EpiDash 1.0 - A GI Case Study

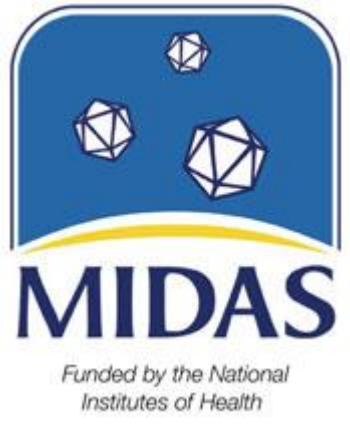
Speaker: Elizabeth Musser
Graduate Research Assistant Virginia Bioinformatics Institute

In collaboration with James Schlitt, Harshal Hayatnagarkar, P. Alexander Telionis, Meredith Wilson MPH, Caitlin Rivers MPH, Dr. Bryan Lewis MPH Ph.D. and Dr. Stephen Eubank Ph.D.

Disclosure

No relationships to disclose

This work was supported by the NIH MIDAS Grant (# 2U01GM070694-09)



Motivations EpiDash 1.0

1) Provide useful epidemiological context for incidence of GI illness within a community, assist in outbreak investigation and contribute in the identification of risk factors associated with disease occurrence
2) Foster awareness of GI illness occurrence within a community and aid in outbreak detection
3) Supplement and compliment conventional surveillance systems with social media and social networking data
4) Detect trends signaling changes in the incidence and prevalence of GI syndrome illness within a health district
5) Provide estimates of the magnitude of morbidity of syndromic illness within a community
6) Contribute significantly to field epidemiologic research directly correlated with the control or prevention of GI syndrome illness.

Challenges for Syndromic surveillance

"Population-level scanning"

The collection and analysis of health data about a clinical syndrome that has a significant impact on public health, which is then used to drive decisions about health policy and health education. The term applies to surveillance of populations and is distinct from active surveillance, which applies to individuals.

Currently consist of lab reports, pharmacy prescription data and doctors visits visualized through systems such as Biosense and Essence.

Challenges for Syndromic surveillance

Limited in scope, missing data, delayed outbreak identification

Limited Health Situational Awareness

"There's the potential for us to idenitfy outbreaks of norovirus much earlier than before, giving us the opportunity to proactively share our advice and guidance with those who might be affected, alert other government departments and industry, and perhaps even help to reduce its spread"

James Baker, FSA

Challenges for Syndromic surveillance

BOX 1. Tasks for evaluating public health surveillance systems for early detection of outbreaks*

Task A. Describe the system
1. Purpose: What is the system designed to accomplish?
2. Stakeholders: Whom does the system serve?
3. Operation: How does the system work?
   a. Systemwide processes
   b. Data sources
   c. Data preprocessing
   d. Statistical analysis
   e. Epidemiologic analysis, interpretation, and investigation

Task B. Provide data demonstrating outbreak detection attributes
1. Timeliness: How early in the outbreak is the event detected?
2. Validity: How well does the system perform in distinguishing outbreak detection of public health significance from less important events or random variations in disease trends?
   a. Sensitivity and predictive value: What percentage of true outbreaks are detected by the system? What percentage of signals by the system are relevant (true positives)? What percentage of negative results are truly negative?
   b. Data quality: How does data quality affect validity of outbreak detection?
      i. Representativeness: How well does the system reflect the population of interest?
      ii. Completeness: What percentage of data are present for each record?

Task C. Describe the system experience
1. System usefulness: In what ways has the system demonstrated value relevant to public health?
2. Flexibility: How adaptable is the system to changing needs and risk thresholds?
3. System acceptability: Have stakeholders been willing to contribute to and use the system?
4. Portability: How readily can the system be duplicated at another location?
5. System stability: How consistent has the system been in providing access to reproducible results?
6. System costs: What are the resource requirements to deploy and maintain the system?

Task D. Summarize conclusions and make recommendations for use and improvement of systems for early outbreak detection

* Source: CDC. Framework for evaluating public health surveillance systems for early detection of outbreaks: recommendations from the CDC working group. MMWR 2004;53(No. RR-5).

Case Definition

The surveillance system is designed to captured statements relevant to bacterial and viral acute gastroenteritis. A case of acute gastroenteritis is defined as a person with diarrhea and/or vomiting and/or abdominal cramps due to either viral or bacterial infection. Diarrhea is defined as two or more loose stools per day or an unexplained increase in the number of bowel movements.

The field epidemiologist….

- True challenge for these tools lies in the successful integration
- Within the context of the public health department challenges are numerous
- Numerous tools have been developed with little feedback, formal evaluation or long term usage
- Tools should evaluated within a well developed framework of procedures and support designed to assist in Epidemiologic analysis, interpretation, and investigation in response to a system signal

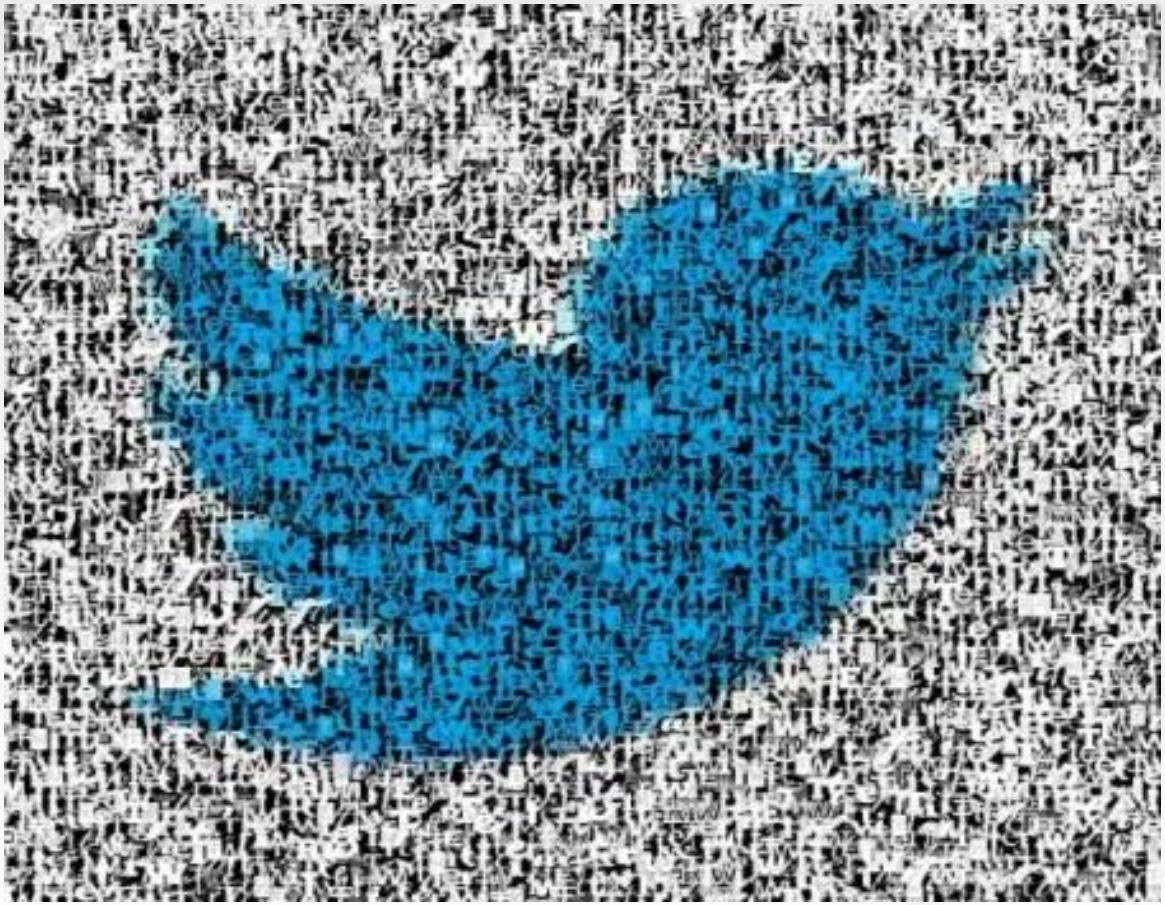
Overview of the Surveillance System

1. Twitter statement occurs in given health district

2. Identified by the ChatterGrabber data miner, tweeters are used as a human sensor network to optimize sensitivity especially in rural areas

3. *S^(u)* is used to make an interference of the P (E^(u)) Where S is the data incoming to the dashboard from the social network and P is the current prevalence of GI syndrome activity. A health statement is relayed to health officials through EpiDash 1.0 Dashboard with epidemiological context for analysis of the data, dissemination of information, and public health action.

