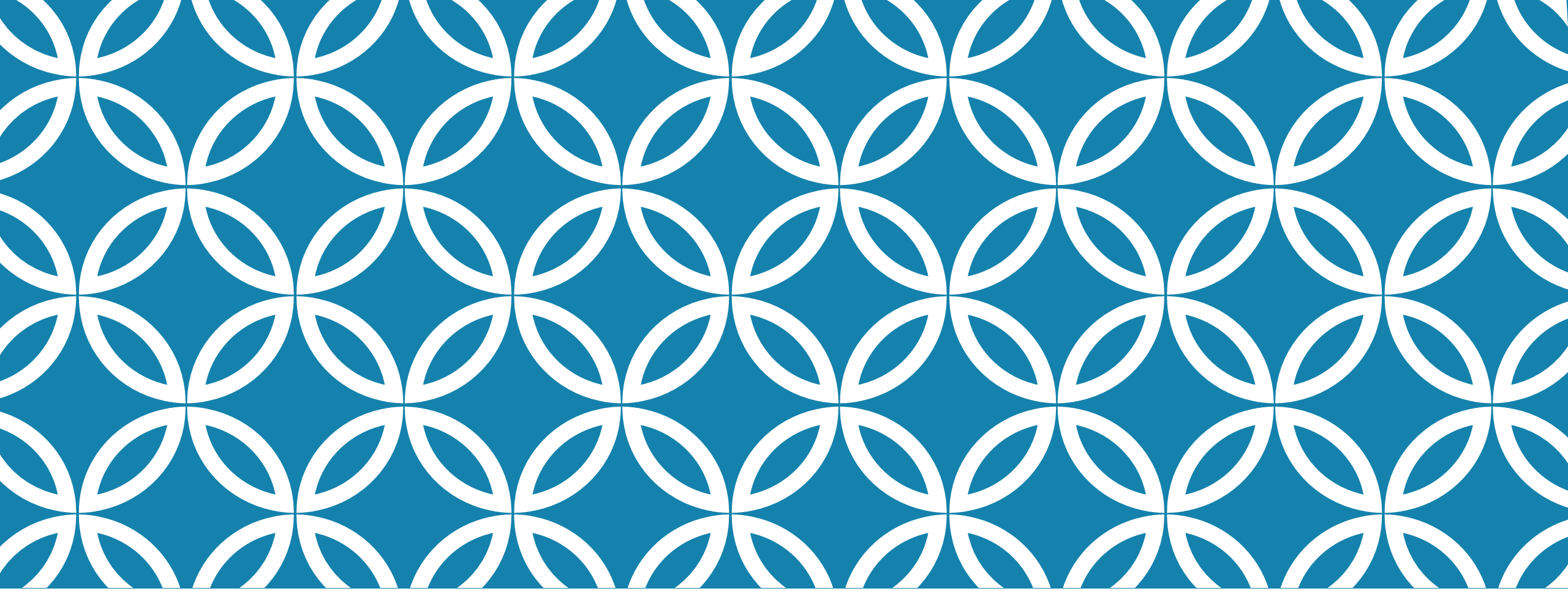


UNIVERSITAT DE BARCELONA

SPECIAL CLASS: BIAS AND FAIRNESS IN MACHINE LEARNING

T12. INTRODUCTION TO MACHINE LEARNING

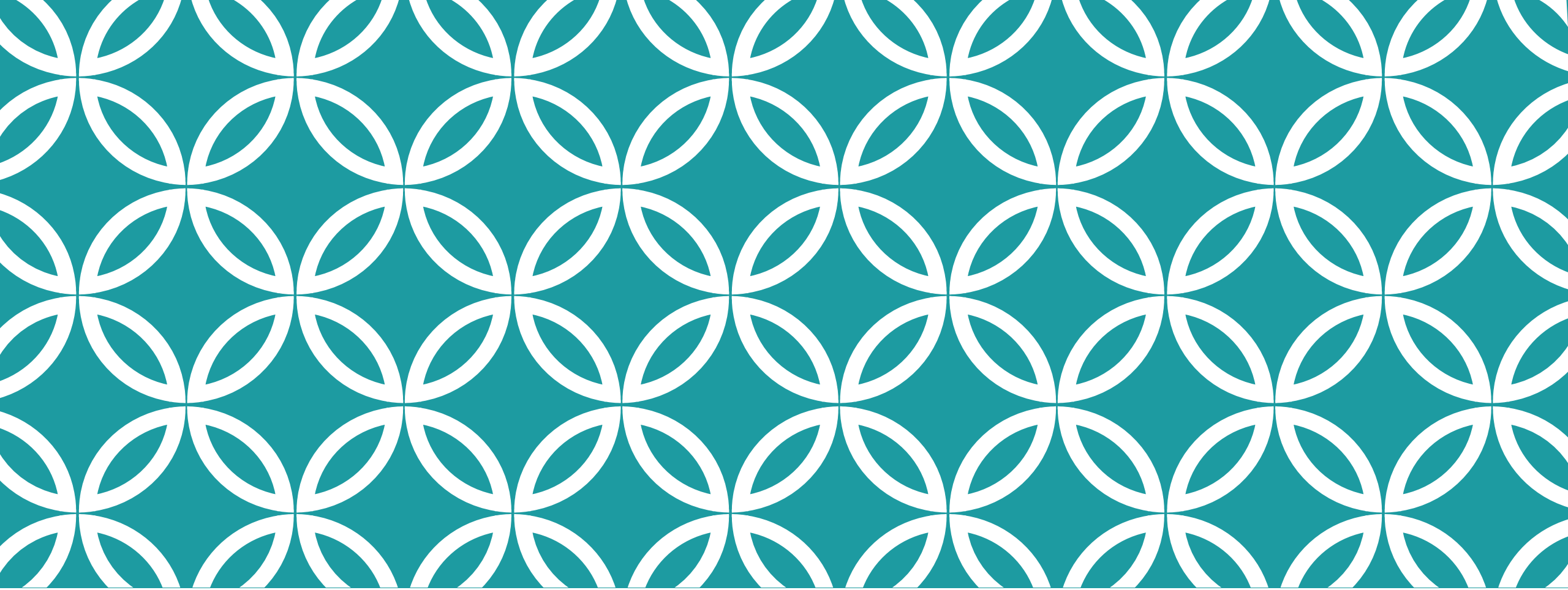
Maria Salamó
Universitat de Barcelona, Spain
maria.salamo@ub.edu

LEARNING OBJECTIVES

- Understand **key concepts** of bias and fairness
- Raise **awareness** on the **importance** and the relevance of considering data and algorithmic bias and fairness issues in Machine Learning
- Showcase approaches that **mitigate bias/fairness** along with the recommendation pipeline and asses their influence on stakeholders
- Play with **recommendation pipelines** and conduct **exploratory analysis** aimed at uncovering sources of bias along them

CONTENT

1. Bias and fairness in AI and Machine Learning
2. Brief introduction to bias and fairness in RecSys
3. Example 1: Bias in calibrated recommendations
4. Example 2: Provider fairness across continents in collaborative recommender systems
5. Hands on Bias in Recommender Systems



BIAS AND FAIRNESS IN AI AND MACHINE LEARNING

WHAT IS AI AND MACHINE LEARNING?

Artificial intelligence is an overall term describing a set of different kinds of techniques to make computers behave in some kind of intelligent fashion. There is no agreed definition of AI, but in general the ability to perform tasks without supervision and to learn so as to improve performance are key parts of AI.

Machine learning is a big topic in AI. Machine learning is a set of algorithms which by themselves learn to make decisions or to structure data. Supervised and unsupervised learning are based on data, while reinforcement learning is where the algorithm uses trial and error to learn to make sequences of decisions.

DEFINITION OF BIAS

"inclination or prejudice for or against one person or group, especially in a way considered to be unfair"

- **Look**! The definition of bias includes the word "**unfair**".
- It's easy to see why the terms bias and fairness get confused for each other a lot.

DEFINITION OF FAIRNESS

"impartial and just treatment or behaviour without favouritism or discrimination"

In the context of Machine Learning

"An algorithm is fair if it makes predictions that do not favour or discriminate against certain individuals or groups based on sensitive characteristics"

Can an Algorithm Hire Better Than a Human?

Claire Cain Miller @clairecm JUNE 25, 2015

Hiring and
jobs to be a
skills that c
social cues.

But people
decisions, c
nothing to
applicant h
likes the sa

That is one
broken. The

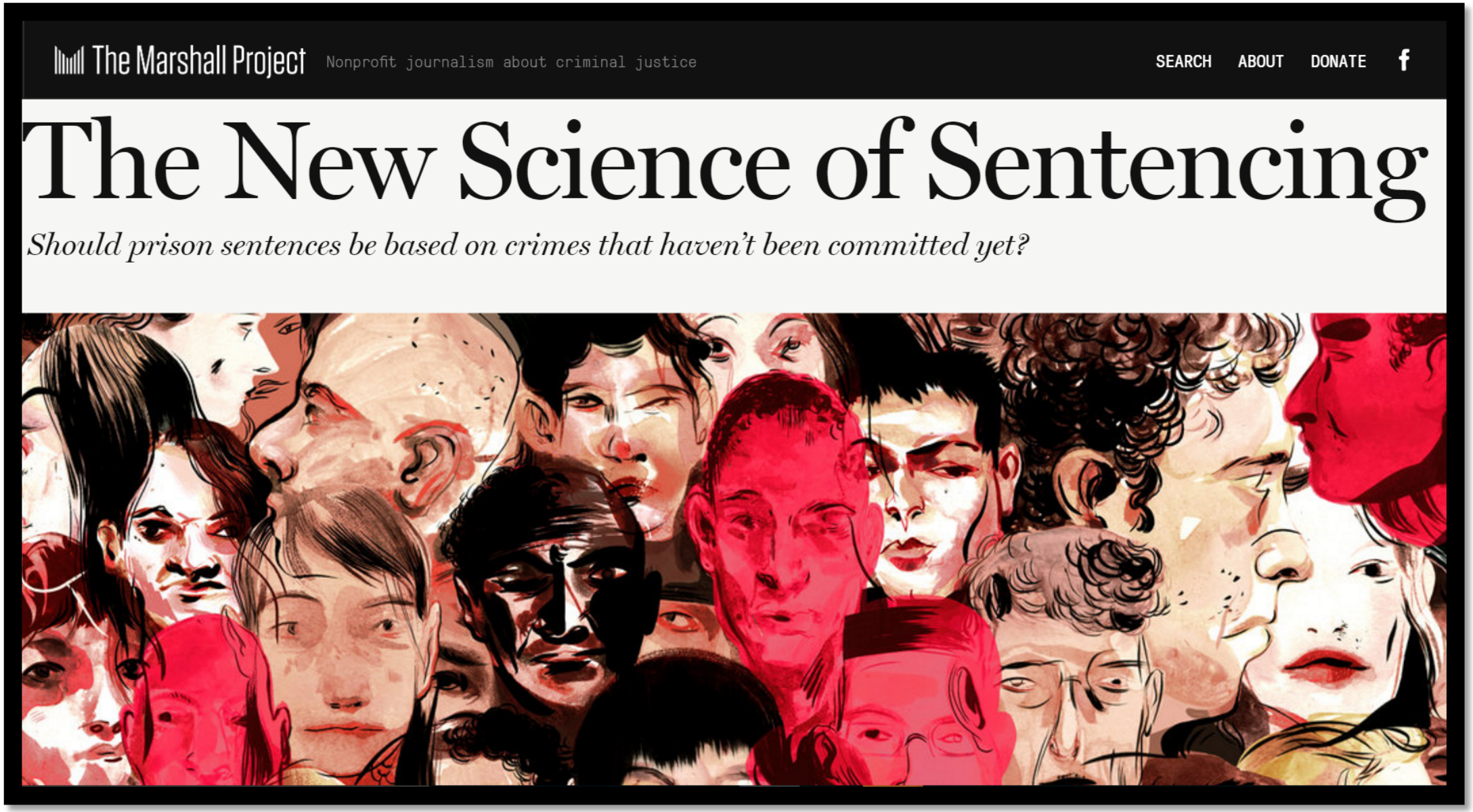
A new wave
and GapJu

The Algorithm That Beats Your Bank Manager

ters make better decisions
ns? One tech firm says they
when it comes to lending

is a short-term lending
at uses an algorithm and
of pieces of information
e in a few seconds whether
another loan shark
by being highly selective
disrupt the space

PREDICTIVE POLICING: USING MACHINE LEARNING TO DETECT PATTERNS OF CRIME



https://www.themarshallproject.org/2015/08/04/the-new-science-of-sentencing

Beauty contest judged by AI and the robots discriminate against dark skin

3 days ago | Published by : Avi

Sept9: The first international
such as facial symmetry and
this year, roughly 6,000 peop
intelligence, supported by co
"human beauty".

But when the results came in
winners: the robots did not li

Is an algorithm any less racist than a human?

ney U.S. + Business Markets Tech Media Personal Finance Small Biz Luxury stock tickers

is racist: How data is driving inequality

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

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of $80.

e

control, determining the
on OkCupid and the
to make decisions about
l policing.

gorithms that rely on data
nfluence. Algorithms are
learning algorithms adjust
sult, say researchers in
reinforce human

showed an ad for high-
wed the ad to women, a
hers found.

s for arrest records were
for distinctively black
eral Trade Commission said
ow-income neighborhoods

10

s Trending

This foldable bike
helmet is made fro
paper

Driving the most
beautiful productio
car ever

t will now cost just
$29 to fix a cracke
Phone

WHEN ALGORITHMS DISCRIMINATE

"There is a widespread belief that software and algorithms that rely on data are objective"

"But software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result … algorithms can reinforce human prejudices"

Article at The New York Times written by Claire Cain Miller (2015)

https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html

Bringing together a growing community of researchers and practitioners concerned with fairness, accountability, and transparency in machine learning

The past few years have seen growing recognition that machine learning raises novel challenges for ensuring non-discrimination, due process, and understandability in decision-making. In particular, policymakers, regulators, and advocates have expressed fears about the potentially discriminatory impact of machine learning, with many calling for further technical research into the dangers of inadvertently encoding bias into automated decisions.

