ULB

INFO-H420
Management of Data Science and Business Workflows

Part II
Fairness

Dimitris SACHARIDIS

2023-2024

How do people feel about interacting with AI systems?

survey of 5,000 consumers by Pegasystems about people interacting with Artificial Intelligence (AI) systems

- 34% say they interact with AI systems
- machines **not trustworthy**
  - only 9% very comfortable interacting
- machines **can be biased**
  - 53% say it’s possible for systems to show bias in their decisions
- machines **cannot be moral**
  - 56% don’t believe it is possible to develop systems that behaves morally

Responsible Data Science

- Responsible data science is the practice of using data and data-driven techniques in a way that is **ethical**, **transparent**, and **respectful of the rights** and interests of **individuals** and **society**

- Key concepts:
  - **Data privacy** and security, to protect sensitive data of individuals
  - **Transparency** about data sources, methods, and limitations
  - Exhibit desirable **ethical** principles
  - Put **human in control**, allow them to understand and control

ChatGPT, December 2022

Responsible AI/ML/Data Science

- Algorithmic Decisions (e.g., from ML/AI models) are ubiquitous, permeate many aspects of **society**
- more and more questions of **responsible use** arise

Fairness Accountability Transparency Ethics

- dedicated conferences
  - ACM Conference on Fairness, Accountability, and Transparency (FAccT)
  - AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)
- topics appear in almost all CS venues
- courses: in Berkeley, Cornell, Princeton
- EU high-level expert group on AI: Ethics Guidelines for Trustworthy AI
- EU upcoming regulation: AI Act

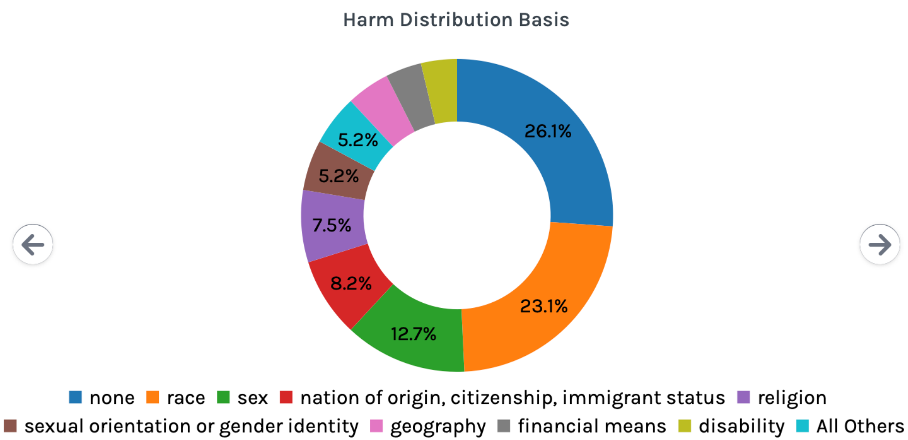
Irresponsible AI

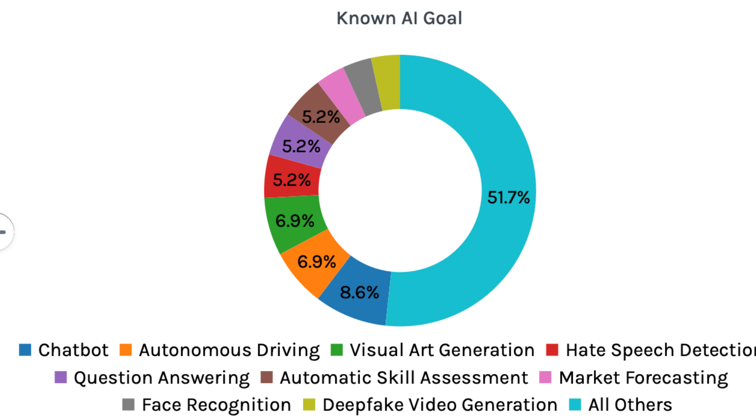
Welcome to the

AI Incident Database

Search over 2000 reports of AI harms

Search | Discover





https://incidentdatabase.ai

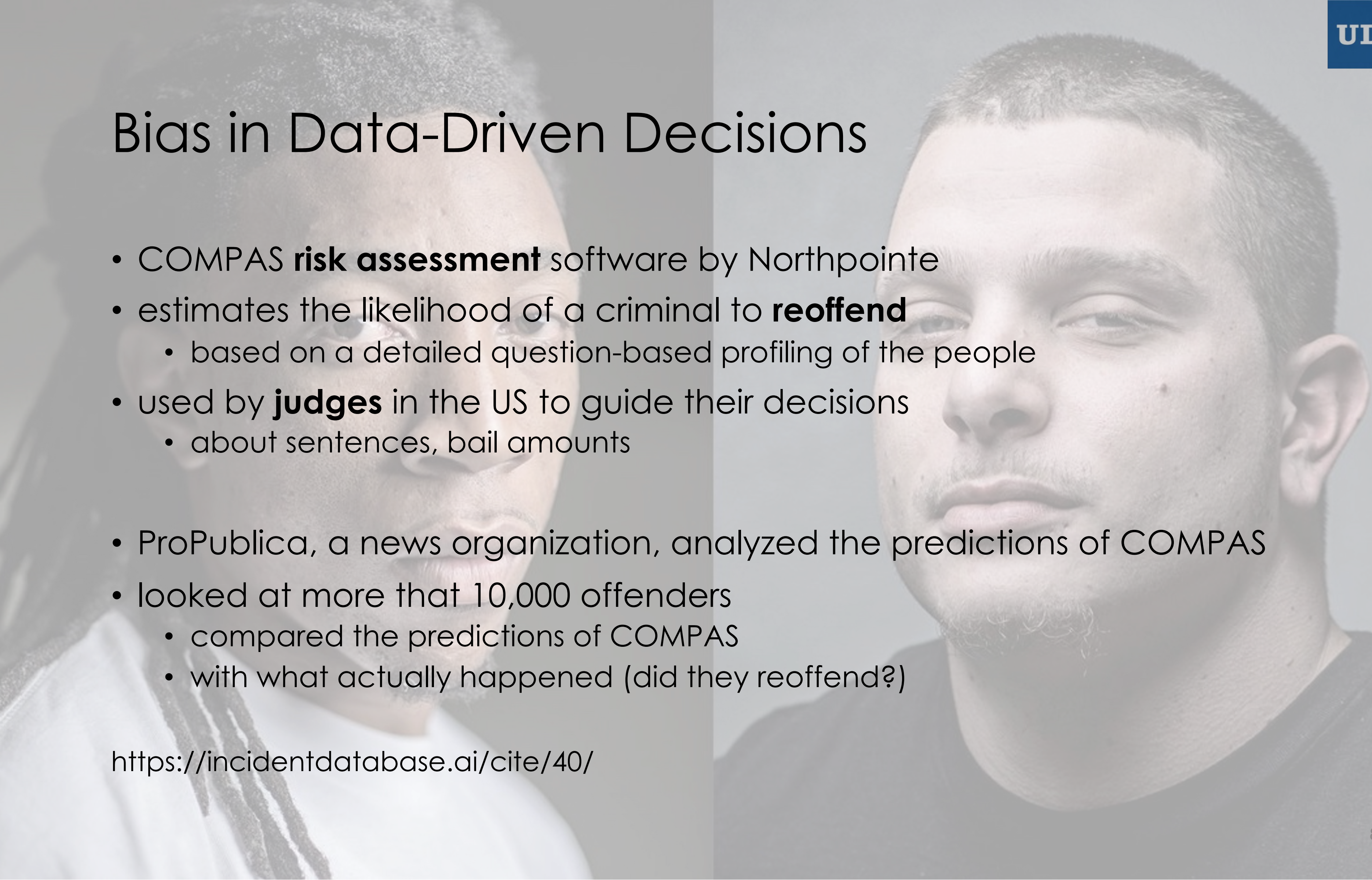
Ethical Principles

- review of 84 AI ethics guidelines
- trust is the "end goal"

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice & fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom & autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

[2019 Nature Machine Intelligence A. Jobin et al.] *Artificial Intelligence: the global landscape of ethics guidelines*

Fairness



Bias in Data-Driven Decisions

- COMPAS **risk assessment** software by Northpointe
- estimates the likelihood of a criminal to **reoffend**
  - based on a detailed question-based profiling of the people
- used by **judges** in the US to guide their decisions
  - about sentences, bail amounts

- ProPublica, a news organization, analyzed the predictions of COMPAS
- looked at more that 10,000 offenders
  - compared the predictions of COMPAS
  - with what actually happened (did they reoffend?)

https://incidentdatabase.ai/cite/40/

Bias in Data-Driven Decisions

results of the study:

- correctly predicts reoffending for black and white at the same rate
    - same **recall** (true positive rate, sensitivity)

but **mistakes** paint a different picture:

- blacks are almost **twice as likely** as whites to be labeled a **higher risk** but not actually reoffend
    - more **false positives** for blacks
- black are **much less likely** than whites to be labeled **lower risk**, but actually reoffend
    - less **false negatives** for blacks

Algorithmic Decisions

- main result of the study:

discrimination against blacks

- but there is some controversy to the study results
  - https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
  - https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html
  - https://www.propublica.org/article/technical-response-to-northpointe
  - http://www.crj.org/assets/2017/07/9_Machine_bias_rejoinder.pdf

Algorithmic Decisions

What are the issues here?

- **fairness** – is the system actually fair to racial groups?
- **transparency** – we don't know what the algorithm does
- **accountability** – who is to hold responsible?

ULB

What is Fairness?

- Dictionary: "***the state of being free from bias or injustice***"

- Political Science: "*distributive justice discusses **fair allocation of resources** among diverse members of a community*"
  - "A Theory of Justice" by J. Rawls (American philosopher)
    - "justice as fairness"; "social cooperation should be fair to all citizens regarded as free and as equals"
  - but what is a *fair allocation*?
    - ***equality of outcome***: each person get the same amount
    - ***equality of opportunity***: equal grounds for competing for resources
    - ***social welfare***: what benefits the society the most

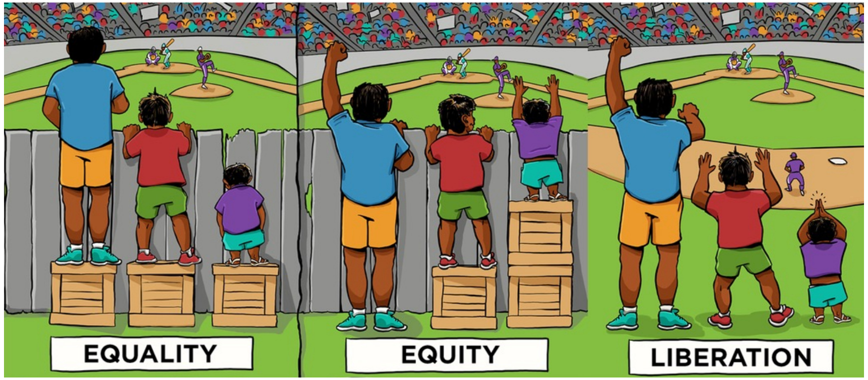
What is Fairness?

