NDBI021, Lecture 8

User preferences, 2/1 ZK+Z,

Wed 12:20 – 13:50 S8

Wed 14:00 – 15:30 SW2 (odd weeks)

https://www.ksi.mff.cuni.cz/~peska/vyuka/ndbi021/2022/

matfyz

siret research group

https://ksi.mff.cuni.cz

Biases in RS

Fairness in evaluation

- Popularity bias (more popular => much more attention)
- Biased historical data (missing not at random) => (unbiased) learning algorithm => biased recommendations
- => biased off-line evaluation (same bias vector => better results)
- => discrepancy between off-line and on-line evaluation
- How to evaluate methods fairly?

Fairness in evaluation

- Inverse propensity score
- Weight results by the inverse to the propensity score
  - (probability of being noticed by the user)
  - Definitions may vary on available information
    - Based on general item's popularity
    - Based on recommended positions
    - Based on user's actions within the page

De-biasing Off-line Evaluation

- https://dl.acm.org/doi/pdf/10.1145/3240323.3240355

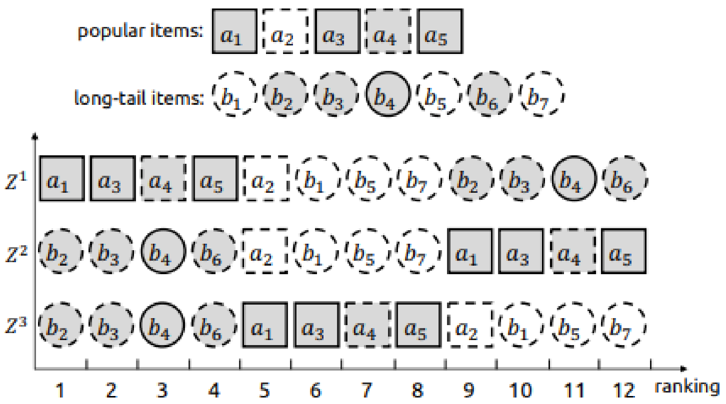


Figure 1: A hypothetical example to illustrate the evaluation bias that results from use of the AOA evaluator. Three recommenders generated distinct lists of recommendations, Z^1, Z^2 and Z^3, for the same user. Among the shaded items that were preferred by the user, the ones with a solid border were observed by recommenders. The performance was measured by DCG, and the results are presented in Table 1.

Table 1: The true and estimated DCG values for three recommenders in Fig. 1. $R(\hat{Z})$ denotes the ground truth, and $\hat{R}_{AOA}(\hat{Z})$ denotes the AOA estimations. The AOA estimator outputs larger values when popular items are ranked higher.

Estimator	Z^1	Z^2	Z^3
$R(\hat{Z})$	0.463	0.463	0.494
$\hat{R}_{AOA}(\hat{Z})$	0.585	0.340	0.390

3.1 Average-over-all (AOA) evaluator

In prior literature, $R(\hat{Z})$ was estimated by taking the average over all observed user feedback S_u^*:

$$\hat{R}_{AOA}(\hat{Z}) = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\frac{1}{|S_u^*|}\sum_{i\in S_u^*} c(\hat{Z}_{u,i}) = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\frac{1}{\sum_{i\in S_u} O_{u,i}}\sum_{i\in S_u} c(\hat{Z}_{u,i})\cdot O_{u,i} \quad (6)$$

nDCG, AUC, MAP,...

3.2 Unbiased evaluator

To conduct unbiased evaluation of biased observations, we leverage the IPS framework [16, 22] that weights each observation with the inverse of its propensity, where the term *propensity* refers to the tendency or the likelihood of an event happening. The intuition is to down-weight the commonly observed interactions, while up-weighting the rare ones. In the context of this paper, the probability $P_{u,i}$ is treated as the pointwise propensity score. Therefore, the IPS unbiased evaluator is defined as follows:

$$\hat{R}_{IPS}(\hat{Z}|P) = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\frac{1}{|S_u|}\sum_{i\in S_u^*}\frac{c(\hat{Z}_{u,i})}{P_{u,i}} = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\frac{1}{|S_u|}\sum_{i\in S_u}\frac{c(\hat{Z}_{u,i})}{P_{u,i}}\cdot O_{u,i} \quad (7)$$

Propensity score

$$\hat{P}_{*,i} \propto (n_i^*)^{\gamma}\cdot n_i, \quad (13$$

where $n_i = \sum_{u\in\mathcal{U}} \mathbf{1}[i\in S_u]$ and $n_i^* = \sum_{u\in\mathcal{U}, i\in S_u^*} O_{*,i}$.

However, empirically, n_i is not directly observable. To addres this problem, we observe that n_i^* is sampled from a binomial distri bution[4] parameterized by n_i, that is, $n_i^* \sim \mathcal{B}(n_i, P_{*,i})$. Therefore, a relationship between n_i and n_i^* can be built by bridging the genera tive model (eqn. 13) with the following unbiased estimator:

$$\hat{P}_{*,i} = \frac{n_i^*}{n_i} \propto (n_i^*)^{\gamma}\cdot n_i \quad (14$$

Therefore, $n_i \propto (n_i^*)^{\frac{1-\gamma}{2}}$. We use this as a replacement for th unobserved n_i in eqn. 13, which results in an unbiased $\hat{P}_{*,i}$ estimato that is determined by only the empirical counts of items:

$$\hat{P}_{*,i} \propto (n_i^*)^{\left(\frac{\gamma+1}{2}\right)} \quad (15$$

De-biasing Off-line Evaluation

- https://link.springer.com/article/10.1007/s10844-021-00651-y
- Alternative: sampling from test data to de-bias them
  - Based on missing-at-random (MAR) vs. Missing-not-at-random (MNAR)
  - Sample from MNAR data to better resemble MAR
  - **Variants:**
    - You have some subsample that is MAR (random recommendations, forced rating), sample from MNAR so that posterior probability is similar to MAR. Finding weight w for each user-item pair

$$P_{mnar}(u,i|\mathcal{O},w) = P_{mar}(u,i|\mathcal{O}) \quad \forall (u,i) \in D^{mnar}$$

$$P_{mnar}(u|\mathcal{O},w) = P_{mar}(u|\mathcal{O}) \quad \forall u \in U$$

$$P_{mnar}(i|\mathcal{O},w) = P_{mar}(i|\mathcal{O}) \quad \forall i \in I$$

$$w_u = \frac{P_{mar}(u|\mathcal{O})}{P_{mnar}(u|\mathcal{O})} \quad \forall u \in U$$

$$w_i = \frac{P_{mar}(i|\mathcal{O})}{P_{mnar}(i|\mathcal{O})} \quad \forall i \in I$$

    - You do not have MAR subsample: assume uniform posterior probability
  - Possible disadvantage: not enough data due to sampling
    - Sample with repetition
  - Possible disadvantage: not enough data from all segments

