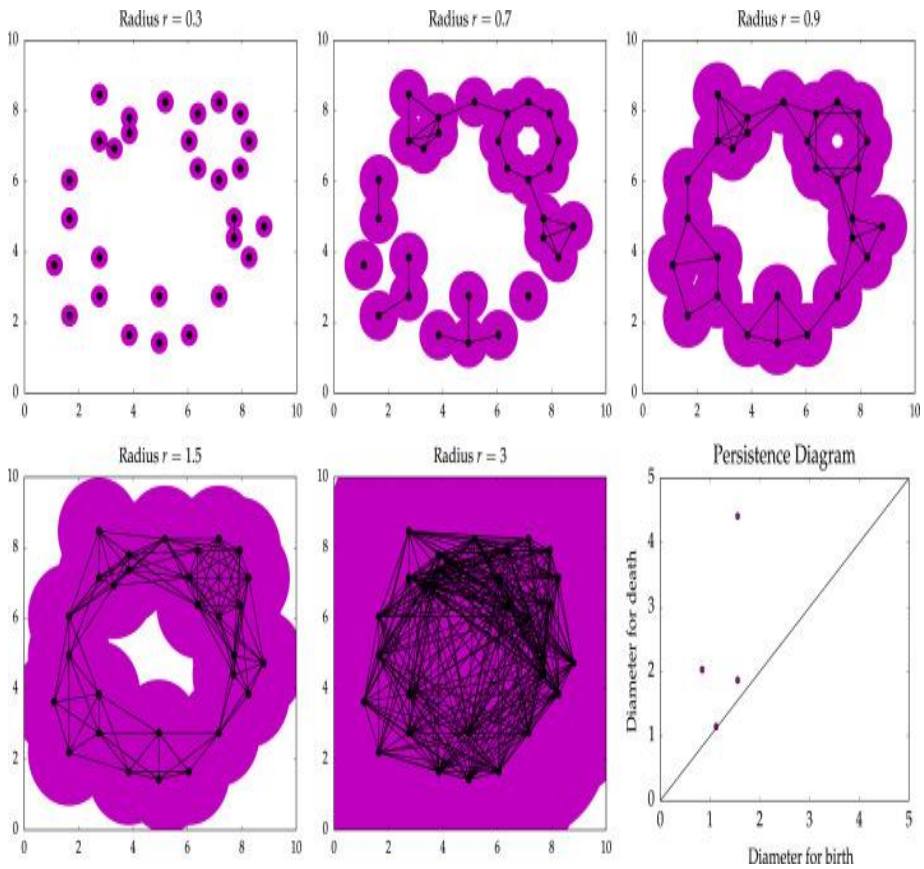
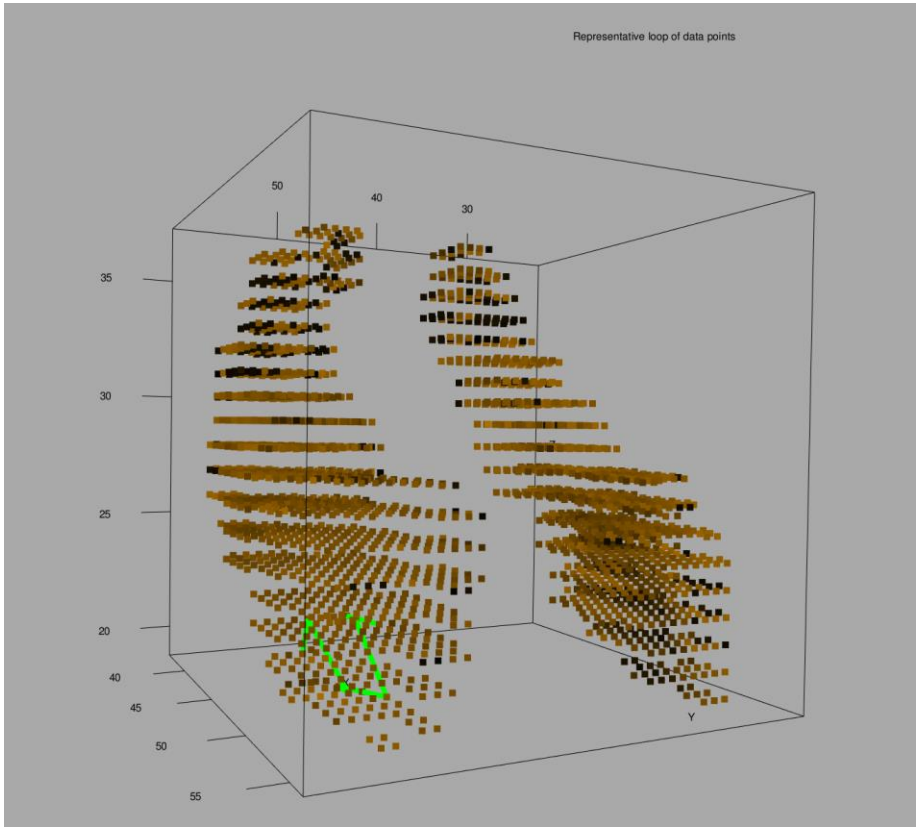
Hassan Abdallah
Wayne State University, Department of Mathematics