- Legal Systems: ***"fairness as non-discrimination"***
  - ***disparate treatment***: intentional discrimination on *protected groups* (defined on race, color, sex, etc.); not "*color-blind*"
    - e.g., only African American applicants are required to take a pre-employment assessment test
  - ***disparate impact***: a procedure that has disproportionate impact on protected groups
    - e.g., all applicants are tested but only African Americans are eliminated based on the results of the assessment.
  - ***affirmative action***: promote non-discrimination and support historically disadvantaged groups; *quota systems*
    - e.g., to address gender imbalance in STEM

Algorithmic Fairness

- Data-Driven Systems often make **decisions** on behalf of humans
  - high-stake decisions: likelihood to reoffend; grant a loan application
  - innocuous decisions: which article to read; what to buy in a store
    - or not so innocuous? fake news, filter bubbles

Algorithmic Fairness: machine made decisions should *not discriminate* against individuals



Group vs Individual Fairness

Group Fairness

- a specific **protected** group defined by a **sensitive** attribute (e.g., race, gender) should be treated/impacted similarly to the other (non-protected) groups
    - typically two groups protected (disadvantaged) and non-protected (advantaged)

Individual Fairness

- a specific **individual** should be treated/impacted similarly **to a similar other** individual

Classification based on Tasks

Fairness depends on the type of task and its outcomes

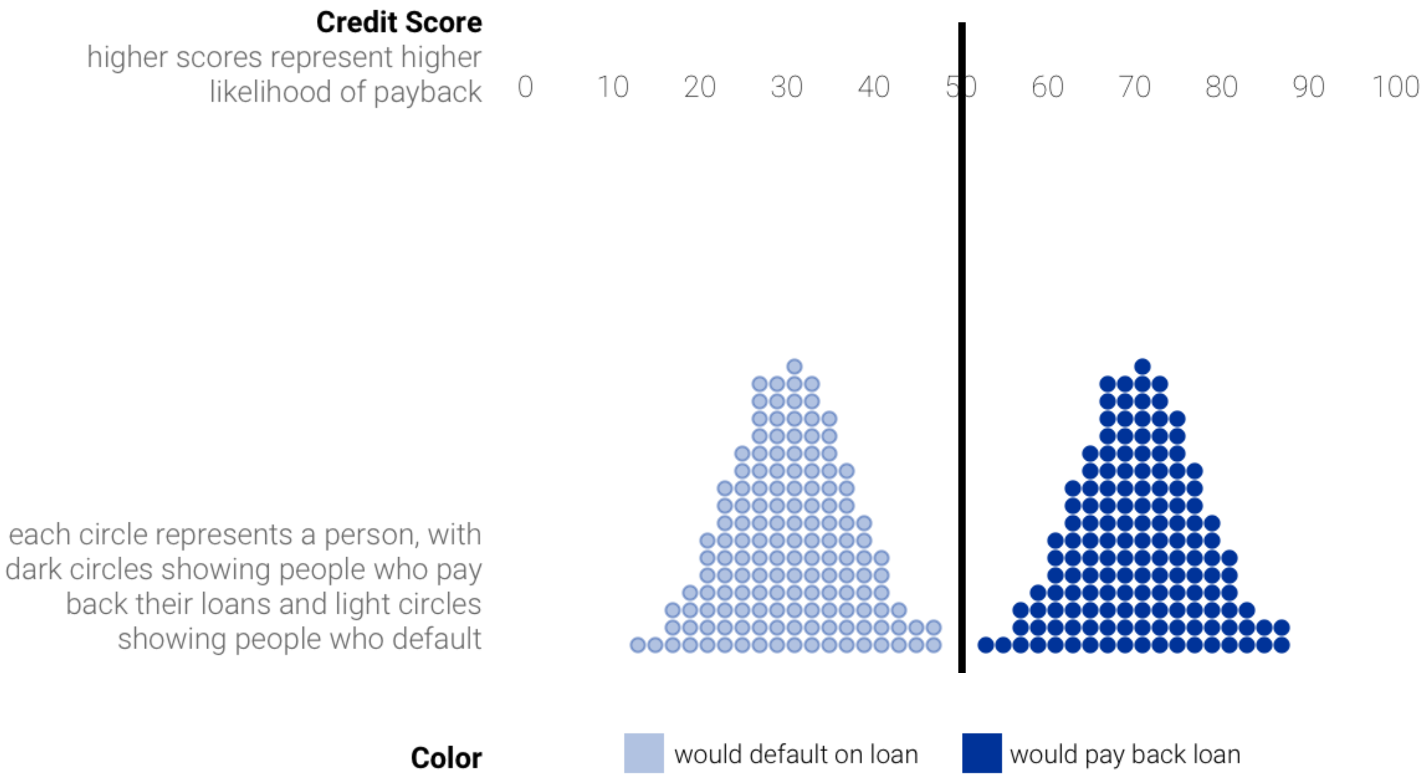
- Classification
- Ranking
- Recommendation

Fairness in Classification

Classification

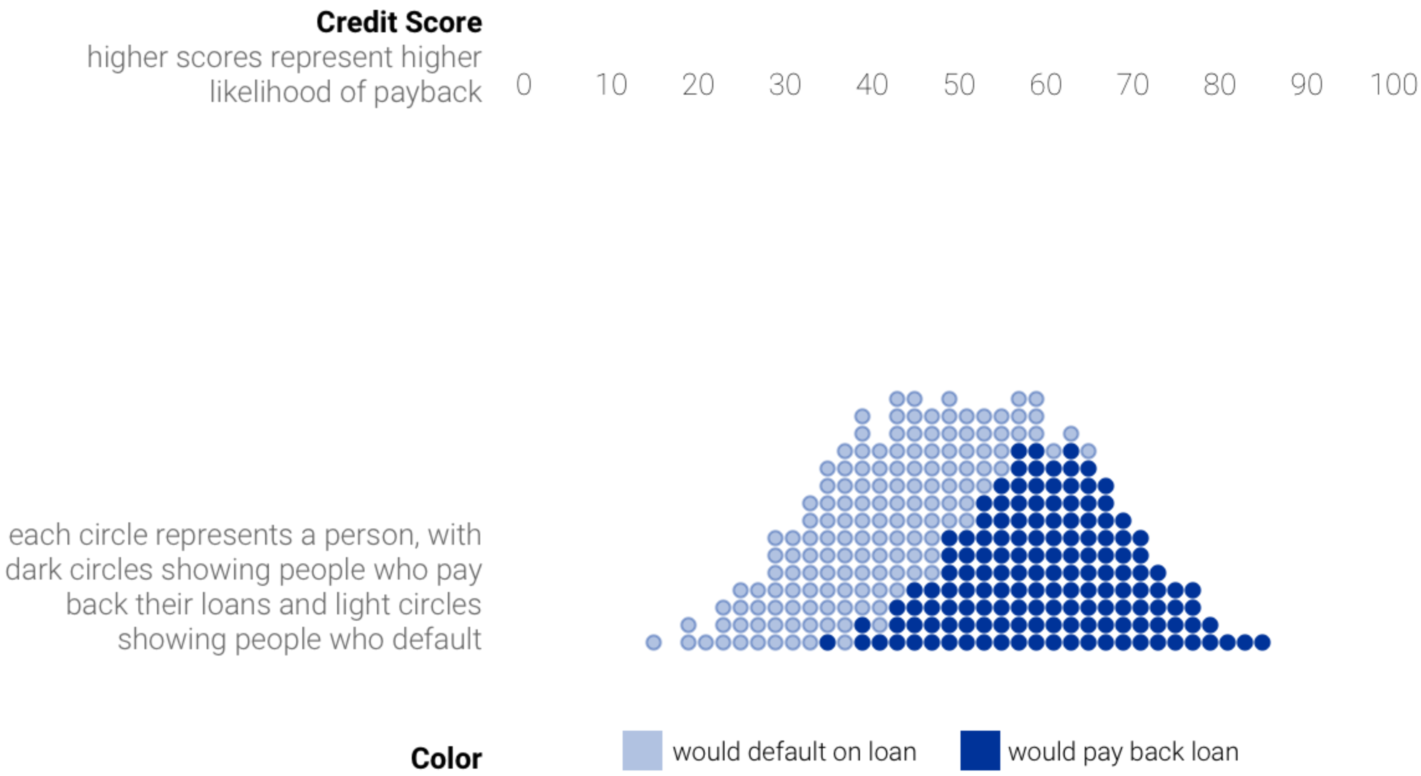
- assume **binary** classification (positive or negative classes)
  - e.g., accept a loan
- assume group fairness: **protected** vs. **non-protected** group
- assume that some **score** predicts the likelihood of being in the positive class
  - e.g., a credit score
- a score **threshold** determines the classification outcome
  - e.g., scores above threshold mean positive

How should we set the threshold?

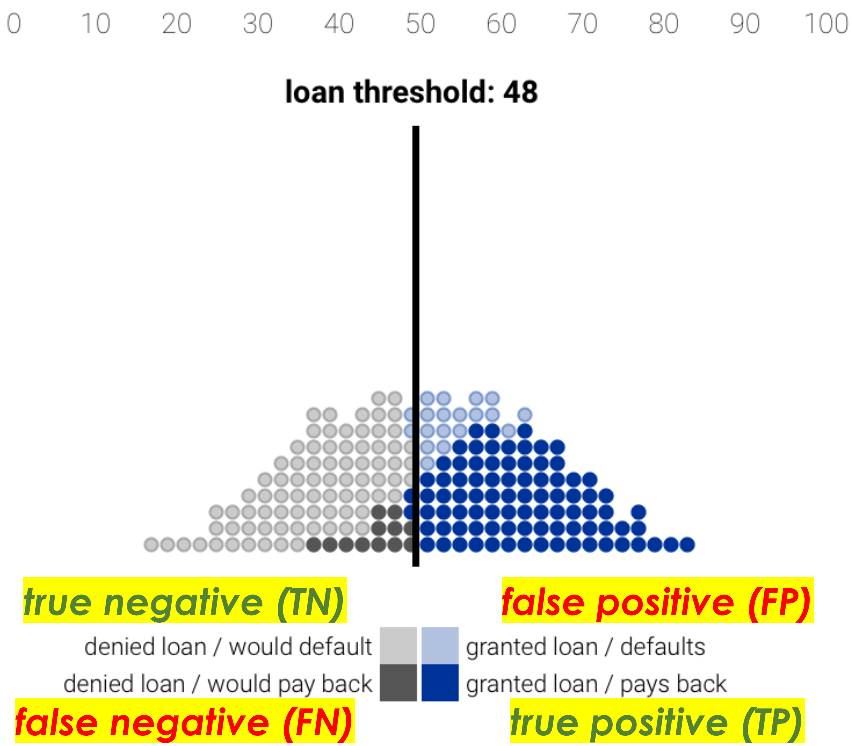


http://research.google.com/bigpicture/attacking-discrimination-in-ml/

How about now?

