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 medalists in total in a given year.
* Identify – show the country with the most medalists 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.

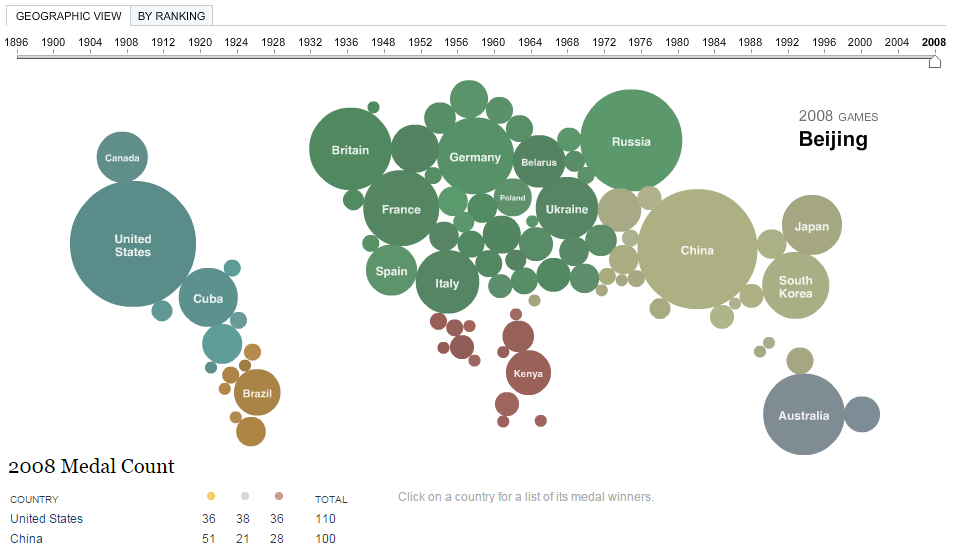


Figure - The pretty but lacking solution we found

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.

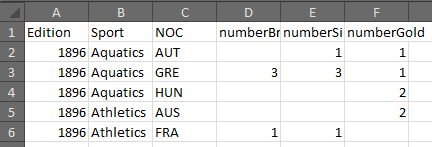


Figure - Our main dataset in its early stages

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 - 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 checkboxes 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.

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.

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.

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.

# 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’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.



Figure 4. Sample of a wide figure. Be sure to place at the top or bottom of the page. Ensure that important information is legible in both black-and-white and color printing. Image: CC-BY-ND ayman on Flickr.

## Table Style

The text of tables will format better if you use the Table Text style (as in Table 1). If you do not use this style, then you may want to adjust the vertical spacing of the text in the tables. To adjust the spacing of text in a table in Word, use Home | Paragraph | Indents and Spacing. Generally, text in each field of a table will look better if it has equal amounts of spacing above and below it, as in Table 1. Table captions should be placed below the table. We recommend table lines be 1 point, 25% black. Minimize use of unnecessary table lines.

For improved accessibility, header rows of tables should be marked. In Word, right-click a header row, and select Table Properties | Row | Repeat as header…

## Subsequent Pages

On pages beyond the first, start at the top of the page and continue in double-column format. The two columns on the last page should be of approximately equal length.



Figure 5. Use high-resolution images, 300+ dpi, legible if printed in color or black-and-white. Number all figures and include captions below, using Insert, Caption.

## References and Citations

Use a numbered list of references at the end of the article, ordered alphabetically by last name of first author, and referenced by numbers in brackets [1,3,4].

| Objects | **Caption – pre-2002** | **Caption – 2003 and afterwards** |
| --- | --- | --- |
| Tables | Above | Below |
| Figures | Below | Below |

Table . Table captions should be placed below the table. We recommend table lines be 1 point, 25% black. Minimize use of unnecessary table lines.

Your references should be published materials accessible to the public. Internal technical reports may be cited only if they are easily accessible (i.e., you provide the address for obtaining the report within your citation) and may be obtained by any reader for a nominal fee. Proprietary information may not be cited. Private communications should be acknowledged in the main text, not referenced (e.g., “[Borriello, personal communication]”).

# LANGUAGE, STYLE AND CONTENT

The written and spoken language of SIGCHI is English. Spelling and punctuation may use any dialect of English (e.g., British, Canadian, US, etc.) provided this is done consistently. Hyphenation is optional. To ensure suitability for an international audience, please:

* Write in a straightforward style.
* Try to avoid long or complex sentence structures.
* Use common and basic vocabulary (e.g., use the word “unusual” rather than the word “arcane”).
* Briefly define or explain all technical terms that may be unfamiliar to readers.
* Explain all acronyms the first time they are used in your text—e.g., “Digital Signal Processing (DSP)”.
* Explain local references (e.g., not everyone knows all city names in a particular country).
* Explain “insider” comments. Ensure that your whole audience understands any reference whose meaning you do not describe (e.g., do not assume that everyone has used an Android phone, or a particular application).
* Explain colloquial language and puns. Understanding phrases like “red herring” may require a local knowledge of English. Humor and irony are difficult to translate.
* Use unambiguous forms for culturally localized concepts, such as times, dates, currencies, and numbers (e.g., “1-5- 97” or “5/1/97” may mean 5 January or 1 May, and “seven o’clock” may mean 7:00 am or 19:00). For currencies, indicate equivalences: “Participants were paid ₩22, or roughly US$29.”
* Be careful with the use of gender-specific pronouns (*he*, *she*) and other gendered words (*chairman*, *manpower*, *man-months*). Use inclusive language that is gender-neutral (e.g., *she* *or* *he*, *they*, *s/he*, *chair*, *staff*, *staff-hours*, *person-years*). See the *Guidelines for Bias-Free Writing* for further advice and examples regarding gender and other personal attributes [9]. Be particularly aware of considerations around writing about people with disabilities.
* If possible, use the full (extended) alphabetic character set for names of persons, institutions, and places (e.g., Grønbæk, Lafreniére, Sánchez, Nguyễn, Universität, Weißenbach, Züllighoven, Århus, etc.). These characters are already included in most versions and variants of Times, Helvetica, and Arial fonts.

