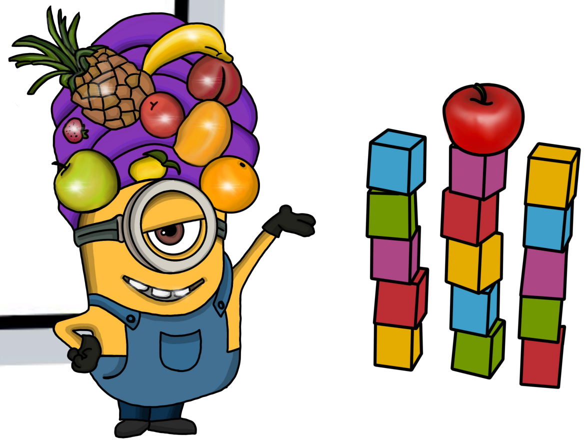
Fairness in Ranking

from values to technical choices & back

Julia Stoyanovich Meike Zehlike Ke Yang

ACM SIGMOD 2023



Fairness in Ranking, Part I: Score-Based Ranking

MEIKE ZEHLIKE, Humboldt University of Berlin, Max Planck Institute for Software Systems, and Zalando Research, Germany
KE YANG, New York University, NY, and University of Massachusetts, Amherst, MA, USA
JULIA STOYANOVICH, New York University, NY, USA

118

In the past few years, there has been much work on incorporating fairness requirements into algorithmic rankers, with contributions coming from the data management, algorithms, information retrieval, and recommender systems communities. In this survey, we give a systematic overview of this work, offering a broad perspective that connects formalizations and algorithmic approaches across sub-fields. An important contribution of our work is in developing a common narrative around the value frameworks that motivate specific fairness-enhancing interventions in ranking. This allows us to unify the presentation of mitigation objectives and of algorithmic techniques to help meet those objectives or identify trade-offs.

In this first part of this survey, we describe four classification frameworks for fairness-enhancing interventions, along which we relate the technical methods surveyed in this article, discuss evaluation datasets, and present technical work on fairness in score-based ranking. In the second part of this survey, we present methods that incorporate fairness in supervised learning, and also give representative examples of recent work on fairness in recommendation and matchmaking systems. We also discuss evaluation frameworks for fair score-based ranking and fair learning-to-rank, and draw a set of recommendations for the evaluation of fair ranking methods.

CCS Concepts: • **Information systems → Data management systems**; • **Social and professional topics → Computing/technology policy;**

Additional Key Words and Phrases: Fairness, ranking, set selection, responsible data science, survey

ACM Reference format:
Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in Ranking, Part I: Score-Based Ranking. *ACM Comput. Surv.* 55, 6, Article 118 (December 2022), 36 pages.
https://doi.org/10.1145/3533379

1 INTRODUCTION

The research community recognizes several important normative dimensions of information technology including privacy, transparency, and fairness. In this survey, we focus on fairness—a broad and inherently interdisciplinary topic of which the social and philosophical foundations are still unresolved [17].

This research was supported in part by NSF Awards No. 1934464, 1916505, and 1922658.
Authors' addresses: M. Zehlike, Humboldt University of Berlin, Max Planck Institute for Software Systems, and Zalando Research, Germany; email: meikezehlike@mpi-sws.org; K. Yang, New York University, NY, and University of Massachusetts, Amherst, MA, USA; email: ky630@nyu.edu; J. Stoyanovich, New York University, NY, USA; email: stoyanovich@nyu.edu.

0360-0300/2022/12-ART118 $15.00
https://doi.org/10.1145/3533379

ACM Computing Surveys, Vol. 55, No. 6, Article 118. Publication date: December 2022.

117

Fairness in Ranking, Part II: Learning-to-Rank and Recommender Systems

MEIKE ZEHLIKE, Humboldt University of Berlin, Max Planck Institute for Software Systems, and Zalando Research, Germany
KE YANG, New York University, NY, and University of Massachusetts, Amherst, MA, USA
JULIA STOYANOVICH, New York University, NY, USA

In the past few years, there has been much work on incorporating fairness requirements into algorithmic rankers, with contributions coming from the data management, algorithms, information retrieval, and recommender systems communities. In this survey, we give a systematic overview of this work, offering a broad perspective that connects formalizations and algorithmic approaches across subfields. An important contribution of our work is in developing a common narrative around the value frameworks that motivate specific fairness-enhancing interventions in ranking. This allows us to unify the presentation of mitigation objectives and of algorithmic techniques to help meet those objectives or identify trade-offs.

In the first part of this survey, we describe four classification frameworks for fairness-enhancing interventions, along which we relate the technical methods surveyed in this article, discuss evaluation datasets, and present technical work on fairness in score-based ranking. In the second part of this survey, we present methods that incorporate fairness in supervised learning, and also give representative examples of recent work on fairness in recommendation and matchmaking systems. We also discuss evaluation frameworks for fair score-based ranking and fair learning-to-rank, and draw a set of recommendations for the evaluation of fair ranking methods.

CCS Concepts: • **Information systems → Data management systems**; • **Social and professional topics → Computing/technology policy;**

Additional Key Words and Phrases: Fairness, ranking, set selection, responsible data science, survey

ACM Reference format:
Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in Ranking, Part II: Learning-to-Rank and Recommender Systems. *ACM Comput. Surv.* 55, 6, Article 117 (December 2022), 41 pages.
https://doi.org/10.1145/3533380

1 INTRODUCTION

This is the second part of a survey on fairness in ranking. In the first part, we argued for the importance of a systematic overview of work on incorporating fairness requirements into algorithmic rankers. Which specific fairness requirements a decision maker will assert depends on the

This research was supported in part by NSF Awards No. 1934464, 1916505, and 1922658.
Authors' addresses: M. Zehlike, Humboldt University of Berlin, Max Planck Institute for Software Systems, and Zalando Research, Germany; email: meikezehlike@mpi-sws.org; K. Yang, New York University, NY, and University of Massachusetts, Amherst, MA, USA; email: ky630@nyu.edu; J. Stoyanovich, New York University, NY, USA; email: stoyanovich@nyu.edu.

0360-0300/2022/12-ART117 $15.00
https://doi.org/10.1145/3533380

ACM Computing Surveys, Vol. 55, No. 6, Article 117. Publication date: December 2022.

https://dl.acm.org/doi/10.1145/3533379

https://dl.acm.org/doi/abs/10.1145/3533380

Example: college admissions

	sensitive attributes		qualification attributes				scores		
	gender	**race**	X_1	X_2	X_3	X_4	Y_1	Y_2	Y_3
b	m	w	4	5	5	cs:0.9, art:0.2	14	9	1
c	m	a	5	3	4	math:0.9, cs:0.5	12	9	1
d	f	w	5	4	2	lit:0.8, math:0.8	11	4	6
e	m	w	3	3	4	math:0.8, econ:0.4	10	7	6
f	f	a	3	2	3	econ:0.9, math:0.8	8	5	8
k	f	b	2	2	3	lit:0.9, art:0.8	7	1	9
l	m	b	1	1	4	lit:0.5, math:0.7	6	6	2
o	f	w	1	1	2	econ:0.9, cs:0.8	4	7	8

τ_1	τ_2	τ_3
b	c	k
c	b	o
d	e	f
e	o	d
f	l	e
k	f	l
l	d	c
o	k	b

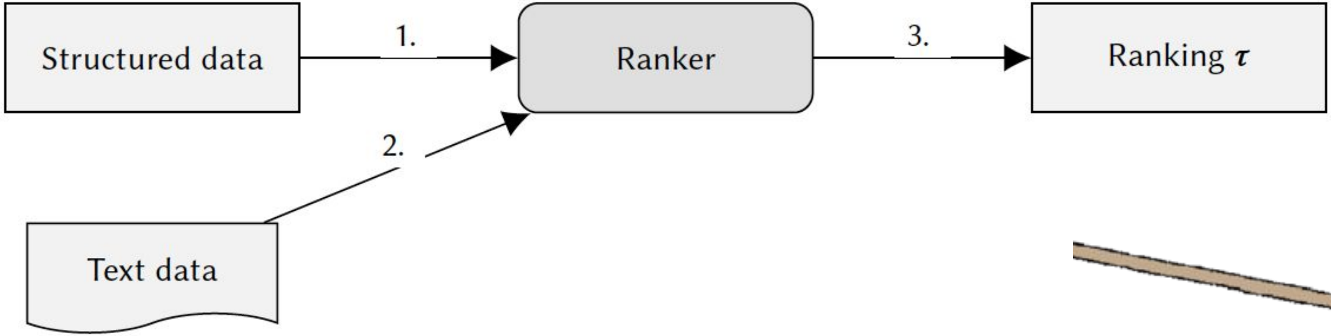
Example: college admissions

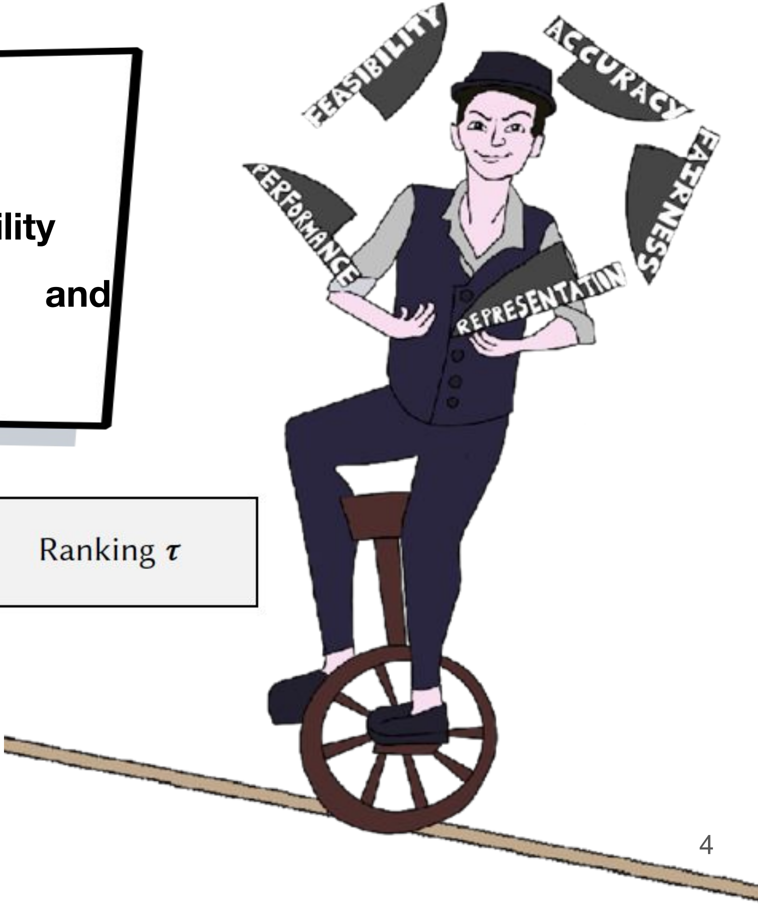
Goal: select candidates who

are likely to succeed (good grades, interested) ~ **utility**

form a demographically diverse group ~ **diversity** **and**

take the data with a grain of salt! ~ **fairness**





Ranking ranking everywhere

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.

Ranking ranking everywhere

theguardian **July 2015**

Women less likely to be shown ads for high-paid jobs on Google, study shows

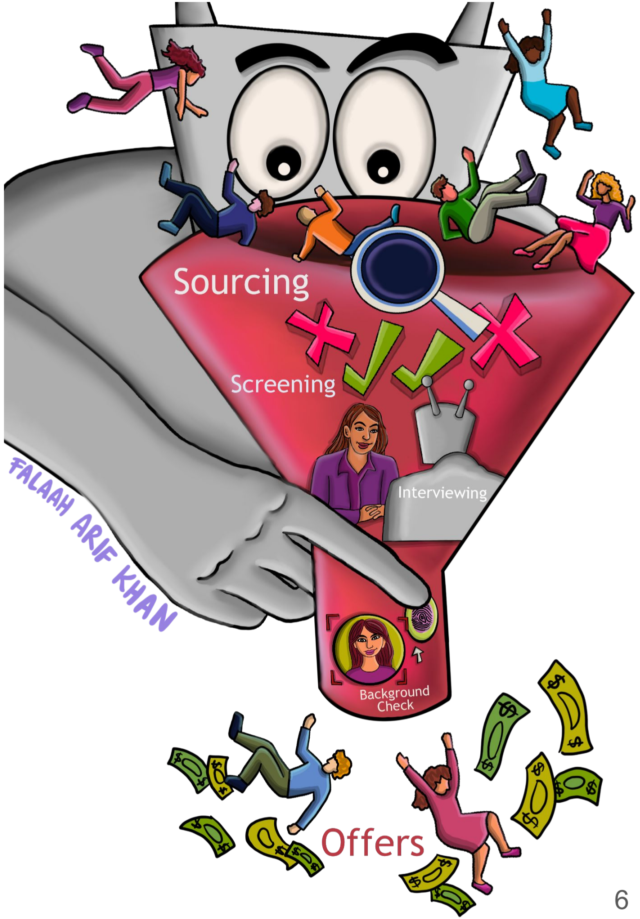
REUTERS **October 2018**

Amazon scraps secret AI recruiting tool that showed bias against women

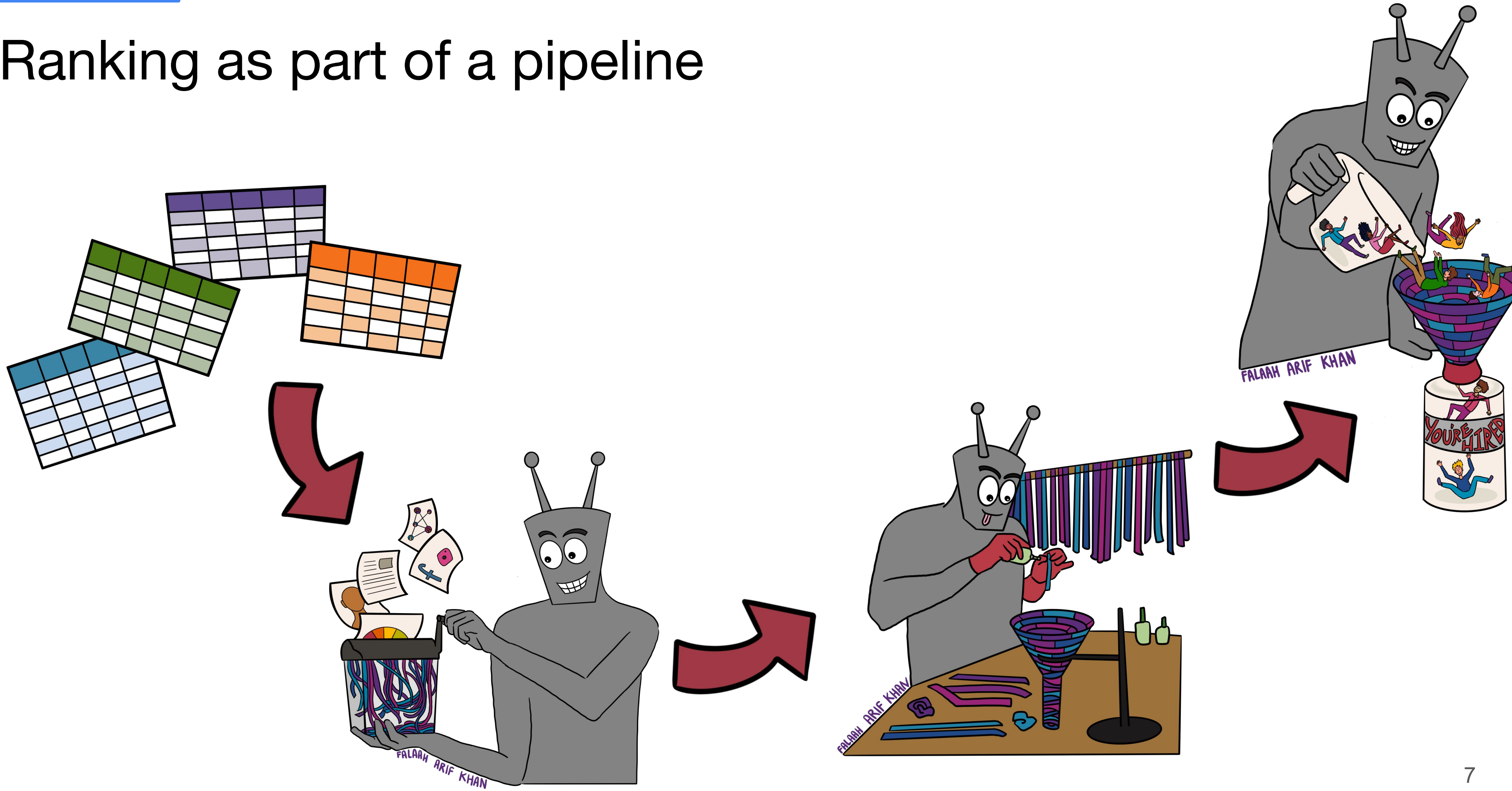
THE WALL STREET JOURNAL. **September 2014**

Are Workplace Personality Tests Fair?

Growing Use of Tests Sparks Scrutiny Amid Questions of Effectiveness and Workplace Discrimination



Ranking as part of a pipeline



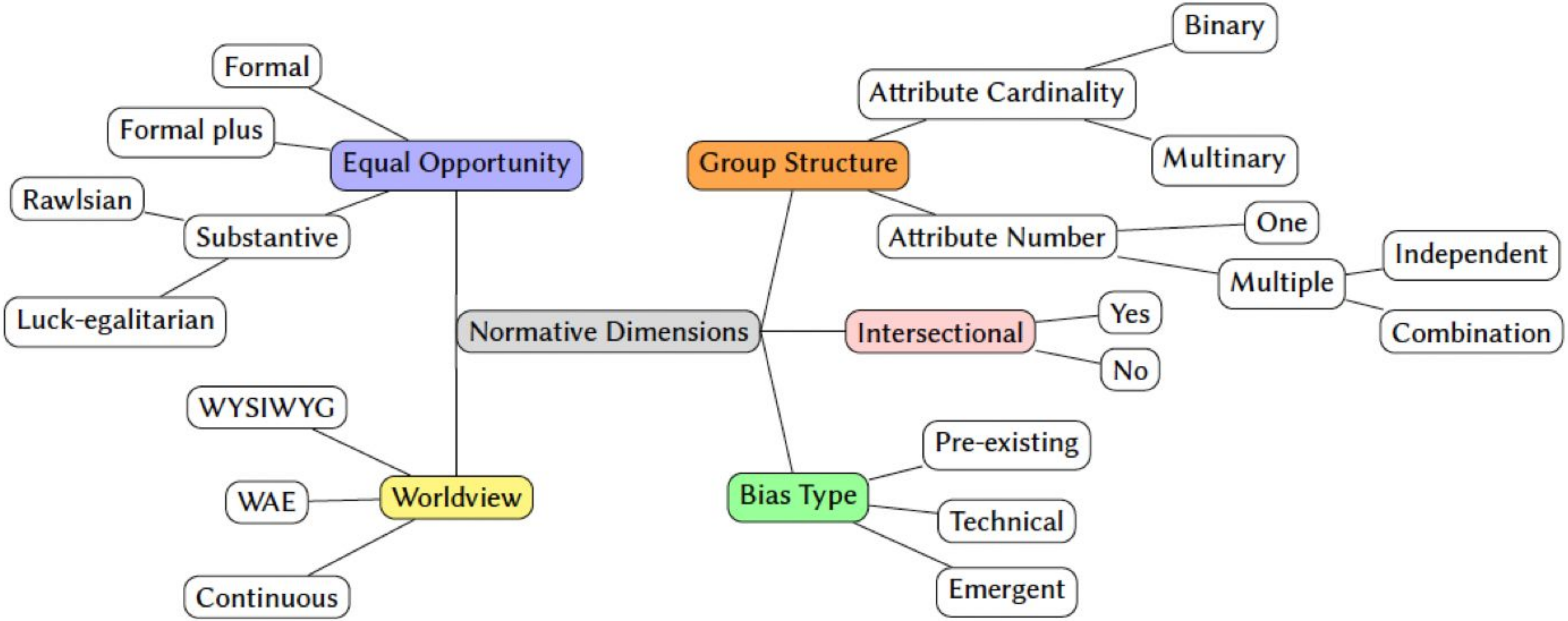
Roadmap

- We present a **classification framework**, unifying fair ranking methods in terms of group structure, type of bias, and mitigation objectives
- We map representative **score-based fair ranking** methods to this framework
- We map representative fair **learning-to-rank methods** to this framework
- We discuss existing **datasets & benchmarks** that have have been used in fair ranking research
- We **conclude** with concrete guidance for practitioners wishing to incorporate fairness objectives into algorithmic rankers

Roadmap

- We present a **classification framework**, unifying fair ranking methods in terms of group structure, type of bias, and mitigation objectives
- We map representative **score-based fair ranking** methods to this framework
- We map representative fair **learning-to-rank methods** to this framework
- We discuss existing **datasets & benchmarks** that have have been used in fair ranking research
- We **conclude** with concrete guidance for practitioners wishing to incorporate fairness objectives into algorithmic rankers

Classification of fair ranking methods



Group structure

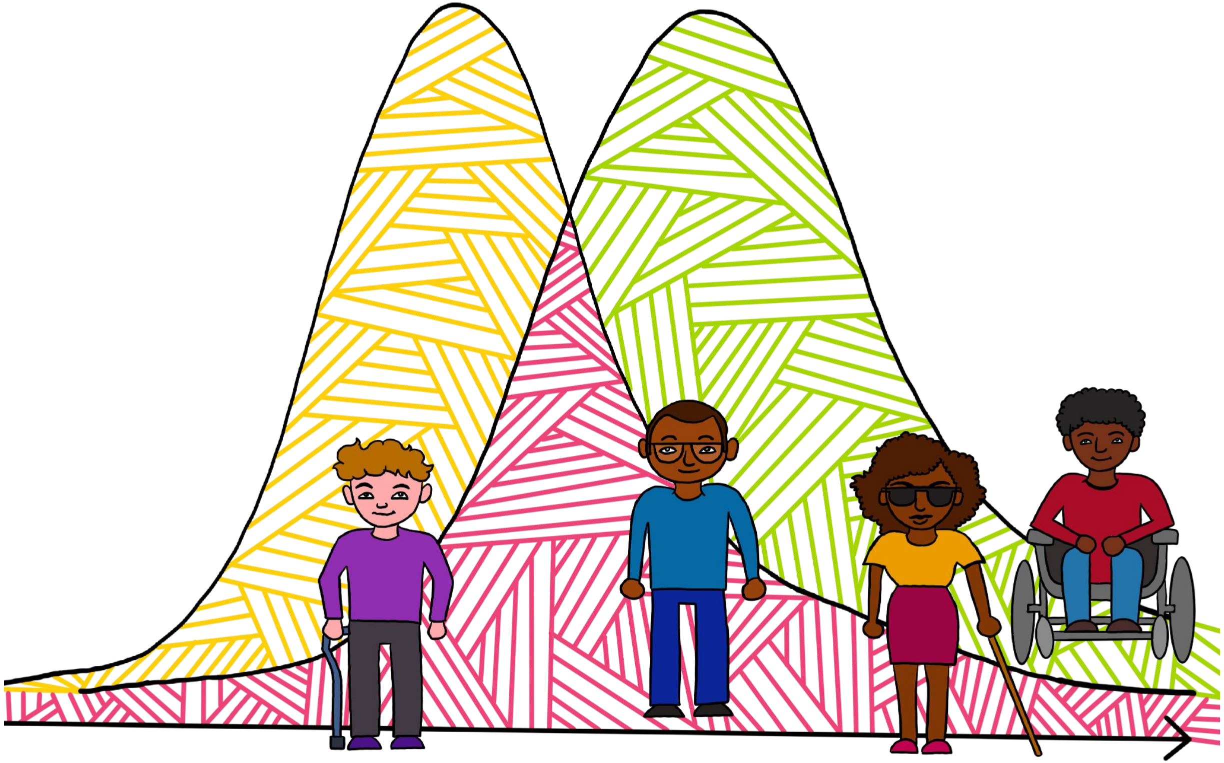
Cardinality of sensitive attributes

- binary (e.g., binary gender, majority / minority ethnicity) vs. multinary
- if multinary, is only one group protected?

