Visualizing Diversity in Hollywood

Jenning Chen[*] Ashwin Srinivasan[†] Lisa Yoo[‡] Cindy Yang[§]

Abstract

Inspired by the #OscarsSoWhite movement in 2015, our visualization experience empowers users to explore diversity in Hollywood, and how the distribution of race and ethnicity relate to different measures of success: movie reception and awards. We bring together datasets on major awards (Oscars, Golden Globes, Screen Actors' Guild) and box office statistics and create digestible visualizations to see how white Hollywood really is.

1 Introduction

The Oscars, Screen Actors Guild Awards, and Golden Globes are some of the most prestigious awards that recognize talent in film and television. However, discussions surrounding the issue of diversity in the film and entertainment industry have been put under spotlight in recent years. While minorities accounted for 40.2% of the US population in 2019, they only accounted for 32.7% of actors, 27.6% of lead actors, and 15.1% of film directors [3]. This suggests that there is a greater issue of racial representation and disparity within the film industry. Our visualization seeks to explore this issue by providing the user with easily digestible information about distributions of race and ethnicity with respect to awards and movie reception.

2 Related Work

Several key events have brought attention to the racial disparities entrenched in the most prominent film award shows. Most notable was the hashtag #OscarsSoWhite, a social justice campaign in response to the fact that all 20 acting nominations for the 2015 Oscars had been given to white actors. The campaign brought attention to long-existing inequities in the film industry that lacked representation for people of color, and the culture that perpetuated this issue. A number of reports and analyses have been subsequently produced in order to explore these issues. Among the most comprehensive is the UCLA Hollywood Diversity Report [3], which conducts an annual study to (i) generate comprehensive research analyses on inclusion in film and television across different roles, (ii) identify best practices for increasing the pipeline of underrepresented groups into the Hollywood industry, and (iii) consider the broader implications of diverse industry practices for society. They study variables including diversity in lead roles, overall cast, writer, directors, and genres. Metrics like global and domestic box office, international market distribution, and ticket buyer demographics are considered for each film. Our work uses similar metrics.

A few visualizations have been produced to communicate data on racial diversity in Hollywood to broader audiences. An impressive one is the 2019 graphic by the Kontinentalist in the article ”Asian representation in movies: have things changed since 1997?” [1]. The article uses a scrolly-telling format to explore Asian representation in blockbuster movies. An interactive graph encodes each film studied as a rectangle as part of a larger bar graph categorizing films by year. They make use of hover tooltips, allowing users to view the racial breakdown of the cast. Useful side texts guides the narrative on the low Asian representation and homogeneity present, i.e. most Asian representation comes from a few key actors who tend to play the same type of role over and over again.

A key design choice in our visualization is the use of a circle graph to encode film award nominees and winners, and our choice of a circle graph is inspired by NYC Foodiverse [4], a stunning visualization which uses a circle graph to visualize food quality at restaurants in NYC. Their choice of encodings allows for easy exploration of individual restaurants' sanitation violations, FourSquare reviews, ratings, and price tiers.

3 Methods

D3.js was used to create all of our interactive visualizations.

3.1 Awards

We decided to create visualizations on the Oscars, Screen Actors Guild Awards, and Golden Globes. We web scraped and gathered data for the three awards over the years into separate CSV files. Each row in the CSVs contained:

- Year of the ceremony/award
- Category
- Name of nominee/winner
- Film
- Whether the individual was a nominee or a winner
- Individual's race

For all three awards, we looked at the following categories: Best Actor, Best Actress, Best Supporting Actor, Best Supporting Actress. Additionally, for the Oscars, we gathered data on the Best Director category. We used D3's circular packing functionality to create our cluster graph visualizations [2]. Since our data included award category and whether the individual was a nominee or winner, we could subsection our visualizations accordingly. Below the circle graph, which displays information about individual movies, we provide summary visualizations. First, we calculate the percentage of minorities in each award category, for the award show chosen, and display the statistic across time. We use ”small multiples”, a group of several line graphs, to show the percentage of minotirites per category. A tooltip allows the user to track the values by year across all of the small graphs, simultaneously. This encoding was chosen over a single graph containing multiple lines for a less cluttered and confusing viewing experience. Below this, we have a line graph comparing the percentage of minorities across the three award shows. The data included changes based on the award categories selected at the top. For both line graphs, we calculated 5-year rolling averages to represent the percentage of minorities in each year, since the percentage fluctuated highly from year to year and it would have been difficult for a user to interpret the overall trends.

