STUDY OF RECOMMENDER SYSTEMS

MANJIT ULLAL

OMKAR WAGHMARE

ARSH IRFAN MODAK

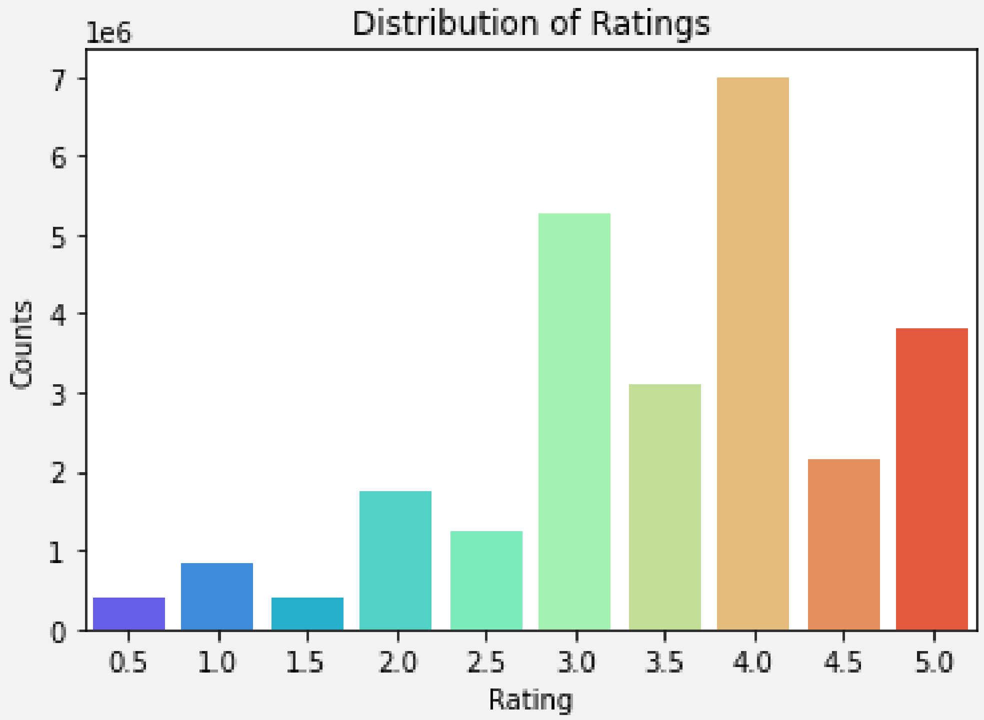
INTRODUCTION

- Recommender systems are widely employed in the industry and their major aim is to help users discover new and relevant items such as movies to watch, text to read or products to buy, find compelling content, so as to create a delightful user experience.

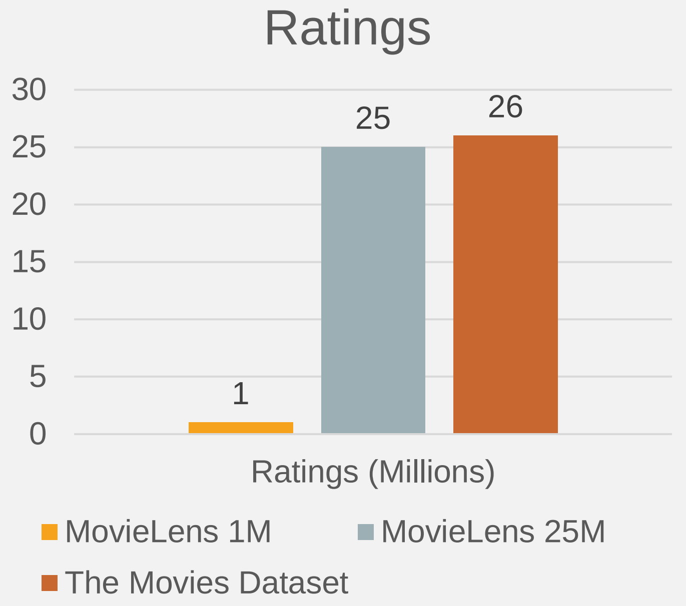
- In this project we implement various Recommender Systems using Similarity Measures, Scikit-Surprise and Neural Networks and try to improve them.

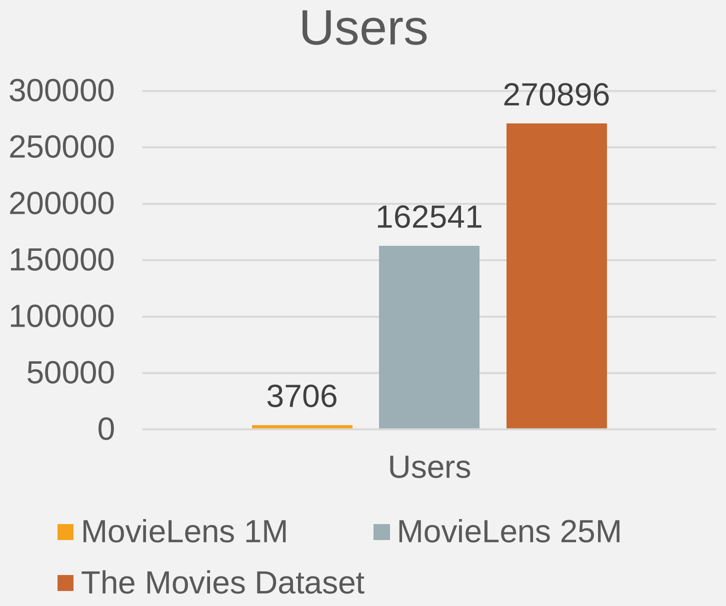
THE DATASETS

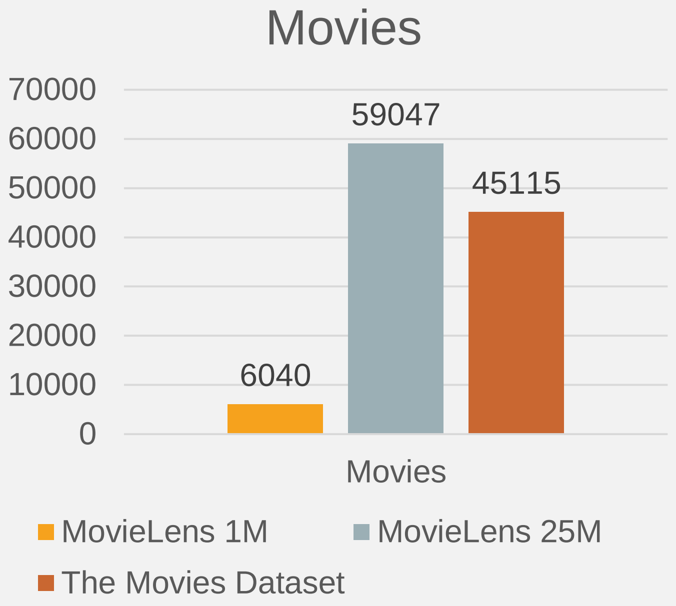
- Our primary dataset is “The Movies Dataset” from Kaggle, which consists of 26 Million ratings from over 270,000 users for over 45,000 movies.
- The ratings are in the range 0.5 to 5
- The dataset also contains a metadata file which has information such as cast, crew, genre, overview, links to posters, etc.



