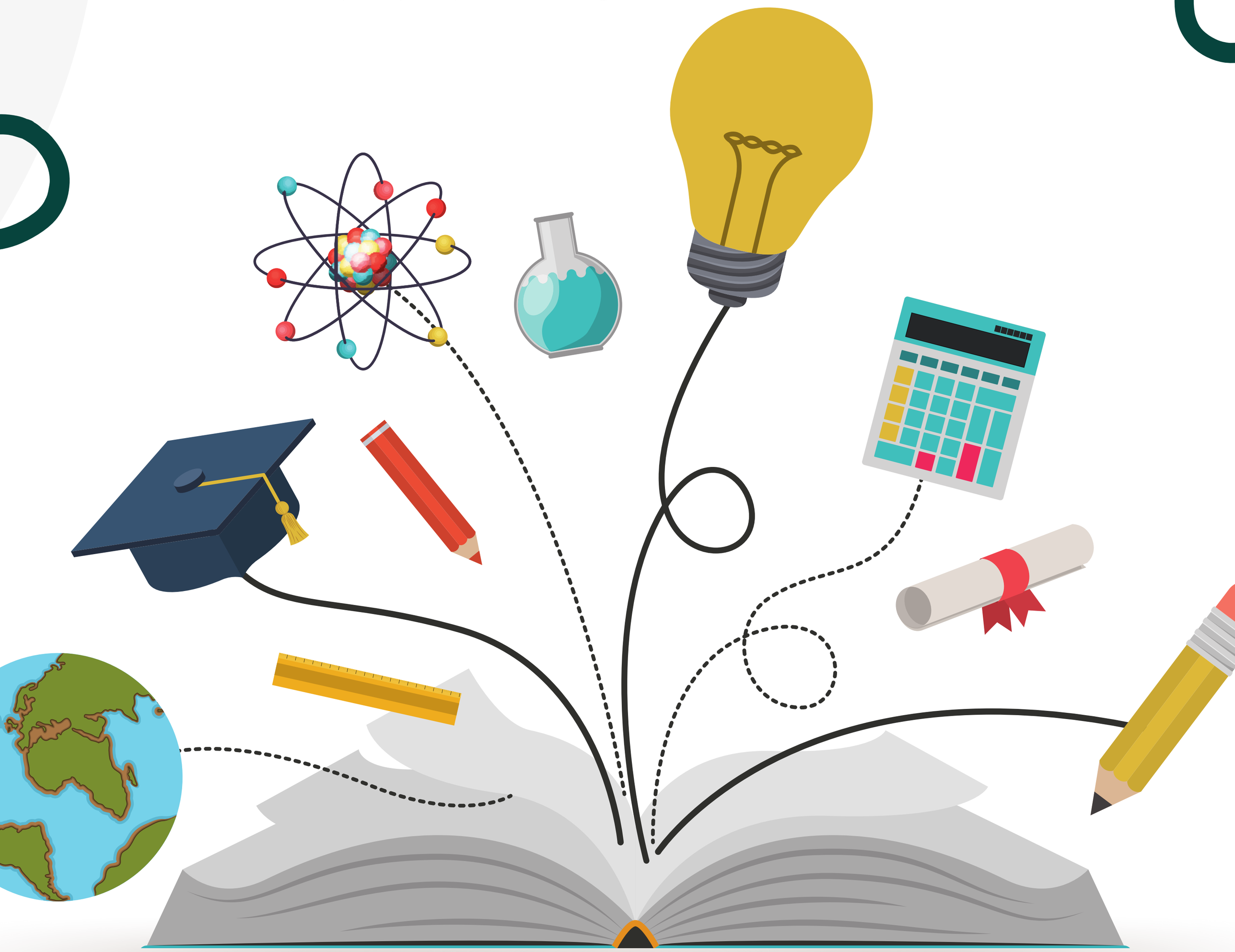
IN PURSUIT OF EDUCATION

WHERE DO EUROPEANS GO ON ERASMUS EXCHANGE?

Process-book for COM-480 Data Visualization at EPFL.

Team “Why Axis”:
Batuhan Faik Derinbay SCIPER 340560
Maciej Styczeń SCIPER 331214
Nevena Drešević SCIPER 321682

GitHub: https: //github.com/com-480-data-visualization/datavis-project-2022-why-axis
Website: https: //com-480-data-visualization.github.io/datavis-project-2022-why-axis
Screencast: https:https://youtu.be/t7a9g44fBnk



INTRODUCTION

The **Erasmus mobility program** (EuROpean Community Action Scheme for the Mobility of University Students) is a program that provides free movement and education between students and universities or institutions. Today, it stands as one of the largest international student exchange programs globally, with more than 3.3 million students since its founding in 1987. For students and staff as participants in the program, it brings many benefits, such as learning, exploring different cultures, forming international connections, language skills, traveling, and personal development. Its significance also reaches institutions participating in exchange programs. It creates a more open and accessible environment, modifying institutions' courses and teaching methods to attract international students and become more modernized. Hence, our project brings more insight and value to the program's student exchanges.

