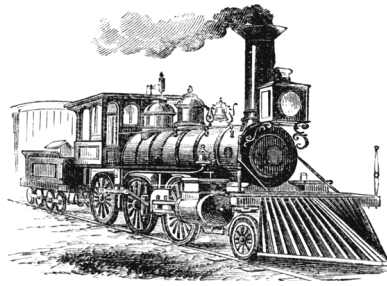# open trains

PROCESS BOOK

RAIL RUNNERS

GIACOMO ORSI

FRANCESCO SALVI

ROBERTO CERAOLO

BACKGROUND

The motivation for our project comes from the realization of the monumental gap that exists between our home and host countries, Italy and Switzerland, on the domain of open data, specifically regarding transportation. In fact, while Switzerland provides detailed information on every means of public transport, in Italy public companies refuse to release any data of that kind. Actually, Italy was fined by the EU for the absence of open data. Fortunately, independent efforts from third-parties have managed to gather a high-quality dataset of historical train trips across Italy. These data, however, have so far just barely been explored, and are far from being ready for public consumption. This is where we come into play.

GOALS
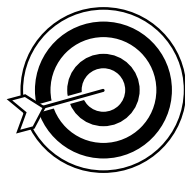
Our work aims to shed light on the efficiency of the Italian railway system, by leveraging a novel dataset of historical train logs from Trenitalia*. We decided to focus on **delay data** from the first three months of 2023, analyzing them both in a macroscopic aggregate way and from a microscopic individual perspective. We do that with the intent of making the data accessible to the general population in a simple and intuitive manner and providing impactful insights on the overall status of the Italian landscape. Primarily, this will **empower the community**, improving the information currently available and helping people in planning more informed trip schedules. Additionally, such a platform can **improve public accountability**, pushing Trenitalia to improve its railway system to decrease delays and malfunctions.

*the primary train operator in Italy, accounting for all the regional trains and about 80% of the high speed trains

THE PATH

FROM OUR FIRST SKETCHES TO THE FINAL RESULT

It's been a ride. Starting from our first researches for sources of inspiration, going through numerous sketches and brainstorming, finally to design decisions and coding. Here we will retrace the main steps and challenges that we encountered in our path while building OpenTrains.
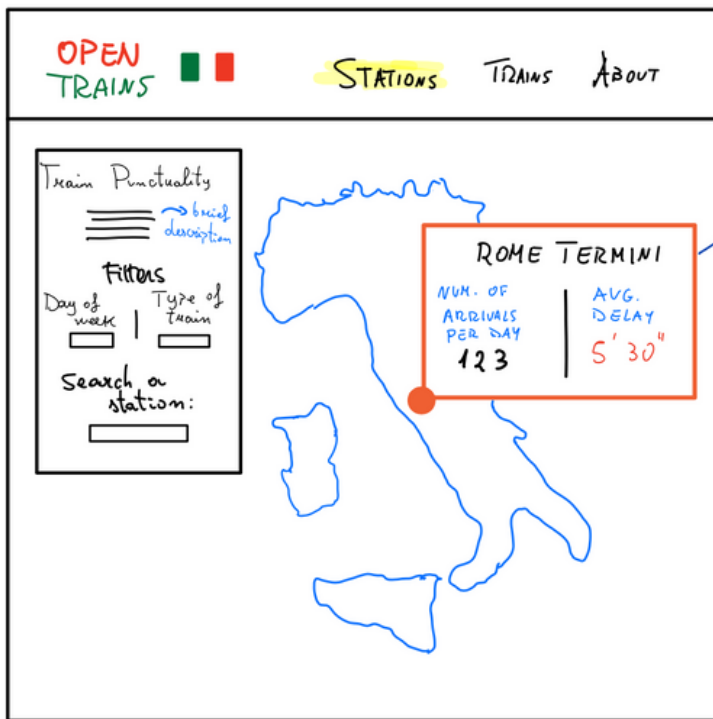
SKETCHES

We started by brainstormed which could be the most effective visualisations to accomplish our goals, sketching several versions of them on a piece of paper until we found an agreement. At the time of Milestone 2, we settled for our website to have two main pages: the "Stations" view and the "Trains" view.


The Stations page would feature a full-screen map of Italy, with a bubble plot on top showing the number of trains and the average delay at each station.

Its main features would be a search functionality for specific stations and filters for day of the week and type of train, dynamically changing the map depending to the users' selections. Hovering or clicking on a station would also show additional aggregate information and a barplot, breaking it down per day of the week.

