in silico Surveillance: Highly detailed Agent-based Models for Surveillance System Evaluation and Design

Bryan Lewis, MPH [1], Allyson Abrams [2], Stephen Eubank, PhD [1], Ken Kleinman, ScD [2]

(1) Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute at Virginia Tech, Research Building XV (0477), 1880 Pratt Drive, Blacksburg, VA 24061.

(2) Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, 133 Brookline Ave., 6th Floor, Boston, MA 02215.

in silico Evaluation and Design of Surveillance Systems

- Detecting outbreaks is a crucial task for public health officials
- Evaluating the actual performance of a surveillance system can be expensive and difficult, if not impossible
- Agent-based models can provide geo-spatially and temporally realistic surveillance data
- We demonstrate a framework for the *in silico* evaluation and design of surveillance systems

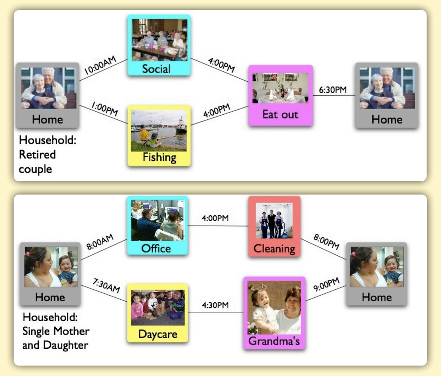
Building an *in silico* Surveillance System

Synthetic Society

Time	Activity	Contacts
6:00	eat out	10
8:00	home	1
11:00	eat out	5
12:00	recreate	15
15:00	transport	2
17:00	eat out	4
20:00	home	1

Synthetic Population :

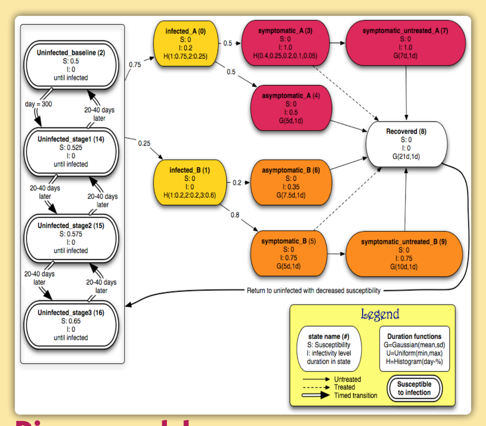
- Individuals created from census data, preserved to census-block level
- Maintains age, gender, household structure, income, for simulation use
- Household locations are geo-located within appropriate census-block to NavTeq home locations



Activity-based Interactions :

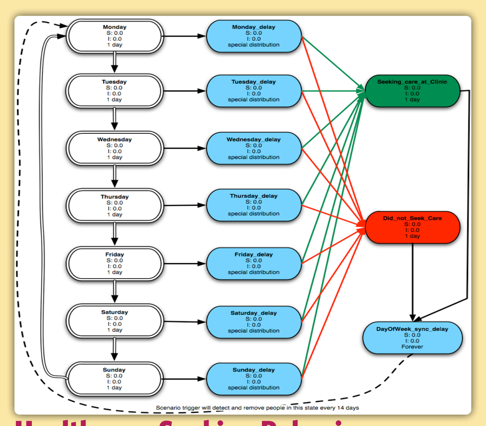
- Activity profiles from time-use surveys are assigned to demographically appropriate individuals
- Dun & Bradstreet data determines what activities occur at each location
- Each individual's daily schedule draws them to activity-appropriate locations, further fitting ensures realistic travel times

Disease and Behavior



Disease model :

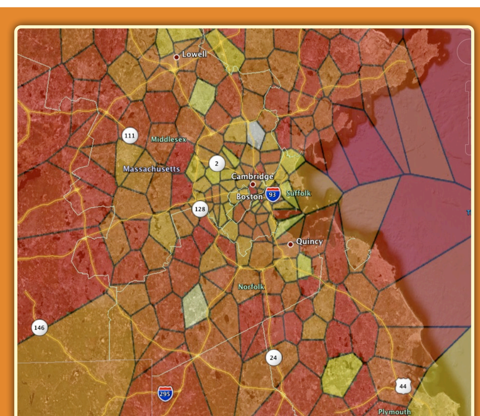
- Finite state machine allows for a flexible representation of disease
- Each state controls level of susceptibility, infectiousness, and symptoms
- Transitions can depend on treatments and any individual's characteristics



