Reel Good

Movie Recommender System

G4

Alireza Hatami

Caitlin Dunne

Jaz Zhou

Precious Orekha

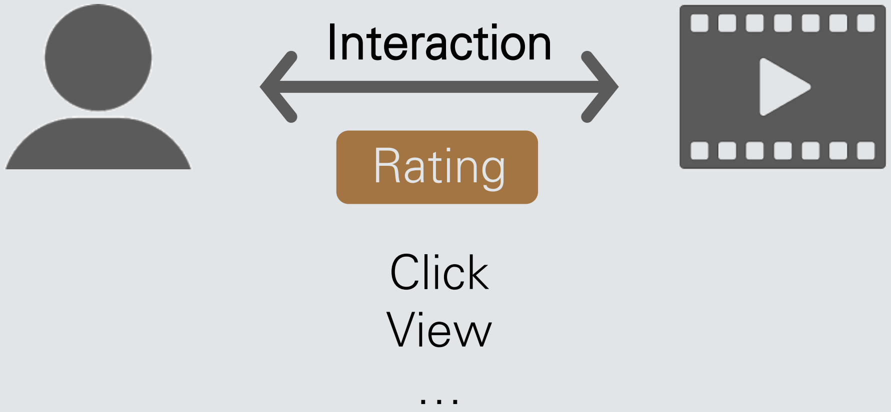
Image Source: Go on a thrill ride with action-packed movies like Extraction, The Gray Man, RRR, and The Old Guard.
. Accessed on [Feb 6 2025].
Available at [https://dnm.nflximg.net/api/v6/2DuQlx0fM4wd1nzqm5BFBi6ILa8/AAAAQRC29H19twWKcTZ9Zpg4biJbGNaHF2GGIYNeZ6fvwugUJbuKxTjjMFPCS-y5P3ZePL57rupDtSkyUIJhv3P8leMJGMzszuG2CHNd65NwWPu5LeKxQkRNfNMHmxAwt7tmQZFk1VIrBd1aXr2AR5DM.jpg?r=5b1]

and interconnect everything.

