

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.

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. We designed it so that users would make their own sentiment judgements based on the contextual information in the visualization alongside a list of top hashtags. An embedded twitter feed also enriched the experience by allowing the user to consider the text of the tweet as well.

Data

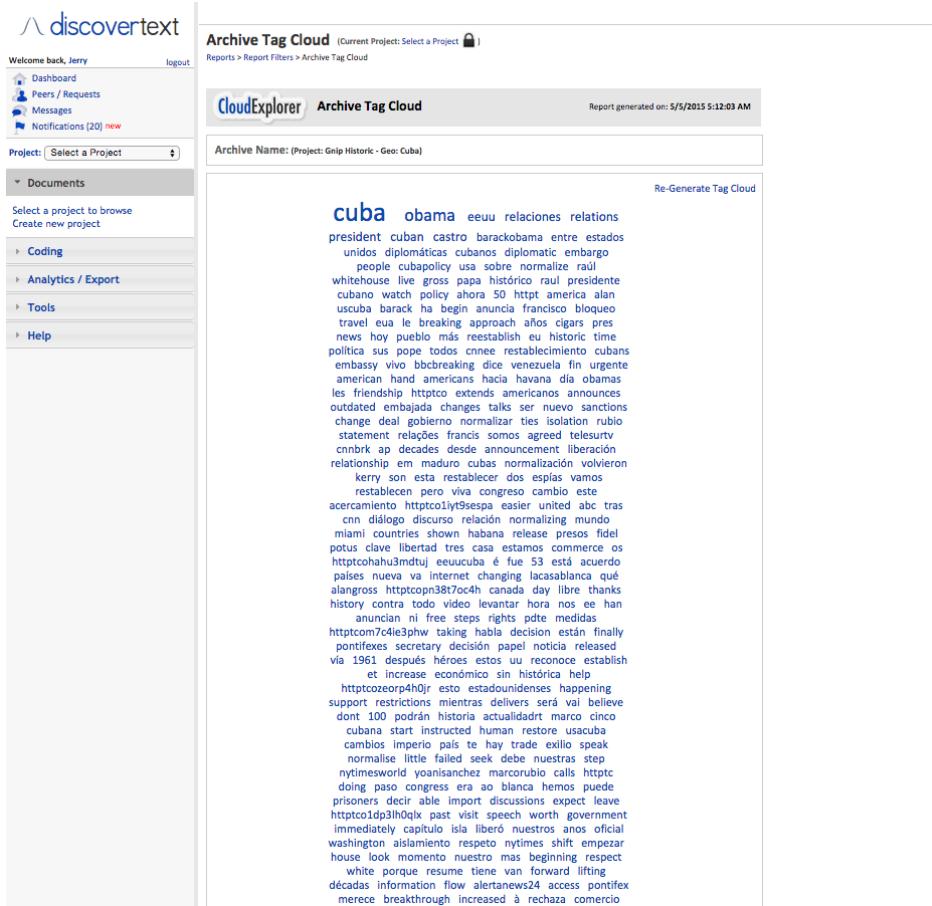
Source, scraping method, cleanup, etc.

