Joint Evaluation of Fairness and Relevance in Recommender Systems with Pareto Frontier

Theresia Veronika Rampisela, Tuukka Ruotsalo,
Maria Maistro, Christina Lioma

1 May 2025
The Web Conference 2025
Sydney, Australia

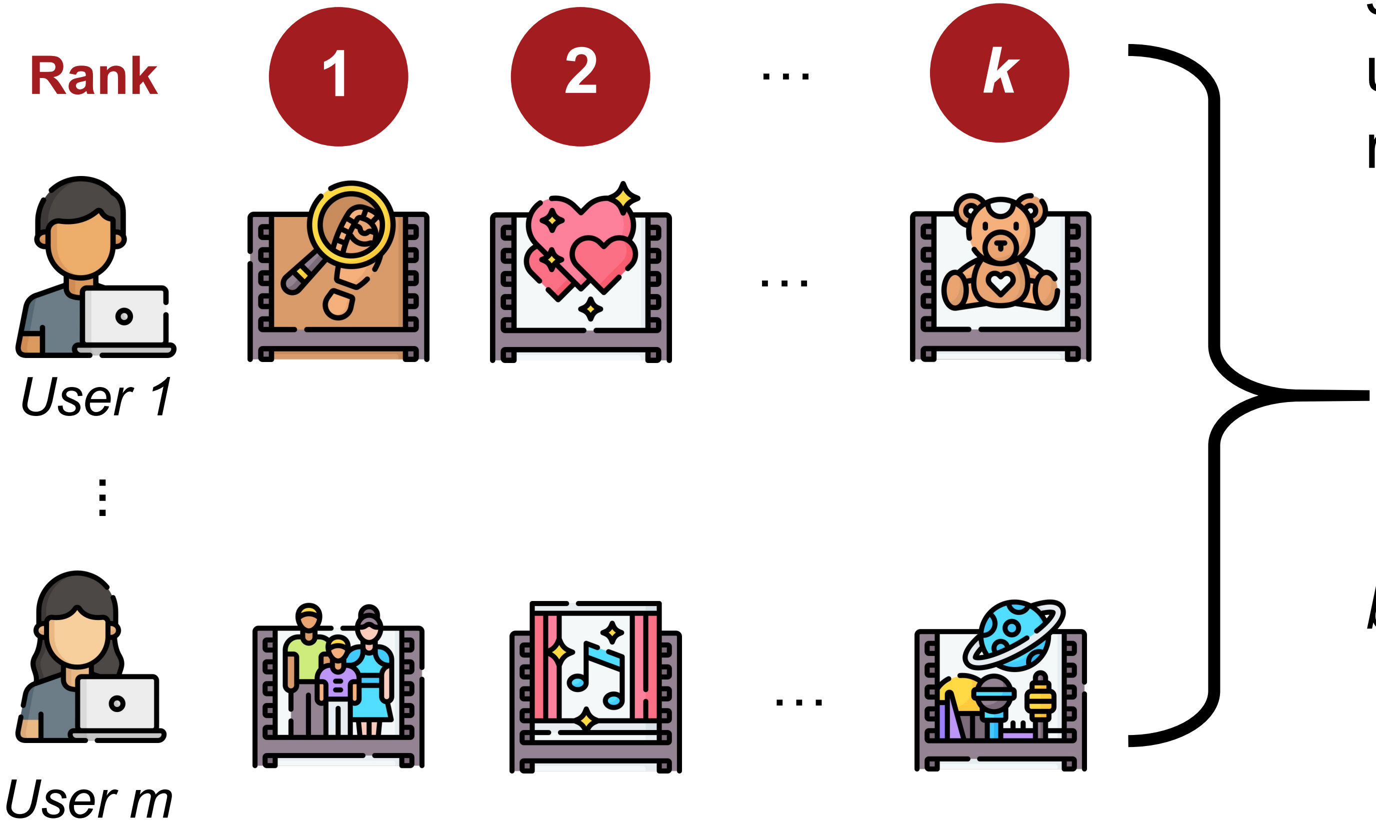
UNIVERSITY OF COPENHAGEN

This work is funded by:
ADD algorithms data & democracy

What is the **maximum achievable fairness and relevance** based on the dataset composition?

How close is the model performance **to an ideal balance of fairness and relevance?**

Background



Recommender systems:
systems that can match items to users such that the recommendations are:

- **Relevant to the users:**
  the users like the items or find them useful

but we also want them to be

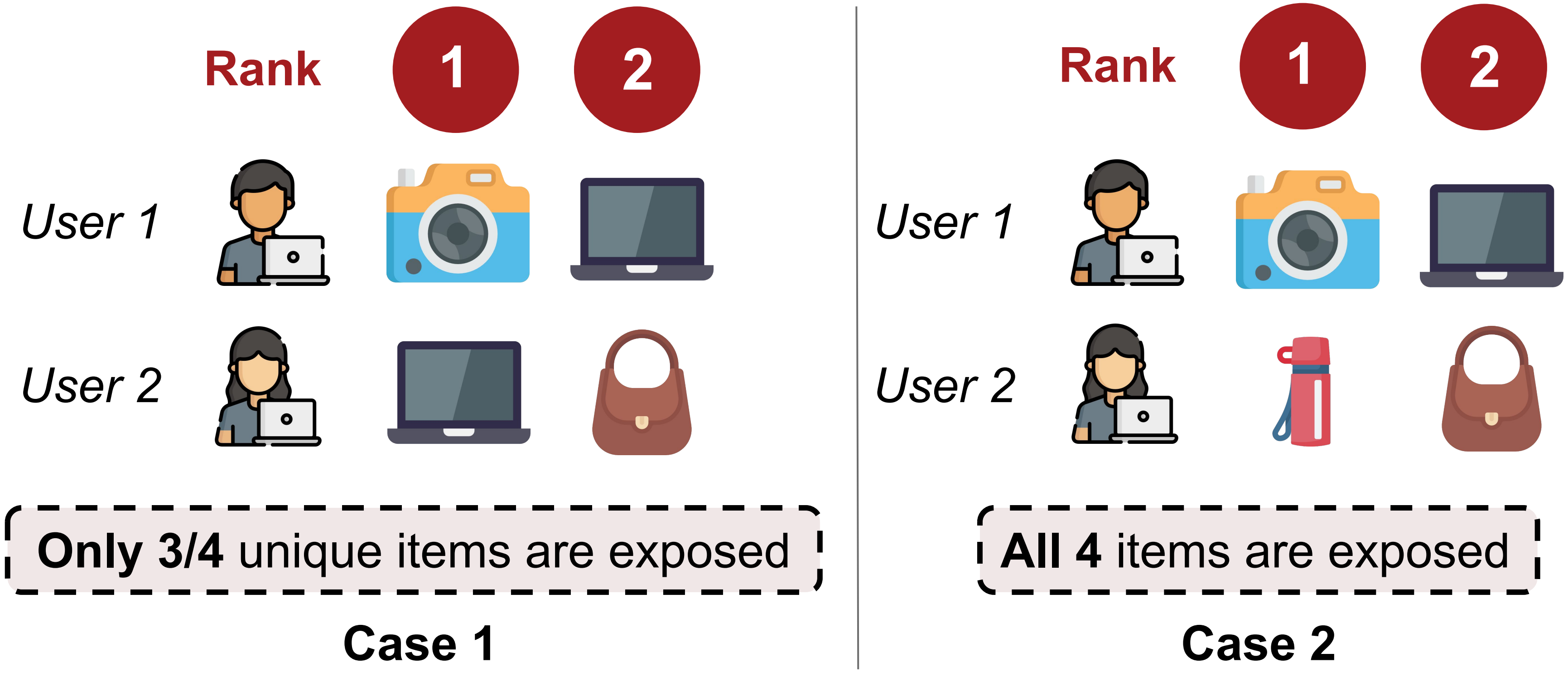
- **Fair to the items:**
  each item get recommended to users for similar amount of times

