DS-GA 3001.009: Responsible Data Science

Algorithmic Fairness (continued)

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

@stoyanoj

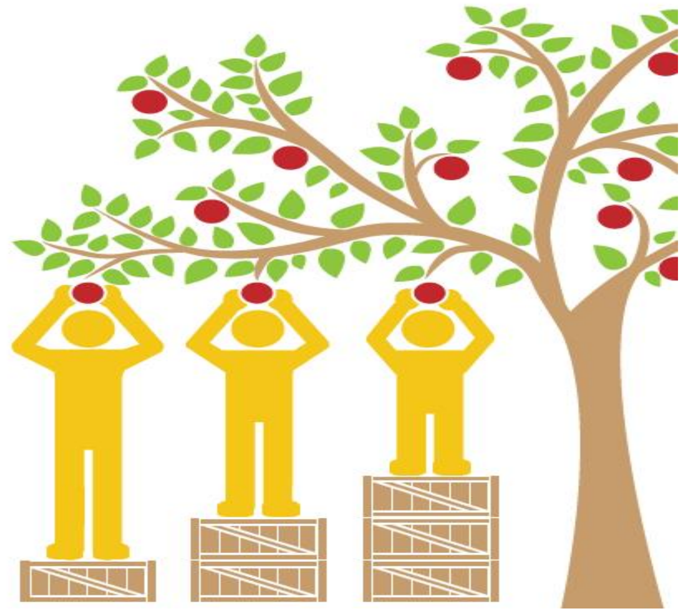
http://stoyanovich.org/
https://dataresponsibly.github.io/

Two notions of fairness

individual fairness

equality

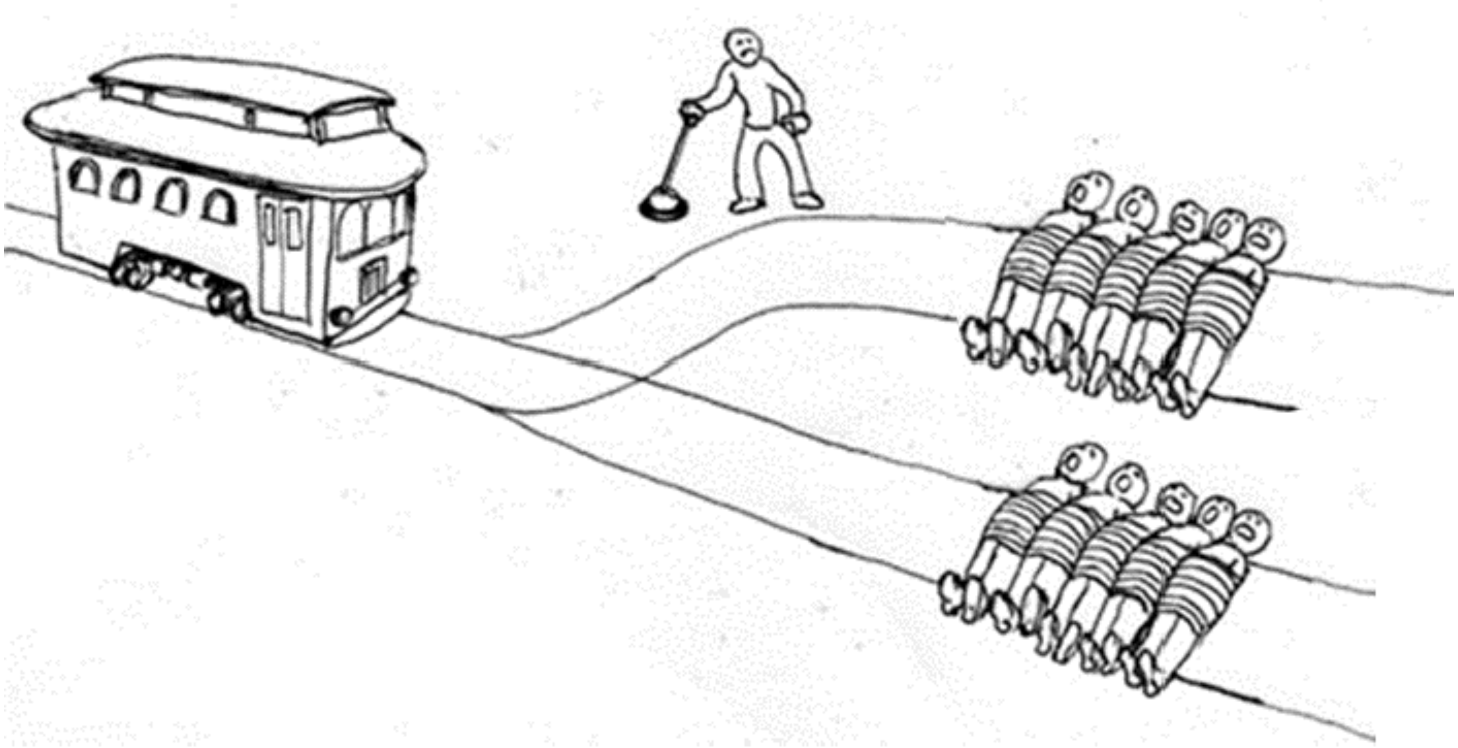
group fairness



equity

two intrinsically different world views

Fairness definitions as “trolley problems”



https://www.helpage.org/silo/images/blogs/16_1391611056.gif

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 errors 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

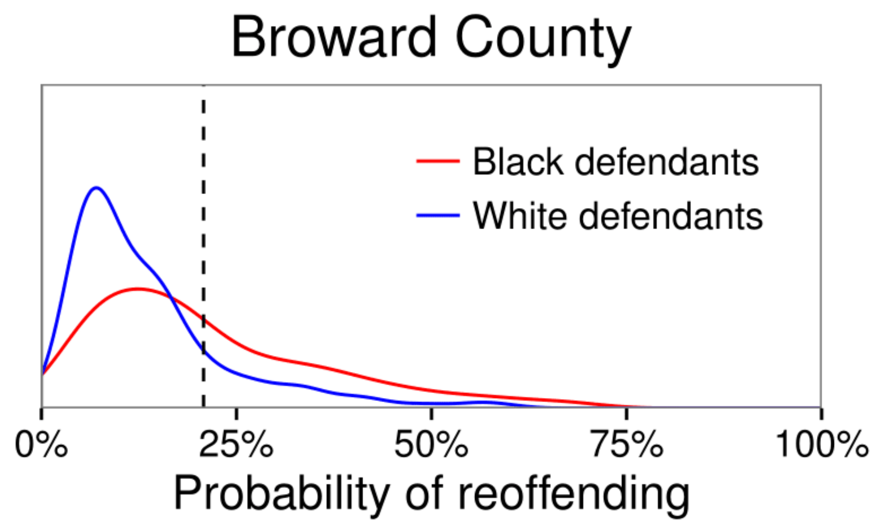
why do we want calibration?

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 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

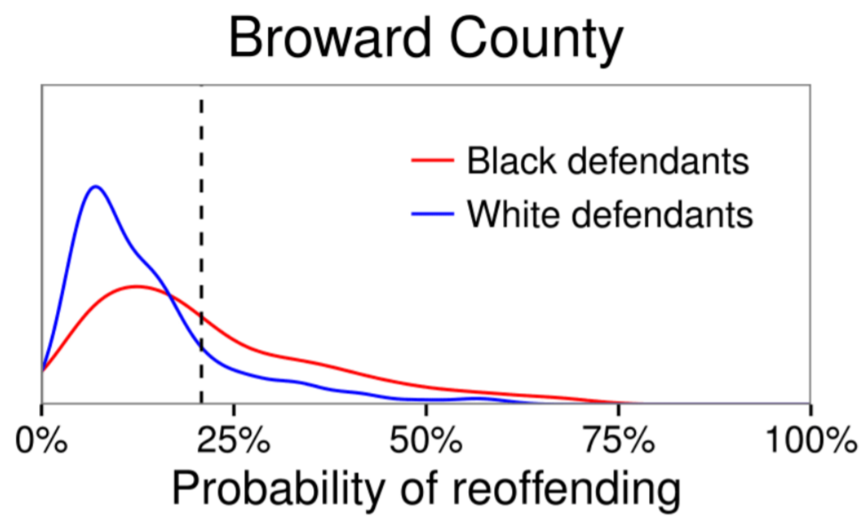
<table>
<tr><td></td><td></td><td colspan="2">True condition</td><td></td><td colspan="2"></td></tr>
<tr><td></td><td>Total population</td><td>Condition positive</td><td>Condition negative</td><td>Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$</td><td colspan="2">Accuracy (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
- Be 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 = ?

Evaluating fairness-aware algorithms

[S. Friedler, C. Scheidegger, S. Venkatsubramanian, S. Chaudhary, E. Hamilton, D. Roth; FAT* *(2019)*]

How do we know how to trade off different fairness objectives, and how to encode them in fairness-aware algorithms? - Societal context + experimental work!

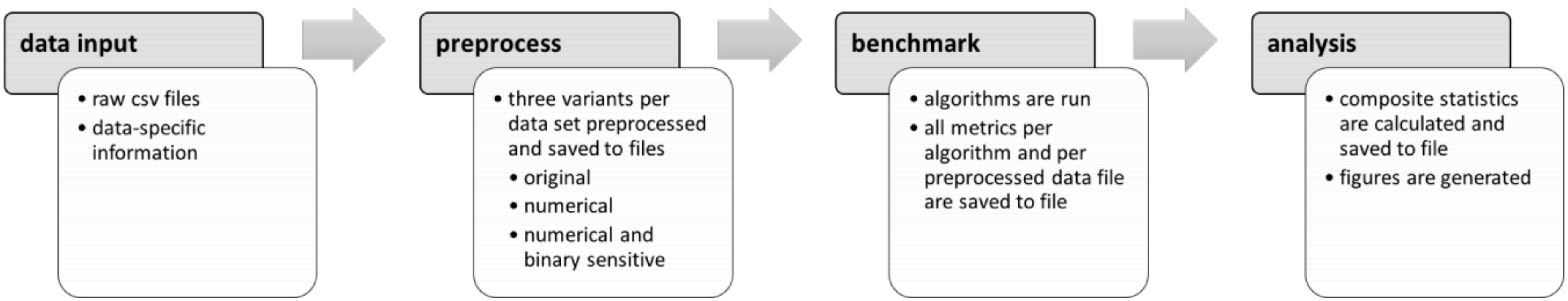


Figure 1: The stages of the fairness-aware benchmarking program: data input, preprocessing, benchmarking, and analysis. Intermediate files are saved at each stage of the pipeline to ensure reproducibility.