Number of attributes

- one sensitive attribute at a time or multiple sensitive attributes simultaneously
- if multiple sensitive attributes, then independently (e.g., fairness for both women and Blacks) vs. in combination (e.g., fairness for Black women)

Intersectional discrimination



Bias type

Pre-existing: independent of the technical system, has origins in society

Technical: introduced or exacerbated by the properties of the technical system

Emergent: arises due to the context of use

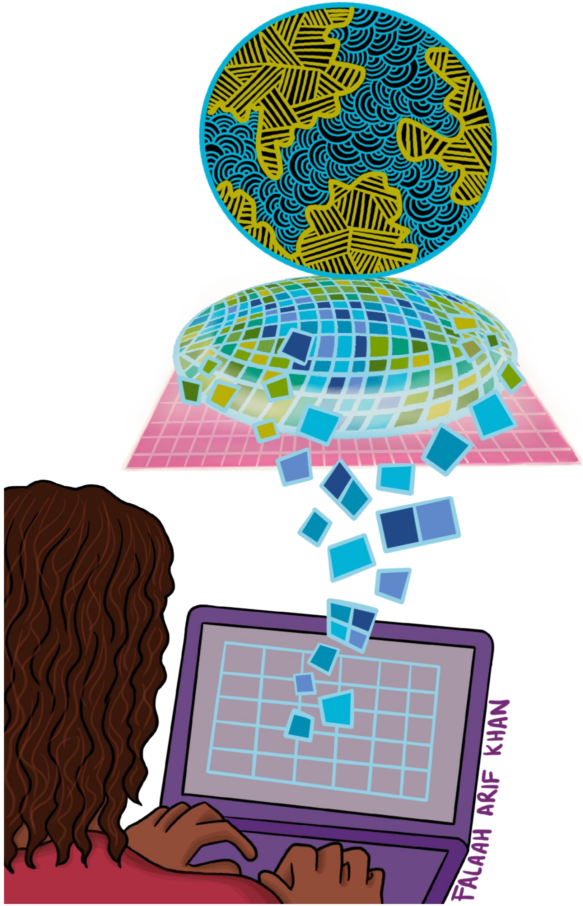
[Friedman & Nissenbaum, 1996]

Bias type: Pre-existing

Pre-existing: independent of the technical system, has origins in society

Technical: introduced or exacerbated by the properties of the technical system

Emergent: arises due to the context of use

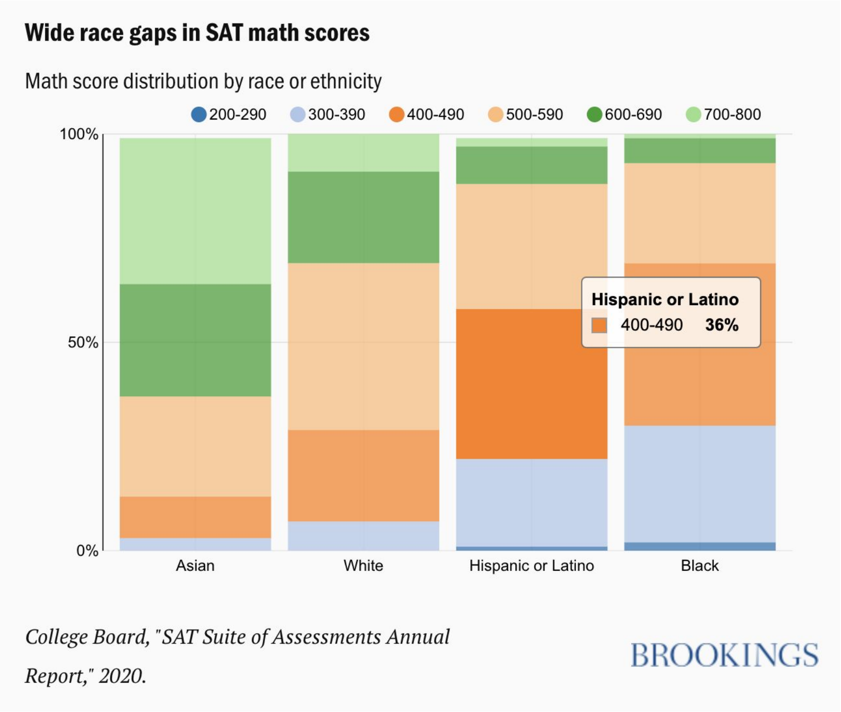


Bias type: Pre-existing

Pre-existing: independent of the technical system, has origins in society

Technical: introduced or exacerbated by the properties of the technical system

Emergent: arises due to the context of use

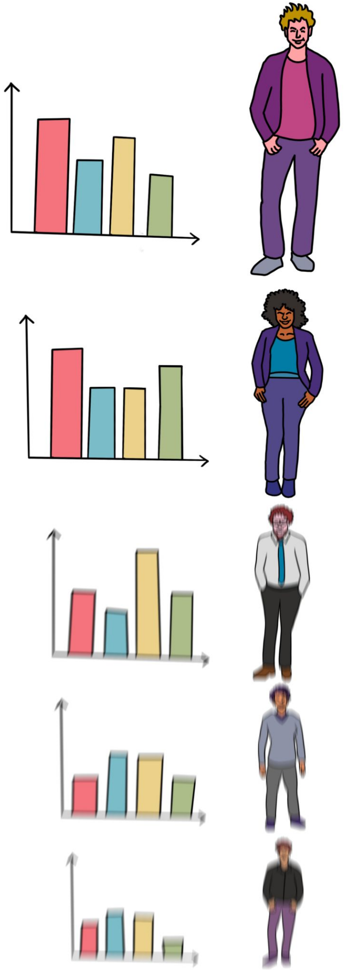


Bias type: Technical

Pre-existing: independent of the technical system, has origins in society

Technical: introduced or exacerbated by the properties of the technical system

Emergent: arises due to the context of use

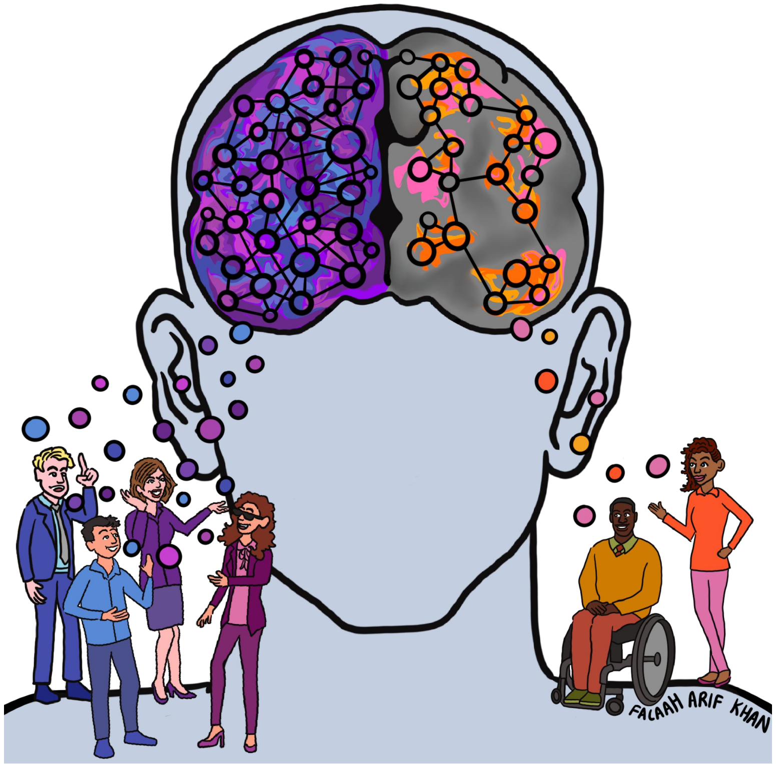


Bias type: Emergent

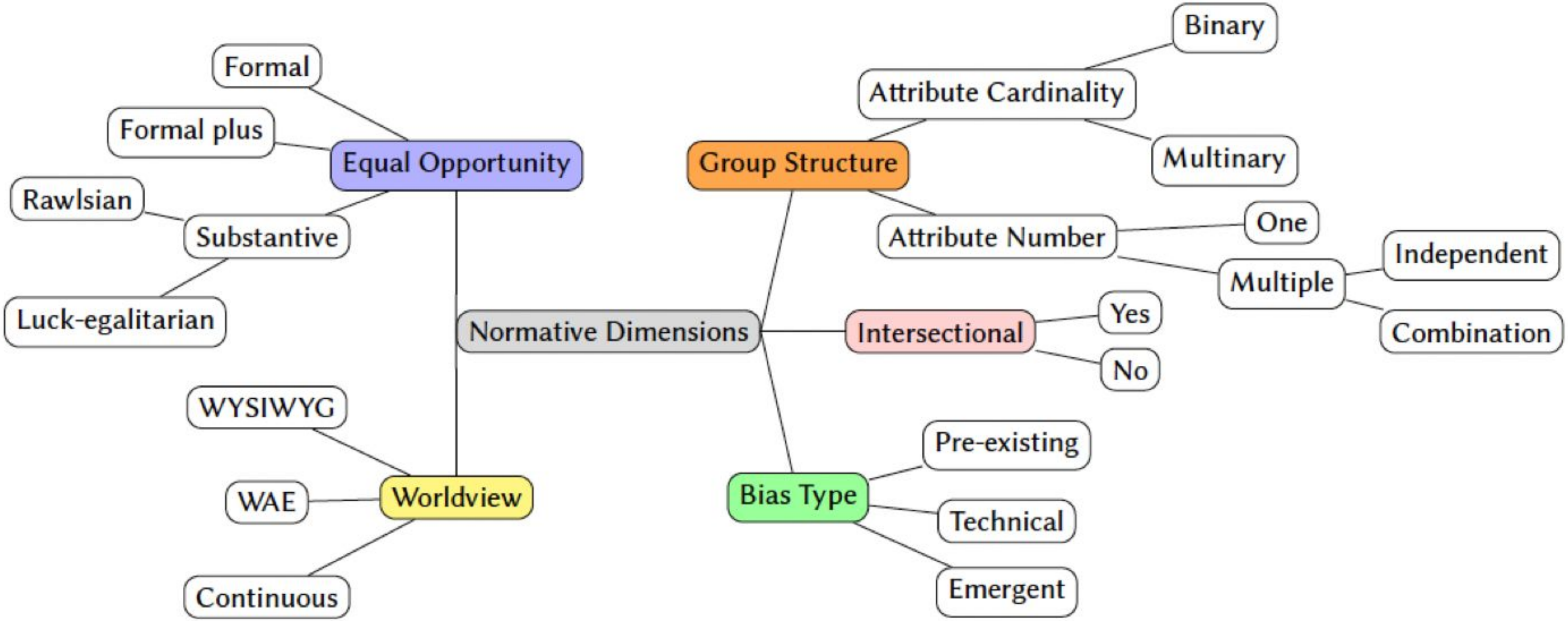
Pre-existing: independent of the technical system, has origins in society

Technical: introduced or exacerbated by the properties of the technical system

Emergent: arises due to the context of use



Classification of fair ranking methods

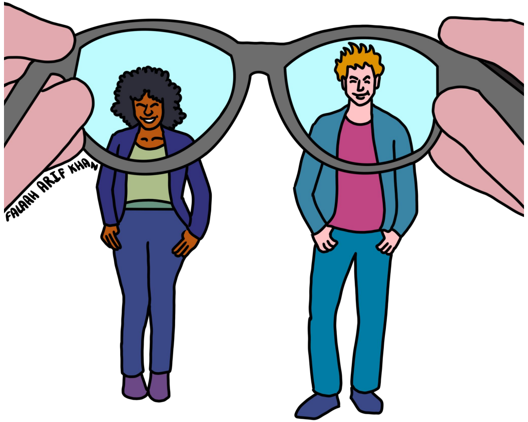


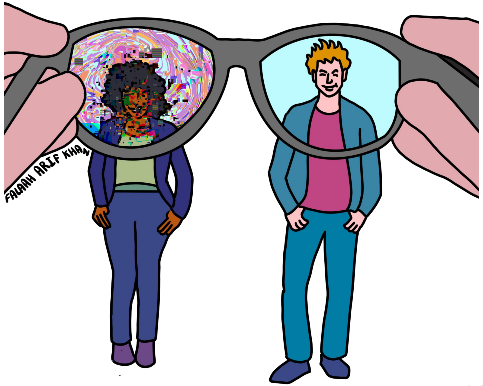
Worldview

WYSIWYG: "What you see is what you get"

WAE: "We are all equal"

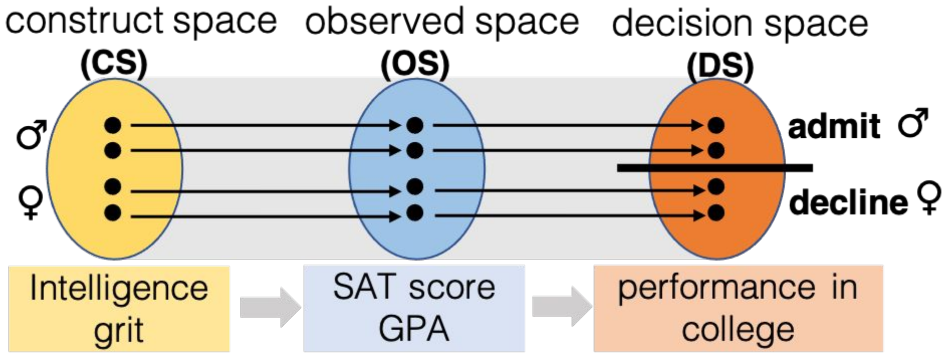
Continuous: interpolating between the two

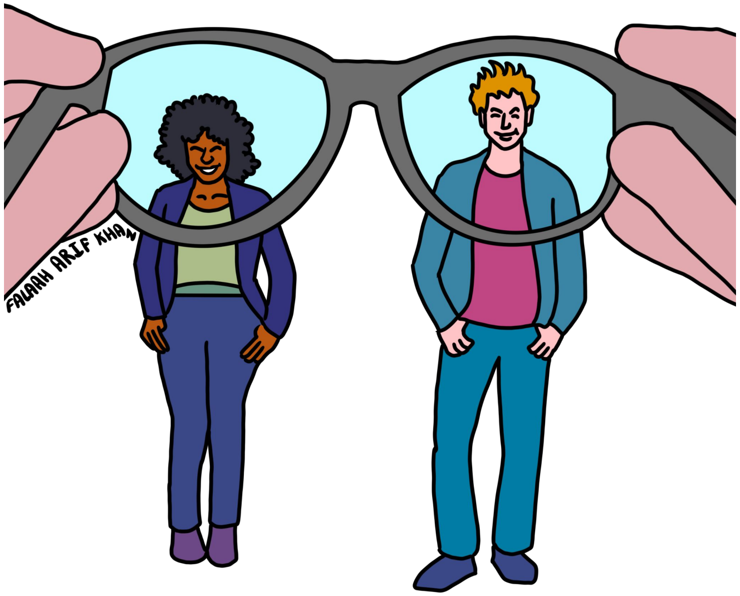




[Friedler, Scheidegger & Venkatasubramanian, 2016]

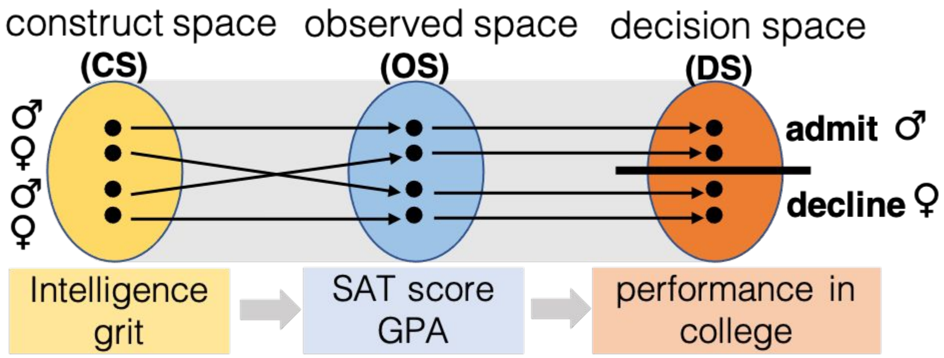
Worldview: WYSIWYG

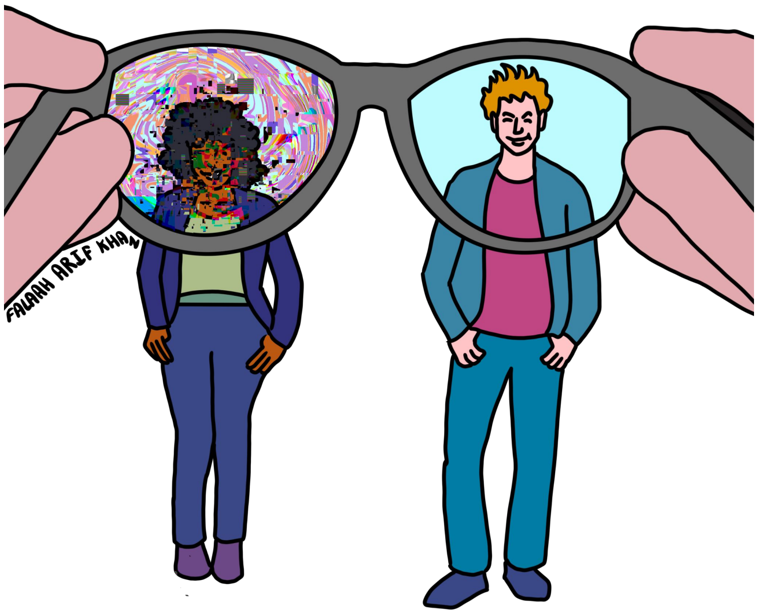




[Friedler, Scheidegger & Venkatasubramanian, 2016]

Worldview: WAE





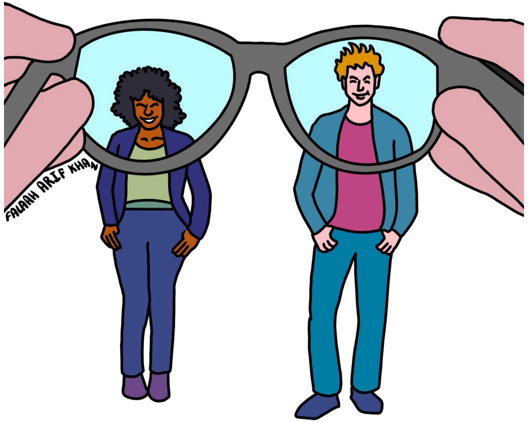
[Friedler, Scheidegger & Venkatasubramanian, 2016]

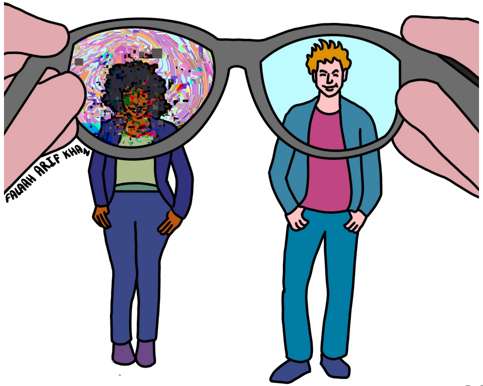
Worldview

WYSIWYG: “What you see is what you get”

WAE: “We are all equal”

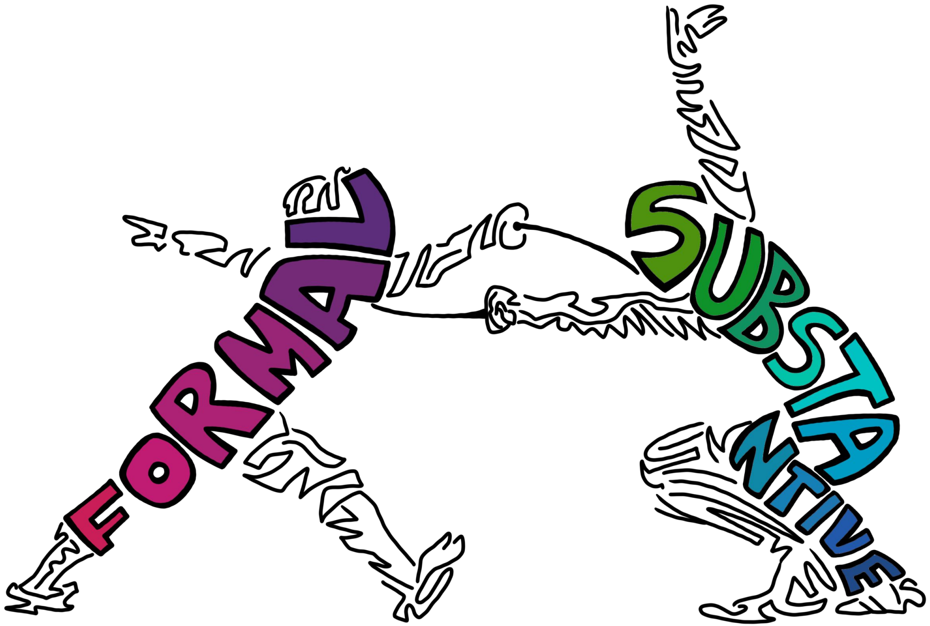
Continuous: interpolating between the two





[Friedler, Scheidegger & Venkatasubramanian, 2016]

Equality of Opportunity (EO) doctrine



[Arif Khan, Manis & Stoyanovich, 2022]

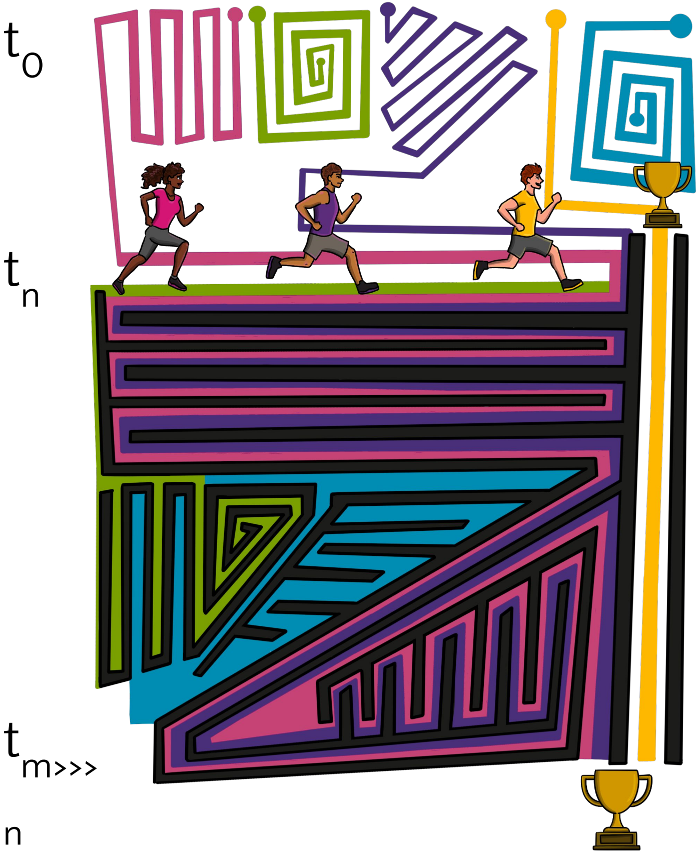
Principles of EO

Fair contests / non-discrimination

Fair life chances (i.e., leveling the playing field)

[Arif Khan, Manis & Stoyanovich, 2022]

Domains of EO



2. Equality in developmental opportunities

1. <u>Fairness at a specific decision point</u>

3. Opportunities over the course of a lifetime

[Arif Khan, Manis & Stoyanovich, 2022]

Formal Equality of Opportunity

"Careers open to talents": applicants should only be judged by relevant qualifications

Fairness through blindness is the most common codification of formal EO

Formal Plus: test performance / validity should not track morally irrelevant disadvantage

[Arif Khan, Manis & Stoyanovich, 2022]

Substantive Equality of Opportunity: Rawls

Equally talented people have equal prospects of success.

Distribute outcomes to improve people's future prospects of success.

