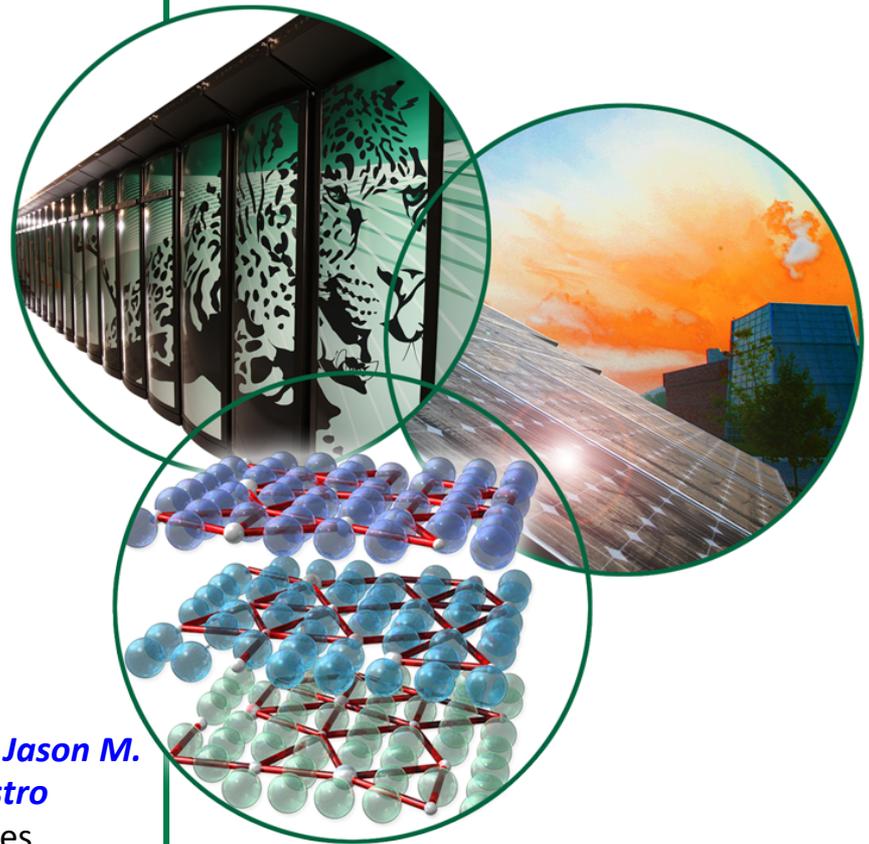


# *Cloud-Based Computational Bio-surveillance Framework for Discovering Emergent Patterns From Big Data*

**Arvind Ramanathan**

[ramanathana@ornl.gov](mailto: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

# Bio-surveillance from Big Data

Data → Discovery → Insights

## 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

- Testable hypotheses

ing  
ck  
oles  
ic  
ives  
ions  
line  
ing  
del  
ck  
ts

tion

)

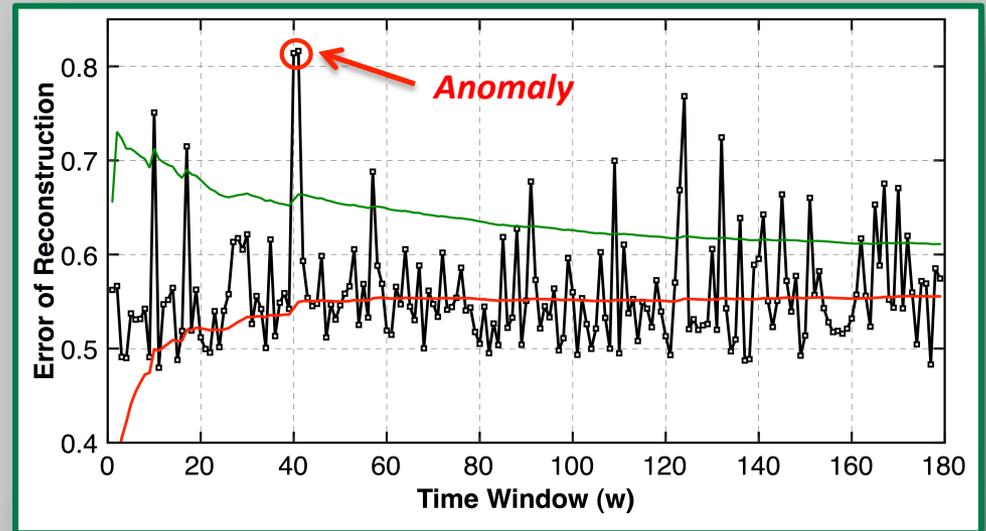
# Analyzing Big Data



www.jolyon.co.uk

- **Event Detection:** time-points where there is deviation from “normal” behavior
- **Multi-scale Feature Extraction:** *intrinsic structure of data*
- **Cluster & Visualize:** simplifying the interpretation for meaningful insights

