The Cuban Thaw Visualization Project

Website & Process Book

Jerry Castro & Ivan Lima

CS 171, Harvard University

May 4th, 2015

Table of Contents

Project Pivot

Our CS171 Final Project took a major pivot from its initial proposal but we strongly believe it more than makes up for it with technical rigor. Before diving more deeply into our final project, we think it is important to briefly explain why we pivoted.

Why the Pivot?

We initially proposed a thorough and rigorous visualization with four orthogonal views, each with several relevant filters, however our premier feature was a US map color coded based on a semantic analysis of the data that would determine whether or not a tweet was for or against the Cuban Thaw. Our TF was concerned that the project was too ambitious and that it might not be possible for us to complete this in time but In fact, that was not the issue. On a technical level, there was sufficient time to do everything we wanted to do.

Rather, as we got deeper into the semantic analysis aspect of the project towards the end of second week, it became apparent that the technology currently available could not provide us with highly reliable sentiment data without extensive optimizations that included custom dictionaries, grammars, and objects of speech. These were considerations that were beyond the time we had available so a change was required.

It must be noted though that this is not something we could have predicted in advance. Twitter's terms of service restricted us to downloading only 100,000 tweets per day, so the full dataset required 10 days to obtain. Furthermore, though our semantic analysis platform was extremely advanced and fully capable of analyzing tweets, it was also extremely costly. The cheapest option was $999 per month. So we had to apply for academic access. It took sometime to get approval but ultimately the relatively small number of academic credits we were granted meant a true test of viability would have to wait until a proper, representative sample from the full dataset could be obtained.

Overview and Motivation

Our projects primary goal is to visualize the Cuban Thaw through a massive dataset of tweets animated and projected with high velocity onto both a map of the world and a of the United States.

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 throughout 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.

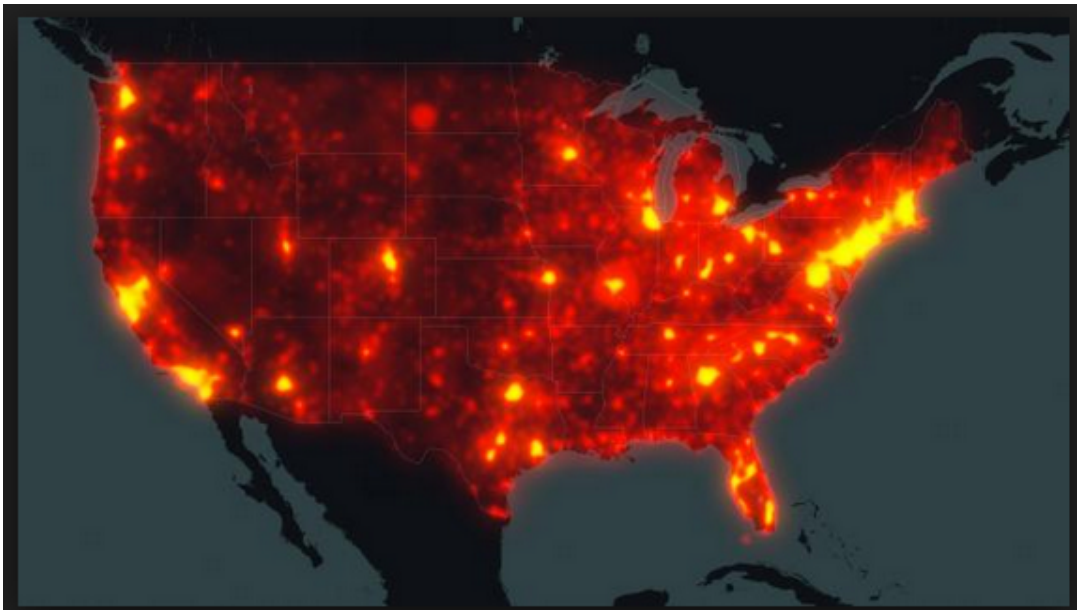
But it has been nearly 25 years since the end of the Cold War. Most American young adults today only know about Fidel Castro and the Soviet Union from history books, that is, if they know about them at all. Political scientists agree that they really aren't sure

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.

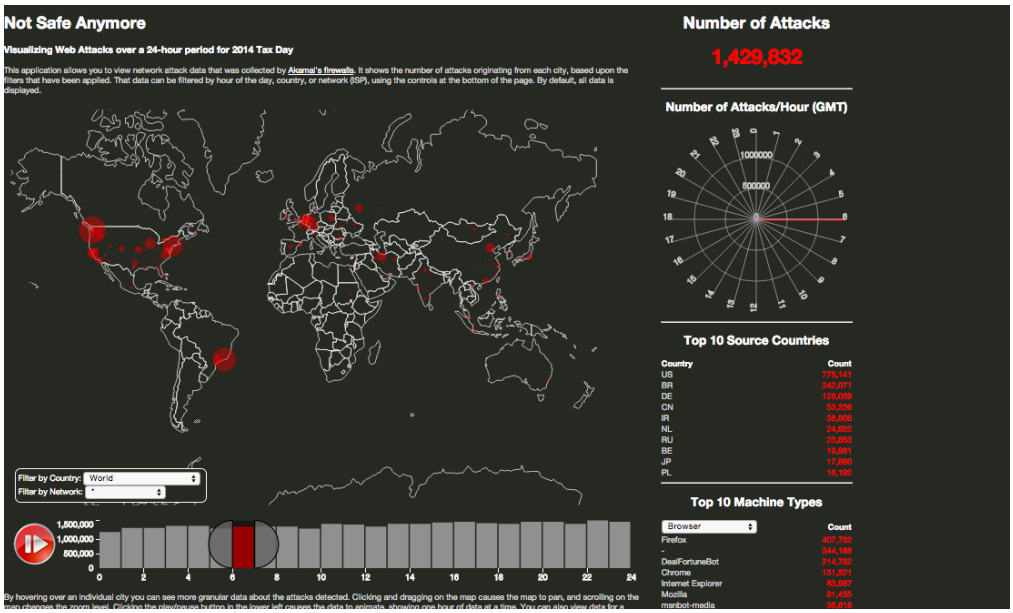
Related Work

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. During our search, the images below also served as inspiration.



This was our mental image of what graphing the Cuban Thaw tweets might look like.



This website gave us some general ideas about actual the implementation and layouts.

Questions

4/04/15, Project Starts

- ➢ Can we gauge public opinion of The Cuban Thaw through tweets?
- ➢ Do the opinions of Millennials match the political tendencies of the state where they are from?
- ➢ How do Young Cuban Americans really feel about the policy?
- ➢ How do the 2016 Presidential candidates feel about the Thaw?
- ➢ How do members of Congress feel about the Thaw?

4/18/15, Midpoint

- ➢ Can we gauge public opinion of The Cuban Thaw through tweets?
- ➢ How does tweet volume vary across the US?
- ➢ What proportion of Republicans in favor of the Thaw come from states with large agricultural industries? (Cuba imports 80% of its food)
- ➢ How do the 2016 Presidential candidates feel about the Thaw?
- ➢ How do members of Congress feel about the Thaw?
- ➢ Do the opinions of Millennials match the political tendencies of the state where they are from?

5/05/15, Project Due

- ➢ How does tweet volume vary across the World & the US?
- ➢ How important did the world find the event?
- ➢ Which countries found it *particularly* more important than others?
- ➢ Which countries appeared unphased?
- ➢ What countries weren't allowed to talk about it (no Twitter access)?
- ➢ How do the 2016 Presidential candidates feel about the Thaw?"

Initially, the project had an audacious goal. We were going to determine how the world, the nation and each state felt about the US's new policy towards Cuba by extracting a sentiment value from each record in a dataset of 1 Million tweets from the date of the announcement and then visualizing this dataset by projecting the tweets onto a map color coded with demographic information (ex. most hispanic/least hispanic counties) and election data (Republican vs Democrat counties). Not only would this include Twitter users, but also politicians.

Our bold and ambitious goal started to look less likely when we noticed that 92% of the data lacked the coordinate-based geo-data (lat/lng or polygon) that we thought we were getting. Our once bulging dataset went from 1,020,000 to ~9,000. But that didn't dampen our enthusiasm much because our map of the US still looked pretty cool and we had 9,000 tweets to use for the sentiment analysis.

Unfortunately, a little more than a week later, it became apparent that our goal was virtually impossible given the time limitations. A leader in semantic analysis and natural language processing, Semantria awarded us a grant of several thousand credits(api calls) to use for our project, but after dozens of hours of tweaking Semantria's sentiment engine with custom dictionaries and other natural language processing optimizations, it was clear that it would take weeks if not months to fully optimize Semantria's algorithms such that the accuracy of the results would be high enough to allow us to make bold political declarations in our visualization. Highly sarcastic and often tangential political comments are rather difficult for many humans to properly understand, much less a computer.

