Trustworthy Recommender Systems for Multiple Stakeholders

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

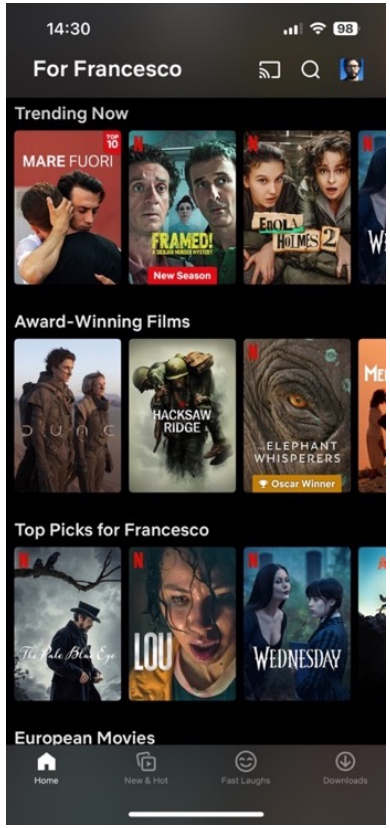
18th International Conference on Advanced Visual Interfaces, AVI 2026, Venice June 8-12

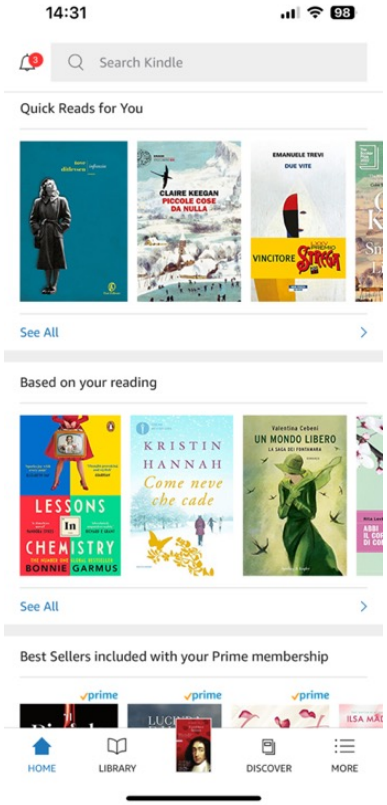
Content

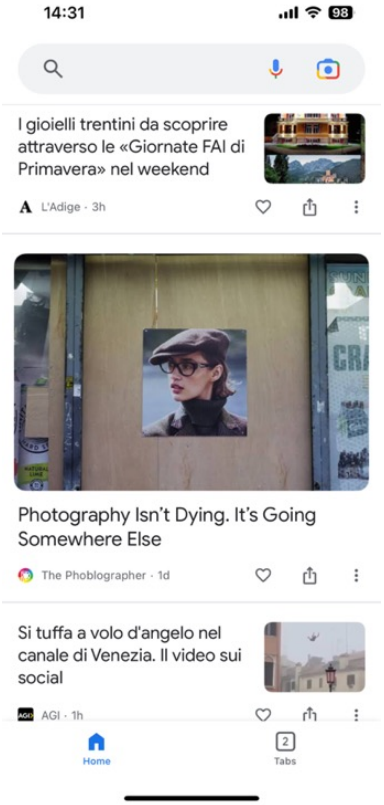
- Recommender systems: how do they **function** and **influence** users
- **Trustworthy** recommender systems
- **Off-line evaluation** of recommender systems
- Model based **simulation** and evaluation of user/RS interaction
- Case study: **sustainable tourism** recommendations
- Approach **validity** and future work

How to avoid crashing real people with real RSs

Recommender Systems



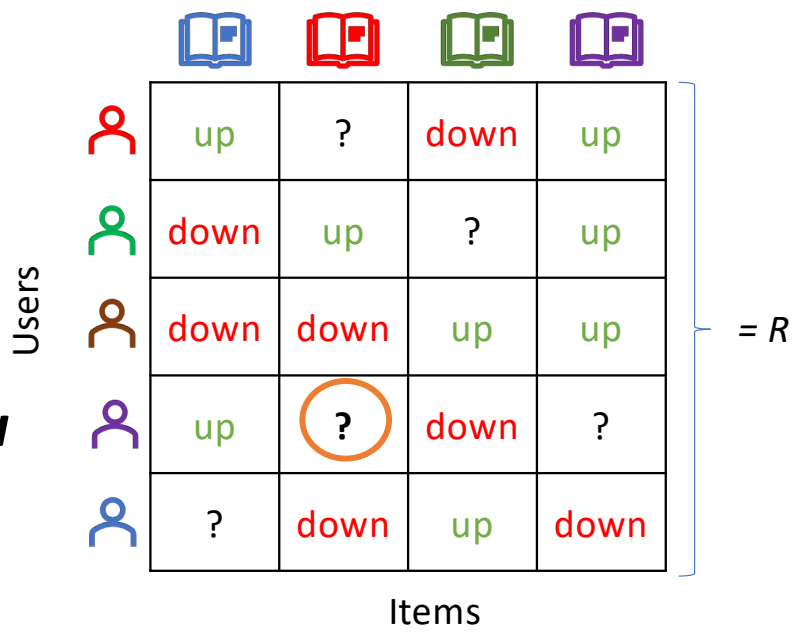




*One of the most successul example of **machine learning** application*

How do they work?

- They use **interaction** data to solve a (*discriminative*) **prediction** problem: will user ***u*** like item ***j***?
- Make use of formal **heuristics** (inductive bias):
  - If ***j*** is liked by users similar to ***u*** then ***u*** will like ***j***
  - If ***u*** likes items similar to ***j*** then ***u*** will like ***j***
  - User (item) representations are correlated, hence the matrix ***R*** could be **aproximated** and **densified** by two lower F-rank matrices (***P*** and ***Q***), describing users and items in a low dimensional space.