Data Collection



Data Collection

- **ChatterGrabber**: A search method based social media data miner developed in Python.
  - GDI Google Docs interface included for simplified partner access.
  - Alternative to free 1% streaming;
  - Specialized hunters pull from GDI Spreadsheets to set run parameters.
  - Multiple logins may be used to increase search frequency during collaborative experiments.
  - No limits on query length.
  - Data sent nightly to dashboard

ChatterGrabber Search Methods

Pure Query Based:

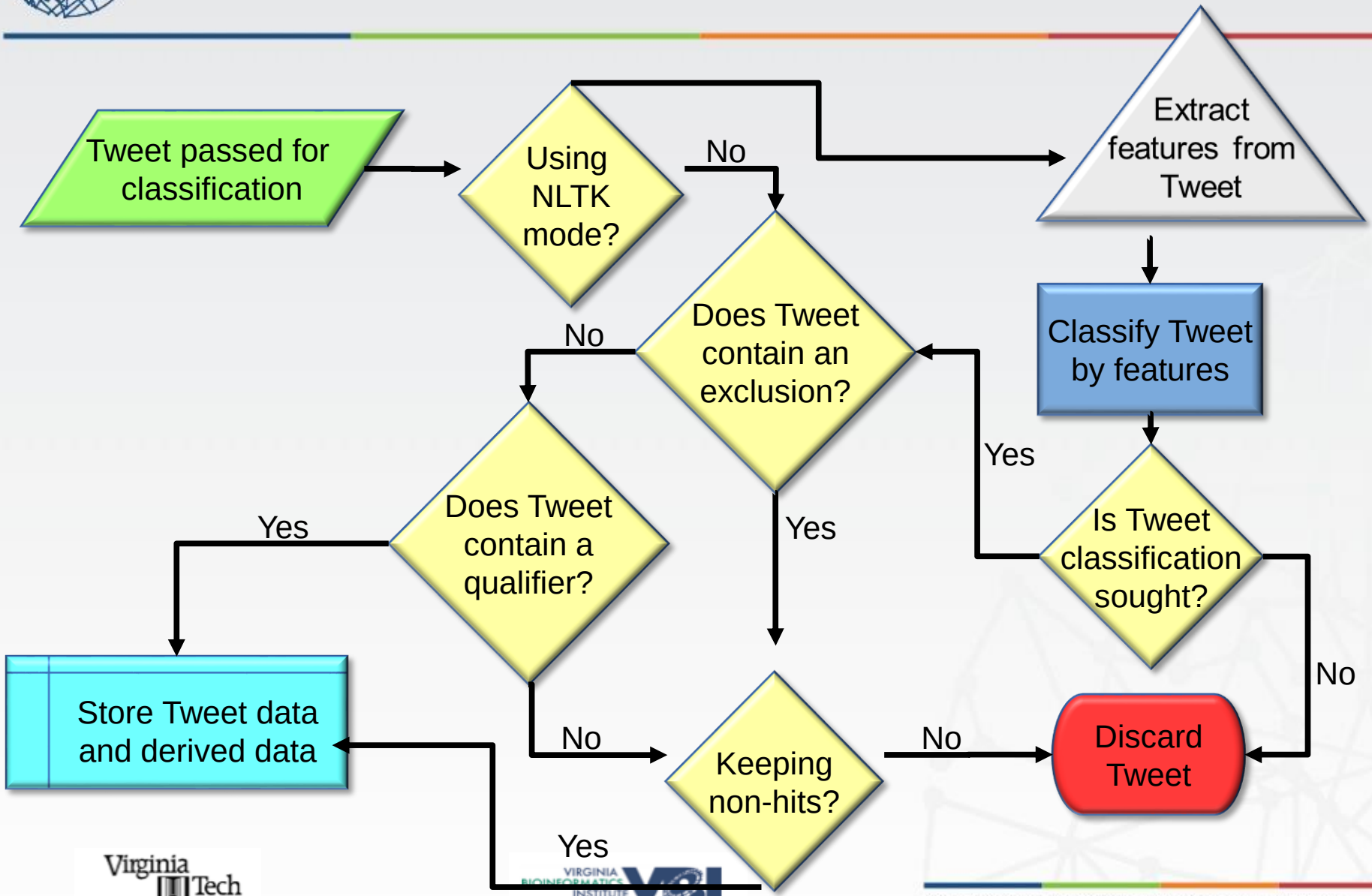
- Conditions, qualifiers, & exclusions.
- Searches by conditions, keeps if qualifier and no exclusions present.
- Simple, easy to setup, but vulnerable to complexities of wording.

NLTK* Based:

- Take output from conditions search, manually classify.
- Train NLTK maxEnt or Naïve Bayesian classifier via content n-grams.
- Classifier discards tweets that don't fit desired categories.
- Powerful, but requires longer setup, representative tweet sample.

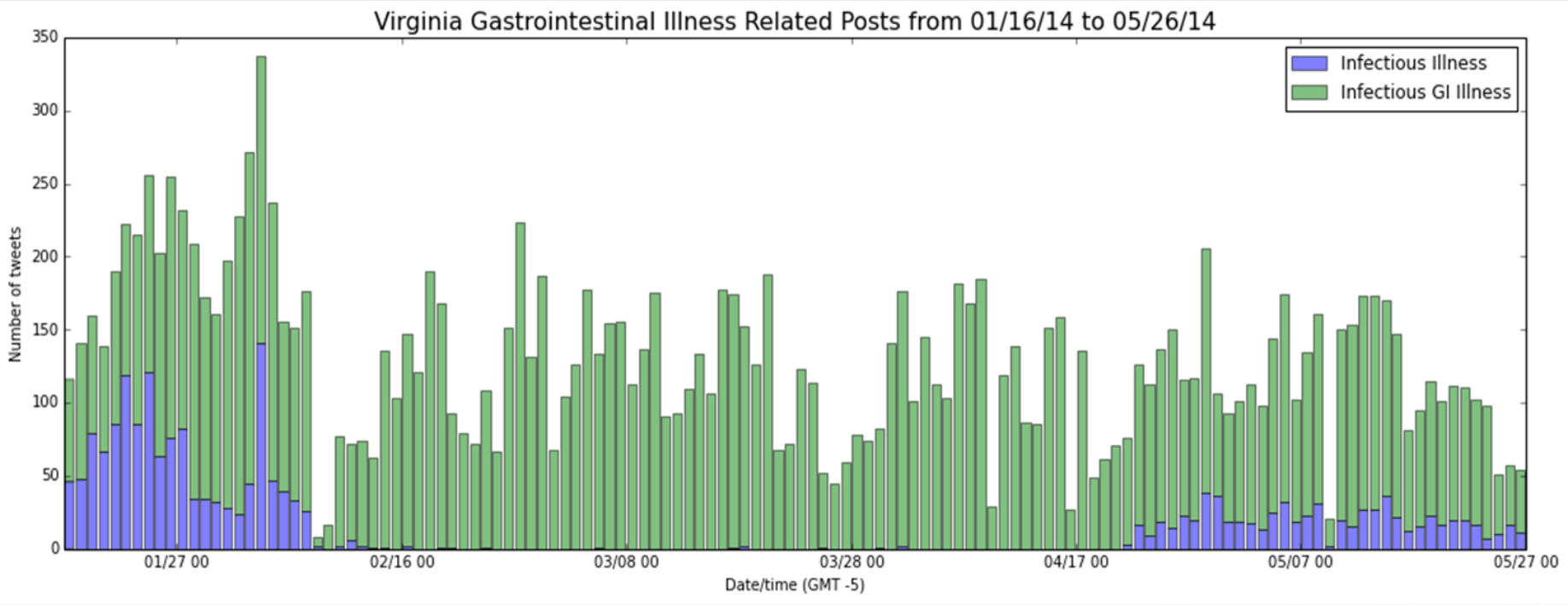
NLTK: Natural Language Tool Kit

Tweet Linguistic Classification



Tweet Linguistic Classification

```
Please enter a test sentence:   just threw up #tgif #yolo #neverdrinkingagain
Query: just threw up #tgif #yolo #neverdrinkingagain
Result: no suspicion of infectious illness
Please enter a test sentence:   my son just threw up
Query: my son just threw up
Result: suspicion of infectious GI illness
Please enter a test sentence:   can't believe I ate so much, just threw up
Query: can't believe I ate so much, just threw up
Result: no suspicion of infectious illness
Please enter a test sentence:   oh god I feel so sick
Query: oh god I feel so sick
Result: suspicion of infectious illness, type unknown
```



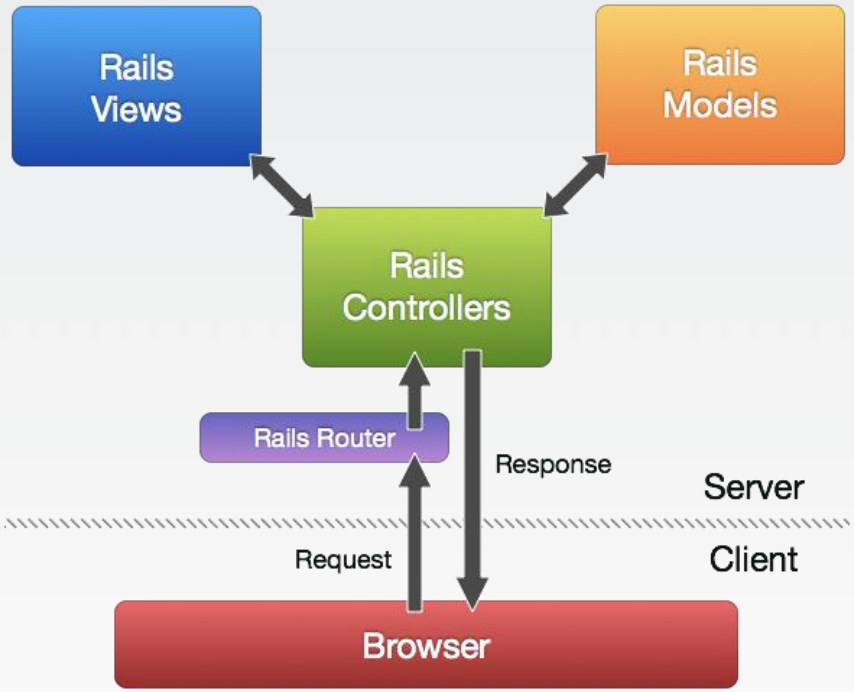
- 74-86% accuracy with 2,000 tweet training set
- 14762 GI illness and 2075 other illness related posts.

