Group 1 Information Visualization Report

Tiago Nascimento
Instituto Superior Técnico
Alameda
70493

Miguel Cruz
Instituto Superior Técnico
Alameda
76102

Daniel Trindade
Instituto Superior Técnico
Alameda
76349

INTRODUCTION

How do the countries that participated in the Olympic Games stand against each other concerning the medals they achieved through the years? Do countries with a greater population also get more medals? How do these standings evolve over time and how do they accumulate in a certain amount of years?

We knew there was data to answer how many medals a country scored for a certain sport in a certain year, and we did find a solution that did that, on the Internet. But we wanted to go a bit further and be able to make comparisons, not just for one sport, not just for one year at a time, and not just counting one or all kinds of medals. So we went further and now we could know, as an example, if Russia had more or less gold and silver medals than the Soviet Union.

We also thought of seeing how many medals each country "owned". That is, for example, how many medals Germany had scored, plus medals Germans playing for other teams scored, minus the ones foreigners playing for Germany scored. Unfortunately, we couldn't find the nationalities of a big amount of athletes, so we decided to leave this feature alone.

The first tasks we proposed to support were, then:

- Browse – display the countries with the most gold medal in total in a given year.
- Identify – show the country with the most medals in a sport of all time.
- Locate – show the position of a country in the overall standings.
- Explore – using the coefficient medals/population (derivative variable), display the countries with the highest coefficient.
- Compare – show the medals each country won.

Our initial thought of showing statistics for "all time" was also changed to a "span of years", where we chose the minimum and maximum years, making our visualization more flexible.

RELATED WORK

When we found our main dataset containing the medal standings, we had only a vague idea of what to do: display bubbles for each country over a few rows, sorted by number of medals.

Then we came across a solution somewhat similar to what we ended up with:

http://www.nytimes.com/interactive/2008/08/04/sports/olympics/20080804_MEDALCOUNT_MAP.html?_r=0

We liked most of what we saw, but as we said before, we wanted to go that step further.

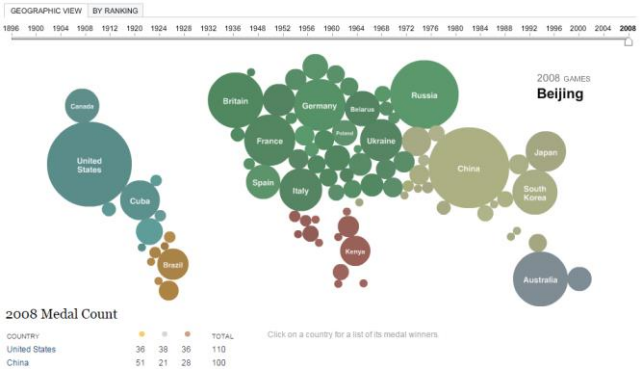

Figure 1 - The pretty but lacking solution we found

The main problems with this solution we found were a few and we wanted to correct most of them in our solution:

- It's difficult to find a country, even if we know its location in the world, since the bubbles are all jumbled together and are not related to a world map.
- It settled for only the amount of medals in a year and didn't allow for side-by-side comparisons.

We drew some inspiration from this solution's bubbles, and the rest of our solution came from our own thoughts.

THE DATA

Over a few days of research on the Internet, we found some documents and datasets that gave us the theme for our visualization and aid its development. The main dataset we found was on a The Guardian article:

http://www.theguardian.com/sport/datablog/2012/jun/25/olympic-medal-winner-list-data#data

At the bottom of the article, we found the data itself:

https://docs.google.com/spreadsheets/d/1zeeZQzFoHE2j_ZrqDkVJK9eF7OH1yvg75c8S-aBcxaU/edit

This was the basis for most of our decisions in the project. It had more information than we needed. What we ended up using of it were the All Medalists sheet (containing information for each medal achieved), and the IOC Country

Codes sheet (containing ISO codes and names for each country.