**Data → Insights → Discovery**



## 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.*

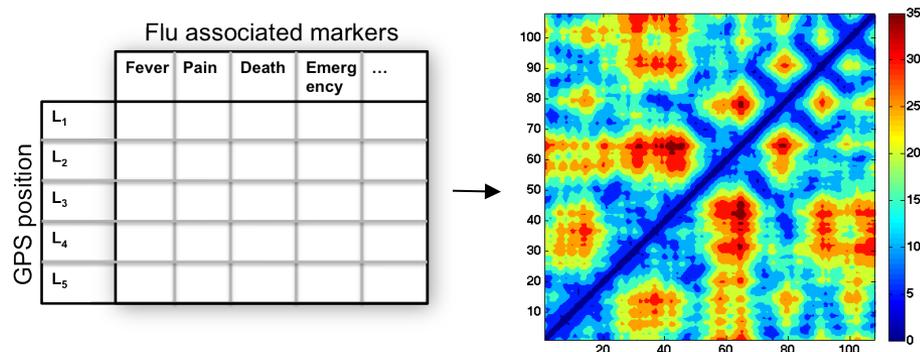
Neoformix: Visualizing Twitter data



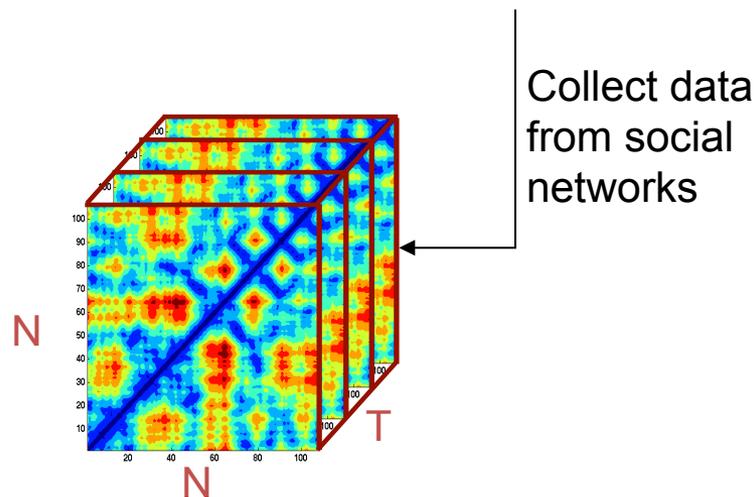
L<sub>5</sub>

# Tensor representation for intra-molecular distances

- Conceptually the data is a collection of matrices



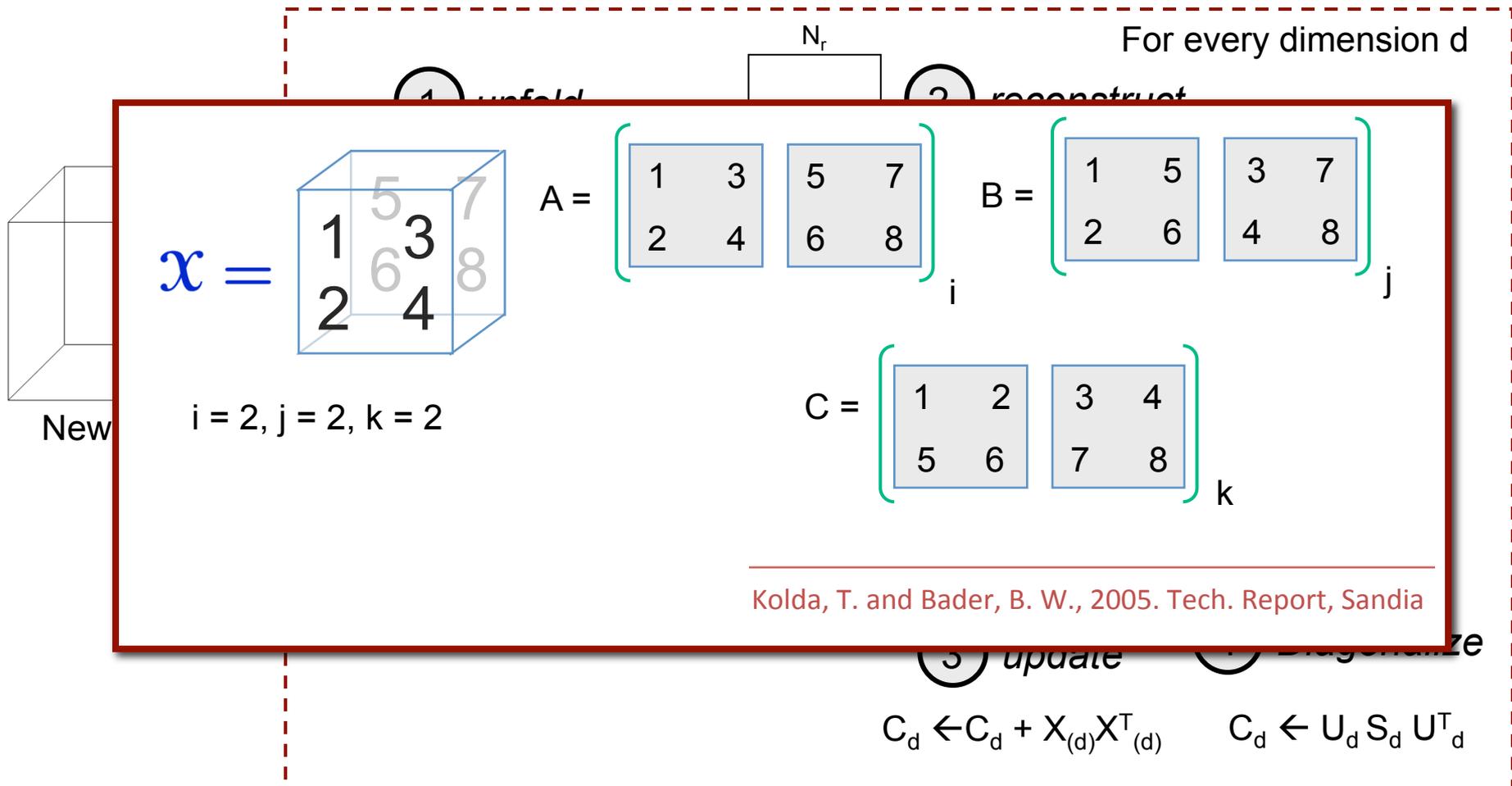
- Conveniently represented as a tensor



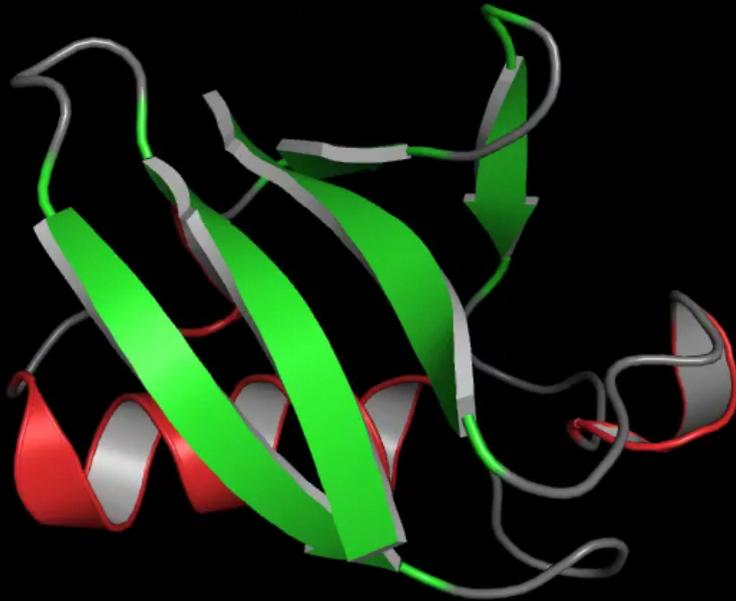
