Towards Real-Time Measurement of Public Epidemic Awareness

Monitoring Influenza Awareness through Twitter

Michael C. Smith
David A. Broniatowski

The George Washington University

Michael J. Paul

University of Colorado, Boulder

Mark Dredze

The Johns Hopkins University

Goal of Research

- Characterize influenza awareness signal (public concern) to determine which factors most influence disease awareness in a population

  - Influenza incidence rate
  - Infection surveillance using Twitter
  - News media regarding influenza

Agenda

- Introduction / Literature Review:
  - **Why disease awareness?**
  - How disease awareness?
- Data
  - Operationalize!
- Analysis
  - What drives disease awareness?
- Discussion and Conclusions

Why Disease Awareness?

- Studies have shown that the **public's awareness** and their reaction to it **may affect disease spread**
    - Funk et al. (2009); Jones and Salathe (2009); Granell, Gomez, and Arenas (2013)

- Public health officials often manage and monitor disease awareness during epidemics or threats
    - Specific context: influenza

- Problem: no effective method exists for efficient, current awareness surveillance
    - Such systems exist for flu **incidence**

- We don't know what drives awareness!

Agenda

- Introduction / Literature Review:
    - Why disease awareness?
    - **How disease awareness?**
- Data
    - Operationalize!
- Analysis
    - What drives disease awareness?
- Discussion and Conclusions

How Disease Awareness?

- Web and social media data!
- Numerous studies on disease surveillance showing benefits: cheap, real-time
    - Search queries:
        - Google Flu Trends (GFT): Ginsberg et al. (2009)
        - Yuan et al. (2013); Santillana et al. (2014); Preis and Moat (2014)
    - Social media
        - Culotta (2010); Aramaki, Maskawa, and Morita (2011); Lampos and Cristianini (2012)
    - Combining multiple sources - state-of-the-art
        - Santillana et al. (2015)
- Gov’t surveillance systems (tracking hospital visits) may not capture awareness (nor might other systems built to emulate)
- Social media might: people discuss and share concern

How Disease Awareness?

- Previous work separated infection from awareness
    - Lamb, Paul, and Dredze (2013); Broniatowski, Paul, Dredze (2013)
    - “I have the flu” vs. “tired of hearing about the flu”
    - “I’m sick with the flu” vs. “it’s flu season – don’t get sick”
    - “My kids gave me the flu” vs. “wash your hands, don’t get the flu”

Twitter flu prediction

Our current system uses a cascade of 3 MaxEnt classifiers:

- **about health** vs **not about health**
- **about flu** vs **not about flu**
- **flu infection** vs **flu awareness**

Training data:
11,900 labeled tweets collected through MTurk

Estimated weekly flu rate:

$$\frac{\text{\# tweets about flu infection that week}}{\text{\# of all tweets that week}}$$

Slide credit: Paul (2015) - "Making Sense of the Web for Public Health using NLP"

Twitter flu prediction

Features:

- Stylometry
  - Retweets, user mentions, URLs, emoticons
- 8 manually created word classes

Infection	getting, got, recovered, have, having, had, has, catching, catch, cured, infected
Disease	bird, flu, sick, epidemic
Concern	afraid, worried, scared, fear, worry, nervous, dread, dreaded, terrified
Treatment/Prevention	vaccine, vaccines, shot, shots, mist, tamiflu, jab, nasal spray
…	…

Slide credit: Paul (2015) - "Making Sense of the Web for Public Health using NLP"

Twitter flu prediction

Features:

- Part of speech templates
  - (subject,verb,object) tuples
    - always a good feature, IMO
  - numeric references
    - “100 more cases of swine flu”
  - whether “flu” is a noun or adjective
    - “tired of the flu” vs “tired of the flu hype”
  - whether “flu” is the subject or object
    - “I have the flu” vs “the flu is going around”
  - … and others

Slide credit: Paul (2015) - “Making Sense of the Web for Public Health using NLP”

How Disease Awareness?

