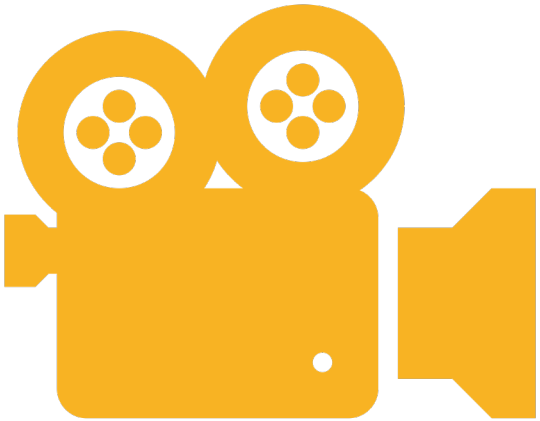
MOVIE RECOMMENDATION SYSTEM



PROJECT TEAM: DERRIE SUSAN VARGHESE
SAYAN BISWAS
SNEHA AGARWAL
VARUN JAGADEESH

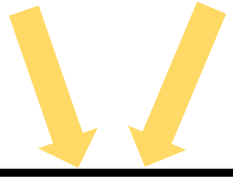
Project Objective

- Building recommendation system to identify most relevant movies for each user

Dataset Description

movieId	title	genres
2	Jumanji (1995)	Adventure|Children|Fantasy
3	Grumpier Old Men (1995)	Comedy|Romance
4	Waiting to Exhale (1995)	Comedy|Drama|Romance

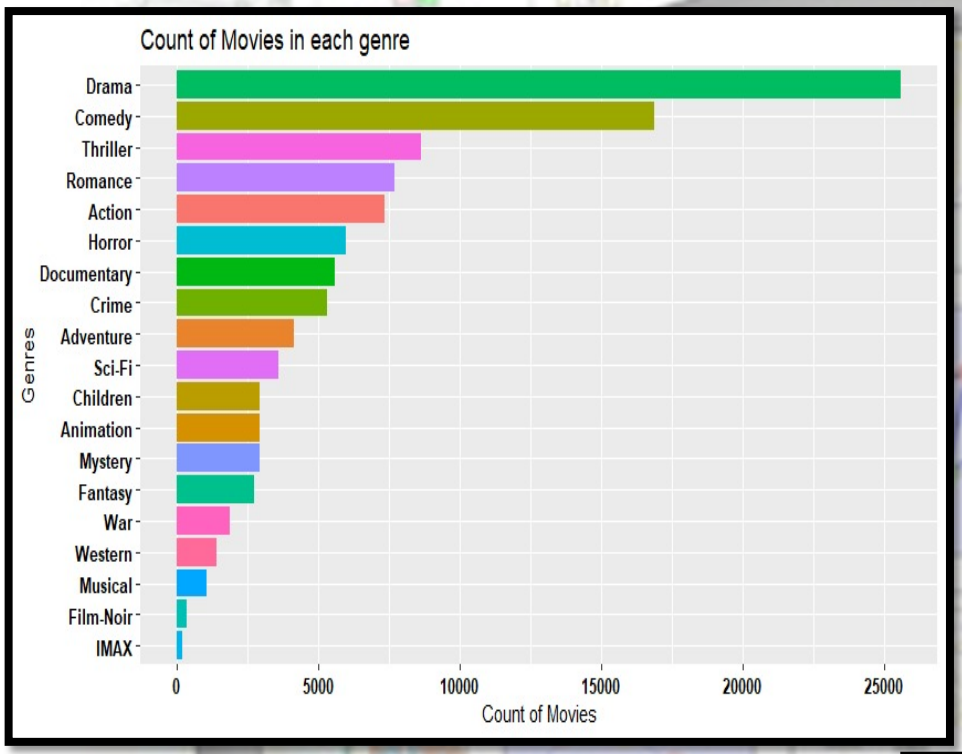
userId	movieId	rating
12	2	2.0
462	3	3.0
75989	4	2.0