Insight 1: pre-processing matters

[S. Friedler, C. Scheidegger, S. Venkatsubramanian, S. Chaudhary, E. Hamilton, D. Roth; FAT* *(2019)*]

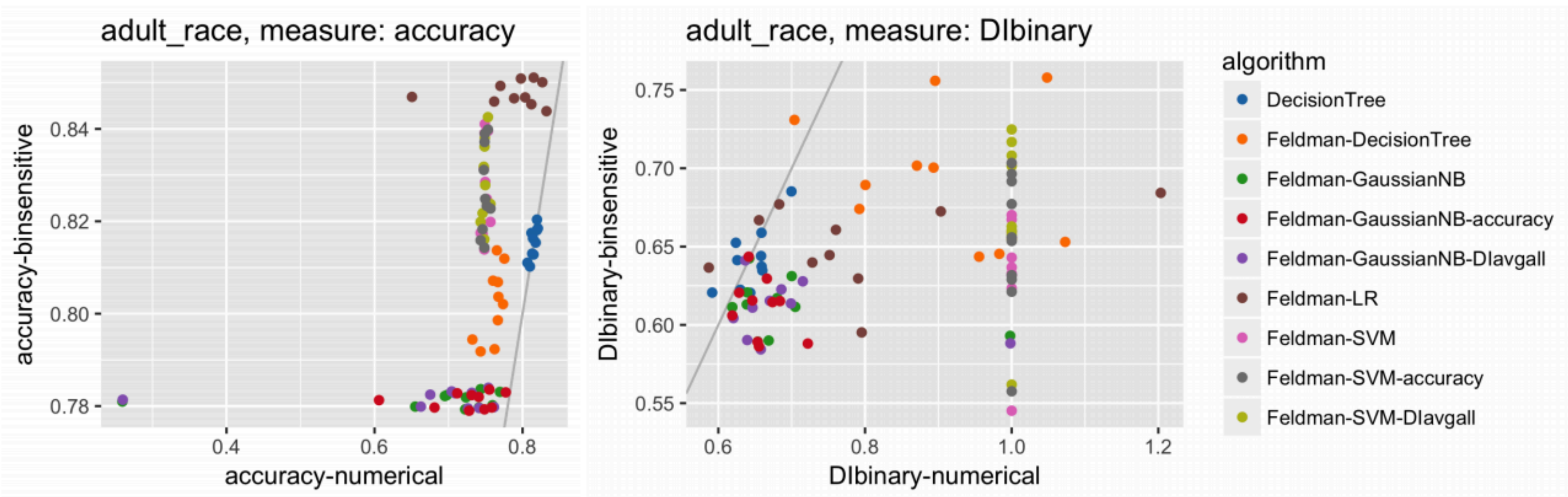
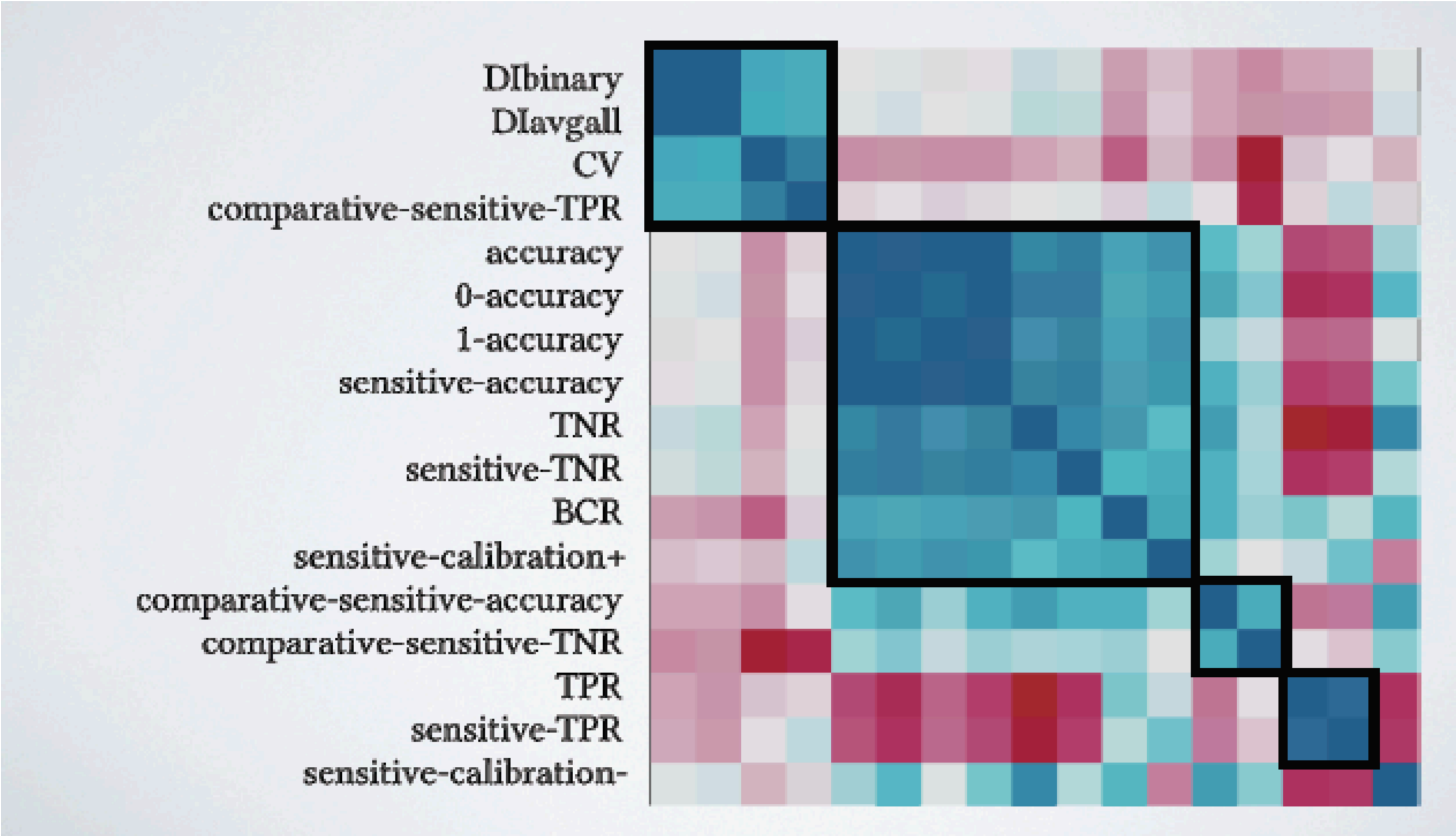


Figure 2: Examining the results of the Feldman et al. [10] algorithm under different preprocessing choices: `numerical` versus `numerical+binary`. Each dot plots the result of a single split of the data in terms of the labeled metric under both preprocessing choices. The gray line shows equality between the preprocessing choices. The model used within the Feldman algorithm is listed, and some variants of the algorithm had the tradeoff parameter optimized for either accuracy or disparate impact value.

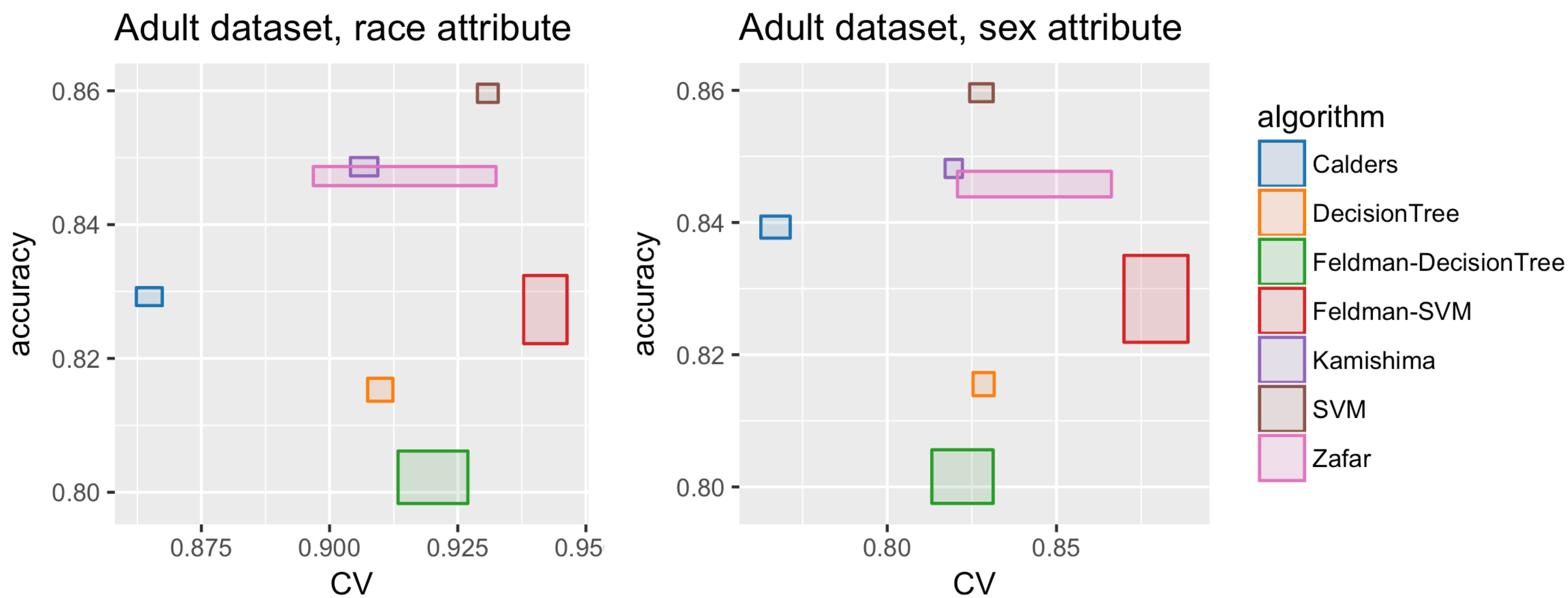
Insight 2: some measures correlate

[S. Friedler, C. Scheidegger, S. Venkatsubramanian, S. Chaudhary, E. Hamilton, D. Roth; FAT* *(2019)*]



Insight 3: beware of variability

[S. Friedler, C. Scheidegger, S. Venkatsubramanian, S. Chaudhary,
E. Hamilton, D. Roth; FAT* *(2019)*]



Feldman et al. varies in accuracy over splits while Zafar et al. varies in fairness.

