Fair and Socially Responsible ML for Recommendations

Hannah Korevaar, Manish Raghavan, Ashudeep Singh

NeurIPS 2022 Tutorial

About Us

Hannah Korevaar
Research Scientist, Meta

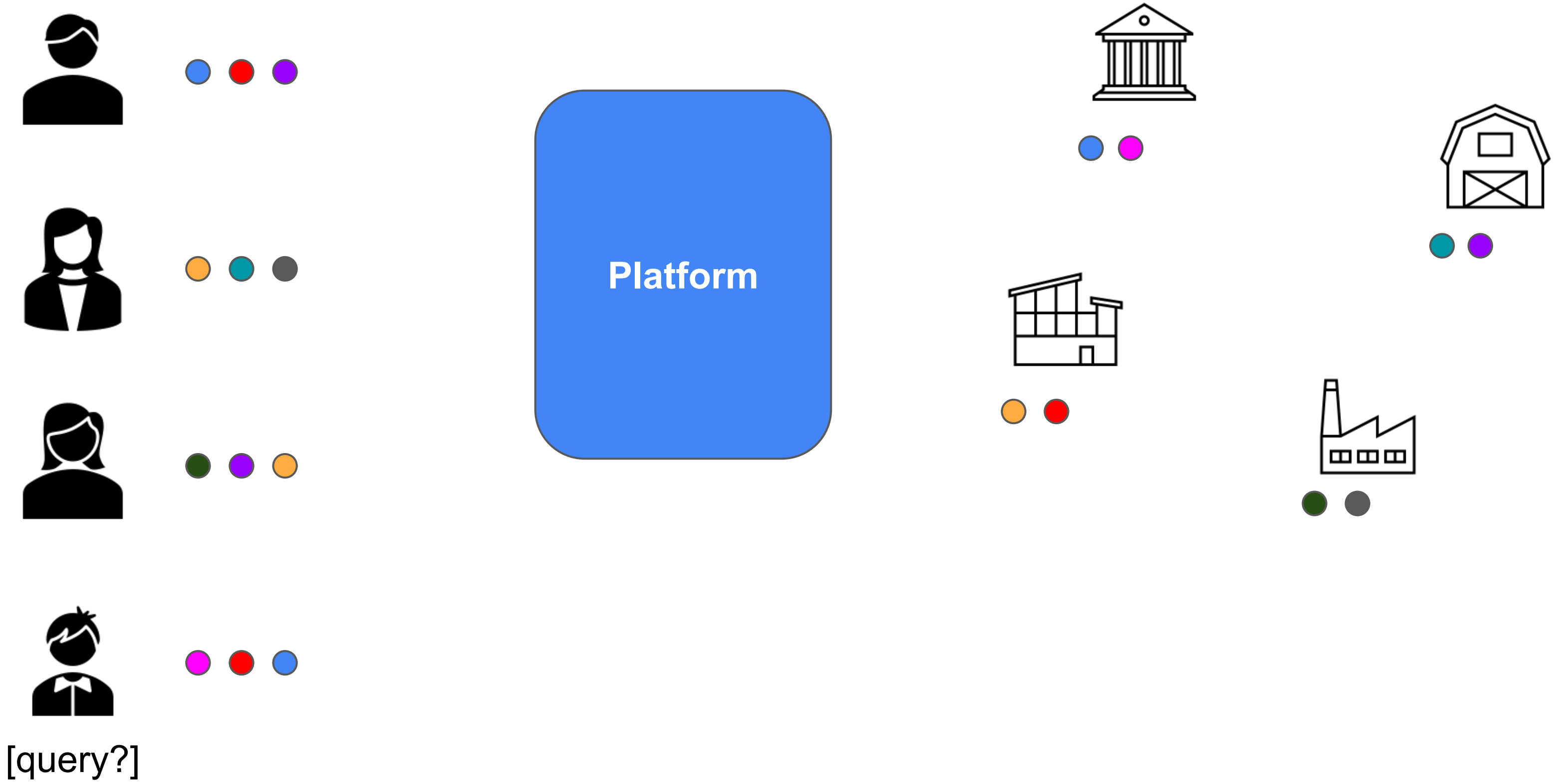
Manish Raghavan
Assistant Professor, MIT

Ashudeep Singh
Applied Scientist, Pinterest

Outline

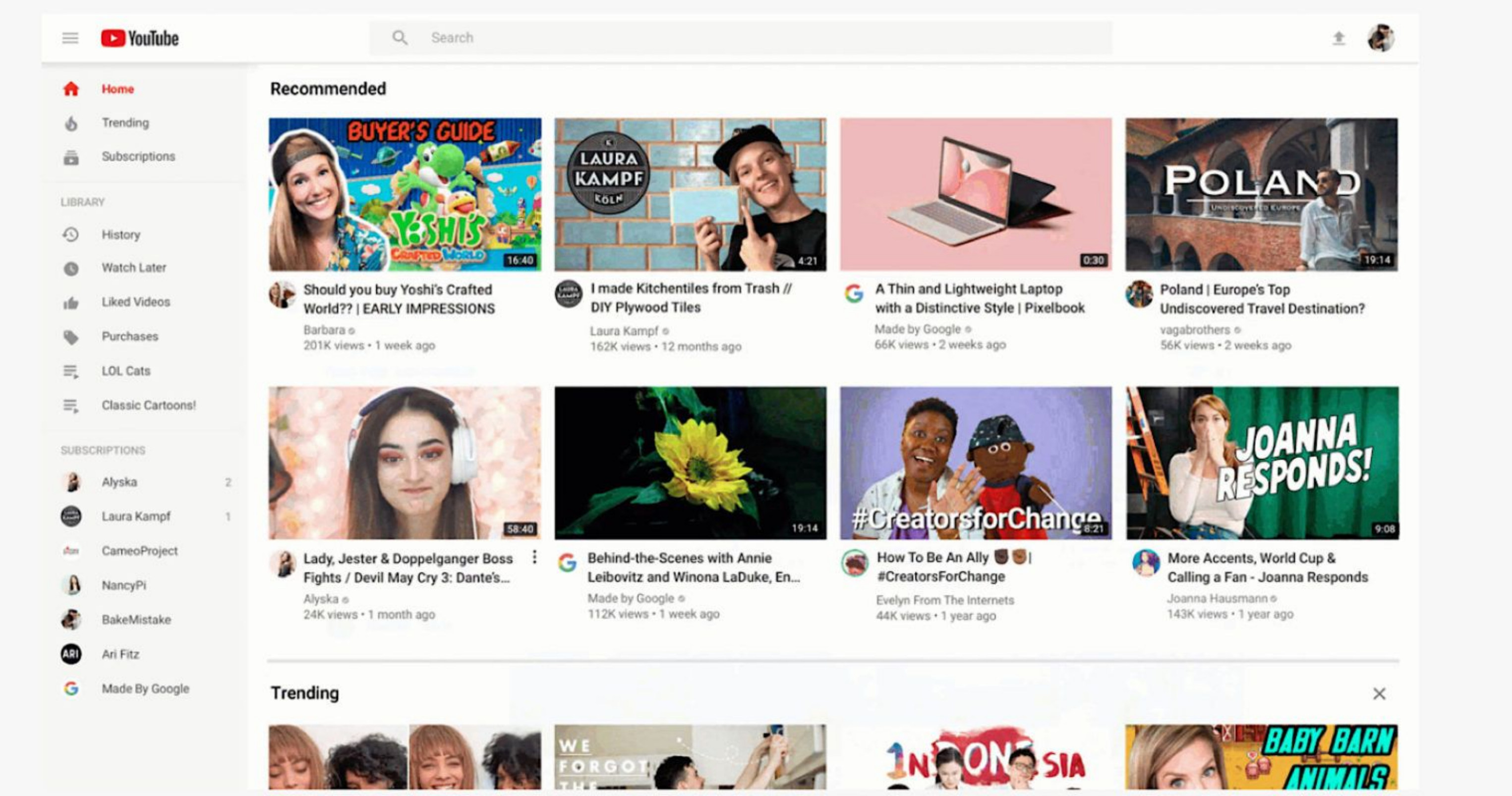
1. Intro to personalized rankings
2. Principles for responsible recommendations
3. Data quality and human behavior
4. Consequences of errors
5. Building & evaluating real-world systems

Personalized rankings

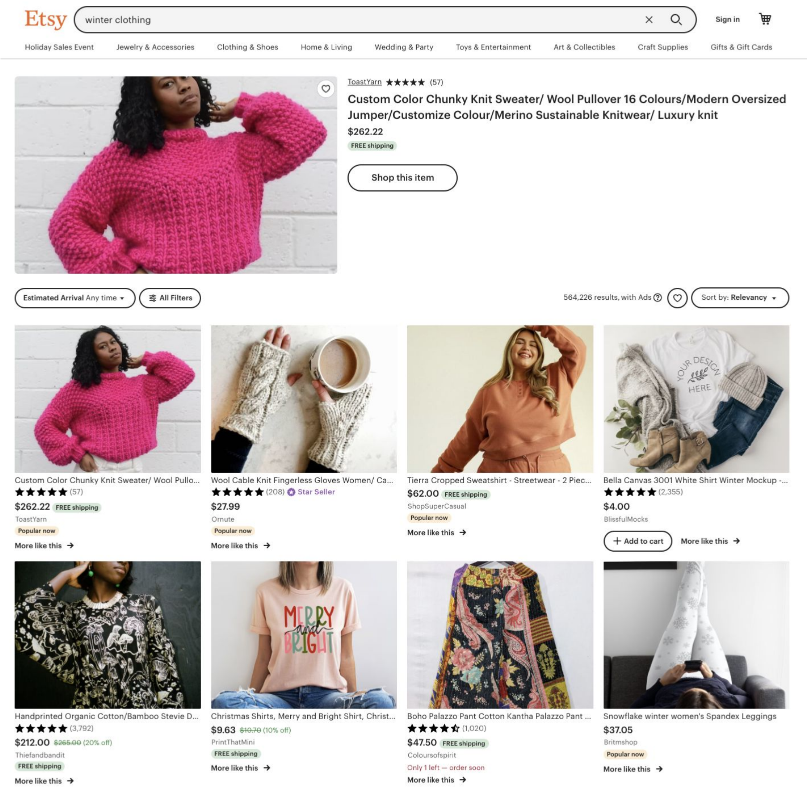


Social media

Entertainment

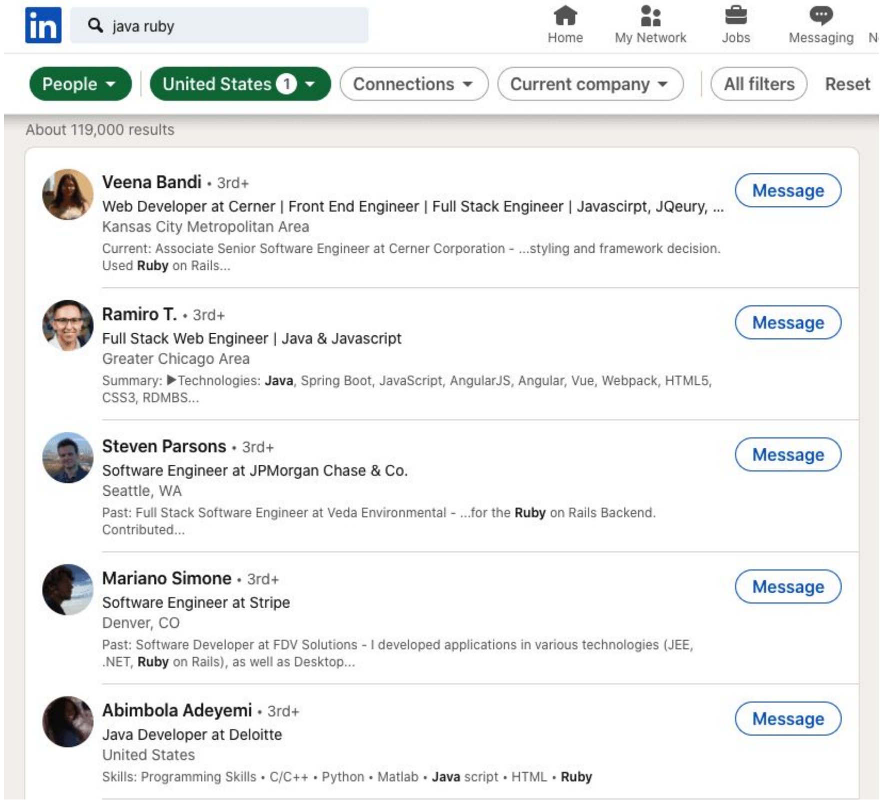


Shopping



Etsy
winter clothing
Sign in
Holiday Sales Event
Jewelry & Accessories
Clothing & Shoes
Home & Living
Wedding & Party
Toys & Entertainment
Art & Collectibles
Craft Supplies
Gifts & Gift Cards
ToastYarn (57)
Custom Color Chunky Knit Sweater/ Wool Pullover 16 Colours/Modern Oversized Jumper/Customize Colour/Merino Sustainable Knitwear/ Luxury knit
$262.22
FREE shipping
Shop this item
Estimated Arrival Any time
All Filters
564,226 results, with Ads
Sort by: Relevancy
Custom Color Chunky Knit Sweater/ Wool Pullo...
(57)
$262.22 FREE shipping
ToastYarn
Popular now
More like this
Wool Cable Knit Fingerless Gloves Women/ Ca...
(208) Star Seller
$27.99
Ornute
Popular now
More like this
Tierra Cropped Sweatshirt - Streetwear - 2 Piec...
$62.00 FREE shipping
ShopSuperCasual
Popular now
More like this
Bella Canvas 3001 White Shirt Winter Mockup -...
(2,355)
$4.00
BlissfulMocks
Add to cart
More like this
Handprinted Organic Cotton/Bamboo Stevie D...
(3,792)
$212.00 $265.00 (20% off)
Thiefandbandit
FREE shipping
More like this
Christmas Shirts, Merry and Bright Shirt, Christ...
$9.63 $10.70 (10% off)
PrintThatMini
FREE shipping
More like this
Boho Palazzo Pant Cotton Kantha Palazzo Pant ...
(1,020)
$47.50 FREE shipping
Coloursofspirit
Only 1 left — order soon
More like this
Snowflake winter women's Spandex Leggings
$37.05
Britmshop
Popular now
More like this

Employment


java ruby
Home
My Network
Jobs
Messaging
People
United States 1
Connections
Current company
All filters
Reset
About 119,000 results
Veena Bandi • 3rd+
Web Developer at Cerner | Front End Engineer | Full Stack Engineer | Javascirpt, JQeury, ...
Kansas City Metropolitan Area
Current: Associate Senior Software Engineer at Cerner Corporation - ...styling and framework decision. Used Ruby on Rails...
Message
Ramiro T. • 3rd+
Full Stack Web Engineer | Java & Javascript
Greater Chicago Area
Summary: ▶Technologies: Java, Spring Boot, JavaScript, AngularJS, Angular, Vue, Webpack, HTML5, CSS3, RDMBS...
Message
Steven Parsons • 3rd+
Software Engineer at JPMorgan Chase & Co.
Seattle, WA
Past: Full Stack Software Engineer at Veda Environmental - ...for the Ruby on Rails Backend. Contributed...
Message
Mariano Simone • 3rd+
Software Engineer at Stripe
Denver, CO
Past: Software Developer at FDV Solutions - I developed applications in various technologies (JEE, .NET, Ruby on Rails), as well as Desktop...
Message
Abimbola Adeyemi • 3rd+
Java Developer at Deloitte
United States
Skills: Programming Skills • C/C++ • Python • Matlab • Java script • HTML • Ruby
Message

A common approach

Predict relevance *r(i, j)* of item *j to* user *i*

For user *i*, show items in descending order of *r(i, j)*

This has been the subject of debate for decades (e.g., Robertson, 1977)

But in practice, it's the still the dominant approach

Key questions

1. How do we measure “relevance”?
    a. Is it single-dimensional? Independent across items?
    b. How do we get good data on it?
2. If we had a good measure of relevance, how should we use it?
    a. What constraints are there?
    b. Is descending-order ranking sufficient?

Challenges

Lots!

- Measuring value is hard
- Inter-item relationships
- Capacity constraints
- Learning from data generated by deployed system (feedback loops)
- Social biases
- Two-sided: consumers & creators
- Utility-maximization vs. fairness
- …

Beyond fairness in ML

“Fair ML” (in particular, group fairness) typically operates in a classification setting:

- You want to predict some outcome Y given inputs X
- You want to do so in a way that is “fair” (by some definition), often across demographic attributes A

This is a rich and nuanced area of research

Some of these ideas are useful here, but miss important features of this setting (e.g., attention, two sided-ness, …)

Principles for responsible ML for recommendations

- Consumers
  - Provide value
  - Respect autonomy
- Creators
  - Provide opportunity
  - Allocate opportunity fairly

Today’s plan

1. Value, preferences, and data
2. Fairness and errors
3. Building and evaluating a real-world system

Today's plan

1. Value, preferences, and data
2. Fairness and errors
3. Building and evaluating a real-world system

Part 1

Value, preferences, and data

“Relevance”

What do we want to measure?

How do we get that data?

Reminder: we’re only talking about **consumers** now. We’ll talk about **producers** in the next parts

Relevance: social media

r(i, j): Will user *i* engage with item *j*?

Engagement: dwell time, watch time, clicks, likes, etc.

Is engagement the (only) goal of the system?

Relevance: entertainment

r(i, j): Will user *i* watch video *j*?

Another goal, perhaps: will user *i* **enjoy** video *j*?

Relevance: shopping

r(i, j): Will user *i* click on item *j*? buy item *j*?

What other goals might a user have? E.g., learn about different products, discover new ones, etc.

Relevance: employment

$r(i, j)$: Will recruiter i (click on | message | hire) person j?

