# InfoVis Final Report

## Group 37-A

**Nuno Amaro** ist181824
**Rafael Benjamim** ist181676
**Francisco Dourado** ist169962

### INTRODUCTION

For the purposes of our visualization, we decided to explore the potential relationships between the movie and videogame industries. As avid gamers and movie enthusiasts, we wanted to visually compare the two industries and showcase them in our visualization. We did this because although there are plenty of sources and tools out there that illustrate some of our ideas, they often do it by means of dull statistics or presented in static and often simplistic visual idioms that, while visually appealing, fall short in describing the actual data.

We found this topic especially relevant since we belong to a university course that is known to have quite a lot of videogame fans, as well as, the blooming videogame industry that seems to keep getting larger every year. We decided to cross-reference our data with the movie industry as we believe it to be the closest medium of storytelling to that of videogames as well as being one of the most comparable in terms of sales figures.

In the beginning of our project, we committed ourselves to focus on four tasks that would be answered through our visualization. With them, we wanted to get a better insight on how both industries behave and potentially interact with each other. What follows is a brief overview of these tasks:

- Exploring the relationship between sale figures and overall popularity of videogames, within different regions.
- Comparing user ratings and critic ratings and how they affect sales values.
- Present both global and regional preferences in videogame genres and compare these with movie genres in a specified time frame.
- Explore the potential relationship between a videogame/movie's title length and sales figures.

These tasks were initially presented accompanied by some examples of questions we wanted to see answered by our visualization. There were a total of six example questions submitted to accompany our proposed tasks, namely:

- Is a good score necessary for a game to be a best seller?
- How much do controversial games sell?
- Does Asia have different taste in videogame genres compared to Europe?
- Does Avatar's (2009) genre have an influence on the popular videogame genres from 2009/2010?
- Can poorly rated games have good sales?
- So short titles favor greater sales?

We feel that questions such as "How much do controversial games sell?" and "Do short titles favor greater sales?" can be answered in the final product. However, we adopted a more broader view of the domain and, because of that, questions that would focus on individual titles, for example "Did Avatar (2009) influence videogame genres in 2009/2010?", were left unanswered even though we could potentially look at the genres within that time frame and check if there has been any change with previous years.

### RELATED WORK

Initially, our inspiration for the visualization were works from previous years contained in the Hall of Fame[1] section of the course's Fénix page. We wanted to understand what kind of visualization was expected and, in a way, get some ideas of idioms that worked well with each other and that could support our data. Some of the takeaways from this initial research lead to our first iterations of the visualization sketches.

While we did try to find similar works to what would be ours, most comparisons between videogames and movies end up being opinion pieces with little or no data visualization to support them. Being two of the biggest entertainment mediums of today, they're subject to ongoing debate.

As for visualizations of each individual industry, we did find some examples that we can look towards, in order to explain why we believe they do not do what we needed. As a side note in the following examples, none of them are similar to our own in the sense that they are not visualizations by themselves, but rather 'article-like' pieces that contain visualizations to support their analysis. This by itself is already a factor we wish to improve upon, by having a standalone dashboard that requires no further explanation or analysis beyond that of the user's own interpretation of the visualizations.

The first example[2] takes reference the videogame industry and contains data not too unlike our own. Our issue with this example is its simplistic approach to the idioms presented. Of all the charts presented, they are all either scatterplots or bar charts and while there is nothing wrong with them, per se, some of them could very well use more visually interesting idioms. We also would note that all the plots are

static, something that we improve upon in our work by allowing more interactivity and dynamic.

A better example of interactivity can be found in the article[3] we found on the movie industry, which contains a much larger variety of idioms, however, we believe that the focus of the domain in this article might be its biggest flaw. By choosing a more individual look into films, they produce charts that have a lot of visual clutter and their interactivity, while present is unintuitive.

Of course, we could not depart from this section without mentioning more broader sources of inspiration that we used: *r/dataisbeautiful*[4] and the D3.js example gallery[5].

## DATA

The data collected to support our visualization was publicly available at the popular Data Science/ Machine Learning online community site *Kaggle*[6]. Separated into two different datasets, each containing information about thousands of videogames[7] and movies[8].