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																												
L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD										
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:	#####	#####ColombiaD	Twitter	for A	16131	http://www.DedicatedC Colombia	NTN24	NTN24	http://www.chefaronna															
2	r.com/chefaronna/statuses/54506994994289152				#####	Rene Beaulieu	Twitter	Web	2869	http://www.Politically ob WI	NowWithale	NowWithale	NowWithale	NowWithale	http://www.Bluegrl_3													
3	r.com/Bueglir_3/statuses/545040697003762688				#####	Daniel Bergin	TweetCaster		1255	http://blogs.Commentate	Nebraska	Wall Street J WSJ	http://www.dbergman39															
4	r.com/dbergman39/statuses/545040697217684480				#####	Lo'Cat Gauthi	Twitter	for A	59	J'a une devi La Ba	Martin Pett	Lemicrofede	http://www.famousred															
5	r.com/famousred/statuses/545040698069115904				#####	Gary J. Esen	Twitter	Web	482	Nashua, NH	Top Conservi	TeaPartyC	http://www.GEsneault															
6	r.com/GEsneault/statuses/5450406988004886914				#####	Gisselle Mori	TweetDeck		3353	http://www.Periodista de Cuba	Escambray	F escambray	http://www.gissellemr															
7	r.com/gissellemr/statuses/5450406988250956				#####	Dani	Twitter	for A	1925	http://www.Vijar y ser Lina - PerA	Noticias SIN	SIN24horas	http://www.dancabrer21															
8	r.com/dancabrer21/statuses/54504069767821312				#####	AmericaNeer	Twitter	Web	1177	http://twitte Serial entrep	North Caroli	Daily Be	the dailybeas	http://www.angel_cortez78														
9	r.com/angel_cortez78/statuses/545040696017024512				#####	Camillo Fdo	V Twitter	Web	3275	Comments	Leonardo Pao	Leonardo_Po	http://www.cam_vallejo															
10	r.com/cam_vallejo/statuses/545040698899174401				#####	caryl mast	Twitter	Web	37252	http://twitte colorado girl	Kenneth Isa	uselephants	http://www.67purple															
11	r.com/67purple/statuses/5450406979954548				#####	Luis Moreira	Plume&forA		26234	The views an Trujillo Alto,	AndrAos Co	andrescolon	http://www.jullopri															
12	r.com/jullopri/statuses/5450406969230360				#####	Miram R.	Twitter	for A	16759	Educarde a Caracas	ManoloReve	ManoloReve	http://www.annirrau															
13	r.com/annirrau_7/statuses/5406971965628416				#####	Tom Scanlon	Twitter	V	10508	http://media Re-elect	Mill Hyattsville	Maryland, USA	http://www.Parcival2															
14	r.com/Parcival2/statuses/54504069734451225				#####	Tom McGev	Twitter	Web	15232	Co-founder	New York	Tom McGev	tmgev	http://www.tmcgev														
15	r.com/tmcgev/statuses/545040697147524608				#####	Conservative	Twitter	iP	1495	http://twitte Proudly follo	for iP	Ban Collectiv	http://www._People															
16	r.com/_People/statuses/545040697147524608				#####	Alison M. D.	Twitter	for A	2494	http://news.Religious	53, Norway	http://www.IslamInNorway	http://www.IslamInNorway															
17	r.com/IslamInNorway/statuses/545040697147524608				#####	Aslihan Yilmaz	Twitter	for A	21199	http://Aspiring sup	Anchorage	A Top Conserv	TeaPartyCat	http://www.AtheisRaven														
18	r.com/AtheisRaven/statuses/545040697147524608				#####	Analog_Zero	Twitter	Web	137347	Comments	Barrenquila	Hassan Nassar	http://www.Analog_Zero															
19	r.com/Analog_Zero/statuses/54504069714779456				#####	CÓZAR Barr	Twitter	Web	9547	http://www.Comienza tu DF	Retired	Retired	Retired	http://www.charrons														
20	r.com/charrons/statuses/545040697157963825				#####	Franklin	Twitter	A	1654	http://actual Chavita, Re Venezuela	RT en EspAñ	Actualidad&R1	http://www.frankalbarra															
21	r.com/frankalbarra/statuses/545040697157963825				#####	Marilyn Oliv	Twitter	for A	14496	http://Dios, Esposa, MÁxicos	Retired	Retired	Retired	http://www.MTMarilyn2														
