

EvaIRS @ CIKM 2022

A Rounded Evaluation of Recommender Systems

Oct. 2022, CIKM, Atlanta

EvalRS at a glance

EvalRS 101

- *EvalRS* was designed as the first-of-its-kind data challenge: models are asked to be not just accurate, but also fair and robust.
- *EvalRS* is based on three principles:
 - **Point-wise metrics are not enough:** operationalize *desiderata* in code
 - **Build in the open:** with the community, for the community
 - **Real-world impact:** re-usable artifacts, reasonable compute



The Team

EvalRS is brought to you with ❤️ by researchers from industry (SPC, Coveo, Microsoft, NVIDIA) and academia (Stanford, Bocconi, NYU).



Jacopo Tagliabue



Patrick John Chia



Federico Bianchi



Tobias Schnabel



Giuseppe Attanasio



Gabriel de Souza P. Moreira

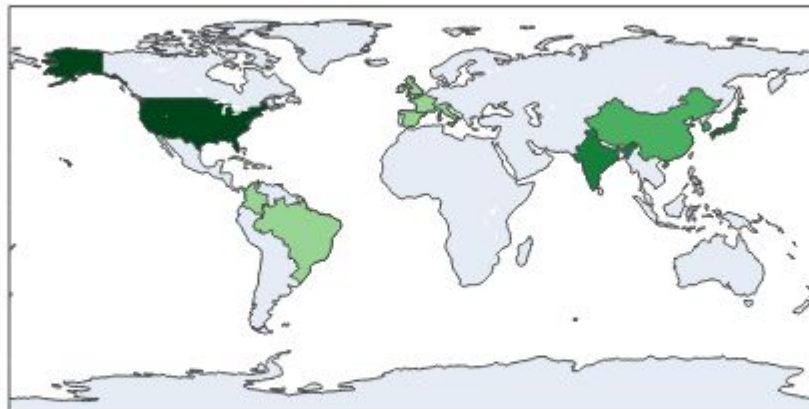


Ciro Greco



Participants

- >130 participants:
 - 63% from industry, 37% from academia
 - 14 countries
 - Tech-first companies of all sizes (e.g. Coveo, NVIDIA, etc.), and traditional ones (e.g. Fidelity, Mastercard)
- > 50 teams
- 770 total submissions across two phases



RecList for EvalRS

EvalRS is possible thanks to our **RecList** library
and its fantastic sponsors: Comet, Neptune and
Gantry!



```
from reclist.datasets import CoveoDataset
from reclist.recommenders.prod2vec import CoveoP2VRecModel
from reclist.reclist import CoveoCartRecList

coveo_dataset = CoveoDataset()

model = CoveoP2VRecModel()
model.train(coveo_dataset.x_train)

# instantiate rec_list object
rec_list = CoveoCartRecList(
    model=model,
    dataset=coveo_dataset
)

# invoke rec_list to run tests
rec_list(verbose=True)
```

If you haven't already, show your appreciation by
adding a ★ to the RecList repository!

Our Speakers



Prof. Dietmar Jannach

Talk: Multi-Objective Recommender Systems

Q&A: Federico Bianchi



Prof. Michael Ekstrand

Talk: Do You Want To Hunt A Kraken? Mapping and Expanding Recommendation Fairness

Q&A: Jacopo Tagliabue

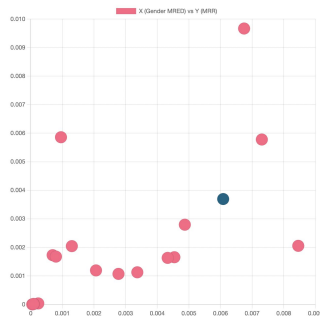


Building better RecSys

The Challenge

- Build a song recommender that is not just accurate, but fair and robust
 - *Fair*: does performance change between user sub-groups?
 - *Robust*: when the recommendation is wrong, is it still reasonable?
- Three types of metrics:
 - IR metrics on the test set
 - IR metrics by important slices
 - Behavioral metrics
- It turns out, **it is pretty hard!**

