

Scientific Analytics & Visualization for Complex Biological Systems

Applications to Biomedical and Healthcare Sciences

Arvind Ramanathan

Health Data Sciences Institute,
Computational Science and Engineering
Division, Oak Ridge National Laboratory,
Oak Ridge, TN 37830

Ph: 865-576-7266

Webpage: <http://ramanathanlab.org>

Email: ramanathana@ornl.gov



Complex Biological Systems: A 100,000 feet overview...

[Nano] Integrating neutron-scattering with molecular simulations for reverse engineering intrinsically disordered protein function



Molecular/cellular Interactions

[Micro] Large-scale analytics and visualization for microbiomes and phylogenomic networks



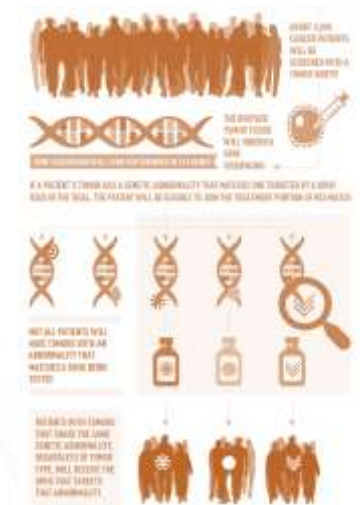
Genome

[Meso] High throughput reconstruction and phenotyping of neuron morphology (BigNeuron)



Communities

[Macro] Population health dynamics @ scale for infectious disease and cancer



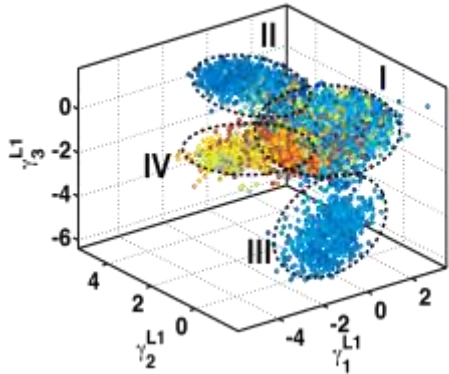
Population/Ecosystems

Overview of Scientific Software & Visualization Tools Developed

Nanoscale

Anharmonic Conformational Analysis

- Higher order statistics for large-scale analysis and visualization from MD: <http://anharmonic.net>
- ~ 100 downloads since 2015



8 papers (since 2011)
Processes large datasets *in memory* and *in situ*
2 Workshops (Telluride, CO)

Microscale

Dtree – Large-scale Phylogenetic Tree Construction + Visualization

- Large-scale phylogenetic trees <http://dtree.ornl.gov>
- About 200-300 users

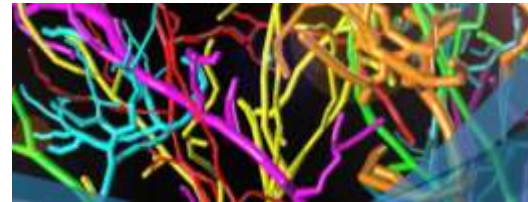


7 papers (since 2014)
Ongoing collaboration with Mike Leuze (CMSD), David Ussery (BSD)

Mesoscale

BigNeuron: Visualization of Neuron Morphology Datasets

- Petabyte-scale data visualization <http://vaa3d.org>
- Visual analytics created specifically for EVEREST visualization center

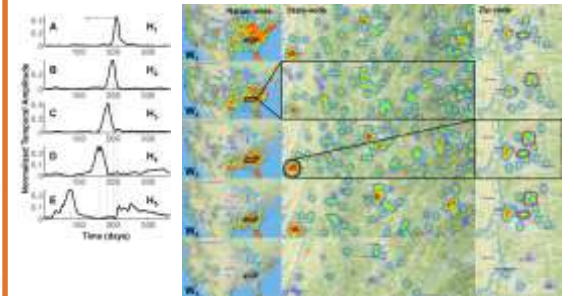


Over 30 different (worldwide) users
Large-scale visualization
2 Workshops (London, ORNL)

Macroscale

Oak Ridge Bio-surveillance Toolkit (ORBiT)

- Analysis of electronic healthcare reimbursement data for public health surveillance
- Analytics support for interactive visualization and simulations



11 papers (since 2011)
ORBiT demonstrated as part of Department of Homeland Security showcase

Outline

[**Nano**] Integrating neutron-scattering with molecular simulations for reverse engineering intrinsically disordered protein function



Molecular/cellular Interactions

[**Micro**] Large-scale analytics and visualization for microbiomes and phylogenomic networks



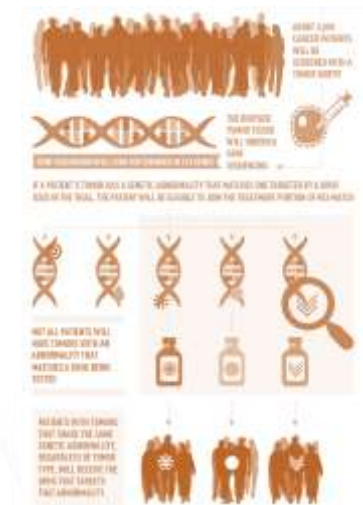
Genome

[**Meso**] High throughput reconstruction and phenotyping of neuron morphology (BigNeuron)



Communities

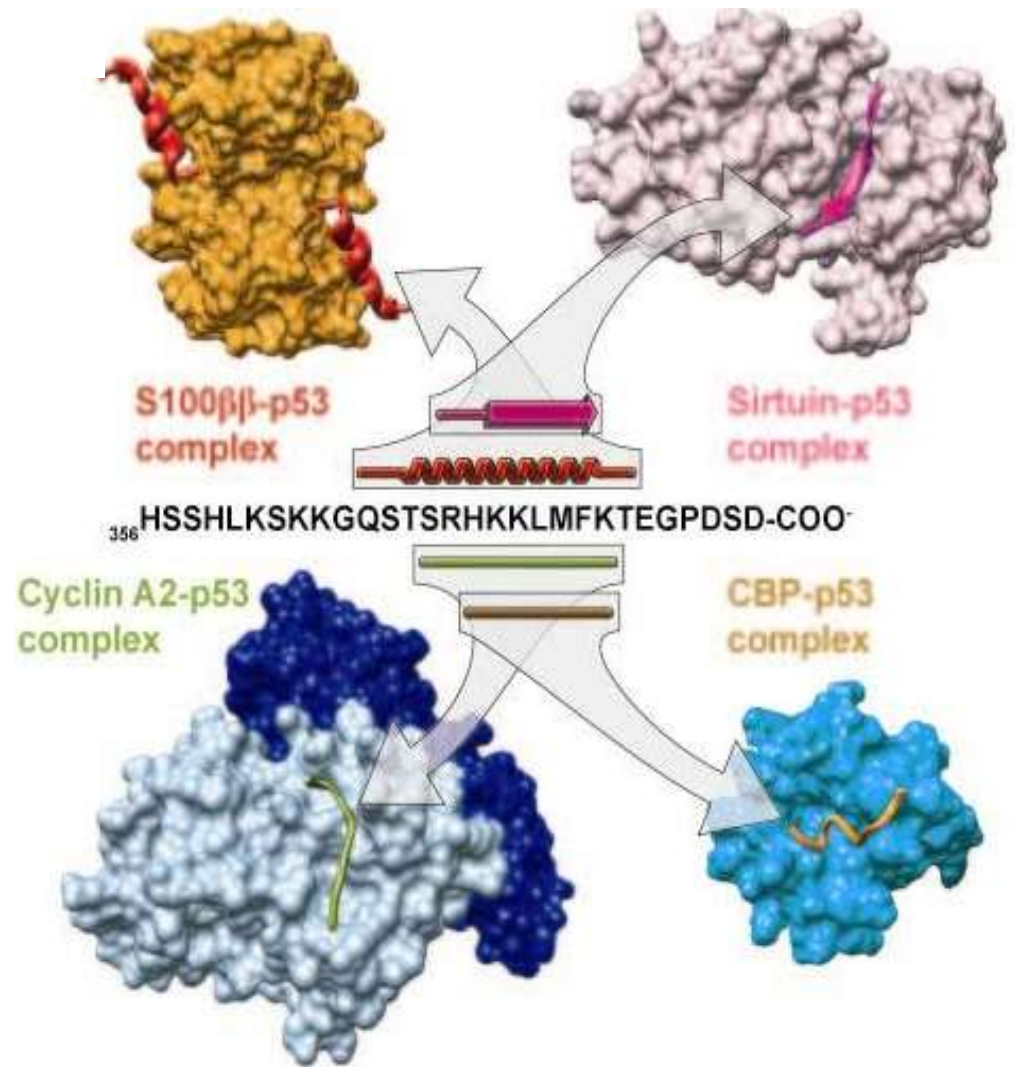
[**Macro**] Population health dynamics @ scale for infectious disease and cancer



Population/Ecosystems

Intrinsically Disordered Proteins

- Proteins without a stable tertiary structure:
 - High flexibility
 - Adaptable binding interfaces
- Over 35% of eukaryotic proteome disordered
- 65% of these proteins are involved in diseases:
 - Neurodegenerative
 - Cardio-vascular
 - Diabetes



Nuclear Co-activator Binding Domain (NCBD)

The Story So Far ...