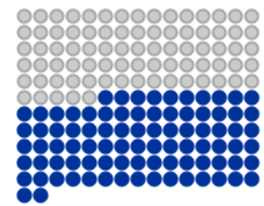


How should we set the threshold?



accuracy

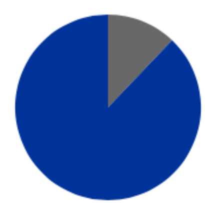
Correct 84%
loans granted to paying applicants and denied to defaulters



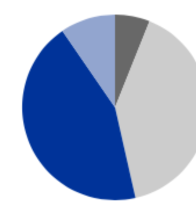
Incorrect 16%
loans denied to paying applicants and granted to defaulters

recall

True Positive Rate 88%
percentage of paying applications getting loans

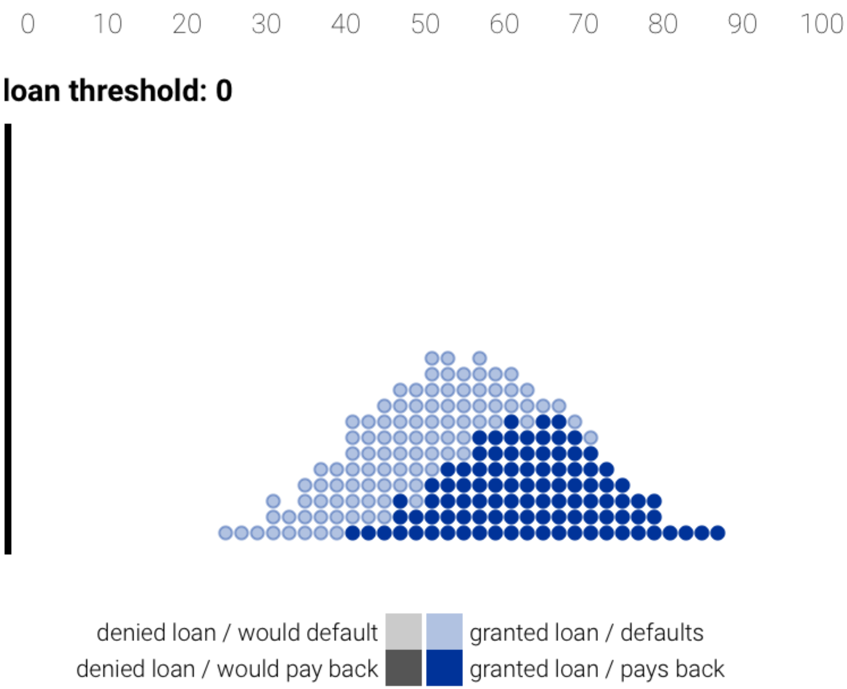


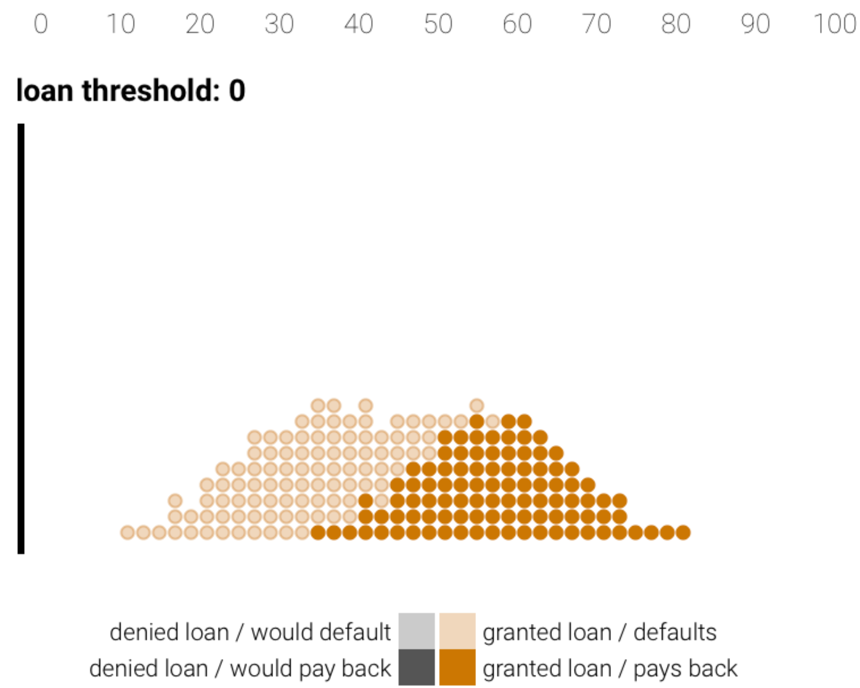
Positive Rate 54%
percentage of all applications getting loans



Two groups: Blue and Orange

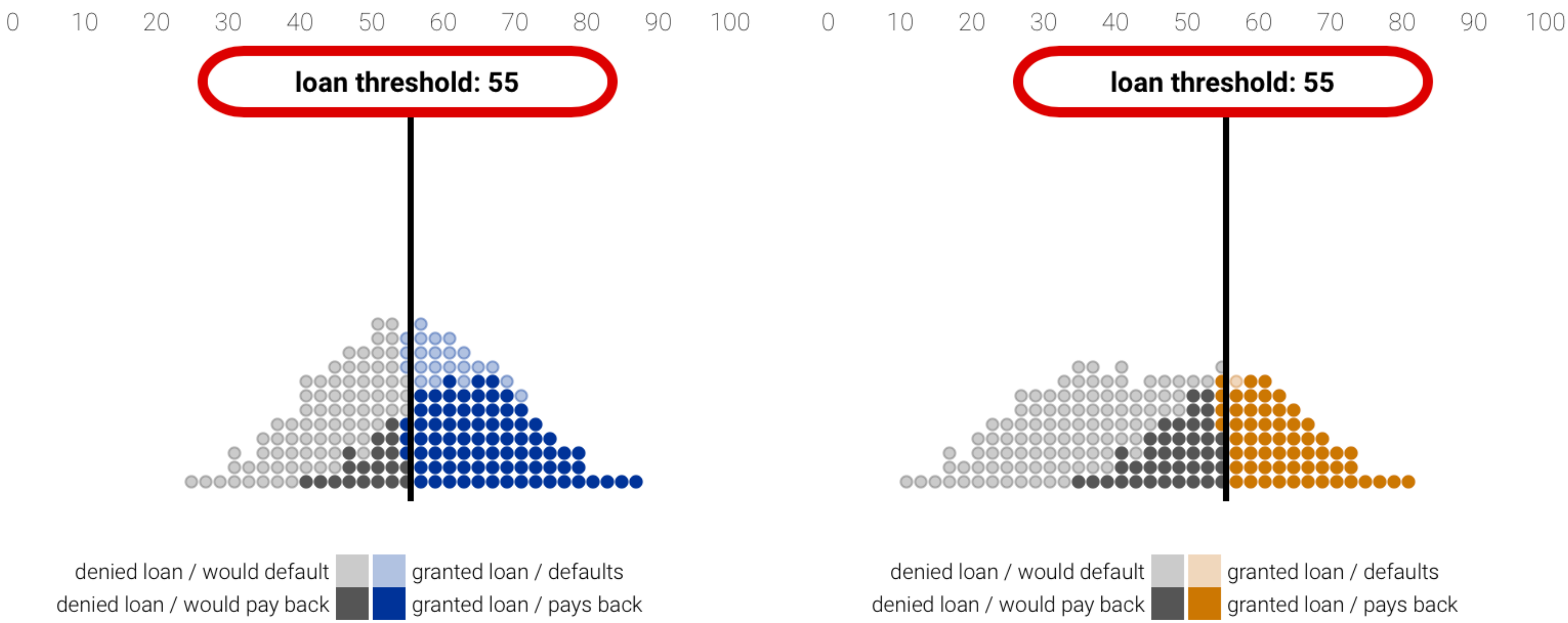
How should we set the thresholds for the two groups?





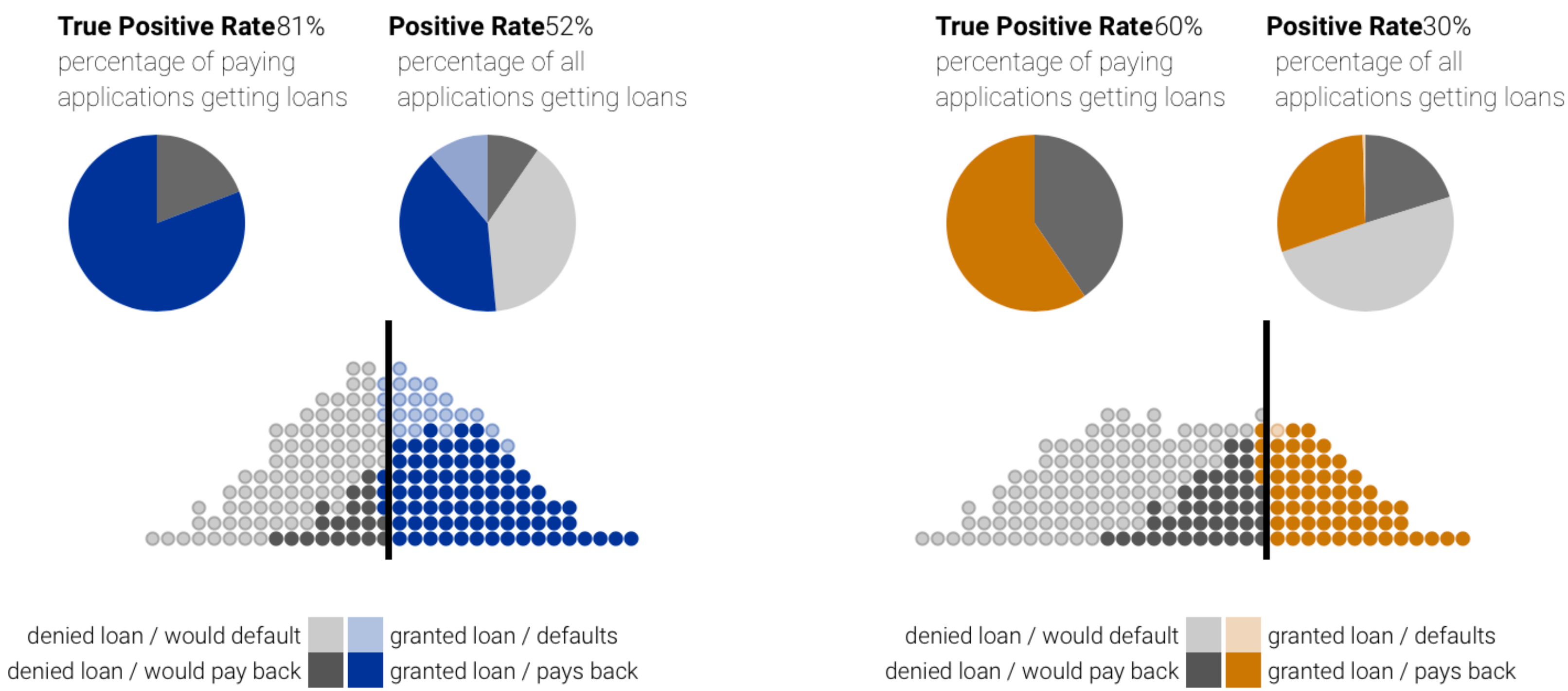
Equal Treatment (Color-Blindness)