22	r.com/MTMarilyn2/statuses/54504069722437377				#####	Sonia Presa	Twitter		15045	http://Dios, Esposa, MÁxicos	Retired	Retired	Retired	http://www.sospr_														
23	r.com/sospr_/statuses/545040698068623236				#####	Ngee	Twitter	A	1165	http://Assalamuloul soul, Korea	MUJAHD	SAFWANMU	http://www.eyranda97															
24	r.com/eyranda97/statuses/545040698068624048				#####	Ramon Andri	Twitter	A	1385	Docente de l'Varcuy Veni Mario Silva	CLahjilljaenT	El Delgado	106	http://www.Delgado_106														
25	r.com/Delgado_106/statuses/545040697632117056				#####	Jorge Moreno	Twitter	Web	4364	Proyecto inn Maracay	MABEL BORG	2405mabel	http://www.proyecto_eps_b															
26	r.com/proyecto_eps_b/statuses/5450406975609472				#####	Dany Suare	Twitter	for IF	10285	Managemen	La Habana, Cuba	12054	http://www.DanaySuarez															
27	r.com/DanaySuarez/statuses/545040697808928780				#####	Beto Moreno	twitterfeed		10646	http://mi nvia y V DF	12055	http://www.betomoreno5																
28	r.com/betomoreno5/statuses/545040697721025664				#####	DamÁn Igle	twitterfeed		10644	http://DESCUBRIMT Ecatepec,	MÁxico	http://www.digilless																
29	r.com/digilless/statuses/5450406976562336				#####	Sonia Presa	twitterfeed		15044	http://Dios, Esposa, MÁxicos	http://www.sospr_																	
30	r.com/sospr_/statuses/545040698467565568				#####	Els Clement	Twitter	Web	15241	http://Socialista, re Venezuela	Fidel Castro	fidelcastro	http://www.Riveroconterreas															
31	r.com/Riveroconterreas/statuses/5450406979251138560				#####	Brie Handgraft	TweetCaster		1455	http:/mone I'm a reporte Rocky Mount	CNNMoney	CNNMoney	http://www.BHandgraft_RMT															
32	r.com/BHandgraft_RMT/statuses/5450406980151692				#####	Finita	Twitter	A	18057	http://twitte Venezuela.	VENEZUELA	El monitor	1Emonitor1B	http://www.Xjesman30														
33	r.com/Xjesman30/statuses/5450406980327552				#####	Conservative	Twitter	Web	14203	http://Passionate	the Texas USA	Katherine Mi	katherinemil	http://www.allovestexas														
34	r.com/allovestexas/statuses/54504069812769472				#####	Marcos Mez	twitterfeed		10164	http://Los sueAos	MÁxico	http://www.mmn9marcos																
35	r.com/mmn9marcos/statuses/5450406980138396384				#####	Proudamer	Twitter	Web	13535	Independent USA	Bill Damato	Bill111; me	http://www.1776betsyRoss															
36	r.com/1776betsyRoss/statuses/5450406980138396384				#####	Jalyn Henton	Twitter	A	12193	Doing my be Del, Alexandria, VA	Yoani SAIN	yoanisoanische	http://www.Longueira_2018															
37	r.com/jalynhenton/statuses/5450406980681375744				#####	Pablo Longo	Twitter	for IF	1745	http://TWITTER de Chile	Yoani SAIN	yoanisoanische	http://www.Longueira_2018															
38	r.com/Longueira_2018/statuses/545040698607527566				#####	M. Farooq Al	Twitter	B	4072	http://Businessman Pakis	Reuters	Top Reuters	http://www.mfarooqafai7															
39	r.com/mfarooqafai7/statuses/545040698452726528				#####	Catherine C.	dirtv		27628	Morelia, MichoacAin	http://www.catherine_czill																	
40	r.com/catherine_czill/statuses/545040698627911361				#####	Pam Margi	dirtv		37082	Amo las flores, Morelia, MichoacAin	http://www.Pamadrigal_																	
41	r.com/digitalcamrona/statuses/5450406980916930680				#####	Jose Luis Car	Twitter	A	2920	http://calimetric Granada, Sp	Teacher	mai_magia	http://www.digitalcamrona															
42	r.com/INGJOSEMANUEL/statuses/54504069835328934				#####	JOSE RODRIC	Twitter	B	17833	Ingenier Int	VENUEZA	http://www.INGJOSEMANUEL																
43	r.com/trilect_noma/statuses/54504069856576756				#####	Nome Drilect	nlvr.it		48656	Electrmei	MÁxicin. DF	nlvrlet_noma	http://www.nlrlt_noma															
44	r.com/nomelabel/statuses/54504069856576756				#####	Has_Geo_Cuba_9-export-20150412			581	41	http://twitte	nlvrlet_noma	nlvrlet_noma	http://www.Has_Geo_Cuba_9-export-20150412														