COMPARISON WITH THE MOVIE LENS DATASET







DATA PRE-PROCESSING

- To keep our data consistent with the Movie-Lens datasets we made sure all users rated at least 20 movies.
- The metadata file was really messy and needed a lot of cleaning to be used for our Context Based Recommender System.
- Columns such as Genres and Overview used to extract useful data from the movies metadata table
- Information such as Actors and Characters were extracted from the Cast Column from the Credits table.
- We also created various samples of our data to tackle a few problems! (more on that later)

BEFORE

'[{'cast_id': 14, 'character': 'Woody (voice)', 'credit_id': '52fe4284c3a36847f8024f95', 'gender': 2, 'id': 31, 'name': 'Tom Hanks', 'order': 0, 'profile_path': '/pQ
Foyx7rp09CJTAb932F2g8Nlho.jpg'}, {'cast_id': 15, 'character': 'Buzz Lightyear (voice)', 'credit_id': '52fe4284c3a36847f8024f99', 'gender': 2, 'id': 12898, 'name': 'T
im Allen', 'order': 1, 'profile_path': '/uX2xVf6pMmPepxnvFWyBtjexzgY.jpg'}, {'cast_id': 16, 'character': 'Mr. Potato Head (voice)', 'credit_id': '52fe4284c3a36847f80
24f9d', 'gender': 2, 'id': 7167, 'name': 'Don Rickles', 'order': 2, 'profile_path': '/h5BcaDMPRVLHLDzbQavec4xfSdt.jpg'}, {'cast_id': 17, 'character': 'Slinky Dog (vo
ice)', 'credit_id': '52fe4284c3a36847f8024fa1', 'gender': 2, 'id': 12899, 'name': 'Jim Varney', 'order': 3, 'profile_path': '/eIo2jVVXYgjDtaHoF19Ll9vtW7h.jpg'}, {'ca
st_id': 18, 'character': 'Rex (voice)', 'credit_id': '52fe4284c3a36847f8024fa5', 'gender': 2, 'id': 12900, 'name': 'Wallace Shawn', 'order': 4, 'profile_path': '/oGE
6JgPP2xH4tN...'

	id	title	overview	genres
0	862	Toy Story	Led by Woody, Andy's toys live happily in his ...	[{'id': 16, 'name': 'Animation'}, {'id': 35, '...
1	8844	Jumanji	When siblings Judy and Peter discover an encha...	[{'id': 12, 'name': 'Adventure'}, {'id': 14, '...
2	15602	Grumpier Old Men	A family wedding reignites the ancient feud be...	[{'id': 10749, 'name': 'Romance'}, {'id': 35, ...
3	31357	Waiting to Exhale	Cheated on, mistreated and stepped on, the wom...	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...
4	11862	Father of the Bride Part II	Just when George Banks has recovered from his ...	[{'id': 35, 'name': 'Comedy'}]

AFTER

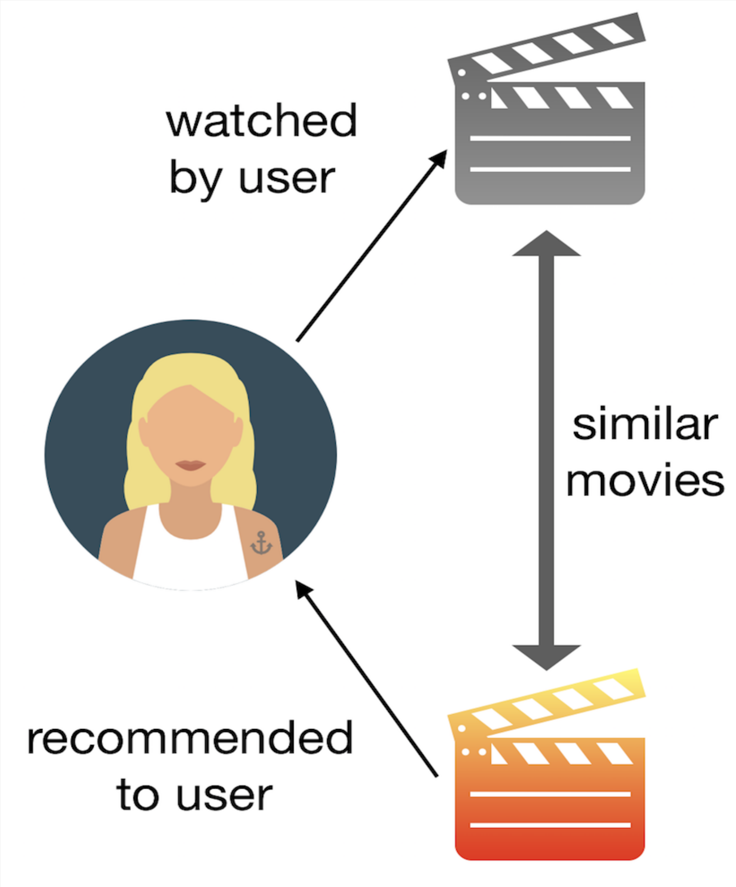
	id	original_title	overview	list_of_actors	list_of_characters	list_of_genres	metadata	overview_genre	overview_actors
0	862	Toy Story	Led by Woody, Andy's toys live happily in his ...	Tom Hanks, Tim Allen, Don Rickles, Jim Varney,...	Woody (voice), Buzz Lightyear (voice), Mr. Pot...	Animation, Comedy, Family	Toy Story, Led by Woody, Andy's toys live happ...	Led by Woody, Andy's toys live happily in his ...	Led by Woody, Andy's toys live happily in his ...
1	8844	Jumanji	When siblings Judy and Peter discover an encha...	Robin Williams, Jonathan Hyde, Kirsten Dunst, ...	Alan Parrish, Samuel Alan Parrish / Van Pelt, ...	Adventure, Fantasy, Family	Jumanji, When siblings Judy and Peter discover...	When siblings Judy and Peter discover an encha...	When siblings Judy and Peter discover an encha...
2	15602	Grumpier Old Men	A family wedding reignites the ancient feud be...	Walter Matthau, Jack Lemmon, Ann-Margret, Soph...	Max Goldman, John Gustafson, Ariel Gustafson, ...	Romance, Comedy	Grumpier Old Men, A family wedding reignites t...	A family wedding reignites the ancient feud be...	A family wedding reignites the ancient feud be...
3	31357	Waiting to Exhale	Cheated on, mistreated and stepped on, the wom...	Whitney Houston, Angela Bassett, Loretta Devin...	Savannah 'Vannah' Jackson, Bernadine 'Bernie' ...	Comedy, Drama, Romance	Waiting to Exhale, Cheated on, mistreated and ...	Cheated on, mistreated and stepped on, the wom...	Cheated on, mistreated and stepped on, the wom...
4	11862	Father of the Bride Part II	Just when George Banks has recovered from his ...	Steve Martin, Diane Keaton, Martin Short, Kimb...	George Banks, Nina Banks, Franck Eggelhoffer, ...	Comedy	Father of the Bride Part II, Just when George ...	Just when George Banks has recovered from his ...	Just when George Banks has recovered from his ...