[*] e-mail: jenning@mit.edu
[†] e-mail: ashwins@mit.edu
[‡] e-mail: ed.grimley@aol.com
[§] e-mail: cxy99@mit.edu

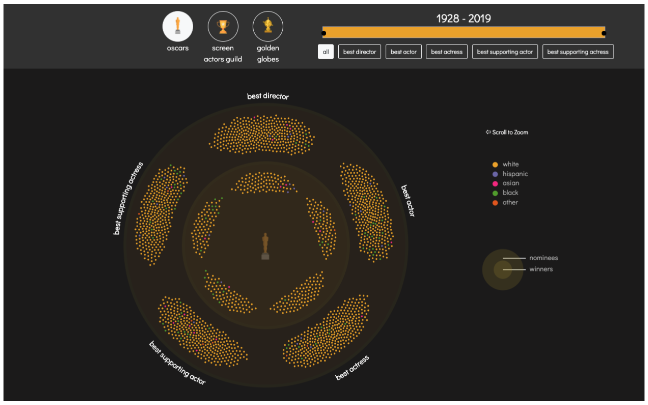


Figure 1: Award circle packing visualization. Inner circle contains winners, outer circle contains nominees. Each circle represents an individual color-encoded by race. Filters on top bar.

3.2 Reception

Taking another angle on examining diversity in Hollywood, we decided to gather data on top movies in the box office. We began with a holistic dataset of movies from IMDB, which contained metadata including gross revenue, budget, year of the film, top-billed actors, IMDB metascore, genre, and more for each movie. To make the resulting dataset a more manageable size with more meaningful datapoints, we filtered for the top 1000 movies by adjusted gross lifetime revenue. Then, using the field for the top-billed actors, we scraped each actor's race from nndb.com using the pandas library in Python in order to calculate the race distribution of the cast.

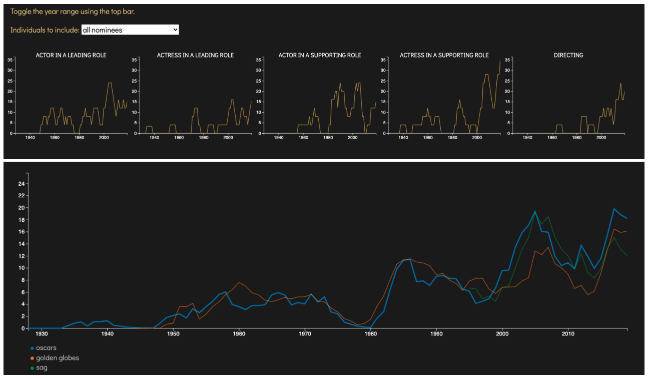
For this dataset, we wanted to create visualizations to show how the variable of cast diversity relates to different metrics: time, revenue, genre. Similar to our awards visualization, we still wanted to engage the user by providing a fine granularity of data, where each data point could represent a movie. With this in mind, we thought to create a bubble clustering visualization, and using position/size encoding to provide more information on time and revenue respectively. Additional movie metadata would appear in a tooltip. This included genre, budget, and the movie poster, which was retrieved through the TMDB API.

To provide a more personalized experience for the user, we wanted to allow them to search for their favorite movies. We searched through our movie database and used simple string matching to provide an autocomplete functionality. We integrated this into visualization where a selected movies' metrics are compared to the industry average for that year. Rather than putting down the raw numbers associated with each of these numbers, to make the comparison easier, we chose to use a number of icons to correspond to each metric, based on a linear scale bounded by maximum possible values.

4 Results

4.1 Awards

For the awards, we have separate pages for the Oscars, Screen Actors Guild, and Golden Globes. The main visualization (Figure 1) on each page consists of a cluster graph, where each individual is encoded as a dot, and the dot's color corresponds to their race. The outer ring around the center contains the nominees, and the inner clusters contain the winners. The dot clusters are also partitioned by award category. If the user hovers over a dot, a tooltip is displayed with more detailed information including the individual, year, and film. The user can filter the data in the visualization using the year slider and category buttons in the top bar. On the right, we have legends that display the race-dot color encodings and the winner/nominee partition. If the user hovers over a particular race in the legend, then only the dots corresponding to the race will be highlighted in the visualization. Similarly, if the user hovers over the nominee or winner section in the legend, then only the nominee or winner dots will be highlighted, respectively. Right below the visualization, we provide a summary statistic stating the percentages of White nominees and winners depending on the current data in the visualization (based on the years and categories the user selected).


