Move fast and test things

Well-rounded evaluation for recommendation engines

Jacopo Tagliabue

ADKDD, 07/08/2023

Ciao!

Serial Entrepreneur

- Founder of Tooso, acquired by TSX:CVO
- Led AI at Coveo from growth to IPO
- Now building Bauplan!

R&D at Reasonable Scale

- 25+ papers in 3 years on ML/NLP/IR (best paper at NAACL 21)
  - Collaborations with Stanford, NVIDIA, Mozilla, Farfetch, Microsoft, etc.
- Organizer of SIGIR eCommerce and EvalRS
- Adj. Prof. of MLSys at NYU

Open source

- ~2k ⭐ in open source projects
- Released 3 massive e-commerce IR datasets
- Trained the 1st industry-aware CLIP, FashionCLIP (~500k downloads in 3 months!)

It takes a (distributed) village

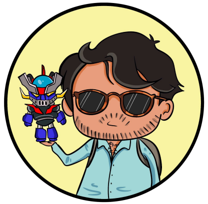
- While I am the only speaker today, Patrick-John, Federico, Chloe and Ciro (and others, unfortunately without a chibi) share with me the credit for whatever value these ideas may have.
- **Obviously, all the remaining mistakes are theirs** 😁

Jacopo

Patrick John

Federico

Chloe



Ciro

Testing ML systems is **hard**

(2d20 malus for IR)


Brenan Keller
@brenankeller

A QA engineer walks into a bar. Orders a beer. Orders 0 beers. Orders 99999999999 beers. Orders a lizard. Orders -1 beers. Orders a ueicbksjdhd.

First real customer walks in and asks where the bathroom is. The bar bursts into flames, killing everyone.

4:21 PM · Nov 30, 2018 · Twitter for iPhone

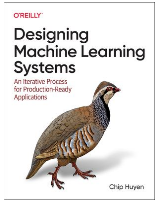
IR is everywhere

watch next

RecSys

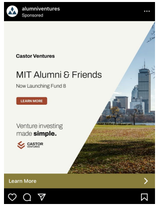
read next

Sponsored Search



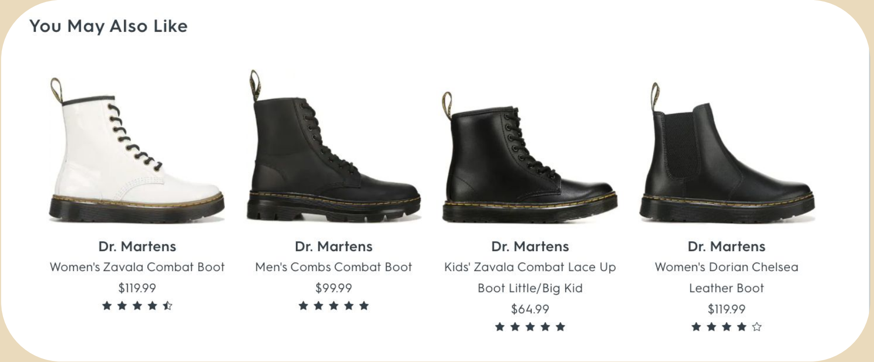
buy next

Digital Ads



38% of users stop shopping if shown non-relevant recommendations.*

*According to people selling RecSys APIs.



70% of people receive irrelevant ads once a month.*

*According to what I googled on my plane.


The New York Times

ADVERTISEMENT

USAA HOMEOWNERS INSURANCE

USAA

NO CLAIMS YES SAVINGS

Why Are You Seeing So Many Bad Digital Ads Now?

Scrolling past ads has rarely been enjoyable. But in recent months, people say the experience seems so much worse.

Testing Matters*

*According to an editorial citing Tagliabue *et al* (2022).

nature machine intelligence

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature machine intelligence › editorials › article

Editorial | Published: 23 February 2023

Algorithmic recommendations, anyone?

Why do we even test?

