



A Comparison of Calibrated and Intent-Aware Recommendations

▶ Mesut Kaya, Derek Bridge

CSCE 670 Paper Breakdown

DHILTON PANDARAPARAMBIL ANTONY

UIN - 327008700

MOTIVATION

Calibrated and intent-aware recommendation are recent approaches to recommendation that have apparent similarities. Both try, to a certain extent, to cover the user's interests, as revealed by her user profile.

In this paper, we compare them in detail. On two datasets, we try to find the extent to which intent-aware recommendations are calibrated and the extent to which calibrated recommendations are diverse. We consider two ways of defining a user's interests, one based on item features, the other based on sub-profiles of the user's profile.

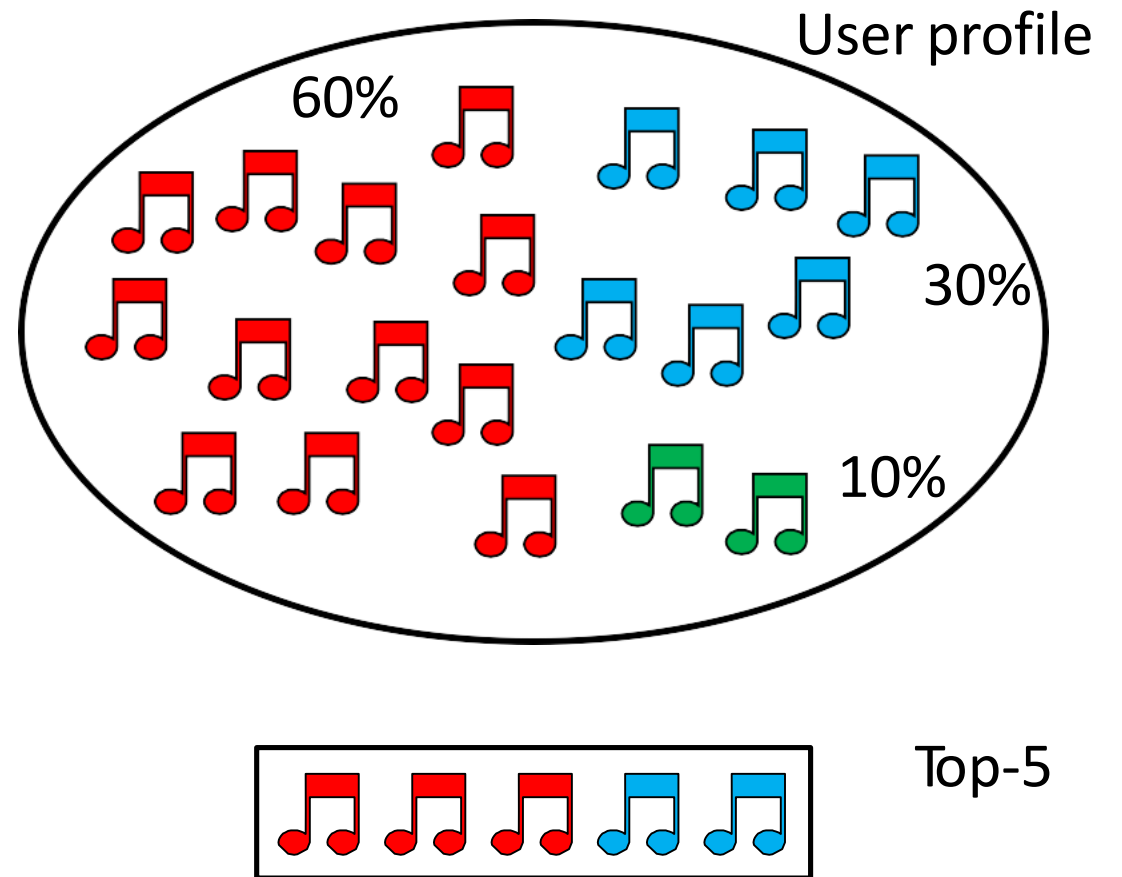
Abstract

This paper compares the two different recommendation approaches to determine how well each achieves three objectives: calibration, relevance and diversity. More specifically, the paper's contributions are:

- Defining a new variant of Steck's calibrated recommender systems, one which calibrates with respect to sub-profiles, rather than item features.
- Defining three new evaluation metrics, corresponding to existing metrics that are defined using item features. The new metrics use sub-profiles, rather than item features. Using the new metrics alongside the existing ones gives a more balanced view of the performance of the recommender algorithms.
- Presenting an empirical comparison using all these metrics on two datasets to see the extent to which calibrated and intent-aware recommenders do produce calibrated recommendations, relevant recommendations and diverse sets of recommendations.

