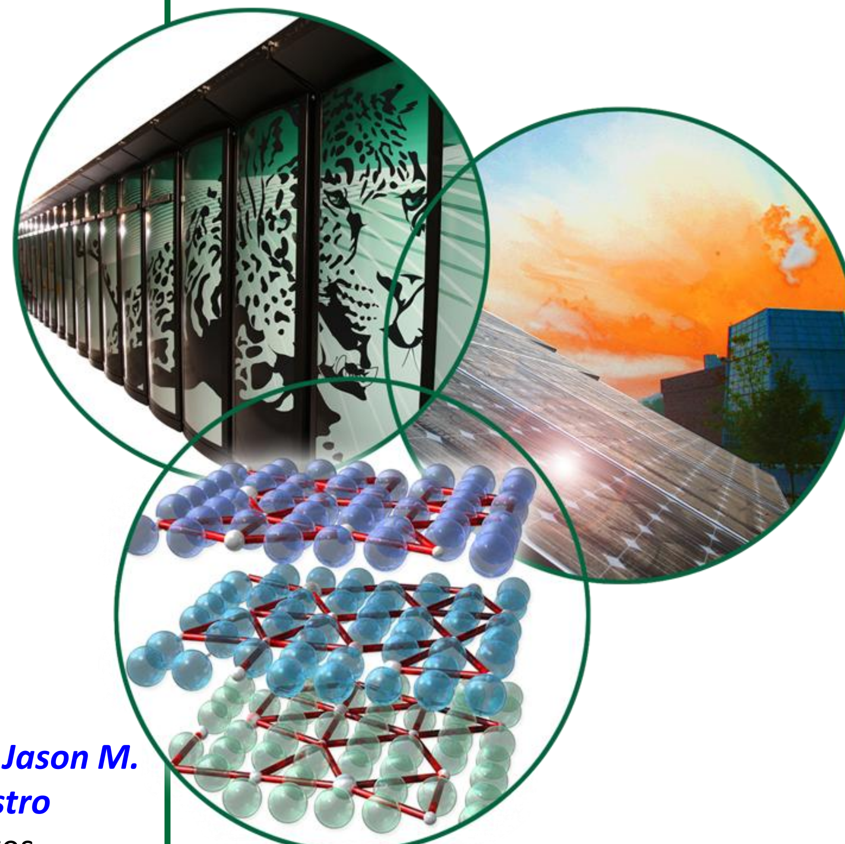
Cloud-Based Computational Bio-surveillance Framework for Discovering Emergent Patterns From Big Data



Arvind Ramanathan

ramanathana@ornl.gov

Computational Data Analytics Group,
Computer Science & Engineering Division,
Oak Ridge National Lab, Oak Ridge, TN

Dr. Chakra S. Chennubhotla
University of Pittsburgh

Shannon Quinn
University of Pittsburgh

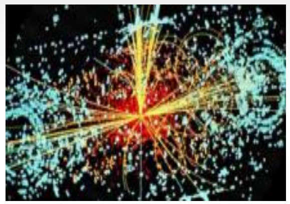
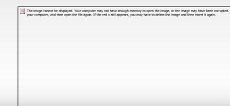
Dr. Jason M. Castro
Bates College

U.S. DEPARTMENT OF ENERGY

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

Bio-surveillance from Big Data: Big Challenges

We generate 2 quintillion bytes (2×10^{18}) of data every day.(IBM)

Experiments	*Simulations*	*Archives*	*Social Media*	*Sensors*
				
• genome scale experiments • proteomics • structural biology, • clinical studies	• disease spread models • molecular dynamics • social networks	• history of communicable diseases • archives of health records	• twitter, • facebook	• environmental monitors, • weather/climate monitors • hospital sensors, • other sources

Information

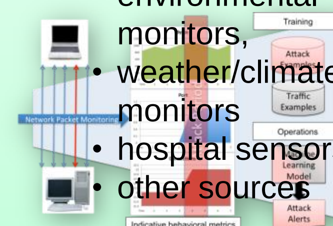
The Challenge

Enable Discovery

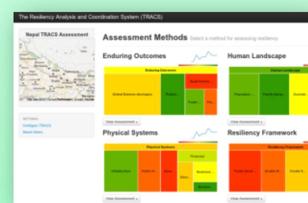
Deliver the capability to mine, search and analyze this data in near real time

Data → Discovery → Insights

VERDE
(Visualizing the electric grid)



Zero Day Attack Detection



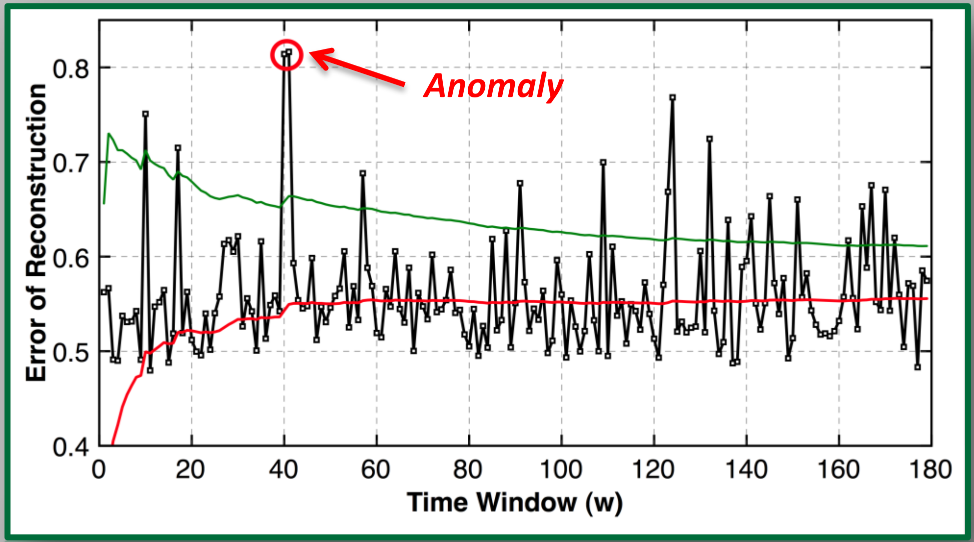
Resiliency Analysis and Coordination System