Health care Seeking Behavior :

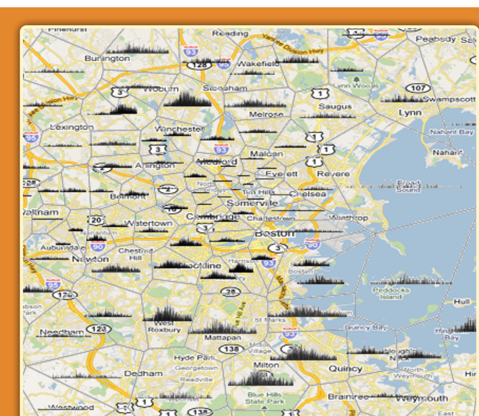
- Identical finite-state machinery as disease model
- States can control actions (stay home, seek care, etc.)
- Accommodates complex combinations of parameter distributions

Surveillance of Cases



Surveilled population selection :

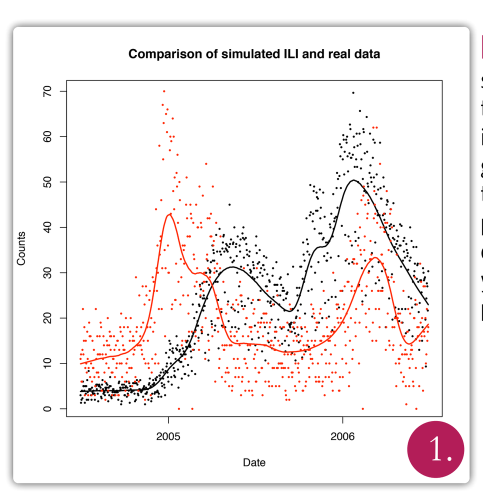
- Can be based on geography, demographics, insurance membership, age distributions, etc.



Detailed information for all cases :

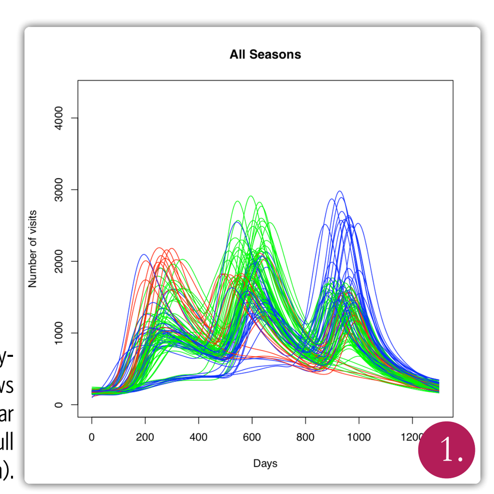
- Cases are individual and thus can be aggregated over any unit of time, space, demographic feature, etc.

Benefits of Simulated Surveillance Data



Realistic Results - By including a similar level of detail and representing many of the processes that generate surveillance data in the real-world, these simulations can generate realistic data that can be calibrated to be indistinguishable from real data. This plot shows both real data and simulated data overlayed, importantly they aren't the same, yet could just as easily be two different time periods from the same source.

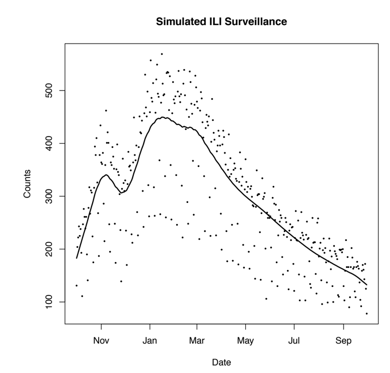
Stochastic Simulations - Highly-detailed agents and a stochastic framework allows each simulation to be highly variable, yet similar within desired constraints. This plot shows the full results of 100 simulations (3 years each).