Reel Good Movie Recommender System

Highlighted Features

Algorithm Choice – Collaborative Filtering

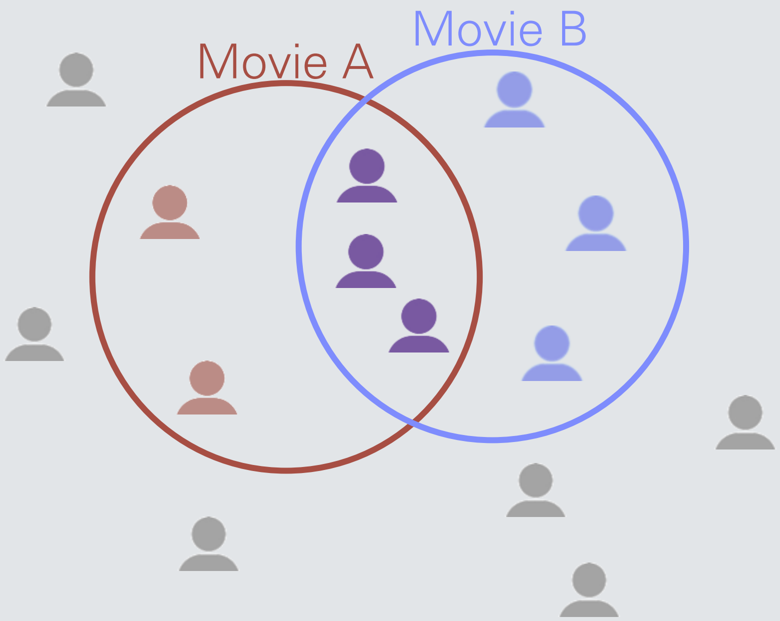


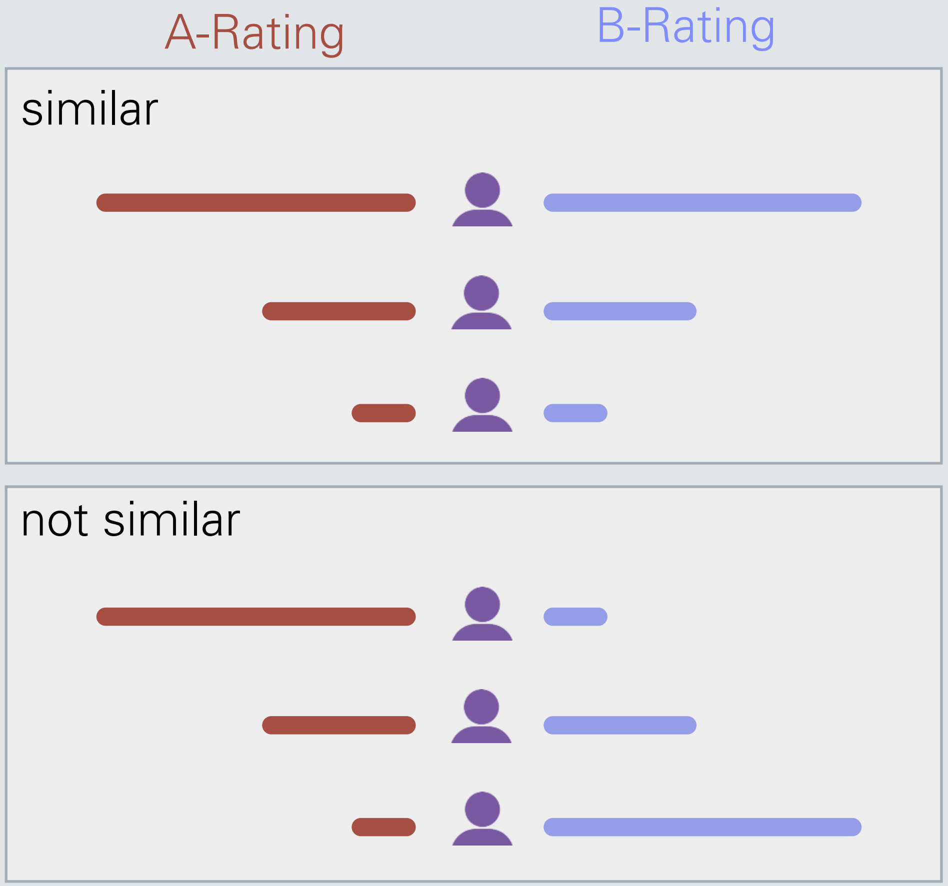
- Domain-free: we don’t need to construct the profiling.
- No privacy concern: we don’t need explicit user profile.

Algorithm Choice – KNN

- Recommend movies that are similar to the ones you already like.

