Automated Detection of Data Entry Errors in a Real Time Surveillance System

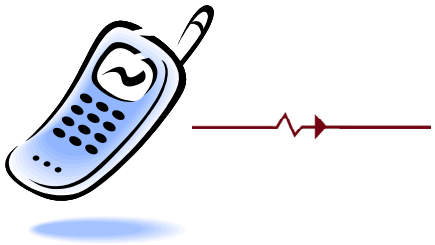
Lujie Chen
Artur Dubrawski
Auton Lab, Carnegie Mellon University

Nuwan Waidyanatha
Chamindu Weerasinghe
LirneASIA, Colombo, Sri Lanka

Presenter:
Maheshkumar Sabhnani
Auton Lab, Carnegie Mellon University

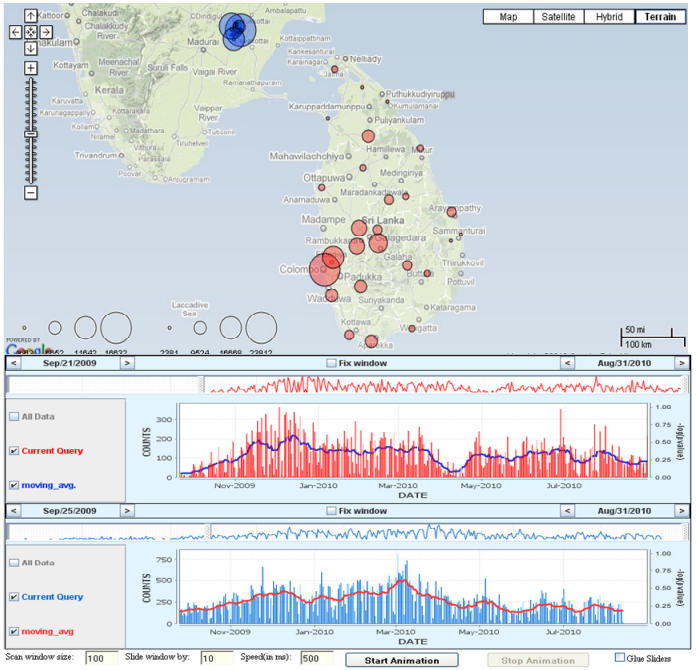
Context: Real-Time Biosurveillance Program (RTBP)

COLLECT

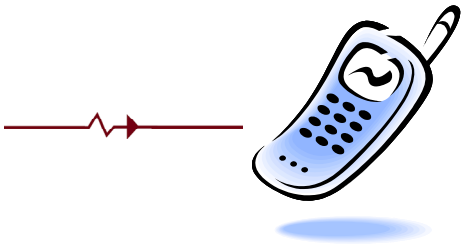


- ✓ **No need for sophisticated infrastructure**
- ✓ **Affordable setup and inexpensive maintenance**

DETECT

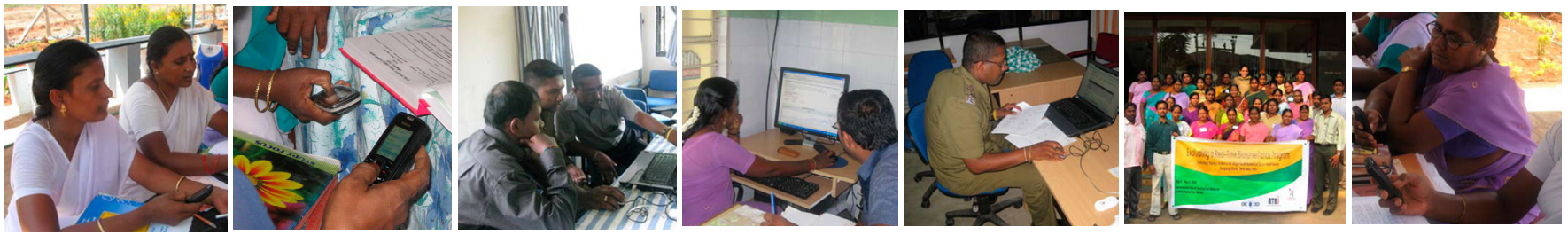


ALERT



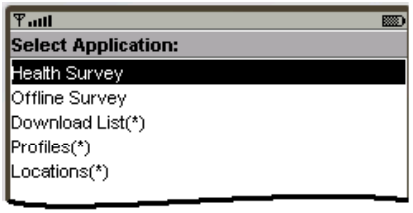
- ✓ **Timely dissemination of alerts**
- ✓ **Rapid response and mitigation of crises**

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

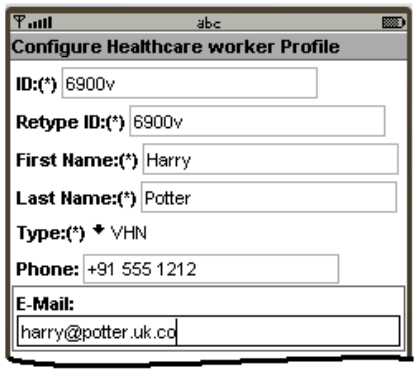


IS Respere | ncbs | Auton Lab | RTBI | Sarvodaya | IDRC CRDI | LIRNEasia | Canada

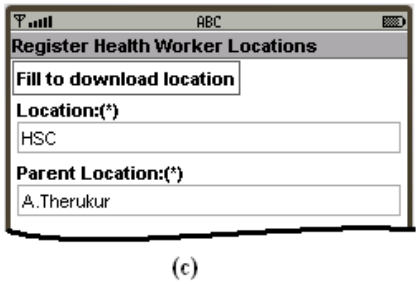
Data entry using Java-enabled mobile phones



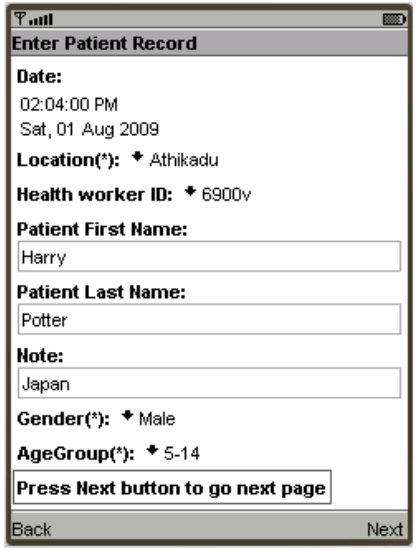
(a)



