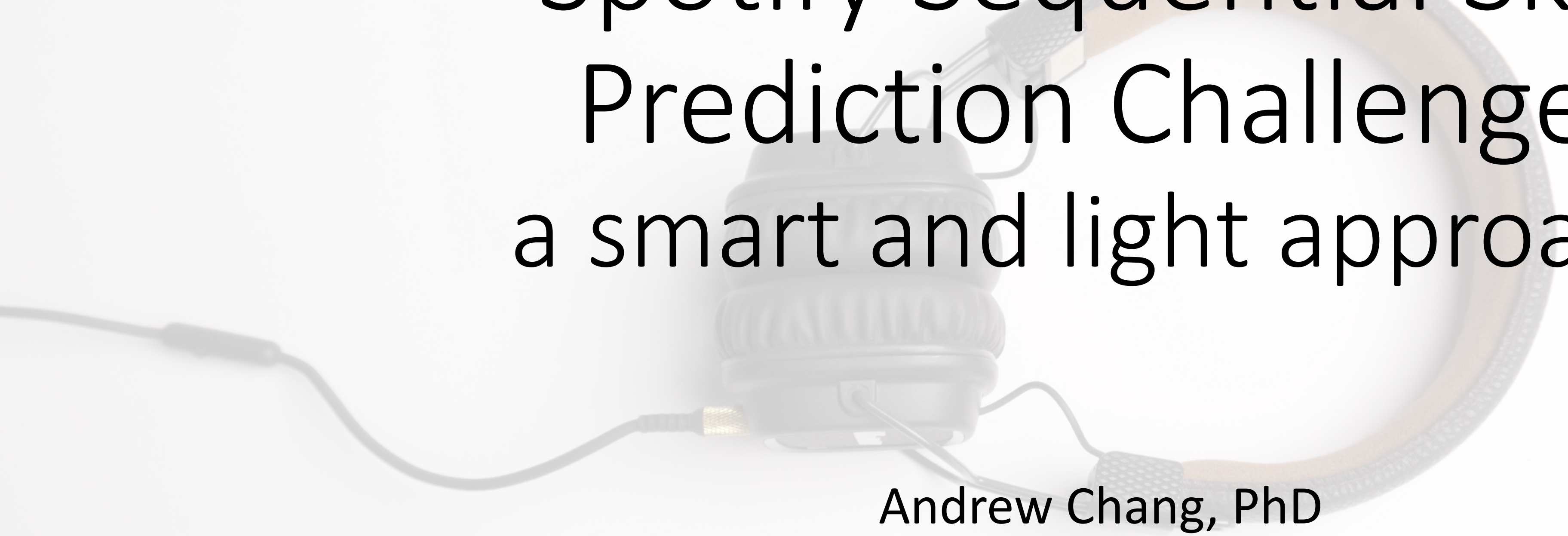
Spotify Sequential Skip Prediction Challenge: a smart and light approach



Andrew Chang, PhD

The right data can be more cost effective than big data.

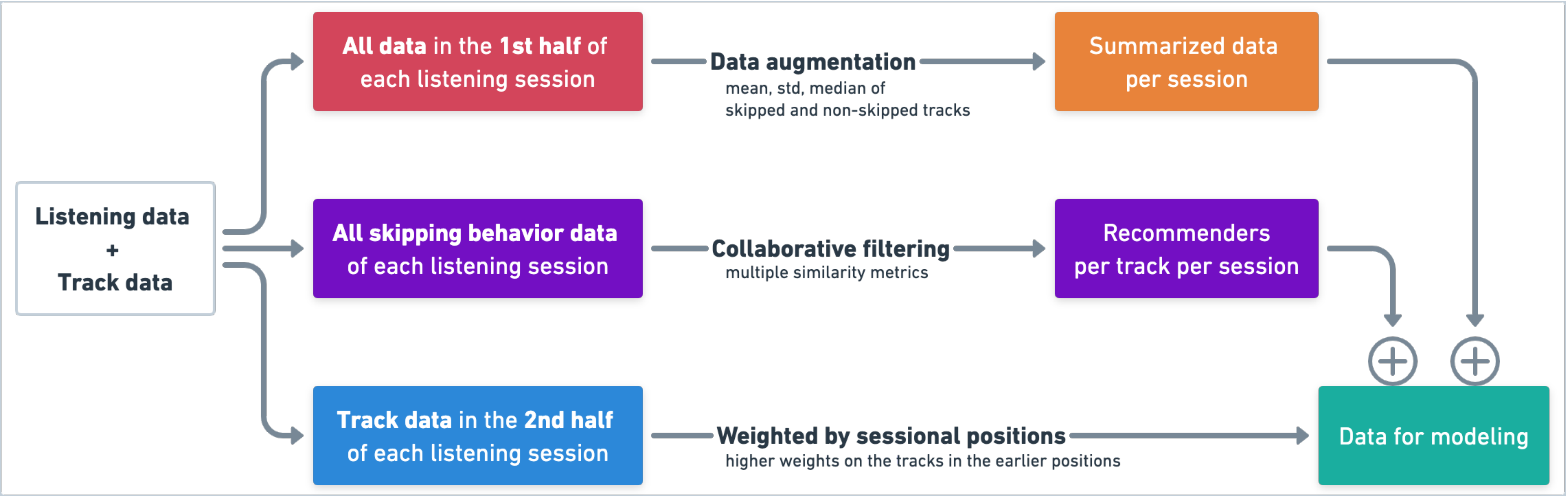
The model trained by the data of 1 day ahead can best predict the skipping behavior on Spotify.