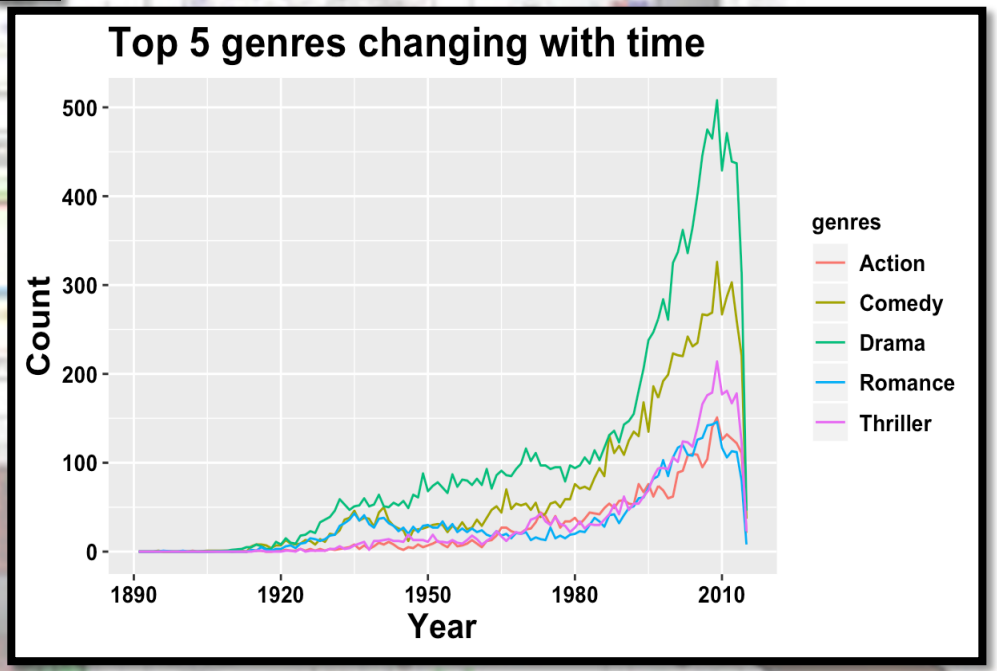
- MovieLens dataset taken from GroupLens
- The data has 25 million ratings given by ~162k users across~62k movies

userId	movieId	rating	title	genres
12	2	2.0	Jumanji (1995)	Adventure|Children|Fantasy
462	3	3.0	Grumpier Old Men (1995)	Comedy|Romance
75989	4	2.0	Waiting to Exhale (1995)	Comedy|Drama|Romance