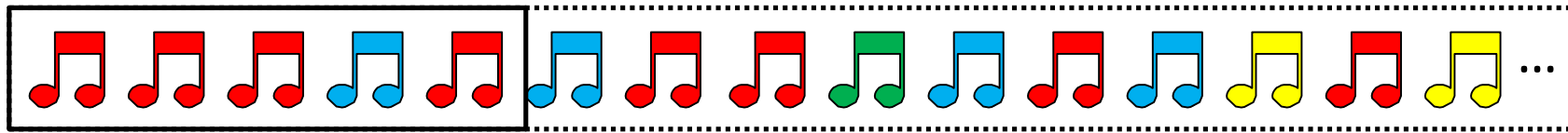
Calibrated Recommendations

- The top- n recommendations from a model trained for accuracy may
 - focus on the user's main interests
 - leaving lesser interests under-represented or absent
- Calibrated recommendations [Steck 2018]
 - the goal is to reflect the various interests of the user in the appropriate proportions



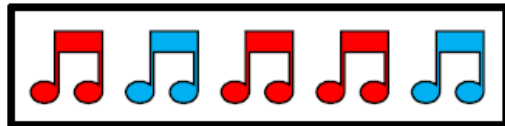
Calibrated Recommendations

- Candidate items, ordered for relevance $s(u, i)$ by a baseline recommender



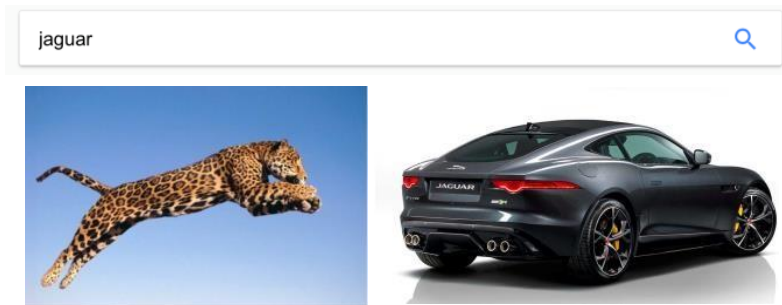
- Greedily re-rank by a linear combination

$$f_{obj}(i, RL) = \overbrace{(1 - \lambda)s(u, i)}^{\text{Relevance}} + \overbrace{\lambda \text{ cal}(i, RL)}^{\text{Calibration}} \quad 0 \leq \lambda \leq 1$$

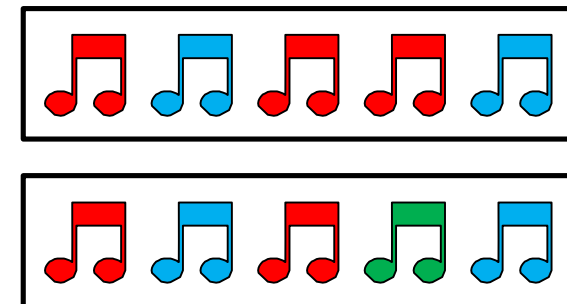


Intent-Aware Diversification

- Early work on diversification
 - greedily re-rank to decrease similarity of items in top- n [Carbonell & Goldstein 1998, Ziegler et al 2005]
- Intent-aware diversification from IR [Santos et al. 2010]
 - top- n contains results for each interpretation of an ambiguous query