TYPES OF RECOMMENDER SYSTEMS

- Content – Based Recommender System
- Collaborative Filtering
- Neural Collaborative Filtering
- Variational Autoencoders
- Recommendation using Clustering (Spark)

CONTENT-BASED RECOMMENDATION

- Content-based filtering uses item features to recommend other items similar to what the user likes, based on their previous actions or explicit feedback.
- A content based recommender works with data that the user provides, either explicitly or implicitly.
- Content-based recommenders suggest similar items based on a particular item. This system uses item metadata, such as genre, director, description, actors, etc. for movies, to make these recommendations



CONTENT-BASED RECOMMENDATION

- To implement this, the first thing we did was to clean the data.
- After we got the data we need we created a TF-IDF Vectorizer and applied it on the data.
- Next, we used Scikit-Learn's linear kernel and cosine similarity to create a similarity matrix.
- This matrix was then used to recommend movies based on the metadata we used.

THE RECOMMENDATIONS (FOR TOY STORY)

OVERVIEW

	original_title
15378	Toy Story 3
3002	Toy Story 2
10317	The 40 Year Old Virgin
24569	Small Fry
23888	Andy Hardy's Blonde Trouble
29265	Hot Splash
43496	Andy Kaufman Plays Carnegie Hall
38543	Superstar: The Life and Times of Andy Warhol
42791	Andy Peters: Exclamation Mark Question Point

ON ACTORS

	original_title
3002	Toy Story 2
15378	Toy Story 3
25847	Toy Story That Time Forgot
21970	Toy Story of Terror!
14686	Ernest Goes to School
14750	Dr. Otto and the Riddle of the Gloom Beam
24569	Small Fry
24567	Hawaiian Vacation
25845	Partysaurus Rex

THE RECOMMENDATIONS (FOR STAR WARS)

ON OVERVIEW

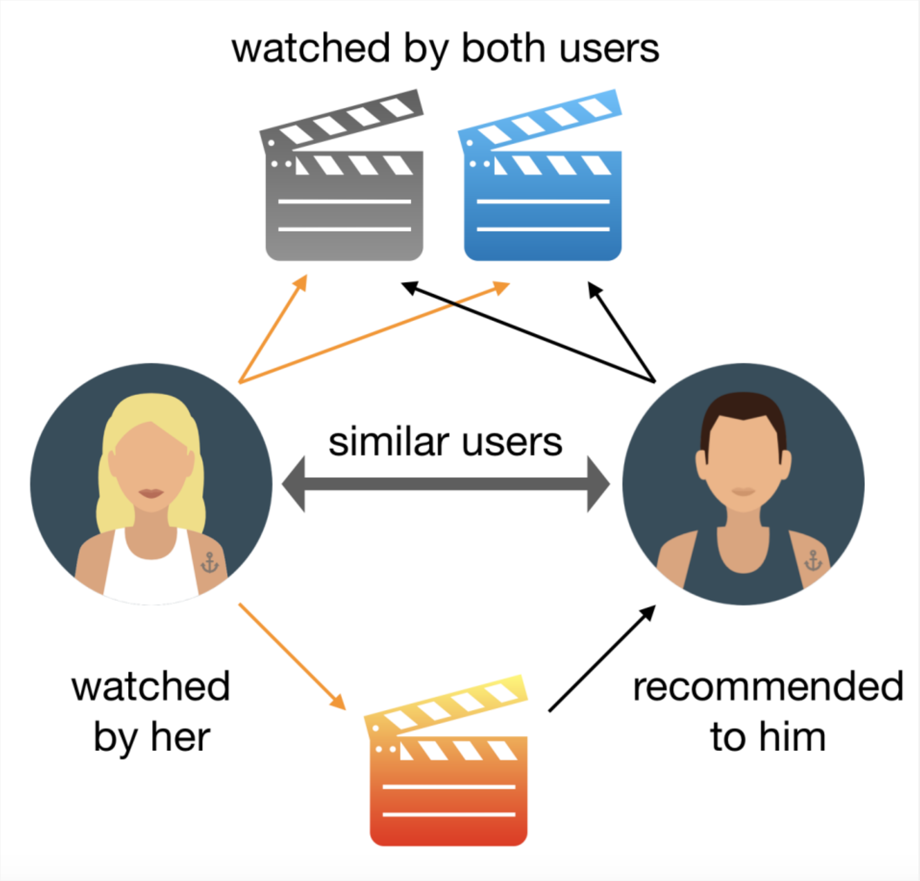
	original_title
256	Star Wars
1157	The Empire Strikes Back
30498	The Star Wars Holiday Special
26616	Star Wars: The Force Awakens
1170	Return of the Jedi
34220	Maciste alla corte del Gran Khan
1270	Mad Dog Time
5195	The Triumph of Love
37901	Dao bing fu
25151	1½ Ritter - Auf der Suche nach der hinreißende...

