

## DISCOVERING POSSIBLE LINKAGES BETWEEN FOOD-BORNE ILLNESS AND THE FOOD SUPPLY USING AN INTERACTIVE ANALYSIS TOOL

Artur Dubrawski		Auton Lab
Lujie Chen		Auton Lab
<u>Robin Sabhnani</u>		<u>Auton Lab</u> ←
Paula J. Fedorka-Cray		USDA ARS
Lynda Kelley		USDA FSIS
Peter Gerner-Smidt	CDC	
Ian Williams		CDC
Mark Huckabee		SAIC
Adrienne Dunham		SAIC

CarnegieMellon

Auton  
Lab

CDC

SAIC  
From Science to Solutions

USDA

International Society for Disease Surveillance  
Eighth Annual Conference, Miami, Florida

Dec 4, 2009

## MOTIVATION

- Salmonellosis is a common food borne illness
  - US: 40,000 cases annually including 600 deaths,
  - Many of them due to consumption of contaminated food
- Insights can be gained by monitoring routinely collected microbiological data for indications of possible causal links between Salmonella isolates in food and outbreaks of human illness.
- Feasibility of such analyses is currently limited
  - Disparateness of the relevant data sources,
  - Limited availability of investigative resources.

## GOALS OF THIS PRESENTATION

- To showcase tools being developed to **improve effectiveness of monitoring and trace-back investigations**.

These tools enable:

- Interactive navigation through multiple streams data
  - **Massive screening for patterns of interest** ← today's focus
- To review examples where **joint analyses** of food safety and public health data from different sources is beneficial
  - To show that computationally efficient analytics can help with food safety monitoring and investigations, even if the available **data is complex, voluminous, and multidimensional**
  - To vouch for the idea of **sharing data** between departments and jurisdictions

## HIGHLY INTERACTIVE VISUAL ANALYTICS WITH THE T-CUBE WEB INTERFACE

### USDA Food Sampling Data

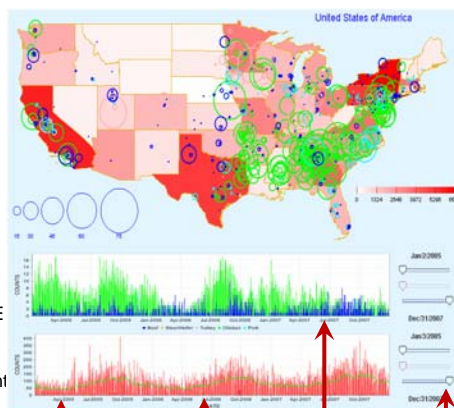
About **150,000 records** related to *Salmonella* across 3 years

**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

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



### CDC PulseNet Human Illness Data

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

**Daily, transactional** temporal resolution

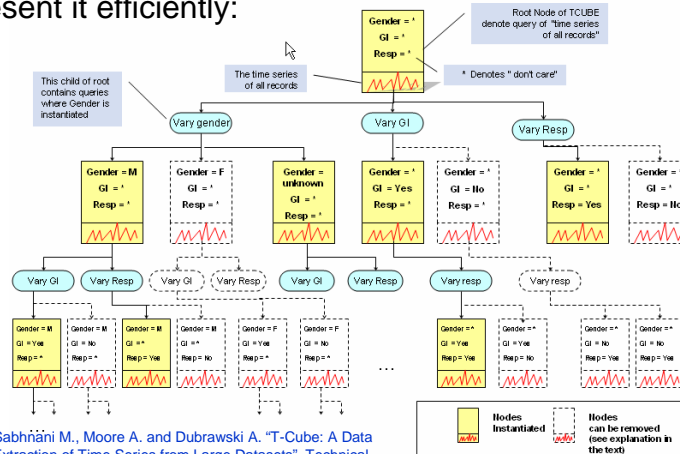
Spatial resolution by **state**

**Key attributes:**  
serotype, PFGE pattern, cluster code, product type, source type, source agency

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

## T-CUBE: EFFICIENT REPRESENTATION OF MULTI-DIMENSIONAL TIME SERIES OF COUNTS

- If we have a set of additive time series annotated with multiple categorical descriptor variables, we can represent it efficiently:



## T-CUBE: EFFICIENT REPRESENTATION OF MULTI-DIMENSIONAL TIME SERIES OF COUNTS

- T-Cube substantially reduces wait time for responses to complex ad-hoc time series queries.