Both datasets required extensive amount of cleaning and processing. Initially, we verified that there were missing data in both. The videogame data, for example, had almost no data in regards of score ratings from before 1996. The movie data, on the other hand, consisted of a large CSV file which contained JSON entries that needed to be properly handled before anything else could be done. To solve most of our data related challenges, we used Python programming language coupled with some third-party libraries[9,10] to parse, clean, process and, in some cases, build new data from the already existing one.

Since we wanted to visualize our data with reference to a specific time frame, we initially thought we could exclude data from before 1985 in order to give videogames a place within the comparisons. This idea, however, was later discarded and replaced with the exclusion of all data prior to 1996 due to the inexistent data from score ratings from that era. Attempting to keep said data would result in half of our visualization being blank, since it required data inputted from ratings.

Certain compromises were also made when processing the movie dataset. There were quite a few genres with very low representation and had to be change to other, equally suitable genres, with a larger population. This was done because of the space available for the star plots. This idiom works well with different categorical data but doesn't scale efficiently for high number of categories. This coupled with the small space we considered for these idioms, lead to our decision.

Of all the data we managed to acquire (16K entries with 16 attributes and 4.8K entries with 20 attributes, for games and movies datasets, respectively), we ended up not finding a suitable regional sales data for movies. Although not exactly what we intended, it's not very problematic since we believe our visualization works well around this issue. As for the remaining data that we required for our purposes, most of it was already contained within the datasets.

A majority of the data was either Nominal or Quantitative. Most nominal/categorical attributes came from the Titles and Genres, while most of the quantitative variables came from Score Ratings and Sale Figures. An interesting, although later unused (due to significant missing values) attribute, ESRB Rating was ordinal.

A lot of variables were initially considered for the visualization but later left unused, primarily because of our focus on a higher-level exploration of the domains. This was due to the most individualistic attributes, such as release dates, not being so useful anymore, and other attributes such as game platform or developer not having a good movie counterpart. To make a solution that blended both domains in a satisfactory way, we had to let these attributes go as including them could potentially create several different visualizations within the final product rather than a single one.

We did, however, derive one measure from the videogame dataset: *Score_Diff*, a diverging type attribute with its central point at 0, which stands for the difference between the score rating given by critics and the score rating given by users. The rating data was already present in the dataset and was originally from Metacritic[11]. This new attribute was used to gauge how "controversial" a game was, i.e., when did critic and user opinions of the game differ the most.

There was some processing of attributes necessary before we could hand them to the visualization. The aforementioned *Score_Diff* attribute, for example, required us to change the User score rating (scale 0.0-10.0) to match the Critic score rating (scale 0-100). Titles needed to be processed to gather their length and videogame sales were originally in units and had to be calculated into monetary values (in dollars) using their release dates and inflation for those years. Since videogame sales span several years and different price points, we considered a slightly lower initial value as well as matching said value to the different platforms since that also affected the initial price of a game.

The data used in our final visualization is much different than the one delivered during Checkpoint II. We thought we could get away using two CSV files, one for videogames and another for movies, serving as our main datasets that would feed our visualization. However, this changed after feedback was given on our delivery and confirmed once we started working on the prototypes. While it could be said that it would be possible, it definitively complicated our work and, step-by-step, we began refactoring our data to ease our transition from concept to code. In essence, we realized that it would be much better to do all the heavy lifting for the data beforehand, than it would be to do it in real time with D3.js[12].

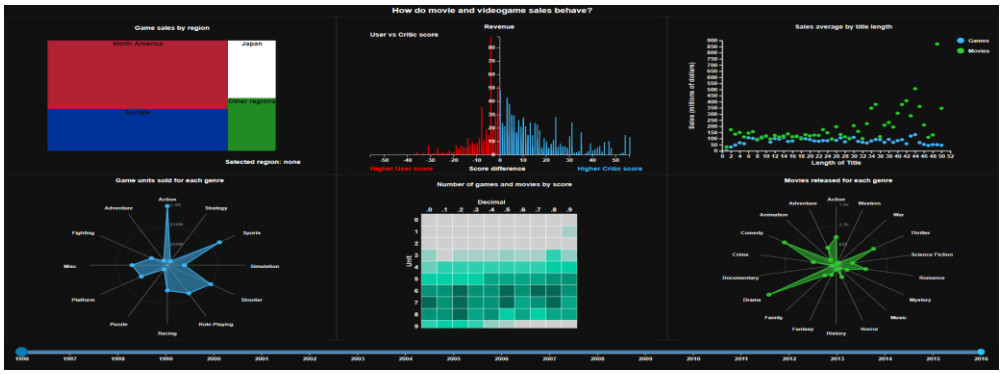


**Figure 1: Visualization Dashboard**

## VISUALIZATION

### Overall Description

Our final solution can be seen in the beginning of this page and contains all proposed idioms using the data collected. As the reader can see, we decided to have as few external selectors (aside from the Timeline) as possible.