CMS Analytics
(Decision from Big Data)

Analyzing Big Data from Bio-surveillance…

What is this talk about …

- Suite of ***statistical and machine learning tools*** for:
  - discovering inherent statistical structure of domain specific big data
  - providing testable hypotheses ("actionable insights")
- ***Challenges*** faced in developing a computational infrastructure:
  - Volume/Velocity
  - Scaling algorithms



Part 1: Online Event Detection

- **Spatio-temporal correlations**
- **Dynamical clustering**

Motivation: Detecting spatio-temporally correlated patterns in real-time data streams (Twitter)

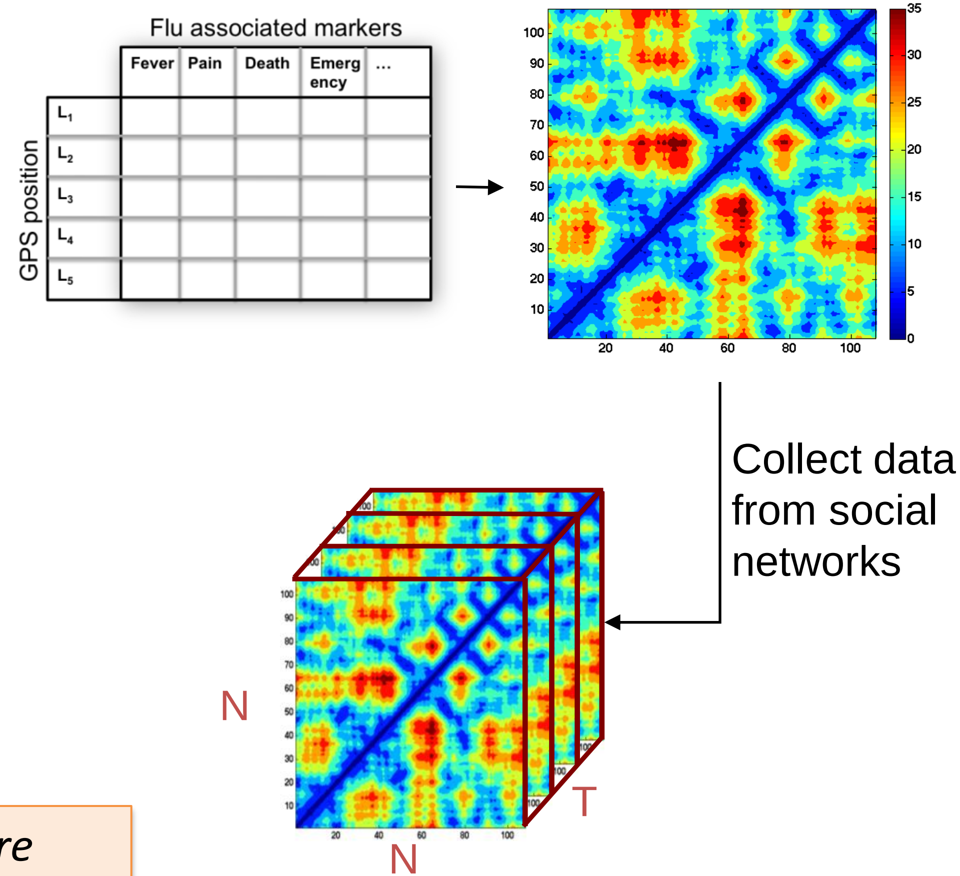
- *Which geographic regions exhibit correlated patterns in twitter patterns?*
  - *Indicative of emergent patterns in spread of disease/ outbreak*
  - *Can be across diseases or regions or along time*
- *At what time-points do these patterns change?*
  - *Anomalies indicative of sudden surges in infections*

varying patterns in disease association.

