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

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

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

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

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

Related Work

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 way of measuring the sentiment of a tweet. Thankfully, those tens of hours engaged in semantic analysis of the tweets meant we both had logged countless hours staring at this data. All of it payed off when we finally realized that for a highly politicized event like the Cuban Thaw, the hashtags themselves are actual flags of sentiment.

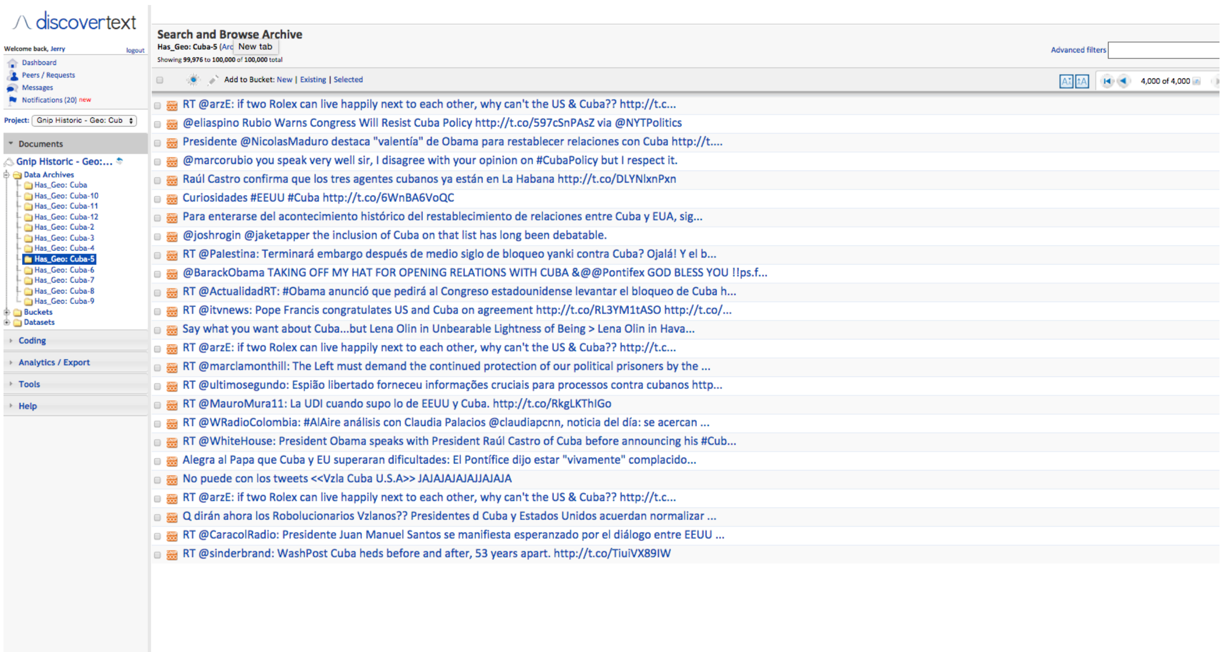
During our brainstorming we also stumbled upon a couple tutorials that explain how to geo-code location strings. We got to work and after the usual repasted failers, we were successful at geo-coding almost 990,000 records. Again we had our full dataset that spanned the world.

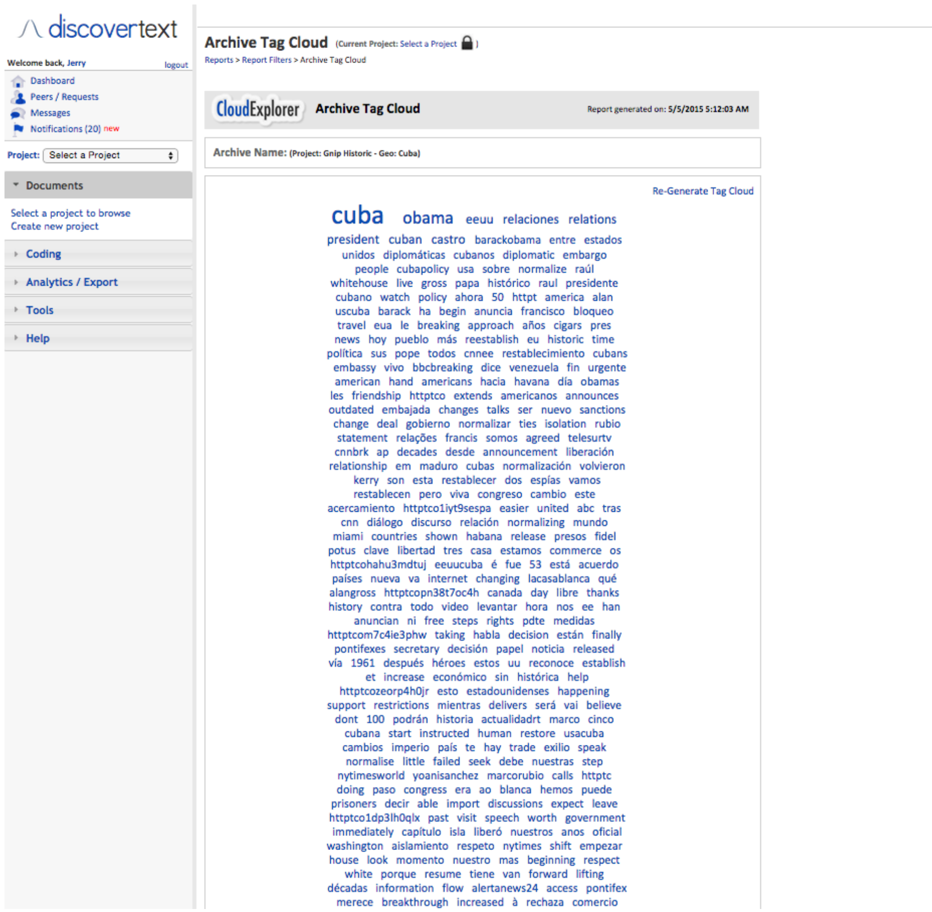
Our solution was to double down on Big Data and put the user's focus on a global tweet map. We designed it so that users would make their own sentiment judgements based on the contextual information in the visualization alongside a list of top hashtags. An embedded twitter feed also enriched the experience 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.





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 import to use multiple geocoding services, so that we did not overload any one free service.

Once available to our web application in CSV and JSON formats, we used standard D3.js geographic projection functions (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 --attributions). When it comes to locations, some users don't provide accurate information, or use the field for metaphorical descriptions. The geocoders may misunderstand intent. *Et cetera.*

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

Human oversight is reasonably important in managing such data pipelines, since you want to incrementally improve the data. It often helps to have someone watching early failures to realize that "You know, a lot of these geocoding failures are on locations that end with a period--and that period does seem out of place here. What if, on locations that fail on the first attempt, we remove the final period and try again?" Or "A lot of people sure do want to make their country #USA a hashtag. Maybe geocoders aren't savvy to that. If we see things that look like a hashtag, let's remove that and try again." Such rule-based cleanups dramatically improve data coding effectiveness.