For the bubble chart we intended to draw over a world map, we needed another dataset with the coordinates for each country. For this, we opted to use this dataset provided by Google:

https://developers.google.com/public-data/docs/canonical/countries_csv

Since we wanted to check if a country's population had some relation with how many medals that country scored (a tendency), we looked for a dataset with population data for countries over the years. We found a dataset with what we wanted:

http://data.worldbank.org/indicator/SP.POP.TOTL?page=6.

It only provided data from 1960 until 2014, but we found those 13 editions in between to still give us a good idea of the tendency we wished to verify.

During the early development of the project we made several changes to the data we had and needed. Since we got our final dataset ready rather early, we could focus on the development of the visualization during the rest of our time.

Setting up and cleaning our data

From the All Medalists sheet, we managed to extract several pieces of data, grouped by arrangements of year, kind of medal, sport and country, using Pentaho. Some of those pieces of data (that we showed in the 2nd Checkpoint) were created as an experiment but never used because the arrangement and grouping of the attributes didn't suit our needs. We realized this only after the 2nd Checkpoint, which fortunately wasn't too late. Also shown in that Checkpoint were the only two pieces of data we used from then, which had the amount of each medal, in each sport that each country had in each year, and summed by country and year. Other attributes such as the gender of the medalist were filtered out for being irrelevant.

	A	B	C	D	E	F
1	Edition	Sport	NOC	numberBr	numberSi	numberGold
2	1896	Aquatics	AUT		1	1
3	1896	Aquatics	GRE	3	3	1
4	1896	Aquatics	HUN			2
5	1896	Athletics	AUS			2
6	1896	Athletics	FRA	1	1	

Figure 2 - Our main dataset in its early stages

Something else we had to do was tidy up the Country Codes table. Some of the country codes didn't match perfectly with the coordinates and population data sets' entries. We checked which codes were missing on those tables using Pentaho and we made matches.

We also faced another problem which was having countries that no longer existed having no coordinates assigned. We added extra entries on the Country Codes and Coordinates tables, so older countries could be displayed on the map, next to their more recent counterparts. An example of this is Russia, which has five bubbles over it on the bubble map: modern Russia, the Soviet Union, the Russian Empire and two agglomerate teams.

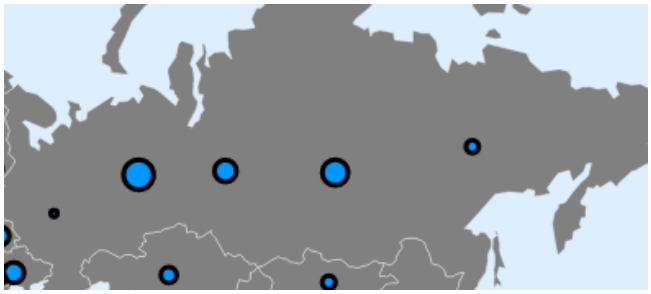

Figure 3 - All the teams related to Russia

Then, through some data joins and confusing normalizations and denormalizations made on Pentaho, we ended up with two almost final datasets. One containing the Olympic medals counts by country, year, sport, joined with the coordinates of that country. The second containing the population number and the total amount of medals by country and year.

Finally, we did simple mathematical operations on those datasets using Excel to obtain additional medal counts on one table and the coefficient on the other table.

This concludes all the work we had to do to have our final datasets.

As we will mention in the fifth section of this report, we ended up not having to worry about scalability issues, even if our datasets were really big, because runtime processing of the data was done much faster than we ever imagined.

VISUALIZATION

Overall Description

The first design we made was the one we stood by. We'd have 3 tabs, each with their own purpose: Standings, Standings Comparison, and Coefficient. We made improvements on them over time, and our final result was a similar but more complex interface.

All those tabs feature similar layouts: a top section for our input, including a timeline and search box, a lower section with the visualization itself. Entering that visualization section, to the top left we have a bubble chart over a map, encoding the number of medals or the coefficient; on the right, we have a bar chart, also encoding the number of medals; to the bottom left, we have a line chart also encoding the number of medals over time, with bubbles on top.