The Trains page, instead, would zoom in at the level of individual trains, showcasing some aggregate statistics on the best and worst trains and allowing users to search for a train of their choice, presenting in-depth insights that could be helpful to someone making a plan for their journey. Particularly, we would show how the train performs at each stop of its route, both on the map and through a barplot.
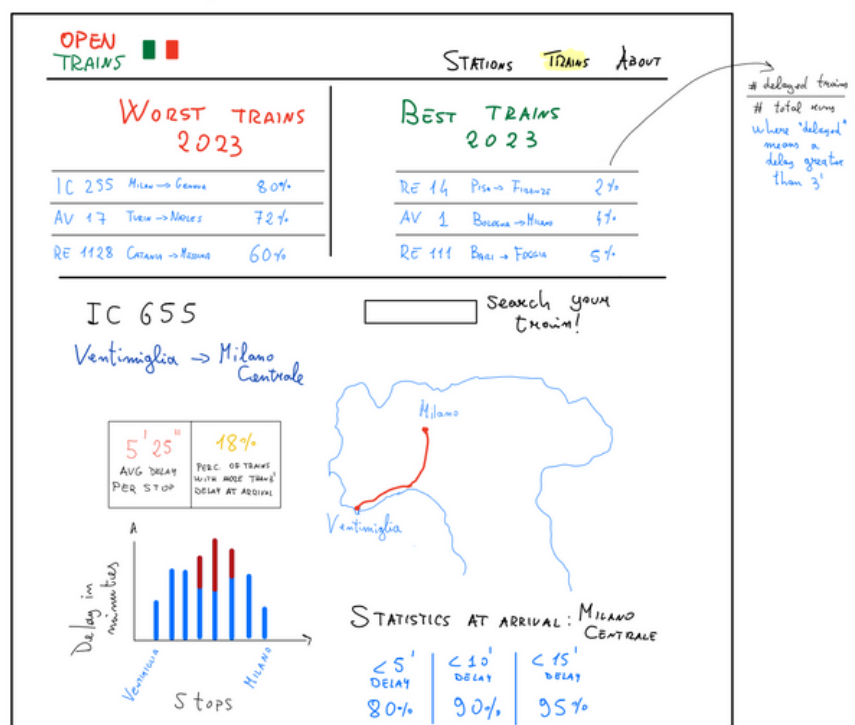
MAKING OF
STATIONS