[Arif Khan, Manis & Stoyanovich, 2022]

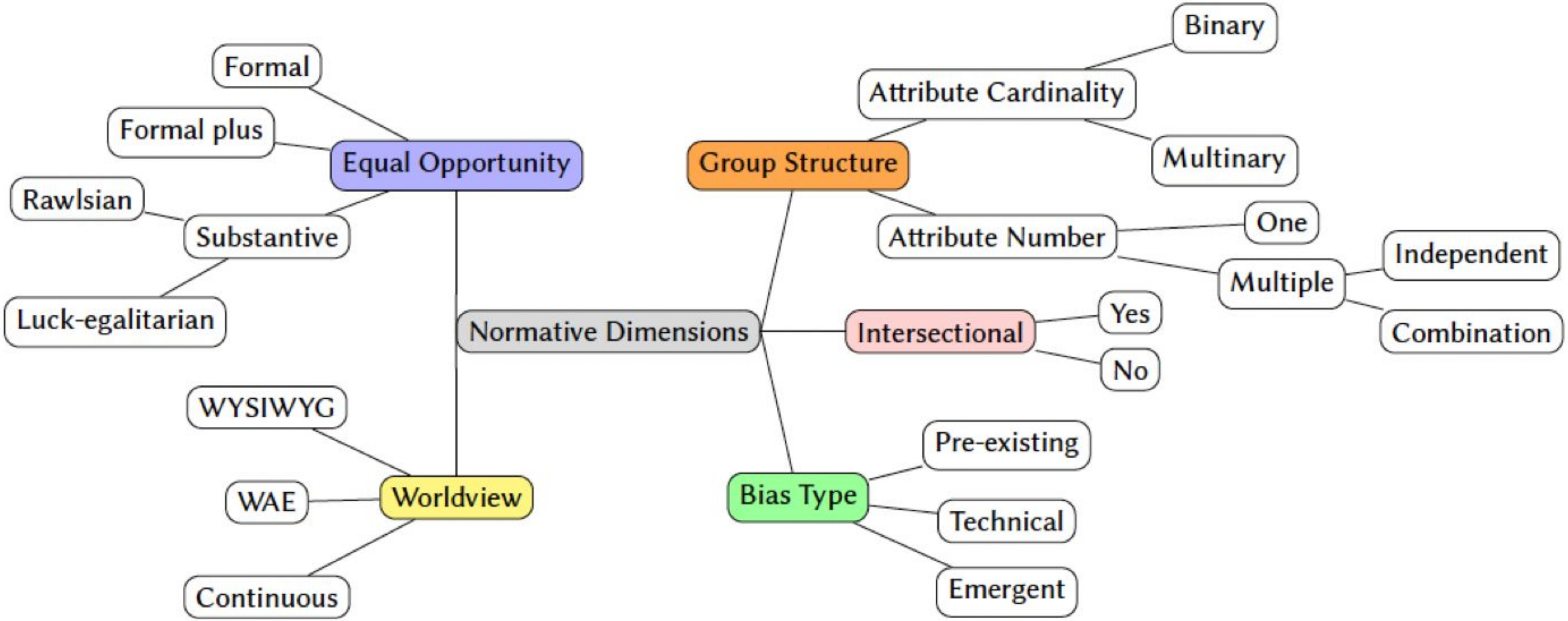
Substantive Equality of Opportunity: luck-egalitarian

Outcomes should only be affected by choice luck (one's responsible choices), not brute-luck (irrelevant circumstance).

But do we make that split?

[Arif Khan, Manis & Stoyanovich, 2021]

Classification of fair ranking methods



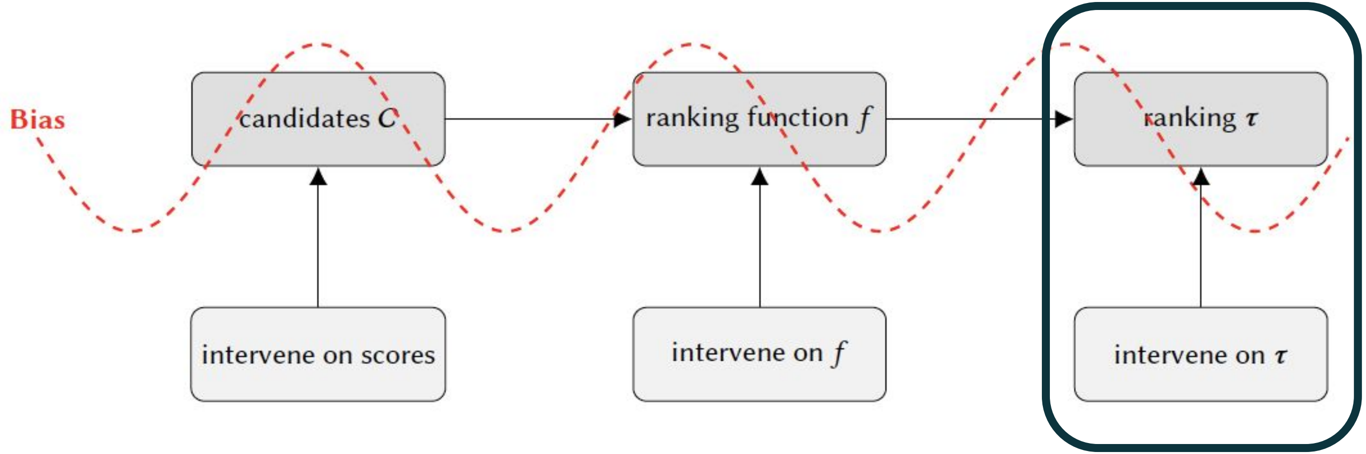
Questions?

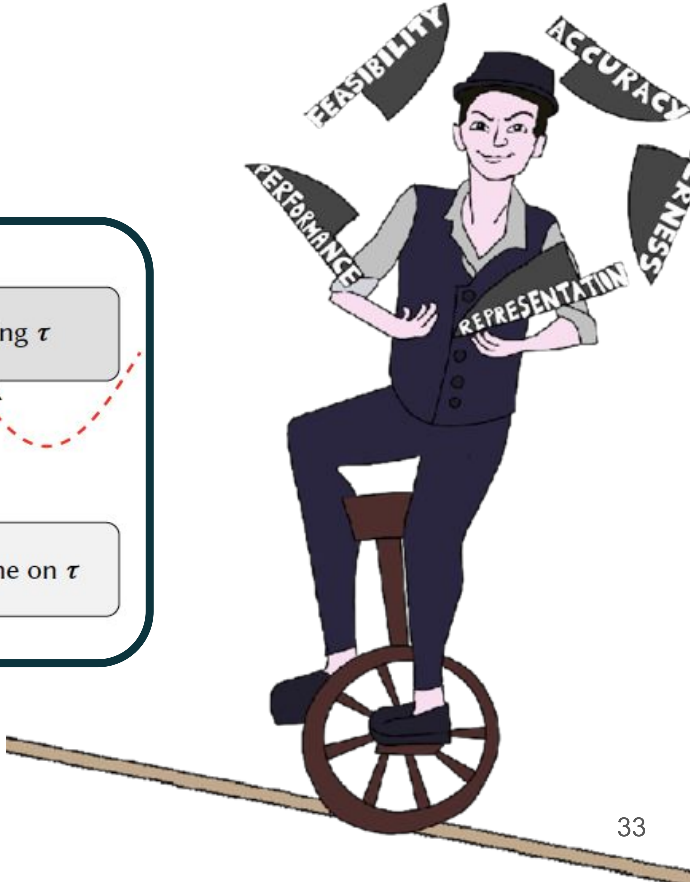
Roadmap

- We present a **classification framework**, unifying fair ranking methods in terms of group structure, type of bias, and mitigation objectives
- We map representative **score-based fair ranking** methods to this framework
- We map representative fair **learning-to-rank methods** to this framework
- We discuss existing **datasets & benchmarks** that have have been used in fair ranking research
- We **conclude** with concrete guidance for practitioners wishing to incorporate fairness objectives into algorithmic rankers

Method	Group structure	Bias	Worldview	EO	Intersectional
Rank-aware proportional representation [80]	one binary sensitive attr.	pre-existing	WAE	luck-egalitarian	no
Constrained ranking maximization [16]	multiple sensitive attrs.; multinary; handled independently	pre-existing	WAE	luck-egalitarian (1 sensitive attr. only)	no
Balanced diverse ranking [78]	multiple sensitive attrs.; multinary; handled independently	pre-existing; technical	WAE	luck-egalitarian	yes
Diverse k-choice secretary [68]	one multinary sensitive attr.	pre-existing	WAE	luck-egalitarian	no
Utility of selection with implicit bias [41]	one binary sensitive attr.	pre-existing; implicit	WAE	N/A	no
Utility of ranking with implicit bias [15]	multiple sensitive attrs.; multinary; handled independently	pre-existing; implicit	WAE	N/A	yes
Causal intersectionally fair ranking [79]	multiple sensitive attrs.; multinary; handled independently	pre-existing	WAE	Rawlsian	yes
Designing fair ranking functions [4]	any	pre-existing	any	any	yes

Bias mitigation methods





Rank-aware proportional representation

τ_1	Y
b	9
c	8
d	7
e	6
f	5
k	4
l	3
o	2

τ_2	Y
b	9
d	7
c	8
f	5
e	6
k	4
l	3
o	2

τ_3	Y
b	9
c	8
d	7
f	5
e	6
l	3
k	4
o	2

Goal: check if candidates' visibility in a ranking depends on their sensitive attributes

Idea:
compute set-wise proportional representation at each prefix of $\boldsymbol{\tau}$

compound values with **position-based discounts**

$$U^k(\boldsymbol{\tau}) = \sum_{i=1}^{k} Y_{\boldsymbol{\tau}(i)} \qquad U^k(\boldsymbol{\tau}) = \sum_{i=1}^{k} \frac{Y_{\boldsymbol{\tau}(i)}}{\log_2(i+1)}$$

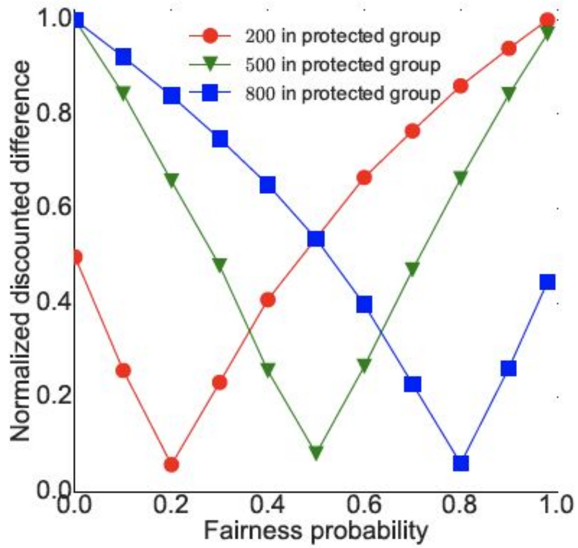
[Yang & Stoyanovich, 2017]

Rank-aware proportional representation

τ_1	Y
b	9
c	8
d	7
e	6
f	5
k	4
l	3
o	2

Idea:
compute set-wise proportional representation at each prefix of $\boldsymbol{\tau}$

compound values with **position-based discounts**



$$\mathrm{rRD}(\tau) = \frac{1}{Z} \sum_{k=10,20,\ldots}^{n} \frac{1}{\log_2 k} \left(\frac{|\tau_{1\ldots k} \cap G_1|}{|\tau_{1\ldots k} \cap G_2|} - \frac{|G_1|}{|G_2|} \right)$$

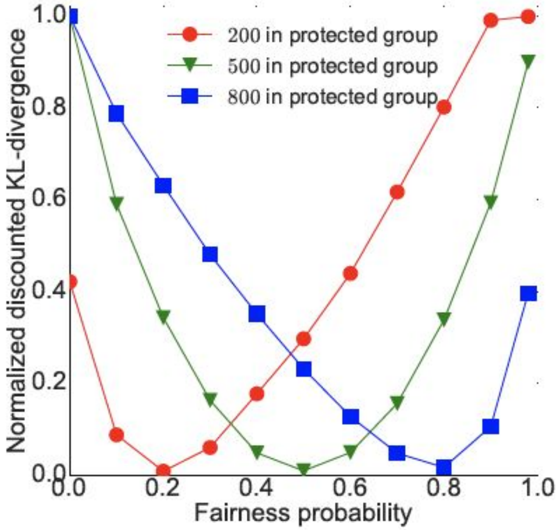
[Yang & Stoyanovich, 2017]

Rank-aware proportional representation

τ_1	Y
b	9
c	8
d	7
e	6
f	5
k	4
l	3
o	2

Idea:
compute set-wise proportional representation at each prefix of $\boldsymbol{\tau}$

compound values with **position-based discounts**



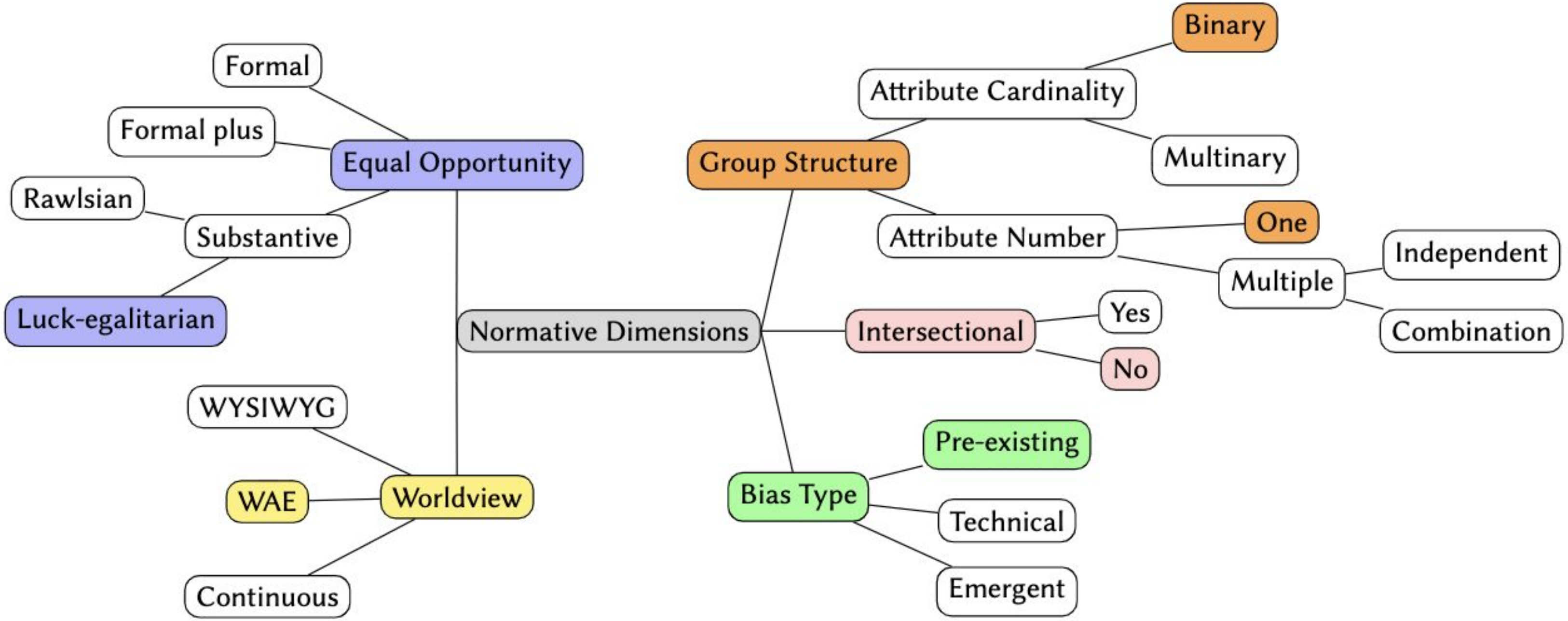
$$P_k = \left(\frac{|\tau_{1...k} \cap \mathcal{G}_1|}{k}, \frac{|\tau_{1...k} \cap \mathcal{G}_2|}{k} \right)$$

$$Q = \left(\frac{|\mathcal{G}_1|}{n}, \frac{|\mathcal{G}_2|}{n} \right)$$

$$\mathrm{rKL}(\boldsymbol{\tau}) = \frac{1}{Z} \sum_{k=10,20,\ldots}^{n} \frac{1}{\log_2 k} D_{KL}(P_k || Q)$$

[Yang & Stoyanovich, 2017]

Rank-aware proportional representation



[Yang & Stoyanovich, 2017]

Constrained ranking maximization

	gender	race	Y
a	m	w	19
b	m	w	18
c	f	w	16
d	f	w	15
e	m	b	11
f	m	b	11
g	f	b	10
h	f	b	9
i	m	a	7
j	m	a	7
k	f	a	6
l	f	a	3

Goals

diversity: pick k=4 candidates, with two of each gender and at least one of each race

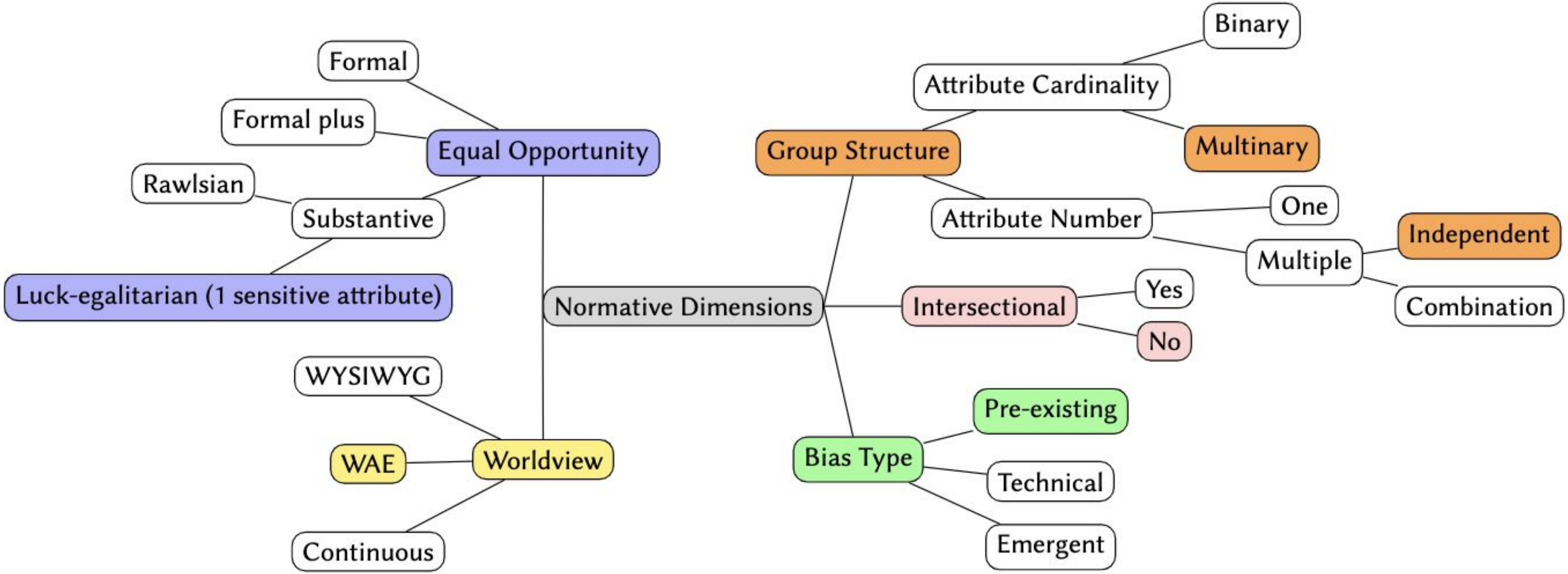
utility: maximize the sum of scores of the selected candidates

Insights

A hard problem when candidates have two or more sensitive attributes

[Celis, Straszak & Vishnoi, 2018]

Constrained ranking maximization



[Celis, Straszak & Vishnoi, 2018]

Balanced diverse ranking

	gender	race	Y	
a	m	w	19	✔
b	m	w	18	✔
c	f	w	16	
d	f	w	15	
e	m	b	11	
f	m	b	11	
g	f	b	10	✔
h	f	b	9	
i	m	a	7	
j	m	a	7	
k	f	a	6	✔
l	f	a	3	

Goals

diversity: pick k=4 candidates, with two of each gender and at least one of each race

utility: maximize the sum of scores of the selected candidates

Problem

Picked the highest scoring male and White candidates (**a** and **b**), but not the highest scoring female (**c** and **d**), Black (**e** and **f**) or Asian (**i** and **j**) candidates.

[Yang, Gkatzelis & Stoyanovich, 2019]

Balanced diverse ranking

	gender	race	Y		
a	m	w	19	✔	✔
b	m	w	18	✔	
c	f	w	16		✔
d	f	w	15		
e	m	b	11		✔
f	m	b	11		
g	f	b	10	✔	
h	f	b	9		
i	m	a	7		
j	m	a	7		
k	f	a	6	✔	✔
l	f	a	3		

[Yang, Gkatzelis & Stoyanovich, 2019]

Goals

diversity: pick k=4 candidates, with two of each gender and at least one of each race

fairness: admit the most qualified candidates of each gender and race

utility: maximize the sum of scores of the selected candidates

Beliefs

effort is relative: scores are more informative within a group than across groups

it is important to **reward effort**

Balancing utility loss: IGF-Ratio, IGF-Agg

c	f	16
d	f	15
g	f	10
h	f	9
k	f	6
l	f	3

highest-scoring skipped

IGF-Ratio(**f**)=10/16

lowest-scoring selected

a	m	19
b	m	18
e	m	11
f	m	11
i	m	7
j	m	7

lowest-scoring selected

IGF-Ratio(**m**)=1

highest-scoring skipped

IGF-Ratio(**w**)=1

IGF-Ratio(**a**)=6/7

IGF-Ratio(**b**)=10/11

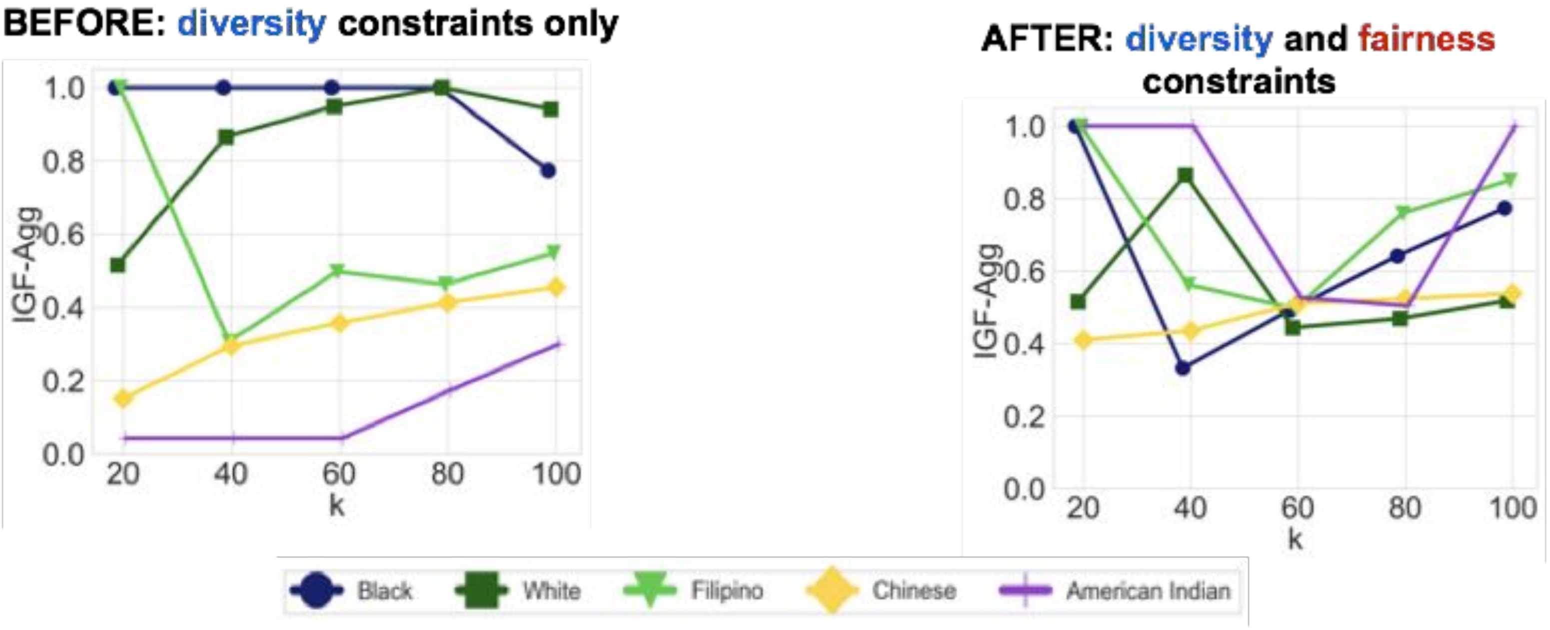
Beliefs

effort is relative: scores are more informative within a group than across groups

it is important to **reward effort**

[Yang, Gkatzelis & Stoyanovich, 2019]