Event-Based Surveillance System Attributes

Attributes of the dashboard development include:

- Timeliness of Data Input
- Reporting Structure
- Timeliness of Detection
- Thresholds for Signal Generation
- Trigger for Dissemination and Analysis

EpiDash 1.0 Architecture



- Ruby on Rails
  - Open source web framework
  - Ruby language
  - Model-view-controller
- MVC
  - Model talks to database.
  - View talks to browser.
  - Controller coordinates between model and view.

EpiDash 1.0 System Attributes

First Look Section:

EpiDash 1.0 System Attributes

First Look Section Analytics:

- Variables to Account for increasing and decreasing levels:
  c= constant level of disease activity (background noise)
- d= day of week variation
- s= seasonal variation

Levels to Account for:

- Red, Yellow, Green.
- Weekly Mean with Standard Deviation:
- The mean will consistent of the corresponding days in the five previous weeks to account for just over a month of data. The standard deviation for the previous five
-

EpiDash 1.0 System Attributes

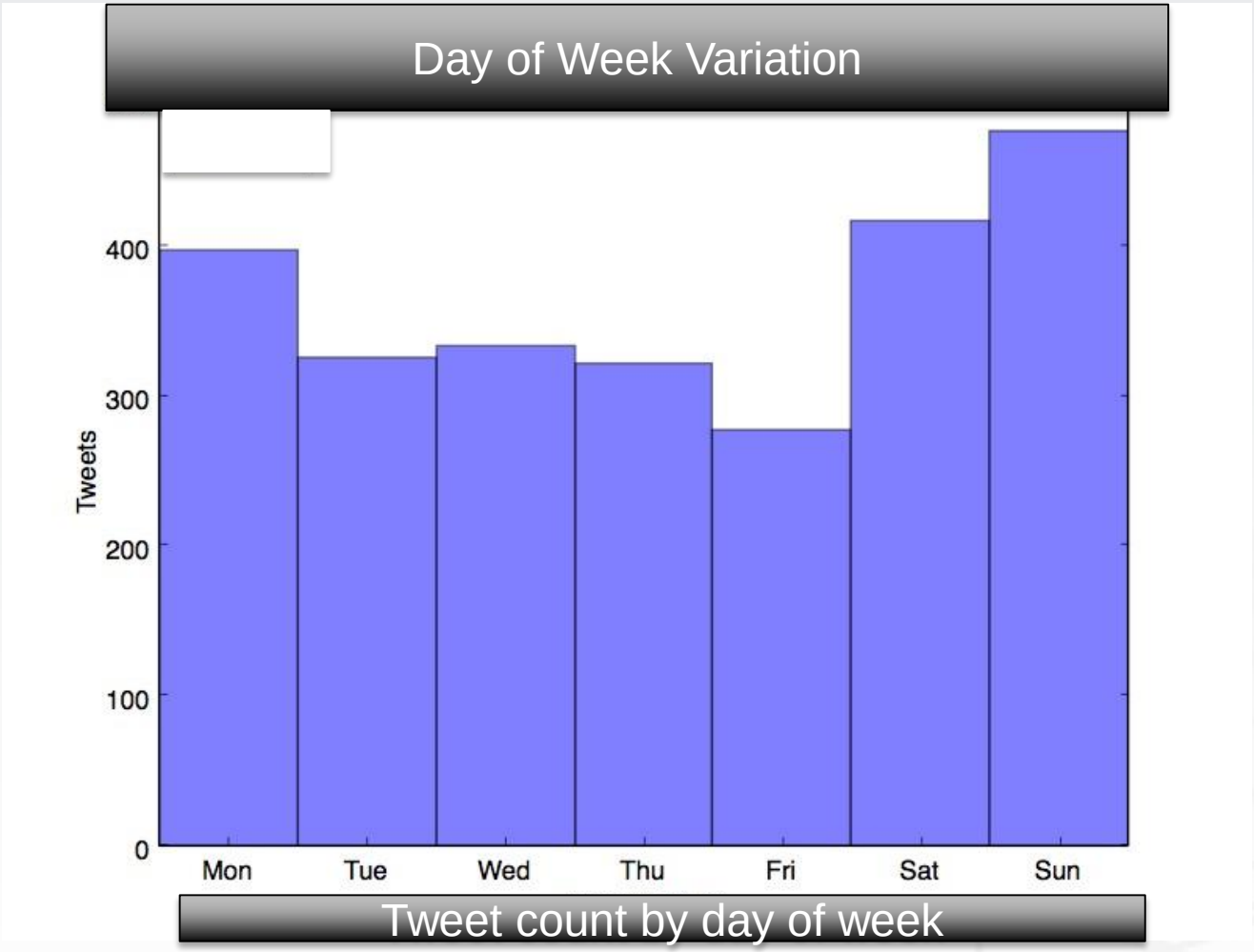
Standard Deviation Scales:

Green: The current days value is no more than one standard deviation from the mean of the previous 5 weeks values.

Yellow: The current days value is greater than one standard deviation from the mean of the previous 5 weeks values but less than two standard deviations from the mean.

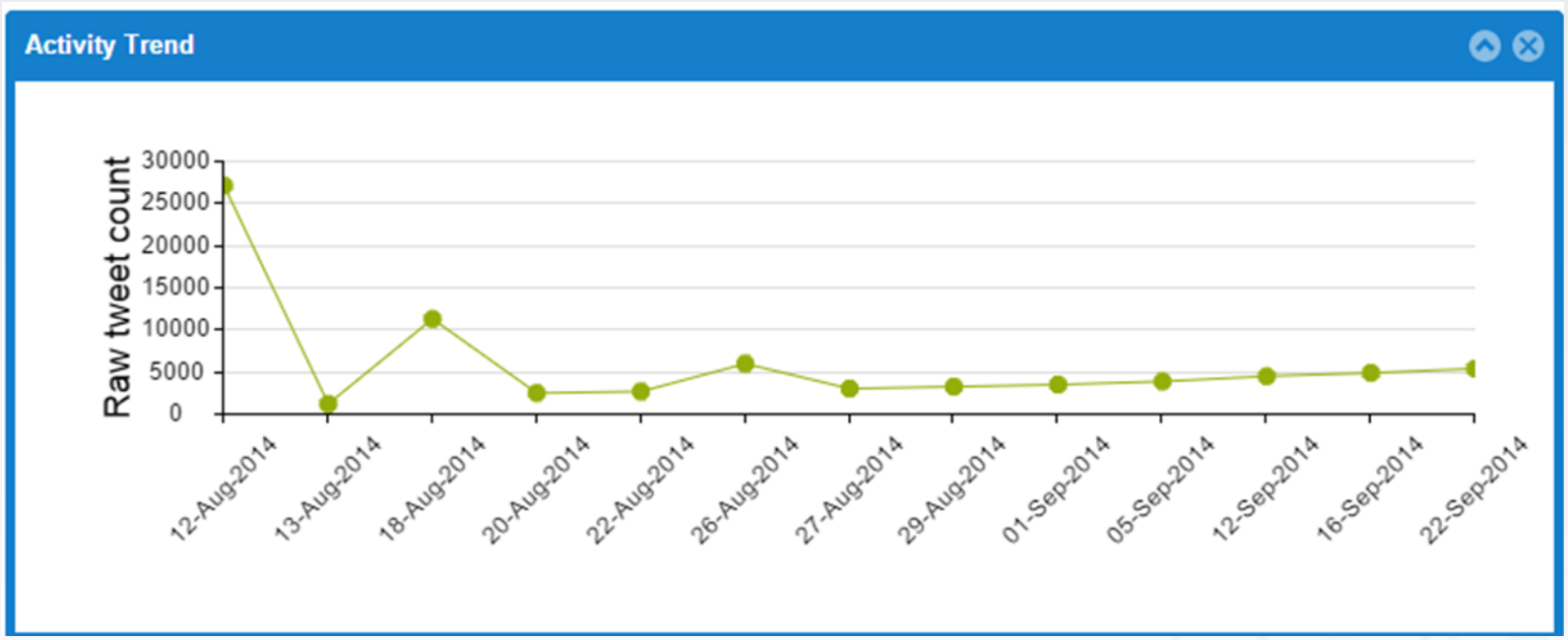
Red: The current days value is two standard deviations or greater from the mean of the previous 5 weeks values.