At the eleventh hour, nervous about the project and unsure of our ability to deliver something "awesome", we brainstormed for alternate ideas.

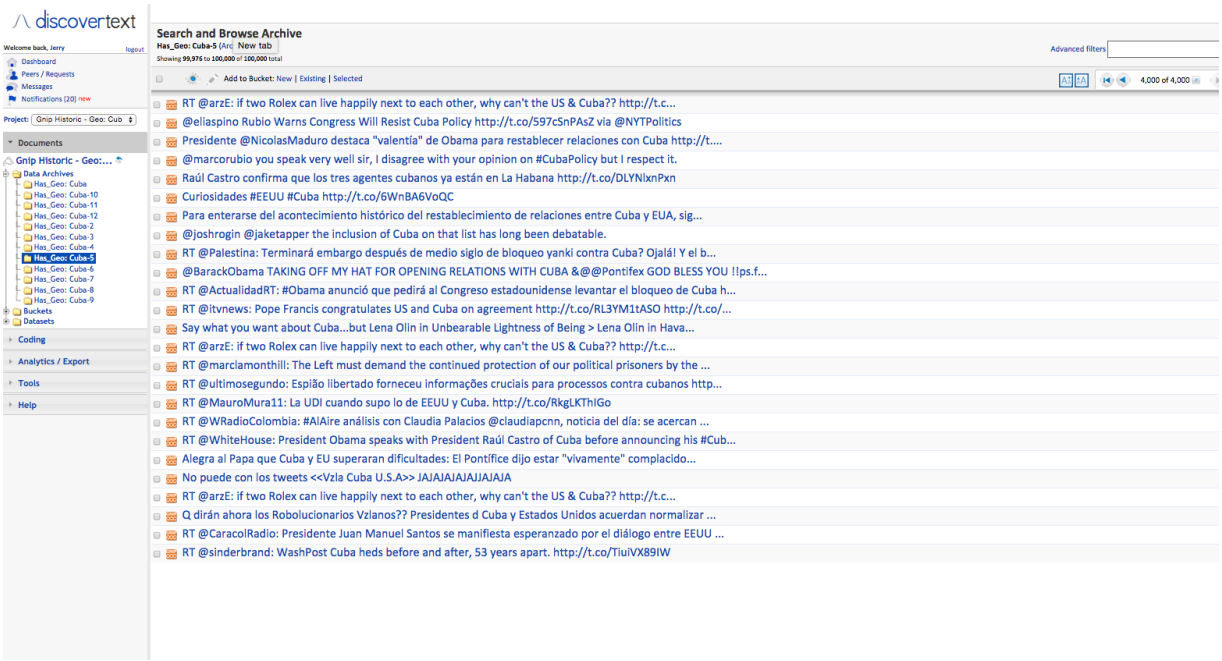
During our brainstorming we tumbled upon a couple tutorials that explain how to geo-code location strings. We got to work and after the usual slow start, we were successful at geo-code almost 990,000 records. Again we had our full dataset that spanned the world.

Our solution was to double down on Big Data and put the user's focus on a global tweet map that visualizes the tweets at 20 times faster than realtime while specifying also if the tweet is original, a retweet, or a favorited tweet. This was a doable but very ambitious goal on both a technical and visual level and thus met our target for an "awesome" project.

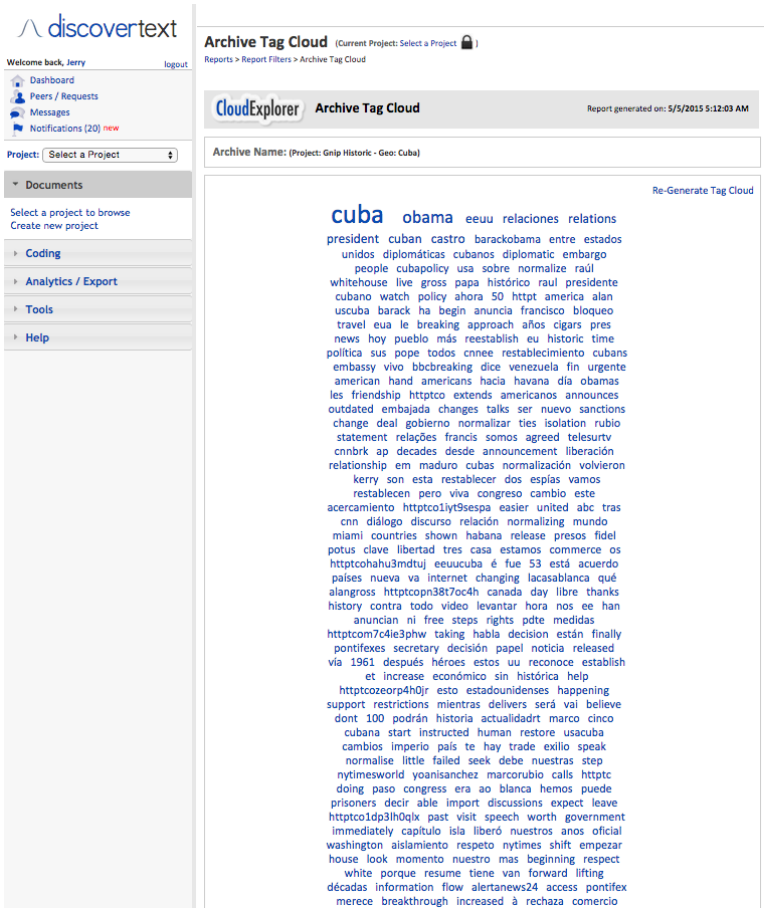
Data

Source, scraping method, cleanup, etc.

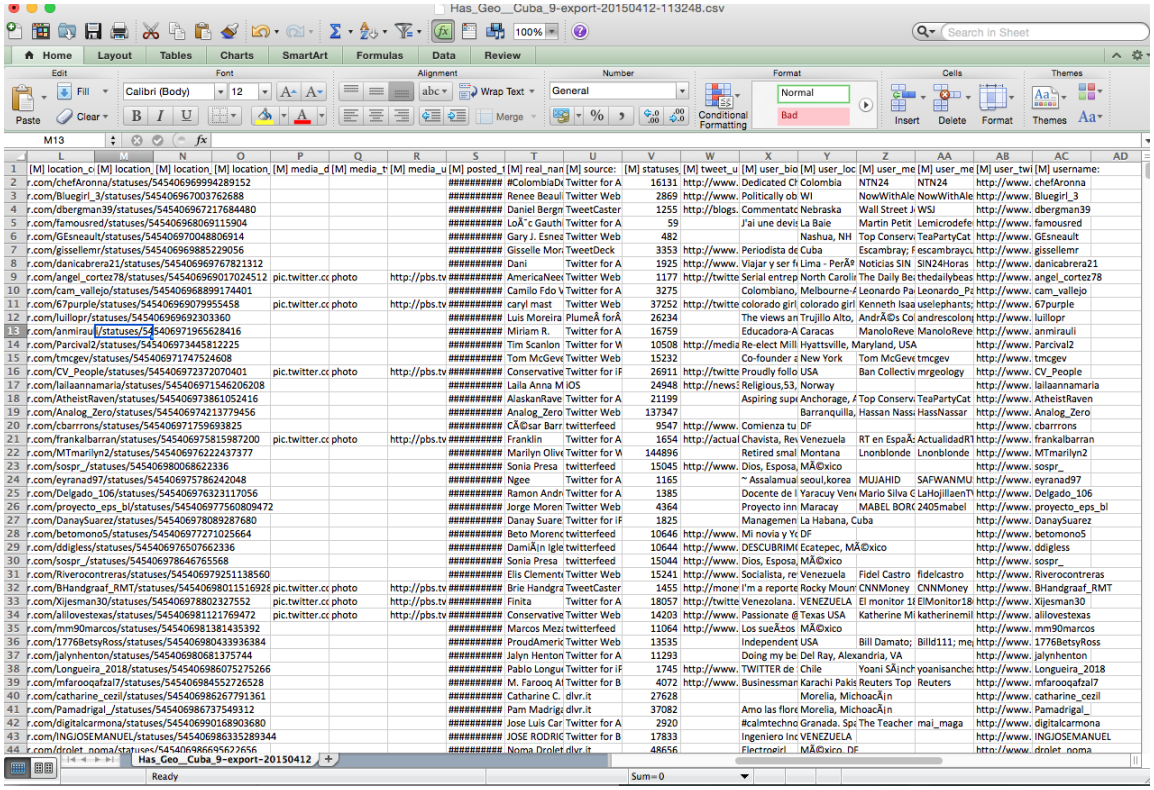
We acquired the data through DiscoverText, an approved distributor of Twitter data. Data consisted of 1,020,000 records each containing a tweet along with extensive metadata.