At the same time, there is increasing alarm that the complexity of machine learning may reduce the justification for consequential decisions to "the algorithm made me do it."

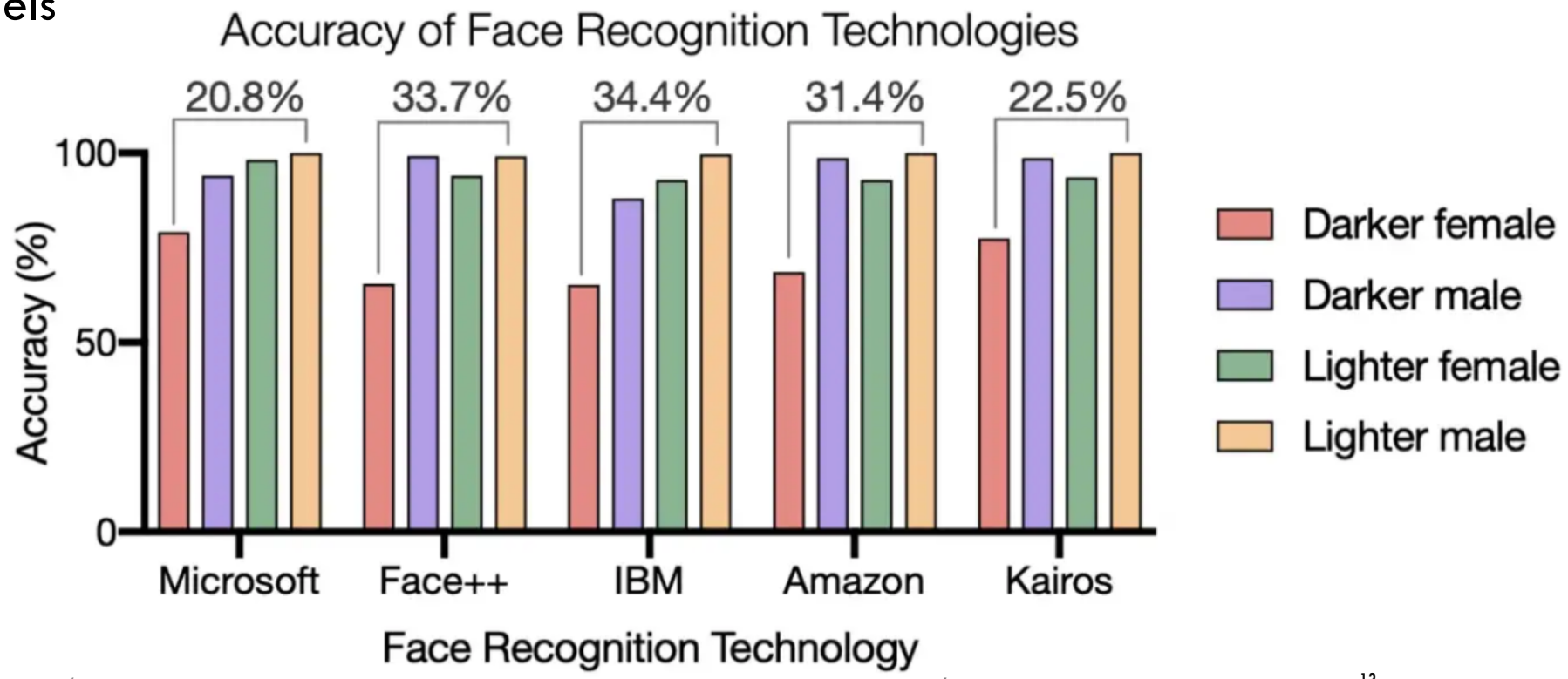
The annual event provides researchers with a venue to explore how to characterize and address these issues with computationally rigorous methods.

https://www.fatml.org/

EXAMPLES OF BIAS IN COMPUTER VISION

In 2018, the **Gender Shades project** studied the accuracy of facial analysis models.

It realized the gender classifier has the lowest performance for darker skin females in many ML (machine learning) models



https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

EXAMPLES OF BIAS IN COMPUTER VISION

An evaluation of four gender classifiers revealed a significant gap exists when comparing gender classification accuracies of:

- females vs males (9–20%) and
- darker skin vs lighter skin (10–21%).

Gender Classifier	Female Subjects Accuracy	Male Subjects Accuracy	Error Rate Diff.
Microsoft	89.3%	97.4%	8.1%
FACE++	78.7%	99.3%	20.6%
IBM	79.7%	94.4%	14.7%

https://www.media.mit.edu/publications/full-gender-shades-thesis-17/

EXAMPLES OF BIAS IN NLP

Word Embeddings

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{computer programmer}} - \overrightarrow{\text{homemaker}}$$

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

https://arxiv.org/pdf/1607.06520.pdf

EXAMPLES OF BIAS IN NLP

New NLP technologies

BERT is a language representation model with impressive accuracy for neural machine translation tasks. If you understand better what people ask, you give us more clues about what you are about to ask. Google says 15% of the Google queries have the correct answer, but it is not clear how accurate it is. The real issues is not on what we do, but on what we don't do. Instead, it is how we define our semantic structure, and it is not clear how we model it. Previously, Google search was mapped to a fixed-size list, but it is not clear whether this is the most accurate representation. But this is far from the truth: in real terms. That is why Google switches over to a semantic map, instead of a fixed size list, and that makes sense because semantic maps can help a lot.

EXAMPLES OF HISTORICAL BIAS

The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer *et al*. find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients.

The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. **Bias occurs because the algorithm uses health costs as a proxy for health needs**. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients. Reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.

https://www.science.org/doi/full/10.1126/science.aax2342

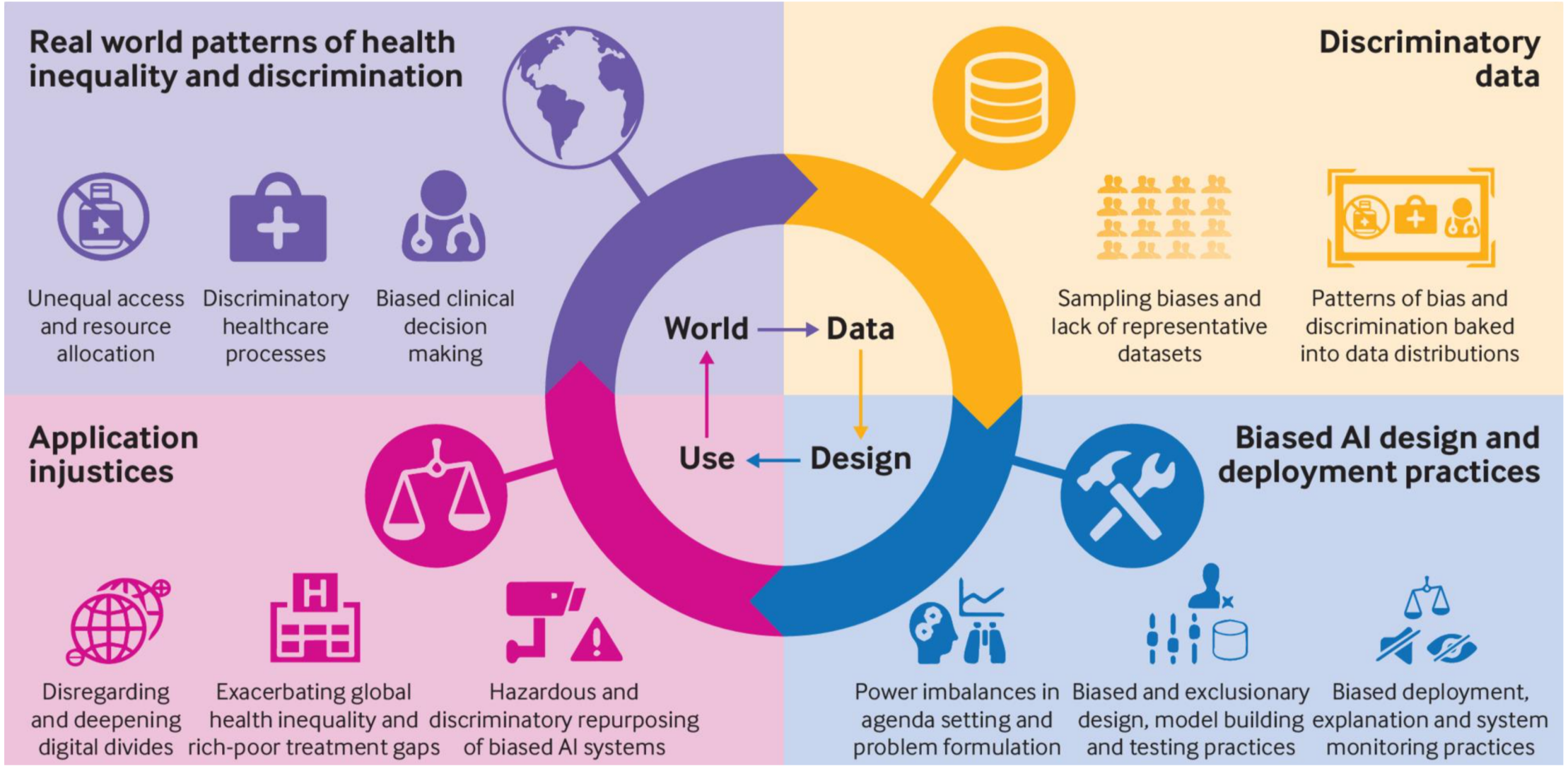
EXAMPLES OF HISTORICAL BIAS

Amazon used a machine learning model to rate its candidates for software jobs. In 2015, it realized **the model was not gender-neutral**.

The model was trained with resumes over the last 10-year. Since male engineers dominated the field, the model was learned to reward resumes for words common in male resumes and discriminate words common in female resumes.

In short, the model taught itself that male candidates were preferable. The model was oversimplified and overfitted with discriminatory data that could not be generalized. It was not complex enough to capture information from the minority group. Instead, it took the easy guess and associates "male-like" resumes to be more successful. Amazon scrapped the project later.

IS TRAINING DATA DISCRIMINATORY IN THE DEVELOPMENT PROCESS?



https://www.bmj.com/content/372/bmj.n304

EXAMPLE OF BIAS IN MACHINE LEARNING

COMPAS is a decision support tool used by some U.S. justice systems. It becomes a textbook case study in fairness.

Judges and parole officers use the system to score criminal defendants' likelihood of reoffending if released. It provides suggestions on sentencing, parole, and bail.

- **Black defendants** were often predicted to be at a higher risk of recidivism than they actually were. The analysis found that black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts (45% vs 23%)
- **White defendants** were often predicted to be less risky than they were. The analysis found that White defendants who re-offended within the next two years were mistakenly labelled low risk almost twice as often as black re-offenders (48% vs. 28%)

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

RELATED TOPICS

RELATED TOPICS

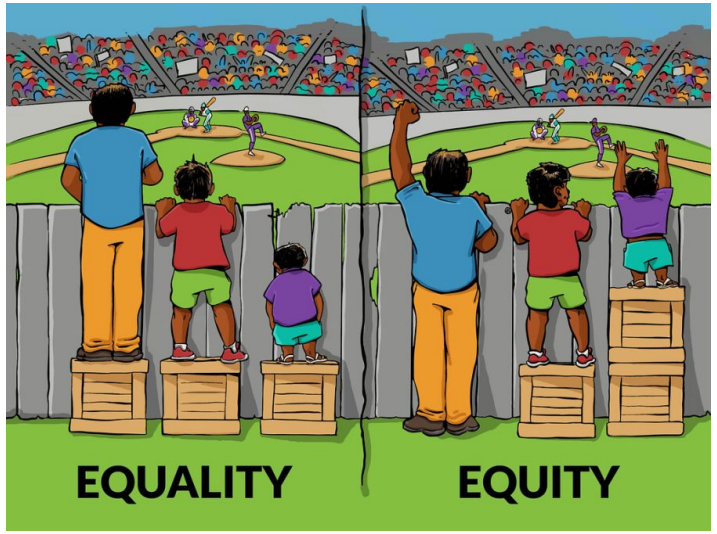
Ethics seeks to answer questions like “what is good or bad”, “what is right or what is wrong”, or “what is justice, well-being or equality”.

As a discipline, ethics involves systematizing, defending, and recommending concepts of right and wrong conduct by using conceptual analysis, thought experiments, and argumentation.

