

Gender in Movies

A photograph of the Hollywood sign, which is a large white letter sculpture on a hillside. The sign is positioned in the middle of the frame, with the word 'HOLLYWOOD' clearly visible. The hillside is covered in dry, brownish vegetation, and the background is a clear blue sky. The sign is supported by metal poles and is surrounded by some trees and shrubs.

HOLLYWOOD

**Martin Milenkoski
Mladen Korunoski
Blagoj Mitrevski**

Introduction

Motivation

In the past few years, there has been a huge movement towards gender equality in Hollywood. This movement was the main motivation for our project. We wanted to explore the evolution of gender representation in the movie industry in the last few decades. We started our project with a set of broad ideas and questions that we want to answer, and with a dataset of 5000 movies available on Kaggle. As our project evolved, we narrowed down the questions to the ones we considered most essential, and we expanded our data by crawling data directly from the site of The Movie Database (TMDB).

Dataset

Originally, we intended to use the dataset "TMDB 5000 Movie Database" available on Kaggle. However, we realized that this dataset was outdated so we decided to crawl TMDB ourselves. With that, we got all the movies in the database including those that were scheduled for production in the near future. There are more than half a million movies, so crawling TMDB took time. We wanted to keep the TMDB 5000 format, so we could easily integrate the new data in our data wrangling pipeline. Although their API is pretty intuitive we instead used the `tmdbsimple` Python library to access the API. It exposes all the end-points with nice wrapper methods that return JSON objects. However, many of the movies were for adults which skewed our visualizations. After filtering a huge portion of the data, we ended up with 100k movies to work with. From a high level, the data wrangling pipeline made sure that the tables were integrated, filtered, the columns properly cast, and the necessary values substituted. We think that they use NoSQL database to store their data in a more flexible format (JSON like), so many of the values were JSON objects. After extracting the appropriate fields from the objects, we ended up with a very clean and tidy table that we exploited using Python and Pandas. The movie database contains many different features for every movie and person. We use only a subset of the features in our visualizations.

Visualization design

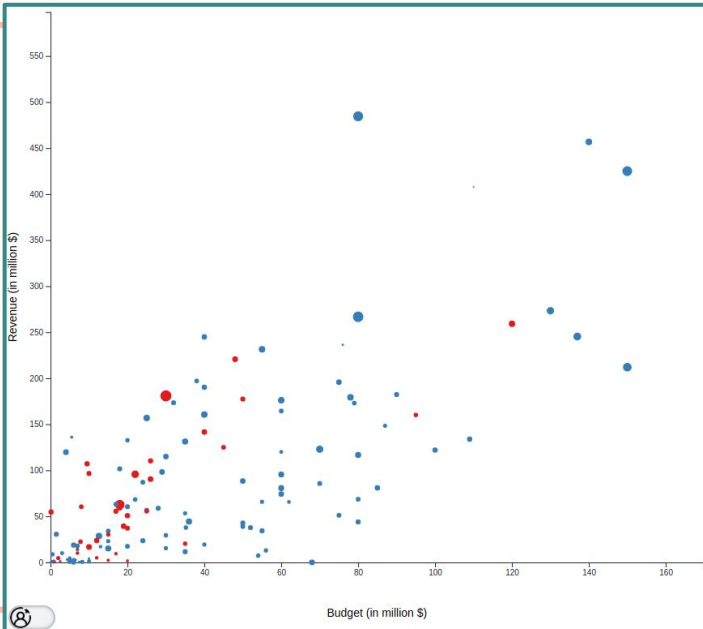
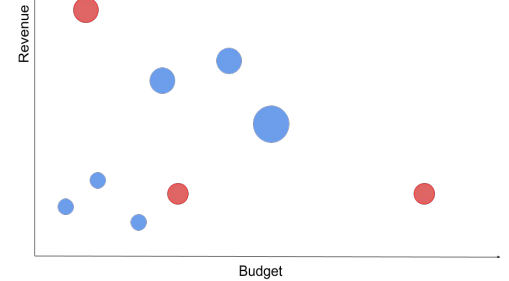
We started off with the idea to create four visualizations that would allow us to explore the relationship between gender and the other dimensions in the dataset:

- **Scatter Plot:** In this visualization, we wanted to show each movie with its budget, revenue, rating and gender of the main actor. In this way, we could see whether gender plays a role in the financing and success of a movie, and how that relationship evolves over time.
- **Stacked Bar Plot:** In this visualization, we wanted to show the distribution of gender in different genres. We had an idea to also show the distribution in different countries, languages and production companies. However, we abandoned this idea because there were many categories in these three features. We did not want to show only a subset of categories because it might have introduced bias in the results. Additionally, the names of the categories were very long and we could not find a good way to show them. For these reasons, we decided to focus only on genres.
- **Network:** In this visualization, we wanted to show the directors and actors in a network where links will represent collaborations between them. During the course of our project, we expanded this visualization and introduced additional features that will be explained in more detail in the corresponding chapter.
- **Word Cloud:** In this visualization, we wanted to show two word clouds generated by the overview of the movies with male and female leading stars. We developed an initial version of the visualization and we realized that there were no interesting insights to be observed. The word clouds seemed random and they were not differentiated. Developing more insightful word clouds would require us to spend less time on the other visualizations. For this reason, we decided to abandon this visualization and focus on developing the other three in more depth.

Once we had implemented the three individual visualizations, we combined them in one dashboard view with a slider at the bottom where users can choose the year of interest. In the following chapters, we will describe the final look of our visualization and explain some of the design choices and challenges we faced along the way.

Scatter Plot

In our first visualization, the initial idea was to create a scatter plot where the x axis showed the movie's budget, the y axis showed the movie's revenue, the size of the circle represented the movie rating and the color represented the gender of the leading actor. The initial sketch is shown on the image on the right.



During our work, we modified some aspects of our initial sketch. We realized that the rating is biased in the cases where there is a small number of reviews. For this purpose, we decided to use the popularity metric for the size of the circles. The popularity is a less biased measure calculated by TMDB. Additionally, we added a button in the bottom left corner, that allows users to choose the role used for coloring the nodes: the leading star or the director. We use blue color for males and red for females. The users can also see the name of the movie by hovering over the corresponding circle.

We realized that many of the movies are clustered near the origin. We did not want to use logarithmic scales because they are less intuitive for many users. For this reason, we implemented a feature for zooming in. The users can select a region using a brush and the selected region will be zoomed in.

On the image to the right you can see a detailed view for the movie Avengers: Endgame. This view can be accessed by clicking on the movie in the view shown above. Here, you can read a short summary of the movie, see the main star and director, as well as watch the movie trailer.



Avengers: Endgame

After the devastating events of Avengers: Infinity War, the universe is in ruins due to the efforts of the Mad Titan, Thanos. With the help of remaining allies, the Avengers must assemble once more in order to undo Thanos' actions and restore order to the universe once and for all, no matter what consequences may be in store.

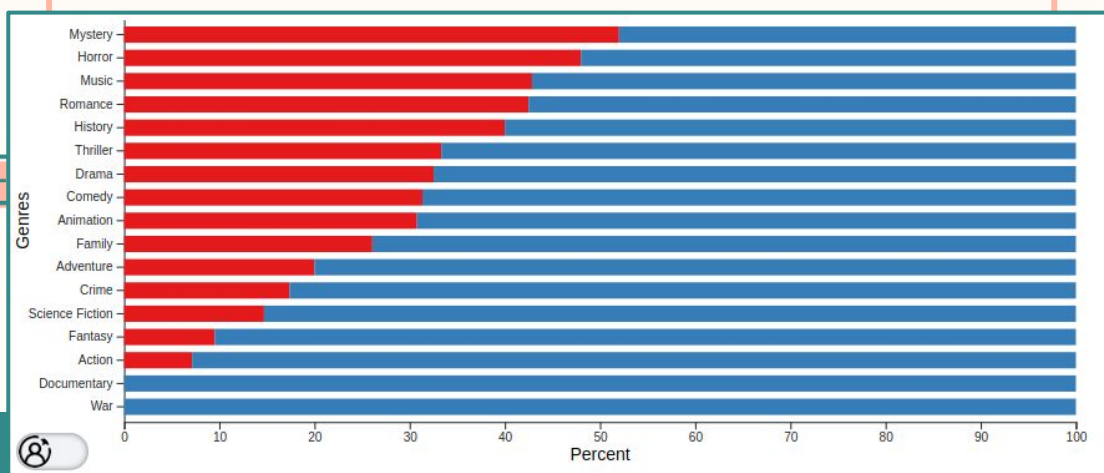
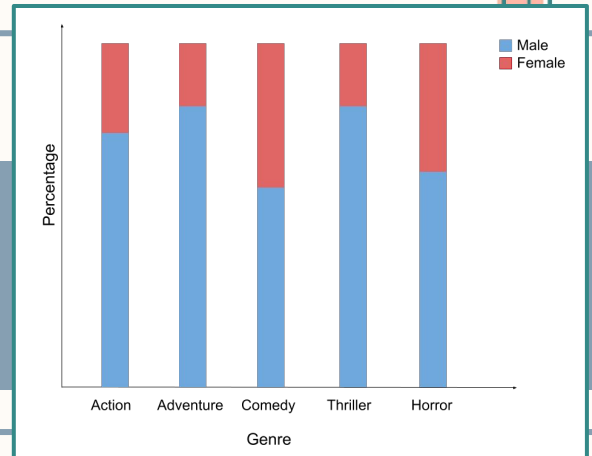
Director: Anthony Russo

