Fairness & Discrimination in Recommendation & Retrieval

Objectives

- Understand key concepts of algorithmic fairness
- Identify stakeholders with fairness concerns in an information access system
- Identify possible sources of unfairness in an information access system
- Assess the applicability of existing metrics and experimental protocols for assessing fairness concerns in a particular system

About Us

Michael Ekstrand - Assistant Professor, People and Information Research Team — Boise State University

Fernando Diaz - Principal Research Manager, Fairness, Accountability, Transparency, and Ethics — Microsoft Research Montreal

Robin Burke - Professor, That Recommender Systems Lab — University of Colorado, Boulder

People and Information Research Team

BOISE STATE UNIVERSITY

Microsoft

Motivating Examples

Embedding Bias and Review Analysis [Speer 2017]

Restaurant reviewing is a common activity

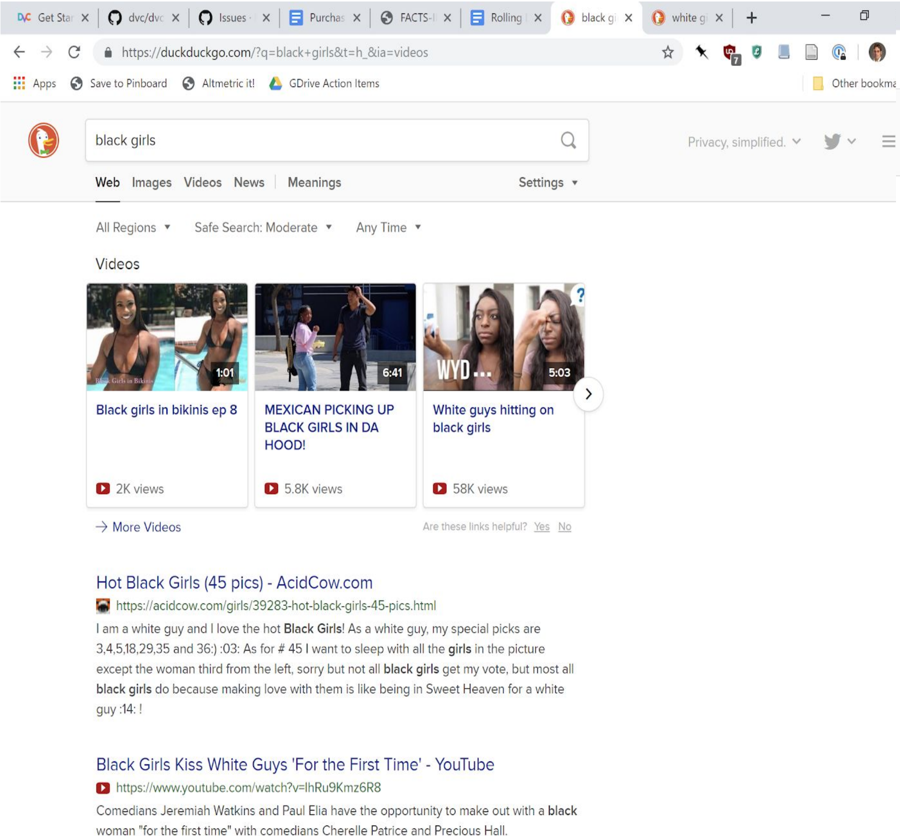
Mine reviews to recommend!

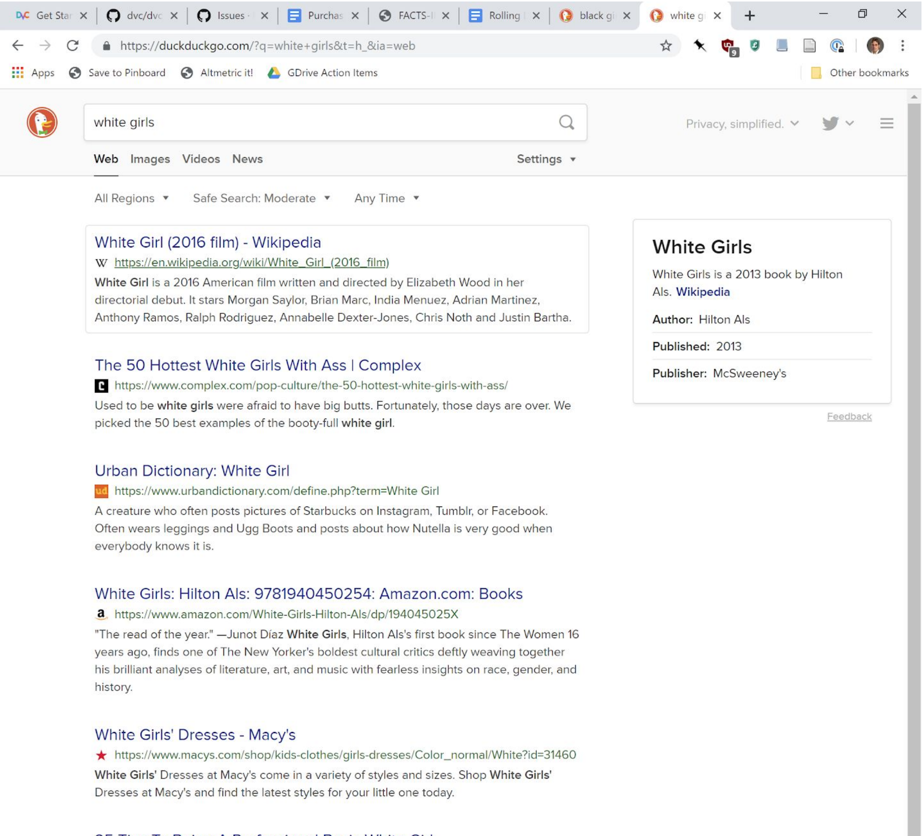
- Sentiment analysis?
- With word embeddings?

Why isn’t the recommender giving me any Mexican recommendations?

S U Noble. *Algorithms of Oppression*. 2018

Result Character [Noble 2018]





Dating Recommendations [Hutson et al. 2018]

If my dating profile says "no racial preference", who should I be recommended?

We quickly see that technical solutions and expertise are insufficient.

Get comfortable with being uncomfortable.

Scholarly Search

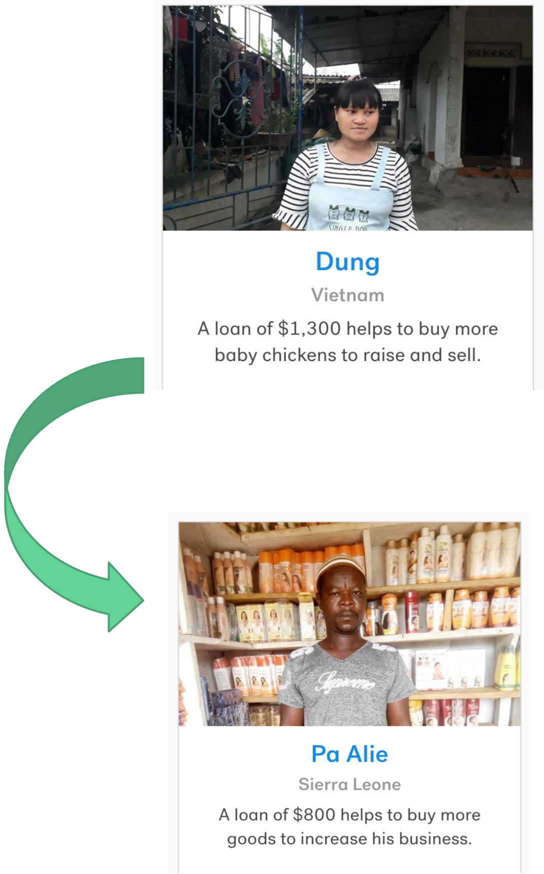
Do major research groups dominate search results?

Are smaller universities or labs disadvantaged in research discoverability?

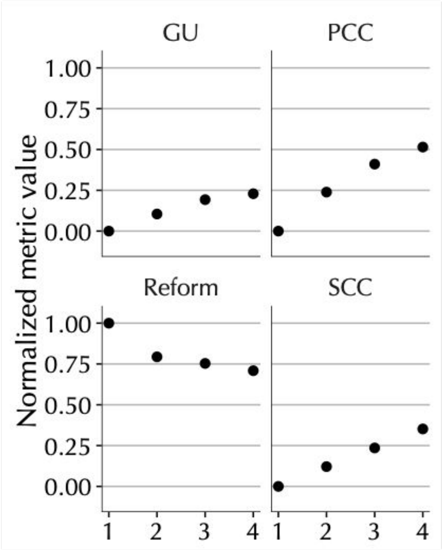
Economic Opportunity

If microloans in southeast Asia are funded more quickly than in sub-saharan Africa, should the system promote loans in Sierra Leone?

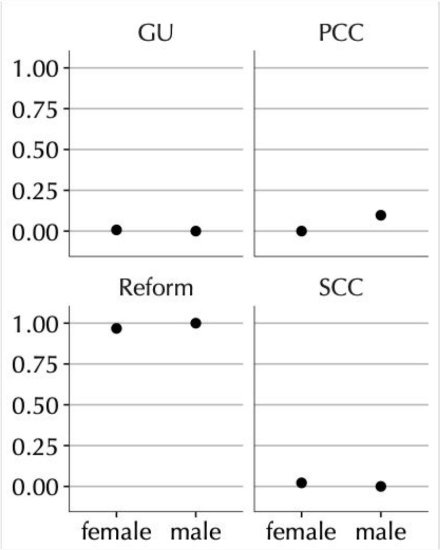
- What projects are "worthy"?
- What if the user has only ever lent to women in Vietnam?
- What is your organizational mission and can you be sure your users share it?



Who Gets Good Results?



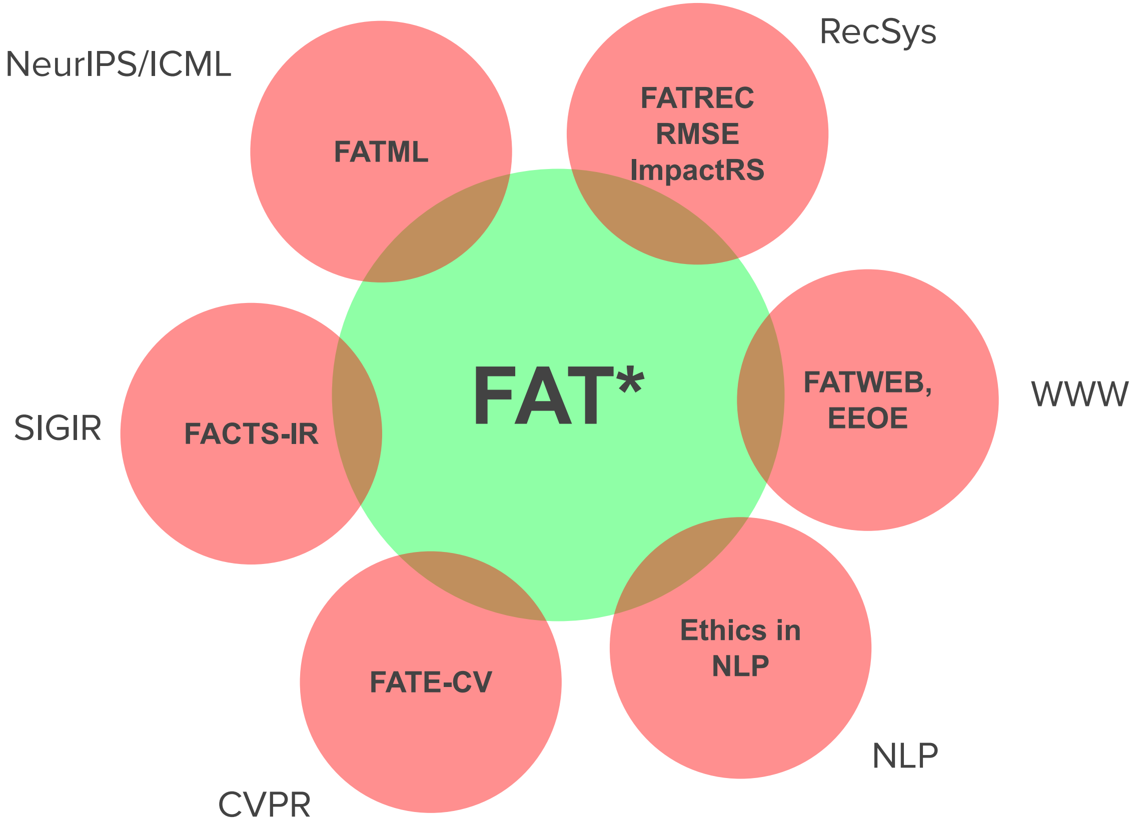
(a) age



(b) gender

web search performance can be biased across different demographic groups

Mehrotra, Sharma, Anderson, Diaz, Wallach, Yilmaz. Auditing search engines for differential satisfaction across demographics, 2017.


NeurIPS/ICML
FATML
RecSys
FATREC
RMSE
ImpactRS
FAT*
SIGIR
FACTS-IR
FATWEB,
EEOE
WWW
FATE-CV
CVPR
Ethics in
NLP
NLP

Overview

Information Access Systems

an information access system mediates an information consumer's interaction with a large corpus of information items.

- generalizes information retrieval and recommendation systems.
  - share interfaces
  - share fairness problems

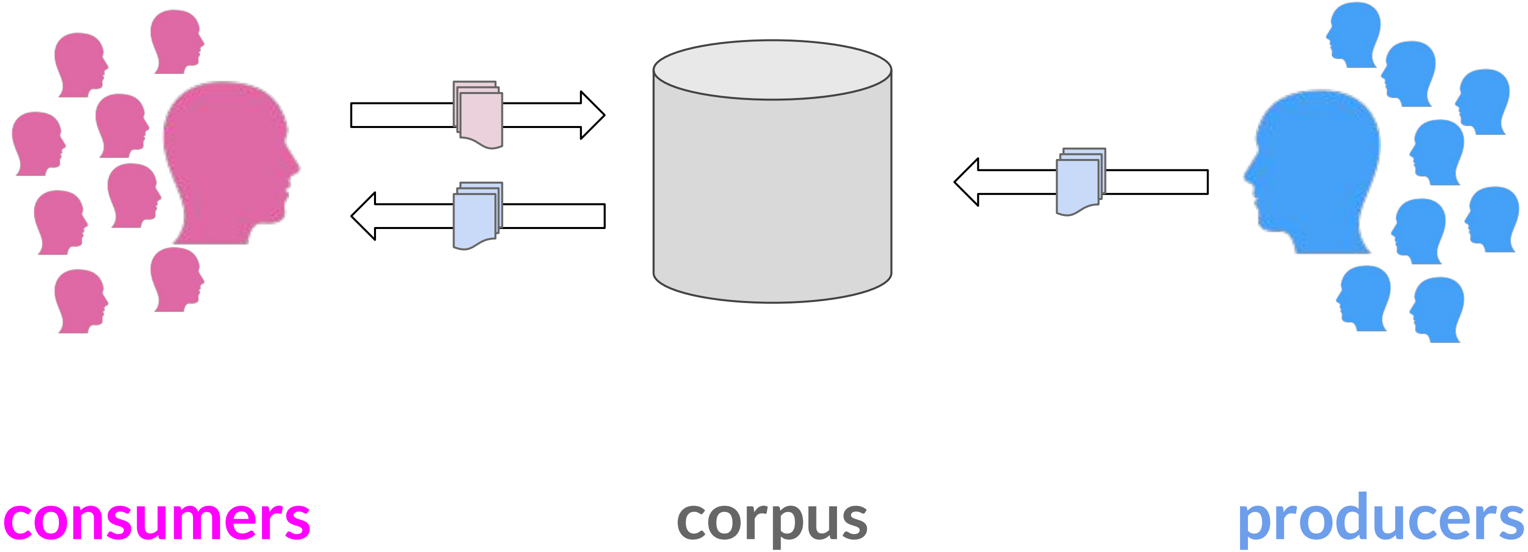
One View: Unified Scoring

$$s(i|u,h,x)$$

$$O(I|u,h,x)$$

- i: item
- u: user (and their historical profile / latent vectors)
- h: explicit task description (e.g. query)
- x: context

Information Access Systems



What is the problem?

