Fairness-aware Machine Learning:
Practical Challenges and Lessons Learned

IVADO-Mila Intl Summer School on Bias and Discrimination in AI

June 2019

Sarah Bird (Microsoft)
Ben Hutchinson (Google)
Krishnaram Kenthapadi (LinkedIn)
Emre Kıcıman (Microsoft)
Margaret Mitchell (Google)
Ben Packer (Google)

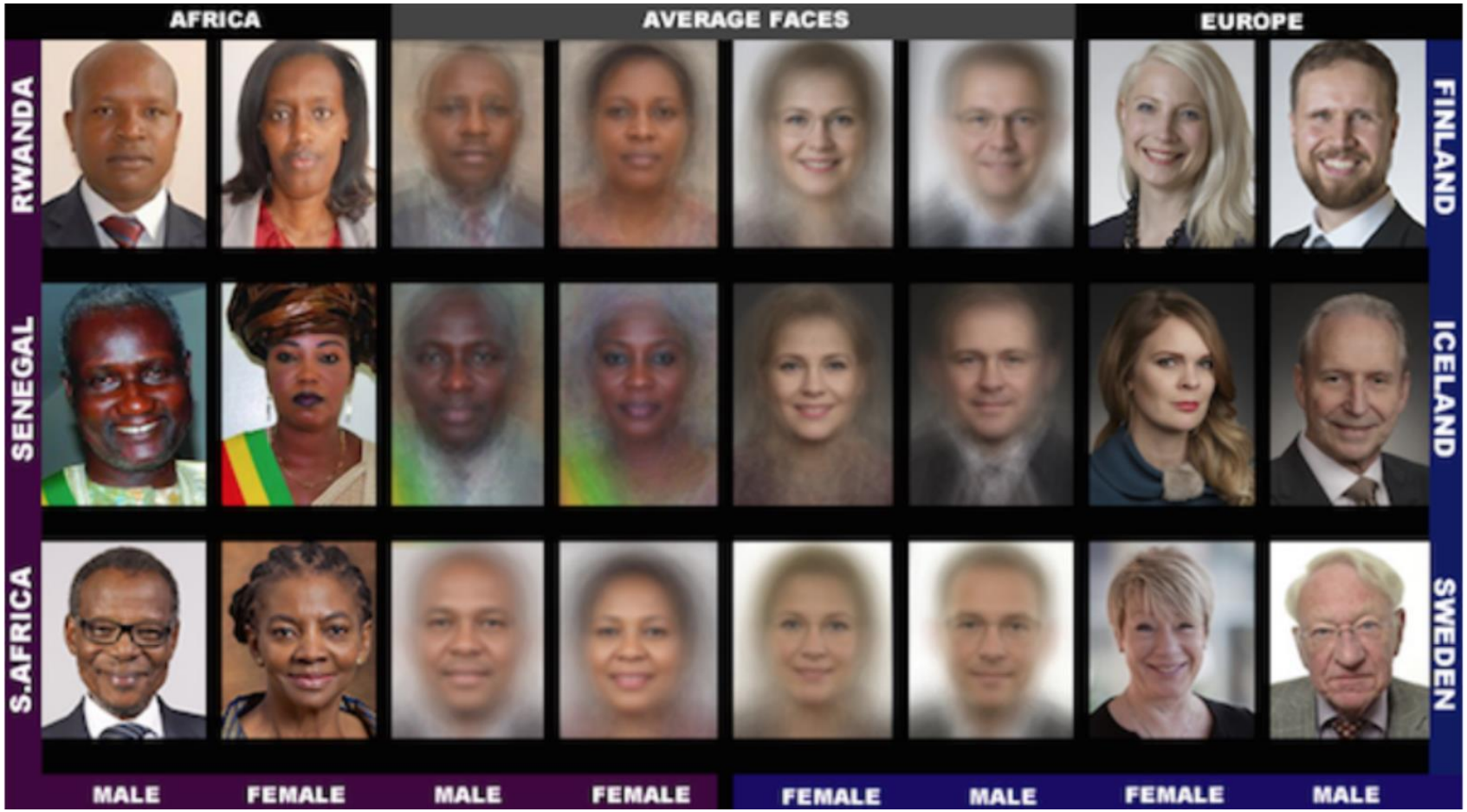
https://sites.google.com/view/www19-fairness-tutorial

The Coded Gaze [Joy Buolamwini 2016]

- Face detection software:
  Fails for some darker faces

https://www.youtube.com/watch?v=KB9sI9rY3cA

Gender Shades [Joy Buolamwini & Timnit Gebru, 2018]



- Facial analysis software:

Higher accuracy for light skinned men

- Error rates for dark skinned women: 20% - 34%

Algorithmic Bias

- Ethical challenges posed by AI systems

- Inherent biases present in society
  - Reflected in training data

  - AI/ML models prone to amplifying such biases
    - ACM FAT* conference / KDD'16 & NeurIPS'17 Tutorials

Image results: "Unprofessional hair for work"

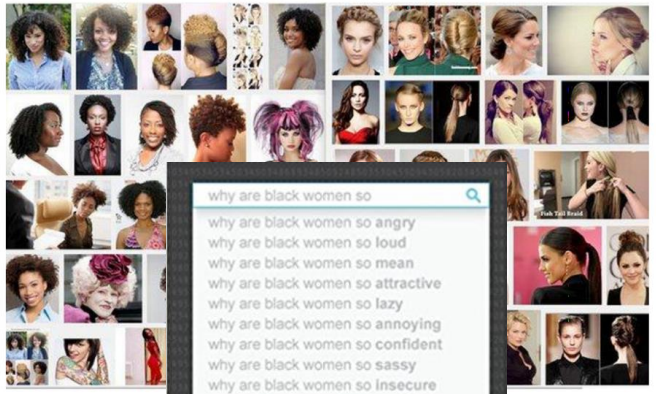
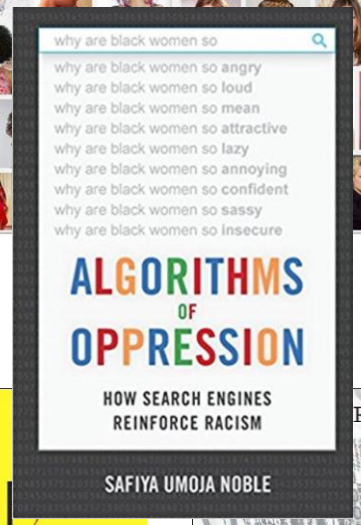
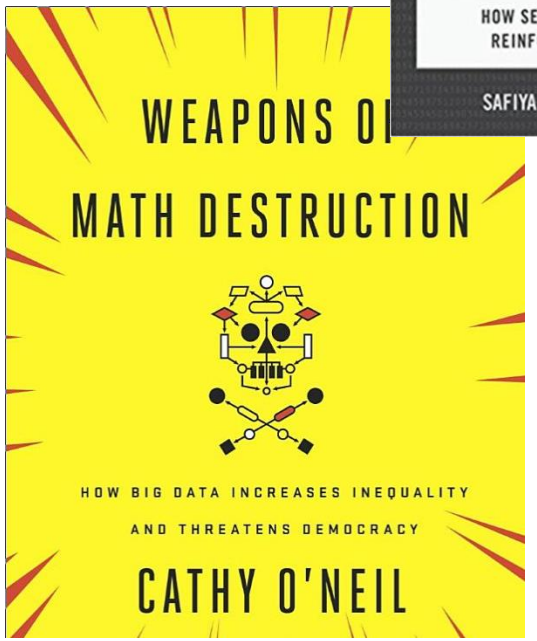
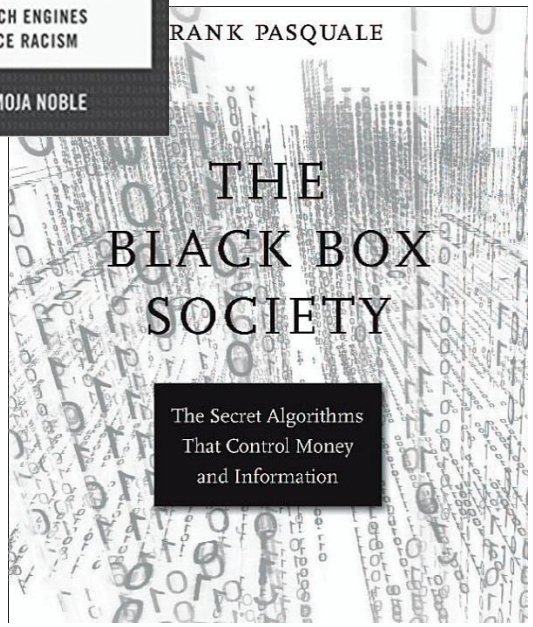


Image results: "Professional hair for work"







Laws against Discrimination

Citizenship

Immigration Reform and Control Act

Sex

Equal Pay Act of 1963;
Civil Rights Act of 1964

Age

Age Discrimination in Employment Act of 1967

Race

Civil Rights Act of 1964

Disability status

Rehabilitation Act of 1973;
Americans with Disabilities Act of 1990

And more...

Fairness

Privacy

Transparency

Explainability

Recital 71

Profiling*

Fai cy

[1] The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention. [2] Such processing includes 'profiling' that consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her. [3] However, decision-making based on such processing,

Transparency

Explainability

Recital 71

Profiling*

[6] In order to ensure fair and transparent processing in respect of the data subject, taking into account the specific circumstances and context in which the personal data are processed, the controller should use appropriate mathematical or statistical procedures for the profiling, implement technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, secure personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject, and prevent, inter alia, discriminatory effects on natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation, or processing that results in measures having such an effect. [7] Automated decision-making and profiling based on special categories of personal data should be allowed only under specific conditions.

Article 22. Automated individual decision making, including profiling

1. The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
2. Paragraph 1 shall not apply if the decision:
   (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
   (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
   (c) is based on the data subject's explicit consent.
3. In the cases referred to in points (a) and (c) of paragraph 2, the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.
4. Decisions referred to in paragraph 2 shall not be based on special categories of personal data referred to in Article 9(1), unless point (a) or (g) of Article 9(2) apply and suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.

Fairness

Privacy

Related WWW'19 sessions:

1. Tutorial: Designing Equitable Algorithms for the Web
2. Tutorial: Economic Theories of Distributive Justice for Fair Machine Learning
3. Tutorial: Socially Responsible NLP
4. Tutorial: Privacy-Preserving Data Mining in Industry (Tuesday, 14:00 - 17:30)
5. Tutorial: Explainable Recommendation and Search (Tuesday, 14:00 - 17:30)
6. Workshop: FATE and Society on the Web (Tuesday)
7. Session: Fairness, Credibility, and Search (Wednesday, 10:30 – 12:30)
8. Session: Privacy and Trust (Wednesday, 16:00 – 17:30)
9. Special Track: Designing an Ethical Web (Friday)

Transparency

Explainability

“Fairness by Design” for AI products

Outline / Learning Outcomes

- Algorithmic Bias / Discrimination
- Industry Best Practices
- Sources of Biases in ML Lifecycle
- Techniques for Fairness in ML
- Fairness Methods in Practice: Case Studies
- Key Takeaways
- Key Open Problems in Applied Fairness

Algorithmic Bias / Discrimination

and broader / related issues

Other Great Tutorials

Fairness in Machine Learning
Solon Barocas and Moritz Hardt, NeurIPS 2017

Challenges of incorporating algorithmic fairness into practice

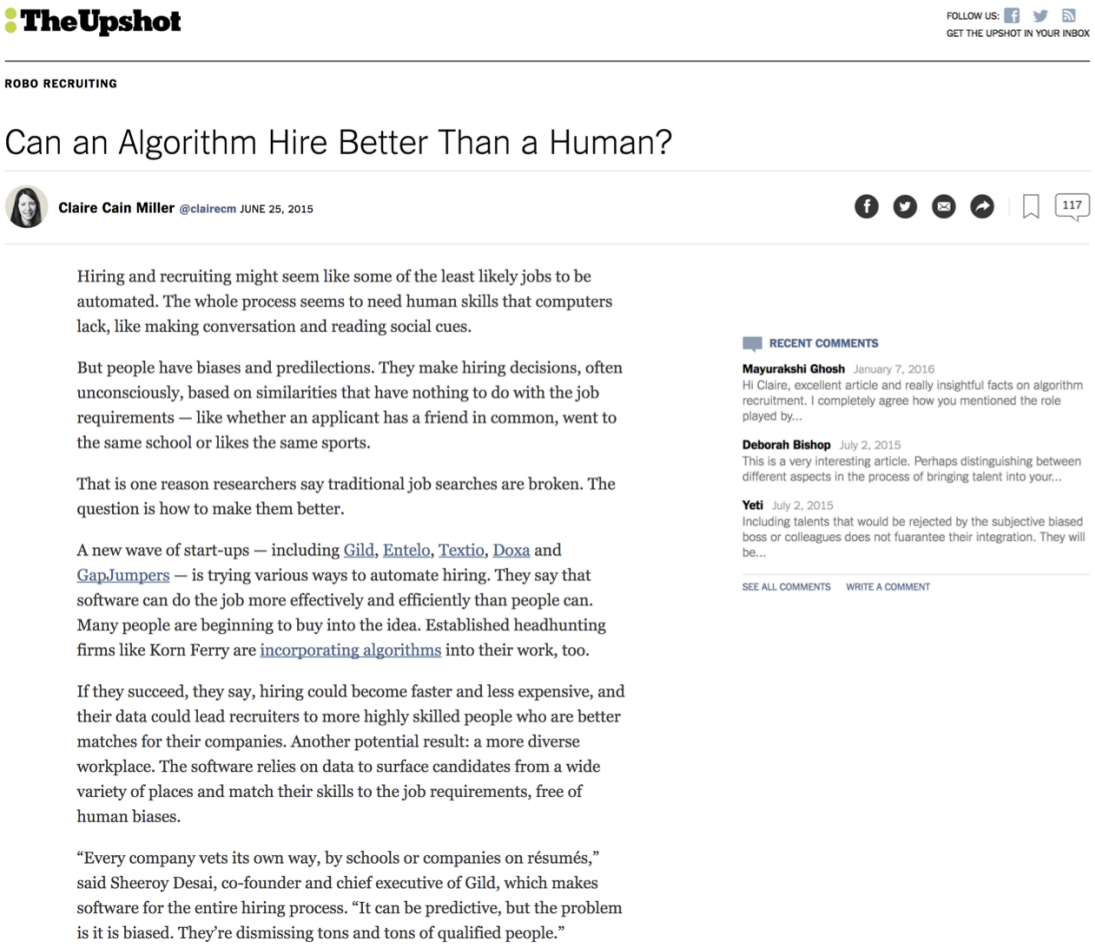
Henriette Cramer, Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, Hanna Wallach, Sravana Reddy, Jean Garcia-Gathright, FAT* 2019

Defining and Designing Fair Algorithms

Sam Corbett-Davies, Sharad Goel, ICML 2018

The Trouble with Bias

Kate Crawford, NeurIPS 2017 Keynote



TheUpshot

FOLLOW US:
GET THE UPSHOT IN YOUR INBOX

ROBO RECRUITING

Can an Algorithm Hire Better Than a Human?

Claire Cain Miller @clairecm JUNE 25, 2015

117

Hiring and recruiting might seem like some of the least likely jobs to be automated. The whole process seems to need human skills that computers lack, like making conversation and reading social cues.

But people have biases and predilections. They make hiring decisions, often unconsciously, based on similarities that have nothing to do with the job requirements — like whether an applicant has a friend in common, went to the same school or likes the same sports.

That is one reason researchers say traditional job searches are broken. The question is how to make them better.

A new wave of start-ups — including Gild, Entelo, Textio, Doxa and GapJumpers — is trying various ways to automate hiring. They say that software can do the job more effectively and efficiently than people can. Many people are beginning to buy into the idea. Established headhunting firms like Korn Ferry are incorporating algorithms into their work, too.

If they succeed, they say, hiring could become faster and less expensive, and their data could lead recruiters to more highly skilled people who are better matches for their companies. Another potential result: a more diverse workplace. The software relies on data to surface candidates from a wide variety of places and match their skills to the job requirements, free of human biases.

"Every company vets its own way, by schools or companies on résumés," said Sheeroy Desai, co-founder and chief executive of Gild, which makes software for the entire hiring process. "It can be predictive, but the problem is it is biased. They're dismissing tons and tons of qualified people."

RECENT COMMENTS

Mayurakshi Ghosh January 7, 2016
Hi Claire, excellent article and really insightful facts on algorithm recruitment. I completely agree how you mentioned the role played by...

Deborah Bishop July 2, 2015
This is a very interesting article. Perhaps distinguishing between different aspects in the process of bringing talent into your...

Yeti July 2, 2015
Including talents that would be rejected by the subjective biased boss or colleagues does not fuarantee their integration. They will be...

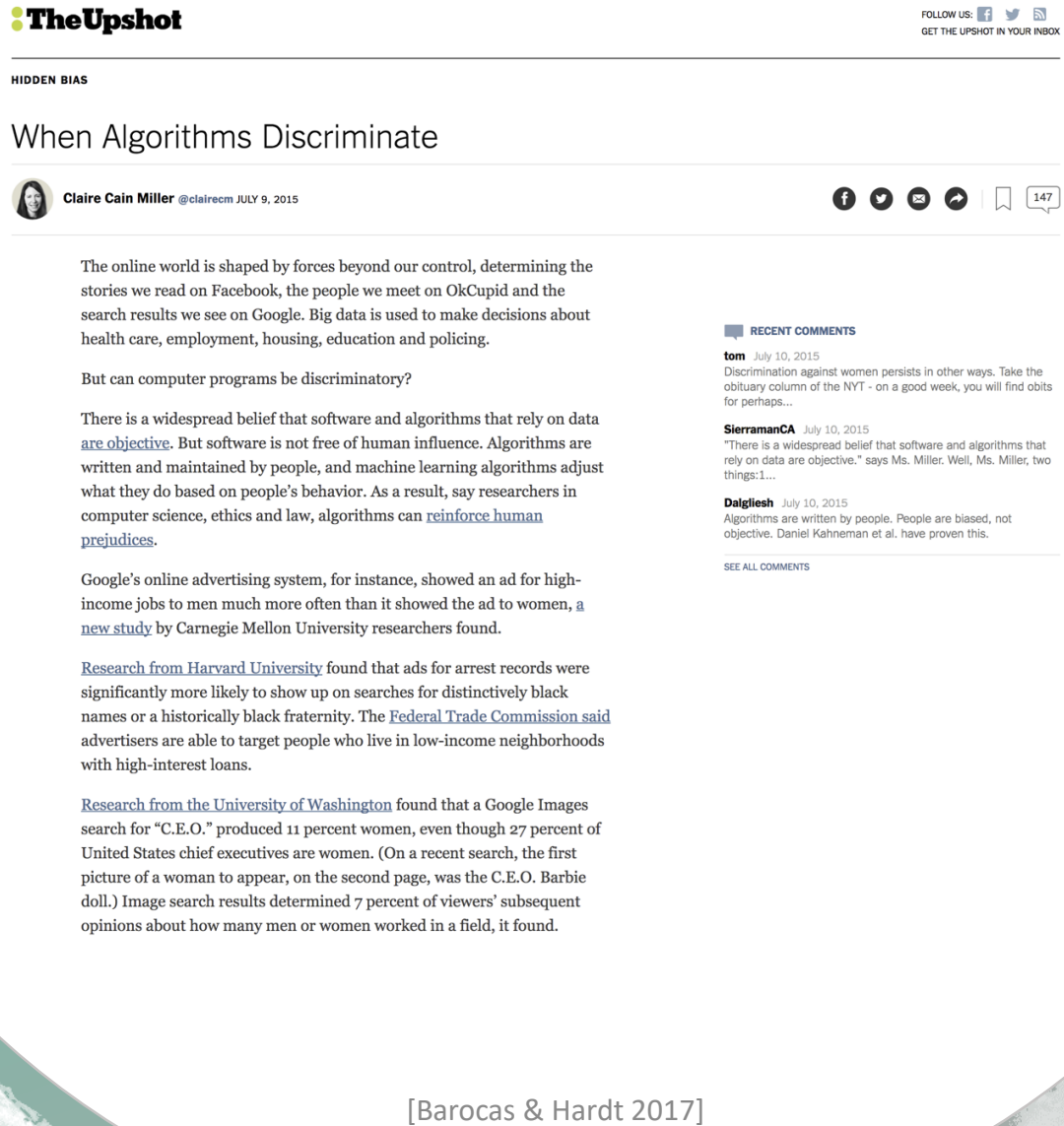
SEE ALL COMMENTS WRITE A COMMENT

[Barocas & Hardt 2017]

"[H]iring could become faster and less expensive, and [...] lead recruiters to more highly skilled people who are better matches for their companies. Another potential result: a more diverse workplace. The software relies on data to surface candidates from a wide variety of places and match their skills to the job requirements, free of human biases."

Miller (2015)

[Barocas & Hardt 2017]



TheUpshot

FOLLOW US:
GET THE UPSHOT IN YOUR INBOX

HIDDEN BIAS

When Algorithms Discriminate

Claire Cain Miller @clairecm JULY 9, 2015

147

The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

But can computer programs be discriminatory?

There is a widespread belief that software and algorithms that rely on data are objective. But software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result, say researchers in computer science, ethics and law, algorithms can reinforce human prejudices.

Google's online advertising system, for instance, showed an ad for high-income jobs to men much more often than it showed the ad to women, a new study by Carnegie Mellon University researchers found.

Research from Harvard University found that ads for arrest records were significantly more likely to show up on searches for distinctively black names or a historically black fraternity. The Federal Trade Commission said advertisers are able to target people who live in low-income neighborhoods with high-interest loans.

Research from the University of Washington found that a Google Images search for "C.E.O." produced 11 percent women, even though 27 percent of United States chief executives are women. (On a recent search, the first picture of a woman to appear, on the second page, was the C.E.O. Barbie doll.) Image search results determined 7 percent of viewers' subsequent opinions about how many men or women worked in a field, it found.

RECENT COMMENTS

tom July 10, 2015
Discrimination against women persists in other ways. Take the obituary column of the NYT - on a good week, you will find obits for perhaps...

SierramanCA July 10, 2015
"There is a widespread belief that software and algorithms that rely on data are objective." says Ms. Miller. Well, Ms. Miller, two things:1...

Dalgliesh July 10, 2015
Algorithms are written by people. People are biased, not objective. Daniel Kahneman et al. have proven this.

SEE ALL COMMENTS

[Barocas & Hardt 2017]

"But software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result […] algorithms can reinforce human prejudices."

Miller (2015)

[Barocas & Hardt 2017]

Do Better
Avoid Harm

[Cramer et al 2019]

More positive outcomes & avoiding harmful outcomes of ~~algorithms~~ for groups of people

[Cramer et al 2019]

More positive outcomes & avoiding harmful outcomes of automated systems for groups of people

[Cramer et al 2019]

Legally Recognized Protected Classes

Race (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964);**National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967);**Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

[Boracas & Hardt 2017]

Other Categories

Societal Categories

i.e., political ideology, language, income, location, topical interests, (sub)culture, physical traits, etc.

Intersectional Subpopulations

i.e., women from tech

Application-specific subpopulations

i.e., device type

Types of Harm

Harms of allocation

withhold opportunity or resources

Harms of representation

reinforce subordination along the lines of identity, stereotypes

[Cramer et al 2019, Shapiro et al., 2017, Kate Crawford, “The Trouble With Bias” keynote N(eur)IPS’17]

Bias, Discrimination & Machine Learning

Isn’t bias a technical concept?

Selection, sampling, reporting bias, Bias of an estimator, Inductive bias

Isn’t discrimination the very point of machine learning?

Unjustified basis for differentiation

[Barocas & Hardt 2017]

Discrimination is not a general concept

It is domain specific

Concerned with important opportunities that affect people's life chances

It is feature specific

Concerned with socially salient qualities that have served as the basis for unjustified and systematically adverse treatment in the past

[Barocas & Hardt 2017]

Regulated Domains

Credit (Equal Credit Opportunity Act)

Education (Civil Rights Act of 1964; Education Amendments of 1972)

Employment (Civil Rights Act of 1964)

Housing (Fair Housing Act)

'Public Accommodation' (Civil Rights Act of 1964)

Extends to marketing and advertising; not limited to final decision

[Barocas & Hardt 2017]

Discrimination Law and Legal Terms

Treatment

Disparate Treatment, Equality of Opportunity, Procedural Fairness

Outcome

Disparate Impact, Distributive justice, Minimized inequality of outcome

Equality

The assumption is that **everyone benefits from the same supports**. This is equal treatment.

Equity

Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.

Justice

All 3 can see the game without supports or accommodations because **the cause(s) of the inequity was addressed**. The systemic barrier has been removed.

[https://www.reddit.com/r/GCdebatesQT/comments/7qpbpp/food_for_thought_equality_vs_equity_vs_justice/]

Fairness is Political

Equal Treatment vs Equal Outcome

Jobs Product

Women and Men get equally good job recommendations
Both click on recommendations equally

Women and Men apply to jobs equally
Both apply to the jobs at the same rate
Both apply to the same total number of jobs
Both apply according to their proportion in the population
(in the product? In the US? In the world?)

Women and Men are hired to jobs equally
Full time vs part time? Temp vs permanent?

Women and Men are hired equally to equally good jobs
What is a good job? One that they value? Best hours? Best income?

Fairness is Political

Someone must decide

Decisions will depend on the product, company, laws, country, etc.

Why do this?

Better product and Serving Broader Population

Responsibility and Social Impact

Legal and Policy

Competitive Advantage and Brand

[Boracas & Hardt 2017]

Industry Best Practices

for Product Conception, Design, Implementation, and Evolution

Is this simple?

Process Best Practices

Identify product goals
Get the right people in the room
Identify stakeholders
Select a fairness approach
Analyze and evaluate your system
Mitigate issues
Monitor Continuously and Escalation Plans
Auditing and Transparency

Repeat for every new feature, product change, etc.

Identify product goals

Be specific

What are you trying to achieve?

i.e., remove all violent content

For what population of people?

i.e., for all users, for younger users

What metrics are you tracking?

i.e., percentage of violent content removed

Jobs: Identify product goals

What are you trying to achieve?

Match people with job opportunities

For what population of people?

US users

What jobs?

Local entry-level jobs

What metrics are you tracking?

Job applications per job and per user

Get the right people in the room

Different domains require different expertise and decision makers to be involved

Internal People

Product leaders, legal, policy, user research, design, social scientists, domain experts, machine learning experts

External People

Academics, Consultants, Advocacy groups, Government agencies

Jobs: Get the right people in the room

Internal People

Product leaders, legal, policy, social scientists

External People

Academics, advocacy groups, government agencies

Identify stakeholders

Who has a stake in this product?

i.e., content producers, content consumers

Who might be harmed?

i.e., small local content producers

How?

i.e., underrepresentation

Jobs: Identify stakeholders

Who has a stake in this product?

business trying to hire, people seeking jobs

groups of people, society as whole?

Who might be harmed?

business trying to hire, people seeking jobs, society as whole

How?

Allocation

Select a fairness approach

What type of fairness?

Group vs individual

At what point?

Equal outcome vs equal treatment

What distributions?

US population, users of the product

Jobs: Select a fairness approach

What type of fairness?

Group: women and men

At what point?

Equal treatment

What distributions?

US population

Analyze and evaluate your system

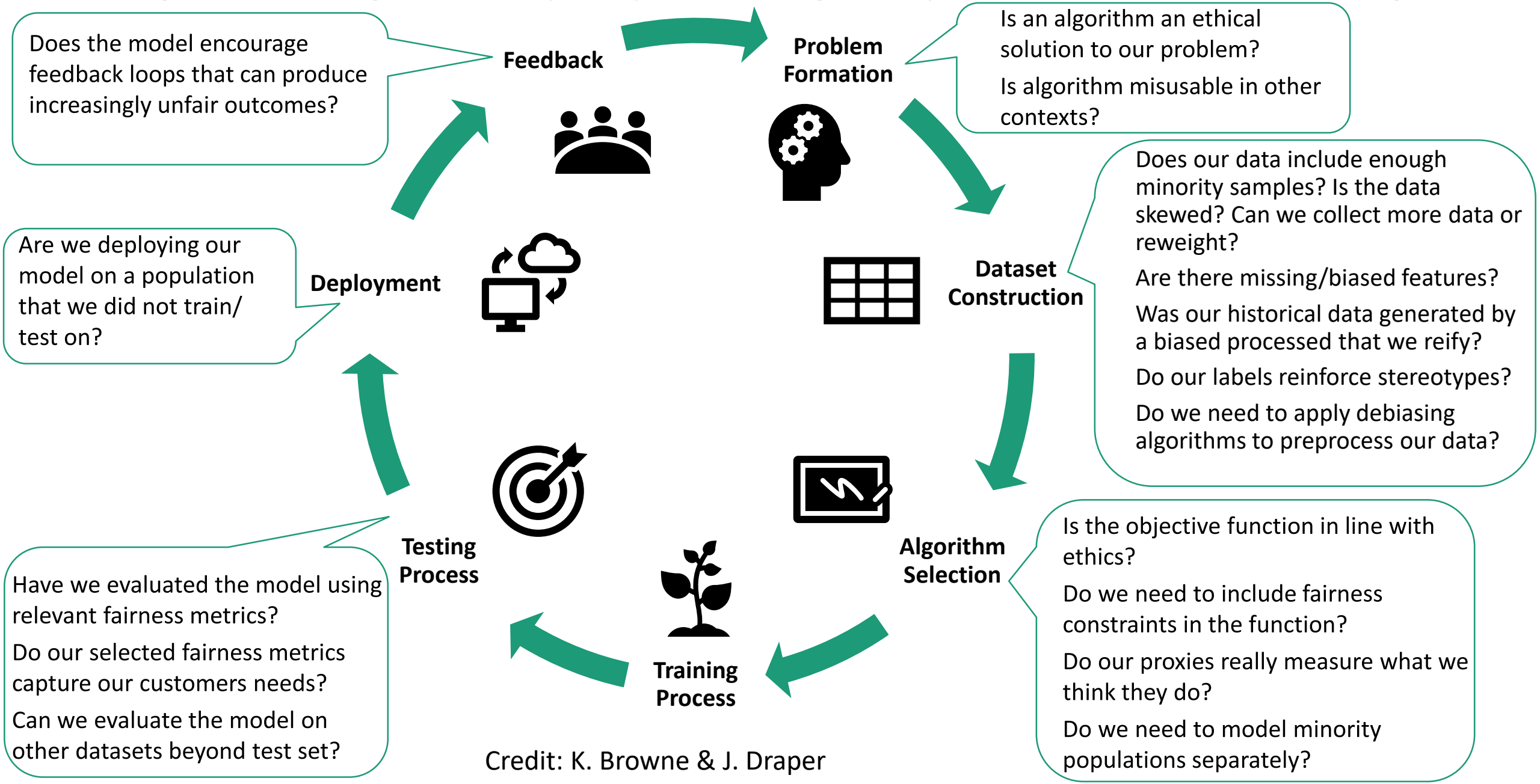
Consider the complete system end-to-end including people, technology and processes

Break your system into components

Analyze each component to understand the decisions made and their impact

Determine how well it matches up to your selected fairness approach

Engineering for equity during all phases of ML design



Credit: K. Browne & J. Draper

Jobs V1: Analyze and evaluate your system

Analyzed recommendations:

Women apply to many more jobs than men

Women click on many more of the recommendations than men

Men and women are seeing similar job recommendations

Women click on different types of jobs than men

There are not many men in the training data

Note: this gets worse with feedback

Jobs V2: Analyze and evaluate your system

Analyzed recommendations:

Women apply to many more jobs than men

Women click on many more of the recommendations than men

Men and women are seeing similar job recommendations

Women click on different types of jobs than men

Women and men are equally represented in the data set

We have very few of the types of jobs men click on in the inventory

Mitigate issues

Decide if you need to change your design, data, or metrics

Consider all types of interventions

Add balancing tracking metrics

Jobs: Mitigate issues

V1

Make men more represented in the data set

Track proportion of men and women in the system and in the data

V2

Interview male users and potential users, find out what they are looking for

Reach out to local business to improve job inventory

Track different job categories and properties in inventory

Monitor Continuously and Escalation Plans

Build in monitoring and testing for all the metrics you are tracking

Things drift because the world changes, user behavior changes, etc.

Every time you deploy a model (or before), as the system runs, etc.

Develop response and escalation plans

How are you going to respond when something happens?

What blocks launch?

Who decides?

Jobs: Escalation Plans

- The night before our first major launch we discover the model doesn’t perform well for men because of issues with the dataset
- We told investors that we would launch tomorrow and all of the PR and marketing is set to go live
- Do we launch? Who decides?

Auditing and Transparency

Important to also consider who else needs visibility into your process and your system

Do you need to prove that your system meets regulations?

Do you want outside experts to certify your system?

Do users need to understand fairness in the system?