GOALS

The project's goal is to visualize and provide insights into the structure of educational mobility in Europe. Some of the key viewpoints of our visualizations include:

1. **An institutional and geographic aspect of mobility:** explore the volume of sent/received students per country, education level, academic year, and country, and also identify strong ties between certain countries/institutions.
2. **The temporal aspect of mobility:** explore and present how the trends in mobility change with time.
3. **The demographic aspect of mobility:** understand the differences in the structure of mobilities for different demographic groups, and show the impact of attributes such as gender, education level, or nationality on the activity in the mobility program.

Our end product aims to offer the user an interactive visualization environment that allows them to explore the data on their own.

DATASET

The dataset we are using is Erasmus mobility statistics collected for the years 2014 to 2019, found on Data Europa's website and created by the Directorate-General for Education, Youth, Sport, and Culture. It contains information on the mobility of both students and teachers, with more than 3 million participants. Each mobility activity is characterized by 24 attributes that show its mobility period data, personal information, and sending and receiving organizations for its program. The practical information on the mobility period includes the academic year, start and end months, duration, and mobility type. Also, there is personal information about each participant, such as gender, age, nationality, the field of education, education level, whether a participant is a learner or staff, if it requires special needs, etc. For sending and receiving organizations participating in the program, we have their codes, city, and names.

PROCESS OVERVIEW

The project development process consisted of the following phases:

1. Perform exploratory data analysis and data wrangling
2. Brainstorm to collect ideas for visualizations
3. Sketch the visualizations by hand
4. Set up a website skeleton with placeholder visualizations
5. Implement visualizations
6. Add visualization interpretations
7. Create screencast and process book

EXPLORATORY DATA ANALYSIS

The dataset is divided into two parts corresponding to two Key Actions: KA1 and KA3. KA1 is the Erasmus+ Exchange program we are all familiar with. KA3 on the other hand, assists in the creation of new policies at the European Union and system level in the domains of education, youth development, and sport. We focus on KA1 since our main point of interest are the student exchanges. The KA1 dataset contains 3.1 million rows containing different types of activities, such as student exchanges, traineeships, and volunteering.

We have also observed several trends in the data:

- We notice a clear increasing trend regarding the evolution of the total number of participants in the program across time. Moreover, the number of participants in the academic year 2014/15 seems very low compared to others, suggesting that the data for that year is incomplete.
- While the majority of participants, including high-school and university students, are classified as “Learners”, there is still a fair number of activities involving staff (e.g., teacher training). We also notice significantly more female than male participants - around a 60-40 split.
- The age distribution of participants shows a visible peak in the 18-25, which corresponds to the typical age of university students. However, there are still many participants in the age bracket 30-60, possibly Ph.D./Postdoc students and academic staff.

WRANGLING CHALLENGES

Before we get started with the analysis, a few preprocessing steps needed to be performed, including:

- Filtering out only activities involving exchange university students in their Bachelor's, Master's, or Ph.D. studies.
- Parsing missing values: there are multiple representations of a missing value in the dataset, including: "-", "Undefined", and "?? Unknown".
- Removing outliers and invalid entries, e.g., negative/very large age.
- Normalizing the values, e.g., some durations are expressed in months, some in days.

- Dealing with duplicate entries, some institutions have multiple labels, e.g., consider TUM and Technical University of Munich.
- Establishing a strategy for dealing with missing values since the majority of the rows in the dataset contain at least one missing value.
- Mapping the initial "education field" division containing over 150 different categories into eight categories of more coarse granularity.
- Preparing a fully-preprocessed dataset for each visualization to minimize the run-time preprocessing on the client side.

WEBSITE

To host our website online we use GitHub pages. The code, templates, and data are stored on GitHub. We used a Bootstrap template named Infinity by StyleShout as a starting point for the website, and we have modified it to fit our desired layout. The visualizations are implemented using D3.js, along with D3 libraries such as d3-force, d3-input, d3-legend, d3-slider, etc. Additionally, we used HTML, CSS, JS stack with jQuery.

VISUALIZATIONS

This section presents the visualizations we deliver in our final product. In particular, for each visualization, we depict:

1. The motivation behind choosing this particular visualization.
2. The initial sketch.
3. The challenges with implementation.
4. A screenshot of the final visualization.
5. Interpretation.