Causal interpretations of fairness

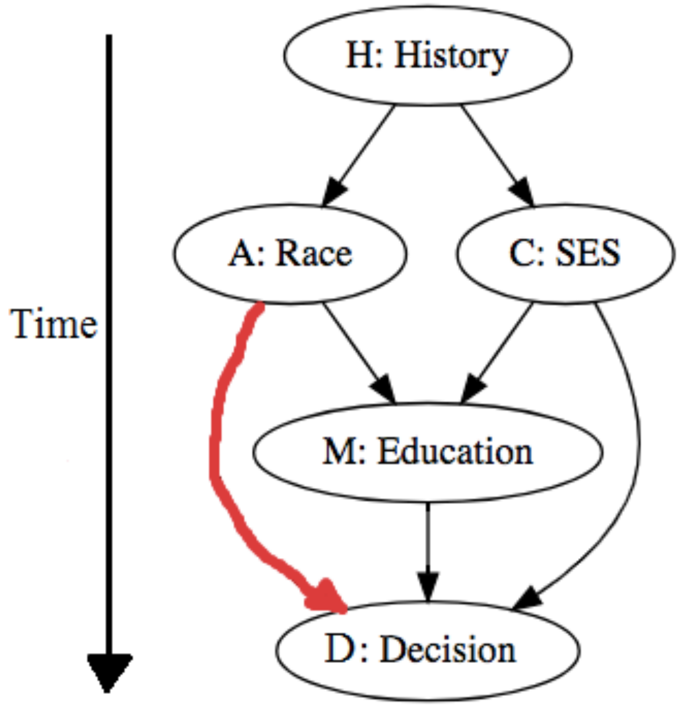
https://shiraamitchell.github.io/fairness/

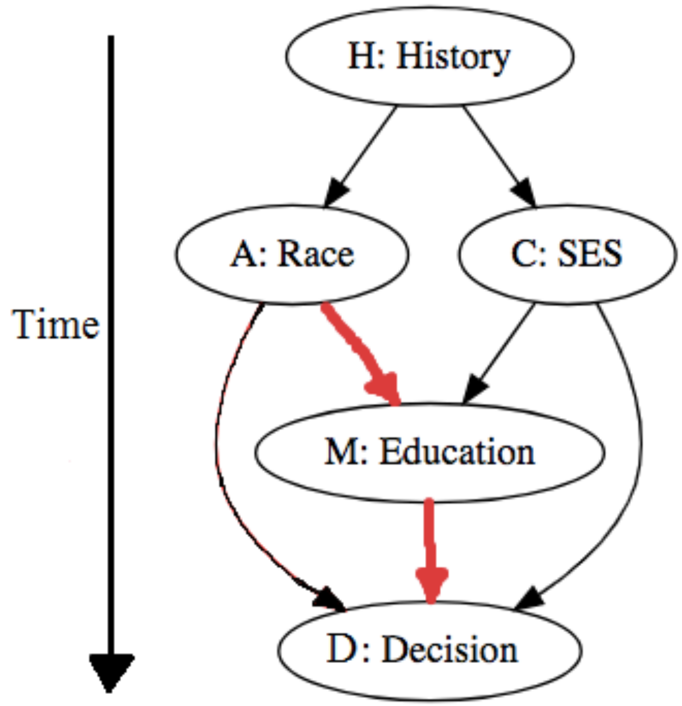
- Will be covered in a **guest lecture by Shira Mitchell on Thursday, February 21** (no lecture that week, so we'll use lab time as lecture time)
- Starting with **counterfactuals**:
  - *Was I not hired because I was black? => Would I have been hired if I were non-black?*
  - *Is there an effect of race on hiring? => Would the rate of hiring be the same if everyone were black? If no-one were?*

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 a **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

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

$f = 0.5$

Rank-aware fairness

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

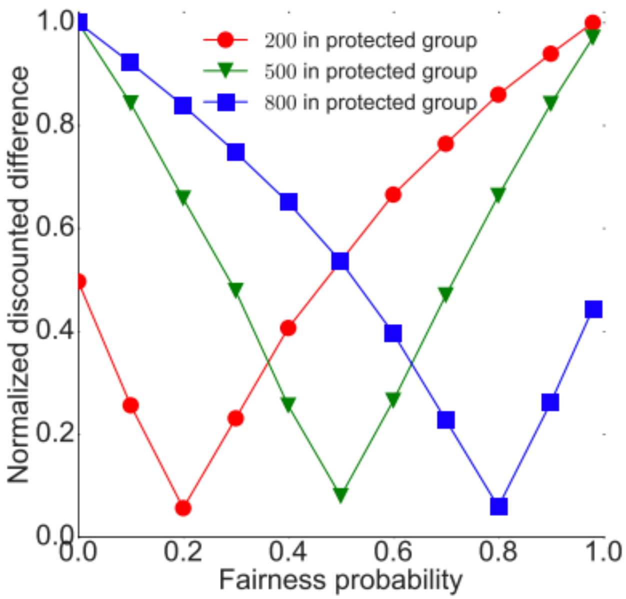


Figure 3: rND on 1,000 items

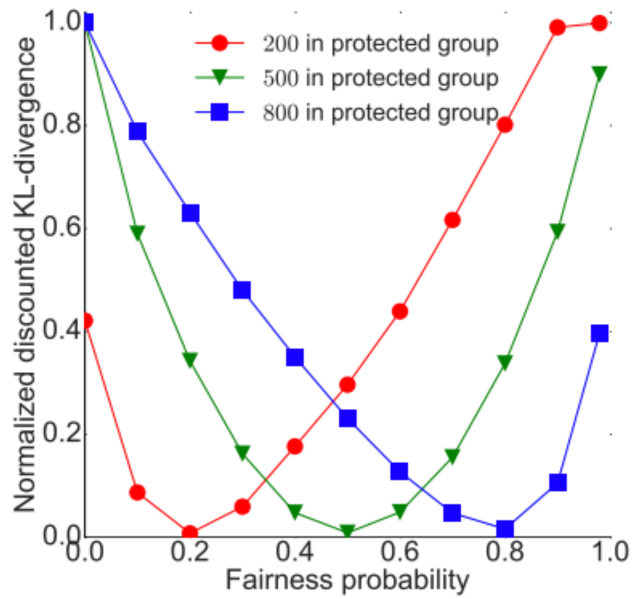


Figure 4: rKL on 1,000 items

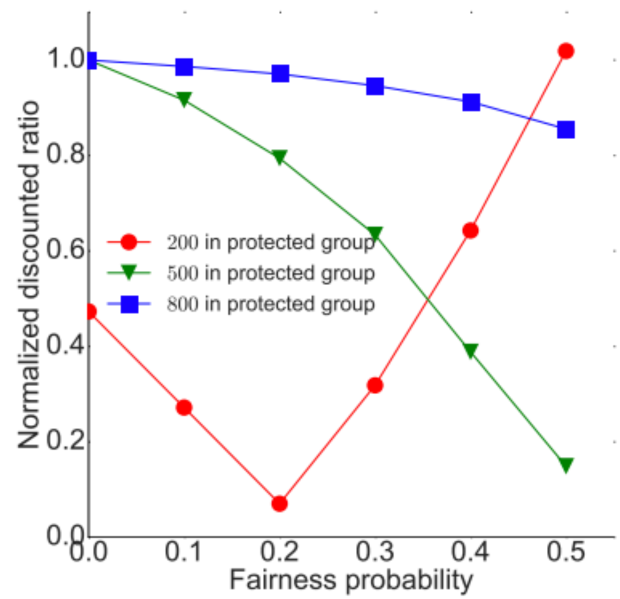


Figure 5: rRD on 1,000 items

In an optimization framework

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

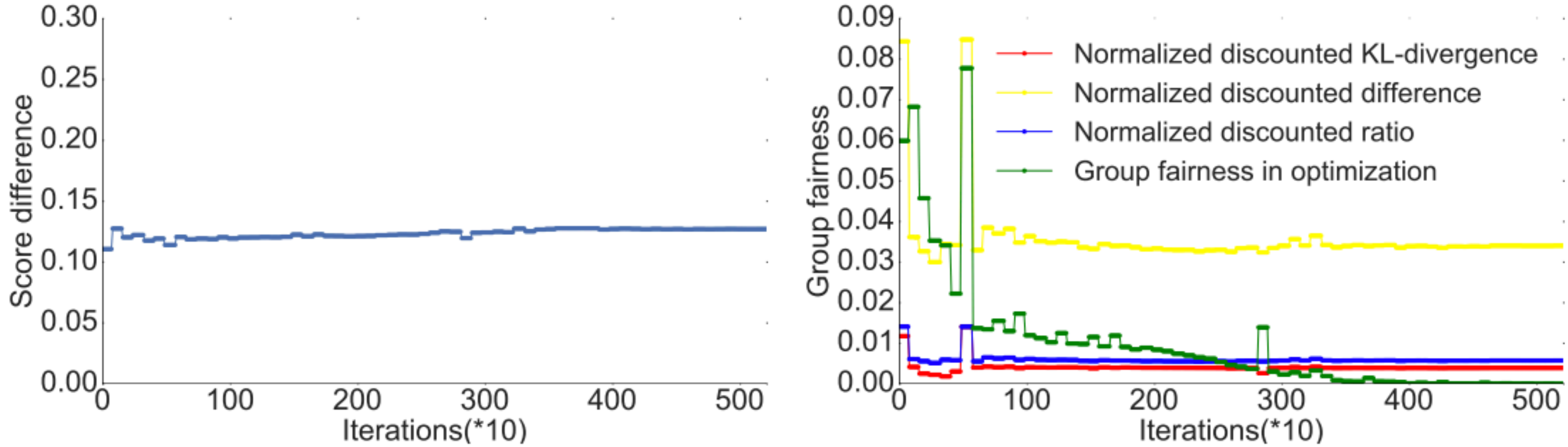
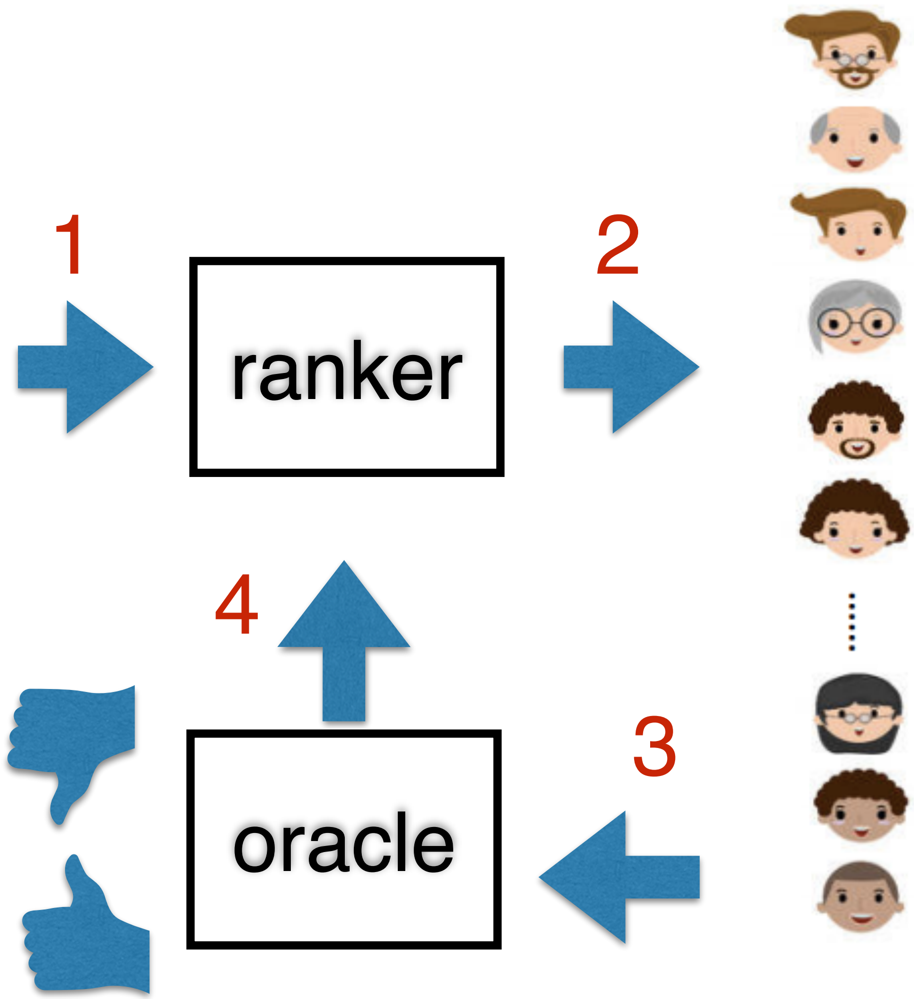