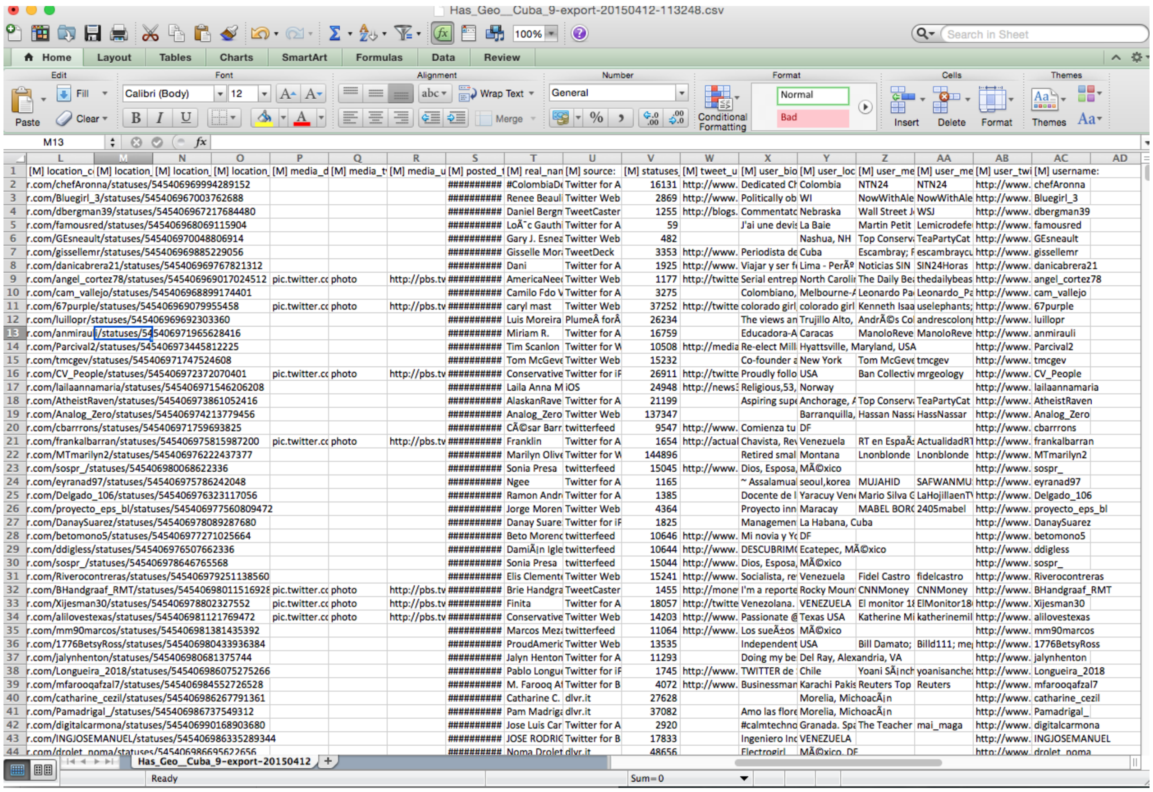
Infrastructure

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

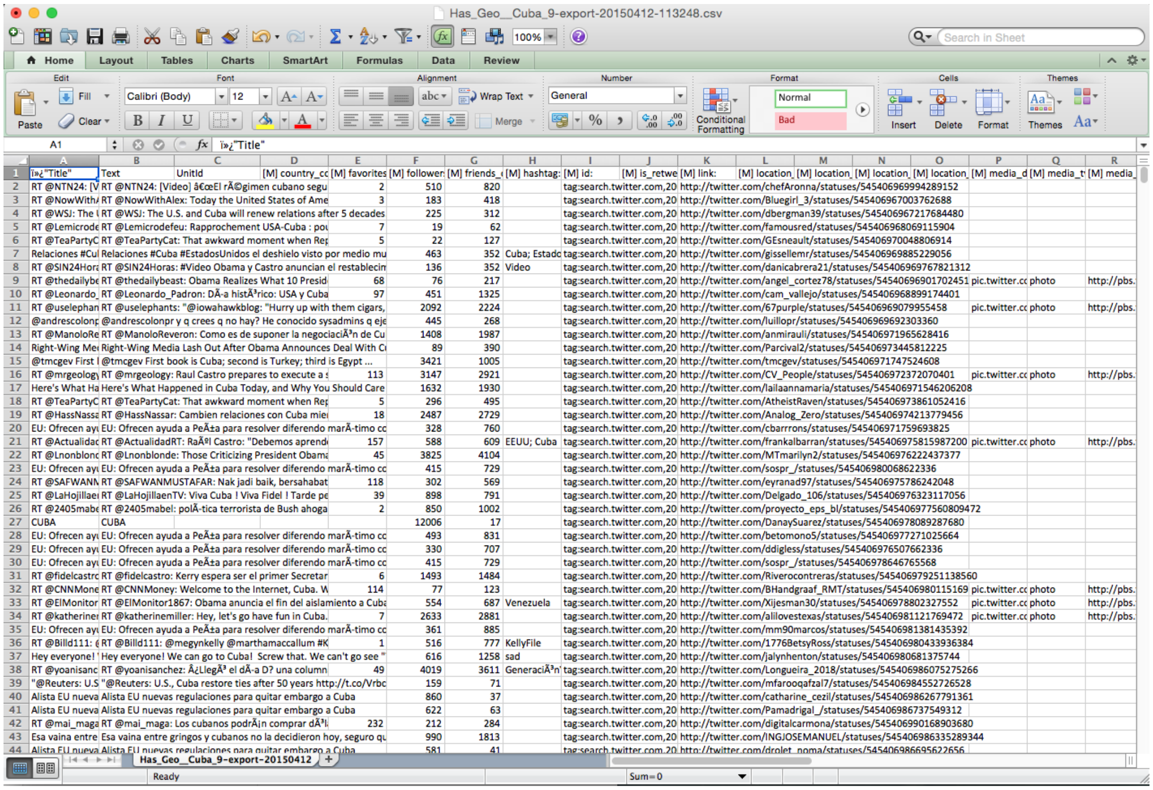
Subdividing the data for the multiple servers was a bit of a chore. We wrote some simple `bash` (Unix shell language) scripts to help automate it, but more work there would be good for dealing with large data sets.

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

Geocoding automation was a huge win, but introduced its own complexities and a bit of variance in the resulting data.



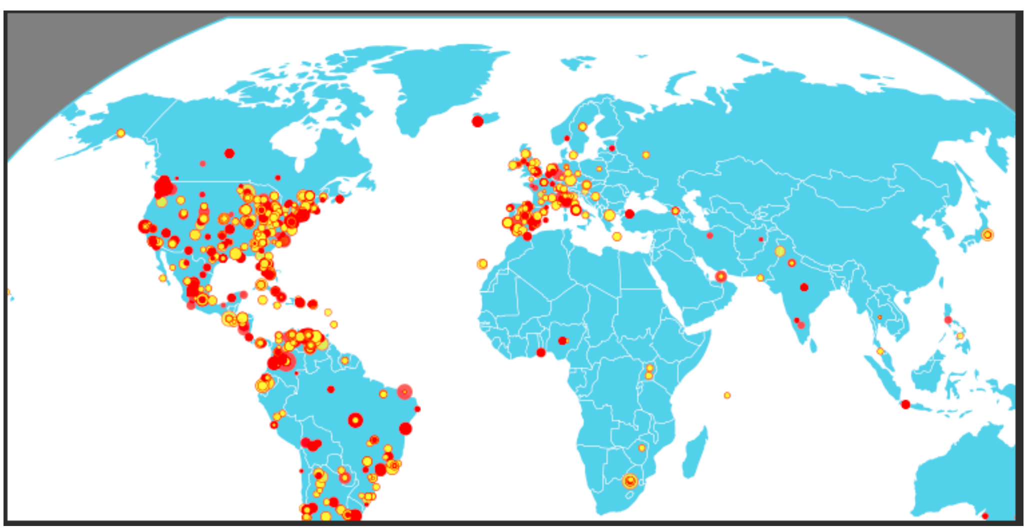
	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	[M] location_c	[M] location_	[M] location_	[M] location_	[M] media_d	[M] media_t	[M] media_u	[M] posted_t	[M] real_nan	[M] source:	[M] statuses_	[M] tweet_u	[M] user_bio	[M] user_loc	[M] user_me	[M] user_me	[M] user_twi	[M] username:	
2	r.com/chefAronna/statuses/545406969994289152							##########	#ColombiaD	Twitter for A	16131	http://www.	Dedicated Ch	Colombia	NTN24	NTN24	http://www.	chefAronna	
3	r.com/Bluegirl_3/statuses/545406967003762688							##########	Renee Beauli	Twitter Web	2869	http://www.	Politically ob	WI	NowWithAle	NowWithAle	http://www.	Bluegirl_3	
4	r.com/dbergman39/statuses/545406967217684480							##########	Daniel Bergm	TweetCaster	1255	http://blogs.	Commentato	Nebraska	Wall Street J	WSJ	http://www.	dbergman39	
5	r.com/famousred/statuses/545406968069115904							##########	LoÃ¯c Gauthi	Twitter for A	59		J'ai une devis	La Baie	Martin Petit	Lemicrodefe	http://www.	famousred	
6	r.com/GEsneault/statuses/545406970048806914							##########	Gary J. Esnea	Twitter Web	482			Nashua, NH	Top Conserv	TeaPartyCat	http://www.	GEsneault	
7	r.com/gissellemr/statuses/545406969885229056							##########	Gisselle Mor	TweetDeck	3353	http://www.	Periodista de	Cuba	Escambray; F	escambraycu	http://www.	gissellemr	
8	r.com/danicabrera21/statuses/545406969767821312							##########	Dani	Twitter for A	1925	http://www.	Viajar y ser fe	Lima - PerÃº	Noticias SIN	SIN24Horas	http://www.	danicabrera21	
9	r.com/angel_cortez78/statuses/545406969017024512				pic.twitter.cc	photo	http://pbs.tw	##########	AmericaNeed	Twitter Web	1177	http://twitte	Serial entrep	North Carolir	The Daily Bea	thedailybeas	http://www.	angel_cortez78	
10	r.com/cam_vallejo/statuses/545406968899174401							##########	Camilo Fdo V	Twitter for A	3275		Colombiano,	Melbourne-A	Leonardo Pa	Leonardo_Pa	http://www.	cam_vallejo	
11	r.com/67purple/statuses/545406969079955458				pic.twitter.cc	photo	http://pbs.tw	##########	caryl mast	Twitter Web	37252	http://twitte	colorado girl,	colorado girl	Kenneth Isaa	uselephants;	http://www.	67purple	
12	r.com/luillopr/statuses/545406969692303360							##########	Luis Moreira	PlumeÂ forÂ	26234		The views an	Trujillo Alto,	AndrÃ©s Col	andrescolon	http://www.	luillopr	
13	r.com/anmirauli/statuses/545406971965628416							##########	Miriam R.	Twitter for A	16759		Educadora-A	Caracas	ManoloReve	ManoloReve	http://www.	anmirauli	
14	r.com/Parcival2/statuses/545406973445812225							##########	Tim Scanlon	Twitter for W	10508	http://media	Re-elect Mill	Hyattsville, Maryland, USA			http://www.	Parcival2	
15	r.com/tmcgev/statuses/545406971747524608							##########	Tom McGeve	Twitter Web	15232		Co-founder a	New York	Tom McGeve	tmcgev	http://www.	tmcgev	
16	r.com/CV_People/statuses/545406972372070401				pic.twitter.cc	photo	http://pbs.tw	##########	Conservative	Twitter for iP	26911	http://twitte	Proudly follo	USA	Ban Collectiv	mrgeology	http://www.	CV_People	
17	r.com/lailaannamaria/statuses/545406971546206208							##########	Laila Anna M	iOS	24948	http://news3	Religious,53,	Norway			http://www.	lailaannamaria	
18	r.com/AtheistRaven/statuses/545406973861052416							##########	AlaskanRave	Twitter for A	21199		Aspiring supe	Anchorage, A	Top Conserv	TeaPartyCat	http://www.	AtheistRaven	
19	r.com/Analog_Zero/statuses/545406974213779456							##########	Analog_Zero	Twitter Web	137347			Barranquilla,	Hassan Nass	HassNassar	http://www.	Analog_Zero	
20	r.com/cbarrrons/statuses/545406971759693825							##########	CÃ©sar Barr	twitterfeed	9547	http://www.	Comienza tu	DF			http://www.	cbarrrons	
21	r.com/frankalbarran/statuses/545406975815987200				pic.twitter.cc	photo	http://pbs.tw	##########	Franklin	Twitter for A	1654	http://actual	Chavista, Rev	Venezuela	RT en EspaÃ	ActualidadRT	http://www.	frankalbarran	
22	r.com/MTmarilyn2/statuses/545406976222437377							##########	Marilyn Olive	Twitter for W	144896		Retired smal	Montana	Lnonblonde	Lnonblonde	http://www.	MTmarilyn2	
23	r.com/sospr_/statuses/545406980068622336							##########	Sonia Presa	twitterfeed	15045	http://www.	Dios, Esposa,	MÃ©xico			http://www.	sospr_	
24	r.com/eyranad97/statuses/545406975786242048							##########	Ngee	Twitter for A	1165		~ Assalamual	seoul,korea	MUJAHID	SAFWANMU	http://www.	eyranad97	
25	r.com/Delgado_106/statuses/545406976323117056							##########	Ramon Andr	Twitter for A	1385		Docente de l	Yaracuy Ven	Mario Silva G	LaHojillaenT\	http://www.	Delgado_106	
26	r.com/proyecto_eps_bl/statuses/545406977560809472							##########	Jorge Moren	Twitter Web	4364		Proyecto inn	Maracay	MABEL BOR(	2405mabel	http://www.	proyecto_eps_bl	
27	r.com/DanaySuarez/statuses/545406978089287680							##########	Danay Suare	Twitter for iP	1825		Managemen	La Habana, Cuba			http://www.	DanaySuarez	
28	r.com/betomono5/statuses/545406977271025664							##########	Beto Morenc	twitterfeed	10646	http://www.	Mi novia y Yo	DF			http://www.	betomono5	
29	r.com/ddigless/statuses/545406976507662336							##########	DamiÃ¡n Igle	twitterfeed	10644	http://www.	DESCUBRIM(	Ecatepec, MÃ©xico			http://www.	ddigless	
30	r.com/sospr_/statuses/545406978646765568							##########	Sonia Presa	twitterfeed	15044	http://www.	Dios, Esposa,	MÃ©xico			http://www.	sospr_	
31	r.com/Riverocontreras/statuses/545406979251138560							##########	Elis Clement	Twitter Web	15241	http://www.	Socialista, re	Venezuela	Fidel Castro	fidelcastro	http://www.	Riverocontreras	
32	r.com/BHandgraaf_RMT/statuses/54540698011516928				pic.twitter.cc	photo	http://pbs.tw	##########	Brie Handgra	TweetCaster	1455	http://mone	I'm a reporte	Rocky Moun	CNNMoney	CNNMoney	http://www.	BHandgraaf_RMT	
33	r.com/Xijesman30/statuses/545406978802327552				pic.twitter.cc	photo	http://pbs.tw	##########	Finita	Twitter for A	18057	http://twitte	Venezolana.	VENEZUELA	El monitor 1	ElMonitor18	http://www.	Xijesman30	
34	r.com/alilovestexas/statuses/545406981121769472				pic.twitter.cc	photo	http://pbs.tw	##########	Conservative	Twitter Web	14203	http://www.	Passionate @	Texas USA	Katherine Mi	katherinemil	http://www.	alilovestexas	
35	r.com/mm90marcos/statuses/545406981381435392							##########	Marcos Meza	twitterfeed	11064	http://www.	Los sueÃ±os	MÃ©xico			http://www.	mm90marcos	
36	r.com/1776BetsyRoss/statuses/545406980433936384							##########	ProudAmeric	Twitter Web	13535		Independent	USA	Bill Damato;	Billd111; me	http://www.	1776BetsyRoss	
37	r.com/jalynhenton/statuses/545406980681375744							##########	Jalyn Henton	Twitter for A	11293		Doing my be	Del Ray, Alexandria, VA			http://www.	jalynhenton	
38	r.com/Longueira_2018/statuses/545406986075275266							##########	Pablo Longu	Twitter for iP	1745	http://www.	TWITTER de	Chile	Yoani SÃ¡nch	yoanisanche	http://www.	Longueira_2018	
39	r.com/mfarooqafzal7/statuses/545406984552726528							##########	M. Farooq Af	Twitter for B	4072	http://www.	Businessman	Karachi Pakis	Reuters Top	Reuters	http://www.	mfarooqafzal7	
40	r.com/catharine_cezil/statuses/545406986267791361							##########	Catharine C.	dlvr.it	27628			Morelia, MichoacÃ¡n			http://www.	catharine_cezil	
41	r.com/Pamadrigal_/statuses/545406986737549312							##########	Pam Madriga	dlvr.it	37082		Amo las flore	Morelia, MichoacÃ¡n			http://www.	Pamadrigal_	
42	r.com/digitalcarmona/statuses/545406990168903680							##########	Jose Luis Car	Twitter for A	2920		#calmtechno	Granada. Spa	The Teacher	mai_maga	http://www.	digitalcarmona	
43	r.com/INGJOSEMANUEL/statuses/545406986335289344							##########	JOSE RODRIG	Twitter for B	17833		Ingeniero Ind	VENEZUELA			http://www.	INGJOSEMANUEL	
44	r.com/drolet_noma/statuses/545406986695622656							##########	Noma Drolet	dlvr.it	48656		Electrogirl	MÃ©xico, DF			http://www.	drolet_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 exploratoration/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?