Image credit:
Icons created by Freepik - Flaticon

Intuitive example: **Individual item fairness in RecSys**

We recommend *k*=2 items from a pool of four items to two users

Items in the dataset:



Rank	**1**	**2**
User 1		
User 2		

Only 3/4 unique items are exposed

Case 1

Rank	**1**	**2**
User 1		
User 2		

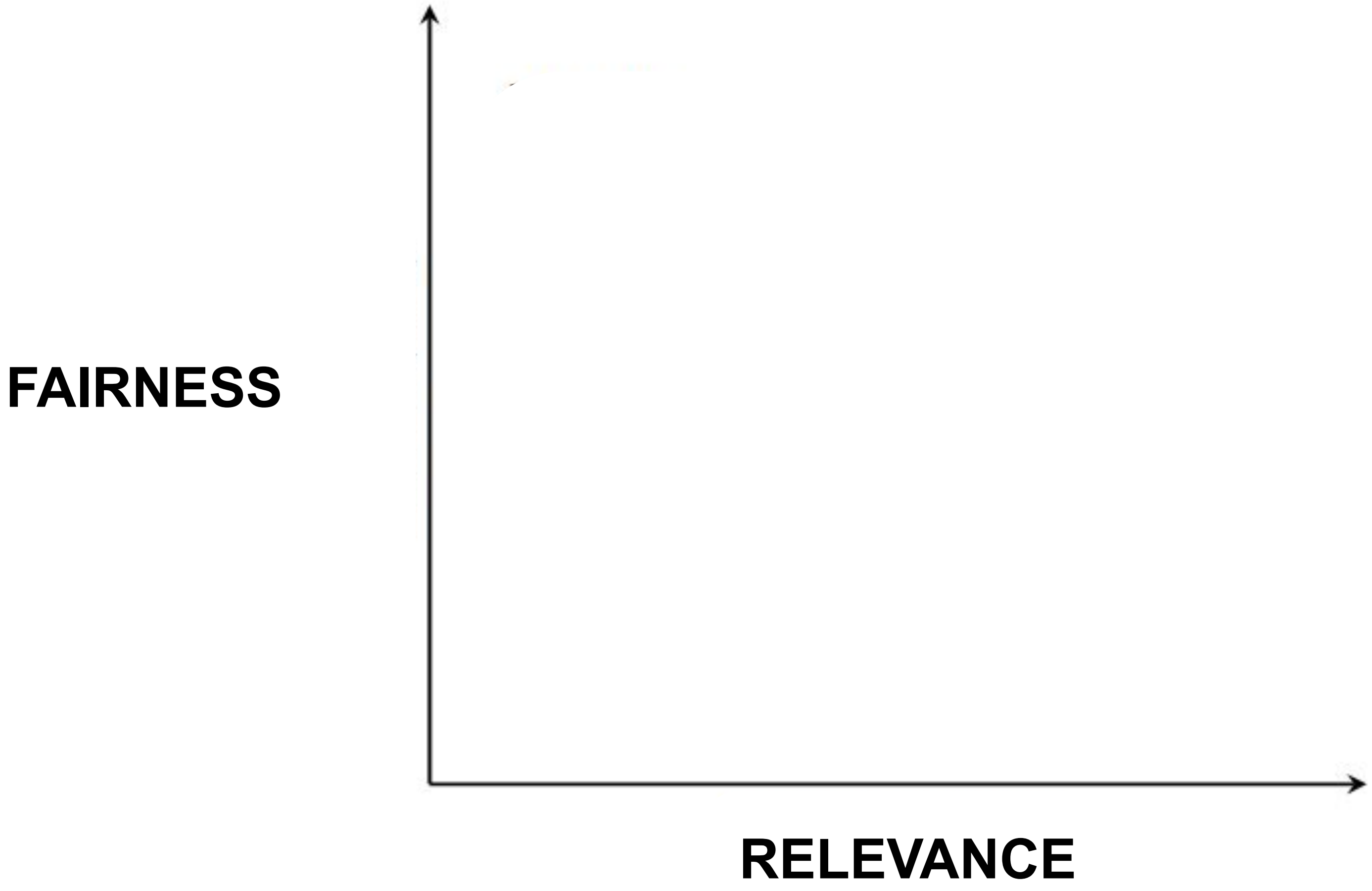
All 4 items are exposed

Case 2

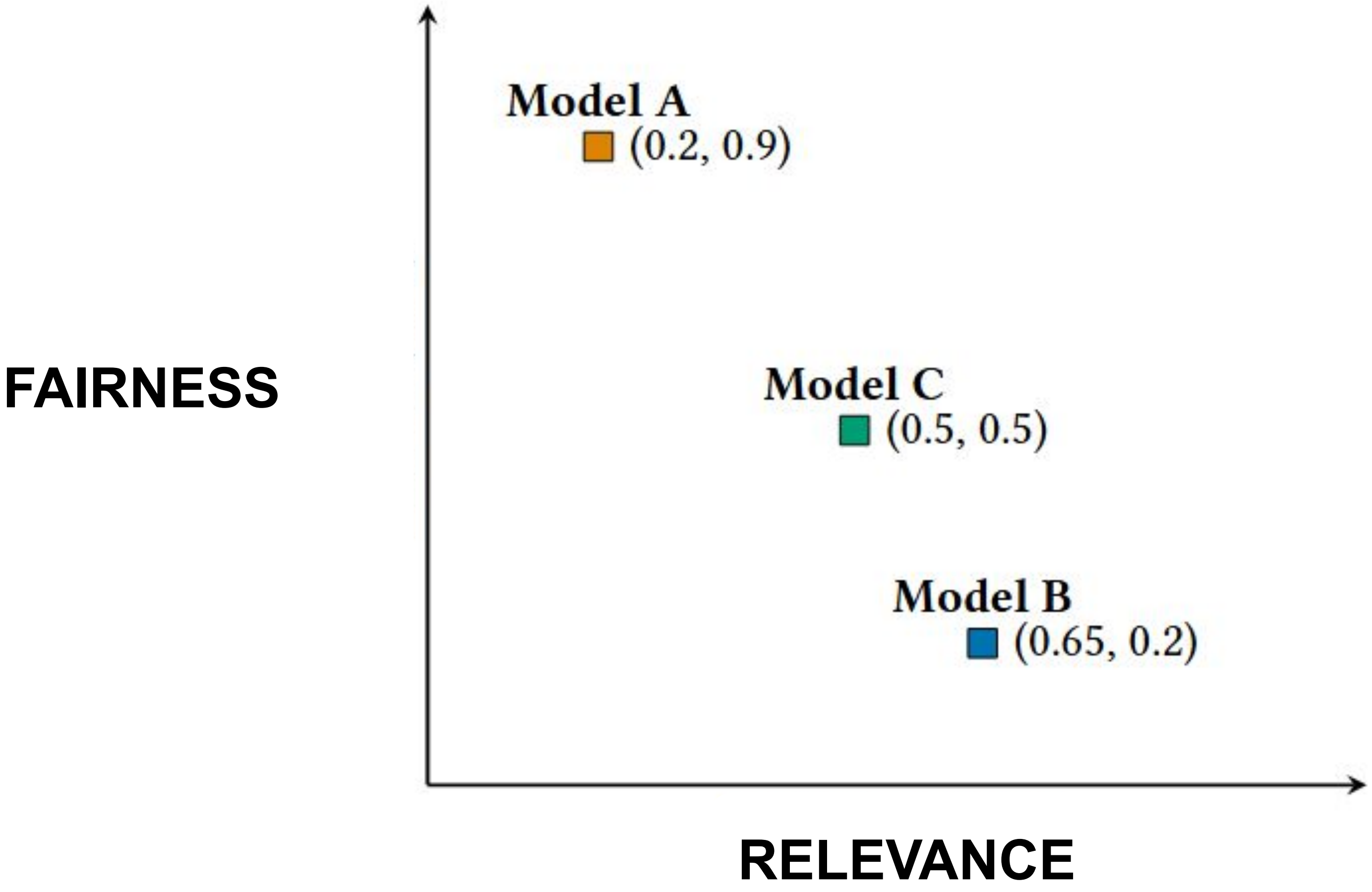
More unique items exposed in Case 2
→ Case 2 is fairer

