The Cuban Thaw Visualization Project:

A Proposal

Jerry Castro & Ivan Lima

CS 171, Harvard University

April 3, 2015

Table of Contents

Background and Motivation.

“Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision.”

When Ivan and I first started looking for a dataset to use for our project, we struggled to find a dataset that really jumped out at us. Many of them just weren’t very interesting. So I stayed up late one night, aimlessly browsing the web until I stumbled upon something both fascinating and heartbreaking. Some time ago, Wikileaks released 500,000 pager intercepts from the day of 9/11. The intercepts showed you the horrors of that day in 5 minute intervals through the communications of ordinary people contacting friends and family.

While this dataset was too dark and painful of a topic, Ivan and I were very intrigued by the idea of visualizing high volume social communication during a single major event. What might a data scientist see if she analyzed and visualized hundreds of thousands of tweets during a major historical news event?

We brainstormed all the events that would have strong Twitter data but none of them seemed to work. I was frustrated and ready to give up until I realized the perfect topic was actually very recent and incredibly historic: The restoration of diplomatic relations between the United States and Cuba on December 17, 2014.

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

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

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

Project Objectives.

"Provide the primary questions you are trying to answer with your visualization. What would you like to learn and accomplish? List the benefits."

The Cuban Thaw Visualization Project has (2) primary questions and by pursuing these questions we will also answer several other questions along the way.

(1) **"How do Millennials feel about the Thaw?"** : Answering this question is our first objective. To do so, we will determine how Twitter users felt about the Cuban Thaw. Given that the average Twitter user is just under 30 years old, our analysis will give us strong insight on how the millennial generation feels about the US embargo of Cuba and consequently the direction the nation over time as millennials begin to increasingly express themselves politically.

(2) **"Can Data Viz decide an election?"** : It is entirely possible that the next Congress and President might try to reverse it or they may instead finally lift all restrictions. This means that the future of US-Cuba relations is directly in the hands of voters in the 2016 election. While this is an amazing opportunity for democracy, it only works if politicians have accurate data on how people really feel about the issue. But since there is no good data on how Millenials feel about the Cuban Thaw, politicians may ignore them or even try to win their votes with completely the wrong message. A battleground state like Florida with 25 electoral college votes is home two Presidential candidates and millions of millenial Cubans and is already well known for deciding elections (Bush v Gore). The issue of the Cuban Thaw might decide the next President of the United States and we want our data visualization make to a difference. Which is entirely possible. In fact, the 2000 Presidential Election was decided on just 537 votes cast in ….. Florida.

The following list includes several other questions that our project may answer. Below each question is the benefit gained by answering it.

Do the opinions Millenials match the politics of the state where they are from? That is, do Millenials from conservative states feel conservative about the issue? Do Millenials from liberal states feel liberal about the issue?

> If the answer is no, then we can be more hopeful about the future because it suggests that the next generation of voters may change the direction of the country and its politics. If the answer is no, then we should get used to the way things are because they aren't likely to change.

"How do Young Cuban Americans really feel about the policy?"

> The benefit to this question is obvious. The answer may decide if the next President is called Bush or Rubio or Clinton or someone else.

"Which Politicians disagree with Millenials about the Cuban Thaw?"

> The benefit to answering this question is that we can begin to piece together a profile for a type of politician that isn't likely to survive in the future once millenials start voting in larger and larger numbers.

"How do the 2016 Presidential candidates feel about the Thaw?"

> Again, this answer might just decide the American Presidency.

Data.

“From where and how are you collecting your data? If appropriate, provide a link to your data sources.”

The datasets described below cover all of the data requirements that we can predict at this time. The first (4) datasets include links and have already been acquired. The remaining (2) datasets do not include links because they must be manually compiled and we have not yet gathered the necessary data points but we know they do exist and we are extremely confident of easily acquiring them through simple google searches.

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

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

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

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

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

(6) “Political Positions of Politicians” : A dataset describing how every member of Congress and every 2016 Presidential candidate feels about the Cuban Thaw, either for or against. These “feelings” will based on public statements. A possible “none” or “no comment” category may also be added.

Visualization.

"How will you display your data? Provide some general ideas that you have for the visualization design. Include sketches of your design."

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

- **"Tweet Map"** : A map of the US color coded to correspond to a geo-located sentiment analysis of tweets about Cuba from December 17 (day of), and December 18 (day after). It has (2) subviews. The first is a comparison of the tweet map to Republican and Democrat voting patterns from the US Presidential Election of 2012. The second is a comparison of the tweet map to the corresponding demographics in each area based on census information.
  - A major visual decision yet to be made is how to visualize difference between two US maps. Solutions include a slider that shows to the left, the tweet map and to the right a map of the 2012 voting results. Another is have the user brush over the area to see the differences. Of course, we could just display two maps next to each other but this is less effective.