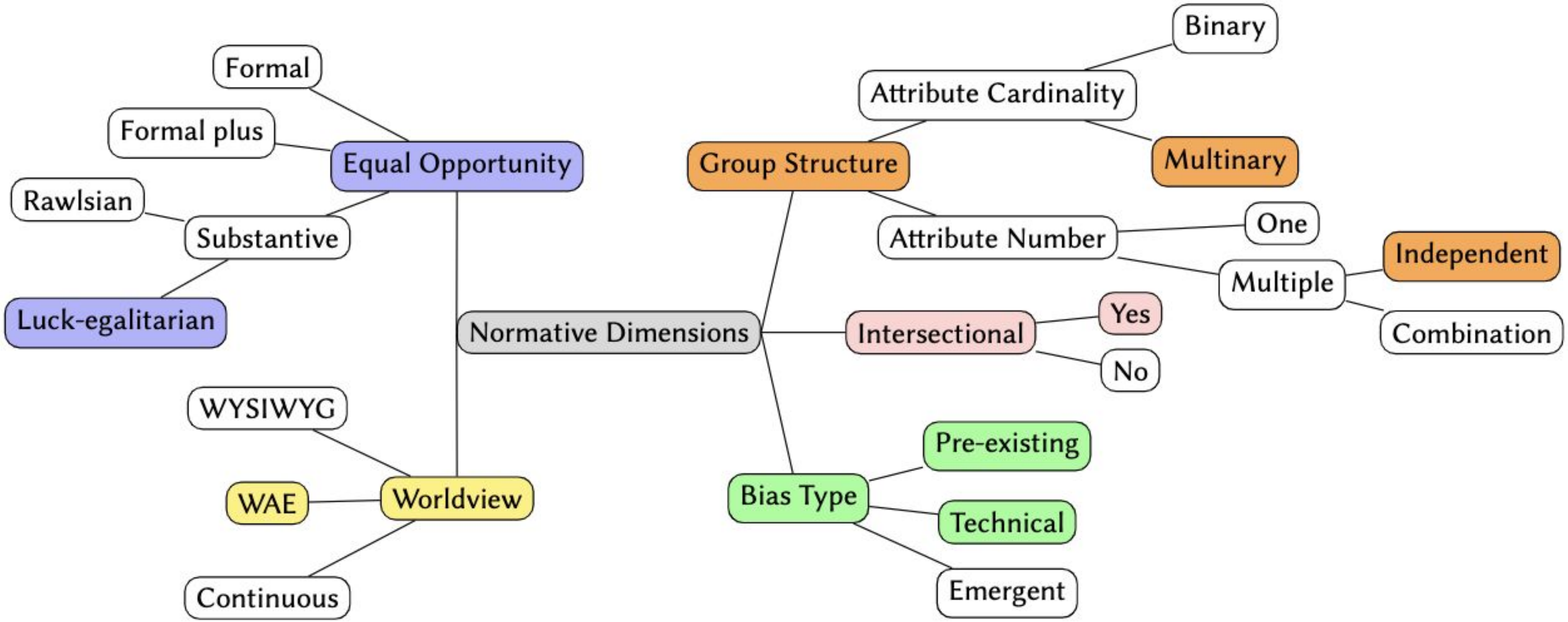
Balancing utility loss: IGF-Ratio, IGF-Agg, ILP magic



MEPS (Medical Expenditure Panel Survey)

[Yang, Gkatzelis & Stoyanovich, 2019]

Balanced diverse ranking

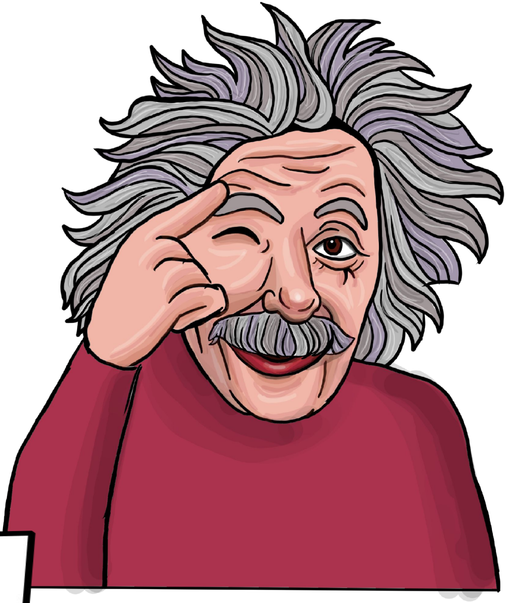


[Yang, Gkatzelis & Stoyanovich, 2019]

Constrained ranking maximization vs. Balanced diverse ranking

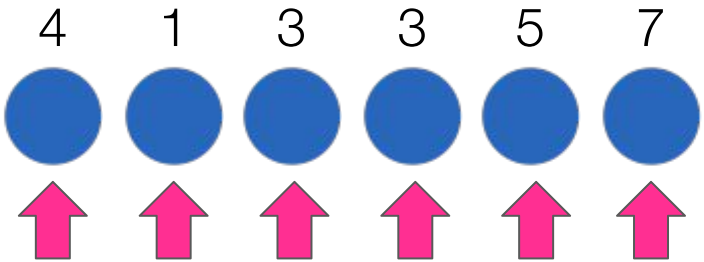
Main difference: assumptions about whether score ("effort") should be measured in absolute terms or per group (relative to "circumstance")

An example where a **small technical difference** encodes a **major difference in values**: substantive EO vs. no EO at all!



Failing to balance utility loss across groups leads to **intersectional discrimination**

Hiring a job candidate



Goal: hire a candidate with a high score

Online setting:

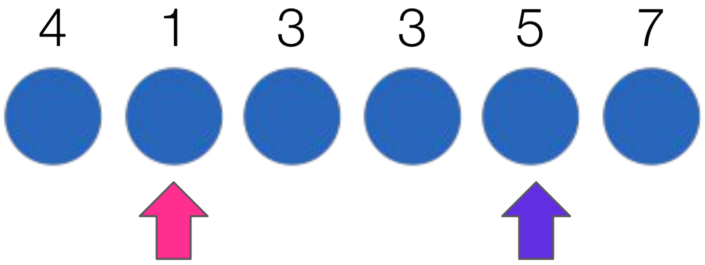
candidates arrive one-by-one, score is revealed when the candidate arrives

candidates arrive in score-independent order

decision to hire or reject must be made before considering the next candidate

[Lindley, 1961; Dynkin, 1963]

The secretary problem



$N = 6$

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

$T = 4$

[Ferguson, 1989]

Goal: pick one element of a randomly ordered sequence to maximize the probability of picking the maximum element of the entire sequence

Online setting:

candidates arrive one-by-one, score is revealed when the candidate arrives

candidates arrive in score-independent order

decision to hire or reject must be made before considering the next candidate

Diverse *k*-choice secretary

Goals

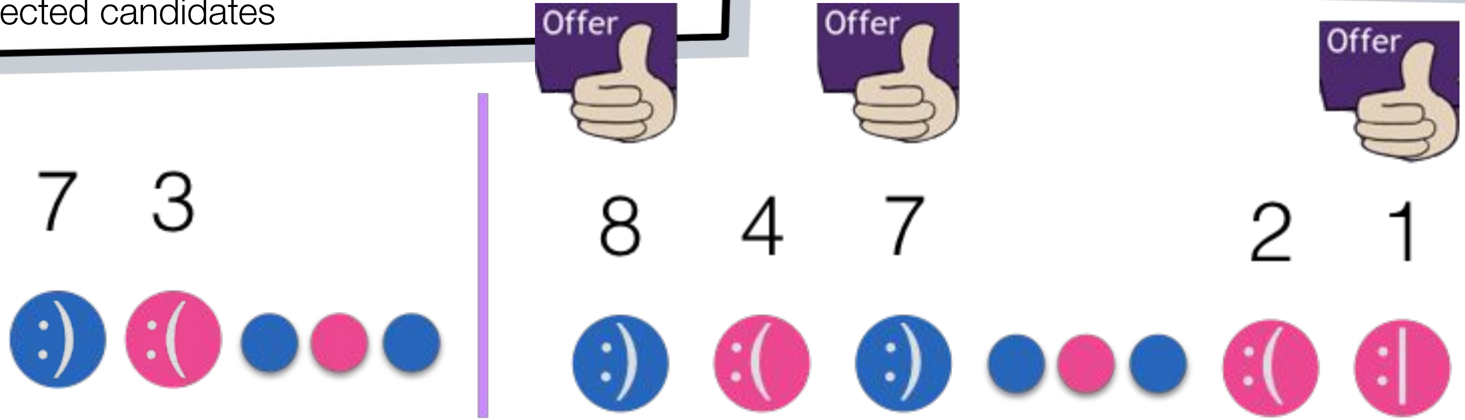
diversity: pick **k=3** candidates, with at least one of each gender

utility: maximize the sum of scores of the selected candidates

Beliefs

effort is relative: scores are more informative within a group than across groups

it is important to **reward effort**



[Stoyanovich, Yang & Jagadish 2018]

Diverse *k*-choice secretary

Goals

diversity: pick **k=3** candidates, with at least one of each gender

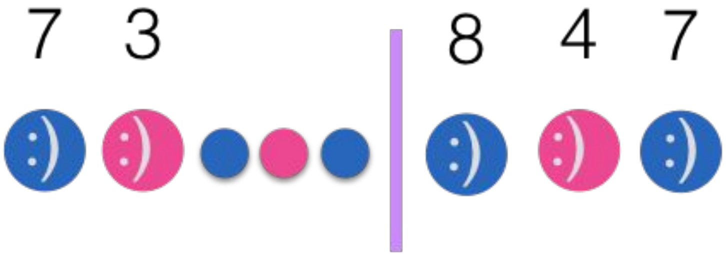
utility: maximize the sum of scores of the selected candidates

Idea: learn what a good candidate looks like separately for each category!

Beliefs

effort is relative: scores are more informative within a group than across groups

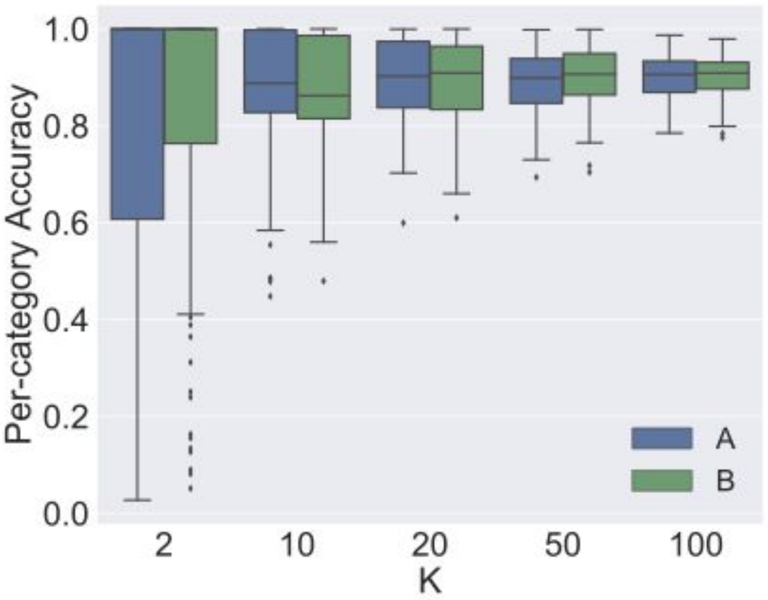
it is important to **reward effort**



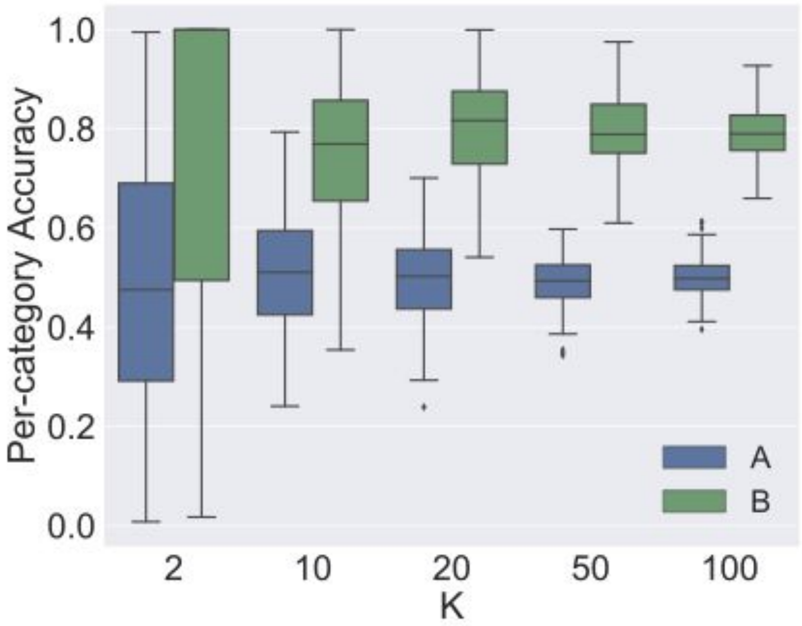
[Stoyanovich, Yang & Jagadish 2018]

Diverse *k*-choice secretary

per-group warm up

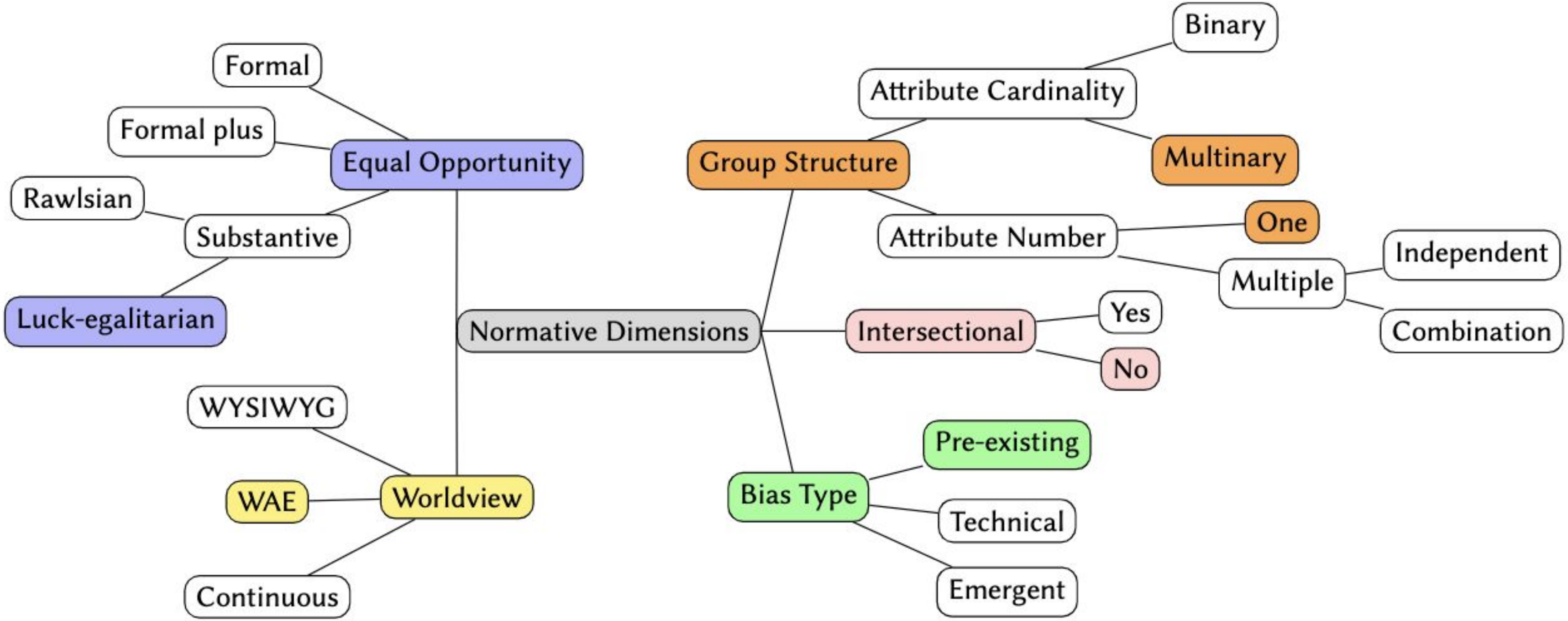


common warm up



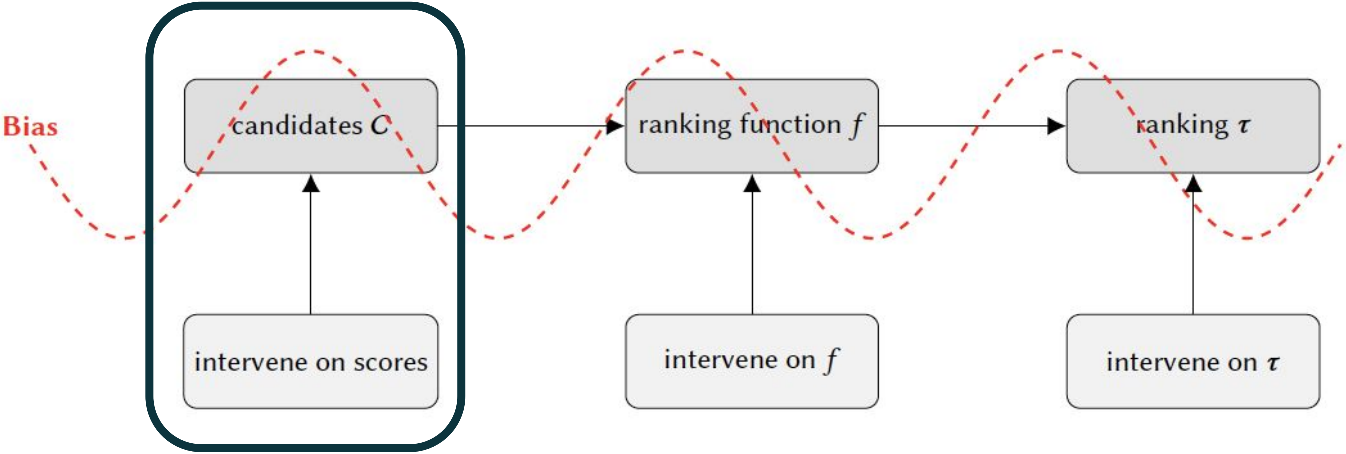
[Stoyanovich, Yang & Jagadish 2018]

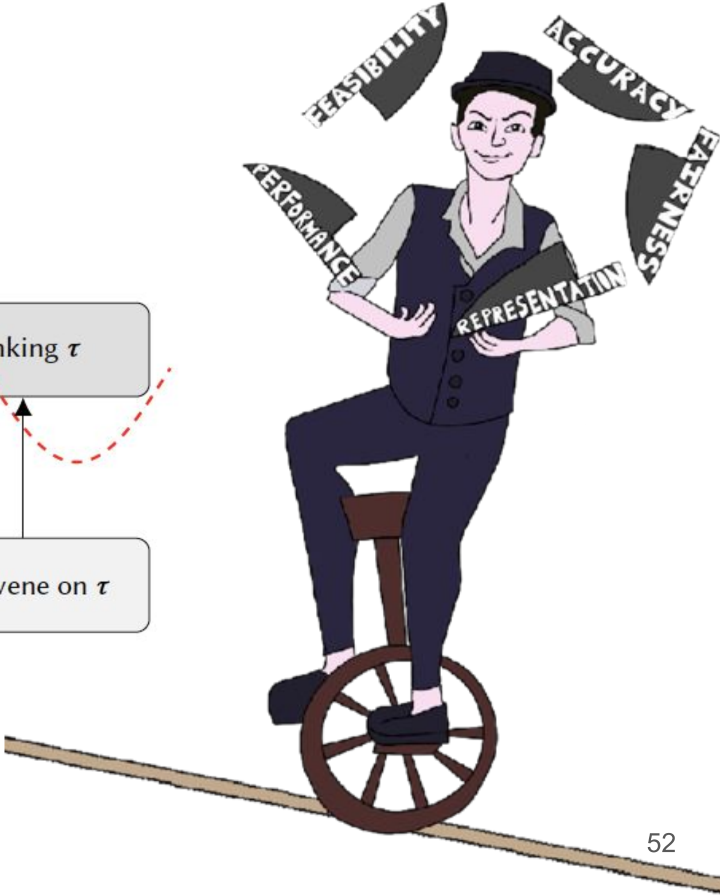
Diverse *k*-choice secretary



[Stoyanovich, Yang & Jagadish 2018]

Bias mitigation method





Set selection with implicit bias

	gender	Y'		Y
b	m	12		12
c	m	9		9
d	f	12	>	8
e	m	7		7
f	f	9	>	6
k	m	5		5
l	m	3		3
o	m	2		2

Goal: pick **k = 2** best-qualified candidates for **an open job position**

Problem: hiring committee uses perceived score **Y** rather than true qualification score **Y'**

Implicit bias: **Y'** → **Y** differently depending on gender

Population factor α: $\alpha = |f| / |m|$, $\alpha < 0$

Bias factor β : $\mathbf{Y = Y'/\beta}$, $\beta > 1$ for female

apply Rooney rule

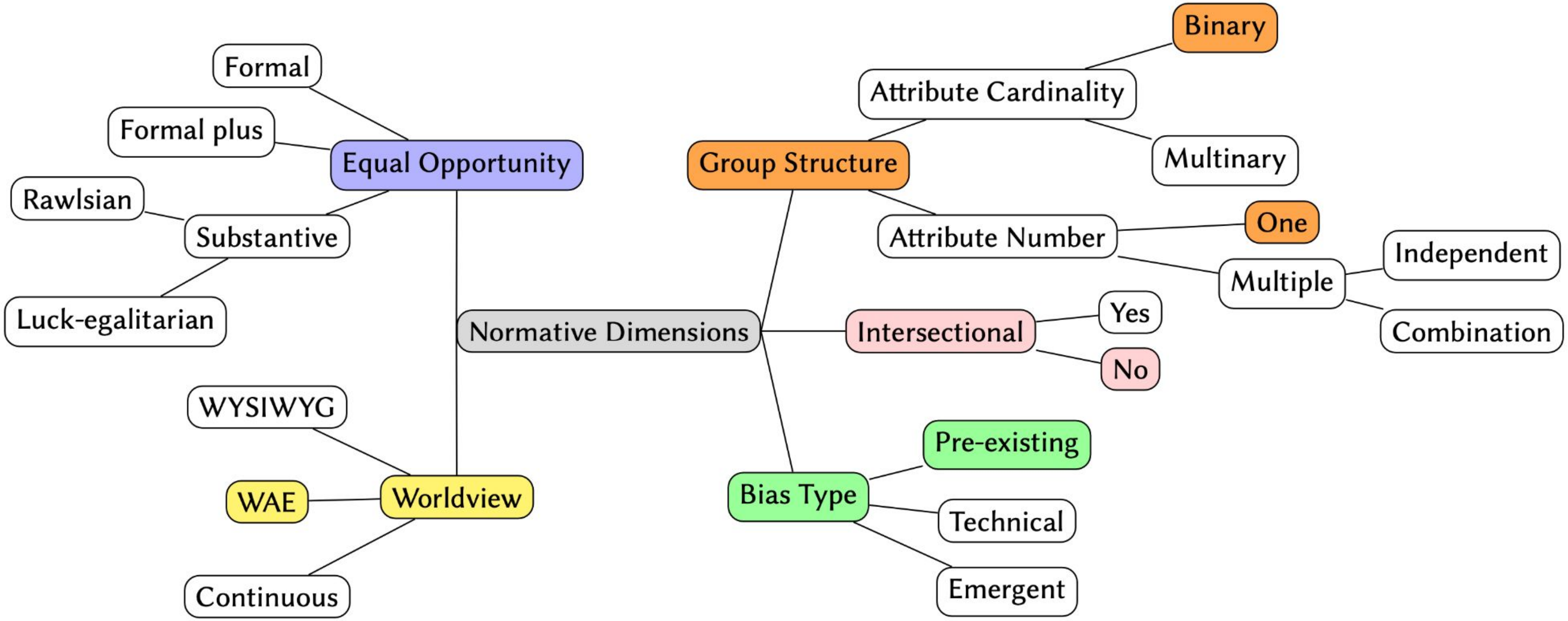
	Y
b	12
c	9

→

	Y
b	12
d	8

[Kleinberg & Raghavan 2018]