- Previously focused on social media infection signal
- Now focus on social media awareness signal
  - 2012-2013 flu season

- Our work builds upon this by analyzing awareness
  - News media potential driver, can lead to overestimates in infection
    - '12-'13 GFT

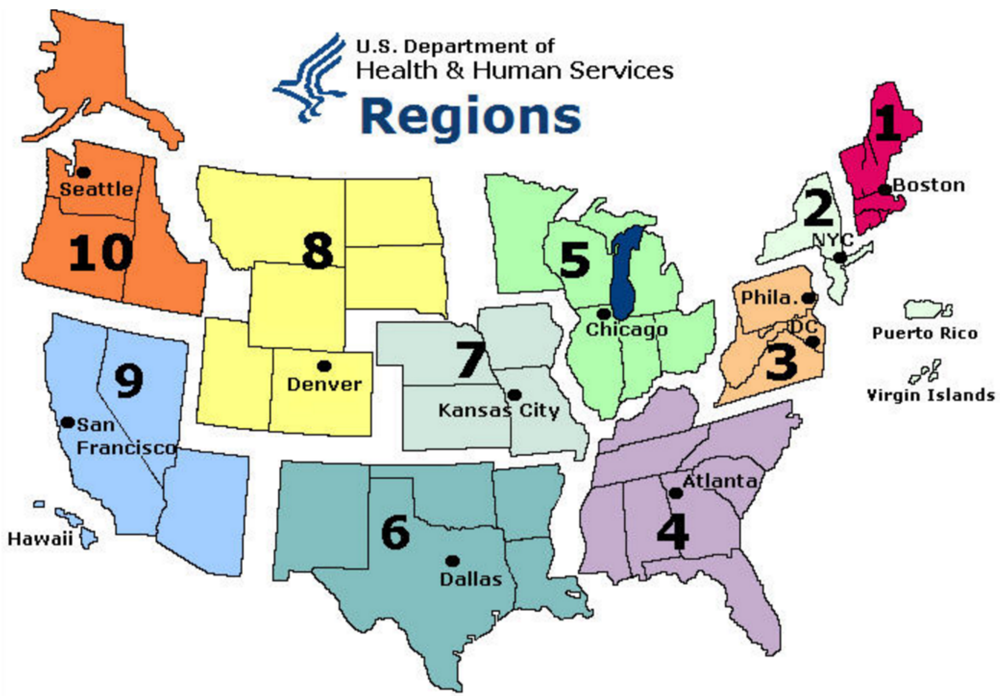
- Studying flu awareness:
  - Yields trends for public health officials
  - Determines drivers for awareness in a population

Agenda

- Introduction / Literature Review:
  - Why disease awareness?
  - How disease awareness?
- **Data**
  - **Operationalize**!
- Analysis
  - What drives disease awareness?
- Discussion and Conclusions

