Fair machine learning in industry: from research to practice

Jean Garcia-Gathright
WebSci ‘19 workshop on Handling Web Bias
June 30, 2019
Boston, MA

Music.
emotional,
personal, social,
(sub)cultural.

Avoiding unintended algorithmic harms.

Allocation
Incl. attention, streams, $

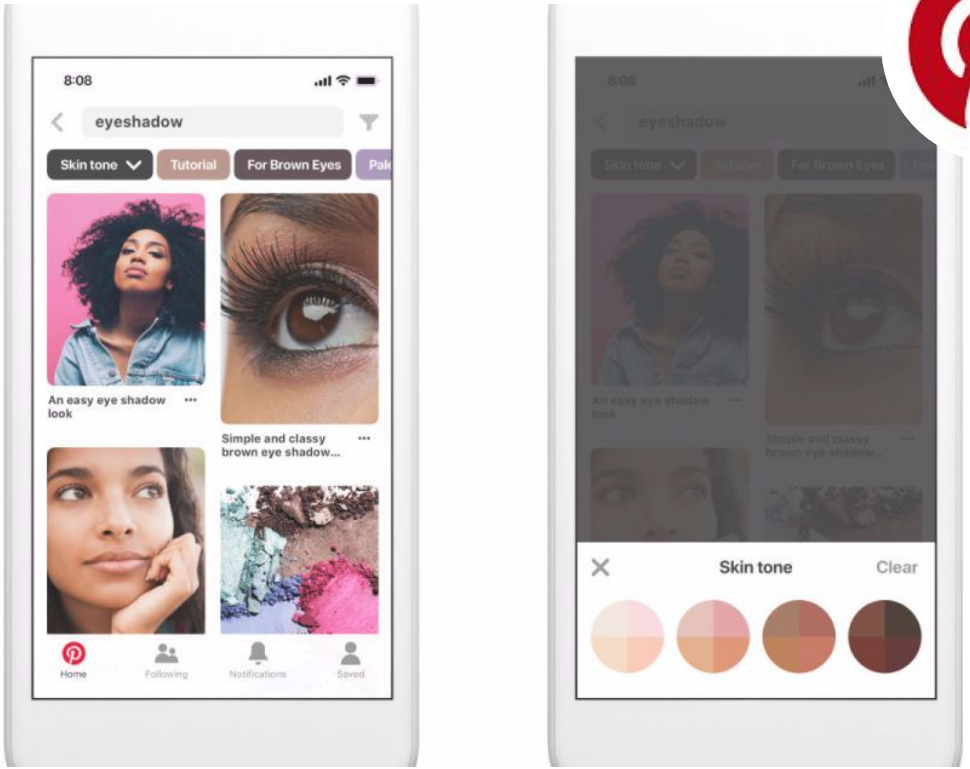
Representation
Incl. stereotypes

Not only legally protected classes, **also categories** like location, interests, (sub)culture. content etc.

Crawford, K. The Trouble With Bias. Keynote at NeurIPS 2017

Why does this matter to practitioners?

1. Better product, serving wider audience(s)



2. Responsibility, social impact & PR

IDEAS • THE ART OF OPTIMISM

Artificial Intelligence Has a Problem With Gender and Racial Bias. Here's How to Solve It

BY JOY BUOLAMWINI FEBRUARY 7, 2019

3. Legal & policy

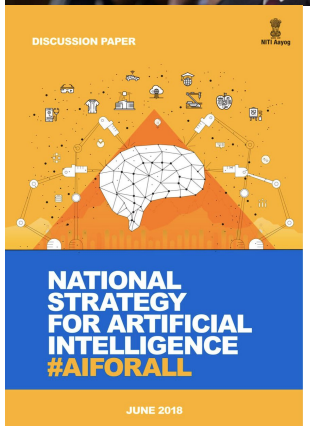
Artificial intelligence: Commission outlines a European approach to boost investment and set ethical guidelines

Brussels, 25 April 2018

MACHINE PERSPECTIVES

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

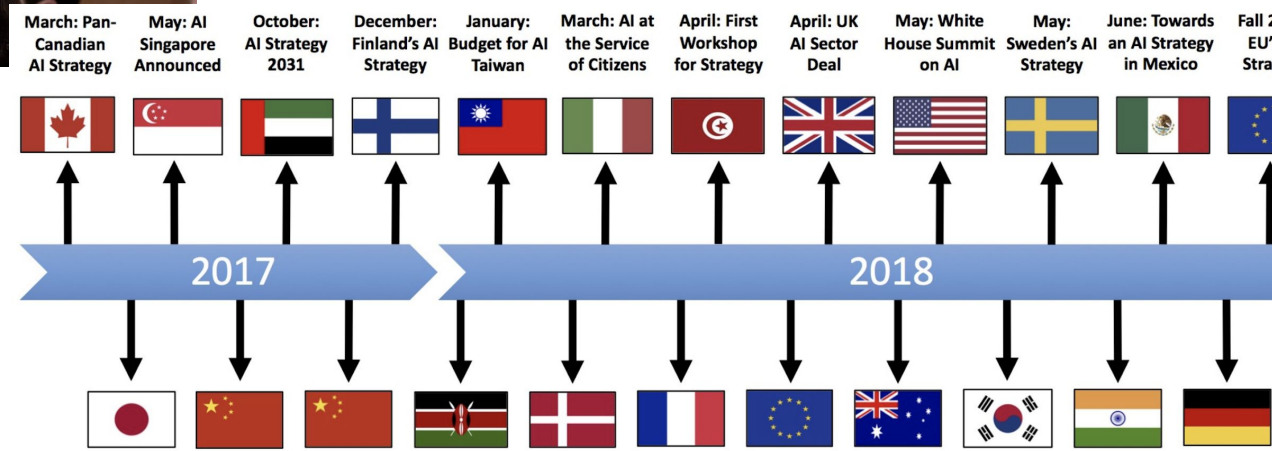
By Dave Gershgorn • September 21, 2018



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.

Guillaume Chaslot @gchaslot · Feb 9

YouTube announced they will stop recommending some conspiracy theories such as flat earth.

I worked on the AI that promoted them by the *billions*.

Here is why it's a historic victory. Thread. 1/

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

By Dave Gershgorn · May 3, 2018

We need to translate research on algorithmic bias to practical approaches for product teams.