Quality vs. volume of signals

Common theme: picking the right measurement is hard

Often, we have some data lying around ("digital exhaust")

- Clicks
- Browsing data
- Upstream outcomes (e.g., profile views, not hires)
- …

Collecting new data is expensive

Quality vs. quantity

Common trade-off

- Survey data vs. clicks
- Hires vs. profile views
- Ratings vs. movie watching
- …

How do we manage this trade-off?

A basic model:

- Suppose you have two measures *A* and *B* of a quantity *y*
- Both of them measure the same thing, but with different noise σ_A and σ_B
- You have *n* and *m* samples of each measure
- Suppose $\sigma_A < \sigma_B$ and $n < m$
  - *A* is high-quality, low-quantity
  - *B* is low-quality, high-quantity

Quality vs. quantity, quantified

More precisely:

$A = (\Sigma_{i = 1 \ldots n} A_i)/n$

$B = (\Sigma_{i = 1 \ldots m} B_i)/m$

$A_i \sim N(y, \sigma_A)$; $B_i \sim N(y, \sigma_B)$

How do you estimate *y*? **Inverse variance.**

$\hat{y} = (A \cdot n/\sigma_A^2 + B \cdot \mathrm{m}/\sigma_B^2) / (n/\sigma_A^2 + \mathrm{m}/\sigma_B^2)$

Does this solve the problem?

Critical assumption! *A* and *B* measure the same thing: **value**

What if this isn't true?

What does value mean?

(And how do we measure it?)

Measuring value

What do people want?

Do we just need to ask them? What can we learn from existing data?

Are items independent?
(We will largely set this aside for now)

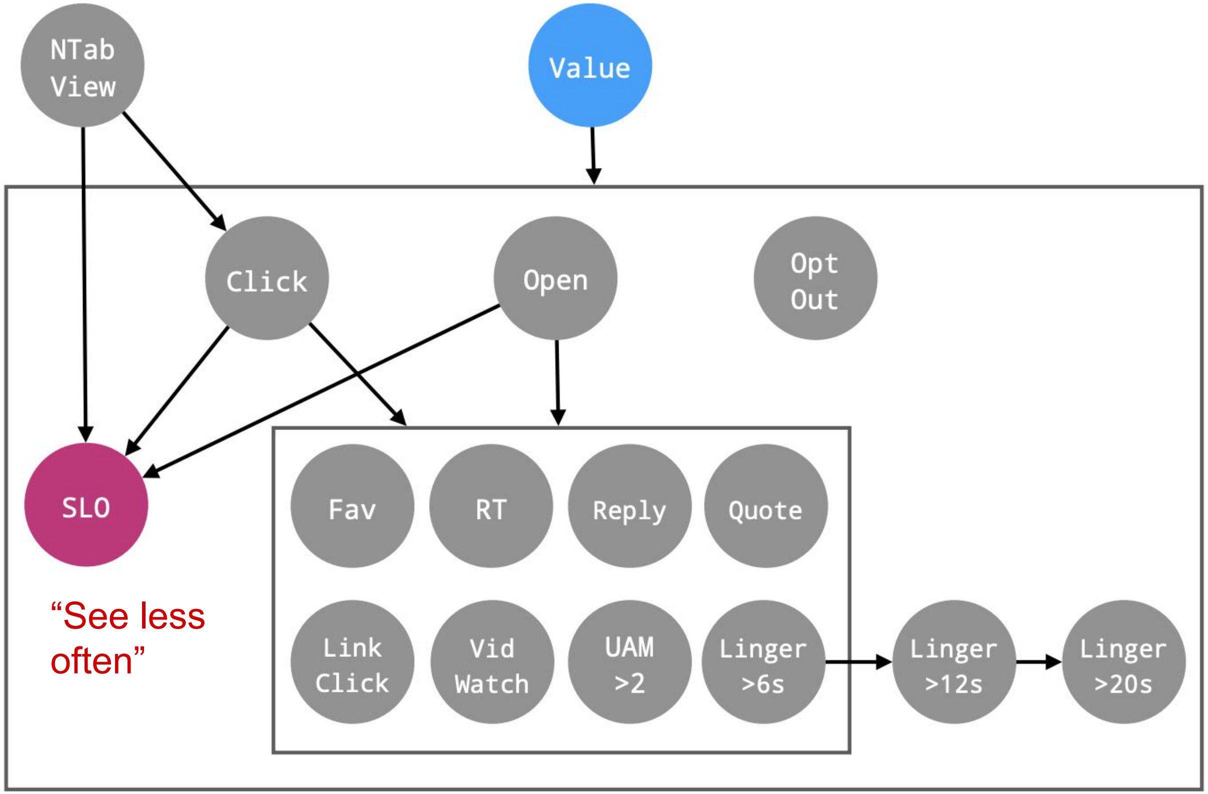
Three perspectives on social media value

- Computational (Milli, Belli, Hardt '21)
- Psychological (Kleinberg, Mullainathan, Raghavan '22)
- Empirical (Agan, Davenport, Ludwig, Mullainathan; forthcoming)

Three perspectives on social media value

- Computational (Milli, Belli, Hardt ‘21)
- Psychological (Kleinberg, Mullainathan, Raghavan ‘22)
- Empirical (Agan, Davenport, Ludwig, Mullainathan; forthcoming)

From Optimizing Engagement to Measuring Value



(Milli, Belli, Hardt '21)

$\mathbb{P}(V = 1 \mid \texttt{Behavior} = 1)$

Behavior	Naive Bayes	Click, Open $\nrightarrow$ SLO	Full Model
OptOut	0	0	0
Click	0	0.316	0.652
Open	0	0.442	0.685
UAM	0	0.157	0.719
VidWatch	0	0.254	0.772
Linger > 6s	0	0.264	0.802
LinkClick	0	0.320	0.836
Reply	0.358	0.570	0.932
Linger > 12s	0	0.245	0.948
Fav	0.579	0.672	0.949
RT	0.680	0.720	0.956
Linger > 20s	0.019	0.296	0.991
Quote	1.0	1.0	1.0

Computational perspective: Inferring value

- Lots of different signals
- Want to know how they relate to “value”
- If you have an “anchor,” you can learn the relationship to other signals
- Note that this is **explicitly** different from our naive model, which said that each signal is a noisy, unbiased measure of “value”

Three perspectives on social media value

- Computational (Milli, Belli, Hardt ‘21)
- Psychological (Kleinberg, Mullainathan, Raghavan ‘22)
- Empirical (Agan, Davenport, Ludwig, Mullainathan; forthcoming)

The Challenge of Understanding What Users Want

Preferences are inconsistent in structured ways (e.g., time-inconsistency)

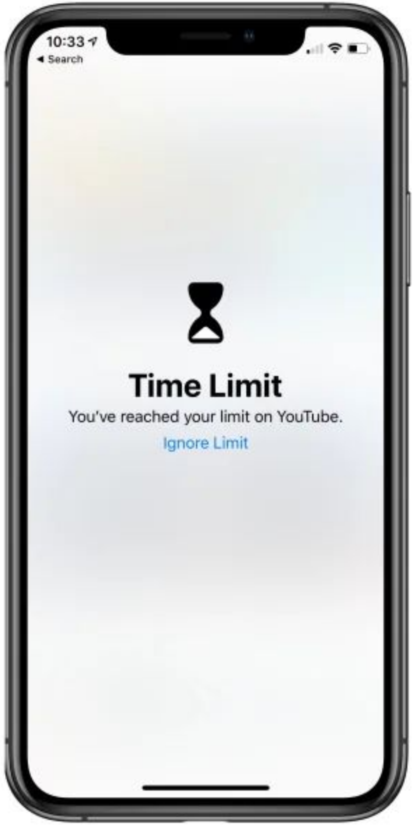
One such structure:

- System 1: fast, impulsive choices
- System 2: slow, deliberative choices

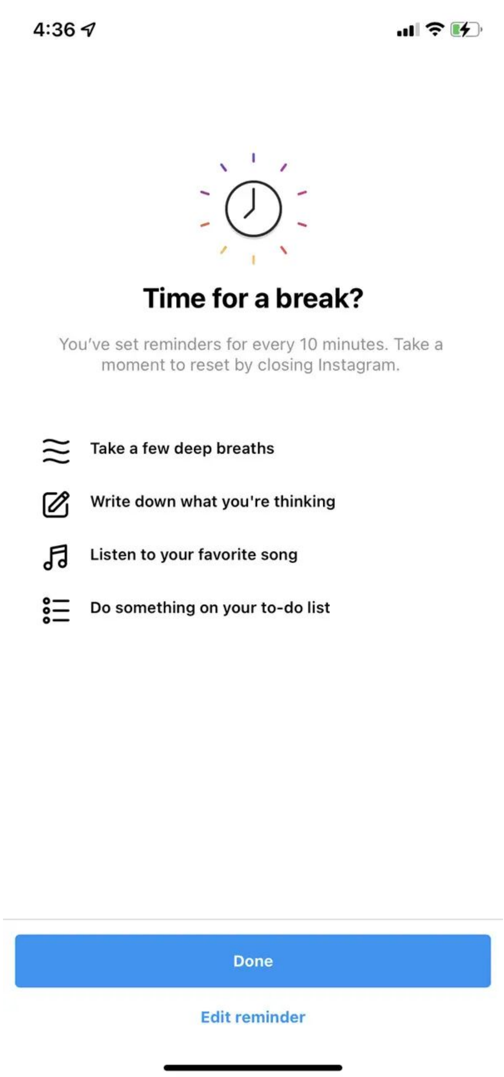
Online behavior reflects a combination of these

Mediated by multiple factors: type of content, platform design, length of session, etc.

(Kleinberg, Mullainathan, Raghavan ‘22)


10:33
◂ Search
Time Limit
You've reached your limit on YouTube.
Ignore Limit


UP NEXT
4:55 / 4:55


4:36
Time for a break?
You've set reminders for every 10 minutes. Take a moment to reset by closing Instagram.
Take a few deep breaths
Write down what you're thinking
Listen to your favorite song
Do something on your to-do list
Done
Edit reminder

Psychological perspective: Impulsivity

- Behavior reflects impulsivity
- Heterogeneous across content
- Influenced by design decisions
- Can we learn what activity is impulsive vs. not?

Three perspectives on social media value

- Computational (Milli, Belli, Hardt ‘21)
- Psychological (Kleinberg, Mullainathan, Raghavan ‘22)
- Empirical (Agan, Davenport, Ludwig, Mullainathan; forthcoming)

Algorithmic Curation Creates Bias

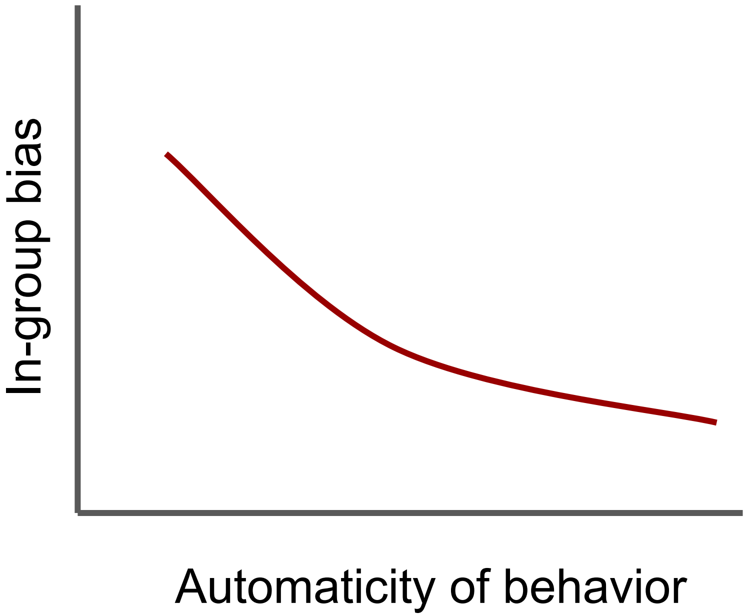
People have in-group bias (e.g., race, ethnicity, religion)

Does this manifest in recommender algorithms?

- Conditioned on explicit preferences, feed algorithm favors in-group
- …but friend suggest algorithm doesn't

Why? **Automaticity**

(Agan, Davenport, Ludwig, Mullainathan; forthcoming)


In-group bias
Automaticity of behavior

Empirical perspective: Automaticity

- Bias increases with automaticity
- Our notion of “value” should reflect this
- The degree to which we trust signals should depend on the automaticity of the underlying actions

The relationship between behavior and value is **structured**

Beyond social media

How should these studies change how we think about:

- Entertainment – can we infer whether people are getting value from binging?
- Shopping – people struggle with impulsivity
- Employment – do more automatic behaviors lead to bias?

Note that this is not just at the objective-choosing level.
It’s at the **algorithmic** level

Behavioral foundations

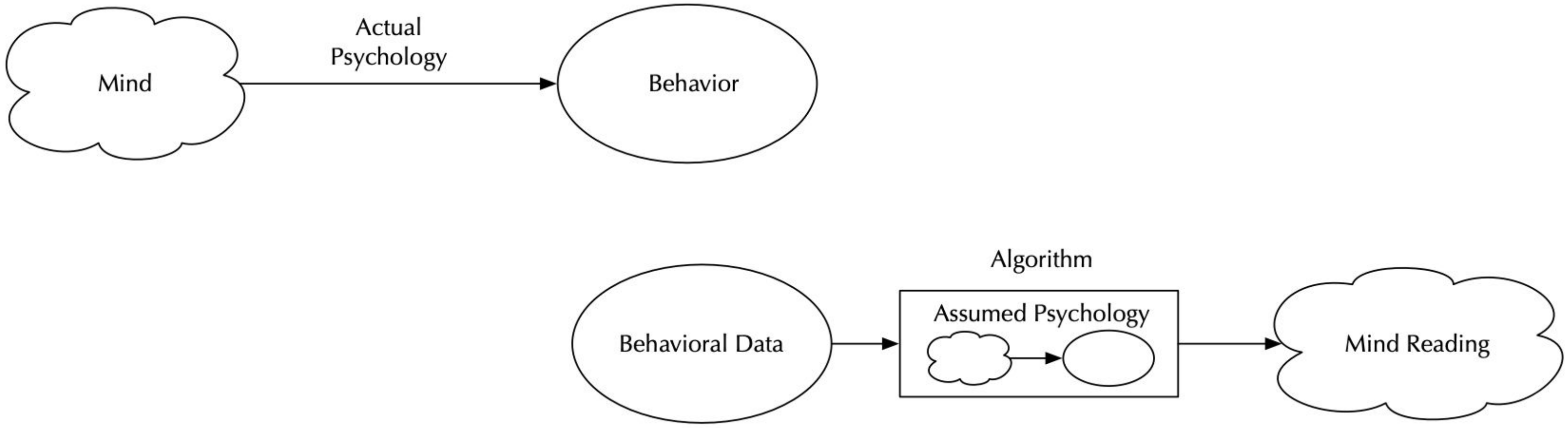
Algorithms learn from data

Data are generated from behavior

→ Algorithms need to account for behavior

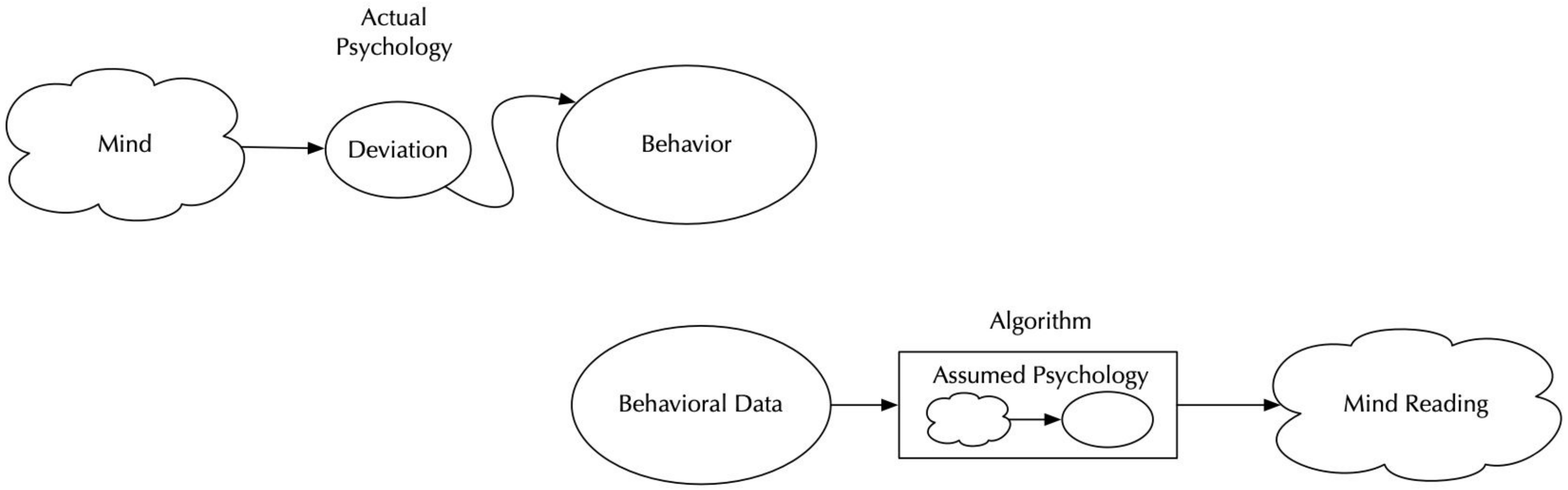
Kleinberg, Ludwig, Mullainathan, Raghavan ‘22

Algorithms invert psychology



Panel A: Algorithm has the right psychology

Algorithms invert psychology



Panel B: Algorithm has the wrong psychology

An example of this in the IR literature: search

An early (wrong) model of search: people pick the best result you show them

A better model: people move down the results list sequentially (e.g., Joachims ‘02)

- Comes from: models of psychology, empirical studies (e.g., Granka et al. ‘04)

This changes the way we design algorithms!

