The ACM Conference Series on Recommender Systems

Tutorial on Offline Evaluation for Group Recommender Systems

Theoretical background

Presenters

Francesco Barile
Maastricht University
Maastricht

Amra Delic
University of Sarajevo
Sarajevo

Ladislav Peska
Charles University
Prague

Agenda

- Theoretical background (60-70 mins)
  - Group Recommender Systems
  - Task Configuration
  - Evaluation
- Hands-on session (80-90 mins)
  - Use case 1: Movielens1M
  - Use case 2: Tourism dataset
- Conclusion

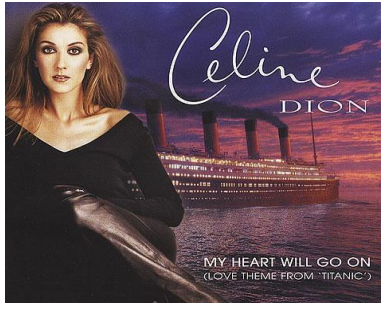
Group Recommender Systems

What are group recommendations?

- In reality various items are experienced in groups rather than individually

Movies

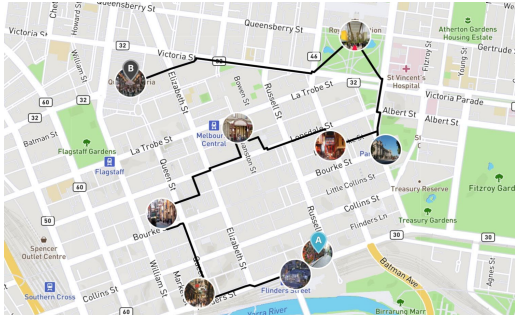
Music



Restaurants

Television program

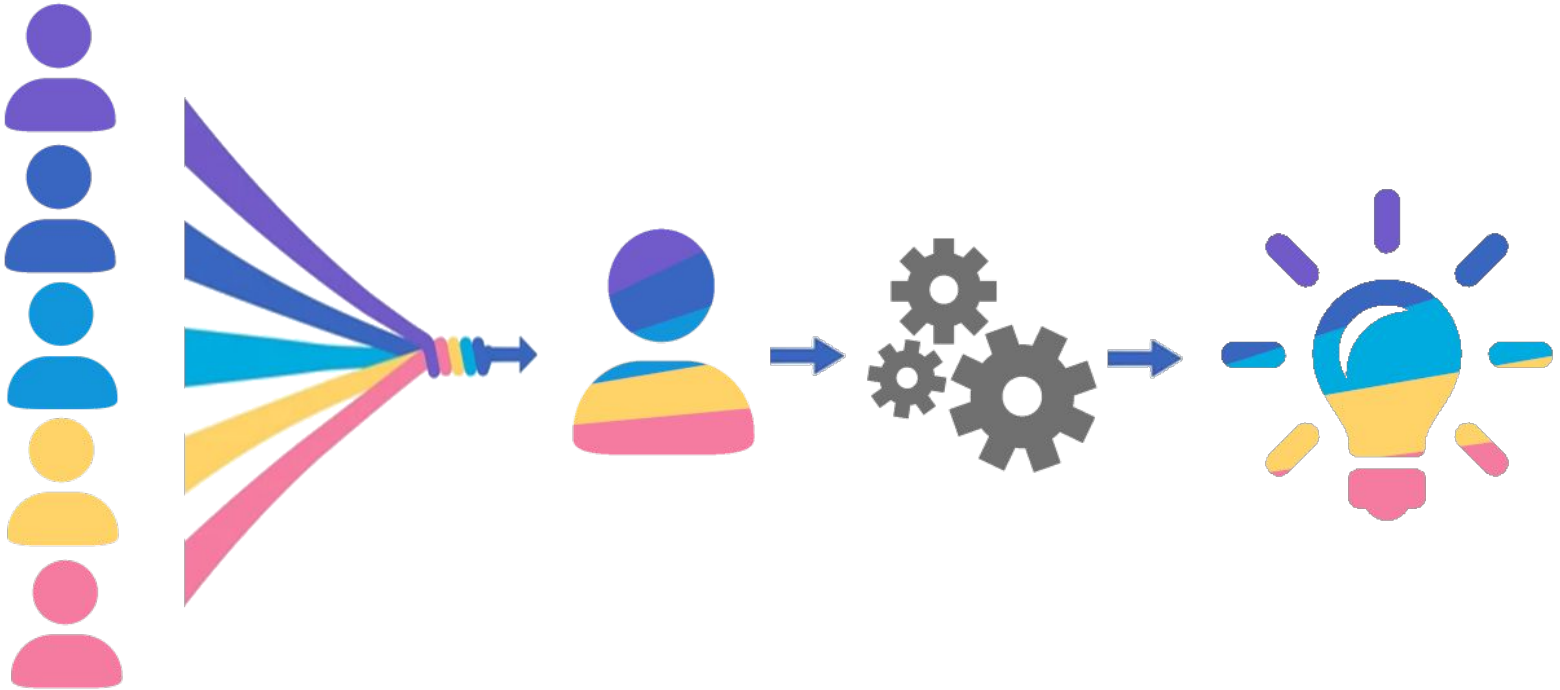
Tours



Destinations

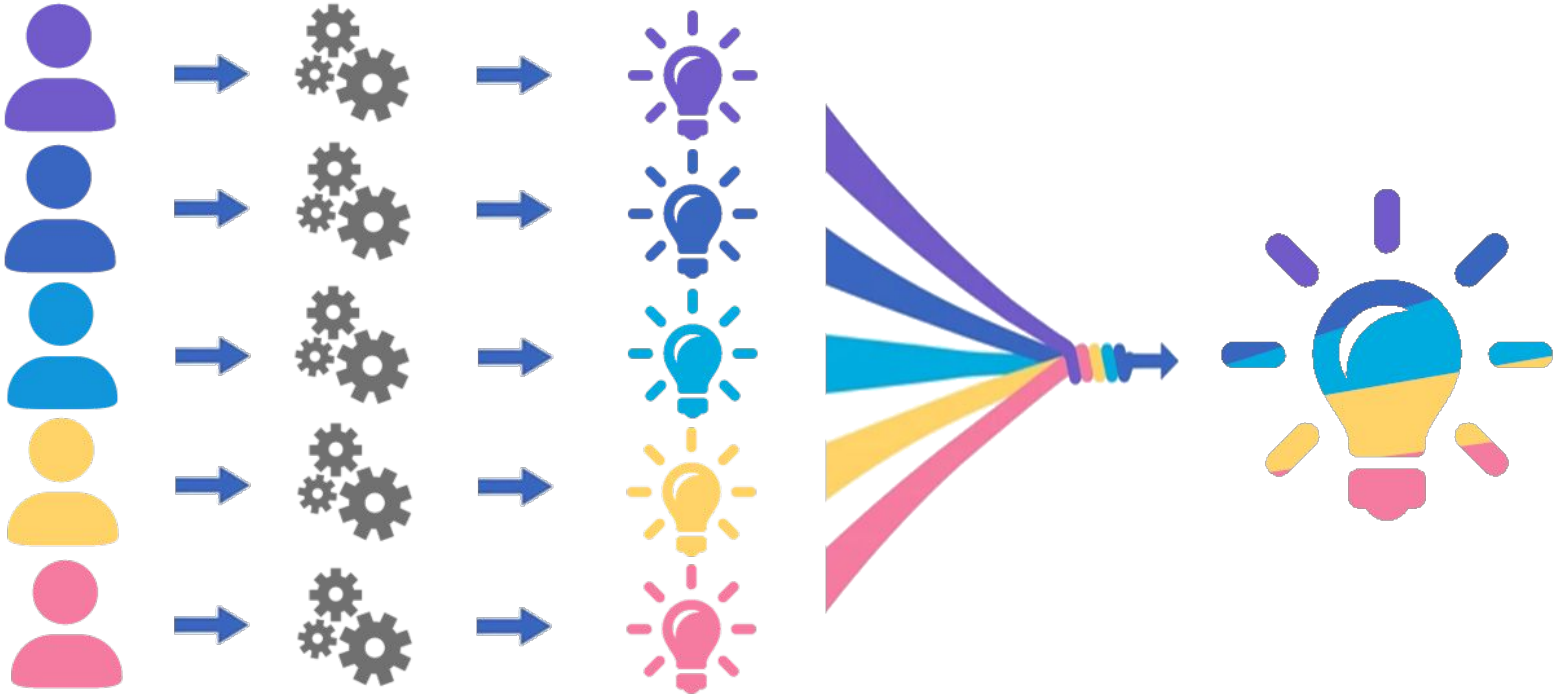
How to obtain group recommendations?

Option I: Combining individual preferences



How to obtain group recommendations?

Option II: Combining recommendations tailored for individuals



How do we combine?

- Methods that combine individual preferences / recommendations are called **aggregation strategies**
- Mainly motivated by the Social Choice theory
  - Plurality voting (majority wins)
  - Average
  - Borda count
  - Least misery
  - Multiplicative
  - …
- More complex models as well
  - Matrix factorization
  - Neural networks (AGREE[1], GroupIM[2])
  - List-wise fairness-preserving strategies
  - …

[1] Cao, D., He, X., Miao, L., An, Y., Yang, C. and Hong, R.: Attentive group recommendation. SIGIR'18

[2] Sankar, A., Wu, Y., Wu, Y., Zhang, W., Yang, H. and Sundaram, H.: Groupim: A mutual information maximization framework for neural group recommendation. SIGIR'20

Aggregation strategies

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Plurality Voting (majority based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list:

- Each group member votes for their most preferred alternative
- The alternative with the most votes wins
- Normally, each voter has one choice
  - This might easily lead to ties
  - In this example, each user votes for the best items for them

Plurality Voting (majority based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list: A

- Each group member votes for his or her most preferred alternative.
- The alternative with the most votes wins
- At the first round, the most voted item is A

Plurality Voting (majority based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list: A, E

- Each group member votes for his or her most preferred alternative.
- The alternative with the most votes wins
- At the second round, the most voted item is E

Plurality Voting (majority based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list: A, E, F

- Each group member votes for his or her most preferred alternative.
- The alternative with the most votes wins
- At the third round, the most voted item is F

Plurality Voting (majority based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list: A, E, F, (D,I)

- Each group member votes for his or her most preferred alternative.
- The alternative with the most votes wins
- At the fourth round, the most voted items are D and I

Plurality Voting (majority based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list: A, E, F, (D,I), H

- Each group member votes for his or her most preferred alternative.
- The alternative with the most votes wins
- At the fourth round, the most voted item is H

Plurality Voting (majority based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list: A, E, F, (D,I), H, J

- Each group member votes for his or her most preferred alternative.
- The alternative with the most votes wins
- At the fourth round, the most voted item is J

Additive/Average (consensus based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list: -

- Ratings are added, and the larger the sum the earlier the alternative appears in the recommendation list
  - Note that the resulting preference list is equivalent if we use the average

Additive/Average (consensus based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6
Group	**21**	**18**	**13**	**22**	**26**	**26**	**17**	**23**	**20**	**22**

Recommendation list: (E, F), H, (D, J), A, I, B, G, C

- Ratings are added, and the larger the sum the earlier the alternative appears in the recommendation list
  - Note that the resulting preference list is equivalent if we use the average

Multiplicative (consensus based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list: -

- Ratings are multiplied, and the larger the result the earlier the alternative appears in the recommendation list

Multiplicative (consensus based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6
Group	**100**	**180**	**48**	**378**	**630**	**648**	**180**	**432**	**210**	**384**

Recommendation list: F, E, H, J, D, I, (B, G), A, C

- Ratings are multiplied, and the larger the result the earlier the alternative appears in the recommendation list

Least misery (borderline)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list: -

- The group rating is the minimum of the individual ratings.
- Items get selected based on such ratings, the higher the sooner

Least misery (borderline)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6
Group	**1**	**4**	**2**	**6**	**7**	**8**	**5**	**6**	**3**	**6**

Recommendation list: F, E, (D, H, J), G, B, I, C, A

- The group rating is the minimum of the individual ratings.
- Items get selected based on such ratings, the higher the sooner

Fairness (consensus based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list: -

- Items are ordered as the group members choose them in turn
  - The order in which they choose is crucial in this case
  - Let’s assume the order is Carl, Bob, Alice

Fairness (consensus based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list: A,

- Items are ordered as the group members choose them in turn
  - The order in which they choose is crucial in this case
  - Let's assume the order is Carl, Bob, Alice
- At the first round, Carl chooses the item A

Fairness (consensus based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list: A,

- Items are ordered as the group members choose them in turn
  - The order in which they choose is crucial in this case
  - Let's assume the order is Carl, Bob, Alice
- At the second round, Bob chooses B, D, F and H
  - Here, for simplicity we assume a user chooses all the items that are equivalent for them. We could also impose to choose only one item.