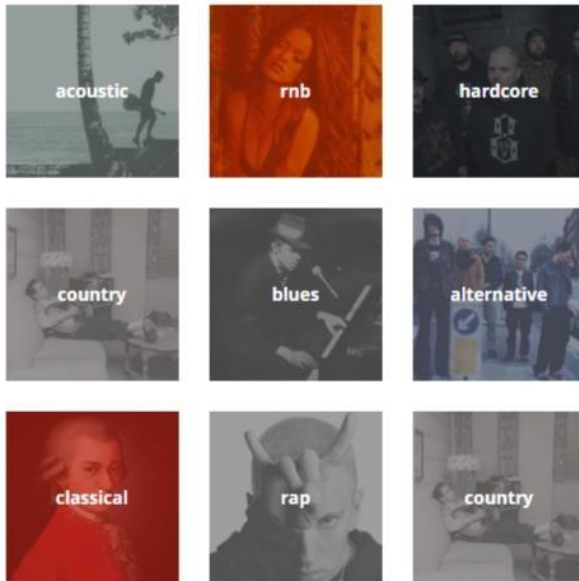


- Intent-aware diversification for recommenders [Vargas 2015]
$$f_{obj}(i, RL) = (1 - \lambda)s(u, i) + \lambda \text{div}_{IA}(i, RL)$$
 - top- n contains results for each of the user's interests (from her profile)
 - this is like calibration but formulated in a different way that *takes relevance into account* as well

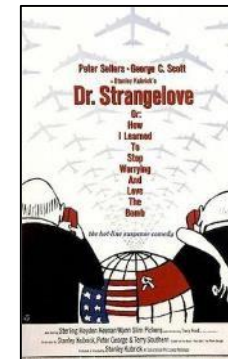


User Interests

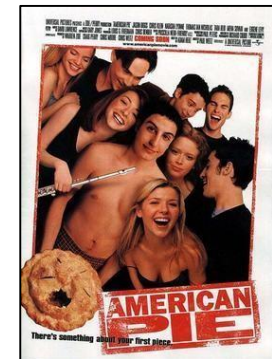
- In [Steck 2018] and [Vargas 2015]
 - user interests are defined by item features, e.g. genres, in the user's profile



- But item features
 - are not available in every domain
 - are often noisy and inconsistently applied
 - may not describe subjective tastes

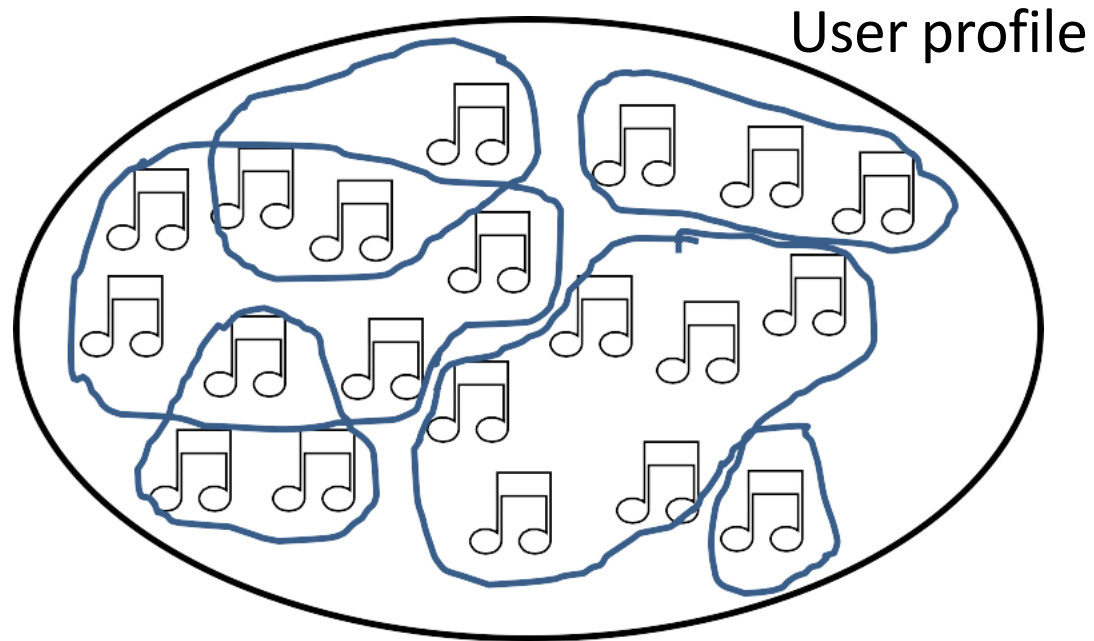


“comedy”






User Interests

- In this paper, we define user interests as *subprofiles*



- Members of a subprofile
 - have similar interactions/ratings
 - no use of item features

