DS-GA 3001.009: Responsible Data Science

Algorithmic Fairness (continued)

Prof. Julia Stoyanovich
Center for Data Science
Computer Science and Engineering at Tandon

@stoyanoj

http://stoyanovich.org/
https://dataresponsibly.github.io/
https://dataresponsibly.github.io/courses/spring20/

fairness in risk assessment

New Jersey bail reform

THE NEW JERSEY PRETRIAL JUSTICE MANUAL

6 or subjected to onerous conditions of release.

Switching from a system based solely on instinct and experience (often referred to as "gut instinct") to one in which judges have access to scientific, objective risk assessment tools could further the criminal justice system's central goals of increasing public safety, reducing crime, and making the most effective, fair, and efficient use of public resources.

Risk Assessment and Release/Detention Decision Making in New Jersey

DECEMBER 2016

ACLU AMERICAN CIVIL LIBERTIES UNION of NEW JERSEY

National Association of Criminal Defense Lawyers NACDL 1958

NEW JERSEY PUBLIC DEFENDER

Fairness in risk assessment

- A risk assessment tool **gives a probability estimate of a future outcome**
- Used in many domains:
  - insurance, criminal sentencing, medical testing, hiring, banking
  - also in less-obvious set-ups, like online advertising
- **Fairness** is concerned with **how different kinds of error are distributed among sub-populations**
  - Recall our discussion on fairness in classification - similar?

Racial bias in criminal sentencing

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

A commercial tool **COMPAS** automatically predicts some categories of future crime to assist in bail and sentencing decisions. It is used in courts in the US.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Desirable properties of risk tools

[J. Kleinberg, S. Mullainathan, M. Raghavan; ITCS *(2017)*]

"risk assessment tool / instrument" = "**risk tool / instrument**"
for brevity in the rest of today's slides

- Calibration
- Balance for the positive class
- Balance for the negative class

can we have all these properties?

Calibration

positive outcomes: do recidivate

	risk score		
	0.2	**0.6**	**0.8**
white	⊖ ⊖ ⊖ ⊖ ⊕	⊖ ⊖ ⊕ ⊕ ⊕	⊖ ⊕ ⊕ ⊕ ⊕
black	⊖ ⊖ ⊖ ⊖ ⊕	⊖ ⊖ ⊖ ⊖ ⊕ ⊕ ⊕ ⊕ ⊕ ⊕	⊖ ⊕ ⊕ ⊕ ⊕

given the output of a risk tool, likelihood of belonging to the positive class is independent of group membership

0.6 means 0.6 for any defendant - likelihood of recidivism

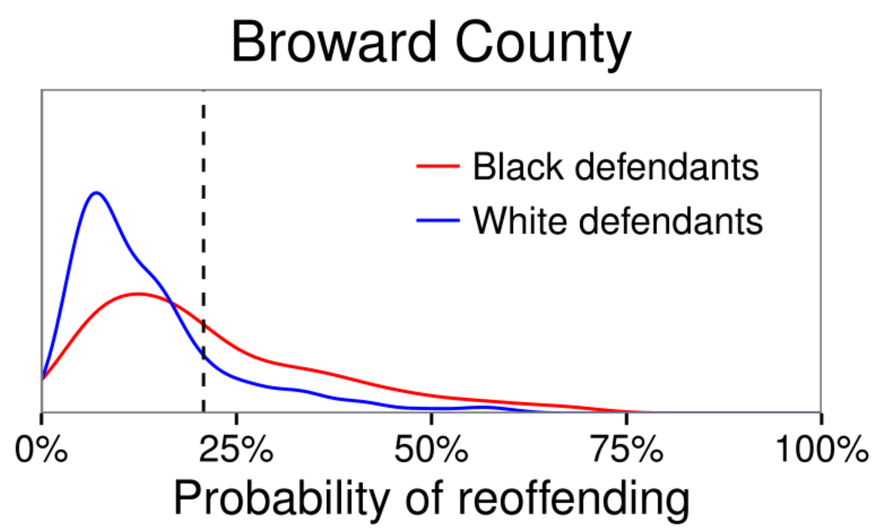
why do we want calibration?

Calibration in COMPAS

[J. Kleinberg, S. Mullainathan, M. Raghavan; *ITCS 2017*]

Predictive parity (also called **calibration**)
an risk tool identifies a set of instances as having probability *x* of constituting positive instances, then approximately an *x* fraction of this set are indeed positive instances, over-all and in sub-populations

COMPAS is **well-calibrated**: in the window around 40%, the fraction of defendants who were re-arrested is ~40%, both over-all and per group.



[plot from Corbett-Davies et al.; *KDD 2017*]

Balance

[J. Kleinberg, S. Mullainathan, M. Raghavan; *ITCS 2017*]

- **Balance for the positive class:** Positive instances are those who go on to re-offend. The average score of positive instances should be the same across groups.
- **Balance for the negative class:** Negative instances are those who do not go on to re-offend. The average score of negative instances should be the same across groups.
- Generalization of: **Both groups should have equal false positive rates and equal false negative rates.**
- Different from statistical parity!

the chance of making a mistake does not depend on race

Desiderata, re-stated

[J. Kleinberg, S. Mullainathan, M. Raghavan; ITCS *(2017)*]

- For each group, a $\mathbf{v_b}$ fraction in each bin **b** is positive
- Average score of positive class same across groups
- Average score of negative class same across groups

can we have all these properties?

Achievable only in trivial cases

[J. Kleinberg, S. Mullainathan, M. Raghavan; ITCS *(2017)*]

- **Perfect information:** the tool knows who recidivates (score 1) and who does not (score 0)
- **Equal base rates:** the fraction of positive-class people is the same for both groups

cannot even find a good approximate solution

a negative result, need tradeoffs

proof sketched out in (starts 12 min in)

https://www.youtube.com/watch?v=UUC8tMNxwV8

Group fairness impossibility result

[A. Chouldechova; arXiv:1610.07524v1 *(2017)*]

If a predictive instrument **satisfies predictive parity**, but the **prevalence** of the phenomenon **differs between groups**, then the instrument **cannot achieve** equal false positive rates and equal false negative rates across these groups

Recidivism rates in the ProPublica dataset are higher for the black group than for the white group

https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

What is recidivism?: Northpointe [*the maker of COMPAS*] defined recidivism as "**a finger-printable arrest** involving a charge and a filing for any uniform crime reporting (UCR) code."

