



IT'S
HOLLYWOOD
NIGHT!

MOVIE INVESTMENT



GROUP NAME: BCD

GROUP MEMBER: LILI WANG
WEIYU CHEN
JINGXUN CHEN

DATE: JUNE 2, 2022

OUR PATH

Our path for this project began with brainstorming and deciding what we wanted to show in our visualization.

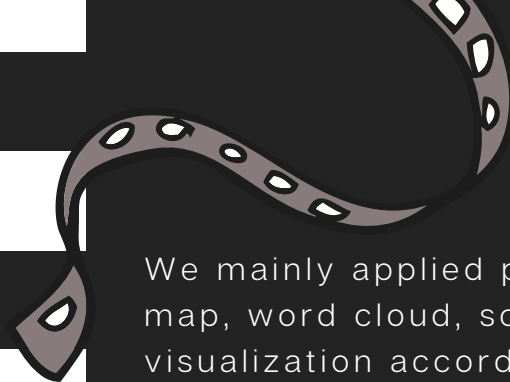
Movies are now an integral part of everyone's lives. Whoever you are, movies can always be the guide leading you to a fancy world, where you can completely release the pressure on your shoulder, and immerse into all kinds of attractive stories. Therefore, we chose the movies as our main topic and ended up with two possible datasets: one about the scripts, characters, potions, and spells in the first three Harry Potter films, and another is an IMDb Movies dataset containing over 5000 movies from 2000 - 2020.

The first criterion we choose the dataset, was having enough data to tell an interesting story rather than just simple statistics, while another important criterion was selecting data from which we could learn something new. Both datasets are great for different reasons, but we chose the IMDb Movies dataset at last.

However, there are too many visualizations based on this kind of movies dataset. Therefore, we tried a new angle to analyze the dataset. Instead of visualization focused on the relationships between votes and movies from common audience's view, our primary concern is whether the movie is worthy of investment.

We have designed our website as a guide for users, the majority of whom are investors, to schedule their own invested movies in order to maximize the profit. If the user would like to utilize current funds and are wondering what to invest, film financing is a brilliant idea because of its importance in people's lives and sustainability along with the development of modern technology. As a result, we choose six aspects in the dataset to focus on, including genre, actor, filming country, movie contents, director, and publish quarter, covering necessary steps during the procedure of planning a filming project. Once we had all the data we need, we started to design and sketch our different visualizations. We focused on relations between the profitability and these six aspects of movies.

We started our implementation by construct the prototype based on our previously decided visualization cores and chart styles. After reading massive number of codes and practicing implementing simple charts and interactions, we built our website with static plots and connected all the pages using simplest linking method.




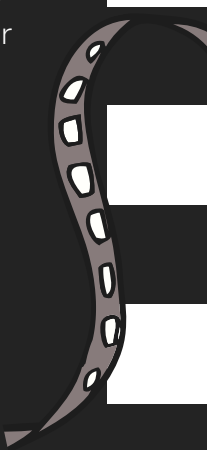
We mainly applied parallel coordinate chart, (grouped) bar charts, map, word cloud, scatter plot as the basic structure of our visualization according to the characteristics of each type of data. We also implemented transitions and animation using d3 and css, including zooming, brushing, and reacting to mouse events, in order to emphasize the point and make the charts clearer and more interactive. We appended a tooltip for each chart as supplement contents besides the original charts.

