

Jerry Castro & Ivan Lima

The Cuban Thaw

Table of Contents

Overview and Motivation

Related Work

Questions

Data

Exploratory

Design Evolution

Implementation

Evaluation

Overview and Motivation

Provide an overview of the project goals and the motivation for it.

Our projects primary goal is to visualize the Cuban Thaw through a massive dataset of tweets purchased through DiscoverText.com using a query on Twitter’s “firehose” data.

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.

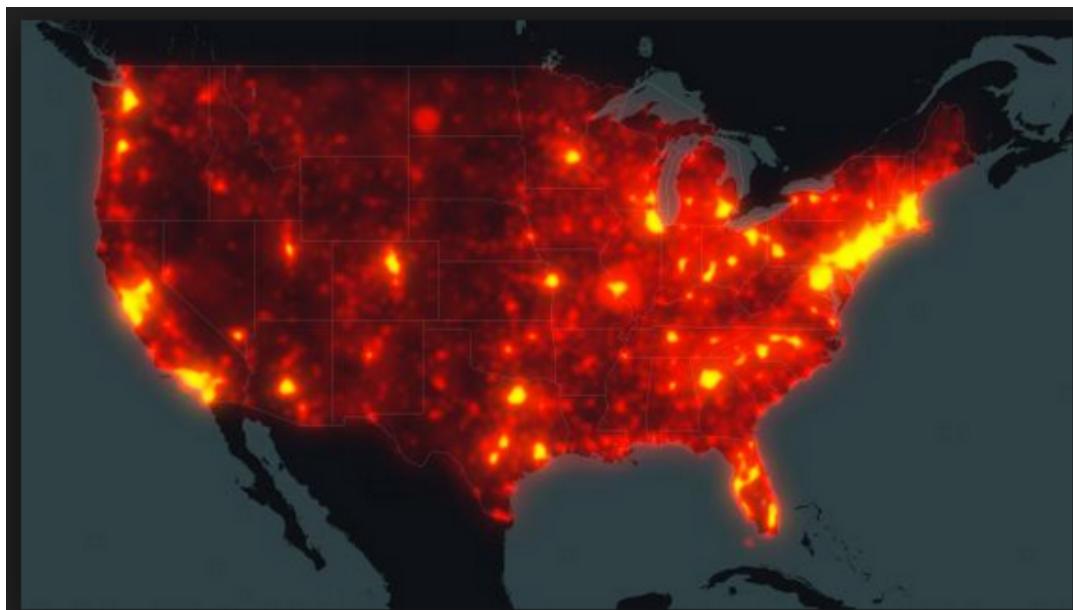
But it has been nearly 25 years since the end of the Cold War. Most American young adults today only know about Fidel Castro and the Soviet Union from history books, that is, if they know about them at all. Political scientists agree that they really aren't sure how this group of millions, the so called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

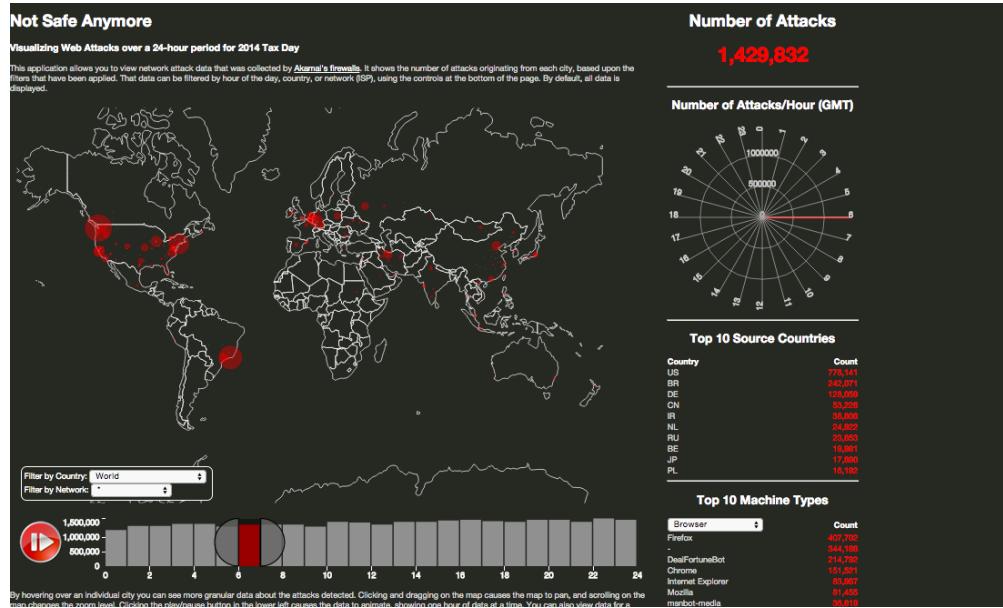
Related Work

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. The images below also served as inspiration.





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?

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?
- What countries didn't appear to find it interesting 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 in real time while specifying if the tweet is original, a retweet, or a favorited tweet. This is a doable but very ambitious goal on both a technical and visual level and thus meets our target for an "awesome" project.

Data

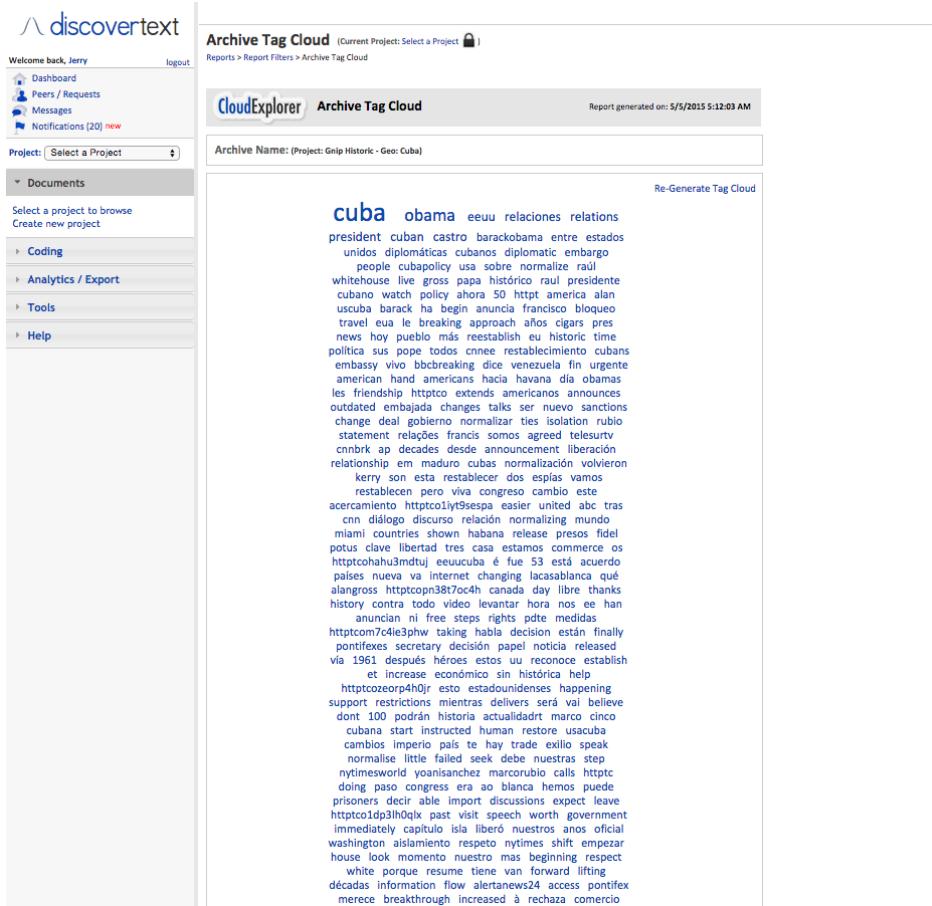
Source, scraping method, cleanup, etc.

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.

The screenshot shows the DiscoverText web application interface. The left sidebar contains a navigation menu with options like 'Dashboard', 'Peers / Requests', 'Messages', 'Notifications (20 new)', 'Documents', 'Data Archives', 'Coding', 'Analytics / Export', 'Tools', and 'Help'. The main area is titled 'Search and Browse Archive' and shows a list of tweets. The search term is 'Has_Geo:Cuba-9' and the results are 'Showing 99,978 to 100,000 of 100,000 total'. The interface includes a toolbar with filters, sorting, and export options. The tweets listed are from various users (@arze, @eliaspin, @NicolasMaduro, @marcorubio, @joshrogin, @RT, @BarackObama, @ActualidadRT, @ltvnews, @MauroMura11, @RadioColombia, @WhiteHouse, @marclamonthill, @ultimosegundo, @sinderbrand) and discuss topics such as US-Cuba relations, Castro, Obama's policies, and the blockade.

