CMU Auton Lab: Introduction

- The Auton Lab was founded in 1993 by Andrew Moore (now director of Google Pittsburgh)
- Central topic of research: scalable, self-adaptive analytic systems with real life impact
- Currently about 18 people

  2 regular+3 affiliated faculty, 1 post-doc, 5 programmers and analysts, 5 PhD students; a few interns; led by Artur Dubrawski and Jeff Schneider
- Currently working on 10+ sponsored projects

  Current and past funding from NSF, DARPA, DHS/HSARPA, ONR, AFRL, NASA, USDA, FDA, CDC, a few Fortune 100 companies, and a number smaller corporate & academic sponsors and partners
- Deliverables
  - Algorithms for fast and scalable statistical machine learning
  - Software for embedding in production systems
  - Software available for download (currently 100 new downloads per month)

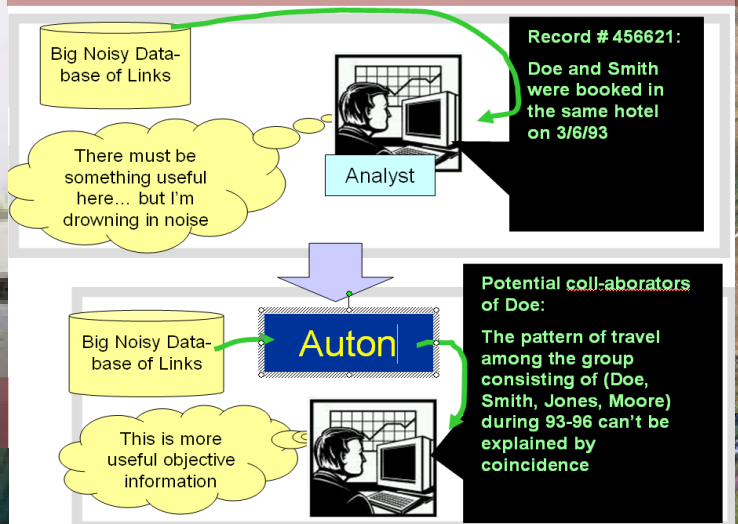
  www.autonlab.org

Nuclear threat detection

Drug discovery

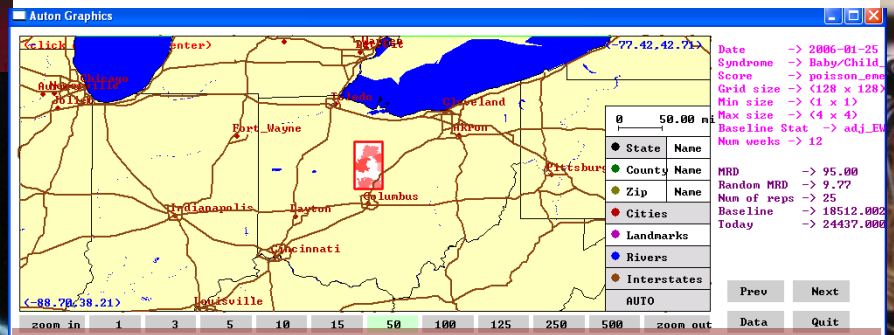
Astrophysics

Human intelligence

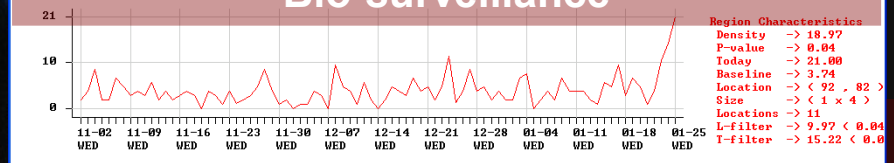


Carnegie Mellon Auton Lab

Recent success stories

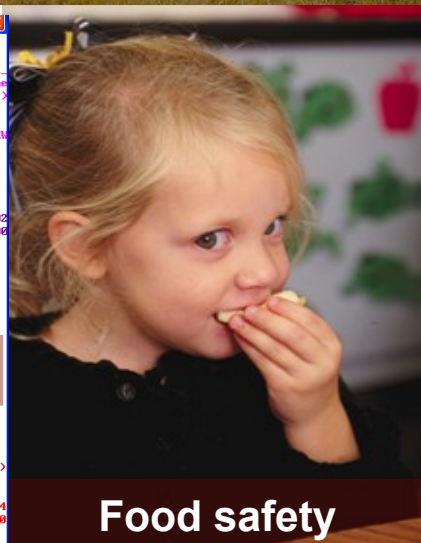


Bio-surveillance



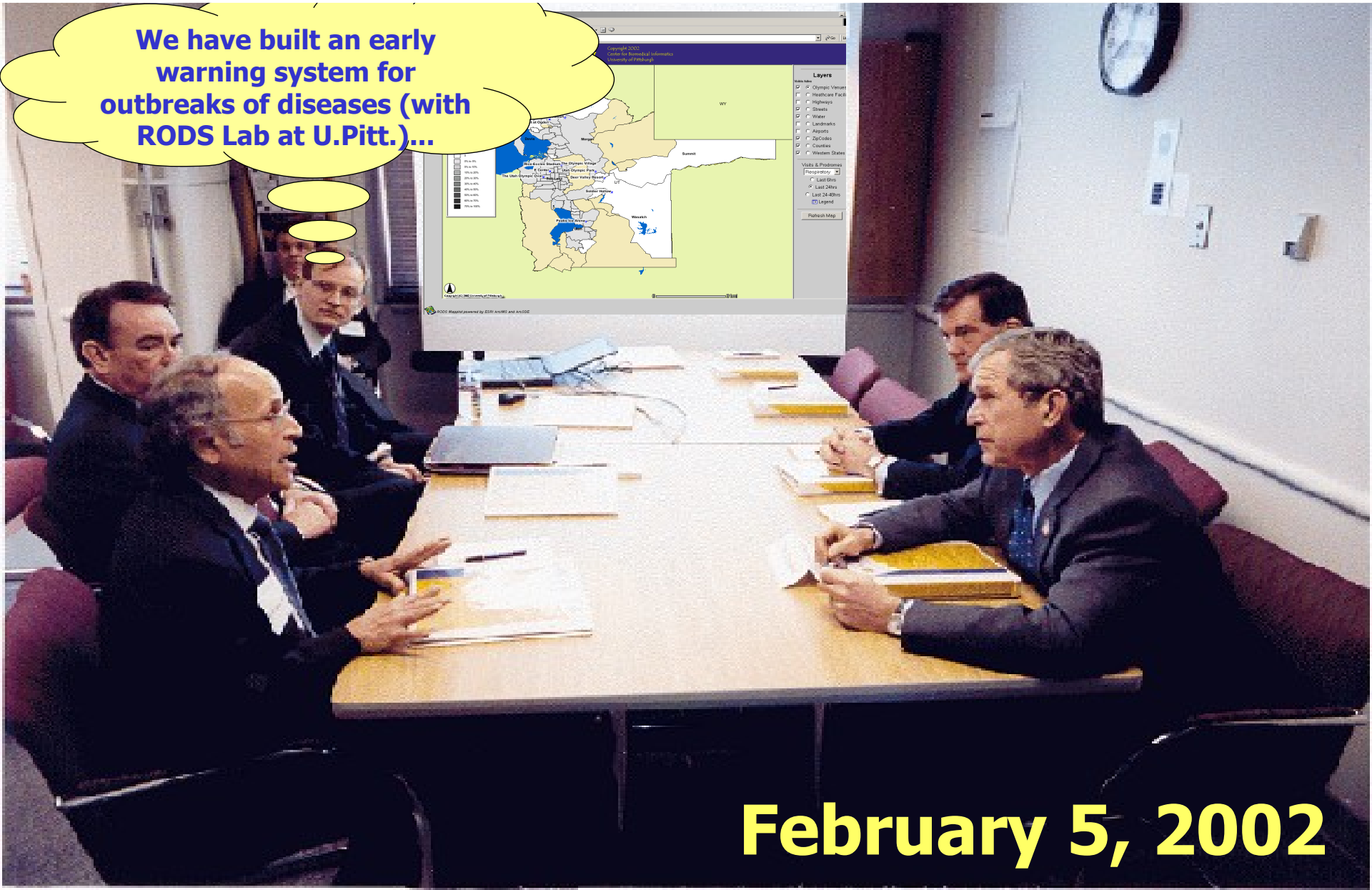
Fleet prognostics

Safety of agriculture



Food safety

Real-time Monitoring of Emergency Department Chief Complaints during 2002 Winter Olympics



Why Do We Care?