Has_Geo_Cuba_9-export-20150412-113248.csv																										
L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL
1	is_a	Title	Text	Untitled	[M] country, cc	[M] favorites	[M] follower	[M] friends	[M] hashtag	[M] id	[M] location	[M] link	[M] location	[M] media_d	[M] media_t	[M] media_u										
2	RT @NTN24: [V RT @NTN24: [Video] &CoEl R@gimen cubano segun	2	102	820	tag:search.twitter.com,20	http://twitter.com/cheffaronna	/statuses/5450406994994289152																			
3	RT @NowWithalex: [NowWithalex: Today the United States of Amer	3	183	418	tag:search.twitter.com,20	http://twitter.com/NowWithalex	/statuses/5450406975609472																			
4	RT @W@TheUnited: The United States of America has been remittances to Cuba over past decades	4	225	312	tag:search.twitter.com,20	http://twitter.com/W@TheUnited	/statuses/545040698607527566																			
5	RT @monolotek: RT @monolotek: Repatriation of Cuban	5	69	62	tag:search.twitter.com,20	http://twitter.com/monolotek	/statuses/5450406980681375744																			
6	RT @TeaPartyC: RT @TeaPartyCat: That awkward moment when Rep	5	22	127	tag:search.twitter.com,20	http://twitter.com/TeaPartyC	/statuses/545040697147524608																			
7	Relaciones #Cuba #Estados Unidos El deshielo visto por medio	463	352	Cuba; Estad	102	tag:search.twitter.com,20	http://twitter.com/Relaciones	/statuses/545040697147524608																		
8	RT @SN24: RT @SN24: Obama: Today he is the first to visit Cuba	136	342	352	Twitter	1005	tag:search.twitter.com,20	http://twitter.com/SN24																		
9	RT @HassNassat: RT @HassNassat: Cambien relaciones con Cuba mi	68	76	217	tag:search.twitter.com,20	http://twitter.com/HassNassat	/statuses/545040698607527566																			
10	RT @Leonardo: RT @Leonardo: Padra -A dA-hi! Rico - USA y Cuba	97																								

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:

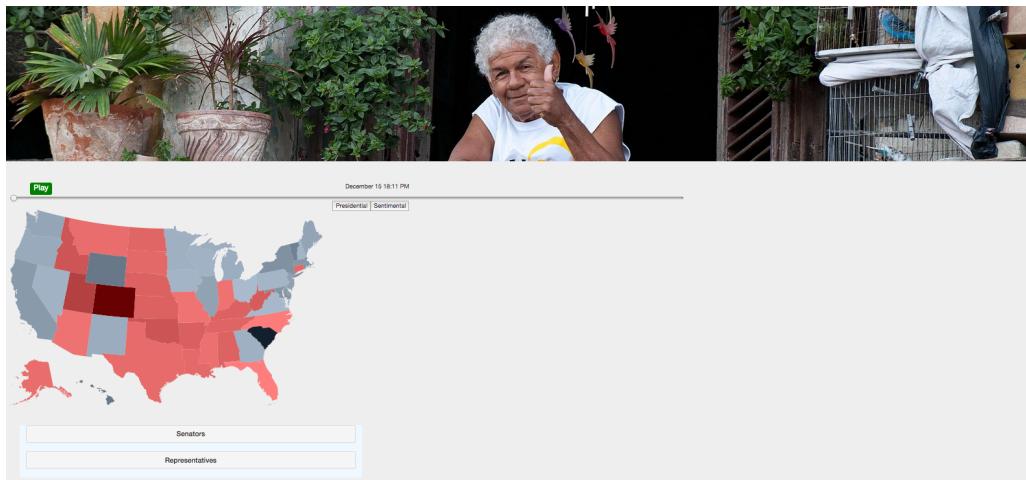
What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?

