Process Book for Milestone 3

Interactive Data Visualization Project

Data Visualization COM-480

Vizionary

David Gauch — 394014 — david.gauch@epfl.ch

Flavia Wallenhorst — 264996 — flavia.wallenhorst@epfl.ch

Arthur Wuhrmann — 344752 — arthur.wuhrmann@epfl.ch

May 30, 2025

Contents

1 Introduction

As part of the Data Visualization course taught by Professor Laurent Gilles Marie Vuillon, our objective was to provide insights on a topic of our choice through web-based data visualizations. We chose to explore the theme of the Olympic Games, focusing on the relationship between athlete attributes and their impact on the likelihood of winning Olympic medals. Our target audience includes sports analysts, data enthusiasts, and Olympic fans. Through our web-based interface, users can explore trends in athlete physiques and medal outcomes over time and across disciplines based on a broad Olympic athletes dataset from Kaggle.

2 Original Idea and Concept

Our initial idea was to focus on the origin of the athletes, specifically the countries they represent. In this version of the project, we planned to explore how an athlete's country of representation might influence the sport they compete in, as well as their likelihood of winning an Olympic medal.

This analysis could have been enriched by examining whether factors such as a country's economic status and sports infrastructures significantly impact both the number of athletes participating in the Olympics and the proportion of those winning medals.The second idea, which is the one we chose, explores the impact of athletes' physical traits on their chances of winning an Olympic medal.

We found this concept more creative, less straightforward, and potentially more engaging for users. While the other idea was also very interesting, we were unsure about the availability and quality of data regarding athletes' origins and the sports infrastructure in different countries. In fact, an athlete's actual origin might be more meaningful than the country they compete for, but this information could be difficult to obtain.

During Milestone 2, we sketched a three-tier bubble navigation (Figure 1):

1. **Games page (Homepage)**: one bubble per Olympic edition.
2. **Event page**: bubbles for each sport within the selected edition.
3. **Sport Detail page**: detailed visualizations for the chosen sport.

During implementation, we noticed that filtering in this way drastically reduced the amount of data, especially for sports that were included in early Olympic editions but have since been removed. Furthermore, by separating the data in this way, we were restricted in how we could represent the evolution of the data over time, as we always had only the data of one sport in one event.

Therefore, in Milestone 3, we decided to remove the Games page and directly arrive at the Discipline page, where clicking a bubble opens the sport's analytics directly. With this approach, we had much more data available, as for every sport we now have data for all athletes across all events. We can thus also include graphs that contain temporal aspects, providing more intuitive and interactive visualizations.

3 Design and Development Process

3.1 Data and Data Visualization

Our project used a broad Olympic athletes dataset from Kaggle, containing over 200'000 records from Olympic Games held between 1896 and 2016. Each record includes athlete names, countries, sports, events, and physical characteristics (age, height, weight). Although a small percentage of values for age (5%), height (10%), and weight (12%) are missing, the dataset had sufficient data to provide insightful information.

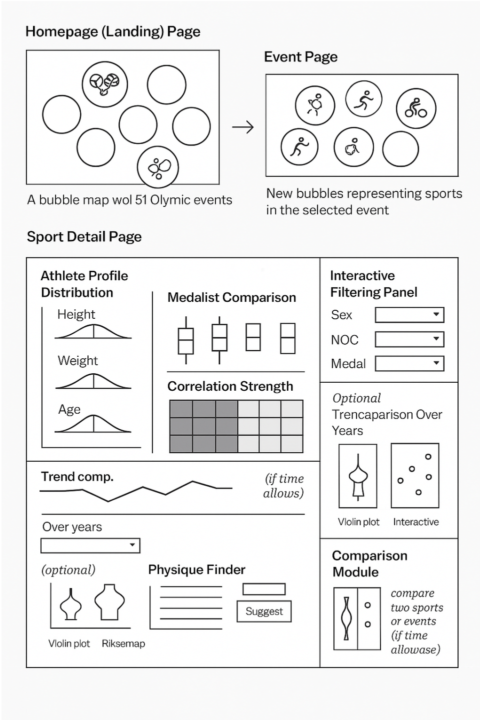

Figure 1: Milestone 2 sketch: Games → Disciplines → Analytics

As a first step, we conducted exploratory analyses to assess whether there was any potential correlation between athletes' physical attributes and their chances of winning an Olympic medal. To do this, we compared the distributions of age, weight, and height between medalists and non-medalists. In certain sports, we observed notable differences, which motivated us to explore further. Specifically, we were motivated to understand which attributes had an influence in which sports, how these patterns varied over time, and what insights could be drawn from these trends.