Spotify: sequential listening data

- From recommender system to **sequential recommender system**?

unthoughtful
big data?

thoughtful
small data



Listening data
+
Track data
All data in the 1st half of each listening session
Data augmentation
mean, std, median of skipped and non-skipped tracks
Summarized data per session
All skipping behavior data of each listening session
Collaborative filtering
multiple similarity metrics
Recommenders per track per session
Track data in the 2nd half of each listening session
Weighted by sessional positions
higher weights on the tracks in the earlier positions
Data for modeling

Different angles of the same data

Summarized data

Session

Position

	S1	S2	S3	S4
11	a	d	c	e
12	b	c	w	a
13	c	a	b	c
14	d	e	a	w

Recommender data

Session

Position

	S1	S2	S3	S4
11	a	d	c	e
12	b	c	w	a
13	c	a	b	c
14	d	e	a	w

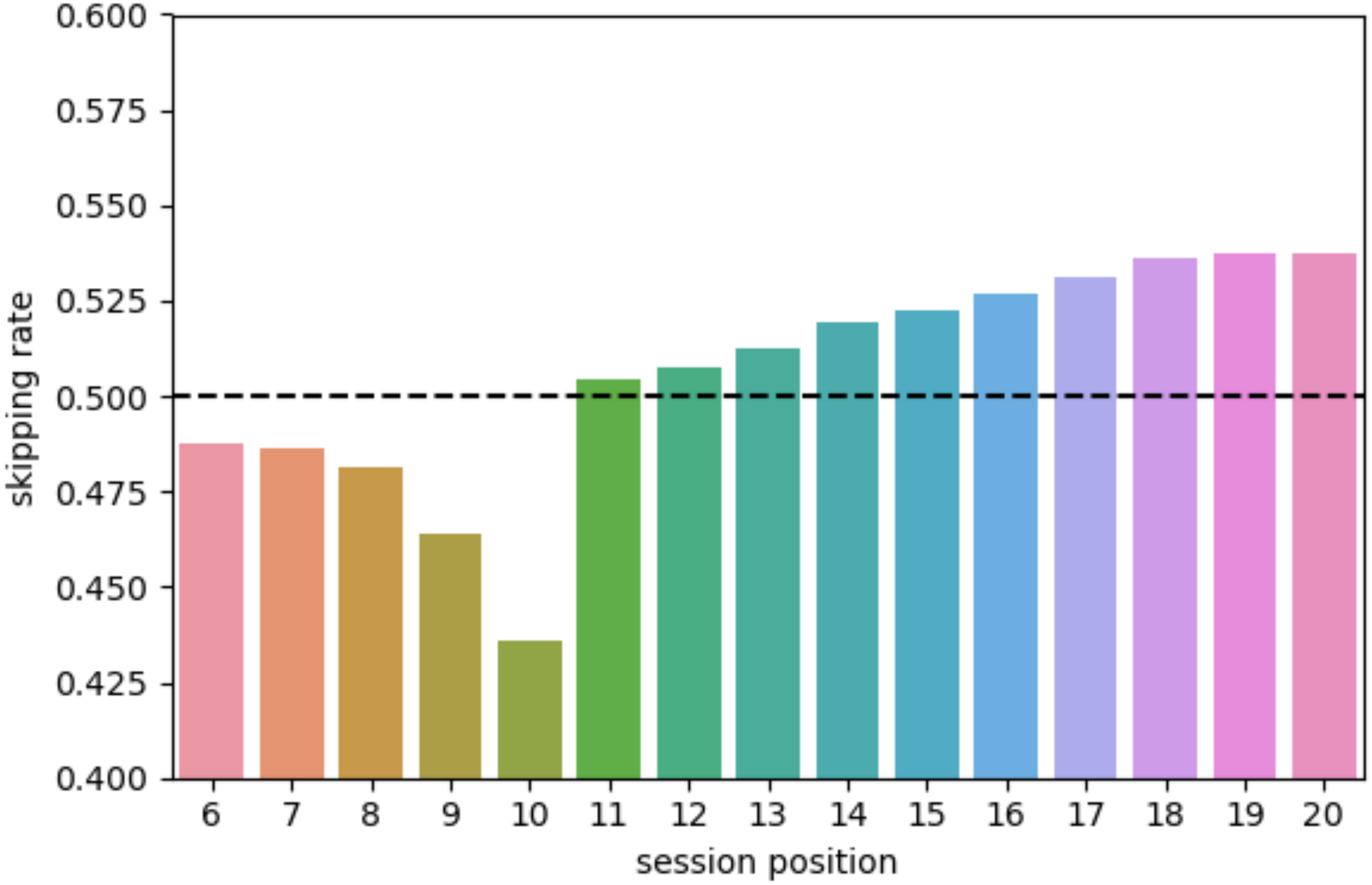
Track data

Session

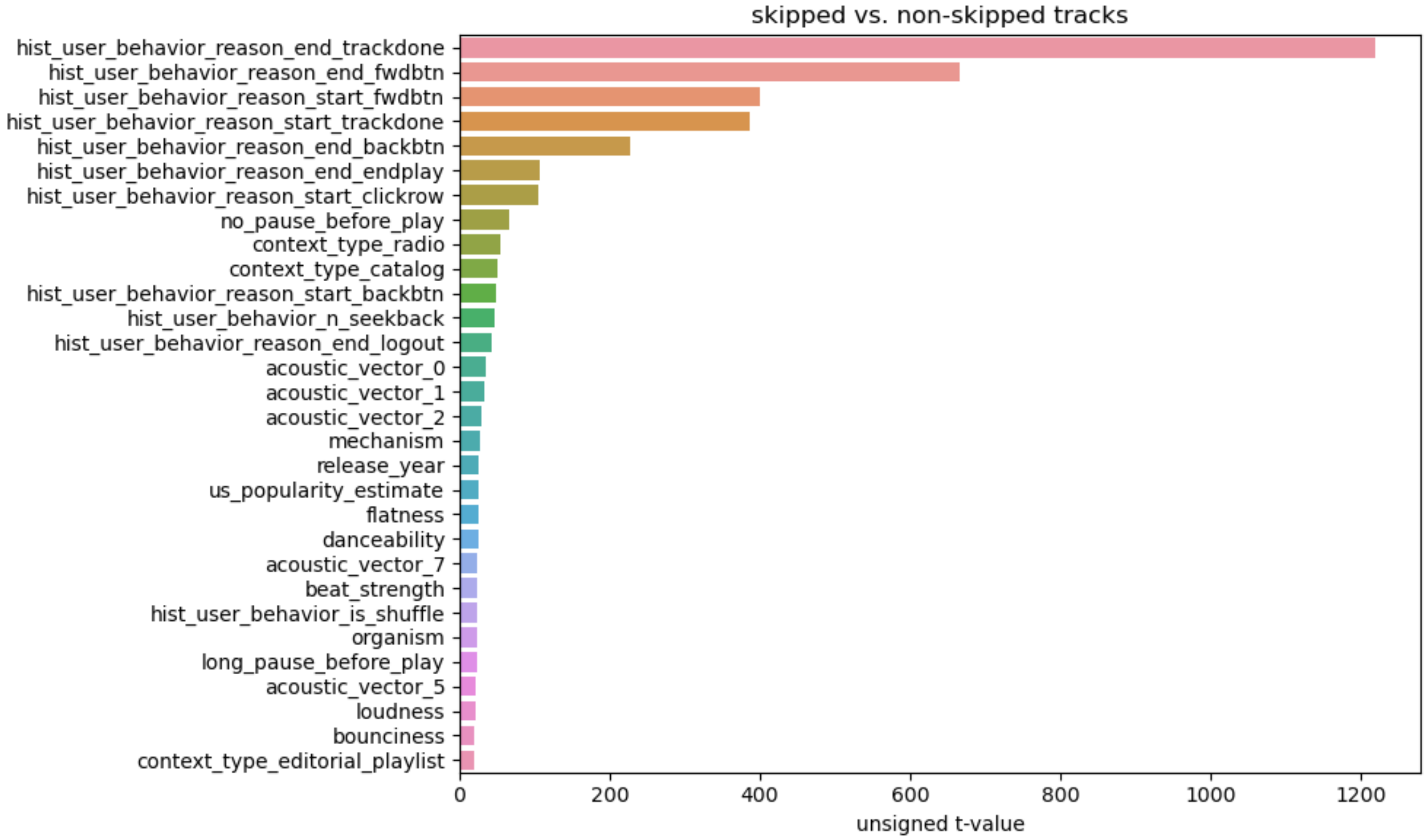
Position

	S1	S2	S3	S4
11	a	d	c	e
12	b	c	w	a
13	c	a	b	c
14	d	e	a	w

The skipping rate by session position



What features were most different between skipped and non-skipped tracks?