The visualization is, from left to right, composed of a treemap, a bar chart and a scatterplot on the first row, a heatmap bordered by star/radar plots on the second row. The third row consists, uniquely, of a Timeline which serves as the only external selector of the visualization and the overall main selector. Although described as the single external selector, that does not mean that data cannot be selected or filtered in other ways, with some of the idioms serving as a "secondary" selectors, allowing interactivity between different idioms, as well as, having certain properties to filter data within themselves.

We decided to keep a fairly simple color scheme in our visualization, with a dark grey background, while games and movies were represented with either blue or green, respectively (more about color in the descriptions of each idiom).

For a better understanding of the dashboard mechanics, we will describe each part of the visualization and demonstrate their full potential.

*Timeline*

**Figure 2: Initial Timeline**

As mentioned previously, the Timeline serves as the main selector of our visualization and allows the user to witness the evolution of our project's domain over time or get an idea of how it behaved within a certain time frame. The user can select a range of different dates or select a single date, which will modify the remaining idioms to the selected time frame.

**Figure 3: Timeline with range of years selected (2002-2013)**

To better appreciate how the evolution of the industries behaved over time, we recommend selecting the timeline year by year and see the different changes within other idioms (an example can be later seen in the "Demonstration of Potential" section).

**Figure 4: Timeline with a single year selected**

*Treemap*

Our first major idiom is located at the top-left corner of the dashboard. The treemap consists of four sections, each representing the four major videogame sale regions of the world: North America (Red), Europe (Blue), Japan (White) and Other (Green, representing the remaining areas of the world).

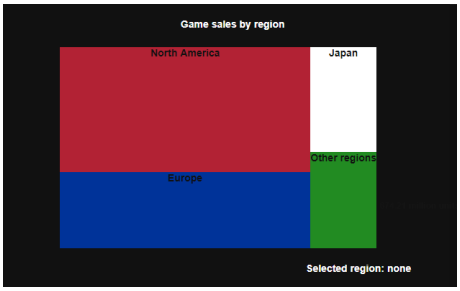


**Figure 5: Treemap**

The colors were not chosen by random, but rather by colors of their flags. North America's color is aptly named "American Red" and is featured in the United States of America (the biggest consumer of the region) flag. Europe is colored in dark blue, the main color of the European Union flag. Japan is white, not only because of the main color of its flag but also so to not confuse the user with North America, if we were to chose the color based on its red center. Other Regions are colored in green, as it symbolizes the rest of the world and was chosen based on the phrase "Green Planet".

The treemap represents the number of videogames sold, in a given time frame. Each inner rectangle's size is calculated with the number of units sold in each region. That size is, therefore, proportional to the global amount of units sold. Upon hovering a given region, a tooltip appears informing the viewer of how many units were sold, in millions (with two decimal points).

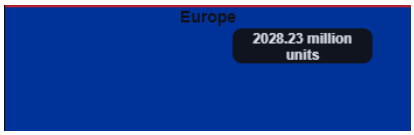


**Figure 6: Europe region being hovered**

Besides being a visual representation of regional sales of videogames, each region of the treemap can also be toggled. By choosing the region, we can update the bar chart and videogame star plot to display relevant values to that region.

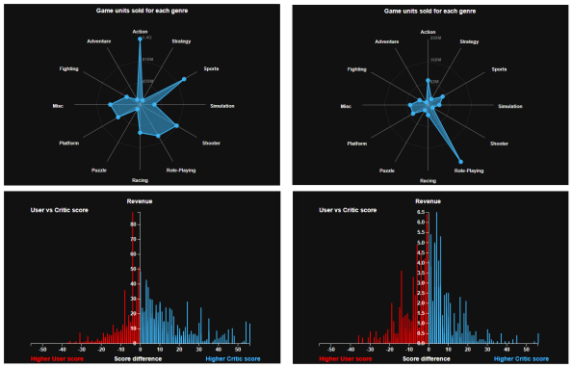


**Figure 7: Comparison between Global on the left and Japanese on the right (1996 to 2016)**

*Bar Chart*

Our bar chart, located just to the right is the idiom responsible to show the viewer how a controversial videogame sells. As mentioned, it can be used to see data within a given time frame and/or specific region (see Figure 7, for an example of this behavior).