Once this analysis was completed, we prepared our data for the later interactive visualizations on the website.

We realized that comparing physical attributes across different genders didn't make much sense, as the distributions and standards vary significantly between male and female athletes. To address this, we added gender information to our JSON data and came up with the idea of allowing users to filter visualizations by gender across the entire website. This approach not only enables more meaningful comparisons, but also aligns with our broader goal of allowing users to explore trends over time, making it possible to analyze how physical profiles and medal distributions have evolved for each gender throughout the years.

Below is a small excerpt of the structure of these sport-specific files:


```
{
  "sport": "Diving",
  "heatmap": {
    "M": {
      "age": {
        "year": [1906, 1908, 1912],
        "corr": [-0.056, 0.05, -0.04]
      }
    }
  }
```

```
}
```

Upon visualizing all our charts for the first time, we noticed that some disciplines lacked sufficient data, leading to poor-quality visualizations. We also realized that, while the interactive bubble charts were fun to explore, they became overwhelming when too many elements were displayed at once. To improve clarity and user experience, we decided to limit our dataset to the 40 sports with the highest number of athletes since 1900.

We also realized that directly presenting the impact of physical attributes was a bit too straightforward. To give users a broader perspective on the Olympic context, we decided to include more general information, such as the number of medals and athletes per country. To represent this data, we chose to use a map and a treemap visualization.

For the map, we used the `countries-110m.json` file, a topographic-style JSON format that allows for geographic representation. This introduced a new challenge: aligning our main dataset with the map data using country names. To resolve this, we used the NOC-to-country dictionary provided in the Kaggle repository.

Later, we also encountered some issues with the treemap. While it's an interesting visualization method, it quickly became cluttered when displaying many countries. To address this, we implemented a threshold to show only the most significant countries individually, while grouping the others under a "Rest of the World" category.

To give a sense of geographic distribution, we highlighted countries from the same continent using the same color. To achieve this, we generated a JSON file containing pairs of countries and their corresponding continents, based on the following dataset: `https://www.kaggle.com/datasets/statchaitya/country-to-continent/data`.

3.2 User interface

We prioritized clarity, consistency, and responsiveness in our visualization design. Our main style choices were:

- **Home Page**: The initial page displays interactive bubbles that move dynamically. The size of each bubble reflects the amount of participant data we have for that sport, the larger the bubble, the more data it represents. By hovering over a bubble, users can see the exact number of participants.
- **Color palette**: On the homepage, we used the color palette *d3.schemeTableau10*. For the graphs, we opted for a concise and consistent color scheme. When appropriate, we used medal colors in the graphs to make them intuitive for users. For instance, in bar charts that represent the average height of gold medalists for a given year, the bars are colored gold.
- **Typography**: We used Google's *Fira Mono* font for body text and bold sans-serif fonts for headings to ensure readability and visual hierarchy.
- **Layout**: For the detailed analytics page, we adopted a grid structure where each graph occupies the same position across different sports. This consistent layout makes it easier for users to find the information they're looking for, as similar data is always located in the same spot across different sports.
- **Interactions**: Most visualizations include hover interactions, both from their title which provides additional information and from their content. Additionally, all of them can be customized according to user preferences such as gender, type of medal, and time period thanks to a dashboard containing toggles and radio buttons.
- **Navigation**: Most of the navigation is done by accessing pages from the homepage and returning using the back arrow. However, it is also possible to navigate directly between related visualizations, from the similar sports.

