Perspectives Breakout Questions

Additional questions:
(Please add any question that you would like to discuss at the workshop)

- How come only few RecSys interfaces allow the user to make autonomous decisions and give explicit feedback or at least allow a true understanding of how the system works?
- Should we have a common repository of shared 'research components', such as simulations, datasets, etc. that meets with FAIR (findability, accessibility, interoperability, and reusability) principles? What would be good metadata to help find 'research component' relevant to specific RS problem characteristics?
- Would sharing code be enough for reproducibility?
- How can we tell if the recommender system (retail) has bias for products that were recently on sale (discounts)? There are two possible scenarios, the products were put into the training dataset or they were not, and people chose those products instead of the ones recommended by the model.
- Is it possible to evaluate a user's product discovery generated by the recommender system? For example, a user always buys cutlery but thanks to the model, they started to buy dishes.
- Is there any gold standard on User Simulation?
- How can we use explanations methods for recommender systems, considering the element of surprise?
- What kinds of evaluation methods can evaluate the UX of recommender systems?
- In recommender system, what effect size is large enough to be of value in a practical sense?
- Do you have any preferred method to find out feature importance for recommender systems? Is it a good idea to evaluate feature importance?
- How to bridge the gap between offline and online evaluation metrics?
- What are good practices in RecSys evaluation and how can we get there?
- How to get rid of common bad practices in RecSys evaluation?
- What is the purpose of RecSys evaluation?
- Who is the target of the evaluation?
- Is counterfactual evaluation an option?
- Future directions?
- When are behavioral/user signals in production not good enough/problematic for evaluation (online, off-policy, etc)?
- How to evaluate the influence of negative examples in triplet loss recommender systems training?
- Do we want to recommend what users are likely to engage with or what users' are likely to enjoy engaging with?
- Can we contextualize engagement data in order to single out data that is generated by compulsive or other behavior that might not represent users' intended behavior? For example if we find through participant case studies, that user intention is gradually eroded during a session (as in Klobas et al 2019; Lukoff et al 2018), can we use

temporal information to discount engagement data generated later in a usage session (in order to preserve user autonomy)?
- Who trusts their offline experiments?

Group 1

Question: How to bridge the gap between offline and online evaluation metrics?

Group participants: Kay Wong, Tiago Cabo, Emilia Gómez, Anastasia Shukhova
- Intro from participants: focus on industry, ML, interest in evaluation
- Online evaluation requires a reduced set of options to be evaluated, vs off-line evaluation where it is possible to try out different setups.
- To decide on the models to test online, you first need to select a reasonable number of candidates in the offline evaluation. However, it can be a challenge to see a similar model performance online to the one you observed offline.

Group 2

Question: Valid Offline Experiments
Group participants: Simen Eide, Zeno Gantner, Eva Zangerle
- You cannot A/B test everything
- Hard to get good correlation between online and offline experiments
    - Many offline protocols do not try hard enough to replicate/simulate the actual user experience. Example 1: many datasets with implicit feedback only contain positive implicit feedback, but no information which items were shown to the user but not interacted with. Example 2: data collected from many places instead of just the specific use case. Not clear whether that data is representative enough.
    - When comparing similar models, it is more likely that there is correlation than when comparing models that work very differently.
- Offline experiments hardly trusted in industry, but widely used in academia
    - Online experiments for academics?
- How much overhead is there at your organization for A/B test?
    - How is an A/B test planned?
    - How many levels of approval?

Group 3

Question:
-

Group 4

Question:
Group participants: Andrea Barraza, Johan Rodríguez, Shaghayegh (Sherry) Sahebi

- How to evaluate long-term metrics and how to consider delayed (unobserved) rewards?
- How to reduce simulators' biases?
- How to use propensity scoring / counterfactual evals to have a fair evaluation and reduce the feedback loop?

Group 5

Question: Who is the target of the evaluation?
Group participants: Leonardo Bonifacio, Bas Vlaming, Daniel Toader, Anna Schröder

- in the end, the business is the main target; customer satisfaction is important, but hard to quantify
- conversion, CTR, margin, order size are most common quantities to optimise for
- user's satisfaction sometimes collected through simple explicit feedback elements (which then multiply the score of items), or through retention
- rarely surveys as an extra step to collect feedback data about recommendation quality (not suitable for all markets though)

Group 6

Question: How to bridge the gap between offline and online evaluation metrics?
Group participants: Ngozi Ihemelandu, Chunpai Wang, Milena Filipovic

How do we bridge the gap between online and offline metrics?
We acknowledged that offline cannot adequately model online but we could incorporate time while splitting data
For offline evaluation, one challenge is to build a perfect user simulator to minic real user behavior. We may use the online evaluation result as the gold standard to refine the simulator so that the offline evaluation results are aligned with online evaluation results.
We could also try different user simulators for offline evaluation.

Group 7

Question: Biases that can affect evaluation/results
Group participants: Fábio Tanaka, Ladislav Peska, Fernando Perez, Jean-Michel D

- What are acceptable and non-acceptable bias
- Is bias impacting your metrics? Is bias impacting the objectives and goals?
- Paper accepted at this workshop from Ladislav
    - How to fight one bias while keeping other parts of the system intact
    - Mostly on popularity bias on MovieLens datasets
- Another difficult question, mostly in Video Game shopping, is how to arrange games while online shopping? how to get the best videogame arrangement by geographical zones?
- What happens with cultural / geographical differences?
    - Complexity of system. Multimodel can learn and deliver recommendations to geographical zones / interests but it significantly increases the complexity of the system. On the other hand, single models might be biased and affect less represented cultural / geographical zones.
    -

Group 8

Question: Do we like the recommendations from products we are using daily?
Group participants: Dilek, Kueichun, Kalle

- Recommenders without prior information like subscription sometimes are not able to capture a user's preference.
- Sometimes it feels like information overload. Too many options make it harder to make decisions
- Would be nice to specify if I am in explorer mode at the moment. My openness to new genres / items varies among products and also in time.

Group 9

Question: Are evaluations domain dependent?
Group participants: Daniel Woolridge, Anton Angwald, Gordon Blackadder, Indrė Žliobaitė

- What makes a "good" recommendation in one domain is not always the same in another domain
- Should we use the same evaluations across domains
  Sampling methods for negative samples for ranking metrics
- If semantic metadata is important for the recommendation problem, benchmark datasets from different domains are insufficient, comparison between domains is difficult

Group 10

Question:
What is the main purpose for increasing purchases in e-commerce using RS? Is that a good purpose? :)

How might that be tailored for the UX of a specific user group?

Group participants:
Maliheh Ghajargar
George Mincu
Fernando Diaz
Mirza