Data query in DiscoverText.com

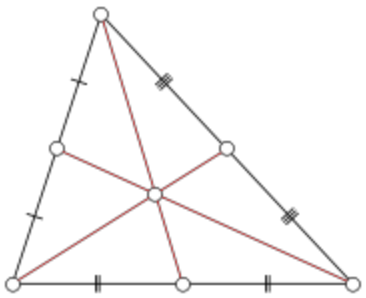


Discovertext had some basic word cloud features that gave us our first exploration into the data, albeit a very broad one.



Of 1 Million records, only 8,000 had exact coordinate data (geo-data in polygon type). The remaining records included only the self-reported location strings found in a user's profile. These strings often had spelling errors and sometimes they were completely nonsensical ("Neverland"). For most of the project, we wrote these records off and accepted them as useless data. And then, with slightly less than a week to go, we stumbled on "geocoding".

Geocoding turned our dataset from 8,000 to 1,000,000+. Geographic information was used with the maximum resolution available from the Twitter data stream. While we requested already-geocoded data from Twitter, surprisingly few tweets came back with legitimate location coordinates. When coordinates were available, they were typically of the "Polygon" style, describing not a point, but an area. Any Polygon location coordinates we collapsed to the geometric center



(http://en.wikipedia.org/wiki/Centroid) of the polygon using standard techniques. While Twitter often failed to provide precise location coordinates, in contrast it almost always provided each user's self-reported location, which can be thought of as each user's "home base." Examples include:

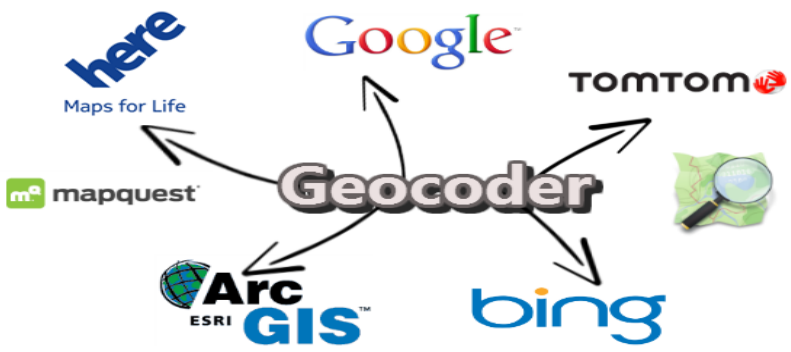
1. Cauquenes - Chile,
2. Boise, Idaho,
3. Johannesburg, SOUTH AFRICA,
4. Scotland.

While most are not precise to a street level, many are quite specific, giving state / regional location; some give an exact city, Others, however, are more whimsical:

1. planet tierra. o earth.
2. The Great State of Texas
3. North of where I came from
4. My underground lair.

Still others appear to be partial addresses, but are not sufficiently well-formed to be automatically converted into geolocations. For some of the missing data, a simple human edit was sufficient to reformat into a form that yielded a good geolocation; others simply had to be omitted as insufficiently geo-located / geo-locatable.

We then used the Python geocoder (https://pypi.python.org/pypi/geocoder) module to access an array of freely available web services provided by companies such as ArcGIS, Google, MapQuest, OpenCage, TomTom, Yahoo, and Yandex.



These services either transformed each textual location name / description into corresponding latitude and longitude coordinates, or reported failure in the attempt. In cases where no textual location was specified, or where the geocoder could not conclude the latitude/longitude from the text provided (occasionally with some manual assistance), we discarded the data point and did not display it on the map.

Because so many locations needed to be generated, it was import to use multiple geocoding services, so that we did not overload any one free service.

Once available to our web application in CSV and JSON formats, we used standard D3.js geographic projection functions (API-Reference#d3geo-geography) to map latitude and longitude into SVG display locations.

The size, color, and styling of display elements are dynamically chosen based on data attributes of each tweet. The radius of a tweet is log-proportional to the number of followers a tweeter has, giving an idea of the "reach" of the message. Whether a tweet is a retweet / retransmission of another message is also visually signaled (in yellow).

Tweet Timeline

Tweets are displayed according to the time that they appear in the Twitter data stream. Tweets may occasionally stop displaying, or appear to pause. This is not an issue with the visualization, but rather a reflection of the fact that the Twitter data stream (at least as exported to us) lacks records for certain seconds.

For example, for one data file, the stream contains records for the following seconds:

0-44,
59-539,
553-562,
568-599

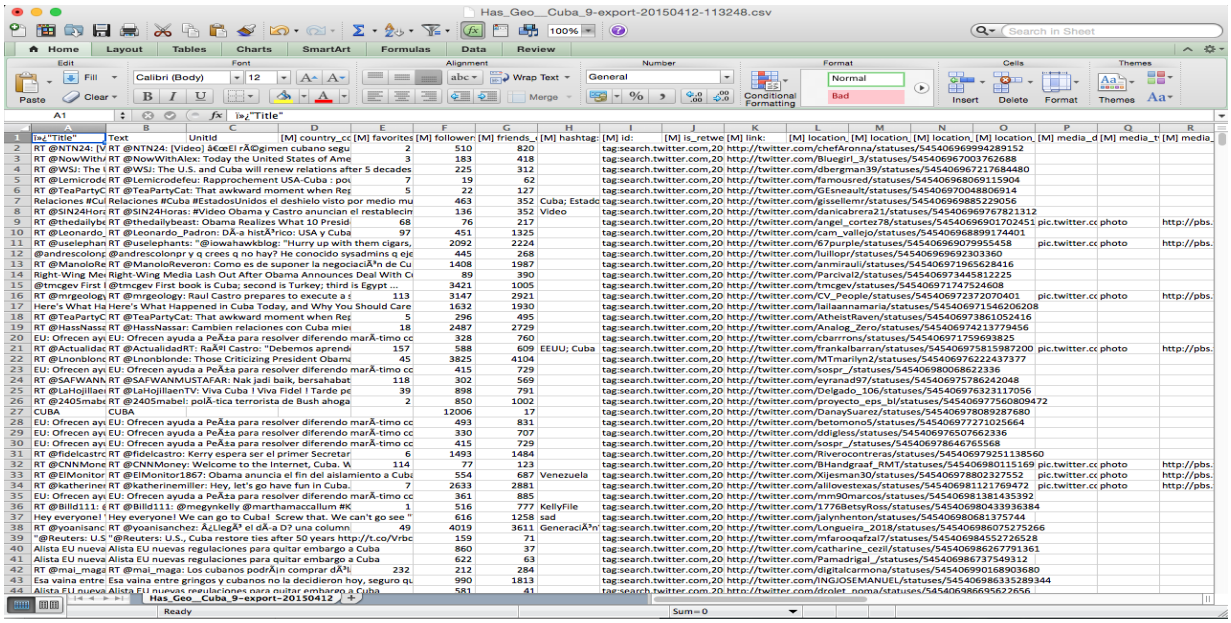
But lacks them for seconds:

45-58,
540-552,
563-567

Data Cleaning and Preparation Pipeline

The goal with any large data set is to have a large, useful, correct, and consistent set of data points. But big data is invariably "dirty." It contains errors, omissions, and inconsistencies. The process of preparing data for visualizations is similar to the "Extract, Transform, Load" (ETL) of the database community. In addition, data has to be prepared in such a way that it's easily consumed and used by the visualization process (which is usually running on a client device, often in the middle of an animation loop) where there is little to no opportunity for significant data cleanup.

While the sparkle and flash of graphical animations are often seen as the high point of visualization, just as important, if not more so, are the quality, quantity, relevance, and impact of the underlying data that is most important. The process of cleaning and structuring the data for visual display is the proverbial "rest of the iceberg" that lies beneath the waterline.



Data files were sourced from Twitter in CSV format.

Example cleanups include:

1. More accurately assess whether a tweet is a retweet or not. Twitter supposedly provides this information in a bespoke field, but even a cursory examination of the data shows it is often wrong. It neglects the text style retweet which starts "RT @userid". While Twitter may wish to consider only retweets using its (newer) native format to be proper retweets, its users clearly still prefer the old style.
2. Locations. Twitter supposedly geo-locatets tweets, but not very well or very often. When it goes to a location, it often does so as a very large geographical area (a polygon) which should be collapsed to a single point for plotting purposes. So we use publicly available geocoding services to translate self-reported user locations into geographical points.

We depend on both the accuracy of the user information (for both retweets and locations) and the accuracy of geocoding services (for locations). Realistically, there will be some errors in the resulting data, no matter how extensively we work to clean it up. Some users will mistype the conventional retweet marker (e.g. "RY @username"). Some will use non-standard, or at least non-English-standard, markers. Some will forget to mention they are copying others' content, or will use alternate quotation means (e.g. good old "quotes" and/or --attributions). When it comes to locations, some users don't provide accurate information, or use the field for metaphorical descriptions. The geocoders may misunderstand intent. *Et cetera.*

A benefit of working with large data sets is the Law of Large Numbers. Yes there may be some errors, but they will generally be overwhelmed by the correct data that is displayed. Humans are pretty good at discarding outliers. The second virtue is visualization, which provides a high-bandwidth mechanism for humans to interact with data. If things are out of kilter, they have a high propensity to notice when there are corresponding visual effects and outcomes--much more so than if the data variances were hidden in otherwise dense textual formats or statistical aggregations.