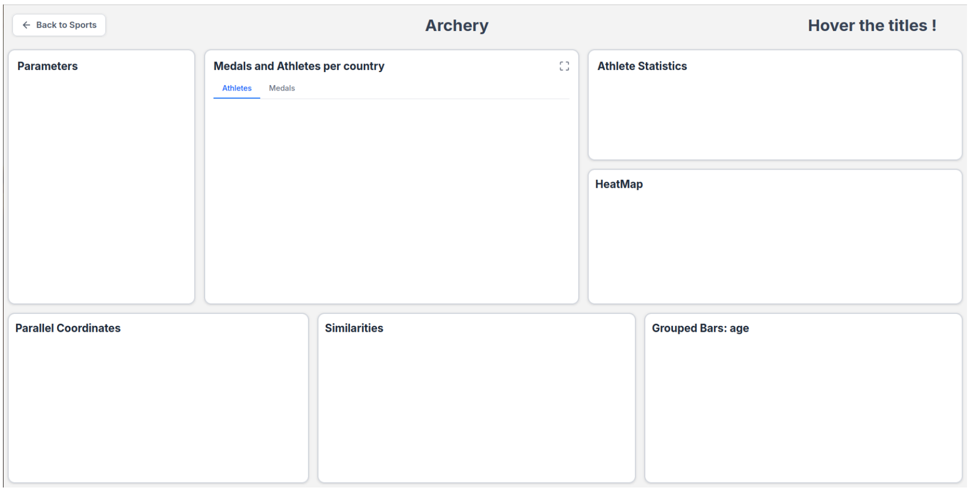

Figure 2: Grid structure of the sport analytics page

3.3 Implementation

We developed the front-end using **Svelte** as a modern reactive framework and styled it with **Tailwind CSS**. All data visualizations were built with **D3.js**, using its low-level APIs for precise and flexible control over SVG rendering.

- **Isolated D3 rendering:** Instead of integrating D3 directly into Svelte's reactive system, we encapsulated D3 code in standalone functions that operate only on specific static SVG containers passed from Svelte components.
- **Direct DOM control:** D3 fully manages SVG element creation, scales, axes, and data updates within these containers, avoiding interference with Svelte's virtual DOM diffing and updates.
- **Reactivity:** In order to make the plots change when changing parameters, we use Svelte 5 powerful runes that transmit the parameters directly to the components, without avoididng unnecessary reloads.

A significant challenge was reconciling D3's imperative DOM manipulations with Svelte's reactive, declarative updates. For example, in our heatmap rendering function, D3 directly creates and updates SVG elements, which can conflict with Svelte's tendency to control the size of the plots. This led to plots overflowing or poor resizing behavior. On top of this technical hurdle, we faced data sparsity issues, especially in the early Games and in niche sports. Some sports had so few athletes or medal records that meaningful visualizations became difficult.

To address the technical conflicts, we confined D3's logic to fixed SVG containers managed by Svelte but left entirely under D3's control. This clear separation of responsibilities let Svelte handle state updates and interface reactivity, while D3 focused solely on precise, low-level visual rendering. As for the data sparsity, we streamlined our dataset by limiting our scope to the top 40 sports with the most participants with this our data became more consistent and the dataset more robust. These careful choices helped ensure that the visualizations remained both meaningful and reliable.

4 Final Product

The final website includes the following components:

- **Home Page:** Users can click on interactive bubbles to navigate to detailed sport pages. Each bubble's size is proportional to the amount of athlete data available for that sport (see Figure 4).
- **Pages by disciplines:** These pages provide in-depth insights into the physical attributes of athletes in a specific sport and how they influence the chance of winning medals, accompanied by several informative graphs. They are located at `/[sport_name]`.
- **Parameters:** a dashboard containing toggles and radio buttons to customize the visualization (Figure 3).
- **Basic sport info:** Displays general statistics about each sport. (Figure 3)
- **World maps:** A dynamic world map highlighting countries that have won medals (Figure 3).
- **Treemaps:** A treemap representing the number of participants per countries (Figure 5).
- **Heatmaps:** Visualize the correlations between height, weight, age, and medal status across different years. (Figure 3)
- **Parallel coordinates:** Allow users to compare multiple physical attributes simultaneously for deeper analysis. (Figure 3)
- **Similar sports:** Recommends other sports that show similar data patterns and correlations regarding physical attribute importance. (Figure 3)
- **Bar charts:** Show average age, height, and weight per Olympic edition, with interactive checkboxes to filter by medal status (e.g., only gold medalists). (Figure 3)

