pollution [e.g., 3], busyness and overwork [e.g., 33], cultural difference and design for "developing" contexts [e.g.,30,48], politics and community-based design [e.g.,12,14], and human values, morality, and the good life

Human (non)-decisions need support.

TONS OF



This tutorial:

Machine learning lifecycle

Better decision making
from the start
is easier than fixing things.

But you'll likely join an existing org,
with existing systems.

This tutorial:

Organizational & domain challenges

Mapping industry challenges

A case study

Pragmatic. Imperfect.
Sharing. Learning.

This 90-min tutorial

Decisions while building

The wider context

Intro

Henriette
(10 mins)

Fairness throughout the machine learning lifecycle

Jenn (25 mins)

Organizational & domain challenges

Mapping industry challenges

Ken (20 mins)

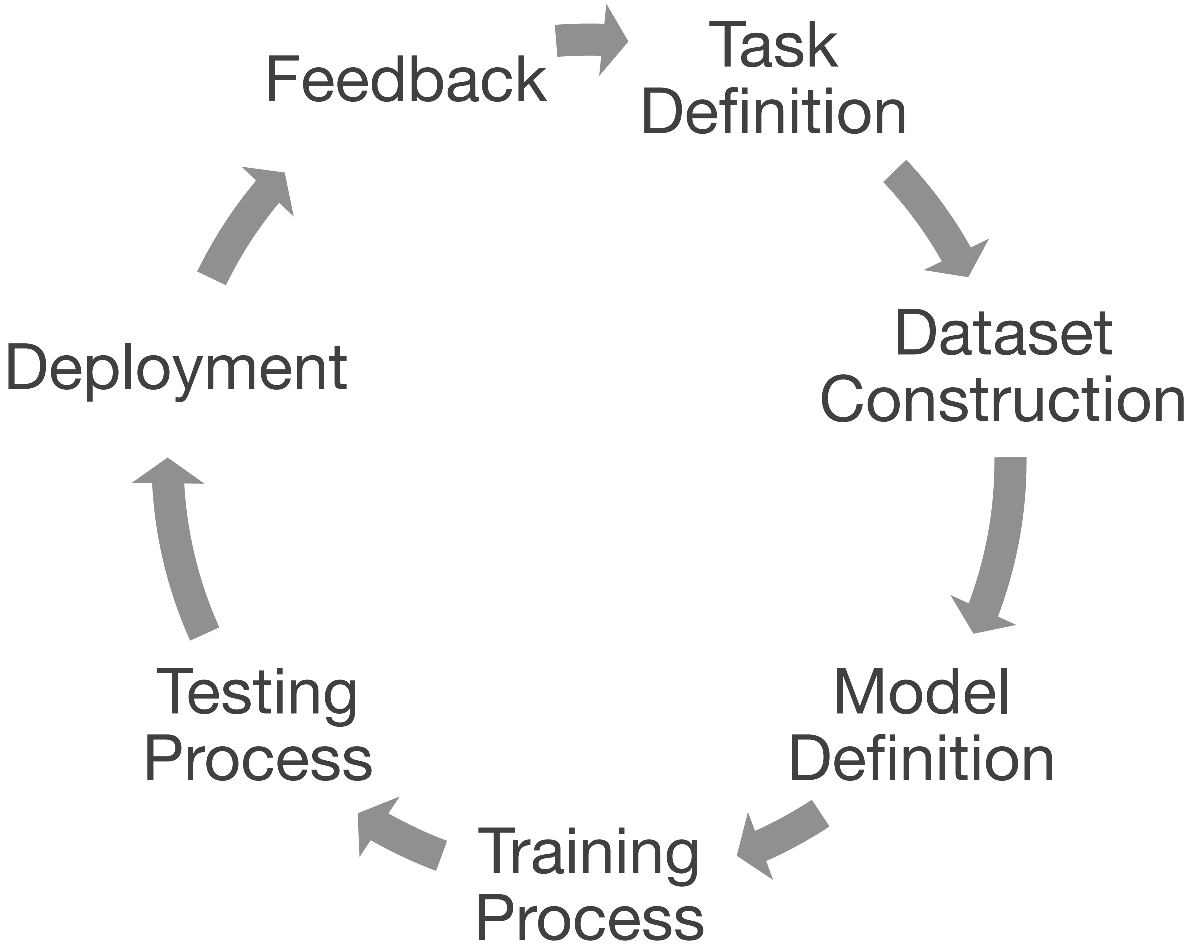
A case study

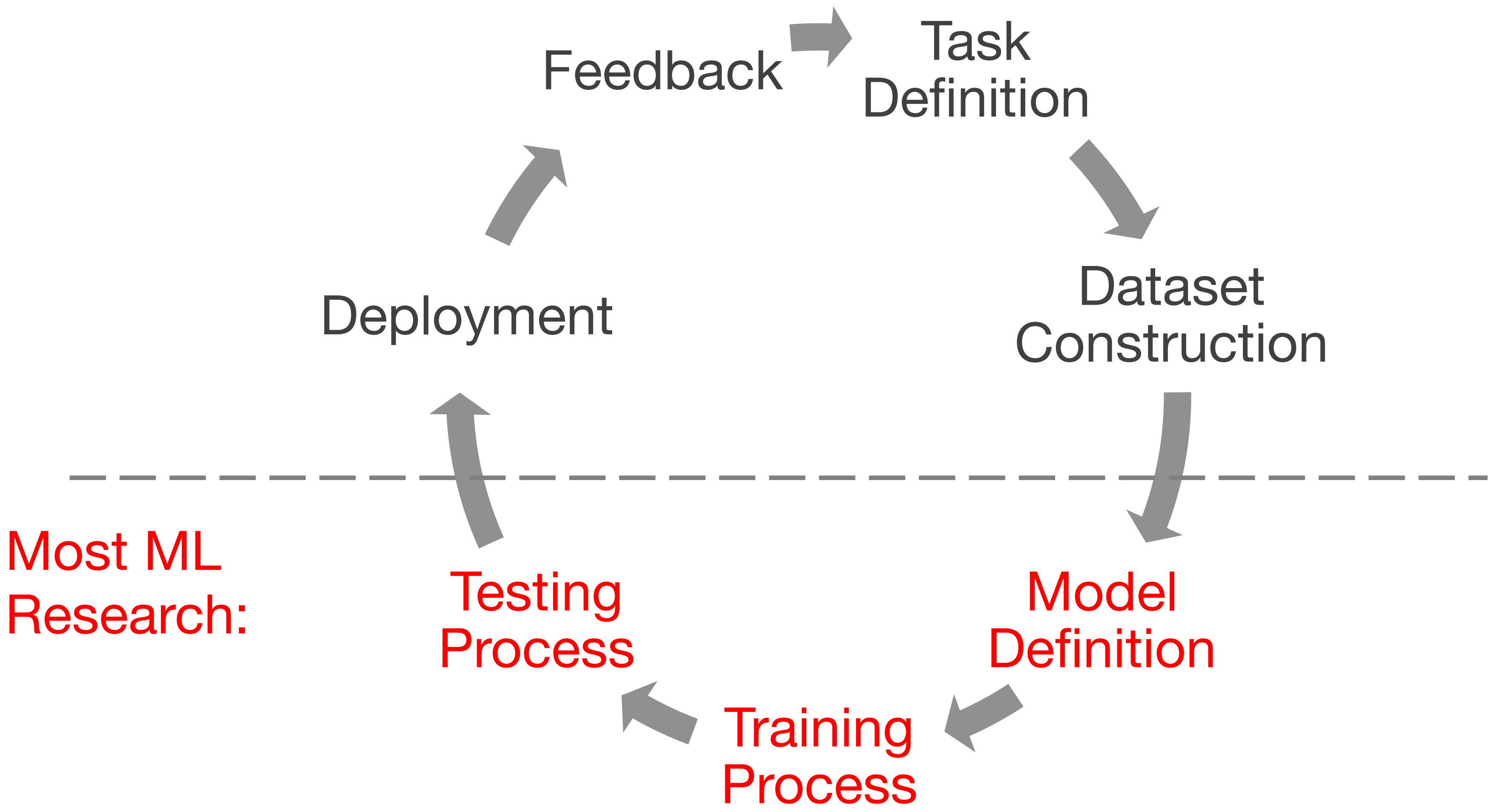
Henriette (20 mins)

Recap

Henriette
(5 mins)

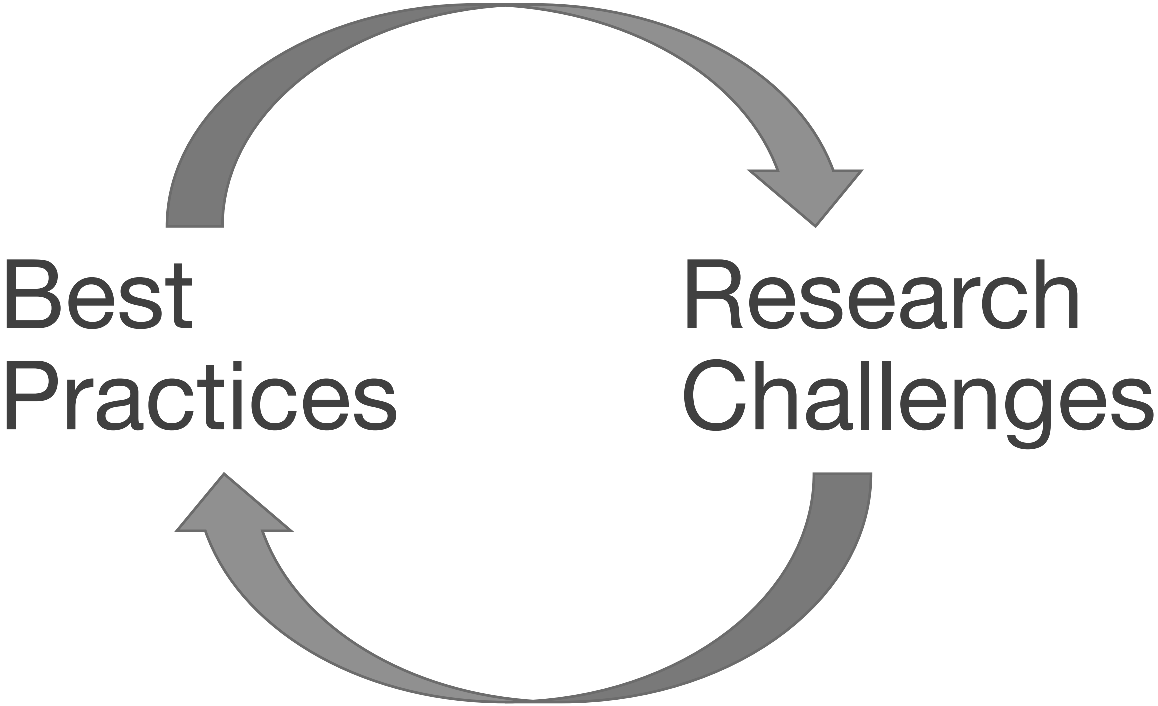
Fairness Throughout the Machine Learning Lifecycle

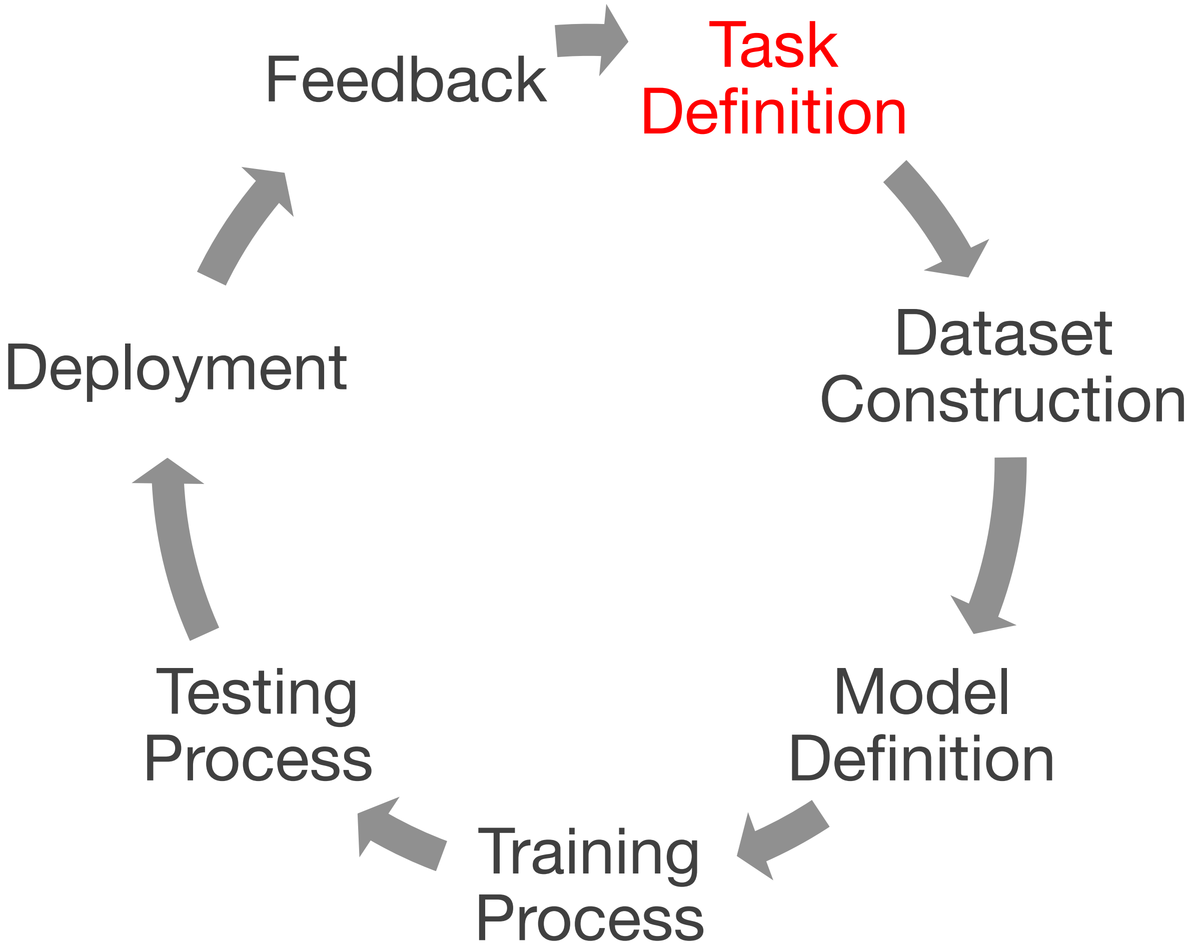

Task
Definition
Dataset
Construction
Model
Definition
Training
Process
Testing
Process
Deployment
Feedback



Feedback
Task Definition
Deployment
Dataset Construction
Most ML Research:
Testing Process
Model Definition
Training Process

Testing ≠ Deployment


Best
Practices
Research
Challenges


Task
Definition
Dataset
Construction
Model
Definition
Training
Process
Testing
Process
Deployment
Feedback

Task Definition

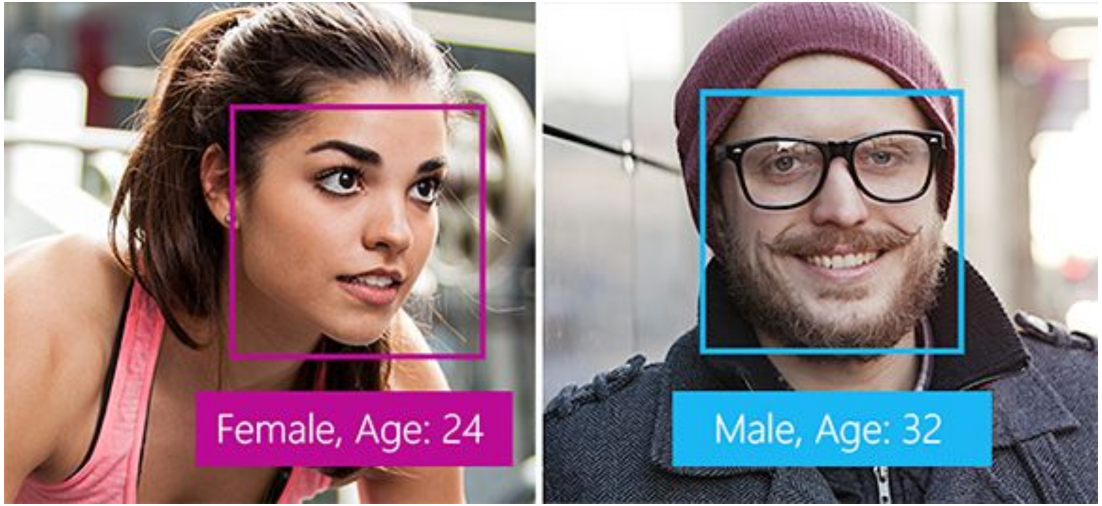
(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Figure 1. Sample ID photos in our data set.

