Trustworthy Recommender Systems for Social Good

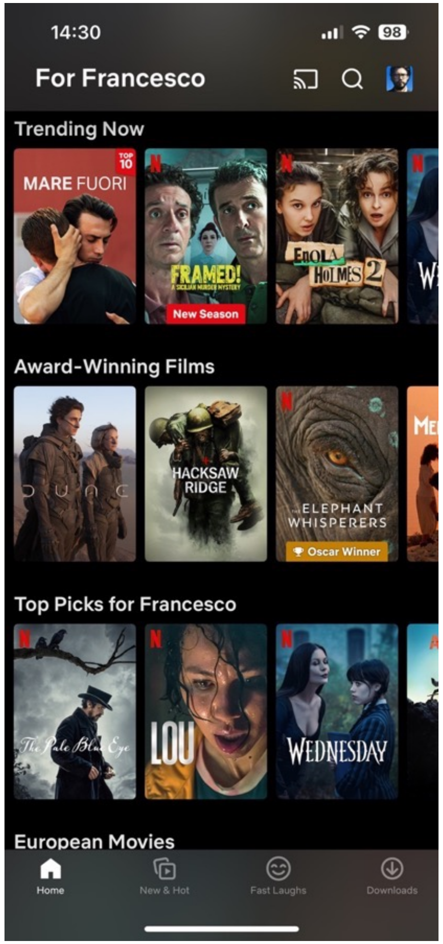
Francesco Ricci
Senior professor - Competence Center on Sustainability
Free University of Bozen-Bolzano, Italy
fmr959@gmail.com

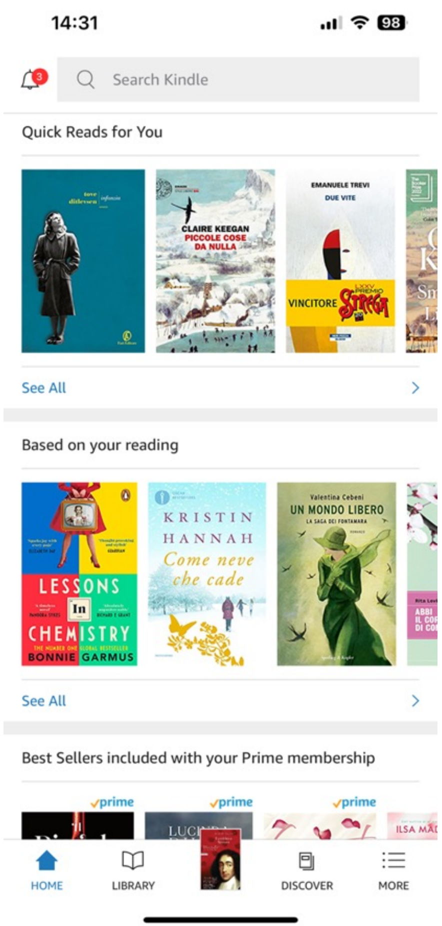
4th ACM Europe Digital Humanism Summer School, September 8th to 12th, at TU Wien, Austria.

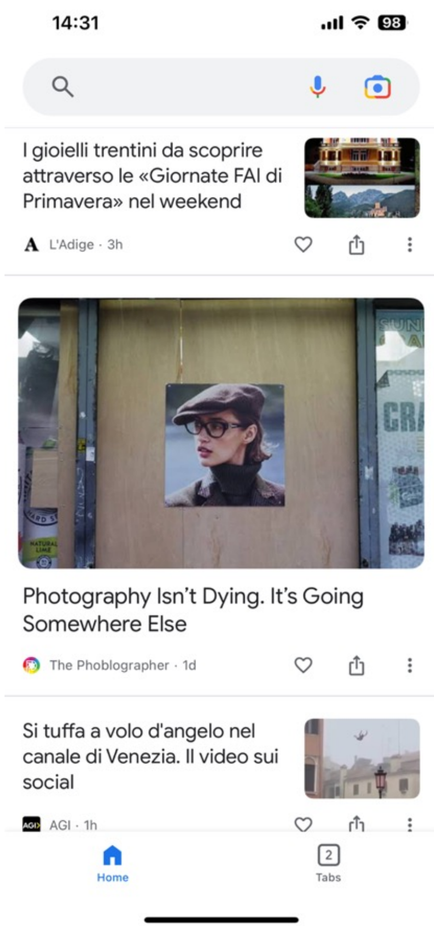
Content

- Recommender systems
- How do they work and how do they influence us
- Trustworthy recommender systems
- European legislation: Digital Service Act
- Evaluating recommender systems
- Model based simulation of user/system interaction
- Case study: sustainable tourism recommendations
- Behaviour change for social good
  - By increasing user awareness of sustainable choices
  - By increasing recommendation salience of sustainable choices
- Positive sum effect for tourists and local community

Recommender Systems







Recommender Systems

- *Provide suggestions for items that are likely to be of interest to a particular user (decision support)*
- Operational Steps:
    1. Preference elicitation (behavioural data collection)
    2. <u>Preference prediction</u>
    3. Selection of recommendations

How do they work?

- They solve a (discriminative) **prediction** problem: will the item ***j*** be interesting for user ***u***?
- Make use of formal **heuristics**:
  - If ***j*** is liked by users similar to ***u*** then it will be liked by ***u***
  - If ***u*** likes items similar to ***j*** then she will like ***j***
  - If the matrix ***R*** can be **aproximated** by two lower F-rank matrices (***P*** and ***Q***), describing users and items, then this decomposition can predict the value of unknown user/item interactions.

