Milestone 3: The Disney Story
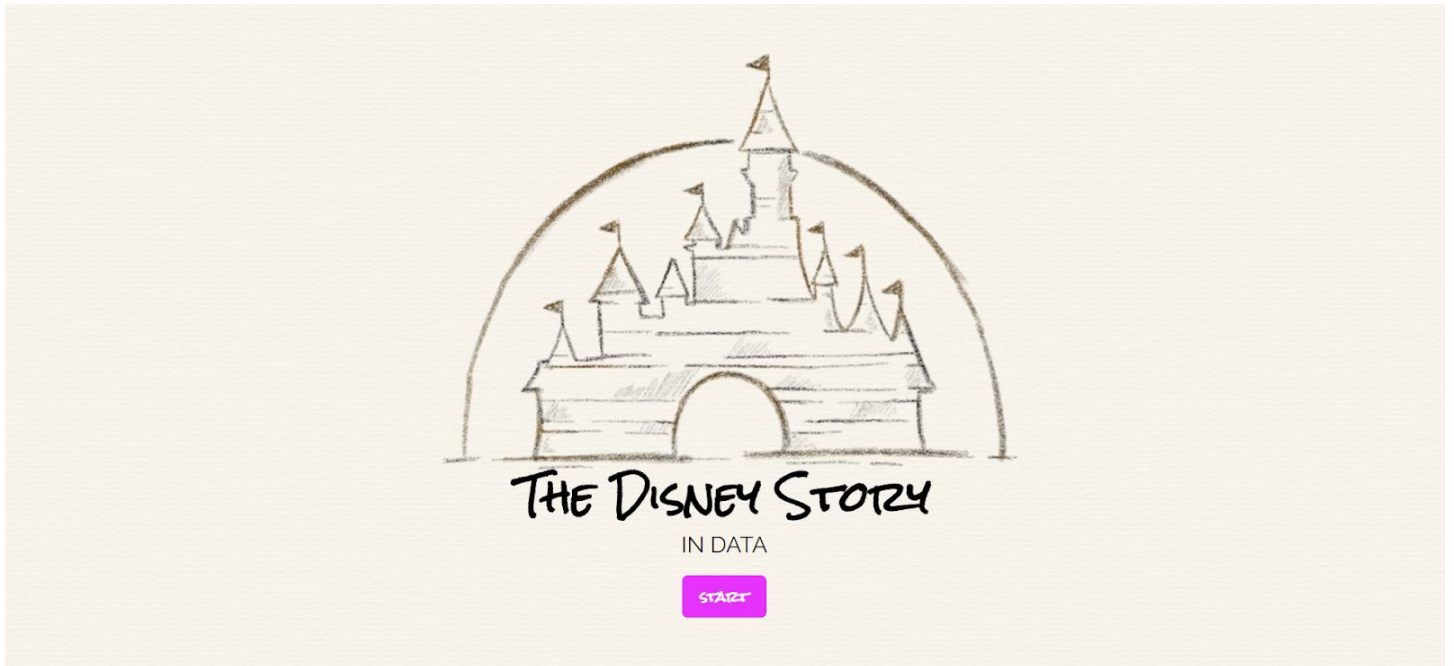
Team CJK:

Bang Chi Duong   Jenessa Tan   Ana Katrina Tan

Overview

Our project uses a scrollytelling approach to tell the story of the Disney animation studio. Our target user is any individual who are fans of Disney animation or animation in general. We provide a background of the animation studio in relation to the Walt Disney company through financial data. Afterwards, we allow our viewers to explore and learn more about the different movies, the eras of disney animation, and some of the actors behind Disney's lovable characters.

Data

Data Sources
Our visualizations use the Disney dataset provided in Data World (https://data.world/dot2/disney-character-data-set-project/workspace/project-summary?agentid=kgarrett&datasetid=disney-character-success-00-16) and the Academy Award dataset from Kaggle (https://www.kaggle.com/theacademy/academy-awards). Missing movie and actor data was taken from IMDB. Character images were taken from https://www.behindthevoiceactors.com/ and https://disney.fandom.com/.