[Wu & Zhang, 2016]

Task Definition

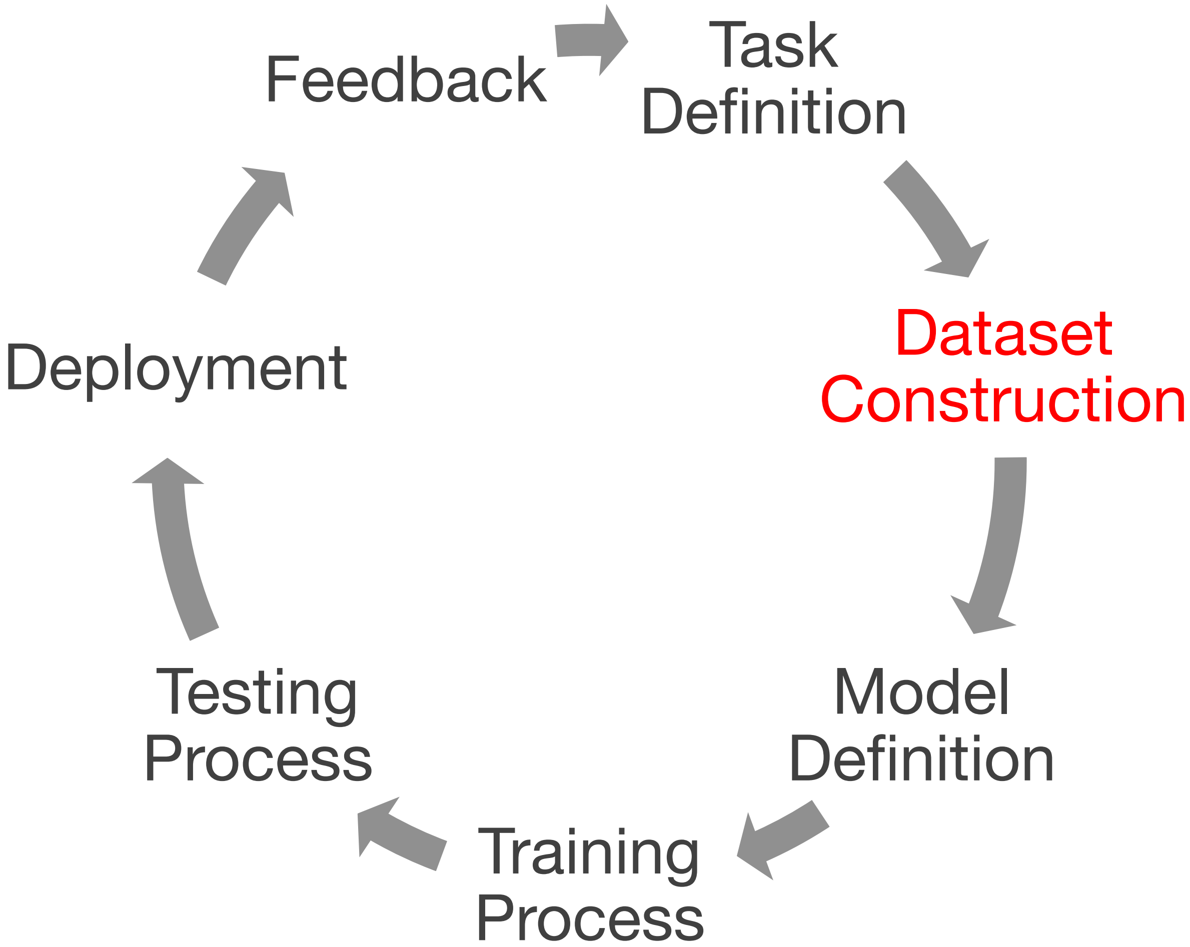


Best Practices: Task Definition

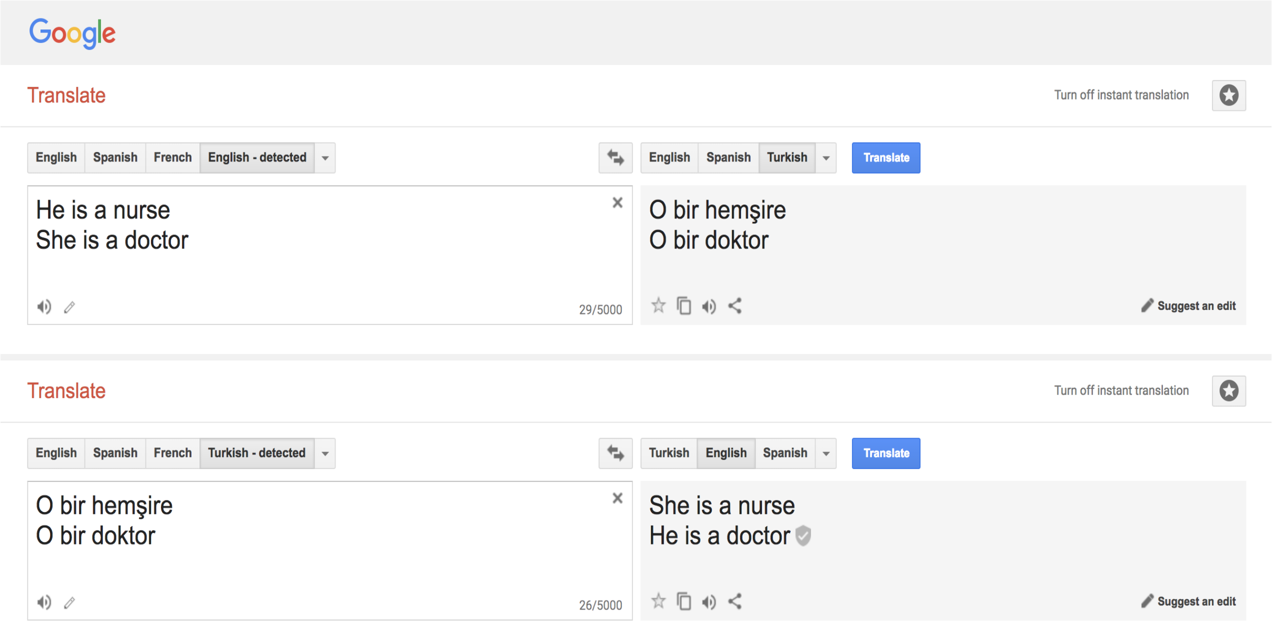
- Clearly define the task & model's intended effects
- Try to identify and document unintended effects & biases
- Clearly define any fairness requirements
- Involve diverse stakeholders & multiple perspectives
- Refine the task definition & be willing to abort

Research Challenges: Task Definition

- What are the most effective ways to elicit diverse opinions? [e.g.,http://techpolicylab.org/diverse-voices/]
- How should decisions be made within companies about which tasks to pursue and which to avoid?
- How should we design processes for uncovering unintended effects and biases before development?

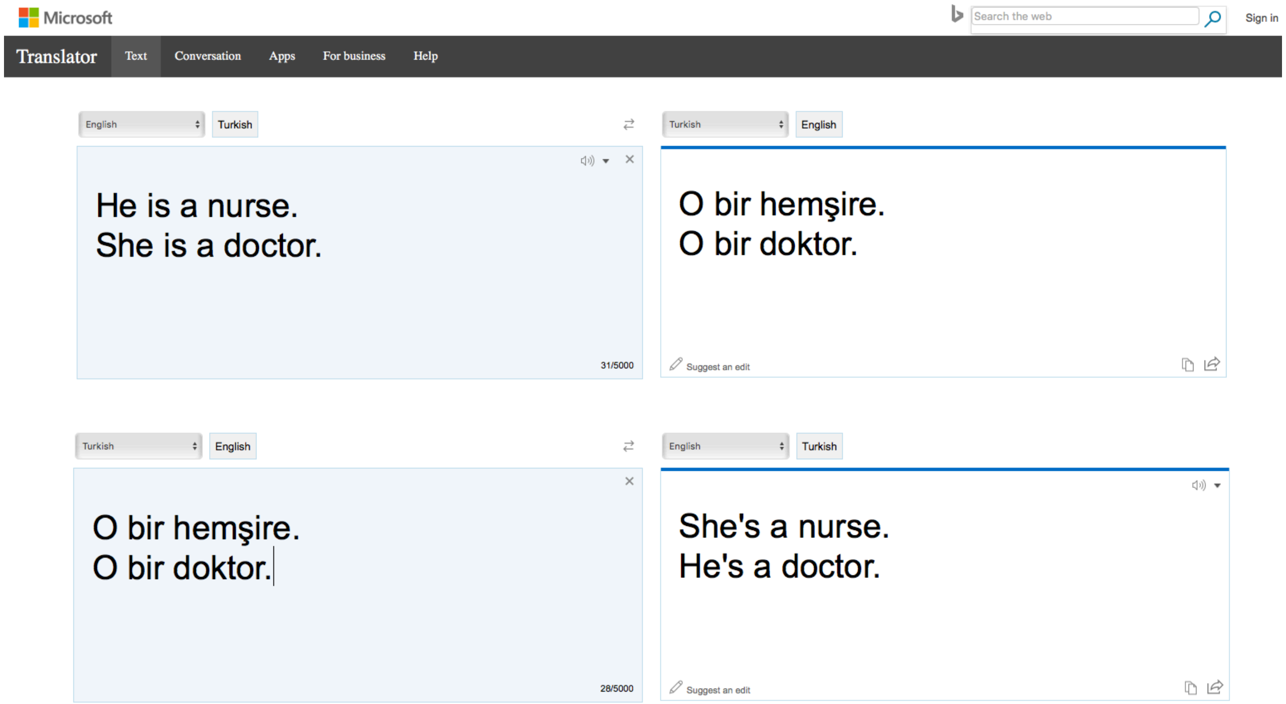

Task
Definition
Dataset
Construction
Model
Definition
Training
Process
Testing
Process
Deployment
Feedback

Data: Societal Bias



[Caliksan et al., 2017]

Data: Societal Bias



[Caliksan et al., 2017]

Data: Skewed Sample

Boston releases Street Bump app that automatically detects potholes while driving

By DAILY MAIL REPORTER

PUBLISHED: 19:37 EST, 20 July 2012 | **UPDATED:** 20:01 EST, 20 July 2012

Share

3
View comments

The next time your car hits a pothole, a new technology could help you immediately tell someone who can do something about it.

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

Best Practices: Data Collection

- 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

Research Challenges: Source/Collection

- Can we develop methods/tools to check for biases in the data source and data collection/sampling process?
- What constitutes “sufficient representation” of subpopulations?
- How can we achieve fairness without putting a tax on already disadvantaged populations?

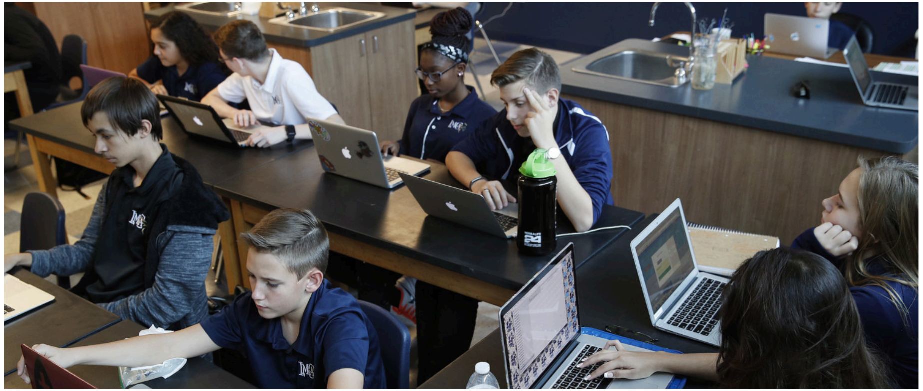
Solutions may be domain-specific!

Data: Labeler Bias

More States Opting To 'Robo-Grade' Student Essays By Computer

June 30, 2018 · 8:13 AM ET
Heard on Weekend Edition Saturday

TOVIA SMITH



Best Practices: Labeling & Preprocessing

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

Research Challenges: Labeling & Preprocessing

- Audit standard preprocessing tools for bias, along the lines of work on word embeddings [Bolukbasi et al. 2016]
- Develop techniques (e.g., training material or post-processing steps) to quantify and reduce the biases introduced by human labelers

Datasheets for Datasets

Motivation for Dataset Creation

Why was the dataset created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)
Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

What (other) tasks could the dataset be used for?
The LFW dataset can be used for the face identification problem. Some researchers have developed protocols to use the images in the LFW dataset for face identification.[2]

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?
Papers using this dataset and the specified evaluation protocol are listed in http://vis-www.cs.umass.edu/lfw/results.html

Who funded the creation of the dataset?
The building of the LFW database was supported by a United States National Science Foundation CAREER Award.

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)
Each instance is a pair of images labeled with the name of the person in the image. Some images contain more than one face. The labeled face is the one containing the central pixel of the image—other faces should be ignored as "background".

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?
There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

How many instances are there? (of each type, if appropriate)?
The dataset consists of 13,233 face images in total of 5749 unique individuals. 1680 of these subjects have two or more images and 4069 have single ones.

[1] All information in this datasheet is taken from one of five sources. Any errors that were introduced from these sources are our fault.
Original paper: http://www.cs.cornell.edu/people/pabo/movie-review-data/; LFW survey: http://vis-www.cs.umass.edu/lfw/lfw.pdf; Paper measuring LFW demographic characteristics : http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf; LFW website: http://vis-www.cs.umass.edu/lfw/.

[2] Unconstrained face recognition: Identifying a person of interest from a media collection: http://biometrics.cse.msu.edu/Publications/Face/BestRowdenetal_UnconstrainedFaceRecognition_TechReport_MSU-CSE-14-1.pdf

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?
Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format. Each image is accompanied by a label indicating the name of the person in the image. While subpopulation data was not available at the initial release of the dataset, a subsequent paper[3] reports the distribution of images by age, race and gender. Table 2 lists these results.

Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version; c) are there access restrictions or fees?
Everything is included in the dataset.

Are there recommended data splits and evaluation measures? (e.g., training, development, testing; accuracy or AUC)
The dataset comes with specified train/test splits such that none of the people in the training split are in the test split and vice versa. The data is split into two views, View 1 and View 2. View 1 consists of a training subset (pairsDevTrain.txt) with 1100 pairs of matched and 1100 pairs of mismatched images, and a test subset (pairsDevTest.txt) with 500 pairs of matched and mismatched images. Practitioners can train an algorithm on the training set and test on the test set, repeating as often as necessary. Final performance results should be reported on View 2 which consists of 10 subsets of the dataset. View 2 should only be used to test the performance of the final model. We recommend reporting performance on View 2 by using leave-one-out cross validation, performing 10 experiments. That is, in each experiment, 9 subsets should be used as a training set and the 10^{th} subset should be used for testing. At a minimum, we recommend reporting the **estimated mean accuracy**, $\hat{\mu}$ and the **standard error of the mean**: S_E for View 2.
$\hat{\mu}$ is given by:

$$\hat{\mu} = \frac{\sum_{i=1}^{10} p_i}{10} \tag{1}$$

where p_i is the percentage of correct classifications on View 2 using subset i for testing. S_E is given as:

$$S_E = \frac{\hat{\sigma}}{\sqrt{10}} \tag{2}$$

Where $\hat{\sigma}$ is the estimate of the standard deviation, given by:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{10} (p_i - \hat{\mu})^2}{9}} \tag{3}$$

The multiple-view approach is used instead of a traditional train/validation/test split in order to maximize the amount of data available for training and testing.