Decisions made during development

Fairness throughout the machine learning lifecycle

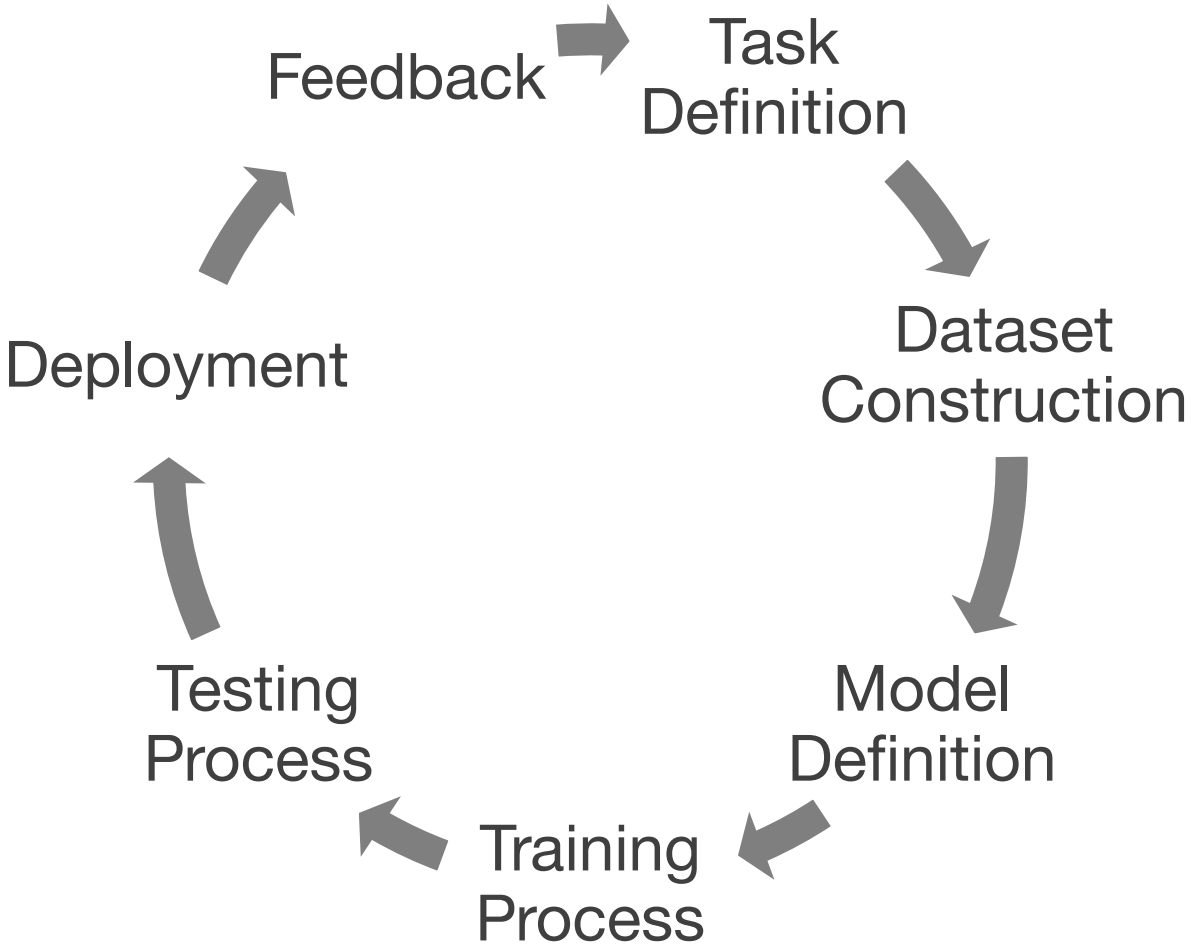
+

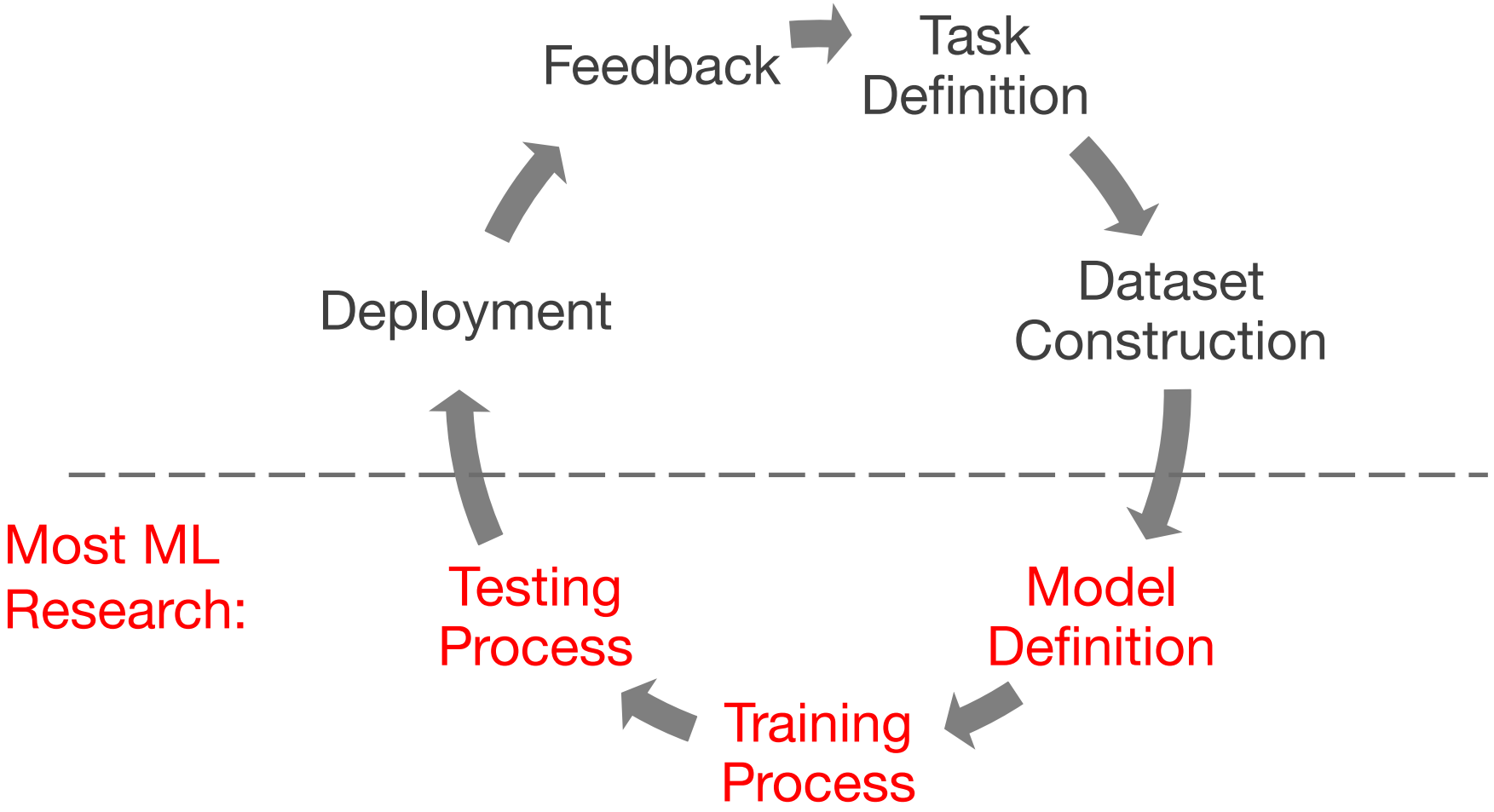
The wider context

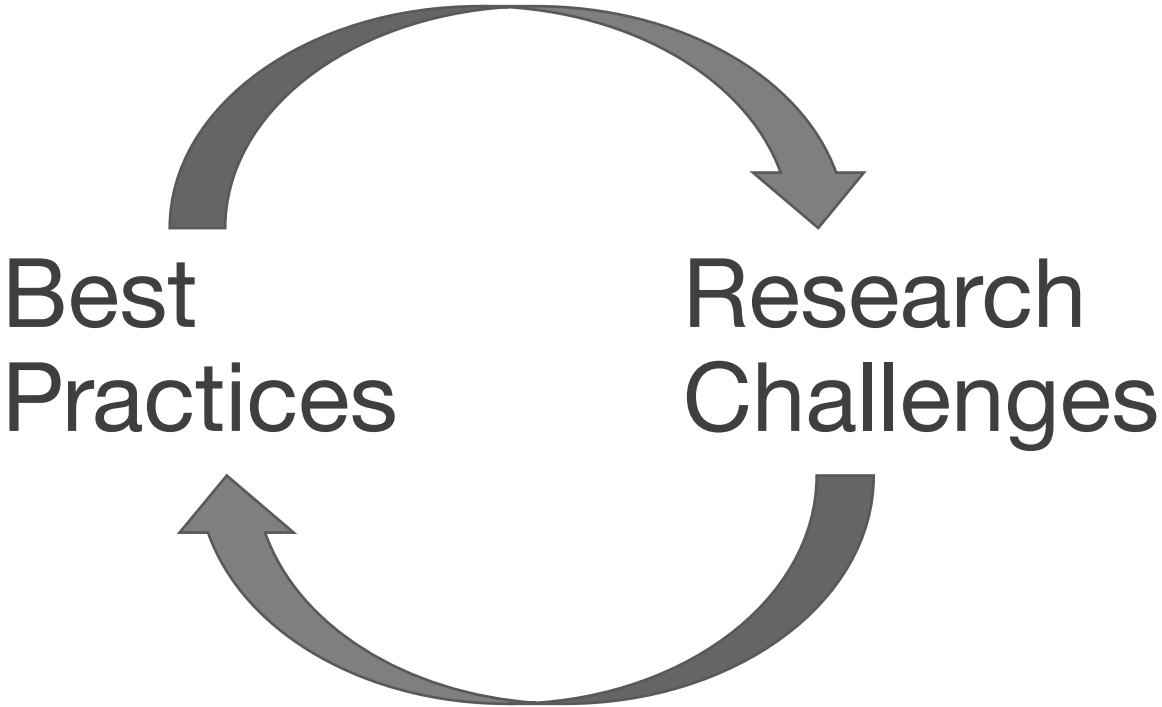
Organizational and domain-specific challenges

Spotify

Fairness throughout the machine learning life cycle


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

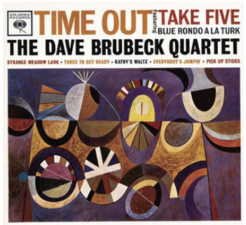

Best
Practices
Research
Challenges

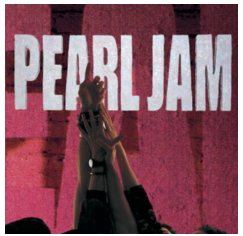
Cool new idea:
classify genre based on album art

