



中国科学技术大学  
University of Science and Technology of China



# Bias Issues and Solutions in Recommender System

Jiawei Chen, Xiang Wang, Fuli Feng, Xiangnan He  
[cjwustc@ustc.edu.cn](mailto:cjwustc@ustc.edu.cn)

slides will be available at: <https://github.com/jiawei-chen/RecDebiasing>

A literature survey based on this tutorial is available at: <https://arxiv.org/pdf/2010.03240.pdf>





## • About Us



**Jiawei Chen**

Postdoc Researcher

University of Science  
and Technology of China

[cjwustc@ustc.edu.cn](mailto:cjwustc@ustc.edu.cn)



**Xiang Wang**

Postdoc Researcher

National University of  
Singapore

[xiangwang@u.nus.edu](mailto:xiangwang@u.nus.edu)



**Fuli Feng**

Postdoc Researcher

National University of  
Singapore

[fulifeng93@gmail.com](mailto:fulifeng93@gmail.com)



**Xiangnan He**

Professor

University of Science  
and Technology of China

[xiangwang@u.nus.edu](mailto:xiangwang@u.nus.edu)

## • Information Seeking

Information explosion problem?

- Information seeking requirements

➤ E-commerce(Amazon and Alibaba)

➤ Social networking(Facebook and Wechat)

➤ Content sharing platforms (Youtube and Pinterest)



**12 million items in Amazon**

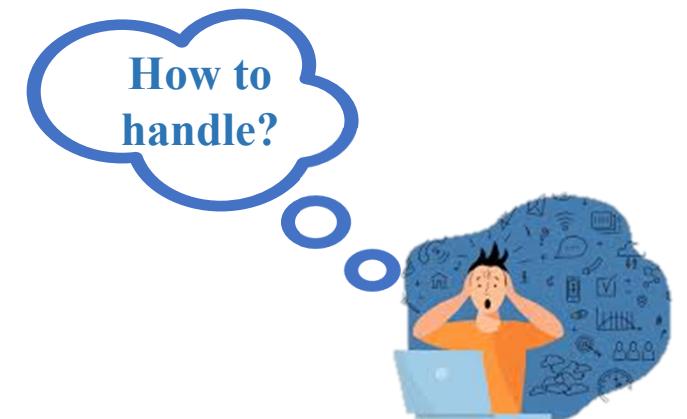
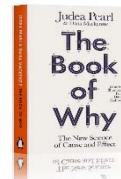
**2.8 billion users in Facebook**

**720,000 hours videos uploaded per day in Youtube**

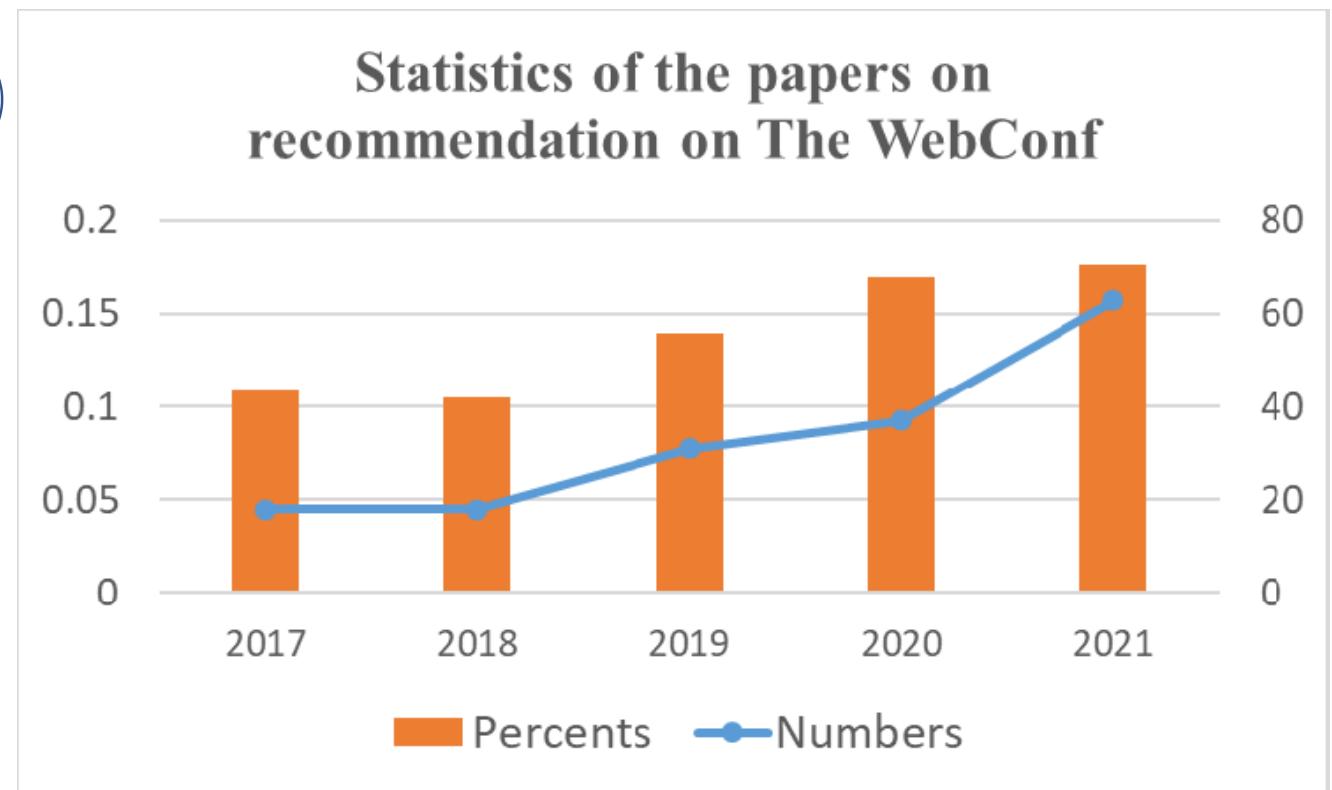
**Recommender system** has been recognized as a powerful tool to address information overload.



*You may like?*



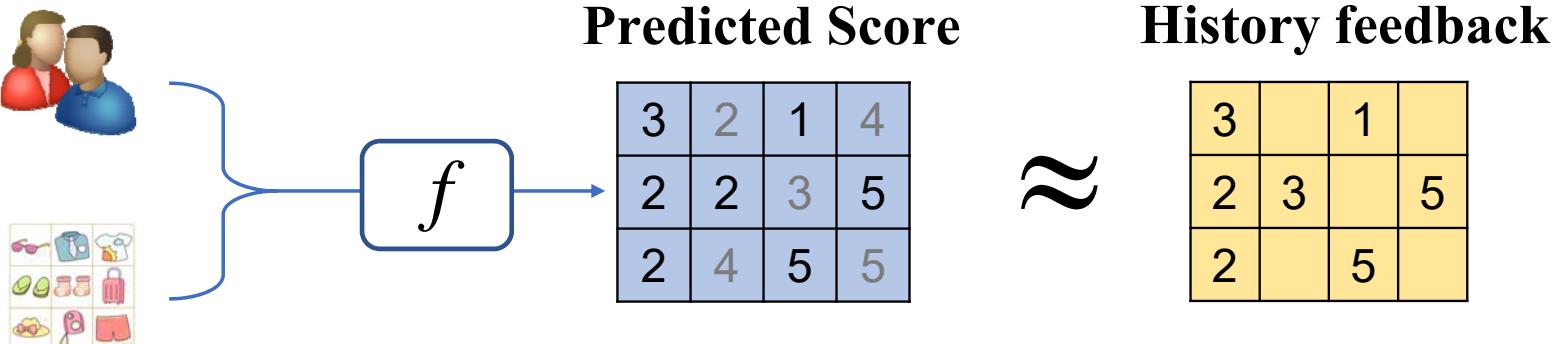
## • Recommendation Has Become Prevalent



## • Classical Problem Setting

- Given:

- a set of users      $U = \{u_1, u_2, \dots, u_n\}$
- a set of items      $I = \{i_1, i_2, \dots, i_m\}$
- users history feedback on items:      $R^o \subseteq \mathbb{R}^{n \times m}$
- To learn a model to predict preference for each user-item pair:      $\hat{R} = f(U, I | \theta)$
- minimizing the difference between the prediction and the observed feedback



## • Mainstream Models

### ➤ Collaborative filtering

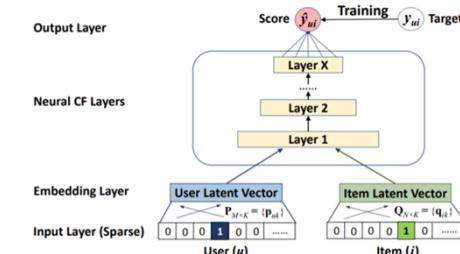
- matrix factorization & factorization machines

Feature vector $\mathbf{x}$										Target $y$										
$x^{(1)}$	1	0	0	...	1	0	0	0	...	0.3	0.3	0.3	0	...	13	0	0	0	0	...
$x^{(2)}$	1	0	0	...	0	1	0	0	...	0.3	0.3	0.3	0	...	14	1	0	0	0	...
$x^{(3)}$	1	0	0	...	0	0	1	0	...	0.3	0.3	0.3	0	...	16	0	1	0	0	...
$x^{(4)}$	0	1	0	...	0	0	1	0	...	0	0	0.5	0.5	...	5	0	0	0	0	...
$x^{(5)}$	0	1	0	...	0	0	0	1	...	0	0	0.5	0.5	...	8	0	0	1	0	...
$x^{(6)}$	0	0	1	...	1	0	0	0	...	0.5	0	0.5	0	...	9	0	0	0	0	...
$x^{(7)}$	0	0	1	...	0	0	1	0	...	0.5	0	0.5	0	...	12	1	0	0	0	...
A	B	C	...	T1	NH	SW	ST	...	T1	NH	SW	ST	...	User	Movie	Time	...	User	Movie	Time

Factorization Machines

### ➤ Deep learning approaches

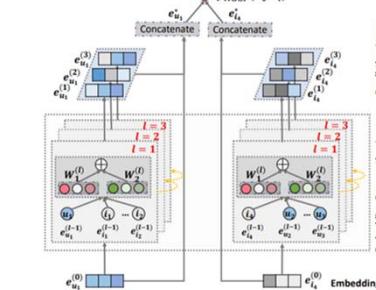
- neural factorization machines & deep interest networks



Neural Collaborative Filtering

### ➤ Graph-based approaches

- leveraging user-item interaction graphs & knowledge graph

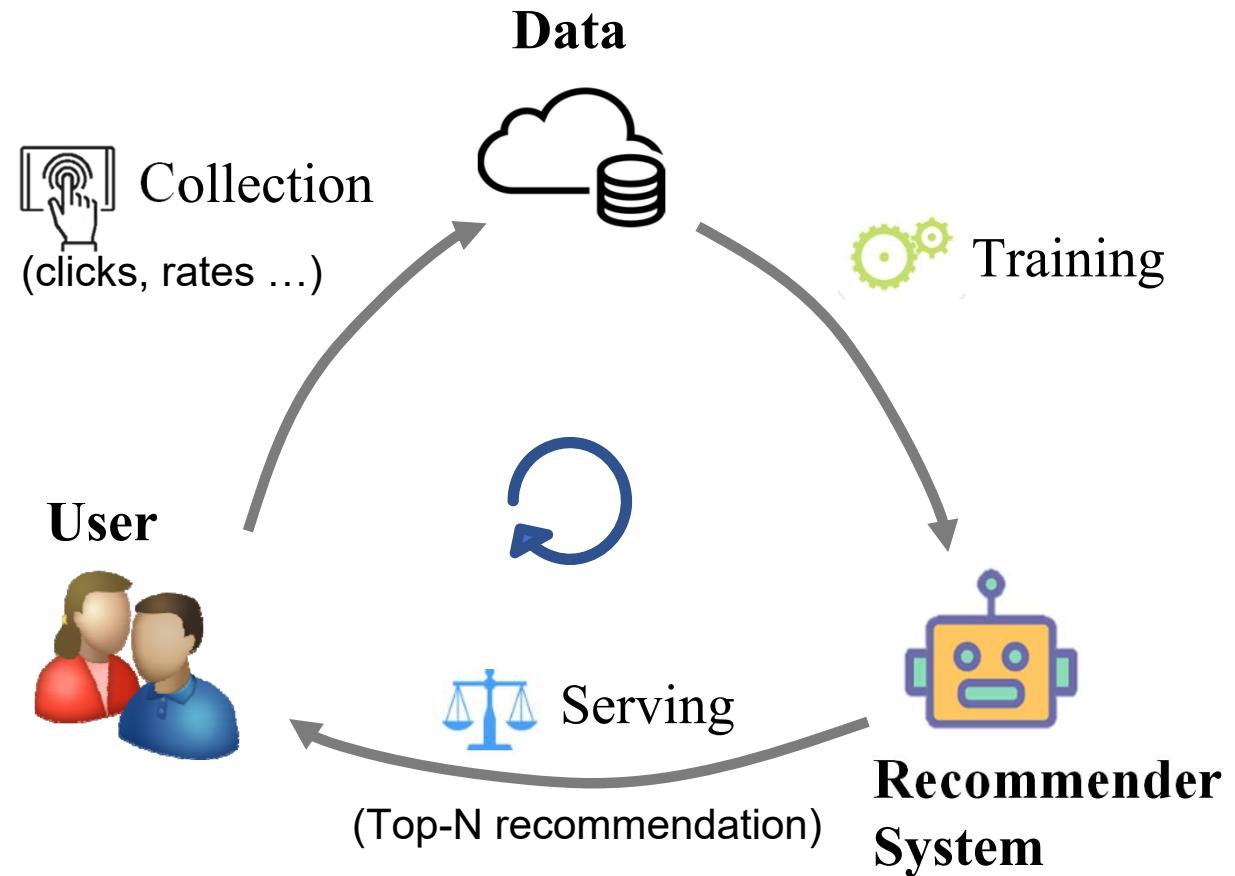


Neural Graph Collaborative Filtering

## • Ecosystem of Recsys

- Working flow of RS

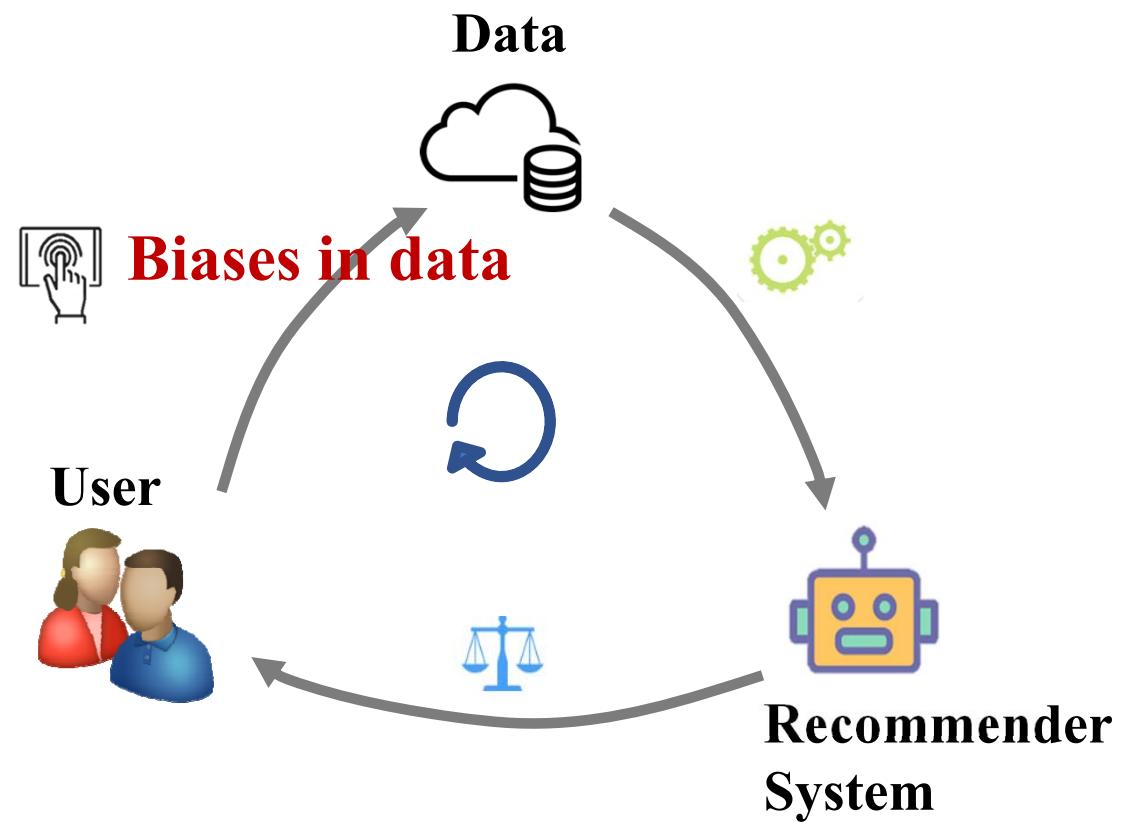
- **Training:** RS are trained or updated on **observed user-item interaction** data.
- **Serving:** RS infers user preference over items and gives **recommendation lists**.
- **Collecting:** User new actions are merged into the **training data**.
- Forming a **feedback loop**



## • Where Bias Comes?

### • Biases in data (Collecting):

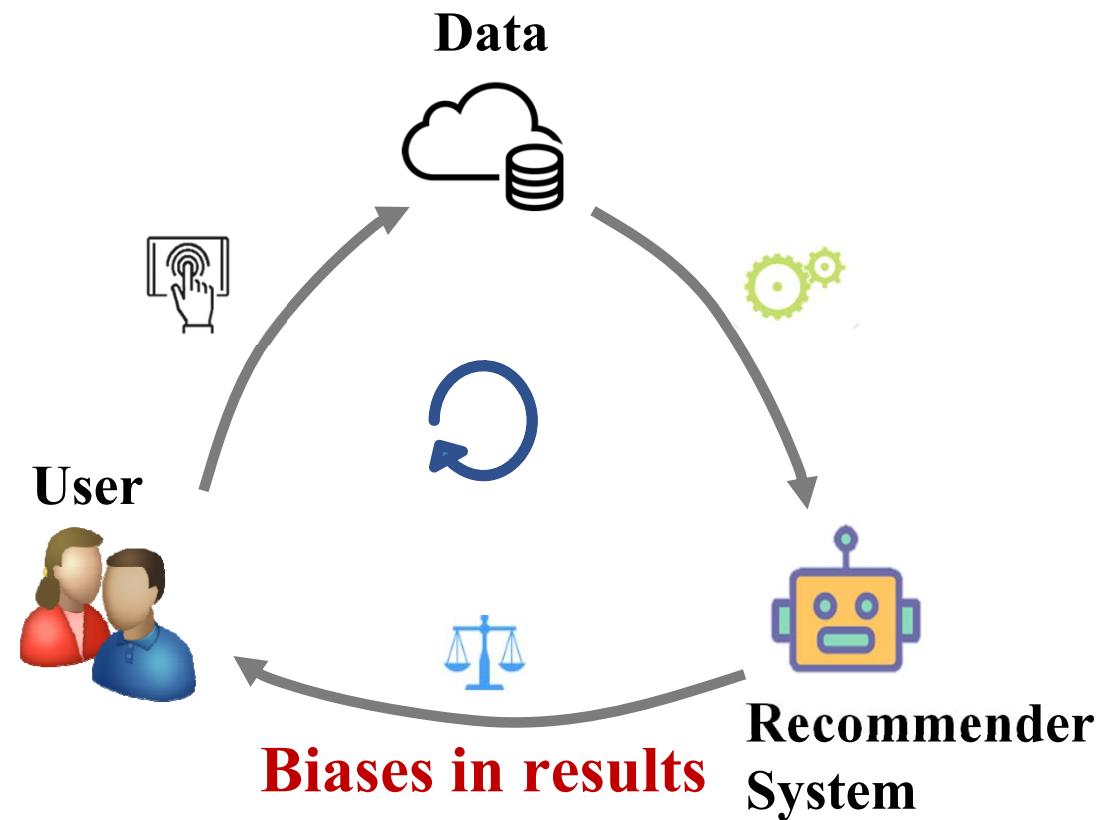
- Data is observational rather than experimental
- Affected by many factors:
  - The exposure mechanism
  - Public opinions
  - Display position
  - .....
- Deviation from reflecting user true preference



## • Where Bias Comes?

### • Biases in results (Serving):

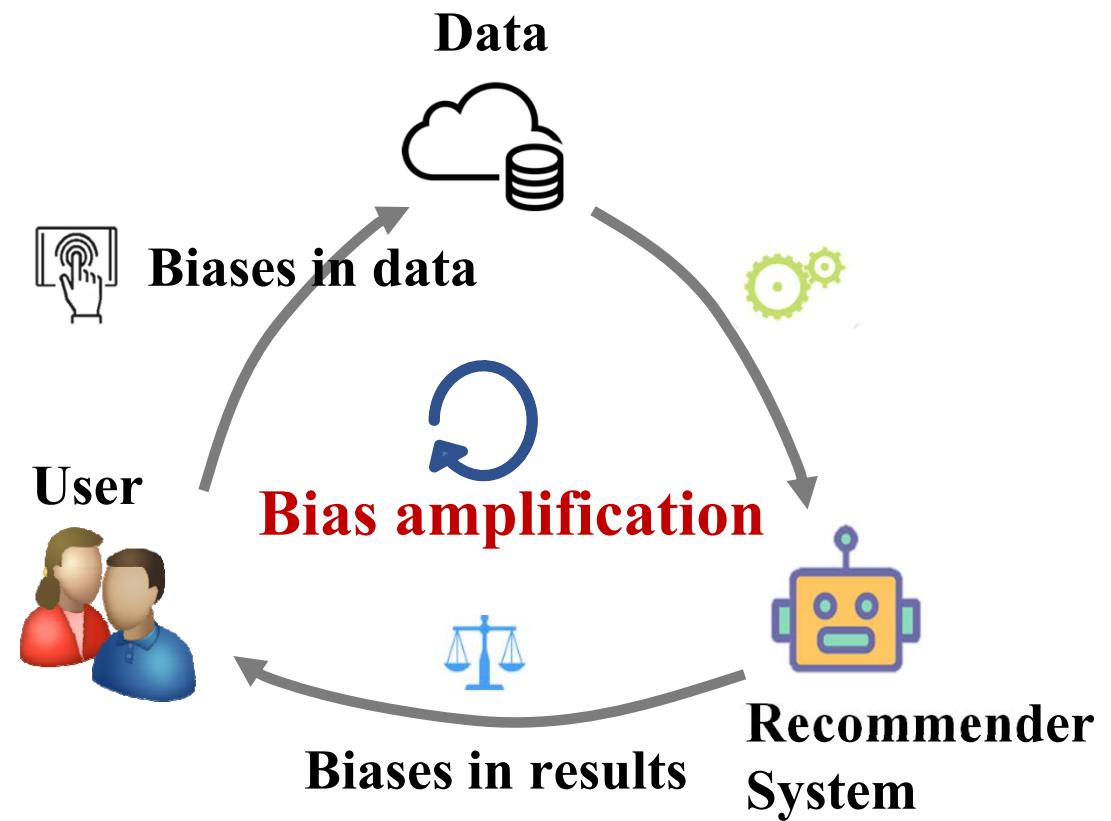
- Unbalanced data
- Recommendations in favor some specific groups
- Resulting in popularity bias and unfairness

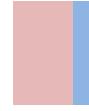


- **Matthew Effect: Bias + Loop**

- Biases amplification along the loop:

- Biases would be circled back into the collected data
- Resulting in “Matthew effect” issue





- **Bias is Evil**

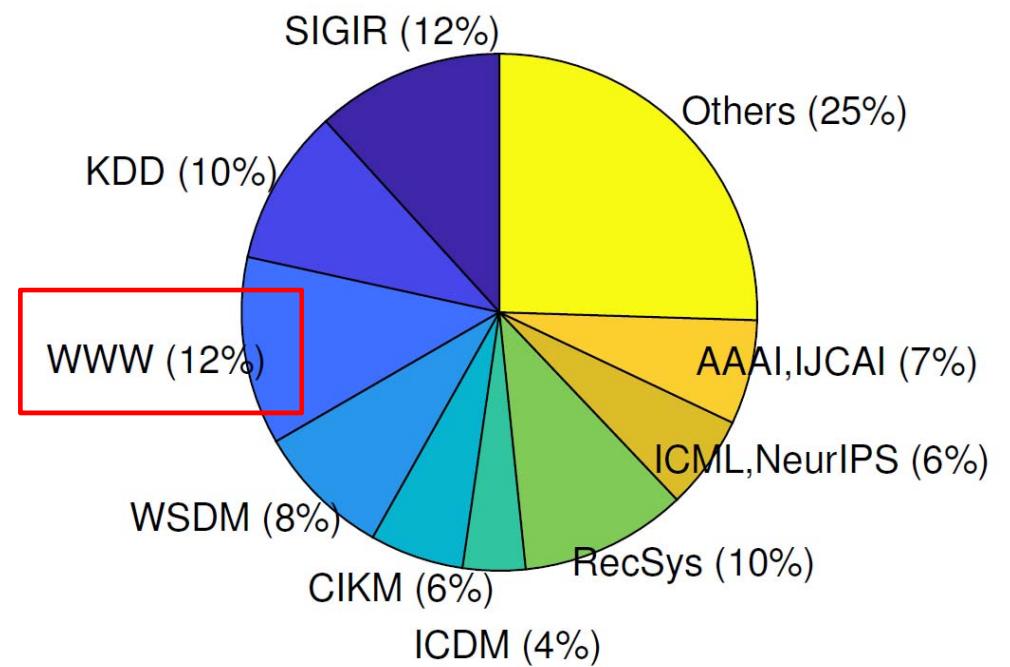
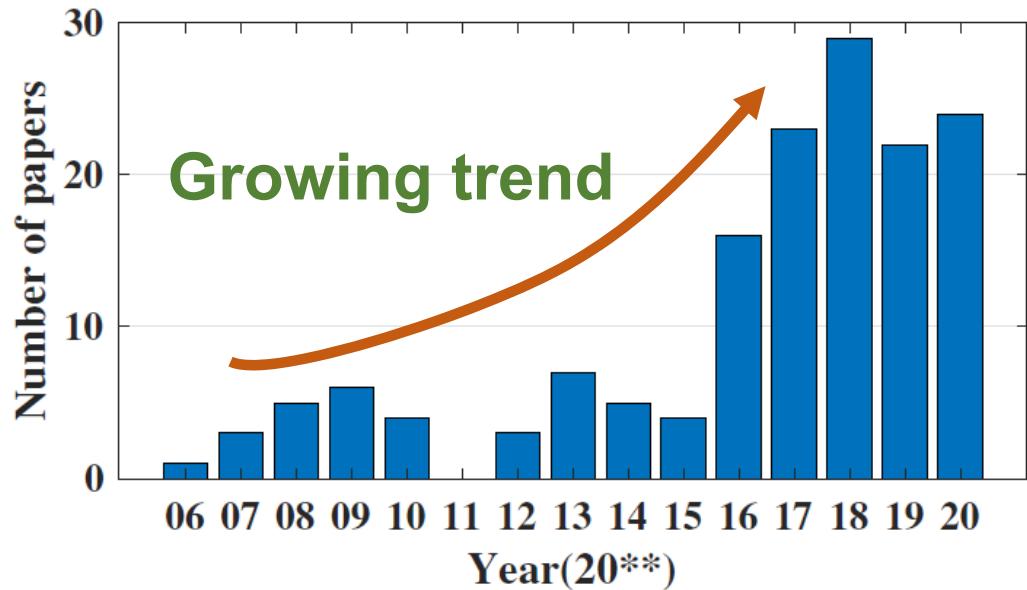
- **Economic**

- Bias affects recommendation accuracy
- Bias hurts user experience, causing the losses of users
- Unfairness incurs the losses of item providers

- **Society**

- Bias can reinforce discrimination of certain user's groups
- Bias decreases the diversity and intensify the homogenization of users

## • Increasing Research in Recsys Bias



Recommendation debiasing becomes a hot topic in top conference

## • Also Better Papers and Challenges



### Best Paper Award

[Controlling Fairness and Bias in Dynamic Learning-to-Rank](#)

Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims.