Users

up	?	down	up
down	up	?	up
down	down	up	up
up	?	down	?
?	down	up	down

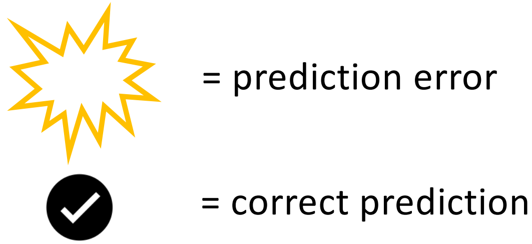
= *R*

Items

$$R \approx P\,Q^T$$

$$R_{uj} \approx \sum_{f=1}^{F} P_{uf} Q_{jf}$$

Testing for Accuracy



Training set

Users

up	?		up
down		?	up
down		up	
up	**?**		?
?	down	up	

Items

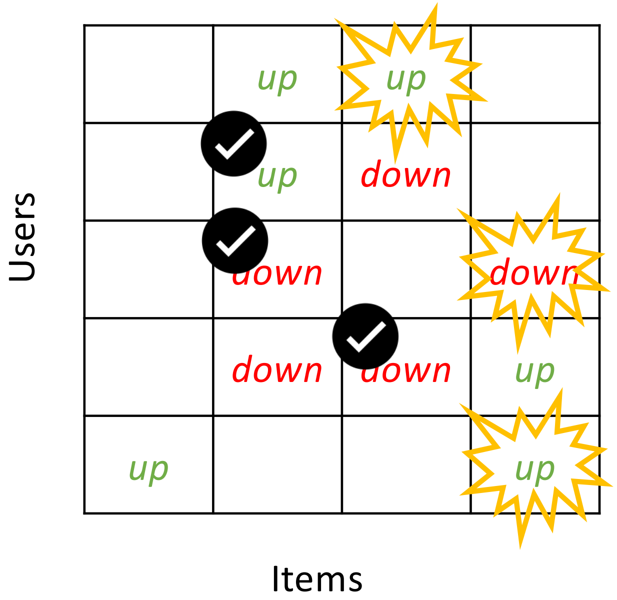
Test set

Users

	?	down	
	up	?	
	down		up
	?	down	?
?			down

Items

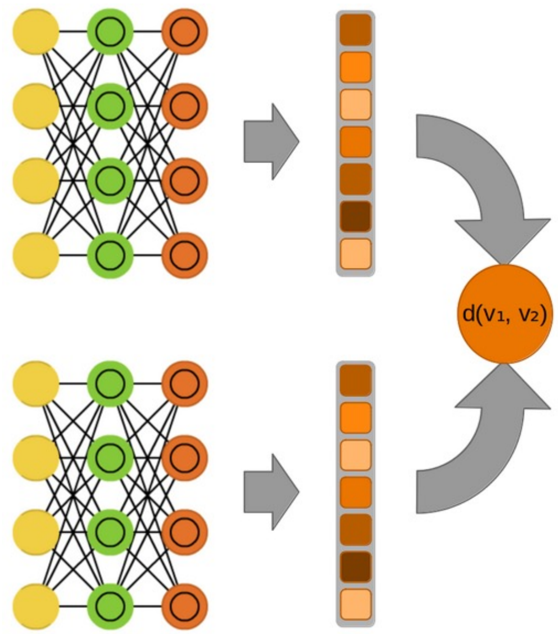
Prediction set



- Error can be estimated only on interaction data present in the test set

Encoder-only LLM Recommendation

- Each item's text content (e.g., title, description, and/or reviews) is treated as a document
- If no explicit query is available from the user, synthesize a query by concatenating the descriptions of a user's recently liked items
- **Dense retrievers** (e.g., BERT) produce a ranked list of documents given a query by evaluating the similarity (e.g., cosine similarity) between the encoder-only LLM document embedding and the query embedding.



[Deldjoo et al., Recommendation with generative models, arXiv:2409.15173, 2024]

Gemini for Travel Planning

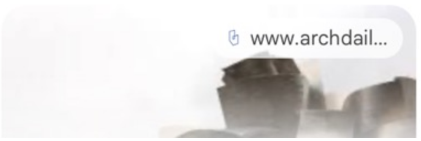
I will be in Bilbao for a couple of days in June. I like photography, and music. What do you think are the must see attractions? Can you schedule their visits in a two day time span?

Show drafts

Bilbao is a vibrant city with a rich history and culture, and it sounds like you've picked the perfect destination for your interests in photography and music! Here's a suggested itinerary for your two-day trip:

Day 1:

- **Morning:**Start your day with a visit to the Guggenheim Museum Bilbao. This world-renowned museum is a masterpiece of modern architecture in itself, and its collection of modern and contemporary art is sure to impress. Allow yourself a couple of hours to explore the exhibits and take photos of the museum's unique exterior.



https://create.microsoft.com/en-us/learn/articles/how-to-use-chatgpt-for-vacation-planning

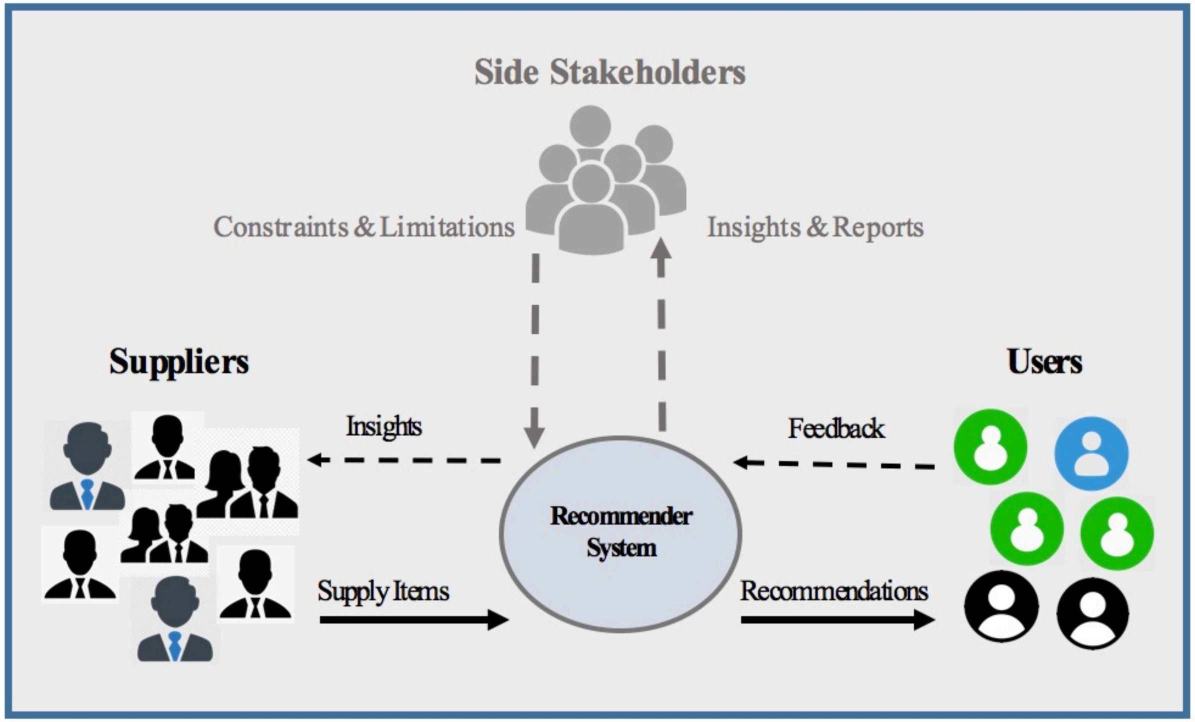
https://www.forbes.com/advisor/credit-cards/travel-rewards/chatgpt-for-travel-planning/

Recipe:
- *Start by providing some context*
- *Ask for what you want*
- *Be specific*
- *Refine your answer*

- Speedup travel planning
- The results look appealing (too much?)
- Possible wrong information
- The results may largely vary depending on the destination
- Miss temporal events (LLM is not frequently updated).

