Visualizing open source machine learning projects on GitHub

There are a lot of open source machine learning projects on GitHub and trying to make sense of them is difficult. We can use visualization techniques to gain more insight into these projects. I want to demonstrate the impact of corporates on open source machine learning projects on GitHub by visualizing how many stars, commits, and contributors the projects have. We also analyze other factors like companies, year and language of development.

Dataset

First we need to find a dataset that we can use for our visualization. Fortunately I have found an open source dataset about popular open source projects related to machine learning. The dataset can be accessed here:

https://data.world/vmarkovtsev/open-ml

Next we need to get a better understanding of the key features in our dataset. These are the important features that I will focus on in this project:

- **Name**: Name of the project
- **Alignment**: Who is mainly responsible for developing the project. It can be either corporate, academia or indie
- **Company**: Name of the company if the alignment is corporate
- **Year**: Project's creation year
- **Stars**: Number of GitHub users who have bookmarked the project. It's a measure of popularity and shows how many people are interested in the project
- **Contributors**: How many users have contributed to the project. It shows how active is the community around the projects
- **Commits**: How many times code has been updated. It shows how active the project is
- **Language**: The programming language used to develop the project

Implementation

I will use the combination of R, Shiny, and r2d3 to implement this visualization project.

This is how the UI looks like:

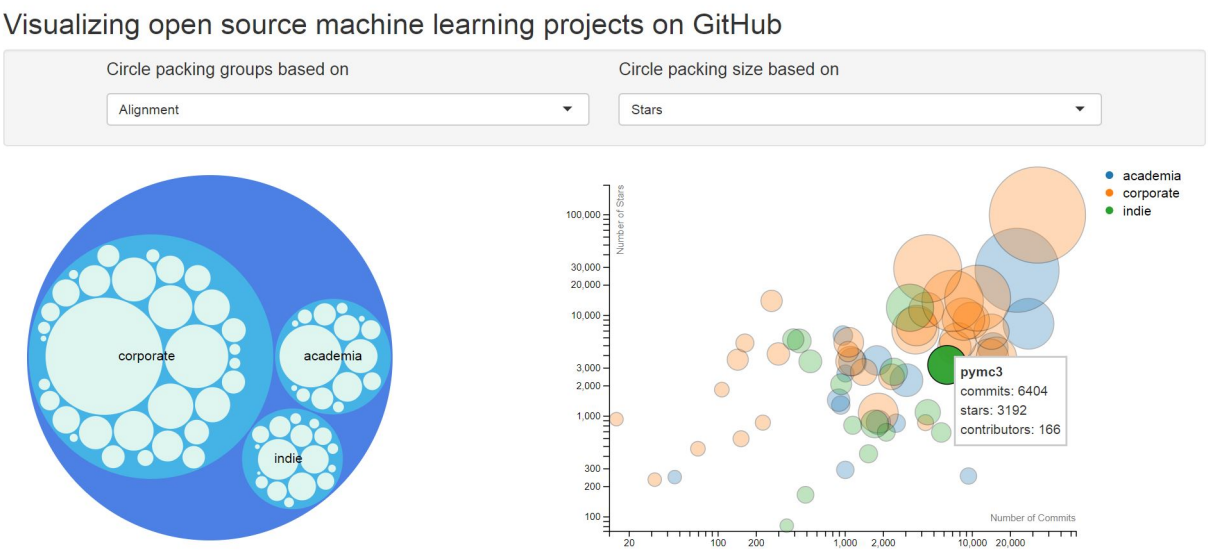

Figure 1

After the Title, I have two drop down list for selecting the parameters that impact how the circle packing is displayed:

1. Circle packing groups based on: We can use different columns in our dataset to divide our data into different groups of circles. It can take values of "Alignment", "Company", "Year", and "Language". In figure 1, the circles are divided to 3 groups based on their "Alignment".
2. Circle packing size based on: We can use other columns in our dataset to determine the size of the circles used in circle packing. In figure 1, the circle sizes show how many stars each project has relative to other projects. Larger circles indicate more stars and thus more popular projects.

On the right in figure 1, we have a scatter plot with number of commits on the x-axis and number of stars on the y-axis. Size of the circles are proportionate to the number of contributors. The large circles on top-right of the scatter plot, indicate popular and active projects with a large community of contributors. On the other hand, small circles on bottom-left, indicate less popular and not very active projects with a small community of contributors. Also by hovering on the circles we can see the exact values as a tooltip.

