Storyboard A3: athern-mahirk

Data domain

We use the Higgs Twitter dataset from Stanford University in this visualization. The data consists of tweet information from before, during, and after the announcement of a new particle that shares many features with the Higgs boson. Messages from June 30th to July 7th are included in the dataset, and contains anonymized user id, tweet time, and whether it is a retweet, reply, or mention. The set also included follower social network information.

Interactive visualization application design

Visualization exploration

Our initial thought when considering different visualizations for this data was to show the frequency of tweets around the world and to find out if there was a geographic pattern with how tweets were spread. However, since all of the data is anonymized, we were unable to obtain location data from the tweets. We began our exploration using Tableau.

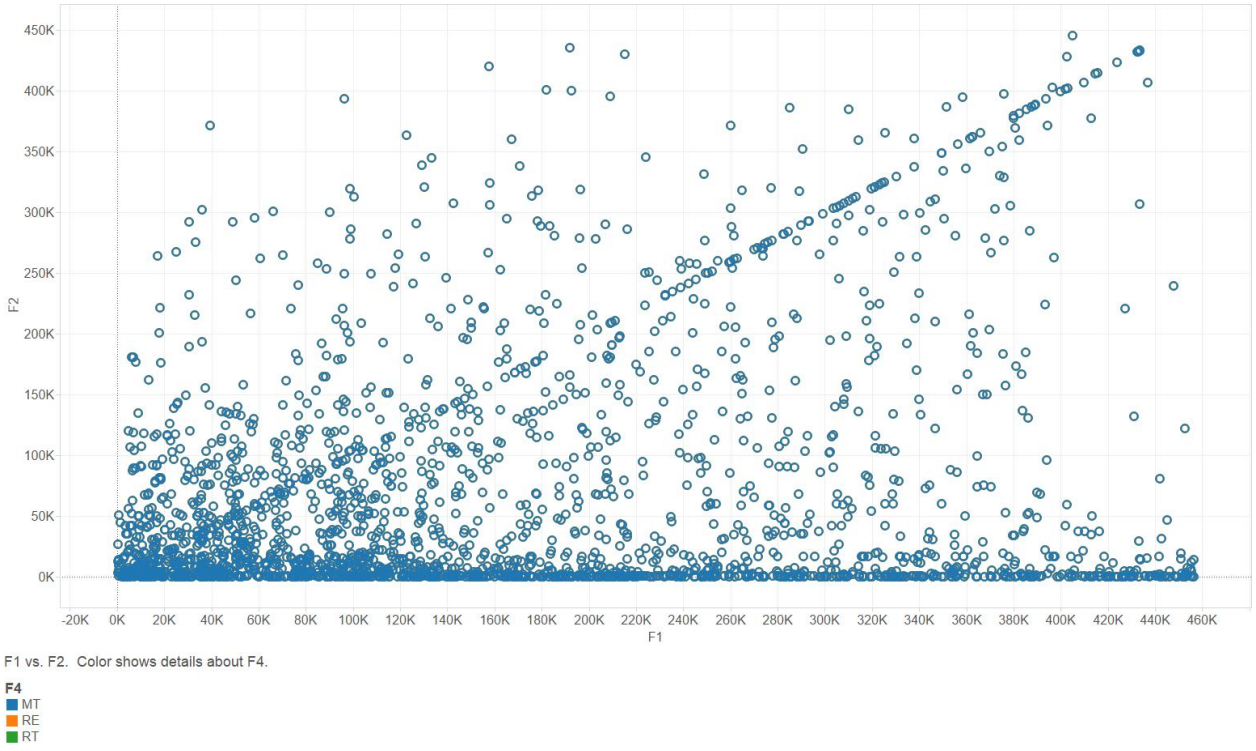

Figure 1: Interactions between users

This first visualization didn't prove to be very helpful. However, it did show us the enormity of the data we were analyzing. Trying to look at all of the data provided from the researchers at Stanford proved to be very hard for Tableau to handle, and it was difficult to produce meaningful visualizations of the entirety of the tweets collected. The spread in this image

seems to be mostly an artifact of the data anonymization, since certain lower numbered users seem to have more activity than higher numbered users, indicating that they may have been information sources.

We decided to narrow the data we looked at, and chose to primarily look at the activity time data. The data contains four periods of time: **Period I** is rumors before the July 2nd announcement, **Period II** is the announcement made on July 2nd indicating the mass of a Higgs particle, **Period III** is the span of time after July 2nd and before July 4th when there were rumors of a Higgs particle discovery, and **Period IV** is the time after the Higgs boson compatible particle discovery announcement on July 4th. We were curious to see how tweets would spike during these four periods, and what sort of tweets they would be.