The first tab (Standings) also features a vertical sidebar on the left with all the Olympic sports, so we can select and highlight in a different color the sport for which data will be shown. On the input section, there are also three medal

shaped checkboxes over a podium that represent the medals we'd like to see counted. The timeline is a double slider bar, where we drag the thumbs to select a single year or an interval to show data for. Then, the search box will highlight in a different color a country's marks if they exist and will push that country's data into the line chart. All these inputs will change the information we're shown when we change their value. The idioms in the visualization section encode the number of medals under the specified circumstances: the more medals, the bigger the width of the bars and the further up the rank they are. For a bigger count of medals, the radius of the bubbles over the map also increases. The line chart encodes the number of medals vertically and that number over time horizontally, with the vertical scale varying according to the maximum number of medals.

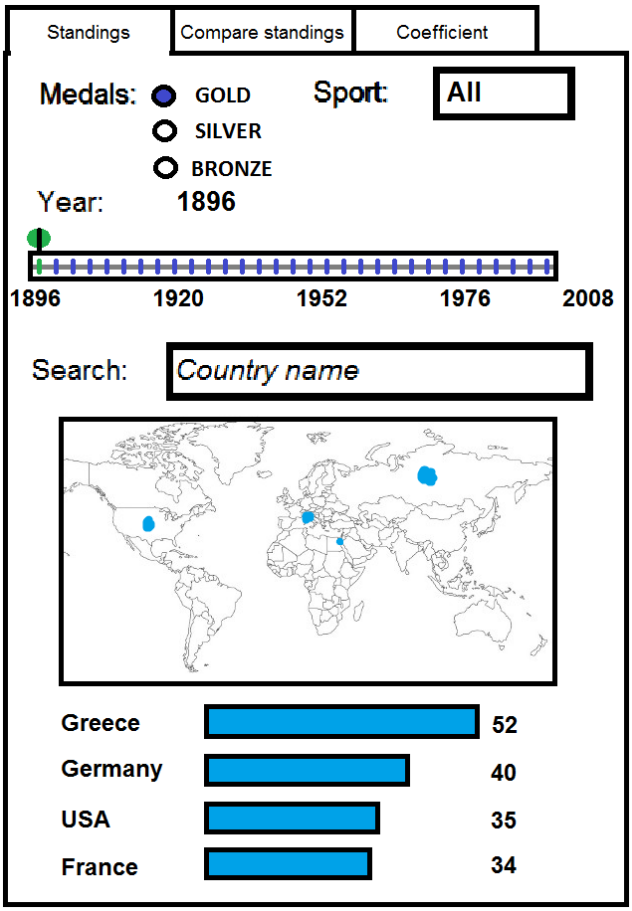

Figure 4 - One of our first prototypes, with the medals still as radio buttons (less flexible) and not checkboxes (more flexible)

The second tab (Standings Comparison) has a layout similar to the first tab's. Since we always compare the two countries concerning their medals in each sport, it's of no use to have the sports sidebar. Since we compare two countries at each time, we have two search text boxes. As for the output, the bubble chart remains the same highlighting the two countries, the bar chart displays the countries scores by sport side by side, and the line chart shows the two lines for the countries. Each country is highlighted in its own color, across the idioms.

The third tab (Coefficient) is also similar to the first tab in layout, but the idioms encode the coefficient of the countries. Since the coefficient is measured using the total of medals and the population, we aren't offered the choice of sport or medal. The timeline is also for only single years, because the population changes each year. The bars and bubbles are bigger for countries with a higher coefficient. The line chart works the same way as in the first tab, encoding coefficient vertically and change over time horizontally.

Figure 5 – Bar chart in the third tab in a mid-development stage, still with poor number format and width differences

In all the idioms there are tooltips popping up when we hover over marks, and displaying the name of the country, country code, number of medals and coefficient where appropriate.

In the first and third tabs, when we click on a country's bar or bubble, the marks on the other idioms are highlighted too, also changing the country for which data is shown on the line chart. If we search for a country, its bar and bubble will be highlighted and, as said before, the country displayed on the line chart changes too.

The bubble chart's map can be zoomed and dragged to see countries in greater detail. The size of the bubbles remains the same on the screen (semantic zoom) so they don't constantly obstruct the countries' borders.