1. Generalization
2. Model comparison

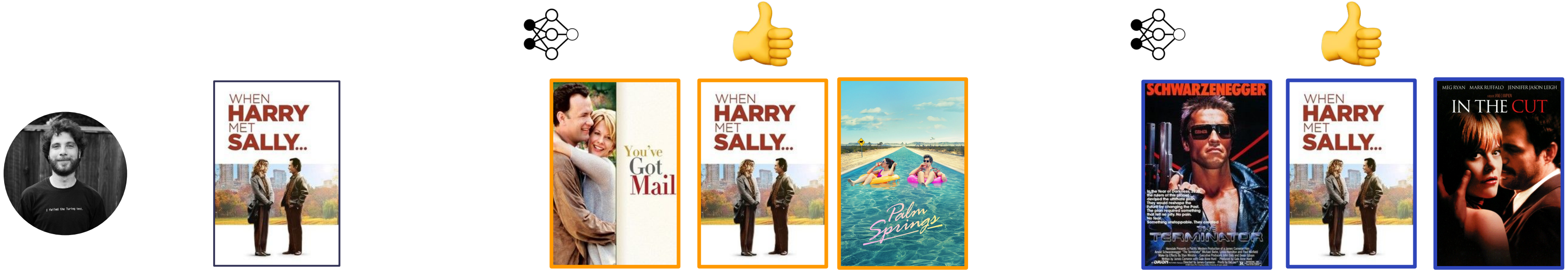
Testing checklist

1. Create a train / test split and train a model
2. Loop over test cases with ground truths and count “successes”
3. Make a decision based on the final number: Model A KPI is 0.424, B is 0.41, *therefore* A > B

Input Model Predictions G. Truth Eval

Lies, big lies, IR metrics

1. Create a train / test split and train a model
2. Loop over test cases with ground truths and count “successes”
3. Make a decision based on the final number: Model A KPI is 0.424, B is 0.41, *therefore* A > B

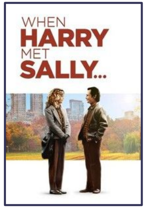


Input G. Truth Predictions Model A Predictions Model B

The importance of being “less wrong”

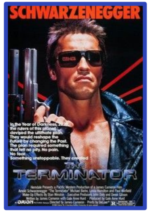
1. Create a train / test split and train a model
2. Loop over test cases with ground truths and count “successes”
3. Make a decision based on the final number: Model A KPI is 0.424, B is 0.41, *therefore* A > B

Input



G. Truth

Model A



Model B

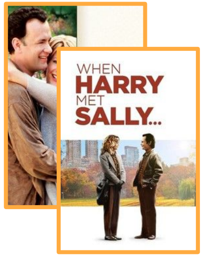
Model C

Hit Rate is deceitful above all things

1. Create a train / test split and train a model
2. Loop over test cases with ground truths and count “successes”
3. Make a decision based on the final number: Model A KPI is 0.424, B is 0.41, *therefore* A > B

50 / 100 Hits

7 / 10 Hits

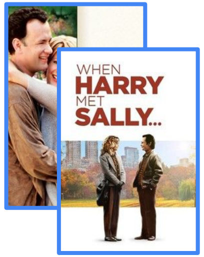


HR 57/100: 0.52

vs

56 / 100 Hits

1 / 10 Hits



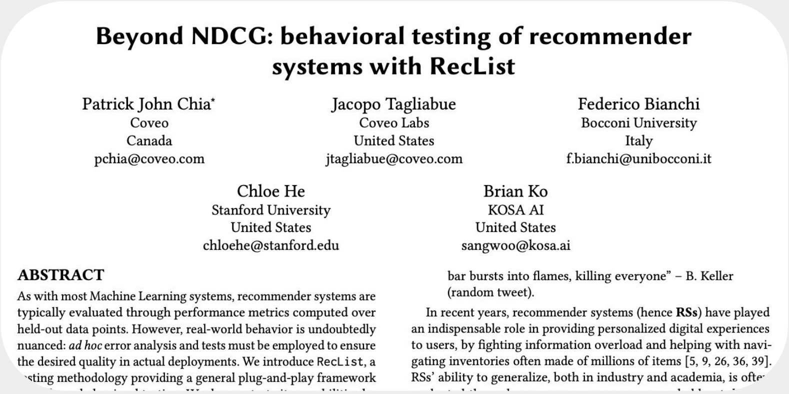
HR 57/100: 0.52

Why do we even test?

1. ~~Generalization~~
2. ~~Model comparison~~

Testing Re-Imagined*

* In theory, there is no difference between theory and practice. In practice, there is.



Beyond NDCG: behavioral testing of recommender systems with RecList

Patrick John Chia*
Coveo
Canada
pchia@coveo.com

Jacopo Tagliabue
Coveo Labs
United States
jtagliabue@coveo.com

Federico Bianchi
Bocconi University
Italy
f.bianchi@unibocconi.it

Chloe He
Stanford University
United States
chloehe@stanford.edu

Brian Ko
KOSA AI
United States
sangwoo@kosa.ai

ABSTRACT

As with most Machine Learning systems, recommender systems are typically evaluated through performance metrics computed over held-out data points. However, real-world behavior is undoubtedly nuanced: *ad hoc* error analysis and tests must be employed to ensure the desired quality in actual deployments. We introduce `RecList`, a testing methodology providing a general plug-and-play framework

bar bursts into flames, killing everyone" – B. Keller (random tweet).