We don't look at color, set the **same threshold.**



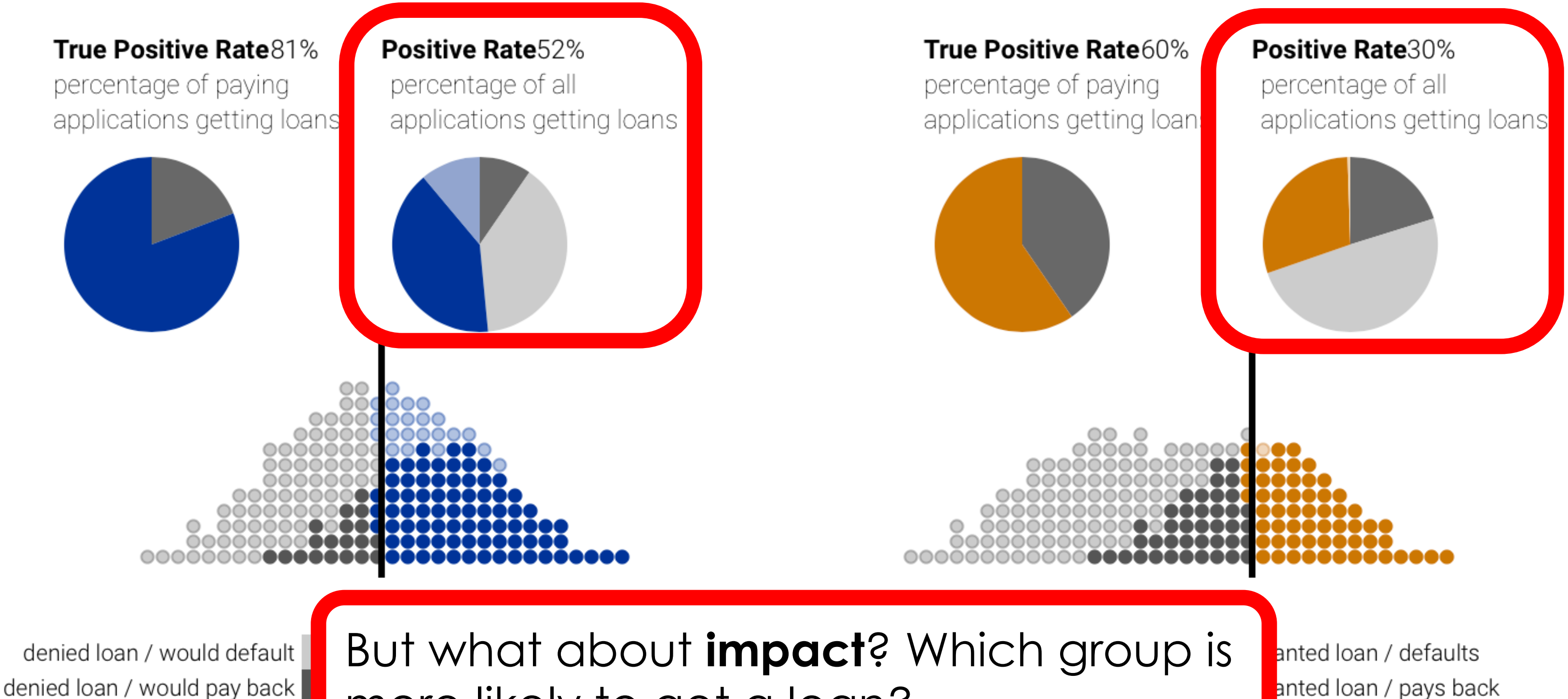
Equal Treatment (Color-Blindness)

We don’t look at color, set the **same threshold.**



Equal Treatment (Color-Blindness)

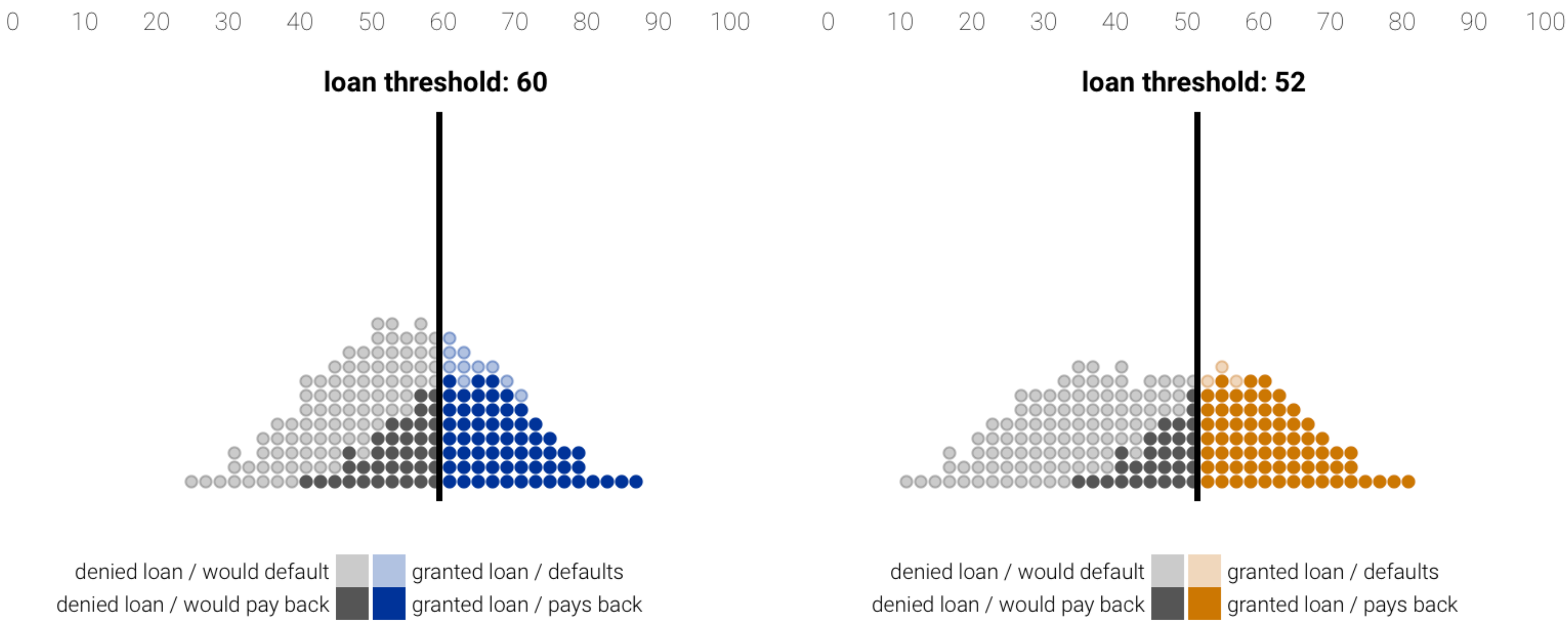
We don't look at color, set the **same threshold.**



But what about **impact**? Which group is more likely to get a loan?

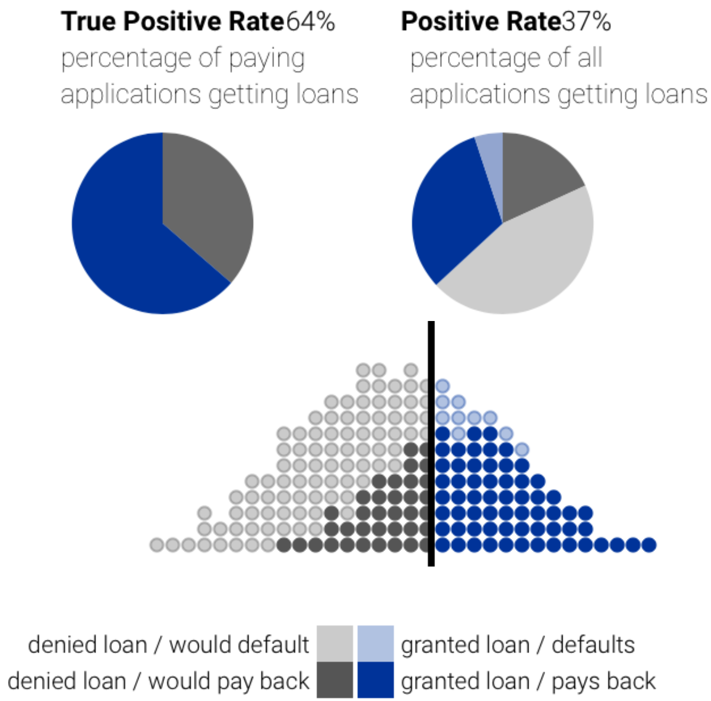
Group (Demographic/Statistical) Parity

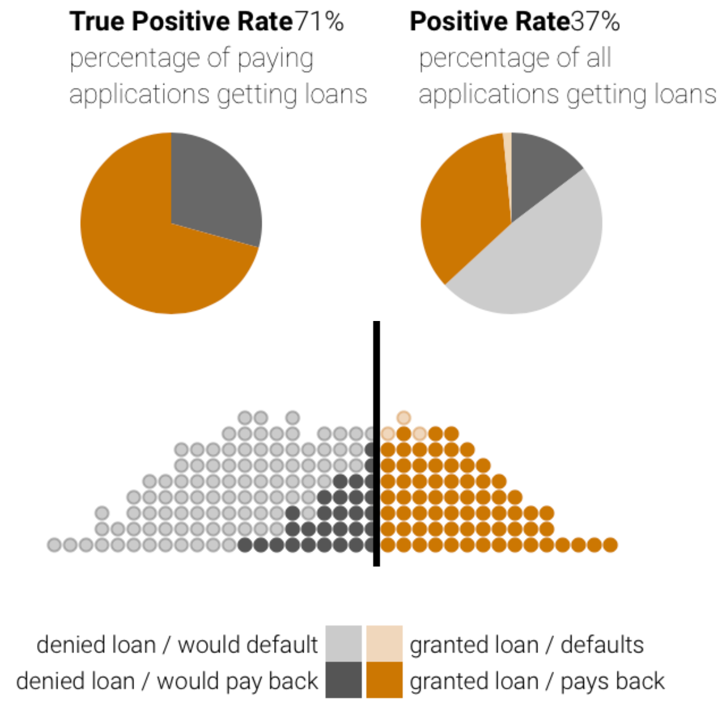
Set **different thresholds** so that groups have equal chance of getting a loan (*equal impact*)