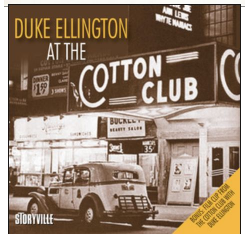
Rock

Jazz

Cool new idea:
classify genre based on album art



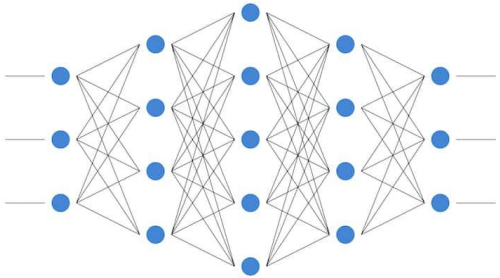




Rock

Jazz

Training data:
images and labels



Model:
convolutional neural network

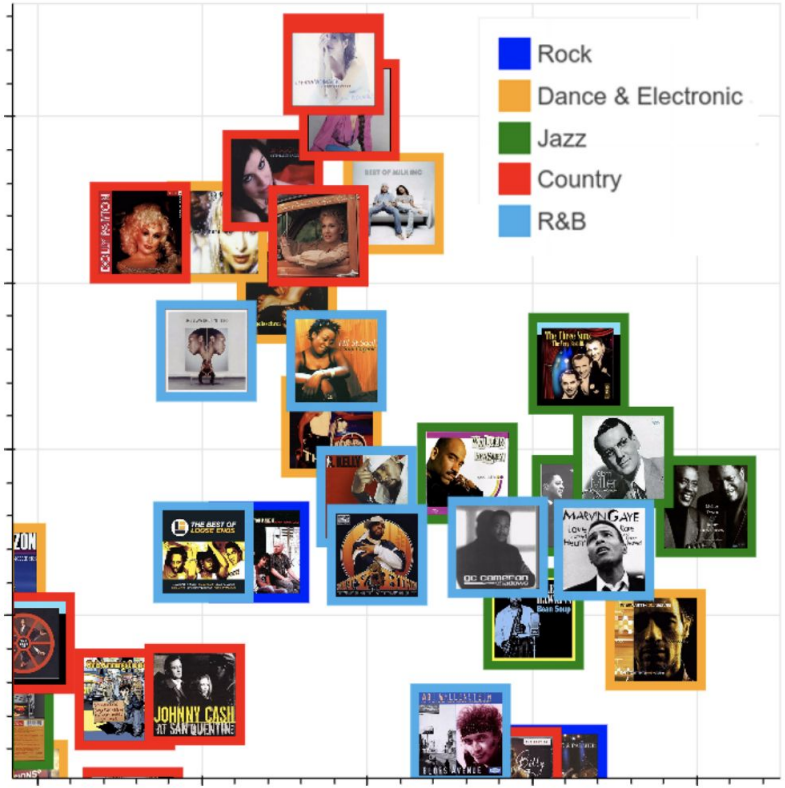
Prediction: Jazz

Prediction: Rock

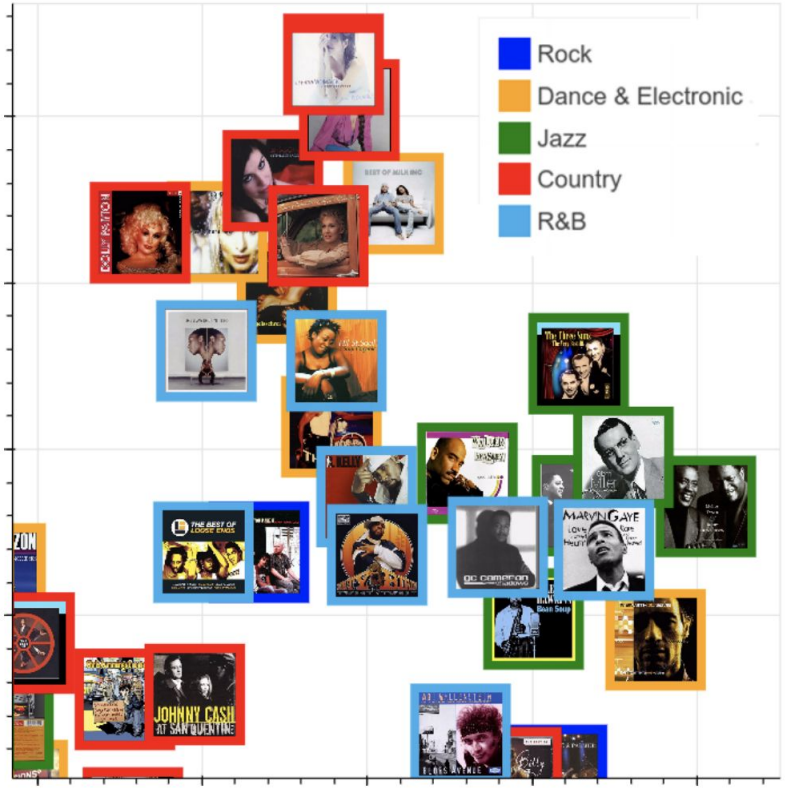
100% accurate!

Testing:
predict labels on hold out set and calculate performance metrics

Scale it up and see what happens!



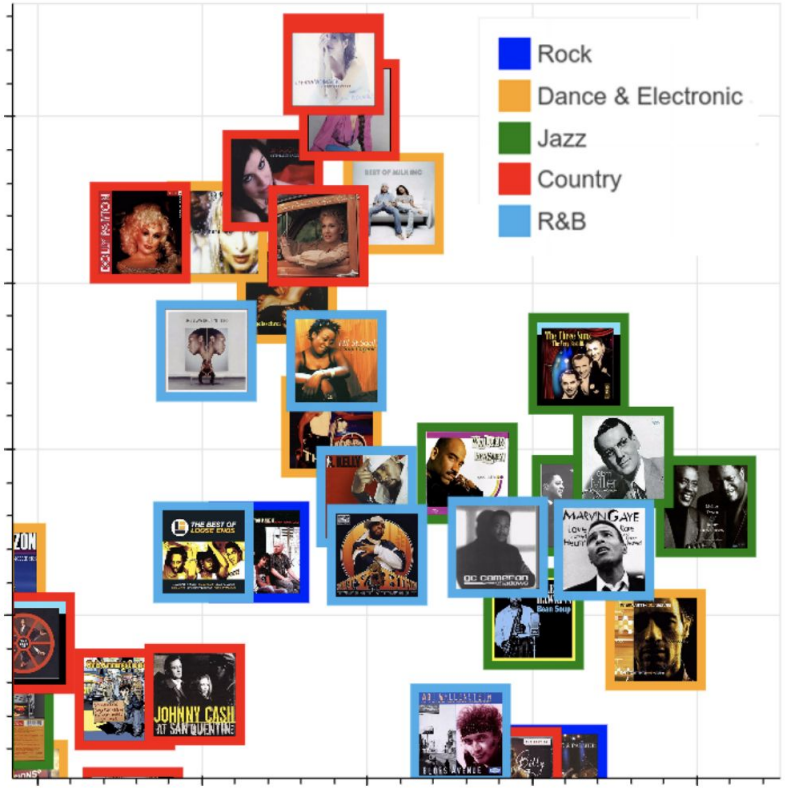
Scale it up and see what happens!



Uh-oh… the "genre" classifier is actually an artist face classifier

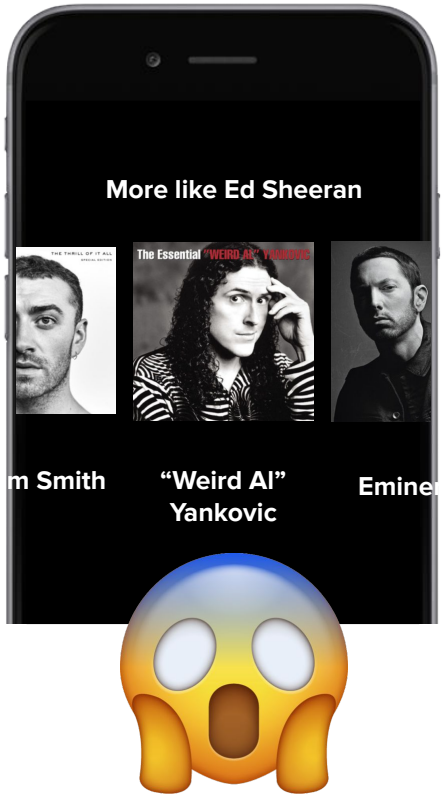
Multi-label genre classification from audio, text, and images using deep features. ICML 2017.

Scale it up and see what happens!