If you want to know more about philosophical reasoning, see this video https://youtu.be/NKEhdsnKKHs by Crash Course Philosophy.)

https://ethics-of-ai.mooc.fi/chapter-1/2-what-is-ai-ethics

RELATED TOPICS

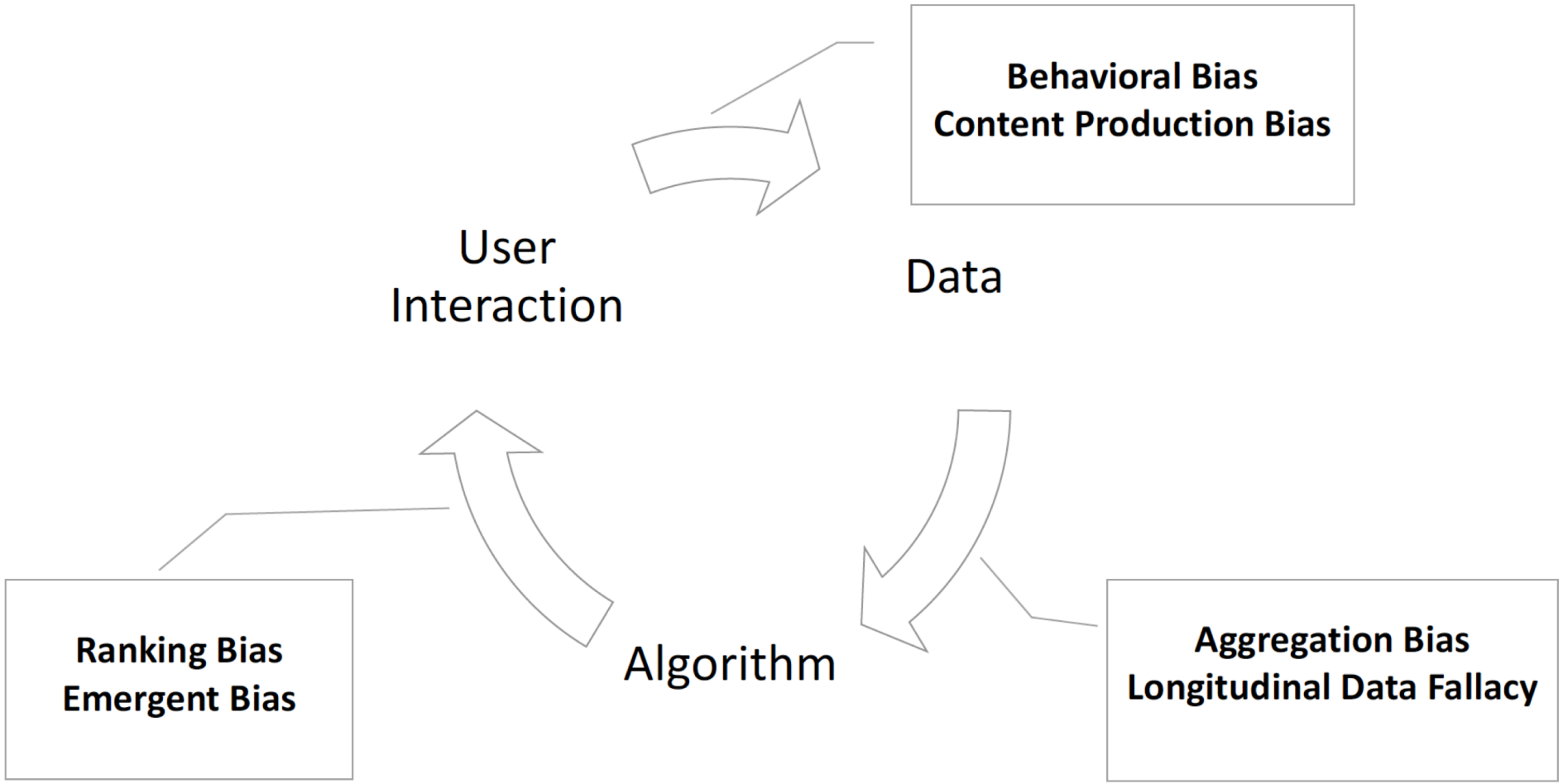


Equality and Equity are not synonyms !!!

Equality and equity are important in decision-making, but can produce very different results.

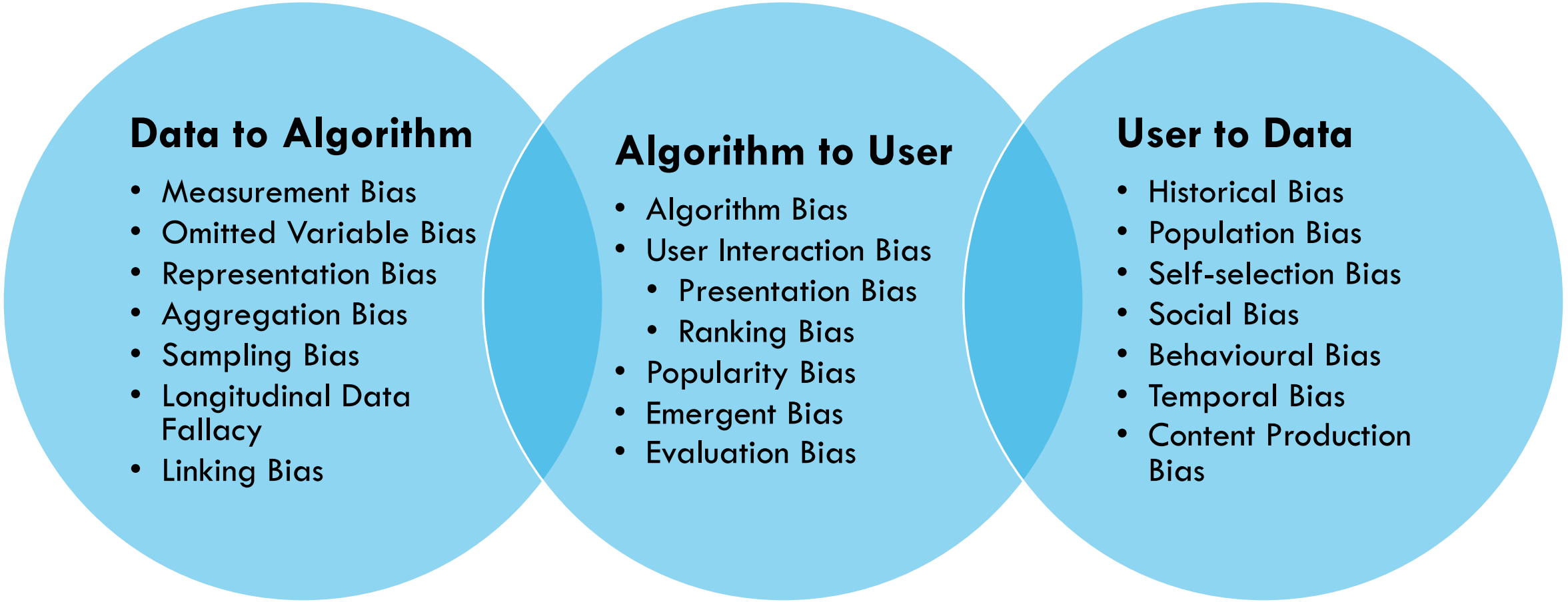
- **Equality**, the same opportunity may be provided to everyone
  - But equality is promoted at the expense of *fairness*
- **Equity**, levels the playing field so that all groups have opportunities to succeed

BIAS IN DATA, ALGORITHMS, AND USER EXPERIENCES



Source: A survey on Bias and Fairness in Machine Learning. Mehrabi et al. 2022. https://arxiv.org/abs/1908.09635

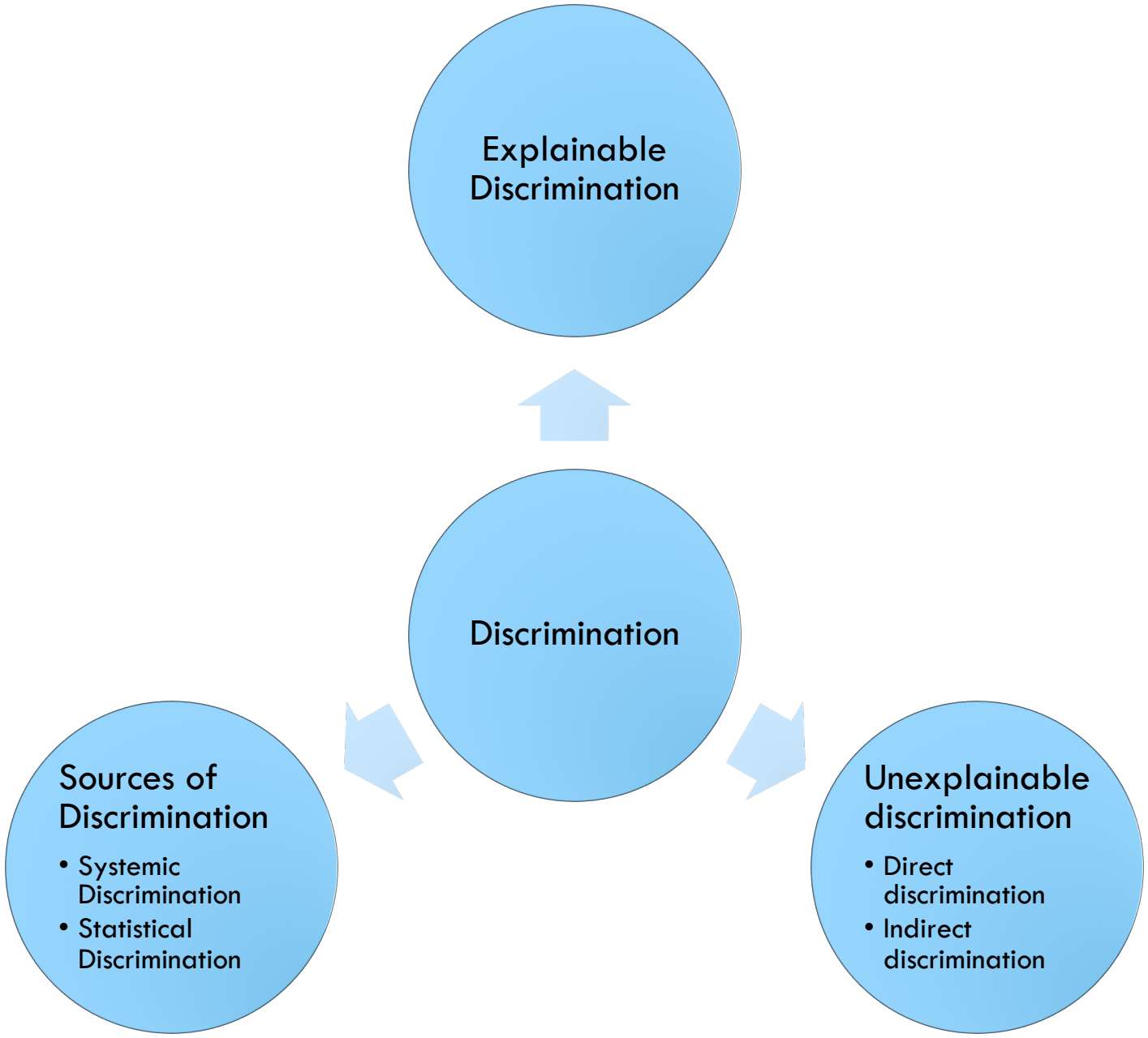
TYPES OF BIAS



DISCRIMINATION

Discrimination can be considered as a source for unfairness that is due to human prejudice and stereotyping based on the sensitive attributes, which may happen intentionally or unintentionally,

while **bias** can be considered as a source for unfairness that is due to the data collection, sampling, and measurement



WHEN MIGHT AI BE UNFAIR OR DISCRIMINATORY?

- Inaccurate or insufficient data
- Variables influence outcomes differently by group (and are under-represented in model development)
- Not fully causally related variables that are correlated to the protected class
  - Typically includes race, ethnicity, religion, sex, disability, familial status, national origin, or age
- Less discriminatory models are available
  - Types of discrimination:
    - Overt Discrimination
    - Disparate treatment
    - Adverse impact or Disparate impact — Demographic Parity

WHAT DOES IT MEANS FOR AI TO BE FAIR?

- Imperfect statistical models (and decisions, generally) are inherently inequitable, but are not necessarily systematically unfair or discriminatory
- Conceptually, Fairness is "The quality of treating people equally or in a way that is reasonable" (*Oxford English Dictionary*)
- There are numerous mathematical definitions of fairness — some of them are contradictory
  - Anti-classification
  - Classification parity
  - Calibration
  - "21 Definitions of Fairness" look at https://youtu.be/jIXIuYdnyyk

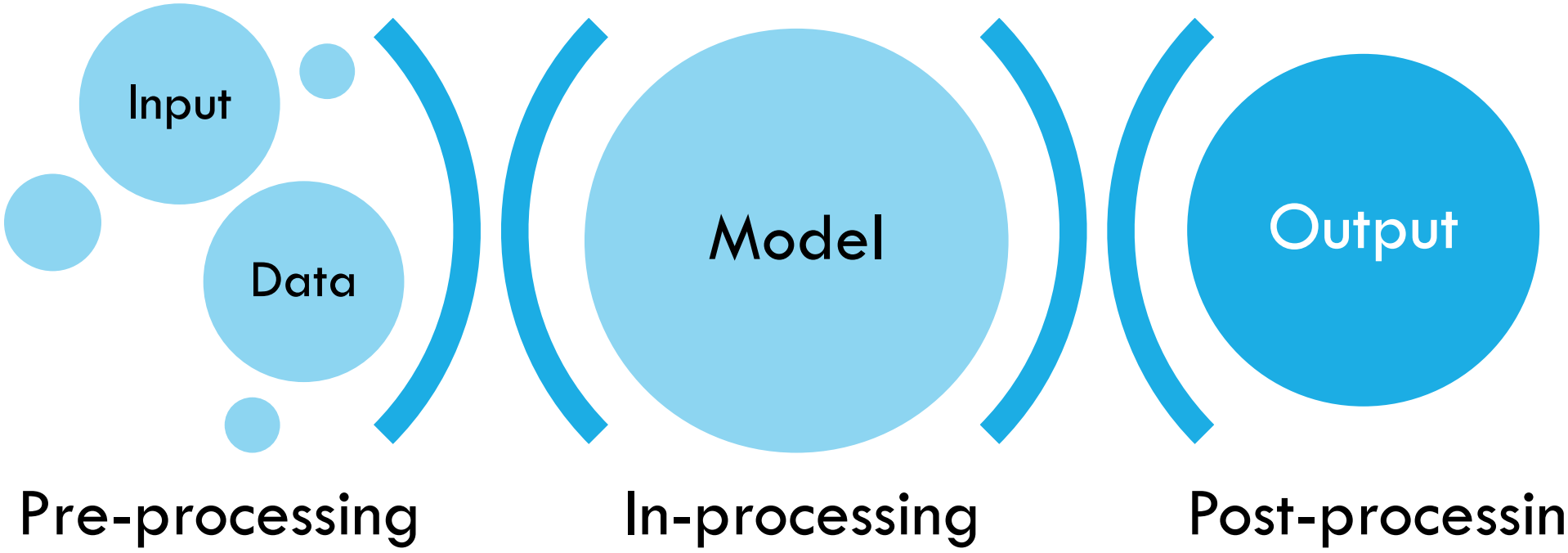
READ: The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. https://arxiv.org/pdf/1808.00023.pdf

WHAT DOES IT MEANS FOR AI TO BE FAIR?