- **"Opinions of Congress"** : A visualization of how each member of Congress (House & Senate) feels about the thaw, either for or against, based on public statements. If no public statement can be found, then a third category of none (or similar), will also be included. This visualization can be filtered by: Political Party, Up for Election in 2016, Not up for Election in 2016, Region (West, South, MidWest, Northeast), Age, Gender.
    - Again, we have not decided for sure what to do on this one. Many different visual ideas of this are possible. We could make simple SVG shapes to correspond to seats in Congress that change colors to indicate certain opinions. Of course, we can also just use a map which works well for Senators since there are only two per state, but some states have a double digit number of representatives in Congress.

- **"Opinions of 2016 Presidential Candidates"** **:** A visualization of how each of the likely 2016 Presidential candidates feels about the Cuban Thaw, either for or against, based on public statements. If no public statement can be found, then a third category of none (or similar), will also be included. This visualization can be filtered by: Political Party, Up for Election in 2016, Not up for Election in 2016, Region (West, South, MidWest, Northeast), Age, Gender
    - Our least defined visualization also has many different solutions. We are particularly interested in the possibility of using actual high resolution pictures of the candidates since there are less than 20 in total and most people do not know them all by name.

- **“Events since the Thaw”**: A 2D Brush on a timeline from December 17th to the present, visualizing the significant events that have transpired since the Thaw was announced. Such events include American businesses opening in Cuba, Conan O’brien hosting a special in Havana, Presidential executive orders, and pending legislation. The timeline is on an ordinal scale with markings for each month. Data points are stacked on top of each other during the month in which they occurred. This visualization allows the user to brush over a data point to see the data point’s state before the thaw. Once the brush is removed, the data point returns to its current state.
    - This one is pretty close to being decided on already. We like the visualization described above.

Data Processing.

“Do you expect to do substantial data cleanup? What quantities do you plan to derive from your data? How will data processing be implemented?”

We do not expect substantial data cleanup. The vast majority of our data has already been structured to our needs. Of course any new unstructured data, will have to be manually structured because automation is not likely to be possible. If it can be done, we will use web scraping to simplify the process but significant use of automation is not likely to be possible on things like political opinions gathered from TV interviews.

The most important derived quantities from our data are the values we assign to each tweet during sentiment analysis. As such, sentiment analysis is one of the most important functions in our projection. While neither Ivan nor I have extensive natural language processing experience, after carefully reading the literature on sentiment analysis we have come to the conclusion that there does not exist a one-size-fits-all solution and as such we will be using a using a variety of open-source and commercial tools in addition to custom software that we will write ourselves. Finally, it is important to note that because accurate sentiment analysis requires good training data, we will spend considerable amounts of time tagging our data by hand and when possible, we will also write scripts that can automate tagging.

Given our twitter dataset is relatively large at one million documents, we will preprocess much of it in order to minimize the amount of data manipulation that will have to be done client side. Because our data is mostly structured, filter and aggregation functions will work just as they normally do after we convert the CSV files to JSON. Therefore, we are not concerned about “Big Data” performance issues because preprocessing and aggregation can turn 1 million tweets with sentiment values into 50 data points representing the averages of each state. Our SQL database in the backend will store many of these derivative datasets for easy retrieval.

Must-Have Features.

"These are features without which you would consider your project to be a failure."

The Cuban Thaw has (4) major features each with several subviews. These features are critical and necessary because together they fulfill the project's mission to empower voters with a visual exploration of a major historical event and its aftermath. Our goal is to use data to tell a story that concludes with a powerful call to action. Each feature is like a chapter in the story.

(1) **"Tweet Map"** : A map of the US color coded to correspond to a geo-located sentiment analysis of tweets about Cuba from December 17 (day of), and December 18 (day after). It has (2) subviews. The first is a comparison of the tweet map to Republican and Democrat voting patterns from the US Presidential Election of 2012. The second is a comparison of the tweet map to the corresponding demographics in each area based on census information.