Fairness (consensus based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list: A, (B, D, F, H), (E, I)

- Items are ordered as the group members choose them in turn
  - The order in which they choose is crucial in this case
  - Let's assume the order is Carl, Bob, Alice
- At the third round, Alice chooses E and I.

Fairness (consensus based)

	A	B	C	D	E	F	G	H	I	J
Alice	10	4	3	6	10	9	6	8	10	8
Bob	1	9	8	9	7	9	6	9	3	8
Carl	10	5	2	7	9	8	5	6	7	6

Recommendation list: A, (B, D, F, H), (E, I)

- Items are ordered as the group members choose them in turn
  - The order in which they choose is crucial in this case
  - Let's assume the order is Carl, Bob, Alice
- The process can continue until all items are selected, or when the group finishes the interaction process

Summary of Aggregation Strategies

Strategy	Preference List	Notes
Plurality Voting	A, E, F, (D, I), H, J, (B, G), C	Majority based
Additive/Average	(E, F), H, (D, J), A, I, B, G, C	Consensus based
Multiplicative	F, E, H, J, D, I, (B, G), A, C	Consensus based
Least Misery	F, E, (D, H, J), G, B, I, C, A	Borderline
Fairness	A, (B, D, F, H), (E, I), C, G	Consensus based (*) changing the ordering we would have different results

Other strategies exist (Approval voting, Borda count, Copeland rule, Dictatorship)

Heuristics-based Methods[1]

Make certain assumptions about group decision-making process, individual and group preferences

Graph-based ranking[2]

- Weighted bipartite graph with users and items as nodes;
  - positive links for items rated above the user's average rating
  - negative links for items rated below the user's average rating
  - a user neighbourhood graph linking users with similar rating patterns
  - an item neighbourhood graph linking items that have been rated similarly
- Group recommendations generated by two random walks:
  - highly visited items over positive links are good options for a group
  - highly visited items over negative links are poor choices for a group

[1] Masthoff, J. and Delić, A., 2022. Group Recommender Systems: Beyond Preference Aggregation. Recommender Systems Handbook.
[2] Kim, H.N., Rawashdeh, M. and El Saddik, A., 2013, March. Tailoring recommendations to groups of users: a graph walk-based approach. IUI'13.

Heuristics-based Methods

Make certain assumptions about group decision-making process, individual and group preferences

Spearman footrule rank aggregation[3]

- Aggregated rank list is a list with minimum distance to individually ranked lists
- Distance between two lists: sum of absolute differences between item ranks

[3] Baltrunas, L., Makcinskas, T. and Ricci, F. Group recommendations with rank aggregation and collaborative filtering. RecSys'10

Deep-learning Methods

AGREE strategy (Cao et al, 2018.):

- Uses a neural network to learn the aggregation strategy and item dependant weights of each group member

MoSAN strategy (Vinh Tran et al, 2019.):

- For each group, a set of sub-attention networks is created, which aims to capture individual decisions of group members, given the decisions of others in that group

GroupIM strategy (Sankar et al, 2020.):

- Trains a neural network to compute probabilities that a group would interact with an item by

(1) minimizing the group recommendation loss between history of group item interactions and the predicted item-probabilities

(2) minimizing the contextually weighted user-item loss, and

(3) maximizing the mutual information between the group and the member user

[3] Baltrunas, L., Makcinskas, T. and Ricci, F. Group recommendations with rank aggregation and collaborative filtering. RecSys'10

List-wise aggregation strategies

- Build list of recommendations incrementally
- Unfairness of partially constructed list may be reflected in next item's selection

GFAR[1]

- Maximizing sums of probabilities that at least one item is relevant for user
  - Optimizing for Hit Rate
  - One's complement (joint probability that all items are irrelevant)

$$p(\text{rel}\,|u,S) = 1 - p(\neg\,\text{rel}\,|u,S)$$
$$= 1 - \prod_{i \in S}(1 - p(\text{rel}\,|u,i))$$

  - Relevance probability as normalized Borda-count w.r.t. user's individual recommendations
    - Borda-count evaluated w.r.t. Top-20 -> swift decrease w.r.t. rank

$$p(\text{rel}\,|u,i) = \frac{\text{Borda-rel}(u,i)}{\sum_{j \in \text{top-}N_u} \text{Borda-rel}(u,j)}$$

- Next item selected based on per-item marginal gain

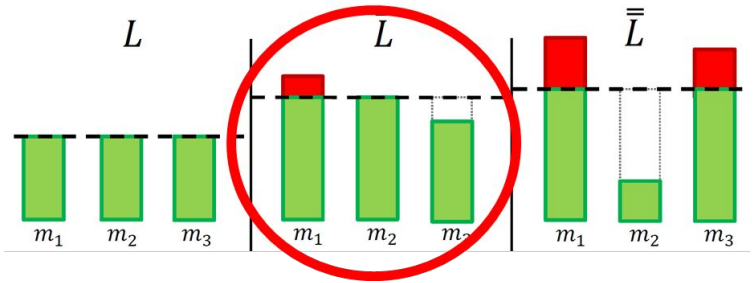
[1] Kaya, M. and Bridge, D. and Tintarev, N.: Ensuring Fairness in Group Recommendations by Rank-Sensitive Balancing of Relevance. RecSys'20

List-wise aggregation strategies

- Build list of recommendations incrementally
- Unfairness of partially constructed list may be reflected in next item's selection

EP-FuzzDA[1]

- Maximizing exactly proportional sum of all group member's gains
  - Next item selected based on per-item marginal gain
  - Optimizing for Average relevance per user



[1] Malecek, L. and Peska, L.: Fairness-preserving Group Recommendations With User Weighting. UMAP'21 LBR

Task Configuration

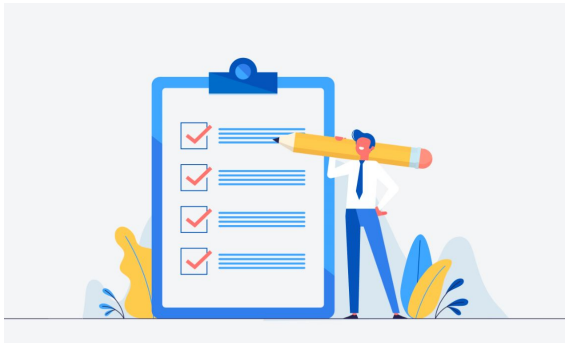
Task Configuration - Group characteristics

1. Group size
    - The larger the group, the more difficult it is to satisfy all members

Task Configuration - Group characteristics

3. Group type
    - One time group vs. long-term groups

Task Configuration - Group characteristics



4. Group decision-making process (group engagement)
    - Group is presented with a list of recommendations
    - Group negotiates the model / attributes of items
        - Conversational or critiquing techniques

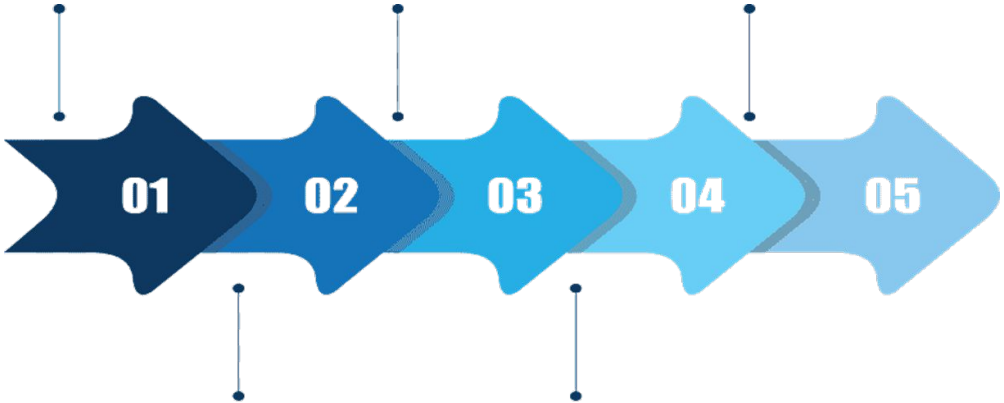
Configurations - System characteristics

5. Individual preferences
    - Long-term preferences vs. developed within the session

Task Configuration - Group characteristics

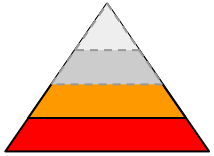
2. Group composition
    - Uniform vs. heterogeneous preferences

Configurations - System characteristics



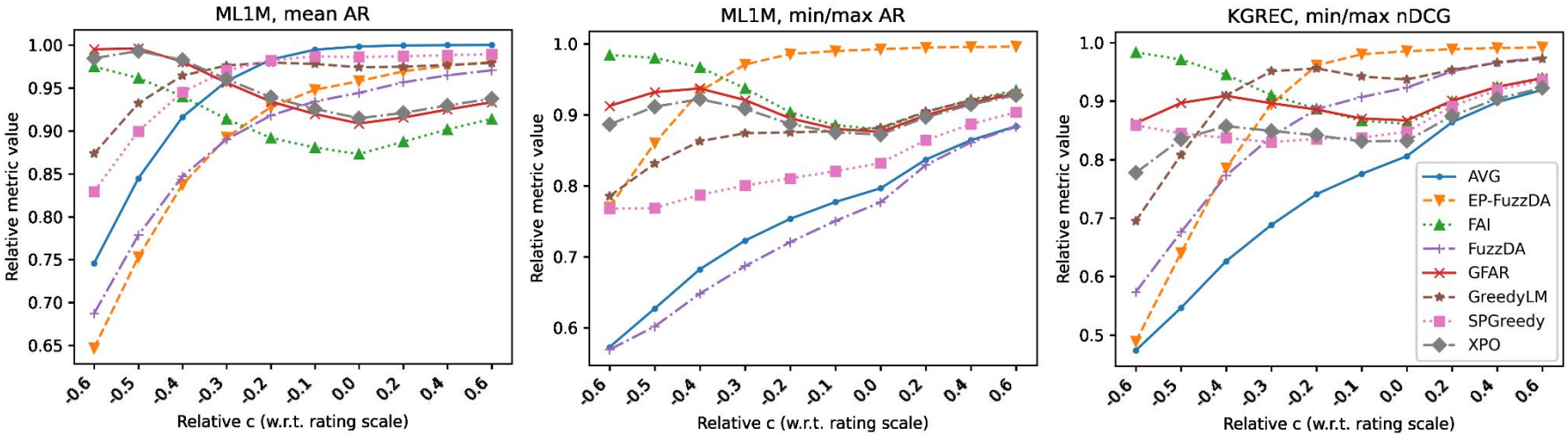
6. Recommendation type
    - One-shot recommendations
    - Repeated one-shot recommendations (long-term groups)
    - Sequence of recommendations