Group (Demographic/Statistical) Parity

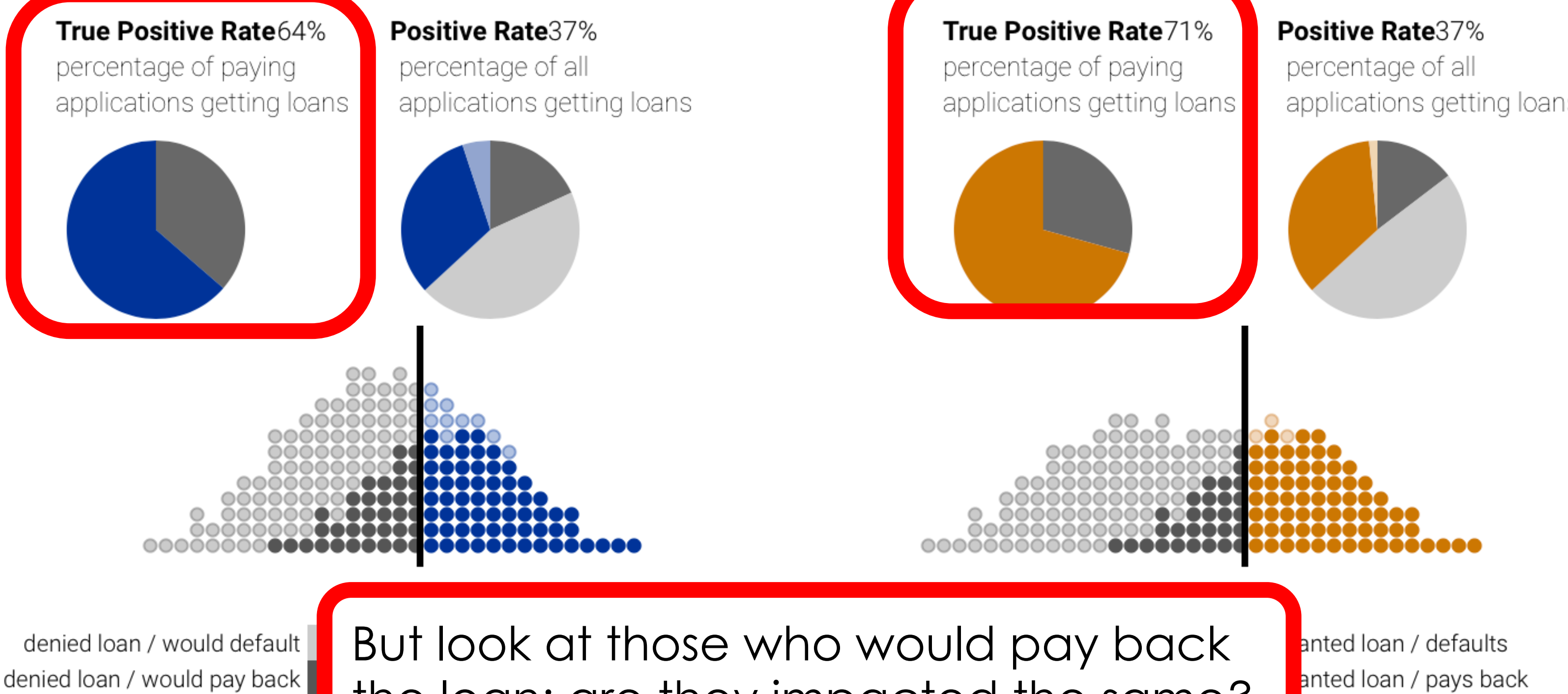
Set **different thresholds** so that groups have equal chance of getting a loan (*equal impact*)





Group (Demographic/Statistical) Parity

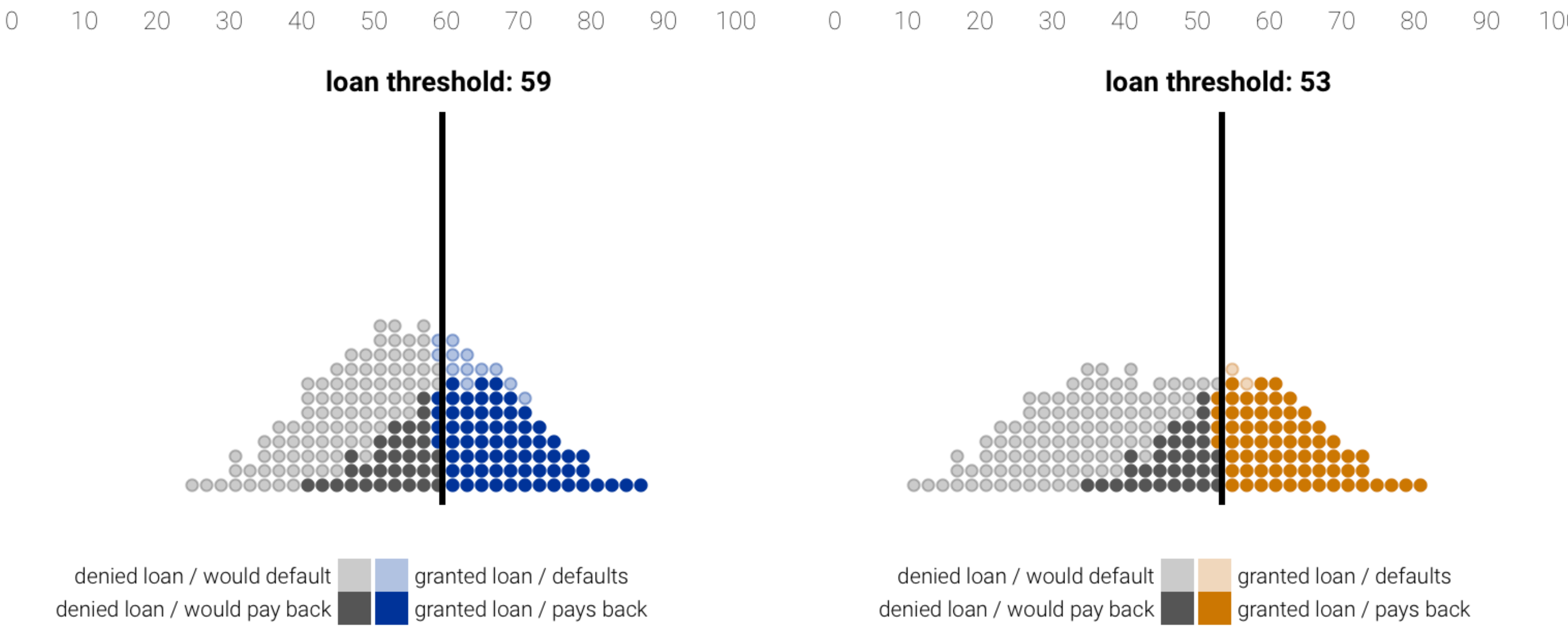
Set **different thresholds** so that groups have equal chance of getting a loan (*equal impact*)



But look at those who would pay back the loan; are they impacted the same?

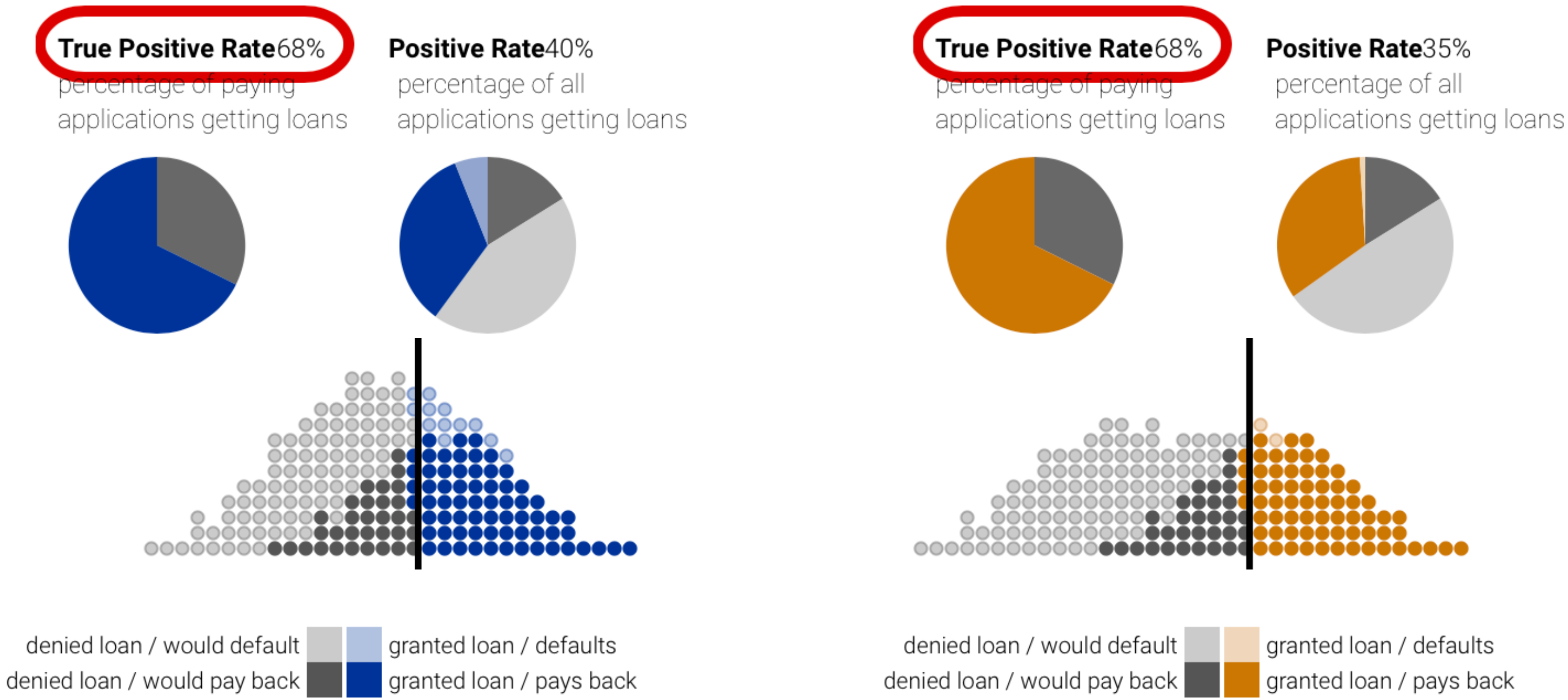
Equal Opportunity

Aim to be fair for those that would pay back the loan



Equal Opportunity

Aim to be fair for those that would pay back the loan



Fairness Notions

- **equal treatment** (color-blindness, fairness by unawareness): same threshold
- **group parity** (demographic or statistical parity): same positive rate
- **equal opportunity**: same true positive rate (recall)

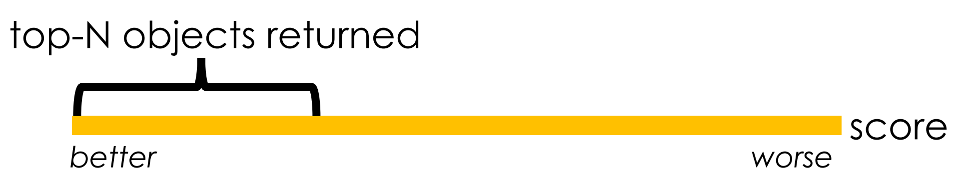
- you cannot have it all!
- there are various impossibility theorems related to fairness in classification

Fairness in Ranking

Ranking

- algorithm computes some **score** that encodes/predicts *goodness*
  - e.g., **merit**, relevance
- we want to **rank** objects (usually people) based on that score
  - e.g., candidates for a job position

- similarity with classification task: algorithm computes scores
- difference with classification task: there is **no threshold**; just the length (cut-off) of the ranked list



- so to achieve fairness we can only change the **scoring function**

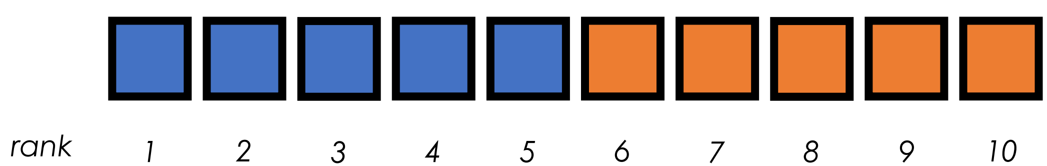
Group Parity

- how does **group parity** translate in ranking?
- remember for classification: group parity = achieve equal positive rate in each group
  - positive (predicted as relevant) = you appear in the ranked list

- so if we look at the top-N objects:
- group parity = *half* from protected, *half* from non-protected

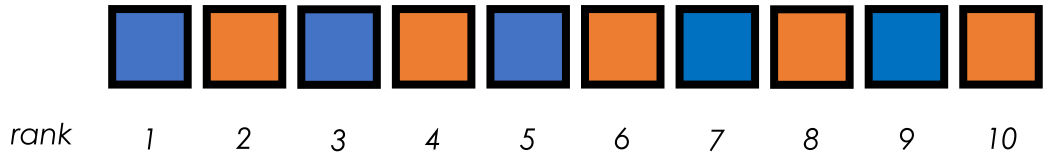
example: rank 10 people from two groups

- is this fair? how can we fix it?