- NCBD hydrophobic core¹⁻⁴ is stable, with only minor rearrangements at micro-/milli-second time-scales:
 - Resembles ACTR-bound state⁵⁻⁶
 - Existence of at least two states: major and minor⁷
- NCBD recognition: conformational selection^{5-7,8,10,11?}
 - Presence of a dominant state that resembles ACTR-bound⁹
- Simulations:
 - No agreement on presence of minor states
 - Number of different mechanisms, with no consensus

1. Demarest, et al., 2002, Nature 415: 549-553

2. Qin, et al, 2005, Structure 13: 1269-1277

3. Lee, et al, 2010, Biochemistry 49: 9964-9971

4. Ebert et al, 2008, Biochemistry. 47:1299– 1308

5. Kjaergaard, et al, 2010, Proc. Natl. Acad. Sci. USA, 107:12535– 12540

6. Kjaergaard, et al, 2012, Biophys. J., 102: 1267-1275

7. Kjaergaard, et al, 2013, Biochemistry,

8. Naganathan, et al, 2011, JACS, 133:12154–12161

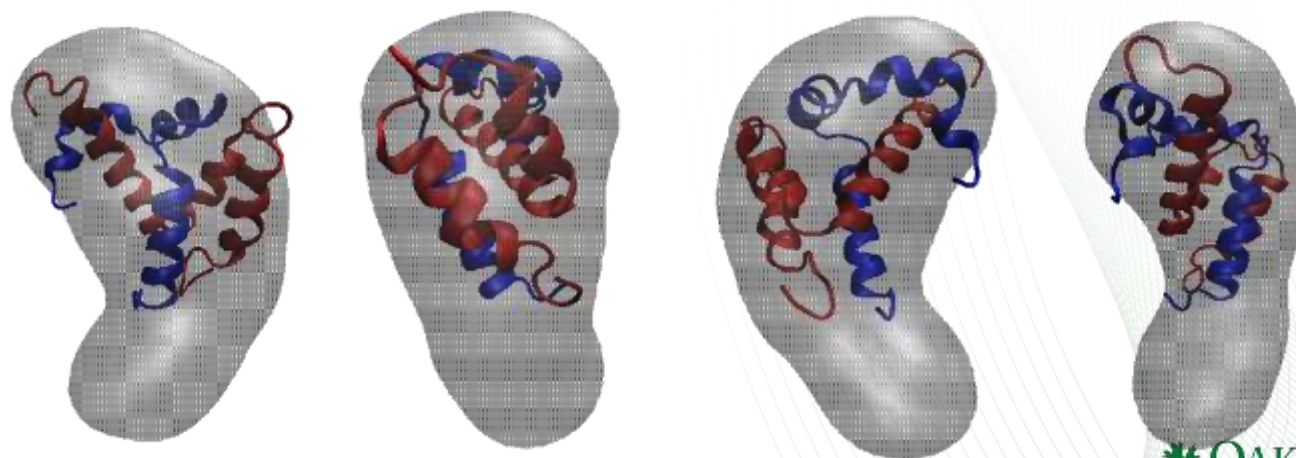
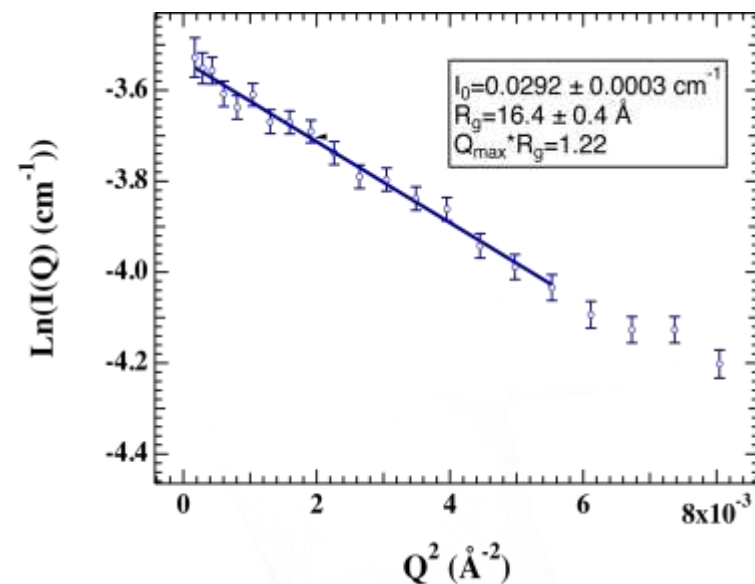
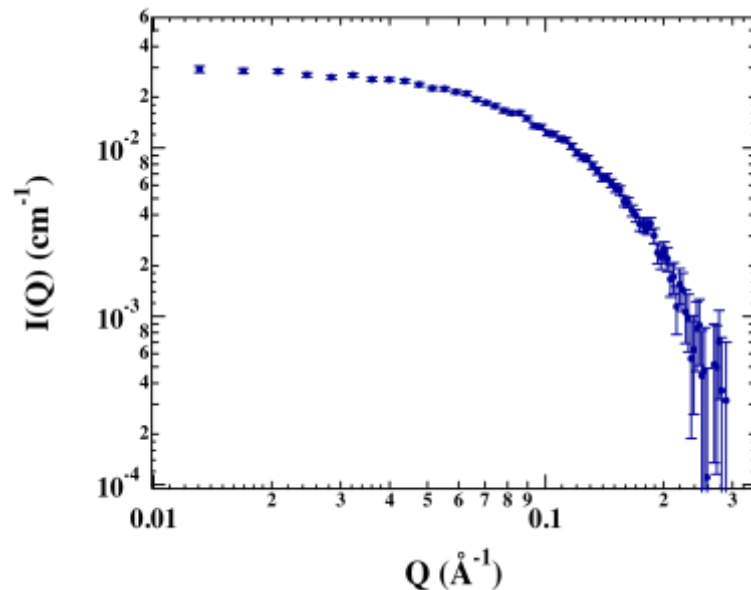
9. Burger, et al, 2012, Pac. Symp. on Biocomput., 17: 70-81

10. Ganguly, et al, 2012, PLoS Comp. Biol.,

11. Knott & Best, 2012, PLoS Comp. Biol.

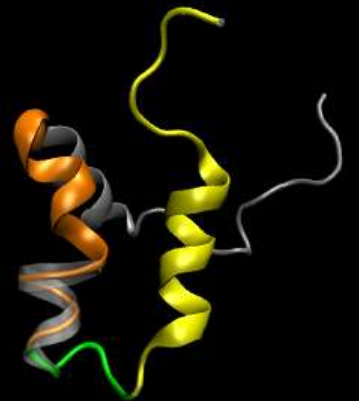
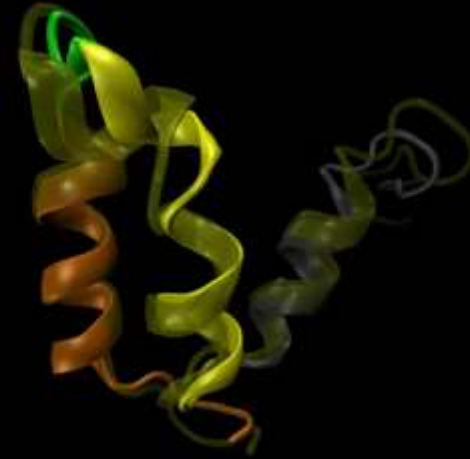
Small Angle Neutron Scattering (SANS) Experiments are Sparse...

- Only overall shape and size information available from SANS experiments
- Challenge:
 - *How to improve the information content / resolution of SANS experiments?*



Long time-scale simulations of IDPs

- Molecular dynamics (MD) simulations:
 - $O(10000)$ - $O(\text{million})$ atoms
 - Integrate over 15 orders of timescales (from femtoseconds to seconds)
- Efficient algorithms for scaling MD simulations on OLCF:
 - GP/GPU + CPU hybrid
- $O(\text{ten})$ days to generate millisecond timescale trajectories