*Tensors are N-dimensional matrices, that are useful to capture multi-way dependencies*

3D tensor of outbreak terms + locations evolving over time

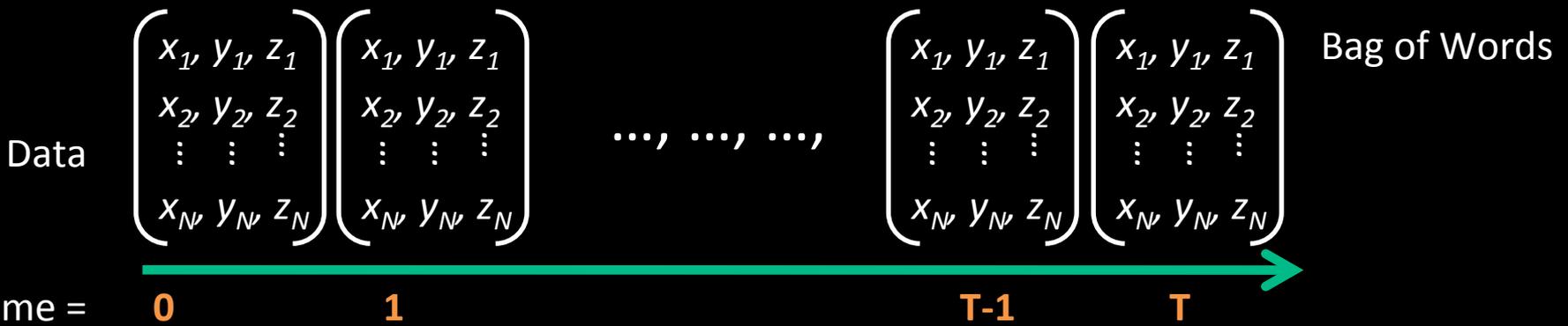
# Online Tensor Analysis



# Translating to a small world!

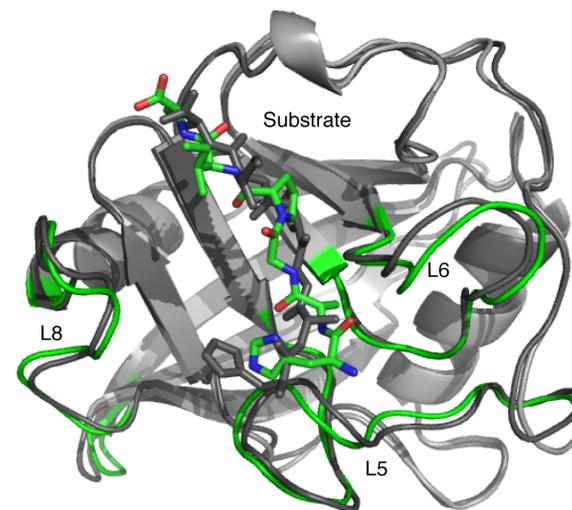
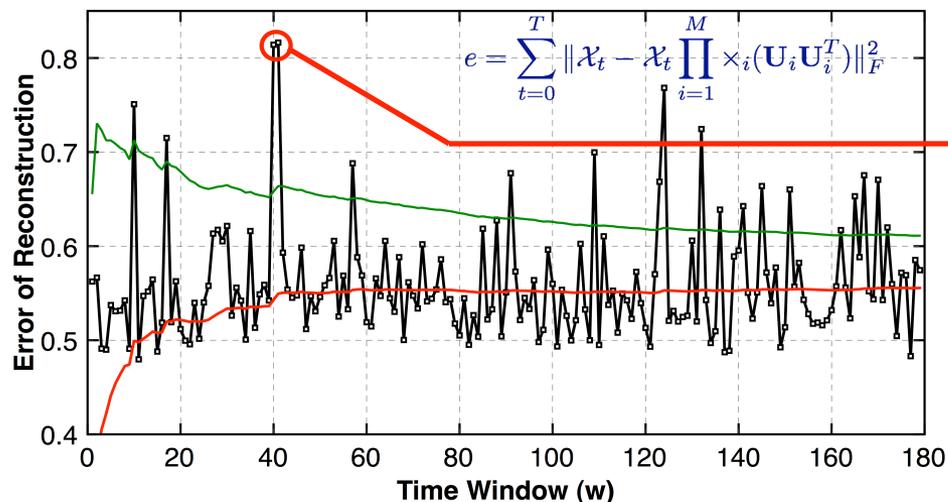


- Which regions of the molecule are moving together?
- At which time-points are the spatio-temporal patterns of motions changing?