Set selection with implicit bias



[Kleinberg & Raghavan 2018]

Ranking with implicit bias

	gender	Y		Y'
b	m	12		12
c	m	9		9
d	f	12	>	8
e	m	7		7
f	f	9	>	6
k	f	8	>	5
l	m	3		3
o	f	2	>	1

τ_1	Y'
b	12
c	9
d	8
e	7
f	6
k	5
l	3
o	1

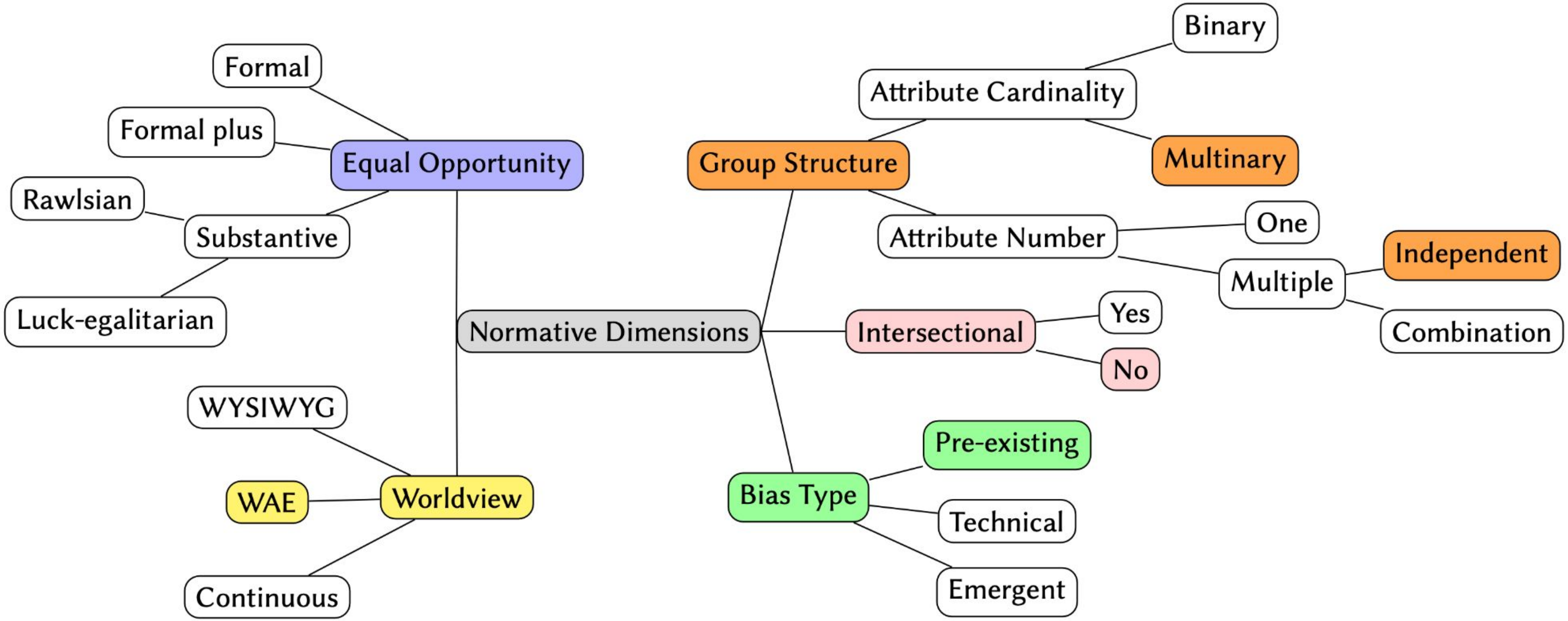
representation constraints

τ_2	Y
b	12
d	12
c	9
f	9
k	8
e	7
l	3
o	2

Insight: representation constraints lead to optimal utility on true qualification score **Y**

[Celis, Mehrotra & Vishnoi 2020]

Ranking with implicit bias



[Celis, Mehrotra & Vishnoi 2020]

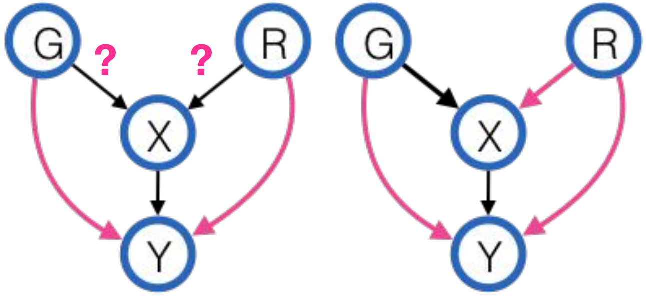
Intersectional causal fairness

	gender	race	X	Y
b	m	w	6	12
c	m	a	5	9
d	f	w	6	8
e	m	w	4	7
f	f	a	3	6
k	f	b	5	5
l	m	b	1	3
o	f	w	1	1

Goal: pick **k = 4** best-qualified candidates to work **at a moving company**

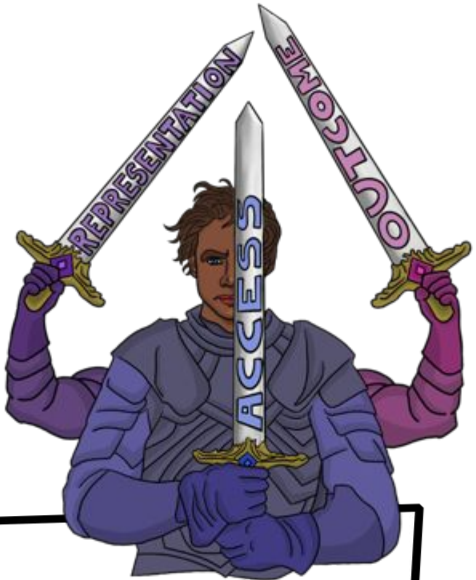
Problem: weight lifting ability **X** maps to qualification score **Y** differently depending on gender

Beliefs



[Yang, Loftus & Stoyanovich 2020]

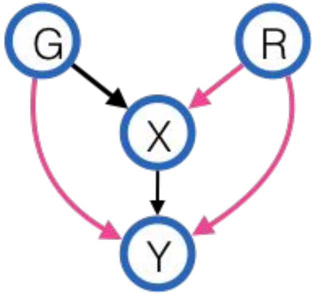
Intersectional causal fairness



Idea: Compute counterfactual scores, treating each individual as though they had belonged to one intersectional group (e.g., Black women).

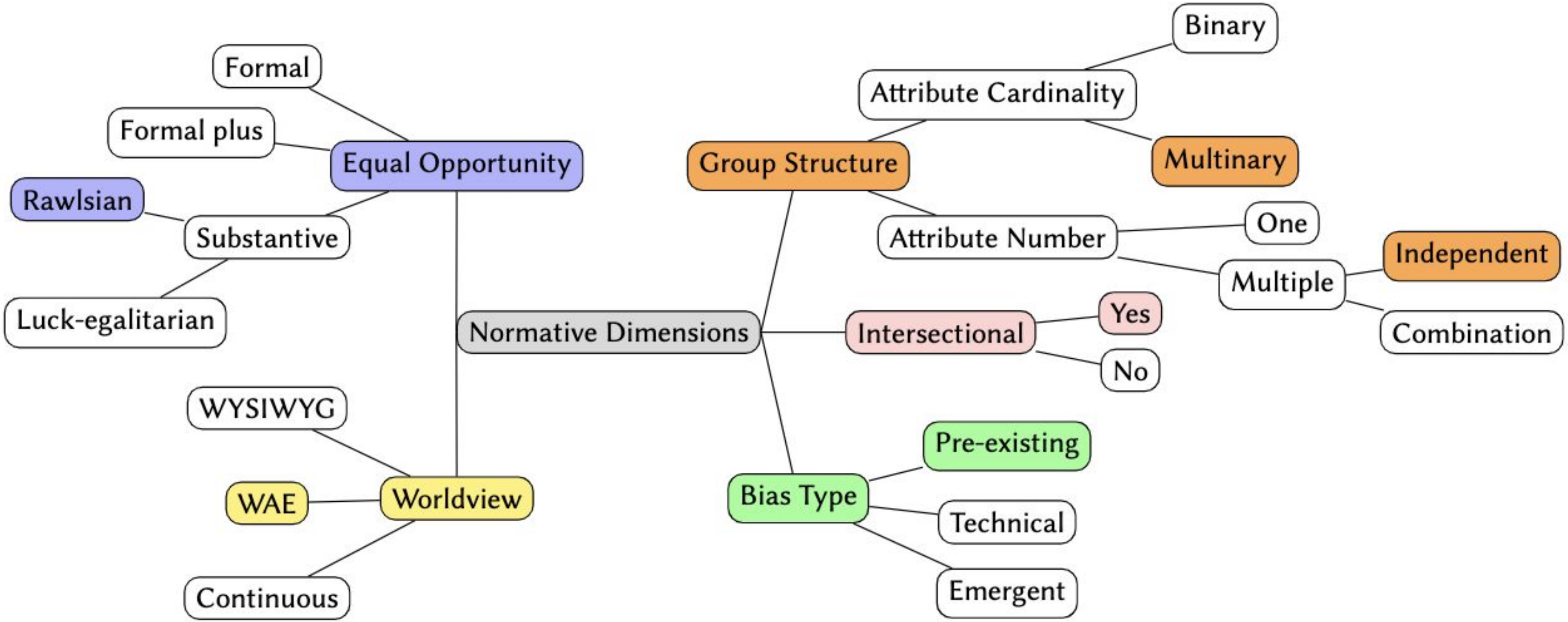
Rank on those scores. This will produce a **counterfactually fair ranking**

Beliefs



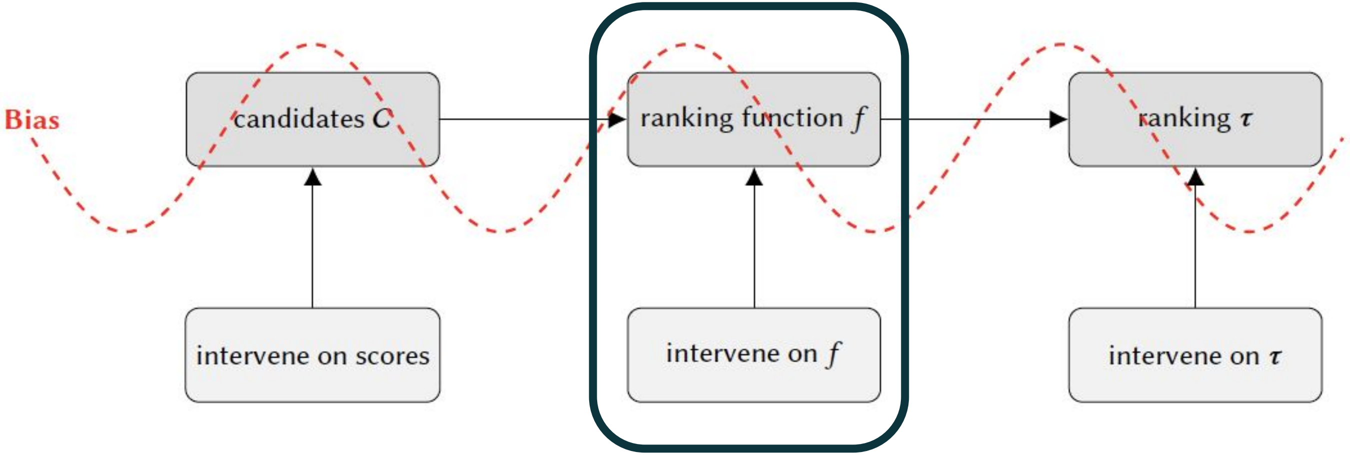
[Yang, Loftus & Stoyanovich 2020]

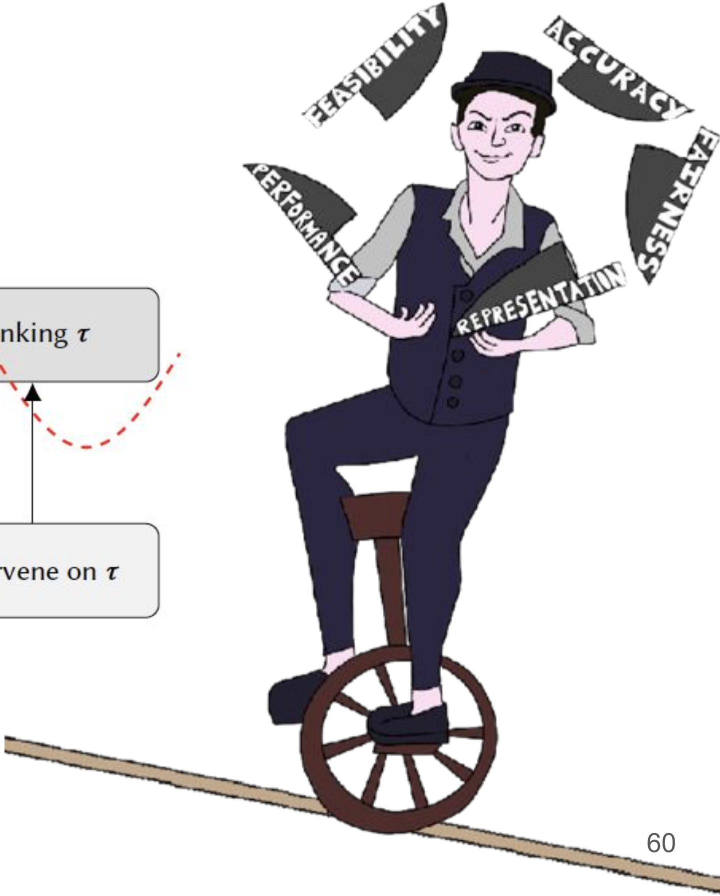
Intersectional causal fairness



[Yang, Loftus & Stoyanovich 2020]

Bias mitigation methods





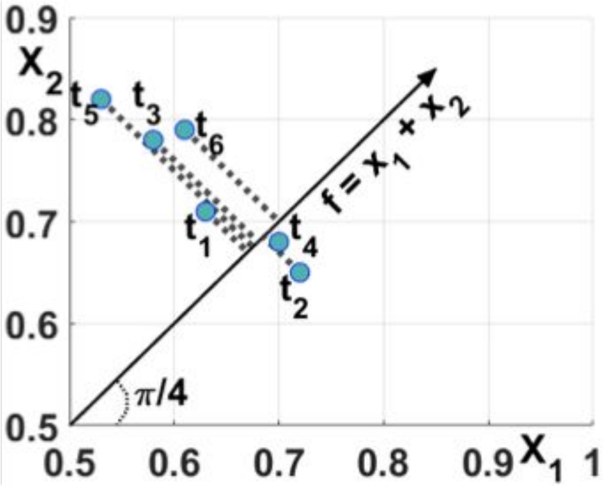
Designing fair rankers

$\mathcal{D}$			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

Goals find a ranking function ***f'***

utility: with similar weights as ***f*** - the function that the human decision-maker had in mind (minimize angular distance)

fairness: ***f'*** should be fair according to an oracle ***O***



[Asudeh, Jagadish, Stoyanovich & Das, 2019]

Designing fair rankers

$\mathcal{D}$			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

Goals find a ranking function

utility: with similar weights as what the human decision-maker had in mind

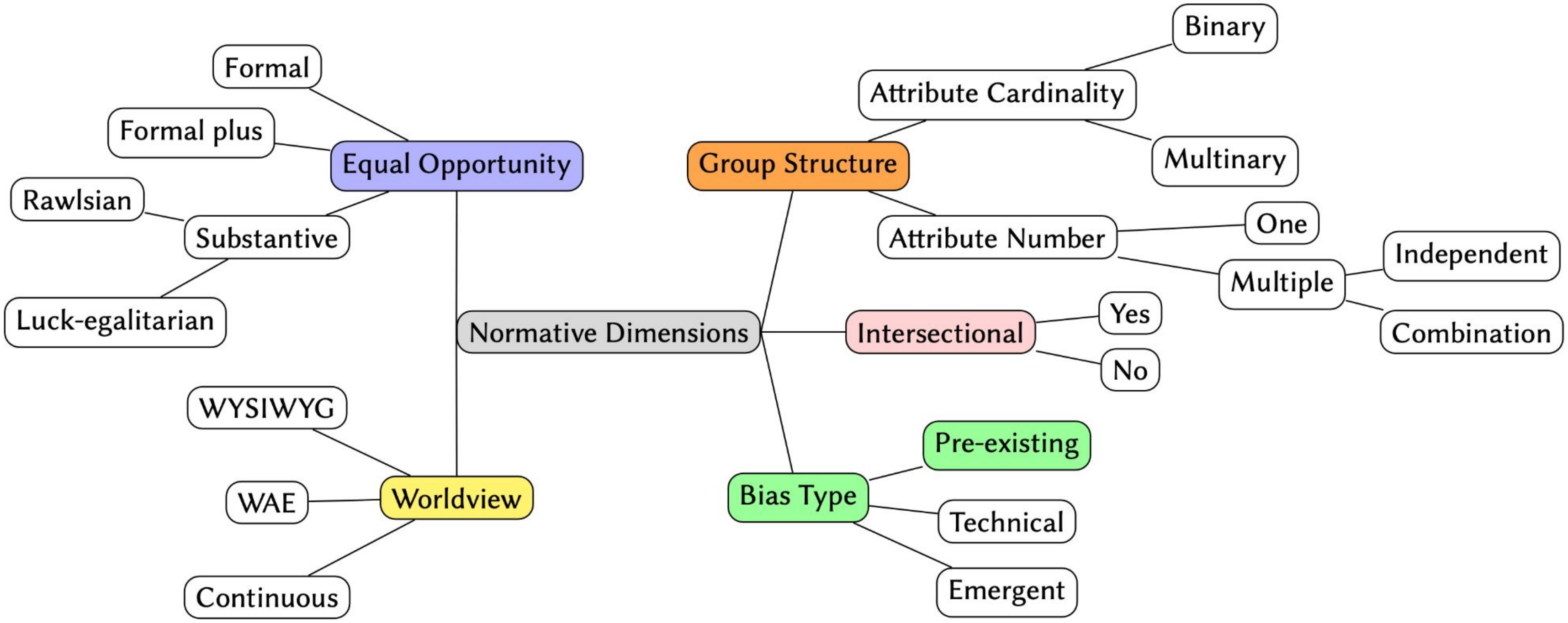
fairness: so that the ranking is fair according to an oracle ***O***

Idea: ordering exchange

Only look at the ranking functions ***f'*** that change the relative order between some pair of points. These are the functions where the oracle may change its mind.

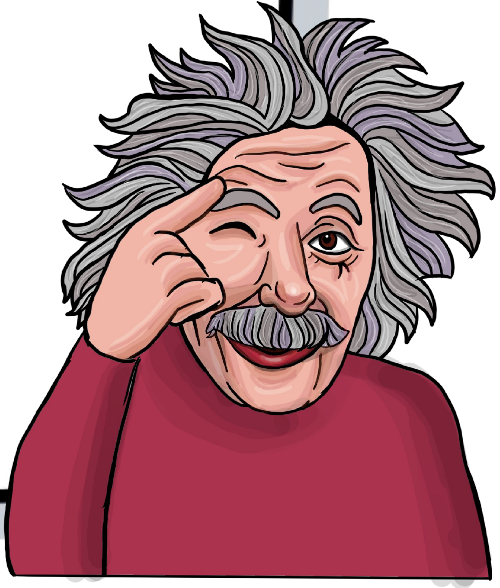
[Asudeh, Jagadish, Stoyanovich & Das, 2019]

Designing fair rankers



[Asudeh, Jagadish, Stoyanovich & Das, 2019]

Questions?



Roadmap

- We present a **classification framework**, unifying fair ranking methods in terms of group structure, type of bias, and mitigation objectives
- We map representative **score-based fair ranking** methods to this framework
- We map representative fair **learning-to-rank methods** to this framework
- We discuss existing **datasets & benchmarks** that have have been used in fair ranking research
- We **conclude** with concrete guidance for practitioners wishing to incorporate fairness objectives into algorithmic rankers

Method	Mitigation Point	Group structure	Bias	Worldview	EO Framework
iFair [26]	pre-proc.	multiple multinary attr.; independent	technical	WYSWYG	formal
DELTR [58]	in-proc.	one binary attr.	pre-existing	WAE	luck-egalitarian
Fair-PG-Rank [43]	in-proc	one binary attr.	technical	WYSIWYG	formal
Pairwise Ranking Fairness [4]	in-proc.	one binary attr.	?	WYSIWYG	formal-plus
FA*IR [57] & [60]	post-proc.	one multinary attr.; combination	pre-existing	continuous	formal / luck-egalitarian
Fair Ranking at LinkedIn [19]	post-proc.	one multinary attr.; combination	pre-existing; technical	continuous	none / luck-egalitarian (1 sensitive attr.)
CFAθ [59]	post-proc.	multiple binary attr.; combination	pre-existing	continuous	formal / substantive
Fairness of Exposure [42]	post-proc.	one binary attr.	pre-existing/ technical	WYSIWYG / WAE	formal / luck-egalitarian
Equity of Attention [6]	post-proc.	one multinary attr.; independent	technical / emergent	WYSIWYG	formal

Roadmap

Taxonomy of fair ranking methods

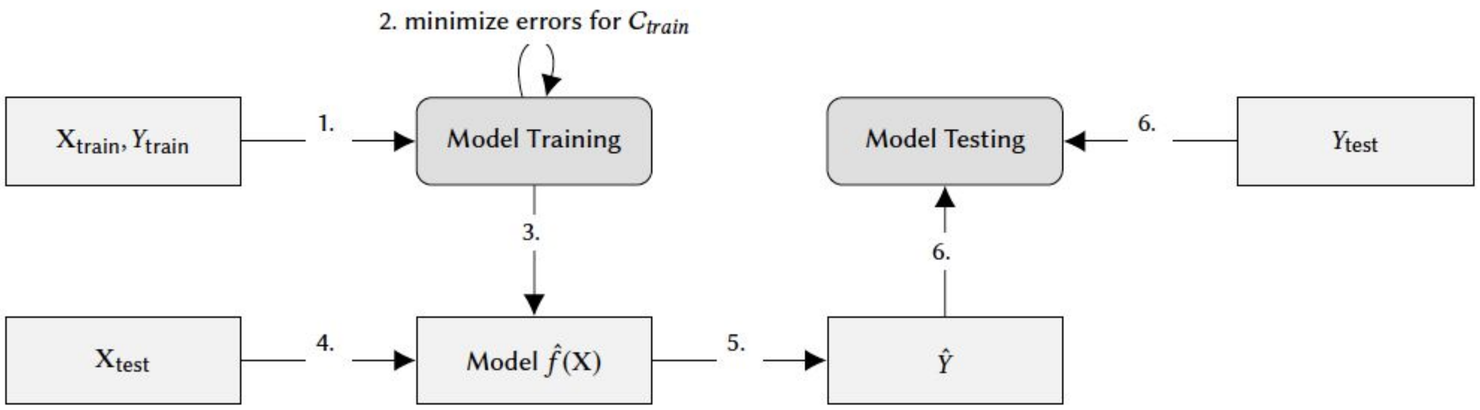
Map representative fair ranking methods: score-based ranker