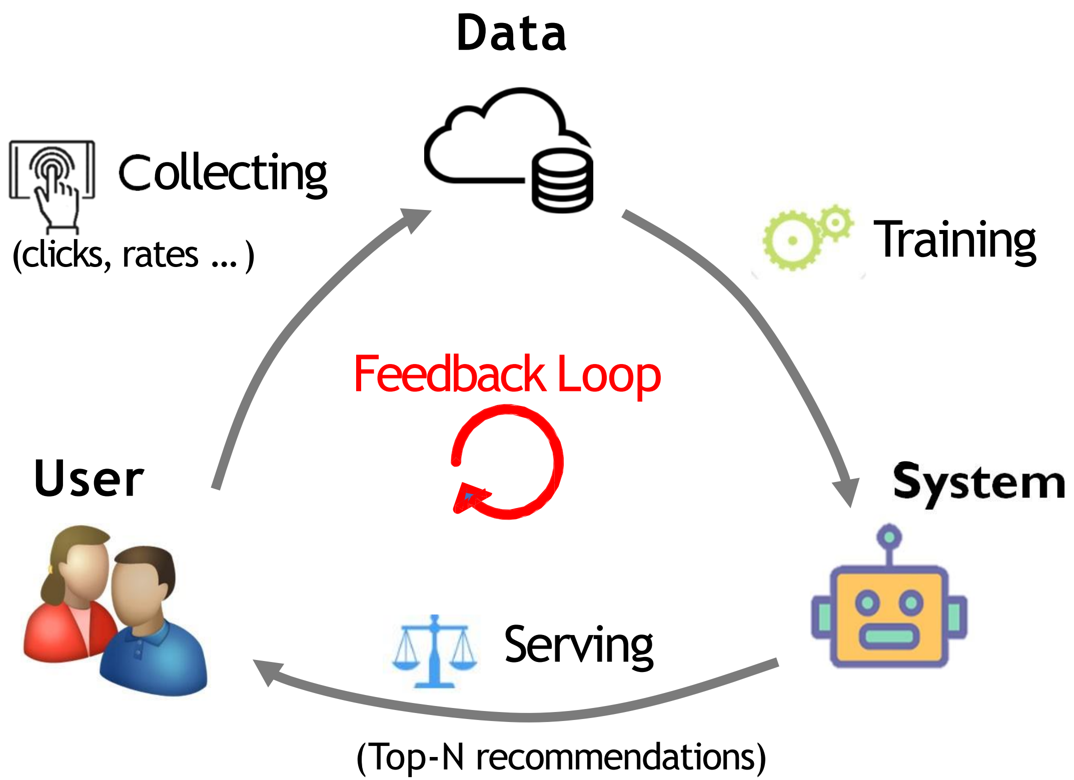
Bias Issues and Solutions in Recommender System

Jiawei Chen, Xiang Wang, Fuli Feng, Xiangnan He
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
.

• Ecosystem of Recsys

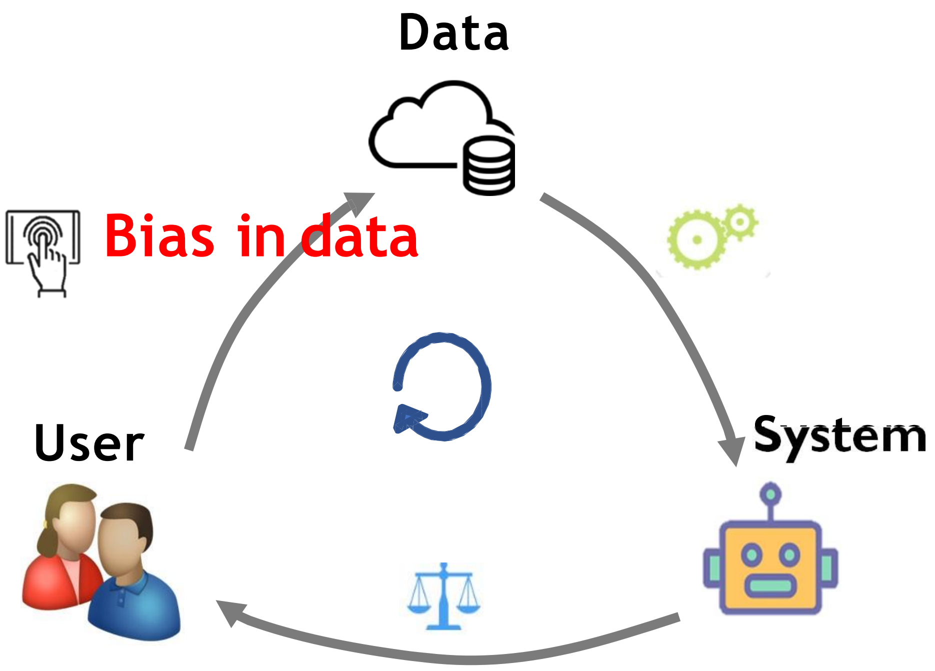
- Workflow of RS
  - **Training**: RS is trained/updated on observed user-item interaction data.
  - **Serving**: RS infers user preference over items and exposes top-n items.
  - **Collecting**: User actions on exposed items are merged into the training data.
- Forming a Feedback Loop



• Where Bias Comes?

- Bias in data (Collecting):
  - Data is observational rather than experimental (i.e., missing-not-at-random)
  - Affected by many factors:
    - The exposure mechanism
    - Public opinions
    - Display position

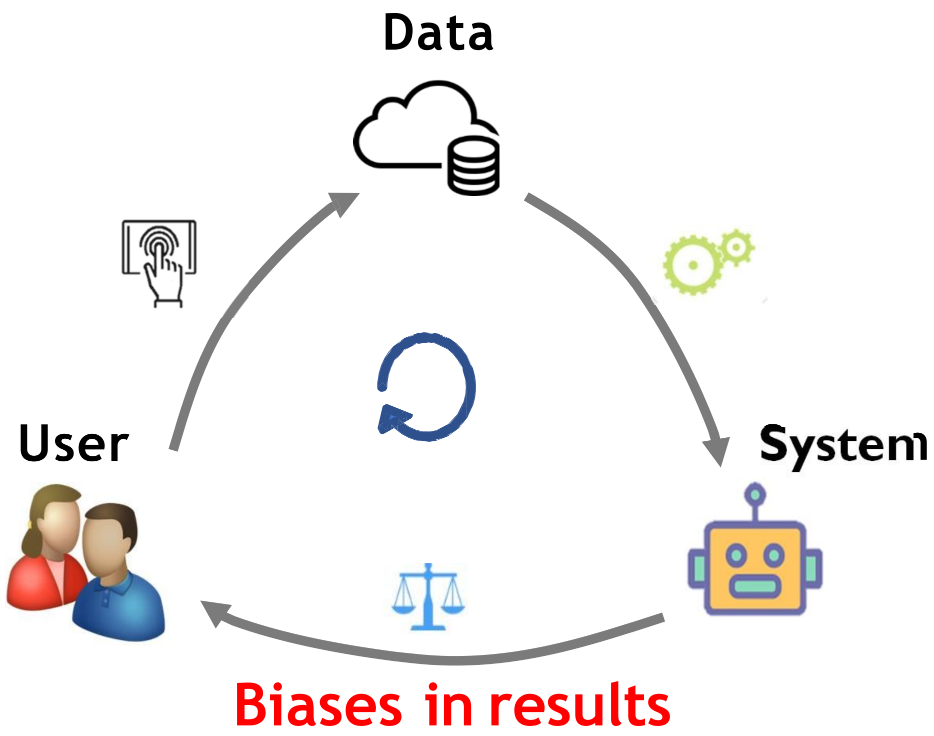
    … …
  - The collected data deviates from user true preference.



- Where Bias Comes?

