Mining Geo-tagged Data to Predict Disease Transmission

Xinyue Liu

CS548 Showcase
Worcester Polytechnic Institute

Reference

Papers

- Sadilek, Adam, Henry A. Kautz, and Vincent Silenzio. "Predicting Disease Transmission from Geo-Tagged Micro-Blog Data." *AAAI*. 2012.
- Sadilek, Adam, Henry A. Kautz, and Vincent Silenzio. "Modeling Spread of Disease from Social Interactions." *ICWSM*. 2012.

Chapters&Books

- Twitter API : Chapter 1, Mining the Social Web (2nd), O'Reilly 2013.
- SVM : Chapter 5, Introduction to Data Mining, 2008.
- SVM : Chapter 9, Data Mining: Concepts and Techniques (3rd Ed), 2012

Web Pages

- Sadilek's research page: http://www.cs.rochester.edu/~sadilek/research/
- *Germtraker*: http://germtracker.org
- *SVM-light*: http://svmlight.joachims.org
- *Python twitter toolset*: https://pypi.python.org/pypi/twitter

Tweets

Geo Data & Illness-Related Message

Tweets

Geo Data & Illness-Related Message

Facts

Twitter launched in 2006

# of total active users = 645,750,000

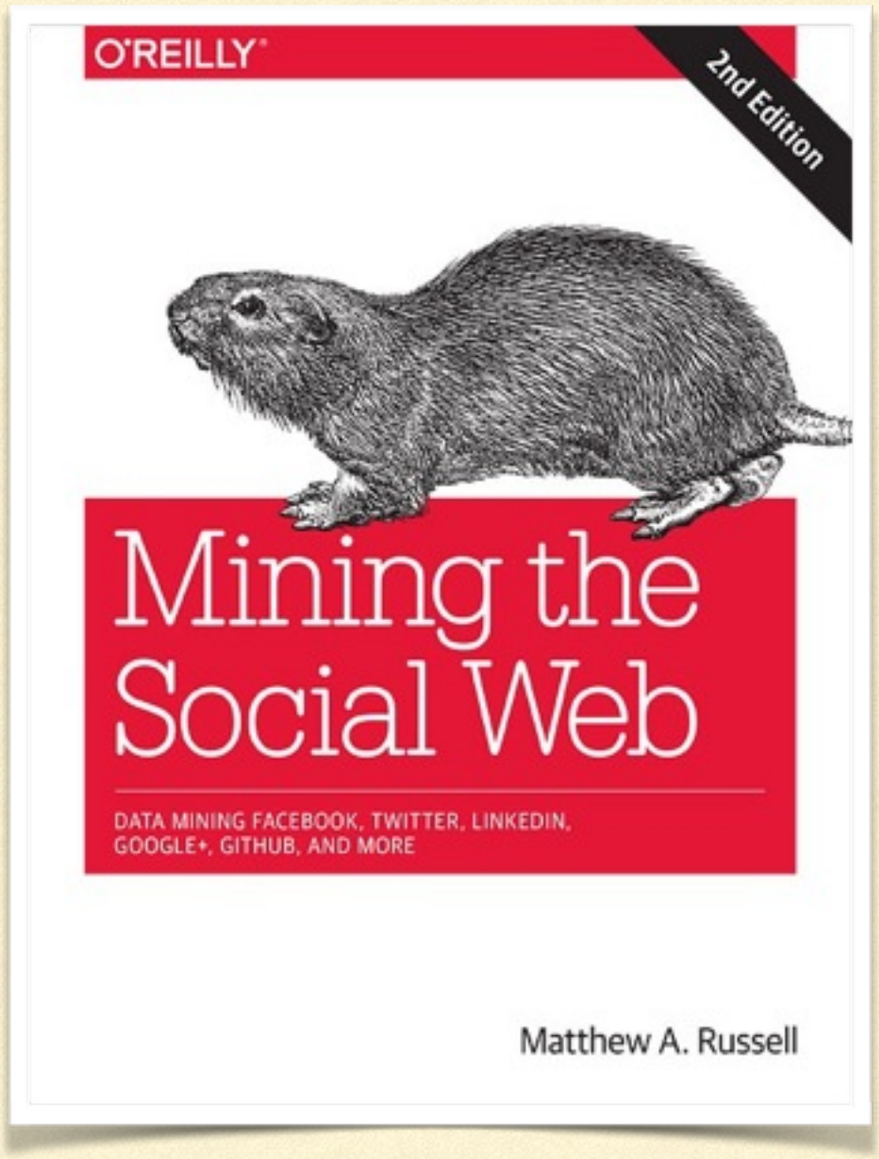
# of tweets every second = 9,100

Big Picture

- Collect data(tweets) from twitter
- Extract illness-related tweets
- Build model for prediction
- Launch experiment to evaluate the model

