PROCESS BOOK

SWISS RACES VISUALIZATION TOOL

STEFANO SAVARÈ
ANASTASIA TKACH

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

COM-480: DATA VISUALIZATION
ECOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

December 2017

1. INTRODUCTION

We explore a dataset of 16 years of Swiss races with over 1M runners.

1.1 MOTIVATION

This visualization is developed for inspiring people to participate in marathons. The idea is to motivate by example, that is by showing millions of Swiss marathon runners and their progress on the trails of Lausanne, Zurich, Luzern, Zermatt, Winterthur and many other Swiss cities.

1.2 TARGET AUDIENCE

The intended use case is for a person considering to run a marathon to be able to see the progress of runners from the past. The user has an opportunity of examining the runners moving on familiar trails and see their gender, age and experience. The dataset contains the runners of both genders, aged from 9 to 90, people running for the first and for the 14-th time. Thus, every person gathering courage to run a marathon can see runners just like them successfully finishing it.

1.3 RELATED WORK AND INSPIRATION

This project is inspired by a related project Hop Swiss implemented in framework of Data Analysis course at EPFL by Stefano Savarè and four other EPFL students.

2. DATASET

We use the same dataset of the Hop Suisse project. We parsed the runner data directly from the datasport.com website. We collected more than 1.5 million entries from 2000 to 2015 in 223 distinct races. We focus in this project on 3 well-recognizable distances: 10km, half marathon and marathon.
We do not want to go deep into details of the parsing process, since it is not the aim of this project. Each entry in the final dataset has the following parameters:

- Race information: i.e. race name, race date, distance, race ID, temperature and weather
- User information: i.e. name, gender, year of birth, user ID.
- Performance: i.e. time, pace final rank and category.

2.1 EXPLORATORY DATA ANALYSIS

The materials from Data Analysis project allowed us to examine the following visualizations of the dataset:

- Number of runners per year classified by gender.
- Number of runners per year classified by distance of the race.
- Number of races with respect to the number of times these races took place.
- The dependency of the time to finish the race on the runner's age.
- Overall age distribution of the runners.
- Age distribution of the runners by gender and by category.
- Towns of residence of the runners.

3. DESIGN

3.1 DESIGNING VISUALIZATION

DATA ABSTRACTION

The input dataset is a historic record of Swiss races. For each race a GPS track a list of participants and their attributes are available.
The data has geometric (spatial) component – the tracks of the races. At the same time, the data has temporal component, that is the races can be simulated in time.
To create an intuitive and natural visualization, we use these two components: we display the data on the map and "replay" the race in time.

TASK ABSTRACTION

The intended use of our visualization is *consuming*, as opposed to *producing* the data. We aim to enable the user to *explore* the dataset. For this purpose, we provide the user with controls in form of *classifiers* and *filters* as well as plot a histogram in time for visual analysis.

INITIAL DESIGN

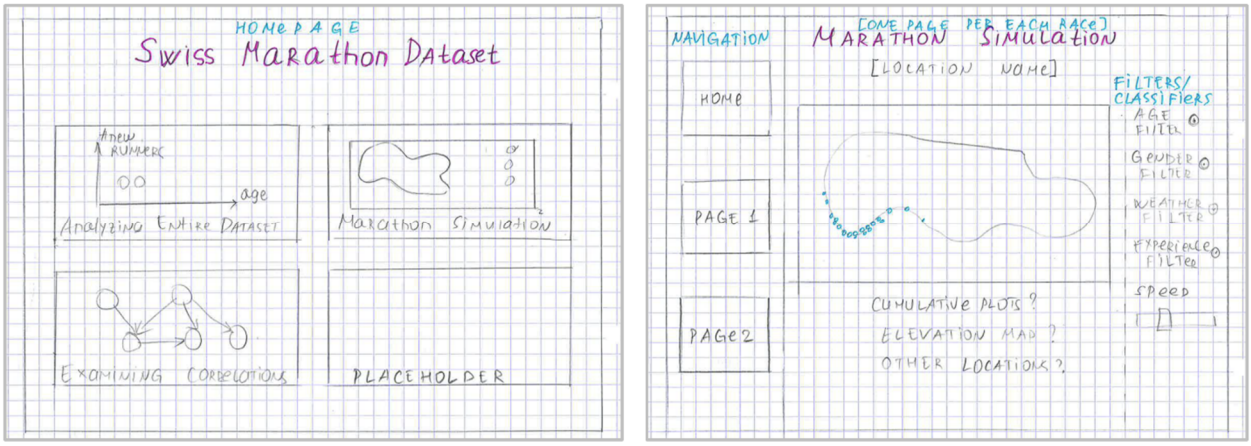

Figure1. Initial design with homepage (left) and a page corresponding to each race (right).