Preprocessing pipeline
There were slight inconsistencies between the Disney datasets in the Data World workspace. Therefore, most of our data preprocessing involved rectifying those inconsistencies. This includes:
- Filtering out voice actors/roles that were not in a Disney animated feature film

- Filtering out live action movies that are not relevant to our project
- Adding missing movie box office revenues
- Adding missing voice actors & roles of Disney animated feature films
- Formatting the dates in the movies dataset for consistency and Javascript compatibility
- Filling in IMDB ratings for the movies
- Joining the Academy Award dataset with the movies dataset to obtain the wins for the movies of interest
- Joining the Academy Award dataset with the voice actors dataset to obtain the wins for the voice actors of interest

Some preprocessing steps were done manually in Excel, and some were done automatically in Python (joining datasets to get awards for movies and actors). The Python codes were included in our github repo as a zip file.The joining process can be laid out as follows:

1. Cleaning the movie titles and actor names in movies, actors, and awards datasets
2. In the awards dataset, columns called "Name" and "Film" are relevant. For each of the movies and actors datasets, two left joins from the award to the movies/actors were carried out and then concatenated to produce a single result. Since we were dealing with strings comparison, I used fuzzy matching (package "fuzzywuzzy") with a threshold of 95% on the matching score.

Data Description
Description of your data in both domain-specific and abstract language (dataset type, scale/cardinality).

Domain-specific description
For our first visualization (area chart), we categorize the Walt Disney Company's different sources of revenue into 5 general categories: Disney Media Networks, Parks & Resorts, Disney Interactive Media, Disney Consumer Products, and Studio Entertainment. We present the increases and decreases in revenue for each category as well the overall revenue (expressed in billions of USD) of the company from 1992-2016 (a period of 24 years).

The second and third visualizations show information about the different movies produced by the Walt Disney Animation Studio.

The dotplot shows the year of release, revenue (in USD), and Disney animation era that each movie belongs to. Through this visualization, the user is also able to see the number of movies released for a particular year by the animation studio.

On the other hand, the node-link graph visualizes the relationship of voice actors and movies, the IMDB rating of each movie, and oscar winners among the movies and voice actors. The links on the graph show which voice actors have worked on multiple movies.

Abstract language description
The dotplot and the area chart both use a table dataset type while the node-link graph uses a network dataset type.

The area chart uses the year as a quantitative attribute for its x-axis and the revenue as another quantitative attribute for the y-axis. There is one categorical attribute with a cardinality of 5 which represents the sources of revenue for the Walt Disney Company. The revenue ranges from $0-80 billion while the years range from

1991-2016. The area chart shows the part-to-whole relationship or the contribution of each sector to the revenue of the company as a whole.

The dotplot uses the year for the x-axis. While the year is a quantitative attribute, to save on pixel real estate we used a categorical scale to display the year, marking areas where there are breaks in the continuity through a line. This is because the year ranges from 1929-2016, which is a period of 87 years. The area of each dot encodes the revenue generated by the movie and this can range from $0-490,000,000. The number of dots above each year indicate the number of movies released that year. Typically, there are about 1-3 movies released in a single year.

The node-link graph shows two different types of nodes: actors and movies. These types are encoded through the start and mickey mouse head shapes respectively. The mickey head's size is coded to visually show the IMDB rating of the movie. While IMDB ratings are between 0-10, the minimum rating for a movie in our database was around 5 and the highest was close to 9. The thick stroke around each node indicates if a movie or actor has won at least one oscar. Links between movie nodes and actor nodes encode the role that the actor plays in the movie. For this dataset, there are roughly 80 movies, 469 voice actors, and 601 roles/links.

Goals and Tasks

Our goal was to provide general users with interesting, interactive visualizations that would allow them to explore and learn about different aspects of the Walt Disney Company and Disney animation.

First we begin by providing the Walt Disney Company's financial information to users. We want users to be able to compare the gross annual revenues from various sectors within the Walt Disney Company [compare values] and to observe the trend of generated revenue for that sector in isolation from the others [discover trends].

Afterwards, we want users to explore the connections between voice actors and the various movies they may have participated in [explore data]. We want them to discover for themselves information about particular movies such as Academy Awards, movie revenues, and movie ratings. We also want them to learn more about voice actors that participated in these movies, to find which voice actor participated in more than one movie, what roles those actors had for each movie, and if an actor was an Academy Award winner.