### WSDM 2021 Best Paper Award Recipient

**38: Unifying Online and Counterfactual Learning to Rank**

Harrie Oosterhuis (University of Amsterdam), Maarten de Rijke (University of Amsterdam & Ahold Delhaize).

Tianchi Academic Competitions

Join the Latest Big Data Competitions and Get Exclusive Awards for University Students



**KDD Cup 2020 Challenges for Modern E-Commerce Platform:  
Debiasing**

## • Tutorial Outline

### ❑ Biases in Data (Jiawei Chen, 50 min)

- ❑ Definition of data biases
- ❑ Categories: Selection bias, Conformity bias, Exposure bias and Position bias
- ❑ Recent solutions for data biases

### ❑ Biases in Results

- ❑ Popularity bias: definition, characteristic and solutions (Fuli Feng, 40 min)
- ❑ Unfairness: definition, characteristic and solutions (Xiang Wang, 50 min)

### ❑ Bias Amplification in Loop and its Solutions (Jiawei Chen, 10 min)

### ❑ Summary and Future Direction (Jiawei Chen, 20 min)

slides will be available at: <https://github.com/jiawei-chen/RecDebiasing>

A literature survey based on this tutorial is available at: <https://arxiv.org/pdf/2010.03240.pdf>

## • Tutorial Outline

### ❑ Biases in Data

- ❑ Definition of data biases
- ❑ Categories: Selection bias, Conformity bias, Exposure bias and Position bias
- ❑ Recent solutions for data biases

### ❑ Biases in Results

- ❑ Popularity bias: definition, characteristic and solutions
- ❑ Unfairness: definition, characteristic and solutions

### ❑ Bias Amplification in Loop and its Solutions

### ❑ Summary and Future Direction

slides will be available at: <https://github.com/jiawei-chen/RecDebiasing>

A literature survey based on this tutorial is available at: <https://arxiv.org/pdf/2010.03240.pdf>

## • Problem Formulation of RS

- Recommender system (RS)



- Goal of RS

$$f\left( \begin{array}{c} \text{User icons} \\ , \end{array} \begin{array}{c} \text{Item icons grid} \\ \rightarrow \end{array} \right) \approx \begin{array}{c} \text{Predicted Score} \\ \begin{array}{|c|c|c|c|} \hline 3 & 2 & 1 & 4 \\ \hline 2 & 2 & 3 & 5 \\ \hline 2 & 4 & 5 & 5 \\ \hline \end{array} \end{array}$$

**Label Matrix**  $R$

3	4	2	5
1	3	2	5
2	3	4	4

- True risk

$$L(f) = E_{P_D(u,i)P_D(R_{ui}|u,i)}[\delta(f(u,i), R_{ui})]$$

## • Problem Formulation of RS

- Training dataset sampled from distribution  $p_T$

$$D_T = \{(u, i, r_{ui}) : u \in U, i \in I, O_{ui} = 1\}$$

- Empirical risk



Training data

3		1	
2	3		5
2		5	

History feedback

3		1	
2	3		5
2		5	

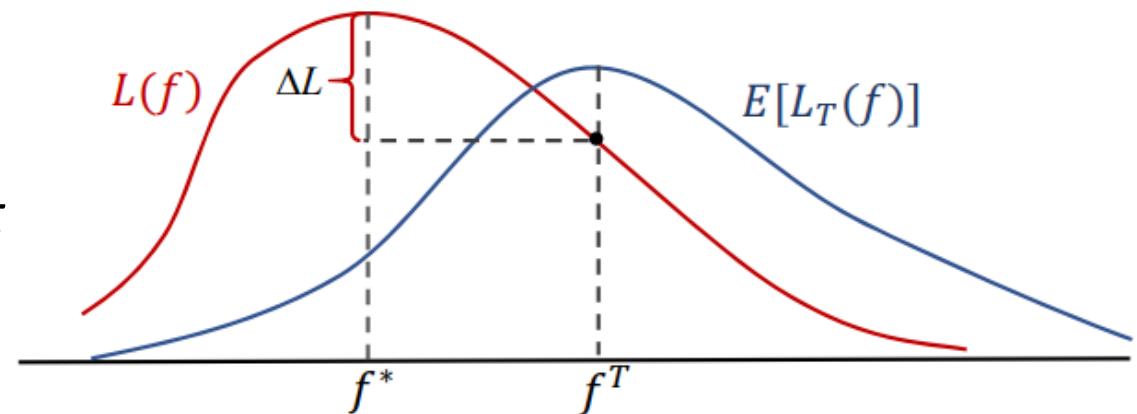
$$\hat{L}_T(f) = \frac{1}{|D_T|} \sum_{(u, i, r_{ui}) \in D_T} [\delta(f(u, i), r_{ui})]$$

PAC theory: If the training data set is sampled from the ideal data distribution, the learned model will be approximately optimal if we have sufficiently large training data.

## • Definition of Data Bias

- What is data bias?

*The distribution for which the training data is collected is different from the ideal data distribution*



Distributional difference  
between  $p_T$  and  $P_D$ .

$$p_T \neq p_D$$

Risk discrepancy  
between  $\hat{L}_T(f)$  and  $L(f)$ .

$$E_{P_T}[\hat{L}_T(f)] \neq L(f)$$

Bias in recommendation  
system.

$$f^* \neq f^T$$

## • Biases in Recommendation Data

Types	Stage in Loop	Data	Cause	Effect
Selection Bias	User→Data	Explicit feedback	Users' self -selection	Skewed observed rating distribution
Exposure Bias	User→Data	Implicit feedback	Users' self-selection; Background; Intervened by systems; Popularity	Unreliable non-positive data
Conformity Bias	User→Data	Both	Conformity	Skewed labels
Position Bias	User→Data	Both	Trust top of lists; Exposed to top of lists	Unreliable positive data

## • Selection Bias

- Definition: *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.*

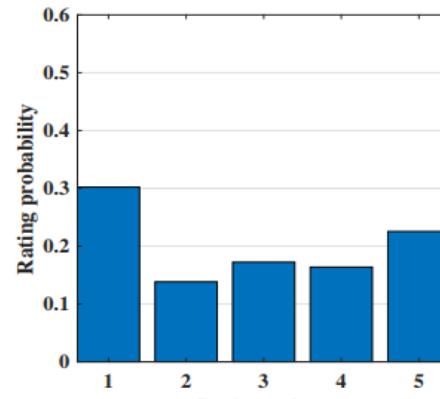
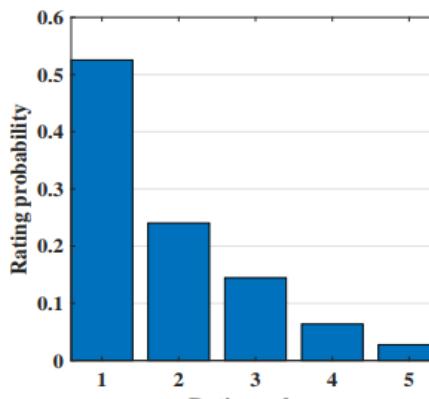
3	4	2	5
1	3	2	5
2	3	4	4

Selection bias

$\xrightarrow{\hspace{1cm}}$

$$p_T(u,i) \neq p_D(u,i)$$

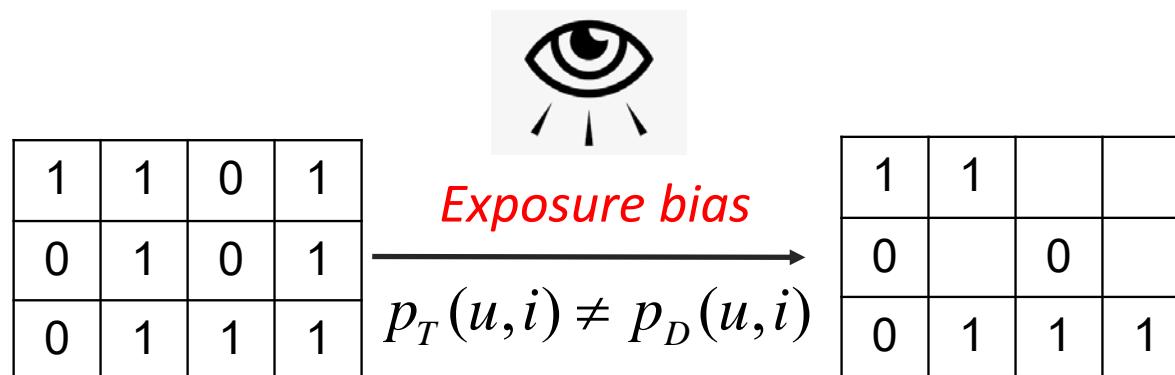
3	4		5
	3		5
	3	4	4



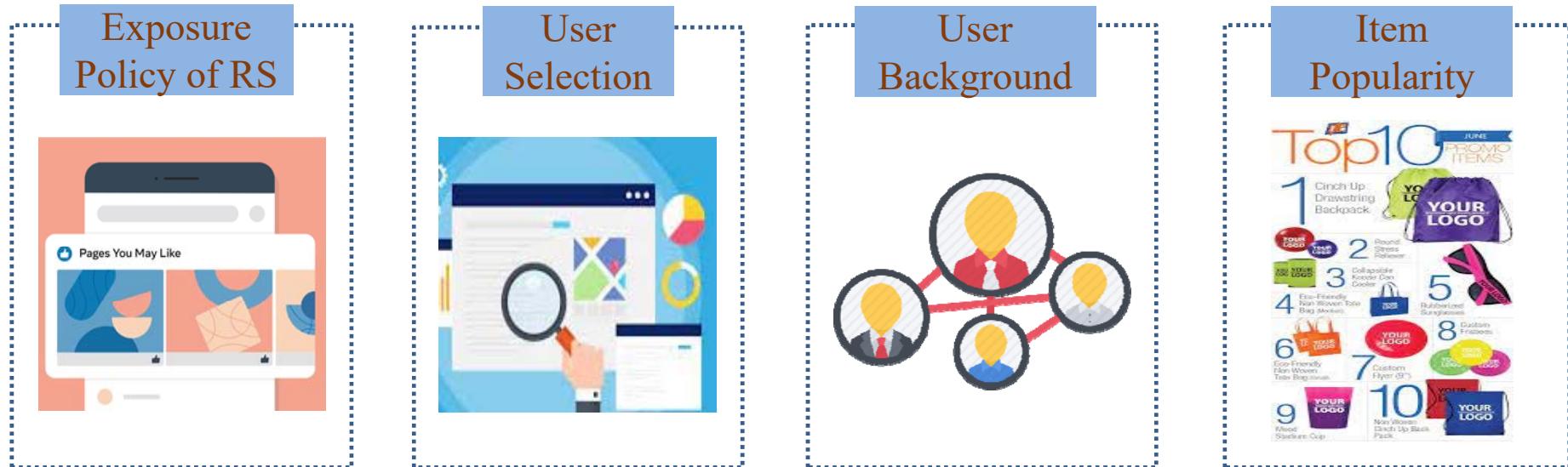
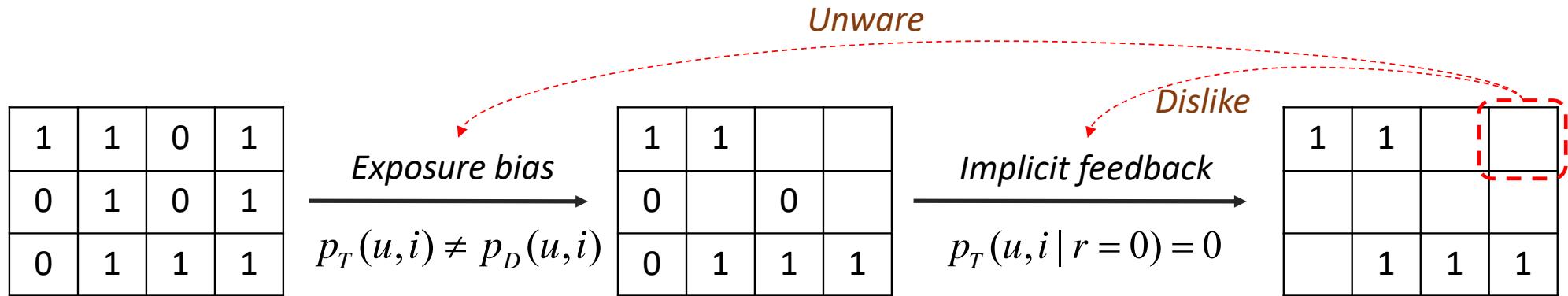
- [1] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In ICML.
- [2] B. M. Marlin, R. S. Zemel, S. Roweis, and M. Slaney, “Collaborative filtering and the missing at random assumption,” in UAI, 2007

## • Exposure Bias

- Definition: *Exposure bias* happens in implicit feedback data as users are only exposed to a part of specific items.
- Explanation: A user generates behaviors on exposed items, making the observed user-item distribution  $p_T(u, i)$  deviate from the ideal one  $p_D(u, i)$ .



## • Exposure Bias



## • Conformity Bias

- Definition: *Conformity bias* happens as users tend to behave similarly to the others in a group, even if doing so goes against their own judgment.

3	4		5
	3		4
	3	4	3

$$p_T(r|u,i) \neq p_D(r | u,i)$$

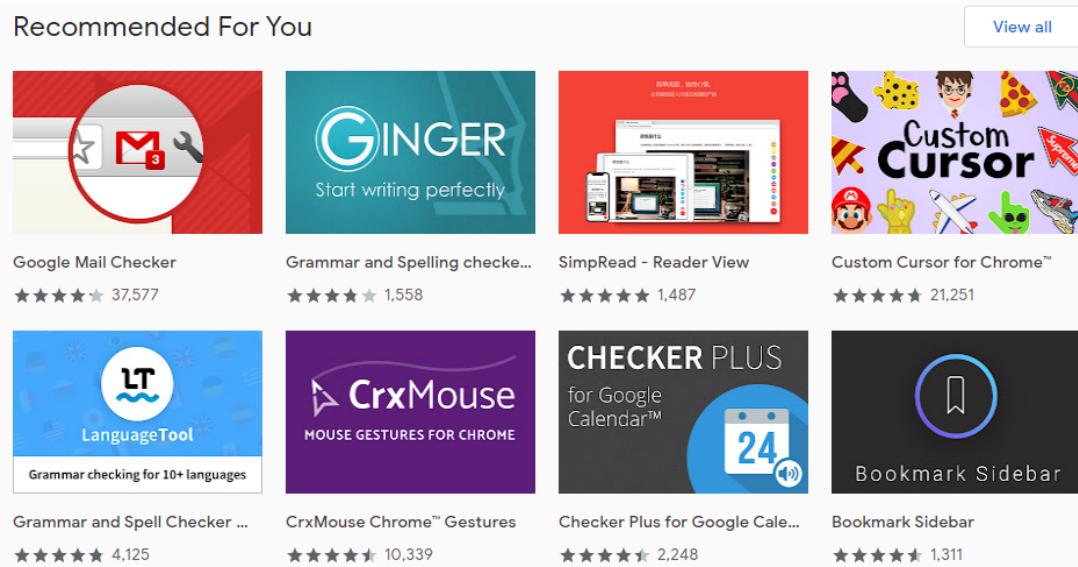
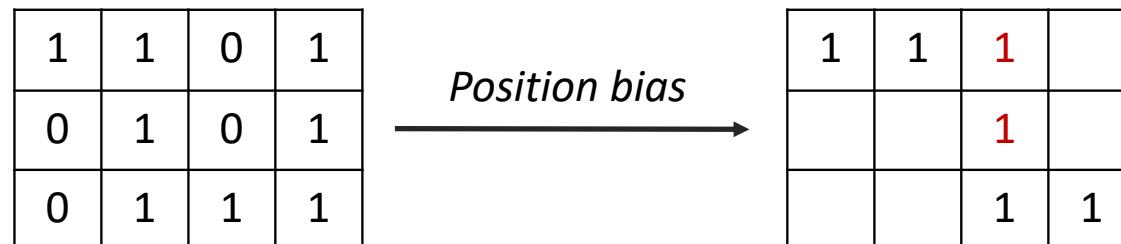
*Conformity bias*

3	4		5
	3		5
	3	3	4



## • Position Bias

- Definition: *Position bias happens as users tend to interact with items in higher position of the recommendation list.*



$$p_T(u, i) \neq p_D(u, i)$$

User exposure will be affected by the position

$$p_T(r | u, i) \neq p_D(r | u, i)$$

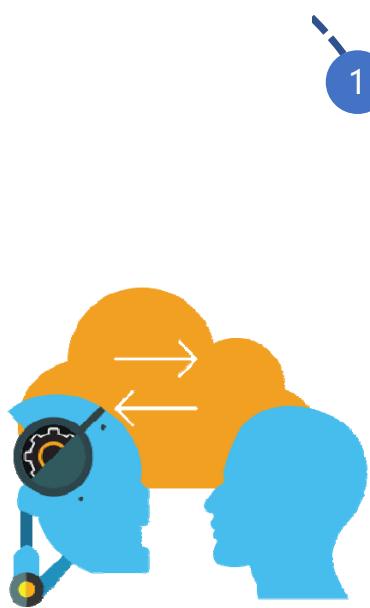
User judgments also will be affected by the position



## • Debiasing Strategies Overview

- Re-weighting
  - giving weights for each instance to re-scale their contributions on model training
- Re-labeling
  - giving a new pseudo-label for the missing or biased data
- Generative modeling
  - assuming the generation process of data and reduces the biases accordingly

# • Debiasing Strategies Overview



1 For Selection Bias

- Data imputation (Relabeling)
- Propensity score (Reweighting)
- Doubly robust model (Reweighting+Relabeling)
- Generative model (Generative modeling)
- Knowledge distillation

2 For Conformity Bias — Disentangling (Relabeling)

3 For Exposure Bias

- Heuristic (Reweighting)
- Sampling (Reweighting)
- Exposure-based generative model (Generative modeling)
- Propensity variant (Reweighting)

4 For Position Bias

- Click model (Generative modeling)
- Propensity score (Reweighting)
- Trust-aware model (Reweighting+Relabeling)

## • Data Imputation for Selection Bias (Relabeling)

### • True Preference

3	4	2	5
1	3	2	5
2	3	4	4

*Selection bias*  
→  
 $p_T(u,i) \neq p_D(u,i)$

### Training data

3	4		5
	3		
2	3	4	4

Data imputation  
→

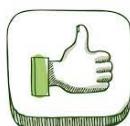
### Imputation data

3	4	2	5
2	3	2	4
2	3	4	4

- Data Imputation: assigns pseudo-labels for missing data.

$$\arg \min_{\theta} \sum_{u,i} \hat{\delta}_{u,i} \left( \begin{matrix} r_{ui}^{o \& i} \\ \underline{r_{ui}} \end{matrix}, f(u,i | \theta) \right) + \text{Reg}(\theta)$$

The imputed labels:  
heuristically or model dictated.



Simple and straightforward.



Sensitive to the imputation strategy.  
Imputing proper pseudo-labels is not easy.

## • Data Imputation for Selection Bias (Relabeling)

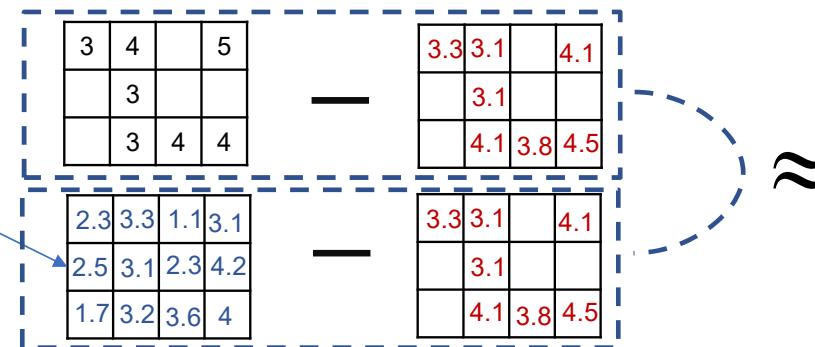
- Heuristic:  
with a specific value  $r$

H. Steck, "Training and testing of recommender systems on data missing not at random," in KDD, 2010, pp. 713–722.

3	4	r	5
r	3	r	r
r	3	4	4

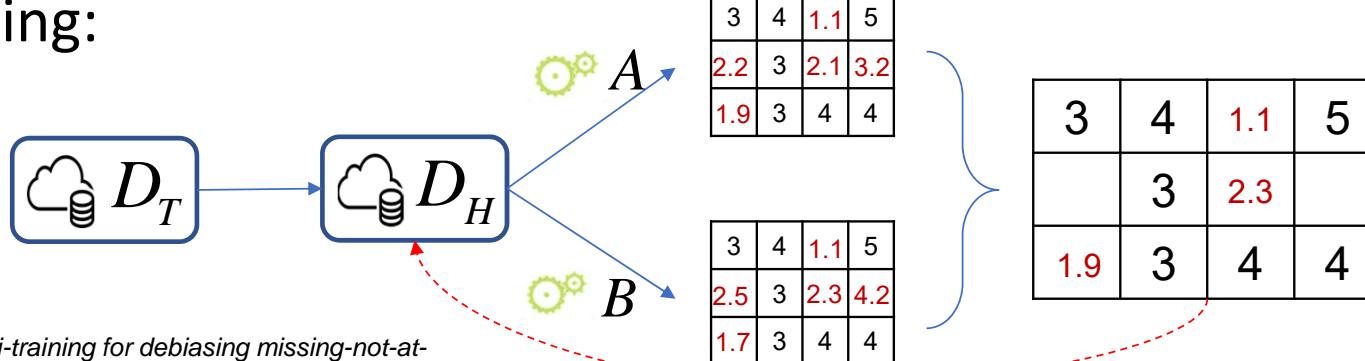
- Model-based:

X. Wang, R. Zhang, Y. Sun, and J. Qi, "Doubly robust joint learning for recommendation on data missing not at random," in ICML, 2019, pp. 6638–6647

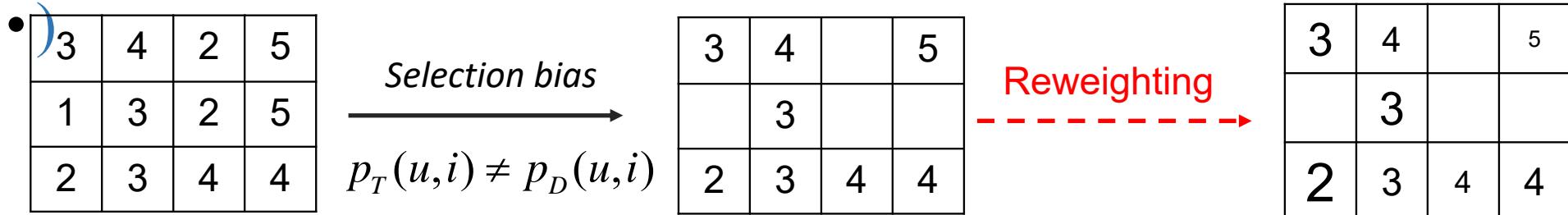


- Tri-training:

Y. Saito, "Asymmetric tri-training for debiasing missing-not-at-random explicit feedback," in SIGIR, 2020, pp. 309–318.



## • Propensity score for Selection Bias (Reweighting)



- Inverse Propensity Score (IPS): reweighs the collected data for an unbiased learning.

$$\operatorname{argmin}_{\theta} \sum_{O_{ui}=1} \frac{\hat{\delta}_{u,i} \left( r_{ui}^o, f(u,i | \theta) \right)}{q_{ui}} + Reg(\theta)$$

Propensity Score

$$q_{ui} = E_{p_T} [O_{ui}]$$



Simple and straightforward.

Relatively robust to the propensity score.



High Variance.

Developing proper propensity strategy is not easy.

## • Doubly Robust for Selection Bias (Relabeling+Reweighting)

3	4	2	5
1	3	2	5
2	3	4	4

*Selection bias*  
 $p_T(u,i) \neq p_D(u,i)$

3	4		5
	3		
2	3	4	4

*Reweighting*  


3	4	2	5
2	3	2	4
2	3	4	4

- Doubly Robust: combines IPS and data imputation for robustness.

$$\hat{L}_{DR}(f) = \sum_{u \in U, i \in I} \left( \delta(f(u, i | \theta), r_{ui}^i) + \frac{O_{ui} d_{ui}}{q_{ui}} \right)$$

Imputation
IPS



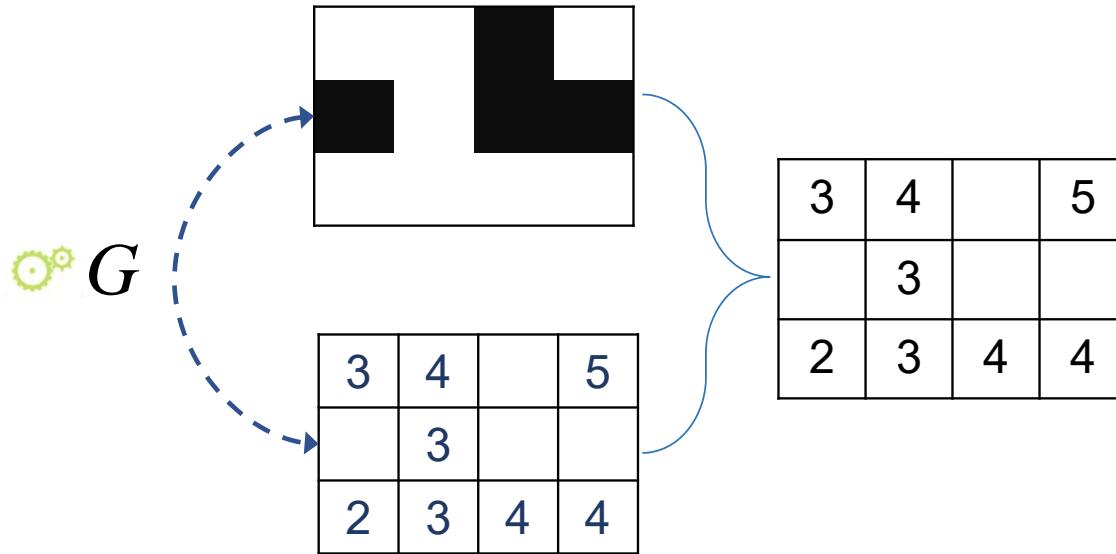
Low Variance.

Relatively robust to the propensity score and imputation value.



Developing proper imputation&propensity strategy is not easy.

## • Generative Model for Selection Bias (Generative modeling)



- Generative Model: jointly modeling rating values and user selection.



Explainable.