- unconstrained, information access systems can reflect the bias inherent in data (e.g. consumer behavior, producer content).
    - biased demographics in user population
    - biased topical distribution in corpus
    - biased language in documents
    - biased opportunity to contribute documents
- algorithms often amplify small preferences and differences

Why is it important?

legal: information access—especially in settings like employment, housing, and public accommodation— potentially is or will be covered by anti-discrimination law.

publicity: disclosure of systematic bias in system performance can undermine trust in information access.

financial: underperformance for large segments of users leads to abandonment.

moral: professional responsibility to provide equal information access.

“The use of information and technology may cause new, or enhance existing, inequities. Technologies and practices should be as inclusive and accessible as possible and computing professionals should take action to avoid creating systems or technologies that disenfranchise or oppress people. Failure to design for inclusiveness and accessibility may constitute unfair discrimination.”

ACM Code of Ethics and Professional Conduct, section 1.4

“In order to promote inclusion and eradicate discrimination, librarians and other information workers ensure that the right of accessing information is not denied and that equitable services are provided for everyone whatever their age, citizenship, political belief, physical or mental ability, gender identity, heritage, education, income, immigration and asylum-seeking status, marital status, origin, race, religion or sexual orientation.”

IFLA Code of Ethics for Librarians and other Information Workers

Why is current practice insufficient?

no way to evaluate: unclear what we mean by fairness or how to measure.

no way to optimize: unclear how to optimize while respecting fairness.

Where do we look for answers?

Fairness is a *social* concept and inherently *normative*

Selbst et al.: fairness "can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms"

Engaging with these problems requires engaging with many disciplines:

- Law
- Ethics / philosophy
- Sociology
- Political science
- Many, many more

Many questions

You will probably leave today with more questions than answers.

That's normal and expected.

Our goal:

- Better questions
- Pointers into the literature to start looking for answers

Agenda

Part 1: Setting the Stage

- Motivating Examples
- Algorithmic Fairness
    - Problems and Concepts
    - Constructs, Metrics, and Results
    - Ensuring Fairness
- What’s Different about RecSys?

☕

Part 2: It Gets Harder

- Fair for Who? (Multisided)
- Fair How?
- Problem Space Taxonomy
- FairRec/IR/Rank Constructs
- Feedback Loops
- Fairness in Production
- Open Problems

Problems and Concepts

Organizing the Space

- Who is experiencing (un)fairness?
- How does that (un)fairness manifest?
- How is that (un)fairness determined?

Common Examples

Finance - system computes credit score/risk, decide to offer loan

Prediction goal: probability of default

Detention (either pretrial or post-conviction) - system computes risk score

Prediction goal: probability of failure-to-appear and/or new crime

College admissions

Prediction goal: likelihood to succeed? (less consistent)

Harm

Distributional harms arise when someone is denied a resource or benefit.

- Prison time
- Job opportunities
- Loans
- Search position
- Quality information

K Crawford. The Trouble with Bias. 2017

Harm

Representational harms arise when someone is *represented incorrectly* in the system or to its users.

- Misgendering
- Racial miscategorization
- Stereotyping (esp. reinforcing negative stereotypes)
- ‘Inverse’ representational harms: who shows up when searching for ‘ceo’?

Can happen to **content creators** or to **users**.

K Crawford. The Trouble with Bias. 2017

Representation Biases

Learning Representational Harms

Representation learning - let's embed {words, products, people} into vector spaces

What are you associated with in the vector space?

- Sentiment analysis - do genders or ethnicities have a sentiment?
- Association - are things like job descriptions embedded in ways that replicate sexism or racism?
    - Occupations project onto a gender axis
    - Goal might be orthogonality

Fair representation learning seeks to mitigate

T Bolukbasi, K-W Chang, J Zou, V Saligrama, A Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. 2016

From Representation to Distribution

- Mine restaurant reviews for recommendation
- Sentiment analysis to interpret reviews (item understanding)
- The embedding learned a negative sentiment for 'mexican'

R Speer. ConceptNet Numberbatch 17.04: better, less-stereotyped word vectors. 2017

Direct and Indirect

Direct discrimination

- Use protected class in decision-making
- Often illegal

Corresponds to *taste-based* in economics

Indirect discrimination

- Protected class affects results through correlates in other variables

Corresponds to *statistical* in economics

Example:

- Increasing insurance premium because you are Black -> direct
- Increasing insurance premium because of your neighborhood, and it is predominantly Black -> indirect

Basis of Fairness

Individual fairness says similar individuals should be treated similarly

- Two applicants with the same ability to repay a loan should receive the same decision

Group fairness says each salient group of people should be treated comparably.

- Black loan applicants should not be denied more often than white
- Often concerned with a *protected class* or *sensitive characteristic*
  - In U.S. context, anti-discrimination law provides this

Why is Individual Fairness Insufficient?

Fundamental reason: historical discrimination + measurement impossible

- Measures of individual merit are skewed
- Prospective outcomes may vary for social reasons

Example: SAT scores predict socioeconomic status.

Scores conflate *aptitude* and *preparation*

Why is Individual Fairness Insufficient?

Fundamental reason: historical discrimination + measurement impossible

- Measures of individual merit are skewed
- Prospective outcomes may vary for social reasons

Example: SAT scores predict socioeconomic status.

Scores conflate *aptitude* and *preparation*

Why should we assume a difference in score is a problem with the people and not the test?

Group Non-Fairness Constructs

Disparate treatment: members of different groups are *treated* differently

Applying different standards to people of different ethnicities

Disparate impact: different groups obtain different *outcomes*

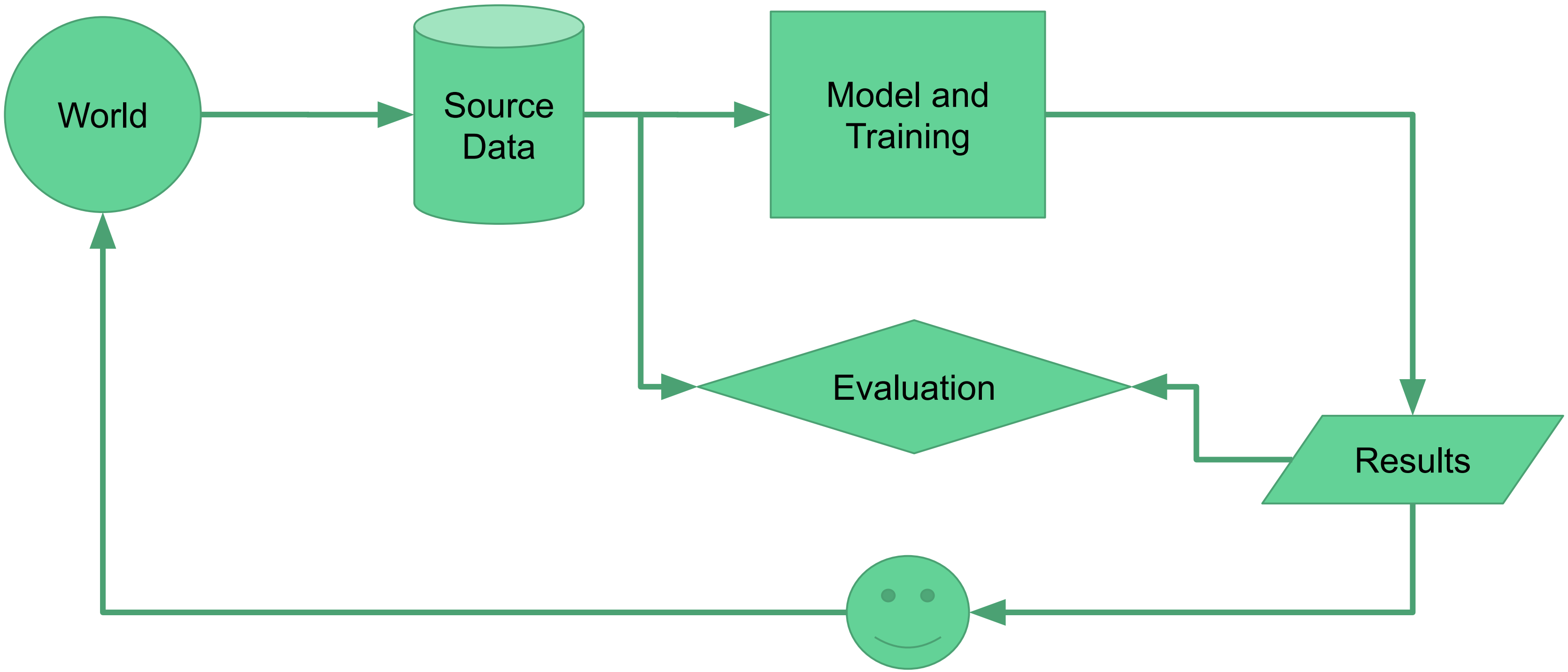
Men pass the employment test at a higher rate than other genders
Foundation of much U.S. anti-discrimination law

Disparate mistreatment: different groups have different *error rates*

A risk assessment tool is more likely to misclassify a black defendant as high-risk

questions?

Where does Unfairness Come From?



Unfairness in the world

- Different group sizes
  - Naive modeling learns more accurate predictions for majority group
- Historical and ongoing discrimination
  - Produces ‘unnatural’ distributions, e.g. redlining in the U.S. skews location, housing
  - Oppression skews social position, socioeconomic status, education, etc.
  - Arises from policy, practice, or both
  - Effects propagate after official practice ends

Unfairness in data

- Sampling strategy - who is included in the data?
- Response bias - who responds / submits data points?
- Proxy selection - valid and unbiased for variable of interest?
- Measurement (in)variance - is instrument consistent across subpopulations?
- Definitions of metrics - what standards or perspectives are reflected?
- Codebook - how is data recorded?
  - Especially important for sensitive variables such as gender, race, and ethnicity
- Cultural understanding - do we understand what the data mean in context?

Unfairness in models

- Using sensitive information (e.g. race) directly + adversely
- Algorithm optimization eliminates “noise”, which might constitute the signal for some groups of users

Unfairness is *usually* an emergent property of data + model.

Unfairness in evaluations

- Definition of Success - who is it good for, and how is that measured?
  - Who decided this? To whom are they accountable?
- How are relevant subgroups measured and aggregated in evaluation?
- All the data issues apply

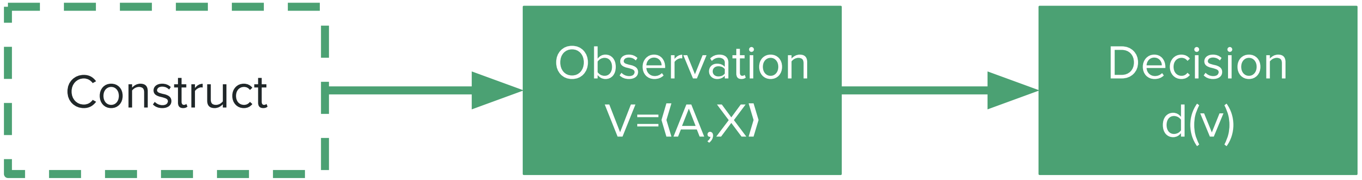
Unfairness in response

- Humans + computers do not compose
  - Does model output skew human response differently?
- Social factors can skew response
  - Community support for loan repayment, making court dates
- Response feeds into next round's training
  - Affects subsequent data collection too!
- Response affects the world (e.g. incarceration rates & distribution, finance access and its effects)

questions?

Constructs, Metrics, Results

Spaces, Skews, & Discrimination



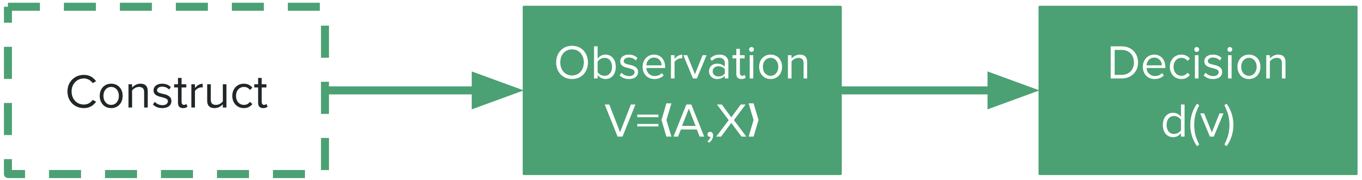
- Subjects have ‘true’ properties in **construct space** (ability to pay, relevance)
- System has access to **observation space**
- Computes results into **decision space**

Unfairness arises through **distortions** between spaces

- Random distortion - fine and recoverable
- Structural bias (e.g. systemic racism) manifests as systemic distortion
    - The *observation process* is skewed (violation of measurement invariance)

S Friedler, C Scheidegger, S Venkatasubramanian. On the (im)possibility of fairness. 2016.
Notation from S. Mitchell, E. Potash, S. Barocas. Prediction-Based Decisions and Fairness: A Catalog of Choices, Assumptions, and Definitions. 2018.

Spaces, Skews, & Discrimination



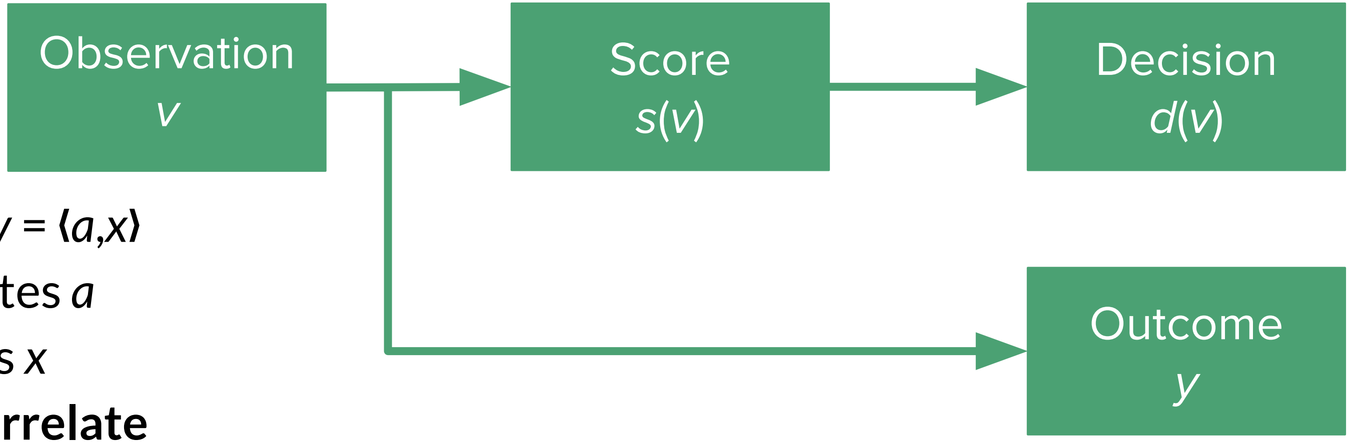
Key results:

- Individual and group fairness operate with incompatible axioms
  - Individual fairness requires ‘what you see is what you get’
  - Group fairness seeks to correct systemic discrimination
- Discrete decision spaces (common!) preclude (individual) fairness

Unclear when ranking or in repeated probabilistic decision processes

S Friedler, C Scheidegger, S Venkatasubramanian. On the (im)possibility of fairness. 2016.

