CS 6630: Visualizing Movies Metadata

Ram Kashyap S and Mohammed Musaddiq

Nov 10, 2017

1 Basic Info

Project Title: Visualizing Movies Metadata

Team member 1: Ram Kashyap S
Email: u1082810@utah.edu
uID: u1082810

Team member 2: Mohammed Musaddiq
Email: mohammed.musaddiq@utah.edu
uID: u1068996

Project repository:
`https://github.com/ramkashyap-s/dataviscourse-pr-movies-viz`

2 Overview and Motivation

The reason we chose this project is the relevance and appeal it would have to a public audience considering the millions of movie fans across the world. Furthermore, being movie buffs ourselves, we are also motivated by personal interest in visualizing different aspects of movies and sharing the same fellow fans.

3 Related Work

We explored many example visualizations online, assessed them according to our project scope and arrived at our final design. We are particularly inspired by this project `https://cips1.engineering.asu.edu/movie_analysis/`

The table and line charts that we implemented in our milestone are similar to the ones that we have seen in class.

We followed the layouts tutorial from the class to implement the node link diagram and added few enhancements to it.

4 Questions

The primary questions we are trying to answer with our visualization are:

- For a given actor/director, view how the following parameters vary over time:
  - Rating of movies they have acted in/directed
  - Gross earnings of movies they have acted in/directed
  - Budget of movies they have acted in/directed
  - Number of User reviews

The above visualization would allow users to get valuable insights into how an actor's/director's movies have fared over time based on various aspects they are interested about and wish to compare.

- For a set of filters such as movie rating, year(s) and genre,
  - Which are the movies that match the criteria? (in the form of a table)
  - Which actor(s) or director(s) are the most prominent in the selection range?
  - How well connected are the actor(s) and director(s)?

This visualization would help users in finding movies based on the genre(s) they like, how recent or old the movie is or how good/bad the movie's ratings are. Further, they can also view node links diagrams between movies, actors and directors.

- Exploring data correlation between rating and attributes like duration, number of user reviews, movie facebook likes and gross by using a scatter plot

5 Data

We have obtained the metadata for 5000+ movies spanning across 100 years in 66 countries from here:

`https://github.com/sundeepblue/movie_rating_prediction/blob/master/movie_metadata.csv`

The data contains 28 variables and close to 5000 movie records. There are 2399 unique director names and thousands of actors/actresses.

5.1 Data Processing

Our data source is from a csv file which has empty values in some columns for few records

- For table, we are checking for empty/missing string values. We are checking for null/empty 'NaN' values in plot.
- In the movie titles column, we have an extraneous character Â which is due to a different encoding. But, it isn't affecting the visualizations.

We decided to handle data cleaning on the fly. We tried to delete the movie records which had empty/zero values in columns. But, this strategy proved to be unhelpful as there are non-empty columns which we could use in our visualizations.

We have removed the movies where movie title, actor and director information are not present.

For node-link diagrams, we had to process the data and create appropriate data structures for creating nodes and edges between them.

6 Visualization Design

Initial Design 1

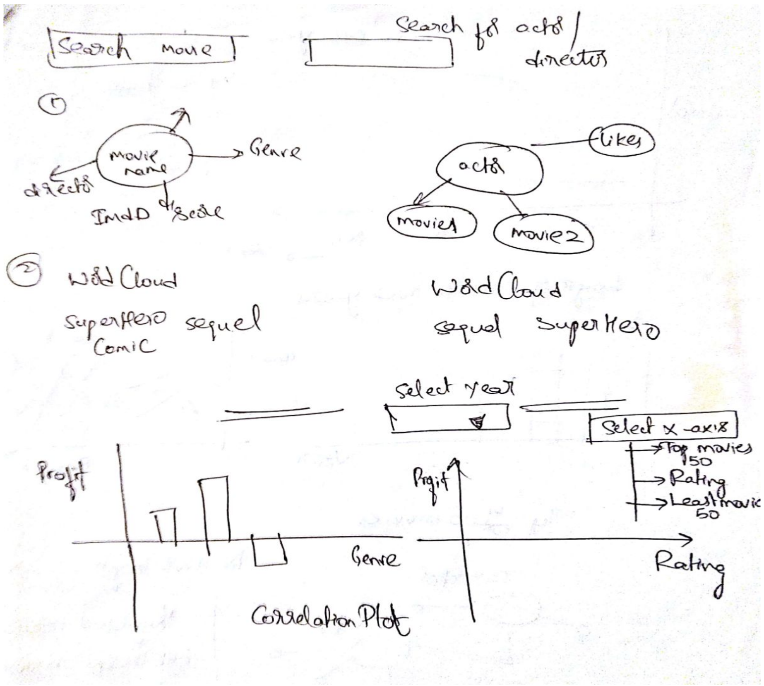


Figure 1: Initial Design 1

Implement a search filter that would allow the user to select a movie and an actor/director. Based on the selection, visualize the movie and actor/director as a graph with the movie and actor/director as central nodes and their associated metadata as child nodes attached to them.

We felt there might be a need to allow users to find/explore movies in other ways instead of restricting them to using a search filter that allows them to filter only by title.

Visualize the movie's plot keywords using a word cloud that would give the user some idea about what the movie is about.

Allow the user to select various movie attributes such as genre, rating using a drop-down list and visualize their correlation with movie profit using bar-chart.

Initial Design 2

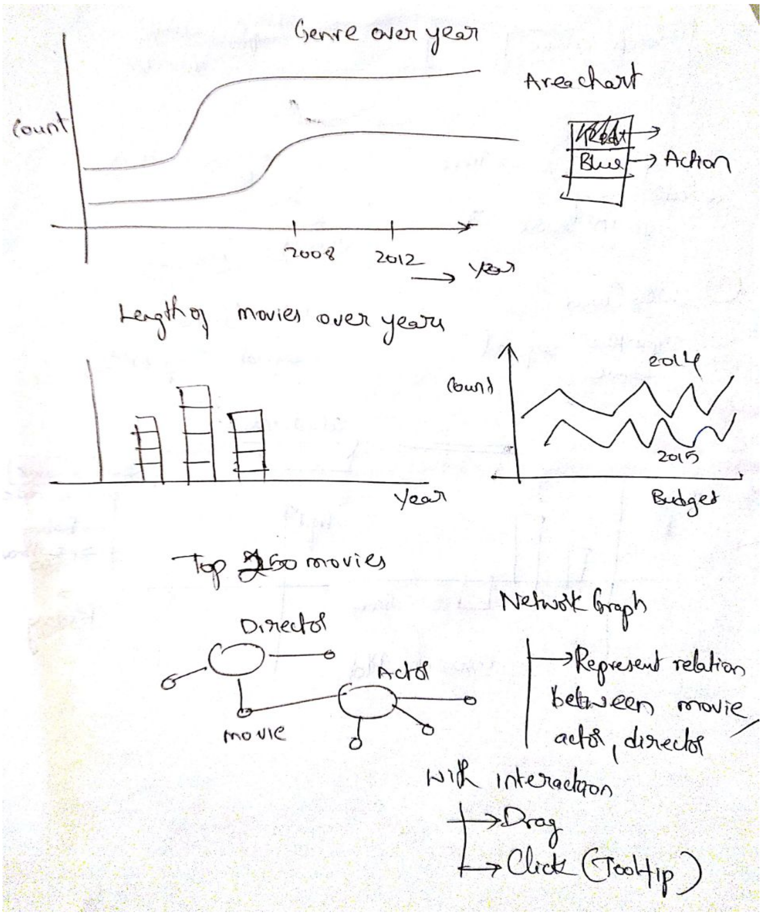


Figure 2: Initial Design 2

For top 250 movies implement a network graph which will represent relations between movie, director and actor. The size of the node for actor and director would be according to the number of connections(degree). The graph would be interactive with drag for visualizing the connections and click for visualizing the details of node.