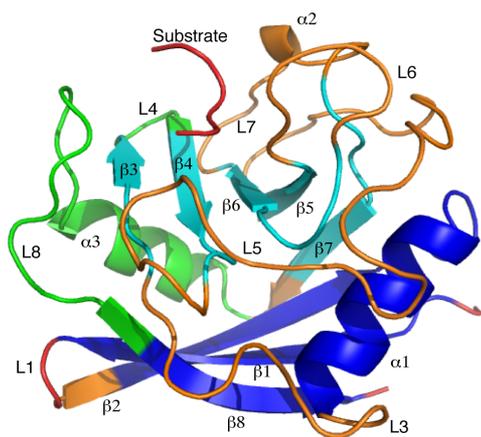


# Data → Insights → Discovery:

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:

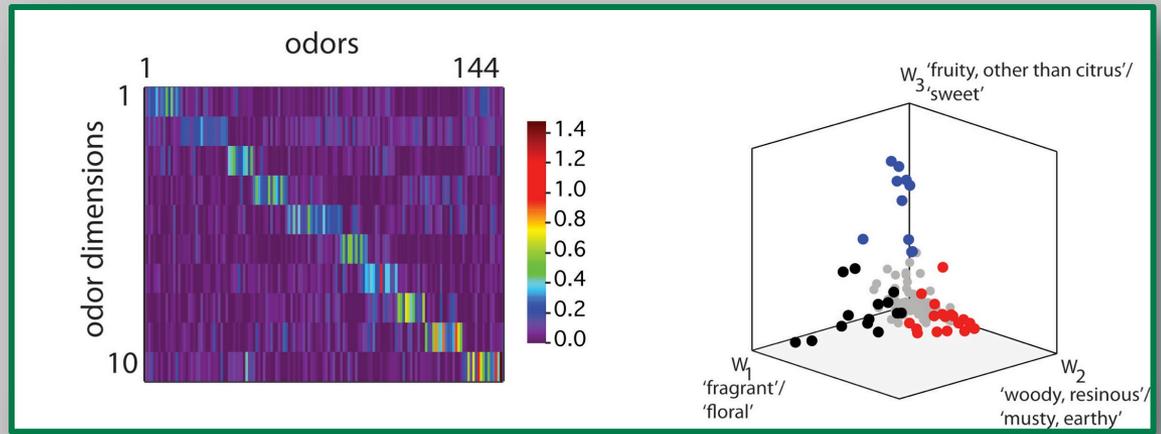
## 1. Anomaly detection:

- *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

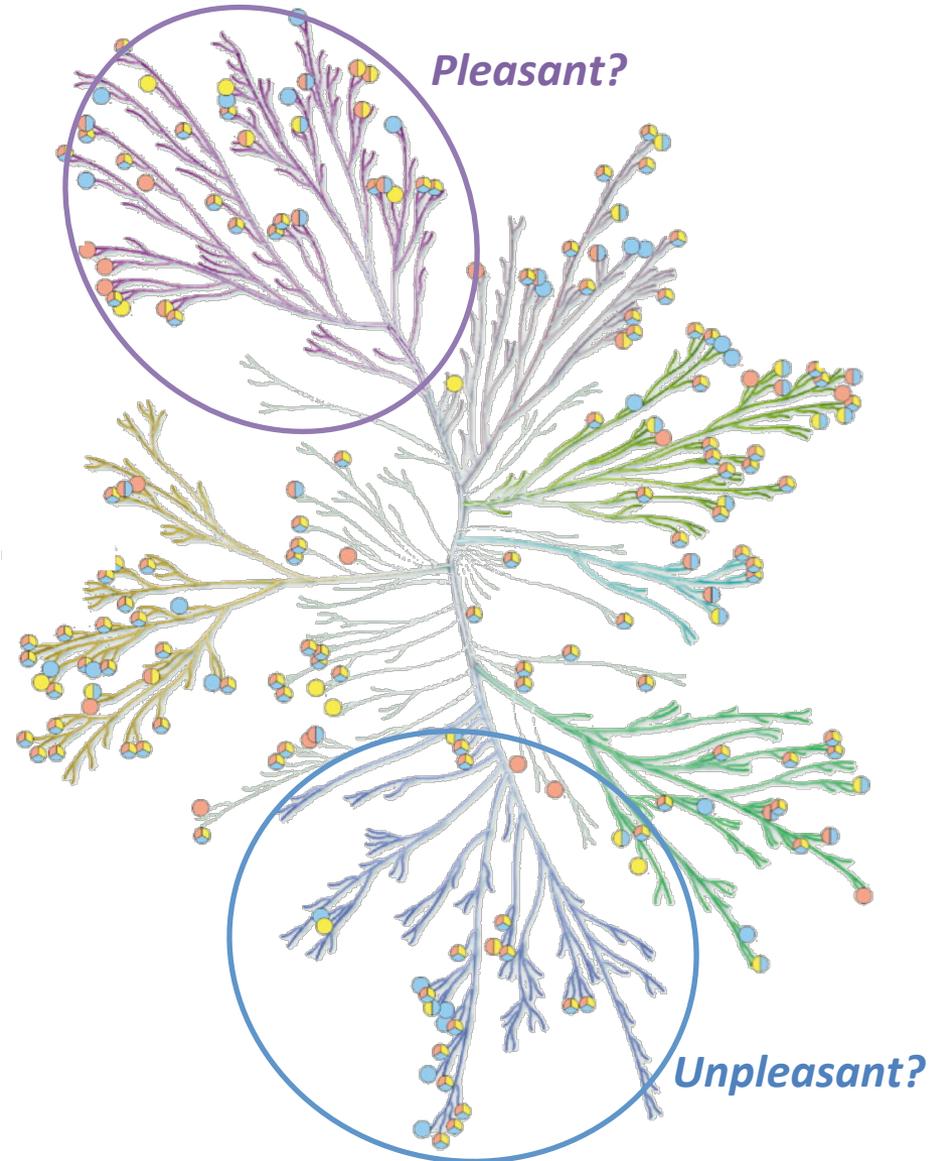


## *Part 2: Discovering inherent statistical structure in big data*

- Organizing high dimensional spaces
- Odor perception

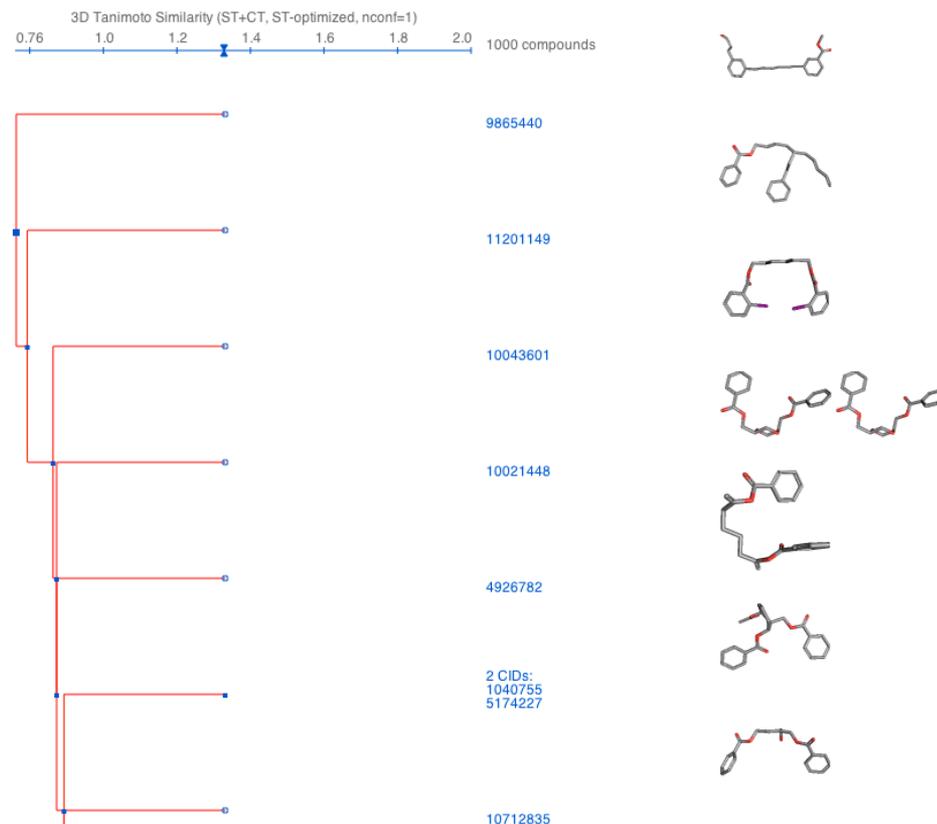
# Motivation: Towards machine olfaction...

- **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



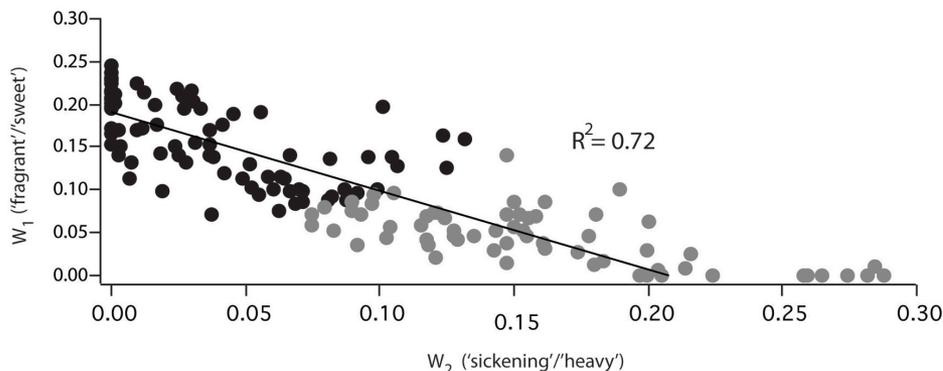
# Using semi-supervised learning to “odor” label the Pubchem