Neoformix: Visualizing Twitter data

Tensor representation for text data streams

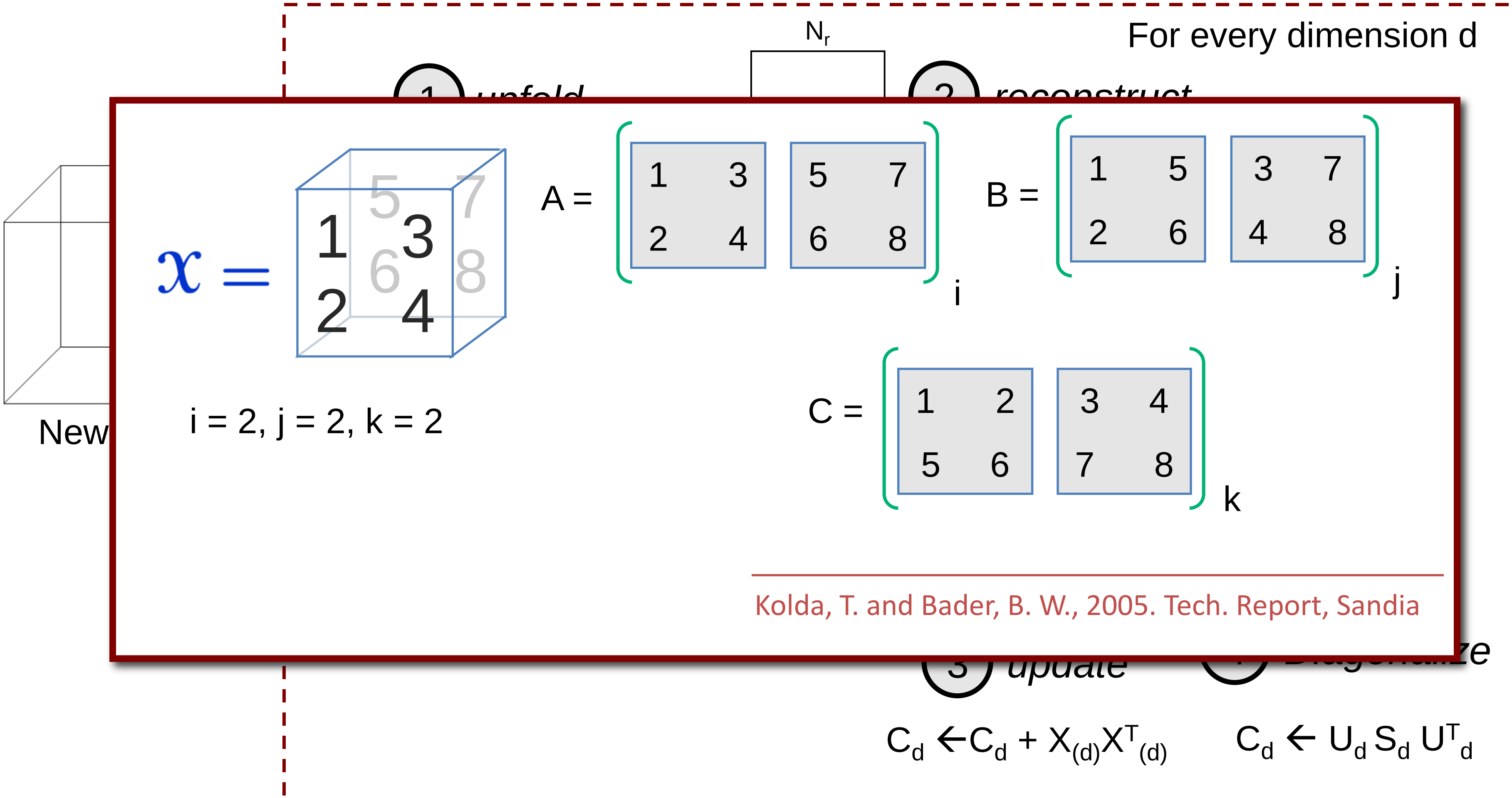
- Conceptually the data is a collection of matrices
- Conveniently represented as a tensor



3D tensor of outbreak terms + locations evolving over time

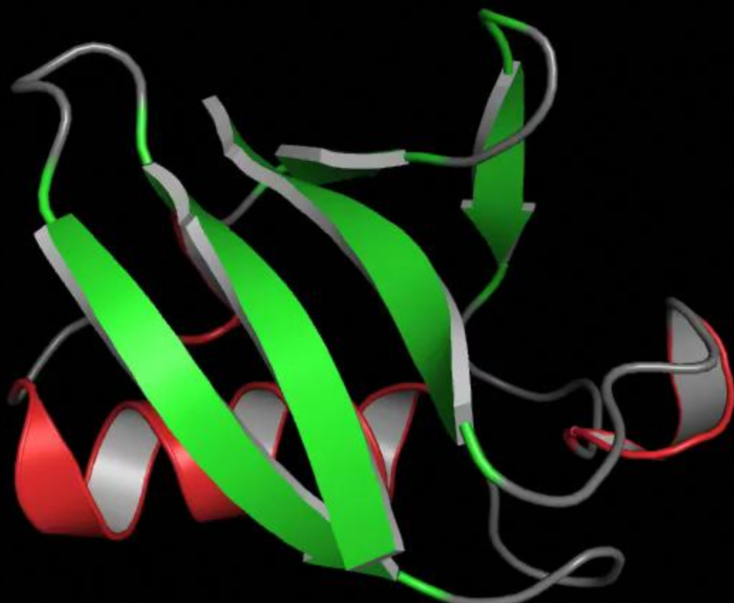
Tensors are N-dimensional matrices, that are useful to capture multi-way dependencies

Online Tensor Analysis



Ramanathan, A., Agarwal, P.K., Kurnikova, M. and Langmead, C., RECOMB 2009.
Sun, J., Faloutsos, C., and Kolda, T., KDD 2006.

Translating to a small world!



- *Which regions of the molecule are moving together?*
- *At which time-points are the spatio-temporal patterns of motions changing?*

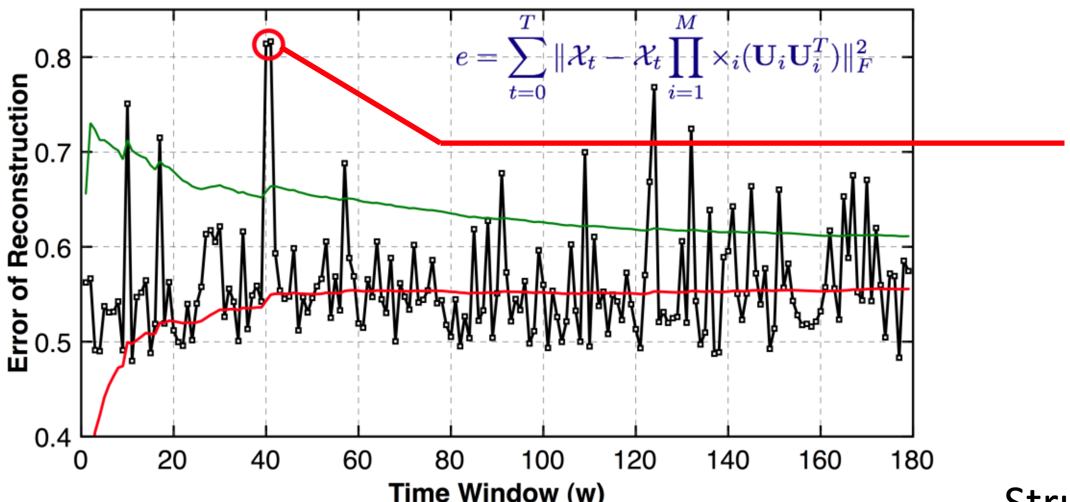
Data
$$\begin{pmatrix} x_1, y_1, z_1 \\ x_2, y_2, z_2 \\ \vdots \quad \vdots \quad \vdots \\ x_N, y_N, z_N \end{pmatrix} \begin{pmatrix} x_1, y_1, z_1 \\ x_2, y_2, z_2 \\ \vdots \quad \vdots \quad \vdots \\ x_N, y_N, z_N \end{pmatrix} \ldots, \ldots, \ldots, \begin{pmatrix} x_1, y_1, z_1 \\ x_2, y_2, z_2 \\ \vdots \quad \vdots \quad \vdots \\ x_N, y_N, z_N \end{pmatrix} \begin{pmatrix} x_1, y_1, z_1 \\ x_2, y_2, z_2 \\ \vdots \quad \vdots \quad \vdots \\ x_N, y_N, z_N \end{pmatrix}$$