Twitter API

- Twitter offers public APIs
- All APIs are RESTful
- OAuth Protocol
- Create App in Twitter Developer (apps.twitter.com)
- API Key / Access Token
- Python Package: twitter



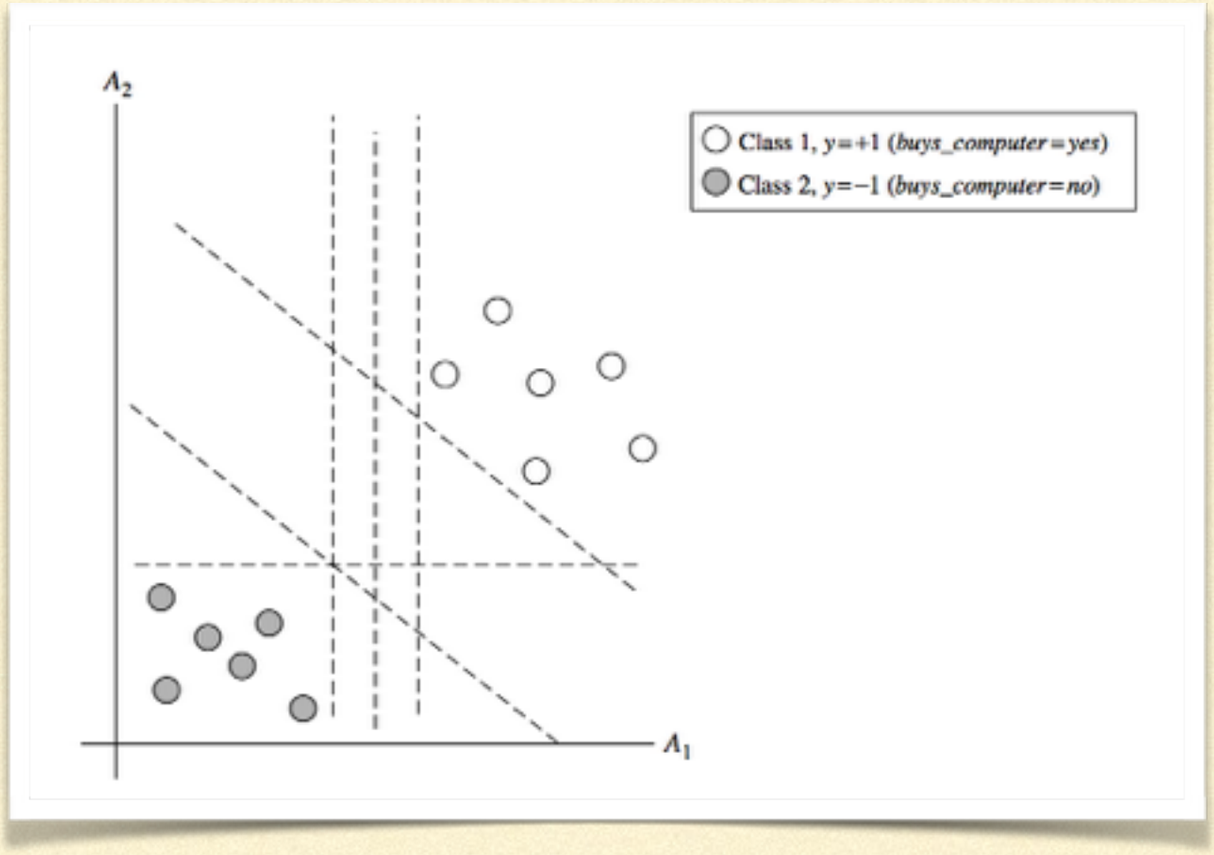
NEW YORK CITY DATASET	
UNIQUE USERS	632,611
UNIQUE GEO-ACTIVE USERS	6,237
TWEETS TOTAL	15,944,084
GPS-TAGGED TWEETS	4,405,961
GTT BY GEO-ACTIVE USERS	2,535,706
GTT BY GEO-ACTIVE USERS SHOWS A SYMPTOM OF AN ILLNESS	2,047
DISTINCT VISITED LOCATION	57,109
FOLLOWS RELATIONSHIP OF GAU	102,739
FRIENDS RELATIONSHIP OF GAU	31,874

Feeling miserable. stomach hurts, headache, and no, I'm not pregnant.