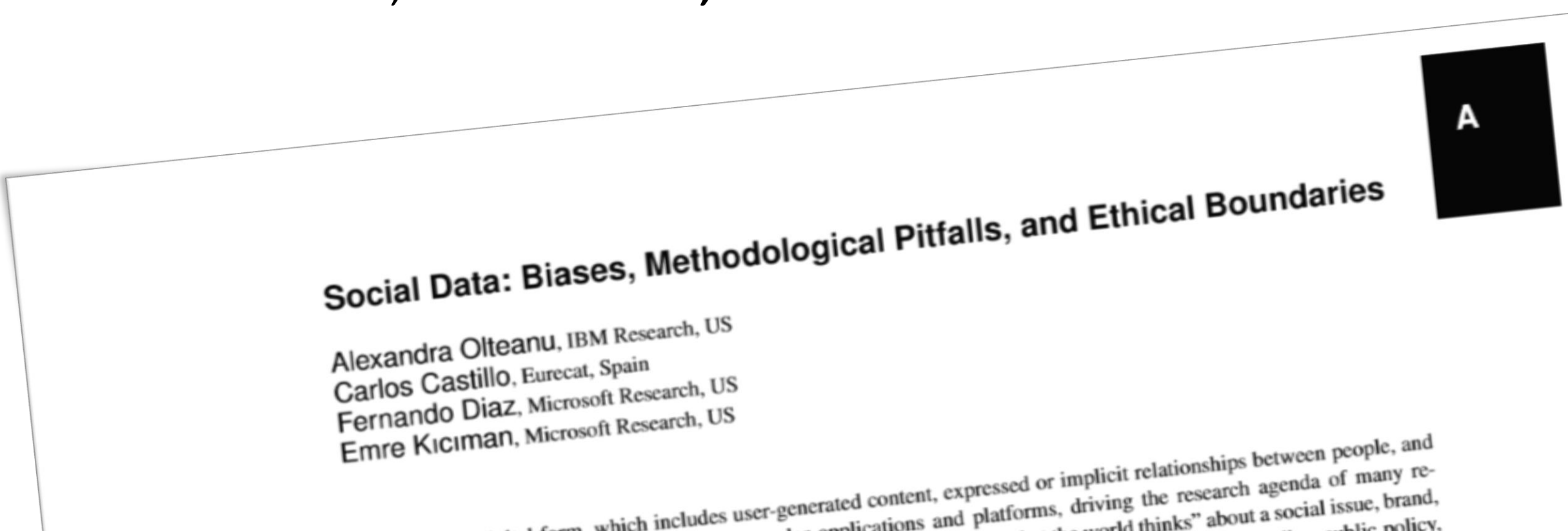
Process Best Practices

Identify product goals
Get the right people in the room
Identify stakeholders
Select a fairness approach
Analyze and evaluate your system
Mitigate issues
Monitor Continuously and Escalation Plans
Auditing and Transparency

Sources of Biases in ML Lifecycle

Collaborators

Much of this section is based on survey paper and tutorial series written by **Alexandra Olteanu**, **Carlos Castillo**, **Fernando Diaz, Emre Kıcıman**

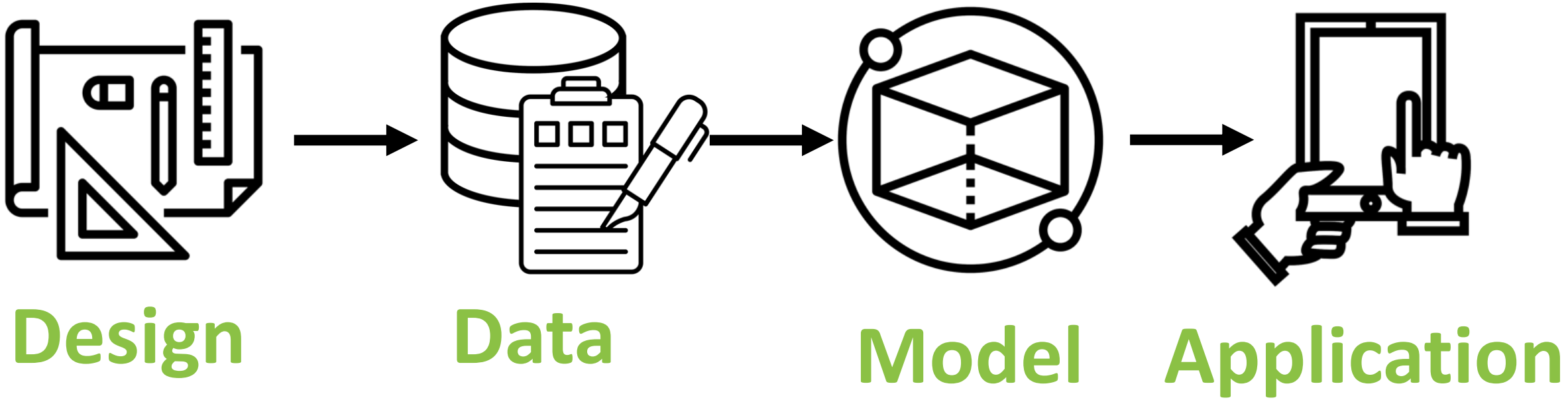


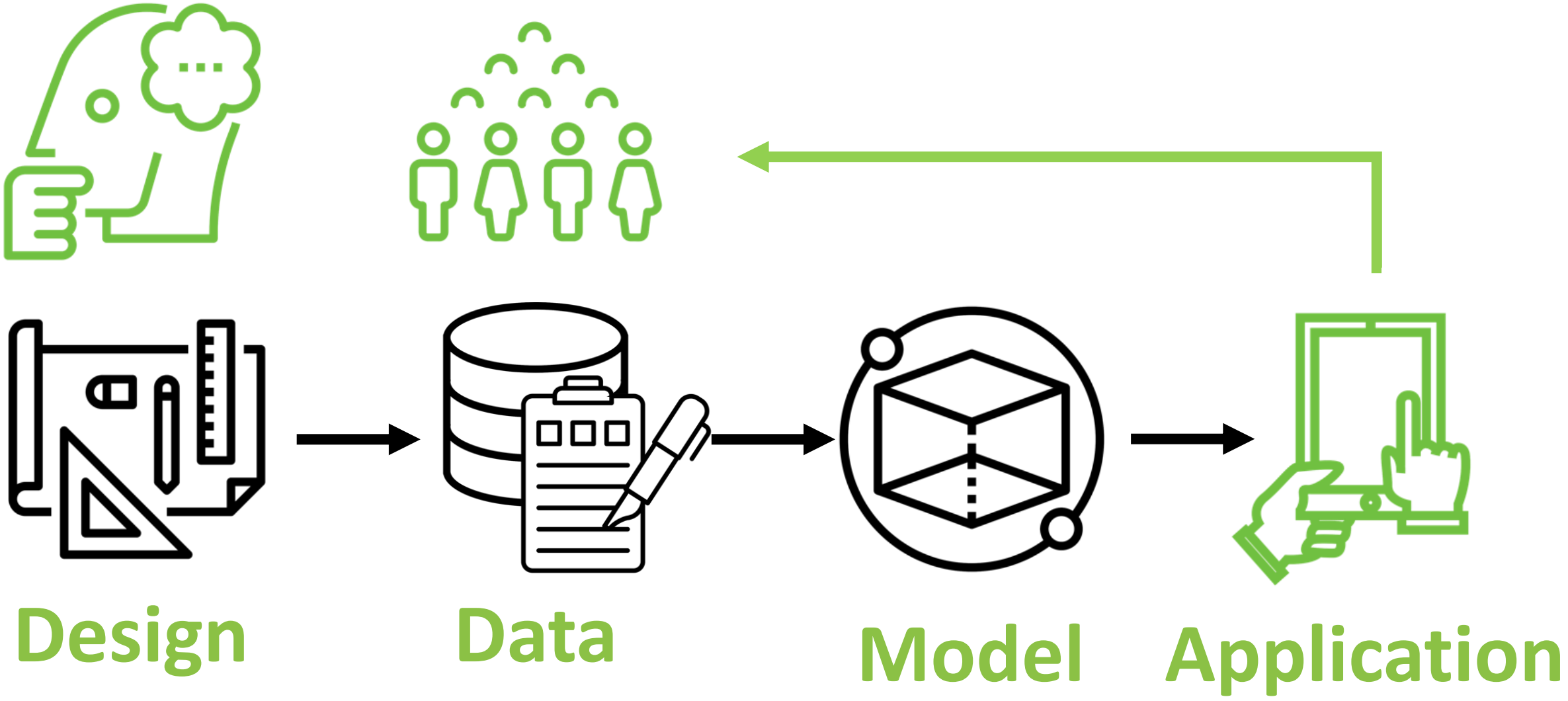
A

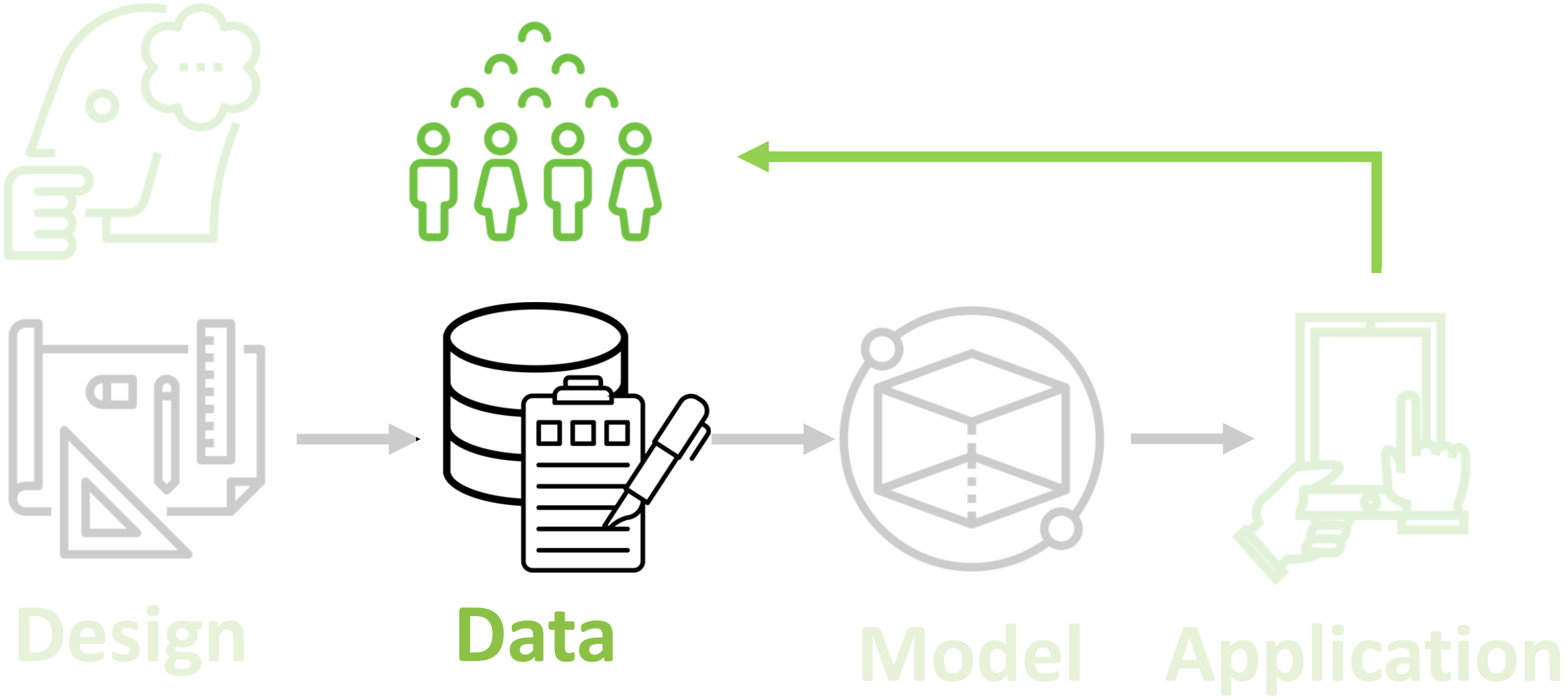
Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Olteanu, IBM Research, US
Carlos Castillo, Eurecat, Spain
Fernando Diaz, Microsoft Research, US
Emre Kıcıman, Microsoft Research, US

form, which includes user-generated content, expressed or implicit relationships between people, and applications and platforms, driving the research agenda of many re- world thinks" about a social issue, brand, public policy,


Design
Data
Model
Application


Design
Data
Model
Application


Design
Data
Model
Application

Data bias: a *systematic distortion* in data that compromises its use for a task.

Note: Bias must be considered relative to task

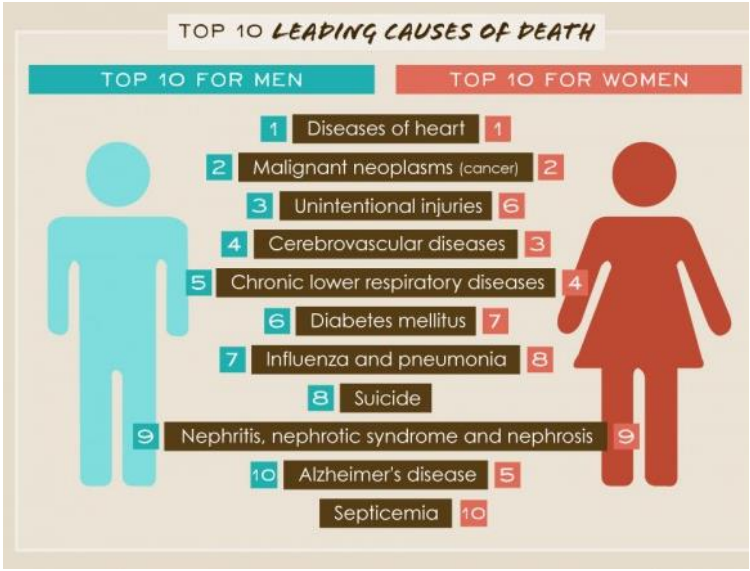
Gender in loan application

FEDERAL TRADE COMMISSION

Mortgage discrimination is against the law.

Gender discrimination is illegal

Gender in medical diagnosis



Gender-specific medical diagnosis is desirable

What does data bias look like?

Measure systematic distortions along 5 data properties

1. Population Biases
2. Behavioral Biases
3. Content Production Biases
4. Linking Biases
5. Temporal Biases

What does data bias look like?

Measure distortions along 5 data properties

1. **Population Biases**
   Differences in demographics or other user characteristics between a user population represented in a dataset or platform and a target population
2. Behavioral Biases
3. Content Production Biases
4. Linking Biases
5. Temporal Biases

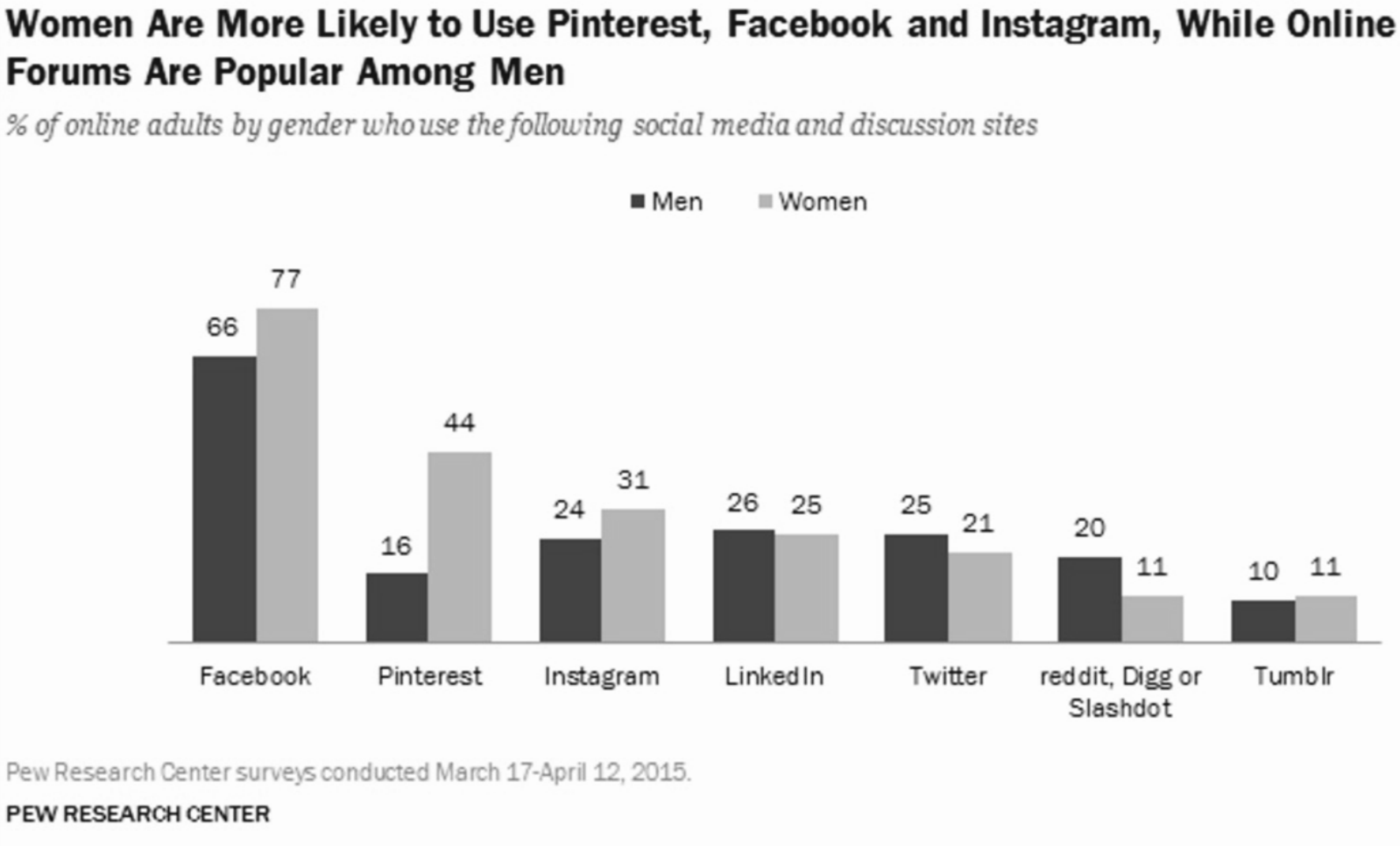


Figure from http://www.pewinternet.org/2016/11/11/social-media-update-2016/

See [Hargittai'07] for statistics about social media use among young adults according to gender, race and ethnicity, and parental educational background.

Systematic distortions must be evaluated in a task dependent way

E.g., for many tasks, populations should **match target population**, to improve external validity

But for some other tasks, subpopulations require approximately **equal representation** to achieve task parity

Gender Shades

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%

http://gendershades.org/

What does data bias look like?

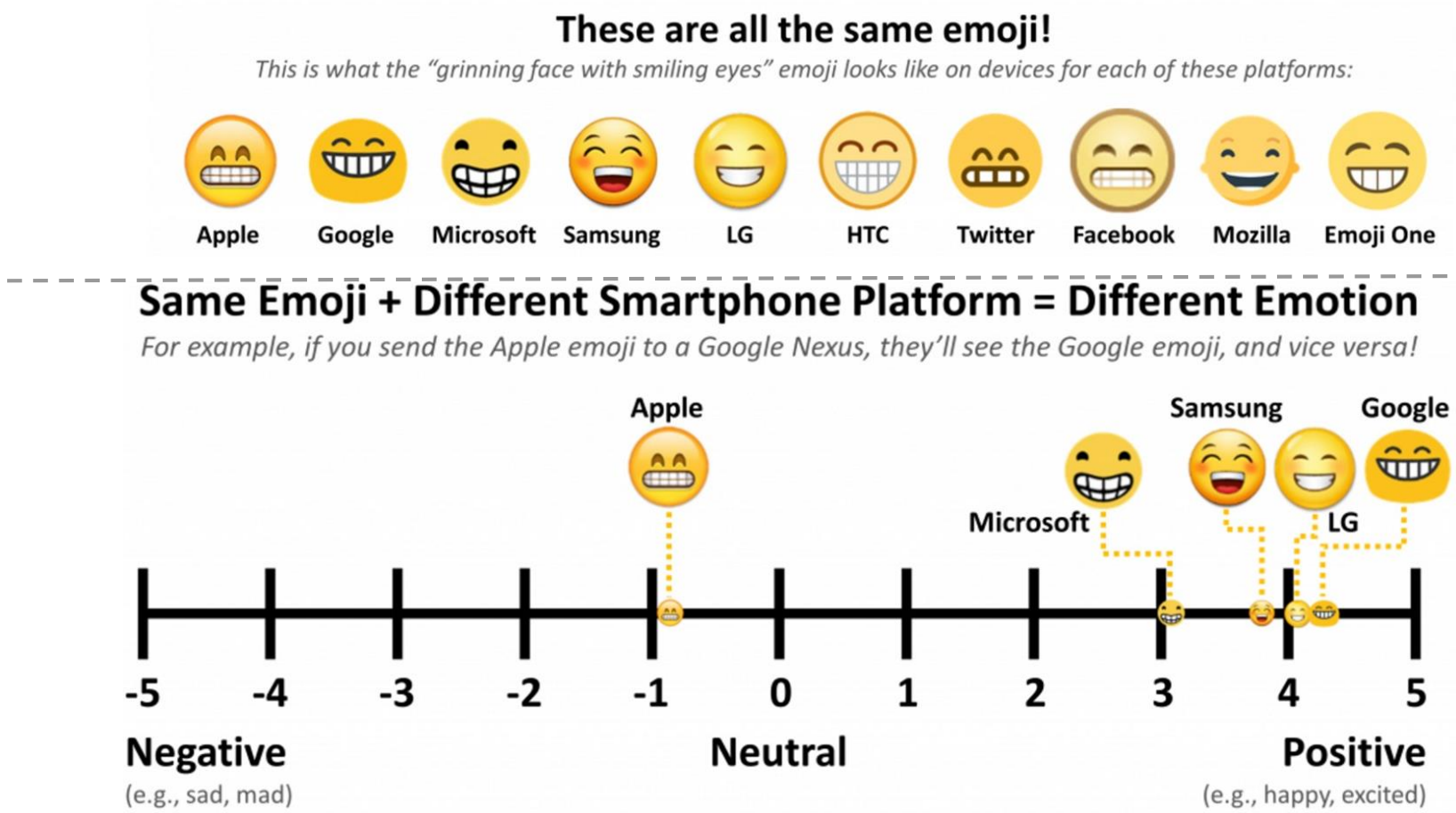
Measure distortions along 5 data properties

1. Population Biases
2. **Behavioral Biases**
   Differences in user behavior across platforms or contexts, or across users represented in different datasets
3. Content Production Biases
4. Linking Biases
5. Temporal Biases

Behavioral Biases from *Functional Issues*

Platform functionality and algorithms influence human behaviors

and our observations of human behaviors

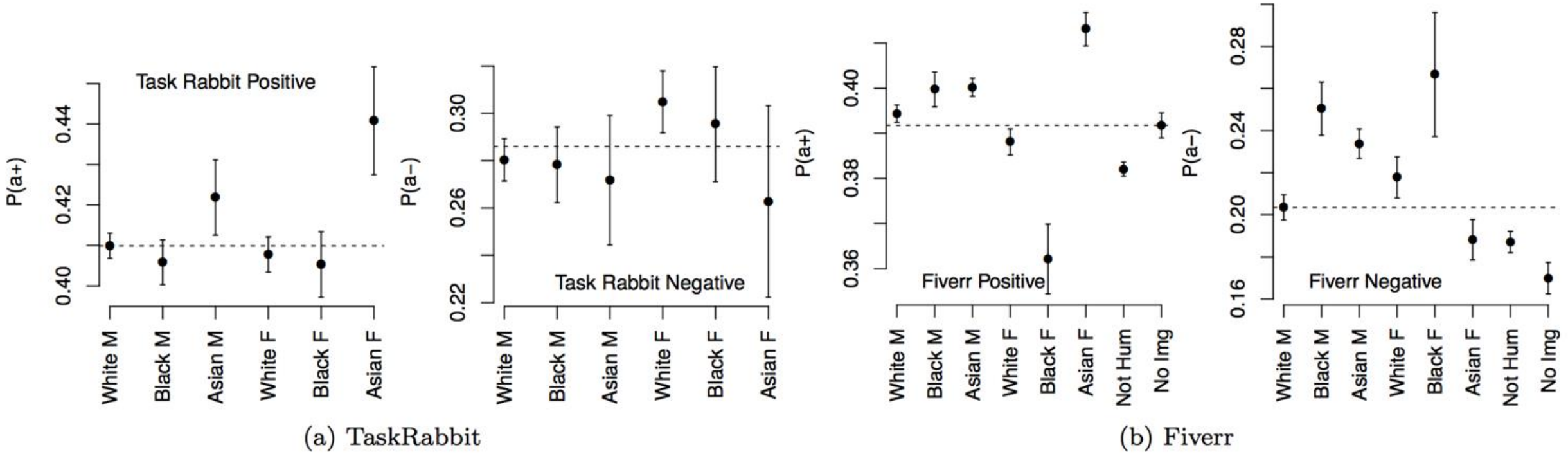


[Miller et al. ICWSM'16]

Figure from: http://grouplens.org/blog/investigating-the-potential-for-miscommunication-using-emoji/

Cultural elements and social contexts are reflected in social datasets

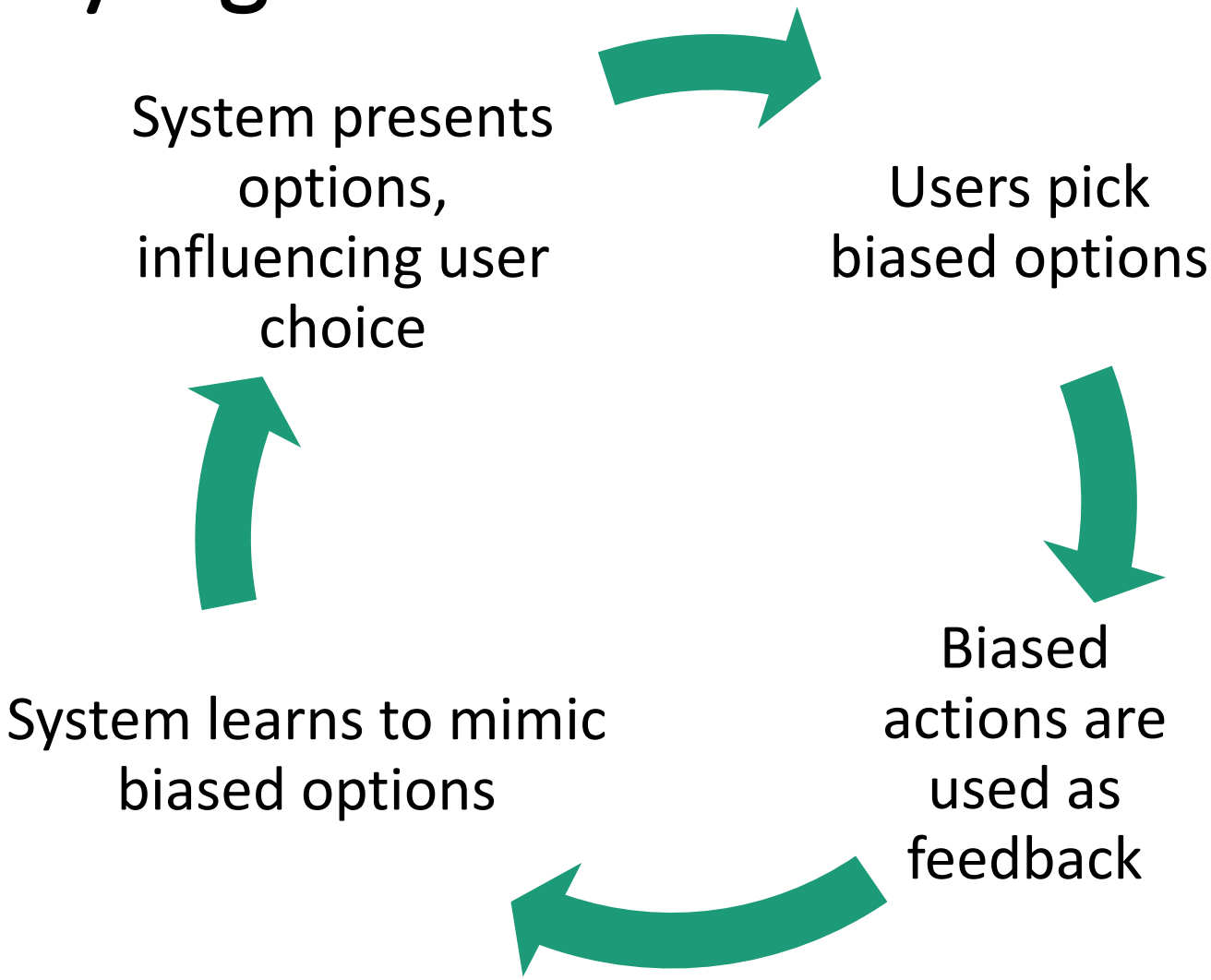
Figure from
[Hannak et al. CSCW 2017]



(a) TaskRabbit (b) Fiverr

Figure 2: Fitted $P(a_+)$ and $P(a_-)$ depending on combinations of gender and race of the reviewed worker. Points show expected values and bars standard errors. In Fiverr, Black workers are less likely to be described with adjectives for positive words, and Black Male workers are more likely to be described with adjectives for negative words.

Societal biases embedded in behavior can be amplified by algorithms



What does data bias look like?

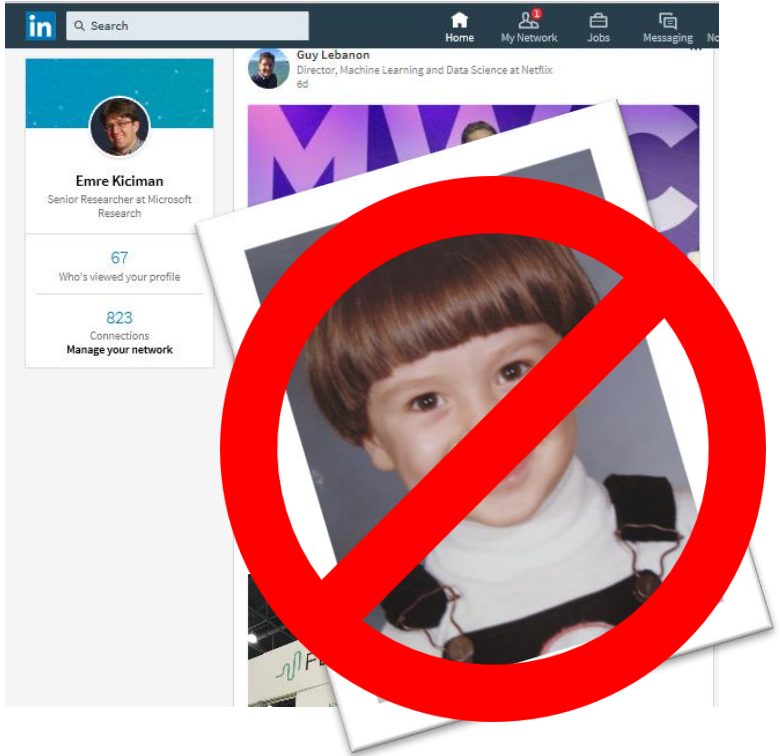
Measure distortions along 5 data properties

1. Population Biases
2. Behavioral Biases
3. **Content Production Biases**
   Lexical, syntactic, semantic, and structural differences in the contents generated by users
4. Linking Biases
5. Temporal Biases

Behavioral Biases from *Normative Issues*

Community norms and societal biases influence observed behavior
and vary across online and offline communities and contexts

What kind of pictures would you share on **Facebook**, but not on **LinkedIn**?



Are individuals comfortable contradicting popular opinions?

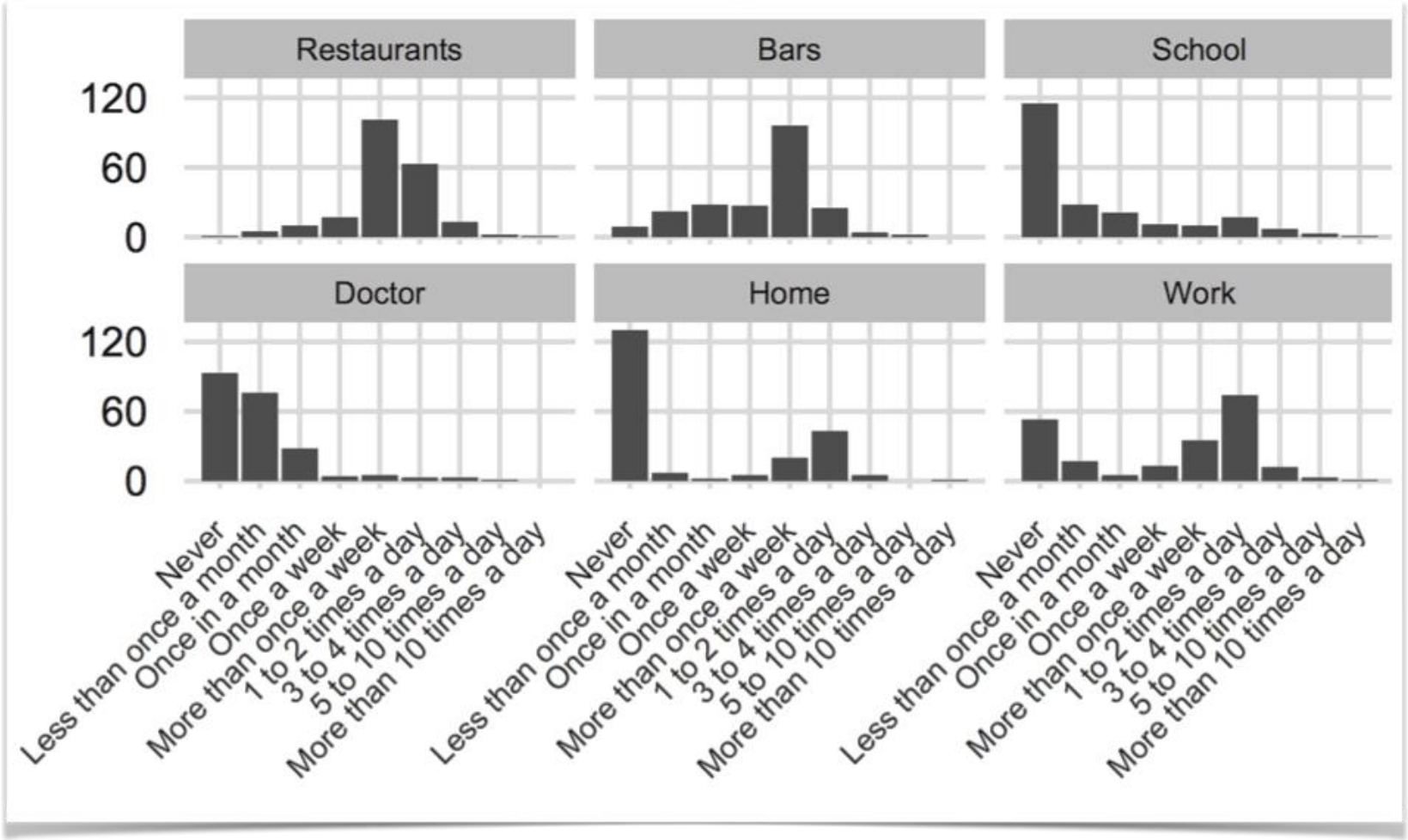
E.g., after singer Prince died, most SNs showed public mourning. But not anonymous site PostSecret

The same mechanism can embed different meanings in different contexts [Tufekci ICWSM'14]

[the meaning of retweets or likes] *"could range from affirmation to denunciation to sarcasm to approval to disgust"*

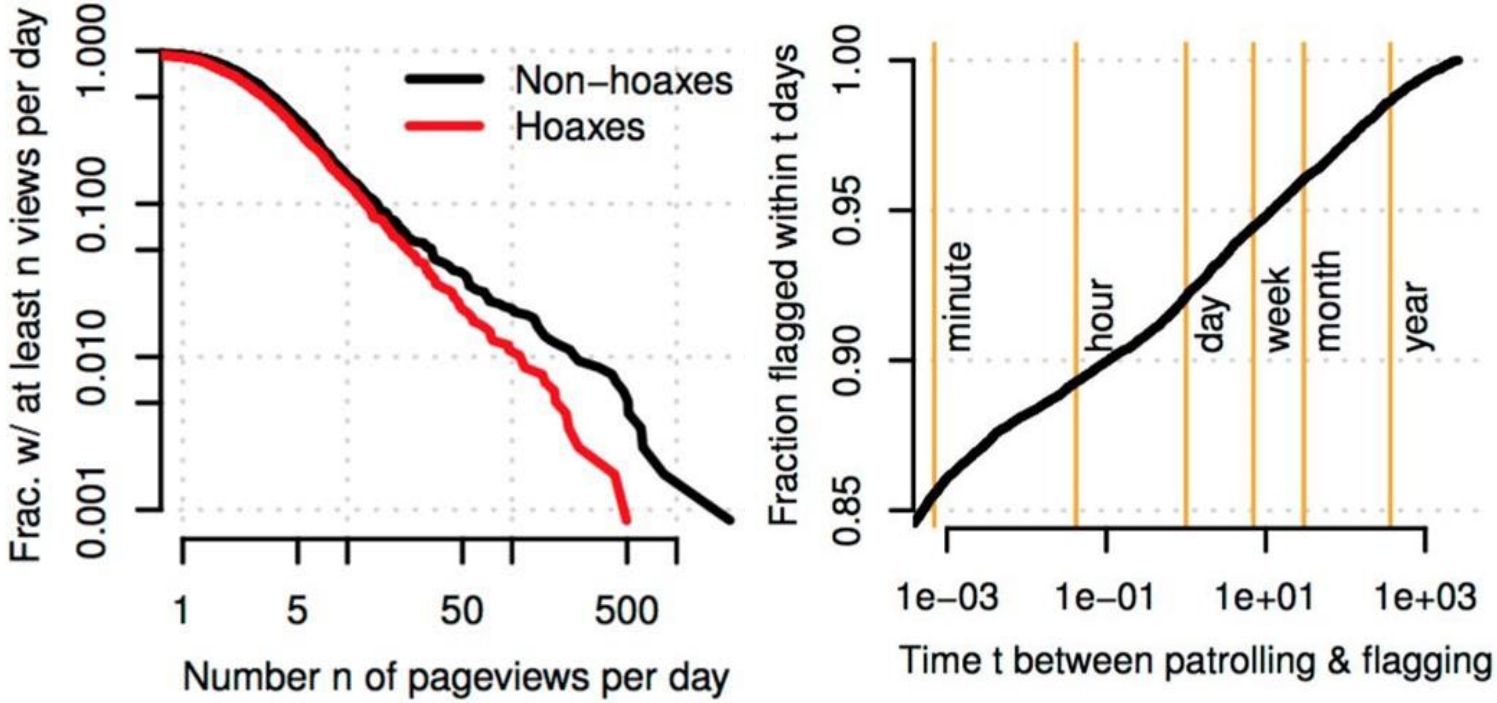
The awareness of being observed by other impacts user behavior: Privacy and safety concerns

Privacy concerns affect what content users share, and, thus, the type of patterns we observe.