We feel that users might not get much information by looking at the connections in the graph.

Exploring data:

Implement area chart for genre trend over the years. Hue is used to encode different genres.

Implement a stacked bar chart for duration of the movies over the years.

Implement a line chart for budget over the years

Initial Design 3

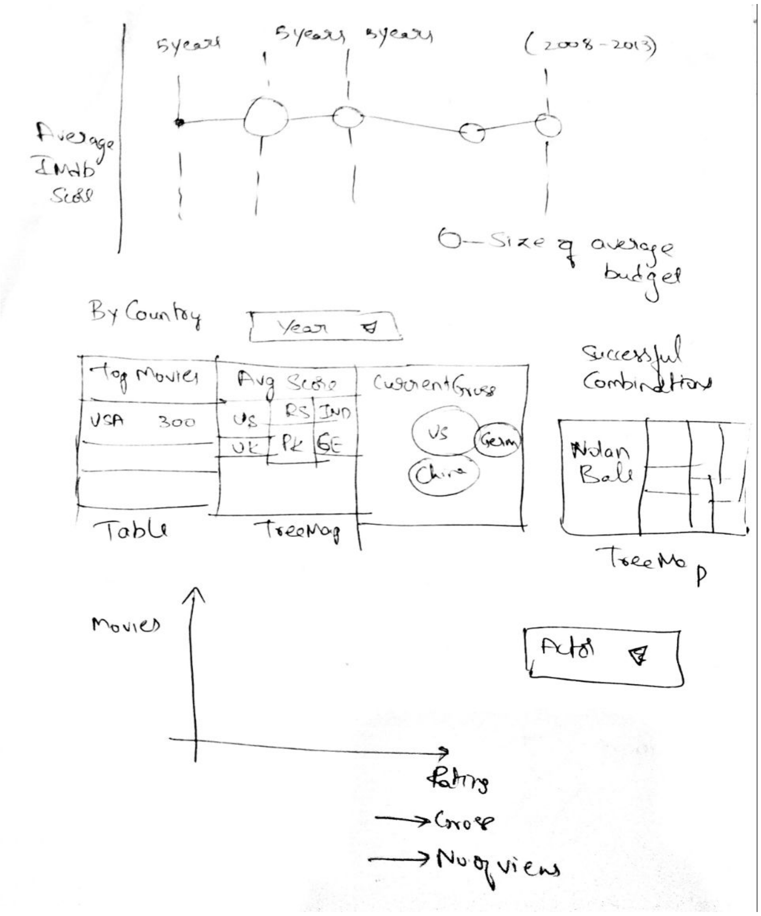


Figure 3: Initial Design 3

Implement a dashboard with different views

View by country: By each country display top twenty movies contributed

Implement a tree map with average scores

Implement a heat map of countries by selected year gross

View by actor and director:

Implement a tree map with successful actor and director combinations

Implement bar chart for top movies by rating, grossing, number of reviews, etc. for an actor or a director

Initial Finalized Design(Proposed in project proposal)

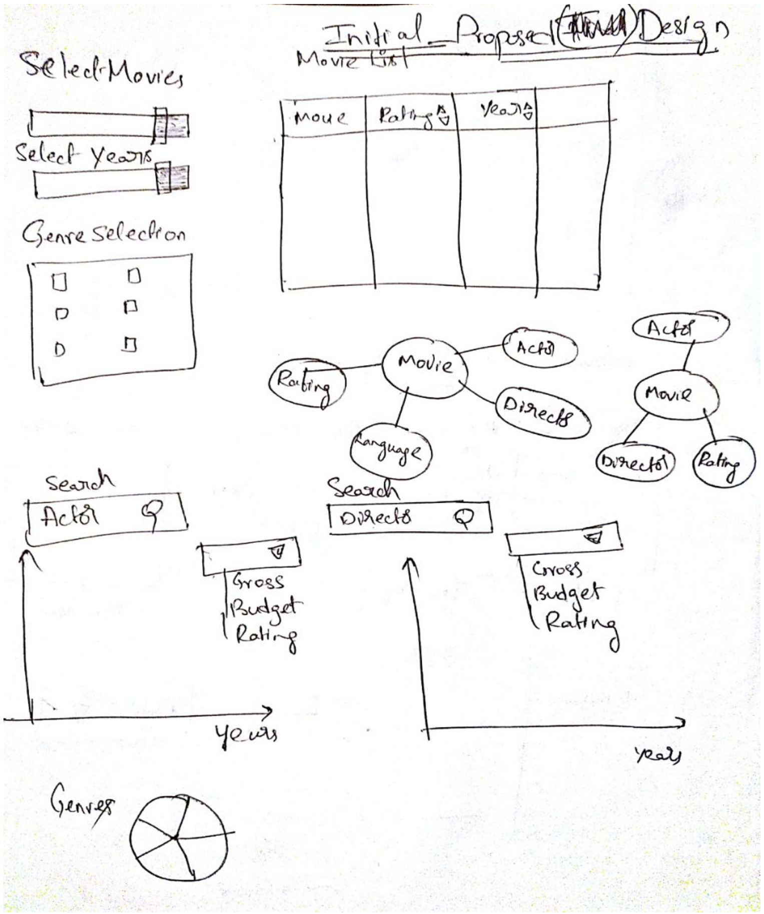


Figure 4: Initial Finalized Design

This is the design which we proposed to implement in our project proposal.

Implement a movie rating filter enable selection of ratings and implement same for the years as well. We will also provide single/multiple genre selection using check-boxes.

We feel this is better than our initial design since earlier, the users did not have the ability to apply certain obvious filters to their search in case they did not know a movie name to begin with and just wanted to explore.

Then, we would visualize all movies matching the above criteria specified by the user using a dynamic table.

Finally, the user can select a movie from the above list he wants to know more about. We will represent that movie as a graph with the movie as a central node and its associated metadata as child nodes attached to it.

Next, for the Actor/Director visualization, implement a search filter that would allow the user to search for and select an Actor/Director of interest. Based on the selection, provide a bar-chart based visualization of how different attributes of their movies have changed over the years. These attributes can be chosen using a drop-down list.

6.1 Changes from project proposal

Node Link diagram:

In the initial design we have proposed to show attributes like rating, content rating and others for each movie using node-link diagrams. Later, we realized that we wouldn't get much inference out of such a design and changed the design.

Now we show connections between movie, actors, and directors with movie as the central node and connections to the three actors and the director from the data set, where size of node corresponds to the degree of that node.

We added additional features such as hover and highlighting neighboring nodes on double-click.

Line Chart: In our visualization deign, we realized that representing movies as circles is much suited for our plots, than a bar chart. So, we are implementing scatter plots and line charts accordingly.

Based on the discussion with professor, we decided to remove the pie chart for genres as a movie can have more than one genre and it would be complicated to get any inference from such a chart.

Filters: For filters, we proposed sliders in the initial design which wouldn't have the ability for selection of values in a range. Now, instead we implemented filters using d3 brush such that we can select a range of values.

Filters: We changed the layout and placed line charts and scatter plots in one view and in another view we have table and node-link diagrams.