Map representative fair ranking methods: learning to rank

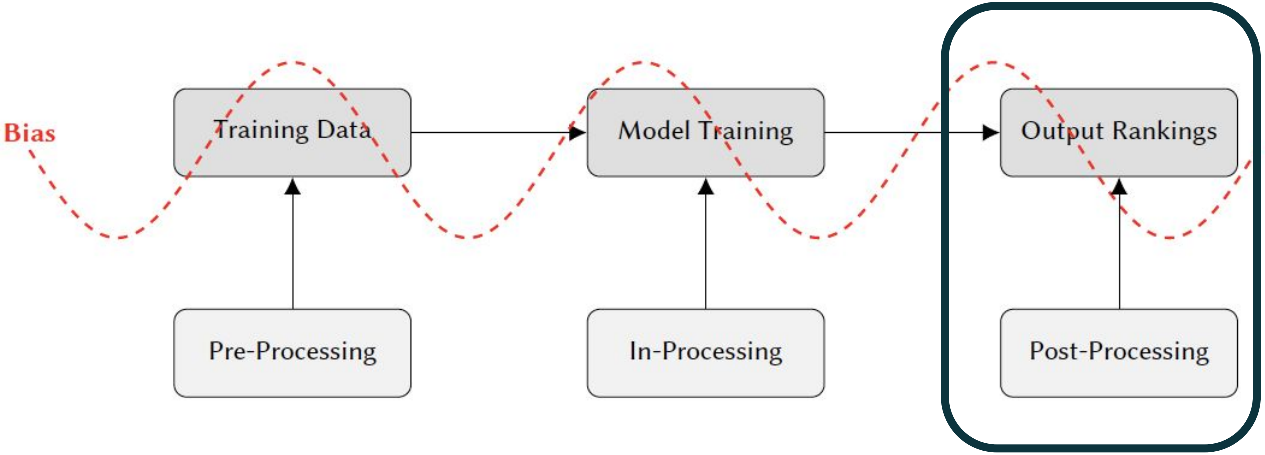
Datasets, benchmark, and framework

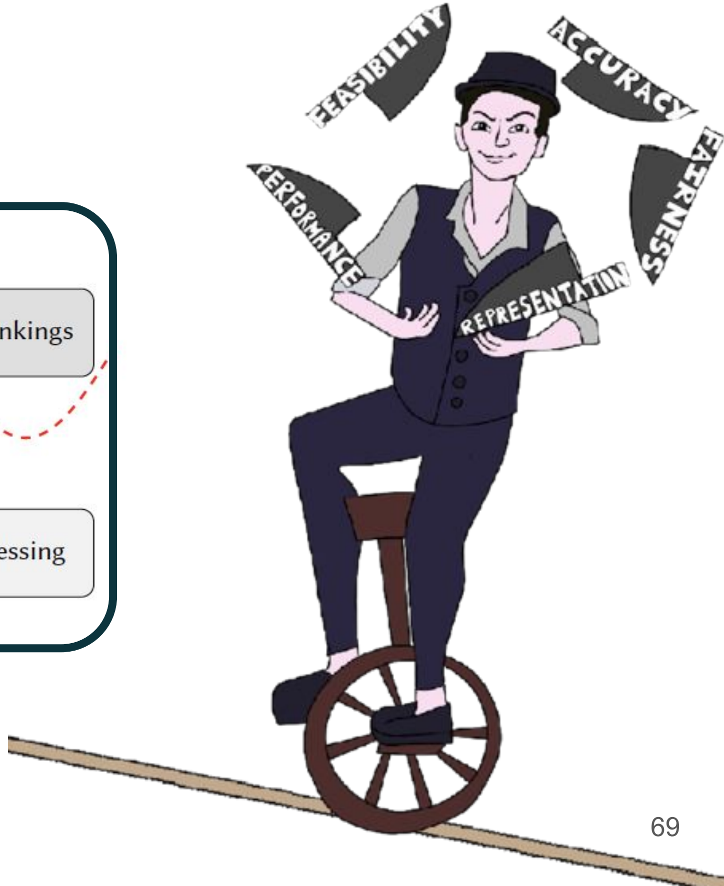
Concrete recommendations

Mitigation methods: learning-to-rank

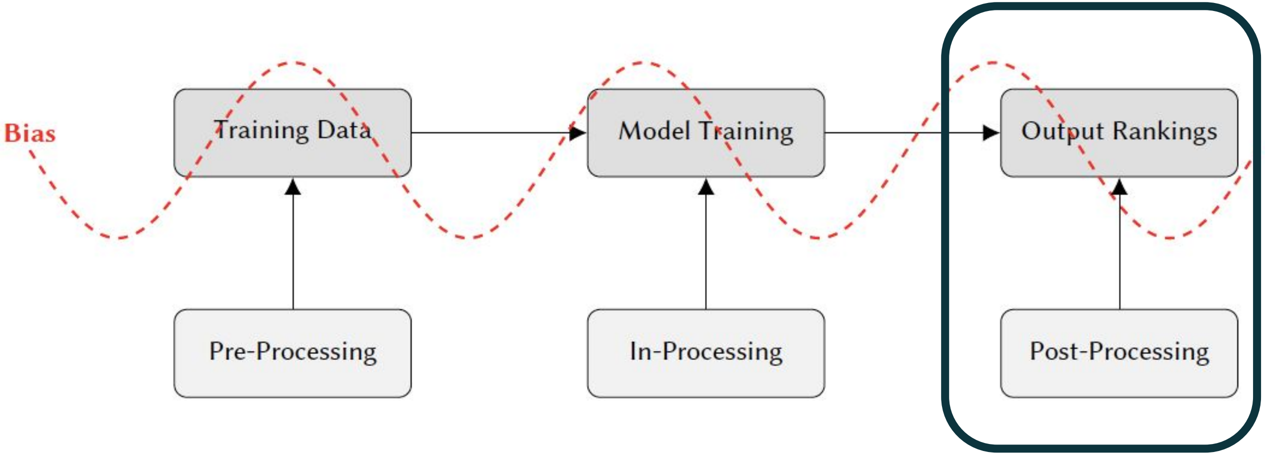


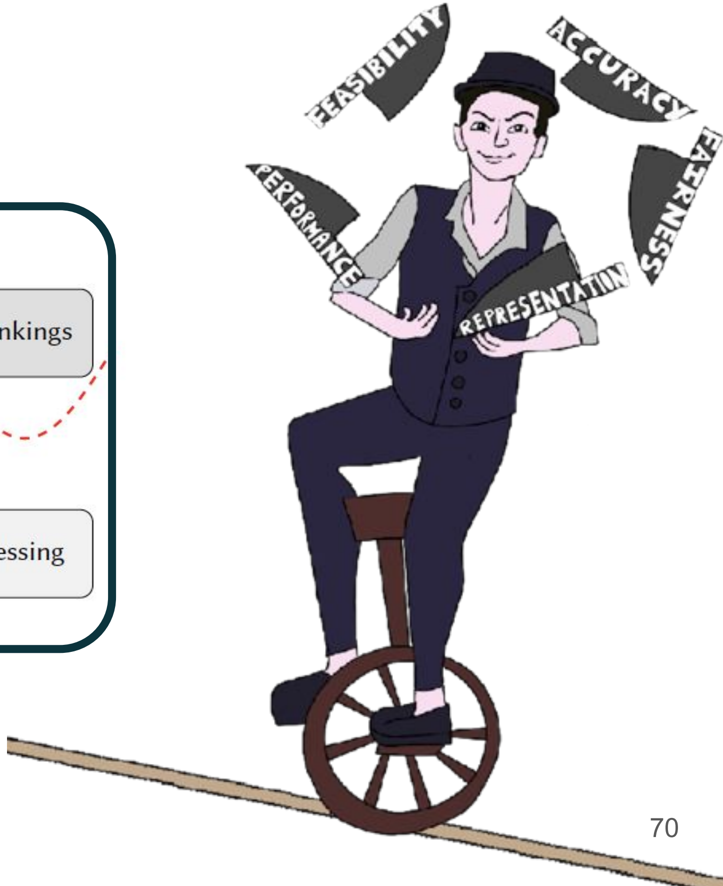
Bias mitigation methods



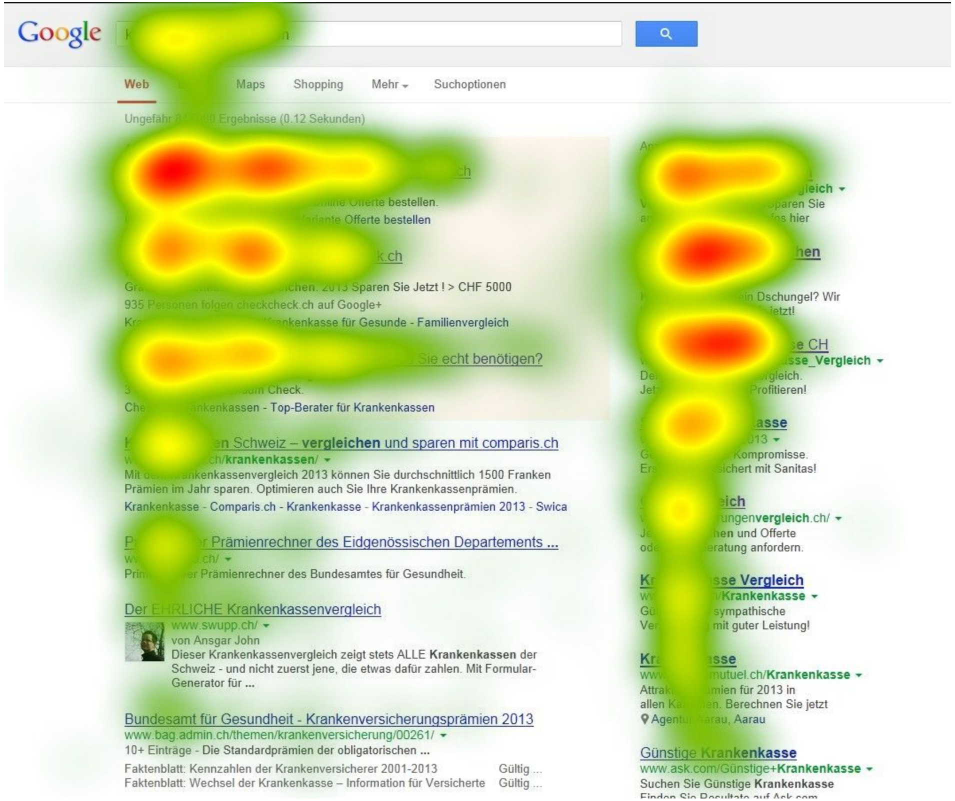


Bias mitigation methods





Exposure-based methods



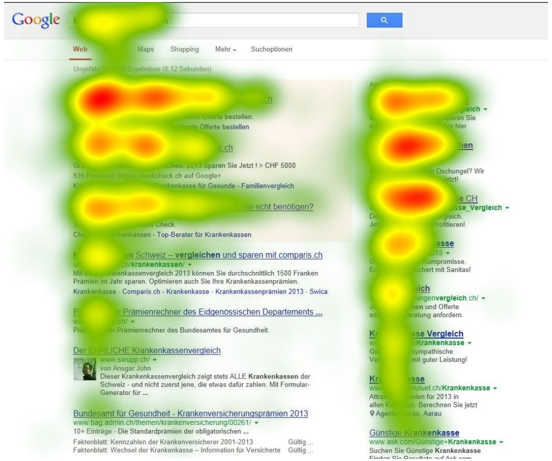
[Castillo, 2022]

Disparate exposure

Exposure: Each position *j* in a ranking has a certain probability v_j of being examined.

This is independent of an item *i*'s utility.

A group's exposure *E(G)* is commonly defined as the average *v* an item *i* ϵ *G receives*



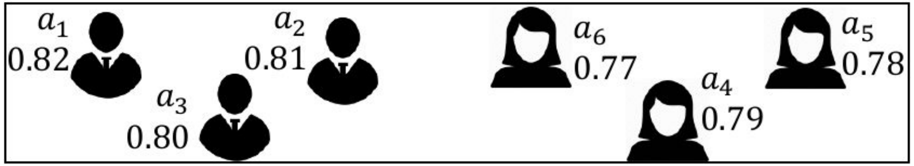
Fairness goal: equalize exposure

A ranking is fair, if

$$E(G_0) \approx E(G_1)$$

[Singh & Joachims, 2018]

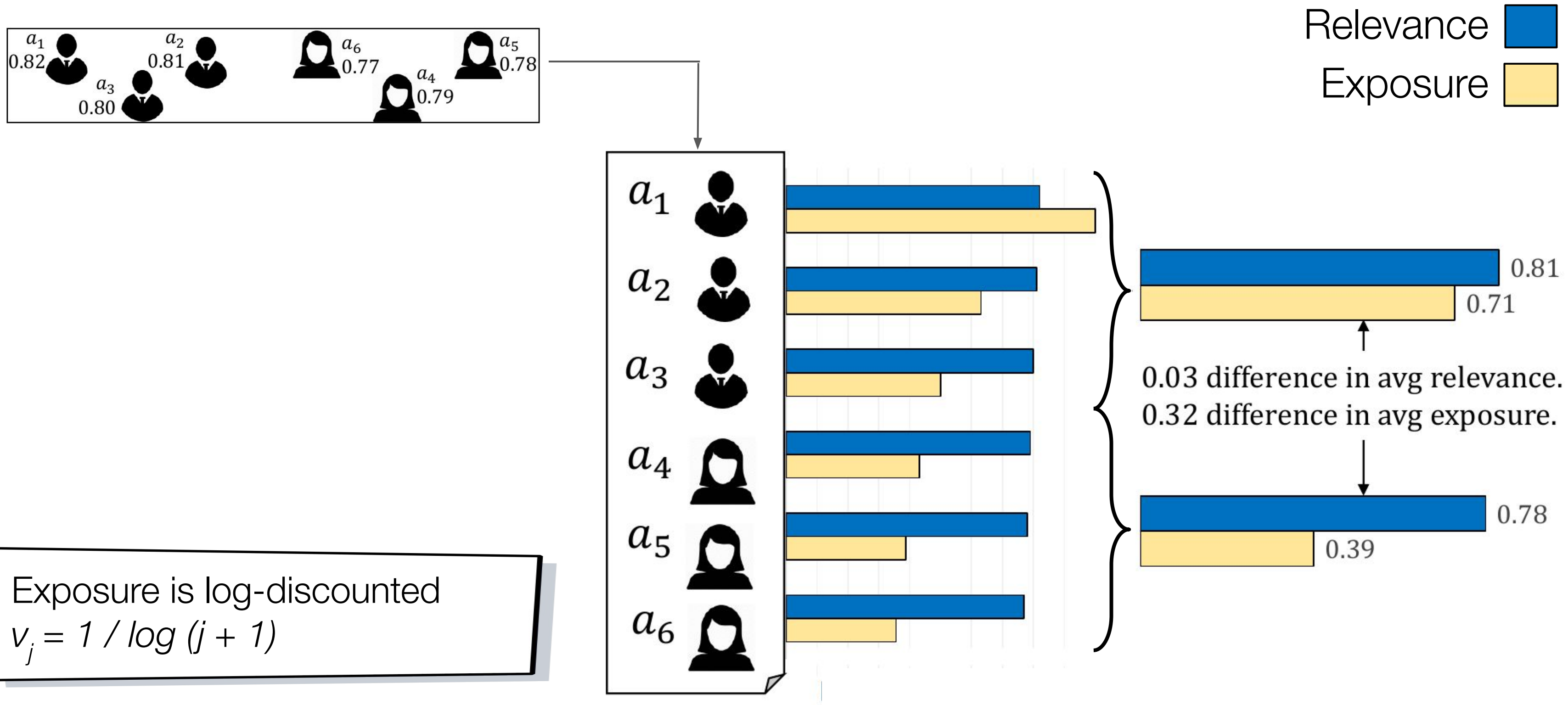
Disparate exposure: example



Candidates
(and their relevance scores)

[Singh & Joachims, 2018]

Disparate exposure: example



Exposure is log-discounted
$v_j = 1 / \log (j + 1)$

[Singh & Joachims, 2018]

Fairness of exposure

Probabilistic ranking $\boldsymbol{P}_{i,j}$: probability to place document *i* at position *j*

v_j is the position bias of position *j*

Group exposure $E(G_k | \boldsymbol{P})$

$$\text{Exposure}(G_k|\mathbf{P}) = \frac{1}{|G_k|} \sum_{d_i \in G_k} \sum_{j=1}^{N} \mathbf{P}_{i,j} \mathbf{v}_j$$

Fairness as demographic parity

A ranking is fair, if

$$E(G_0 | \boldsymbol{P}) \approx E(G_1 | \boldsymbol{P})$$

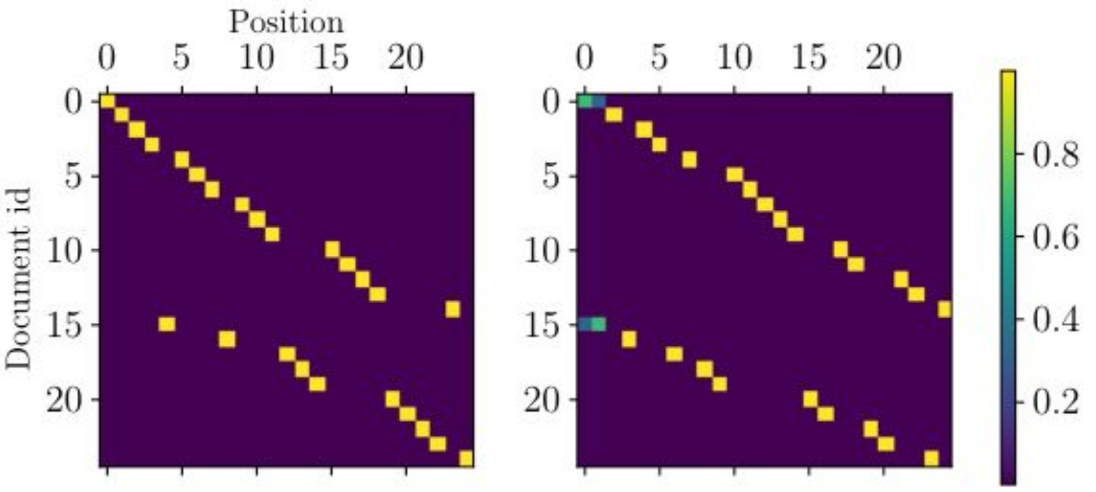
[Singh & Joachims, 2018]

Fairness of exposure

Experimental results, two groups

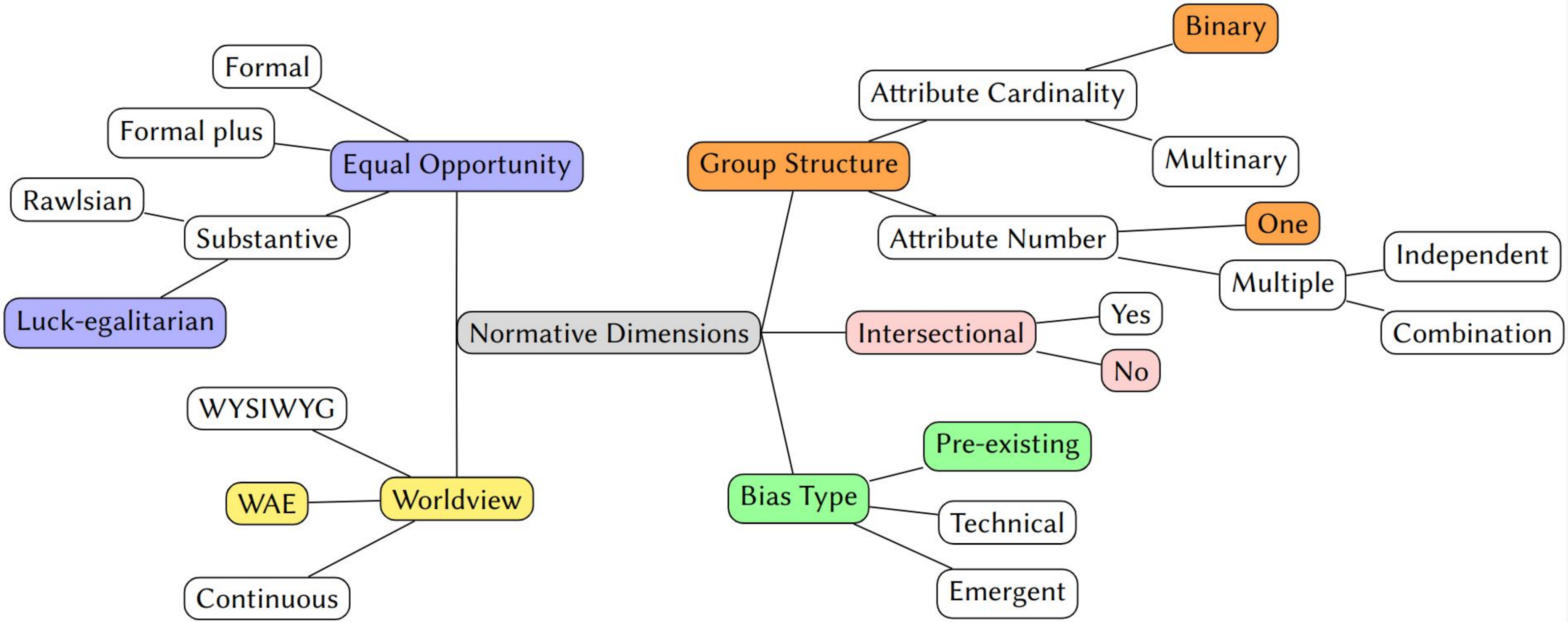
Doc id 0-14 is unprotected
Doc id 15-24 is protected

(a) Unconstrained
(b) Fair Ranking



(a) DCG=5.2027 (b) DCG=5.1360

[Singh & Joachims, 2018]


Normative Dimensions
Equal Opportunity
Formal
Formal plus
Substantive
Rawlsian
Luck-egalitarian
Worldview
WYSIWYG
WAE
Continuous
Group Structure
Attribute Cardinality
Binary
Multinary
Attribute Number
One
Multiple
Independent
Combination
Intersectional
Yes
No
Bias Type
Pre-existing
Technical
Emergent

Utility-normalized fairness

Ranking Utility

$$\mathrm{U}(\mathbf{P}|q) = \sum_{d_i \in \mathcal{D}} \sum_{j=1}^{N} \mathbf{P}_{i,j}\, u(d_i|q)\, \mathbf{v}_j$$

"Disparate treatment ratio"

$$\mathrm{DTR}(G_0, G_1|\mathbf{P}, q) = \frac{\mathrm{Exposure}(G_0|\mathbf{P})/\mathrm{U}(G_0|q)}{\mathrm{Exposure}(G_1|\mathbf{P})/\mathrm{U}(G_1|q)}$$

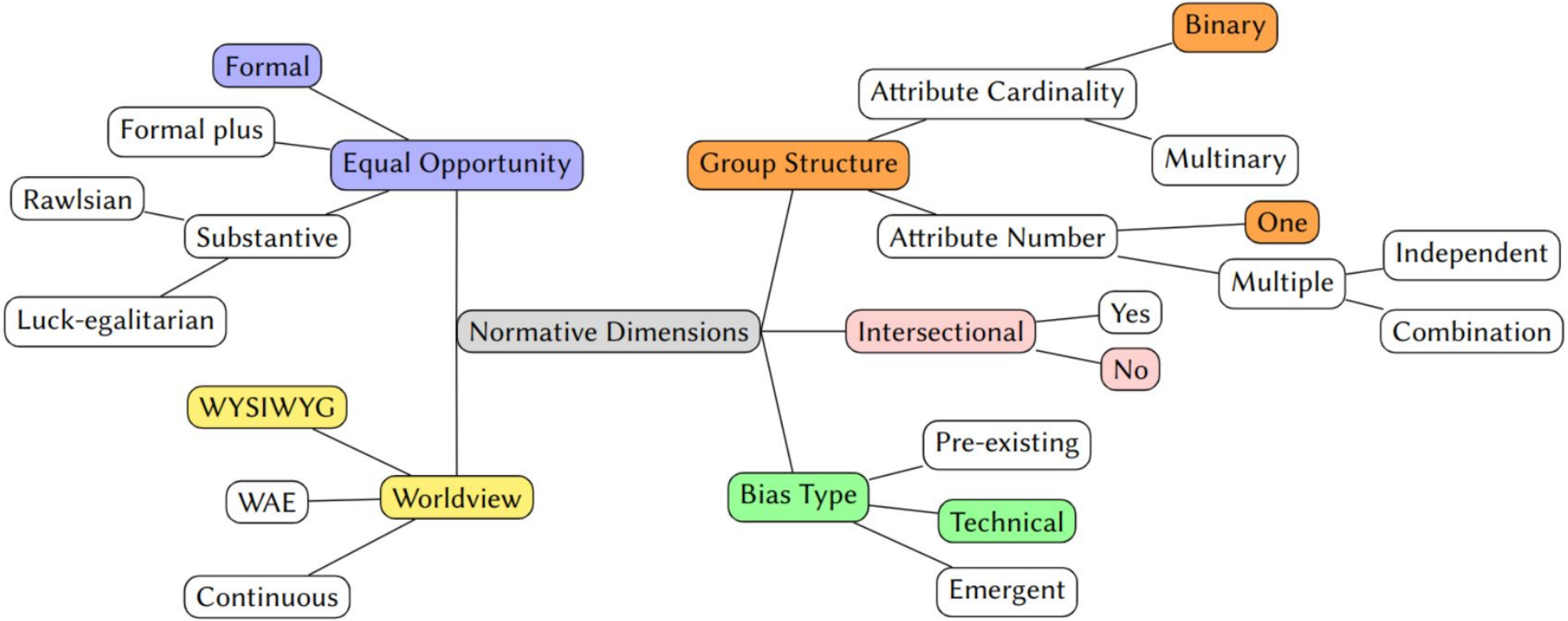
$$\mathrm{Exposure}(G_k|\mathbf{P}) = \frac{1}{|G_k|} \sum_{d_i \in G_k} \sum_{j=1}^{N} \mathbf{P}_{i,j}\mathbf{v}_j$$