Data

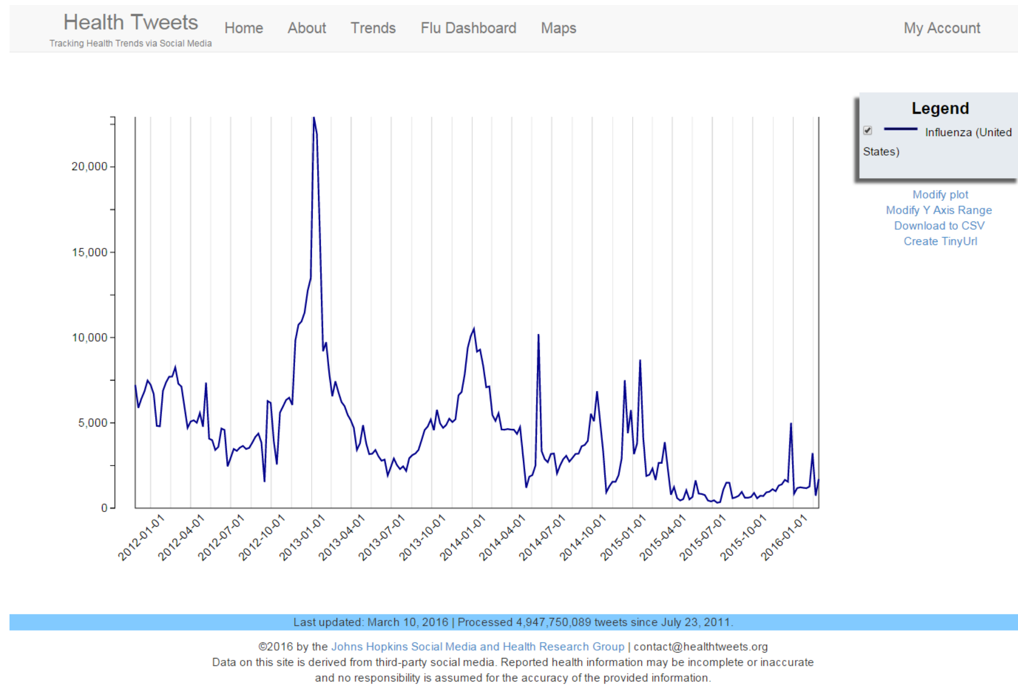
- 2012-2013 flu season
  - September 30, 2012 to May 25, 2013
- National and regional levels



Map of Health and Human Services regions
Source:

Data – Twitter Awareness

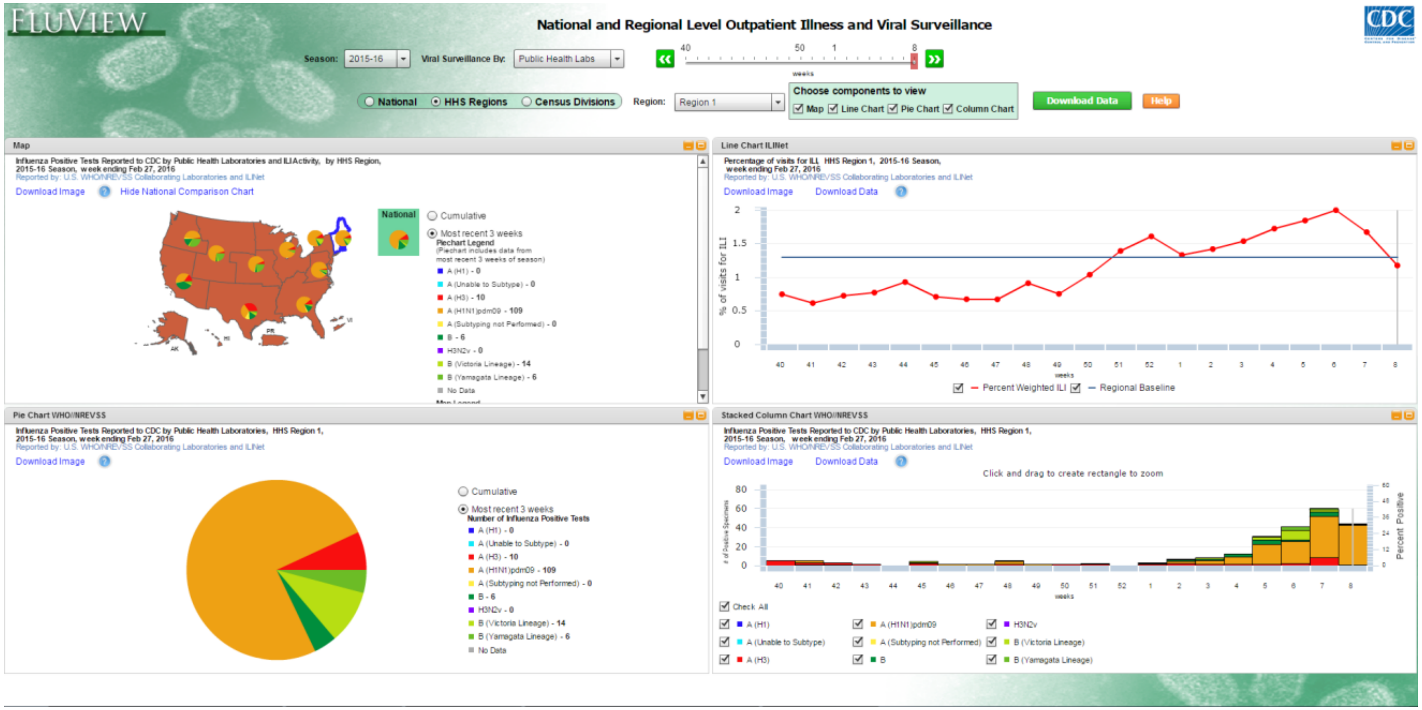
- Healthtweets.org (Dredze et al. 2014)
- Lamb et al. (2013); Broniatowski et al. (2013)
    - Normalized by public tweet counts
    - US tweets, and state level tweets
    - Also downloaded Twitter infection signal