Fairness for whom?

based on a slide by Arvind Narayanan

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

Defendant: how likely am I to be incorrectly classified high-risk?

Society: (think positive interventions) is the selected set demographically balanced?

	labeled low-risk	labeled high-risk
did not recidivate	TN	**FP**
recidivated	**FN**	TP

different metrics matter to different stakeholders

https://www.propublica.org/article/propublica-responds-to-companys-critique-of-machine-bias-story

Impossibility theorem

based on a slide by Arvind Narayanan

Metric	Equalized under
Selection probability	Demographic parity
Pos. predictive value	Predictive parity
Neg. predictive value	
False positive rate	Error rate balance
False negative rate	Error rate balance
Accuracy	Accuracy equity

Chouldechova paper

All these metrics can be expressed in terms of FP, FN, TP, TN

If these metrics are equal for 2 groups, some trivial algebra shows that the prevalence (in the COMPAS example, of recidivism, as measured by re-arrest) is also the same for 2 groups

Nothing special about these metrics, can pick any 3!

Ways to evaluate binary classifiers

based on a slide by Arvind Narayanan

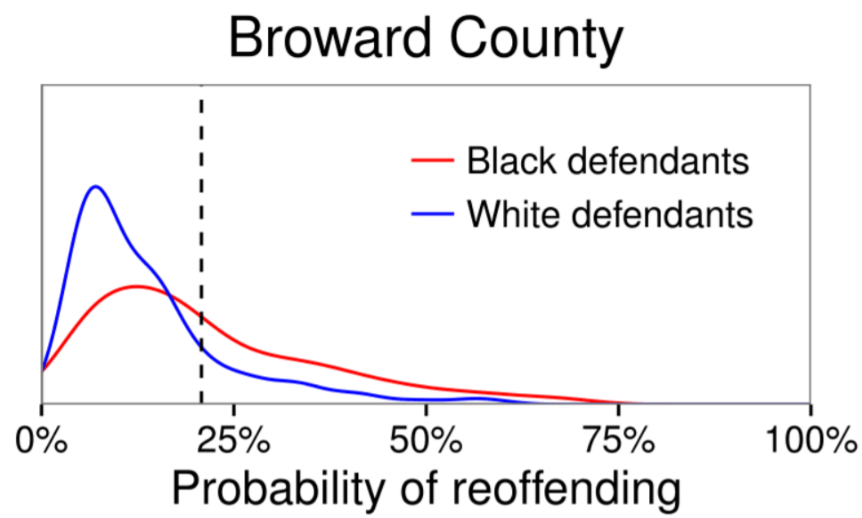
<table>
<tr><td></td><td></td><td colspan="2">True condition</td><td></td><td colspan="2"></td></tr>
<tr><td></td><td>Total population</td><td>Condition positive</td><td>Condition negative</td><td>Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$</td><td colspan="2">Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$</td></tr>
<tr><td rowspan="2">Predicted condition</td><td>Predicted condition positive</td><td>True positive, Power</td><td>False positive, Type I error</td><td>Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$</td><td colspan="2">False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$</td></tr>
<tr><td>Predicted condition negative</td><td>False negative, Type II error</td><td>True negative</td><td>False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$</td><td colspan="2">Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$</td></tr>
<tr><td></td><td></td><td>True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$</td><td>False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$</td><td>Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$</td><td rowspan="2">Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$</td><td rowspan="2">$F_1$ score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$</td></tr>
<tr><td></td><td></td><td>False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$</td><td>True negative rate (TNR), Specificity (SPC) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$</td><td>Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$</td></tr>
</table>

364 impossibility theorems :)

Individual fairness

based slides by Arvind Narayanan

Individual fairness: assuming scores are calibrated, we cannot pick a single threshold for 2 groups that equalizes both the False Positives Rate and the False Negatives Rate



What's the right answer?

There is no single answer!

Need transparency and public debate

- Consider harms and benefits to different stakeholders
- Being transparent about which fairness criteria we use, how we trade them off
- Recall "Learning Fair Representations": a typical ML approach

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

group fairness **individual fairness** **utility**

apples + oranges + fairness = ?

causal interpretations of fairness

Effect on sub-populations

Simpson's paradox
disparate impact at the full population level disappears or reverses when looking at sub-populations!

		grad school admissions	
		admitted	**denied**
gender	**F**	1512	2809
	M	3715	4727

positive outcomes

35% of women

44% of men

UC Berkeley 1973: it appears men were admitted at higher rate.

Effect on sub-populations

Simpson's paradox
disparate impact at the full population level disappears or reverses when looking at sub-populations!

Department	Men		Women		favored group
	Applicants	Admitted	Applicants	Admitted	
A	825	62%	108	**82%**	women
B	560	63%	25	**68%**	women
C	325	**37%**	593	34%	men
D	417	33%	375	**35%**	women
E	191	**28%**	393	24%	men
F	373	6%	341	**7%**	women

whole population

35% of women

44% of men

UC Berkeley 1973: women applied to more competitive departments, with low rates of admission among qualified applicants.

Correlation is not causation!

Cannot claim a causal relationship based on observational data alone. Need a story.

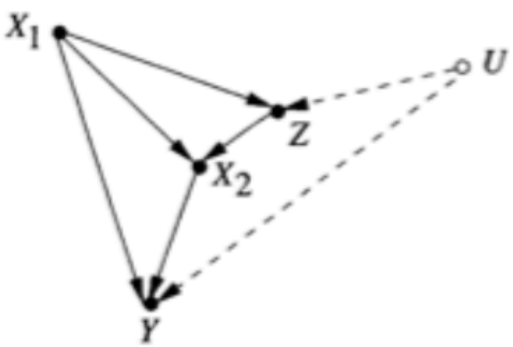


Figure 4.9 Causal relationships relevant to Berkeley's sex discrimination study. Adjusting for department choice (X_2) or career objective (Z) (or both) would be inappropriate in estimating the direct effect of gender on admission. The appropriate adjustment is given in (4.10).