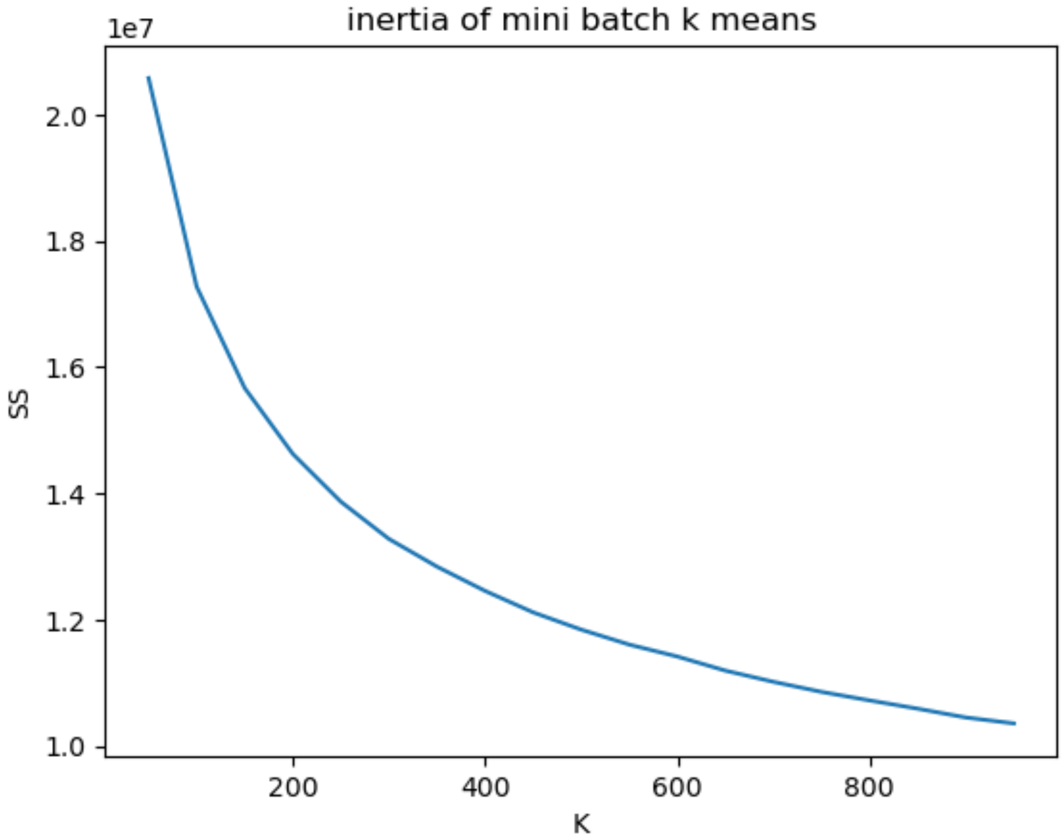
Recommender system

- **Content-based filtering**: If a user skips Track A, he/she is likely skipping other tracks which share the similar features as Track A.
- **Collaborative filtering**: Given that User X skips Track A, B and C, and User Y skips Track A and B, considering the similar skipping behavior between these two users, User Y is likely to skip track C too as User X did.
- **Hybrid**: According to Wikipedia, “there is no reason why several different techniques of the same type could not be hybridized.”

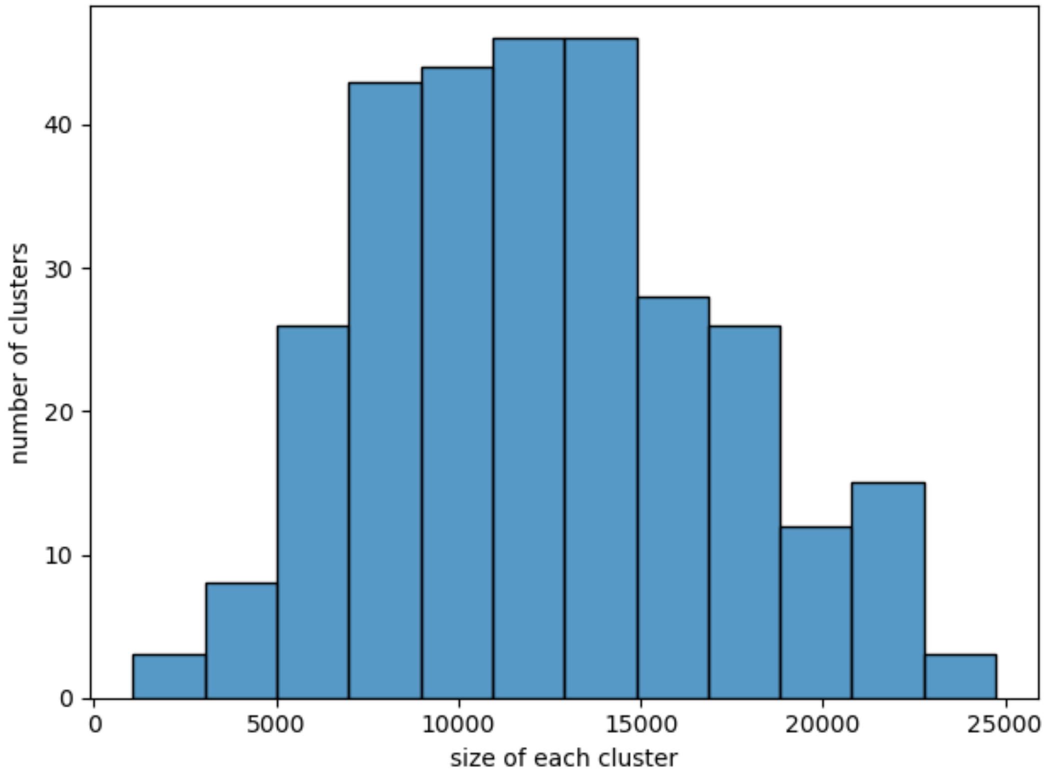
Challenges to build a recommender system