[3] http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf

A Database for Studying Face Recognition in Unconstrained Environments — Labeled Faces in the Wild

Training Paradigms: There are two training paradigms that can be used with our dataset. Practitioners should specify the training paradigm they used while reporting results.

- **Image-Restricted Training** This setting prevents the experimenter from using the name associated with each image during training and testing. That is, the only available information is whether or not a pair of images consist of the same person, not who that person is. This means that there would be no simple way of knowing if there are multiple pairs of images in the train/test set that belong to the same person. Such inferences, however, might be made by comparing image similarity/equivalence (rather than comparing names). Thus, to form training pairs of matched and mismatched images for the same person, one can use image equivalence to add images that consist of the same person.

  The files pairsDevTrain.txt and pairsDevTest.txt support image-restricted uses of train/test data. The file pairs.txt in View 2 supports the image-restricted use of training data.

- **Unrestricted Training** In this setting, one can use the names associated with images to form pairs of matched and mismatched images for the same person. The file people.txt in View 2 of the dataset contains subsets of of people along with images for each subset. To use this paradigm, matched and mismatched pairs of images should be formed from images in the same subset. In View 1, the files peopleDevTrain.txt and peopleDevTest.txt can be used to create arbitrary pairs of matched/mismatched images for each person. The unrestricted paradigm should only be used to create training data and not for performance reporting. The test data, which is detailed in the file pairs.txt, should be used to report performance. We recommend that experimenters first use the image-restricted paradigm and move to the unrestricted paradigm if they believe that their algorithm's performance would significantly improve with more training data. While reporting performance, it should be made clear which of these two training paradigms were used for particular test result.

What experiments were initially run on this dataset? Have a summary of those results.
The dataset was originally released without reported experimental results but many experiments have been run on it since then.

Any other comments?
Table 1 summarizes some dataset statistics and Figure 1 shows examples of images. Most images in the dataset are color, a few are black and white.

Property	Value
Database Release Year	2007
Number of Unique Subjects	5649
Number of total images	13,233
Number of individuals with 2 or more images	1680
Number of individuals with single images	4069
Image Size	250 by 250 pixels
Image format	JPEG
Average number of images per person	2.30

Table 1. A summary of dataset statistics extracted from the original paper: Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

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%

Table 2. Demographic characteristics of the LFW dataset as measured by Han, Hu, and Anil K. Jain. *Age, gender and race estimation from unconstrained face images.* Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5) (2014).

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API)
The raw images for this dataset were obtained from the Faces in the Wild database collected by Tamara Berg at Berkeley[4]. The images in this database were gathered from news articles on the web using software to crawl news articles.

Who was involved in the data collection process? (e.g., students, crowdworkers) and how were they compensated (e.g., how much were crowdworkers paid)?
Unknown

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame of the instances?
Unknown

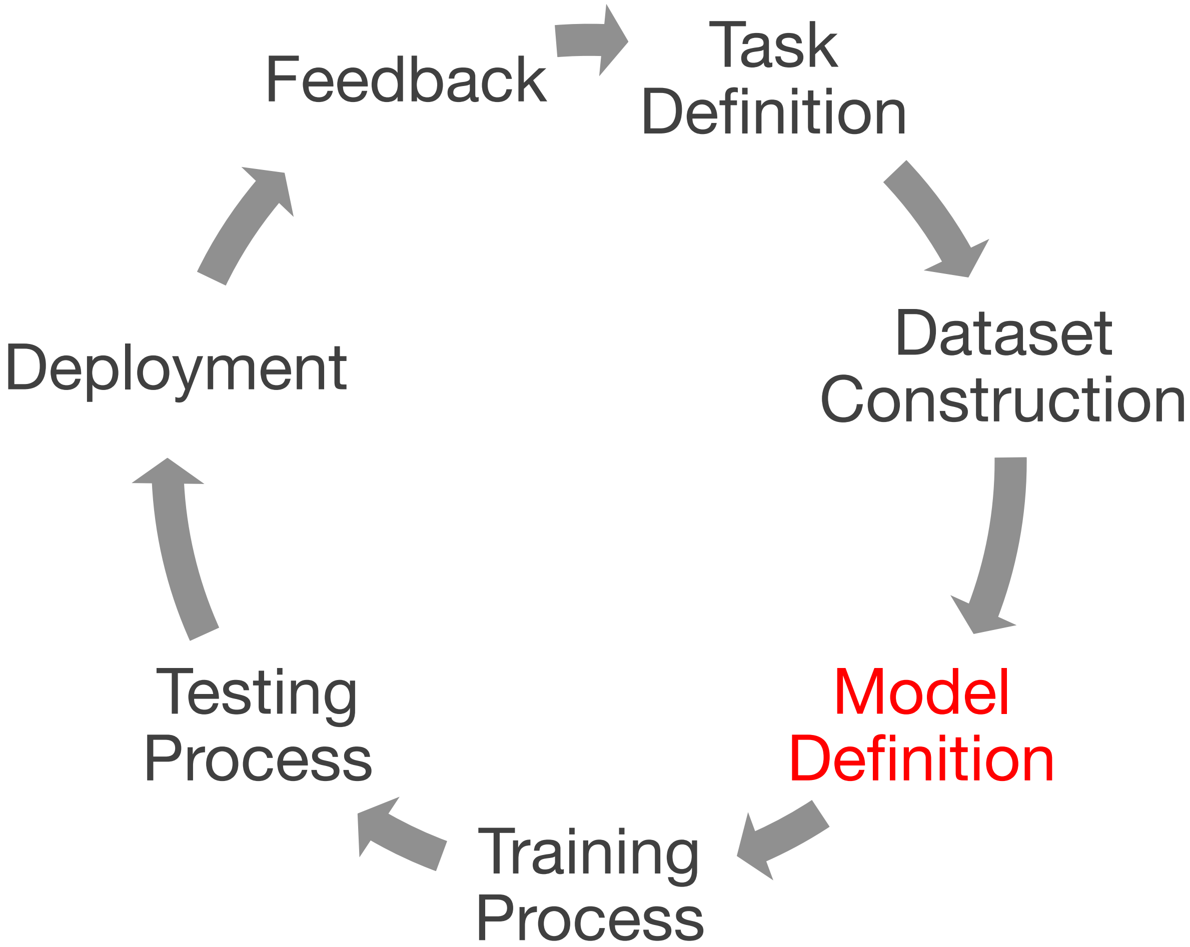
[Gebru et al., 2018]

Research Challenges: Datasheets

- What is the right set of questions?
  – How best to handle continually evolving datastreams?
  – Are there legal or PR risks to creating datasheets?

- What is the right process for making a datasheet?
  – How best to incentivize developers & PMs?
  – How much (if anything) should be automated?

[Gebru et al., 2018]



Feedback
Task Definition
Dataset Construction
Model Definition
Training Process
Testing Process
Deployment

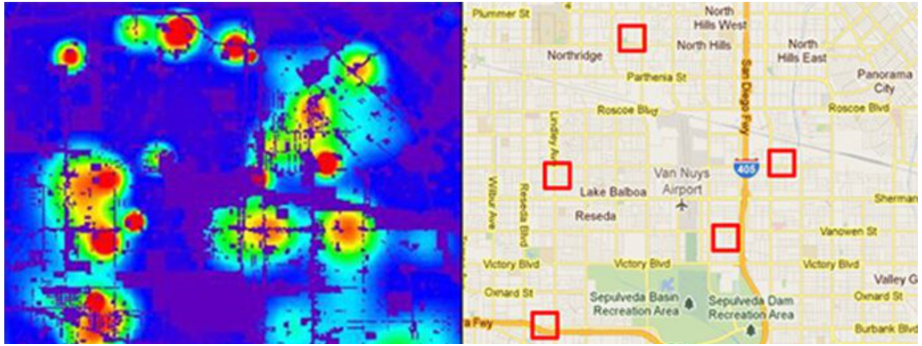
What is a model?

$$\begin{aligned}\text{price of house} = \; & w_1 \text{ * number of bedrooms +} \\ & w_2 \text{ * number of bathrooms +} \\ & w_3 \text{ * square feet +} \\ & \text{a little bit of noise}\end{aligned}$$

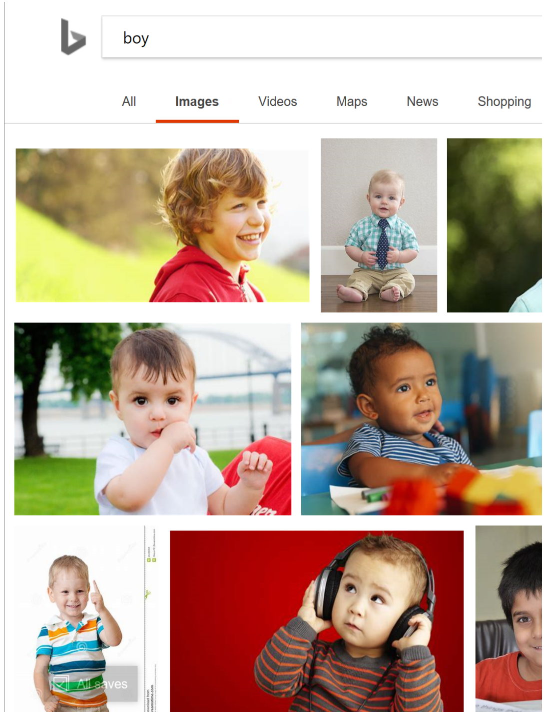
Model: Assumptions

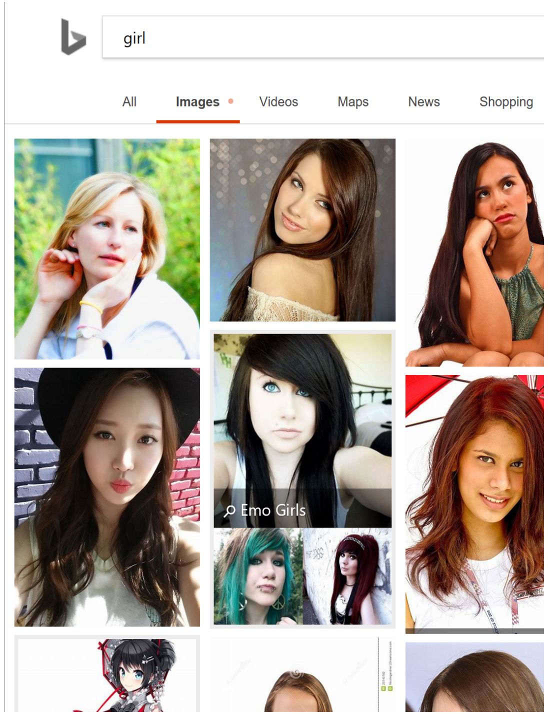
Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?

The software is supposed to make policing more fair and accountable. But critics say it still has a way to go.



Model: Objective Function



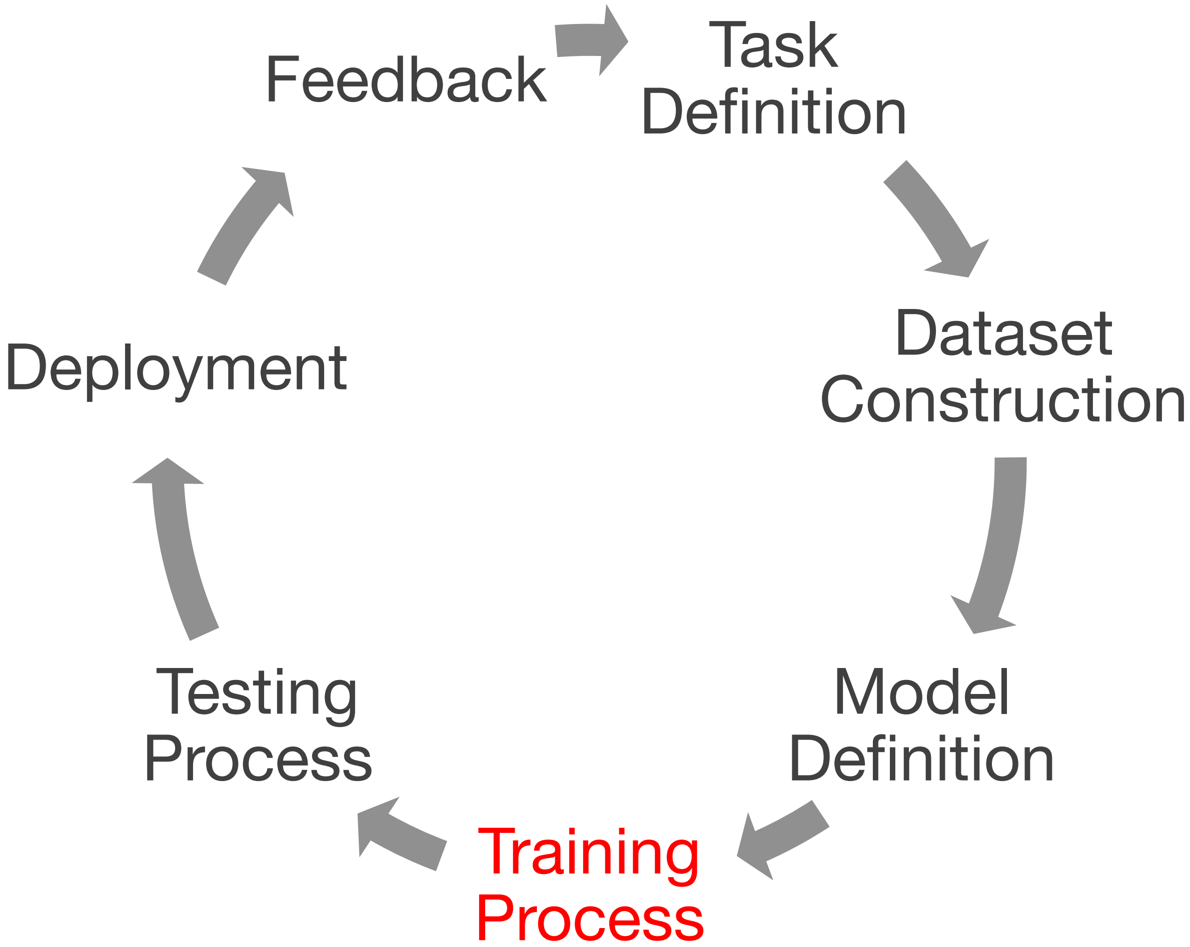


Best Practices: Model Definition

- 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

Research Challenges: Model Definition

- Identify biases in common modeling assumptions (in consultation with domain experts)
- Explore ways in which some measure of “fairness” might be included in the objective function—but be thoughtful about the limitations of this approach! [e.g., Corbett-Davies and Goel, 2018]
- Move beyond supervised learning