- Structural understanding of what a click **means**
- We design algorithms to **invert** this behavioral model by accounting for position bias

Takeaways: value, preferences, & data

- We often want to provide value, but measuring value is hard
- Data do not always reflect preferences
- …but these differences can manifest in **systematic** ways
- Before we can responsibly allocate attention, we must know what people **value**

Part 2

Fairness and errors

Outline

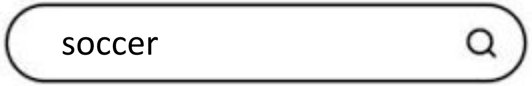
- Fairness
  - Group-level fairness
  - Framework for fairness considerations in AI
  - Classification example
    - Fairness dimensions
    - Evaluation: outcomes
    - Evaluation: models
- Personalized ranking
  - Problem space
  - Optimization framework
  - Measurement challenges
- Evaluation: outcomes
- Evaluation: models

Fairness in classification

Soccer or not soccer?

Fairness in classification

Soccer or not soccer?



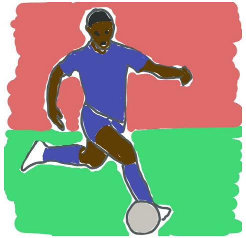
Fairness in classification

Soccer or not soccer?

soccer












Fairness questions

- Product policy
  - What is the product meant to do?
- Labeling policy
  - What are the labeling rules?
- Labels
  - Are they accurate?
  - Are there enough?
- Models
  - Are they accurate?
  - What types of errors do they make?
- Outcomes
  - How representative are the images?

Fairness measurements

- Errors: Assume the system design remains unchanged. Do models or components make errors more frequently for one group (of content/creator/user) over another?
- Design decisions: What impact does including this model, component, target metric etc. have on the representation and value obtained for different groups from the product? These tend to be questions of tradeoffs rather than clear-cut questions of fair or unfair.

Fairness Dimension Examples

	Design Decisions/Tradeoffs
Product policy	Alternative product design/goals; balancing stakeholder interests; taking on goals related to diversity or inclusion.
Label policy	Label guidelines do not align with label policy; alternative labeling rules or labeling policies; balancing specificity and complexity.

	Errors/Mistakes	Design Decisions/Tradeoffs
Labels	Mis-labeled or inaccurate labels.	Sampling frame for model training.
Models	Mis-classification.	Model architecture, optimization structure, thresholds; balancing performance for different groups, balancing inclusion and errors.

	Design Decisions/Tradeoffs
Outcomes	What is the diversity or representation in the system? How do changes in the rows above manifest in changes to outcomes or representation?

Fairness Dimension Examples

	Design Decisions/Tradeoffs
Product policy	Alternative product design/goals; balancing stakeholder interests; taking on goals related to diversity or inclusion.
Label policy	Label guidelines do not align with label policy; alternative labeling rules or labeling policies; balancing specificity and complexity.

	Errors/Mistakes	Design Decisions/Tradeoffs
Labels	Mis-labeled or inaccurate labels.	Sampling frame for model training.
Models	Mis-classification.	Model architecture, optimization structure, thresholds; balancing performance for different groups, balancing inclusion and errors.

	Design Decisions/Tradeoffs
Outcomes	What is the diversity or representation in the system? How do changes in the rows above manifest in changes to outcomes or representation?

	Design Decisions/Tradeoffs
Product policy	Alternative product design/goals; balancing stakeholder interests; taking on goals related to diversity or inclusion.
Label policy	Label guidelines do not align with label policy; alternative labeling rules or labeling policies; balancing specificity and complexity.

	Errors/Mistakes	Design Decisions/Tradeoffs
Labels	Mis-labeled or inaccurate labels.	Sampling frame for model training.
Models	Mis-classification.	Model architecture, optimization structure, thresholds; balancing performance for different groups, balancing inclusion and errors.

	Design Decisions/Tradeoffs
Outcomes	What is the diversity or representation in the system? How do changes in the rows above manifest in changes to outcomes or representation?

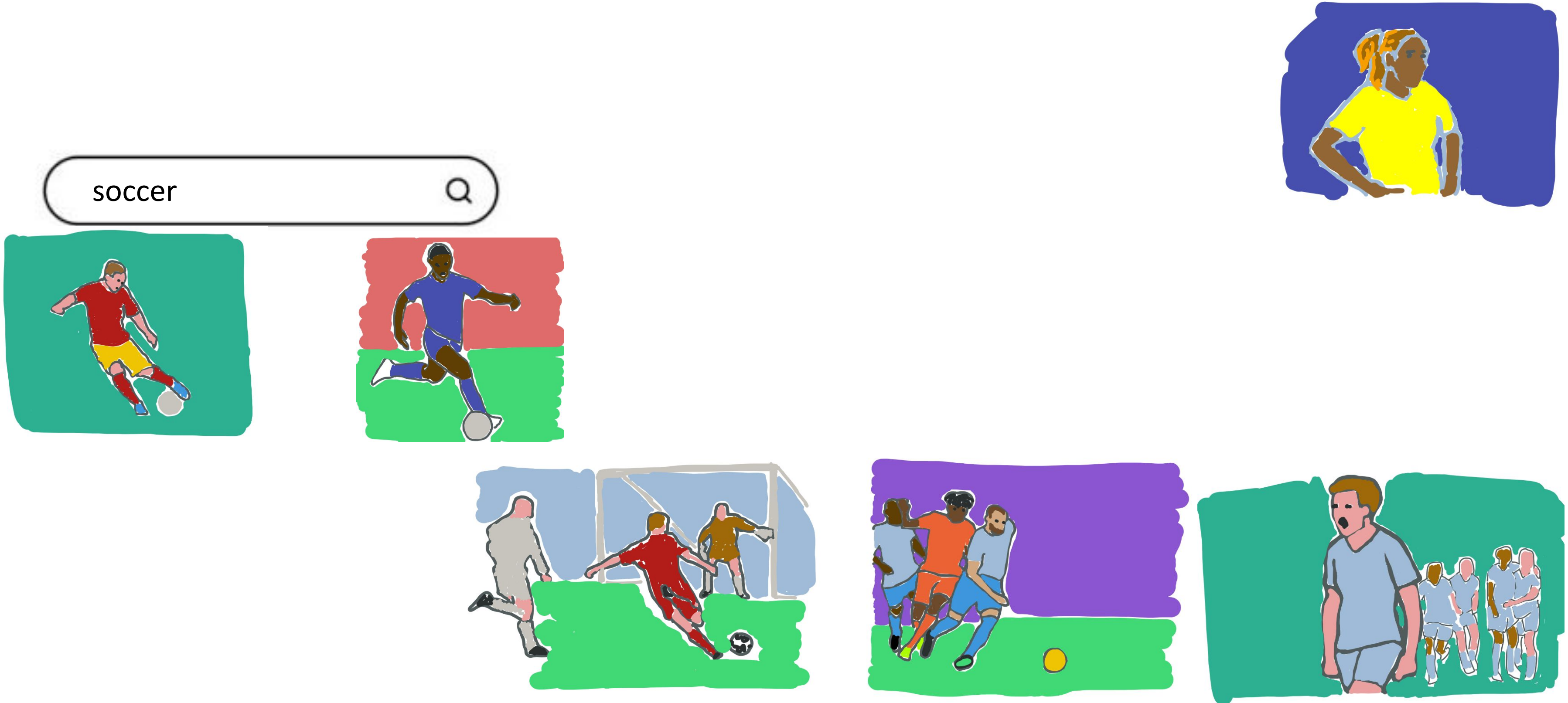

?
soccer
?
?

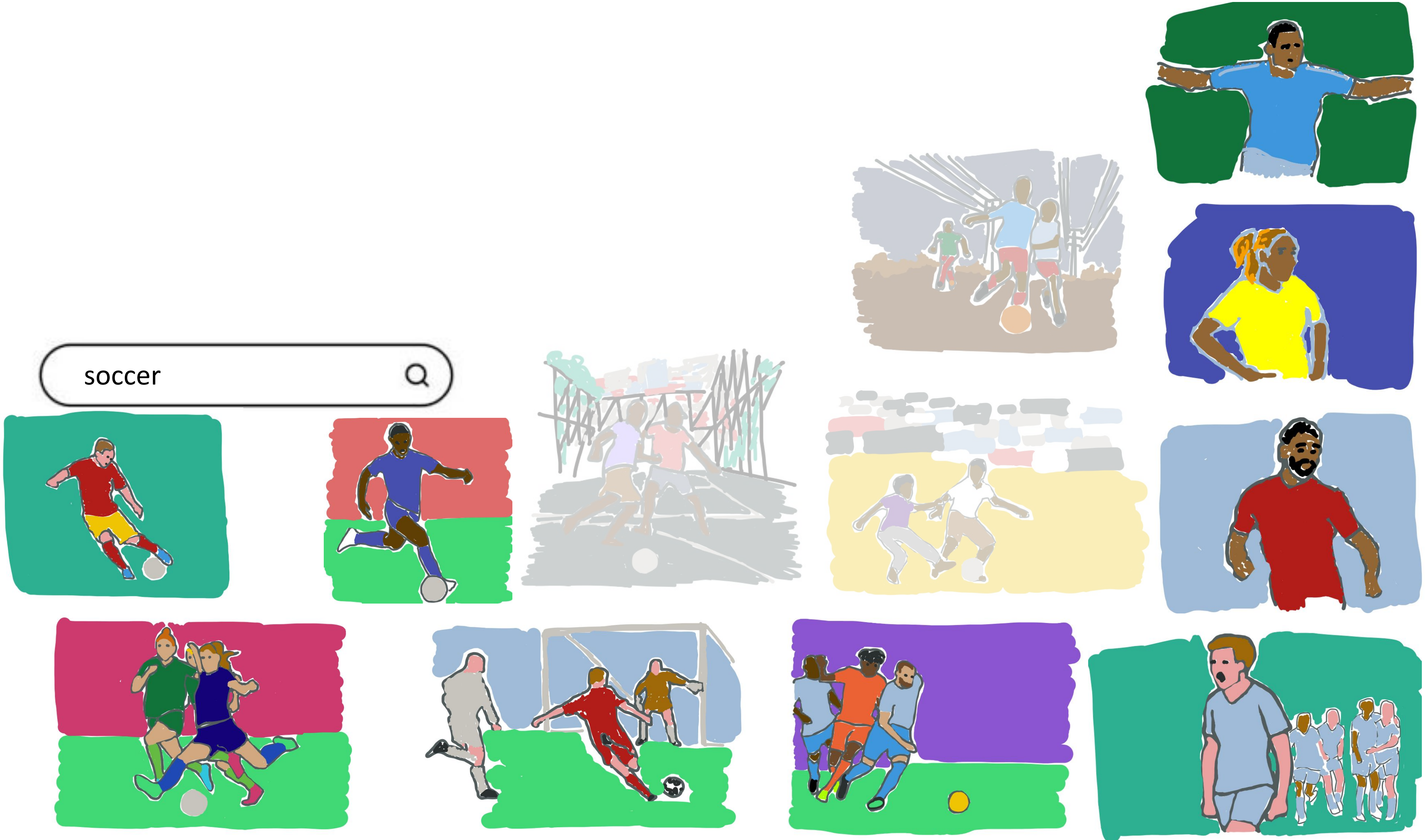

soccer

	Design Decisions/Tradeoffs
Product policy	Alternative product design/goals; balancing stakeholder interests; taking on goals related to diversity or inclusion.
Label policy	Label guidelines do not align with label policy; alternative labeling rules or labeling policies; balancing specificity and complexity.

	Errors/Mistakes	Design Decisions/Tradeoffs
Labels	Mis-labeled or inaccurate labels.	Sampling frame for model training.
Models	Mis-classification.	Model architecture, optimization structure, thresholds; balancing performance for different groups, balancing inclusion and errors.

	Design Decisions/Tradeoffs
Outcomes	What is the diversity or representation in the system? How do changes in the rows above manifest in changes to outcomes or representation?


soccer


soccer

	Design Decisions/Tradeoffs
Product policy	Alternative product design/goals; balancing stakeholder interests; taking on goals related to diversity or inclusion.
Label policy	Label guidelines do not align with label policy; alternative labeling rules or labeling policies; balancing specificity and complexity.

	Errors/Mistakes	Design Decisions/Tradeoffs
Labels	Mis-labeled or inaccurate labels.	Sampling frame for model training.
Models	Mis-classification.	Model architecture, optimization structure, thresholds; balancing performance for different groups, balancing inclusion and errors.

	Design Decisions/Tradeoffs
Outcomes	What is the diversity or representation in the system? How do changes in the rows above manifest in changes to outcomes or representation?


soccer

soccer

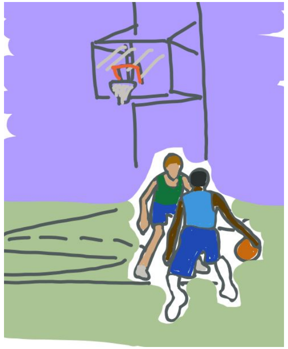
	Design Decisions/Tradeoffs
Product policy	Alternative product design/goals; balancing stakeholder interests; taking on goals related to diversity or inclusion.
Label policy	Label guidelines do not align with label policy; alternative labeling rules or labeling policies; balancing specificity and complexity.

	Errors/Mistakes	Design Decisions/Tradeoffs
Labels	Mis-labeled or inaccurate labels.	Sampling frame for model training.
Models	Mis-classification.	Model architecture, optimization structure, thresholds; balancing performance for different groups, balancing inclusion and errors.

	Design Decisions/Tradeoffs
Outcomes	What is the diversity or representation in the system? How do changes in the rows above manifest in changes to outcomes or representation?


soccer





soccer

	Design Decisions/Tradeoffs
Product policy	Alternative product design/goals; balancing stakeholder interests; taking on goals related to diversity or inclusion.
Label policy	Label guidelines do not align with label policy; alternative labeling rules or labeling policies; balancing specificity and complexity.

	Errors/Mistakes	Design Decisions/Tradeoffs
Labels	Mis-labeled or inaccurate labels.	Sampling frame for model training.
Models	Mis-classification.	Model architecture, optimization structure, thresholds; balancing performance for different groups, balancing inclusion and errors.

	Design Decisions/Tradeoffs
Outcomes	What is the diversity or representation in the system? How do changes in the rows above manifest in changes to outcomes or representation?

	Design Decisions/Tradeoffs
Product policy	Alternative product design/goals; balancing stakeholder interests; taking on goals related to diversity or inclusion.
Label policy	Label guidelines do not align with label policy; alternative labeling rules or labeling policies; balancing specificity and complexity.

	Errors/Mistakes	Design Decisions/Tradeoffs
Labels	Mis-labeled or inaccurate labels.	Sampling frame for model training.
Models	Mis-classification.	Model architecture, optimization structure, thresholds; balancing performance for different groups, balancing inclusion and errors.

	Design Decisions/Tradeoffs
Outcomes	What is the diversity or representation in the system? How do changes in the rows above manifest in changes to outcomes or representation?

Algorithmic Fairness Metrics

- Models, labels, errors
  - Based on scores/predictions and labels
- Outcomes
  - Based on predicted class

Algorithmic Fairness Metrics: Models I

Equalized Odds

- TP / (TP + FN)
  what proportion of actual positives are labeled positive
- TN / (FP + TN)
  what proportion of actual negatives are labeled negative

Precision, Recall

- Precision = TP / (TP + FP)
  positive predictive value; how many of the retrieved items are relevant?
- Recall = TP / (TP + FN)
  sensitivity; how many relevant items are retrieved?

		Actual values	
		Positive	Negative
Predicted values	Positive	TP	FP
	Negative	FN	TN

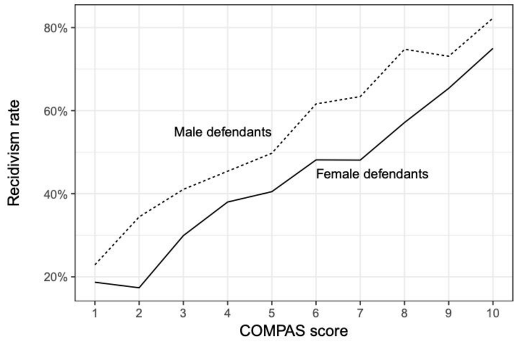
Algorithmic Fairness Metrics: Models II

Calibration

The Measure and Mismeasure of Fairness:
A Critical Review of Fair Machine Learning*

Sam Corbett-Davies, Stanford University
Sharad Goel, Stanford University

August 14, 2018



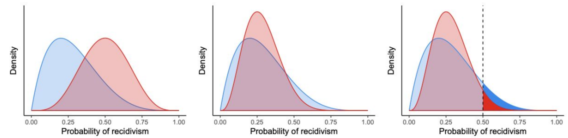
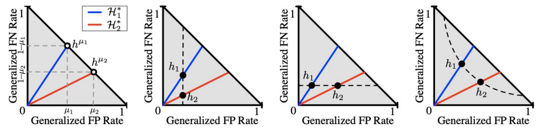


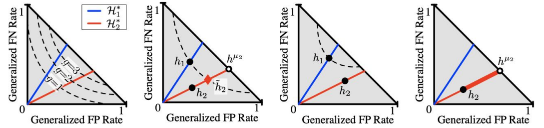
Figure 2: *Hypothetical risk distributions and a decision threshold (in the right-most plot). When risk distributions differ, infra-marginal statistics—like the precision and the false positive rate of a decision algorithm—also differ, illustrating the problem with requiring classification parity.*

- Compare (binned) scores with average outcomes
- Calibration accounts for differences in risk distributions
- Calibration is not compatible with constraints in except in cases of perfect prediction



(a) Possible cal. classifiers $\mathcal{H}_1^*, \mathcal{H}_2^*$ (blue/red). (b) Satisfying cal. and equal F.P. rates. (c) Satisfying cal. and equal F.N. rates. (d) Satisfying cal. and a general constraint.