- ❖ 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.
- ❖ We learned this was a massive dataset.
- ❖ Semantic Analysis is very very hard. Entities vs Documents, phrases, language detection (n-grams).
- ❖ Realization that a proper semantic analysis would not be feasible led to the pivot.

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.

- ❖ Struggles choosing colors for tweets and maps
- ❖ Deciding between greater precision and greater aesthetics.
- ❖ Deciding whether to aggregate and how to aggregate.
- ❖ 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?”
- ❖ Choosing a “One page” website design over a traditional multi-page design in terms of story telling.
- ❖ 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.



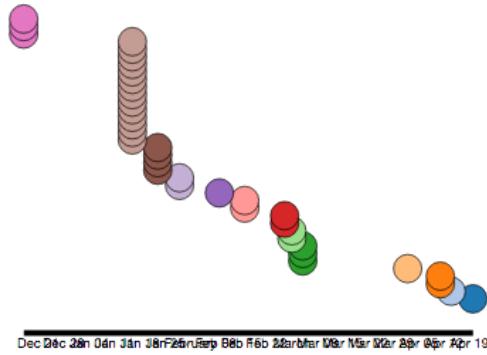
The Next President & Cuba



TheCubanThaw-master.zip
Download error



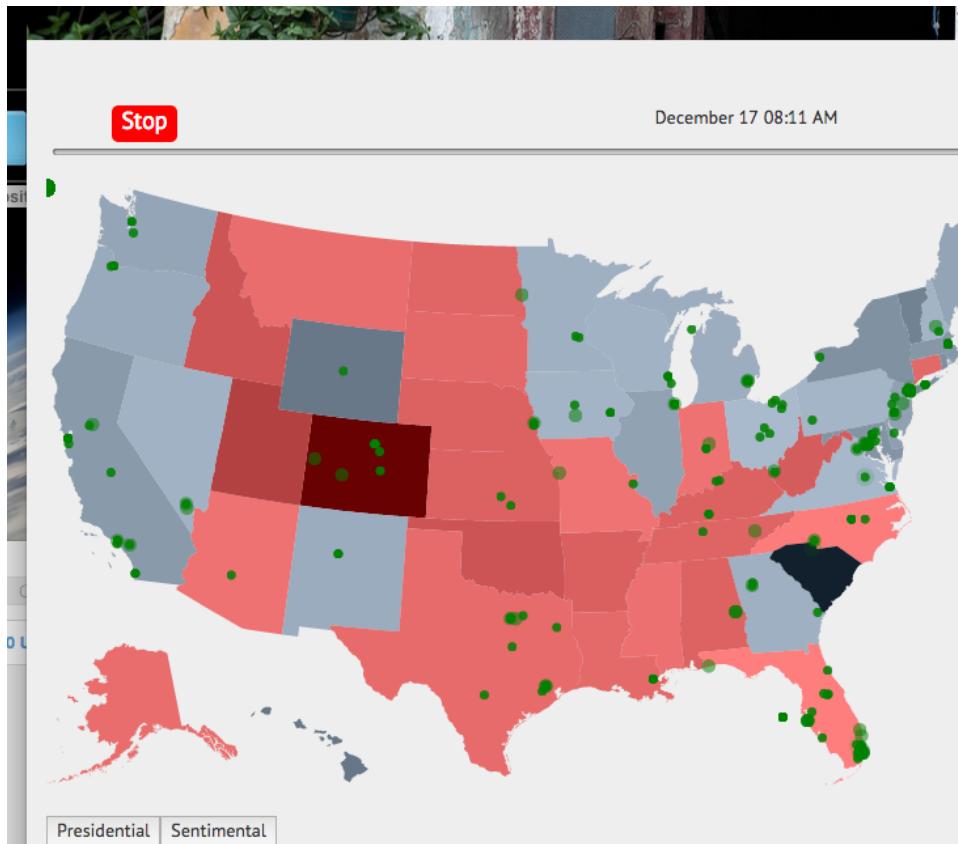
Thaw Events



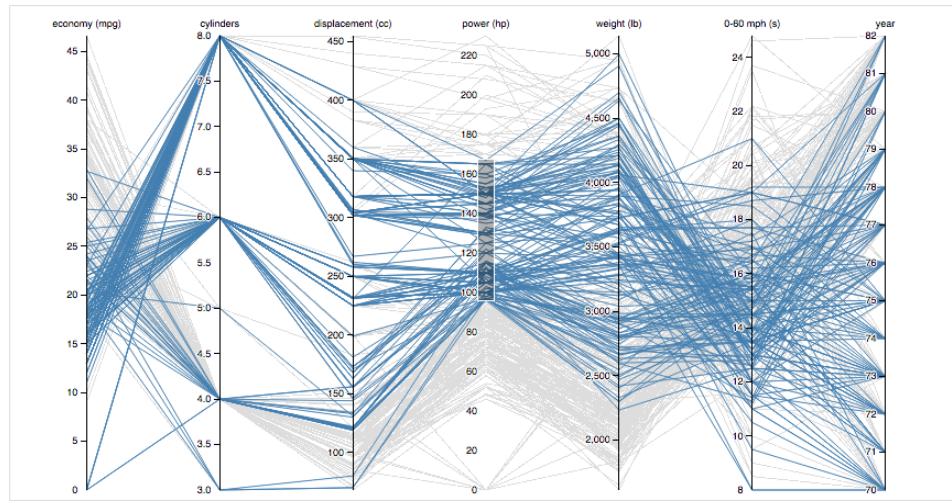
Screenshot of an Excel spreadsheet titled "ivan_sentiment_copy.xlsx". The spreadsheet contains a large dataset of tweets with columns for ID, Source Text, Summary, Detected Language, Document Language, Phrase Intensity, Phrase Sentiment, and various entity and entity type columns. The data spans from row 1 to approximately row 2000.