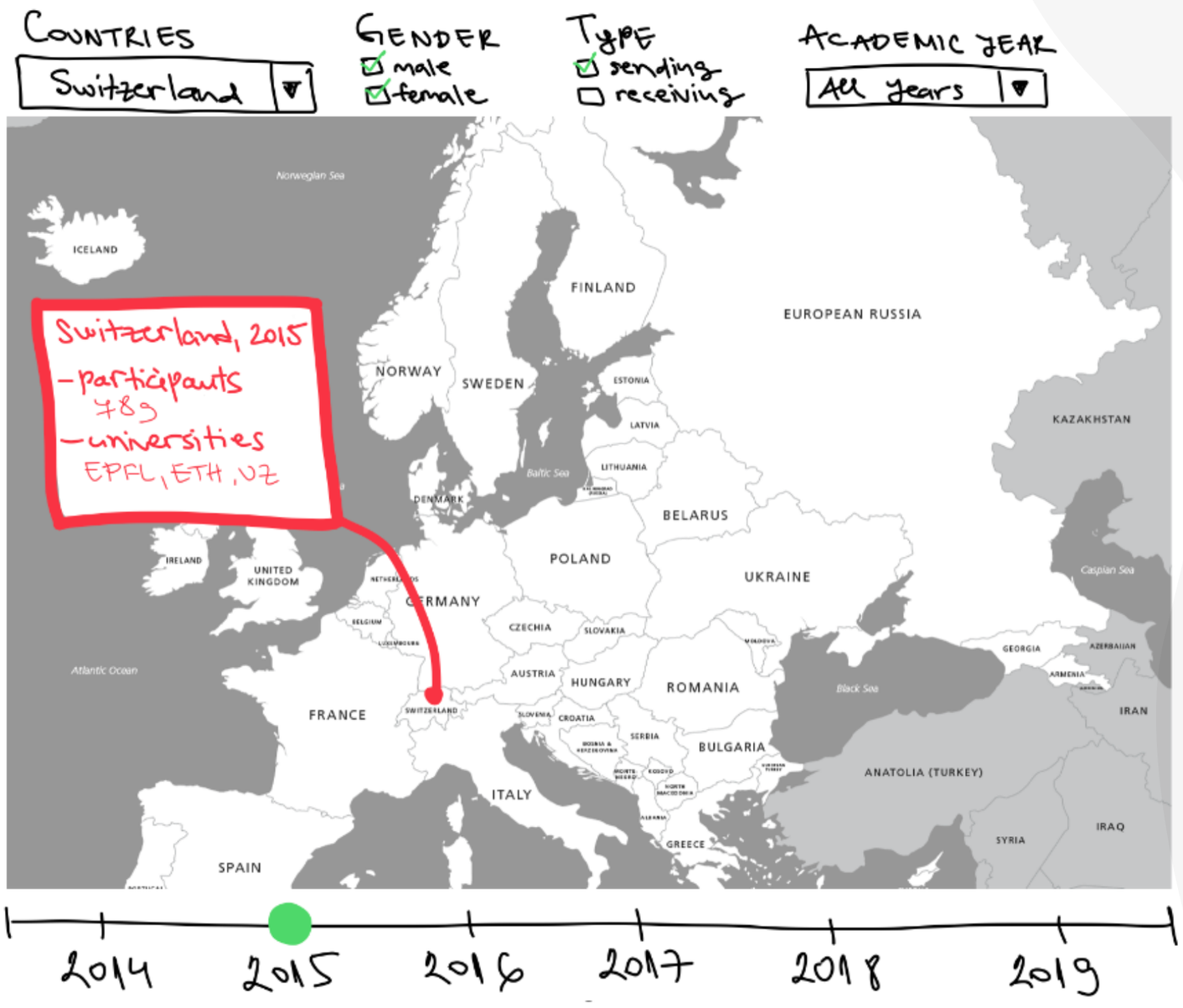
VIZ I: INTERACTIVE EXCHANGE MAP

The map shows the evolution of the Erasmus program in terms of geographic, temporal, and demographic points of view. The data is displayed per one program year, chosen from 2014 to 2019. Additionally, it is possible to filter and investigate participants' gender, type of exchange for a country (sending/receiving), and education level. The world map represents a heatmap corresponding to the number of program participants that conform to specified filters.