# Accessibility

The Executive Council of SIGCHI has committed to making SIGCHI conferences more inclusive for researchers, practitioners, and educators with disabilities. As a part of this goal, the all authors are asked to work on improving the accessibility of their submissions. Specifically, we encourage authors to carry out the following five steps:

1. Add alternative text to all figures
2. Mark table headings
3. Generate a tagged PDF
4. Verify the default language
5. Set the tab order to “Use Document Structure”

For more information and links to instructions and resources, please see: <http://chi2016.acm.org/accessibility>.

# Page Numbering, Headers, and Footers

Your final submission should not contain footer or header information at the top or bottom of each page. Specifically, your final submission should not include page numbers. Initial submissions may include page numbers, but these *must* be removed for camera-ready. Page numbers will be added to the PDF when the proceedings are assembled.

# Producing and testing PDF files

We recommend that you produce a PDF version of your submission well before the final deadline. Your PDF file must be ACM DL Compliant. The requirements for an ACM Compliant PDF are available at:

<http://sheridanprinting.com/typedept/ACM-distilling-settings.htm>

When creating your PDF from Word, ensure that you generate a tagged PDF from improved accessibility. This can be done by using the Adobe PDF add-in, also called PDFMaker. Select Acrobat | Preferences from the ribbon and ensure that “Enable Accessibility and Reflow with tagged Adobe PDF” is selected. You can then generate a tagged PDF by selecting “Create PDF” from the Acrobat ribbon. Test your PDF file by viewing or printing it with the same software the publisher will use, Adobe Acrobat Reader Version 10, which is widely available at no cost. Note that most reviewers will use a North American/European version of Acrobat Reader, so please check your PDF accordingly.

# Conclusion

It is important that you write for the SIGCHI audience. Please read previous years’ proceedings to understand the writing style and conventions that successful authors have used. State clearly what you have done, not merely what you plan to do, and explain how your work is different from previously published work, i.e., *the unique contribution that your work makes to the field*. Please consider what the reader will learn from your submission, and how they will find your work useful. If you write with these questions in mind, your work is more likely to be successful, both in being accepted into the conference, and in influencing the work of our field.

# ACKNOWLEDGMENTS

Sample text: We thank all the volunteers, and all publications support and staff, who wrote and provided helpful comments on previous versions of this document. Authors 1, 2, and 3 gratefully acknowledge the grant from NSF (#1234-2012-ABC). This is just an example.

# References format

References must be the same font size as other body text. References should be in alphabetical order by last name of first author. Example reference formatting for individual journal articles [3], articles in conference proceedings [7], books [9], theses [10], book chapters [11], an entire journal issue [6], websites [1,4], tweets [1], patents [5], and online videos [8] is given here. This formatting is a slightly edited version of the format automatically generated by the ACM Digital Library (http://dl.acm.org) as “ACM Ref”. More details of reference formatting are available at:

<http://www.acm.org/publications/submissions/latex_style>

Note that the Hyperlink style used throughout this document uses blue links; however, URLs that appear in the references section may appear in black.

# REFERENCES

1. @\_CHINOSAUR. 2014. VENUE IS TOO COLD. #BINGO #CHI2016. Tweet. (1 May, 2014). Retrieved February 2, 2014 from https://twitter.com/\_CHINOSAUR/status/461864317415989248
2. ACM. How to Classify Works Using ACM’s Computing Classification System. 2014. Retrieved August 22, 2014 from [http://www.acm.org/class/how\_to\_use.html](http://www.acm.org/class/how_to_use.html%20)
3. Ronald E. Anderson. 1992. Social impacts of computing: Codes of professional ethics. *Soc Sci Comput Rev* 10, 2: 453-469.
4. Anna Cavender, Shari Trewin, Vicki Hanson. 2014. Accessible Writing Guide. Retrieved August 22, 2014 from <http://www.sigaccess.org/welcome-to-sigaccess/resources/accessible-writing-guide/>
5. Morton L. Heilig. 1962. Sensorama Simulator, U.S. Patent 3,050,870, Filed January 10, 1961, issued August 28, 1962.
6. Jofish Kaye and Paul Dourish. 2014. Special issue on science fiction and ubiquitous computing. *Personal Ubiquitous Comput*. 18, 4 (April 2014), 765-766. <http://dx.doi.org/10.1007/s00779-014-0773-4>
7. Scott R. Klemmer, Michael Thomsen, Ethan Phelps-Goodman, Robert Lee, and James A. Landay. 2002. Where do web sites come from?: capturing and interacting with design history. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '02), 1-8. <http://doi.acm.org/10.1145/503376.503378>
8. Psy. 2012. Gangnam Style. Video. (15 July 2012.). Retrieved August 22, 2014 from <https://www.youtube.com/watch?v=9bZkp7q19f0>
9. Marilyn Schwartz. 1995. *Guidelines for Bias-Free Writing.* Indiana University Press.
10. Ivan E. Sutherland. 1963. *Sketchpad, a Man-Machine Graphical Communication System*. Ph.D Dissertation. Massachusetts Institute of Technology, Cambridge, MA.
11. Langdon Winner. 1999. Do artifacts have politics? In *The Social Shaping of Technology* (2nd. ed.), Donald MacKenzie and Judy Wajcman (eds.). Open University Press, Buckingham, UK, 28-40.