# The Cuban Thaw Visualization Project

Process Book

Jerry Castro & Ivan Lima

CS 171, Harvard University

April 17, 2015

# Table of Contents

# Overview and Motivation

***Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.***

Our projects primary goal is to predict the impact of the the thawing of relations between the United States and Cuba on the 2016 Presidential Election. In order to achieve this we must successfully meet our smaller goals which are: to determine whether twitter users in each state are in favor of or opposed to the thaw and to determine which presidential candidates are in favor of and opposed to the thaw. By combining knowledge of each state's twitter users position on the thaw with that state's political leanings and then comparing that data with the stances of democratic and republican Presidential candidates on the thaw we can hopefully predict how the thaw will impact voting behavior.

Hours of attempting to find an event in the twitter era that we deemed historic and that would have an appropriate sample size of tweets brought us to the realization that the perfect topic was actually very recent. The restoration of diplomatic relations between the United States and Cuba on December 17, 2014.

This historic warming of relations between the US and Cuba was a monumental shift in foreign policy between two countries stuck in a bitter cold war for half of a century. The negotiations, surprisingly brokered by Pope Francis, made news on December 17th, 2014 and instantly captured everyone's attention. Republicans, Democrats, and politicians from around the world immediately put out statements to the media. Some were furious, others supportive. Many were cautious. Reports of the "Cuban Thaw" were all over the news. Footage from Miami, a heavily Cuban city, showed angry protesters in the streets holding up signs that said things like "This is treason" or "Obama is a traitor".

A pattern quickly emerged that continued throughout the day. The loudest opinions that were repeated over and over all over the media, belonged to people in middle age and old age who still vividly remember the Cold War and the Soviet Union and the constant threat of nuclear war. Some of those Miami protesters lived in Cuba during the revolution that brought Communism to the country and destroyed countless families.

But its been nearly 25 years since the end of the Cold War. Most young Americans today adults only know about Fidel Castro and the Soviet Union from history books, that is, if they know about them at all. Political analysts state that they really aren't sure how how this group of millions, the so called "millennials", feel about the embargo. Of course they don't. The answer is on Twitter, hidden inside a million tweets.

# Related Work

***Anything that inspired you, such as a paper, a web site, visualizations we discussed in class, etc.***

When we first started looking for a dataset to use for our project we struggled to find one that caught our attention. One night while browsing the web I stumbled on something both fascinating and heartbreaking. Wikileaks had released 500,000 pager intercepts from the day of 9/11. They showed reactions from that day in 5 minute intervals through peoples communications with their friends and family.

We ultimately decided that 9/11 was too dark and painful of a topic we realized that it was a good jumping off point for discovering a different topic. We were both intrigued by the idea of visualizing high volume social communication during a single major event. While it didn't exist during 9/11, we realized that twitter was the perfect medium for finding such data. Looking through hundreds of thousands of thousands of tweets on a single event one is sure to find fascinating trends.

## Questions

***What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?***

The two main questions we are trying to answer are: “How do millennials view the Thaw?” and “Can data visualization determine an election?”

Over the course of the project we realized that twitter was primarily used by millennials as the age of the average twitter user is 29 years old. So we realized that by figuring out whether the majority of tweets about the Thaw were favorable or unfavorable we would have a pretty accurate opinion of the millennial demographic. We also realized that we could contrast the overall opinion of a states tweets about the Thaw with the states political leanings.

We also realized that as the Thaw could be an important topic in the 2016 election we could potentially use tweets on the thaw to gather millennial opinion on that topic that when paired with knowledge of presidential candidates positions on the Thaw could potentially predict who millennials in each state might vote for. The millennial vote will have a significant impact on the election and therefore we realized that we may be able to use data visualization to predict the election.

Other questions considered were “Do the opinions Millennials match the politics of the state where they are from?” and “How do young Cuban-Americans feel about the policy?”

# Data

***Source, scraping method, cleanup, etc.***

The datasets described below cover all of the data requirements that we can predict at this time.

(1) “Sample Tweet with Metadata”: About 1,000,000 tweets from December 16 18 2014, each containing geographic data and the string “cuba”. Tweets come from around the world and were acquired through a historical tweet retriever service called Sifter. The raw data is structured, with metadata, in a CSV file.

(2) "US Presidential Voting Results 2012": Structured dataset on official US Presidential voting results from 2012 at the county level. That data is in CSV format.

(3) "US Demographics at County Level": Structured dataset from the US Census Bureau regarding US demographic data at the county level with respect to age, sex, gender, race, and hispanic origin. The data is from 2013 and in CSV format.