Notation



Observed variables *v* = ⟨*a*,*x*⟩

Sensitive attributes *a*

Other attributes *x*

***x* and *a* often correlate**

Outcome *y*

Decision *d*(*v*), often based on score *s*(*v*)

Goal: *d*(*v*) = *y* (e.g. *d*(*v*) = 1 to offer a loan, and *y* = 1 if it is repaid)

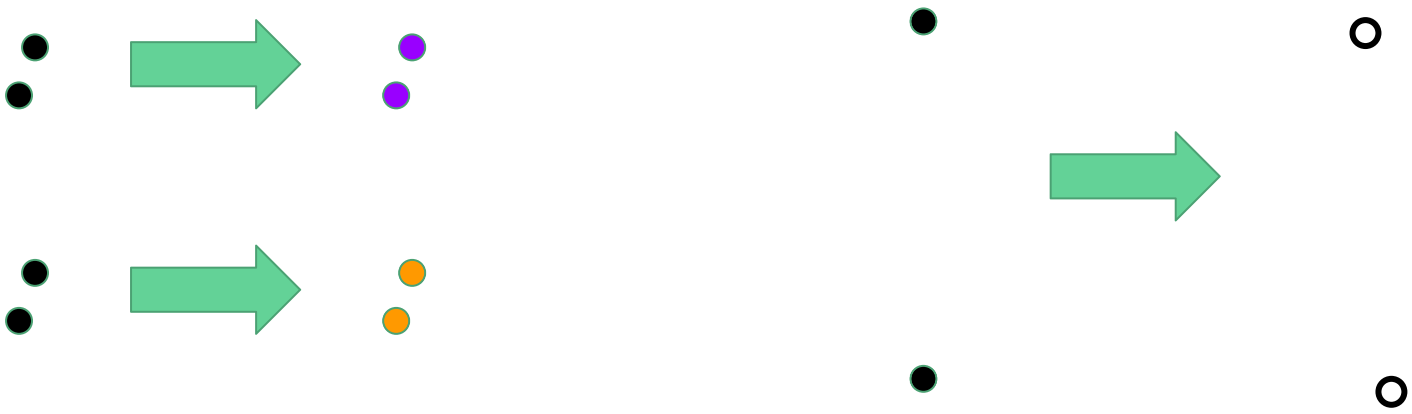
S. Mitchell, E. Potash, S. Barocas. Prediction-Based Decisions and Fairness: A Catalog of Choices, Assumptions, and Definitions. 2018.

Discrimination Types

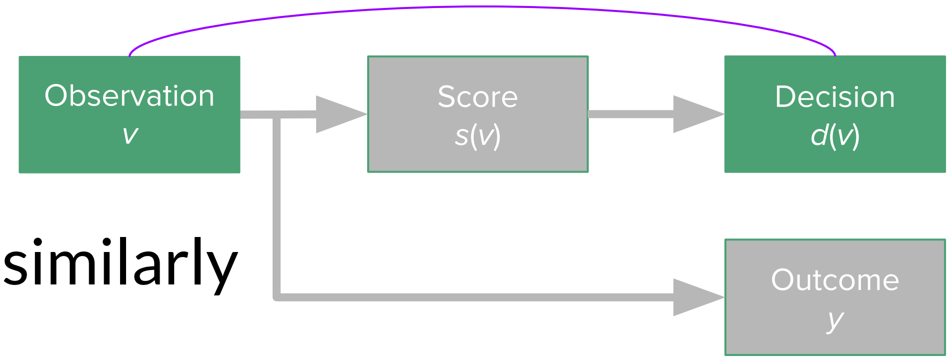
Direct discrimination - use the sensitive attribute

Indirect discrimination - arises from redundancies between sensitive & insensitive attribute

Individual Fairness



Individual Fairness



Goal: treat similar individuals similarly

Prerequisite: task-specific distance metric $m(v_1, v_2)$

decision distribution metric $m'(d(v_1), d(v_2)))$

Definition of Fair: $\forall v_1, v_2 . m'(d(v_1), d(v_2)) \leq m(v_1, v_2)$

If two individuals are similar, they receive similar outcomes

Says nothing about dissimilar individuals

C Dwork, M Hardt, T Pitassi, O Reingold, R Zemel. Fairness through awareness. 2012

Similarity and Recommendation

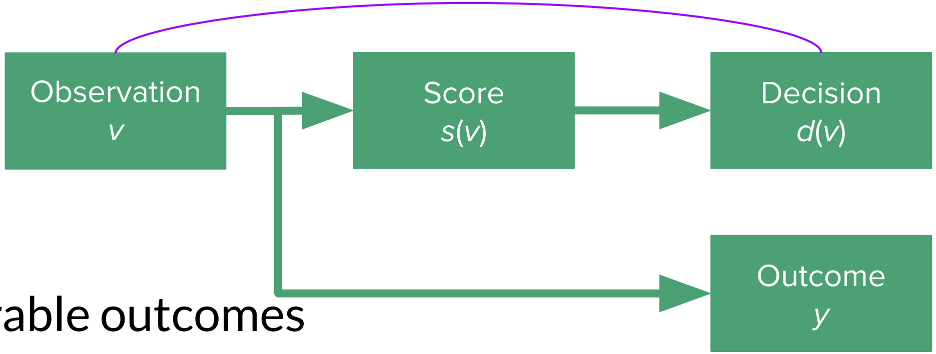
People with the same financial situation should receive the same loan decision.

Should similar documents both be recommended?

- Single ranking - diversity says no!
- Multiple rankings - maybe they get the same chance, but don’t appear together?

More on this later.

Statistical Parity



Goal:

different groups experience comparable outcomes

outcome is statistically independent of sensitive attribute

Prerequisite: sensitive attribute or group membership (e.g. race)

Definition of Fair: $E[d(v)|a] = E[d(v)]$

Disparate Impact Standard (U.S. law):

$$\Pr[d(v) = 1|a = 0] \geq 0.8 \cdot \Pr[d(v) = 1|a = 1]$$

Key insight [Dwork]: **group-blindness does not ensure equitable group outcomes**

C Dwork, M Hardt, T Pitassi, O Reingold, R Zemel. Fairness through awareness. 2012

M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian. Certifying and removing disparate impact. 2015

Why Statistical Parity?

It’s unconditioned on outcome or predictive variables - why is this ok?

- Predictive variables correlate - should we have a strong prior correlated components being irrelevant?
- Non-sensitive covariates are an opportunity to hide or launder bias

Partially inherited from U.S. law

One framing: statistical parity reflects a **strong prior** that advantaged and disadvantaged people are **fundamentally the same** in their relevant characteristics.

Error Parity

Goal:

different groups experience different misclassification rates

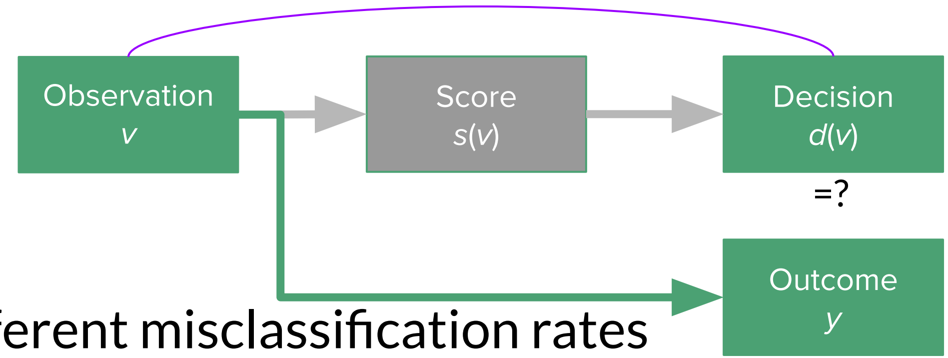
Example: recidivism prediction

- Defendant info ⇒ risk classification (‘high risk’)
- FPR: classified as high-risk when would not recidivate
- FNR: low-risk when would recidivate

If FPR_black > FPR_white, then the system is more likely to **falsely accuse** a black defendant than a white defendant

FNR_white > FNR_black: system more likely to let white defendant **off the hook**

Error Parity



Goal:

different groups experience different misclassification rates

Prerequisites: protected class / attributes

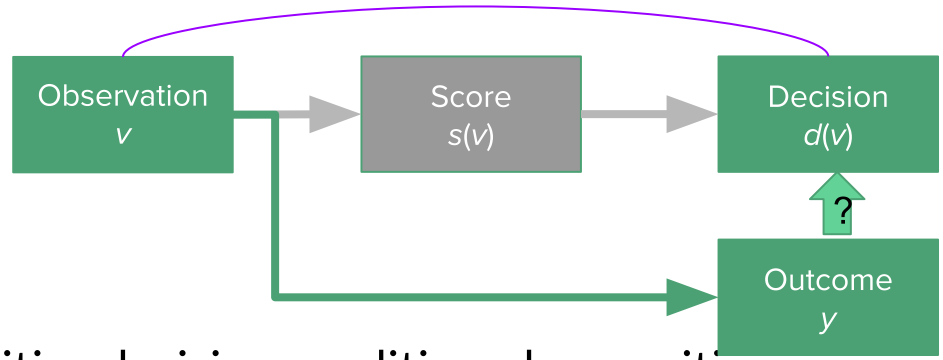
Definition of Fair (FPR):

$$\Pr[d(v) = 1 | y = 0, a] = \Pr[d(v) = 1 | y = 0]$$

Violations are **disparate mistreatment.**

M B Zafar, I Valera, M Gomez Rodriguez, K Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. 2017

Recall Parity



Goal:

Groups have equal likelihood of positive decision conditioned on positive outcome ("equal opportunity" for creditworthy people to get loans)

Prerequisites: protected class / attributes

Definition of Fair: $\Pr[d(v) = 1 | y = 1, a] = \Pr[d(v) = 1 | y = 1]$

In practice - requires time travel.

Suitable for supervised learning (but needs constant review)

M Hardt, E Price, N Srebro. Equality of Opportunity in Supervised Learning. 2016

Calibration and Predictive Value Parity

Goal:

Judgments are equally predictive across groups

Scores are equally predictive across groups

Definition of Fair:

equal PPV: $\Pr[y|d(v), a] = Pr[y|d(v)]$

calibration: $\Pr[y|s(v), a] = Pr[y|s(v)]$

A Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. 2017

J Kleinberg, S Mullainathan, M Raghavan. Inherent Trade-offs in the Fair Determination of Risk Scores. 2017

Expanding the Concept Space

Any marginal of the confusion matrix can be used to define a fairness metric:

- Equal accuracy
- Equality of any error metric

We can also look at scores within any category:

- Balance for positive class - scores for positive cases should be equivalent between groups $E[s(v)|y = 1, a] = E[s(v)|y = 1]$

S Mitchell, E Potash, S Barocas. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. 2018

questions?

Tradeoffs

If *base rates* are different, you cannot simultaneously equalize:

- False positive rate
- False negative rate
- Positive predictive value

System *will* be unfair, by some definition.

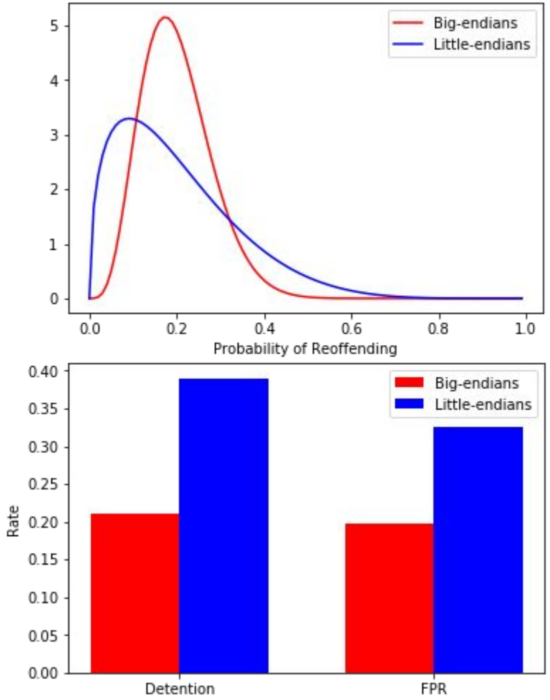
A Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. 2017
J Kleinberg, S Mullainathan, M Raghavan. Inherent Trade-offs in the Fair Determination of Risk Scores. 2017

Tradeoffs (continued)

If base rates are the same, parity can still be violated.

Model: subject has criminality score *p*; recidivates with probability *p*.

- Same mean probability (0.2)
- Perfectly calibrated
- Threshold: detain 30% (no disparate treatment)
- Unequal detention rates (**disparate impact**)
- Unequal FPR (**disparate mistreatment**)



S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, A. Huq. Algorithmic Decision Making and the Cost of Fairness. 2017.
Figures from our replicated simulation.

What Does This Mean?

Different concepts of (in)justice map to different metrics

Disparate treatment — *people should be treated the same regardless of group*

Use the same model and thresholds

Disparate impact — *groups should experience equal outcome likelihood*

Statistical parity metrics

Disparate mistreatment — *groups should experience equal unjustified adversity*

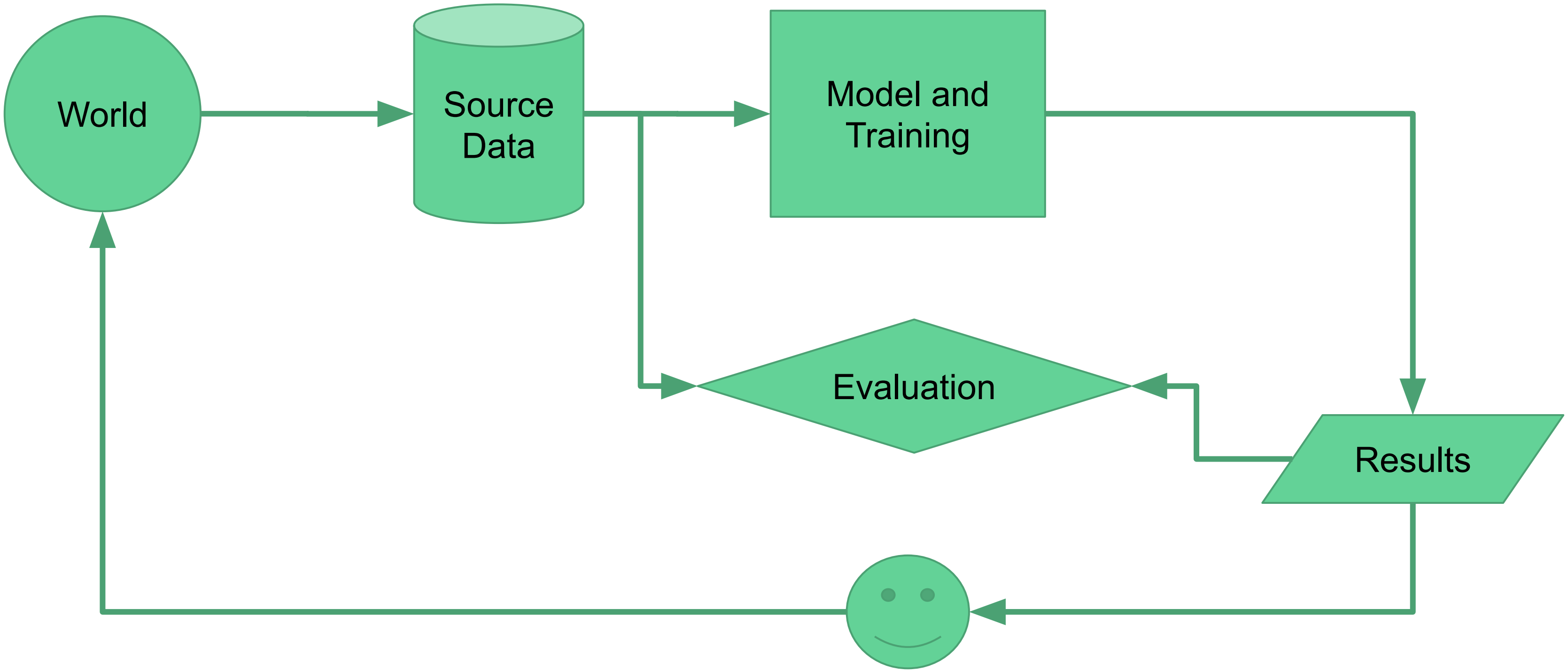
Error parity metrics

You cannot have it all. Applications will differ in what is most important.