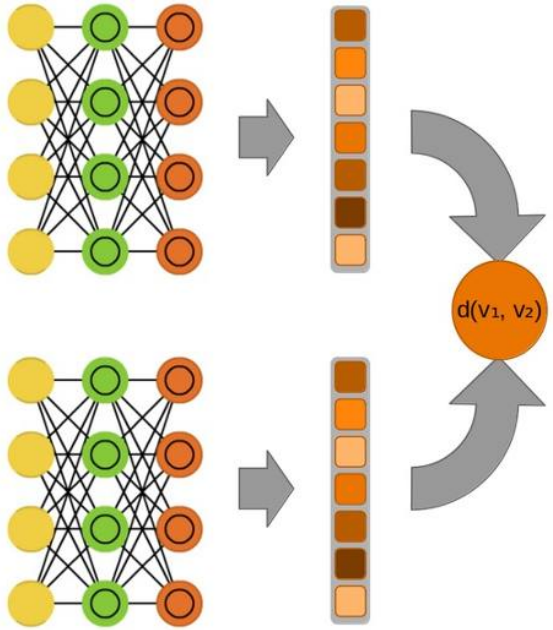


$$R \approx P\,Q^T$$

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

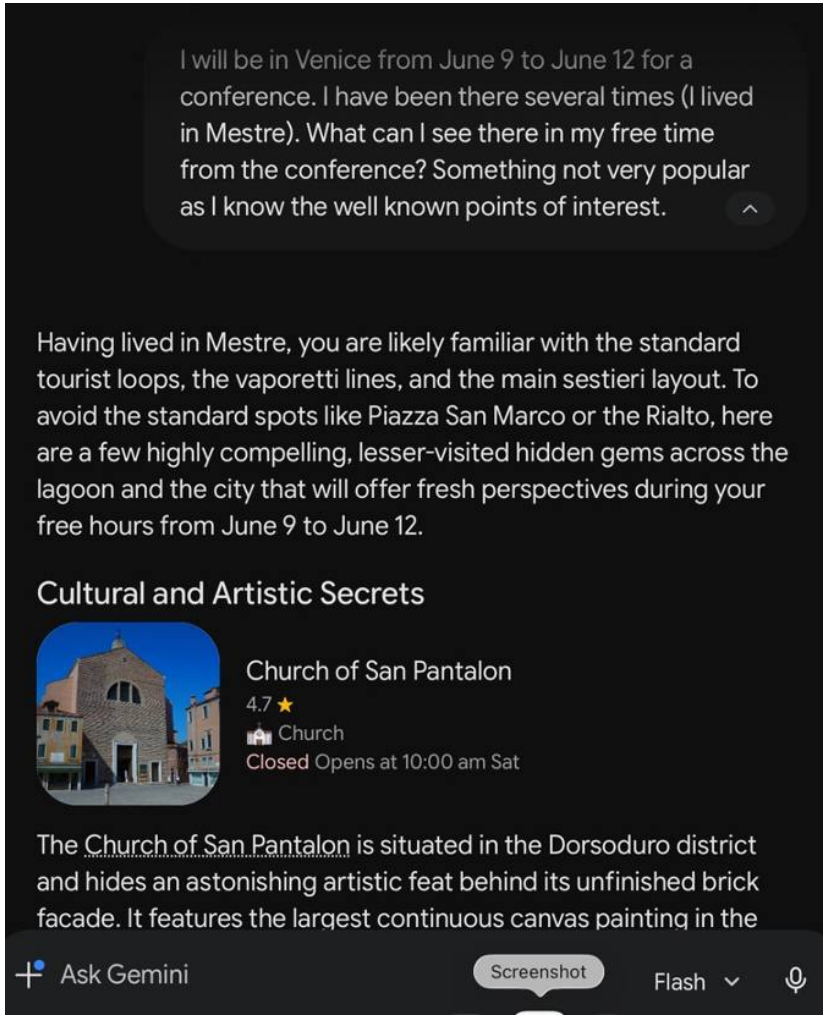
Encoder-only LLM Recommendation

- Each item's text content (e.g., title, description, and/or reviews) is treated as a **document**
- If no explicit query is entered by 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** between the encoder-only LLM **document** embedding and the **query** embedding.



[Deldjoo Y., et al., Recommendation with generative models. *Foundations and Trends in Information Retrieval*, Vol. 20 No. 1-2 pp. 1–176, 2026]

LLM for Travel Planning



- Speedup travel planning
- Result looks appealing (too much?)
- Often wrong information
- Results quality largely varies depending on the destination
- Miss recent 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]