[Volchek, K. and Ivanov, S., ChatGPT as a Travel Itinerary Planner. In ENTER e-Tourism Conference (pp. 365-370). 2024]

Multistakeholder Systems



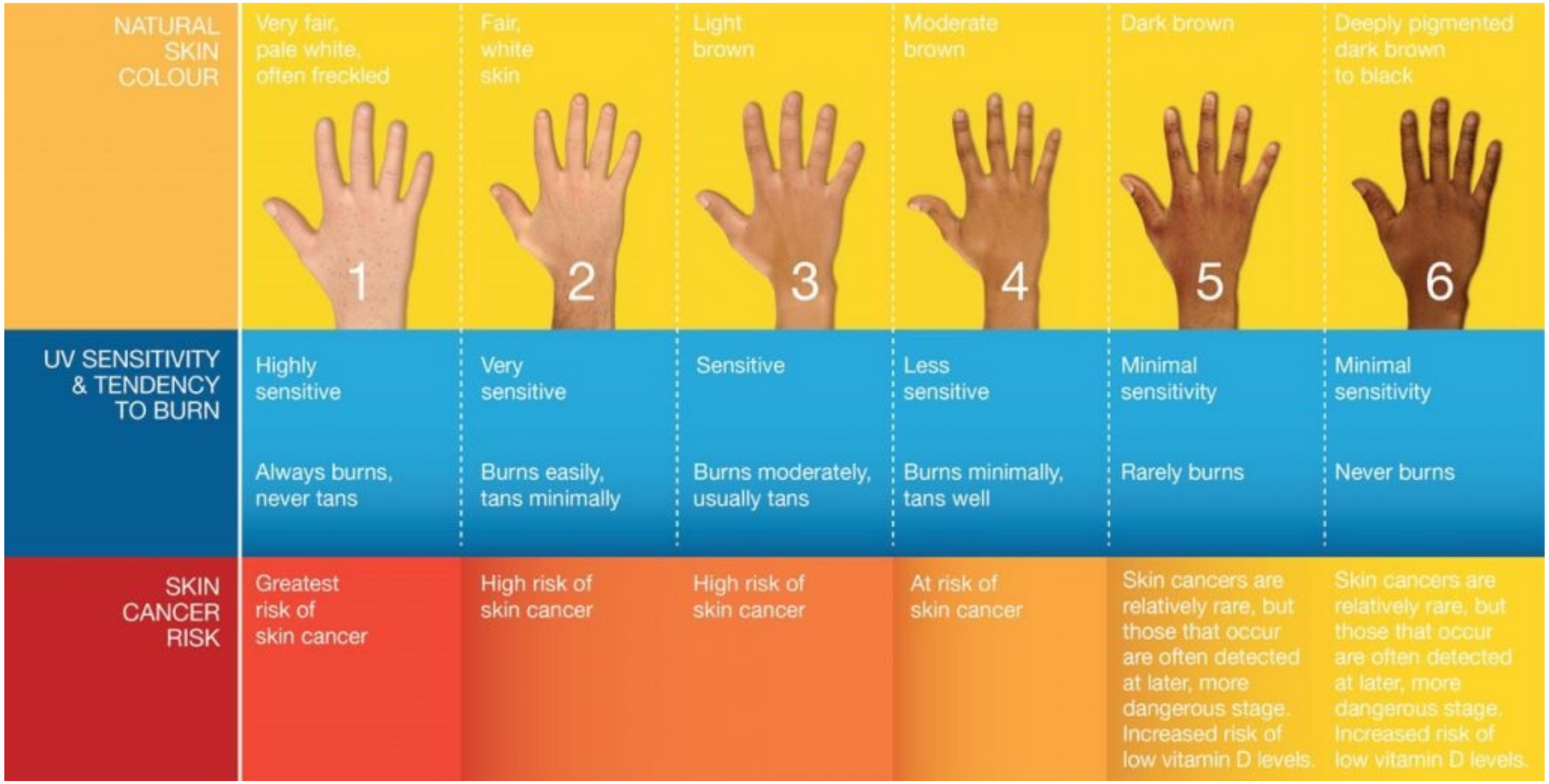
[H. Abdollahpouri & R. Burke, Multistakeholder Recommender Systems, in Recommender Systems Handbook, 647-677, 2022]

Influencing

- User preferences are constructed while consuming the recommendations
- The RS may have a hidden agenda
- Influencing is easy [Adomavicius et al. 2013]
- But deliberately influencing users to change their preferences or behaviours is not easy
    - *For instance, exposing users to diversity does not produce choice diversity* [Helberger et al. 2018]

[G. Adomavicius et al., Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. Inf. Syst. Res. 24, 4, 956–975, 2013]
[N. Helberger et al., Exposure diversity as a design principle for recommender systems. Information, Communication & Society 21, 191–207, 2018]

Exposure vs Effect (Choice)



Effects on Humans

- If my friends are recommended to eat pizza, they will tend to eat pizza and I will be recommended to eat pizza as well - **delayed effects**
- We are becoming more and more predictable since we are forced to be predictable
- Novelty and creativity could decrease
- The most reliable and effective models can only be built by the big Internet players – they know how we react to their recommendations
- Recognition of the importance of the experts is decreasing – there are no more experts.

Trustworthy Recommender Systems

- *Trustworthy Recommender Systems (TRSs) aim to **better serve users** in the complex and challenging cyberspace*
- ***Stakeholders**, including users, owners and regulators of RSs not only demand recommendation accuracy but also need trustworthiness, including robustness, fairness, explainability and privacy preservation*
- ***Trustworthiness** is even more important than **accuracy** in some **critical and sensitive domains**, including finance and medicine, where highly reliable RSs are required.*

[S. Wang, X. Zhang, Y. Wang, F. Ricci: Trustworthy Recommender Systems. ACM Trans. Intell. Syst. Technol. 15(4): 84:1-84:20 (2024)]

Digital Service Act

- Article 34 (Risk assessment): *when conducting **risk assessments**, providers ... shall take into account ... : (a) the design of their recommender systems and any other relevant algorithmic system*
- Article 37 (Independent audit): *providers ... shall be subject, at their own expense and at least once a year, to independent **audits** to assess compliance with ...*
- Article 38 (Recommender systems): *... providers ... that use recommender systems shall provide at least one option for each of their recommender systems which is **not based on profiling***
- Article 40 (Data access and scrutiny): *providers ... shall, at the request ... **explain** the design, the logic, the functioning and the testing of their algorithmic systems, including their recommender systems.*


QUALITY
IMPROVEMENT
CRITERIA
EVALUATION
PERFORMANCE
LOADING...
ANALYSIS
RESULTS
ASSESSMENT
AUDIT

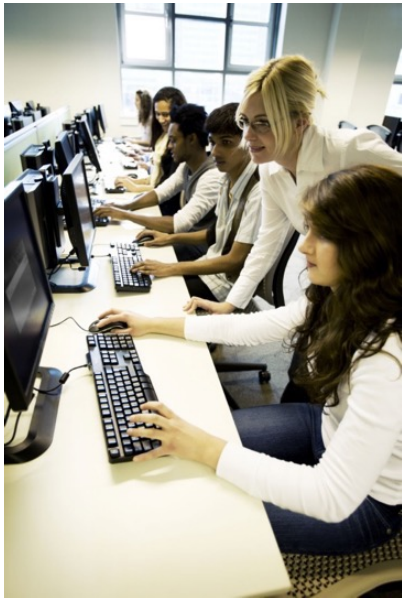
Measuring Performance

- User preference
- Prediction accuracy
- Coverage
- **Confidence**
- **Trust**
- Novelty
- Serendipity
- Diversity
- **Fairness**
- **Risk**
- **Robustness**
- **Privacy**
- Adaptivity
- Scalability
- **Explainability**
- **Transparency**

- *There are trade-off*
- *The effect of these properties (quantitative) on user experience may be unclear*
- *The effect may depend on the user*
- *They must be tuned to the specific user/application.*

User Studies

- We flight (online) only reasonable systems – *"is a candidate good enough?"*
- Collect a group of test subjects and ask them!
- **User study**
  - A group of (paid) volunteers **tries** one or many **alternatives**
  - Experiment can be done in a **controlled environment**
  - Subjects can **answer** questions about the test, **before** and **after**
  - Subjects can provide **additional** relevant **information** that is not available for real users (e.g., marital status).