Image credit:
Icons created by Freepik - Flaticon

Two **evaluation aspects:** fairness and relevance

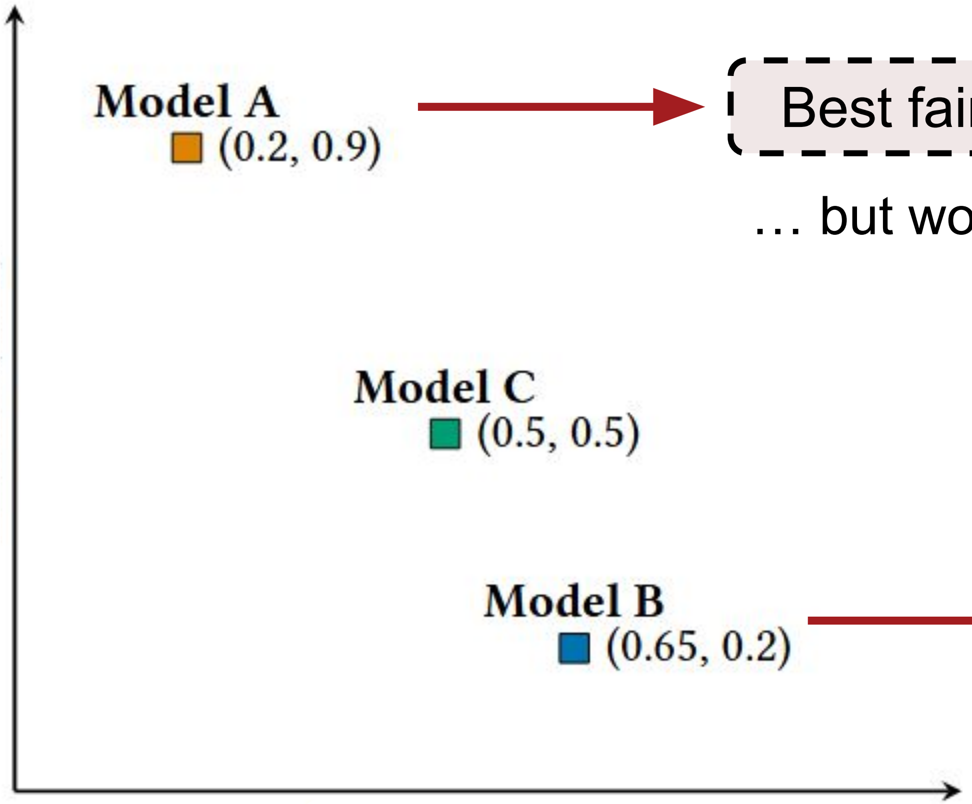


Suppose that we have the scores from three models…



If we measure fairness and relevance **separately**…

FAIRNESS



RELEVANCE

Best fairness

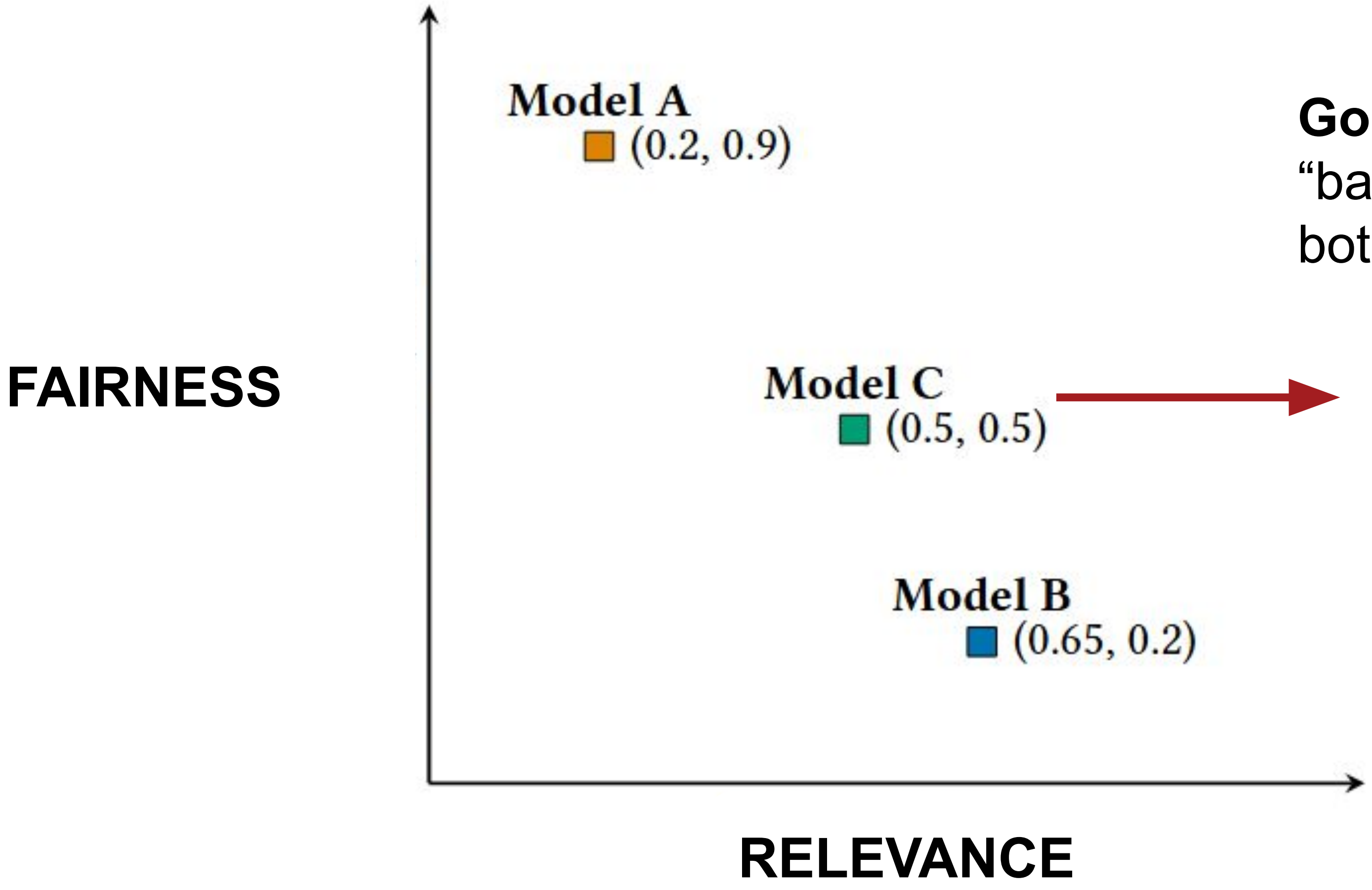
… but worst relevance

Best relevance

… but worst fairness

Assumption: both fairness and relevance measures range in [0,1]

What if we want to measure **fairness** and **relevance jointly**?