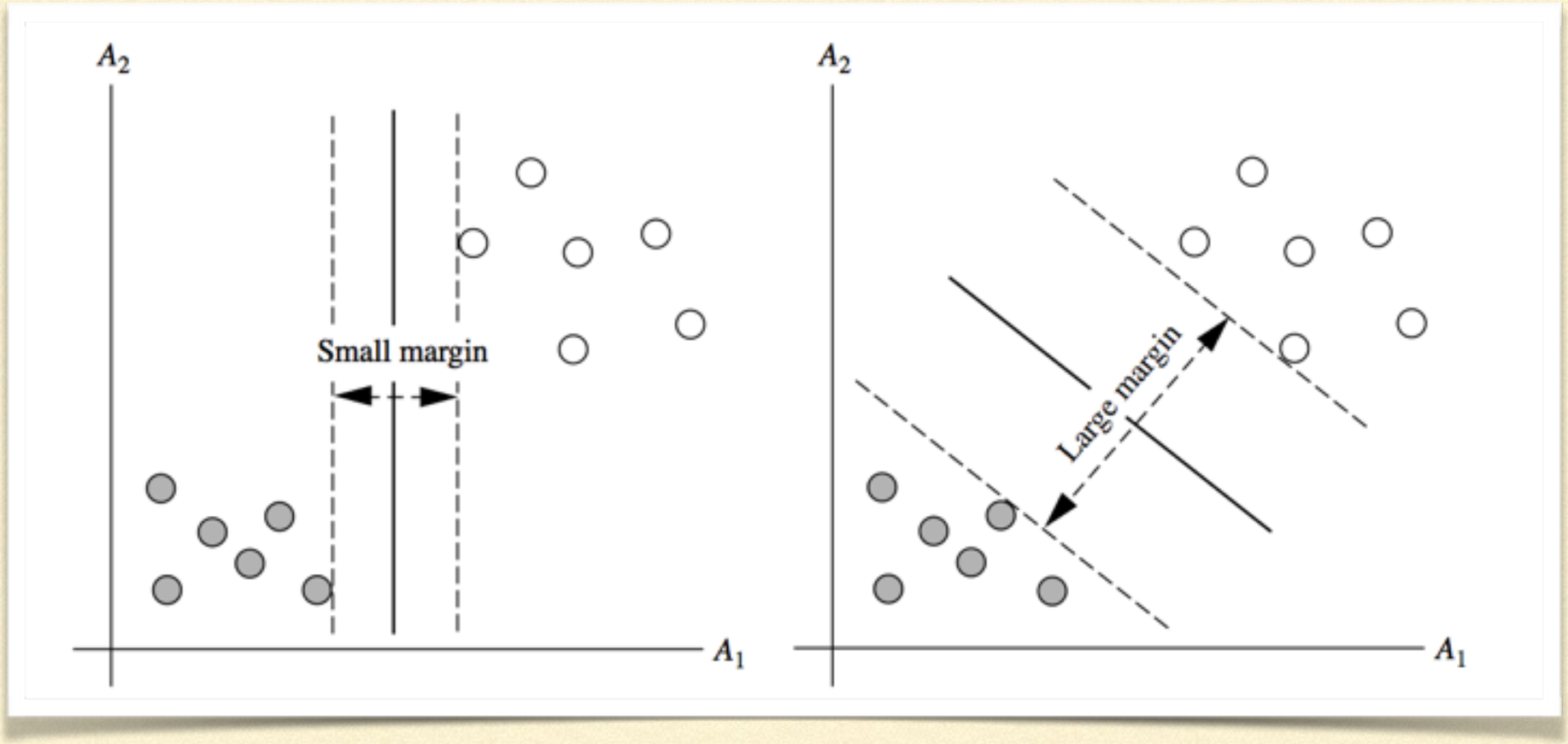
Meh I actually have to go to school tomorrow.. #sick