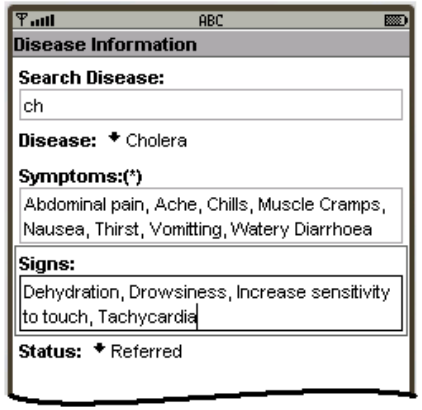
(b)



(c)



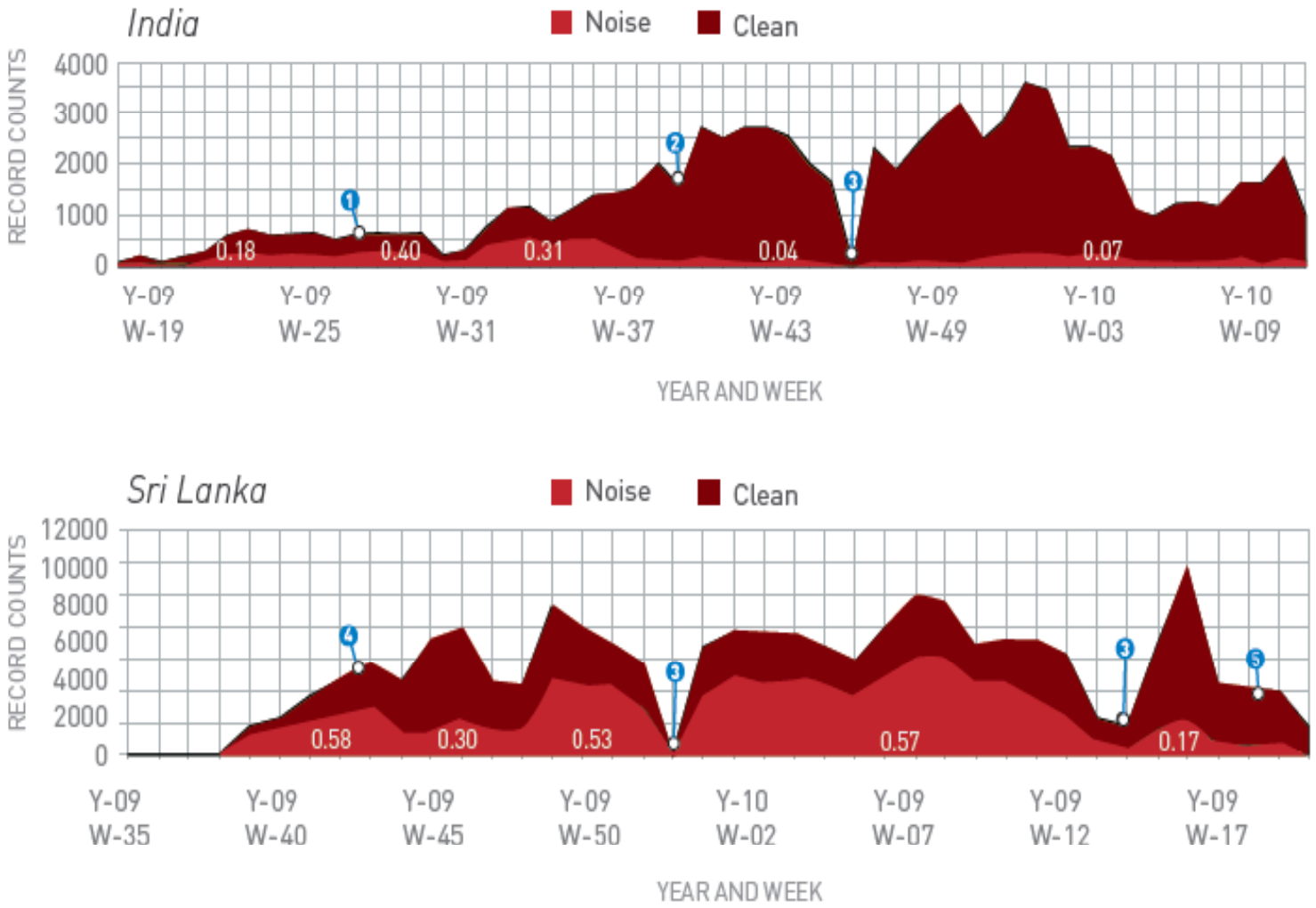
(d)



(e)

- **mHealthSurvey application is a pull-down menu driven interface designed for entering case data**
- **It does nor require fancy phones or technically savvy personnel**
  - The duty in India performed by village nurses (medically trained, but initially not always fluent in mobile phone usage)
  - In Sri Lanka – by low cost personnel with limited medical knowledge, but technically able
- **Practical utility of RTBP hinges on reliability of data**

