Fairness among New Items in Cold Start Recommender Systems

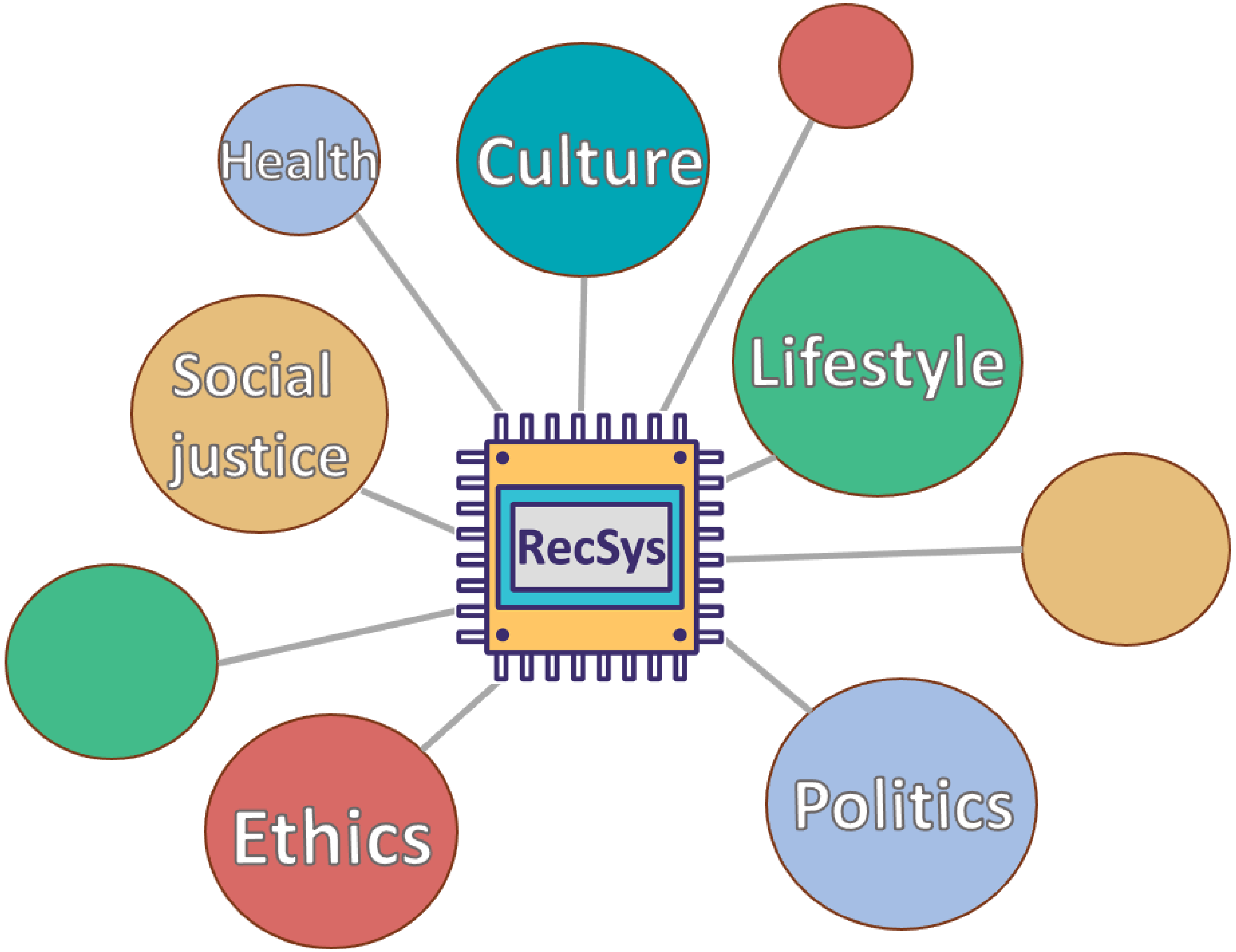
Ziwei Zhu, Jingu Kim*, Trung Nguyen*, Aish Fenton*, and James Caverlee

Texas A&M University

*Netflix

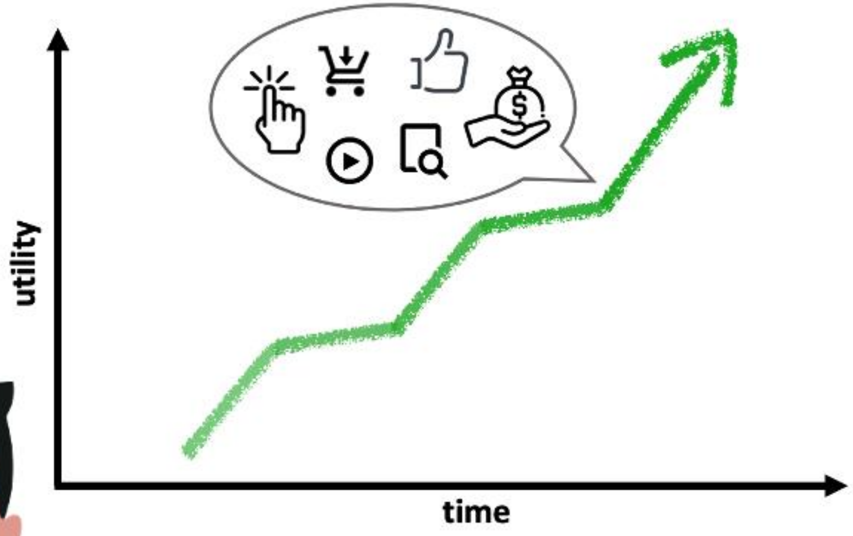
About me

Fifth-year PhD student working on **responsible recommender systems**.



About me

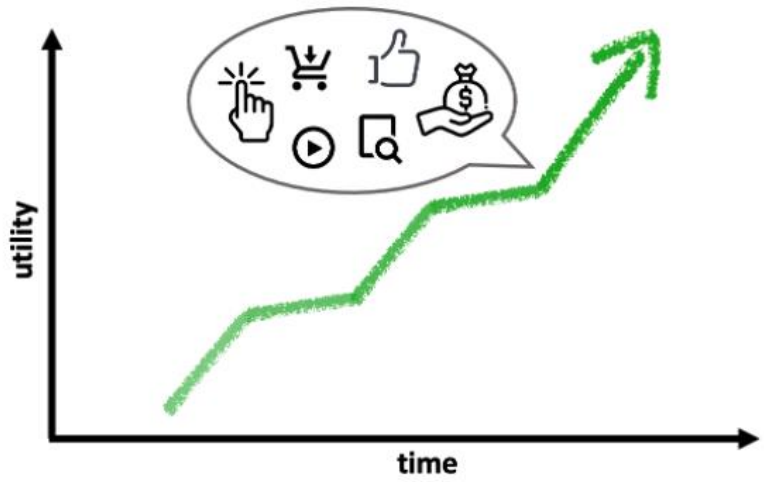
Fifth-year PhD student working on **responsible recommender systems**.



About me

Fifth-year PhD student working on **responsible recommender systems**.

Lifestyle, health,
culture, politics, ethics,
social justice, ...



About me

- Counteract **Exposure Bias** in User-item interaction Data
- Identify and Mitigate **Popularity-opportunity Bias**
- Measure and Enhance **Item Recommendation Fairness**
- Identify and Mitigate **Mainstream Bias** on Users

http://people.tamu.edu/~zhuziwei/

About me

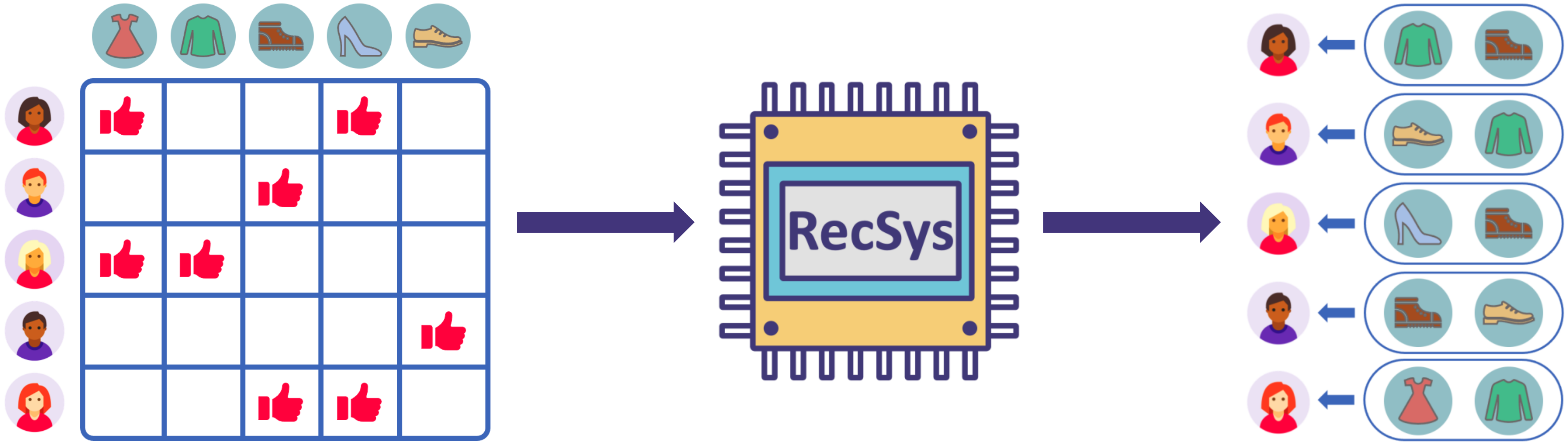
- Counteract **Exposure Bias** in User-item interaction Data
- Identify and Mitigate **Popularity-opportunity Bias**
- Measure and Enhance **Item Recommendation Fairness**
- Identify and Mitigate **Mainstream Bias** on Users