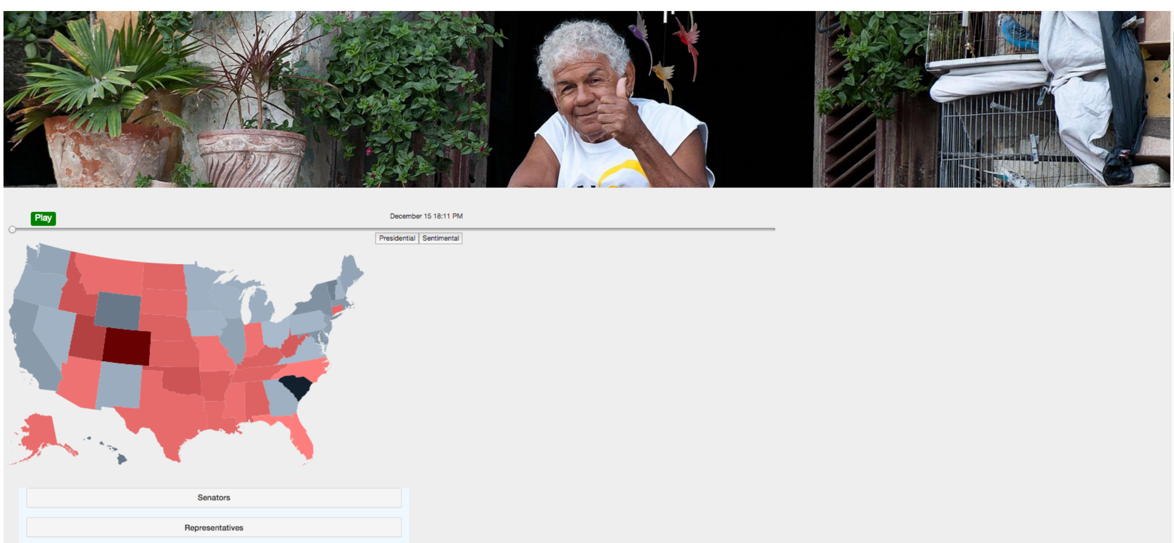
- For most of the project, we were unable to do proper data visualization during exploration because data set was orders of magnitude larger than anything we were prepared for, because most lacked reliable/easily usable geoData/.
- Main visualization for most of project was a slice of tweets mapped against 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.

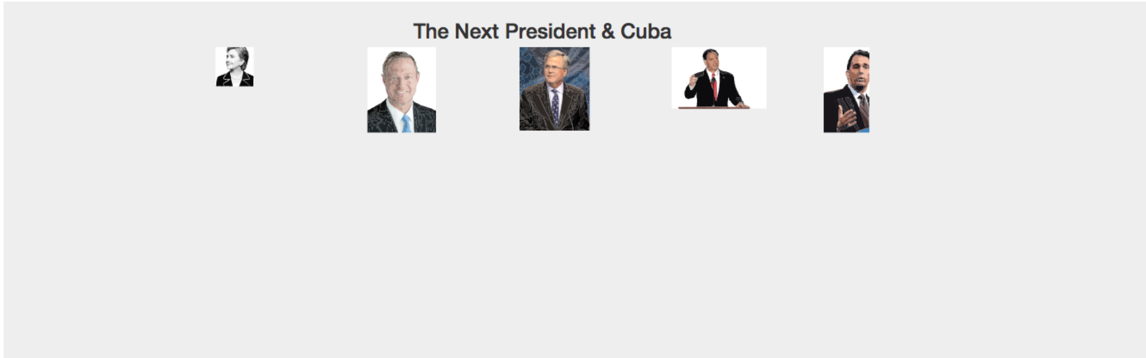
Design Evolution:

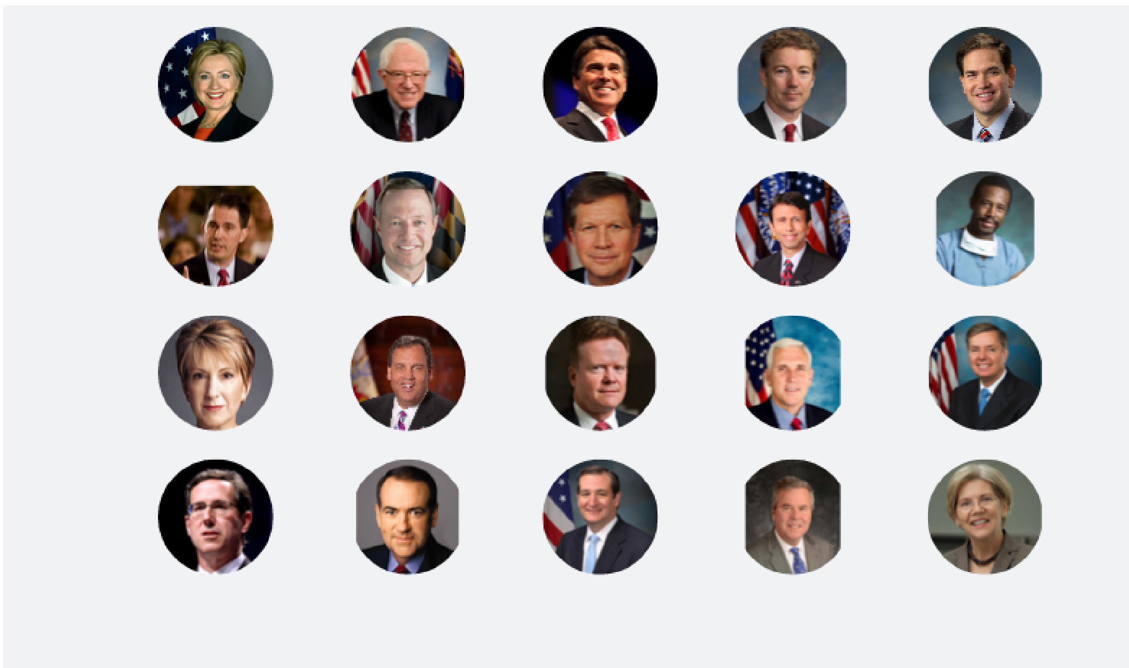
What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course.