- 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

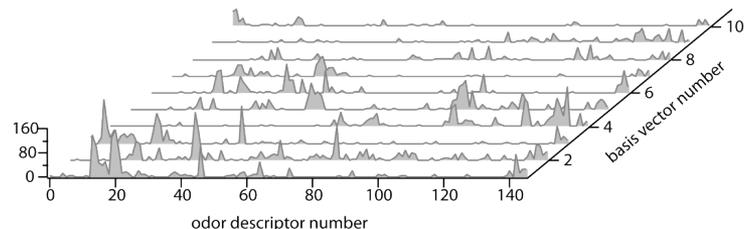


# Building a perceptual model of odors on Atlas of Odor Chemical Percepts (AOCP)

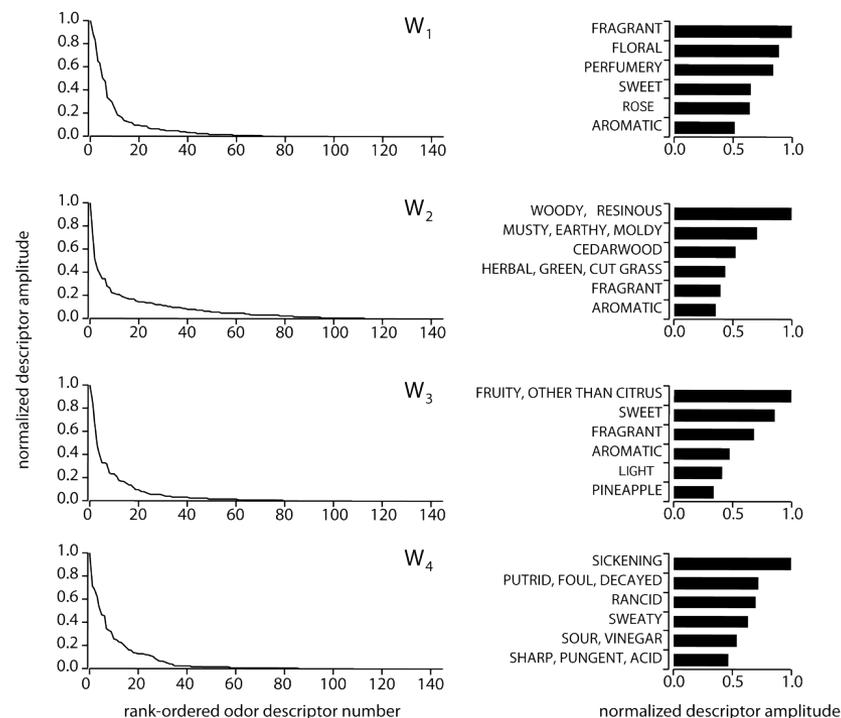
- 144 odors; ~150 odor descriptors
- Use non-negative matrix factorization for dimensionality reduction
  - Use bi-clustering to associate odors with perception



B

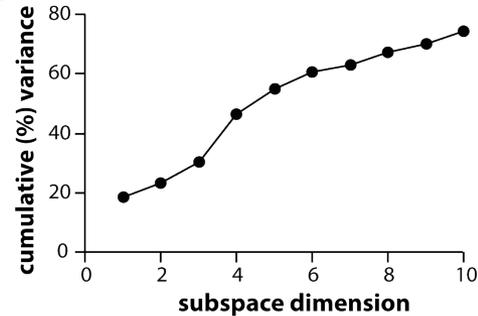
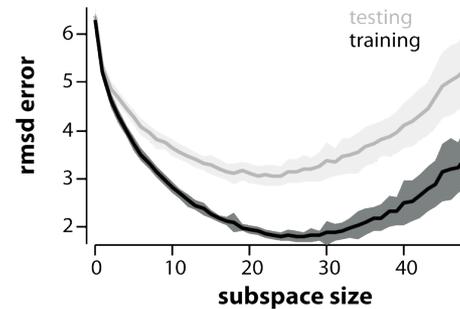
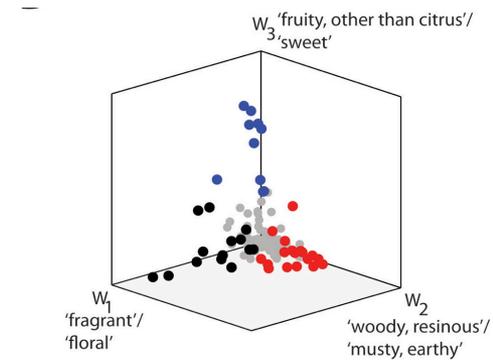
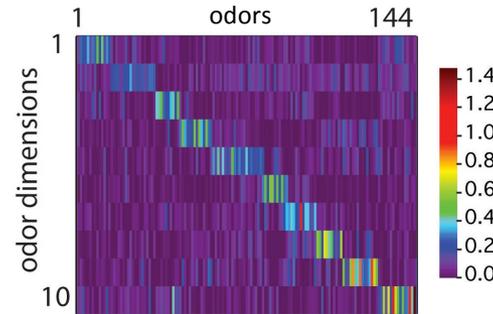


C



# Clustering odors based on perceptual qualities

- Odors separated into clear odor percepts:
  - (Fruity, sweet) different from (Floral, sweet)
  - Putrid different from Sewer, etc.
- Rigorous cross validation