Figure 6: Accuracy and fairness on German Credit, ranked by *sum of normalized attribute values*, with $k = 10$.

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



Score-based rankers

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

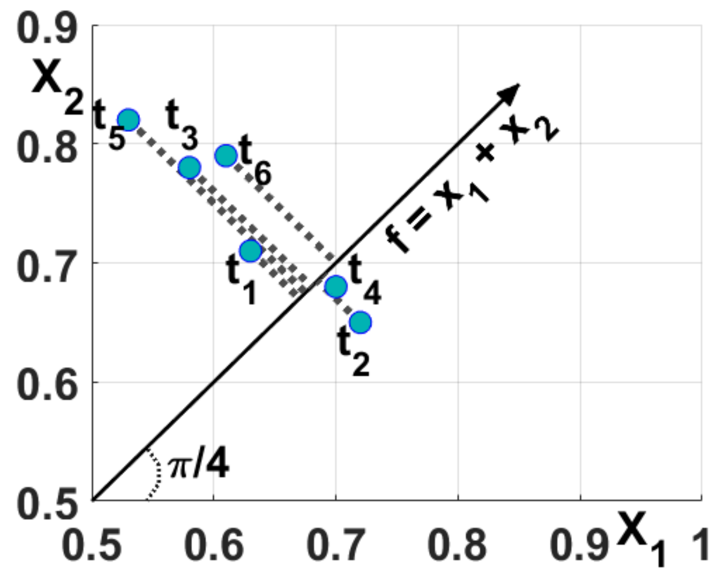
- tuple x in D; score(x): sum of attribute values, with non-negative weights (a common special case of monotone aggregation)
- weights subjectively chosen by a user: 0.5 g+ 0.5s, where g - normalized GPA, s - normalized SAT; why not 0.45 g + 0.55 s?

$\mathcal{D}$			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 $\mathcal{O} : \nabla_f(\mathcal{D}) \rightarrow \{\top, \bot\}$, 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 $\mathcal{O}(\nabla_{f'}(\mathcal{D})) = \top$ and the angular distance between $\vec{w}$ and $\vec{w}'$ is minimized.*

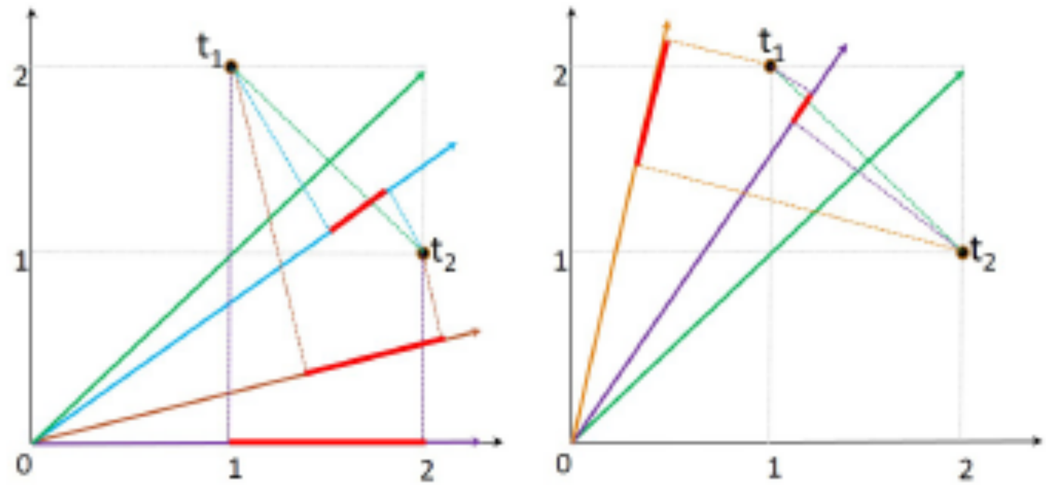
How might we approach this? Why is this difficult?

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.

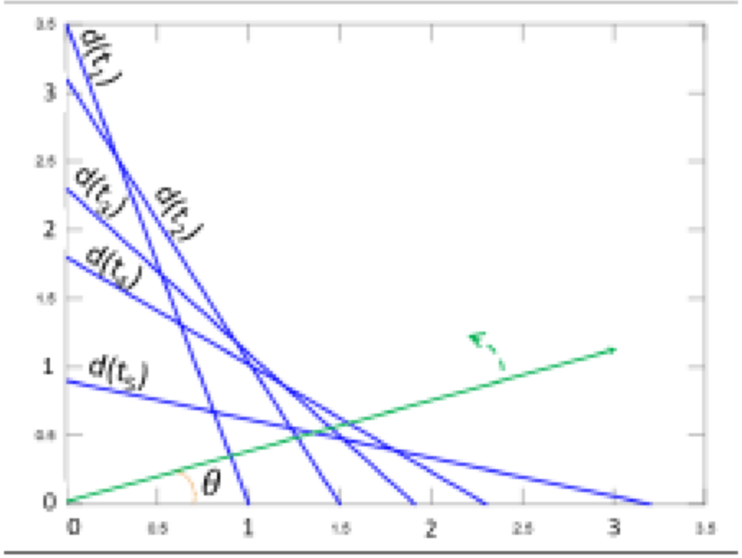
Outline of approach

[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

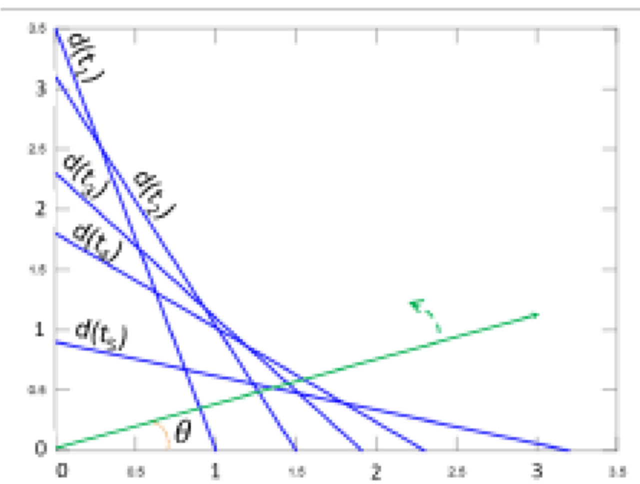


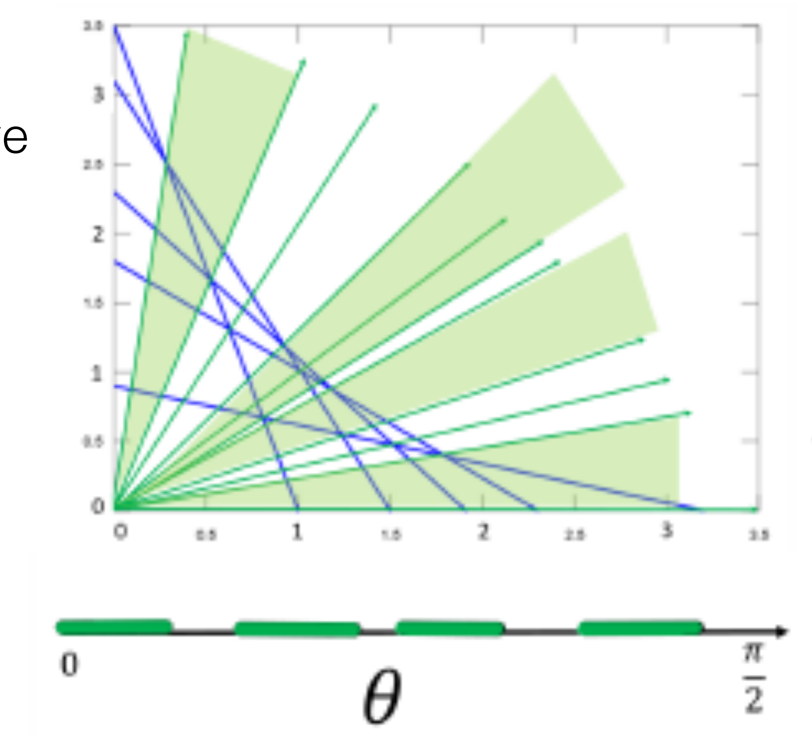
Outline of approach

[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

Diversity

Selection in presence of bias

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

Marianne Bertrand

Sendhil Mullainathan

AMERICAN ECONOMIC REVIEW
VOL. 94, NO. 4, SEPTEMBER 2004
(pp. 991-1013)

We study race in the labor market by sending fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, resumes are randomly assigned African-American- or White-sounding names. White names receive 50 percent more callbacks for interviews. Callbacks are also more responsive to resume quality for White names than for African-American ones. The racial gap is uniform across occupation, industry, and employer size. We also find little evidence that employers are inferring social class from the names. Differential treatment by race still appears to still be prominent in the U.S. labor market. (JEL J71, J64).