MRR and Gender MRED

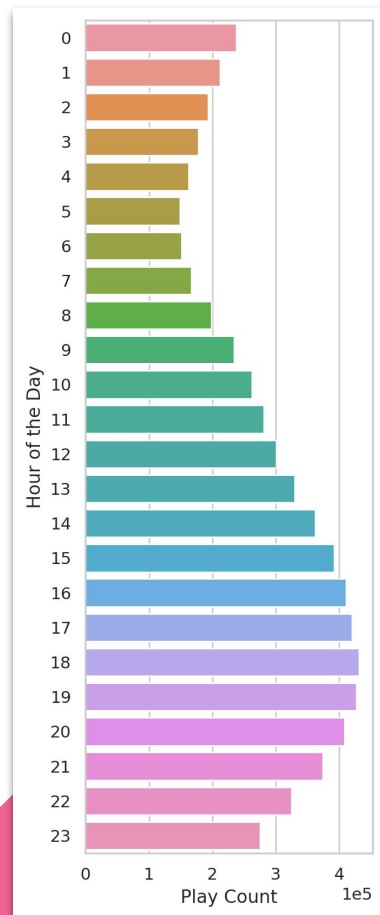
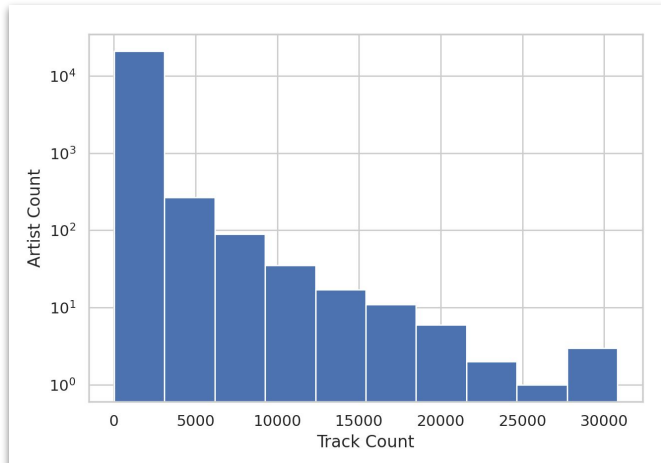
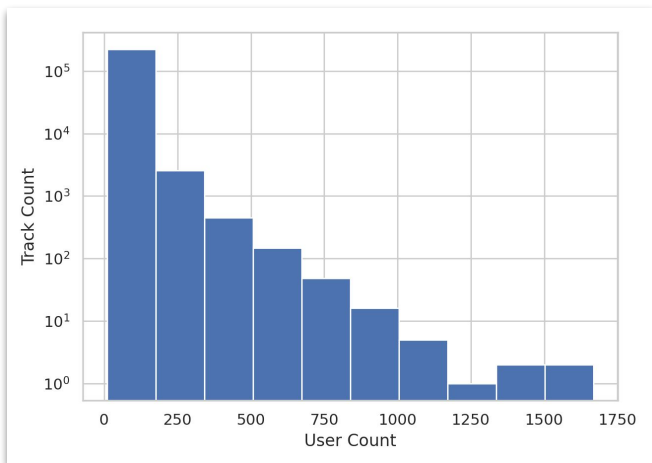


MRR and Country MRED



The Dataset

- We prepared a cleaned version of LFM-1b
 - 120K users, 820K tracks, 38M listening events
 - Fold bootstrapping evaluation