Recommenders connect users to items

User-item interaction data

Recommendation model

Recommendation results

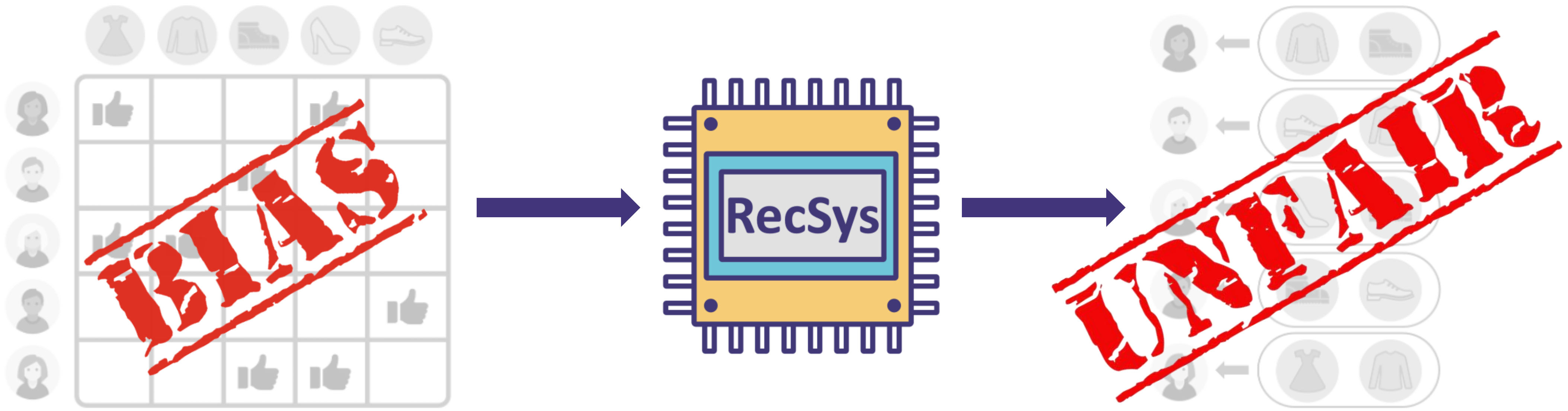


RecSys inherit or intensify data bias and produce unfairness

User-item interaction data

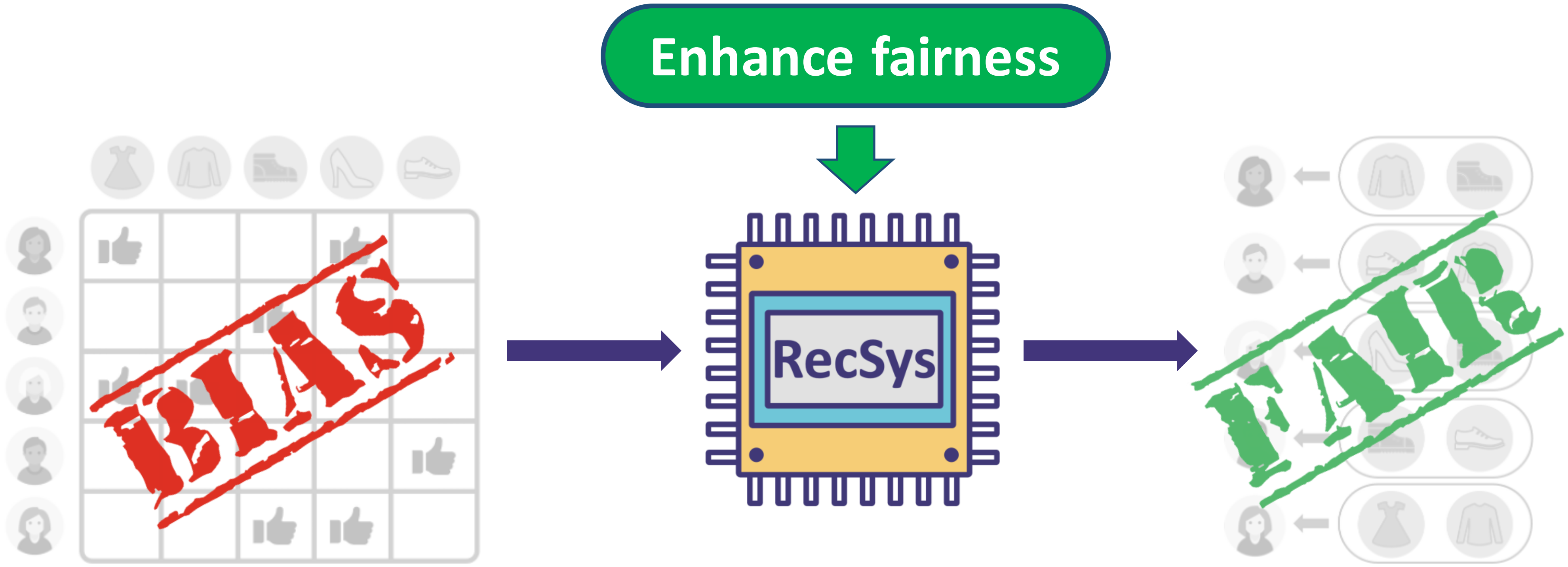
Recommendation model

Recommendation results

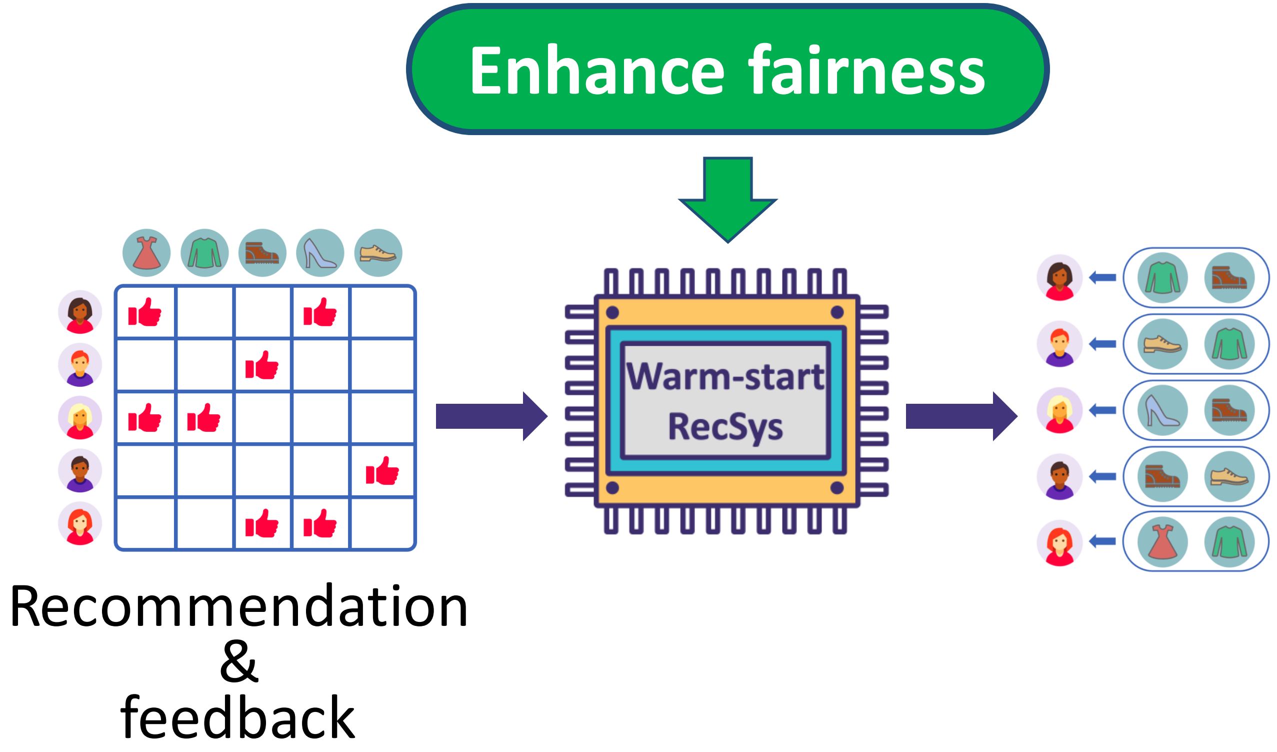


Prior works designed for fairness-enhanced RecSys

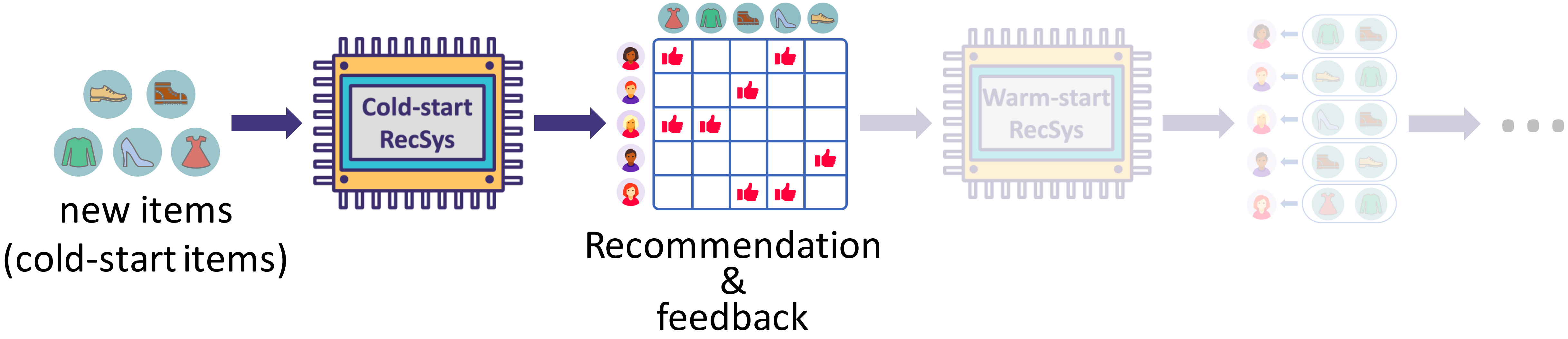
NeurIPS17, KDD18, CIKM18,
SIGIR18, KDD19, SIGIR21 ...



Only consider the fairness at the middle of life cycle of items

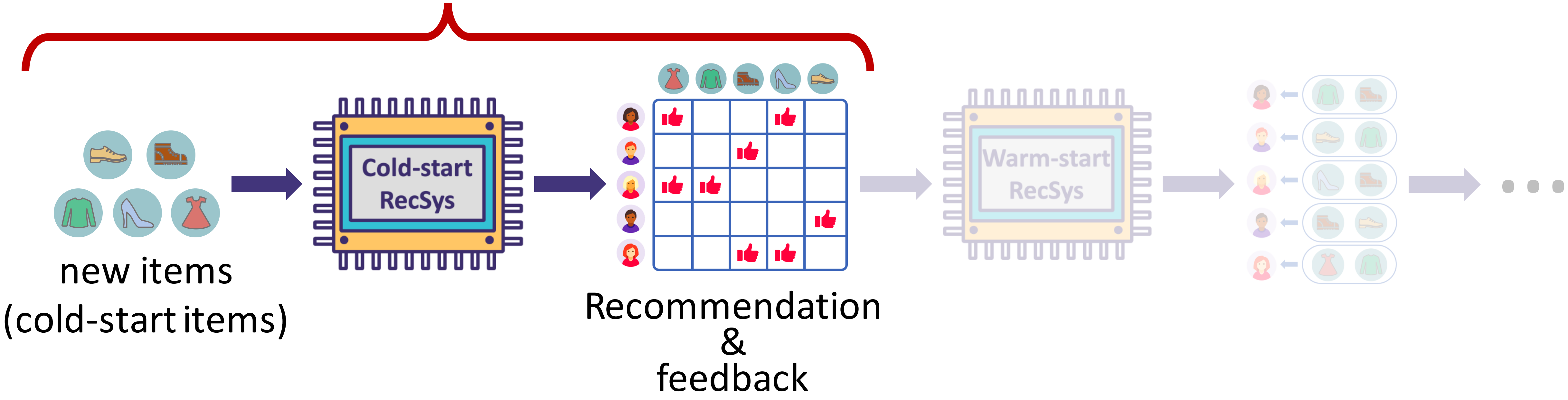


Only consider the fairness at the middle of life cycle of items



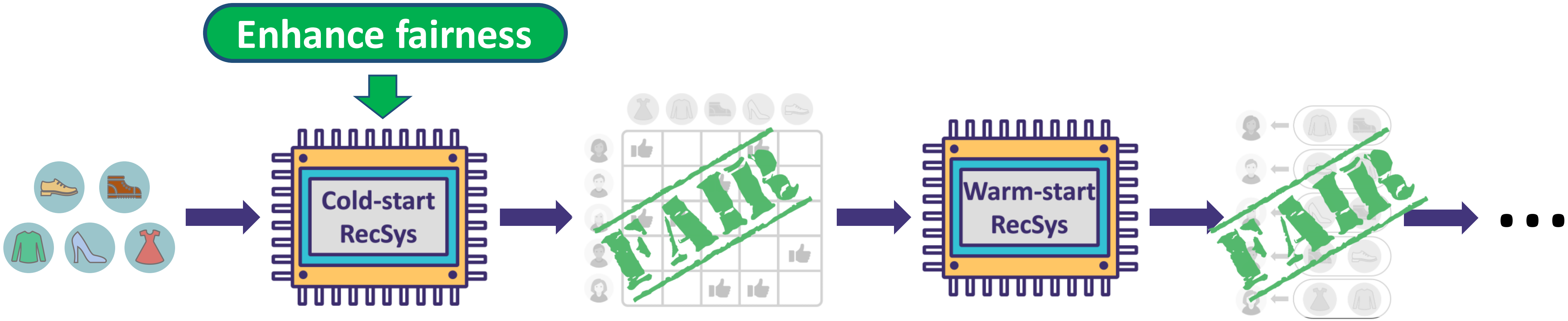
The fairness issue at the cold-start step is ignored

Are recommendations fair among these new items?



Fairness among new items is important

- Unfairness introduced by cold-start RecSys will be **perpetuated and accumulated** through the entire life cycle of items.
- Instead, providing fair recommendations among new items could give rise to a **virtuous circle** of collecting (relatively) unbiased feedback and training fairer models later in the life cycle.



Contributions

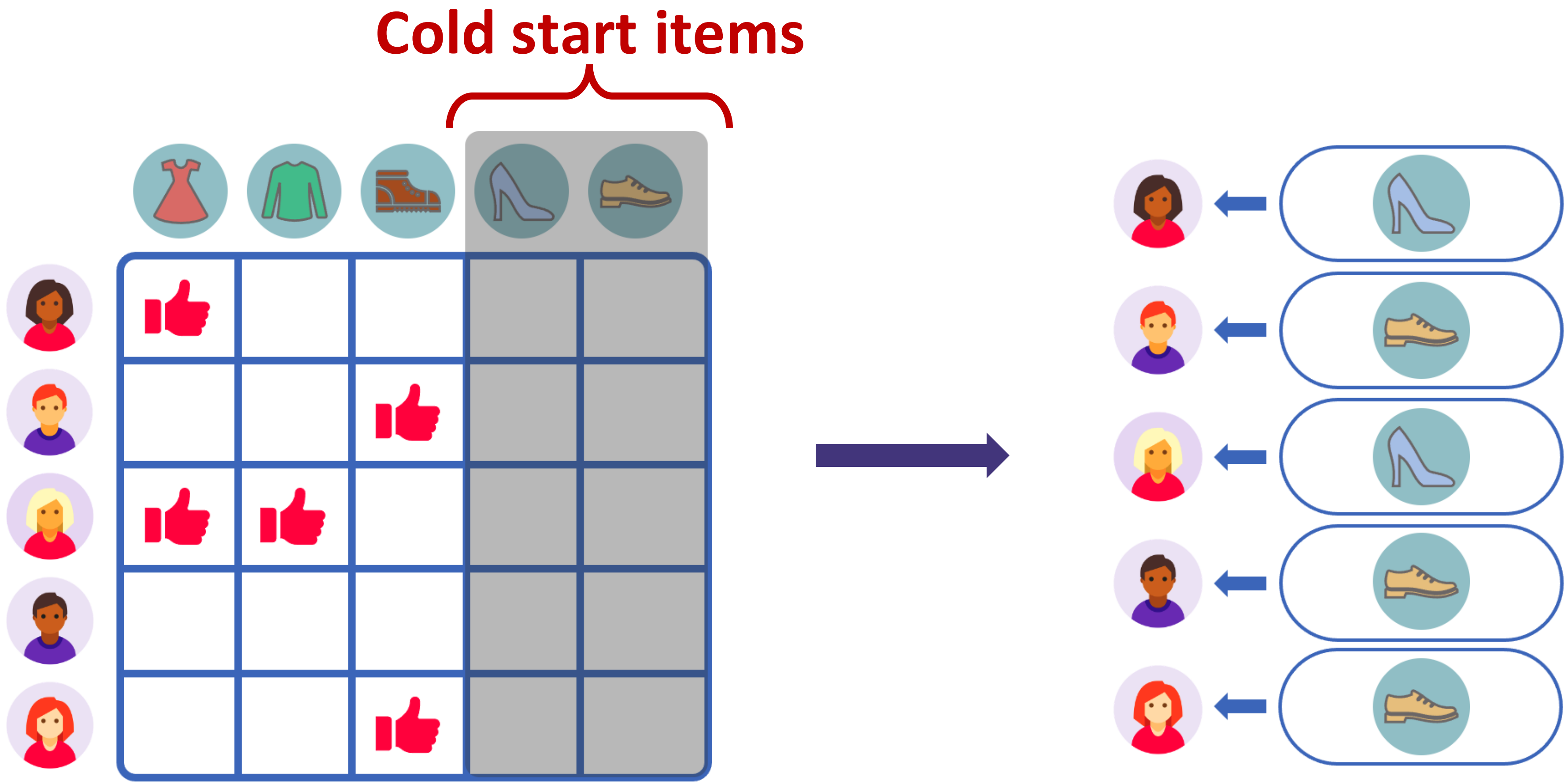
- Introduce the problem of **fairness among new items** in cold-start scenarios;
- Conduct a **data-driven study** to demonstrate the prevalence of unfairness among new items in cold-start RecSys.
- Propose a novel **learnable post-processing framework** as a solution blueprint. Based on this blueprint, we demonstrate two concrete approaches: a score **scaling** method and a **joint-learning generative** method.
- Extensive experiments show the **effectiveness** of the proposed methods.