User	Tweet Content
@arze	If two Rolex can live happily next to each other, why can't the US & Cuba??
@eliaspin	Rubio Warns Congress Will Resist Cuba Policy
@NicolasMaduro	destaca "valentía" de Obama para restablecer relaciones con Cuba
@marcorubio	you speak very well sir, I disagree with your opinion on #CubaPolicy but I respect it.
@joshrogin	@akettappere the inclusion of Cuba on that list has long been debatable.
@RT	@Palestina: Terminará embargo después de medio siglo de bloqueo yanqui contra Cuba? Ojalá! Y el b...
@BarackObama	TAKING OFF MY HAT FOR OPENING RELATIONS WITH CUBA & @Pontifex GOD BLESS YOU ! ips.f...
@ActualidadRT	#Obama anunció que pedirá al Congreso estadounidense levantar el bloqueo de Cuba h...
@ltvnews	Pope Francis congratulates US and Cuba on agreement
@MauroMura11	La UDI cuando supo lo de EEUU y Cuba.
@RadioColombia	#AlAire análisis con Claudia Palacios @claudiapcmn, noticia del dia: se acercan ...
@WhiteHouse	President Obama speaks with President Raúl Castro of Cuba before announcing his #Cub...
@marclamonthill	The Left must demand the continued protection of our political prisoners by the ...
@ultimosegundo	Espíñola libertado informações cruciais para processos contra cubanos
@sinderbrand	WashPost Cuba heds before and after, 53 years apart.



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. (<http://upload.wikimedia.org/wikipedia/commons/thumb/5/5e/Triangle.Centroid.svg/182px-Triangle.Centroid.svg.png>) 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 important 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 (<https://github.com/mbostock/d3/wiki/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).

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 projects, it really is the quality, quantity, relevance, and impact of 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-locates 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 --attribution). 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.

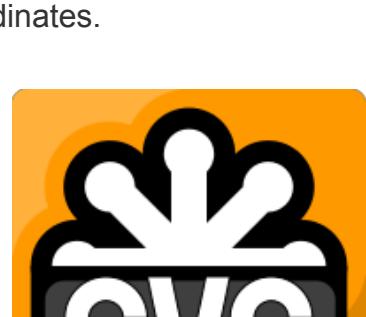
Infrastructure

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.