GenAI and Tourism

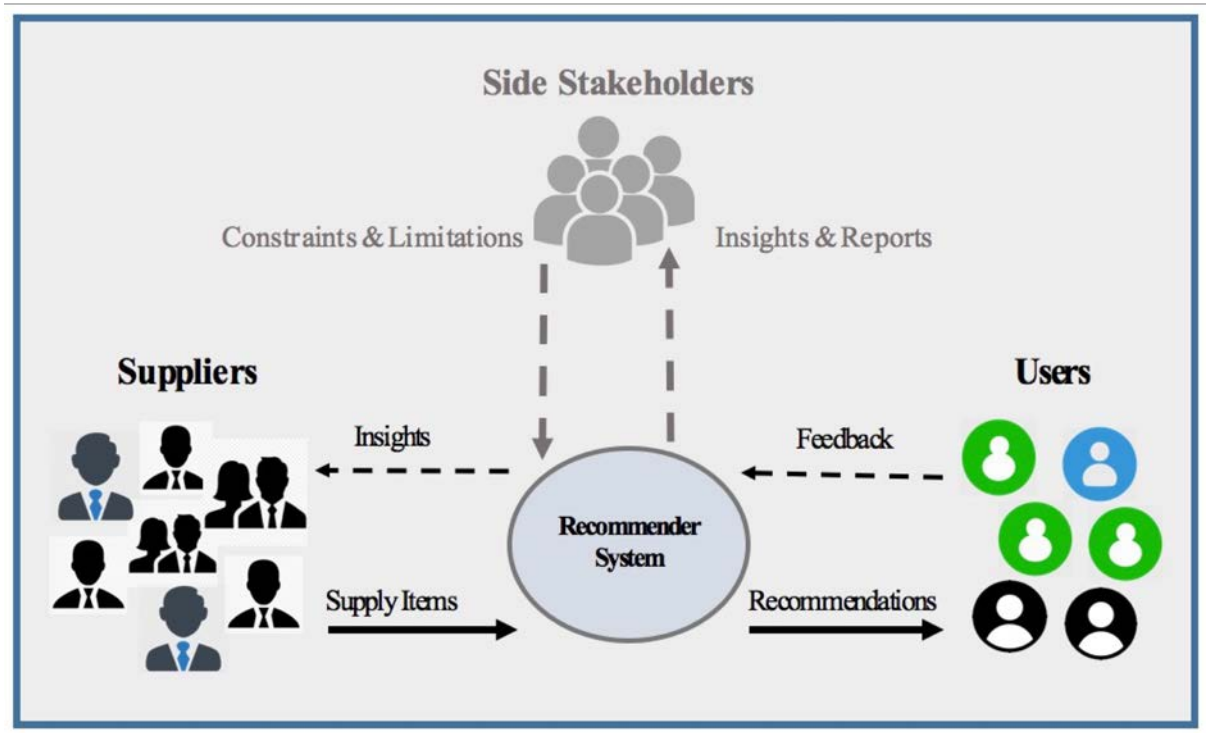
Booking.com

- Big market players, such as Booking.com, are evaluating the prospect of introducing GenAI
- 91% of the respondents to one of their survey say that they are **excited about AI** and 79% are **familiar with the technology**
- But, only 6% fully **trust AI** and the majority (91%) have at least one **concern about its implications**
- Only 12% of consumers are comfortable with AI **making decisions independently**, but still 89% of consumers **want to use AI** in future travel planning, with AI assistants (24%)
- AI is now considered a **more trusted** source than travel **bloggers** (19%) or social media **influencers** (14%).

https://news.booking.com/bookingcom-releases-the-global-ai-sentiment-report/

Multistakeholder Systems

Do we have a similar structure for GenAI?



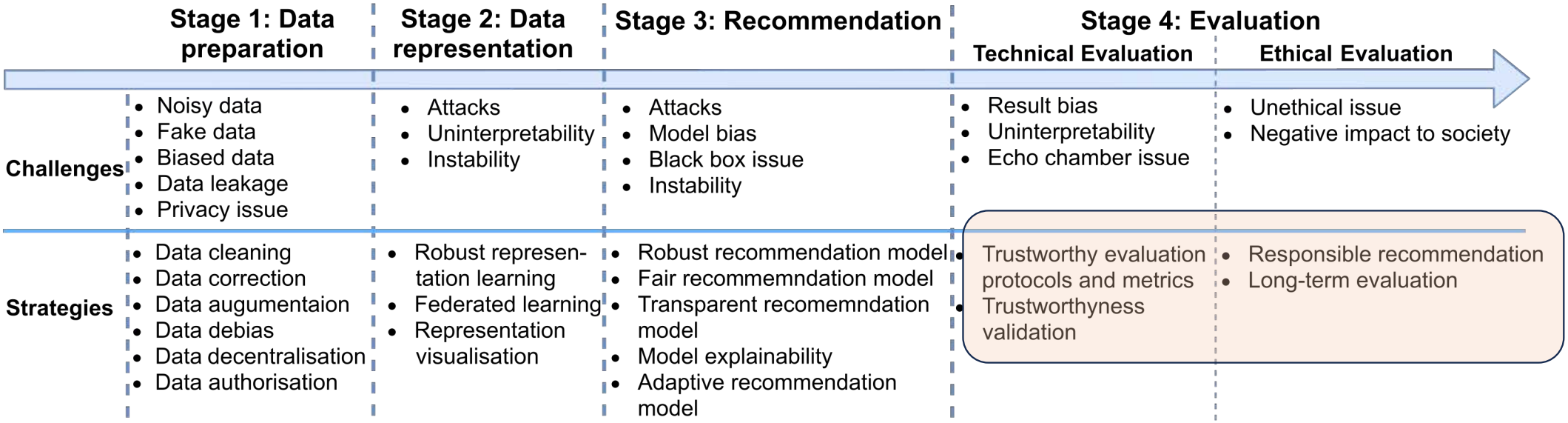
[H. Abdollahpouri & R. Burke, Multistakeholder Recommender Systems, in Recommender Systems Handbook, 647-677, 2022]

Influencing Users

- User preferences are **constructed** while consuming the recommendations
- The RS has **its own agenda** (merging the goals of its stakeholders)
- **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]

Building Trustworthy RSs



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

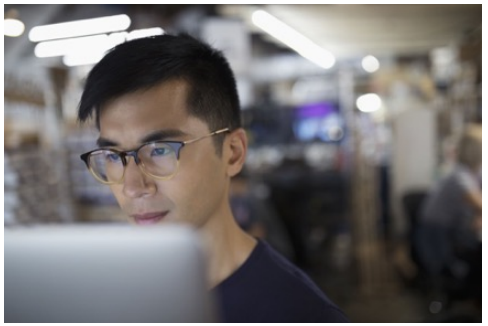
Evaluation methods are independent from RS technology – they are applicable to all types of RSs, both traditional and LLM based.

Beyond Recommendation Accuracy

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

- There are **trade-off**
- Some of these metrics are **not computed on any test data** or ground truth (e.g. coverage and diversity)
- The true **effect** of maximizing these metrics **on user experience** remains unclear (e.g., dependence on personality).