X2 (choice) - "resolving variable", then the effect of X1 on Y through X2 is "fair"

the direct effect of X1 on Y is unfair

X_1 = applicant's gender;
X_2 = applicant's choice of department;
Z = applicant's (pre-enrollment) career objectives;
Y = admission outcome (accept/reject);
U = applicant's aptitude (unrecorded).

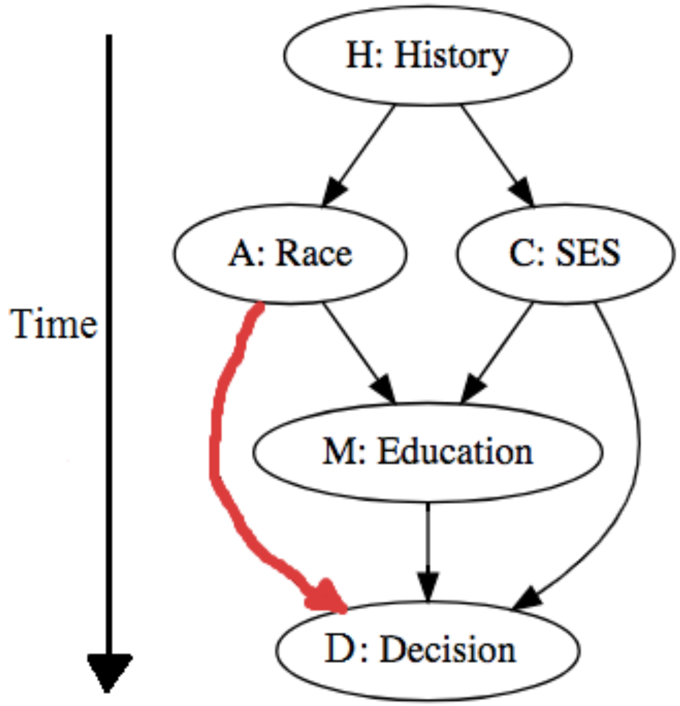
from Pearl's "Causality", page 129

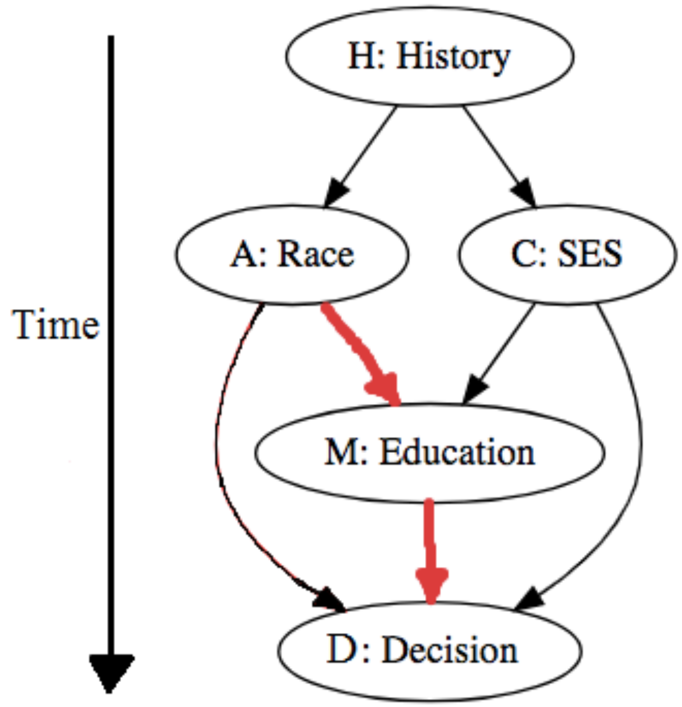
Note that U affects applicant's career objective and also the admission outcome Y (say, through verbal skills (unrecorded)).

Causal interpretations of fairness

[T.J. VanderWeele and W.R. Robinson; Epidemiology *(2014)*]

arrows represent possible causal relationships





we (society) decide which of these are "OK"

fairness in ranking

Fairness in ranking

[K. Yang & J. Stoyanovich, FATML (2016)]

Input: database of items (individuals, colleges, cars, …)

Score-based ranker: computes the score of each item using **known formula**, then sorts items on score

Output: permutation of the items (complete or top-k)

id	sex	race	age	cat
a	F	W	25	T
b	F	B	23	S
c	M	W	27	T
d	M	B	45	S
e	M	W	60	U

ranker

What is a positive outcome in a ranking?

Idea: Rankings are relative, fairness measures should be rank-aware

The order of things

THE NEW YORKER

DEPT. OF EDUCATION FEBRUARY 14 & 21, 2011 ISSUE

THE ORDER OF THINGS

What college rankings really tell us.

By Malcolm Gladwell

1. Chevrolet Corvette 205
2. Lotus Evora 195
3. Porsche Cayman 195

1. Lotus Evora 205
2. Porsche Cayman 198
3. Chevrolet Corvette 192

1. Porsche Cayman 193
2. Chevrolet Corvette 186
3. Lotus Evora 182

Rankings are not benign!

THE NEW YORKER

DEPT. OF EDUCATION FEBRUARY 14 & 21, 2011 ISSUE

THE ORDER OF THINGS

What college rankings really tell us.

By Malcolm Gladwell

Rankings are not benign. They enshrine very particular ideologies, and, at a time when American higher education is facing a crisis of accessibility and affordability, we have adopted **a de-facto standard of college quality** that is uninterested in both of those factors. And why? Because a group of magazine analysts in an office building in Washington, D.C., decided twenty years ago to **value selectivity over efficacy**, to **use proxies** that scarcely relate to what they're meant to be proxies for, and to **pretend that they can compare** a large, diverse, low-cost land-grant university in rural Pennsylvania with a small, expensive, private Jewish university on two campuses in Manhattan.