Selection in presence of bias

HARVARD | BUSINESS | SCHOOL

WORKING KNOWLEDGE

Business Research for Business Leaders

17 MAY 2017 RESEARCH & IDEAS

Minorities Who 'Whiten' Job Resumes Get More Interviews

by Dina Gerdeman

African American and Asian job applicants who mask their race on resumes seem to have better success getting job interviews, according to research by **Katherine DeCelles** and colleagues.

Selection in presence of bias

Blacks get more job interview callbacks when they "whiten" their resumes. Graphic by Blair Storie-Johnson (*Source: "Whitened Resumes: Race and Self-Presentation in the Labor Market"*)

AI’s White Guy Problem

The New York Times

Artificial Intelligence’s White Guy Problem

By KATE CRAWFORD JUNE 25, 2016

Like all technologies before it, artificial intelligence will reflect the values of its creators. So **inclusivity matters** — from who designs it to who sits on the company boards and which ethical perspectives are included.

Otherwise, **we risk constructing machine intelligence that mirrors a narrow and privileged vision of society**, with its old, familiar biases and stereotypes.

problems are beyond AI, whatever your definition of AI

Step 1: The Rooney Rule

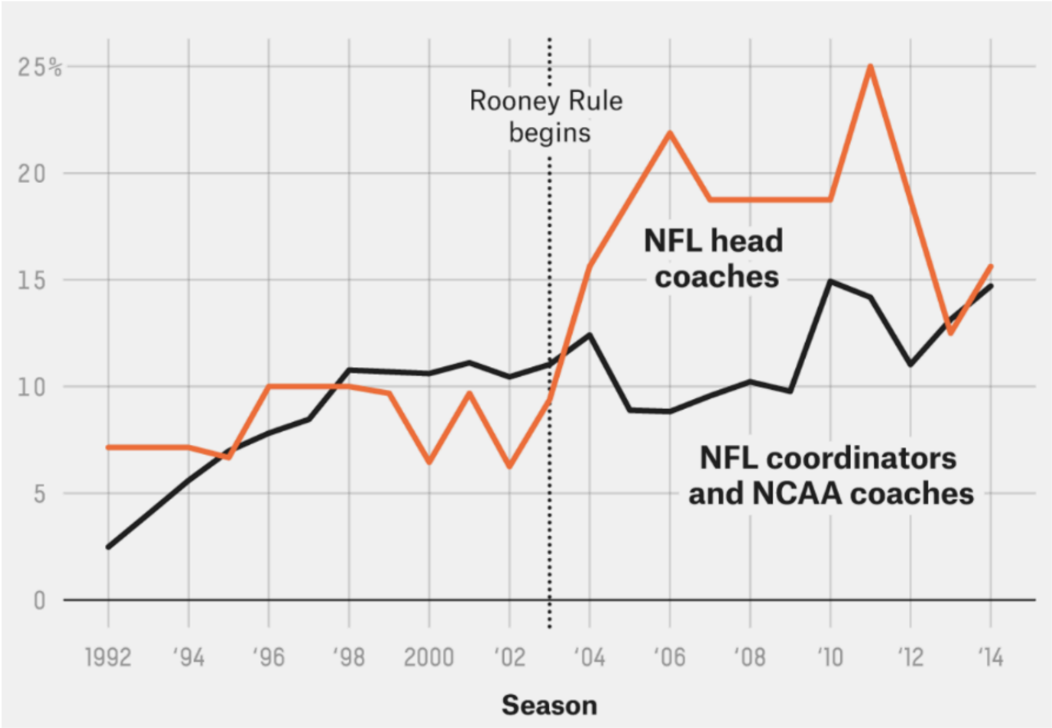
Named for a protocol adopted by the National Football League (NFL) in 2002, to increase the number of African-American head coaches

Requires that **at least one minority candidate** be interviewed for a position

Currently also used by the tech giants, to increase hiring of women and members of under-represented minorities (URM)

Push-back based on a **utility argument**: does the quality of the hired candidate / candidates decrease if the Rooney Rule is implemented?

Step 1: The Rooney Rule



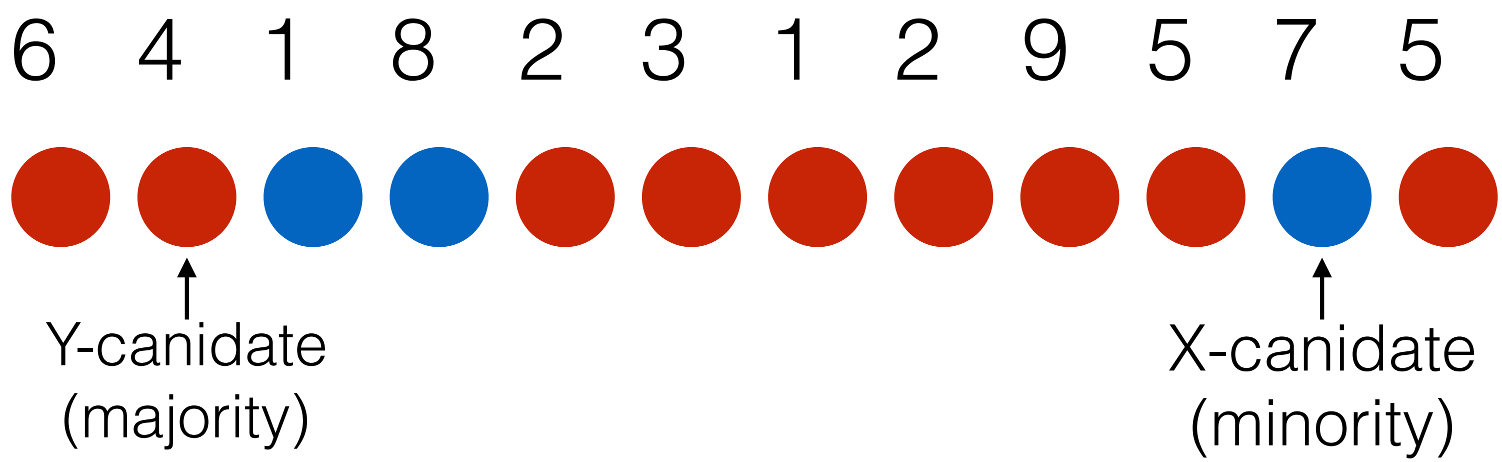
DuBois 2016

Selection with implicit bias

[J. Kleinberg, M. Raghavan, ITCS (*2018*)]

Goal: Given some estimates of the **extent of bias**, and the **prevalence** of available minority candidates, develop a mathematical model to quantify the **expected quality** of the candidates interviewed by a hiring committee.

Potential: of each candidate drawn from Z, the **same power law distribution for X and Y**!



α proportion of X-candidates

δ exponent of the power law

Bias in scoring and ranking

MENU
nature
International journal of science

Commentary | Published: 22 May 1997

Nepotism and sexism in peer-review

Christine Wennerås & Agnes Wold

Nature **387**, 341–343 (1997) | Download Citation

In the first-ever analysis of peer-review scores for postdoctoral fellowship applications, the system is revealed as being riddled with prejudice. The policy of secrecy in evaluation must be abandoned.

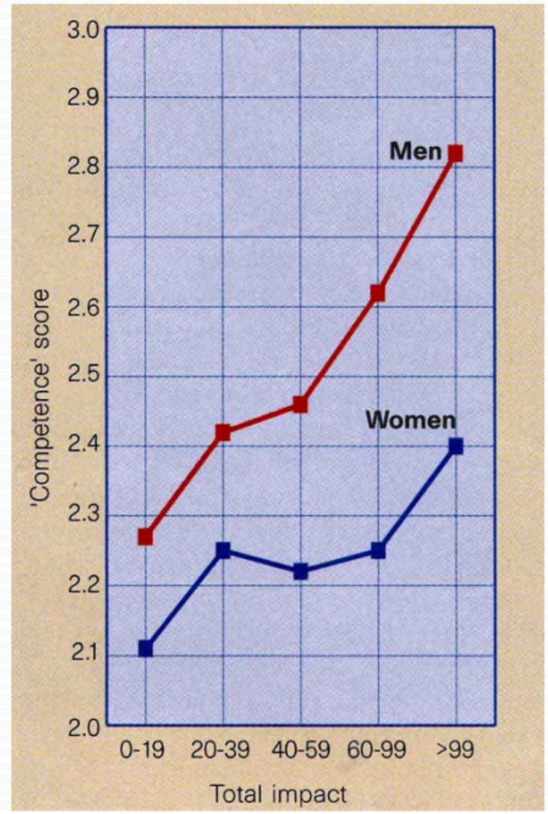


Figure 1 **The mean competence score given to male (red squares) and female (blue squares) applicants by the MRC reviewers as a function of their scientific productivity, measured as total impact. One impact point equals one paper published in a journal with an impact factor of 1. (See text for further explanation.)**

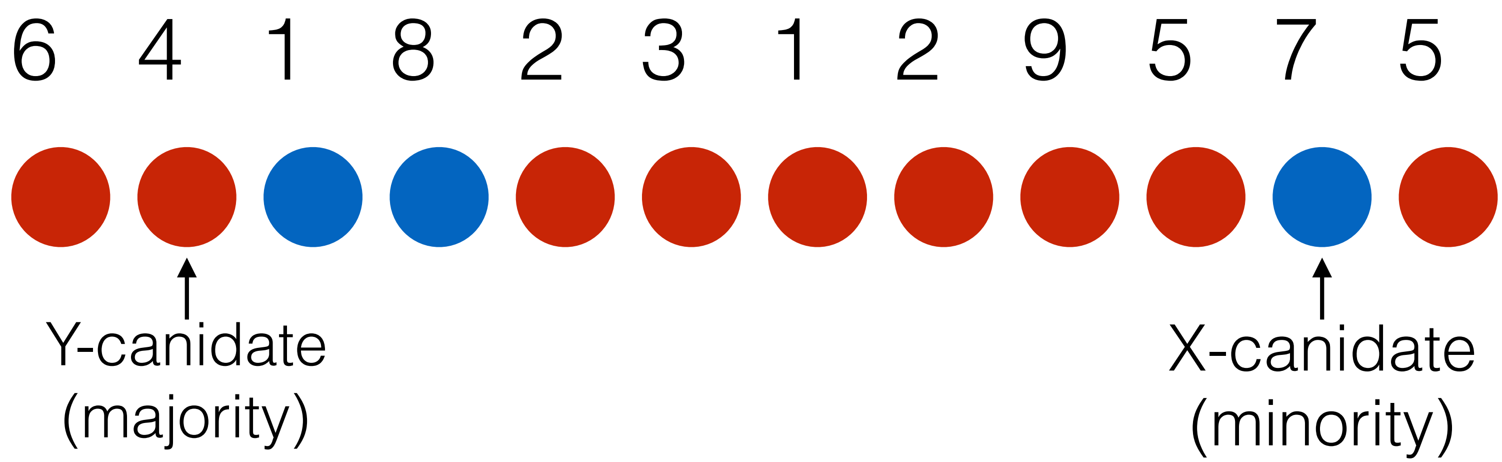
Multiplicative bias

[J. Kleinberg, M. Raghavan, ITCS (*2018*)]

Goal: Pick **k** finalists to interview, maximizing expected utility: sum of potentials of the chosen candidates.