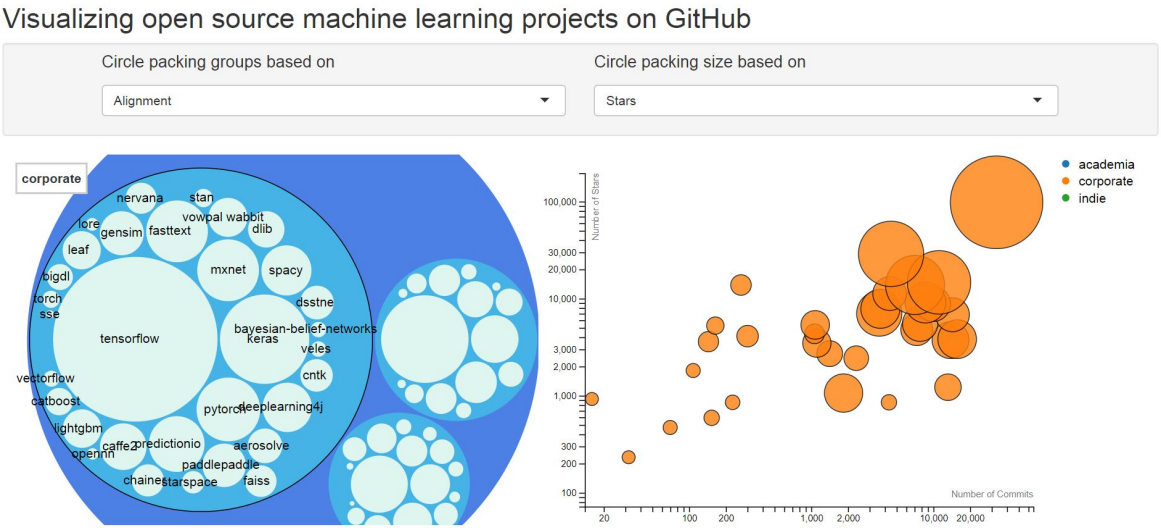

Figure 2

In figure 2, you can see the effect of zooming. By hovering the mouse over a circle on the left circle packing chart, the circle is highlighted by a black line around it. After clicking on the circle it will zoom on it, and the name of the group appears on top-left of the chart (corporate in this example). On the right, only the projects belonging to the zoomed group is displayed.

Alignment Analysis

Now that we have this tool, we can use it to gain some insight. To show the impact of corporates on the open source machine learning projects, I will use “Alignment” for circle packing grouping and vary the circle size option.