Since the beginning, it was clear that we wanted to welcome users landing on our website with a full-screen map of Italy, providing an immediate overview of the country. However, we had to deal with a peculiar feature of its geography, which made it challenging to effectively organize the space: its **verticality**. We decided to work around it by placing the country in the center of the screen, and filling the space around it with a color legend on one side and an overlay box on the other, where users can control the map through the dynamic filters. We faced a second more technical challenge during the implementation phase: initially, we decided to implement the map using a combination of **Mapbox GL JS** and **d3.js**, the former taking care of the geographical backbones and the latter of the bubble plot and all the
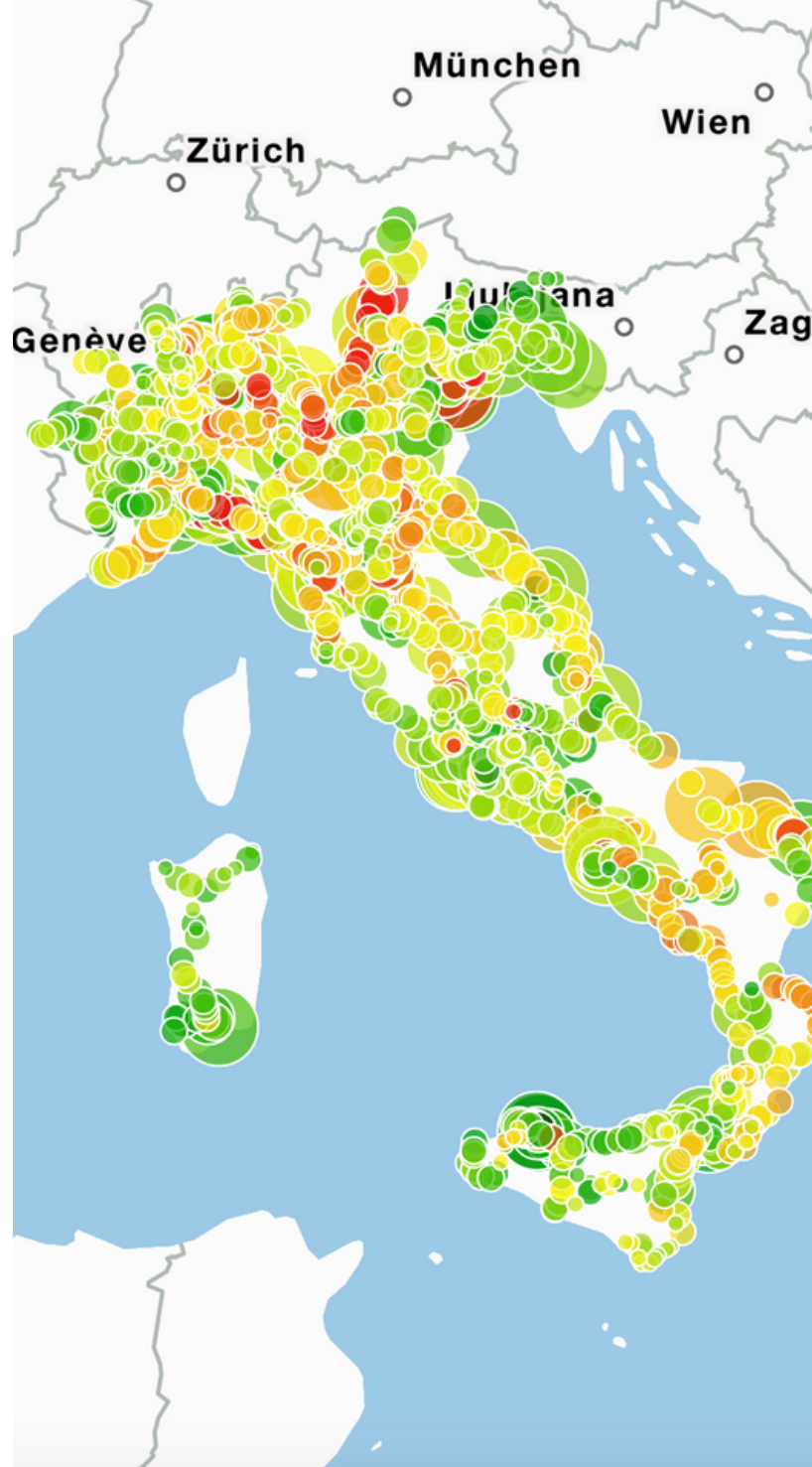
additional functionalities. However, we later realised that the interaction between the two frameworks was quite inefficient, because a projection of all the dots had to be recomputed every time an user scrolled or zommed on the map. Therefore, in the final version we decided to re-implement everything from scratch using Mapbox, achieving a much higher efficiency and smoothness in the movements.

The final product looks very similar to our initial sketches. The only major difference from the sketches is that we decided to abandon the idea of having histograms on the hover of stations, because the hover box would become way too large, making it hard to navigate the map and finally deteriorating the user experience. Instead, we opted for a **minimalistic approach**, showing only the number of trains and the average delay for each station.

For the Trains page, we spent a lot of time in trying to decide on how to best present the information that we wanted to show, to make a **smooth user experience** out of it. In the end, we decided to follow along the sketches, separating the page into multiple boxes: on the top, users can see some aggregate data and the path of the train. In order to make the paths adhere to the actual railroad, the lines are reconstructed using the Bast and Brosi algorithm.
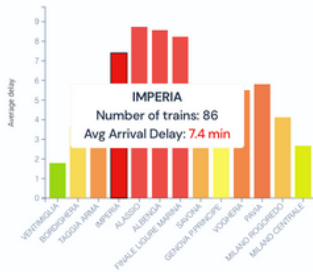
On the bottom, users can see statistics regarding single stations along the path. Instead of showing the full distribution of delays, which is likely to be well understood only by people with some kind of technical background, we decided to only provide a few quantiles, summarizing the same information in a more intuitive way.

A key challenge that we had to face for the Trains page concerned the data that we used to bring it to life. Again, differently from the virtuous example of Switzerland and SBB, Italy's Trenitalia does not adhere, ad least for its public APIs, to **GTFS**, the international standard format for public transportation schedules. As a result, our dataset completely lacked identifiers that could help us to connect different stops of the same trip or different trips of the same kind, only prodiving for each day a single sequence of stops that comprehend all the trains that ran on that day. Therefore, we had to reconstruct ourselves the timetable data, and specifically scheduled route for each train number. We did so by making a simplifying assumption: each train always follows the same route, except for few exceptional instances where some stops are added or skipped, or a deviation is followed, due to extraordinary circumstances. As a consequence, we counted the sequences of stops followed by each train number in our timerange of interest, and only took the most popular one as the scheduled route.