		Has_Geo_Cuba_9-export-20150412-113248.csv																										
		Home	Layout	Tables	Charts	SmartArt	Formulas	Data	Review	Font	Calibri (Body)	12	A A	abc	Wrap Text	Format	Number	General	Conditional Formatting	Normal	Bad	Cells	Themes	Insert	Delete	Format	Themes	
1	[M] location_c [M] location	[M] location	[M] location	[M] media_d [M] media_t [M] media_u [M] posted_1 [M] real_nan [M] source:	##### #ColombiaDx Twitter for A	16131	http://www. Dedicated Colombia	NTN24	NTN24	http://www. chefAronna																		
2	r.com/chefAronna/statuses/5450696994289152				##### Rene Beaul Twitter Web	2869	http://www. Politically ob WI	NowWithale NowWithale	NowWithale	http://www. Bluegrl_3																		
3	r.com/Bulegir_3/statuses/545406967003762688				##### Daniel Berg TwitterCaster	1255	http://blogs. Commentate Nebraska	Wall Street J WSJ	Wall Street J WSJ	http://www. dberman39																		
4	r.com/dberman39/statuses/545406967217684480				##### Lo'Ac Gauthi Twitter for A	59	J'ai une devi La Ba	Martin Pett Lemicodefi	Martin Pett	http://www. famousred																		
5	r.com/famousred/statuses/545406968069115904				##### Gary J. Esnei Twitter Web	482	Nashua, NH	Top Conservi TeaPartyC	Top Conservi	http://www. GESeault																		
6	r.com/GEsneault/statuses/5454069688069164				##### Giselle Mori TweetDeck	3353	http://www. Periodista de Cuba	Escambray	F Escambray	http://www. gissellemr																		
7	r.com/gissellemr/statuses/5454069698829056				##### Dani Twitter for A	1925	http://www. Vajar y ser Lima - PerÃ	Noticias SIN SIN24horas	Vajar y ser Lima - PerÃ	http://www. diancabrera21																		
8	r.com/diancabrera21/statuses/54540696976821312				##### AmericaNeer Twitter Web	1177	http://twitte Serial entrep North Caroli	The Daily Be	the daily beans	http://www. angel_cortez78																		
9	r.com/angel_cortez78/statuses/545406969017024512	pic.twitter.cc photo	http://pbs.tv #####		##### Camilo Fdo V Twitter for A	3275	Colombiano, Melbourne - Leonardo Pa Leonardo_Pc	Leonardo Pa Leonardo_Pc	Colombiano	http://www. cam_vallejo																		
10	r.com/cam_vallejo/statuses/545406968899174401	pic.twitter.cc photo	http://pbs.tv #####		##### mast Twitter Web	37252	http://twitte colorado girl colorado girl	Kenneth Isa uselephants	colorado girl	http://www. 67purple																		
11	r.com/67purple/statuses/5454069688995458	pic.twitter.cc photo	http://pbs.tv #####		##### Luis Moreira Plume&forÃ	26234	The views an Trujillo Alto, AndrÃo Co andrescolon	Jullopi	Jullopi	http://www. lullopi																		
12	r.com/lullopi/statuses/54540696962303360	pic.twitter.cc photo	http://pbs.tv #####		##### Miriam R. Twitter for A	16759	Educadora A Caracas	ManoloReve ManoloReve	Educadora A Caracas	http://www. annirauli																		
13	r.com/annirauli/statuses/545406971965628416	pic.twitter.cc photo	http://pbs.tv #####		##### Tom Scanlon Twitter for V	10508	http://media Re-elect Mill Hyattsville, Maryland, USA	Parcial2	Re-elect Mill Hyattsville, Maryland, USA	http://www. Parcial2																		
14	r.com/Parcial2/statuses/545406973445812225	pic.twitter.cc photo	http://pbs.tv #####		##### Tom McGevit tmgev Twitter Web	15232	Co-founder N York	Tom McGevit tmgev	Co-founder N York	http://www. tmgev																		
15	r.com/tmgev/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Conservative Twitter for iF	1695	http://twitte Proudly follo USA	Ban Collectiv meology	Proudly follo USA	http://www. _People																		
16	r.com/_People/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Alisa Anna M Twitter for A	2448	http://news Religious,53, Norway	Islam in Norway	Islam in Norway	http://www. Islam_in_Norway																		
17	r.com/Islam_in_Norway/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Aslana S. Twitter for A	21189	Aspiring sup Anchorage, AK	Top Conservi TeaPartyCat	Aspiring sup Anchorage, AK	http://www. AthiesBaven																		
18	r.com/AthiesBaven/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Analog_Zero Twitter Web	173747	Colombiano, Melbourne - Leonardo Pa Leonardo_Pc	Analog_Zero	Analog_Zero	http://www. Analog_Zero																		
19	r.com/Analog_Zero/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### CÃ©sar Barr Twitterfeed	9547	http://www. Comienza tu DF	Barraguilla	Hassan Nassar HassNassar	http://www. cbarrons																		
20	r.com/cbarrons/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Franklin Twitter for A	1654	http://actual Chavita, Res Venezuela	Rt en EspÃ ActualidadadR1	Retired small Montana	http://www. frankalbarra																		
21	r.com/frankalbarra/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Marilyn Oliv Twitter for V	144896	http://www. Dios, Esposa, MÃxico	Lynblonde Lynblonde	Retired small Montana	http://www. MTmarlyn2																		
22	r.com/MTmarlyn2/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Sonia Presa Twitterfeed	15045	http://www. Los sueÃos MÃxico	Los sueÃos MÃxico	Assalamu korea	http://www. sospr																		
23	r.com/sospr_/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Ngee Twitter for A	1165	Docente de I Varcup Veni Mario Silva C LaJilljaenT	Retired small Montana	Assalamu korea	http://www. sospr																		
24	r.com/sospr/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Ramon Andri Twitter for A	1385	Docente de I Varcup Veni Mario Silva C LaJilljaenT	Delgado, 106	Docente de I Varcup Veni Mario Silva C LaJilljaenT	http://www. Delgado, 106																		
25	r.com/Delgado_106/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Jorge Moren Twitter Web	4364	Proyecto inn Maracay	MABEL BORK 2405mabel	Proyecto inn Maracay	http://www. proyecto_eps_b1																		
26	r.com/proyecto_eps_b1/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Danay Suarez Twitter for IF	1825	Managemen La Habana, Cuba	DanaySuarez	Managemen La Habana, Cuba	http://www. DanaySuarez																		
27	r.com/DanaySuarez/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Beto Moren twitterfeed	10646	http://www. Mi novia y Yo DF	betomoren	betomoren	http://www. betomoren																		
28	r.com/betomoren/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### DamÃ¡n Igles Twitterfeed	10644	http://www. DESCUBRIMI Ecatepec, MÃxico	DESCUBRIMI Ecatepec, MÃxico	DESCUBRIMI Ecatepec, MÃxico	http://www. ddiless																		
29	r.com/ddiless/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Sonia Presa tweeterfeed	10644	http://www. Dio, Esposa, MÃxico	Dio, Esposa, MÃxico	Dio, Esposa, MÃxico	http://www. sospr_																		
30	r.com/sospr_/_statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Elis Clement Twitter Web	15241	http://www. Socialista, re Venezuela	Fidel Castro	fidelcastro	http://www. Riverocontreras																		
31	r.com/Riverocontreras/statuses/54540697174724608	pic.twitter.cc photo	http://pbs.tv #####		##### Catherine C. dritv.it	1455	http://none I'm a reporte Rocky Mount CNNMoney	CNNMoney	CNNMoney	http://www. BHandgraaf_RMT																		
32	r.com/BHandgraaf_RMT/statuses/5454069801516928	pic.twitter.cc photo	http://pbs.tv #####		##### Finita Twitter for A	18057	http://twitte Venezuela.	El monitor 1EElmonitor1B	El monitor 1EElmonitor1B	http://www. Xjesman30																		
33	r.com/Xjesman30/statuses/5454069802327552	pic.twitter.cc photo	http://pbs.tv #####		##### Passionate Twitter Web	14203	http://www. Passionate Texas USA	Katherine Mi katherinemil	Katherine Mi katherinemil	http://www. allovestexas																		
34	r.com/allovetexas/statuses/545406981121769472	pic.twitter.cc photo	http://pbs.tv #####		##### Conservative Twitter Web	1064	http://www. Los sueÃos MÃxico	Los sueÃos MÃxico	Los sueÃos MÃxico	http://www. mn90marcos																		
35	r.com/mn90marcos/statuses/545406981131435392	pic.twitter.cc photo	http://pbs.tv #####		##### Marcos Mez twitterfeed	1064	http://www. Los sueÃos MÃxico	Los sueÃos MÃxico	Los sueÃos MÃxico	http://www. mn90marcos																		
36	r.com/1776BettyRoss/statuses/545406980433936384	pic.twitter.cc photo	http://pbs.tv #####		##### Proudamerik Twitter Web	13535	Independent USA	Bill Damato, Bill111; me	Independent USA	http://www. 1776BettyRoss																		
37	r.com/jalynhenton/statuses/545406980681375744	pic.twitter.cc photo	http://pbs.tv #####		##### Jalyn Henton Twitter for A	11293	Doing my be Del Ray, Alexandria, VA	Yoani SÃ¡inz yoanisanche	Yoani SÃ¡inz yoanisanche	http://www. jalynhenton																		
38	r.com/Longueira_2018/statuses/545406980670527566	pic.twitter.cc photo	http://pbs.tv #####		##### Pablo Longui Twitter for IF	1745	http://www. TWITTER de Chile	Yoani SÃ¡inz yoanisanche	Yoani SÃ¡inz yoanisanche	http://www. Longueira_2018																		
39	r.com/mfarooqqaifal7/statuses/54540698458452726528	pic.twitter.cc photo	http://pbs.tv #####		##### M. Farooq Al Twitter for B	1072	http://www. Businessman Pakis Reuters	Reuters	Reuters	http://www. mfarooqqaifal7																		
40	r.com/catherine_cecil/statuses/5454069862677911361	pic.twitter.cc photo	http://pbs.tv #####		##### Catherine C. dritv.it	27628	Morelia, MichoacÃn	Morelia, MichoacÃn	Morelia, MichoacÃn	http://www. catherine_cel																		
41	r.com/PamManggdrivit/statuses/545406986303527552	pic.twitter.cc photo	http://pbs.tv #####		##### Pam Manggdrivit	37082	Amo las flores Morelia, MichoacÃn	Amo las flores Morelia, MichoacÃn	Amo las flores Morelia, MichoacÃn	http://www. PamManggdrivit																		
42	r.com/digitalcamroni/statuses/545406986303527552	pic.twitter.cc photo	http://pbs.tv #####		##### Jose Luis Car Twitter for A	2920	calimetrico Granada, Sp! Teacher mai_magia	calimetrico Granada, Sp! Teacher mai_magia	calimetrico Granada, Sp! Teacher mai_magia	http://www. digitalcamroni																		
43	r.com/INGJOSEMANUEL/statuses/545406986303527552	pic.twitter.cc photo	http://pbs.tv #####		##### JOSE RODRIC Twitter for B	17833	Ingeniero Int Pakis Neutra	Neutra	Ingeniero Int Pakis Neutra	http://www. INGJOSEMANUEL																		
44	r.com/ftrelt_noma/statuses/545406986303527552	pic.twitter.cc photo	http://pbs.tv #####		##### Noma Drielt nrlv.it	48656	Electrmrl MÃxico, DF	Electrmrl MÃxico, DF	Electrmrl MÃxico, DF	http://www. ftrelt_noma																		



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 twitter data is actually very messy. We spent equal or more time in data wrangling / analysis than in visualization.

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

While Semantria is the industry leader, but dataset proved too unruly to handle in a limited amount of time. But great for exploration/derivative semantic measures. Had to get a virtual windows 8. Datafiles so big, excel would routinely crash. Crazy amount of cleaning had to be done just to do the semantic analysis. Limited api calls means, have to put serious thought into it.

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 during exploration because data set was orders of magnitude larger than anything we were prepared for, because most lacked reliable/easily usable geoData/.
- ❖ Main visualization for most of project was a slice of tweets mapped against US shaded to Obama Romney election results.
- ❖ 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 real time.

TopMeta Discovery

x

Top values for: country_code:

◀ ▶ 1 of 6 ↗ ↘

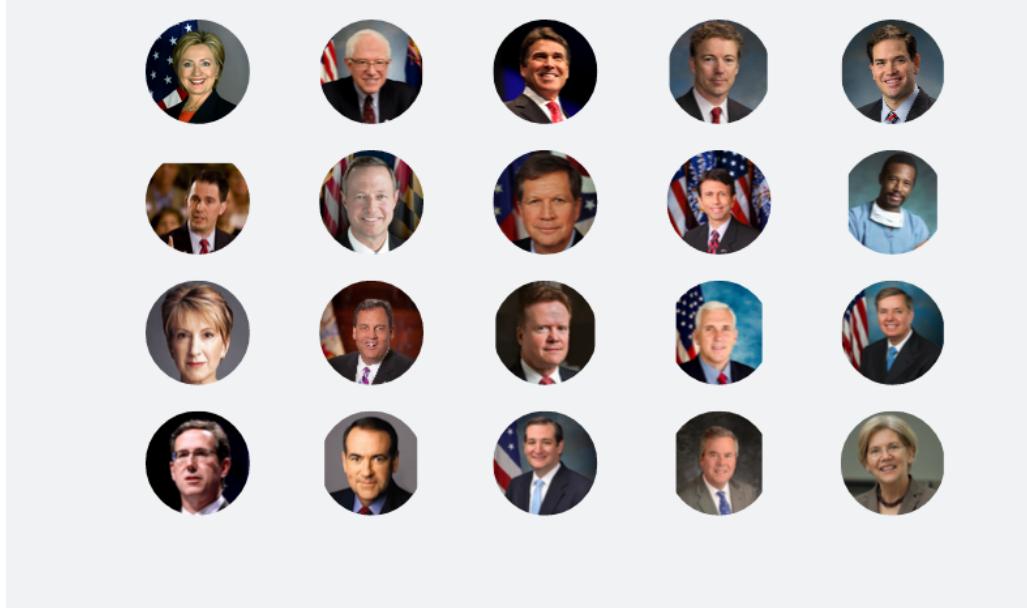
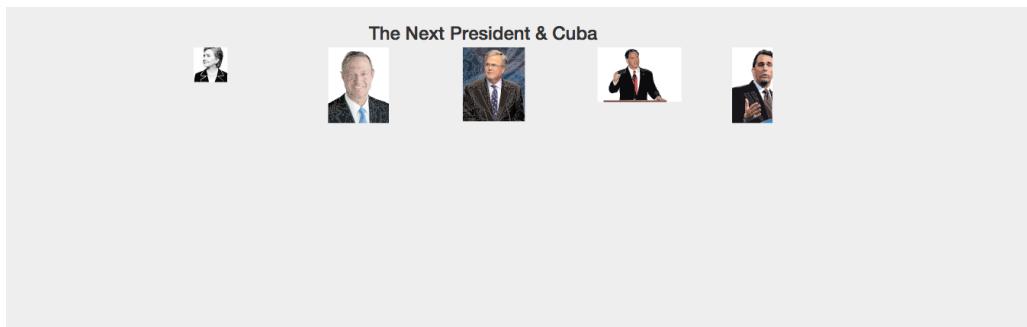
Meta Value	Total ▾	Filter
United States	746	✖️ +
Venezuela	199	✖️ +
Brasil	195	✖️ +
España	128	✖️ +
Chile	114	✖️ +
México	93	✖️ +
Colombia	91	✖️ +
Argentina	85	✖️ +
United Kingdom	51	✖️ +
Canada	45	✖️ +