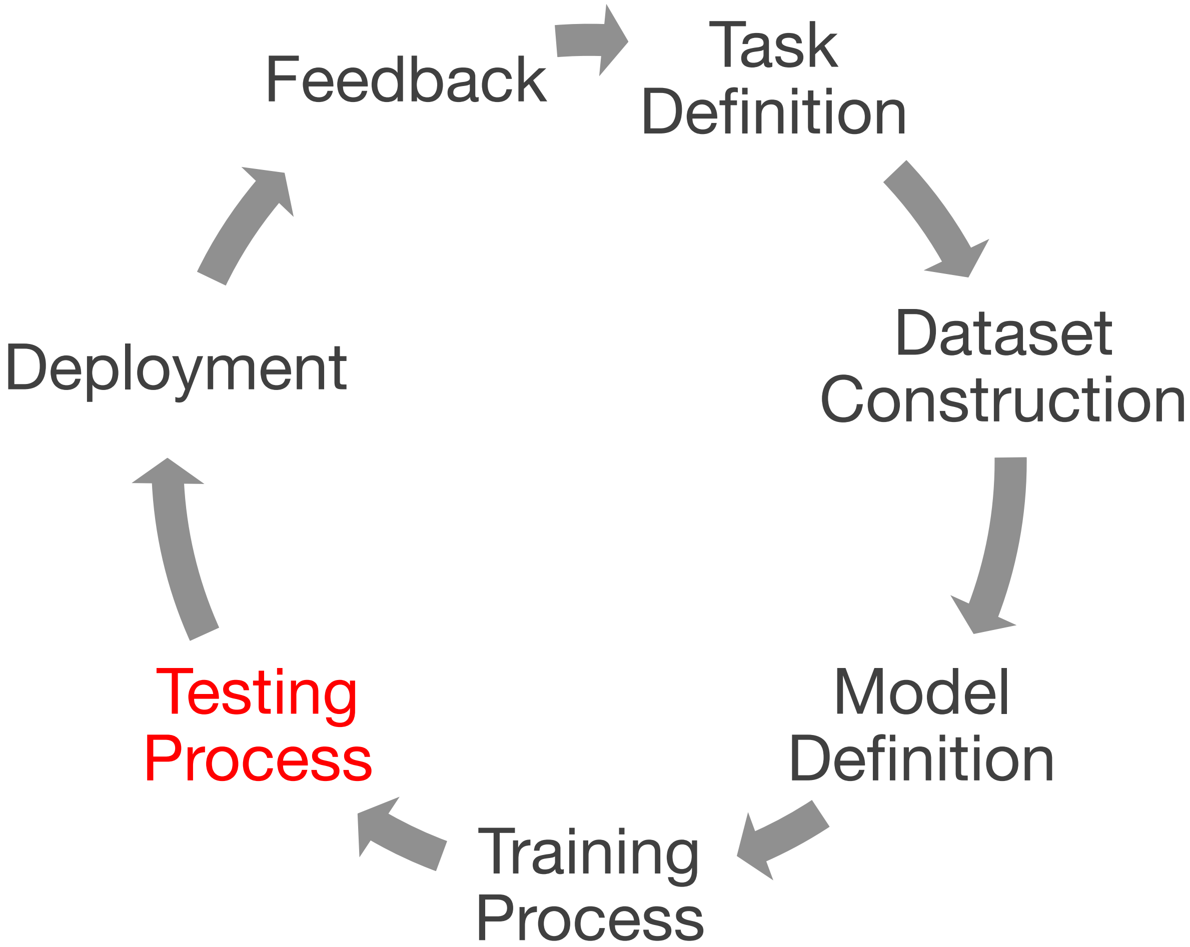

Feedback
Task Definition
Dataset Construction
Model Definition
Training Process
Testing Process
Deployment

What is training?

$$\text{price of house} = w_1 * \text{number of bedrooms} + w_2 * \text{number of bathrooms} + w_3 * \text{square feet} + \text{a little bit of noise}$$

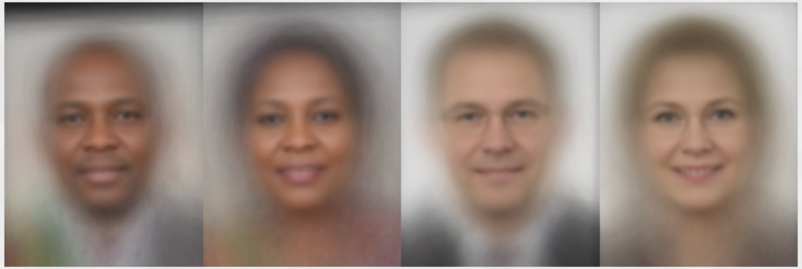
Training Process

```
if schedule['Lambda_SKCC'] <= self.total_iter:
    start = time.time()

    shp_SKCC[:] = np.outer(W_d_C, W_d_C)
    shp_SKCC[:, :, bool_diag_CC] = W_a_C * W_d_C
    shp_SKCC *= W_K[None, :, None, None]
    shp_SKCC *= W_S[:, None, None, None]
    post_shp_SKCC = shp_SKCC + Y_SKCC

    if mask.ndi[illegible]2:
        mask_N[illegible]k
        zeta_S[illegible]p.[illegible](Theta_NC.T, np.dot(ma[illegible]NN, Theta_NC))
        zeta_[illegible]TS.sum(axis=0)[:, None, None]
    else:
        mask_TNN = mask
        zeta_TNC = np.einsum('tij,jd->tid', mask_TNN, Theta_NC)
        zeta_TCC = np.einsum('tid,ic->tcd', zeta_TNC, Theta_NC)
        zeta_SCC = np.einsum('tcd,ts->scd', zeta_TCC, Psi_TS)
    post_rte_SKCC = d + zeta_SCC[:, None, :, :]
```


Task
Definition
Dataset
Construction
Model
Definition
Training
Process
Testing
Process
Deployment
Feedback

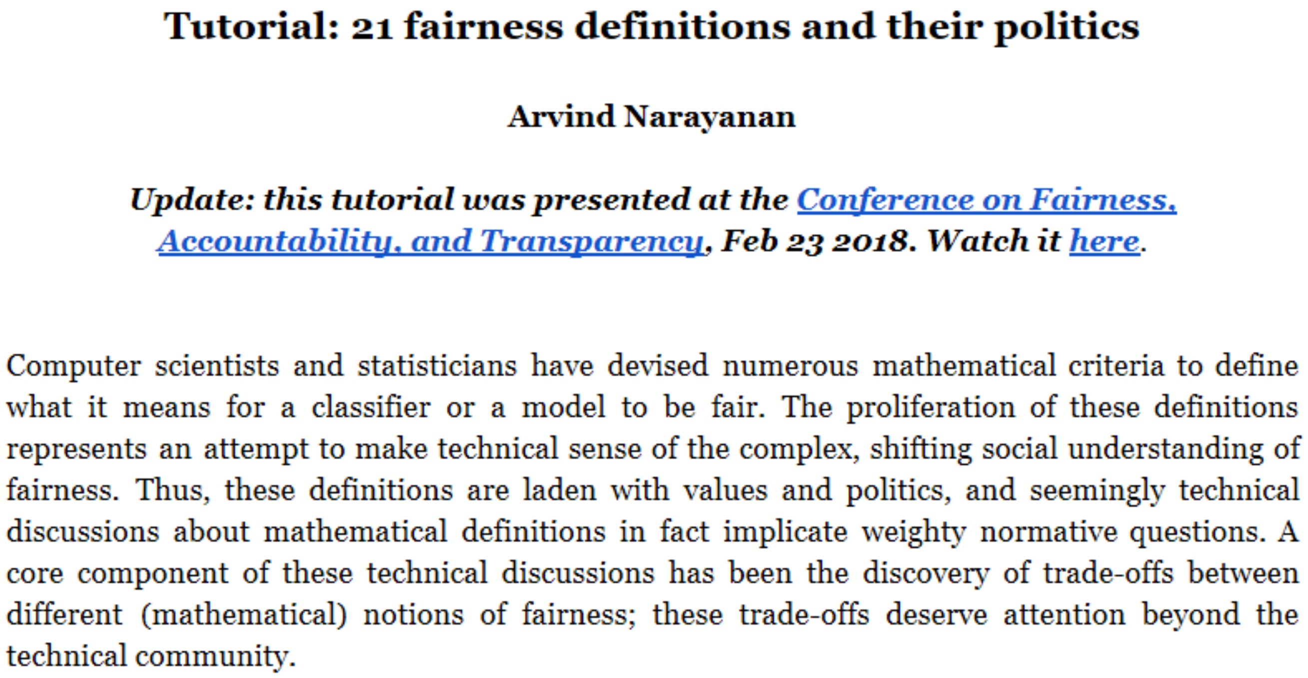
Testing: Data

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



[Buolamwini & Gebru, 2018]

Testing: Metrics


Tutorial: 21 fairness definitions and their politics

Arvind Narayanan

Update: this tutorial was presented at the Conference on Fairness, Accountability, and Transparency, Feb 23 2018. Watch it here.

Computer scientists and statisticians have devised numerous mathematical criteria to define what it means for a classifier or a model to be fair. The proliferation of these definitions represents an attempt to make technical sense of the complex, shifting social understanding of fairness. Thus, these definitions are laden with values and politics, and seemingly technical discussions about mathematical definitions in fact implicate weighty normative questions. A core component of these technical discussions has been the discovery of trade-offs between different (mathematical) notions of fairness; these trade-offs deserve attention beyond the technical community.

Metrics: Points to Consider

Fairness is a non-trivial sociotechnical challenge

- » Many types of harm relate to a broader cultural context than a single decision-making system
- » Many aspects of fairness not captured by metrics

No free lunch! Can't satisfy all metrics [Kleinberg et al. 2017]

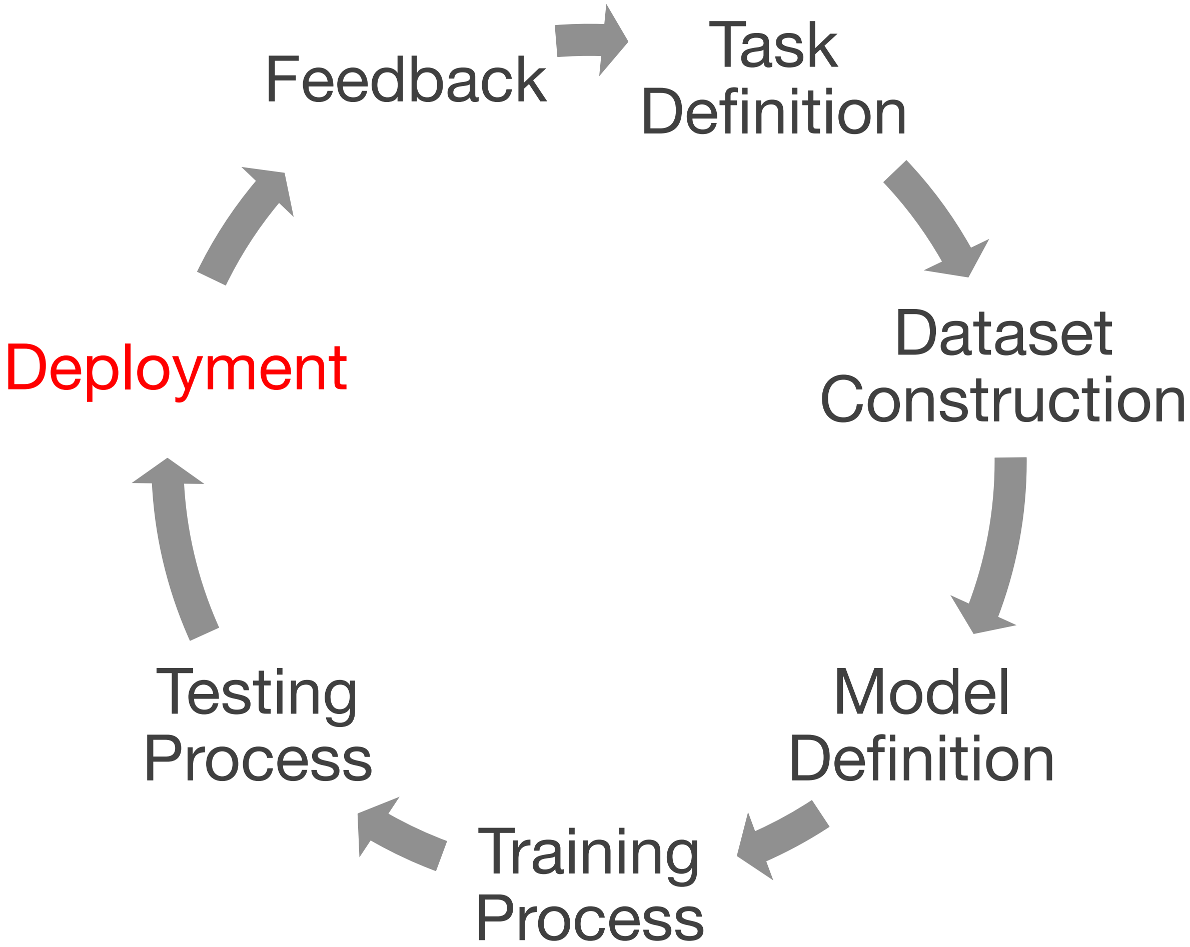
- » Need to make different tradeoffs in different contexts

Best Practices: Testing

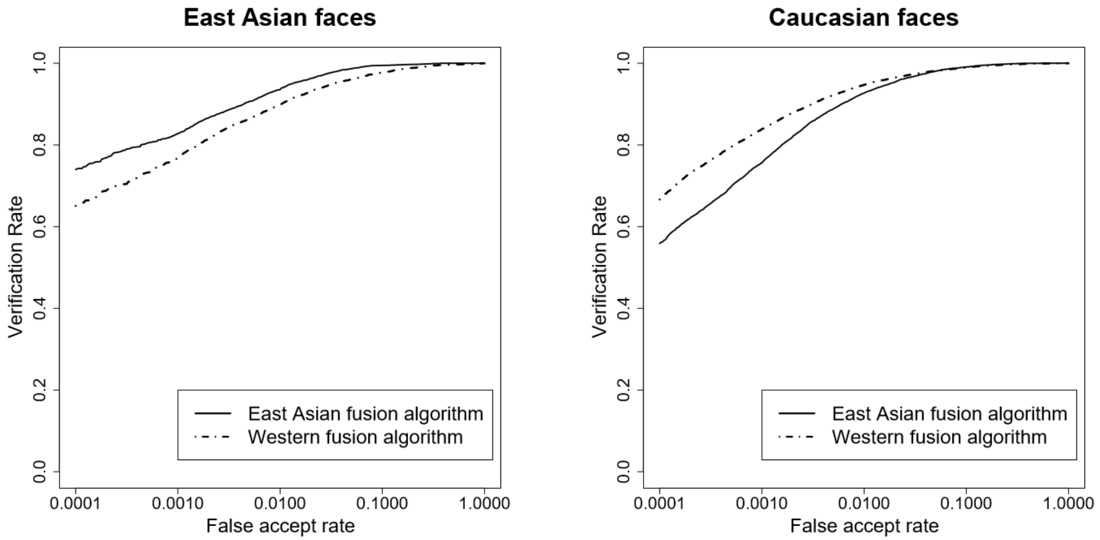
- Check that test data matches deployment context
- Ensure test data has sufficient representation
- Continue to involve diverse stakeholders
- Revisit all fairness requirements
- Use metrics to check that requirements are met

Research Challenges: Testing

- What constitutes “sufficient representation” of subpopulations for test data in different domains?
- What are the subpopulations of interest for testing?
- Which fairness metrics are appropriate in which scenarios?
- What are the right fairness metrics for unsupervised learning, RL, or complex systems like chatbots?