Hassan Abdallah
Wayne State University, Department of Mathematics

Challenges:

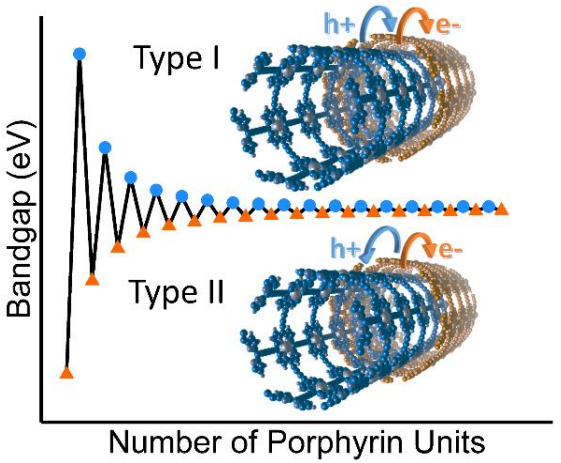
- Current tools are memory-hungry
- How best to parallelize?
- Most efficient method depends on question being asked. How to cover most bases in single design?

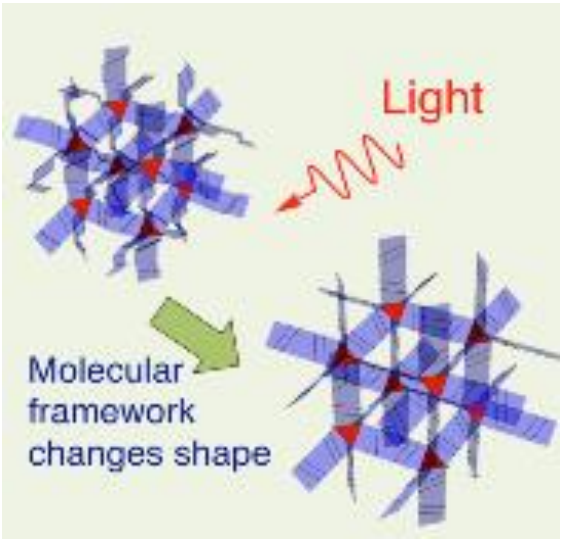
What I do...