I am so sick of school and I have another month left

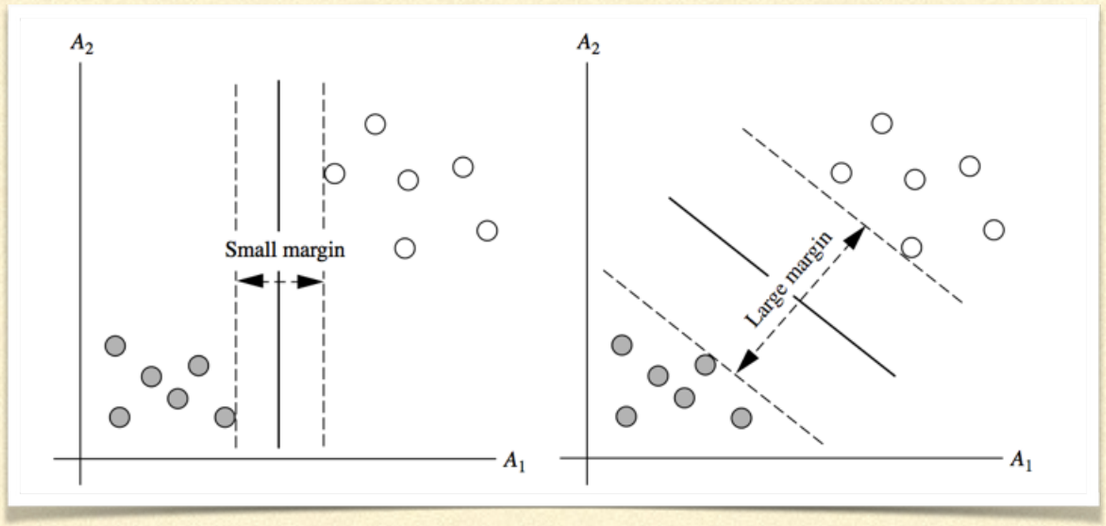
SVM



SVM



SVM



$$X^T = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{bmatrix}$$

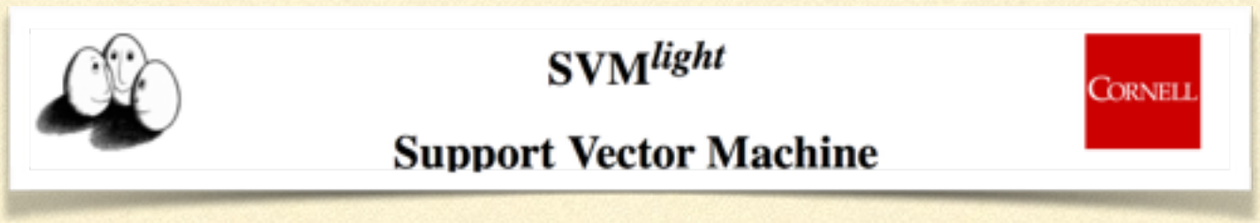
X_i : support vector

α_i : Lagrange multiplier

y_i : +1 or -1

$$d(X^T) = \sum_{i=1}^{l} y_i \alpha_i X_i X^T$$

SVM^{light}



SVM^{light} is an implementation of Support Vector Machines (SVMs) in C

SVM^{light} is open-source and free to use

Researcher use SVM^{light} train the model to label tweets

Some Alternatives:

Weka + LibSVM

SVM^{struct}

SVM^{perf}

Process Tweets

I fell sick

Process Tweets

i fell sick

Process Tweets

i fell sick

i	x_1
fell	x_2
sick	x_3
i fell	x_4
fell sick	x_5
i fell sick	x_6

Process Tweets

i fell sick

(1,1,1,1,1,1,0,0,…,0)