- **Cold start**: As there isn't enough data from the new sessions and tracks to do collaborative filtering, grouping them with the other existing tracks with similar features can make them "not new" and use the existing data.
- **Sparsity**: Given that there was max 20 tracks per session in the current dataset, but there are 3,706,388 tracks in the data. Therefore, the overlap between tracks and sessions will be very sparse if not using clustering.
- **Scalability**: The dataset is way too big (> 350 GB) to be computed on any single local machine. Grouping the tracks into clusters can largely reduce the data dimension and speed up the computation. However, it comes with a price that all the tracks within the same cluster will be considered the same for the recommender system.

Use k-means clustering to group the track dataset

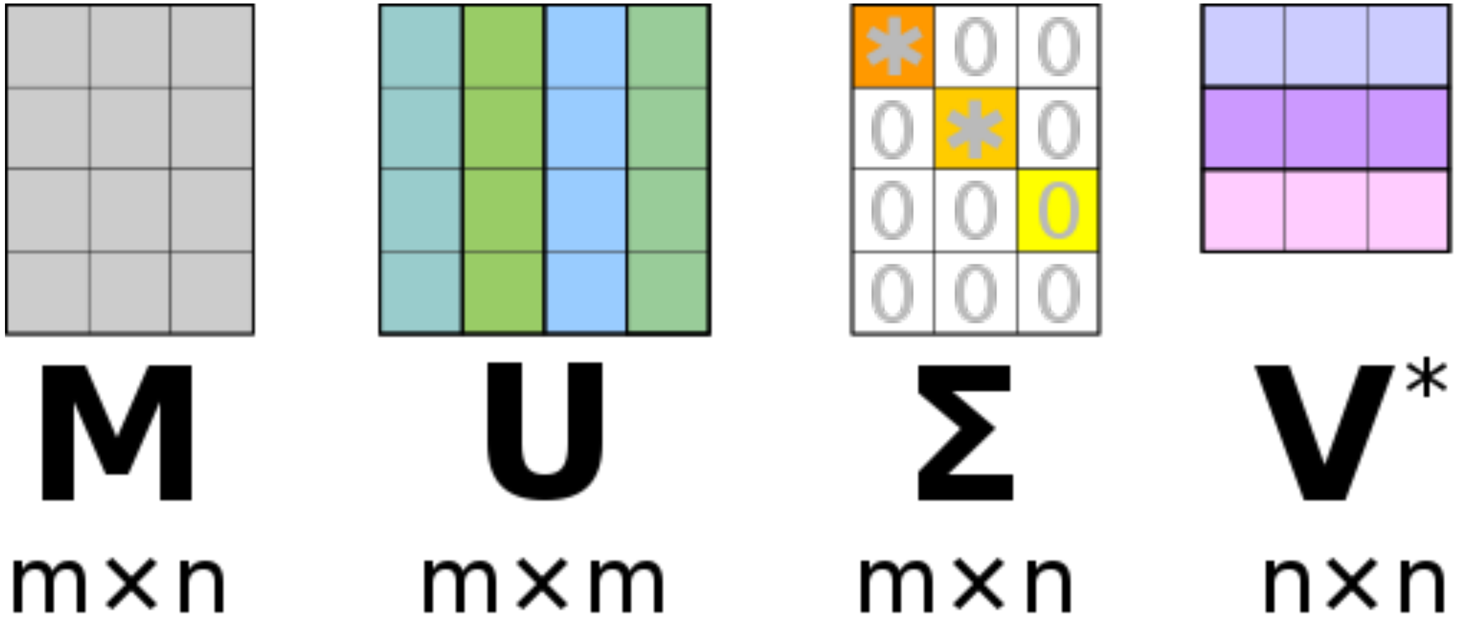


K = 300



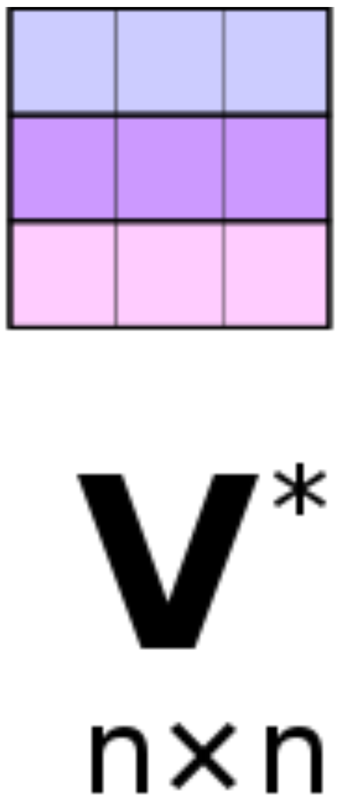
Data reduction using singular value decomposition (SVD)