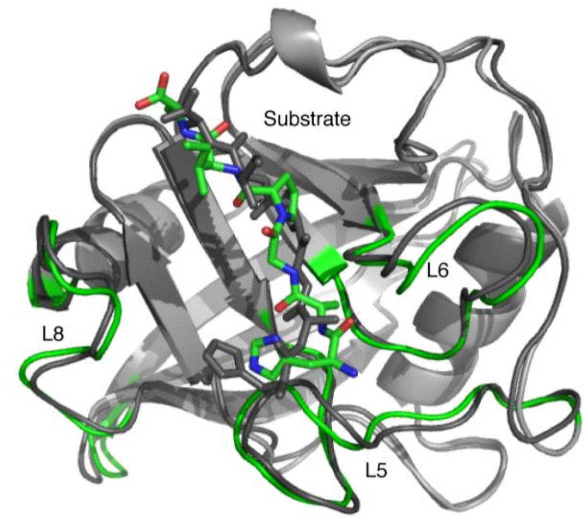
Bag of Words

Time = **0** **1** **T-1** **T**

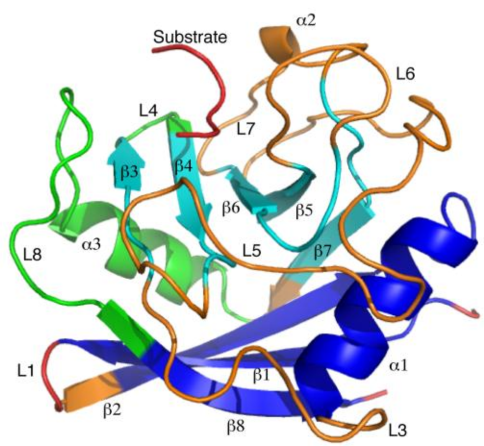
Data → Insights → Discovery:

Time-points where spatio-temporal correlations change can be used to control simulations





Structural differences shown in green



Clustering spatial regions in the enzyme showing similar patterns of motion

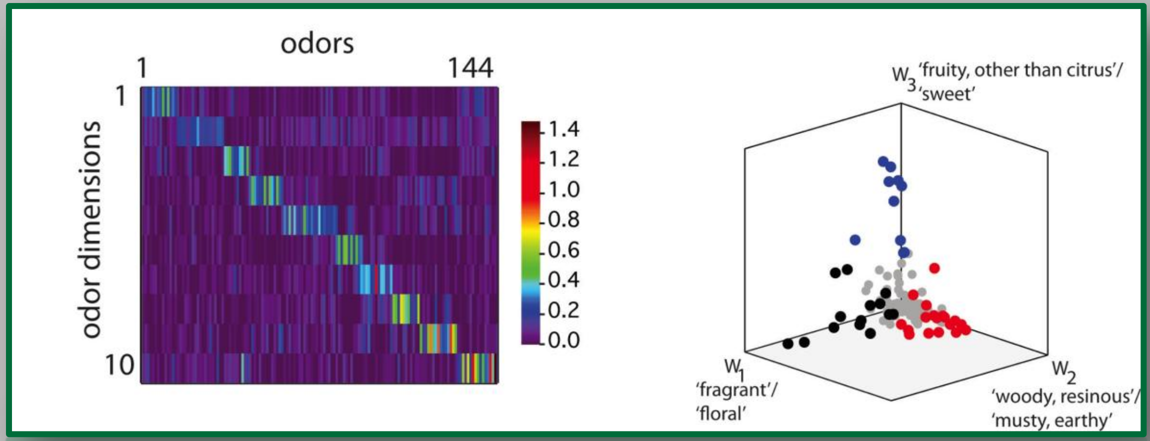
Key Contributions

An *online* tool for data mining:

1. **Anomaly detection:**
   – ***time points where social media patterns change***
   – ***Can be used to track disease outbreak***

2. **Spatio-temporal pattern discovery:**
   – ***cluster geographical regions based on media patterns***