Implemented Design

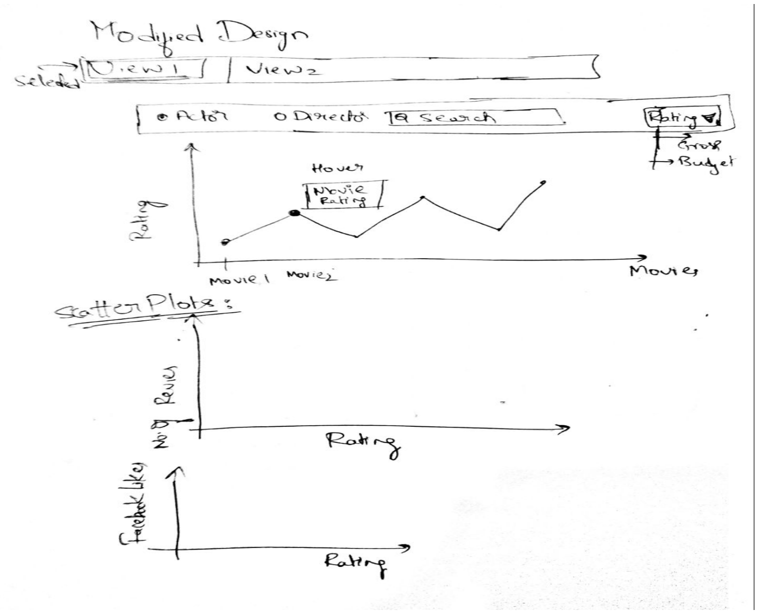


Figure 5: View-1 Implemented Design

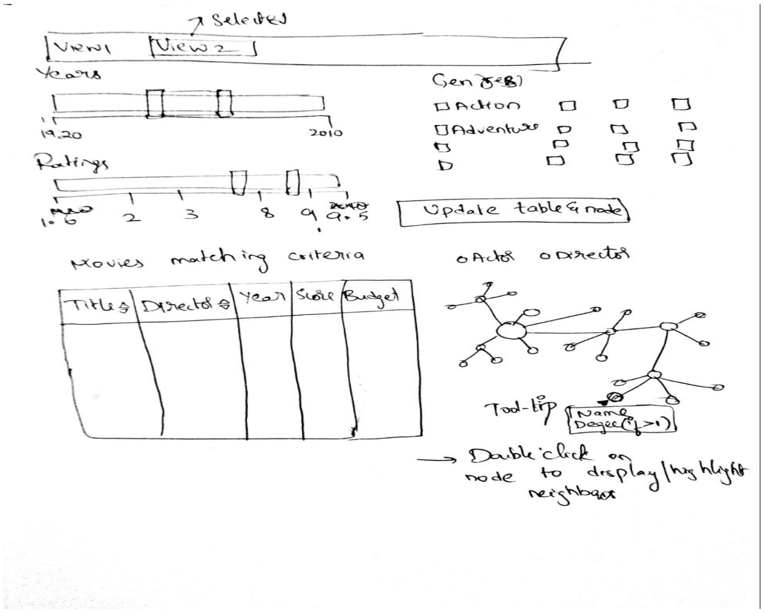


Figure 6: View-2 Implemented Design

7 Exploratory Data Analysis

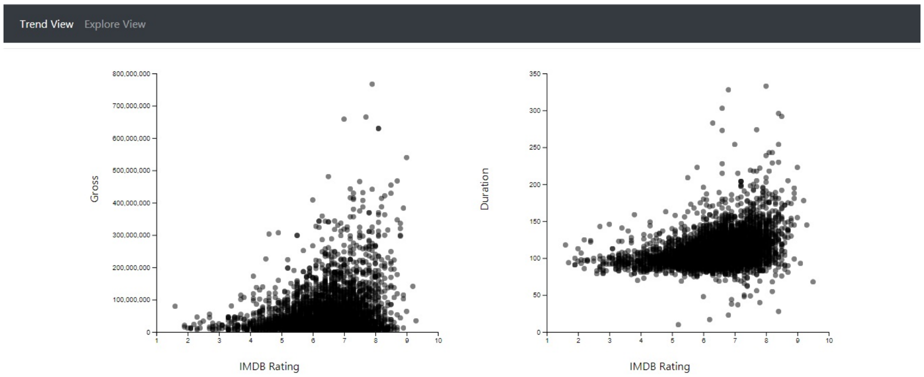


Figure 7: Plot - rating vs duration

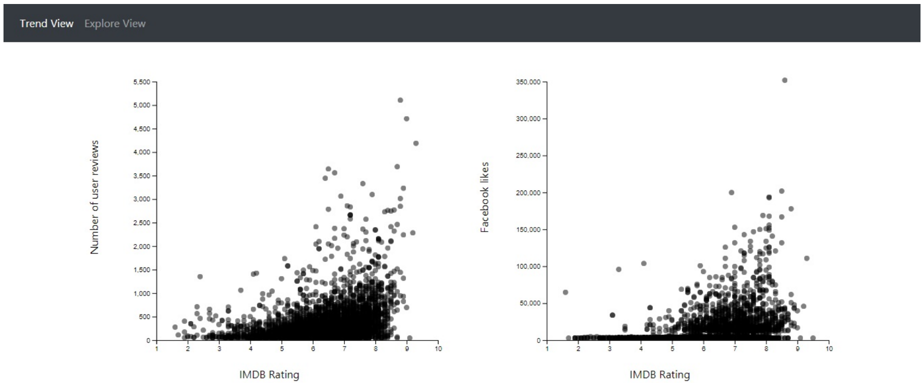


Figure 8: Plot - rating vs number of user reviews

We have created few scatter plots like the ones shown above.

We observed that rating is correlated with the attributes such as number of reviews, number of movie facebook likes, gross and duration

Also, we observed that there are many outliers present in the data.

8 Peer feedback

We received suggestions to implement some of the features such as:

Graph connections based on directors and actors to compare two movies from the list.

A node-link diagram that shows connections between actors and directors

Word cloud with a summary of the movie

Integrate data with the Oscar nominations data

Pie chart for genres might not work as each movie might have more than one genre

Add gross and budget to the metrics

Questions asked during peer feedback:

Based on the feedback, we improved some features such as node-link diagrams and some components in line chart.

We have implemented word cloud, but we observed that the plot words of movie are not repeating enough to generate a word cloud based visualization. So we dropped it from our final submission.

9 Design Evolution:

9.1 Filters

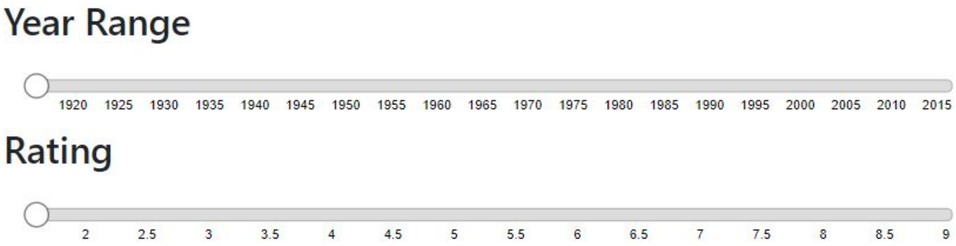


Figure 9: Selectors

The filters enable the user to narrow down what kind of movies they want to look at by having them choose their preferences. The filters we have chosen are parameters that someone would most commonly look at when trying to explore movies.

These include:

(i). Selecting years (ii). Selecting ratings (iii). Selecting genres

We created sliders such that a range of values are considered i.e. from the maximum year possible to the current slider position.

Also, we have populated the check boxes required for selecting the genres from the dataset.