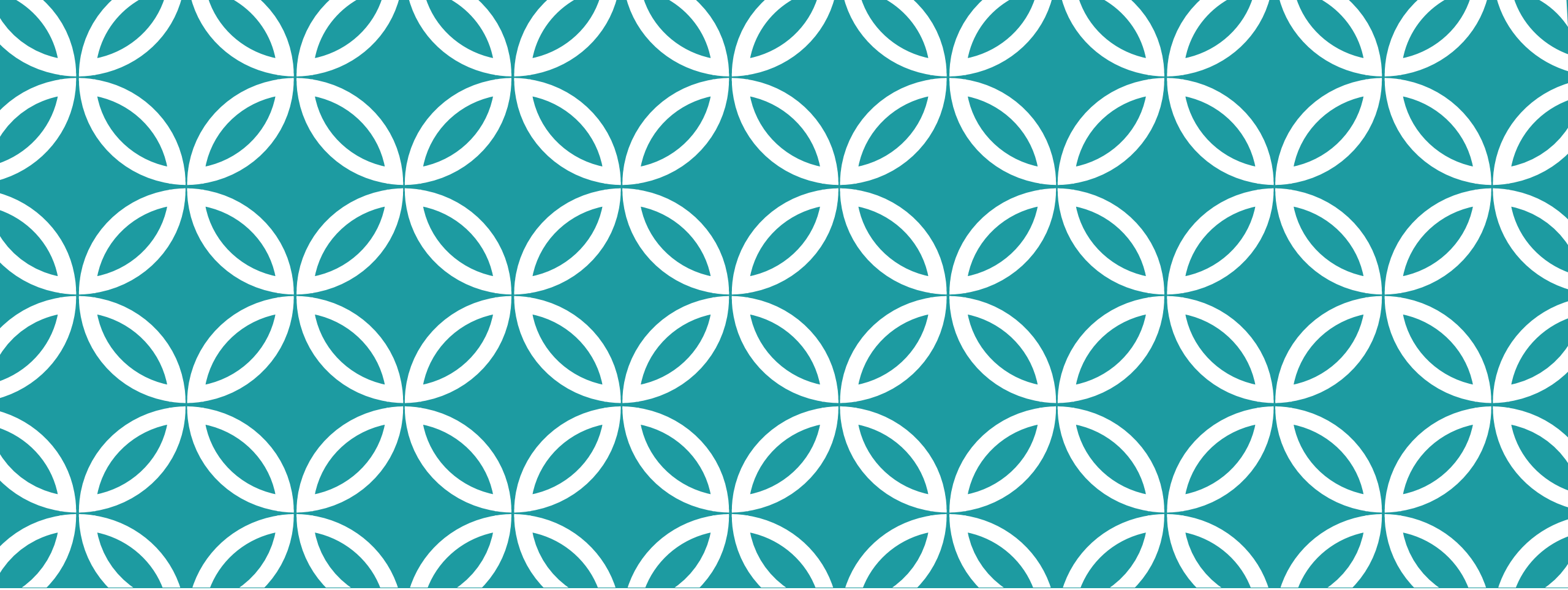
- Classes of fairness definitions to provide equitable risk assessment
  - Anti-classification, stipulates that risk assessment algorithms not consider protected characteristics—like race, gender
  - Classification parity, requires that certain common measures of predictive performance be equal across groups defined by the protected attributes
  - Calibration, requires that outcomes are independent of protected attributes after controlling for estimated risk

READ: The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. https://arxiv.org/pdf/1808.00023.pdf

METHODS FOR FAIR MACHINE LEARNING

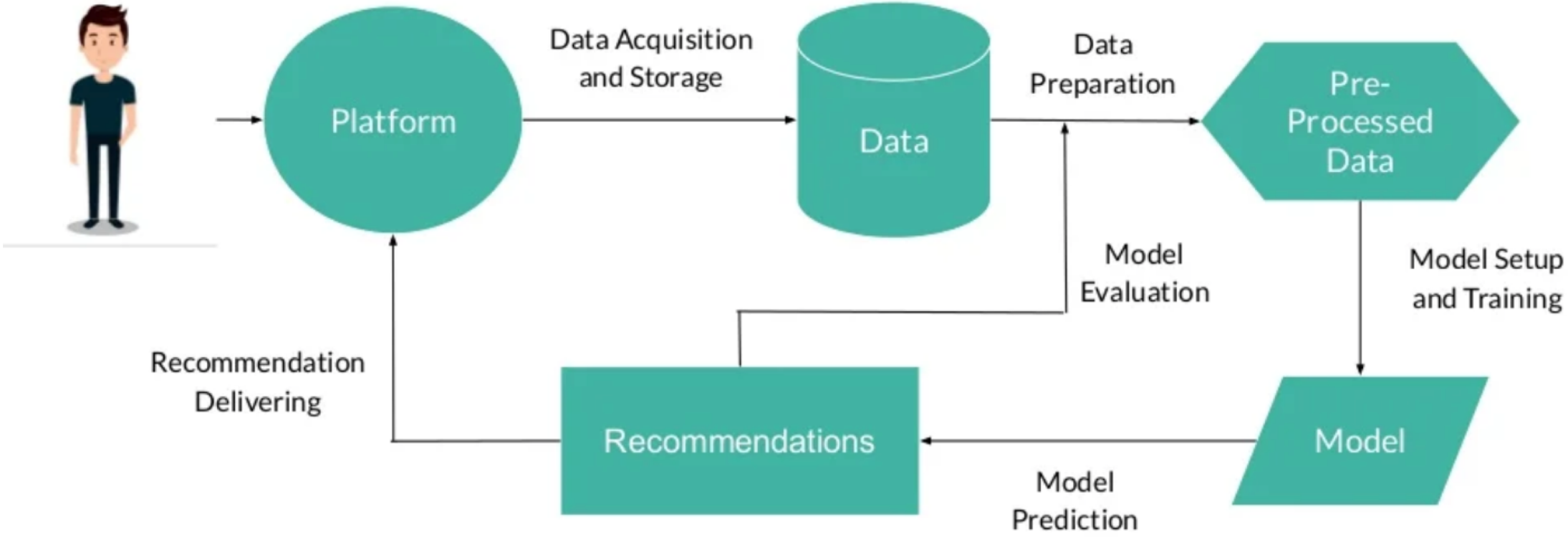


READ: A survey on Bias and Fairness in Machine Learning. Mehrabi et al. 2022. https://arxiv.org/abs/1908.09635



BRIEF INTRODUCTION TO BIAS AND FAIRNESS IN RECOMMENDER SYSTEMS

RECOMMENDATION PIPELINE



Source: https://www.slideshare.net/MirkoMarras/tutorial-on-bias-in-rec-sys-umap2020-238273774
UMAP2020 Tutorial: Hands on Data and Algorithmic Bias in Recommender Systems

CORE STAKEHOLDERS IN RECOMMENDATION

A recommendation stakeholder is any group or individual that can affect, or is affected by, the delivery of recommendations to users

Consumers

Providers

System

A SAMPLE MULTI-SIDED SCENARIO

Consumers

Students

Providers

Teachers

Udemy

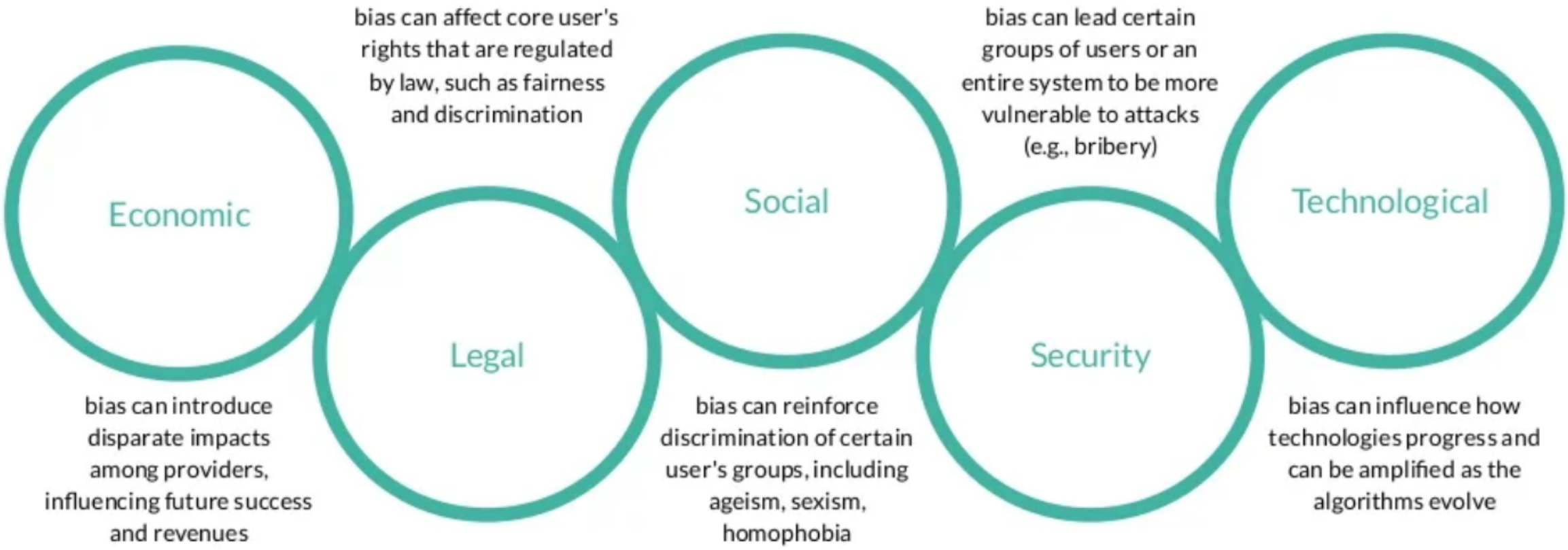
coursera

edX

System

Online Course Platform

PERSPECTIVES IMPACTED BY BIAS



ETHICAL ASPECTS INFLUENCED BY BIAS

Content

recommendation of inappropriate content

Privacy

unauthorised data collection, data leaks, unauthorised inferences

Autonomy and Identity

behavioural traps and encroachment on sense of personal autonomy

Opacity

black-box algorithms, uninformative explanations, feedback effects

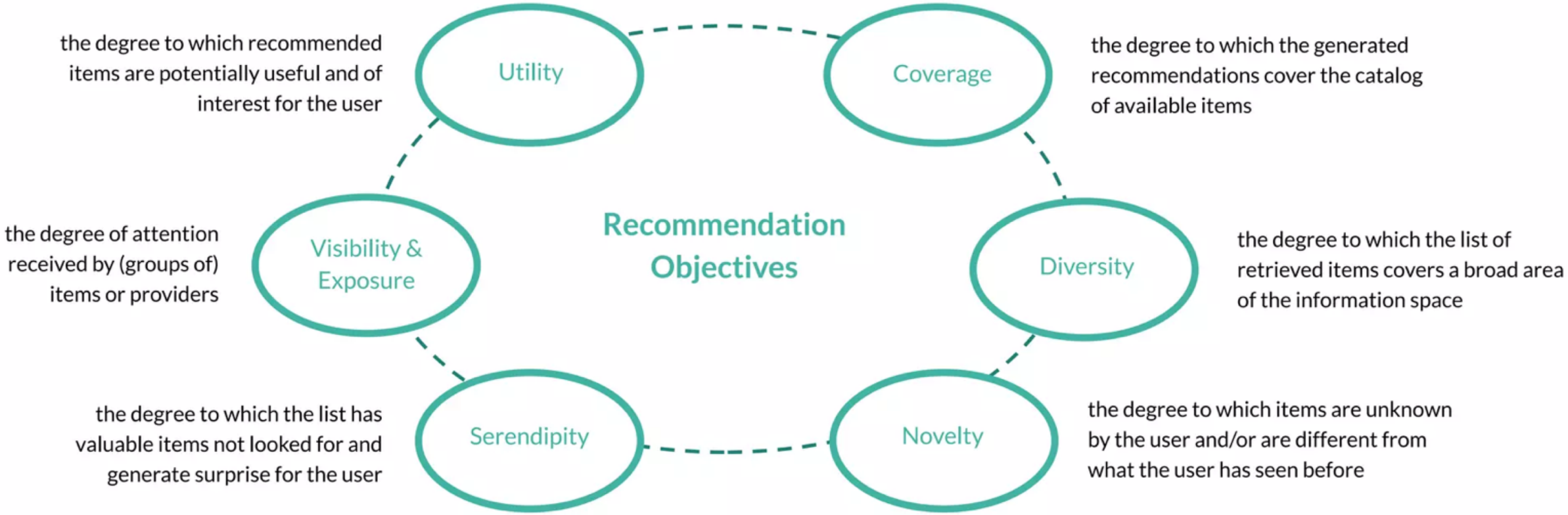
Fairness

observation bias, population imbalance

Social

lack of exposure to contrasting viewpoints, feedback effects

OBJECTIVES INFLUENCED BY BIAS



BIAS IN DATA RECOMMENDATION

- **Selection bias** happens as users are free to choose which items to rate, so that the observed ratings are not a representative sample of all ratings. In other words, the rating data is often missing not at random (MNAR)
- **Conformity bias** happens as users tend to rate similarly to the others in a group, even if doing so goes against their own judgment, making the rating values do not always signify user true preference
- **Exposure bias** happens as users are only exposed to a part of specific items so that unobserved interactions do not always represent negative preference
- **Position bias** happens as users tend to interact with items in higher position of the recommendation list regardless of the items' actual relevance so that the interacted items might not be highly relevant