The gains:

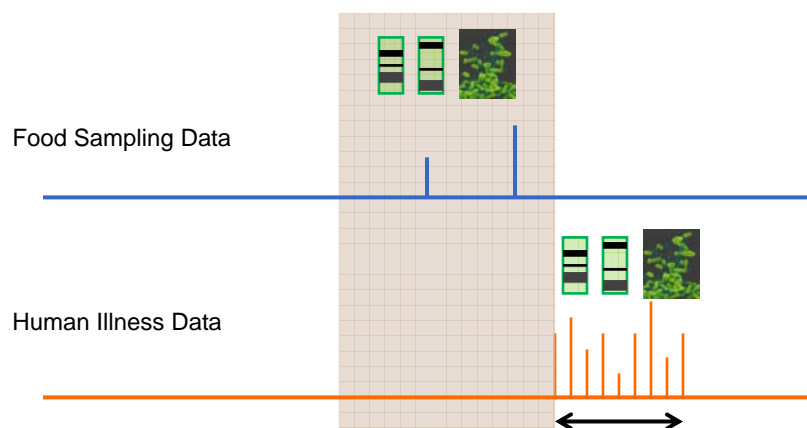
- Support interactive visualizations
  - Support rapid drill-downs, summarizations (roll-ups)
  - Support data-intensive analyses
  - Enable comprehensive monitoring of highly dimensional data for indications and tracking of known, emerging or unexpected patterns.
- Combined benefits improve attainable situational awareness, and they can make crisis monitoring and trace-back investigations more effective.

## SEARCHING FOR POTENTIAL LINKAGES BETWEEN FOOD AND HUMAN ILLNESS

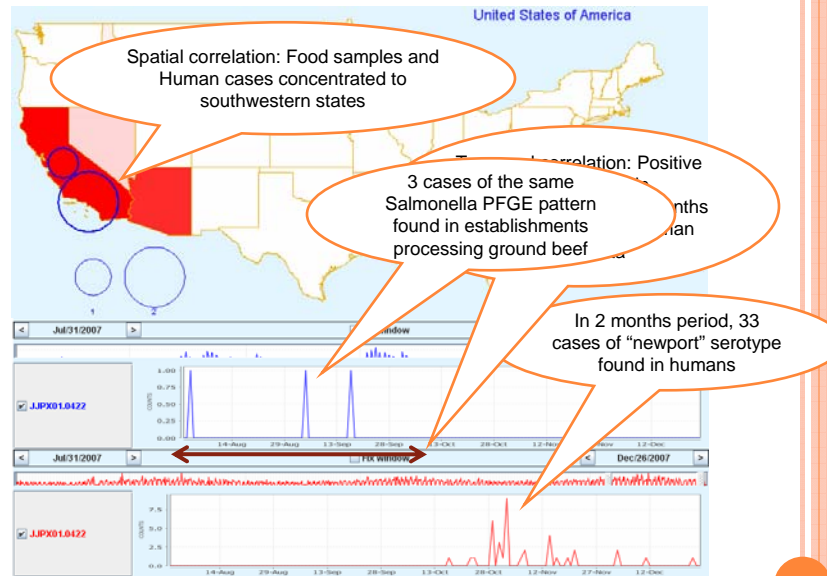
- USDA and CDC analysts are on the constant outlook for any evidence that can attribute identified clusters of human illness to sources in food supply
  - Lessons learned from identified, epidemiologically confirmed, and understood linkages, inform avoidance of future problems
- Some of those efforts can be supported with data-driven analytics
- T-Cube Web Interface has already proven utility in facilitating manual searches for potential linkages
- It also implements a few automated search algorithms identified by the analysts as the most useful:
  - Cluster pattern
  - Pattern not-in-cluster
  - Rare pattern
  - Recent pattern spike

### (1) CLUSTER PATTERN

- For outbreaks found in the Human Illness data, is there linkage evidence to establishment inspections?