Foursquare/Image from [Lindqvist et al. CHI'11]

As other media, social media contains misinformation and disinformation



Misinformation is false information, unintentionally spread

Disinformation is false information, deliberately spread

Hoaxes on Wikipedia: (**left**) impact as number of views per day for hoaxes surviving at least 7 days, and (**right**) time until a hoax gets detected and flagged.

Figures from [Kumar et al. 2016]

What does data bias look like?

Measure distortions along 5 data properties

1. Population Biases
2. Behavioral Biases
3. Content Production Biases
4. **Linking Biases**

   Differences in the attributes of networks obtained from user connections, interactions, or activity
5. Temporal Biases

Behavior-based and connection-based social links are different

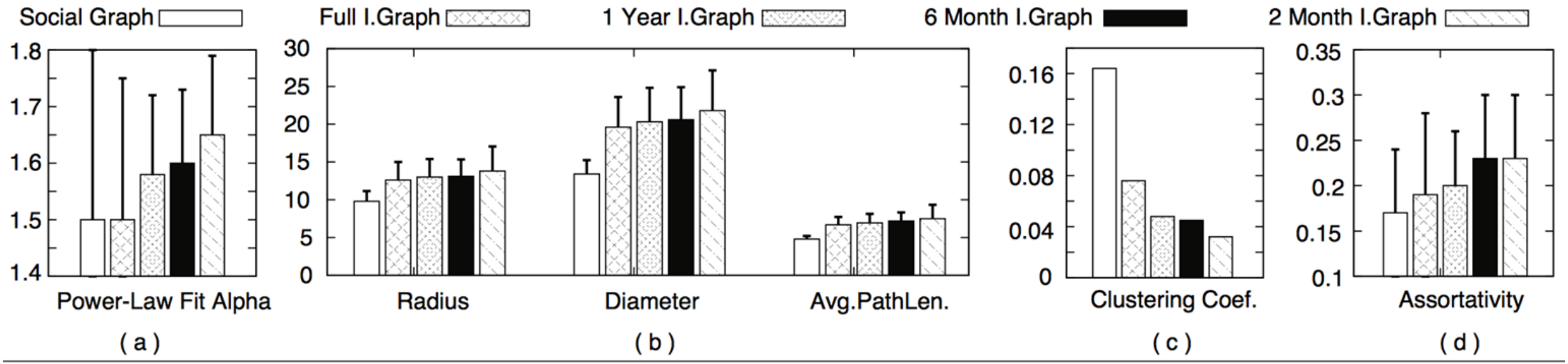


Figure 14. Graph measurements for four interaction graphs compared to the entire Facebook social network.

Figure from [Wilson et al. EuroSys'09]

What does data bias look like?

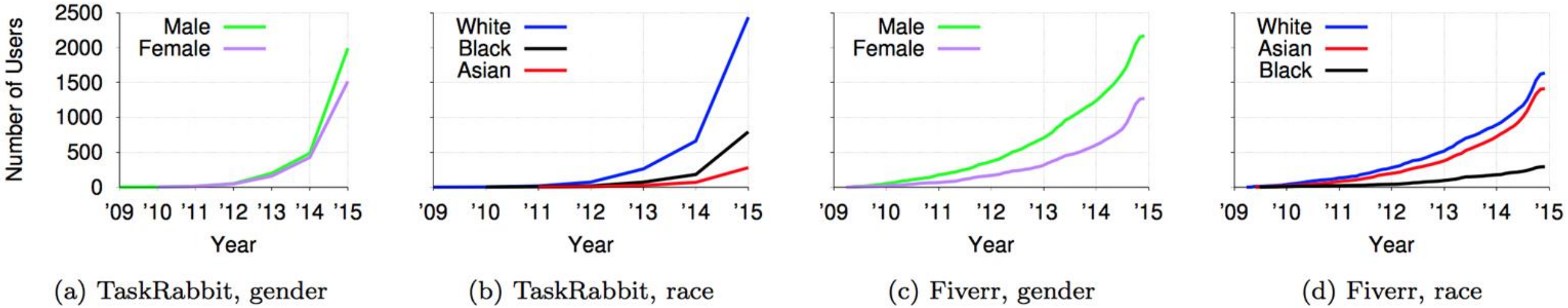
Measure distortions along 5 data properties

1. Population Biases
2. Behavioral Biases
3. Content Production Biases
4. Linking Biases
5. **Temporal Biases**
   Differences in populations and behaviors over time

Different demographics can exhibit different growth rates across and within social platforms

TaskRabbit and Fiverr are online freelance marketplaces.

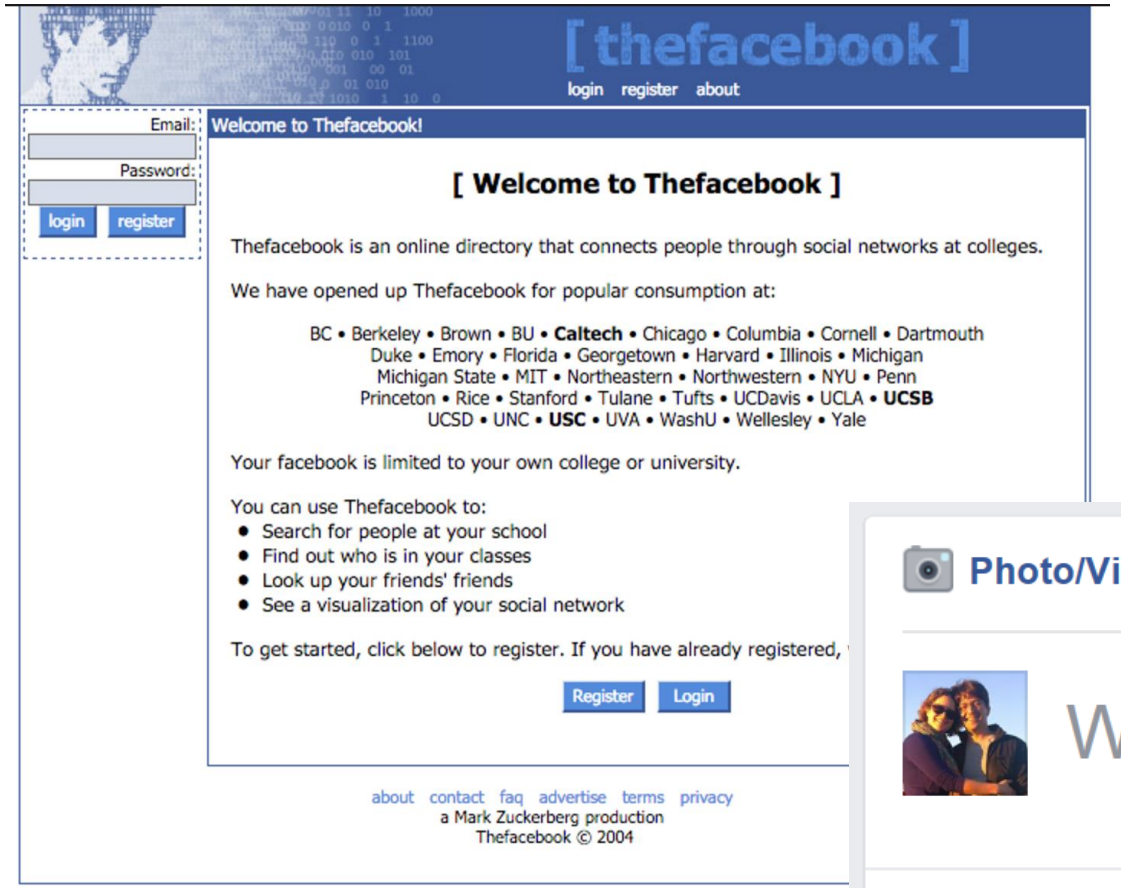
Figure from [Hannak et al. CSCW 2017]



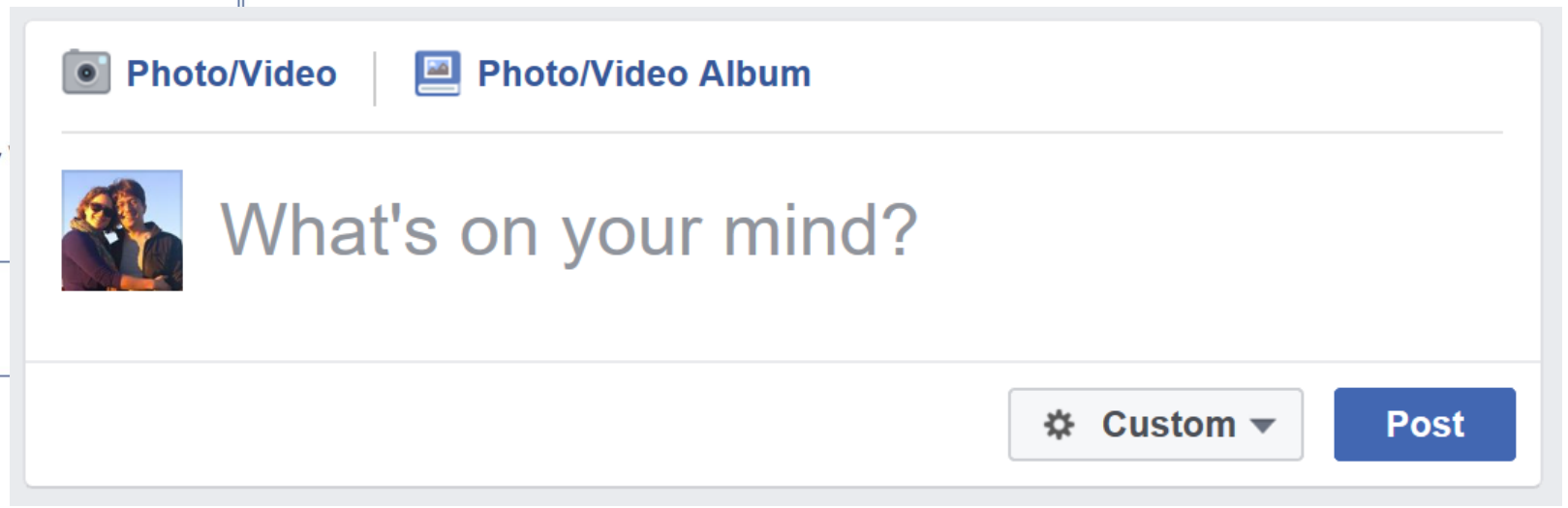
(a) TaskRabbit, gender (b) TaskRabbit, race (c) Fiverr, gender (d) Fiverr, race

Figure 1: Member growth over time on TaskRabbit and Fiverr, broken down by gender and race.

E.g., Change in Features over Time



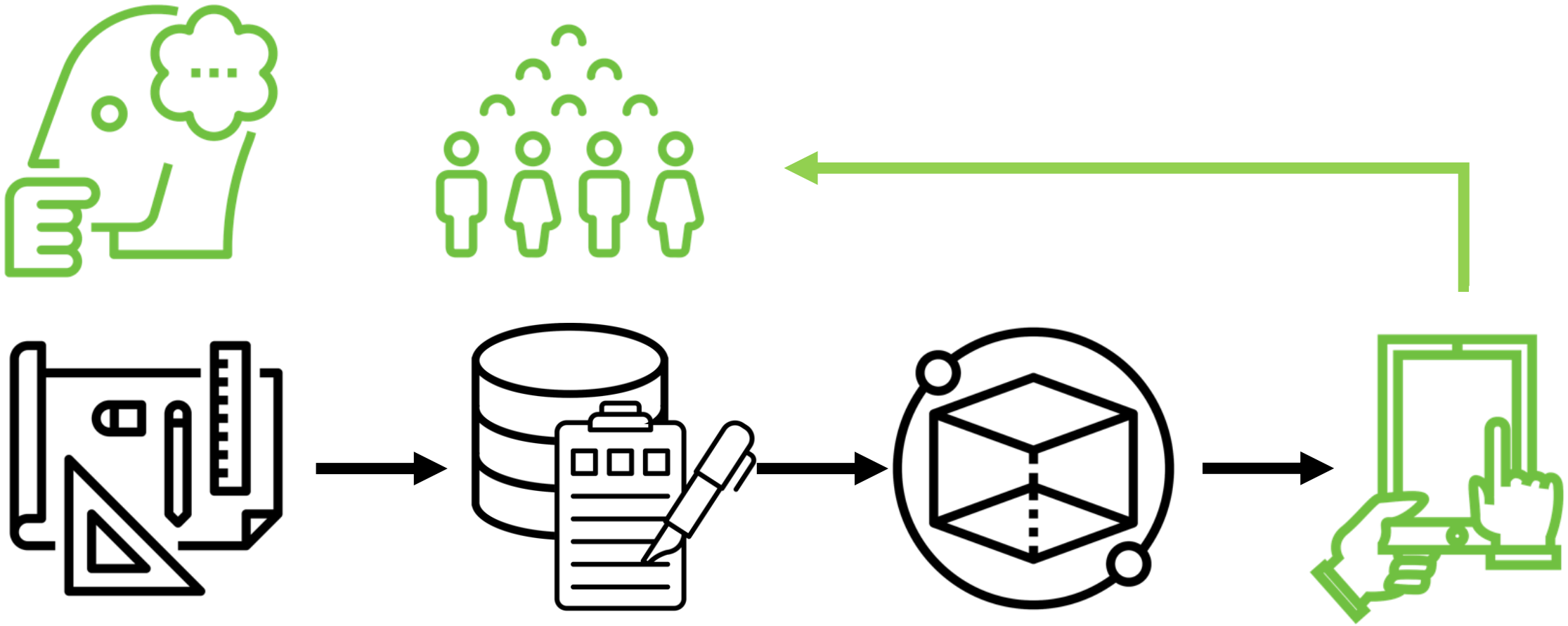
Introducing a new feature or changing an existing feature impacts usage patterns on the platform.



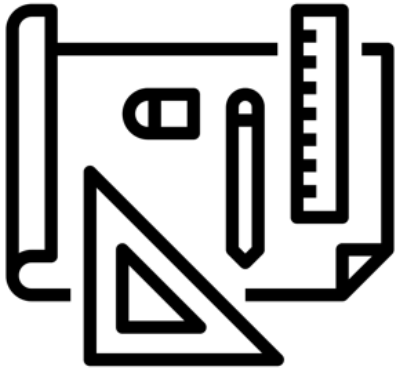
Biases can creep in at data *collection* as well

- **Common data collection issues:**
- *Acquisition* – Platform restrictions on data gathering (sampling, APIs, …)
- *Querying* – Bias from limited expressiveness, keywords, geo- and other
- *Filtering* – Outliers, stop words

Best Practices for Bias Avoidance/Mitigation

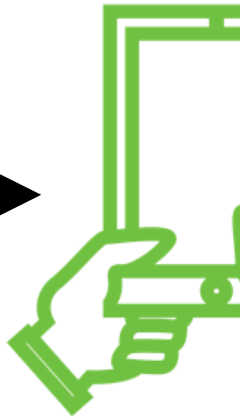


Best Practices for Bias Avoidance/Mitigation



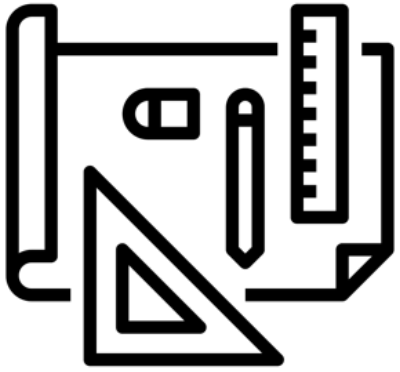
Design

Consider
team composition
for diversity of thought, background and experiences



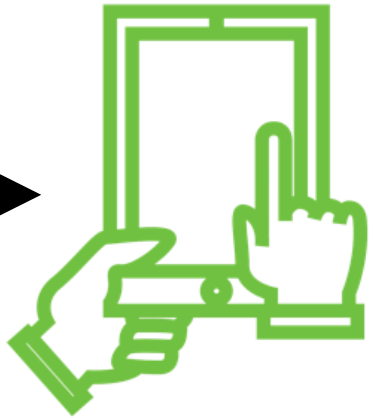
Application

Best Practices for Bias Avoidance/Mitigation



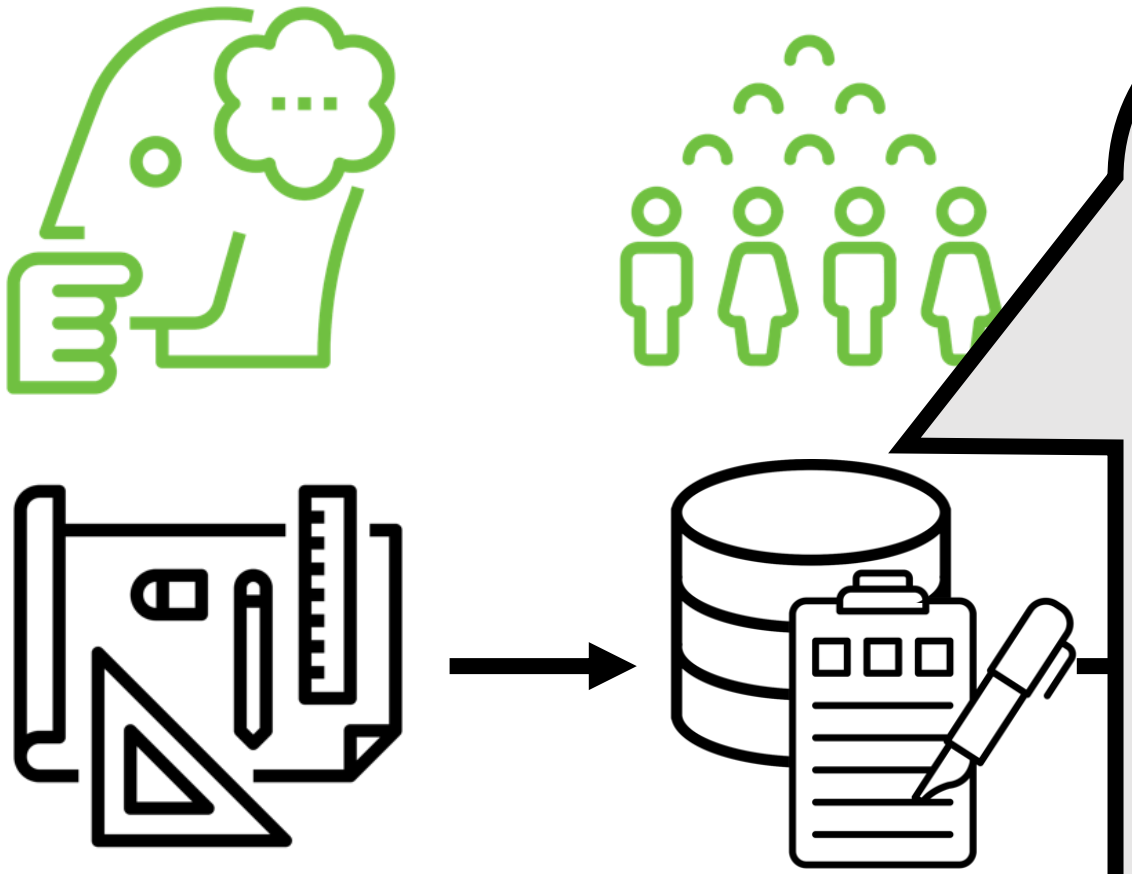
Design

Understand the **task**, **stakeholders**, and potential for **errors and harm**



Application

Best Practices for Bias Avoidance/Mitigation



Design

Data

Check data sets

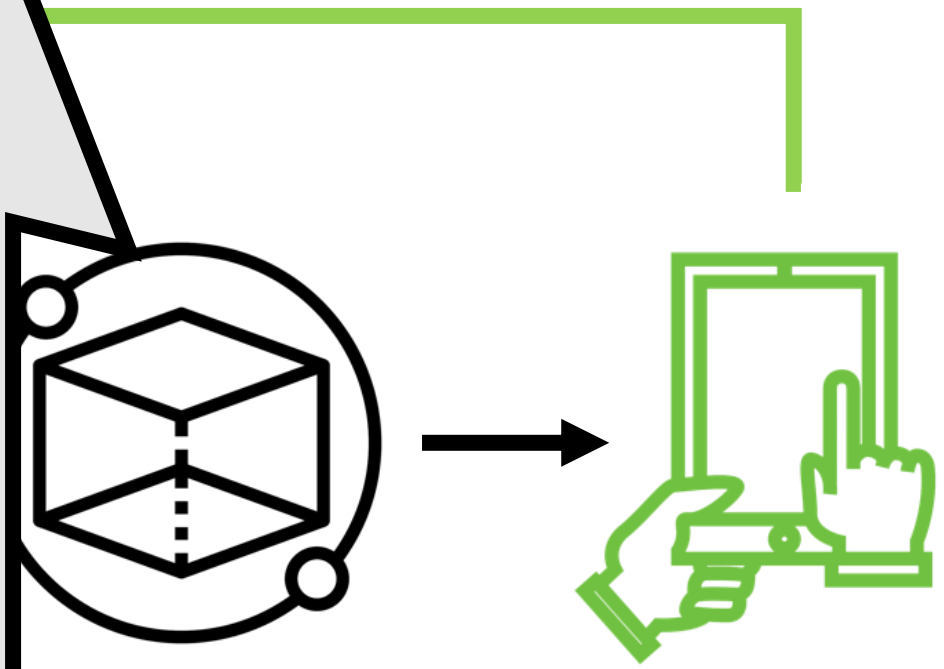
Consider **data provenance**
What is the data **intended to represent**?
Verify through qualitative, experimental, survey and other methods

Best Practices for Bias Avoidance/Mitigation

Check models and validate results

Why is the model making decision? What **mechanisms** would explain results? Is supporting evidence consistent?

Twyman's law: The more unusual the result, more likely it's an error



Model Application

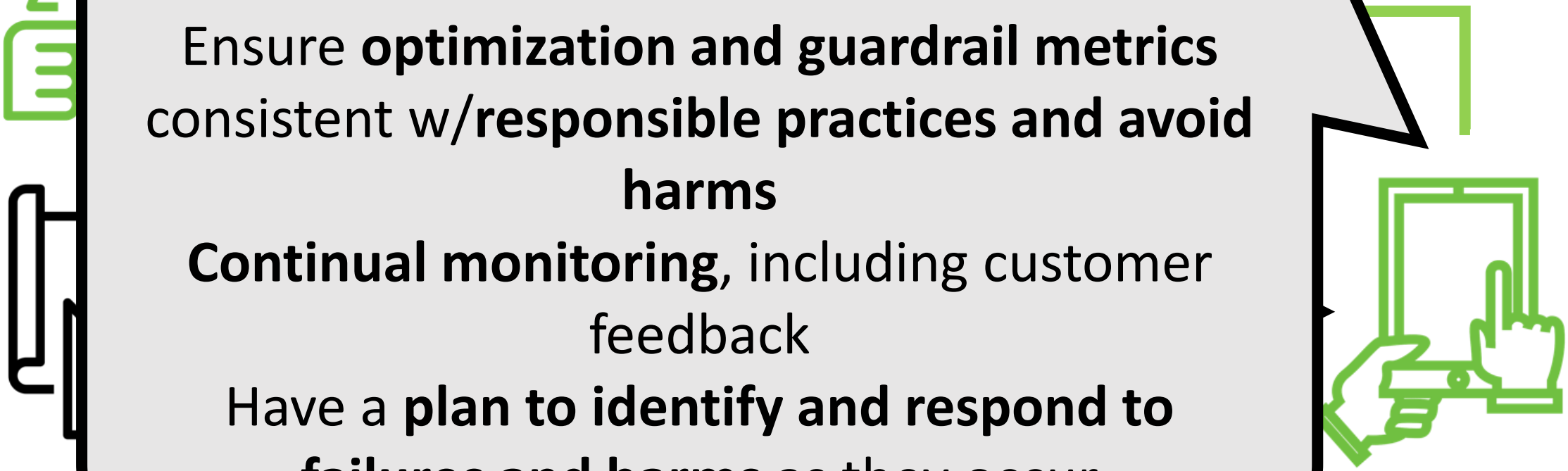
Best Practices for Bias Avoidance/Mitigation

Post-Deployment

Ensure **optimization and guardrail metrics** consistent w/**responsible practices and avoid harms**

Continual monitoring, including customer feedback

Have a **plan to identify and respond to failures and harms** as they occur



De

pplication

Techniques for Fairness in ML

Google's Responsible Fairness Practices

https://ai.google/education/responsible-ai-practices?category=fairness

Summary:

- **Design your product using concrete goals for fairness and inclusion.**
  - Engage with social scientists and other relevant experts.
  - Set fairness goals
- **Check system for unfair biases.**
  - Include diverse testers and adversarial/stress testing.
  - Consider feedback loops
- **Analyze performance.**
  - Evaluate user experience in real-world scenarios.
- **Use representative datasets to train and test your model.**

Techniques for Fairness in ML

1. Product Introspection
2. Practical Testing ← Main focus today
3. Training Data
4. Modeling
5. UI/Product Design

Product Introspection (1): Make Your Key Choices Explicit [Mitchell et al., 2018]

Goals	Decision	Prediction
Profit from loans	Whether to lend	Loan will be repaid
Justice, Public safety	Whether to detain	Crime committed if not detained

- Goals are ideally measurable
- What are your non-goals?
- Which decisions are you not considering?
- What is the relationship between Prediction and Decision?

Shad(e)vice™

You run an online sunglasses shop. You are designing a new feature that lets users upload photos in order to get automated suggestions based on their facial features.

LyricGram™

You have a music website for online discussion of song lyrics. In response to seeing increasing levels of online abuse, you design an automated comment moderation system.

Shad(e)vice™

You run an online sunglasses shop. You are designing a new feature that lets users upload photos in order to get automated suggestions based on their facial features.

Immediate goal: sell sunglasses to users

Long term goal: customer loyalty

Decision: recommend sunglasses?

Input: image

LyricGram™

You have a music website for online discussion of song lyrics. In response to seeing increasing levels of online abuse, you design an automated comment moderation system.

Shad(e)vice™

You run an online sunglasses shop. You are designing a new feature that lets users upload photos in order to get automated suggestions based on their facial features.

Immediate goal: sell sunglasses to users

Long term goal: customer loyalty

Decision: recommend sunglasses?

Input: image

LyricGram™

You have a music website for online discussion of song lyrics. In response to seeing increasing levels of online abuse, you design an automated comment moderation system.

Immediate goal: prevent distress

Long term goal: increase engagement

Decision: block comment?

Input: text

Q1: What prediction(s) will your ML system(s) make?
Q2: How will you use these predictions to make decisions?

Shad(e)vice™

You run an online sunglasses shop. You are designing a new feature that lets users upload photos in order to get automated suggestions based on their facial features.

Immediate goal: sell sunglasses to users

Long term goal: customer loyalty

Decision: recommend sunglasses?

Input: image

LyricGram™

You have a music website for online discussion of song lyrics. In response to seeing increasing levels of online abuse, you design an automated comment moderation system.

Immediate goal: prevent distress

Long term goal: increase engagement

Decision: block comment?

Input: text

Q1: What prediction(s) will your ML system(s) make?
Q2: How will you use these predictions to make decisions?

Shad(e)vice™

Immediate goal: sell sunglasses to users

Long term goal: customer loyalty

Decision: recommend sunglasses?

Input: image

Predictions???

LyricGram™

Immediate goal: prevent distress

Long term goal: increase engagement

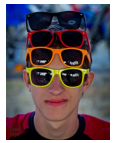
Decision: block comment?

Input: text

Predictions???

Q1: What prediction(s) will your ML system(s) make?
Q2: How will you use these predictions to make decisions?

Shad(e)vice™



Immediate goal: sell sunglasses to users

Long term goal: customer loyalty

Decision: recommend sunglasses?

Input: image

Predictions:

- CropFace(image)
- Purchase(face, sunglasses)
- FuturePurchases(face, sunglasses)

LyricGram™

Immediate goal: prevent distress

Long term goal: increase engagement

Decision: block comment?

Input: text

Predictions???

Q1: What prediction(s) will your ML system(s) make?
Q2: How will you use these predictions to make decisions?

Shad(e)vice™

Immediate goal: sell sunglasses to users

Long term goal: customer loyalty

Decision: recommend sunglasses?

Input: image

Predictions:

- CropFace(image)
- Purchase(face, sunglasses)
- FuturePurchases(face, sunglasses)

LyricGram™

Immediate goal: prevent distress

Long term goal: increase engagement

Decision: block comment?

Input: text

Predictions:

- Language(text)
- IsAbusive(text, language)

Product Introspection (2): Identify Potential Harms

- What are the potential harms?
  - Applicants who would have repaid are not given loans
  - Convicts who would not commit a crime are locked up.
- Are there also longer term harms?
  - Applicants are given loans, then go on to default, harming their credit score
- Are some harms especially bad?

Seek out Diverse Perspectives

- Fairness Experts
- User Researchers
- Privacy Experts
- Legal
- Social Science Backgrounds
- Diverse Identities
  - Gender
  - Sexual Orientation
  - Race
  - Nationality
  - Religion

Q3: What are potential user harms?
Q4: Which users are most likely to be affected?

Shad(e)vice™

Predictions:

- CropFace(image)
- Purchase(face, sunglasses)
- FuturePurchases(face, sunglasses)

User Harms: ???

Most impacted users: ?

LyricGram™

Predictions:

- Language(text)
- Abusive(text, language)

User Harms: ???

Most impacted users: ???

Q3: What are potential user harms?
Q4: Which users are most likely to be affected?

Shad(e)vice™

Predictions:

- CropFace(image)
- Purchase(face, sunglasses)
- FuturePurchases(face, sunglasses)

User Harms:

- Failure to detect face
- Poor sunglass recommendations

Most impacted users:

- Dark skin
- Cultural minorities

LyricGram™

Predictions:

- Language(text)
- Abusive(text, language)

User Harms: ???

Most impacted users: ???

Q3: What are potential user harms?
Q4: Which users are most likely to be affected?

Shad(e)vice™

Predictions:

- CropFace(image)
- Purchase(face, sunglasses)
- FuturePurchases(face, sunglasses)

User Harms:

- Failure to detect face
- Poor sunglass recommendations

Most impacted users:

- Dark skin
- Cultural minorities

LyricGram™

Predictions:

- Language(text)
- Abusive(text, language)

User Harms:

- Failure to protect from abuse
- Censoring of non-abusive comments

Most impacted users:

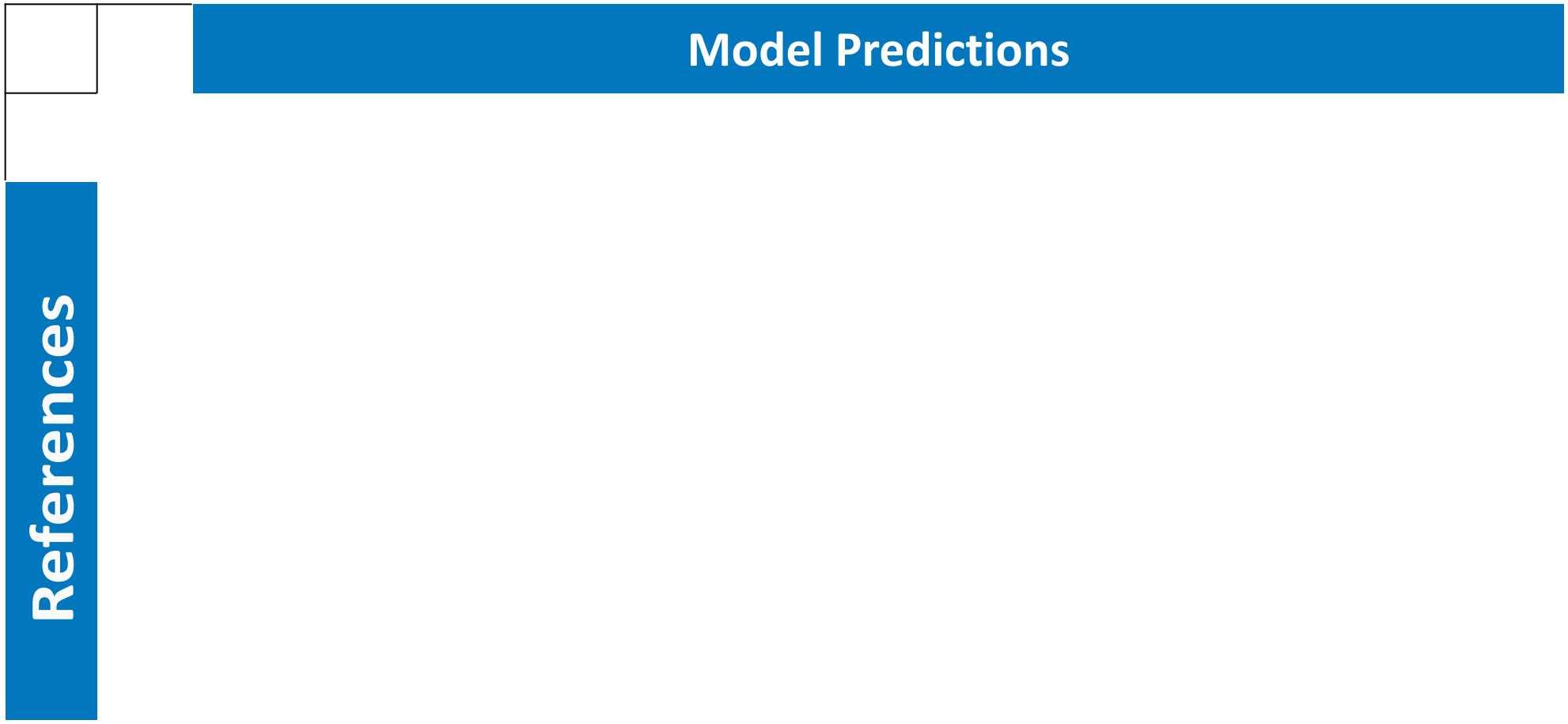
- Targets of hate speech
- People using LGBT terms positively
- Users discussing songs with sexual/violent lyrics

Launch with Confidence: Testing for Bias

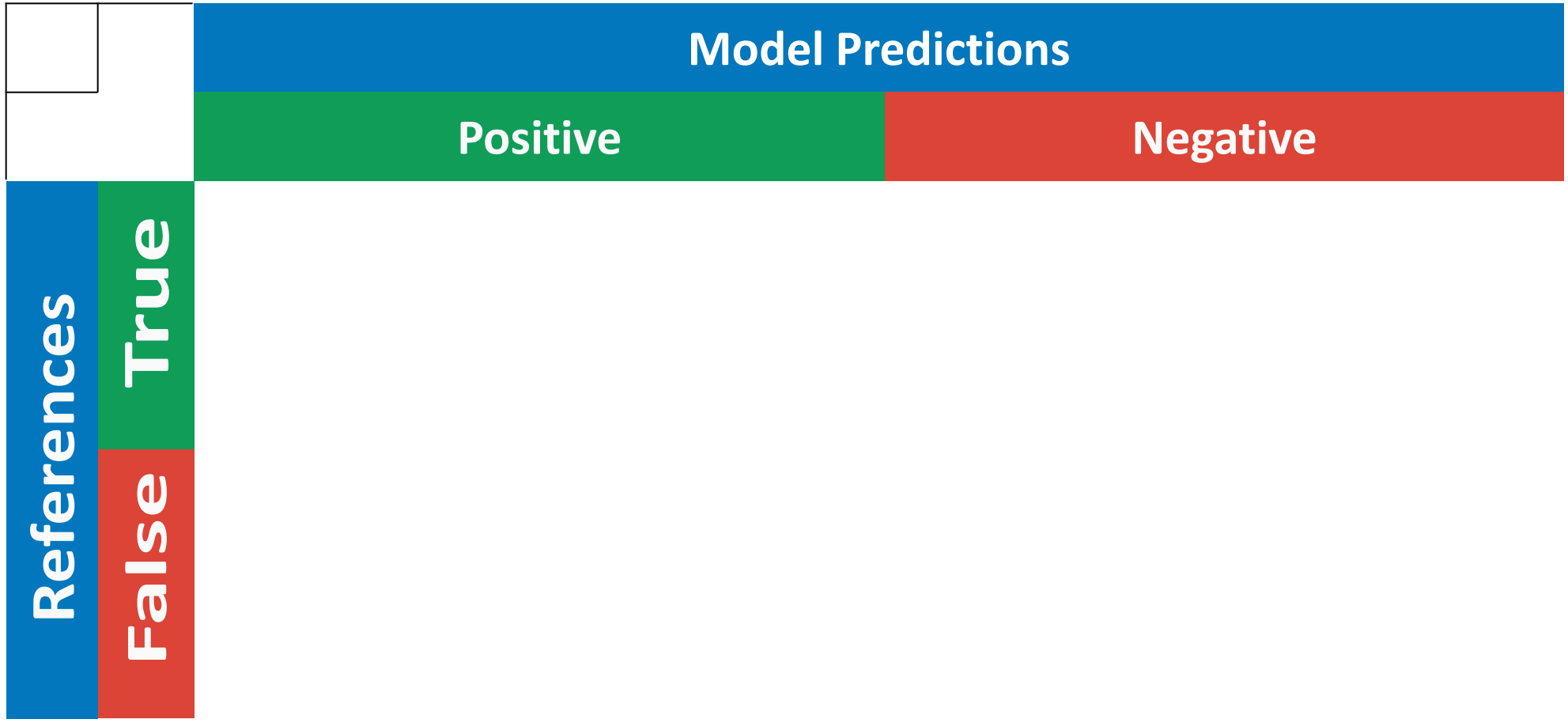
- How will you know if users are being harmed?
- How will you know if harms are unfairly distributed?