Bias: Committee correctly estimates the potential of Y-candidates and they under-estimate the potential of X-candidates.

$$\boxed{\begin{array}{l}\tilde{X}_i = X_i / \beta \\ \beta > 1\end{array}}$$



Process: Estimate potentials of all candidates, rank, pick the best **k**.

Main result

[J. Kleinberg, M. Raghavan, ITCS (*2018*)]

▶ **Theorem 1.** *For $k = 2$ and sufficiently large n, the Rooney Rule produces a positive expected change if and only if $\phi_2(\alpha, \beta, \delta) > 1$ where*

$$\phi_2(\alpha, \beta, \delta) = \frac{\alpha^{1/(1+\delta)}\left[1 - (1+c^{-1})^{-\delta/(1+\delta)}\left[1 + \frac{\delta}{1+\delta}(1+c)^{-1}\right]\right]}{\frac{\delta}{1+\delta}(1+c)^{-1-\delta/(1+\delta)}} \tag{1}$$

and $c = \alpha\beta^{-(1+\delta)}$. Moreover, $\phi_2(\alpha, \beta, \delta)$ is increasing in β, so for fixed α and δ there exists β^* such that $\phi_2(\alpha, \beta, \delta) > 1$ if and only if $\beta > \beta^*$.

Selecting a seemingly sub-optimal candidate can improve utility!

Illustration: Infinite bias (all Y ranked higher than all X), pick k=2 candidates. Rooney rule improves utility if and only if

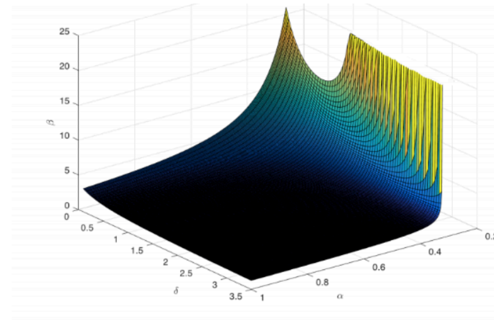
$$\alpha > \left(\frac{\delta}{1+\delta}\right)^{1+\delta}$$

Intuition

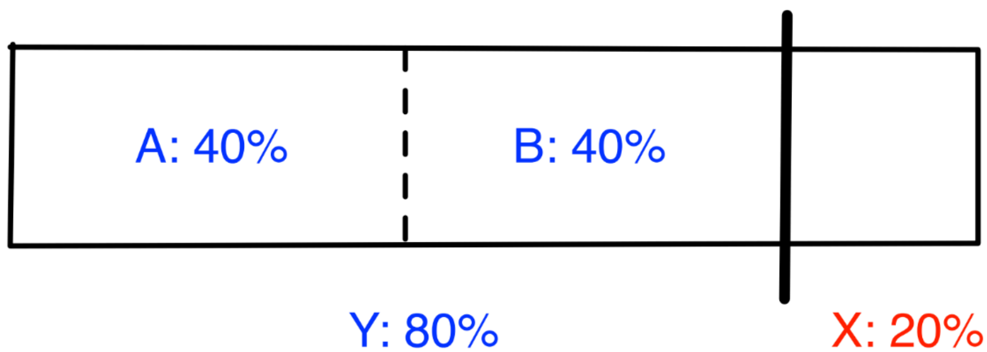
[J. Kleinberg, M. Raghavan, ITCS (*2018*)]

For which (α, δ) pairs does the Rooney Rule improve utility as $\beta \to \infty$?

- When should we reserve a slot for an X-candidate in the case of infinite bias? (Let's focus on $k = 2$.)

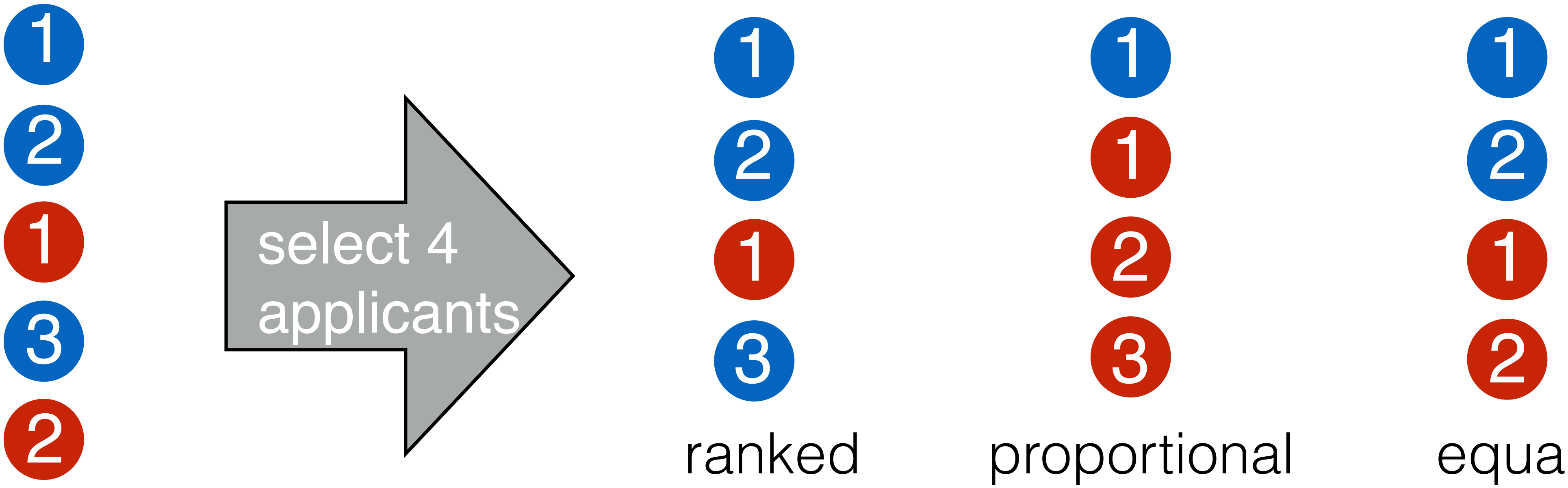


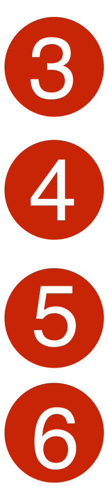
Surprising fact: No matter how small the fraction of X-candidates ($\alpha > 0$), there is a small enough power-law exponent ($\delta > 0$) so that the Rooney Rule improves utility.



slide from Jon Kleinberg's FAT* 2019 keynote

Another take: Online job applicant selection



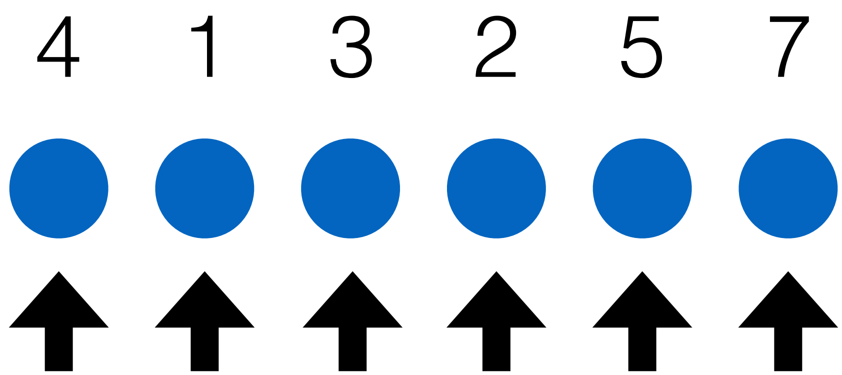


Can state all these as constraints:

for each category ***i***, pick ***K**ᵢ* elements, with $floor_i \leq K_i \leq ceil_i$

Hiring a job candidate

Goal: Hire a candidate with a high score



Candidates arrive one-by-one

A candidate's score is revealed when the candidate arrives

Decision to accept or reject a candidate made on the spot

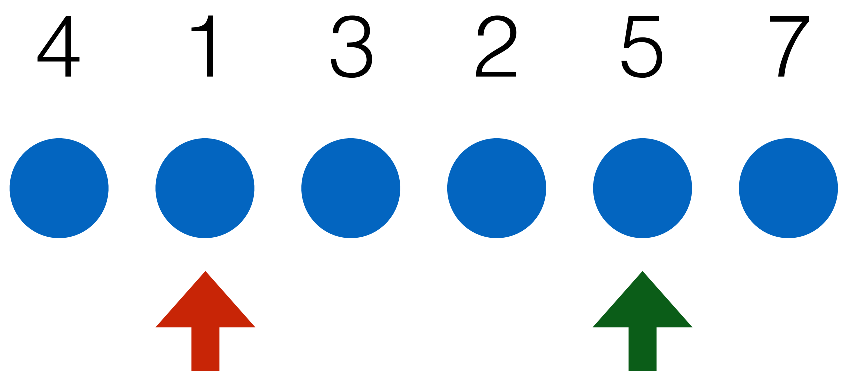
The Secretary Problem

Goal: Design an algorithm for picking **one** element of a **randomly ordered** sequence, to maximize the probability of picking the **maximum element** of the entire sequence.

$N = 6$

$S = \left\lfloor N/e \right\rfloor = 2$

$T = 4$



Competitive ratio

$1/e$

the best possible!