Outline

- Motivations
- **Problem Formalization**
  - **Cold start recommendation**
  - **Item recommendation fairness**
- Data-driven Study
- Fairness-enhancing Approaches
- Fairness-enhancing Experiments

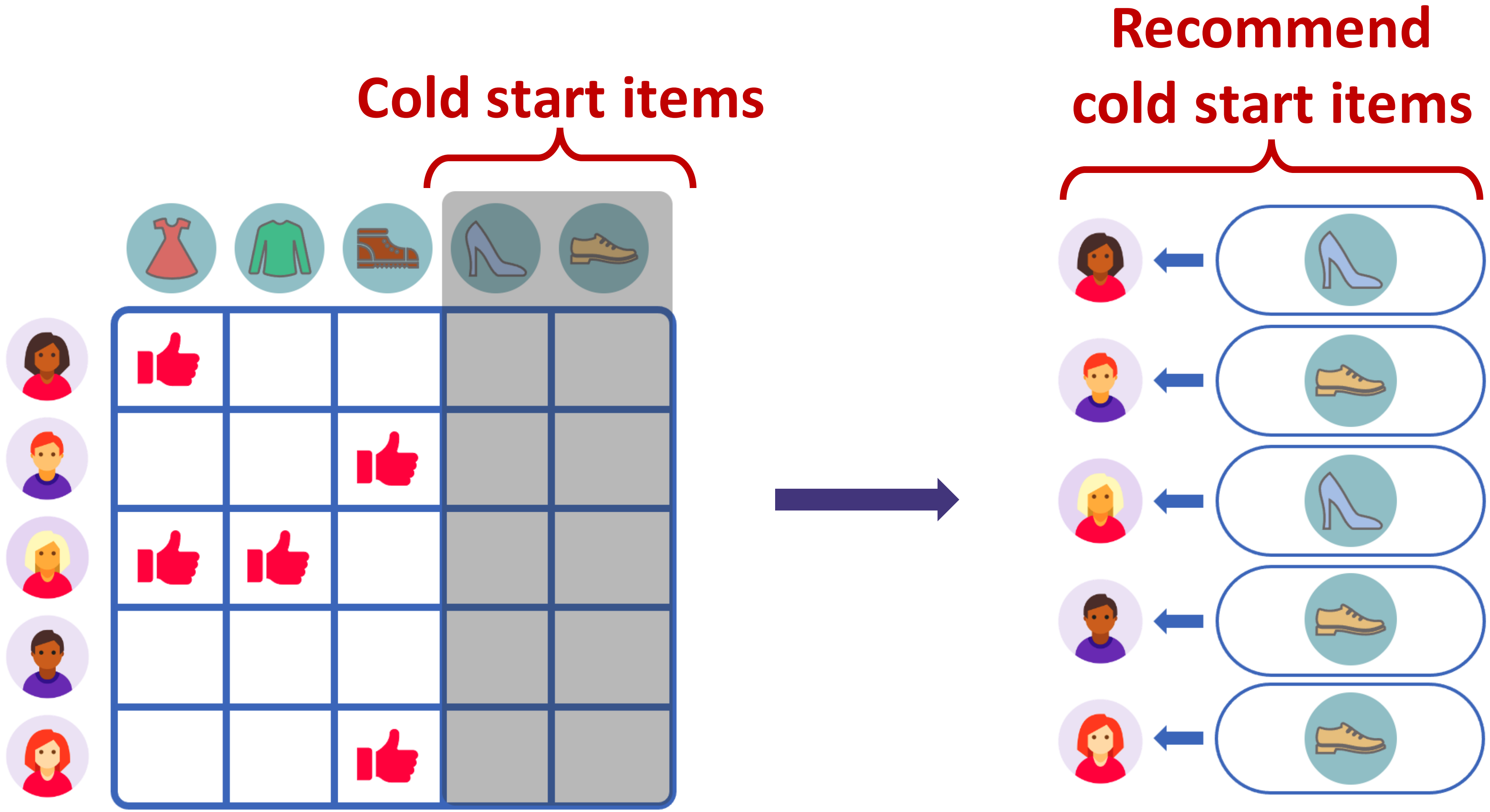
Formalize cold start recommendation

How to accurately recommend cold start items, which do not have any historical feedback, to users.

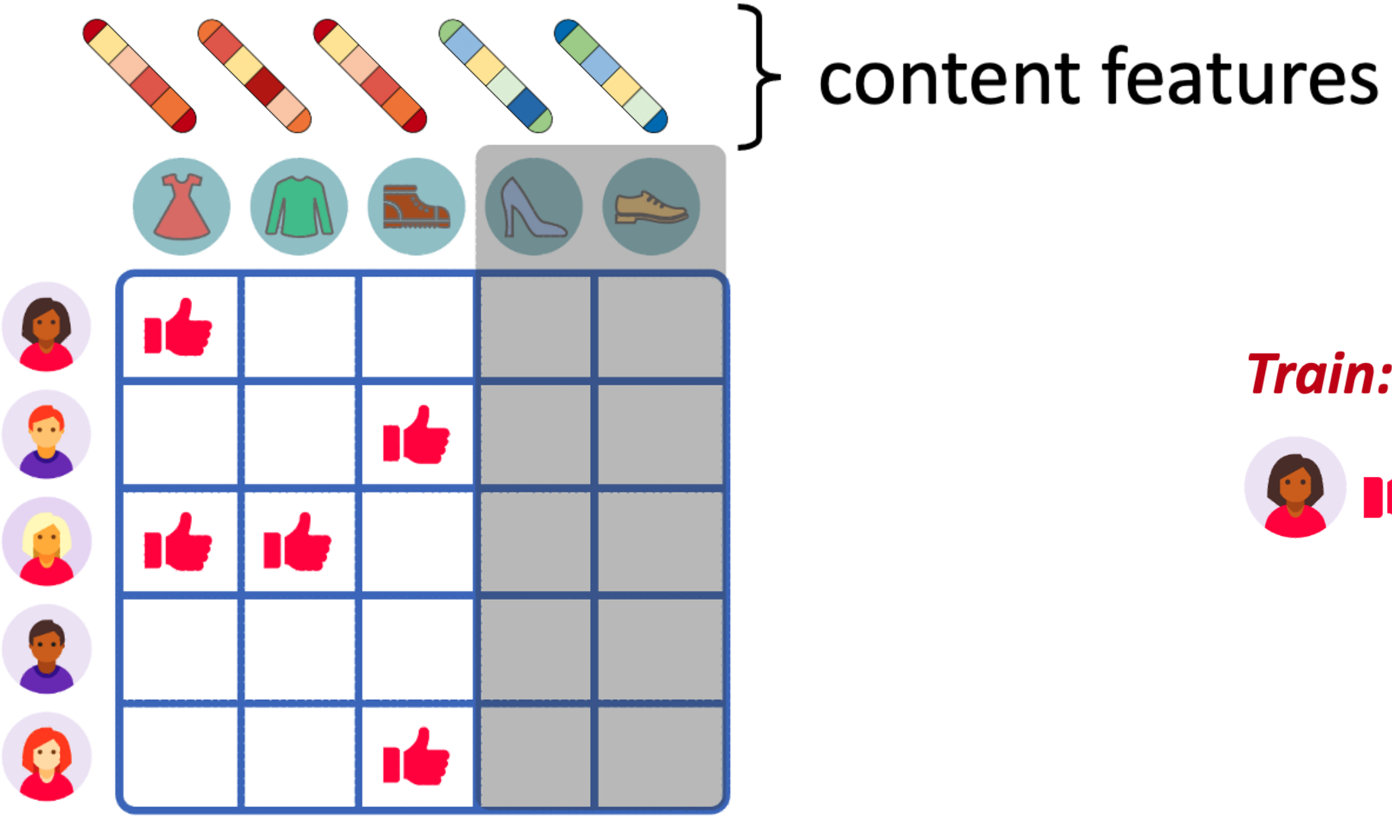


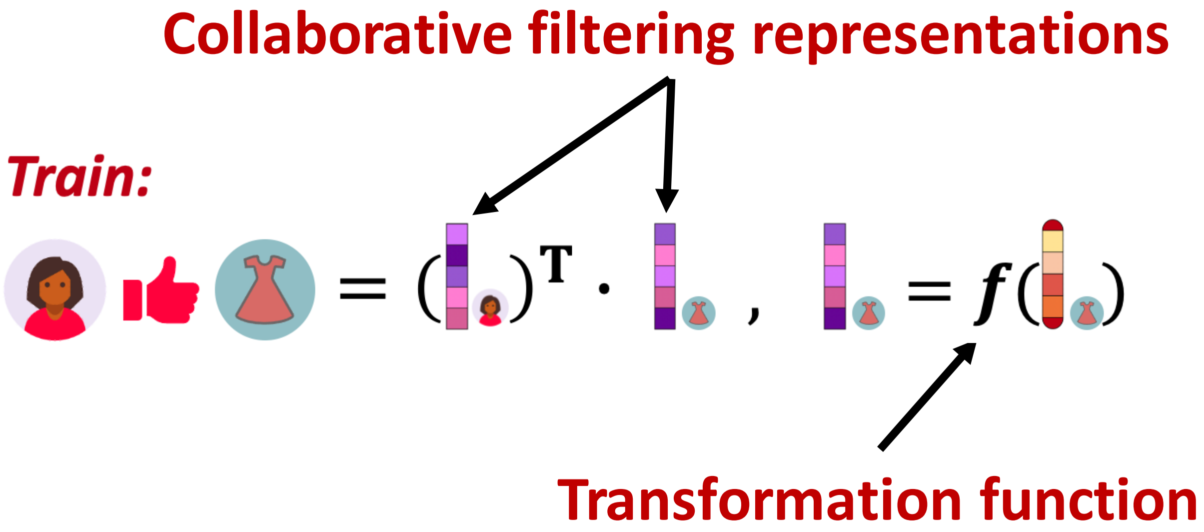
Formalize cold start recommendation

How to accurately recommend cold start items, which do not have any historical feedback, to users.