Exploratory Data Analysis



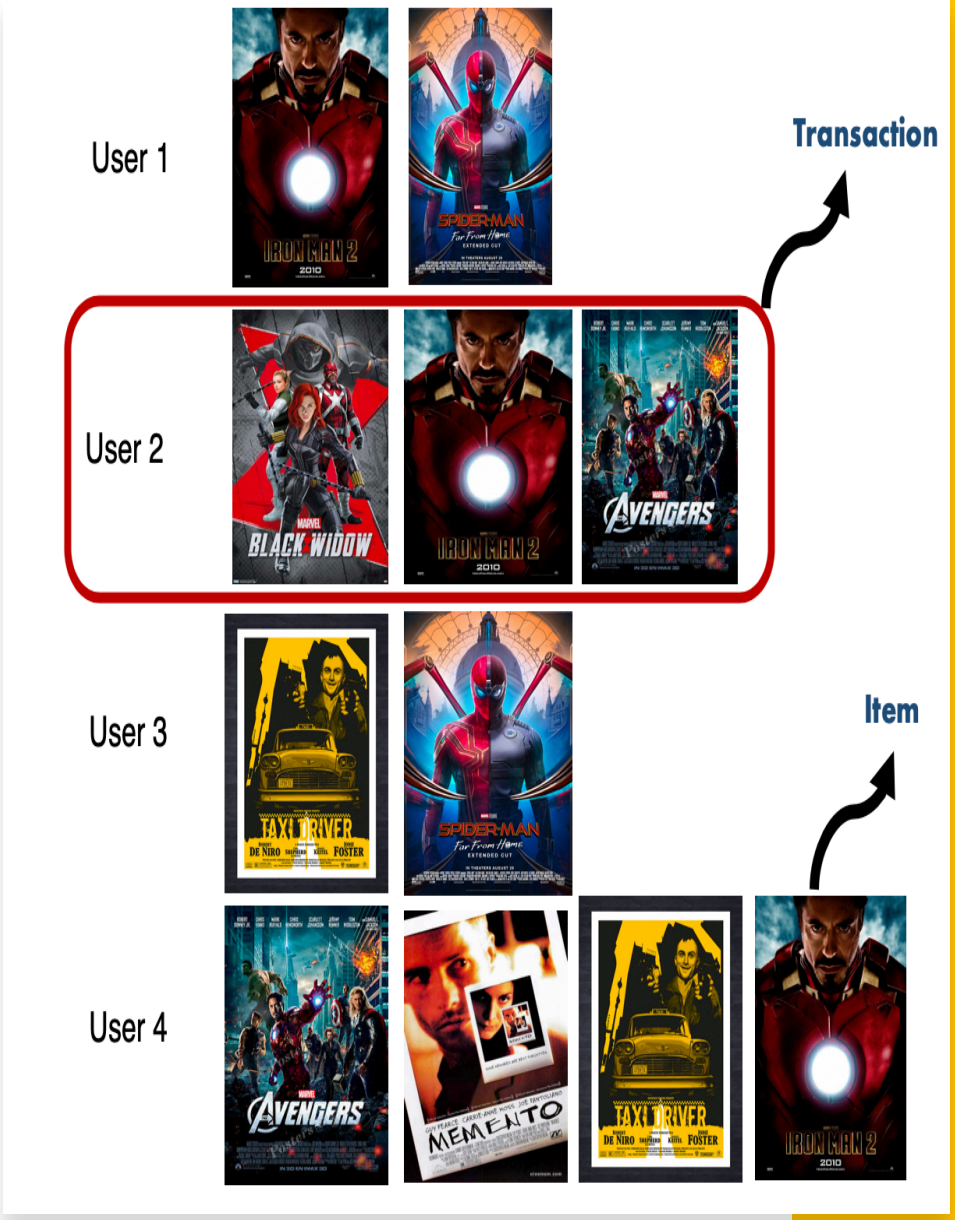
Maximum number of movies belong to drama genre followed by comedy



All the genres show an increasing trend in number of movies with time

Market Basket Analysis

- Movies frequently watched together are identified using MBA
- Generate Association Rules using Apriori & FP growth
  - Parameters: support : 0.05
    confidence : 0.7
    lift >= 1.2
- Rules with highest conviction and lift are selected
- FP growth runs faster, uses less memory



Association Rules using Market Basket

antecedents	consequents	support	confidence	lift	leverage	conviction
(5952.0, 7153.0, 260.0)	(4993.0)	0.051365	0.932901	5.350606	0.041765	12.304874
(5952.0, 7153.0, 1210.0)	(4993.0)	0.058114	0.927568	5.320017	0.047190	11.398841
(5952.0, 7153.0, 6539.0)	(4993.0)	0.053841	0.925585	5.308646	0.043699	11.095129
(5952.0, 7153.0, 1196.0)	(4993.0)	0.050756	0.925036	5.305500	0.041189	11.013966

Movies a user has watched:

movieId	title	year	genres
260	Star Wars: Episode IV - A New Hope	1977	Action|Adventure|Sci-Fi
5952	Lord of the Rings: The Two Towers, The	2002	Adventure|Fantasy
7153	Lord of the Rings: The Return of the King, The	2003	Action|Adventure|Drama|Fantasy

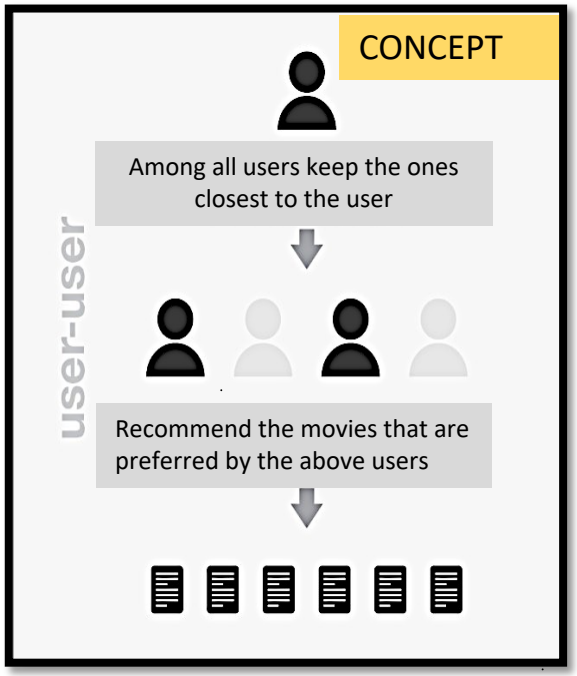
Our Recommendation:

movieId	title	year	genres
4993	Lord of the Rings: The Fellowship of the Ring,...	2001	Adventure|Fantasy