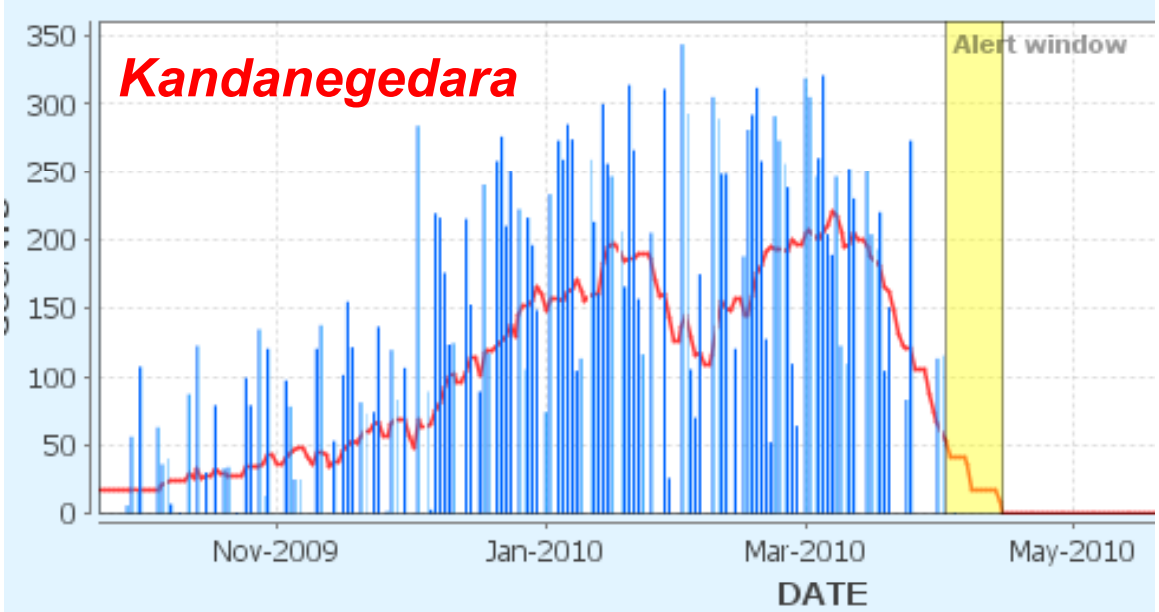
Fidelity of the digitized data



Some of the errors can be addressed relatively straightforwardly by post-processing of text

PROBLEM	EXAMPLE
Use of synonyms	goal fever =jail fever = typhus fever dementia = memory loss enteric fever = typhoid fever; encephalitis = meningitis
Inserting symbols and extra spacing between words	body ache =body-ache, body pain = body pain
Changing the order of words	muscle weakness = weakness in muscle stomach pain = pain in the stomach
Inclusion and exclusion of adjectives	'severe' memory loss vs memory loss
Using local language when terms are unknown	leg vettuthal (Tamil) = broken leg (English)
Using preposition and conjunctions between different terms	nasal stuffiness *or* sneezing over bleeding *with* abdominal pain
Long sentences	not able to identify color white and shining patches without any sense
Mistaking treatment for diagnosis	oral pils, remove catheter, vaccination
Prepopulated instructions in text boxes propagating to database"	please specify details specify symptoms
UK vs USA spelling	diarrhoea = diarrhea vomiting = vommitting
Test results as symptoms or signs	BP 140/90, BP 120/100
Singular vs plural	fit / fits , cut / cuts
Inconstancies in the verb tense	faint, fainted, fainting

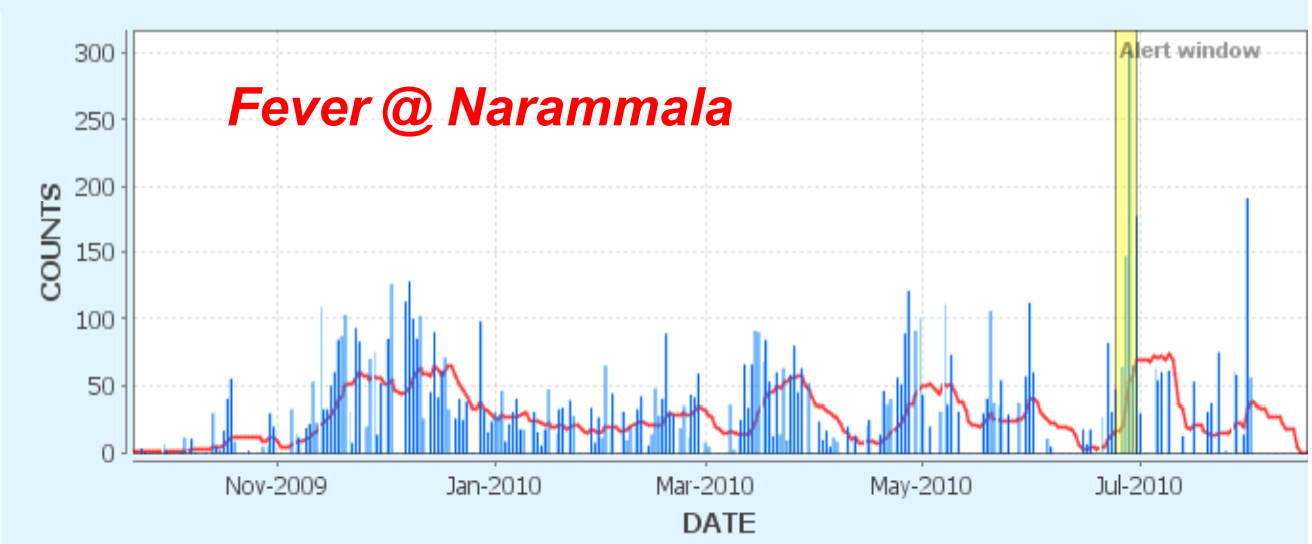
One type of a systematic problem



- This graph depicts time series of daily counts of disease records collected at one particular location

