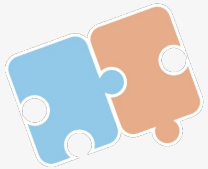
04

Introduction to Datarec



What is DataRec?

DataRec [5] is an open-source Python library aimed at streamlining, unifying, and simplifying data preparation pipelines in recommender systems.

[5] Mancino et al., SIGIR 2025: “DataRec: A Python Library for Standardized and Reproducible Data Management in Recommender …”

Why DataRec?

Integration

Works with standalone projects and reproducibility frameworks

Simplicity

Unifies and streamlines data processing pipelines

Evaluation

Prepares datasets for offline recommendation evaluation

Reproducibility

Ensures traceability, versioning, and consistent pipelines

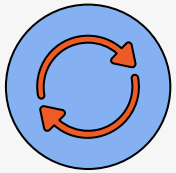
DataRec Components

Data Model

Processing Module

Splitting Module

Interoperability



Reproducibility

Tabular

user	item	rating	timestamp
1	10	5.0	1580
1	20	3.0	1639
2	30	5.0	1677

Inline

1, 10, 20, 30
2, 30
3, 10, 40
4, 10, 20, 50

JSON


```
[{"user": 1,
"item": 20,
"rating": 5.0,
"timestamp":
1580},
...]
```


Each data format ha a dedicated reader and writer to maximize compatibility.

RawData Class

The RawData class serves as a unified interface for data input and output.

Decouples I/O from dataset storage formats, allowing flexible handling of different data sources.

DataRec Class

The core class of the DataRec library, all data is represented as a DataRec object.

Built on Pandas DataFrame, enhanced to integrate with other DataRec modules and pipelines.

A DataRec object contains dataset characteristics.

#users	#items	#transactions
space_size	shape	density
gini_user	gini_item	
ratings_per_user	ratings_per_item	

Think of a DataRec as a dataset with superpowers!

data + metrics + reproducible pipelines

all in one object.

Built-in Datasets

The most common recommendation datasets are built into DataRec, with automated download, traceability, and versioning.

DataRec implements the most common preprocessing techniques from recommender system literature, consistently transforming one DataRec object into another.

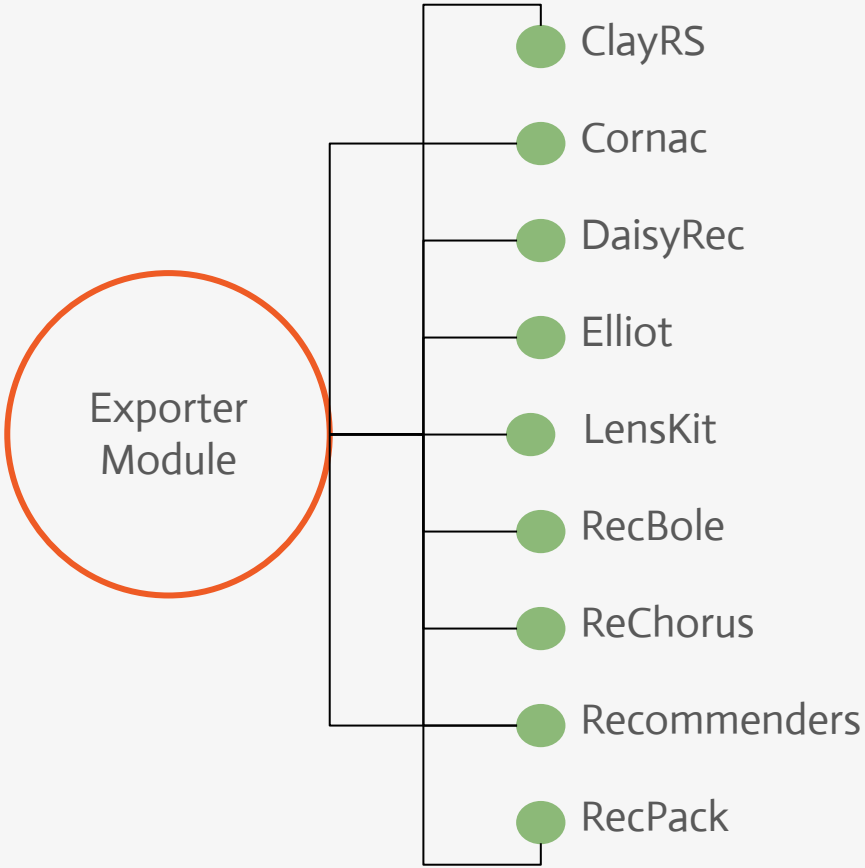
DataRec → *Preprocessing* → *DataRec*

K-core

Binarization

⋮

DataRec implements the most common splitting strategies from recommender system literature, producing separate DataRec objects for training, validation, and test.



DataRec ensures interoperability with popular recommendation frameworks through a dedicated export module, which outputs datasets in a compatible format.

DataRec automatically records all transformations, keeping a complete history for transparent and reproducible data processing.

pipeline.yml

```yaml
pipeline:
- name: load
  operation: MovieLens
  params:
    version: 1m
- name: process
  Operation: UserItemIterativeKCore
  params:
    cores: 5
- name: split
  operation: UserStratifiedHoldOut
  params:
    seed: 42
    test_ratio: 0.25
    val_ratio: 0.25
- name: export
  operation: Elliot
  params:
    output_path: ./elliot
```

Re-run Share

The transformation history can be exported as a YAML configuration.

This enables easy sharing, ensures consistency, and guarantees reproducibility.