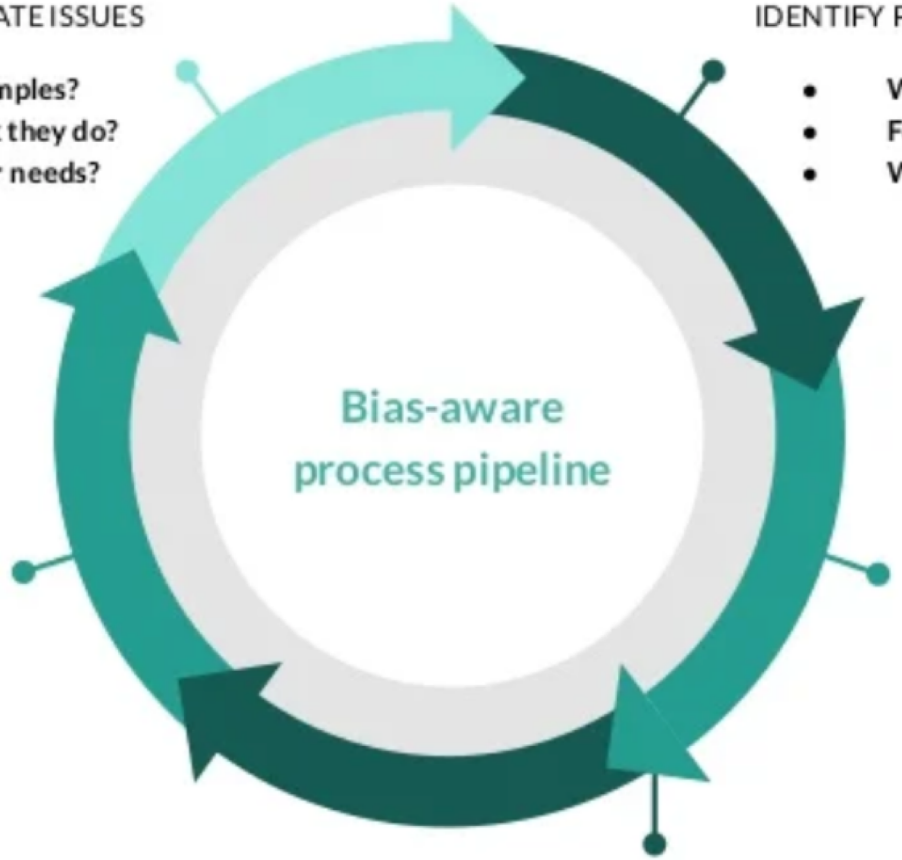
BIAS IN MODEL RECOMMENDATION

- **Inductive bias** denotes the assumptions made by the model to better learn the target function and to generalize beyond training data
- **Popularity bias**. Popular items are recommended even more frequently than their popularity would warrant
- **Unfairness**. The system systematically and unfairly discriminates against certain individual or groups of individuals in favour others

BIAS ASSOCIATED TO USERS

- **Population biases**, differences in demographics or other user characteristics, between a population of users represented in a dataset/platform and a target population
- **Behavioural biases**, differences in user behaviour across platforms or contexts, or across users represented in different datasets
- **Content biases**, behavioural biases that are expressed as lexical, syntactic, semantic, and structural differences in the contents generated by users
- **Linking biases**, behavioural biases that are expressed as differences in the attributes of networks obtained from user connections, interactions or activity
- **Temporal biases**, differences in populations or behaviours over time

BIAS-AWARE PROCESS PIPELINE



MITIGATE ISSUES

- Does data include enough minority samples?
- Do our proxies measure what we think they do?
- Does the bias notion capture customer needs?

IDENTIFY PRODUCT GOALS

- What are you trying to achieve?
- For what population of people?
- What metrics are you tacking?

DEVELOP AND ANALYZE THE SYSTEM

- How well the system matches product goals?
- To what degree bias is still present?
- How decisions impact on each stakeholder?

IDENTIFY STAKEHOLDERS

- Who has a stake in this product?
- Who might be harmed?
- How?

DEFINE A BIAS NOTION

- What type of bias? At what point?
- What distributions?

TECHNIQUES FOR BIAS TREATMENT

Pre-processing

before model training

Pre-processing techniques try to **transform the data** so that the bias is mitigated. If the algorithm is allowed to modify the training data, pre-processing can be used

In-processing

during model training

In-processing techniques try to modify **learning algorithms** to mitigate bias during training process. If it is allowed to change the learning procedure, in-processing can be used

Post-processing

after model training

Post-processing is performed by **re-ranking items** of the lists obtained after model training. If the algorithm can treat the learned model as a black box, post-processing can be used

FAIRNESS VARIES ACCORDING TO THE STAKEHOLDERS

Consumer Fairness

We talk about unfairness for consumers when their experience in the platform differs:

- in terms of service effectiveness (results' relevance, user satisfaction)
- resulting outcomes (exposure to lower-paying job offers)
- participation costs (differential privacy risks)

Provider Fairness

Providers experience unfairness when a platform/service creates:

- different opportunities for their items to be consumed
- different visibility or exposure in the ranking
- different participation costs

Consumer-Provider Fairness

It might be needed that the platform guarantees fairness for both consumers and providers, e.g.,:

- people matching
- property/business recommendation
- user skills and job matching
- and so on...

GRANULARITY OF DISCRIMINATION

when a system gives unfairly different predictions to individuals who are considered similar for that task

when a system systematically treats individuals who belong to different groups unfairly

when a system systematically discriminate individuals over a large collection of subgroups

For much more, see “Fairness & Discrimination in Retrieval & Recommendation” tutorial at https://fair-ia.ekstrandom.net/sigir2019

UNAP
UNIVERSIDAD ARTURO PRAT
DEL ESTADO DE CHILE

UNIVERSITAT DE BARCELONA

EXAMPLE 1: BIAS IN CALIBRATED RECOMMENDATIONS

International Workshop on Advances in Bias and Fairness in Information Retrieval (2022)

Carlos Rojas, David Contreras, {carlrojasc,david.contreras}@unap.cl

Universidad Arturo Prat, Avenida Arturo Prat, 2120, Iquique, Chile

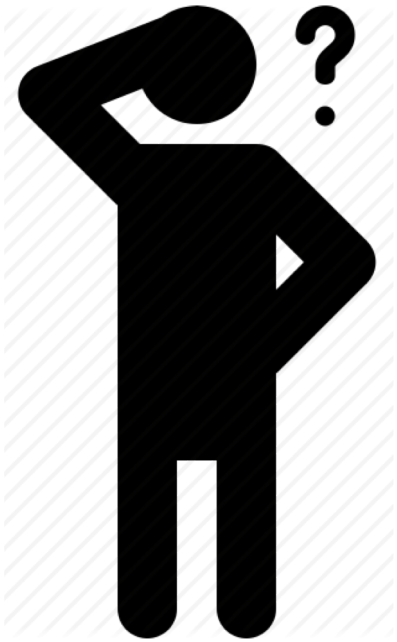
Maria Salamó, maria.salamo@ub.edu

Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Gran Via de les Corts Catalanes, 585, Barcelona, Spain

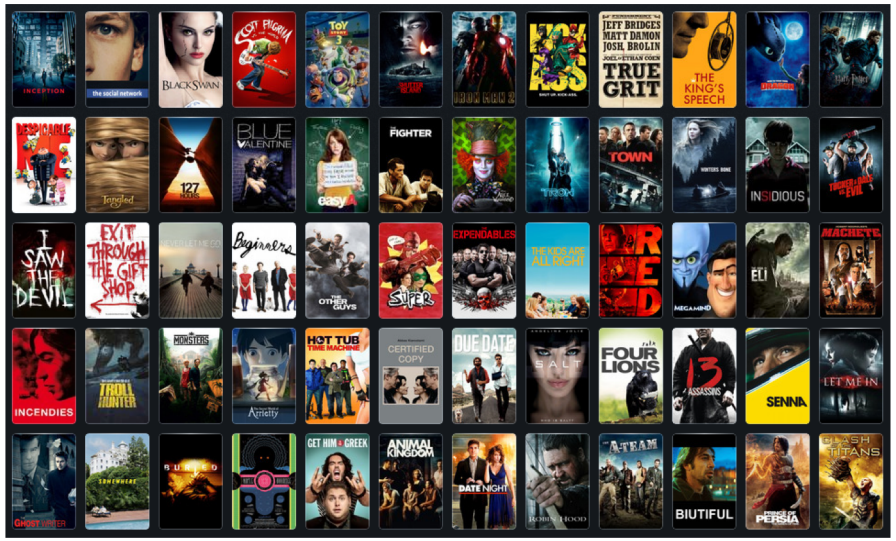
Institute of Complex Systems (UBICS), Universitat de Barcelona, Barcelona, Spain

INTRODUCTION

RECOMMENDER SYSTEMS



What movie do I watch?

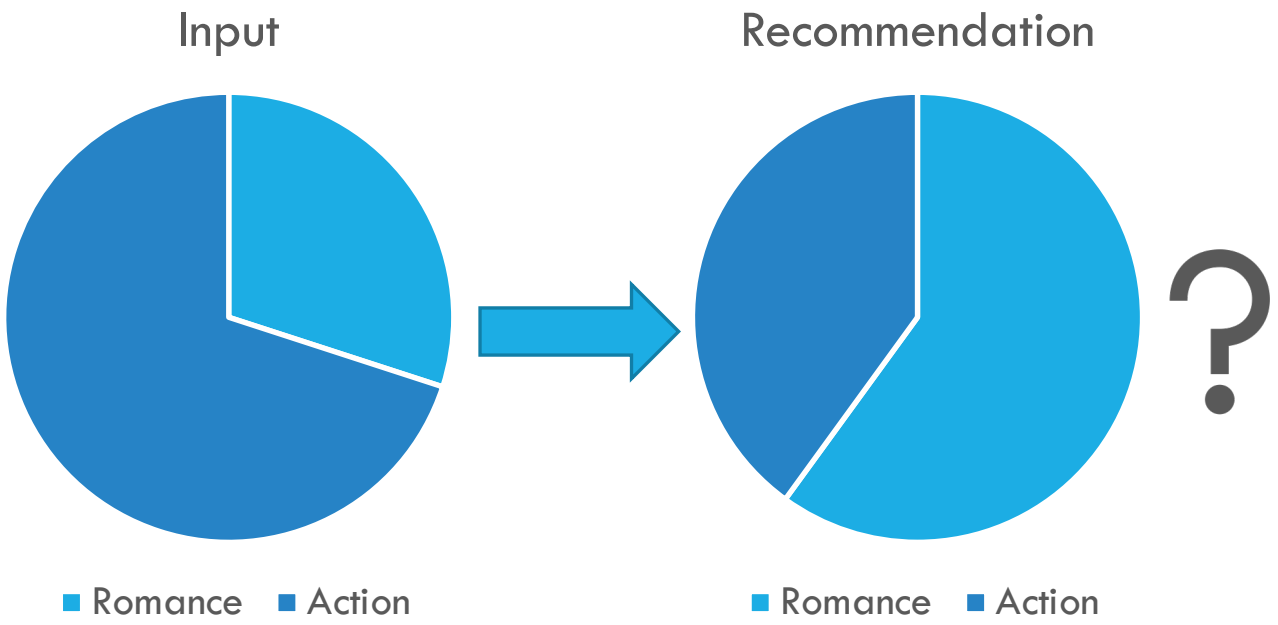


INTRODUCTION

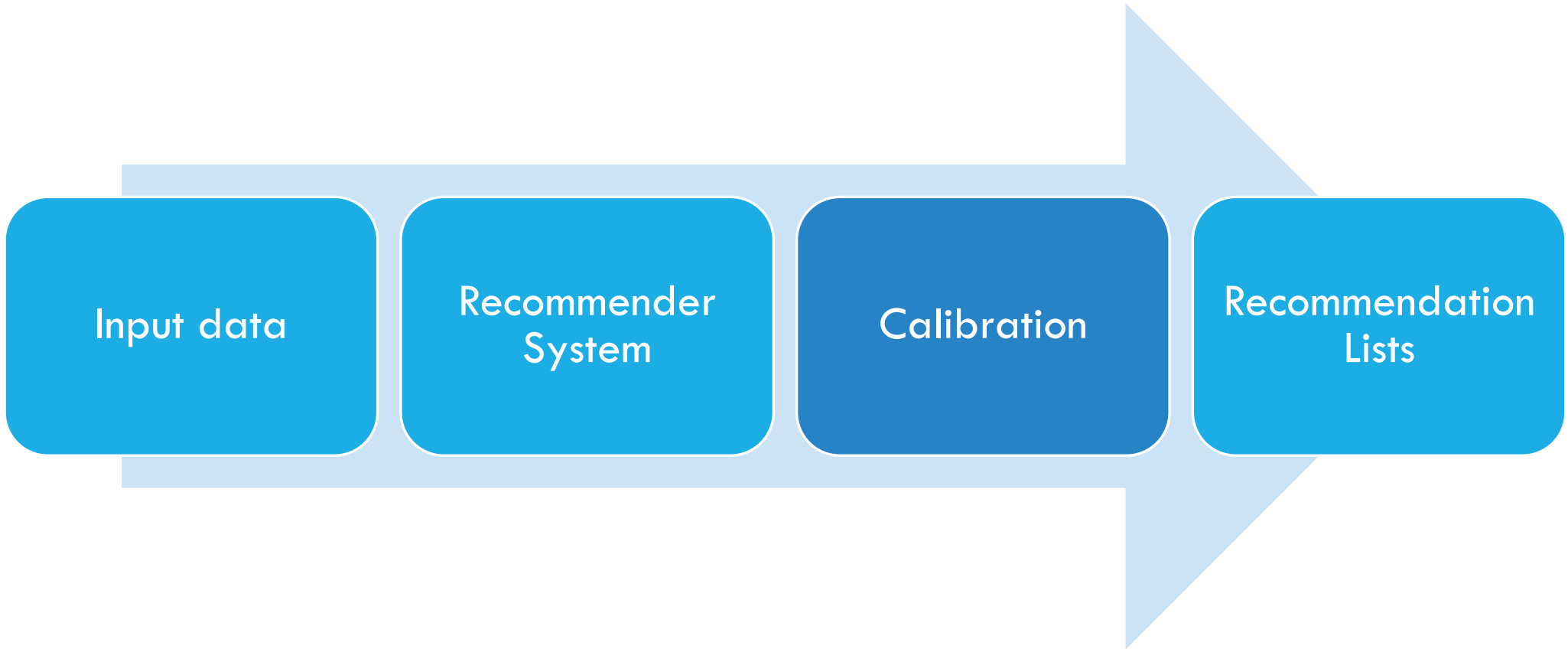
BIAS PROBLEM

Imbalances in the input data can lead to biases in the results of recommendation models.