Circle size = Stars

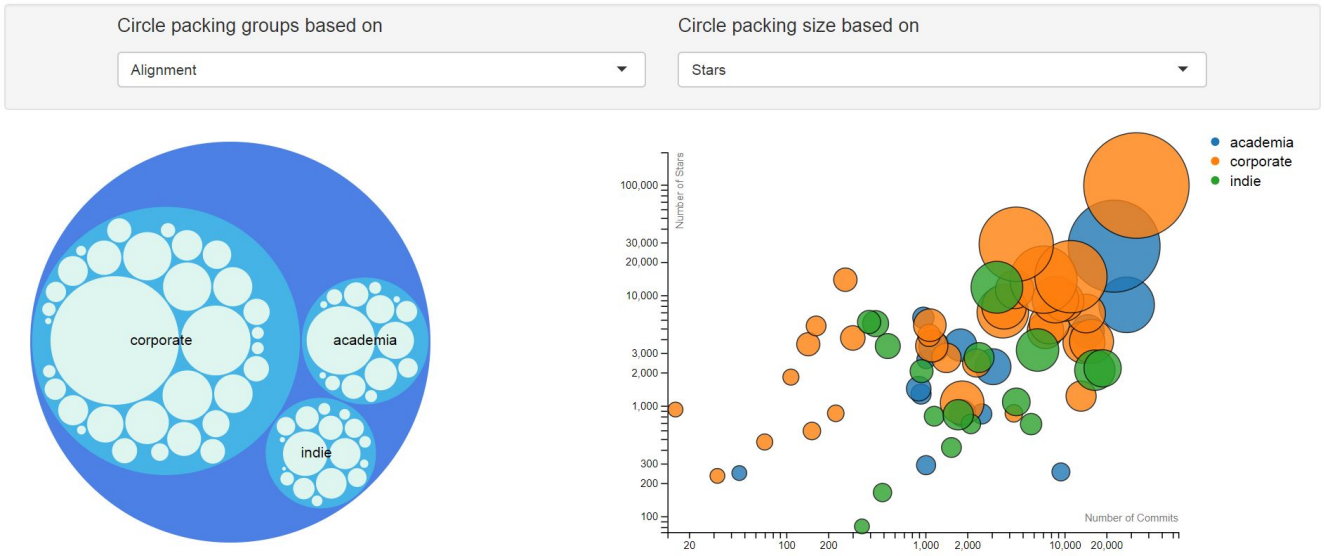

Figure 3

Circle size = Commits

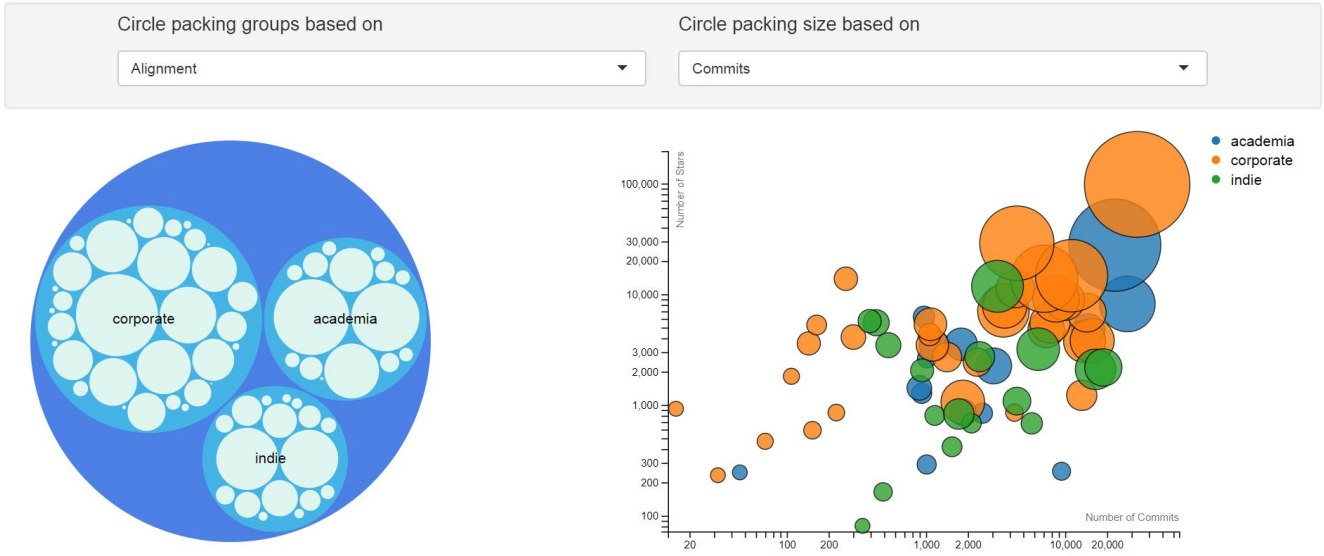

Figure 4

Circle size = Contributors

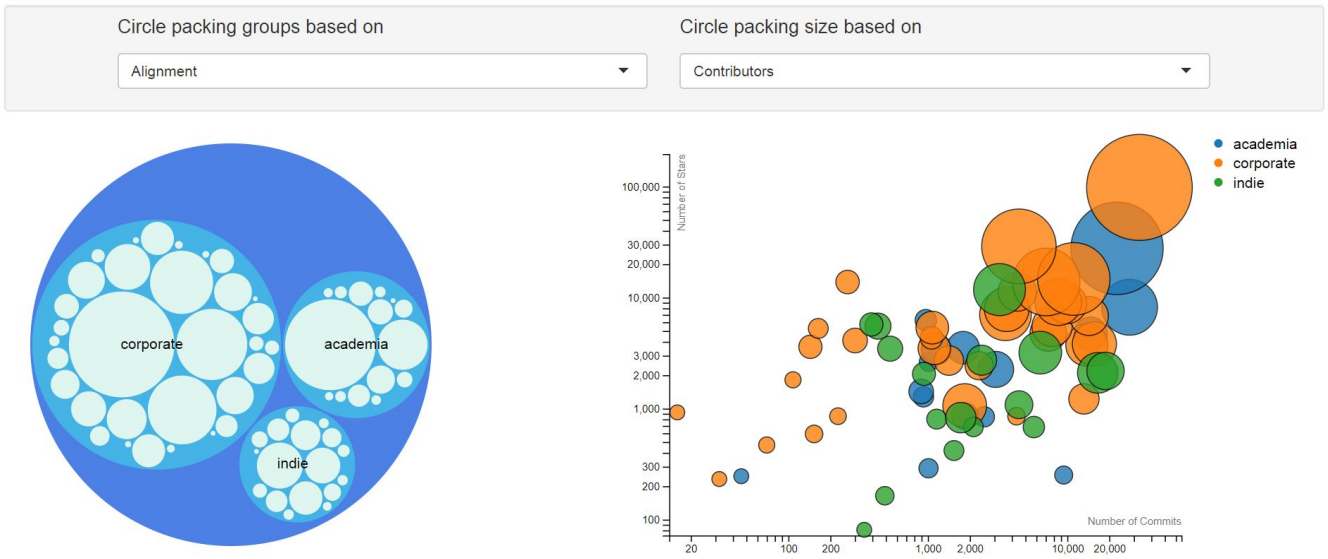

Figure 5

As we can see, in terms of stars and contributors, the corporate projects are clearly ahead, which means they are more popular and have larger community of contributors. In terms of commits, the difference is not very pronounced, which means that academic and indie projects are not far behind the corporate projects when it comes to maintaining an active project.
Also on the scatter plot, we see that the orange color for the corporate is dominating in terms of large circles and leaning more toward top-right, which is desirable.