Biases in Decoupled Evaluation



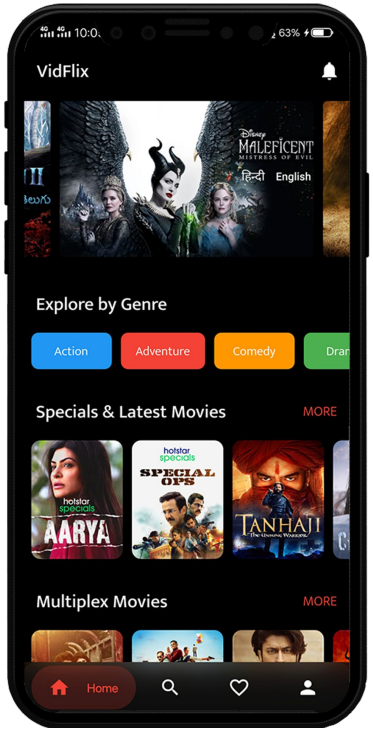
Mitigation strategy:

- Normalize predicted scores[1], e.g., max(0, score + c)



[1] Dokoupil, P. and Peska, L.: Robustness Against Polarity Bias in Decoupled Group Recommendations Evaluation. GMAP@UMAP 2022

Configurations - System characteristics



7. Recommendations are experienced vs. presented
    - e.g., music vs. movies

Configurations - System characteristics

8. Goals of the system
    - Item relevance vs. individual satisfaction vs. fairness…
    - Survey individual preferences vs. try to find items with overall agreement

Configuration implications

- Sequence of recommendations / long-term groups enables long-term evaluation
  - What is the farines w.r.t. group members after several recommendation sessions?
  - Can we balance against previous mistreatment?
  - Long-term effect of exploratory techniques?

- Experienced recommendations
  - Evaluate w.r.t. existing / developed individual preferences

- Presented recommendations
  - If possible, evaluate w.r.t. group decisions & individual satisfaction with them
  - Individual preferences might not be enough as they change during the group negotiation process

Configuration implications

- Group size & composition
  - Larger & more heterogeneous groups limits applicability of consensus seeking methods
  - Should be reflected in evaluation metrics as well
  -
- Negotiation based (interactive) scenarios
  - Hard to evaluate off-line in general
  - In some work simulations are used [1, 2]

[1] Nguyen, T. N., & Ricci, F. (2018). Situation-dependent combination of long-term and session-based preferences in group recommendations: an experimental analysis.
[2] Rossi, S., Di Napoli, C., Barile, F., & Liguori, L. (2016). A multi-agent system for group decision support based on conflict resolution styles. In International Workshop on Conflict Resolution in Decision Making.

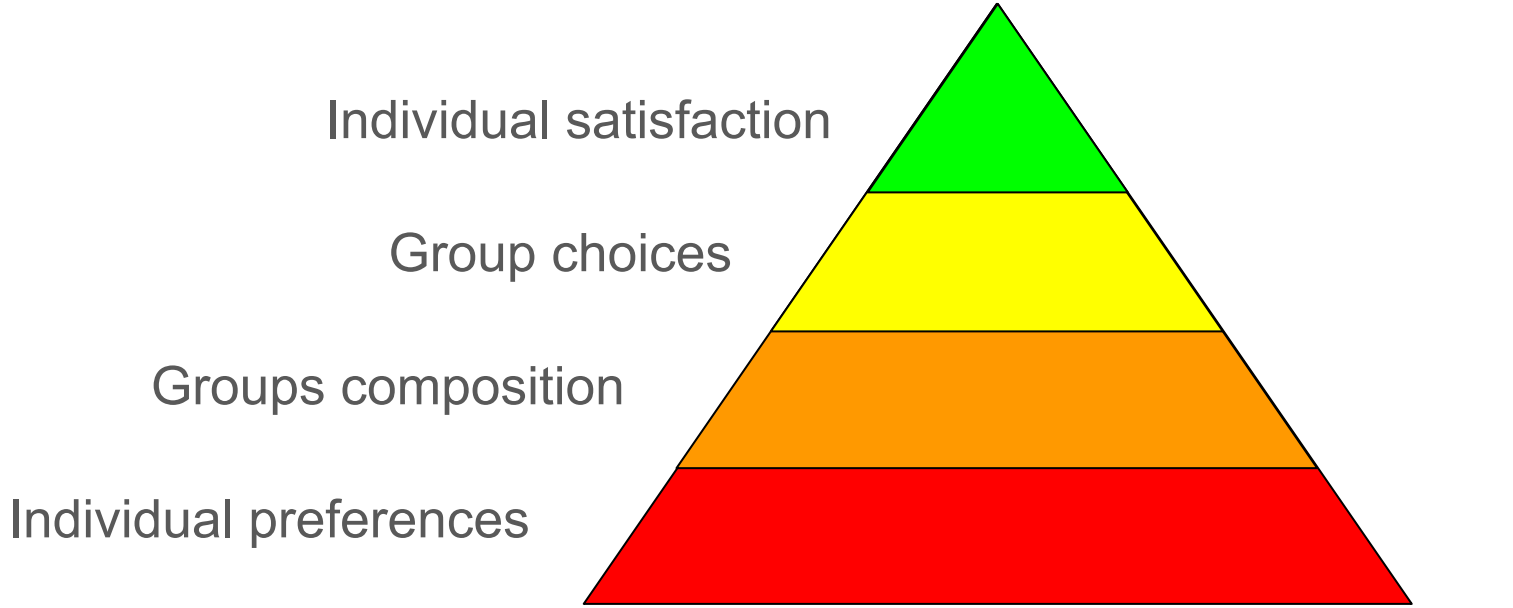
Evaluation

Evaluation

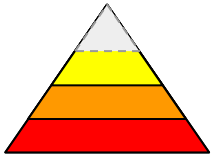
- Define the configuration
- Consider what data is available (possible to collect)
  - Real groups
    - Individual preferences + group choices + choice satisfaction
    - Individual and group characteristics
    - Group negotiation / decision process
  - Synthetic groups
    - No group data is available
    - How to generate groups (consider group related configuration dimensions)
- Coupled vs. decoupled evaluation
  - Bias mitigation strategies
- Evaluation metrics / goals of the system

Data availability pyramid

Group characteristics, group decision process…

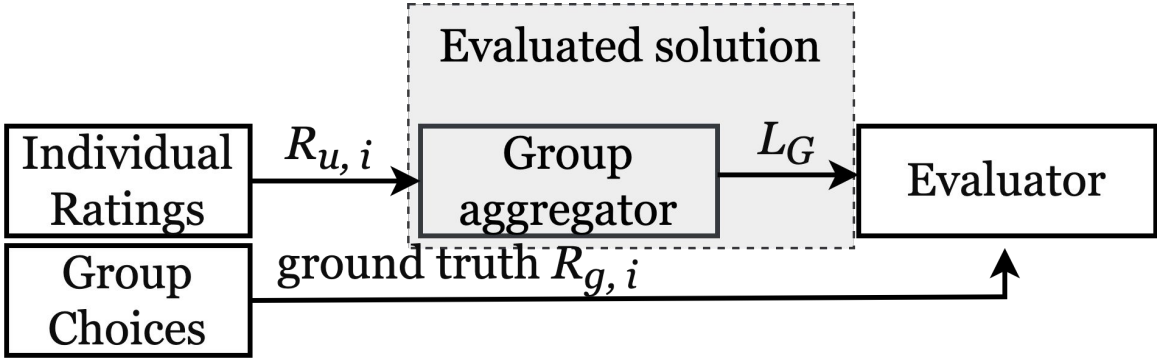


Evaluation: known choices of the group

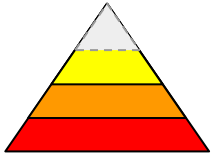


Focus on individual preferences of group members and actual group choices

- Known individual and group preferences

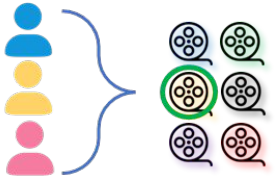


Evaluation: known choices of the group

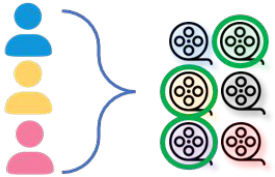


Accuracy-based metrics w.r.t. selected options:

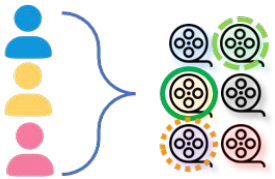
- Only one option selected
  - Hit rate @ top-k
  - Mean reciprocal rank (MRR)



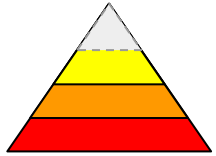
- Multiple selected choices *(e.g., not a final decision, sequential consumption…)*
  - Recall@top-k
  - Mean average precision (MAP)



- Multiple selected choices with graded relevance
  - Normalized discounted cumulative gain (nDCG)

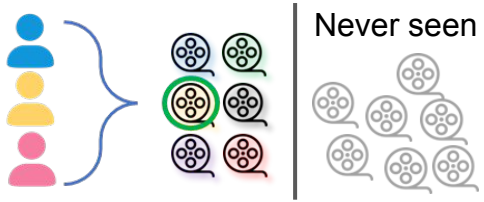


Evaluation: known choices of the group



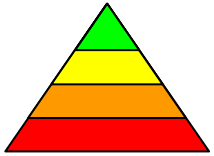
Beware of the presentation bias!