Millions * 300 (M) → 300 * 300 (V*)

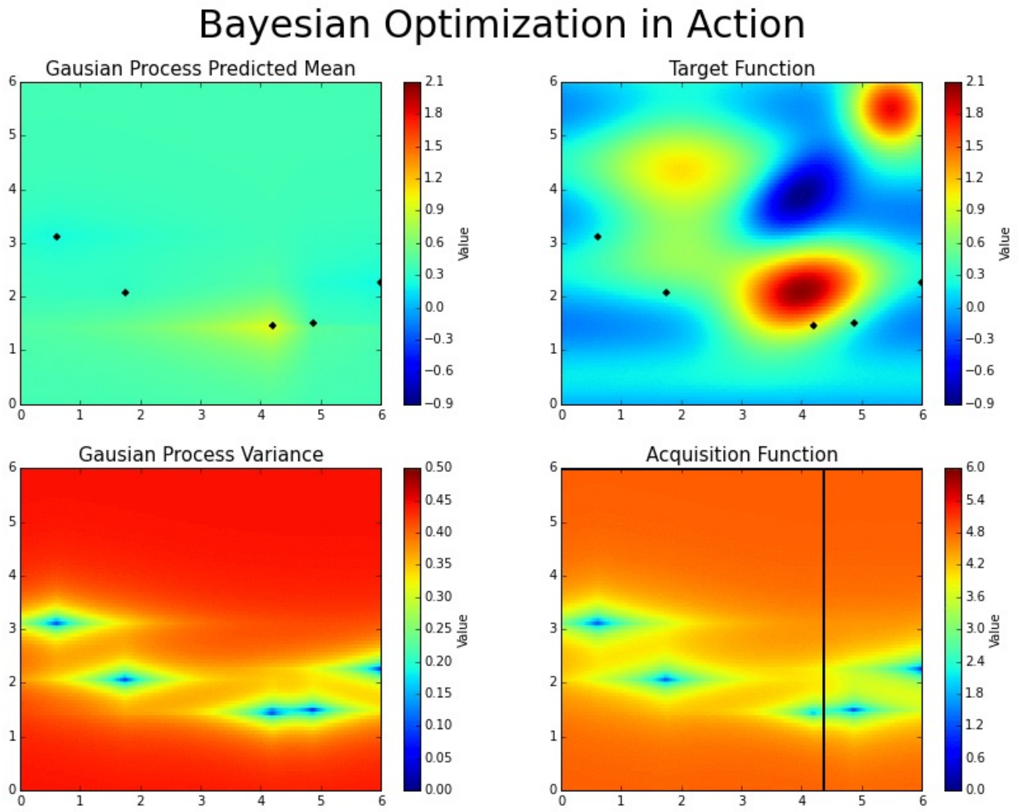


Calculating similarity or distance using multiple metrics

- cosine similarity
- Pearson correlation
- Spearman correlation
- Kendall correlation
- Manhattan distance
- Canberra distance
- Hamming distance
- Chebyshev distance
- Bray–Curtis dissimilarity



Bayesian parameter tuning



Spotify evaluation metrics: AP

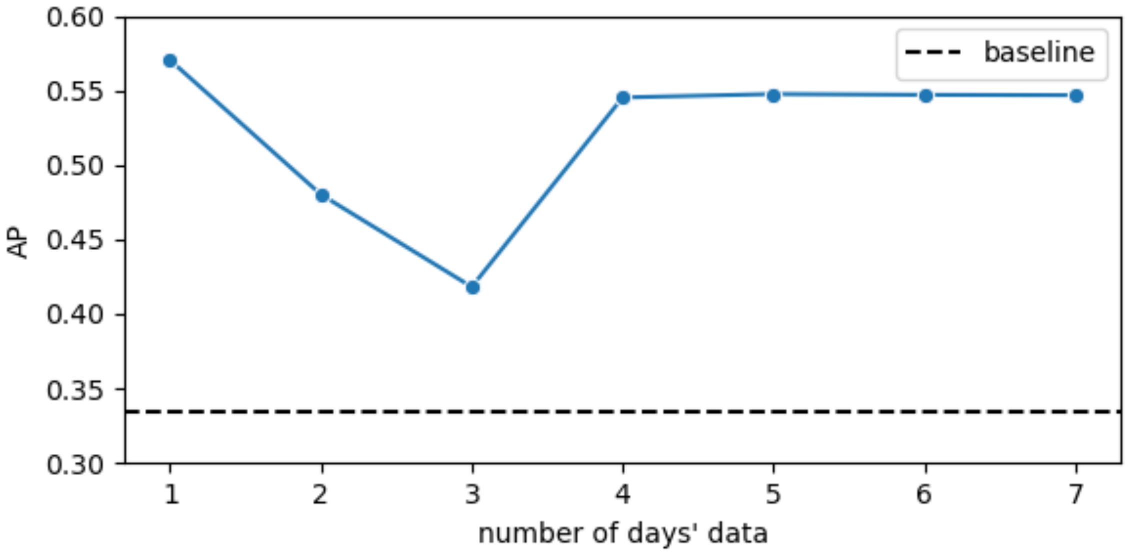
AP puts more weights on the earlier tracks than the later tracks