Task Definition
Dataset Construction
Model Definition
Training Process
Testing Process
Deployment
Feedback

Deployment: Context



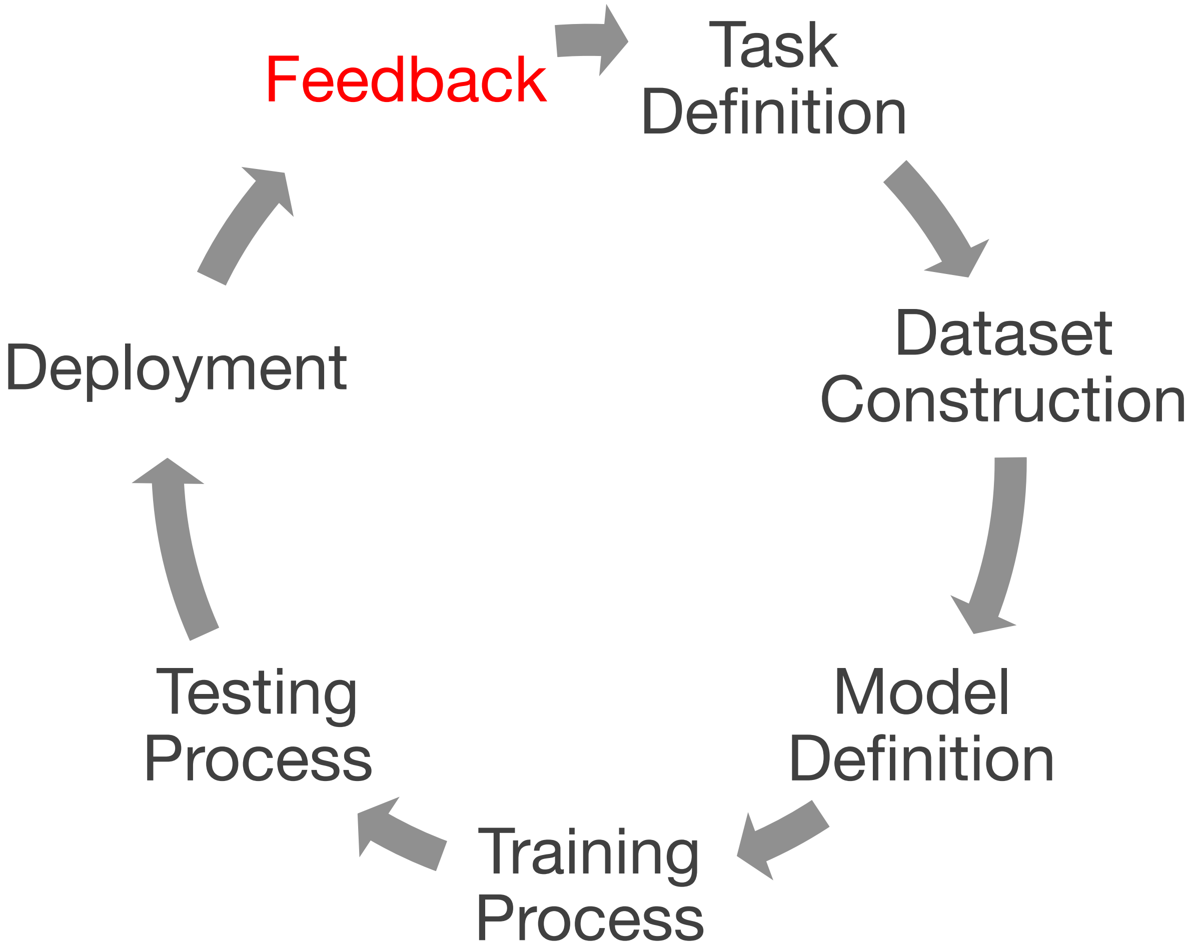
[Phillips et al., 2011]

Best Practices: Deployment

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

Research Challenges: Deployment

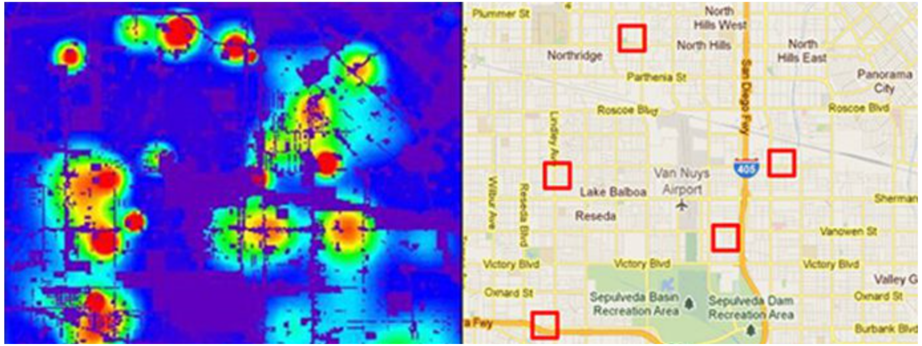
- Methods/tools to audit for shifts in population
- Methods/tools to determine whether a particular error is a one-off issue or is indicative of a systemic problem
- Audit existing system for biases (in collaboration with the teams that built the systems whenever possible)


Feedback
Task Definition
Dataset Construction
Model Definition
Training Process
Testing Process
Deployment

Feedback: Non-Adversarial

Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?

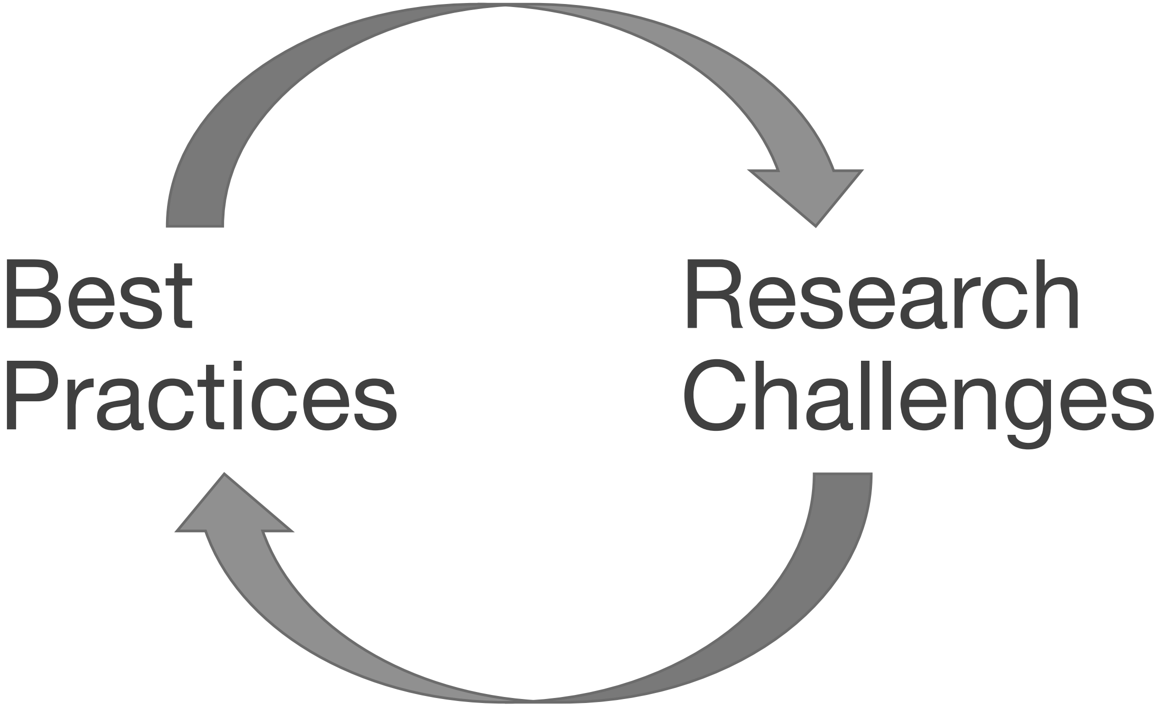
The software is supposed to make policing more fair and accountable. But critics say it still has a way to go.



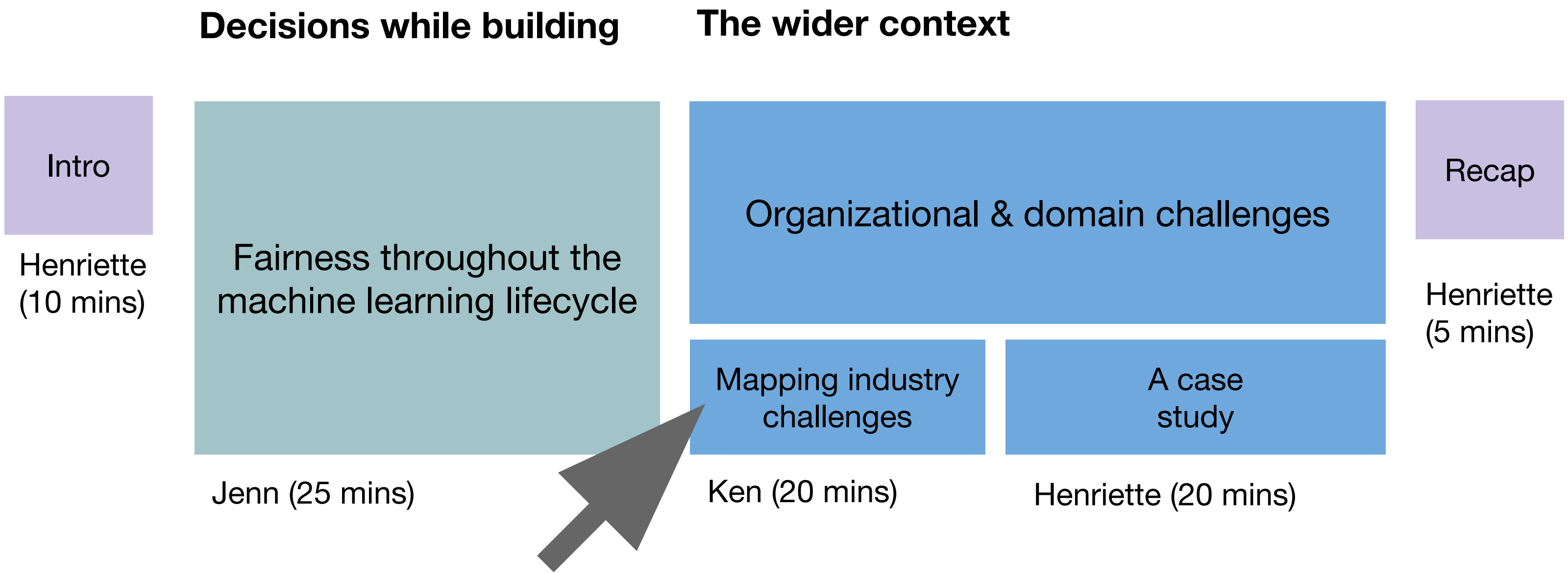
Feedback: Adversarial

Best Practices: Feedback

- Continue to monitor
  - match between training data, test data, and instances you encounter in deployment
  - fairness metrics
  - user reports & user complaints
- Monitor users' interactions with system
- Consider prohibiting some types of interactions


Best
Practices
Research
Challenges

This 90-min tutorial



Improving fairness in ML systems: What do industry practitioners need?

(Holstein, Wortman Vaughan, Daumé III, Dudík, & Wallach, in press)

“...it would be so valuable to have more researchers want to embed on certain problems with product groups … so there's a shared sense of success by solving as opposed to [...] sitting outside of the problem and critiquing it...”

- anonymous interviewee

Initial, exploratory interviews with product managers (PMs) for each of **6 product teams** at a major technology company

Domains
Machine Translation
Computer Vision
Speech and Voice
Content Personalization / Optimization
Natural Language Understanding
Face Recognition and Classification

Initial, exploratory interviews with product managers (PMs) for each of 6 product teams at a major technology company

→ Disconnects between research and practice

Domains
Machine Translation
Computer Vision
Speech and Voice
Content Personalization / Optimization
Natural Language Understanding
Face Recognition and Classification

Main interview study

Technology Area	Roles of Participants	Participant IDs
Adaptive Tutoring & Mentoring	Chief Data Scientist, CTO, Data Scientist, Research Scientist	R10, [R13, R14], R30
Chatbots	CEO, Product Manager, UX Researcher	[R17, R18], R35
Vision & Multimodal Sensing	CTO, ML Engineer, Product Manager, Software Engineer	[R2, R3, R4], R6, R7, R9, R26
General-purpose ML (e.g., APIs)	Chief Architect, Director of ML, Product Manager	R25, R32, R34
NLP (e.g., Speech, Translation)	Data Manager, Data Collector, Domain Expert, ML Engineer, PM, Research Software Eng., Technical Mgr., UX Designer	R1, [R15, R16, R19, R20, R21, R22], R24, [R27, R29], R28, R31
Recommender Systems	Chief Data Scientist, Data Scientist, Head of Diversity Analytics	R8, R12, R23, R33
Web Search	Product Manager	R5, R11

Series of semi-structured interviews with an additional 29 ML practitioners across 25 product teams from 10 major technology companies

Main interview study

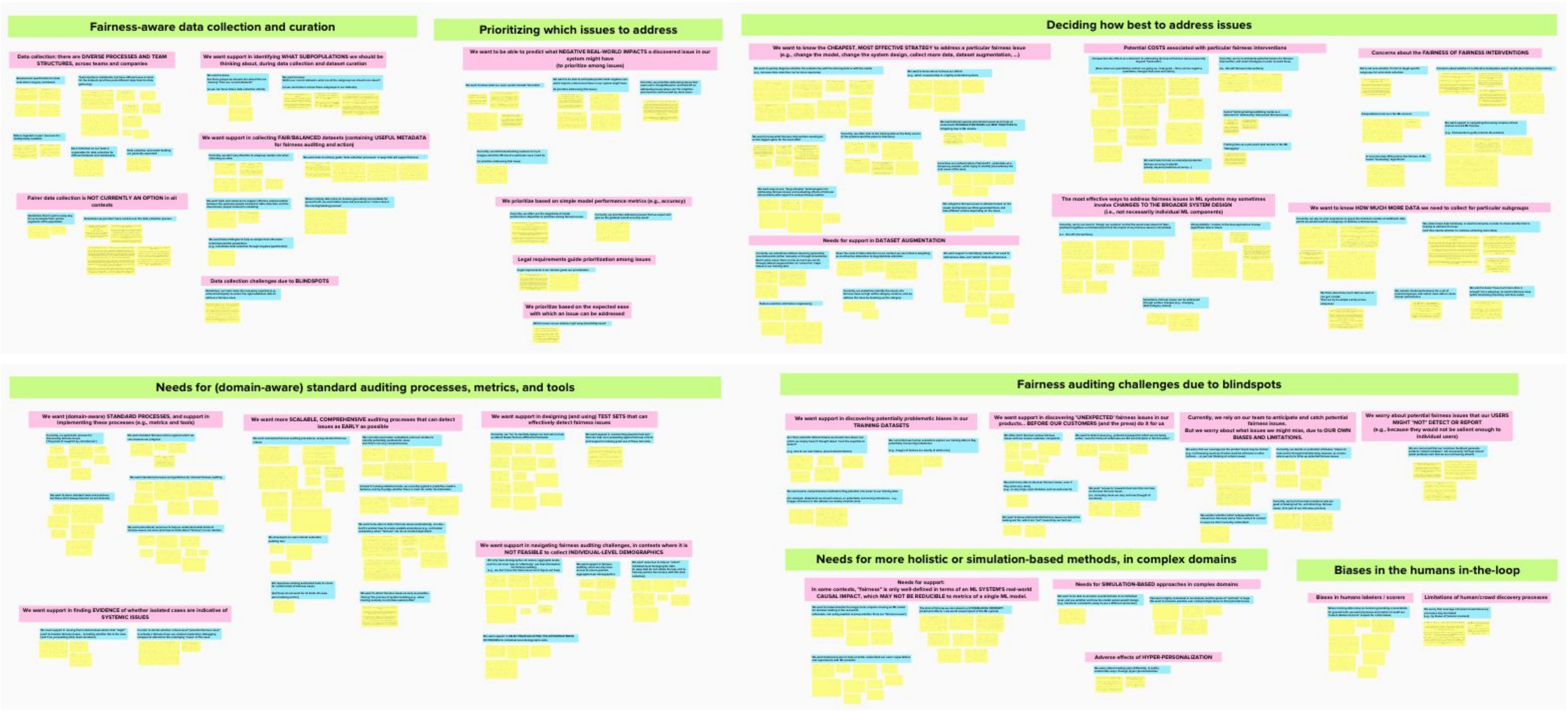
Current practices and challenges

e.g., “Can you recall times you or your team ____________________? … Can you walk me through *how* your team ____________________?”