- Bias in results (Serving):
  - Unbalanced training data
  - Recommendations are in favor of some item groups
  - E.g., popularity bias, category-aware unfairness
  - Hurting user experience and satisfaction

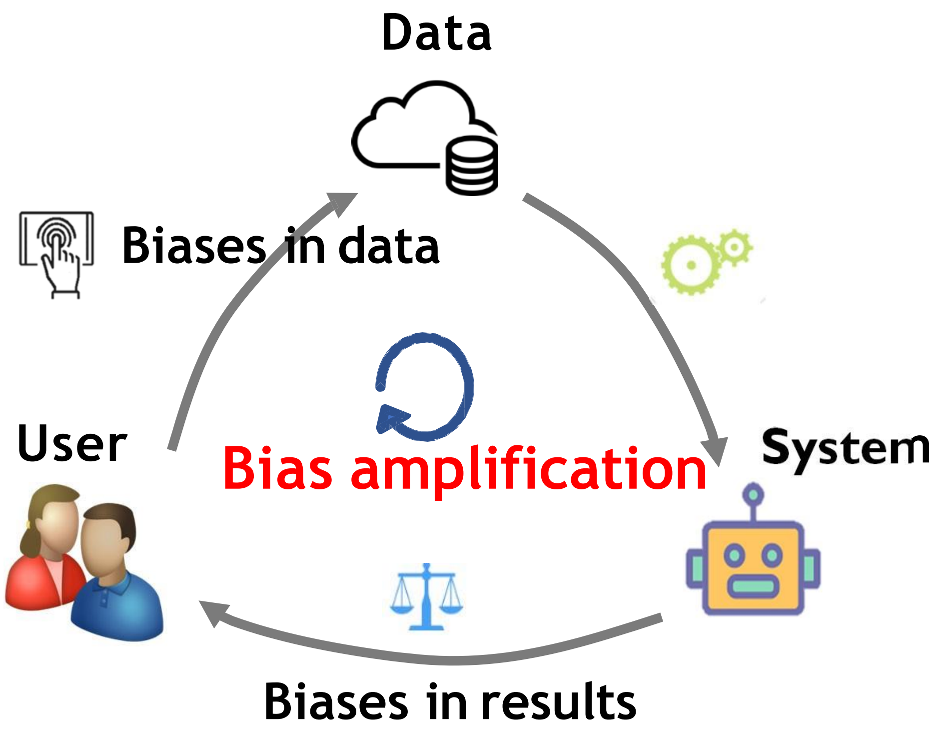
Fairness intervention strategies from previous lectures



• Matthew Effect: Bias + Loop

- Biases amplification along the loop:
  - Biases would be circled back into the collected data
  - Resulting in "Matthew effect" issue: the rich gets richer
  - Damaging the ecosystem of RS

Managable through exploration promotion



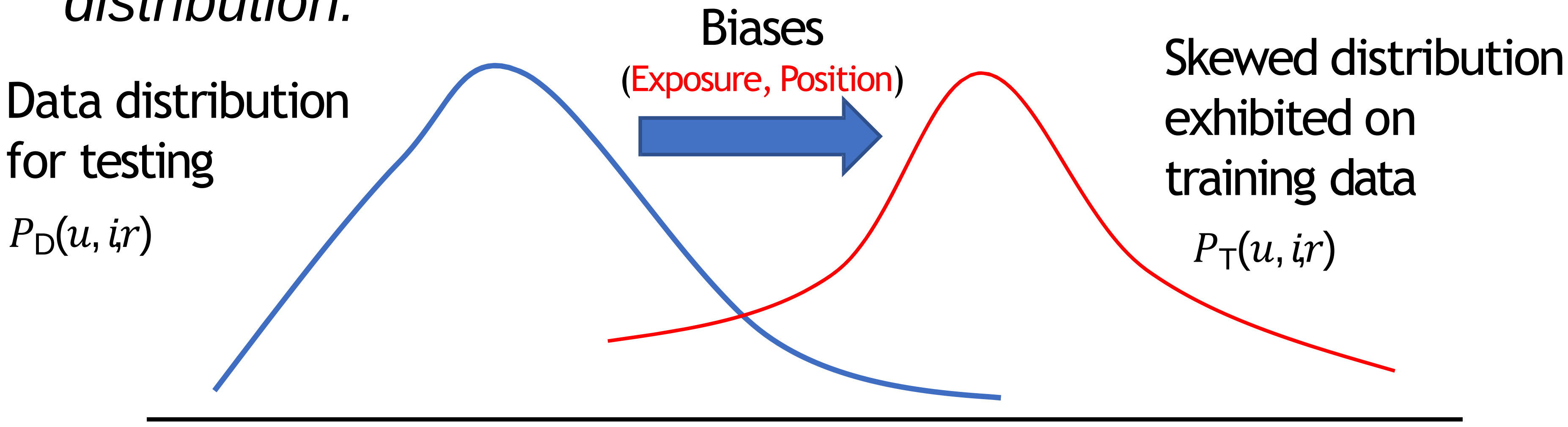
• Bias is Evil

- 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

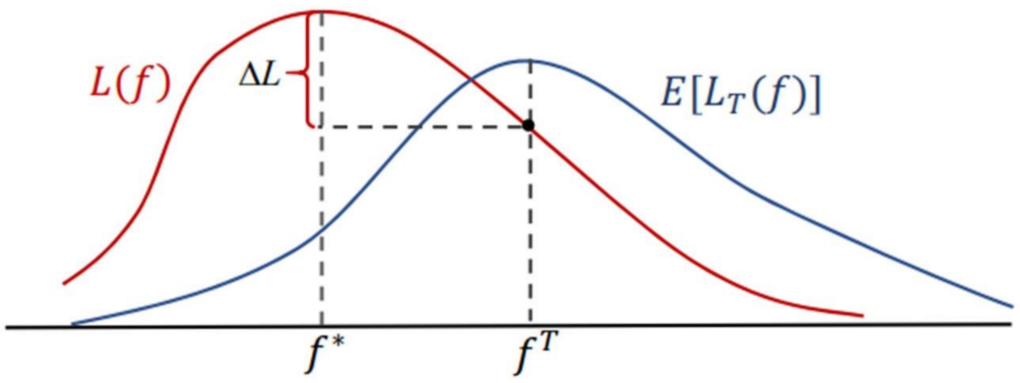
- **What is data bias?**

Data bias: *The distribution for which the training data is collected is different from the ideal data distribution.*



• Impact of Data Bias

- Data bias causes model training towards wrong direction.



Distributional difference between p_T and P_D . → Risk discrepancy between $\hat{L}_T(f)$ and $L(f)$. → Suboptimal results.

$p_T \neq p_D$ $E_{P_T}[\hat{L}_T(f)] \neq L(f)$ $f^* \neq f^T$

- True risk. $L(f) = E_{P_D(u,i)P_D(R_{ui}|u,i)}[\delta(f(u,i), R_{ui})]$
- Empirical risk. $\hat{L_T}(f) = \frac{1}{|D_T|} \sum_{(u,i,r_{ui}) \in D_T} [\delta(f(u,i), r_{ui})]$

- **Selection Bias**

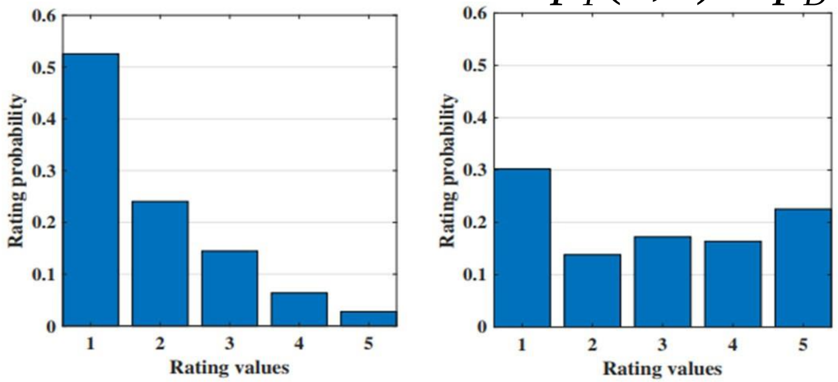
  - Definition: *Selection bias happens in explicit feedback data as users are free to choose which items to rate, so that the observed ratings are not a representative sample of all ratings.*

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

Selection bias →

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

3	4		5
	3		5
	3	4	4



(a) Random (b) User-selected

1Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as Treatments: Debiasing Learning and Evaluation. In ICML.
2B. 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 $p_D(u,i)$.