EpiDash 1.0 System Attributes



EpiDash 1.0 System Attributes

Time Series Weekly Overview:



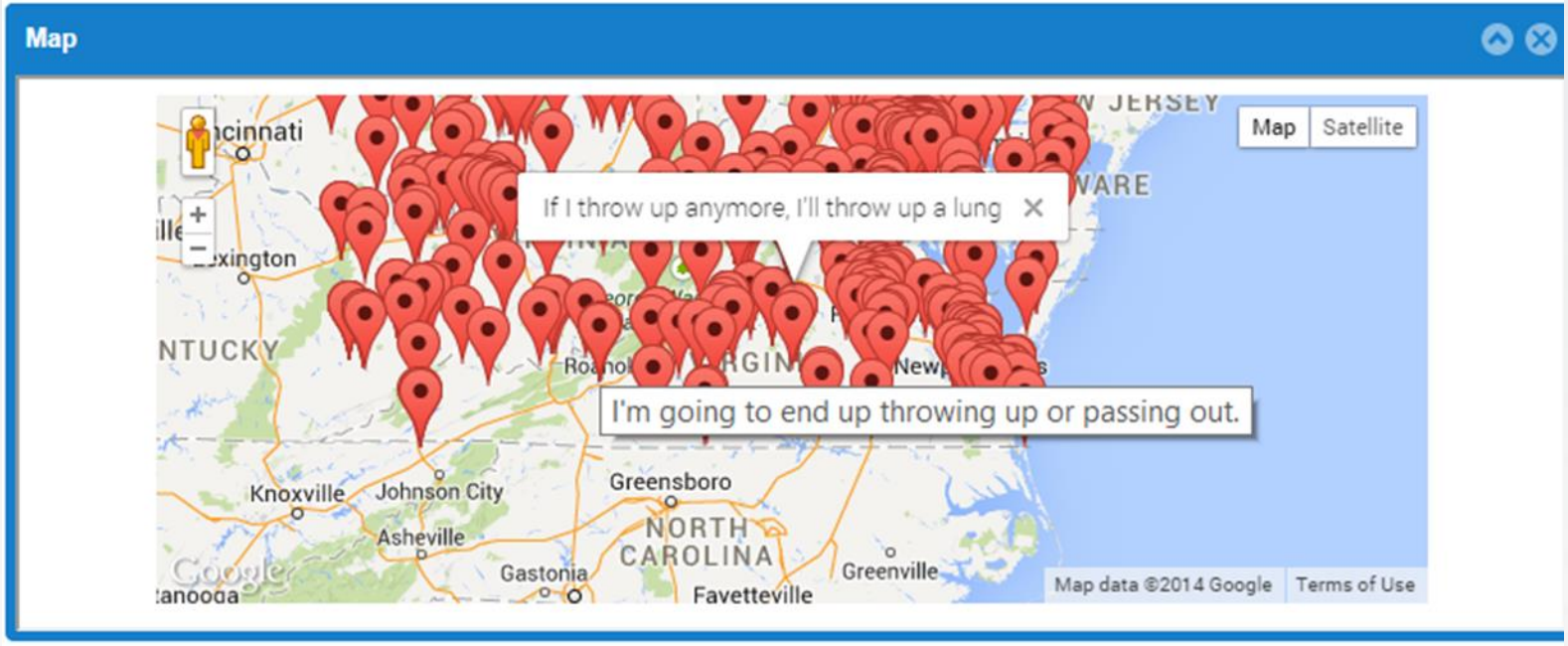
EpiDash 1.0 System Attributes

Time Series Analytics:

-Seasonal Trending Identification
-Keyword trending for Event based surveillance

EpiDash 1.0 System Attributes

GIS Map:

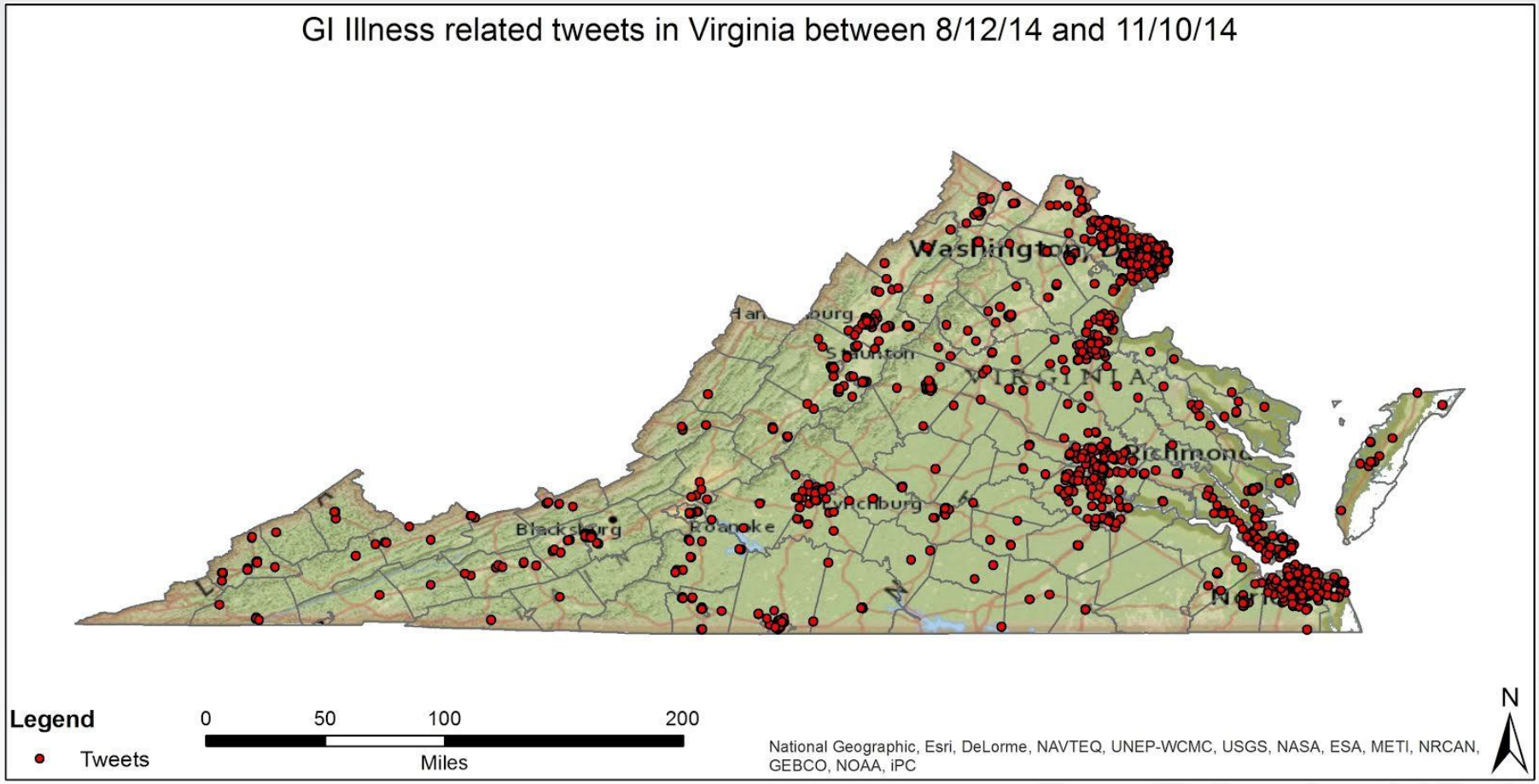


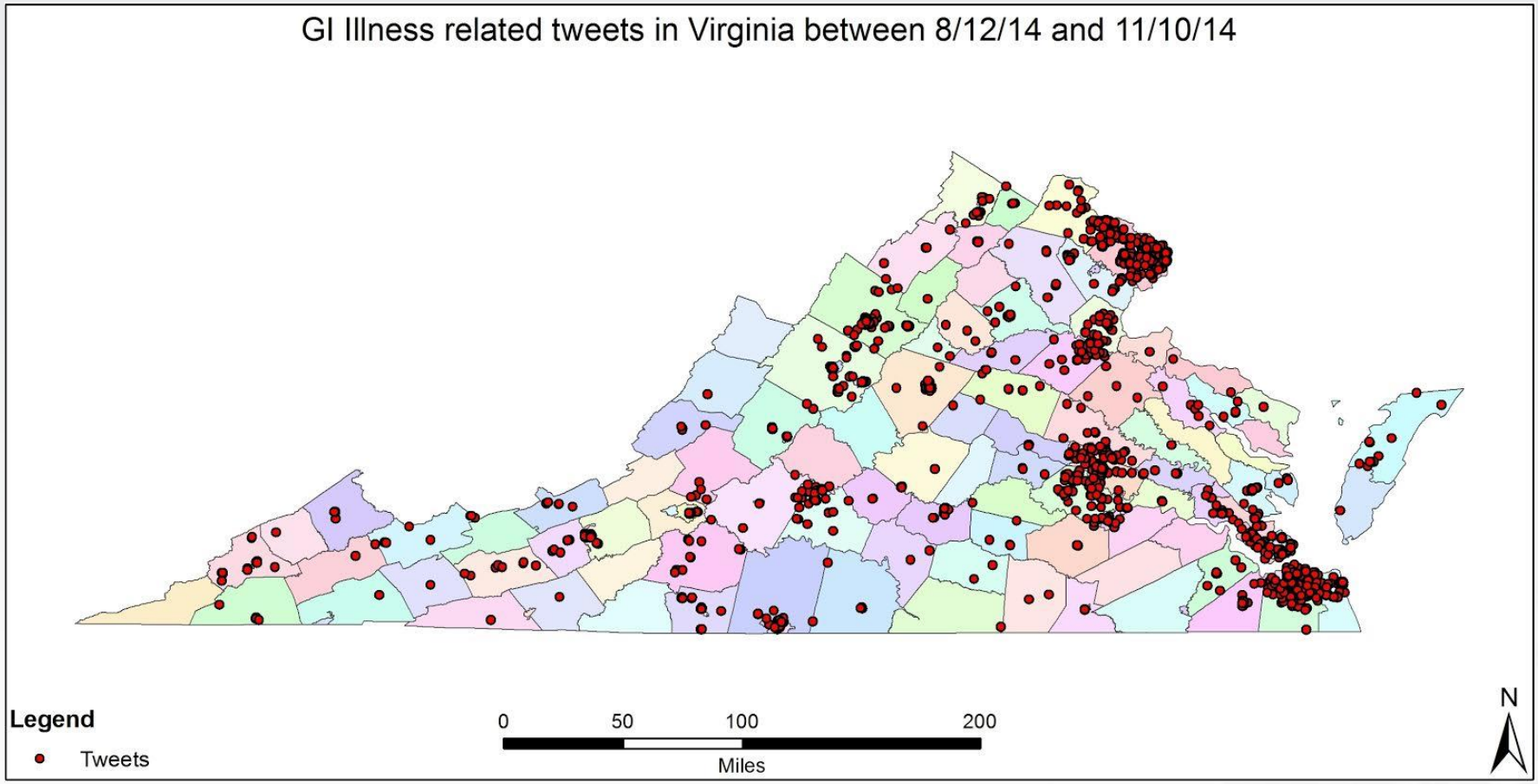
EpiDash 1.0 System Attributes

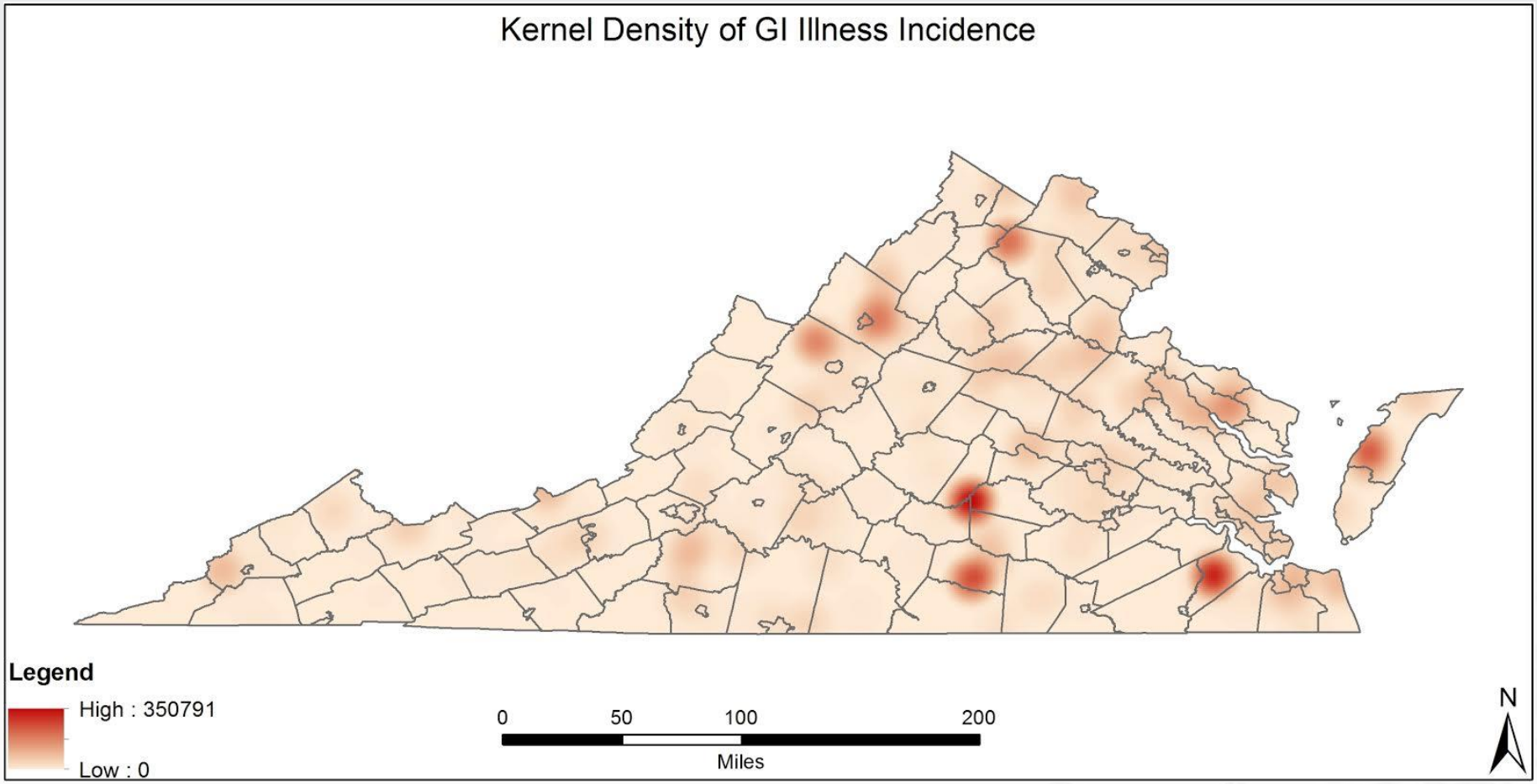
GIS Map Analytics:

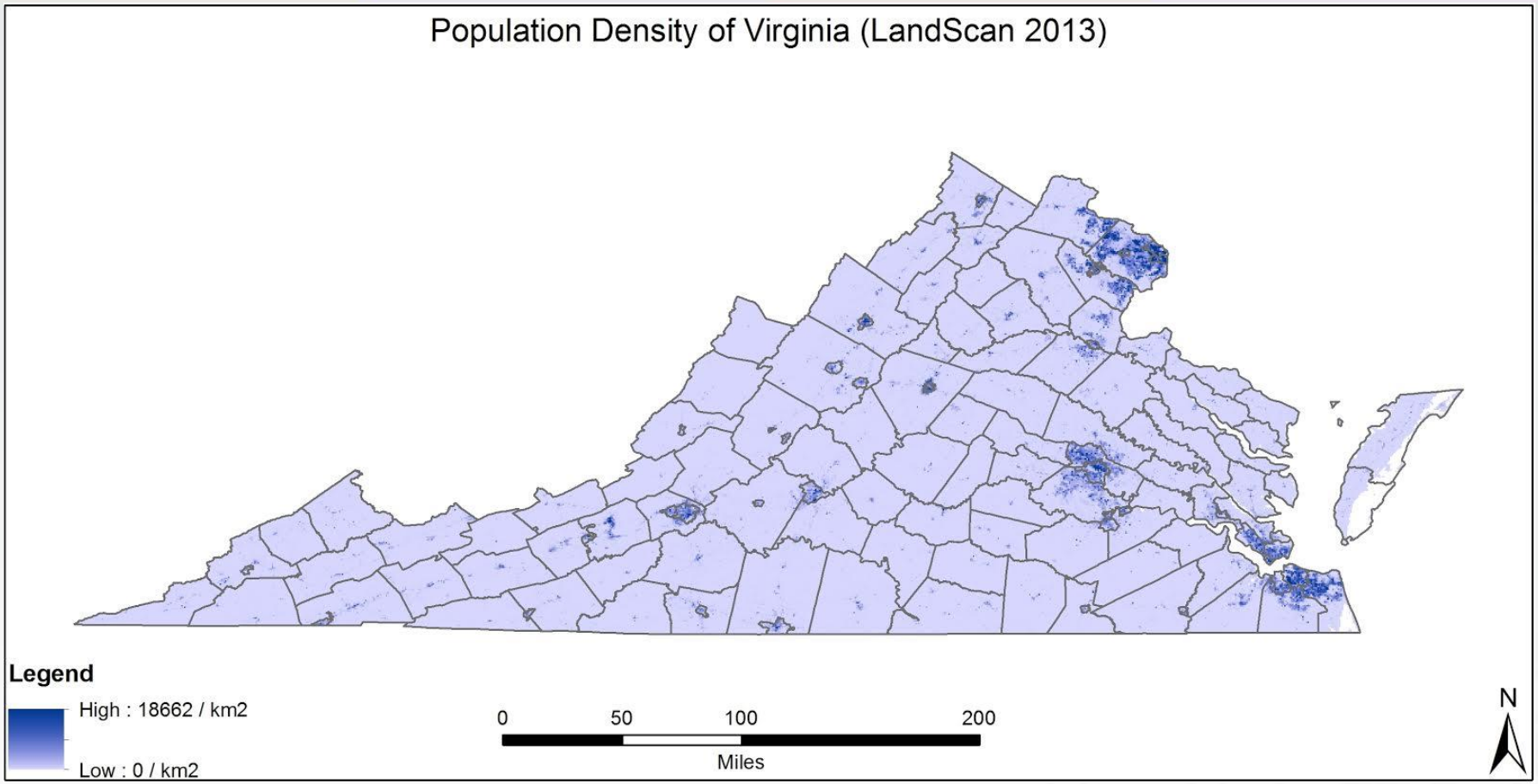
-Cluster Analysis

-Identification of hotspots for illness (eating establishments, venues, events etc.)