Long Time-scale simulations are challenging for data analytics

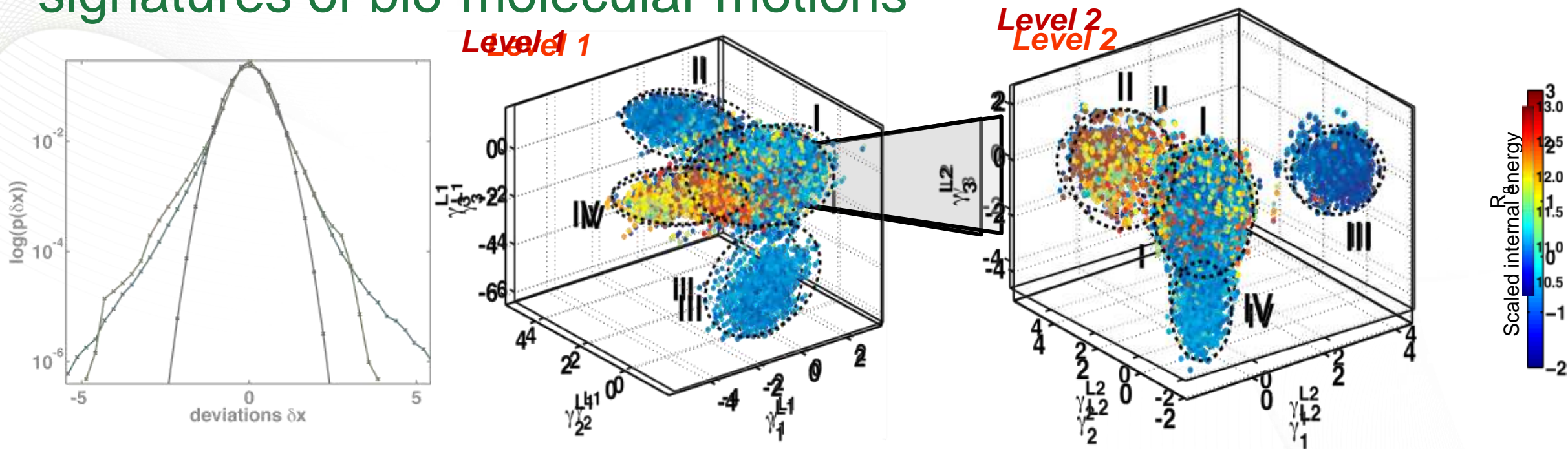
- High dimensional data
- Analyze simulations to identify biologically relevant information
- Integrate information from simulations with experiments
 - Have little/no knowledge of how simulations relate to experiments!

Solutions Developed

- **Event detection techniques** for storyboard organization of simulations
- Using **higher-order statistical signatures** (Anharmonic Conformational Analysis) to organize high-dimensional simulation datasets
- **Bayesian methods** to integrate and improve neutron scattering experiment resolution

Ramanathan, A., et al, Acct. Chemi. Res. (2014)
Ramanathan, A. et al, Scientific Rep. (2015)
Ramanathan, A. et al, Biochemistry (2016)

Anharmonic Conformational Analysis reveals intrinsic signatures of bio-molecular motions



- Long tailed fluctuations in the conformational landscape of IDPs
- Higher order signatures in the landscape give rise to conformational substates similar in structure, dynamics and energetic features

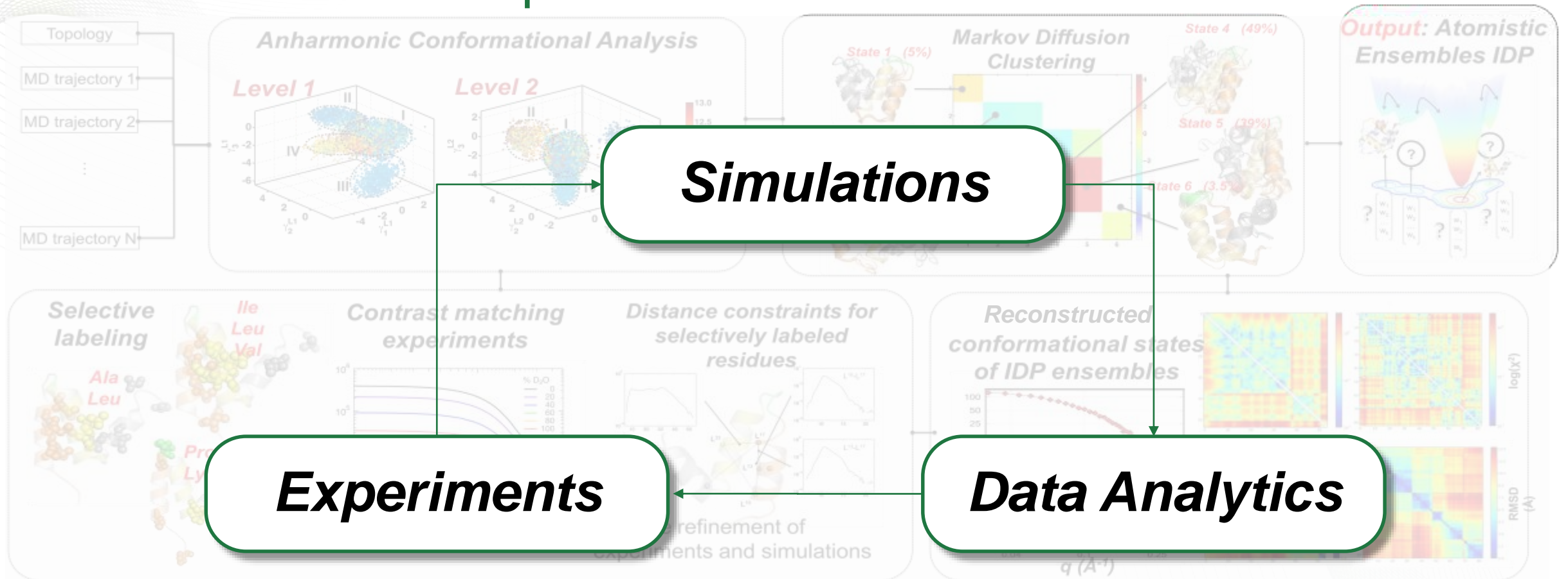
Ramanathan, A., Agarwal, P.K., Kurnikova, M.G., Langmead, C.J., (2010) J. Comp. Biol.

Ramanathan, A., J. Oh-Yoo, Langmead, C.J., (2011) J. Chem. Theory & Comput

Ramanathan, A., Agarwal, P.K., (2011) PLoS Biology

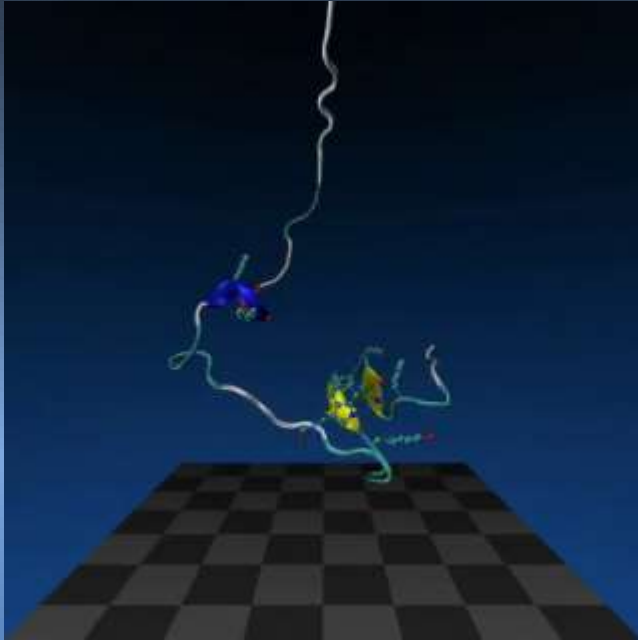
Burger, V.M., Ramanathan, A., Savol, A.J., Stanley, C.B., Agarwal, P.K., Chennubhotla, C.S., (2012) Pacific Symposium on Biocomputing

Bayesian Framework to Integrate and Inform Higher Resolution SANS Experiments

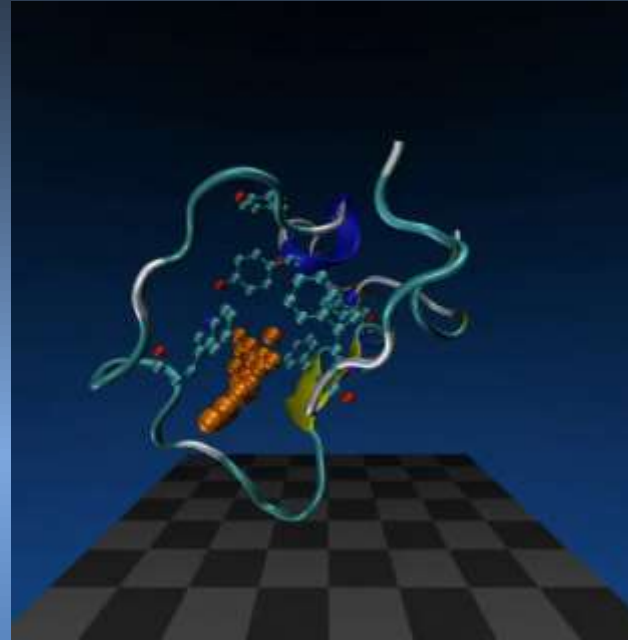


- Improvement of SANS resolution by integrating simulations with data analytics to drive experiments
- Effective use of heterogeneous computing resources
- Open Source Python-based software