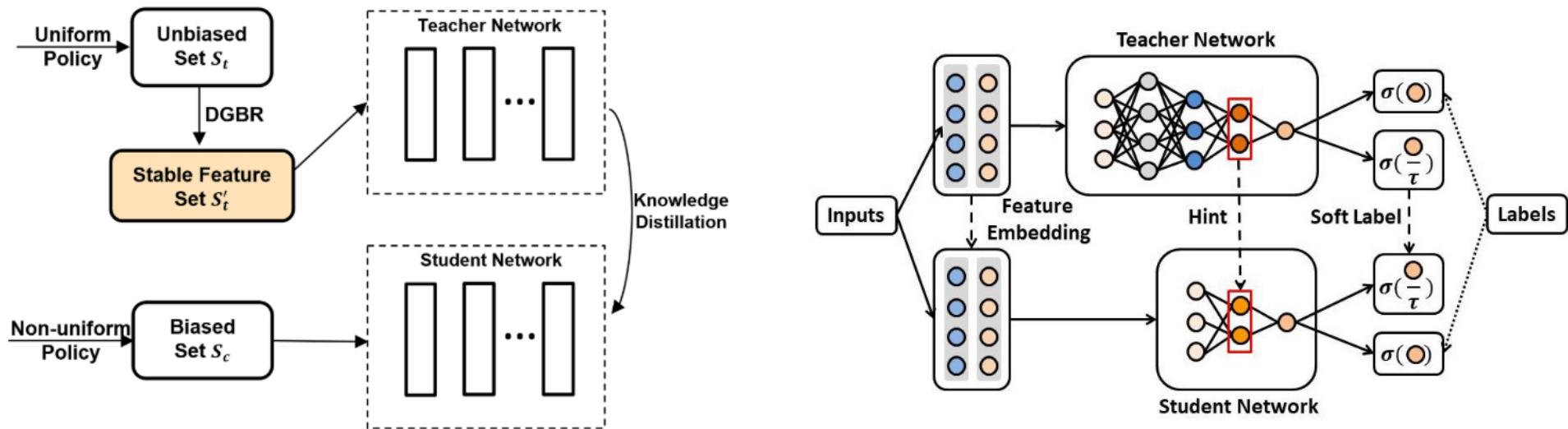
Complex and sophisticated models.  
Hard to train.

B. M. Marlin and R. S. Zemel, "Collaborative prediction and ranking with non-random missing data," in RecSys, 2009, pp. 5–12.

J. M. Hernández-Lobato, N. Houlsby, and Z. Ghahramani, "Probabilistic matrix factorization with non-random missing data," in ICML, 2014, pp. 1512–1520.

J. Chen, C. Wang, M. Ester, Q. Shi, Y. Feng, and C. Chen, "Social recommendation with missing not at random data," in ICDM. IEEE, 2018, pp. 29–38.

## • Knowledge Distillation for Selection Bias



- Knowledge distillation: distill the knowledge from unbiased data to the normal training on biased data.



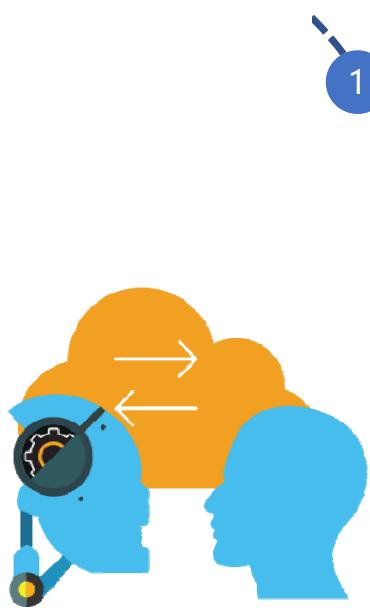
Effective.



Rely on uniform data.  
High variance.

D. Liu, P. Cheng, Z. Dong, X. He, W. Pan, and Z. Ming,  
“A general knowledge distillation framework for counterfactual  
recommendation via uniform data,” in SIGIR, 2020, pp. 831–840.

## • Debiasing Strategies Overview



1

For Selection Bias

2

For Conformity Bias

3

For Exposure Bias

4

For Position Bias

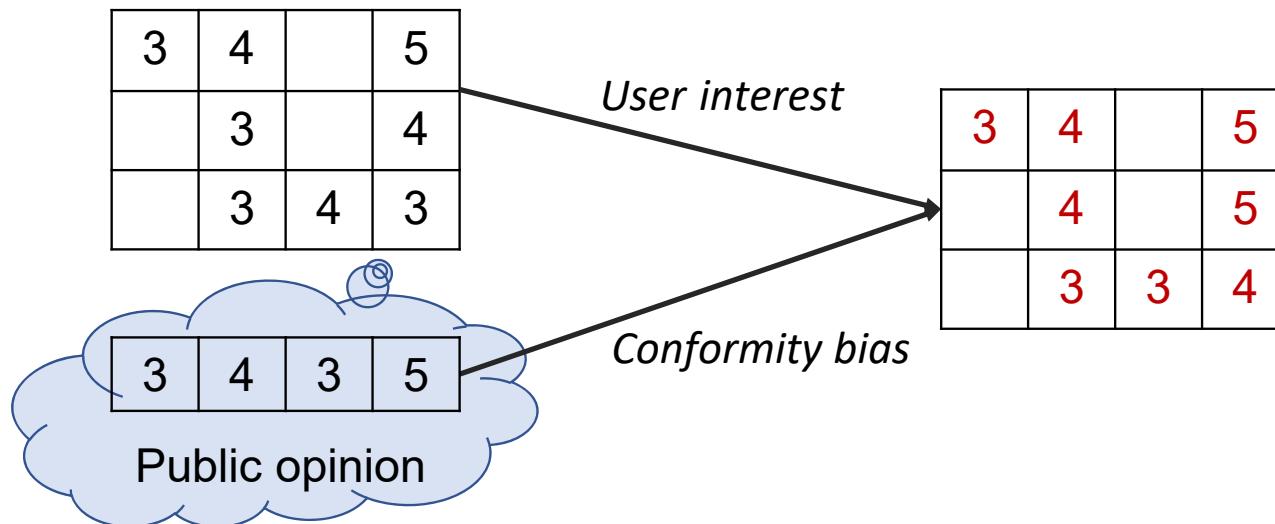
{ Data imputation (Relabeling)  
Propensity score (Reweighting)  
Doubly robust model (Reweighting+Relabeling)  
Generative model (Generative modeling)  
Knowledge distillation

— Disentangling (Relabeling)

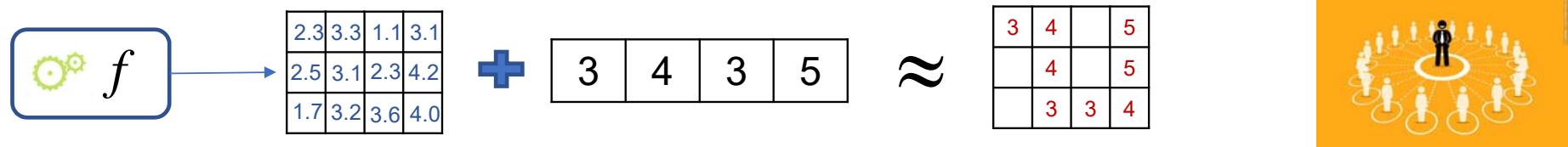
{ Heuristic (Reweighting)  
Sampling (Reweighting)  
Exposure-based generative model (Generative modeling)  
Propensity variant (Reweighting)

{ Click model (Generative modeling)  
Propensity score (Reweighting)  
Trust-aware model (Reweighting+Relabeling)

## • Disentangling for Conformity Bias (Relabeling)



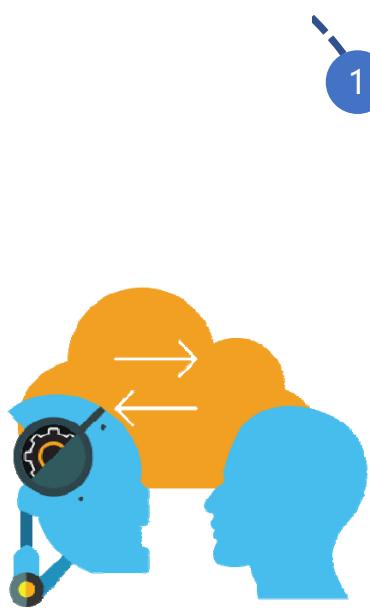
- Disentangling: disentangle the effect of user interest and conformity.



Y. Liu, X. Cao, and Y. Yu, "Are you influenced by others when rating?: Improve rating prediction by conformity modeling," in RecSys. ACM, 2016, pp. 269–272.

A.~J. Chaney, D.~M. Blei, and T.~Eliassi-Rad, ``A probabilistic model for using social networks in personalized item recommendation," in RecSys, ACM, 2015

## • Debiasing Strategies Overview



1

For Selection Bias

2

For Conformity Bias

3

For Exposure Bias

4

For Position Bias

{ Data imputation (Relabeling)  
Propensity score (Reweighting)  
Doubly robust model (Reweighting+Relabeling)  
Generative model (Generative modeling)  
Knowledge distillation

— Disentangling (Relabeling)

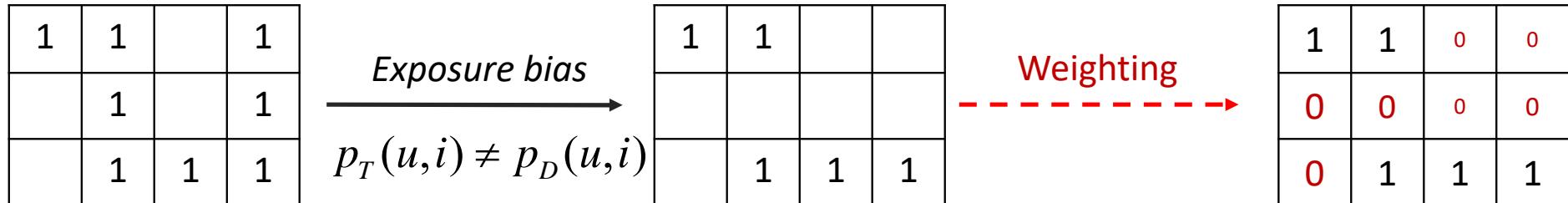
{ Heuristic (Reweighting)  
Sampling (Reweighting)  
Exposure-based generative model (Generative modeling)  
Propensity variant (Reweighting)

{ Click model (Generative modeling)

Propensity score (Reweighting)

Trust-aware model (Reweighting+Relabeling)

## • Heuristic weighting for Exposure Bias (Reweighting)



$$\operatorname{argmin}_{\theta} \sum_{ui} W_{ui} \delta(r_{ui}, f(u, i | \theta)) + Reg(\theta)$$

- Heuristic weighting: setting negative weighting in a heuristic way.

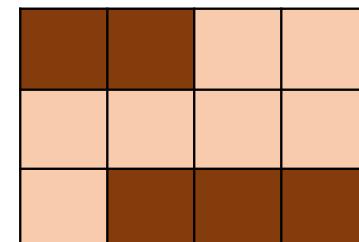


Simple.

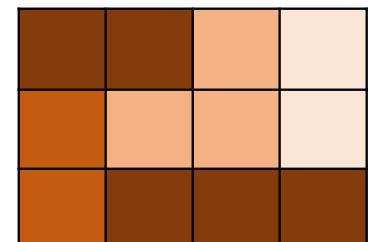


Impersonalized.  
Relying on human knowledge.

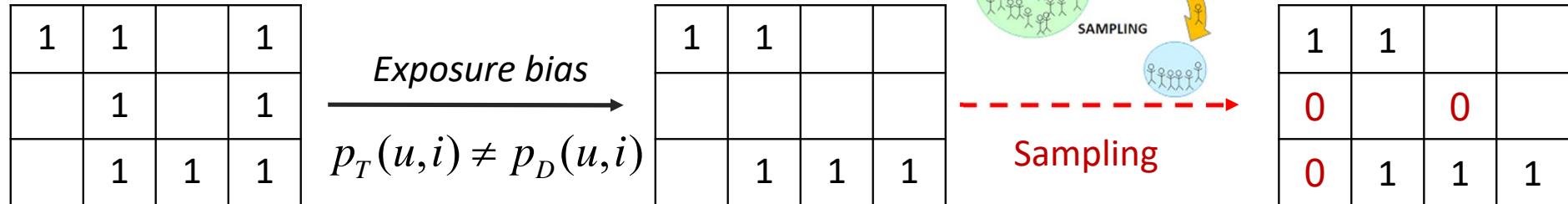
Uniform weights



Item popularity  
-based weights



## • Sampling for Exposure Bias (Reweighting)



$$E_{(u,i) \sim p} [\delta(r_{ui}, f(u, i | \theta))] = \sum_{u,i} p_{ui} \delta(r_{ui}, f(u, i | \theta))$$

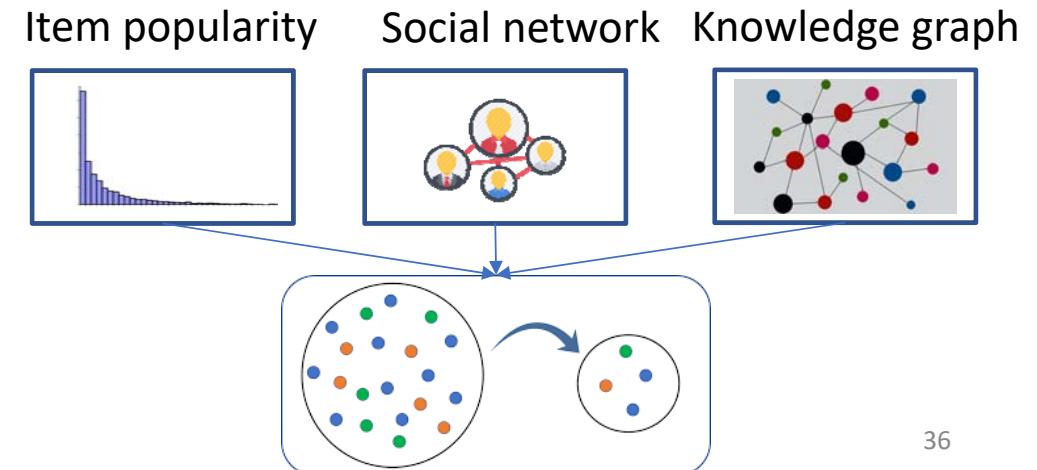
- Sampling: performing sampling to scale the data contribution.



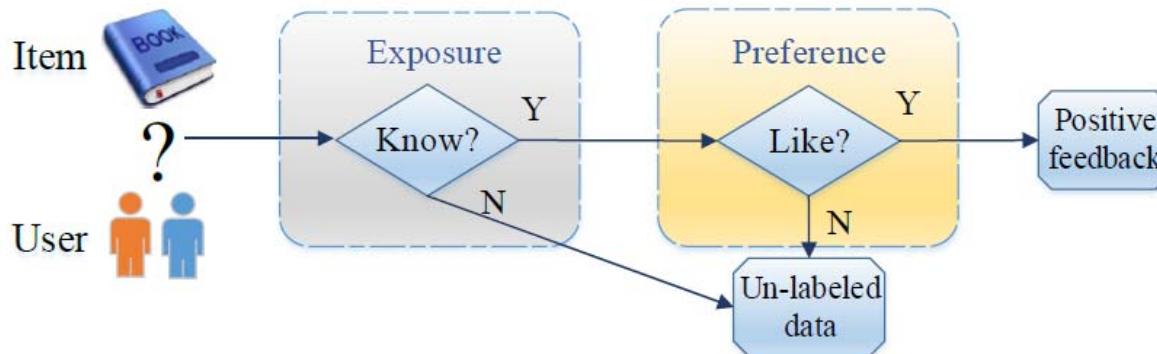
Efficient.



Relying on human knowledge or side information.



## • Exposure Model for Exposure Bias (Generative modeling)



$$a_{ui} \sim \text{Bernoulli}(\eta_{ui})$$

$$(r_{ui} | a_{ui} = 1) \sim \text{Bernoulli}(f(u, i | \theta))$$

$$(r_{ui} | a_{ui} = 0) \sim \delta_0$$

$$\operatorname{argmin}_{\theta, \gamma} \sum_{ui} \gamma_{ui} \delta(x_{ui}, f(u, i | \theta)) + \sum_{ui} g(\gamma_{ui}) \quad \gamma_{ui} \approx p(a_{ui} | r_{ui})$$

- Generative model: jointly modeling both user exposure and preference.



Personalized.

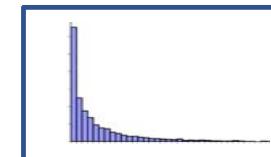


Learnable.

Hard to train.

Relying on strong assumptions.

Item popularity



Social network

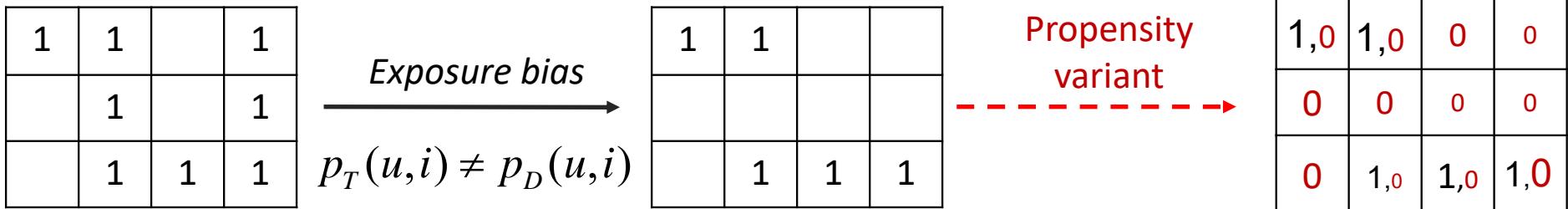


User community



D. Liang, L. Charlin, J. McInerney, and D. M. Blei, "Modeling user exposure in recommendation," in *WWW*. 2016  
 J. Chen, C. Wang, S. Zhou, Q. Shi, Y. Feng, and C. Chen, "Samwalker: Social recommendation with informative sampling strategy," in *The World Wide Web Conference*. ACM, 2019, pp. 228–239.

## • Propensity Variant for Exposure Bias (Reweighting)



$$L_{IPSV} = \sum_{O_{ui}=1} \frac{\hat{\delta}_{u,i}(r_{ui}, f(u,i | \theta))}{q_{ui}} + \sum_{u \in \mathcal{U}, i \in \mathcal{I}} \left(1 - \frac{O_{ui}}{q_{ui}}\right) \delta(f(u,i), 0)$$

- Propensity variant: scaling both positive and negative instances with propensity-based weights.



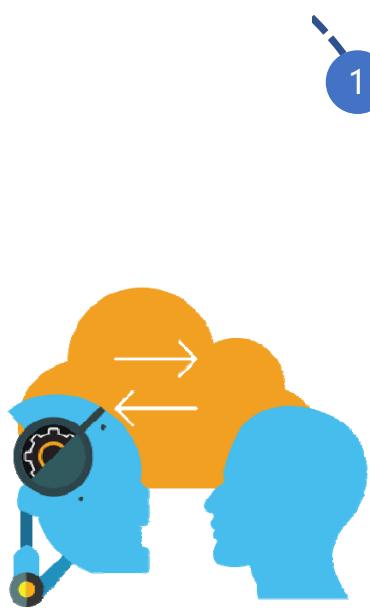
Theoretical soundness.



High variance.

Developing proper propensity strategy is not easy.

## • Debiasing Strategies Overview



1

For Selection Bias

2

For Conformity Bias

3

For Exposure Bias

4

For Position Bias

{ Data imputation (Relabeling)  
Propensity score (Reweighting)  
Doubly robust model (Reweighting+Relabeling)  
Generative model (Generative modeling)  
Knowledge distillation

— Disentangling (Relabeling)

{ Heuristic (Reweighting)  
Sampling (Reweighting)  
Exposure-based generative model (Generative modeling)  
Propensity variant (Reweighting)

{ Click model (Generative modeling)

Propensity score (Reweighting)

Trust-aware model (Reweighting+Relabeling)

## • Click model for Position Bias (Generative modeling)



$$\begin{aligned}
 P(C = 1 | u, i, p) &= \underbrace{P(C = 1 | u, i, E = 1)}_{r_{ui}} \cdot \underbrace{P(E = 1 | p)}_{h_p} \\
 P(E_{p+1} = 1 | E_p = 0) &= 0 \\
 P(E_{p+1} = 1 | E_p = 1, C_p) &= 1 - C_p \\
 P(C_p = 1 | E_p = 1) &= r_{u_p, i}
 \end{aligned}$$

- Click model: making hypotheses about user browsing behaviors and learn true preference (or relevant) by optimizing likelihood of the observed clicks .



Explainable.



Requiring a large quantity of clicks.  
Requiring strong assumptions.

O. Chapelle and Y. Zhang, “A dynamic bayesian network click model for web search ranking,” in *WWW*, 2009, pp. 1–10.

F. Guo, C. Liu, A. Kannan, T. Minka, M. Taylor, Y.-M. Wang, and C. Faloutsos, “Click chain model in web search,” in *WWW*, 2009, pp. 11–20.

Z. A. Zhu, W. Chen, T. Minka, C. Zhu, and Z. Chen, “A novel click model and its applications to online advertising,” in *WSDM*, 2010, pp. 321–330.

## • Propensity Score for Position Bias (Reweighting)



$$\begin{aligned} L_{IPW}(S, q) &= \sum_{x \in \pi_q} \Delta_{IPW}(x, y \mid \pi_q) \\ &= \sum_{x \in \pi_q, o_q^x = 1, y = 1} \frac{\Delta(x, y \mid \pi_q)}{P(o_q^x = 1 \mid \pi_q)} \end{aligned}$$

- Propensity: weighting each instance with a position-aware value.



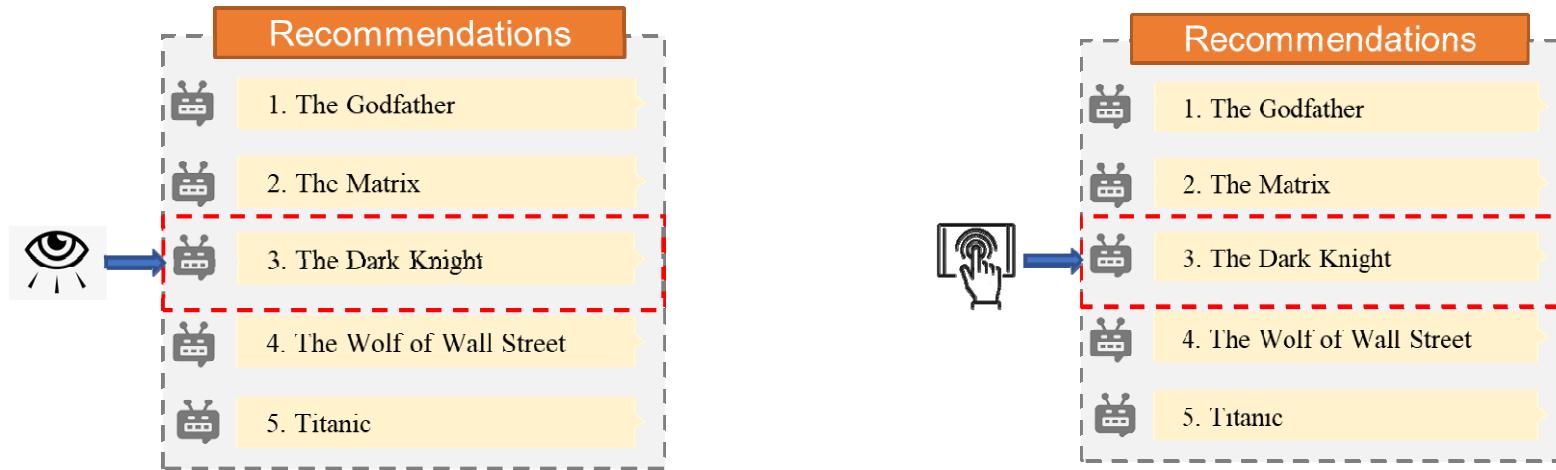
Simple and straightforward.



High variance.  
Specifying proper propensity is not easy.

T. Joachims, A. Swaminathan, and T. Schnabel, "Unbiased learning-to-rank with biased feedback," in WSDM, 2017, pp. 781–789

## • Trust-aware Model for Position Bias



- Trust-aware model: explicitly modeling the effect of position on user judgement.

$$\begin{aligned} & \Pr(\tilde{R} = 1 \mid E = 1, q, d, k) \\ &= \Pr(\tilde{R} = 1 \mid R = 1, E = 1, k) \Pr(R = 1 \mid q, d) \\ &+ \Pr(\tilde{R} = 1 \mid R = 0, E = 1, k) \Pr(R = 0 \mid q, d) \end{aligned}$$

## • Trust-aware Model for Position Bias

### ➤ Trust-aware weighting (Reweighting)

$$\sum_{(q,d,k,c=1) \in \mathcal{L}} \frac{1}{\theta_k} \frac{\epsilon_k^+}{\epsilon_k^+ + \epsilon_k^-} f(q, d, \Omega_q) \quad \begin{aligned} \epsilon_k^+ &= \Pr(\tilde{R} = 1 \mid R = 1, E = 1, k) \\ \epsilon_k^- &= \Pr(\tilde{R} = 1 \mid R = 0, E = 1, k) \end{aligned}$$

Agarwal, Aman, Xuanhui Wang, Cheng Li, Michael Bendersky, and Marc Najork. "Addressing trust bias for unbiased learning-to-rank." In *The World Wide Web Conference*, pp. 4-14. 2019.

### ➤ Affine model (Reweighting+Relabeling)

$$\hat{\Delta}_{\text{affine}}(f) = \frac{1}{N} \sum_{i=1}^N \sum_{(d,k) \in y_i} \frac{c_i(d) - \theta_k \epsilon_k^-}{\theta_k (\epsilon_k^+ - \epsilon_k^-)} \cdot \lambda(d \mid q_i, f)$$

Vardasbi, Ali, Harrie Oosterhuis, and Maarten de Rijke. "When Inverse Propensity Scoring does not Work: Affine Corrections for Unbiased Learning to Rank." *CIKM*, 2019, pp. 1475-1484. 2020.

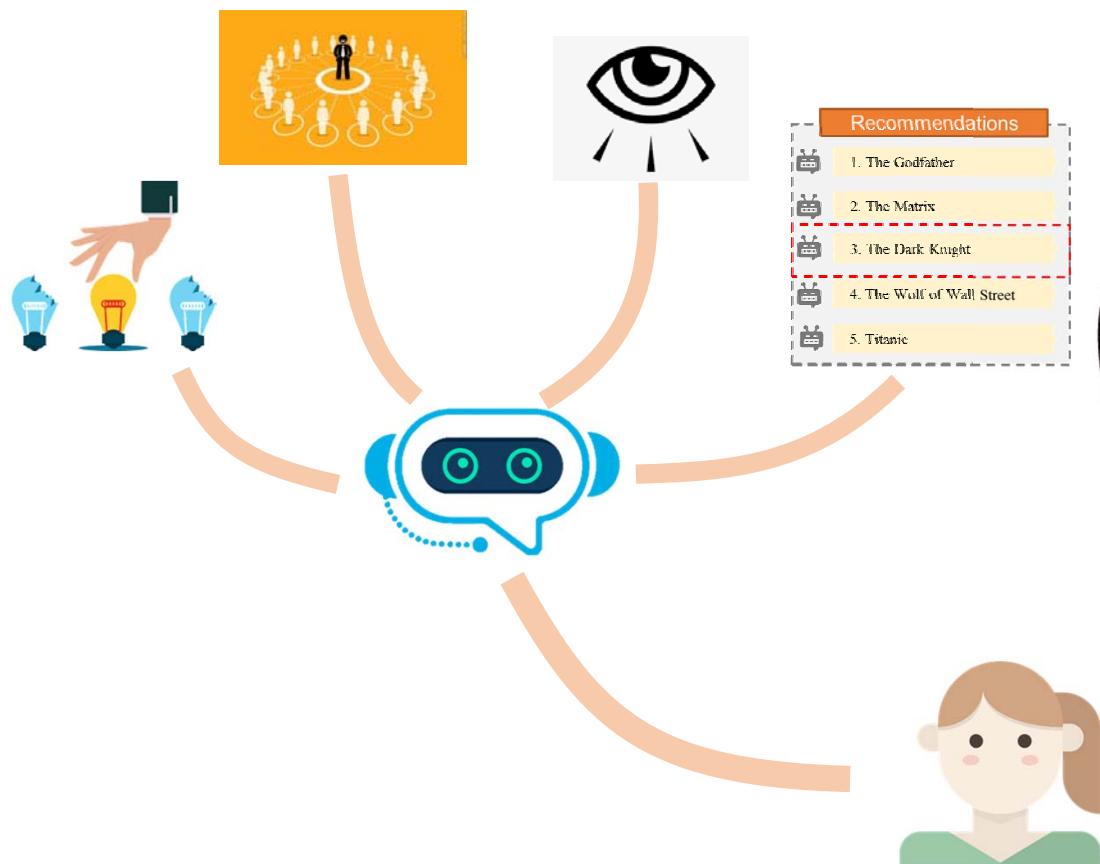
### ➤ Neural network model

$$\mathbb{P}(c_j \mid X, P) = f(g(r(x_j)) + h(e(x_j, p_j, X_n P)))$$

Zhuang, Honglei, Zhen Qin, Xuanhui Wang, Mike Bendersky, Xinyu Qian, Po Hu, and Chary Chen. "Cross-Positional Attention for Debiasing Clicks." *WWW*, 2021.