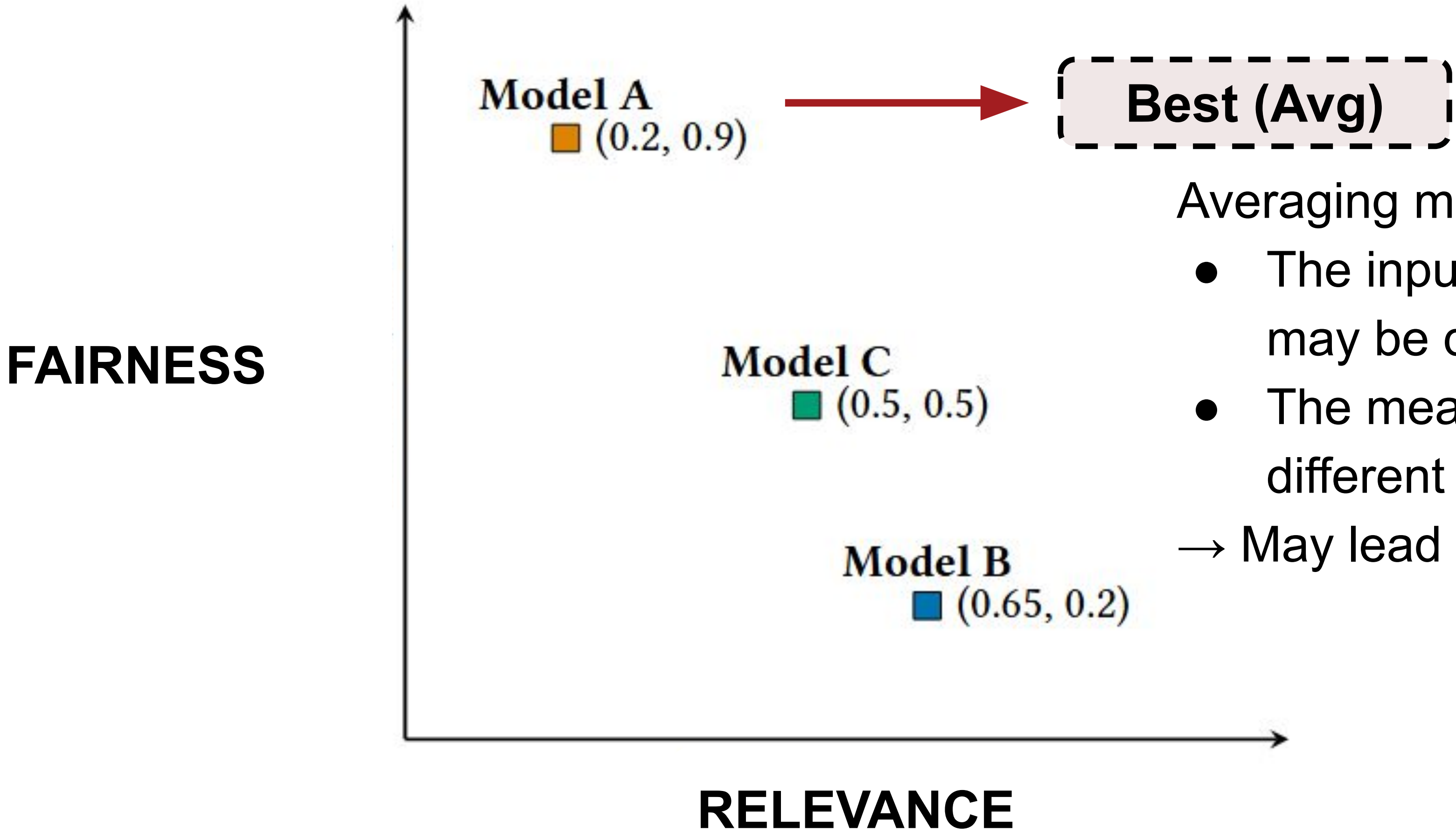


Goal: we want to have a “balanced” model in terms of both fairness and relevance

Best ?

but how?

Averaging the fairness and relevance scores?

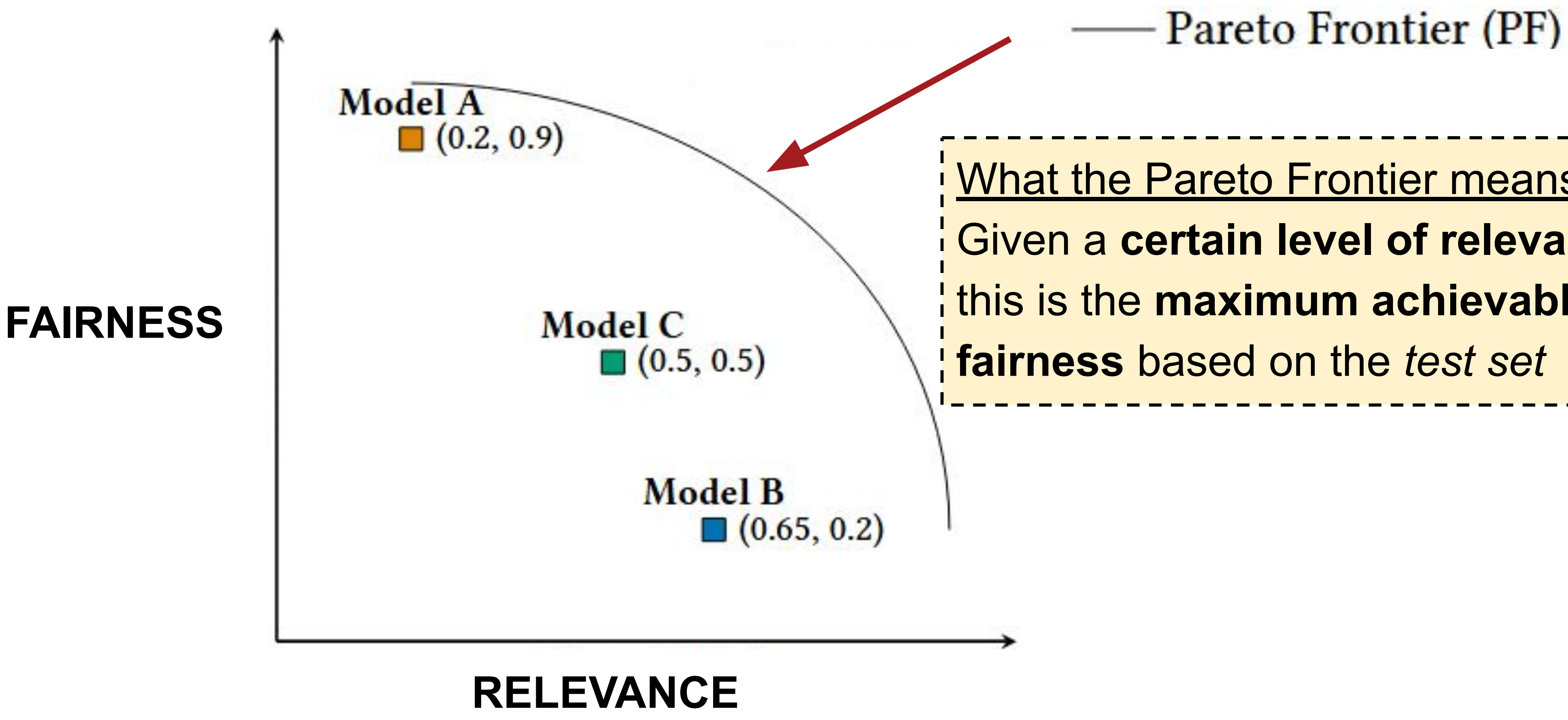


Averaging may be problematic:

- The input to the two measures may be different
- The measures may have different scales/distributions

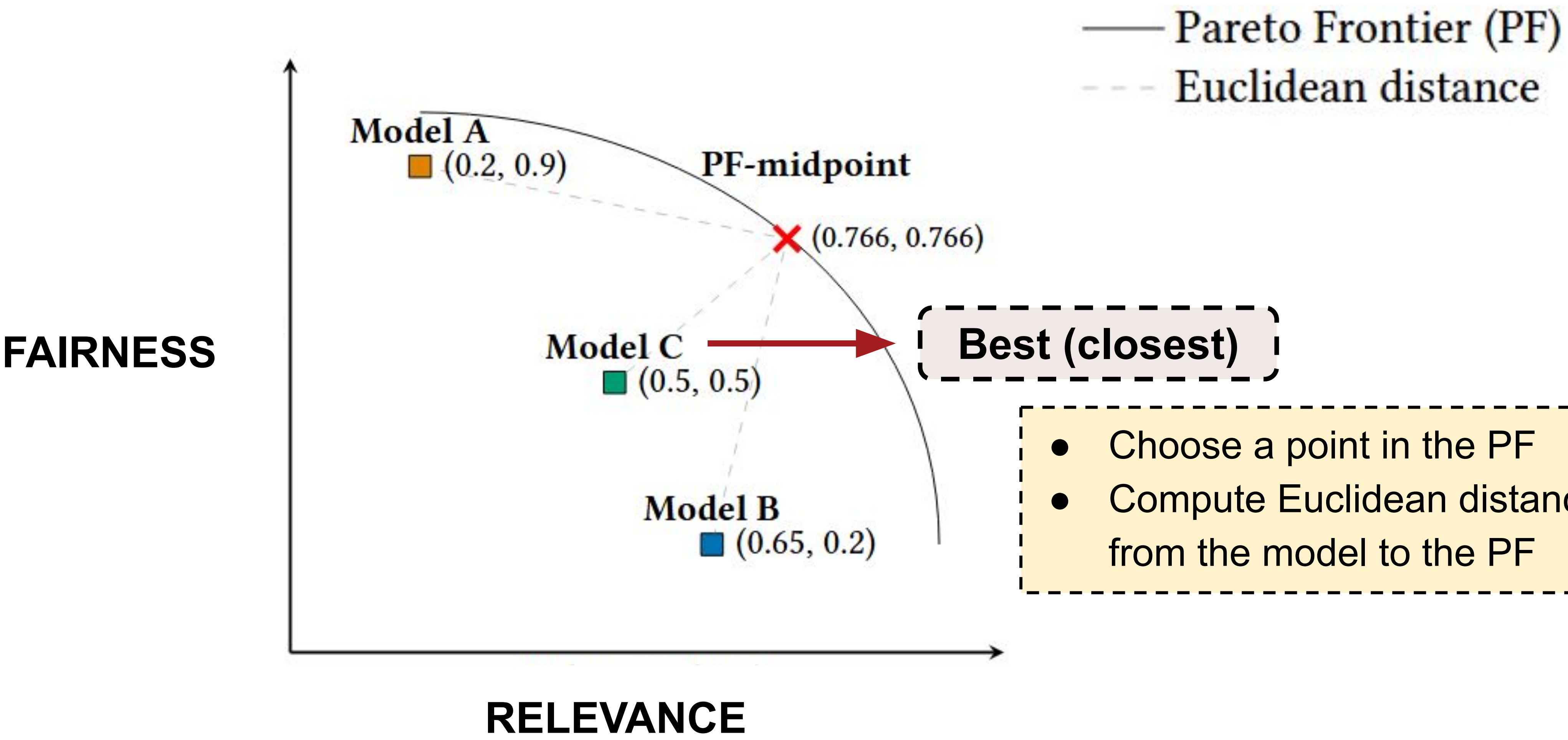
→ May lead to incorrect conclusions

Q1. What is the **maximum achievable fairness and relevance** based on the dataset composition?



What the Pareto Frontier means:
Given a **certain level of relevance**, this is the **maximum achievable fairness** based on the *test set*

Q2. **How close** are the models to an **ideal balance of fairness & relevance?**



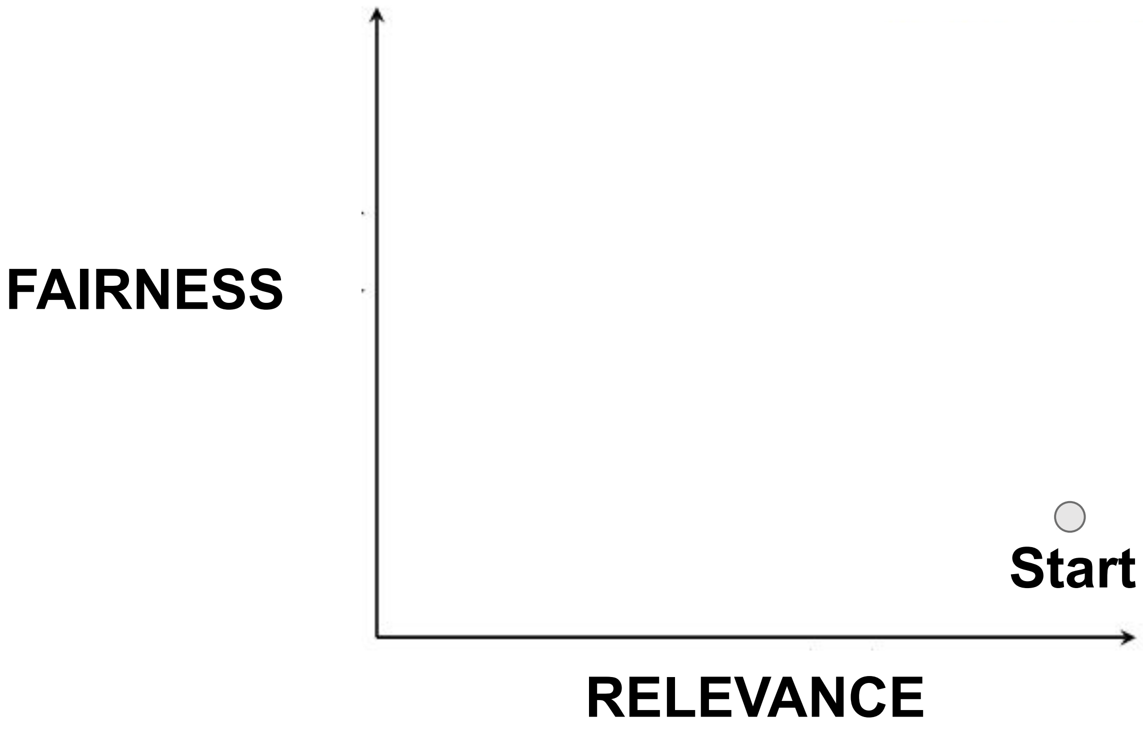
Best (closest)

- Choose a point in the PF
- Compute Euclidean distance from the model to the PF

How to **generate the Pareto Frontier** from the test set?