Human oversight is reasonably important in managing such data pipelines, since you want to incrementally improve the data. It often helps to have someone watching early failures to realize that "You know, a lot of these geocoding failures are on locations that end with a period--and that period does seem out of place here. What if, on locations that fail on the first attempt, we remove the final period and try again?" Or "A lot of people sure do want to make their country #USA a hashtag.

Maybe geocoders aren't savvy to that. If we see things that look like a hashtag, let's remove that and try again." Such rule-based cleanups dramatically improve data coding effectiveness.

Infrastructure

We used cloud servers to run the geocoding process, so that we could have multiple systems working on the problem at a time. We used 4 external servers at most times, bursting to 8 late in the process as our automation scripts became more mature.

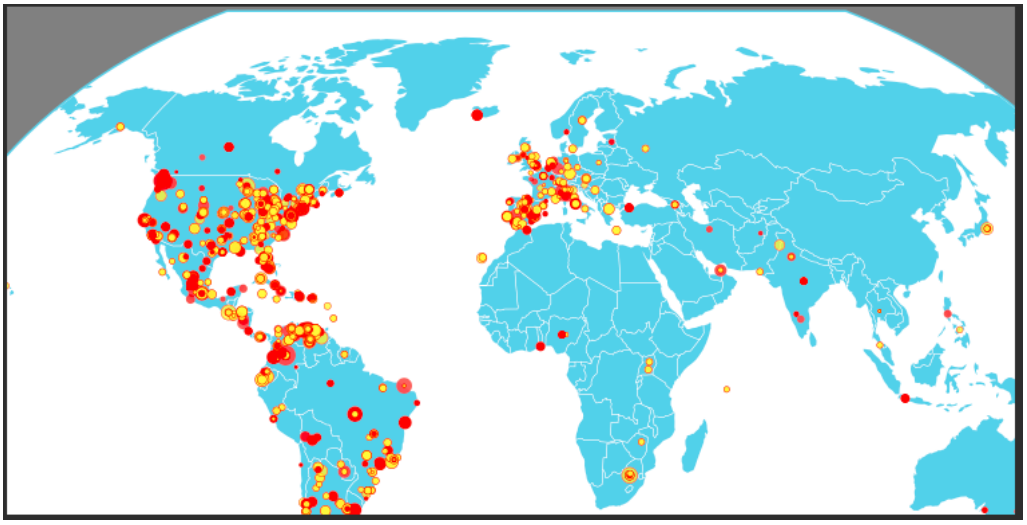
Subdividing the data for the multiple servers was a bit of a chore. We wrote some simple `bash` (Unix shell language) scripts to help automate it, but more work there would be good for dealing with large data sets.

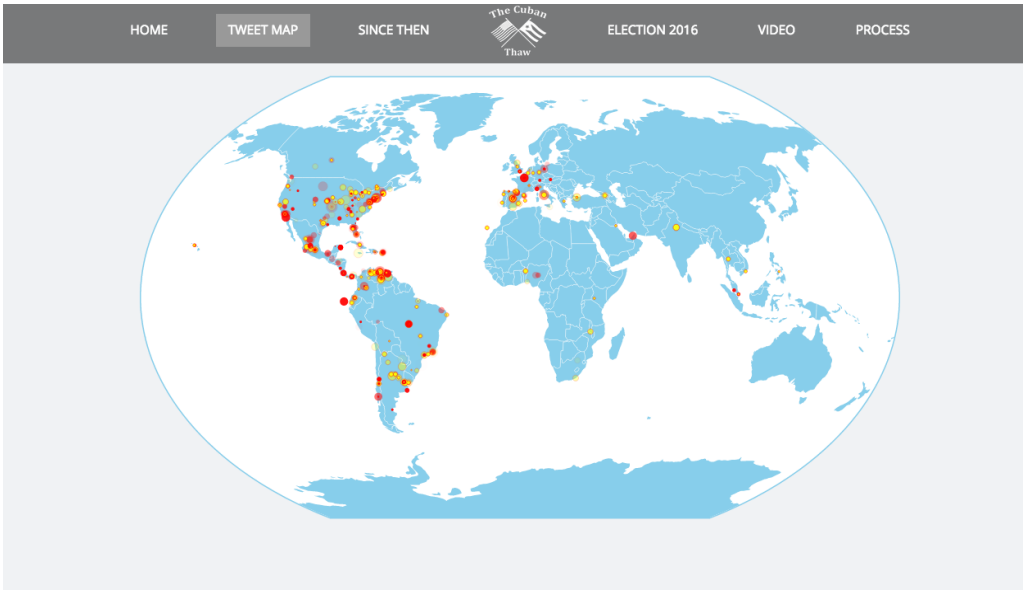
Spreading processing over multiple geocoding services was a net win, allowing us to process a very large number of requests and geocode almost a million tweets (where the original data only had a few thousand geocodes present). But dealing with multiple services introduces some variance. Google, MapQuest, and Google may not agree on the exact location of a place, for instance. They generally are very close--but it's not exact. Also, some geocoders can code some locations that others have trouble with. Geocoding automation was a huge win, but introduced its own complexities and a bit of variance in the resulting data.

Animation Engine

The map-based animation of tweet occurrence uses the D3.js framework to plot SVG shapes against a geographic background. D3.js handles most of the heavy lifting of mathematically converting from latitude and longitude onto various map projections and thence onto SVG display coordinates.

We use two map projections, a conic conformal projection for the US map, and a Kavrayskiy 7 pseudo-cylindrical for the world map. Choosing a map projection is a balance between optimizing multiple competing geometric properties and achieving pleasing visual aesthetics.





The primary animation routine is quite short, basically pulling in a second-by-second array of animatable events via JSON and displaying them according to various properties. Larger tweets reflect larger follower counts for the tweeter, thus a larger addressable audience. Retweeted content ("someone else said...") is displayed in a lighter color (yellow) to reflect the lesser originally. Purple borders are applied for tweets that are favorited, indicating enthusiasm of reception.

Keeping the tweet animation "fed" is a significant responsibility, since it plays at upwards of 20x real-time display rates. Thus, if there are 15 or 20 geocoded and animatable tweets in a given second--not an unreasonable estimate--there will be 300 to 400 new animation objects added per second. Animations last up to 2 seconds, considering their emergence and then incremental dissipation, so there can be easily be 600 to 800 animations in progress at any given instant. Feeding this requires an optimized JSON format, which is prepared by our backend data cleanup.

But a vast number of optimizations were required before we could start to visualize Big Data.

We expected clean, uniform, and categorized data but unfortunately twitter data is actually very messy. We spent equal or more time in data wrangling / semantic analysis than in visualization.

The tools for semantic analysis are many, but we went with Semantria because they are strong supporters of visualization education and data research.

Though Semantria is the industry leader in text analytics, our dataset proved too unruly to handle in the limited amount of time afforded to us to complete the project. That time was further reduced because we had to acquire and install windows 8 on a virtual machine in order to use Semantria's tools in Excel. A virtual Windows 8 environment cost us a lot of ram and our programs ran much slower and less predictable. In fact, loading the csv data files would often cause excel to crash. Furthermore, before we could run a viable semantic analysis of the data we had to do even more cleaning to remove special characters that gave Semantria trouble in addition to writing an N-gram text categorization algorithm to identify and sort the tweets by language in order to improve Semantria's accuracy when running a single configuration of the engine over a multi-language dataset.

All of this slowed our progress, but the fact that we only had a limited number of API calls meant, that we had to put serious thought into every batch and that realization reduced our agility.

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?

- For most of the project, we were unable to do proper data visualization in our exploration because the data set was orders of magnitude larger than anything we were prepared for. We also lacked consistent, reliable, and easily usable geoData. DiscoverText had basic exploration functionality that we used primarily in the first 7-10 of the project.
- Our main visualization for the first two weeks or so was a slice of tweets mapped against the US shaded to Obama Romney election results.
- We learned that the semantic analysis of political tweets is very, very hard. You must develop custom dictionaries and weights for a variety of different objects such as entities, documents, phrases in addition to ensuring robust language detection (n-gram categorization).
- Our realization that a proper semantic analysis would not be feasible led to the pivot to try and visualize all the of the tweets in an accelerated time frame.