In recent years, recommender systems (hence **RSs**) have played an indispensable role in providing personalized digital experiences to users, by fighting information overload and helping with navigating inventories often made of millions of items [5, 9, 26, 36, 39]. RSs' ability to generalize, both in industry and academia, is ofte

From point-wise to bin-wise metrics

- Instead of reporting just HR over the full distribution, report HR per frequency!

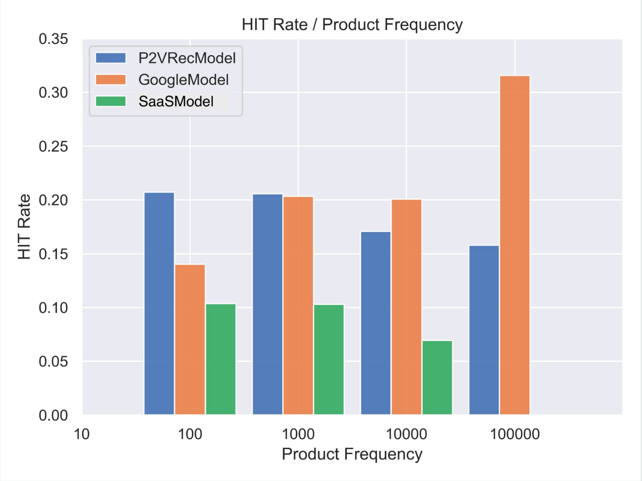


Table 1: Results for a complementary *RecList*.

Test	P2V	GOO	S1
HR@10	0.197	**0.199**	0.094
MRR@10	0.091	**0.102**	0.069
Coverage@10	1.01e-2	**1.99-e2**	3.00e-3
Popularity Bias@10	**9.91e-5**	1.41e-4	1.20e-4

From point-wise to category-wise metrics

- Instead of reporting just HR over the full distribution, report HR per item type!

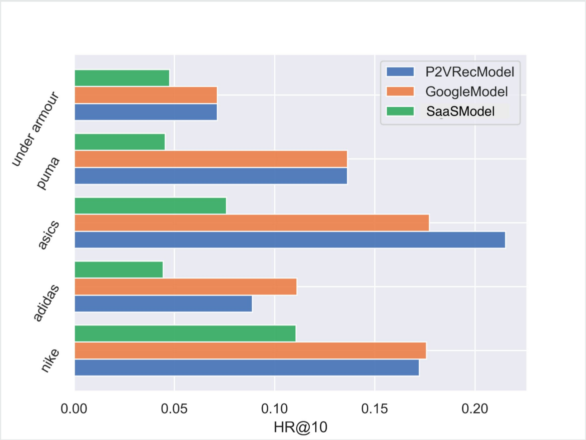
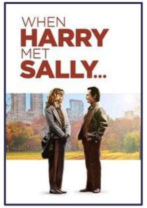


Table 1: Results for a complementary *RecList*.

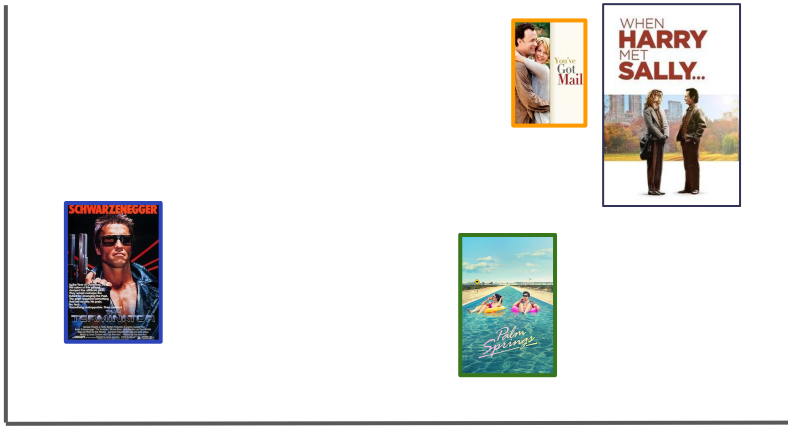
Test	P2V	GOO	S1
HR@10	0.197	**0.199**	0.094
MRR@10	0.091	**0.102**	0.069
Coverage@10	1.01e-2	**1.99-e2**	3.00e-3
Popularity Bias@10	**9.91e-5**	1.41e-4	1.20e-4

How do we know when something is “less wrong”?

- Ideally, we would just ask people to provide similarity judgements for the mistakes! But that does not scale:
    - Use representational learning and approximate relevance as distance in the underlying space (2022);
    - Ask a Large Language Model (2023).



