Health Care and Big Data

What data is available?

- Survey data.
- Panel data. Sources Like Nielson.
- Social Network Data.

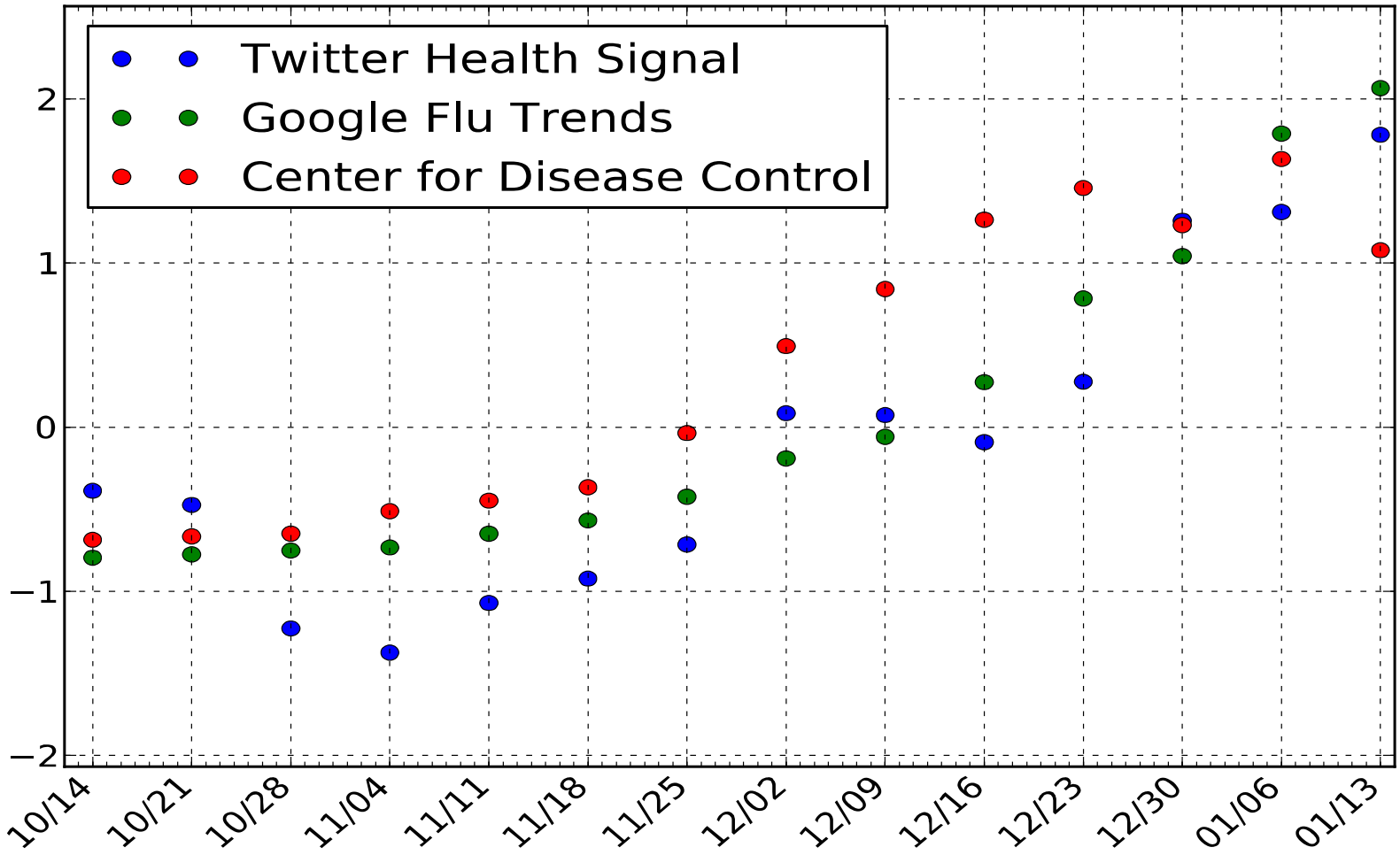
Possibilities:

- Predict probabilities of getting ill at various metropolitan cities at any given time. (This can be used to help designate more attendants at hospitals when sickness chances are greater)
- Analyze the spread of illness after specific incidents. (like the July 2010 toxic material leak in the Pacific Ocean near Long Island)

Problems with use of Social Network data:

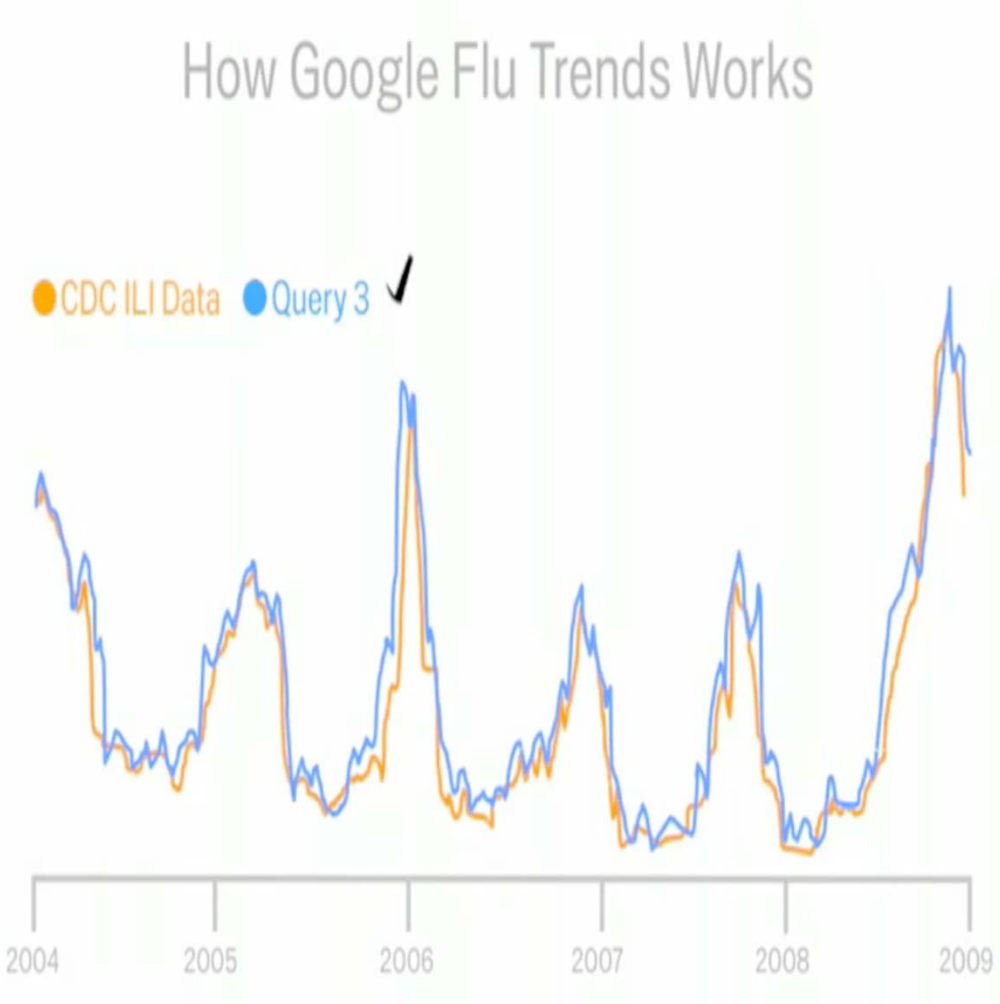
- Relevant data is very small depending on the project. (Consider tracking tweets from a group of people who visit a certain location at a certain time period and tweet from that location.)
- Data is not always accurate. We need to account for uncertainty.
- Data is not always easily interpretable. Consider: “I am so sick”vs “I am having a flu”.

Accuracy of Social Network data:



Google Flu Trends:

- Methodology: When people are sick they google the word “FLU” or their friends Google the word “FLU”.
- Timely Prediction: The CDC usually takes 2 weeks to gather data on flu and predict the trends.
- Accuracy: They are as accurate:



General Twitter Learning Overview:

- Supervised Learning on twitter data using trial and test sets to generate a probability rubric for each word. For the learning data must be labeled manually first.
- Plotting these tweets either on a real life map, tree map, heat map (whatever suffices the purpose of the experiment) with live points that show the tweet and the color of the point based on the probability of the tweet being a sick tweet.
- Analyzing these plots to conclude the correctness of the hypothesis.

Getting Relevant Data:

- Use Data Providers like Datasift (we currently use them).
  - They provide labeled data.
  - Historical data from all parts of the world are stored on their server.
  - They are very expensive ($85K a year for average usage)
- Download your own data:
  - Need to make a Crawler that processes and parses twitter webpage data to get relevant data.
  - Difficult to get Historical Timely data, it(needed for research purposes) is stored in descending order of time.

Extracting data from Twitter:

- Downloading data through a Machine Learner using the Twitter Api to extract relevant information.
- Twitter API has references to parse Regex to help extract relevant information for web developers (This is very useful when designing your own crawler).