→ New algorithm: Oracle2Fair

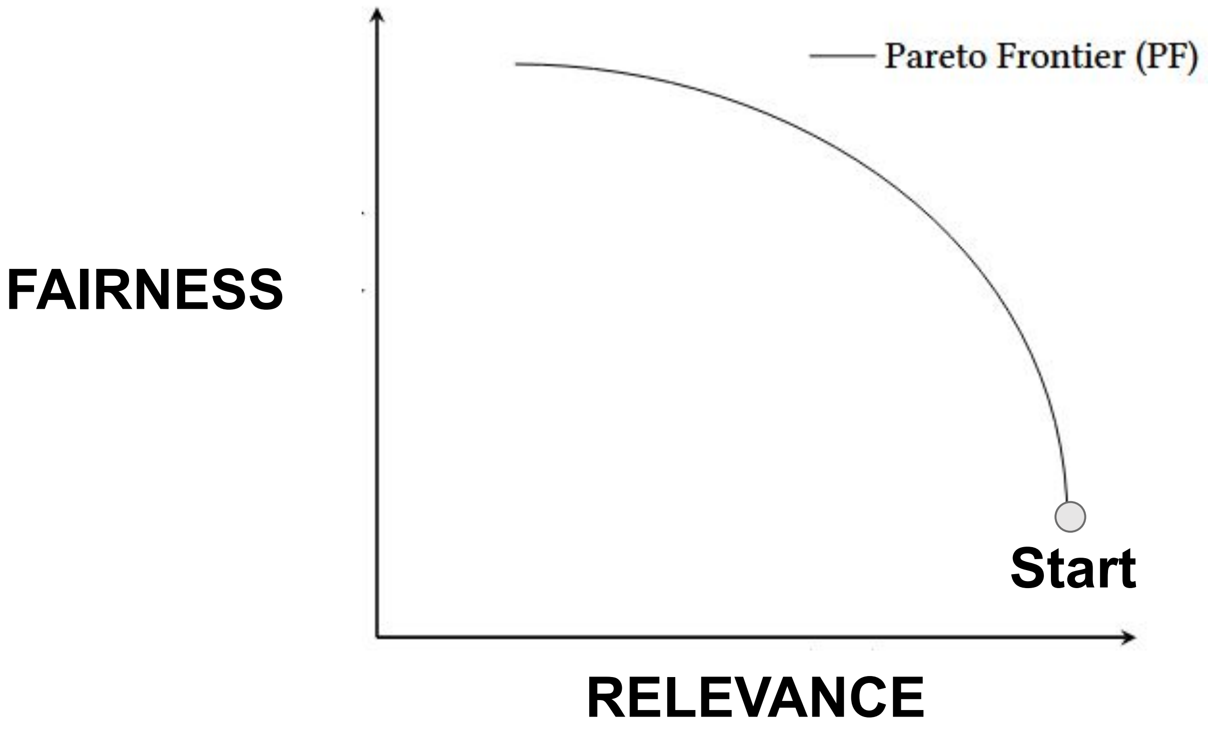
Generating the Pareto Frontier (Oracle2Fair Algorithm)



Start: create maximally relevant recommendations by recommending items in the test split

(and compute fairness and relevance measures)

Generating the Pareto Frontier (Oracle2Fair Algorithm)

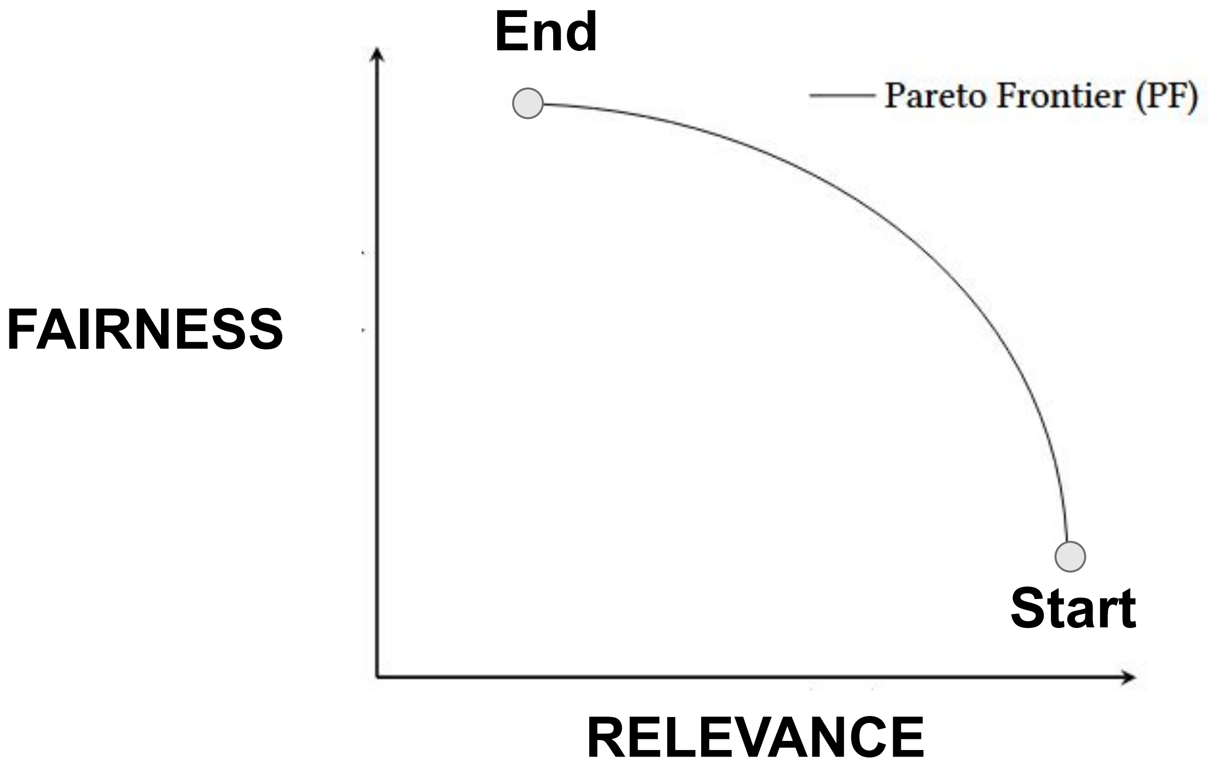


Start: create maximally relevant recommendations by recommending items in the test split

Iteratively replace **most popular** items with **least popular items** to increase fairness (sacrificing relevance)

(and compute fairness and relevance measures every replacement)

Generating the Pareto Frontier (Oracle2Fair Algorithm)