- With impression data
  - Restrict results to options that were experienced by the group
- Without impression data
  - Try to predict what was experienced by the group
  - Inverse propensity score etc.[1]



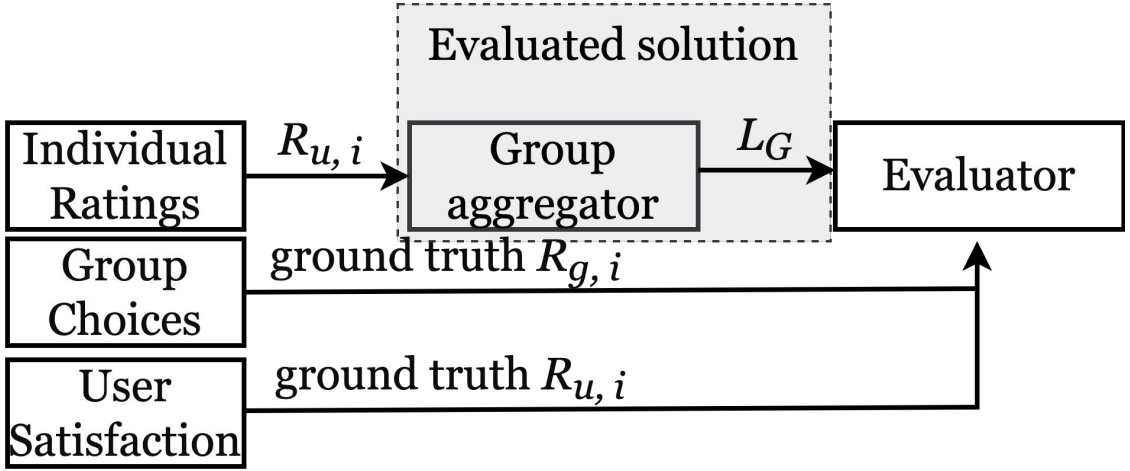
[1] RecSys2021 tutorial: Bias Issues and Solutions in Recommender System, https://lds4bias.github.io/

Evaluation: known choices + choice satisfaction

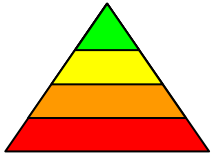


Focus on individual preferences, actual group choices and individual satisfaction

- Known individual and group preferences, as well as individual satisfaction with group choices

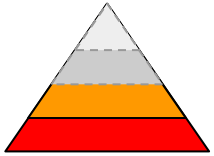


Evaluation: known choices + choice satisfaction



- Decision process fairness
  - More relevant for on-line evaluation scenarios
  - Manipulate group towards higher & more fairly distributed satisfaction
- Agreement between individual preferences and post-decision satisfaction
  - Reliability of fairness-aware approaches working with individual preferences
- Compare differences in performance w.r.t. different ground truth “levels”
- Multiple selected choices + choice satisfaction
  - Evaluate whether options with higher and/or more fairly distributed satisfaction are better ranked
  - No known dataset available

Evaluation: without known choices of the group



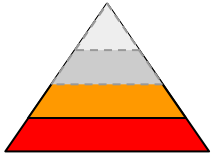
Focus on individual preferences of group members

- Known (i.e., hold-out data) => Coupled evaluation
- Estimated (i.e., given by individual RS) => Decoupled evaluation

Typically, systems have two goals:

- Overall satisfaction of the group
- Satisfaction distribution among users (i.e., fairness)

Evaluation: without known choices of the group

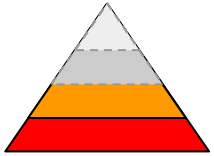


Overall satisfaction of the group

- Evaluate individual satisfaction with the items recommended to the group
  - Any metric(s) suitable for individual RS
    - Depending on assumed model of user's satisfaction
    - Precision/recall@k, MAP, nDCG, ...
    - beyond accuracy metrics (novelty, diversity, serendipity…)
  - Utilize
    - Raw metrics scores
    - Normalized scores w.r.t. recommendations that would be given to individual users[1]
- Aggregate over all group members (mean)

[1] Stratigi, M. et al: Sequential group recommendations based on satisfaction and disagreement scores. J. Intell. Inf. Syst. 2022

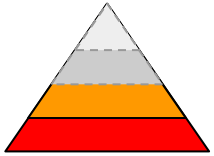
Evaluation: without known choices of the group



Satisfaction distribution among users (i.e., group fairness)

- One-hit fairness *(one relevant item per user is sufficient)*
  - Zero-recall: fraction of group members with no relevant item recommended
    - The lower the better
  - Discounted First Hit (DFH):
    - Only the first relevant item per user counts
    - Its contribution is discounted by $\log_2$ of its rank

Evaluation: without known choices of the group

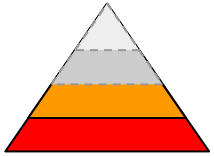


Satisfaction distribution among users (i.e., group fairness)

- m-hit fairness *(a fixed volume of relevant items per user is sufficient)*
  - Envy-freeness[1]
    - Item is *envy-free* for the user, if it is within the set of his/her top-k most prefered items
    - List of recommendations is *m-envy-free* for the user, if it contains at least *m* envy-free items
    - *m–Envy-freeness* is the fraction of group members for whom the recommendation is m-envy-free
  - Example: 3-envy-freeness of 1.0 means all group members got at least three relevant items in the recommendations

[1] Serbos, D. et al: Fairness in Package-to-Group Recommendations. WWW 2017

Evaluation: without known choices of the group



Satisfaction distribution among users (i.e., group fairness)

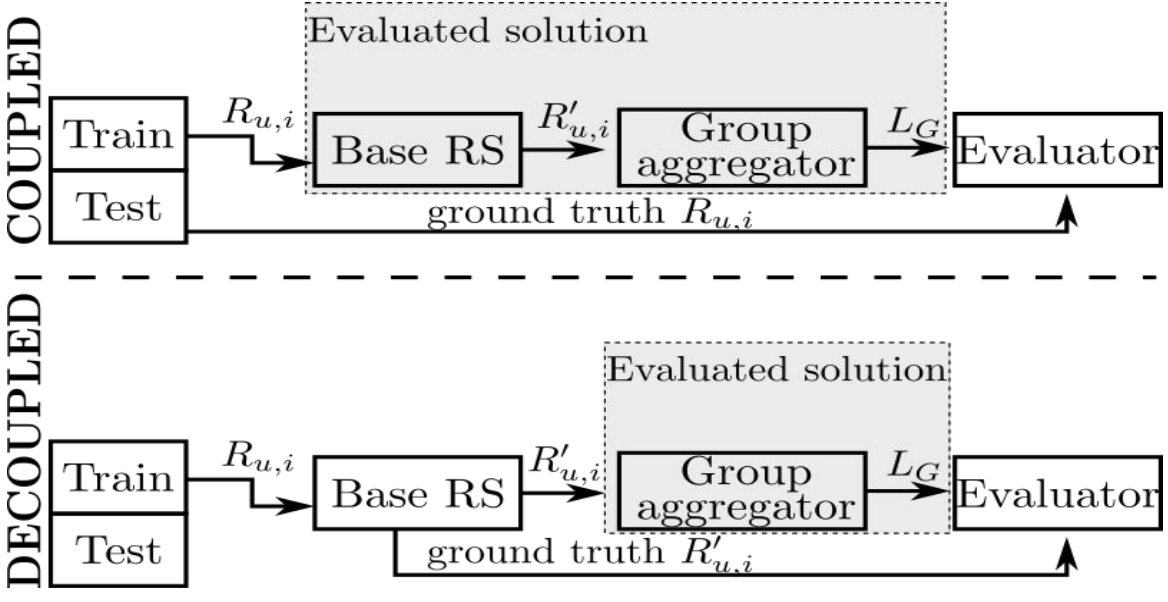
- Distribution-based fairness *(the whole list affects user's satisfaction)*
  - Select target satisfaction metric (e.g., nDCG)
  - Evaluate metric scores for all group members
    - Optionally, normalize w.r.t. recommendations that would be given to individual users[1]
  - Return some statistics of the per-group scores distribution
    - Minimum score per group
    - Minimum / Maximum score ratio
    - *Depict the distribution itself, e.g. a histogram*[2]
    - *Compare to the ideal distribution (e.g., uniform)*

[1] Stratigi, M. et al: Sequential group recommendations based on satisfaction and disagreement scores. J. Intell. Inf. Syst. 2022
[2] Ekstrand, M. and Carterette, B. and Diaz, F.: Evaluating Recommenders with Distributions. Perspectives@RecSys 2021

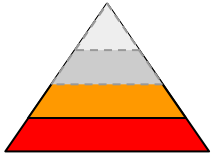
Evaluation: without known choices of the group

Focus on individual preferences of group members

- Known preferences (i.e., hold-out data) => Coupled evaluation
- Estimated preferences (i.e., given by individual RS) => Decoupled evaluation



Coupled vs. Decoupled Evaluation



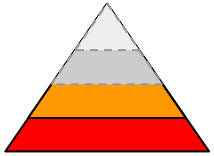
Coupled

- Suitable both for combined recommendations and combined preferences
- Measures combined performance of the whole GRS solution
- Affected by the performance of base RS
- Affected by biases in test data

Decoupled

- Suitable for combined recommendations
- Measures individual performance of recommendations aggregator
- Affected by the biases present in recommendations

Biases in Coupled Evaluation



Popularity bias in test data + popularity bias in most RS algorithms

- Feedback on popular items is known more often
- Highly popular items are often selected as top choices for most users

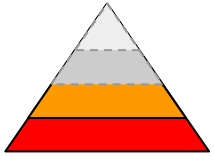
=> Unfair advantage for aggregators mostly displaying user's top recommended items

Mitigation strategies:

- Evaluate w.r.t. Inverse Propensity Score re-weighting[1]
- Re-ranking based on Value-aware Ranking[2]

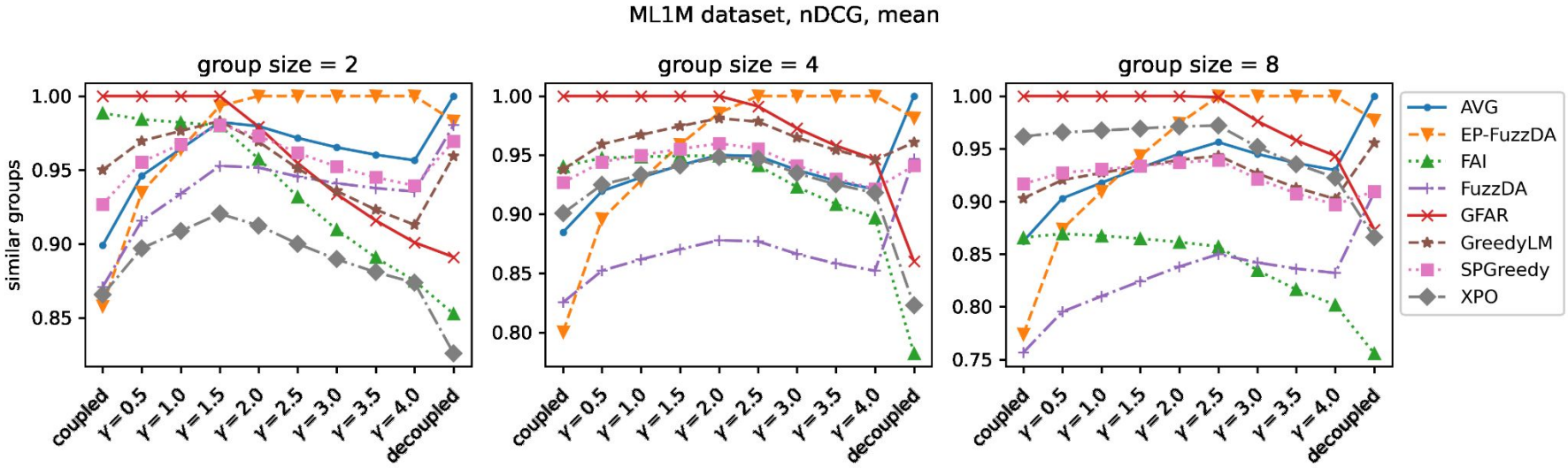
[1] Peska, L. and Malecek, L.: Coupled or Decoupled Evaluation for Group Recommendation Methods? Perspectives@RecSys 2021
[2] Yalcin, E. and Bilge, A.: Investigating and counteracting popularity bias in group recommendations. Inf. Process. Manage. September 2021

