Overview
The Internet has become a hotbed of social activism and political organization. We created a classifier that takes publicly available data (Google Trends and market sentiment surveys) and predicts whether a mass movement will occur in a specific metropolitan area. We defined mass movement to be any large movement (1000+ people) gathered for a political/social cause.

Features

Google Trends Keywords: Identified 30 major protests in 26 metropolitan areas (in GB, CA, and USA), and generated the previous three months’ worth of Google Trends data for each specific locality, corresponding to over 20 related to protest movements. We generated a similar amount of data for non-protest movements. This resulted in about 78,000 datapoints.

Market Sentiment: Measures the percentage of individual investors who are bullish, bearish, and neutral on the stock market for the next six months. Sourced 2011-2016 weekly data from Quandl

Unemployment: Weekly unemployment rates, sourced from Quandl

Methodology

Feature Selection (Forward Search)
Using forward search, we selected combination with the best precision scores. This is due to the relative rarity of protests occurring in a calendar year. By doing so, we penalize models with a high rate of false positives.

Keyword Combination with Top Precision: protests, militarization, civil resistance, movements, progressive, oppression, strike, debate, civil rights, boycott, march, ceremony

Binary / Multiclass Labeling
For binary classification, we classified whether an example was within a specific date range. For multiclass, we classified which date interval preceding the protest date the example fell into (0-3 days, 4-6 days, etc)

Classification example: 2014-08-02
Protest: 2014-08-09 (Black Lives Matters, Cincinnati)
Binary Class (with interval 10 days): is a protest
Multi Class (with intervals 3,7,10 days): 7 days

Results

Gradient Boosting consistently produced the best results. With a test set of 537, we were able to achieve 77% accuracy, 76% precision, and 75% recall on our binary classifier. While this was not our best accuracy score, it produces the best precision.

Our original set was very skewed with only 10% of examples labeled positive. This resulted in high accuracy and low recall. By narrowing our training and test sets, we were able to dramatically improve recall and precision.

References
Quandl, www.quandl.com
Google Trends, www.google.com/trends

Moving Forwards
We hope to increase accuracy by analyzing social media interaction graphs (Facebook, Instagram). This will allow us to look at individuals’ actions rather than regional trends. We also hope to apply NLP techniques to the massive GDelt database to analyze daily events across the country.