

Basic Info.

- **The project title:** Movies Visualization
- **Project repository:** <https://github.com/SaharMehrpour/dataviscourse-pr-MoviesVisualization>
- **Members:**
 - Sahar Mehrpour, u1078019, mehrpour@cs.utah.edu
 - Sunipa Dev, u1077658, sunipadev@gmail.com
 - Siddhartha Ravichandran, u1015163, siddhartha1191@gmail.com
 - Zahra Fahimfar, u0900547, zahra.fahimfar@gmail.com

Background and Motivation.

Watching movies is one of the most fun activities people do in their free time. However, selecting a movie to watch without prior decision is not an easy task, especially with the volume of movies churned out every year. Every individual has specific preferences in terms of genres, themes, actors or even directors. Searching the potential favorite movies in the net takes plenty of time. A good and informative visualization of movies helps users in the manner of an indirect recommender system.

Project Objectives.

- What are the list of movies within particular [genres](#), [actors](#), [directors](#), [years/time periods](#) etc?
- Who are the [director](#), [actors](#) for a specific movie? What are its [genres](#)? What is its [gross](#), [budget](#), [rating](#)?
- What are the movies related by [“keywords”](#)?
- In which movies did this [“actor”](#) and this [“director/actor”](#) work together?
- [Region/language](#) distribution of movies in a given [year](#).

Data.

We obtained the main dataset for movies from [Kaggle](#) website. Here is the link to the dataset: <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset>. A (free) account is needed to download the dataset.

We also have an additional dataset for Academy Awards downloadable from: <https://cs.uwaterloo.ca/~s255khan/oscars.html#download>.

We are also scouring for other possible datasets which are larger than the ones we have now.

Data Processing.

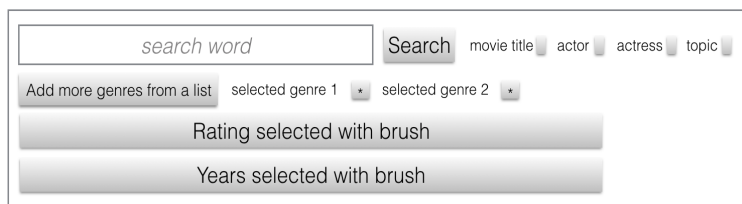
The data from the main dataset is a table with around 5000 movies and 28 columns. About half of the information is not informative (such as the number of likes in Facebook). So, we don't have any plan to use them. Other than that, the dataset is neat and ready to use.

The second dataset has four separate tables (actors, actress, directors, pictures). Each table has 88 rows and 10 columns. We need to check whether the title of movies, genres spelling, and actors' name match with the first dataset. As there is not much data ($88 * 4$) we can manually verify that. This dataset misses the data for two recent years which we can manually add it to the tables ($2 * 4$ rows).

Visualization Design.

Design 1: (The preferred one)

We designed four primary views with a search toolbar. There are a lot of movies and filtering data before visualizing is necessary. To filter the data we designed a toolbar. The primary design uses a textbox for a search word, checkboxes to select the scope of the search, a list of genres, a rating interval, and a year interval.



There are many words as a query, so the trivial way is to enter the query word directly. Instead of four textboxes for four attributes, we designed four checkboxes. The number of genres is limited, so they fit in a drop down menu. Selected genres are displayed on the side with the option of unselecting. The rating and years are quantitative variables, so it seems natural to use an interval to display them. Selecting years and ratings can be done with brushing.

The result of the search can be displayed in a table. The table is sortable based on its columns.

Movie Title	Year	Rating	Budget	Gross	...
Movie title 1
Movie title 2
...

Since not all the information can be visualize in the table, we design a new view which displays the details of the selected movie. We use the Academy Award dataset to visualize whether a movie/director/actor ever won the award. We designed a tooltip which is called on the academy award statue.

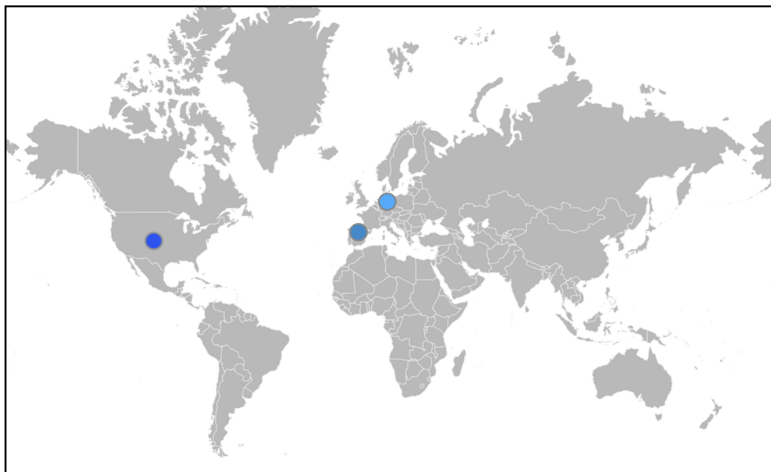
selected movie title 🏆

- director 🏆
- actor
 - actor 1
 - actor 2 🏆
 - actor 3
- budget
- gross
- rating
- ...