Offline Testing an RS



- ***Collect data:*** *platform's users' likes to items are collected –* ***ground truth GT***
- ***Split data:*** *for each user, partition their likes in train and test sets*
- ***Train:*** *the RS by using the full train set and its GT (all users)*
- ***Test:*** *for each user,* ***compare*** *user's recommendations with likes in the user's test set GT*
- ***Problem 1 - completeness****: is the ground truth* ***sufficient*** *for evaluating all the recommendations? No, most of the recommendations are not in the test set.*
- ***Problem 2 - reliability:*** *are the data in the ground truth an* ***unbiased*** *sample of users' preferences? No, they are not sampled at random.*

Can *old likes* be used to test a *new RS*?

- **RS exposure affects user behavior - users change their preferences**
- User's likes or behavior data **collected in the past** – when users were exposed to a different RS strategy - are not a safe ground truth for evaluating a **new "intervention" RS**
- We need a **better method** for evaluating the **long-term effect** of the **exposure to an RS**

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

Crash and Safety Test

https://www.youtube.com/watch?v=jo9llxyBC8Y

Crash and Safety Test

- **Crash test dummies** (users) are augmented with **sensors** measuring their *reaction* to car crash
- Different dummies are used to assess the crash effects on people of **different sizes and proportions** (e.g. children vs adults)
- Tests are **repeatable** and **alternative contexts** and conditions can be considered
- Prior to the development of crash test dummies, automobile companies tested using **human cadavers**, **animals** and **live volunteers**.

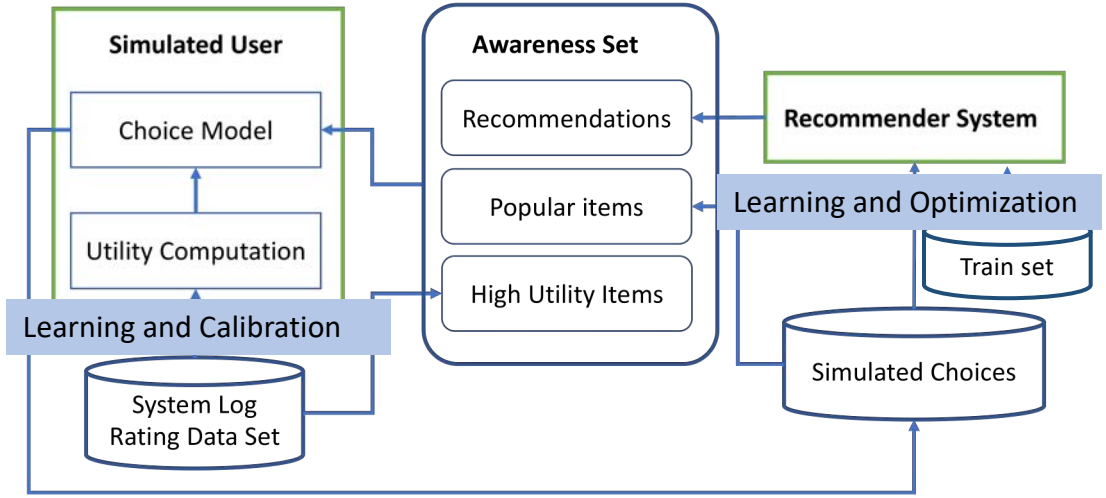
Model-based Simulations

- Learn a **model** that *captures* the **choice behavior of a population** of users, while being exposed to recommendations
- Use this model to simulate user feedback while testing a novel RS

- **Pros:**
  - We can simulate as many interactions as needed with any candidate RS
- **Cons:**
  - We must guarantee that the choice model **faithfully captures** the **users' behavior** (*better than assuming that items liked in the past will be liked again*).

Simulation of Choices

- **Simulated users** estimate items' utility (learned preferences)
- Users are **aware** of:
  a) recommended items,
  b) popular items, and
  c) some items that strongly match their preferences
- They **choose** with a stochastic model: *the larger the item utility is, the larger the choice probability is*



- RS salience effect may be simulated by **increasing 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)

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

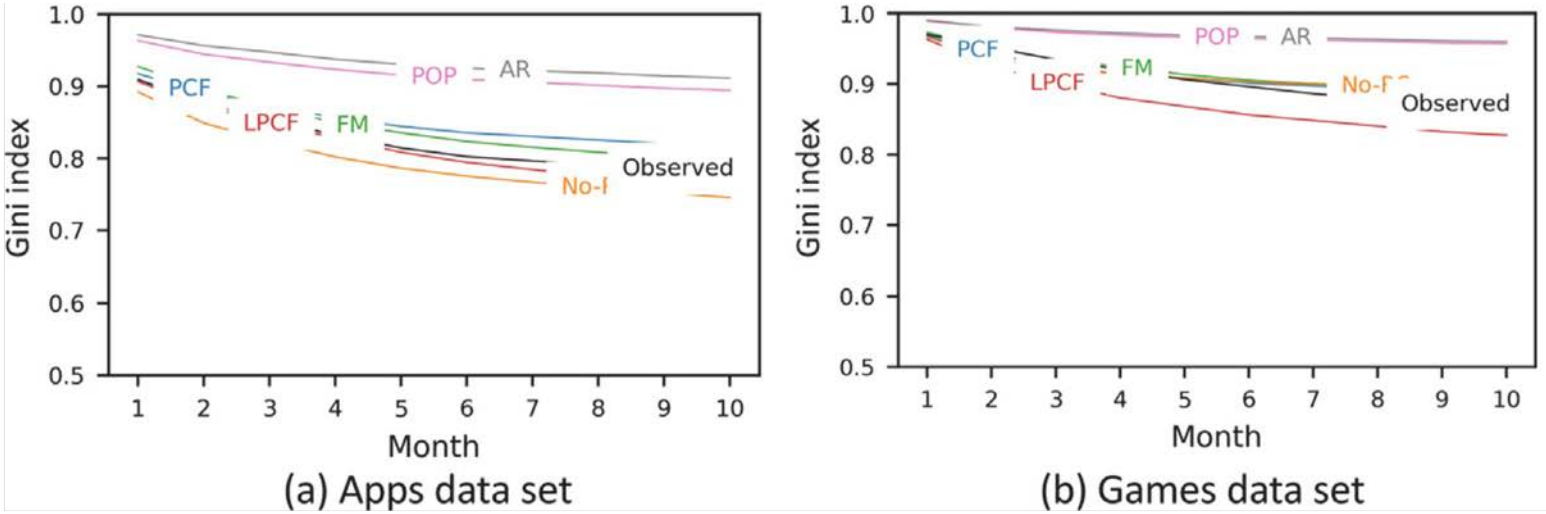
The choice model and the RS are trained differently: correctly simulating choices vs optimizing a performance criteria