Z Lipton, J McAuley, A Chouldechova. Does mitigating ML's impact disparity require treatment disparity? 2018

What About The People?

Scores are rarely the end of the line!



What About The People?

Scores are rarely the end of the line!

Response is biased

- Skews in-practice outcomes
- Biases subsequent model retraining
  - Ex: Impossible to learn that a rejected option would have been good after all
- Presence of risk scores can **increase** decision disparity

B Green, Y Chen. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. 2019

Pitfalls of Fairness

Selbst et al. identify 5 *abstraction traps*:

- Framing - isolating technical components from their surrounding sociotechnical contexts and human response
- Portability - assuming equivalence between social contexts
- Formalism - assuming operationalizations can fully capture social concepts
- Ripple Effect - overlooking changes to the social system that arise from introducing or modifying technology
- Solutionism - assuming technology is the best (or even a good) solution

A Selbst, d boyd, S Friedler, S Venkatasubramanian, J Vertesi. Fairness and Abstraction in Sociotechnical Systems. 2019

Pitfalls 2

- Asymmetric feedback (we don't learn from denied loans)
    - D Ensign, S Friedler, S Neville, C Scheidegger, S Venkatasubramanian. Decision making with limited feedback: Error bounds for predictive policing and recidivism prediction. 2018.
- Secondary effects of decision processes
    - What effect does incarceration have on crime?
    - What effect does representation in book authorship have on future production?

Give up?

While our systems are running, lives are materially impacted.

Or lost.

More Reading

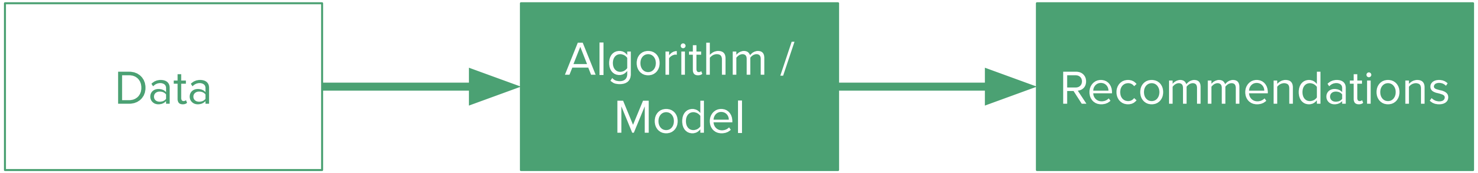
- 21 Definitions of Fairness and Their Politics [Narayanan 2018]
- Mirror Mirror [Mitchell]
- Prediction-Based Decisions and Fairness [Mitchell et al. 2018]
- 50 Years of Test (Un)fairness [Hutchinson and Mitchell 2019]
- Fairness and Abstraction in Sociotechnical Systems [Selbst et al. 2019]
- Where Fairness Fails [Hoffman 2019]

All in the bibliography.

questions?

Fairness Methods

Pre-processing



If bias is present in the data, we can de-bias before building a model

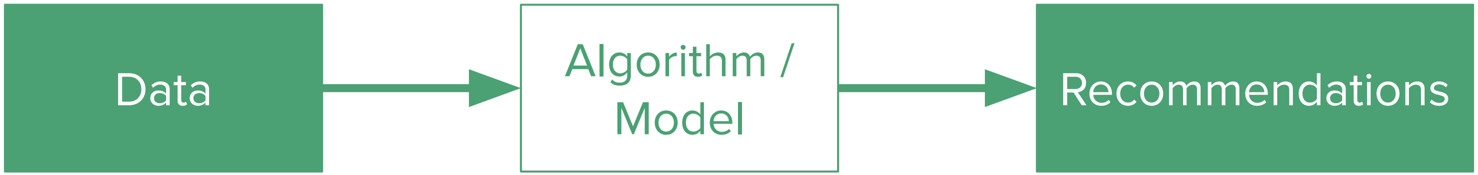
Data relabeling/repair to remove disparate impact [Feldman, et al. 2015, Kamiran et al. 2012, Salimi et al. 2019]

Can go as far as to obscure data within variables!

Data sampling [Hajian & Domingo-Ferrer 2013]

D Ensign, S Friedler, S Neville, C Scheidegger, S Venkatasubramanian. Runaway Feedback Loops in Predictive Policing. 2018
M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian. Certifying and removing disparate impact. 2015
F Kamiran, T Calders. Data preprocessing techniques for classification without discrimination. 2012
S Hajian, J Domingo-Ferrer. A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. 2013

Modifying the Algorithm



Alter the objective of the algorithm to emphasize fairness

Typically by adding regularization

T Kamishima, S Akaho, H Asoh, J Sakuma. Considerations on Fairness-Aware Data Mining. 2012
T Kamishima, S Akaho, H Asoh, I Sato. Model-Based Approaches for Independence-Enhanced Recommendation. 2016
R Burke, N Sonboli, A Ordonez-Gauger. Balanced Neighborhoods for Multi-sided Fairness in Recommendation. 2018

Post-processing Algorithm Scores

Example: risk prediction

Problem: same threshold results in disparate impact

Solution: use per-group thresholds

Solution: re-engineer test / features (but see tradeoffs above!)

Post-processing Algorithm Outputs

Example: word embeddings

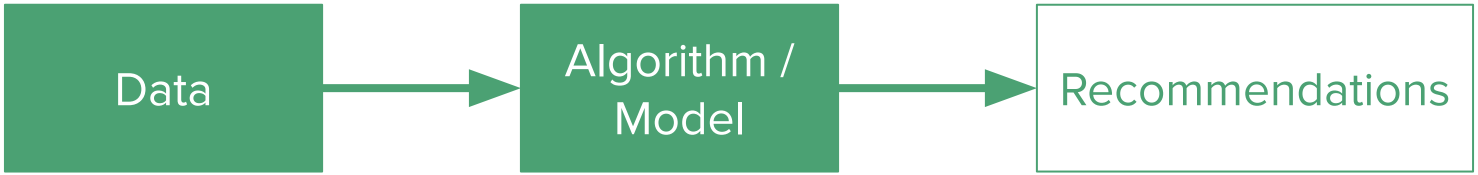
Problem: word embeddings encode sexist & racist skews

Demonstration: project ‘neutral’ words onto a gender axis

Solution: learn a transformation to re-embed words, preserving inner products subject to orthogonality constraints on target words

T Bolukbasi, KW Chang, J Zou, V Saligrama, A Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. 2016

Post-processing Recommendations



Re-ranking recommendation results for enhanced fairness

Greedy methods [Zehlike et al. 2017, Liu et al. 2019]

Constraint-satisfaction methods [Singh & Joachims, 2018]

Post-processing Decision Feedback

Example: predictive policing

Problem: bandit setting amplifies small differences into large ones (allocate all police to an area with 5% more crime)

Solution: sample feedback with inverse probability scoring

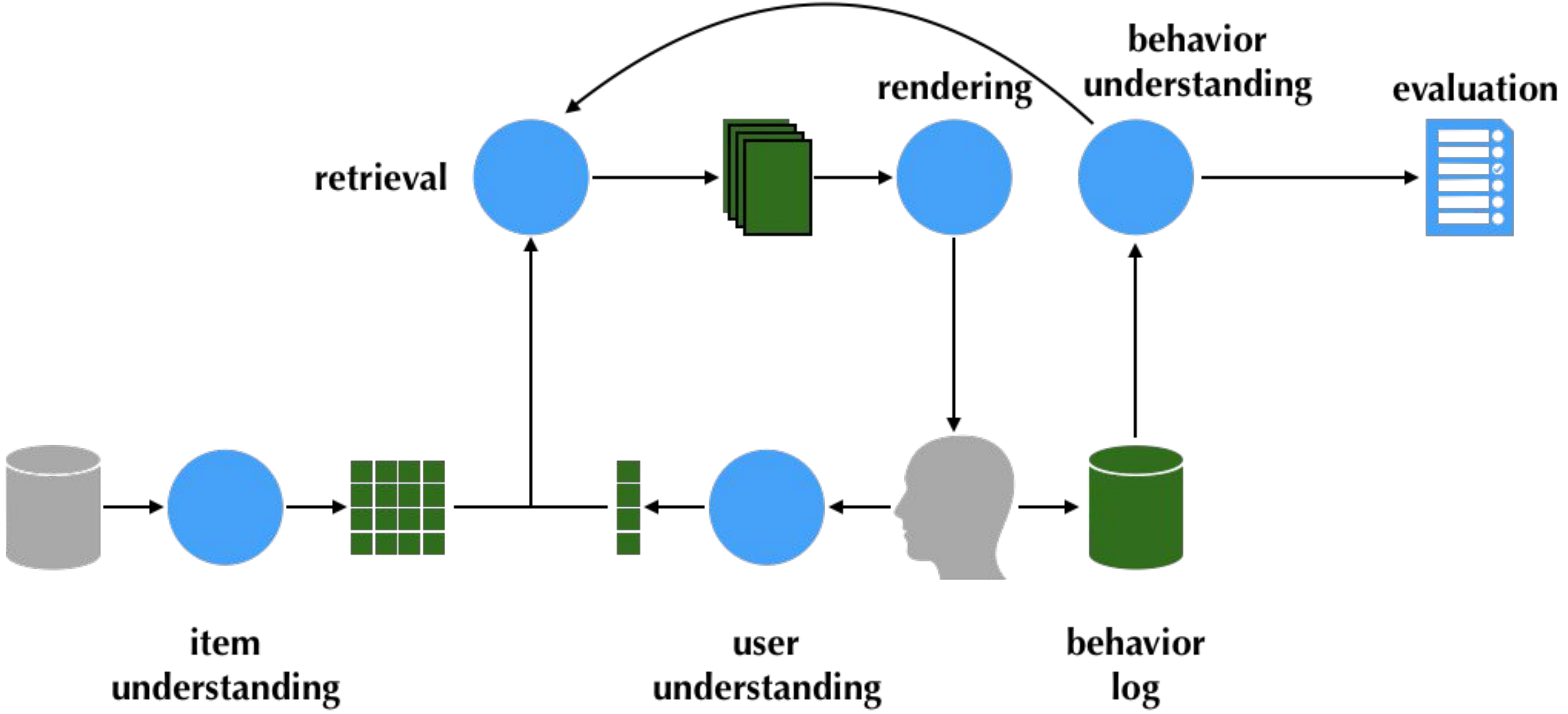
Solution: reevaluate structure of policing system

D Ensign, S Friedler, S Neville, C Scheidegger, S Venkatasubramanian. Runaway Feedback Loops in Predictive Policing. 2018

questions?

RecSys: What’s Different

Information Access Pipeline



“Classical” fairness setting

Mitchell et al:

- Classification: high/low risk of [crime, default, job failure]
- Decisions and consequences are individual and independent
- One-shot process

Also:

- Process independent of “user”/decision-maker (e.g. loan officers are interchangable)

Exceptions to many of the above… (e.g. selective college admissions, reinforcement learning)

S Mitchell, E Potash, S Barocas. Prediction-Based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions. 2018

Retrieving and recommending

- Evaluating ranked lists involves a user model
    - Ranking, not classification - violates independence
    - Classification views have small, fixed number of positive decisions (e.g. P@k)
- Queries are repeated - more than one opportunity for decisions
    - Opportunity to address first point
- Outcome (relevance / utility) is subjective and personalized
    - Different users have different knowledge, styles, informational preferences
    - Components of relevance are pure personal preference, esp. in recommendation
- Multiple sets of stakeholders with fairness concerns

SOME ETHICAL AND POLITICAL IMPLICATIONS
OF THEORETICAL RESEARCH IN INFORMATION SCIENCE

Nicholas J. Belkin
The City University
London, England

and

Stephen E. Robertson
University College
London, England

We have suggested both reasons and means for limiting theoretical investigations in information science. Much of this paper has been based on our own social ideology, which, although we think it correct for the situation, we must admit is arguable. We argue here for the necessity of making explicit a social ideology, such as ours, and acting upon it. By such action, we mean, for instance, attempting to develop a science which cannot be used for malign purposes. We would like to finish by emphasizing two points: we must perform such self-conscious examination and limitation in order to keep our theoretical activities related to their social context; and, we must do this before our developing theories reach a point where they might be misapplied - for by then it will be too late to prevent their being misapplied.

Belkin, Robertson, Some Ethical Implications of Theoretical Research in Information Science, 1976.

coffee

Agenda

Part 1: Setting the Stage

- Motivating Examples
- Algorithmic Fairness
  - Problems and Concepts
  - Constructs, Metrics, and Results
  - Ensuring Fairness
- What’s Different about RecSys?

☕

Part 2: It Gets Harder

- Fair for Who? (Multisided)
- Fair How?
- Problem Space Taxonomy
- Fair IR/Rec/Rank Constructs
- Feedback Loops
- Fairness in Production
- Open Problems

Fair for Who?

Multisided Fairness

Different stakeholders have **different concerns**

- **Consumers** want quality of service, access to information
- **Producers** want opportunity

How are these fairly allocated?

Different applications give rise to different tradeoffs.

R Burke. Multisided Fairness for Recommendation. 2017

Who does Information Access Affect?

Users

Vendors

Stockholders

Authors

Publishers

Society

Consumer Fairness

Consumer fairness is violated if user experience differs in an unfair way

- Quality of service (result relevance, user satisfaction)
- Resulting information (different, lower-paying job listings)
- Costs of participation (differential privacy risks)

Group recommendation has long been concerned with fairness across group members

Provider Fairness

Provider fairness is violated if content creators are treated unfairly

- Different opportunity to be read/purchased/cited
- Different visibility
- Different costs of participation

Publishers and authors are both providers, with different concerns.

Diversity and Subject Fairness

Subject fairness is violated if information subjects are not fairly represented or not fairly treated

- News results omitting rural issues
- Medical results not representative of population
    - Scholarly papers skewed towards particular populations
    - Diseases disproportionately affecting certain populations underrepresented
- Image search results not representative of population

Closely related to diversity, but stems from **different normative concerns**

Fair How?

Individual Fairness

- Each user gets comparable quality of service
  - Already standard practice
- Each provider gets comparable opportunity for user engagement
  - Conditioned on relevance
  - Different attention/relevance curves induce disparities

Group Fairness

- System does not systematically underserve groups of users
- System does not disadvantage groups of providers
- System does not disadvantage groups of subjects

questions?

Consumer Fairness

Consumer Fairness

Consumer fairness is violated if user experience differs in an unfair way

- Quality of service (result relevance, user satisfaction)
- Resulting information (different, lower-paying job listings)
- Costs of participation (differential privacy risks)

Not the most widely-studied

Quality of Service: Fundamental Cause

Aggregate quality/accuracy/response emphasizes majority populations.

User have different...

- ... reading fluency
- ... background knowledge
- ... tastes and preferences
- ... contexts of use

Warm-up: Fairness in Group Recommendations

- Recommending for a group of people
- How do you elicit preferences?
- How do you balance group member utilities?

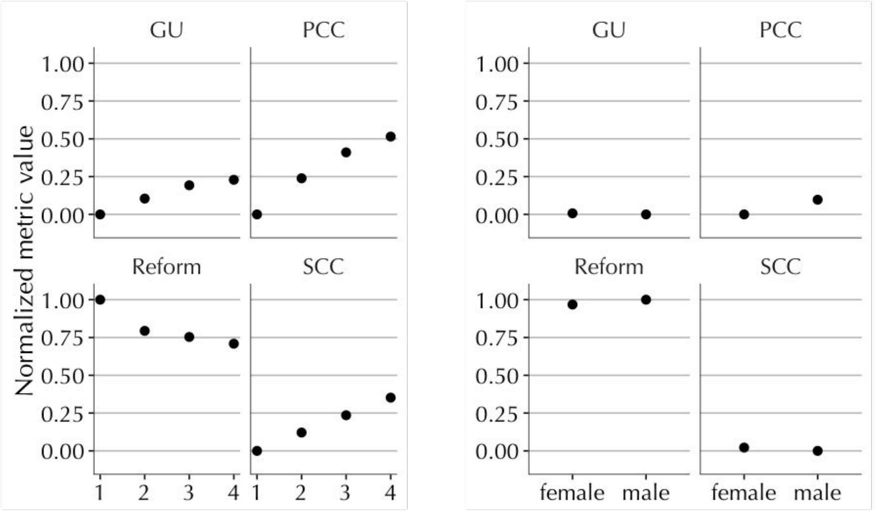
Long-studied in group recommender systems.