## • Open Problem and Future Direction

- A universal solution.

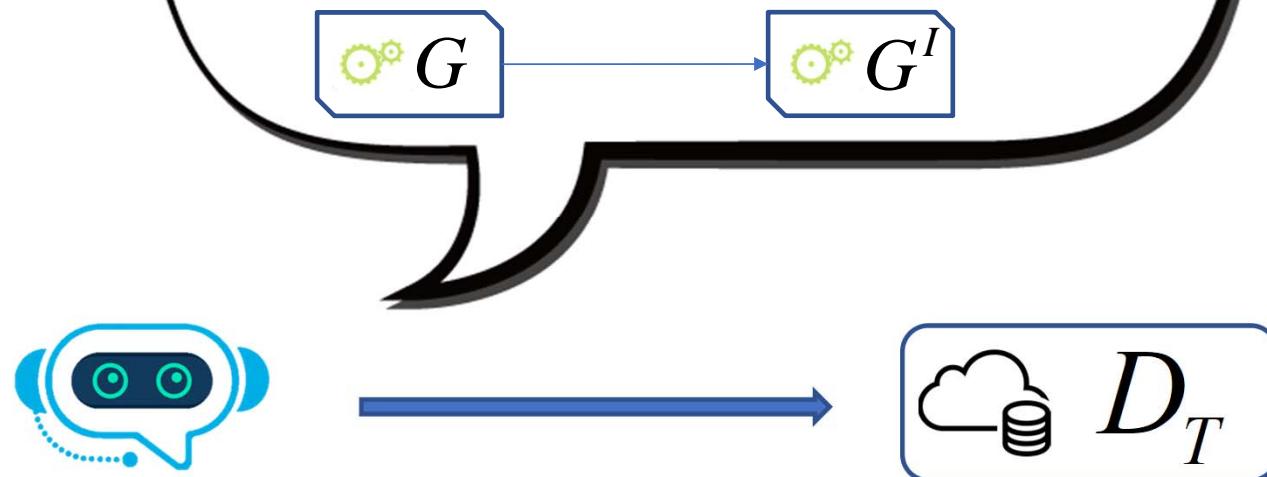


How to develop a universal solution that accounts for multiple biases and their combinations?

## • Open Problem and Future Direction

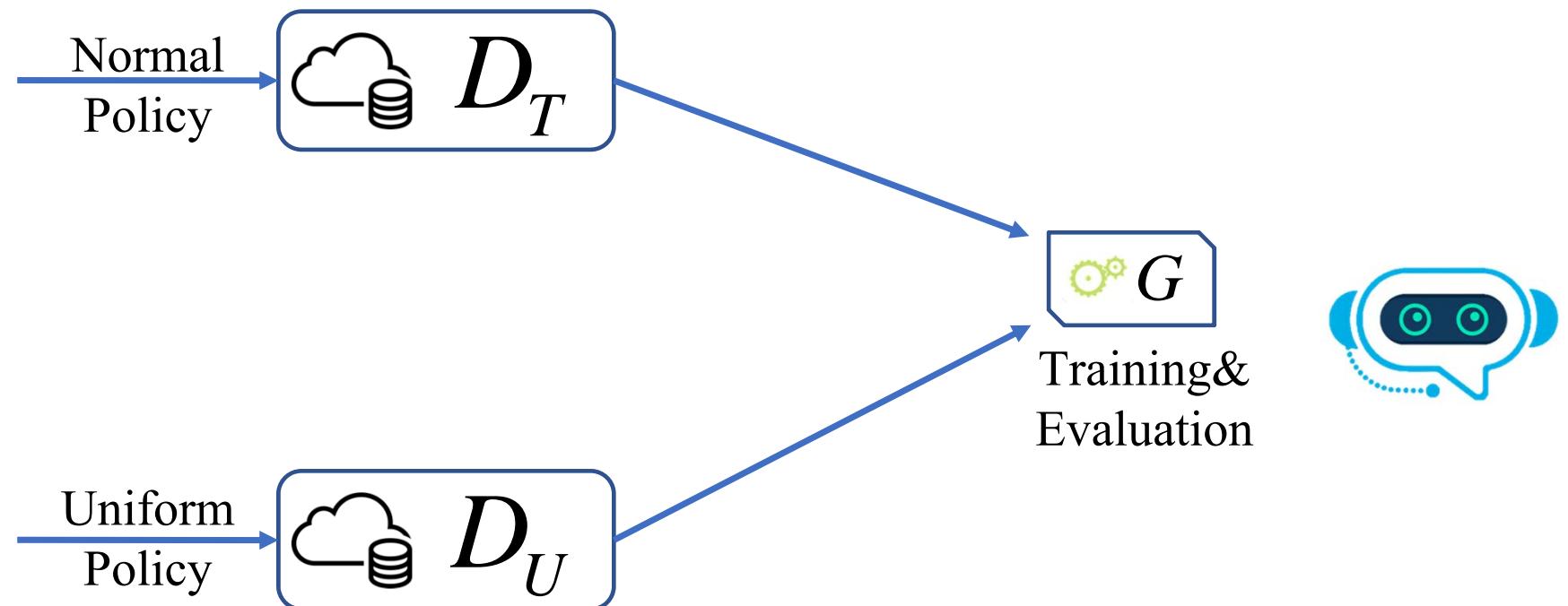
- An adaptive solution.

I guess there exist position bias and selection bias on the data. I need introduce a proper propensity score.



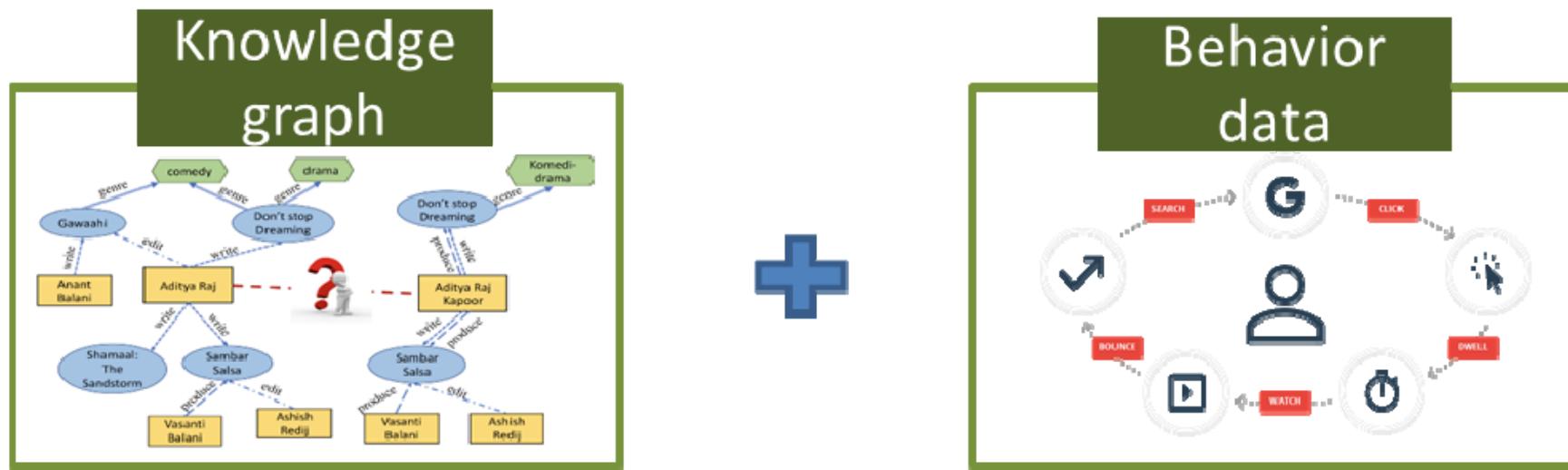
## • Open Problem and Future Direction

- Leveraging uniform data.



## • Open Problem and Future Direction

- Leveraging knowledge graph.



- Human prior knowledge plus data knowledge.



## • Tutorial Outline

### ❑ Biases in Data

- ❑ Definition of data biases
- ❑ Categories: Selection bias, Conformity bias, Exposure bias and Position bias
- ❑ Recent solutions for data biases

### ❑ Biases in Results

- ❑ Popularity bias: definition, characteristic and solutions
- ❑ Unfairness: definition, characteristic and solutions

### ❑ Bias Amplification in Loop and its Solutions

### ❑ Summary and Future Direction

slides will be available at: <https://github.com/jiawei-chen/RecDebiasing>

A literature survey based on this tutorial is available at: <https://arxiv.org/pdf/2010.03240.pdf>



# Popularity Bias

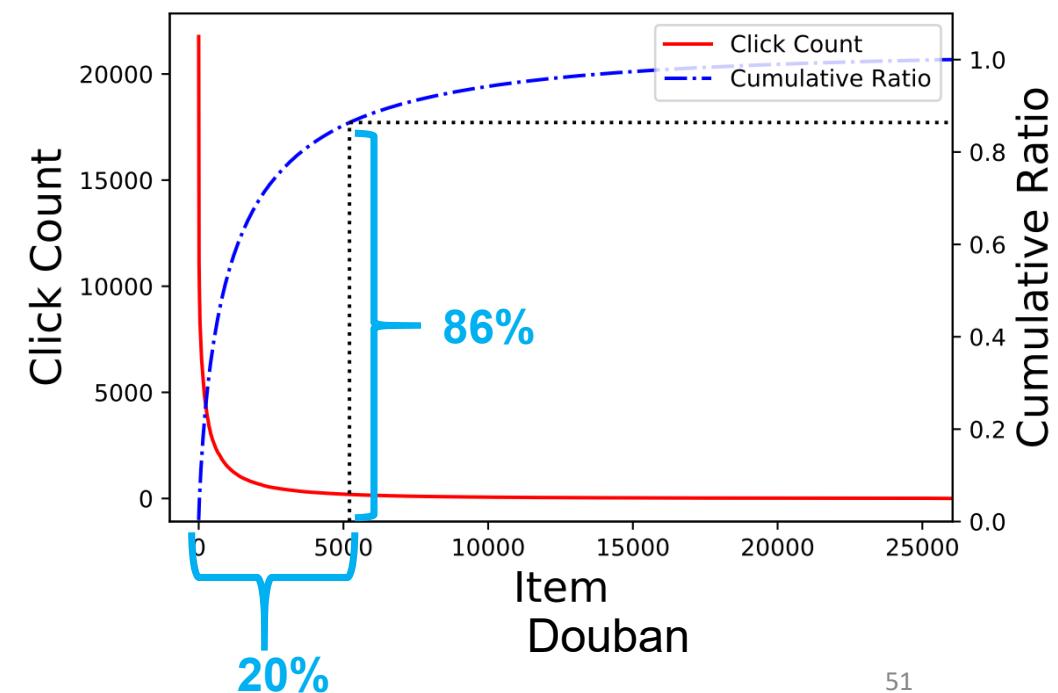
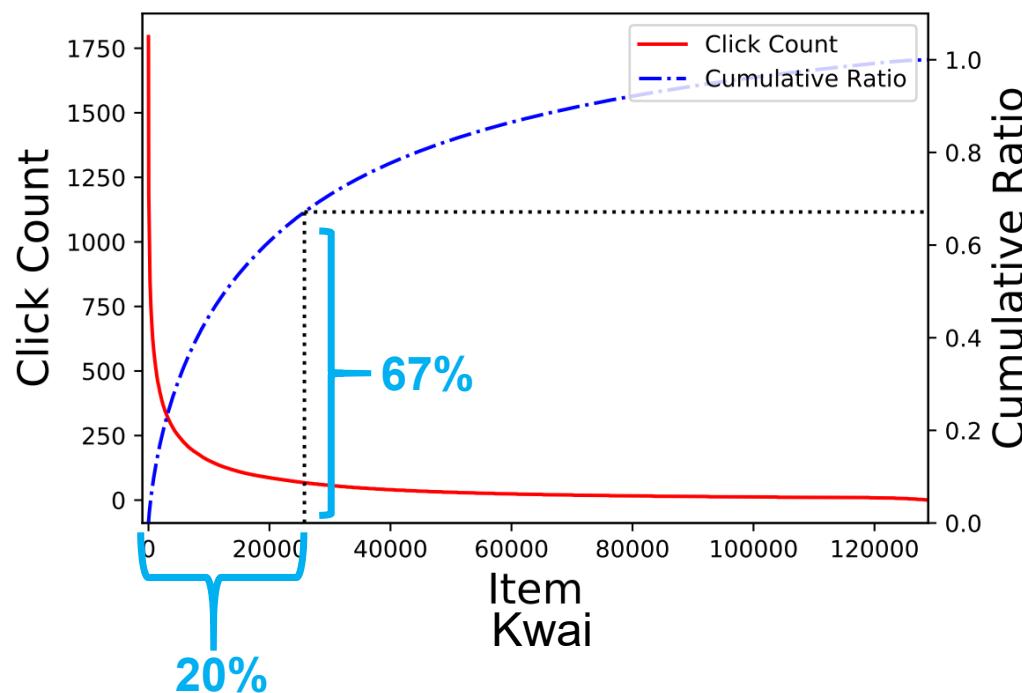
➤ Definitions [1]:

- Popularity bias refers to the problem where the recommendation algorithm **favors a few popular items** while not giving deserved attention to the majority of other items.
  
- Popularity bias is a well-known phenomenon in recommender systems where popular items are recommended even more frequently than their popularity would warrant, **amplifying** long-tail effects already present in many recommendation domains.

# Source of Popularity Bias

## ➤ The Underlying Data

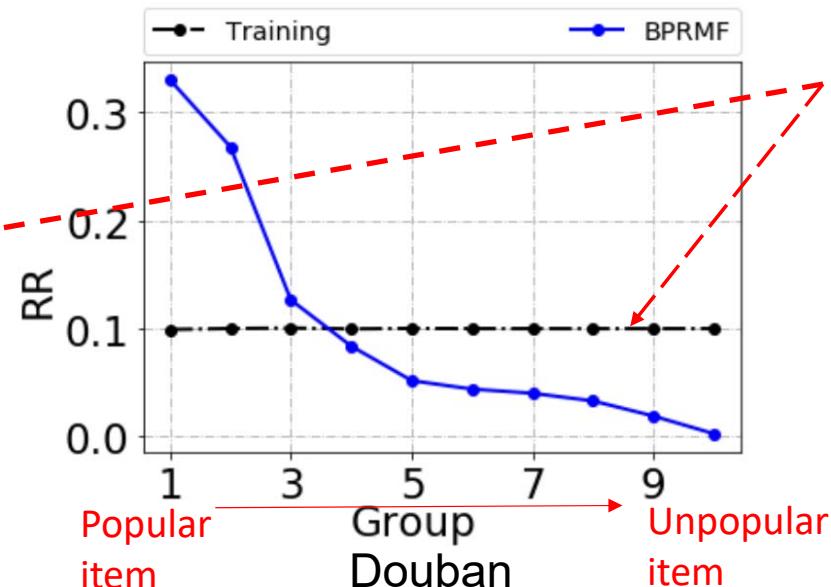
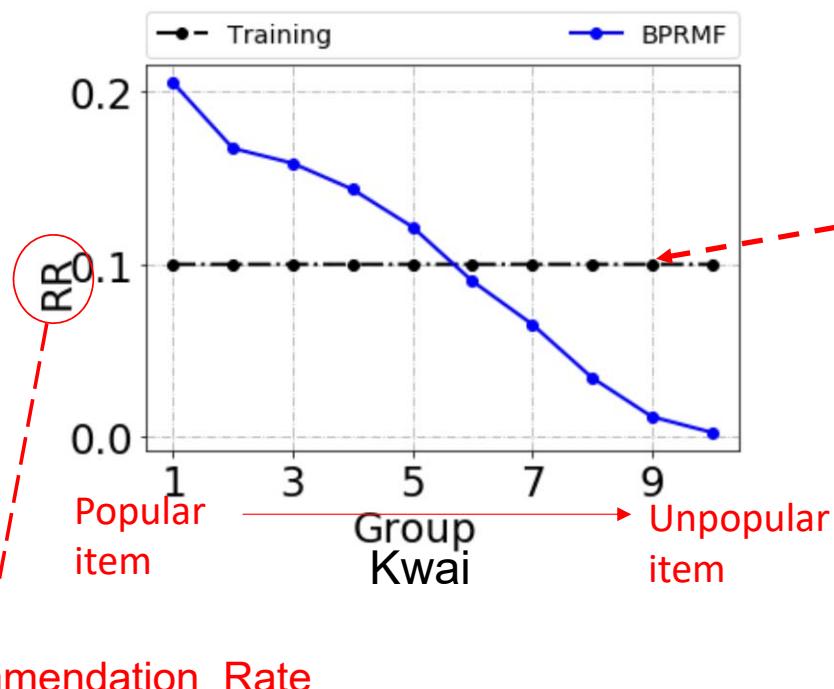
- ☐ Few popular items which take up the majority of rating interactions while the majority of the items receive small attention from the users.



# Source of Popularity Bias

## ➤ Algorithmic Bias

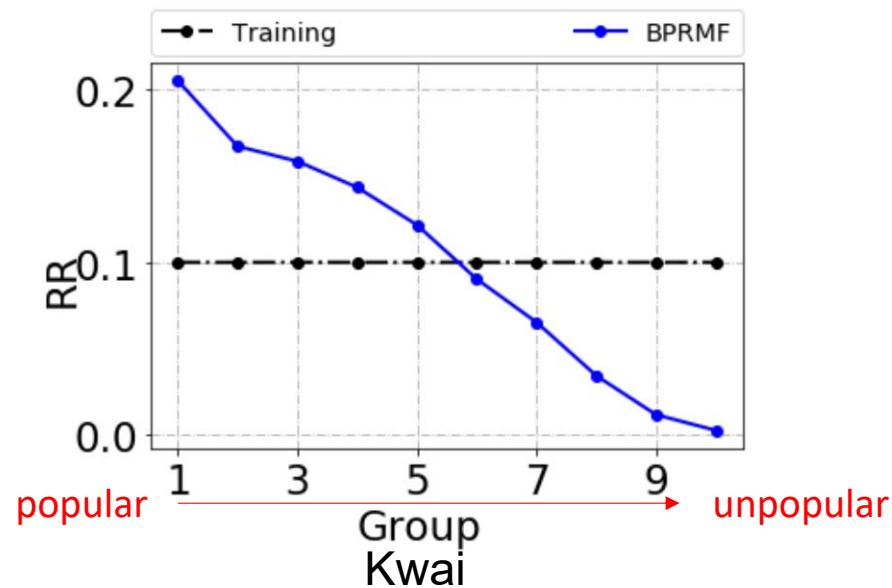
- Not only inherit bias from data, but also amplify the bias.  
— the rich get richer and the poor get poorer



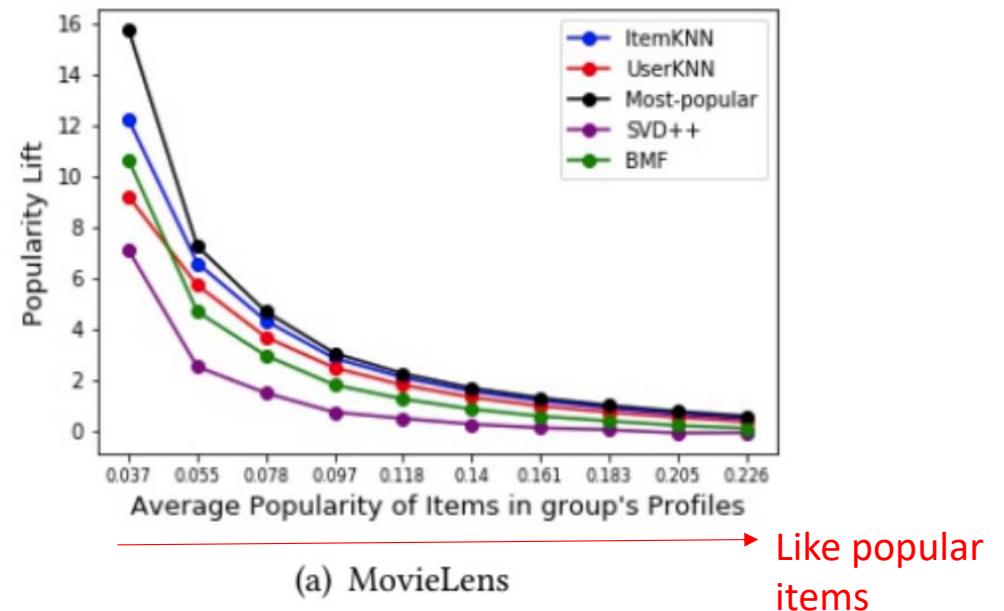
Each group has the same number of interactions in the training set

# Impacts of Popularity Bias

## ➤ Item-side



## User-side [1]



Matthew effect; Amplified interests for popular items; Unfairness for both users and items

[1]. Abdollahpouri, Himan, et al. "The Connection Between Popularity Bias, Calibration, and Fairness in Recommendation." Fourteenth ACM Conference on Recommender Systems. 2020.



# Methods for Popularity Bias

- Ranking Adjustment --- balance the recommendation lists
  - Regularization
  - Re-ranking
- Causal Embedding --- utilize causal-specific data
  - Disentanglement
- Causal Inference --- control the causal-effect of popularity
  - Inverse Propensity Score
  - Intervention
  - Counterfactual

# Ranking Adjustment

## ➤ Regularization

Key: push the model towards ‘balanced’ recommendation lists by regularization

$$\min_{\{P,Q\}} L_{acc}(P, Q) + \lambda L_{pop\_reg}(P, Q)$$

Recommendation Loss

Regularization term for adjusting recommendation list

### □ $L_{pop\_reg}$

✓ **Fairness-aware** [1] :  $tr(Q^\top L_D Q) \Rightarrow \min_{L_u} \frac{1}{N(N-1)} \sum_{ij \in R_u} d_{ij}$

where  $R_u$  is recommendation list, and  $D_{ij} = d_{ij} = \begin{cases} 0, & pop(i) \neq pop(j) \\ 1, & pop(i) = pop(j) \end{cases}$

✓ **Decorrelation** [2] :  $PCC(\hat{R}, pop(I))^2$

where  $\hat{R} = P^\top Q$ ,  $pop(I)$  is the popularity of  $I$ ,  
and PCC is Pearson Correlation Coefficient

[1]. Abdollahpouri, Himan et.al. "Controlling popularity bias in learning-to-rank recommendation." In RecSys 2017.

[2]. Ziwei zhu et.al. "Popularity-Opportunity Bias in Collaborative Filtering." In WSDM 2021.

# Ranking Adjustment

## ➤ Re-ranking

Key: Modify the ranking score to adjust the ranking list

$$\operatorname{argmax}_i \hat{R}_{int}(u, i) + \lambda \hat{R}_{pop}(u, i)$$

model score      adjusting score

### □ $\hat{R}_{pop}$

- ✓ **Popularity Compensation [1]** :  $C_{u,i} * \frac{n_u}{m_u}$

Where  $C_{u,i} = \frac{1}{pop(i)}(\hat{R}_{int}(u, i)\beta + 1 - \beta)$ ,  $\frac{n_u}{m_u}$  is the re-scaling coefficient

- ✓ **List smoothing [2]** :  $\sum_{c \in \{F, F'\}} P(c|u)p(i|c) \prod_{j \in S} (1 - P(j|c, S))$

$F, F'$ : popular or unpopular     $P(c|u)$ : user interests for the popular (unpopular)

$p(i|c)$ : category of item i       $\prod_{j \in S} (1 - P(j|c, S))$ : list state regarding popularity

[1] Ziwei zhu et.al. "Popularity-Opportunity Bias in Collaborative Filtering." In WSDM 2021.

[2] Abdollahpouri et.al. "Managing popularity bias in recommender systems with personalized re-ranking." In FLAIRS 2019.



# Methods for Popularity Bias

- Ranking Adjustment --- balance the recommendation lists
  - Regularization
  - Re-ranking
- Causal Embedding --- utilize causal-specific data
  - Disentanglement
- Causal Inference --- control the causal-effect of popularity
  - Inverse Propensity Score
  - Intervention
  - Counterfactual

# Causal Embedding

## ➤ Bias-free uniform data

Key: utilizing causal-specific data to guide model learning [1]

□ Even data(Causale) :

On even data	On biased data
$\min_{\mathcal{W}_c, \mathcal{W}_t} \frac{1}{ S_c } \sum_{(i,j) \in S_c} \ell(y_{ij}, \hat{y}_{ij}^c)$	$+ \frac{1}{ S_t } \sum_{(i,j) \in S_t} \ell(y_{ij}, \hat{y}_{ij}^t) +$
$\lambda_c R(\mathcal{W}_c) + \lambda_t R(\mathcal{W}_t) +$	$\lambda_{tc}^{CausE} \ \mathcal{W}_t - \mathcal{W}_c\ _F^2,$

Guiding term

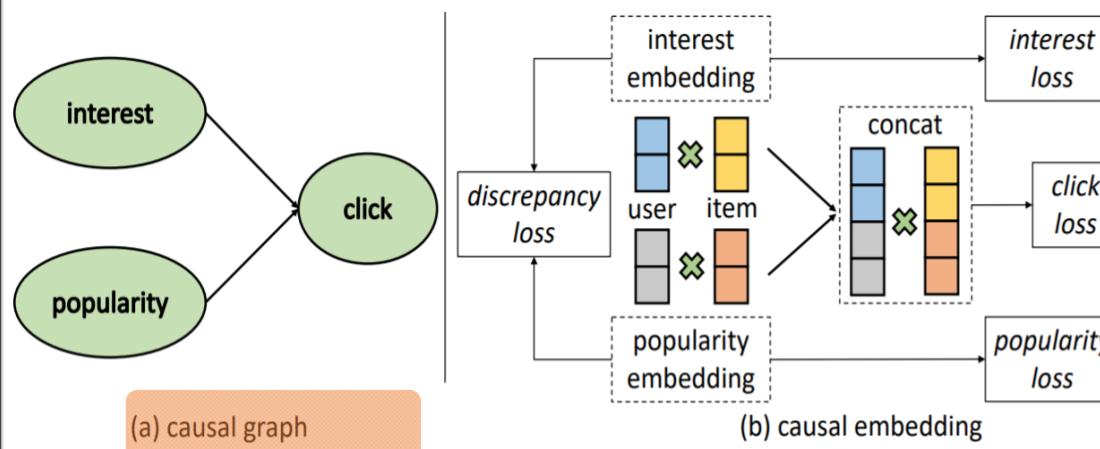
[1] Bonner, Stephen et.al. "Causal embeddings for recommendation." In RecSys 2018.