		
×		
	×	×
		×
×	×	×
	×	×

Summary So Far

	User interests from item features	User interests from subprofiles
Calibrated Recommenders	Calibrated recommendation using features - CR_F [Steck 2018]	Calibrated recommendation using subprofiles - CR_S
Intent-Aware Recommenders	Query Aspect Diversification framework (xQuAD) [Vargas 2015]	Subprofile-Aware Diversification (SPAD)

Experiments

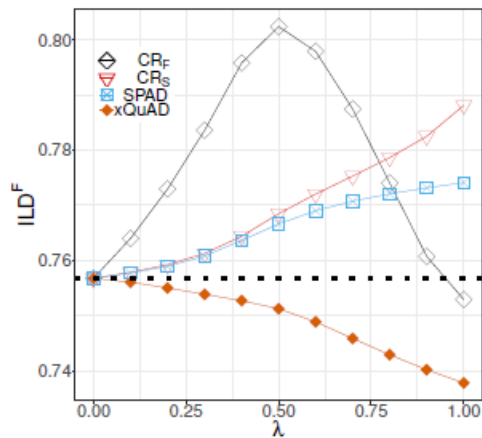
We use two datasets - MovieLens 20 Million dataset and the Taste Profile Subset dataset and then compare the two forms of calibrated recommendation (CRF and CRs) to two forms of intent-aware recommendation (xQuAD and SPAD).

We measure calibration, precision and diversity. In the case of diversity, we show results for four different metrics, and we explore the trade-off the recommenders make between precision and diversity.

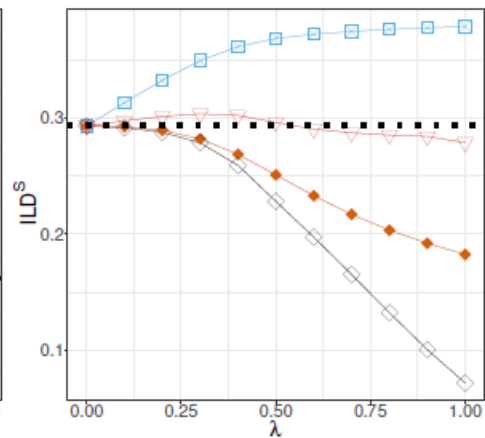
Evaluation metrics

- For recommendation relevance, we measure Precision.
- For diversity, we use Intra-List Diversity (ILD) and α -nDCG. α -nDCG is based on nDCG but it is aspect and redundancy aware, which makes it a measure of diversity.
- To measure the degree of calibration, we use CKL (Kullback-Leibler divergence of the two distributions)
- The objective function is a linear combination of the relevance of the items in the recommendation set and the diversity of that set, the trade-off between the two being controlled by a parameter λ ($0 \leq \lambda \leq 1$)