Consider, and reject, the first *S* candidates

Record *T*, the best seen score among the first *S* candidates

Accept the next candidate with score better than *T*

K-choice Secretary

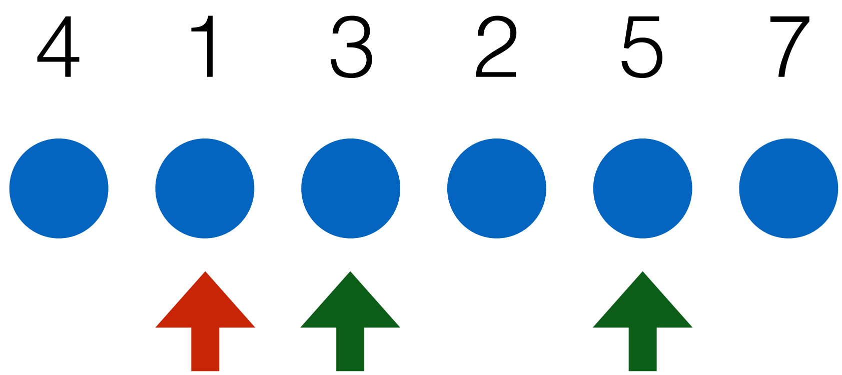
[Babaioff et al., 2007]

Goal: Design an algorithm for picking **K** elements of a **randomly ordered** sequence, to maximize their **expected sum**.

$N = 6 \quad K = 2$

$S = \left\lfloor N/e \right\rfloor = 2$

$T = \{1, 4\}$



Competitive ratio

$1/e$

far from optimal

Consider, and reject, the first ***S*** candidates

Record ***K*** best scores among the first ***S*** candidates, call this ***T***

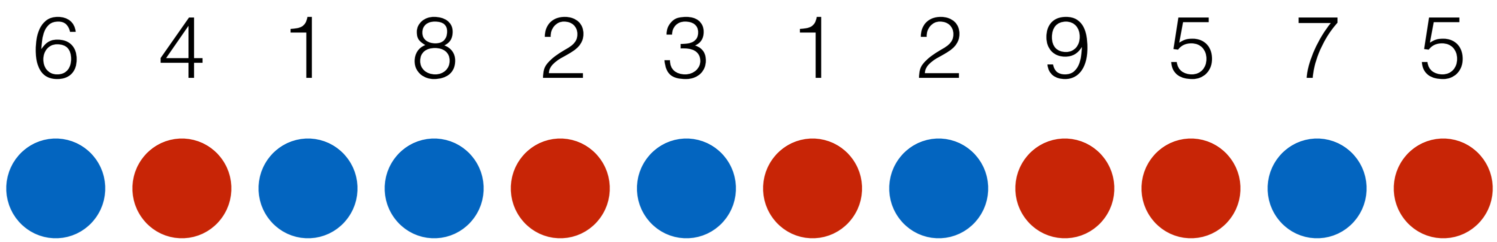
Whenever a candidate arrives whose score is higher than the minimum in ***T***, accept the candidate and delete the minimum from ***T***

K-choice Secretary

[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]

Goal: Design an algorithm for picking **K** elements of a **randomly ordered** sequence, to maximize their **expected sum**.

For each **category** *i*, pick K_i elements, with $floor_i \leq K_i \leq ceil_i$



$$N_{red} = N_{blue} = 6$$

$$K = 3$$

$$1 \leq K_{red}, K_{blue} \leq 2$$

Accept *floor* items for each category from per-category streams

$$slack = K - (floor_{red} + floor_{blue})$$

Accept the remaining *slack* items irrespective of category membership, but subject to *ceil*

Diverse K-choice Secretary

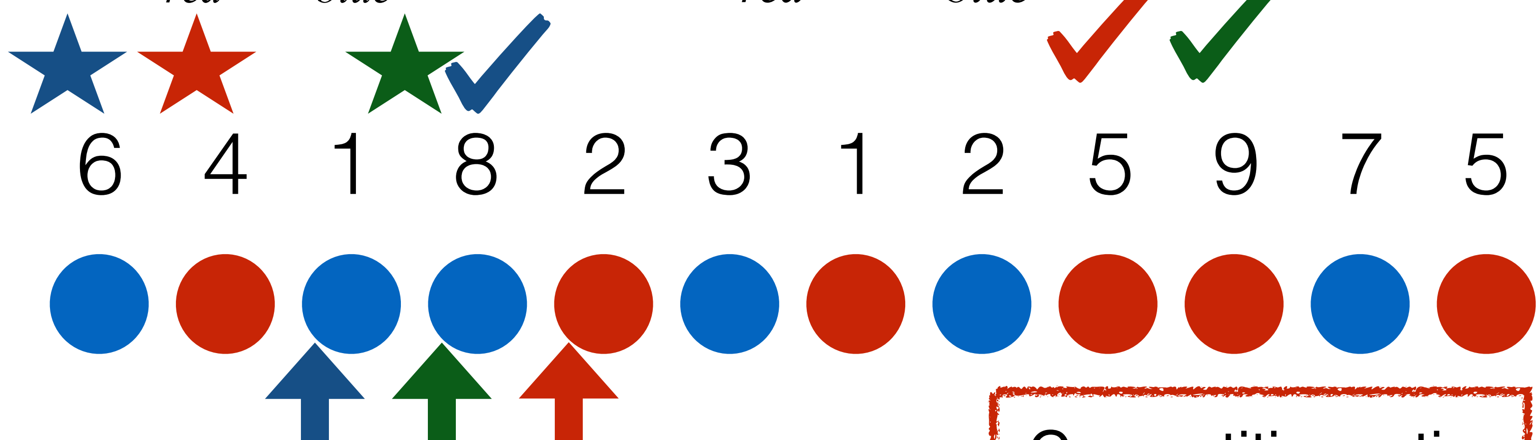
[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]

$N_{red} = N_{blue} = 6$

$K = 3 \quad 1 \leq K_{red}, K_{blue} \leq 2$

$slack = 1$

$S_{red} = S_{blue} = 2 \quad S = 4$



Competitive ratio

$1/e$

far from optimal

Adding a deferred list

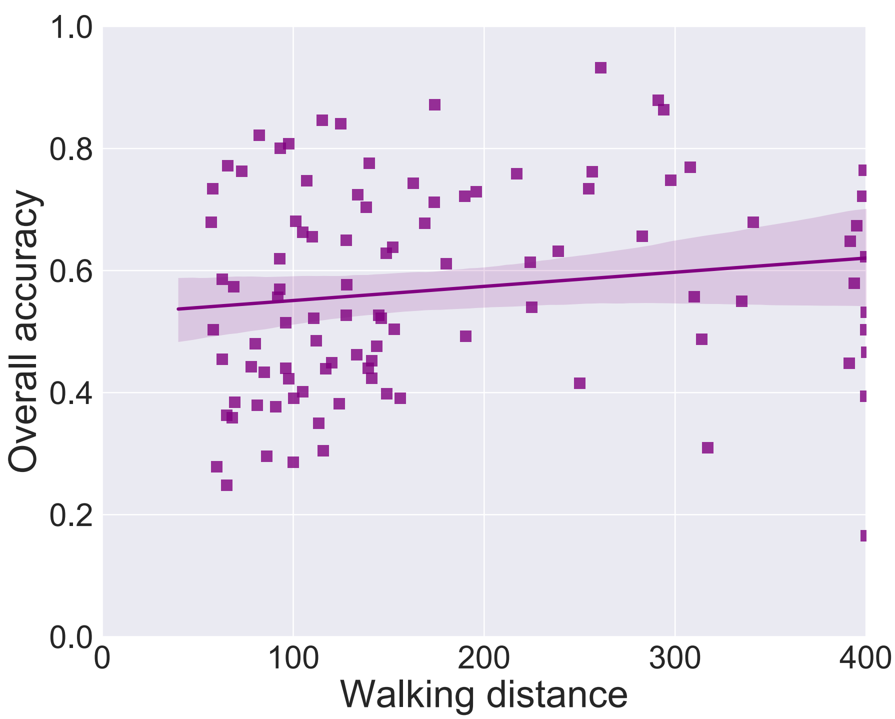
[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]

- An improvement on Diverse K-choice Secretary
- Do not immediately reject or accept items: keep a deferred list D_i per category i of size up to $\textbf{\textit{ceil}}_i$
- Stop reading the input, post-warm up, once all $\textit{floor}_i$ constraints are met, and once there are K items in the union of the deferred lists
- Main advantage: often avoids reading items from the end of the stream

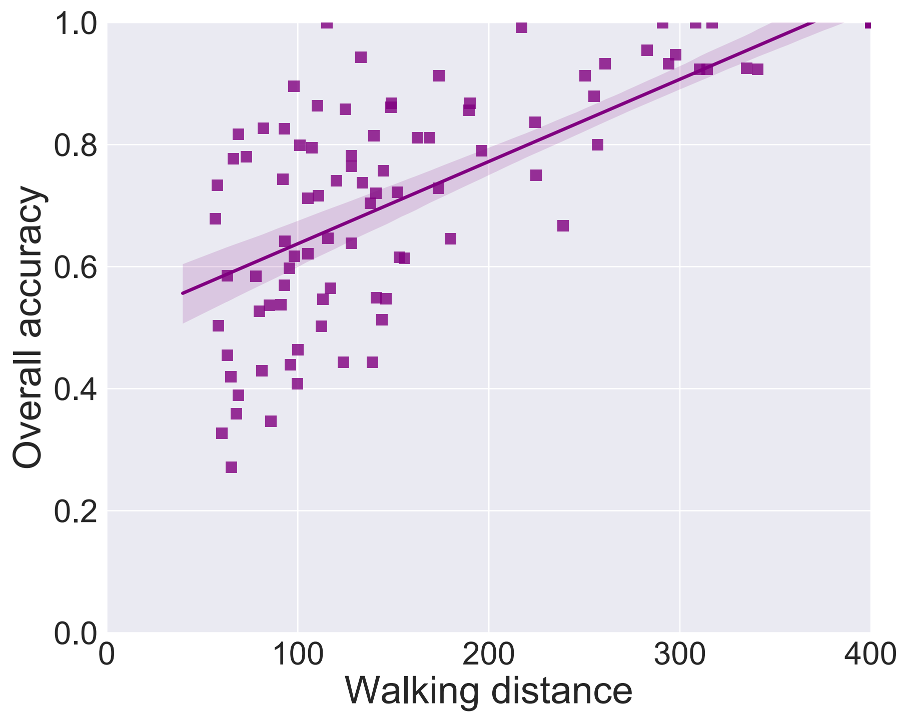
Diversity is achievable

[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]