Example trends from HealthTweets.org: weekly United States influenza counts

Data – Government Flu Incidence

- Center for Disease Control and Prevention (CDC)
    - US Outpatient Influenza-like Illness Surveillance Network (ILINet) yields Influenza-like Illness rates (ILI)
- Publicly available on the CDC's flu dashboard:



Screenshot of the CDC's flu dashboard. Source: http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

Data – News Media

Example search of NewsLibrary.com. Source: http://nl.newsbank.com/nl-search/we/Archives/?p_action=customized&s_search_type=customized&p_product=NewsLibrary&p_theme=newslibrary2&d_sources=location&d_place=United%20States&p_nbid=&

Data – Summary

- Twitter awareness
- Infection signals:
  - Twitter infection
  - CDC ILI
- News Media

Agenda

- Introduction / Literature Review:
  - Why disease awareness?
  - How disease awareness?
- Data
  - Operationalize!
- **Analysis**
  - **What drives disease awareness?**
- Discussion and Conclusions

Analysis – Overview

- Compare ILI to Twitter awareness, Twitter infection via Pearson correlations
  - How related are they?

- Plot weekly regional awareness, Twitter infection
  - How similar are the trends?

- Plot weekly awareness, ILI, infection, media
  - What can we glean?

- Regressions: drivers of awareness
  - Infection signals, news media
  - Regional, national