Company Analysis

To show the impact of companies, we can select company for our circle packing grouping. Here, I will fix the circle size to stars only, which means we only look at the effect of companies on project popularity.

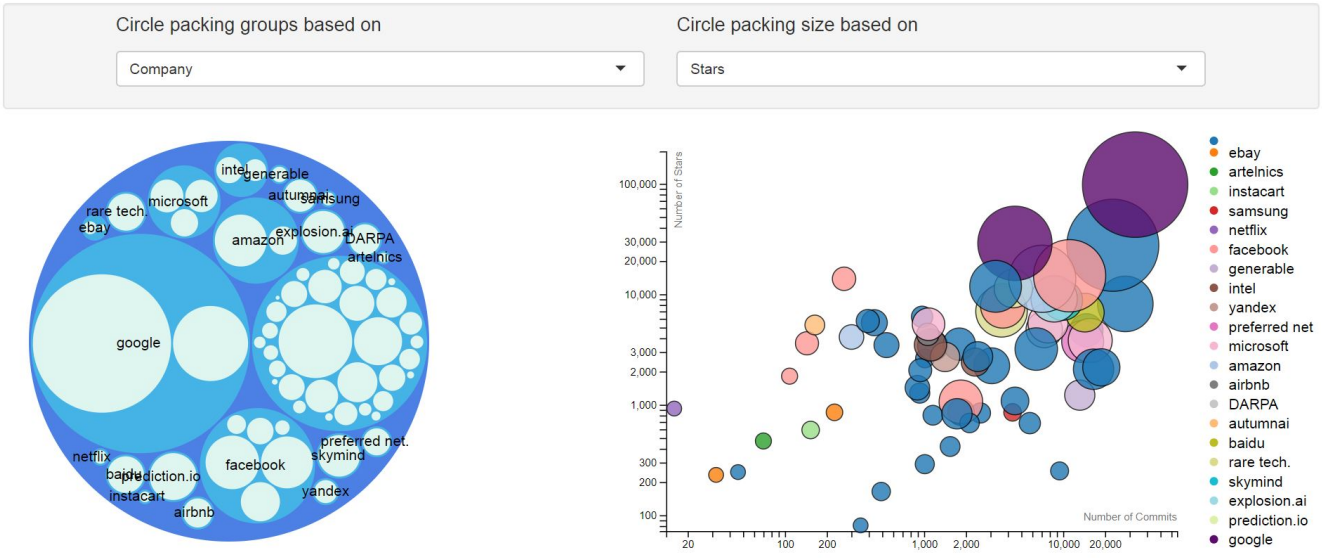

Figure 6

As we can see, Google, with only 2 projects has a really big impact. Behind it we also see Facebook, Amazon, and Microsoft, which are to be expected from these big companies, but Google seems to be far ahead in terms of popular projects. Also we can see the big unnamed circle on the right, which are all the projects that are not affiliated with any company, and they are competing with Google in popularity. On the scatter plot, the two big purple circles of Google projects are very prominent.