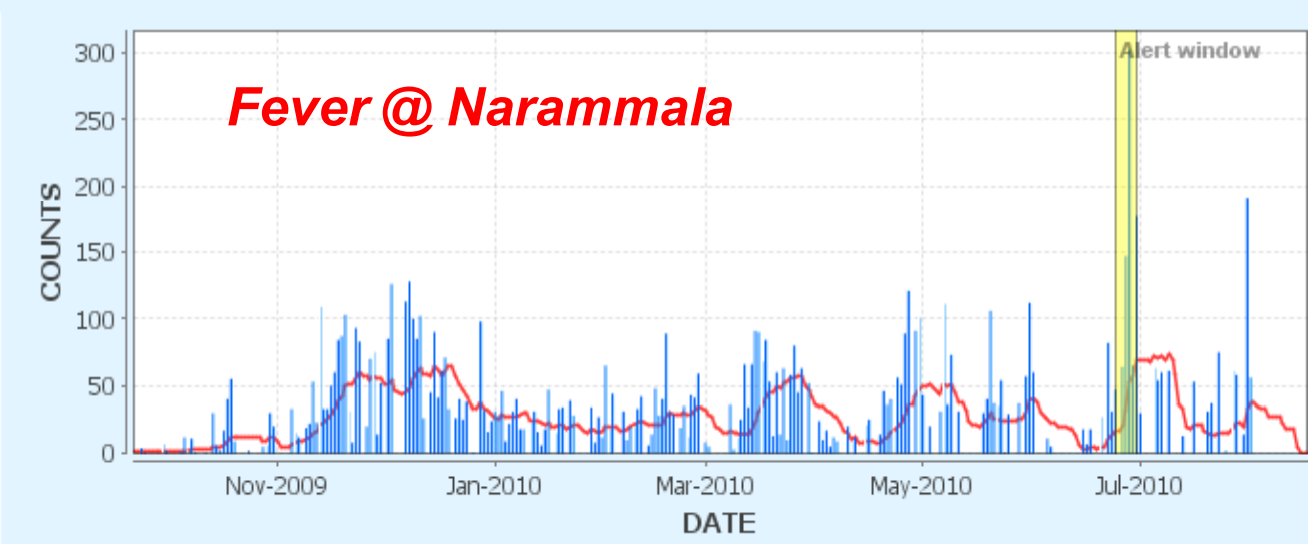
- Luckily, a sudden drop in data quantity can be recognized using one of the standard event detection algorithms (such as CuSum or temporal scan)
- This type of irregularity, if not accounted for, could bias baselines and artificially lower expectations of disease rates, potentially leading to increased rates of false positive alerts in the future.

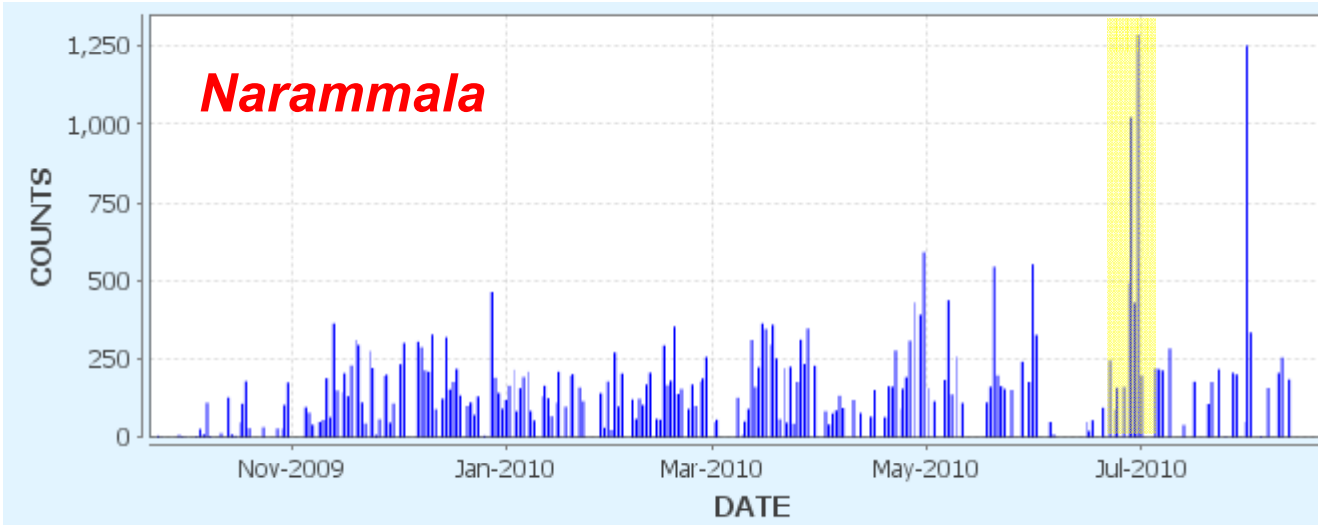
Another type of issue: Batch data entry



- RTBP detected an unexpected escalation of fever cases in one city by end of June2010
- However, the total count of all diseases also has peaked at that location at the same time