Finally, we want users to gain insight into the various Disney eras and the movies that defined them. They can observe the distribution of movie releases and compare box office revenues between different movies [discover distribution and make comparisons].

Visualization
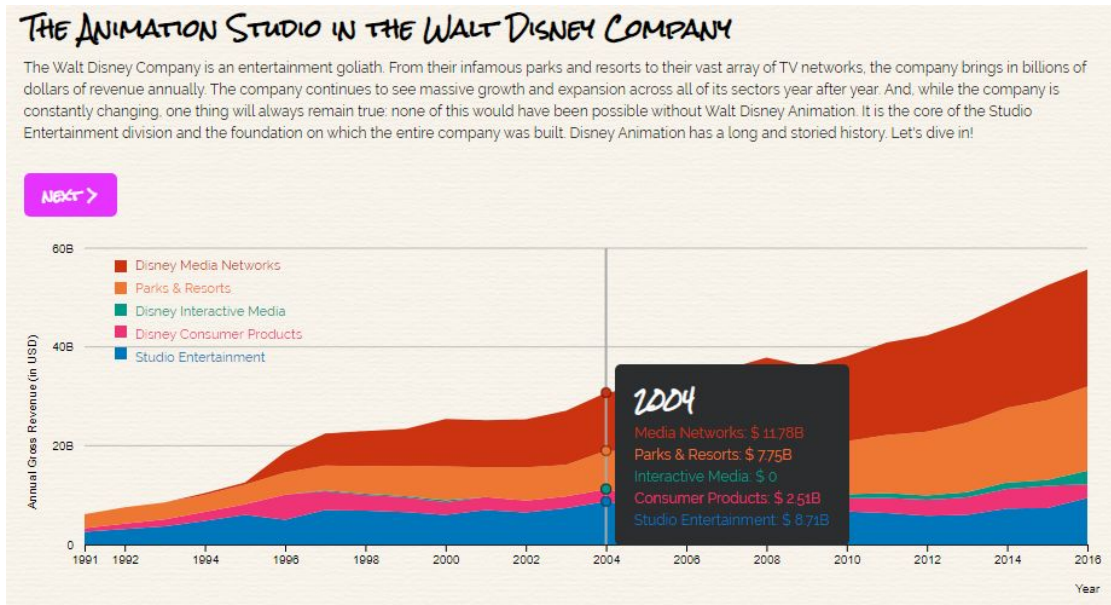
We started off with an overview of the animation studio as part of the corporate Walt Disney Company. To do this, we showed the contribution of the animation studios to the company revenue over time. We utilized an area chart to show the part-to-whole relationship of the animation studio to the overall company revenue. The revenue is a quantitative value that ranges from USD 0-60 Billion. We considered using either a clustered bar or stacked bar chart but the years range from 1992 - 2016 (i.e. a period of 24 years) and we felt that there were too many years to opt to use the clustered bar chart. In addition, we felt that area charts were great since it expresses the continuous flow of time so we opted for that rather than the stacked bar chart. The revenue is broken down into 5 categories so we used a categorical color scheme. Our main focus is the animation studio, so we aligned it on the bottom axis so that viewers are still able to easily read the values it contributed on its own using the y-axis. Users can view a detailed yearly breakdown of the annual revenues across different sectors by hovering over the year. Hovering over the sector name in the upper left legend, users would be able to observe the trend of generated revenue for that sector in isolation from the others.
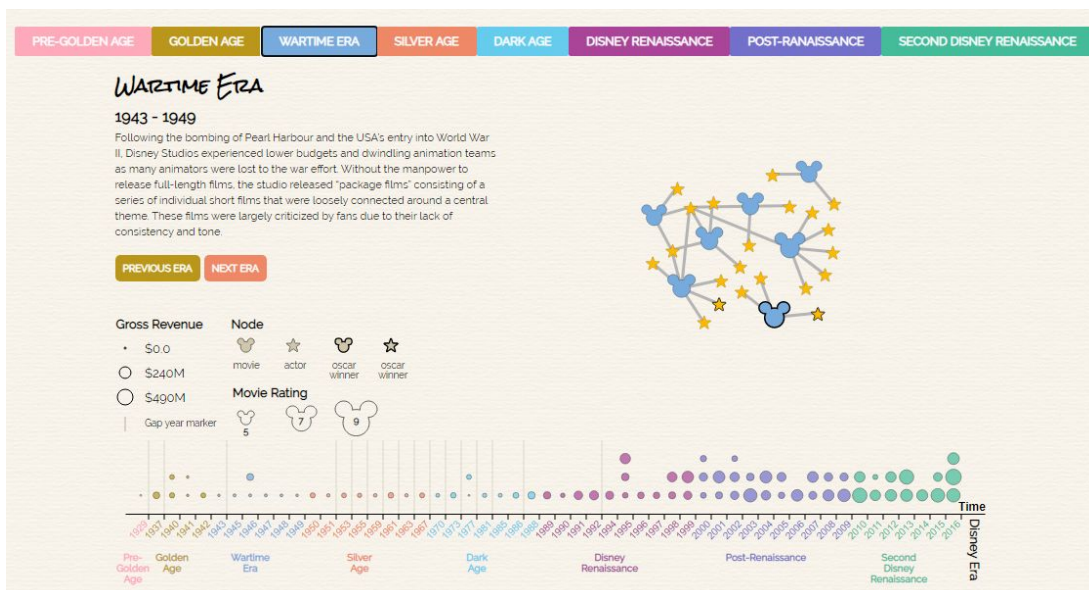
After showing the financial part of the animation studio, we wanted users to be able to explore the data related to individual movies and the voice actors behind some of the main characters in the movie. We decided to