Location-location-location

[K. Yang & J. Stoyanovich, FATML (2016)]

gender is the sensitive attribute, input is balanced

Algorithm 1 Ranking generator

Require: Ranking $\boldsymbol{\tau}$, fairness probability f.
{Initialize the output ranking $\boldsymbol{\sigma}$.}

$\boldsymbol{\sigma} \leftarrow \emptyset$
$\boldsymbol{\tau}^{+} = \boldsymbol{\tau} \cap S^{+}$
$\boldsymbol{\tau}^{-} = \boldsymbol{\tau} \cap S^{-}$
while $(\boldsymbol{\tau}^{+} \neq \emptyset) \wedge (\boldsymbol{\tau}^{-} \neq \emptyset)$ **do**
  $p = random([0, 1])$
  if $p < f$ **then**
    Pop an item from the top of the list $\boldsymbol{\tau}^{+}$.
    $\boldsymbol{\sigma} \leftarrow pop(\boldsymbol{\tau}^{+})$
  else
    Pop an item from the top of the list $\boldsymbol{\tau}^{-}$.
    $\boldsymbol{\sigma} \leftarrow pop(\boldsymbol{\tau}^{-})$
  end if
end while
$\boldsymbol{\sigma} \leftarrow \boldsymbol{\tau}^{+}$
$\boldsymbol{\sigma} \leftarrow \boldsymbol{\tau}^{-}$
return $\boldsymbol{\sigma}$

rank	gender
1	M
2	M
3	M
4	M
5	M
6	F
7	F
8	F
9	F
10	F

$f = 0$

rank	gender
1	M
2	M
3	F
4	M
5	M
6	F
7	M
8	F
9	F
10	F

$f = 0.3$

rank	gender
1	M
2	F
3	M
4	F
5	M
6	F
7	M
8	F
9	M
10	F

$f = 0.5$

parity in outcomes

Rank-aware fairness

[K. Yang & J. Stoyanovich, FATML (2016)]

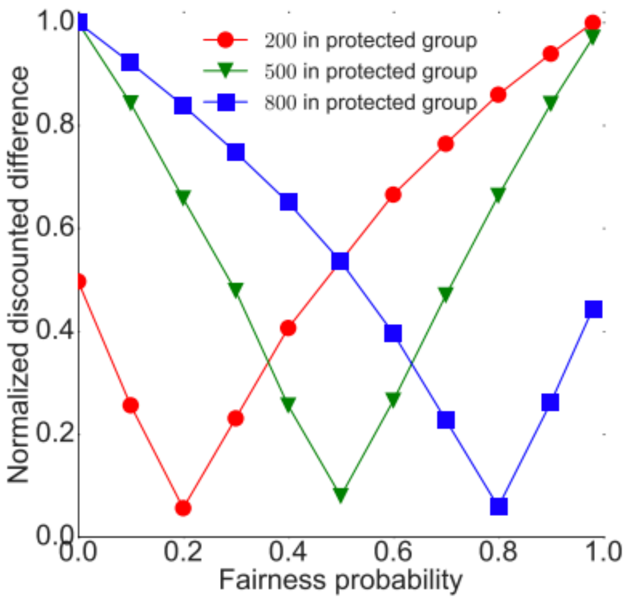


Figure 3: rND on 1,000 items

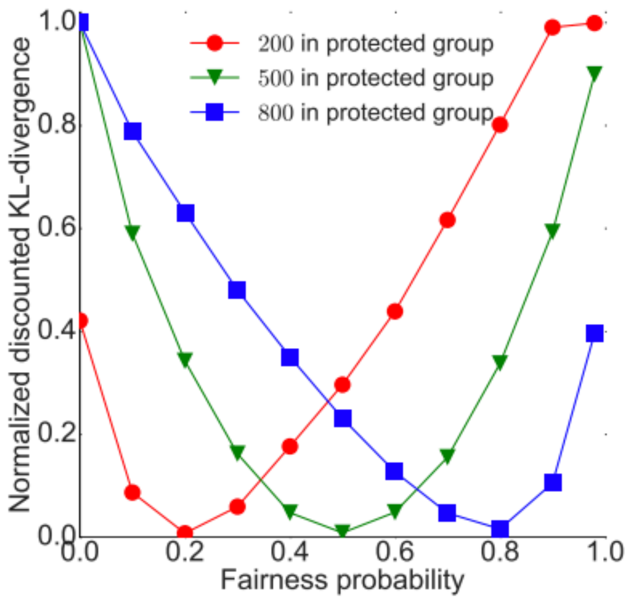


Figure 4: rKL on 1,000 items

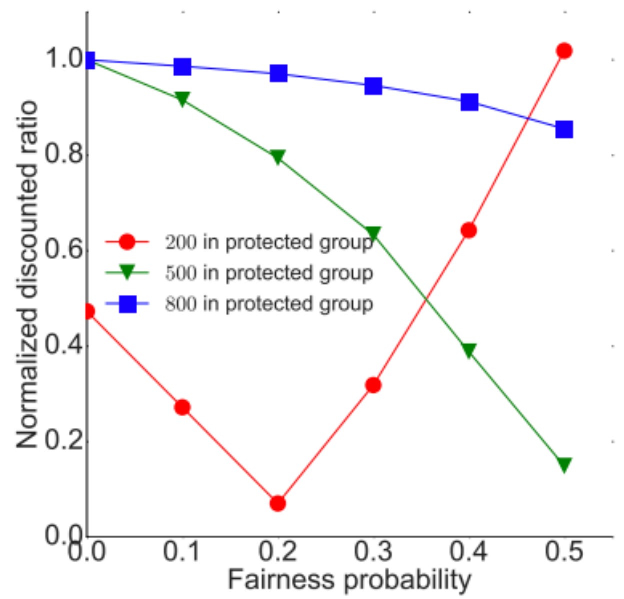


Figure 5: rRD on 1,000 items

In an optimization framework

[K. Yang & J. Stoyanovich, FATML (2016)]