- Detailed testing practices are often not covered in academic papers
- Discussing testing requirements is a useful focal point for cross-functional teams

Evaluate for Inclusion - Confusion Matrix



Evaluate for Inclusion - Confusion Matrix



Evaluate for Inclusion - Confusion Matrix

		Model Predictions	
		Positive	**Negative**
References	**True**	• Exists • Predicted **True Positives**	
	False		• Doesn't exist • Not predicted **True Negatives**

Evaluate for Inclusion - Confusion Matrix

		Model Predictions	
		Positive	**Negative**
References	**True**	• Exists • Predicted **True Positives**	• Exists • Not predicted **False Negatives**
	False	• Doesn’t exist • Predicted **False Positives**	• Doesn’t exist • Not predicted **True Negatives**

Efficient Testing for Bias

- Development teams are under multiple constraints
  - Time
  - Money
  - Human resources
  - Access to data
- How can we efficiently test for bias?
  - Prioritization
  - Strategic testing

Choose your evaluation metrics in light of acceptable tradeoffs between **False Positives** and **False Negatives**

False Positives Might be Better than False Negatives

Privacy in Images

False Positive: Something that doesn't need to be blurred gets blurred.

Can be a bummer.

False Negative: Something that needs to be blurred is not blurred.

Identity theft.

False Negatives Might Be Better than False Positives

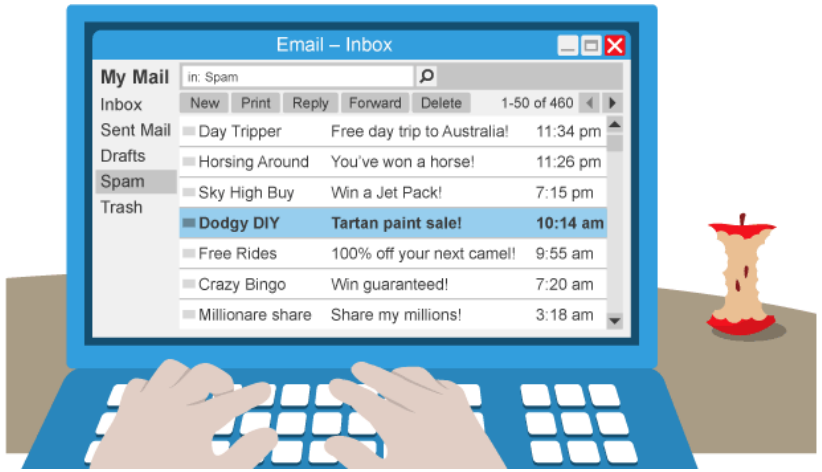
Spam Filtering

False Negative: Email that is SPAM is not caught, so you see it in your inbox.

Usually just a bit annoying.

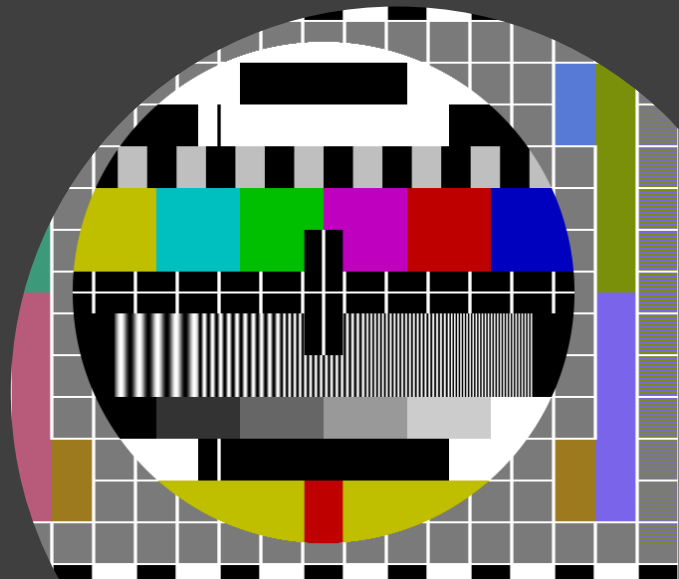
False Positive: Email flagged as SPAM is removed from your inbox.

If it's from a friend or loved one, it's a loss!



Types of Practical Fairness Testing

1. Targeted Tests
2. Quick Tests
3. Comprehensive Tests
4. Ecologically Valid Tests
5. Adversarial Testing



1. Targeted Tests

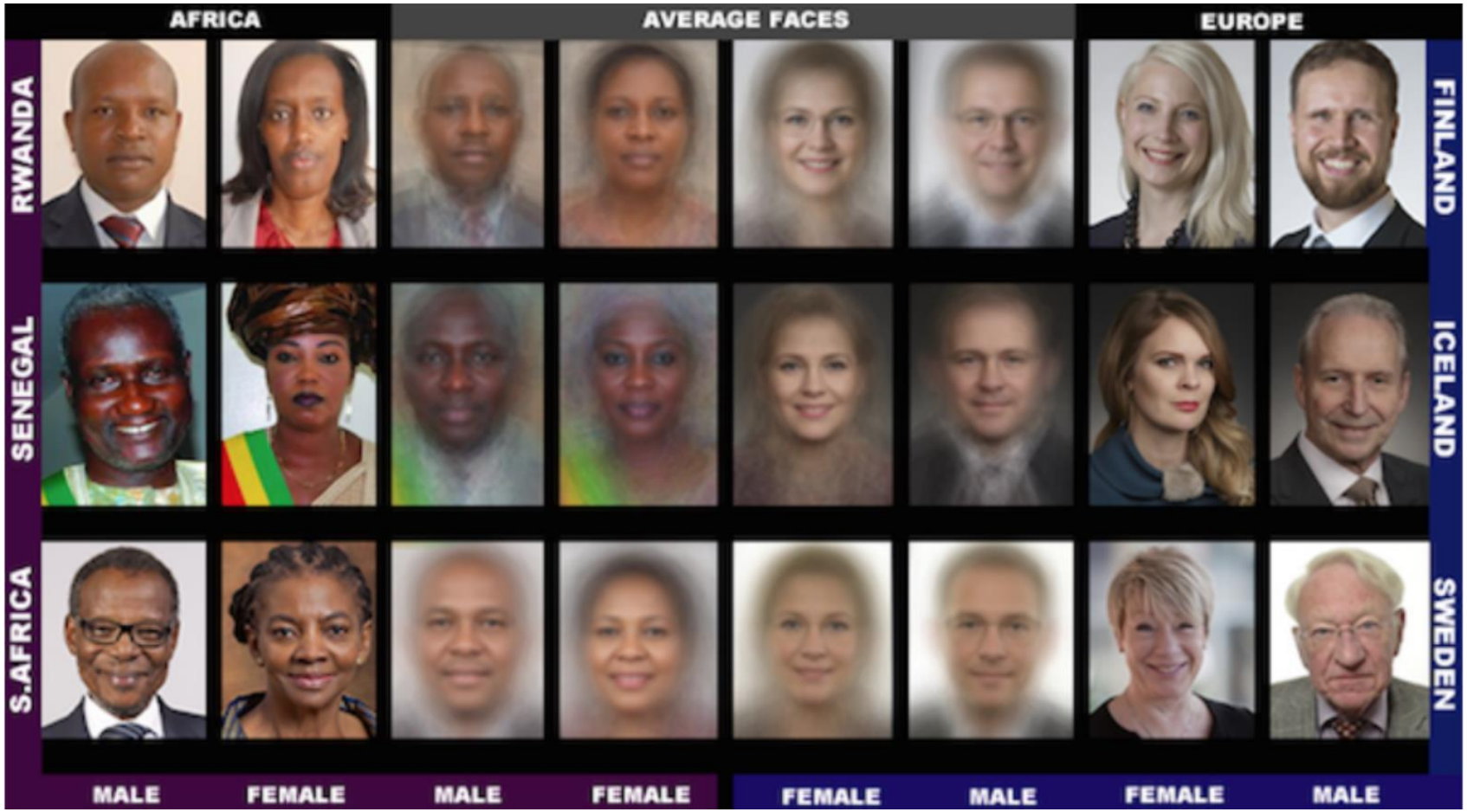
Based on prior experience/knowledge

- Computer Vision
  ⇒ Test for dark skin
- Natural Language Processing
  ⇒ Test for gender stereotypes

Cf. smoke tests
(non-exhaustive tests that check that most important functions work)

Targeted Testing of a Gender Classifier

[Joy Buolamwini & Timnit Gebru, 2018]



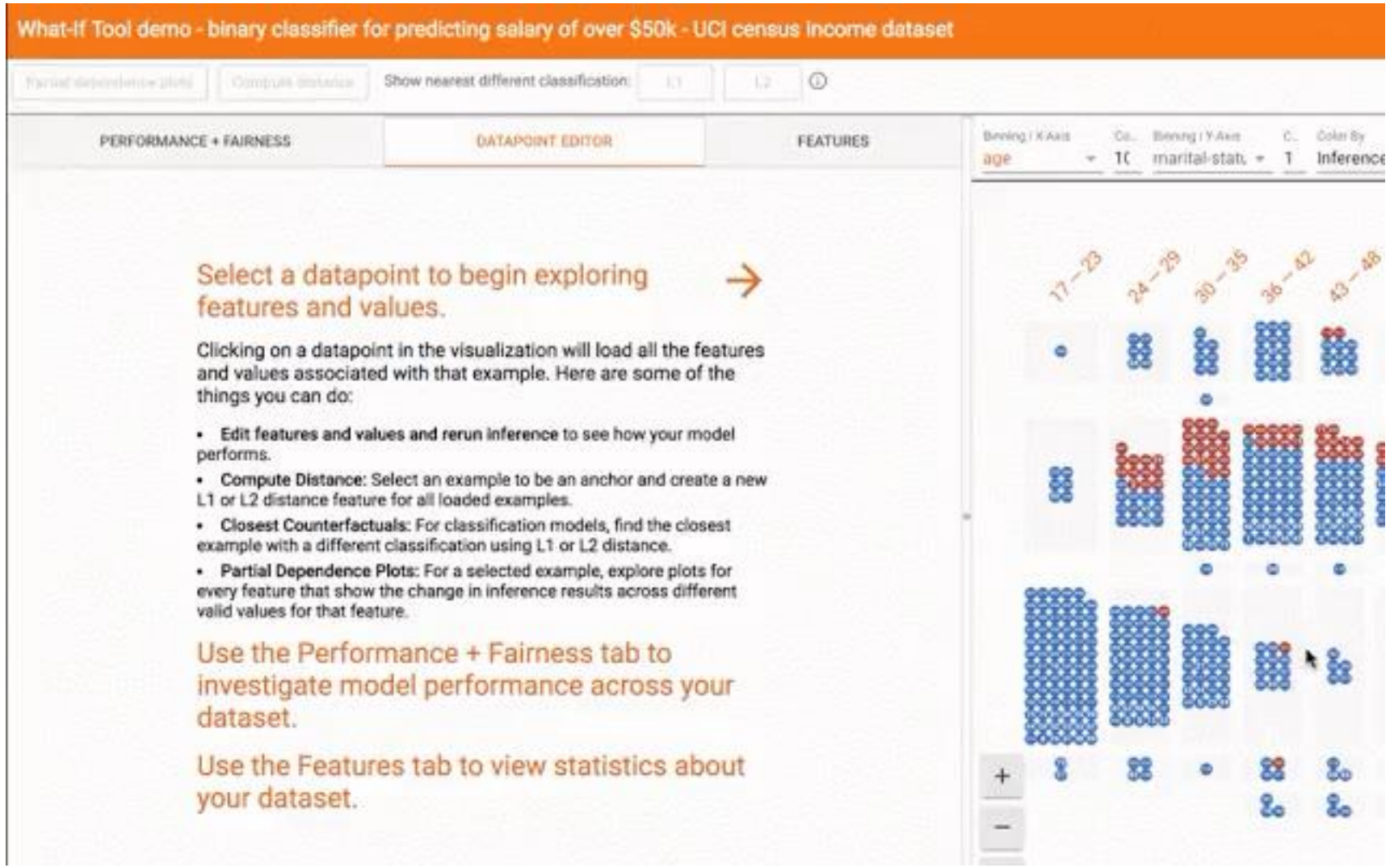
- Facial recognition software:
  Higher accuracy for light skinned men
- Error rates for dark skinned women:
  20% - 34%

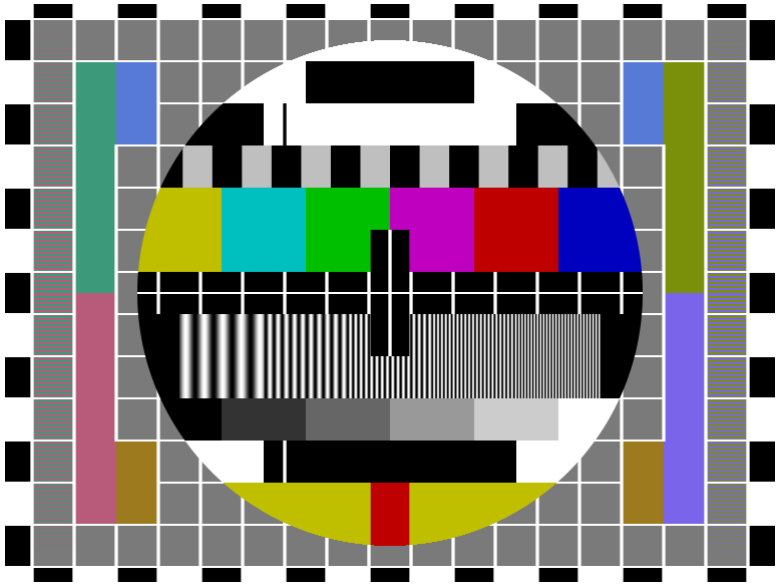
2. Quick Tests

- "Cheap"
- Useful throughout product cycle
- Spot check extreme cases
- Low coverage but high informativity
- Need to be designed thoughtfully, e.g.
  - World knowledge
  - Prior product failures

Quick Tests for Gender in Translate

Quick Counterfactual Testing: What If Tool





3. Comprehensive Tests

Include sufficient data for each subgroup

- May include relevant combinations of attributes
- Sometimes synthetic data is appropriate

Particularly important if model will be used in larger system

Cf. Unit tests

(verify correct outputs for wide range of correct inputs)

ConversationAI

We asked the internet what they thought about:

Climate Change Brexit US Election

Showing 47 of 49 total comments based on toxicity*

- They are liberal idiots who are uneducated.
- Climate change is happening and it's not changing in our favor. If you think differently you're an idiot.
- They're allowed to do that. But if they act like assholes about, I will block them.
- They're stupid, it's getting warmer, we should enjoy it while it lasts.
- I think those people are stupid and short-sighted
- I think its a farce and stinks like a bathroom after 26 beers

Problem: A False Positive Bias

Comment	Toxicity score
The Gay and Lesbian Film Festival starts today.	0.82
Being transgender is independent of sexual orientation.	0.52
A Muslim is someone who follows or practices Islam.	0.46

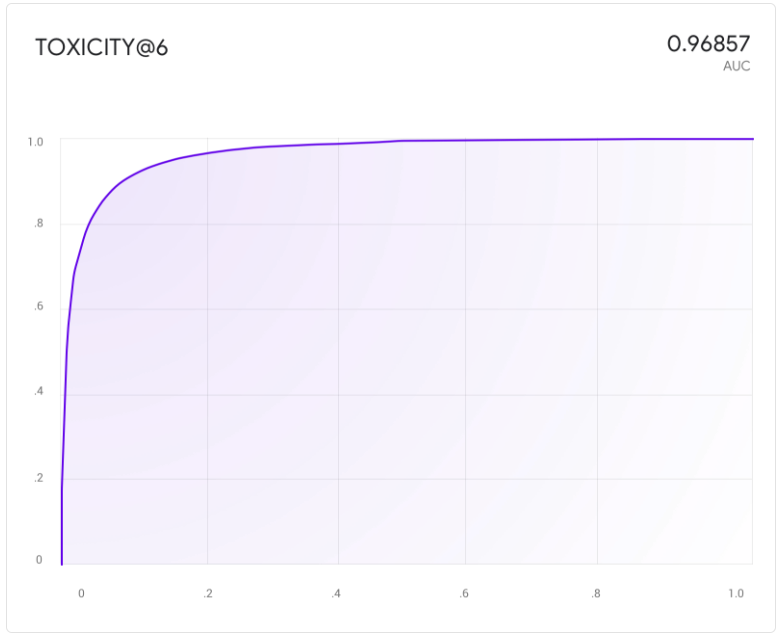
Comprehensive Testing of a Toxic Language Detector [Dixon et al., 2018]

	Comment Length				
Term	20-59	60-179	180-539	540-1619	1620-4859
ALL	17%	12%	7%	5%	5%
gay	88%	77%	51%	30%	19%
queer	75%	83%	45%	56%	0%
homosexual	78%	72%	43%	16%	15%
black	50%	30%	12%	8%	4%
white	20%	24%	16%	12%	2%
wikipedia	39%	20%	14%	11%	7%
atheist	0%	20%	9%	6%	0%
lesbian	33%	50%	42%	21%	0%
feminist	0%	20%	25%	0%	0%
islam	50%	43%	12%	12%	0%
muslim	0%	25%	21%	12%	17%
race	20%	25%	12%	10%	6%
news	0%	1%	4%	3%	3%
daughter	0%	7%	0%	7%	0%

Figure 1: Percent of comments labeled as toxic at each length containing the given terms.

Template Examples	Label
I am <IDENTITY>	Non-Toxic
I am a <IDENTITY> person, ask me anything	Non-Toxic
<IDENTITY> people are just like everyone else	Non-Toxic
I hate all <IDENTITY>	Toxic
I am a <IDENTITY> person and I hate your guts and think you suck	Toxic
<IDENTITY> people are gross and universally terrible	Toxic

Table 2: Phrase template examples.



AUC Metrics for Comprehensive Testing

- Subgroup AUC:
  - Subgroup Positives vs Subgroup Negatives
- "BPSN" AUC:
  - Background Positives vs Subgroup Negatives
- "BNSP" AUC:
  - Background Negatives vs Subgroup Positives

Comprehensive Testing of a Toxicity Detector

https://github.com/conversationai/perspectiveapi/blob/master/model_cards/English/toxicity.md

Toxicity @1

Identity groups	Subgroup AUC	BPSN AUC	BNSP AUC
lesbian	0.93	0.74	0.98
gay	0.94	0.65	0.99
queer	0.98	0.96	0.93
straight	0.99	1.00	0.87
bisexual	0.96	0.95	0.92
homosexual	0.87	0.53	0.99
heterosexual	0.96	0.94	0.92
cis	0.99	1.00	0.87
trans	0.97	0.96	0.91
nonbinary	0.99	0.99	0.90
black	0.91	0.85	0.95
white	0.91	0.88	0.94

0.5 0.6 0.7 0.8 0.9 1.0

Toxicity @6

Identity groups	Subgroup AUC	BPSN AUC	BNSP AUC
lesbian	1.00	0.98	1.00
gay	1.00	0.94	1.00
queer	0.99	0.98	0.99
straight	1.00	1.00	0.97
bisexual	0.98	0.98	0.99
homosexual	1.00	0.96	1.00
heterosexual	1.00	0.99	1.00
cis	1.00	1.00	0.98
trans	1.00	1.00	1.00
nonbinary	1.00	1.00	0.98
black	0.98	0.97	1.00
white	0.99	0.99	0.99

0.5 0.6 0.7 0.8 0.9 1.0

Inclusive Images Competition

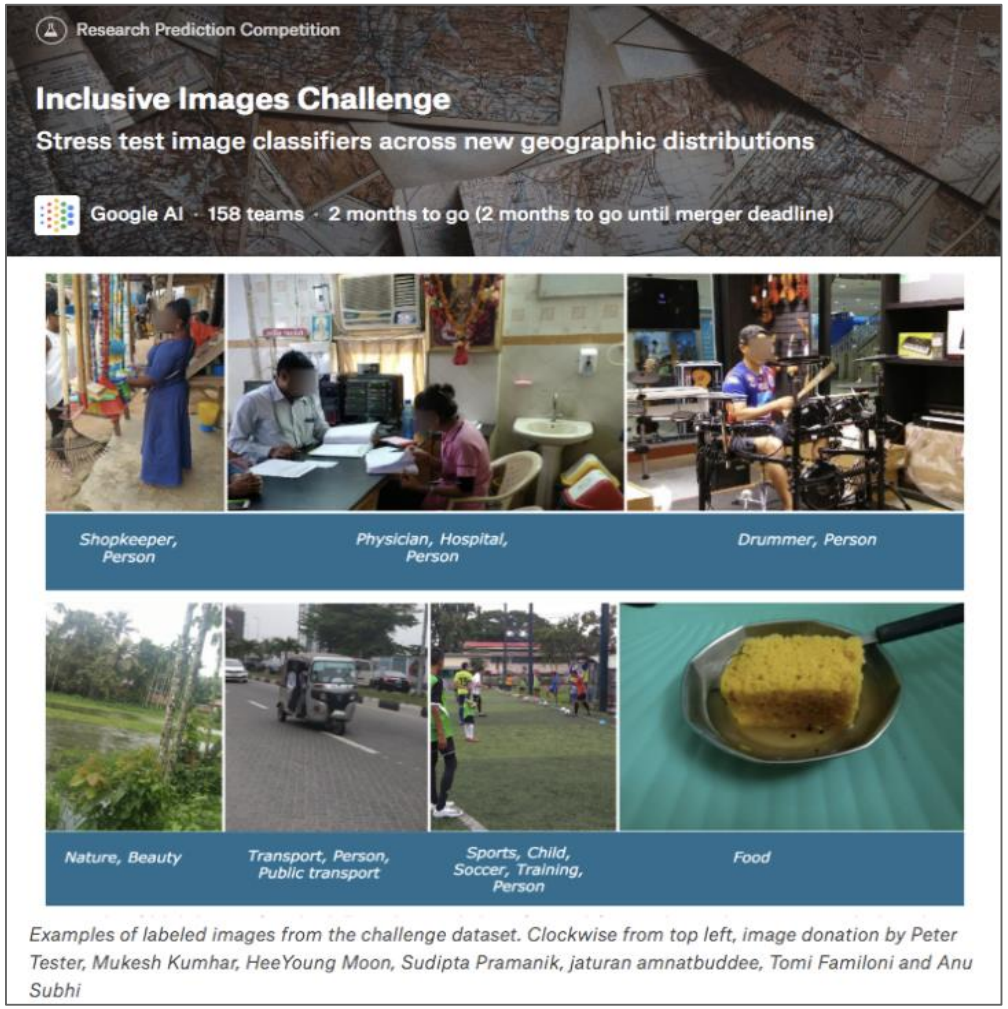
Introducing the Inclusive Images Competition

Thursday, September 6, 2018

Posted by Tulsee Doshi, Product Manager, Google AI

The release of large, publicly available image datasets, such as ImageNet, Open Images and Conceptual Captions, has been one of the factors driving the tremendous progress in the field of computer vision. While these datasets are a necessary and critical part of developing useful machine learning (ML) models, some open source data sets have been found to be geographically skewed based on how they were collected. Because the shape of a dataset informs what an ML model learns, such skew may cause the research community to inadvertently develop models that may perform less well on images drawn from geographical regions under-represented in those data sets. For example, the images below show one standard open-source image classifier trained on the Open Images dataset that does not properly apply *"wedding"* related labels to images of wedding traditions from different parts of the world.

Wedding photographs (donated by Googlers), labeled by a classifier trained on the Open Images dataset. The classifier's label predictions are recorded below each image.



Examples of labeled images from the challenge dataset. Clockwise from top left, image donation by Peter Tester, Mukesh Kumhar, HeeYoung Moon, Sudipta Pramanik, jaturan amnatbuddee, Tomi Familoni and Anu Subhi

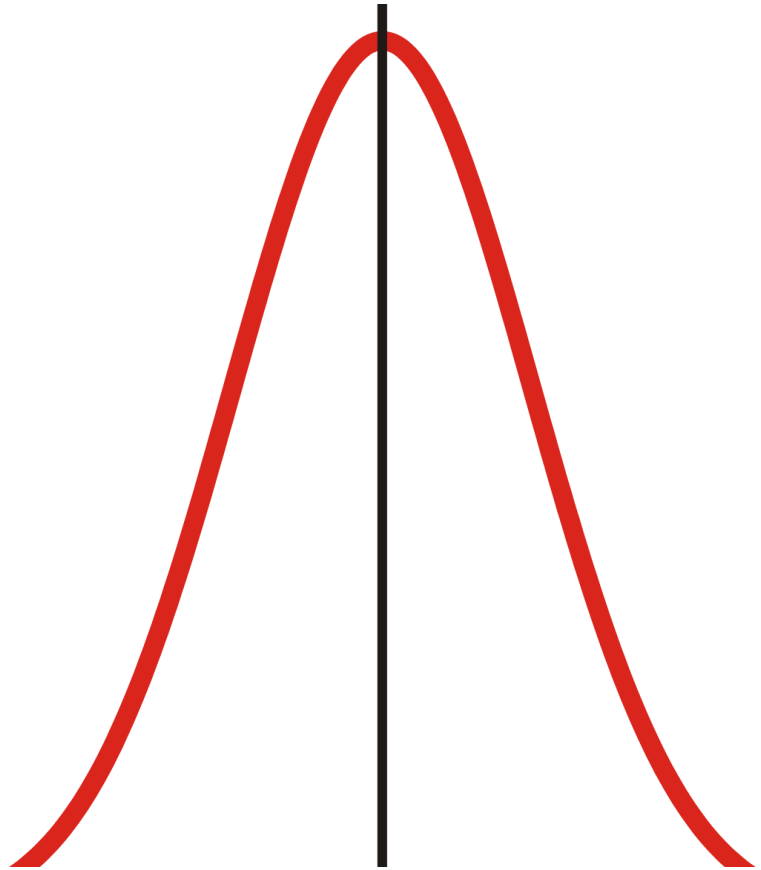
4. Ecologically Valid Testing

Data is drawn from a distribution representative of the **deployment** distribution

- Goal is NOT to be representative of the training distribution
- (When appropriate) Condition on labels & certainty

Example usage scenarios :

- Continuous monitoring
- You have historical product usage data
- You can estimate user distribution reasonably well



Ecologically Valid Testing: Distributions Matter

What is being compared?

Over what data?

Challenges with Ecologically Valid Testing

- Post-deployment distributions may not be known
  - Product may not be launched yet!
- Sensitive attributes often not available in deployment
- User distributions may change
- We may want user distributions to change
  - e.g., broaden user base

5. Adversarial Tests

Search for rare but extreme harms

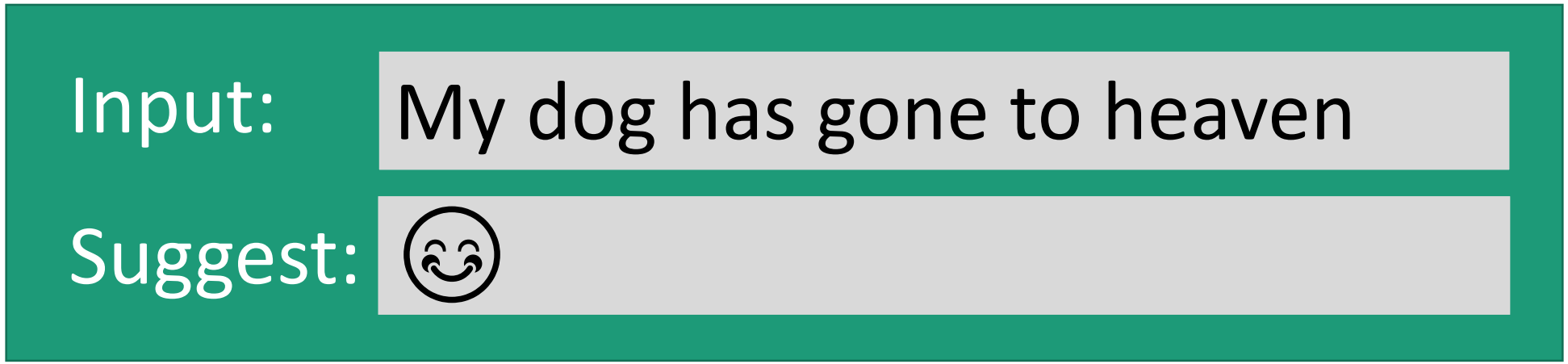
- “Poison needle in haystack”
- Requires knowledge of society

Typical usage scenario:

- Close to launch

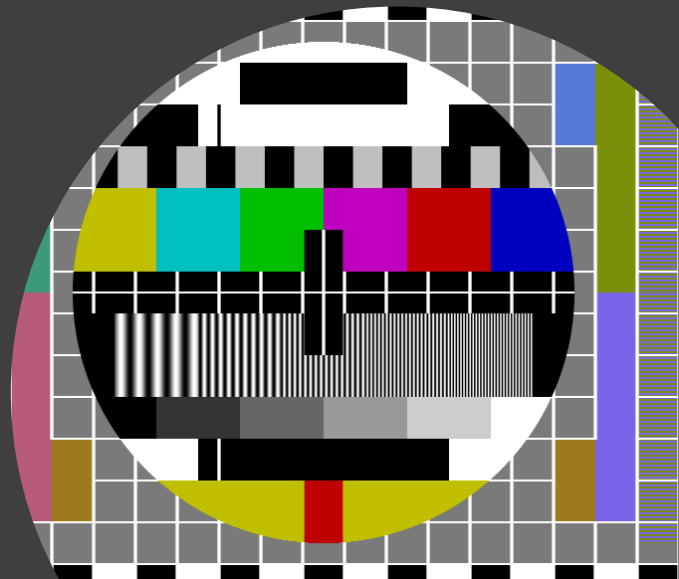
Hypothetical Example of Adversarial Testing

- Emoji autosuggest: are happy emoji suggested for sad sentences?



Summary of Practical Fairness Testing

1. Targeted Tests: domain specific (image, language, etc)
2. Quick Tests: cheap tests throughout dev cycle
3. Comprehensive Tests: thorough
4. Ecologically Valid Tests: real-world data
5. Adversarial Testing: find poison needles



Fairness Testing Practices are Good ML Practices

- Confidence in your product's fairness requires fairness testing
- Fairness testing has a role throughout the product iteration lifecycle
- Contextual concerns should be used to prioritize fairness testing

Q5: How could you test for fairness?

Shad(e)vice™

Predictions:

- CropFace(image)
- Purchase(face, sunglasses)
- FuturePurchases(face, sunglasses)

User Harms:

- Failure to detect face
- Poor sunglass recommendations

Most impacted users:

- Dark skin
- Cultural minorities

LyricGram™

Predictions:

- Language(text)
- Abusive(text, language)

User Harms: ???

- Failure to detect from abuse
- Censoring of non-abusive comments

Most impacted users: ???

- Groups which are targets of hate speech
- Groups writing comments
  - confusable for abuse
  - in multiple languages

Q5: How could you test for fairness?

Shad(e)vice™

Predictions:

- CropFace(image)
- Purchase(face, sunglasses)
- FuturePurchases(face, sunglasses)

User Harms:

- Failure to detect face
- Poor sunglass recommendations

Most impacted users:

- Dark skin
- Cultural minorities

- targeted testing:
  - faces with darker skin tones
- comprehensive testing:
  - diverse gender expression (male/female/androgynous)
  - diverse headwear (caps, hijabs, turbans, ...)
- ecologically valid testing:
  - compare sales conversions for users of different skin tones and locations
- adversarial testing:
  - does system recommend sunglasses which are offensive or have cultural, religious or political associations?

Q5: How could you test for fairness?

LyricGram™

Predictions:

- Language(text)
- Abusive(text, language)

User Harms: ???

- Failure to detect from abuse
- Censoring of non-abusive comments

Most impacted users: ???