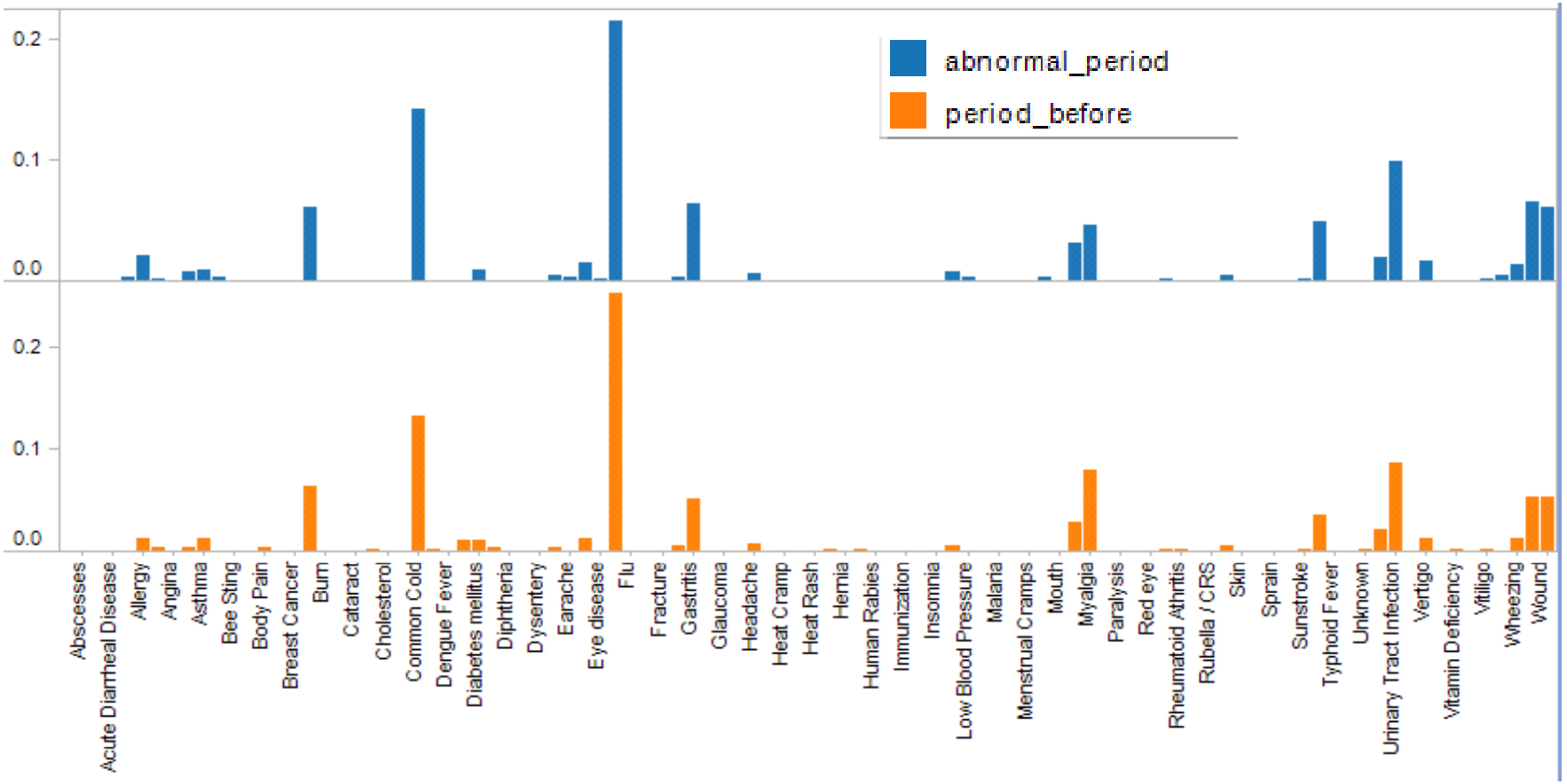
Another type of issue: Batch data entry





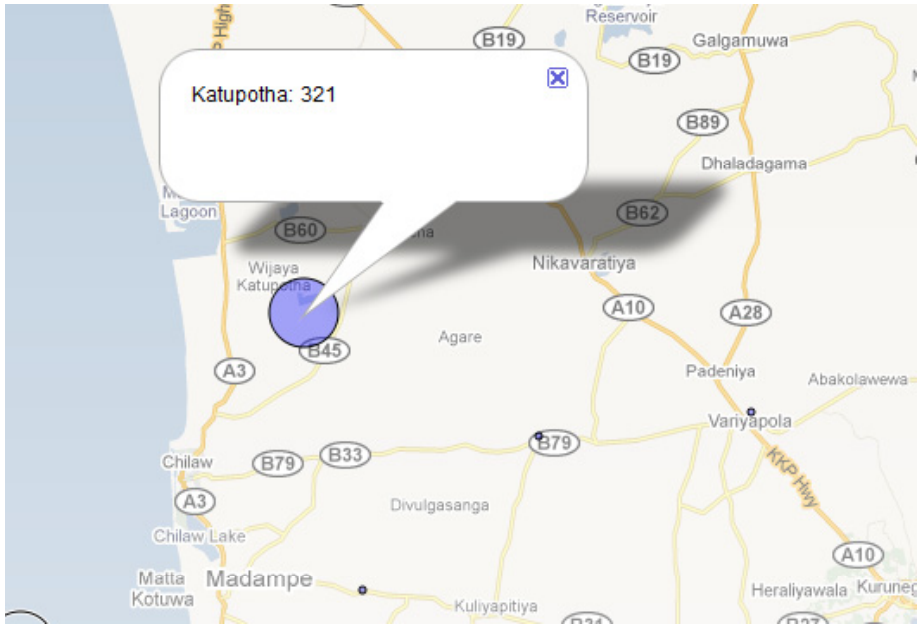
- RTBP detected an unexpected escalation of fever cases in one city by end of June2010
- However, the total count of all diseases also has peaked at that location at the same time
- Turns out, the data was batch-entered with the date of entry incorrectly replacing the date of patient's visit

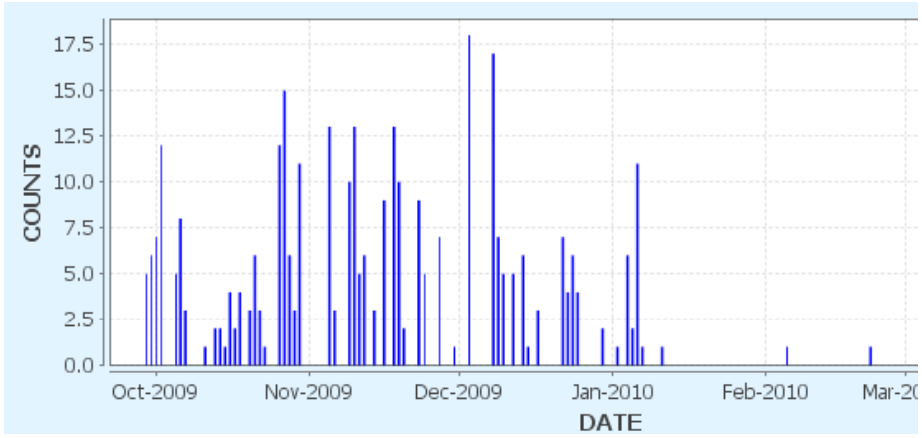
Batch data entry



- That hypothesis can be further confirmed by comparing disease rate distributions observed during the period of suspicion and in the past
- They are almost identical
- This confirms the nature of the detection to be a data entry problem, not a disease outbreak