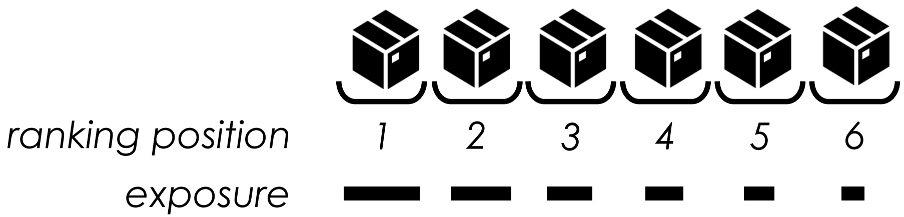
Group Parity

- stronger requirement: **for every prefix** of the ranked list, there should be group parity



- generalize a bit: assume we have a **target ratio**/mix of groups
  - e.g., 40% blue, 60% orange
- how can you achieve this?

Exposure in rankings



- a ranked list **exposes** items to the user
- the amount of exposure depends on the **ranking position**
  - top positions give higher exposure (aka *position bias*)

- the exposure of an item **does not depend** on relevance, utility, user satisfaction, etc.; just on ranking position

Fairness of Exposure

- a ranking **exposes** objects at varying degrees
  - **position bias**: higher ranked items are exposed more

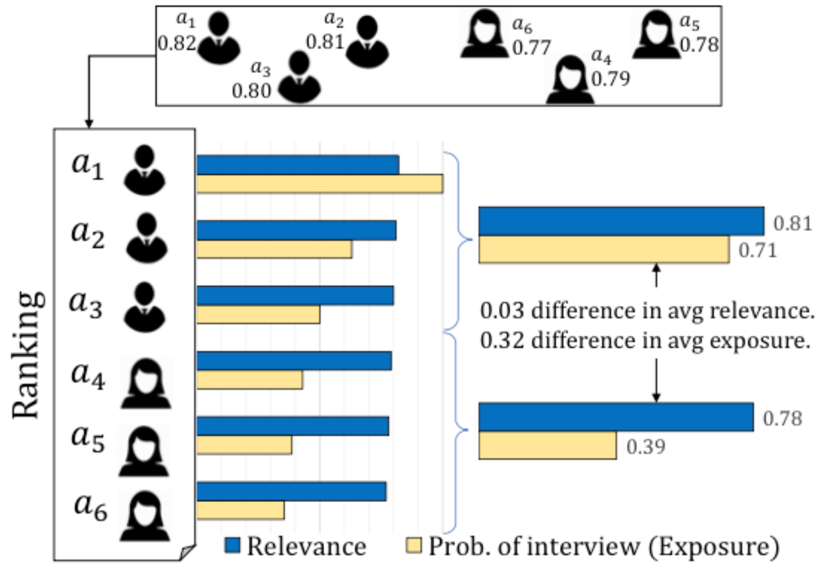
- fairness aspect: **exposure** should be **proportional** to **merit**
  - if an item is x times better than another, it should receive x times more exposure

Fairness of Exposure

- fairness = exposure ~ merit (relevance)
- can we achieve this? almost never
  - exposure is **fixed**
  - merit **depends** on ranking scores

job applicants

ranking by merit



high discrepancy in group exposure

[2018 KDD A. Singh T. Joachims] *Fairness of Exposure in Rankings*

Fairness of Exposure

- what can we do then?
  - be fair in the long-term: **amortized fairness**
  - or be fair **probabilistically**: probabilistic rankings

- formulate an optimization problem
  - **maximize** relevance **subject to** equal exposure, or
  - **minimize** exposure discrepancy **subject to** relevance quality drop

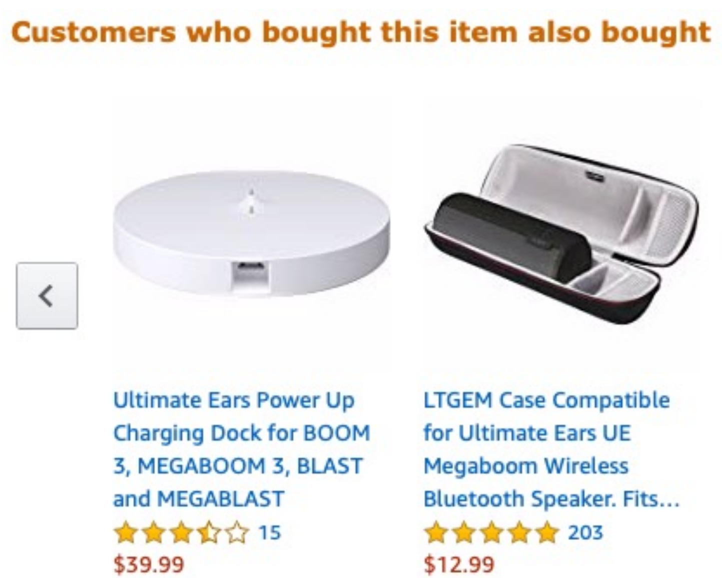
[2018 SIGIR A. Biega et al.] *Equity of Attention: Amortizing Individual Fairness in Rankings*
[2018 KDD A. Singh T. Joachims] *Fairness of Exposure in Rankings*

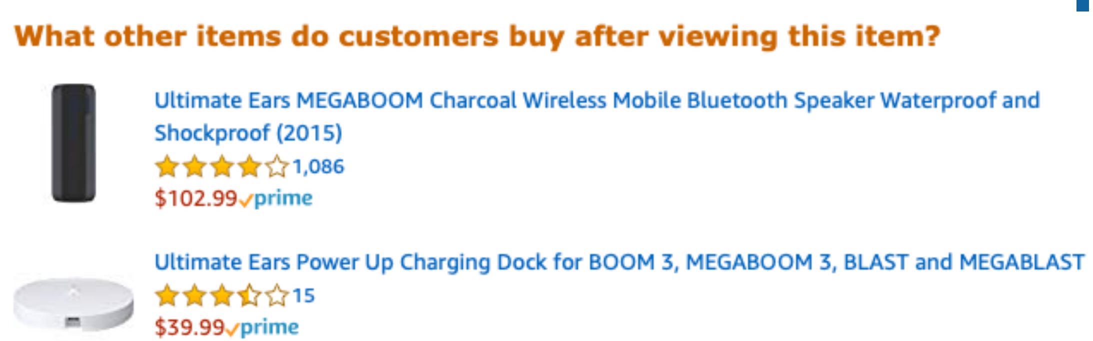
Fairness in Recommendations

What are recommender systems?

We Have Recommendations for You

Sign in to see personalized recommendations

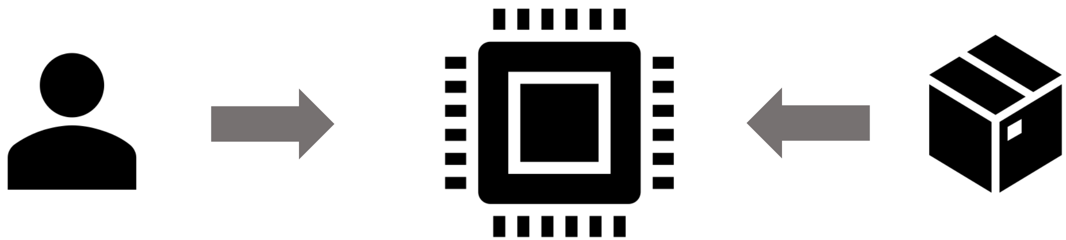




Stakeholders in recommender systems

recommenders = match **users** with **items**

users *recommender* *items*



consumers **system provider** **producers**

a multi-sided market with possibly conflicting interests
and various **fairness concerns**

Fairness in recommendations

fairness is the **fair distribution** of a **resource**

to make it less abstract and cyclic, answer three questions:

- fair **for whom?** (which are the *protected groups*)
  - **consumers** (end users, buyers)
  - **producers** (item providers, creators, sellers)
- distribution **of what** resource?
  - **accuracy** (effectiveness, utility, satisfaction, quality of service, etc.)
    - e.g., items are good matches for users; users are good matches to items
  - **exposure** (attention)
    - e.g., what users see; to whom items are recommended
- **when** is the distribution fair?
  - specify the **optimal state** of fairness
  - quantify how far the optimal state is = define some (**un-**)**fairness measure**

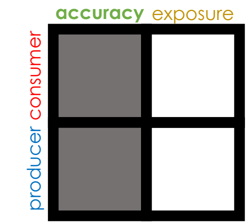
Taxonomy of fairness definitions in recommenders

of what?

for whom?

	accuracy	**exposure**
consumer	prediction accuracy [2017 NeurIPS S. Yao et al.] ranking accuracy [2018 FAT M. Ekstrand et al.]	equal exposure [2018 FAT R. Burke et al.]
producer	ranking accuracy [2019 KDD A. Beutel et al.]	exposure coverage [2018 CIKM, 2019 RecSys] equal exposure [2018 FAT R. Burke et al.] calibrated exposure [2018 RecSys]

some representative works



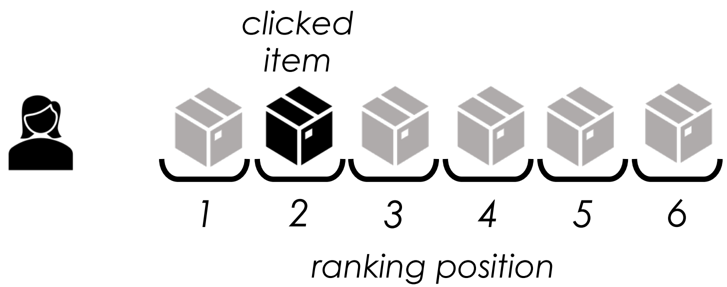
Accuracy in recommender systems

view 1: the recommender **matches** users with items

prediction accuracy: how good is the matching (regression task)

