02

Data Handling and Processing in Recommendation Research

Data processing is a fundamental step in any recommendation task, with strategies that vary from paper to paper.

Identifying unique data processing steps is not possible, as they may strongly depend on the recommendation task or on the specific research questions being addressed.

To identify the most common approaches in state-of-the-art research, we surveyed recent papers from top-tier search and recommendation conferences.

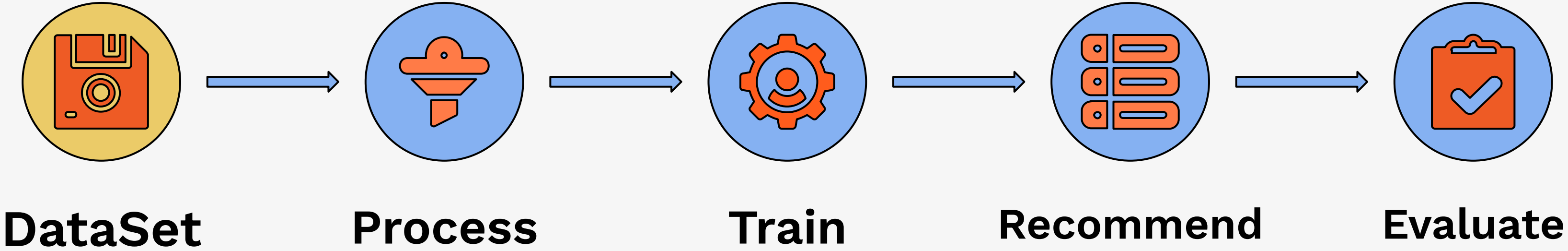
55 surveyed papers

Between 2020 and 2024

To identify the most common approaches in state-of-the-art research, we surveyed recent papers from top-tier search and recommendation conferences.

Recommendation Datasets

Offline Evaluation Pipeline



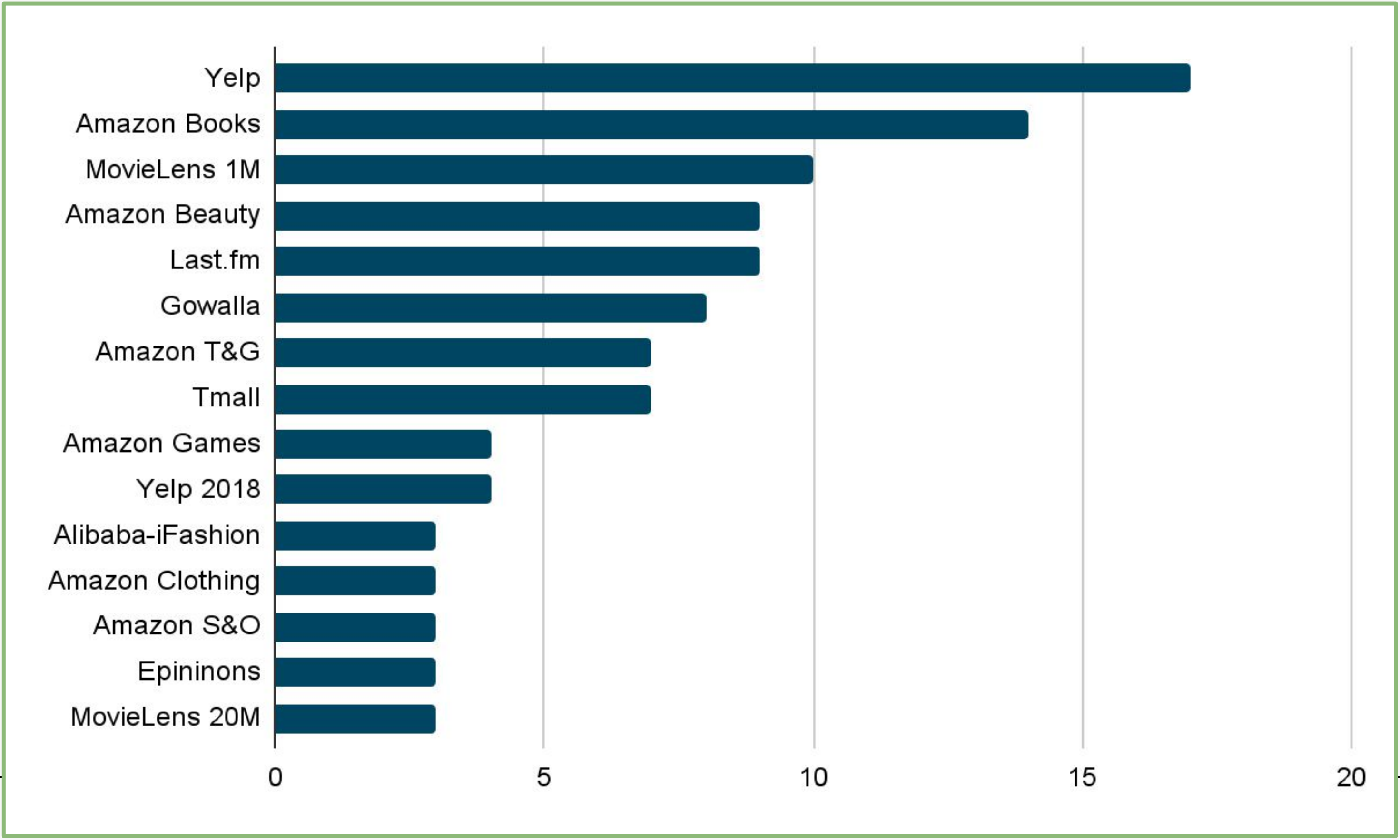
Typical pipeline for offline evaluation in recommender systems.

79 distinct datasets

over 55 survey papers

On average, each paper introduces more than one dataset that is not used in other studies.
This highlights the strong heterogeneity of recommendation datasets in offline evaluation.