CHANGES

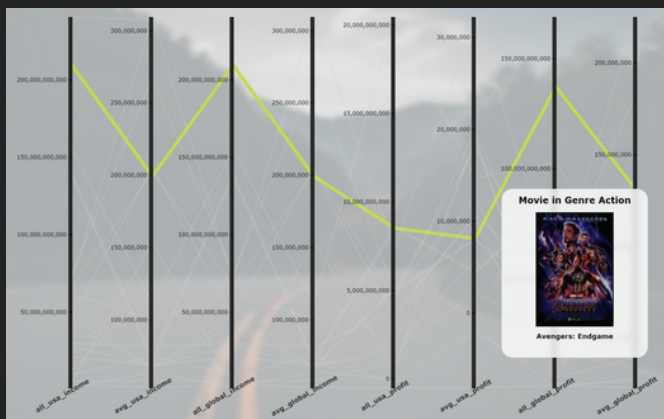
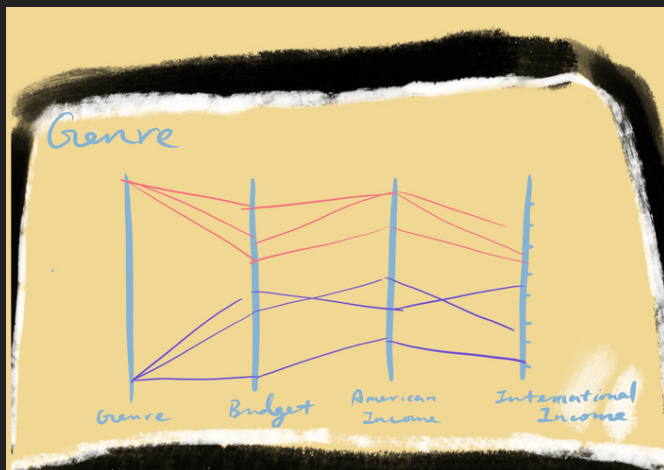
In milestone 1, we still focus on the votes for the movies. However, we found that factors that determine movie ratings are not a new topic. On the other hand, we hardly found some research on investment based on this dataset. Moreover we have obtained budgets and income of the movies which allowed us to analyze the data from this perspective. Therefore, our main focus changed to the relationship between the profits and movies.

We changed the general architecture of the website compared to our milestone 2.

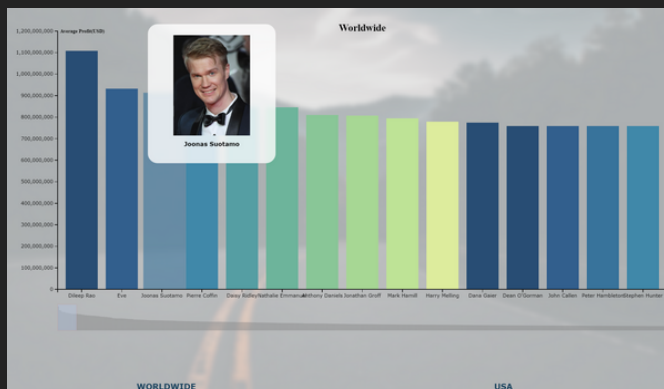
In our milestone 2, we designed our home page as a direct road to Hollywood, and the user is sitting in the car. The six signs along the road represent the movie's six different aspects. The user needs to click on each guideboard to get into an entrance page and then click one more time to get into the information page. The operation is too lengthy and useless, so we simplify it in a clearer and better-looking way. The user can use the navigation bar on the top to go to pages of their interest with side buttons connecting previous and next pages for smooth transition between pages.



GENRE



The original parallel coordinator only has four columns, "genre", "budget", "USA income" and "international income". In the final version, we decided to remove the column "genre" and show each line's genre when hovering on it. This way makes the chart more concise, and users can focus more on the numerical data. Then we added average and all income and profits both worldwide and in the USA to provide users more chance to compare each genre's possible profit. In addition, we add a floating window to show one film poster for each genre, this is the film with the highest profit in this genre.