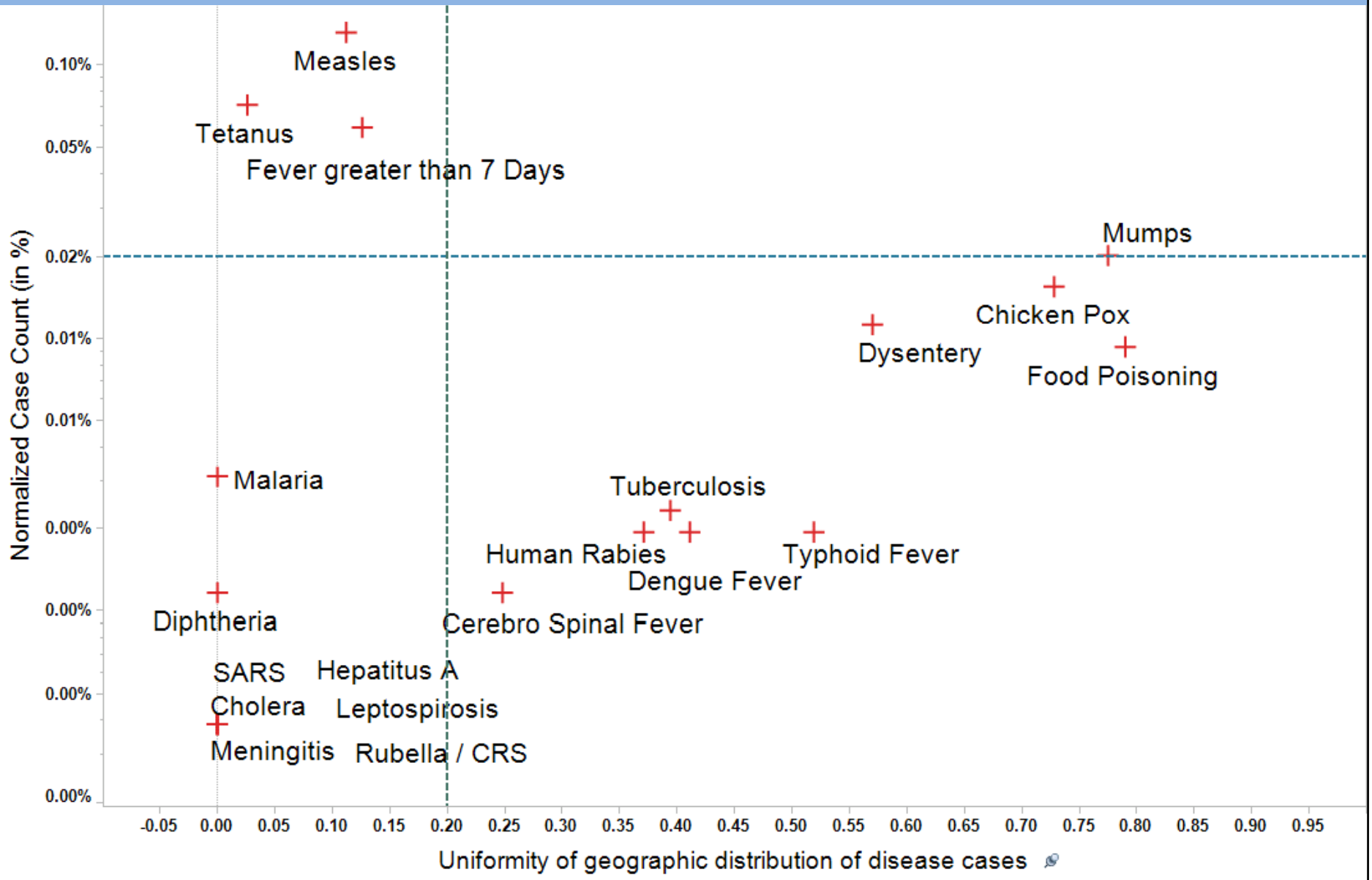
Miscoding





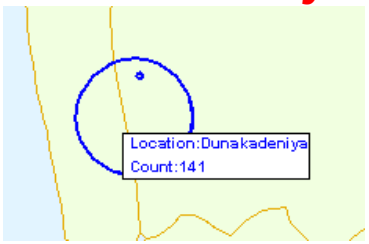
- **RTBP reveals 340 cases Measles coming almost exclusively from a single location**
- **Measles is a notifiable disease in Sri Lanka and it is thought to be rather rare**
- **Can we possibly face a real outbreak?**
- **It turns out that the data entry staff meant to enter “muscle pain”, not “measles”**
- **How could this type of error be automatically flagged?**

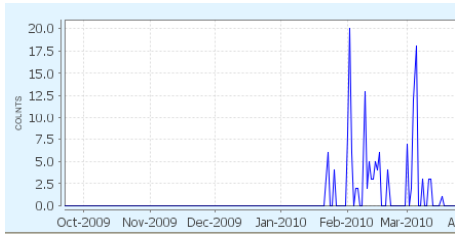
"Disease Map" for notifiable diseases in Sri Lanka



Items from the upper left quadrant of the disease map

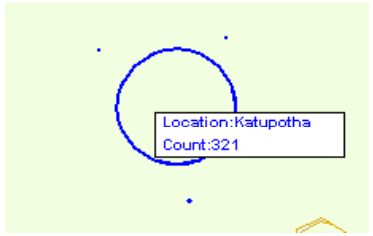
Fever > 7 Days

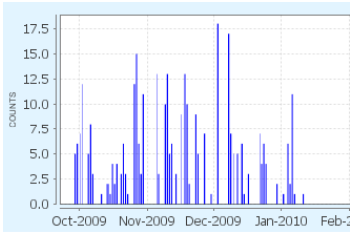




Over 150 cases escalated in February and March of 2010, affecting primarily a single location, during the non-rainy season, which was not so common. But, RTBP personnel was unable to identify any data fidelity issues. It was a legitimate health event.

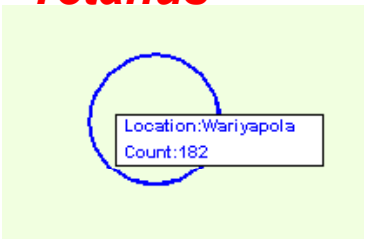
Measles

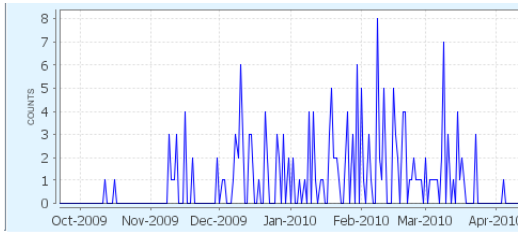




The likelihood of a measles outbreak emerging in a single location without spreading to other areas is low. The data was miscoded. The true diagnosis was "muscle pain".