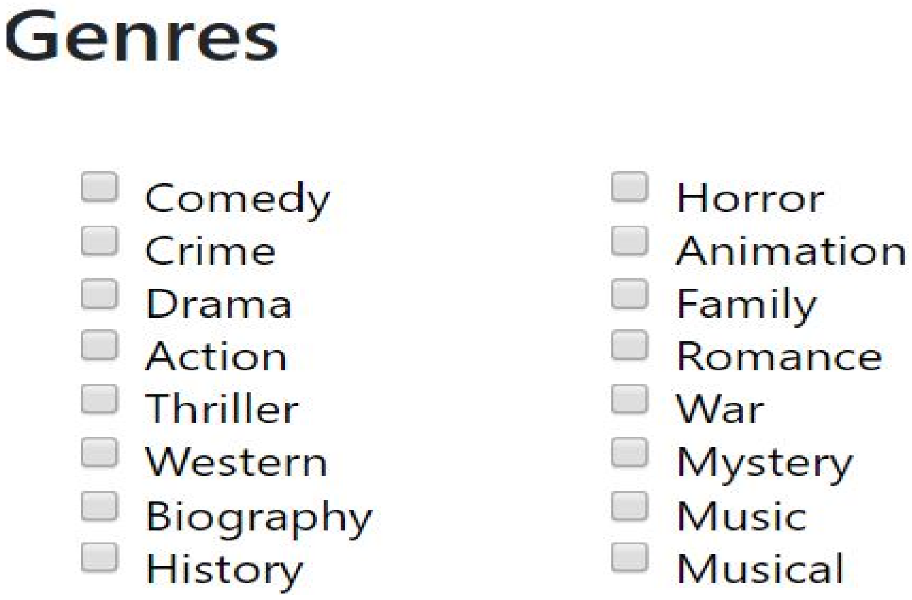
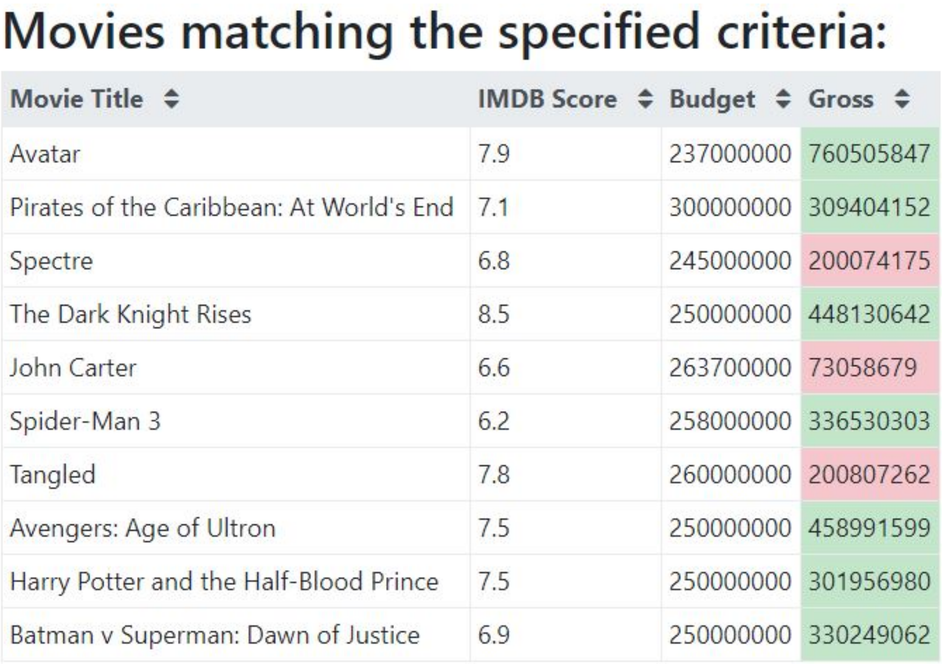


Figure 10: Selectors

9.2 Table

The intent behind the table is to list all the movies (along with attributes such as IMDB rating, Budget and Gross) that match the filter criteria specified by the user. As for the interactive functionality, we have enabled sorting on each column that would allow the user to easily sort the entries in ascending/descending order.



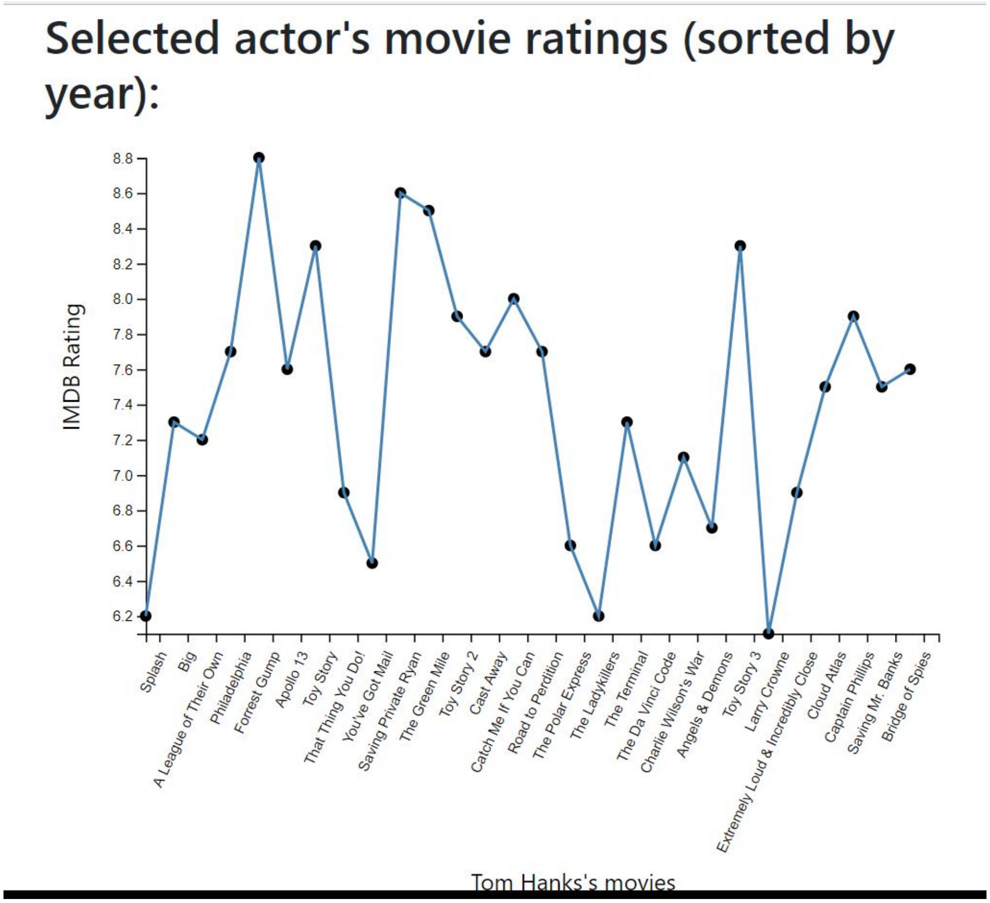
Movies matching the specified criteria:

Movie Title	IMDB Score	Budget	Gross
Avatar	7.9	237000000	760505847
Pirates of the Caribbean: At World's End	7.1	300000000	309404152
Spectre	6.8	245000000	200074175
The Dark Knight Rises	8.5	250000000	448130642
John Carter	6.6	263700000	73058679
Spider-Man 3	6.2	258000000	336530303
Tangled	7.8	260000000	200807262
Avengers: Age of Ultron	7.5	250000000	458991599
Harry Potter and the Half-Blood Prince	7.5	250000000	301956980
Batman v Superman: Dawn of Justice	6.9	250000000	330249062

Figure 11: Table

9.3 Line chart

The intent behind the line chart is to analyze how a particular actor's or director's movies have performed over the years using their respective IMDB ratings. As for the interactive functionality, we implemented a search filter that would allow selection of actor and line chart would update dynamically. Also, we have a tool-tip to show the value for a movie.