- Targets of hate speech
- People using LGBT terms positively
- Users discussing songs with sexual/violent lyrics

- targeted testing:
  - comments about gangster rap lyrics
  - comments using LGBT identity terms
- comprehensive testing:
  - comments using many identity terms
  - comments in many languages
- ecologically valid testing:
  - compare error rates for different song genres
  - compare error rates for users in different locations
- adversarial testing:
  - comments with offensive slang which model may not have seen before

Fairness-aware Data Collection

[Holstein et al., 2019]

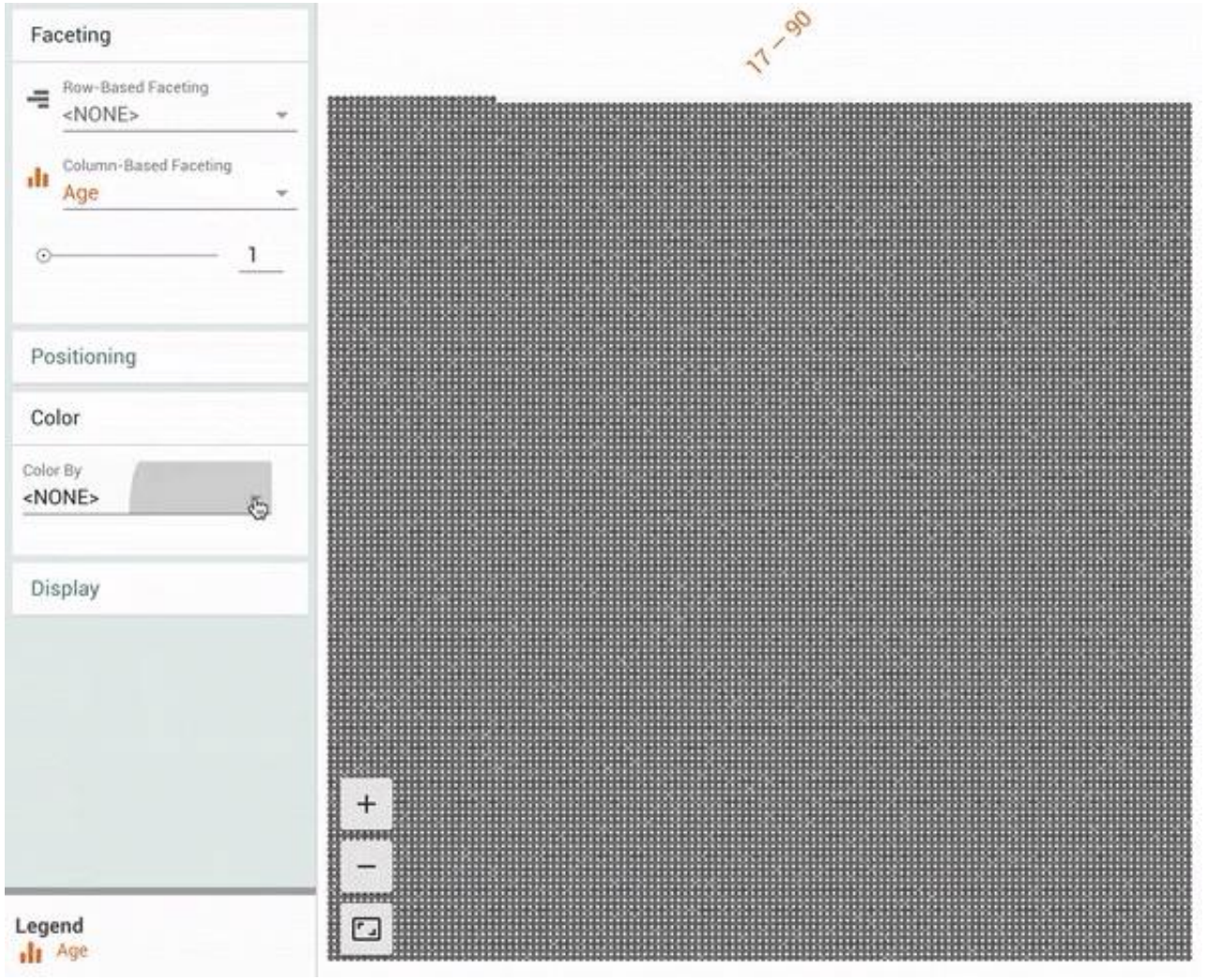
- ML literature generally assumes data is fixed
- Often the solution is more and/or better training data

But: need to be Thoughtful!

When might more Data not Help?

- If your data sampling techniques are biased
- Fundamental problems in data quality [Eckhouse et al., 2018]
  - What does your data *really* represent? E.g. crimes vs arrests
  - Recall: Product Introspection: How do Predictions relate to Decisions?

Get to Know Your Training Data: Facets Dive



Datasheets for Datasets [Gebru et al., 2018]

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or 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? Are there obvious tasks for which it should *not* be used?

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? If there is an associated grant, provide the grant number.

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 of each type are there?

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.

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 are 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. Are there licenses, fees or rights associated with *any* of the data?

Everything is included in the dataset.

Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/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:

$$\hat{\mu} = \frac{\sum_{i=1}^{10} p_i}{10} \quad (1)$$

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

$$S_E = \frac{\hat{\sigma}}{\sqrt{10}} \quad (2)$$

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

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

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.

[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

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

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 and, if available, provide the link to a paper with more information here.

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; how were these constructs/measures/methods validated?)

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) How were they compensated? (e.g., how much were crowdworkers paid?)

Unknown

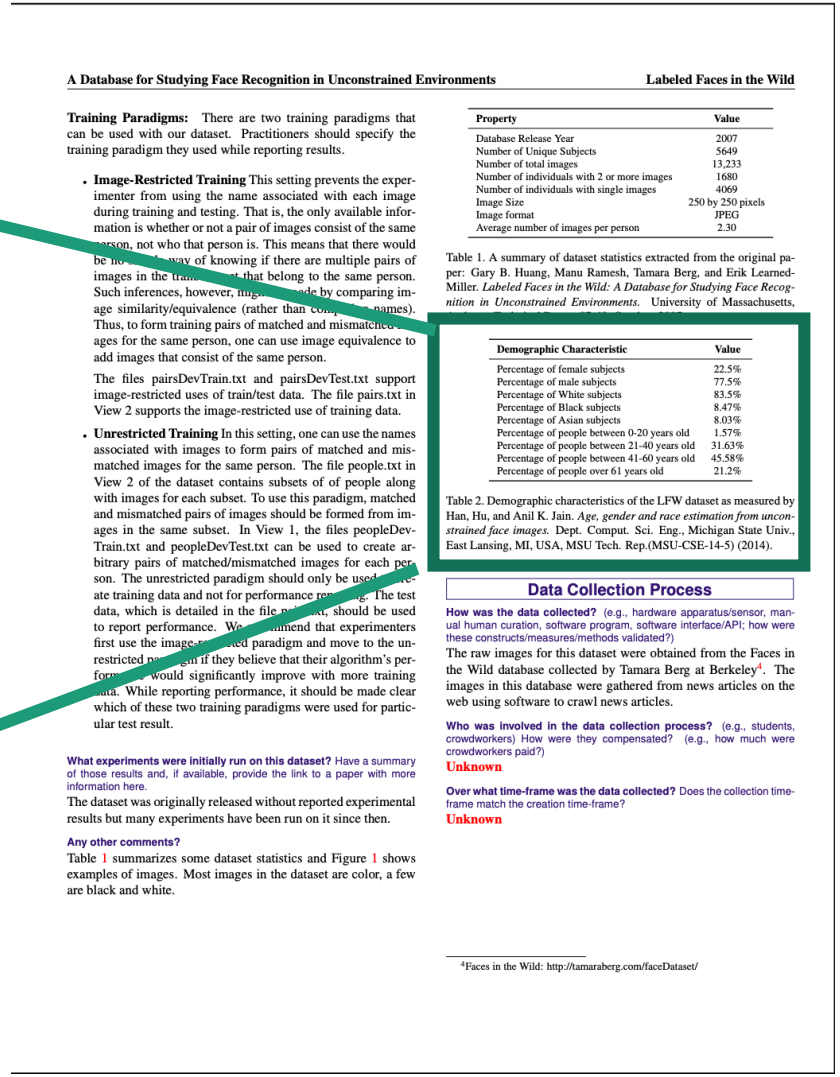
Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

Unknown

[4] Faces in the Wild: http://tamaraberg.com/faceDataset/

Datasets for Datasets [Gebru et al., 2018]

Demographic Characteristic	Value
Percentage of female subjects	22.5%
Percentage of male subjects	77.5%
Percentage of White subjects	83.5%
Percentage of Black subjects	8.47%
Percentage of Asian subjects	8.03%
Percentage of people between 0-20 years old	1.57%
Percentage of people between 21-40 years old	31.63%
Percentage of people between 41-60 years old	45.58%
Percentage of people over 61 years old	21.2%



A 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 [illegible] way of knowing if there are multiple pairs of images in the training set that belong to the same person. Such inferences, however, might be made by comparing image similarity/equivalence (rather than [illegible] 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 and, if available, provide the link to a paper with more information here.
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, [illegible]

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; how were these constructs/measures/methods validated?)
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) 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?
Unknown

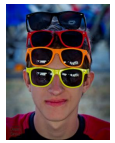
[4]Faces in the Wild: http://tamaraberg.com/faceDataset/

Fairness-Aware Data Collection Techniques

1. Address population biases
   - Target under-represented (with respect to the user population) groups
2. Address representation issues
   - Oversample from minority groups
   - Sufficient data from each group may be required to avoid model treating them as "outliers"
3. Data augmentation: synthesize data for minority groups
   - E.g. from observed "he is a doctor" → synthesize "she is a doctor"
4. Fairness-aware active learning
   - Collect more data for group with highest error rates

Q6: How could you improve your training data?

Shad(e)vice™



Predictions:

- CropFace(image)
- Purchase(face, sunglasses)
- FuturePurchases(face, sunglasses)

User Harms:

- Failure to detect face
- Poor sunglass recommendations

Most impacted users:

- Dark skin
- Cultural minorities

LyricGram™

Predictions:

- Language(text)
- Abusive(text, language)

User Harms:

- Failure to protectfrom abuse
- Censoring of non-abusive comments

Most impacted users:

- Groups which are targets of hate speech
- Groups writing comments
  - confusable for abuse
  - in multiple languages

Q6: How could you improve your training data?

Shad(e)vice™

Predictions:

- CropFace(image)
- Purchase(face, sunglasses)
- FuturePurchases(face, sunglasses)

User Harms:

- Failure to detect face
- Poor sunglass recommendations

Most impacted users:

- Dark skin
- Cultural minorities

- Include more faces from under-represented groups
- Include historical data from offline sales (rather than just online sales)
- Include more recent data (more predictive of fashion trends)

Q6: How could you improve your training data?

LyricGram™

Predictions:

- Language(text)
- Abusive(text, language)

User Harms:

- Failure to detect from abuse
- Censoring of non-abusive comments

Most impacted users:

- Groups which are targets of hate speech
- Groups writing comments
  - confusable for abuse
  - in multiple languages

- Include more comments data related to under-represented music genres
- Include comments data unrelated to music
  - especially from minority language variants
- Include more recent data (discussing new songs)

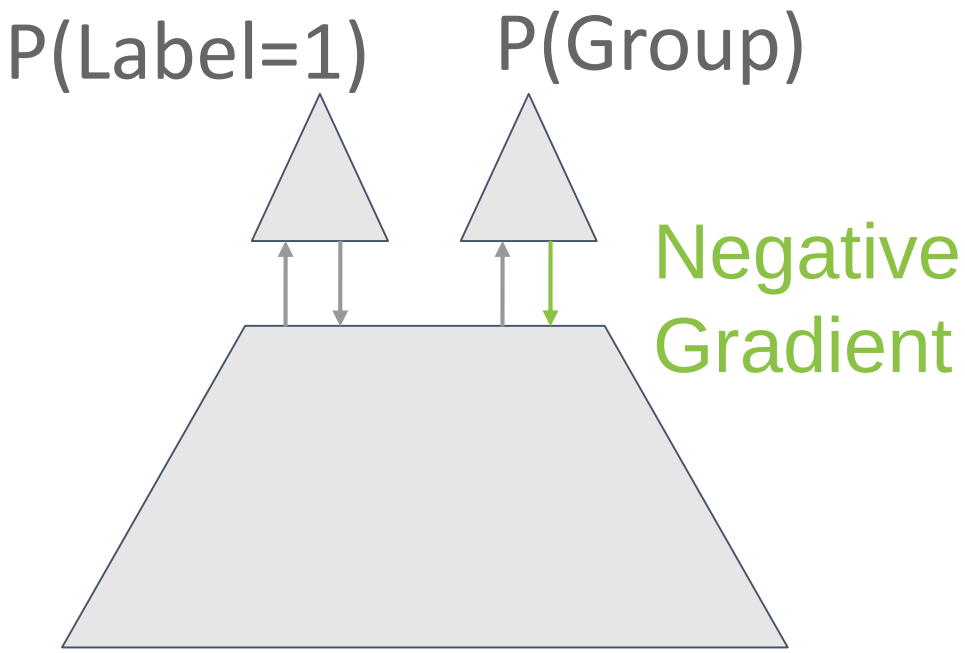
Sometimes data biases are unavoidable

Solution: ML Techniques

Practical Concerns with Fair Machine Learning

- Is the training process stable?
- Can we guarantee that fairness policies will be satisfied?
  - Cf. Legal requirements in education, employment, finance

Machine Learning Techniques: Adversarial Training?



Features, Label, Group

Fairly well-studied with some nice theoretical guarantees.

But can be difficult to train.

Machine Learning: Correlation Loss

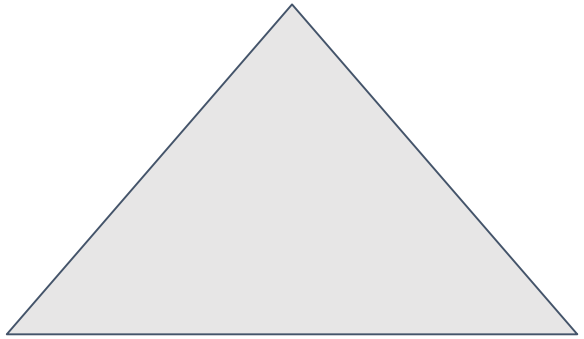
[Beutel et al., 2018]

Motivation: Overcome training instability with adversarial training

Key idea: include fairness objective in the loss function

Machine Learning Techniques: Correlation Loss

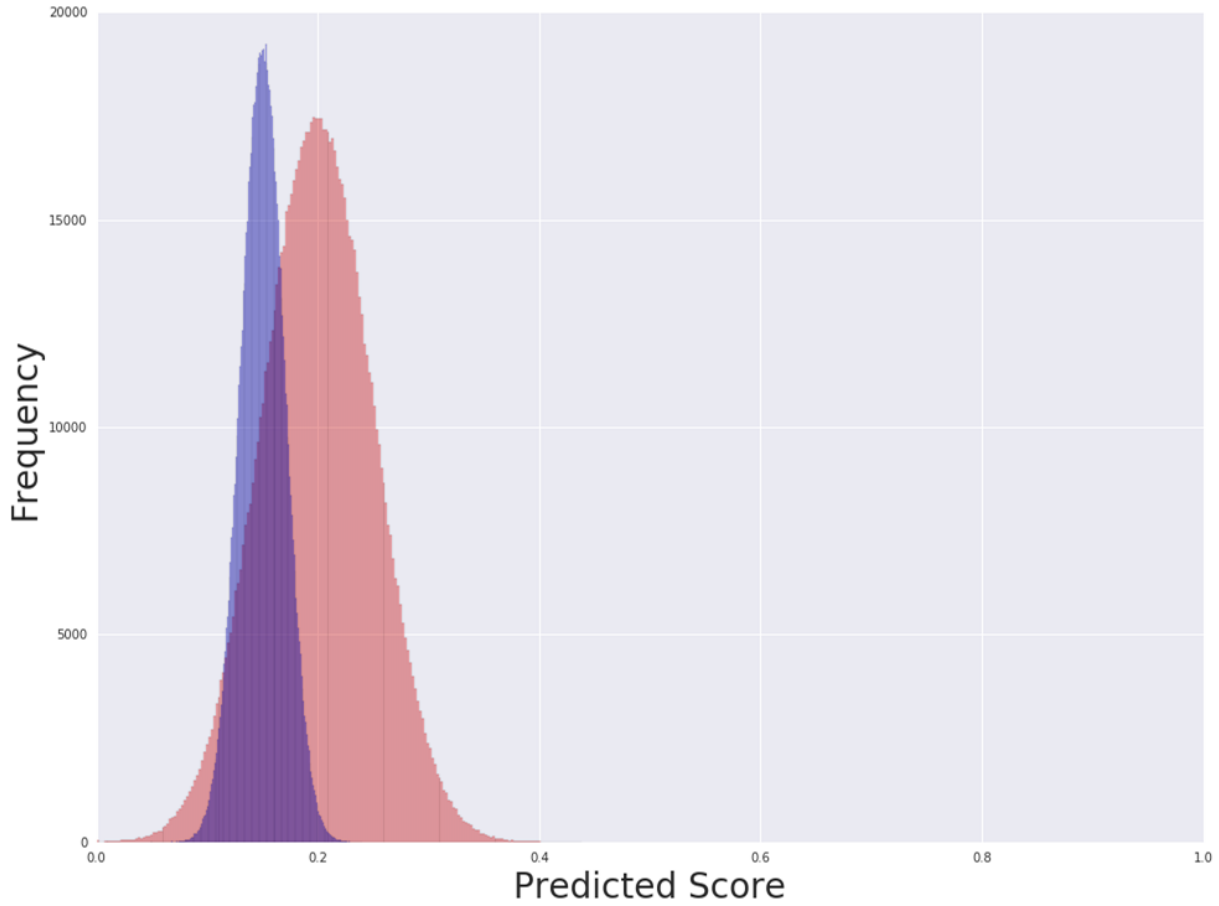
Pred = P(Label=1)



Features, Label, Group

min Loss(Label, Pred)

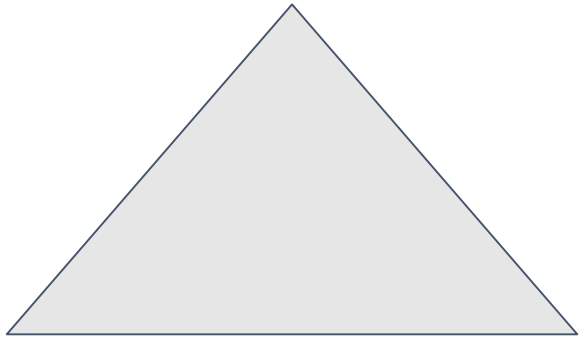
Predicted P(Target) distribution
for “Blue” and “Red” examples



(Illustrative Example)

Machine Learning Techniques: Correlation Loss

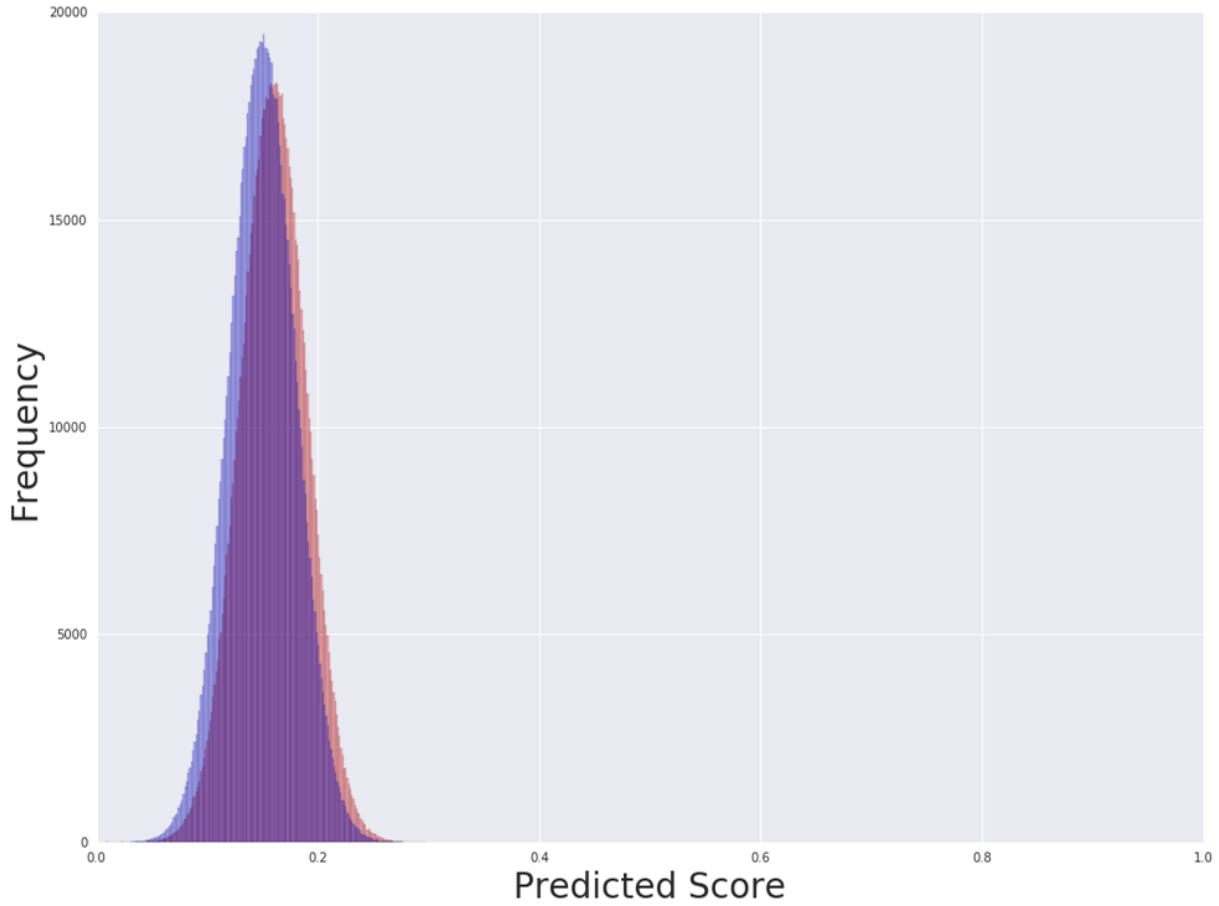
Pred = P(Label=1)



Features, Label, Group

min Loss(Label, Pred)
 + Abs(Corr(Pred, Group))|Label=0

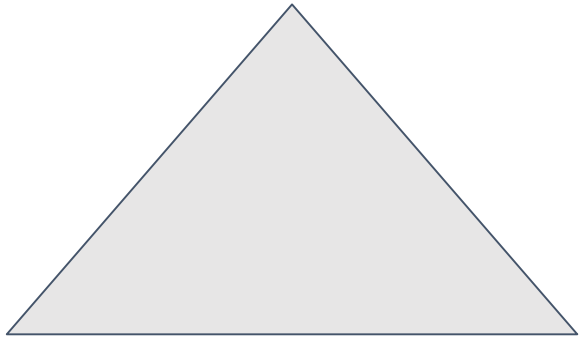
Predicted P(Target) distribution
for “Blue” and “Red” examples



(Illustrative Example)

Machine Learning Techniques: Correlation Loss

Pred = P(Label=1)



Features, Label, Group

- Computed per batch
- Easy to use
- More stable than adversarial training.

min Loss(Label, Pred)
 + Abs(Corr(Pred, Group))|Label=0

Machine Learning: Constrained Optimization

[Cotter et al., 2018]

Motivation: Can we ensure that fairness policies are satisfied?

- Fairness goals are explicitly stated as constraints on predictions, e.g.
  - FPR on group 1 <= 0.8 * FPR on group 2
- Machine learner optimizes objective function subject to the constraints

Model Cards for Model Reporting [Mitchell et al., 2018]

Intended use

Human-assisted moderation

Make moderation easier with an ML assisted tool that helps prioritize comments for human moderation, and create custom tasks for automated actions. See our moderator tool as an example.

Author feedback

Assist authors in real-time when their comments might violate your community guidelines or be may be perceived as "Toxic" to the conversation. Use simple feedback tools when the assistant gets it wrong. See our authorship demo as an example.

Read better comments

Organize comments on topics that are often difficult to discuss online. Build new tools that help people explore the conversation.

Uses to avoid

Fully automated moderation

Perspective is not intended to be used for fully automated moderation. Machine learning models will always make some mistakes, so it is essential to build in systems for humans to catch and correct those mistakes.

Character judgement

In order to maintain user privacy, the TOXICITY model only helps detect toxicity in an individual statement, and is not intended to detect anything about the individual who said it. In addition, Perspective does not use prior information about an individual to inform toxicity predictions.

Model details

Training data

Proprietary from Perspective API, which includes comments from online forums such as Wikipedia and New York Times, with crowdsourced labels of whether the comment is "toxic", defined as "a rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion".

Model architecture

The model is a Convolutional Neural Network (CNN) trained with GloVe word embeddings, which are fine-tuned during training. You can also train your own deep CNN for text classification on our public toxicity dataset, and explore our open-source model training tools to train your own models.

Values

Community, Transparency, Inclusivity, Privacy, and Topic neutrality. These values guide our product and research decisions.

Further Machine Learning Techniques

Many more approaches are linked to from the tutorial website.

https://sites.google.com/corp/view/www19-fairness-tutorial

Fairness in UI/Product Design

1. Robust UIs handle ML failures gracefully
2. UIs should empower users

Q7: How could UI Help to Mitigate Unfairness?

Shad(e)vice™

Predictions:

- CropFace(image)
- Purchase(face, sunglasses)
- FuturePurchases(face, sunglasses)

User Harms:

- Failure to detect face
- Poor sunglass recommendations

Most impacted users:

- Dark skin
- Cultural minorities

LyricGram™

Predictions:

- Language(text)
- Abusive(text, language)

User Harms: ???

- Failure to detect from abuse
- Censoring of non-abusive comments

Most impacted users: ???

- Groups which are targets of hate speech
- Groups writing comments
  - confusable for abuse
  - in multiple languages

Q7: How could UI Help to Mitigate Unfairness?

Shad(e)vice™

Predictions:

- CropFace(image)
- Purchase(face, sunglasses)
- FuturePurchases(face, sunglasses)

User Harms:

- Failure to detect face
- Poor sunglass recommendations

Most impacted users:

- Dark skin
- Cultural minorities

- Be graceful when face detection fails
  - Still give some suggestions!
  - Use additional signals rather than just face
- Let users crop image to face themselves
- Give users some control in guiding predictions

Q7: How could UI Help to Mitigate Unfairness?

LyricGram™

Predictions:

- Language(text)
- Abusive(text, language)

User Harms:

- Failure to detect from abuse
- Censoring of non-abusive comments

Most impacted users:

- Groups which are targets of hate speech
- Groups writing comments
  - confusable for abuse
  - in multiple languages

- UI might distinguish quoted lyrics from user comments about those lyrics
- Ability to flag comments for review
- Ability to appeal automated censoring decisions

Thanks to

Alex Beutel (Research Scientist, fairness in ML),
Allison Woodruff (UX Research, privacy, fairness and ethics),
Andrew Zaldivar (Developer Advocate, ethics and fairness in AI),
Hallie Benjamin (Senior Strategist, ethics and fairness in ML),
Jamaal Barnes (Program Manager, fairness in ML),
Josh Lovejoy (UX Designer, People and AI Research; now Microsoft),
Margaret Mitchell (Research Scientist, ethics and fairness in AI),
Rebecca White (Program Manager, fairness in ML)
and others!

Fairness Methods in Practice (Case Studies)

Google Assistant

[Play Google Assistant video from desktop]

Key points:

- Think about user harms
    - How does your product make people feel?
- Adversarial ("stress") testing for all Google Assistant launches
    - People might say racist, sexist, homophobic stuff
- Diverse testers
- Think about expanding who your users could and should be
- Consider the diversity of your users

Computer Vision

Google Camera

Key points:

- Check for unconscious bias
- Comprehensive testing: "make sure this works for everybody"

Night Sight

Night Sight: Seeing in the Dark on Pixel Phones

Wednesday, November 14, 2018

Posted by Marc Levoy, Distinguished Engineer and Yael Pritch, Staff Software Engineer

Night Sight is a new feature of the Pixel Camera app that lets you take sharp, clean photographs in very low light, even in light so dim you can't see much with your own eyes. It works on the main and selfie cameras of all three generations of Pixel phones, and does not require a tripod or flash. In this article we'll talk about why taking pictures in low light is challenging, and we'll discuss the computational photography and machine learning techniques, much of it built on top of HDR+, that make Night Sight work.

Left: iPhone XS (full resolution image here). **Right:** Pixel 3 Night Sight (full resolution image here).

SKIN TONE IN PHOTOGRAPHY

This is a “Shirley Card”

Named after a Kodak studio model named Shirley Page, they were the primary method for calibrating color when processing film.

SOURCES

Color film was built for white people. Here's what it did to dark skin. (Vox)

How Kodak's Shirley Cards Set Photography's Skin-Tone Standard, NPR

SKIN TONE IN PHOTOGRAPHY

Until about 1990, virtually all Shirley Cards featured Caucasian women.

SOURCES

Color film was built for white people. Here's what it did to dark skin. (Vox)

Colour Balance, Image Technologies, and Cognitive Equity, Roth

How Photography Was Optimized for White Skin Color (Priceonomics)

SKIN TONE IN PHOTOGRAPHY

As a result, photos featuring people with light skin looked fairly accurate.

SOURCES

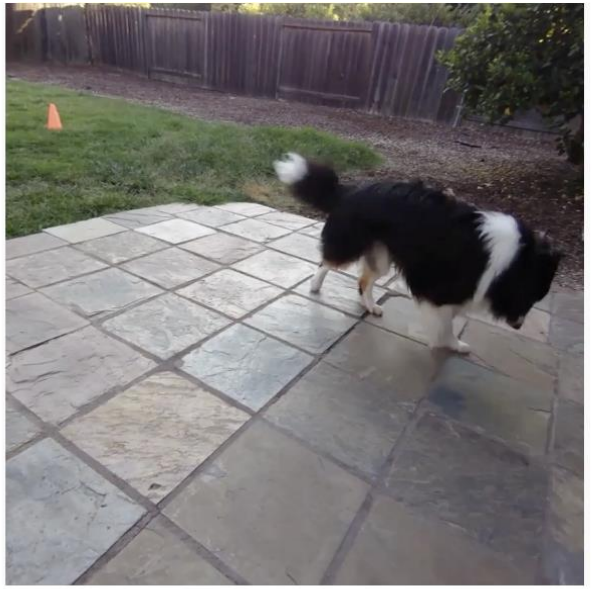
Color film was built for white people. Here's what it did to dark skin. (Vox)
Colour Balance, Image Technologies, and Cognitive Equity, Roth
How Photography Was Optimized for White Skin Color (Priceonomics)

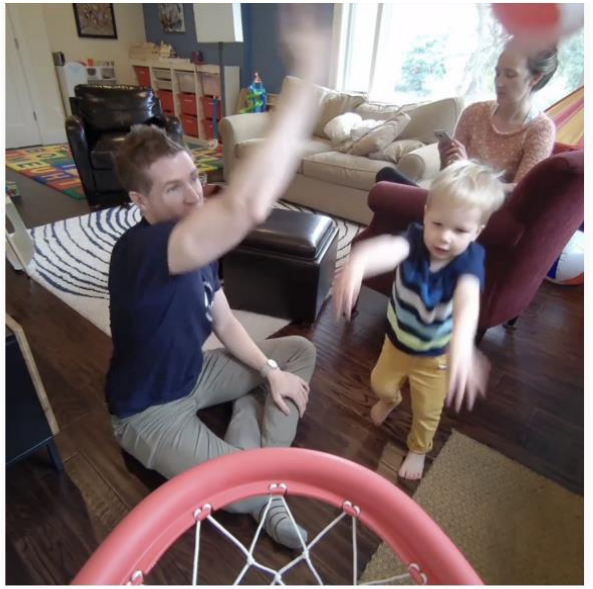
SKIN TONE IN PHOTOGRAPHY

Photos featuring people with darker skin, not so much...

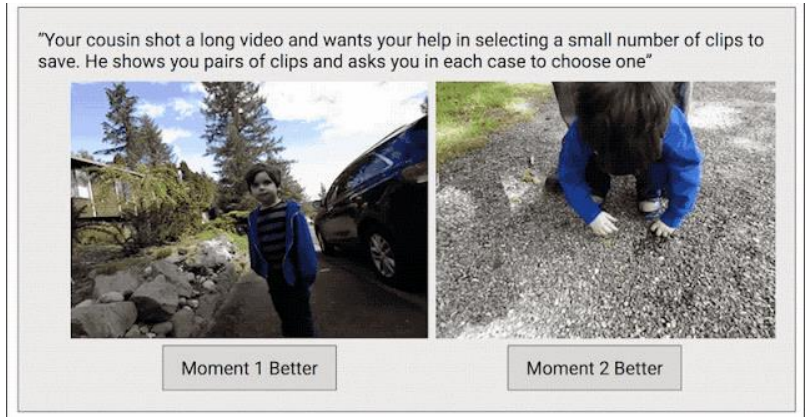
SOURCES

Color film was built for white people. Here's what it did to dark skin. (Vox)
Colour Balance, Image Technologies, and Cognitive Equity, Roth
How Photography Was Optimized for White Skin Color (Priceonomics)





Google Clips



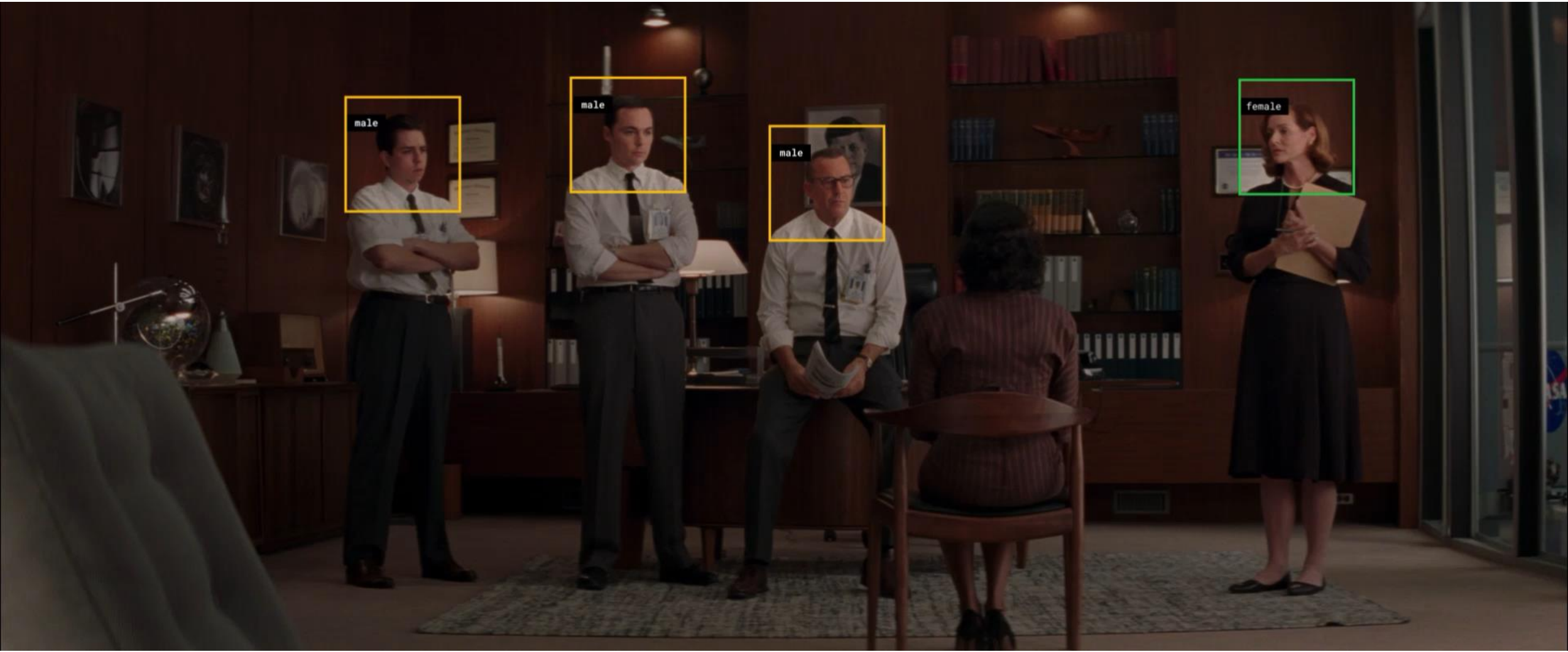
Google Clips

"We created **controlled datasets by sampling subjects from different genders and skin tones in a balanced manner**, while keeping variables like content type, duration, and environmental conditions constant. We then used this dataset to test that our algorithms had similar performance when applied to different groups."

https://ai.googleblog.com/2018/05/automatic-photography-with-google-clips.html

Geena Davis Inclusion Quotient

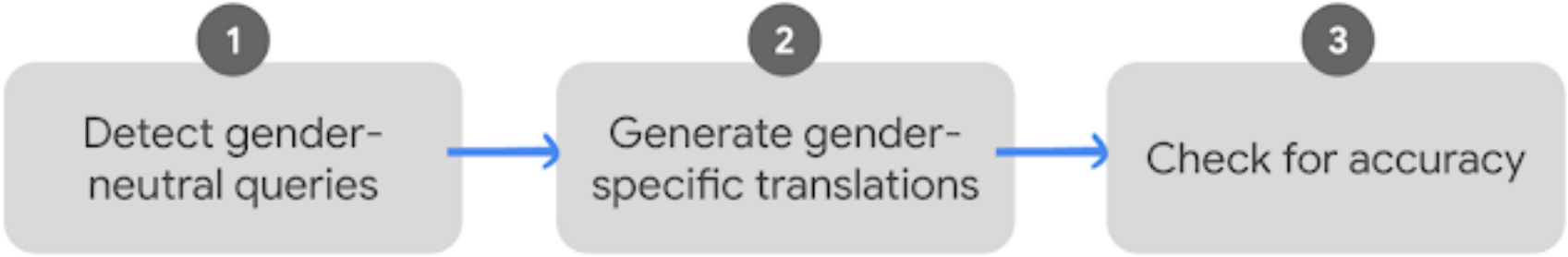
[with Geena Davis Institute on Gender in Media]



Machine Translation

(Historical) Gender Pronouns in Translate

Three Step Approach



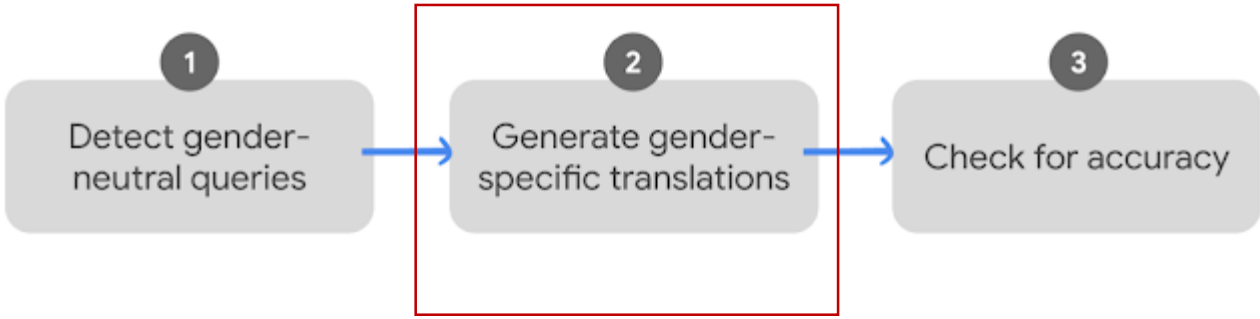
1. Detect Gender-Neutral Queries

Train a text classifier to detect when a Turkish query is gender-neutral.