Oral history: giving vetos leads to least-offensive

A Delcic, J Neidhardt, T T Nguyen, F Ricci, L Rook, H Werthner, M Zanker. Observing Group Decision Making Processes. 2016
D Serbos, S Qi, N Mamoulis, E Pitoura, P Tsaparas. Fairness in Package-to-Group Recommendations. 2017
L Xiao, Z Min, Z Yongfeng, G Zhaoquan, L Yiqun, M Shaoping. Fairness-aware Group Recommendation with Pareto-Efficiency. 2017

Differential Satisfaction

User satisfaction should be independent of protected attribute.



(a) age (b) gender

R Mehrotra, A Anderson, F Diaz, A Sharma, H Wallach, E Yilmaz. Auditing search engines for differential satisfaction across demographics. 2017

DS Analysis: Context Matching

Causal inference technique simulating a matched pairs experiment

- For each data point in one group, find match in another
- Isolates effect of group membership

Matched intent w/ final success result (navigational only)

- Controls for query + intent
- Limits data + generalizability

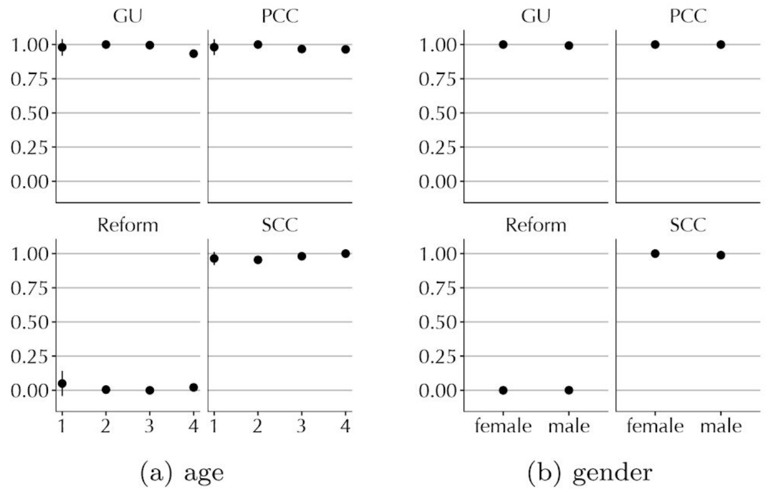


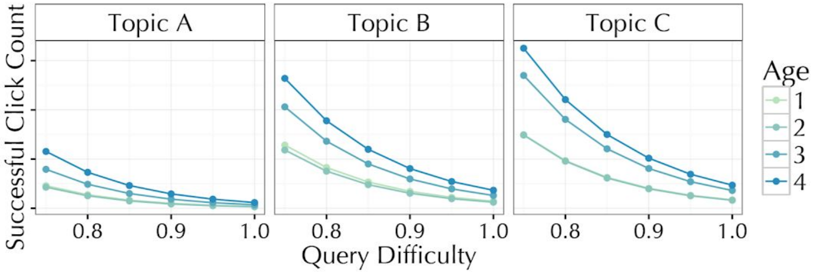
Figure 3: Context-matched normalized query-averaged values for each metric by age groups (a) and genders (b). "GU" denotes graded utility; "PCC" denotes page click count; "Reform" denotes reformulation rate; "SCC" denotes successful click count. Error bars (one standard error) are present in all plots, but are mostly so small that they cannot be seen.

R Mehrotra, A Anderson, F Diaz, A Sharma, H Wallach, E Yilmaz. Auditing search engines for differential satisfaction across demographics. 2017

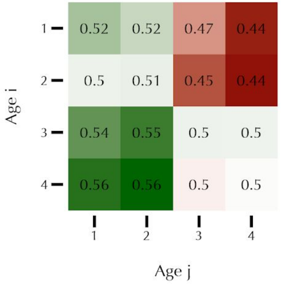
DS Analysis: Linear Modeling

Dependent variable: metric or pairwise satisfaction ordering (Si > Sj)

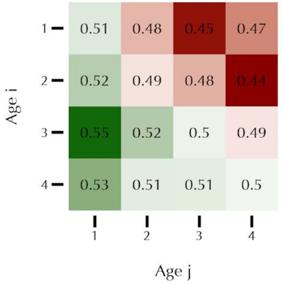
Independent variables: age, gender, query difficulty (for metric model)



(d) successful click count (hardest queries)



(a) Bing



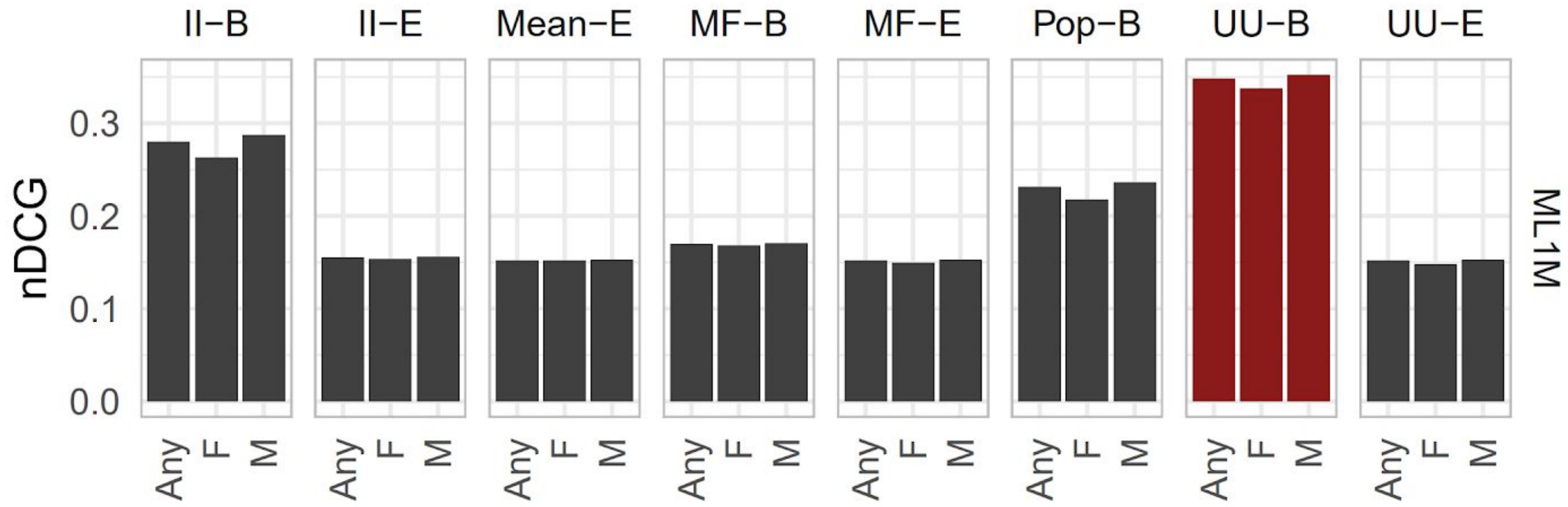
(b) comScore

R Mehrotra, A Anderson, F Diaz, A Sharma, H Wallach, E Yilmaz. Auditing search engines for differential satisfaction across demographics. 2017

Recommendation Accuracy

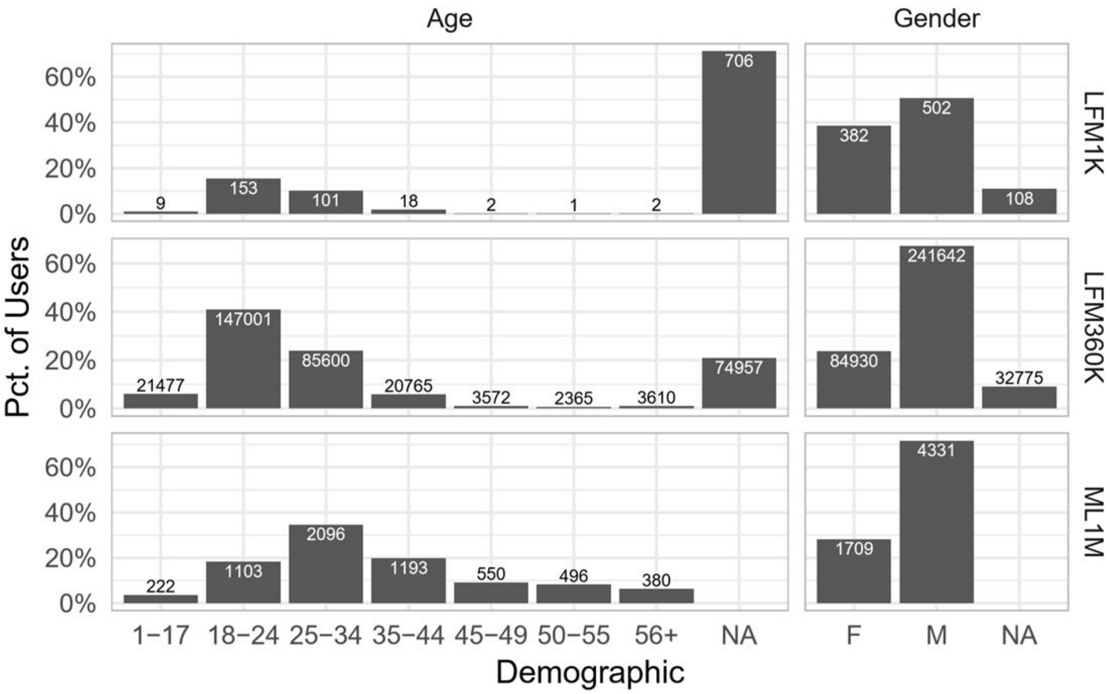
Stratify offline recommender evaluation by demographic

g



M Ekstrand, M Tian, I Madrazo Azpiazu, J Ekstrand, O Anuyah, D McNeill, M S Pera. All The Cool Kids, How Do They Fit In: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. 2018

Recommendation Data



M Ekstrand, M Tian, I Madrazo Azpiazu, J Ekstrand, O Anuyah, D McNeill, M S Pera. All The Cool Kids, How Do They Fit In: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. 2018

Significant Differences

Some groups have better performance

- Men in MovieLens, women in Last.FM 1K
- Young & old in Last.FM 360K

Not simple ‘biggest group ⇒ most benefit’ story

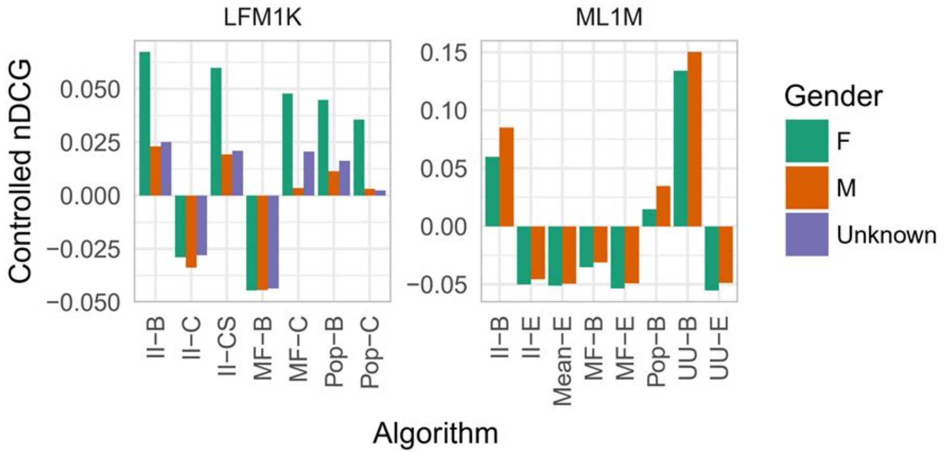
- MovieLens differences correlate with # of users
- Last.FM differences anti-correlate

M Ekstrand, M Tian, I Madrazo Azpiazu, J Ekstrand, O Anuyah, D McNeill, M S Pera. All The Cool Kids, How Do They Fit In: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. 2018

Controls

What drives this?

- Profile size? Controlled with linear model
  - Differences persist
- Number of users? Resampled data set
  - Differences drop below significance



M Ekstrand, M Tian, I Madrazo Azpiazu, J Ekstrand, O Anuyah, D McNeill, M S Pera. All The Cool Kids, How Do They Fit In: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. 2018

Confounds and Limitations

Popularity bias - U1R correction (Bellogin) scrambles age differences

Profile size - negative correlation with accuracy

- Suspect larger profiles had already rated more ‘easy’ recs

No examination of **result character**

M Ekstrand, M Tian, I Madrazo Azpiazu, J Ekstrand, O Anuyah, D McNeill, M S Pera. All The Cool Kids, How Do They Fit In: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. 2018

Collaborative Filtering Parity

Goal: equal **predictive accuracy**

Metric: difference in rating prediction error between advantaged & disadvantaged group; four types:

- Signed value
- Absolute value
- Underestimation
- Overestimation

Each admits a regularizer

Insight: allow different results with comparable quality.

S Yao, B Huang. Beyond Parity: Fairness Objectives for Collaborative Filtering. 2017.

Difficulties

Multiple comparisons - we're looking at a lot of differences

Causality

Interaction with other effects, like popularity bias

Getting data, and issues such as gender binarization

Potential mitigations

- Prioritize challenges affecting underserved groups
  - May improve service for everyone!
- Build specialized services to meet users' needs
- Infer user type / need class
  - Ok for some (e.g. kids)
  - Problematic for others

But first, **study the problem**!

Different classes of users or needs have different ethical concerns

What does this mean?

Not all users experience the system in the same way

- Measure! Measure! Measure!
- How does what you see align with business or social goals?

Different concerns bring contradictory pictures

- Correcting for popularity bias ⇒ changed demographic picture
- Which is ‘right’? Need more research!

Delivering: open area of research

questions?

Provider Fairness

CEO

Kay, Matuszek, Munson, Unequal Representation and Gender Stereotypes in Image Search Results for Occupations, 2015

What happens to authors?

Provider fairness

- unfair representation of providers in neutral queries/contexts
- sources
    - **provider composition**: biases in representation of providers
    - **user behavior**: biases in user feedback can affect learned targets
    - **system design**: biases in what data are filtered in/out

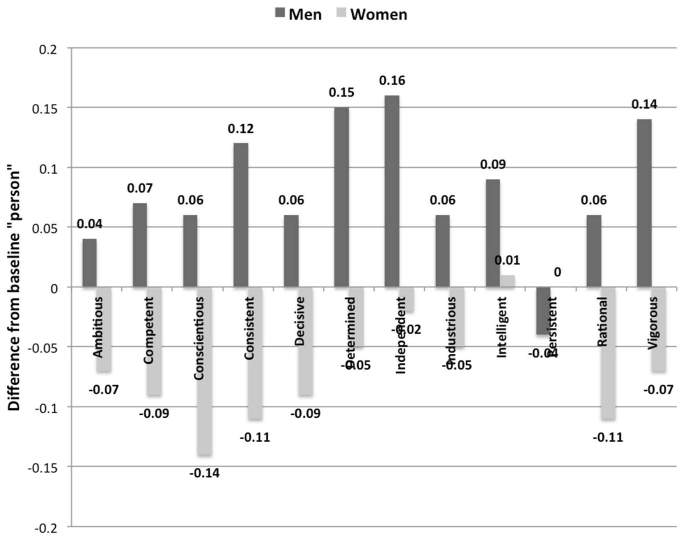


Figure 4: Proportion of agentic images conveying power (difference from respective "person" baseline).