From the beginning we decided that the visualization will consist of a homepage (Figure 1, left) and pages corresponding to each race of the dataset (Figure 1, right). Initially we considered showing several different views at the data.

- *View 1:* simulation of the race with runners moving on the map (Figure 1, right). The user is provided with filters and classifiers for exploring the data.
- *View 2:* playing out the dependency of percentage of new runners on the average pace with the age of runners as time dimension.
- *View 3:* a graph of correlation of the attributes for each runner, such as gender, age, experience, speed and weather at the day of the race.

FINAL DESIGN

Eventually we decided to concentrate on the most interesting view – *View 1*. The rationale was presenting the data in a linear way and making *View 1* more compelling by concentrating all the effort on it. The final design is presented at the Figure 2; the control elements are highlighted in orange. Compared to the initial design we changed the following:

- Instead of containing the controls for accessing *View 1, View 2* and *View 3*, the homepage contains a map of Switzerland with clickable tracks of the races.
- To engage the user, the visualization starts with a short introductory clip, that can be dismissed at will.
- Since some tracks overlap on the map, the user also has an access to the list of all tracks from the drop-off tracks menu.
- In addition to the simulation of runners moving on the track, *View 1* also contains a set of moving distribution of the runners' position by the current classification attribute (gender, age or experience).

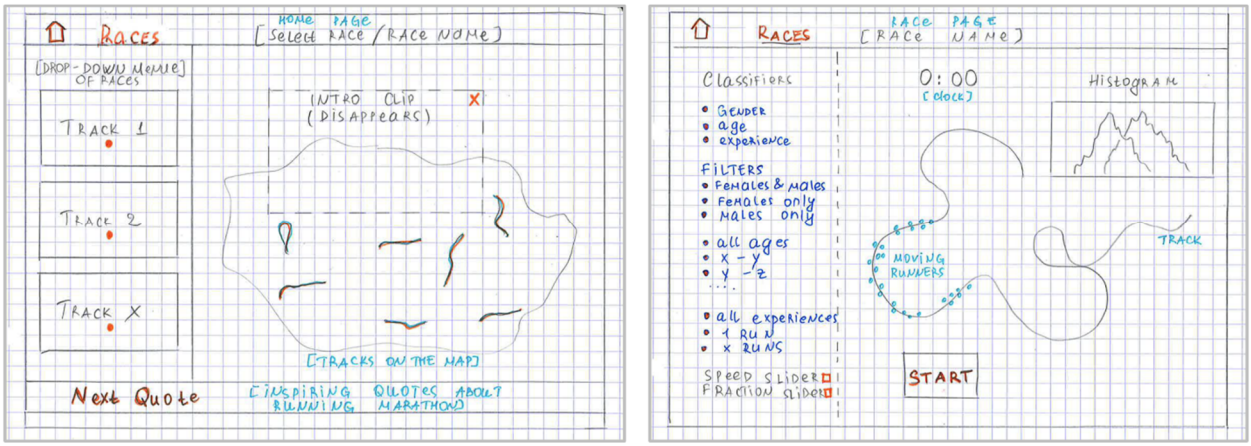

Figure 2. Final design with homepage (left) and a page corresponding to each race (right).

3.2 MARK AND CHANNELS

To make the perception of the dataset more natural, we implement a simulating of each actual race, representing the runners as moving circles. The position of the circle encodes their progress in the race. The color of the circle encodes their personal qualities, such as age, gender and experience. This is a simple way for a user to not only look at aggregated quantities of the dataset, but to see the individuals.
We also provide a histogram for the ease of analysis, while the individual runners visualization demonstrates that there are outliers in every category and that a person cannot be exactly represented by an averaged performance score of their group.