(4) "US Congress": Structured dataset containing political, biographical, and other properties for every current Congressional legislator (House & Senate).

(5) “Events since the Thaw”: A dataset on all the significant events since the Thaw on December 17th, 2014. “Events” include developments like Netflix and Mastercard operating in Cuba and the loosening of travel restrictions.

(6) “Political Positions of Politicians”: A dataset describing how every member of Congress and every 2016 Presidential candidate feels about the Cuban Thaw,

either for or against. These “feelings” will based on public statements. A possible “none” or “no comment” category may also be added.

Scraping method:

To scrape data off of the New York Times Api, given that it only returns a single page of your search query of 10 results per page, In order to get the full return we had to first return a single page to find out from results how many total other pages there were behind that one, and then we had to encapsulate our api call in a four loop over the total number of pages in order to bring down the full results.

Clean Up:

We did not expect much clean up at all. We figured the 1 million plus tweets would give us the answers we needed. Instead only 20,000 of the million tweets were actually geo-located. The rest of the locations were based off of the location section of the twitter profiles of those who wrote the tweets. The 20,000 tweets will be sorted by state and a favorable or unfavorable view of the Thaw and those views compared to state party representation.

# Exploratory Data Analysis

***What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?***

We looked at our data with a map and a table. We have not yet started the sentiment analysis because we are still writing the code for it.

Insight gained was that Florida was the state with the most tweets and that Hillary Clinton, the front runner for the Democratic Party nomination is in support of the thawed relations with Cuba while Jeb Bush and Scott Walker, the two main candidates for the Republican party nomination are both opposed.

This insight made sense given Florida's proximity to Cuba. This made us realize that tweet density would be a very useful tool in determining which areas had the strongest opinions regarding The Thaw and that we should have a map that shows tweet location over the period from December 16-December 18.

This insight about the views of potential Presidential Candidates toward the thaw influenced our design because it convinced us that we needed to mark each state by its 2012 voting results. Because the Democratic frontrunner is in favor of The Thaw while the Republican candidates are opposed, party affiliation clearly influences opinion on The Thaw. This will make it easier to make a prediction on how the Thaw will affect the election.

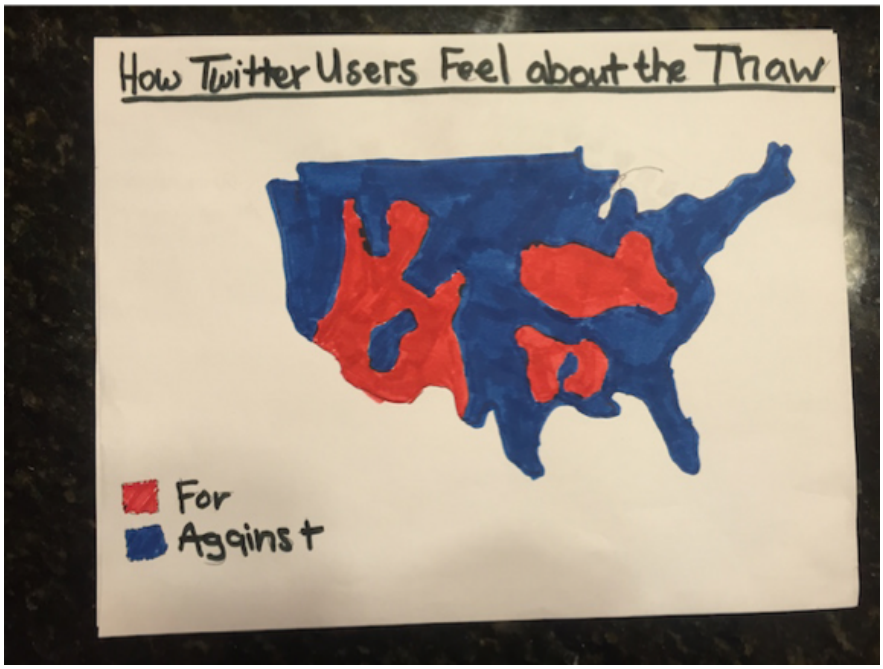
# Design Evolution

***What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course.***

We have not yet finalized visual design decisions but we do have some general ideas for how each major feature might look.