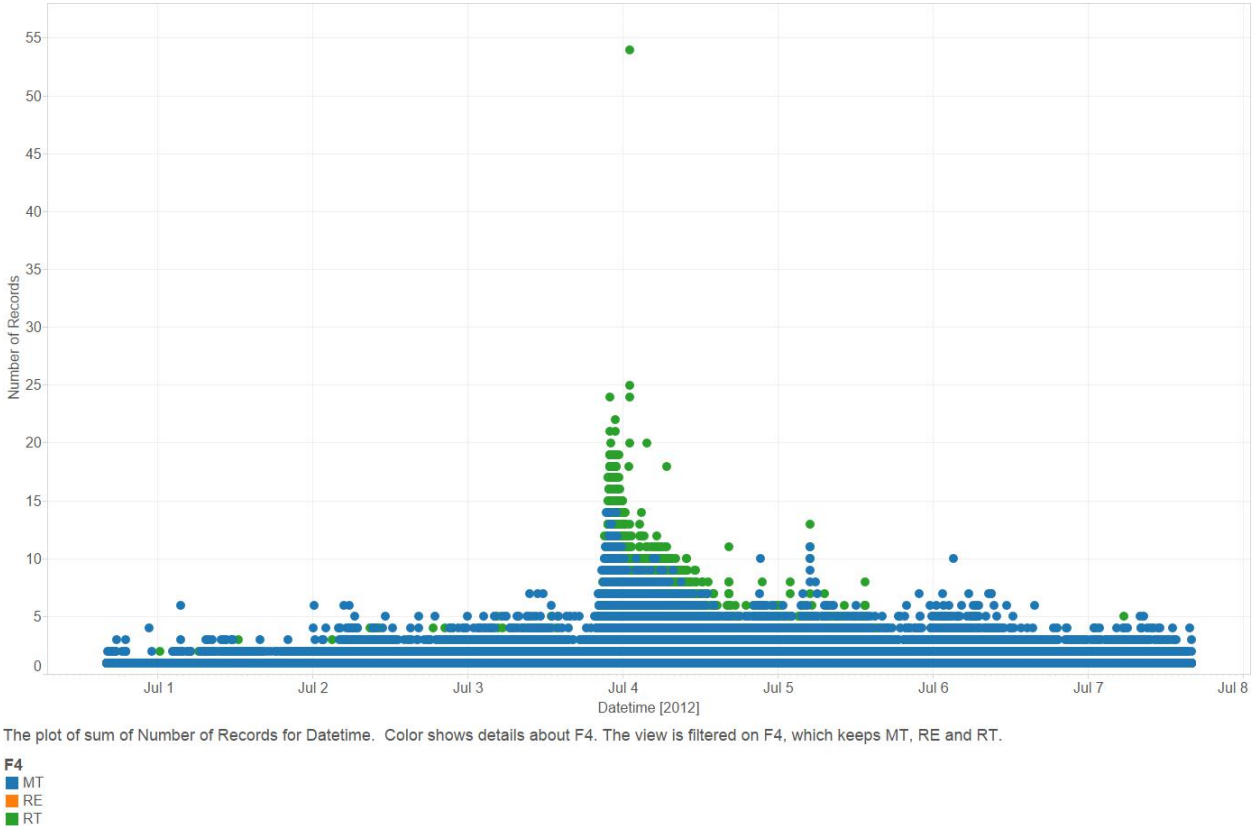

Figure 2. Number of different types of tweets over time

This visualization plotted tweets in exact time and was separated based on its activity type (retweet, mention, or reply). However, with the amount of data presented, there were many occlusions, especially with the mention data, which is made evident in Figure 3. However, even when removing the mention tweets, the graph still didn't seem very useful. It gave the false look of a stacked graph, almost, which is not what was actually occurring. Clearly, we wouldn't be able to plot every tweet. We would have to bin them instead.