Clicking on a bubble on the line chart will change the year of the visualization. The bubbles contained in the range of years or just the year selected will be filled, otherwise they will be empty.

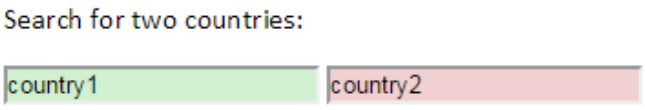

Figure 6 - The mid-development look of the search boxes in Tab 2, color coded according to the color of the marks of the countries

Additionally, we can see the data changing on the bubble chart and bar chart over the years, if we click on the play button. In the first and second tabs, the progression will start at the position of the Minimum year thumb and end automatically in 2008 unless we click the Play/Pause button again. In the third tab, the progression will start at the position of the single slider and end automatically in 2008 unless we click the Play/Pause button.

Rationale

Our choices were made aiming for the simplicity of the visualizations. We wanted to make sure the user understood all the information he saw, so a higher level of complexity could overwhelm them. We also attempted to choose the best colors for the various kinds of colorblindness; the shades of blue, red and green we used still seemed distinguishable.

One thing we had done initially was having text labels over the bubbles over the map and next to the bars. It worked well with the bars and we kept them, but the ones over the bubbles on the map made that idiom much more confusing and didn't have much of a positive impact.

Figure 7 - Labels over the bubbles could become messy if in great amount, so we removed them

To solve that, we had added text that popped up when we hovered over any bubble on the map, bar or bubble of a line chart. It worked well, but those took a while to pop up (since they used SVG primitives), so we made different ones that popped up immediately upon hovering.

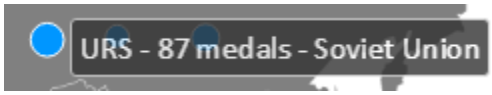

Figure 8 - The new tooltips we used

The map bubble chart was chosen because it was easy to find a country on it, if we knew its approximate location, and the radius of the bubble was intuitively associated with the number of medals. If they didn't know that location, they could use the search box instead. This is only an amuse, cultural and direct way of viewing the data in a World Map. Even if there are a lot of bubbles, we can zoom in and reduce the mess on the screen.

The bar chart was chosen for the same kind of simplicity concerning the width of the bars. Some text labels to the sides of the bars would help the user know what country it belongs to and the count of medals or coefficient. In the first and third tabs, the bars were sorted by the number of medals or coefficient, since what we'd like to see is a rank of the countries. In the second tab, we sorted the bars by sport, alphabetically, since it makes it easier to find the sport we'd like to see. Even though the bars are side by side and we can see bigger differences in size, smaller differences are problematic. The label showing the amount of medals solves that issue. A growing amount of bars doesn't affect the visualization too much since we can just scroll down for more bars.

Our final visualization – the line chart – was used because we wanted more than just seeing the information of a country in a certain year. Using the line chart, we could easily see the number of medals or the coefficient of the country over the years, which we found to be a useful and interesting feature. The line chart doesn't seem to get too filled with more entries.

Simplicity and effectiveness were important in the development of our visualizations and we think we achieved that (plus one Gold for Portugal, in 2015!). There could be room for improvement, of course, if we had an extra month and €3000.

Demonstrate the potential

So here are the five tasks we promised to give answers for, plus a few other special cases after.

What countries had the most gold medals in the first games, in 1896?

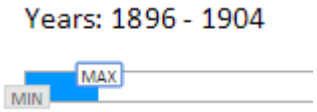

Figure 9 – In Tab 1, Standings, we first we drag the Min and Max sliders to 1896

Figure 10 - We select only the gold medal on the podium

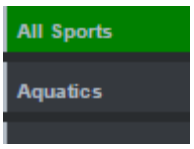

Figure 11 - On the sidebar, we select All Sports

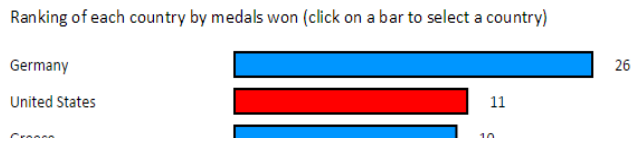

Figure 12 – It's Germany that had the most gold medals in 1896

What country has the most medals in Judo?