Offline Testing

- Problem: offline testing must **mimic** recommendation setting and user’s behaviors as closely as possible
- Typical scenario in machine learning (e.g., classification)
  - ***Train offline:*** *on plates correctly labelled by humans*
  - ***Test offline:*** *on some plates not considered in training*
  - ***Offline test plates images*** *are distributed exactly as those recorded by the deployed camera*
- Offline train/test is **not** valid for rating prediction in RSs:
  - *RS generates many recommendations for a user and we should assess the goodness of all them*
  - But we know only the quality of the recommendations that **the user rated and are in the test set**.

Offline Testing Issues

- **RecSys affects user behavior**
- Hence user ratings or behavior data **collected in the past** - with a particular RS strategy - may not be a safe “ground truth” for designing a **new “intervention” RS**
- In practice all the **data sets** (user choices/rating) that we use have been acquired while the users were exposed to **unknown treatments.**

Can you estimate the effect of **novel drug** by using the historical data of subjects **possibly** treated with some **unknown drug**?

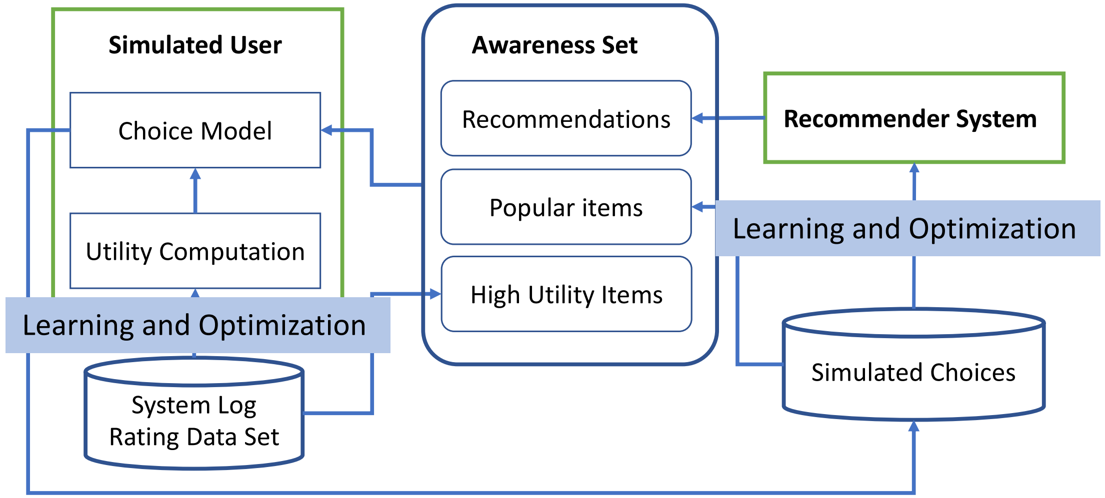
Model-based Simulations

- Learn a **model** that capture the **behavior of users** (choice model) while being exposed to recommendations
- Use this model (not the test data) to simulate user feedback while testing a novel RS
- **Pros:**
  - We can run as many interactions as needed with a novel candidate RS

- **Cons:**
  - We must be able to guarantee that the choice model truly captures users' behaviors (*better than assuming that the items the user chose in the past will be chosen again*).

Simulation of Choices

- **Simulated users** estimate items' utility (learned preferences)
- Users are **aware** of:
  a) recommended items,
  b) popular items, and c) items that match their preferences



- They **choose** with a probabilistic model: *the larger the item utility is, the larger the choice probability is*
- RecSys is supposed to **increase the user's estimated utility** of the recommended items.

[N. Hazrati, F. Ricci: Recommender systems effect on the evolution of users' choices distribution. Inf. Process. Manag. 59(1): 102766 (2022)]

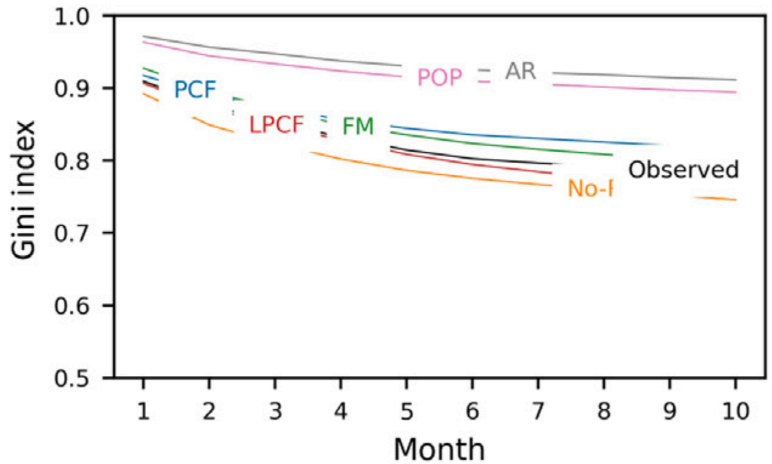
Choice Model (softMax)

$$p(u\ chooses\ j) = \frac{e^{v_{uj}}}{\sum_{k \in A_u} e^{v_{uk}}}$$

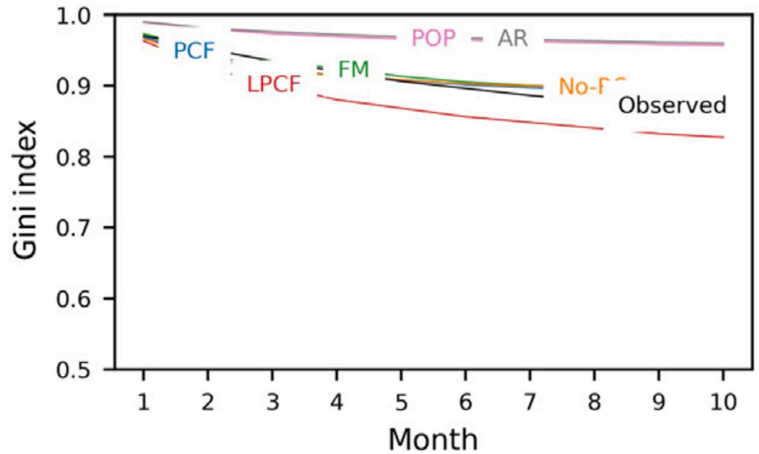
- A_u is the **Awareness set**: items that the user may consider to choose because are **known**
- v_{uk} is the **estimated expected utitility** of the item *k* for the user *u* (computed with a predictive model)
- When an item *k* is recommended the **user updates the expected utility** of *the item* (salience):

  $v_{uk} := 2 * v_{uk}$
- We can simulate the interaction of the user with **any item** and artificially simulate the **salience effect of the RS.**

Simulation Results



(a) Apps data set



(b) Games data set

- Gini is a measure of diversity of the choices: *larger Gini means lower diversity*
- AR = average rating RS; POP = most popular RS; FM = Matrix Factorization RS; PCF = most frequent in the neighbors RS; No-R = no recommendations; LPCF = most frequent in the neighbors but also non popular.