Otterbacher, Bates, Clough. Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results. 2017

scenario	queries	document	relevance	producers
news	keyword	article	topical	authors
academic	keyword	article	topical	authors
citation recommendation	draft	article	topical	authors
medical academic	keyword	article	topical	research subjects
book search	keyword	book	topical	authors
book recommendation	keyword	book	topical	authors
employment	job description	candidate	skill	candidates
community QA	question	answer	topical	answerers
music recommendation	keysong	track	entertainment	musicians
movie recommendation	keyfilm	movie	entertainment	directors

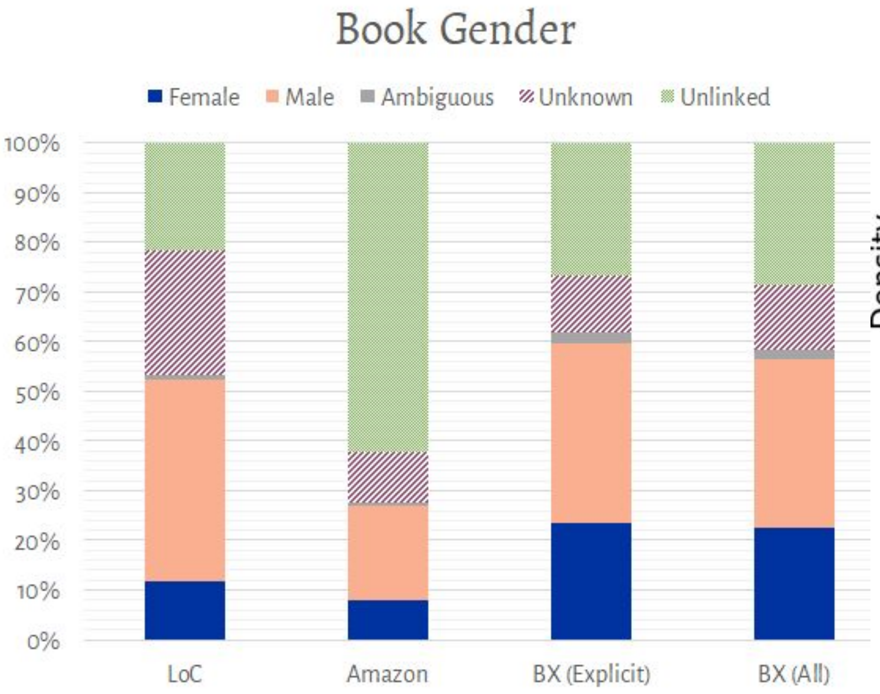
Calibration

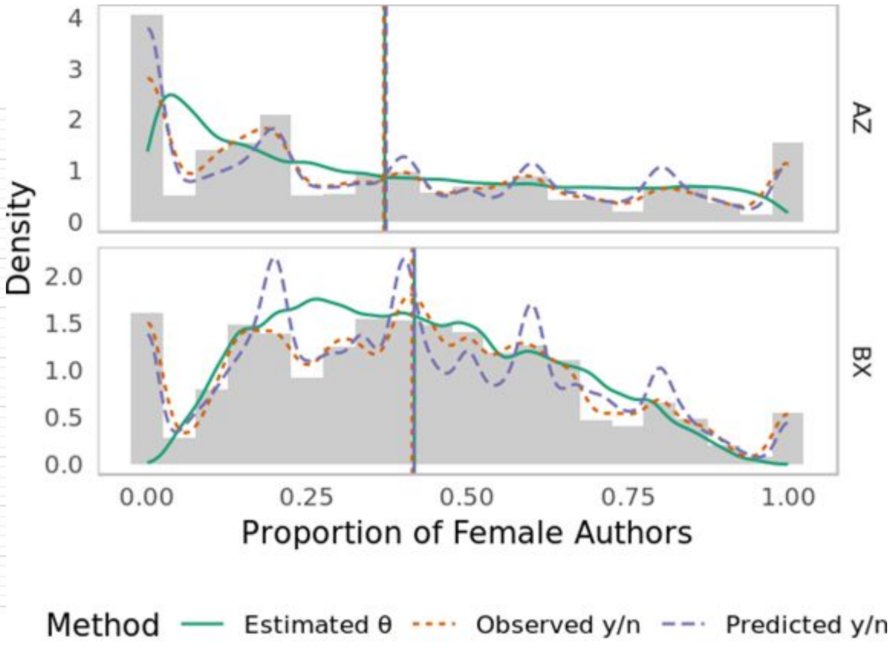
Results are fair if they achieve **fair representation**.

- Results are *evenly balanced*?
- Results *reflect population*?
- Results *reflect user historical data*?

H Steck. Calibrated recommendations. 2018
M Ekstrand, M Tian, M Kazi, H Mehrpouyan, D Kluver. Exploring author gender in book rating and recommendation. 2018

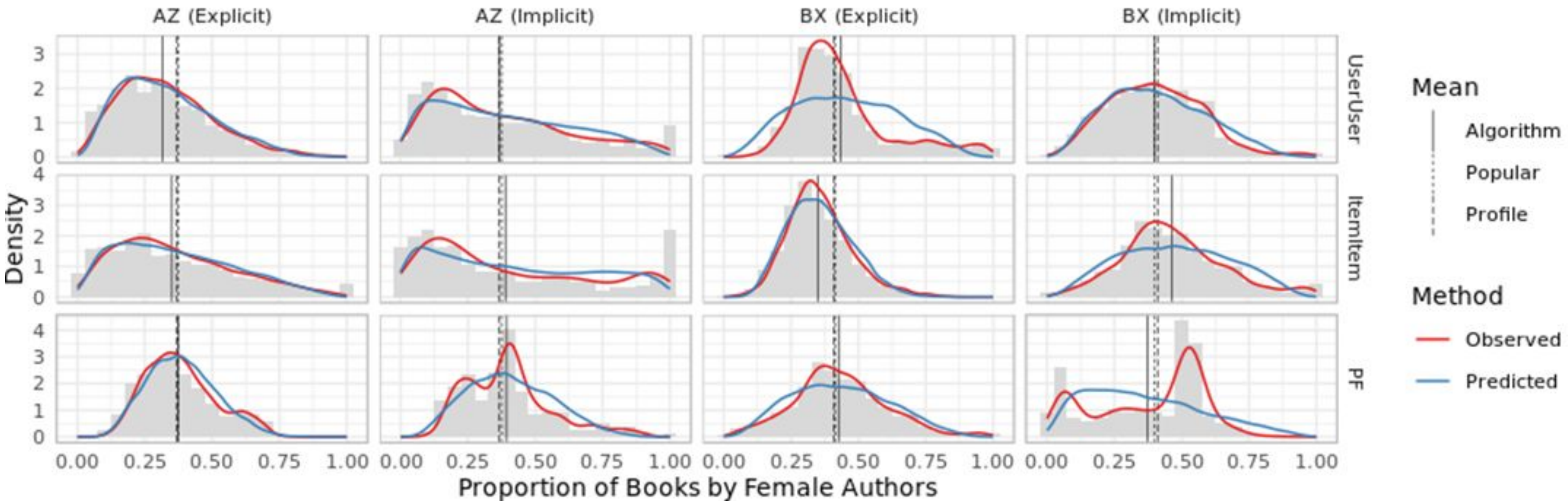
Book Gender - Ratings



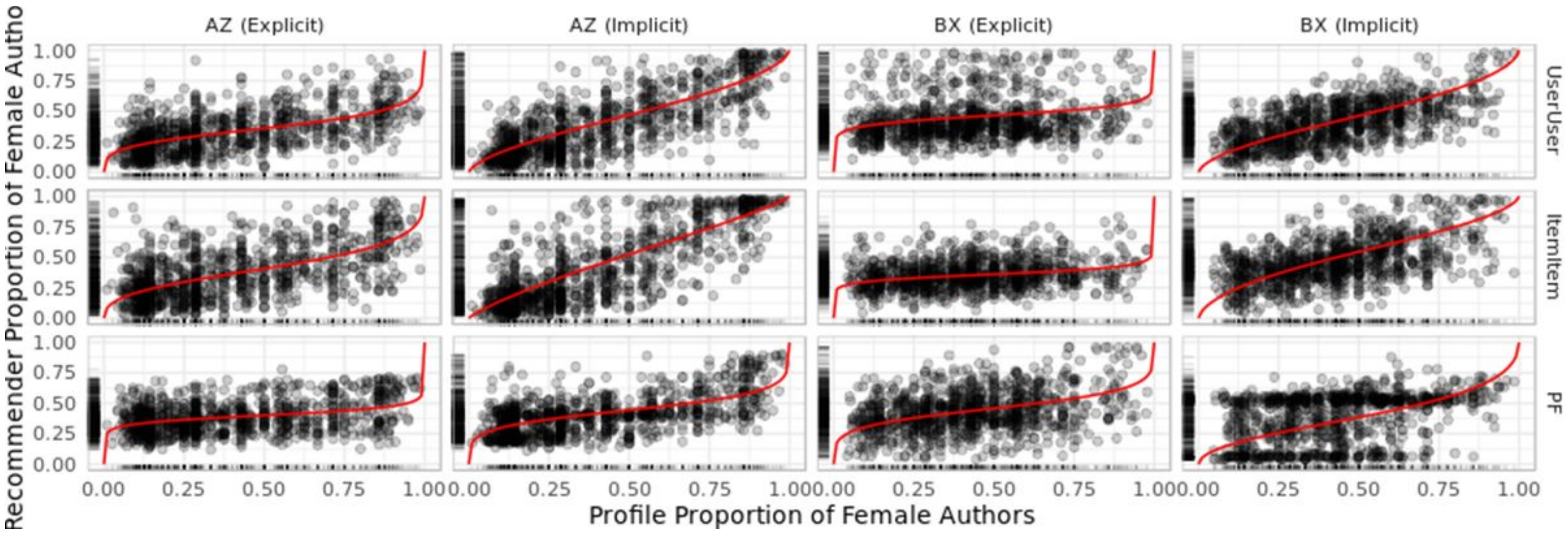


M Ekstrand, M Tian, M Kazi, H Mehrpouyan, D Kluver. Exploring author gender in book rating and recommendation. 2018

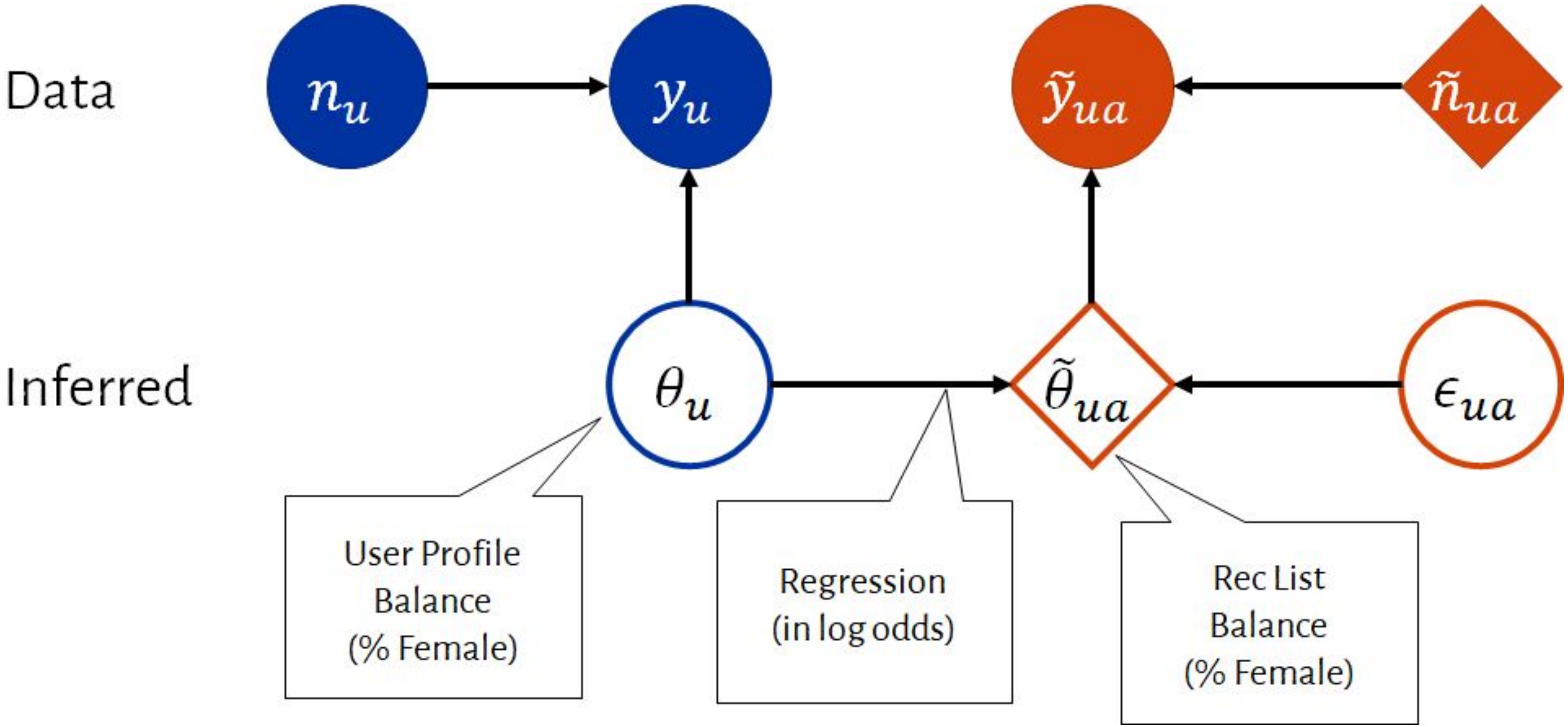
Book Gender - Recommendations



Book Gender - Propagation



Modeling User Calibration



Fairness for Probabilistic Models

Results are fair if,

$$P(R|d) = P(R|\tilde{d})$$

$\tilde{d}$ document d without sensitive attributes

Achieving this:

- Regularization term penalizing non-independence
- Extend recommendation model to incorporate sensitive attribute

T Kamishima, S Akaho, H Asoh, J Sakuma. Recommendation Independence. 2018.

$L_{1/2}$-Fairness

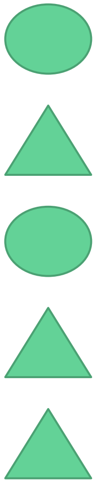
$$\ell_{\frac{1}{2}}\text{-fairness} = \left(\sum_{a \in \mathcal{A}} \sqrt{P(a|\pi_{\leq k})}\right)^2$$

$\pi_{\leq k}$ top k elements in π

A fair ranking has good representation for different groups *a*.

Mehrotra, McInerney, Bouchard, Lalmas, Diaz, Towards a Fair Marketplace, 2018.

Population-Sensitive Ranking Fairness



Is it fair?

Next metrics: a ranking is **fair** if its **composition** reflects the **population**

- Population 35% female => rankings 35% female

What is population?
How do you count proportion in rankings?

Rank-based fairness measures

Assume a binary protected attribute:

Population: full ranking turned into a set

$$\mathrm{rND}(\pi) = \frac{1}{Z} \sum_{i=1}^{|\mathcal{D}|} \delta_i \left| P(a|\pi_{\leq i}) - P(a|\mathcal{D}) \right|$$

Counting: average composition of ranking prefixes

$$\mathrm{rKL}(\pi) = \frac{1}{Z} \sum_{i=1}^{|\mathcal{D}|} \delta_i D_{\mathrm{KL}}(P(A|\pi_{\leq i}) || P(A|\mathcal{D})) \quad \delta_i = \frac{1}{\log_2 i}$$

If whole list is 50% women, first 10 should be 50% women

$$\mathrm{rRD}(\pi) = \frac{1}{Z} \sum_{i=1}^{|\mathcal{D}|} \delta_i \left| \frac{P(a|\pi_{\leq i})}{P(\overline{a}|\pi_{\leq i})} - \frac{P(a|\mathcal{D})}{P(\overline{a}|\mathcal{D})} \right|$$

Yang, Stoyanovich. Measuring fairness in ranked outputs. 2017.