[2] Liu, Dugang, et al. "A general knowledge distillation framework for counterfactual recommendation via uniform data." In SIGIR 2020.

# Causal Embedding

## ➤ Pairwise causal-specific data — DICE

Key: Disentangle user interest and item popularity.



A click record reflects one or both aspects:

- the item's characteristics match the user's interest
- The item's popularity matches the user's conformity

$$L_{\text{interest}} = \sum_{(u,i,j) \in \mathcal{O}_1} \text{BPR}(\langle u^{(\text{int})}, i^{(\text{int})} \rangle, \langle u^{(\text{int})}, j^{(\text{int})} \rangle).$$

$$L_{\text{click}} = \sum_{(u,i,j) \in \mathcal{O}} \text{BPR}(\langle u^{(\text{int})} \| u^{(\text{pop})}, i^{(\text{int})} \| i^{(\text{pop})} \rangle, \langle u^{(\text{int})} \| u^{(\text{pop})}, j^{(\text{int})} \| j^{(\text{pop})} \rangle)$$

$$L_{\text{popularity}}^{(1)} = \sum_{(u,i,j) \in \mathcal{O}_1} -\text{BPR}(\langle u^{(\text{pop})}, i^{(\text{pop})} \rangle, \langle u^{(\text{pop})}, j^{(\text{pop})} \rangle),$$

$$L_{\text{popularity}}^{(2)} = \sum_{(u,i,j) \in \mathcal{O}_2} \text{BPR}(\langle u^{(\text{pop})}, i^{(\text{pop})} \rangle, \langle u^{(\text{pop})}, j^{(\text{pop})} \rangle),$$

$$L_{\text{popularity}} = L_{\text{popularity}}^{(1)} + L_{\text{popularity}}^{(2)}.$$

•  $\mathcal{O}$ : whole training set of triplets  $(u, i, j)$ : user, pos item, neg item

•  $\mathcal{O}_1$ : positive samples are less popular than negative samples

•  $\mathcal{O}_2$ : positive samples are more popular than negative samples

$$\mathcal{O} = \mathcal{O}_1 + \mathcal{O}_2$$



# Methods for Popularity Bias

- Ranking Adjustment --- balance the recommendation lists
  - Regularization
  - Re-ranking
- Causal Embedding --- utilize causal-specific data
  - Disentanglement
- Causal Inference --- control the causal-effect of popularity
  - Inverse Propensity Score
  - Intervention
  - Counterfactual

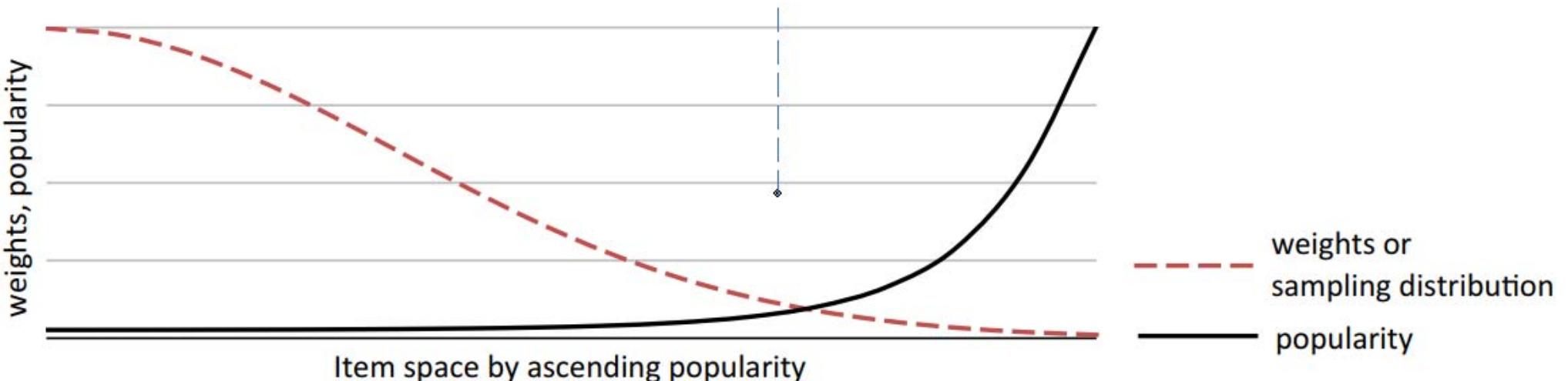
# Causal Inference

## ➤ Inverse Propensity Scoring (IPS)[1,2]

Key: adjust the distribution of training data

$$Loss = \frac{1}{N} \sum \frac{1}{ps(i)} \delta(u, i)$$

Impose lower weights on popular items, and boost unpopular items



[1] Jannach, Dietmar, et al. "What recommenders recommend: an analysis of recommendation biases and possible countermeasures." *User Modeling and User-Adapted Interaction* 25.5 (2015): 427-491.

[2] Schnabel, Tobias, et al. "Recommendations as treatments: Debiasing learning and evaluation." *international conference on machine learning*. PMLR, 2016.

# Causal Inference

## ➤ Basic Concepts in Causal Theory [1]

### □ Causal Graph:

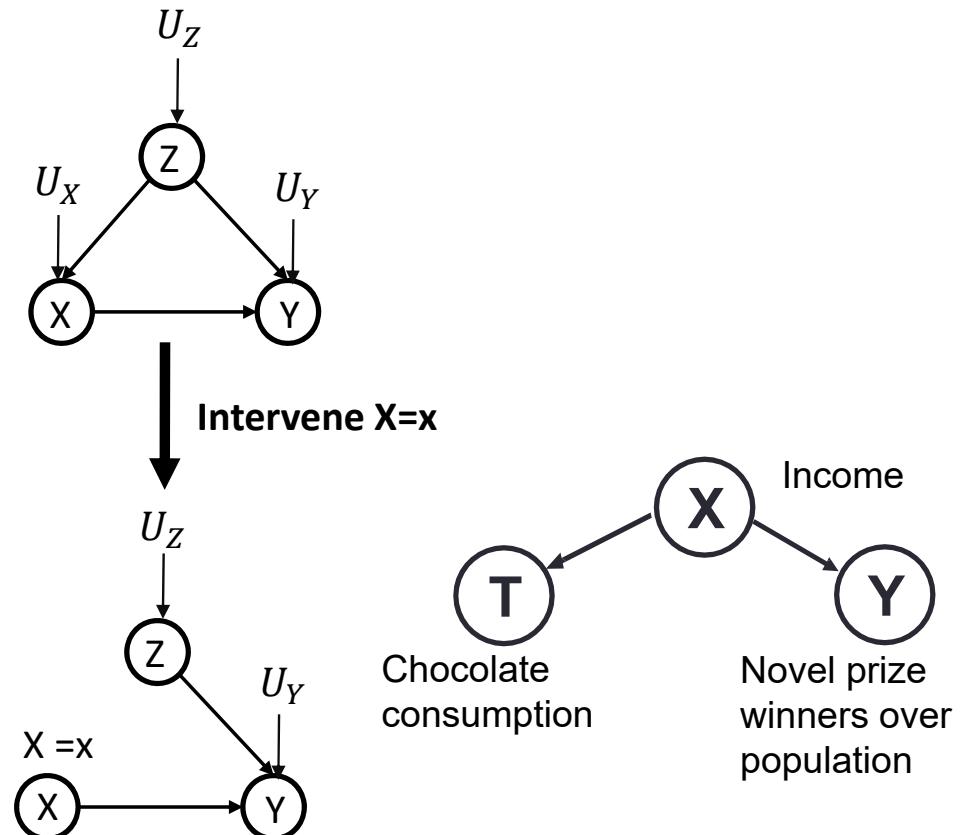
Graphical models used to encode assumptions about the data-generating process.

### □ Intervention on X [ term: $\text{do}(X=x)$ ]

Study specific causal relationships between X and the target variable.

Randomized controlled trial.

In graph: Cut off the paths that point into X



[1]. Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

# Causal Inference

## ➤ Basic Concepts in Causal Theory [1]

### □ Causal Effect:

$$P(Y | \text{do}(X=x)) - P(Y | \text{do}(X=x_{ref}))$$

measures the expected increase in Y as the treatment changes from  $X = x$  to  $X=x_{ref}$

**General causal effect:**  $P(Y | \text{do}(X=x))$

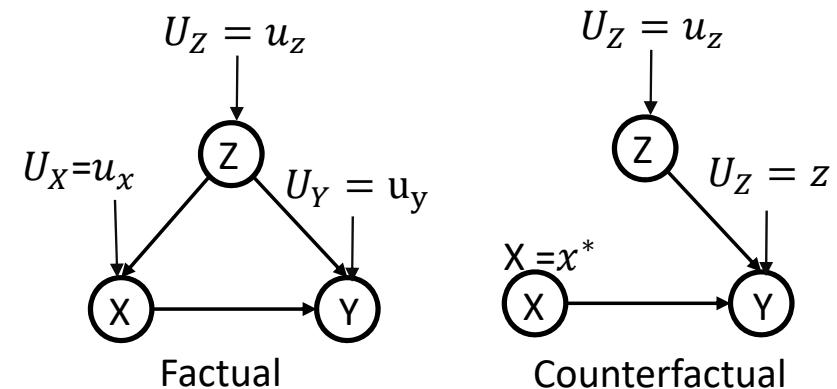
**Others:** NIE, NDE, TIE ...

### □ Counterfactual

Imagine a world that does not really existed, given existed information.

Observed  $Y=y_1$ , assume the  $X$  is  $x^*$ , what will the  $Y$  is?

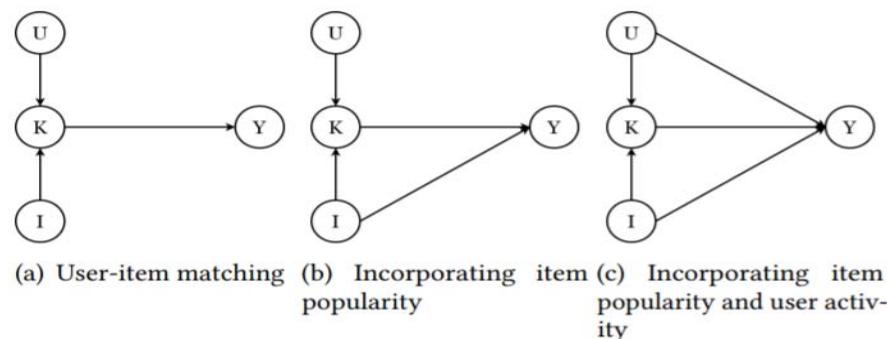
- **Abduction:** Based on  $Y=y_1$ , inference  $U_Y = u_y, U_Z = u_z$
- **Action:** Let  $X=x^*$
- **Prediction:**  $Z = f_z(u_z), X = x^*, Y = f_Y(f_z(u_z), x^*, u_y)$



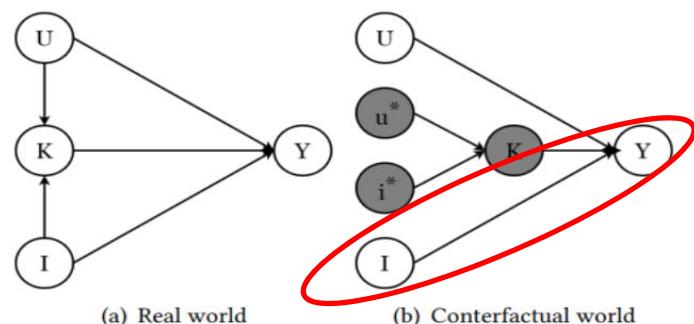
[1]. Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

# Causal Inference

## ➤ Counterfactual Inference — MACR



**Figure 2: Causal graph for (a) user-item matching; (b) incorporating item popularity; and (c) incorporating item popularity and user activity.**  
**I:** item. **U:** user. **K:** matching features between user and item.  
**Y:** ranking score (e.g., the probability of interaction).



- Counterfactual inference:

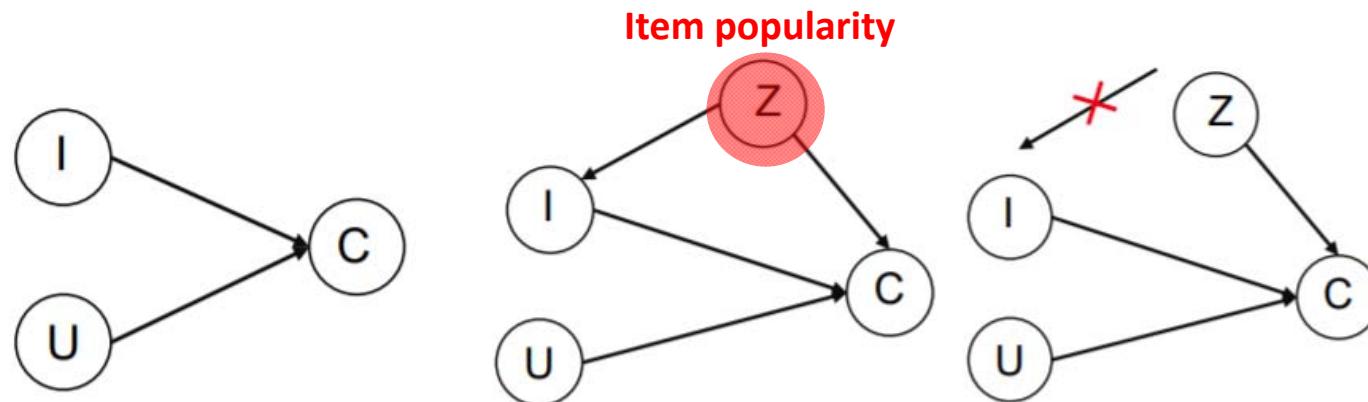
$$\frac{\hat{y}_k * \sigma(\hat{y}_i) * \sigma(\hat{y}_u) - c * \sigma(\hat{y}_i) * \sigma(\hat{y}_u)}{\text{Factual prediction} \quad \text{Counterfactual prediction}}$$

# Causal Inference

- De-confounding —— Popularity De-confounding(PD) and Adjusting (PDA)

Key: item popularity is a confounder, both bad and good effect of popularity exist.

Leverage popularity bias instead of blindly removing.



(a) Causal graph of traditional methods.

(b) Causal graph that considers item popularity.

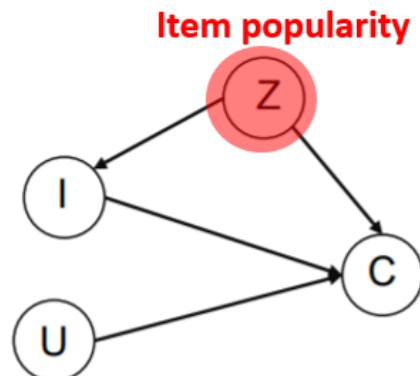
(c) We cut off  $Z \rightarrow I$  for model training

We estimate the user-item matching as  $P(C|do(U, I))$  based on figure (c)

“Causal Intervention for Leveraging Popularity Bias in Recommendation.” under submission

# Causal Inference

## ➤ PD --- Popularity De-confounding



**Causality:**

$$P(C|do(U, I)) = \sum_Z P(C|U, I, Z)P(Z)$$

**vs**

**Correlation:**

$$P(C|U, I) = \sum_Z P(C|U, I, Z)P(Z|U, I)$$

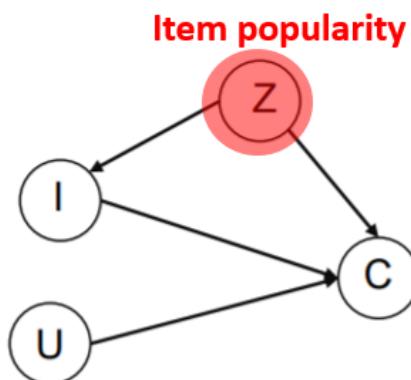
### □ De-confounding --- estimate $P(C|do(U, I))$ :

- **Step 1.** estimate  $P(C|U, I, Z)$ 
    - $P_\Theta(c = 1|u, i, m_i^t) = ELU'(f_\Theta(u, i)) \times (m_i^t)^\gamma$
    - $m_i^t$ : the popularity of item  $i$  in timestamp  $t$
    - $f_\Theta(u, i)$ : user-item matching, such as MF
    - Learning this component from data
  - **Step 2.** computing  $P(C|do(U, I))$ 
    - $\sum_Z P(C|U, I, Z)P(Z) \propto ELU'(f_\Theta(u, i))$
    - ranking with this term
- ✓ In pursuit of real interests instead of even state!  
Higher popularity because of better quality.

# Causal Inference

## ➤ PDA --- Popularity De-confounding and Adjusting

- We have estimated  $P(C|do(U, I))$ , which does not chase the even state but the real interests.
- Is it enough?
  - No... In some time, we need inject some desired popularity.
  - Such as we can recommend more item that will be popular if we can know the trends of popularity.



Introducing popularity bias by intervention:

$$P(C|do(U, I), do(Z = \tilde{Z})) = P(C|U, I, \tilde{Z})$$
$$P(C|U, I, \tilde{Z}) = ELU'(f_{\Theta}(u, i)) \times (\tilde{Z}_i)^{\gamma}$$

$\tilde{Z}$ : predicted by the trends of item popularity.

# Causal Inference

## ➤ Experimental Setting

### ■ Datasets:

Dataset	#User	#Item	#Interaction	#Sparsity	#type
Kwai	37,663	128,879	7,658,510	0.158%	Click
Douban	47,890	26,047	7,174,218	0.575%	Review
Tencent	80,339	27,070	1,816,046	0.084%	Like

### ■ Data Splitting:

Temporal splitting --- split each into 10 time stages according to timestamp.  
0-8th stages: training, 9th stage: validation & testing.

### ■ Evaluation Setting:

PD: directly test

PDA: Most recent stages can be utilized to predict future popularity.

### ■ Baselines:

PD: MostPop, BPRMF, xQuad(2019FLAIRS), BPR-PC(2021WSDM), DICE(2021WWW)

PDA: MostRecent(2020SIGIR), BPRMF(t)-pop(2017RecTemp@ RecSys), BPRMF-A, DICE-A

# Causal Inference

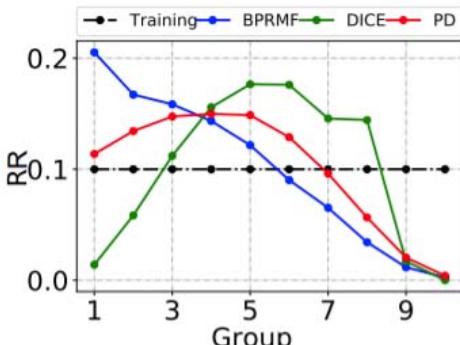
## ➤ Results for PD

Dataset	Methods	Top 20				Top 50			
		Recall	Precision	HR	NDCG	RI	Recall	Precision	HR
Kwai	MostPop	0.0014	0.0019	0.0341	0.0030	632.4%	0.0040	0.0021	0.0802
	BPRMF	0.0054	<u>0.0057</u>	0.0943	0.0067	146.3%	0.0125	<u>0.0053</u>	0.1866
	xQuad	0.0054	<u>0.0057</u>	0.0948	0.0068	145.0%	0.0125	<u>0.0053</u>	0.1867
	BPR-PC	<u>0.0070</u>	0.0056	<u>0.0992</u>	<u>0.0072</u>	125.0%	<u>0.0137</u>	0.0046	0.1813
	DICE	0.0053	0.0056	0.0957	0.0067	147.8%	0.0130	0.0052	0.1872
	PD	<b>0.0143</b>	<b>0.0138</b>	<b>0.2018</b>	<b>0.0177</b>	-	<b>0.0293</b>	<b>0.0118</b>	<b>0.3397</b>
Douban	MostPop	0.0218	0.0297	0.2373	0.0349	75.4%	0.0490	0.0256	0.3737
	BPRMF	0.0274	<u>0.0336</u>	0.2888	0.0405	47.0%	0.0581	<u>0.0291</u>	0.4280
	xQuad	0.0274	<u>0.0336</u>	<u>0.2895</u>	0.0391	48.3%	0.0581	<u>0.0291</u>	<u>0.4281</u>
	BPR-PC	<u>0.0282</u>	0.0307	0.2863	0.0381	51.6%	<u>0.0582</u>	0.0271	0.4260
	DICE	0.0273	<u>0.0336</u>	0.2845	<u>0.0421</u>	46.2%	0.0513	0.0273	0.4000
	PD	<b>0.0453</b>	<b>0.0454</b>	<b>0.3970</b>	<b>0.0607</b>	-	<b>0.0843</b>	<b>0.0362</b>	<b>0.5271</b>
Tencent	MostPop	0.0145	0.0043	0.0684	0.0093	340.8%	0.0282	0.0035	0.1181
	BPRMF	0.0553	<u>0.0153</u>	0.2005	0.0328	27.1%	0.1130	<u>0.0129</u>	0.3303
	xQuad	0.0552	<u>0.0153</u>	0.2007	0.0326	27.3%	0.1130	<u>0.0129</u>	0.3302
	BPR-PC	<u>0.0556</u>	<u>0.0153</u>	<u>0.2018</u>	<u>0.0331</u>	26.5%	<u>0.1141</u>	0.0128	<u>0.3322</u>
	DICE	0.0516	0.0149	0.1948	0.0312	32.8%	0.1010	0.0132	0.3312
	PD	<b>0.0715</b>	<b>0.0195</b>	<b>0.2421</b>	<b>0.0429</b>	-	<b>0.1436</b>	<b>0.0165</b>	<b>0.3875</b>

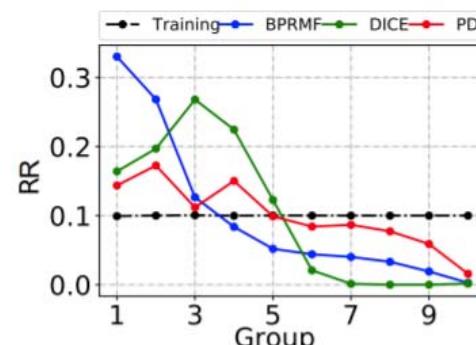
The power of de-confounded estimation !!

# Our Works

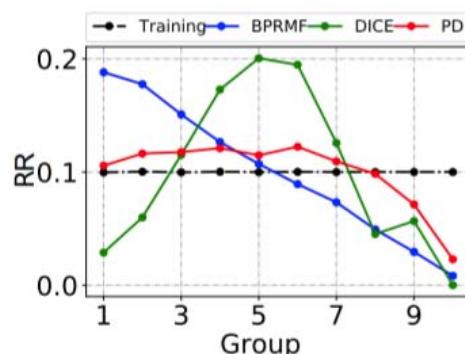
## ➤ PD —— Recommendation Analysis.



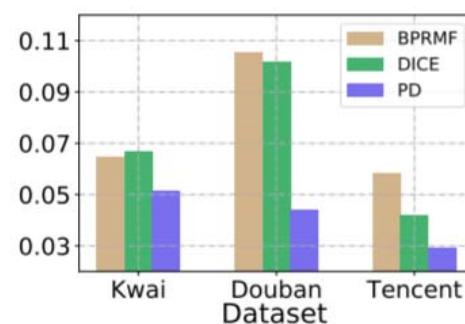
(a) Kwai



(b) Douban



(c) Tencent



(d) std. dev.

Figure 4: Recommendation rate(RR) over item groups.

- Less amplification for most popular groups compared with BPRMF
- Do not **over-suppress** the most popular groups compared with DICE
- More flat lines and standard deviations over different groups
  - **relative fair recommendation opportunities** for different group (refer to training set)
- Better performance
  - remove bad effect but keep good effect of popularity bias

# Causal Inference

## ➤ Results for PDA

**Table 2: Top-K recommendation performance with popularity adjusting on Kwai, Douban, and Tencent Datasets.**

Datasets		Kwai				Douban				Tencent			
Methods		top 20		top 50		top 20		top 50		top 20		top 50	
		Recall	NDCG										
MostRecent		0.0074	0.0096	0.0139	0.011	0.0398	0.0582	0.0711	0.0615	0.0360	0.0222	0.0849	0.0359
BPRMF(t)-pop		0.0188	0.0241	0.0372	0.0286	0.0495	0.0682	0.0929	0.0760	0.1150	0.0726	0.2082	0.1001
BPRMF-A	(a)	0.0191	0.0249	0.0372	0.0292	0.0482	0.0666	0.0898	0.0744	0.1021	0.0676	0.1805	0.0905
	(b)	0.0201	0.0265	0.0387	0.0306	0.0486	0.0667	0.0901	0.0746	0.1072	0.0719	0.1886	0.0953
DICE-A	(a)	0.0242	0.0315	0.0454	0.0363	0.0494	0.0681	0.0890	0.0736	0.1227	0.0807	0.2161	0.1081
	(b)	0.0245	0.0323	0.0462	0.0370	0.0494	0.0680	0.0882	0.0734	0.1249	0.0839	0.2209	0.1116
PDA	(a)	<u>0.0279</u>	<u>0.0352</u>	<u>0.0531</u>	<u>0.0413</u>	<u>0.0564</u>	<b>0.0746</b>	<b>0.1066</b>	<b>0.0845</b>	<u>0.1357</u>	<u>0.0873</u>	<u>0.2378</u>	<u>0.1173</u>
	(b)	<b>0.0288</b>	<b>0.3364</b>	<b>0.054</b>	<b>0.0429</b>	<b>0.0565</b>	<u>0.0745</u>	<b>0.1066</b>	<u>0.0843</u>	<b>0.1398</b>	<b>0.0912</b>	<b>0.2418</b>	<b>0.1210</b>

- Introducing desired popularity bias can improve the recommendation performance.
- Our method achieves the best performance.



# Conclusion & Future Work

## ➤ Conclusion

- Heuristic methods -- Not best
- Uniform/unbiased data -- Hard to get these data
- Causal perspective
  - IPS -- Hard to estimate Propensity Scores
  - Counterfactual & Intervention -- Extra assumption of causal graph
- Eliminate the bad effect of bias, leverage the good effect of bias.

## ➤ Potential directions

- Comprehensive causal graph.
- Accurate estimation of causal effect.
- Popularity bias with features of users and items.
- Considering popularity bias at finer-grain.