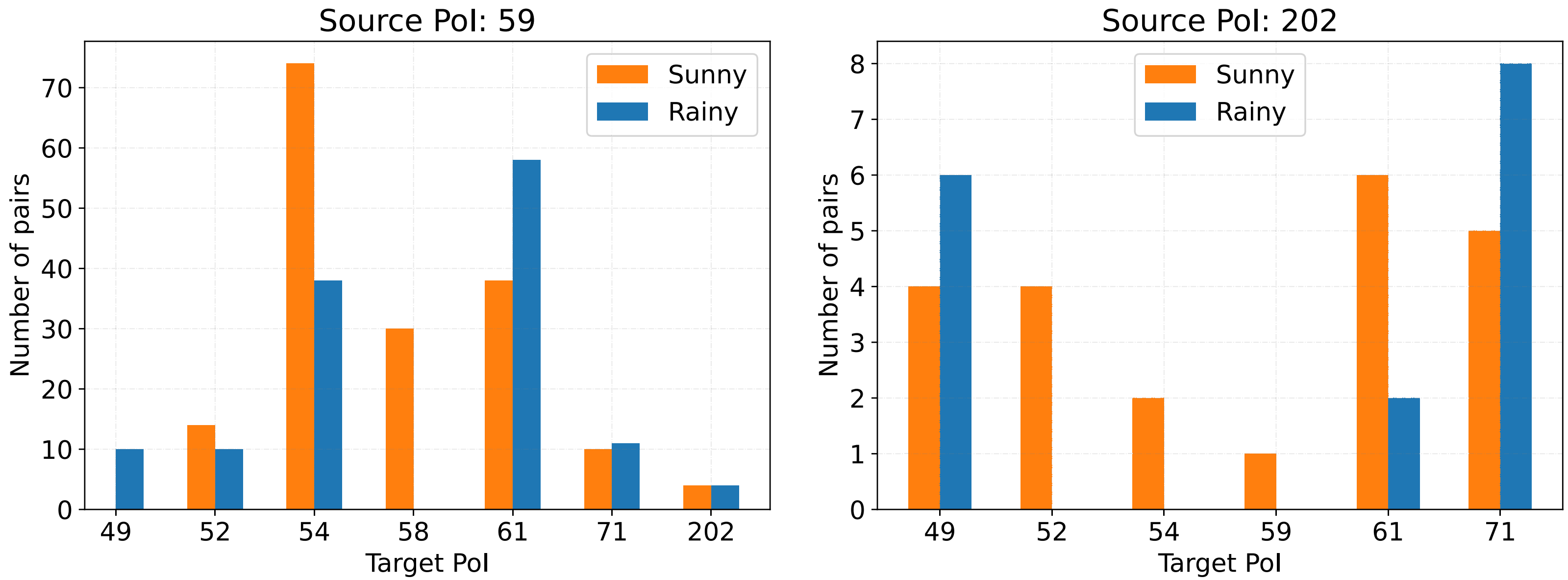
Overtourism

Application Example: Overtourism

- **Overcrowding** from an excess of tourists, resulting in **conflicts with locals**
- **Destination managers** try to prevent popular/central POIs from being overly crowded – tourists want to visit them
- Approaches
  - Hard rules (close an area)
  - Time rules
  - Entrance fees
  - **Multicriteria sustainable recommendations?**

Venezia, 19/2/2025 - Carnaval

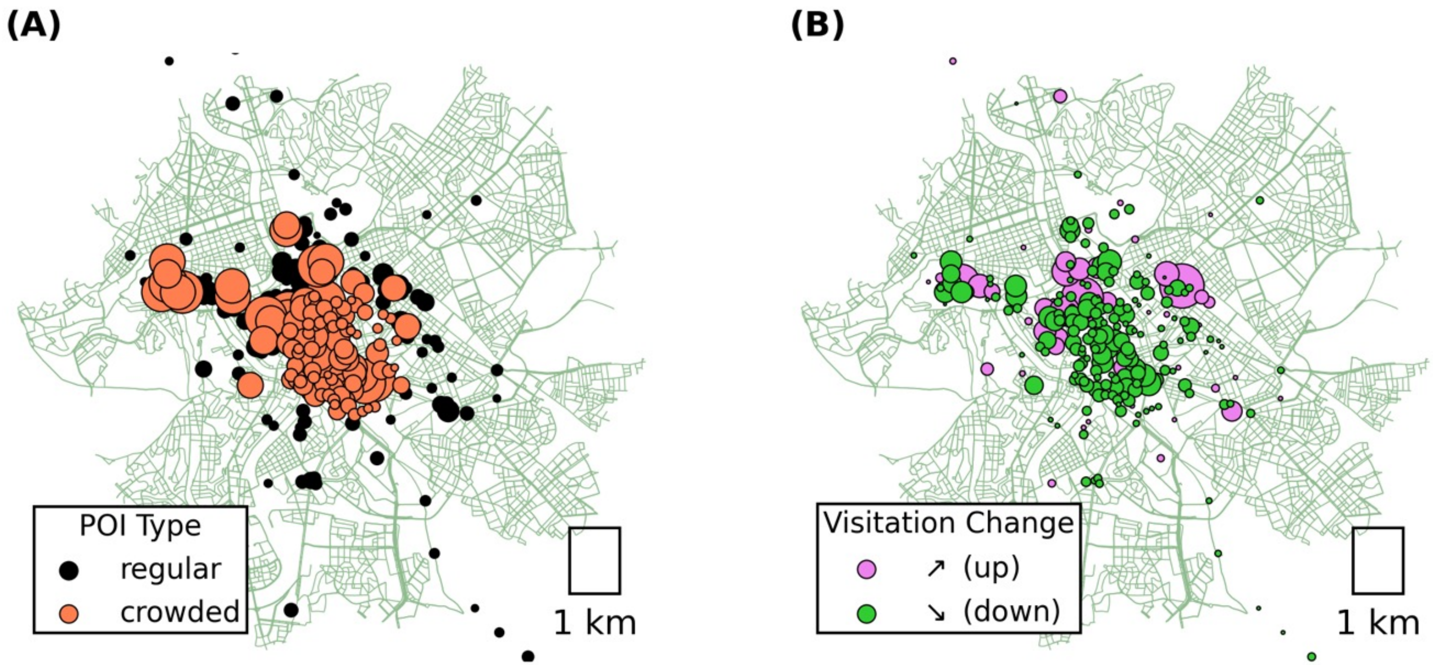
Popularity Bias



Most of the RSs will recommend to a tourist in Verona at point of interest 59 to move to POI 54 if sunny and to POI 61 if rainy

[S. Migliorini, A. Dalla Vecchia, A. Belussi, and E. Quintarelli: ARTEMIS: a Context-Aware Recommendation System with Crowding Forecaster for the Touristic Domain, Inf. Systems Frontiers, 2024]