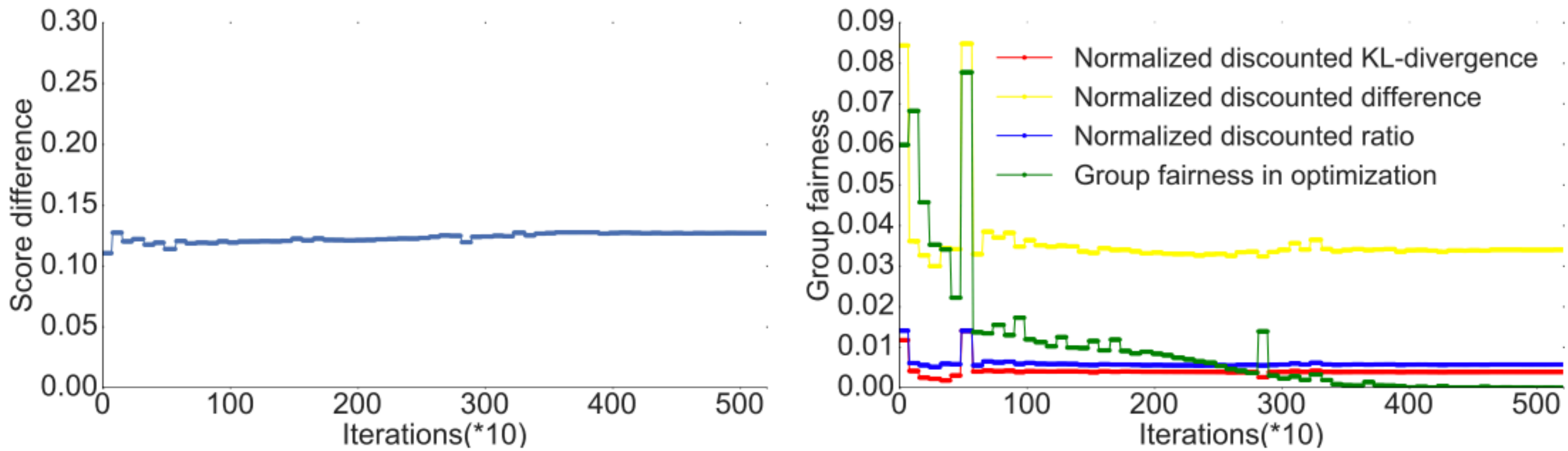
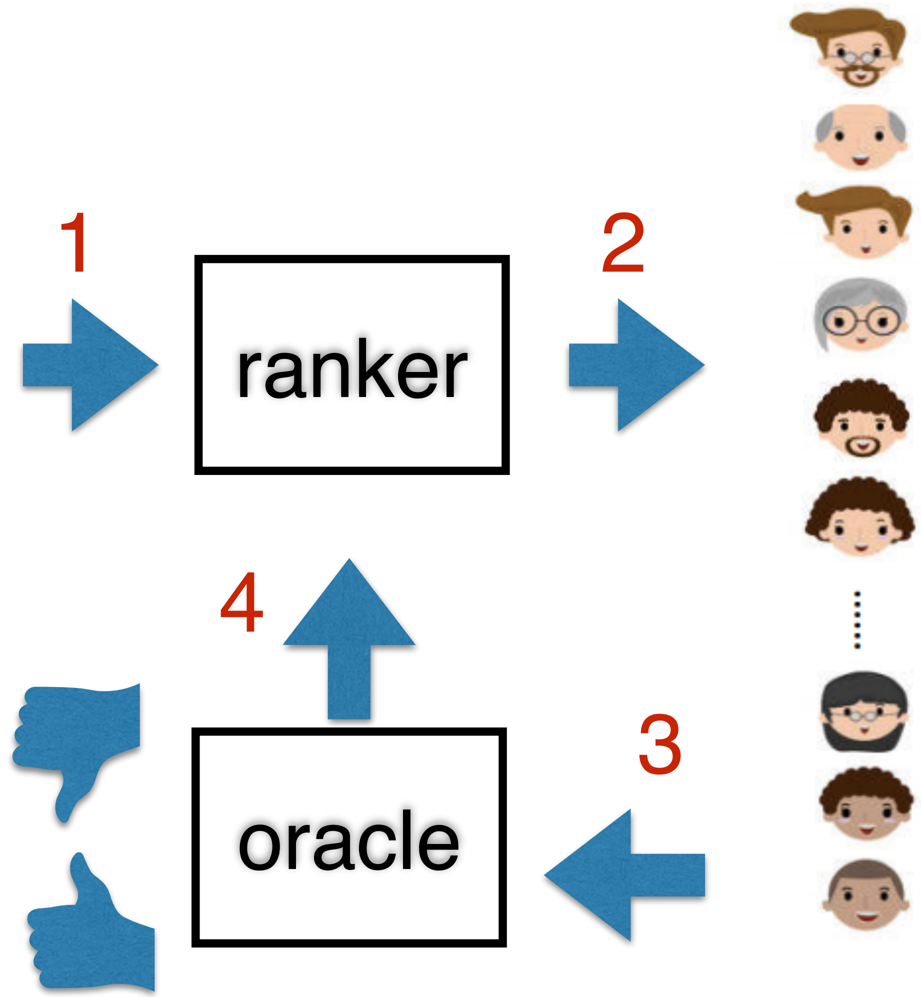


Figure 6: Accuracy and fairness on German Credit, ranked by *sum of normalized attribute values*, with $k = 10$.

Designing fair rankers

[A. Asudeh, HV Jagadish, J. Stoyanovich, G. Das; ACM SIGMOD *(2019)*]

id	sex	race	age	cat
a	F	W	25	T
b	F	B	23	S
c	M	W	27	T
d	M	B	45	S
e	M	W	60	U



More fairness in ranking

Designing Fair Ranking Schemes

Abolfazl Asudeh†, H. V. Jagadish†, Julia Stoyanovich‡, Gautam Das††
†University of Michigan, ‡Drexel University, ††University of Texas at Arlington
†{asudeh, jag}@umich.edu, ‡stoyanovich@drexel.edu, ††gdas@uta.edu

ACM SIGMOD 2019

ABSTRACT

Items from a database are often ranked based on a combination of multiple criteria. A user may have the flexibility to accept combinations that weigh these criteria differently, within limits. On the other hand, this choice of weights can greatly affect the fairness of the produced ranking. In this paper, we develop a system that helps users choose criterion weights that lead to greater fairness.

We consider ranking functions that compute the score of each item as a weighted sum of (numeric) attribute values, and then sort items on their score. Each ranking function can be expressed as a vector of weights, or as a point in a multi-dimensional space. For a broad range of fairness criteria, we show how to efficiently identify regions in this space that satisfy these criteria. Using this identification method, our system is able to tell users whether their proposed ranking function satisfies the desired fairness criteria and, if it does not, to suggest the smallest modification that does. We develop user-controllable approximation that and indexing techniques that are applied during preprocessing, and support sub-second response times during the online phase. Our extensive experiments on real datasets demonstrate that our methods are able to find solutions that

impact processes that are directly designed and validated by humans. Perhaps the most immediate example of such a process is a score-based ranker. In this paper we consider the task of *designing a fair score-based ranking scheme*.