- metrics: absolute error (MAE), root mean square error (RMSE)

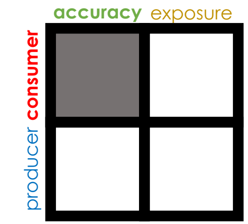
view 2: the recommender **ranks** items for a user



reciprocal rank (RR) = 1/2

ranking accuracy: how good is the ranking (ranking task)

- metrics: reciprocal rank (RR), discounted cumulative gain (DCG)

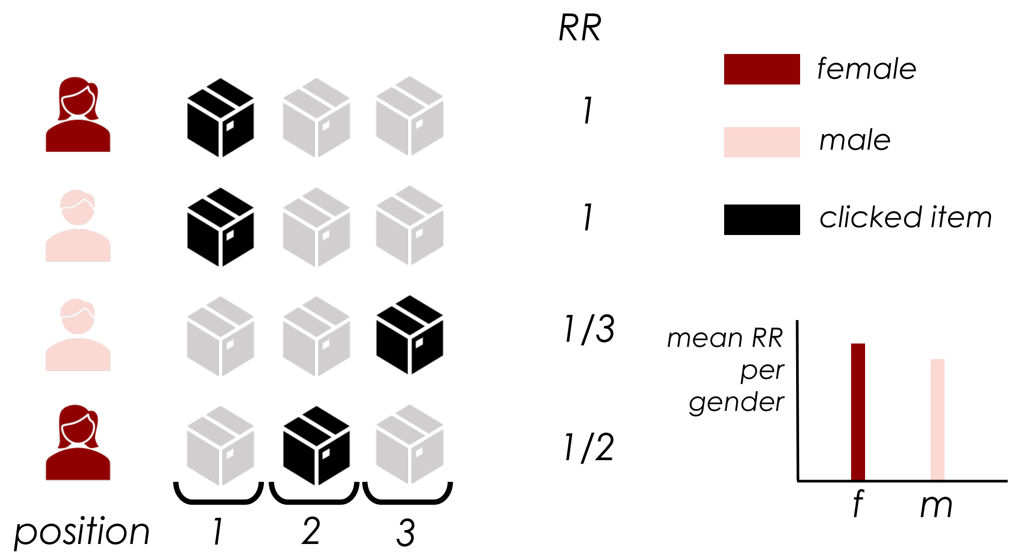


Fairness of **accuracy** for **consumers**

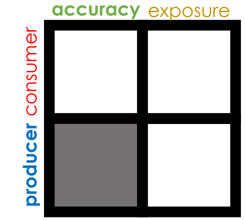
*“**Males** and **females** should experience the same quality of service.”*

operationalized as:

1. **group** users **by gender**
2. measure **total accuracy** per gender
3. **assess** if distribution is **fair**



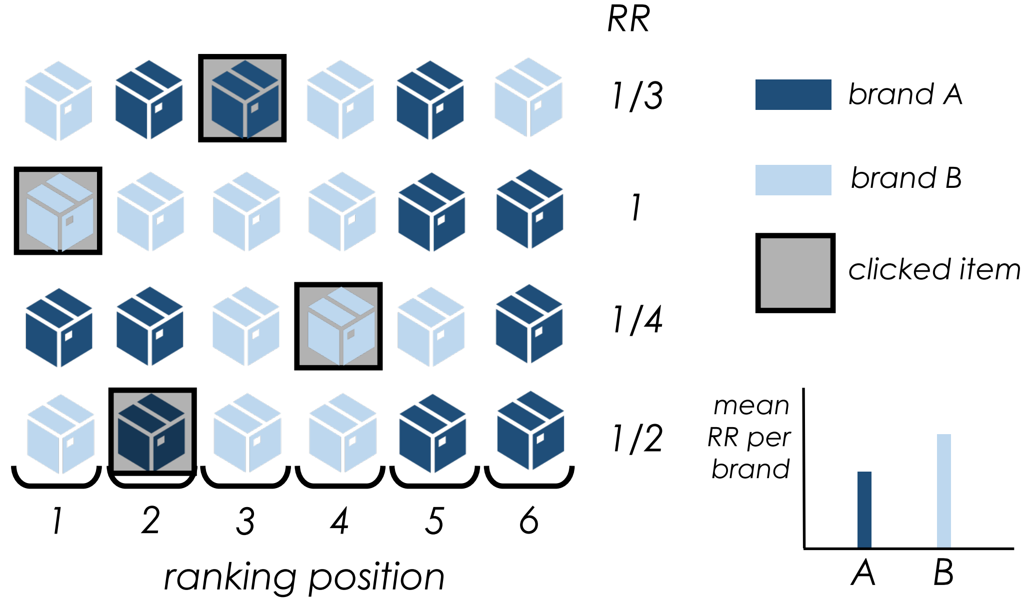
Fairness of accuracy for producers

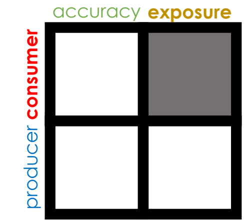


*"When recommending products, **all brands** should have similar accuracy."*

operationalized as:

1. **group** items **by brand**
2. measure **total accuracy** per brand
3. **assess** if distribution is **fair**



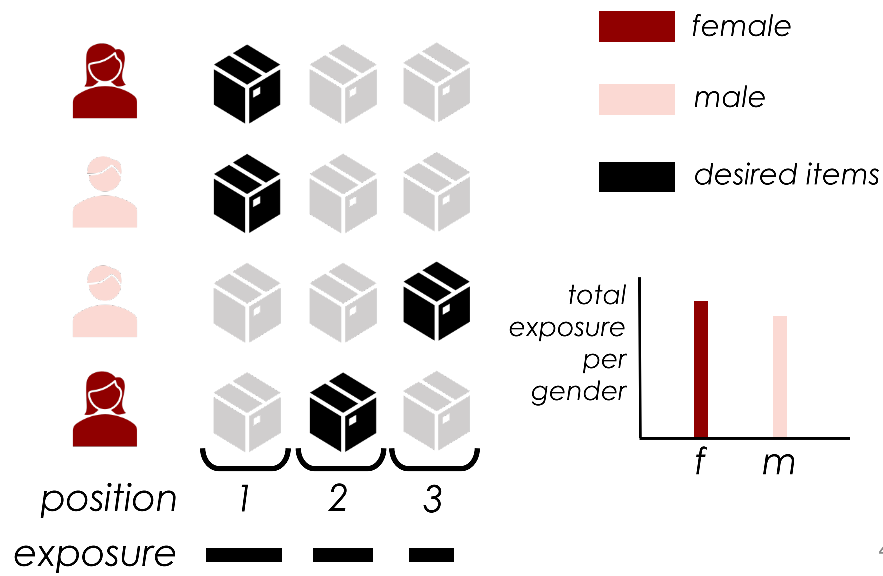


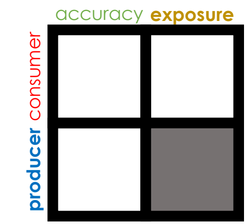
Fairness of exposure for consumers

"When recommending jobs, ***males*** *and* ***females*** *should see the same number of executive openings."*

operationalized as:

1. **group** users **by gender**
2. measure **total exposure** per gender
3. **assess** if distribution is **fair**



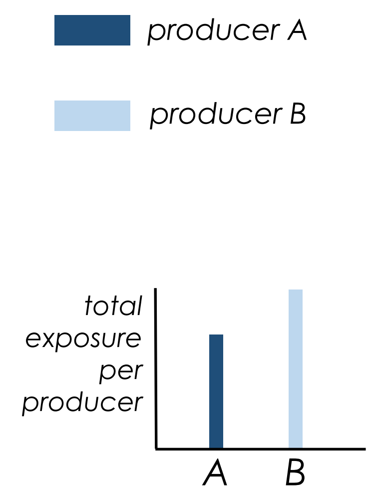


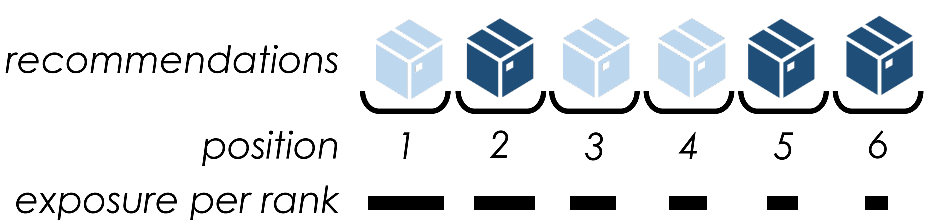
Fairness of **exposure** for **producers**

*"When recommending products, **all brands** should be fairly exposed."*

operationalized as:

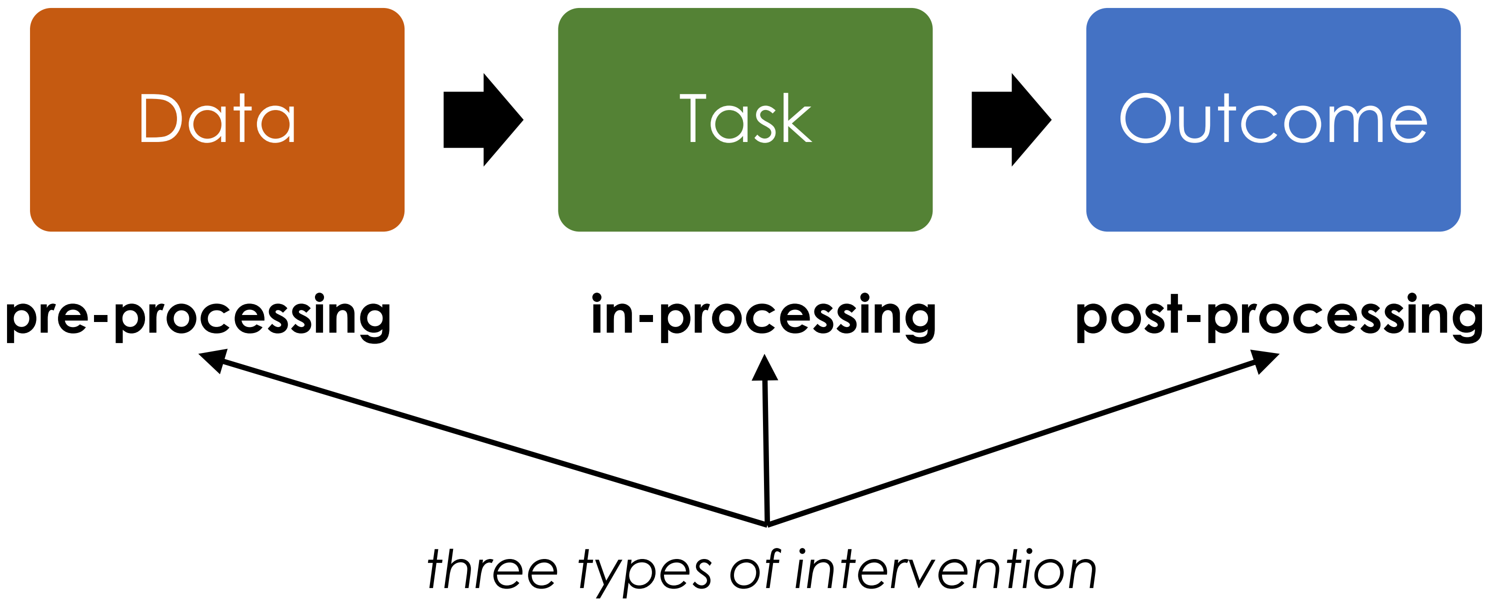
1. **group** items **by producer**
2. measure **total exposure** per producer
3. **assess** if the distribution is **fair**