GI Illness related tweets in Virginia between 8/12/14 and 11/10/14
Legend
Tweets
0
50
100
200
Miles
N
National Geographic, Esri, DeLorme, NAVTEQ, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, iPC


GI Illness related tweets in Virginia between 8/12/14 and 11/10/14
Legend
Tweets
0
50
100
200
Miles
N


Kernel Density of GI Illness Incidence
Legend
High : 350791
Low : 0
0
50
100
200
Miles
N

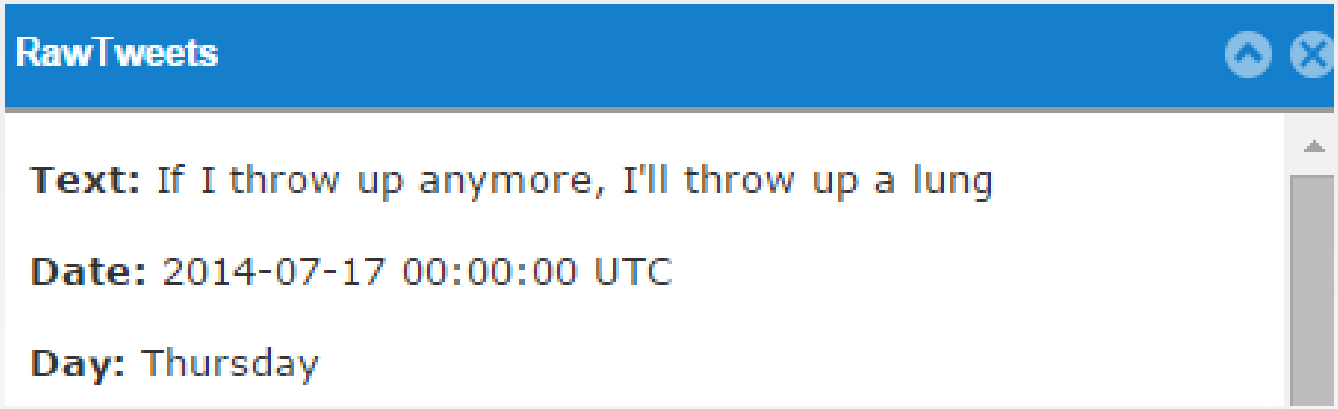

Population Density of Virginia (LandScan 2013)
Legend
High : 18662 / km2
Low : 0 / km2
0
50
100
200
Miles
N

EpiDash 1.0 System Attributes

Word Cloud:

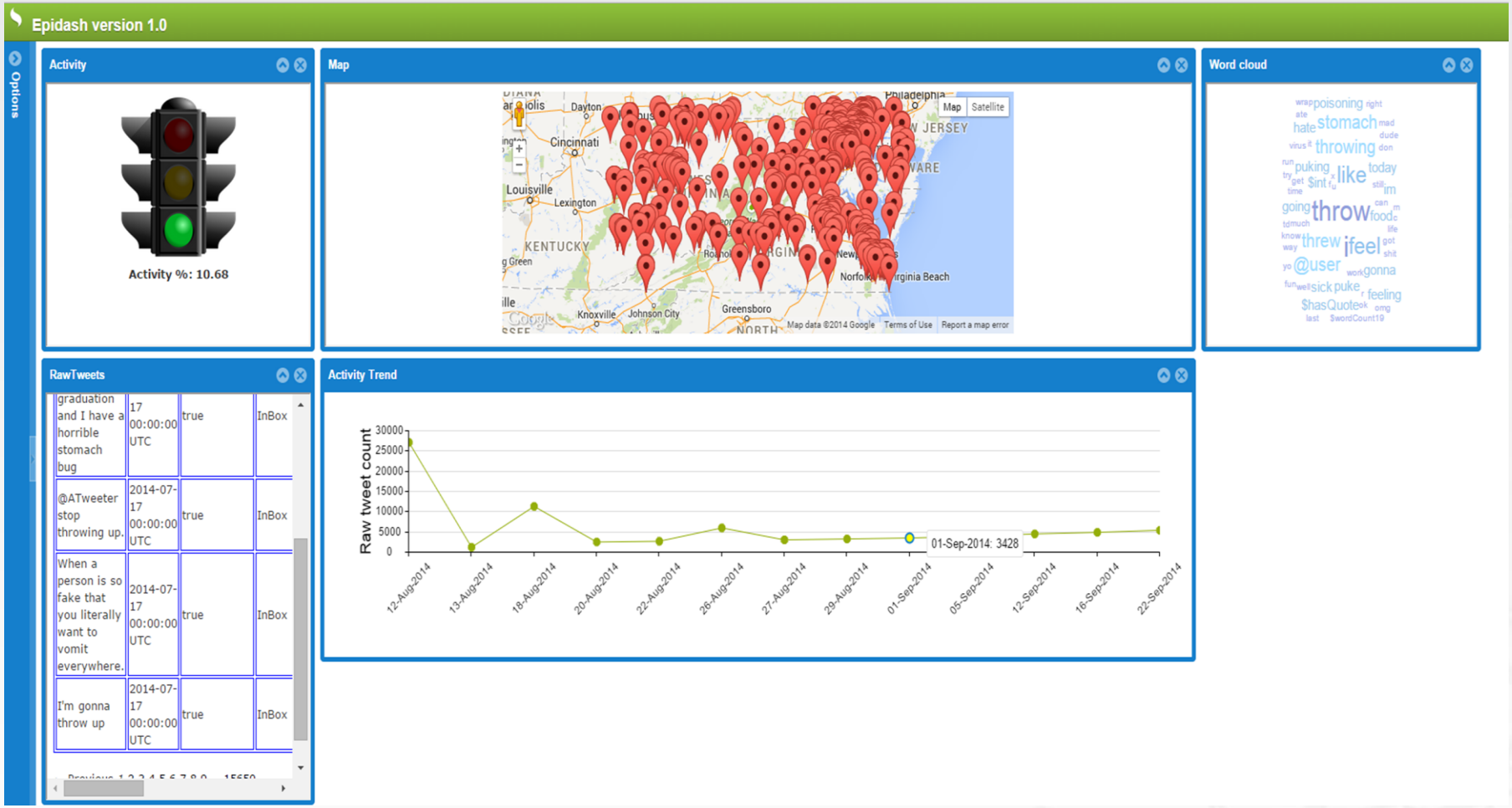
EpiDash 1.0 System Attributes

Searchable Keyword Matching and Raw Data:



- Keyword pair matching, linguistical pattern recognition and matching

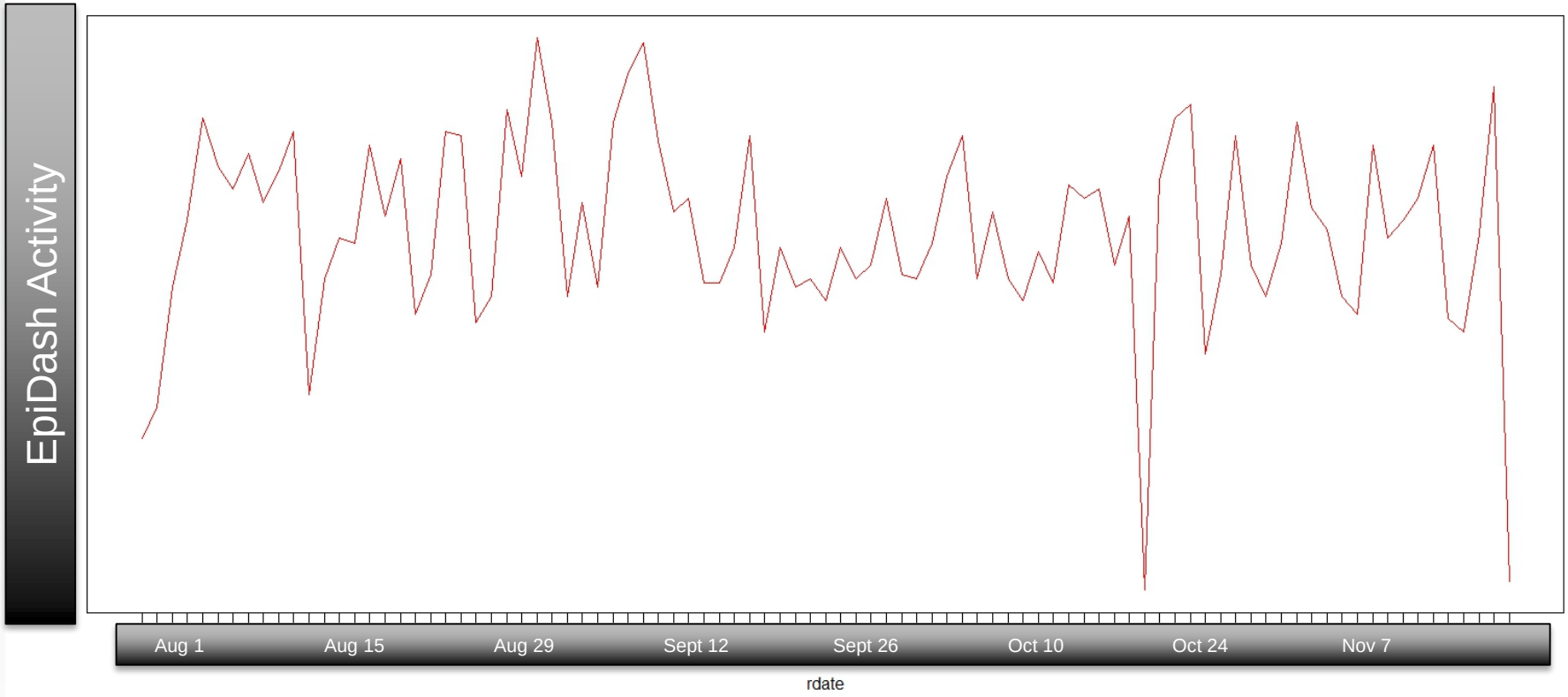
EpiDash 1.0


Epidash version 1.0
Options
Activity
Activity %: 10.68
Map
Map
Satellite
Cincinnati
Louisville
Lexington
KENTUCKY
Knoxville
Johnson City
Greensboro
Virginia Beach
Map data ©2014 Google
Terms of Use
Report a map error
Word cloud
poisoning
stomach
throwing
puking
like
today
throw
feel
@user
gonna
sick
puke
feeling
RawTweets
graduation and I have a horrible stomach bug
17 00:00:00 UTC
true
InBox
@ATweeter stop throwing up.
2014-07-17 00:00:00 UTC
true
InBox
When a person is so fake that you literally want to vomit everywhere.
2014-07-17 00:00:00 UTC
true
InBox
I'm gonna throw up
2014-07-17 00:00:00 UTC
true
InBox
Activity Trend
Raw tweet count
30000
25000
20000
15000
10000
5000
0
01-Sep-2014: 3428
12-Aug-2014
13-Aug-2014
18-Aug-2014
20-Aug-2014
22-Aug-2014
26-Aug-2014
27-Aug-2014
29-Aug-2014
01-Sep-2014
05-Sep-2014
12-Sep-2014
16-Sep-2014
22-Sep-2014