Highly motivational example: Adverse bio-events

- Terrorist acts (Anthrax, Smallpox, ...)
  - 100 kg of anthrax released in DC may kill 1—3 million people (WHO)
  - "For the life of me, I cannot understand why terrorists have not attacked our food supply because it is so easy to do" (Resignation speech by departing U.S. Health Secretary Tommy Thompson, December 3, 2004)
- Naturally occurring events, including emerging threats (SARS, Avian Influenza, West Nile Virus, Mad Cow Disease, Foot-and-Mouth Disease, E.coli, Salmonella, ...)
  - Lower-bound estimate of death toll of Avian Influenza pandemic: 2—7 million lives
- Unintentionally introduced events (such as accidental contamination of food at processing plant)
  - Each year in the US: 76 million cases of food borne diseases (5,000+ of them fatal)

Why Do We Care?

Detection based on definitive diagnoses may involve too much latency:

- A 2-day gain in detection time during an incident of inhalational Anthrax release could reduce fatalities by a factor of six (DARPA)
- Improvements of even an hour over current detection capabilities could reduce economic impact of a bioterrorist anthrax attack by hundreds of millions of dollars (Wagner et al. 2004)
- US agricultural industry: $1T of economic activity, $60B in food exports, root cause identification is apparently very difficult: 82% of food-borne disease cases are of unknown origin.

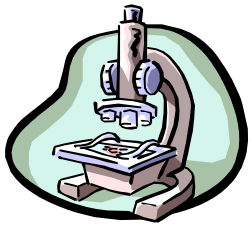
How we can help:

- Exploit available early signals
- Build algorithms and supporting data structures for rapid detection

What Kind of Data Carries Early Signals?