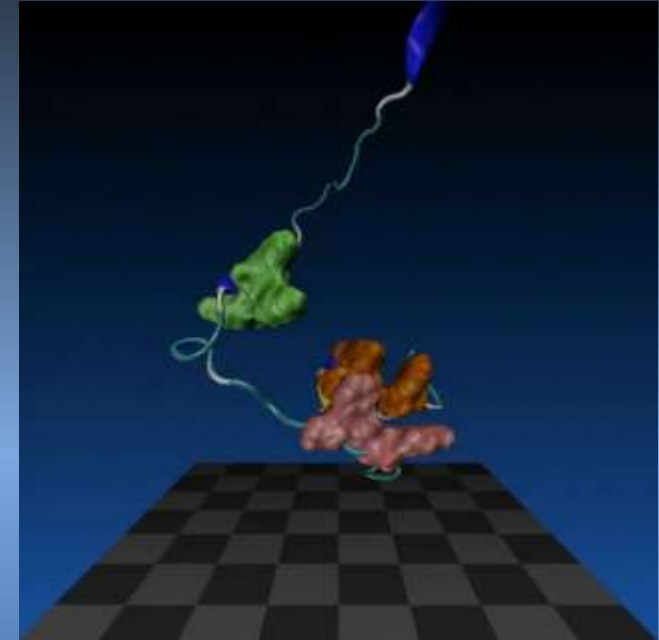
Experiments & Simulations can lead to better drug discovery



P27 Apo-state simulations (~10 μ s)



P27 Holo-state simulations (~15 μ s)



P27 Story Board

- Long time-scale simulations and molecular docking lead to biophysical insights for the intrinsically disordered protein p27
- Designed molecular scaffolds against which wet-lab experiments could validate such insights

Outline

[**Nano**] Integrating neutron-scattering with molecular simulations for reverse engineering intrinsically disordered protein function



Molecular/cellular Interactions



[**Micro**] Large-scale analytics and visualization for microbiomes and phylogenomic networks



Genome



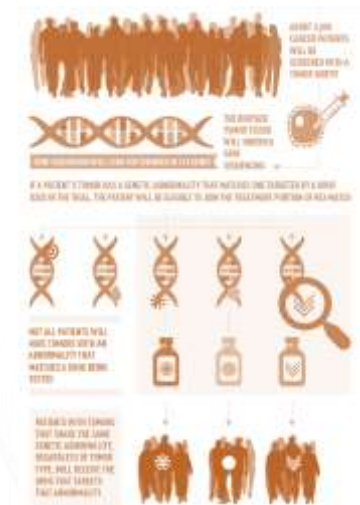
[**Meso**] High throughput reconstruction and phenotyping of neuron morphology (BigNeuron)



Communities

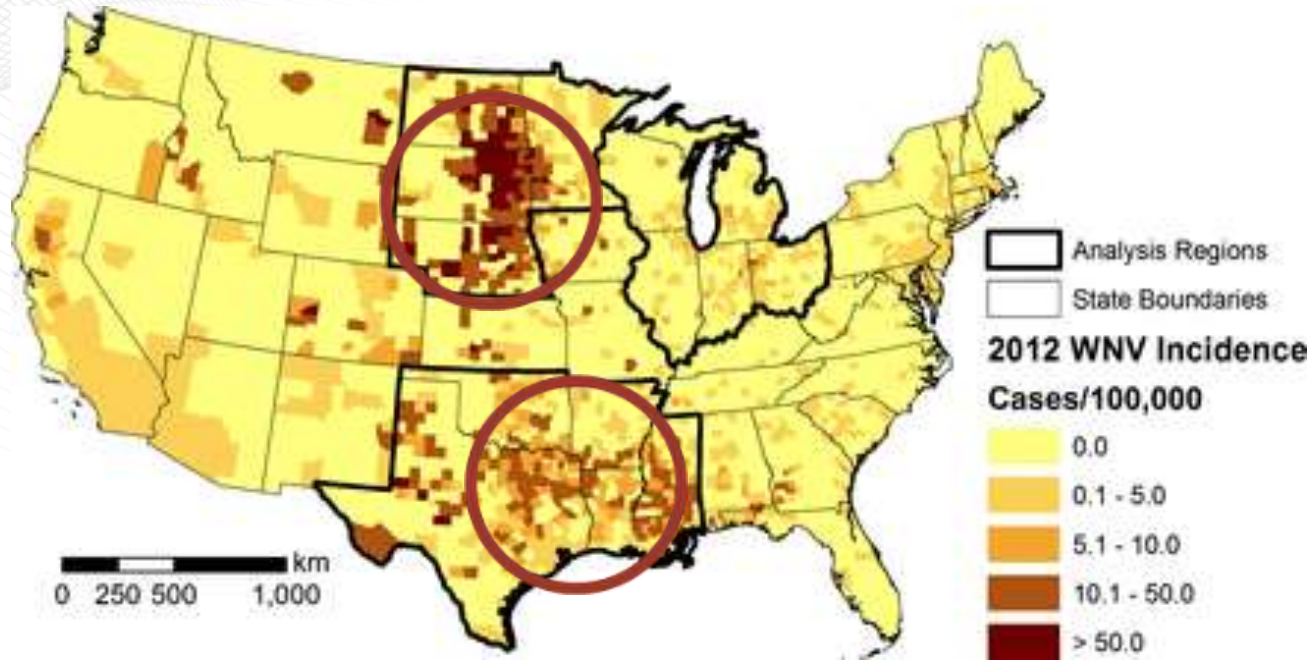


[**Macro**] Population health dynamics @ scale for infectious disease and cancer



Population/Ecosystems

Motivation: How do ecological factors affect viral evolution?



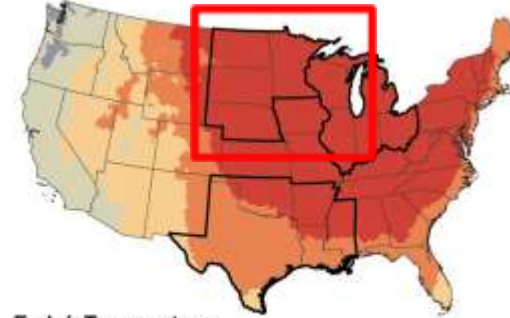
A January Temperature



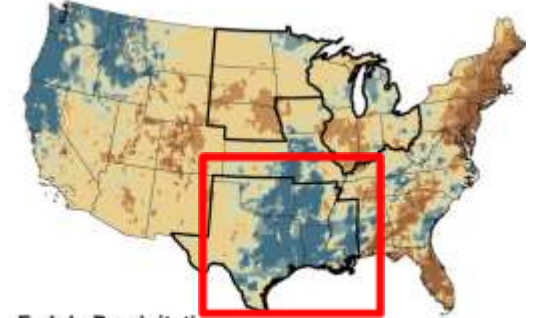
B January Precipitation



C March Temperature



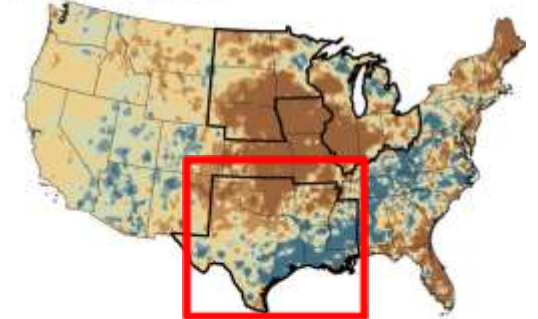
D March Precipitation



E July Temperature

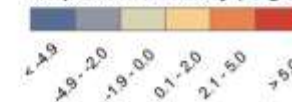


F July Precipitation

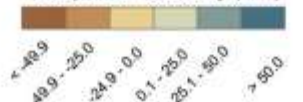


Analysis Regions
State Boundaries

Temperature Anomaly (Degrees C)



Precipitation Anomaly (mm)

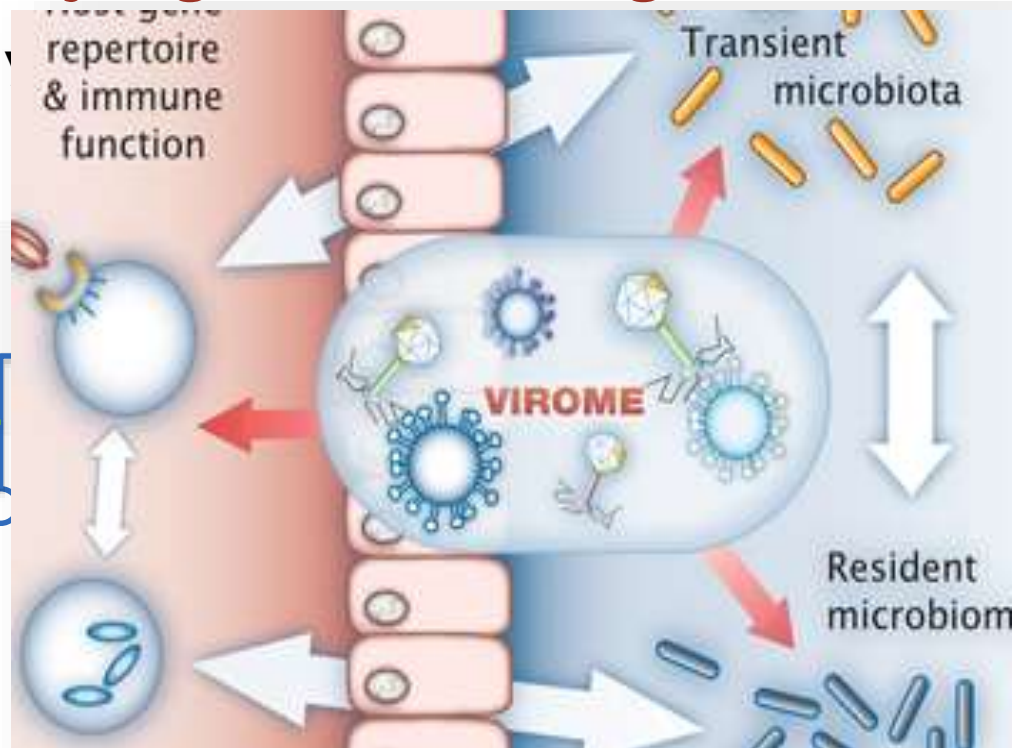


- Higher winter temperatures \rightarrow WNV incidence
- Anomalies in precipitation \rightarrow WNV incidence