TopMeta Discovery

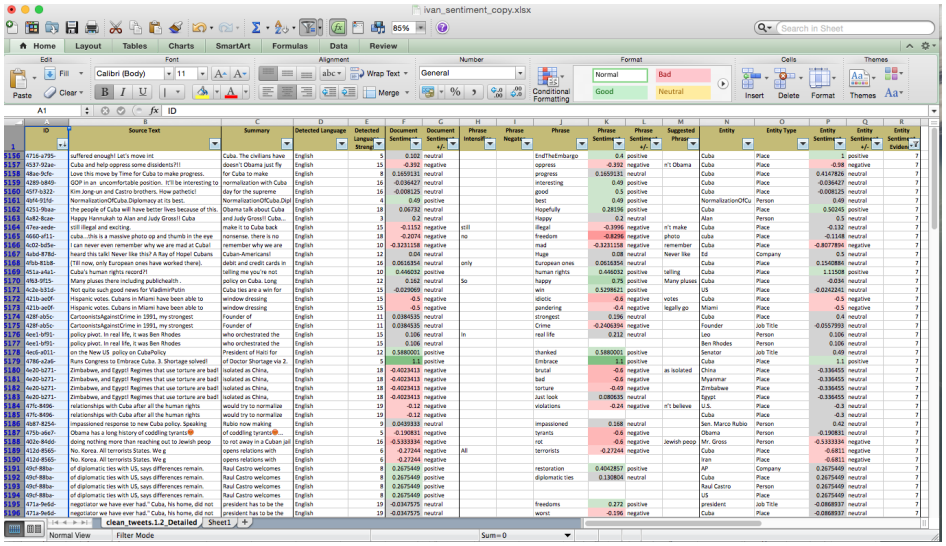
Top values for: **country_code:**

1 of 6

Meta Value	Total ▾	Filter
United States	746	
Venezuela	199	
Brasil	195	
España	128	
Chile	114	
México	93	
Colombia	91	
Argentina	85	
United Kingdom	51	
Canada	45	

Showing **1** to **10** of **57** total

DisoverText's exploration by country code

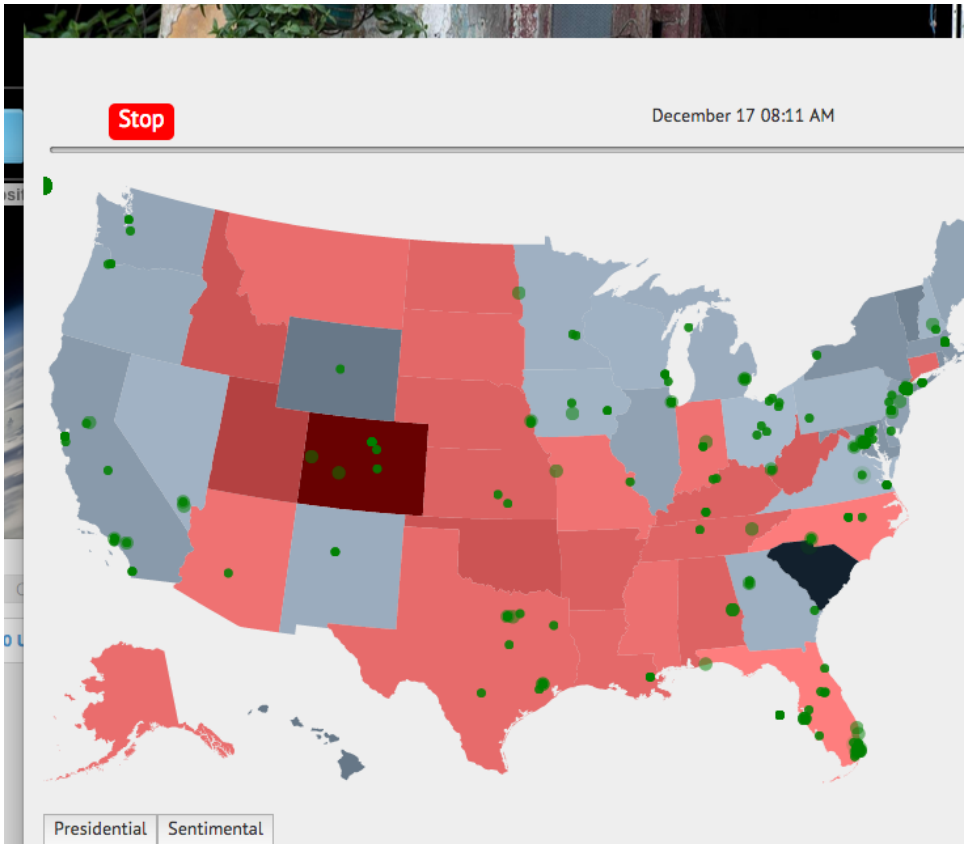


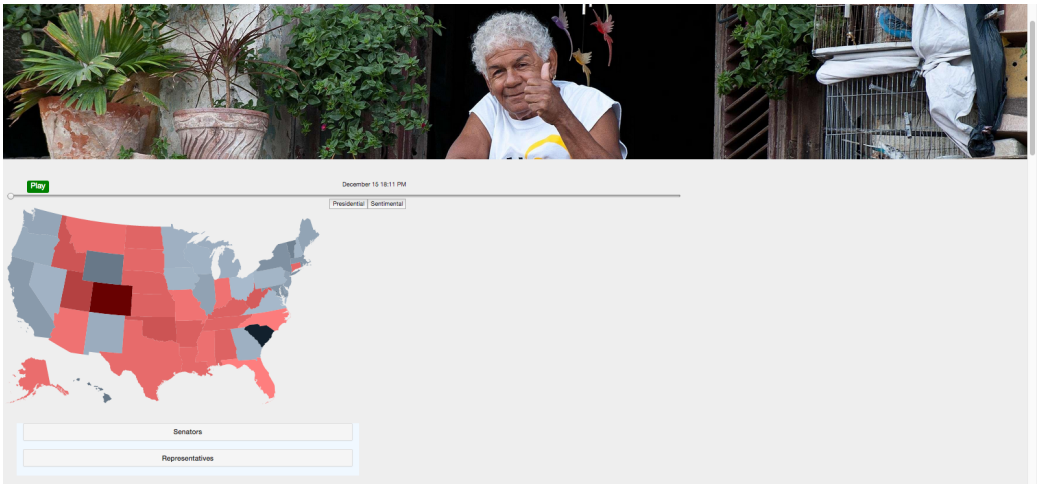
Semantria's tools also allowed us to see the common phrases, objets, languages, and emotions in the data.

Design Evolution:

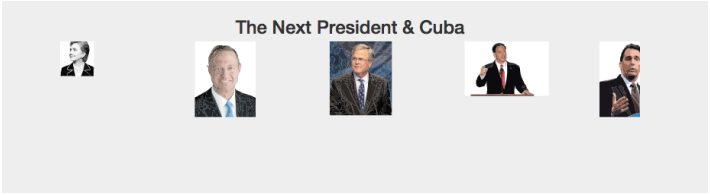
- We struggled choosing colors for tweets and maps. Ultimately we decided to keep the maps a single color and the tweets could be one of 3 color combinations. This decision was based on the fact that with tens of thousands of tweets populating the screen, multiple colors on multiple backgrounds would be too much.
- Deciding between greater precision and greater aesthetics. Because a small subset of the tweets used a rectangular area as its location, we decided on resolving this to a single lat/lng point because it improves aesthetics. The visual benefit is more important than a slight deviation in the integrity of the data.
- Deciding whether to aggregate and how to aggregate. We did not aggregate any tweets although we did do some aggregation of tweets over time. That is, if a group of tweets occurred extremely close to each other, we would display them as if they occurred at the same time. This improved performance considerably due to caching.
- Semantic analysis posed difficult aggregation questions. “If your visualization is correct on average but internally it is very volatile and wrong, is it a good visualization?” While attempting the semantic a
- Choosing a “One page” website design over a traditional multi-page design in terms was a decision that improved storytelling by forcing a linear path for the user’s eyes as they travel down the site.

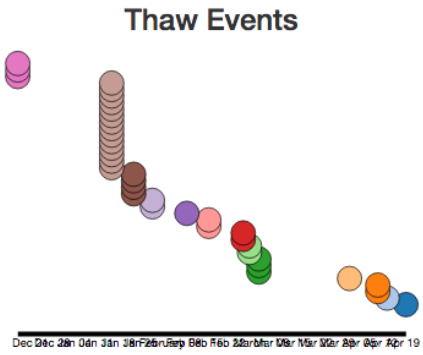
The image below was our main visualization for most of the project, a map of the US with using a small subset of the full dataset (8,000 vs 1,000,000+). Colors were meant to represent either states Obama won or Romney won, but the actual color is just a basic prototype choice.