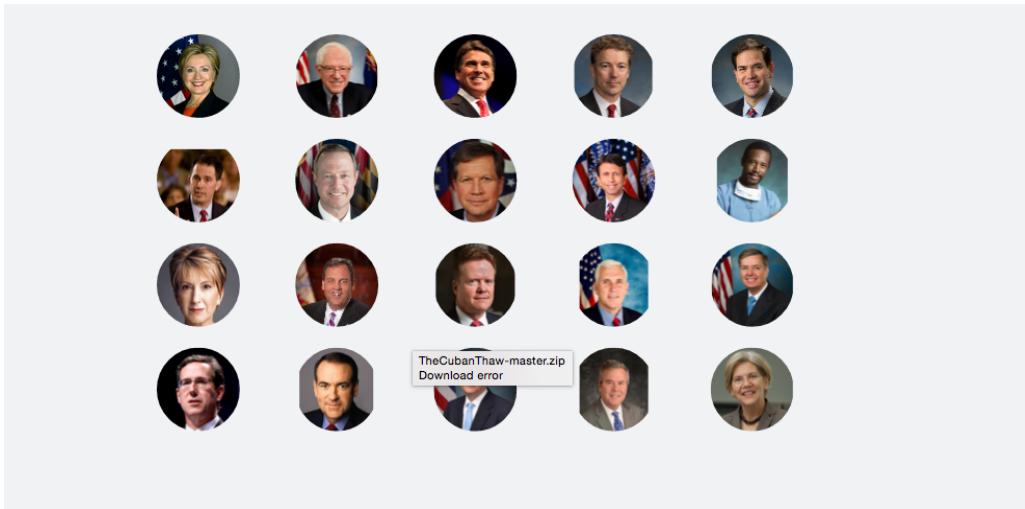
Showing 1 to 10 of 57 total

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.

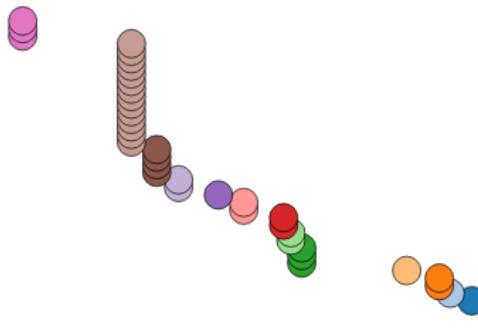
- ❖ We struggled choosing colors for tweets and maps. Ultimately we decided to keep the maps a single color and the tweets could be one of 3 color combinations. This decision was based on the fact that with tens of thousands of tweets populating the screen, multiple colors on multiple backgrounds would 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.





TheCubanThaw-master.zip
Download error

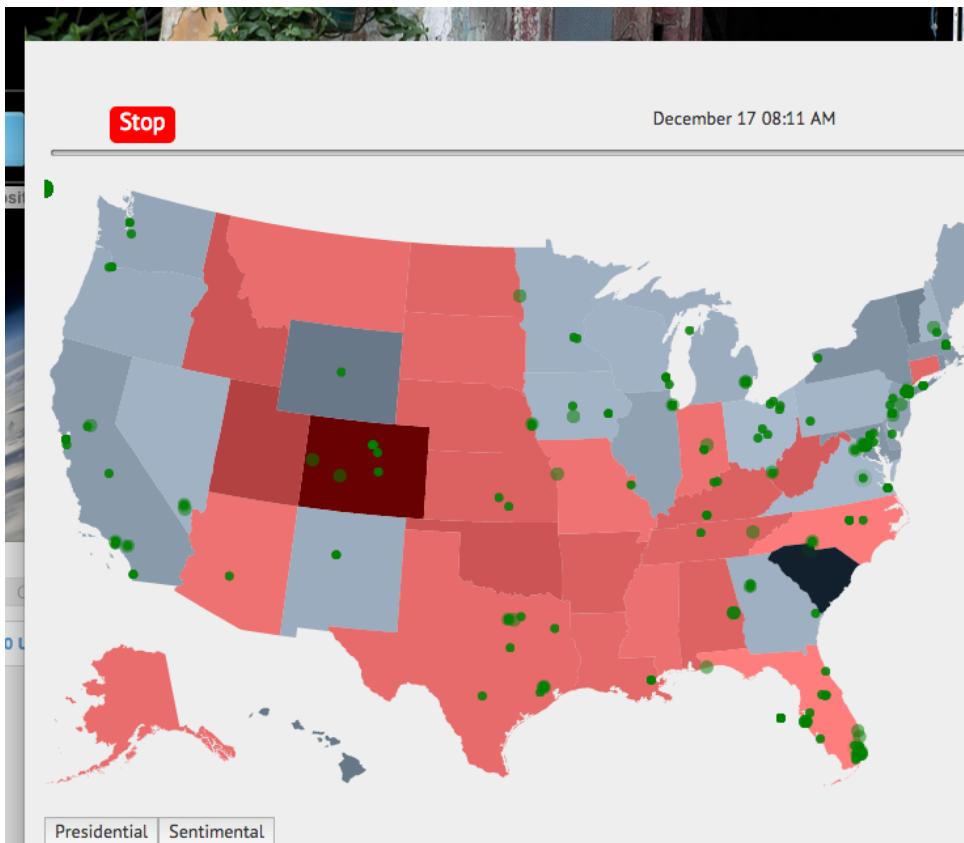
Thaw Events



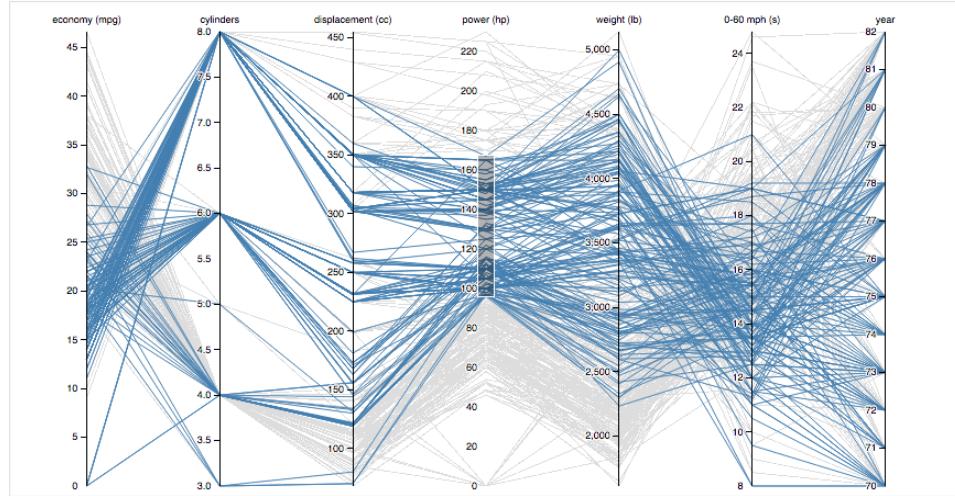
Dec Dec 28 Jan 38 Feb 29 Mar 1 Mar 15 Mar 29 Apr 29 Apr 19