**Thanks!**

## • Tutorial Outline

### ❑ Biases in Data

- ❑ Definition of data biases
- ❑ Categories: Selection bias, Conformity bias, Exposure bias and Position bias
- ❑ Recent solutions for data biases

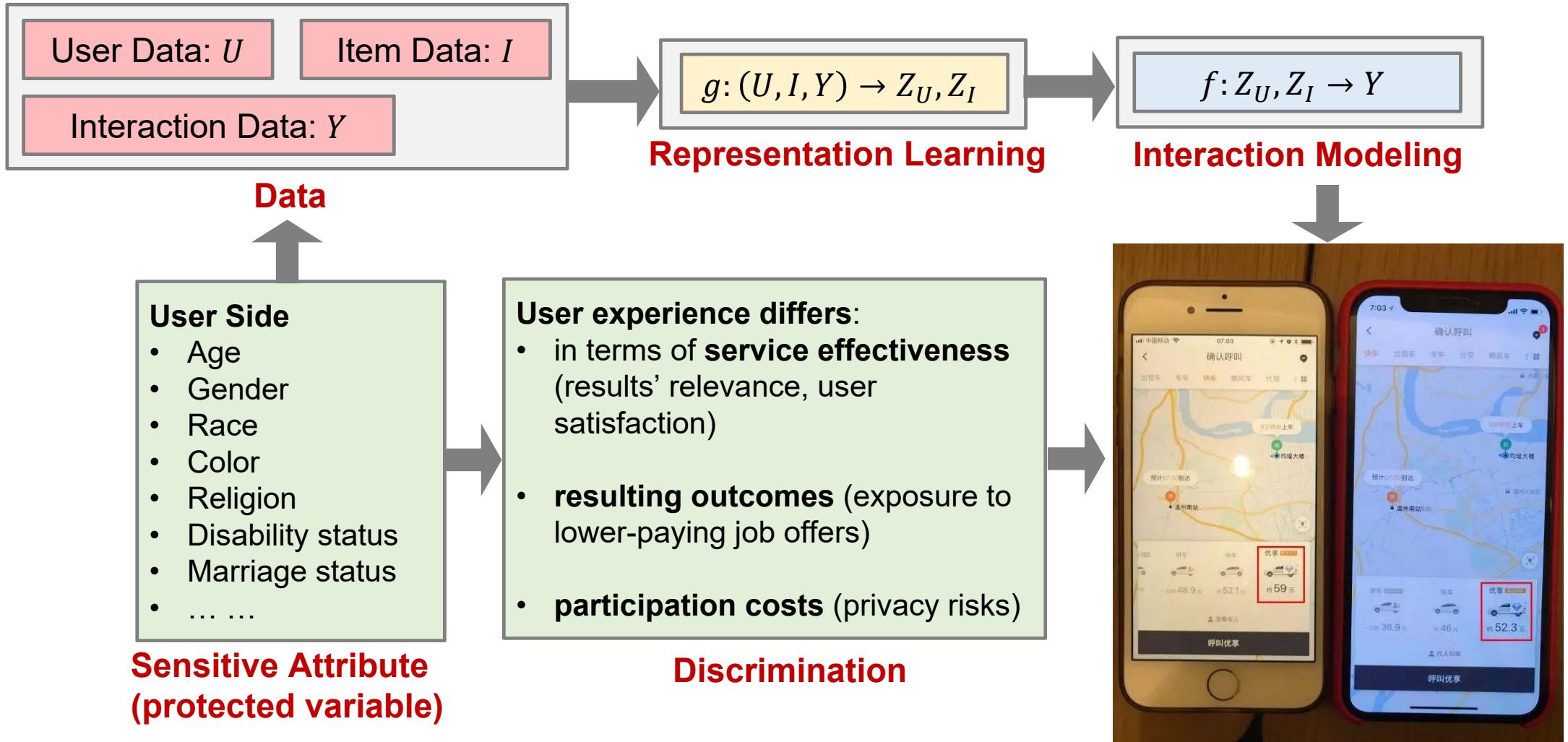
### ❑ Biases in Results

- ❑ Popularity bias: definition, characteristic and solutions
- ❑ Unfairness: definition, characteristic and solutions
- ❑ Bias Amplification in Loop and Its Solutions
- ❑ Summary and Future Direction

slides will be available at: <https://github.com/jiawei-chen/RecDebiasing>

A literature survey based on this tutorial is available at: <https://arxiv.org/pdf/2010.03240.pdf>

## • Sensitive Attributes in Fairness



## • Unfairness Leads to Discrimination

### Individual Fairness

“Similar individuals treated similarly”



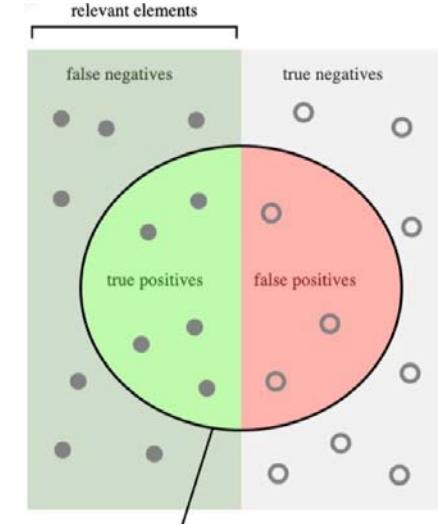
Basketball



Ping-pong ball

### Group Fairness

“Similar Classifier Statistics Across Groups”



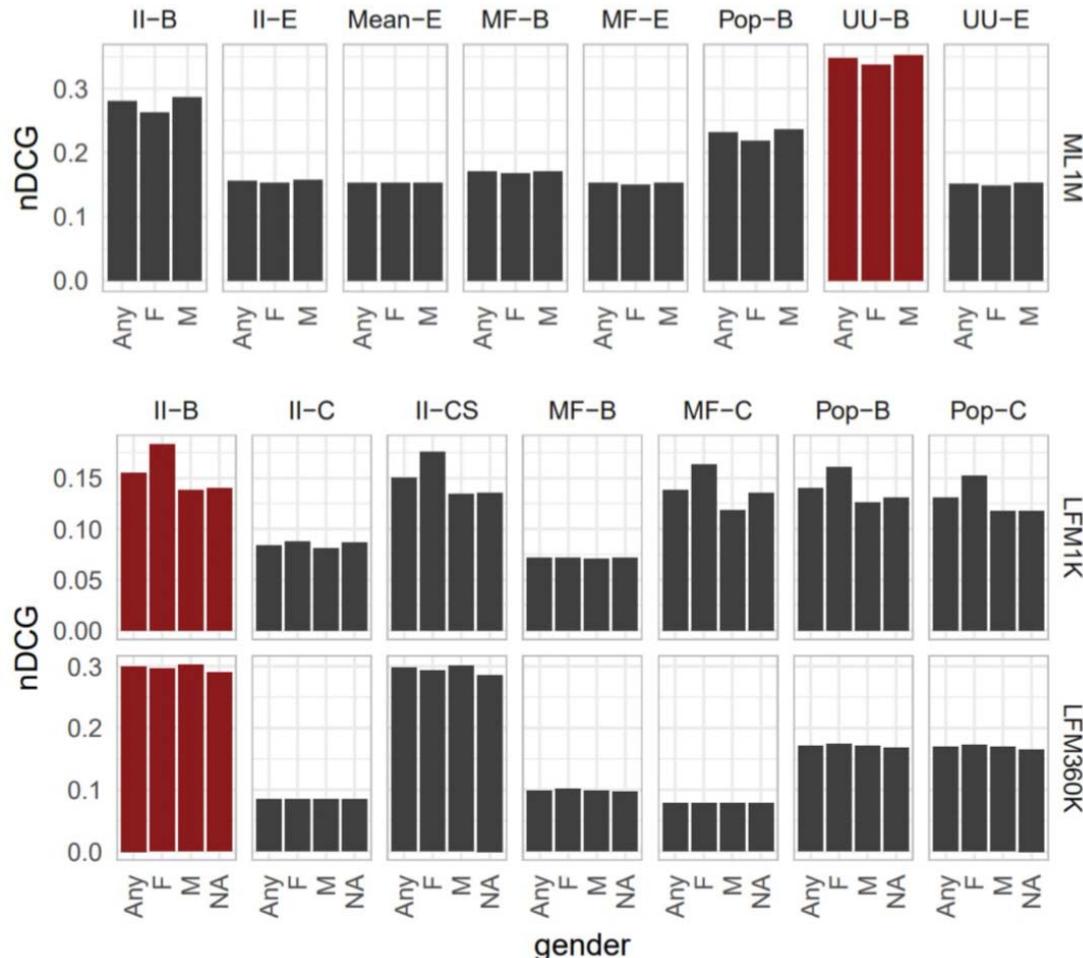
### Individual Discrimination

A model gives unfairly different predictions to similar individuals

### Group Discrimination

A model systematically treats individuals who belong to different groups unfairly

## • Case 1 in Recommendation



### Motivation

Investigating whether **demographic groups obtain different utility** from recommender systems in LastFM and MovieLens 1M datasets

### Findings

- MovieLens 1M & LastFM 1K have statistically-significant differences between **gender groups**
- LastFM 360K has significant differences between **age brackets**

## • Case 2 in Recommendation

User group	UserItemAvg	UserKNN	UserKNNAvg	NMF
LowMS	42.991***	49.813***	46.631***	<b>38.515***</b>
MedMS	33.934	42.527	37.623	<b>30.555</b>
HighMS	40.727	46.036	43.284	<b>37.305</b>
All	38.599	45.678	41.927	<b>34.895</b>

**Table 1.** MAE results (the lower, the better) for four personalized recommendation algorithms and our three user groups. The worst (i.e., highest) results are always given for the LowMS user group (statistically significant according to a t-test with  $p < .005$  as indicated by \*\*\*). Across the algorithms, the best (i.e., lowest) results are provided by NMF (indicated by bold numbers).

### Motivation

Investigating **three user groups** from Last.fm based on how much their listening preferences deviate from the most popular music:

- low-mainstream users
- medium-mainstream users
- high-mainstream users

### Findings

- Different user groups are treated differently
- **Low-mainstream user group** significantly receives the **worst** recommendations

## • Definitions of Fairness

### Fairness through Unawareness

A model is fair if **any sensitive attribute is not explicitly used** in the decision-making process

### Equal Opportunity

A model is fair if the groups have **equal true positive rates**

$$P(\hat{Y} = 1 | A = 0, Y = 1) = P(\hat{Y} = 1 | A = 1, Y = 1)$$

### Demographic Parity

A model is fair if **the likelihood of a positive outcome** should be the same regardless of the group

$$P(\hat{Y}|A = 0) = P(\hat{Y}|A = 1)$$

### Individual Fairness

a model is fair if it gives similar predictions to **similar individuals**

$$\hat{Y}(X(i), A(i)) \approx \hat{Y}(X(j), A(j)), \text{if } |X(i) - X(j)| \leq \varepsilon$$

### Equalized Odds

A model is fair if the groups have **equal rates for true positives and false positives**

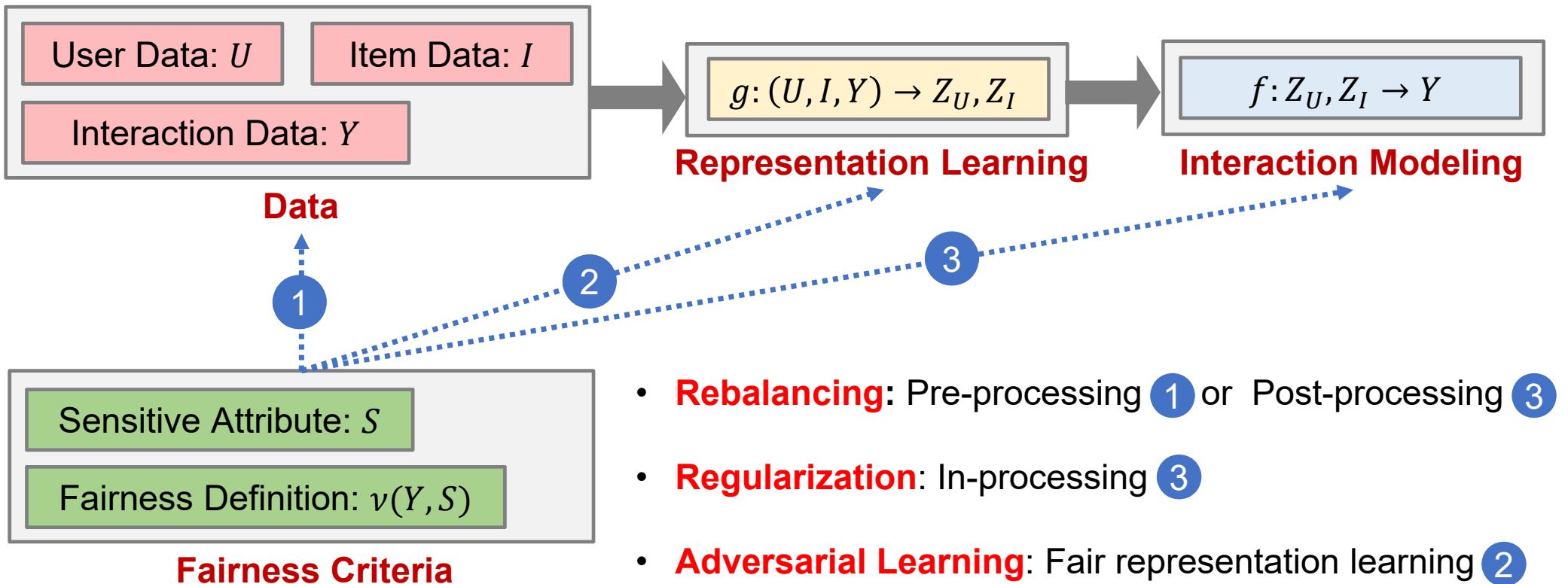
$$P(\hat{Y} = 1 | A = 0) = P(\hat{Y} = 1 | A = 1)$$

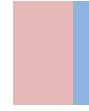
### Counterfactual Fairness

A model is fair towards an individual if it is the same **in both the actual world and a counterfactual world** where the individual belonged to a different demographic group

$$P(\hat{Y}|X = x, do(A) = 0) = P(\hat{Y}|X = x, do(A) = 1)$$

## • Four Research Lines towards Fairness

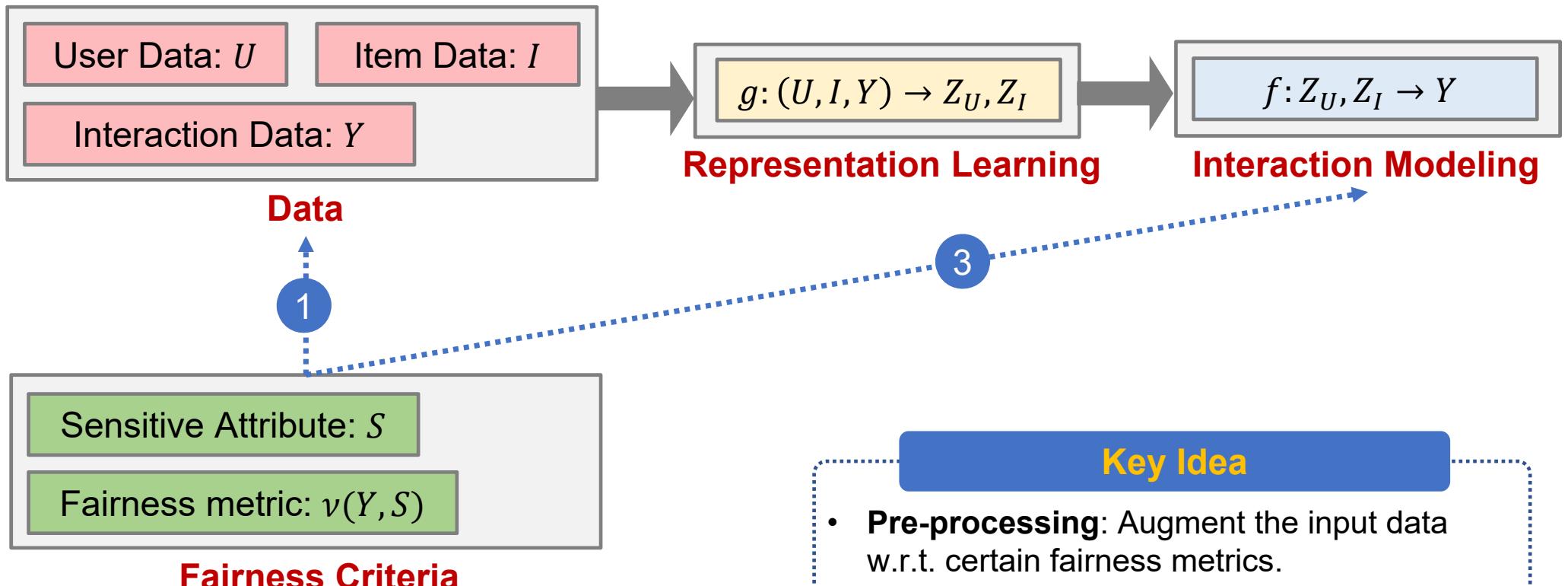




- Four Research Lines towards Fairness

- Rebalancing
- Regularization
- Fair Representation Learning

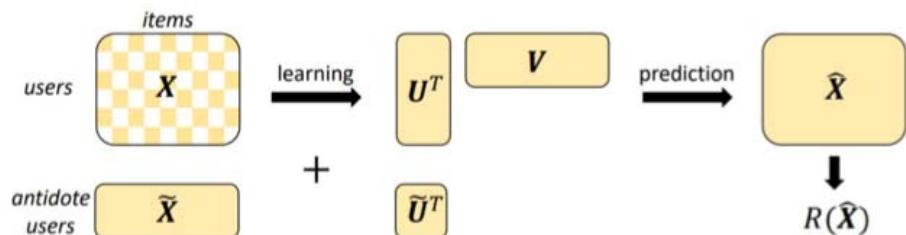
## • Line 1: Rebalancing



### Key Idea

- **Pre-processing:** Augment the input data w.r.t. certain fairness metrics.
- **Post-processing:** Balance the results w.r.t. certain fairness metrics.

## • Example 1: Pre-processing — Using Antidote Data



### Idea

Augmenting the input with **additional “antidote” data** can improve the social desirability of recommendations

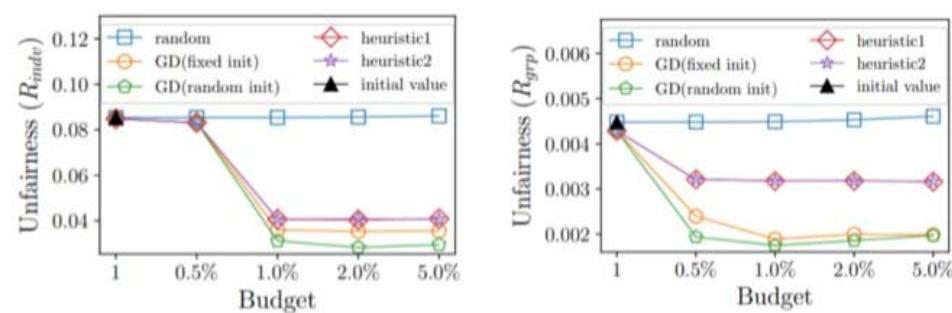


Figure 3: Improving fairness.

### Algorithms

MF family of algorithms

### Findings

- The small amounts of antidote data (typically on the order of 1% new users) can generate a dramatic improvement (on the order of 50%) in the polarization or the fairness of the system's recommendations

## • Example 2: Post-processing — Fairness-Aware Re-ranking

Personalization score determined by the base recommender

$$\max_{v \in R(u)} \underbrace{(1 - \lambda)P(v|u)}_{\text{personalization}} + \lambda \tau_u \sum_c P(\mathcal{V}_c) \mathbb{1}_{\{v \in \mathcal{V}_c\}} \prod_{i \in S(u)} \mathbb{1}_{\{i \notin \mathcal{V}_c\}},$$

personalized fairness

Importance of the group with attribute c

coverage of  $\mathcal{V}_c$  for the current generated re-ranked list  $S(u)$

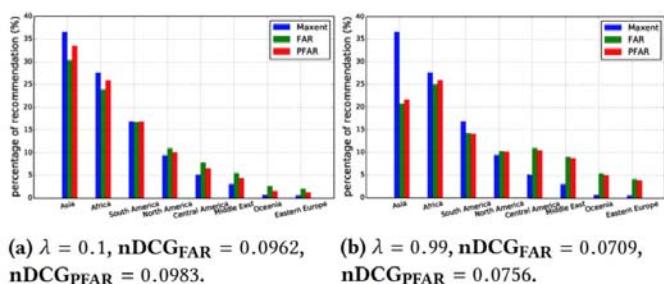
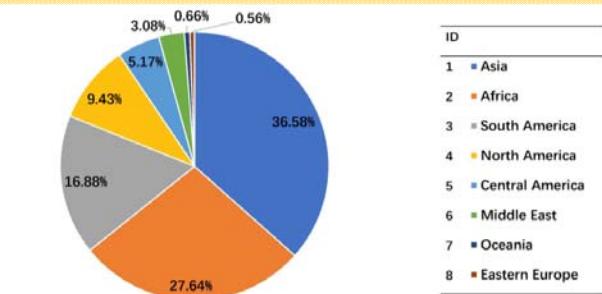


Figure 4: Recommendation percentage of each region.

[Liu et al.: Personalized fairness-aware re-ranking for microlending. RecSys 2019)]

### Idea

Combining a **personalization-induced term** & a **fairness-induced term** to promote the loans of currently uncovered borrower groups

### Algorithms

RankSGD, UserKNN, WRMF, Maxent

### Findings

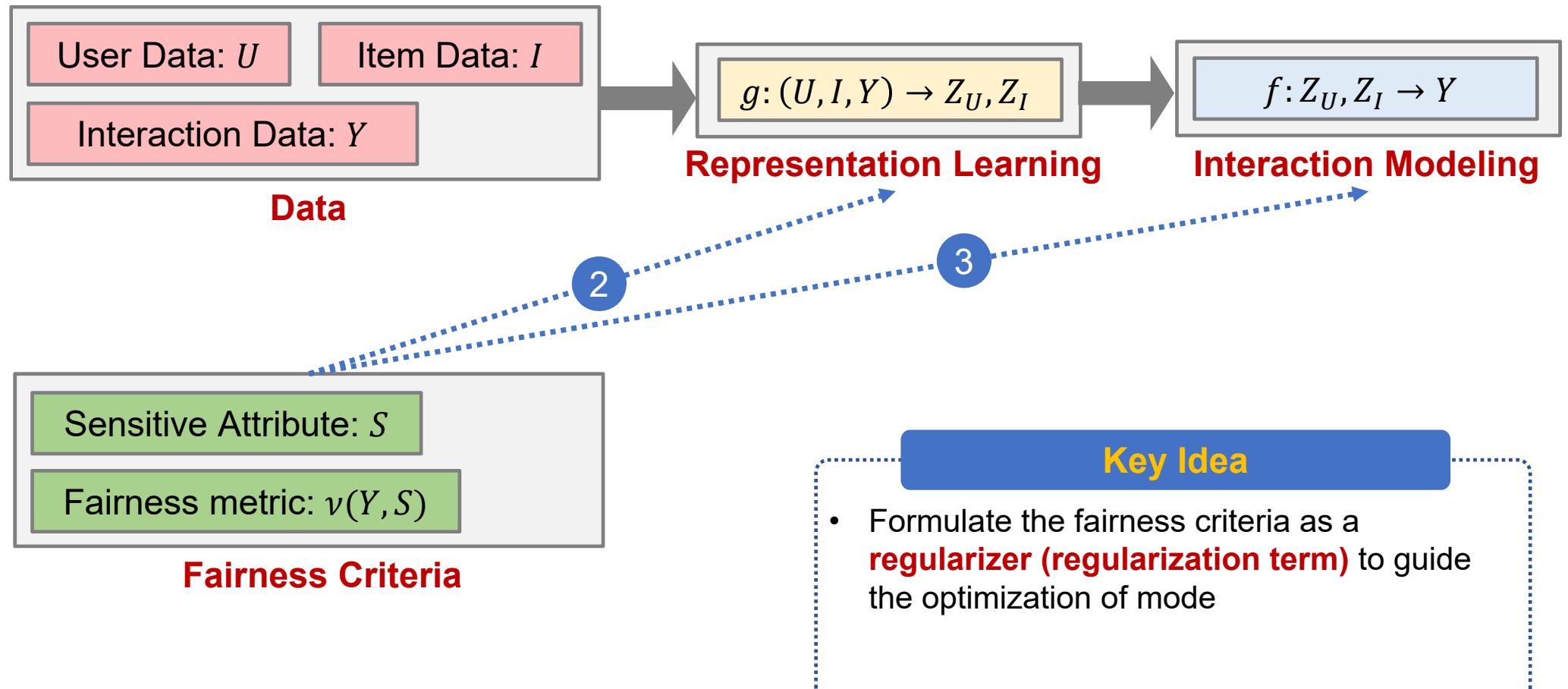
- A balance between the two terms
- **Recommendation accuracy** (nDCG) remains at a high level after the re-ranking
- **Recommendation fairness** is significantly improved → loans belonging to less-popular groups are promoted.



- Four Research Lines towards Fairness

- Rebalancing
- Regularization
- Fair Representation Learning

## • Line 2: Regularization



# • Example 1: Learned Fair Representation (LFR)

**Reconstruction loss**  
between input data X and representations R

$$\min \mathcal{L} = \alpha C(X, R) + \beta D(R, A) + \gamma E(Y, R)$$

**Prediction error** in generating prediction Y from R

**Regularization term** that measures the dependence between R and sensitive attribute A

**Fairness criteria** (e.g., demographic parity)

$$D(R, A) = |\mathbb{E}_R P(R|A = 1) - \mathbb{E}_R P(R|A = 0)|$$

**Distance** of representation R and the centroid representation of the group where A = 1

## Idea

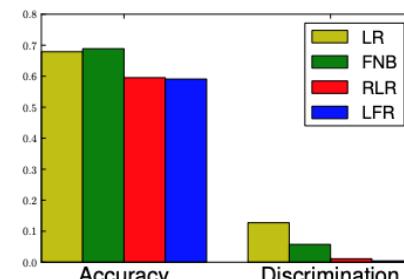
Representation Learning

- **encode insensitive attributes** of data
- **remove any information about sensitive attributes** w.r.t. the protected subgroup

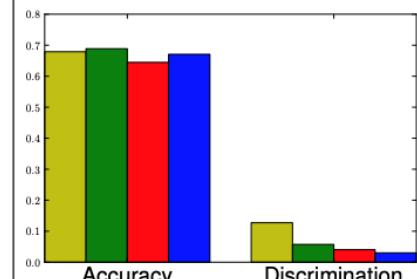
## Findings

- pushing the discrimination to very low values
- while maintaining fairly high accuracy

Min. Discrimination



Max. Delta



## • Example 2: Neutrality-Enhanced Recommendation

Loss of predicting ratings (e.g., squared error)

L2 regularization on model parameters

$$\mathcal{L}(\mathcal{D}) = \sum_{(x_i, y_i, s_i, v_i) \in \mathcal{D}} (s_i - \hat{s}(x_i, y_i, v_i))^2 + \eta I(\hat{s}; v) + \lambda R$$

**Neutrality function** to quantify the degree of the information neutrality from a viewpoint variable

**Independence** between the predictions & sensitive attributes → negative mutual information

$$-I(\hat{s}; v) = \sum_{v \in \{0,1\}} \int \Pr[\hat{s}, v] \log \frac{\Pr[\hat{s}|v]}{\Pr[\hat{s}]} d\hat{s}$$

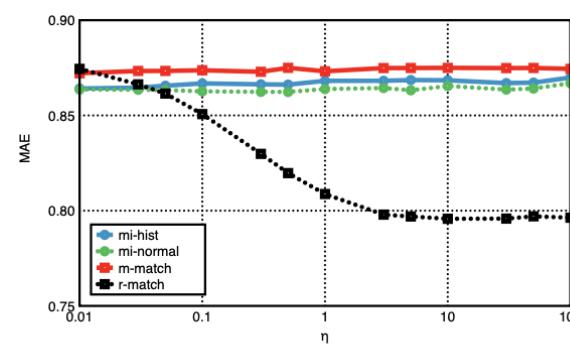
### Idea

Regularization term

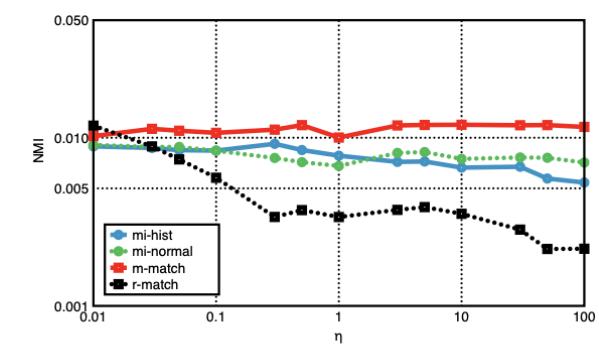
- **Negative mutual information** between sensitive attribute A and prediction Y

### Findings

- enhances the independence toward the specified sensitive attribute

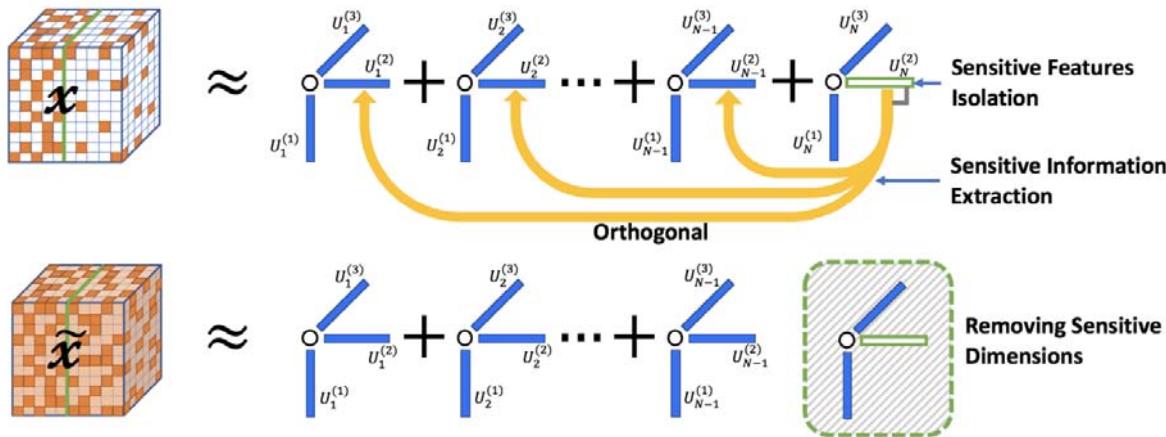


(c) Prediction error (MAE) for Gender data



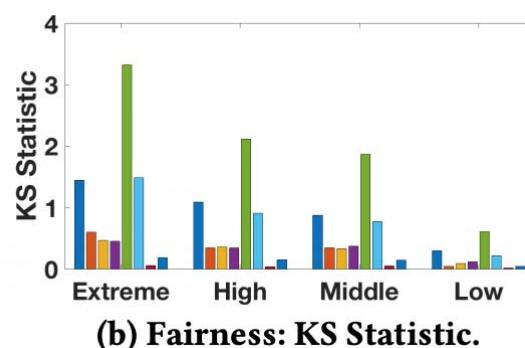
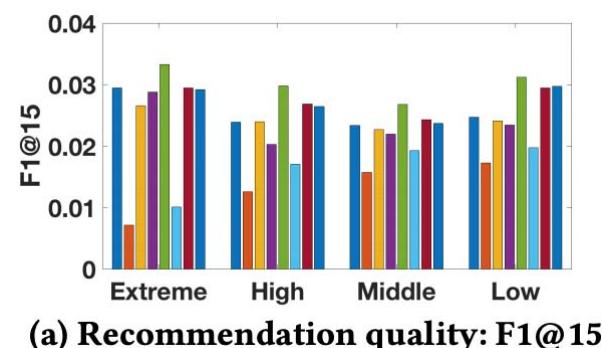
(d) Degree of neutrality (NMI) for Gender data

## • Example 3: Fairness-Aware Tensor-based Rec (FATR)



### Idea

- Use **sensitive latent factor matrix** to isolate sensitive features
- Use a regularizer to **extract sensitive information which taints other factors**.