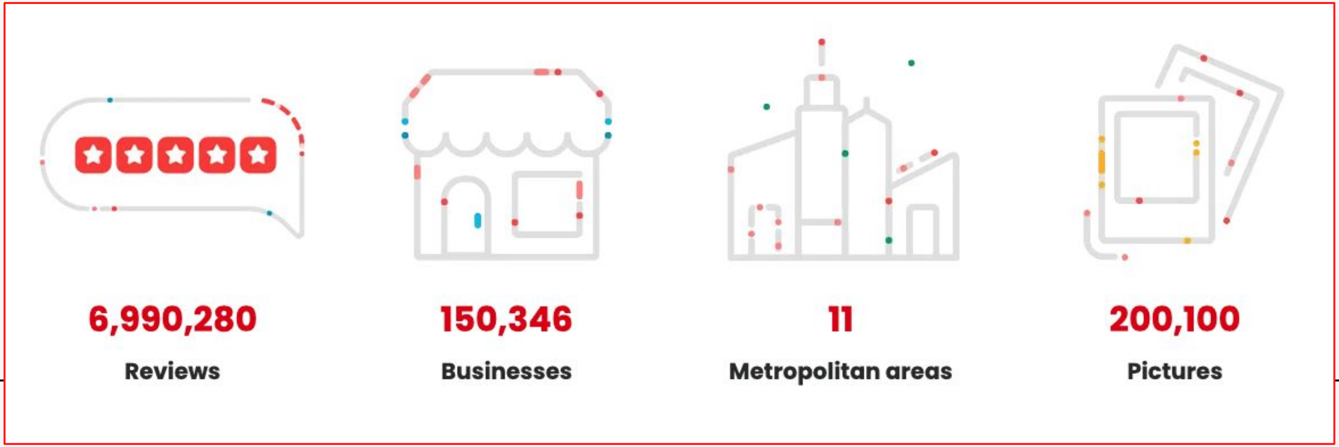
Datasets used in at least 3 papers



yelp

The Yelp Open Dataset contains information about businesses. User feedback are stored as reviews.

DATASET



6,990,280	150,346	11	200,100
Reviews	Businesses	Metropolitan areas	Pictures

{"review_id":"AqPFMleE6RsU23_auESxiA",
"user_id":"_7bHUi9Uuf5__HHc_Q8guQ",
"business_id":"kxX2SOes4o-D3ZQBkiMRfA",
"Stars":5.0,
"Useful":1,"funny":0,"cool":1,
"text":"Wow! Yummy, different, delicious.   Our favorite is the lamb curry and korma. With 10 different kinds of naan!!! Don't let the outside deter you (because we almost changed our minds)...go in and try something new!   You'll be glad you did!","
date":"2015-01-04 00:01:03"}

Amazon Reviews'23

WEBSITE

Amazon Reviews is a project started by Julian McAuley and Jure Leskovec at the SNAP laboratory at Stanford in 2013 [19].

Users’ feedback is collected from the Amazon website as textual reviews.

Each review is linked to the corresponding Amazon product ID and user ID.

The dataset is divided into 26 categories and has been updated in 2014, 2018 and 2023.

[19] McAuley & Leskovec, RecSys 2013, Hidden factors and hidden topics: understanding rating dimensions with review text

Amazon Reviews'23

Sample extracted from the Toys & Games category from the 2023 release.

user_id	parent_asin	rating	timestamp
AFKZENTNBQ7A7V7UXW5JJI6UGRYQ	B07XRSD5R9	5.0	1580949719154
AFKZENTNBQ7A7V7UXW5JJI6UGRYQ	B06XYKSKQP	3.0	1639855230760
AFKZENTNBQ7A7V7UXW5JJI6UGRYQ	B09QH7QJS7	5.0	1677939664713
AGKASBHYZPGTEPO6LWZPVJWB2BVA	B006GBITXC	3.0	1452647382000
AGKASBHYZPGTEPO6LWZPVJWB2BVA	B00TLEMSVK	4.0	1454675785000
AGKASBHYZPGTEPO6LWZPVJWB2BVA	B00SO7HF6I	3.0	1454676014000

movielens

It was firstly published in 1998 by the GroupLens group.

MovieLens is an online platform with research purposes that builds recommendations of movies.

WEBSITE

It is one of the most well-known datasets. Several papers have analyzed and discussed its properties [20, 21, 22].

It has been published in multiple versions.

[20] Harper & Konstan, TIIS 2016, The MovieLens Datasets: History and Context
[21] Fan et al., TOIS 2024, Our Model Achieves Excellent Performance on MovieLens: What Does It Mean?
[22] Forouzandeh et al., Multimedia Tools and Applications 2021, Presentation of a recommender system with ensemble learning and graph embedding: a case ...

movielens

userID	movieID	rating	timestamp
1	1193	5	978300760
1	661	3	978302109
1	914	3	978301968
1	3408	4	978300275
1	2355	5	978824291
1	1197	3	978302268

It was published in 2011 by the GroupLens group.

last.fm is an online platform that collects music preferences to produce statistics, recommendations and lets users get in touch with people with similar tastes.

WEBSITE

The dataset contains not only user preferences for artists but also user friendships and user tags. The GroupLens group is the same that released the MovieLens dataset.

last.fm

userID	artistID	weight
2	51	13883
2	52	11690
2	53	11351
2	54	10300
2	55	8983
2	56	6152

天猫

理想生活上天猫

Has been published in 2015 by the Alibaba group at the IJCAI conference as part of a conference challenge.

WEBSITE

Tmall, operated by Alibaba Group, is China’s leading B2C platform connecting local and global brands with over 500M users. It serves both as an e-commerce marketplace and a channel for brand awareness.

The Gowalla Dataset was published by the SNAP research group from Stanford in 2011 [13].

Gowalla is a location-based social network from which **6+ million check-ins** have been collected between Feb. 2009 and Oct. 2010.

It consists of two datasets: a friendship network and a collection of users' check-ins.

GOWALLA

[13] Cho et al., KDD 2011, Friendship and mobility: user movement in location-based social networks

iFashion is a part of the Taobao e-commerce website, which belongs to the Alibaba group.

WEBSITE

It was published in 2019 at KDD.

It contains not only user feedback for items but also their textual and visual descriptions.

[23] Chen et al., KDD 2019, POG: Personalized Outfit Generation for Fashion Recommendation at Alibaba iFashion

iFashion

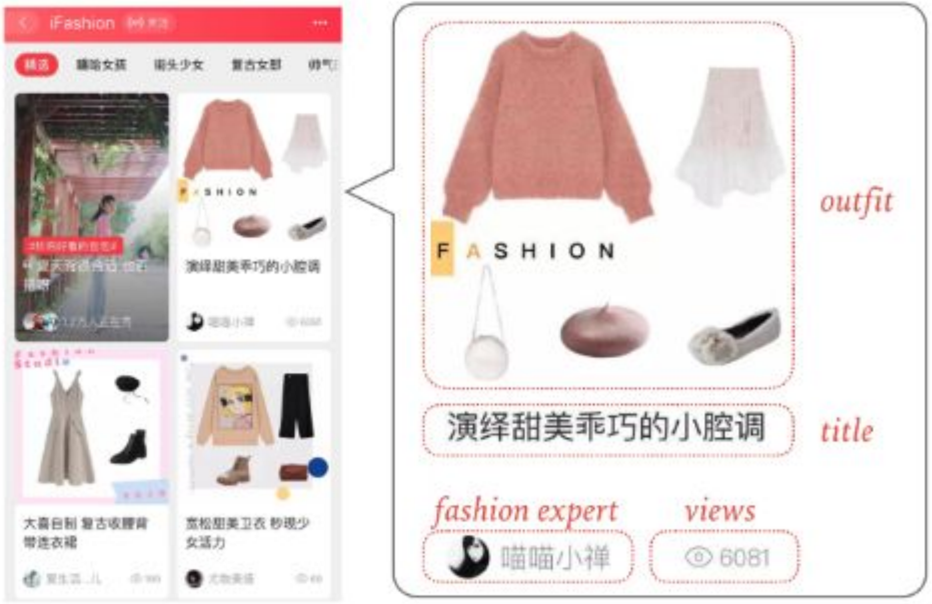


Figure 1: A sample of iFashion application in Taobao. We recommend fashion outfits (sets of fashion items which interact with each other) to users.