Figure 1: Calibration, trivial classifiers, and equal-cost constraints – plotted in the false-pos./false-neg. plane. $\mathcal{H}_1^*, \mathcal{H}_2^*$ are the set of cal. classifiers for the two groups, and h^{μ_1}, h^{μ_2} are trivial classifiers.



(a) Level-order curves of cost. Low cost implies low error rates. (b) Usually, there is a calibrated classifier $\tilde{h}_2$ with the same cost of h_1. (c) Cal. and equal-cost are incompatible if h_1 has high error. (d) Possible cal. classifiers for G_2 (bold red) by mixing h_2 and h^{μ_2}.

Figure 2: Calibration-Preserving Parity through interpolation.

On Fairness and Calibration

Geoff Pleiss*, Manish Raghavan*, Felix Wu, Jon Kleinberg, Kilian Q. Weinberger
Cornell University, Department of Computer Science
{geoff,manish,kleinber}@cs.cornell.edu,
{fw245,kwq4}@cornell.edu

Algorithmic Fairness Metrics: Outcomes

Strict parity: $TP_a + FP_a = TP_b + FP_b$

Representation: $(TP_a + FP_a) / N = \mathbf{N}_a / \mathbf{N}$

		Actual values	
		Positive	Negative
Predicted values	Positive	TP	FP
	Negative	FN	TN

Algorithmic Fairness Metrics: Models vs Outcomes

- Fairness typically rooted in model errors rather than model outcomes
- Calibration is most in line with *equal treatment* or *equality of opportunity*
  - Similar items receive similar treatment independent of group membership
  - For now we are focused on equality, not equity
- Outcome metrics still provide useful signals
  - Products may have an interest in diversity in addition to equal treatment
  - Outcome metrics are often used to assess system health and can guide products through evaluating trade-offs

Personalized Ranking

Why is personalized ranking so challenging?

- Fairness for creators/providers/items in systems designed for viewers/consumers

Why is personalized ranking so challenging?

- Defining relevance
- Position + consumer bias
- People Problems

Why is personalized ranking so challenging?

1. Defining relevance
    a. The task is inherently less well-defined, no universal ground truth for each item
    b. A plethora of sparse data to choose from
    c. What is success for the product? How does that map to user experience?
    d. The conversion of certain *qualitative values* into *numerical values*
2. Position + consumer bias
    a. Present items in a ranked order (descending order of “relevance”)
    b. Complex systems, feedback loops, dependencies
    c. Session/composition/temporal effects, attention degrading etc.
    d. Potential correlations between consumer groups and creator group
3. People Problems
    a. A blurry line between preference and unfairness
    b. Preferences are not *fixed*
    c. Multi-stakeholder systems

Measuring Commonality in Recommendation of Cultural Content: Recommender Systems to Enhance Cultural Citizenship

Andres Ferraro
andresferraro@acm.org
McGill University
Montréal, Canada

Gustavo Ferreira
gustavo.ferreira@mila.quebec
McGill University
Montréal, Canada

Fernando Diaz
Canadian CIFAR AI Chair
Google
Montréal, Canada
diazf@acm.org

Georgina Born
University College London
London, United Kingdom
g.born@ucl.ac.uk

	Design Decisions/Tradeoffs
Product policy	Alternative product topline metrics, product goals, prioritizing one stakeholder (e.g. consumers, producers, items) group over another.
Ranking policy	Alternative ranking rules, inclusion of different components

	Errors/Mistakes	Design Decisions/Tradeoffs
Labels	Mis-labeled (position bias) or unreliable labels (human behavior).	Sampling for training (sessions vs viewers, timeframe).
Models	Mis-classification (incorrect position, mis-predicted event).	Model architecture, optimization structure, thresholds, interdependent tasks (event prediction)

	Design Decisions/Tradeoffs
Outcomes	What is the diversity or representation in the system? How do changes in the rows above manifest in changes to outcomes or representation?

Evaluation: Outcomes

Metrics: Measuring outcomes

- Parity, Skew @ k, Representation @ k
- Regression frameworks
- Gini, Atkinson, Ratios
- Comparison to long term holdouts

Parity, Skew @ rank k, Rep @ rank k

- Google images
  - Parity to population
- LinkedIn
  - Skew @ k: At rank k, how representative is the ranked list relative to an appropriate benchmark
- Netflix
  - Genre consistency at t and t+1

Less personalized → More personalized

Unequal Representation and Gender Stereotypes in Image Search Results for Occupations

Matthew Kay
Computer Science
& Engineering | dub,
University of Washington
mjskay@uw.edu

Cynthia Matuszek
Computer Science & Electrical
Engineering, University of
Maryland Baltimore County
cmat@umbc.edu

Sean A. Munson
Human-Centered Design
& Engineering | dub,
University of Washington
smunson@uw.edu

Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search

Sahin Cem Geyik, Stuart Ambler, Krishnaram Kenthapadi
LinkedIn Corporation, USA

Calibrated Recommendations

Harald Steck
Netflix
Los Gatos, California
hsteck@netflix.com

Parity, Skew @ rank k, Rep @ rank k

- How do you select a benchmark?
- What about personalization?
  - Base on follows, previous plays, previous recommendations
    - All affected by the recommendation system
  - What about quality-weighting?
  - What about dynamic preferences?
- What about unconnected recommendations?

Regression frameworks

- Regression models or covariate rebalancing
- Average outcomes (e.g. plays, clicks) for producer groups
- Rebalance or regress covariates that might impact outcomes (e.g. genre, number of songs, production quality) and re-assess averages
- Open Questions:
  - What kind of variables to include?
  - What about feedback effects?

Gini, Atkinson, Ratios

Measuring Disparate Outcomes of Content Recommendation Algorithms with Distributional Inequality Metrics

Tomo Lazovich[1*], **Luca Belli**[1], **Aaron Gonzales** [1], **Amanda Bower** [1], **Uthaipon Tantipongpipat** [1], **Kristian Lum** [1], **Ferenc Huszar** [2†], **Rumman Chowdhury** [1]

[1] Twitter, Inc.
[2] University of Cambridge

- Measures of inequality
- Tend to be difficult to adapt to group-level fairness
- Includes qualitative (interpretability) and empirical (stability and effect detection) considerations

Comparison to long term holdouts

- Compare outcomes of interest between users in ranked products versus users in unranked products (e.g. chronological feeds)

Algorithmic amplification of politics on Twitter

Ferenc Huszár[a,b,c,1,2], Sofia Ira Ktena[a,1,3], Conor O'Brien[a,1], Luca Belli[a,2], Andrew Schlaikjer[a], and Moritz Hardt[d]

- Key findings:
  - Ranked feeds amplify political content
  - Right leaning media amplified more than left leaning

Metrics: Measuring outcomes

- General pitfalls
  - Setting the right benchmark or comparison groups
  - Does not tell us *why* differences exist
  - Difficult to separate *success* from historical system bias
- General value
  - Diagnostic of potential representative harms
  - Even perfectly calibrated systems can lead to wide gaps in outcomes
  - Intuitive (but potentially misleading)

Evaluation: Models

	Design Decisions/Tradeoffs
Product policy	Alternative product topline metrics, product goals, prioritizing one stakeholder (e.g. consumers, producers, items) group over another.
Ranking policy	Alternative ranking rules, inclusion of different components

	Errors/Mistakes	Design Decisions/Tradeoffs
Labels	Mis-labeled (position bias) or unreliable labels (human behavior).	Sampling for training (sessions vs viewers, timeframe).
Models	Mis-classification (incorrect position, mis-predicted event).	Model architecture, optimization structure, thresholds, interdependent tasks (event prediction)

	Design Decisions/Tradeoffs
Outcomes	What is the diversity or representation in the system? How do changes in the rows above manifest in changes to outcomes or representation?

Ranking fairness measurements

- Problem set up
  - How are items scored?
  - Consumer bias
  - Position bias
- Measuring models offline
- Measuring models online

How are items scored?

- Some combination of proxies for relevance
- Model composed of many parts
- Hundreds of features as well as past engagement data

What’s the problem?

- Consumer bias
    - Scores are continuous and depend on session and consumer so they are not cross-session or cross-viewer compatible
    - Tastes and demographics are likely correlated, there will be spillover in performance between viewers and items
- Position bias
    - Attention degrades with position, this can lead to feedback loops where lower ranked items stay ranked lower (and the rich get richer)
    - Positions are zero sum, unlike classifications
    - Each individual event model can be assessed, but lists are rarely in the order of one model

Consumer bias

😀

country music		indie music
90 predicted		90 predicted
70 actual		68 actual
	<	
Calibration ratio **1.28**		Calibration ratio **1.32**

Consumer bias

	country music		indie music
🤠	75 predicted 60 actual Cal ratio **1.25**		15 predicted 13 actual Cal ratio **1.15**
		>	
👨🏽‍🎤	15 predicted 10 actual Cal ratio **1.5**		75 predicted 55 actual Cal ratio **1.36**

Position bias

- Salganik et al (2006)
  - Experimental music market shows impact of popularity rank on outcomes
  - Lists increase impact of social influence
  - More inequality, randomness under social influence conditions
- Singh and Joachims (2018)
  - Lack of proportionality
  - Small differences in estimated relevance lead to large differences in exposure
- Agarwal et al (2019)
  - Demonstrate decay in propensity to click on items by swapping items in first position with items in position k

Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market

Matthew J. Salganik,[1,2*] Peter Sheridan Dodds,[2*] Duncan J. Watts[1,2,3*]

Fairness of Exposure in Rankings

Ashudeep Singh
Cornell University
Ithaca, NY
ashudeep@cs.cornell.edu

Thorsten Joachims
Cornell University
Ithaca, NY
tj@cs.cornell.edu

Estimating Position Bias without Intrusive Interventions

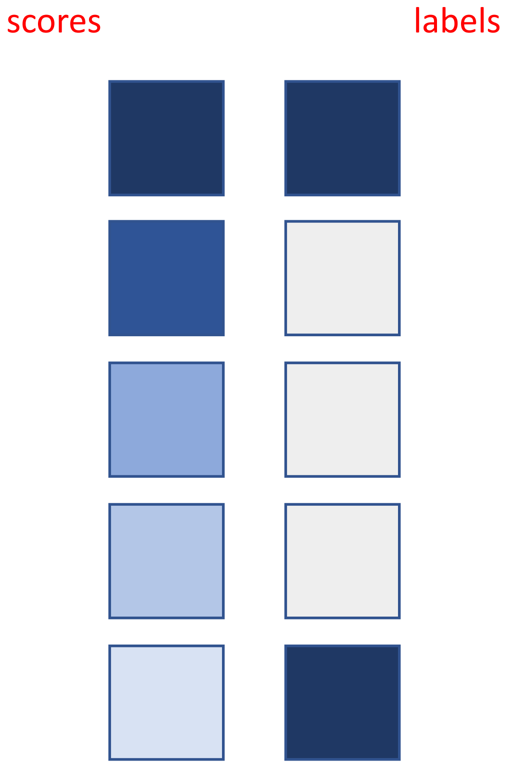
Aman Agarwal
Cornell University
Ithaca, NY
aa2398@cornell.edu

Ivan Zaitsev
Cornell University
Ithaca, NY
iz44@cornell.edu

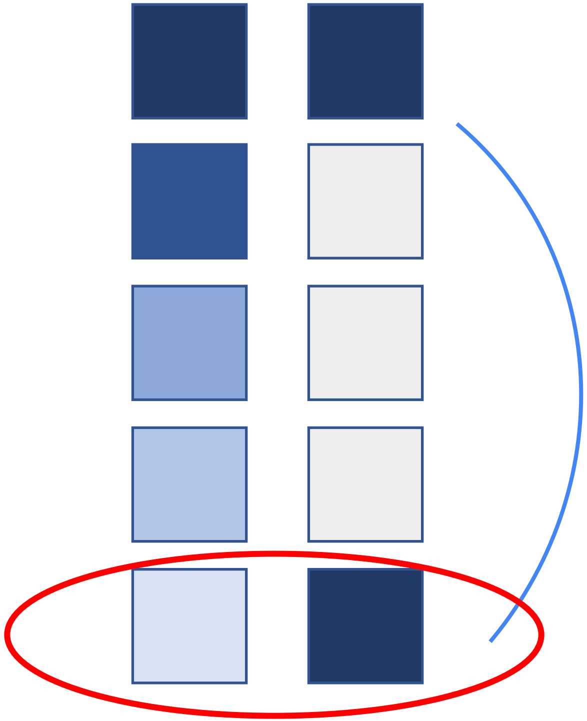
Xuanhui Wang, Cheng Li, Marc Najork
Google Inc.
Mountain View, CA
{xuanhui,chgli,najork}@google.com

Thorsten Joachims
Cornell University
Ithaca, NY
tj@cs.cornell.edu

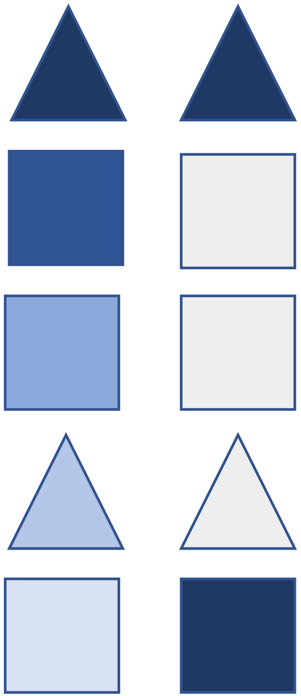
What is an error?



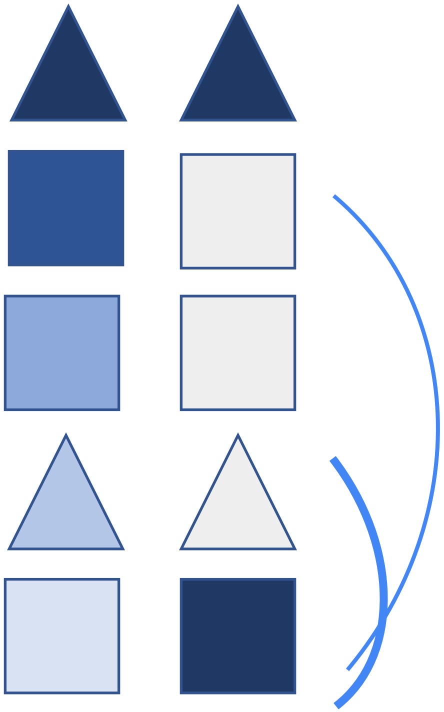
What is an error?



What is an error with multiple groups?



What is an error with multiple groups?



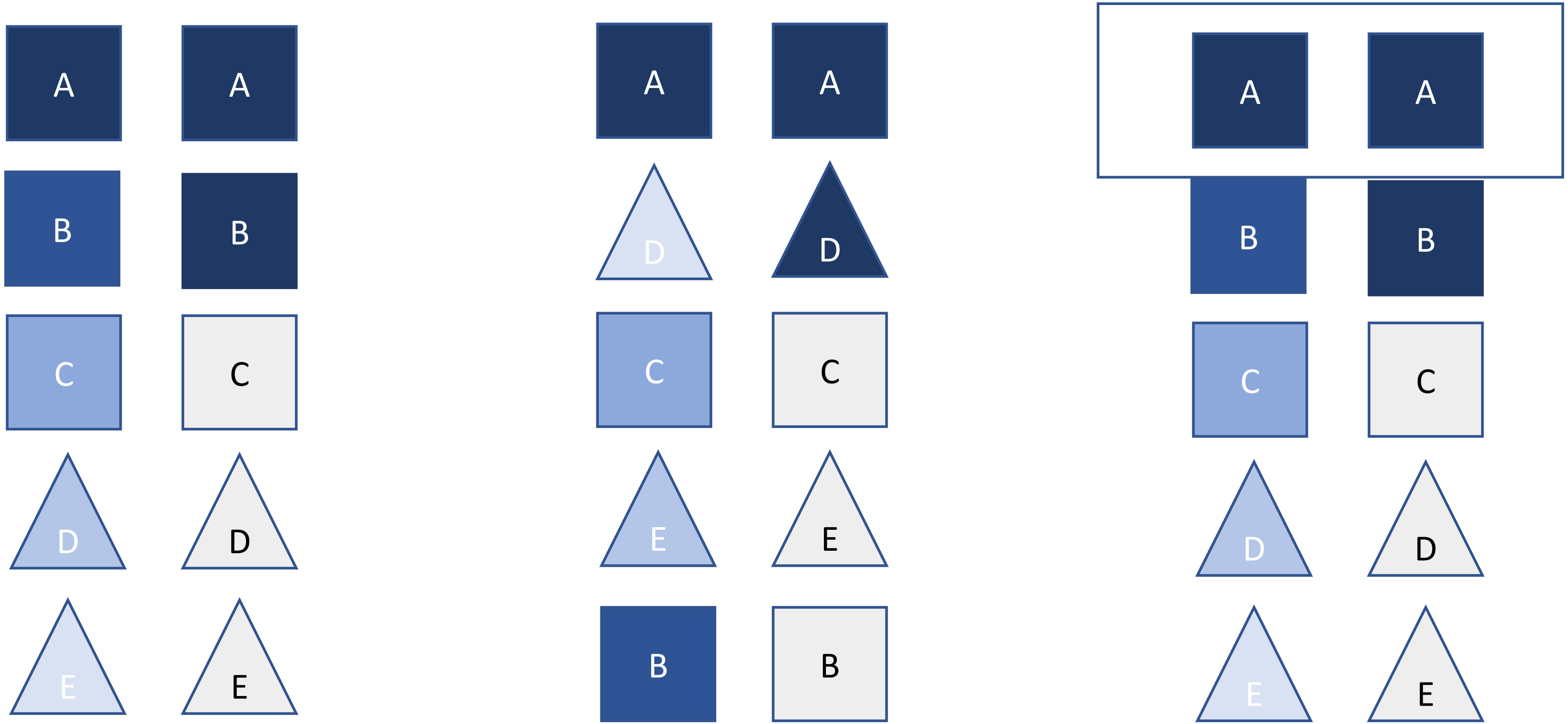
Fairness in Recommendation Ranking through Pairwise Comparisons