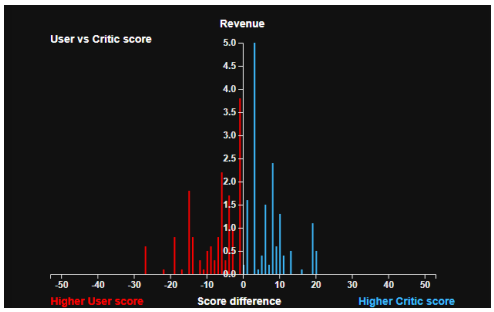


**Figure 8: 2007's global bar chart**

The vertical axis, which scales to better represent any selections made, represents the total revenue of videogames in millions of dollars.

The horizontal axis is, in the case of this idiom, the most interesting. As mentioned before, *Score_Diff* was an attribute derived from the user and critic score rating of each videogame. This attribute, the difference between Critic score and User score, is shown in the horizontal axis and represents the controversy surrounding a game. This controversial score means, in other words, that there wasn't a clear consensus around the game's score. Values to the left signify that users gave a higher rating than industry critics and, by consequence, the right portion of the chart represents games with a higher critic score than the general community.

*Scatterplot*

Unlike the rest of the idioms, the scatterplot always presents a global view of how the length of a title impacted the sale figures of each industry.

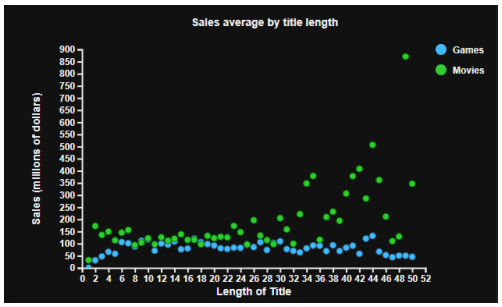


**Figure 9: Initial Scatterplot**

As mentioned before, games and movies are here differentiated by color, blue and green respectively. The legend, at the top-right corner serves both as a visual guide and selector, being able to toggle the visibility of each element by clicking on the legend circle. If toggled, the legend circle will grey-out and the respective element disappear. The now grey circle can be toggled once more to bring back the elements.

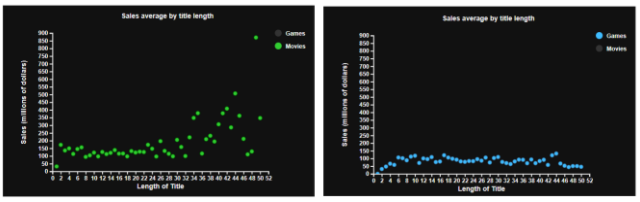


**Figure 10: Different views of the scatterplot**

The scatterplot also features the ability to be panned and zoomed in/out. This can be achieved by holding down the mouse button and dragging along, in the case of panning, or by use of the mouse's scroll wheel to zoom.

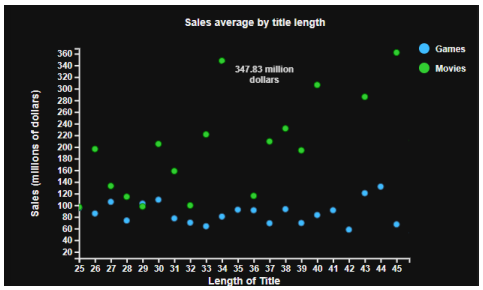


**Figure 11: Example of scatterplot after zoom and pan (with hovered element)**

*Star plots*

Our star plots, otherwise known as radar charts, are our way of representing the different genres of movies and videogames. As not all genres have a direct match between

industries (example: Role Playing Game), we opted to build two distinct star plots.


**Figure 12: Star plots for 1996-2016**

These plots can be controlled either by manipulating the timeline, selecting a region in the treemap, as illustrated in Figure 7, and/or selecting scores in the heatmap (Figure 15). The videogame plot represents the amount of games sold and the movie star plot contains the number of movies released, within a certain genre for a given time frame.