Epinions.com

The Epinion Dataset was published by the SNAP research group from Stanford in 2003 [13].

It was derived from the epinions.com website (no longer active).
It was a general consumer review website.

It contains trust relationships between the users of the platform.

WEBSITE

[24] Richardson et al., ISWC 2003, Trust Management for the Semantic Web

Epinions.com℠

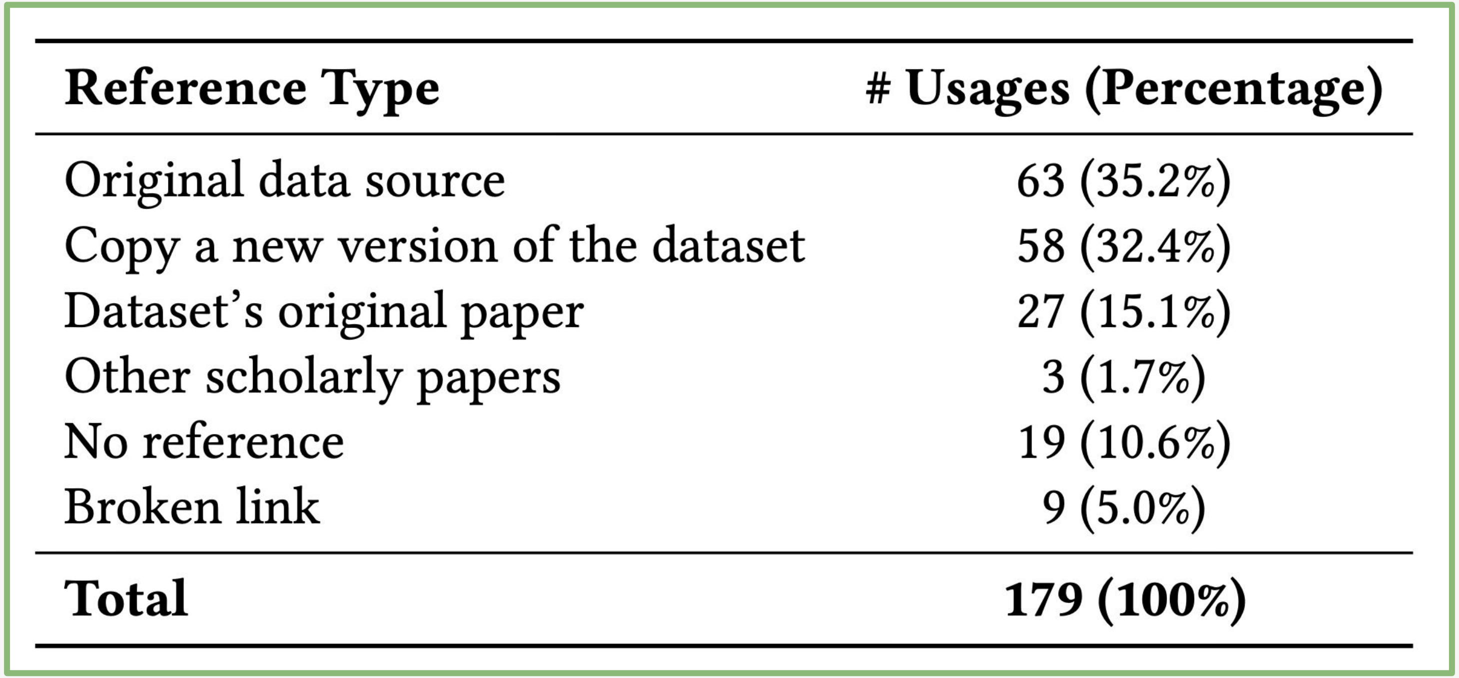
Stored as an directed graph.

FromNodeId	ToNodeId
0	4
0	5
0	7
0	8
0	9
0	10

Having such a wide variety of datasets available is evidence of the maturity of recommender systems, which can be easily adapted to different domains.

However, this makes the comparability of results more challenging and makes their sharing more fragmented, which is prone to errors (see the case of Gowalla).

For this reason, we systematically analysed how datasets are referenced and shared within scientific articles.



Reference Type	# Usages (Percentage)
Original data source	63 (35.2%)
Copy a new version of the dataset	58 (32.4%)
Dataset's original paper	27 (15.1%)
Other scholarly papers	3 (1.7%)
No reference	19 (10.6%)
Broken link	9 (5.0%)
Total	**179 (100%)**

Referencing correctly

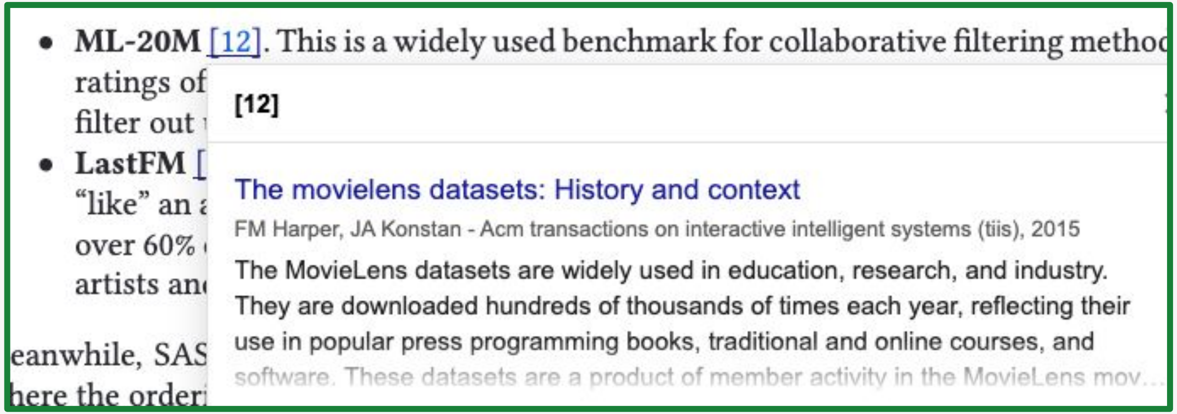
35,2% of the survey papers refer to the original data source. Authors usually attach a link to the data source in a footnote.