(2) **"Opinions of Congress"** : A visualization of how each member of Congress (House & Senate) feels about the thaw, either for or against, based on public statements. If no public statement can be found, then a third category of none (or similar), will also be included. This visualization can be filtered by:

- Political Party
- Up for Election in 2016
- Not up for Election in 2016
- Region (West, South, MidWest, North East)
- Age
- Gender

(3) **“Opinions of 2016 Presidential Candidates”** **:** A visualization of how each of the likely 2016 Presidential candidates feels about the Cuban Thaw, either for or against, based on public statements. If no public statement can be found, then a third category of none (or similar), will also be included. This visualization can be filtered by:

- Political Party
- Up for Election in 2016
- Not up for Election in 2016
- Region (West, South, MidWest, North East)
- Age
- Gender

(4) **“Events since the Thaw”** **:** A 2D Brush on a timeline from December 17th to the present, visualizing the significant events that have transpired since the Thaw was announced. Such events include American businesses opening in Cuba, Conan O’brien hosting a special in Havana, Presidential executive orders, and pending legislation. The timeline is on an ordinal scale with markings for each month. Data points are stacked on top of each other during the month in which they occurred. This visualization allows the user to brush over a data point to see the data point’s state before the thaw. Once the brush is removed, the data point returns to its current state.

Optional Features.

“Those features which you consider would be nice to have, but not critical.”

To tell a coherent, powerful story that satisfies our mission, we require all of the major features described in the section above; but if we have enough time, we can give the audience an even richer experience with more context by implementing the optional features described below.

Optional Feature 1. A visualization of the history of US-Cuba relations starting from the Spanish-American War to the Cuban Thaw.

Optional Feature 2. A visualization of the US vs Cuba based on various World Bank statistics using the Caribbean / Latin America averages as a baseline.

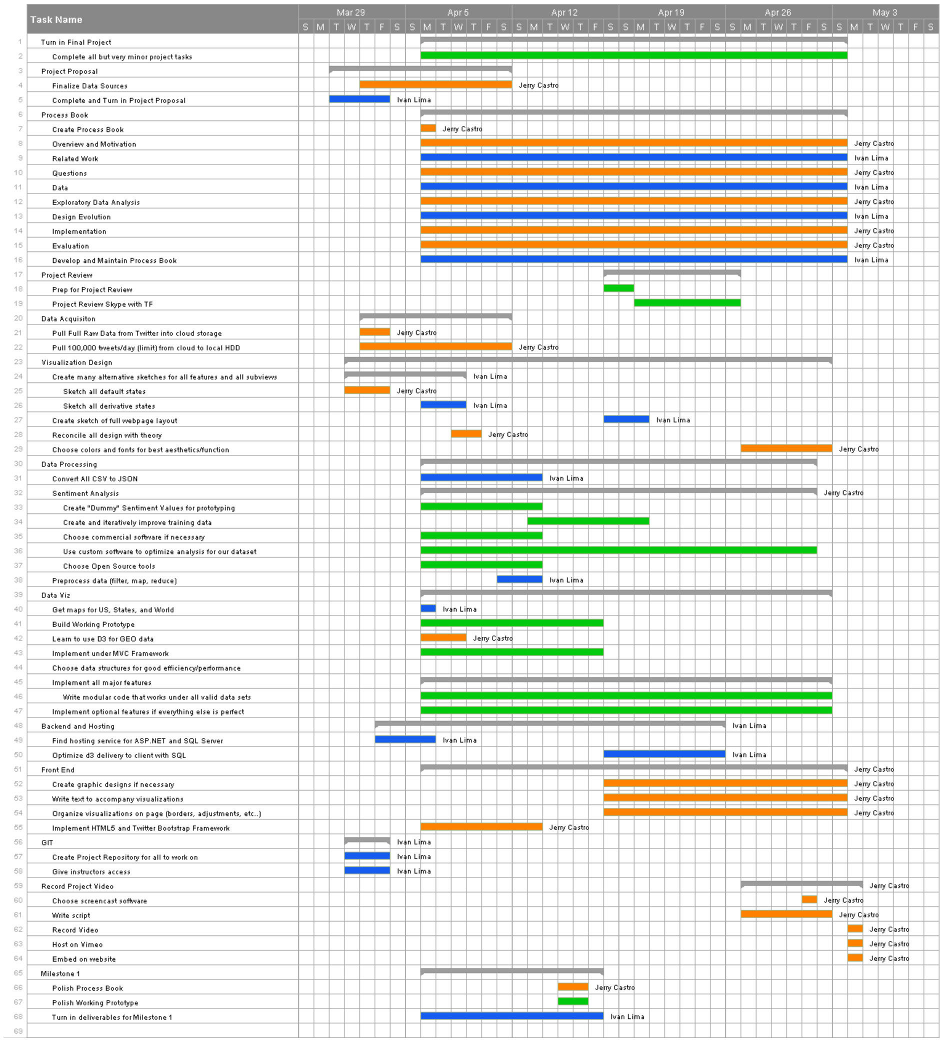
Optional Feature 3. A tweet map of the entire world showing how the rest of the world feels about the Cuban Thaw compared to how people feel in the US.