ID	Source Text	Summary	Detected Language	Document Language	Phrase Intensity	Phrase Sentiment	Entity	Entity Type	Entity	Entity Type	Entity	Entity Type
1	afforded enough! Let's move on	Cuba	English	English	5	0.092 neutral	End/Thembargo	Cuba	Place	1	positive	
4232-275-	Cuba and help express some dissidents!!	Cuba	English	English	15	-0.092 negative	express	0.092 negative	r7 Obama	Place	-0.95 negative	
1518	Love this move by Time for Cuba to make progress	Cuba	English	English	8	0.1695131 neutral	progress	0.1695131 neutral	Cuba	Place	0.4147926 neutral	
4280-3869-	GDP is an uncomfortable position. It'll be interesting to	Cuba	English	English	16	-0.036427 neutral	interesting	0.49 positive	Cuba	Place	-0.036427 neutral	
1519	normalize with Cuba	Cuba	English	English	16	-0.036427 neutral	good	0.49 positive	Cuba	Place	-0.036427 neutral	
4281-3916-	Tim Zougmari and his group are great!	Cuba	English	English	4	0.49 positive	best	0.6 positive	Normalizaci&onOfCu	Person	0.49 neutral	
1520	Normalizaci&onOfCu Dignity at its best.	Cuba	English	English	4	0.49 positive	Hopefully	0.28196 positive	Cuba	Place	0.50245 positive	
4251-39aa-	The people of Cuba will have better lives because of	Cuba	English	English	18	0.06732 neutral	Happy	0.2 neutral	Alan	Person	0.15 neutral	
4282-39ae-	Happy Hammock to Alan and Judy Gross! Cuba	Cuba	English	English	3	0.2 neutral	Happier	0.2 neutral	Alan	Person	0.15 neutral	
1521	make a difference	Cuba	English	English	15	-0.11322 negative	still	0.092 negative	r7 make	Place	-0.11322 negative	
4660-4f11-	cuba...this is a nice photo op and thumb in the eye	Cuba	English	English	18	-0.2074 negative	no	0.092 negative	photo	Cuba	Place	-0.1148 neutral
1522	nonsense, there why we are	Cuba	English	English	10	-0.231158 negative	mad	-0.231158 negative	remember	Cuba	Place	-0.231158 negative
4283-4f5e-	remained why we are	Cuba	English	English	12	-0.231158 neutral	no	0.092 neutral	Company	Company	0.092 neutral	
1523	can never even remember why we are mad at Cuba!	Cuba	English	English	12	-0.231158 neutral	no	0.092 neutral	Cuba	Place	0.092 neutral	
4284-4f61-	Many places there including publications	Cuba	English	English	12	-0.231158 neutral	no	0.092 neutral	Cuba	Place	0.092 neutral	
1524	telling me you're not	Cuba	English	English	10	0.446032 positive	human rights	0.446032 positive	telling	Cuba	Place	1.1598 positive
4524-a61-	Cuba's human rights record?	Cuba	English	English	12	0.162 neutral	so	0.75 positive	Many places	Cuba	Place	0.034 neutral
1525	Many places there including publications	Cuba	English	English	12	-0.231158 neutral	so	0.75 positive	Cuba	Place	0.034 neutral	
4285-4f71-	Time for Cuba to make progress	Cuba	English	English	15	-0.231158 neutral	so	0.75 positive	Cuba	Place	0.034 neutral	
1526	Time for Cuba to make progress	Cuba	English	English	15	-0.231158 neutral	so	0.75 positive	Cuba	Place	0.034 neutral	
4212-4f72-	Hispanic vote, Cubans in Miami have been able to	Cuba	English	English	15	-0.6 negative	idiotic	-0.6 negative	votes	Cuba	Place	-0.5 negative
1527	window dressing	Cuba	English	English	15	-0.5 negative	pandering	-0.4 negative	legally go	Miami	Place	-0.5 negative
4213-4f73-	Hispanic vote, Cubans in Miami have been able to	Cuba	English	English	15	-0.5 negative	so	0.092 neutral	real	Cuba	Place	0.092 neutral
1528	window dressing	Cuba	English	English	15	-0.5 negative	so	0.092 neutral	real	Cuba	Place	0.092 neutral
4287-4f74-	CartoonistAga<img alt="Crime in 1993, my strongest	Cuba	English	English	11	0.0384535 neutral	Crime	-0.240394 neutral	Founder	Job Title	0.037953 neutral	
1529	policy pivot. In real life, it was Ben Rhodes	Cuba	English	English	15	0.106 neutral	in	0.212 neutral	Leo	Person	0.106 neutral	
4661-4f91-	who orchestrated the	Cuba	English	English	15	0.106 neutral	real	0.212 neutral	Ben Rhodes	Person	0.106 neutral	
1530	policy pivot. In real life, it was Ben Rhodes	Cuba	English	English	15	0.106 neutral	real	0.212 neutral	Sen. Marco Rubio	Person	0.106 neutral	
4288-4f91-	who orchestrated the	Cuba	English	English	15	0.106 neutral	real	0.212 neutral	Obama	Person	0.106 neutral	
1531	policy pivot. In real life, it was Ben Rhodes	Cuba	English	English	15	0.106 neutral	real	0.212 neutral	Sen. Marco Rubio	Person	0.106 neutral	
4786-4f91-	Runs Congress to Embraze Cuba 3. Shortage solved	Cuba	English	English	5	0.11 positive	Embrace	1.1 positive	Cuba	Place	1.1 positive	
1532	of Doctor Shortage v.2	Cuba	English	English	18	-0.231158 negative	brutal	-0.6 negative	as isolated	China	Place	-0.36453 neutral
4289-4f91-	Zimbabwe, and Egypt Regimes that use torture are still	Cuba	English	English	18	-0.231158 negative	bad	-0.6 negative	as isolated	Zimbabwe	Place	-0.36453 neutral
1533	isolated at China,	Cuba	English	English	18	-0.231158 negative	torture	-0.49 negative	as isolated	Zimbabwe	Place	-0.36453 neutral
4290-4f91-	Zimbabwe, and Egypt Regimes that use torture are still	Cuba	English	English	18	-0.231158 negative	Just look	0.080835 neutral	Egypt	Place	-0.36453 neutral	
1534	isolated at China,	Cuba	English	English	18	-0.231158 negative	violations	-0.12 negative	r7 believe	U.S.	Place	-0.12 neutral
4291-4f91-	relationships with Cuba after all the human rights	Cuba	English	English	18	-0.231158 negative	violations	-0.12 negative	U.S.	Place	-0.12 neutral	
1535	violations	Cuba	English	English	18	-0.231158 negative	violations	-0.12 negative	U.S.	Place	-0.12 neutral	
4284-4f91-	impositioned response to new Cuba policy. Speaking	Cuba	English	English	9	0.0493933 neutral	impositioned	0.168 neutral	Sen. Marco Rubio	Person	0.42 neutral	
1536	Rubio now making	Cuba	English	English	15	0.106 neutral	tyrants	-0.6 negative	Obama	Person	0.42 neutral	
4287-4f91-	of coding tyrants	Cuba	English	English	15	0.106 neutral	tyrants	-0.6 negative	Sen. Marco Rubio	Person	0.42 neutral	
1537	of coding tyrants	Cuba	English	English	15	0.106 neutral	tyrants	-0.6 negative	Sen. Marco Rubio	Person	0.42 neutral	
4288-4f91-	No. Korea. All terrorists States. We g	Cuba	English	English	6	0.27244 negative	All	-0.27244 negative	Iran	Place	-0.8811 negative	
1538	opens relations with	Cuba	English	English	6	0.27244 negative	terrorists	-0.27244 negative	Iran	Place	-0.8811 negative	
4289-4f91-	opens relations with	Cuba	English	English	6	0.27244 negative	terrorists	-0.27244 negative	Iran	Place	-0.8811 negative	
1539	of diplomatic ties w/US, say differences remain.	Cuba	English	English	8	0.2757449 positive	restoration	0.404857 positive	U.S.	Place	0.2757449 neutral	
4290-4f91-	of diplomatic ties w/US, say differences remain.	Cuba	English	English	8	0.2757449 positive	diplomaticies	0.139804 neutral	Raul Castro	Person	0.2757449 neutral	
1540	of diplomatic ties w/US, say differences remain.	Cuba	English	English	8	0.2757449 positive	restoration	0.404857 positive	U.S.	Place	0.2757449 neutral	
4291-4f91-	of diplomatic ties w/US, say differences remain.	Cuba	English	English	8	0.2757449 positive	diplomaticies	0.139804 neutral	Raul Castro	Person	0.2757449 neutral	
1541	negotiator we have ever had - Cuba, he home, did not	Cuba	English	English	19	0.0493933 neutral	freedom	0.272 positive	president	Job Title	0.0493933 neutral	
4292-4f91-	negotiator we have ever had - Cuba, he home, did not	Cuba	English	English	19	0.0493933 neutral	worst	-0.196 negative	Cuba	Place	0.0493933 neutral	