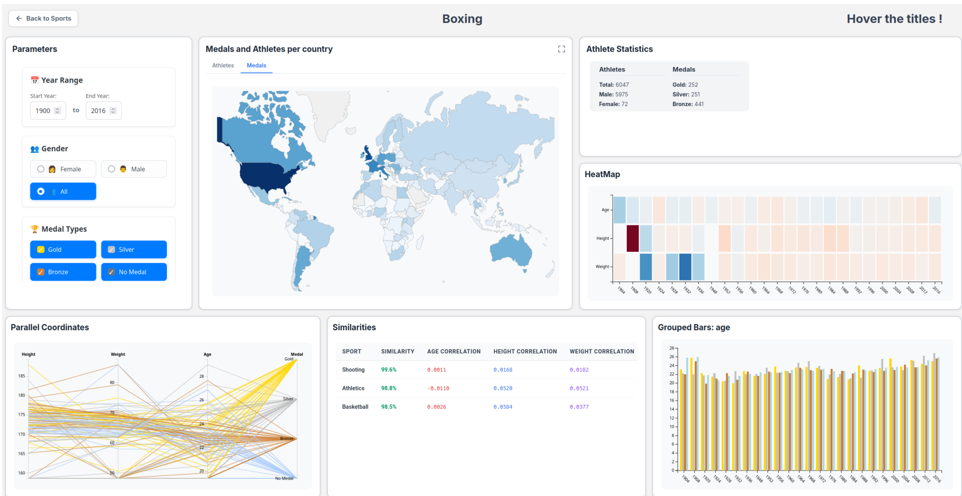

Figure 3: Dashboard of the website, with the tab on world map (see Figure 5 for the treemap, and Figure 4 for the homepage).

5 Reflections and Future Work

This project underscored the importance of carefully balancing creative design with technical rigor. One of our main lessons was the value of clear separation of concerns between D3 and Svelte. By keeping D3's rendering logic isolated in dedicated containers, we maintained smooth interactions and precise visual updates without compromising Svelte's reactivity. From a learning perspective, we also discovered how websites can be significantly enhanced through interactive data integration, using JSON files and appropriate visualization libraries.

Looking forward, we see promising directions for extending this work:

- **Live data feeds:** Integrate real-time updates for upcoming Olympic Games.
- **Event-level drilldowns:** Provide more granular comparisons between finals, qualifiers, and other rounds.
- **Country-level data:** Enrich insights by including economic and demographic factors via APIs.
- **Predictive modeling:** Add models (e.g., logistic regression) to forecast medal probabilities.
- **Interactive avatars:** Implement the avatar creation idea from our milestone 1 where users could input their physical attributes (age, weight, height), watch the avatar's physique change in real time, and discover which Olympic sports might best fit them. This feature would make the data even more relatable and engaging.

Overall, the project's interactive visualizations reveal how athlete attributes relate to Olympic success, while opening the door to new and exciting applications of these insights.

6 Peer Assessment

David Gauch (394014) Data sourcing/cleaning, Sport Bubble page, Heatmap, Bar chart and process book.

Flavia Wallenhorst (264996) Data sourcing/cleaning, Map, Treemap, Screencasts scripting/editing and process book review.

Arthur Wuhrmann (344752) Parallel coordinates, similarities, Svelte integration and website reactivity.