15,1% cite the paper associated to the dataset publication. This approach avoids ambiguities and connects datasets and papers.

50% of the analysed papers included appropriate references to the datasets.

Referencing correctly

🙂



- **ML-20M** [12]. This is a widely used benchmark for collaborative filtering methoc ratings of filter out
- **LastFM** ["like" an a over 60% artists and

eanwhile, SAS here the orderi

[12]

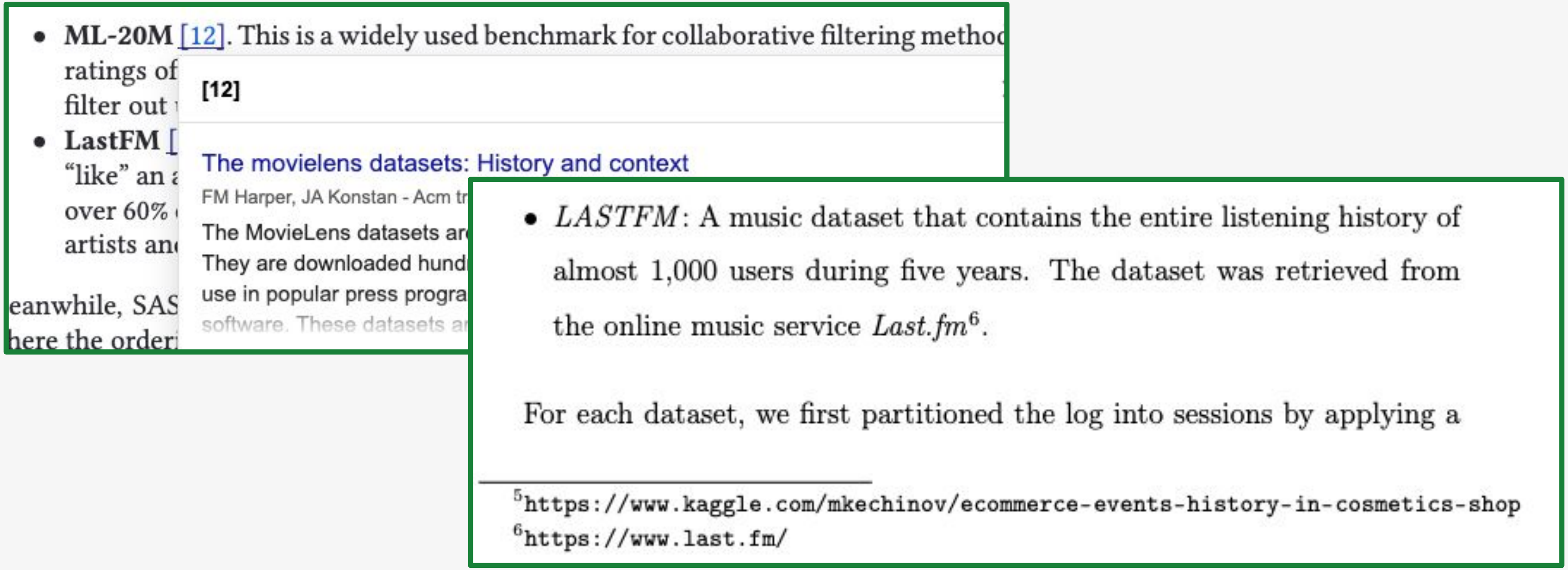
The movielens datasets: History and context

FM Harper, JA Konstan - Acm transactions on interactive intelligent systems (tiis), 2015

The MovieLens datasets are widely used in education, research, and industry. They are downloaded hundreds of thousands of times each year, reflecting their use in popular press programming books, traditional and online courses, and software. These datasets are a product of member activity in the MovieLens mov…

Referencing correctly

🙂



- **ML-20M** [12]. This is a widely used benchmark for collaborative filtering method
ratings of
filter out
- **LastFM** [
"like" an a
over 60%
artists and

eanwhile, SAS
here the orderi

[12]

The movielens datasets: History and context

FM Harper, JA Konstan - Acm tr

The MovieLens datasets are
They are downloaded hund
use in popular press progra
software. These datasets ar

- *LASTFM*: A music dataset that contains the entire listening history of almost 1,000 users during five years. The dataset was retrieved from the online music service *Last.fm*[6].

For each dataset, we first partitioned the log into sessions by applying a

[5] https://www.kaggle.com/mkechinov/ecommerce-events-history-in-cosmetics-shop

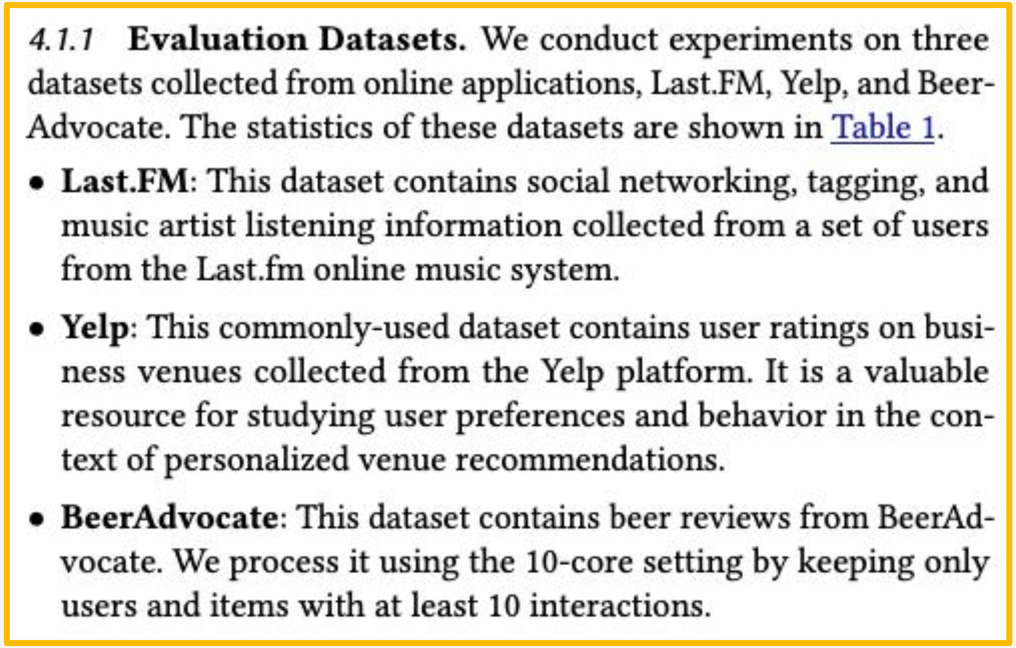
[6] https://www.last.fm/

Sharing a copy of the dataset 😐

32.4% of the papers shared a copy of the dataset. Usually, the dataset is uploaded to the repository containing the code used for the experiments.