3. **Data summarization**

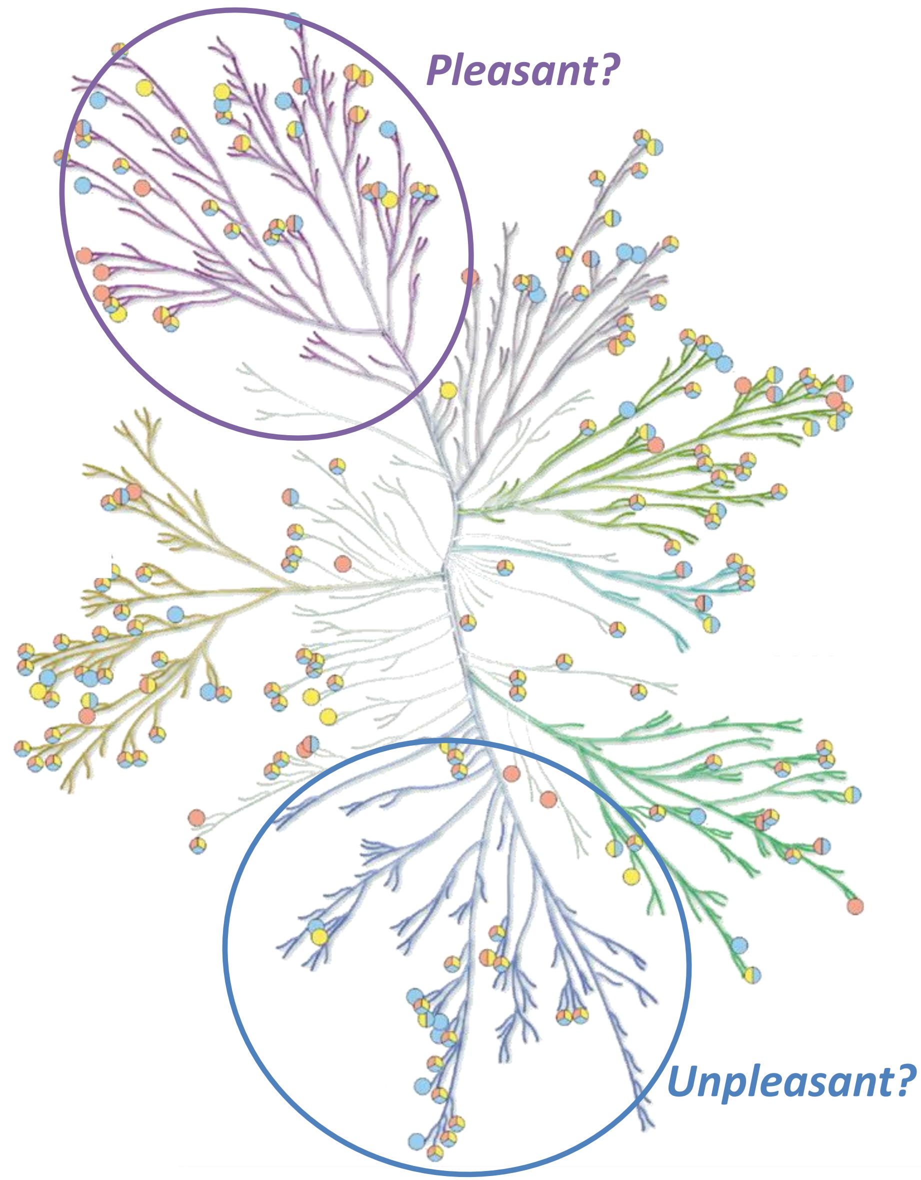


Part 2: Discovering inherent statistical structure in big data

- **Organizing high dimensional spaces**
- **Odor perception**

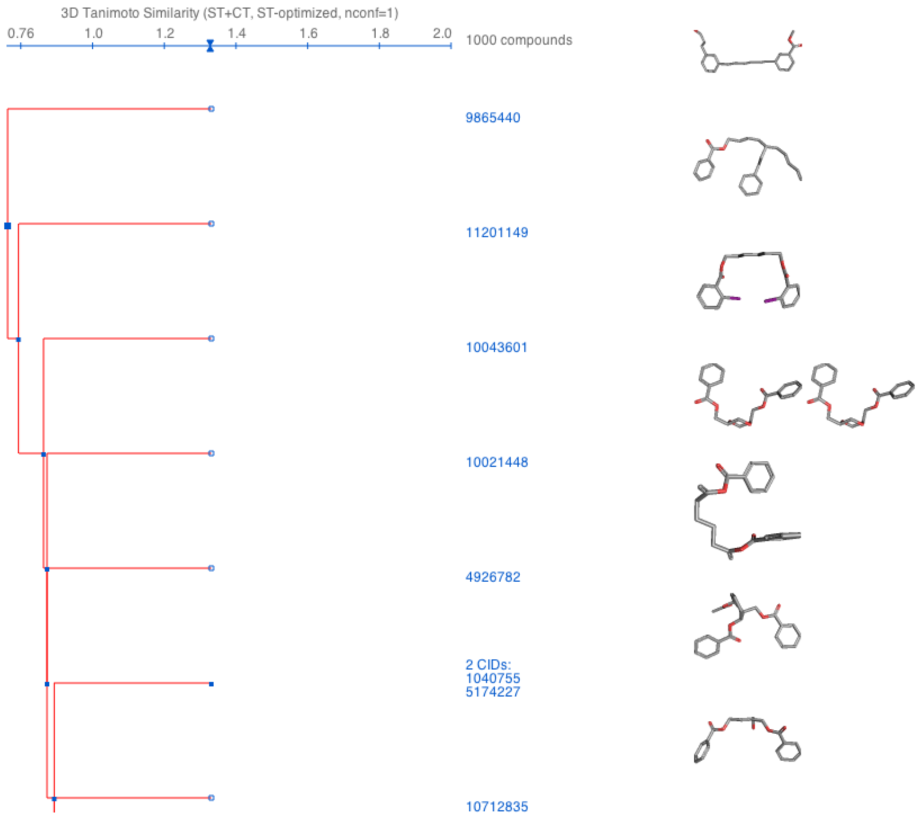
Motivation: Towards machine olfaction…

- **Odor perception:**
  - *What is the perceptual space of the human olfactome?*
- 31 million molecules from Pubchem!!
  - Big Data: How to organize this space?
- We don't have this organization:
  - Can we build this from data?
  - Statistical characteristics from both psychophysics & chemical spaces