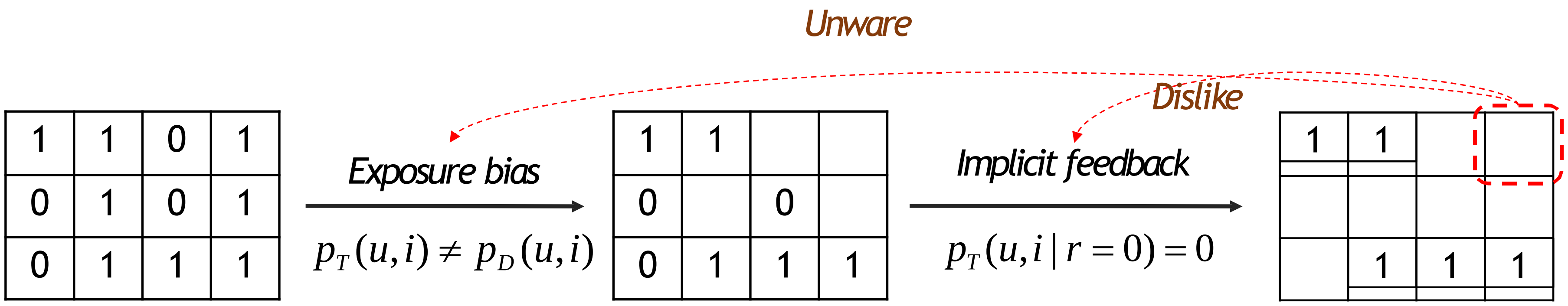
1	1	0	1
0	1	0	1
0	1	1	1

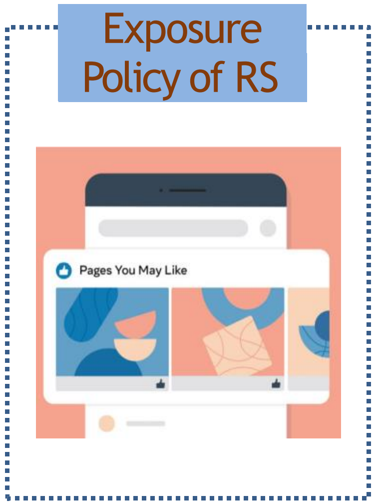
Exposure bias

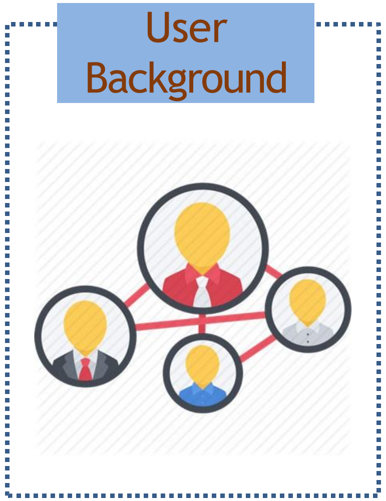
$p_T(u,i) \neq p_D(u,i)$

1	1		
0		0	
0	1	1	1

• 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

Conformity bias →

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

3	4		5
	3		5
	3	3	4

3	4	3	5

Public opinions

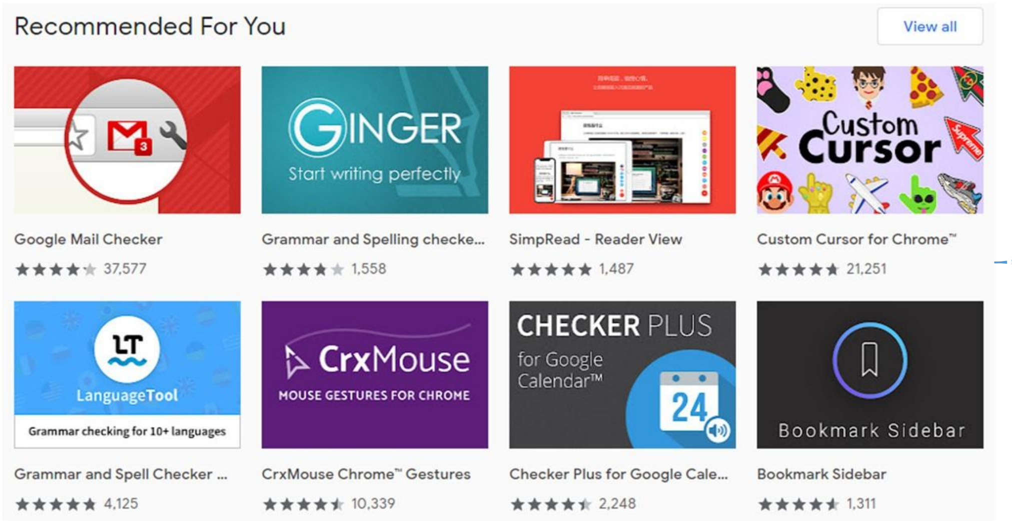
Position Bias

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

1	1	0	1
0	1	0	1
0	1	1	1

Position bias →

1	1	1	
		1	
		1	1



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

User exposure will be affected by the position

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

User judgments also will be affected by the position

• Debiasing Strategies Overview

IPS – can be applied for training as well as evaluation

- 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

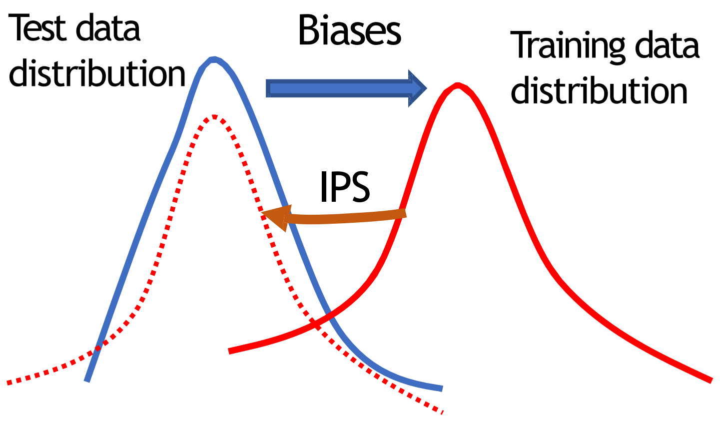
• Re-weighting Strategies

- Basic idea: change data distribution by sample reweighting:

$$L_{ipc} = \sum_{(u,i) \in D_T} \frac{1}{\rho_{ui}} \delta(r_{ui}, r_{ui})$$

- Mainly addressing the deviation of $p(u, i)$

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



- Properly defining weights can lead to *unbiased estimator* of the ideal:

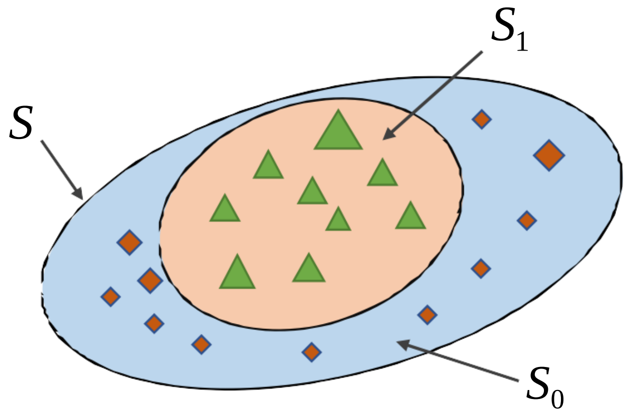
$$L(f) = E_{P_D(u,i) P_D(r|u,i)}[\delta(r_{ui}, \hat{r}_{ui})] \neq E[\hat{L}_T(f)] = E_{P_T(u,i) P_T(r|u,i)}[\delta(r_{ui}, \hat{r}_{ui})]$$

$$\frac{P_D(u,i)}{P_T(u,i)} = \frac{1}{\rho_{ui}}$$ Inverse propensity Scores (IPS)

$$= E[\hat{L}_{IPS}(f)] = E_{P_T(u,i) P_T(r|u,i)}\left[\frac{P_D(u,i)}{P_T(u,i)} \delta(r_{ui}, \hat{r}_{ui})\right]$$

• Limitation of Reweighting: Requiring positivity

- Just leveraging propensity score is insufficient:



$S: \{(u,i,r): p_U(u,i,r) > 0\}$

$S_0: \{(u,i,r): p_U(u,i,r) > 0, p_T(u,i,r) = 0\}$

$S_1: \{(u,i,r): p_U(u,i,r) > 0, p_T(u,i,r) > 0\}$

▲ :Training data

◆ :Imputed data

  - Due to the data bias, training data distribution P_T may only provide the partial data knowledge of the region S (S_O is not included)
  - IPS cannot handle this situation

- Imputing pseudo-data to the region S_O:

$$L_T = \sum_{(u,i)\in D_T} w_{ui}^{(1)} \delta(r_{ui}, r_{ui}) + \sum_{u\in U, i\in I} w_{ui}^{(2)} \delta(m_{ui}, r_{ui})$$

• 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 instance
- Generative Modeling
  - Assuming the generation process of data and reduces the biases accordingly

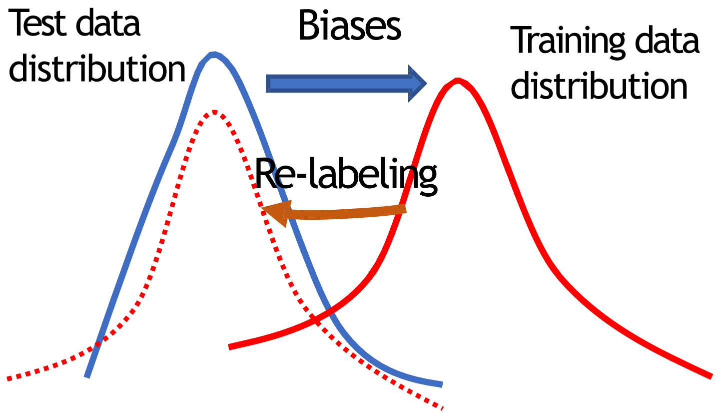
• Re-labeling Strategies

- Basic idea: change data distribution by imputing pseudo-labels:

$$L_{DI} = \sum_{(u,i)\in D_T \vee D_n} \delta(r_{ui} \backslash m_{ui}, r_{ui})$$

- Could address the deviation of $p(u, i)$ and $p(r|u,i)$

$$p_T(u,i) \neq p_D(u,i) \qquad p_T(r \mid u,i) \neq p_D(r \mid u,i)$$



- Properly defining pseudo-labels can lead to *unbiased estimator* of the ideal:

For $p_T(r \mid u,i) \neq p_D(r \mid u,i)$ ⟹ $L_{DI} = \sum_{(u,i,r)\in D_T} \delta(m_{ui}, \hat{r}_{ui}), m_{ui} \sim p_D(r \mid u,i)$

For $p_T(u,i) \neq p_D(u,i)$ ⟹ $L_{DI} = \sum_{(u,i,r)\in D_T} \delta(r_{ui}, \hat{r}_{ui}) + \sum_{(u,i)-D_T} \delta(m_{ui}, \hat{r}_{ui})$

• 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		5
2	3	4	4

Data imputation - - - →

Imputation data

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

- Relabeling: assigns pseudo-labels for missing data.

$$\arg\min_{\theta} \sum_{u,i} \hat{\delta}\left(r_{ui}^{o\&i}, f(u,i \mid \theta)\right) + \mathrm{Reg}(\theta)$$

Simple and straightforward.

Sensitive to the imputation strategy. Imputing proper pseudo-labels is more difficult.

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

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

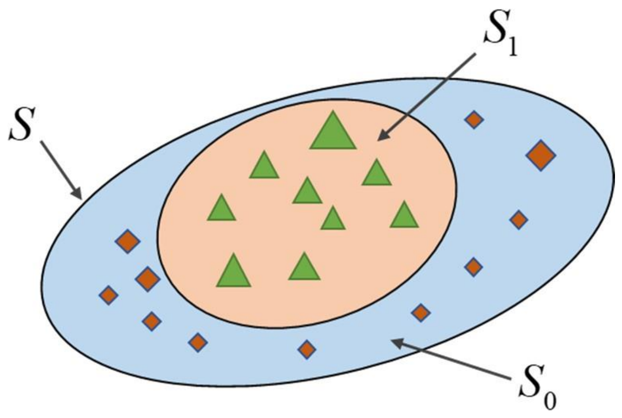
• Relabeling+Reweighting

- Reweighting:
  - Relatively Robust
  - High variance; Requires positivity

+

- Relabeling:
  - General
  - Sensitive to pseudo-labels

$$L_T = \sum_{(u,i)\in D_T} w_{ui}^{(1)} \delta(r_{ui}, r_{ui}) + \sum_{u\in U, i\in I} w_{ui}^{(2)} \delta(m_{ui}, r_{ui})$$



• 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

Relabeling+ Reweighting - - - →

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

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

$$\hat{L}_{DR} = \sum_{(u,i)\in D_T} \frac{1}{\rho_{ui}} \left(\delta(\hat{r}_{ui}, r_{ui})\right) + \sum_{u\in U, i\in I} \left(1-\frac{O_{ui}}{\rho_{ui}}\right)\delta(\hat{r}_{ui}, m_{ui})$$

IPS

Imputation

$$O_{ui} = \mathbf{I}[(u,i) \in D_T]$$

Low Variance.
Relatively robust to the propensity score and imputation value.

Requires proper imputation or propensity strategy.

Xiaojie Wang, Rui Zhang, Yu Sun, and Jianzhong Qi. 2019. Doubly robust joint learning for recommendation on data missing not at random. In ICML.

• Relabeling+Reweighting for Exposure Bias

1	1		1
	1		1
	1	1	1

Exposure bias →

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

1	1		
	1	1	1

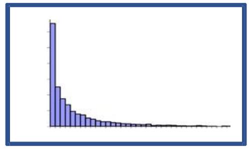
Imputation
+Weighting

1	1	0	0
0	0	0	0
0	1	1	1

$$L_w = \sum_{(u,i)\in D_T} \frac{1}{\rho_{ui}} \delta\left(r_{ui}, \hat{r}_{ui}\right) + \sum_{u\in U, i\in I} w_{ui}^{(2)} \delta\left(0, \hat{r}_{ui}\right)$$

- Imputing zero for unobserved data and downweight their contribution.
- $w_{ui}^{(2)}$ reflects how likely the item is exposed to the user.