Taming Overtourism



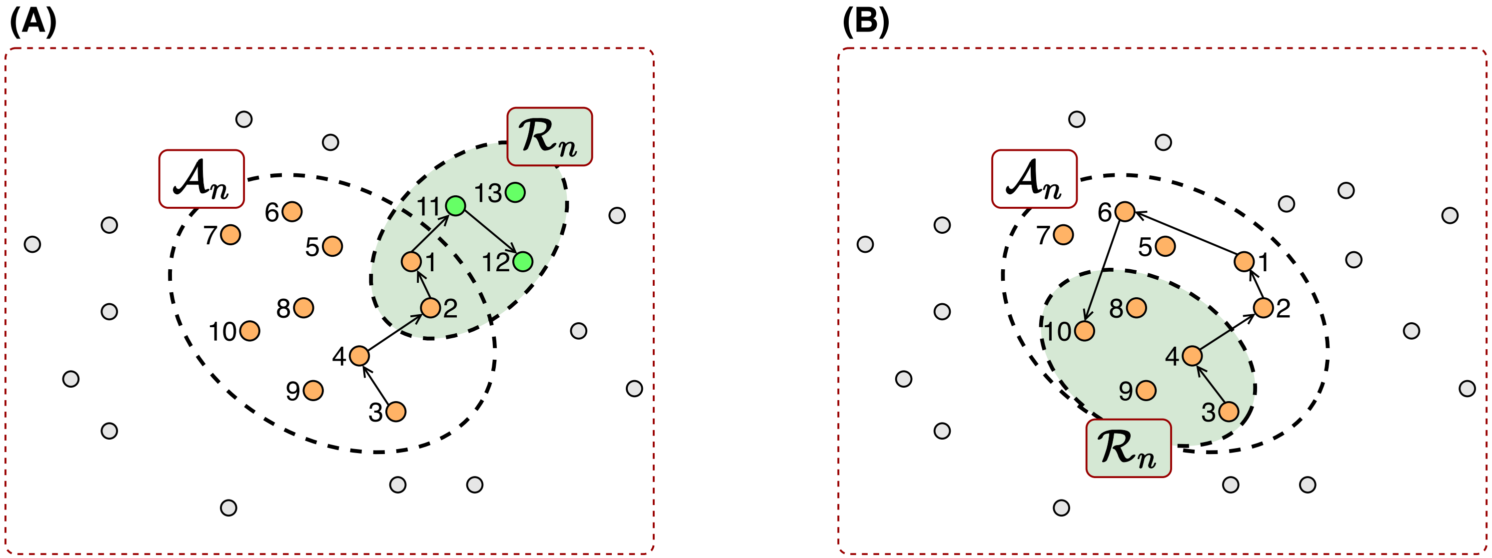
- Applying a proper recommendation policy can reduce the number of tourists visiting the most crowded points of interest (Rome)

[P. Merinov, F. Ricci: Positive-Sum Impact of Multistakeholder Recommender Systems for Urban Tourism Promotion and User Utility. RecSys 2024]

Positive Sum Impact

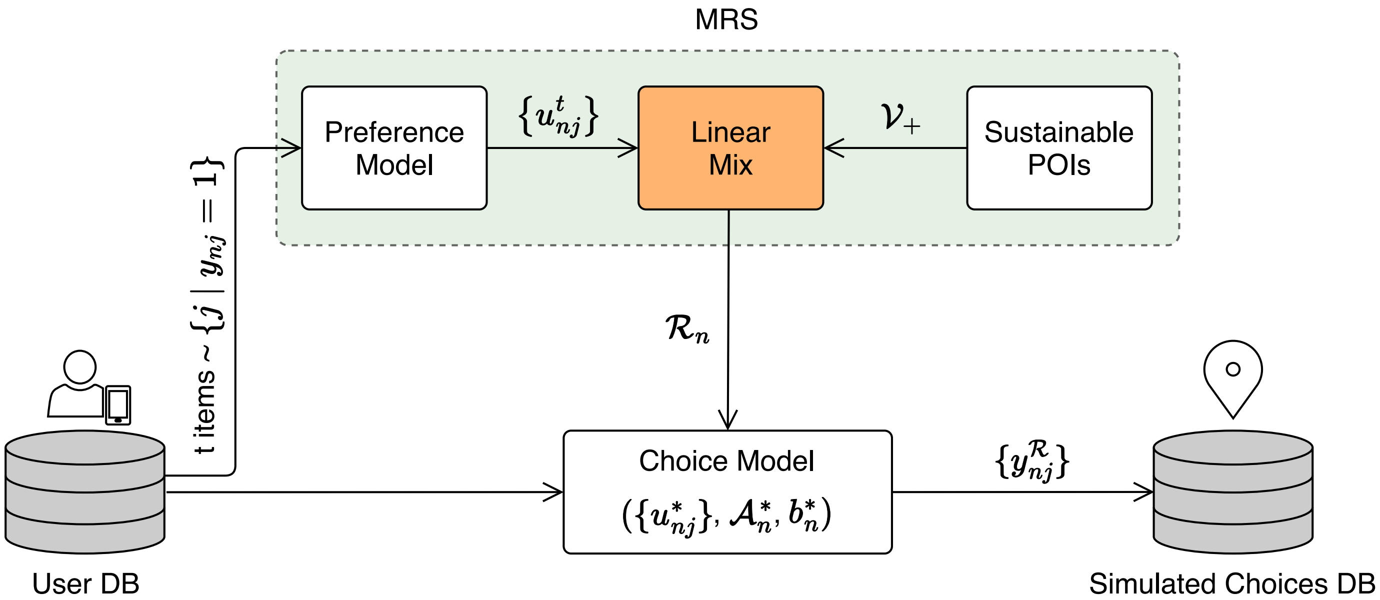
- *Can an RS identify not crowded POIs that the user does not yet know and will like?*
- Such recommendations may benefit both stakeholders: tourist and destination
- Ingredients:
  - Predict the (**limited) user knowledge of the catalogue** (awareness) and how tourists make their choices
  - Build an optimized and **reliable** (trustful) data-driven **choice simulation protocol**
  - Simulate the impact of the RS policy for alternative configuration parameters.

Awareness and Recommendations



- When recommendations for user n (R_n) includes POIs that the user does not know (i.e., not in A_n) there is a possibility that the behaviour departs from the organic one.

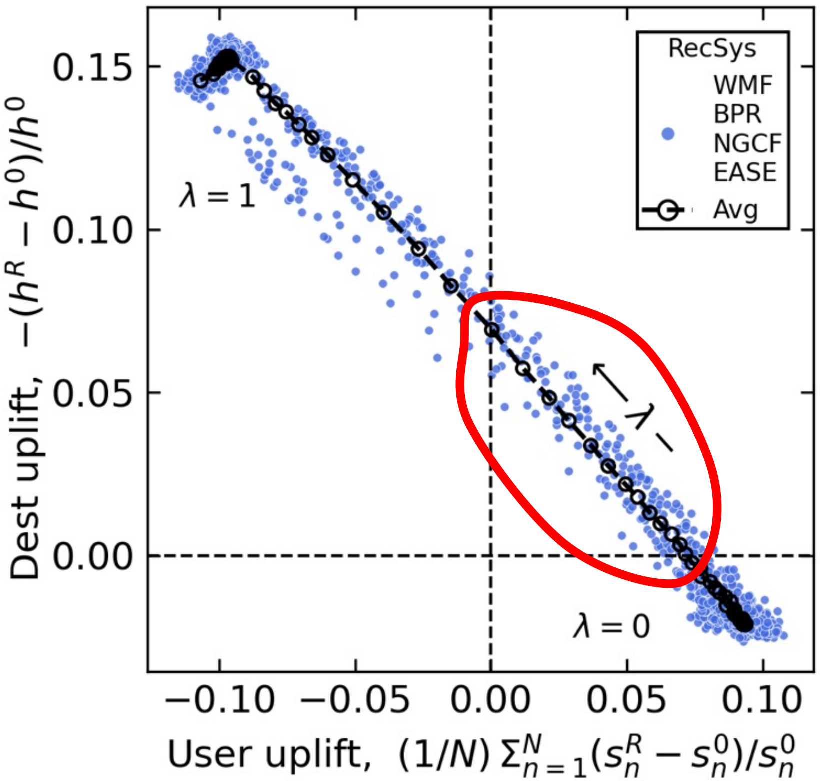
Simulating tourist interactions with an MRS