Toggle the year range using the top bar.

Individuals to include: all nominees

Figure 2: Line graph visualizations breaking down dataset by category and award.

Below the main visualization, we have two line graphs to provide more information about diversity across the different award categories (Figure 2). The first line visualization contains separate graphs of the percentage of minorities in each category of the award (depending on which awards page the user is on) over the years. There is a dropdown where the user can specify which individuals they would like included in the graphs (all nominees, winners, nominees excluding winners). The x-axis of these graphs is synchronized with the year slider. There is also a tooltip when the user hovers over any of the graphs, where they can see the percentage of minorities in each category for that year. The second line visualization is a single line graph that compares the percentage of minorities in the current award with the other two awards. The line corresponding to the current award is thicker than the other two awards' lines so that the user can easily identify which line is the current award. There is also a tooltip that shows the percentage value on the line of the current award. The color encoding of the awards is the same on all three award pages to prevent possible confusion with the user. This graph is also synchronized with the year slider and category buttons on the top bar. For the category buttons the user selects, the proportions are averaged among those categories for each award. The y-axis scale of the graph dynamically changes as the user changes the categories/years so that they can best observe the differences between the awards at any given setting.

4.2 Reception

On the reception page, we have two main visualizations. The first one is a bubble clustering plot (Figure 3). Each bubble on the visualization corresponds to a movie, and its position is plotted on a decade row, based on the year it came out, and on the x-axis based on the cast's percentage of white actors. The size of the bubble encodes the adjusted lifetime revenue. When the user hovers over a bubble, there's a tooltip on the sidebar which shows a movie poster, along with some metrics on the selected movie (genre, year, revenue, budget, metascore). There is a also a breakdown of the top-billed actors by race as a small bar chart in the sidebar. To provide an extra level of interactivity, users can change the race on the x-axis and also filter the movies by genre.

The second visualization is an comparative icon-based graph (Figure 4). The user searches for a movie via the dropdown autocomplete search bar. When a movie is selected, this populates two side-by-side panels: on the left is the selected movie, and on the right is the industry average among movies that year. Each panel has identical metrics (metascore, revenue, budget, percent minority cast), meant to be compared side-by-side. Each metric highlights a number of icons out of ten, based on a linear scale bounded by the maximum possible value. The more precise values for these metrics are also displayed.

5 Discussion

Ultimately, the goal of our visualization is to allow the user to explore and evaluate for themselves, how race relates to success in Hollywood in two contexts: awards and box office reception. This consolidated data is not readily available to the general public, and through our visualizations, we hope to make these insights on diversity in Hollywood more accessible.

From the awards visualizations, there are a few high-level conclusions that the audience can draw:

- Sheer amount of white actors/directors across all awards and categories: this is evident in the high number of gold-colored circles in the circle-packing visualization. The summary statistics on percent white nominees and winners also demonstrate this.

- Gradual increase of minority nominees and winners: by choosing different intervals of years, the users can see how the distribution of race evolves. This insight is even clearer in the line graphs by category and award below the main visualization.

The users gain these insights at a high-level through the category and award line graphs on the page, and leverage the interactive filters to explore more hidden trends in combinations of specific categories/years/awards.

From the reception visualizations, we're hoping that the user can observe the following:

- Lack of diversity among their favorite movies: only one movie on the cluster plot shows has less than a majority white cast. Both exploration via the cluster plot and the comparison visualization show the lack of diversity among top movies.

- The positive relationship between the proportion of white cast members and the gross revenue of the film, demonstrated in Figure 3.

- Trends in race diversity among movie genres (i.e., western movies feature virtually only White cast members, and no Asian cast members).

Our visualizations not only make it easy to track Hollywood-wide trends across movies, but they also allow users to see where individual movies fall along the spectrum of diversity and these other metrics.