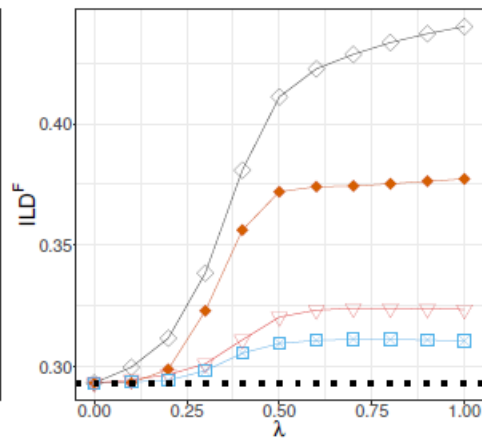
Results



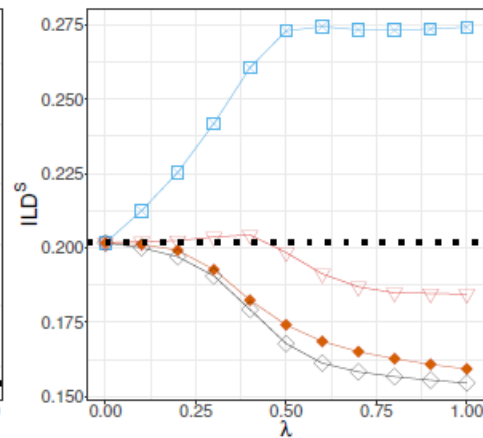
(a) ILD^F on ML20M



(b) ILD^S on ML20M



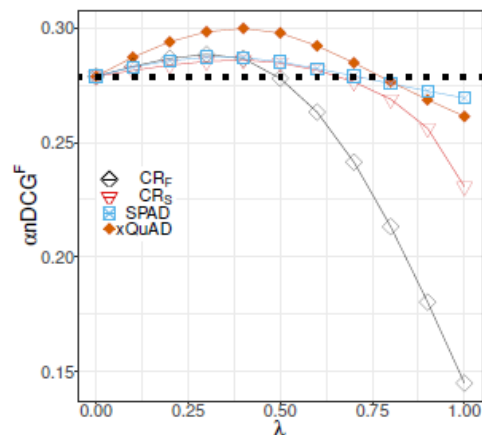
(c) ILD^F on TasteProfile



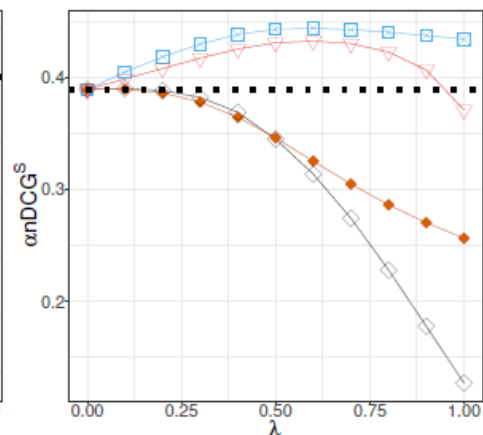
(d) ILD^S on TasteProfile

For ML20M and TasteProfile, ILD measured using features and sub-profiles for different values of λ

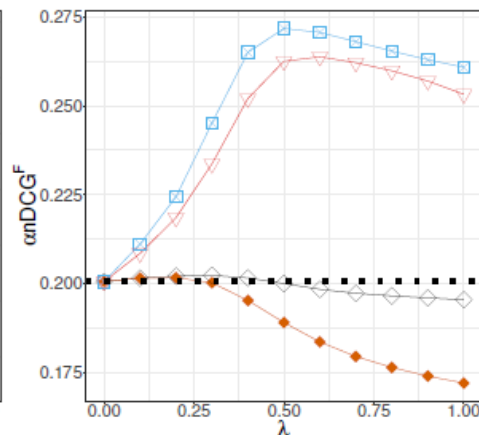
For ML20M and TasteProfile, α -nDCG measured using features and subprofiles for different values of λ