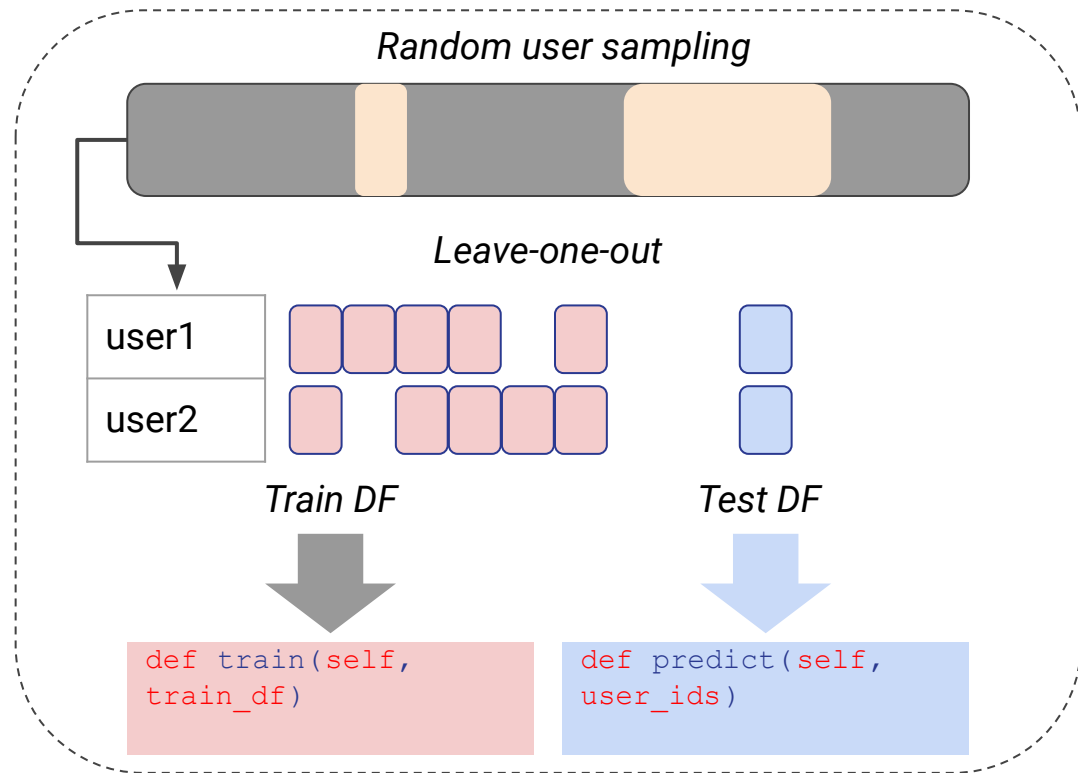
Power laws motivate fairness in underrepresented groups!

Our Methodology

- **Challenge #1:** dataset is public, so we need to avoid overfitting to a known test set.
- **Challenge #2:** a rounded evaluation needs to harmonize for the leaderboard the three types of tests. How do we combine individual scores in a way that is intuitive and actionable?

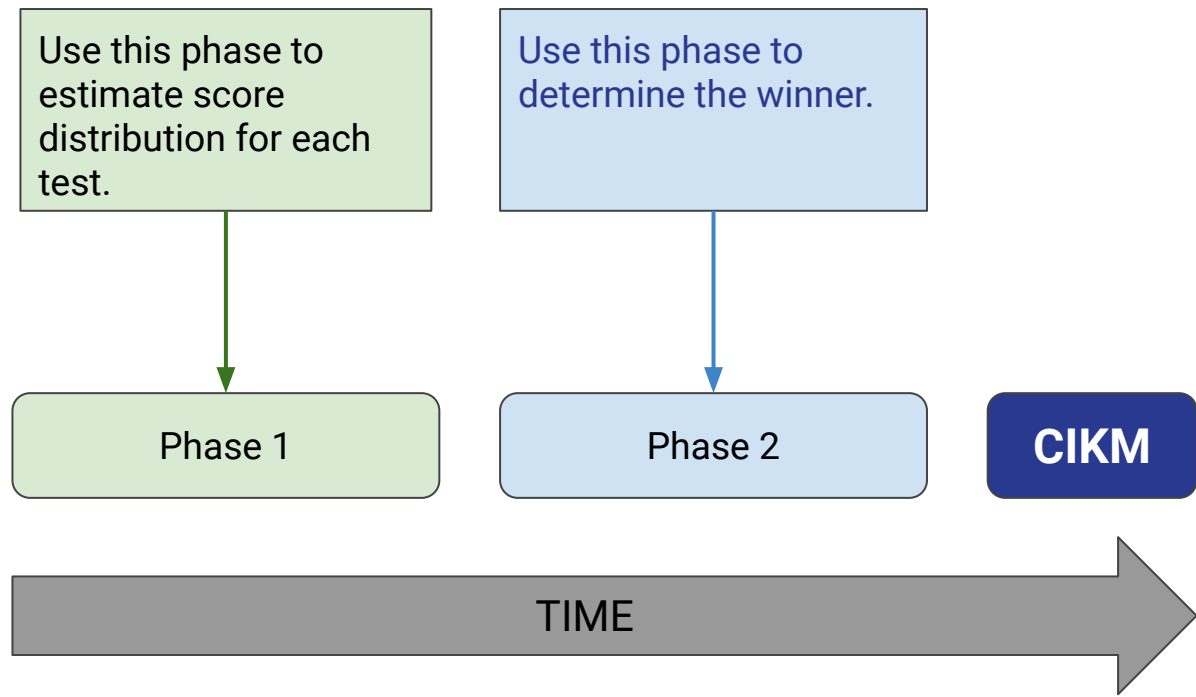


Our Methodology



- **Challenge #1:** repeat 4 times the train / test split, with leave-one-out for each user.