- *B* recommendations ($x_k = 1$) maximise the linear combination of estimated user utility (u^t_{nj}) and destination utility: v_j is positive if the POI *j* is not overcrowded (0 elsewhere).

$$\mathcal{R}_n(t, \lambda) = \underset{(x_1,\ldots,x_J):\sum x_i = B}{\arg\max} \sum_j x_j \left((1 - \lambda)u^t_{nj} + \lambda v_j\right)$$

Positive Sum Impact



- **User and destination uplift** measure the variation of the (user and destination) utilities with respect to the organic behaviour (no recommendations)
- Recommendations **do not receive** any expected-utility **increase (no salience)**
- **Points** represent the **performance of an MRS** built with a particular λ value and a particular algorithm (WMF, BPR, NGCF, and EASE) used to estimate user preferences
- There exist a range of λ values that produce a positive uplift for both stakeholders.

Behaviour shift

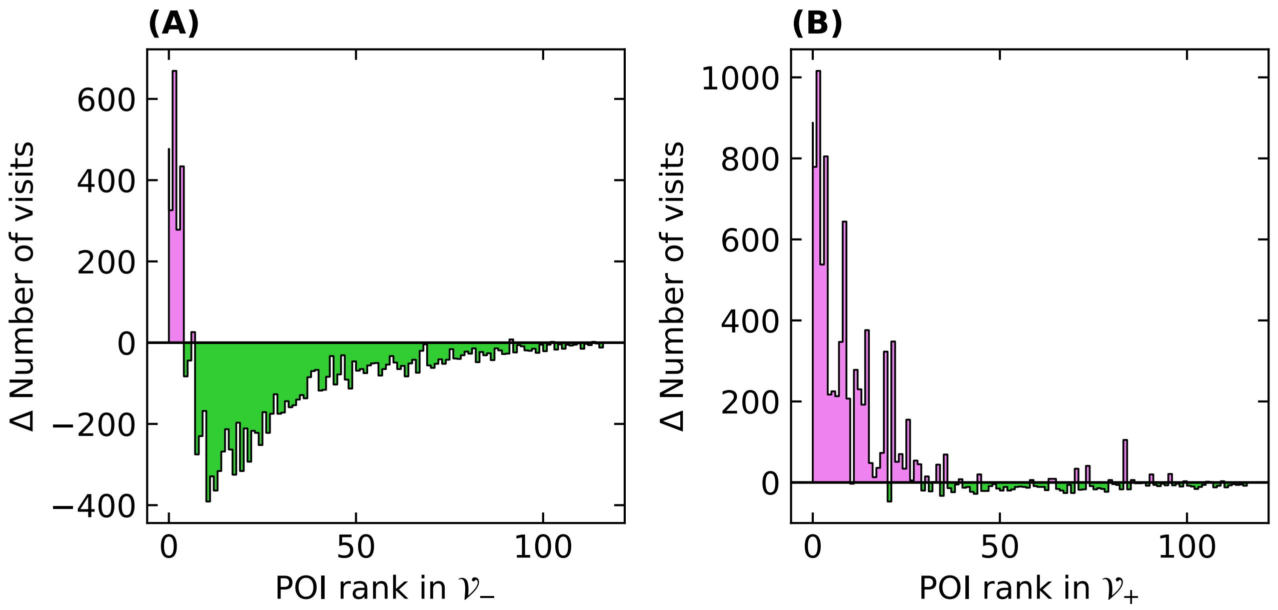


Figure 11: Recommendation policy in action. (A) Visits shift in $\mathcal{V}_-$ (POIs sorted by decreasing value of popularity) after applying the proposed MRS policy. Most critical central POIs exhibit reduced crowding (green bars), while a few popular POIs experience an increase in visits (violet bars). (B) Visits shift in $\mathcal{V}_+$ (POIs sorted by popularity) after applying the proposed MRS policy. Overall, the desired periphery promotion policy for the $\mathcal{V}_+$ POIs is attained with a notable 7% increase in $\mathcal{V}_+$ visits (hence, decrease in $\mathcal{V}_-$ visits).

The most popular POIs still see an increase in the number of visits since their user utitlity is high, they are known by tourists, and there are no alternative less crowded POIs that have larger utility