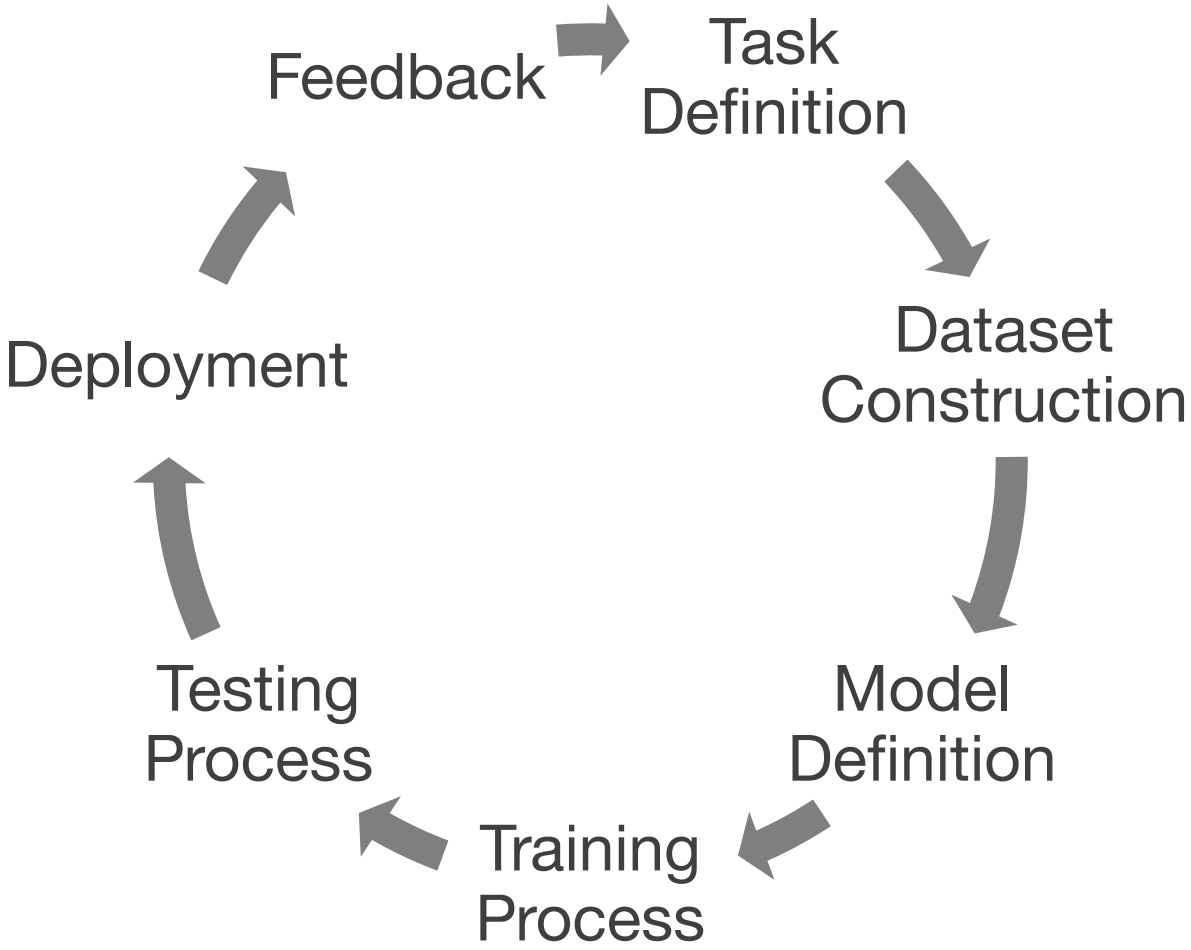


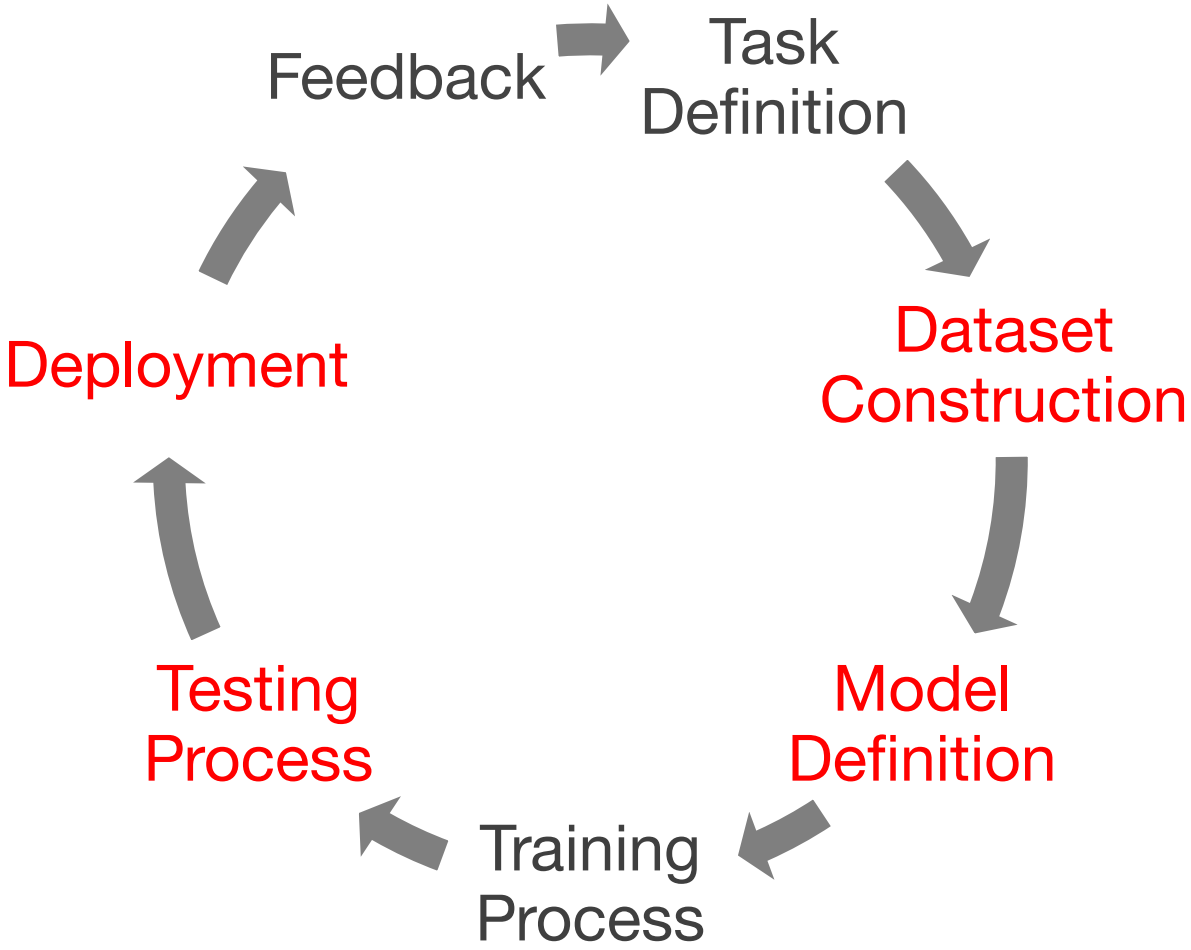
Uh-oh… the “genre” classifier is actually an artist face classifier

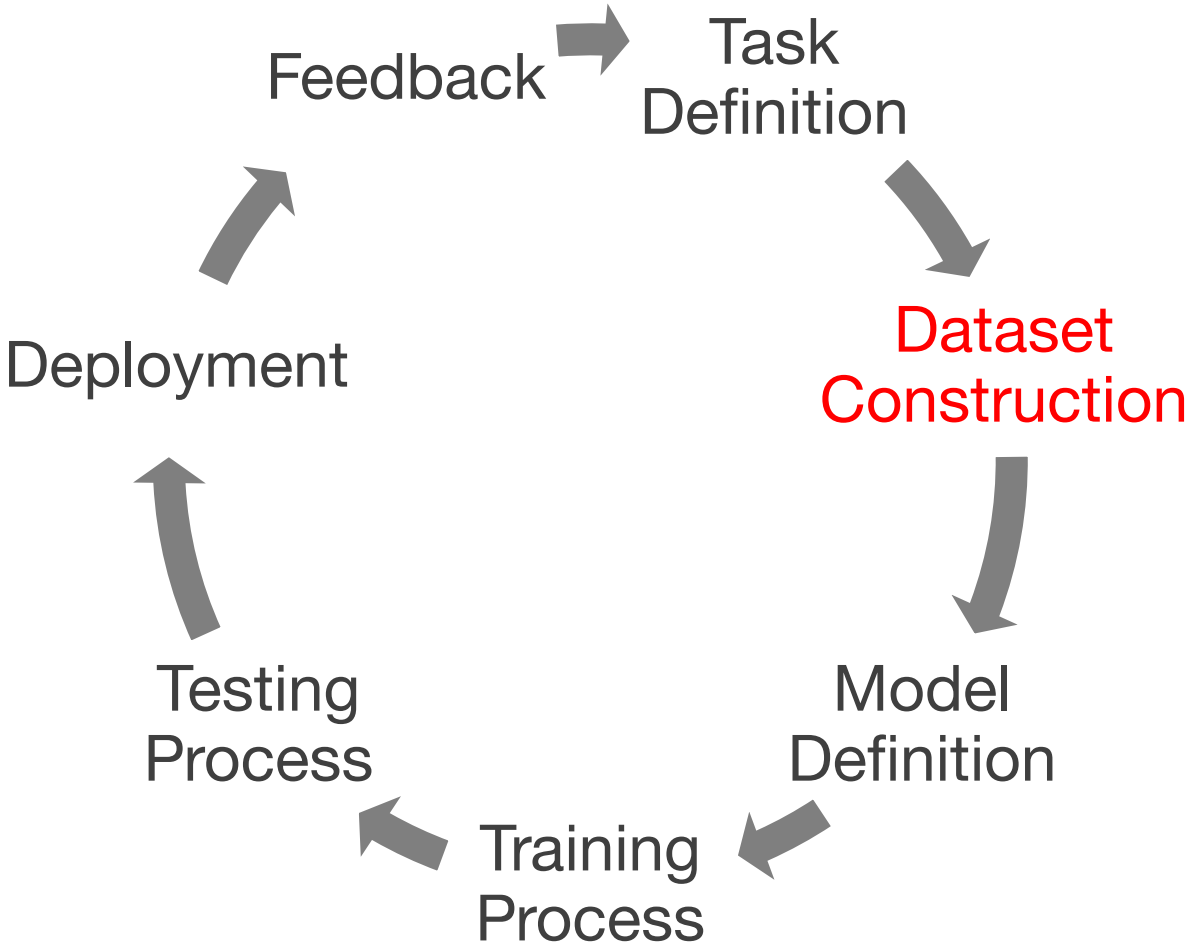
Multi-label genre classification from audio, text, and images using deep features. ICML 2017.