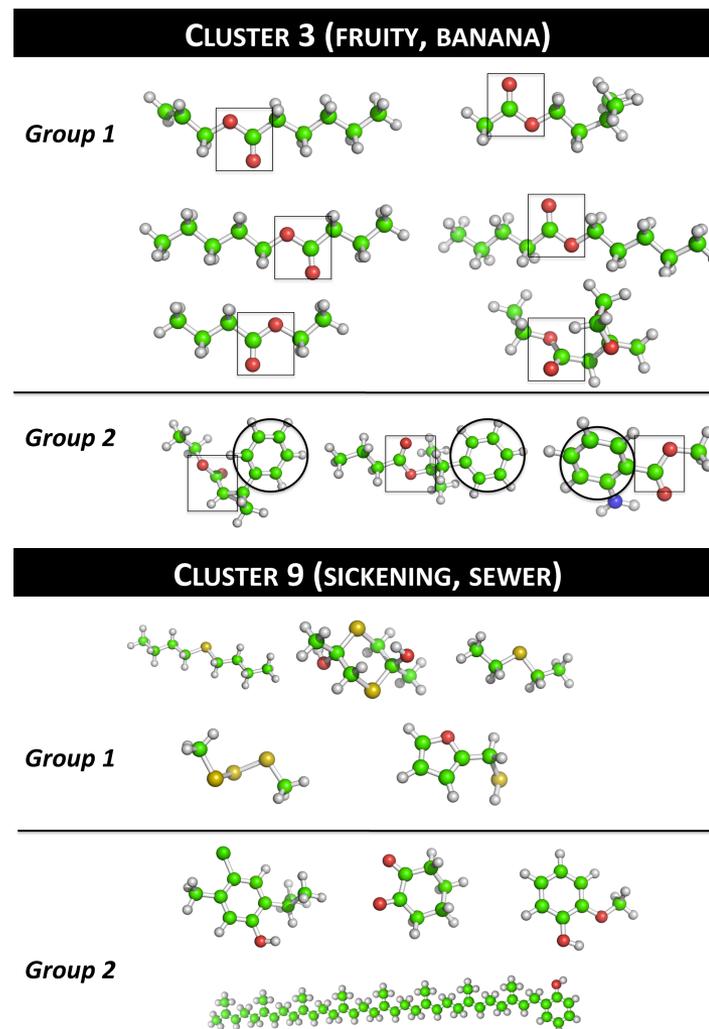


Castro, J. M., Ramanathan, A., Chennubhotla, C.S., PLoS ONE (submitted)

# Data → Insights → Discovery

***Odors with similar perception share unique physio-chemical signatures***

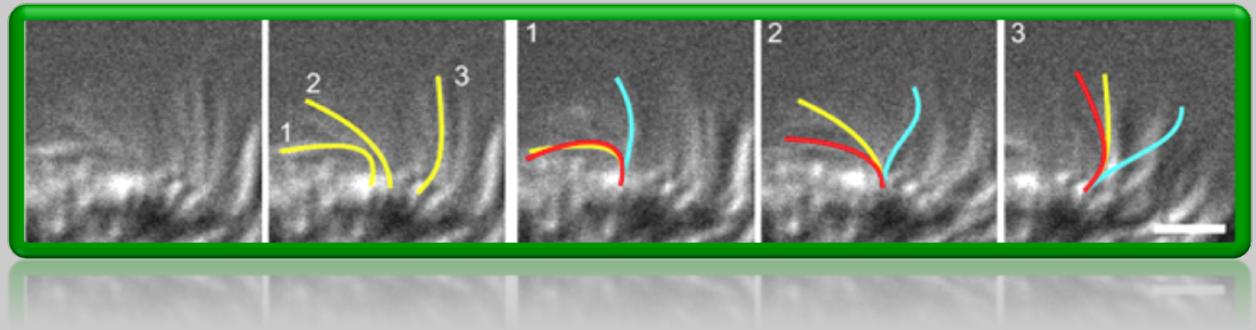
- Fruits and sewer have distinct chemical features:
  - nRCOOCR
  - nS
- Identified automatically from over 1600 physio-chemical features



# Key Contributions & Future Work

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

- Discovery of underlying statistical structure within large-scale datasets
  - linking “chemical nature” to “odor perception”
  - linking “odor perception” to “chemical signatures”
- Organizing odors into a perceptual frame of reference using novel machine learning tools
  - integration with psycho-physics experiments
  - expanding the compounds to include a larger chemical repertoire