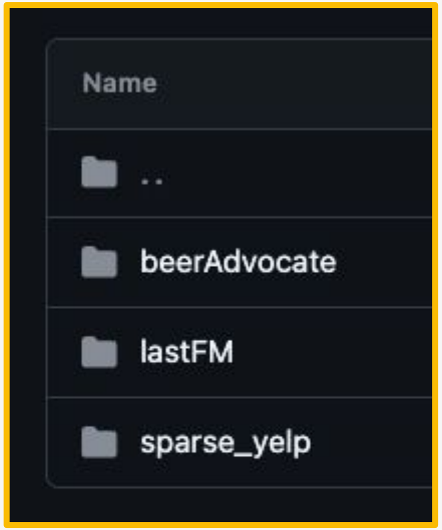
The shared dataset is often a processed or pre-split version. While this supports reproducibility, it undermines standardisation, since others may treat these versions as the original source, propagating errors.

Sharing a copy of the dataset 😐



4.1.1 **Evaluation Datasets.** We conduct experiments on three datasets collected from online applications, Last.FM, Yelp, and BeerAdvocate. The statistics of these datasets are shown in Table 1.

- **Last.FM**: This dataset contains social networking, tagging, and music artist listening information collected from a set of users from the Last.fm online music system.
- **Yelp**: This commonly-used dataset contains user ratings on business venues collected from the Yelp platform. It is a valuable resource for studying user preferences and behavior in the context of personalized venue recommendations.
- **BeerAdvocate**: This dataset contains beer reviews from BeerAdvocate. We process it using the 10-core setting by keeping only users and items with at least 10 interactions.



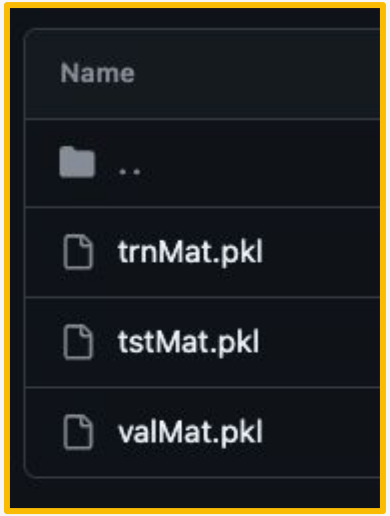
Name

..

beerAdvocate

lastFM

sparse_yelp



Name

..

trnMat.pkl

tstMat.pkl

valMat.pkl

Irretrievable datasets ☹

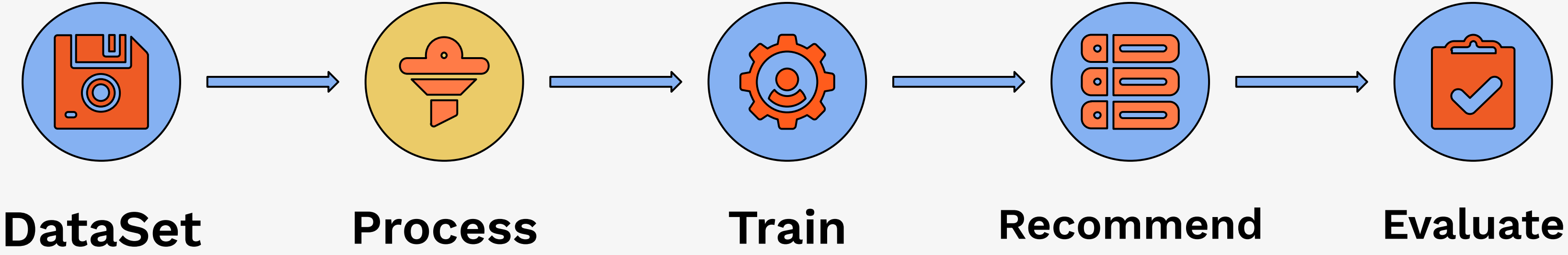
15,6% of the datasets are irretrievable due to the absence of any reference or due to broken links.

This highlights the importance of tracking and versioning recommendation datasets.

Our recommendation is to refer to the dataset paper (when available), to the data source, and to share the processing pipeline with the code used for the experiments.

Data Processing

Offline Evaluation Pipeline



Typical pipeline for offline evaluation in recommender systems.

Rarely in recommender systems research are datasets used for training without any pre-processing.