express this through a node-link graph. We shape-coded the nodes where voice actors were represented as stars and the movies used mickey icons. We wanted to use these shapes to bring the Disney theme into our design. The movie node is size-coded to show its IMDB rating. Users can determine which movies were not rated as high compared to other movies in the same era or if movies in that era had similar ratings. We decided to use area since it is a channel that was available and can be accurately perceived by a viewer. Viewers can choose which era they would like to view using navigation buttons at the top of the graph or the buttons below the era descriptions. We also used color hue to encode the era each disney movie belongs to. This is to be consistent with the dot plot graph below it. We have identified 8 different categorical eras (technically there are 9 from our research but we have no movies for the silent era which dates before 1928 so we have decided to ignore it). We are once again utilizing a categorical color palette. The thicker stroke on the movies and voice actor nodes indicates if it is an Oscar Winner.


Hovering over the nodes and links brings up a tooltip that provides more information about the movies and voice actors as well as the relationship between the two.

Hovering over a node highlights its one-hop neighbours and fades the other nodes and links. In this manner, users would be able to explore which movies shared voice actors and how many movies a voice actor participated in. If the user hovers over a movie node, the corresponding dot in the dotplot will also be highlighted, fading the other dots in a similar fashion. Users may also drag nodes to reposition them if the initial simulation arranged the links in a confusing manner.