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, Cristos Goodrow
alexbeutel,jilinc,tulsee,hqian,liwei,wuyish,heldt,zhezhao,lichan,edchi,cristos@google.com
Google

Measuring models offline

- Calibration
- Pairwise comparisons
  - Intragroup pairwise errors
  - Intergroup pairwise errors
  - Matched pair calibration

Calibration



Pairwise comparisons I

- Good summary of challenges

Intergroup accuracy

- A model is considered to obey inter-group pairwise fairness if the likelihood of a clicked item being ranked above another relevant unclicked item from the opposite group is the same independent of group, conditioned on the items have been engaged with the same amount

Intragroup accuracy

- A model is considered to obey intra-group pairwise fairness if the likelihood of a clicked item being ranked above another relevant unclicked item from the same group is the same independent of group, conditioned on the items have been engaged with the same amount

Fairness in Recommendation Ranking through Pairwise Comparisons

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, Cristos Goodrow
alexbeutel,jilinc,tulsee,hqian,liwei,wuyish,heldt,zhezhao,lichan,edchi,cristos@google.com
Google

Pairwise comparisons II

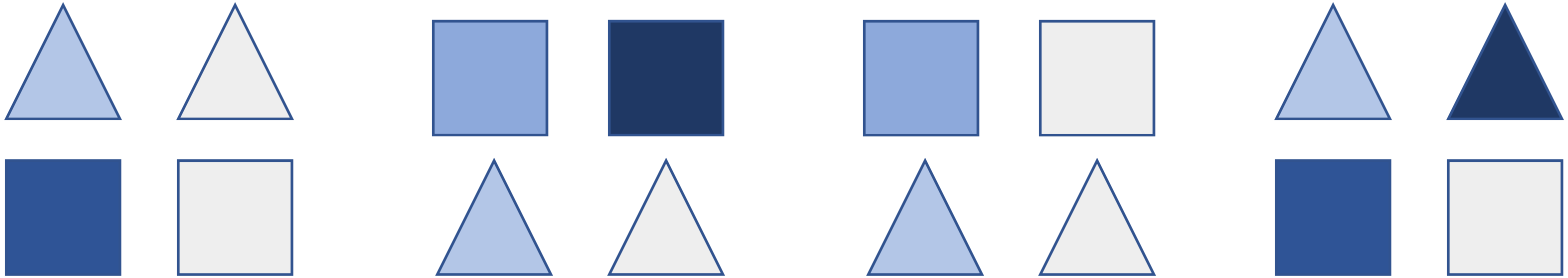
- Average label of adjacent items when group A is ahead versus when group B is ahead

An Outcome Test of Discrimination for Ranked Lists

Jonathan Roth
jonathanroth@brown.edu
Brown University
Providence, RI, USA

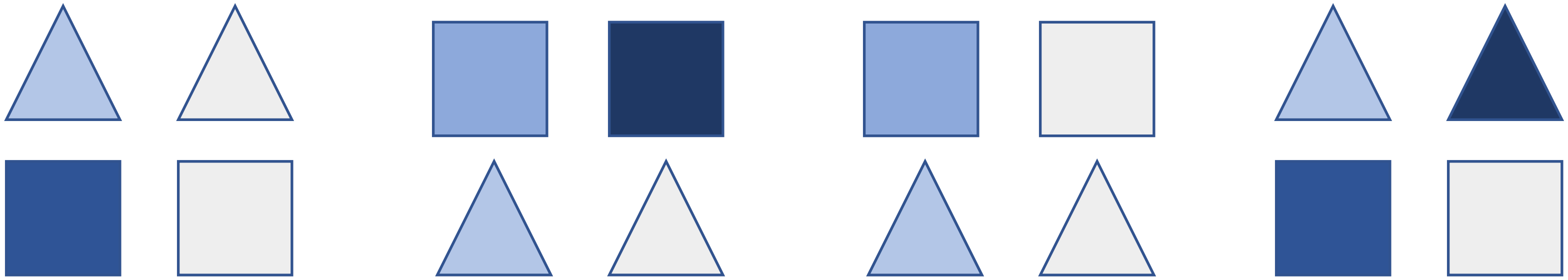
Guillaume Saint-Jacques
guillaume.saintjacques@gmail.com
Apple
USA

YinYin Yu
yinyyu@linkedin.com
LinkedIn
USA



Pairwise Comparisons III

- A calibration extension of pairwise comparisons with score matching.
- Match on score and adjacency in the ranked list.
- We can then compare average labels in this balanced set.
- A higher average label indicates the system has under-ranked items from that group.



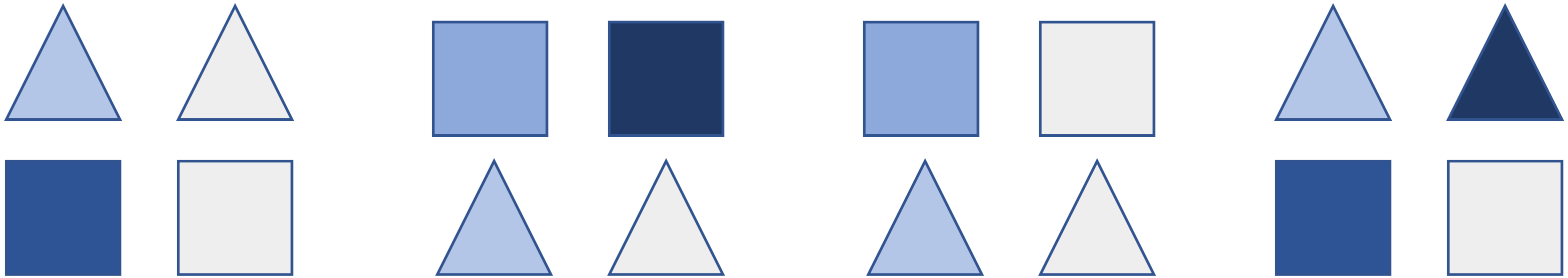
Pairwise Comparisons I & II

Pros

- Eliminates issues with cross user and cross session variation in model score by relying on position
- Isolates to key set of comparisons
- Relatively computationally efficient

Cons

- Ignores scores which makes intervening at the model level difficult.
- Underlying cause is unknown



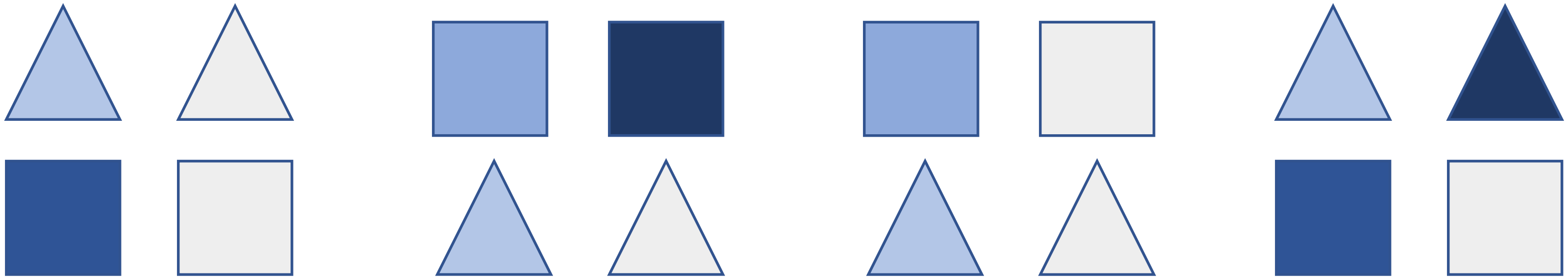
Pairwise Comparisons III

Pros:

- Uses model scores, more akin to calibration

Cons:

- It's hard to know if items even lower in the list would also have higher average labels.
- Scale differences in score versus label space may cause misleading results
- Score matching limits our data to places where there are ties
- Decisions we care about
- Might be lower in the list



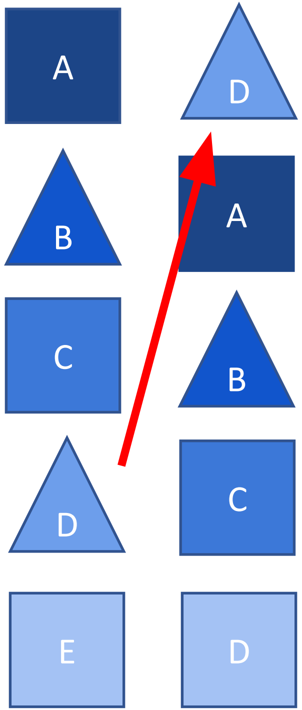
Measuring models offline: should you do it?

- Pros
  - Safer, less risk to the systems
  - Better than not measuring
- Cons
  - Less reliable signal
  - Risk that findings will not match production
  - Limited ability to address position bias
  - No counterfactual data (e.g. with different ranking outcomes)

Measuring models online

- Calibration with boosts
- Pairwise Perturbations
- Counterfactual group analysis

Calibration with boosts



- Boost from k to position 0 and assess calibration
- Swap(1, k) – interventions, create propensity estimation to adjust for position bias
- Addresses position bias
- How large to set k?

Estimating Position Bias without Intrusive Interventions

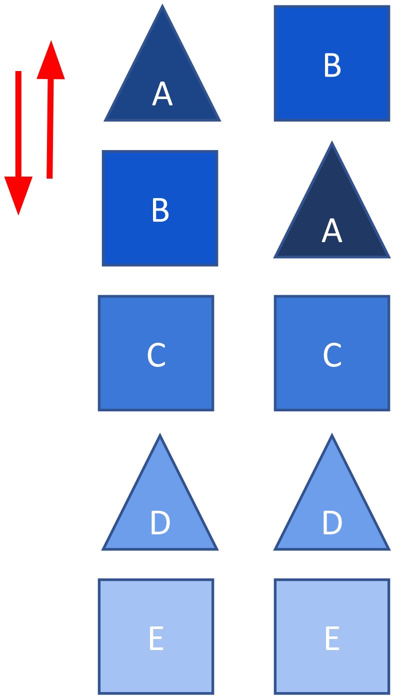
Aman Agarwal
Cornell University
Ithaca, NY
aa2398@cornell.edu

Ivan Zaitsev
Cornell University
Ithaca, NY
iz44@cornell.edu

Xuanhui Wang, Cheng Li, Marc Najork
Google Inc.
Mountain View, CA
{xuanhui,chgli,najork}@google.com

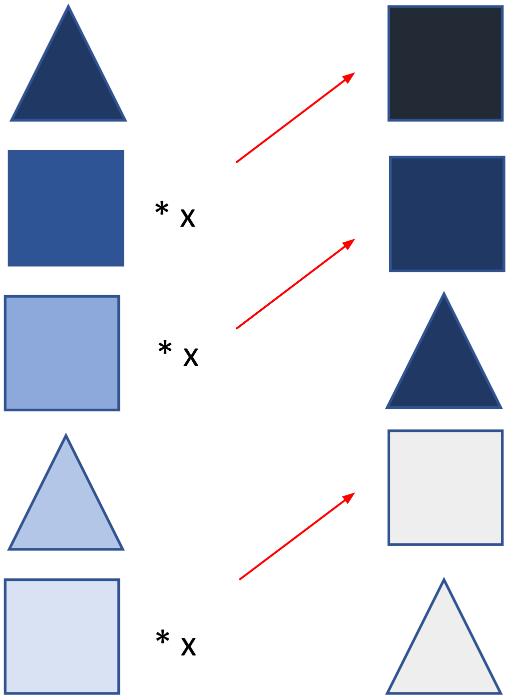
Thorsten Joachims
Cornell University
Ithaca, NY
tj@cs.cornell.edu

Pairwise perturbations



- Swap two items, collect labels
- Assess the impact of position bias, position by position
- This also allows for online measurement of the matched pair metric
- Low risk of harm to user experience, minimal estimation of full impact of feedback effects
- Requires very good logging

Counterfactual group analysis



- Search a grid of potential group-level score changes
- If you can obtain a higher product metric value with nonzero changes to group specific scores/positions, the ranker is unfair.

An Outcome Test of Discrimination for Ranked Lists

Jonathan Roth* Guillaume Saint-Jacques† YinYin Yu‡

November 16, 2021

Selection Problems in the Presence of Implicit Bias

Jon Kleinberg, Cornell University
Manish Raghavan, Cornell University

Becker's (1957) taste-based discrimination

Rooney Rule (2003)

Measuring models online: should you do it?

- Pros
  - More reliable information
  - Could theoretically translate quickly to mitigations
- Cons
  - More product and user experience risk
  - Policy and legal complications

Methods Review

- Outcomes
    - Parity, skew @ k
    - Covariate adjusted parity
    - Long term holdouts
- Models
    - Offline
        - Calibration
        - Pairwise Comparisons
    - Online
        - Calibration with boosts
        - Pairwise Perturbations
        - Counterfactual group analysis

Methods Review

- Outcomes
  - Use with a strong notion of desirable benchmark
  - Overall health and diversity in a system
  - Even well-calibrated systems can have large outcome gaps
- Models
  - Measures variation in system performance
  - Calibration most consistent with the AI Fairness literature is challenging in the ranking setting
  - Trade-offs between highly localized measures (pairwise) of fairness and the potential to disrupt user experiences (exploring more variety in ranking policies)

Design Decisions

Design decisions revisited

- The space is nearly infinite, but here are some real-world examples:
    - Product policy
        - Additional tools for users
            - Skin tone filters on Pinterest
            - Chronological Feed on Instagram
        - Diversity criteria
            - Inclusion of balanced perspectives in news aggregation on Google News
    - Ranking policy
        - Boosting/Re-ranking
            - Increase demographic representation in candidate search on LinkedIn
        - Shift in value model
            - Meaningful Social Interaction on Facebook News Feed
    - Label policy
        - Casual Conversations Data

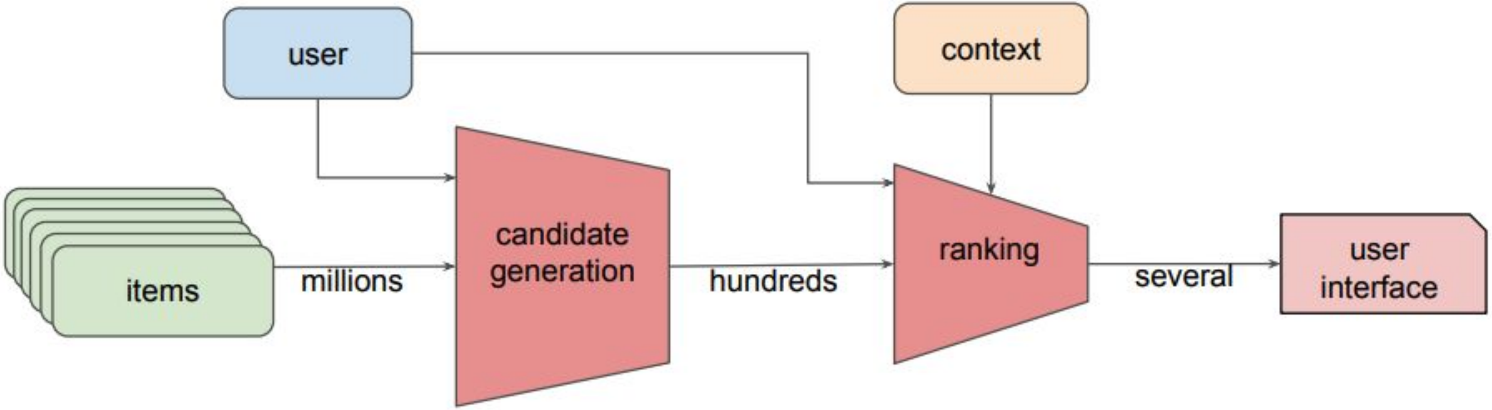
Closing thoughts and open questions

- You can’t get signal on items you never show, so some amount of randomness is always good (and may have good fairness qualities)
- Measuring other system components (e.g. sourcing or candidate retrieval)
- How much measurement is enough?
- Learning to rank with fairness in mind (up next)

Part 3

Fairness in Learning-to-Rank and Collaborative Filtering

Large-scale Recommender Systems



Low latency, High recall.

e.g., Nearest Neighbor search on embeddings, Collaborative Filtering.

High precision, can afford more computation per item.

e.g. Learning-to-rank.

Part 3: Outline

How to train a fair recommender system?

- Collaborative Filtering
- Learning-to-Rank
- Online Learning, Contextual bandits, Sequential decision making (RL)

X

- Selection Bias
- User Fairness
- Item Fairness
- Multistakeholder perspective
- Feedback loops

X

- Evaluation
- Pre-processing
- In-processing
- Post-processing

Part 3: Outline

How to train a fair recommender system?

- Collaborative Filtering
- Learning-to-Rank
- Online Learning, Contextual bandits, Sequential decision making (RL)

X

- Selection Bias
- User Fairness
- Item Fairness
- Multistakeholder perspective
- Feedback loops

Collaborative Filtering

- Collaborative filtering uses similarities between users and items simultaneously to provide recommendations, i.e.,
  - recommend an item to user A based on the interests of a "similar" user B.
- Common method: Matrix Factorization of the user-item rating matrix.

Given a dataset of user item ratings: $Y_{u,i}$,

Find a user and item embedding matrix (U and V), so that the $U^T V$ is as close to the ratings matrix.