Post-Diagnosis Data

Specific, available after several weeks

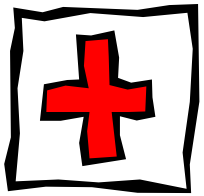


Lab Reports

Mortality Reports

Pre-Diagnosis Data

Non-specific, available relatively early



ER Chief Complaints

Pharmacy OTC Sales

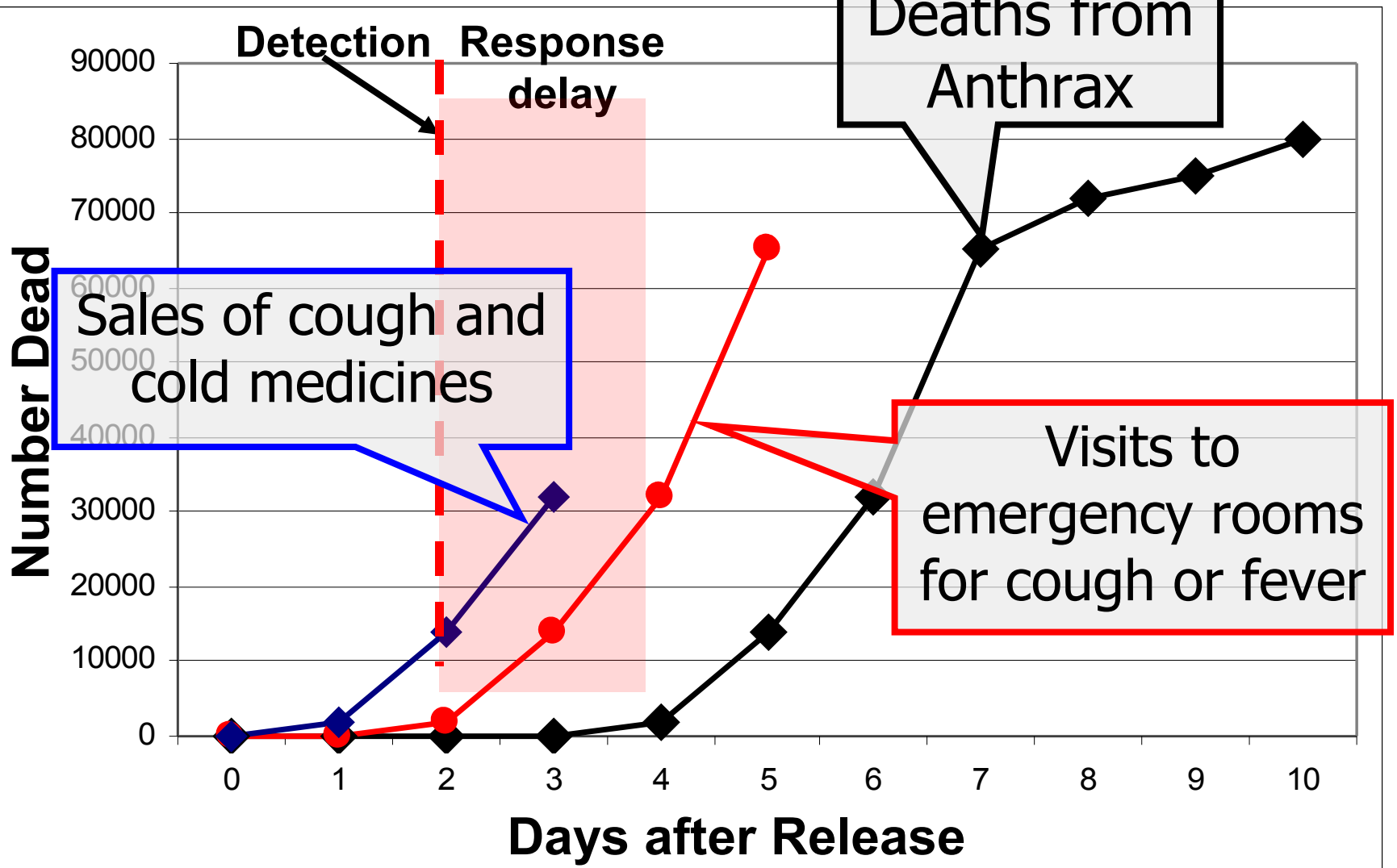
School Absenteeism

Ambulance Response Logs

Veterinarian Data

Consumer Complaints

Potential Utility of Non-specific Data



Auton Lab's Field-tested Algorithms and Data Structures for Rapid Detection of Anomalies and Emerging Patterns

- What's Strange About Recent Events (WSARE)

  Monitoring hospital Emergency Rooms chief complaints for unusually high counts of patients from specific sub-populations who report with similar symptoms

- Fast Spatial Scan
  - Expectation-based
  - Multivariate Bayesian

  Monitoring volumes of non-prescription drug sales and/or ER data for spatio-temporal over-densities

- Multi-Stream Temporal Monitor

  Combining evidence from multiple streams of time series surveillance data for detection of anomalies or specific patterns of interest

- Tip Monitor

  Monitoring food consumer complaints for small sets of similar reports which may be attributed to the same underlying cause

- T-Cube

  In-memory data structure enabling huge speedups in time series extraction and aggregation

What's Strange About Recent Events

(Wong's Ph.D. thesis, 2003)

Representative Surveillance Data

Date	Time	Hospital	ICD9	Prodrome	Gender	Age	Home Location	Many more…
6/1/03	9:12	1	781	Fever	M	20s	NE	…
6/1/03	9:45	1	787	Diarrhea	F	40s	SE	…
:	:	:	:	:	:	:	:	:

Standard Approach

Select in advance which subpopulations to monitor (e.g., each county, zip)