UNIVERSITY OF CALIFORNIA UCRIVERSIDE

Density functional theory

$$\hat{\mathrm{H}}\,|\Psi\rangle = E|\Psi\rangle$$





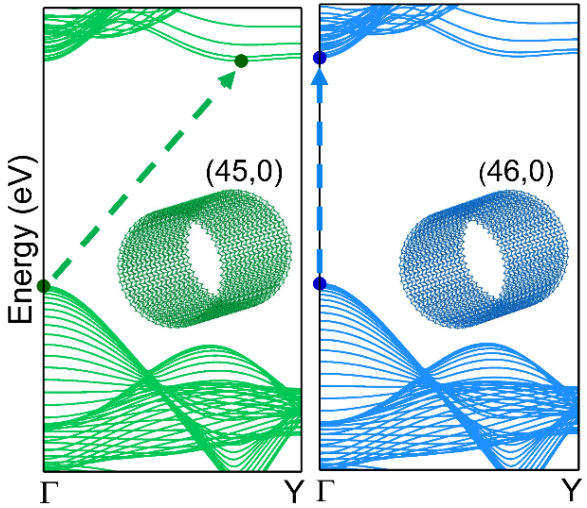
Computational materials science

Electron dynamics

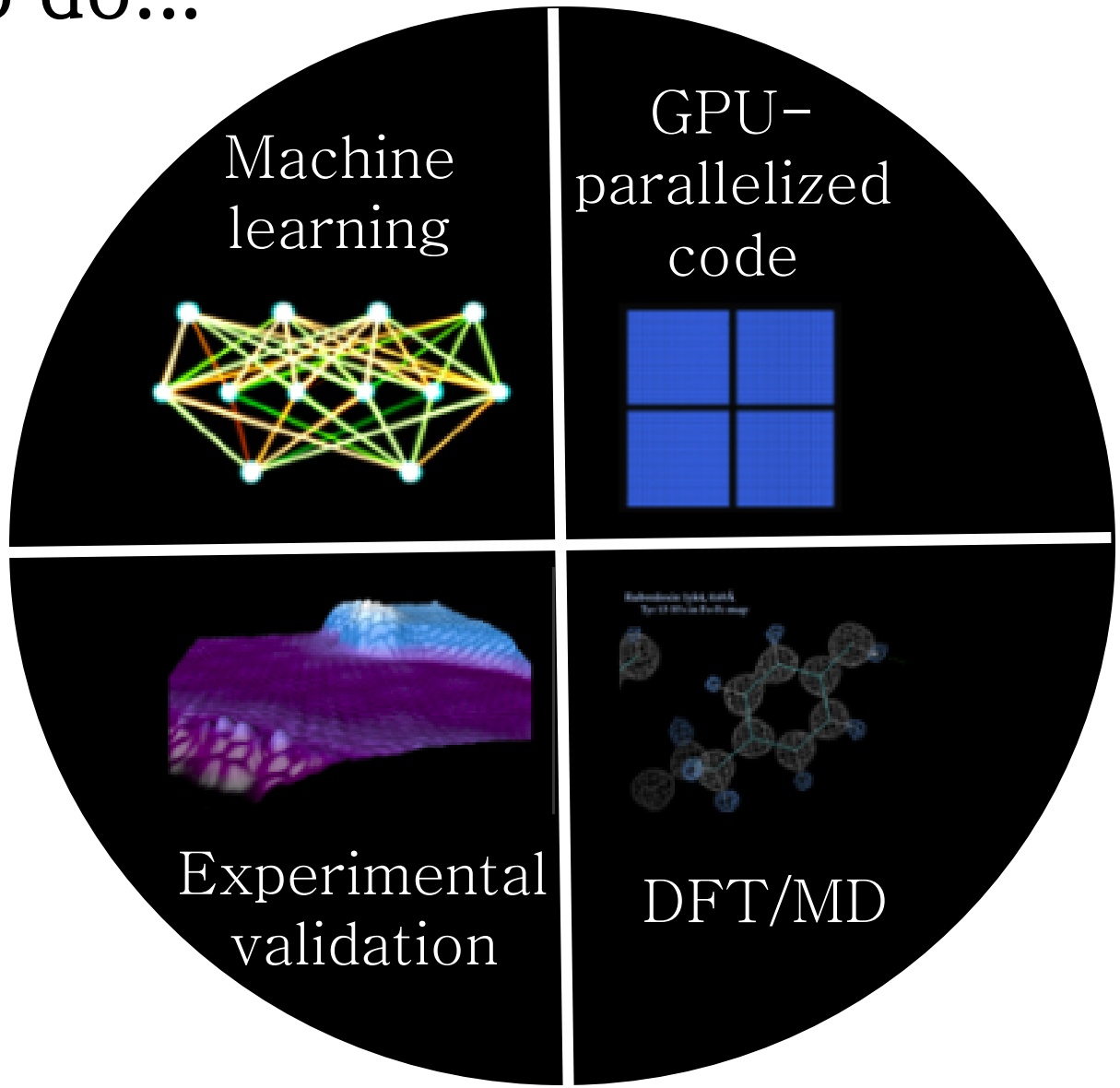
$$i\hbar\frac{\partial}{\partial t}|\Psi(\mathbf{r},t)\rangle = \hat{H}|\Psi(\mathbf{r},t)\rangle$$

Molecular dynamics

$$\vec{F} = m\vec{a}$$



What I want to do...