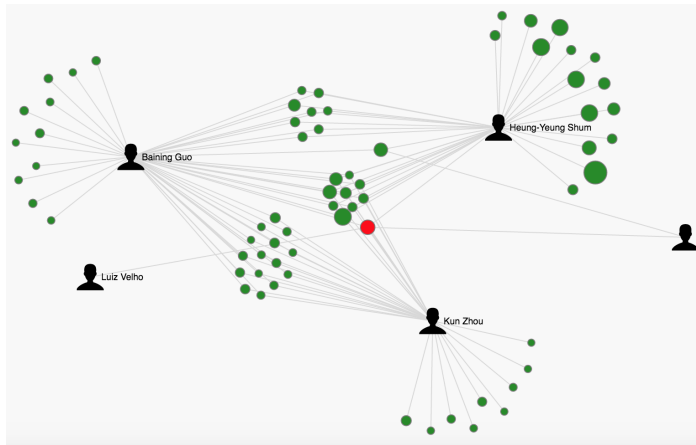
[tooltip appears on
hover over Academy
Award statue]

- year
- movie
- ...

To display the regional information, we draw the world map in another view. The main purpose here is to visualize the distribution of movies in the world. So the mark is points and one of the channels is position. Compared to other countries, annually, the US build a lot of movies. So, using size is not a valid choice here. So we use color scales to visualize the number of movies. Clicking each node on the map will fill out a side table with list of movies and some information.



To display the movies in which the selected actor and director collaborated, we may use either a graph or a table in a separate view. The table is pretty much similar to the one we designed earlier. So we decided to design a different visualization and incorporate a graph. Selecting a movie in one the visualization will create a graph similar to the one in [Siggraph Visualization](#), where nodes are either actors/directors or movies.



Design 2:

A version of Scrollytelling with all movies arranged in a horizontal space arranged chronologically.

Clicking on one movie shows links, and highlights all other movies similar to it based on selected filter (genre, director, actor, etc.). We can also filter the number of movies shown on the scroll. Then, we repeat the Scrollytelling technique for actors and directors too. Links here would show collaborations between actors and directors.

This would be accompanied by a panel or a sidebar with details and the official poster of the movie, more filters and sort options.

We prefer this design less than the first one, since this would make the initial visualization page more cluttered and would take the stress off of recommendations. Also, this would be more complicated to execute.

Design 3:

Select a year or a range of years. The main screen would have bubbles around a single central point of different radii and varied intersections between them. Hovering over the bubbles or their intersections would give tooltips with number of movies in that division and examples. By clicking on a bubble, the browser jumps to a table which could be further filtered and sorted. Clicking on a movie or an actor in this table would then render the graph of collaborations.

We are not sure if this design has the potential to cover all the information we want to show.

Must-Have Features.

- Since there are a lot of movies in the dataset, it does not seem practical to visualize all the data at the same time. So, we must implement a search toolbar by which movies are filtered based on actors, directors, genres, year, etc.
- We must have the option of sorting the filtered movies alphabetically, year, genres, gross/budget, rating, etc.
- We must also visualize a complete information of a selected movie in a separate view.
- Connections or collaborations between actors and directors.
- Visualizing the information about the countries in which movies are made on a map view.

We have added some rough drafts of the visualizations we want in the project.

Optional Features.

- Use [Word Clouds](#) instead of lists of items for nominal variables like genres.
- Extract some information (probably the official poster) from [imdb](#) links for movies (links exist in the dataset).
- Add tooltip to the points on the map such that it displays more information about the movies built in the country corresponding the point.
- Enable zoom and drag on the map.

Project Schedule.

- Week One: Research on different ways to implement our target visualizations, find the best layout for visualization. Survey other visualizations with similar purpose.
- Week Two - Four: Implement the main parts of the visualization
- Week Five and Six: clean and improve the main views, implement extra views.
- Week Six: Finalize the report