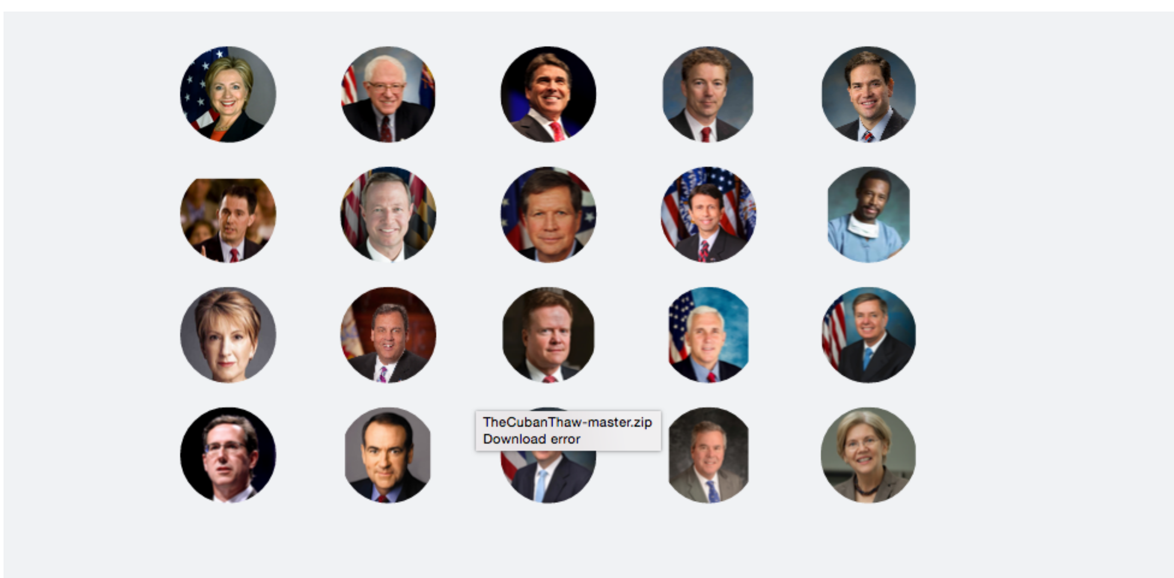
[Include pictures and use theory from book]

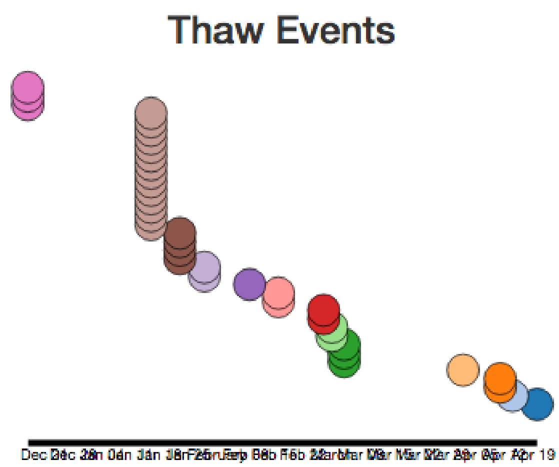
- struggles with color and maps
- deciding between greater honesty and greater utility/ greater aesthetics.
- deciding whether to aggregate and how to aggregate. Semantic analysis posed difficult aggregation questions. “If your visualization is correct at the aggregate but internally it is very volatile and wrong, is it a good visualization?” / “Meaning of average”
- “One page” website design vs 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.

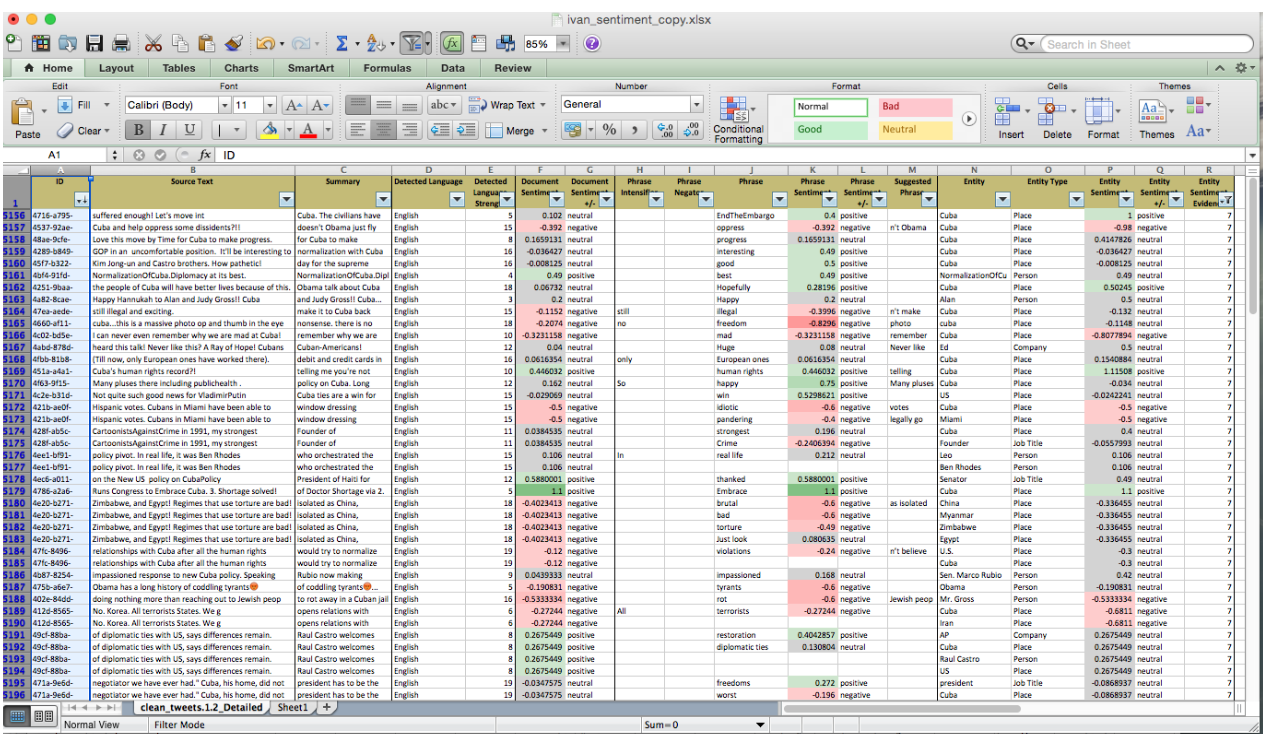



The Next President & Cuba




TheCubanThaw-master.zip
Download error


Thaw Events


ivan_sentiment_copy.xlsx
clean_tweets.1.2_Detailed
Sheet1

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 very that man If you truly understand this dataset, you can know the minds of hundreds of thousands of people in the US and around the world.

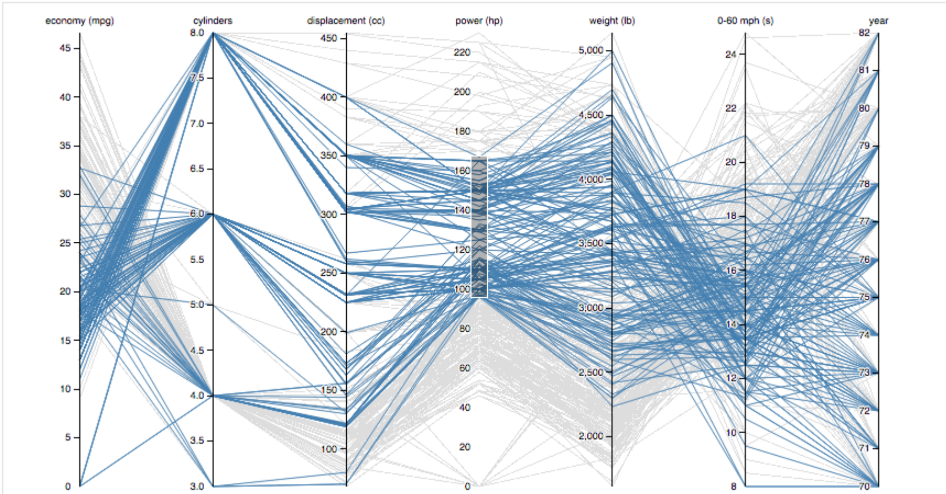
Improvements:

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 countries color, thus country takes on the color of most popular hashtag

Our main visualization for most of the project. Map of the US with a small subset of 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

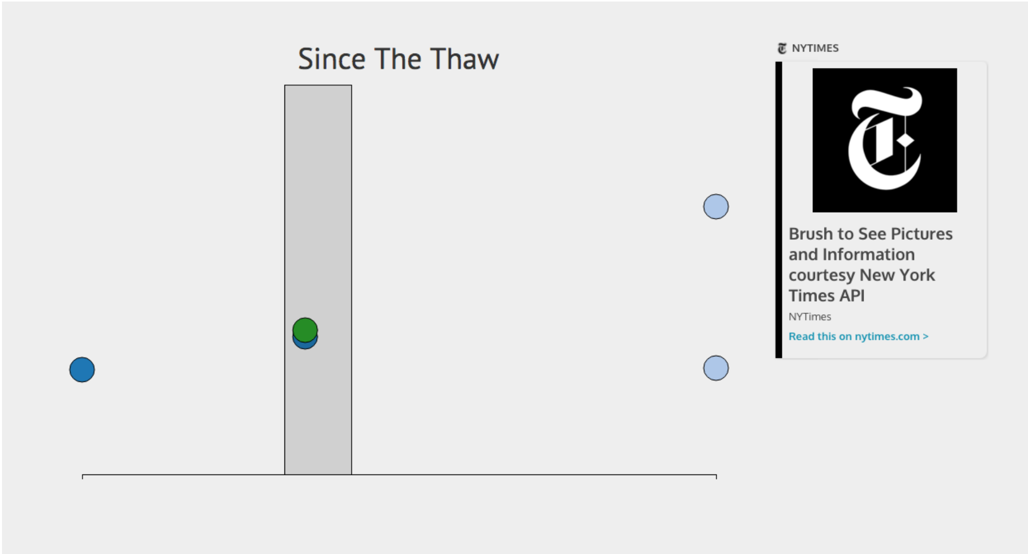


This is a version of Mike Bostock's parallel coordinates example, modified to include reorderable axes.