Input G. Truth



Meet RecList*

* Talk is cheap, show me the code

```python
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)
```

The RecList project

- <u>RecList</u> spawned a popular open source package, the CIKM 2022 data challenge, the EVALRS23@KDD workshop, and three papers.

nature machine intelligence

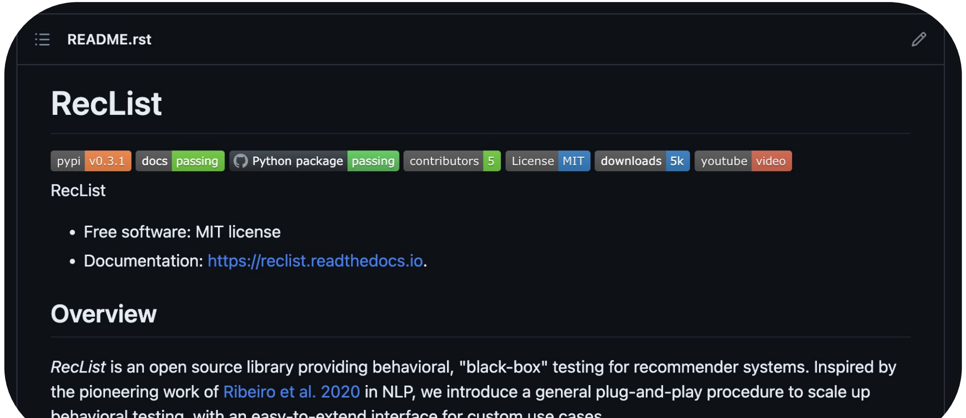
Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature › nature machine intelligence › challenge accepted › article

Challenge Accepted | Published: 26 January 2023

A challenge for rounded evaluation of recommender systems

Jacopo Tagliabue, Federico Bianchi, Tobias Schnabel, Giuseppe Attanasio, Ciro Greco, Gabriel de Souza Moreira & Patrick John Chia



README.rst

RecList

pypi v0.3.1 | docs passing | Python package passing | contributors 5 | License MIT | downloads 5k | youtube video

RecList

- Free software: MIT license
- Documentation: https://reclist.readthedocs.io.

Overview

RecList is an open source library providing behavioral, "black-box" testing for recommender systems. Inspired by the pioneering work of Ribeiro et al. 2020 in NLP, we introduce a general plug-and-play procedure to scale up behavioral testing, with an easy-to-extend interface for custom use cases.

README.md

EvalRS-KDD-2023

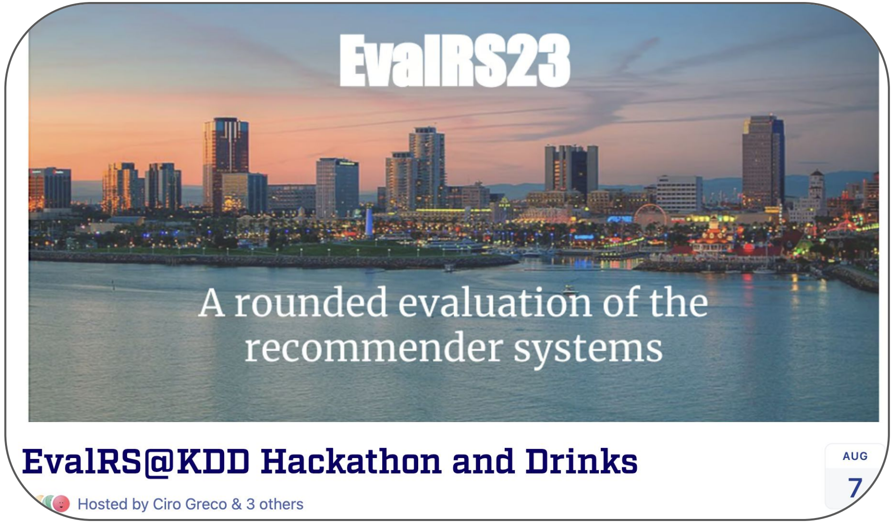
Official Repository for EvalRS @ KDD 2023, the Second Edition of the workshop on well-rounded evaluation of recommender systems.

Open in Colab

EvalRS @ KDD: papers, hackathon, party

- 2 keynotes
- 5 talks
- 1 new open dataset
- $2500 in hackathon prizes
- Unlimited* drinks

* Conditions apply!

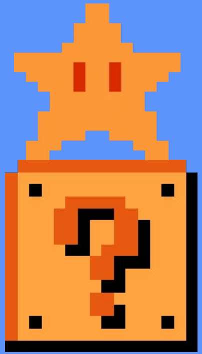


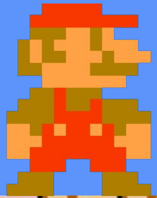
https://reclist.io/kdd2023-cup/

Come for the papers,
stay for the drinks!

Check out / share / add a star to our open source **projects**!

Wanna work with us? Get in touch!





BAUPLAN