Concretely, in most cases, the recommendations produced by a model differ from users' history.



INTRODUCTION
CALIBRATION



RESEARCH QUESTIONS

RQ1. How does a calibration algorithm deal with the bias associated with the recommended items' distribution, when considering several recommendation models?

RQ2. Can the calibration algorithm impact the effectiveness of different recommendation models?

PRELIMINARIES

GENERAL PROCESS

MovieLens 1M Dataset

- 6040 users
- 3600 movies
- Complimentary genre dataset (such as Comedy, Horror, among others)

Preprocessing

- Focus on implicit data

Elliot Framework (1)

- Use of k-nearest neighbors and latent factor models
- 3456 recommendations generated

Steck Calibration (2)

$$C_{KL}(p,q) = KL(p||\tilde{q}) = \sum_g p(g|u) \log \frac{p(g|u)}{\tilde{q}(g|u)} \qquad \tilde{q}(g|u) = (1-\alpha) \cdot q(g|u) + \alpha \cdot p(g|u)$$

(1) Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation, Anelli et al., July 2021

(2) Calibrated Recommendations, Steck H., September 2018

EXPERIMENTS

Recommender models

- Latent Factors: MF, BPRMF, WRMF, SVDpp
- Neighbor based: UserKNN, ItemKNN

We aggregated the results regarding the 10% sub-population of test users who received recommendation with the worst calibration metric

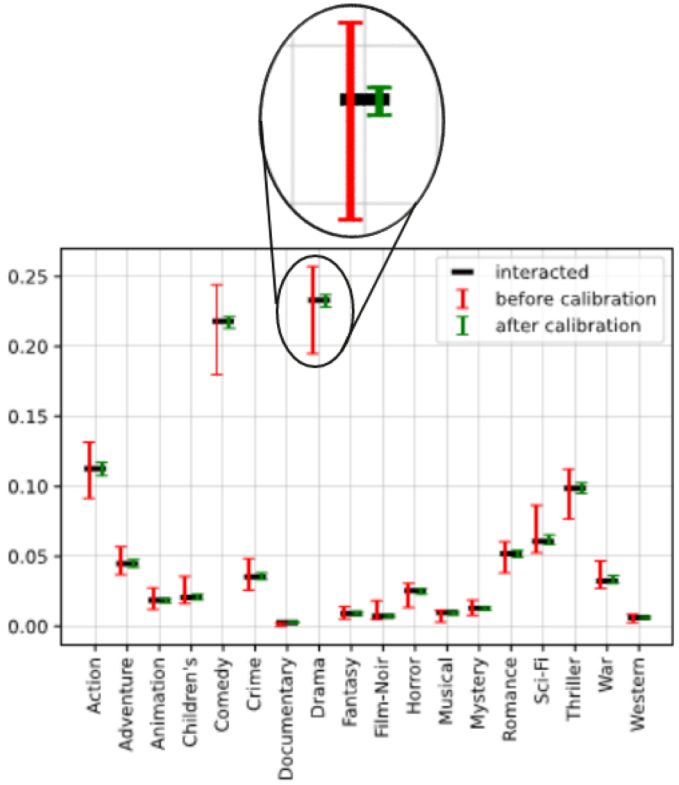
- Averaged positive and negative differences between genre distribution in recommendations and interactions

Is important to note…

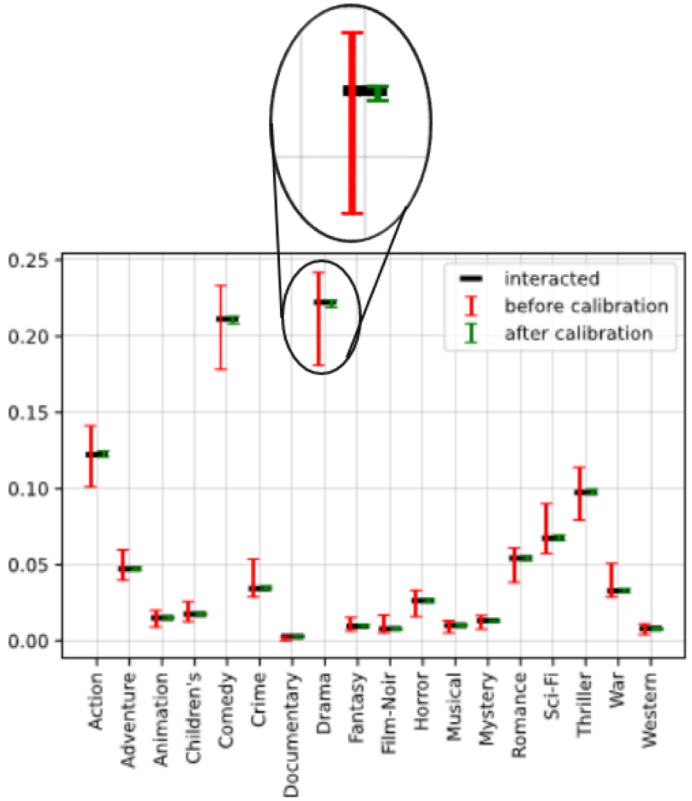
- The 10% sub population of test users is defined by calibration metric
- This is calculated using interactions and recommendation distributions
- The latter is dependent on the results of each recommendation model
- Subset is different for each model, in consequence, average too

RESULTS

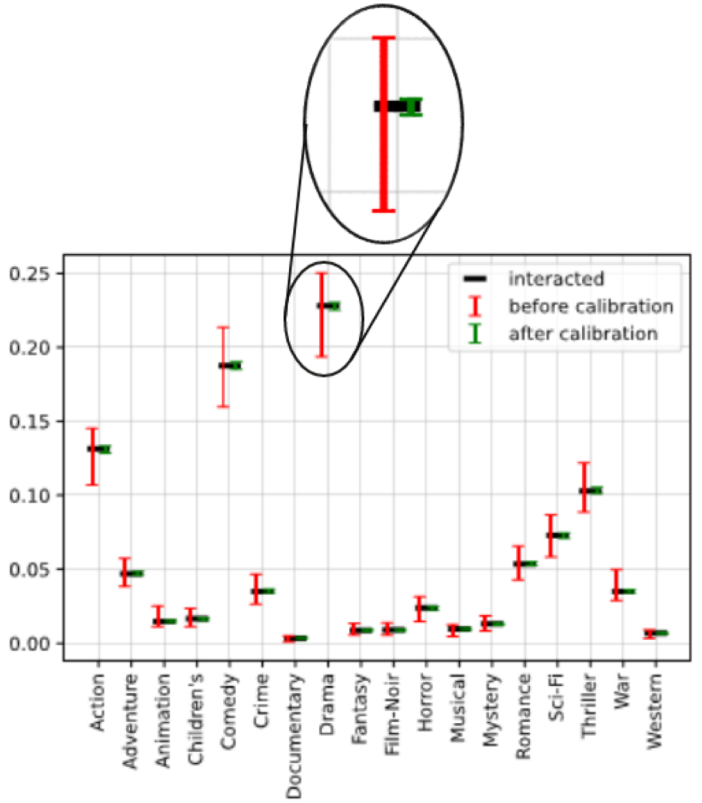
ANALYSIS OF BIAS IN THE CALIBRATION ALGORITHM



BPRMF



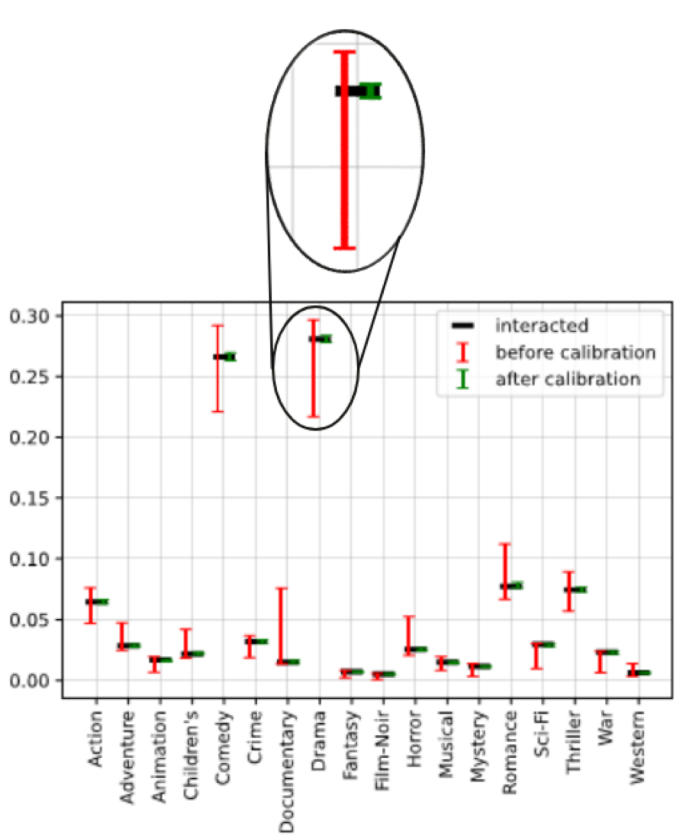
WRMF



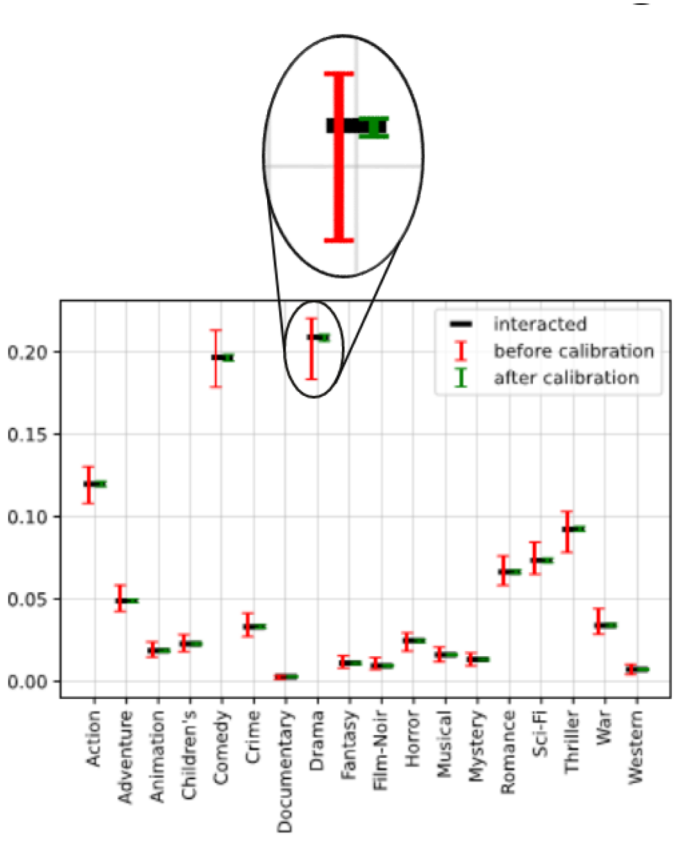
MF

RESULTS

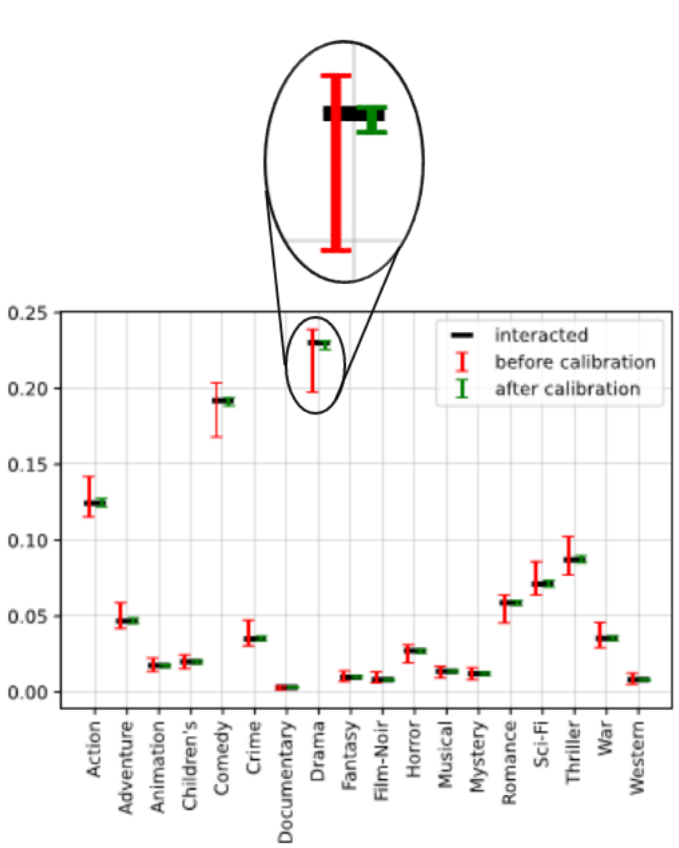
ANALYSIS OF BIAS IN THE CALIBRATION ALGORITHM