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.



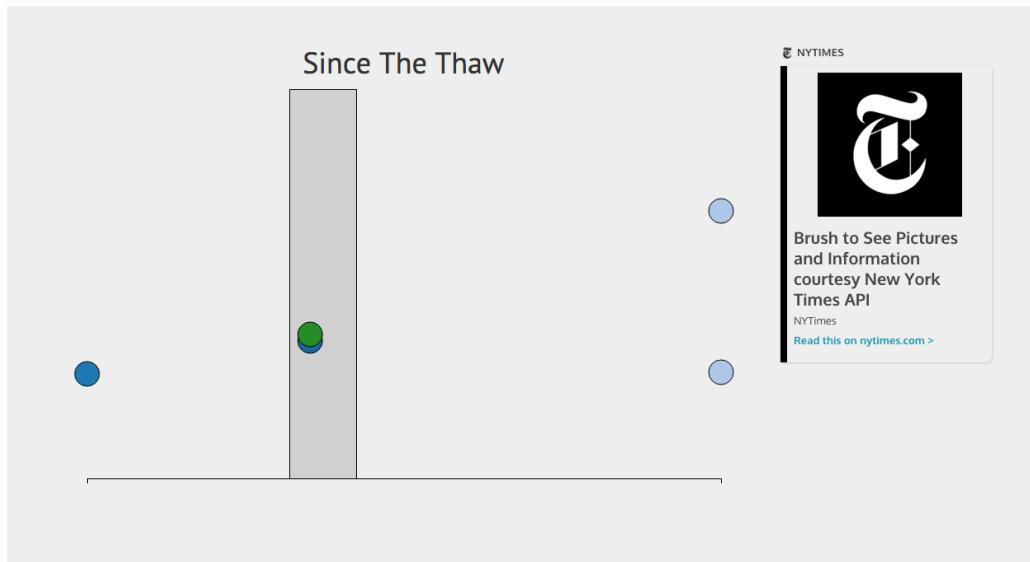
Parallel Coordinates



This is a version of Mike Bostock's [parallel coordinates example](#), modified to include reorderable axes.

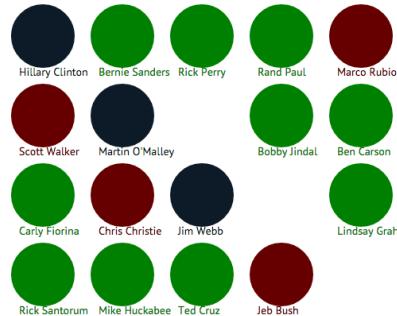
[Open in a new window.](#)

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

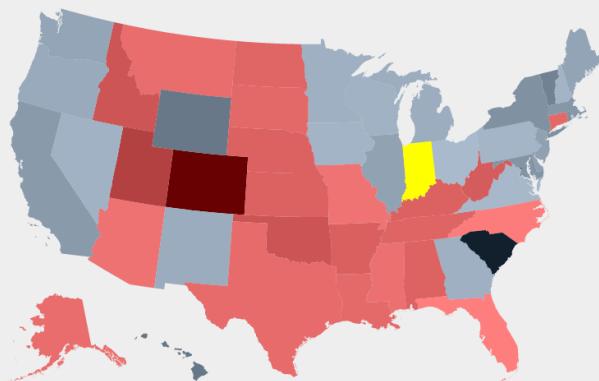


[For | Against]

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

Play

December 15 18:11 PM



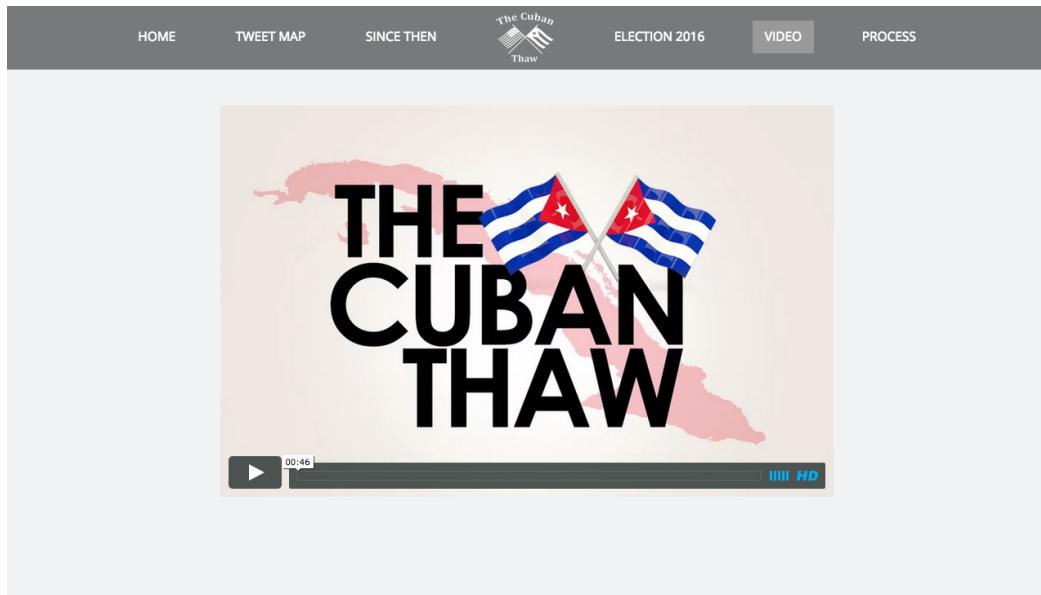
[Presidential | Sentimental]