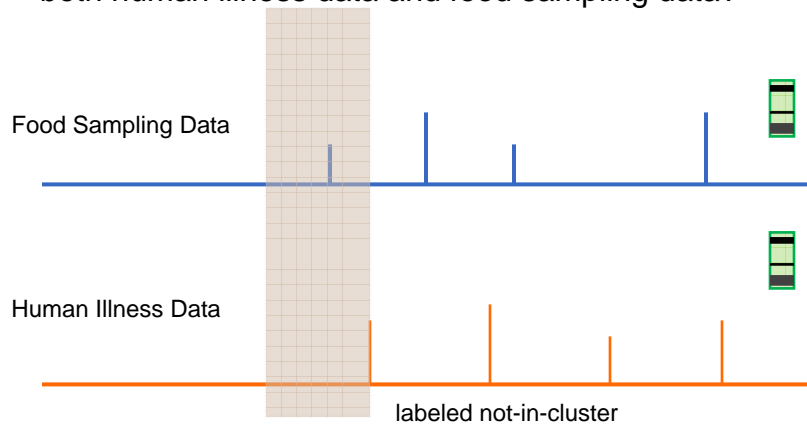


## (1) CLUSTER PATTERN: EXAMPLE



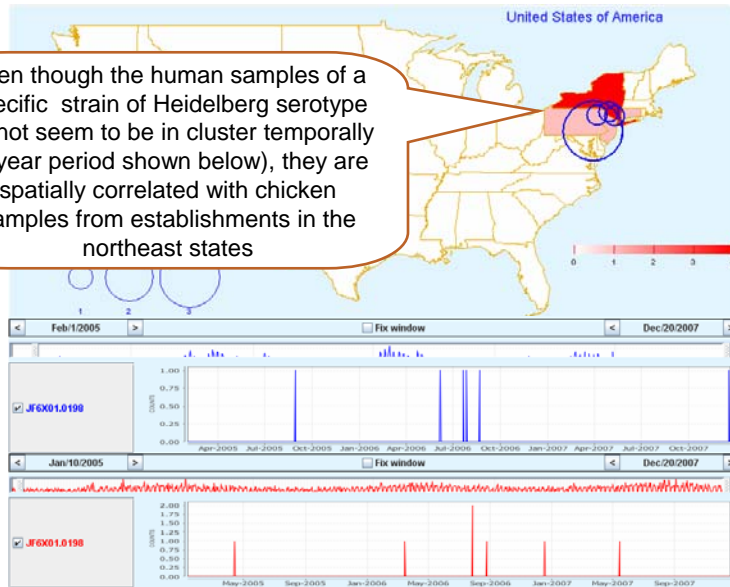
## (2) PATTERN NOT-IN-CLUSTER

- For records **not labeled as clusters** in Human Illness data, are there any PFGE patterns that were found in both human illness data and food sampling data?



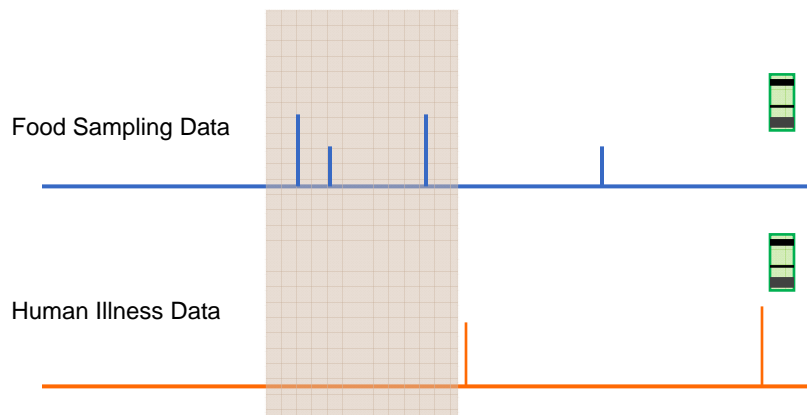
## (2) PATTERN NOT-IN-CLUSTER: EXAMPLE

Even though the human samples of a specific strain of Heidelberg serotype do not seem to be in cluster temporally (3 year period shown below), they are spatially correlated with chicken samples from establishments in the northeast states