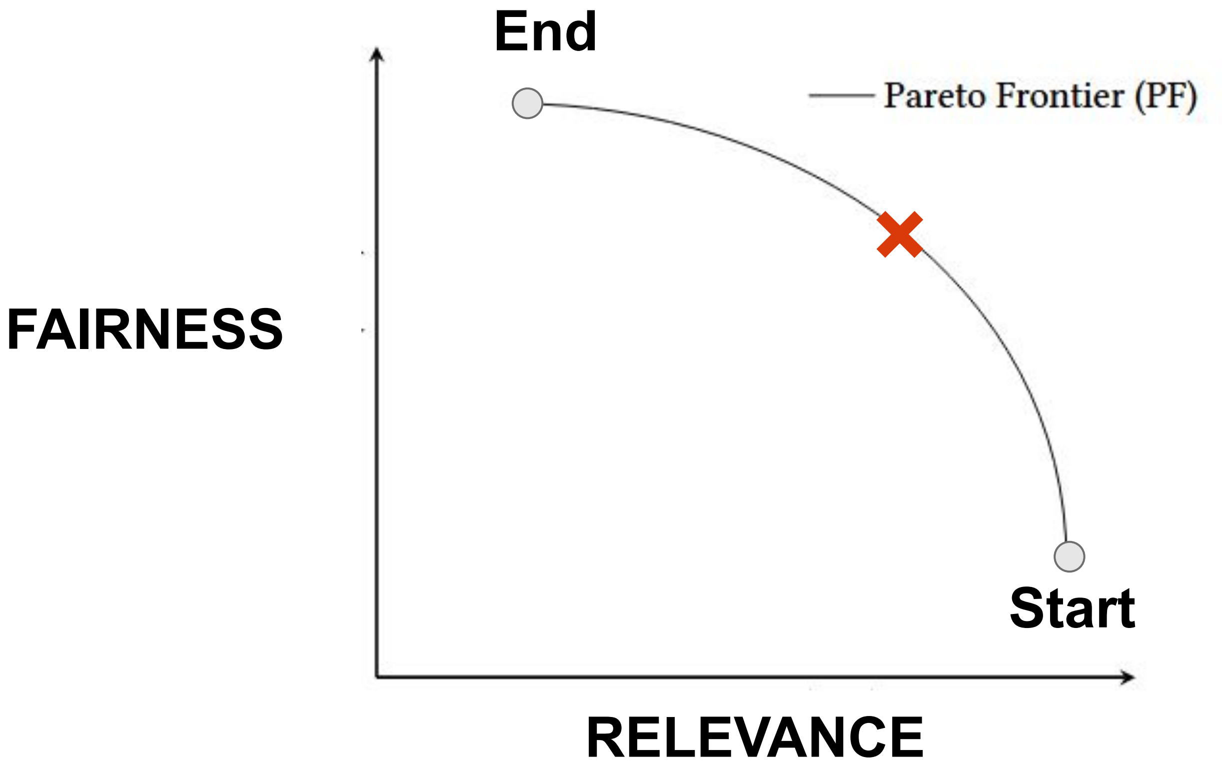
Start: create maximally relevant recommendations by recommending items in the test split

Iteratively replace **most popular** items with **least popular items** to increase fairness (sacrificing relevance)

End: fairest possible recommendation

(and compute fairness and relevance measures)

Computing the reference point ✖



Select a point in the PF based on α:
α controls the relative position between the start & end points

- α=0 only considers relevance
- α=1 only considers fairness

Compute the distance between the model score to the PF as **Distance to Pareto Frontier (DPFR)**

Experiment

Datasets: 6 publicly available data (various domain, sparsity, size)

Recommenders:

- 4 models: ItemKNN, BPR, MultiVAE, NCL
- 3 fair rerankers: Greedy Substitution (GS), COMBMNZ (CM), Borda Count (BC)

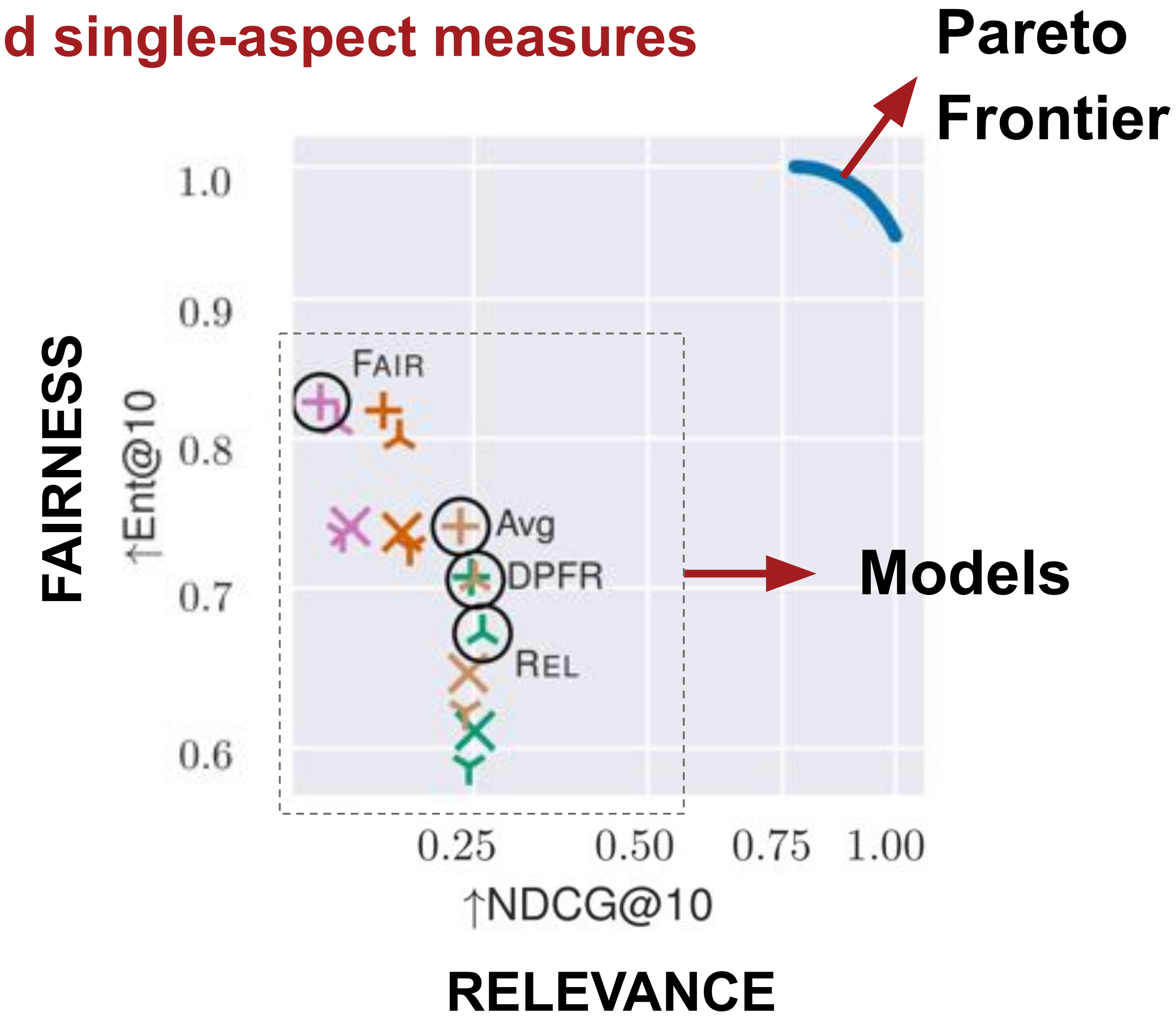
Evaluation:

- **Single-aspect measures**: 6 relevance (REL) + 5 fairness (FAIR)
- **Joint measures** of relevance & fairness:
    - 5 existing joint measures of fairness w.r.t. relevance
    - Avg: Averaging relevance + fairness score
- **DPFR**: Distance to Pareto Frontier, combining 6 x 5 REL-FAIR measure pairs

Finding #1: Comparison between DPFR and single-aspect measures

For all datasets and all measure pairs:

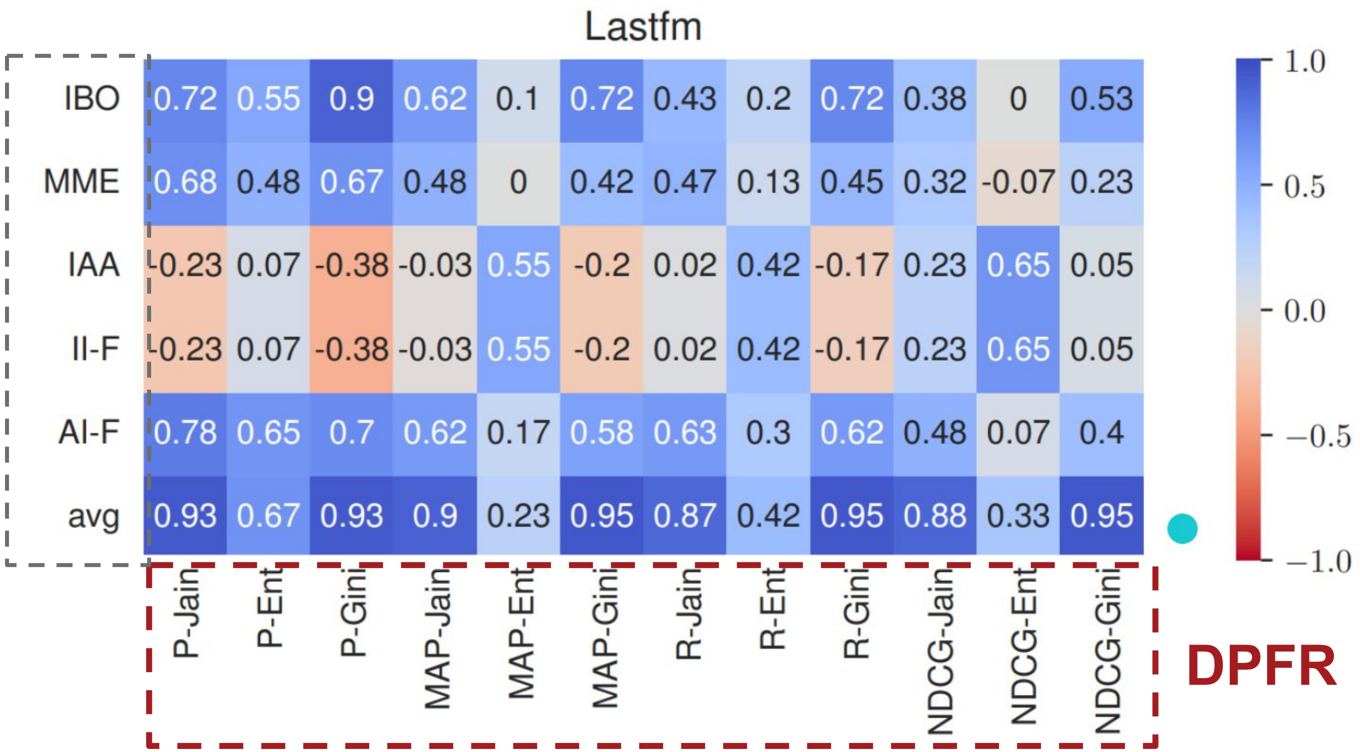
- The **best model based on DPFR is always different** from the best model based on **relevance** measures
- Half the time, the **best model as per DPFR is different from the best model** based on **fairness** measures



Can we use existing measures to reach the same conclusion as DPFR?

Finding #2: Comparison between DPFR and existing joint evaluation baselines

- We compute Kendall's τ correlations between the ranking of models given by DPFR and by the joint evaluation baselines
- If Kendall's Tau ≥ 0.9, we consider the rankings equivalent
- **Existing measures** that jointly quantify fairness w.r.t. relevance do not consistently rank models equivalently to DPFR → **they are not a reliable proxy for DPFR!**



Joint evaluation baselines

	P-Jain	P-Ent	P-Gini	MAP-Jain	MAP-Ent	MAP-Gini	R-Jain	R-Ent	R-Gini	NDCG-Jain	NDCG-Ent	NDCG-Gini
IBO	0.72	0.55	0.9	0.62	0.1	0.72	0.43	0.2	0.72	0.38	0	0.53
MME	0.68	0.48	0.67	0.48	0	0.42	0.47	0.13	0.45	0.32	-0.07	0.23
IAA	-0.23	0.07	-0.38	-0.03	0.55	-0.2	0.02	0.42	-0.17	0.23	0.65	0.05
II-F	-0.23	0.07	-0.38	-0.03	0.55	-0.2	0.02	0.42	-0.17	0.23	0.65	0.05
AI-F	0.78	0.65	0.7	0.62	0.17	0.58	0.63	0.3	0.62	0.48	0.07	0.4
avg	0.93	0.67	0.93	0.9	0.23	0.95	0.87	0.42	0.95	0.88	0.33	0.95

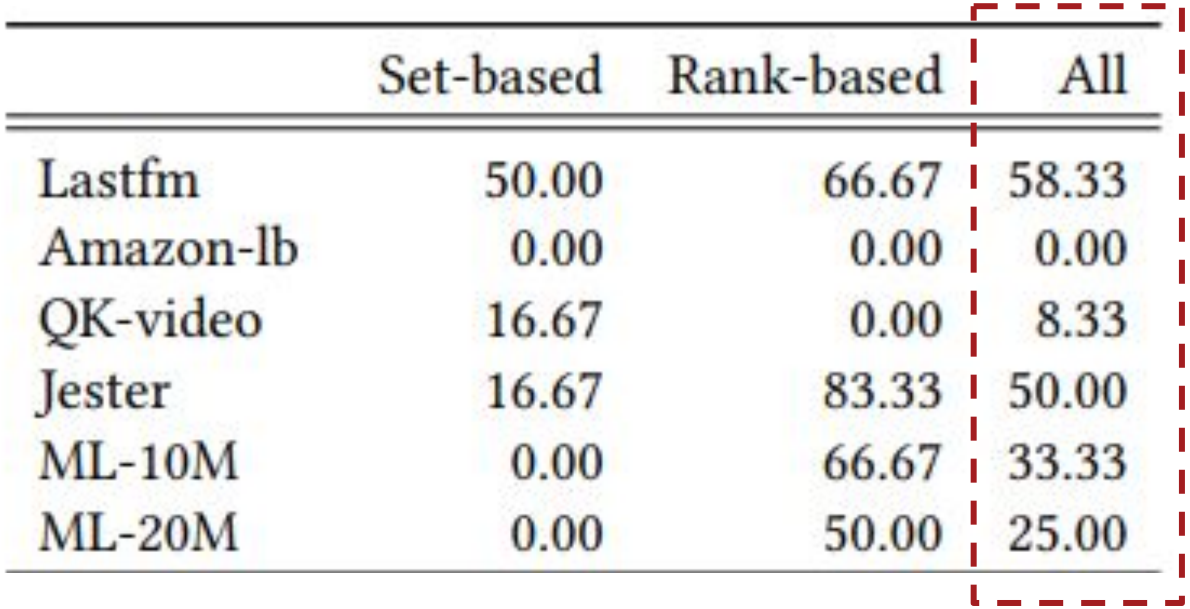
DPFR

strong agreement

strong disagreement

Finding #3: Comparison between DPFR and Averaging Fairness + Relevance (Avg)

- We count the percentage of times the best model based on Avg differs from DPFR
- **The best model based on Avg differs from DPFR up to 83% of the time**



	Set-based	Rank-based	All
Lastfm	50.00	66.67	58.33
Amazon-lb	0.00	0.00	0.00
QK-video	16.67	0.00	8.33
Jester	16.67	83.33	50.00
ML-10M	0.00	66.67	33.33
ML-20M	0.00	50.00	25.00

Huge range of variability across datasets (0–58.33%)

→ averaging fairness & relevance scores **cannot be guaranteed** to get the same result as DPFR

Summary

We contributed **DPFR, a new Pareto-optimal-based evaluation** approach

… to evaluate **fairness and relevance jointly**

… by **measuring the distance** from the **model performance** to an **ideal balance of fairness and relevance**

… based on **existing measures** for **relevance** and for **fairness**

… and the approach is model-agnostic (as it is based on the test set composition)

Paper

Code

SIGIR’24 paper

Related paper (SIGIR’24):

- Deeper investigation into fairness measures that consider both item exposure and item relevance
- Extended version coming soon!

Thank you!