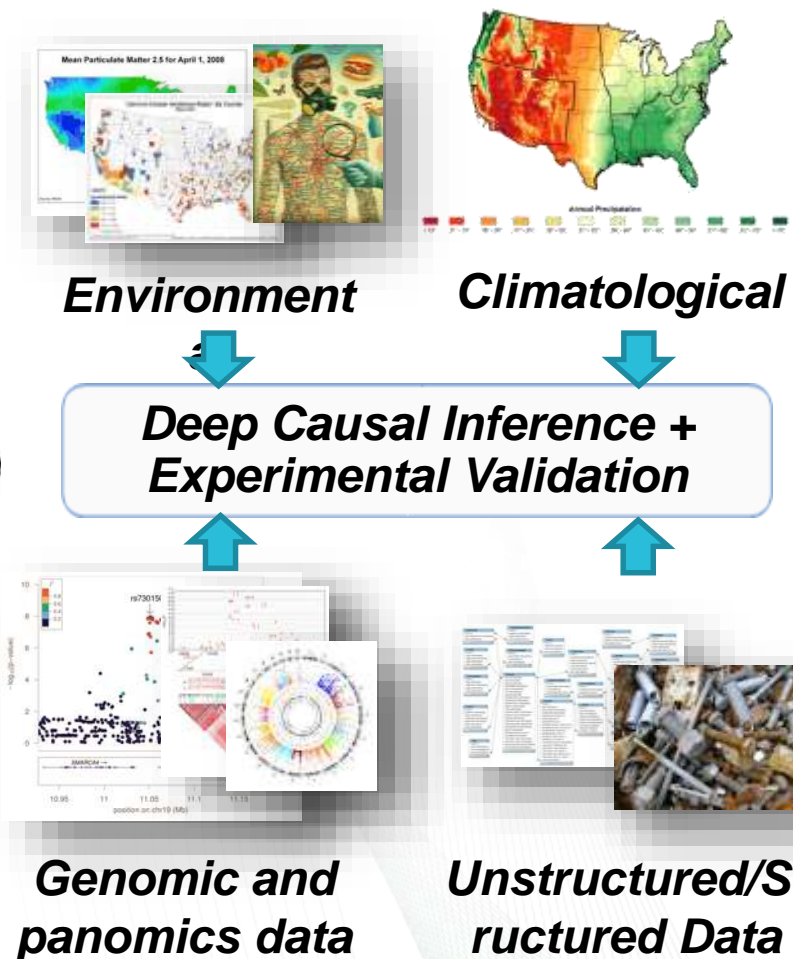
Wimberly, et al, Am. J. Trop. Med. Hyg. (2014): 91 (4), pp 677-684

Looking beyond the genome: Microbiome

Phylogenetic lineage of WNV



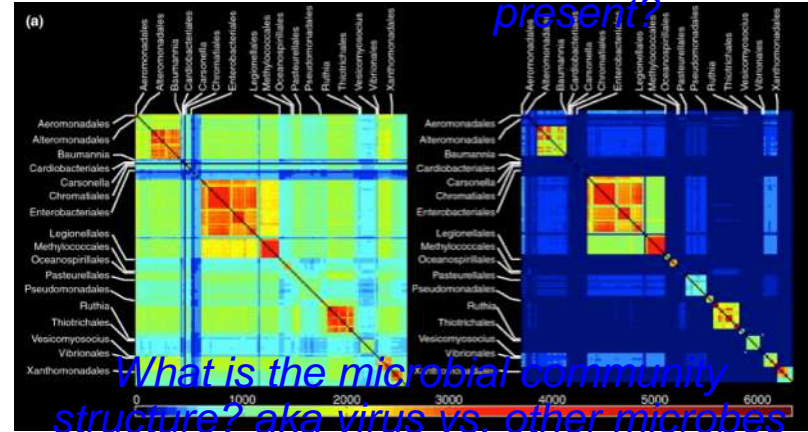
- **Clim**
- **Environment:** Chemicals, particulate matter
- **Viral-Vector/Reservoir Dynamics:** Microbiome



Scalable ML Approaches to Extract Metagenomic Signatures and Phylogenomic Networks

- Within a viral + vector/reservoir sample, identify what species are present:
 - metagenomic signatures
 - individual species level or taxa level
- Construct phylogenomic networks¹:
 - identify which species/ taxa “talk to each other” in the sample
 - Shared genes or proteins determined from metagenomic signatures
 - Comparison of phylogenomic networks → changes in microbiome interactions

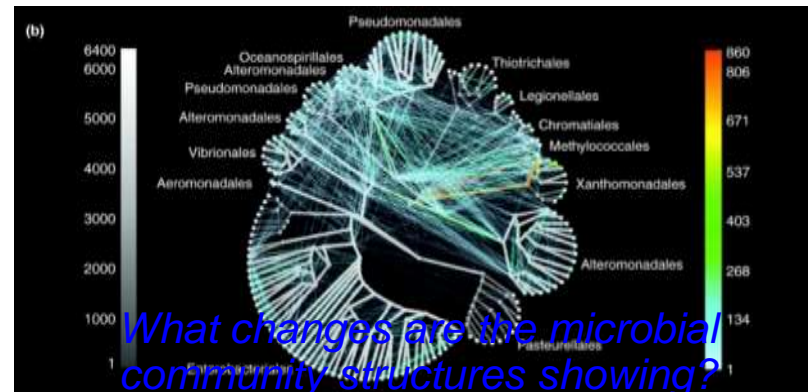
What species are present?



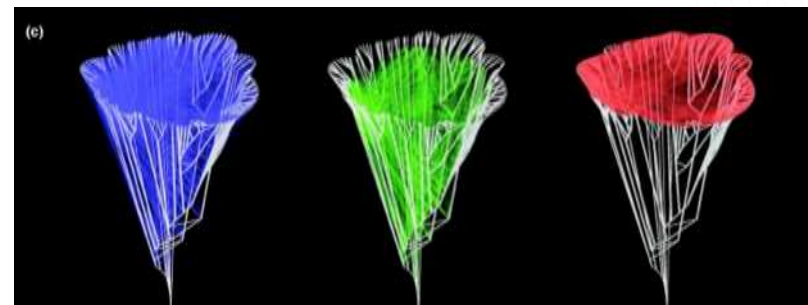
What is the microbial community structure? aka virus vs. other microbes



dtree.ornl.gov



What changes are the microbial community structures showing?



Outline

[**Nano**] Integrating neutron-scattering with molecular simulations for reverse engineering intrinsically disordered protein function



Molecular/cellular Interactions



[**Micro**] Large-scale analytics and visualization for microbiomes and phylogenomic networks



Genome



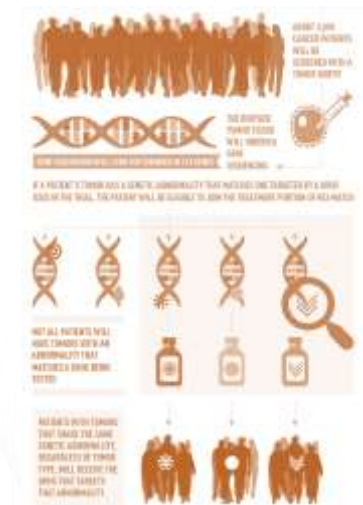
[**Meso**] High throughput reconstruction and phenotyping of neuron morphology (BigNeuron)



Communities



[**Macro**] Population health dynamics @ scale for infectious disease and cancer



Population/Ecosystems

BigNeuron: High-throughput imaging for neuron morphology reconstruction and phenotyping

- A community wide effort to:
 - evaluate the state-of-the-art of single neuron reconstruction
 - standardize the protocols
 - establish a big data resource for neuroscience
- Over 200 terabytes of processed datasets
- Initial testing of 30,000 neuronal reconstructions across 26 different algorithms

<http://alleninstitute.org/bigneuron/about/>



ABOUT

OVERVIEW

DATA

ALGORITHMS

HACKATHONS & WORKSHOPS

SUPERCOMPUTING

HOW TO PARTICIPATE

FAQ

TERMS & CONDITIONS

CONTACT

GET INVOLVED >

ABOUT | OVERVIEW | DATA | ALGORITHMS | HACKATHONS & WORKSHOPS | SUPERCOMPUTING | HOW TO PARTICIPATE | FAQ | TERMS | CONTACT