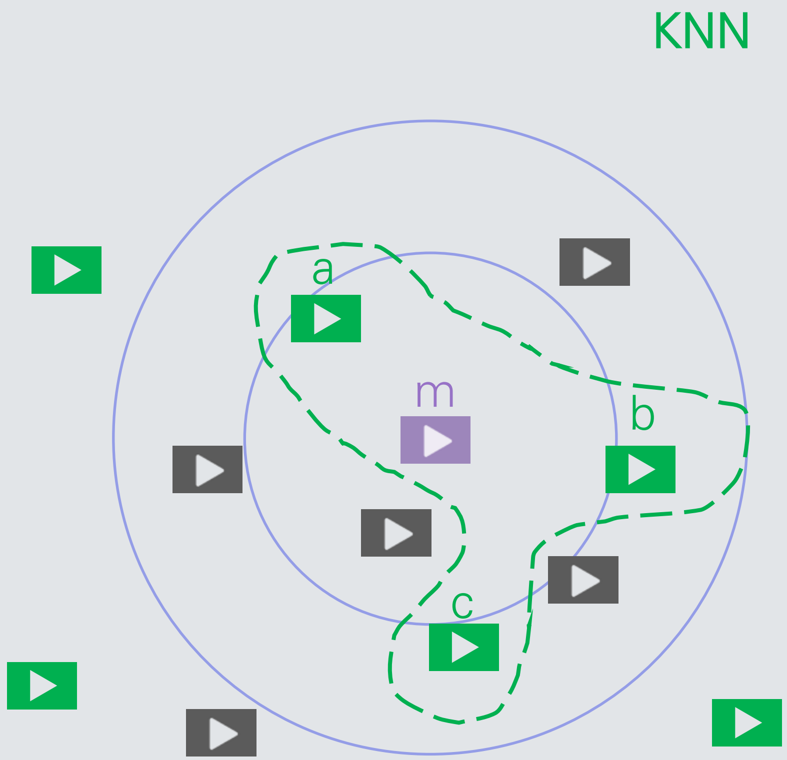
KNN





Algorithm Choice – KNN

- Recommend movies that are similar to the ones you already like.



$$\hat{r}_m = S_{am} r_a + S_{bm} r_b + S_{cm} r_c$$

Algorithm Choice – SVD++

Rating Matrix

	M1	M2	M3	M4	M5
U1	1	1	1	/	/
U2	3	/	2	/	/
U3	4	4	4	/	5
U4	5	/	5	/	0
U5	/	2	/	4	4
U6	/	/	/	5	1
U7	0	1	/	2	2

=

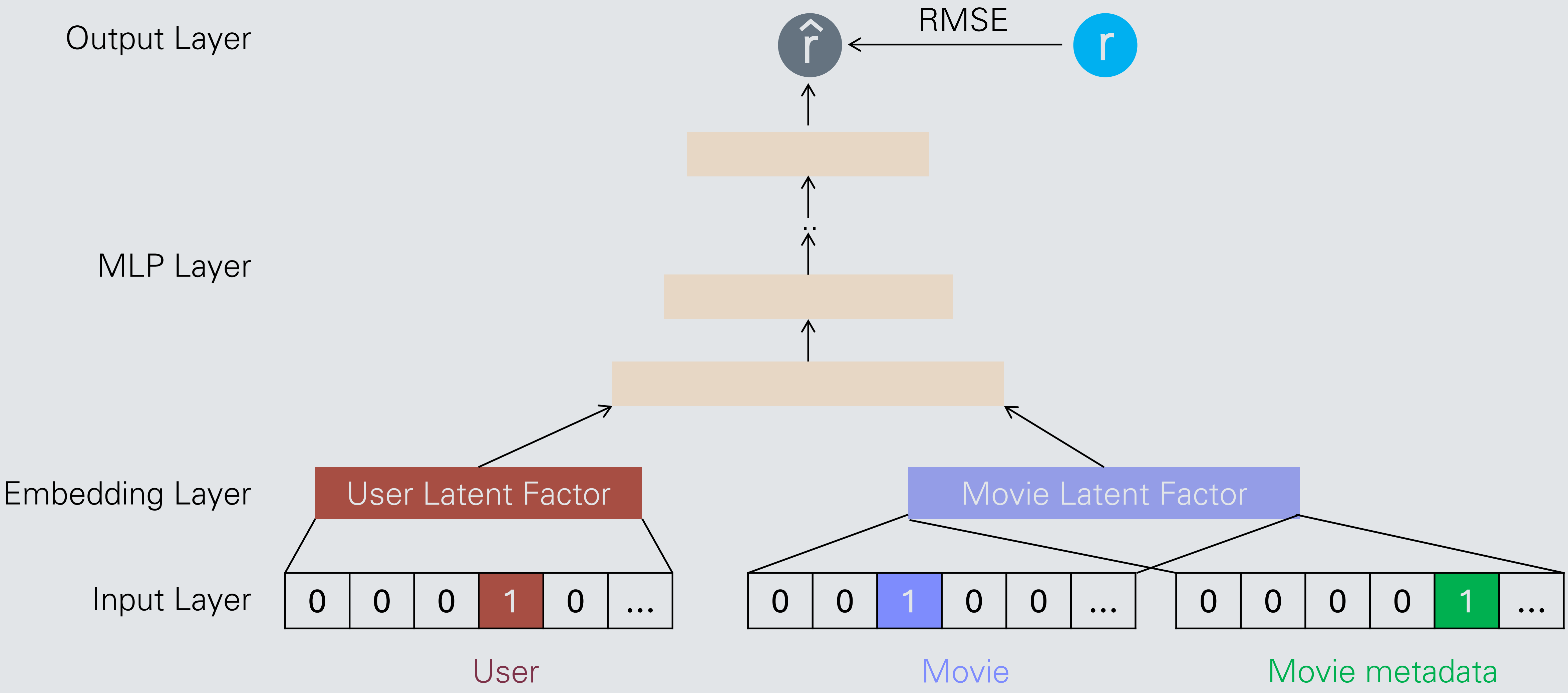
Latent Factor Space (k dim)