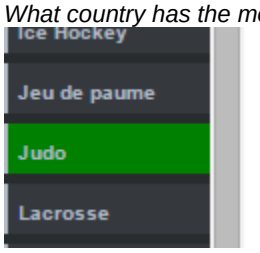

Figure 13 - First we can choose Judo (the order doesn't matter)

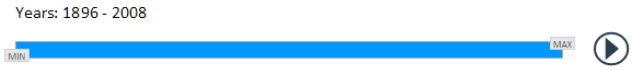

Figure 14 - We can drag both sliders to span over all the years now

Figure 15 - And then all the medals

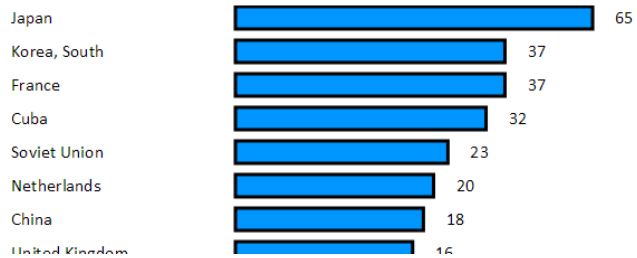

Figure 16 - Japan has the most Judo medals

What are the standings of the USSR in 1964?

Figure 17 - We drag both sliders to 1964

Figure 18 – We've seen how to choose all the medals and sports, so we did that. We know where the Soviet Union was centered so we can zoom and click on it to highlight the bubble and the corresponding bar

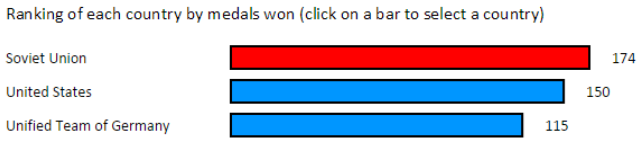

Figure 19 – The Soviet Union's bubble and bar turned red and we can see it was at the top of the ranking in that year, closely followed by the USA

See the countries with the most medalists per capita in 2008.

Figure 20 - In Tab 3, Coefficient, we can begin by dragging the slider to 2008

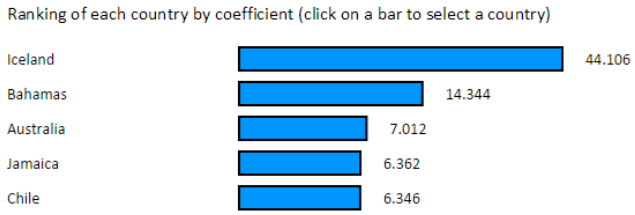

Figure 21 - Iceland had 44 million medallists per capita. That's more than how many popes there are per square km, in the Vatican City (2.3)

How do the USSR and Russia's cumulative scores compare?

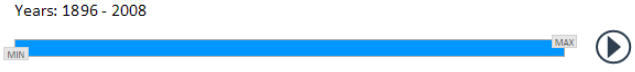

Figure 22 - In Tab 2, Standings Comparison, we can start by dragging the sliders across all years

Figure 23 - Again, we select all the medals

Figure 24 - We search for the two countries in the boxes

Figure 25 - And now we can see on the map the bubbles of the two compared countries