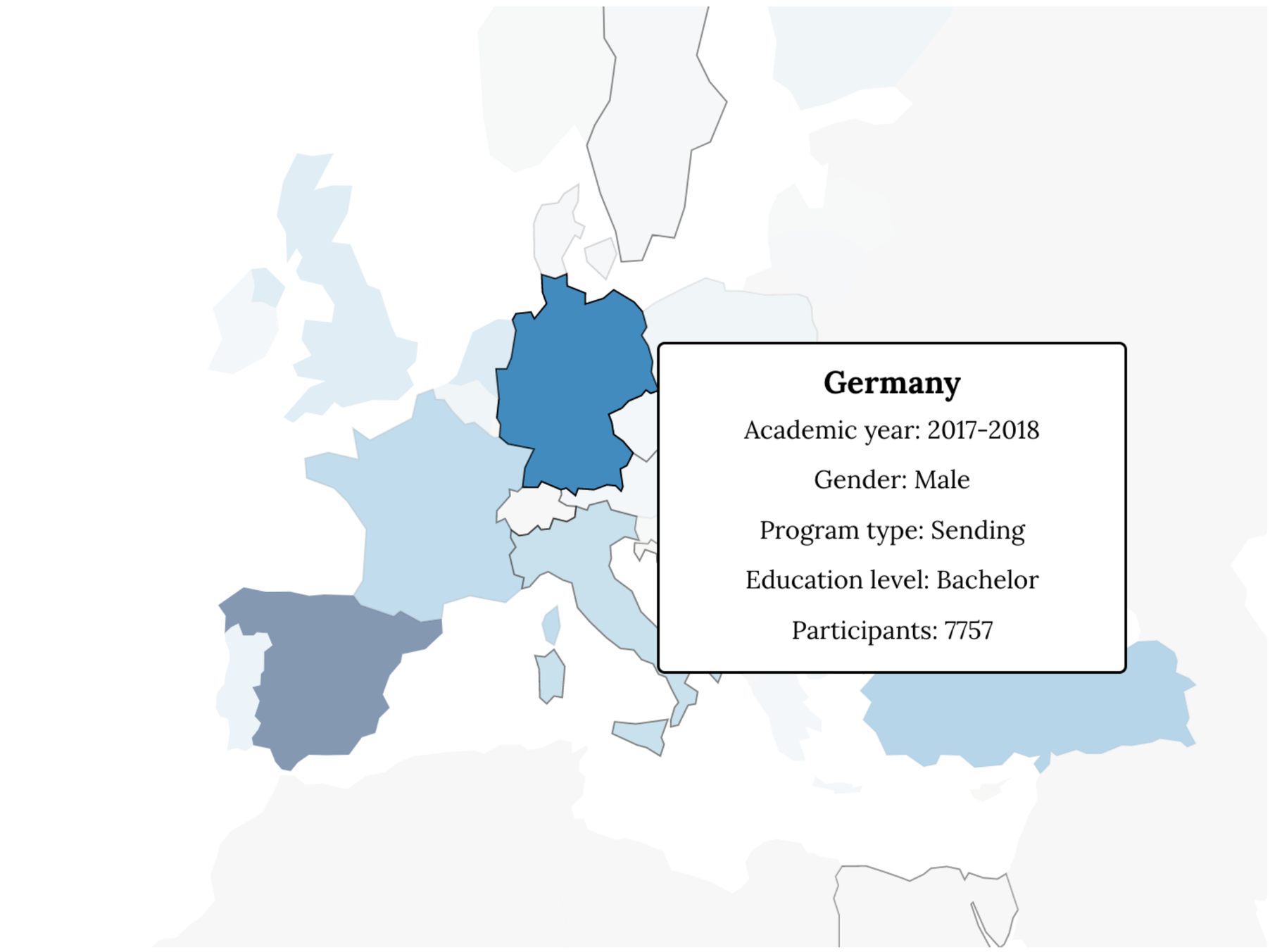
The initial sketch was transformed into an actual implementation on our website with minor changes. We decided to improve our world map by removing the drop-down box intended to choose a country, and making it a heatmap based on the number of participants in the Erasmus program. The number of participants is calculated based on the programs that fit the user's specified criteria. Countries' colors are interpolated from white to blue, correspond to program's participants from lowest to highest. Hence, the intensity of the blue represents the county's popularity in programs, while hovering over the country provides the exact details: country name, academic year, participant gender, program type, education level, and the number of participants.

The main challenges of this visualization were to choose and implement the best way to represent the country's participation based on color and information. We tried a few variations of color schemes and their changes on events before picking the final version. Additionally, it was a bit complex to preprocess the data and integrate all the user filters in the design.

Sketch:



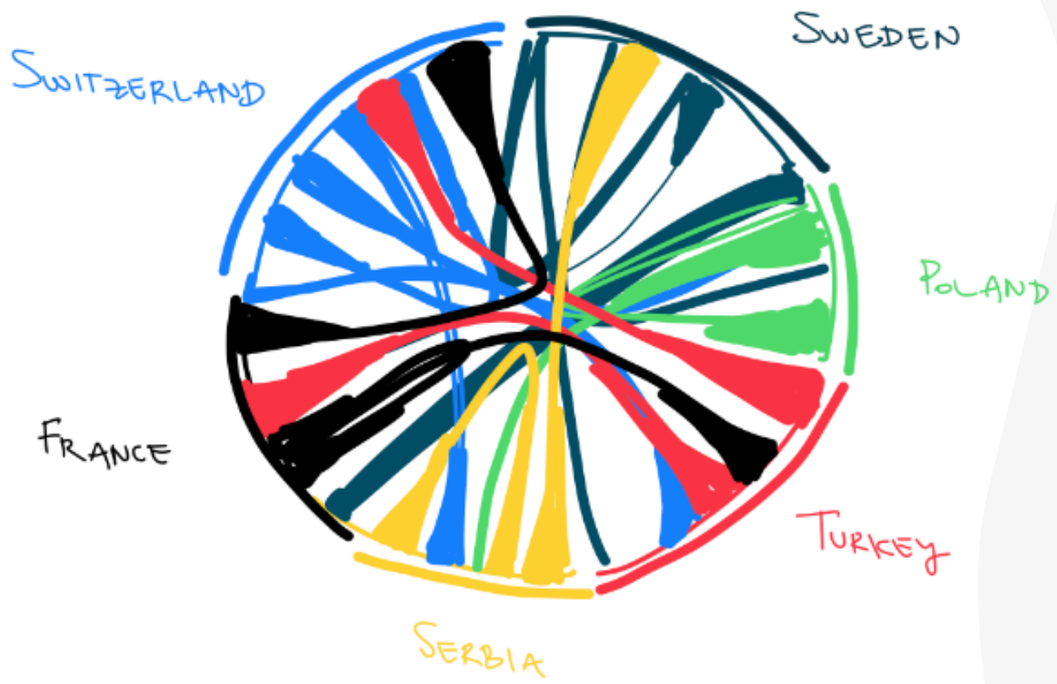
Final visualization:



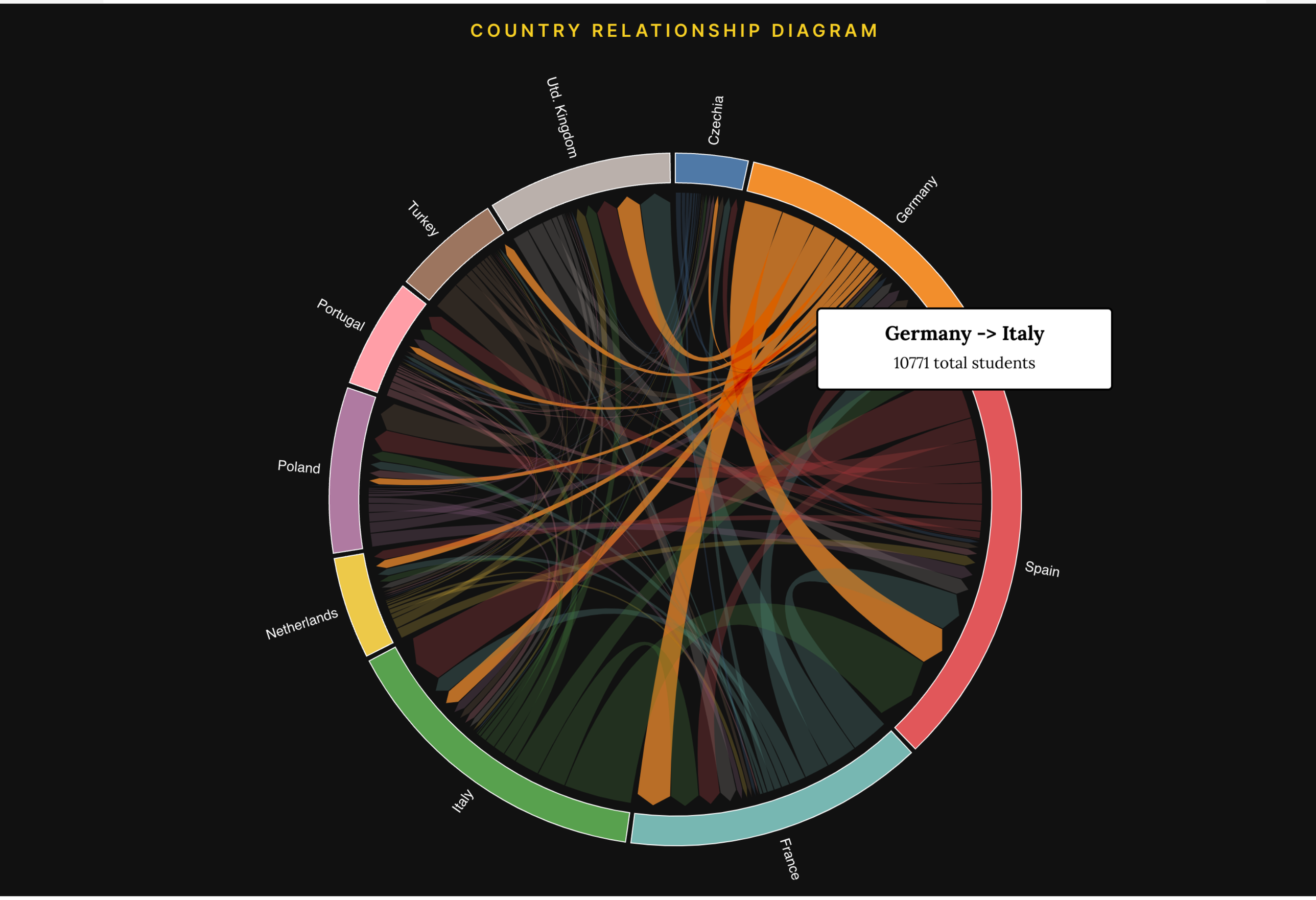
VIZ II: STUDENT FLOW CHORD DIAGRAM

To go beyond the absolute number of participants from each country and provide a deeper understanding of student exchange dynamics, we would like to observe interdependencies between countries. To do so, we visualize the flow of students between countries on a chord diagram. This chart helps us identify strong ties between pairs of countries and identify countries with a disproportion between incoming and outgoing students.

Sketch:



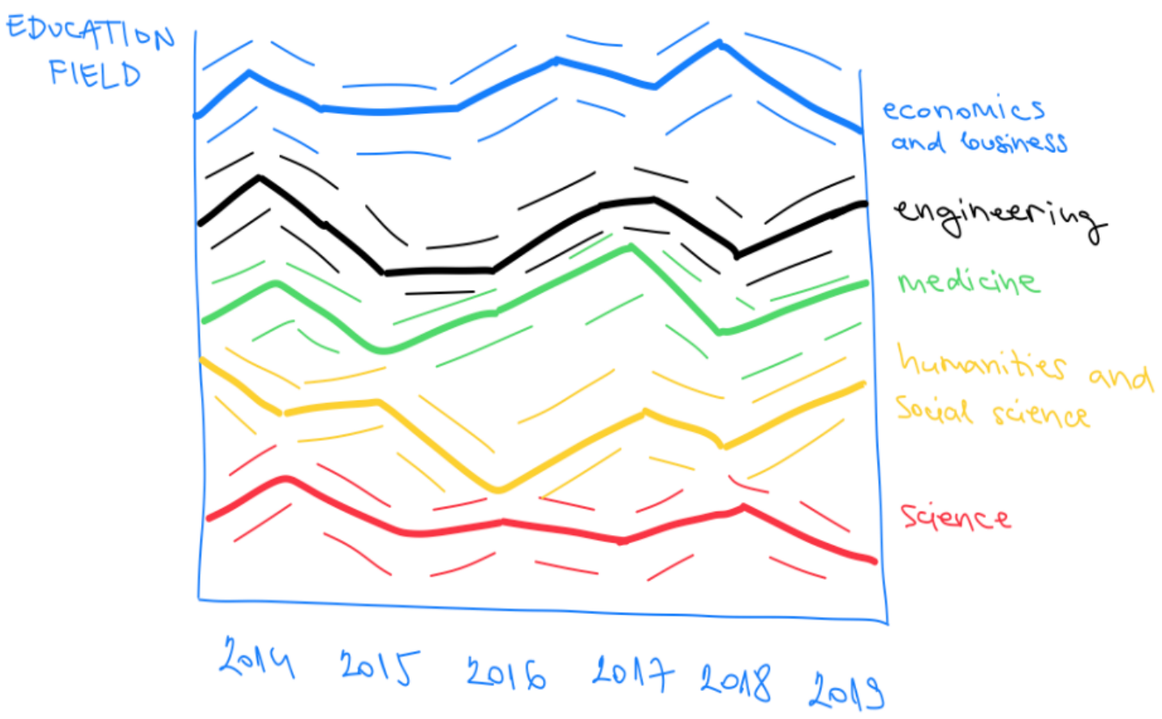
Final visualization:



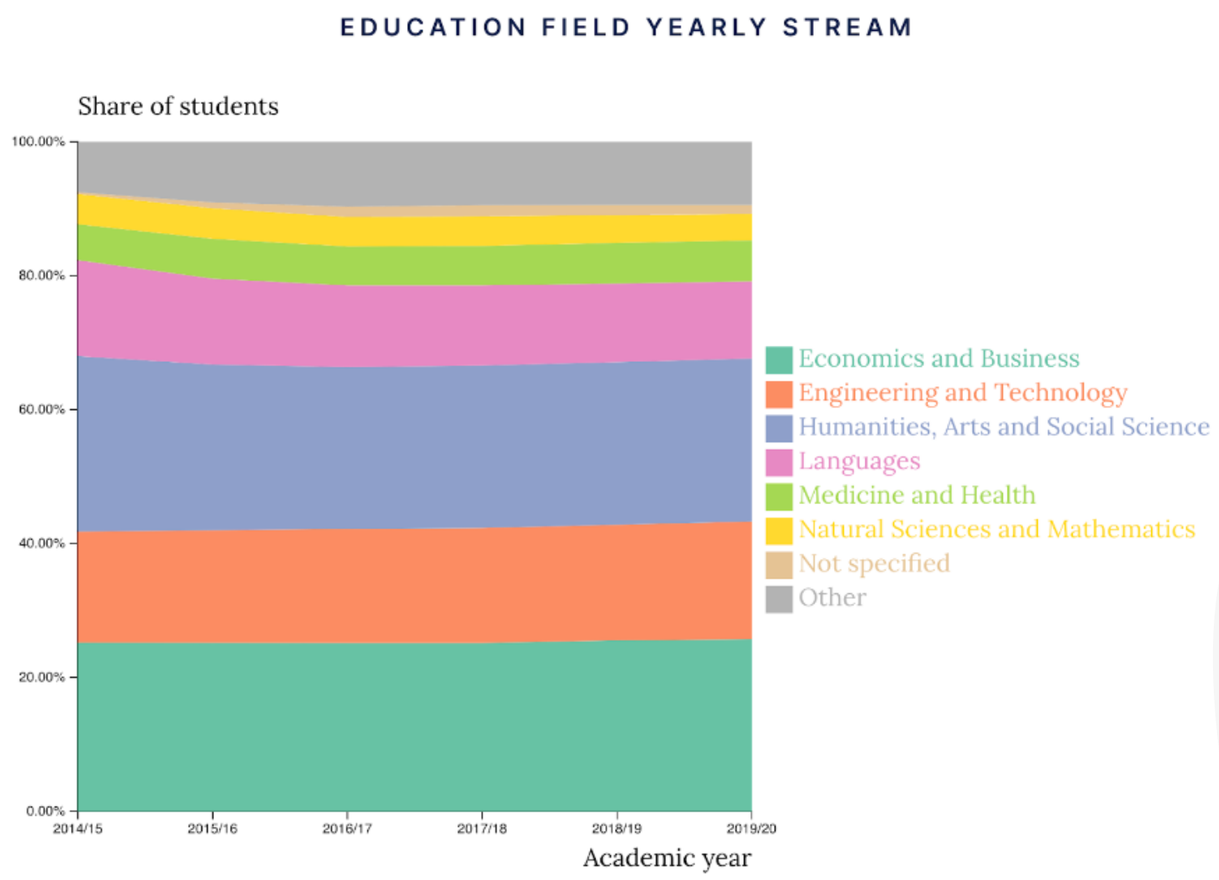
The diagram above allows for interactive exploration of the interdependencies between pairs of countries in terms of student exchanges. Hovering over the edges (ribbons) will display information about the number of students going between the two countries. Hovering over the circle cutouts (arcs) will display information about the total incoming/outgoing students for a country. Visualizing flows for all 34 countries was not possible - the sheer number of edges flowing in the graph made it completely unreadable. We decided to sort the countries based on the total number of participants (in and outgoing) and only present the flow between the top 10 countries.

VIZ III: FIELD POPULARITY STACKED CHART

Sketch:



Final visualization:



The stacked area chart shows the evolution of popularity of education fields over time. On hover, it highlights the area corresponding to the selected section. The plot shows almost no trends or changes in the education field's popularity over time. It might be that the observed period is too short for us to spot any major shifts, or maybe we cannot see clear trends because the categories are too general, and stronger trends can only evolve on a smaller scale category, e.g., Computer Science. To check that, we would need more precise data about education fields. The quality of the categorical data for education fields was not the highest - the main issue was varying granularity, with some exchanges having very precisely formulated categories as “Food processing”. In contrast, others only had a general category, e.g., “Engineering”. We grouped these labels into six main categories. The number of samples per year also varied significantly as data for 2014 and 2019 was incomplete, so we normalized the data for each year.