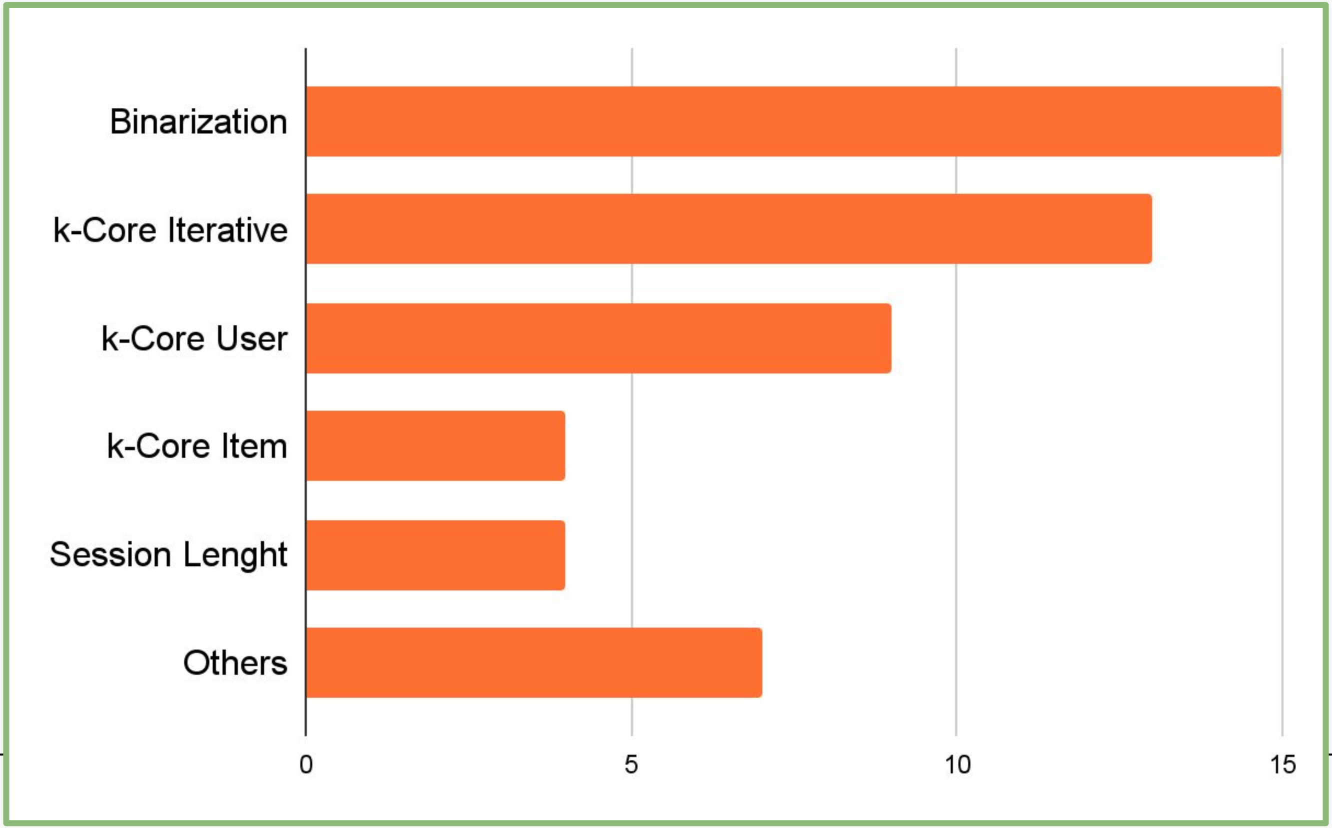
55% surveyed papers

declared to have processed datasets

Data transformation is an essential step to align the dataset with the experimental design.

Due to their key role, their standardisation and reproducibility is fundamental to ensuring a correct and comparable evaluation.

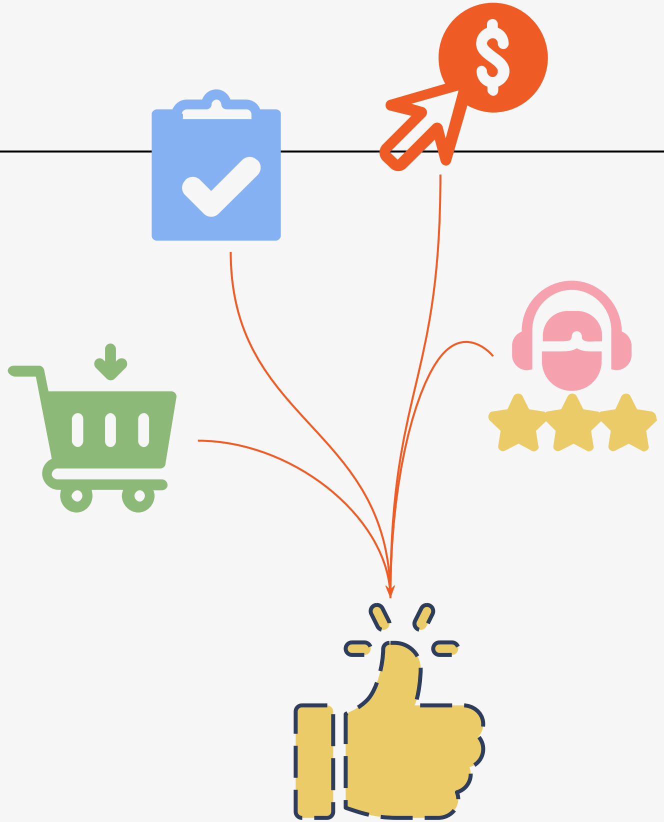
Most frequently used processing techniques.



Binarization

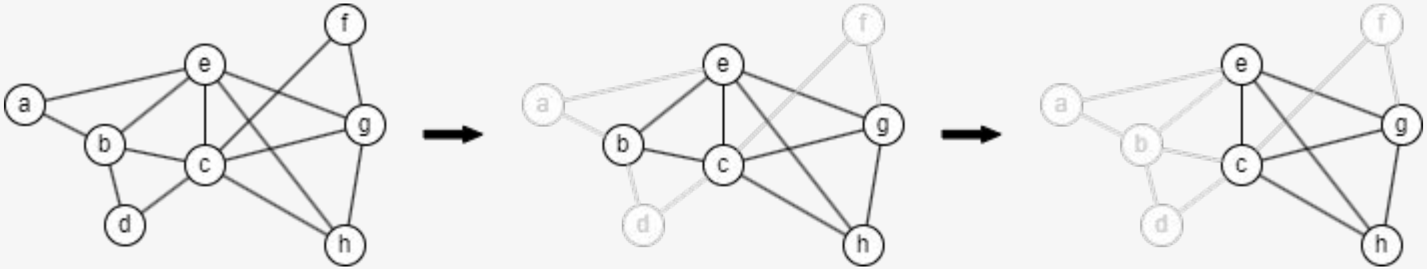
Binarization is the most common approach.
This could be related to the large number of papers that study implicit feedback recommendation.

Binarization consists of associating an event, a rating or a review to an implicit feedback of a user to an item.



k-Core

The k-Core algorithm finds the largest subgraph of a graph where all the nodes have a degree of at least k.

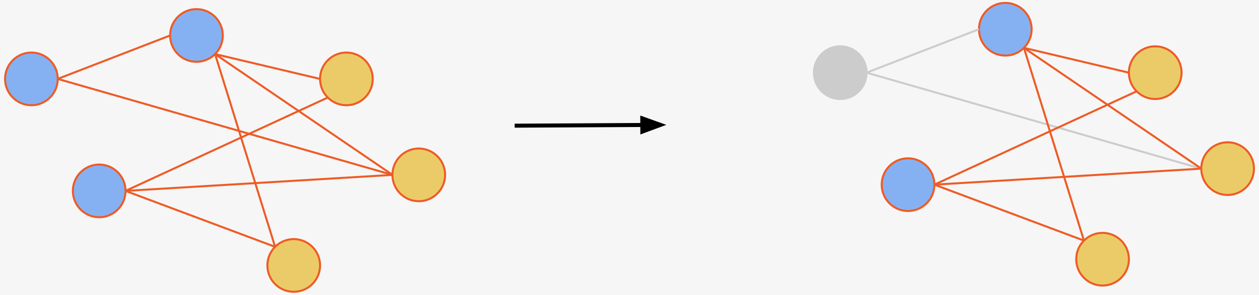


Since recommendation data can be interpreted as a bipartite undirected graph, it can also be applied in recsys.