What can we do
to avoid this?


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


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


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

Best Practices: Data

Identify societal and cultural biases in data source

What do we (not) amplify?

Best Practices: Data

Identify societal and cultural biases in data source

What do we (not) amplify?

Ensure sufficient representation of subpopulations

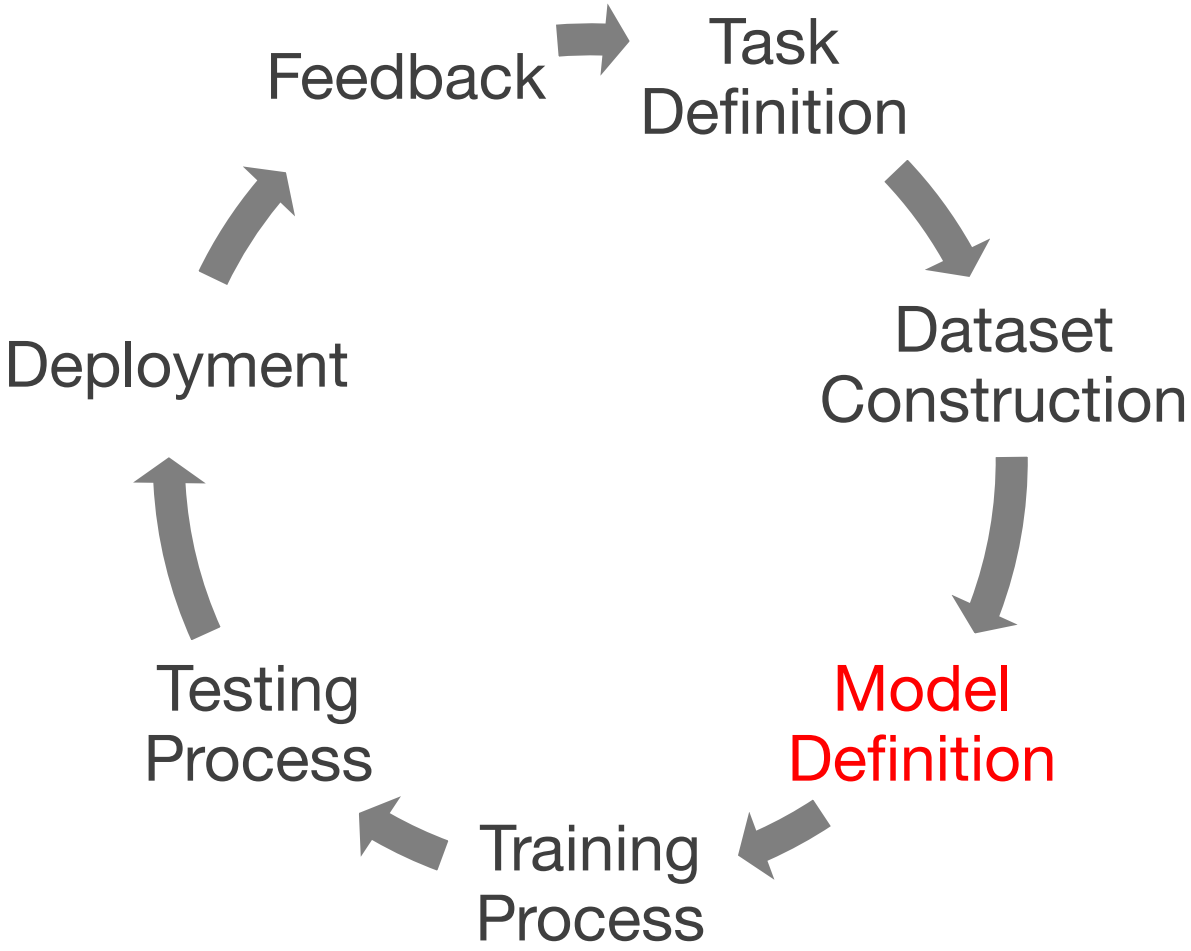
Seek variety in training examples.

Research challenges: Data

Methods and tools for “fairness-aware” data collection and curation?

What constitutes “sufficient representation” of subpopulations?

What is “fair and ethical” data collection?


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

Best practices: Model Definition

Clearly define the assumptions of the model

Fairness through unawareness?

Best practices: Model Definition

Clearly define the assumptions of the model

Fairness through unawareness?

Consider an inspectable model

(a) Original Image

(b) Explaining *Electric guitar*

"Why Should I Trust You?" Explaining the Predictions of Any Classifier. KDD 2016.

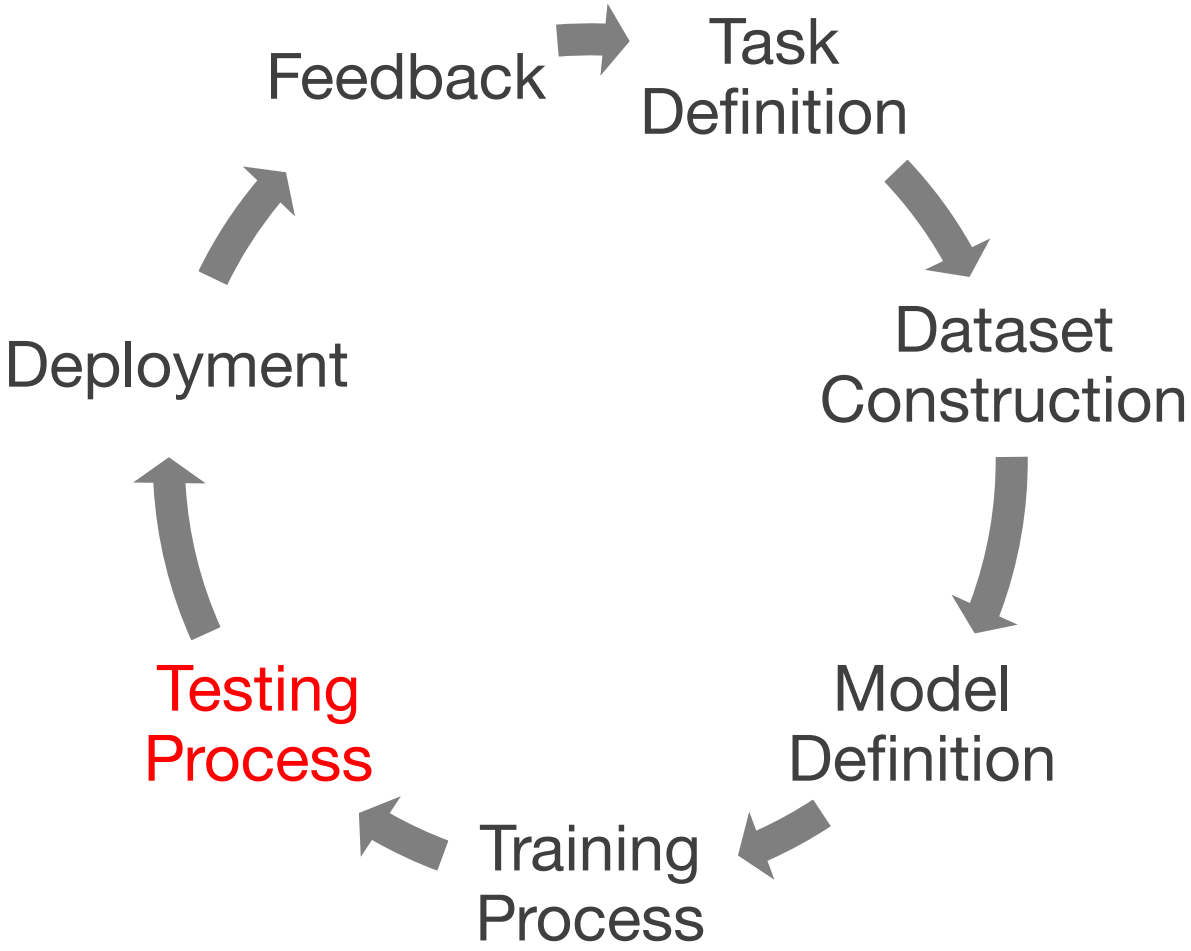
Research challenges: Model definition

What are the biases in common modeling assumptions in a particular domain?