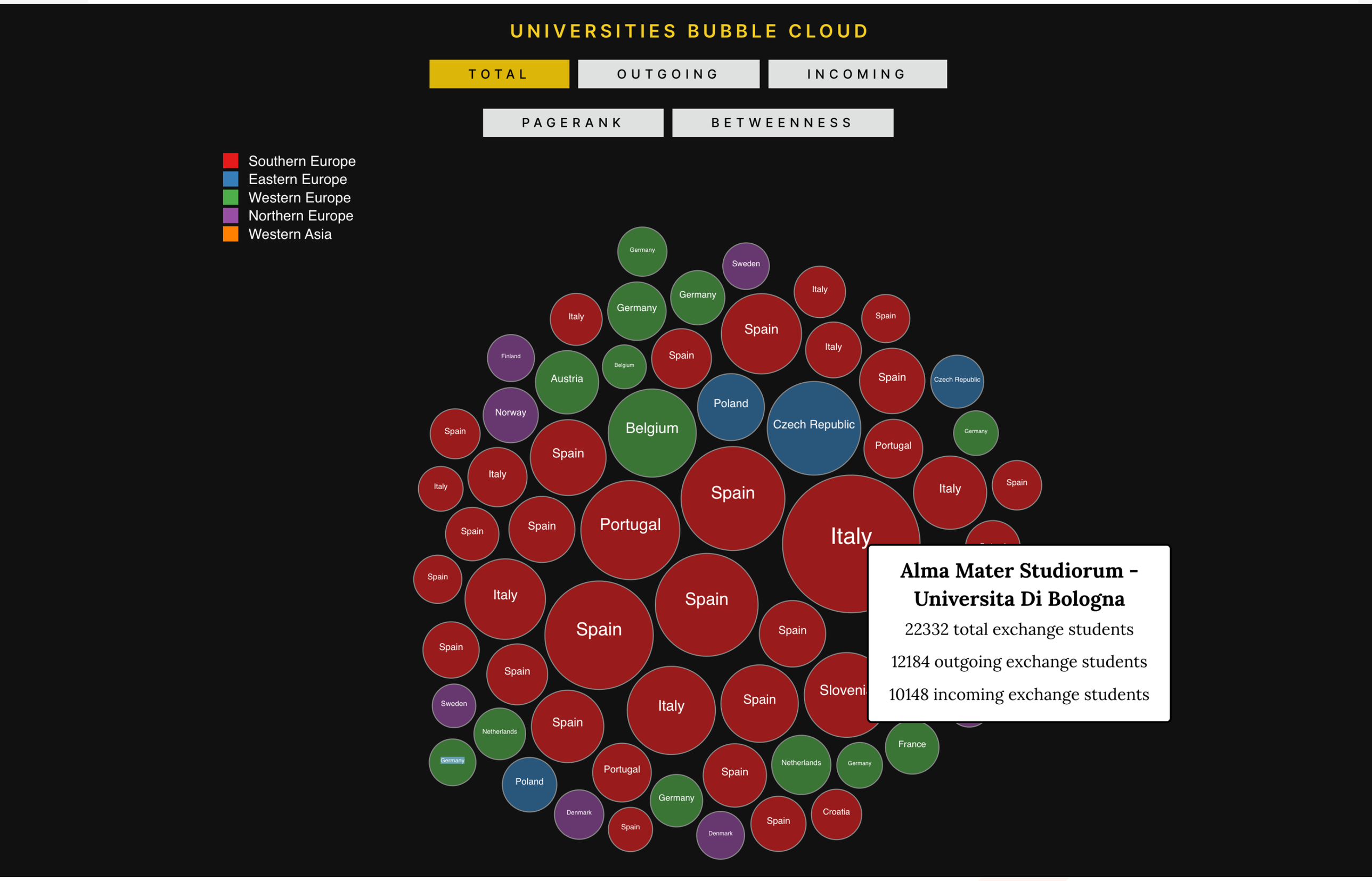
VIZ IV: UNIVERSITIES BUBBLE CLOUD

Sketch:

We use a bubble plot to explore which universities are the most active members of Erasmus mobility exchanges. Each university is represented as a colored bubble, while the attribute chosen from a filter determines the size of the bubble. This attribute can be the number of outgoing or incoming participants, the total number of participants, pagerank, or university graph betweenness measure. All the data is calculated by summing the provided data for 6 years. We decided to replace university names with country names to better represent a country's importance in the Erasmus exchange. We kept the university name and all other important information in the tooltip.

Each university's Erasmus connections are shown. It shows the total, outgoing, incoming exchange students and Pagerank and Betweenness centrality of universities as bubble sizes. The color represents the UN region. It shows that southern Europe has more exchange students and central universities. The betweenness centrality tab shows Western Asian universities. Pagerank centrality shows Northern European universities.

Final visualization:



REFLECTIONS

Throughout this journey, from dataset exploration to div item centering in flex boxes, we have learned a lot about how one can visually represent data in several ways and how important it is to use correct visualizations paired with suitable design choices. The most difficult challenges were on the JS side as the team did not have prior experience. However, we believe that we handled it with great success and teamwork.

CONTRIBUTIONS

This project involved great collaboration and equal individual contributions from all team members. We met usually on a bi-weekly basis starting from the initial dataset search until producing the final product. Each of us brought different strengths to the project, such as creativity, organization, programming expertise, etc. We worked together for all sections on Milestone 1, while the further work can be devided as follows:

Nevena: Milestone 2 text & sketches, Viz I, process book, screencast.
Maciej: Milestone 2 extra ideas, data wrangling, Viz III, Viz IV, process book, screencast.
Batuhan: Milestone 2 website setup, Viz IV, process book, screencast.