OVERVIEW



Integration of synaptic
state of the organism,
finding how neurons

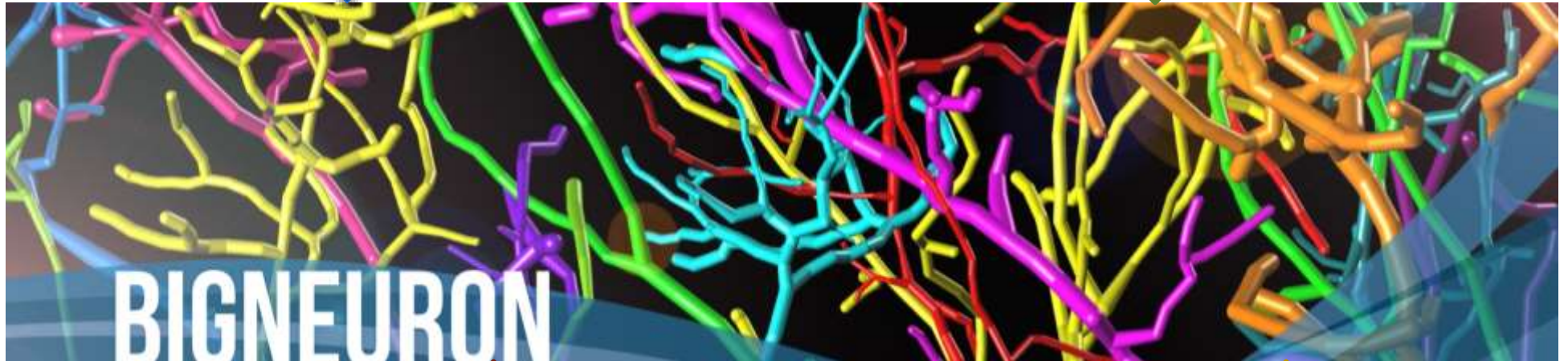
is from dozens of different
ita. Dozens of different
of neurons from labs
nce field needs standards
table for analysis, and for
ing the field.

available single 3D neuron image data sets acquired by several light

BigNeuron Overview

Algorithms ($M > 20$)
Algorithm porting and data
analysis hackathons (4)

Neuron Images ($N > 30K$) Annotation
workshop to establish “gold standard”
manual reconstruction

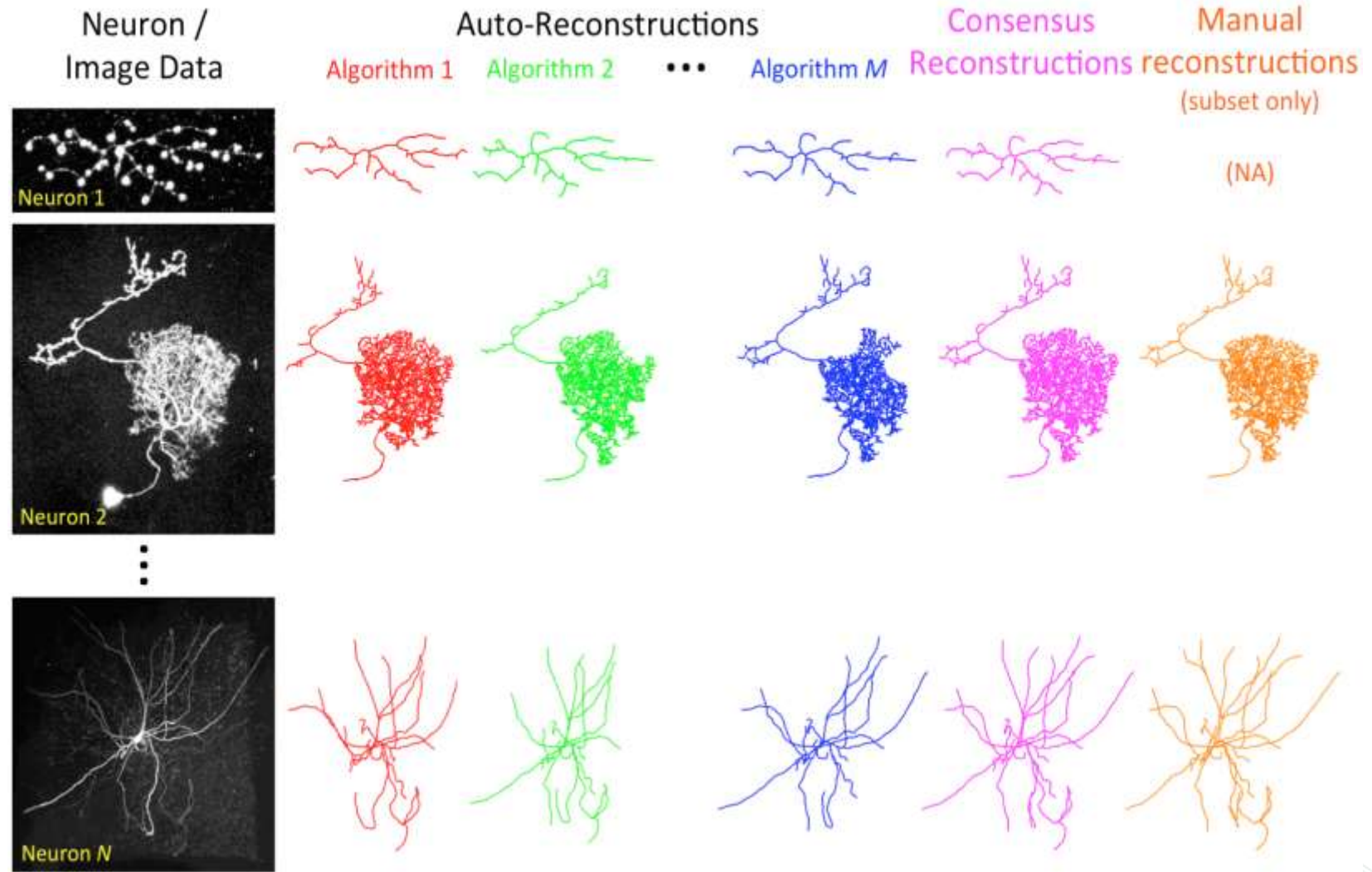


Supercomputing (4 centers) Bench testing of
algorithms across all images

Data hosting
(1-3 mirror sites)

Supercomputing (Big Data Analytics) + Visualization = Processing O(millions) of images in < O(hours)!!

- Novel use of visualization capabilities at ORNL for neuroscience data
- Interactive analytics enables improvement of neuroscience related algorithmic workflows



A. Ramanathan, H. Peng, Neuroinformatics (2016)

Outline

[**Nano**] Integrating neutron-scattering with molecular simulations for reverse engineering intrinsically disordered protein function



Molecular/cellular Interactions



[**Micro**] Large-scale analytics and visualization for microbiomes and phylogenomic networks



Genome



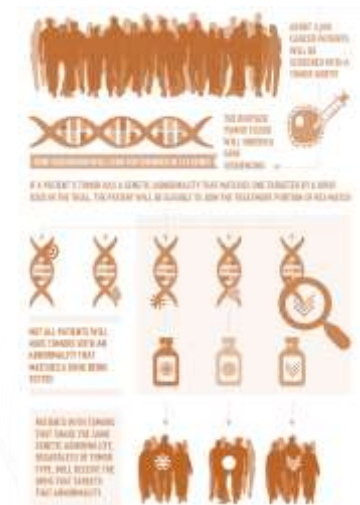
[**Meso**] High throughput reconstruction and phenotyping of neuron morphology (BigNeuron)