Using semi-supervised learning to "odor" label the Pubchem

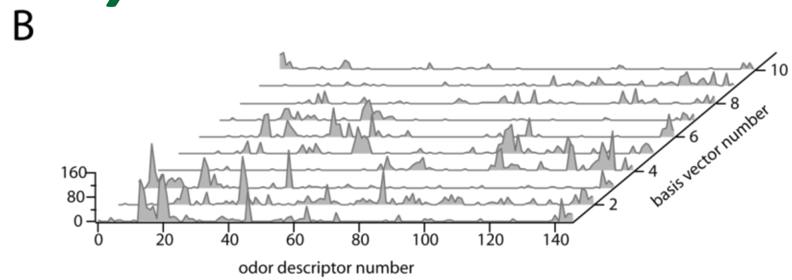
- Label small portion of the data with odor percepts
  - Derive physio-chemical features from labeled data
- Graph-kernel approaches to quickly compare compounds
- Propagate labels on successively to larger data sets (flavornet, superscent)
- Test / Validate / Refine

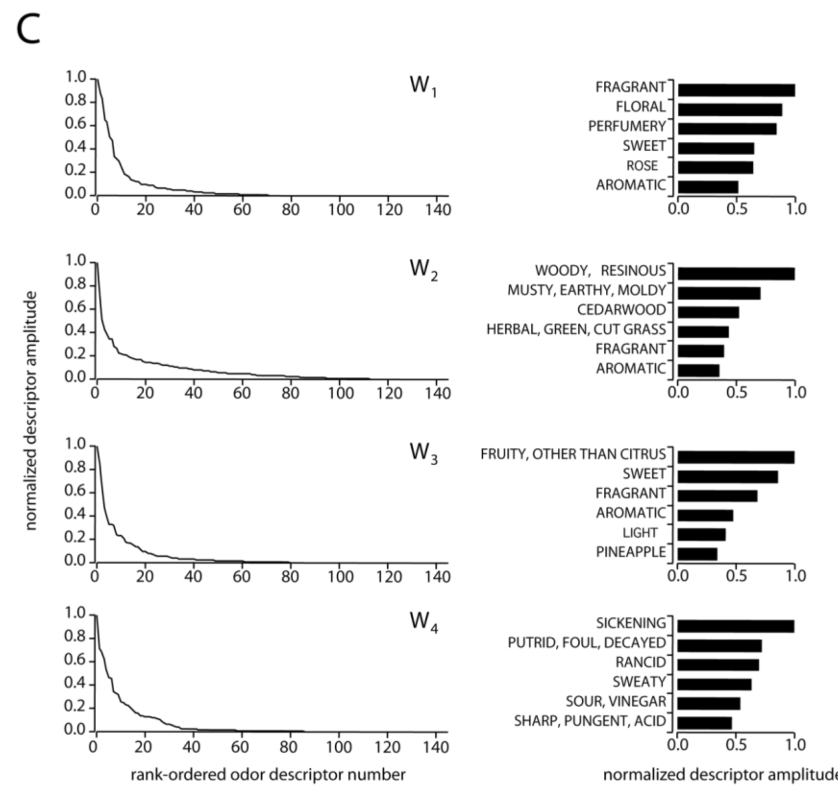


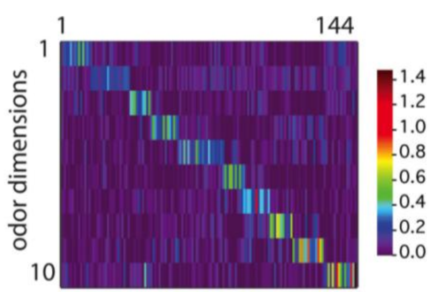
Castro, J.B., Ramanathan, A., Chennubhotla, C.S. (2012) PLoS One (in preparation)

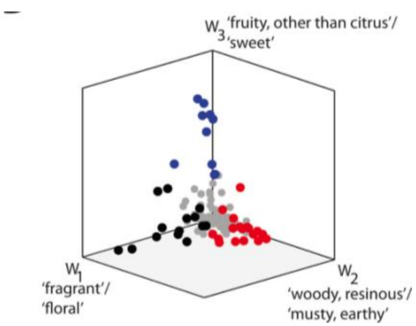
Building a perceptual model of odors on Atlas of Odor Chemical Percepts (AOCP)

- 144 odors; ~150 odor descriptors
- Use non-negative matrix factorization for dimensionality reduction
- Rigorous cross validation