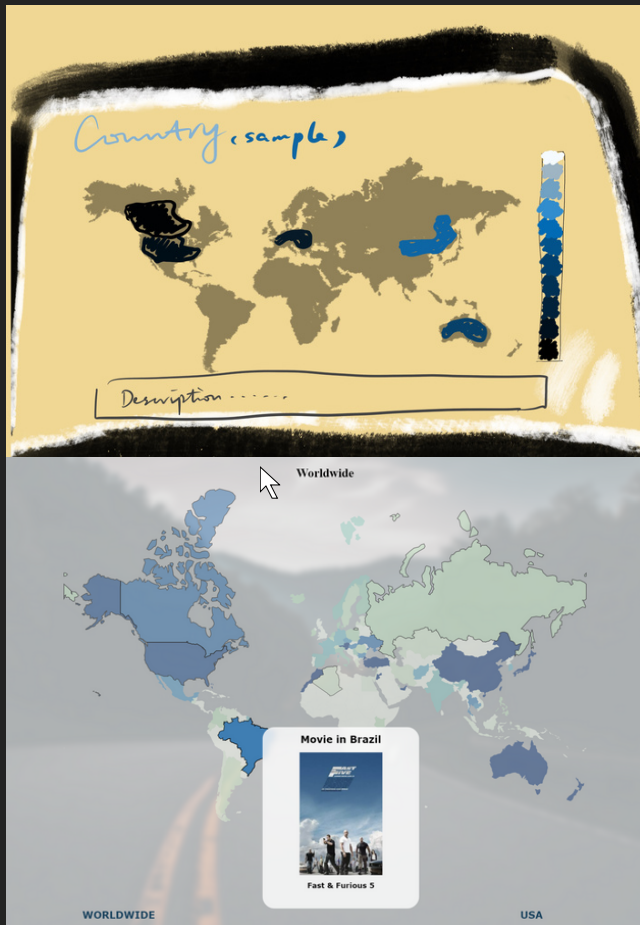


The original chart has bars to show each actor's average film profit. The number of actors in our data set is over 20 thousand. To better view the actors with higher profits, we decided only to extract the top 500 actors, sort them by profits, and make the chart draggable. We also add gradient colors to show the differences between the actors.

We also include a window to show each actor's photo in this chart.

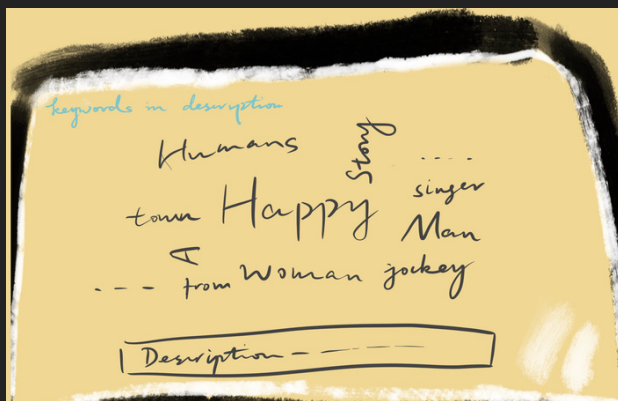
ACTOR

COUNTRY



The original map shows each country's average film profit. We didn't change much in the final version. Since we have both global and the USA average profits, we make two separate charts for each range of areas. The color shade represents the profit level. The darker the color, the higher the profit.

In the floating window, we show one film poster for each country, this is the film with the highest profit in this country.

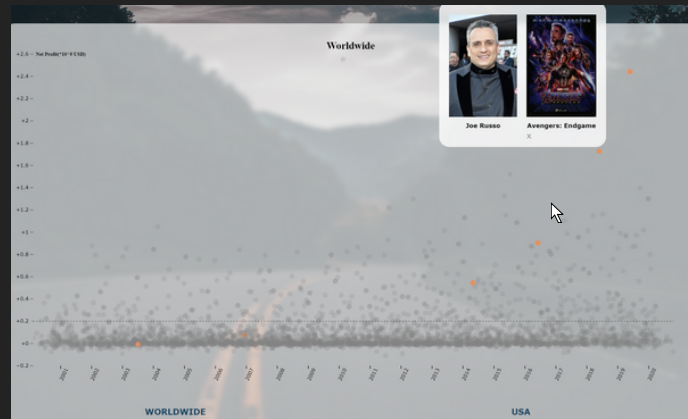
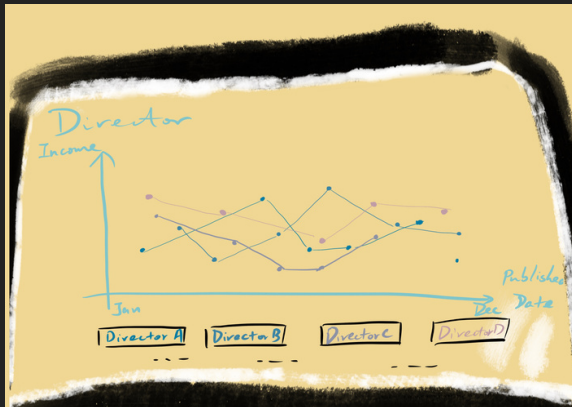


We have implemented the word cloud just as the one in milestone2's sketch, along with tooltip and mouse event to interact with the user. We have used color scheme from grey to dark blue indicating its profits, together with light yellow as event highlight color.