As mentioned, each runner has 4 categorical attributes in total, however, for clearness of representation only one quantitative attribute and one categorical or ordered attribute is

shown at any particular time. We explain our design choices below. The user is provided with the controls for changing the displayed attributes.

Attributes Available in the Dataset

- Age – quantitative;
- Gender – categorical;
- Experience (derived from the dataset) – quantitative;
- Running time – quantitative.

Chosen Encoding of the Attributes

We processed age and experience attributes to change their type from quantitative to ordered because we considered the selected granularity level to be sufficient for the purpose of visualization. Moreover, it facilitates reading the attributes.

- Age – ordered;
- Gender – categorical;
- Experience (derived from the dataset) – ordered;
- Running time – quantitative.

Explaining the Choice of Channels

Since we aim to simultaneously represent several attributes, we need to carefully select the channels that can be used together.

- *Quantitative attributes:* The position channel is pre-selected to use for the qualitative attribute "running time".
- *Ordered attributes:* Moving along the efficiency scale of the ordered attribute (T. Munzner, Visualization Analysis and Design, 2014), we select a color as a channel for "age" and "experience" attributes. We do not use the higher efficiency options length, angle, size and 3D position. This choice is dictated by the fact that the circles are continuously and overlapping moving in 2D, thus the listed attributes would be hard to discern.
- *Categorical attributes:* for the categorical attributes we use "color hue", which is the second option along the efficiency scale; as with the ordered attributes, the most efficient option "spatial region" is already used for representing "running time".

Absence/Filtering as additional modality of data

We enable "intersection filtering" of the attributes. In more details, the user has an option of only displaying the runners with a given value of the attribute; for example, only the runners under 20. The filtering allows to implicitly annotate the data with the value of the attribute, since the user can infer the value of the attribute in question for all the displayed data points.

Transparency and random shift to encode density

The "fraction" control (see *Implementation* Section) allow the user to either examine the density of the runners or follow the progress of individual runners.

The "fraction" slider changes the displayed fraction of the race participants. When the fraction is high, multiple runners could be displayed at the same point without any user feedback for the runners' density. We solve this problem in the following way: we assign to each runner a random shift along the width of the track. Add the same time we add 35% percent of transparency to each circle. The combination of random shift and transparency serves as a proxy to the density of the runners.

3.3 Perception

Emphasizing navigation elements with Color

While for color-coding the attributes of the runners we carefully selected the smoky colors that match the palette of the underlying map, all the navigation elements are emphasized in "Coral Red" color. "Coral Red" color is selected as a bright color harmonious with the existing color palette.

Selecting Color-Coding

We provide a legend that explains color-coding of the runners' attributes. However, we also believe that intuitive interpretability of the colors is important for the expressiveness of the visualization. The chosen encoding is presented on the Figure 6 (left). For the age attribute we use light green color for representing young runners. For the experience attribute we use green color to represent people running for the first time. For the gender attribute we use the standard blue and magenta colors.

3.4 Storytelling

The high-level storyline of visualization is zooming in from cumulative statistics of Swiss races to the individual participants.

Storytelling Model

We choose a "Data Scientist Model" of story-telling, as opposed to "Journalist Model", that is we add very little annotation and no written conclusions, but rather we allow the user to draw his own conclusions from the data.

Storyline

Our visualization has a linear storyline in the first part and allows the user to explore freely in the second part.
The visualization starts with a short non-interactive display of cumulative numbers about races in Switzerland. At the next step the user can examine the locations of races tracks and select a track for further exploration. After the track is selected, the visualization switches to a different view. The chosen track is scaled to the full screen and a big button "START" grabs attention of the user. Once the "START" button is pressed, the simulation of a race begins. At this point the user has access to the set of filters that enable color-coding the runners by their gender, age or experience as well as filtering them by the same criteria.

NAVIGATION/INTERACTION ELEMENTS

We provide a limited set of intuitive interaction elements to facilitate linear development of the storyline:

- Selecting a race
- Starting a race
- Classifying/Filtering runners
- Returning to the main page

4. Implementation

4.1 Homepage: Map with Clickable Races Tracks

The homepage of the visualization is a map with clickable races tracks.