Our first visualization will be a “Tweet Map,” a map of the United States which will be color coded in correspondence with the geolocated sentiment analysis of tweets about Cuba collected on December 17 (the day of) and December 18 (the day after). This visualization will have two subviews. The first will be a comparison of the tweet map to Republican and Democrat voting patterns from the US Presidential Election of 2012. The second will be a comparison of the tweet map to the corresponding demographics in each area based on census information.

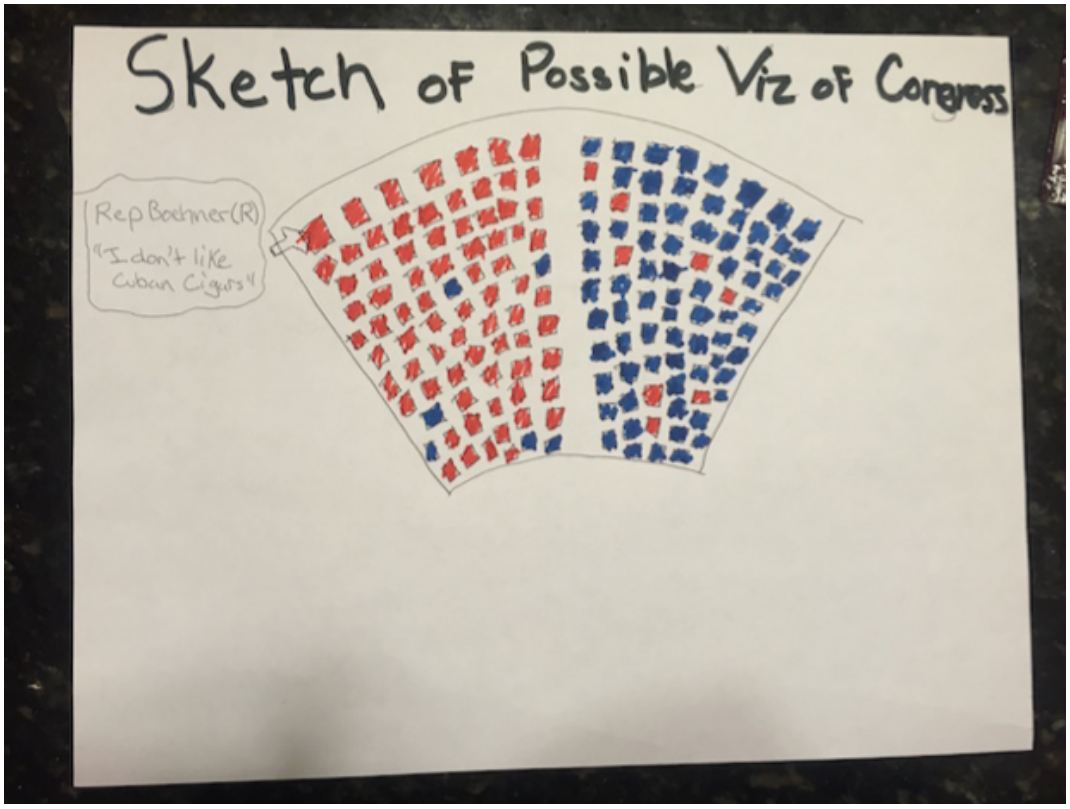
Here is a primitive sketch of our “Tweet Map” visualization:




➢ *A major visual decision yet to be made is how to visualize difference between two US maps. Solutions include a slider that shows to the left, the tweet map and to the right a map of the 2012 voting results. Another is have the user brush over the area to see the differences. Of course, we could just display two maps next to each other but this is less effective.*

Another visualization will be a visualization of the “Opinions of Congress," which will illustrate the stance of each member of Congress (House & Senate) as positive or negative, based on their public statements. If no public statement can be found for any of the candidates, then a third category will be included to reflect that candidates indifference. This visualization can be filtered by: Political Party, Up for Election in 2016, Not up for Election in 2016, Region (West, South, MidWest, Northeast), Age, Gender.