ON ACTORS

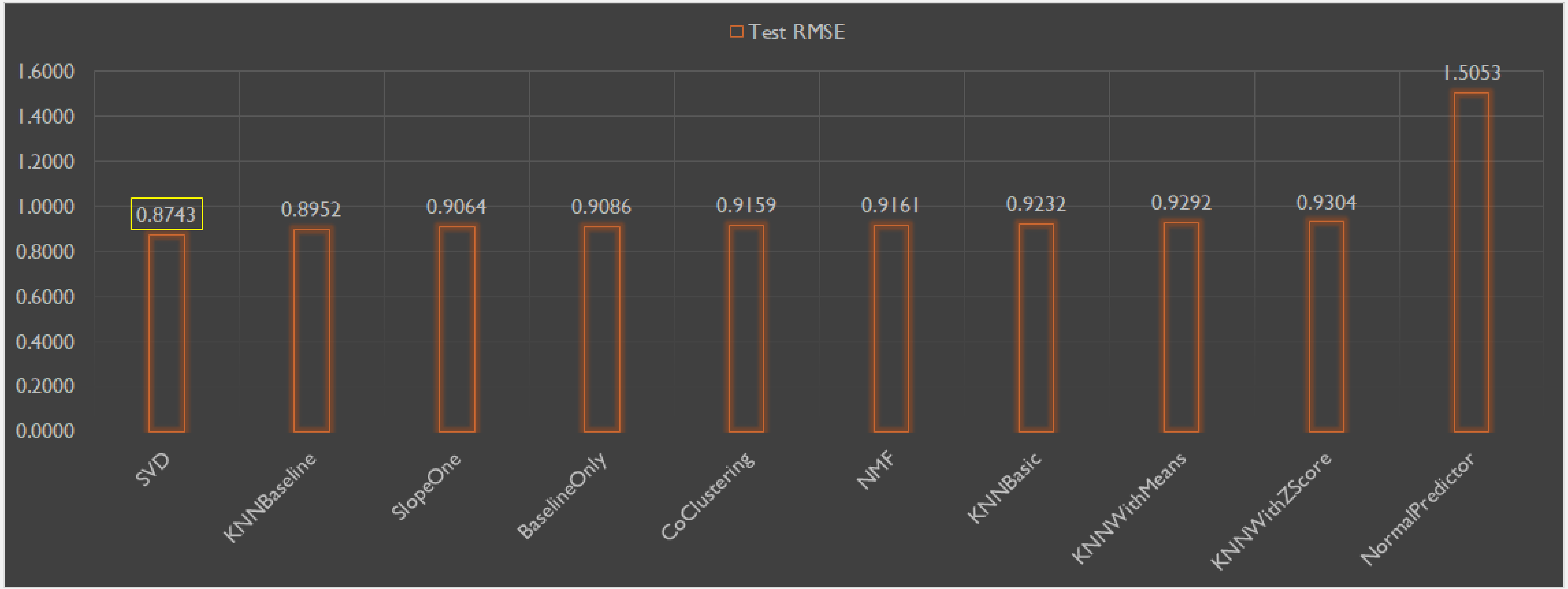
	original_title
256	Star Wars
1157	The Empire Strikes Back
39126	Elstree 1976
15483	Empire of Dreams: The Story of the Star Wars T...
1170	Return of the Jedi
30498	The Star Wars Holiday Special
733	Dr. Strangelove or: How I Learned to Stop Worr...
1159	Raiders of the Lost Ark
3331	Funny Bones
927	Around the World in Eighty Days

COLLABORATIVE FILTERING USING SCIKIT-SURPRISE

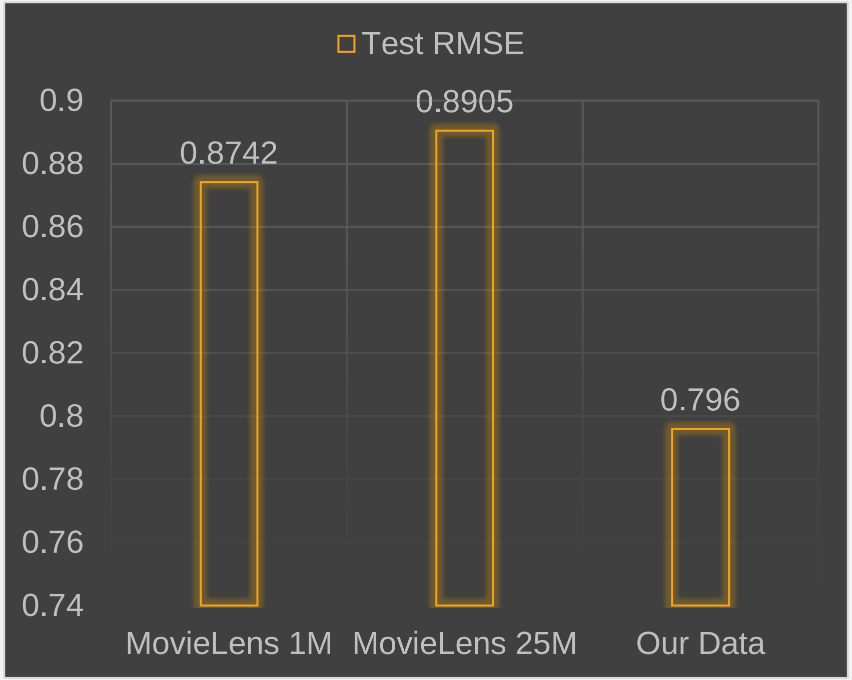
- Scikit-Surprise is a recommendation system library python which lets us implement the following models for Collaborative Filtering:
- Singular Value Decomposition (SVD)
- Non-Negative Matrix Factorization (NMF)
- NormalPredictor
- BaselineOnly
- KNN Based Models (Baseline, Basic, With Means, With Z-score
- SlopeOne
- Co-clustering