Salience effect

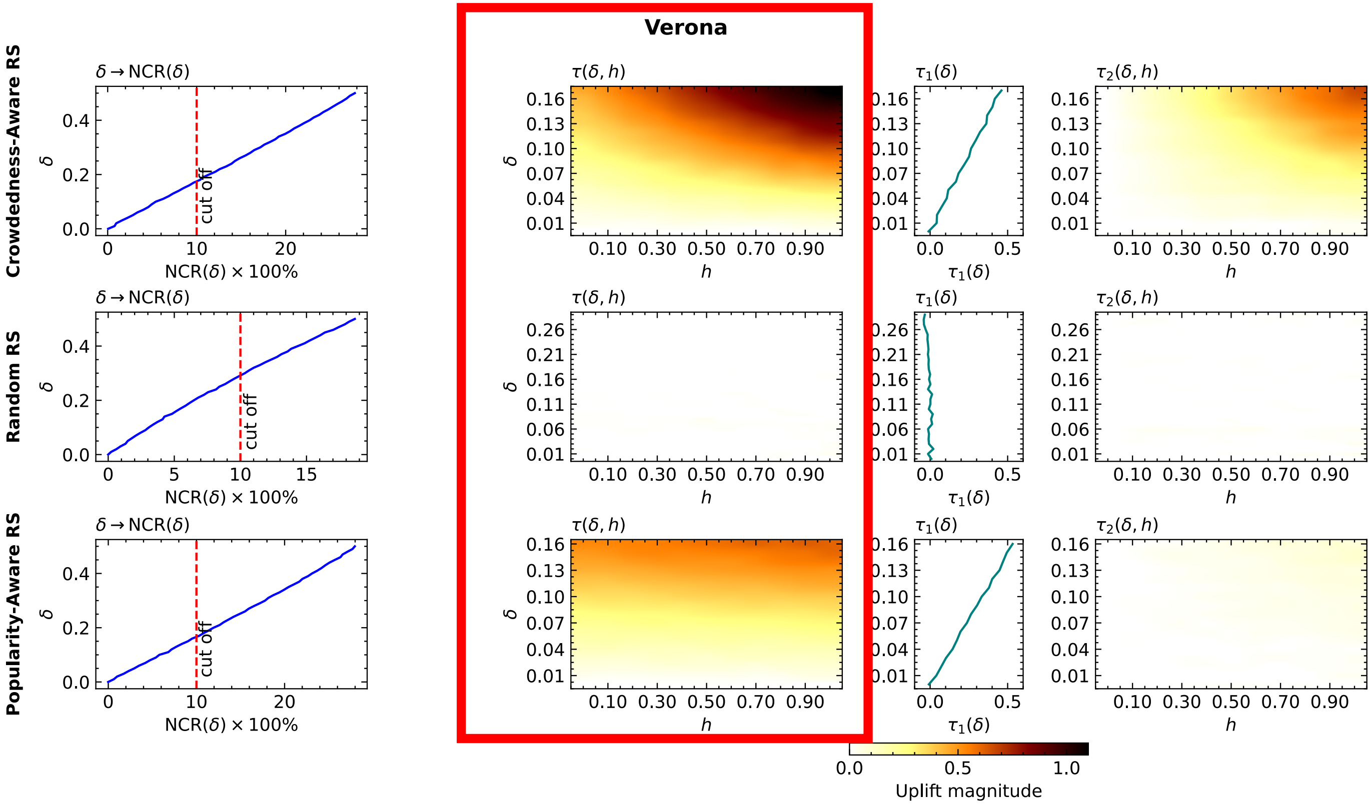
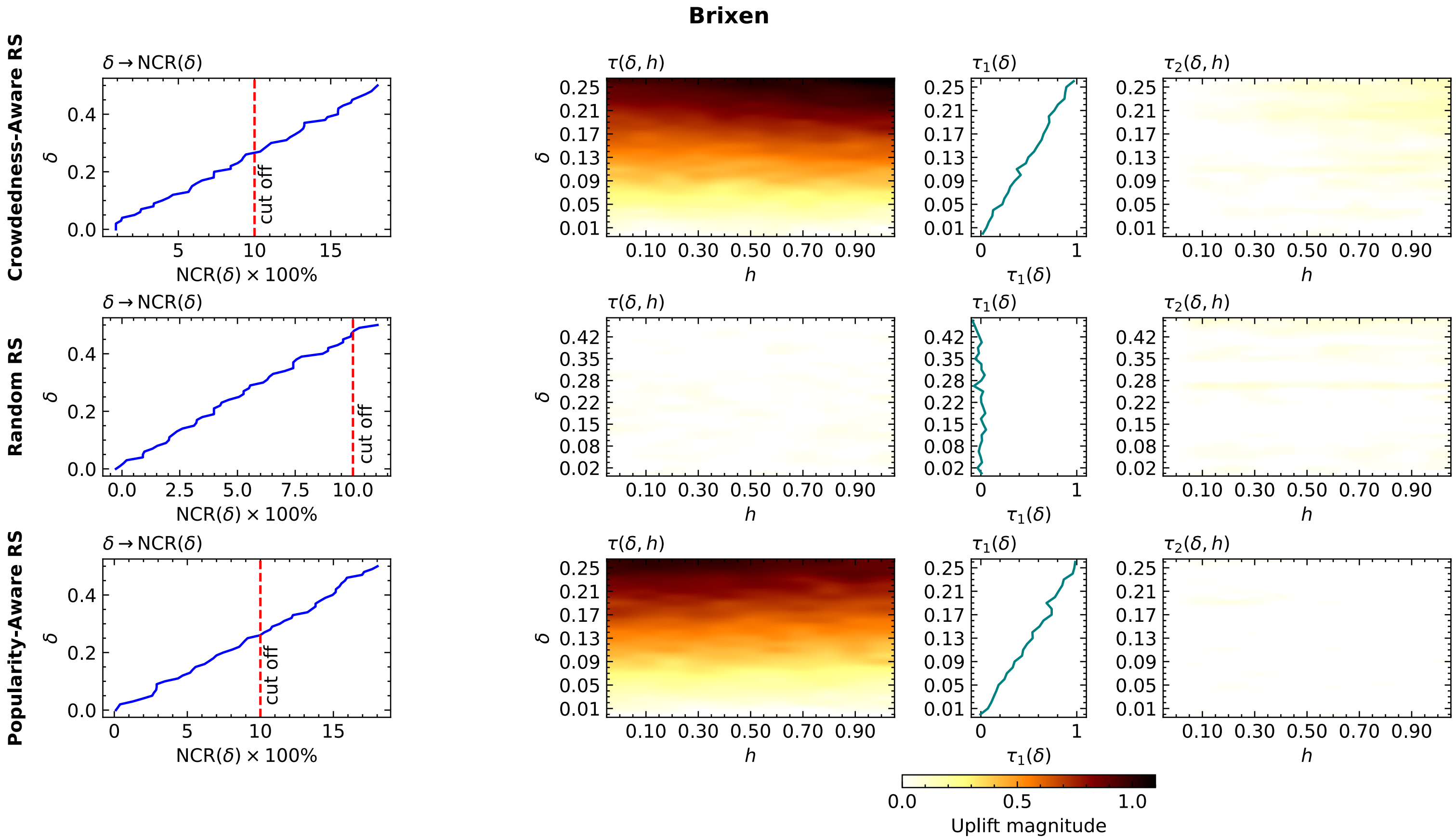


Fig. 8. VeronaCard simulated uplift surface. Salience δ and penalty h are free parameters. For each RS, the left subplot shows the simulated NCR to salience δ dependency. The dependence curves for small δ values (here range $[0, 0.5]$ is covered) are approximately linear. Right subplots report in a row: $\tau(\delta, h)$, $\tau_1(\delta)$, and $\tau_2(\delta, h)$ uplifts.

Recommended items are simulated to be evaluated by the user with δ increment in utility – increasing the probability of being chosen

[Merinov & Ricci, Shaping Sustainable Tourist Experience: Simulating the Impact of Recommendation Salience, to be published, ACM trans. Rec. Sys.]

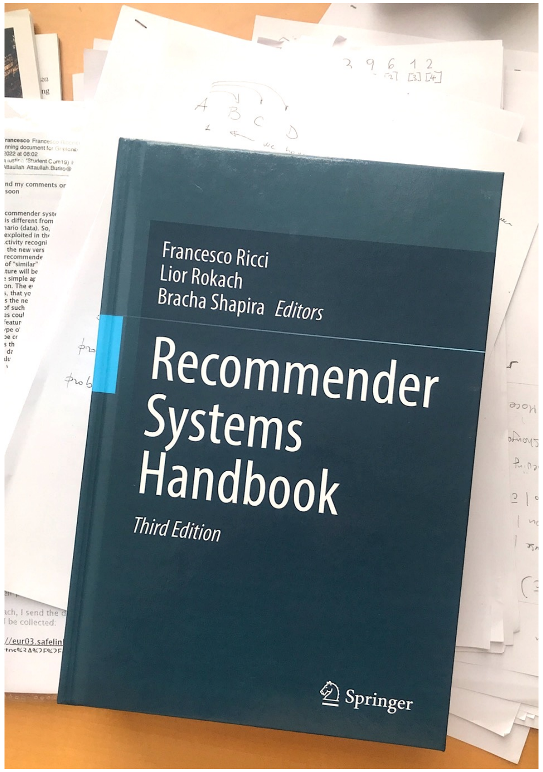
Salience effect



Merinov, P., Ricci, F. (2025). Simulating the Impact of Recommendation Salience on Tourists Experienced Utility. In: Recommender Systems for Sustainability and Social Good. RecSoGood 2024. Springer.

Is that reliable?

- The choice model is optimized to faithfully replicate the organic behaviour and minimise a specific bias (difference between utility obtained in real choices vs simulated choices)
- The **expected utility** estimated by the user at decision time may be influenced by a number of hard to control contextual parameters (e.g. GUI effect, decision biases as the attraction effect)
- It is hard to precisely know what the users of an RS really know about the items.


Francesco Ricci
Lior Rokach
Bracha Shapira Editors
Recommender Systems Handbook
Third Edition
Springer

Discussion