Mining Health Data for Informative Patterns

Artur Dubrawski

Auton Lab
The Robotics Institute
Carnegie Mellon University
www.autonlab.org
awd@cs.cmu.edu

Carnegie Mellon

Auton Lab

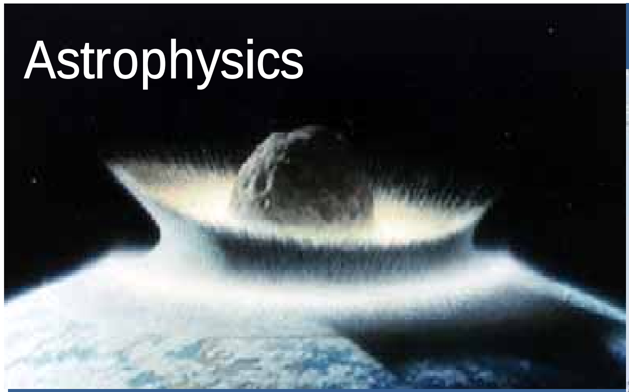
CMU Auton Lab: Research and applications

- Central topic of our research: **scalable**, **self-adaptive analytic systems** with **real-life impact**

- **20+ people:** 2 regular+3 affiliated faculty, 2 post-docs, 8 analysts and programmers, 10 graduate students; + interns; led by Artur Dubrawski and Jeff Schneider

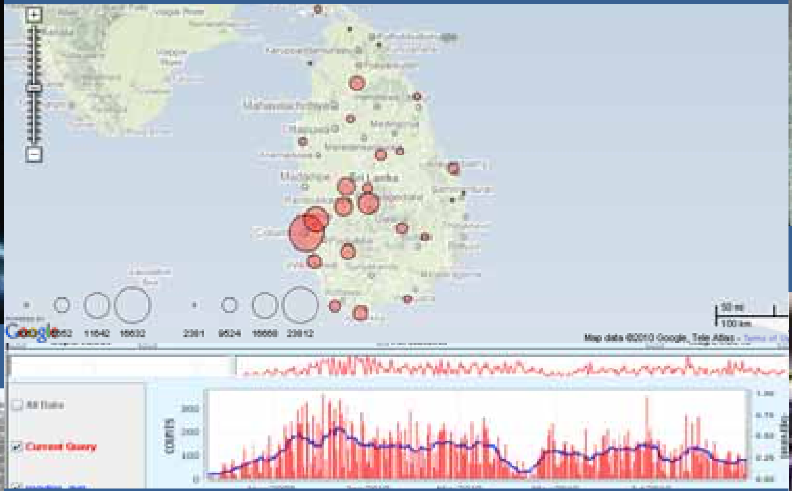
- **Working on 10+ sponsored projects.** Current and past funding from NSF, DARPA, DHS, DoD, HSARPA, IARPA, NASA, USDA, CDC, FDA, IDRC, a few Fortune 100 companies, and a number of smaller corporate & academic sponsors and partners

- **Our deliverables:**
  - **Algorithms** for fast and scalable statistical machine learning and analytics
  - **Software** for embedding in production systems
  - Software available for download at www.autonlab.org



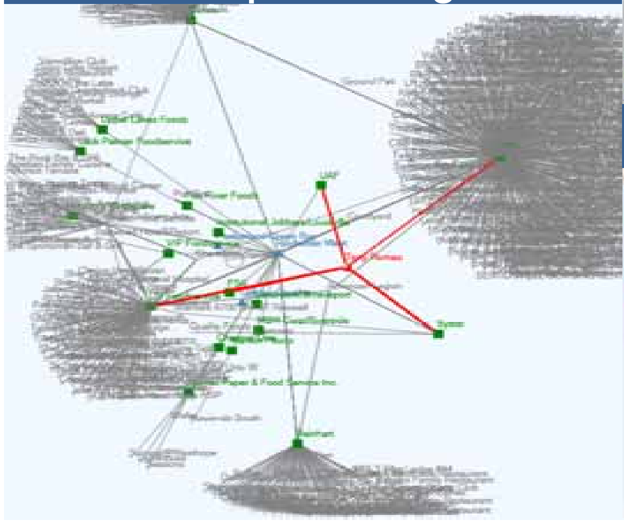
Astrophysics

Public health surveillance

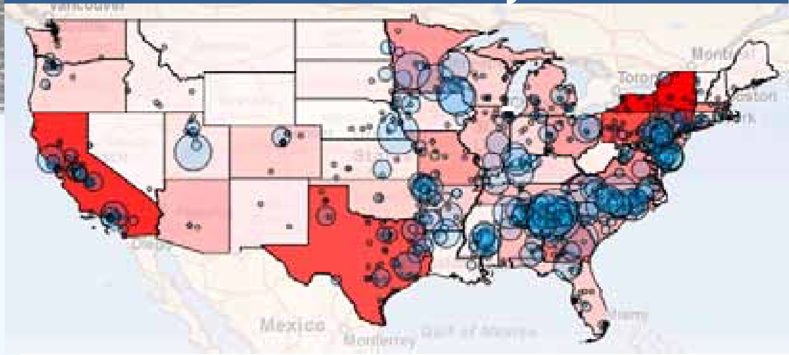


Saving sea turtles

Graph mining



Interactive analytics



Food safety

Nuclear threat assessment

Fleet health management

Safety of agriculture

Bottom Line Up Front

☺ **We can develop and deploy societally-beneficial analytic systems**

But, how to sustain them in a long term?

→ **Still looking for exact answers…**

Three examples from Healthcare Informatics:

- **Different target applications:**
  Public Health, Food Safety, Clinical Care
- **At different stages in their lifetime**
  Know-how → Show-how → Use-how → Sustain-how
- **All rely on the same type of enabling technology**
  Cached sufficient statistics data structures that facilitate massive-scale comprehensive searches for patterns in large sets of multi-dimensional data

Motivating example: Data that may contain patterns useful for monitoring emergence of crises in public health

Fragment of a multidimensional record of disease cases:

↓ Individual patient cases (time-stamped)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
	Categorical descriptors →																				
date	location	disease	age_group	gender	sign_Joint_pain	sign_Muscle_pain	sign_Red_eyes	sign_Light_colored_stool	sign_Diarrhea	sign_Rash	sign_Vomiting	sign_Dark_urine	sign_Abdominal_pain	sign_Chills	sign_other	symptom_Loss_of_appetite	symptom_Fever	symptom_Fatigue	symptom_Nausea	symptom_Headache	symptom_other
16-Dec-06	Colombo	Dengue_fever	20-24	Female	1	1	0	0	1	1	1	0	1	0	0	0	1	0	1	1	0
16-Dec-06	Colombo	Dengue_fever	Above_45	Female	1	1	0	0	1	1	1	0	1	0	0	0	1	0	1	1	0
16-Dec-06	Colombo	Dengue_fever	35-39	Male	1	1	0	0	1	1	1	0	1	0	0	0	1	0	1	1	0
16-Dec-06	Colombo	Dysentery	40-45	Female	0	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0	0
16-Dec-06	Kandy	Dengue_fever	20-24	Male	1	1	0	0	1	1	1	0	1	0	0	0	1	0	1	1	0
16-Dec-06	Kandy	Dysentery	0-1	Male	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	0
16-Dec-06	Kandy	Dysentery	30-34	Female	0	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0	0
16-Dec-06	Matale	Viral_Hepatitis	Above_45	Female	0	0	0	1	0	0	1	1	1	0	0	1	1	1	1	0	0
16-Dec-06	Nuwara_Eliya	Viral_Hepatitis	Above_45	Male	0	0	0	1	0	0	1	1	1	0	0	1	1	1	1	0	0
16-Dec-06	Galle	Dengue_fever	20-24	Male	1	1	0	0	1	1	1	0	1	0	0	0	1	0	1	1	0
16-Dec-06	Hambantota	Typhus_fever	Above_45	Female	1	1	0	0	0	0	1	0	0	1	1	0	1	0	1	1	0
16-Dec-06	Matara	Dengue_fever	Above_45	Male	1	1	0	0	1	1	1	0	1	0	0	0	1	0	1	1	0
16-Dec-06	Matara	Leptospirosis	20-24	Female	0	1	1	0	1	1	1	0	1	1	0	0	1	0	0	1	1
16-Dec-06	Vavuniya	Dengue_fever	30-34	Female	1	1	0	0	1	1	1	0	1	0	0	0	1	0	1	1	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Challenges:

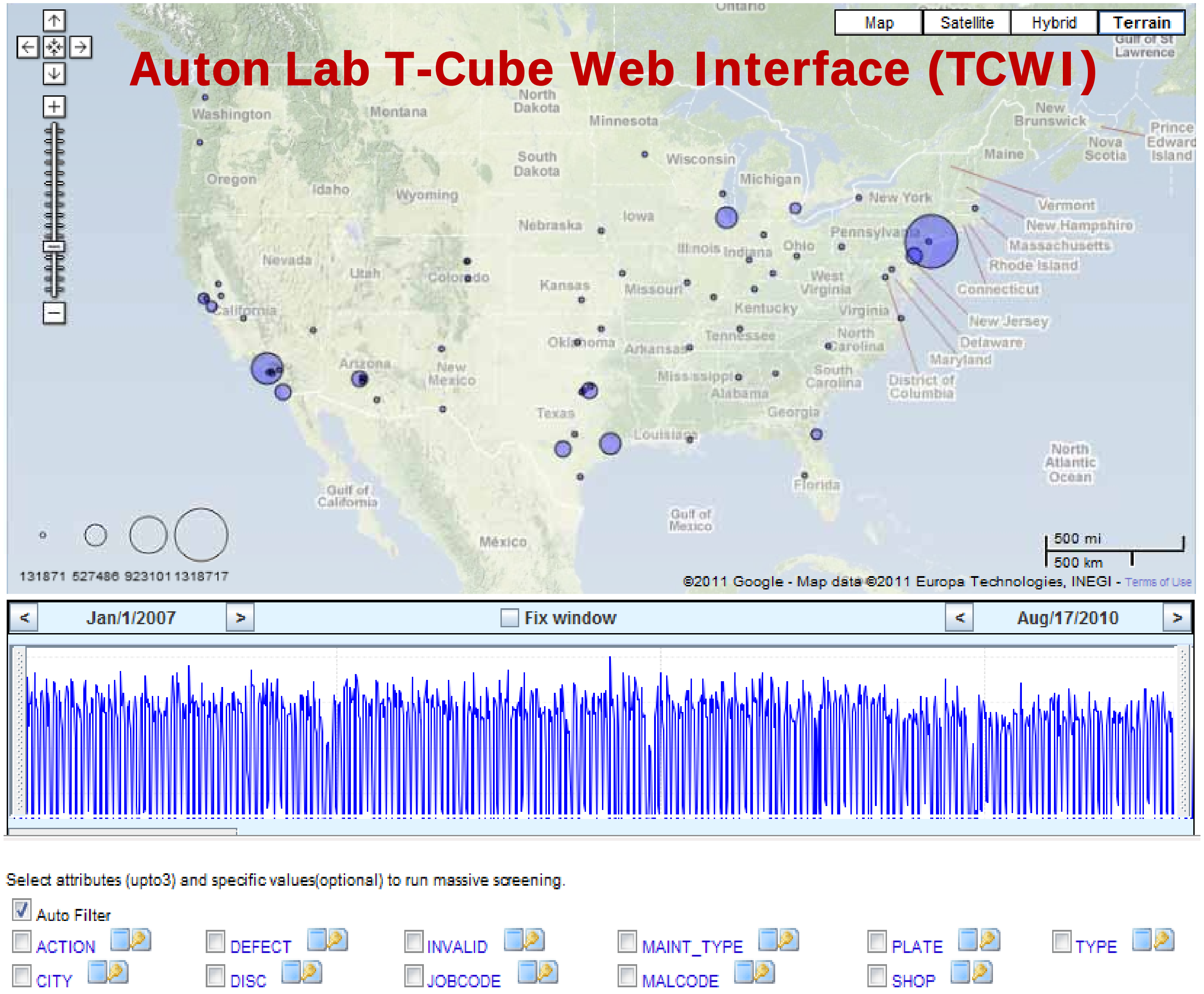
→ Complexity and multiplicity of potentially interesting patterns

→ Can epidemiologists afford monitoring of all possible subpopulations?

The concept of Massive Screening

- **Given: Spatio-temporal transactional data**
  - o Each transaction characterized by several attributes including location and date/time

- **Goal: Find instances of surprising increases (changepoints)**
  - o Comprehensively check many projections of data



Auton Lab T-Cube Web Interface (TCWI)

Select attributes (upto3) and specific values(optional) to run massive screening.

☑ Auto Filter

□ ACTION	□ DEFECT	□ INVALID	□ MAINT_TYPE	□ PLATE	□ TYPE
□ CITY	□ DISC	□ JOBCODE	□ MALCODE	□ SHOP	

Carnegie Mellon

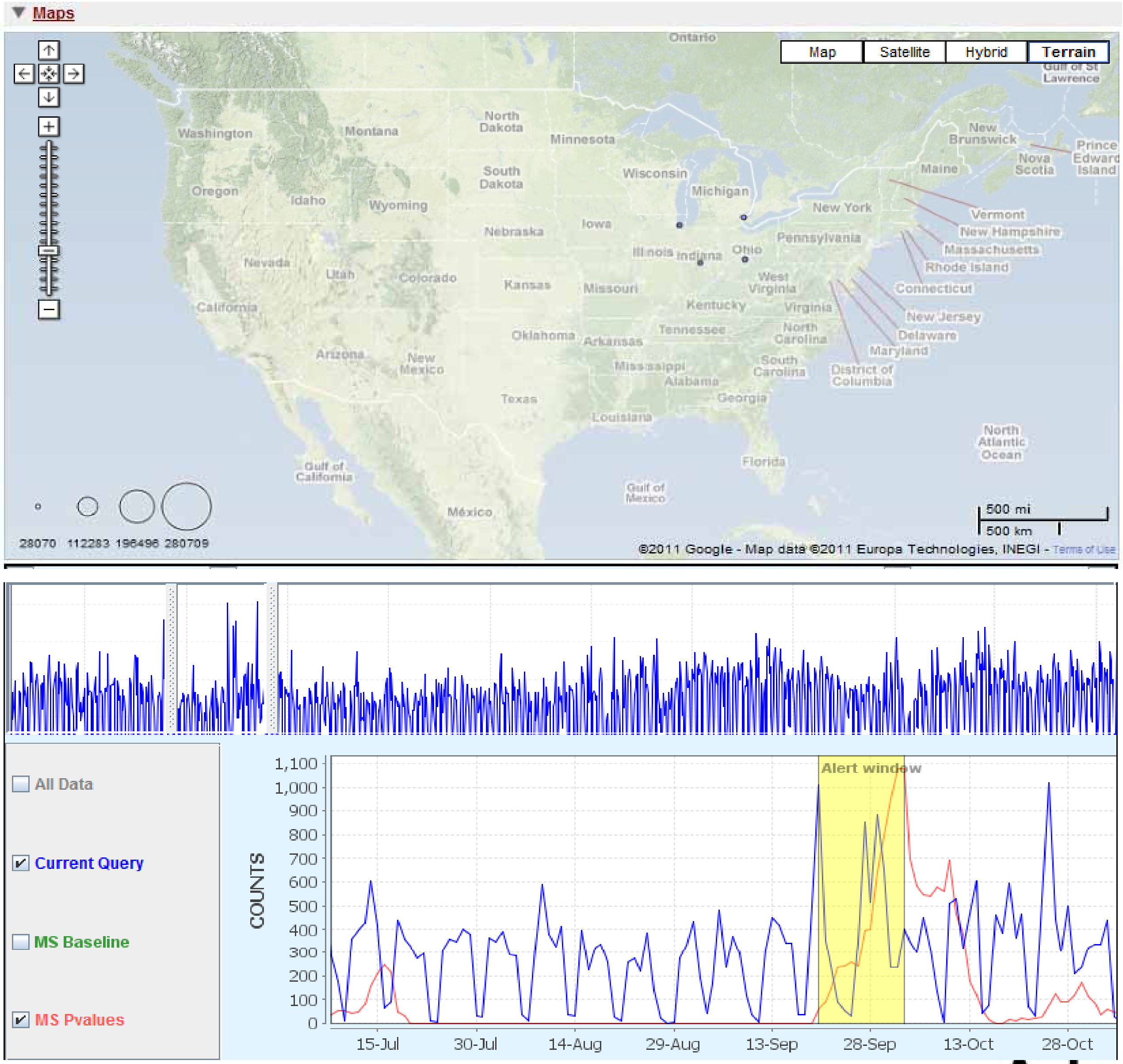
Auton Lab

The concept of Massive Screening

- **Example: Reports of bloody stools among preschool children in the Midwest almost doubled starting in September 2007**

- **How do we evaluate this as a potential alert?**

- **We want to score them all and produce a ranking list…**



▼ Maps

☐ All Data

☑ Current Query

☐ MS Baseline

☑ MS Pvalues

Carnegie Mellon

Auton Lab

Temporal Scan

One of applicable detectors of temporal anomalies (familiar alternative: CuSum):