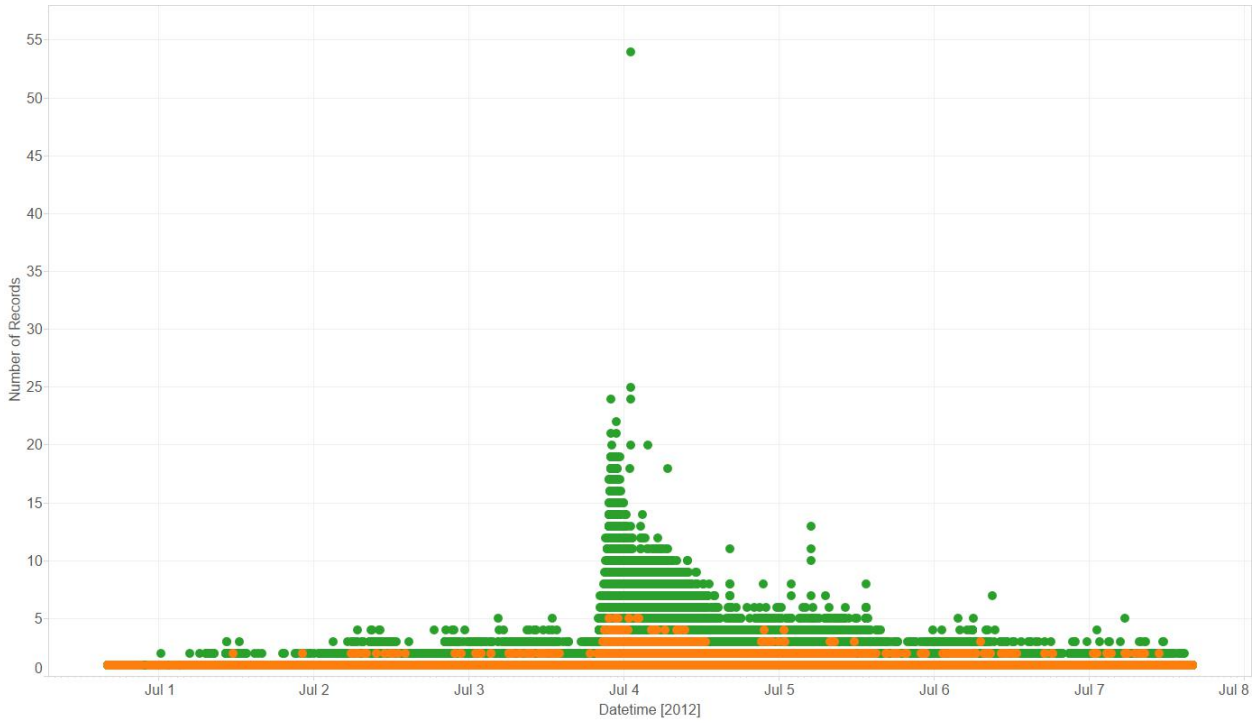

Figure 3. Number of tweets over time, no mentions

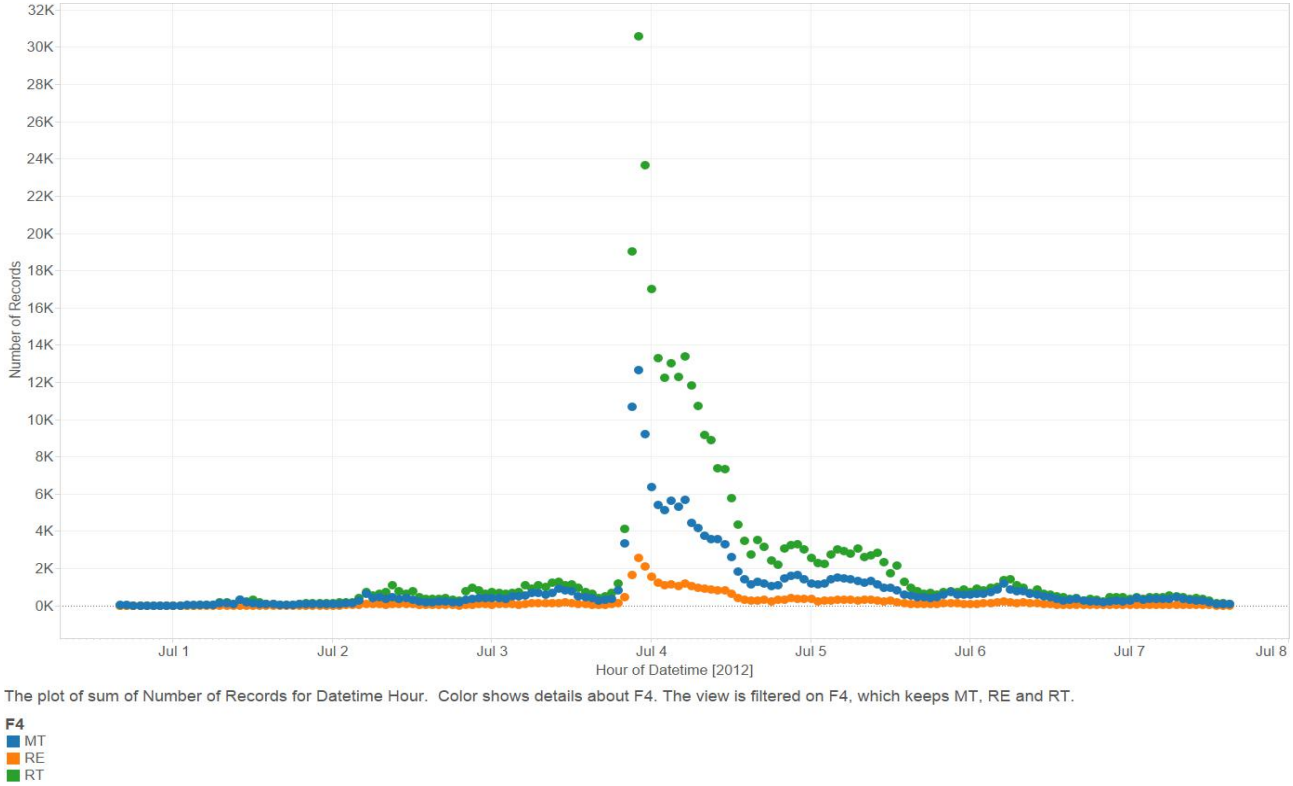

Figure 4. Number of different types of tweets by hour

Looking at the tweets by hour gave us a much better look at how the various periods affected the twitter community. The giant spike in tweets on July 4th is still as clear as in the previous visualizations, but we can also see smaller spikes on July 2nd in Period II. This was really interesting, but the immense amount of tweets on July 4th makes most of the rest of the timeline seem fairly static, even though there are some changes.