i	x_1
fell	x_2
sick	x_3
i fell	x_4
fell sick	x_5
i fell sick	x_6

The final model is in 1.7 million dimensions

Problem

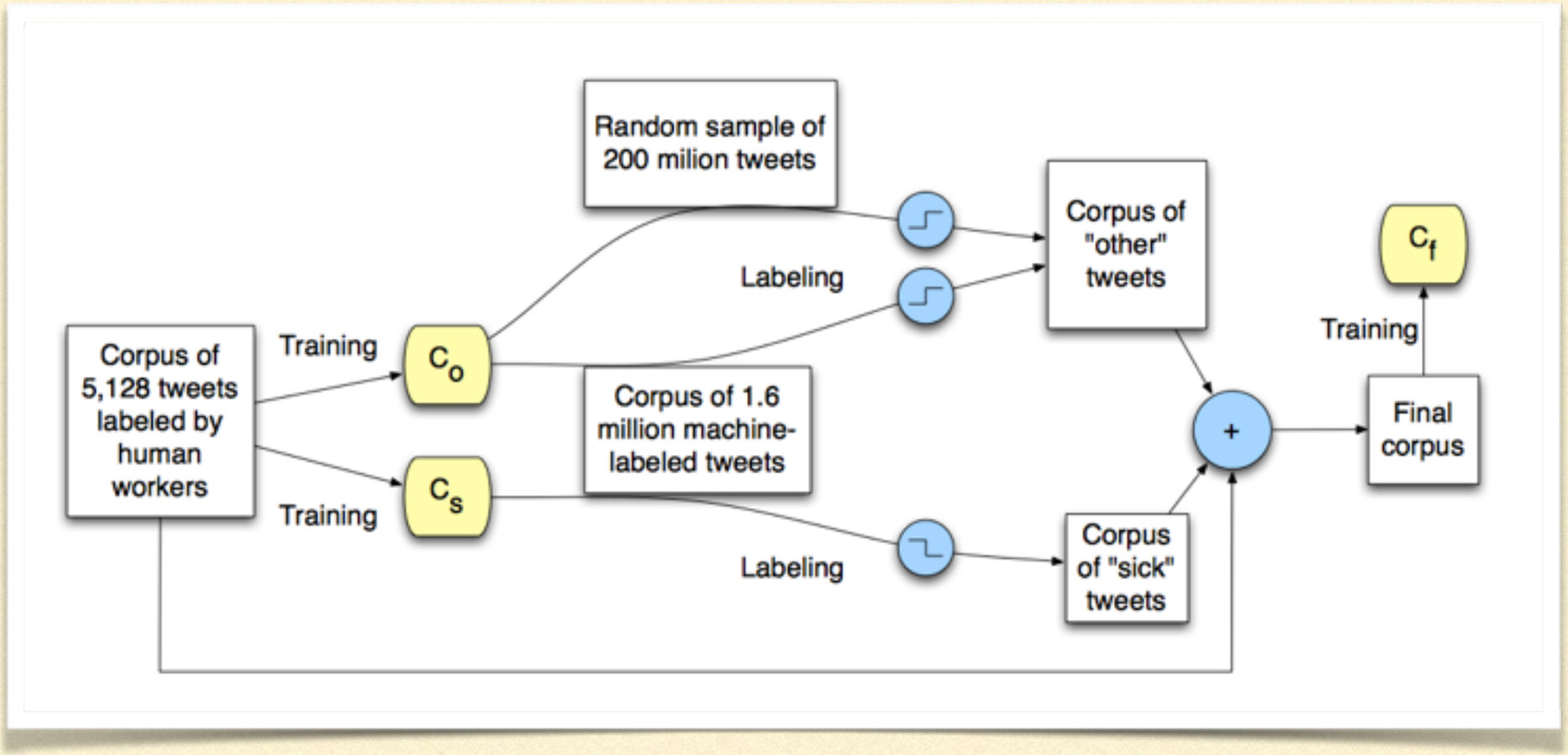
We collected millions of tweets

but only 2,047 tweets are illness related

It seems not enough to train a good model

Bootstrapping!

Bootstrapping



SVM Model

Positive Features		Negative Features	
Feature	**Weight**	**Feature**	**Weight**
sick	0.9579	sick of	−0.4005
headache	0.5249	you	−0.3662
flu	0.5051	of	−0.3559
fever	0.3879	your	−0.3131
feel	0.3451	lol	−0.3017
cough	0.3062	who	−0.1816
feeling	0.3055	u	−0.1778
coughing	0.2917	love	−0.1753
throat	0.2842	it	−0.1627
cold	0.2825	her	−0.1618
home	0.2107	they	−0.1617
still	0.2101	people	−0.1548
bed	0.2088	shit	−0.1486
better	0.1988	smoking	−0.0980
being	0.1943	i'm sick of	−0.0894
being sick	0.1919	so sick of	−0.0887
stomach	0.1703	pressure	−0.0837
and my	0.1687	massage	−0.0726
infection	0.1686	i love	−0.0719
morning	0.1647	pregnant	−0.0639