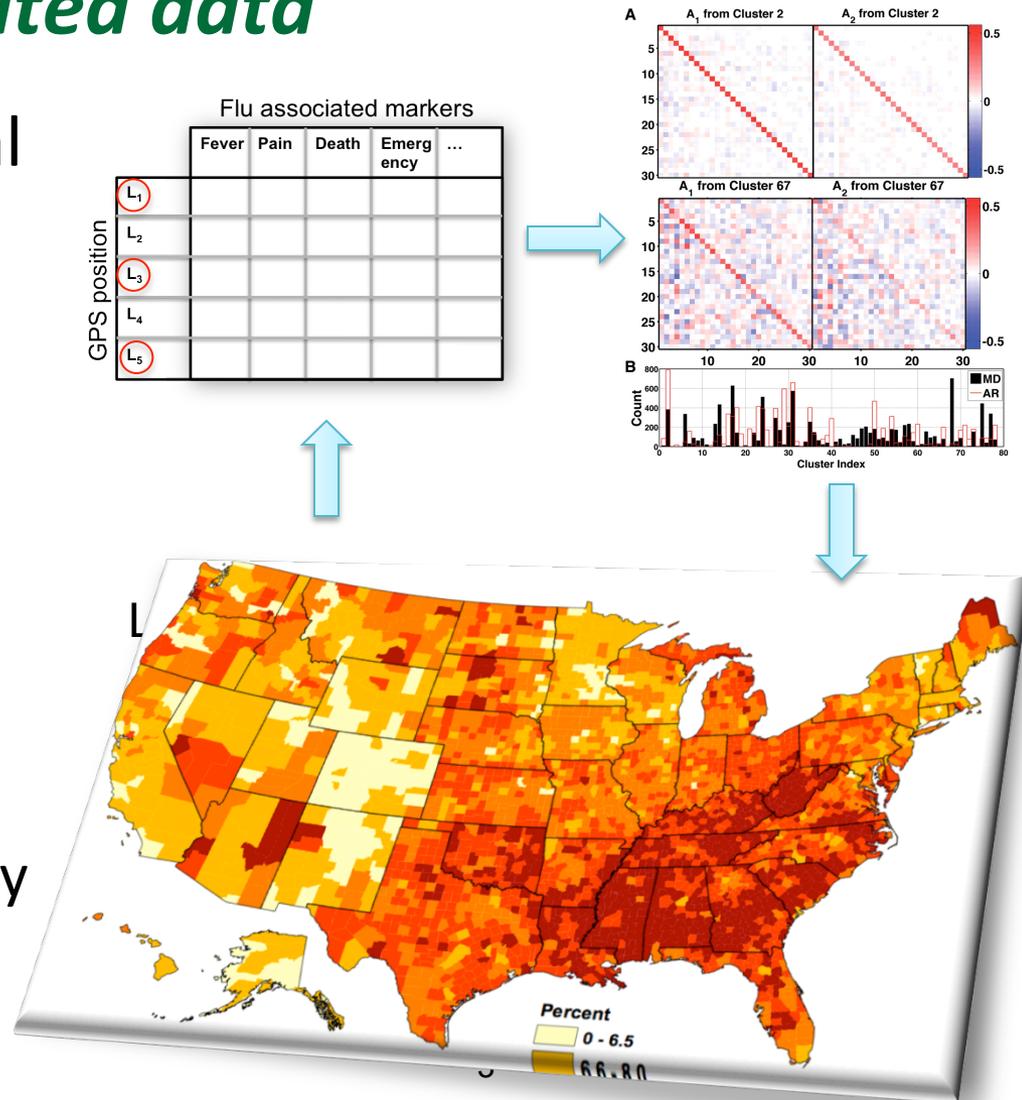


## *Part 3: Moving to the cloud...*

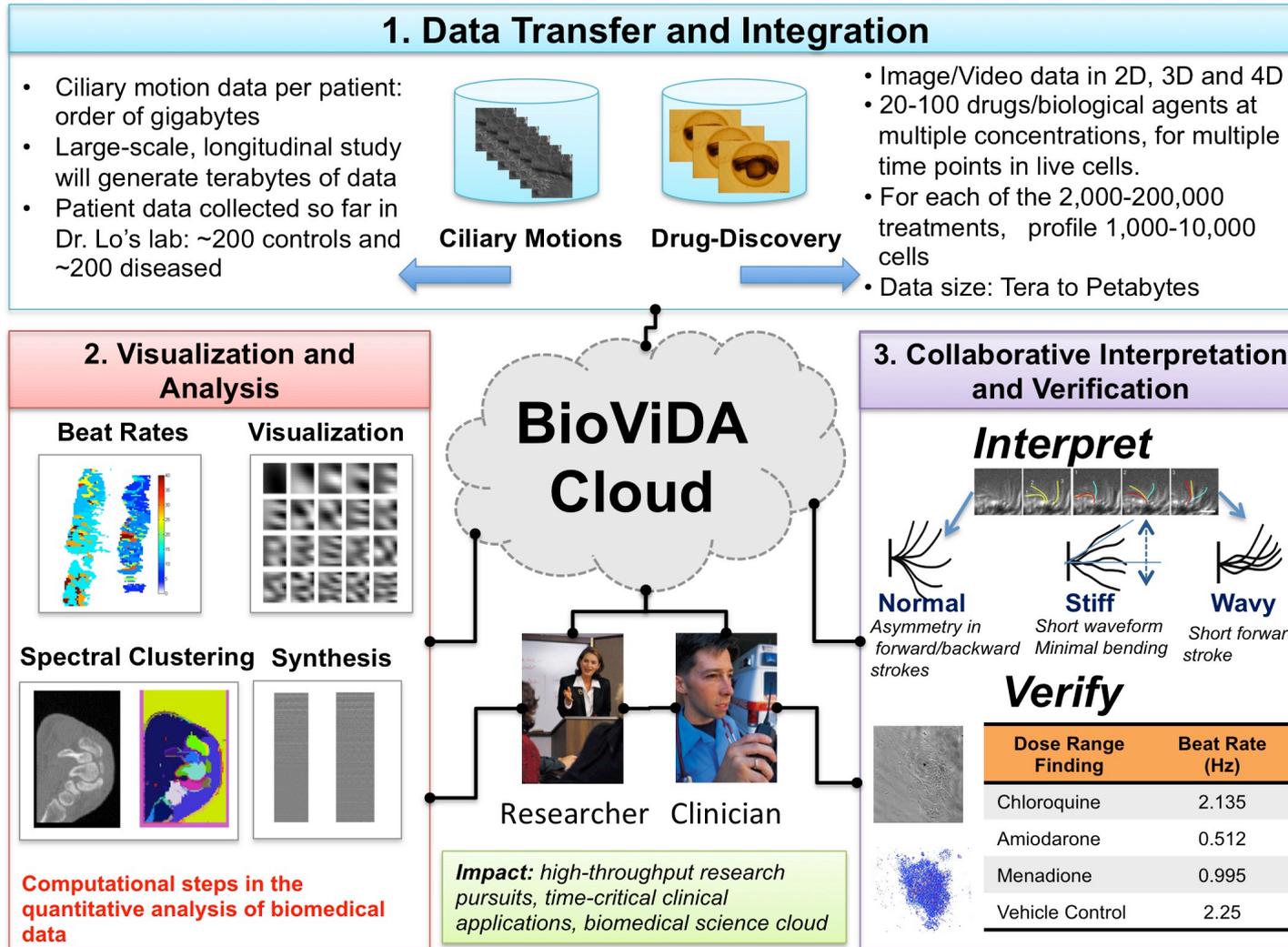
- Organizing high dimensional spaces
- Auto-regressive models
- Bio-medical imaging applications

# 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



# Example Implementation: Disease Diagnostics using BioViDA



# Bio-surveillance and the Cloud

## Bio-surveillance data

- is BIG and NOISY



- requires repetitive analysis in chunks



- modeling involves linear algebra and statistics



# 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](mailto:ramanathana@ornl.gov)