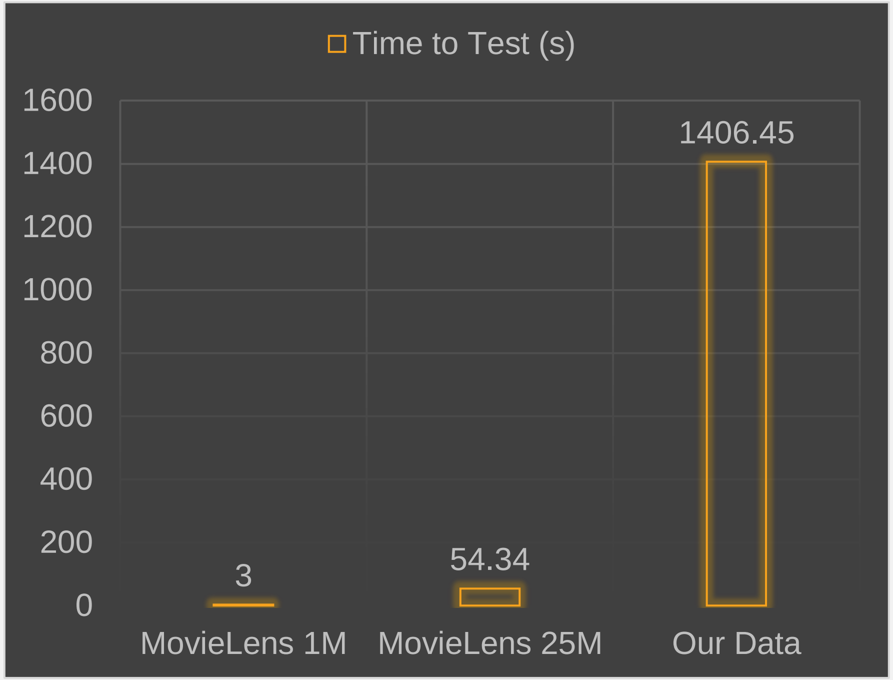


MOVIE-LENS 1 MILLION



COMPARISON OF SVD ON DIFFERENT DATA





PREDICTING USER RATINGS USING SVD ON MOVIE-LENS 25 MILLION

USER ID: 1

	Movies	Ratings
45626	Planet Earth II (2016)	4.584791
24081	John Mulaney: New In Town (2012)	4.572865
33706	Little Dorrit (2008)	4.508264
7865	Before Sunset (2004)	4.482566
21092	Grand Budapest Hotel, The (2014)	4.471425
47391	I Am So Proud of You (2008)	4.461771
661	Some Folks Call It a Sling Blade (1993)	4.460846
1195	Harold and Maude (1971)	4.459788
45478	Band of Brothers (2001)	4.457491
5636	Professional, The (Le professionnel) (1981)	4.432907

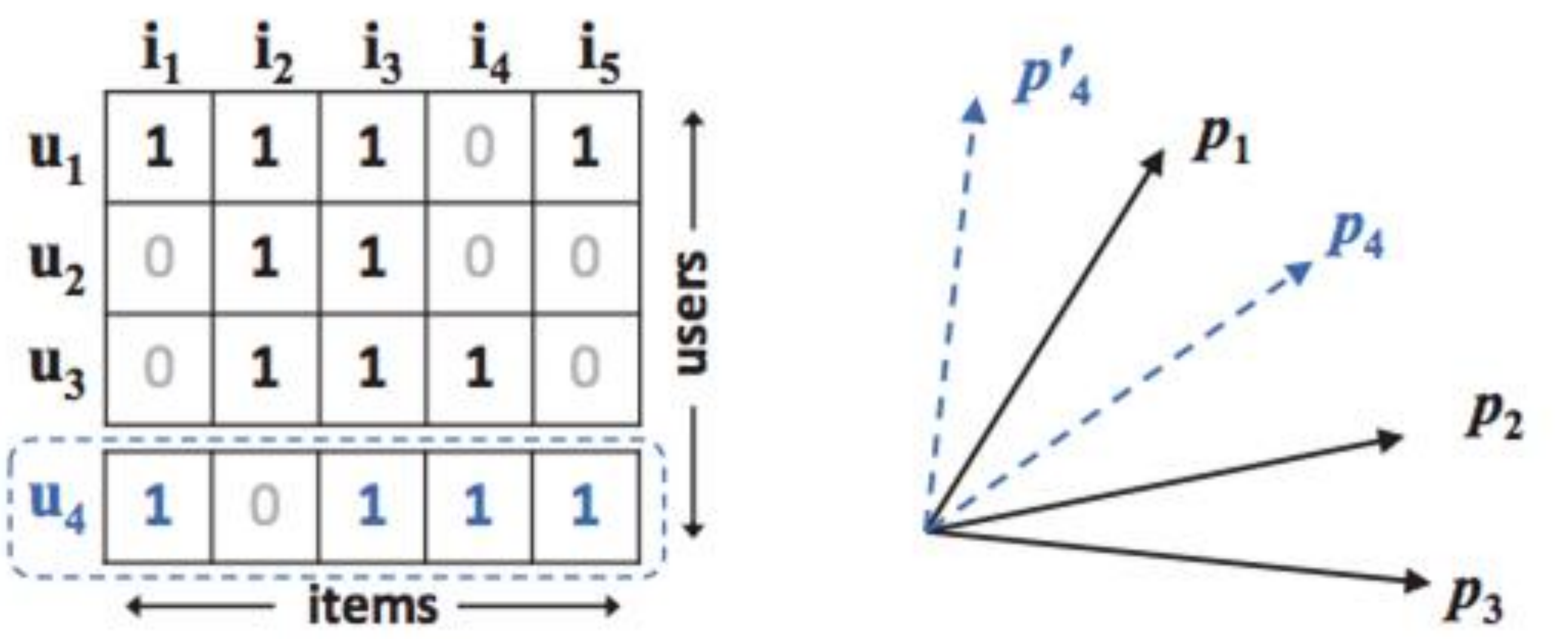
USER ID: 777

	Movies	Ratings
49470	Blue Planet II (2017)	5.0
11041	Fear City: A Family-Style Comedy (La cité de l...	5.0
23652	Bill Burr: I'm Sorry You Feel That Way (2014)	5.0
35465	Winter on Fire: Ukraine's Fight for Freedom (2...	5.0
4163	Rififi (Du rififi chez les hommes) (1955)	5.0
35201	The Adventures of Sherlock Holmes and Dr. Wats...	5.0
35200	The Adventures of Sherlock Holmes and Doctor W...	5.0
35182	Seventeen Moments in Spring (1973)	5.0
35163	The Adventures of Sherlock Holmes and Doctor W...	5.0
10887	Army of Shadows (L'armée des ombres) (1969)	5.0