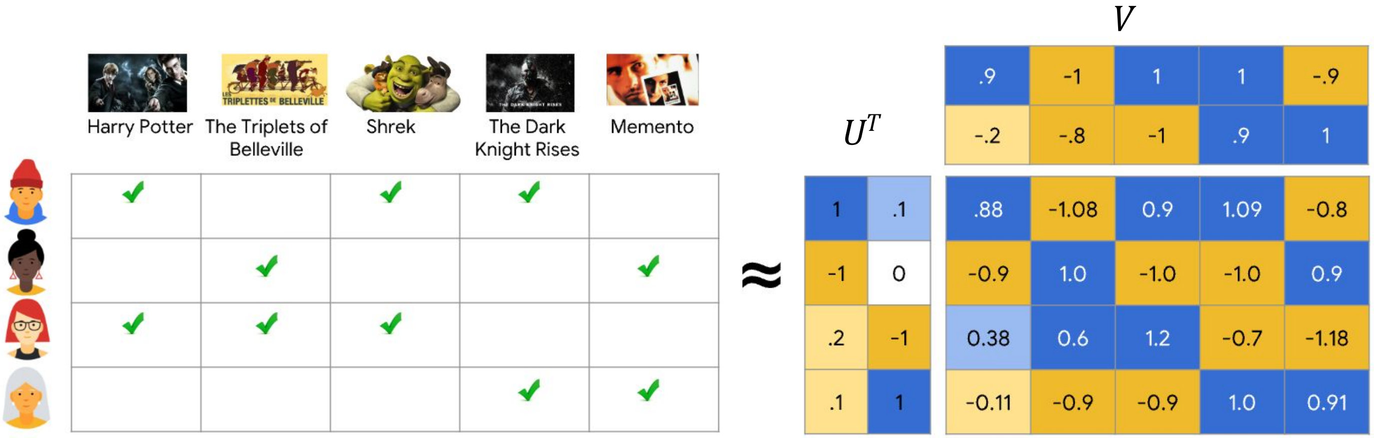


Image source: link

Missing data in Collaborative filtering

- Conventional loss function assumes ratings are **missing completely at random (MCAR)**, i.e.,
  - *Pr[$Y_{u,i}$ is observed]* is equal for all u, i.
- Other types of missing data:
  - **Missing at random (MAR)**: missingness depends on observable features
  - **Missing not at random (MNAR)**: missingness may depend on observable features, unobservable features and the rating itself.

[Little & Rubin 2002]
[Marlin & Zemel 2009]

- Ignoring the missingness mechanism,
  - causes evaluation to be biased,
  - the ML model predictions could be biased/skewed.

[Schnabel et al. ICML 2016]

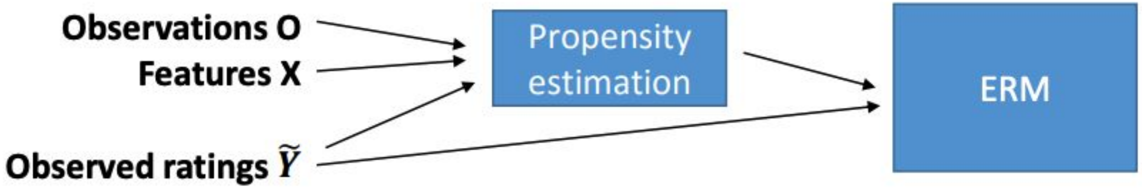
Handling missing data in Collaborative filtering

- Use inverse propensity scoring loss function

$$\hat{Y}^{ERM} = \underset{V,W}{\operatorname{argmin}} \left\{ \sum_{O_{u,i}=1} \frac{1}{P_{u,i}} \left(Y_{u,i} - V_u W_i\right)^2 + \lambda\left(\|V\|_F^2 + \|W\|_F^2\right) \right\}$$

propensity weight

- Propensity Estimation:
  Build a discriminative model using the given information to predict $\hat{P}(O_{u,i} = 1 | X_{u,i})$.



[Schnabel et al. ICML 2016]

Part 3: Outline

How to train a fair recommender system?

- Collaborative Filtering
- Learning-to-Rank
- Online Learning, Contextual bandits, Sequential decision making (RL)

X

- Selection Bias
- User Fairness
- Item Fairness
- Multistakeholder perspective
- Feedback loops

User Fairness

Yao & Huang (NIPS 2017) define fairness metrics based on the discrepancy between the prediction behavior for *disadvantaged* users and *advantaged* users. (Group Fairness)

- **Value Fairness**: Difference in signed error of advantaged and disadvantaged groups.
- **Absolute Fairness**: Difference in absolute errors of advantaged and disadvantaged groups.
- **Underestimation unfairness**: inconsistency in how much the predictions underestimate the true ratings.
- **Overestimation unfairness**: inconsistency in how much the predictions overestimate the true ratings

Average predicted score from disadvantaged users

Average ratings from disadvantaged users

Average predicted score from advantaged users

Average ratings from advantaged users

$$\text{Value Fairness:} \quad U_{\text{val}} = \frac{1}{n} \sum_{j=1}^{n} \left| \left(\boxed{\mathrm{E}_g [y]_j} - \boxed{\mathrm{E}_g [r]_j} \right) - \left(\boxed{\mathrm{E}_{\neg g} [y]_j} - \boxed{\mathrm{E}_{\neg g} [r]_j} \right) \right|$$

Train using a joint objective

$$\min_{\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{u}, \boldsymbol{v}} \boxed{J(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{u}, \boldsymbol{v})} + \boxed{U}$$

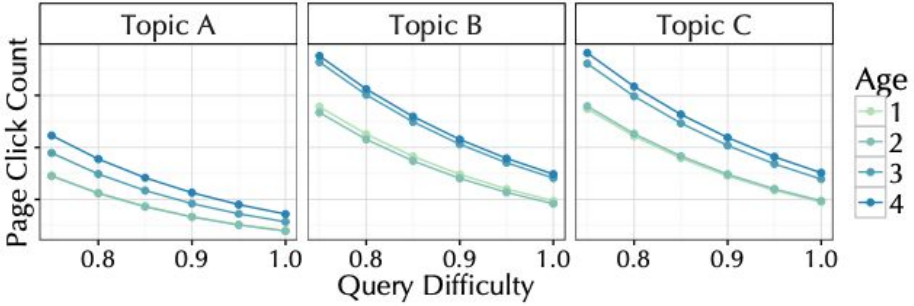
Loss for recommender model

Fairness constraint

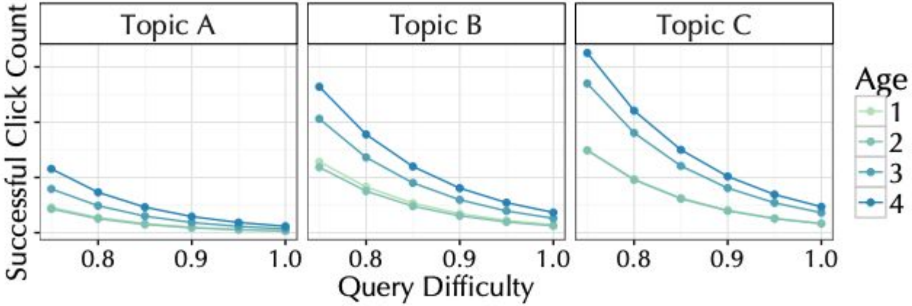
Equal access across user demographics

- Auditing search and recommender systems for equal access is more complicated than comparing engagement metrics across demographics.
- Dataset sizes differ significantly across demographics.
- Differences in engagement metrics and latent satisfaction are confounded by differences in usage across genders and age groups.

Mostly open question: How do we compare metrics across user groups?



page click count (hardest queries)



successful click count (hardest queries)

[Mehrotra et al. WWW 2017, Ekstrand et al. FAT* 2018]

Part 3: Outline

How to train a fair recommender system?

- Collaborative Filtering
- Learning-to-Rank
- Online Learning, Contextual bandits, Sequential decision making (RL)

X

- Selection Bias
- User Fairness
- Item Fairness
- Multistakeholder perspective
- Feedback loops

Item Fairness

Inter-group pairwise accuracy:

$$A_{G_i > G_j} := P(f(x) > f(x') \mid y > y', (x, y) \in G_i, (x', y') \in G_j),$$

- A ranking model f obeys **intergroup pairwise fairness** if the likelihood of **correctly ranking** a more relevant item x (of group G) over a less relevant item x' of another group is equal for all groups G. [Beutel et al. 2019, Narasimhan et al. 2019]
- Beutel et al. propose a regularizer that minimizes the correlation between the group membership and the model's predictions.
- Zhou et al. 2019 propose a post-processing method using a monotonic transformation of the scoring function.

Part 3: Outline

How to train a fair recommender system?

- Collaborative Filtering
- Learning-to-Rank
- Online Learning, Contextual bandits, Sequential decision making (RL)

X

- Selection Bias
- User Fairness
- Item Fairness
- Multistakeholder perspective
- Feedback loops

Probability Ranking Principle (PRP)

Robertson (1977)

- ”if a reference retrieval system’s response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request,
- where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose,
- the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data.”

THE PROBABILITY RANKING PRINCIPLE IN IR

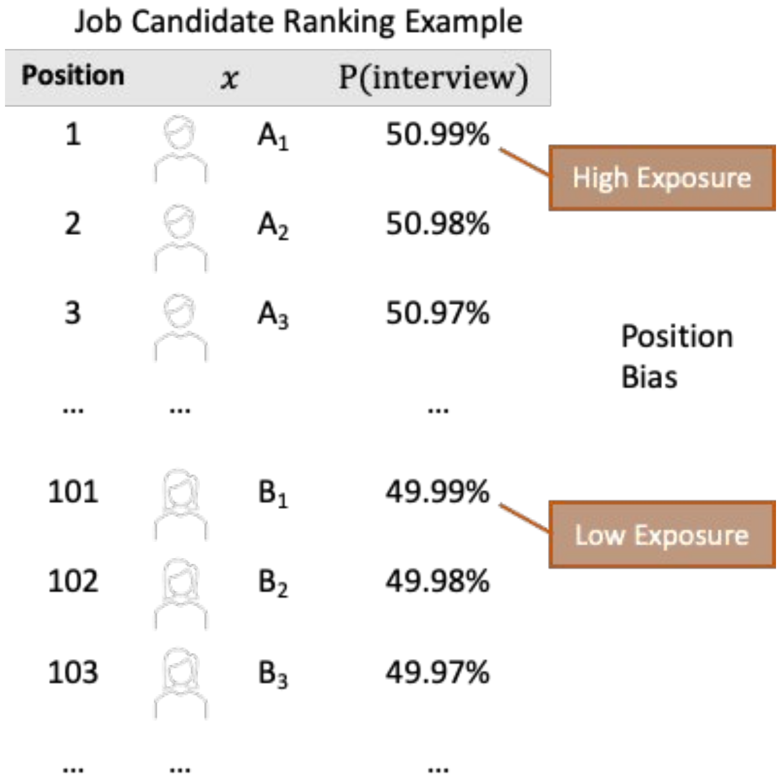
S. E. ROBERTSON
School of Library, Archive, and Information Studies, University College London

The principle that, for optimal retrieval, documents should be ranked in order of the probability of relevance or usefulness has been brought into question by Cooper. It is shown that the principle can be justified under certain assumptions, but that in cases where these assumptions do not hold, the principle is not valid. The major problem appears to lie in the way the principle considers each document independently of the rest. The nature of the information on the basis of which the system decides whether or not to retrieve the documents determines whether the document-by-document approach is valid.

PRP in a two-sided system

- In two-sided markets, PRP might be inadequate since it does not explicitly consider the item-side utility.

- Examples:
  - Job Candidate Ranking
    - Amplifies existing societal biases.

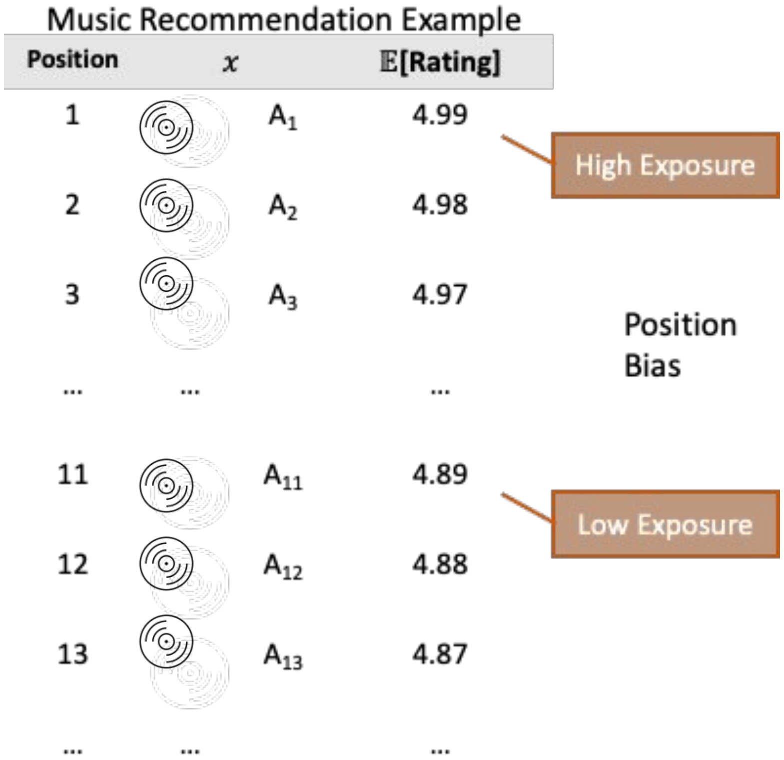


[Singh & Joachims 2018, Biega et al. 2018]

PRP in a two-sided system

- In two-sided markets, PRP might be inadequate since it does not explicitly consider the item-side utility.

- Examples:
  - Job Candidate Ranking
    - Amplifies existing societal biases.
  - Music Recommendation
    - Winner-takes-all!



[Singh & Joachims 2018, Biega et al. 2018]

PRP in a two-sided system

- In two-sided markets, PRP might be inadequate since it does not explicitly consider the item-side utility.
- Examples:
  - Job Candidate Ranking
    - Amplifies existing societal biases.
  - Music Recommendation
    - Winner-takes-all!
  - News Ranking
    - Polarization of the platform.

News Ranking Example

Position	x		$P(read)$	
1	R	R_1	50.99%	High Exposure
2	R	R_2	50.98%	
3	R	R_3	50.97%	
...	...		...	Position Bias
101	T	T_1	49.99%	Low Exposure
102	T	T_2	49.98%	
103	T	T_3	49.97%	
...	...		...	

[Singh & Joachims 2018, Biega et al. 2018]

In online platforms,

Exposure $\longrightarrow$ Opportunity

Hence,

Fairness $\longrightarrow$ Fair Allocation of Exposure

Position-based Model of Exposure

Exposure e_k is the probability a user observes the item at position k.

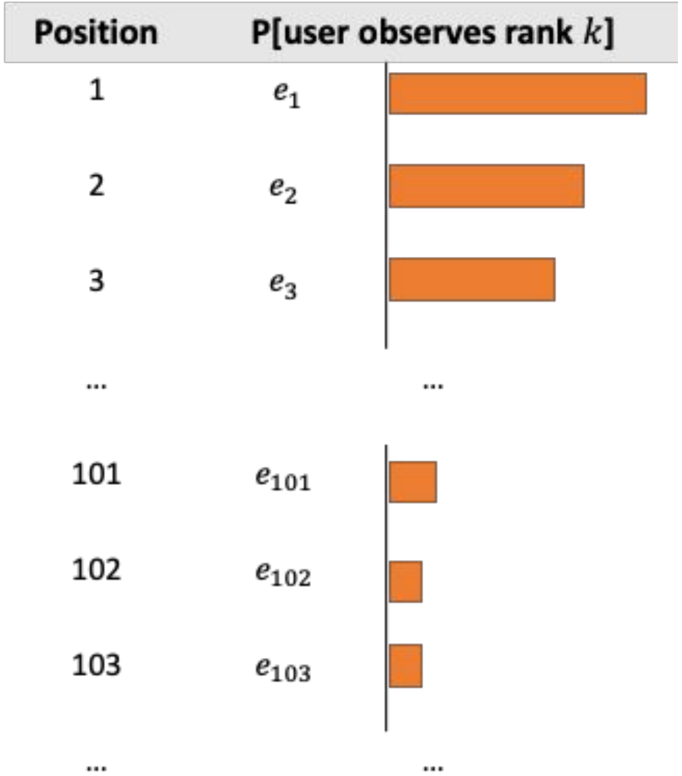
Exposure of a group of items (e.g., seller, artist, etc.)

$$Exp(G|y) = \sum_{y(k) \in G} e_k$$

Other user-click models: Cascading click model (CCM), etc. *[Chuklin et al. 2015]*

How to estimate?

- Eye tracking *[Joachims et al. 2007]*
- Intervention studies *[Joachims et al. 2017]*
- Intervention harvesting *[Agarwal et al. 2019]*



Fairness of Exposure

Goal: Enable the explicit statement of how exposure is allocated relative to the value or merit of the items in the group.

For example: Exposure for each individual/group should be proportional to the relevance of the group.

[Singh & Joachims 2018, Biega et al. 2018]

Equal Expected Exposure

For tasks with graded relevance (e.g., movie ratings — 1 to 5, binary relevance — 0, 1), define equal expected exposure as:

No item has less or more expected exposure as compared to other items in the same relevance grade.

[Diaz et al 2019]

Disparate Exposure & Impact

Disparate exposure: Allocate exposure proportional to relevance per group

$$\text{Exposure} \propto \text{Relevance}$$

$$\frac{Exp(G_0|x)}{Exp(G_1|x)} = \frac{Rel(G_0|x)}{Rel(G_1|x)}$$

Disparate impact: Allocate expected clickthrough rate proportional to relevance per group

$$\frac{\sum_{d \in G_0} Exp(d|x)\, Rel(d|x)}{\sum_{d \in G_1} Exp(d|x)\, Rel(d|x)} = \frac{Rel(G_0|x)}{Rel(G_1|x)}$$

[Singh & Joachims, KDD 2018]

Fairness of Exposure

Objective: Given relevance scores, find a ranking that optimizes user utility while satisfying fairness of exposure constraints, e.g., exposure proportional to average relevance.