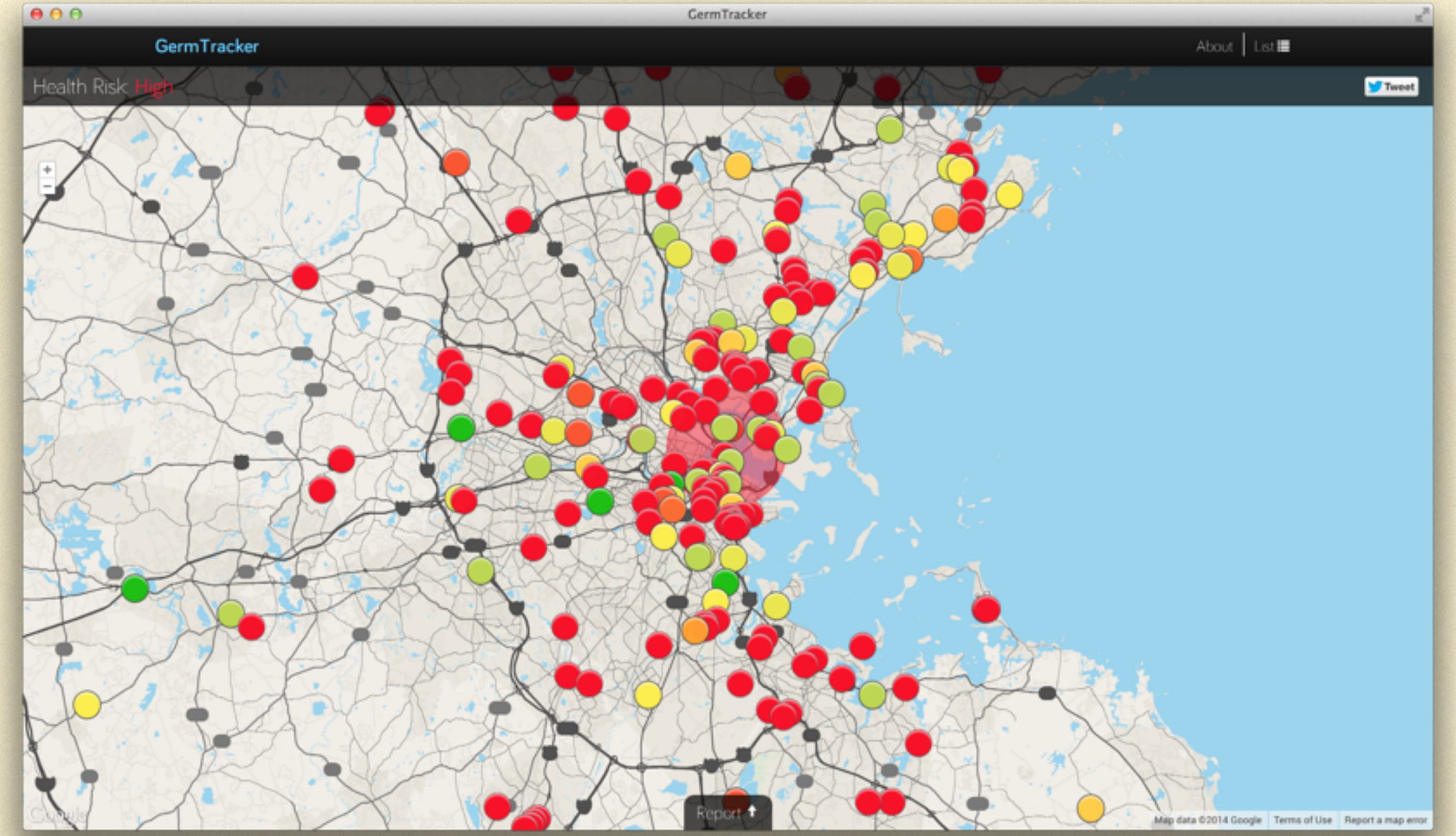
SVM Model

Precision = 0.98

Recall = 0.97

Positive Features		Negative Features	
Feature	**Weight**	**Feature**	**Weight**
sick	0.9579	sick of	−0.4005
headache	0.5249	you	−0.3662
flu	0.5051	of	−0.3559
fever	0.3879	your	−0.3131
feel	0.3451	lol	−0.3017
cough	0.3062	who	−0.1816
feeling	0.3055	u	−0.1778
coughing	0.2917	love	−0.1753
throat	0.2842	it	−0.1627
cold	0.2825	her	−0.1618
home	0.2107	they	−0.1617
still	0.2101	people	−0.1548
bed	0.2088	shit	−0.1486
better	0.1988	smoking	−0.0980
being	0.1943	i'm sick of	−0.0894
being sick	0.1919	so sick of	−0.0887
stomach	0.1703	pressure	−0.0837
and my	0.1687	massage	−0.0726
infection	0.1686	i love	−0.0719
morning	0.1647	pregnant	−0.0639

GermTacker


GermTracker
GermTracker
About
List
Health Risk High
Tweet
Report
Map data ©2014 Google
Terms of Use
Report a map error

GermTracker

Real-time

Mobile Compatible