Image source: https://www.ultipa.com/docs/graph-analytics-algorithms/k-core

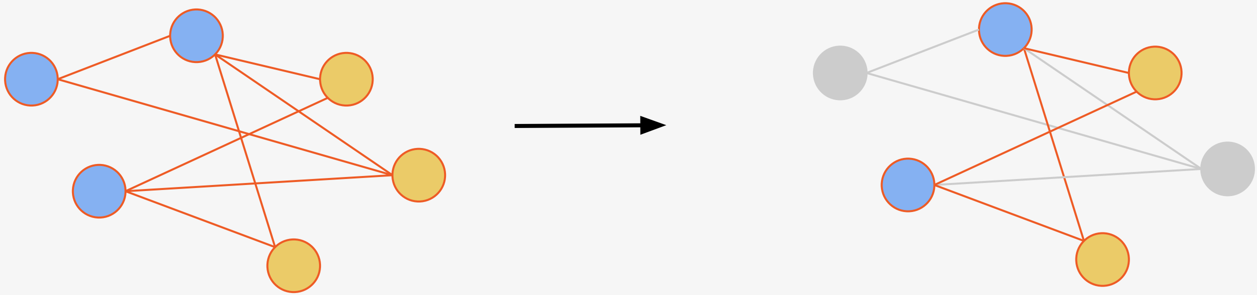
k-Core

The k-Core algorithm has been adapted in recommender systems to a user- and item-based k-Core.



Instead of retaining all the nodes with at least k node degree, it focusses on retaining only user/item nodes with a node degree greater than or equal to k.

k-Core



A common mistake is to perform a k-core on users and items in a non-iterative way.

There are no guarantees that the node degree will remain the same when nodes are removed.

Iterative k-Core

When applying the k-core to both user and item nodes, it is necessary to adopt the iterative version. Without using the iterative algorithm, there are no guarantees that node degrees will be preserved.

The k-Core is used as a strategy to filter out cold users and cold items.

Session Length

In session-based and sequential recommendation, filtering strategies differ from those used in other domains.

Sessions (sequences) may be retained according to a minimum or maximum length, or, for example, only events occurring within a specific time window may be retained.

Data Splitting

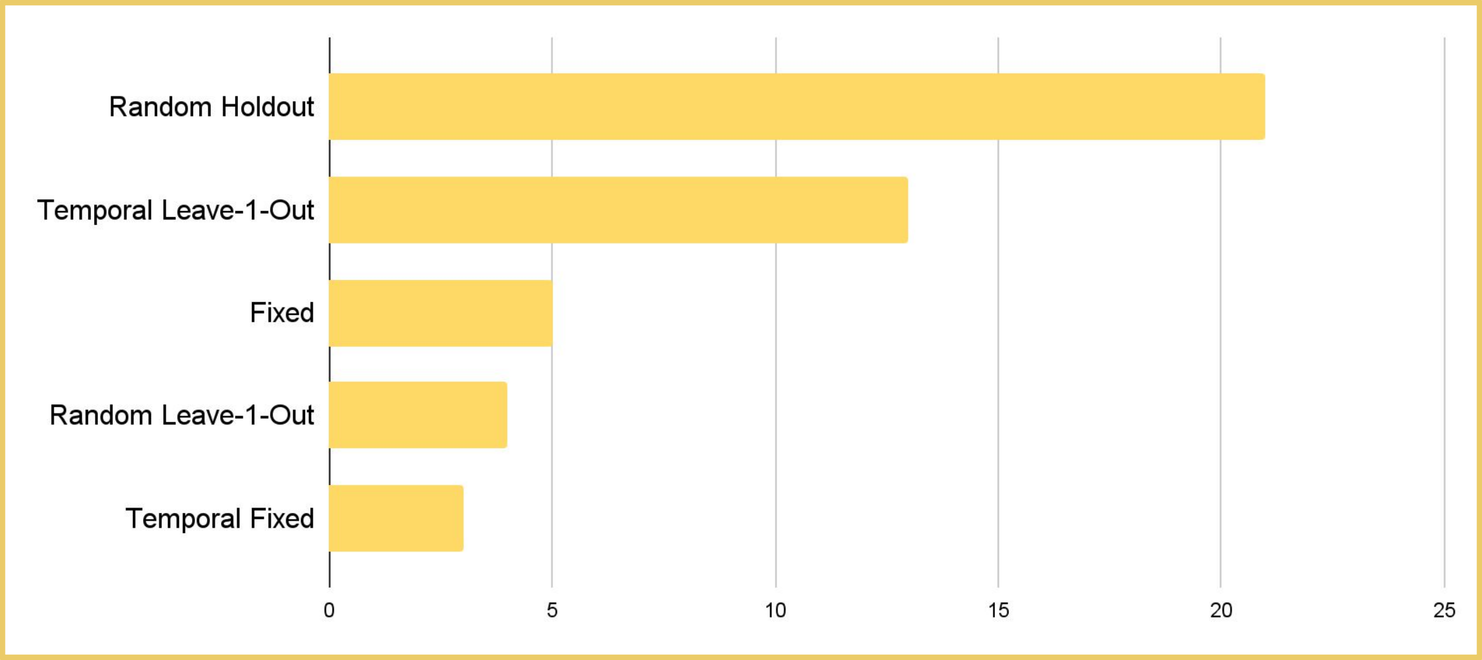
Data splitting consists of partitioning a dataset for training, evaluating and testing models.

Our survey confirms results already published [23, 24]: there is no universally accepted splitting strategy.

The adoption of different splitting strategies between studies contributes to creating inconsistent rankings, even when the same metrics and datasets are adopted.

[23] Meng et al., RecSys 2020, Exploring Data Splitting Strategies for the Evaluation of Recommendation Models
[24] Sun, SIGIR 2023, Take a Fresh Look at Recommender Systems from an Evaluation Standpoint

Most frequently used splitting strategies



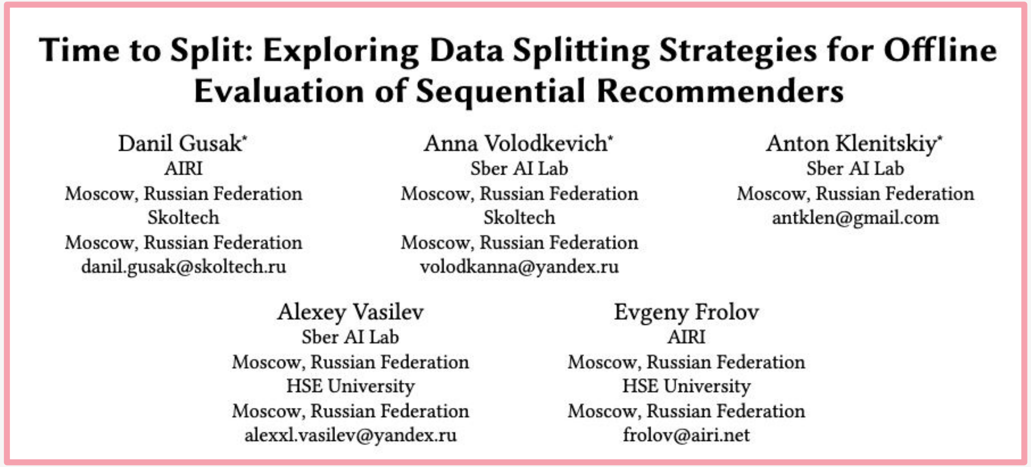
Random Hold-out is the most common approach: due to its randomness, there is no guarantee of comparability of results.