How to Achieve Fairness

How to Achieve Fairness



Post-Processing for Group Parity in Ranking

Step 1. Determine **minimum representation** from the **protected** group in each **prefix**

- Suppose we want at least $\lfloor 40\% N \rfloor$ orange at every top-N

N	1	2	3	4	5	6	7	8	9	10
# orange	0	0	1	1	2	2	2	3	3	4

Step 2. Create two sublists of objects (blue and orange), sorted on score

a b c d e f

better

g h i j k l

better

Step 3. Build the ranked list incrementally

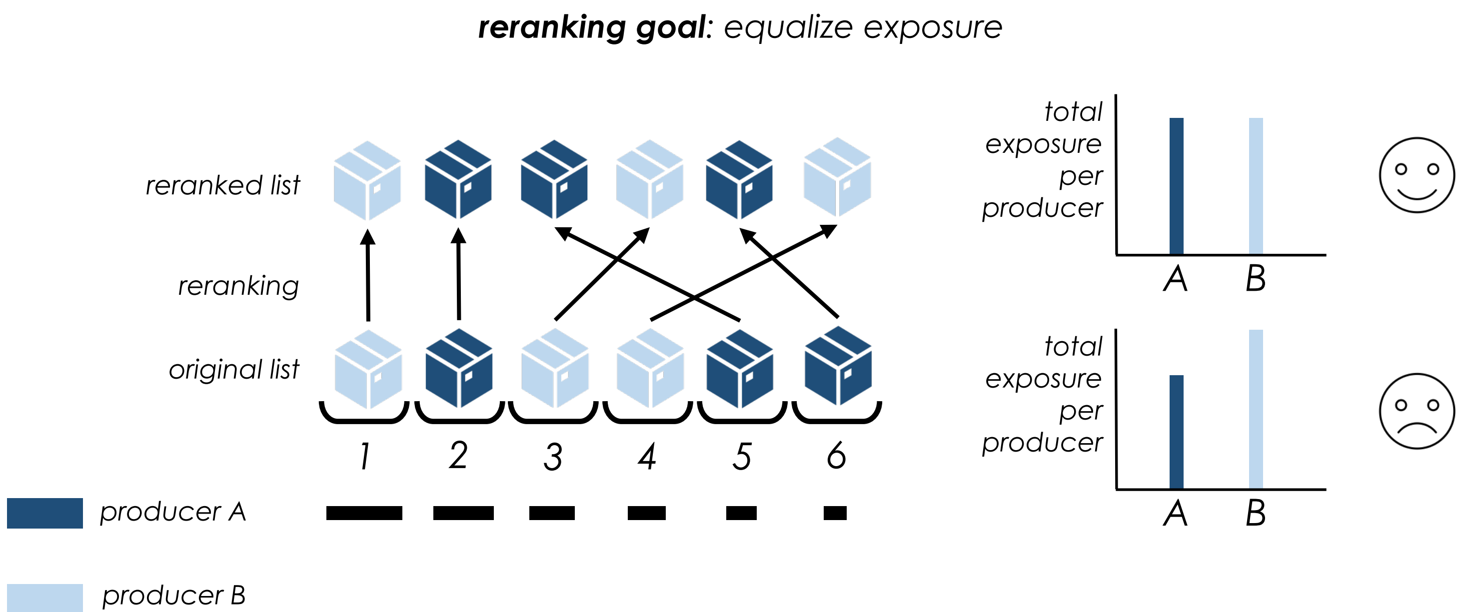
- at each rank choose the **best** from either group (why best?),
- **unless** you *must* choose from the protected (to ensure minimum representation)

rank	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
	a	b	g	c	h	d	e	i	f	j

Post-Processing for Fair Exposure in Ranking

change the position of items to improve fairness

trade-off between **fairness** and **accuracy**



Post-Processing for Ranking

the **original list** optimizes some **internal measure** of **utility**

- e.g., relevance, click-through rate

but to increase **fairness**, the list has to be reranked

- at the expense of **utility**

reranking must **trade-off** two objectives: **utility** and **fairness**

possibilities:

maxU	maximize **utility** given a constraint on **fairness**
maxF	maximize **fairness** given a constraint on **utility**
U+F	maximize a combination of **utility** and **fairness**

Other Fairness Challenges

Continuous Protected Attributes

- All definitions compare a **protected group** (e.g., blacks, women) against the **non-protected**

- What happens in the case of continuous protected attributes, e.g., **age**, **income**, ***location***

- Groups are not defined beforehand, and defining them can lead to **gerrymandering**
  - purposefully setting the group boundaries to **hide discrimination**

Continuous Protected Attributes

- All definitions compare a **protected group** (e.g., blacks, women) against the **non-protected**

- What happens in the case of continuous protected attributes, e.g., **age**, **income**, ***location***

- Groups are not defined beforehand, and defining them can lead to **gerrymandering**
  - purposefully setting the group boundaries to **hide discrimination**

Spatially Fair Distribution of Outcomes

X	O	X	O
O	X	O	X
X	O	X	O
O	X	O	X

Spatially Unfair Distribution of Outcomes

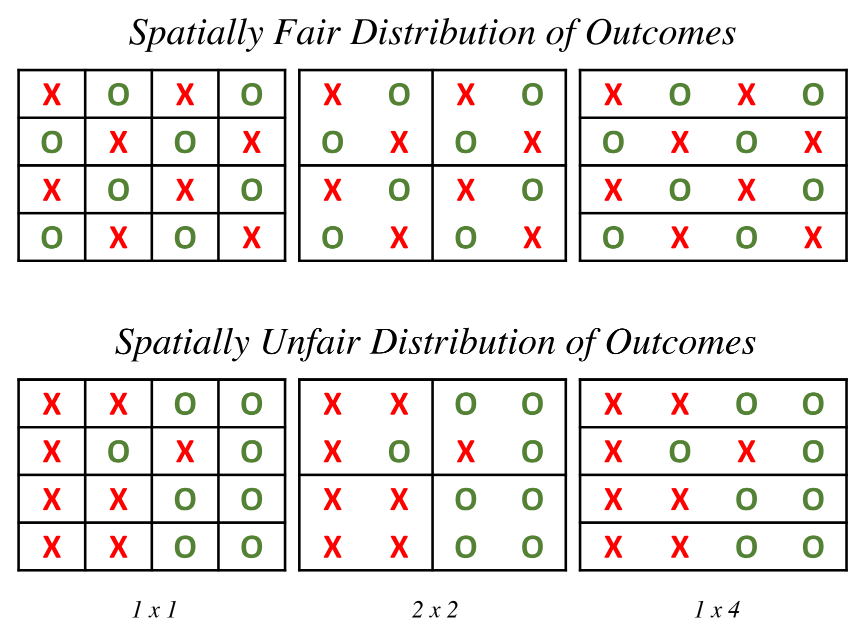
X	X	O	O
X	O	X	O
X	X	O	O
X	X	O	O

Continuous Protected Attributes

- All definitions compare a **protected group** (e.g., blacks, women) against the **non-protected**

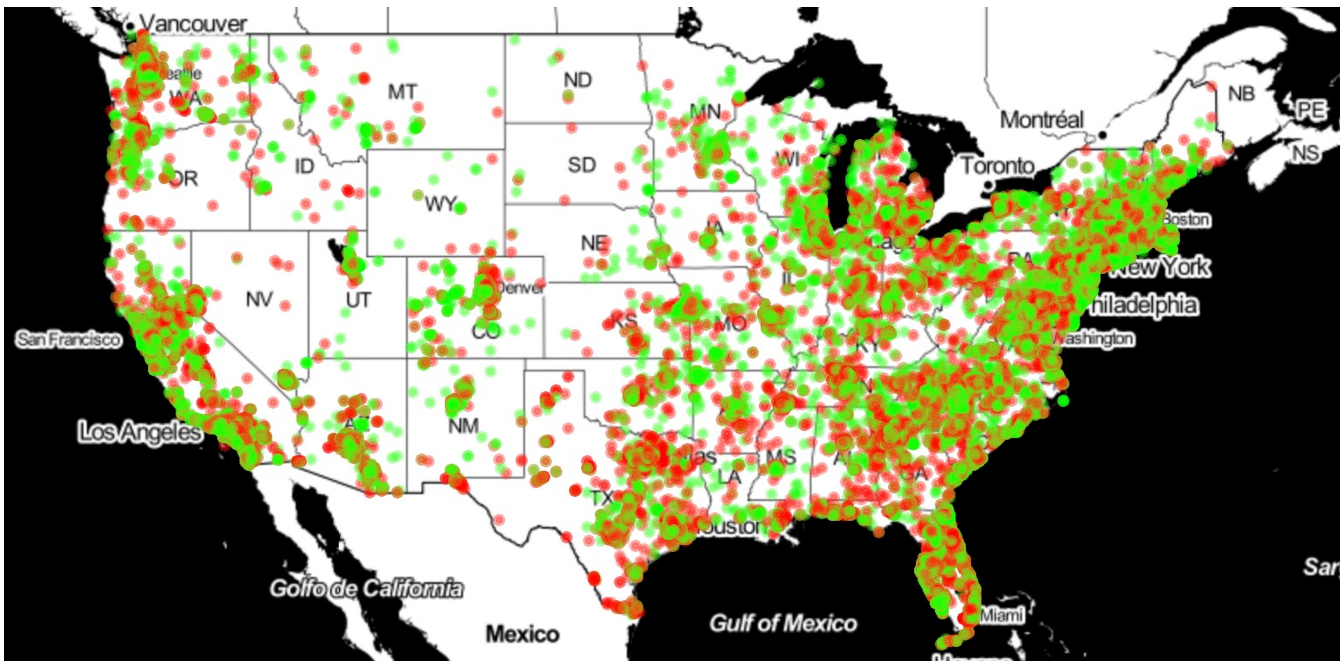
- What happens in the case of continuous protected attributes, e.g., **age**, **income**, ***location***

- Groups are not defined beforehand, and defining them can lead to **gerrymandering**
  - purposefully setting the group boundaries to **hide discrimination**



Location Fairness

- We want **outcomes** to be **independent of location**
- Consider **Loan Application** Accepts and Rejects per location
- Is this map **fair**?



Discover Unfairness via Explanations

- Previous definitions require the groups (protected vs. non-protected) to be fixed and known in advance
- Can you **discover** groups (or sub-groups) where discrimination occurs?

Counterfactual Explanations

- A person x **does not receive** a loan
  - x = {Race=Caucasian, Gender=Female, Income=Low}
- What are the **minimal changes** to **receive** the loan?
  - {Gender=Female} → {Gender=Male}

- This shows **discrimination** based on gender!