METHODS

1. Simulate Surveillance Data

Disease model parameterized for disease of interest. Calibrate to approximate global disease levels, add other influences like seasonality and disease interventions. Health care seeking and other behavioral modeling fine tune the simulation results to the desired surveillance system



Demonstration Study Specifics:

A. Calibrate ILI disease for endemicity

B. Make global adjustments to transmissibility to create seasonal peaks

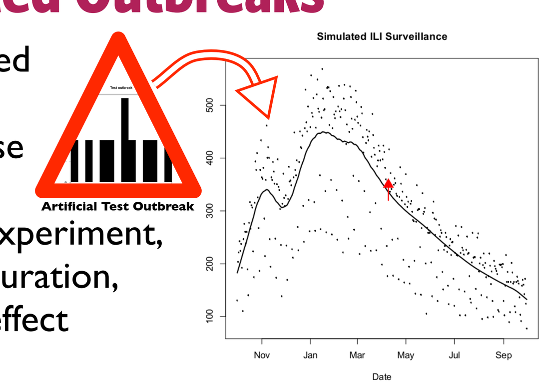
C. Determine if and when a case will seek health care using delay to care and day of week bias

D. Determine if this person is a member of the surveillance system

E. Sum surveilled cases by zip code for each day

2. Insert Standardized Outbreaks

Artificial outbreaks are inserted into the synthetic disease surveillance data stream. These test outbreaks factor into the specific design of the *in silico* experiment, as true-positives. The shape, duration, method of insertion, etc. can effect detection.



Demonstration Study Specifics:

A. Select a random day for artificial outbreak insertion

B. Scale outbreak case numbers to reflect coverage level of surveillance system

C. Select random location and 2 neighboring locations for insertion

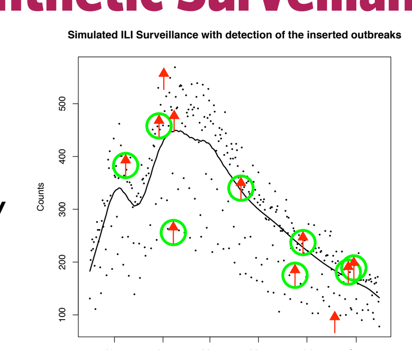
D. Add outbreak cases to surveillance data

E. Remove inserted cases, and repeat 11 times.

F. Retain all 12 data sets with inserted cases and original simulated surveillance data for analysis.

3. Detect Outbreaks in Synthetic Surveillance Data

Surveillance signals with the inserted test outbreaks are evaluated with outbreak detection algorithms. The synthetic surveillance signal without any inserted outbreaks is also evaluated, these outbreak events can be considered false-positives.



Demonstration Study Specifics:

A. Perform SaTScan analysis for every day of an entire simulated ILI season as well as all independent inserted outbreaks

B. Merge SaTScan identified clusters that overlap in time and space under appropriate detection thresholds

C. Identify which SaTScan identified clusters correspond to inserted outbreaks

D. Evaluate surveillance system performance

Demonstration Study:
Is Outbreak Detection More Sensitive to Coverage or Geography?

Synthetic Society:

- Boston Metropolitan Area
- 4.1 million individuals
- 2.1 million unique gathering places housed in nearly 1 million physical locations
- 22.8 million individual activities
- Includes Massachusetts counties: Essex, Middlesex, Norfolk, Plymouth, & Suffolk; New Hampshire counties: Rockingham & Strafford

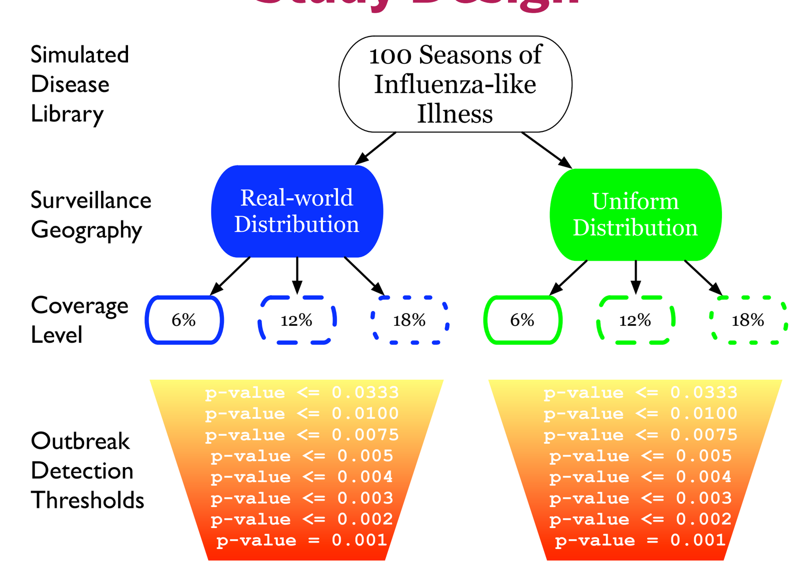
Disease and Behavior:

- Influenza-like Illness (ILI), multiple manifestations: short and very infectious, long and less infectious
- Health care seeking in an outpatient clinic setting, includes health care seeking delays from onset of symptoms as well as a day-of-the-week bias

Surveillance of Cases:

- Harvard Pilgrim Health Care - Membership information used to generate a realistic representation of HPHC ILI surveillance gathered by its network of clinics
- Surveillance data presented as counts per day per home zip-code location (210 unique zip codes)
- Additional surveillance systems created by expanding coverage of HPHC surveillance and changing the geographic distribution of these "real-world" based surveillance systems

Study Design



Probability Detected: Proportion of the inserted artificial outbreaks detected by the system, expressed as a probability.

Detection Delay: Time from day of the first case in the inserted outbreak to the first day the outbreak detection algorithm identified the outbreak.

Tools and Data

- Synthetic Population Construction - Oracle, Dun & Bradstreet, NavTeq, 2000 Census, and National Household Travel Survey
- Agent-based modeling - EpiSimdemics v2.0
- Outbreak detection - SaTScan v7.0.3 (satscan.org)
- Analysis - R v2.9.0 (r-project.org) with packages spatstat v1.15-0 and deldir v0.0-12
- Computing resource - "pecos" an SGI cluster with 768 3GHz cores, 1.5TB shared RAM, 32 TB storage

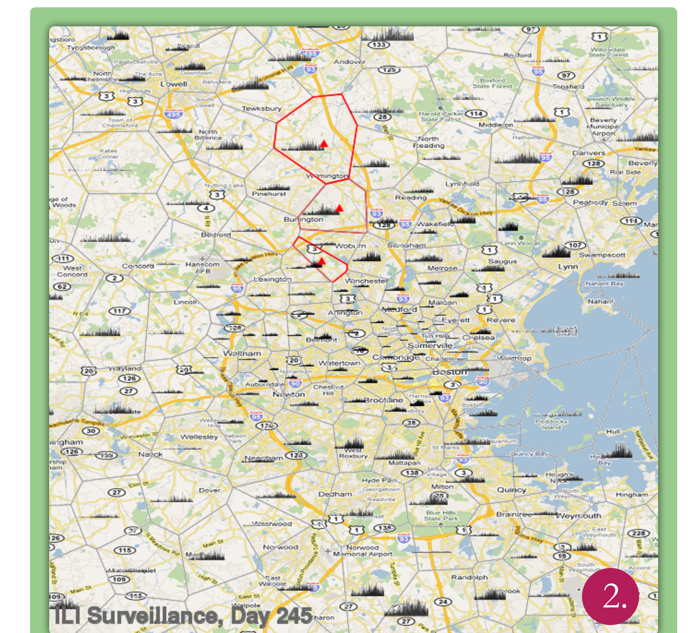
References

1. Atkins K, Barrett CL, Beckman R, Bisset K, Chen J, Eubank S, et al. An Interaction Based Composable Architecture for Building Scalable Models of Large Social, Biological, Information and Technical Systems. CTWatch Quarterly 2008;4(1).
2. Barrett C, Bisset K, Eubank S, Feng X. EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. Proceedings of the 2008 ACM/IEEE conference on Supercomputing 2008.
3. Barrett C, Bisset K, Leidig J, Marathe A, Marathe M. An integrated modeling environment to study the co-evolution of networks, individual behavior and epidemics. AI Magazine 2010.
4. Barrett CL, Beckman RJ, Khan M, Kumar VA, Marathe MV, Stretz PE, et al. Generation and Analysis of Large Synthetic Social Contact Networks. Proceedings of the 2009 Winter Simulation Conference 2009:1003-1014.
5. Eubank S, Guclu H, Kumar VSA, Marathe MV, Srinivasan A, Toroczkai Z, et al. Modelling disease outbreaks in realistic urban social networks. Nature 2004;429(6988):180-4.
6. Halloran ME, Ferguson NM, Eubank S, Longini IM, Cummings DAT, Lewis B, et al. Modeling targeted layered containment of an influenza pandemic in the United States. Proc Natl Acad Sci USA 2008;105(12):4639-44.
7. Kleinman K, Abrams A, Katherine Yih W, Platt R, Kulldorff M. Evaluating spatial surveillance: detection of known outbreaks in real data. Stat Med 2006;25(5):755-69.
8. Kleinman KP, Abrams AM. Assessing surveillance using sensitivity, specificity and timeliness. Stat Methods Med Res 2006;15(5):445-64.
9. Kleinman KP, Abrams AM. Assessing the utility of public health surveillance using specificity, sensitivity, and lives saved. Stat Med 2008;27(20):4057-68.
10. Kleinman KP, Abrams AM, Kulldorff M, Platt R. A model-adjusted space-time scan statistic with an application to syndromic surveillance. Epidemiol Infect 2005;133(3):409-19.