Do not pay close attention to effect of multiple testing

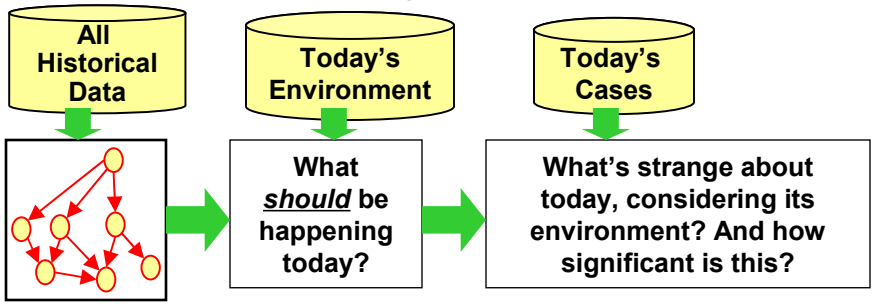
WSARE Approach

Monitor hundreds of thousands of subpopulations

Pay close attention to effect of multiple testing

Method

- Search over 100,000s of subpopulations
- For each subpopulation, use as good of a model as can be created to predict expected counts
- The model will have a form of a multi-component rule, e.g. ***"There is a surprisingly large number of children with respiratory problems today"***
- Compute p value, taking into account multiple testing



Evaluation

Detailed comparison on 2,000 simulated scenarios and Western PA ER Data

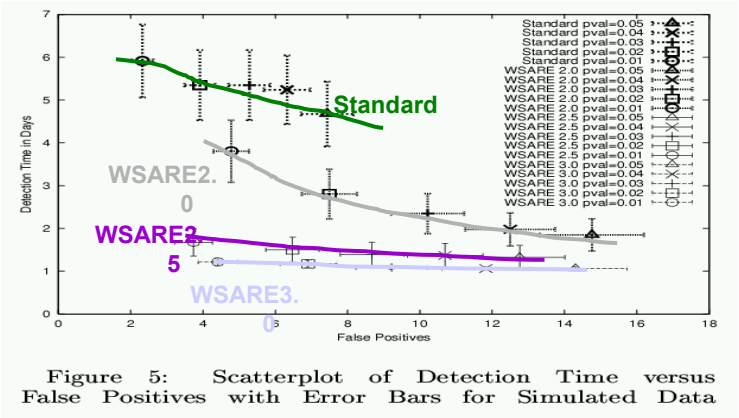


Figure 5: Scatterplot of Detection Time versus False Positives with Error Bars for Simulated Data

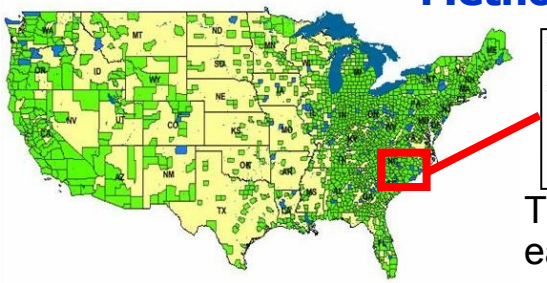
Significance

- WSARE is designed to detect small clusters of illness in healthcare workers, age groups, workplaces...
- Israeli Center for Disease Control evaluated WSARE retrospectively using an unusual outbreak of influenza type B that occurred in an elementary school in central Israel. It detected the outbreak on the second day from its onset.
- Retrospective analysis of Walkerton case: alerted one day ahead of the issuance of boil water advisory (at the expected rate of 2 false positives per year).

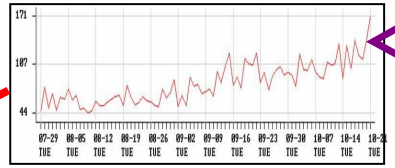
Fast Spatial Scan

(Neill's Ph.D. thesis, 2006)

Method (Expectation-Based Variant of FSS)



OTC drug sales



Time series of counts for each zip code (at least 3 months of historical data).

- This increase could be due to an outbreak, or due to chance.
- Which regions of increase are significant?
- Search all rectangular regions on the grid.