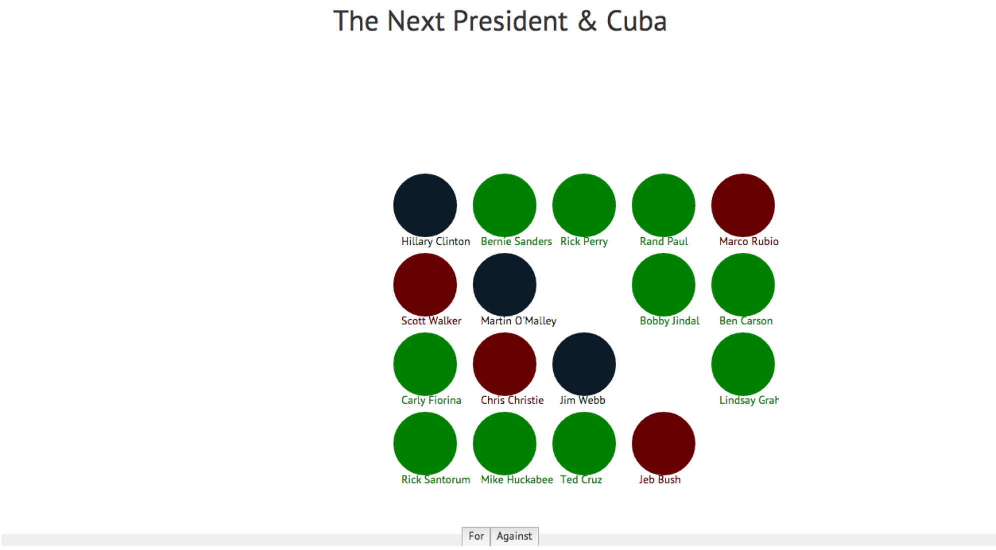
Open in a new window.

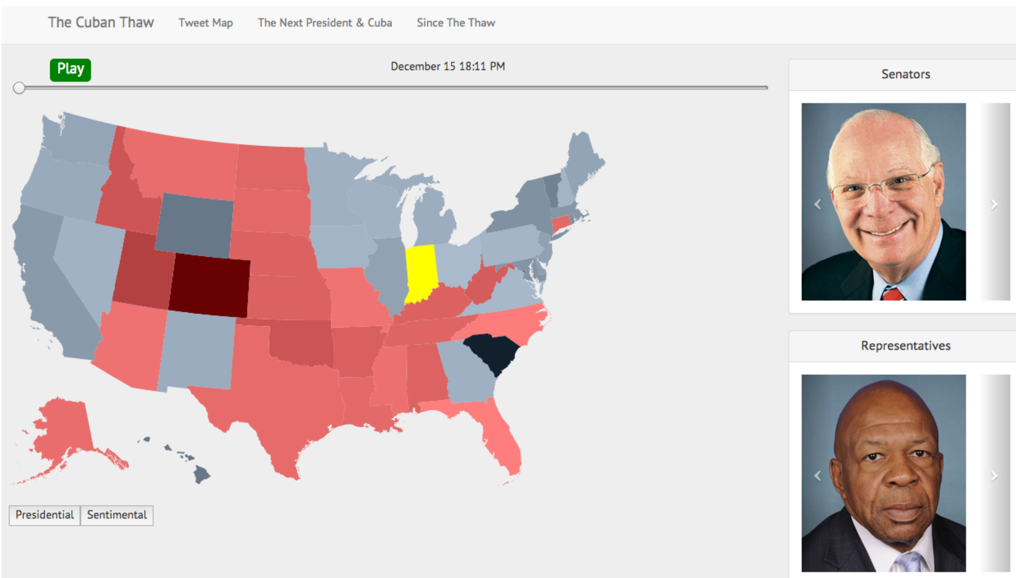
Discussed with TA about turning visualization of presidential candidates into a Parallel Coord vis, but while technically interesting and visually complex, we feel that this is too advanced for our expected users and their use cases. It is neither the most intuitive solution nor the most promising of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

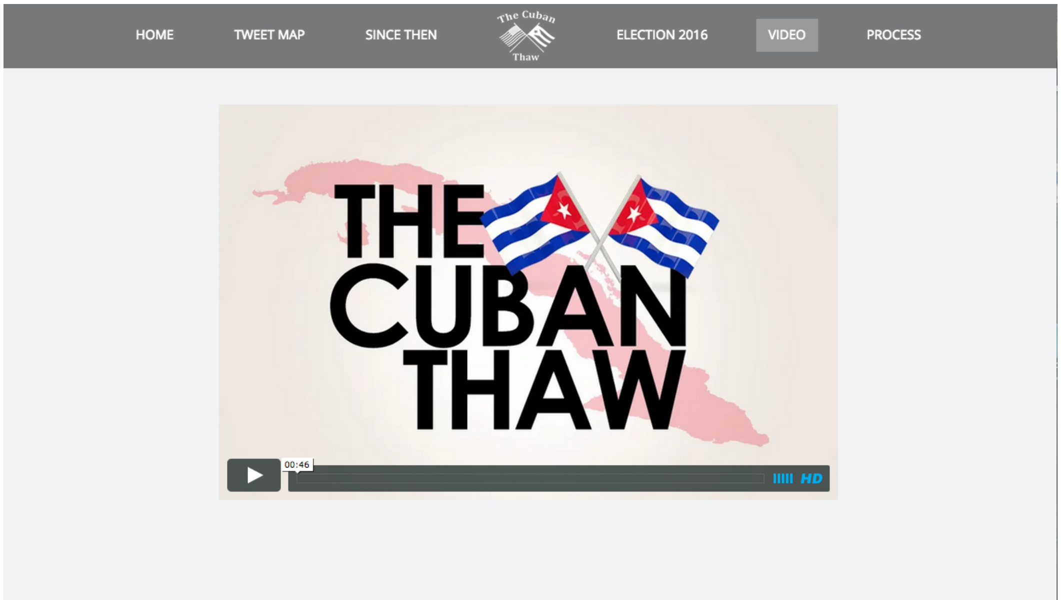
.

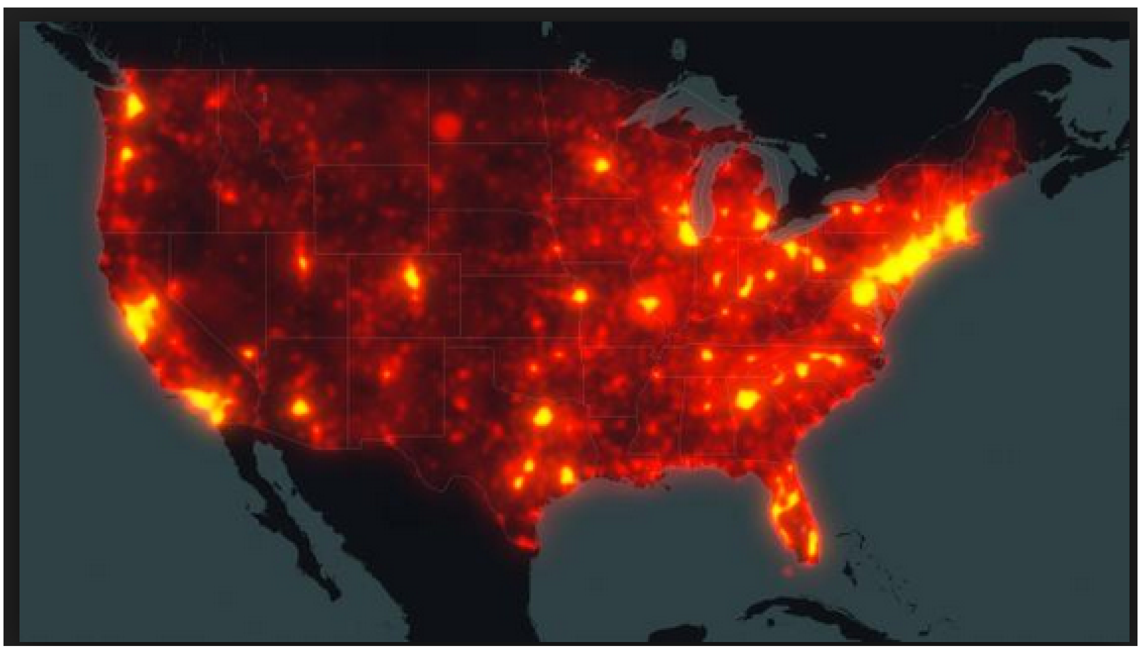


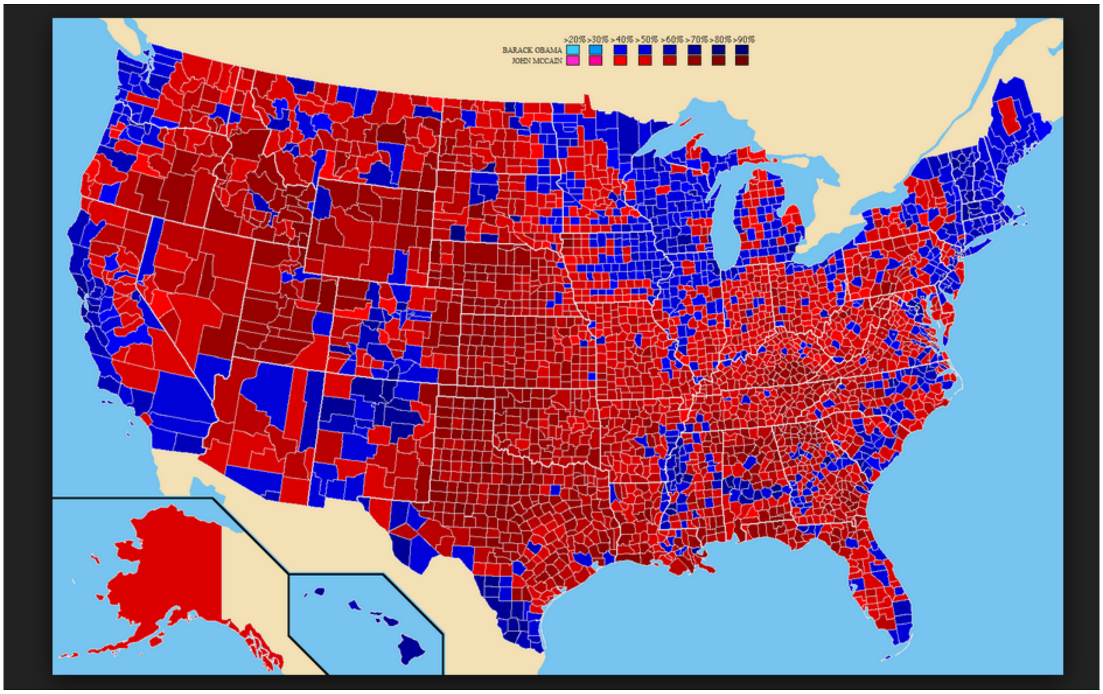
Purpose of this visualization was to communicate to the user the immediate consequences of the Thaw through events happening in the news. This brush queries the NYTimes for articles related to Cuba between the brush's extent.