- *A simple, well studied, and easily **interpretable** choice model*
- A_u is the **set** items the user *u* is **aware** of – choice/awareness set

$$v_{uk} = \beta\, util(u,k)$$

- *util(u,k)* is the (estimated) **expected utitility** of the item *k* for user *u* - predicted rating
- β is a free parameter: **calibrated** to minimize a target bias (e.g. minimize the difference in the distribution of simulated choices w.r.t observed choices)
- If the item *k* is **recommended** to the user, one **can simulate that the user gives more utility** to the item (recommendation salience): *util(u,k) := 2 util(u,k)*

Simulation Results



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

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

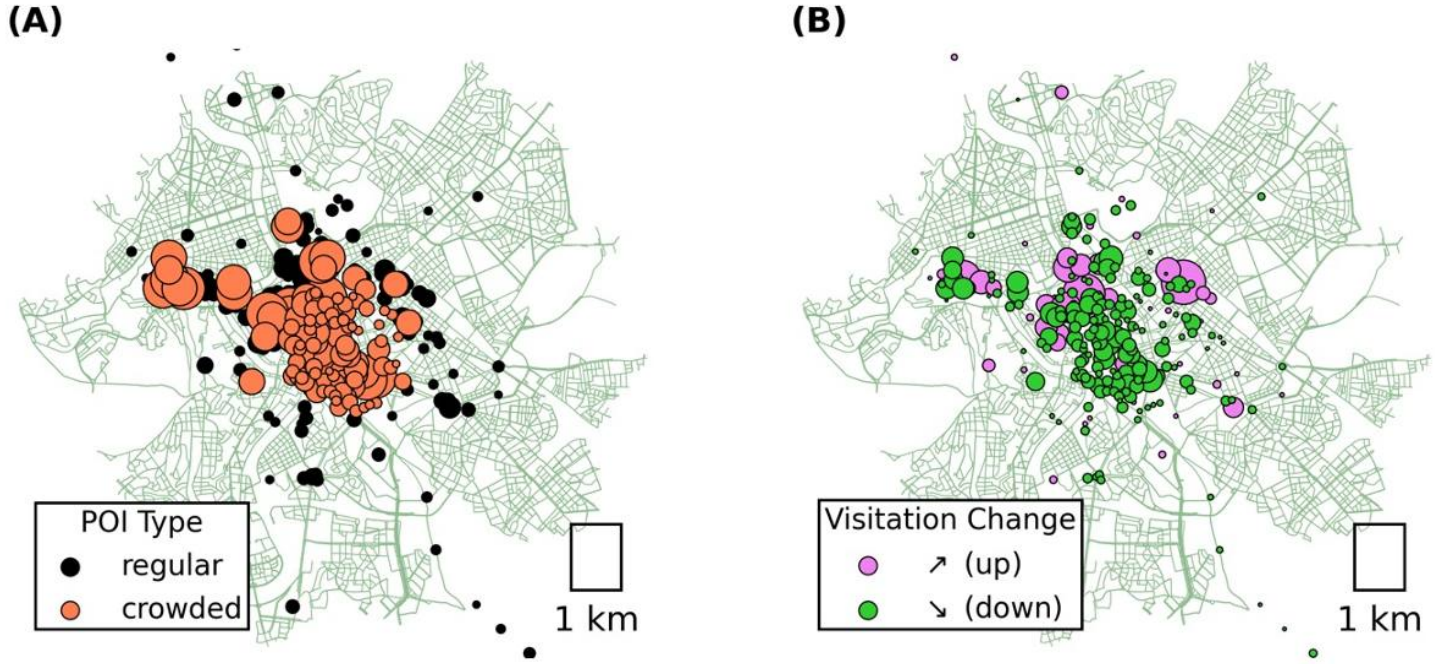
Overtourism

Application Example: Overtourism

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

Venezia, 19/2/2025 - Carnaval

Taming Overtourism



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

[P. Merinov, F. Ricci: Positive-sum impact of multistakeholder recommendations for urban tourism. Appl. Soft Comput. 190: 114582 (2026)]