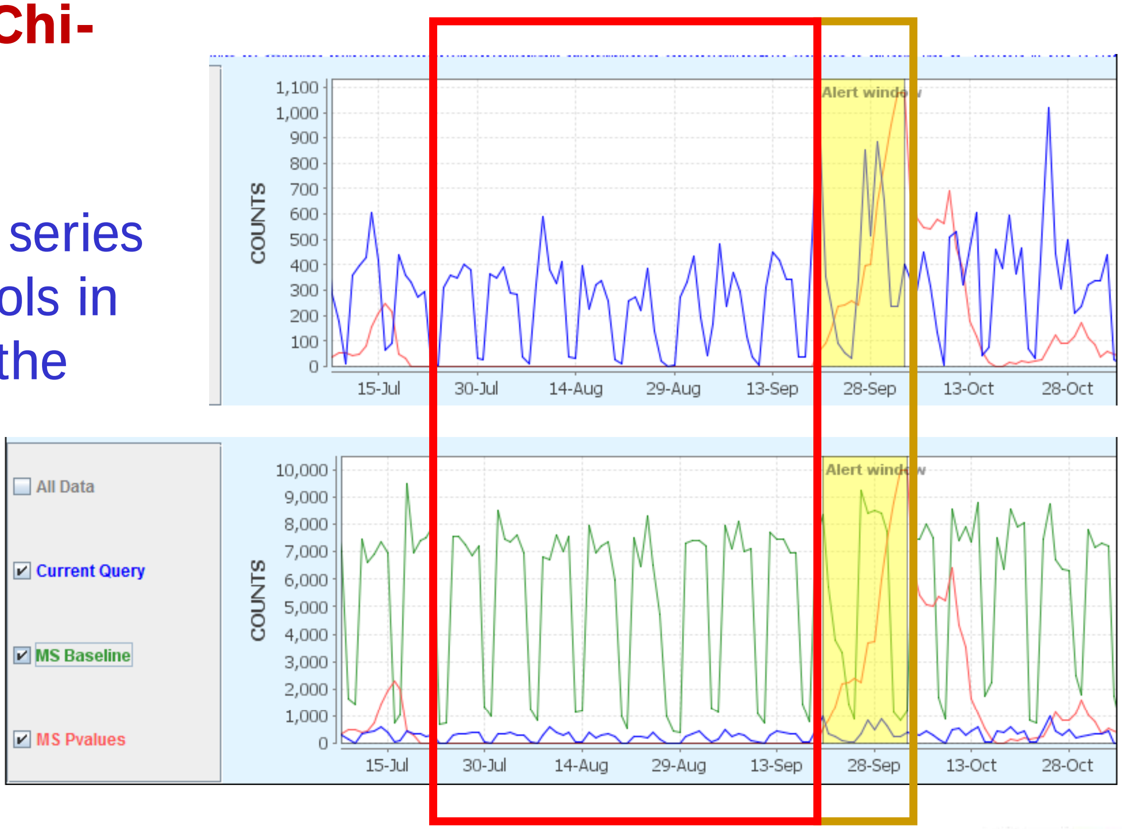
1. **Create a 2x2 contingency table**
2. **Perform a Fisher's exact test (if counts are low) or a Chi-square test**

reference time window (56 days)

target time window (14 days)

target time series (bloody stools in children in the Midwest)

baseline time series (all patient visits)



Carnegie Mellon

Auton Lab

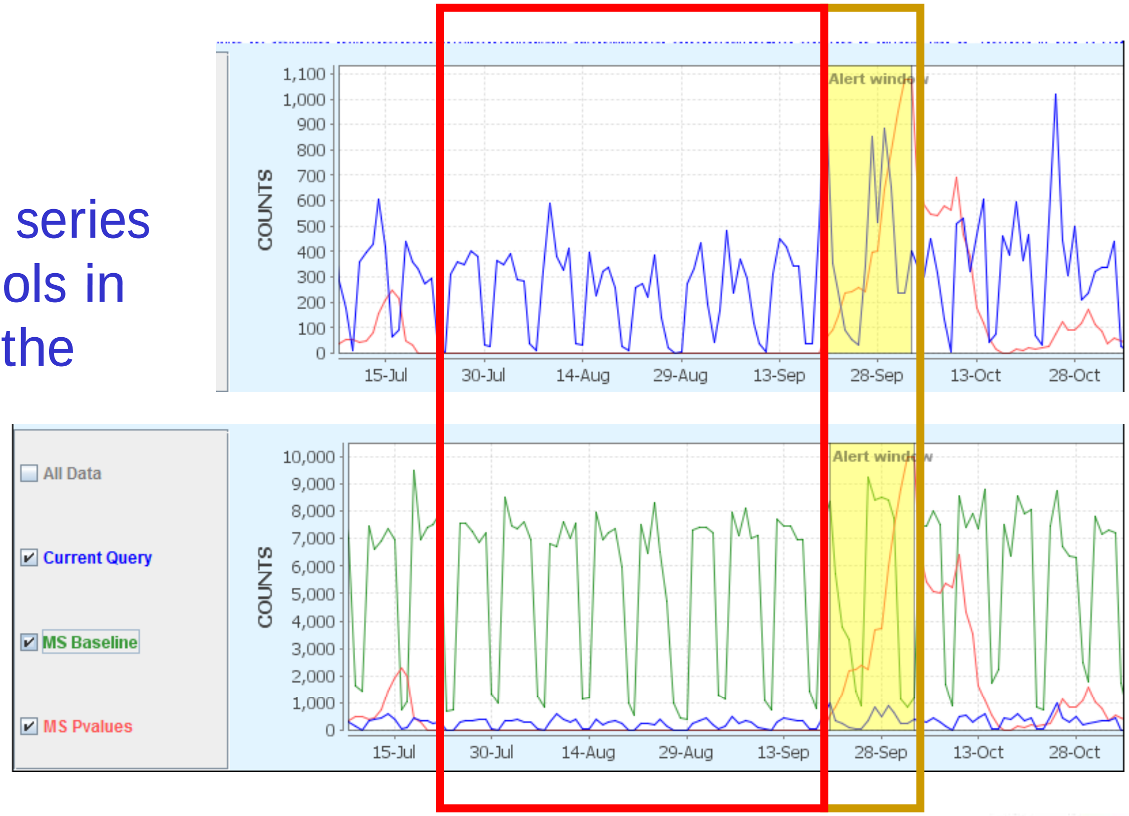
Evaluating a potential detection

	reference	target
target	14K	6K
baseline	300K	70K

p-value < 10^{-20}

reference time window (56 days)

target time window (14 days)



☐ All Data

☑ Current Query

☑ MS Baseline

☑ MS Pvalues

target time series (bloody stools in children in the Midwest)

baseline time series (all patient visits)

Merging scan statistics with massive attribute screening

Plugging-in the spatial context:

- **Consider the cross-product of every possible region with every possible target query on attributes**
- **How many regions are there?**
  - **All subsets of locations: O(2^n)**
  - **All rectangular regions: O(n^4)**
  - ✓ **All locations plus up to k of their nearest neighbors: O(nk)**
- **Use the 2 x 2 contingency table statistics**
- **Report the results sorted in order of p-value**

- **So many hypotheses to test…**
- **Can we afford computing everything?**

Plausible idea: Replace raw data with sufficient statistics

→ Pre-compute key statistics about data ahead of its extensive analyses in order to amortize the bulk of the costs of future computations

Example: Using Contingency Tables to represent categorical data

- Mining categorical data is very often all about counting (co-)occurrences
- Precomputing counts makes the future costs of analyses independent on the data size
- **E[P(CricketFan | Indian)] =**

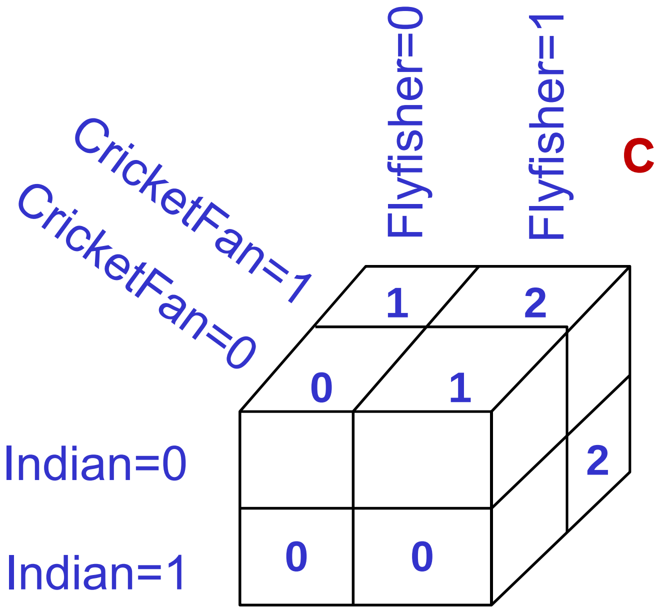
  = NumberOf(Indian CricketFans) / NumberOf(Indians)

Raw data

N=7 Records

Indian?	Flyfisher?	CricketFan?
1	0	1
0	1	0
0	1	1
0	0	1
1	1	1
0	1	1
1	1	1

M=3 Attributes



Contingency table

Plausible idea: Replace raw data with sufficient statistics

→ Pre-compute key statistics about data ahead of its extensive analyses in order to amortize the bulk of the costs of future computations

<u>Example</u>: Using Contingency Tables to represent categorical data

Complaint:

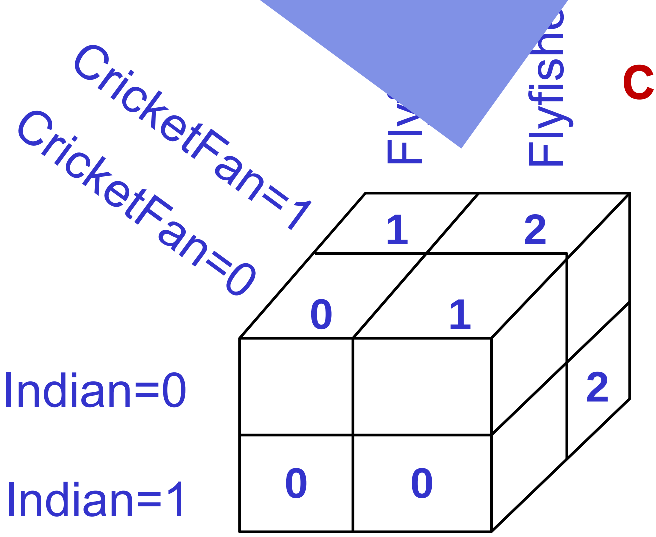
Contingency Tables can reach enormous sizes (numbers of cells) if the underlying data is highly dimensional and if the involved variables can assume many different values

Raw data

N=7 Records

Indian?	Flyfisher?	CricketFan?
1	0	1
0	1	0
0	1	1
0	0	1
1	1	1
0	1	1
1	1	1

M=3 Attributes



Contingency table

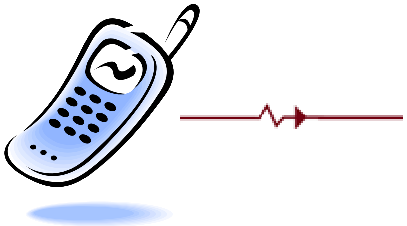
Real-time monitoring of Emergency Department chief complaints during 2002 Winter Olympics

February 5, 2002

Auton Lab
Carnegie Mellon

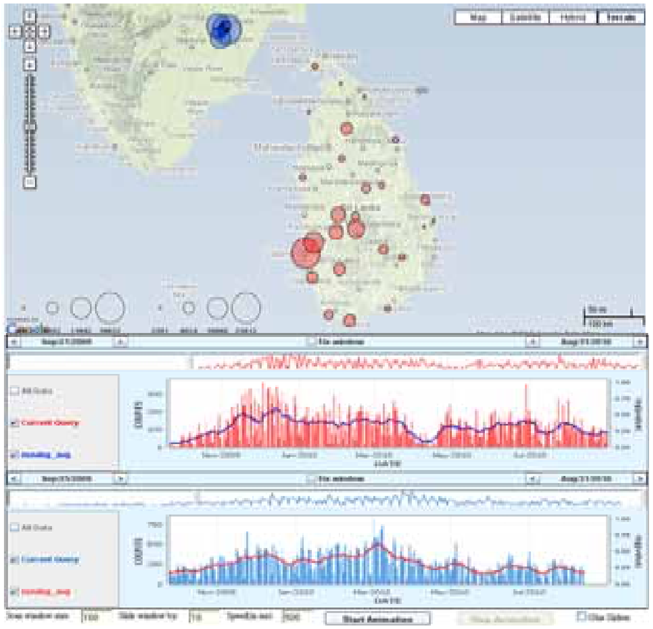
Monitoring Public Health Crises in Developing Countries using T-Cube Web Interface

COLLECT

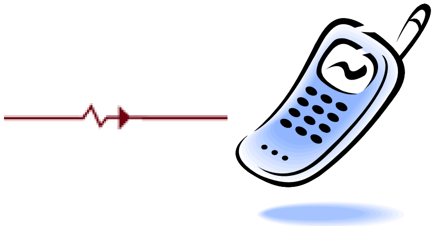


- ✓ No need for sophisticated infrastructure
- ✓ Affordable setup and inexpensive maintenance

DETECT



ALERT



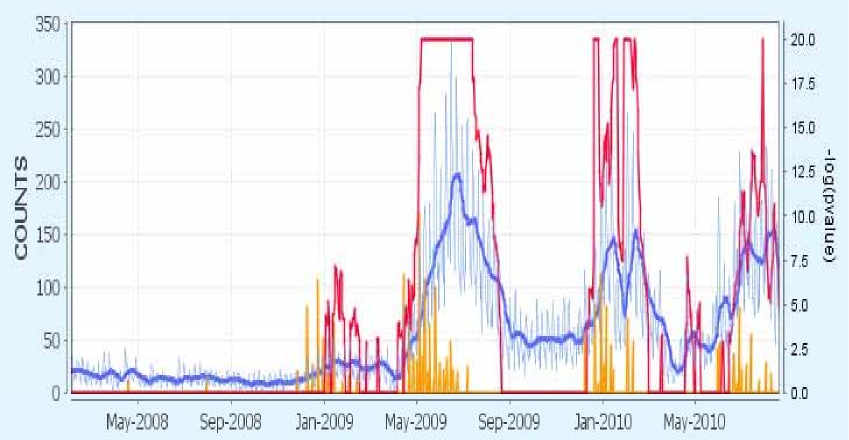
- ✓ Timely dissemination of alerts
- ✓ Rapid response and mitigation of crises

- ✓ Reliable advanced analytics & Intuitive, highly interactive interface
- ✓ Automation of routine screenings & Support of manual evaluations by human experts

Respere | ncbs | Auton Lab | RTBI | Sarvodaya | IDRC CRDI | LIRNEasia | Canada

TCWI: Statistical tool

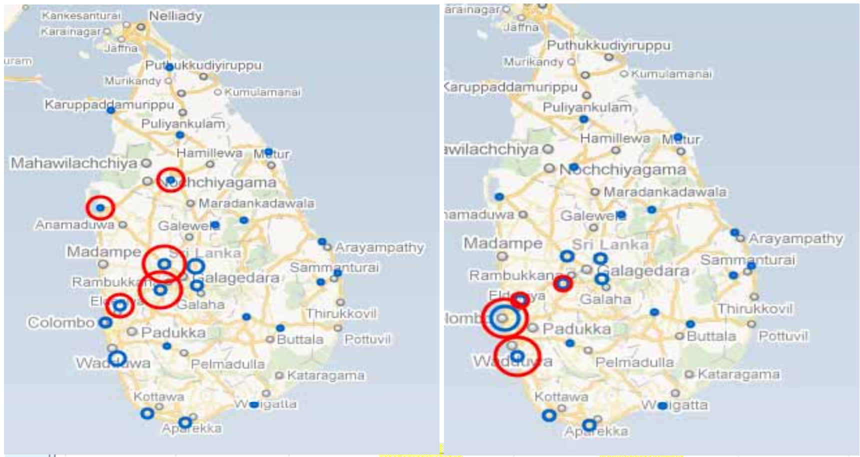
Example: Dengue Fever in Sri-Lanka

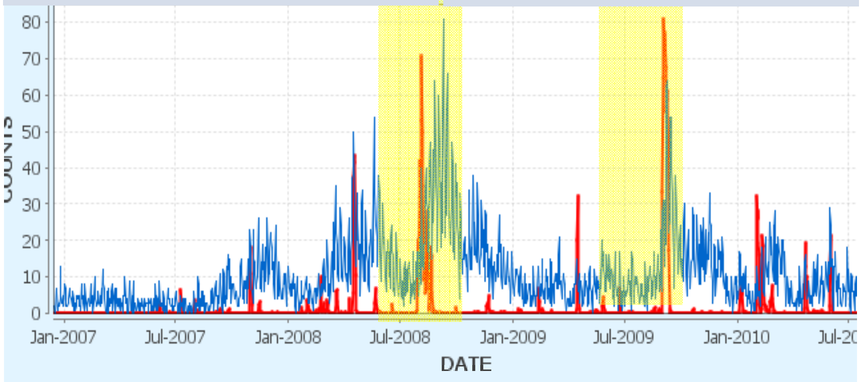


Blue: disease counts Red: temporal scan alerts Orange: CuSum alerts

- Dengue fever outbreaks in Sri-Lanka in 2009 and 2010 are thought to be the worst in history
- The one in 2009 amounted to 35,007 cases and 346 deaths
- TCWI would have issued warnings in early 2009 about that year event, when dengue cases just began to escalate, and it would have continued to issue alerts through the outbreak period. Early warning would have given health officials more time to prepare response and to mitigate consequences