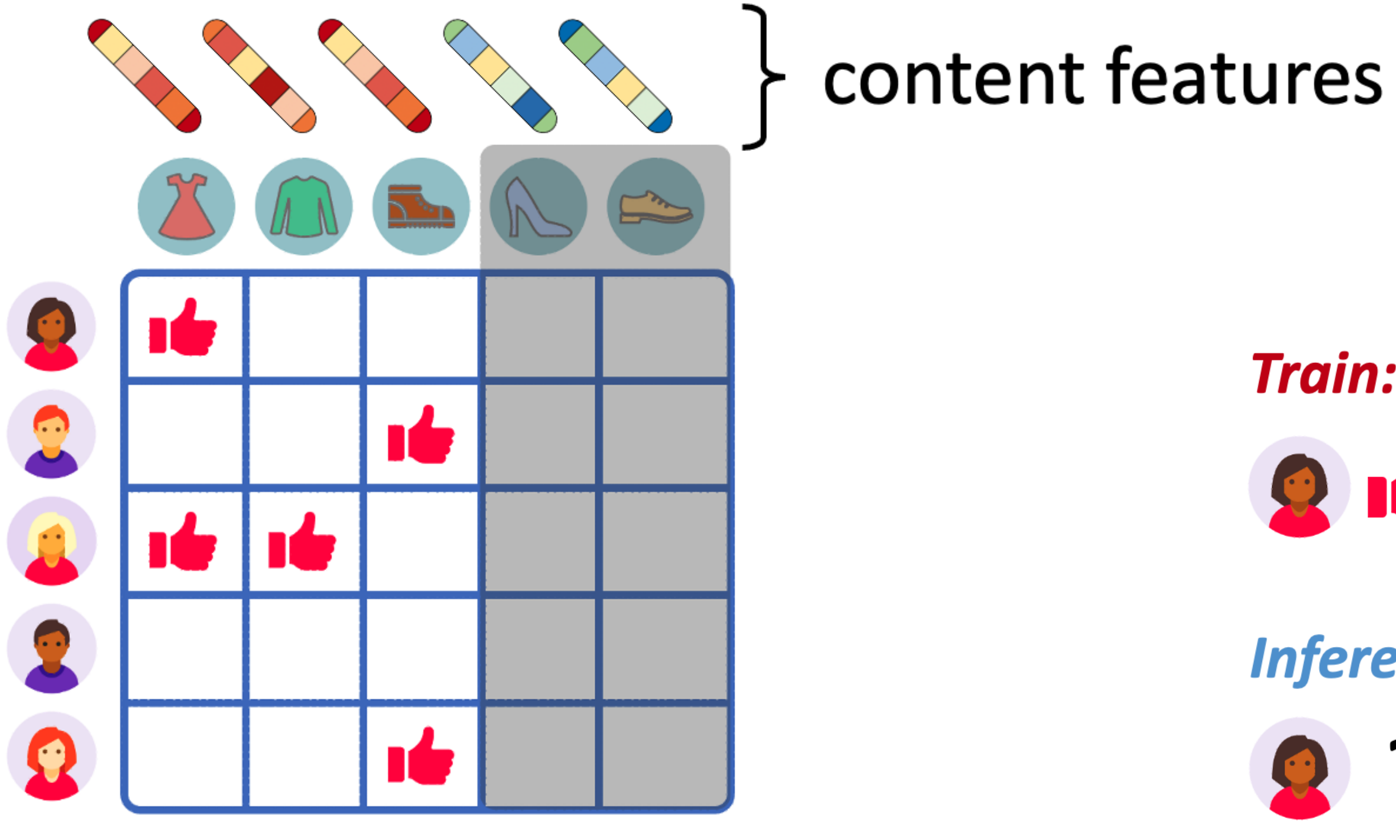


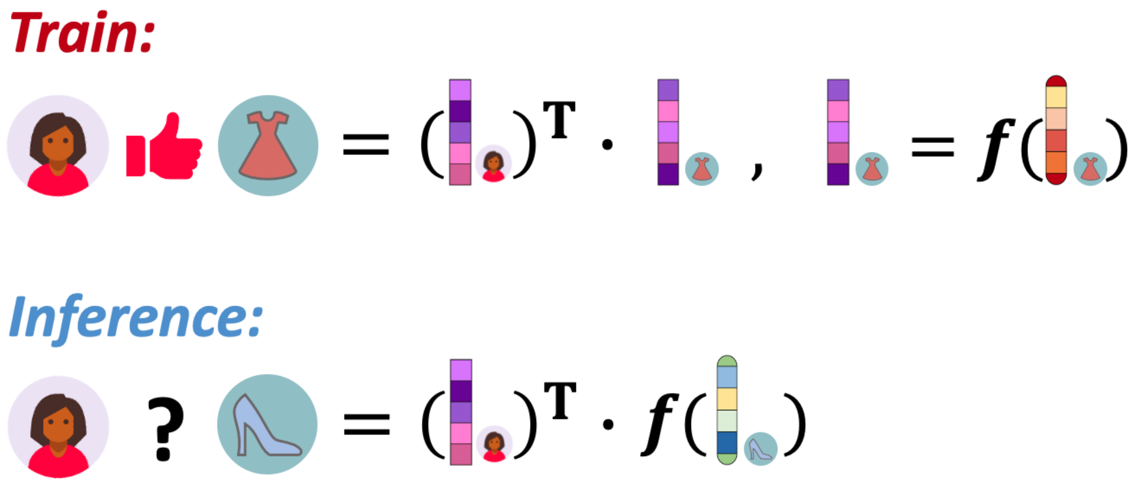
Formalize cold start recommendation





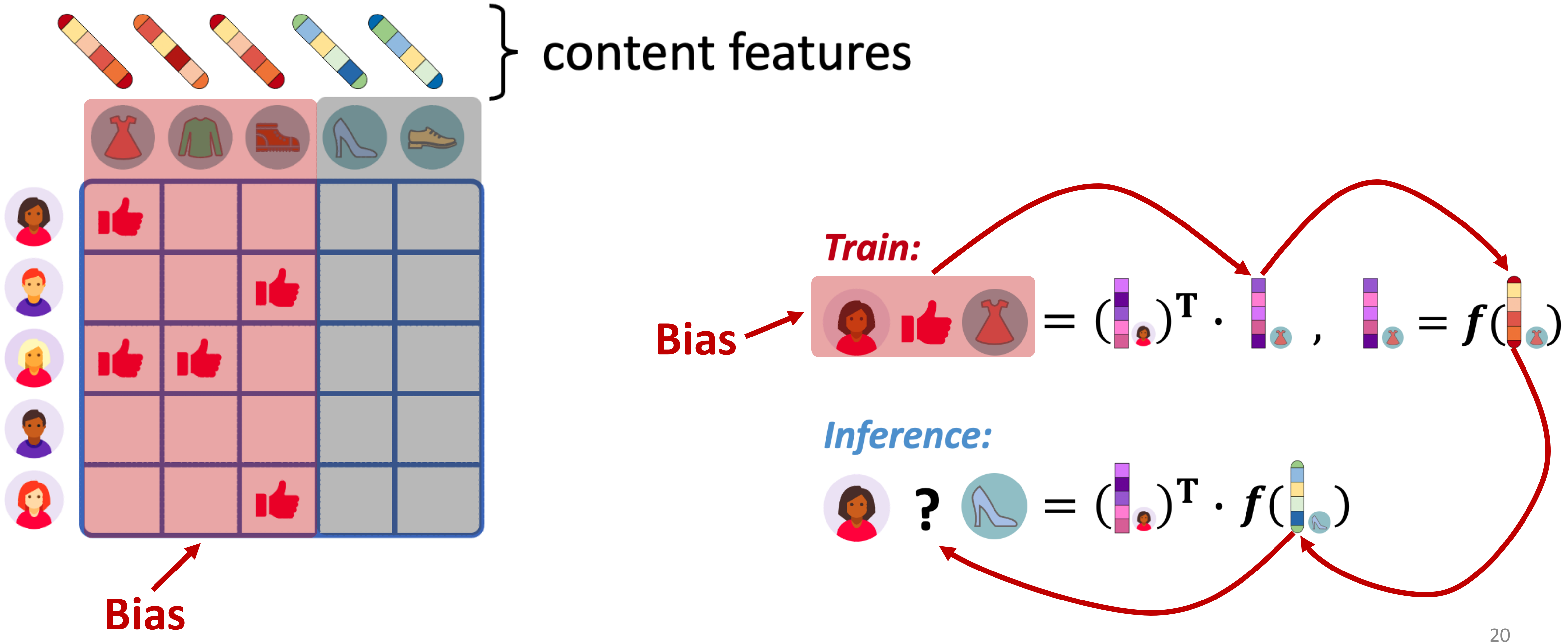
Formalize cold start recommendation





Formalize cold start recommendation

Bias in data for warm start items will be transferred to the recommendations for cold start items through the content features by the learned transformation function.



Formalize Fairness

Following the well-known concepts of **equal-opportunity** and **Rawlsian Max-Min fairness principle**:

- Measuring fairness: the **true positive rate** of **worst-off** items
- Enhancing fairness: **maximizing** the true positive rate of the worst-off items

Formalize Fairness

True positive rate of an item in recommendation: the expected exposure the item gets to matched users in testing set (users who will click the item once recommended). Define true positive rate metric **Mean Discounted Gain** (MDG) for an item *i:*

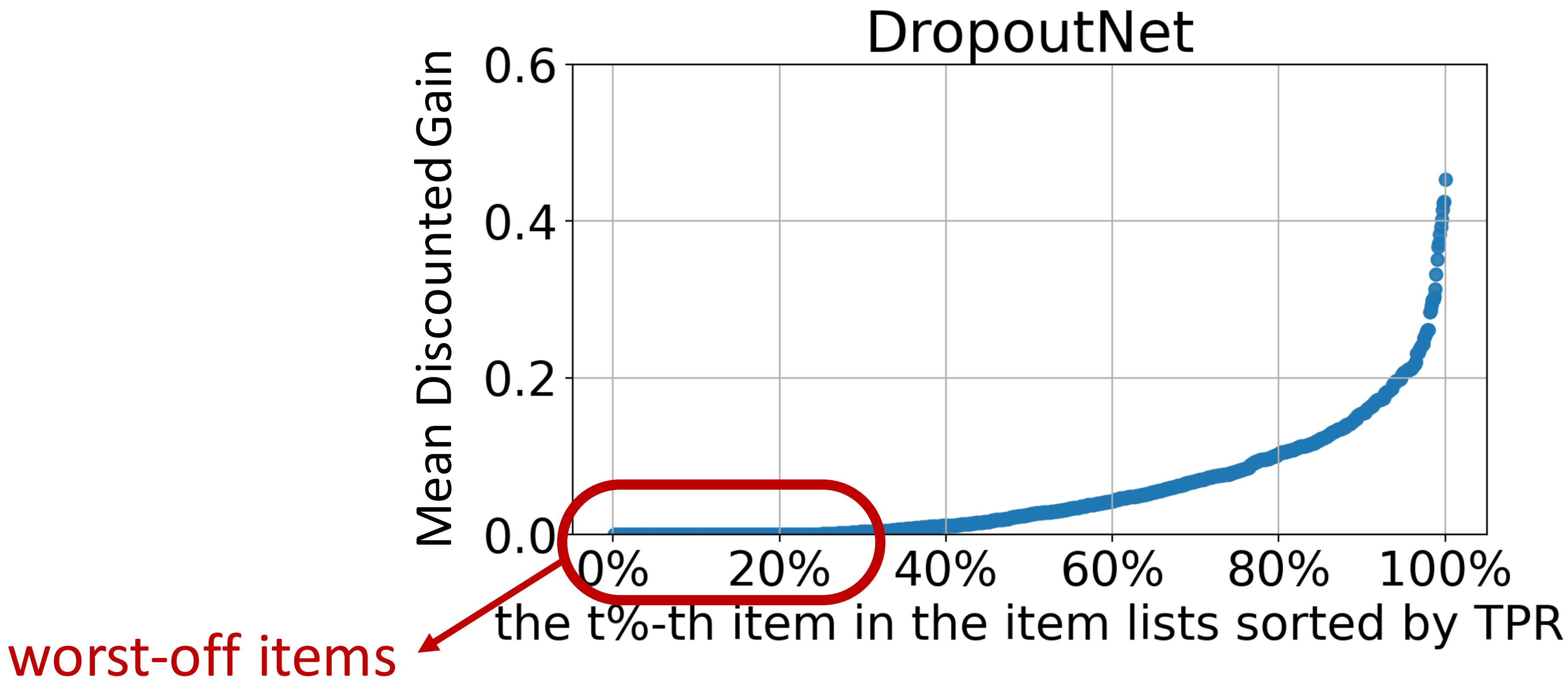
$$MDG_i = \frac{1}{|\mathcal{U}_i^+|} \sum_{u \in \mathcal{U}_i^+} \frac{1}{log(1 + \widehat{z}_{u,i})}$$

matched users of item *i*

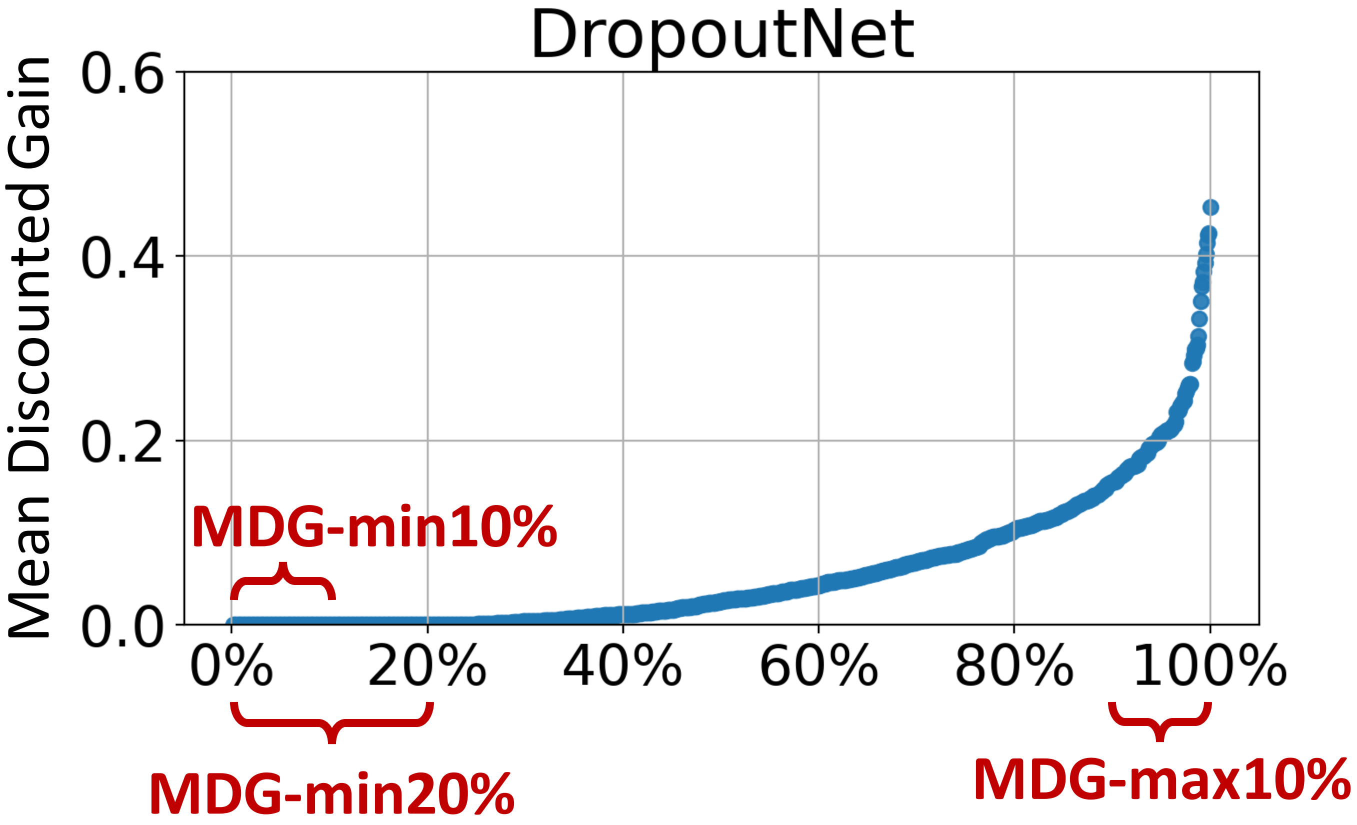
the ranking position of *i* for *u*

Formalize Fairness

- Measuring fairness: the **true positive rate** of **worst-off** items
- Enhancing fairness: **maximizing** the true positive rate of the worst-off items



Formalize Fairness



Measure the fairness by the average MDG of *t%* worst-off items: **MDG-min10%** and **MDG-min20%**. Besides, we also calculate **MDG-max10%** for comparison.