SVDpp



UserKNN

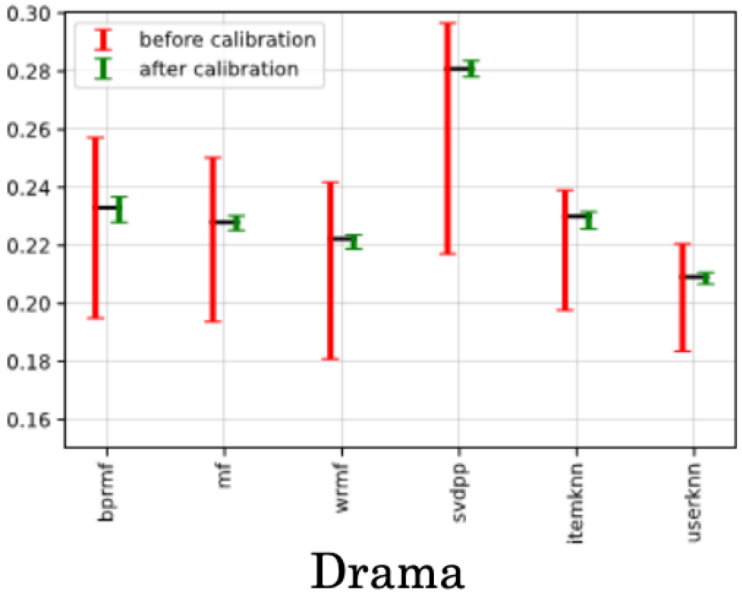


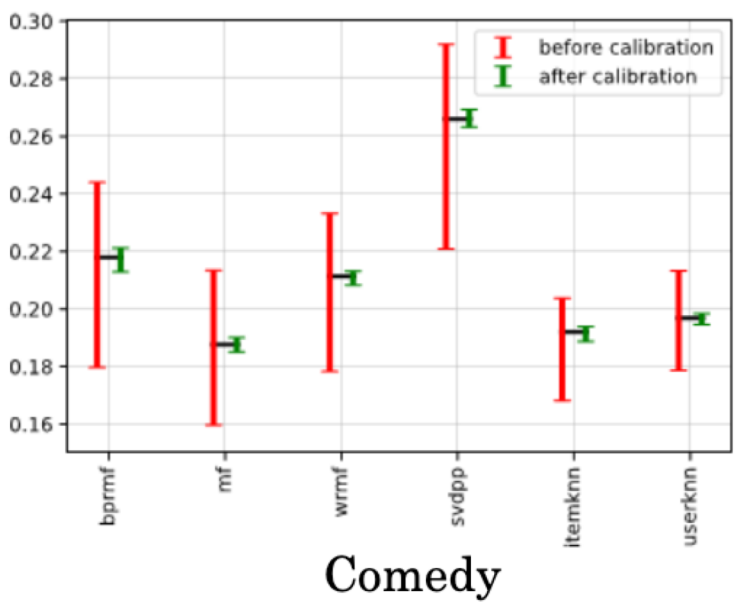
ItemKNN

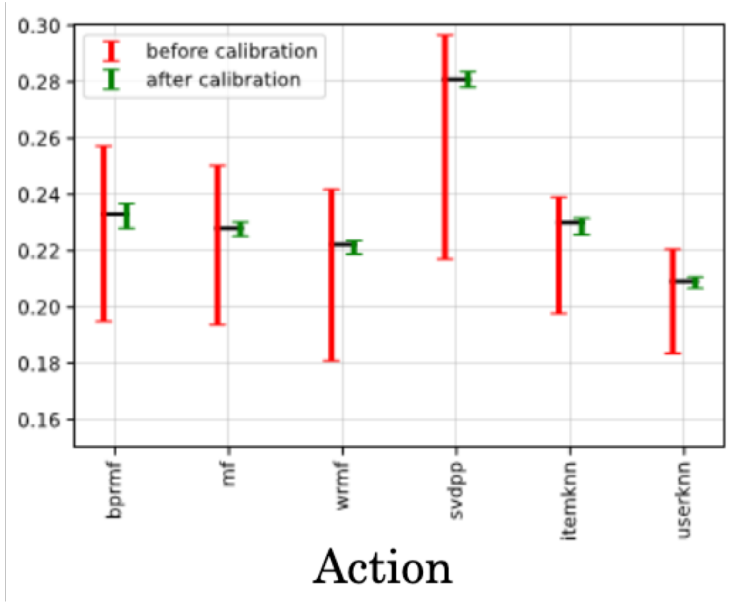
RESULTS

ANALYSIS OF BIAS IN THE CALIBRATION ALGORITHM

Most representative genres



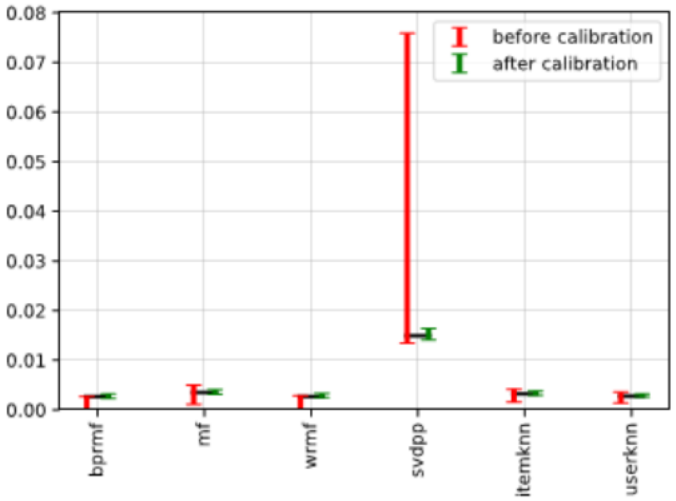




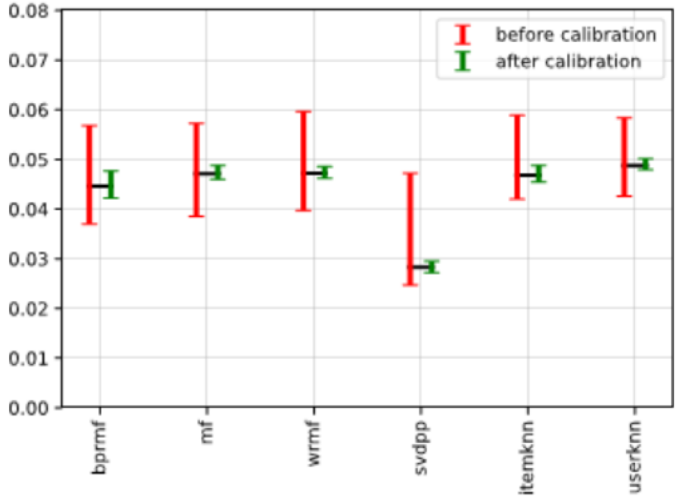
RESULTS

ANALYSIS OF BIAS IN THE CALIBRATION ALGORITHM

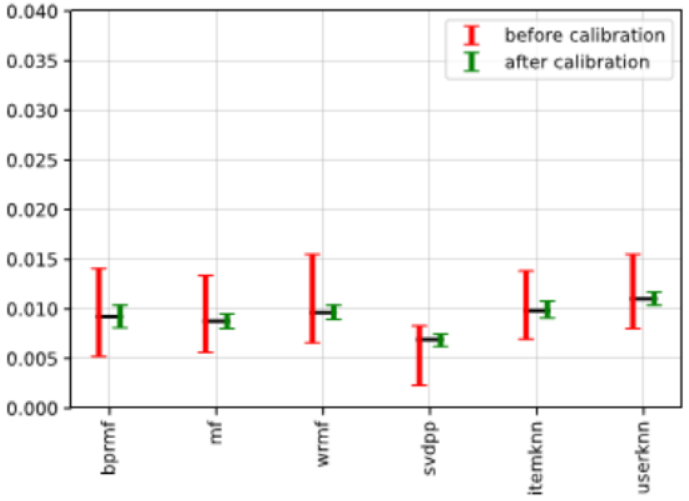
Least representative genres



Documentary



Adventure



Fantasy

The calibration algorithm still shows a slightly biased results that depends of recommendation models (**RQ1**)

RESULTS

ANALYSIS OF RECOMMENDATION ACCURACY IN CALIBRATION PROCESS

To evaluate how the calibration affects the recommendation accuracy, we computed the NDCG before and after the calibration for all models, alongside the difference and percentage increases or decreases.

The **decrease of accuracy is in the range of 5 to 20%**, meaning the calibration has a negative impact in the accuracy of the model.

Models	NDCG (before)	NDCG (after)	δ	Percentage
BPRMF	0.36585	0.32986	-0.03599	-9.84%
MF	0.34709	0.32641	-0.02068	-5.96%
WRMF	0.38617	0.30890	-0.07726	-20.01%
SVDpp	0.01276	0.06175	0.04898	79.33%
UserKNN	0.37995	0.30451	-0.07544	-19.86%
ItemKNN	0.37781	0.34120	-0.03661	-9.69%

The calibration has a negative impact in the accuracy of the recommendation model. **(RQ2)**

CONCLUSIONS AND FUTURE WORK

Results show that there is a bias in the calibration algorithm

Calibration impacts the recommendation accuracy

Future work will study the interplay between calibration and algorithmic fairness

EXAMPLE 2:
PROVIDER FAIRNESS ACROSS CONTINENTS IN COLLABORATIVE RECOMMENDER SYSTEMS

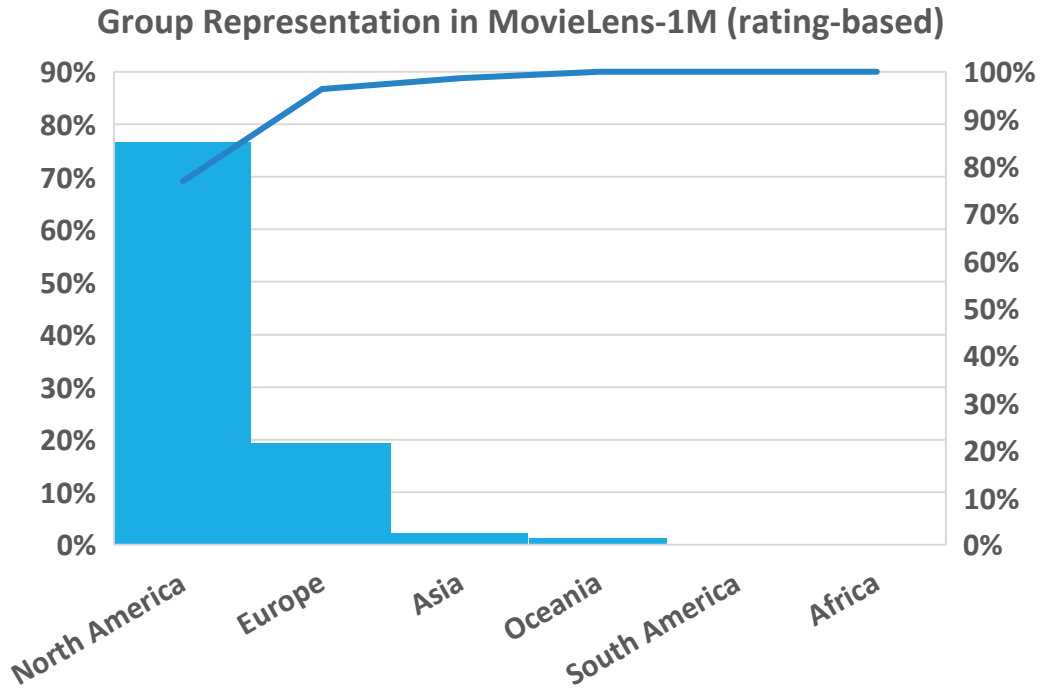
Information Processing & Management Journal (2022),
Women in RecSys Journal Paper of the Year Awards, junior category

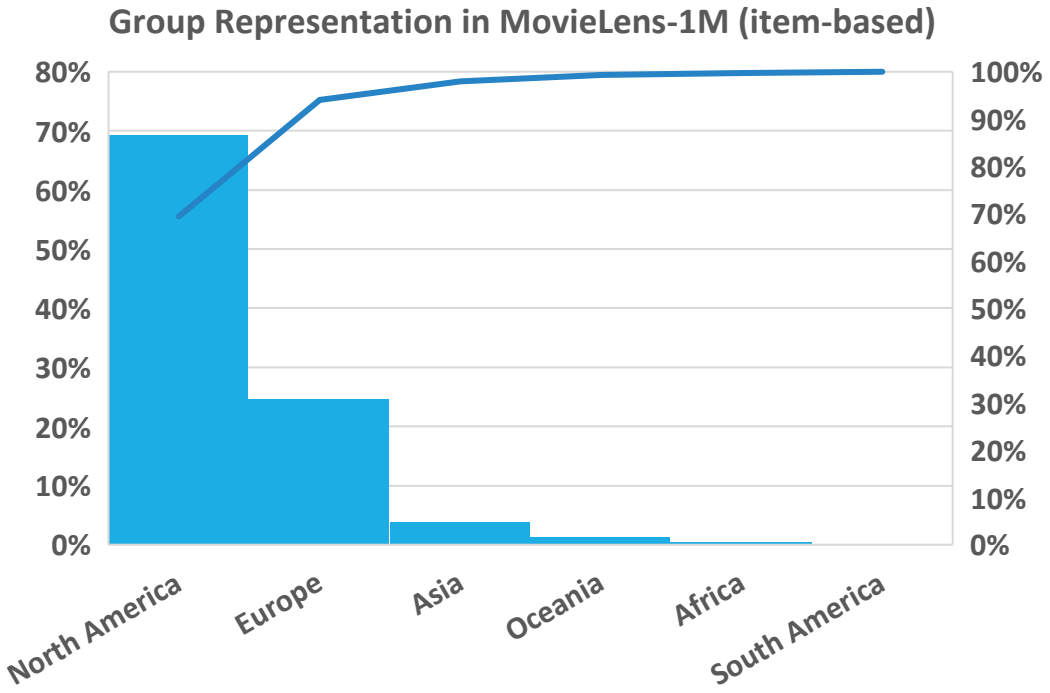
Elizabeth Gómez
Universitat de Barcelona, Spain
egomezye13@alumnes.ub.edu

Ludovico Boratto
University of Cagliari, Italy
ludovico.boratto@acm.org

Maria Salamó
Universitat de Barcelona, Spain
maria.salamo@ub.edu

PROBLEM: IMBALANCES IN THE DATA DISTRIBUTION





Natural imbalances in the data might be embedded in the patterns in recommender systems

Data imbalances might be inherently connected to the way an industry is composed

FOCUS: PROVIDER UNFAIRNESS ACROSS CONTINENTS

- We study group unfairness, based on the **geographic provenience of the providers**
- We divide items into several groups according to the geographic area, in a **multi-group setting**
- We study the **visibility and exposure given to providers** of the different groups when recommending items to users