7 Appendix

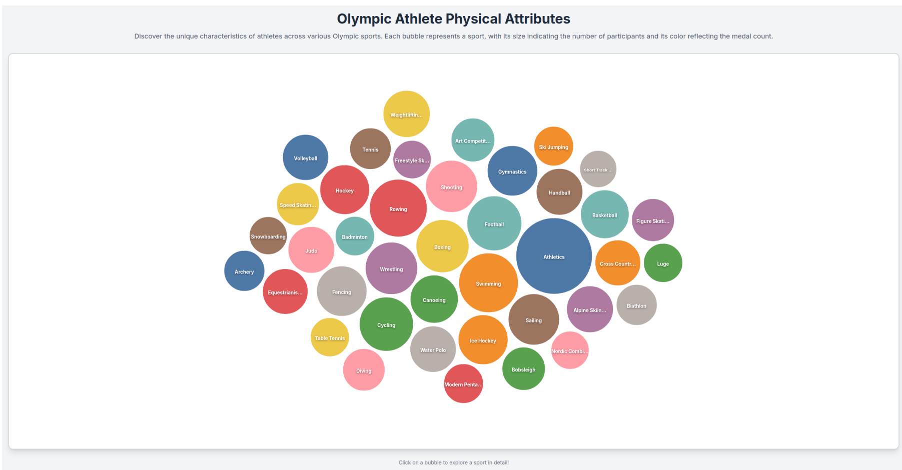

Figure 4: Main page.

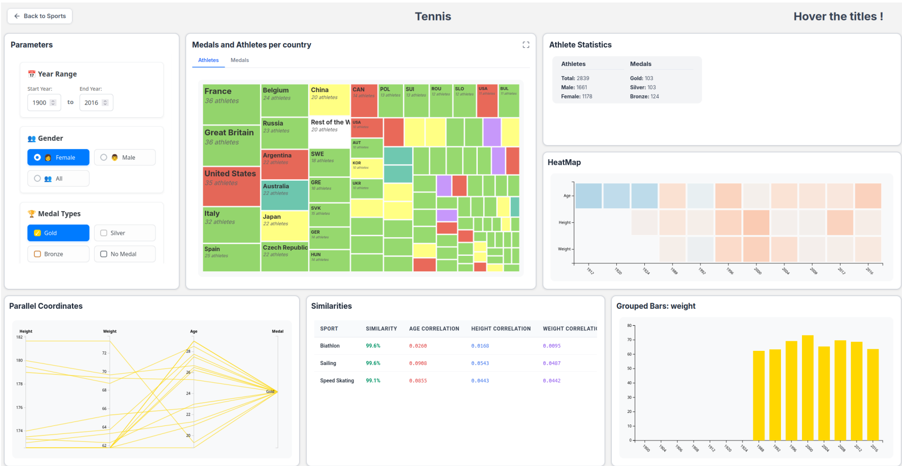

Figure 5: Website dashboards with a different sport and other parameters, also showing the threemap this time.

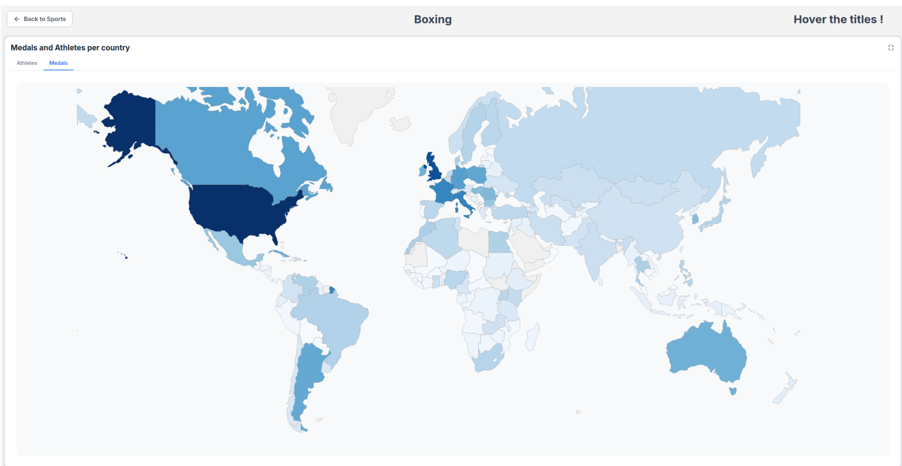

Figure 6: Full website, when zooming on the map.