Main star: Robert Downey Jr.



Stacked Bar Plot

In our second visualization, the initial idea (shown on the right) was to create a stacked bar chart where the x axis showed the different genres. There were two groups per genre showing the percentage of movies with male and female leading stars in the corresponding genre.



During our work, we decided to show the bars horizontally because it was a better fit for our dashboard. Additionally, we decided to give the users an option to choose the role for coloring: the leading star or the director. This can be done using the button on the bottom left corner. We realized that it might be difficult to read the exact percentage, so we show it when users hover over a bar. Finally, we show the genres sorted in decreasing order of women percentage in the spirit of our interest in women inclusion.

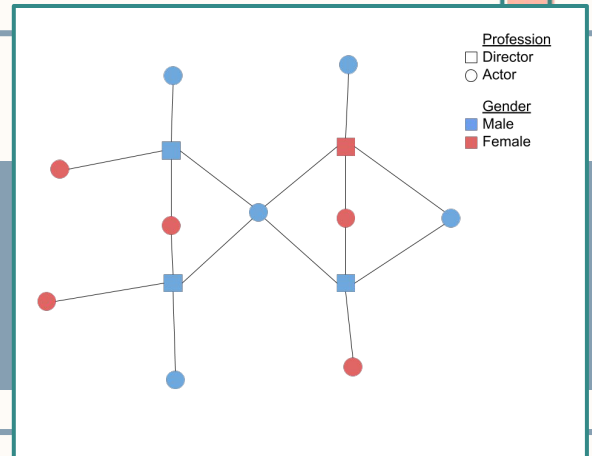
Top action movies with male stars



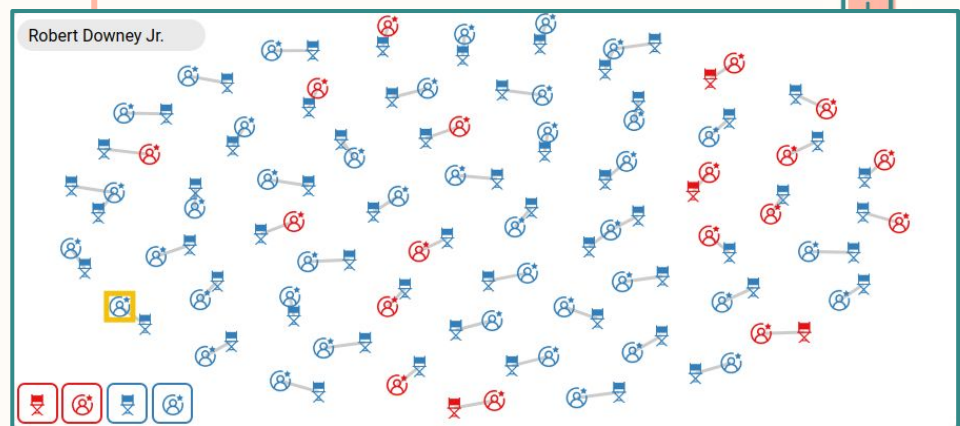
Clicking on one of the bars will open the detailed view shown above. In this view, the users can see the 5 most popular movies in the group corresponding to the selected bar. In this example, we can see the 5 most popular action movies with male leading stars. We decided to show posters instead of names because we believe it is a more visual approach. We could not show more than 5 posters without compromising readability.