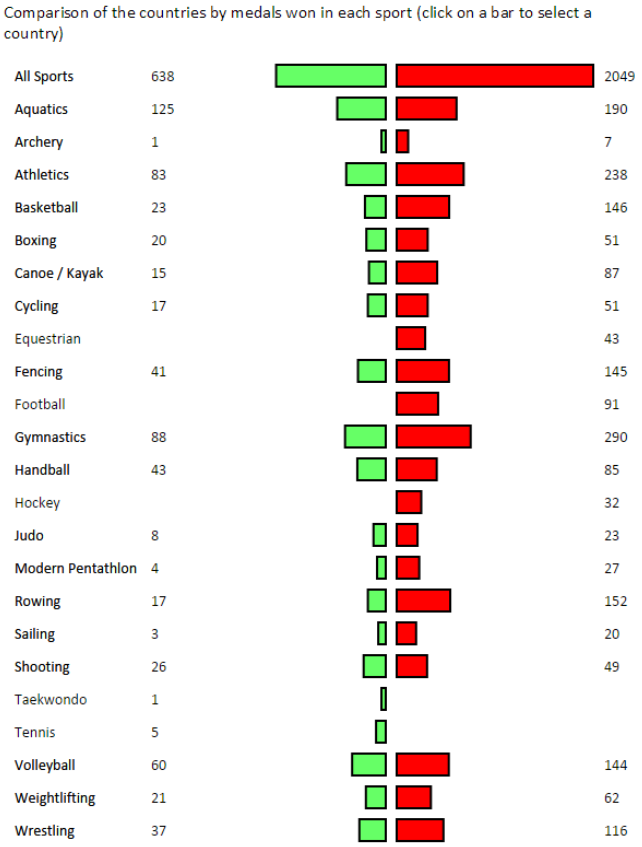

Figure 26 - To the right, the Soviet Union, to the left, Russia. The Soviet Union had way more medals overall and in every sport

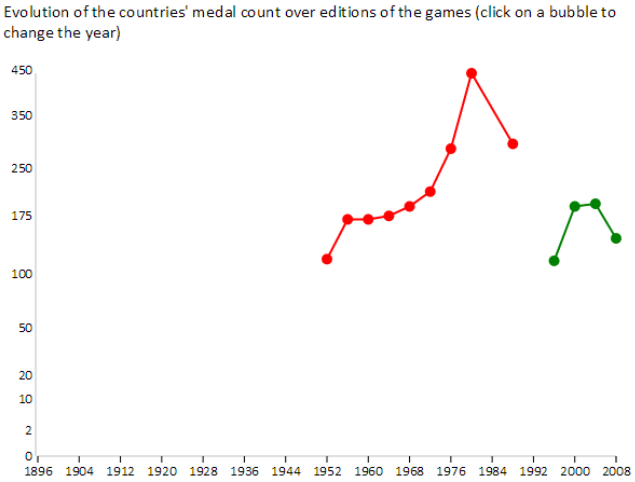

Figure 27 - And on the line graph, how the medal winnings change through the years

Additional 1 – See the distribution of silver medals on the map in 1940.

Figure 28 - In Tab 1, Standings, we select the silver medal. We've already seen how to select a single year and all sports.

Figure 29 - Oh no! The world was at war in 1940 so there were no games

Additional 2 – See how Zimbabwe's overall medals compare roughly with surrounding countries.

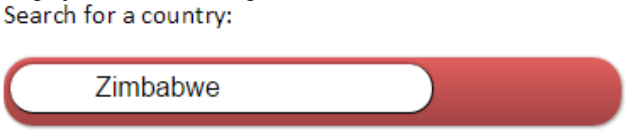

Figure 30 - After selecting All Sports, every medal and years from 1896 to 2008, we just don't know where Zimbabwe is, so we can search for it

Figure 31 - There's Zimbabwe, in red, being an average country concerning the medals won

Additional 3 – See the evolution of Portugal's coefficient over the years.

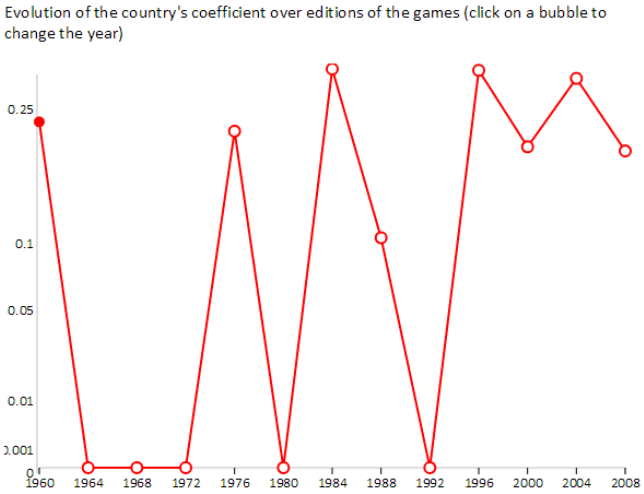

Figure 32 – After searching for Portugal, we can see the line graph pertaining to Portugal: very disappointing

IMPLEMENTATION DETAILS

The implementation of the most complicated aspects of the visualization weren't in fact ridiculously difficult.