Rank-Aware Calibration

Population: generalized *population estimator*

Counting: probability of picking a group member going down the list, discounted

$$\text{fairness}_\lambda = \phi\left(\sum_{i=1}^{|\mathcal{D}|} \delta_i^\lambda P(A|\pi_i), P(A|\mathcal{D})\right)$$

δ_i^λ parameterized rank discount

P Sapiezynski, W Zeng, R Robertson, A Mislove, C Wilson. Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists. 2019.

Pairwise Fairness

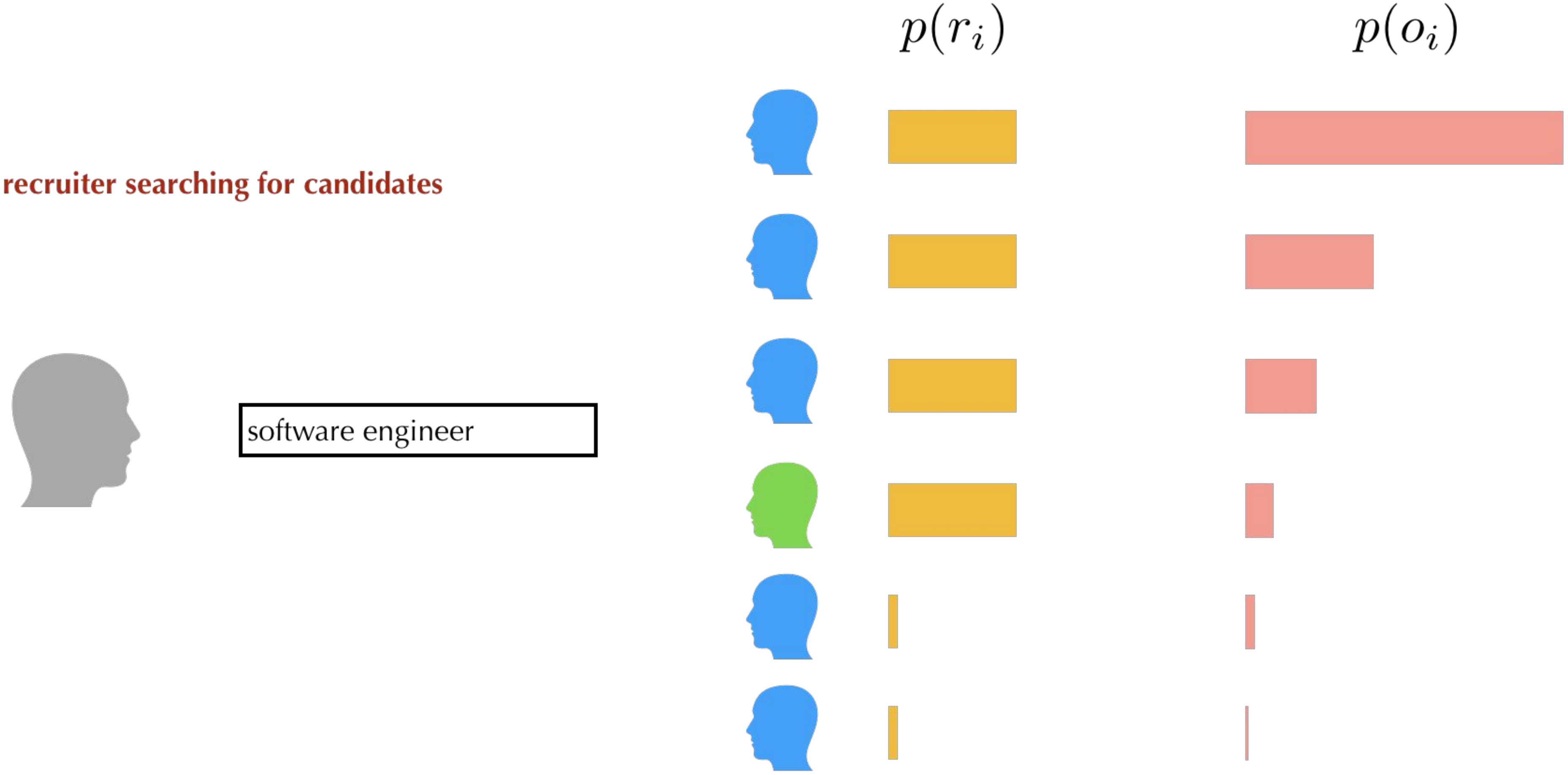
$$P(d \succ d' | f^*(d) > f^*(d'), A_d = a) = P(d \succ d' | f^*(d) > f^*(d'), A_d = \bar{a}) \quad \text{pairwise fairness}$$

$$P(d \succ d' | f^*(d) > f^*(d'), A_d = A_{d'} = a) = P(d \succ d' | f^*(d) > f^*(d'), A_d = A_{d'} = \bar{a}) \quad \text{intra-group pairwise fairness}$$

$$P(d \succ d' | f^*(d) > f^*(d'), A_d = a, A_{d'} = \bar{a}) = P(d \succ d' | f^*(d) > f^*(d'), A_d = \bar{a}, A_{d'} = a) \quad \text{inter-group pairwise fairness}$$

A ranking is fair if probability of correct ranking (relevant over irrelevant) is independent of protected class.

Beutel, Chen, Doshi, Qian, Wei, Wu, Heldt, Zhao, Hong, Chi, Goodrow. Fairness in Recommendation Ranking through Pairwise Comparisons. 2019.



Biega, Gummadi, Weikum, Equity of Attention: Amortizing Individual Fairness in Rankings, 2018.

$$\mathcal{A}_i = \sum_{q \in \mathcal{Q}} a_i^q$$

accumulated **attention** for user i

over a sequence of queries $\mathcal{Q}$

$$\mathcal{R}_i = \sum_{q \in \mathcal{Q}} r_i^q$$

accumulated **relevance** for user i

over a sequence of queries $\mathcal{Q}$

Equity of Amortized Attention

$$\frac{\mathcal{A}_i}{\mathcal{R}_i} = \frac{\mathcal{A}_j}{\mathcal{R}_j}, \forall i, j$$

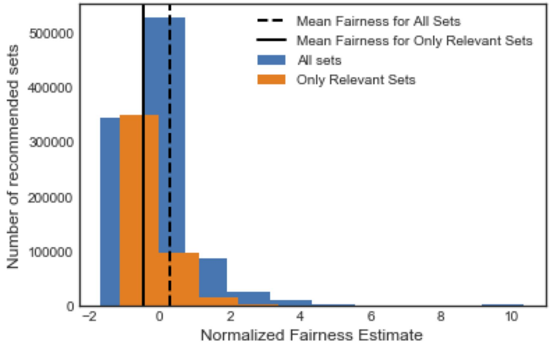
Biega, Gummadi, Weikum, Equity of Attention: Amortizing Individual Fairness in Rankings, 2018.

Relationship to Diversity in Information Retrieval

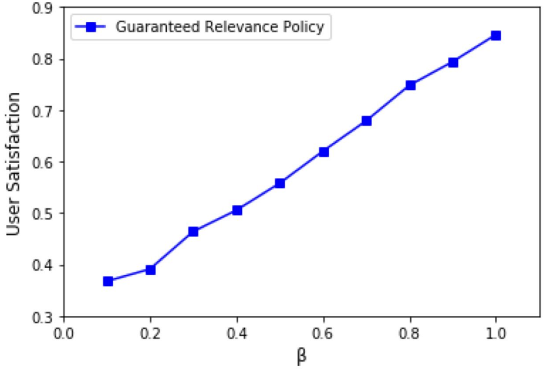
- diversity in information retrieval
  - topic composed of multiple subtopics
  - document can be composed of zero or more subtopics
  - measures promote exposure of many subtopics early in ranking
- fairness in information access
  - producer population composed of multiple intersecting attributes
  - document (usually) associated with one producer
  - measures promote exposure of many intersecting subgroups early in ranking

Combining Fairness with Effectiveness

- fairness metrics do not include effectiveness information.



- fairness and effectiveness trade off.



Mehrotra, McInerney, Bouchard, Lalmas, Diaz, Towards a Fair Marketplace, 2018.

Linear Interpolation

$$\mu^*(\pi) = \beta\mu_{\mathrm{rel}}(\pi) - (1-\beta)\mu_{\mathrm{fairness}}(\pi)$$

Mehrotra, McInerney, Bouchard, Lalmas, Diaz, Towards a Fair Marketplace, 2018.

Fairness Maximal Marginal Relevance (FMMR)

$\mathcal{A}$ set of protected attribute values

$\phi_a(d, d')$ similarity between two documents based on protected attribute value a of providers of d and d'

Karako, Manggala. Using Image Fairness Representations in Diversity-Based Re-ranking for Recommendations. 2018.

Fairness Maximal Marginal Relevance (FMMR)

relevance redundancy

$$\pi_k = \mathrm{argmax}_{d \in \mathcal{D} - \pi_{<k}} \lambda f(d) - (1 - \lambda) \max_{d' \in \pi_{<k}} \sum_{a \in \mathcal{A}} \phi_a(d, d')$$

1. Star Wars
2. Frozen
3. Iron Man
4. *Star Wars IV?*
   Avengers?
   La La Land?

f ranking function

Direct application of diversity concepts!

Karako, Manggala. Using Image Fairness Representations in Diversity-Based Re-ranking for Recommendations. 2018.

Challenges in Fair Ranking

- Joint optimization of consumer and producer
- Non-uniform consumer tolerance to diversity
- Optimization with competing or adversarial services

Mehrotra, McInerney, Bouchard, Lalmas, Diaz, Towards a Fair Marketplace, 2018.

questions?

Feedback Loops

Runaway Feedback Loops

Feedback loops amplify small differences

Be careful with:

- Relevance feedback
- Collaborative filtering inputs
- **Learning from click data**

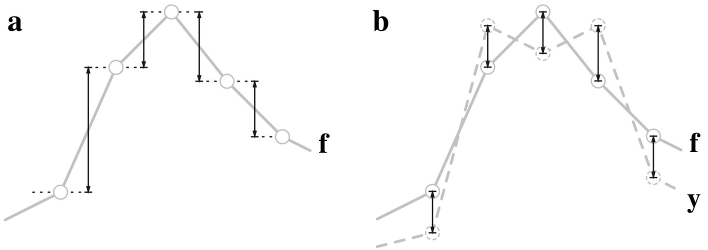
If D1 is a little more relevant than D2, should it receive a lot more exposure?

What if D2 is by an underrepresented author?

Ensign, Friedler, Neville, Scheidegger, Venkatasubramanian. Runaway Feedback Loops in Predictive Policing. 2018.

Iterative Prediction and Fairness

- recommendation systems, especially those based on ML, **increase the consistency** in recommendations across different users.
- how does this consistency between users change over multiple iterations compared with prediction error?



inter-user consistency

user error

Chaney, Stewart, Engelhardt. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. 2018.

Iterative Prediction and Fairness

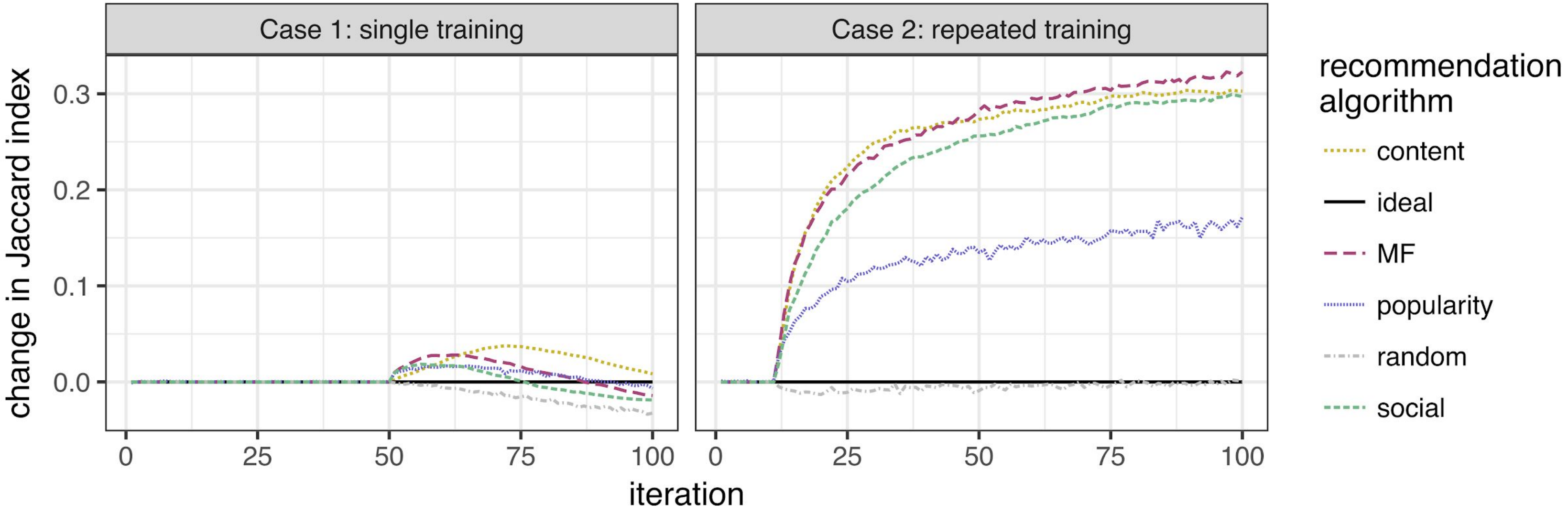
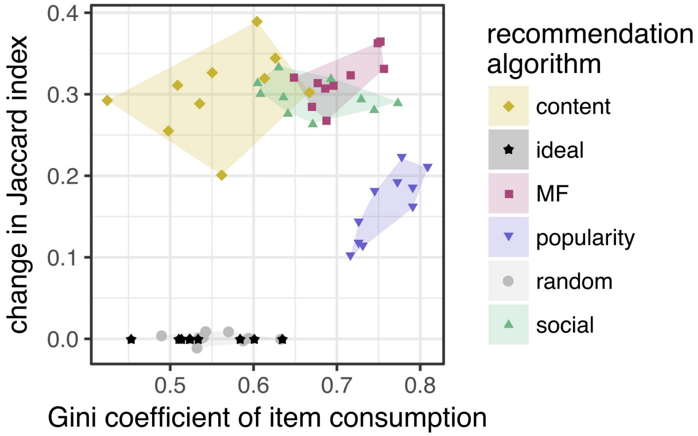


Figure 3: Change in Jaccard index of user behavior relative to ideal behavior; users paired by cosine similarity of θ. On the left, mild homogenization of behavior occurs soon after a single training, but then diminishes. On the right, recommendation systems that include repeated training homogenize user behavior more than is needed for ideal utility.

Chaney, Stewart, Engelhardt. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. 2018.

Iterative Prediction and Fairness

- does consistency uniformly impact all items in the corpus?
- Gini coefficient: measures inequity of exposure.
- some algorithms increase user-user consistency relative to optimal and increase inequity relative to optimal.



Chaney, Stewart, Engelhardt. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. 2018.

Iterative Prediction and User Churn

- When the protected group labels are latent, we cannot monitor fairness.
- Even initially-fair models can converge to unfair models.
- How bad is the situation if we assume that under-performance leads to user churn?