MAKING OF STATISTICS

During the development of the "Trains" page, we soon realised that one page was not enough to contain all the information that we wanted to show, In our initial sketches, we hypothesized to link the search functionality to a small leaderboard of the best and worst trains, so that an user clicking on one of the trains in the leaderboard could toggle the detailed train view on it. However, we then decided to incorporate this element in a separate page, that we call the "Statistics" view. This was done mainly to improve the clarity of user experience, since we found that the previous page was not visually appealing and a bit complex to understand. In addition, we wanted to have a separate place to put various other static elements, as a sort of Data Science-ish **insights hub**.
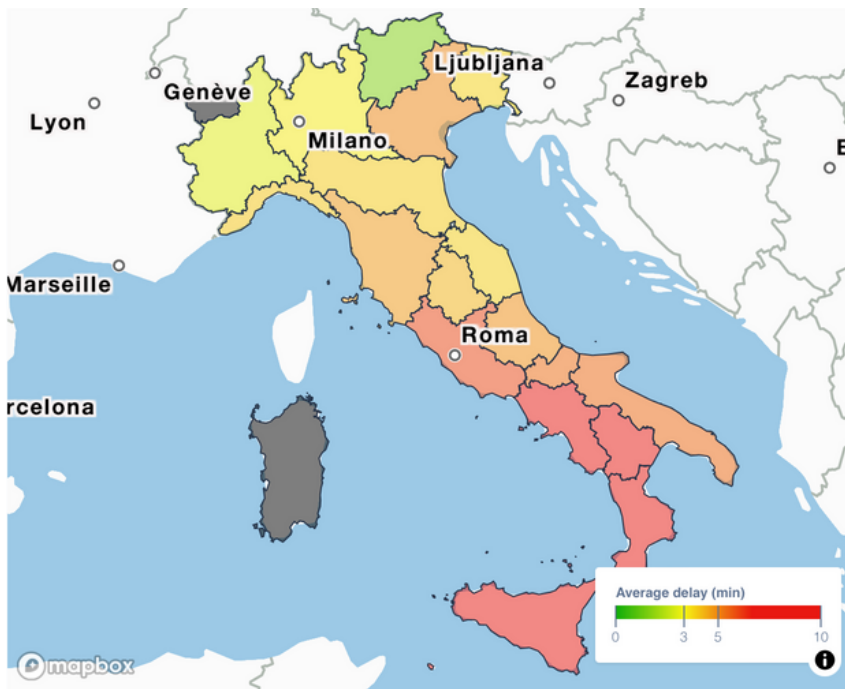
With that spirit, we included in the new Statistics page a regional breakdown of Italy, representing average arrival delays. For some types of trains, like the Intercity shown in the image on the left, we found a confirmation of the common belief that the North tends to be more efficient than the South. For others, however, we were surprised to see that there is not much regional variation. Very democratically, delays on regional trains seem to be equally bad from North to South!

FUTURE IMPROVEMENTS 🚀

With more time, we want to expand the Statistics page to turn it into a more comprehensive visual data story, investigating more temporal and local patterns . Also, we would like to incorporate **real-time data**, to show the current position of trains in the map, together with their delays. We believe that this dataset, previously almost unexplored, has still much potential for plenty of other analyses, and we are excited to see what others will build with it!

PEER ASSESSMENT

ALONE, WE CAN DO SO LITTLE; TOGETHER, WE CAN DO SO MUCH.

In general, every team member contributed equally to the overall planning, design and vision for the website, which is probably where most of the time was spent. We also worked together on the backbones of the Map and Trains views on joint sprints. However, to avoid diffusion of responsibility, we then designated individual leaders for some specific tasks.

GIACOMO - Data pre-processing and analysis, Mapbox integration, Statistics page
FRANCESCO - Hovers and legends, M2 report, Process Book
ROBERTO - Sketches, search, autocomplete and filter functionalities, Screencast

ENDING NOTES

We would like to thank the staff of TrainStats, that gathered train data for years and made it publicly available. This allowed us to create OpenTrains, even if Trenitalia does not officially publish historical data.
Another special mention goes to the Swiss Railway company, SBB, that together with the Belgian Mobility Dashboard inspired us in creating impactful visualizations in the realm of public transportation.