Unprocessed Tweet:

- {"count":697,"hash":"18c936829e1c4a8fca05","hash_type":"historic","id":"f952cf77b09efae442ec987966a5e233","delivered_at":"Tue, 15 Oct 2013 07:14:33 +0000","interactions":[{"interaction":{"author":{"id":"\
- 243318676","link":"http://twitter.com/243318676","name":"\u0645\u062D\u0645\u062F \u062F\u0627\u0648\u062F","username":"mohammed_dawod"},"content":"@dudu_cba ;)","created_at":"Thu, 21 Jul 2011 22:49:11 +\
- 0100","geo":{"latitude":40.66552753200000,"longitude":-73.78504682000001},"id":"1e0b3e3448aaa580e07487dd4fc20001","link":"http://twitter.com/mohammed_dawod/statuses/94162026816995329","schema":{"version"\
- :2},"source":"oauth:129032","type":"twitter"}

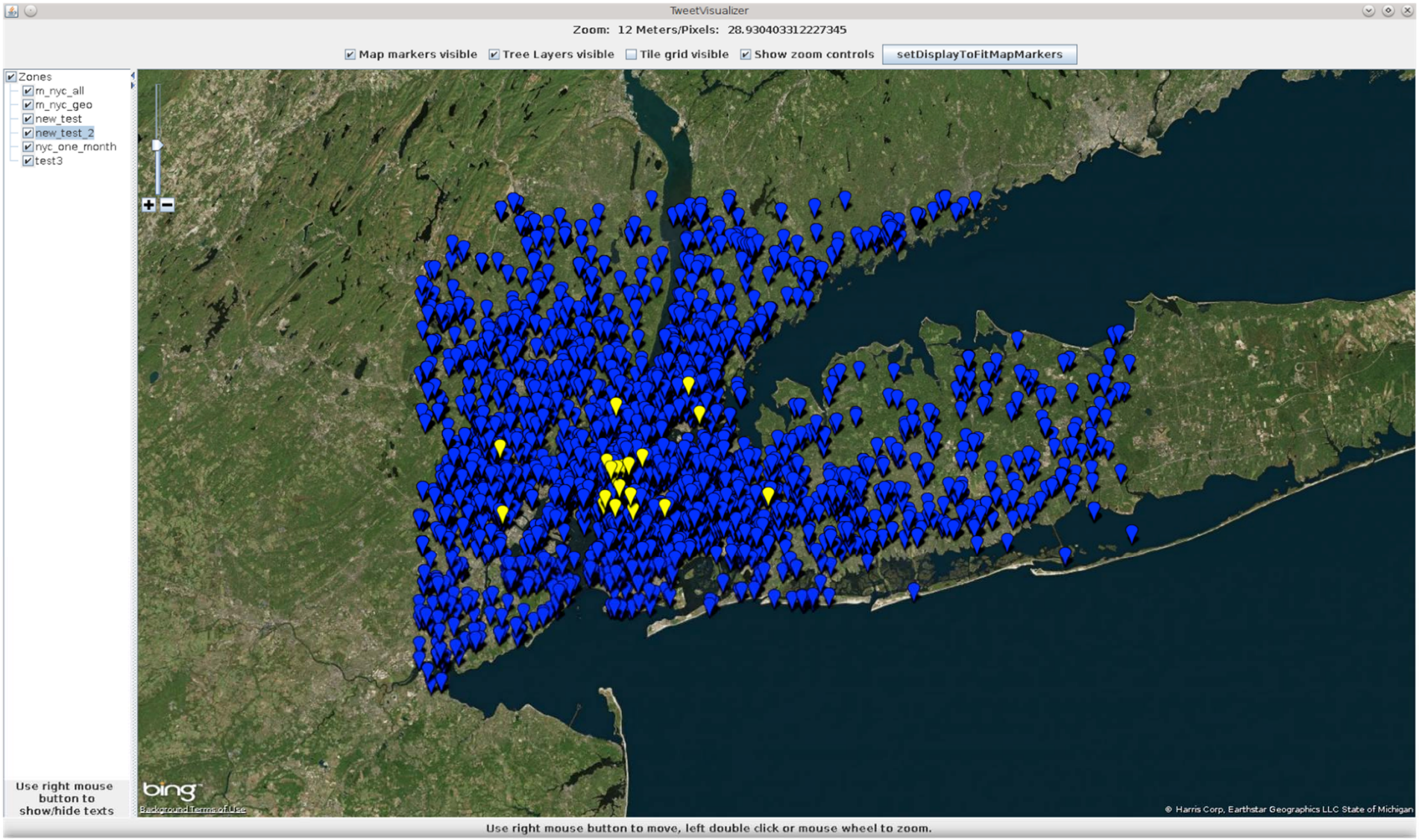
Twitter Health at University of Rochester: Table Rubric Making

- Learn a table of marginal probability contribution for each word or group of words for a tweet.