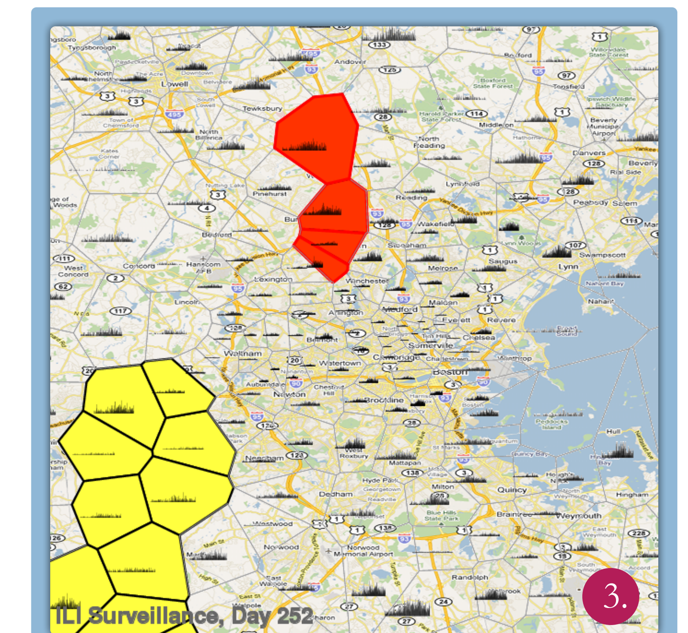
RESULTS

Outbreak Detection

Surveillance System at work - The following figures demonstrate a single synthetic outbreak insertion and its subsequent detection. The counts per day per location are visualized as epicurves centered on each location. Locations that are identified as part of a cluster are filled in with a color ranging from red to pale yellow denoting the statistical significance of the identified cluster (red = p-value of 0.001 and pale yellow = p-value of 0.03333). Inserted artificial outbreaks are shown as red triangles in the epicurves and the locations are also outlined in red.



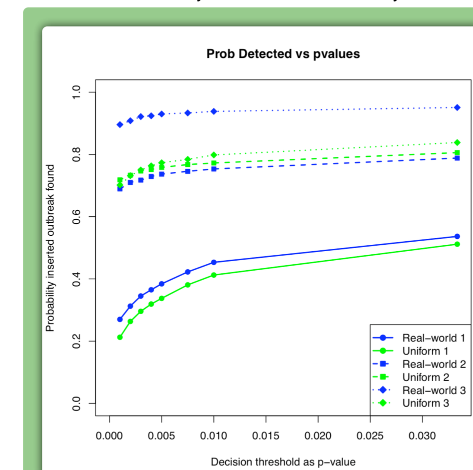
Insert Standardized Outbreak - Standardized artifical outbreak inserted on day 245 of the simulation into 3 locations in the north-central area of Boston. Surveillance counts per day are presented as epicurves centered on the home location of the reported cases.



Detect Outbreaks - Cluster identified where artificial outbreak was inserted with a detection threshold of p = 0.001, a less significant "false-positive" cluster identified as outbreak independent of any inserted artificial outbreak at a detection threshold of p-value = 0.0075.