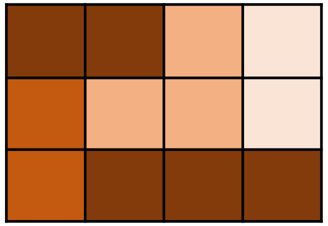
Item popularity



Social network

User community

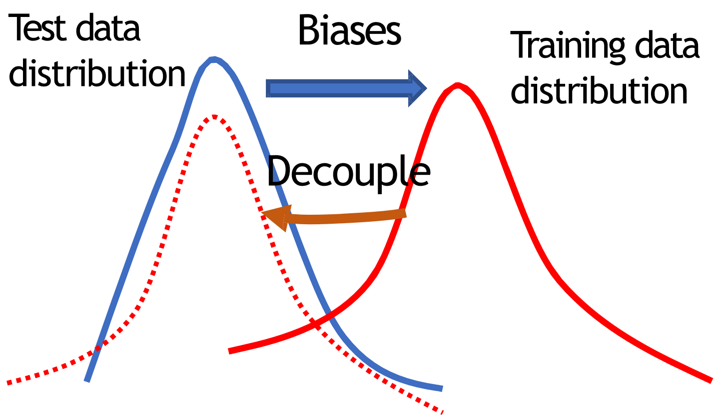
weighting

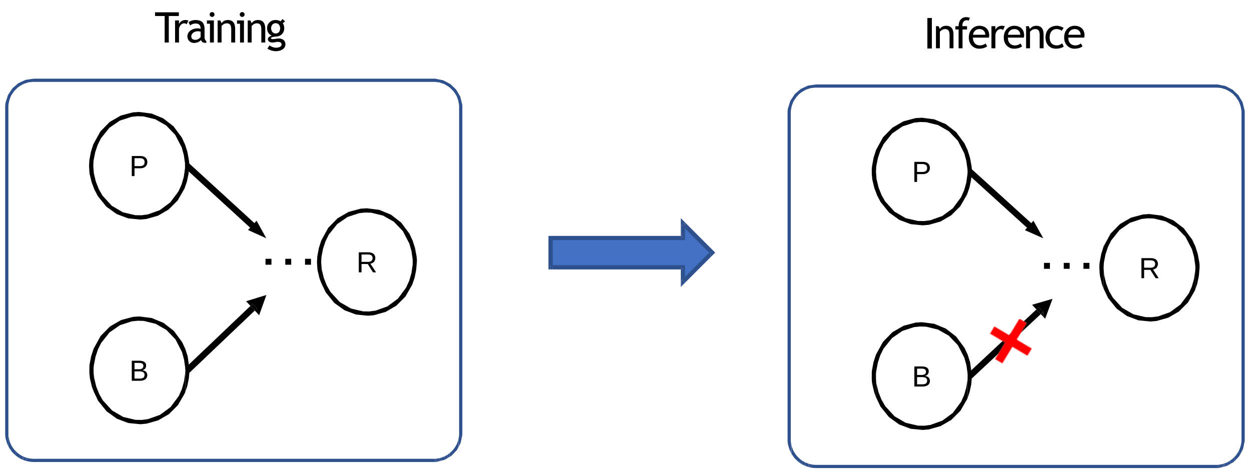


• Debiasing Strategies Overview

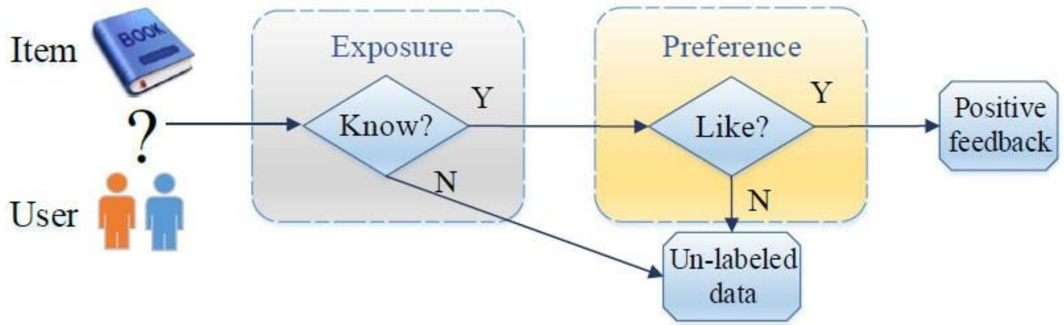
- Re-weighting
  - Giving weights for each instance to re-scale their contributions on model training
- Re-labeling
  - Giving a new pseudo-label for the missing or biased instance
- Generative Modeling
  - Assuming the generation process of data and reduces the biases accordingly

- Generative Modeling
  - Basic idea: assuming the generation process of data to decouple the effect of user true preference from the bias.





• Exposure Model for Exposure Bias (Generative modeling)



$$a_{ui} \sim Bernoulli(\eta_{ui})$$
$$(r_{ui} \mid a_{ui} = 1) \sim Bernoulli(f(u,i \mid \theta))$$
$$(r_{ui} \mid a_{ui} = 0) \sim \delta_0$$

$$\underset{\theta,\gamma}{\operatorname{argmin}} \sum_{ui} \gamma_{ui} \delta\left(r_{ui}, f(u,i \mid \theta)\right) + \sum_{ui} g(\gamma_{ui}) \qquad \gamma_{ui} \approx p(a_{ui} \mid r_{ui})$$

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

Personalized.

Learnable.

Hard to train.

Relying on strong assumptions.

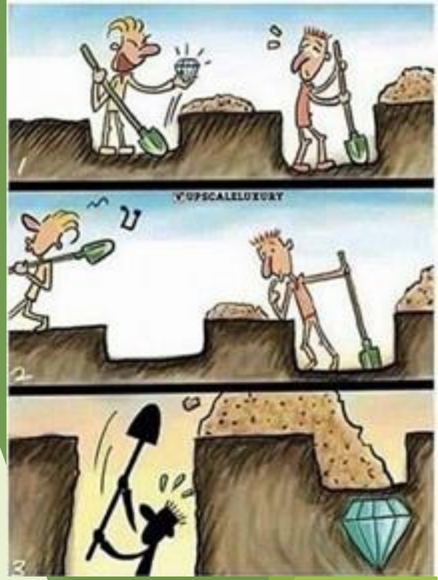
D. Liang, L. Charlin, J. McInerney, and D. M. Blei, "Modeling user exposure in recommendation," in WWW. 2016

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.

Long vs. short-term evaluation

- Exploration vs. Exploitation tradeoff

  - Purely exploitational RS: high target values in short-term, but possibly low target values in long-term
  - Problematic evaluation
    - No exploration in the train data => no way to learn it => no exploration in the test data => Penalization of exploration-oriented RS



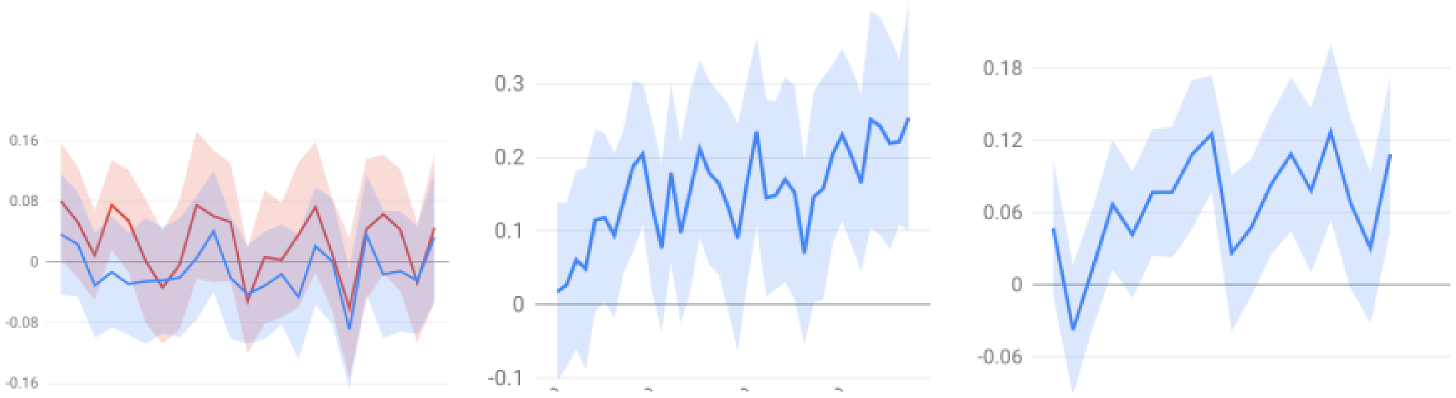
Exploration vs. Exploitation

- Values of User Exploration in Recommender Systems
  https://dl.acm.org/doi/pdf/10.1145/3460231.3474236
  - Reinforcement learning based RS (learning through rewards given for each recommendation)
  - Reward shaping / Intrinsic motivation (improved reward for relevant items from previously unknown interest clusters)