Kevin Andrew
University of Vermont

The BREE Project

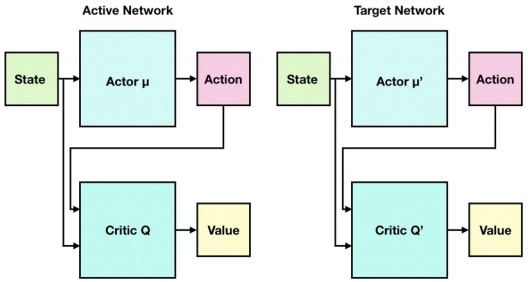
- Focused on the Lake Champlain Water Basin (NY, VT, and QC)
- Climate & Environmental Modeling to Advise Policy-making

(VT EPSCoR, *Funding Provided by NSF OIA 1556770*)

My Work

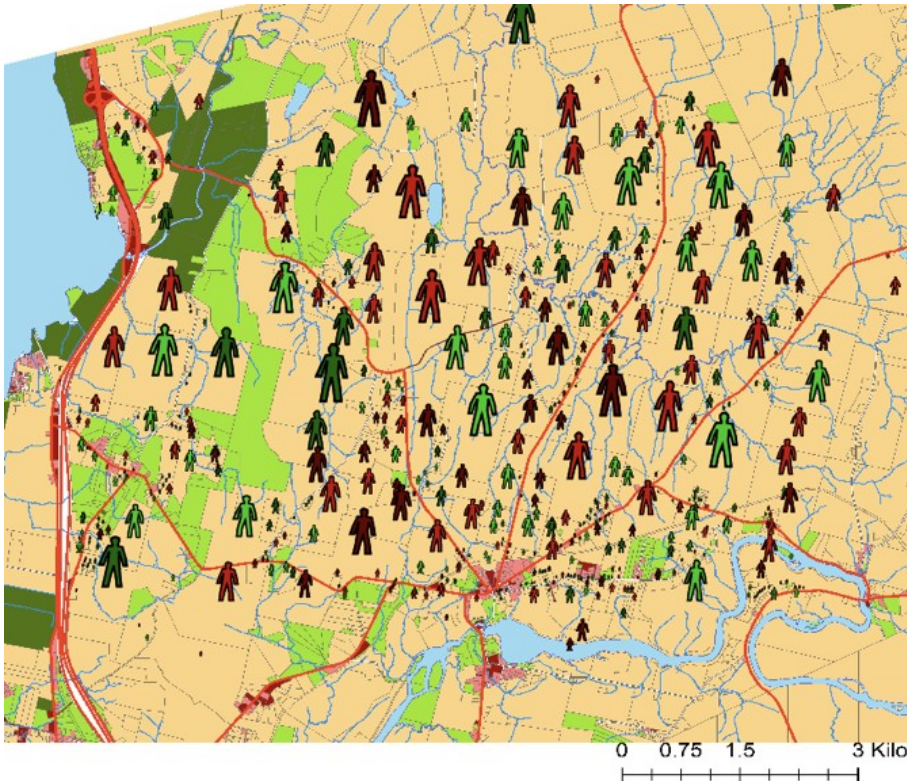
- "Embedding Learning"
- Training Human, Governmental, and NGO Agents on Historic Data, so We Can Model Potential Future Behavior

- I'm Primarily Using **Deep Reinforcement Learning** with **DDQN** to Train Agents