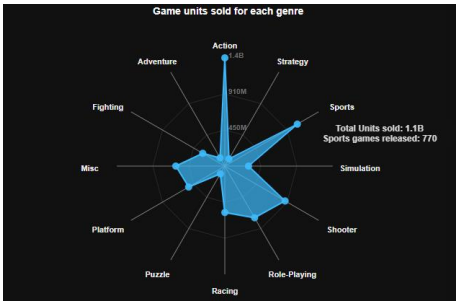


**Figure 13: Hovering Sports**

*Heatmap*

Our next and final idiom of our visualization consists of a heatmap containing the number of games and movies within a certain time frame that had a specific score rating from 0.0 to 9.9.

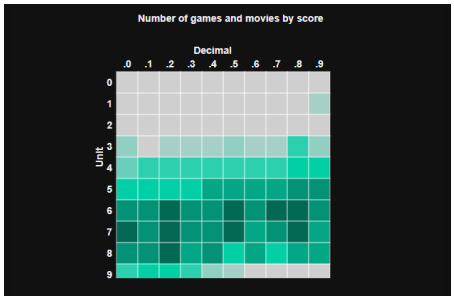


**Figure 14: Initial Heatmap**

This technique is, therefore, a 10x10 grid where the vertical axis represents the unit of the score, and the horizontal axis represents the decimal point of the score. As example, if we wanted to know how many games/movies had an 8.4 score, we would hover the element located at the '8' tick on the vertical axis and '.4' on the horizontal axis.

The heatmap also serves as a secondary selector for the star plots, being able to select multiple scores by hand or dragging along a set of scores. These selections will then update the star plots with the respective elements that had said scores.

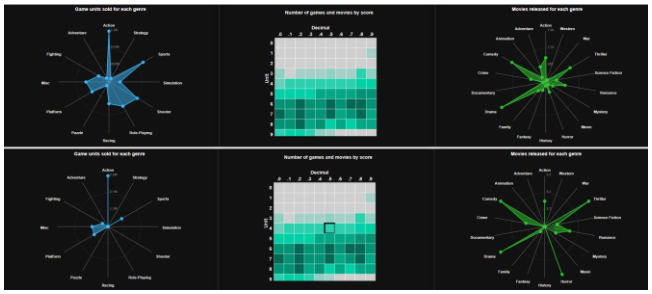

**Figure 15: No Score Selected vs. '4.5' Score Selected**

As the heatmap contains data from both videogames and movies, we thought it would be fitting to account for this in the color of the elements, choosing a turquoise color as a mix of green (movies) and blue (videogames).

### Rationale

It is important to realize that our final visualization was the product of design iteration and didn't always look the way it looks presently. The first sketches had nothing to do with the final product. For instance, when thinking about how we would present different regional sales to the user, our thoughts almost immediately remembered a lot of examples from the Hall of Fame and thought: "map".

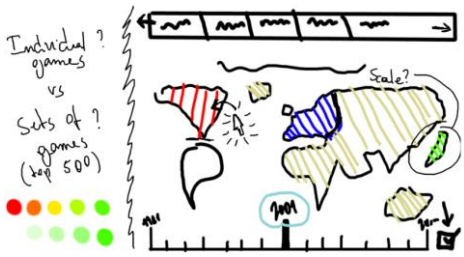


**Figure 16: Initial sketch for regional sales. On the left: our indecision on the level we wanted to present the domain**

We eventually realized that this technique was too much for what we wanted to accomplish and ended up going with a treemap, as seen in the final visualization. At every step of the design process, we would consider how each idiom would behave and how it could affect the others. An example of this can be seen in Figure 17.

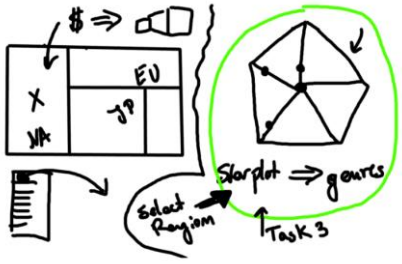


**Figure 17: Interactivity Sketches**

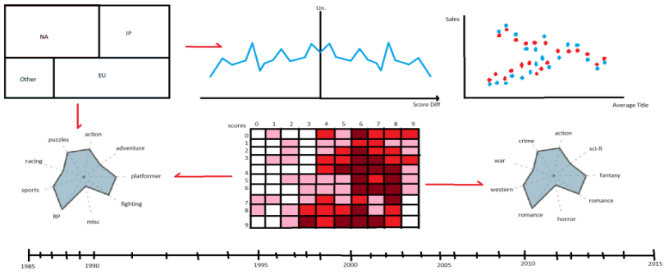


**Figure 18: One of our final sketches. Red arrows represent interactivity between idioms.**

It is worth noting that we did not stop changing the look and feel of the visualization once we started code production. As it can be seen in Figure 18, our initial idea was to have a line chart for the *Score_Diff* plot. However, as can be seen in Figure 19, our plans changed once we realized that the data did not pair well with the idiom.