"Disparate impact ratio"

$$\mathrm{DIR}(G_0, G_1|\mathbf{P}, q) = \frac{\mathrm{CTR}(G_0|\mathbf{P})/\mathrm{U}(G_0|q)}{\mathrm{CTR}(G_1|\mathbf{P})/\mathrm{U}(G_1|q)}$$

$$\mathrm{CTR}(G_k|\mathbf{P}) = \frac{1}{|G_k|} \sum_{i \in G_k} \sum_{j=1}^{N} \mathbf{P}_{i,j}\mathbf{u}_i\mathbf{v}_j$$

[Singh & Joachims, 2018]


Normative Dimensions
Equal Opportunity
Formal
Formal plus
Substantive
Rawlsian
Luck-egalitarian
Worldview
WYSIWYG
WAE
Continuous
Group Structure
Attribute Cardinality
Binary
Multinary
Attribute Number
One
Multiple
Independent
Combination
Intersectional
Yes
No
Bias Type
Pre-existing
Technical
Emergent

Amortized attention

Ranking elements *a* and *b* should enjoy equal attention discounted by their utility

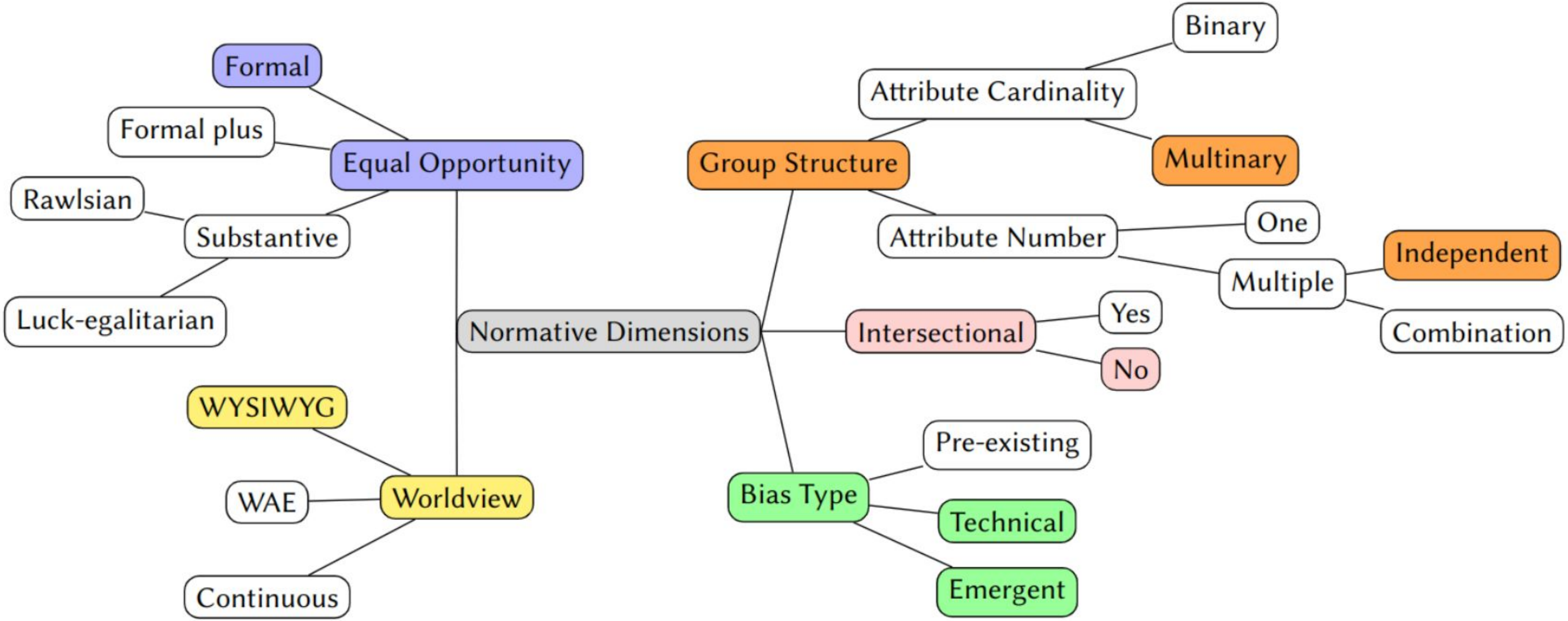
This equality shall be achieved over *m* rankings $\boldsymbol{\tau}$

$$\frac{\sum_{i=1}^{m} att(\boldsymbol{\tau}_i, a)}{\sum_{i=1}^{m} U(\boldsymbol{\tau}_i, a)} = \frac{\sum_{i=1}^{m} att(\boldsymbol{\tau}_i, b)}{\sum_{i=1}^{m} U(\boldsymbol{\tau}_i, b)}$$

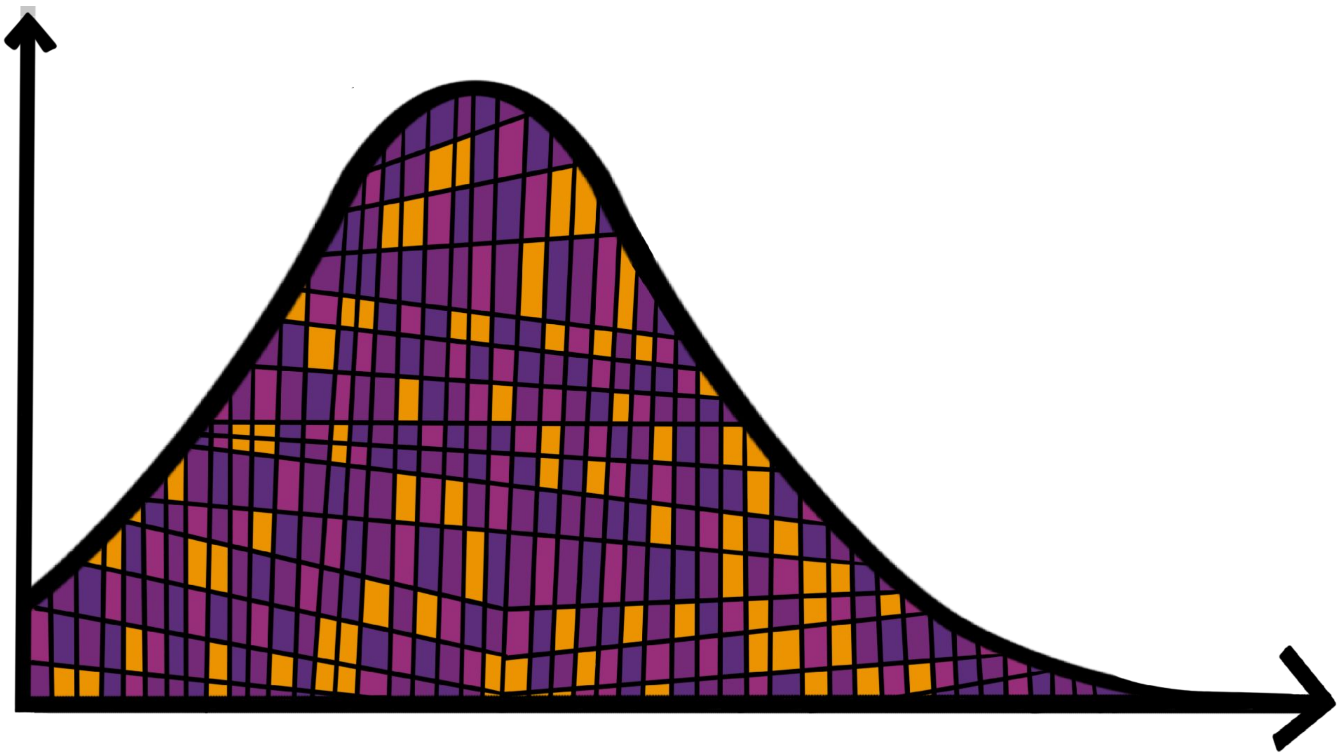
Unfairness is measured as the **accumulated difference in attention**

$$\text{unfairness}(\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_m) = \sum_{a=1}^{n} \left| \sum_{i=1}^{m} att(\boldsymbol{\tau}_i, a) - \sum_{i=1}^{m} U(\boldsymbol{\tau}_i, a) \right|$$

[Biega, Gummadi & Weikum, 2018]


Normative Dimensions
Equal Opportunity
Formal
Formal plus
Substantive
Rawlsian
Luck-egalitarian
Worldview
WYSIWYG
WAE
Continuous
Group Structure
Attribute Cardinality
Binary
Multinary
Attribute Number
One
Multiple
Independent
Combination
Intersectional
Yes
No
Bias Type
Pre-existing
Technical
Emergent

Probability-based methods



Probability-based vs. exposure-based methods

Probability-based methods measure the probability that a ranking was created according to some statistic process (e.g., tossing a coin)

Thus they fail immediately at the position where the condition does not hold anymore

Exposure-based methods are usually based on a cumulative measure

Thus they allow to make up unfair placement on the top at later positions in the ranking

FA*IR: fair representation condition

Given minimum proportion p, significance level α and a **set** of size k

Let $F(x;p,k)$ be the cumulative distribution function of a binomial distribution with parameters p, k

A ranking of k elements having x protected elements satisfies the **fair representation condition** with probability p and significance α if $F(x;p,k) > \alpha$

[Zehlike, Bonchi, Castillo, Hajian, Megahed & Baeza-Yates, 2017]

Example: fair representation condition

Suppose p=0.5, k=10, α=0.10

$F(1, 0.5, 10) = 0.01 < 0.10 \Rightarrow$ if 1 protected element, **fail**

$F(2, 0.5, 10) = 0.05 < 0.10 \Rightarrow$ if 2 protected elements, **fail**

$F(3; 0.5, 10) = 0.17 > 0.10 \Rightarrow$ if 3 protected elements, **pass**

$F(4; 0.5, 10) = 0.37 > 0.10 \Rightarrow$ if 4 protected elements, **pass**

[Zehlike, Bonchi, Castillo, Hajian, Megahed & Baeza-Yates, 2017]

FA*IR: ranked group fairness condition

Given parameters p, α and a **list** of size k

The list satisfies the **ranked group fairness** condition if

for every $i \leq k$

the prefix of size i of the list satisfies the **fair representation condition** for i, p, α

Problem: **multiple hypotheses testing**

Solution: adjust α

p \ k	1	2	3	4	5	6	7	8	9	10	11	12
0.1	0	0	0	0	0	0	0	0	0	0	0	0
0.2	0	0	0	0	0	0	0	0	0	0	1	1
0.3	0	0	0	0	0	0	1	1	1	1	1	2
0.4	0	0	0	0	1	1	1	1	2	2	2	3
0.5	0	0	0	1	1	1	2	2	3	3	3	4
0.6	0	0	1	1	2	2	3	3	4	4	5	5
0.7	0	1	1	2	2	3	3	4	5	5	6	6

[Zehlike, Bonchi, Castillo, Hajian, Megahed & Baeza-Yates, 2017]

Probability-based measure

Given a ranking of k elements …

… and a significance α:

its **ranked group fairness is the maximum *p*** such that the ranking passes ranked group fairness at p, α

… and a probability p:

its ranked group fairness is the minimum α such that the ranking passes ranked group fairness at p, α

[Zehlike, Bonchi, Castillo, Hajian, Megahed & Baeza-Yates, 2017]

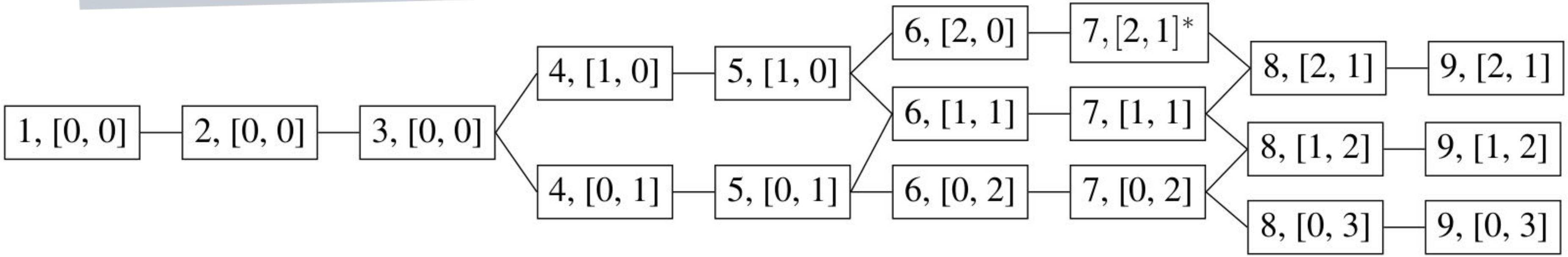
Multiple protected attributes

Extending previously seen definitions to the general case of n-1 protected groups: results in *mTree*

Any path through the tree is a valid configuration of a fair ranking according to the **ranked group fairness condition**

Shown here for $p_1 = 0.4$ and $p_2 = 0.2$ $(\alpha = 0.1)$

Read *-node as: by position 7 put at least 2 candidates from group 1 and 1 candidate from group 2



[Zehlike, Sühr, Baeza-Yates, Bonchi, Castillo & Hajian, 2022]

The FA*IR algorithm

Rank candidates of all protected groups p_i and non-protected separately

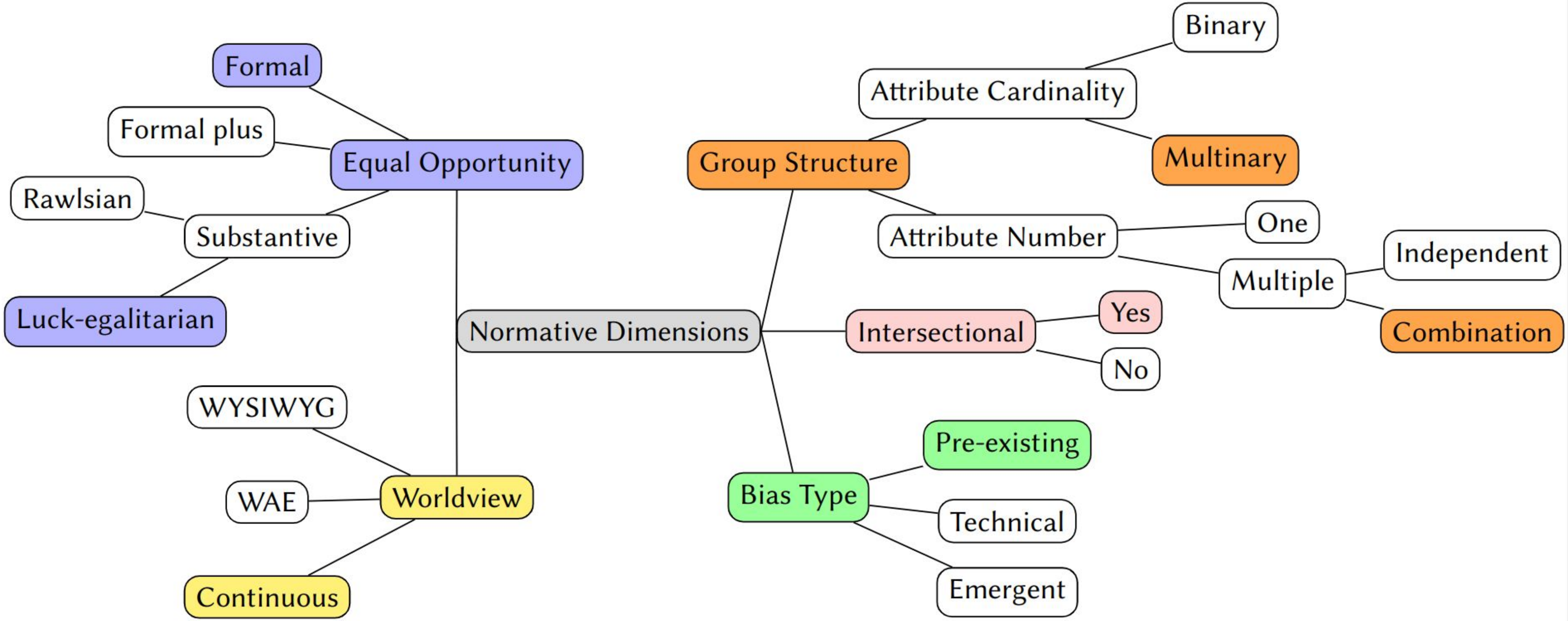
Determine the *minimum number* of protected elements required at every ranking position using *p_i, α* (that is, compute mTree)

For every position

If *enough* protected elements from all groups: pick next from best of all candidates

else: randomly choose next branch in mTree and put protected candidate from respective group

[Zehlike, Sühr, Baeza-Yates, Bonchi, Castillo & Hajian, 2022]


Normative Dimensions
Equal Opportunity
Formal
Formal plus
Substantive
Rawlsian
Luck-egalitarian
Worldview
WYSIWYG
WAE
Continuous
Group Structure
Attribute Cardinality
Binary
Multinary
Attribute Number
One
Multiple
Independent
Combination
Intersectional
Yes
No
Bias Type
Pre-existing
Technical
Emergent

The DetGreedy algorithm

Input: ranking of length k,

n groups of items, n-1 are protected,

$p_{2 \ldots n}$ proportions of protected groups

Fairness Definition: In a fair ranking, the number of protected items from each group shall neither fall below **nor exceed** the respective $p_{2 <= i <= n}$ at any point in the ranking

[Geyik, Ambler & Kenthapadi, 2022]

The DetGreedy algorithm

Rank candidates of all protected groups p_i and non-protected separately

For every position:

Check for all groups if they have not yet met their minimum, nor exceeded their maximum

If *enough* protected elements from all groups: pick next from best of all candidates

else: pick best candidate among all that have not reached their maximum yet

[Geyik, Ambler & Kenthapadi, 2022]

FA*IR vs. DetGreedy

Both are post-processing methods

Input and thus interface is almost the same

Re-ranking procedures also very similar

DetGreedy:

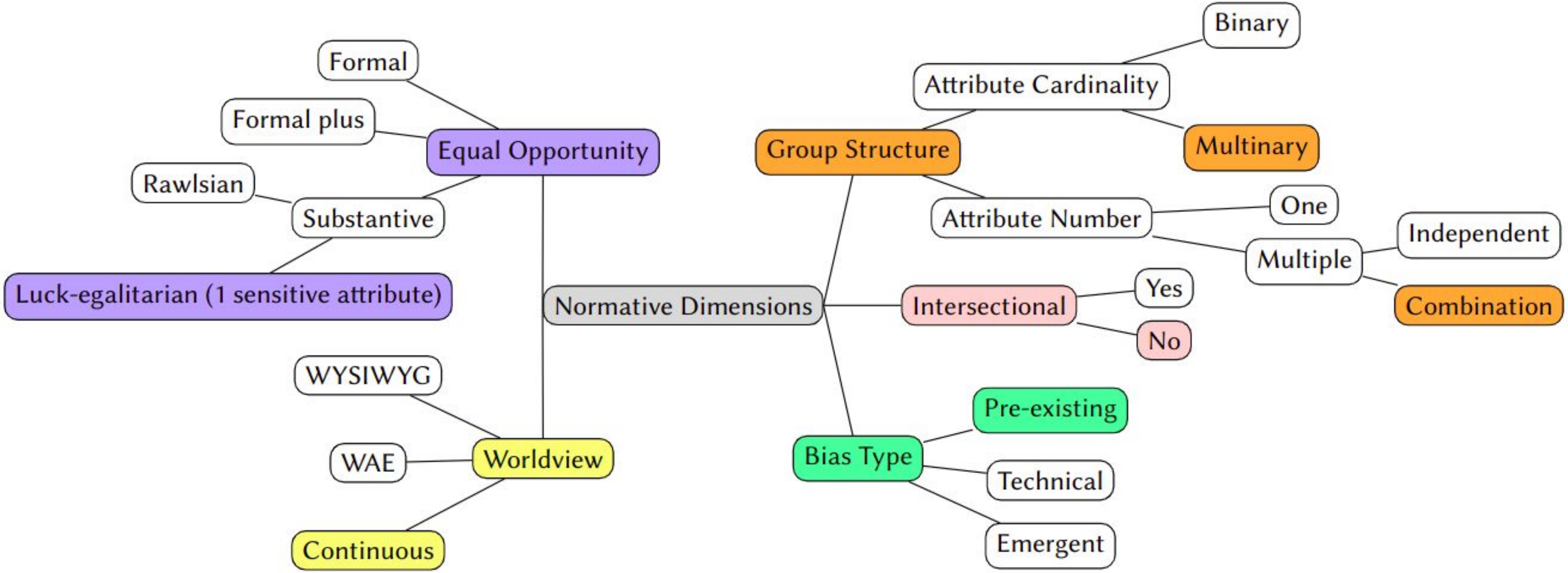
Can run into dead ends during re-ranking

Compares across protected candidates, thus **unsuitable for intersectionality**

FA*IR:

Only infeasible if not enough candidates

Does not ever compare candidates across groups, thus **suitable for intersectionality**


Normative Dimensions
Equal Opportunity
Formal
Formal plus
Substantive
Rawlsian
Luck-egalitarian (1 sensitive attribute)
Worldview
WYSIWYG
WAE
Continuous
Group Structure
Attribute Cardinality
Binary
Multinary
Attribute Number
One
Multiple
Independent
Combination
Intersectional
Yes
No
Bias Type
Pre-existing
Technical
Emergent

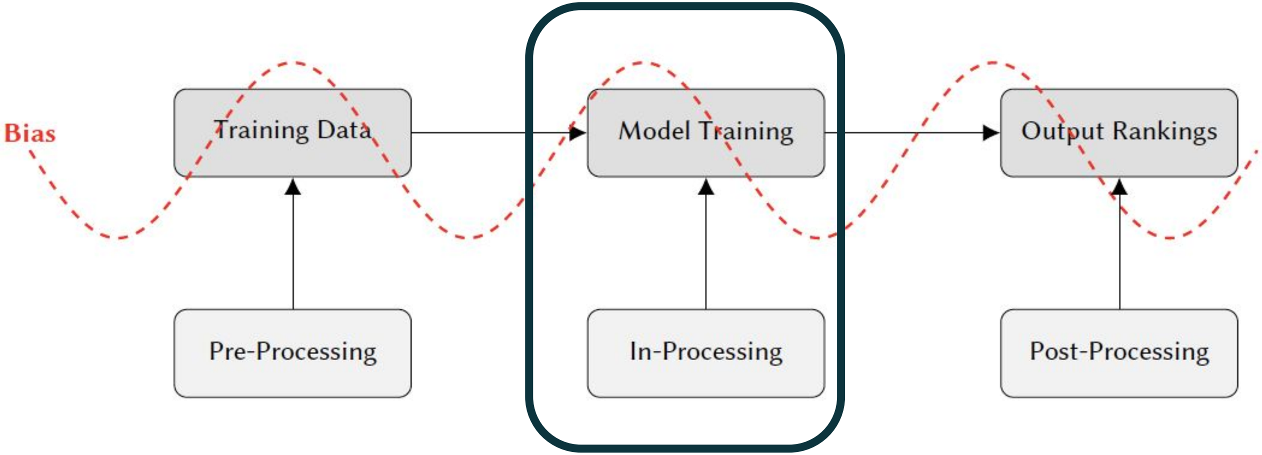
Why should I care?