	ID	Source Text	Summary	Detected Language	Detected Language Strength	Document Identifier	Phrase Identifier	Phrase Neglect	Phrase	Phrase Sentiment	Phrase Score	Entity	Entity Type	Entity Identifier	Entity Neglect	Entity Score	Entity Sentiment
1	475-1795	lunched enough! Let's move on!	Cuba, The cubans have	English	0.100000				End/Head/Emph	positive	0.29	Cuba	Place	Place	positive	0.29	positive
1137	4337-024n	Cuba and hello express some disorders!!	Obama just fly	English	-0.2302				express	0.39	n't Obama	Place	Place	neutral	0.39	neutral	
1138	4848-02fc	Love this move by Time for Cuba to make progress.	for Cuba to make	English	0.165931				progress	0.165931	neutral	Time	Person	0.0417932	neutral	7	
1139	4289-0489	GDP on an uncomfortable position. It'll be interesting to	normalization with Cuba	English	-0.036427				interesting	0.49	positive	Time	Person	-0.036427	neutral	7	
1140	4289-0490	Stay on and Cuba brothers. How pathetic!	day the normalization	English	-0.000598				normalization	0.5	positive	Time	Person	-0.000598	neutral	7	
1141	4624-011d	Normalizing the relationship has been	the people of Cuba	English	0.69	positive			best	0.49	positive	Normalization	Organization	0.69	neutral	7	
1142	4551-0bae	the people of Cuba will have better lives because of this.	Obama talk about Cuba	English	0.18	0.0732			Hopefully	0.28196	positive	Cuba	Place	0.50245	positive	7	
1143	4842-0c4e	Happy Hammill to Alan and Judy Gross!! Cuba	Judy Gross!! Cuba	English	0.2	neutral			Happy	0.2	neutral	Alan	Person	0.3	neutral	7	
1144	4402-050e	I can't even imagine why we are at Cuba	the normalization	English	-0.110000	still			freedom	-0.829	negative	n't make	Place	-0.380000	neutral	7	
1145	4660-0111	cuba...this is a massive photo op and thumb in the eye	normance, there is no	English	-0.2074	negative	no		mad	-0.323158	negative	remember	Cuba	Place	-0.1148	neutral	7
1146	4700-011b	I can't even imagine why we are at Cuba	debt and credit cards in	English	0.16	0.063354	negative	only	European ones	0.063354	neutral	Cuba	Place	0.1540864	neutral	7	
1147	4512-04a1	Cuba's human rights record?	telling me you're not	English	0.10	0.446032	positive		human rights	0.446032	positive	telling	Cuba	Place	0.11508	positive	7
1148	4793-0151	Many places there including publications	politics on Cuba's org	English	0.12	0.1562	neutral	So	happy	0.7	positive	Many places	Cuba	Place	-0.0361	neutral	7
1149	4210-0271	Alan and Judy Gross!! Cuba	Alan and Judy Gross!!	English	0.15	0.020313	positive		so	0.528000	neutral	Cuba	Place	-0.034232	neutral	7	
1150	4212-0e0f	Hispanic votes, Cubans in Miami have been able to	winnow dressing	English	0.15	-0.5	negative		shots	-0.4	negative	votes	Cuba	Place	-0.5	negative	7
1151	4213-0e0f	Hispanic votes, Cubans in Miami have been able to	winnow dressing	English	0.15	-0.5	negative		shots	-0.4	negative	legally go	Miami	Place	-0.5	negative	7
1152	4281-020c	Ben Rhodes	Ben Rhodes	English	0.11	0.038493	positive		pandering	0.230000	neutral	Cuba	Place	0.230000	neutral	7	
1153	4281-020d	Carroone's grand Crime in 1993, my strongest	Founder of	English	0.11	0.038493	neutral		Crime	-0.246393	negative	Founder	Person	0.0557993	neutral	7	
1154	4281-020e	policy plan. In real life, it was Ben Rhodes	who orchestrated the	English	0.15	0.106	neutral	In	real life	0.212	neutral	Leo	Person	0.106	neutral	7	
1155	4281-020f	policy plan. In real life, it was Ben Rhodes	who orchestrated the	English	0.15	0.106	neutral		Ben Rhodes	0.106	neutral	Person	Person	0.106	neutral	7	
1156	4281-020g	Ben Rhodes	Ben Rhodes	English	0.15	0.106	neutral		Ben Rhodes	0.106	neutral	Person	Person	0.106	neutral	7	
1157	4798-024e	Runs Congress to Embrace Cuba. 3. Shortage solved	of Doctor Shortage via 2.	English	5	1.3	positive		thanked	0.588000	positive	Denator	Person	0.60	neutral	7	
1158	4240-0271	Zimbabwe, and Egypt! Regimes that use torture are bad	isolated as China,	English	18	-0.002313	negative		Embrace	1.1	positive	Cuba	Place	0.11 positive	7		
1159	4240-0272	Zimbabwe, and Egypt! Regimes that use torture are bad	isolated as China,	English	18	-0.002313	negative		brutal	-0.4	negative	as isolated	China	Place	-0.336355	neutral	7
1160	4240-0273	Zimbabwe, and Egypt! Regimes that use torture are bad	isolated as China,	English	18	-0.002313	negative		brutal	-0.4	negative	as isolated	Cuba	Place	-0.336355	neutral	7
1161	4240-0274	Zimbabwe, and Egypt! Regimes that use torture are bad	isolated as China,	English	18	-0.002313	negative		torture	-0.49	negative	Zimbabwe	Place	-0.336355	neutral	7	
1162	4240-0275	Zimbabwe, and Egypt! Regimes that use torture are bad	isolated as China,	English	18	-0.002313	negative		Just look	0.06063	neutral	Egypt	Place	-0.336355	neutral	7	
1163	4287-0490	relationships with Cuba after all the human rights	wanted to normalize	English	19	-0.127000	negative		violations	-0.26	negative	r'n't believe	U.S.	Place	-0.3	neutral	7
1164	4287-0491	relationships with Cuba after all the human rights	wanted to normalize	English	19	-0.127000	negative		violations	-0.26	negative	Cuba	Place	-0.3	neutral	7	
1165	4287-0492	impassioned response to new Cuba policy. Speaking	Rubio now making	English	9	0.049333	neutral		impassioned	0.116	neutral	Sen. Marco Rubio	Person	0.42	neutral	7	
1166	4287-0493	impassioned response to new Cuba policy. Speaking	of codding tyrants	English	5	0.130831	negative		tyrants	-0.4	negative	Obama	Person	-0.220000	neutral	7	
1167	4287-0494	Obama has a long history of codding tyrants	of codding tyrants	English	14	0.130831	negative		tyrants	-0.4	negative	Sen. Marco Rubio	Person	0.42	neutral	7	
1168	4287-0495	Obama has a long history of codding tyrants	of codding tyrants	English	14	0.130831	negative		tyrants	-0.4	negative	President	Person	-0.220000	neutral	7	
1169	4212-0655	No. Korea. All terrorist States. We g	opens relations with	English	6	-0.2734	negative	All	terrorists	-0.7274	negative	Jewish people	Person	-0.221100	neutral	7	
1170	4212-0656	No. Korea. All terrorist States. We g	opens relations with	English	6	-0.2734	negative		terrorists	-0.7274	negative	Cuba	Place	-0.611	negative	7	
1171	4240-0276	of diplomatic ties with US, says differences remain.	Raul Castro welcomes	English	8	0.267549	positive		restoration	0.0402857	positive	Iran	Place	-0.811	negative	7	
1172	4240-0277	of diplomatic ties with US, says differences remain.	Raul Castro welcomes	English	8	0.267549	positive		restoration	0.0402857	positive	Company	Person	-0.811	negative	7	
1173	4240-0278	of diplomatic ties with US, says differences remain.	Raul Castro welcomes	English	8	0.267549	positive		restoration	0.0402857	positive	Cuba	Place	0.2757449	neutral	7	
1174	4240-0279	of diplomatic ties with US, says differences remain.	Raul Castro welcomes	English	8	0.267549	positive		restoration	0.0402857	positive	Raul Castro	Person	0.2757449	neutral	7	
1175	4240-0280	of diplomatic ties with US, says differences remain.	Raul Castro welcomes	English	8	0.267549	positive		restoration	0.0402857	positive	U.S.	Place	0.2757449	neutral	7	
1176	4240-0281	of diplomatic ties with US, says differences remain.	Raul Castro welcomes	English	19	-0.024797	neutral		freedoms	0.272	positive	president	Job Title	0.8191	neutral	7	
1177	4240-0282	negotiator we have ever had." Cuba, his home, did not	the president has to be the	English	19	-0.024797	neutral		worst	-0.196	negative	Cuba	Place	0.0886937	neutral	7	