Imbalanced Input data

Recommendations

OVERVIEW

- We assess **unfairness for groups of providers** belonging to different **geographic continents,** considering state-of-the-art recommendation models
- We propose a **re-ranking algorithm** to introduce provider fairness **for multiple groups**
  - It mitigates unfairness by providing equity of **visibility** and **exposure**
  - It introduces items that cause **minimum possible loss** in terms of effectiveness
- We evaluate our algorithm in two recommendation domains and study its effectiveness at producing **fair but effective recommendations**

ALGORITHM

```
1 define optimizeContinentsVisibilityExposure (recList, targetProportions)
2 begin
      // mitigation to target the desired visibility
3     reRankedList ← mitigationContinent(recList, "visibility", targetProportions);
      // mitigation to regulate the exposure
4     reRankedList ← mitigationContinent(reRankedList, "exposure", targetProportions);
5     return reRankedList ; // re-ranked list adjusted by visibility and exposure
6 end
```

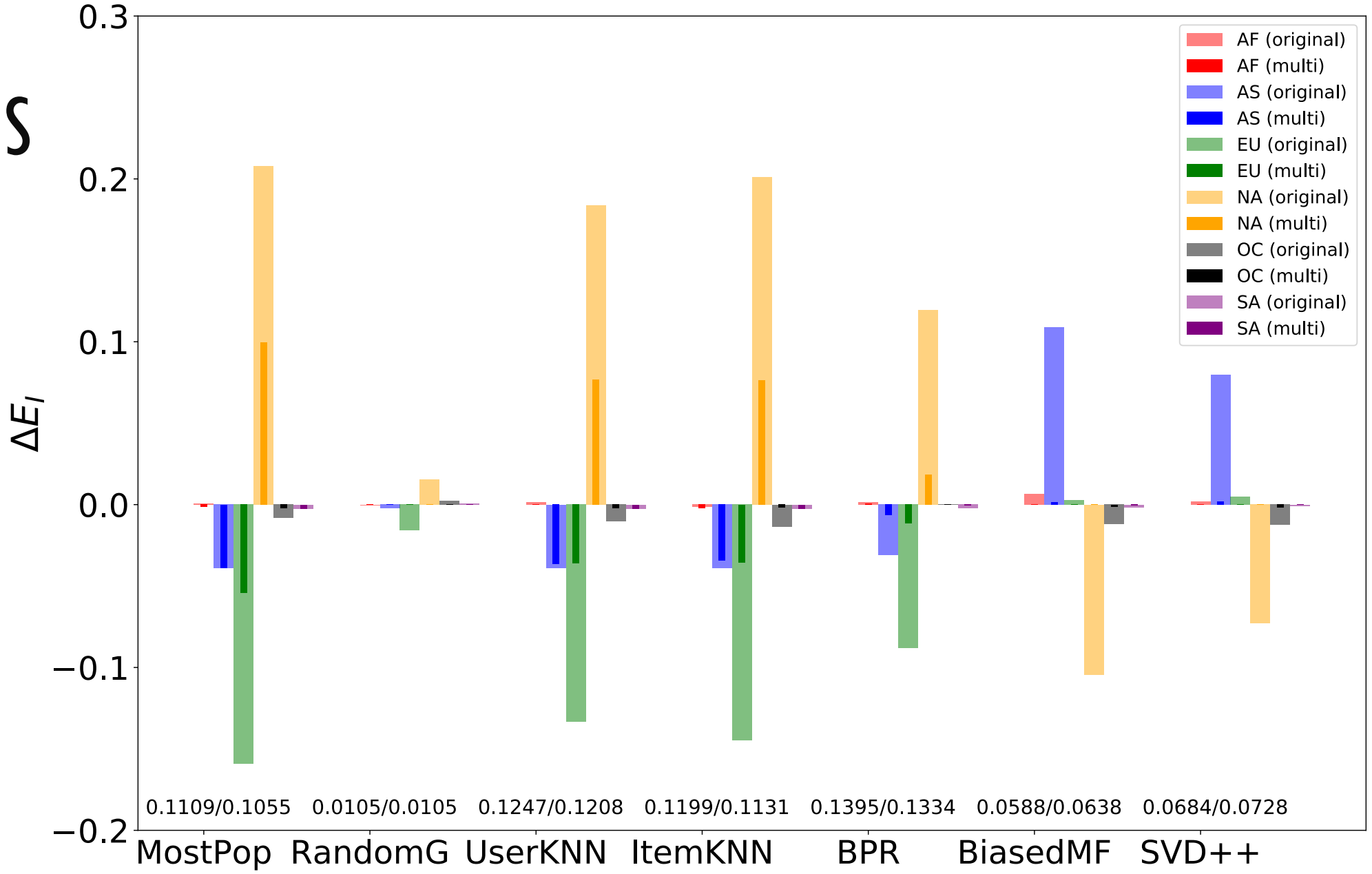
The foundation behind our mitigation algorithm is to **promote** in the recommendation list the item, that considering all the users, **minimizes the loss** in prediction.

The **optimizeContinentsVisibilityExposure** executes the mitigation:

1. To regulate the **visibility** of the disadvantaged groups (by adding their items from the **top-n** to the top-k)
2. To regulate the **exposure** (by moving their items up in the **top-k**)

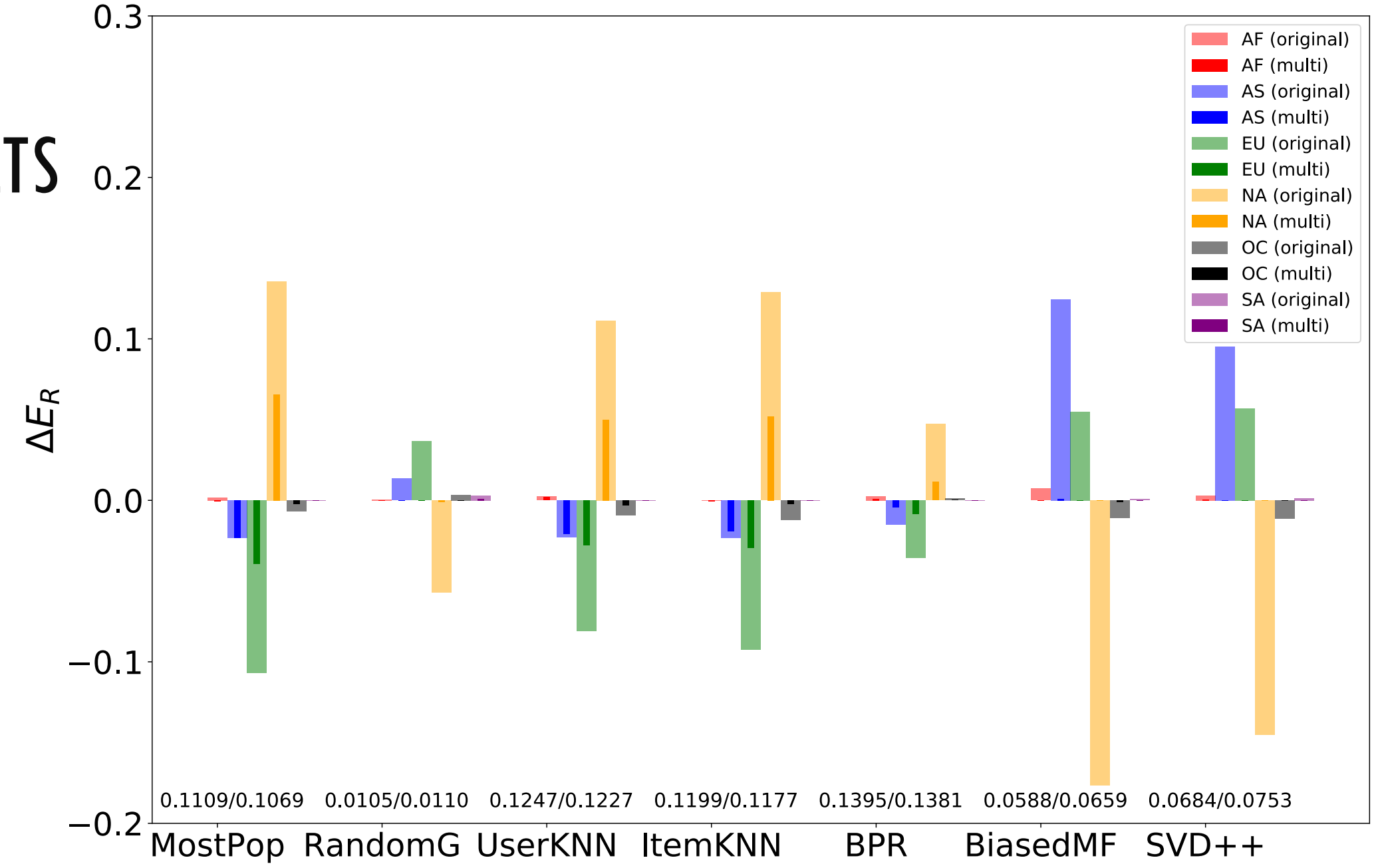
RESULTS

Visibility



RESULTS

Exposure



CONCLUSIONS

- We study the scenario in which **imbalances are associated with the way an industry is composed,** with certain geographic areas that produce more items of certain types.
- We **assess how recommender systems deal with data imbalances** and possible unfairness phenomena emerging from the way recommendations are distributed.
- We have proposed a **multi-group re-ranking approach** that re-distributes the recommendation across provider groups based on a notion of equity.
- Experimental results show that our approach introduces **provider fairness** without affecting **recommendation effectiveness.**

HANDS ON BIAS IN RECSYS

Maria Salamó, maria.salamo@ub.edu

Dept. de Matemàtiques i Informàtica, Universitat de Barcelona, Gran Via de les Corts Catalanes, 585, Barcelona, Spain

Institute of Complex Systems (UBICS), Universitat de Barcelona, Barcelona, Spain

Source: https://github.com/biasinrecsys/ecir2021-tutorial

STEPS ON THIS HANDS ON

Data Load

We load data from publicly available datasets, specifically focusing on Movielens 1M (movies)

Data Pre-Processing

We process data to be fed into the model and we prepare training samples, focusing on pairwise data

Model Definition and Train

We define the architecture of the model, setup the training parameters and run the model training process

Relevance Computation

Given a pre-trained model, we compute the user-item relevance scores across all the user-item pairs

Model Evaluation

We compute accuracy and beyond-accuracy metrics, such as coverage, novelty, and diversity

DISCLAIMERS

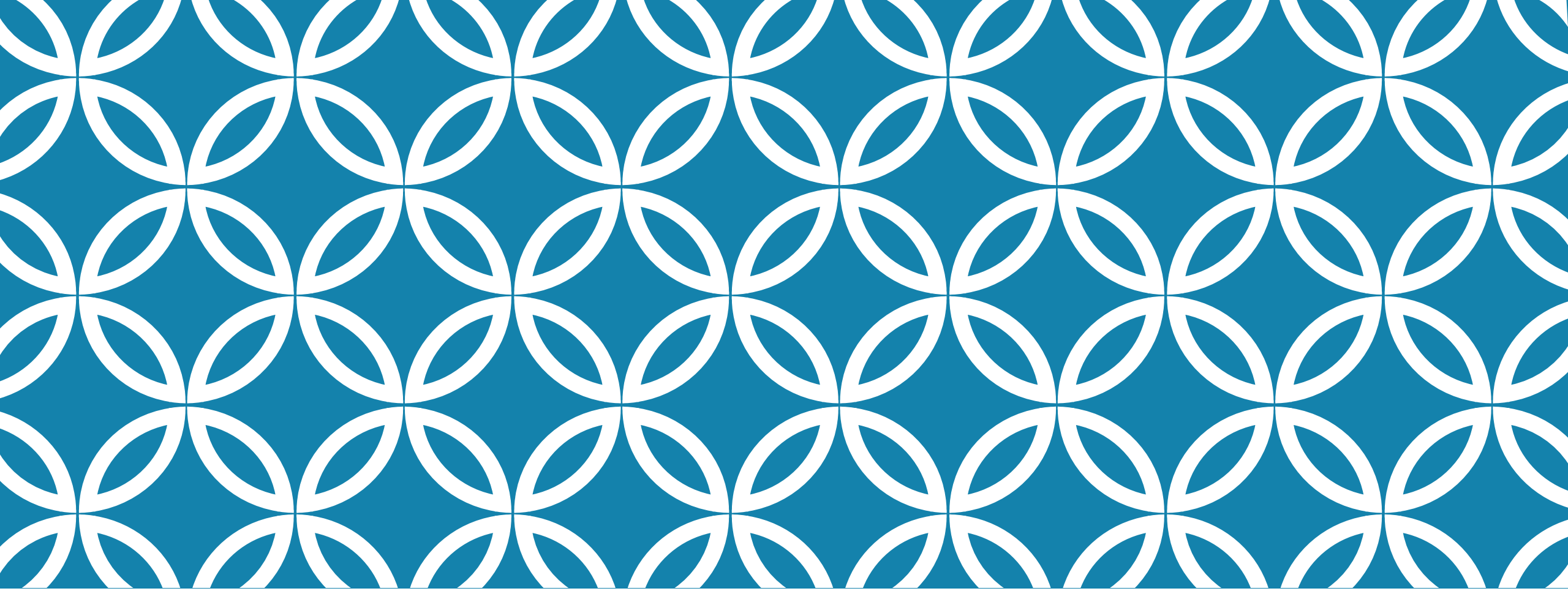
- In this tutorial, the authors do not aim to show how to fine-tune algorithms
- Due to the time constraints, authors decided to reduce the optimization part
- The pre-trained models do not represent fine-tuned baselines
- The goal is to get familiar with an environment where it is easier to control the whole Recommender System process

HANDS ON ITEM POPULARITY BIAS

https://github.com/biasinrecsys/ecir2021-tutorial

Select Notebook **#1**: Designing and Evaluating a recommendation Algorithm

Select Notebook **#2**: Popularity bias in personalized rankings



UNIVERSITAT DE BARCELONA

SPECIAL CLASS: BIAS AND FAIRNESS IN MACHINE LEARNING

T12. INTRODUCTION TO MACHINE LEARNING

Maria Salamó
Universitat de Barcelona, Spain
maria.salamo@ub.edu