deferred list



with deferred list

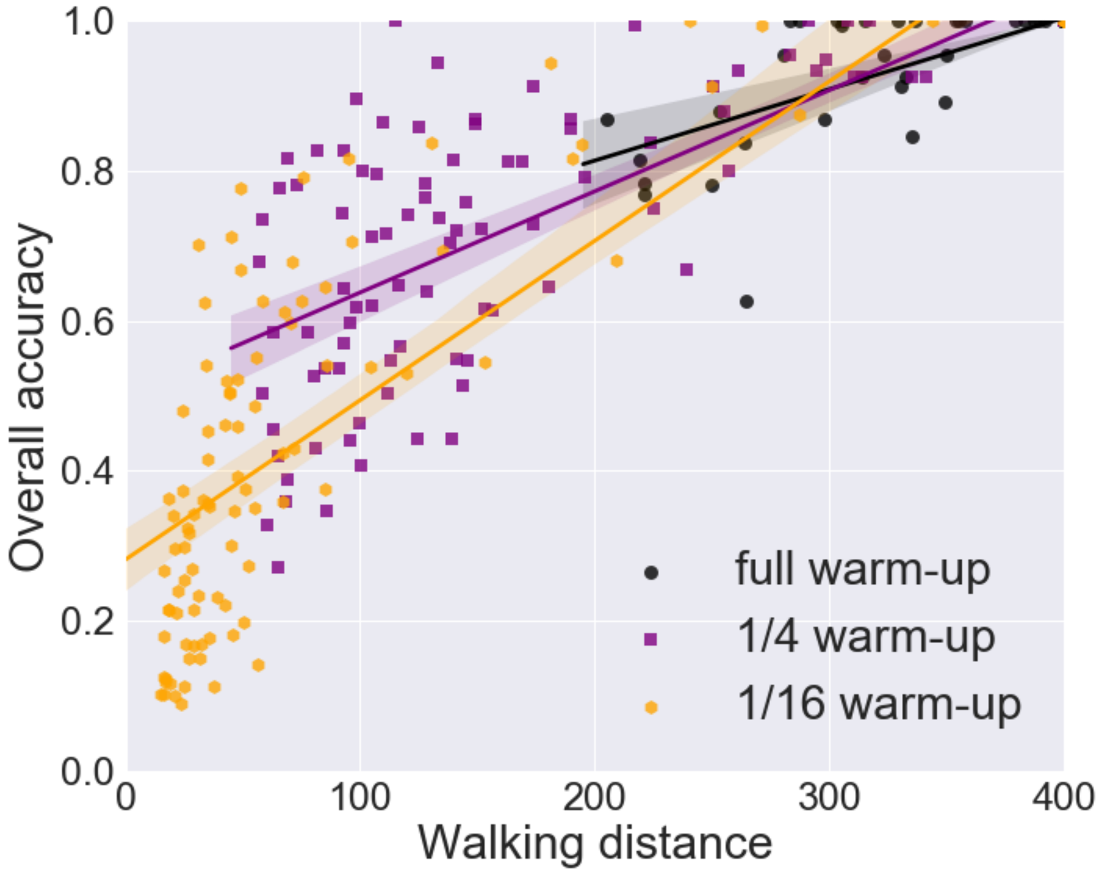


Forbes US Richest: N=400, K=4 (27 female, 373 male)

diversity on gender: select 2 per gender

Warm-up can be shorter

[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]

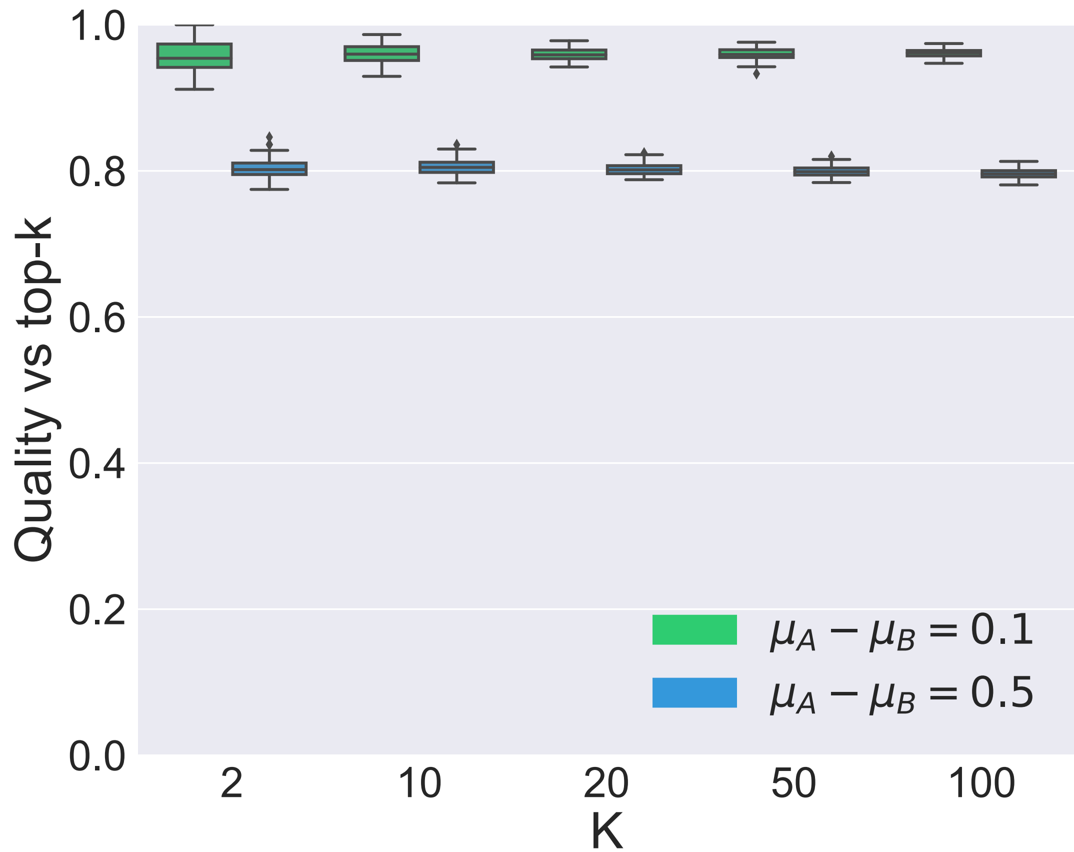


Forbes US Richest: N=400, K=4 (27 female, 373 male)

deferred list variant, diversity on gender: select 2 per gender

The cost of diversity

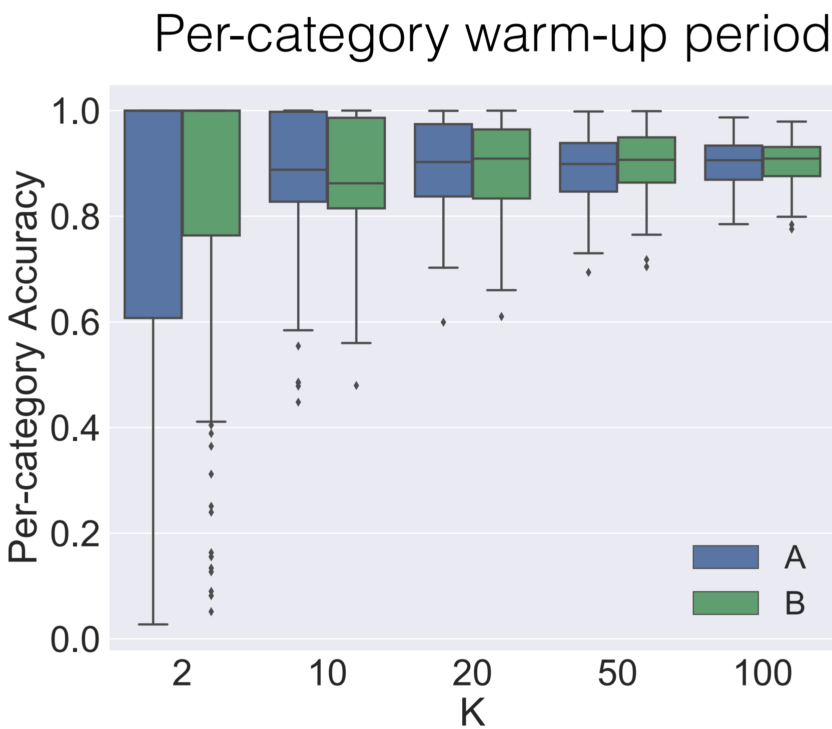
[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]

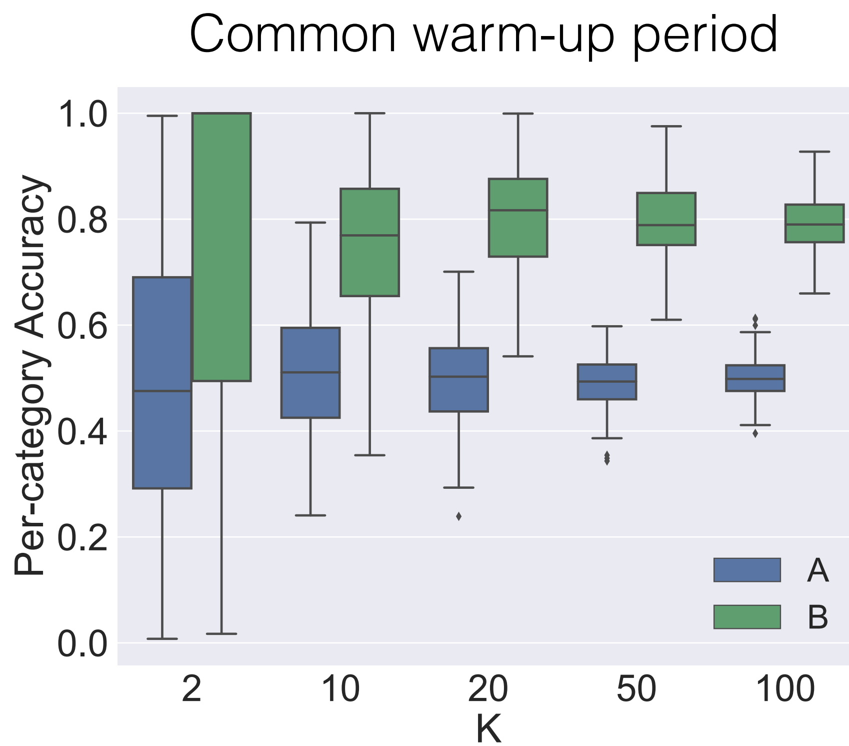


static variant (see paper), synthetic data in categories A and B, score lower for B

Per-category warm-up is crucial

[J. Stoyanovich, K. Yang, HV Jagadish, EDBT (2018)]





synthetic data with categories A and B, score depends on category, lower for A

diversity by design

Why is diversity important?

- Unlike fairness, there is **no legal reason** to enforce diversity
- However, there are strong **utilitarian reasons**: diversity leads to better user satisfaction (IR, recommendation), higher quality of results (crowdsourcing, team formation), more efficient resource allocation (matchmaking)
- Further, diversity levels the playing field and **improves fairness in the long run**