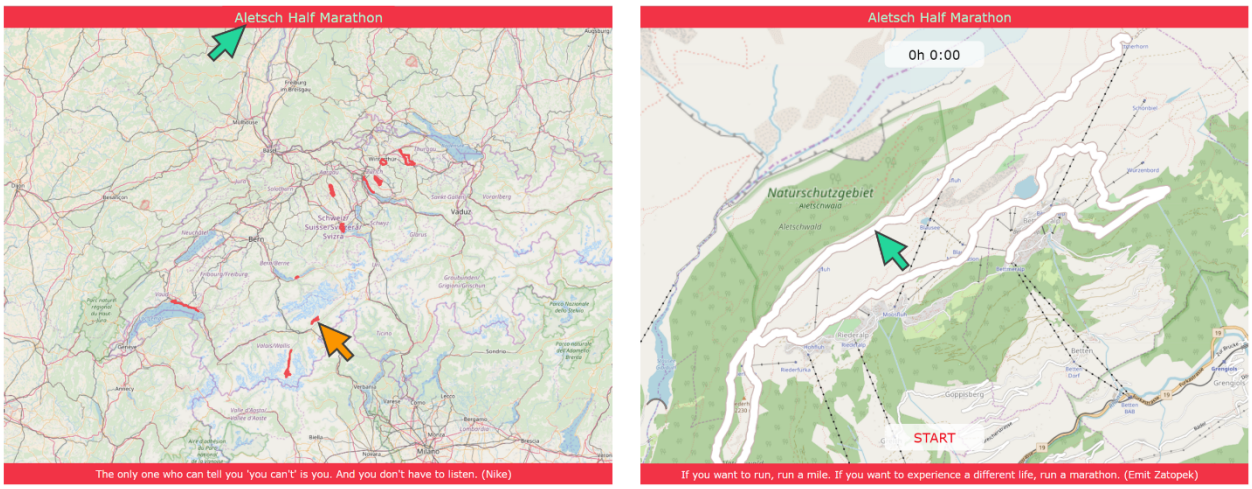

Figure 3. Left – homepage with clickable races tracks. Right - race page opens when a track is clicked.

Displaying the map

To display the map, we used open-source JavaScript library Leaflet. Leaflet is the leading open-source map library. It has a strong community and it offers all the tools that we need. The annotation of the map with points and line is easy and well documented. These are the reasons behind our choice.