$$R_t(s_t, a_t) = \begin{cases} c \cdot R_t^e(s_t, a_t) & \text{if recommending } a_t \text{ under } s_t \text{ leads to discovery of previously unknown user interests;} \\ R_t^e(s_t, a_t) & \text{otherwise.} \end{cases} \quad (6)$$

Here $c > 1$ is a constant multiplier.

  - Promotes serendipity
  - How to transfer this for different algorithms?



(a) Entropy Regularization (b) Intrinsic motivation (c) Intrinsic Motiv. + Actionable Repre.

Figure 3: Overall user enjoyment improvement by comparing (a) Entropy regularization vs base REINFORCE; (b) Intrinsic motivation vs base REINFORCE; (c) Intrinsic motivation + Actionable representation vs Intrinsic motivation.

Exploration vs. Exploitation

- Multiarmed bandits alg. for recommendation
  - Arm = item / arm = recommending algorithm
  - https://dl.acm.org/doi/pdf/10.1145/3172944.3172967
  - Each recommended slot selected via Thompson sampling
    - Beta distribution: rewarded vs. Trials

- ***So, is this another example of the same type of solution?***

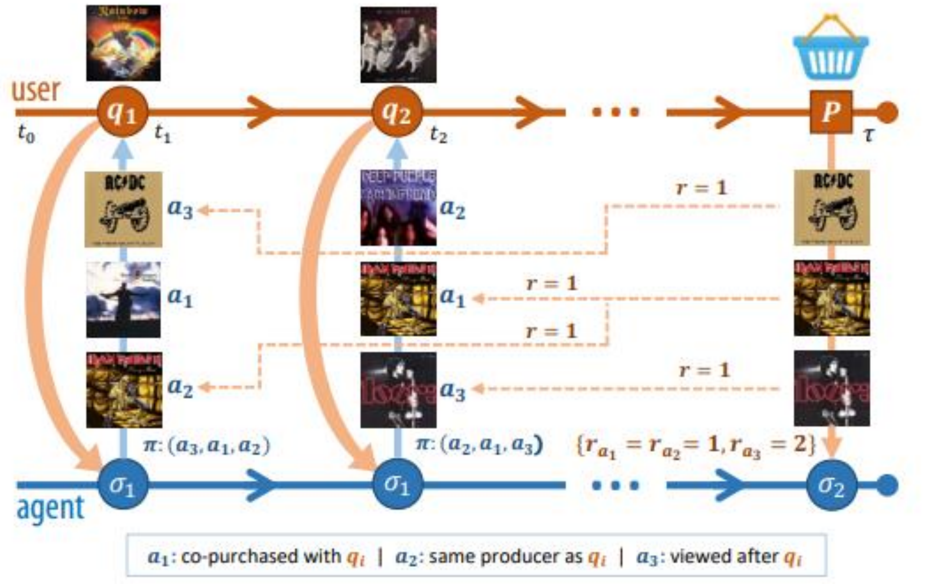
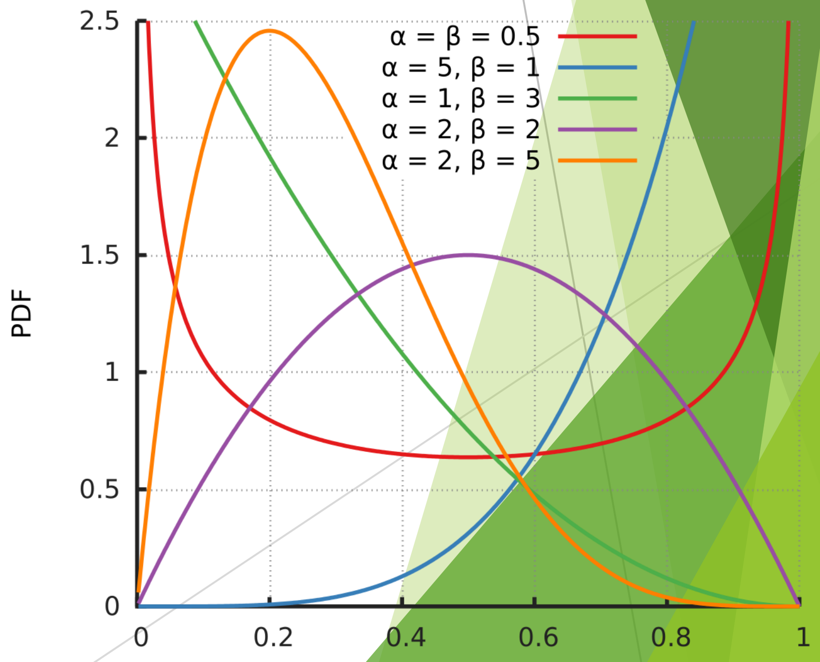
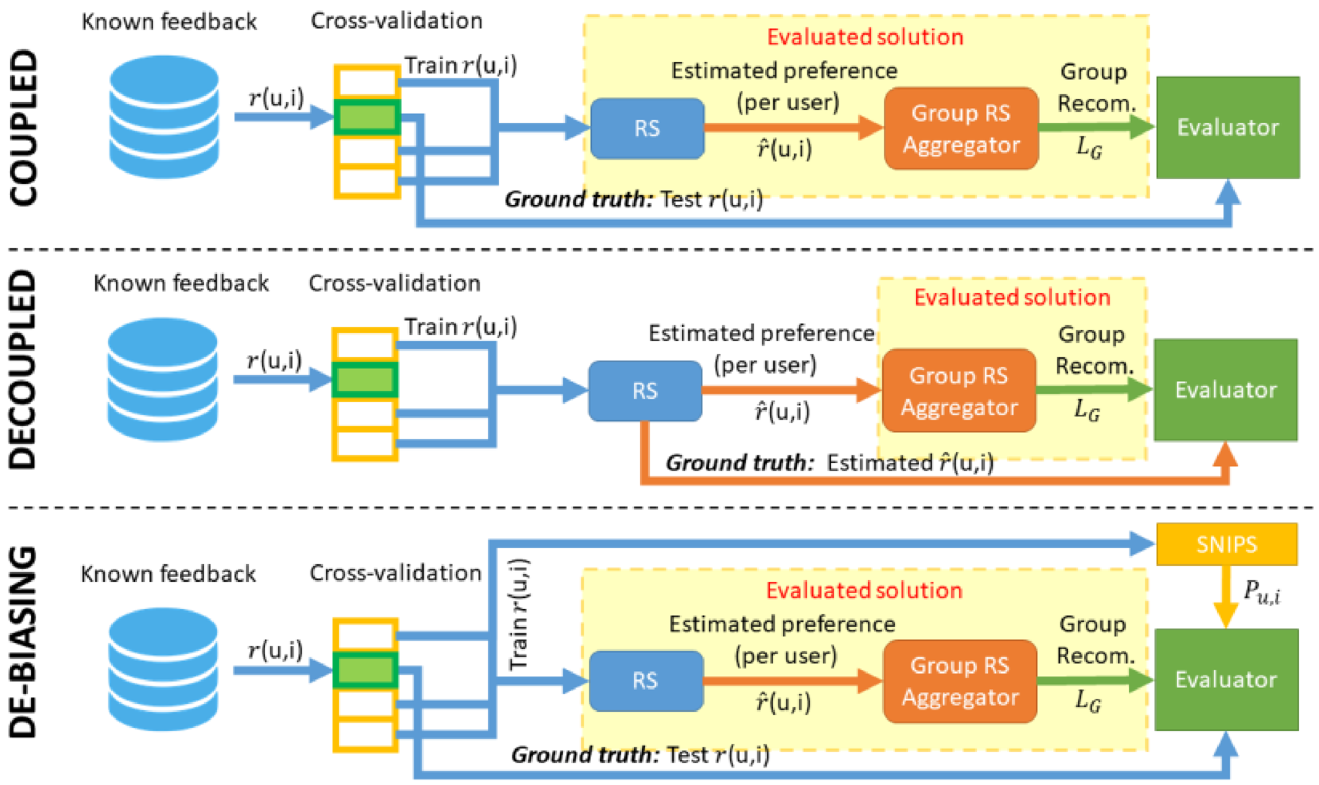