Ranking of individuals is ubiquitous, and is used, for example, to establish credit worthiness, desirability for college admissions and employment, and attractiveness as dating partners. A prominent family of ranking schemes are score-based rankers, which compute the score of each individual from some database $\mathcal{D}$, sort the individuals in decreasing order of score, and finally return either the full ranked list, or its highest-scoring sub-set, the top-k. Many score-based rankers compute the score of an individual as a linear combination of attribute values, with non-negative weights. Designing a ranking scheme amounts to selecting a set of weights, one for each feature, and validating the outcome on the database $\mathcal{D}$.

Our goal is to assist the user in designing a ranking scheme that both reflects a user's a priori notion of quality and is fair, in the sense that it mitigates *preexisting bias with respect to a protected feature* that is embodied in the data. In line with prior work [17, 27, 31–33], a protected feature denotes membership of an individual

parity in outcomes

Score-based rankers

[A. Asudeh, HV Jagadish, J. Stoyanovich, G. Das; ACM SIGMOD *(2019)*]

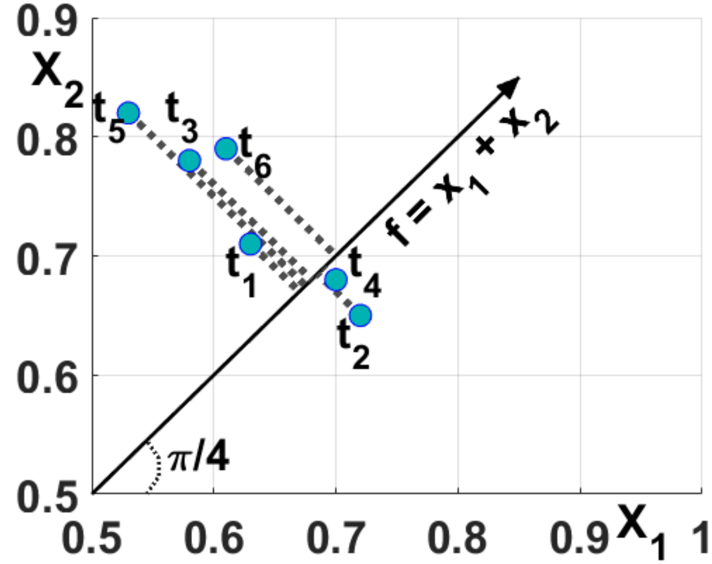
- tuple x in D; score(x): sum of attribute values, with non-negative weights (a common special case of monotone aggregation)
- weights subjectively chosen by a user: 0.5 g+ 0.5s, where g - normalized GPA, s - normalized SAT; why not 0.45 g + 0.55 s?

$\mathcal{D}$			f
id	x_1	x_2	$x_1 + x_2$
t_1	0.63	0.71	1.34
t_2	0.72	0.65	1.37
t_3	0.58	0.78	1.36
t_4	0.7	0.68	1.38
t_5	0.53	0.82	1.35
t_6	0.61	0.79	1.4

Geometry of a (2D) ranker

[A. Asudeh, HV Jagadish, J. Stoyanovich, G. Das; ACM SIGMOD *(2019)*]

$\mathcal{D}$			f
id	x_1	x_2	$x_1 + x_2$
t_1	0.63	0.71	1.34
t_2	0.72	0.65	1.37
t_3	0.58	0.78	1.36
t_4	0.7	0.68	1.38
t_5	0.53	0.82	1.35
t_6	0.61	0.79	1.4



- tuples are points in 2D, scoring functions are rays starting from the origin
- to determine a ranking of the points, we read it off from the projections of the points onto the ray of the scoring function, walking the ray towards the origin
- examples: $f(x) = x_1 + x_2 \quad f(x) = x_1 \quad f(x) = x_2$

Goal: find a satisfactory function

[A. Asudeh, HV Jagadish, J. Stoyanovich, G. Das; ACM SIGMOD *(2019)*]

Closest Satisfactory Function: *Given a dataset $\mathcal{D}$ with n items over d scalar scoring attributes, a fairness oracle $O : \nabla_f(\mathcal{D}) \rightarrow \{\top, \bot\}$, and a linear scoring function f with the weight vector $\vec{w} = \langle w_1, w_2, \cdots, w_d \rangle$, find the function f' with the weight vector $\vec{w}'$ such that $O(\nabla_{f'}(\mathcal{D})) = \top$ and the angular distance between $\vec{w}$ and $\vec{w}'$ is minimized.*

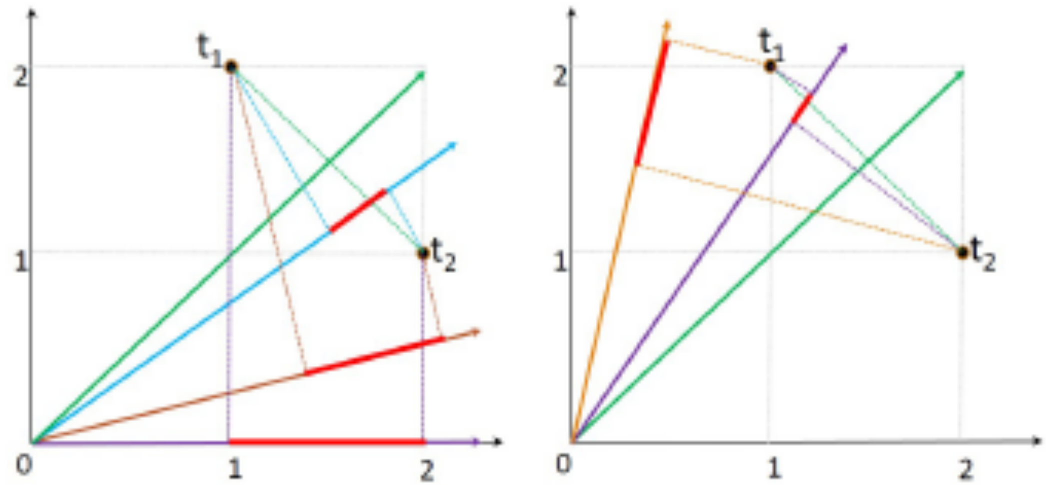
How might we approach this? Why is this difficult?