The Next President & Cuba
Hillary Clinton
Bernie Sanders
Rick Perry
Rand Paul
Marco Rubio
Scott Walker
Martin O'Malley
Bobby Jindal
Ben Carson
Carly Fiorina
Chris Christie
Jim Webb
Lindsay Grah
Rick Santorum
Mike Huckabee
Ted Cruz
Jeb Bush
For
Against


The Cuban Thaw
Tweet Map
The Next President & Cuba
Since The Thaw
Play
December 15 18:11 PM
Senators
Representatives
Presidential
Sentimental


HOME
TWEET MAP
SINCE THEN
The Cuban Thaw
ELECTION 2016
VIDEO
PROCESS
THE CUBAN THAW
00:46
HD





z - W

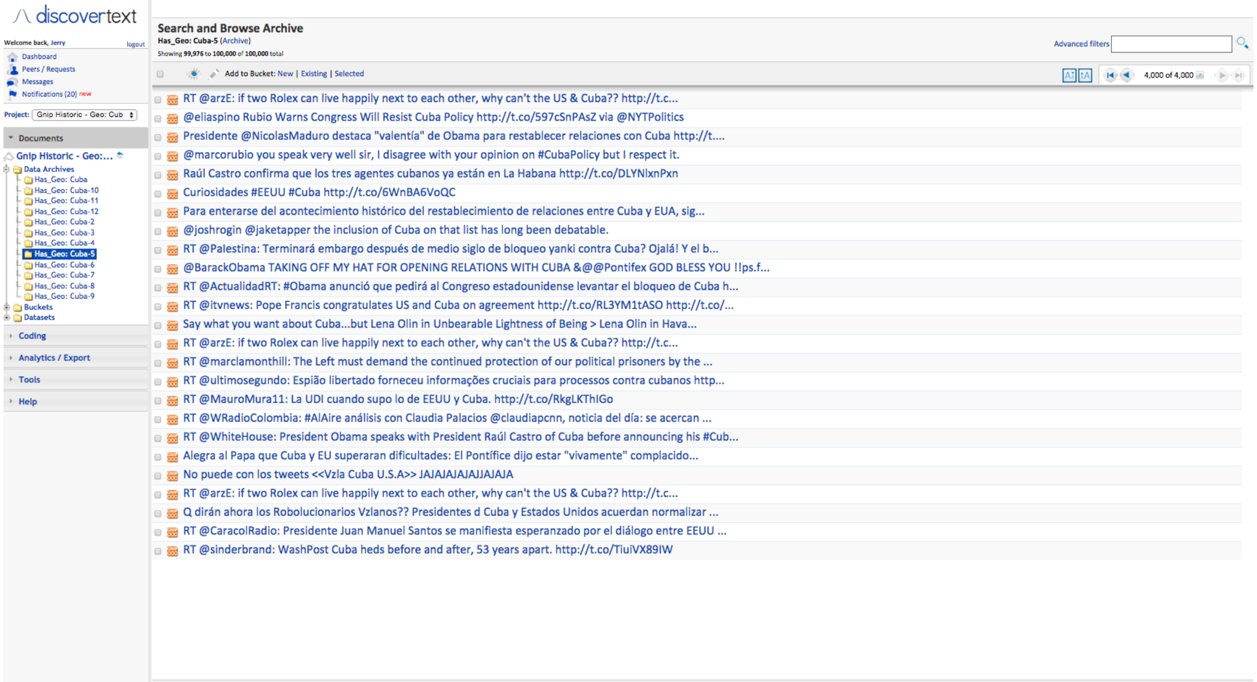
TopMeta Discovery ✖

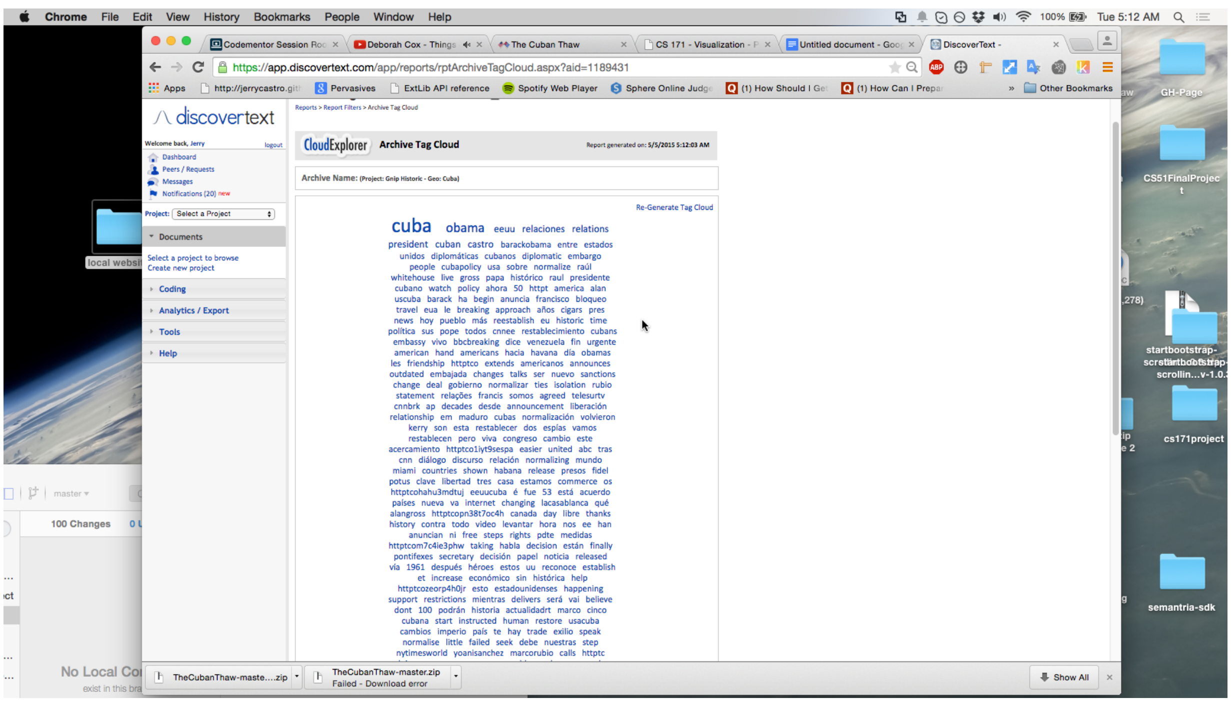
Top values for: **country_code:**

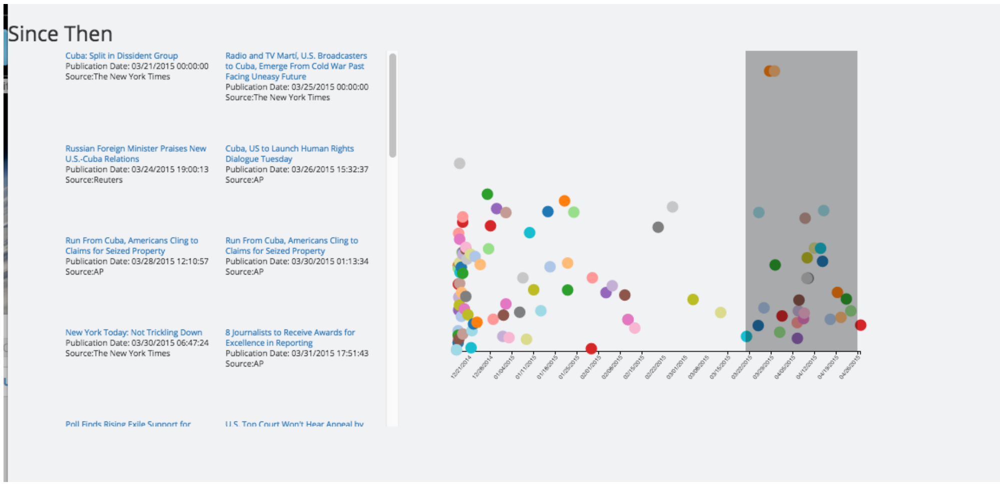
1 of 6

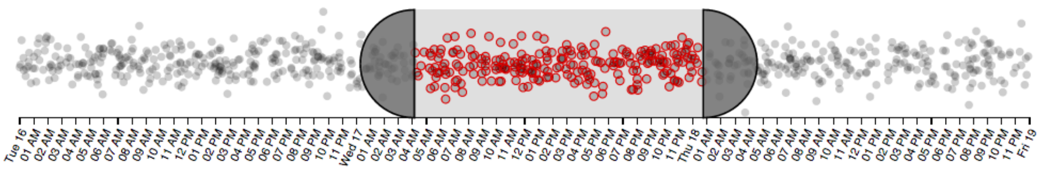
Meta Value	Total ▼	Filter
United States	746	
Venezuela	199	
Brasil	195	
España	128	
Chile	114	
México	93	
Colombia	91	
Argentina	85	
United Kingdom	51	
Canada	45	