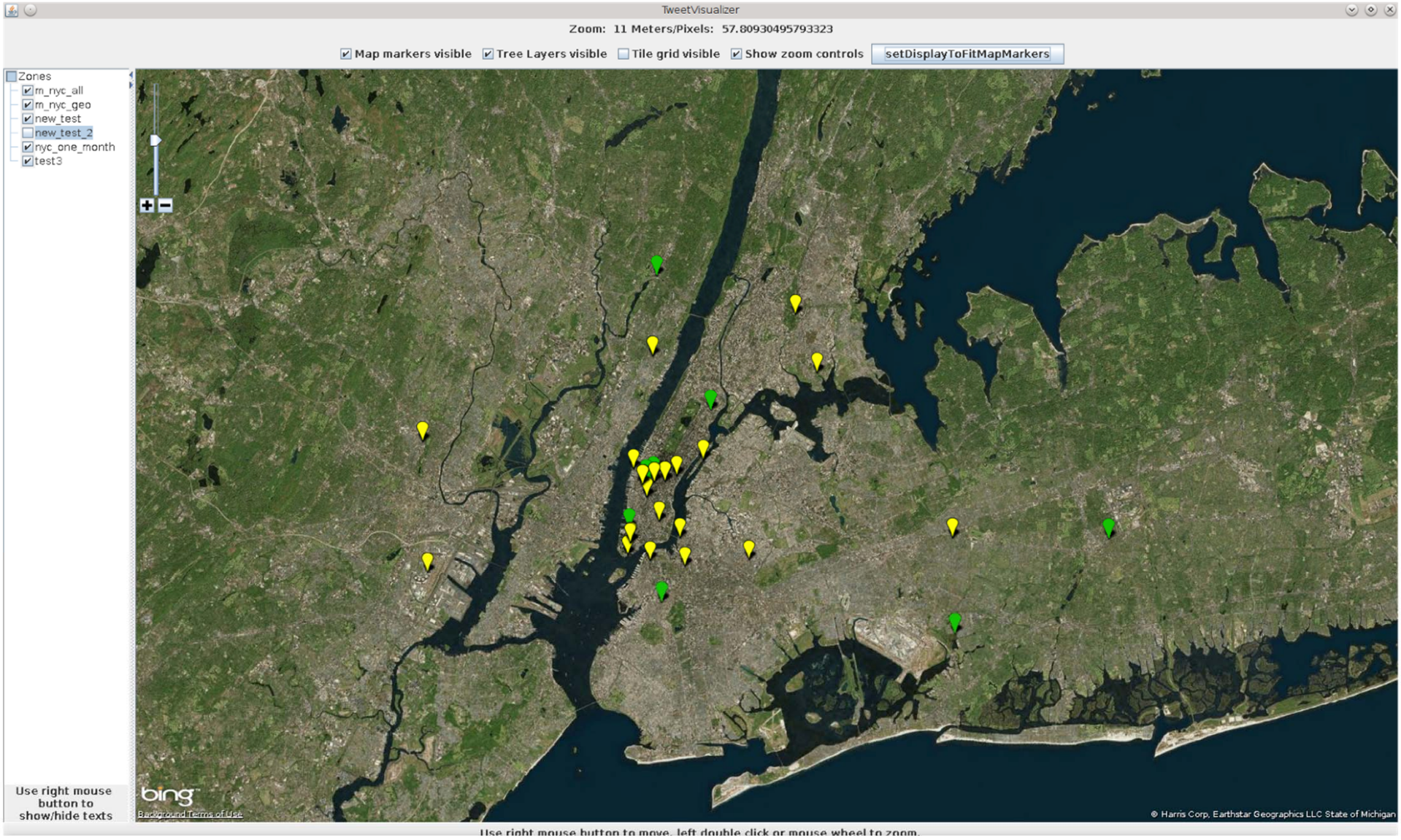
Word attribution table				
Positive Drivers			**Negative Drivers**	
Word	**Probability drive**		**Word**	**Probability Drive**
sick	0.2		sick of	-0.05
flu	0.4		bieber fever	-0.39
ache	0.12		sick of life	-0.32
fever	0.08		so sick	-0.09
Cutoff: 0.65				

Twitter Health at University of Rochester: Plotting Findings

Twitter Health at University of Rochester: Concluding correctness eg1



Twitter Health at University of Rochester: Concluding correctness eg2



Troubles with Visualization:

- Plotting large amounts of data on a pixelated format can be detrimental to the RAM.
- Data querying is expensive. When you click on a Pin, we need to display the tweet information and stuff. Requerying is log n after sorting (binary search).
- Storing the data is very expensive, mostly impossible. Storing a Million Strings in Static Memory + Accessing them.

References:

- References: Towards understanding global spread of disease from everyday personal interaction, Henry Kautz, Adam Sadilek and Sean Brennan. http://www.cs.rochester.edu/u/kautz/papers/ijcai2013airports.pdf, accessed on oct 23 2013
- Collaborators: John Hinkel (UG), Henry Kautz, Jiebo Luo, Tianran Hu, Roya et al.