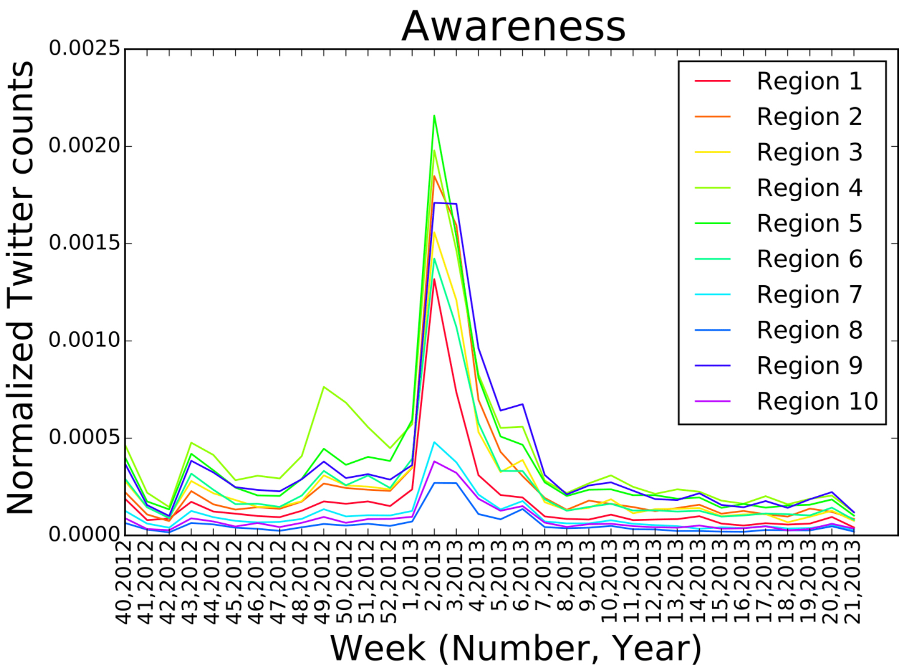
Analysis – Compare to Gold Standard

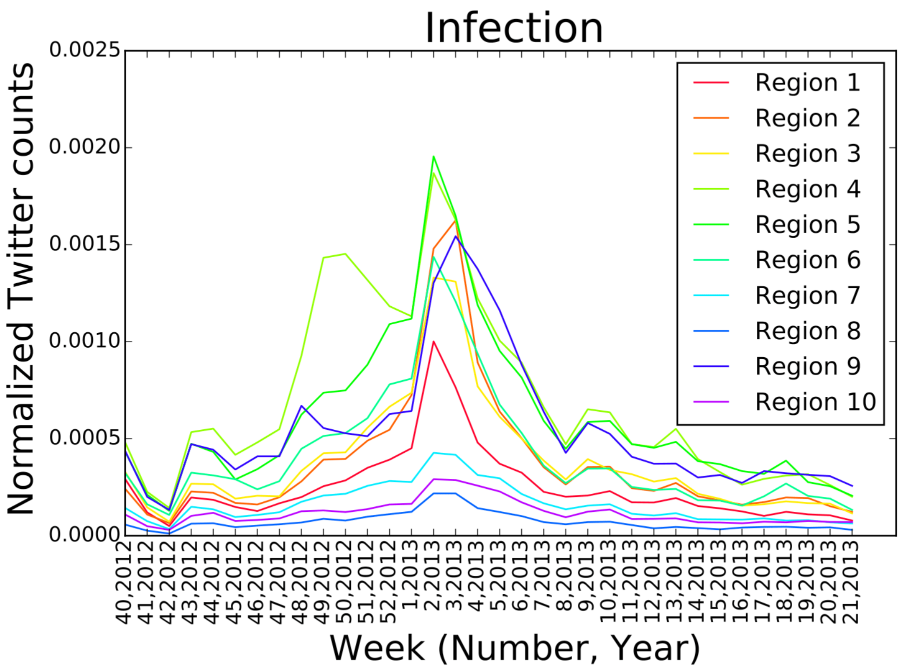
- Compare ILI to both awareness, Twitter infection
- Awareness significantly lower than Twitter infection (p=.029) nationally

Region	Infection	Awareness
1	.802	.588
2	.804	.620
3	.815	.575
4	.812	.489
5	.818	.547
6	.868	.633
7	.885	.626
8	.869	.667
9	.778	.548
10	.846	.658
National	.827	.555

Table 1: Correlations between the Twitter infection and awareness data and the CDC's ILINet influenza prevalence data.