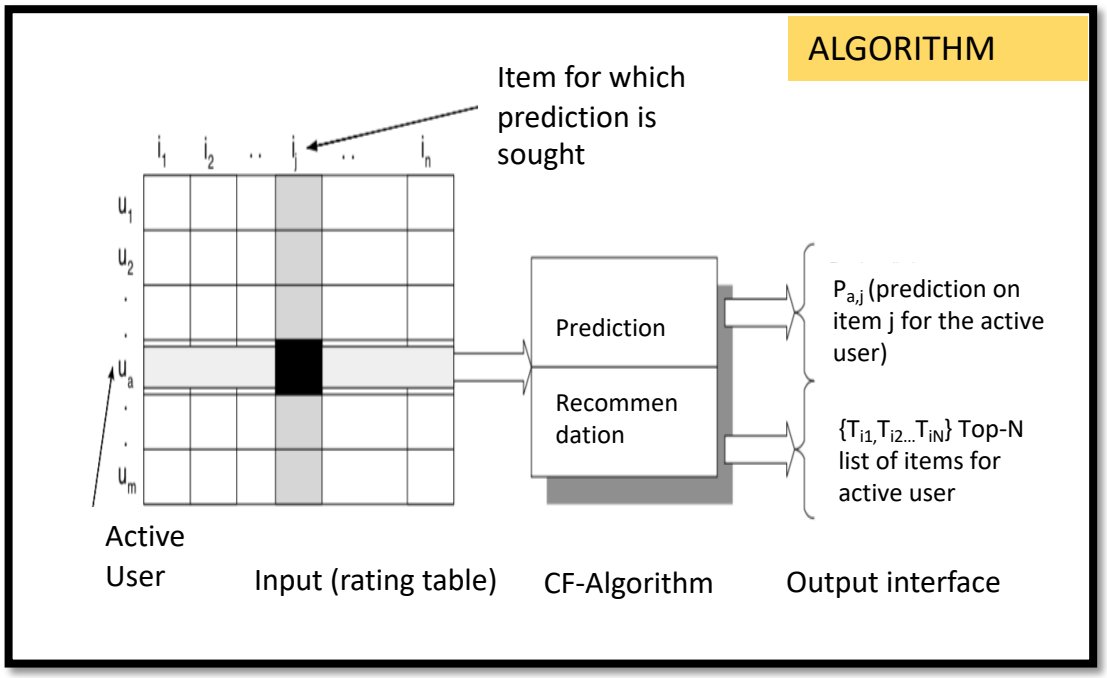
leverage(A→C) = support(A→C)−support(A)×support(C), range: [-1,1]
conviction(A→C) = 1−support(C)/1−confidence(A→C), range: [0,∞]