How might “fairness” be included in the objective function?*

Move beyond supervised learning?

*Corbett-Davies and Goel, 2018. The measure and mismeasure of fairness: A critical review of fair machine learning.


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

Best Practices: Testing

Ensure sufficient representation of subpopulations

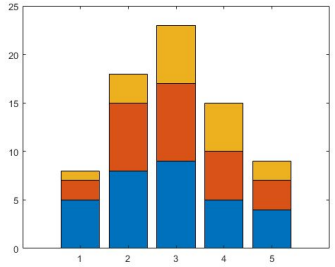
Seek variety in test examples.

Best Practices: Testing

Ensure sufficient representation of subpopulations

Seek variety in test examples.

Consider a variety of evaluation metrics, including “fairness”



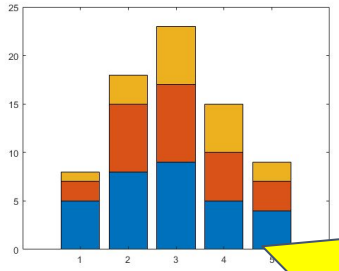
Gender shades: Intersectional accuracy disparities in commercial gender classification. PMLR 2018.

Best Practices: Testing

Ensure sufficient representation of subpopulations

Seek variety in test examples.

Consider a va… evaluation m… including "fairne…



Gender shades: Intersectional acc… disparities in commercial ge… classification. PMLR 2018.

Fairness is a non-trivial socio-technical challenge

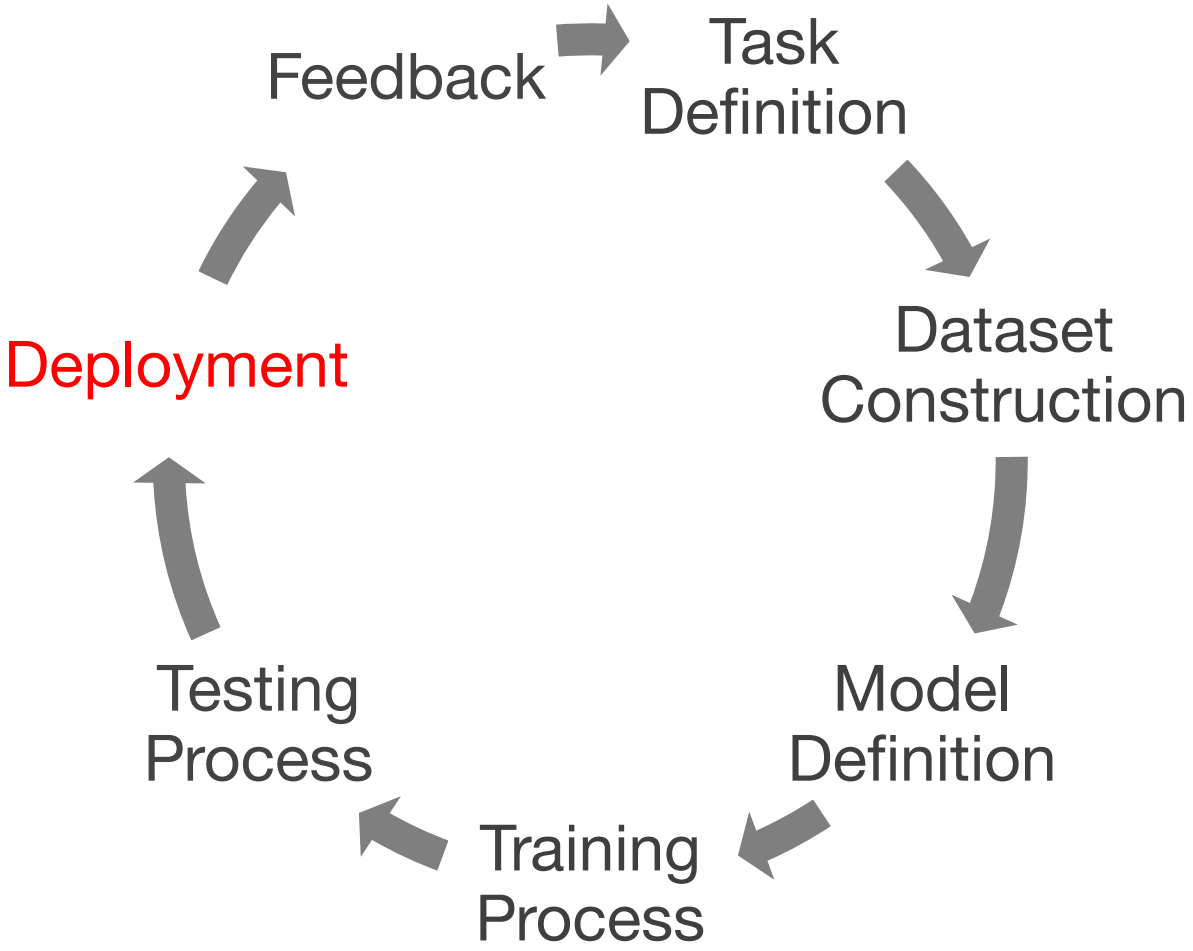
No free lunch: can't satisfy all metrics

Research challenges: Testing

What are the subpopulations of interest for testing?

Which fairness metrics are appropriate in which scenarios?

What are the right fairness metrics for complex systems?


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

Best Practices: Deployment

Check that training and test data match deployment context

New markets?

Best Practices: Deployment

Check that training and test data match deployment context

New markets?

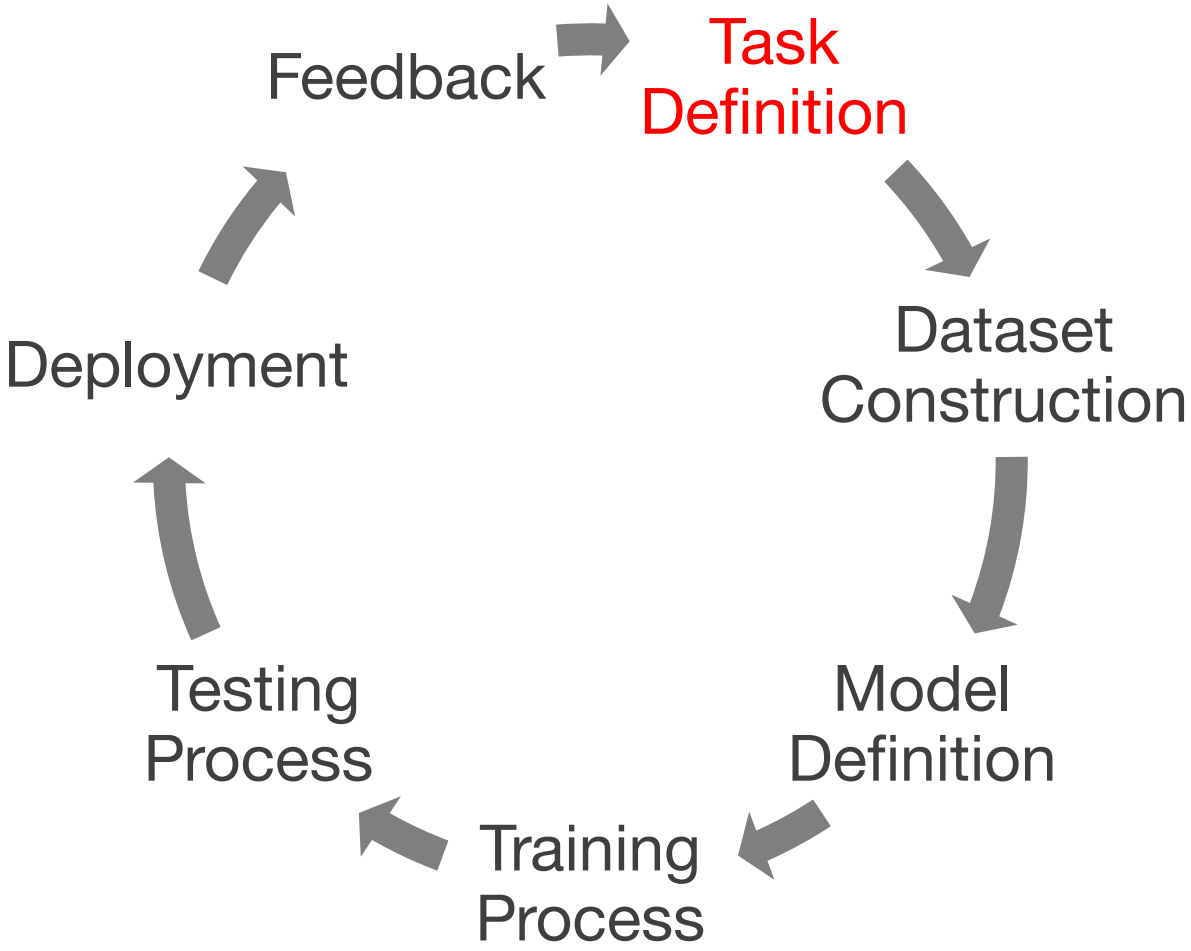
Monitor user reports and complaints

Research challenges: Deployment

Methods and tools to audit for shifts in population?