Ordering exchange

[A. Asudeh, HV Jagadish, J. Stoyanovich, G. Das; ACM SIGMOD *(2019)*]

Key idea: only look at scoring functions that change the relative order between some pair of points. These are the only points where the fairness oracle may change its mind!

$t_1\langle 1,2\rangle \quad t_2\langle 2,1\rangle \qquad t_2 \succ_x t_1 \qquad \boxed{t_2 =_{x+y} t_1} \qquad t_2 \prec_y t_1$



An **ordering exchange** is a set of functions that score a pair of points equally. In 2D, it corresponds to a single function.

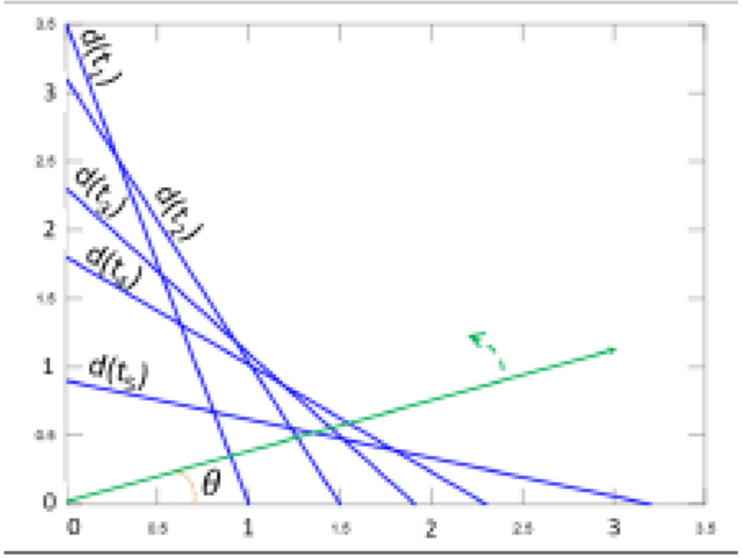
Outline of approach

[A. Asudeh, HV Jagadish, J. Stoyanovich, G. Das; ACM SIGMOD *(2019)*]

t_1	1	3.5
t_2	1.5	3.1
t_3	1.91	2.3
t_4	2.3	1.8
t_5	3.2	0.9

Pre-processing

- Transform the original space into the dual space (in 2D, points become lines)
- Sort points per f(x)=x; compute ordering exchanges between adjacent pairs of points
- Sweep the space with a ray from the x-axis to the y-axis, find satisfactory regions

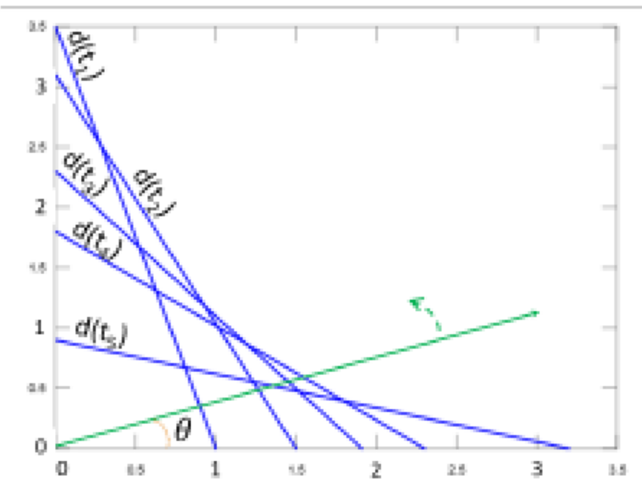


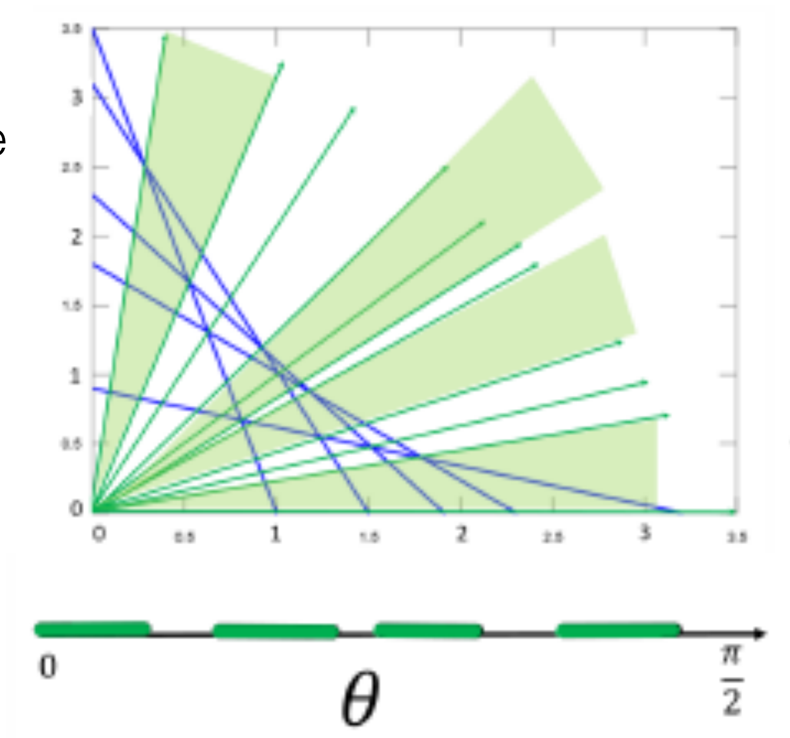
Outline of approach

[A. Asudeh, HV Jagadish, J. Stoyanovich, G. Das; ACM SIGMOD *(2019)*]

At query time

- Look for a satisfactory region closest to the query function
- In 2D, this is simply binary search
- Beyond 2D, everything is hard, and expensive to compute





And lots more algorithmic + systems work

[A. Asudeh, HV Jagadish, J. Stoyanovich, G. Das; ACM SIGMOD *(2019)*]

- Multi-dimensional indexing methods for “arrangement construction”
- Sampling of items (does work), sampling of functions (doesn’t work) to speed up index construction
- Experiments on COMPAS and on US Department of Transportation (DOT) - flights / airlines - datasets

Follow-up work on designing fair ranking functions

Looking at trade-offs

[K. Yang, V. Gkatzelis, J. Stoyanovich, IJCAI (2019)]