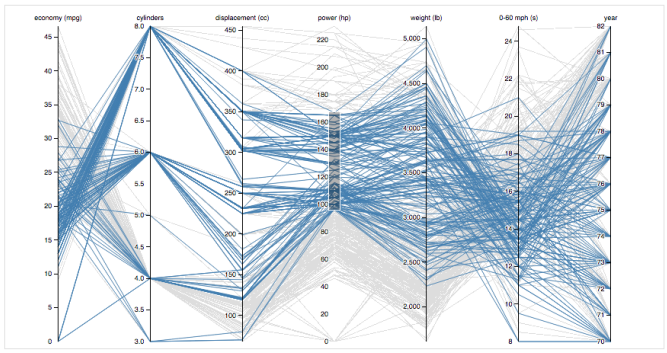


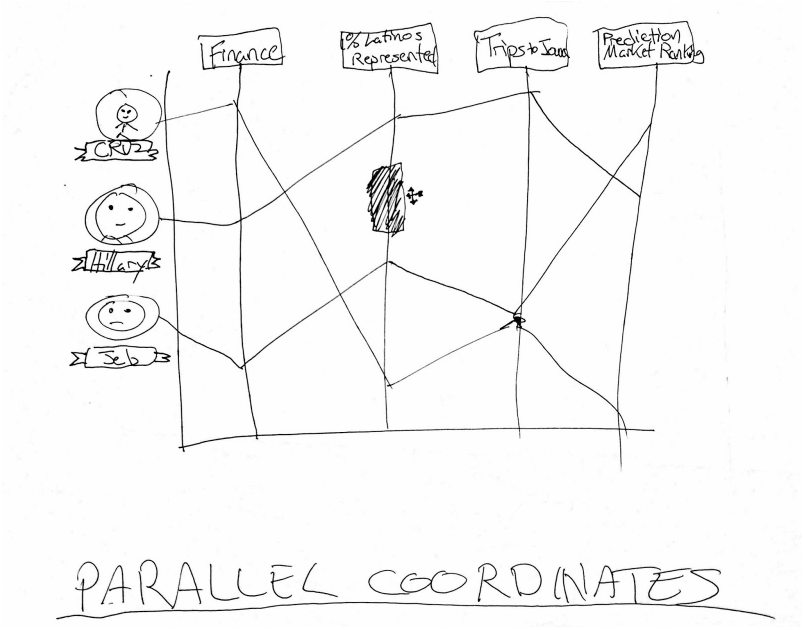


The other visualizations, while a part of the project, were more ornamental and ancillary to our main goals. Seen below in rudimentary form, the President and NYT API Brush visualizations still made it into our final project.







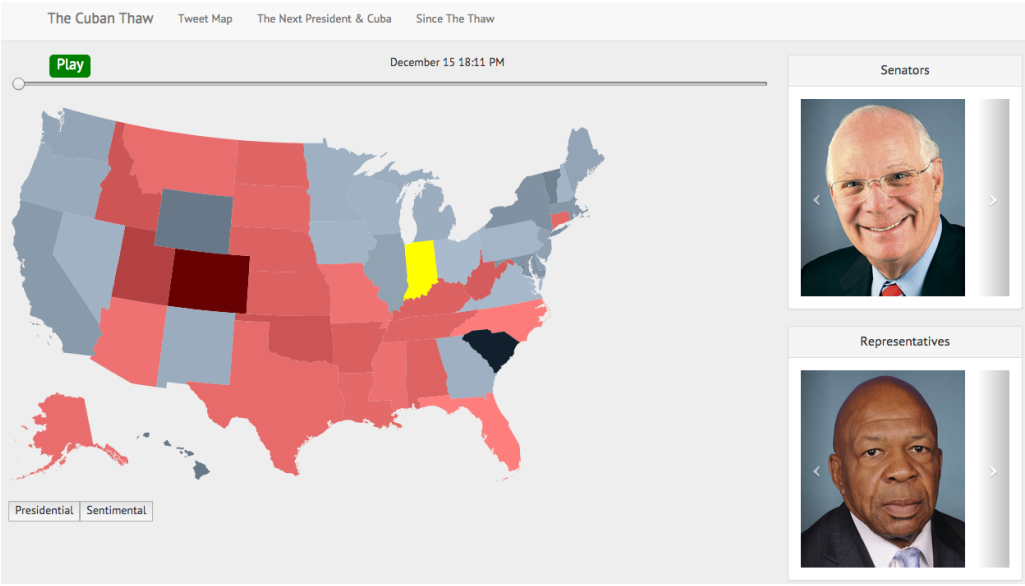


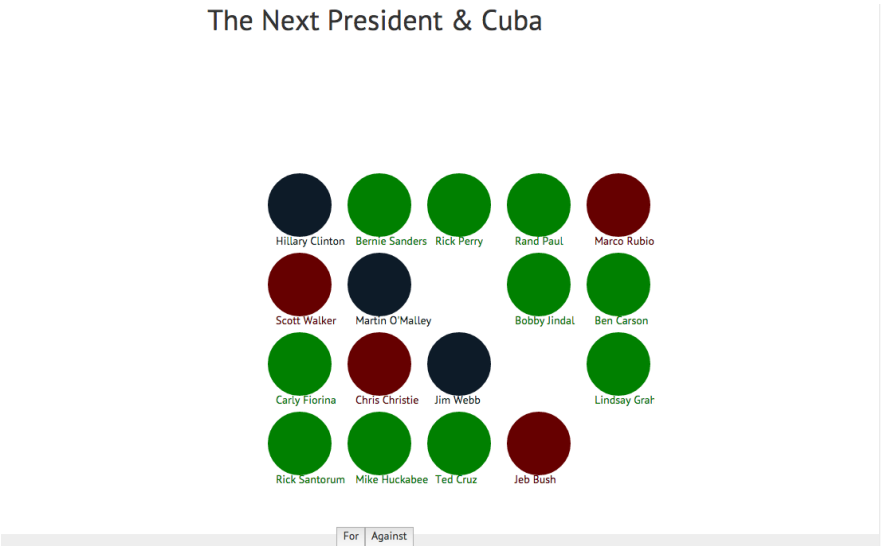
During our milestone one meeting, we discussed with our TF about turning the visualization of presidential candidates into a Parallel Coordinates visual. While technically interesting and visually complex, we feel that this is interface is too advanced for our expected users and their use cases. It is neither the most intuitive solution nor the most promising of insights given that we know apriori that party affiliation has by far the strongest predictive ability given how most candidates agree very closely with party

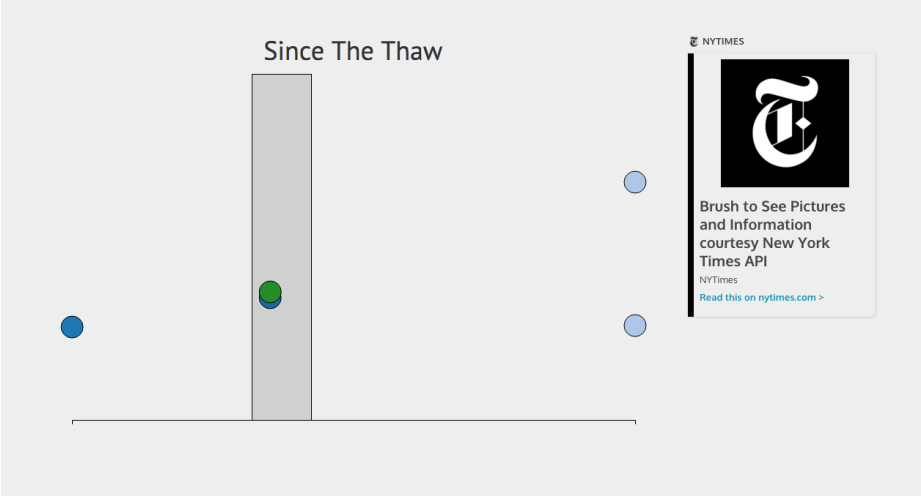
line. Of all the potential presidential candidates, only one (Rand Paul) goes against his party's official opinion (He supports it).

Directly below are those same visualizations albeit in a more intermediate state.

Note: The Congress visualization was dropped because after extensive research made it clear that there was not enough variance within the parties to be able to show meaningful data.