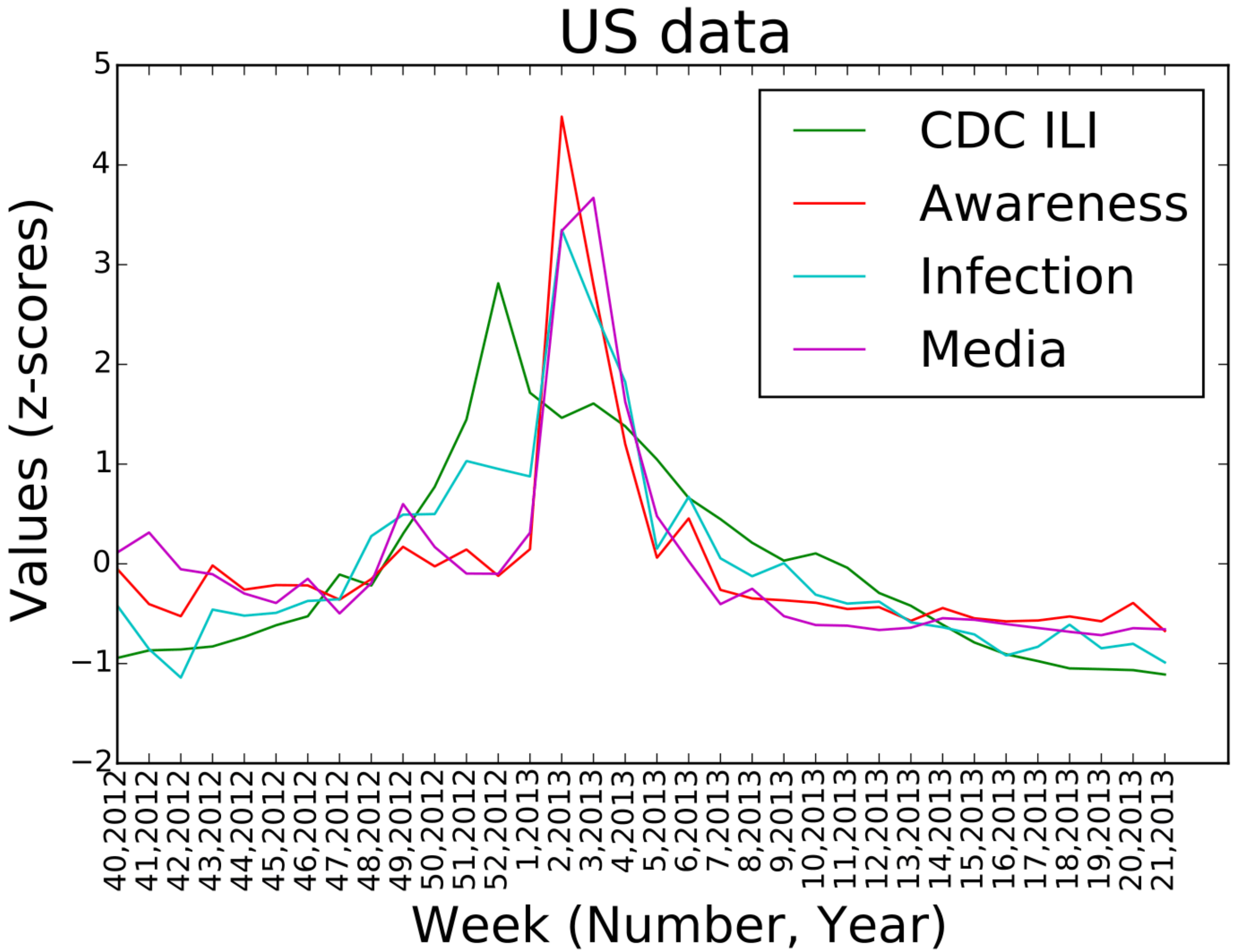
Analysis – Awareness Signal





- Awareness more similar than Twitter infection
- Differences in peak activity, off-peak activity