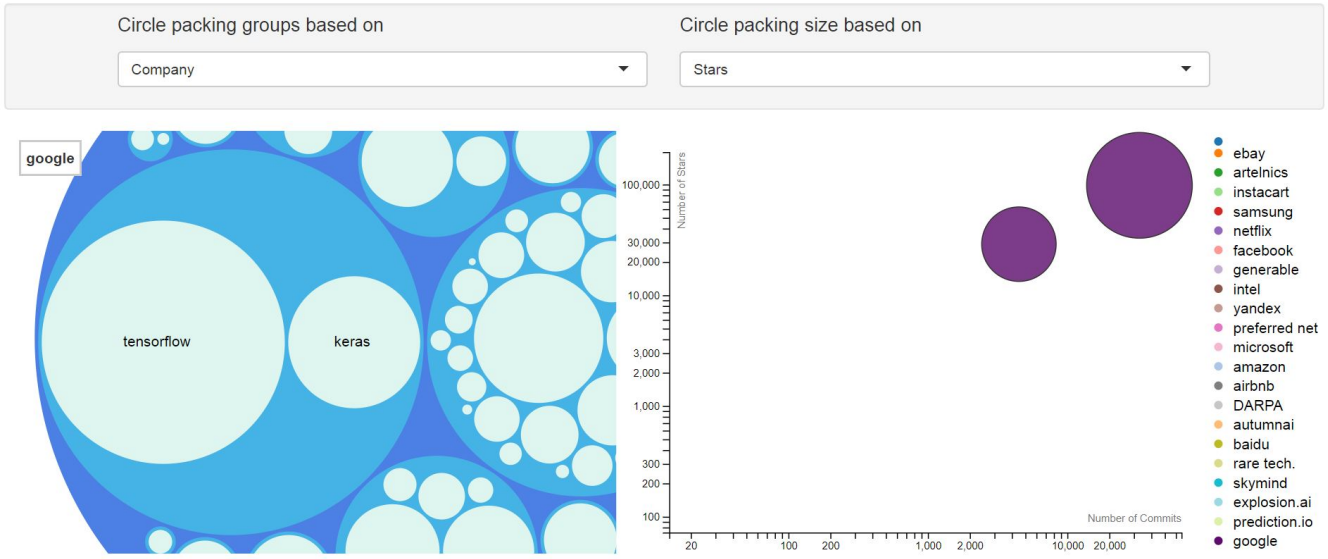

Figure 7

By zooming on Google circle, we see that these two popular projects of Google, are tensorflow and keras.

Year Analysis

To show the important years in development of popular open source machine learning projects, we can select year for our circle packing grouping. Again, I will fix the circle size to stars only, which means we only look at the project popularity.