Collaborative Filtering – User based

CONCEPT



ALGORITHM



RESULT

```
# Movies user 1 would like
User_based_recom
```

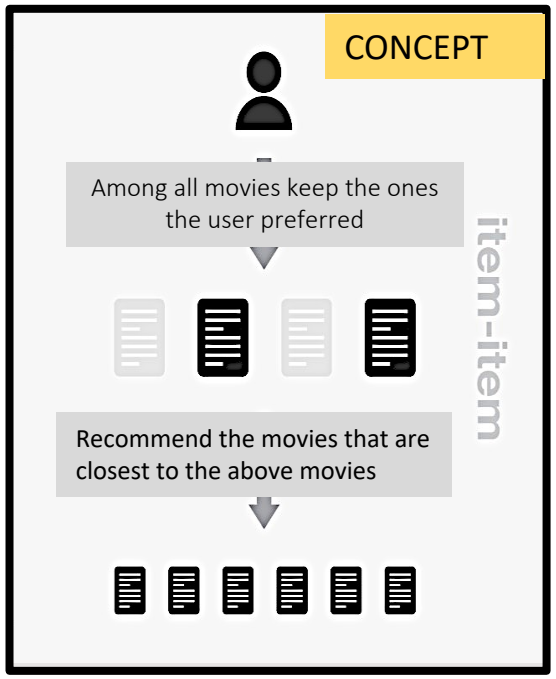
	movieId	score	title	genres
0	881	1.255747	First Kid (1996)	Children|Comedy
1	766	0.702344	I Shot Andy Warhol (1996)	Drama
2	854	0.702344	Ballad of Narayama, The (Narayama Bushiko) (1958)	Drama
3	805	0.667178	Time to Kill, A (1996)	Drama|Thriller
4	231	0.255747	Dumb & Dumber (Dumb and Dumber) (1994)	Adventure|Comedy
5	235	0.255747	Ed Wood (1994)	Comedy|Drama

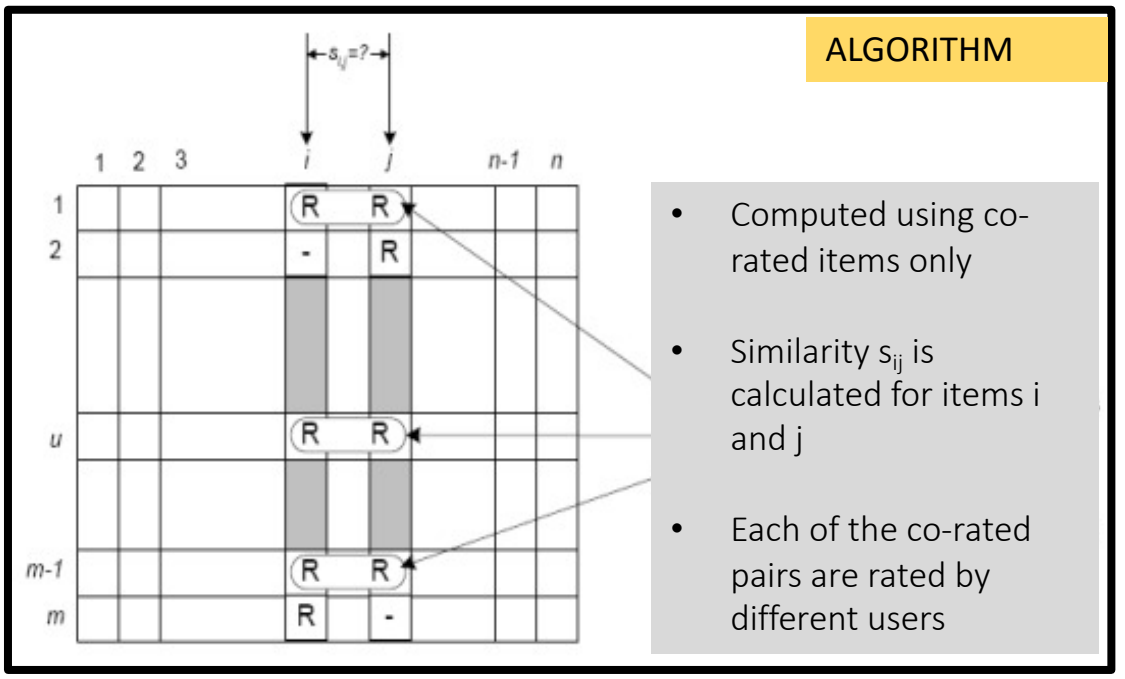
Based on ratings of similar users, user Id 1 is recommended the movies on the left in decreasing order of scores

$$R_U = (\sum_{u=1}^{n} R_u * S_u)/(\sum_{u=1}^{n} S_u)$$

R_U = Predicted rating
S_u= Similarity score as weights
R_u = Rating of other similar users

Collaborative Filtering- Item based





RESULT

```
1 pd.merge(movie_based_recom('Spider-Man (2002)'), df_movies, on='title').head(10)
```