	Sci-fi	Romantic		
U1	0.13	0.02	..	-0.01
U2	0.41	0.07	..	-0.03
U3	-0.55	0.09	..	-0.04
U4	0.68	0.11	..	0.05
U5	0.15	-0.59	..	0.65
U6	0.07	0.73	..	-0.67
U7	0.07	-0.29	..	0.32

x

	M1	M2	M3	M4	M5
Sci-fi	0.56	0.59	0.56	0.09	0.09
Romantic	0.12	-0.02	0.12	0.64	-0.69
	..	..	..	..	..
	0.40	-0.80	0.40	0.09	0.09

Algorithm Choice – MLP



The Dataset

- **Dataset:** Sourced from Kaggle, containing metadata and user ratings for 45,000 movies released on or before July 2017.
- **Details Included:** Cast, crew, plot keywords, budget, revenue, release dates, languages, production companies, and countries.
- **Reason for Selection:** Offers comprehensive metadata compared to typical movie datasets.
- **This dataset includes three main files:**

✓ movies_metadata.csv – Details for 45,000 movies (title, genre, budget, revenue, etc.).

✓ credits.csv – Cast and crew information.

✓ ratings.csv – 26 million ratings from 270,000 users.

Preprocessing

Movies Metadata:

	adult	belongs_to_collection	budget	genres	homepage
0	False	{'id': 10194, 'name': 'Toy Story Collection', ...	30000000	[{'id': 16, 'name': 'Animation'}, {'id': 35, '...	http://toystory.disney.com/toy-story
1	False	NaN	65000000	[{'id': 12, 'name': 'Adventure'}, {'id': 14, '...	NaN
2	False	{'id': 119050, 'name': 'Grumpy Old Men Collect...	0	[{'id': 10749, 'name': 'Romance'}, {'id': 35, ...	NaN
3	False	NaN	16000000	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...	NaN
4	False	{'id': 96871, 'name': 'Father of the Bride Col...	0	[{'id': 35, 'name': 'Comedy'}]	NaN

	id	imdb_id	original_language	original_title	overview	popularity	poster_path
0	862	tt0114709	en	Toy Story	Led by Woody, Andy's toys live happily in his ...	21.946943	/rhIRbceoE9lR4veEXuwCC2wARtG.jpg
1	8844	tt0113497	en	Jumanji	When siblings Judy and Peter discover an encha...	17.015539	/vzmL6fP7aPKNKPRTFnZmiUfciyV.jpg
2	15602	tt0113228	en	Grumpier Old Men	A family wedding reignites the ancient feud be...	11.7129	/6ksm1sjKMFLbO7UY2i6G1ju9SML.jpg
3	31357	tt0114885	en	Waiting to Exhale	Cheated on, mistreated and stepped on, the wom...	3.859495	/16XOMpEaLWkrcPqSQqhTmeJuqQl.jpg
4	11862	tt0113041	en	Father of the Bride Part II	Just when George Banks has recovered from his ...	8.387519	/e64sOI48hQXyru7naBFyssKFxVd.jpg