Example: Leptospirosis in Sri-Lanka





Blue: disease counts Red: Spatial Scan Alerts

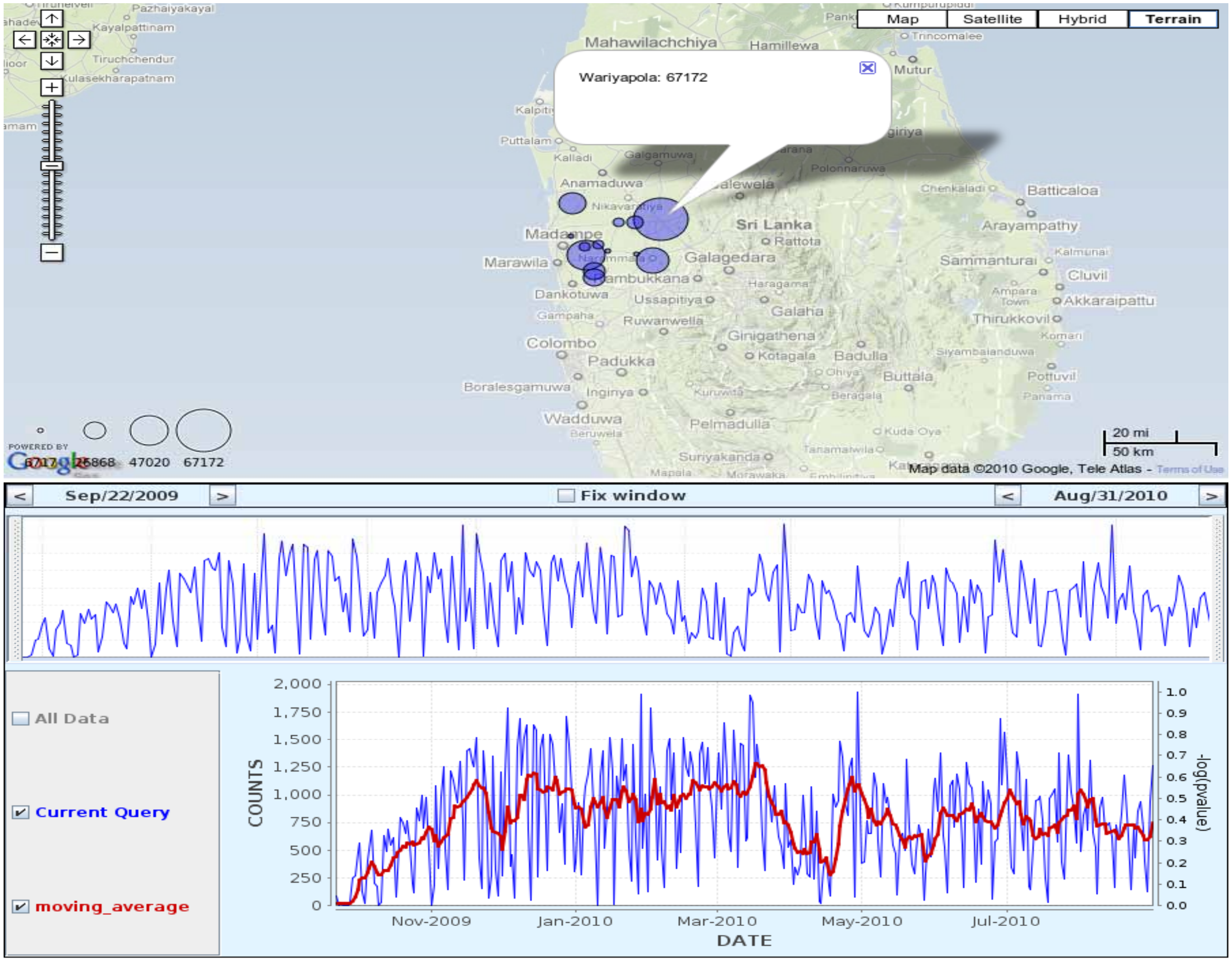
- Our spatial scan analysis of Leptospirosis counts in mid 2008 and late 2009 revealed spatial clusters
- This type of insight could hardly be obtained using traditional data collection and analysis methods

TCWI: Dynamic spatio-temporal visualization

Real-time spatial and temporal visualization of large-scale data

Highly interactive user interface

Executes complex queries and runs statistical tests instantaneously



Wariyapola: 67172

< Sep/22/2009 > □ Fix window < Aug/31/2010 >

□ All Data

☑ Current Query

☑ moving_average

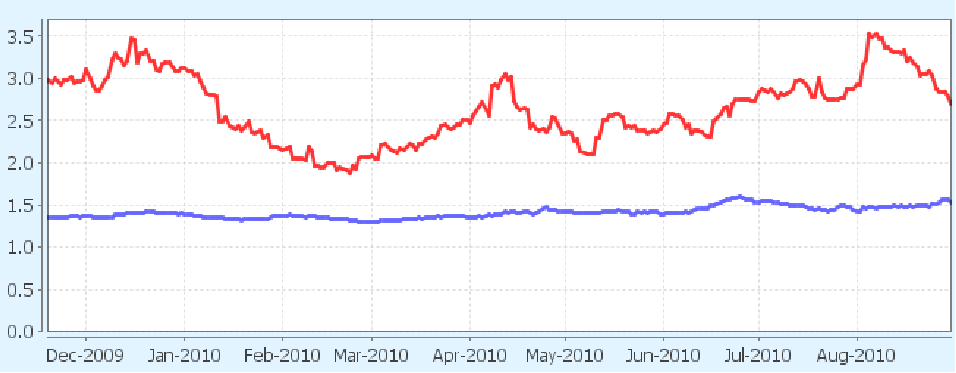
Disease data collected from Kurunegala district in Sri Lanka analyzed and visualized using TCWI

Additional benefits

Ability to monitor all diseases and syndromes

- Real-Time Biosurveillance Program data collection and analytic capabilities allow to monitor many more diseases than before
- Reported signs and symptoms enable syndromic surveillance
- The data collected may contribute to research of emerging, non-notifiable, as well as chronic diseases

Example: Hypertension in Sri-Lanka: A gender division pattern



Blue: all patient visits Red: Hypertension cases

Female / male ratio based on data from RTBP

- Hypertension appears to be 2-3 times more prevalent in female than in male patients
- RTBP data and statistical analysis capabilities support making such discoveries

Additional benefits

Qualitatively better timeliness of reporting and analysis

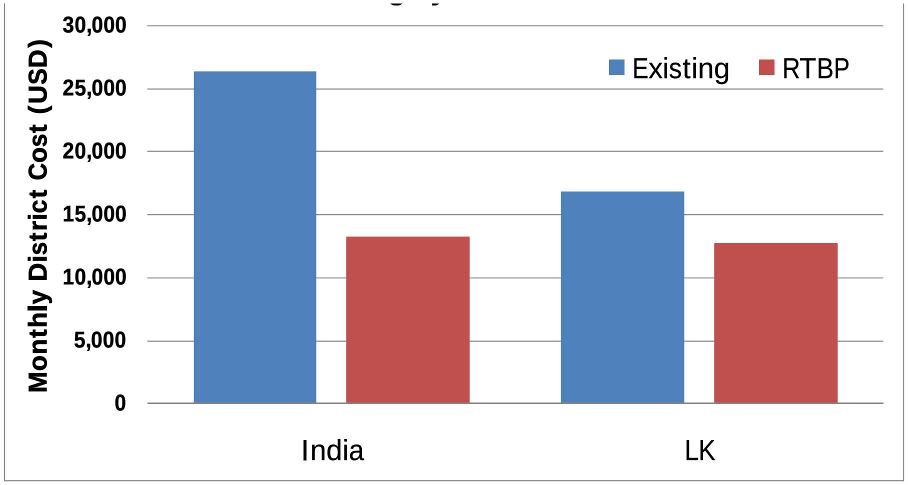
- Information updated daily as opposite to monthly+

Higher level of detail

- Case-level information vs. weekly-by-disease aggregates

Maintainability and cost-effectiveness

- RTBP relies on widely available inexpensive mobile technology
- Service of these phones is readily available even in rural areas
- Total costs of operation are lower than with the currently used paper-based notifiable disease reporting systems
- 50% and 30% cost avoidance attainable respectively in India and Sri Lanka



Another application: Mitigation of crises in food safety

Fundamental motivation: Fighting food-borne diseases

- **Burden on societies and economies worldwide**

  In the USA: 76 million cases, 320,000 hospitalizations and 5,000 deaths **each year**

- **Some cases are attributable to preventable contamination of food**

- **Sometimes evidence can be found in routinely collected data**, e.g.:
  - Records of microbial testing of samples of food taken at food factories
  - Results of regulatory inspections of food facilities
  - Results of monitoring health of animals arriving at slaughter houses
  - ...etc.

- **Data like that can be analyzed to:**
  - Detect adverse events
  - Support post-event investigations
  - Survey pathways of pathogen transmission
  - Assess effectiveness of countermeasures
  - Assess risk and reallocate resources to proactively address threats
  - ...etc.