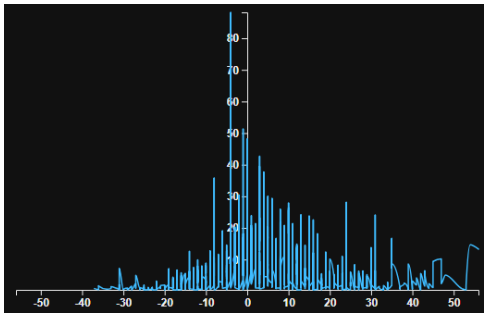


**Figure 19: A very early prototype of what would soon be our bar chart**

The treemap was chosen based on how we could tailor the proportions of the inner rectangles to the amount of units sold. This allowed us to not only show how different regions compared with each other, but also how they compared in a global scale (the borders of the treemap) without explicitly having a Global element.

The idiom representing the sales figures per title length was initially thought of as a bar or line chart, being changed after receiving feedback that a scatterplot would probably be a better fit.

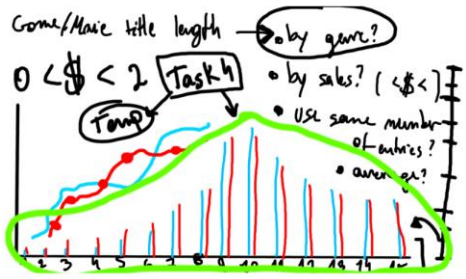


**Figure 21: First draft of what would become our Scatterplot**

For most of the design iteration, the heatmap and star plots have been present. We viewed the heatmap as our best option to present several different scores, while the star plot gave us the freedom to have several categories (genres) present at a given time.

In particular, the star plots were inspired by Pokémon videogames, which have very similar charts to present different attributes of a Pokémon.

### Demonstration of Potential

Having now gone through all the individual idioms that compose our visualization, we can now move towards answering the tasks we exposed in the beginning of this report. We did find some interesting insights we did not expect as well as other insights that are not completely unexpected but confirm our suspicions.

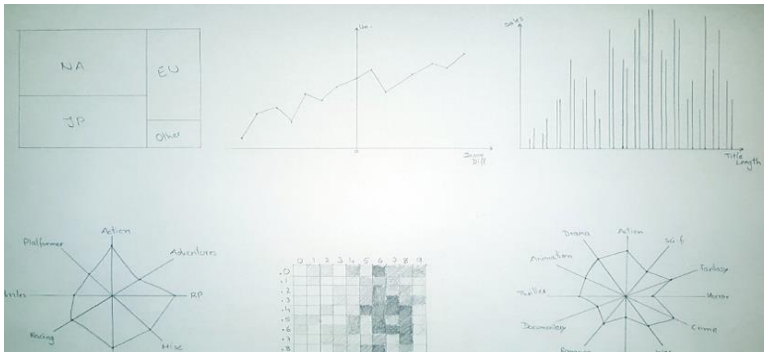


**Figure 20: A hand-drawn sketch in our presentation**

For example, one of our tasks was to compare how different title lengths sold, i.e., if a videogame/movie with a shorter title sold more or less than one with a bigger title. As we can see in Figure 21, we found that larger titles grant larger sales for movies, however, it seems that title length makes little difference when it came to videogames. It seems as if once we hit a certain halfway point, movie sales actually increase with the length of its title.

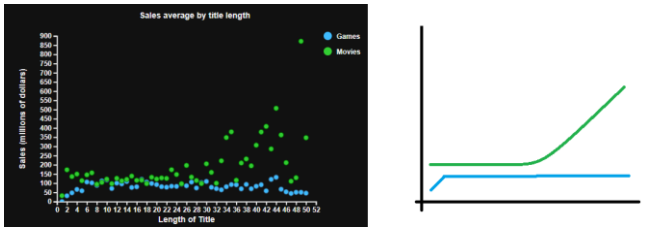


**Figure 22: Illustration of how movie and game sales differ**

Another task we proposed dealt with the question of how different regions have different tastes in videogame genres. An example of this would be a comparison between North America and Japan. This can be done via the star plots for each region (by selecting the treemap).