USER ID: 8634

	Movies	Ratings
2223	Life Is Beautiful (La Vita è bella) (1997)	5.000000
312	Shawshank Redemption, The (1994)	5.000000
17648	Intouchables (2011)	4.947167
48341	Black Mirror	4.904947
7830	Notebook, The (2004)	4.901970
349	Forrest Gump (1994)	4.872719
1632	Good Will Hunting (1997)	4.867236
14225	3 Idiots (2009)	4.864804
9339	Chorus, The (Choristes, Les) (2004)	4.861174
21268	Cranford (2007)	4.830718

LIMITATIONS OF MATRIX FACTORIZATION



(a) user–item matrix (b) user latent space

Considering u1,u2,u3: $S_{23} > S_{12} > S_{13}$

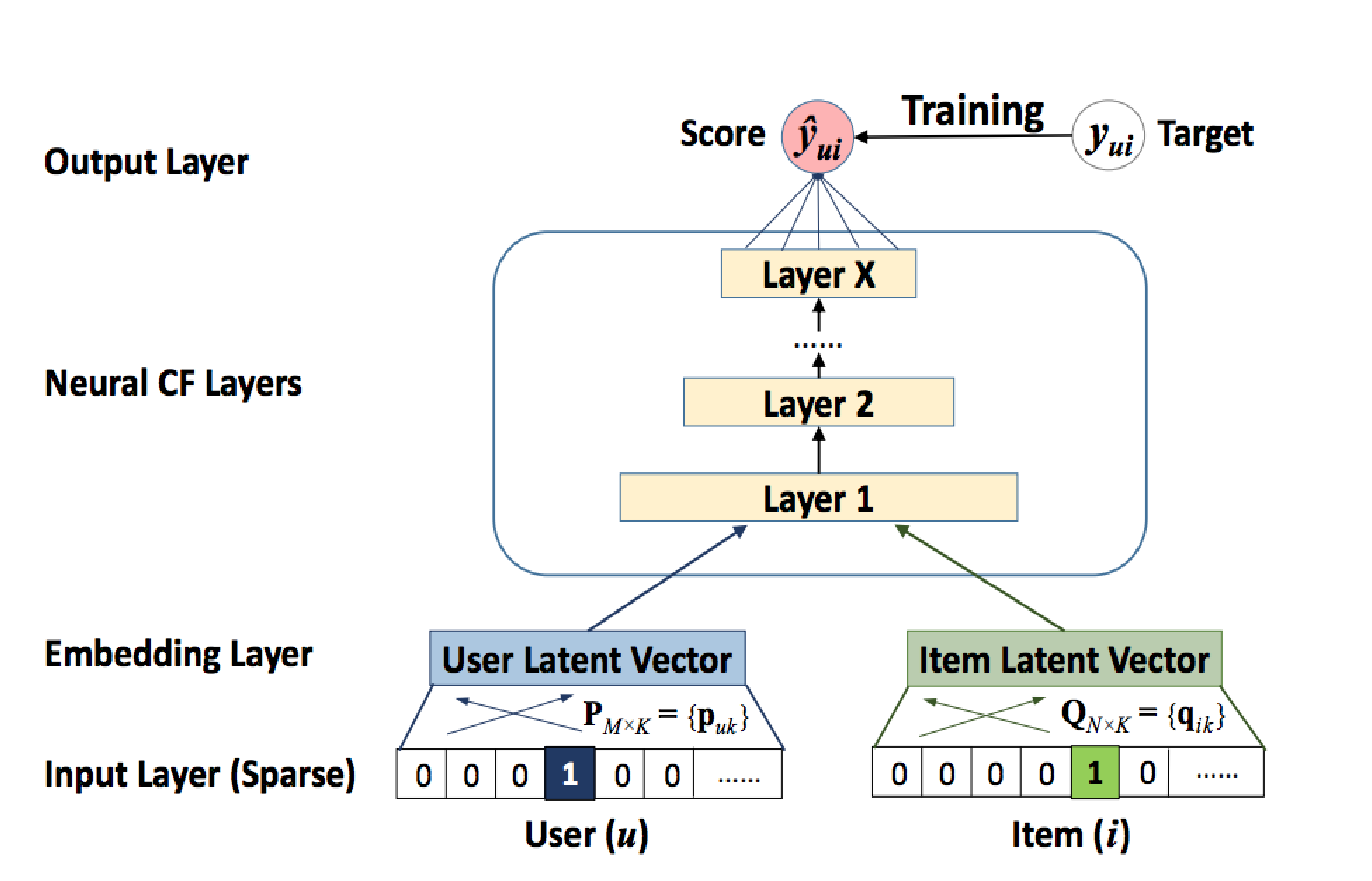
Considering u4 : $S_{41} > S_{43} > S_{42}$

MF model places p4 closest to p1

Incur a large ranking loss

slide taken from Kung-hsaing Huang NCF

NCF ARCHITECTURE



Input Layer: binarise a sparse vector for a user and item identification where:

Item (i): 1 means the user u has interacted with Item(i)
User (u): To identify the user

Embedding layer: is a fully connected layer that projects the sparse representation to a dense vector. The obtained user/item embeddings are the latent user/item vectors.

Neural CF layers: use Multi-layered neural architecture to map the latent vectors to prediction scores.