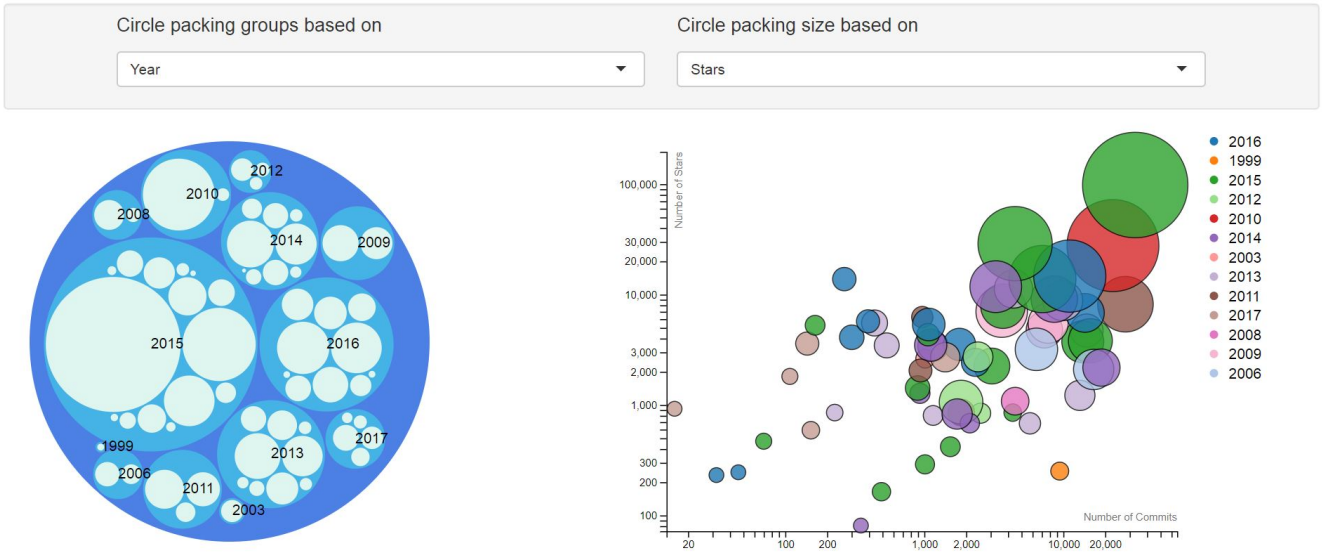

Figure 8

As we can see, the most important year in development of popular open source machine learning projects was 2015, followed by 2016.

Language Analysis

To show the important programming languages in developing popular open source machine learning projects, we can select language for our circle packing grouping. Again, I will fix the circle size to stars only, which means we only look at the project popularity.

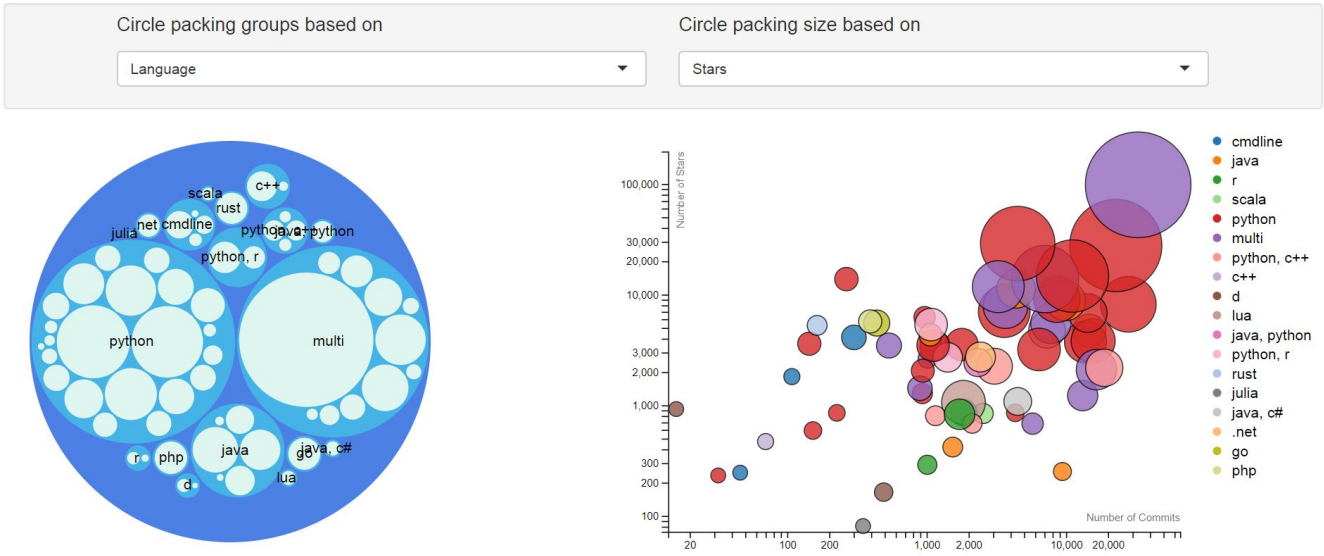

Figure 9

As we can see Python and multi are dominating languages. Here multi, indicate that multiple languages were used, so we can say that for a single language, Python is the most important programming language for developing open source machine learning projects.

Sources

I used the zoom functionality from here:

https://bl.ocks.org/davo/cd7261bd67581f284b6601fa4dd652b1

I also used “d3 SVG Legend” library for scatter plot legend:

https://d3-legend.susielu.com/