Network

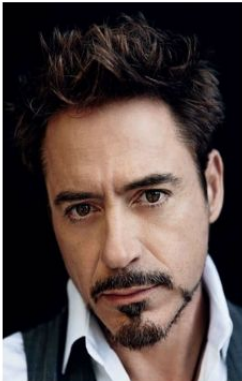
In our third visualization, the initial idea (shown on the right) was to create a network representation of the collaboration between actors (shown in circles) and directors (shown in squares). The gender color scheme is same as before.



In our final visualization, directors are represented with chairs and leading stars are represented with user icons with a little star. We believe that this is a more intuitive representation. Additionally, we decided to show a link between a star and a director only when the actor/actress is a leading star in the movie. The reason for this is because including too many nodes in the network slowed down our whole visualization. Furthermore, we added a functionality to search for a specific person and have it highlighted as shown in the picture on the right. This allows users to track a person across different years. The search field has an autocomplete feature. Users can also show or hide some types of nodes using the buttons at the bottom. This allows the users to focus on the representation of specific roles and genders.



Clicking on one of the nodes in the network will open the detailed view for that person (shown on the right). Here we present an image of the person and some personal information like the name, birthday, place of birth, gender and popularity.



Robert Downey Jr.

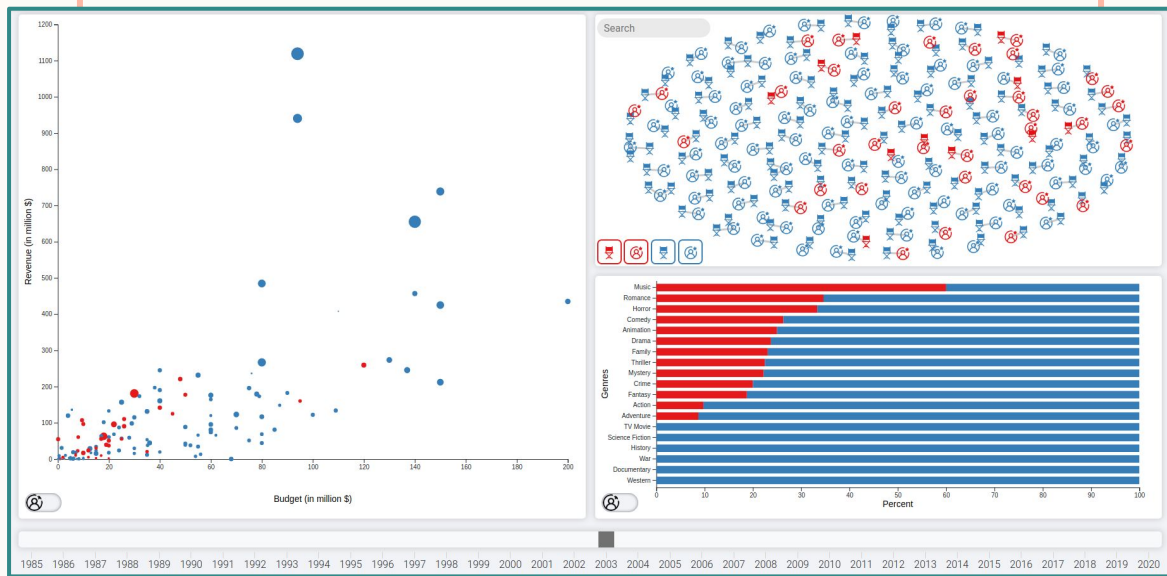
Birthday: 1965-04-04

Place of birth: Manhattan, New York City, New York, USA

Gender: Male

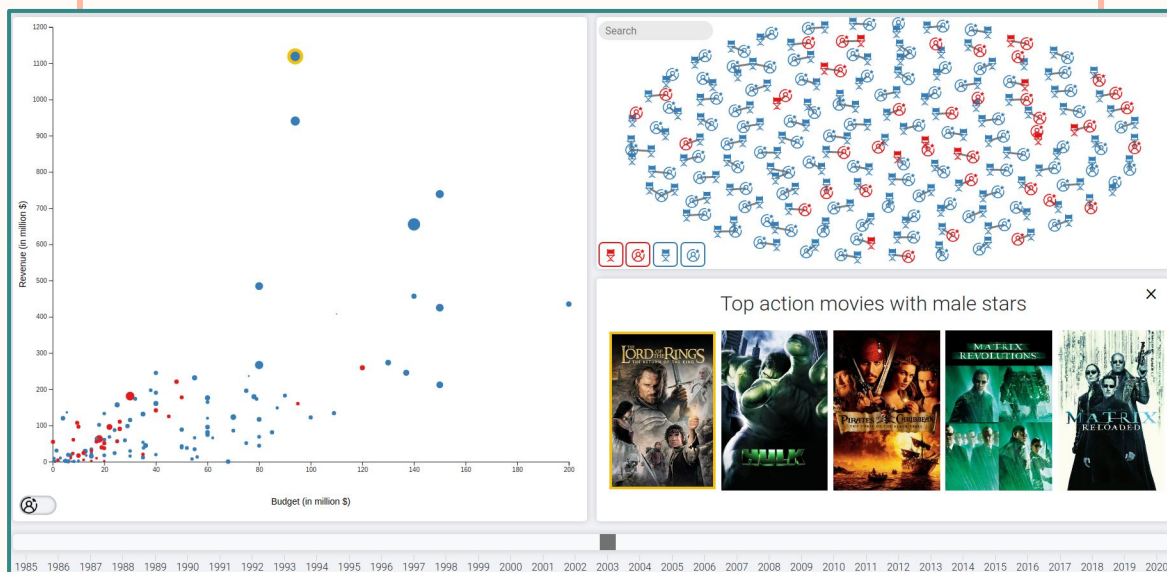
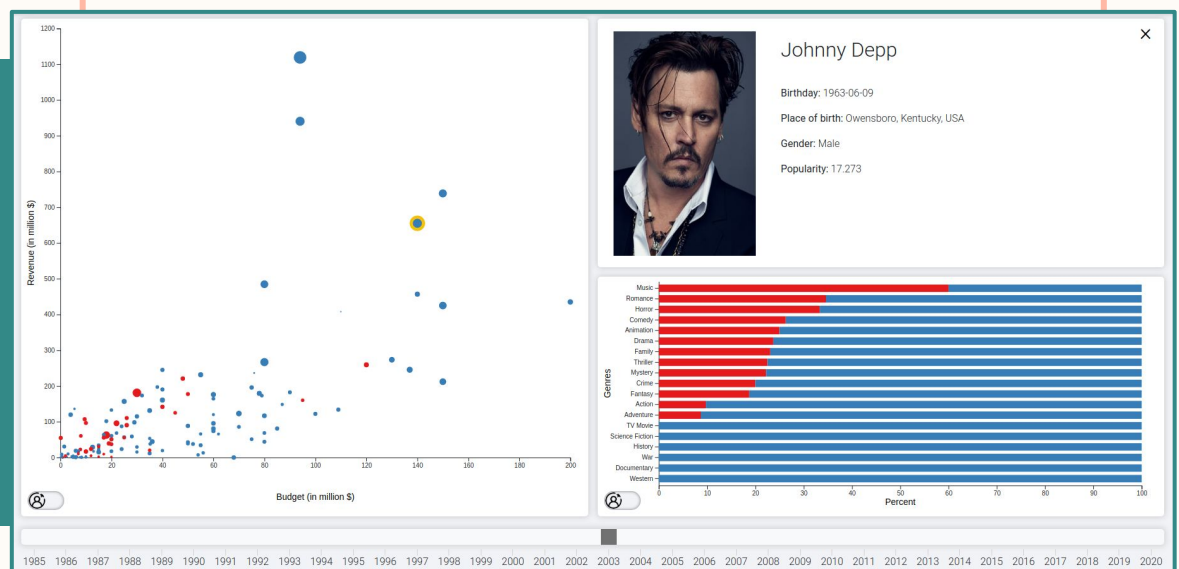
Popularity: 35.822

Dashboard & Gallery



We combine the three visualizations in one dashboard. At the bottom of the dashboard there is a slider where the user can select a specific year. All three visualizations show data only from the selected year. We decided to start in 1985 for cleaner visualization.