Multivariate Bayesian variant of this method analyzes multiple streams of data to maximize detection power while enabling disambiguation among possible causes of outbreaks

- Solution: We learn the expected count for each area from historical data.
- Then we find regions where the recent **counts are significantly higher than expected**, accounting for anticipated spatial and temporal variations

Evaluation

- Current version typically turns multi-day analysis into **20 minutes** for daily counts from over 20,000 drugstores nationwide
- Searches *all* rectangular regions.
- Results are *exact* (not approximations).

Significance

- Traditional spatial scan is very expensive, especially with randomization tests of significance
- A few hours difference may actually matter.
- Retrospective analysis of Walkerton case: alerted two days ahead of the issuance of boil water advisory.

Algorithm	Search space	# of regions	Search time	Time / region	Likelihood ratio
SaTScan	Circles centered at datapoints	150 billion	16 hours	400 ns	413.56
exhaustive	Axis-aligned rectangles	1.1 trillion	45 days	3600 ns	429.85
FSS	Axis-aligned rectangles	1.1 trillion	81 minutes	4.4 ns	429.85

ER dataset (600,000 records), 1000 replicas; for SaTScan: M=17,000 distinct spatial locations; for exhaustive/fast: 256 x 256 grid.

Tip Monitor

(Dubrawski et al., IAAI 2006)

Application:

Monitoring food consumers complaints for emerging patterns of public health significance

Method:

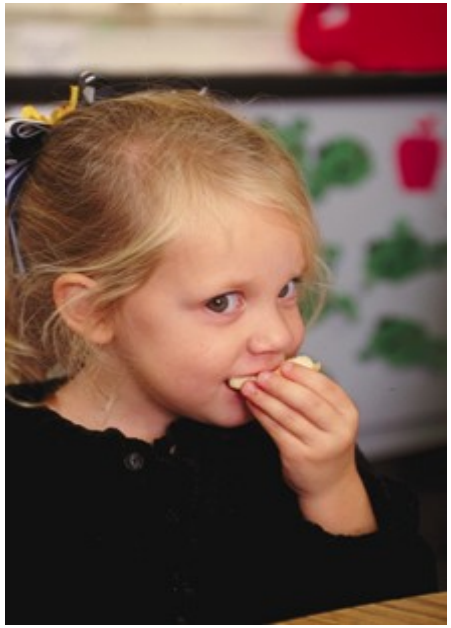
4. Hypothesize that the two complaints have been generated by the same underlying process (selected from a list of predefined possible causes)

- Estimate probability of such an event using a model partially learned from historical data and partly obtained from experts
- Scan through all potentially relevant cases from the recent past and through predefined probabilistic models of causal scenarios
- Report a few top matches to human analysts for further evaluation

$$Q_{ik} = \frac{I_{kn} \cdot R_{ik} \cdot K_{ikn}}{G_n\left(1 - \sum_k \sum_j R_{jk}\right) + \sum_k \left(I_{kn} \cdot \sum_j \left(R_{jk} \cdot K_{jkn}\right)\right)}$$