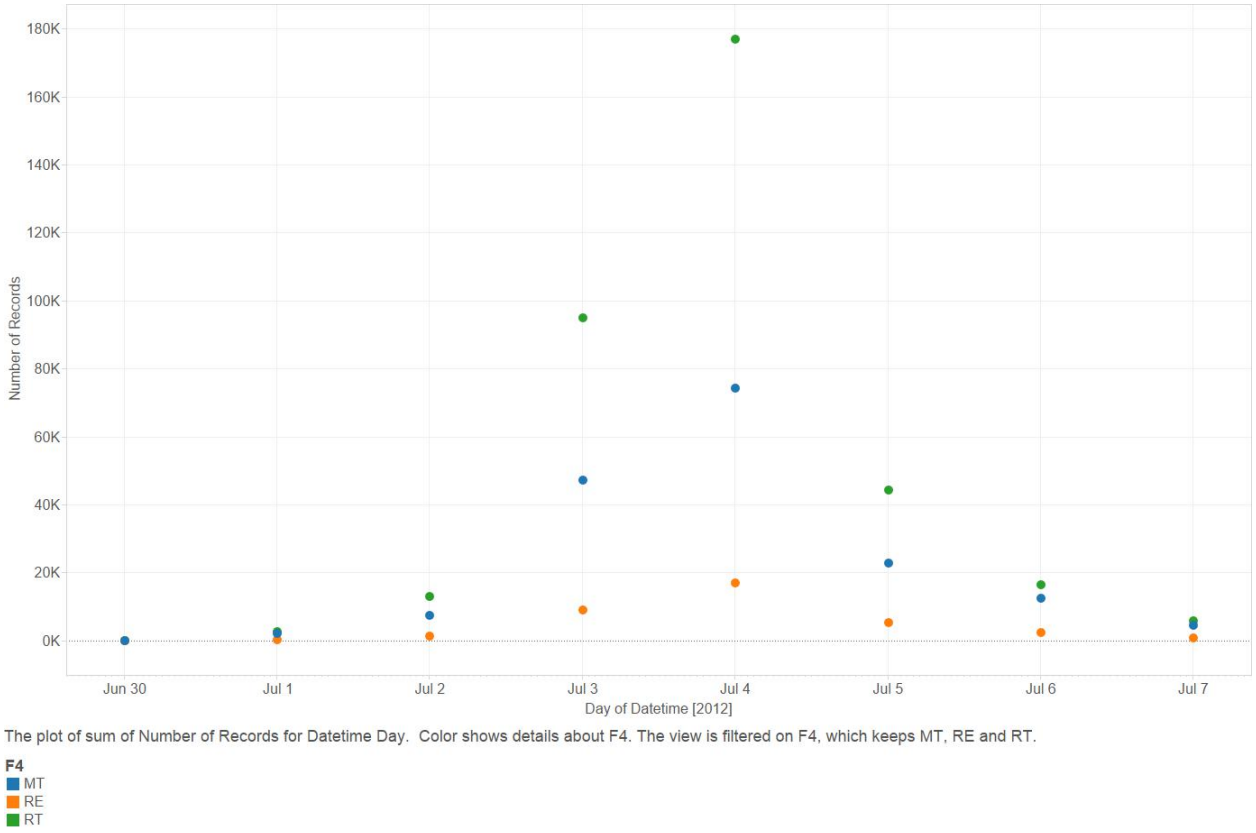

Figure 5. Number of different tweets by day

Binning the data down into day increments shows the change in tweets much better than the previous graph, even though there are less points. We lose some of the decrease in tweets some hours after the announcement on July 4th, but we gain a better insight on the surrounding days.It is much easier to see the spike in tweets on July 3rd, for example, which blended into the July 4th spike before. We can see that the rumors of the discovery of a Higgs like particle really increased the day before the big announcement. We also can see that the amount of tweets a few days after the announcement went down back to levels similar to Period II. This is the graph we decided to visualize.

Design

Now that we had picked the graph we wanted to visualize we had to answer questions regarding what style we would pick, in terms of the x and y axis scales, the colors/opacity we pick for the data points as well as the animation and transitions we pick for the points. Animations and transitions were ruled out as we believed they wouldn't help in viewing the

visualization. The axis that was pick ranged from 0 to 2000 for the y axis, so we could scale the number of tweets down by 100x, this helped scale the data better while still presenting a scale that is proportional to the scales presented from the Tableau mock-ups.