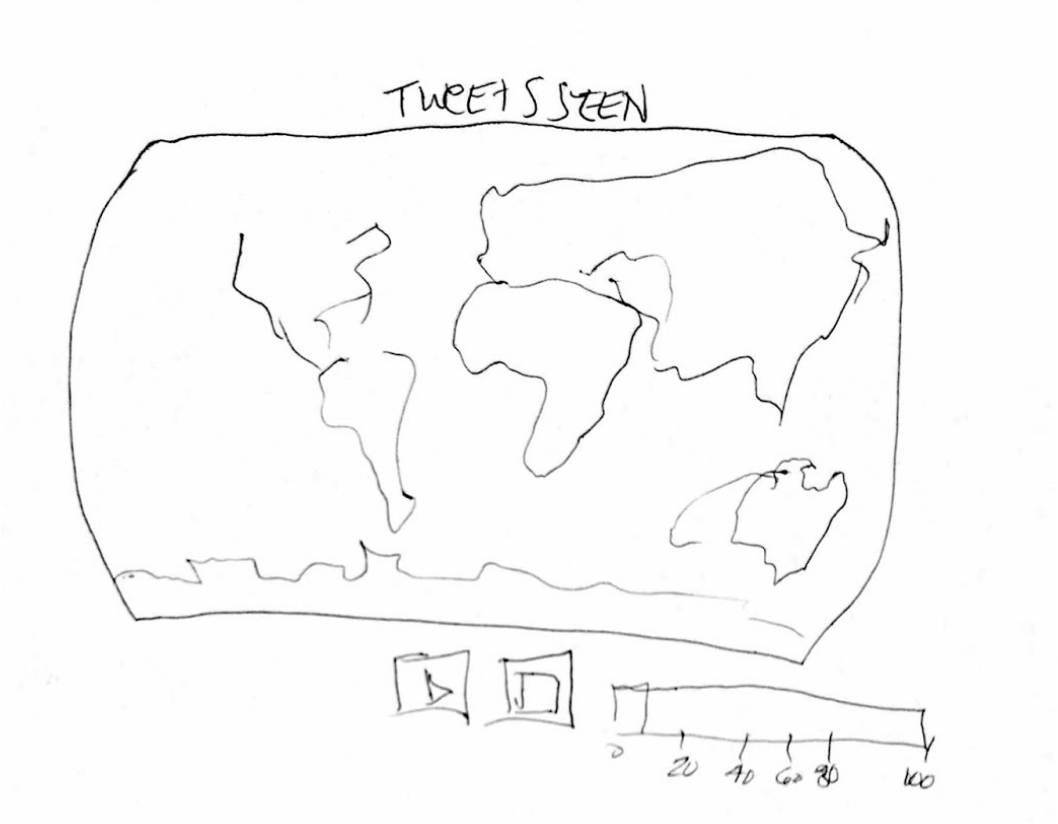




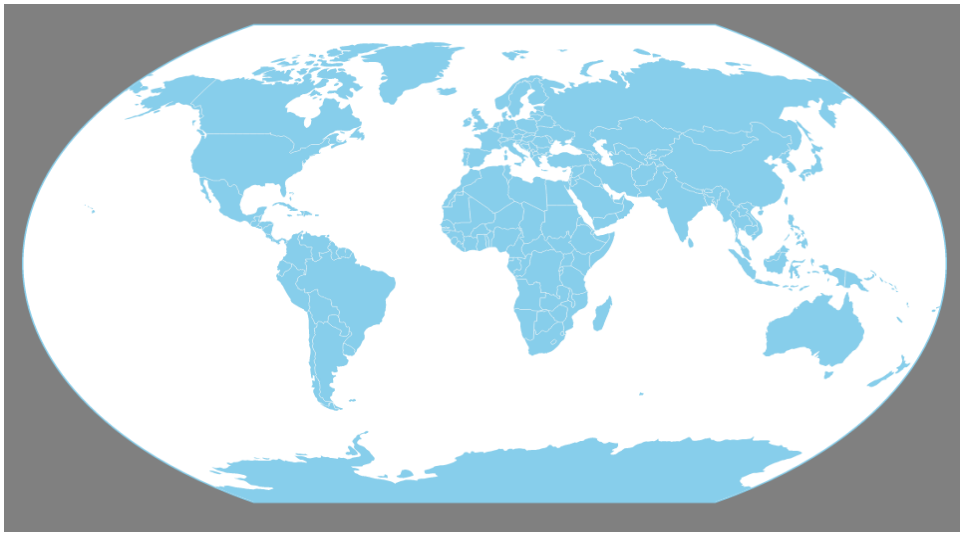
This brush queries the NYTimes for articles related to Cuba between the brush's extent.

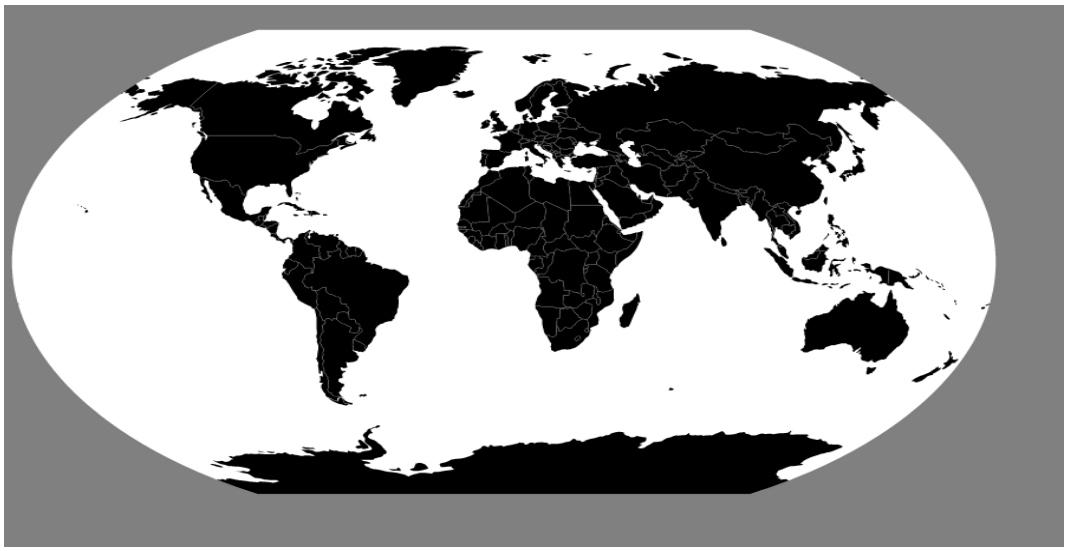
The Project Pivot.

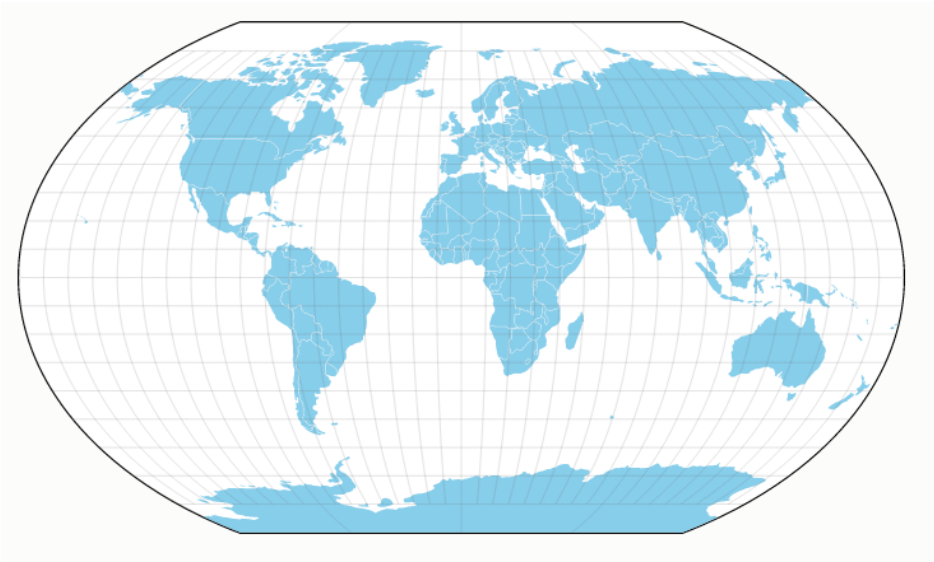
Below this line are images we produced after the project changed course from the initial propoal.



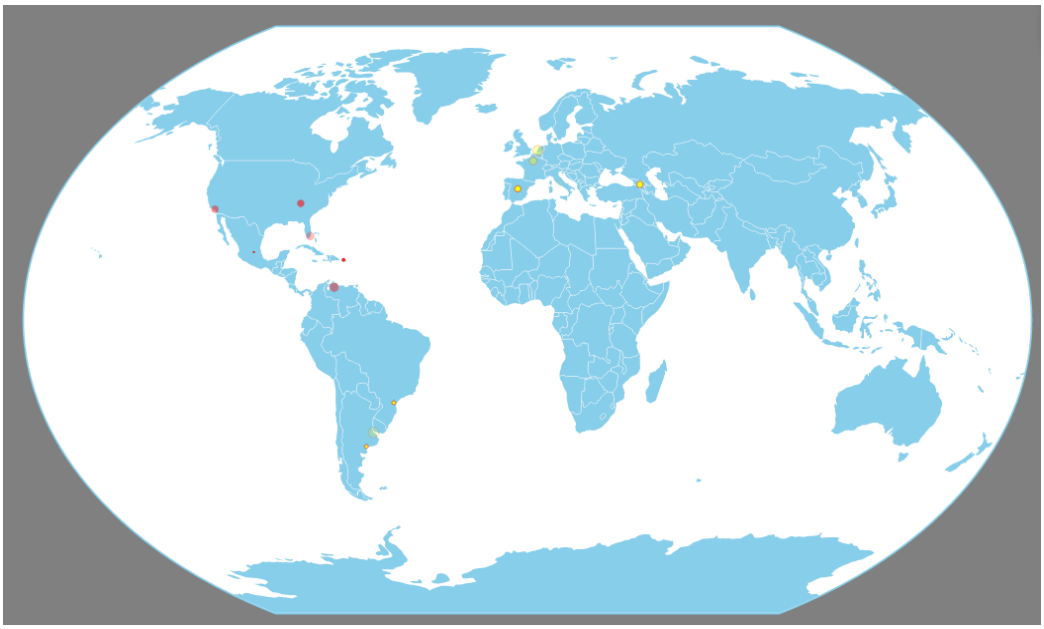
An initial, basic sketch of our map visualization.