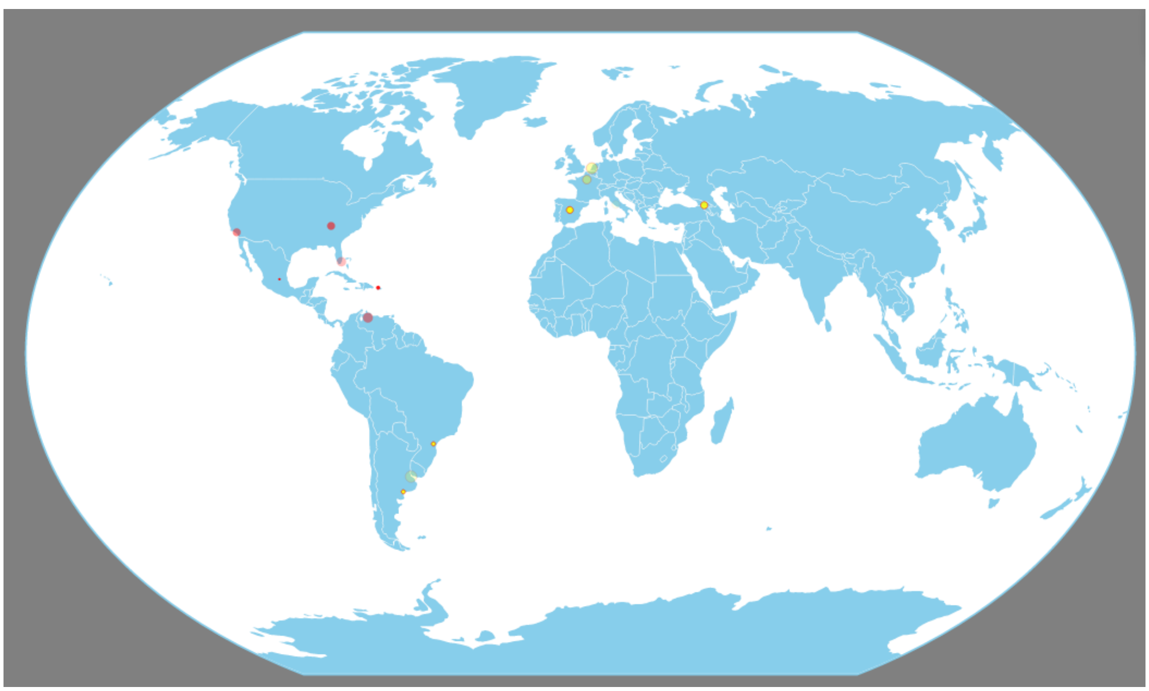
Showing **1** to **10** of **57** total

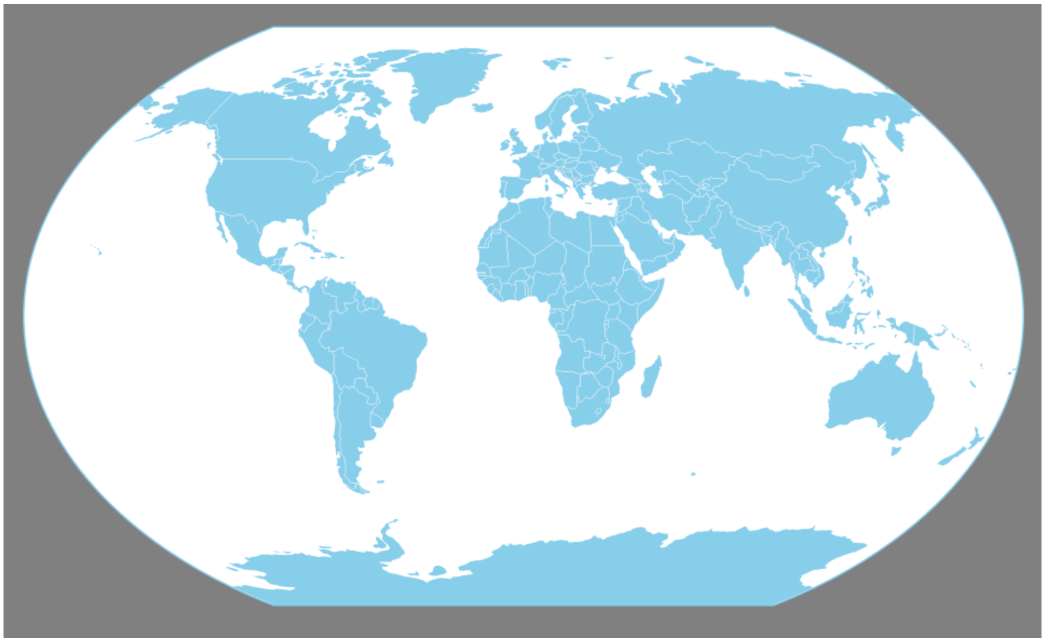

discovertext
Welcome back, Jerry
logout
Dashboard
Peers / Requests
Messages
Notifications (20) new
Project: Gnip Historic - Geo: Cub
Documents
Gnip Historic - Geo:...
Data Archives
Has_Geo: Cuba
Has_Geo: Cuba-10
Has_Geo: Cuba-11
Has_Geo: Cuba-12
Has_Geo: Cuba-2
Has_Geo: Cuba-3
Has_Geo: Cuba-4
Has_Geo: Cuba-5
Has_Geo: Cuba-6
Has_Geo: Cuba-7
Has_Geo: Cuba-8
Has_Geo: Cuba-9
Buckets
Datasets
Coding
Analytics / Export
Tools
Help
Search and Browse Archive
Has_Geo: Cuba-5 (Archive)
Showing 99,976 to 100,000 of 100,000 total
Advanced filters
Add to Bucket: New | Existing | Selected
4,000 of 4,000
RT @arzE: if two Rolex can live happily next to each other, why can't the US & Cuba?? http://t.c...
@eliaspino Rubio Warns Congress Will Resist Cuba Policy http://t.co/597cSnPAsZ via @NYTPolitics
Presidente @NicolasMaduro destaca "valentía" de Obama para restablecer relaciones con Cuba http://t....
@marcorubio you speak very well sir, I disagree with your opinion on #CubaPolicy but I respect it.
Raúl Castro confirma que los tres agentes cubanos ya están en La Habana http://t.co/DLYNlxnPxn
Curiosidades #EEUU #Cuba http://t.co/6WnBA6VoQC
Para enterarse del acontecimiento histórico del restablecimiento de relaciones entre Cuba y EUA, sig...
@joshrogin @jaketapper the inclusion of Cuba on that list has long been debatable.
RT @Palestina: Terminará embargo después de medio siglo de bloqueo yanki contra Cuba? Ojalá! Y el b...
@BarackObama TAKING OFF MY HAT FOR OPENING RELATIONS WITH CUBA &@@Pontifex GOD BLESS YOU !!ps.f...
RT @ActualidadRT: #Obama anunció que pedirá al Congreso estadounidense levantar el bloqueo de Cuba h...
RT @itvnews: Pope Francis congratulates US and Cuba on agreement http://t.co/RL3YM1tASO http://t.co/...
Say what you want about Cuba...but Lena Olin in Unbearable Lightness of Being > Lena Olin in Hava...
RT @arzE: if two Rolex can live happily next to each other, why can't the US & Cuba?? http://t.c...
RT @marclamonthill: The Left must demand the continued protection of our political prisoners by the ...
RT @ultimosegundo: Espião libertado forneceu informações cruciais para processos contra cubanos http...
RT @MauroMura11: La UDI cuando supo lo de EEUU y Cuba. http://t.co/RkgLKThIGo
RT @WRadioColombia: #AlAire análisis con Claudia Palacios @claudiapcnn, noticia del día: se acercan ...
RT @WhiteHouse: President Obama speaks with President Raúl Castro of Cuba before announcing his #Cub...
Alegra al Papa que Cuba y EU superaran dificultades: El Pontífice dijo estar "vivamente" complacido...
No puede con los tweets <<Vzla Cuba U.S.A>> JAJAJAJAJAJJAJAJA
RT @arzE: if two Rolex can live happily next to each other, why can't the US & Cuba?? http://t.c...
Q dirán ahora los Robolucionarios Vzlanos?? Presidentes d Cuba y Estados Unidos acuerdan normalizar ...
RT @CaracolRadio: Presidente Juan Manuel Santos se manifiesta esperanzado por el diálogo entre EEUU ...
RT @sinderbrand: WashPost Cuba heds before and after, 53 years apart. http://t.co/TiuIVX89IW