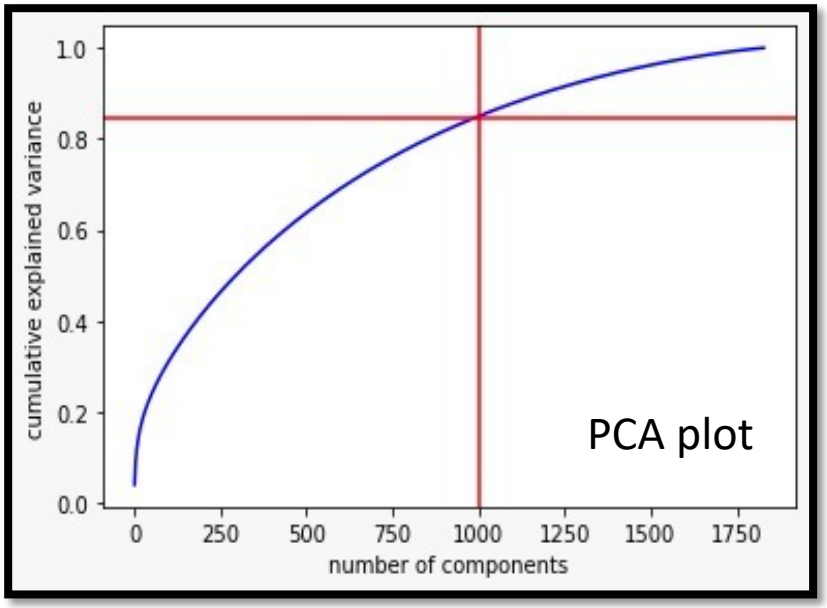
	title	cosine_sim	movieId	genres
0	Spider-Man (2002)	1.000000	5349	Action|Adventure|Sci-Fi|Thriller
1	Spider-Man 2 (2004)	0.692453	8636	Action|Adventure|Sci-Fi|IMAX
2	X-Men (2000)	0.635151	3793	Action|Adventure|Sci-Fi
3	Minority Report (2002)	0.632360	5445	Action|Crime|Mystery|Sci-Fi|Thriller
4	Pirates of the Caribbean: The Curse of the Bla...	0.625582	6539	Action|Adventure|Comedy|Fantasy
5	X2: X-Men United (2003)	0.625271	6333	Action|Adventure|Sci-Fi|Thriller

Movie recommendations for people who like Spider Man based on Item Based Collaborative Filtering

Clustering Users

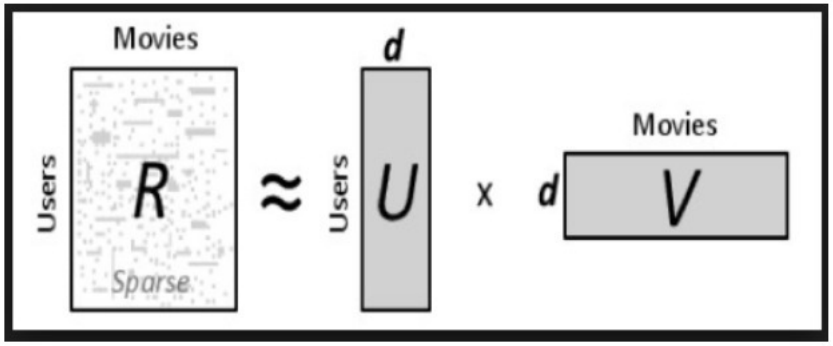
movieId	1	2	3	5	6	7
userId						
548	4.5	4.0	NaN	NaN	4.0	NaN
1748	4.5	3.0	2.5	4.0	4.0	3.5
2177	4.0	3.0	4.0	3.0	3.0	4.0
8619	4.5	2.5	2.5	2.0	4.5	4.0
12244	4.0	4.5	NaN	NaN	4.0	NaN

Snapshot of user-movie rating matrix used for clustering users



- Original no. of features = 1830
- No. of features explaining ~80% variance = 1000

- Both KMeans and GMM performed poorly on the sampled data because of high sparsity
- Performed SVD to convert sparse matrices to dense matrices