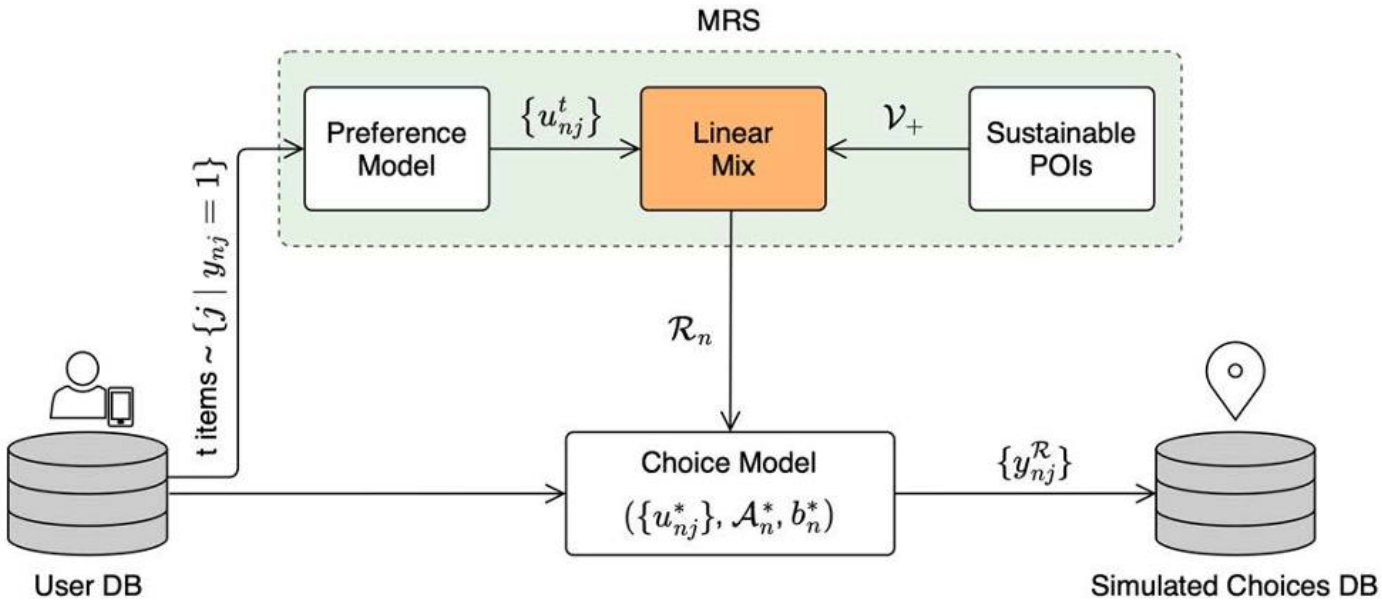
Positive Sum Impact

- *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:

train
- Estimate the (**limited) user knowledge of the POIs catalogue** (awareness)
- Build an optimized and **reliable** (trustworthy) data-driven **choice simulation protocol**

test
- **Simulate** the impact of the RS policy for alternative configuration parameters
- **Measure** the obtained benefit for tourists and destination.

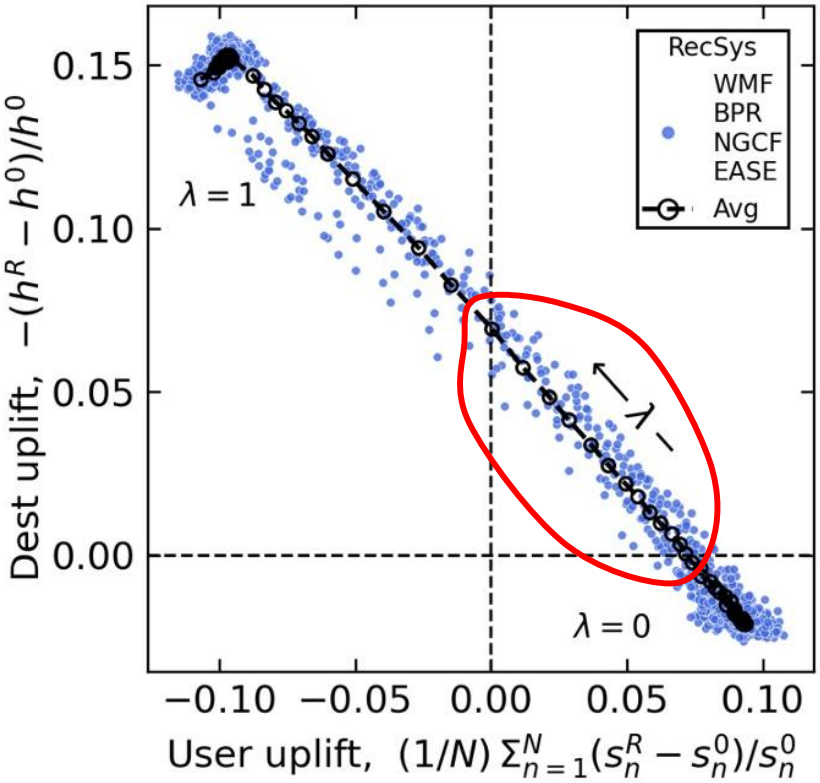
Simulating tourist interactions with an MRS



- *B* recommendations ($x_k = 1$) that 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)
- **Points** represent the **performance of an MRS** built with a particular λ value and a particular algorithm (WMF, BPR, NGCF, and EASE)
- The choice model was previously calibrated (for utility unbiased estimation)
- There exist a range of λ values that produce a positive uplift for both stakeholders.

[P. Merinov, F. Ricci: Positive-sum impact of multistakeholder recommendations for urban tourism. Appl. Soft Comput. 190: 114582 (2026)]