Balanced Ranking with Diversity Constraints

Ke Y

[1]New York Un… …gineering

[2]Dre…

ky630@… …du

IJCAI 2019

Abstract

Many set selection and ranking …
cently been enhanced with *diver*…
aim to explicitly increase repres…
cally disadvantaged populations…
overall representativeness of th…
unintended consequence of thes…
ever, is reduced *in-group fairnes*…
didates from a given group m…
ones, and this unfairness may n…
across groups. In this paper …
nomenon using datasets that con…
sitive attributes. We then introd…
straints, aimed at balancing th…
across groups, and formalize the…
tion problems as integer linear…
these programs, we conduct an…
uation with real datasets, and …
ble trade-offs between balance…
mance in the presence of divers…

… sociologists and political sci-
…ki, 2005]. Last but not least,
…ed to ensure dataset represen-
…electing a group of patients to
…edical treatment, or to under-
…edical services [Cohen *et al.*,
…isit in this paper.

…evaluate and mitigate an unin-
…diversity constraints may have
…tion and ranking algorithms.
…at these algorithms do not sys-
… items in particular groups. In
…-up more precise.

… associated with multiple sen-
… a quality score (or utility), a
…o select k of these items aim-
…tility, computed as the sum of
…s. The score of an item is a
…omputed and stored as a phys-
…nputed on the fly. The output
…

parity in outcomes **loss balance**

Ranking with diversity constraints

[K. Yang, V. Gkatzelis, J. Stoyanovich, IJCAI (2019)]

Goal: pick k=4 candidates, including 2 of each gender, and at least one candidate per ethnicity, maximizing the total score of the selected candidates.

	Male		Female	
White	A (99)	B (98)	C (96)	D (95)
Black	E (91)	F (91)	G (90)	H (89)
Asian	I (87)	J (87)	K (86)	L (83)

score=373

Table 1: A set of 12 individuals with sensitive attributes `race` and `gender`. Each cell lists an individual's ID, and score in parentheses.

Problem: **In-group fairness fails** for Female (C and D not picked, which G and K are), Black (E and F are not picked, while G is), and Asian (I and J are not picked, while K is). **In-group fairness holds** for White and Male groups though (those with higher scores)!

Ranking with diversity constraints

[K. Yang, V. Gkatzelis, J. Stoyanovich, IJCAI (2019)]

Goal: pick k=4 candidates, including 2 of each gender, and at least one candidate per ethnicity, maximizing the total score of the selected candidates.

	Male		Female	
White	A (99)	B (98)	C (96)	D (95)
Black	E (91)	F (91)	G (90)	H (89)
Asian	I (87)	J (87)	K (86)	L (83)

score=372

Table 1: A set of 12 individuals with sensitive attributes `race` and `gender`. Each cell lists an individual's ID, and score in parentheses.

Problem: **In-group fairness fails** for Female (C and D not picked, which G and K are), Black (E and F are not picked, while G is), and Asian (I and J are not picked, while K is). **In-group fairness holds** for White and Male groups though (those with higher scores)!

Insight: while in-group fairness will inevitably fail to some extent because of diversity constraints, this loss should be **balanced** across groups.

Trading off utility, diversity and fairness

[K. Yang, V. Gkatzelis, J. Stoyanovich, IJCAI (2019)]

Goal: select and rank ***k*** items

Utility: each item has a score, maximize the sum of scores of selected items

Diversity: items have labels, pick at least $K_{v,p}$ items for each label v in each prefix of length $p < k$

Fairness: ensure that **loss is balanced** across all groups. We call this in-group fairness, **IGF**.

What's a good IGF measure?

Balancing loss across groups

[K. Yang, V. Gkatzelis, J. Stoyanovich, IJCAI (2019)]

ordered list of **female students**

female student	C	D	G	H	K	L
score	95	95	90	86	83	83

$$IGFRatio(Female) = \frac{score(K)}{score(D)} = 0.91$$

highest-scoring skipped item

lowest-scoring selected item

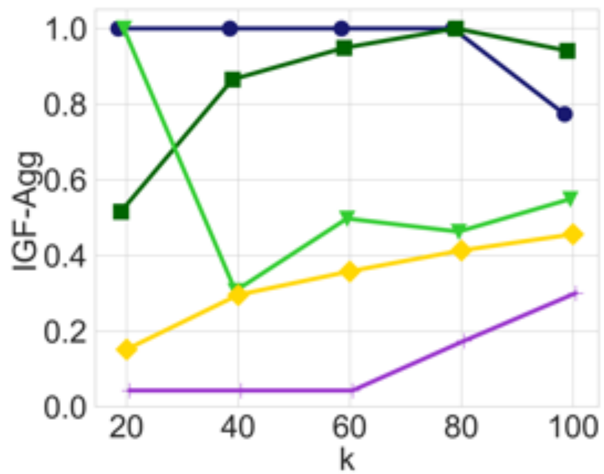
use an ILP, to maximize utility, subject to diversity and IGF loss constraints

Also propose another measure, *IGFAgg*, see paper.

Exploring the trade-off

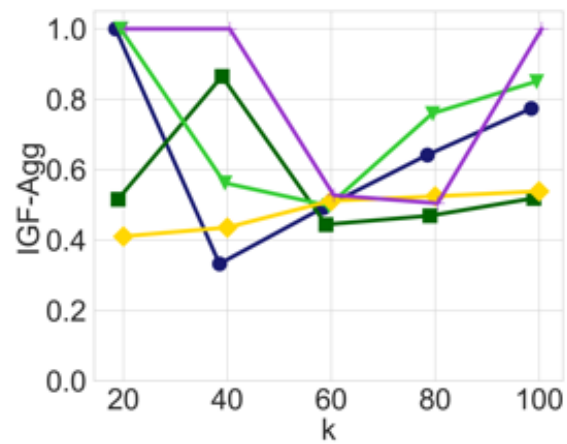
[K. Yang, V. Gkatzelis, J. Stoyanovich, IJCAI (2019)]

BEFORE: diversity constraints only



fairness **differs** among groups

AFTER: diversity and fairness constraints



fairness **balanced** across groups

MEPS (Medical Expenditure Panel Survey): sensitive attributes race and age