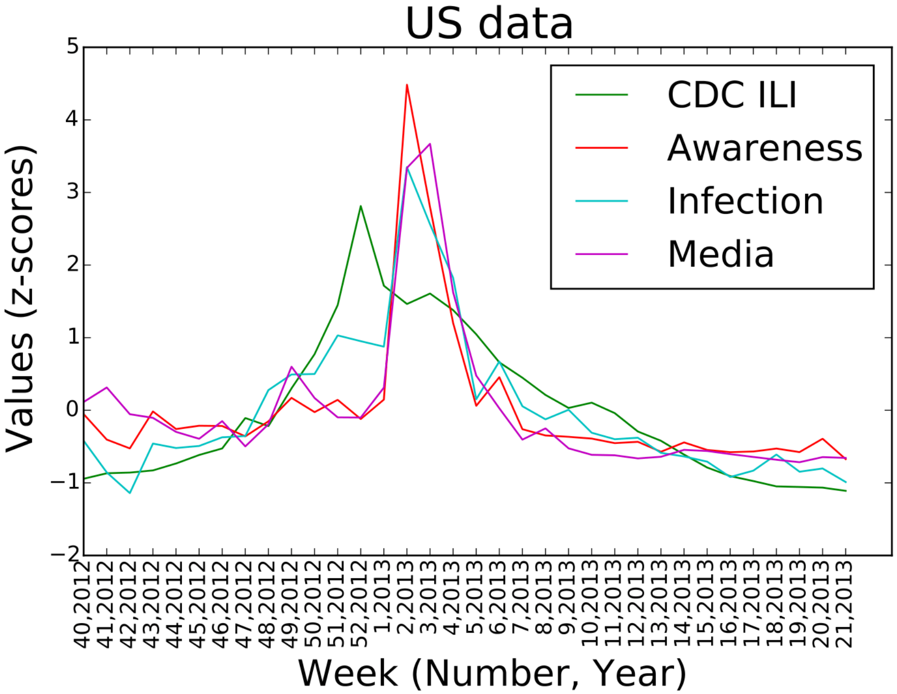
Analysis – Awareness Signal



Analysis – Effect of News Media

Awareness Correlation with:	Mean Regional Correlation	National Correlation:
Media	.905 (SD.044)	.940
Twitter Infection	.907 (SD .026)	.900

- Can we better define the relationship?



Analysis – Effect of News Media

- Bivariate linear regression model at the regional level:

$$awareness_{rw} = \beta_{r0} + \beta_{r1}\text{infection}_{rw} + \beta_{r2}\text{media}_{rw} + \epsilon_{rw}$$

	Twitter Infection		*CDC ILINet*	
Region	**Infection**	**Media**	**Infection**	**Media**
1	.898	.034	.098	.761
2	.484	.514	.143	.873
3	.614	.359	.196	.776
4	.386	.652	.192	.869
5	.547	.431	.073	.847
6	.173	.818	-.003	.978
7	.341	.645	.119	.852
8	.580	.401	.161	.785
9	.490	.531	.186	.834
10	.561	.435	.228	.741
National	.340	.645	.0281	.924

Table 2: Coefficients learned from two bivariate regression models that estimate each week's flu awareness level (as measured from Twitter) as a linear combination of the week's flu infection level and the week's level of media attention (as measured by newspaper volume). The first model uses the Twitter-based estimate of flu infection, while the second model uses the CDC's ILINet estimate.

Analysis – National / Regional Media

- What about national media?
- Similar regression model:
  - Weekly regional awareness as regional media and national media
  - Generated 10 regional models with coefficients for regional media and national media

Mean Regional Media:	Mean National Media:
-.088 (SD .582)	1.026 (SD.561)

Mean regression coefficients across all ten HHS regions

- National media explains much more than regional media!

Agenda

- Introduction / Literature Review:
    - Why disease awareness?
    - How disease awareness?
- Data
    - Operationalize!
- Analysis
    - What drives disease awareness?
- **Discussion and Conclusions**

Discussion

- Distinction between awareness and infection important for flu surveillance
- Awareness more a function of media than infection
    - National media more than regional media
- Little variation in regional awareness
- Awareness does not rise until the flu becomes severe
    - Drops sharply after peak, though infection still high

Conclusions

- Awareness more function of media than infection

- Opportunity: target only certain national distribution channels
    - National media levels contribute more than regional media

- Additional study:
    - Relationship more complex than regression model
    - News media