Behaviour shift

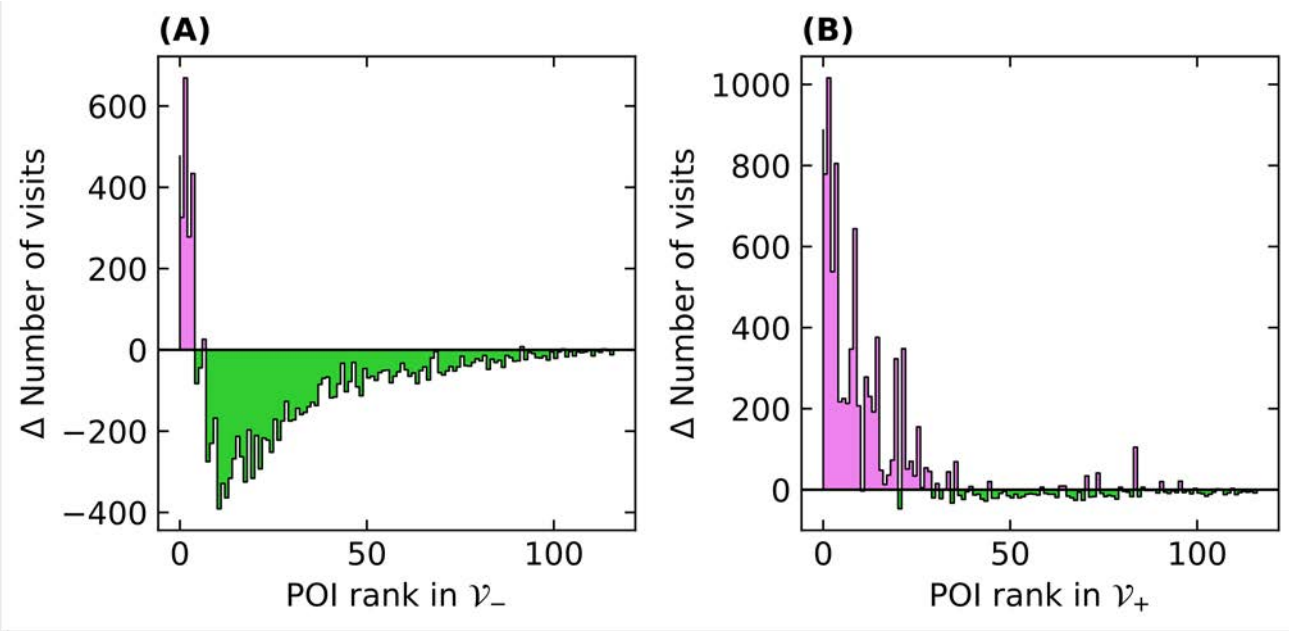


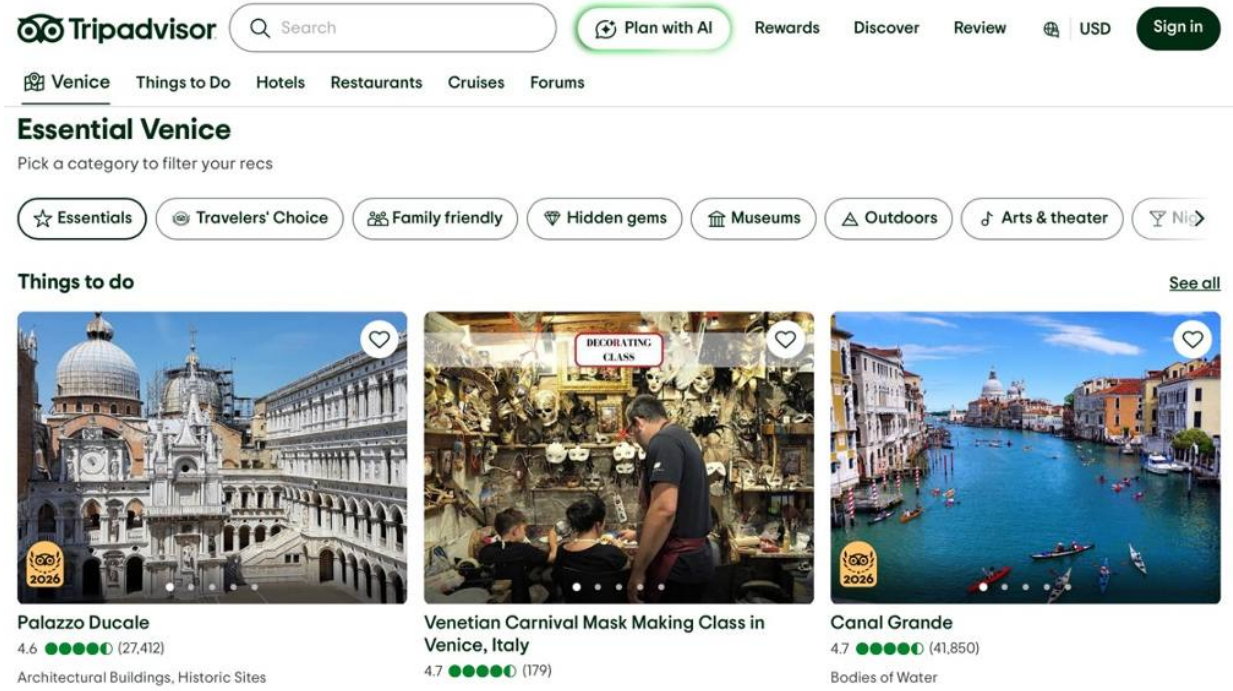
Figure 11: Recommendation policy in action. (A) Visits shift in $\mathcal{V}_-$ (POIs sorted by decreasing value of popularity) after applying the proposed MRS policy. Most critical central POIs exhibit reduced crowding (green bars), while a few popular POIs experience an increase in visits (violet bars). (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 top popular POIs become even more popular: their *utility* is high, they are well known, and there are no alternative less crowded POIs that have larger utility.

Reducing the number of visits to extremely popular POIs can be obtained by RSs that decrese the expected utility of crowded places (persuasion).

Manipulating Recommendation Salience

- Recommendations are meant to **increase the salience** of the selected items (influencing choice)
- Users often underestimate the negative effect of crowding and decide to visit a POI even if it is know to be overcrowded.