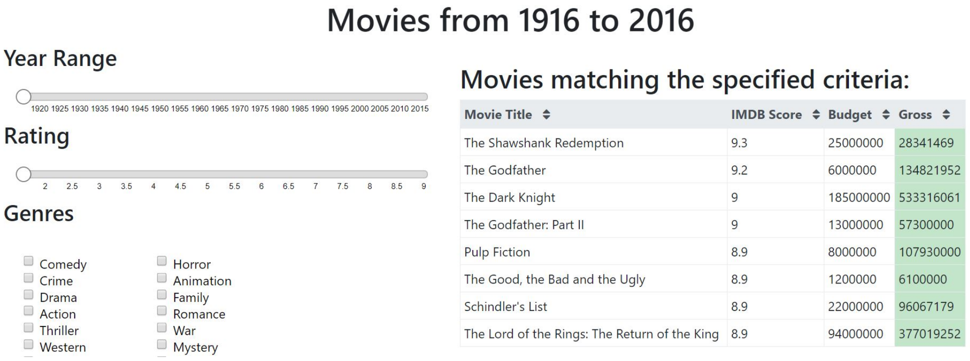


Selected actor's movie ratings (sorted by year):

Figure 12: Actor/Director plot

9.4 Layout

This was our initial layout of filters and table when we submitted our milestone:



Movies from 1916 to 2016

Year Range

Rating

Genres

☐ Comedy
☐ Crime
☐ Drama
☐ Action
☐ Thriller
☐ Western

☐ Horror
☐ Animation
☐ Family
☐ Romance
☐ War
☐ Mystery

Movies matching the specified criteria:

Movie Title ⬍	IMDB Score ⬍	Budget ⬍	Gross ⬍
The Shawshank Redemption	9.3	25000000	28341469
The Godfather	9.2	6000000	134821952
The Dark Knight	9	185000000	533316061
The Godfather: Part II	9	13000000	57300000
Pulp Fiction	8.9	8000000	107930000
The Good, the Bad and the Ugly	8.9	1200000	6100000
Schindler's List	8.9	22000000	96067179
The Lord of the Rings: The Return of the King	8.9	94000000	377019252

Figure 13: Initial Layout

9.5 Modifications

9.5.1 Sliders

We have realized that we would need second slider handle for selecting movies in a range. So, we have implemented a second slider handle.

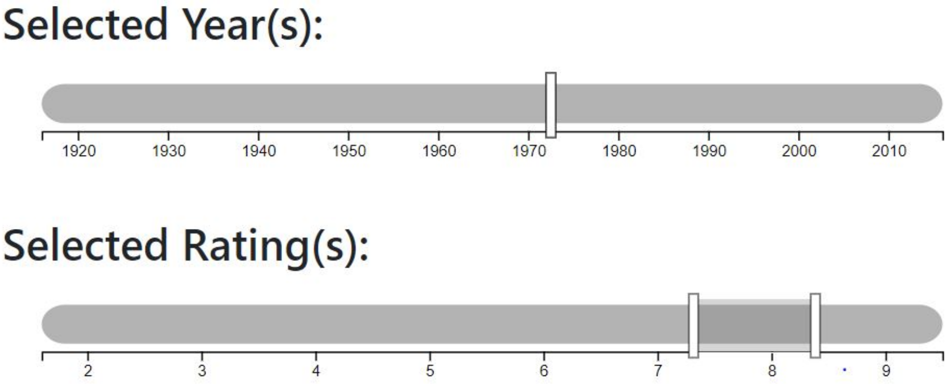


Figure 14: Actor/Director plot

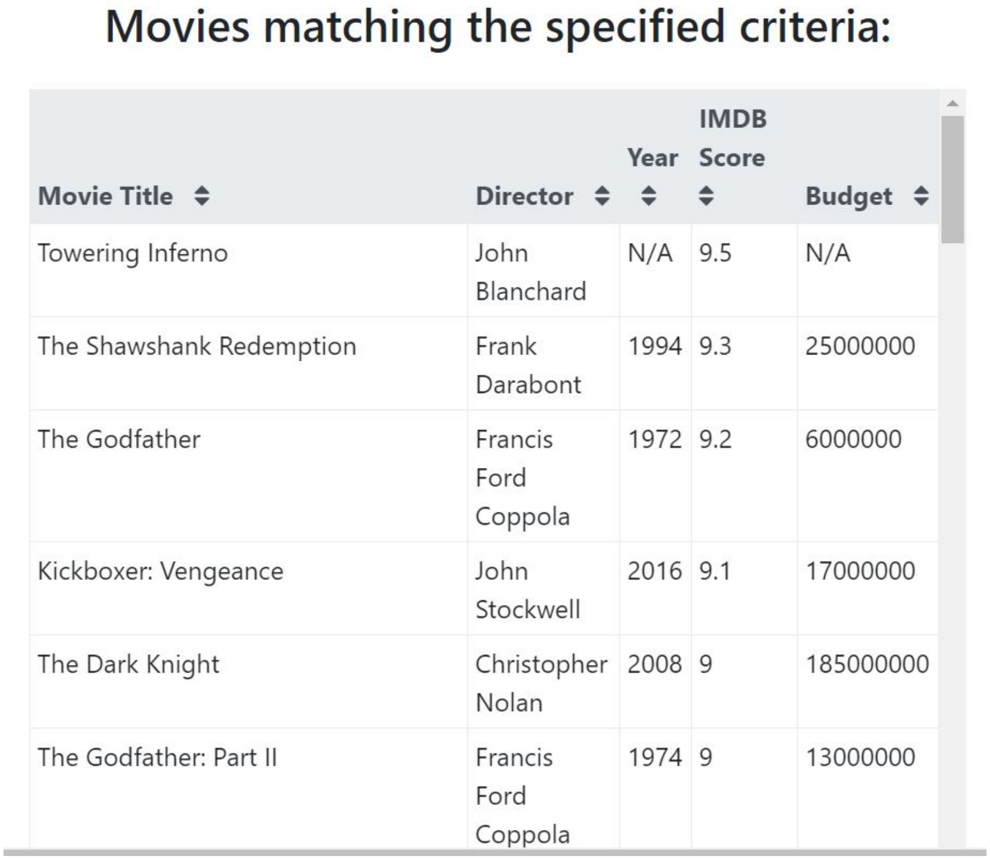
Problems faced:

We faced a problem while trying to visualize a selection of single year or rating as d3 brush is generally meant for a range of selection. The brush disappears as the start and end of the brush are the same.

Finally, we resolved this issue and the user would be able to select and visually see a single selection of year or rating.

9.5.2 Table

We have thought about the scenario where the number of movies matching the user's filter criteria could be too high to visualize efficiently. So we improved this table by implementing a scroll feature that would allow the user to scroll through any number of resultant entries without negatively impacting the screen space.



Movies matching the specified criteria:

Movie Title	Director	Year	IMDB Score	Budget
Towering Inferno	John Blanchard	N/A	9.5	N/A
The Shawshank Redemption	Frank Darabont	1994	9.3	25000000
The Godfather	Francis Ford Coppola	1972	9.2	6000000
Kickboxer: Vengeance	John Stockwell	2016	9.1	17000000
The Dark Knight	Christopher Nolan	2008	9	185000000
The Godfather: Part II	Francis Ford Coppola	1974	9	13000000

Figure 15: Actor/Director plot

9.5.3 Line chart

We improved this feature by including a selection of other attributes such as budget/gross to be analyzed for the selected actor.