Outline

- Motivations
- Problem Formalization
- **Data-driven Study**
- Fairness-enhancing Approaches
- Fairness-enhancing Experiments

Data-driven study on ML1M

Random ranking

Optimal result using test data

Four SOTA cold start recommendation models

		Heater	DropoutNet	DeepMusic	KNN	Optimal	Random
utility	NDCG@30	0.5332	0.5316	0.5167	0.4226	1.0000	0.0586
Fairness	MDG-min10%	0.	0.	0.	0.0001	0.1388	0.0118
	MDG-min20%	0.	0.	0.0001	0.0020	0.1498	0.0145
	MDG-max10%	0.2272	0.2294	0.2323	0.2091	0.2471	0.0386

Heater: Zhu, Ziwei, et al. "Recommendation for New Users and New Items via Randomized Training and Mixture-of-Experts Transformation." *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020.

DropoutNet: Volkovs, Maksims, Guang Wei Yu, and Tomi Poutanen. "DropoutNet: Addressing Cold Start in Recommender Systems." *NIPS*. 2017.

DeepMusic: Van Den Oord, Aäron, Sander Dieleman, and Benjamin Schrauwen. "Deep content-based music recommendation." *Neural Information Processing Systems Conference (NIPS 2013)*. Vol. 26. Neural Information Processing Systems Foundation (NIPS), 2013.

KNN: Sedhain, Suvash, et al. "Social collaborative filtering for cold-start recommendations." *Proceedings of the 8th ACM Conference on Recommender systems*. 2014.

Data-driven study on ML1M

Four SOTA cold-start recommendation models produce **near-zero** MDG for 10% and 20% worst-off items.

		Heater	DropoutNet	DeepMusic	KNN	Optimal	Random
utility	NDCG@30	0.5332	0.5316	0.5167	0.4226	1.0000	0.0586
Fairness	MDG-min10%	0.	0.	0.	0.0001	0.1388	0.0118
	MDG-min20%	0.	0.	0.0001	0.0020	0.1498	0.0145
	MDG-max10%	0.2272	0.2294	0.2323	0.2091	0.2471	0.0386

Data-driven study on ML1M

Big gap between **MDG-max10%** and MDG-min *t*%.

		Heater	DropoutNet	DeepMusic	KNN	Optimal	Random
utility	NDCG@30	0.5332	0.5316	0.5167	0.4226	1.0000	0.0586
Fairness	MDG-min10%	0.	0.	0.	0.0001	0.1388	0.0118
	MDG-min20%	0.	0.	0.0001	0.0020	0.1498	0.0145
	MDG-max10%	$\boxed{0.2272}$	$\boxed{0.2294}$	$\boxed{0.2323}$	$\boxed{0.2091}$	0.2471	0.0386

Data-driven study on ML1M

Personalized models are even worse than **Random** method.

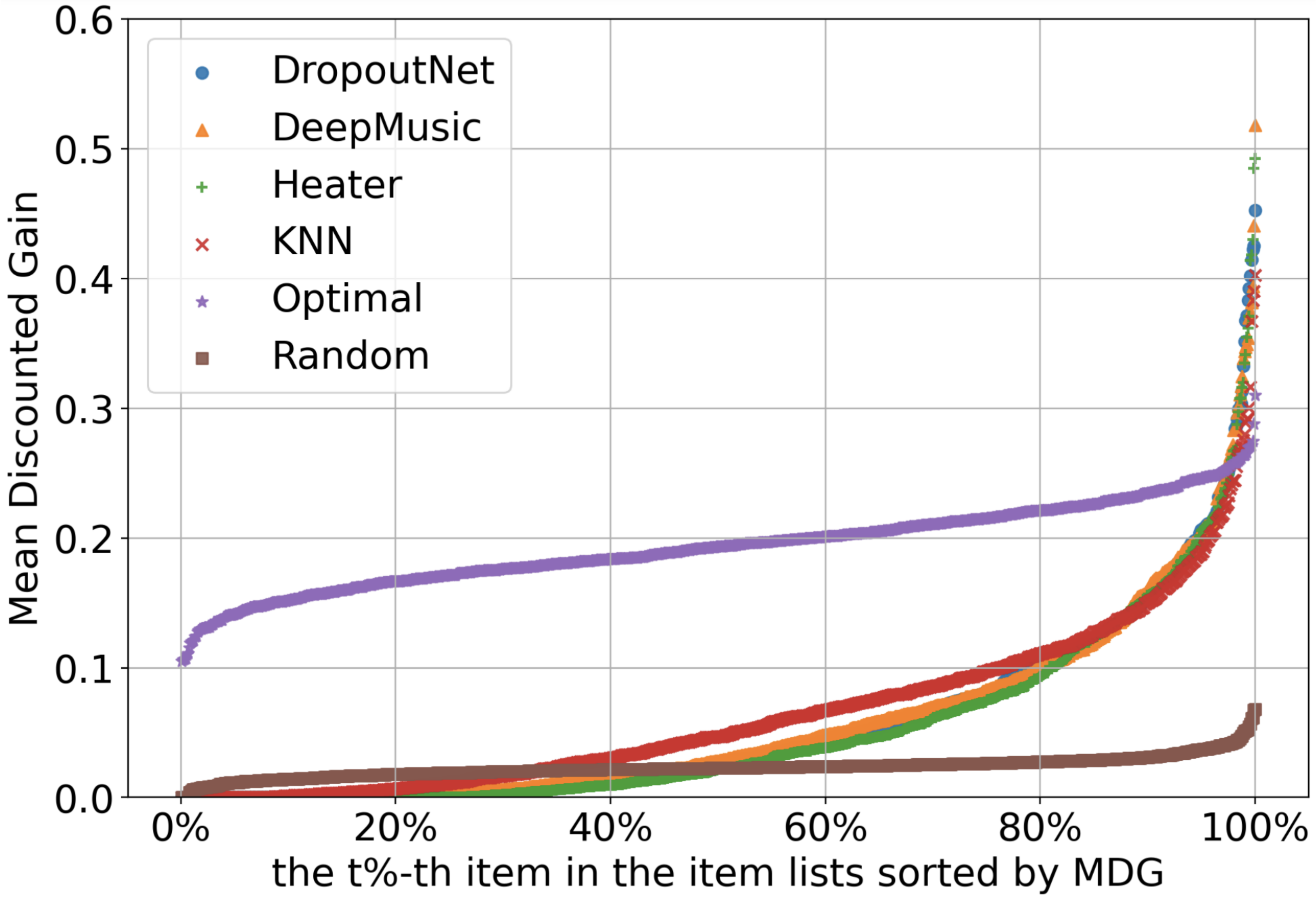
		Heater	DropoutNet	DeepMusic	KNN	Optimal	Random
utility	NDCG@30	0.5332	0.5316	0.5167	0.4226	1.0000	0.0586
Fairness	MDG-min10%	0.	0.	0.	0.0001	0.1388	0.0118
	MDG-min20%	0.	0.	0.0001	0.0020	0.1498	0.0145
	MDG-max10%	0.2272	0.2294	0.2323	0.2091	0.2471	0.0386

Data-driven study on ML1M

Result of Optimal method shows the goal.

		Heater	DropoutNet	DeepMusic	KNN	Optimal	Random
utility	NDCG@30	0.5332	0.5316	0.5167	0.4226	1.0000	0.0586
Fairness	MDG-min10%	0.	0.	0.	0.0001	0.1388	0.0118
	MDG-min20%	0.	0.	0.0001	0.0020	0.1498	0.0145
	MDG-max10%	0.2272	0.2294	0.2323	0.2091	0.2471	0.0386

Data-driven study on ML1M



Outline

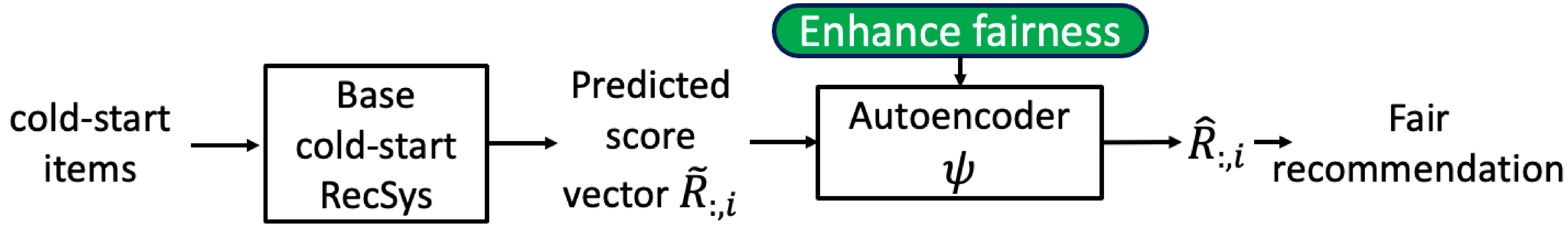
- Motivations
- Problem Formalization
- Data-driven Study
- **Fairness-enhancing Approaches**
  - **Learnable post-processing framework**
  - **A score scaling method**
  - **A joint-learning generative method**
- Fairness-enhancing Experiments

Three ways to enhance fairness

- Pre-processing (data augmentation)
  - Model agnostic; Challenging; need to re-train existing models.
- In-processing
  - Promising performance; coupled with specific models; need to re-train existing models;
- Post-processing (heuristic re-ranking)
  - Flexible to be applied to existing models; limited performance.

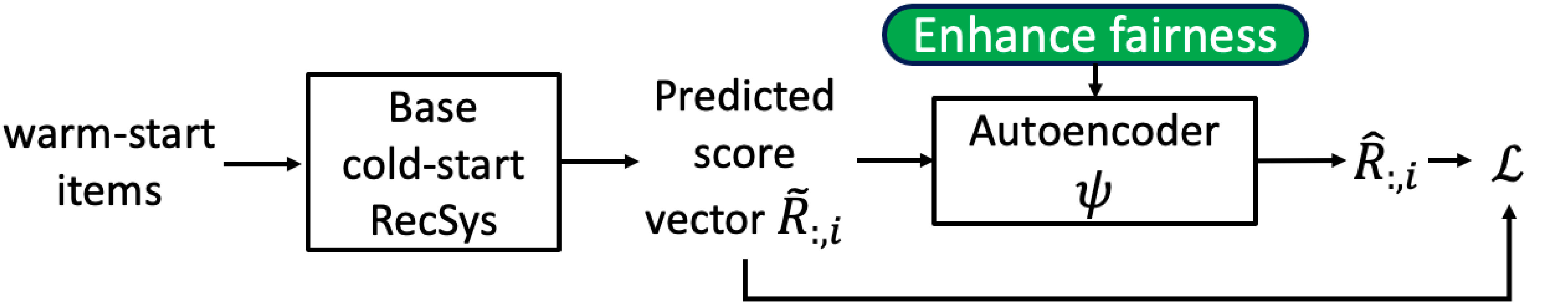
Learnable post-processing framework

During inference:



Learnable post-processing framework

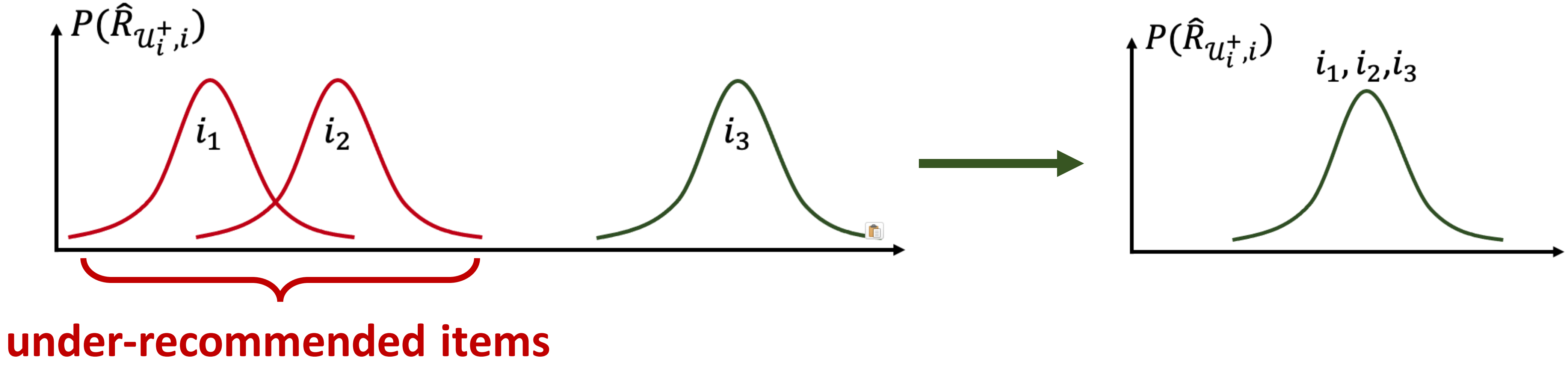
During training:



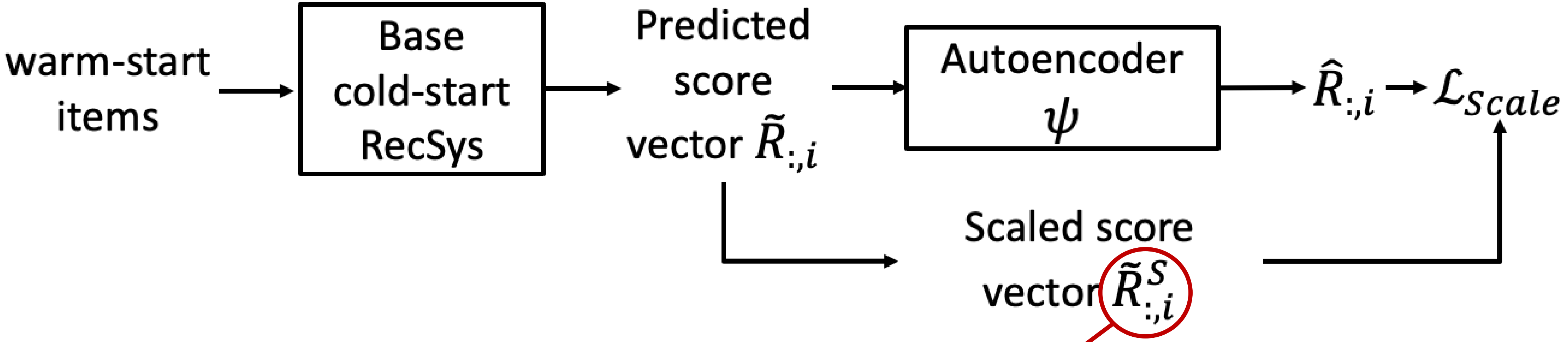
Intuition: how to enhance fairness

- Enhancing fairness: **maximizing** the true positive rate of the worst-off items

- During training, the distribution of the predicted scores for matched users $P(\hat{R}_{\mathcal{U}_i^+,i})$ (or noted as $P(\hat{R}_{u,i}|R_{u,i}=1)$) to be the same across items.