Optional Feature 4. Allow users to “mouseover” a region on the tweet map to see a sample tweet from that particular area.

Each one of these optional features can stand on its own. That is, any one the optional features can be added to the project with no requirement that we also implement any of the other ones. Therefore, we will implement as many optional features as time allows but it must be noted that there is no expectation that we will have extra time so it is in fact very possible that none of these optional features will be implemented by the May 5th deadline.

Project Schedule.

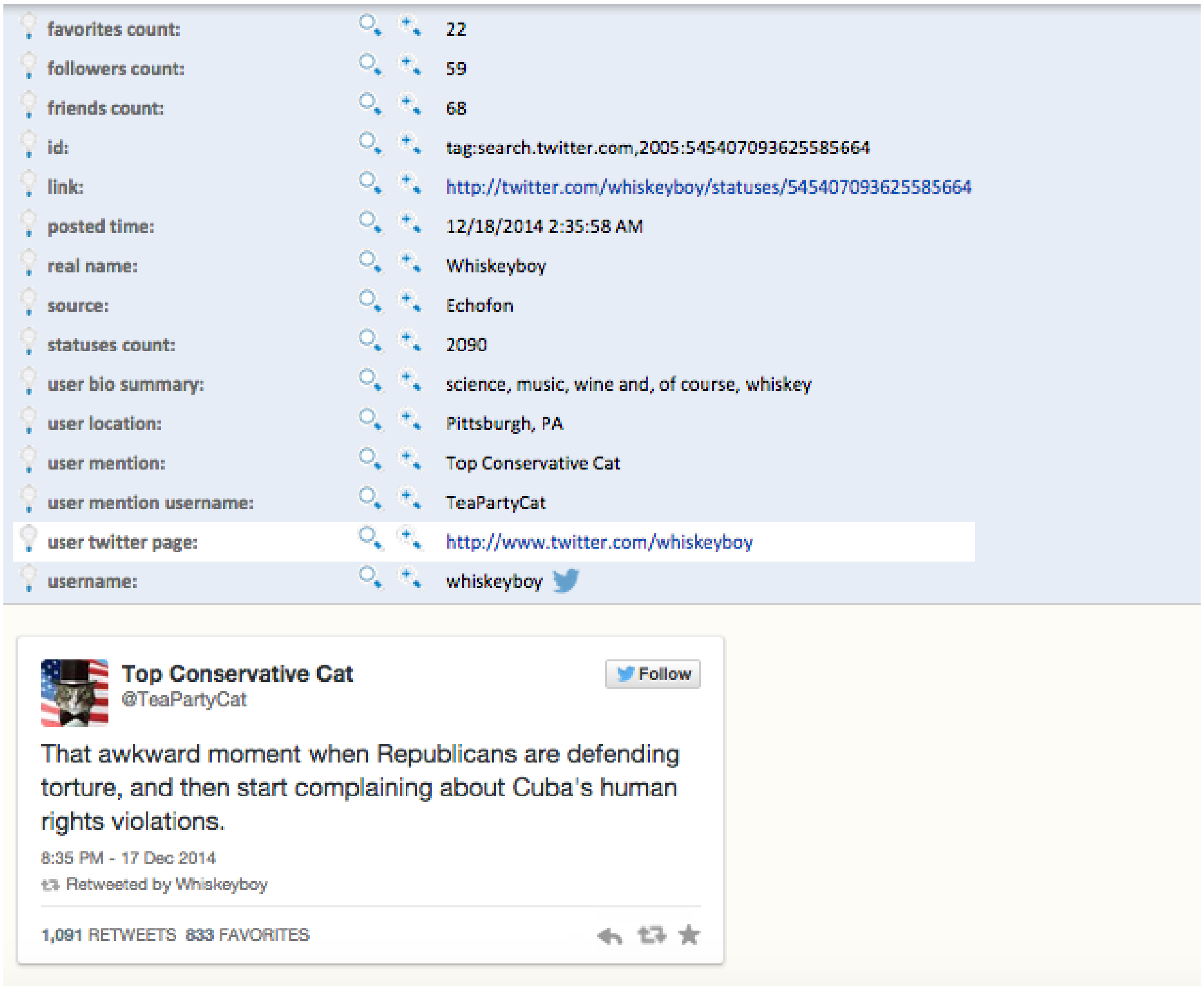
Make sure that you plan your work so that you can avoid a big rush right before the final project deadline, and delegate different modules and responsibilities among your team members. Write this in terms of weekly deadlines.