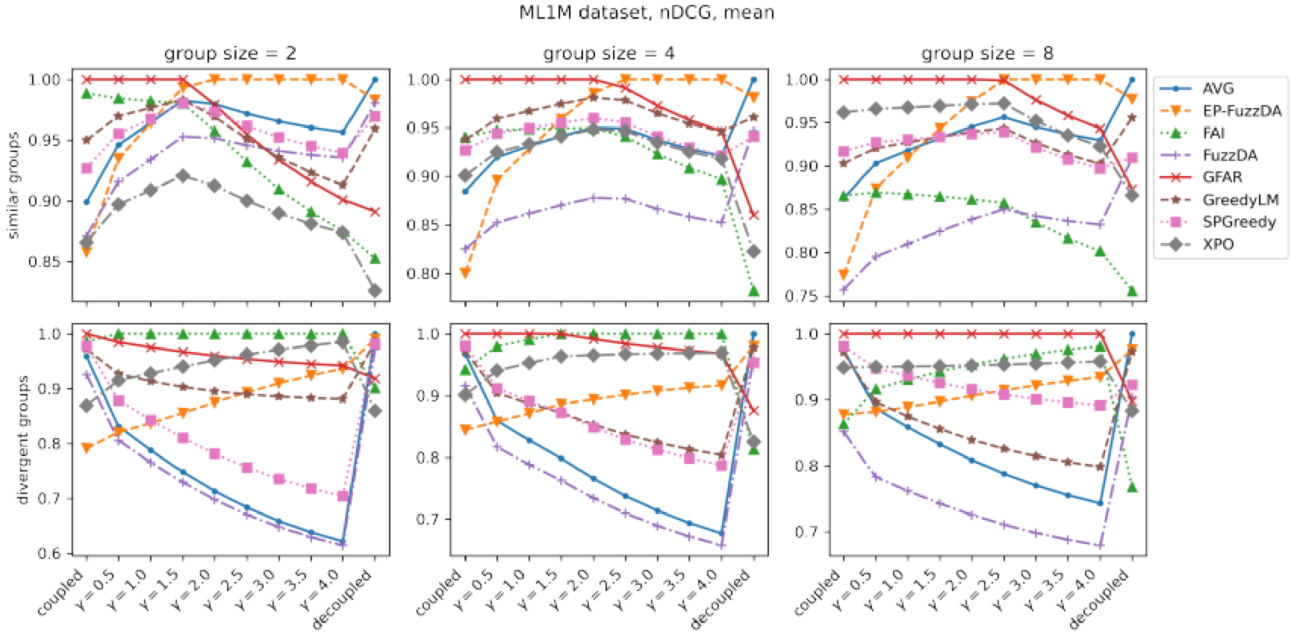
Figure 1. ***i2i*** **session-based recommendations with explainable actions**



Biases in metrics

- GFAR vs. FuzzDA – Group RS:
  - What to evaluate for group RS?
  - Decoupled evaluation depends on estimated ratings (their absolute differences)
    - Lower values / higher score differences favor „best-per-user“ algos.
    - Higher values /smaller differences may favor algorithms seeking items best in average
  - Scale [0:10]
  - $\widehat{r_{u,i1}}$ = [4, 7, 3, 4, 6] vs.
    $\widehat{r_{u,i2}}$ = [9, 1, 0, 2, 9]
  - $\widehat{c_{u,i1}}$ = [100, 20, 150, 100, 40] vs.
    $\widehat{c_{u,i2}}$ = [1, 600, 1000, 500, 5]
    - Which one is better?
    - Average estimated relevance vs. Borda count





Biases in metrics

- How to evaluate multiple metrics?
  - Recap: diversity, novelty, popularity bias, relevance

$$div_{sim}(u) = \frac{\sum_{\forall o_i, o_j \in O_u; i \neq j} 1 - sim(o_i, o_j)}{|O_u| * (|O_u| - 1)}$$

$$MMR = \arg\max_{D_i \in R \setminus S} \left[\lambda \, Sim_1(D_i, Q) - (1 - \lambda) \max_{D_j \in S} Sim_2(D_i, D_j) \right]$$

$$IP = \frac{\text{number of users who have rated the item}}{\text{number of users}} \quad (6)$$

An item's novel value (INV) is then measurable by taking the log of the inverse IP:

$$INV = -\log_2(IP) \quad (7)$$

$$DCG_{pos} = rel_1 + \sum_{i=2}^{pos} \frac{rel_i}{\log_2 i}$$

$$PopLift = \frac{mPop_{rec} - mPop_{data}}{mPop_{data}} \quad (13)$$

The $mPop_{rec}$ and $mPop_{data}$ stands for the mean popularity of items that were recommended and items that occurs in the dataset respectively. Formally, suppose to have a list of positive feedback events in a dataset $f_i(u, o) \in \mathcal{F}^+$. Each event is triggered by a user u on an item o. We can use the notation $o_j \in f_i$ meaning that the item o_j is a target in the event f_i. Then popularity of an item is defined as

$$pop(o_j) = \frac{|\{f_i : o_j \in f_i\}|}{|\mathcal{F}^+|}$$

Now, suppose that O_{rec} contains a concatenated list of all recommendations (irrespective of users) and O_{data} contains a list of target items for all events $f_i(u, o) \in \mathcal{F}^+$. Then

$$mPop_{rec} = \frac{\sum_{o_j \in O_{rec}} pop(o_j)}{|O_{rec}|} \text{ and } mPop_{data} = \frac{\sum_{o_j \in O_{data}} pop(o_j)}{|O_{data}|}.$$

Biases in metrics

- How to evaluate multiple metrics?
  - Is it good to trade 0.1 increase in diversity for 0.05 decrease in nDCG?
  - What about methods ranking?
    - But this is affected by the selection of evaluated cases
  - Pareto optimality
    - Hard to find in reality
    - Probabilistic approach: for randomly selected aggregated utility from the set of plausible ones, what is the chance that A1 is better than A2 (idea from https://dsachar.github.io/publication/2019-sac-sac/2019-sac-sac.pdf)
      - Then again, how the plausible set of utilities looks like?