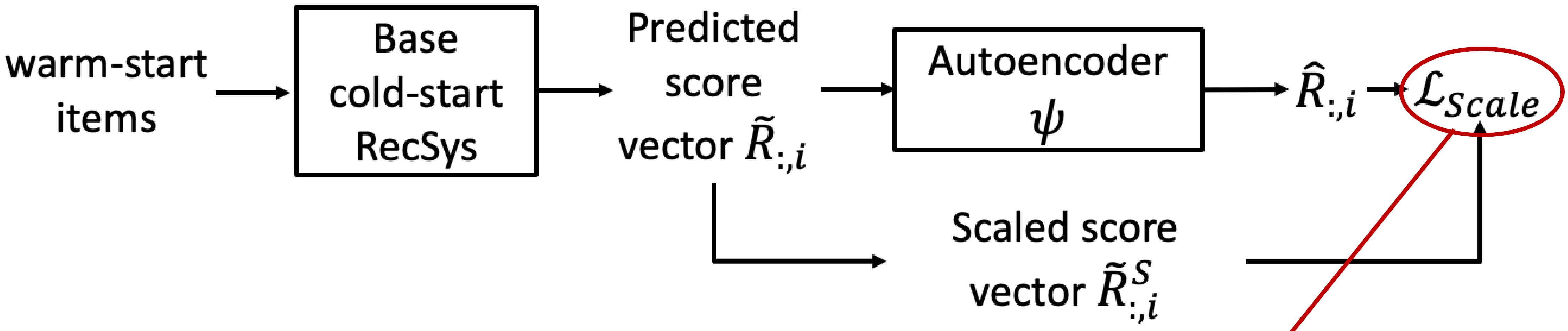
Score scaling method



$$\tilde{R}^{S}_{:,i} = \beta_i \times \tilde{R}_{:,i} \qquad \beta_i \propto \frac{1}{Mean(\tilde{R}_{\mathcal{U}_i^+,i})}$$

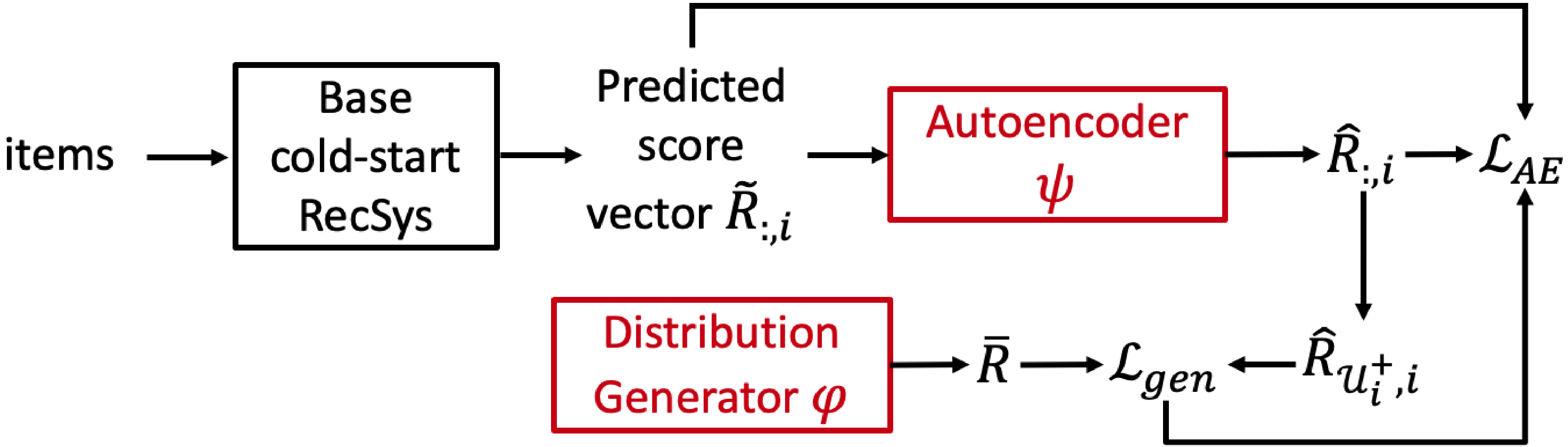
Scale the scores based on the mean of the original scores for matched users

Score scaling method

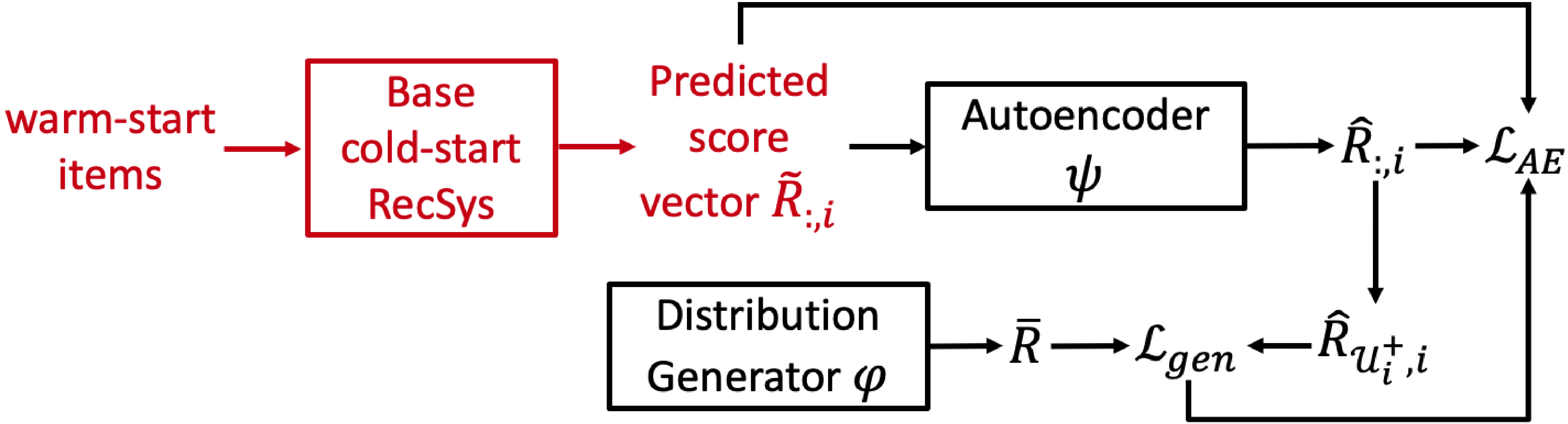


$$\min_{\psi} \mathcal{L}_{Scale} = \sum_{i \in \mathcal{I}_w} \|\widetilde{R}^S_{:,i} - \widehat{R}_{:,i}\|_{\mathrm{F}} + \lambda \|\psi\|_{\mathrm{F}}$$