	Task Name	
1	Turn in Final Project	
2	Complete all but very minor project tasks	
3	Project Proposal	
4	Finalize Data Sources	Jerry Castro
5	Complete and Turn in Project Proposal	Ivan Lima
6	Process Book	
7	Create Process Book	Jerry Castro
8	Overview and Motivation	Jerry Castro
9	Related Work	Ivan Lima
10	Questions	Jerry Castro
11	Data	Ivan Lima
12	Exploratory Data Analysis	Jerry Castro
13	Design Evolution	Ivan Lima
14	Implementation	Jerry Castro
15	Evaluation	Jerry Castro
16	Develop and Maintain Process Book	Ivan Lima
17	Project Review	
18	Prep for Project Review	
19	Project Review Skype with TF	
20	Data Acquisiton	
21	Pull Full Raw Data from Twitter into cloud storage	Jerry Castro
22	Pull 100,000 tweets/day (limit) from cloud to local HDD	Jerry Castro
23	Visualization Design	
24	Create many alternative sketches for all features and all subviews	Ivan Lima
25	Sketch all default states	Jerry Castro
26	Sketch all derivative states	Ivan Lima
27	Create sketch of full webpage layout	Ivan Lima
28	Reconcile all design with theory	Jerry Castro
29	Choose colors and fonts for best aesthetics/function	Jerry Castro
30	Data Processing	
31	Convert All CSV to JSON	Ivan Lima
32	Sentiment Analysis	Jerry Castro
33	Create "Dummy" Sentiment Values for prototyping	
34	Create and iteratively improve training data	
35	Choose commercial software if necessary	
36	Use custom software to optimize analysis for our dataset	
37	Choose Open Source tools	
38	Preprocess data (filter, map, reduce)	Ivan Lima
39	Data Viz	
40	Get maps for US, States, and World	Ivan Lima
41	Build Working Prototype	
42	Learn to use D3 for GEO data	Jerry Castro
43	Implement under MVC Framework	
44	Choose data structures for good efficiency/performance	
45	Implement all major features	
46	Write modular code that works under all valid data sets	
47	Implement optional features if everything else is perfect	
48	Backend and Hosting	Ivan Lima
49	Find hosting service for ASP.NET and SQL Server	Ivan Lima
50	Optimize d3 delivery to client with SQL	Ivan Lima
51	Front End	Jerry Castro
52	Create graphic designs if necessary	Jerry Castro
53	Write text to accompany visualizations	Jerry Castro
54	Organize visualizations on page (borders, adjustments, etc..)	Jerry Castro
55	Implement HTML5 and Twitter Bootstrap Framework	Jerry Castro
56	GIT	Ivan Lima
57	Create Project Repository for all to work on	Ivan Lima
58	Give instructors access	Ivan Lima
59	Record Project Video	Jerry Castro
60	Choose screencast software	Jerry Castro
61	Write script	Jerry Castro
62	Record Video	Jerry Castro
63	Host on Vimeo	Jerry Castro
64	Embed on website	Jerry Castro
65	Milestone 1	
66	Polish Process Book	Jerry Castro
67	Polish Working Prototype	
68	Turn in deliverables for Milestone 1	Ivan Lima
69		

Attachments

Sample Tweet with Metadata.



favorites count:	22
followers count:	59
friends count:	68
id:	tag:search.twitter.com,2005:545407093625585664
link:	http://twitter.com/whiskeyboy/statuses/545407093625585664
posted time:	12/18/2014 2:35:58 AM
real name:	Whiskeyboy
source:	Echofon
statuses count:	2090
user bio summary:	science, music, wine and, of course, whiskey
user location:	Pittsburgh, PA
user mention:	Top Conservative Cat
user mention username:	TeaPartyCat
user twitter page:	http://www.twitter.com/whiskeyboy
username:	whiskeyboy