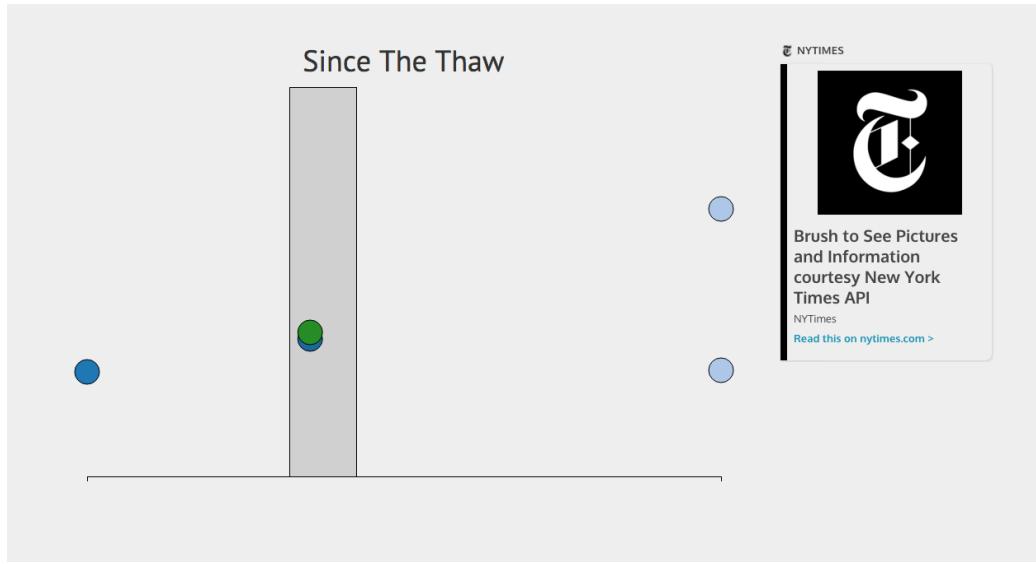
Our main visualization for most of the project, a map of the US with using a small subset of the full dataset (8,000 vs 1,000,000+). Colors were meant to represent either states Obama won or Romney won, but the actual color is just a basic prototype choice.



Parallel Coordinates

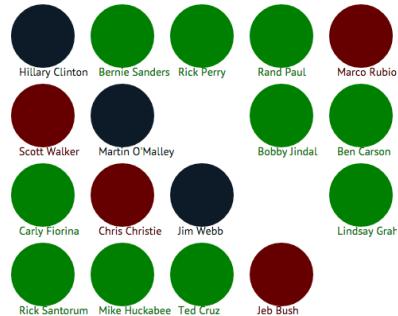


Discussed with TA about turning visualization of presidential candidates into a Parallel Coord vis, but while technically interesting and visually complex, we feel that this is too advanced for our expected users and their use cases. It is neither the most intuitive solution nor the most promising of insight 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



Purpose of this visualization was to communicate to the user the immediate consequences of the Thaw through events happening in the news. This brush queries the NYTimes for articles related to Cuba between the brush's extent.

The Next President & Cuba

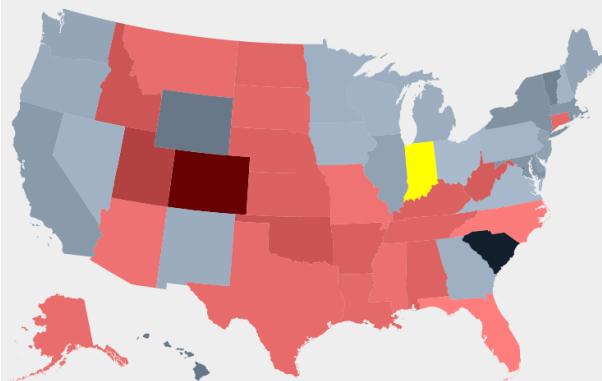


[For | Against]

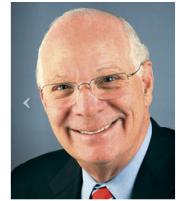
The Cuban Thaw Tweet Map The Next President & Cuba Since The Thaw

Play

December 15 18:11 PM

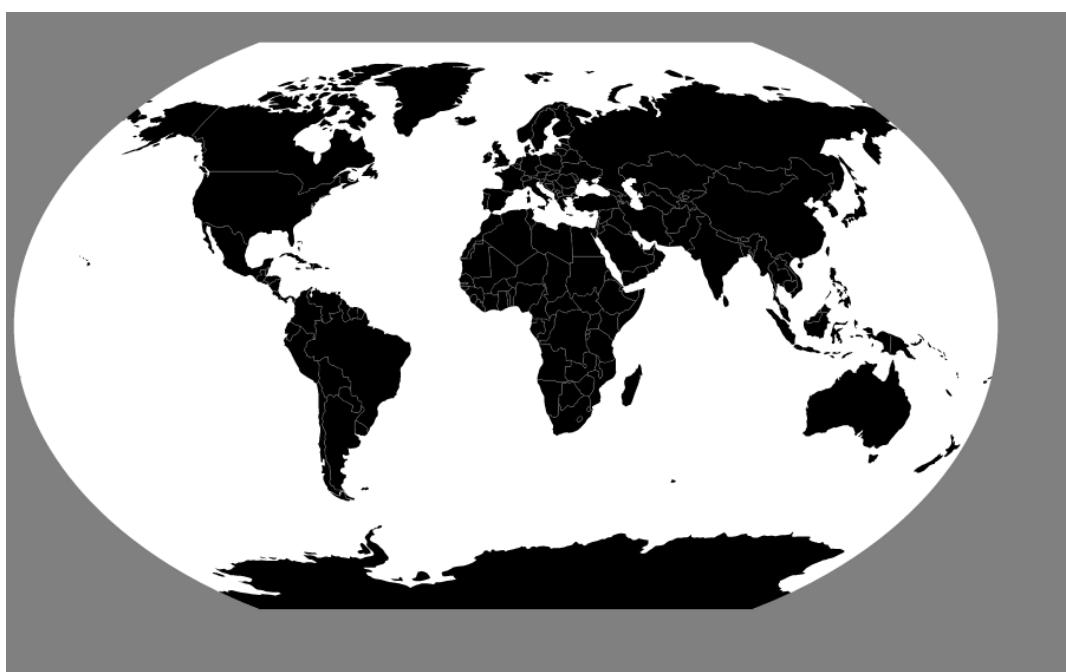
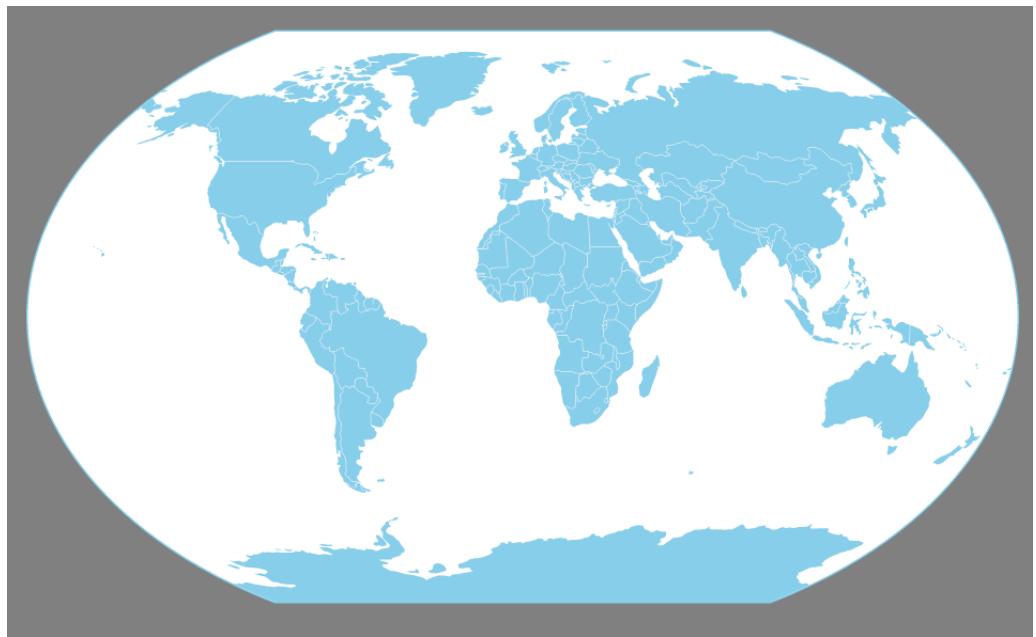