We kept the colors as a "*flat*" color variation of RGB with an opacity of 0.5 (50%), however after experimenting with the Spectrum chrome extension, we learnt that the colors may not work well for all types of color blindness, especially Deuteranope and its variations. The initial scheme for the data points is below:

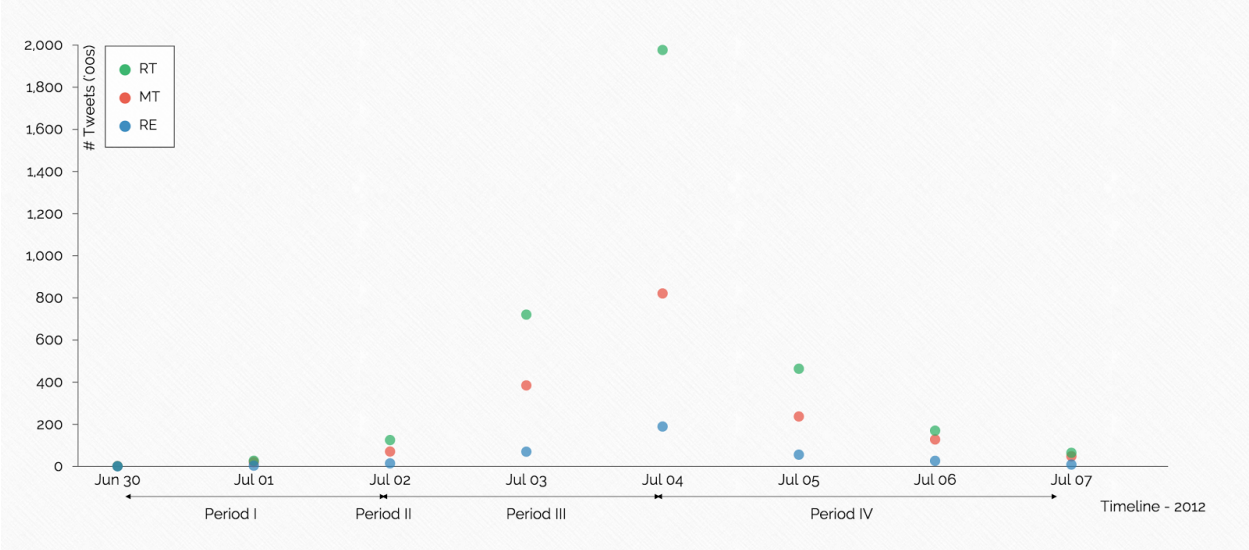

Figure 6: Original colors on the final visualization

Adding interaction

There are many different ways we could approach adding interactivity to this design. One interaction we plan to implement is allowing the user to toggle on and off the different types of tweets. This way, they can view only the trend of retweets over time if they choose. It is easy to look at the general trend of the data and not notice the unique aspects of each type of tweet. By looking at each type individually, it will be clearer how different periods beget different types of tweets.

We also want the user to be able to hover over each data point to see the exact number of tweets that that point represents. This will prevent some false assumptions that lower points, since they are much closer to 0, mean that practically no one was tweeting those days. It also makes it much easier to compare between the different types of tweets when their tweet counts are so close that the points overlap on the graph. Along with that, we decided to add interactivity to the labels below the x-axis, which showed which periods each section was, however we had not decided how we wanted to present the same.