Opening the detailed view for a person will highlight the movies in which he/she is a leading star or a director on the scatter plot to the left. This allows users to track how well a person's movies do across different years.



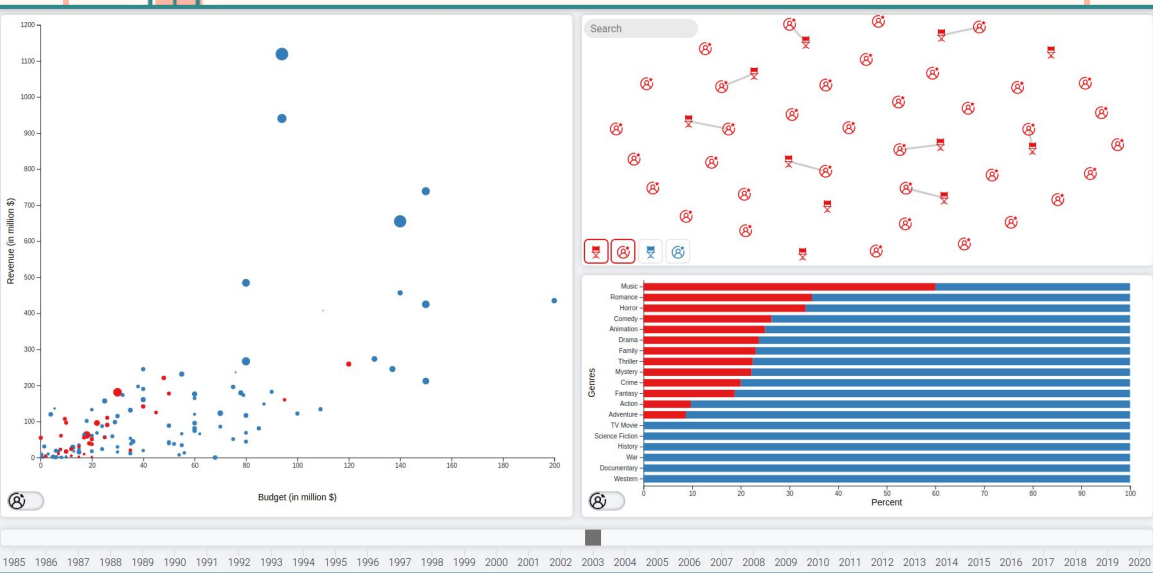
Clicking on a movie poster in the detailed view of the bar visualization will highlight the poster in that view and the movie in the scatter plot to the left. Double clicking on the poster will open the detailed view for the movie on the scatter plot.