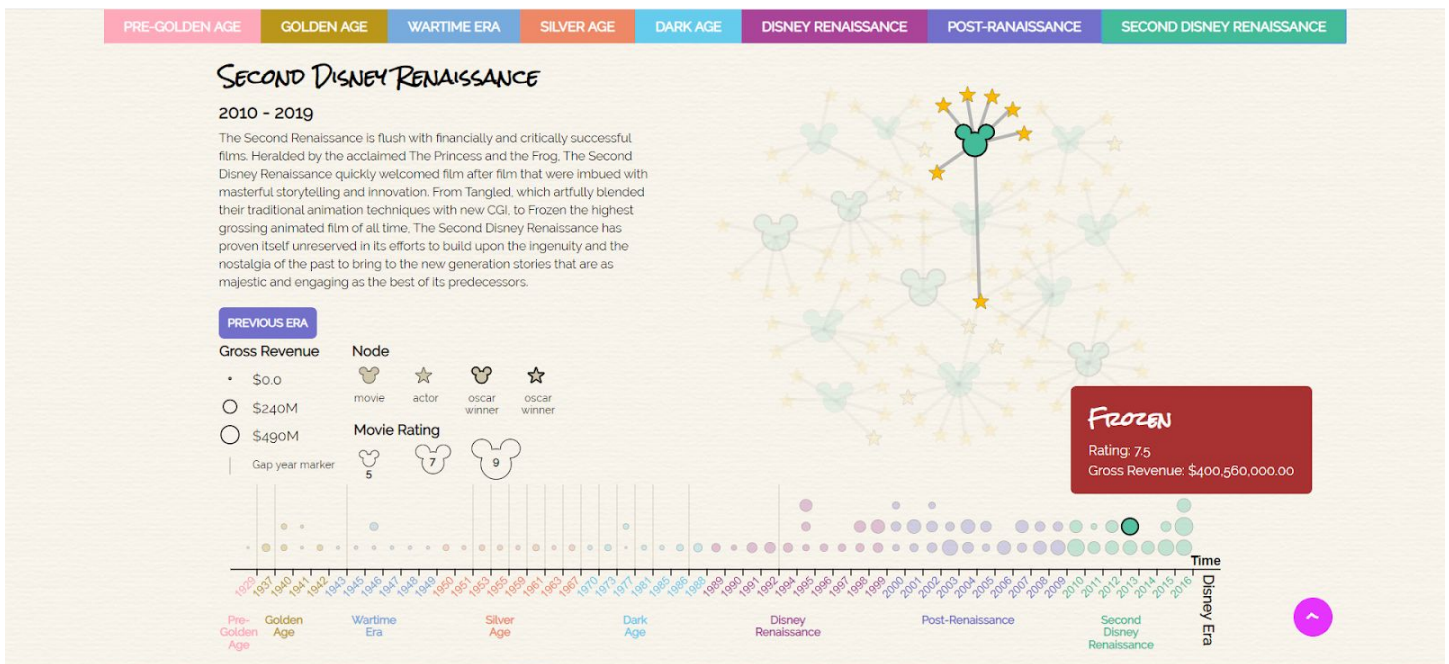
From going into detail about the movies and their related voice actors, we decide to take a step back and provide another overview of movies. We utilized a dot plot and size-coded the dots to express its gross revenue. As previously mentioned, we use color hue to encode the era each movie belongs to. We decided to use a dot plot because each dot still expresses a single movie and the user will be able to see the number of movies released each year. Typically there's only about 1-3 movies released per year and our years range from 1928-2016 (approximately 87 years). We chose the dot plot visualization over a bar chart because being able to see the movies individually is important. Similar to the hover functionality mentioned in the node-link graph, hovering over a movie's dot here will highlight the corresponding movie and its one-hop neighbours in the node-link graph, fading the other nodes and links.
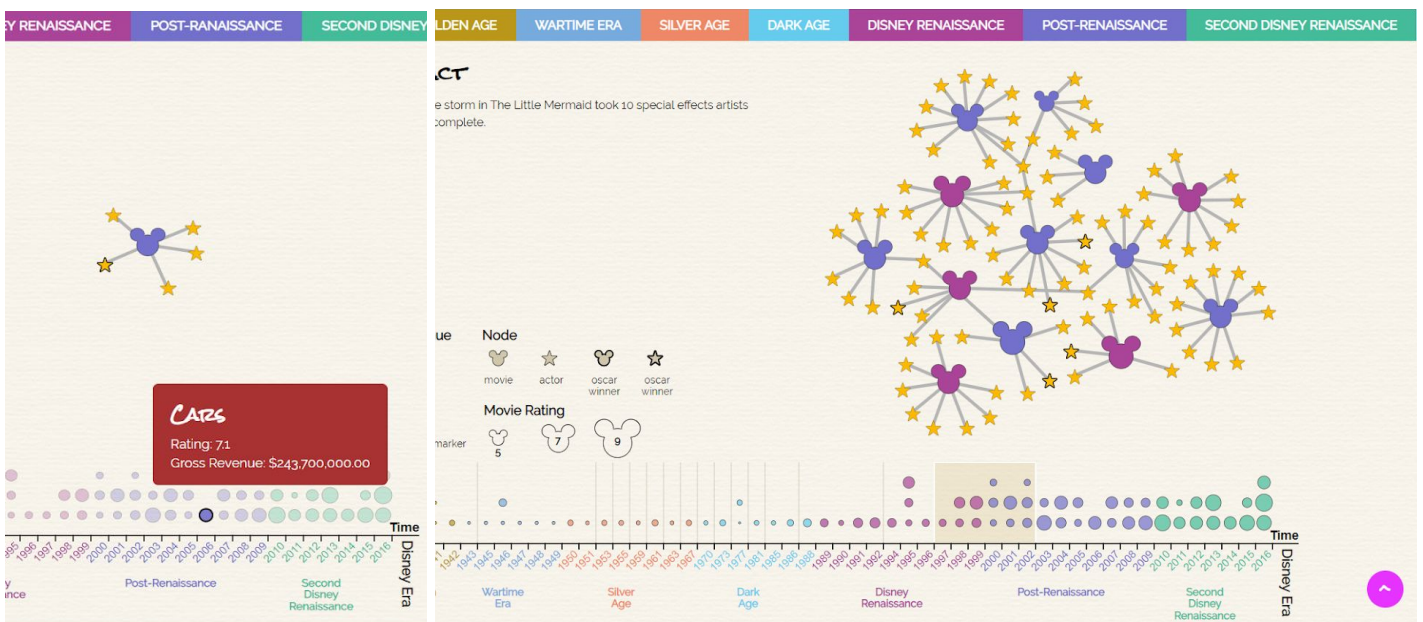
This dot plot serves as an interactive navigation for users to see different movies in the node-link graph in two different ways. First, the user may select a movie's dot. This will select the movie and its voice actors in the node-link graph and remove the other nodes and links from the view. This helps a user to focus on the movie and its cast of voice actors. Second, the user may use a brush selection to choose a custom time range. This