Biases in Coupled Evaluation



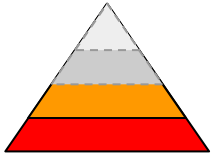
Effect of inverse propensity score weighting on evaluation results[1]

- Increasing γ denote more severe penalty for recommending popular items



[1] Peska, L. and Malecek, L.: Coupled or Decoupled Evaluation for Group Recommendation Methods? Perspectives@RecSys 2021

Biases in Decoupled Evaluation



Scores predicted by RS (i.e. ground truth) may be biased compared to the true user preferences

- RS may **over-estimate** or under-estimate true preferences

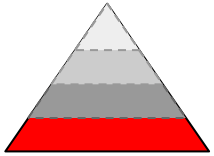
  Selection bias (feedback on positive cases is provided more often)

  ⇒ mean(feedback) > mean(preference)
  ⇒ Unbiased RS: mean(predict. scores) = mean(feedback) > mean(preference)

Over-estimation provides an unfair advantage to consensus seeking algorithms

- Items seem more relevant than they really are

Evaluation: without actual groups

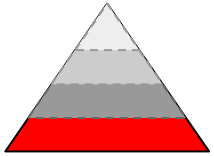


Synthetic groups generation

- Group size
- User - group assignment method
  - Based on user - user feedback similarity
    - Similar[1] / dissimilar[1,2] / combination of previous[2]
  - Based on some shared property
    - Spatio-temporal proximity in POIs systems[3]
  - Random

[1] Kaya, M. and Bridge, D. and Tintarev, N.: Ensuring Fairness in Group Recommendations by Rank-Sensitive Balancing of Relevance. RecSys'20
[2] Stratigi M. et al.: Sequential group recommendations based on satisfaction and disagreement scores. J. of Intelligent Information Systems 2022
[3] Sankar, A. et al.: GroupIM: A Mutual Information Maximization Framework for Neural Group Recommendation. SIGIR'20

Evaluation: without actual groups

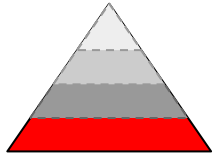


Synthetic groups generation based on user - user feedback similarity[1]

- Calculate user - user similarity metric (e.g., Pearson Correlation Coefficient)
- Select first member at random
- While group is not full:
  - **Similar:** Add user, whose similarity is > threshold to some of the existing group members
    - Other options possible, e.g. similarity > threshold for all members, or selection from top-k most similar users
  - **Divergent**: Add user, whose similarity is < threshold to some / all existing group members
  - **Minority groups:** two or more subgroups, similar within, diverse to other groups

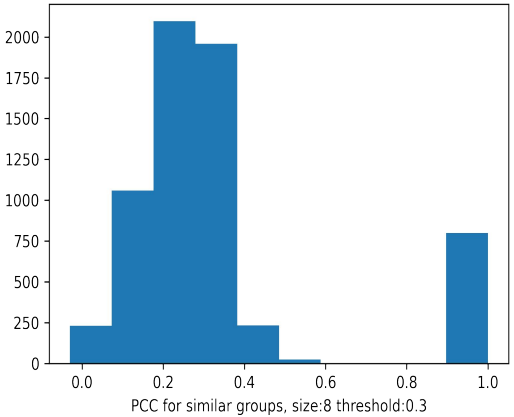
[1] Kaya, M. and Bridge, D. and Tintarev, N.: Ensuring Fairness in Group Recommendations by Rank-Sensitive Balancing of Relevance. RecSys'20

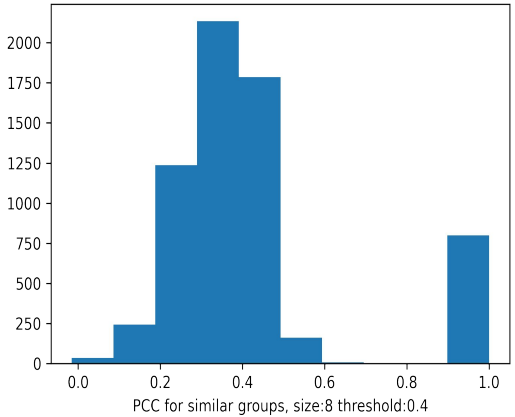
Evaluation: without actual groups

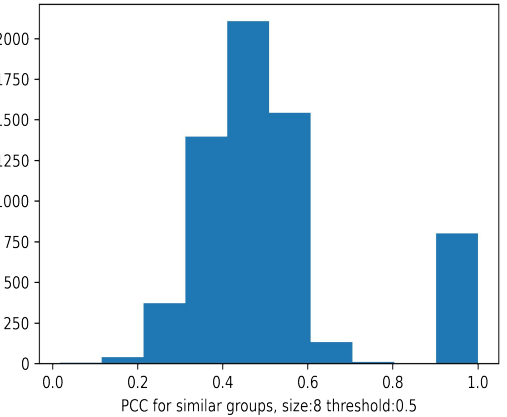


Synthetic groups generation based on user - threshold matters

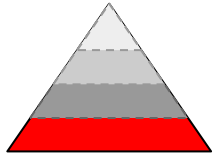
- Similar groups





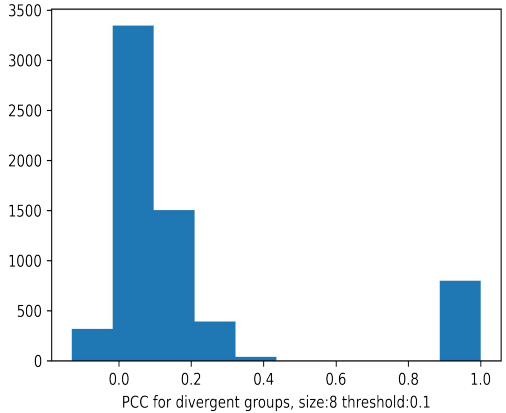


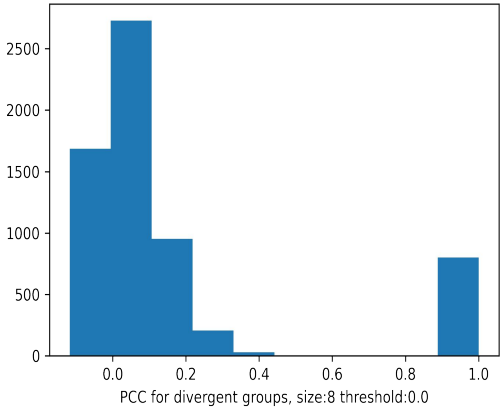
Evaluation: without actual groups

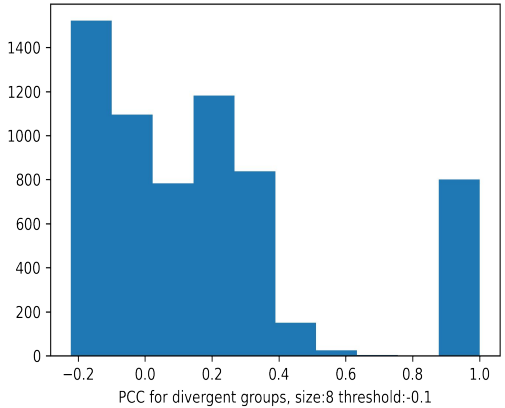


Synthetic groups generation based on user - threshold matters

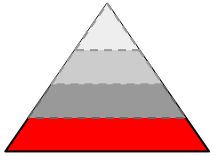
- Divergent groups







Evaluation: without actual groups



Synthetic groups generation based on shared property

- Ephemeral groups in POIs social networks[1]

  Define (ephemeral) group as:
  - Users who checked at the same POI within a certain time-frame (15 mins)
  - Users are at each other's list of friends

[1] Sankar, A. et al.: GroupIM: A Mutual Information Maximization Framework for Neural Group Recommendation. SIGIR'20

Long-term Evaluation

Evaluate multiple recommending sessions per-group

- Focus on overall utility and perceived fairness after *K* steps…
- (Dis)allow algorithms to take into account results of previous sessions
  - E.g., increase importance of previously under-represented users[1,2]

Data partitioning

- None: i.e., recommendations are consumed sequentially *(music in shared environment)*
- Random splitting[1]
- Multiple time-based partitions[2]
  - Event (timestamp) vs. Session level[3]
  - How to temporally align group members? *(in case of synthetic groups)*

[1] Malecek, L. and Peska, L.: Fairness-preserving Group Recommendations With User Weighting. UMAP 2021 LBR
[2] Stratigi M. et al.: Sequential group recommendations based on satisfaction and disagreement scores. J. of Intelligent Information Systems 2022
[3] Quadrana, M. and Cremonesi, P. and Jannach, D.: Sequence-Aware Recommender Systems. ACM Comput. Surv. 2019

Evaluation concerns

- Datasets unavailable (no ground truth)
- Synthetic data mostly used
  - What is considered as a group choice (ground truth)
- Metrics
- No agreed (solid) baselines
- No comparison between different (more complex) methods
- …

Hands on

Use-cases

MovieLens

Tourism Dataset

Use case 1: MovieLens 1M Dataset

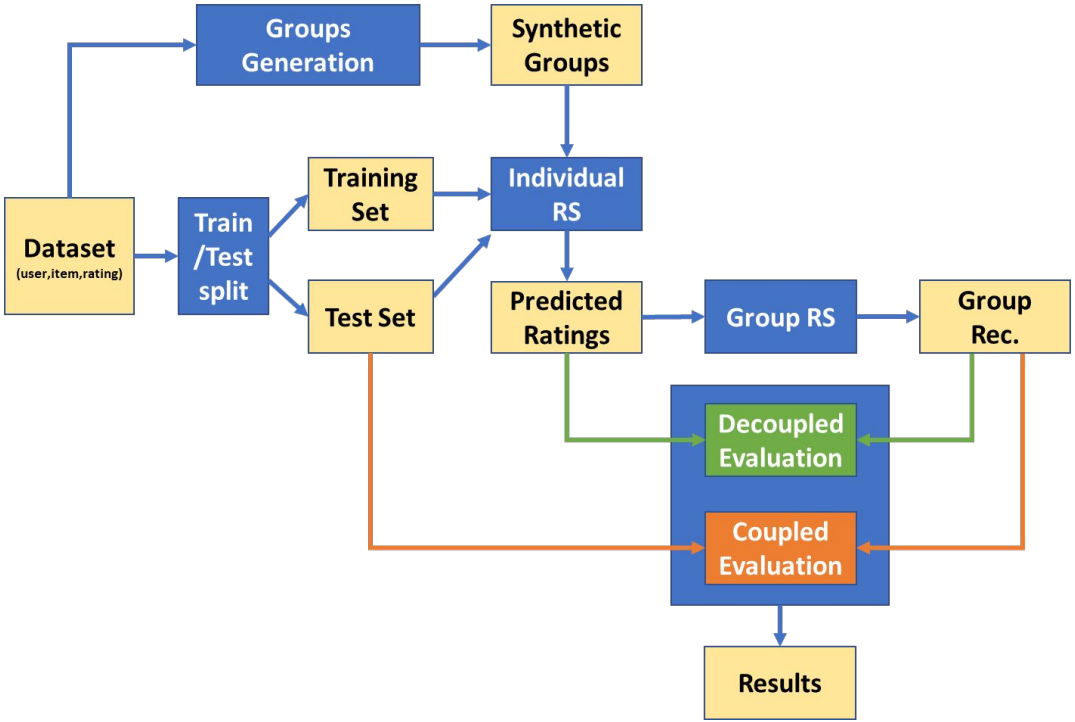
https://grouplens.org/datasets/movielens/1m/