### Findings

- Eliminate sensitive information & provides fair recommendation with respect to the sensitive attribute.
- Maintain recommendation quality

## • Some Tradeoffs when Comparing These Fairness Approaches

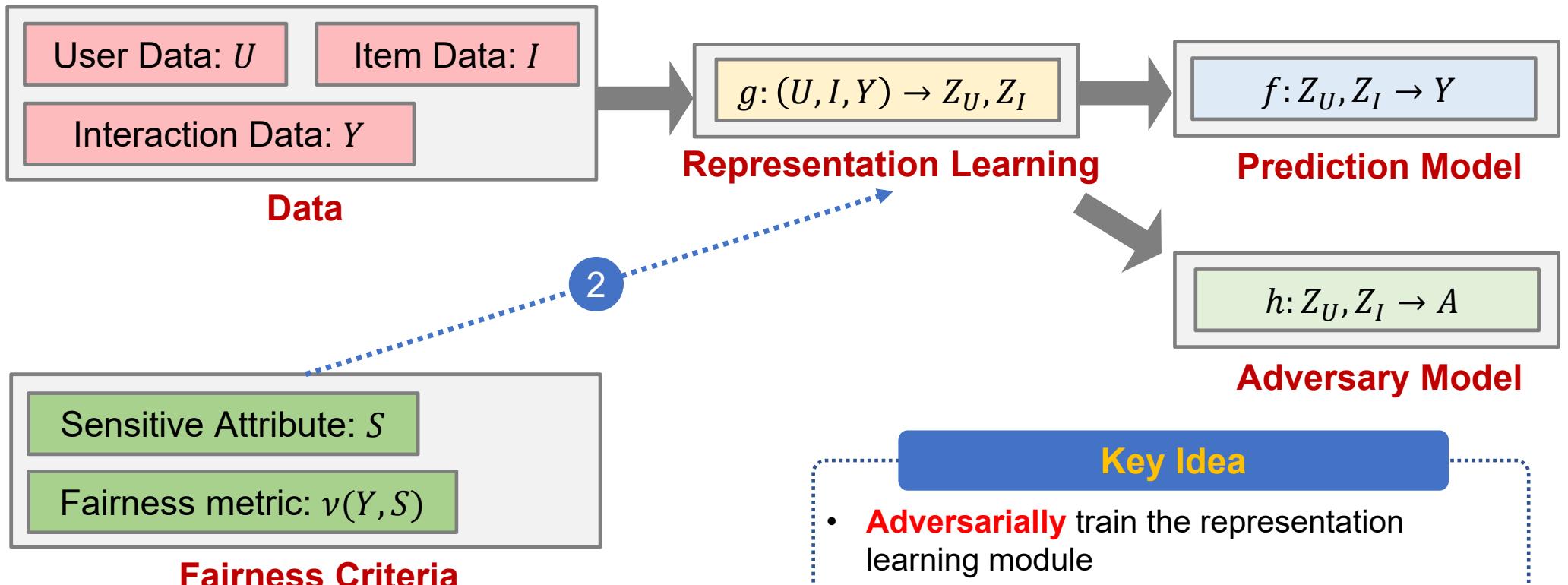
	Ease of implementation and (re-)use	Scalability	Ease of auditing	Fairness / Performance tradeoff	Generalization
Pre-processing, e.g., representation learning	✗	✗	✗		✗
In-processing, i.e., joint learning and fairness regulation			✗	✗	✗
Post-processing, e.g., threshold adjustment		✗	✗		



- Four Research Lines towards Fairness

- Rebalancing
- Regularization
- Fair Representation Learning

## • Line 3: Adversarial Learning → Fair Representation Learning



### Key Idea

- **Adversarially** train the representation learning module
- Based on the representations, the **prediction & adversary models** play a min-max game

# • Example 1: Adversarial Learned Fair Representation (ALFR)

**Reconstruction loss**  
between input data X and  
representations R

**Prediction error** in  
generating prediction Y  
from R

$$\max_{\phi} \min_{\theta} \mathcal{L} = \alpha C_{\theta}(X, R) + \beta D_{\theta, \phi}(R, A) + \gamma E_{\theta}(Y, R)$$

**Training an adversary model** to encourage the independence between the representation R and the sensitive attributes A, **rather than a regularization term**

Predicting sensitive attributes from the representations R

$$D = \mathbb{E}_{X, A} A \cdot \log(f(R)) + (1 - A) \cdot \log(1 - f(R))$$

Cross entropy for binary sensitive attribute

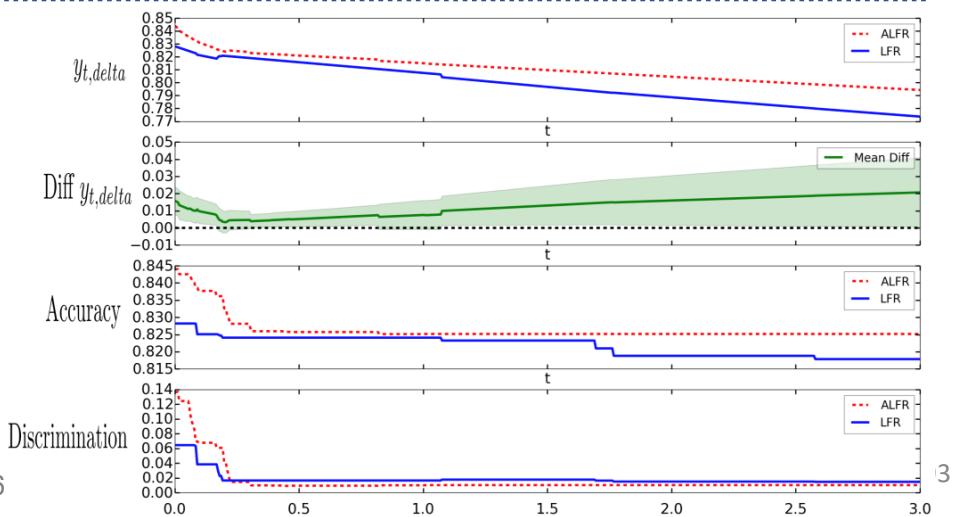
## Idea

Adversarial Representation Learning

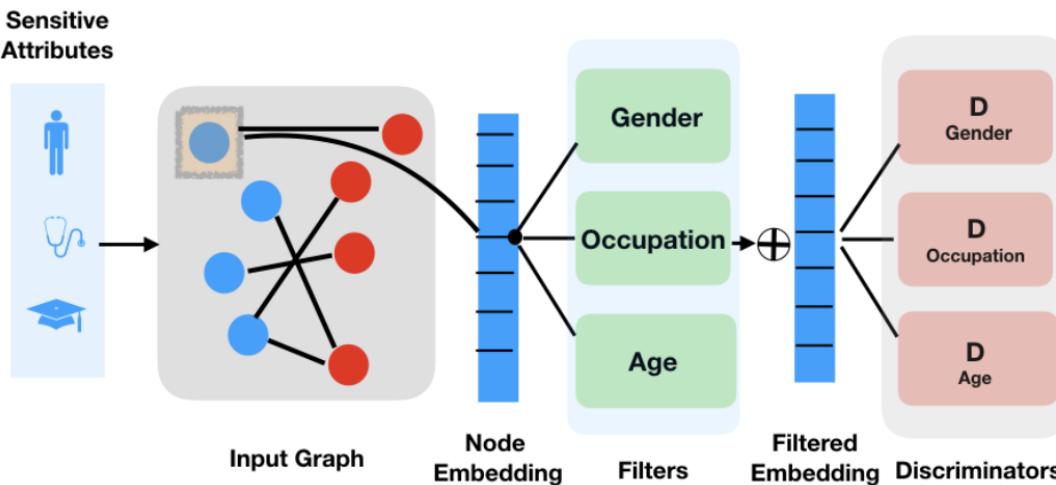
- **encode insensitive attributes** of data
- **remove any information about sensitive attributes**

## Findings

- Achieve better performance & fairness than LFR (regularization)



## • Example 2: Compositional Fairness Constraints for Graph Embeddings



### Idea

Based on ALFR

- Focusing on **graph structured data**
- Flexibly accommodate **different combinations of fairness constraints** → **compositional fairness**

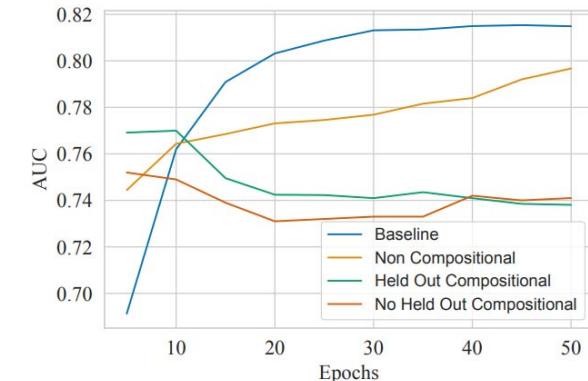


Figure 3. Performance on the edge prediction (i.e., recommendation) task on the Reddit data. Evaluation is using the AUC score, since there is only one edge/relationship type.

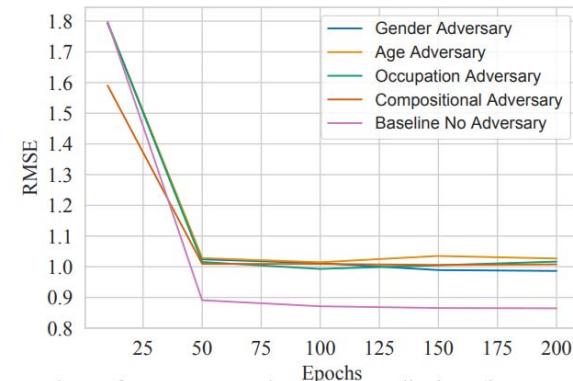
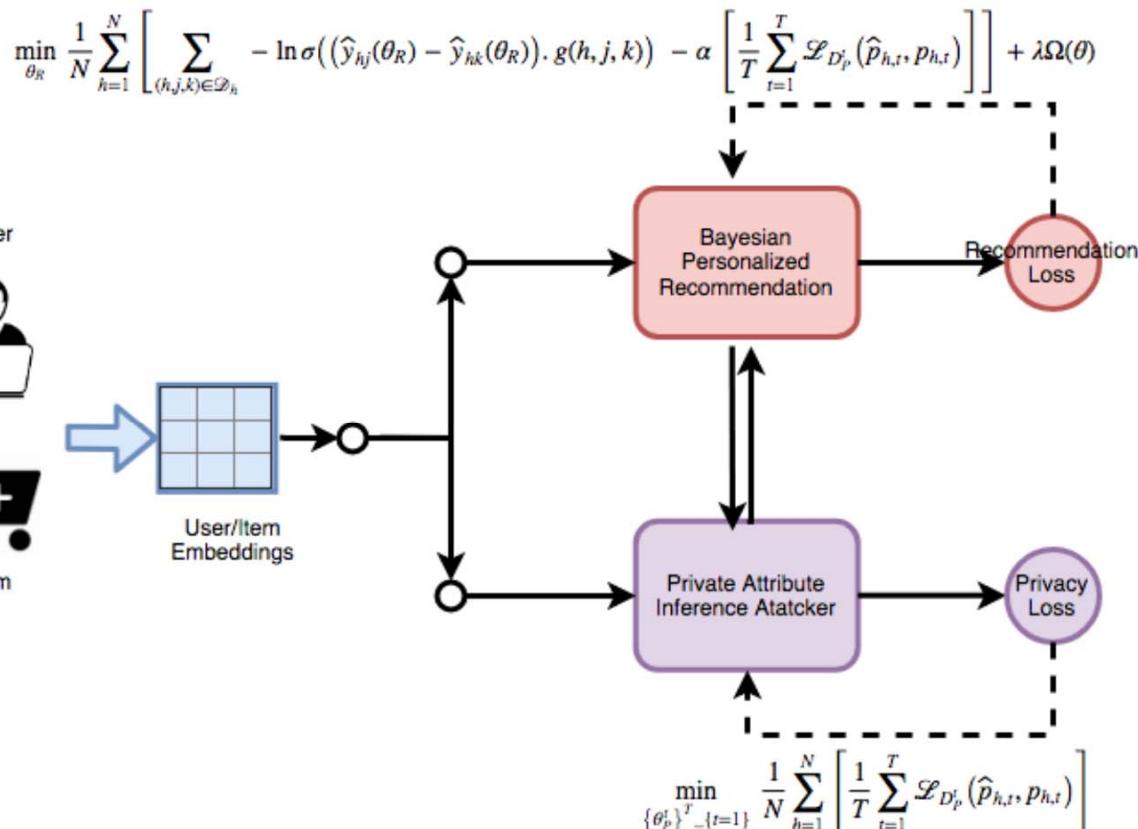


Figure 2. Performance on the edge prediction (i.e., recommendation) task on MovieLens, using RMSE as in Berg et al. (2017).

## • Example 3: Recommendation with Attribute Protection (RAP)



### Idea

Based on ALFR

- Focusing on **recommendation scenarios**
- **Prediction model** → BPR
- **Adversary model** → Private attribute inference attacker

Model	35				$P@K$	$R@K$
	Gen	Age	Occ			
<b>ORIGINAL</b>	0.7662	0.7050	0.8332	0.156	0.156	
<b>LDP-SH</b>	0.6587	0.6875	0.8076	0.071	0.071	
<b>BLURMe</b>	0.6266	0.6177	0.7614	0.118	0.118	
<b>RAP</b>	<b>0.6039</b>	<b>0.5397</b>	<b>0.7319</b>	<b>0.152</b>	<b>0.152</b>	



## • Summary

### Pros:

- Representation learning can centralize fairness constraints
- Representation learning can simplify and centralize the task of fairness auditing
- Learned representations can be constructed to satisfy multiple fairness measures simultaneously
- Learned representations can simplify the task of evaluating the fairness/performance tradeoff, e.g., using performance bounds

### Cons:

- Less precise control of fairness/performance tradeoff, than joint learning ...
- May lead to fairness overconfidence ...

## • Tutorial Outline

### ❑ Biases in Data

- ❑ Definition of data biases
- ❑ Categories: Selection bias, Conformity bias, Exposure bias and Position bias
- ❑ Recent solutions for data biases

### ❑ Biases in Results

- ❑ Popularity bias: definition, characteristic and solutions
- ❑ Unfairness: definition, characteristic and solutions

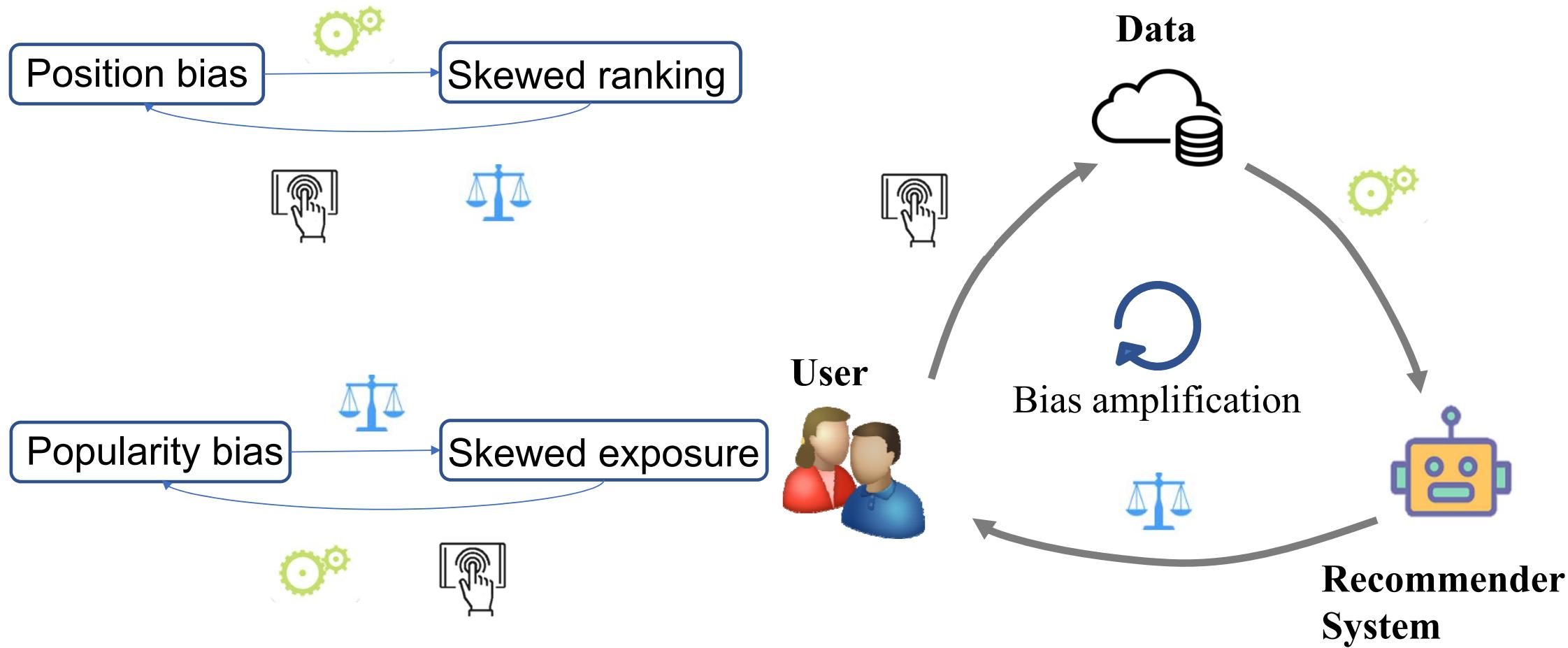
### ❑ Bias Amplification in Loop and Its Solutions

### ❑ Summary and Future Direction

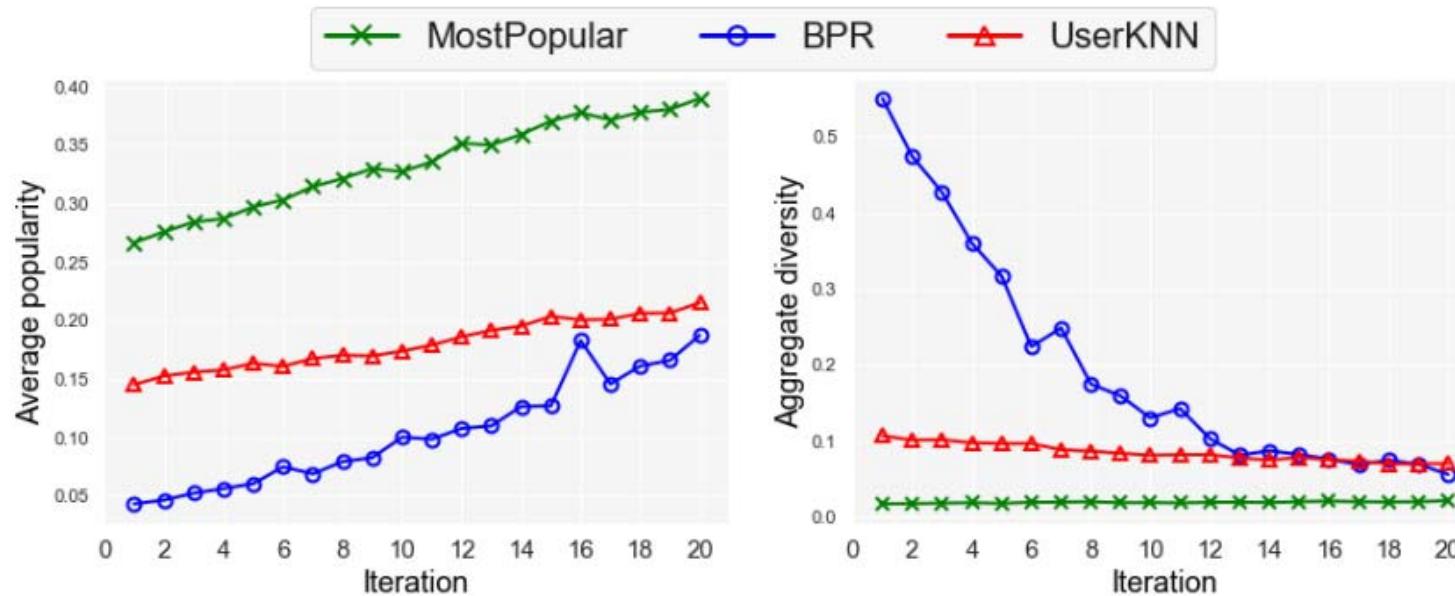
slides will be available at: <https://github.com/jiawei-chen/RecDebiasing>

A literature survey based on this tutorial is available at: <https://arxiv.org/pdf/2010.03240.pdf>

## • Feedback Loop Amplifies Biases



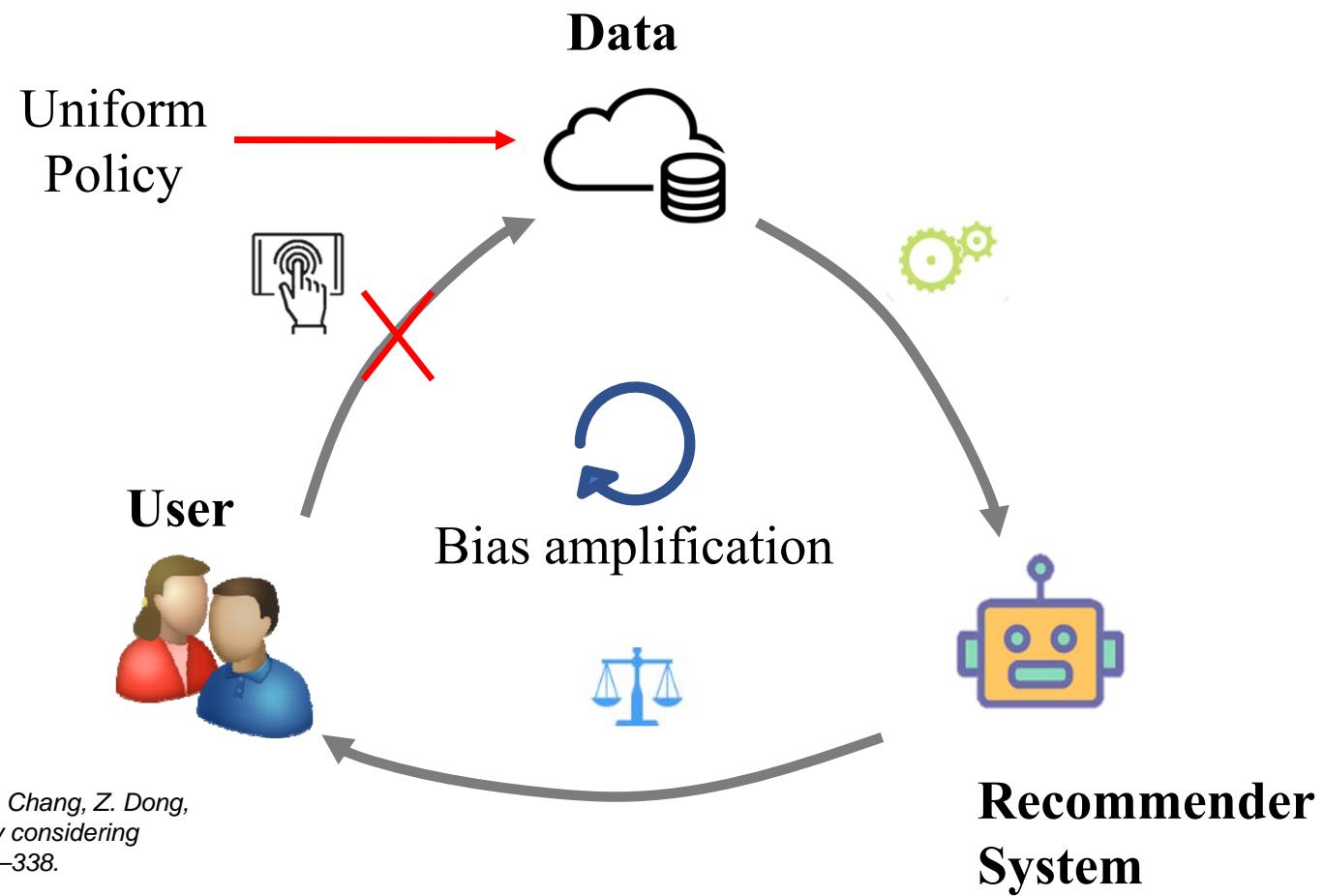
- Feedback Loop Amplifies Biases



The average popularity of the recommendation items are increasing while the diversity are decreasing along the feedback loop.