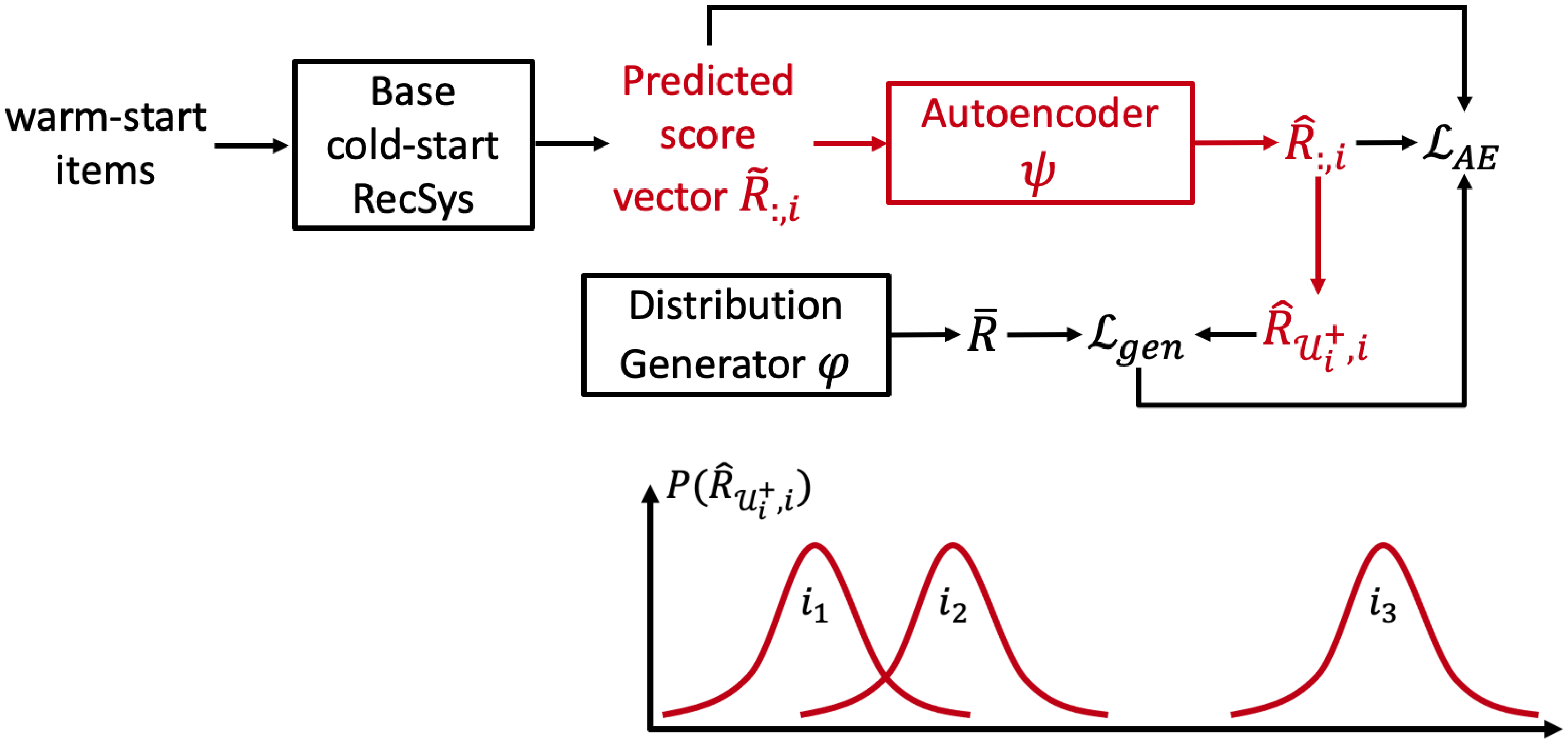
Joint-learning generative method



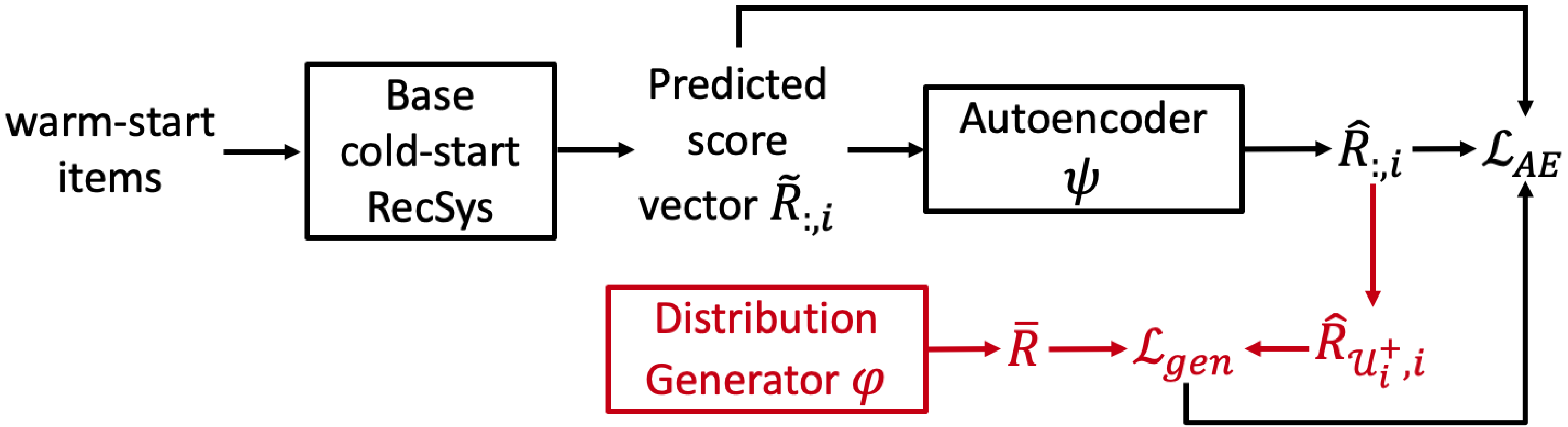
Joint-learning generative method



Joint-learning generative method

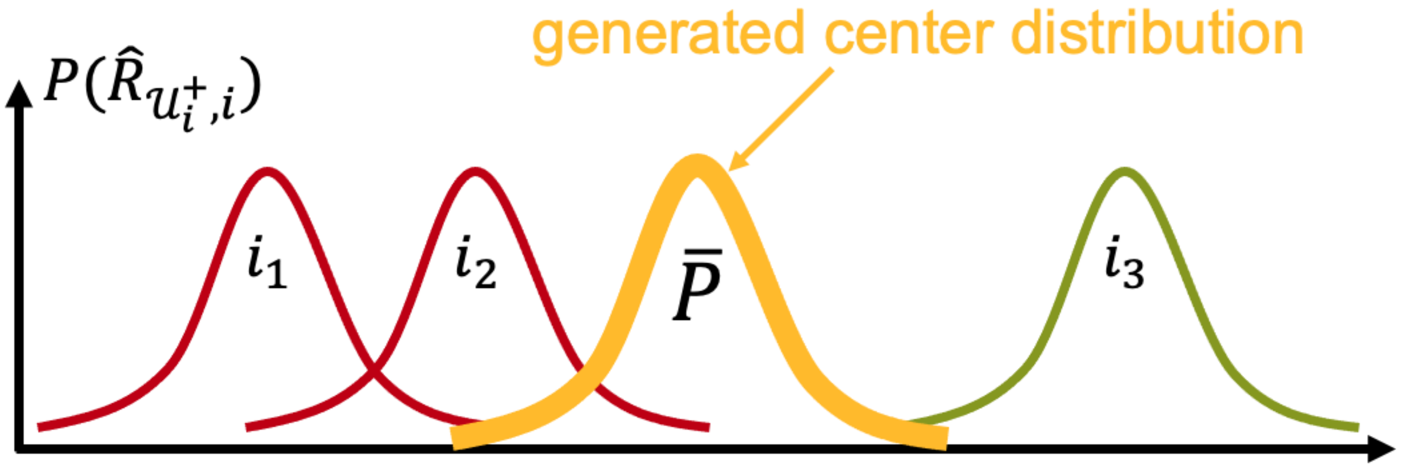


Joint-learning generative method

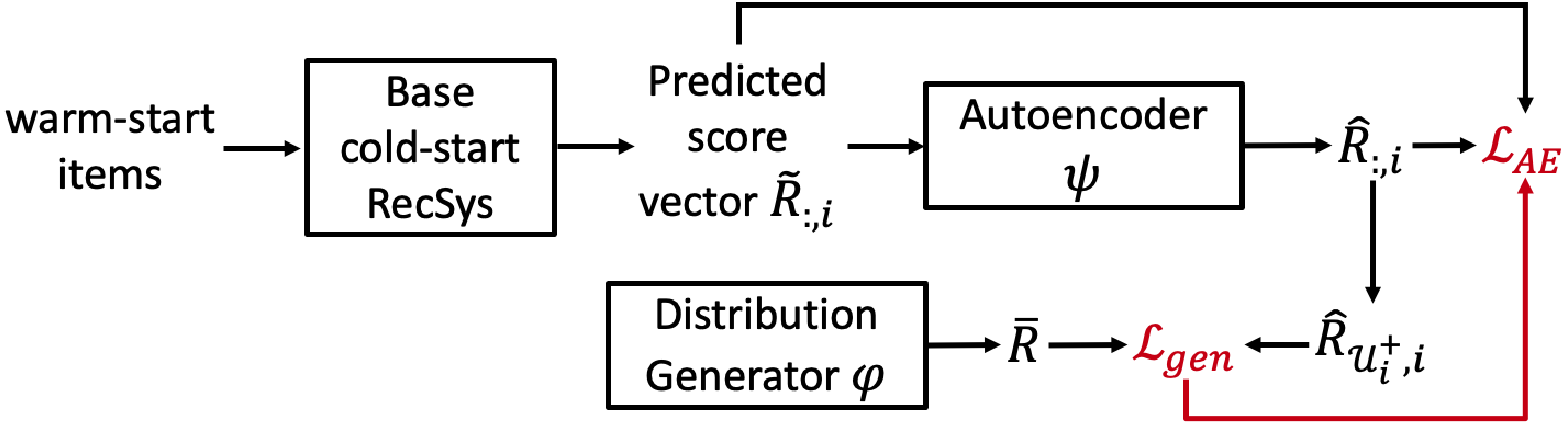


$$\min_{\varphi} \mathcal{L}_{gen} = \sum_{i\in\mathcal{I}_w} \underline{MMD}(\bar{R}, \hat{R}_{\mathcal{U}_i^+,i})$$

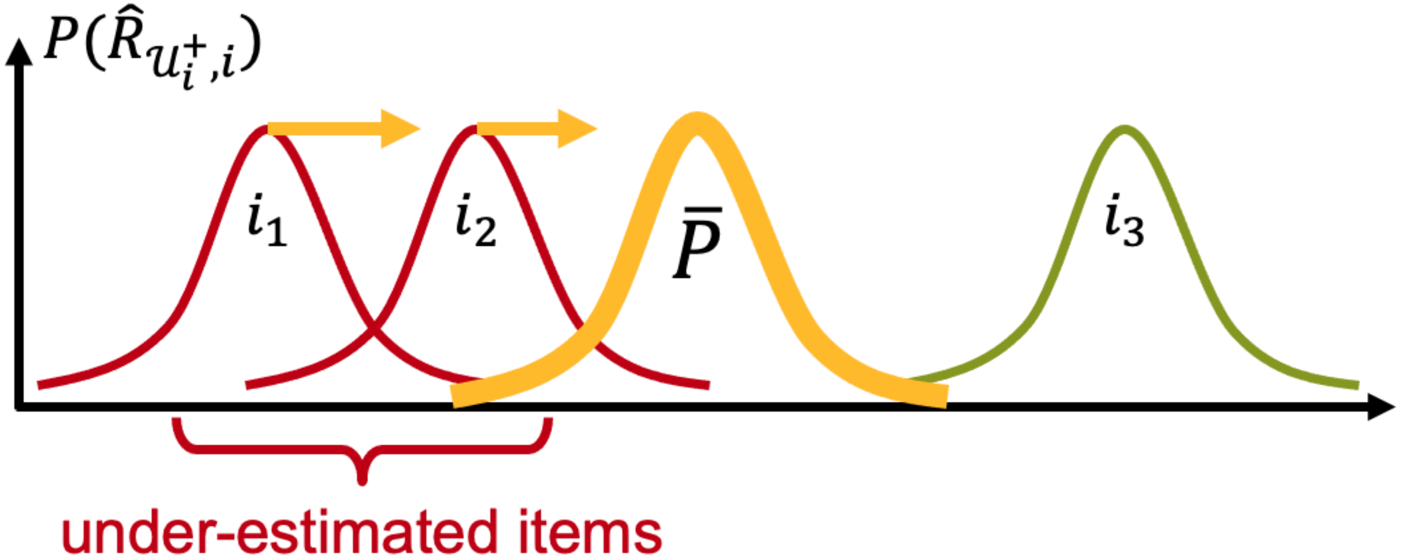
Maximum Mean Discrepency



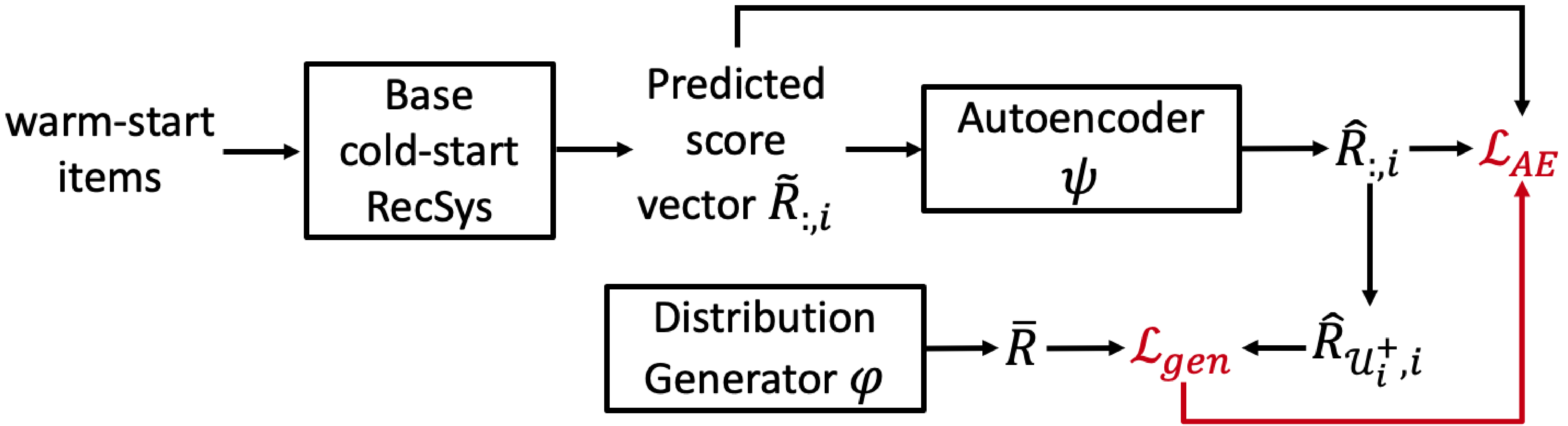
Joint-learning generative method



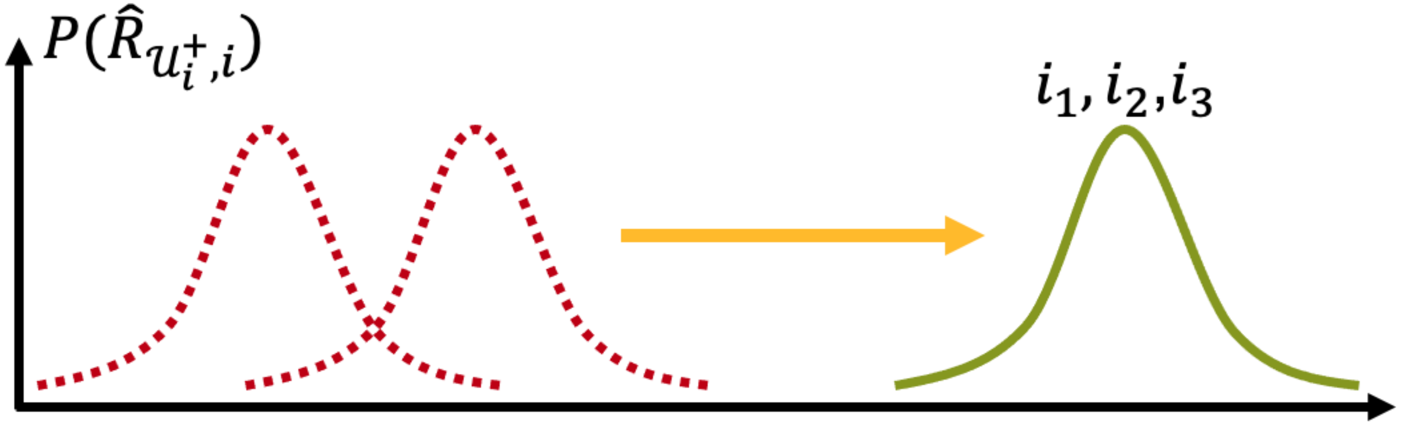
$$\min_{\psi} \sum_{i\in\mathcal{I}_w} MMD(\bar{R}, \widehat{R}_{\mathcal{U}_i^+,i}) \cdot \underbrace{\delta(i \in \mathcal{I}_{UE})}_{\text{under-estimated items}}$$



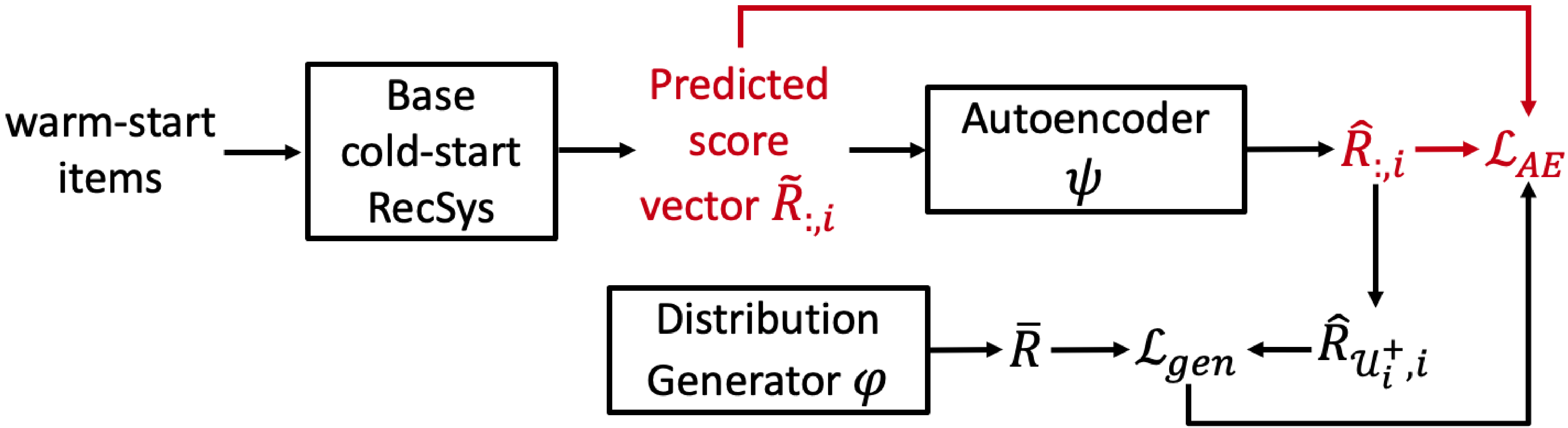
Joint-learning generative method



$$\min_{\psi} \sum_{i \in \mathcal{I}_w} MMD(\bar{R}, \hat{R}_{\mathcal{U}_i^+,i}) \cdot \delta(i \in \mathcal{I}_{UE})$$