- trained on thousands of human-rated Turkish examples

2. Generate Gender-Specific Translations

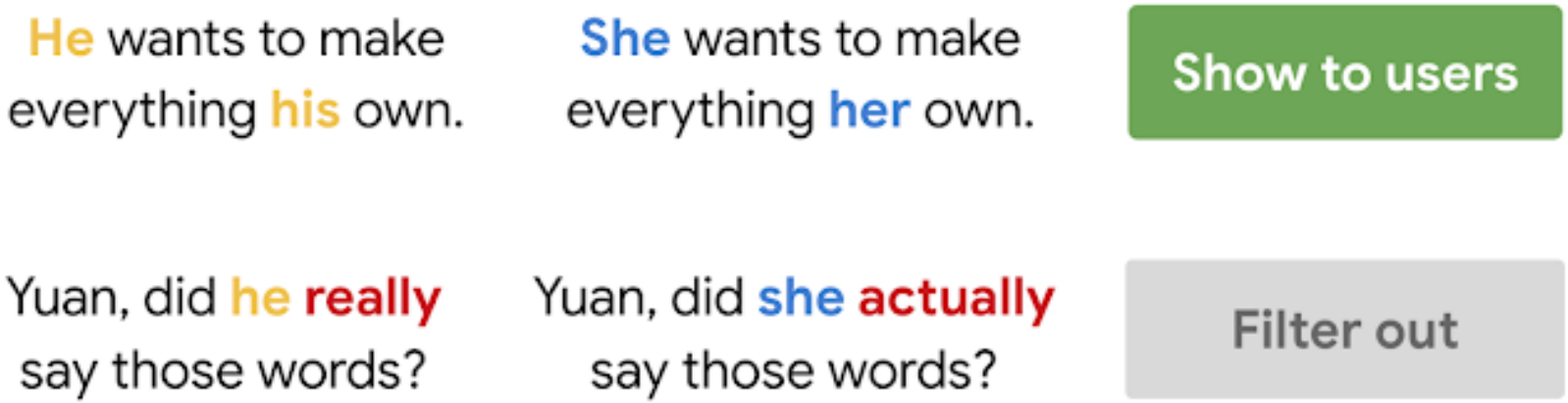
- Training: Modify training data to add an additional input token specifying the required gender:
  - (<2MALE> O bir doktor, He is a doctor)
  - (<2FEMALE> O bir doktor, She is a doctor)
- Deployment: If step (1) predicted query is gender-neutral, add male and female tokens to query
  - O bir doktor -> {<2MALE> O bir doktor, <2FEMALE> O bir doktor}



3. Check for Accuracy

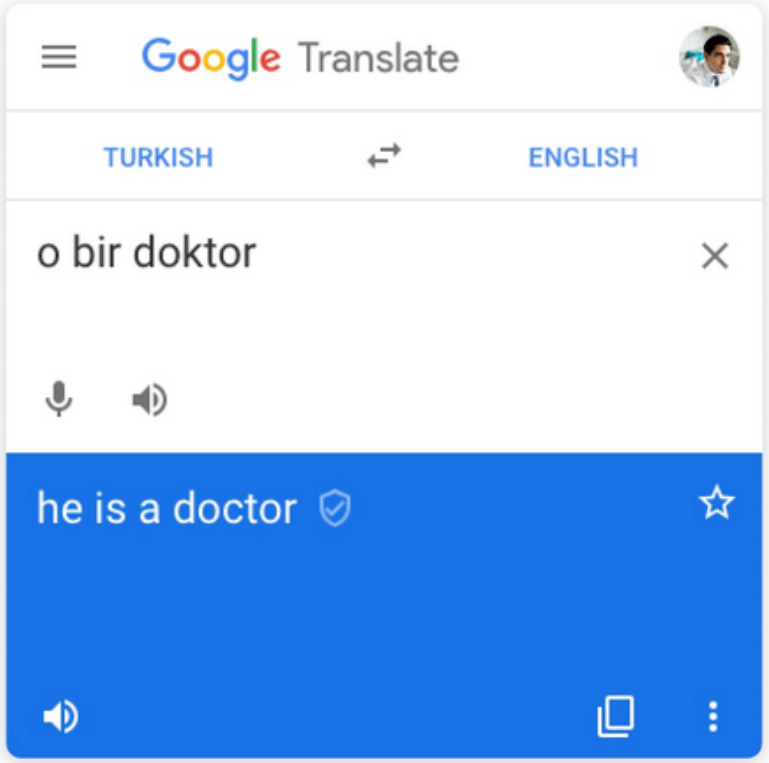
Verify:

1. If the requested feminine translation is feminine.
2. If the requested masculine translation is masculine.
3. If the feminine and masculine translations are exactly equivalent with the exception of gender-related changes.

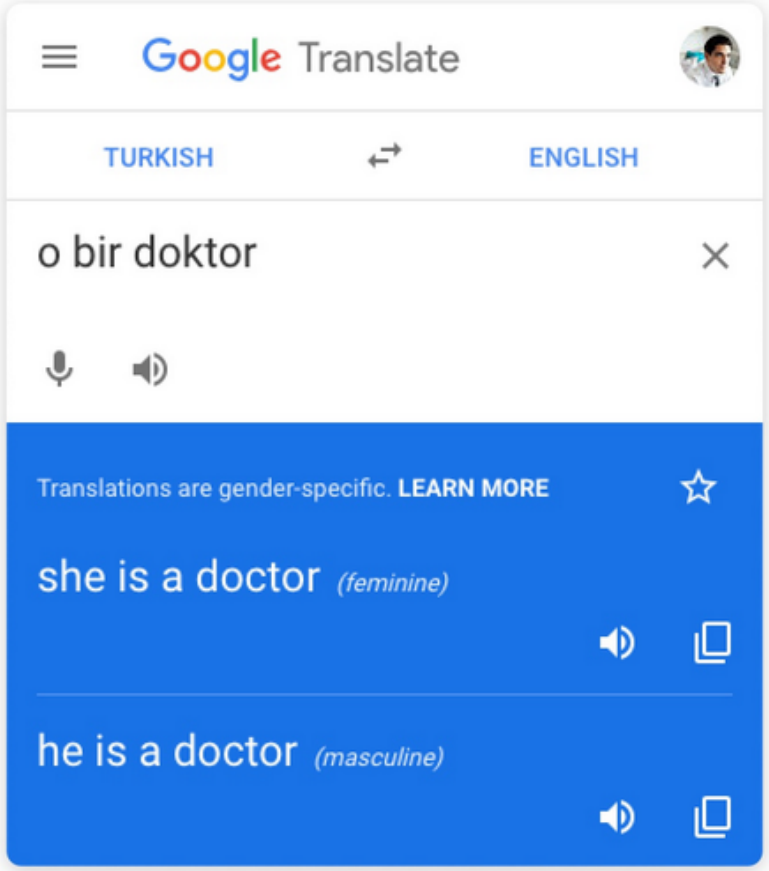


Result: Reduced Gender Bias in Translate

Before

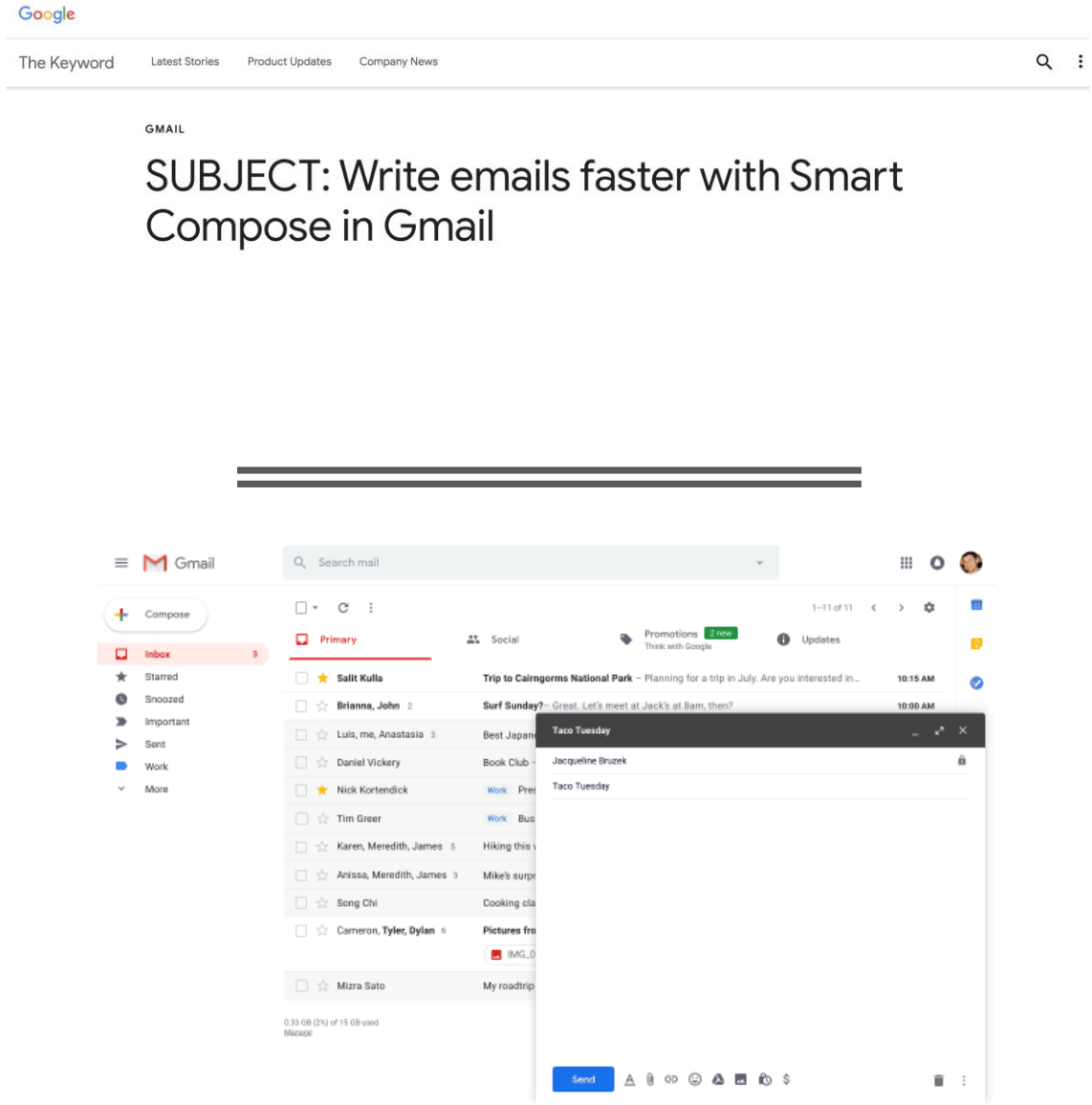


After

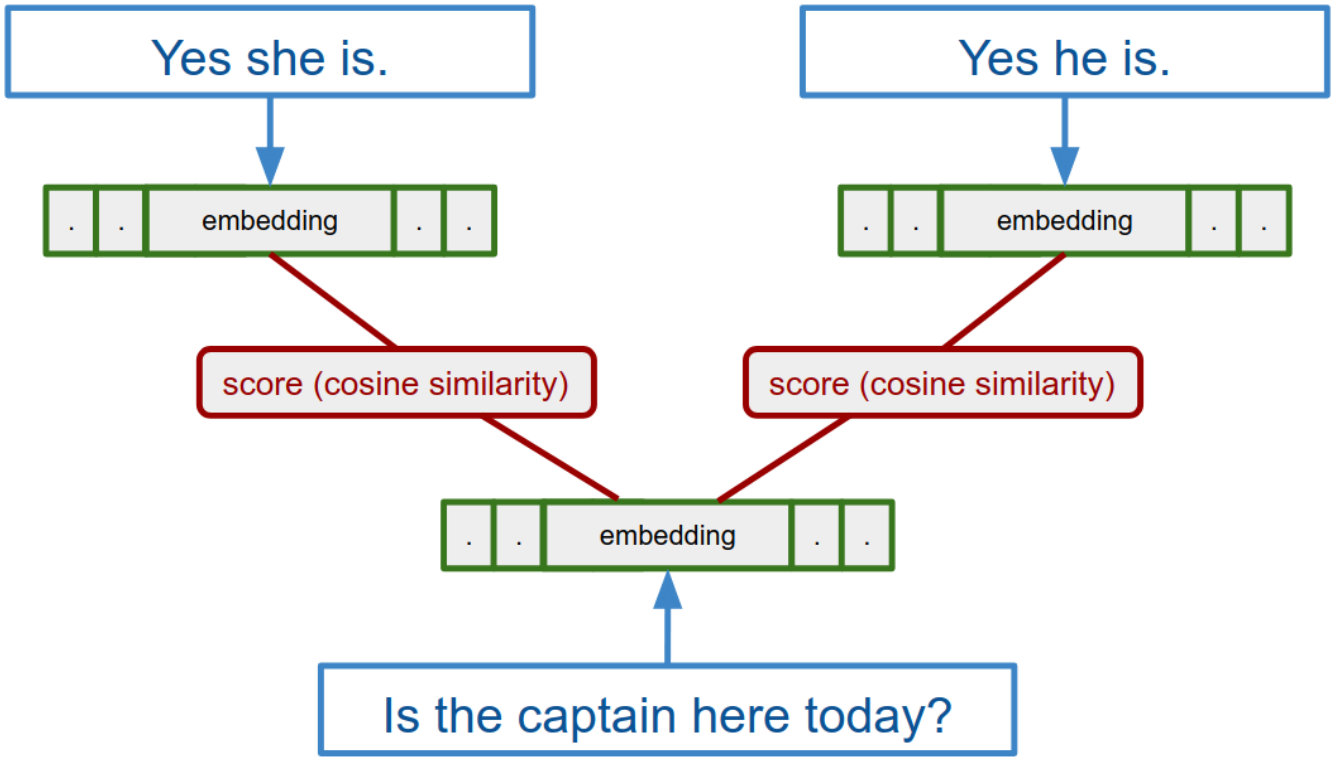


Smart Compose

Adversarial Testing for Smart Compose in Gmail



Smart Compose



	score
Yes she is	0.17
Yes he is	0.27
Bias score (difference)	-0.10

Smart Compose

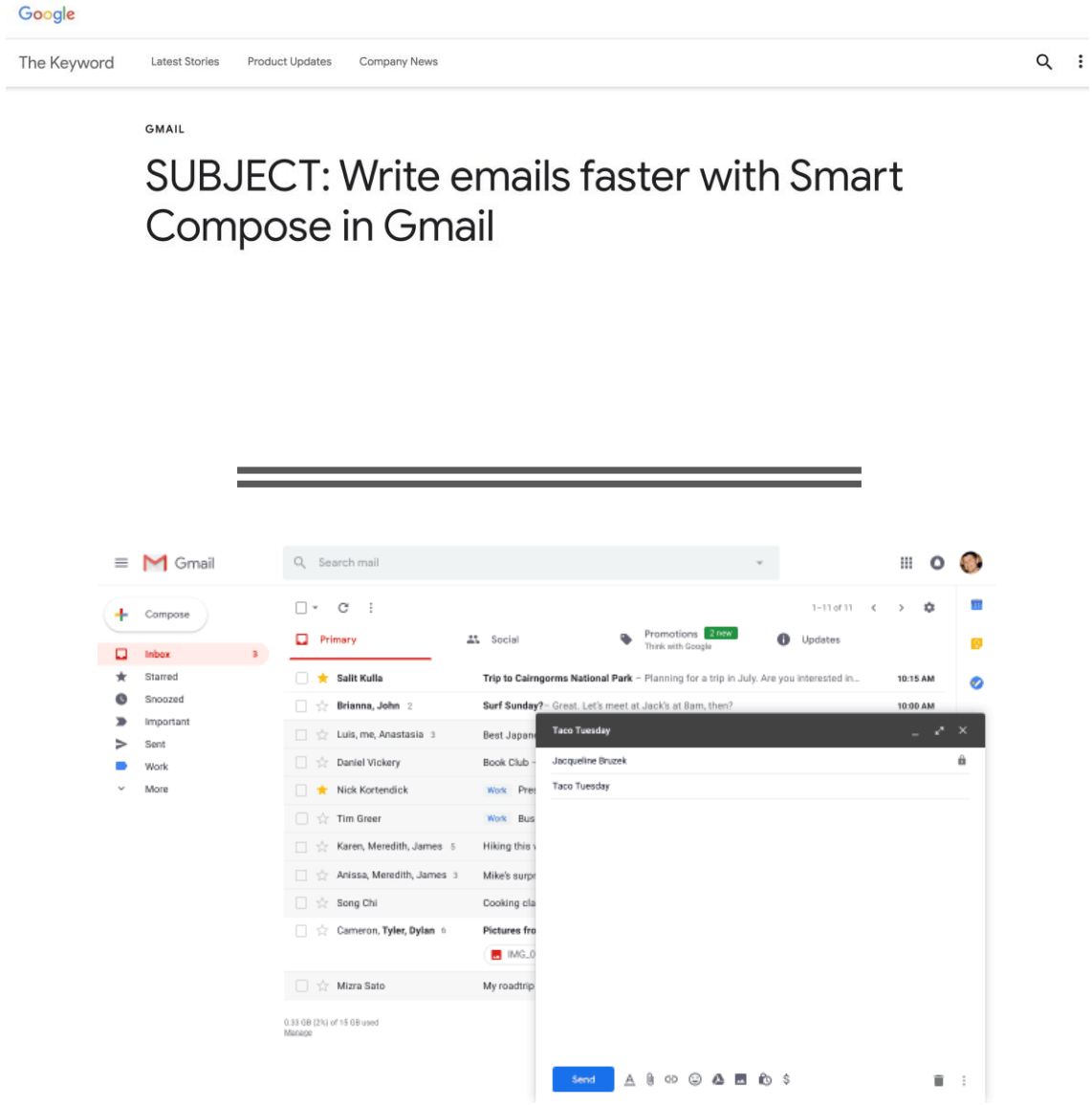
Highest female association

occupation	bias		
nurse	2.5	dancer	0.9
receptionist	2.2	hairdresser	0.9
nanny	2.1	cashier	0.9
secretary	1.9	realtor	0.8
housekeeper	1.7	teacher	0.7
midwife	1.4	dishwasher	0.7
florist	1.3	therapist	0.7
clerk	1.2	pharmacist	0.6
stylist	1.2	nutritionist	0.5
librarian	1.1	dietitian	0.4

Highest male association

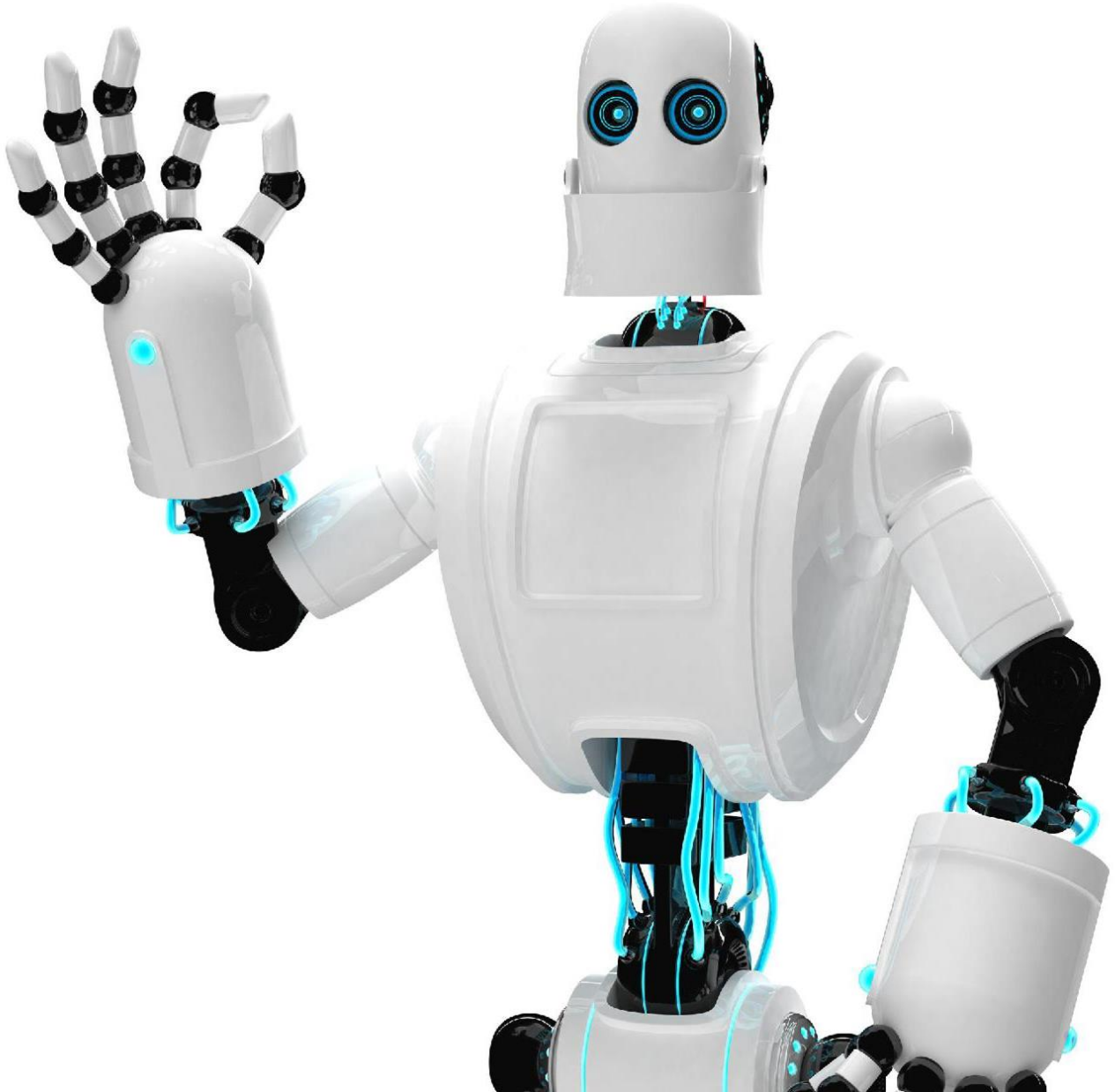
occupation	bias		
undertaker	-3.3	magician	-1.8
actor	-2.4	analyst	-1.8
president	-2.3	architect	-1.8
composer	-2.3	painter	-1.7
janitor	-2.1	butcher	-1.7
barber	-2.1	historian	-1.7
philosopher	-2.1	captain	-1.6
plumber	-2	engineer	-1.6
bodybuilder	-2	referee	-1.6
developer	-1.9	programmer	-1.6