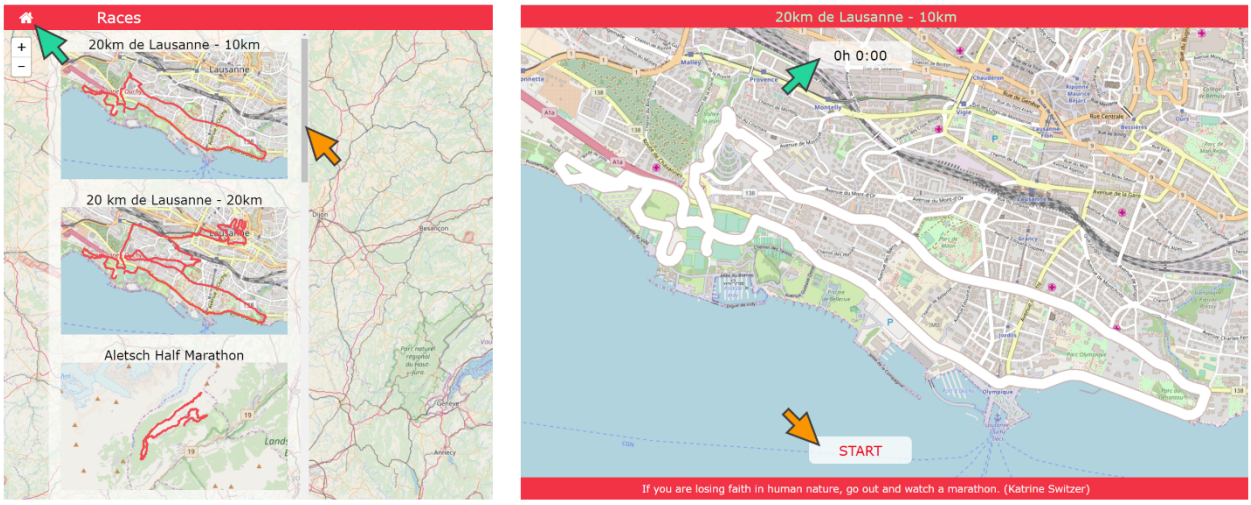

Figure 4. Left – drop-down menu with the list of races. Right – race page with start button and a clock.

DISPLAYING RACES TRACKS

We found online the GPS coordinates of each race in GPX format. Using the library leaflet-gpx, available on Github, we annotated the map with a line for each race.

INTERACTION ELEMENTS

Hover: when the user moves a mouse over the track (Figure 3, left, orange arrow), the name of the corresponding race is displayed at the top of the window (Figure 3, left, green arrow).
Mouse-click: if a track is clicked, the rage page for that track is opened (Figure 3, right). The track is shown as a thick white line (Figure 3, right, green arrow).
Home button: for returning to the home page, we provide an easily noticeable home button at the top left corner of every page (Figure 4, left, green arrow).
Drop-down menu: since some tracks are co-located on the map, we also provide a list of all races in a drop-down menu (Figure 4, left, orange arrow). The user can scroll down to examine all the available races.

4.2 RACE SIMULATION

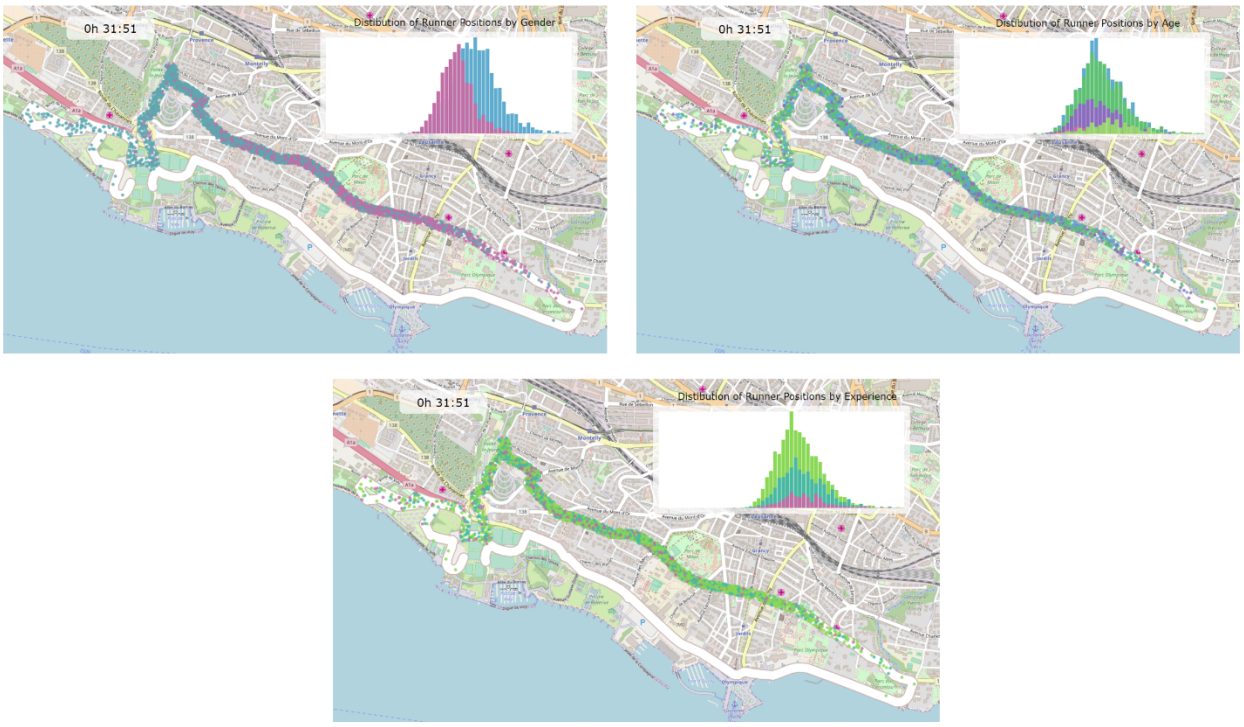

Figure 5. Classification of the runners by their attributes. Left – classified by gender; right – classified by age; bottom – classified by experience.