Also, we have attached a tool-tip for the movies.

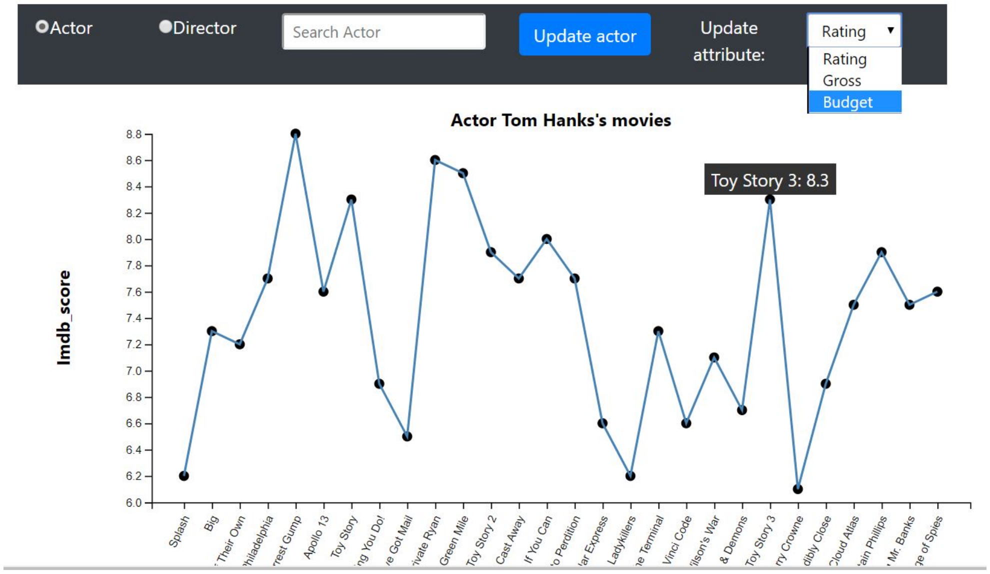


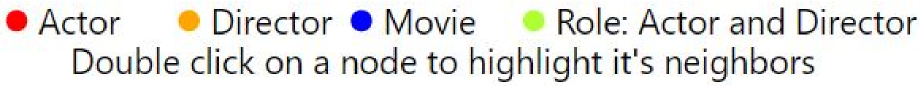
Figure 16: Actor/Director plot

Issues faced:

We faced performance problems while implementing search filter for actor or director. Since there are thousands of actors and directors.

We resolved this issue by storing a list of actors and directors as a global variable and this significantly improved thee performance.

9.5.4 Force Directed graph



● Actor ● Director ● Movie ● Role: Actor and Director

Double click on a node to highlight it's neighbors

Figure 17: Legend

We have created a legend assigning colors for nodes: directors, actors, movies and for a person acting as both actor and director

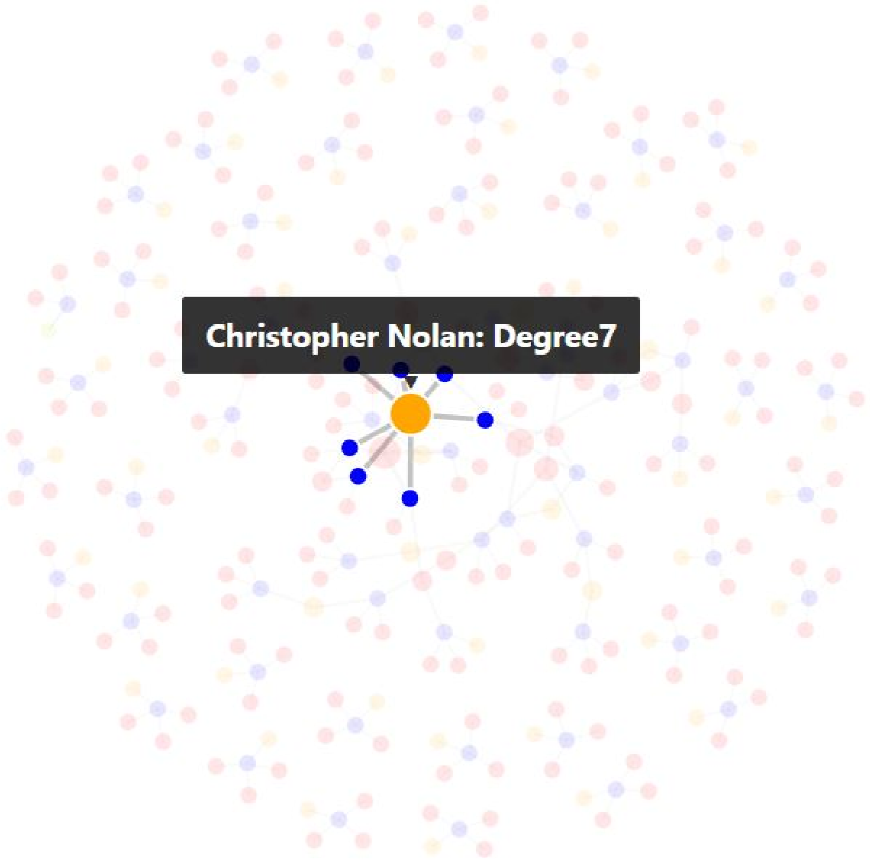


Figure 18: Force Directed Graph

When a user double clicks on a particular node we get all its connected nodes and we make rest of the nodes opaque.

We have created force directed graph with movies as central node and connections to actors and director. The size of the nodes correspond to the degree of the node.

We used d3 force layout for creating this force directed graph with charge and forces adjusted accordingly to suit out layout.