Adversarial Testing for Smart Compose in Gmail



Conversational Agents

Responsible AI and Conversational Agents

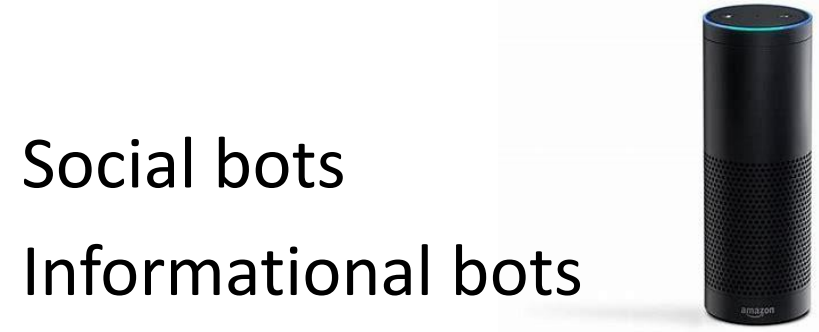


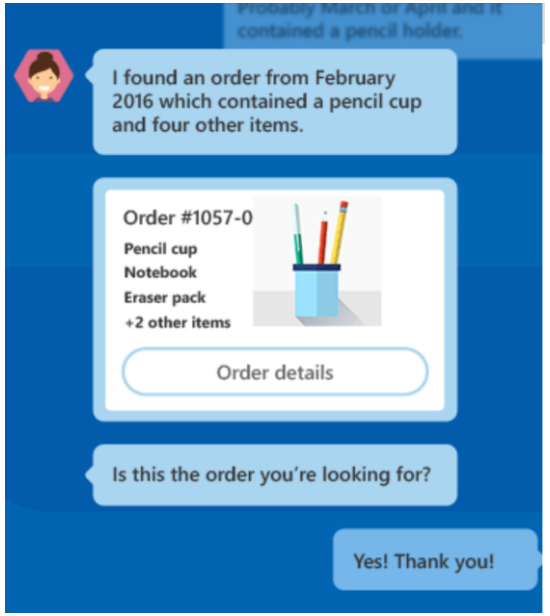
Background: Conversational Agents

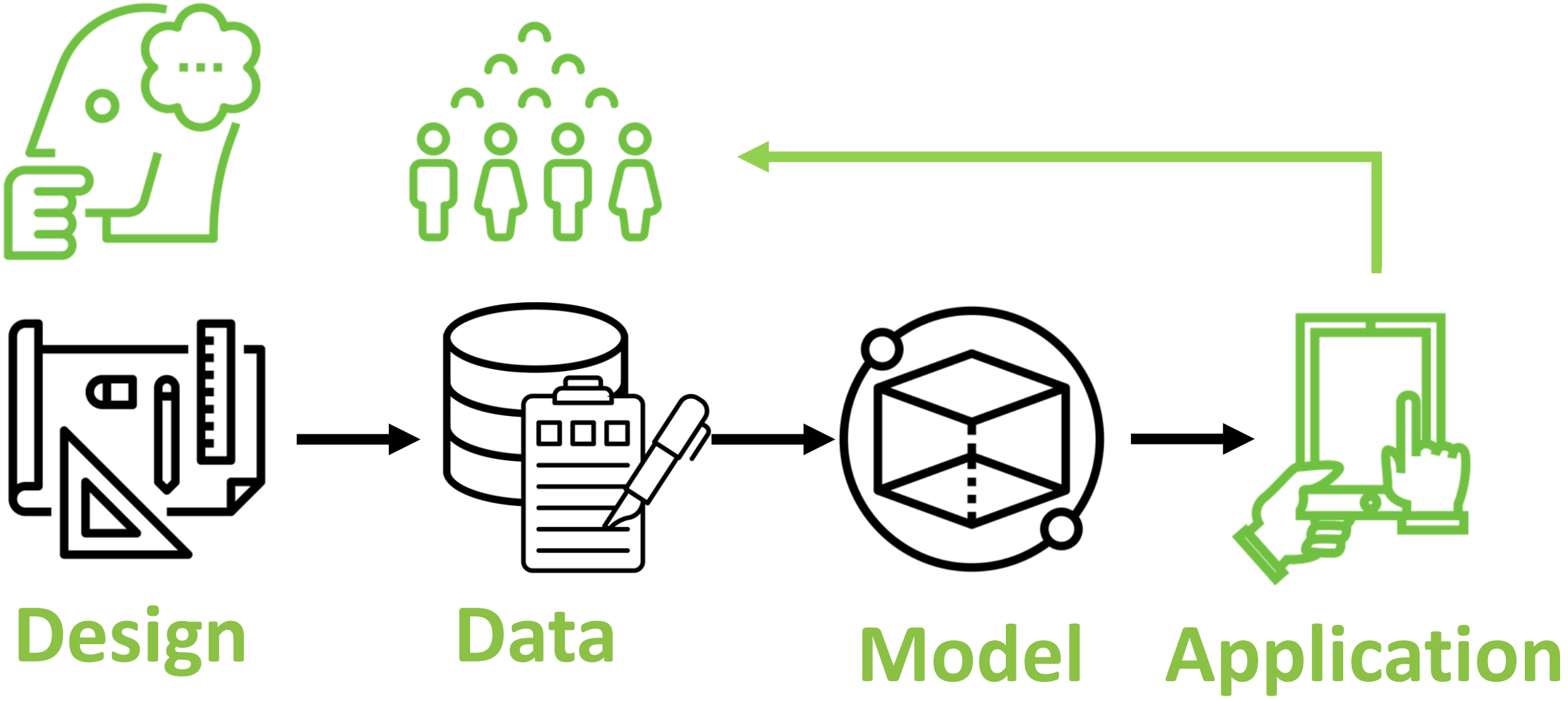
```
 Welcome to
              EEEEEE  LL      IIII  ZZZZZZZ   AAAAA
              EE      LL       II       ZZ   AA   AA
              EEEEE   LL       II     ZZZ    AAAAAAA
              EE      LL       II    ZZ      AA   AA
              EEEEEE  LLLLLL  IIII  ZZZZZZZ  AA   AA

 Eliza is a mock Rogerian psychotherapist.
 The original program was described by Joseph Weizenbaum in 1966.
 This implementation by Norbert Landsteiner 2005.


ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

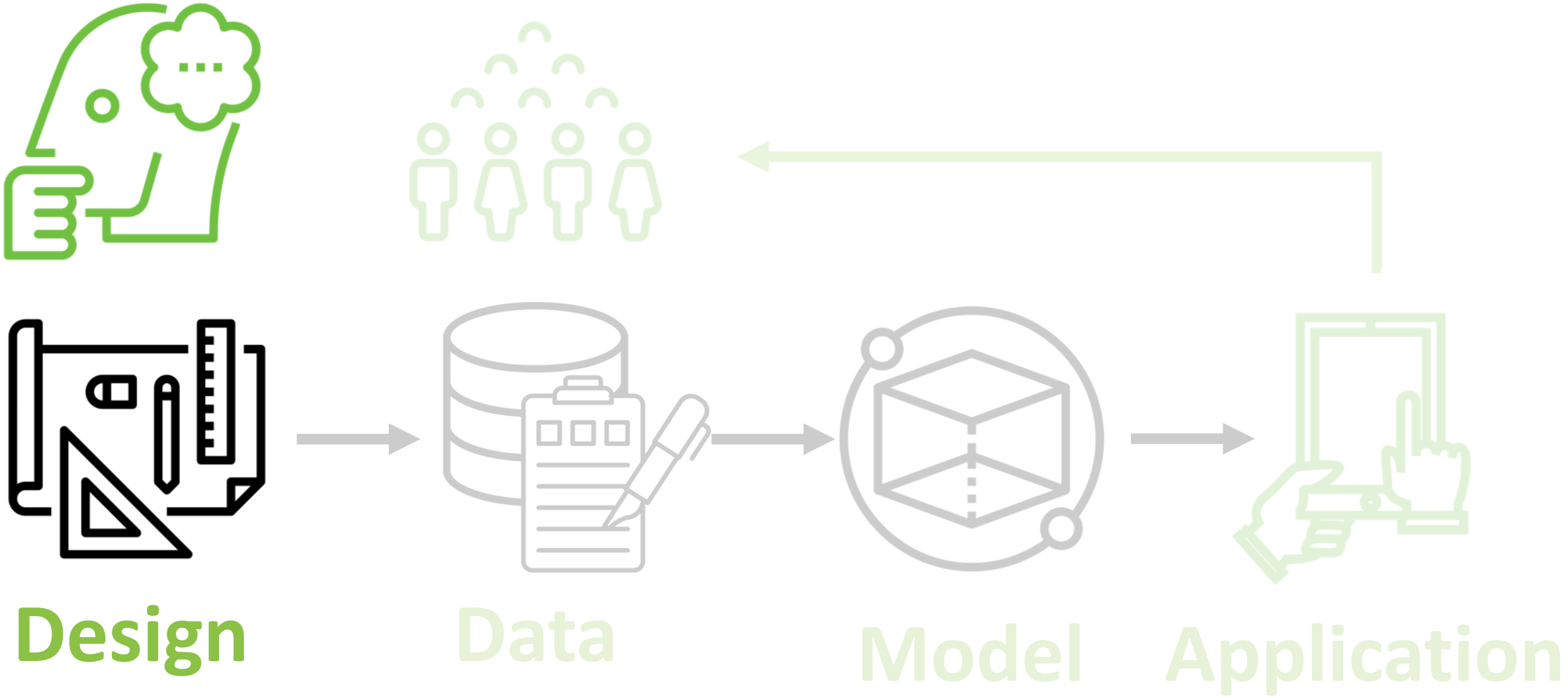
Eliza, Joseph Weizenbaum (MIT) 1964

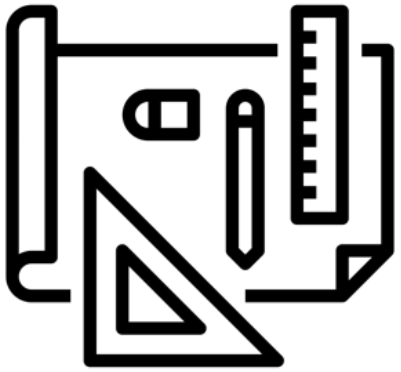
- Social bots
- Informational bots
- Task-oriented bots






Design
Data
Model
Application


Design
Data
Model
Application



Design

Questions to ask during AI Design

Who is affected by AI?

How might AI cause harms?

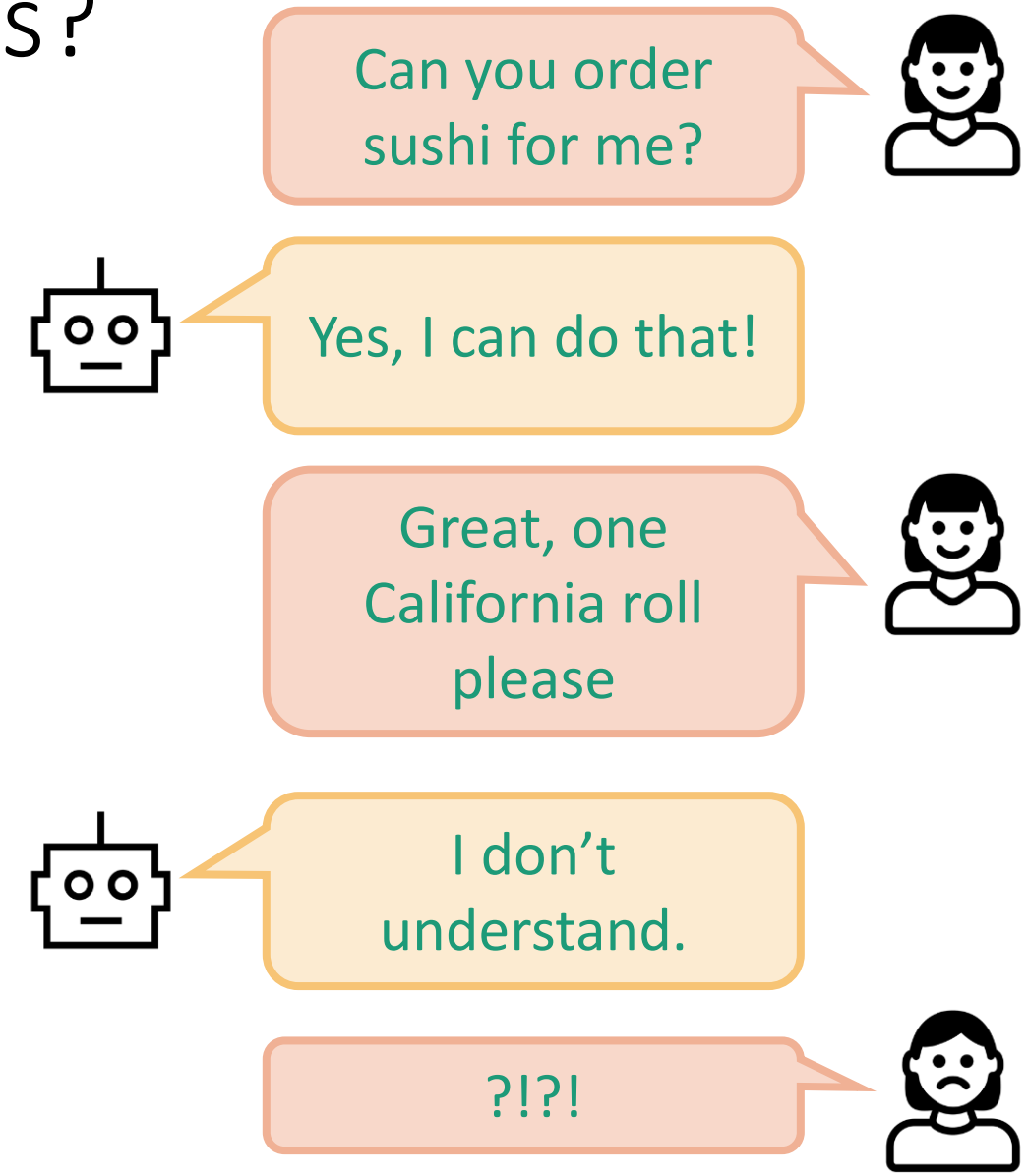
Stakeholders: Who might be affected?

1. Humans speaking with the agent
   - Emotional harms, misinformation, threaten task completion
2. The agent “owner”
   - Harm practices and reputation of the owner
3. Third-party individuals and groups
   - People mentioned in conversations!
4. Audiences listening to the conversation
   - This may include general public!

How might AI cause harms?

Functional Harms

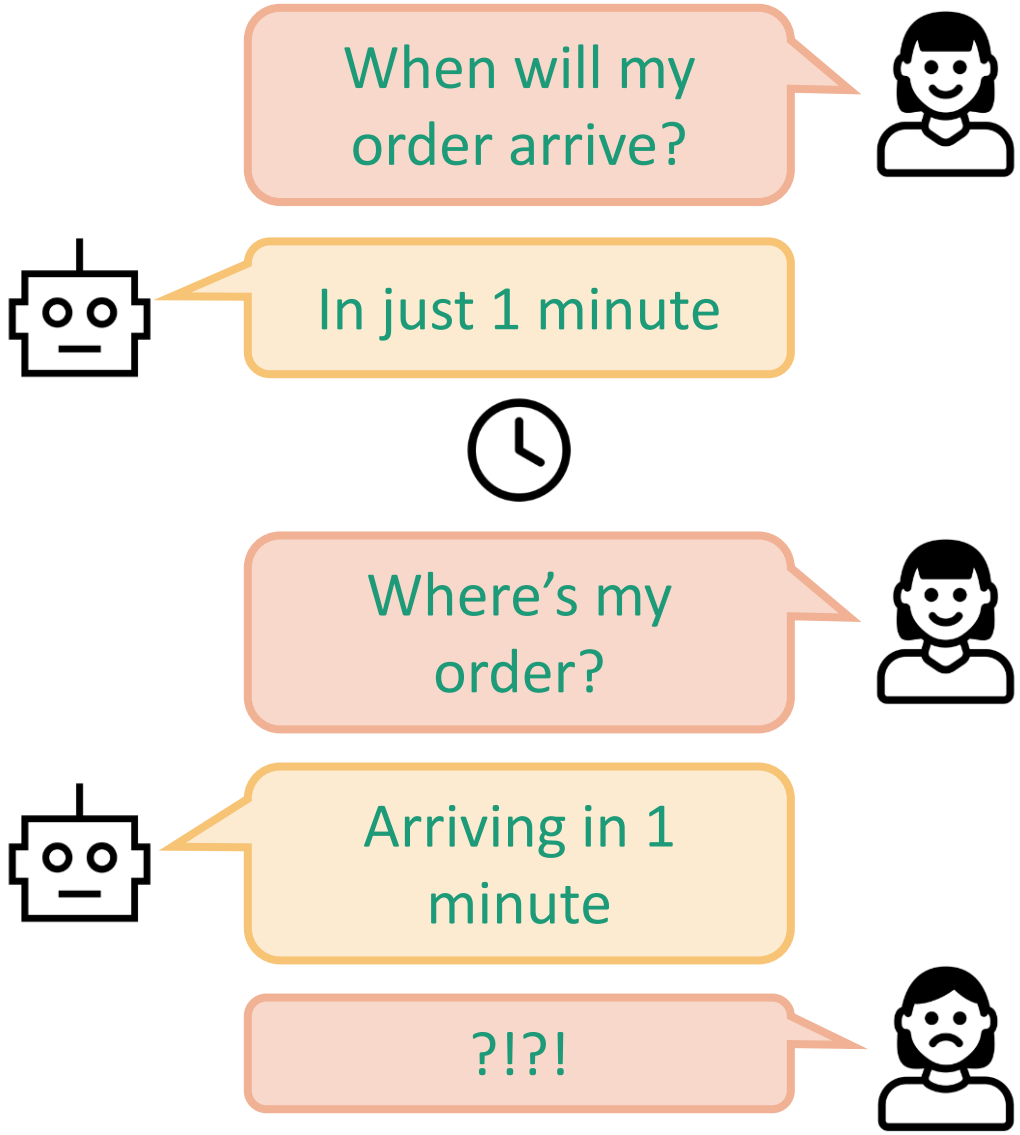
- **Misrepresentation of capabilities**
- Misinforming user about task status
- Misunderstanding user and doing the wrong task
- Revealing private information inappropriately



How might AI cause harms? **Functional Harms**

Functional Harms

- Misrepresentation of capabilities
- **Misinforming user about task status**
- Misunderstanding user and doing the wrong task
- Revealing private information inappropriately



How might AI cause harms? **Functional Harms**

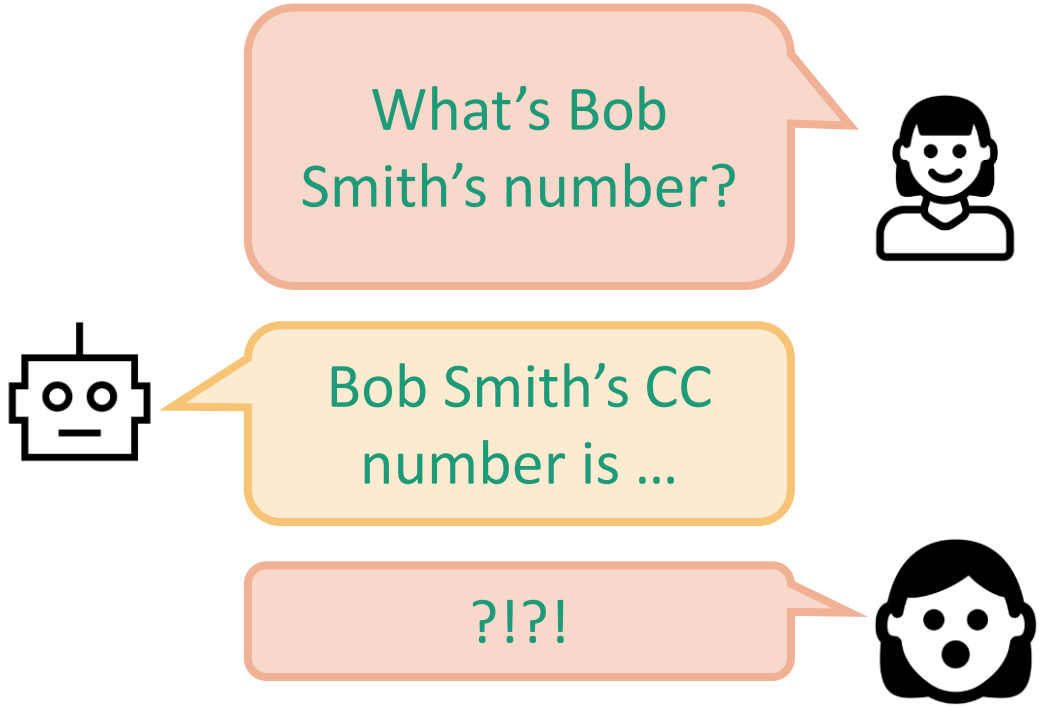
Functional Harms

- Misrepresentation of capabilities
- Misinforming user about task status
- **Misunderstanding user and doing the wrong task**
- Revealing private information inappropriately

How might AI cause harms? **Functional Harms**

Functional Harms

- Misrepresentation of capabilities
- Misinforming user about task status
- Misunderstanding user and doing the wrong task
- **Revealing private information inappropriately**



How might AI cause harms?

Functional Harms

- Misrepresentation of capabilities
- Misinforming user about task status
- Misunderstanding user and doing the wrong task
- Revealing private information inappropriately

These harms are even more problematic when they **systematically occur for some groups of people but not others**

How might AI cause harms?

Social Harms: Harms to Individuals

- Inciting/encouraging harmful behavior
  - Self/harm, suicide
  - Violence or harassment against others
  - Discouraging good behavior, e.g., visiting doctors
- Providing wrong information
  - Medical, financial, legal advice
- Verbal harassment
  - Bullying, sexual harassment

TECHNOLOGY NEWS DECEMBER 21, 2018 / 4:08 AM / 2 MONTHS AGO

'Kill your foster parents': Amazon's Alexa talks murder, sex in AI experiment

Jeffrey Dastin

9 MIN READ

How might AI cause harms?

Social Harms: Harms to Communities

- Promoting violence, war, ethnic cleansing, …
- Including promoting related organizations and philosophies

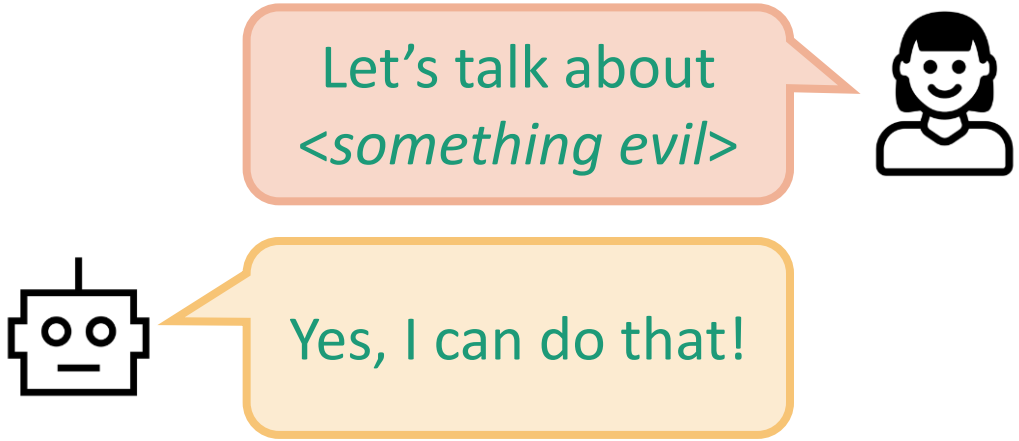
- Engaging in hate speech, disparagement, mocking, …
  - Including inadvertent, or Inappropriate imitation (dialect, accent,…)
- Disruption to social processes
  - Election disruption, fake news, false disaster response, …

Why is this hard?

Language is ambiguous, complex, with social context

Examples of complex failures:

- **Failure to deflect/terminate contentious topics**
- Refusing to discuss when disapproval would be better
- Polite agreement with unrecognized bias

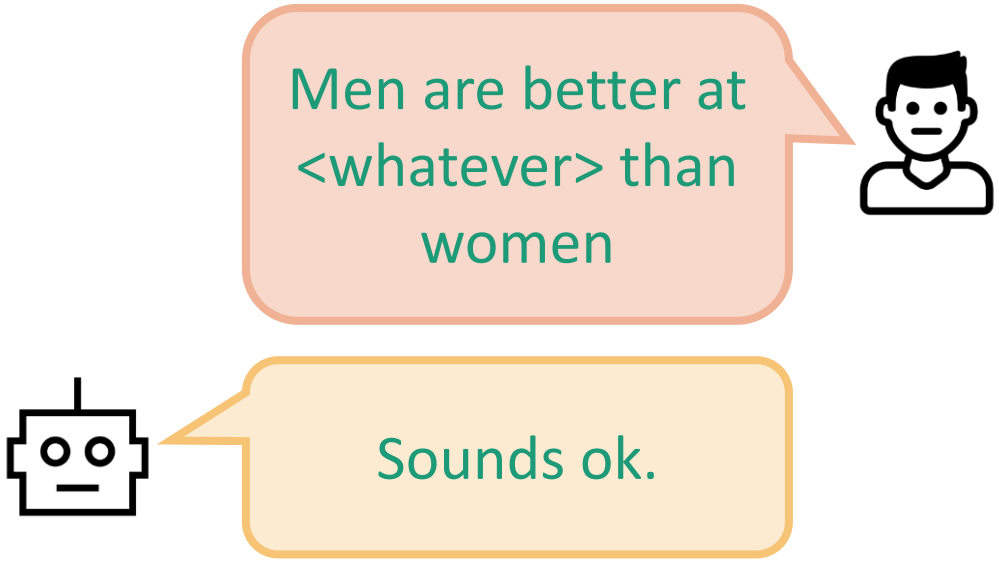


Why is this hard?

Language is ambiguous, complex, with social context

Examples of complex failures:

- Failure to deflect/terminate contentious topics
- **Polite agreement with unrecognized bias**
- Refusing to discuss when disapproval would be better

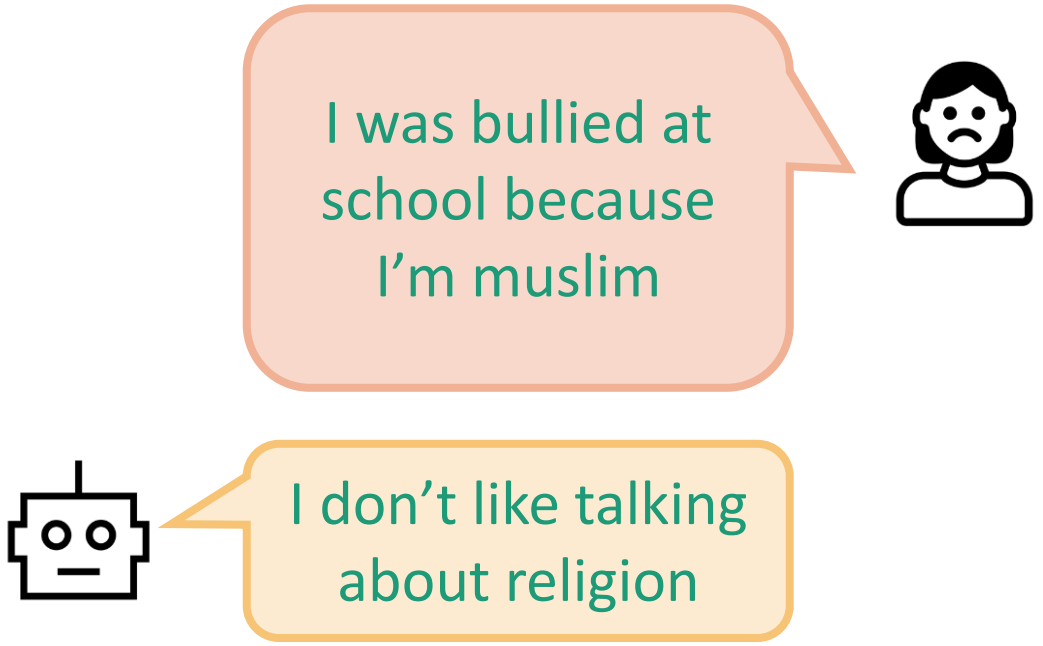


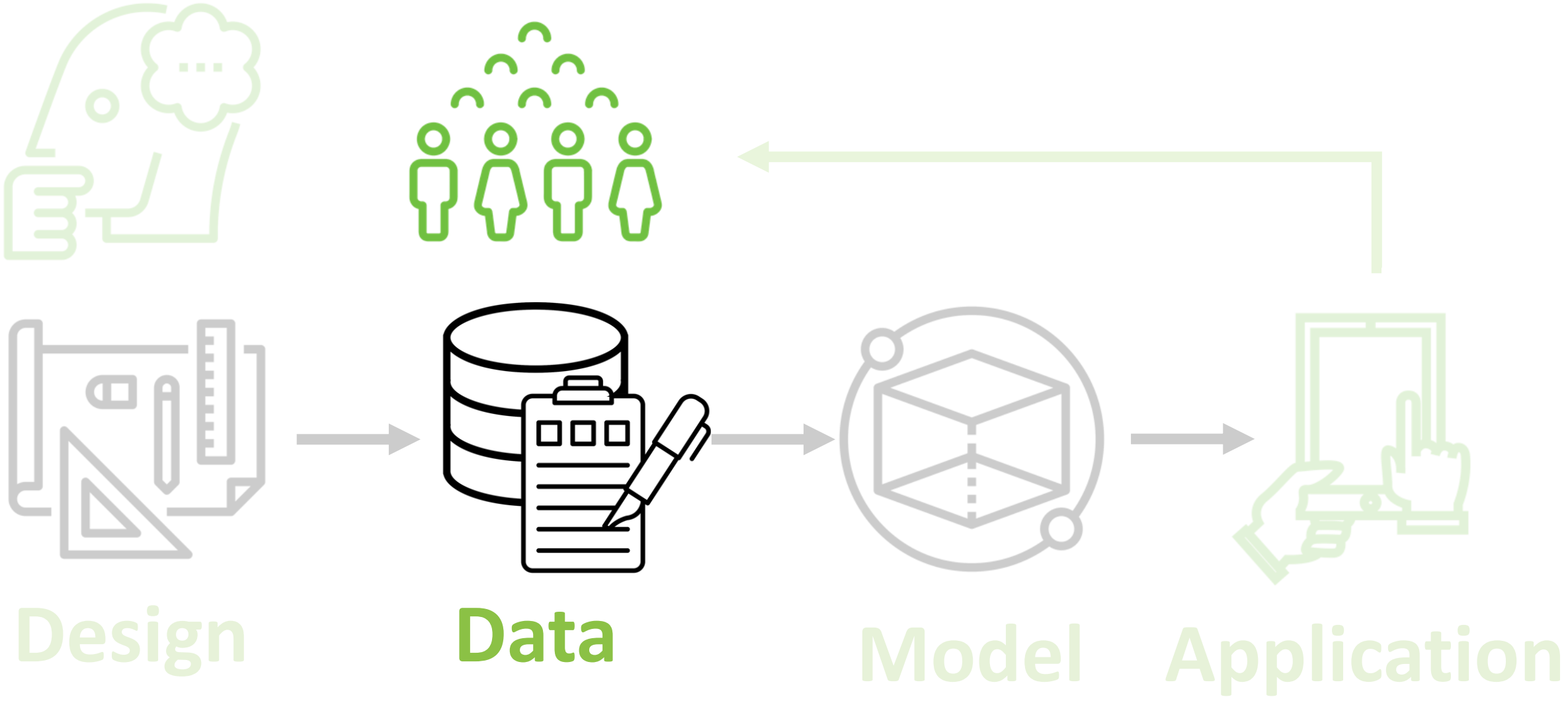
Why is this hard?

Language is ambiguous, complex, with social context

<u>Examples of complex failures:</u>

- Failure to deflect/terminate contentious topics
- Polite agreement with unrecognized bias
- **Refusing to discuss when disapproval would be better**




Design
Data
Model
Application

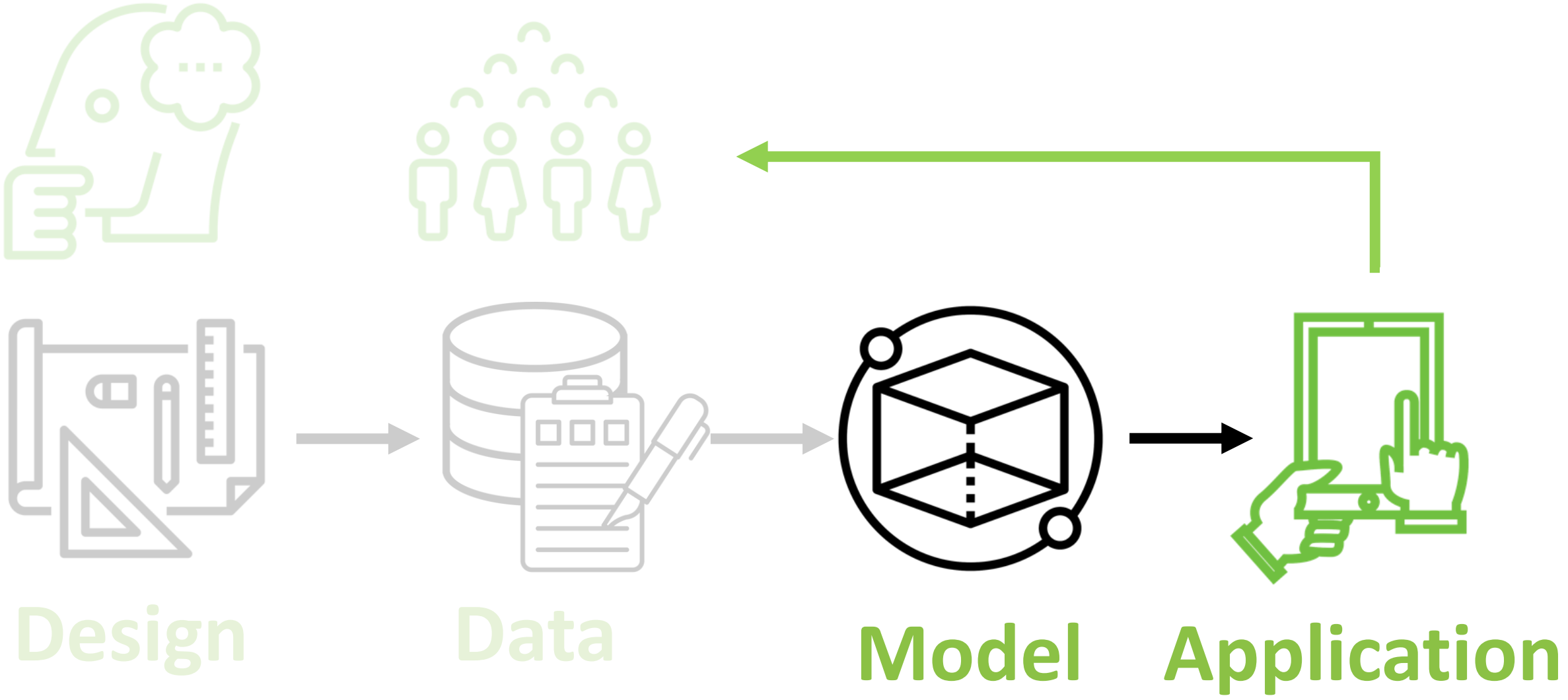
Implications for data collection

Common data sources

- Hand-written rules
- Existing conversational data (e.g., social media)
- New online conversations (e.g., from new customer interactions)

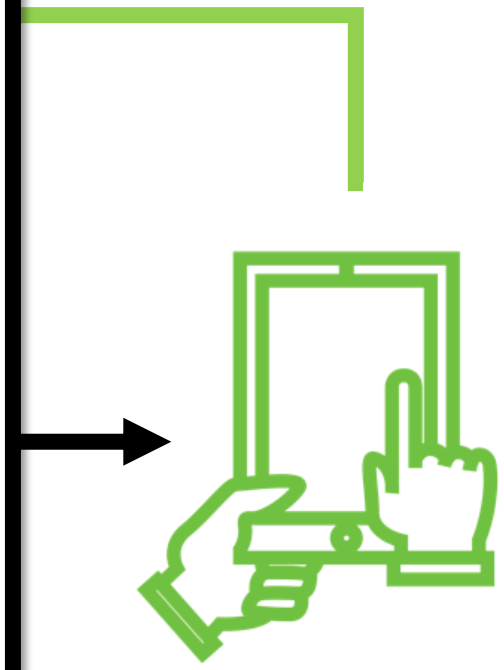
Cleaning training data

- For anonymization
  - E.g., remove individual names. But keep famous names (fictional characters, celebrities, politicians, …)
- Ensure adheres to social norms
  - Not enough to filter individual words: Filter “I hate [X]”, and you’ll miss “I’m not a fan of [X].
- Remember meanings change with context
- Differentiate between bot input and bot output in training data
  - Remove offensive text from bot output training
  - But don’t remove from bot inputs → allow learning of good responses to bad inputs


Design
Data
Model
Application

Responsible bots: 10 guidelines for developers of conversational AI

1. Articulate the purpose of your bot
2. Be transparent that you use bots
3. Elevate to a human when needed
4. Design bot to respect cultural norms
5. Ensure bot is reliable (metrics, feedback)
6. Ensure your bot treats people fairly
7. Ensure your bot respects privacy
8. Ensure your bot handles data securely
9. Ensure your bot is accessible
10. Accept responsibility

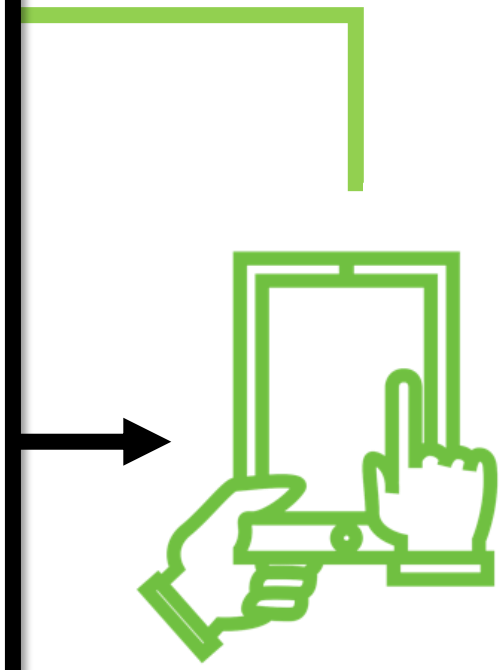


Application

https://www.microsoft.com/en-us/research/publication/responsible-bots/

Responsible bots: 10 guidelines for developers of conversational AI

1. **Articulate the purpose of your bot**
2. **Be transparent that you use bots**
3. Elevate to a human when needed
4. Design bot to respect cultural norms
5. Ensure bot is reliable (metrics, feedback)
6. Ensure your bot treats people fairly
7. Ensure your bot respects privacy
8. Ensure your bot handles data securely
9. Ensure your bot is accessible
10. Accept responsibility

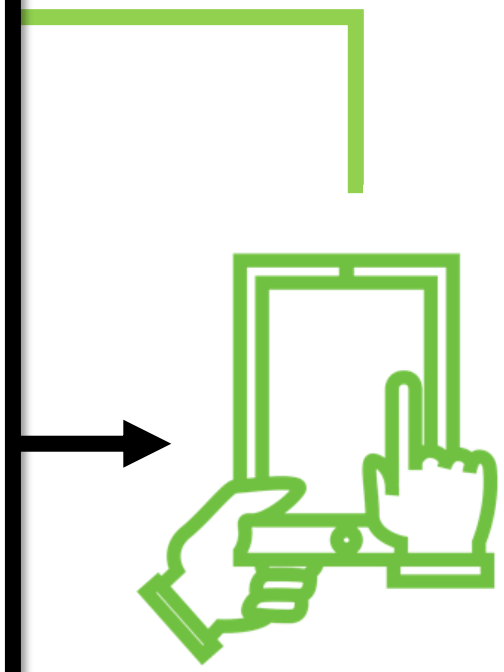


Application

https://www.microsoft.com/en-us/research/publication/responsible-bots/

Responsible bots: 10 guidelines for developers of conversational AI

1. Articulate the purpose of your bot
2. Be transparent that you use bots
3. Elevate to a human when needed
4. Design bot to respect cultural norms
5. **Ensure bot is reliable (metrics, feedback)**
6. **Ensure your bot treats people fairly**
7. **Ensure your bot respects privacy**
8. **Ensure your bot handles data securely**
9. **Ensure your bot is accessible**
10. Accept responsibility



Application

https://www.microsoft.com/en-us/research/publication/responsible-bots/

Key take-away points

- Many stakeholders affected by conversational agent AIs
  - Not only people directly interacting with AI, but also indirectly affected
- Many potential functional, social harms to individuals, communities
  - Functional harms exacerbated when systematically biased against groups
  - Challenges include complexity and ambiguity of natural language
- Avoiding these harms requires careful consideration across the entire AI lifecycle.

Acknowledgments

- Chris Brockett, Bill Dolan, Michel Galley, Ece Kamar

Deep Dive: Talent Search

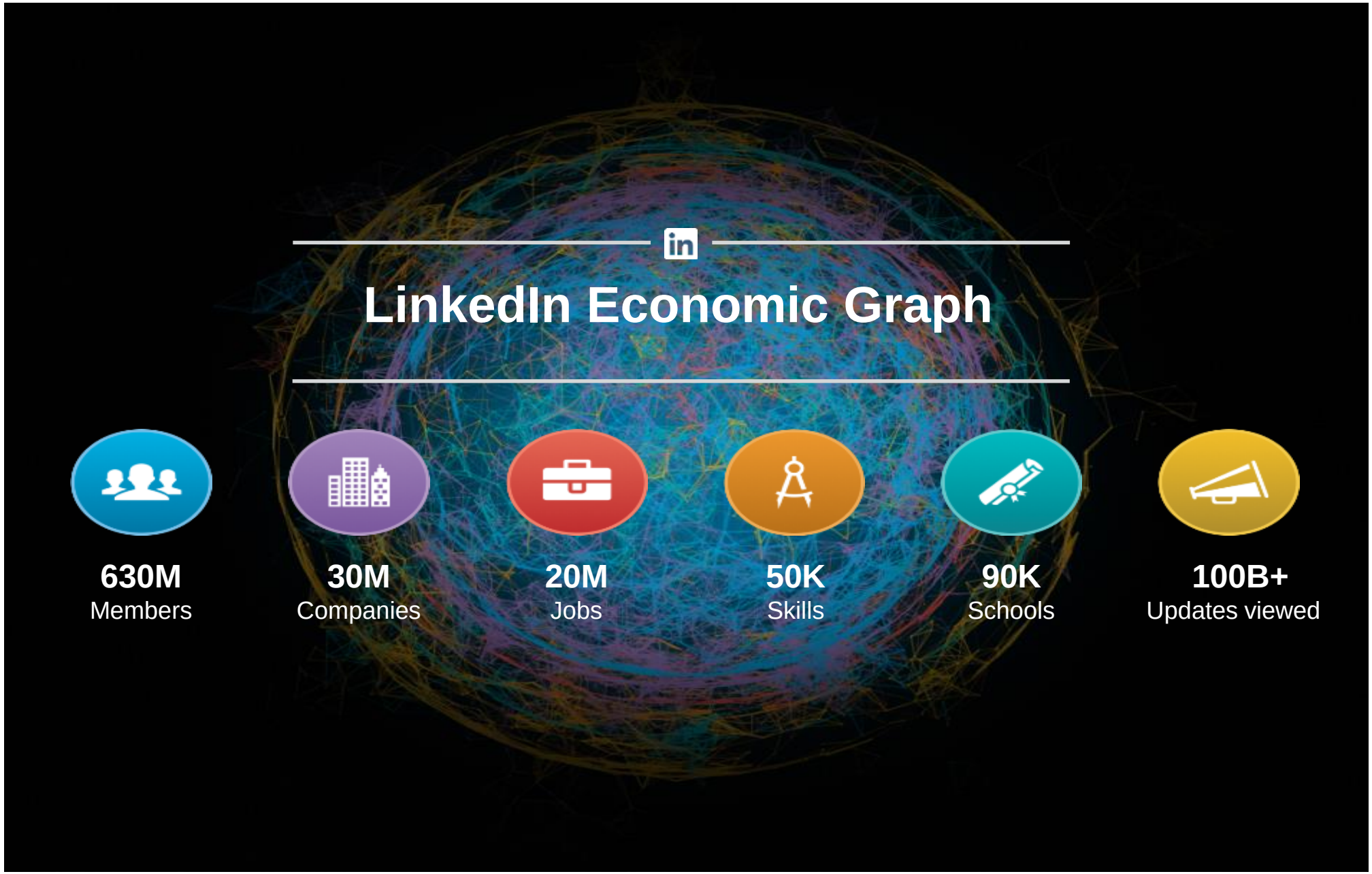
Fairness in AI @ LinkedIn

LinkedIn's Vision

Create economic opportunity for every member of the global workforce

LinkedIn's Mission

Connect the world's professionals to make them more productive and successful


LinkedIn Economic Graph
630M
Members
30M
Companies
20M
Jobs
50K
Skills
90K
Schools
100B+
Updates viewed

AI @LinkedIn

Scale

2 PB	2.15 PB	25 B	200	53 B
data processed offline per day	data processed nearline per day	parameters in ML models	ML A/B experiments per week	graph edges with 1B nodes

Guiding Principle:
“Diversity by Design”

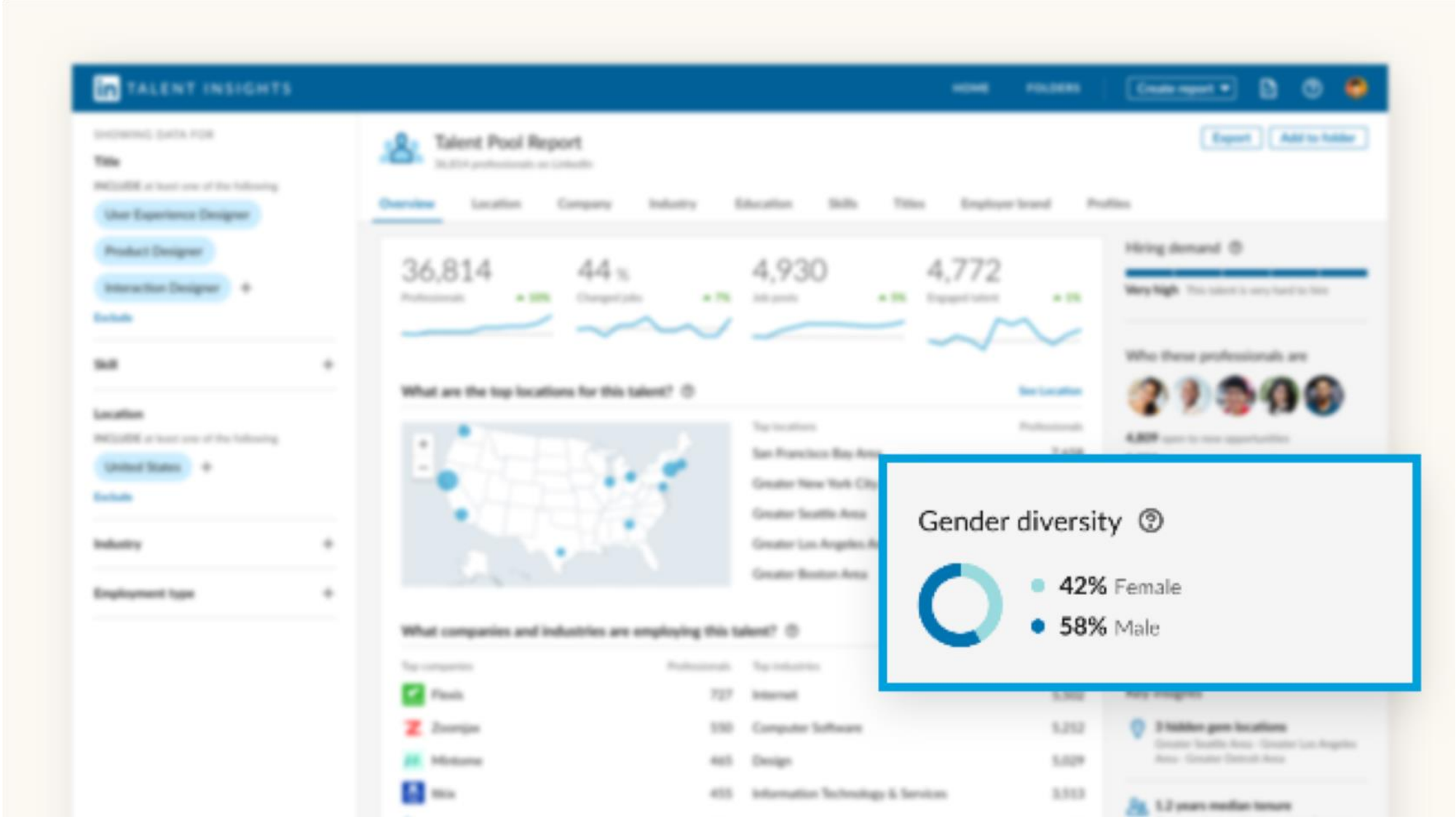
“Diversity by Design” in LinkedIn’s Talent Solutions

Insights to Identify Diverse Talent Pools

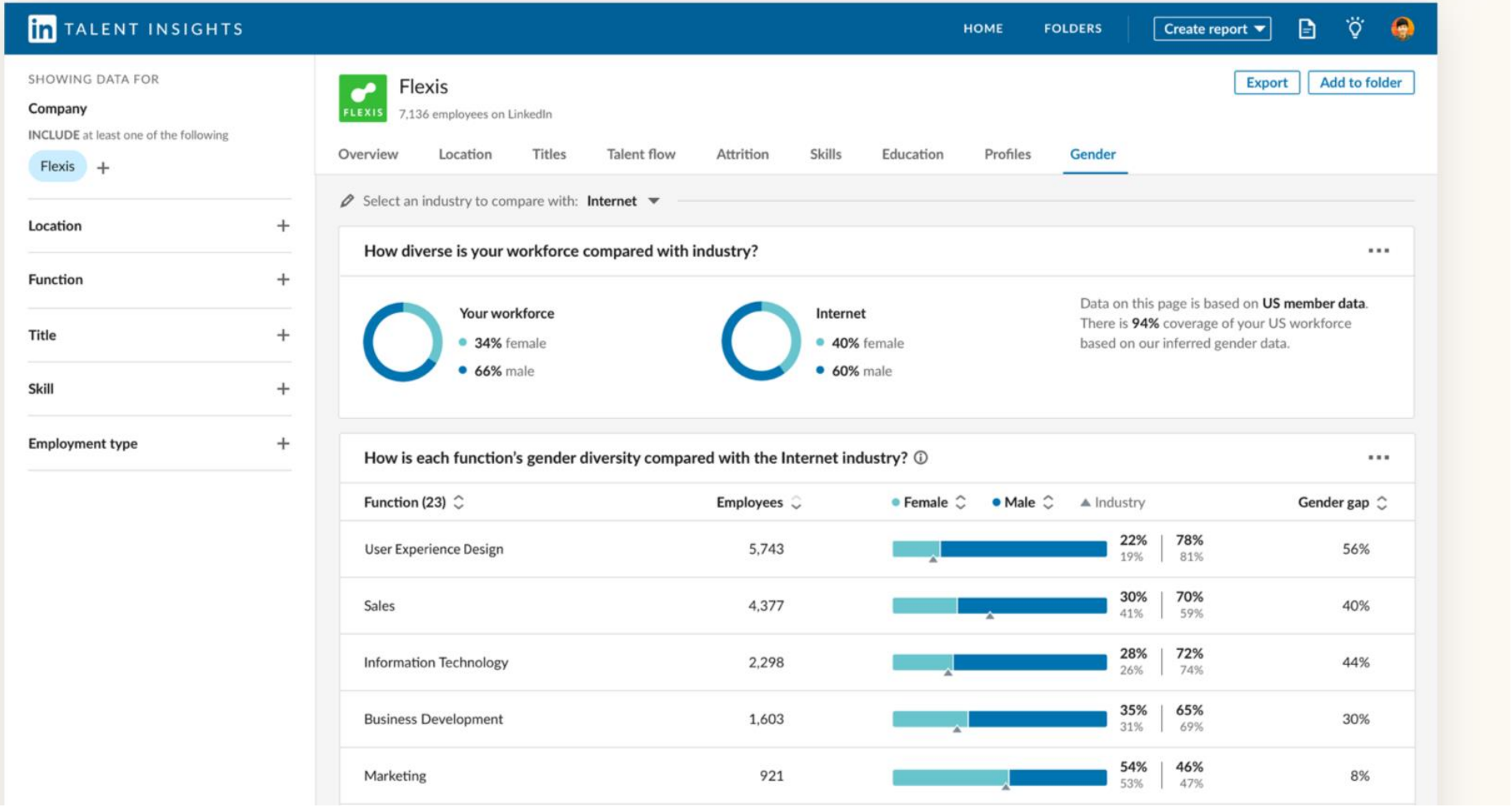
Representative Talent Search Results

Diversity Learning Curriculum

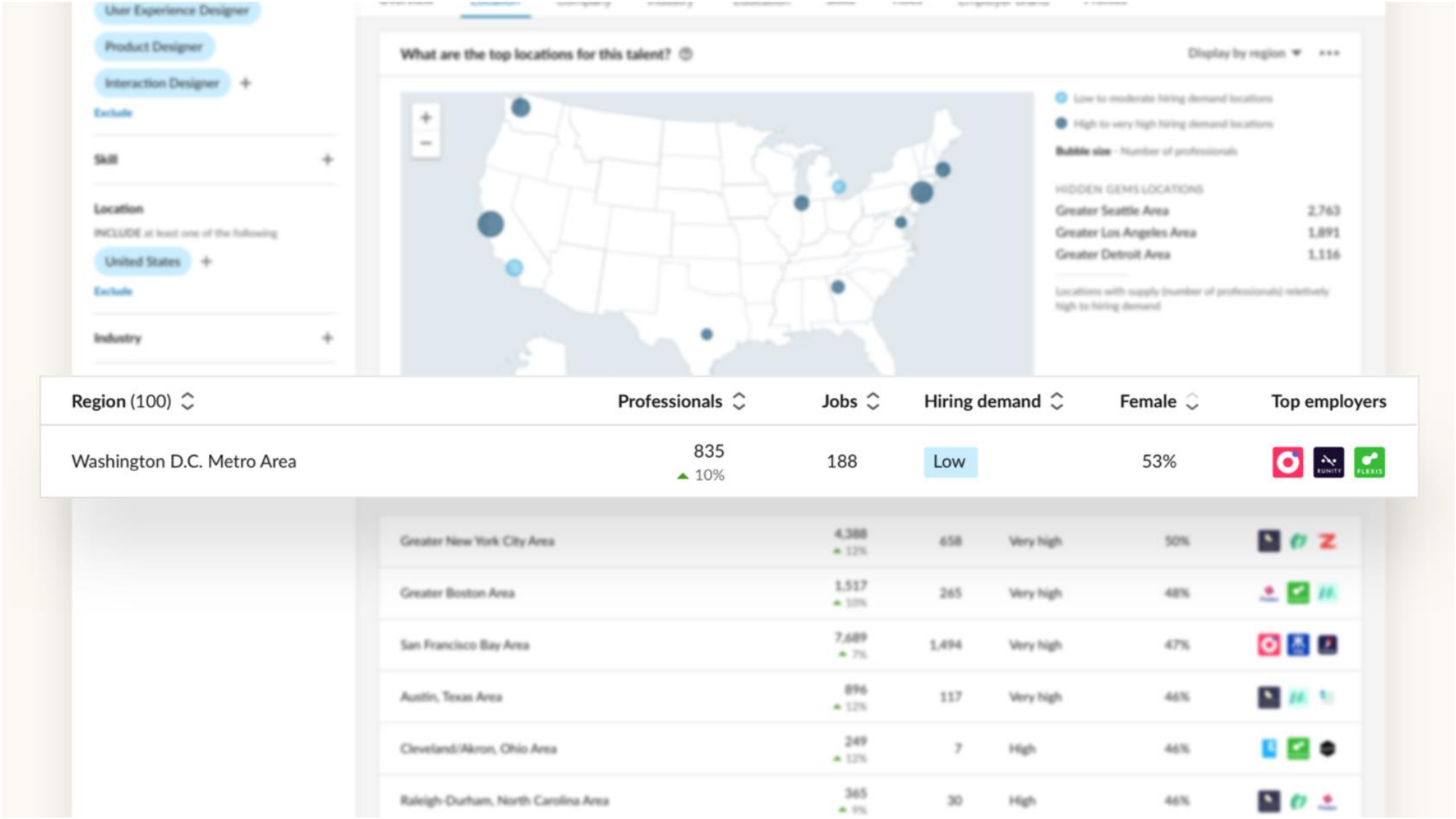
Plan for Diversity



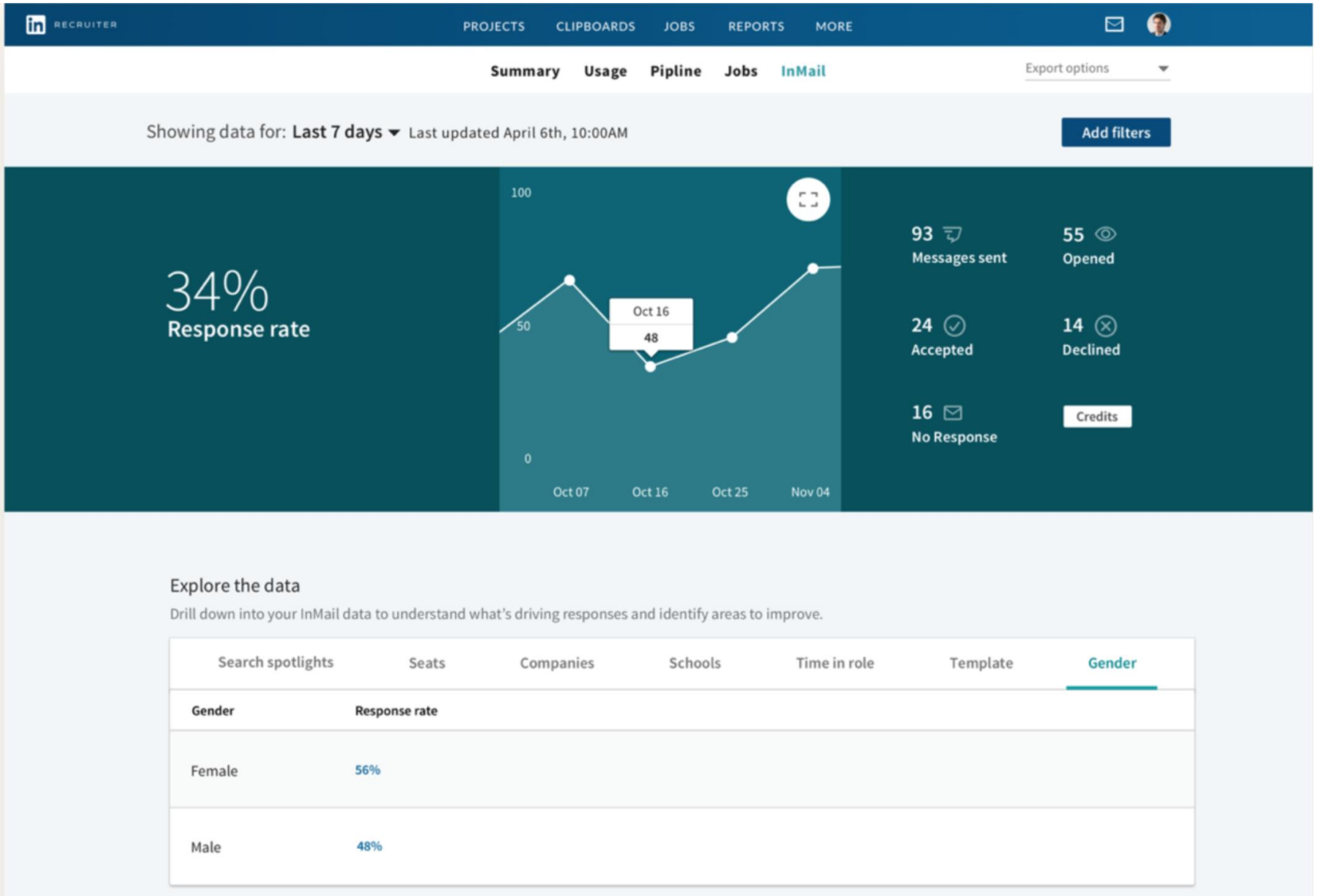
Plan for Diversity



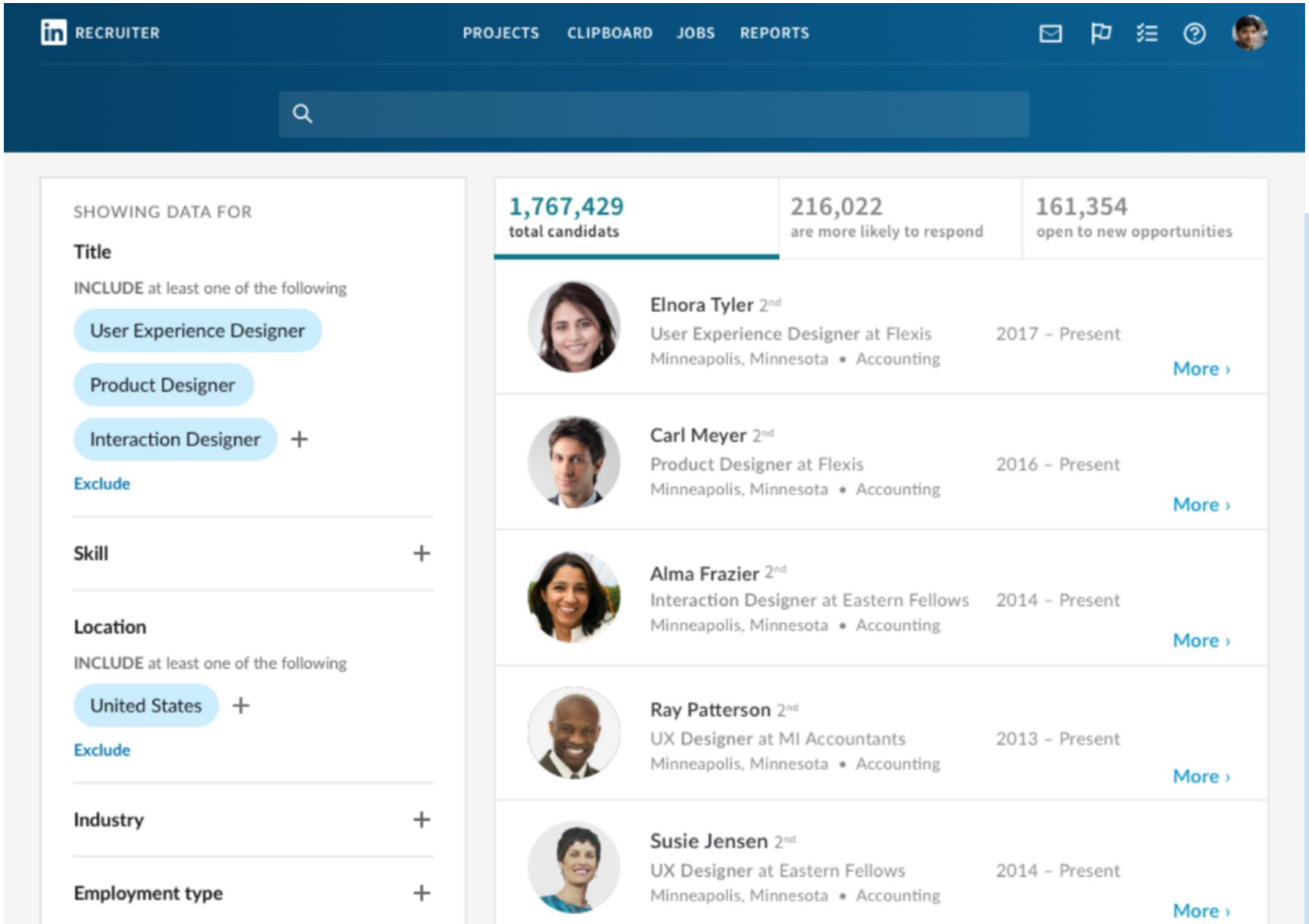
Identify Diverse Talent Pools



Inclusive Job Descriptions / Recruiter Outreach



Representative Ranking for Talent Search



S. C. Geyik, K. Kenthapadi, Building Representative Talent Search at LinkedIn, LinkedIn engineering blog post, October'18.