	production_companies	production_countries	release_date	revenue	runtime	spoken_languages
0	[{'name': 'Pixar Animation Studios', 'id': 3}]	[{'iso_3166_1': 'US', 'name': 'United States o...	1995-10-30	373554033.0	81.0	[{'iso_639_1': 'en', 'name': 'English'}]
1	[{'name': 'TriStar Pictures', 'id': 559}, {'na...	[{'iso_3166_1': 'US', 'name': 'United States o...	1995-12-15	262797249.0	104.0	[{'iso_639_1': 'en', 'name': 'English'}, {'iso...
2	[{'name': 'Warner Bros.', 'id': 6194}, {'name'...	[{'iso_3166_1': 'US', 'name': 'United States o...	1995-12-22	0.0	101.0	[{'iso_639_1': 'en', 'name': 'English'}]
3	[{'name': 'Twentieth Century Fox Film Corporat...	[{'iso_3166_1': 'US', 'name': 'United States o...	1995-12-22	81452156.0	127.0	[{'iso_639_1': 'en', 'name': 'English'}]
4	[{'name': 'Sandollar Productions', 'id': 5842}...	[{'iso_3166_1': 'US', 'name': 'United States o...	1995-02-10	76578911.0	106.0	[{'iso_639_1': 'en', 'name': 'English'}]

	status	tagline	title	video	vote_average	vote_count
0	Released	NaN	Toy Story	False	7.7	5415.0
1	Released	Roll the dice and unleash the excitement!	Jumanji	False	6.9	2413.0
2	Released	Still Yelling. Still Fighting. Still Ready for...	Grumpier Old Men	False	6.5	92.0
3	Released	Friends are the people who let you be yourself...	Waiting to Exhale	False	6.1	34.0
4	Released	Just When His World Is Back To Normal... He's ...	Father of the Bride Part II	False	5.7	173.0

Credits:

	cast	crew	id
0	[{'cast_id': 14, 'character': 'Woody (voice)',...	[{'credit_id': '52fe4284c3a36847f8024f49', 'de...	862
1	[{'cast_id': 1, 'character': 'Alan Parrish', '...	[{'credit_id': '52fe44bfc3a36847f80a7cd1', 'de...	8844
2	[{'cast_id': 2, 'character': 'Max Goldman', 'c...	[{'credit_id': '52fe466a9251416c75077a89', 'de...	15602
3	[{'cast_id': 1, 'character': "Savannah 'Vannah...	[{'credit_id': '52fe44779251416c91011acb', 'de...	31357
4	[{'cast_id': 1, 'character': 'George Banks', '...	[{'credit_id': '52fe44959251416c75039ed7', 'de...	11862

Ratings:

	userId	movieId	rating	timestamp
0	1	110	1.0	1425941529
1	1	147	4.5	1425942435
2	1	858	5.0	1425941523
3	1	1221	5.0	1425941546
4	1	1246	5.0	1425941556

Preprocessing – Feature Selection

- **Movies metadata:**

✓ id: Primary key for table joins.

✓ imdbId: Used for retrieving missing data via IMDB API.

✓ genre, release_date, original_language: Key for filtering and capturing user preferences.

✓ title: Ensures meaningful recommendations.

- **Credits:**

✓ cast: Helps recommend movies with favorite actors.

✓ crew: Only the director is retained for relevance.

- **Ratings:**

✓ Since we focus on CF algorithms, ratings naturally become the main feature

Preprocessing – Handling Duplicates, Missing Values, Feature Cleaning

Duplicate Removal:

- Duplicate entries are identified and removed to ensure data integrity.

Handling Missing Values

- Missing values exist in both movies_metadata.csv and credits.csv, we retrieve missing values using imdbId as a key from the IMDB API.

- While a large portion of missing data is recovered, a few values are unavailable on IMDB. The missing data are minimal and, therefore, dropped without significant impact on the dataset.

Feature Cleaning

- Raw feature data is cleaned for improved usability. For example, we needed to convert a nested list of dictionaries into a list of genre names:

- "[{'id': 16, 'name': 'Animation'}, {'id': 35, 'name': 'Comedy'}]" → ['Animation', 'Comedy'].

Preprocessing – Movies Metadata

1. ‘id’, ‘imdbId’, ‘title’, ‘original_language’:

- Convert to string format to ensure consistency.