Needs for additional support

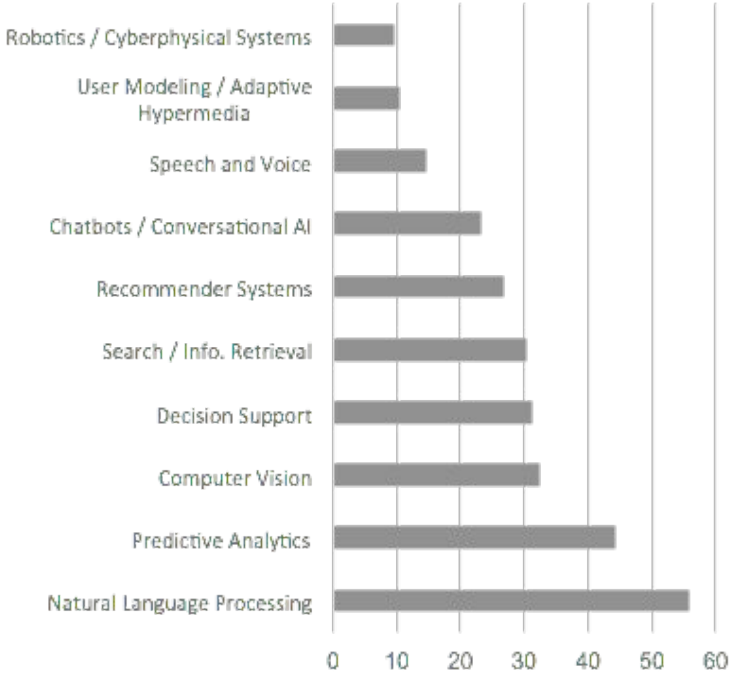
e.g., “Imagine you’d had access to a magical, all-knowing oracle, and could ask it anything you wanted, to help your team ____________________ …”

Bottom-up, Iterative Affinity Diagramming

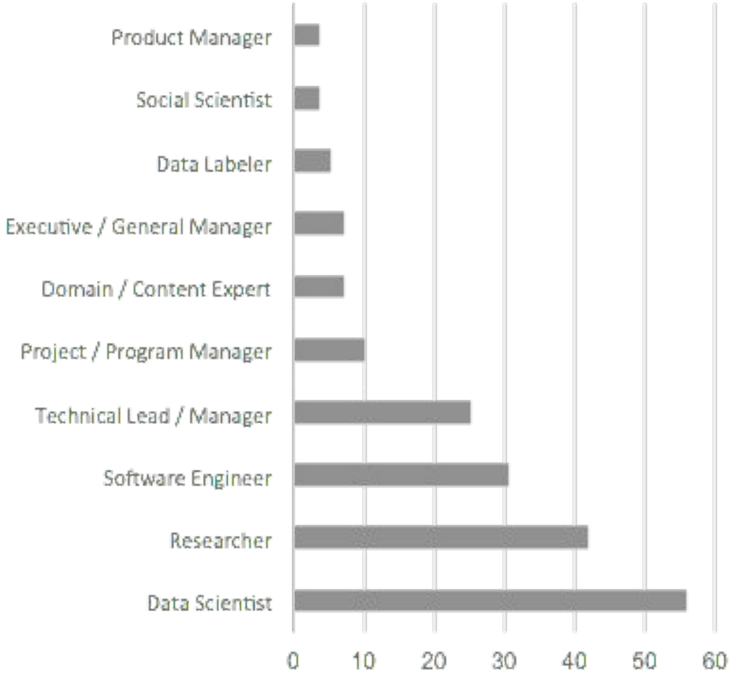


Anonymous survey (n=267)

Technology areas



Team roles



Disconnects

Models vs. Data

- ML literature generally assumes data is given and focuses on fair models and/or algorithms to optimize fairness metrics.

- Industry practitioners more often turn to the data first
  - 65% of survey respondents reported having control over data collection or curation
  - 73% of respondents who had tried to address fairness issues had focused on collecting more training data

Models vs. Data

- Needs for support in creating datasets that support fairness downstream
  - e.g., tools to diagnose whether a given fairness issue might be addressed by **collecting more training data** from a particular subpopulation … and to predict **how much more** data is needed

> **"I always would just really want to know how much was enough." - R4**

(cf. Chen, Johansson, & Sontag, 2018; Nushi, Kamar, & Horvitz, 2018)

Models vs. Data

- Needs for support in creating datasets that support fairness downstream
  - e.g., tools to help **actively guide** data collection / curation processes

> **"To score African American students fairly, they need examples of [these] students scoring highly. But in the data [the data collection team] collect[s], this is very rare.**
>
> **[...We need] some kind of way to indicate [which schools] to collect from [...] or what to bother spending the extra money to score."** - R19

(cf. Chen, Johansson, & Sontag, 2018; Nushi, Kamar, & Horvitz, 2018)

Blind Spots

- ML literature often assumes subpopulations of interest are given (e.g., based on race, gender, age, religion), but several interviewees highlighted needs for support in identifying relevant subpopulations
  - 62% of survey respondents said it would be very/extremely useful

Blind Spots

- ML literature often assumes subpopulations of interest are given (e.g., based on race, gender, age, religion), but several interviewees highlighted needs for support in identifying relevant subpopulations
  - 62% of survey respondents said it would be very/extremely useful

> **"It's just everyone's collecting all the things that they can think of that could be offensive and testing for it"** - R2

> **"...you know, no one person on the team are experts in all types of bias or offense... especially when you take into account different cultures and different parts of the world"** - R4

Blind Spots

"[although people tend to] start thinking about attributes like [ethnicity and gender], the biggest problem I found is that these [subpopulations] should be defined based on the domain and problem." - R32

Blind Spots

“It’d be nice to have a central place to kind of know where we could potentially go wrong…

Otherwise, you just have to put your model out there, and then you know if there’s fairness issues if someone raises hell...” - R7

Blind Spots

"It'd be nice to have a central place to kind of know where we could potentially go wrong...

Otherwise, you just have to put your model out there, and then you know if there's fairness issues if someone raises hell..." - R7

- Scaffolding fairness-aware test set design
  - (e.g., sharing test sets across teams, facilitating rapid dataset annotation)

(cf. Yang, Suh, Chen, & Ramos, 2018)

Blind Spots

- Interviewees shared stories in which they were hampered in **addressing** issues by their teams' cultural blind spots

Blind Spots

- Interviewees shared stories in which they were hampered in **addressing** issues by their teams' cultural blind spots

> **"If I noticed that there's some celebrity from Taiwan that doesn't have enough images in there, I actually don't know what they look like to go and fix that. [...]**
>
> **But, Beyoncé, I know what she looks like."** - R4

Blind Spots

- Team diversity
- Fairness-focused interview questions
- Ad-hoc recruitment of diverse, team-external “experts”
  (for specific tasks requiring team-external knowledge)

(cf. Retelny, Robaszkiewicz, To, Lasecki, Patel, Rahmati, Doshi, Valentine, & Bernstein, 2014)

UX Side Effects of Fairness Interventions

- Needs for tools and processes that can help teams **anticipate trade-offs** between particular aspects of fairness and other desiderata for an ML system (beyond 'fairness vs accuracy' – e.g., user satisfaction)

(cf. Dove, Halskov, Forlizzi, & Zimmerman, 2017; Friedman & Nissenbaum, 1996; Selbst, Friedler, Venkatasubramanian, & Vertesi, 2019)

UX Side Effects of Fairness Interventions

- Needs for tools and processes that can help teams **anticipate trade-offs** between particular aspects of fairness and other desiderata for an ML system (beyond 'fairness vs accuracy' – e.g., user satisfaction)

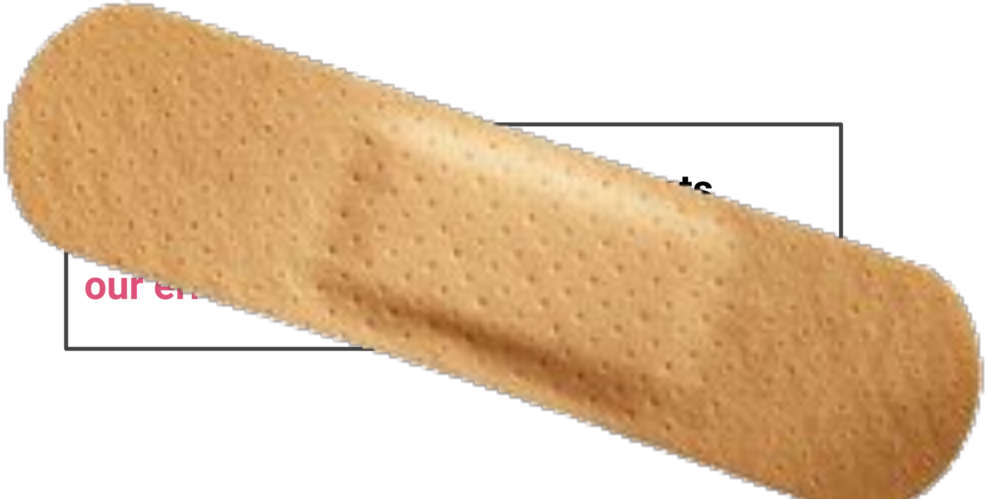
> **"...we had a couple of deployments which regressed in serious ways which our error rate did not reflect..."** - R1

> **"...even if your scores come out better... at the end of the day, it's really just different from what you had before... and [customers] notice that for their particular scenario... it's different in a negative way..."** - R4

(cf. Dove, Halskov, Forlizzi, & Zimmerman, 2017; Friedman & Nissenbaum, 1996; Selbst, Friedler, Venkatasubramanian, & Vertesi, 2019)

UX Side Effects of Fairness Interventions

- Teams often reported implementing local, **“band-aid” solutions** to avoid risk of system-wide side effects



“So the idea really is fix the problem... for the [specific] case under investigation but try not to break anything else” **- R1**

Limitations of Existing ML Methods

- Most fairness metrics designed for classification (bail/no bail, hire/no hire), while product groups face a much richer space of applications (chatbots, adaptive tutoring, search)
    - Interviewees reported **struggling to use existing fairness research**
    - Applications less amenable to de-contextualized fairness metrics of isolated ML system components

Limitations of Existing ML Methods

- Most fairness metrics designed for classification (bail/no bail, hire/no hire), while product groups face a much richer space of applications (chatbots, adaptive tutoring, search)
  - Interviewees reported **struggling to use existing fairness research**
  - Applications less amenable to de-contextualized fairness metrics of isolated ML system components

> **"[with] contextual kinds of responses [it is] harder to [...] predict all the outcomes [... It would help to] find ways to automate the identification of risky conversation patterns that emerge." - R17**

(cf. Jain, Pecune, Matsuyama, & Cassell, 2018)

Limitations of Existing ML Methods

- Most fairness metrics designed for classification (bail/no bail, hire/no hire), while product groups face a much richer space of applications (chatbots, adaptive tutoring, search)
  - Interviewees reported **struggling to use existing fairness research**
  - Applications less amenable to de-contextualized fairness metrics of isolated ML system components

"[with] contextual kinds of responses [it is] harder to [...] predict all the outcomes [... It would help to] find ways to automate the identification of risky conversation patterns that emerge." - R17

"If we think about educational interventions as analogous to medical interventions or drug trials [...] we know and [expect] a particular intervention will have different effects on different subpopulations." - R30

(cf. Friedman & Nissenbaum, 1996; Selbst, Friedler, Venkatasubramanian, & Vertesi, 2019)

Limitations of Existing ML Methods

- ML literature generally assumes individual-level access to sensitive attributes, which many teams lack
  - Needs for support in effectively and efficiently monitoring fairness with access only to coarse-grained, partial, or indirect information (e.g., neighborhood- or organization-level statistics)

(cf. Kilbertus et al., 2018; Veale & Binns, 2018)

Limitations of Existing ML Methods

- ML literature generally assumes individual-level access to sensitive attributes, which many teams lack
  - Needs for support in effectively and efficiently monitoring fairness with access only to coarse-grained, partial, or indirect information (e.g., neighborhood- or organization-level statistics)

> **“If we had more people who we could throw at this... ‘Can we leverage this fuzzy [coarse-grained] data to [audit]?’ that would be great [...]**
>
> **It’s a fairly intimidating research problem I think, for us.” - R21**

(cf. Kilbertus et al., 2018; Veale & Binns, 2018)

Limitations of Existing ML Methods

- ML literature generally assumes individual-level access to sensitive attributes, which many teams lack
  - Needs for support in effectively and efficiently monitoring fairness with access only to coarse-grained, partial, or indirect information (e.g., neighborhood- or organization-level statistics)

> **"We called it the SETHtimator, a sex and ethnicity estimator. [...with] one dataset, we [only] had a list of people's names and their IP addresses.**
>
> **So we were able to sort of cross-reference their IP addresses with a name database, and from there use a [classifier] to list a probability that someone with that name in that region would have a certain gender or ethnicity. [...]" - R23**

(cf. Kilbertus et al., 2018; Veale & Binns, 2018)

Biases in the Humans in the Loop

- Several interviewees mentioned biases in the humans embedded at different stages of the machine learning pipeline (e.g., crowdworkers who annotate data)
  - 69% of survey respondents said tools to reduce the influence of biases from humans in the loop would be very/extremely useful

- This contrasts the common attitude that teams should just add a human in the loop to combat undesirable biases

(cf. Kamar, Kapoor, & Horvitz, 2015)

Major Needs

- Research on how to support practitioners in "**fairness-aware" data collection and curation**

- **Application- and domain-specific** tools and resources

- Research on how to support fairness auditing given only **partial demographic information** (e.g., neighborhood- or organization-level demographics)

- Useful and usable tools for **fairness debugging**
  (e.g., determining whether a customer complaint represents a "one-off" or is indicative of a systemic issue … or diagnosing the cause(s) of particular unfair behaviors in multi-component ML systems)

- New tools and approaches for **prototyping ML systems**
  (beyond existing UX prototyping methods)

For more…

Computer Science > Human-Computer Interaction

Improving fairness in machine learning systems: What do industry practitioners need?

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, Hanna Wallach

(Submitted on 13 Dec 2018 (v1), last revised 7 Jan 2019 (this version, v2))

Improving fairness in practice requires co-design and participatory approaches to research*

"...it would be so valuable to have more researchers want to embed on certain problems with product groups … so there's a shared sense of success by solving as opposed to [...] sitting outside of the problem and critiquing it..."

- anonymous interviewee

* But external critiques can be extremely impactful!
(e.g., Buolamwini & Gebru, 2018; Raji & Buolamwini, 2019)

This 90-min tutorial

Decisions while building

The wider context

Intro

Henriette (10 mins)

Fairness throughout the machine learning lifecycle

Jenn (30 mins)

Organizational & domain challenges

Mapping industry challenges

Ken (20 mins)

A case study

Henriette (20 mins)

Recap

Henriette (5 mins)

Translation, tracks & data:

Lessons learnt while setting up an algorithmic bias effort, in a specific domain.

[Cramer et al.,
CHI'19 case study]

From a research perspective to ‘product’ perspective.

Empower teams to **assess & address algorithmic bias** and better serve **underserved audiences.**

Music.

emotional,
personal,
social,
(sub)cultural.

One very specific effort & domain.

Lessons learnt from establishing a common framework

1) Organizational activities
2) Checklists and other tools

Lessons learnt from auditing

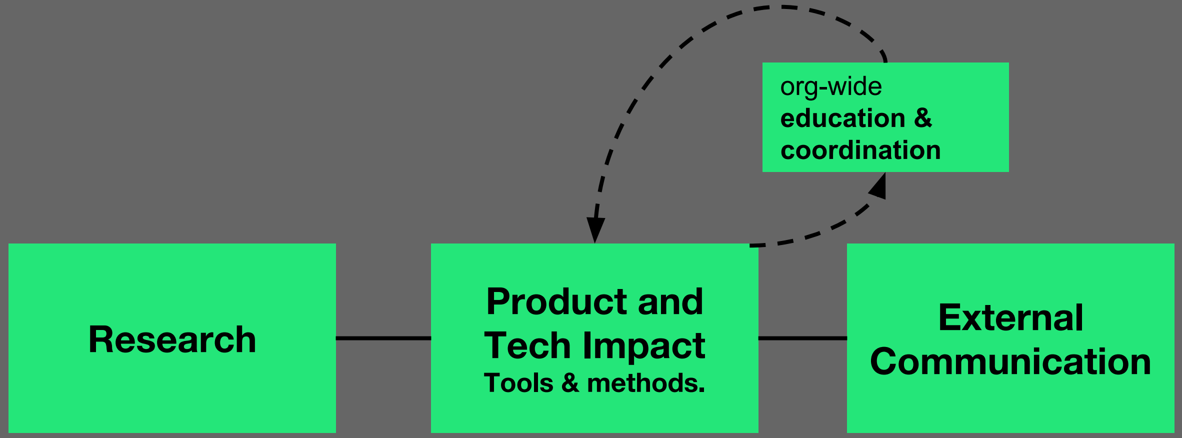
3) A case study in voice / recommendation products

A shared framework.