- 1,000,209 anonymous ratings
- approx 3,900 movies
- 6,040 MovieLens users

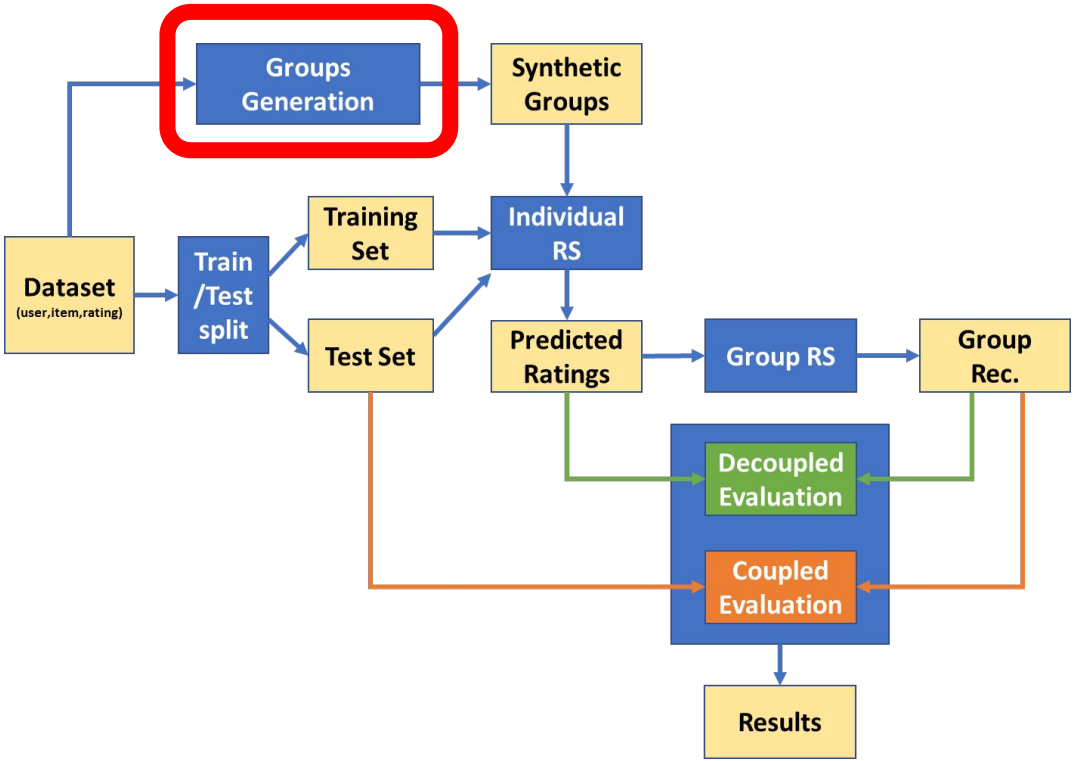
No information about groups

- We need to generate synthetic groups

Evaluation pipeline



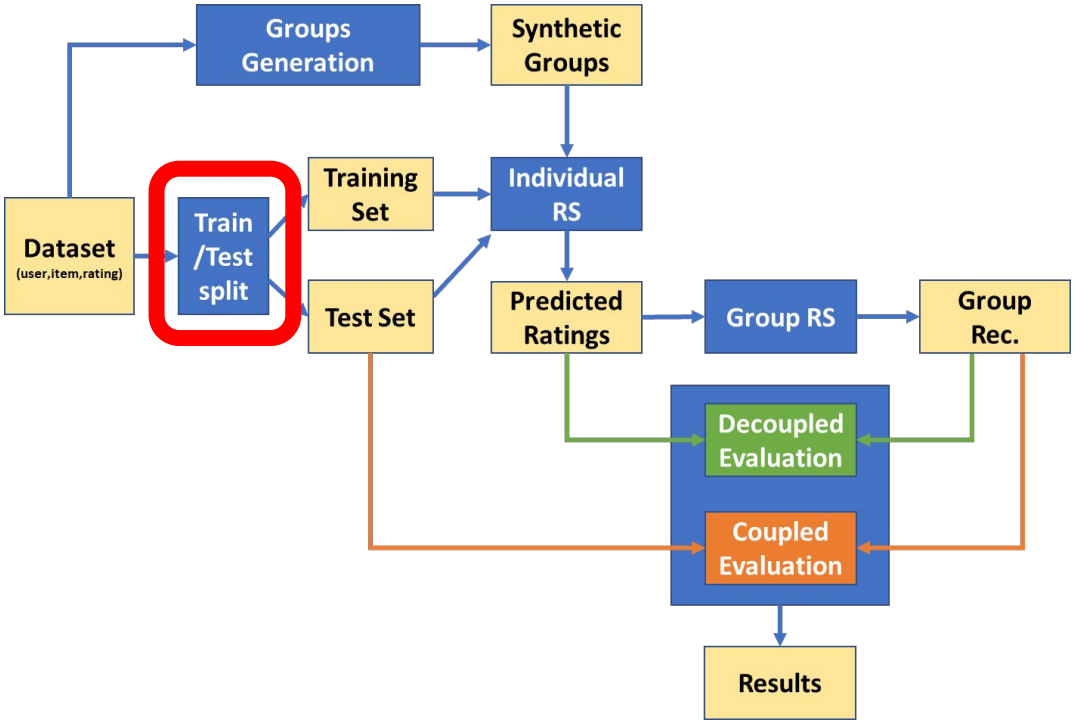
Evaluation pipeline



Groups Generation

- Which types of groups will we generate?
- Which sizes we want to evaluate?
- Similarity metric?
- How many groups?

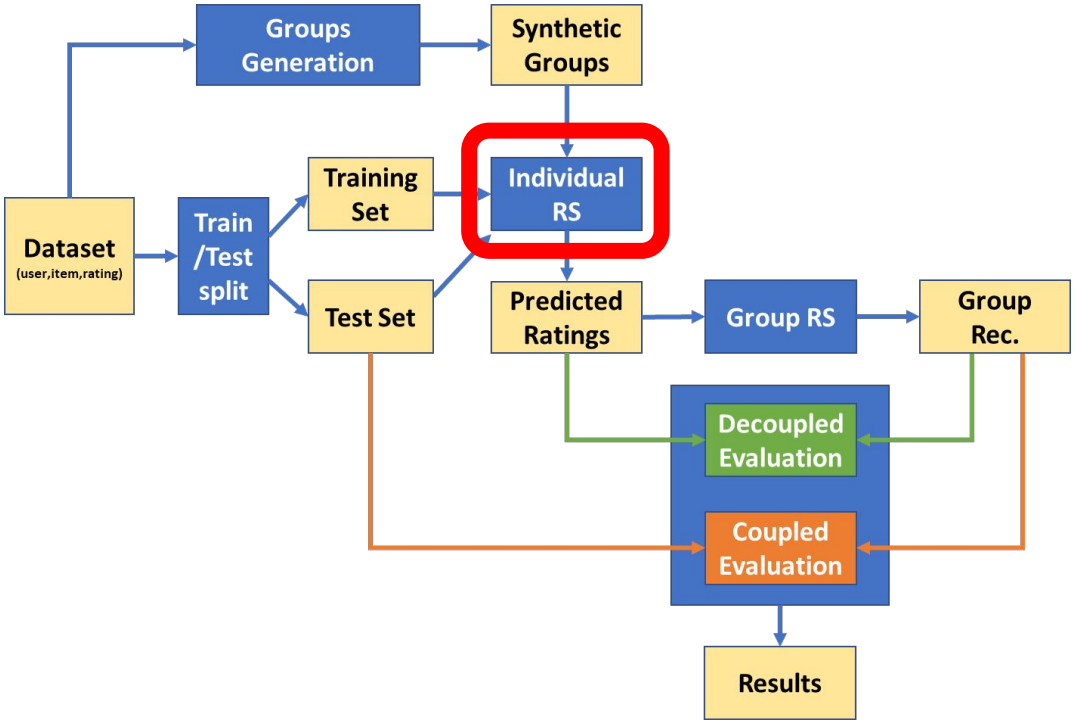
Evaluation pipeline



Train/Test Split

- Which strategy?
  - K-fold validation
- Stratified split?

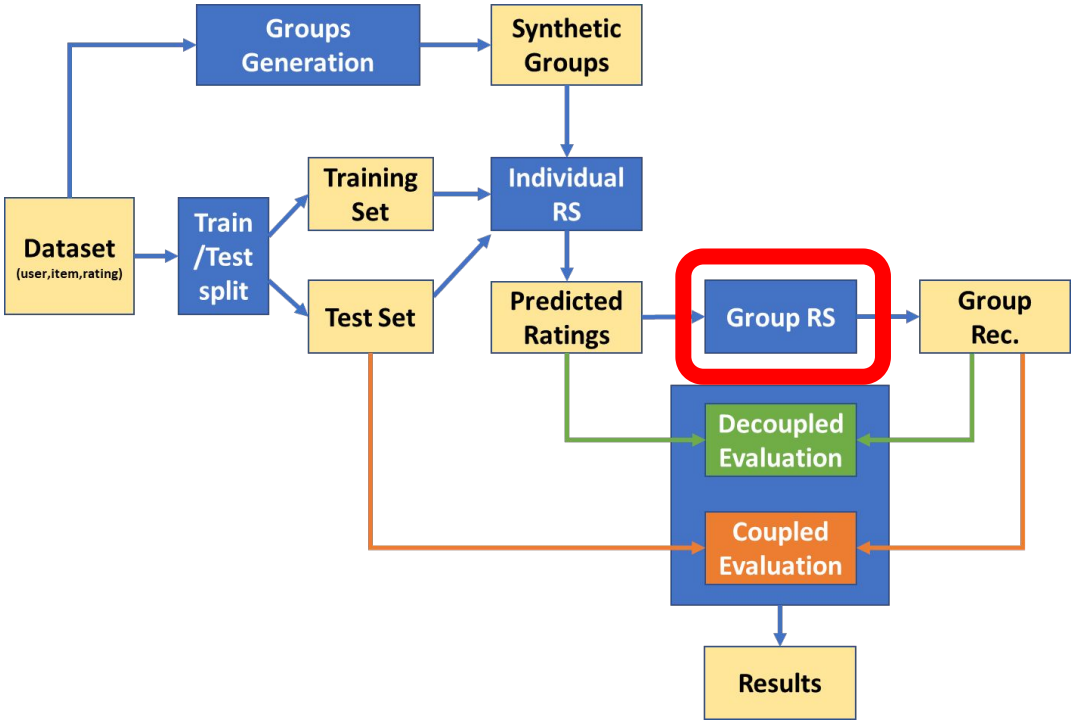
Evaluation pipeline



Individual RS

- Which strategy to use?
- Which items we consider for computing the predicted ratings?