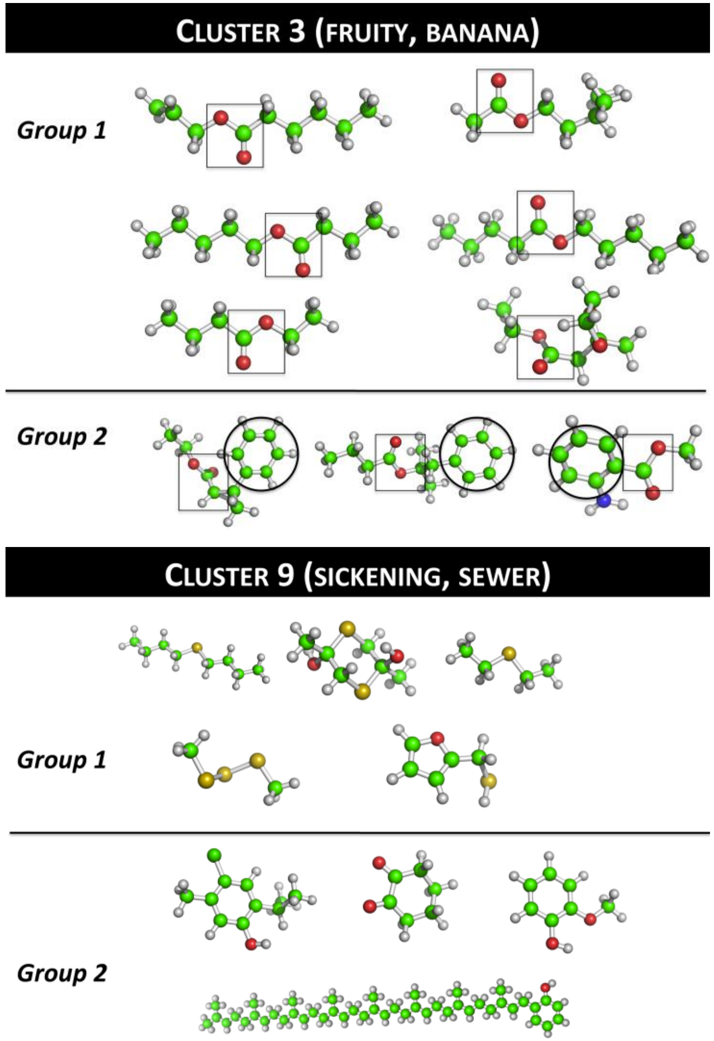




Data → Insights → Discovery

Odors with similar perception share unique physio-chemical signatures

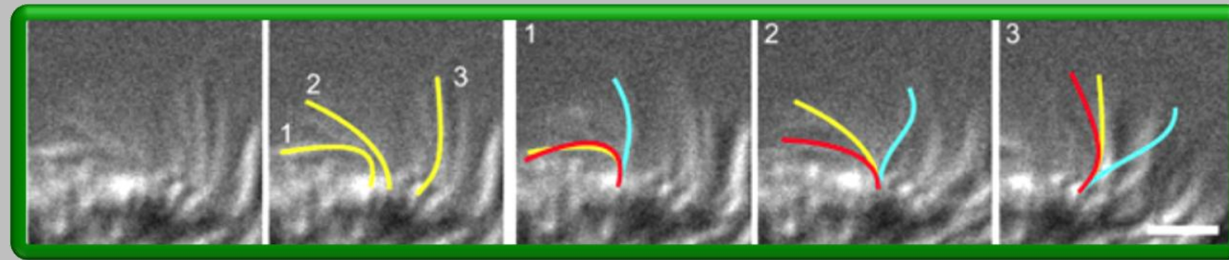
- Fruits and sewer have distinct chemical features:
  - nRCOOCR
  - nS
- Identified automatically from over 1600 physio-chemical features



Key Contributions & Future Work

A machine learning framework to relate chemicals to their odor percepts:

- Discovery of underlying statistical structure within large-scale datasets
  - how do people perceive odors?
  - linking "odor perception" to "chemical signatures"
- Organizing odors into a perceptual frame of reference: ***Olfactome***: using novel machine learning tools
  - integration with psycho-physics experiments
  - expanding the compounds to include a larger chemical repertoire

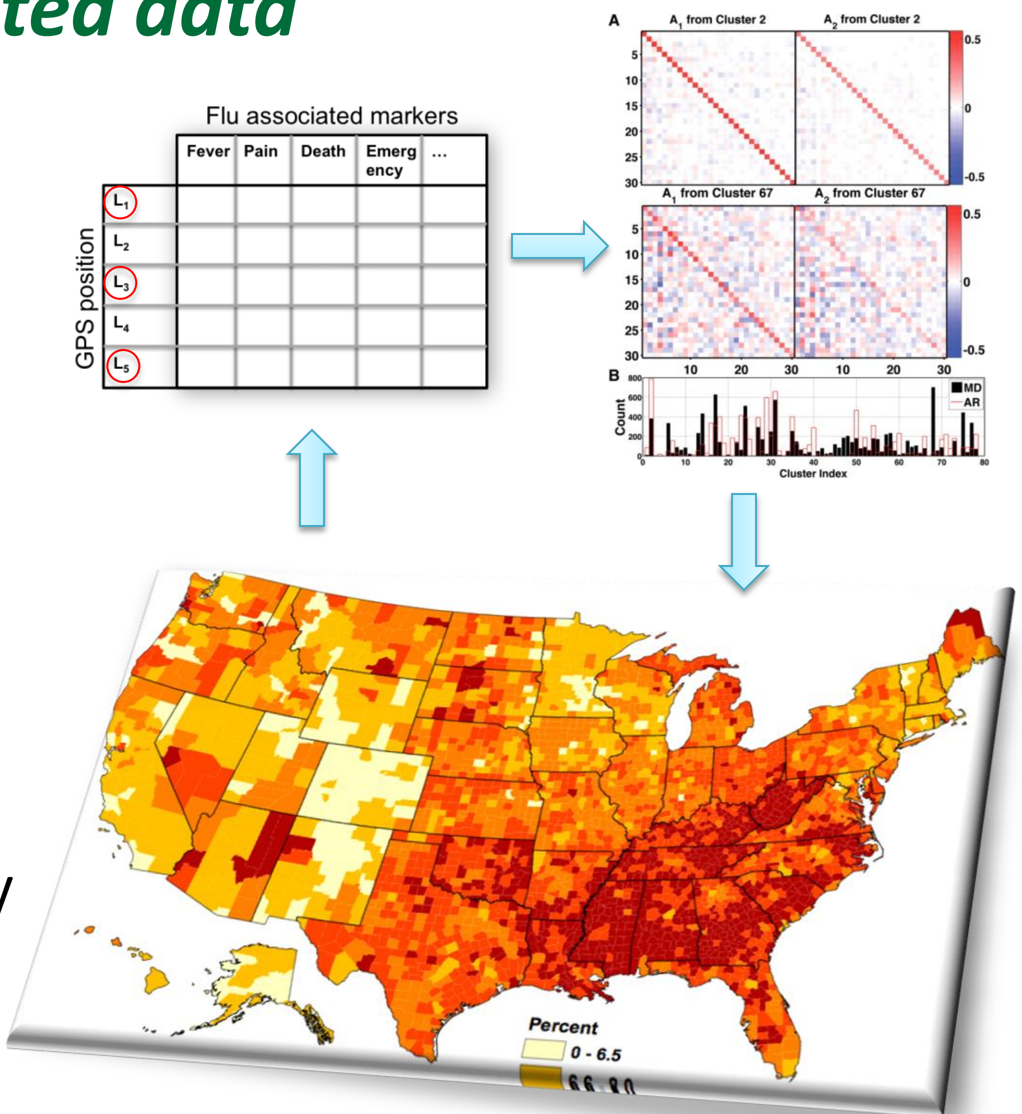


Part 3: Moving to the cloud…

- **Organizing high dimensional spaces**
- **Auto-regressive models**
- **Bio-medical applications**