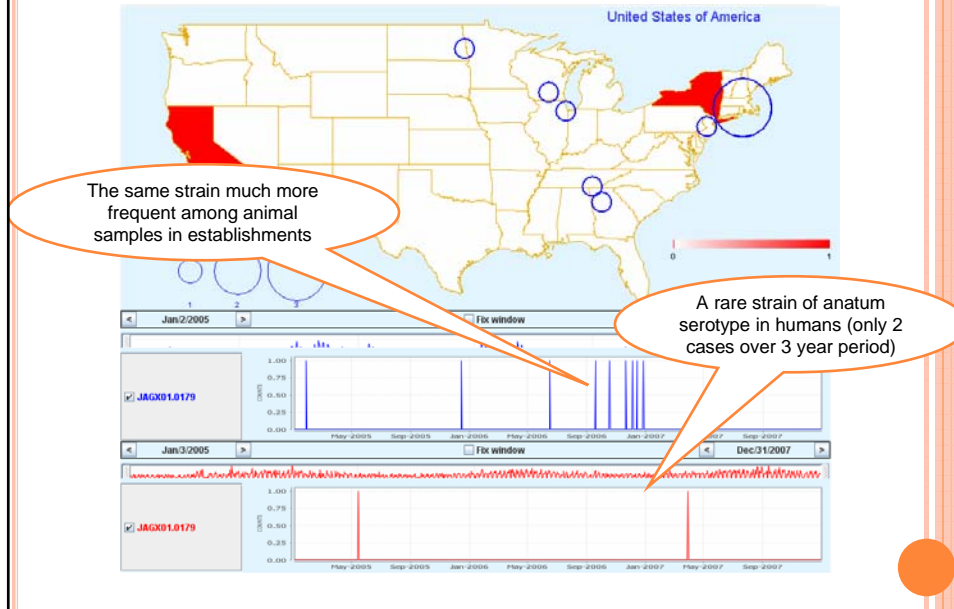


## (3) RARE PATTERN

- For rarely occurring PFGE patterns in Human Illness data, is there linkage evidence to the Food Sampling data?

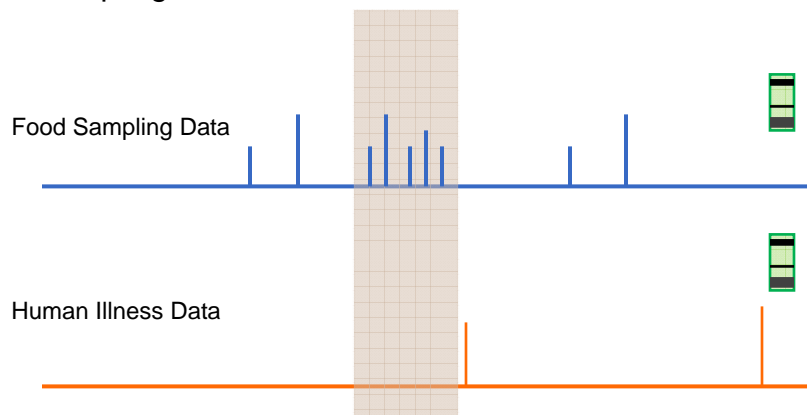


### (3) RARE PATTERN: EXAMPLE

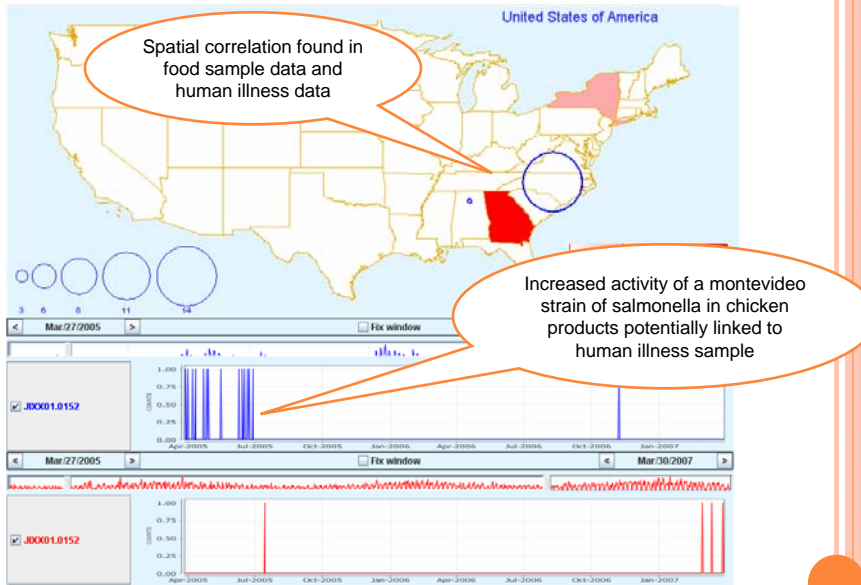


### (4) RECENT PATTERN SPIKE

- For every PFGE pattern found in Humans, find whether same PFGE pattern has recently spiked in Food sampling data?