For this reason, some scholars use fixed splits. This is a good way to ensure comparability, provided that the splits are shared and accessible to everyone.

Temporal splitting with a global timestamp is, from our perspective, the best option since it is easy to replicate and does not lead to information leakage.

Selecting the best splitting strategy is still an open question.

Apart from data leakage, another aspect to consider is the consistency of the model performance with real-life scenarios.



Time to Split: Exploring Data Splitting Strategies for Offline Evaluation of Sequential Recommenders

Danil Gusak*
AIRI
Moscow, Russian Federation
Skoltech
Moscow, Russian Federation
danil.gusak@skoltech.ru

Anna Volodkevich*
Sber AI Lab
Moscow, Russian Federation
Skoltech
Moscow, Russian Federation
volodkanna@yandex.ru

Anton Klenitskiy*
Sber AI Lab
Moscow, Russian Federation
antklen@gmail.com

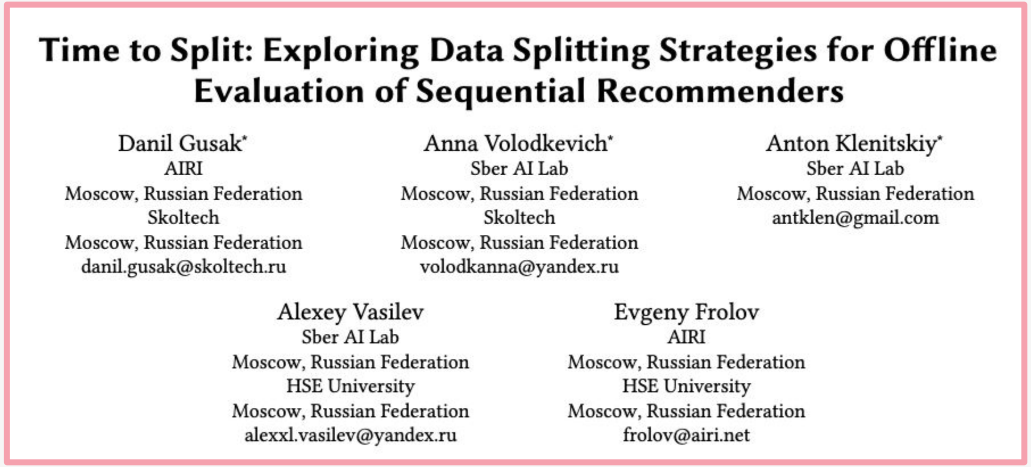
Alexey Vasilev
Sber AI Lab
Moscow, Russian Federation
HSE University
Moscow, Russian Federation
alexxl.vasilev@yandex.ru

Evgeny Frolov
AIRI
Moscow, Russian Federation
HSE University
Moscow, Russian Federation
frolov@airi.net

[25] Gusak et al., RecSys 2025, Time to Split: Exploring Data Splitting Strategies for Offline Evaluation of Sequential Recommenders

Selecting the best splitting strategy is still an under-discussion question.

Apart from data leakage
u
c
c
m
w



Time to Split: Exploring Data Splitting Strategies for Offline Evaluation of Sequential Recommenders

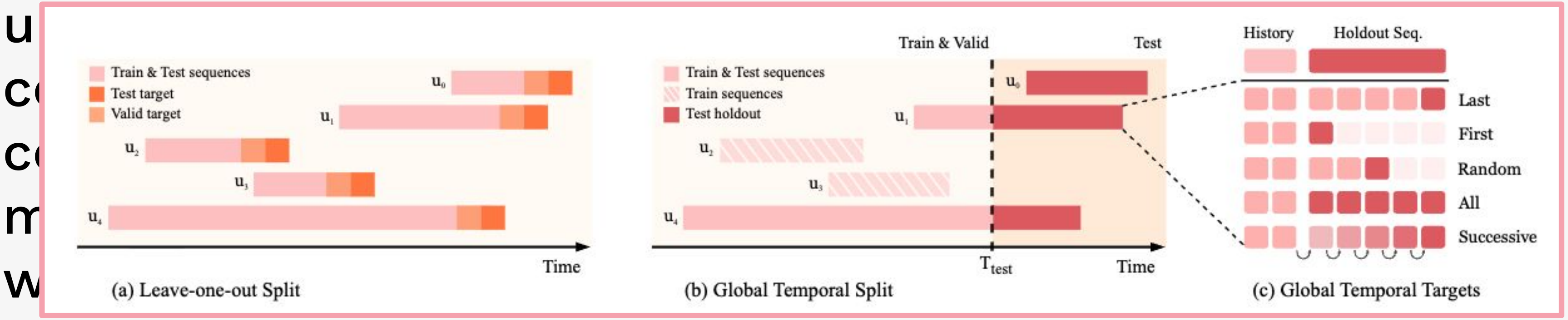
Danil Gusak*
AIRI
Moscow, Russian Federation
Skoltech
Moscow, Russian Federation
danil.gusak@skoltech.ru

Anna Volodkevich*
Sber AI Lab
Moscow, Russian Federation
Skoltech
Moscow, Russian Federation
volodkanna@yandex.ru

Anton Klenitskiy*
Sber AI Lab
Moscow, Russian Federation
antklen@gmail.com

Alexey Vasilev
Sber AI Lab
Moscow, Russian Federation
HSE University
Moscow, Russian Federation
alexxl.vasilev@yandex.ru

Evgeny Frolov
AIRI
Moscow, Russian Federation
HSE University
Moscow, Russian Federation
frolov@airi.net



(a) Leave-one-out Split (b) Global Temporal Split (c) Global Temporal Targets

[25] Gusak et al., RecSys 2025, Time to Split: Exploring Data Splitting Strategies for Offline Evaluation of Sequential Recommenders