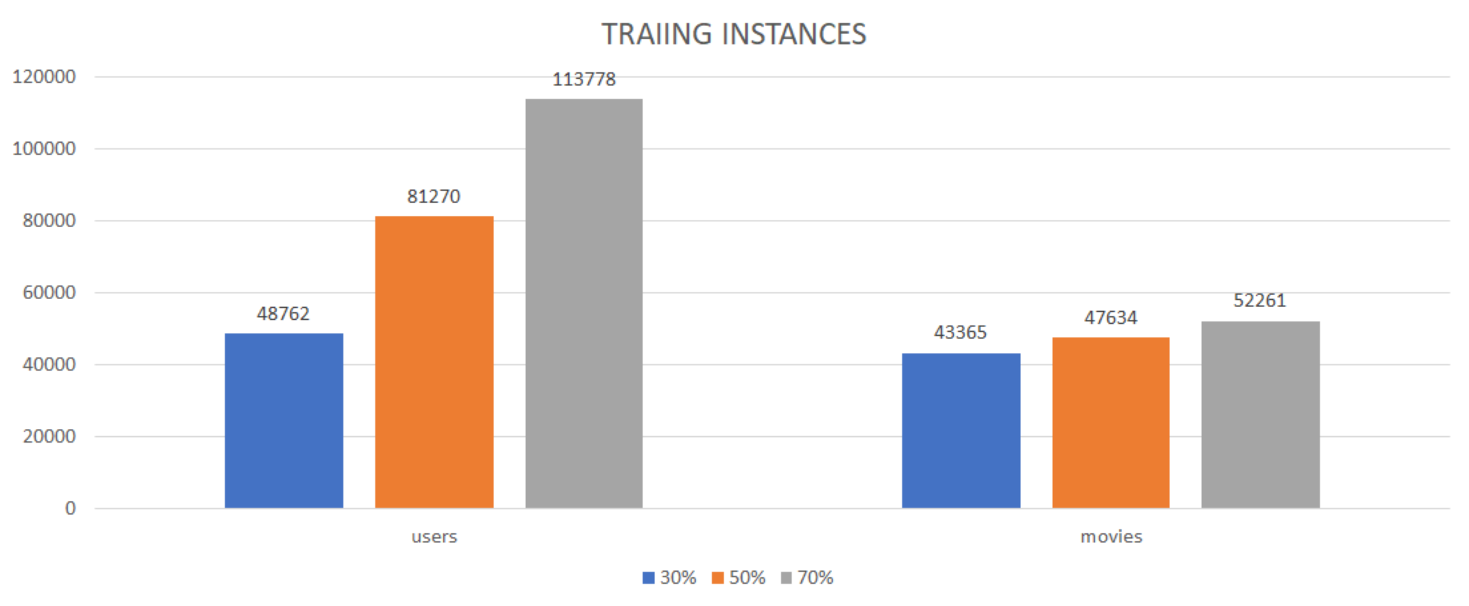
Output layer: returns the predicted score. In our case we are using a sigmoid function.

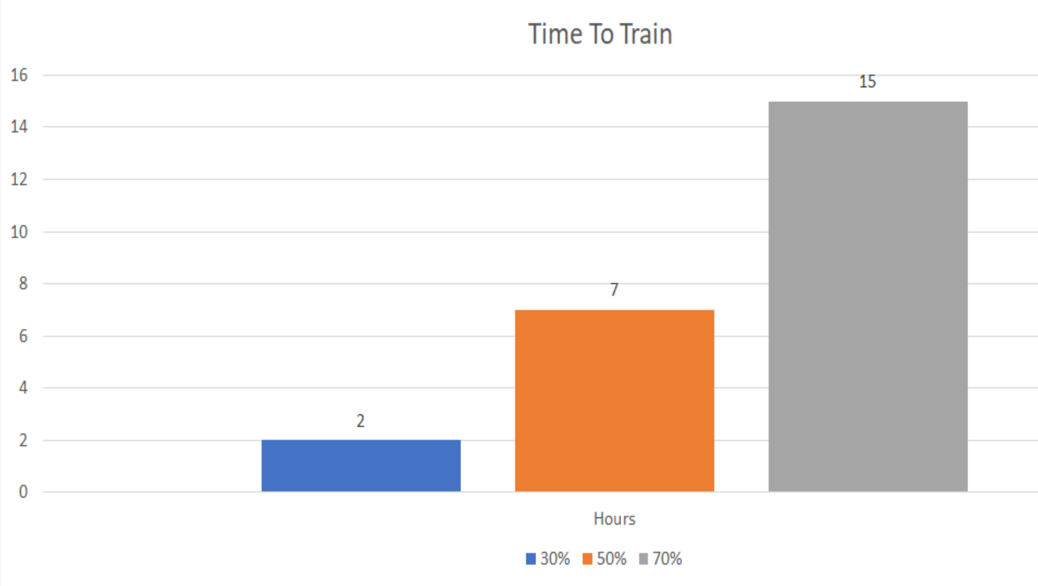
slide taken from Kung-hsaing Huang NCF

METRIC: HIT RATIO

- For each user, randomly select 99 items that the user has not interacted with
- Combine these 99 items with the test item (the actual item that the user interacted with). We now have 100 items.
- Run the model on these 100 items, and rank them according to their predicted probabilities
- Select the top 10 items from the list of 100 items. If the test item is present within the top 10 items, then we say that this is a hit.
- Repeat the process for all users. The Hit Ratio is then the average hits.
- This evaluation protocol is known as Hit Ratio @ 10, and it is commonly used to evaluate recommender systems.

PERFORMANCE





SAMPLE OUTPUT

```
Showing recommendations for user: 181
====================================
Movies with high ratings from user
--------------------------------
Babe (1995) : Children|Drama
Forrest Gump (1994) : Comedy|Drama|Romance|War
Firm, The (1993) : Drama|Thriller
Jurassic Park (1993) : Action|Adventure|Sci-Fi|Thriller
Ghost (1990) : Comedy|Drama|Fantasy|Romance|Thriller
--------------------------------
Top 10 movie recommendations
--------------------------------
Star Wars: Episode IV - A New Hope (1977) : Action|Adventure|Sci-Fi
Shawshank Redemption, The (1994) : Crime|Drama
Godfather, The (1972) : Crime|Drama
American History X (1998) : Crime|Drama
Matrix, The (1999) : Action|Sci-Fi|Thriller
Fight Club (1999) : Action|Crime|Drama|Thriller
Amelie (Fabuleux destin d'Amélie Poulain, Le) (2001) : Comedy|Romance
Lord of the Rings: The Fellowship of the Ring, The (2001) : Adventure|Fantasy
Departed, The (2006) : Crime|Drama|Thriller
Dark Knight, The (2008) : Action|Crime|Drama|IMAX
```

COLLABORATIVE FILTERING USING VAE (VARIATIONAL AUTOENCODERS)

Learn non linear dependencies in the data by modeling user-item implicit feedback data using auto-encoders.

Create a latent representation of input space of user rating and infer missing rating by non-linear modelling of dependencies.