Interactive analyses and visualizations of multiple data streams using T-Cube Web Interface (for USDA)

USDA Food Sampling Data

About **150,000 records related to *Salmonella*** across 3 years

Daily, transactional temporal resolution

Spatial resolution by **unique establishment**

Key attributes: test result (positive, negative), serotype, PFGE pattern, antibiotic resistance pattern, product type, establishment production profile

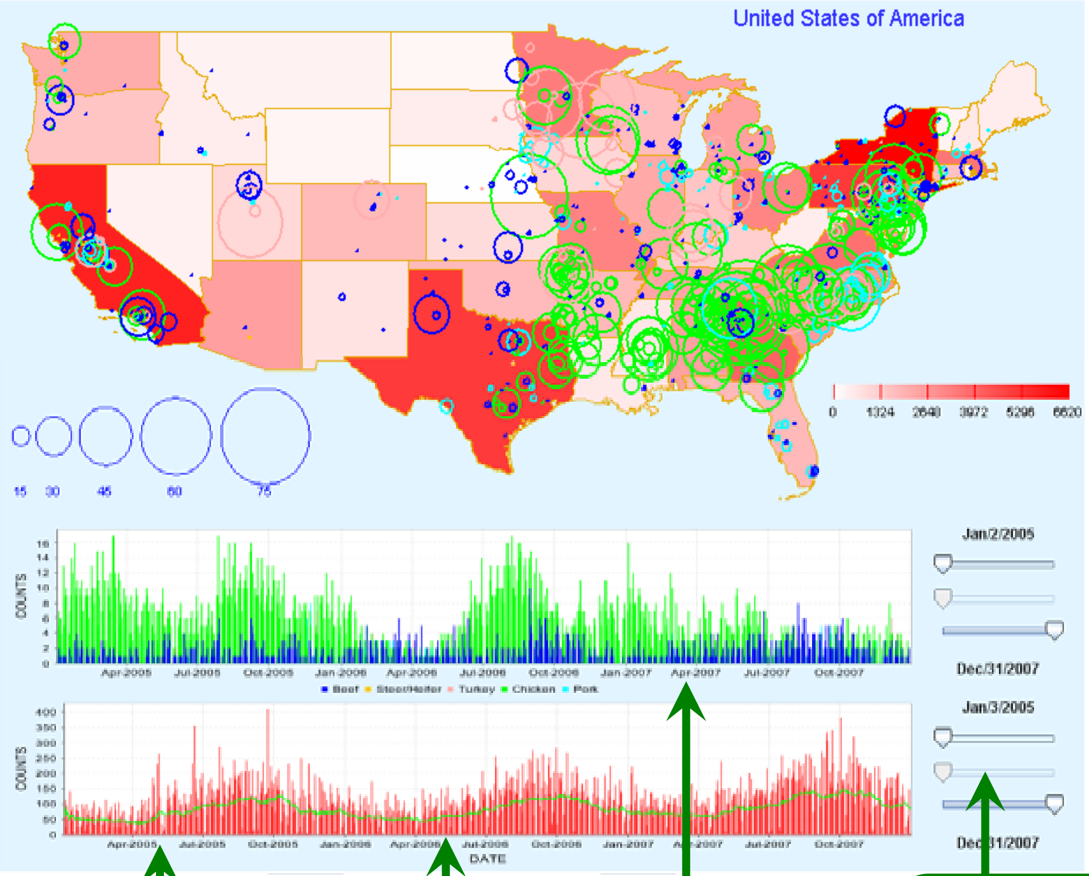
CDC PulseNet Human Illness Data

About **100,000 records related to *Salmonella*** across 3 years

Daily, transactional temporal resolution

Spatial resolution by **state**

Key attributes: serotype, PFGE pattern , outbreak ID

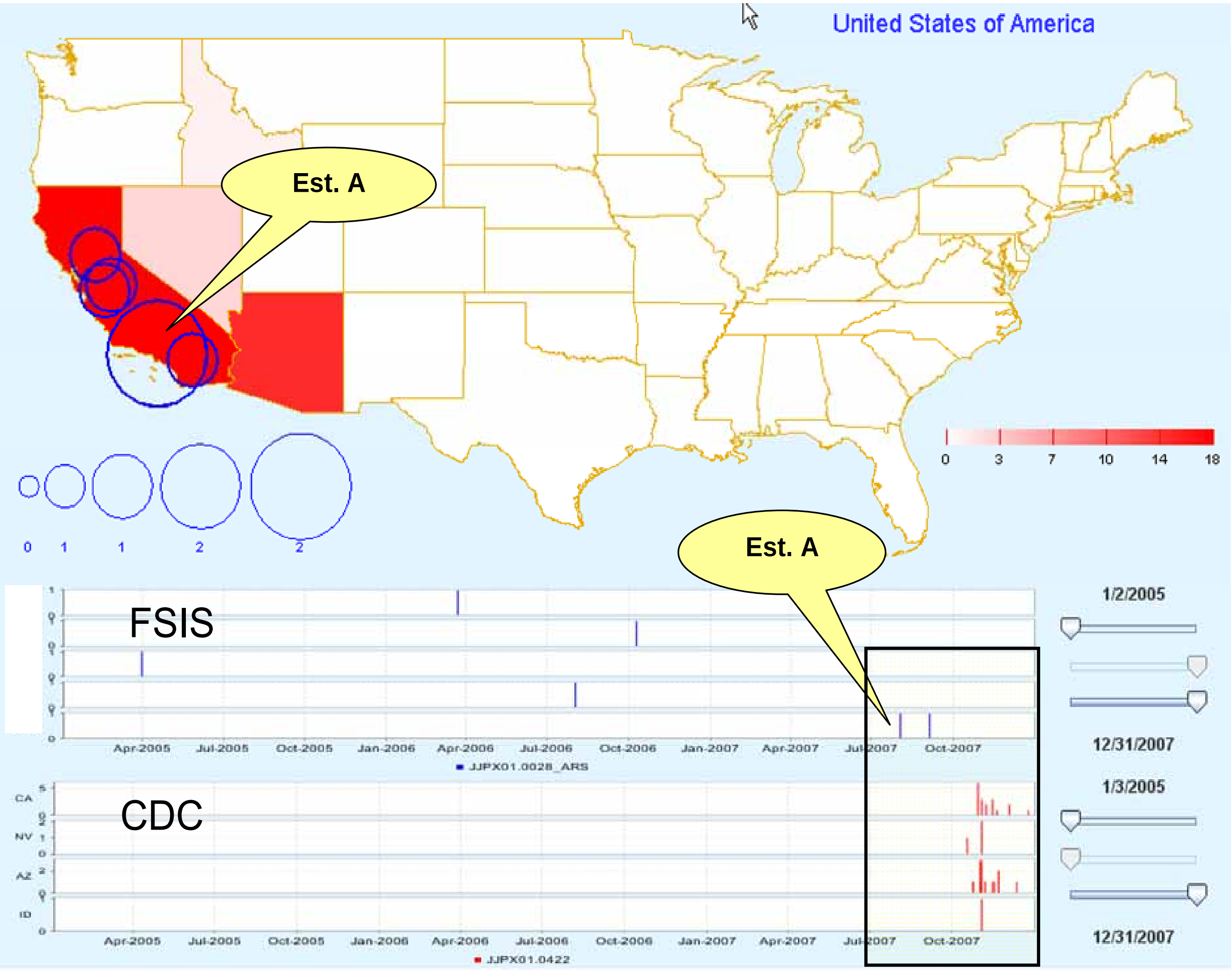


Certain statistics such as moving average are **computed/updated on-the-fly**

Color coding is used to show **multiple categories of data**

Controls such as time window sliders **make visualization interactive**

Example of a successful detection of a pattern of a potential linkage between human illness and food supply



Human health data contained a **cluster of human cases of salmonellosis** for which microbiological tests identified the same PFGE pattern

After correlating that data against **records of microbial sampling of food** at food factories, Establishment A was found to report positives with a similar PFGE immediately prior to the emergence of human cluster

It was located in the same geographic area as the human cases

Third application: Bed-side informatics

Motivation:

- Medical personnel more prone to making mistakes when on the rush

Ideas:

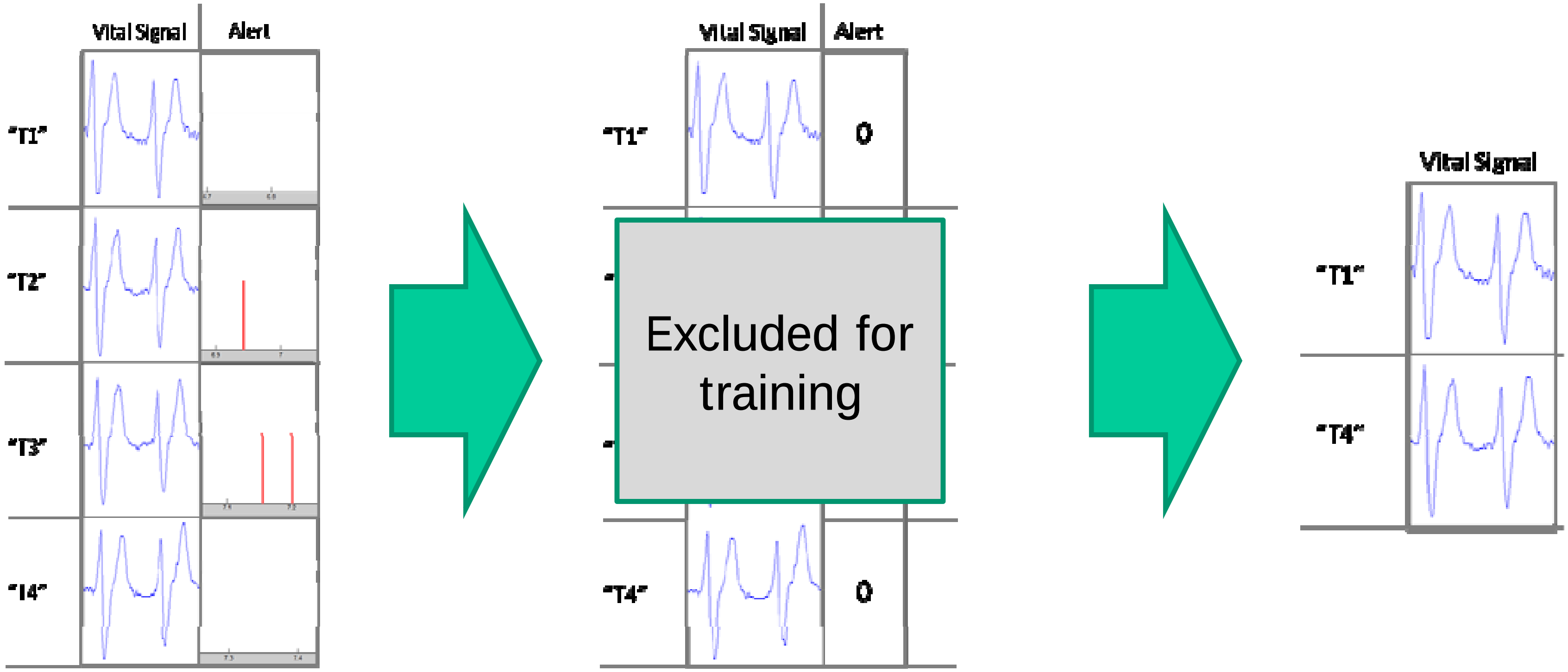
- Leverage routinely collected data to identify leading indicators of emerging health crises
- Use them to issue early warnings, giving medical personnel more time to respond