Methods and tools to determine whether an error is indicative of a systemic problem?

Audit existing system for biases (in collaboration with the teams that built the systems)


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

Best Practices: Task Definition

Clearly define the task; understand the model's (un)intended effects

Best Practices: Task Definition

Clearly define the task; understand the model's (un)intended effects

Refine the task definition & be willing to abandon

Research challenges: Task definition

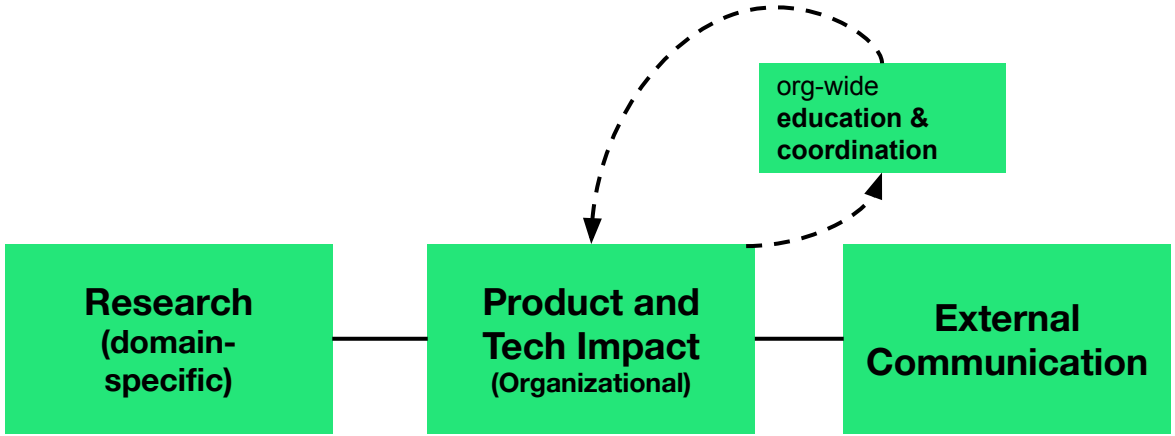
How should decisions be made about which tasks to pursue and which to avoid?

How should we design processes for uncovering unintended effects and biases before development?

What are the most effective ways to elicit diverse opinions?

Organizational and domain-specific challenges

An algorithmic bias mitigation effort in practice

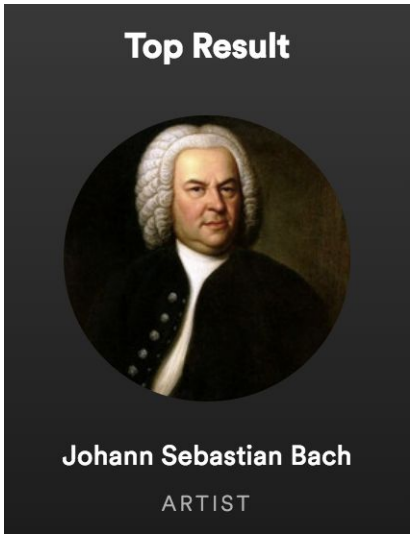


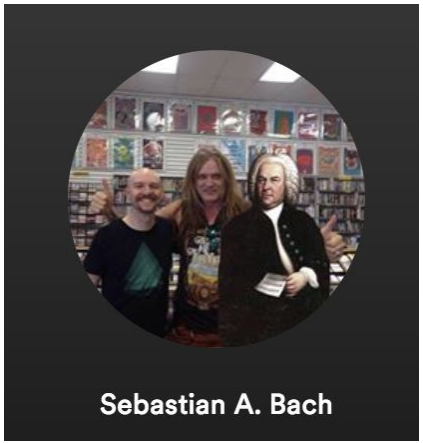
Domain-specific challenges: case study in Voice

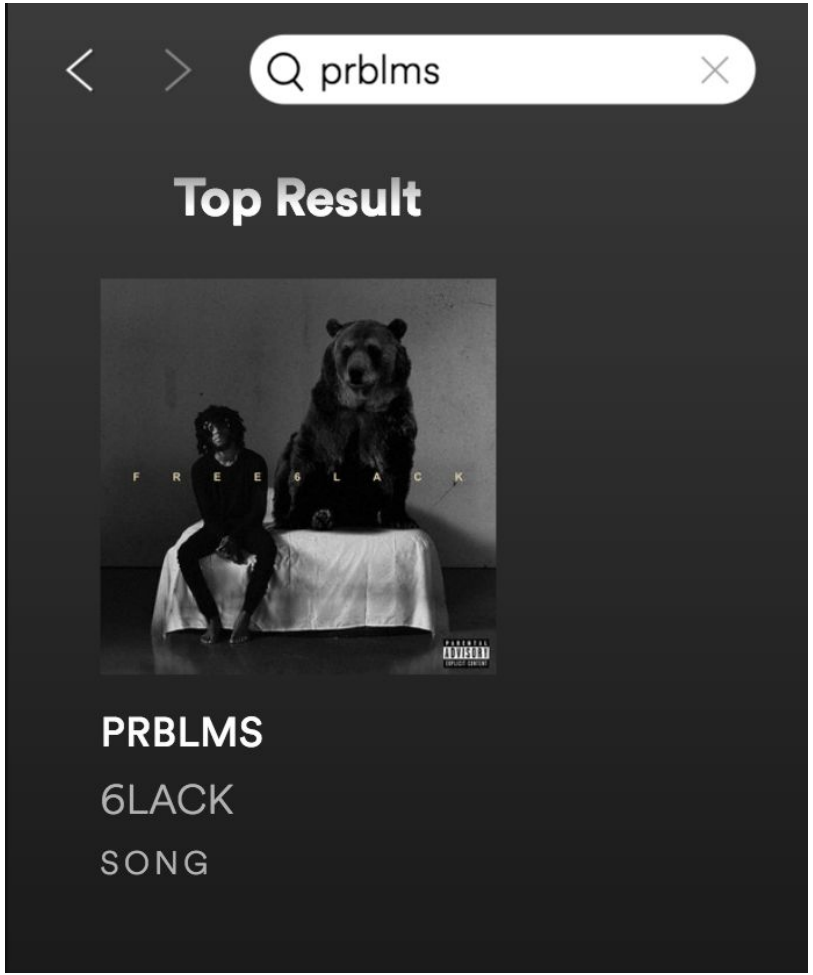
Voice Experiences

"Play my Discover Weekly"

Voice amplifies what's on top







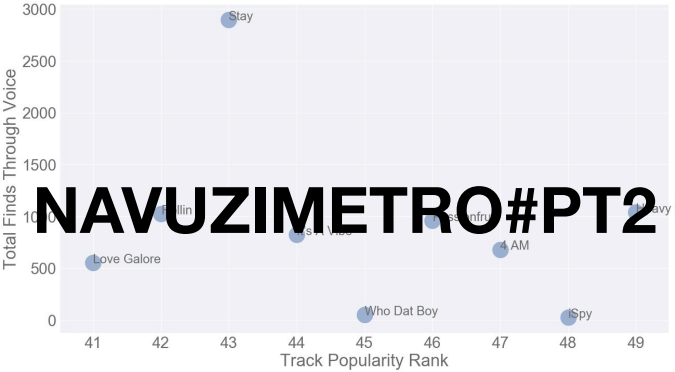
What becomes inaccessible when using voice?

“Play PRBLMS”

“Play Problems by Black”

“Play P.R.B.L.M.S. by six lack”


bear
@6LACK
Follow
if y'all catch anybody pronouncing it six lack or 6-black, enlighten em
5:48 PM - 21 May 2016



NAVUZIMETRO#PT2

Springer, A. and Cramer, H., 2018, April. Play PRBLMS: Identifying and Correcting Less Accessible Content in Voice Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 296). ACM.