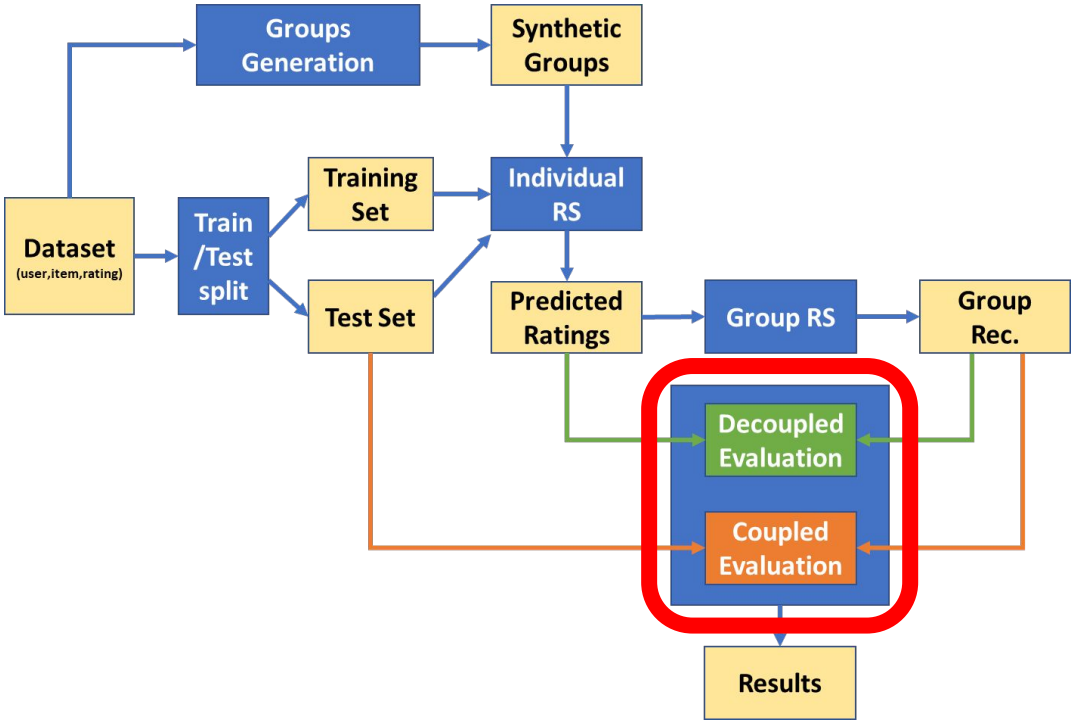
Evaluation pipeline



Group RS

- Which strategies we want to evaluate?

Evaluation pipeline



Evaluation

- Metrics to use?
- Coupled or Decoupled?

A quick look at the code

https://github.com/barnap/group-recommenders-offline-evaluation

Brainstorming and hands on session

- Divide in groups
- Work on setting up the evaluation
  - Which groups will we generate? How many?
  - Which methodology (train/test split)?
  - Which individual recommender?
  - How do we generate group recommendations?
  - Which metrics do we evaluate?
  - What evaluation setting (coupled/decoupled)?
- Play with the pre-computed results
- After we will analyze some of the obtained results

Results Analysis

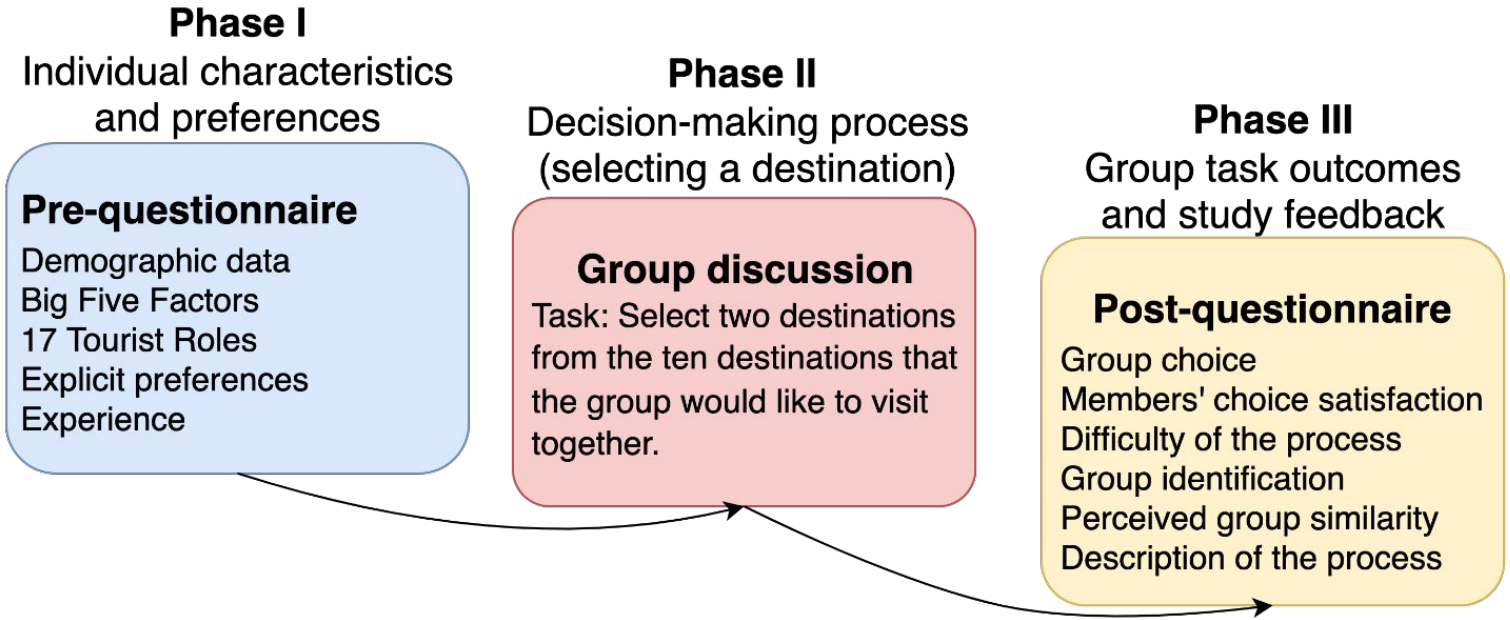
Use case 2: Tourism dataset

- Collected in 2018 among students (TU Wien, TU Delft, Uni Leiden, Uni Sarajevo)
  - 282 participants in 79 groups
  - deciding on a travel destination to visit together
- For details on data collection refer to [1] and [2]

[1] Delic, A., Neidhardt, J., Nguyen, T.N. and Ricci, F.: An observational user study for group recommender systems in the tourism domain. Information Technology & Tourism 2018.
[2] Delic, A., Neidhardt, J., Nguyen, T.N., Ricci, F., Rook, L., Werthner, H. and Zanker, M.: Observing group decision making processes. RecSys'16