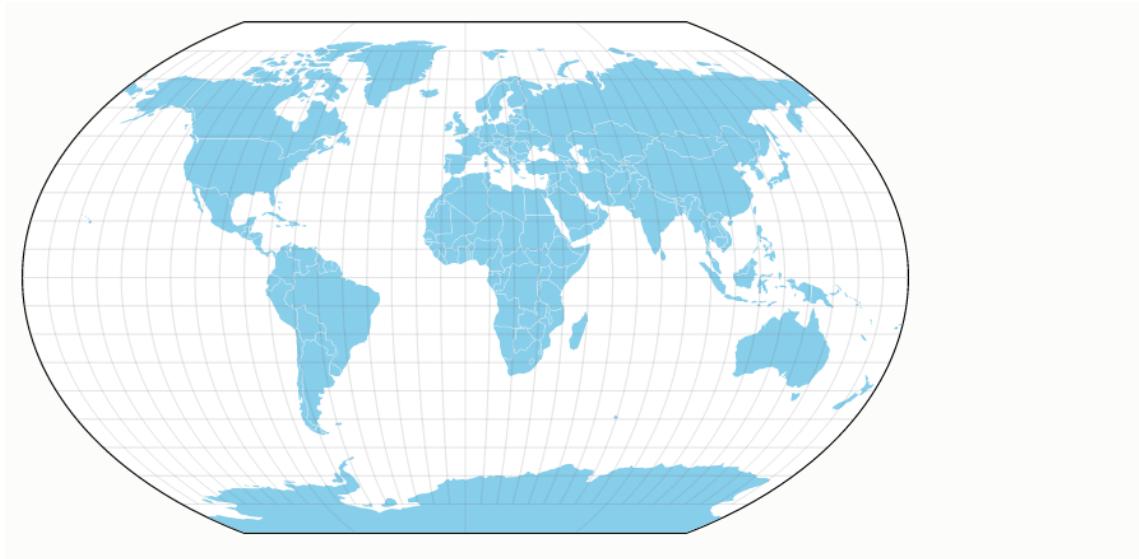
Senators

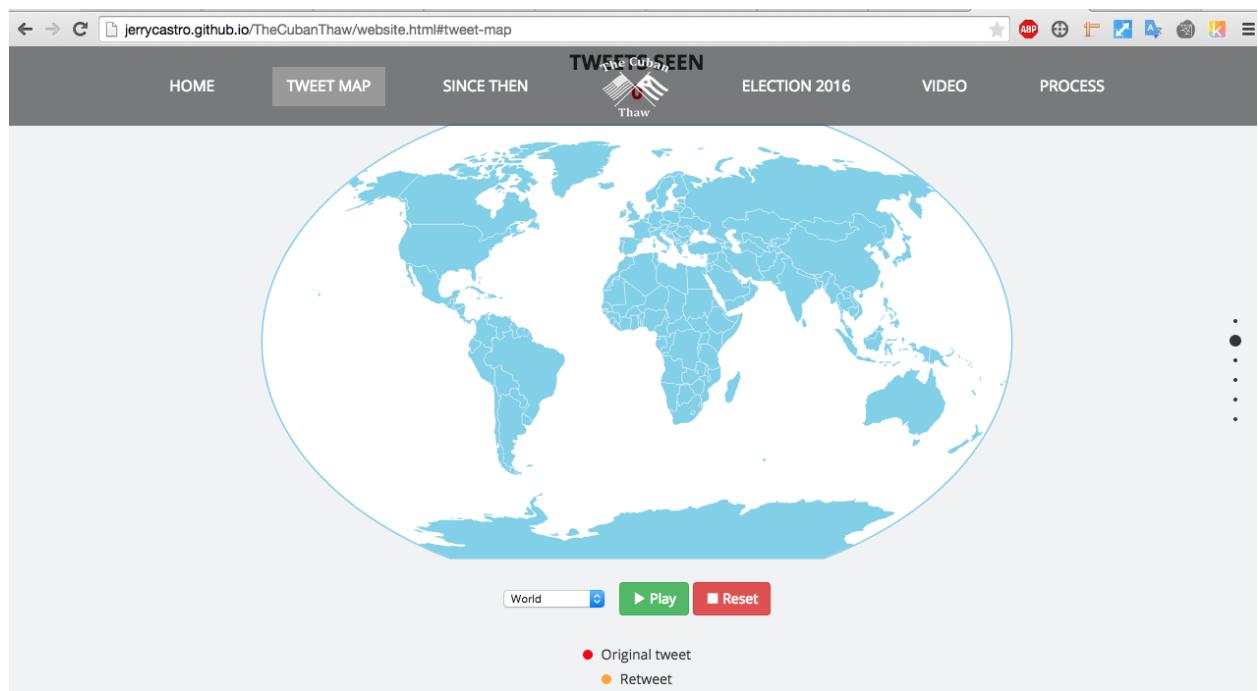
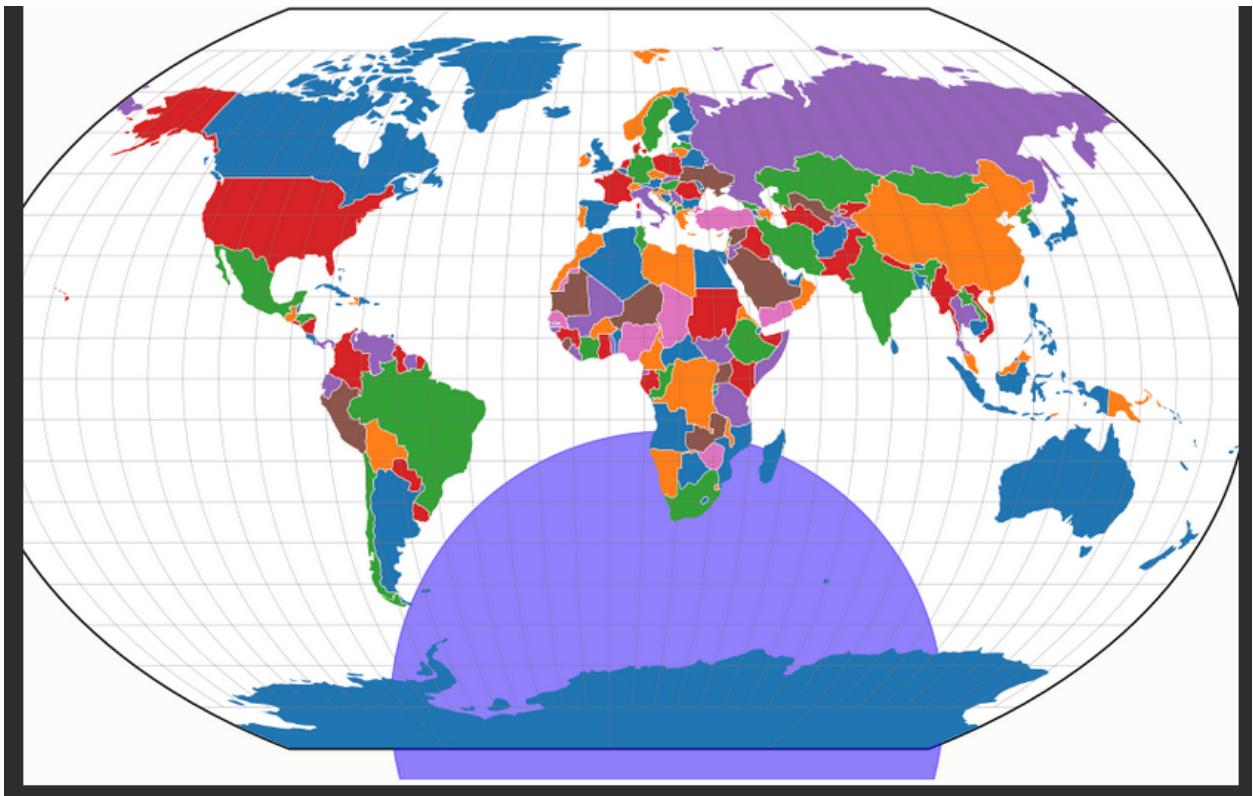