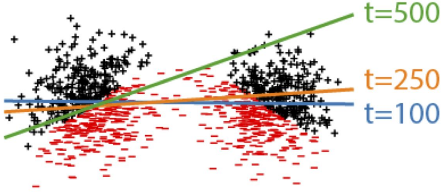
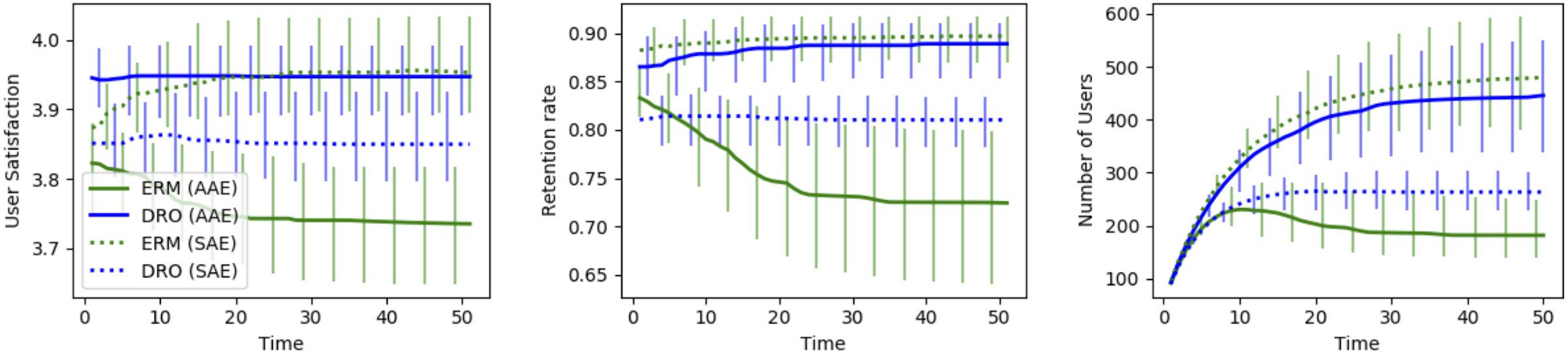


Figure 1. An example online classification problem which begins fair, but becomes unfair over time.

Hashimoto, Srivastava, Namkoong, Liang. Fairness Without Demographics in Repeated Loss Minimization. 2018.

Iterative Prediction and User Churn



(a) User satisfaction (b) User retention (c) User count

ERM: Expected Risk Minimization (standard learning approach)
DRO: Distributionally Robust Optimization

AAE: African American English (under-represented group)
SAE: Standard American English (over-represented group)

Hashimoto, Srivastava, Namkoong, Liang. Fairness Without Demographics in Repeated Loss Minimization. 2018.

Challenges in Feedback Loops

- Temporal reasoning/delayed reward
- Modeling and understanding the consumers/world
- Two-sided feedback loops
- Feedback loop dependence on number of substitutable services

Pragmatics: Data for Studying Fairness

The Problem

We want to study distribution of opportunity, quality, etc. by sensitive attribute

We have lots of data sets… most of which don't have sensitive attributes

For *much* more, see *Limits of Social Data* tutorial and paper below.

- http://www.aolteanu.com/SocialDataLimitsTutorial/

A Olteanu, C Castillo, F Diaz, E Kıcıman. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. 2019

Consumer fairness data

RecSys rating data with user demographics:

- MovieLens (100K and 1M)
- Last.FM collected by Celma

Infer other fairness-relevant characteristics, e.g.:

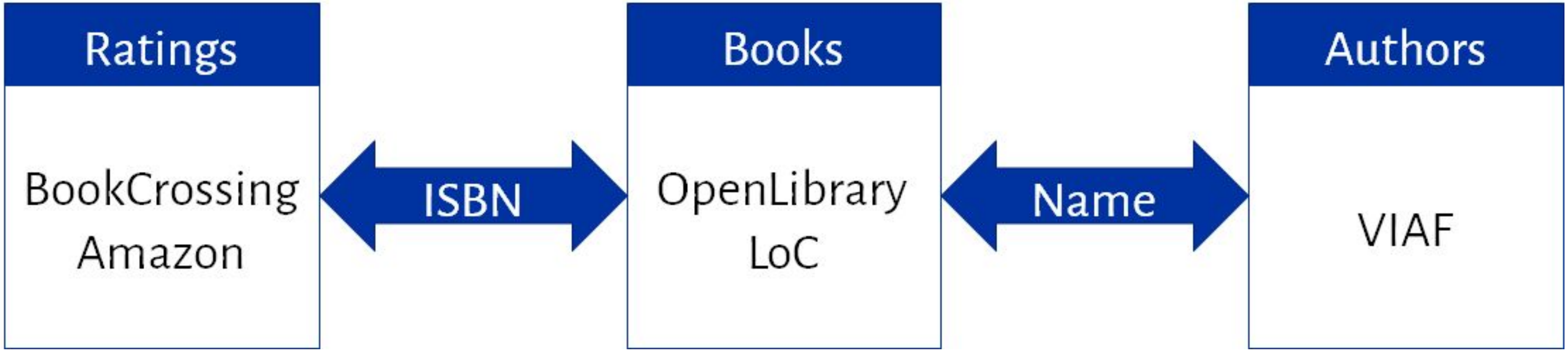
- Position in taste space
- Account age / activity level

Probably don't try to infer demographics (gender, race)

Producer fairness data

Easier, because producers tend to be more public than consumers

Example Pipeline



Producer fairness data

Easier, because producers tend to be more public than consumers

Pitfalls:

- Imprecise data linking
- Problematic operationalizations (e.g. VIAF enforcing binary gender)

Sources:

- Books: Library of Congress
- Scholars: mine open corpus data (for some characteristics)

Be careful *distributing*

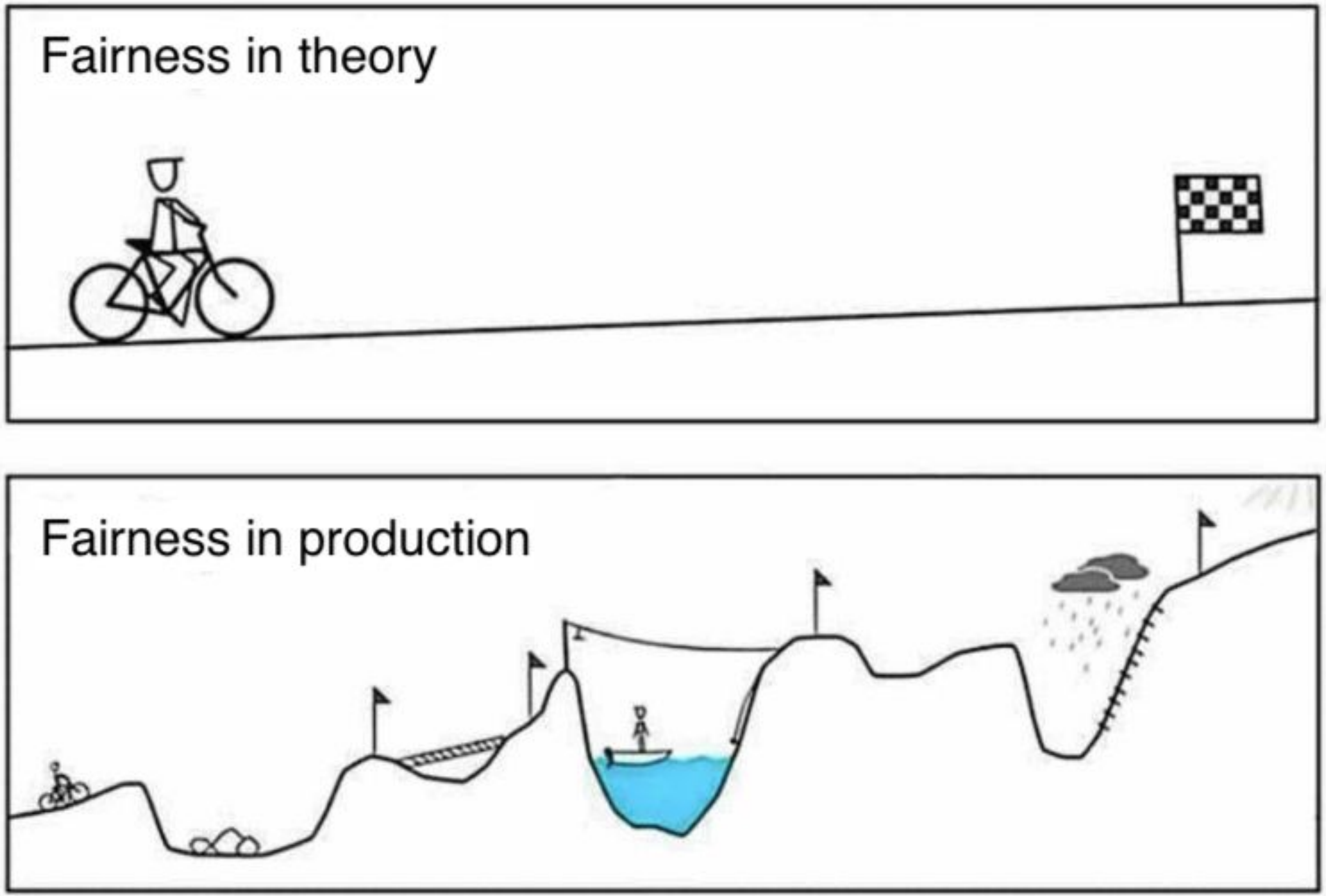
More Challenges

- Public data is hard to find
  - How was it defined + assembled?
- Inference is deeply problematic
  - Reinforces stereotypes
  - Inaccurate in biased ways
  - Program for Cooperative Cataloging specifically prohibits assuming gender from pictures or names
- Reasonably accurate data may not be distributable
  - Making target lists easy to find - increase risk
  - Propagates errors on an individual level

A Olteanu, C Castillo, F Diaz, E Kıcıman. Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. 2019
Report of the PCC Ad Hoc Task Group on Gender in Name Authority Records. 2016
A L Hoffmann. Data Violence and How Bad Engineering Choices Can Damage Society. 2018

questions?

Fairness in Production

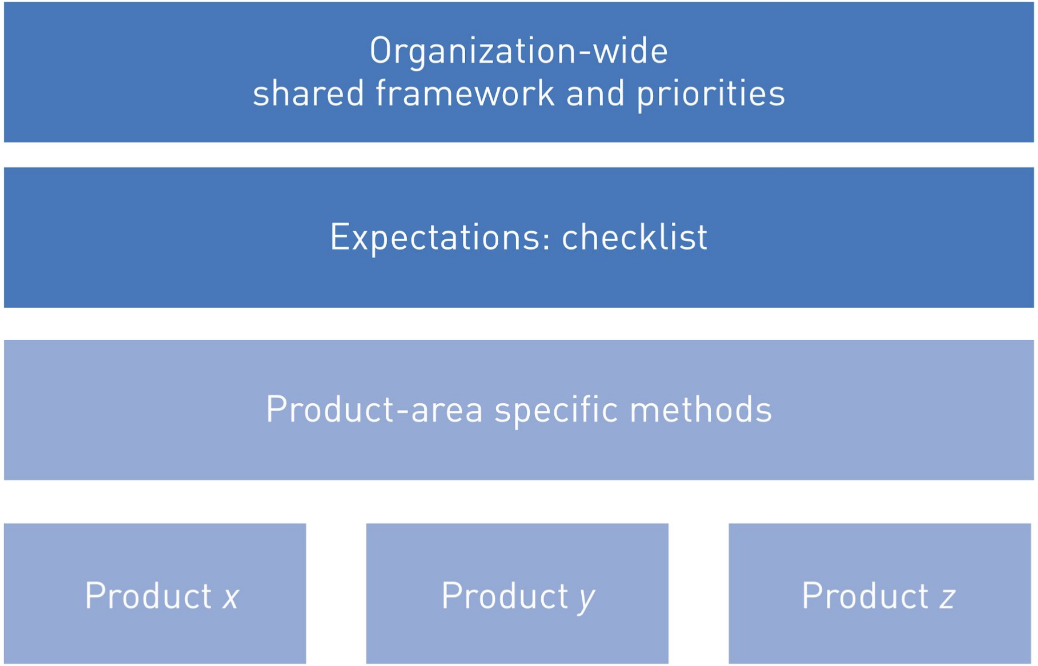

Fairness in theory
Fairness in production

Fairness in Production

- **Data collection**: training data often biased and causes downstream fairness issues.
- **Blind spots**: sensitive group definition poorly-understood/absent
- **Audit protocol**: current approaches *reactive* to user complaints
- **Audit scale**: current approaches atomistic, ignoring system-level fairness
- **Remedies**: current treatments incompatible with production realities
- **Human bias**: values embedded throughout the design process

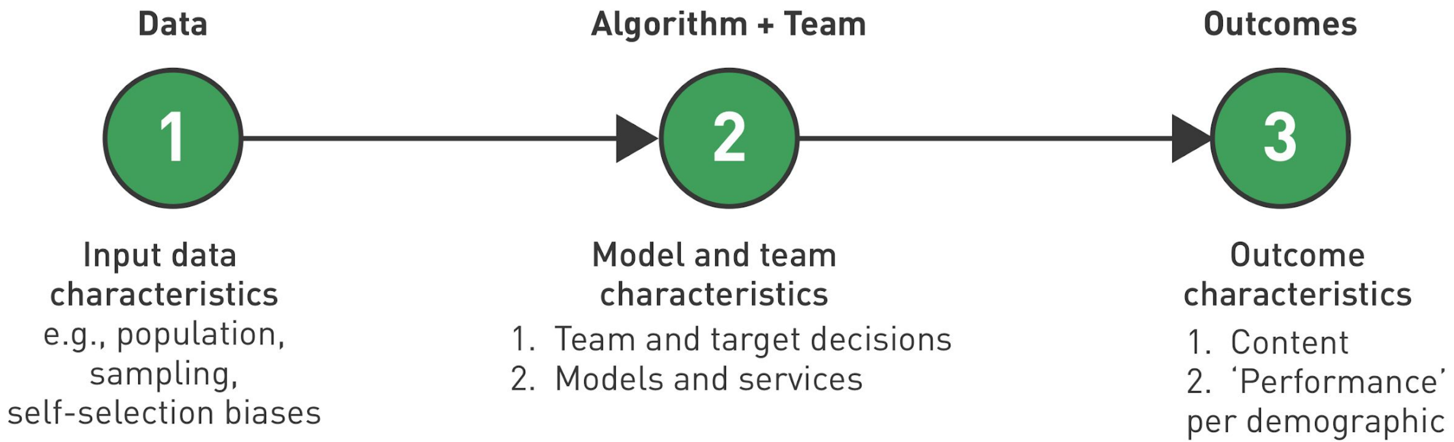
Holstein, Wortman Vaughan, Daumé III, Dudík, Wallach, Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?, 2019.

Fairness in Production



Cramer, Garcia-Garthright, Springer, Reddy. Assessing and Addressing Algorithmic Bias in Practice, 2018.

Fairness in Production



Cramer, Garcia-Garthright, Springer, Reddy. Assessing and Addressing Algorithmic Bias in Practice, 2018.

Tutorial: Challenges of incorporating algorithmic fairness into industry practice

H. Cramer, K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudík, H. Wallach, S. Reddy, J. Garcia-Gathright

https://www.youtube.com/watch?v=UicKZv93SOY

Open Problems

Reference Points for Fairness

What *should* result lists look like?

- What accurately represents the world?
- What accurately represents the world as it could or should be?

Fairness in UX

How do interface & interaction affect fairness outcomes?

- Result presentation
- Feedback / preference elicitation

Most FAT* interface work focused on transparency / explainability

Operationalizing Justice

How do we translate socially-relevant goals into measurable (and optimizable?) properties of information access systems?

- What are the relevant concepts of justice, fairness?
- How do they manifest in information access?
- How do we measure them?

A lot focuses on what we *can* measure.

Accountability

“FAT*” is Fairness, Accountability, and Transparency

What does accountability look like for information access?

- To whom do IA systems & their operators answer?
- Who decides relevant fairness constructs?
- What are mechanisms for seeking redress for violations?

More Resources

From us:

- Slides
- Bibliography

https://fair-ia.ekstrandom.net

Paper in progress

Elsewhere:

- FACTS-IR workshop Thursday
- FATREC workshop ‘17-’18
- Papers in FAT*, RecSys, SIGIR
- TREC track!

Questions?

https://fair-ia.ekstrandom.net

Thanks to:

- Amifa Raj & the People & Information Research Team
- NSF (based in part on work supported by IIS 17-51278)
- Himan Abdollahpouri & Nasim Sonboli & That Recommender Systems Lab