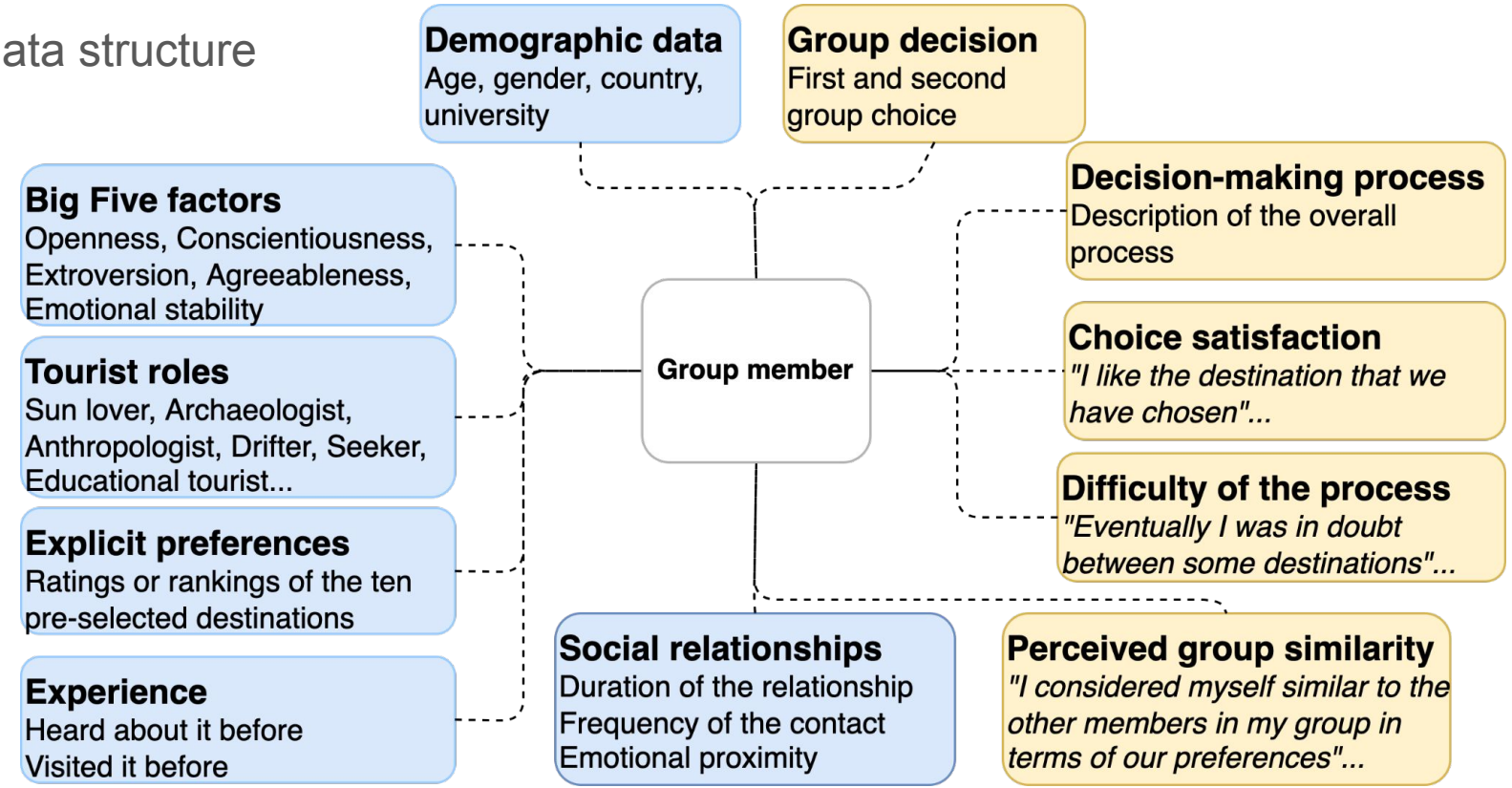
Use case 2: Tourism dataset

Data collection procedure



Use case 2: Tourism dataset

Data structure

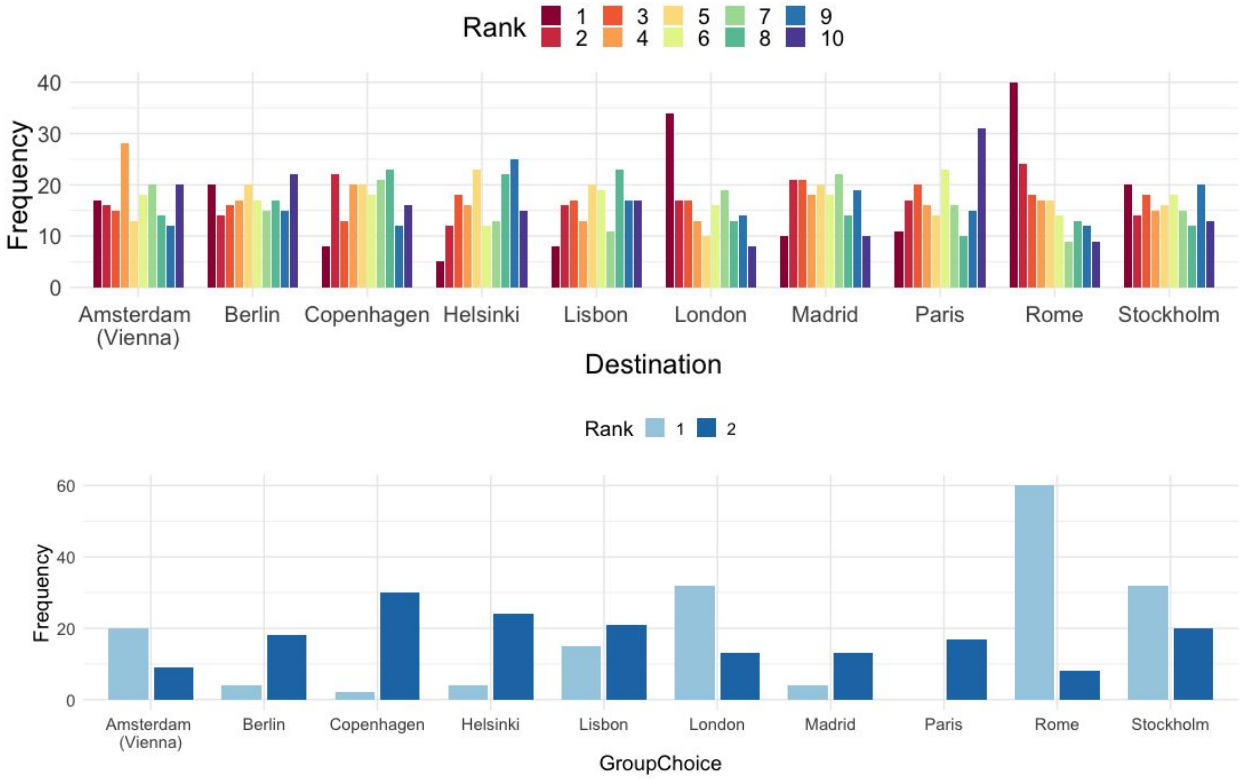


Use case 2: Tourism dataset

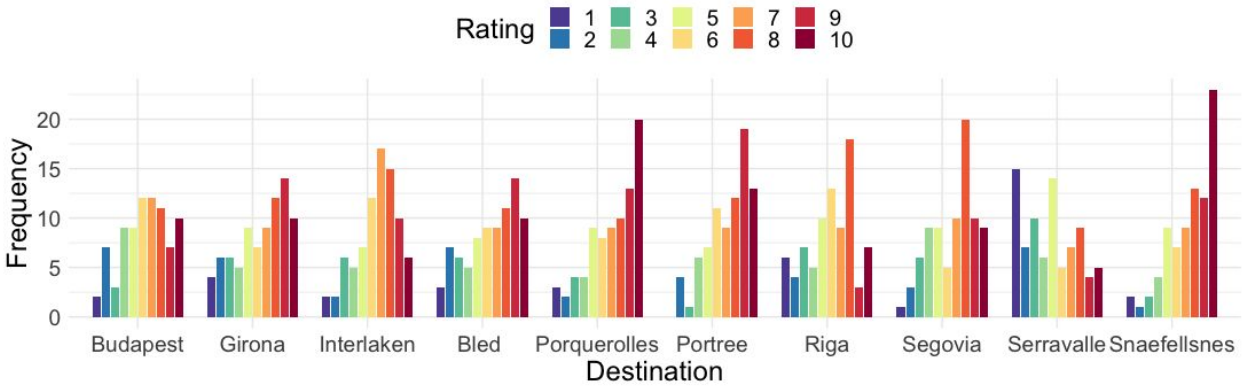
Provided data:

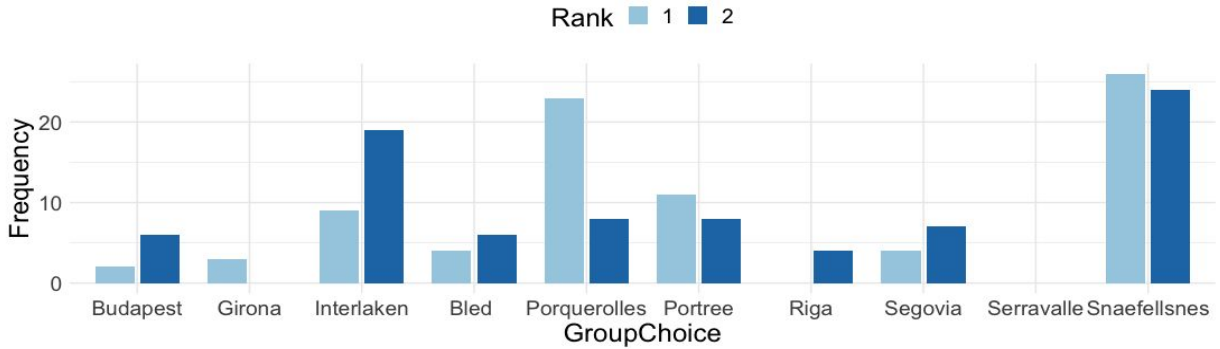
- Ratings
  - individual rating for the 10 travel destinations
- Groups composition
  - information about the group members
- Group choices
  - Outcome of the group discussion
  - 2 selected travel destinations (first and second choice) for each group
- Individual Feedback
  - Individual evaluation of the group choices and group decision-making process
  - Choice satisfaction, difficulty of the process, identification with the group

Use case 2: Tourism dataset (descriptives)



Use case 2: Tourism dataset (descriptives)

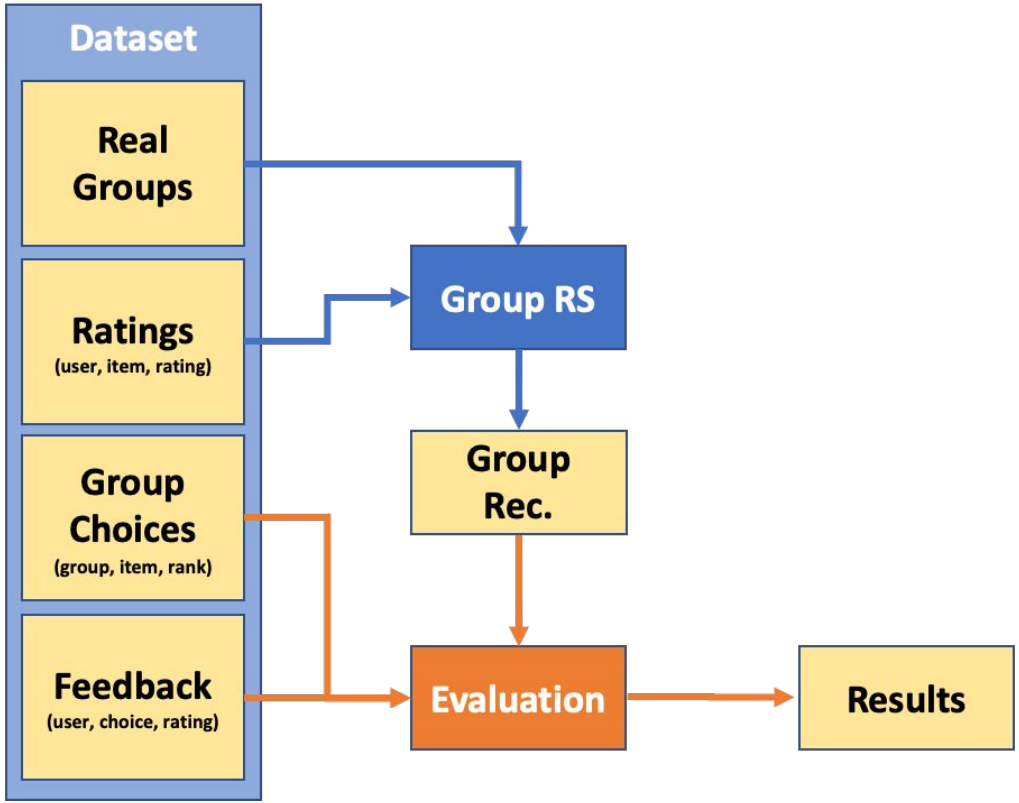




Use case 2: Tourism dataset (descriptives)

Group size	**2**	**3**	**4**	**5**	**SUM**
Rankings					
#Groups	5	13	21	8	**47**
#Participants	10	39	84	40	**173**
Ratings					
#Groups	5	5	13	1	**24**
#Participants	10	15	52	5	**82**

Evaluation pipeline



A quick look at the code

https://github.com/barnap/group-recommenders-offline-evaluation

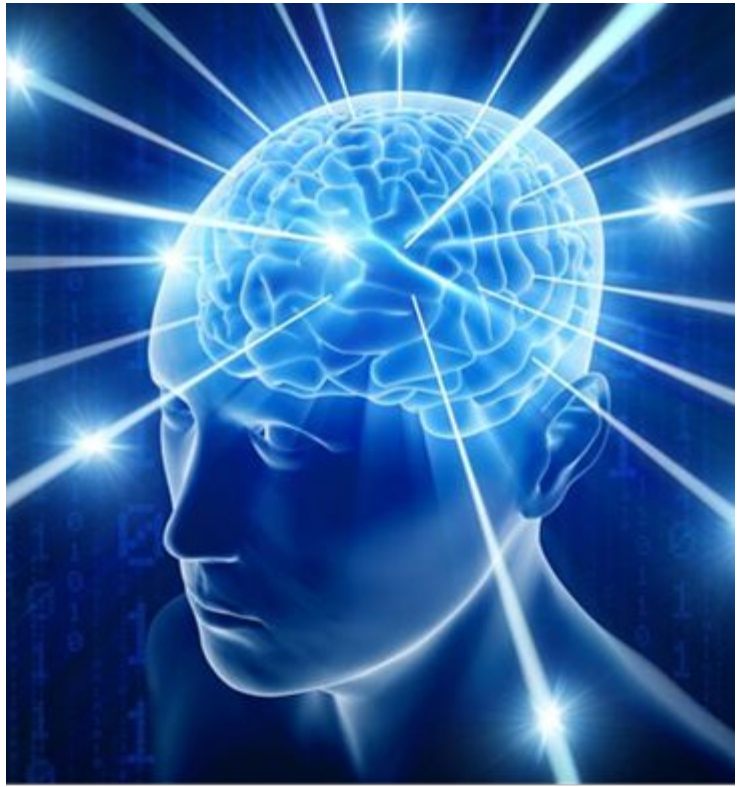
Brainstorming and hands on session

- Divide in groups
- Work on setting up the evaluation
  - How do we generate group recommendations?
  - Which metrics do we evaluate?
- Try to generate results with the provided data
- After we will analyze some of the obtained results

Conclusion

Conclusion

- Difficulty of Offline Evaluation
  - Available data impacts what we can evaluate
  - Relate your results to the evaluation choices
- Improve reproducibility
  - Report and motivate your choices
  - Lack of reproducibility works in the field
- Share collected datasets



Thanks for your attention!

For any questions feel free to contact us:

- Barile Francesco
  f.barile@maastrichtuniversity.nl
- Amra Delic
  adelic@etf.unsa.ba
- Ladislav Peska
  Ladislav.Peska@matfyz.cuni.cz

Tutorial on Offline Evaluation for Group Recommender Systems

The ACM Conference Series on Recommender Systems