Salience Effect on Experienced Utility

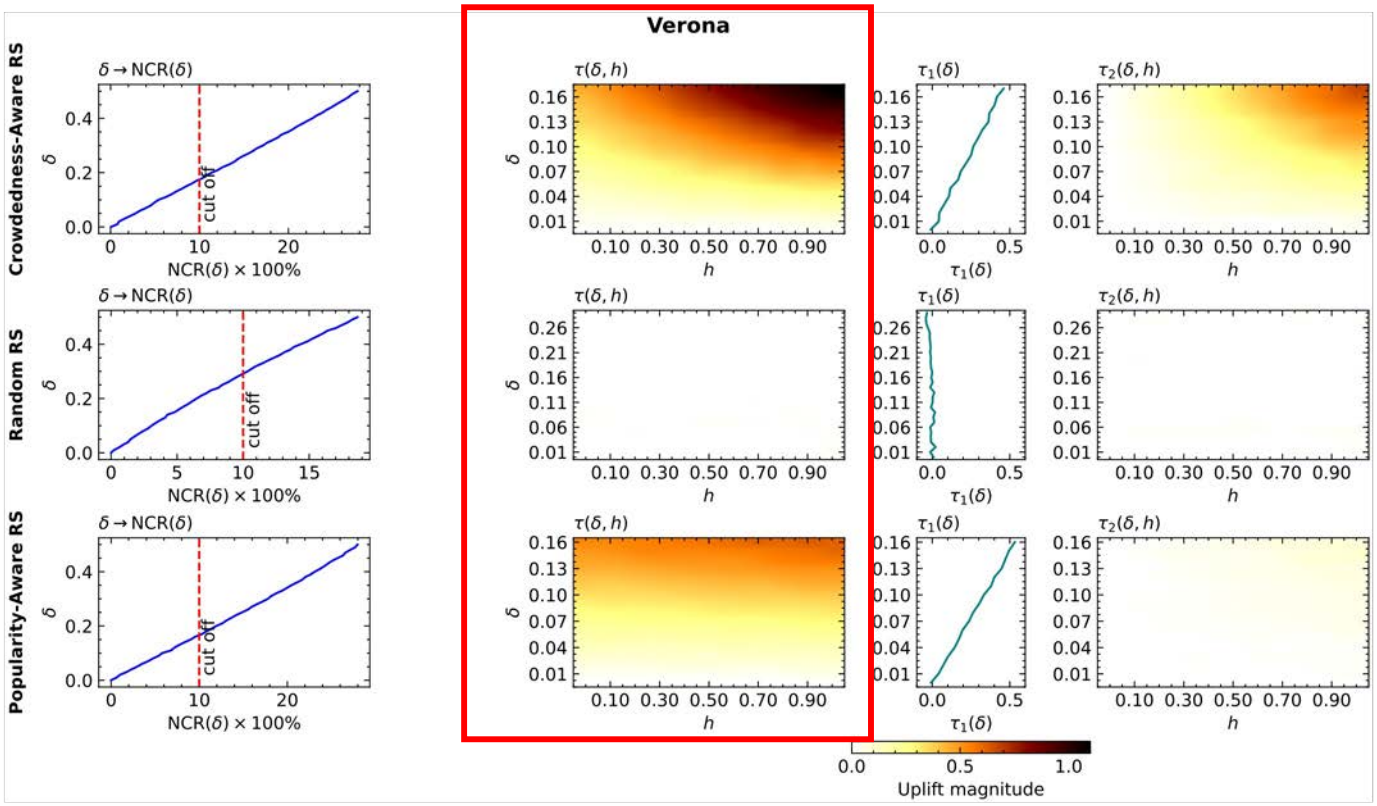


Fig. 8. VeronaCard simulated uplift surface. Salience δ and penalty h are free parameters. For each RS, the left subplot shows the simulated NCR to salience δ dependency. The dependence curves for small δ values (here range $[0, 0.5]$ is covered) are approximately linear. Right subplots report in a row: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

Recommended items receive a δ **increment in expected utility** – increasing the probability of being chosen

Crowdedness produces a **decrement** *h* of **experienced** utility

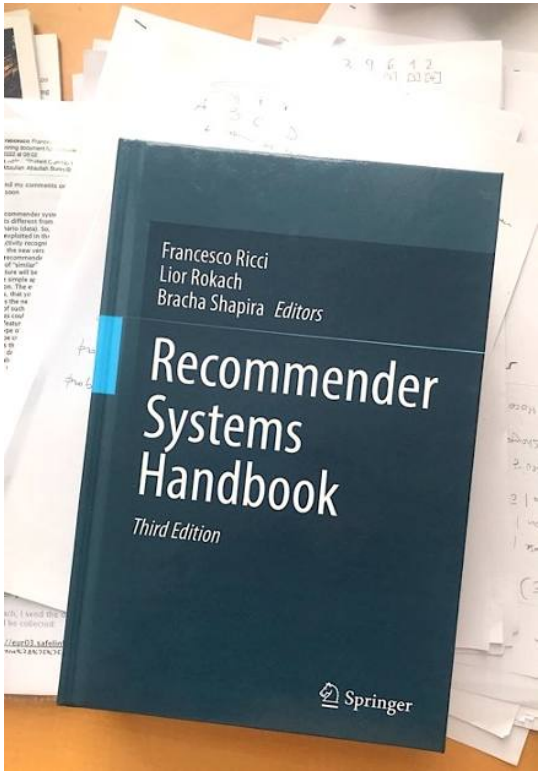
Users have **complete awareness** of the catalogue.

[Merinov & Ricci, Shaping Sustainable Tourist Experience: Simulating the Impact of Recommendation Salience, ACM Trans. Rec. Sys., 2025]

Is the Simulation Reliable?

- The **expected utility** of an item, used by the choice model, is influenced by a number of contextual conditions:
  - Graphical user interface
  - Item description
  - Explanations
  - Conversational interaction
  - Decision biases as the attraction effect
  - User intent and personality
- It is hard to know what the users of an RS really know about the domain items (awareness).

Future Work



Apologies for not crashing real people with a real GUI ☺

Simulation does NOT replace user studies

Special thanks to **Naieme Hazrati** and **Pavel Merinov** – PhD students