6 Conclusion and Future Work

Future extensions of the visualization experience include supplementing the dataset with roles beyond actors. Currently, the awards page only examines the award categories corresponding to actors (the Oscars also includes director), while the reception page only analyzes the diversity of the cast, where the cast is the list of the top-billed actors. To each of these datasets, other roles such as director, cinematographer, producer, writer, composer, makeup artist, and set designer could be accounted for as well. This would provide the option to analyze a more holistic view of Hollywood, rather than only of the faces on screen.

Another extension is to create additional visualizations with the existing datasets. Such a visualization could analyze the resources allocated to a film based on its diversity; for example, how is the race of a movie's director or producer related to the budget that the film is given?

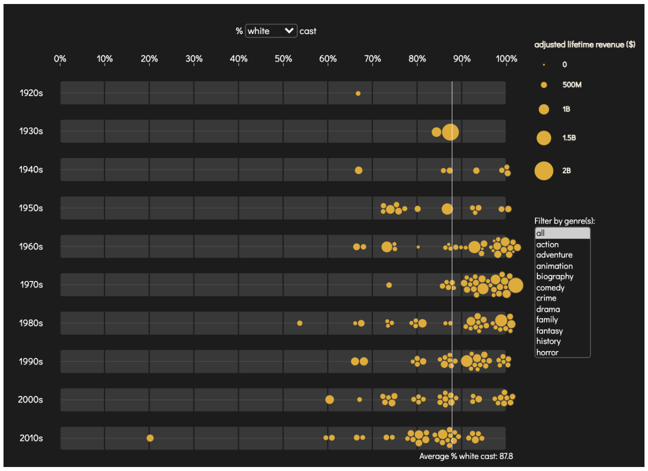
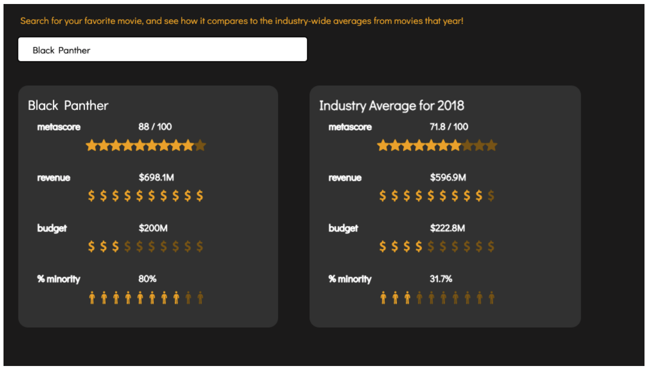


Figure 3: Reception bubble visualization. Larger bubbles, representing movies with higher gross revenue, have a larger percent White cast (located on or right of the average line).



Search for your favorite movie, and see how it compares to the industry-wide averages from movies that year!

Black Panther

Black Panther

metascore 88 / 100

revenue $698.1M

budget $200M

% minority 80%

Industry Average for 2018

metascore 71.8 / 100

revenue $596.9M

budget $222.8M

% minority 31.7%

Figure 4: Side-by-side comparison of selected movie with industry average of movies from that year, via icon-based visualization.

Improvements could be made to the existing experience as well through gamification. The viewer could be asked to enter measures of their own demographics (race, sex, age, etc.) for an evaluation of their likelihood of making it into a top movie in Hollywood. Search for a particular movie could also be added to the bubble visualization on the reception page to allow for easier location by movie.

Overall, the visualization experience aims to inform the user on the progression of Hollywood diversity over time to its present status quo. In both the awards and reception visualizations, it is clear that diversity has increased since the start of cinema, especially in recent decades with the advent of high-grossing films like Black Panther and Crazy Rich Asians. However, there is still a long way to go, as majority of high-grossing films have a top-billed cast that is majority, if not all, white, and minorities receive disproportionately few acting awards and nominations.

Acknowledgments

The authors wish to thank the students of MIT's 6.859 Spring 2021 class for their feedback through the course of the visualization's development, the course staff, especially TA Doga Dogan for his insightful suggestions, and Professor Arvind Satyanarayan for his detailed instruction and guidance.

References

[1] I. Chua. Asian representation in movies: have things changed since 1997? *Kontinentalist*, 2019.

[2] Y. Holtz. Most basic circular packing, 2018.

[3] D. Hunt, A.-C. Ramon, M. Tran, C. Chang, A. Stevenson, and K. Tambree. Hollywood diversity report 2020. 2020.

[4] W. Su. Nyc foodiverse. *Information is Beatiful Awards*, 2017.