removes the limitation of seeing voice actor-movie relationships only for movies within the same era and allows users to explore our dataset better.
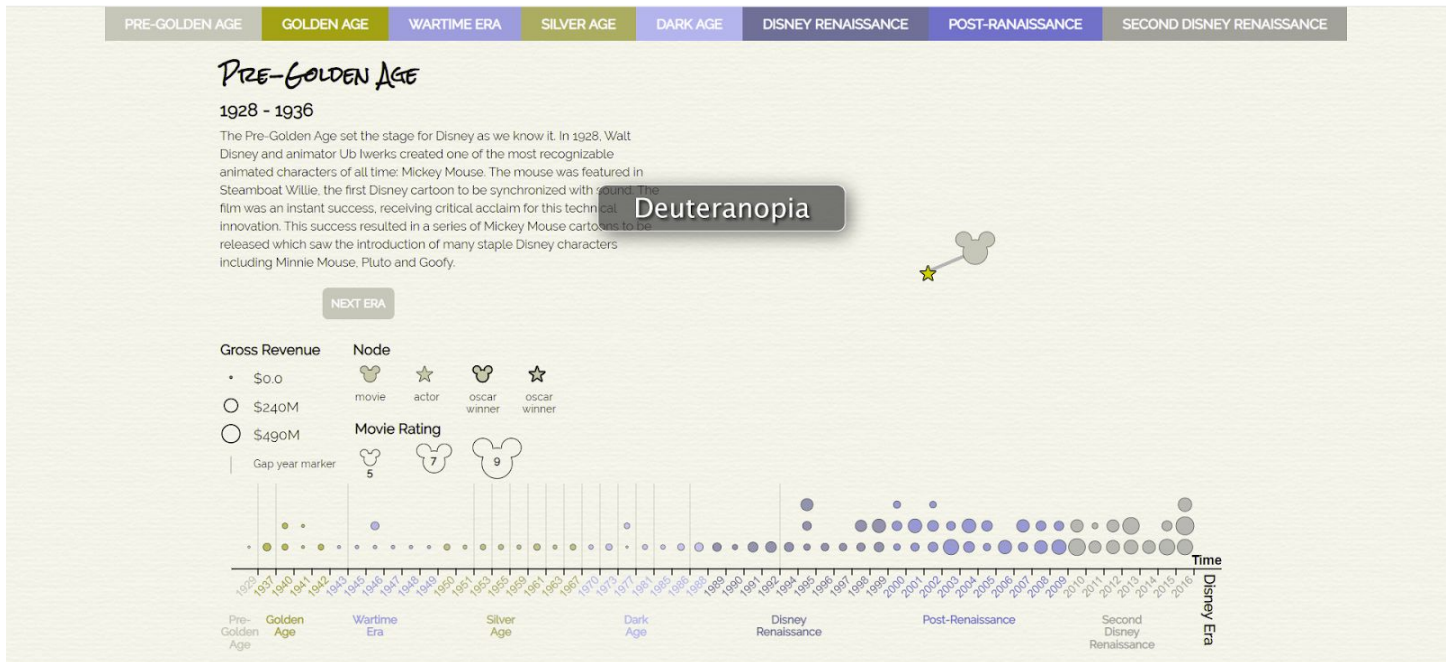
All the color palettes used for these visualizations are categorical. While the area chart requires only 5 colors, it was more difficult to choose colors for the node-link graph and the dotplot because it required 9 colors (8 eras and the voice actors). After some research, we found Paul Tol's suggested color palettes (https://personal.sron.nl/~pault/#sec:qualitative) and modified it a little to match the style of our site. We tested the visualization further using Color Oracle (http://www.colororacle.org/manualwin.html) which shows the display as viewed by someone with color blindness. For the colorblind screens, we mostly focused on getting adjacent eras to have distinguishable colors. See the color scheme below with the different simulated views for colorblindness.
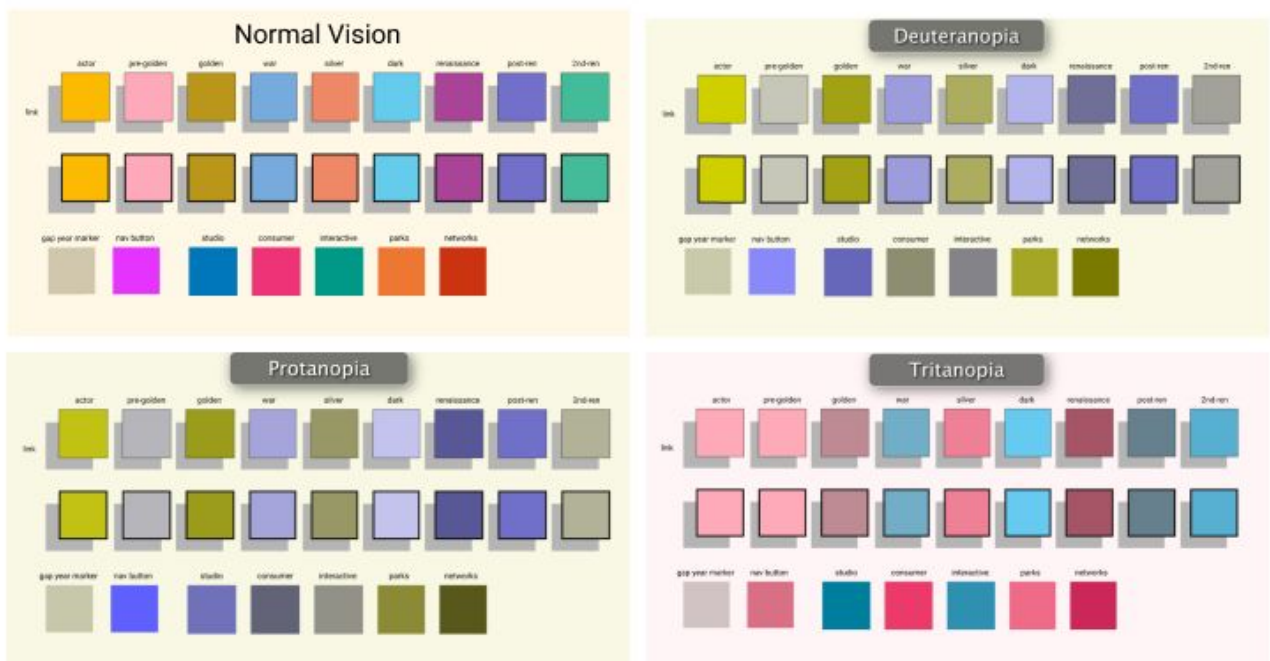
Reflection