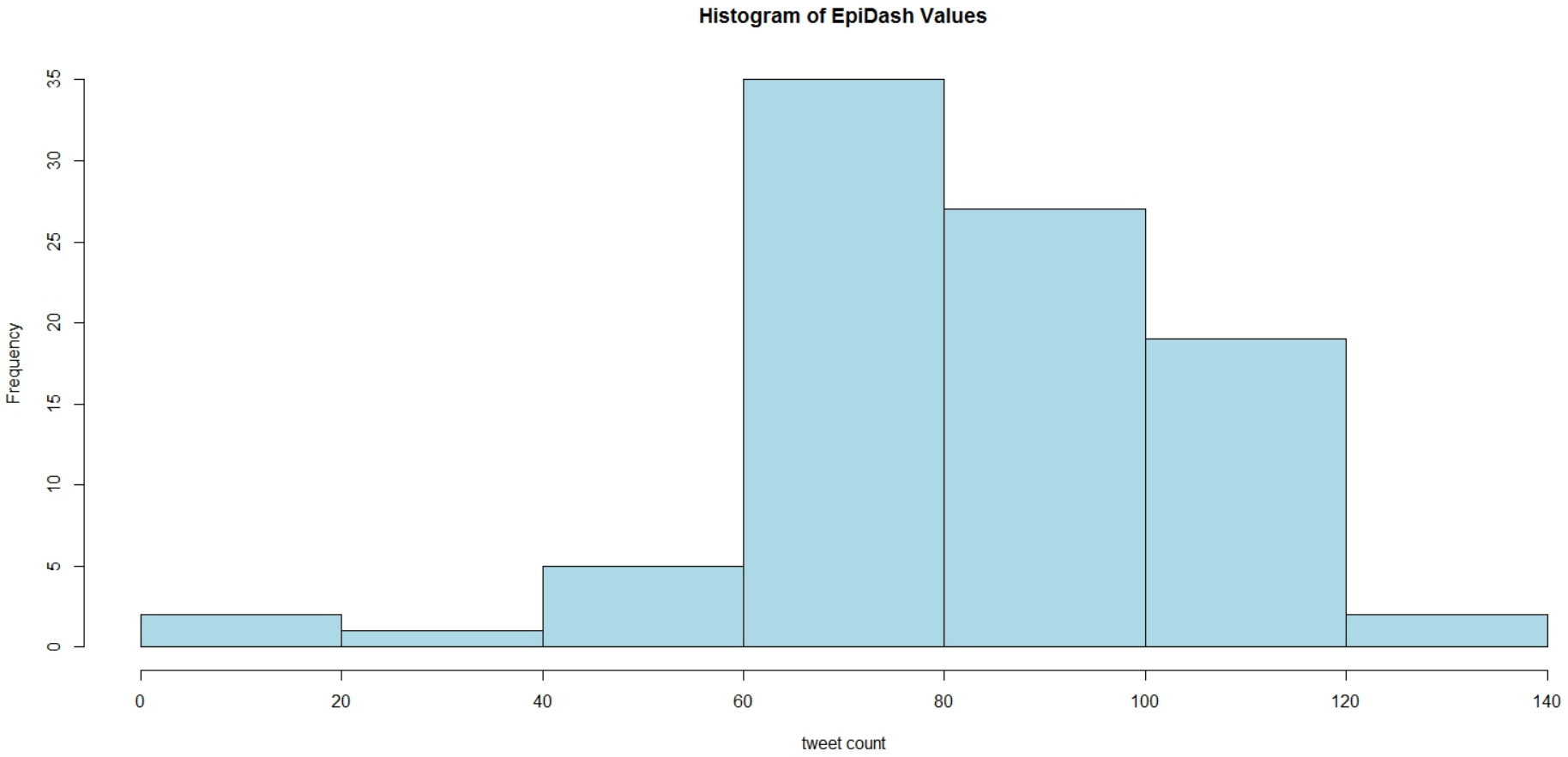
Pre-Deployment

- Pre-Surveillance Intake Assessment:
- VDH state district epidemiologists
  - Average 7-9 hours working on syndromic surveillance
  - "knowledgeable" but with limited "technical skills"
  - On a scale from 1-10 syndromic surveillance was described as "10" or highly pertinent.
  - Estimated interactions between user and interface was 3-5 times per week
  - Difficulty of usage, system portability and lack of data completeness were cited most frequent as reasons surveillance systems failed to integrate in standard surveillance protocol.
  - **All users identified social media as a critical part of current syndromic surveillance**

Deployment Activity



Deployment Activity



86% accuracy and is cross validated in training set

Deployment Support

-Dashboard User Guide

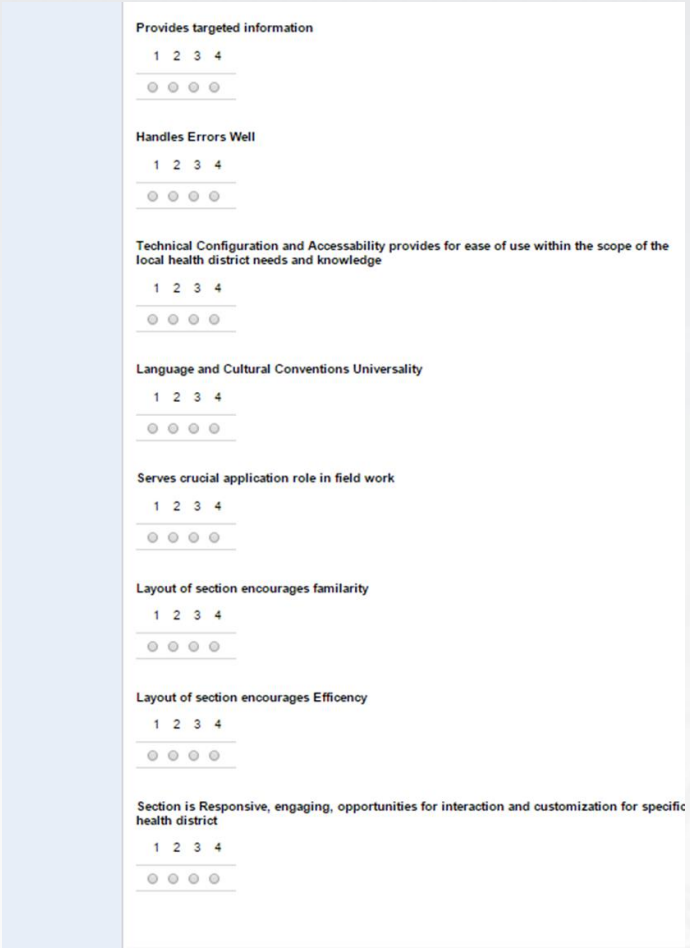
-Dashboard Guided Tutorial Trainings

-Interactive Case Study Based Workshops

- Network Collaboration Capabilities with other State Epidemiologists utilizing Dashboard Tool