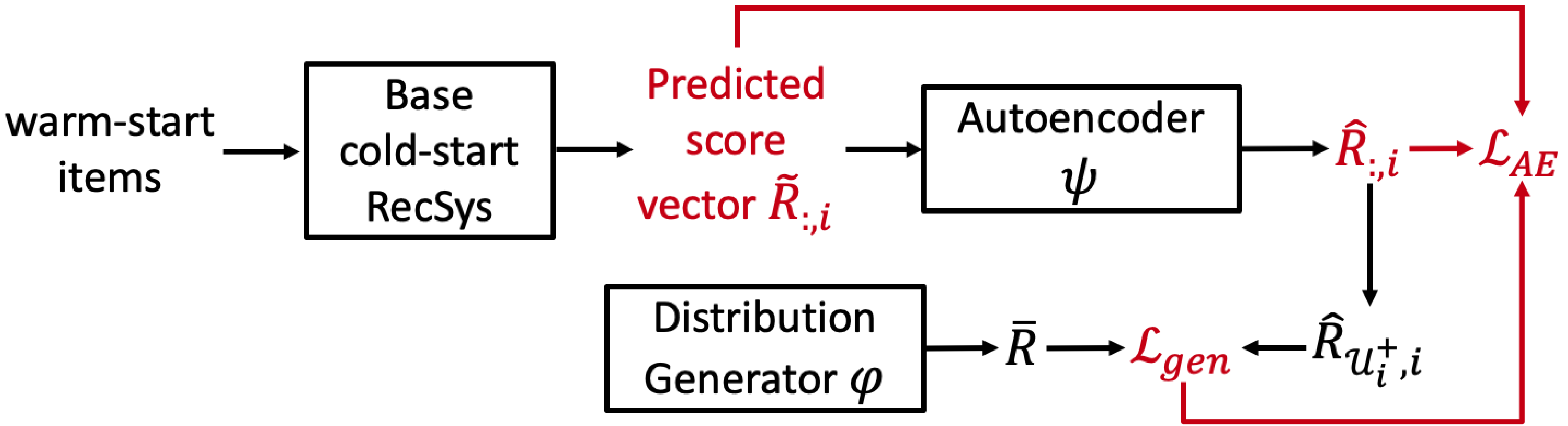


Joint-learning generative method



$$\min_{\psi} \sum_{i \in \mathcal{I}_w} \|\widetilde{R}_{:,i} - \widehat{R}_{:,i}\|_{\mathrm{F}}$$

Joint-learning generative method



$$\min_{\psi} \mathcal{L}_{AE} = \sum_{i \in \mathcal{I}_w} \left(\|\widetilde{R}_{:,i} - \widehat{R}_{:,i}\|_{\mathrm{F}} + \alpha \left(MMD(\overline{R}, \widehat{R}_{\mathcal{U}_i^+,i}) \cdot \delta(i \in \mathcal{I}_{UE}) \right) \right)$$

Outline

- Motivations
- Problem Formalization
- Data-driven Study
- Fairness-enhancing Approaches
- ➢ **Fairness-enhancing Experiments**

Fairness-enhancing experiment

Different cold-start recommendation models as base model

	NDCG @30	Fairness: MDG		
		min10%	min20%	max10%
Heater	0.5332	0.	0.	0.2272
Noise	0.4084	0.0017	0.0046	0.1730
Scale	0.5135	0.0015	0.0066	0.2025
Gen	0.5206	0.0073	0.0136	0.2036
DropoutNet	0.5316	0.	0.	0.2294
Noise	0.4420	0.0010	0.0037	0.1876
Scale	0.5150	0.0015	0.0069	0.2057
Gen	0.5175	0.0075	0.0138	0.2055
DeepMusic	0.5167	0.	0.0001	0.2323
Noise	0.4304	0.0007	0.0032	0.1937
Scale	0.4946	0.0010	0.0047	0.2140
Gen	0.5024	0.0027	0.0071	0.2136
KNN	0.4226	0.0001	0.0020	0.2091
Noise	0.3378	0.0016	0.0053	0.1643
Scale	0.4027	0.0023	0.0084	0.1791
Gen	0.4002	0.0075	0.0140	0.1831

Fairness-enhancing experiment

Baseline: add random noise to scores to improve fairness

	NDCG @30	Fairness: MDG		
		min10%	min20%	max10%
Heater	0.5332	0.	0.	0.2272
Noise	0.4084	0.0017	0.0046	0.1730
Scale	0.5135	0.0015	0.0066	0.2025
Gen	0.5206	0.0073	0.0136	0.2036
DropoutNet	0.5316	0.	0.	0.2294
Noise	0.4420	0.0010	0.0037	0.1876
Scale	0.5150	0.0015	0.0069	0.2057
Gen	0.5175	0.0075	0.0138	0.2055
DeepMusic	0.5167	0.	0.0001	0.2323
Noise	0.4304	0.0007	0.0032	0.1937
Scale	0.4946	0.0010	0.0047	0.2140
Gen	0.5024	0.0027	0.0071	0.2136
KNN	0.4226	0.0001	0.0020	0.2091
Noise	0.3378	0.0016	0.0053	0.1643
Scale	0.4027	0.0023	0.0084	0.1791
Gen	0.4002	0.0075	0.0140	0.1831

Fairness-enhancing experiment

Score scaling method

Joint-learning generative method

	NDCG @30	Fairness: MDG		
		min10%	min20%	max10%
Heater	0.5332	0.	0.	0.2272
Noise	0.4084	0.0017	0.0046	0.1730
$\boxed{\text{Scale}}$	0.5135	0.0015	0.0066	0.2025
$\boxed{\text{Gen}}$	0.5206	0.0073	0.0136	0.2036
DropoutNet	0.5316	0.	0.	0.2294
Noise	0.4420	0.0010	0.0037	0.1876
Scale	0.5150	0.0015	0.0069	0.2057
Gen	0.5175	0.0075	0.0138	0.2055
DeepMusic	0.5167	0.	0.0001	0.2323
Noise	0.4304	0.0007	0.0032	0.1937
Scale	0.4946	0.0010	0.0047	0.2140
Gen	0.5024	0.0027	0.0071	0.2136
KNN	0.4226	0.0001	0.0020	0.2091
Noise	0.3378	0.0016	0.0053	0.1643
Scale	0.4027	0.0023	0.0084	0.1791
Gen	0.4002	0.0075	0.0140	0.1831

Fairness-enhancing experiment

	NDCG @30	Fairness: MDG		
		min10%	min20%	max10%
Heater	0.5332	0.	0.	0.2272
Noise	0.4084	0.0017	0.0046	0.1730
Scale	0.5135	0.0015	0.0066	0.2025
Gen	0.5206	0.0073	0.0136	0.2036
DropoutNet	0.5316	0.	0.	0.2294
Noise	0.4420	0.0010	0.0037	0.1876
Scale	0.5150	0.0015	0.0069	0.2057
Gen	0.5175	0.0075	0.0138	0.2055
DeepMusic	0.5167	0.	0.0001	0.2323
Noise	0.4304	0.0007	0.0032	0.1937
Scale	0.4946	0.0010	0.0047	0.2140
Gen	0.5024	0.0027	0.0071	0.2136
KNN	0.4226	0.0001	0.0020	0.2091
Noise	0.3378	0.0016	0.0053	0.1643
Scale	0.4027	0.0023	0.0084	0.1791
Gen	0.4002	0.0075	0.0140	0.1831

Goal: to improve MDG-min10% and MDG-min20%.

Fairness-enhancing experiment

	NDCG @30	Fairness: MDG		
		min10%	min20%	max10%
Heater	0.5332	0.	0.	0.2272
Noise	0.4084	0.0017	0.0046	0.1730
Scale	0.5135	0.0015	0.0066	0.2025
Gen	0.5206	0.0073	0.0136	0.2036
DropoutNet	0.5316	0.	0.	0.2294
Noise	0.4420	0.0010	0.0037	0.1876
Scale	0.5150	0.0015	0.0069	0.2057
Gen	0.5175	0.0075	0.0138	0.2055
DeepMusic	0.5167	0.	0.0001	0.2323
Noise	0.4304	0.0007	0.0032	0.1937
Scale	0.4946	0.0010	0.0047	0.2140
Gen	0.5024	0.0027	0.0071	0.2136
KNN	0.4226	0.0001	0.0020	0.2091
Noise	0.3378	0.0016	0.0053	0.1643
Scale	0.4027	0.0023	0.0084	0.1791
Gen	0.4002	0.0075	0.0140	0.1831

Gen outperforms other methods for enhancing fairness.

Fairness-enhancing experiment

	NDCG @30	Fairness: MDG min10%	Fairness: MDG min20%	Fairness: MDG max10%
Heater	0.5332	0.	0.	0.2272
Noise	0.4084	0.0017	0.0046	0.1730
Scale	0.5135	0.0015	0.0066	0.2025
Gen	0.5206	0.0073	0.0136	0.2036
DropoutNet	0.5316	0.	0.	0.2294
Noise	0.4420	0.0010	0.0037	0.1876
Scale	0.5150	0.0015	0.0069	0.2057
Gen	0.5175	0.0075	0.0138	0.2055
DeepMusic	0.5167	0.	0.0001	0.2323
Noise	0.4304	0.0007	0.0032	0.1937
Scale	0.4946	0.0010	0.0047	0.2140
Gen	0.5024	0.0027	0.0071	0.2136
KNN	0.4226	0.0001	0.0020	0.2091
Noise	0.3378	0.0016	0.0053	0.1643
Scale	0.4027	0.0023	0.0084	0.1791
Gen	0.4002	0.0075	0.0140	0.1831

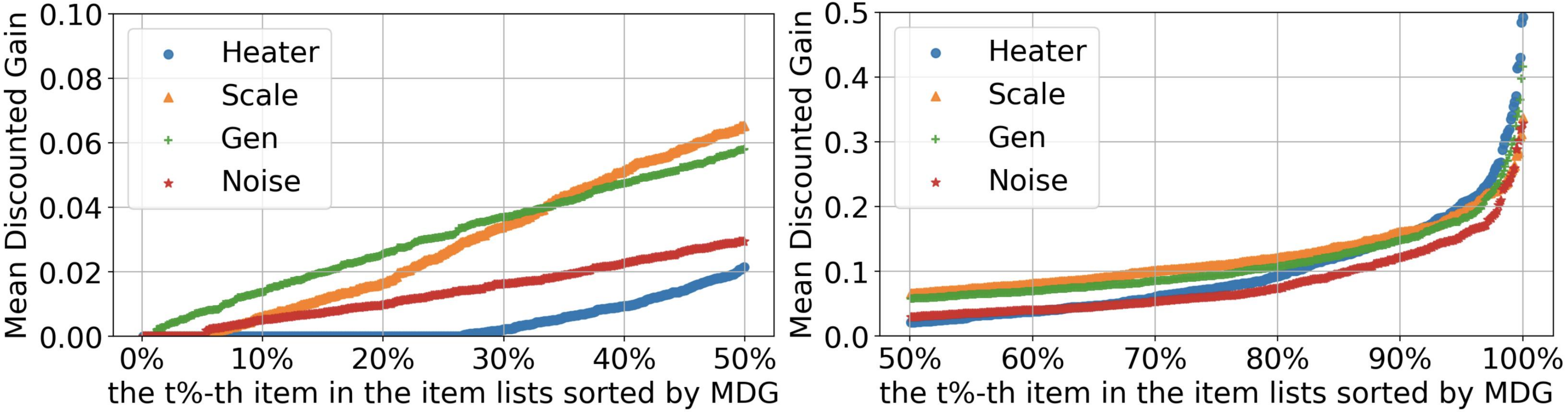
Gen preserves utility for best-served items more effectively.

Fairness-enhancing experiment

	NDCG @30	Fairness: MDG		
		min10%	min20%	max10%
Heater	0.5332	0.	0.	0.2272
Noise	0.4084	0.0017	0.0046	0.1730
Scale	0.5135	0.0015	0.0066	0.2025
Gen	$\boxed{0.5206}$	0.0073	0.0136	0.2036
DropoutNet	0.5316	0.	0.	0.2294
Noise	0.4420	0.0010	0.0037	0.1876
Scale	0.5150	0.0015	0.0069	0.2057
Gen	$\boxed{0.5175}$	0.0075	0.0138	0.2055
DeepMusic	0.5167	0.	0.0001	0.2323
Noise	0.4304	0.0007	0.0032	0.1937
Scale	0.4946	0.0010	0.0047	0.2140
Gen	$\boxed{0.5024}$	0.0027	0.0071	0.2136
KNN	0.4226	0.0001	0.0020	0.2091
Noise	0.3378	0.0016	0.0053	0.1643
Scale	0.4027	0.0023	0.0084	0.1791
Gen	$\boxed{0.4002}$	0.0075	0.0140	0.1831

Gen preserve recommendation utility more effectively.

Fairness-enhancing experiment



Conclusions

- Propose to study the recommendation **fairness among new items** in cold-start RecSys;
- Empirically show the **prevalence of unfairness** among new items in cold-start RecSys;
- Propose the **learnable post-processing framework** as the solution blueprint. And based on the blueprint, we propose the **score scaling method** and **joint-learning generative model** to enhance the fairness;
- Extensive experiment to show the **effectiveness** of the proposed method.

Thank You!

Ziwei Zhu, Jingu Kim*, Trung Nguyen*, Aish Fenton*, and James Caverlee

Texas A&M University

*Netflix