 MOTIVATION
• Influenza (flu) is a contagious respiratory illness caused by influenza viruses.
• 5 to 20% of population gets flu.
• ~200,000 people are hospitalized from flu-related complications.
• Worldwide death toll is 250,000 to 500,000.
• Results in substantial economic loss and triggers public panic.
• Early detection of the disease outbreak is critical for effective intervention to contain epidemics.

BACKGROUND
• Center for Disease Control and Prevention (CDC) monitors influenza-like illness (ILI) cases, by collecting data from sentinel medical practices, collating the reports and publishing them on a weekly basis.
• As diagnoses are made and reported by doctors, the system is almost entirely manual, resulting in a 1-2 weeks delay between the time a patient is diagnosed and the moment that data point becomes available in aggregate ILI reports.
• Google Flu Trends uses aggregated search data on flu indicators to estimate current flu activity around the world in real time.

OUR APPROACH
• OSNs like Twitter and Facebook have emerged as popular platform for people to make connections, share information and interact.
• OSN represents a previously untapped data source for detecting onset of epidemics & predicting its spread.
• (I am down with flu, ‘got flu’) messages posted by users provides early, robust predictions.
• User demographics such as age, gender, location, affiliations, etc., can be obtained or inferred from profile data.
• Provides snapshot of existing epidemic situation and preview on what to expect daily or even on hourly bases.

DATA ANALYSIS

DATA CLEANING
• Retweets: A retweet is a post originally made by one user that is forwarded by another user.
• Syndrome elapsed time: To avoid multiple counting of a single episode of illness, we discount subsequent tweets from same user until more than six wee Data show strong correlation (Pearson correlation coefficient 0.8907) between the Twitter data set and ILI rates from CDC, providing a strong base for accurate prediction of ILI.
• Remove all re-tweets and tweets originating from the same user within a syndrome elapsed time, since they do not indicate new ILI cases.

CONCLUSIONS
• Investigated use of messages posted on Twitter to track and predict influenza epidemic situation in the real world.
• Number of flu related tweets are highly correlated with ILI activity in CDC data.
• Build auto-regressive models to predict number of ILI cases as percentage of visits to physicians in successive weeks.
• Tested our regressive models with the historic CDC data and verified that Twitter data effectively improves model’s accuracy in predicting ILI cases.
• Opportunity to significantly enhance public health preparedness among the masses for influenza epidemic and other large scale pandemics.

RESEARCH LINK
http://www.cs.uml.edu/~bliu/SNEFT/

This work was supported in parts by the National Institute of Health under grant IR43LM010766-01 and National Science Foundation under grant CNS-0953620.