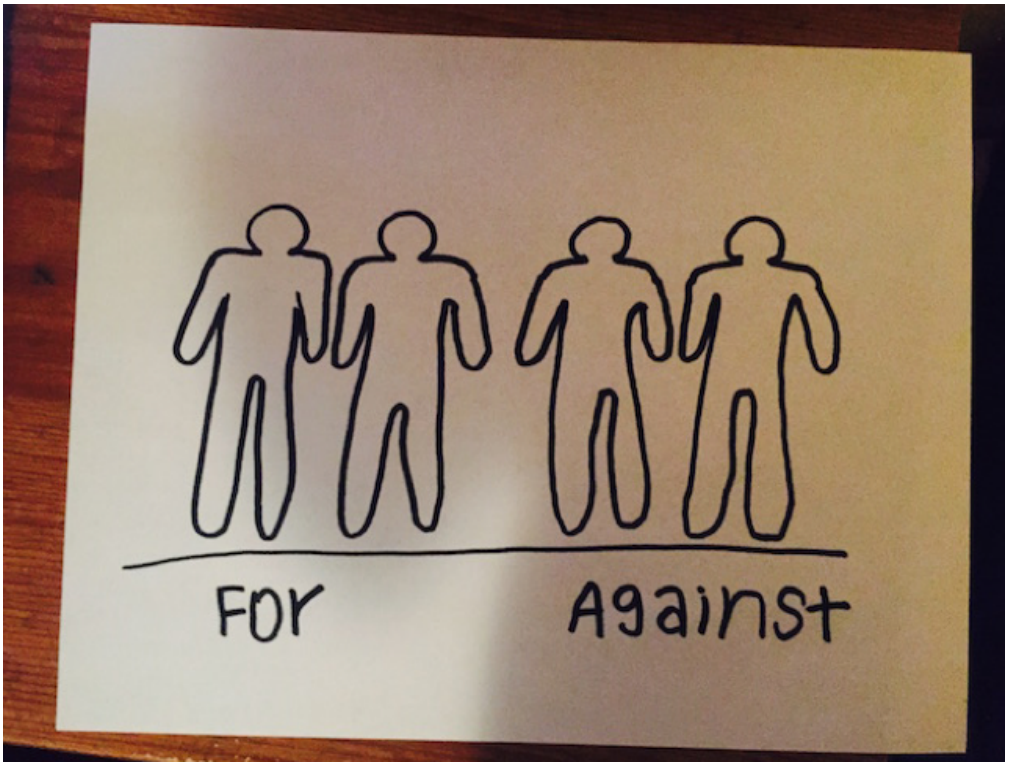
The following is a sketch of our initial idea for this visualization:



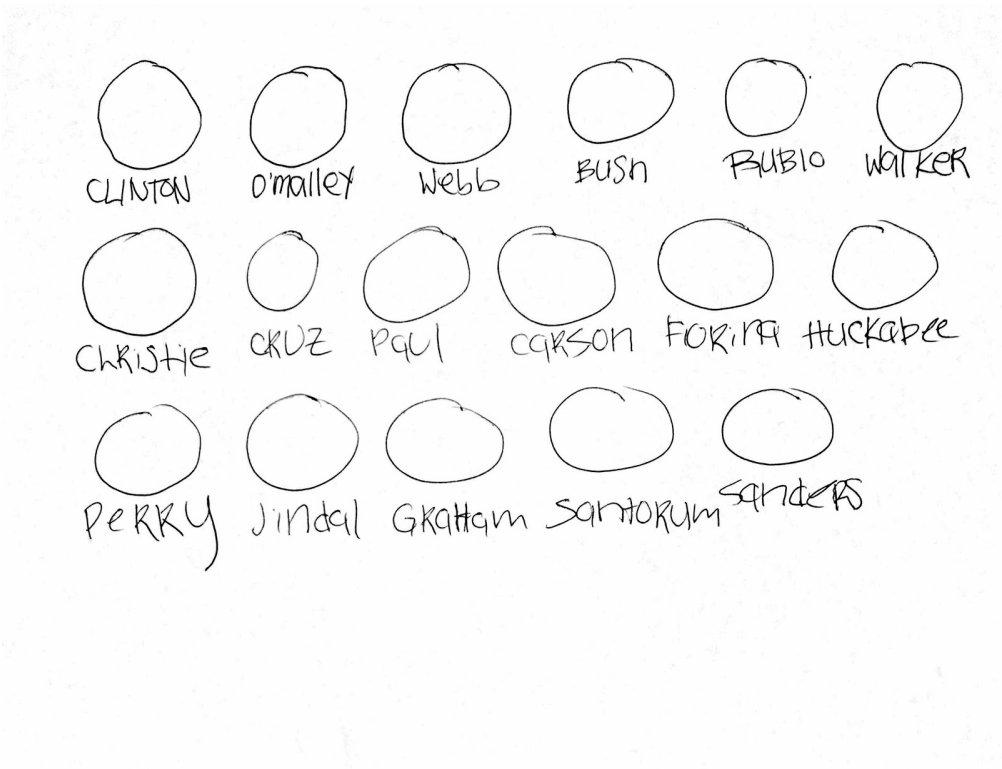

➢ *Again, we have not decided for sure what to do on this one. Many different visual ideas of this are possible. We could make simple SVG shapes to correspond to seats in Congress that change colors to indicate certain opinions. Of course, we can also just use a map which works well for Senators since there are only two per state, but some states have a double digit number of representatives in Congress.*