Representatives





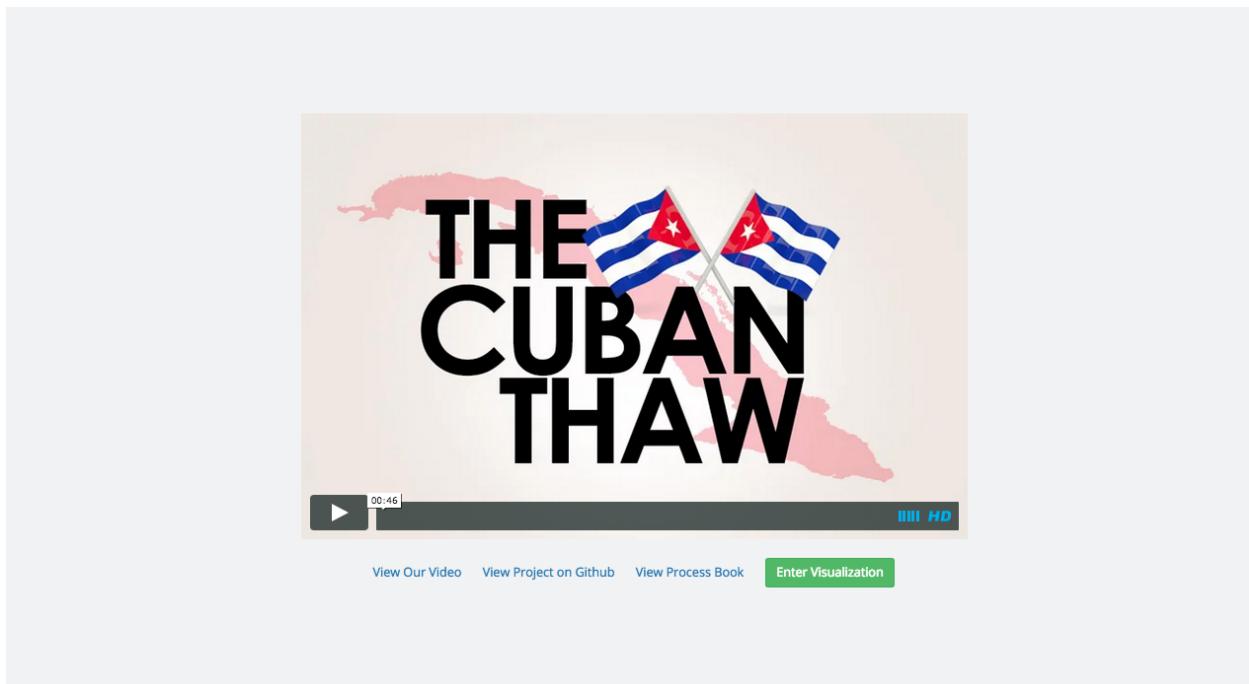




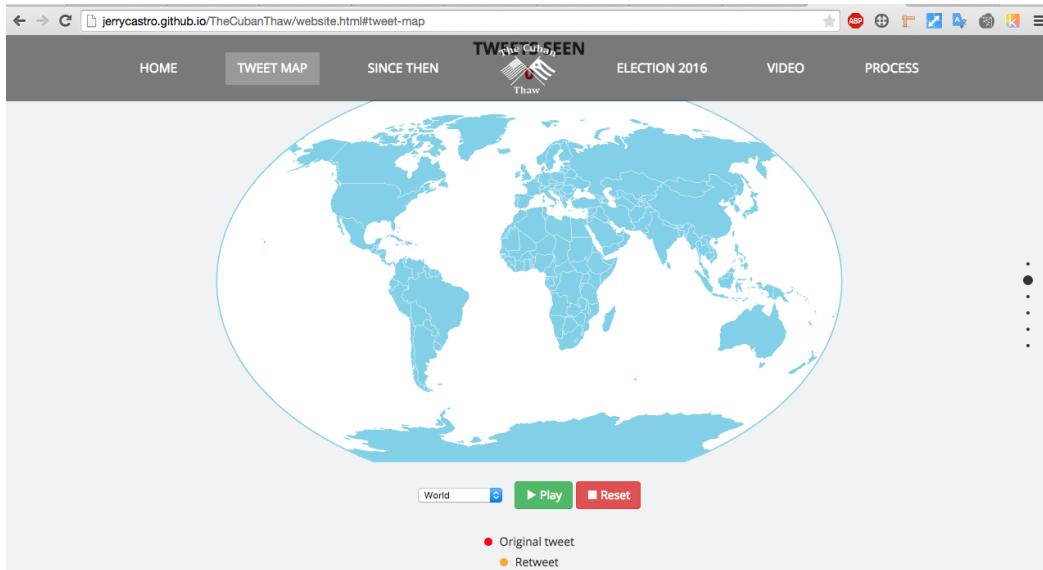
Implementation

Describe the intent and functionality of the interactive visualizations you implemented.

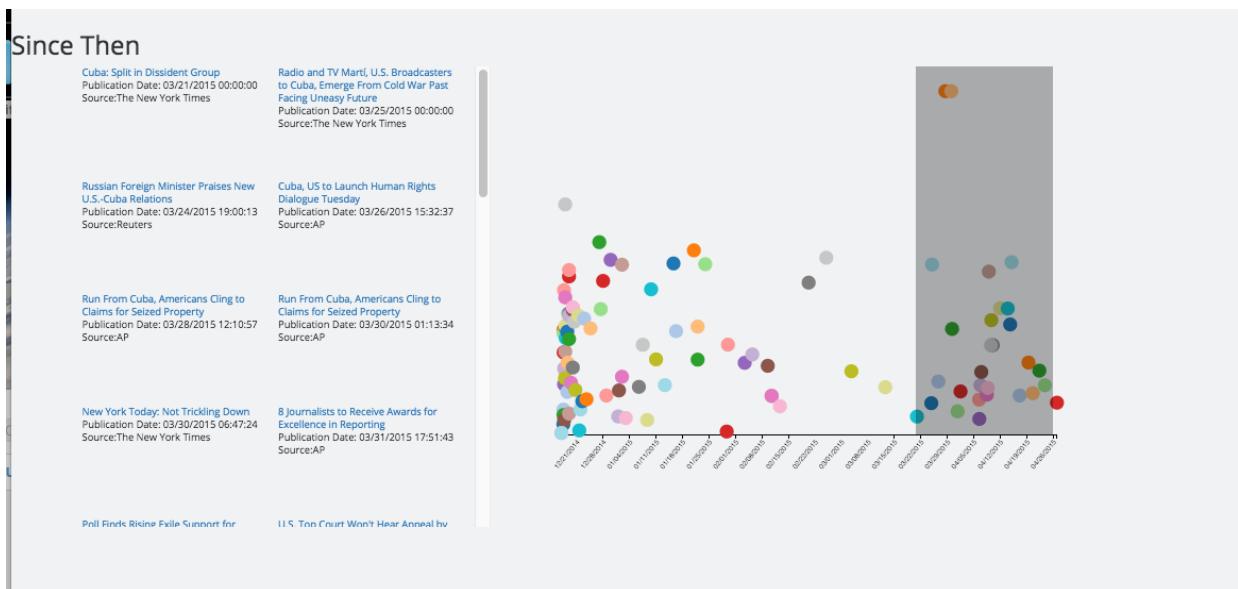
Provide clear and well-referenced image



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



This is the tweet map of our dataset in real time (accelerated , but still real relevant to the user's frame of reference. 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:

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?

The US embargo of cuba was a very big deal to people in Latin America and Europe. The Thaw crossed into the mainstream conversation with many ordinary people commenting on it.. Also, even though Twitter skews young, Florida still erupted with activity indicating a young cuban population is politically engaged which could prove dangerous to some politicians.

We had no idea Venezuela's Twitter has such an extensive presence. as much as it did. Venezuela has been Cuba's closest ally over last 10 years. That ended with death of Hugo Chavez and now the collapse of oil which Venezuela's economy depends on as one of the largest exporters of petroleum in the world.

We chose this dataset because we felt strongly that our dataset is truly a significant part of history. Not only did it capture a historical detente, but it did 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