Communities

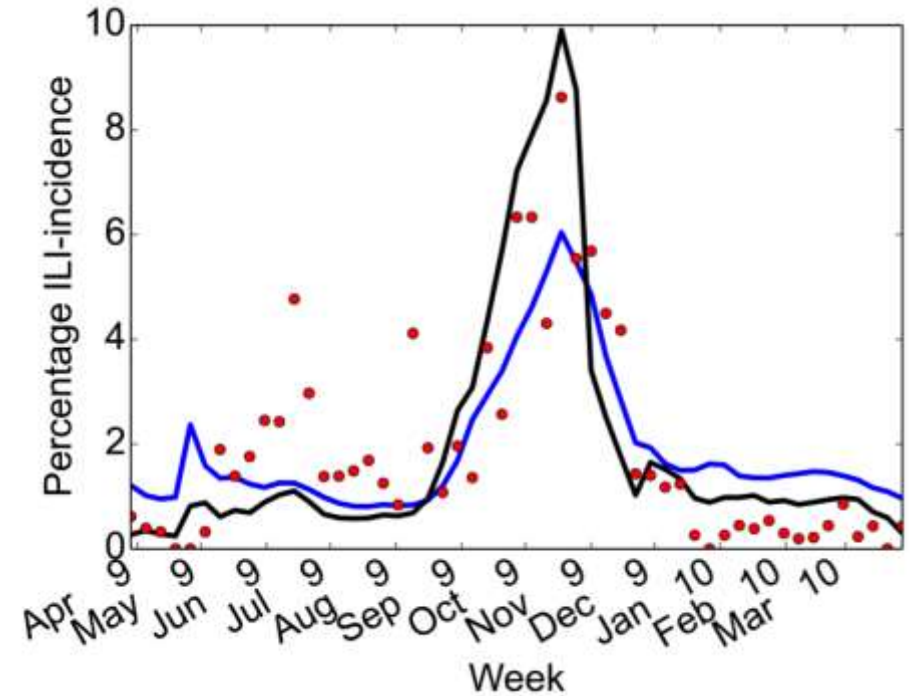
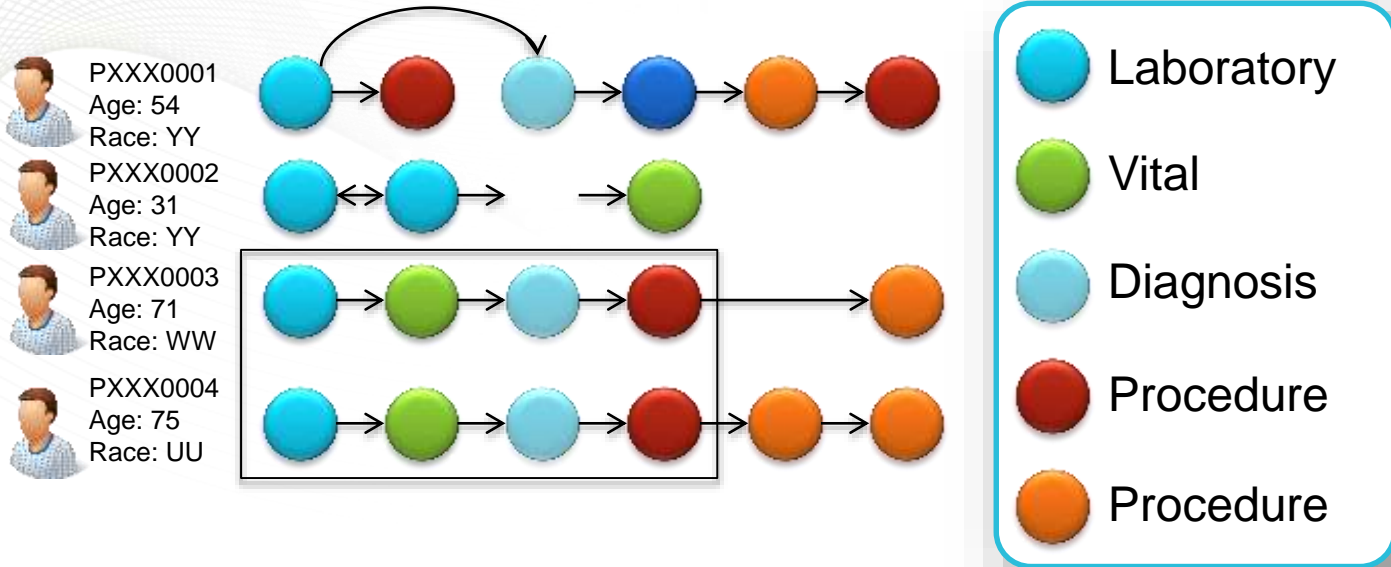


[**Macro**] Population health dynamics @ scale for infectious disease and cancer



Population/Ecosystems

Motivation: Public Health Dynamics using Big Data Analytics



- Given EMR data of patients, how to find:
 - sequentially ordered sets of commonly carried out clinical procedures?
 - find patterns of common clinical procedures across two or more clinical conditions?

Electronic Healthcare Reimbursement Claims (eHRC) are complementary to Electronic Medical Records

Payers adjudicates claim for reimbursement

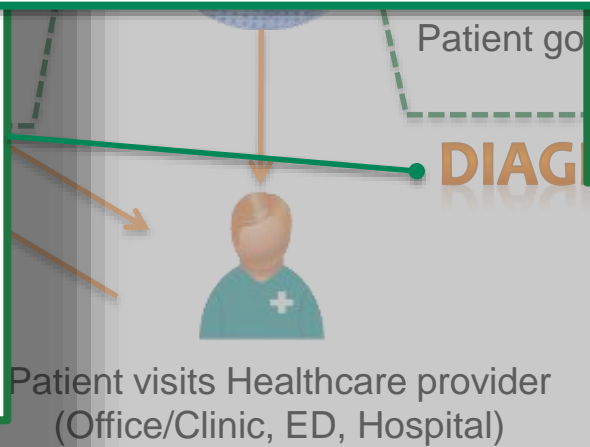
Adjudicate the claim for the payer

Adjudicate the claim for

Pharmacy Data (Rx)

Can eHRC data be used for public health dynamics?

- Date of service
- Age, gender, zip
- ICD-9 diagnoses
- Procedure codes
- Provider identifier
- Provider specialty
- **Provider zip**



- ~70-80% of all dispensed Rx
- ~200 million patients

- Anonymized patient data
- HIPAA compliant
- High fidelity
- Thorough coverage

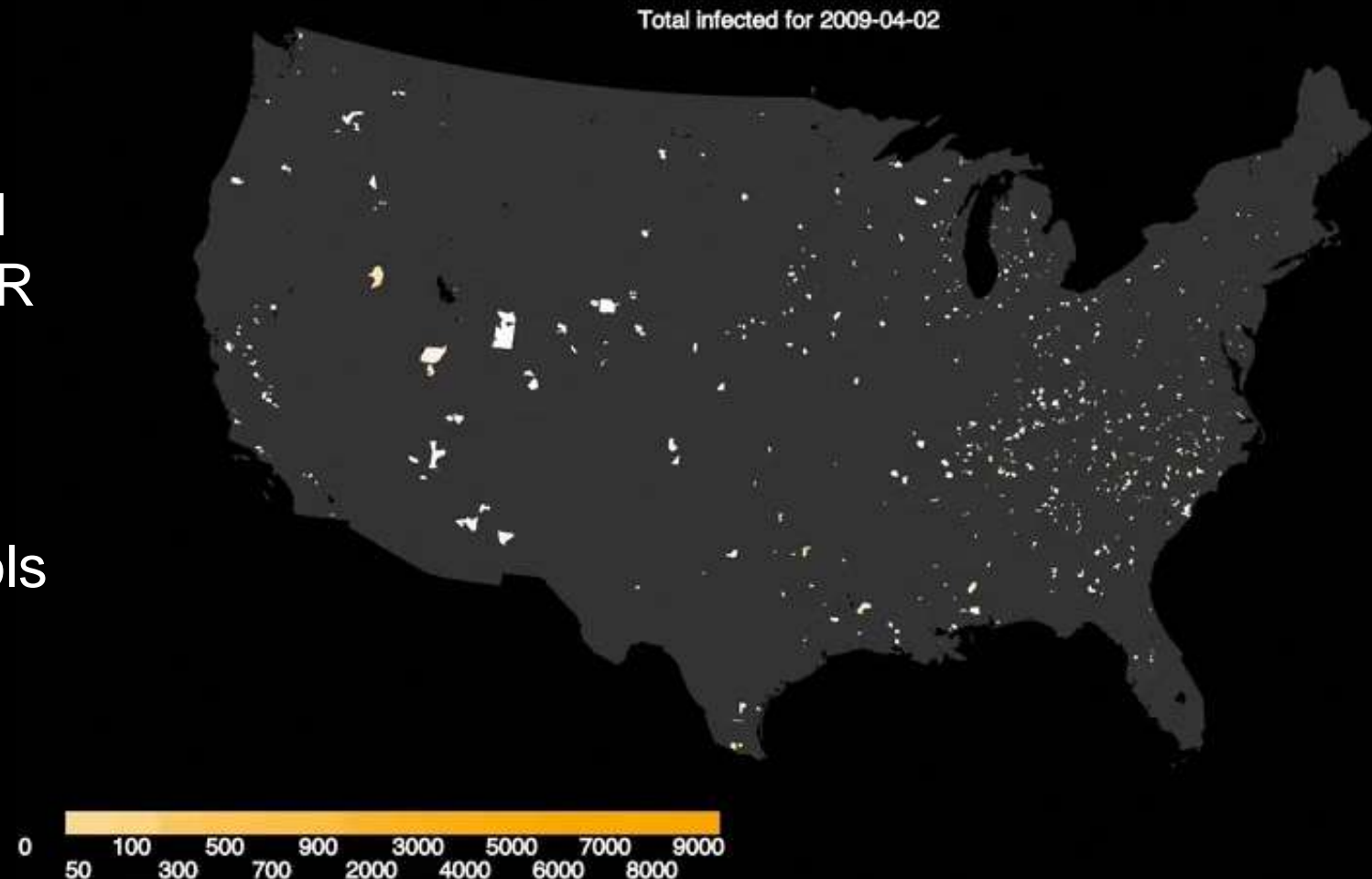
Extracting Meaning from eHRC data: Geo-temporal Patterns from the 2009-2010 Pandemic Flu Season

Challenges

- Meaningful spatial and temporal pattern extraction
- Interactive visualization and analytic framework for eHCR datasets