## • Solution for Bias Amplification

- Leveraging uniform data.



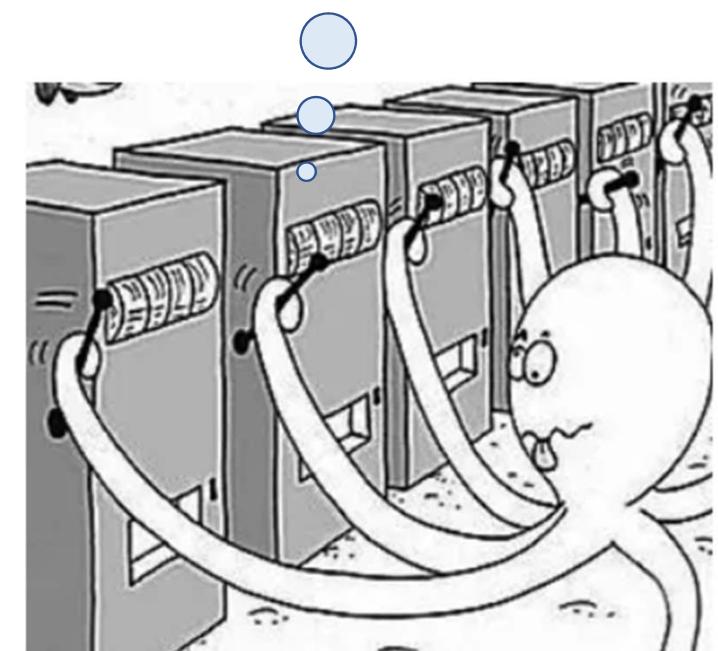
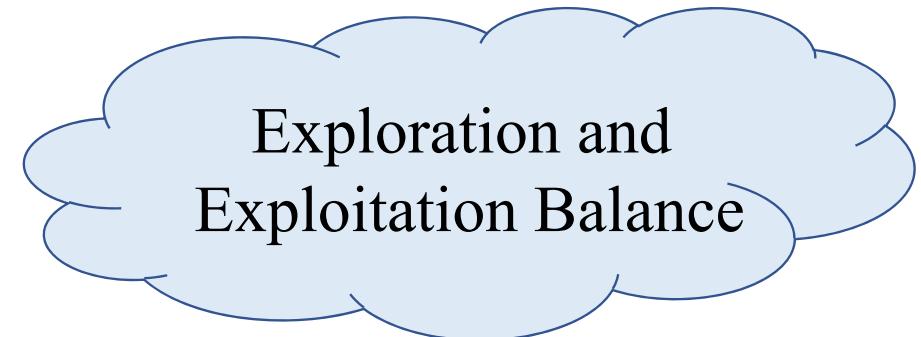
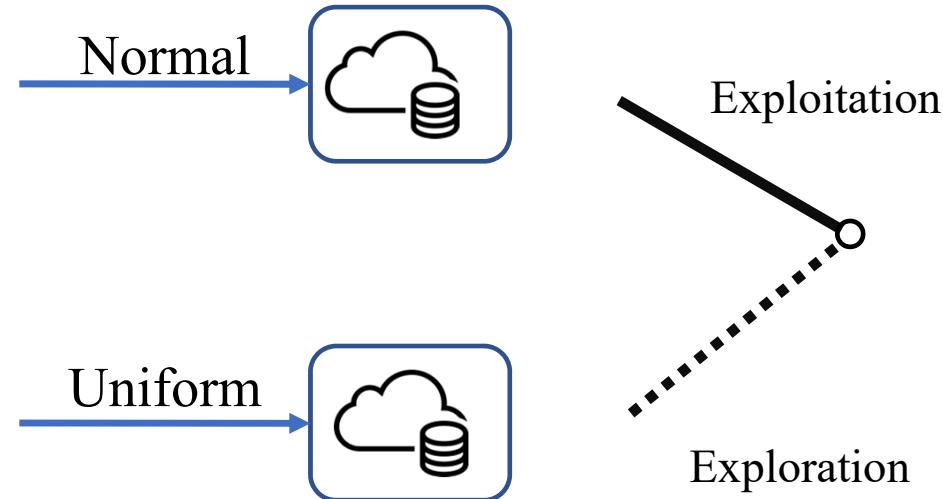
B. Yuan, J.-Y. Hsia, M.-Y. Yang, H. Zhu, C.-Y. Chang, Z. Dong, and C.-J. Lin, “Improving ad click prediction by considering nondisplayed events,” in CIKM, 2019, pp. 329–338.

S. Bonner and F. Vasile, “Causal embeddings for recommendation,” in RecSys, 2018, pp. 104–112.

## • Solution for Bias Amplification

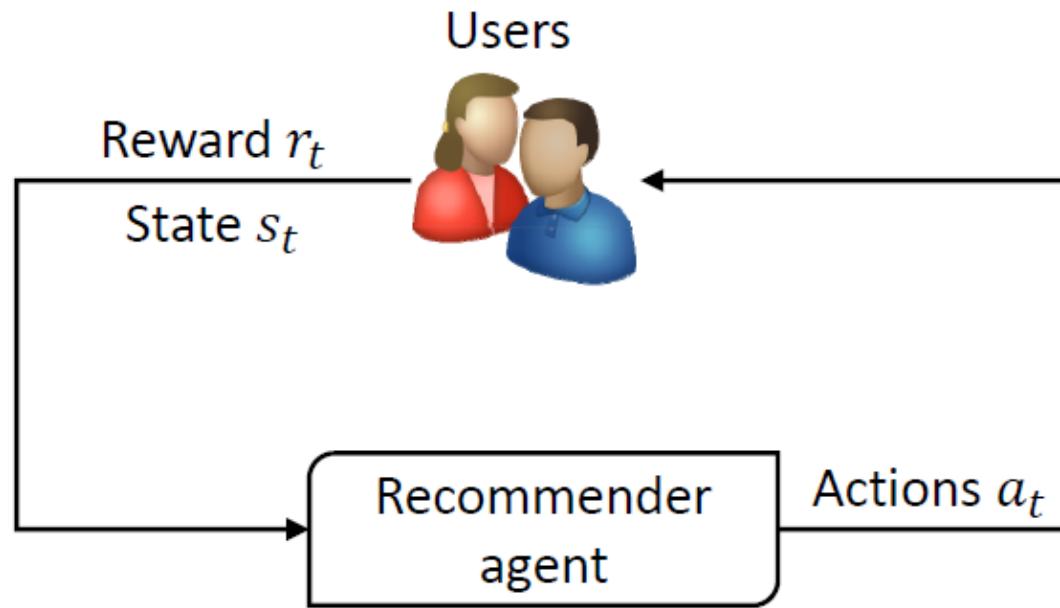
- Interactive recommendation.

a recommender system can interact with a user and dynamically capture his preference



*Bandit*

## • Solution for Bias Amplification



RL agent → Recommender system  
Reward → User feedback  
Environment → User  
Policy → Which items to be recommended

## • Tutorial Outline

### ❑ Biases in Data

- ❑ Definition of data biases
- ❑ Categories: Selection bias, Conformity bias, Exposure bias and Position bias
- ❑ Recent solutions for data biases

### ❑ Biases in Results

- ❑ Popularity bias: definition, characteristic and solutions
- ❑ Unfairness: definition, characteristic and solutions

### ❑ Bias Amplification in Loop and Its Solutions

### ❑ Summary and Future Direction

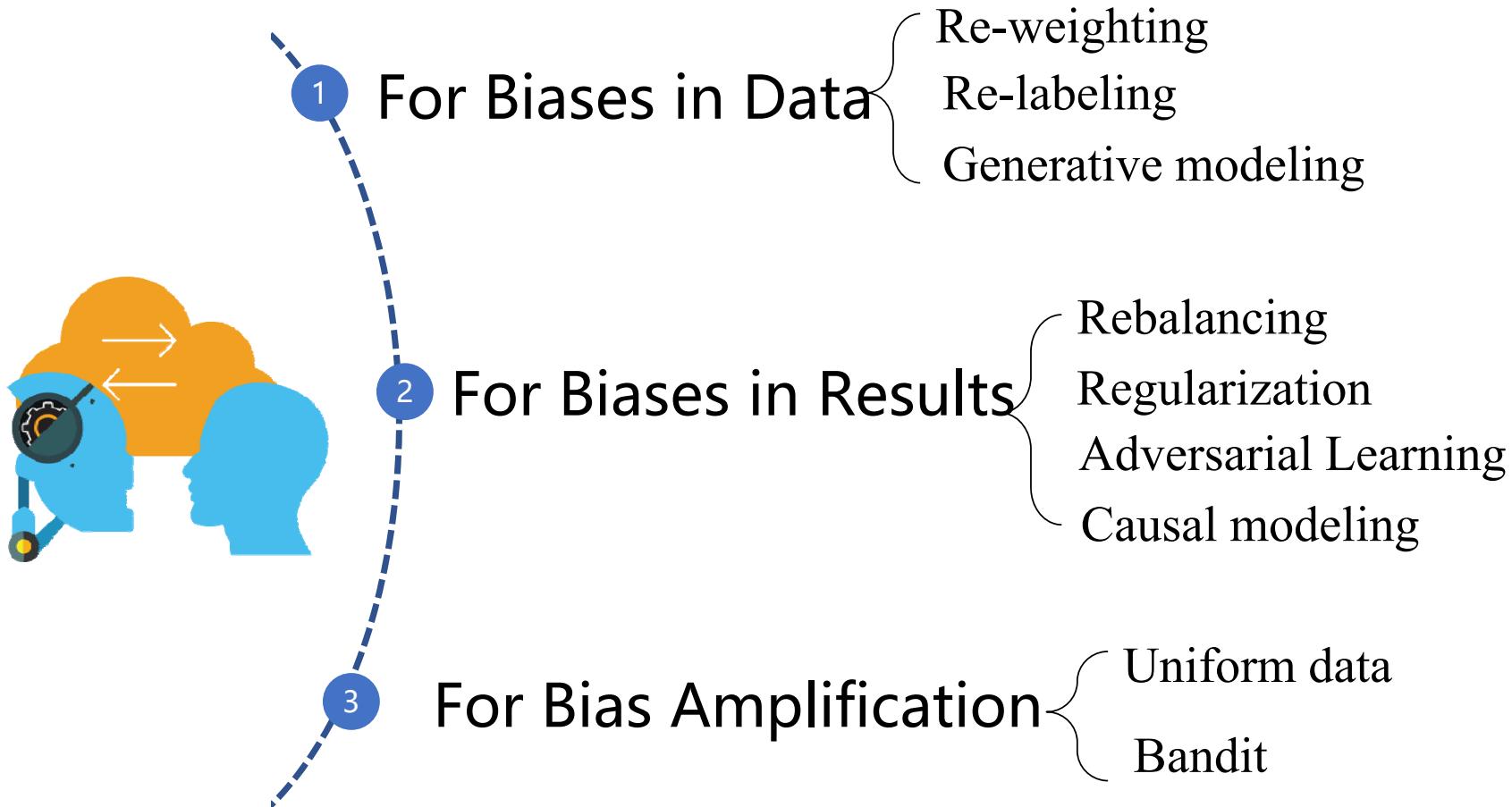
slides will be available at: <https://github.com/jiawei-chen/RecDebiasing>

A literature survey based on this tutorial is available at: <https://arxiv.org/pdf/2010.03240.pdf>

## • Summary

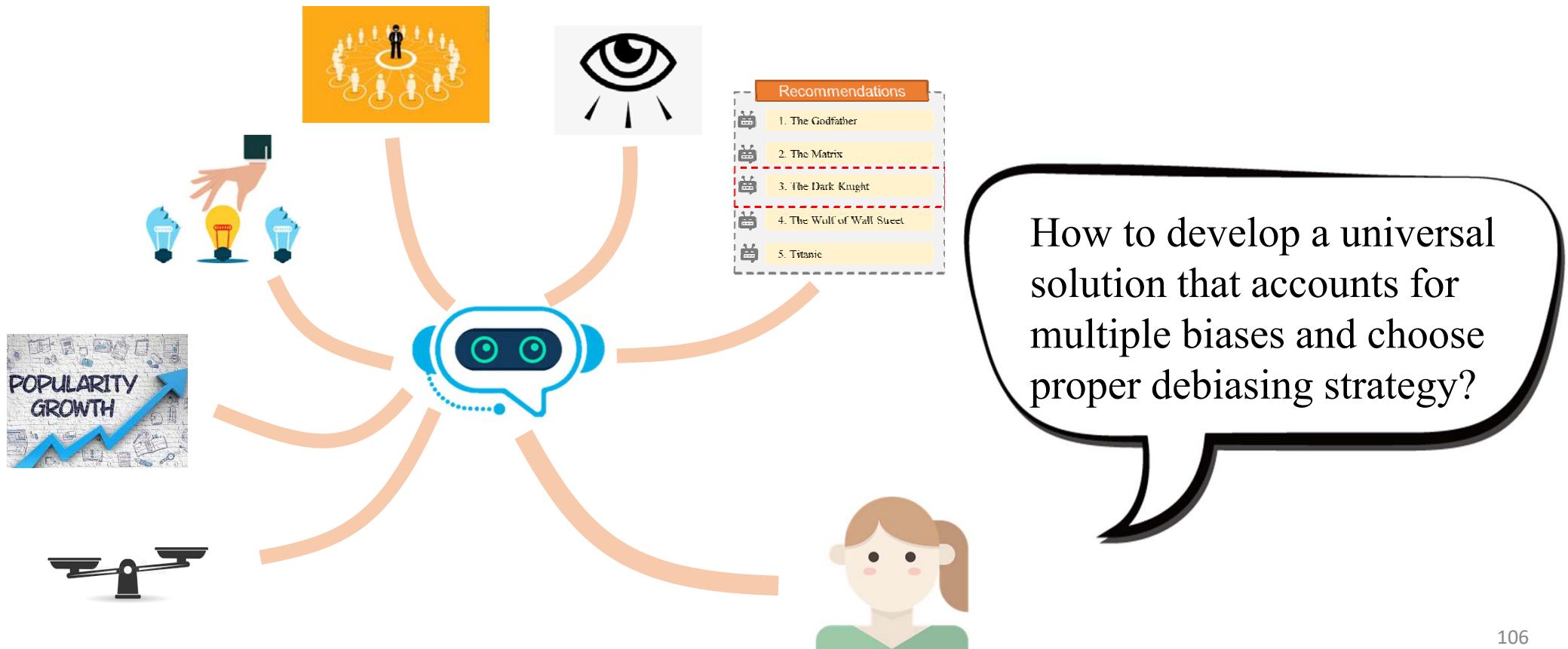
Types	Stage in Loop	Data	Cause	Effect
Selection Bias	User→Data	Explicit feedback	Users' self -selection	Skewed observed rating distribution
Exposure Bias	User→Data	Implicit feedback	Users' self-selection; Background; Intervened by systems; Popularity	Unreliable non-positive data
Conformity Bias	User→Data	Both	Conformity	Skewed rating values
Position Bias	User→Data	Both	Trust top of lists; Exposed to top of lists	Unreliable positive data
Popularity Bias	Model→User	Both	Algorithm and unbalanced data	Matthew effect
Unfairness	Model→User	Both	Algorithm and unbalanced data	Unfairness for some groups
Bias amplification in Loop	All	Both	Feedback loop	Enhance and spread bias <sup>4</sup>

## • Summary



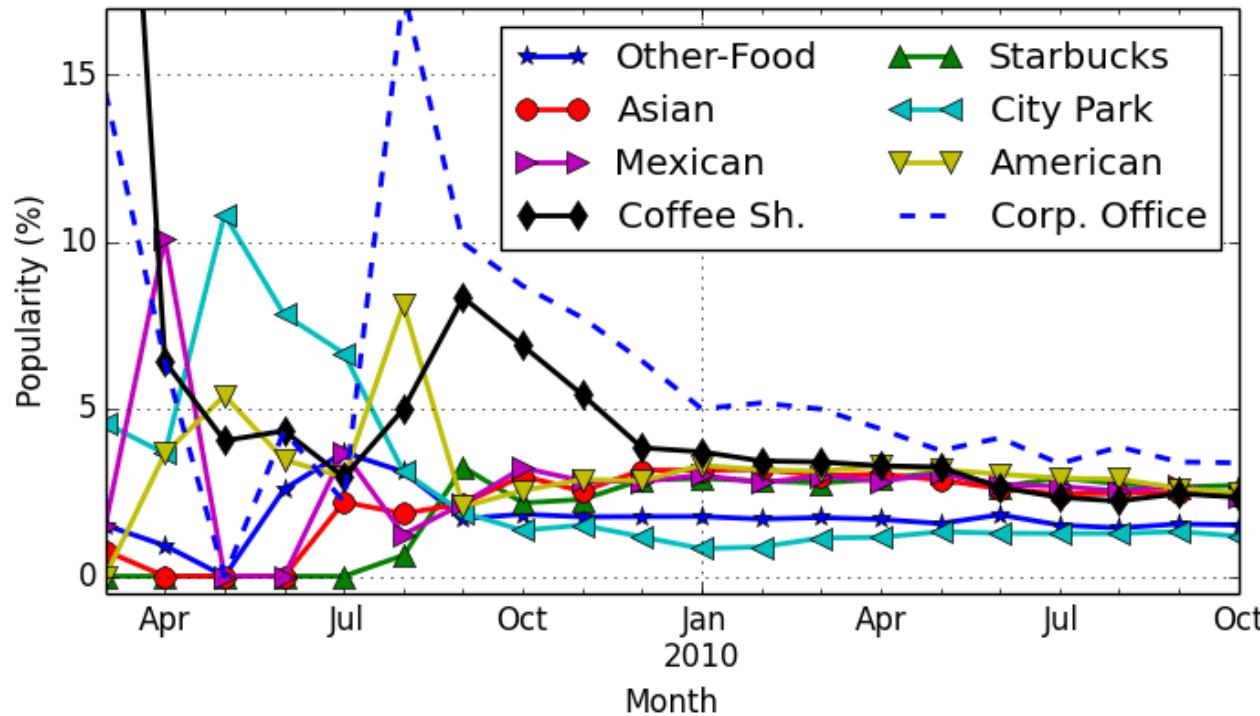
## • Future Direction

- A universal and adaptive debiasing framework.



## • Future Direction

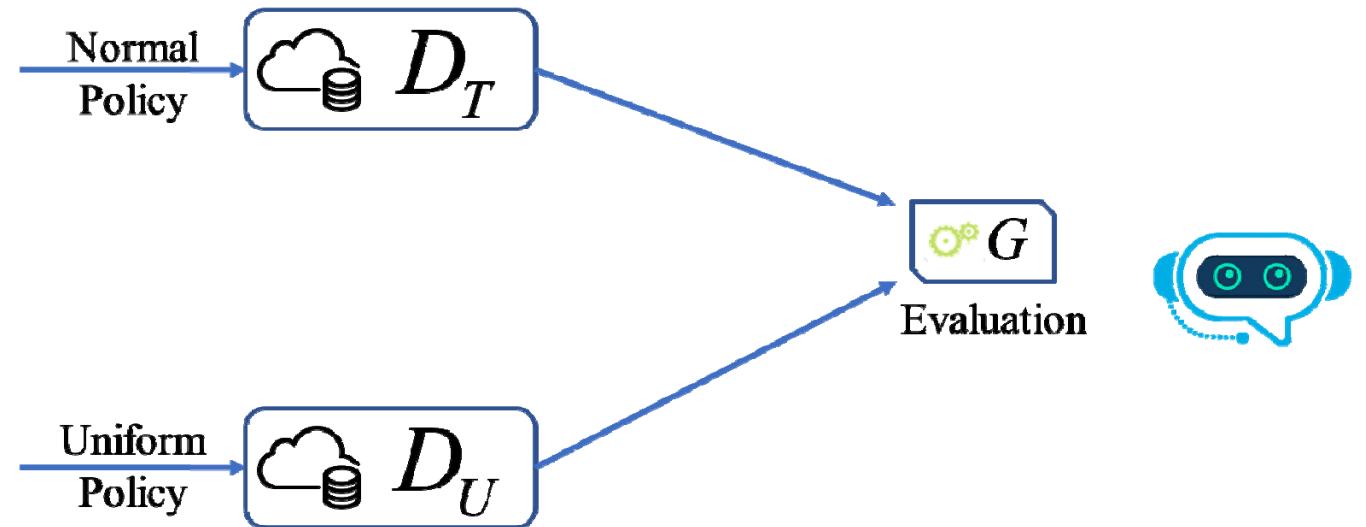
- Dynamic bias.



- Biases are usually dynamic rather than static.
- Online updating of debiasing strategies.

## • Future Direction

- Better evaluation.



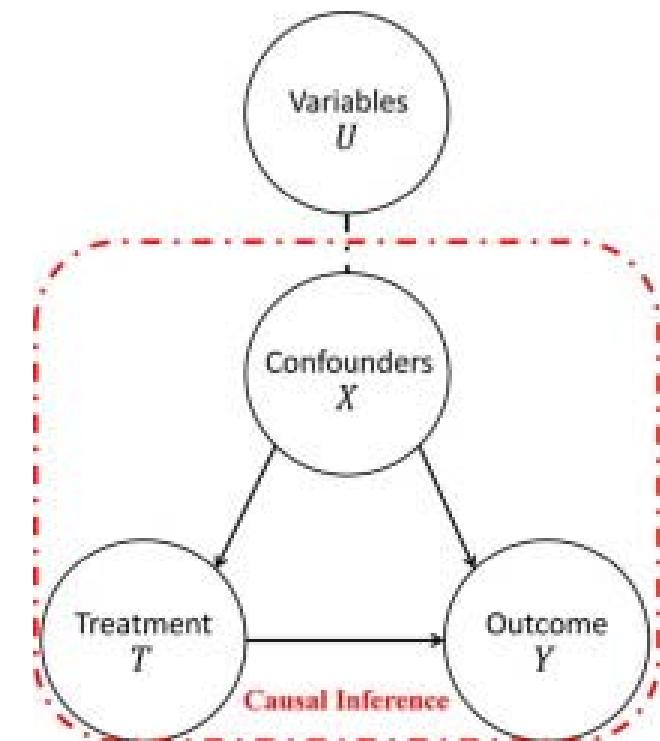
- Benchmark datasets and evaluation metrics.

## • Future Direction

- Understanding biases from causality.

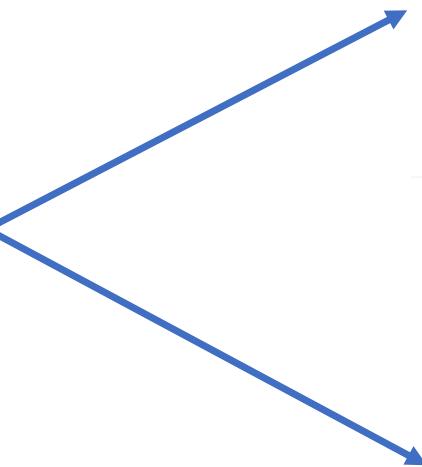
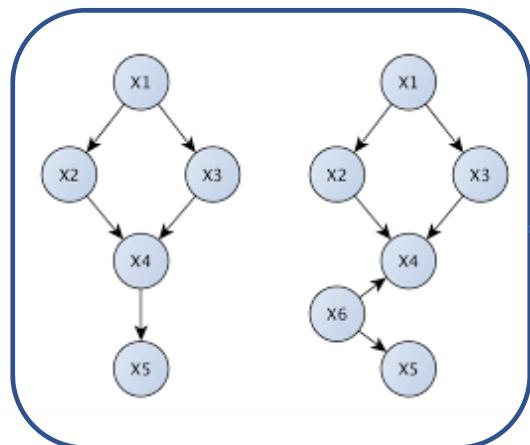


- Various biases can be understood as the confounders in the causal graph.
- Integrating user prior assumptions.

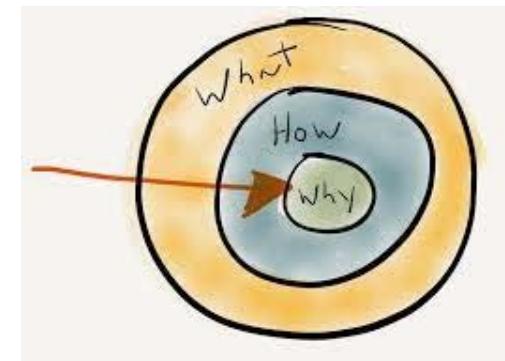
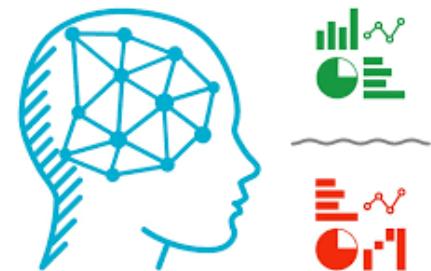


## • Future Direction

- Explanation and reasoning with causal graph.



BIAS?





# Bias and Debias in Recommender System: A Survey and Future Directions

Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, Xiangnan He

**Abstract**—While recent years have witnessed a rapid growth of research papers on recommender system (RS), most of the papers focus on inventing machine learning models to better fit user behavior data. However, user behavior data is observational rather than experimental. This makes various biases widely exist in the data, including but not limited to selection bias, position bias, exposure bias, and popularity bias. Blindly fitting the data without considering the inherent biases will result in many serious issues, e.g., the discrepancy between offline evaluation and online metrics, hurting user satisfaction and trust on the recommendation service, etc. To transform the large volume of research models into practical improvements, it is highly urgent to explore the impacts of the biases and perform debiasing when necessary. When reviewing the papers that consider biases in RS, we find that, to our surprise, the studies are rather fragmented and lack a systematic organization. The terminology “bias” is widely used in the literature, but its definition is usually vague and even inconsistent across papers. This motivates us to provide a systematic survey of existing work on RS biases. In this paper, we first summarize seven types of biases in recommendation, along with their definitions and characteristics. We then provide a taxonomy to position and organize the existing work on recommendation debiasing. Finally, we identify some open challenges and envision some future directions, with the hope of inspiring more research work on this important yet less investigated topic.

**Index Terms**—Recommendation, Recommender System, Collaborative Filtering, Survey, Bias, Debias, Fairness

---

<https://arxiv.org/pdf/2010.03240.pdf>

## paper/code link

Papers	Taxonomy 1	Taxonomy 2	Taxonomy 3	Date	Conference	Code
Collaborative filtering and the missing at random assumption	Bias in data	Bias in explicit feedback data	Selection Bias	2007	UAI	Python
Probabilistic matrix factorization with non-random missing data	Bias in data	Bias in explicit feedback data	Selection Bias	2014	PMLR	Python
Evaluation of recommendations: rating-prediction and ranking	Bias in data	Bias in explicit feedback data	Selection Bias	2013	RecSys	
Why amazon's ratings might mislead you: The story of herding effects	Bias in data	Bias in explicit feedback data	Conformity Bias	2014	Big data Volume: 2 Issue 4: December 15, 2014	
Are you influenced by others when rating?: Improve rating prediction by conformity modeling	Bias in data	Bias in explicit feedback data	Conformity Bias	2016	RecSys	
A methodology for learning, analyzing, and mitigating social influence bias in recommender systems	Bias in data	Bias in explicit feedback data	Conformity Bias	2014	RecSys	Python

<https://github.com/jiawei-chen/RecDebiasing>



**THANK YOU!**