Tetanus

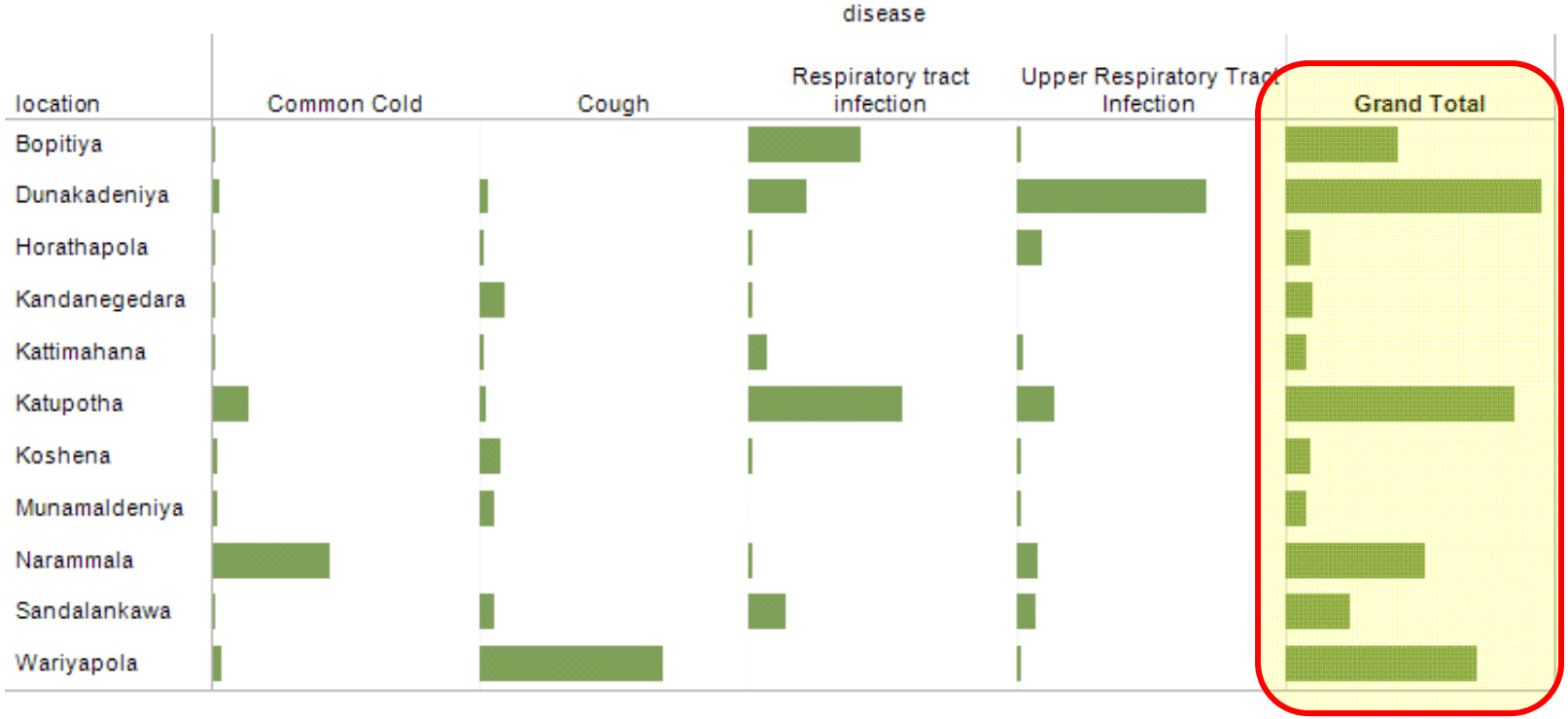




The assistant entering the data had submitted data for Toxide vaccination as if it represented Tetanus disease cases.

Another issue: Biases in preliminary diagnoses

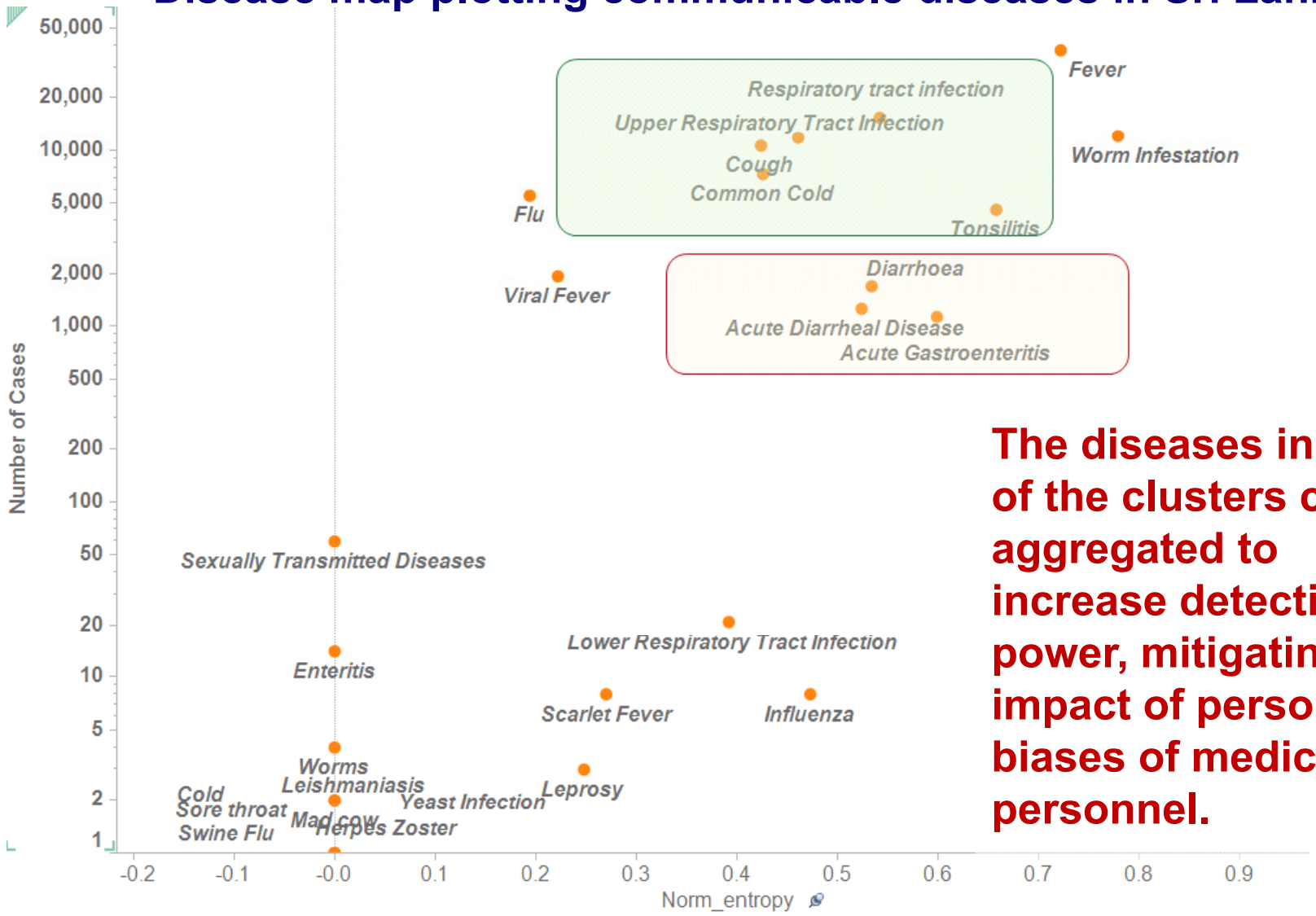
We observe variance in doctors preferences regarding issuance of preliminary diagnoses of certain diseases which manifest in similar way, for example flu-like diseases:



This issue may dilute signals and reduce our ability to quickly detect emerging outbreaks.

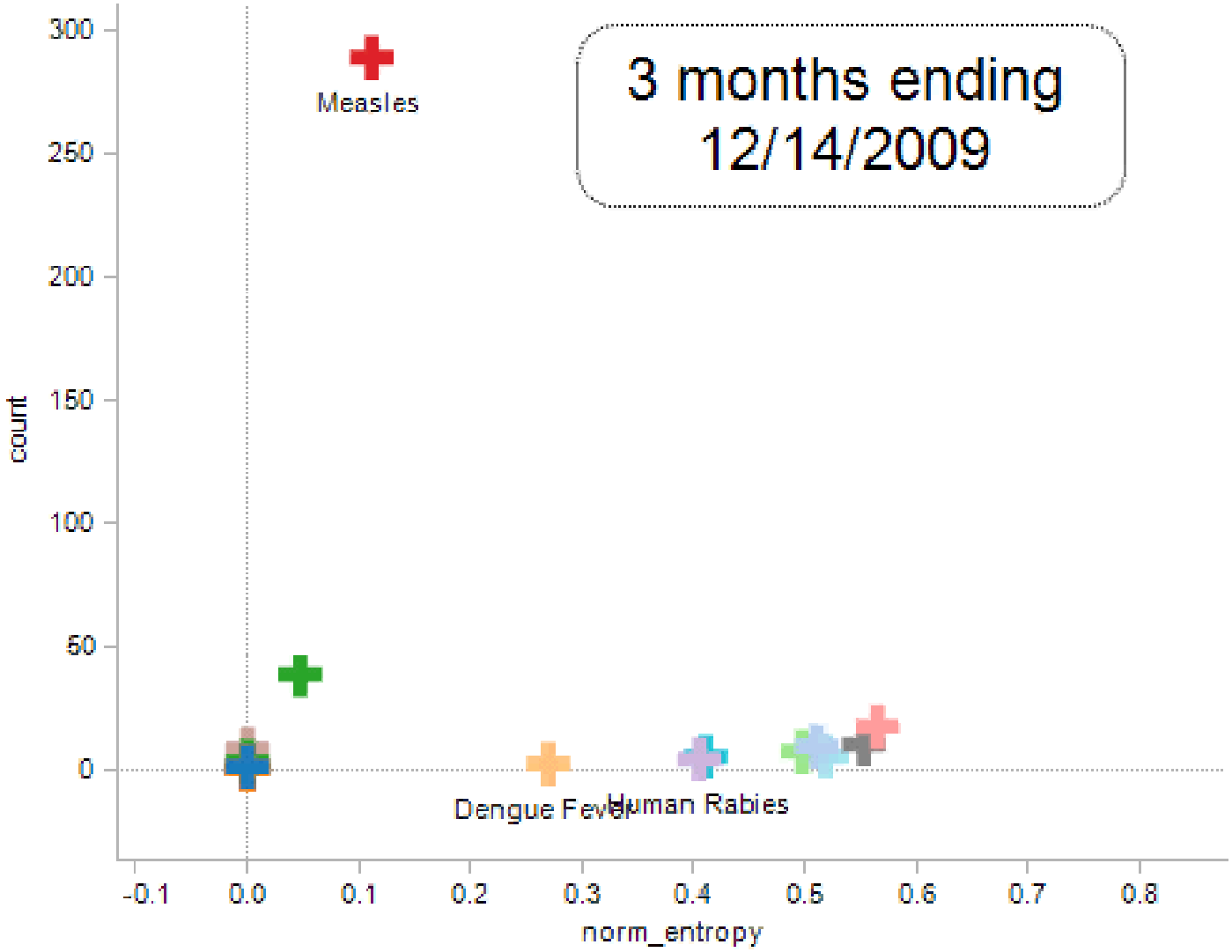
Biases in preliminary diagnoses

Disease map plotting communicable diseases in Sri Lanka



The diseases in each of the clusters can be aggregated to increase detection power, mitigating the impact of personal biases of medical personnel.

Disease map in motion



Summary

- **Low quality of data can easily invalidate health surveillance and epidemiological analyses**
- **Implementers of biosurveillance systems must be prepared to face challenges of data quality, especially in environments with limited resources**
- **Automated algorithms capable of detecting systematic data quality issues can be of help**
- **Using a few practical scenarios, we have shown how to automatically detect certain types of avoidable systematic data entry errors in support of real time health surveillance**
- **We are developing a toolkit to guide agile management of health data quality**