Top Conservative Cat
@TeaPartyCat

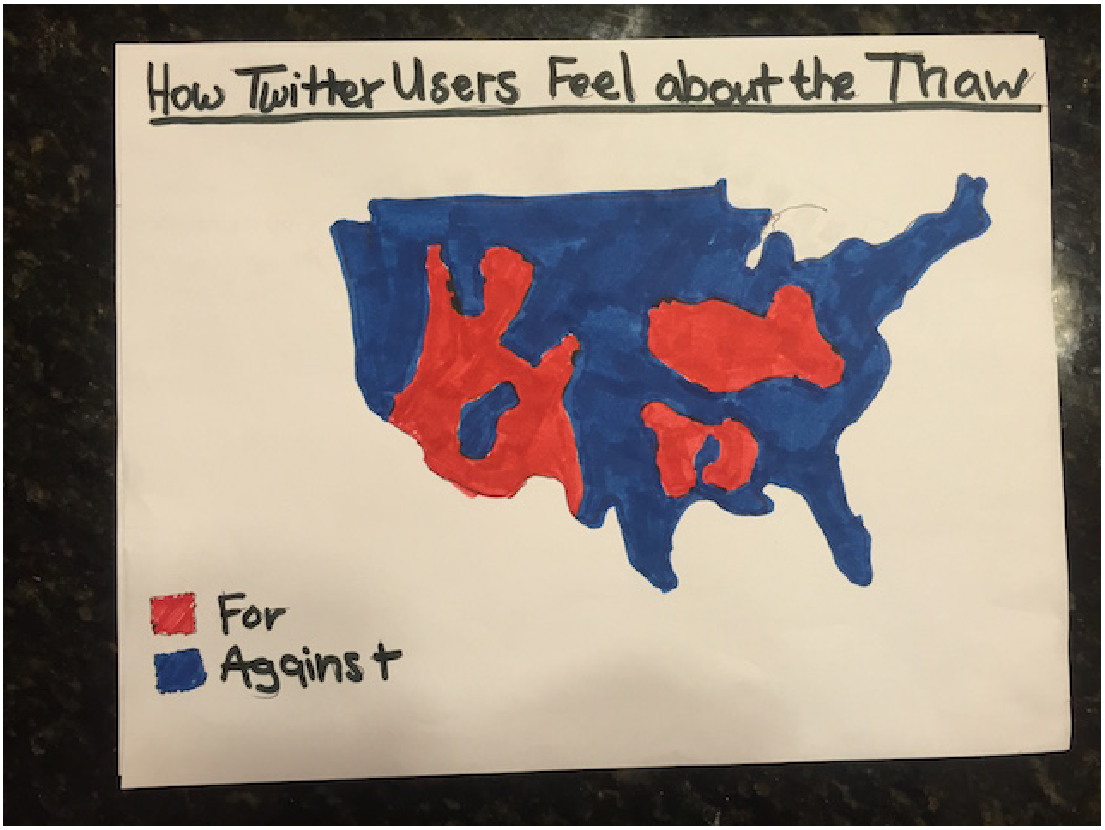
That awkward moment when Republicans are defending torture, and then start complaining about Cuba's human rights violations.