Test environment:

- Intensive Care Unit (ICU)
  - Vital signs (heart rate, ECG, oxygen intake, ...)
  - Collected at high frequencies (125 Hz in our case)

Preparation of data

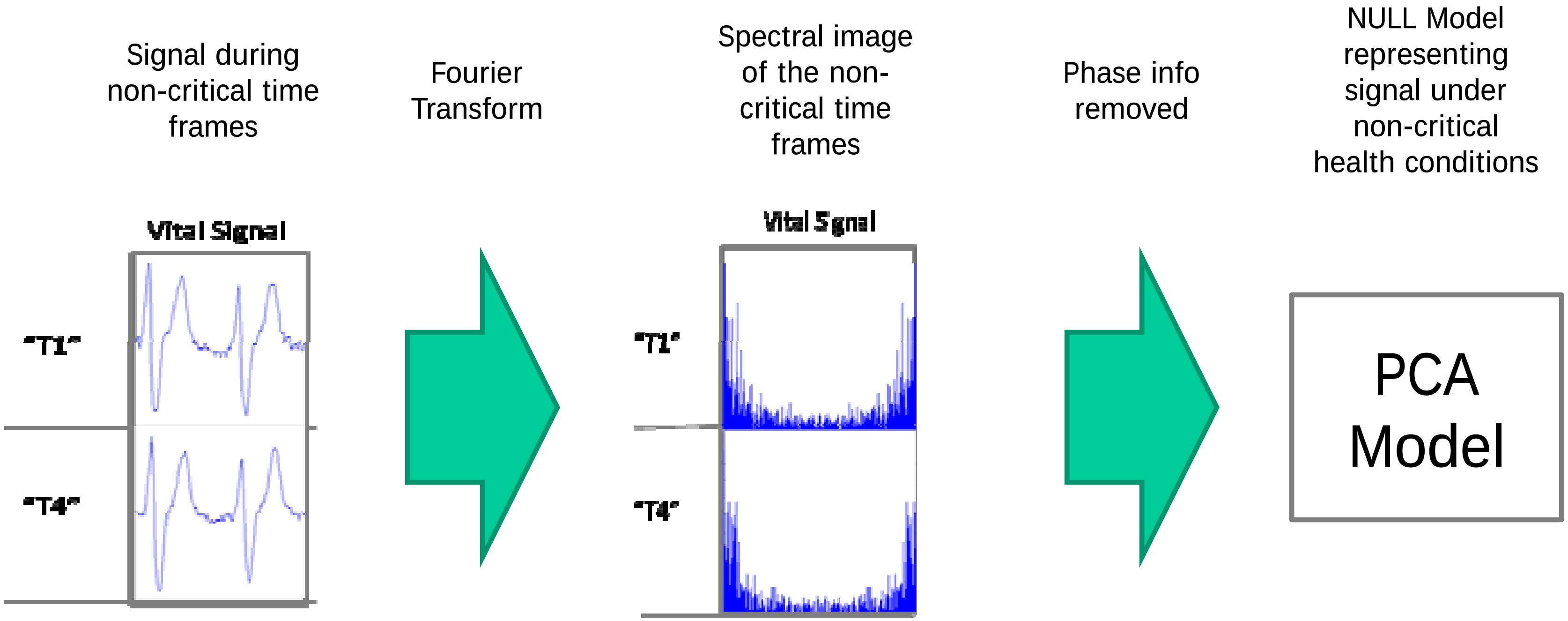
1. Each signal segmented into two-second disjoint frames
2. Each frame correlated with the presence or absence of a crisis
3. Negative examples used to learn the NULL model



Learning the model

1. **Perform Fourier Transform on each negative example to obtain its spectral representation**
2. **Learn Principal Component (PCA) model for spectral data**

   ✓ **This model reflects the expected spectral profile of a non-critical health status**