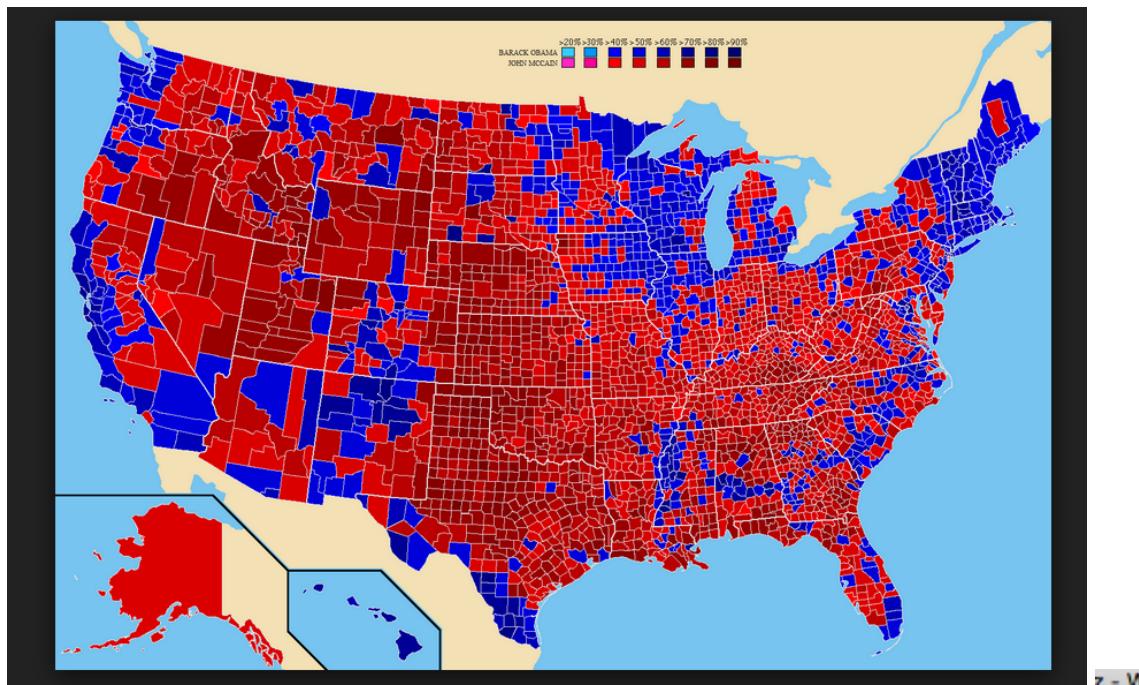
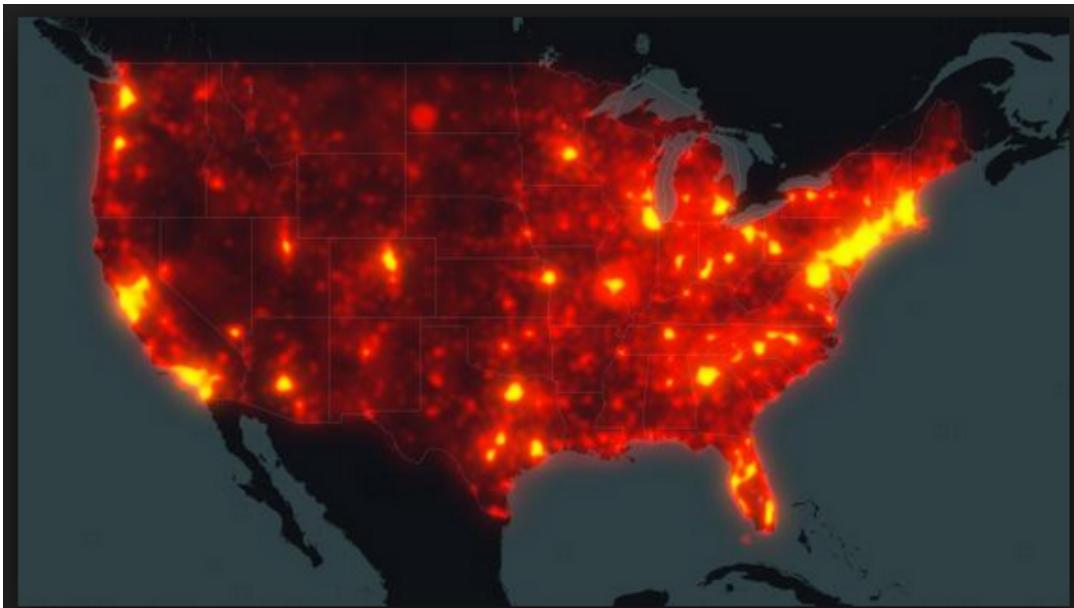
Senators



Representatives







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:

- 1) Take 2-3 months and dive extremely deeply into semantic analysis (not recommended, there are better ways)
- 2) 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.
- 3) We could have tweets of a certain hashtag increase the proportion of that color in a country's color, thus country takes on the color of most popular hashtag

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

Since Then

Cuba Split in Dissident Group
Publication Date: 03/21/2015 00:00:00
Source: The New York Times

Radio and TV Martí, U.S. Broadcasters
to Cuba, Emerge From Cold War Past
Facing Uncertain Future
Publication Date: 03/25/2015 00:00:00
Source: The New York Times

Russian Foreign Minister Praises New
U.S.-Cuba Relations
Publication Date: 03/24/2015 19:00:13
Source: Reuters

Cuba, US to Launch Human Rights
Dialogue Tuesday
Publication Date: 03/26/2015 19:32:37
Source: AP

Run From Cuba, Americans Cling to
Claims for Seized Property
Publication Date: 03/28/2015 12:10:57
Source: AP

Run From Cuba, Americans Cling to
Claims for Seized Property
Publication Date: 03/30/2015 01:13:34
Source: AP

New York Today: Not Trickling Down
Publication Date: 03/30/2015 06:47:24
Source: The New York Times

8 Journalists to Receive Awards for
Excellence in Reporting
Publication Date: 03/31/2015 17:51:43
Source: AP

Poll Finds Rishie Singh's Support for

U.S. Tom Cruise Won't Hear About It