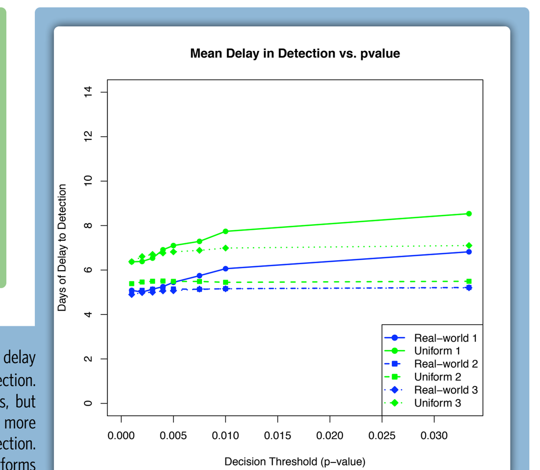
Surveillance System Evaluation

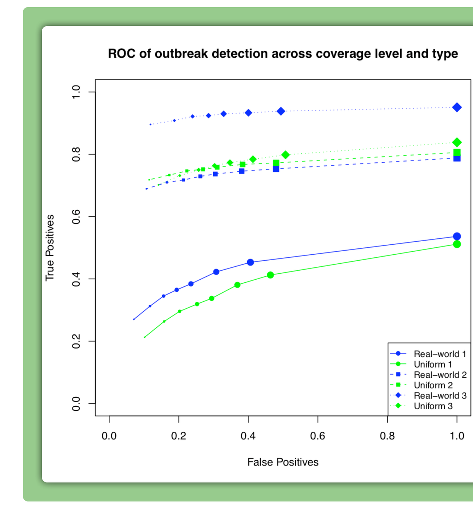
Evaluation - Comparison of the outbreak detection algorithm under the 6 different surveillance systems using 8 different decision thresholds. For this study the main metrics analyzed were the Probability Detected and the Detection Delay.



Probability Detected - Probability an inserted outbreak will be detected by the six surveillance systems over the range of decision thresholds. As coverage increases so does the likelihood of detection, to a limit. Performance is similar between the surveillance geographies at the first two levels of coverage, but the uniform surveillance geography reaches a ceiling and the 3x coverage performs roughly the same as the 2x. The real-world surveillance geography with ~18% of the population surveilled is able to identify ~90% of all inserted outbreaks.

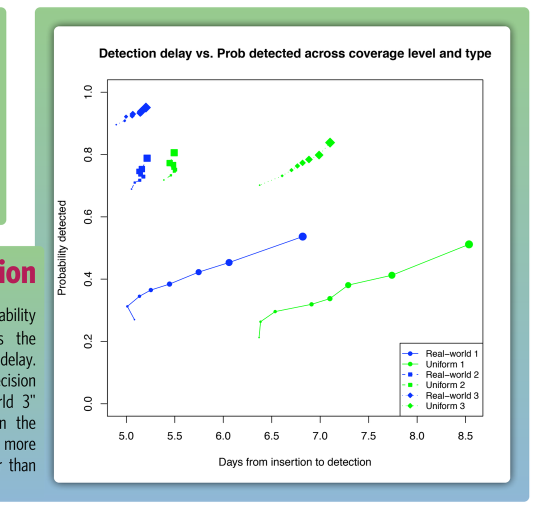
Detection Delays - Mean delay in days from outbreak insertion to its detection. As coverage increases, the delay decreases, but the performance improvement is even more limited than with the probability of detection. Interestingly, the "Uniform 3" system performs similarly to "Uniform 1."





ROC curve - Performance of the surveillance systems presented as a psuedo reciever operator characteristic. True positives are the proportion of inserted outbreaks detected. False positive proportions were based on the number of false positives detected at the highest decision threshold. The decision threshold influences the proportion of false-positives more than the probability of detection

Probability & Detection Delay - Combined detection probability and delay to detection analysis shows the interplay between detection likelihood and delay. The point-sizes are based on the decision threshold. Unsurprisingly, the "Real-world 3" system is shown to perform better than the others, the combined plot reveals with more clarity that "Real-world 2" performs better than the "Uniform 2 & 3" systems.



Next Steps

- Evaluate other outbreak detection algorithms
- Test sensitivity to artificial outbreak characteristics
- Evaluate the impact of additional data for outbreak detection protocols (age, work/school location, etc.)
- Offer *in silico* surveillance data construction through webservice to allow 3rd party experimentation

Conclusions

- Agent-based models can help design and evaluate surveillance systems
- How a surveillance system is distributed over geography can be more important than the level of coverage
- This approach can be used to improve current outbreak detection algorithms by applying machine learning techniques which was previously impossible to do with limited training data sets.

Models of Infectious Disease Agent Study
MIDAS

Department of Population Medicine

VIRGINIA BIOINFORMATICS INSTITUTE AT VIRGINIA TECH

More information available:
http://ndssl.vbi.vt.edu/insilicoSurveillance

Contact info:
Bryan Lewis | blewis@vbi.vt.edu | ndssl.vbi.vt.edu