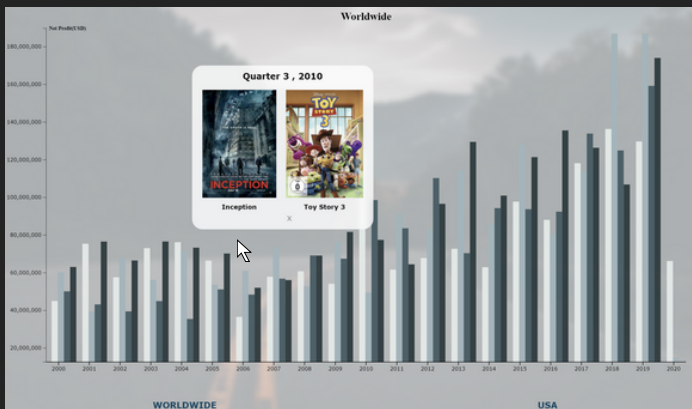
We have provided the highest-voted movie with certain keyword, to help the user do brainstorming on given words.

CONTENT

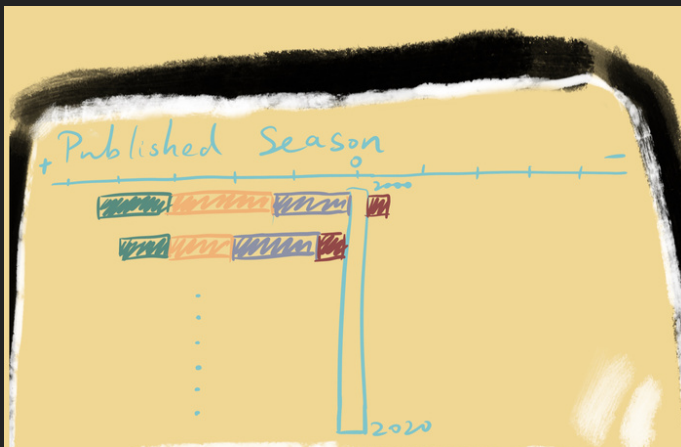
DIRECTOR



We have implemented the scatter plot without the path connecting films belonging to the same director. Since we had 3900+ movies and over a thousand directors, simply plot all the circles will cause severe lagging. Thus, we gave up drawing the lines, and instead, using mouse event to select and highlight corresponding points with tooltip providing extra information.



We have changed original stacked bar to grouped bar, basically because of we found that stacked bar chart was not clear enough for comparing profits in each quarter in the same year as well as profits in the same quarter across a range of years.



In addition, we have added extra information using tooltip to help user grasp the meaning of the charts by providing movie instances. This is also a way to make the visualization more close to the user with real life instances, instead of pure bars and numbers.

PUBLISH QUARTER

CHALLENGES

Data processing

- *Currency unit unifying*

The budget, global income, and the USA income in our data set have currency other than US dollars, such as DEM and GBP. What's more, the cells with other currency's data type are in string type instead of numbers. As a result, we filtered out non-USD values, and then applied CurrencyConverter, a python currency change library, CurrencyConverter.

- *No images resources*

We would like to present relevant images such as posters and director photos to make the website visually richer. At first we found a corresponding IMDB API for obtaining images and introductions. However, soon we figured it out that there was a strict limitation on daily call, exactly 100 calls per day. But we have got over 10,000 objects waiting for pairing images, thus we were forced to give it up.

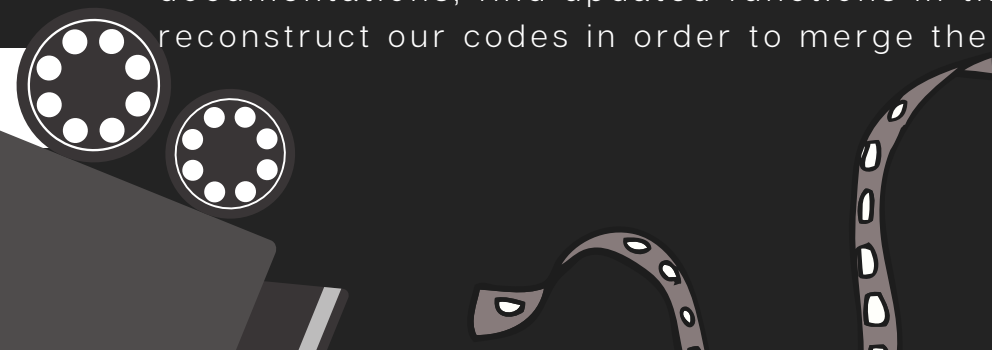
The second idea was to utilize wikipedia API for searching relevant terms. It was not too long before we got into a dilemma again, caused by movies titles in mixed-up languages, mismatched terms with the same names and terms not included in wikipedia.