- Future work needed: generalize to other flu seasons

- Thoughts? Feedback?
    - mikesmith@gwu.edu

References

•Aramaki, Eiji, Sachiko Maskawa, and Mizuki Morita. 2011. “Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter.” In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1568–76. EMNLP ’11. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2145432.2145600.
•Broniatowski, David A., Michael J. Paul, and Mark Dredze. 2013. “National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic.” *PLOS ONE* 8 (12): e83672. doi:10.1371/journal.pone.0083672.
•Culotta, Aron. 2010. “Towards Detecting Influenza Epidemics by Analyzing Twitter Messages.” In *Proceedings of the First Workshop on Social Media Analytics*, 115–22. SOMA ’10. New York, NY, USA: ACM. doi:10.1145/1964858.1964874.
•Dredze, Mark, Renyuan Cheng, Michael J. Paul, and David Broniatowski. 2014. “HealthTweets.org: A Platform for Public Health Surveillance Using Twitter.” In . Quebec City. http://www.cs.jhu.edu/~mdredze/publications/2014_w3phi_healthtweets.pdf.
•Funk, Sebastian, Erez Gilad, Chris Watkins, and Vincent A. A. Jansen. 2009. “The Spread of Awareness and Its Impact on Epidemic Outbreaks.” *Proceedings of the National Academy of Sciences* 106 (16): 6872–77. doi:10.1073/pnas.0810762106.
•Granell, Clara, Sergio Gomez, and Alex Arenas. 2013. “Dynamical Interplay between Awareness and Epidemic Spreading in Multiplex Networks.” *Physical Review Letters* 111 (12). doi:10.1103/PhysRevLett.111.128701.
•Jones, James Holland, and Marcel Salathé. 2009. “Early Assessment of Anxiety and Behavioral Response to Novel Swine-Origin Influenza A(H1N1).” *PLOS ONE* 4 (12): e8032. doi:10.1371/journal.pone.0008032.
•Lamb, Alex, Michael J. Paul, and Mark Dredze. 2013. “Separating Fact from Fear: Tracking Flu Infections on Twitter.” In *HLT-NAACL*, 789–95. http://www.aclweb.org/anthology/N/N13/N13-1097.pdf.
•Lampos, Vasileios, and Nello Cristianini. 2012. “Nowcasting Events from the Social Web with Statistical Learning.” *ACM Trans. Intell. Syst. Technol.* 3 (4): 72:1–72:22. doi:10.1145/2337542.2337557.
•Paul, Michael J. 2015. “Making Sense of the Web for Public Health Using NLP.” National Cancer Institute, January 21. http://cmci.colorado.edu/~mpaul/files/slides_nih_1-21-15.pdf.
•Preis, Tobias, and Helen Susannah Moat. 2014. “Adaptive Nowcasting of Influenza Outbreaks Using Google Searches.” *Royal Society Open Science* 1 (2): 140095. doi:10.1098/rsos.140095.
•Santillana, Mauricio, André T. Nguyen, Mark Dredze, Michael J. Paul, Elaine O. Nsoesie, and John S. Brownstein. 2015. “Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance.” *PLOS Comput Biol* 11 (10): e1004513. doi:10.1371/journal.pcbi.1004513.
•Santillana, Mauricio, D. Wendong Zhang, Benjamin M. Althouse, and John W. Ayers. 2014. “What Can Digital Disease Detection Learn from (an External Revision To) Google Flu Trends?” *American Journal of Preventive Medicine* 47 (3): 341–47. doi:10.1016/j.amepre.2014.05.020.
•Yuan, Qingyu, Elaine O. Nsoesie, Benfu Lv, Geng Peng, Rumi Chunara, and John S. Brownstein. 2013. “Monitoring Influenza Epidemics in China with Search Query from Baidu.” *PLOS ONE* 8 (5): e64323. doi:10.1371/journal.pone.0064323.