Once a race from a drop-down menu is clicked, the race page opens. As mentioned, it contains a track displayed on a map, as well as a start button (Figure 4, right, orange arrow) and a clock that shows simulated race time (Figure 4, right, green arrow).

Simulation Algorithm

With a press of start button, race simulation is launched. The positions of the moving runners are represented with circles; their attributes are color-coded. To efficiently update the positions of the runners on the map, we create a lookup table that provides the GPS coordinates of the runner, given the distance covered so far.
We chose to display the positions of individual runners instead of providing the density of runners on the track. This is because we enable color-coded classification of the runners and their filtering on multiple attributes. Thus, the color channel that could have being used for encoding the density is already used for classification. An alternative solution could be aggregating the runners of the same class in the bigger circles. However, since the runners move relative to each other the circles would have to be split and merged. This would introduce additional clutter of the scene; moreover, implementing smooth transitions for split and merge is too involved.
To improve runtime of the algorithm, we introduce a parameter "fraction". On a desktop thousands of runners are simulated in real time. However, on a laptop the user may choose to display only a fraction of the runners for better performance.

4.1 Classifiers and Filters

Figure 6. Left – classifiers and filters for exploring the dataset. Right – moving distributions of the runner positions, the runners are split on three groups by experience.

Classifiers

We provide a set of classifiers (Figure 6, left, purple arrow) that allow to color-code the runners by their gender, age of experience (Figure 5). The legend with explanation of the colors for the classifiers is provided below then at the filters panel.

Filters

Applying filters allows to see a subset of runners, for example only females or only males (Figure 6, left, orange arrow). Applying several filters simultaneously gives an intersection

of the filtered groups. For example, on the Figure 7 (left) the male runners over 60 are color-coded by their experience.
On the Figure 7(right) the female runners with experience of more than 3 runs are color-coded by their age. On the Figure 7 (bottom) the runners under 20 that run for the first time are color-coded by their gender.

4.2 Runner Positions Histogram

Some tracks are quite winding and it might be hard to see the distribution of runners positions from the track. Thus, we provide set of histograms that show the distribution of the runners positions by current classification (Figure 5). The histograms move along with the runners (Figure 6, right). On the Figure 5 (left) the runners are classified by gender and the histogram window displays two moving distributions – the positions of female runners and the positions of male runners. On the Figure 5 (right) and Figure 5 (bottom) a set of distributions for age and experience attributes is shown.

Figure 7. Applying several filters to the runners simultaneously.

4.3 Quotes about Running a Marathon

To enhance emotional involvedness of the users, that is to remind that running a marathon is very challenging, we provide inspiring marathon quotes. To see new quite one should press “Next Quote” button (Figure 6, left, green arrow).

5. Evaluation

5.1 Learning from the Data

Our visualization allowed as to see some known facts: males in general run faster than females, and some surprising facts: the speed of the runners does not drastically change with age and experience. It was inspiring to learn that a fraction of runners is older than 60, more over the are some runners with age up to 90.

5.2 Future Work

Some ideas for future work are listed below.

- Making the runners clickable and displaying information about the runner on hover or on click;
- Further optimizing the algorithm to enable displaying a larger fraction of runners in real time on a laptop;
- Adding several more types of distribution plots that evolve in time with the progress of the race.

6. Peer Assessment

- Preparation – were they prepared during team meetings?
  Stefano – yes
  Anastasia – yes
- Contribution – did they contribute productively to the team discussion and work?
  Stefano – yes
  Anastasia – yes
- Respect for others' ideas – did they encourage others to contribute their ideas?
  Stefano – yes
  Anastasia – yes
- Flexibility – were they flexible when disagreements occurred?
  Stefano – yes
  Anastasia – yes