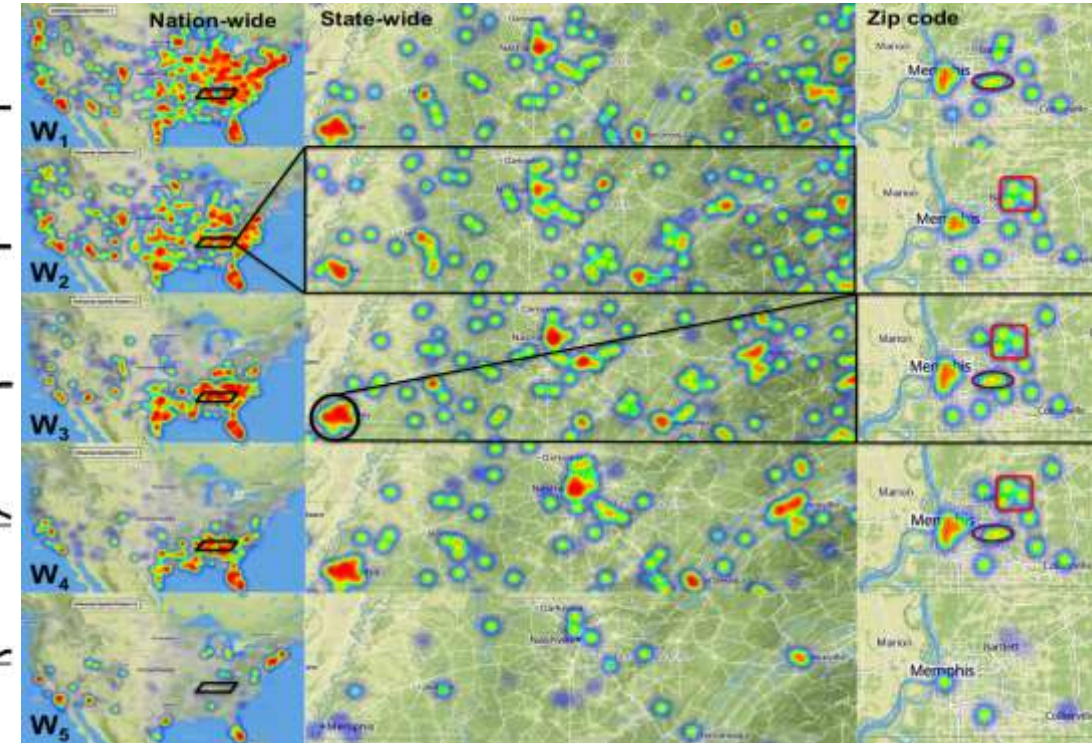
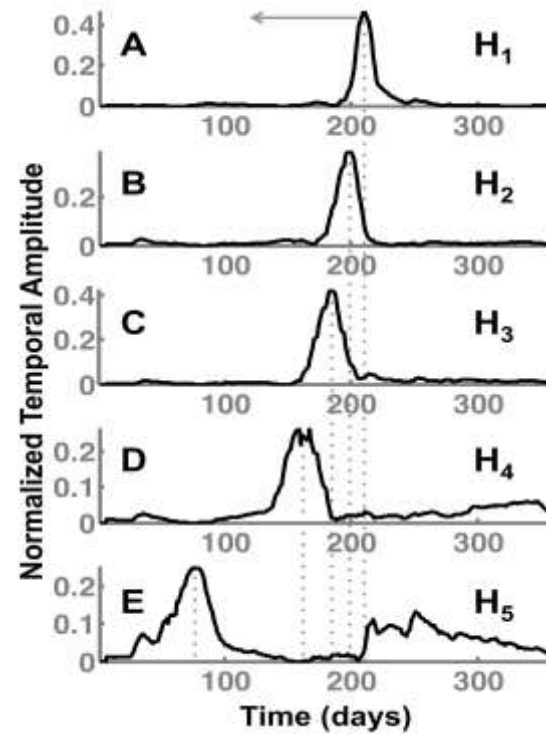
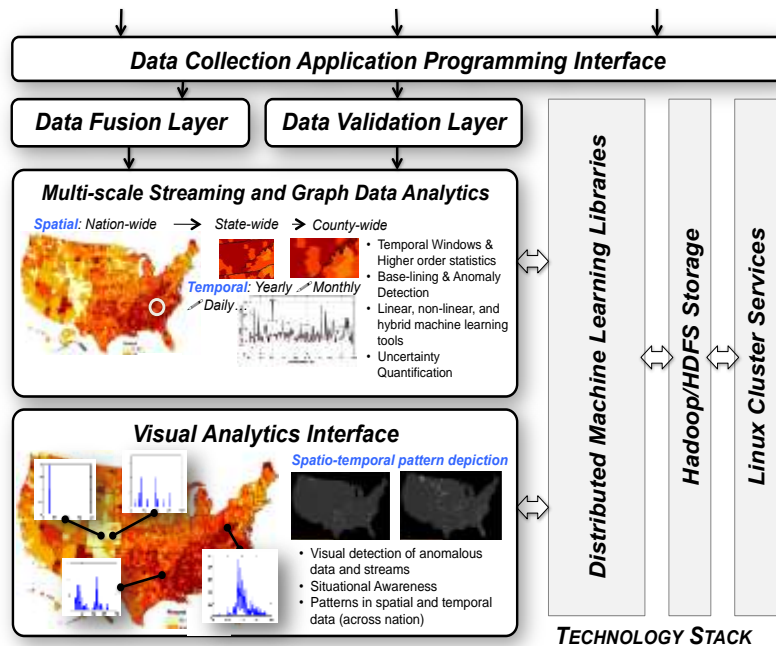
Opportunity

- Developing scalable ML tools to analyze large volumes of data



Oak Ridge Bio-surveillance Toolkit (ORBiT)

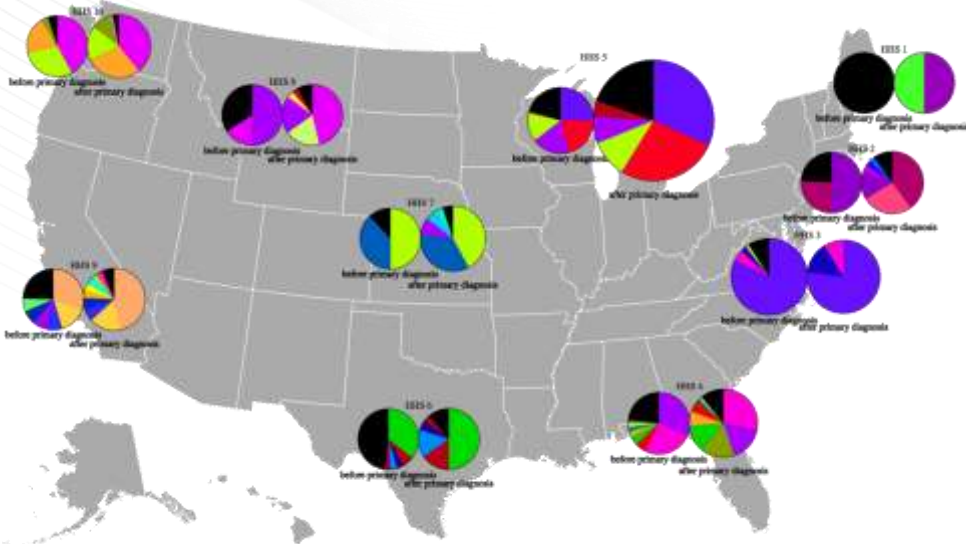
eHCR Datasets



- ORBiT: Automated tools to extract, analyze, and visualize eHCR data for public health surveillance.
- Input to modeling and simulation toolkits to appropriately bound runtime parameter settings
- Data analytic tools highlighted as part of DHS technology showcase

Novel Insights from Sequential Mining of eHRC datasets

Clinical Patterns for Heart Disease



- Sequential pattern mining can be informative regarding:
 - which clinical procedural sequences are commonly followed for patients
 - what are the costs before and after treatment
 - heterogeneity and general lack of consensus on how patients are treated
- Future work:
 - Understanding clinical outcomes based on procedures followed
 - Clinical effectiveness studies
 - Optimizing costs and assessing risks based on patient outcomes

Summary

- Data science is central for Quantitative Biology:
 - Need novel algorithms [theory]
 - Scalable tools [engineering]
 - Hybrid platforms that can provide seamless transitions both from programming and deployment perspectives
- Vignettes illustrating diverse bio-medical/biological applications:
 - Feedback between experiments, simulations and theory for improving resolution of experimental observables
 - Data-driven simulations that can be used to guide “what-if” scenarios

Funding & Acknowledgements

- **Funding:**

- ORNL SEED (7278, 7280),
- LDRD (7417),
- NIH GHU CTSA

- **People:**

- **Students:** James C. Pino (Vanderbilt), Thomas Pospiech, Corrine Nief, Patrick O'Brien (Yale), Gabe Vacaluic, Townes Dean-Bouchard, Elizabeth Tuggle, Marianne Catanho
- **Collaborators:** Pratul K. Agarwal, Chakra S. Chennubhotla, Michael Leuze, David Ussery, Georgia Tourassi, Laura L. Pullum, Shannon P. Quinn