2. ‘genre’:

- Convert nested lists of dictionaries into a list of genre names.

- Example: "[{'id': 16, 'name': 'Animation'}, {'id': 35, 'name': 'Comedy'}]" → ['Animation', 'Comedy’]

3. ‘release_date’:

- Extract only the year from the full date format.

- Example: "1994-06-15" → "1994"

4. ‘cast’:

- Convert nested lists of dictionaries into a list of actor names (three actors).

- Example: "[{'cast_id': 14, 'name': 'Tom Hanks'}, {'cast_id': 2, 'name': 'Tim Allen'}]" → ['Tom Hanks', 'Tim Allen’]

5. ‘crew’:

- Extract only the director’s name from the list of crew members.

- Example: "[{'job': 'Director', 'name': 'Joe Johnston'}, {'job': 'Producer', 'name': 'Jane Doe'}]" → ['Joe Johnston']

```python
import ast
def extract_genres(genres):
    try:
        genres_list = ast.literal_eval(genres)
        return [genre['name'] for genre in genres_list]
    except (ValueError, TypeError):
        return []
```

```python
def get_first_three_actors(cast):
    try:
        cast_list = ast.literal_eval(cast)
        return [actor['name'] for actor in cast_list[:3]]
    except (ValueError, TypeError):
        return []
```

Preprocessing – ratings

Downsizing

- Drop movies that are not in the metadata.
- Drop users that has less than 20 ratings.

	userId	movieId	rating	timestamp
0	1	110	1.0	1425941529
1	1	147	4.5	1425942435
2	1	858	5.0	1425941523
3	1	1221	5.0	1425941546
4	1	1246	5.0	1425941556
..	..	..	..	..
..	..	..	..	..

`26024289(100%) rows`

	userId	movieId	rating	timestamp
0	4	223	4.0	1042668576
1	4	415	4.0	1042667925
2	4	648	4.0	1042674800
..	..	..	..	..
..	..	..	..	..

`10212749(~40%) rows`

Preprocessing – ratings

Leave-Last-Out Splitting

- **Mimics Real-World Prediction** – Trains on past interactions, tests on the most recent ones.
- **Prevents Data Leakage** – Ensures the model doesn't "see" future interactions during training.
- **Standard Benchmarking** – Common in research, enabling direct performance comparison.

Training	the first N-2 items	Train size	9972455
Validation	the (N-1)-th item	Validation size	120147
Testing	the N-th item	Test size	120147

Preprocessing – ratings

Training Matrix

movieId \ userId	4	7	8	9	11	12	15	16	20	22	...	270879	270881	270883	270885	270887	270891	270892	270893	270894	270896
2	/	/	/	/	/	/	/	/	/	/	...	3.5	/	/	/	5.0	/	/	/	/	/
3	/	/	/	/	/	/	/	/	/	/	...	/	/	/	/	4.0	3.0	/	/	/	/
5	/	/	/	/	/	/	/	/	/	/	...	/	/	/	/	/	/	/	/	/	/
6	/	/	/	/	/	/	4.0	/	/	/	...	/	/	/	/	5.0	/	/	/	/	/
11	/	/	/	/	/	/	/	/	/	/	...	/	/	/	/	4.0	4.0	/	/	/	3.5
..	..	..	..	..	..	..	..	..	..	..	...	..	..	..	..	..	..	..	..	..	..

```
Shape: (7486 movies, 120147 users)
```

```
Sparsity: 98.89%
```

Next Stages

EDA

- Correlations between movie metadata vs user preference – Justify metadata features choice.
- Rating distribution per item – Identify highly-rated vs. poorly-rated items. b_i
- Rating distribution per user – Detect users who rate too generously or harshly. b_u
- Time-based trends – Check if ratings change over time (e.g., new releases get higher ratings).
- Cluster similar items– See if similarity measurement makes sense.
- …

Feature Engineering

- Similarity matrix of movies → KNN
- Global bias (μ)
- Item biases (b_i) → SVD++
- User biases (b_u)
- One-hot encoding of genre/language → MLP
- TDIDF of actors/director
- …