A third visualization that we will be using is a visualization of the “Opinions of 2016 Presidential Candidates.” Much like the congressional visualization, this visualization will depict how each of the 2016 Presidential candidates feels about the Cuban Thaw, either for or against, based on public statements. If no public statement can be found, then a third category of none (or similar), will also be included. This visualization can be filtered by: Political Party, Region (West, South, MidWest, Northeast), Age, Gender.

The following is a sketch of our initial idea for the “2016 Presidential Candidate” visualization. Our idea for this visualization has since evolved.




Here is a sketch of our current plan for this visualization:



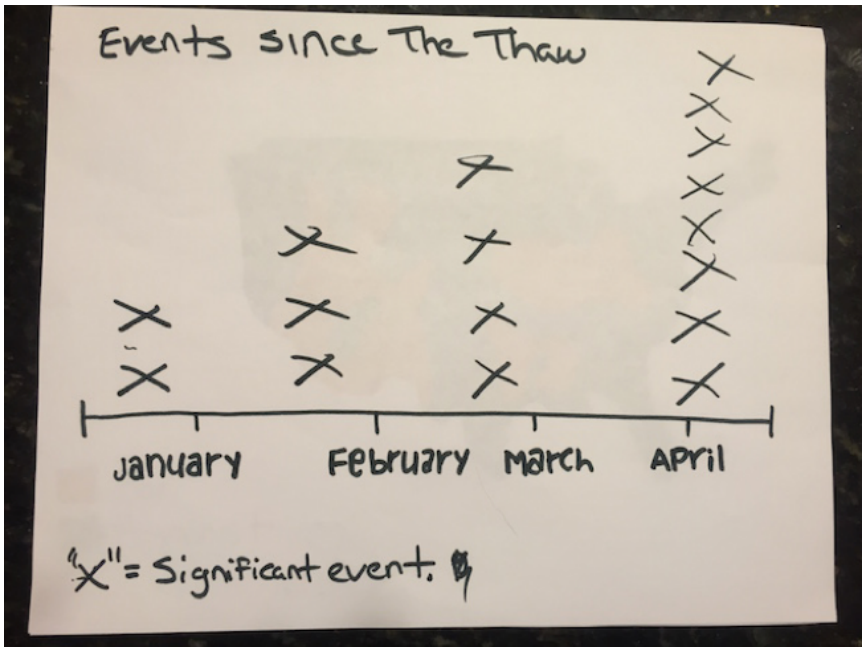

Each circle in the above drawing represents a campaign button like this one:

These buttons will be filtered by party, stance on the Cuban Thaw, region, age or gender.

Our next visualization will focus on the “Events since the Thaw,” and will show A 2D Brush on a timeline from December 17th to the present, visualizing the significant events that have transpired since the Thaw was announced. Such events include American businesses opening in Cuba, Conan O’brien hosting a special in Havana, Presidential executive orders, and pending legislation. This timeline will be on an ordinal scale with markings for each month. Data points will be stacked on top of each other during the month in which they occurred. This visualization will allow the user to brush

over a data point in order to see the data point's state before the thaw. Once the brush is removed, the data point will return to its current state.

Here is a sketch of our original plan for this visualization:

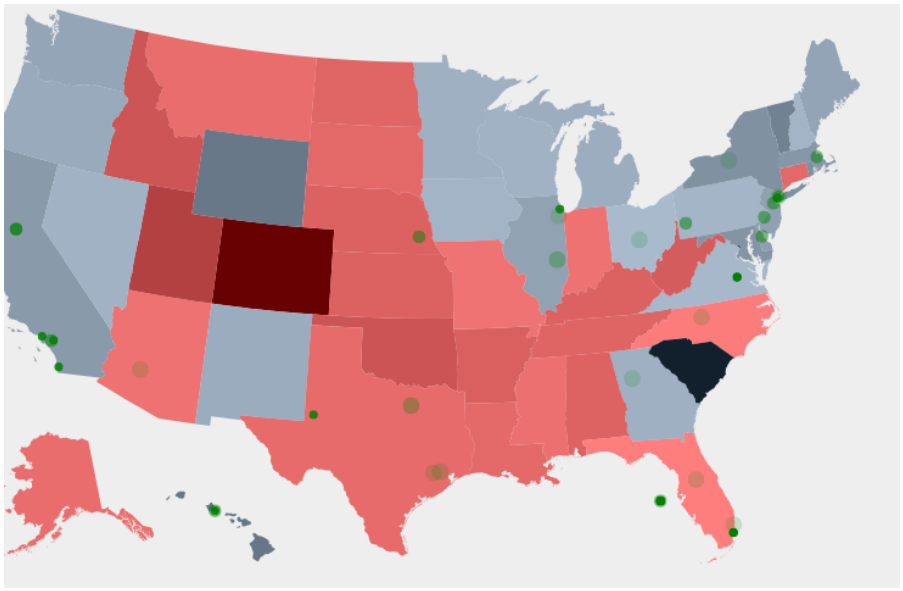


# Implementation

***Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.***

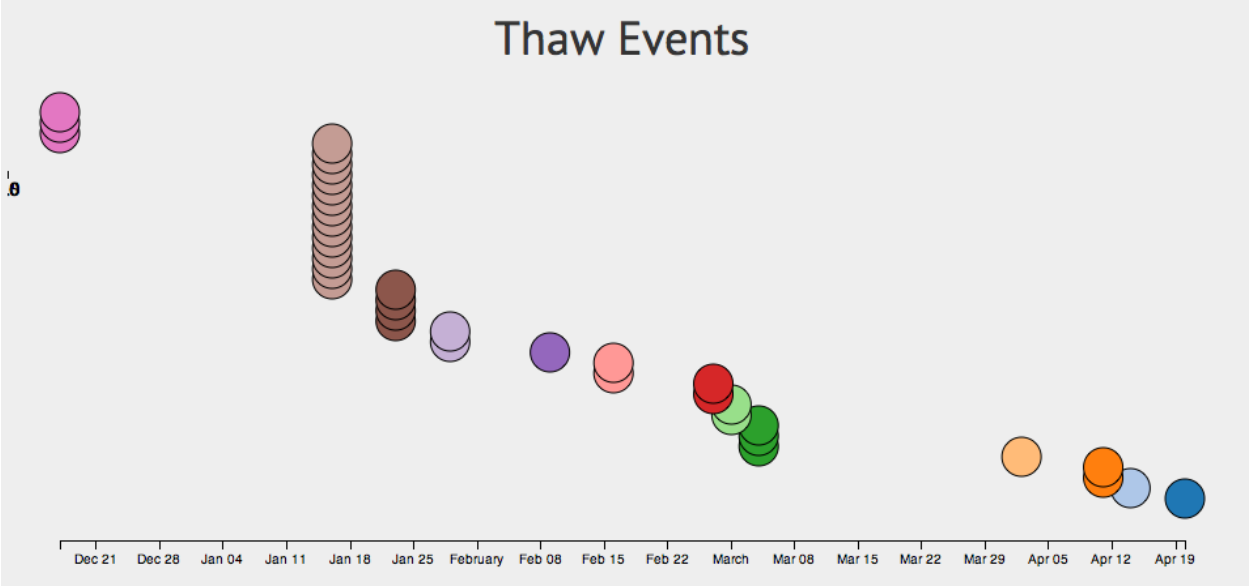
The first interactive visualization that we implemented on our website was the tweet map. The states are divided into blue and red states, according to how they voted in the Obama Romney election. When a user presses play, green marks appear to represent that somebody in that location has tweeted about the Cuban Thaw. When a user selects a state, the representatives from that state will appear on the right side of the screen. Future iterations will show the political positions of each member of congress regarding the new policies in Cuba.

The following is what this visualization currently looks like on our page:



We are currently working polishing up our “Tweet Map” visualization, as well as implementing our visualization of the “Opinions of the 2016 Presidential Candidates.”

We have also begun to implement our visualization of the events since the thaw, which currently looks like this:




When we are finished, the user will be able run the brush across the visualization in order to reveal photographs and a link to an article on The New York Times about the event.

# Evaluation

**What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?**

Currently, we haven't learned a lot by using our visualizations, however we have learned a lot by cleaning our visualizations. We have discovered that there is a lot of interest about Cuba internationally- possibly thanks in part to the population of Cuban cigars.

For being a couple weeks in we have a good grasp on where we are going but we can definitely see that there is a lot that we can do in the next two weeks. the bugs that we are currently experiencing will be resolved, giving way to a much more sophisticated system.