9.5.5 Layout changes

We used bootstrap and modified our layout as below

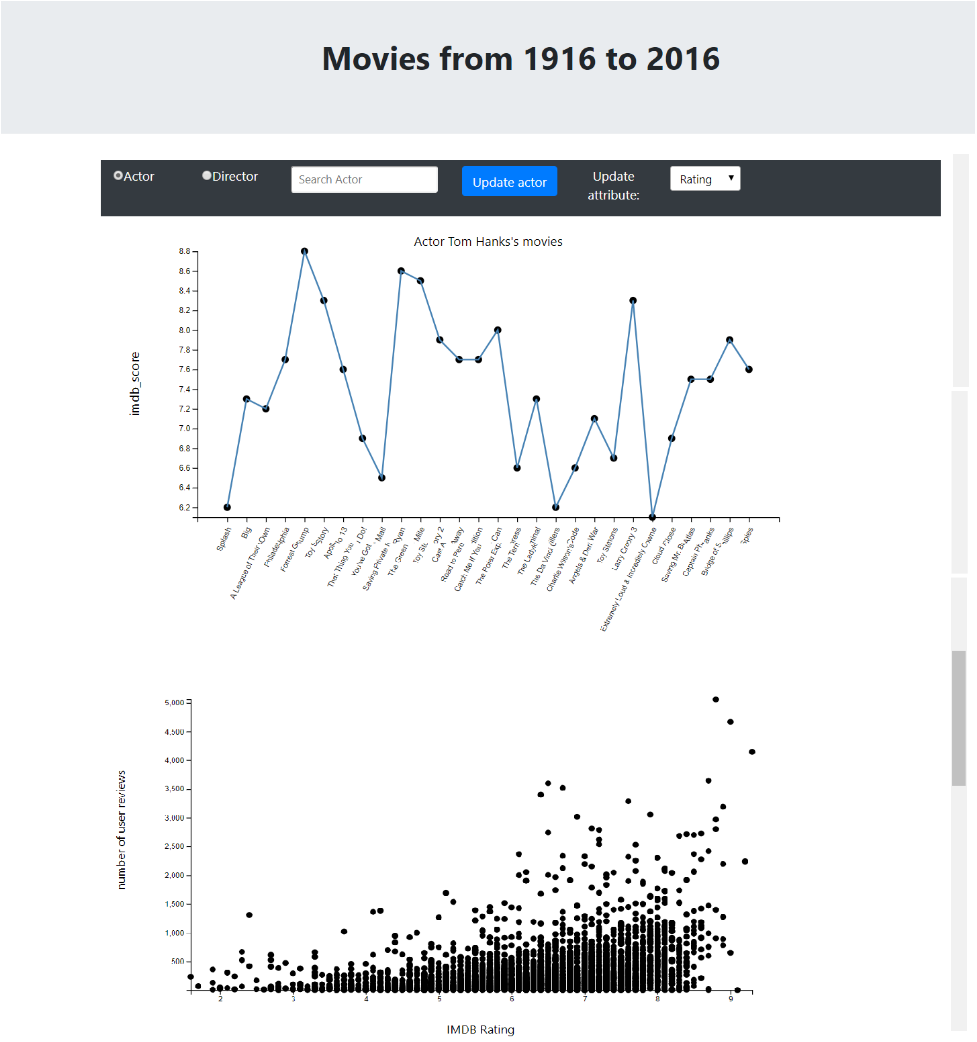
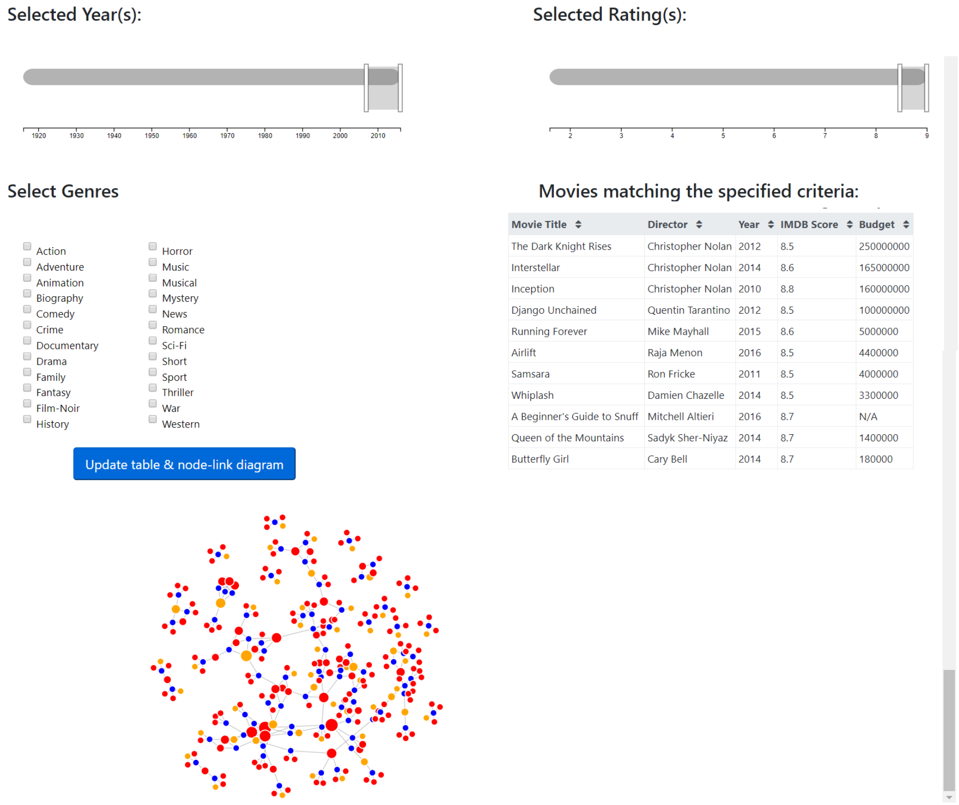


Figure 19: Modified Layout



Selected Year(s):

Selected Rating(s):

Select Genres

□ Action
□ Adventure
□ Animation
□ Biography
□ Comedy
□ Crime
□ Documentary
□ Drama
□ Family
□ Fantasy
□ Film-Noir
□ History
□ Horror
□ Music
□ Musical
□ Mystery
□ News
□ Romance
□ Sci-Fi
□ Short
□ Sport
□ Thriller
□ War
□ Western

Update table & node-link diagram

Movies matching the specified criteria:

Movie Title	Director	Year	IMDB Score	Budget
The Dark Knight Rises	Christopher Nolan	2012	8.5	250000000
Interstellar	Christopher Nolan	2014	8.6	165000000
Inception	Christopher Nolan	2010	8.8	160000000
Django Unchained	Quentin Tarantino	2012	8.5	100000000
Running Forever	Mike Mayhall	2015	8.6	5000000
Airlift	Raja Menon	2016	8.5	4400000
Samsara	Ron Fricke	2011	8.5	4000000
Whiplash	Damien Chazelle	2014	8.5	3300000
A Beginner's Guide to Snuff	Mitchell Altieri	2016	8.7	N/A
Queen of the Mountains	Sadyk Sher-Niyaz	2014	8.7	1400000
Butterfly Girl	Cary Bell	2014	8.7	180000

Figure 20: Modified Layout

Issues:

Still, we had to scroll three times to see all the visualizations.

Final Layout:
So, we have split our visualizations into two views: Trend View and Explore View

- Trend View has Actor/Director line charts
- Explore View has Table and Node Link diagrams that would get populated according to the options selected by the filters.