**Figure 23: North America (left) vs. Japan (right)**

As shown in Figure 23, there seems to be a preference of RPG games in Japan, and by contrast, North America seems to have a larger audience for action videogames.

Another insight we can take away from our visualization is how the gaming industry has evolved over time. We can see this when comparing data from the 90's compared with data from the 21st century.

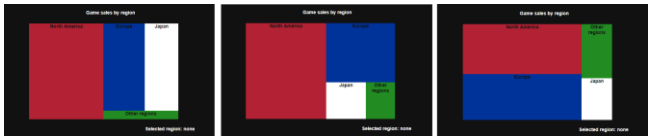

**Figure 24: 1997 (left) vs. 2002 (middle) vs. 2014 (right)**

As we can see in Figure 24, the European market for videogames starts to rise. In the late 90's both regions have similar numbers, however, in the early years of the 21st century, Europe at times has more than triple (as the example above shows) what Japan has. If we explore even later years, like 2014, we can see that Europe now matches North America numbers.

## IMPLEMENTATION DETAILS

Our biggest challenge while building our visualization was getting the initial D3 code to play nice with our data. For example, the original treemap algorithm[13] only expects a single 'root node' (much like the heatmap[14]). Our data, however, had several roots (years) and could potentially handle data from several roots in one go (array of years). The solution to this problem consisted in carefully planned data, some D3 knowledge and good old JavaScript to be able to make the algorithms believe we always had one 'root node' at any given time.

The initial scatterplot code[15] had no pan and zoom mechanic and had to be implemented using similar examples[16] we found. It also did not have the ability to select and deselect elements from the legend, which we later added.

The scale and categories of our data, namely genres, required us to have two separate star plots and required some tinkering of the original algorithm[17].

The links between views were done via regular array operations and variable settings which would then help the corresponding idioms to update accordingly. With these updates, we tried to keep the views consistent by scaling them to new inputs, which allows the user to see data referring to a given selection with few elements just as well as one with a larger amount.

It is safe to say that all idioms suffered various changes to better accommodate our purposes and data.

## CONCLUSION

In the end we believe we managed to do a decent job at responding to our initially proposed tasks, as well as, giving some extra insight into the industries.

Our first and third tasks heavily rely on the treemap to select the regions. On one hand we can compare how the region changes our bar chart, which gives us an indication of popularity. Secondly, we can view how the star plot updates in order to understand regional preferences when it comes to videogame genre.

The second task had 'controversy' as its main focus. The bar chart provides us with such insight, by visualizing how the distribution of the difference in opinions varies.

As for how the length title of a videogame or movie might affect its sales, we can simply look towards our scatterplot for a better understanding.

### Future Work

After looking at the visualization, it is clear that we explored a more "industry" level of the domains, preferring to compare both. An interesting proposal to enrich our solution would be to also allow a more individualistic look at elements within the industries, that is, to allow the user to select and compare individual titles of videogames and movies.

## REFERENCES


1. https://fenix.tecnico.ulisboa.pt/disciplinas/VI/2018-2019/1-semestre/hall-of-fame
2. https://towardsdatascience.com/using-tableau-to-visualize-the-list-of-greatest-video-games-from-wikipedia-eee804459f29
3. https://towardsdatascience.com/exploring-movie-data-with-interactive-visualizations-c22e8ce5f663
4. https://www.reddit.com/r/dataisbeautiful/
5. https://github.com/d3/d3/wiki/Gallery
6. https://www.kaggle.com/
7. https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings
8. https://www.kaggle.com/adktyakirloskar/movies
9. https://pandas.pydata.org/pandas-docs/stable/index.html
10. http://www.numpy.org/
11. https://www.metacritic.com/
12. https://d3js.org/
13. https://bl.ocks.org/mbostock/6bbb0a7ff7686b124d80
14. http://bl.ocks.org/tjdecke/5558084
15. http://bl.ocks.org/weiglemc/6185069
16. https://bl.ocks.org/aleereza/d2be3d62a09360a770b79f4e5527eea8
17. http://bl.ocks.org/nbremer/6506614