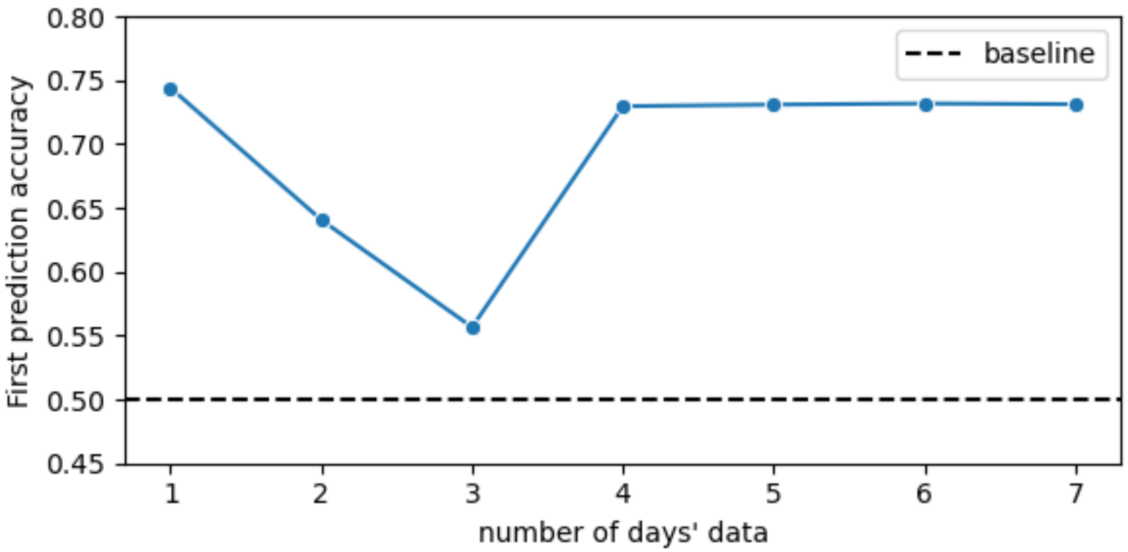
- [1,1,1,0,1] → AP = 0.910
- [0,1,1,1,1] → AP = 0.543

$$AP = \frac{1}{n}\sum_{i=1}^{n}\frac{\sum_{j=1}^{i} y_j}{i}$$

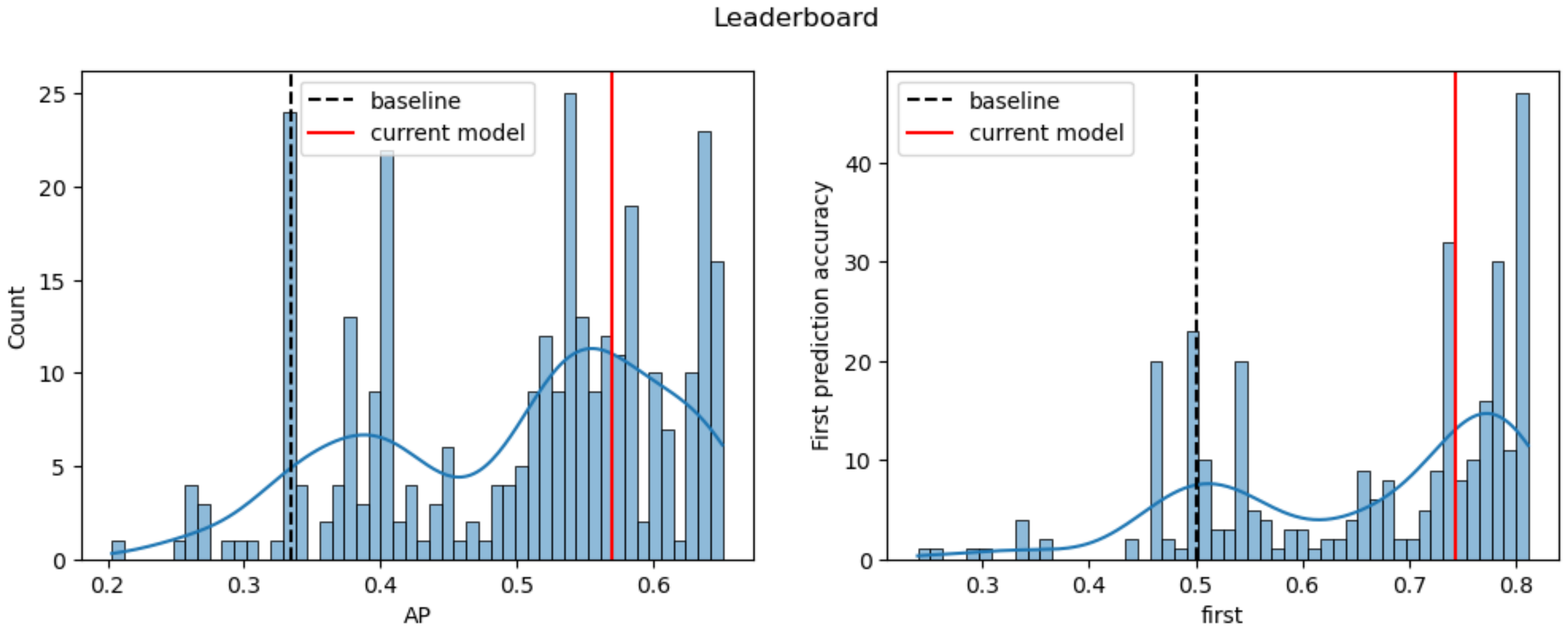
The data of how many days should be included to train a model?