Our final solution to this problem was to crawl images from Bing image using selenium and requests, which cost about 8 hours to retrieve 7800+ image links.

Technical implementation

- *D3 version conflicts*

Since our codes were collaborated by two of us, we had met the conflict of source codes versions, caused in the incompatibility in some of the functions, and consequently crash down of the whole websites. The only solution to this problem is to read relevant documentations, find updated functions in the new version and reconstruct our codes in order to merge them together.



CHALLENGES

- **Bar chart**

The number of unique actors in our dataset is 27352, which is too large to be shown by a bar chart. So we decided to sort actors by their average profit for each film and then select the 500 actors with the highest average profit. 500 bars still seem crowded for one chart. To solve this problem, we decided to add a draggable box with a thumbnail under the main chart, and the main chart only shows a reasonable number of bars. By dragging the little box, users can easily see the actors' average profit.

-

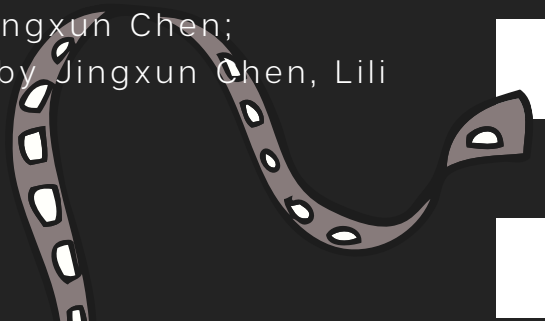
- **Element overlapping**

It is hard to check the order of each element when there are too many nested divs, svg and shapes. When the front element completely covers another selection which is used to receive cursor events, it becomes unreasonably difficult to accurately target the problematic element. It can only be solved by getting our ideas into shape step by step, analyzing the whole structure of the website using DevTools in the chrome to check the elements and functions running results.

PEER ASSESSMENT

We break down the project as following:

- Dataset and basic structure including core data visualization of the website are decided after the group discussion;
- Data processing were done by Lili and Weiyu, while reports are written by Jingxun;
- Data visualization core: actor, genre, country were implemented by Lili; word cloud, director, publish quarter were done by Weiyu.
- Screencast are done by Jingxun Chen;
- Process book are written by Jingxun Chen, Lili Wang and Weiyu Chen.



REFERENCES

- [1]Unsplash.com. 2016. Road in Great Smoky Mountains by Jake Blucker on Unsplash. [online] Available at: <<https://unsplash.com/photos/OJX7gIU3E6U>> [Accessed 3 June 2022].
- [2]Bostock, M., 2020. Gallery. [online] Observable. Available at: <<https://observablehq.com/@d3/gallery>> [Accessed 3 June 2022].
- [3]codepen. n.d. Arrows. [online] Available at: <<https://codepen.io/cbrst/pen/ebxwLJ>> [Accessed 3 June 2022].
- [4]W3schools.com. CSS Tutorial. [online] Available at: <<https://www.w3schools.com/css/>> [Accessed 3 June 2022].
- [5]2022. [Online]. Available: <https://soshace.com/mapping-the-world-creating-beautiful-maps-and-populating-them-with-data-using-d3-js/>. [Accessed: 03- Jun- 2022].
- [6]"Scroll Bar Chart", Blocks.org, 2022. [Online]. Available: <http://blocks.org/cdagli/728e1f4509671b7de16d5f7f6bfee6f0>. [Accessed: 03- Jun-2022].
- [7]R. Engineering, "D3.js Bar Chart Tutorial: Build Interactive JavaScript Charts and Graphs - RisingStack Engineering", RisingStack Engineering, 2022. [Online]. Available: <https://blog.risingstack.com/d3-js-tutorial-bar-charts-with-javascript/>. [Accessed: 03- Jun- 2022].