8:35 PM - 17 Dec 2014

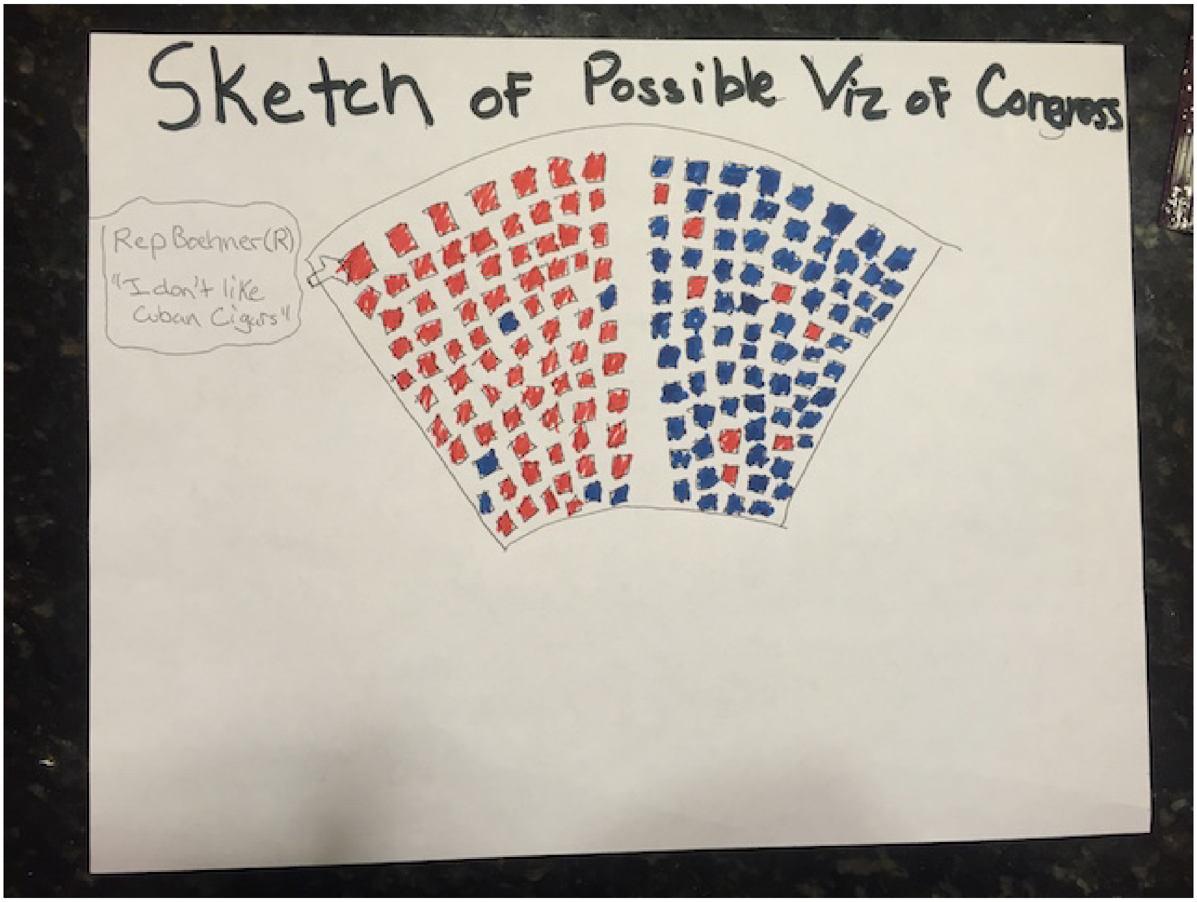
Retweeted by Whiskeyboy

1,091 RETWEETS 833 FAVORITES

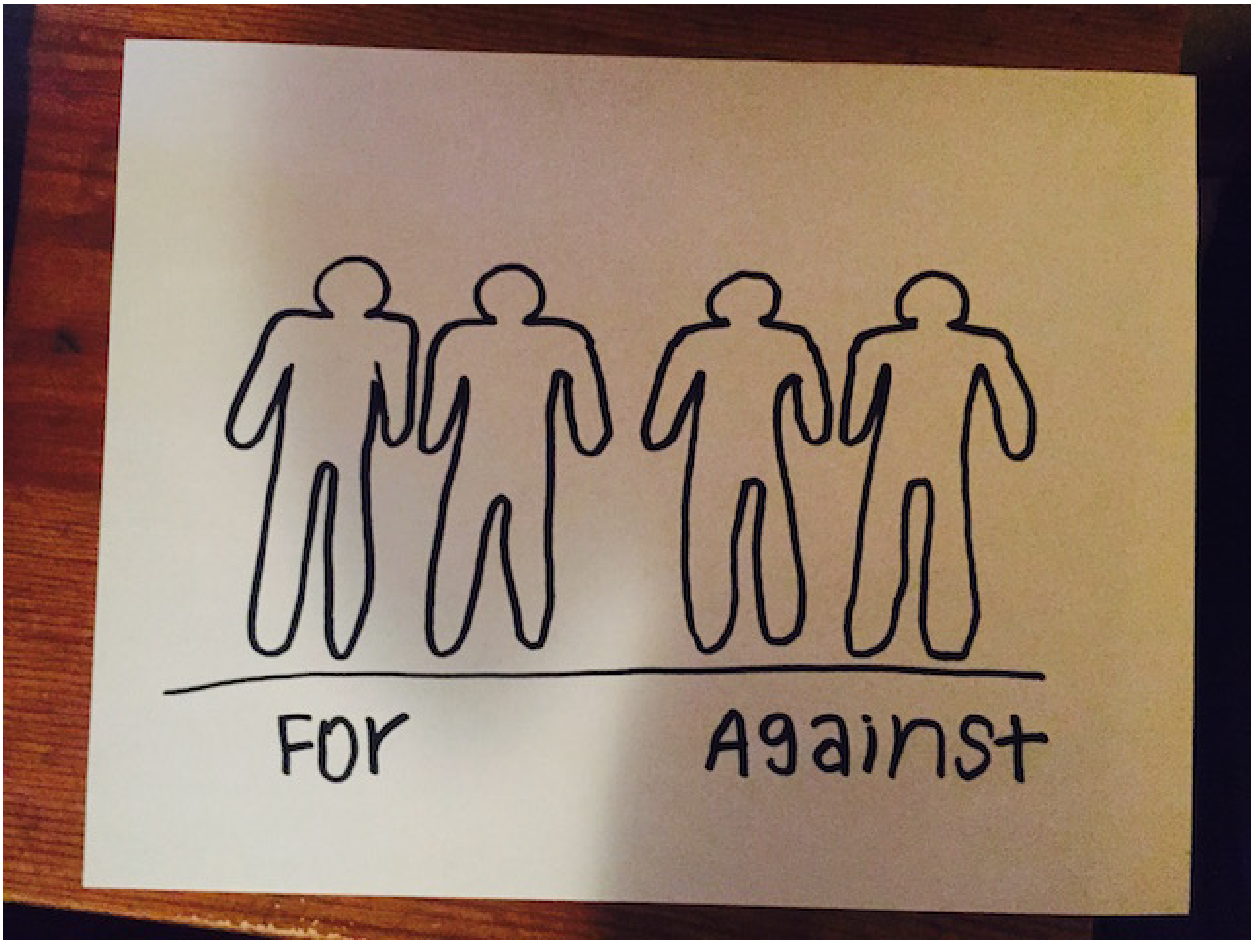
Tweet Map



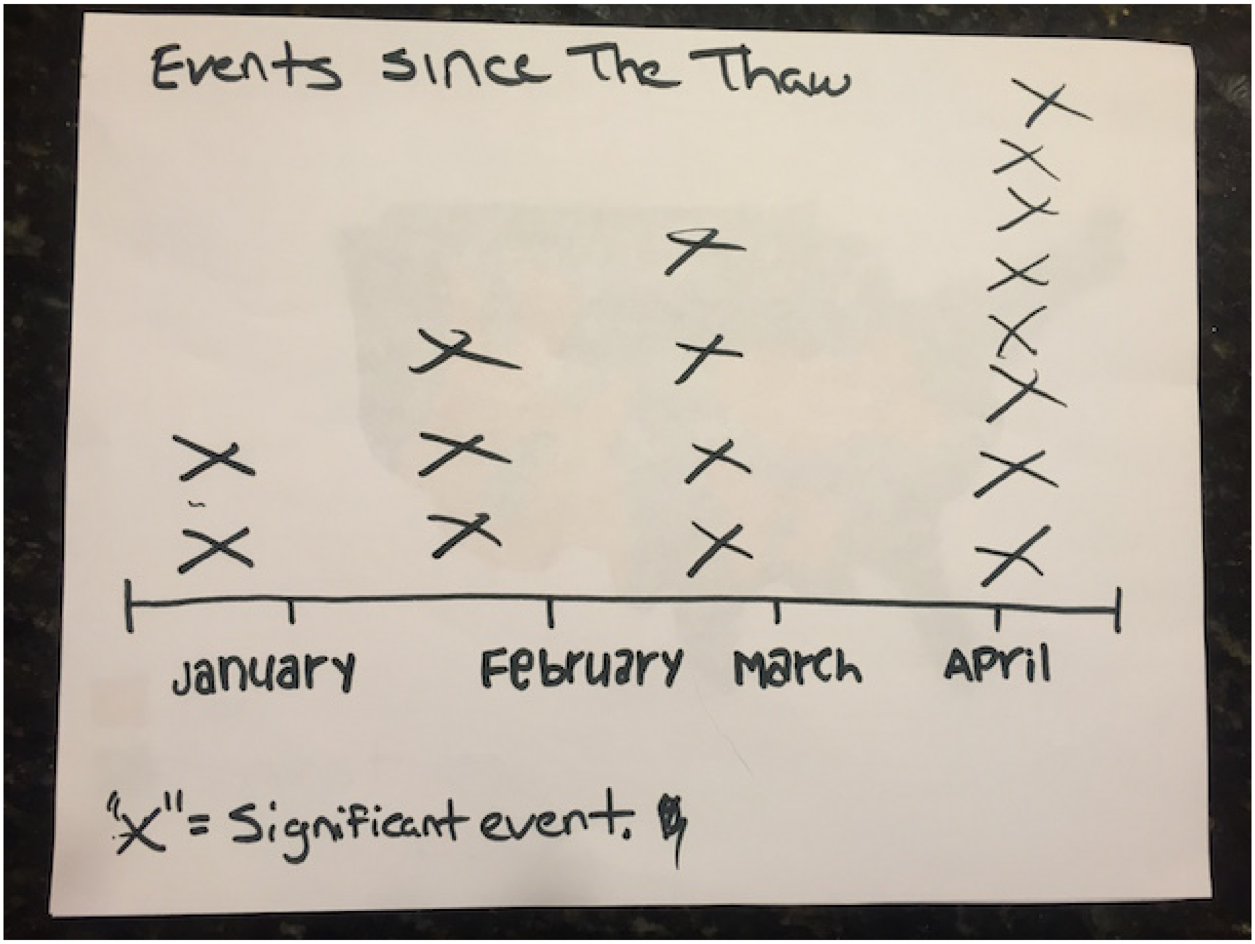
Opinions of Congress


Sketch of Possible Viz of Congress
Rep Boehner (R)
"I don't like Cuban Cigars"

Opinions of 2016 Presidential Candidates



Events since the Thaw


Events since The Thaw
January
February
March
April
"X" = Significant event.