## (4) RECENT PATTERN SPIKE: EXAMPLE



## COMPUTATIONAL SUMMARY

	#Matches	#Time series processed	Response time
Cluster pattern	116	651,923	22
Pattern not-in-cluster	38	1,286,331	38
Rare pattern	39	1,449,896	39
Recent pattern spike	23	197,694	12

## CONCLUSION

- We discussed a few search scenarios that show potential linkage between Food Supply data and Human Illness data.
- Finding such linkages in diverse data streams can help improve disease outbreak detection
- Using efficient algorithms can help both investigation and data mining.
- Sharing data sources among multiple departments is the key to the future of bio-surveillance.
- **SALMONELLA DATA ARE USED HERE AS AN EXAMPLE. THE PRESENTED FRAMEWORK IS EASILY ADAPTABLE TO USE WITH OTHER PATHOGENS, OTHER DATA SOURCES AND OTHER DOMAINS INVOLVING SPATIOTEMPORAL DATA STREAMS. WE BELIEVE THAT DEVELOPMENT OF INTERACTIVE ANALYSIS TOOLS LIKE THIS WILL HELP OFFICIALS TO QUICKLY AND PROACTIVELY IDENTIFY AND RESPOND TO PUBLIC HEALTH THREATS.**

## TAKE HOME MESSAGE

- There is **huge benefit** in **simultaneously analyzing** the **Food Supply data** collected by USDA-FSIS and **Human Illness data** collected by CDC
- We strongly recommend **sharing data sources** with linked dimensions to significantly improve syndromic surveillance
- Acknowledgements  
This work was partially supported by the USDA (award 1040770), CDC (R01-PH000028), and NSF (IIS-0911032).

## FOOD SAMPLING DATA

- Food Safety and Inspection Services (FSIS), a department of USDA, routinely **inspects various meat establishments** across the nation for possible sources of Salmonella infections.
- This “Food Supply” data contains **multiple lab test results** per establishments and records the Salmonella Serotype found, if any, in the food.
- The data also contains the Salmonella DNA finger-print using pulsed-field gel electrophoresis (**PFGE**) analysis for positive test cases
- A lot of **meat product recalls** are based off evidence from this data stream

## FOOD SAMPLING DATA DIMENSIONS

- We received 145K Salmonella lab test results from FSIS in the time period between Jan, 2005 to Dec, 2007.
- Each record contains the following dimensions:
  - Inspection date
  - Establishment number (3544)
  - Corporate group (18) – Tyson, Kraft, Hormel ...
  - PFGE pattern (1955)
  - CDC matched PFGE pattern (301)
  - Salmonella serotype (132)
  - Product category (7) – Beef, Chicken, Turkey, ...
  - Product type (2) – Ground, not Ground
  - Test result (2) – Positive, Negative

## HUMAN ILLNESS DATA

- Center for Disease Control and Prevention (CDC) collects reports of Salmonella cases found in Humans
- For each case in the data, they record the Salmonella serotype and PFGE pattern
- CDC looks cases of Salmonella outbreaks in humans that are spatially co-located and found in clusters. All such clusters of records are given unique outbreak identifier

## HUMAN ILLNESS DATA DIMENSIONS

- We received 93K records from CDC in the time period between Jan, 2005 to Dec, 2007.
- Each record contains the following dimensions:
  - Report date
  - State (49)
  - Product type (13) – Dairy, Meat, Fruit/Vegetable, ...
  - Outbreak identifier (664)
  - Salmonella serotype (635)
  - PFGE pattern (10155)
  - Agency (3) – FDA, USDA-FSIS, unknown
  - Source (5) – Animal, Food, Human, Environmental, ...

## WHY SALMONELLA LINKAGE?

- Salmonella is among the most common food-borne illnesses in the USA with about 40,000 cases reported annually including 600 deaths
- Many of these deaths are believed due to consumption of contaminated food
- By linking food supply to human illness data, we hope to early detect emerging Salmonella outbreaks and hence reduce the impact of these outbreaks to human health