Motivation: Automate detection of patterns from disparate, distributed data

- Data: Twitter Feed / Social media
  - Globally distributed data
  - Large volume
- Temporal models:
  - patterns in disease spread
- Generative models:
  - predicting how disease may spread

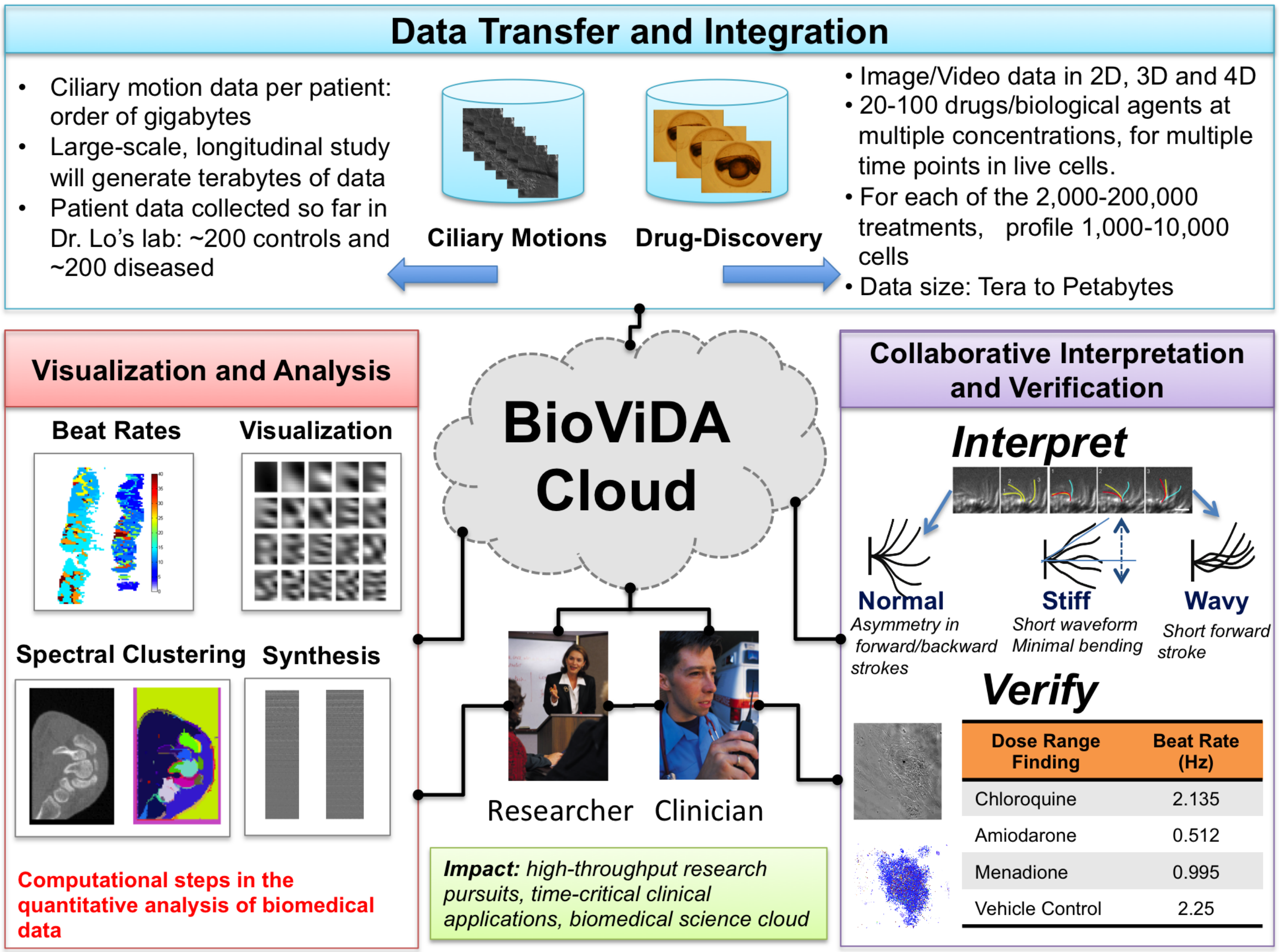


Bio-surveillance and the Cloud

Bio-surveillance data

- **is BIG and NOISY**
- **requires repetitive analysis in chunks**
- **modeling involves linear algebra and statistics**

Example: Biological Visualization & Data Analytics for Disease Diagnostics



Dose Range Finding	Beat Rate (Hz)
Chloroquine	2.135
Amiodarone	0.512
Menadione	0.995
Vehicle Control	2.25

- Data@ University of Pittsburgh:
  - Dr. Cecilia Lo's lab
  - Drug-discovery Institute
- Compute Cloud: Qloud@CMU-Qatar, Dr. Majd Sakr

Summary

- An overview of a computational infrastructure that implements ***scalable machine learning algorithms*** to:
  - ***discover inherent structure*** from various sources of bio-surveillance data
  - ***provide near real-time feedback*** for end-users on emerging patterns
- ***Challenges*** include:
  - Seamlessly fusing multiple data sources
  - Standards across the globe differ!

Acknowledgements

- **Computational Data Analytics Group**
  - *Dr. Thomas Potok (Group Leader)*
  - *Dr. Laura Pullum*
  - *Dr. Bryan Gorman*
  - *Dr. Mallikarjun Shankar*
- **Computer Science Research**
  - *Dr. Pratul K. Agarwal*

Thank You !!!

Questions/ Comments: ramanathana@ornl.gov