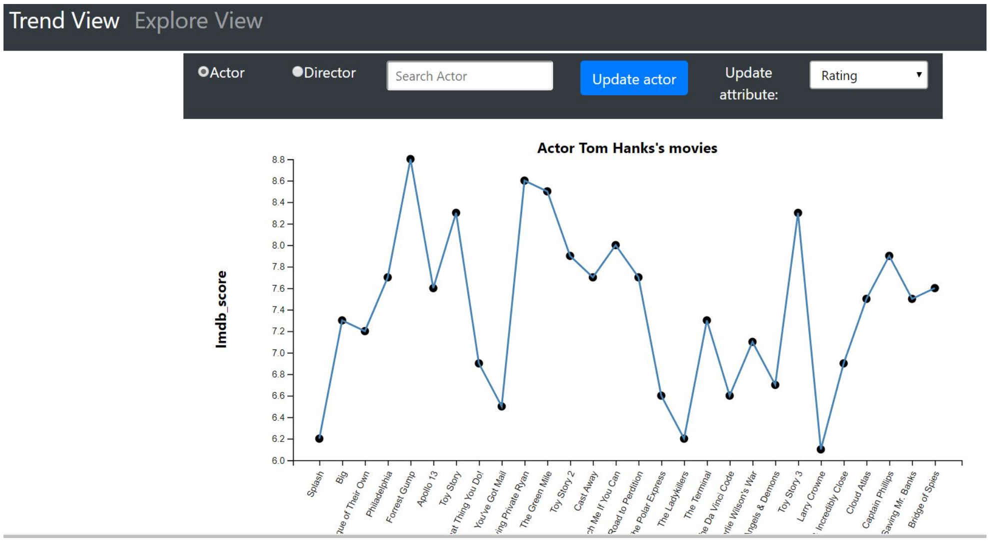


Figure 21: Final Layout: Trend View

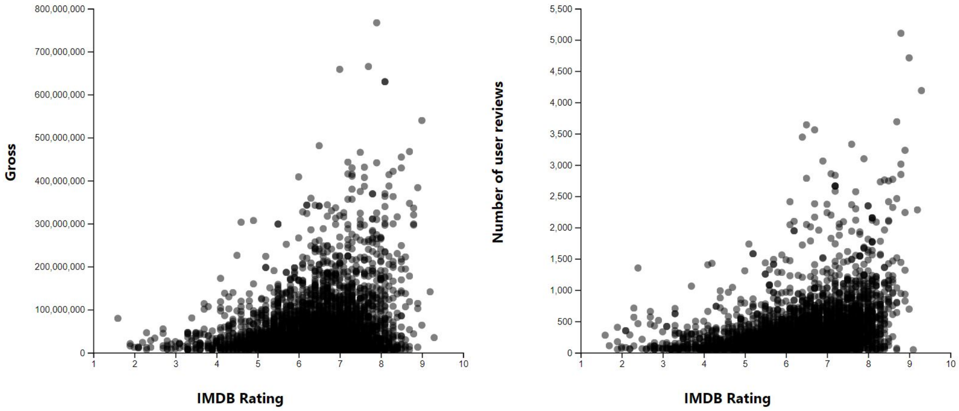
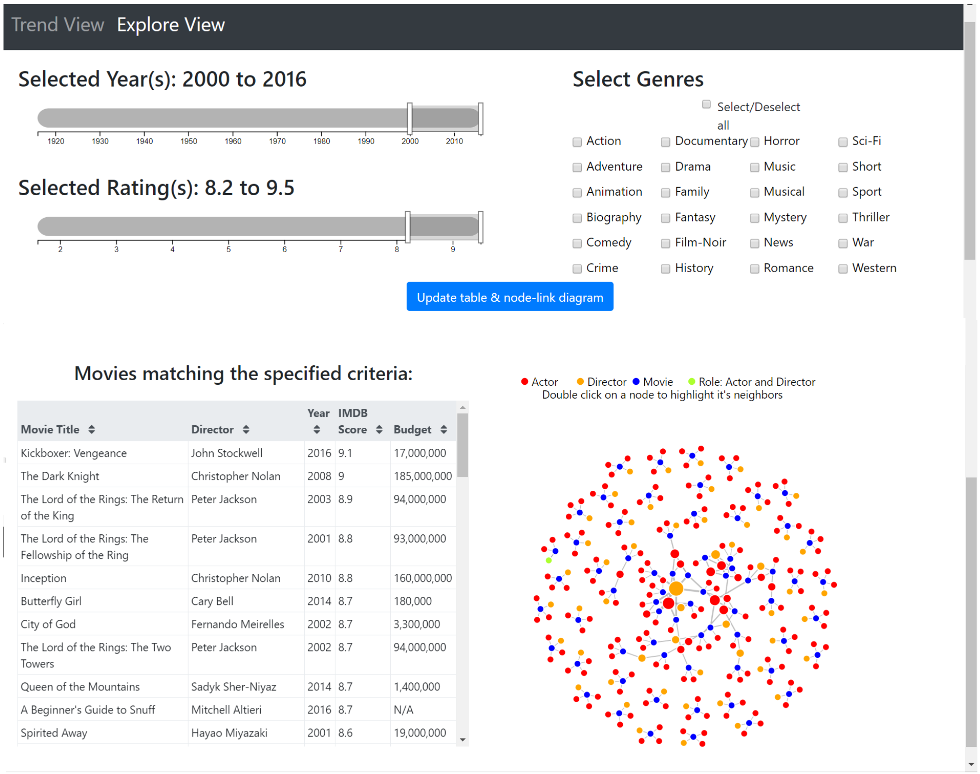


Figure 22: Final Layout: Trend View



Selected Year(s): 2000 to 2016

Selected Rating(s): 8.2 to 9.5

Movies matching the specified criteria:

Movie Title ⬍	Director ⬍	Year ⬍	IMDB Score ⬍	Budget ⬍
Kickboxer: Vengeance	John Stockwell	2016	9.1	17,000,000
The Dark Knight	Christopher Nolan	2008	9	185,000,000
The Lord of the Rings: The Return of the King	Peter Jackson	2003	8.9	94,000,000
The Lord of the Rings: The Fellowship of the Ring	Peter Jackson	2001	8.8	93,000,000
Inception	Christopher Nolan	2010	8.8	160,000,000
Butterfly Girl	Cary Bell	2014	8.7	180,000
City of God	Fernando Meirelles	2002	8.7	3,300,000
The Lord of the Rings: The Two Towers	Peter Jackson	2002	8.7	94,000,000
Queen of the Mountains	Sadyk Sher-Niyaz	2014	8.7	1,400,000
A Beginner's Guide to Snuff	Mitchell Altieri	2016	8.7	N/A
Spirited Away	Hayao Miyazaki	2001	8.6	19,000,000

● Actor ● Director ● Movie ● Role: Actor and Director
Double click on a node to highlight it's neighbors

Figure 23: Final Layout: Explore view

10 Evaluation and Future Scope

During the course of building this visualization, we learned that our visualization could have greatly improved if our dataset had more variety of attributes and also had valid values for all of the existing attributes. However, we tried to utilize the existing attributes in an efficient way for our visualization. As a result, the user would be able to learn interesting things about actors and directors matching his preferences.

We thought of comparing different actors and directors for attributes like gross but we realized that for comparing different actors/directors we need years of the directors to be compared to be released in same years.

So, we have implemented line charts with transitions following the design principles such as simplicity and displaying just the data and we feel that for displaying trend line chart would be the most suitable method.

We didn't start our y-axis values from zero as, we are missing out on interesting trends. So, we represented our data faithfully by taking minimum and maximum values as scale for y-axis and displaying tool-tip to avoid confusion.

We observed few trends from the line charts such as

- Clint Eastwood, as an actor has higher rated movies and grossed movies than when he was a director.

- Woody Allen is highly successful both as an actor and director based on this ratings and gross

We also, implemented force directed graph to show the most prominent actor or director in the filter selection.

We feel that choosing force directed graph is better for visualizing node link relations in a small sized dataset.

We implemented several pop out effects like drag functionality that enables user to inspect the sub section of graph and it connected nodes separately

Also, we have implemented pop out effect to highlight connected nodes.

We have implemented several effects like these to make the visualizations user friendly and attractive.

When we applied the filters to get the top hundred movies from 2000 to 2016 and imdb rating of 8.2 to 9.5, we found some interesting facts such as

- The most prominent director is Christopher Nolan

- The most prominent actor is Christian Bale

- Clint Eastwood is the only person who acted as both director and actor in the top selected movies

We have followed the principle of simplicity while designing tool-tip. We attached tool-tip to line chart and nodes in node link diagrams to show information and make it easy of the user to get information.

As for the future work, we feel that it would be great to link actors, director and movies from the node link to imdb profiles so that we can show movie poster, actor and director profiles

We have implemented word cloud, but we realized that the plot words of movies are not repeating significantly for the word cloud visualization to be defended. So, it would be great if we did some kind of natural language processing and tagged similar words together and visualize a word cloud.

11 References

We mostly used stack over flow answers for our issues

$Reference[1]$: http://dataviscourse.net/tutorials/lectures/lecture-d3-layouts/
$Reference[2] : https : //github.com/sundeepblue/movie_rating_prediction/blob/master/movie_metadata.csv$