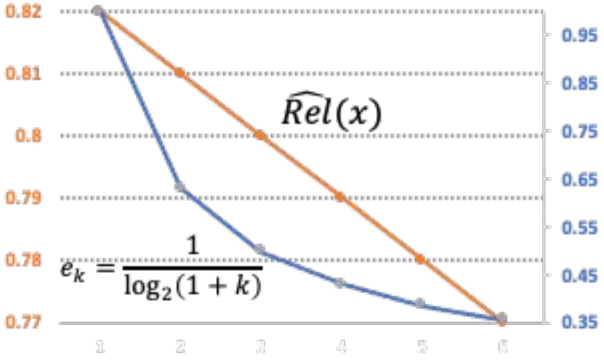
Items	$\hat{h}(x)$
A_1	0.82
A_2	0.81
A_3	0.80
B_1	0.79
B_2	0.78
B_3	0.77

$\times$

Exposure@k
e_1
e_2
e_3
e_4
e_5
e_6

Problem:

- Exposure drops off at a different rate than relevance.
- Rankings are discrete combinatorial objects.
  - Exponential solution space!



[Singh & Joachims, KDD 2018]

Key Idea 1: Stochastic Ranking Policies

- Ranking Policy

  $\pi(y|x)$ is the conditional distribution over rankings of items under query x.

y_1	y_2	y_3	y_4
A_1	A_1	A_1	B_1
A_2	B_1	A_2	A_1
A_3	A_2	B_1	B_2
B_1	B_2	A_3	A_2
B_2	A_3	B_2	B_3
B_3	B_3	B_3	A_3
0.40	0.40	0.16	0.04

Define Utility

$$U(\pi|x) = \sum_y U(y|x) \cdot \pi(y|x)$$

Define Exposure

$$Exp(d|\pi) = \sum_k e_k \cdot P(rank(d) = k \mid \pi)$$

[Singh & Joachims, KDD 2018]

Key Idea 2: Doubly Stochastic Matrices

Represent a Stochastic Ranking π as a Marginal Rank Distribution $\mathbb{P}$.

$$\begin{array}{cc} & \text{Rank} \\ \text{Item} & \begin{pmatrix} \cdot & \cdot & \cdot \\ & \cdot & \cdot \\ \cdot & \mathbb{P}_{i,k} & \\ & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ & \cdot & \cdot \end{pmatrix} \end{array}$$

$\mathbb{P}_{i,k}$ = Probability of item i at position k.

Utility (e.g., DCG, Avg Precision) and Exposure can be expressed as a Linear function of the matrix.

For example, $\mathrm{DCG}(\mathbb{P}) = \sum_i \mu_i \sum_k \frac{\mathbb{P}_{i,k}}{\log(1+k)}$.

Optimization problem of finding $\mathbb{P}$ that optimizes utility U and satisfies fairness constraints → Linear Program

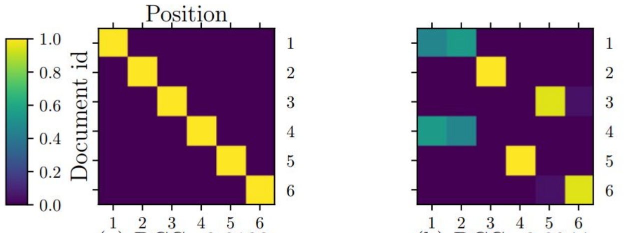
[Singh & Joachims, KDD 2018]

Example: Exposure Proportional to Relevance

Items	$\hat{h}(x)$
A_1	0.82
A_2	0.81
A_3	0.80
B_1	0.79
B_2	0.78
B_3	0.77

$\times$

Exposure@k
e_1
e_2
e_3
e_4
e_5
e_6



(a) DCG=3.8193, DTR=1.7483

(b) DCG=3.8044, DTR=1.0000

Without Fairness Constraint

$\mathbb{P}_{\text{fair}}$: Proportional Exposure

Problem setup: Maximize Utility (e.g., DCG) while fulfilling the fairness constraint (exposure proportional to relevance).

Solution: Ranking Policy

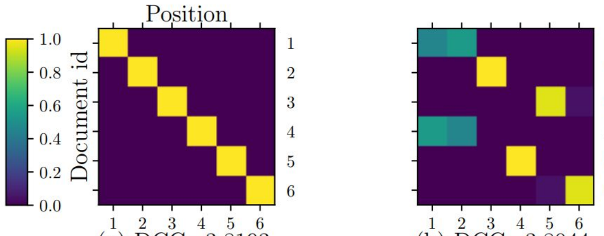
[Singh & Joachims, KDD 2018]

Example: Exposure Proportional to Relevance

Items	$\hat{h}(x)$
A_1	0.82
A_2	0.81
A_3	0.80
B_1	0.79
B_2	0.78
B_3	0.77

$\times$

Exposure@k
e_1
e_2
e_3
e_4
e_5
e_6



(a) DCG=3.8193, DTR=1.7483

(b) DCG=3.8044, DTR=1.0000

Without Fairness Constraint

$\mathbb{P}_{\text{fair}}$: Proportional Exposure

Solution: Ranking Policy

What if these relevance predictions are biased?

How to incorporate these constraints into a learning to rank framework?

[Singh & Joachims, KDD 2018]

Learning-to-Rank with fairness constraints

For a query x, rank a candidate set $\mathcal{S}_x = \{d_1, d_2, d_3, \ldots\}$ of items

- d_i represented by features $\psi(d_i|x)$, and
- d_i has a merit score (e.g., relevance—whether a user would click it or not).

Ranking Policy π maps $\mathcal{S}_x$ to a ranking.

[Singh & Joachims, NeurIPS 2019]

Learning-to-Rank with fairness constraints

For a query x, rank a candidate set $\mathcal{S}_x = \{d_1, d_2, d_3, \dots\}$ of items

- d_i represented by features $\psi(d_i|x)$, and
- d_i has a merit score (e.g., relevance—whether a user would click it or not).

Ranking Policy π maps $\mathcal{S}_x$ to a ranking.

Learning objective: Find policy π that maximizes expected utility U with small disparity D

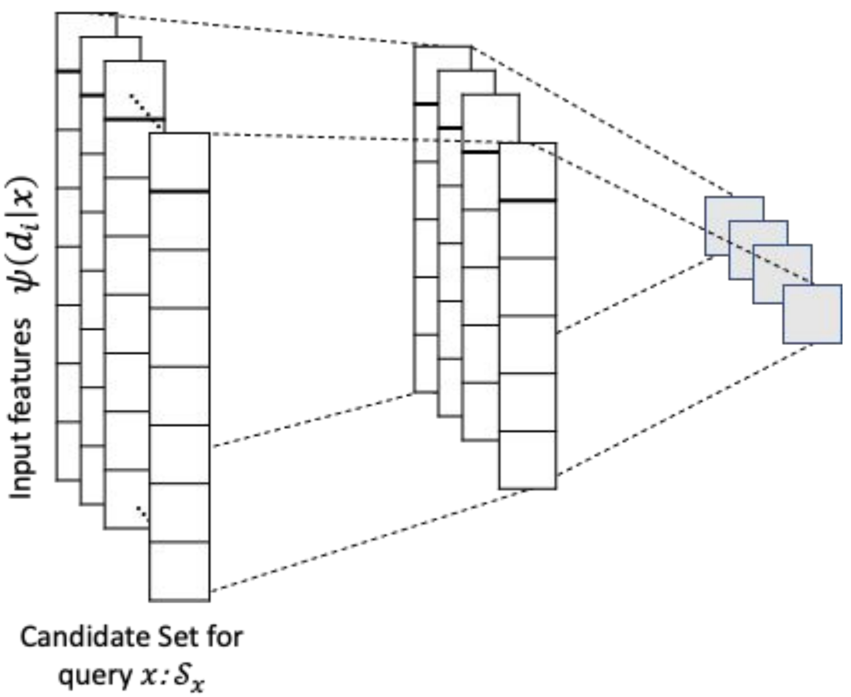
$$\pi^* = \mathrm{argmax}_\pi \ \mathbb{E}_x[U(\pi|x)] \text{ s.t. } \mathbb{E}_x[D(\pi|x)] \leq \delta.$$

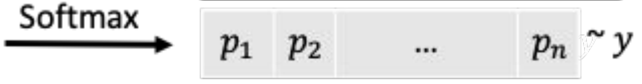
Empirical Risk Minimization with Lagrange multiplier:

$$\pi^* = \mathrm{argmax}_\pi \frac{1}{n} \sum_{i=1}^{n} U(\pi|x_i) - \lambda \cdot D(\pi|x_i)$$

[Singh & Joachims, NeurIPS 2019]

Stochastic Ranking Policy (π)



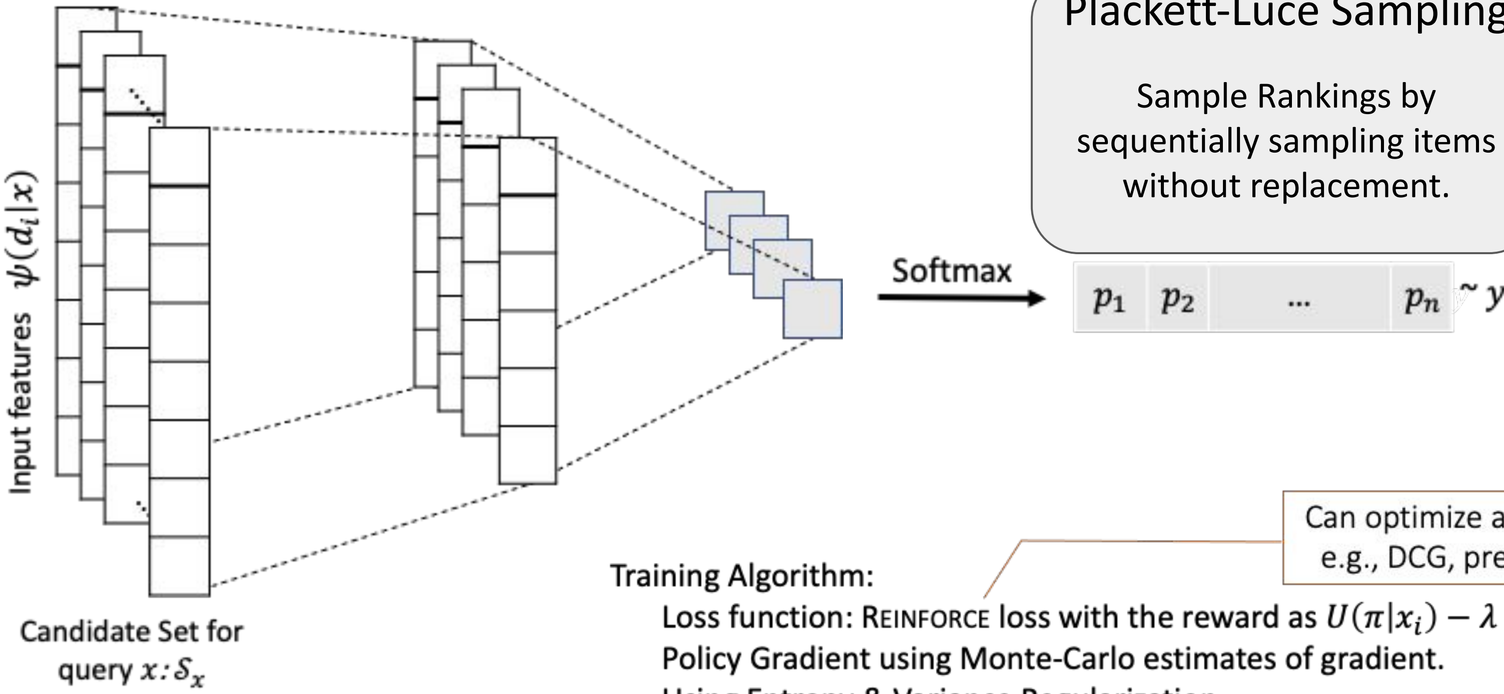


Plackett-Luce Sampling

Sample Rankings by sequentially sampling items without replacement.

[Singh & Joachims, NeurIPS 2019]

Stochastic Ranking Policy (π)



Plackett-Luce Sampling

Sample Rankings by sequentially sampling items without replacement.

Can optimize an arbitrary metric U, e.g., DCG, prec@k, average-rank.

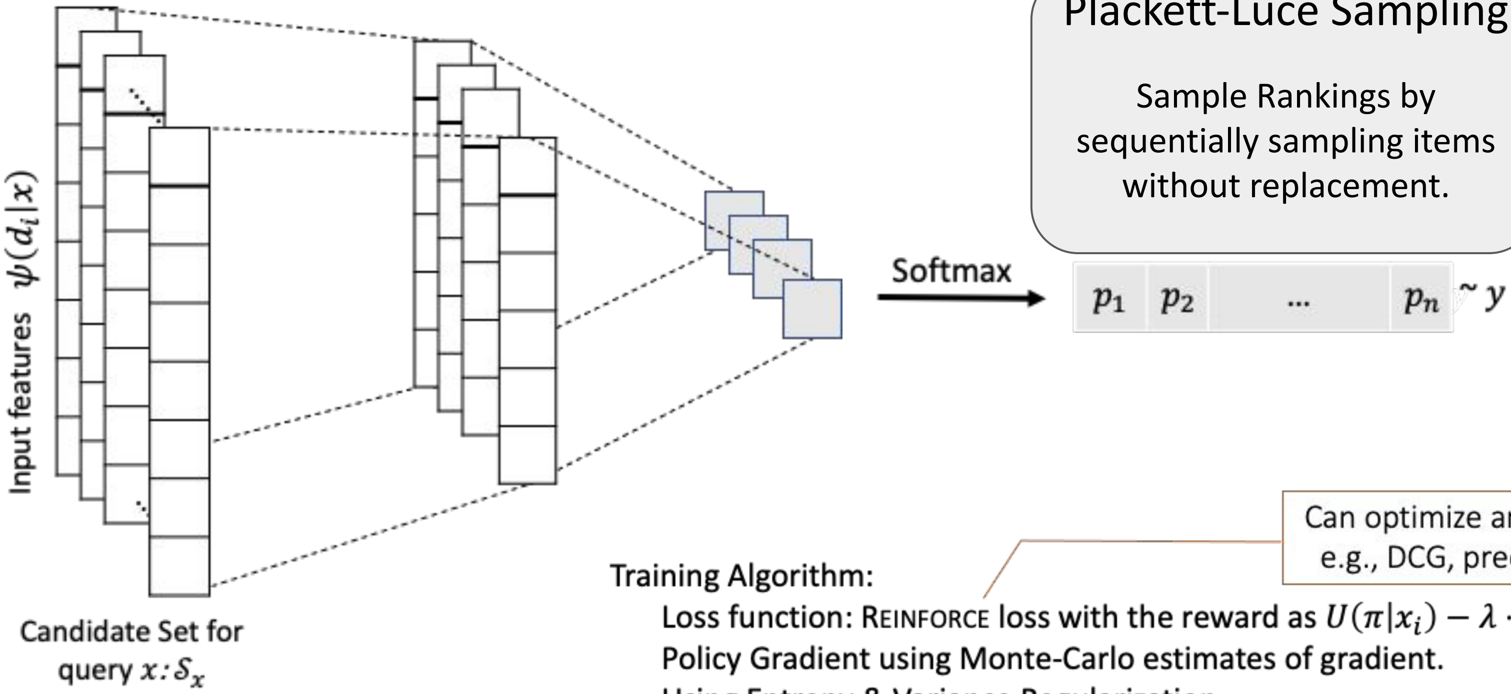
Training Algorithm:

- Loss function: REINFORCE loss with the reward as $U(\pi|x_i) - \lambda \cdot D(\pi|x_i)$.
- Policy Gradient using Monte-Carlo estimates of gradient.
- Using Entropy & Variance Regularization.

[Singh & Joachims, NeurIPS 2019]

Stochastic Ranking Policy (π)

Sequentially sampling one item at a time is slow in practice.



Plackett-Luce Sampling

Sample Rankings by sequentially sampling items without replacement.

Can optimize an arbitrary metric U, e.g., DCG, prec@k, average-rank.

Training Algorithm:

Loss function: REINFORCE loss with the reward as $U(\pi|x_i) - \lambda \cdot D(\pi|x_i)$.

Policy Gradient using Monte-Carlo estimates of gradient.

Using Entropy & Variance Regularization.

[Singh & Joachims, NeurIPS 2019]

Learning-to-Rank with Stochastic Rankings

Sequential sampling to construct a ranking can be expensive, and policy gradient updates can have high variance.

1. Reparametrize the probability distribution by adding independently drawn noise samples G_i from a Gumbel distribution

$$\tilde{p}(d_i) = \frac{\exp\left(\mathrm{y}_{d_i} + G_i\right)}{\sum_{d_j \in \mathcal{D}} \exp\left(\mathrm{y}_{d_j} + G_j\right)}$$

2. Sort by $\tilde{p}(d_i)$ to obtain a ranking.

Can be used for learning as well as deploying stochastic rankings.

[Diaz et al. CIKM 2020; Bruch et al. WSDM 2020]

Part 3: Outline

How to train a fair recommender system?