-Interactive feedback and response opportunities for Q&A

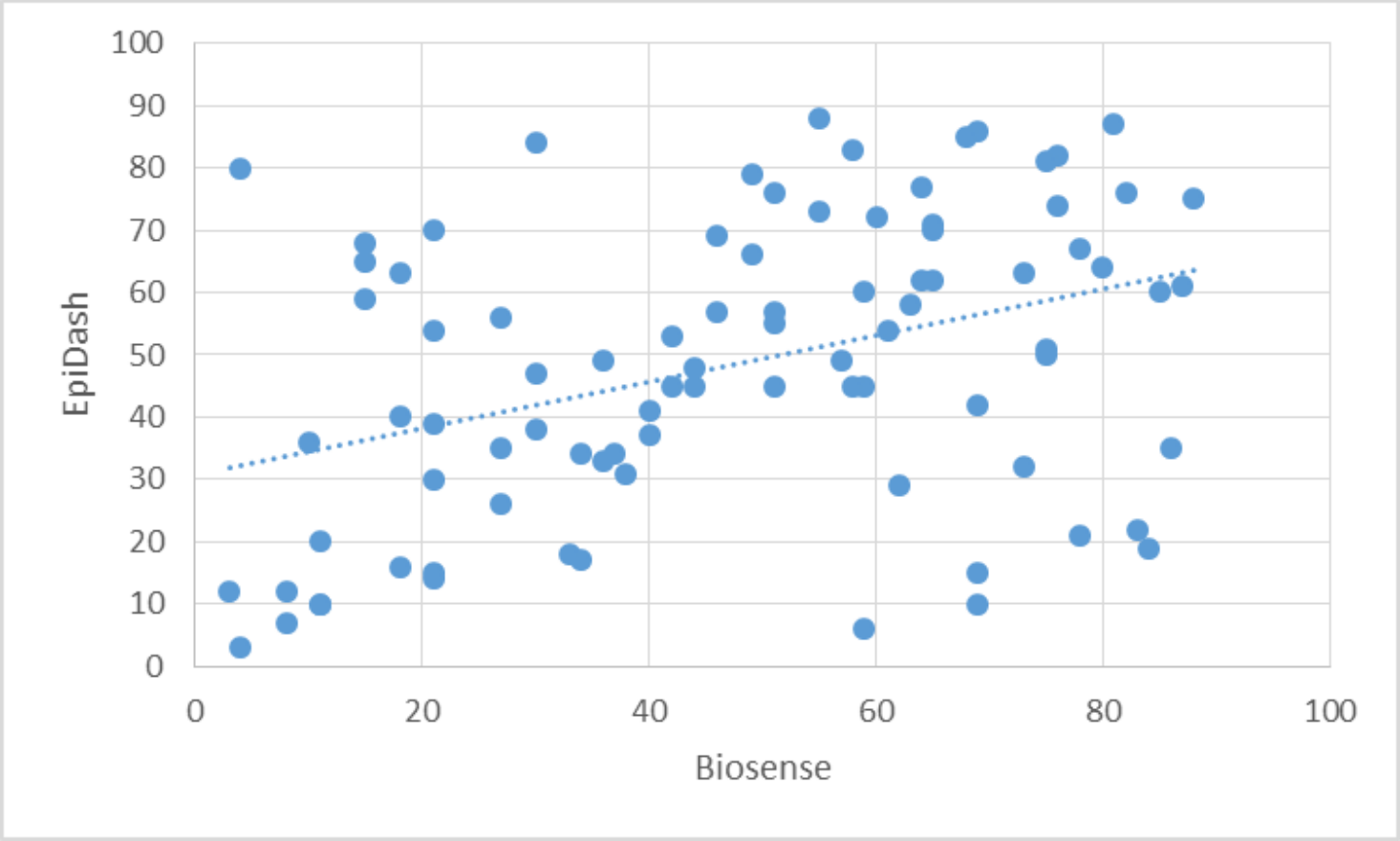
-Protocol for Refinement



Quantitative Evaluation

- BioSense 2.0 provides a mechanism to collect and share information on emergency department visits, hospitalizations, and other health related data from multiple sources, including the Department of Veterans Affairs (VA), the Department of Defense (DoD), and civilian hospitals from around the country.

Clues from Biosense…



Other measures….

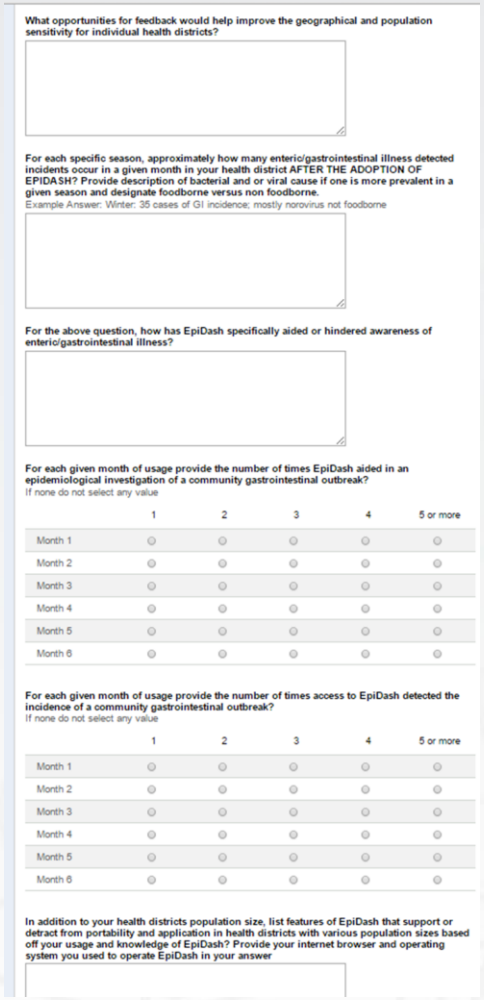
- VDH Enteric Disease Data
- Health Map

Evaluation

Post-Surveillance Evaluation:

VDH state district epidemiologists

- Data Quality
- Level of Health situational awareness
- System utility
- Timeliness of data
- System stability
- Portability
- User-flow-time percentage
- Detection vs. Investigation



What opportunities for feedback would help improve the geographical and population sensitivity for individual health districts?

For each specific season, approximately how many enteric/gastrointestinal illness detected incidents occur in a given month in your health district AFTER THE ADOPTION OF EPIDASH? Provide description of bacterial and or viral cause if one is more prevalent in a given season and designate foodborne versus non foodborne.

Example Answer: Winter: 35 cases of GI incidence; mostly norovirus not foodborne

For the above question, how has EpiDash specifically aided or hindered awareness of enteric/gastrointestinal illness?

For each given month of usage provide the number of times EpiDash aided in an epidemiological investigation of a community gastrointestinal outbreak?

If none do not select any value

	1	2	3	4	5 or more
Month 1					
Month 2					
Month 3					
Month 4					
Month 5					
Month 6					

For each given month of usage provide the number of times access to EpiDash detected the incidence of a community gastrointestinal outbreak?

If none do not select any value

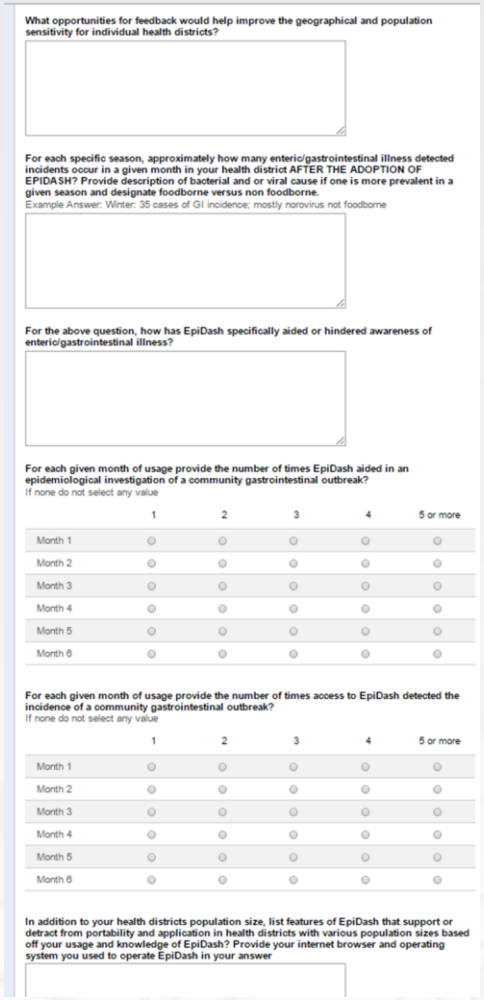
	1	2	3	4	5 or more
Month 1					
Month 2					
Month 3					
Month 4					
Month 5					
Month 6					

In addition to your health districts population size, list features of EpiDash that support or detract from portability and application in health districts with various population sizes based off your usage and knowledge of EpiDash? Provide your internet browser and operating system you used to operate EpiDash in your answer

Evaluation

- Sectional Evaluation:

- Section provides a Simple Format, Easy to Navigate, Minimalist Design to highlight critical information
- Provides targeted information
- Handles Errors Well
- Technical Configuration and Accessibility provides for ease of use within the scope of the local health district needs and knowledge
- Language and Cultural Conventions Universality
- Serves crucial application role in field work
- Layout of section encourages familiarity
- Layout of section encourages Efficiency
- Section is Responsive, engaging, opportunities for interaction and customization for specific health district



What opportunities for feedback would help improve the geographical and population sensitivity for individual health districts?

For each specific season, approximately how many enteric/gastrointestinal illness detected incidents occur in a given month in your health district AFTER THE ADOPTION OF EPIDASH? Provide description of bacterial and or viral cause if one is more prevalent in a given season and designate foodborne versus non foodborne.
Example Answer: Winter: 35 cases of GI incidence; mostly norovirus not foodborne

For the above question, how has EpiDash specifically aided or hindered awareness of enteric/gastrointestinal illness?

For each given month of usage provide the number of times EpiDash aided in an epidemiological investigation of a community gastrointestinal outbreak?
If none do not select any value

	1	2	3	4	5 or more
Month 1					
Month 2					
Month 3					
Month 4					
Month 5					
Month 6					

For each given month of usage provide the number of times access to EpiDash detected the incidence of a community gastrointestinal outbreak?
If none do not select any value

	1	2	3	4	5 or more
Month 1					
Month 2					
Month 3					
Month 4					
Month 5					
Month 6					

In addition to your health districts population size, list features of EpiDash that support or detract from portability and application in health districts with various population sizes based off your usage and knowledge of EpiDash? Provide your internet browser and operating system you used to operate EpiDash in your answer

Points of Further Discussion

- Assessment of evaluation and further refinement to tailor and customize the dashboard to meet local and state health districts needs.
- Deployment in local Emergency and Urgent medical care facilities, schools. Etc.
- The inclusion of a broader media feed to include: local online news, newspapers and Facebook activity.
- Continue to foster awareness of ethical issues surrounding social media data collection
- **Building supportive infrastructure for surveillance system integration.**
- **NLP optimized classifiers: Public Emergency Tracker (fire, weather, crime), Norovirus, Firearms Violence, Tick Zoonoses, Vaccine Sentiment, and Ebola sentiments, unrest, rumors, & misinformation**

Contact Information

Beth Musser:

Virginia Bioinformatics Inst.
1880 Pratt Drive Blacksburg VA 24061

emusser0@vbi.vt.edu or emusser0@vt.edu

(336) 813-2677