

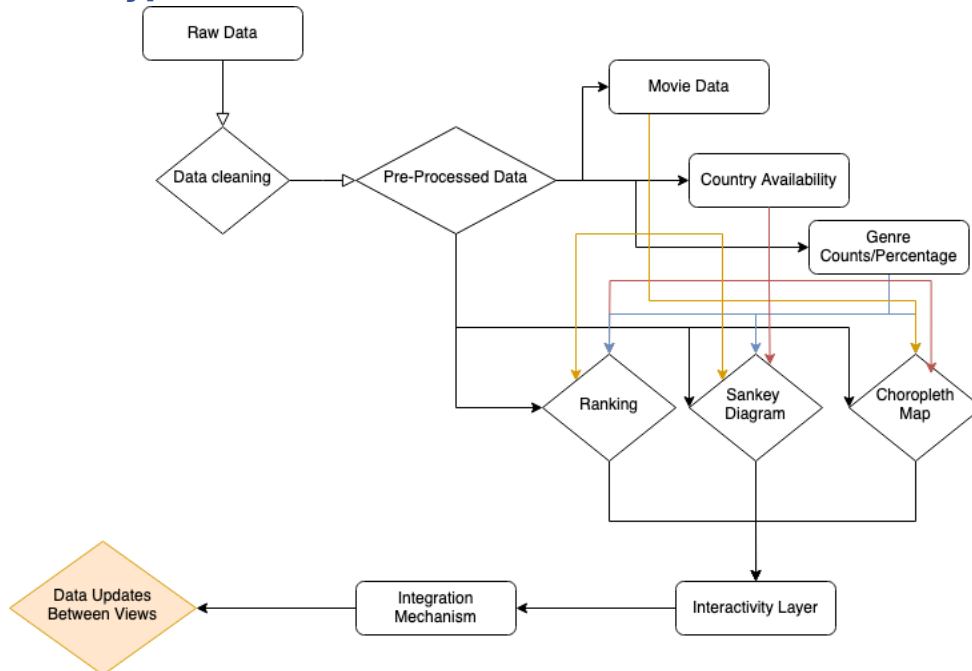


Checkpoint III: First Prototype

Group: 30

Date: <2024/10/05>

Prototype Architecture



Prototype Overview

1. Data Input & Cleaning:

- The initial dataset ('do-not-touch-netflix-rotten-tomatoes-metacritic-imdb.csv') is cleaned, resulting in 'cleaned_netflix_data.csv'. This serves as the base for further visualizations.

2. Data Preprocessing:

- Data is processed to analyze **country availability**, **genre counts**, and **genre percentages** over time. This prepares data for faster visualization once we apply filters such as only TV Shows, only certain genres and so on.

3. Visualization Modules:

- **Choropleth Map:** Shows Netflix content availability by country with region-based zoom and shortcuts, the genres are needed for the pie charts on each countries
- **Sankey Diagram:** Visualizes genre distribution and supports interactivity (hovering shows percentages, clicking filters content by TV shows).
- **Small Multiples List:** Displays top content (movies/TV shows) by region or genre with details like director and scores.
- **Year Slider:** Filters visualizations by year, adjusting the content displayed in other views.

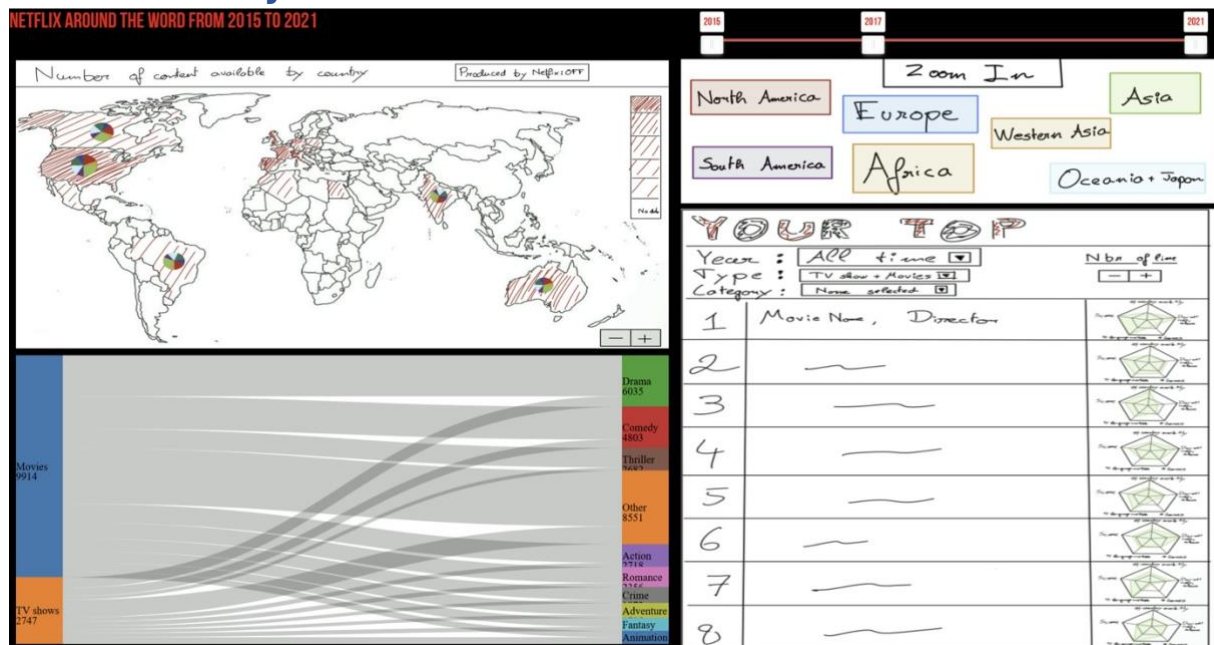
4. Interactivity:

- Hovering on the **Sankey Diagram** reveals genre numbers and percentages.
- Clicking on "TV Shows" updates genre information.
- Zooming on the Map, Map Shortcuts
- Adjusting the **Year Slider** updates all relevant visualizations.

5. Integration:

- A central event dispatcher links interactions across views. For example, selecting a year or genre updates all connected charts.
- The system is designed to be modular, allowing for easy future extensions and new visualizations.

Dashboard Layout



As of now we are still currently implementing the other idioms and the CSS of the page, but all the divs are correctly set up. We are of course trying to take example of the Netflix homepage for our background, we will have the shortcuts in the top rights and the map on the left as it was in our sketches, colors of the Sankey diagram **are not final** since they really do not match any palette as of now.

Data Processing

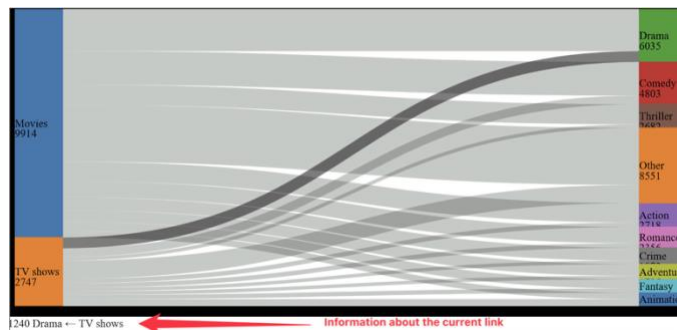
We did some preprocessing which consisted of building 9 datasets for the visualization purpose, for example we had : `tv_shows_genre_percentages` which consists of 9 columns,

Genre	2021	2020	2019	2018	2017	2016	2015	Overall
-------	------	------	------	------	------	------	------	---------

This dataset has computed all the percentage by year and overall in regard to all the genres, but of course we will not use the whole dataset for this, since some genres have maybe 10 movies or tv shows, hence we did a threshold arbitrarily $< 4\%$ in our JavaScript files and it allows us to categorize those small percentages to an Other category.

Chart Interaction

We currently have set up a few interactions, such as hovering on one of the links of the Sankey diagram



which highlights it of course but also displays information about the number of movies in this category.

Please note that if you sum, the counts in the genres it will be higher than the total of TvShows + Movies and it makes sense because some movies are in multiple categories.

We also have click on a category and it applies a filter, for example if you click on one of the genres it will filter and adapt the views of our other uncoded diagrams (not yet implemented because the other idioms are not set up).

Lastly we have a slider as we can see in Figure 1, which allows us to filter the years on which movies appeared on Netflix, do we want to see the % of TvShows between 2019 and 2021 in the action genre for example, if so how many were there? This slider need still improvement, as for now only the left one select a single year, set of years (2015-2021) are not supported but it will display the 2015 information (in this specific case).

Chart Integration

We are still to find some mechanisms to ease the change/add of a chart, but for now we have 3 things implemented. First, we have a unique start function, which load the data sets and will begin all the start function, for now it's only two of them (the Sankey diagram and the slider). Then we use global variables, like "yearMin" or "yearMax" and the loaded data. With that it is way easier to access important data in other javascript scripts, like "uptade.js". Finally the html file is made with modulization in mind, all elements are connected and rely on their parent for basic settings, like the font (here the Netflix font).