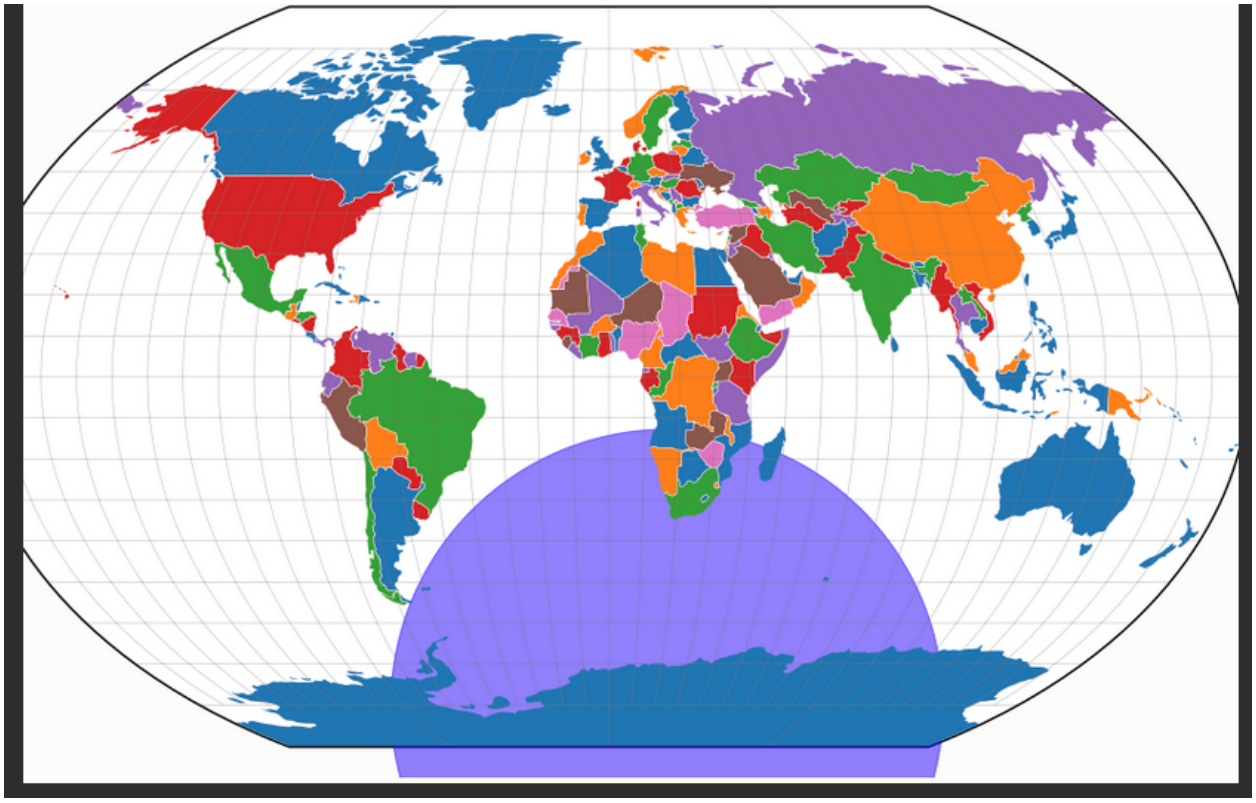




Choosing what kind of projection and what colors to use was a non-trivial decision, but ultimately we based it on aesthetics.



Our first test visualizing a couple points onto a global projection.



Sometimes things happen in testing that you can't explain….

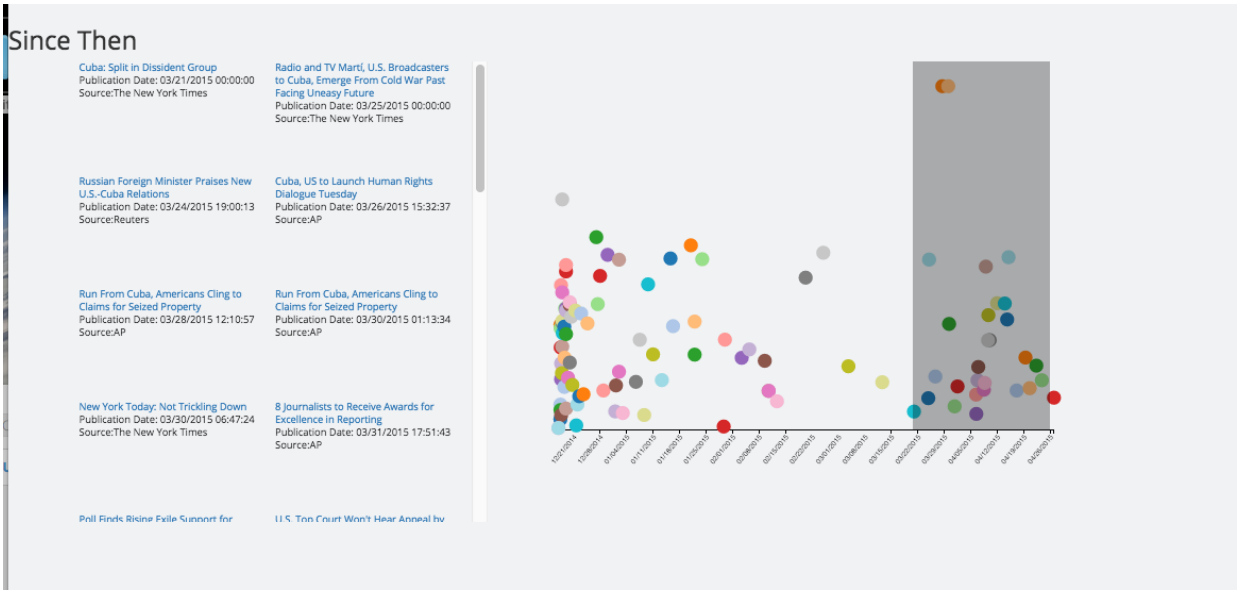
A near final representation.

Implementation

Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced image

This intro video gives all users some historical context while creating excitement for the visualization. Links are provided for those users that know exactly what they want to see.

This is the tweet map of our dataset playing at 20x faster than real time. Users may select whether they want to see a world or a US-only map.



Brush against this graph of NY Times API results based on a query to find all articles related to the Thaw. Articles appear on the left with links and source data.

Hover over a Presidential candidate to see whether or not she approves of the new Cuba Policy.

Evaluation

The Thaw crossed into the mainstream conversation with many ordinary people commenting on it and the historical significance of the US embargo of Cuba was a very big deal to people in Latin America and Europe. We also noticed that even though Twitter skews young, Florida still erupted with activity indicating a young Cuban population is likely to be politically engaged which could prove dangerous to some politicians.

We had no idea that Venezuela's Twitter had such an extensive presence until we finally saw it in the visualization. Venezuela has been Cuba's closest ally over last 10 years. That relationship ended first with death of Hugo Chavez and now the collapse of oil which Venezuela's economy depends on as it is one of the largest exporters of petroleum in the world. With the Cuban Thaw occurring on top of the economic havoc and political turmoil in that country, it is obvious in hindsight that the people of Venezuela would be especially vocal on Twitter, but of course it took the visualization for us to see this.

Ultimately, we chose this dataset because we felt strongly that it is truly a significant piece of history. Not only does it capture a historical detente in Americas, but it does so using the natural expressiveness and stream of consciousness style that Twitter is famous for. This dataset will one day allow future historians to read the mind of humanity which we think is the future of historical record keeping.

Improvements

- Take 2-3 months and dive extremely deeply into semantic analysis (not recommended, there are better, less technical ways albeit with less originality.
- Given that tweets can be either retweets to your followers or just an original comment or thought, we can build a graph of tweets to show the flow of information. Flow of tweets between cities. If Miami's tweets are indeed against the new policy, than interesting links would emerge between Miami and more rural anti-Obama areas.
- We could have tweets of a certain hashtag increase the proportion of that color in a countries color, thus a country takes on the color of it's most popular hashtag.
- The slider can be upgraded to be made much more responsive but there is a limit to what we can do because with the full data playing, we are certainly approaching the outer edges of animation in the browser.

.