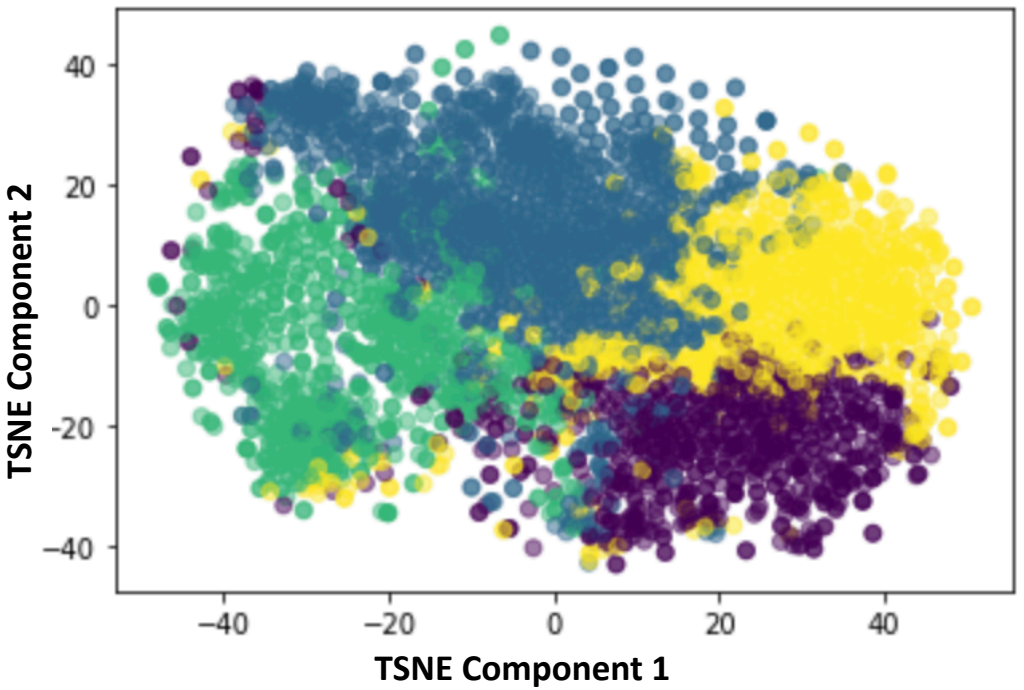


SVD to get dense matrices, d=50

Clustering Users

- K-means did not give satisfactory results
- Clustered using GMM
- GMM works better when clusters are overlapping
- Least BIC score for K=4

Clustering results for 4 clusters



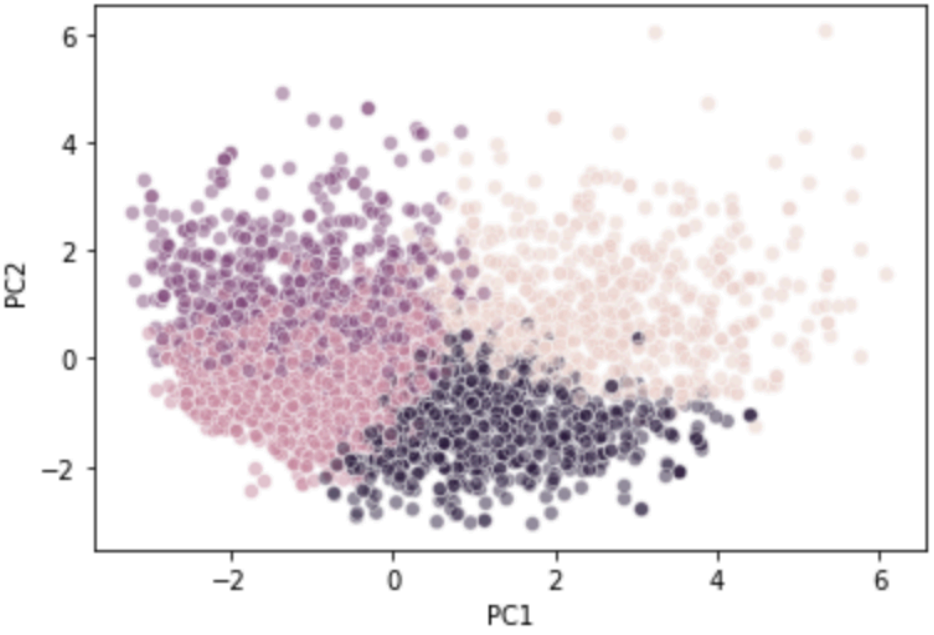
Recommendation Method:

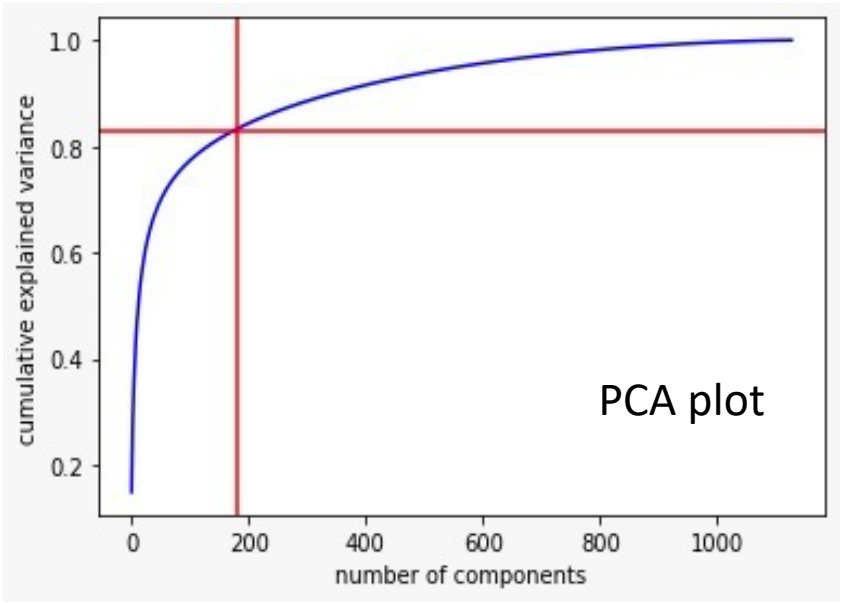
- Filter all the users of the cluster the user belongs to
- Filter all the movies the users mentioned above have watched
- Filter the top-rated movies from these
- Recommend those movies user has not already watched

Clustering Movies

tag movieId	007	007 (series)	18th century	1920s	1930s
176887	0.03325	0.04325	0.30975	0.09550	0.63375
247	0.02250	0.02200	0.06075	0.11125	0.09250
127218	0.02575	0.03400	0.02975	0.07550	0.08525
5810	0.03225	0.02775	0.02300	0.03850	0.06400
1002	0.02500	0.02800	0.02550	0.01575	0.05775

Snapshot of movie-tag score matrix used for clustering movies





- Original no. of features = 1128
- No. of features explaining ~80% variance = 200

Recommendation Method:

- Select cluster the user has watched maximum number of movies from
- Top-rated movies of this cluster are recommended.

Clustering Results for Recommendation

```
x=recommendation(1652)
x['user_watched']
```

	movieId	title
52395	593	Silence of the Lambs, The (1991)
167670	195159	Spider-Man: Into the Spider-Verse (2018)
195394	16	Casino (1995)
541703	4432	Sweet Smell of Success (1957)
747819	48516	Departed, The (2006)
794633	4262	Scarface (1983)
1362366	1653	Gattaca (1997)
1483612	1259	Stand by Me (1986)
1755348	81932	Fighter, The (2010)
1958766	74545	Ghost Writer, The (2010)

Movies user ID 1652 has watched

```
x['movies_recommended']
```

	movieId	title
1679066	858	Godfather, The (1972)
4564028	8183	Educating Rita (1983)
5565177	4292	Norma Rae (1979)
6580544	50	Usual Suspects, The (1995)
1340639	3467	Hud (1963)
151051	1235	Harold and Maude (1971)
1640850	858	Godfather, The (1972)
1861245	4226	Memento (2000)
6699104	79132	Inception (2010)
551467	91529	Dark Knight Rises, The (2012)

Movie recommendations for user ID 1652

Conclusion

- Basic recommendation system can be built using CF and cosine similarity
- K-means works best for non-overlapping data points
- GMM works best when data points are overlapping
- Huge sparse matrices often must be converted to dense matrices for better performance
- Clustering is better suited for large-scale problem in recommender systems

Proposed Work

- Hybrid approach:

  Merge multiple criteria to build a more robust system
- Compare the different clustering algorithms
- Evaluate the accuracy of our recommendation system
- Deploy an interactive web application on cloud