Signal during non-critical time frames

Fourier Transform

Spectral image of the non-critical time frames

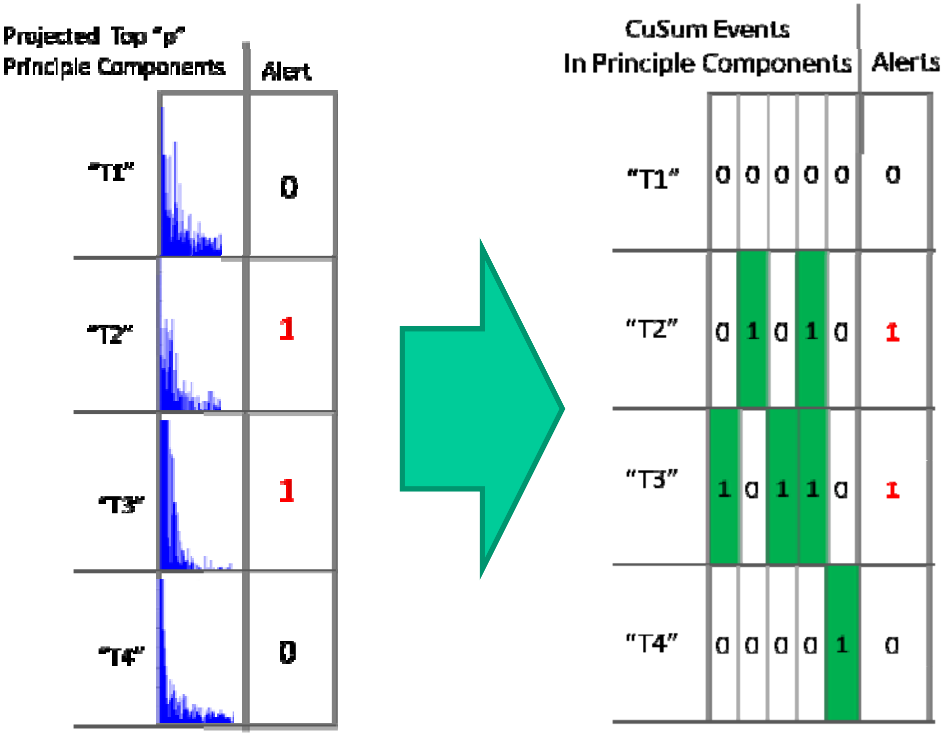
Phase info removed

NULL Model representing signal under non-critical health conditions

PCA Model

Extracting potential indicators

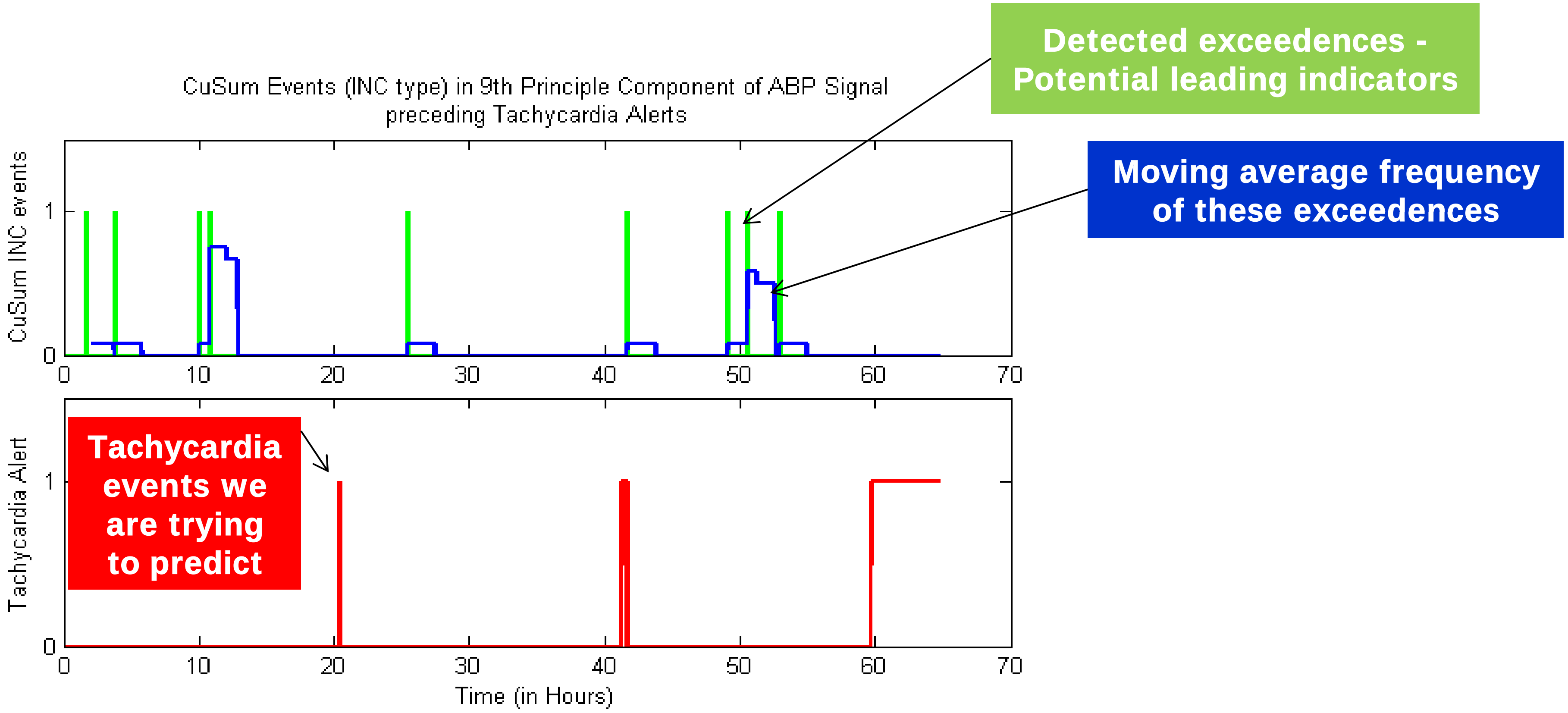
1. Consider a set of patients separate from training data
2. Segment their signals, perform Fourier transform, project onto principal components of the learned NULL model
3. Use a Control Chart (e.g. CuSum or Temporal Scan) to identify significant exceedences of the observed signal w.r.t. expected magnitudes of its principal components
4. Hypothesize that any such exceedence can be indicative of an upcoming heath crisis



	Projected Top "p" Principle Components	Alert
"T1"		0
"T2"		1
"T3"		1
"T4"		0

	CuSum Events In Principle Components					Alerts
"T1"	0	0	0	0	0	0
"T2"	0	1	0	1	0	1
"T3"	1	0	1	1	0	1
"T4"	0	0	0	0	1	0

Example result



Detected exceedences - Potential leading indicators

Moving average frequency of these exceedences

Tachycardia events we are trying to predict

Observation:

Frequency of detected exceedences of this particular type visibly increases a few hours prior to onset of Tachycardia events

More example results (from massive screening)

Lift = P(Outcome | Evidence) / P(Outcome)

We try multiple combinations of parameters (window widths, offsets, principal components, signals) to identify which exceedences precede health crises with statistical regularity

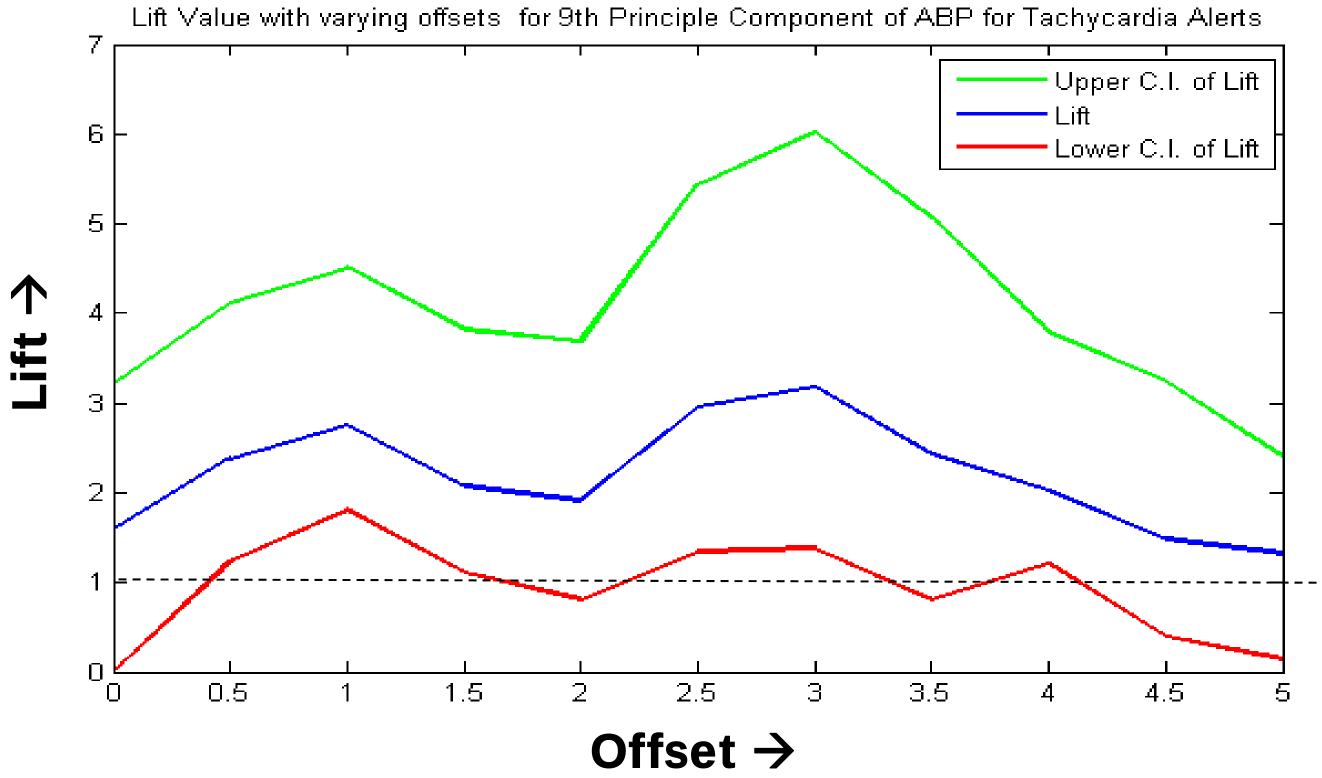
→ Several thousand trials per signal (scalability is a must!)

Principal Component	Trend	CuSum Threshold	Lift (CI)	Evidence Window (Minutes)	Outcome Window (Minutes)	Offset (Minutes)
5	Up	3	3.887 (1.420 6.964)	15	45	15
9	Up	3	3.143 (1.959 5.381)	30	15	60
5	Up	7	3.111 (1.761 5.979)	60	60	60
9	Up	3	3.009 (1.834 5.028)	45	15	60
1	Down	3	2.978 (1.309 4.811)	45	15	60
1	Down	3	2.762 (1.255 4.559)	45	30	60

A sample of significant results from the Lift Analysis of Arterial Blood Pressure (ABP) for a set of patients experiencing episodes of Tachycardia

More example results

- **This graph depicts observed Lift as a function of offset between the evidence and outcome windows (for one of the principal components)**
- **Lift peaks at above 3.0 around 2.5 to 3 hours ahead of Tachycardia episodes**



Evidence Window: 45 minutes

Outcome Window: 45 minutes

Conclusion

☺ **We can develop and deploy societally-beneficial analytic systems**

- ✓ **We have seen examples in many domains, including Healthcare Informatics**

But, how to ensure their long term sustainment…

→ **Organically public ownership model?**

→ **Public-private partnership?**

→ **Commercial service delivered to the public sector?**

→ **… ?**

Candidate theme for a large-scale Indo-US collaboration

Comprehensive Environment for Health Surveillance

Objectives?

1. To study:
   - All conceivable sources of information
   - All conceivable modalities for data acquisition
   - All conceivable usage modalities
     - Public health / Epidemiology / Environmental medicine
     - Chronic / Emerging diseases
     - Clinical Health Care
     - Food / Agriculture
     - ...
2. To develop:
   - Analytics capable of delivering useful information in the most palatable manner
   - Principled, scalable infrastructure to ensure comprehensiveness of solutions
   - Pathways to deployment (technical, policy, culture aspects)
   - ...

Thank you!

Acknowledgements: **USDA**(1040770), **CDC**(R01-PH000028), **NSF**(IIS-0325581,IIS-0911032), **IDRC**(Canada,105130)