Any dataset, any algorithmic outcome is ‘biased’*

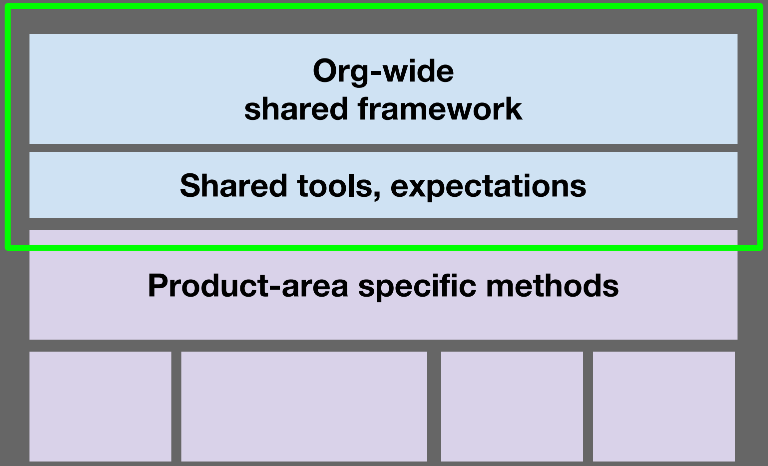
* has characteristics influenced by (non-)decisions

Algorithmic bias effort with different types of activities & talents



Shared framework & education

Complemented with specific deep-dives.



‘Checklist’ effort

First step:
help teams think concretely about ‘entry points for bias’ in their products

Springer, Garcia-Garthright, Cramer, UX of AI ’18

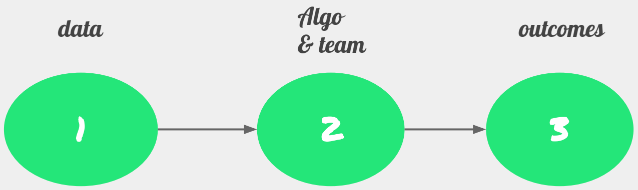
Combining existing resources into a ‘checklist’ for teams?

Internal discussions

pre-survey examples + Team diversity

Main external frameworks

Data	Social data biases (Olteanu et al.,‘16) Dataset nutrition label (Chmielinksi et al.’17) Datasheets for datasets (Gebru et al. \`18)
Models	Modelcards for model reporting (Mitchell et al. ‘18)
Outcomes	Preexisting, Technical Bias, Emergent Bias (Friedman & Nissenbaum, ‘97) Types of harm (Crawford’17)
Cycle	Bias on the Web cyclical model (Baeza-Yates ‘16) ML Life cycle. (Wallach & Wortman Vaughan ‘19)



- (un)intended characteristics?
- How would you assess this?
- Prioritization & action items?

Springer, Garcia-Garthright, Cramer,
UX of AI ’18, ACM Interactions ‘18

DATA

- **Why was the dataset created?** (e.g., was there a specific intended task gap that needed to be filled?)
- **Who funded the creation of the dataset?**
- **What preprocessing/cleaning was done?** (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances)
- **If it relates to people, were they told what the dataset would be used for and did they consent?** If so, how? Were they provided with any mechanism to revoke their consent in the future or for certain uses?
- **Will the dataset be updated?** How often, by whom?

Gebru et al, '18
Datasheets for datasets

Dataset Fact Sheet

Metadata

Cj CC-0 .csv

Title COMPAS Recidivism Risk Score Data

Author Broward County Clerk's Office, Broward County Sherrif's Office, Florida

Email browardcounty@florida.usa

Description Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat.

DOI 10.5281/zenodo.1164791

Time Feb 2013 - Dec 2014

Keywords risk assessment, parole, jail, recidivism, law

Records 7214

Variables 25

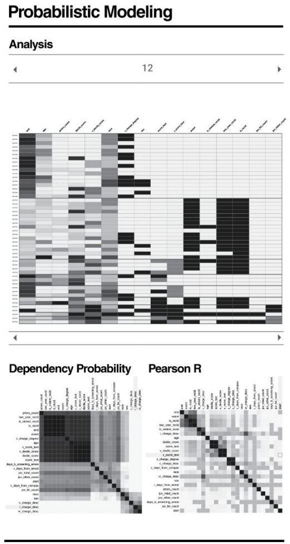
priors_count: *Ut enim ad minim veniam, quis nostrud exercitation* — numerical

two_year_recid: *Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.* — nominal

. . .

Missing Units 15452 (8%)

This dataset contains variables named "age," "race," and "sex."



Chmielinksi et al, '18
datanutrition.media.mit.edu

MODELS

Model Card

- **Model Details**. Basic information about the m
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results

Mitchell et al, '18
Modelcards for model reporting

We tried to summarize this all …

DATA

Population bias: Are there differences between the data population's demographics [...] and the target population?

Behavioral bias: Are there differences in user behavior across platforms (mobile, voice?) or contexts (work, party, family) [...]

Temporal bias Are there differences in populations or behaviors over time?

Redundancy Are there data items that appear in multiple copies, or are near duplicates, or happen artificially often (bots)?

Content production bias Are there lexical, syntactic, semantic, or structural differences in how content is produced vs the content that you want to surface?

Linking bias Are there differences in the attributes of networks, or user connections that affect your data?

Interface Bias Are there biases that result from UI design or presentation? (e.g. position/ranking bias)

Sampling Biases: Are there any biases resulting from data sampling choices?

Self-Selection Bias: Who would *not* participate in this product?

ALGO & TEAM

Algorithmic parameters bias
Do you expect any side-effects from your model, and (hyper) parameter choices?

Team composition
Are there any knowledge/experience gaps within the team, i.e. would you be able to recognize 'obvious' problems?

OUTCOMES

CONTENT/CREATOR OUTCOMES
Which **content gaps*** are intended or expected?[…]

Which unintended content gaps do you want to avoid / test for?

USER OUTCOMES
Which **performance or satisfaction gaps** are intended or expected?I.e. for which users is this going to work very well, and for whom will it not *[..]*?

What do you want to avoid/ test for?

Incl. aspects from a.o:
-Social data biases (Olteanu et al.,'16)
-Bias on the Web (Baeza-Yates '16)
-Types of harm (Crawford'17)
-Dataset nutrition label (Chmielinksi et al.'17)
-Datasheets for datasets (Gebru et al. `18)

We tried to summarize this all …

Incl. aspects from a.o:
-Social data biases (Olteanu et al.,'16)
-Bias on the Web (Baeza-Yates '16)
-Types of harm (Crawford'17)
-Dataset nutrition label (Chmielinksi et al.'17)
-Datasheets for datasets (Gebru et al. `18)

DATA

Population bias: Are there differences between the data population's demographics [...] and the target population?

Behavioral bias: Are there differences in user behavior across platforms (mobile, voice?) or conte

Temporal bias Are there differe
over time?

Redundancy Are there data ite
or are near duplicates, or happe

Content production bias Are t
structural differences in how co
that you want to surface?

Linking bias Are there differen
user connections that affect you

Interface Bias Are there biases
presentation? (e.g. position/ranking

Sampling Biases: Are there any biases resultir
sampling choices?

Self-Selection Bias: Who would *not* participate in this product?

ALGO & TEAM

Algorithmic parameters bias
Do you expect any side-effects from your model, and (hyper) parameter choices?

expected?I.e. for which users is this going to work very well, and for whom will it not *[..]*?

What do you want to avoid/ test for?

Simplify, and simplify some more.
A didactic tool isn't necessarily a practical day-to-day tool.

General frameworks educate, but do not surface domain-specific priorities or goals to help decision making.

{Data, model/API, product} ownership is just as important; who can fix / break things?

Help teams figure out which subpopulations and outcomes to focus on

Streaming outcomes.
Representation.

Creator streams

Gender
Popularity
Genre
Locality

Avriel Epps

Does this product work for listeners?

Music taste
Subculture
Gender
Age
New Markets

Jasmine McNealy

From data engineering to data auditing

What can you make centrally accessible?

Pipelines
Dashboards

OUTCOMES

CONTENT/CREATOR OUTCOMES
Which **content gaps*** are intended or expected?[…]

Which unintended content gaps do you want to avoid / test for?


Lessons learnt from auditing and dashboarding

[Cramer et al.,
CHI’19 case study]

Practical and scalable models are also needed

MODELS

OUTCOMES

Demonstrating positive (or at least non-negative) impact

Explore, Exploit, and Explain: Personalizing Explainable Recommendations with Bandits

James McInerney, Ben Lacker, Samantha Hansen, Karl Higley, Hugues Bouchard, Alois Gruson, Rishabh Mehrotra

Towards a Fair Marketplace: Counterfactual Evaluation of the trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems

Rishabh Mehrotra[1], James McInerney[1], Hugues Bouchard[1], Mounia Lalmas[1], Fernando Diaz[2*]
[1]Spotify Research, [2]Microsoft Research
{rishabhm,jamesm,hb,mounial}@spotify.com,diazf@acm.org

ABSTRACT

Two-sided marketplaces are platforms that have customers not only on the demand side (e.g. users), but also on the supply side (e.g. retailer, artists). While traditional recommender systems focused specifically towards increasing consumer satisfaction by providing relevant content to consumers, two-sided marketplaces face the problem of additionally optimizing for supplier preferences, and visibility. Indeed, the suppliers would want a *fair* opportunity to be presented to users. Blindly optimizing for consumer relevance may have a detrimental impact on supplier fairness. Motivated by this problem, we focus on the trade-off between objectives of consumers and suppliers in the case of music streaming services, and consider the trade-off between *relevance* of recommendations to the consumer (i.e. user) and *fairness* of representation of suppliers (i.e. artists) and measure their impact on consumer *satisfaction*.

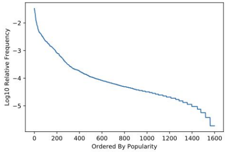


Figure 1: Exposure of artist playlists on a music app. A small number of artists receive the highest relevance score for most users.

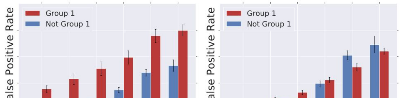
1 INTRODUCTION

Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, Ed H. Chi
{alexbeutel, jilinc, tulsee, hqian, woodruff, cmluu, kreitmann, bischof, edchi}@google.com
Google

Abstract

As more researchers have become aware of and passionate about algorithmic fairness, there has been an explosion in papers laying out new metrics, suggesting algorithms to address



Challenges in showing data & assessing ‘fairness’

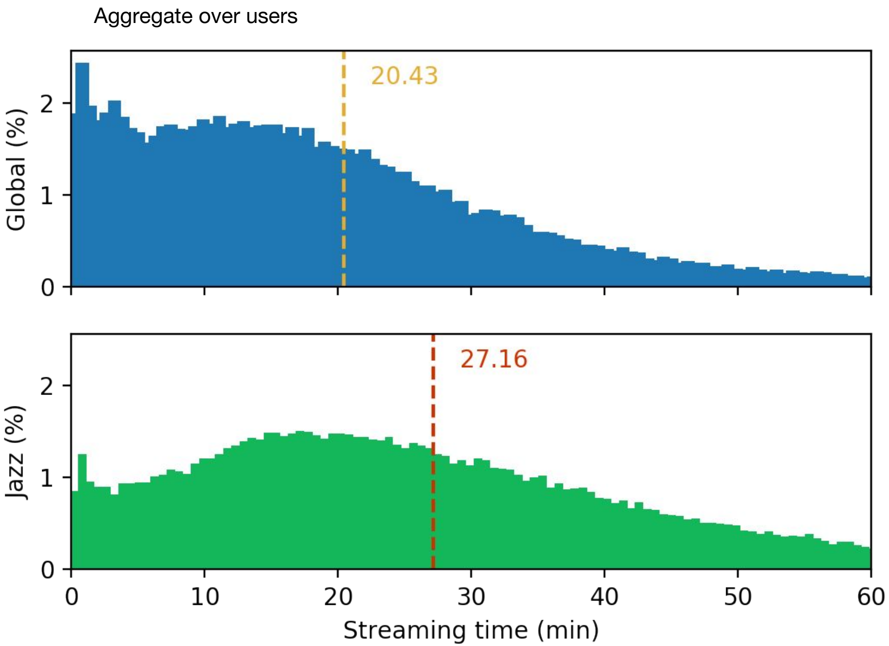
Some content gaps & biases are **intentional:**

- New music playlists: recency bias

Some content gaps & biases can be argued to be **unfair:**

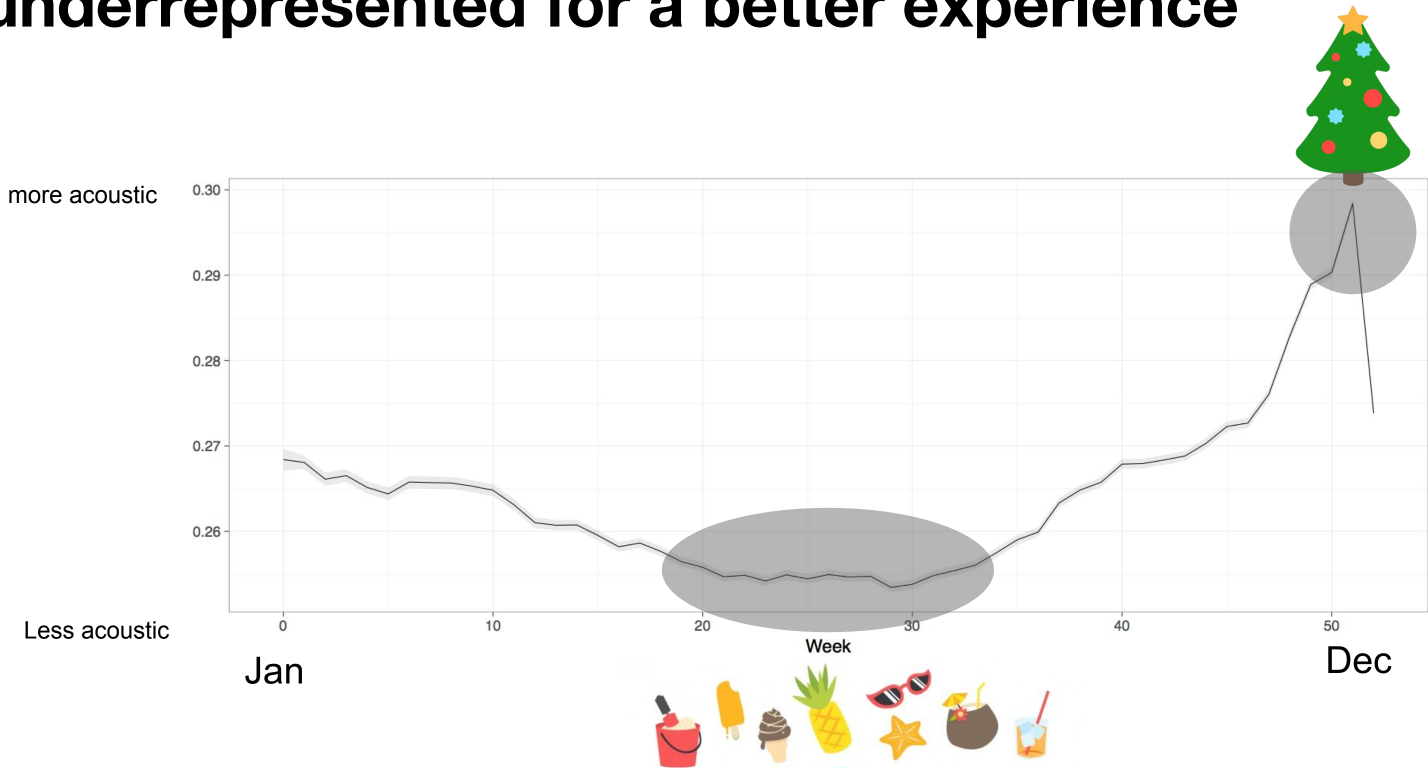
- Under-index of certain genres over others

‘Success’ metrics differ between groups & genres



Jazz listeners consume Jazz and other playlists for longer period than average.