Our vision for this project has not changed since our proposal. We did have to make slight adjustments to our original plan from a usability and structural design perspective. For example, we had initially planned to display all of the movie and actor data in the node-link graph and allow users to selectively filter. However, we did not take into account the sheer size of that dataset when drafting our plan. To avoid a hairball problem, we chose to only display the nodes and links of the currently selected era/user selected time period.

We believe that minor changes like this were reasonable choices to improve the usability of our application. They did not impact the visualization goals of allowing users to explore and learn about different aspects of the Walt Disney Company and Disney animation that we set out to achieve. We were able to implement our solution in such a way that closely aligned with our original plans, accomplishing both our intended visualization and technical goals. As such, we would collectively agree that our original proposal was very realistic given the tools at our disposal.

If given unlimited time to make this project again, we would want to do something very similar but add more functionality and interactivity, such as:
- Search functionality where users can search for actors names and have the node-link populate with the actor and all of the movies they were involved in, i.e. a dynamic node-link plot
- Expand to show different roles associated with the movies (such as directors or writers) so that users could find common links between their favourite films
- Enable users to select multiple movies that may not be temporally close to each other (for example, they could select a movie from 1970, 1999 and 2009)

Additions like these would further add to a users' ability to learn and explore the different aspects of Disney animation.

Team Assessment

Bang Chi
- processed the Academy Award data to get the award information for each of the movies and actors of interest, cleaning and joining multiple datasets as described above
- Built the dotplot view of the Disney movies broken down by year and Disney era including interactive tooltip
- Created the legends for the dotplot and node-link graphs
- Improvement on tooltip selection for node-link graph
- Implemented the drag-and-drop functionality of the node-link graph

Jenessa:
- Cleaned and processed original Disney DataWorld datasets - filtering out irrelevant movies and actors, filling in missing data, ensuring proper formatting
- Built the area view of the Disney corporate revenue including interactive tooltip
- Implemented the linked hover interactivity between the dotplot and node-link views
- Wrote narrative of the app and implemented the scrollytelling functionality
- Searched for character icon files used in the node-link tooltip

Ana Katrina:
- Processed the original Disney DataWorld datasets ensuring consistency between the voice actor and movie datasets
- Built the node-link view of movie and voice actors including interactive tooltips
- Implemented the brush selection tool on the dotplot to selectively filter movies displayed in the node-link graph
- Implemented filtering by era functionality
- Styled the app overall

C = Bang Chi Duong ; J = Jenessa Tan ; K = Ana Katrina Tan

Milestone	Date	Actual Date	Est Hours	Actual Hours	Team Member
Narrative	Mar 10	Mar 23	3	4	J
Data Processing	Mar 14	Mar 25	12	12	CJK
Glyph Files	Mar 15	Mar 15	3	0.25	K
Static Area View	Mar 18	Mar 19	7	7	J
Area Interactivity	Mar 28	Mar 28	8	8	J
Static Node View	Mar 20	Mar 27	27	24	K
Static Dot Plot View	Mar 20	Mar 24	17	13	C
Tooltip Creation	Mar 31	Mar 25	9	9	CK
Midterm Writeup	Mar 23	Mar 24	4	4	CJK
M2 Due	Mar 25		---------		
Data Filtering by Era	Mar 28	Mar 28	17	6	K
Brush Selection Interactivity	Mar 28	Apr 6	17	5	K
Linkage Between Node & Dot Plot	Mar 28	Mar 31	12	6	J
Character Icon Files	Mar 30	Apr 3	3	20	J
Legend Creation	Mar 31	Apr 6	4	6	C
Character images in tooltip	Apr 2	Apr 3	0.25	0.25	J
Scrollytelling	Apr 4	April 7	4	4	J
Drag/Drop on Node-link	Apr 5	Apr 5	4	2	C
Movie selection via dotplot	Apr 6	Apr 3	5	3	K
Color Palette	Apr 6	Apr 8	2	2	K

Oscar award parsing in tooltip	Apr 7	Apr 7	4	4	C
Tech Debt & Testing	Apr 7	Apr 8	15	9	CJK
Final Writeup	Apr 7	Apr 8	6	6	CJK
M3 Due	Apr 10		----------		