1 day!

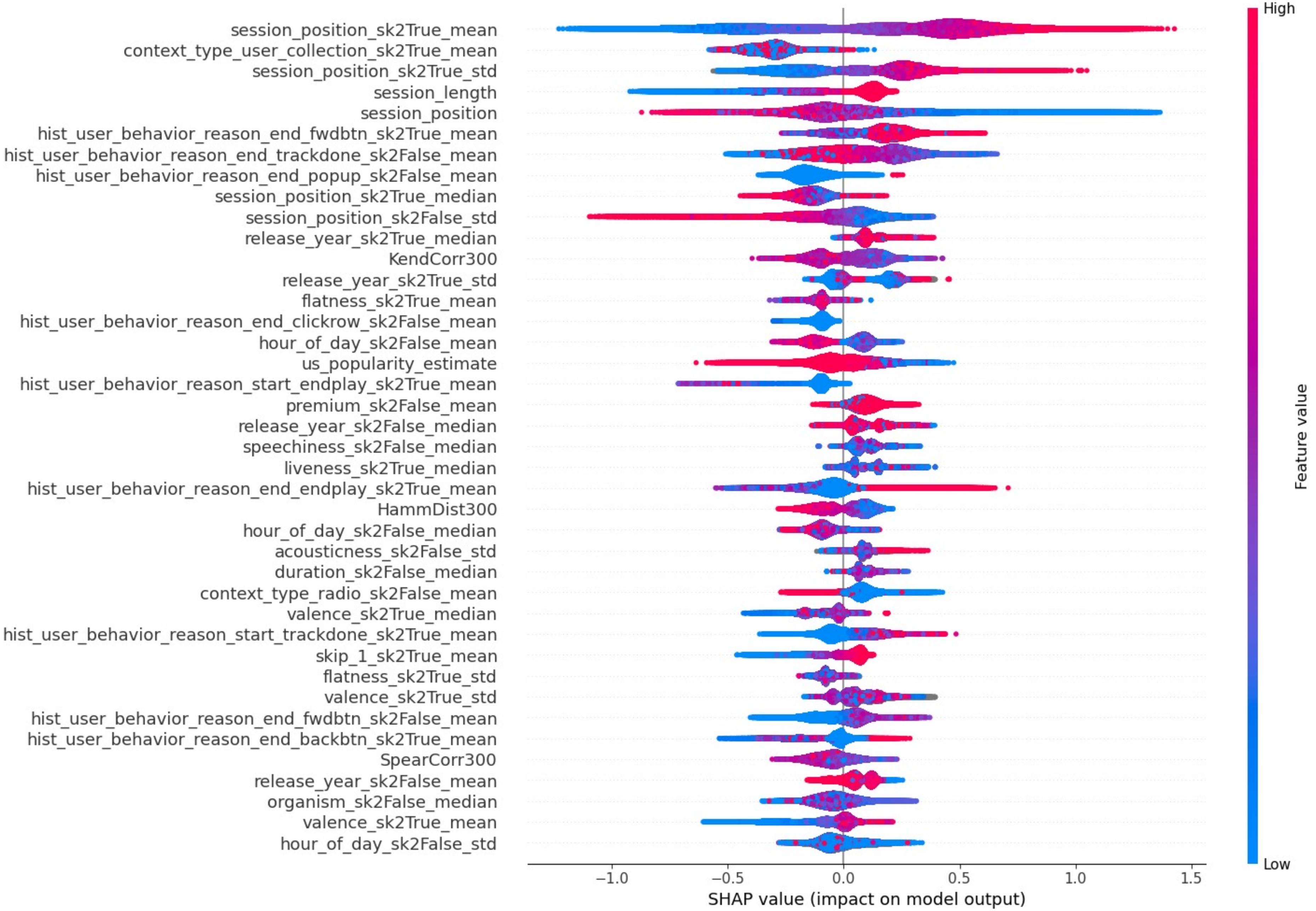




Compared to the leaderboard



What features contribute to the prediction?



How to minimize skipping rate?

- Build a more precise recommender system based on multiple similarity/distance metrics
- Recommend the tracks which have higher general popularity.

Conclusions

- The right data is more useful than big data
- Proper data engineering on the right data guided by domain knowledge can be more cost effective
- Users' listening behavior can change from day to day, the data from a day before can make the best prediction