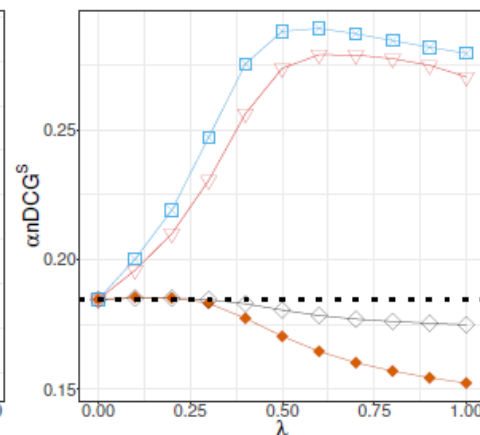
(a) α -nDCG F on ML20M



(b) α -nDCG S on ML20M

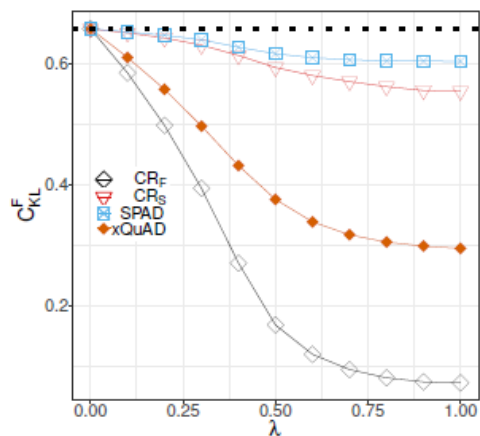


(c) α -nDCG F on TasteProfile

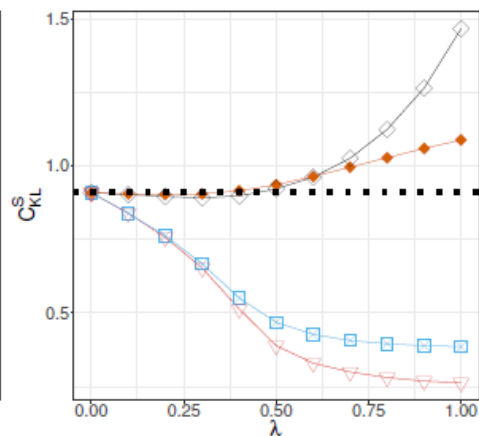


(d) α -nDCG S on TasteProfile

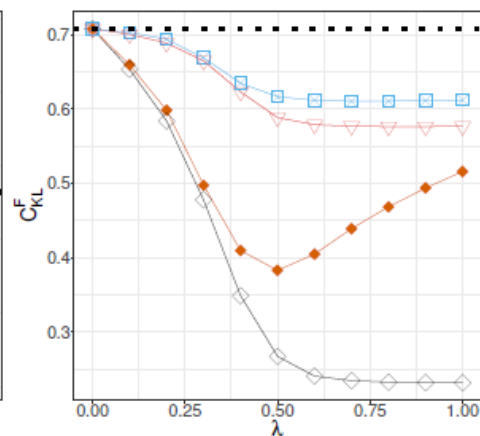
Results



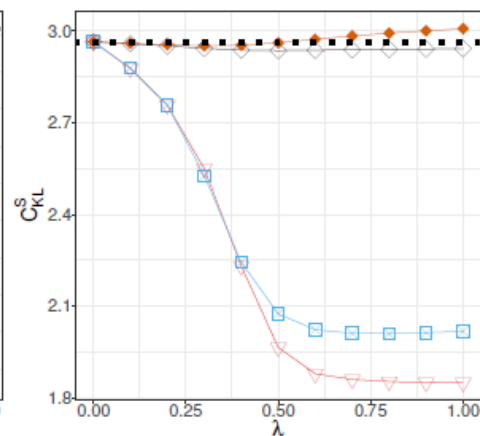
(a) C_{KL}^F on ML20M



(b) C_{KL}^S on ML20M



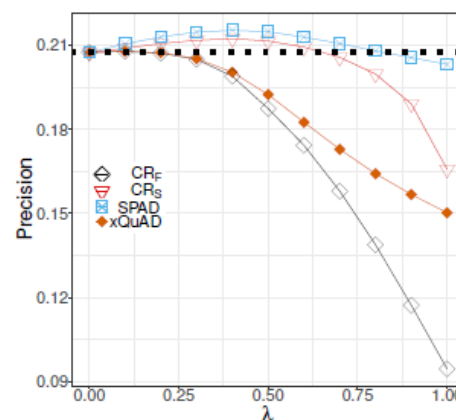
(c) C_{KL}^F on TasteProfile



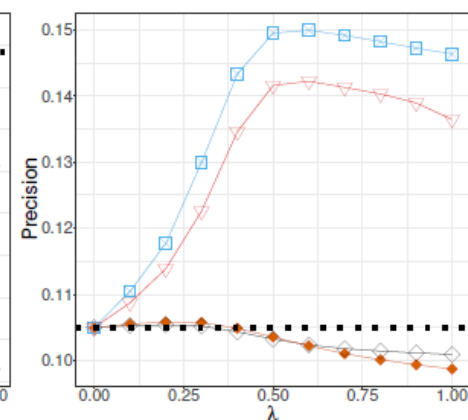
(d) C_{KL}^S on TasteProfile

For ML20M and TasteProfile, C_{KL} measured using features and subprofiles for different values of λ

Precision values for different values of λ for ML20M and TasteProfile



(a) Precision on ML20M



(b) Precision on TasteProfile

Concluding Remarks

- On these datasets, the approaches that use sub-profiles (CR_S and SPAD) achieve
 - the highest precision
 - better than baseline calibration according to both calibration metrics
 - good diversity according to α -nDCG
- SPAD also achieves
 - better than baseline diversity according to both ILD metrics
 - suffers least from the relevance/diversity trade-off

Future Directions

Based on the results that we got, the following can be the next course of action :

- investigate how users perceive calibrated / diversified recommendations
- apply subprofiles to tasks other than calibration/ diversification
- develop the idea that calibration in general (and these approaches in particular) could be used for fairness in recommendations

Thank You