Every technical choice is also always a normative choice

Small differences in technical choices can have tremendous normative implications

The values we encode in our technical choices should match our intended values for the task at hand

Bias mitigation methods



Listwise fairness (exposure-based)

Based on ListNet

Combination of two losses:

L = loss due to difference between ranking predictions and training elements

U = loss due to expected different exposure

DELTR

Exposure differences between two groups

U is not utility discounted

Fair-PG-Rank

Exposure differences between two candidates or two groups

U is utility discounted

[Singh & Joachims, 2019]
[Zehlike & Castillo, 2020]

Pairwise fairness

Idea based on fairness metrics that were proposed for classification (“equal opportunity”)

Pairwise accuracy should be the same across groups

$$P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b, A_a = 0\right) = P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b, A_a = 1\right)$$

Distinguishes between intra- and inter-group fairness

$$\boxed{\begin{aligned}&P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b, A_a = A_b = 0, z_a = \tilde{z}\right) = \\ &P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b, A_a = A_b = 1, z_a = \tilde{z}\right) \forall \tilde{z}\end{aligned}}$$

$$\boxed{\begin{aligned}&P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b, A_a = 0, A_b = 1, z_a = \tilde{z}\right) = \\ &P\left(\hat{f}(\mathbf{X}_a) > \hat{f}(\mathbf{X}_b) \mid Y_a > Y_b, A_a = 1, A_b = 0, z_a = \tilde{z}\right) \forall \tilde{z}\end{aligned}}$$

[Beutel, Chen, Doshi, Qian, Wei, Wu, Heidt, Zhao, Hong, Chi & Goodrow, 2019]

Questions?

Roadmap

- We present a **classification framework**, unifying fair ranking methods in terms of group structure, type of bias, and mitigation objectives
- We map representative **score-based fair ranking** methods to this framework
- We map representative fair **learning-to-rank methods** to this framework
- We discuss existing **datasets & benchmarks** that have have been used in fair ranking research
- We **conclude** with concrete guidance for practitioners wishing to incorporate fairness objectives into algorithmic rankers

Datasets

Name	Size	Sensitive attributes	Scoring attributes
AirBnB	10,201 houses	gender of host	rating, price
COMPAS	7,214 people	gender, race	risk scores
CS departments	51 departments	size, location	# publications in CS areas
DOT	1.3 million flights	airline name	departure delay, arrival delay, taxi-in time
Engineering students	5 queries, 650 students per query	gender, high school type	academic performance after first year
Forbes richest U.S.	400 people	gender	net worth

Datasets

Name	Size	Sensitive attributes	Scoring attributes
German credit	1,000 people	gender, age	credit amount, duration
IIT-JEE	384,977 students	birth category, gender, disability status	test scores
LSAC	21,792 students	gender, race	LSAT scores
MEPS	15,675 people	gender, race, age	# visits requiring medical care
NASA astronauts	357 astronauts	major in college	flight hours
Pantheon	11,341 people	occupation	popularity of Wiki page
SAT	1.6M students	gender	SAT score

Datasets

Name	Size	Sensitive attributes	Score
StackExchange	253,000 queries, 6M documents	domains	document relevance
SSORC	8,975,360 papers	gender of authors	number of citations
W3C experts	60 queries, 200 experts per query	gender	probability of being an expert
XING	40 candidates	gender	years of experience, education
Yahoo LTR	26,927 queries, 638,794 docs	N/A	relevance
Yow news	unknown	source of news	relevance

Fair ranking benchmark at TREC

Started in 2019

2022 track "focuses on fairly prioritising Wikimedia articles for editing to provide fair exposure to articles from different groups"

Resource allocation task with **exposure-based fairness** metrics

Explicitly mentions **intersectional** fairness

TREC 2022 Fair Ranking Track

The TREC Fair Ranking track evaluates systems according to how well they *fairly* rank documents.

The 2022 track focuses on fairly prioritising Wikimedia articles for editing to provide a fair exposure to articles from different groups.

TIMELINE

- **May, 2022**: guidelines released.
- **June, 2022**: training queries and corpus released
- **July, 2022**: evaluation queries released
- **31st August, 2022**: submissions due
- **September, 2022**: evaluated submissions returned

DOWNLOADS

The TREC 2022 Fair Ranking Track participation guidelines, experimentation protocol, data and evaluation scripts will be made available here.

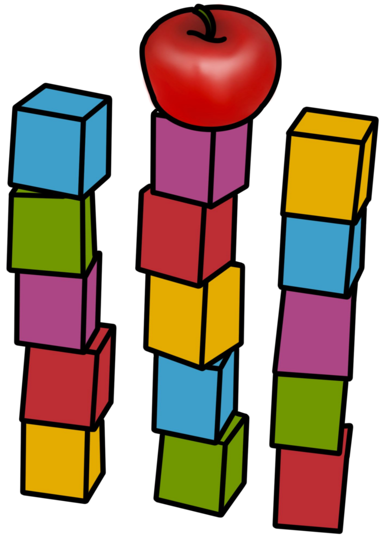
- Participant Instructions
- Corpus
- 2022 Topics and Metadata
- 2022 Eval Topics

Fair ranking benchmark at TREC: data

Many different **fairness attributes** to select from:

- Geographic location (topic and source)
- Gender and occupation (biographies)
- Age of topic and article
- Article popularity
- Article languages
- Alphabetical order of topics

Limitation: English-language only



Fair ranking benchmark at TREC: tasks

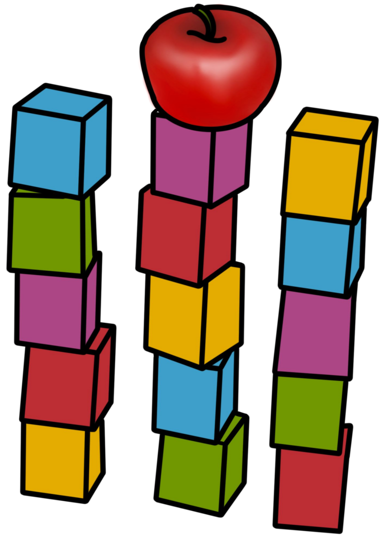
Task 1

WikiProject coordinators who search for articles needing work and produce a ranked list per topic

Outputs a **single ranking per query**

Relevance as nDCG for topic

Attention-weighted rank fairness: compares cumulative group exposure with target distribution (not relevance discounted)



Fair ranking benchmark at TREC: tasks

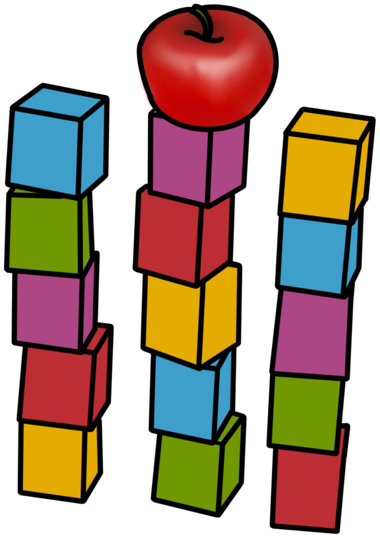
Task 2

Wikipedia editors looking for work associated with a project

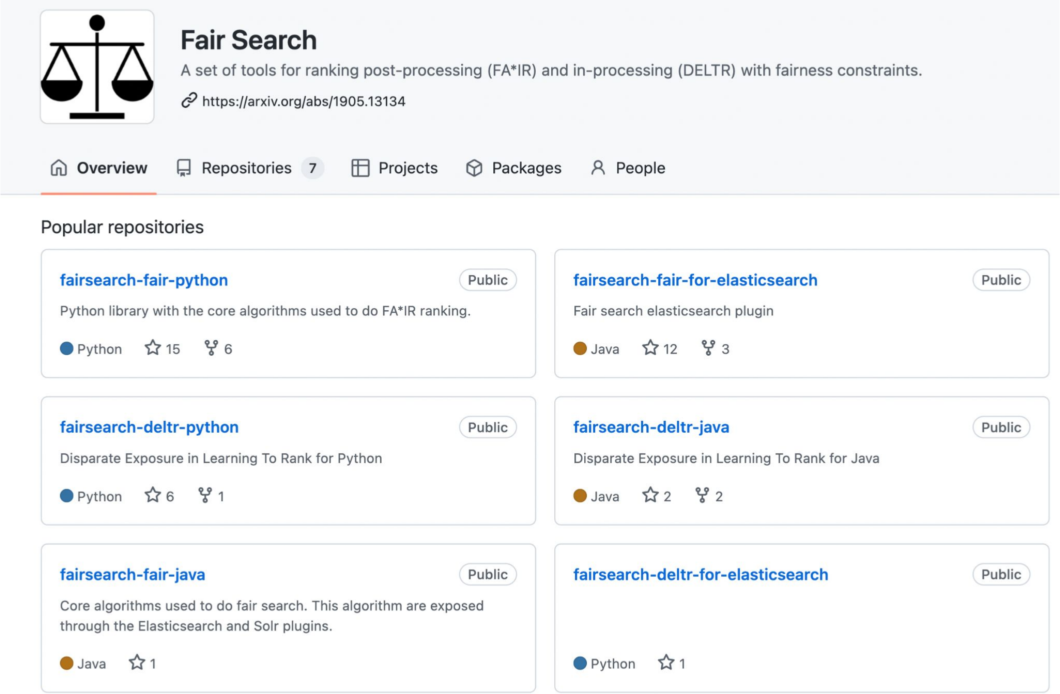
Outputs **100 rankings per query** (20 articles)

Relevance as nDCG for topic and work needed

Fairness as expected exposure over multiple rankings (relevance discounted)



Fair Search, an open source API



[Zehlike, Sühr, Castillo, & Kitanovski 2019]

Roadmap

- We present a **classification framework**, unifying fair ranking methods in terms of group structure, type of bias, and mitigation objectives
- We map representative **score-based fair ranking** methods to this framework
- We map representative fair **learning-to-rank methods** to this framework
- We discuss existing **datasets & benchmarks** that have have been used in fair ranking research
- We **conclude** with concrete guidance for practitioners wishing to incorporate fairness objectives into algorithmic rankers

Key questions

How do we select or design fairness & diversity metrics?

- What values and beliefs do we want to encode?
- What is the legal and practical context of use?

How do we show that our method works?

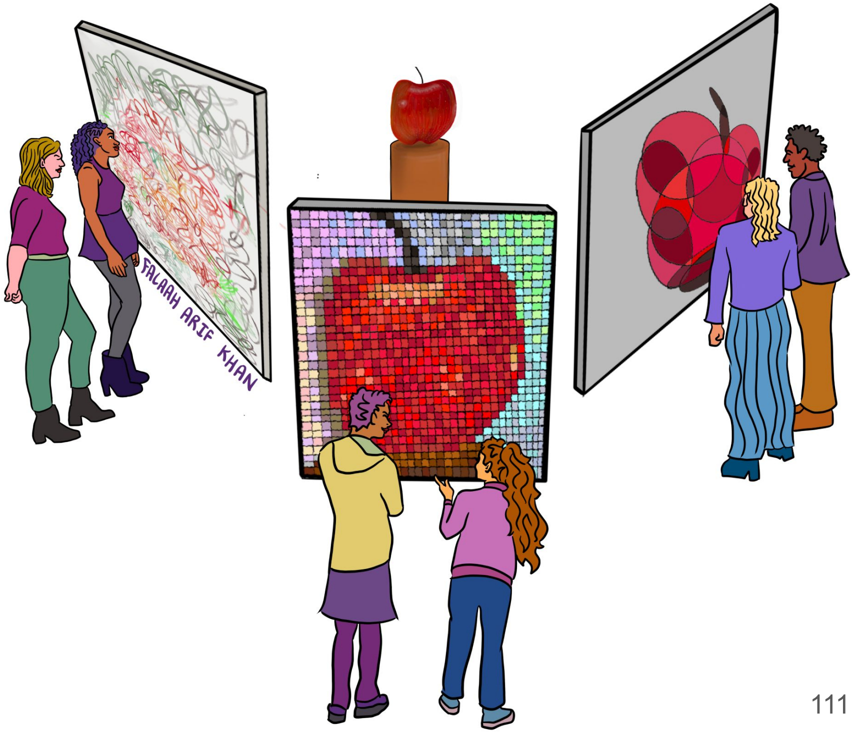
- With which methods should we compare?
- What dataset should we experiment on?

How do we publish our results?

- By being upfront about the limitations, and about the pote for misuse

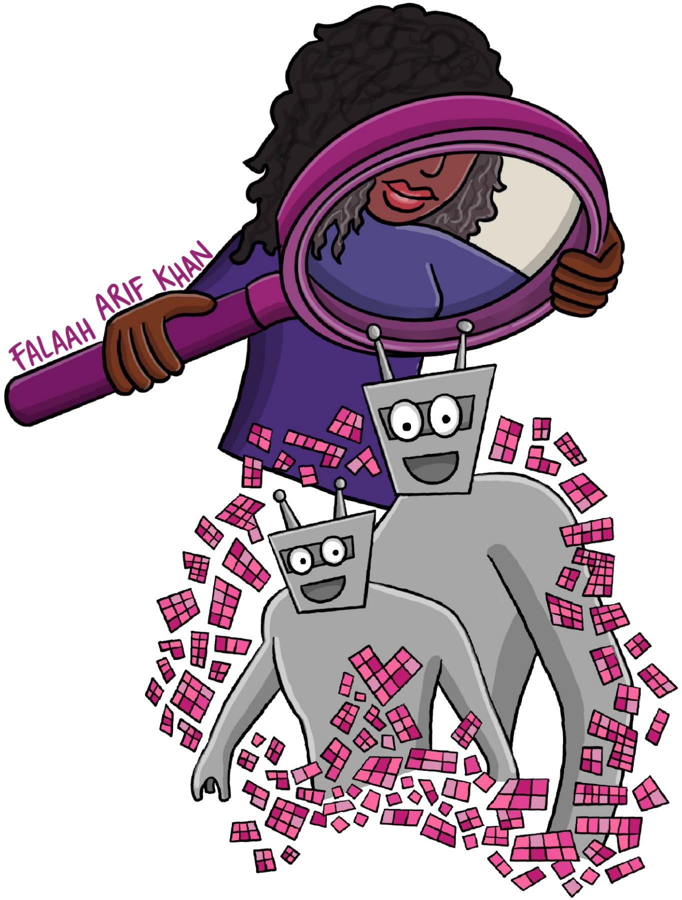
Recommendation 1

Make **context of use** explicit



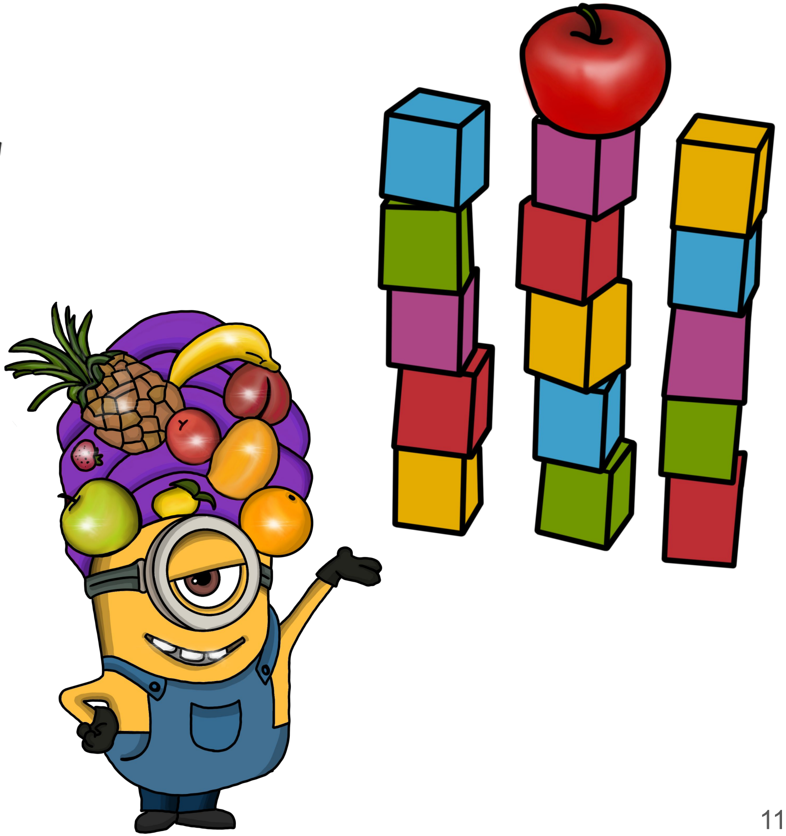
Recommendation 2

Surface **normative** consequences of **technical** choices

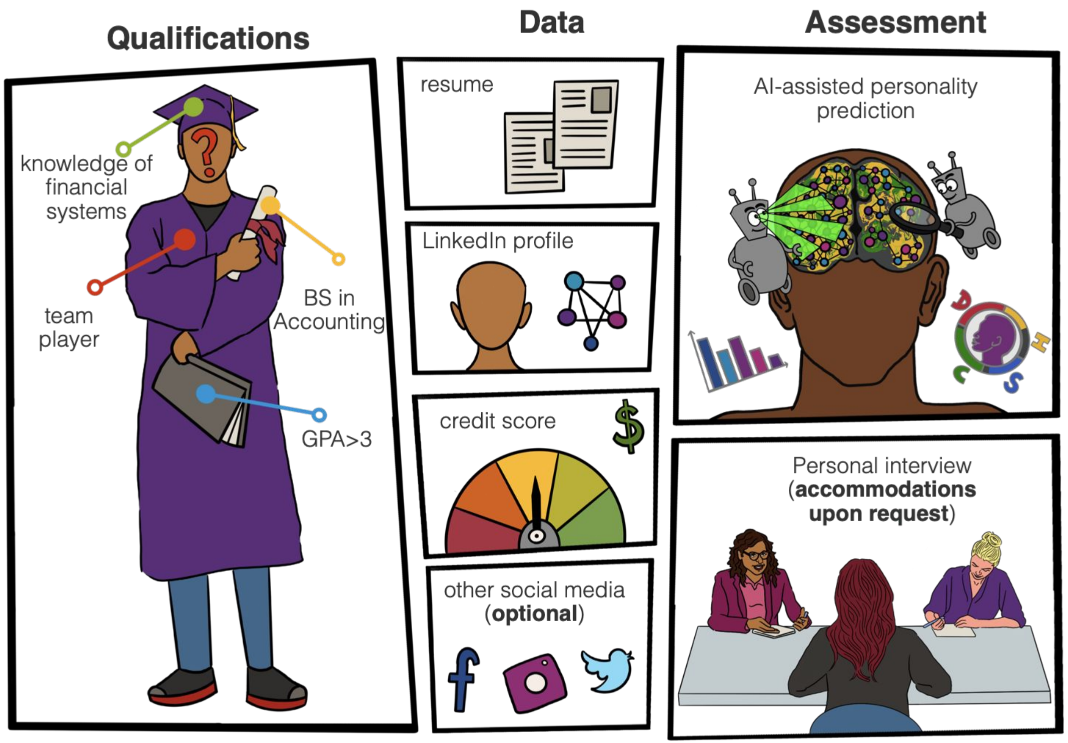


Recommendation 3

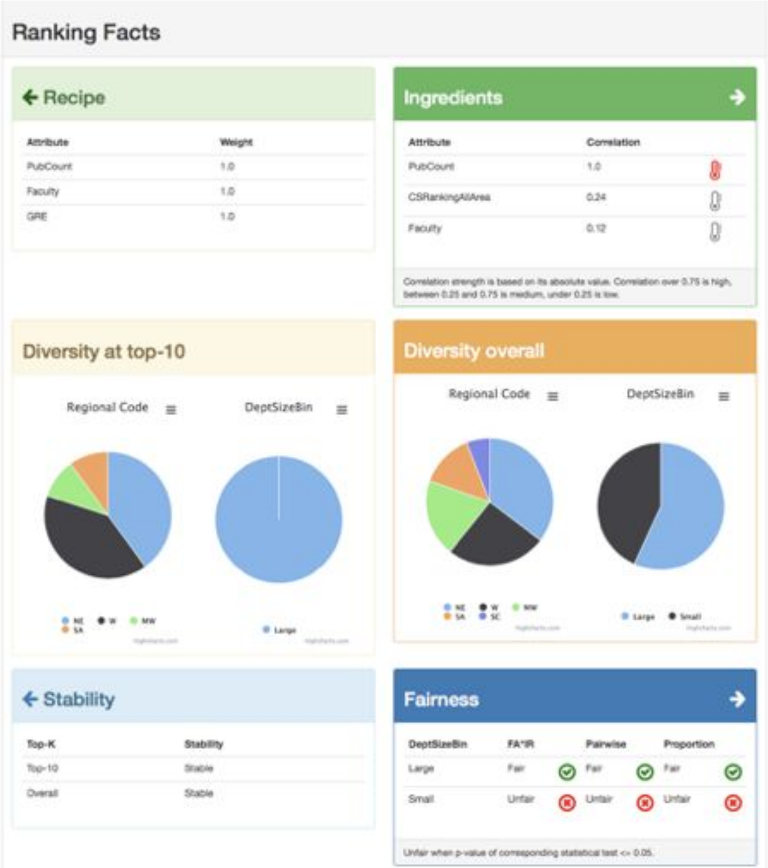
Draw **meaningful comparisons**



Beyond fairness: transparency & interpretability



Ranking Facts



[Yang, Stoyanovich, Asudeh, Howe, Jagadish & Miklau 2018]

Ranking Facts, a “nutritional label” for rankings

comprehensible: short, simple, clear

consultative: provide actionable info

comparable: implying a standard

computable: incrementally constructed

[Stoyanovich & Howe 2019]

Beyond fairness: stability

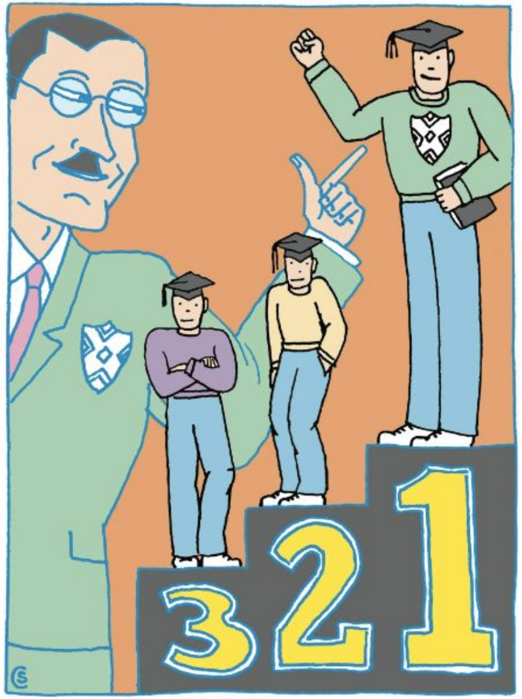
THE NEW YORKER

DEPT. OF EDUCATION FEBRUARY 14 & 21, 2011 ISSUE

THE ORDER OF THINGS

What college rankings really tell us.

By Malcolm Gladwell



Rankings depend on what weight we give to what variables. Illustration by SEYMOUR CHWAST

Designing stable rankers

Goals

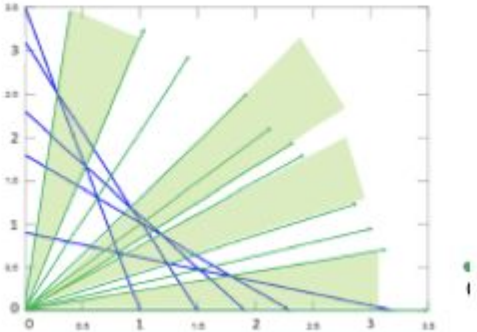
utility: with similar weights as what the human decision-maker had in mind

stability: so that the ranking doesn't reshuffle when weights change slightly

$\mathcal{D}$			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

Belief

stable rankings are more **trustworthy**



[Asudeh, Jagadish, Miklau & Stoyanovich 2018]

Beyond fairness: privacy

Thank you!

Questions?

Julia Stoyanovich Meike Zehlike Ke Yang

ACM SIGMOD 2023