Hypothesized scenario: X_n is a "noisy copy" of some X_i given H C_k

Alternative: X_n is not a "noisy copy" of any past case given $\{C_k\}$



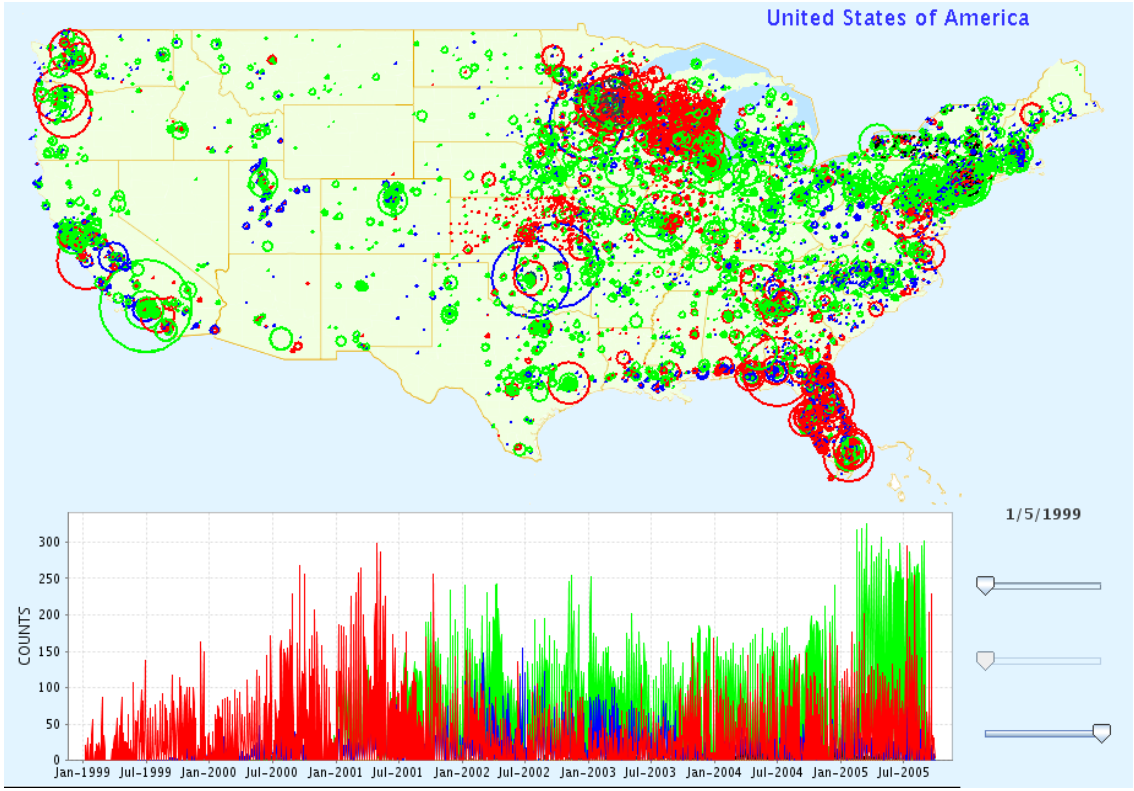
Significance:

- As the part of the USDA Consumer Complaint Monitoring System, Tip Monitor is able to detect interesting low amplitude signals in sparse and noisy data which comes in a short supply
- When tested on historical CCMS data, Tip Monitor rapidly identified several related E.Coli O157:H7 cases. Originally, the relationship was not realized until two weeks into the problem.

T-Cube Web Interface

(Ray et al., AMIA 2008)

- A prototype interface designed to support rapid, interactive navigation of multi-dimensional spatio-temporal data
- Considered for use by the USDA, FDA, CDC and the USAF
- Test version available to public: http://www.autonlab.org/T-Cube/
- We anticipate adopting and customizing it for the purposes of RTBP



Analytics in RTBP: Tasks for Consideration

1. Automated Data Monitoring and Event Detection

- Will provide alerting and reporting functionality
- Key tradeoff: timeliness of detection vs. frequency of alerts
- Eventual utility will strongly depend on accuracy, timeliness and contents/comprehensiveness of the available data
  - Shall we limit data collection to only what is subjectively interpreted as communicable diseases?
  - Could all pre-diagnostic information be entered into the database as soon as it is available (the diagnosis would follow if and when it becomes available)?
  - Could we receive data at the individual transaction level?

2. Interactive Data Navigation via Web-based Interface

- Drill-downs, roll-ups, spatio-temporal visualization
- Support trace-backs and retrospective investigations
- Support attribution and explanation of detections

3. Maintainability of the Statistical Models (optional)

- Surveillance models will age with time, they will need to be periodically re-trained or made adaptive (using Machine Learning)
  - Is there any pre-existing historical data that could be used for this purpose?

Contact Information

Carnegie Mellon University
5000 Forbes Avenue, NSH 3121
Pittsburgh, PA 15213-3890, USA

www.autonlab.org

Artur Dubrawski
Director, Auton Lab
Systems Scientist, Robotics Institute
Adjunct Professor, Heinz School of Public Policy and Management

Tel: 412-268-6233
Fax: 412-268-7350
E-mail: awd@cs.cmu.edu