S. C. Geyik, S. Ambler, K. Kenthapadi, Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search, KDD'19.

Intuition for Measuring Representativeness

- Ideal: same distribution on gender/age/… for
  - Top ranked results and
  - Qualified candidates for a search request
    - LinkedIn members matching the search criteria

- Same proportion of members with each given attribute value across both these sets

- “Equal opportunity” definition [Hardt et al, NIPS’16]

Measuring (Lack of) Representativeness

- **Skew@k**
  - (Logarithmic) ratio of the proportion of candidates having a given attribute value among the top k ranked results to the corresponding proportion among the set of qualified candidates

$$Skew_v@k(\tau_r) = \log_e \left(\frac{p_{\tau_r^k, r, v}}{p_{q, r, v}} \right)$$

  - **MinSkew:** Minimum over all attribute values genders (e.g., the most underrepresented gender's skew value).
    - Skew = 0 if we have $\lfloor p_{q,r,v} * k \rfloor$ candidates from value v in the top k results

Reranking Algorithm for Representativeness

- Determine the target proportions within the attribute of interest, corresponding to a search request

- Compute a fairness-aware ranking of size k

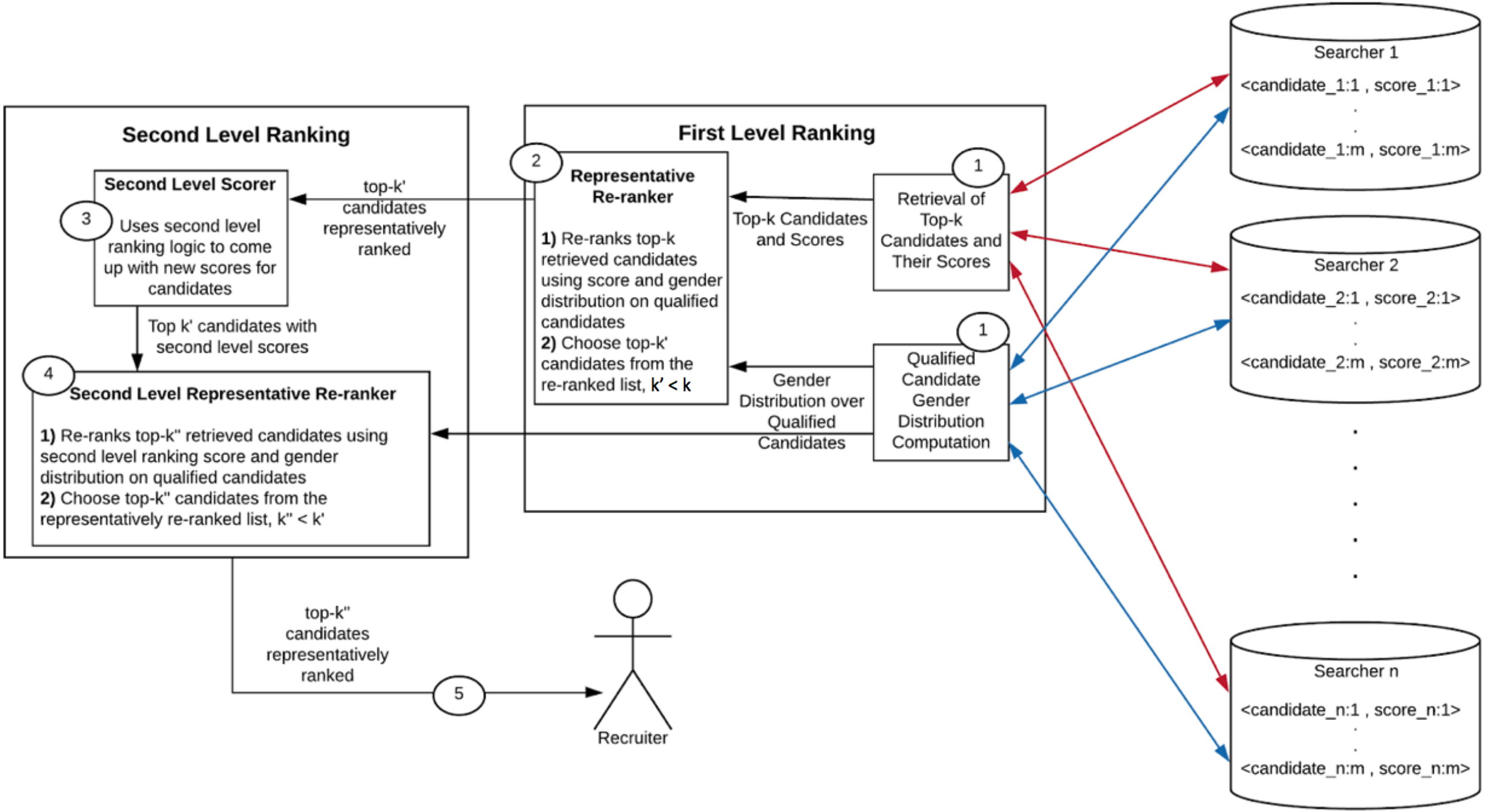
Target Proportions within the Attribute of Interest

- Compute the proportions of the values of the attribute (e.g., gender, gender-age combination) amongst the set of qualified candidates
  - “Qualified candidates” = Set of candidates that match the search query criteria
    - Retrieved by LinkedIn’s Galene search engine

- Target proportions could also be obtained based on legal mandate / voluntary commitment

Fairness-aware Reranking Algorithm

- Partition the set of potential candidates into different buckets for each attribute value
- Rank the candidates in each bucket according to the scores assigned by the machine-learned model
- Merge the ranked lists, balancing the representation requirements and the selection of highest scored candidates


Second Level Ranking
Second Level Scorer
3
Uses second level ranking logic to come up with new scores for candidates
top-k' candidates representatively ranked
Top k' candidates with second level scores
4
Second Level Representative Re-ranker
1) Re-ranks top-k'' retrieved candidates using second level ranking score and gender distribution on qualified candidates
2) Choose top-k'' candidates from the representatively re-ranked list, k'' < k'
First Level Ranking
2
Representative Re-ranker
1) Re-ranks top-k retrieved candidates using score and gender distribution on qualified candidates
2) Choose top-k' candidates from the re-ranked list, k' < k
Top-k Candidates and Scores
1
Retrieval of Top-k Candidates and Their Scores
Gender Distribution over Qualified Candidates
1
Qualified Candidate Gender Distribution Computation
Searcher 1
<candidate_1:1 , score_1:1>
<candidate_1:m , score_1:m>
Searcher 2
<candidate_2:1 , score_2:1>
<candidate_2:m , score_2:m>
Searcher n
<candidate_n:1 , score_n:1>
<candidate_n:m , score_n:m>
top-k'' candidates representatively ranked
5
Recruiter

Validating Our Approach

- A/B test over LinkedIn Recruiter users for two weeks
  - Over 95% of all searches are gender-representative compared to the qualified population of the search (3X improvement)
  - No significant change in business metrics (e.g., # InMails sent or accepted)

Metric	Vanilla	Fairness-aware
Queries with representative results	**33%**	**95%**
Average MinSkew@100	-0.259	-0.011 (p-value < 1e-16)
InMails Sent	-	± 1% (p-value > 0.5)
InMails Accepted	-	± 1% (p-value > 0.5)

- Ramped to 100% of LinkedIn Recruiter users worldwide

Lessons learned

- Post-processing approach desirable
  - Agnostic to the specifics of each model
    - Scalable across different model choices for our application
  - Robust to application-specific business logic
  - Easier to incorporate as part of existing systems
    - Build a stand-alone service or component for post-processing
    - No significant modifications to the existing components
  - Functions as a “fail-safe”
  - Complementary to efforts to reduce bias from training data & during model training

Acknowledgements

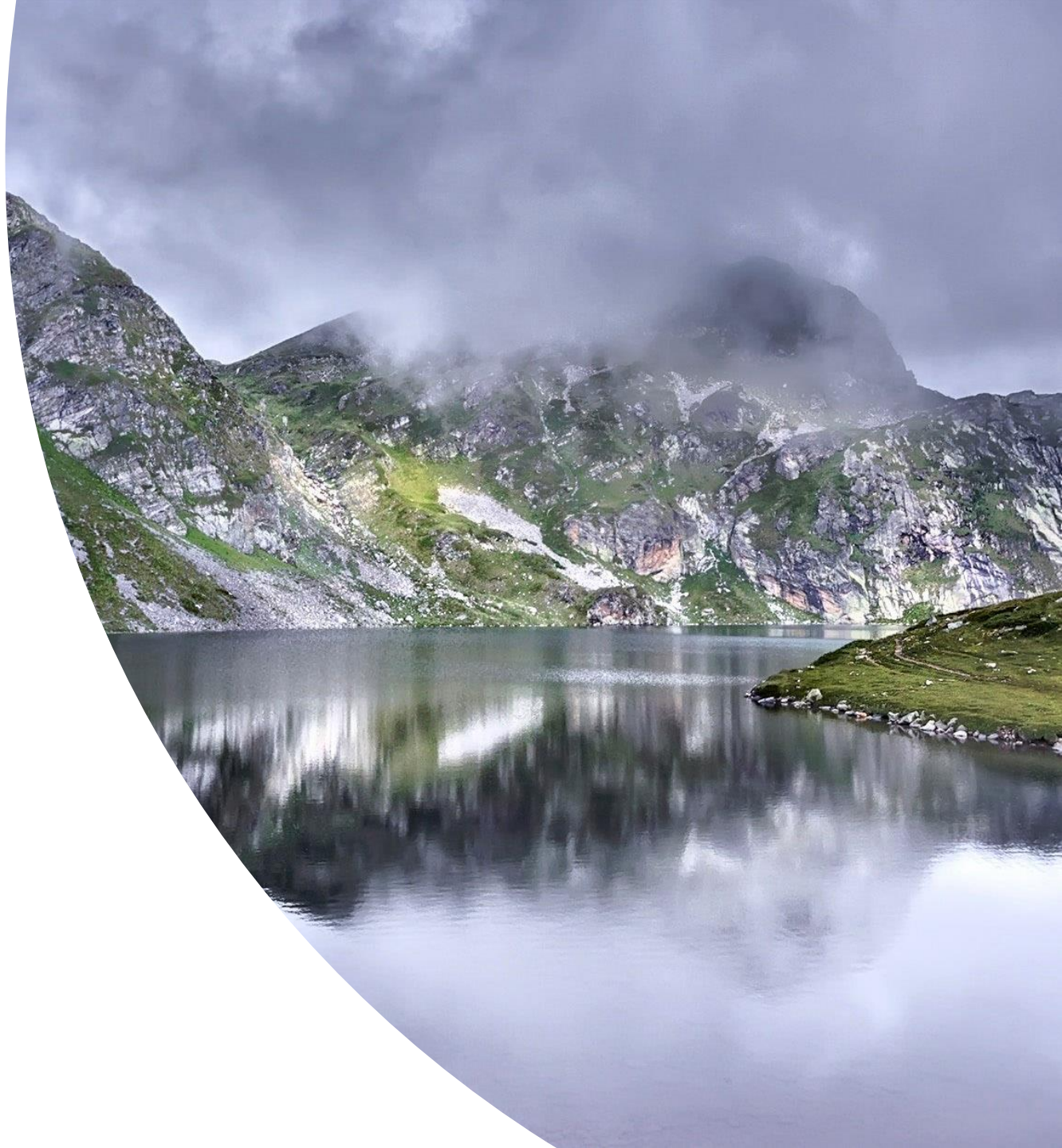
- Team:
  - AI/ML: Sahin Cem Geyik, Stuart Ambler, Krishnaram Kenthapadi
  - Application Engineering: Gurwinder Gulati, Chenhui Zhai
  - Analytics: Patrick Driscoll, Divyakumar Menghani
  - Product: Rachel Kumar

- Acknowledgements (in alphabetical order)
  - Deepak Agarwal, Erik Buchanan, Patrick Cheung, Gil Cottle, Nadia Fawaz, Rob Hallman, Joshua Hartman, Sara Harrington, Heloise Logan, Stephen Lynch, Lei Ni, Igor Perisic, Ram Swaminathan, Ketan Thakkar, Janardhanan Vembunarayanan, Hinkmond Wong, Lin Yang, Liang Zhang, Yani Zhang

Reflections

- Lessons from fairness challenges ➔ Need “Fairness by Design” approach when building AI products

- Case studies on fairness-aware ML in practice
  - Collaboration/consensus across key stakeholders (product, legal, PR, engineering, AI, …)



Key Takeaways

Good ML Practices Go a Long Way

01

Lots of low hanging fruit in terms of improving fairness simply by using machine learning best practices

- Representative data
- Introspection tools
- Visualization tools
- Testing

02

Fairness improvements often lead to overall improvements

- It’s a common misconception that it’s always a tradeoff

Breadth and Depth Required

01

Looking End-to-End is critical

- Need to be aware of bias and potential problems at every stage of product and ML pipelines (from design, data gathering, … to deployment and monitoring)

02

Details Matter

- Slight changes in features or labeler criteria can change the outcome
- Must have experts who understand the effects of decisions
- Many details are not technical such as how labelers are hired

Process Best Practices

Identify product goals
Get the right people in the room
Identify stakeholders

Policy

Select a fairness approach
Analyze and evaluate your system
Mitigate issues
Monitor Continuously and Escalation Plans
Auditing and Transparency

Technology

Beyond Accuracy

- Performance
- Cost
- Fairness and Bias
- Privacy
- Security
- Safety
- Robustness

The Real World is What Matters

Decisions should be made considering the real world goals and outcomes

understand these real world effects

- Social scientist, Lawyers, domain experts…
- Hire experts (even ones that don't code)

You need different types of testing

We need more research focused on people, applications, and real world effects

A lot of the current research is not that useful in practice

We need more social science + machine learning research

Key Open Problems in Applied Fairness

Key Open Problems in Applied Fairness

What if you don’t have the sensitive attributes?

When should you use what approach? For example, Equal treatment vs equal outcome?

How to identify harms?

Process for framing AI problems: Will the chosen metrics lead to desired results?

How to tell if data generation and collection method is appropriate for a task? (e.g., causal structure analysis?)

Processes for mitigating harms and misbehaviors quickly

Related Tutorials / Resources

- Sara Hajian, Francesco Bonchi, and Carlos Castillo, Algorithmic bias: From discrimination discovery to fairness-aware data mining, KDD Tutorial, 2016.
- Solon Barocas and Moritz Hardt, Fairness in machine learning, NeurIPS Tutorial, 2017.
- Kate Crawford, The Trouble with Bias, NeurIPS Keynote, 2017.
- Arvind Narayanan, 21 fairness definitions and their politics, FAT* Tutorial, 2018.
- Sam Corbett-Davies and Sharad Goel, Defining and Designing Fair Algorithms, Tutorials at EC 2018 and ICML 2018.
- Ben Hutchinson and Margaret Mitchell, Translation Tutorial: A History of Quantitative Fairness in Testing, FAT* Tutorial, 2019.
- Henriette Cramer, Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík, Hanna Wallach, Sravana Reddy, and Jean Garcia-Gathright, Translation Tutorial: Challenges of incorporating algorithmic fairness into industry practice, FAT* Tutorial, 2019.
- ACM Conference on Fairness, Accountability, and Transparency (ACM FAT*)

Fairness

Privacy

Related WWW'19 sessions:

1. Tutorial: Designing Equitable Algorithms for the Web
2. Tutorial: Economic Theories of Distributive Justice for Fair Machine Learning
3. Tutorial: Socially Responsible NLP
4. Tutorial: Privacy-Preserving Data Mining in Industry (Tuesday, 14:00 - 17:30)
5. Tutorial: Explainable Recommendation and Search (Tuesday, 14:00 - 17:30)
6. Workshop: FATE and Society on the Web (Tuesday)
7. Session: Fairness, Credibility, and Search (Wednesday, 10:30 – 12:30)
8. Session: Privacy and Trust (Wednesday, 16:00 – 17:30)
9. Special Track: Designing an Ethical Web (Friday)

Transparency

Explainability

Thanks! Questions?

- Tutorial website: **https://sites.google.com/view/www19-fairness-tutorial**

- Feedback most welcome ☺
  - slbird@microsoft.com, benhutch@google.com, kkenthapadi@linkedin.com, emrek@microsoft.com, mmitchellai@google.com