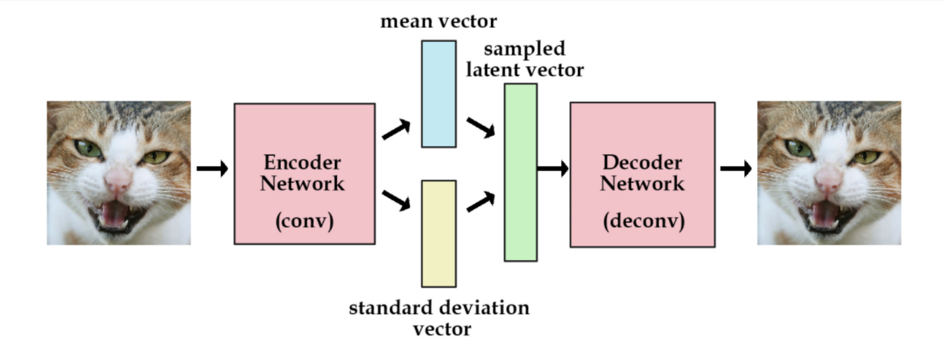


image taken from http://kvfrans.com/

RMSE on Test dataset

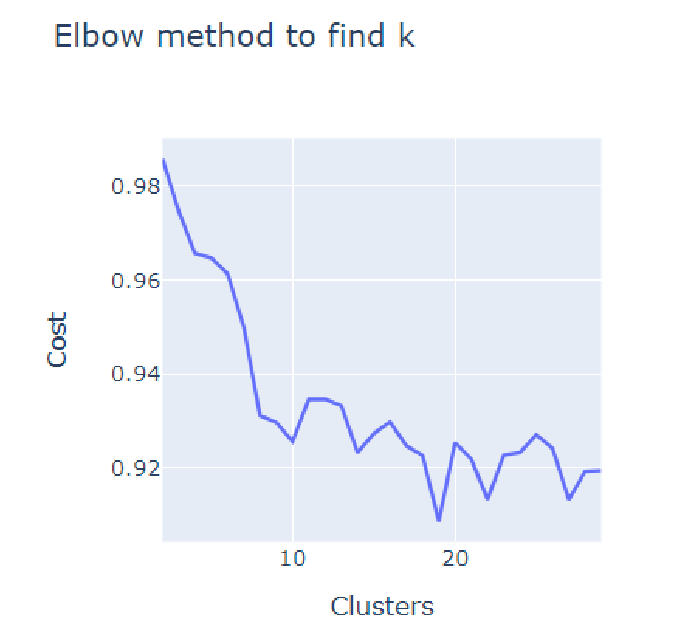
achieved best results compared to the benchmarks

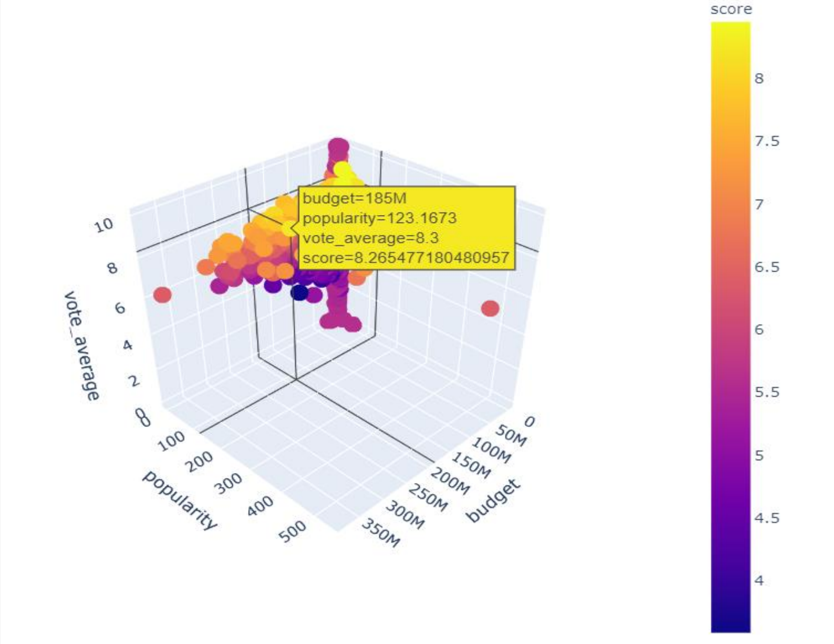
```
Test best model with test set!
Val: N@1 0.401, N@5 0.387, N@10 0.368, R@1 0.401, R@5 0.380, R@10 0.359: 100% 4/4 [00:00<00:00, 38.39it/s]
{'Recall@100': 0.573666023731232, 'NDCG@100': 0.4347800090909004, 'Recall@50': 0.45674996823072433, 'NDCG@50': 0.389883354306221,
```

RECOMMENDATION USING CLUSTERING

- Use Spark distributed framework's Kmeans clustering algorithm to find movies which are similar on content.
- Create Spark context and parallelize the Dataset, which is scalable and can process Big data.
- Use Spark ML lib Kmeans to find cluster of assignments to each movie.
- Find best 'k' cluster using elbow method.

K-MEANS USING APACHE Spark





	adult	budget	popularity	revenue	vote_average	vote_count	score
23742	False	3300000	64.3	13092000.0	8.3	4376.0	8.205405
586	False	19000000	4.30722	272742922.0	8.1	4549.0	8.015676
1632	False	10000000	15.0648	225933435.0	7.9	2880.0	7.779907
891	False	878000	13.9161	10462500.0	7.9	1462.0	7.674918
582	False	100000000	22.6617	520000000.0	7.7	4274.0	7.624880

REMAINING AND FUTURE WORK

- Hyperparameter Tuning of SVD on our Dataset (Extremely Time Consuming)
- Hyperparameter Tuning of KNN Based Algorithms on MovieLens 1M
- Calculating User-Item Rating Bias
- Experimenting with different configurations of Neural Networks for Neural Collaborative Filtering.
- Experimenting on a smaller subset of our data for Content-Based Recommendation on various metadata (separately and combined).
- Resolving cold start problem using deep learning.
- Creating new metrics for evaluating recommender systems.

THANK YOU


*Le Recommender System


NARCOS
DAREDEVIL
GOTHAM
ARROW
THE 100
BLUE MOUNTAIN STATE
BETTER CALL Saul
WENTWORTH
CLUB DE CUERVOS
archer
the inbetweeners
how to get away with Murder
MARCO POLO
That '70s Show
BLACKLIST
FORENSIC FILES COLLECTION
BLACK MIRROR
SUITS
BROOKLYN NINE-NINE
BLOODLINE
JANE THE VIRGIN
THE IT CROWD
Grace and Frankie
UNBREAKABLE KIMMY SCHMIDT