The first challenge was choosing the scales for the widths of the bars, radii of the bubbles and coordinates of the lines. We had to choose values that would display the number of medals and coefficient clearly but without great disparity between smaller and larger values. After that, we had to choose the best tick values for the line graph's Y axis.

The second challenge was being able to sum the medals of all sports for a country in a certain year, and also adding all the medals of the same sport over the years. For both we had to use a few iterative cycles nested by one another and sum all the medals.

The third challenge was making the connection between idioms. The HTML and SVG elements we used had some attributes (their "id" containing the country code) that made matching bubbles and bars easy. The hard part was coding the logic to highlight the marks for a chosen country and stop highlighting the previously chosen country's marks. For that we had to save the name of the previous country.

The fourth challenge was implementing the range sliders for the timeline and the animation that progresses through the years. For the range timeline, we used a J-Query solution we found online and modified for our fitting. Then we had to figure out a cycle that would increase the years and update the visualization and the sliders.

The fifth challenge was making the tooltips over the bubbles and bars. We initially used a graphical element that would pop up over a mark after one second, but then opted for a simple solution we found online that popped up immediately after hovering over the mark.

The sixth challenge was making the drawing of the map more efficient. Initially we had it being drawn every time we changed the bubbles, which was the most straining thing in our visualization. We had to find a way around D3's logic to allow a one-time drawing of the map bug free.

What did not end up being a challenge was optimization. We expected the sum of medals over the years and sports to be a very slow process. We'd have to do it every time we selected a range of years or to show stats for all sports. But that sum process was in fact rather quick under any circumstances, and did not spoil the experience. Thus, no optimization was needed on that.

In the end, most of the problems we had arose from not knowing the JavaScript language's quirks well enough, initially. We made all the idioms ourselves without ever using D3's examples page.

CONCLUSION & FUTURE WORK

Having finished our visualization with all that we had planned, and even during the development, we learned a few things, some related to the data we displayed, some unrelated to that, as well as. Some extra features beyond the tasks we first committed to answer to were also added to enrich our visualization, and those taught us a few more things.

Things we learned

We didn't know there hadn't been any games during the World Wars. We had to display an image instead of data, when we found out there were no bubbles or bars during those years.

We learned that computers are really fast. We had a few thousands of entries to process during runtime, and our

code just skimmed through them when the time came to pump out graphical elements.

We learned that the Soviet Union achieved three times as many medals as modern Russia has, in the span of only around twice the editions of the Olympics.

We found out about obscure games such as Jeu de Paume, Tug of War and Basque Pelota where only a couple countries got medals. Jeu de Paume is a French game but none of the medalists were from a French team.

Things to improve on

If we were to start over, we would have skipped the useless data processing experiments we did and gone straight to the data that mattered. But experimenting was probably a useful step to help us figure out how each arrangement of data is best for what we want to draw.

Furthermore, we would have organized our code better from the beginning, to facilitate the increments we made throughout the development. For example, the tooltips weren't something we had originally planned and they required some dark magic using SVG, HTML and JavaScript in ways they should probably not be mixed. The animation through the years also required some structure changes in the code, and we still had to correct some small bugs.

If we had more resources to work on our visualization, we'd like to first perform a small style overhaul to the interface and general look of our solution. We would also allow for a choice of an arbitrary amount of sports, instead of forcing the user between choosing either one or all of them. We would also attempt to merge the two tabs Standings and Standings Comparison into one, after we figured out how to organize the views on the page. We'd like to improve the general feel of the zooming and panning of the bubble chart map.

We'll finish off by saying this was enriching project that went as we planned, in some aspects better than we expected, that we would like to do some further improvements on, and that taught us skills in building visualizations as well as facts about the Olympic Games.