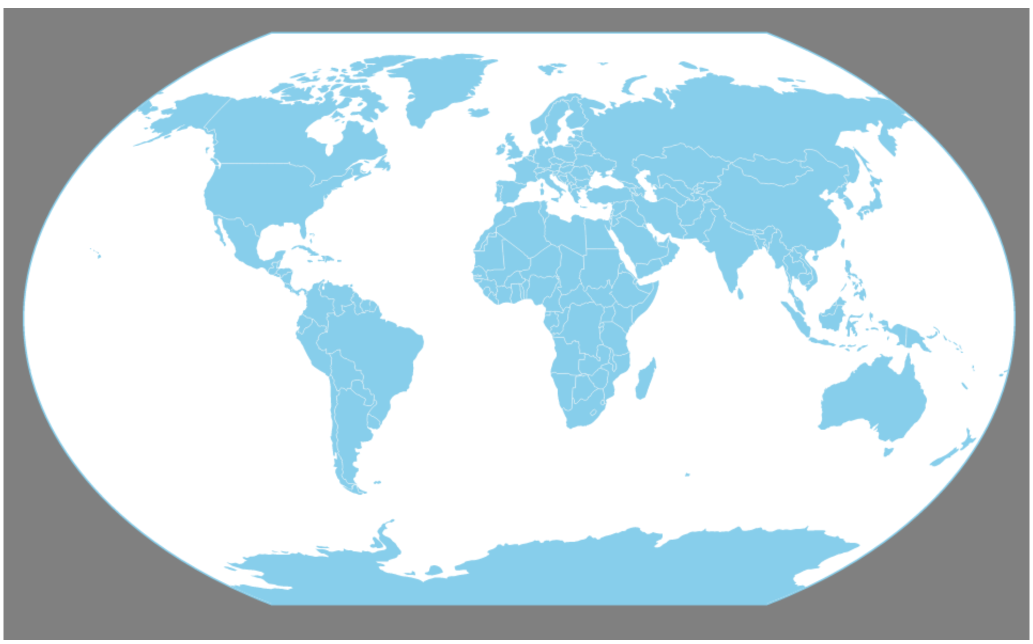

Chrome File Edit View History Bookmarks People Window Help
100% Tue 5:12 AM
Codementor Session Roo
Deborah Cox - Things
The Cuban Thaw
CS 171 - Visualization
Untitled document - Goo
DiscoverText -
https://app.discovertext.com/app/reports/rptArchiveTagCloud.aspx?aid=1189431
Apps
http://jerrycastro.git
Pervasives
ExtLib API reference
Spotify Web Player
Sphere Online Judge
(1) How Should I Ge
(1) How Can I Prepa
Other Bookmarks
discovertext
Welcome back, Jerry
logout
Dashboard
Peers / Requests
Messages
Notifications (20) new
Project: Select a Project
Documents
Select a project to browse
Create new project
Coding
Analytics / Export
Tools
Help
Reports > Report Filters > Archive Tag Cloud
CloudExplorer
Archive Tag Cloud
Report generated on: 5/5/2015 5:12:03 AM
Archive Name: (Project: Gnip Historic - Geo: Cuba)
Re-Generate Tag Cloud
cuba obama eeuu relaciones relations
president cuban castro barackobama entre estados
unidos diplomáticas cubanos diplomatic embargo
people cubapolicy usa sobre normalize raúl
whitehouse live gross papa histórico raul presidente
cubano watch policy ahora 50 httpt america alan
uscuba barack ha begin anuncia francisco bloqueo
travel eua le breaking approach años cigars pres
news hoy pueblo más reestablish eu historic time
politica sus pope todos cnnee restablecimiento cubans
embassy vivo bbcbreaking dice venezuela fin urgente
american hand americans hacia havana día obamas
GH-Page
CS51FinalProject
startbootstrap-
cs171project
semantria-sdk
master
100 Changes
No Local Co
TheCubanThaw-maste....zip
TheCubanThaw-master.zip
Failed - Download error
Show All

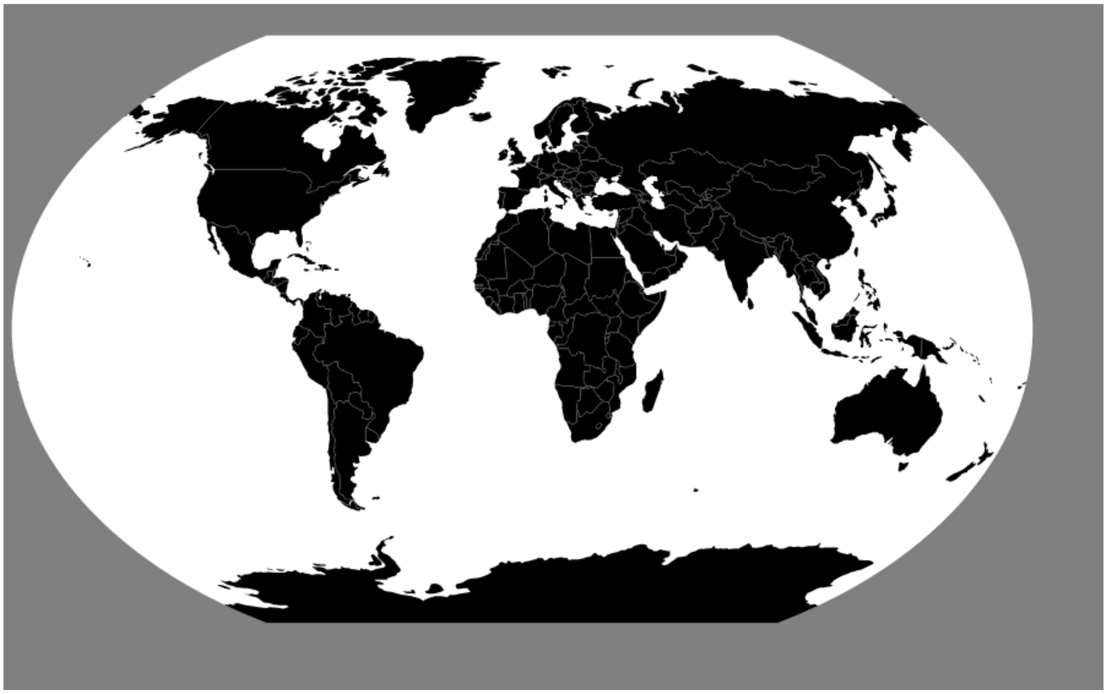

Since Then
Cuba: Split in Dissident Group
Publication Date: 03/21/2015 00:00:00
Source:The New York Times
Radio and TV Martí, U.S. Broadcasters to Cuba, Emerge From Cold War Past Facing Uneasy Future
Publication Date: 03/25/2015 00:00:00
Source:The New York Times
Russian Foreign Minister Praises New 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 15:32:37
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 Rising Exile Support for
U.S. Top Court Won't Hear Appeal by

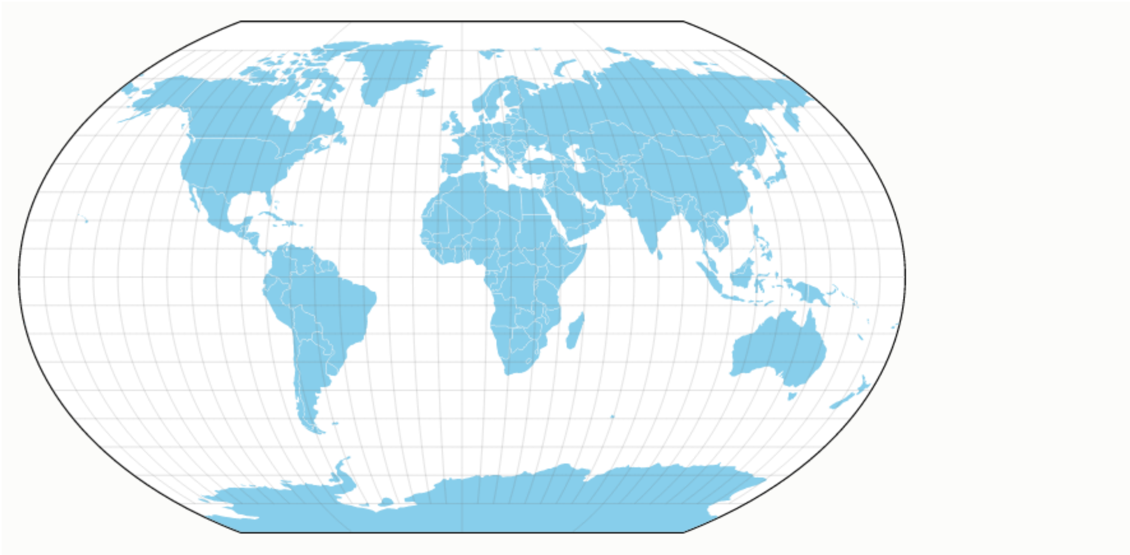

Tue 16
Wed 17
Thu 18
Fri 19

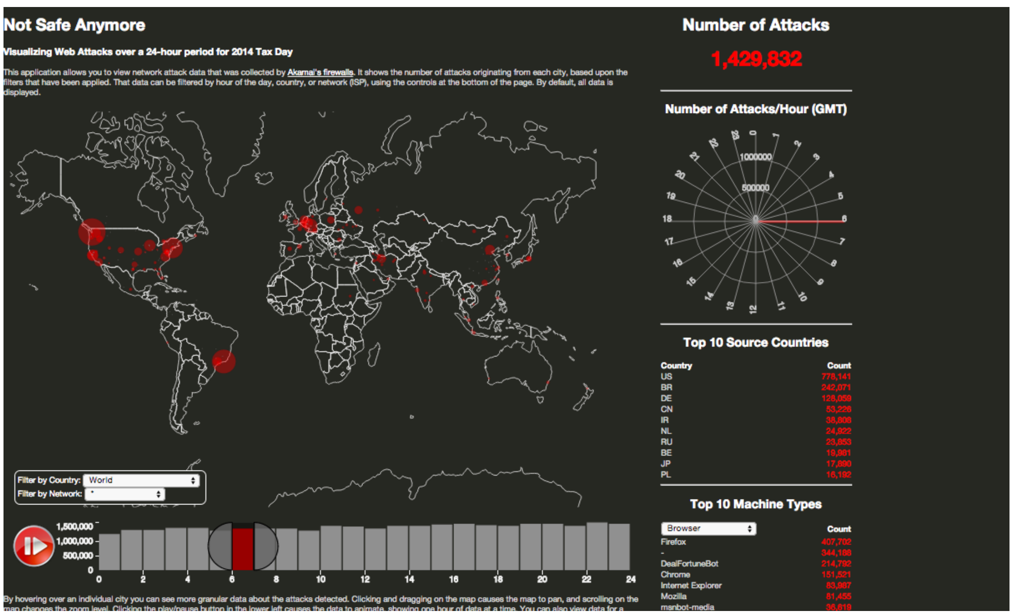











Not Safe Anymore
Visualizing Web Attacks over a 24-hour period for 2014 Tax Day
This application allows you to view network attack data that was collected by Akamai's firewalls. It shows the number of attacks originating from each city, based upon the filters that have been applied. That data can be filtered by hour of the day, country, or network (ISP), using the controls at the bottom of the page. By default, all data is displayed.
Filter by Country: World
Filter by Network: *
1,500,000
1,000,000
500,000
0
0 2 4 6 8 10 12 14 16 18 20 22 24
By hovering over an individual city you can see more granular data about the attacks detected. Clicking and dragging on the map causes the map to pan, and scrolling on the
Number of Attacks
1,429,832
Number of Attacks/Hour (GMT)
1000000
500000
Top 10 Source Countries
Country Count
US 778,141
BR 242,071
DE 128,059
CN 53,226
IR 38,808
NL 24,922
RU 23,853
BE 19,981
JP 17,890
PL 16,192
Top 10 Machine Types
Browser Count
Firefox 407,702
- 344,188
DealFortuneBot 214,792
Chrome 151,521
Internet Explorer 83,987
Mozilla 81,455
msnbot-media 36,619