Dashboard & Gallery

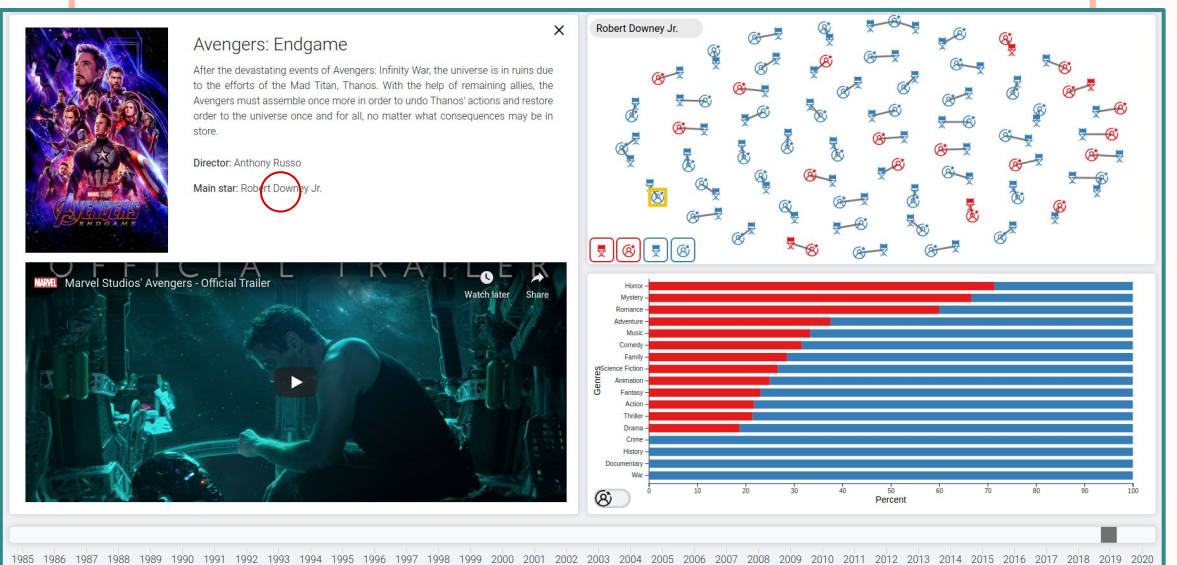
On this image we show the autocomplete feature of our search bar. We found an implementation that matched only people whose name starts with the typed input, and we modified it so it matches any name that contains the input. In this way, users can search by surname or partial name.



Here we can see the effect of turning off the male directors and actors in the network visualization. In this way, users can more easily track the evolution of women inclusion in the movie industry through the years..



Clicking on the name of the main star or director in the detailed view of the scatter plot on the left will autofill the search bar in the network on the right and highlight the person in the network. This allows users to save time by not having to type the name themselves.



Conclusion

Additional challenges

During our project, we faced numerous challenges and design decisions. We described some of them in the previous sections. In this section, we will try to describe some of the main challenges that were not mentioned previously.

- **Animations** - every visualization has its own update animation that is played whenever the data is updated, either through choosing a year in the slide bar or clicking a button for updating the specific visualization. Additionally, there are animations that play when users hover over a bar, a dot in the scatter plot or a node in the graph. We faced difficulties making the animations not interrupt each other. For example, one issue we faced was when users hover over a bar while the update animation of the bar chart is not yet finished. This resulted in the first animation stopping and never continuing. We fixed this issue by giving different names to different animations. Additionally, letting the network animation run endlessly slowed down our whole page. For this reason, we only run the animation for 5 seconds which is enough time for the nodes to separate well from each other.
- **Media** - We did not want to load all the data for movie posters, trailers and personal details from the start when most of it would not be used at all. For this reason, we use the TMDB API to load the necessary data for the detailed view of every visualization whenever users want to see it. In this way, we sped up our visualization.
- **Filtering buttons** - Every visualization has its own filtering buttons. We decided to add these buttons later on in the project, so we have not designed our code with them in mind. Therefore, adding them later on was not a straightforward process. Adding the functionality for showing and hiding specific nodes in the network visualization proved to be especially difficult and required significant refactoring of our initial code. However, when we went through the process once, adding additional functionalities like the search bar was much easier.
- **Stacked bar updates** - Updating the stacked bar chart was especially difficult. Due to the fact that not every genre is present in every year, and the way we implement the stacked bars as two separate groups of bars for the two genders, we could not find a way to update the existing bars when we want to update the visualization. For this reason, in our final visualization we erase the existing bars and redraw them whenever we need to update the chart.

Final words

In this process book, we focused on describing the features of our visualization, the challenges we faced and the design decisions we made. We believe that we created a tool that will allow users to explore the gender representation on their own and make their own conclusions. However, in this final section, we want to point out that through working on this visualization, we noticed that gender representation in the movie industry did get better in the last few decades, but the situation is far from perfect. The industry still seems to be dominated by male actors and directors and there is certainly a tendency to cast women as leading stars in genres like Music, Romance and Drama, as opposed to the male dominated genres like Action and Adventure.

Peer Assessment

Martin Milenkoski

Implementation of the network visualization

Writing the process book and recording the screencast

Mladen Korunoski

Implementation of the scatter plot visualization

Data crawling and preprocessing

Blagoj Mitrevski

Implementation of the bar plot visualization

Implementation of the dashboard