Sometimes genres *should* be underrepresented for a better experience

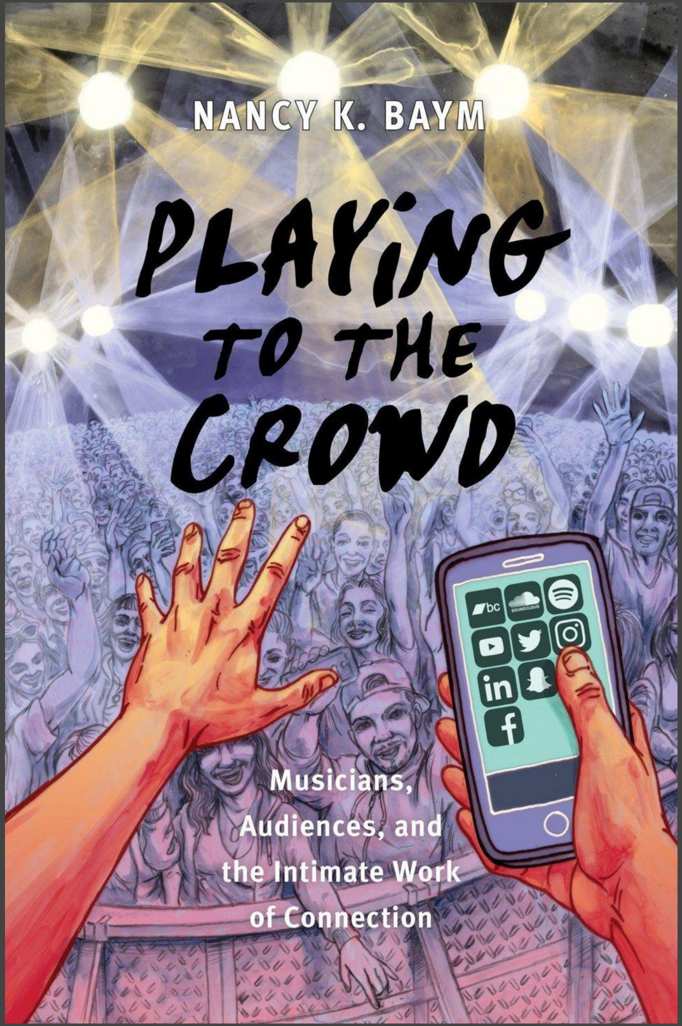


Minsu Park,Jenn Thom, Henriette Cramer, Sarah Mennicken, Michael Macy.
Nature Human Behavior ‘19

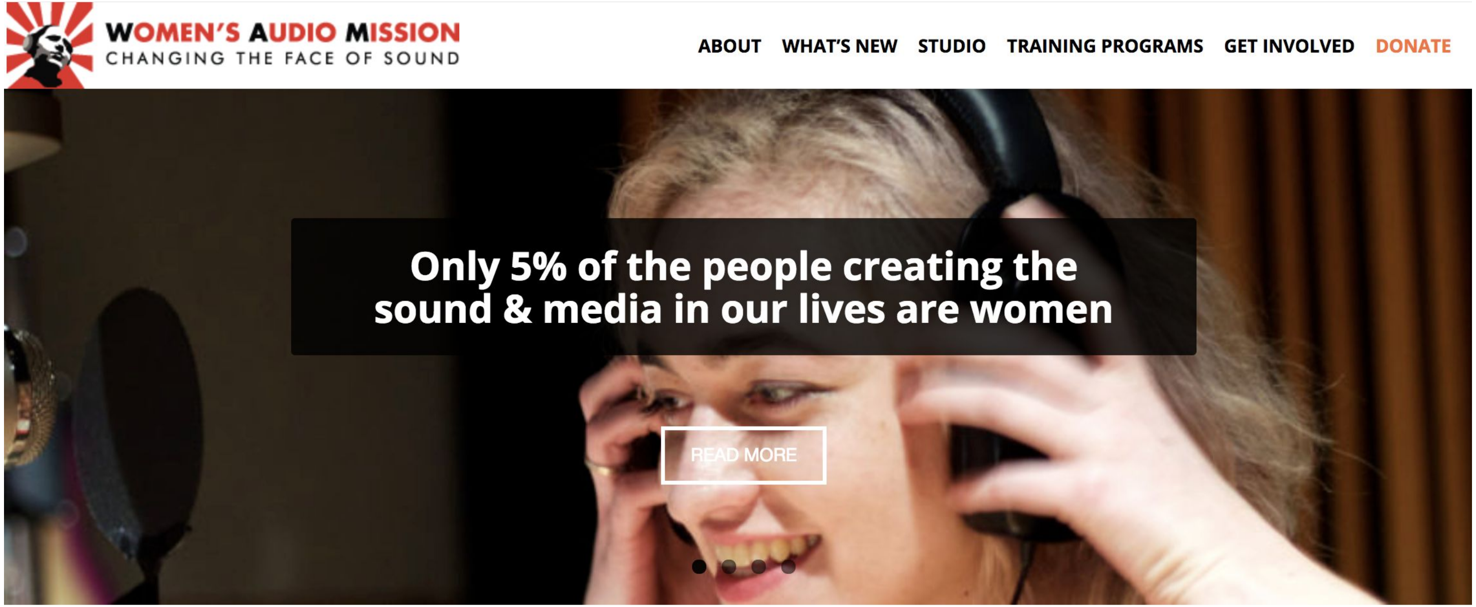
We also need to measure long-term impact

Being recommended once, vs. gaining a lifelong fan.

This should influence prioritization & measurement.



Know your baselines.

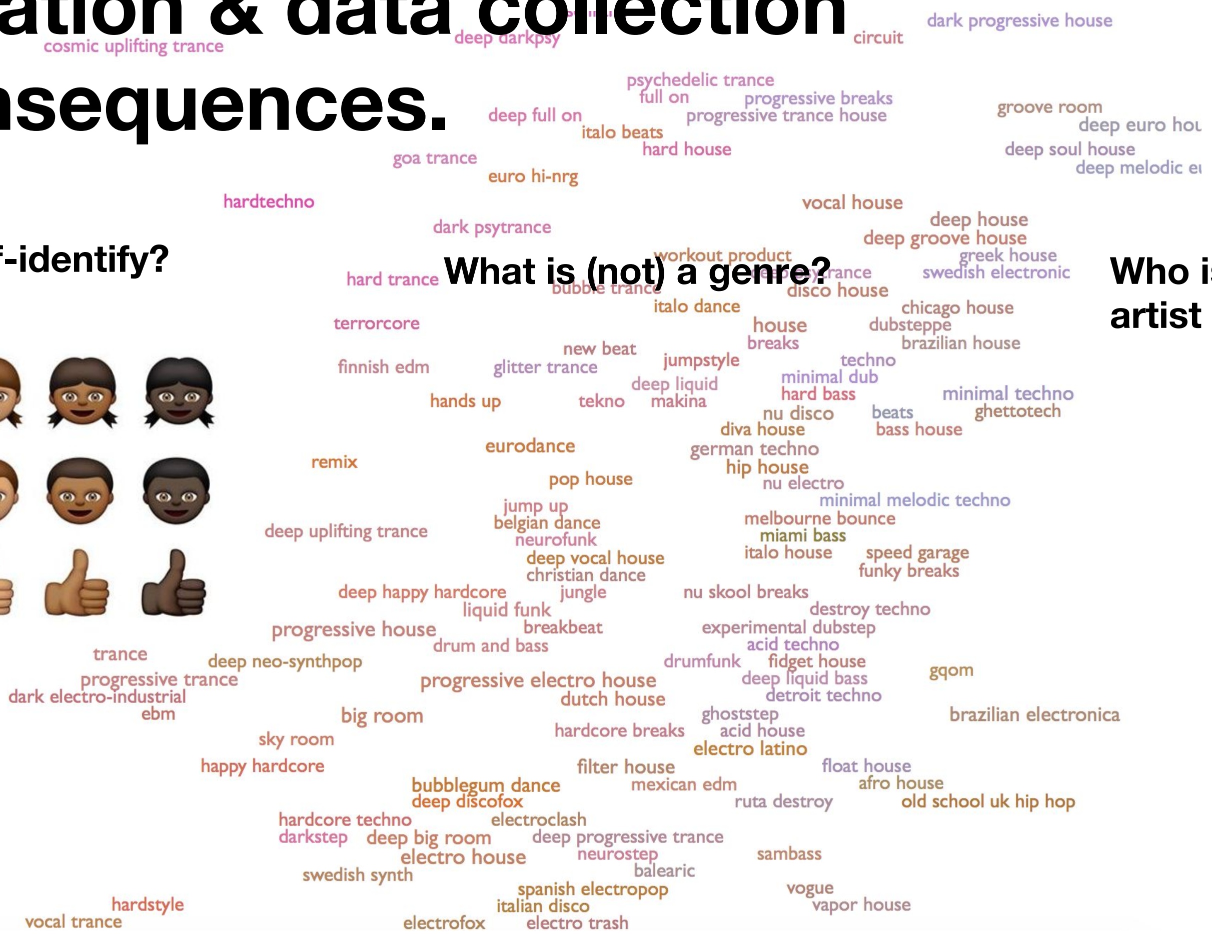


- The music industry isn't balanced.
- Comparing 'recommended' to 'explicitly asked for' is one baseline
- Data will be missing on intersections with popularity. This can misrepresent results if you don't show missing data.

Classification & data collection have consequences.

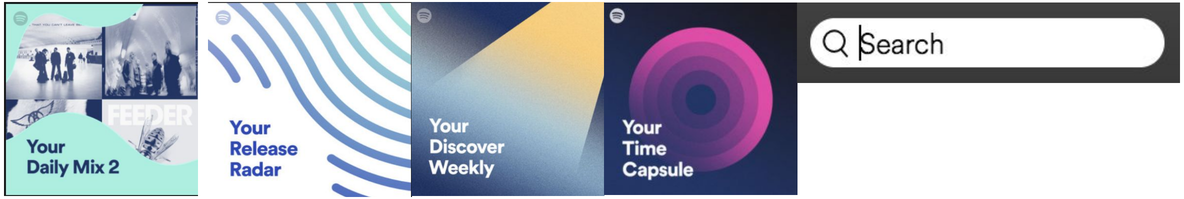
Don’t collect? Self-identify? ‘Internationalize’?

What is (not) a genre?



Who is a ‘local’ artist or not?

Machines don’t know what machines don’t know.
You need an human perspective.



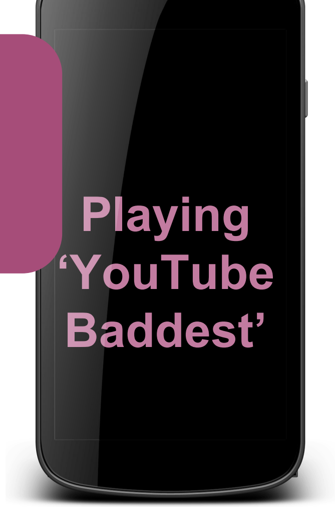
algotorial = algorithmic + editorial

Editors
Data curators
Employee resource groups
Product teams & grassroots reports

Auditing case study: voice

Dialects

Play you da baddest



Creative expression

Play PRBLMS by 6lack

NAVUZIMETRO

Lil Uzi Vert

PΣNT▲GR▲MΦPHΦNΣ

V▲LH▲LL

oOoOO

Sunn O)))

Multilingual & code switching

Play Dile Que Tu Me Quieres

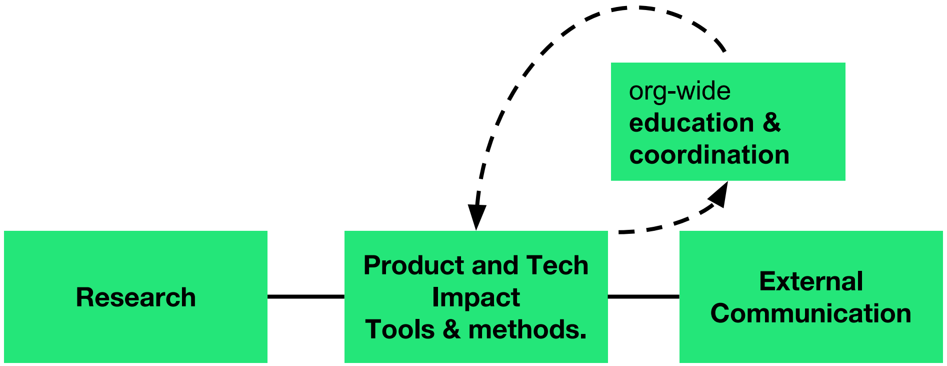
Springer & Cramer, CHI'18

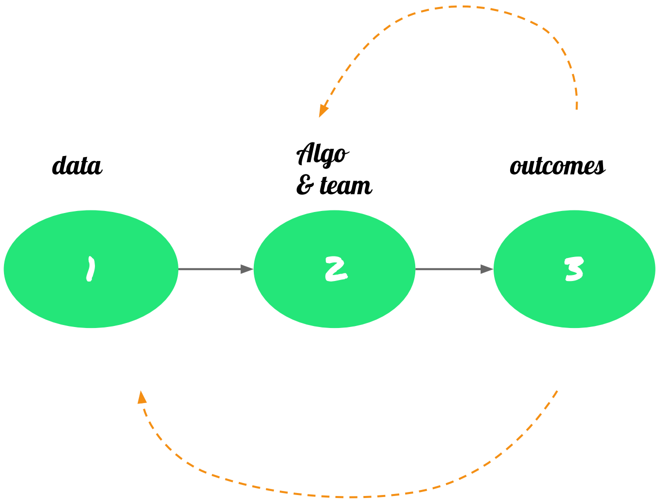
Log-voice-findability = log(streams/voice finds)

Building models for each of these issues is time-intensive & fragile.

Your detection methods will have limitations too.

Track popularity rank





Lessons Learnt

Domains like audio & music & entertainment you also run into big open challenges.

Self-serve tools are useful, but org-wide tools help track concrete impact; and require lots of hidden work.

All tools were helpful, and inspiration, but there was still a gap in checklist, dashboarding, case studies (open research challenge!).

This 90-min tutorial

Intro

Henriette
(10 mins)

Decisions while building

Fairness throughout the machine learning lifecycle

Jenn (25 mins)

The wider context

Organizational & domain challenges

Mapping industry challenges

Ken (20 mins)

A case study

Henriette (20 mins)

Recap

Henriette
(5 mins)

To recap this tutorial

Decisions while building

Fairness throughout the machine learning lifecycle

+

The wider context

Organizational & domain challenges

Mapping industry challenges

A case study

Decisions are made at every point of the pipeline.

Those decisions need support.

Concrete examples or pragmatic advice help.

Organizational work is as crucial as advanced ML-methods.

Shared frameworks / checklists are useful didactics, but each product & domain needs specific methods.

A lot of issues are 'known'. That doesn't mean there is easy-to-digest advice available for practitioners.

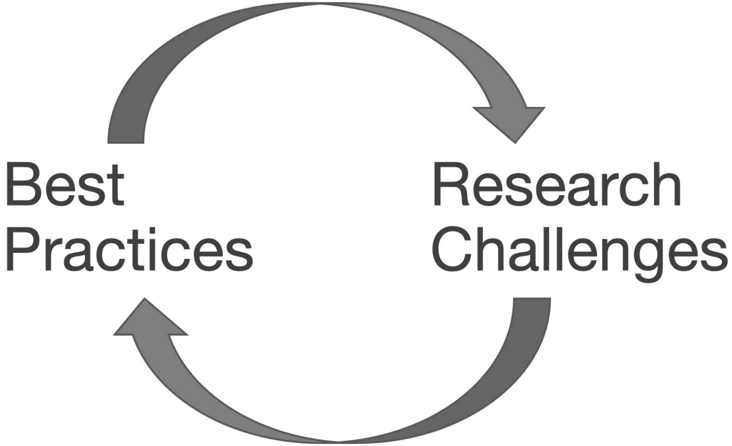
What if ... ?

Assessing & addressing algorithmic bias requires navigating uncertainty.

Who to involve, what to prioritize, how to assess & address, & predicting interventions' impact.

Enable organizing & sharing.

Let's make the community + work accessible.



Practitioner? Please come chat with us!