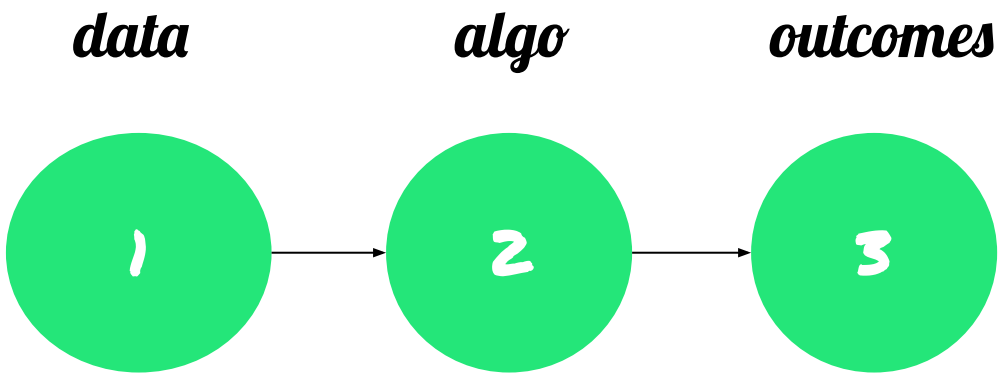
CrowdFlower

Nav Uzi Metro Hashtag Part Two

~~**Naboo See Metro Hash Tag Part Two**~~

Organizational challenges: “the checklist”

Help teams think concretely about ‘entry points for bias’ in their products

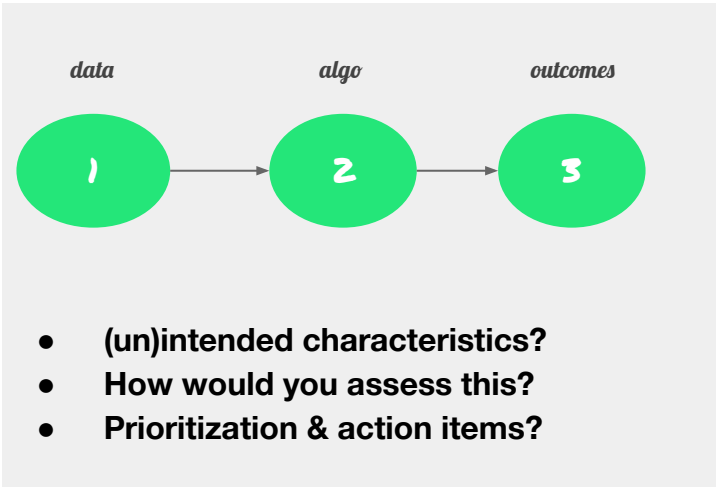


Springer et al. Assessing and addressing algorithmic bias -- but before we get there... AAAI Spring Symposium 2018.

We combined existing resources into a ‘checklist’ for teams...

Main external frameworks

Data	**Biases in social data (Olteanu et al., 2016)** Dataset nutrition label (Holland et al., 2017) Datasheets for datasets (Gebru et al., 2018)
Models	Model cards for model reporting (Mitchell et al., 2018)
Outcomes	Preexisting, Technical Bias, Emergent Bias (Friedman & Nissenbaum, 1997) **Types of harm (Crawford, 2017)**
Cycle	**Bias on the Web cyclical model (Baeza-Yates 2016)** ML Life cycle. (Wallach & Wortman Vaughan 2019)



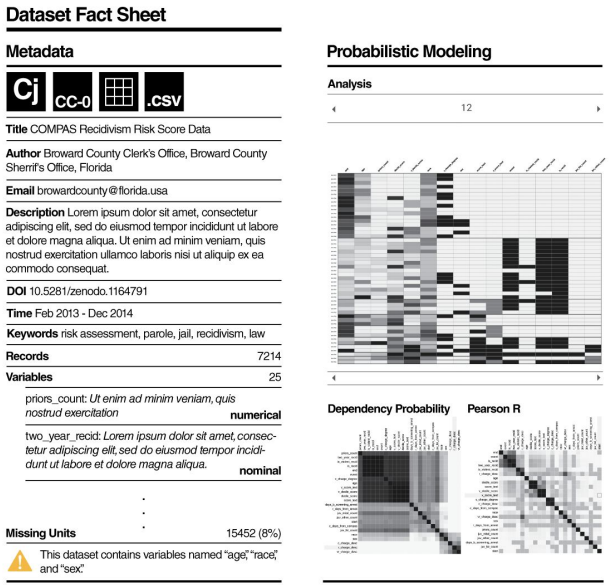
- **(un)intended characteristics?**
- **How would you assess this?**
- **Prioritization & action items?**

Cramer et al.
Assessing and addressing algorithmic bias in practice.
ACM Interactions 2018.

Existing frameworks for documenting data and models

- **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.
Datasheets for datasets.
FATML 2018.



Dataset Fact Sheet

Metadata

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."

Probabilistic Modeling

Analysis

12

Dependency Probability Pearson R

Holland et al.
Dataset nutrition label, 2018.
datanutrition.media.mit.edu

Model Card

- **Model Details**. Basic information about the model.
  - 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

Mitchell et al.
Model cards for model reporting.
FAT* 2019.

Our attempt to translate fairness literature into actionable self-interrogation...

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

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?

Our attempt to translate fairness literature into actionable self-interrogation...

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 (mobil

Temporal bia over time?

Redundancy or are near du

Content prod structural diffe that you want

Linking bias user connecti

Interface Bias presentation? (e.g.

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

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

ALGO

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

n the team, blems?

?[…]

avoid /

re 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?

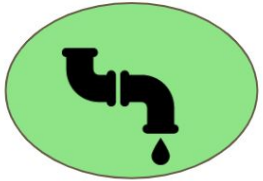
Limitations of checklists:

- Barriers to adoption
- Exhaustiveness vs. actionability
- Capturing interdependencies between systems

Lessons learned

Teams need to be aligned on priorities and interdependencies.

Mitigation efforts must compete with other prioritized deliverables



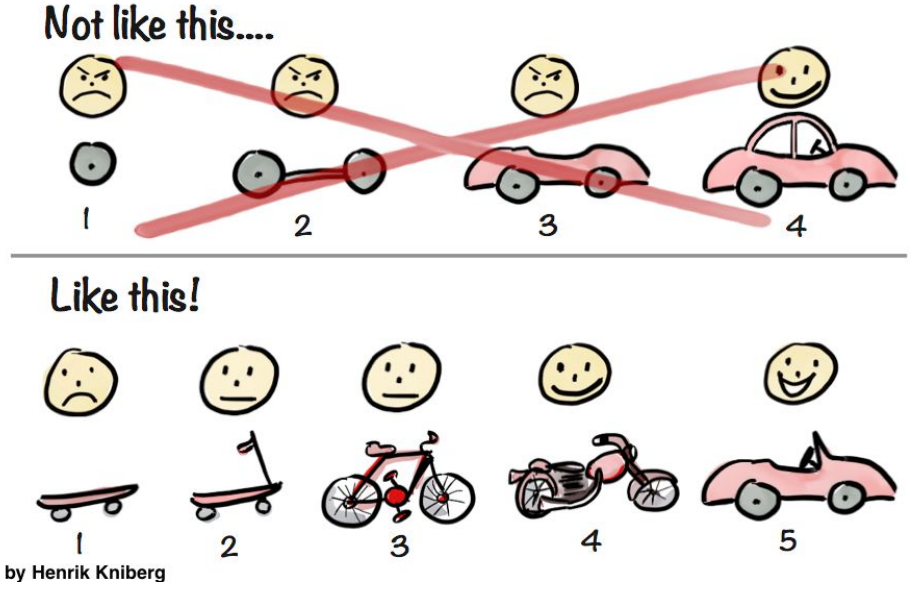
Established products may have upstream and downstream dependencies

It’s most effective to appeal to the desire to make a better product

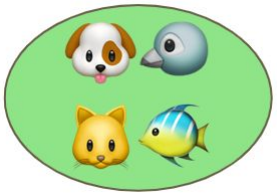
Must get organizational support through education and evangelism

Mitigation efforts should fit into ways-of-working in product.

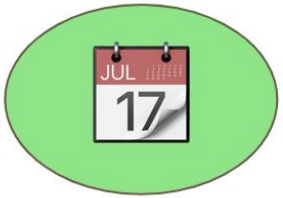
- What would an MVP for algorithmic bias assessment and mitigation look like?



Address technical debt through cultural changes.



More inclusive product development through diverse teams



Integrate algorithmic bias assessment into everyday workflow

Summary

Decisions made during development

Fairness throughout the machine learning lifecycle

+

The wider context

Organizational and domain-specific challenges

Decisions are made at every point of the pipeline.

Collaboration between research and industry is needed.

Organizational work is as crucial as advanced ML-methods.

References

- H. Cramer, K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudík, H. Wallach, S. Reddy, J. Garcia-Gathright. Challenges of incorporating algorithmic fairness into practice. Tutorial at FAT* 2019.
- K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudík, H. Wallach. Improving fairness in machine learning systems: What do industry practitioners need? CHI 2019.
- Cramer, H., Garcia-Gathright, J., Reddy, S., Springer, A. and Takeo Bouyer, R. Translation, Tracks & Data: an Algorithmic Bias Effort in Practice. CHI case studies 2019.
- A. Springer, J. Garcia-Garthright, H. Cramer. Assessing and Addressing Algorithmic Bias – But Before We Get There. AAAI Spring Symposium, UX of AI workshop 2018.
- A. Springer, H. Cramer. Play PRBLMS: Identifying and Correcting Less Accessible Content in Voice Interfaces. CHI 2018.

Any dataset or algorithmic outcome is ‘biased’

***has characteristics we can influence**

jean@spotify.com
@scinoise