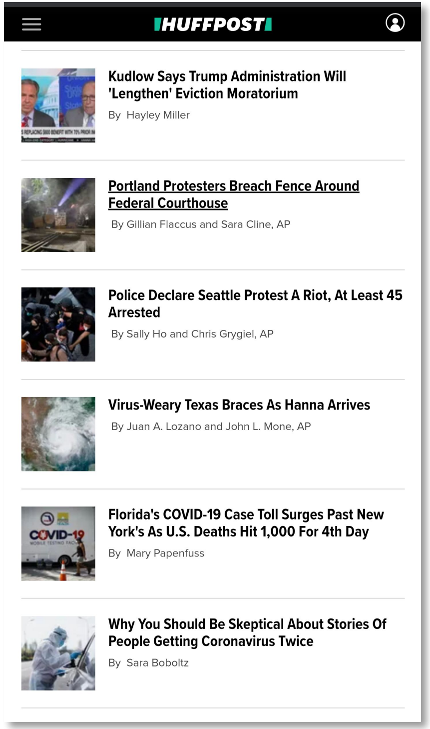
- Collaborative Filtering
- Learning-to-Rank
- Online Learning, Contextual bandits, Sequential decision making (RL)

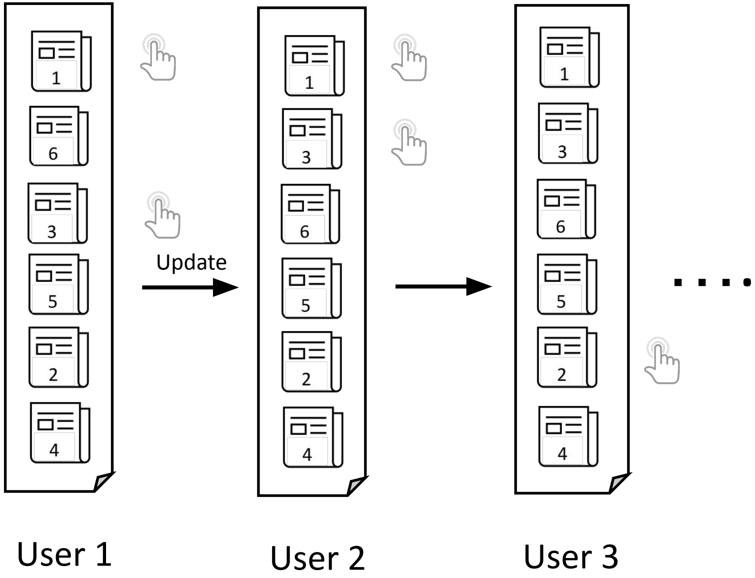
X

- Selection Bias
- User Fairness
- Item Fairness
- Multistakeholder perspective
- Feedback loops

Dynamic Learning-to-Rank

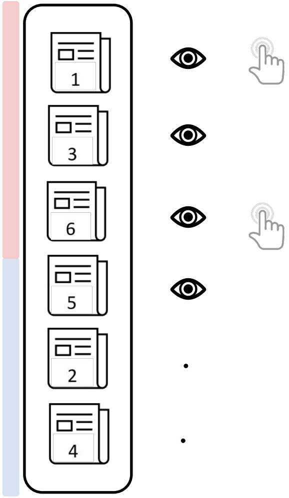
How to train a ranking policy that adapts the ranking to user interactions?





[Morik*, Singh*, Hong & Joachims. SIGIR 2020]

Dynamic Learning-to-Rank



Position Bias

Problem 1: Selection bias due to position

- Click count is not a consistent estimator of relevance.
  - Lower positions get lower attention.
  - Less attention means fewer clicks.
- Click feedback is **biased** by:
  - the deployed ranking function
  - user's position bias

Rich-get-richer dynamic: What starts at the bottom has little opportunity to rise in the ranking.

Problem 2: Exposure disparity between groups

- Ranking solely by relevance may cause some groups to get most of the exposure on the platform.
  - For the news homepage example, this may make the platform seem biased.

[Morik, Singh*, Hong & Joachims. SIGIR 2020]*

Estimating Relevance from Clicks

Question: Clicks → Relevance?

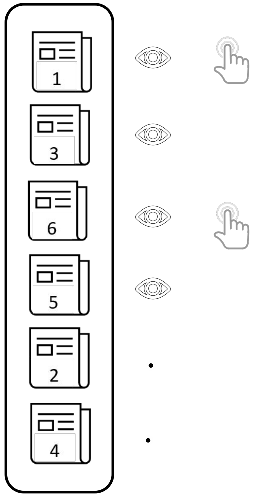
Key Idea [Joachims 2017]: Understand the Observation Mechanism.

Assume a Position-based Model:

$$click(d) = 1 \leftrightarrow (obs(d) = 1) \wedge (rel(d) = 1)$$

Problem:

$$click(d) = 0 \leftrightarrow \color{red}{(obs(d) = 0)} \vee (rel(d) = 0)$$



Estimating Relevance from Clicks

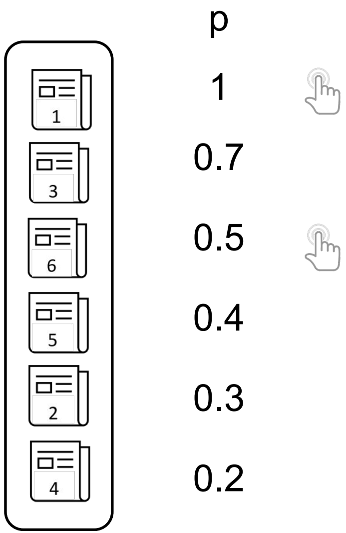
Propensity: $p(d) = P[obs(rank(d)) = 1 \mid y]$

- Can use position-based exposure e_j as an estimate.

Inverse Propensity Score (IPS) Weighting

$$\hat{R}_\tau^{IPS}(d) = \frac{1}{\tau}\sum_{i=1}^{\tau}\frac{click_t(d)}{p_t(d)}.$$

Unbiased estimator of relevance



[Joachims et al. WSDM 2017]

Estimating Relevance from Clicks

$$\mathcal{L}^c(w) = \sum_{t=1}^{\tau} \sum_{d} \hat{R}^w(d|x_t)^2 + \frac{c_t(d)}{p_t(d)}\left(c_t(d) - 2\,\hat{R}^w(d|x_t)\right)$$

- Train a neural network by minimizing $\mathcal{L}^c(\mathrm{w})$.
- $\mathcal{L}^c(\mathrm{w})$ is unbiased i.e., in expectation, $\mathcal{L}^c(\mathrm{w})$ is equal to a full information squared loss (with no position bias).

[Morik*, Singh*, Hong & Joachims. SIGIR 2020]

Fairness Controller (FairCo) LTR Algorithm

FairCo: Ranking at time τ for query x

$$\sigma_\tau = \mathrm{argsort}_{d \in \mathcal{D}} \left(\hat{R}(d|x) + \lambda \, \mathrm{err}_\tau(d) \right)$$

P-Controller:
Linear feedback control system where correction is proportional to the error.

$\hat{R}(d|x)$: Estimated Conditional Relevance

$\lambda > 0$

$err_\tau(d) = (\tau - 1) \max_{G_i}(\hat{D}_\tau^E(G_i, G(d)))$

Handles Selection Bias (Problem 1)

Handles Exposure Disparity (Problem 2)

[Morik, Singh*, Hong & Joachims. SIGIR 2020]*

Part 3: Outline

How to train a fair recommender system?

- Collaborative Filtering
- Learning-to-Rank
- Online Learning, Contextual bandits, Sequential decision making (RL)

X

- Selection Bias
- User Fairness
- Item Fairness
- Multistakeholder perspective
- Feedback loops

X

- Evaluation
- Pre-processing
- In-processing
- Post-processing

Part 3: Outline

How to train a fair recommender system?

- Collaborative Filtering
- Learning-to-Rank
- Online Learning, Contextual bandits, Sequential decision making (RL)

X

- Selection Bias
- User Fairness
- Item Fairness
- Multistakeholder perspective
- Feedback loops

X

- Evaluation
- Pre-processing
- In-processing
- Post-processing

However, real world recommender systems have other complexities that affect the applicability of these approaches.

Practical Recommender Systems

- ↪ Fairness under composition
- ↪ Two-stage recommender systems
- ↪ Repeated Training

Practical Recommender Systems

↪ **Fairness under composition**

- Real world recommender systems are composed of multiple models trained separately.
- Composition of fair models may not lead to a fair model.
- **Goal**: make the end-ranking meet fairness goals.

Even if two predictors are fair, the composition of their predictions can still be unfair.
[Fairness under Composition, *Dwork and Ilvento, ITCS 2019*]

Example: $E[\text{rating}] = P(\text{click}) \times E[\text{rating}|\text{click}] = pCTR \times pRating.$

Component	Author demographics: non-white	non-white	white	white
$pCTR$	0.1	0.4	0.2	0.3
$pRating$	0.4	0.1	0.3	0.2
$pCTR \times pRating$	0.04	0.04	0.06	0.06

Ranking by *pCTR* or *pRating* leads to <*nw, w, w, nw*>, but ranking by their product leads to <*w, w, nw, nw*>.

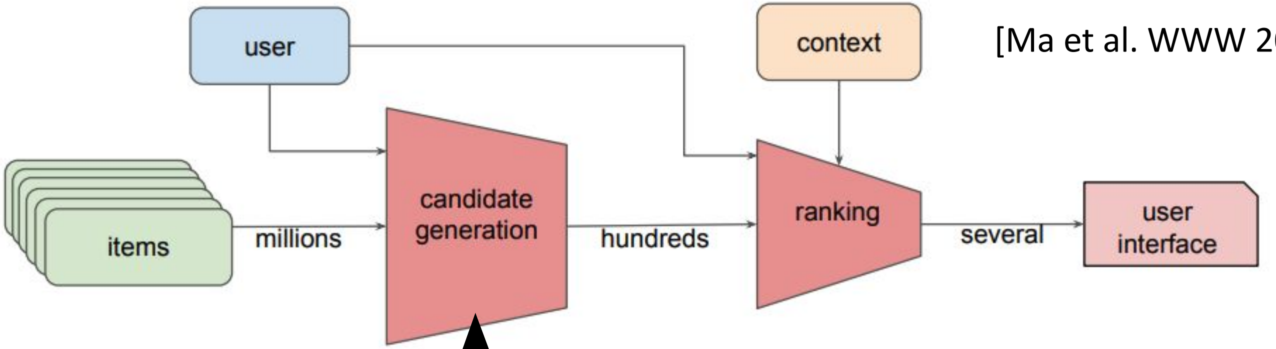
[Wang et al. WSDM 2021]

Practical Recommender Systems

↪ Fairness under composition
↪ **Two-stage recommender systems**

Two stage Recommender systems:

- Candidate generation → Ranking (→ User)



[Ma et al. WWW 2020]

Lack of diversity at candidate generation may lead to unfair results overall

[Wang & Joachims. *forthcoming*. 2022]

Practical Recommender Systems

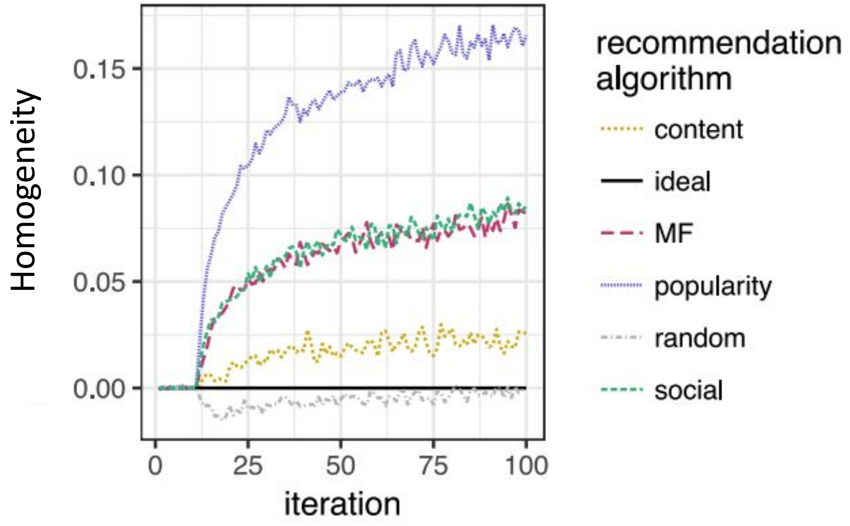
↪ Fairness under composition
↪ Two-stage recommender systems
↪ **Repeated Training**

Models undergo repeated training (daily, weekly, monthly).

Retraining is done using data that is confounded by algorithmic recommendations from a previously deployed system.

Consequences:

- “The recommendation feedback loop causes **homogenization of user behavior**”
- “Users experience **losses in utility** due to homogenization effects; these losses are **distributed unequally**”
- “The feedback loop **amplifies the impact of recommendation systems** on the distribution of item consumption”

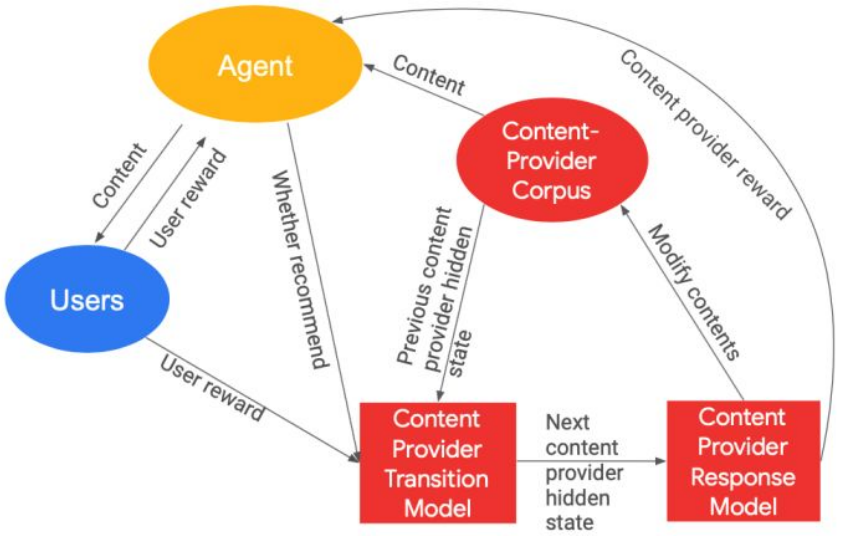


Homogeneity of content recommended increases with repeated training.

[Chaney et al. RecSys 2018]

Fairness in Sequential Recommender Systems

- Sequential Recommender Systems such as RL based systems may need to consider
  - content provider dynamics in addition to user dynamics.
  - optimize for long term content provider reward.



[“Towards Content Provider Aware Recommender Systems”, Zhan et al. WWW’21]

Challenges and Open Questions

- Open Questions:
  - How do users and item providers experience and perceive “unfairness”?
  - Maintaining legality: How can we ensure group fairness without violating constraints around model inputs (e.g. without using protected attributes)?
- What did we not cover but is also important?
  - Privacy
  - User safety and trust
  - Explainability and transparency

Thank you

Fair and Socially Responsible ML for Recommendations

https://fair-recs-tutorial.github.io/neurips-2022-tutorial/

Hannah Korevaar
Research Scientist, Meta

Manish Raghavan
Assistant Professor, MIT

Ashudeep Singh
Applied Scientist, Pinterest

References

[1] Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan. 2022. The Challenge of Understanding What Users Want: Inconsistent Preferences and Engagement Optimization. arXiv:2202.11776. https://arxiv.org/abs/2202.11776

[2] Jonathan Roth, Guillaume Saint-Jaques, Yinyin Yu. 2021. An Outcome Test of Discrimination for Ranked Lists. arXiv:2111.08779v1. https://arxiv.org/abs/2111.07889

[3] Tomo Lazovich, Luca Belli, Aaron Gonzales, Amanda Bower, Uthaipon Tantipongpipat, Kristian Lam, Ferenc Huszar, Rumman Chowdhury. 2022. Measuring Disparate Outcomes of Content Recommendation Algorithms with Distributional Inequality Metrics. arXiv:2022.01615v1. https://arxiv.org/abs/2202.01615

[4] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichen Hong, Ed H. Chi, Cristos Goodrow. Fairness in Recommendation Ranking through Pairwise Comparisons. 2019. In Proceedings of KDD ‘19. https://arxiv.org/pdf/1903.00780.pdf

[5] Lequn Wang, Thorsten Joachims. User Fairness, Item Fairness, and Diversity for Rankings in Two-Sided Markets. 2021. arXiv:2010.01470v3. https://arxiv.org/abs/2010.01470

[6] Ashudeep Singh, Thorsten Joachims. Fairness of Exposure in Rankings. 2018. In Proceedings of KDD ‘18. https://www.cs.cornell.edu/~tj/publications/singh_joachims_18a.pdf

[7] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, Ben Carterette. Evaluating Stochastic Rankings with Expected Exposure. arXiv:2004.13157

[8] Matthew J. Salganik, Peter Sheridan Dodds, Duncan J. Watts. Experimental Study of Inequality in an Online Cultural Market. 2006. Science. DOI: 10.1126/science.1121066. https://www.science.org/doi/10.1126/science.1121066

References

[9] Camelia Simoiu, Sam Corbett-Davies, Sharad Goel. 2016. The Problem of Infra-marginality in Outcome Tests for Discrimination. arXiv:1607.05376. https://doi.org/10.48550/arXiv.1607.05376

[10] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, and Moritz Hardt. Algorithmic amplification of politics on Twitter. 2021. PNAS. 119(1)e2025334119. https://doi.org/10.1073/pnas.2025334119

[11] Ashudeep Singh, David Kempe, Thorsten Joachims. 2021. NeurIPS Proceedings. Fairness in Ranking Under Uncertainty. https://proceedings.neurips.cc/paper/2021/file/63c3ddcc7b23daa1e42dc41f9a44a873-Paper.pdf

[12] Solon Barocas, Moritz Hardt, Arvind Narayanan. 2019. Fairness and Machine Learning. Fairmlbook.org. http://www.fairmlbook.org

[13] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshiro Kamishima, Jan Krasnodebski, Luiz Pizzato. 2019. Beyond Personalization: Research Directions in Multistakeholder Recommendation. arXiv:1905.01986v2. https://arxiv.org/pdf/1905.01986.pdf

[14] Asia J. Biega, Krishna P. Gummadi, Gerhard Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. arXiv:1805.01788

[15] Ferraro, A., Ferreira, G., Diaz, F., & Born, G. (2022). Measuring Commonality in Recommendation of Cultural Content: Recommender Systems to Enhance Cultural Citizenship. *arXiv preprint arXiv:2208.01696*.

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