Our Methodology



- **Challenge #2:** use a first phase to calculate the tests' individual distribution, and use the weighting scheme in the second phase.



And the winner is...

Student Awards



Wei-Wei Du



Flavio Giobergia



Wei-Yao Wang



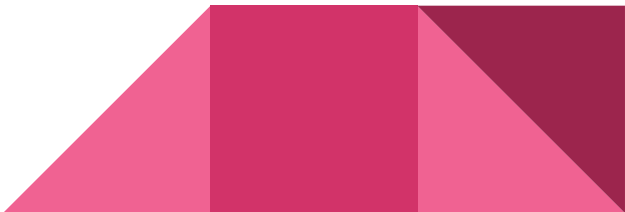
Jinhyeok Park



Dain Kim



coveo™



Leaderboard Awards

- **First position:** lyk team, Yankai Liu, Junlan Feng, Chao Deng and Haitao Zeng
- **Second position:** ML team, Jinhyeok Park, Dain Kim and Dongwoo Kim

# ^{↑↓}	Name	Score [↕]	Hit Rate [↕]	MRR [↕]	Country (MRED) [↕]	User Act. (MRED) [↕]	Track Pop. (MRED) [↕]	Artist Pop. (MRED) [↕]	Gender (MRED) [↕]	Being Less Wrong [↕]	Later Divers
1	lyk	1.702570	0.015484	0.005859	-0.004070	-0.006932	-0.002044	-0.001688	-0.000956	0.424817	-0.1216
2	ML	1.552977	0.016065	0.001727	-0.003727	-0.002913	-0.002307	-0.001047	-0.000692	0.363927	-0.2964
3	fgiobergia	1.330388	0.015565	0.001677	-0.004036	-0.003504	-0.004444	-0.000867	-0.000797	0.281863	-0.2725
4	wwweiwei	1.184669	0.021619	0.002044	-0.005366	-0.004417	-0.003191	-0.001542	-0.001299	0.320594	-0.3173
5	Sunshine	1.138580	0.018819	0.001071	-0.005213	-0.005174	-0.005043	-0.001234	-0.002774	0.280727	-0.2444


Special Awards

Best Paper

Item-based Variational Auto-encoder for Fair Music Recommendation, by Jinhyeok Park, Dain Kim and Dongwoo Kim.

Committee comment:

“The paper is solid, well argued and reports interesting findings that lead to fairer recommendations and are valuable for the community (e.g. that item-based VAEs outperform the more widespread user-based architecture). The authors also capture the spirit of this data challenge proposing an interesting addition to the RecList suite with a new evaluation metric for fairness which does point out real shortcomings of the evaluation method chose for EvalRS.”



Special Awards

Best Test

Variance agreement, by Flavio Giobergia

Committee comment:

“We evaluated each team's proposed test under the following criteria: *originality, relevance, and code quality*. According to these criteria, we believe the best new test is the *Variance Agreement* from team *fgiobergia*.”






What's next?

Looking Back, Going Forward

Lessons learned

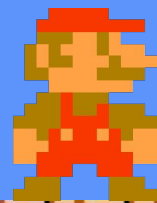
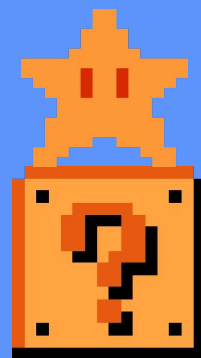
- Building rounded systems is *hard*.
- Evaluating rounded systems is even *harder*.

What's next

- **Practically:** wrap-up EvalRS and disseminate the artifacts in the community. Don't ask what RecList can do for you, but...
 - **Philosophically:** trading off different “tests” seems to some extent arbitrary: can we make it less so?
- 

Check out [RecList on Github](#) and give us a star!

Wanna contribute to the project? Get in touch!





See you, RecList cowboys