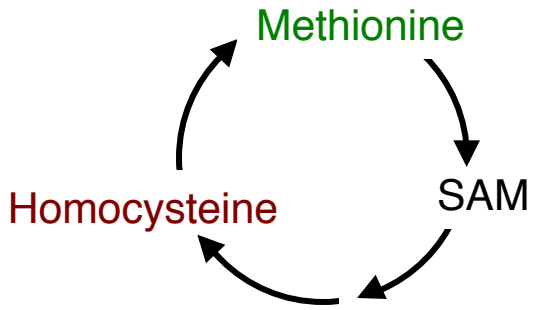
Kevin Andrew
University of Vermont

- Lots and agents and ANNs make my models very memory intensive.
- Runtimes can be very long, and doing multiple runs for validation and variance analysis only makes this problem worse.
- Right now, only some parts are parallelized with CPUs, but parallelizing more components and taking advantage of GPU programming is a goal.

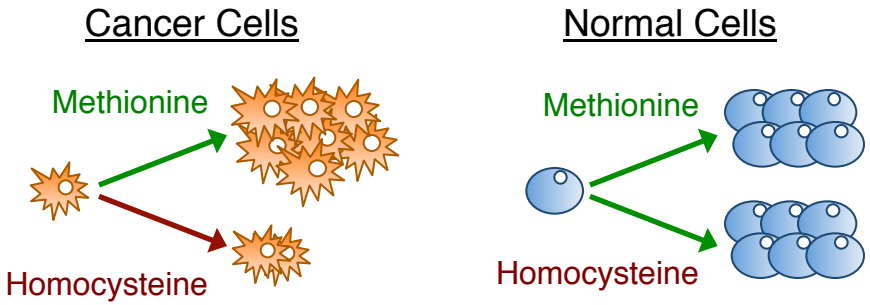


Stacey Borrego - University of California, Irvine

Metabolism of Nutrients



Differences between CANCER and NORMAL cells



Next Generation Sequencing

RNA-seq
- RNA
- Gene expression, splicing

ChIP-seq
- DNA
- Genes associated with histone modifications (methylation, acetylation)

ATAC-seq
- DNA
- Accessibility of specific genes for transcription
- Complementary with RNAseq and ChIPseq data

Methyl-seq
- DNA
- Methylation on DNA

Stacey Borrego - University of California, Irvine

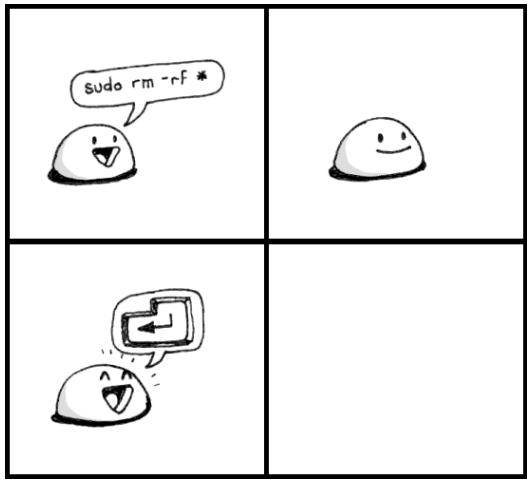
Bioinformatics Support Group

- Help introduce the UCI HPC and linux environment
- Introduce different types of NGS analyses
  - RNA-seq
  - ChIP-seq
  - Metagenomics
- Provide useful tools
  - Regular expressions
  - Task arrays
  - BioLinux with UCI RCIC

Why am I Here?

To learn about computing systems to be more:

- Efficient
  - parallelization
- Effective
  - learn more tools to improve analysis
- Aware of how to troubleshoot linux and cluster issues



Direct Numeric Simulations: Turbulent Flows

Challenges in Turbulence

Small scale structures require high resolution

Nonlinear term is relevant

Reynolds = Inertial / Viscous Forces = Captures Nonlinearity

$$\frac{\partial \vec{u}}{\partial t} + \vec{u} \cdot \nabla \vec{u} = \cdots + \frac{1}{Re} \nabla^2 \vec{u}$$

Approach

Extremely simplified physics: 2D, Incompressible, periodic BC, homogeneous

Solve in Fourier space

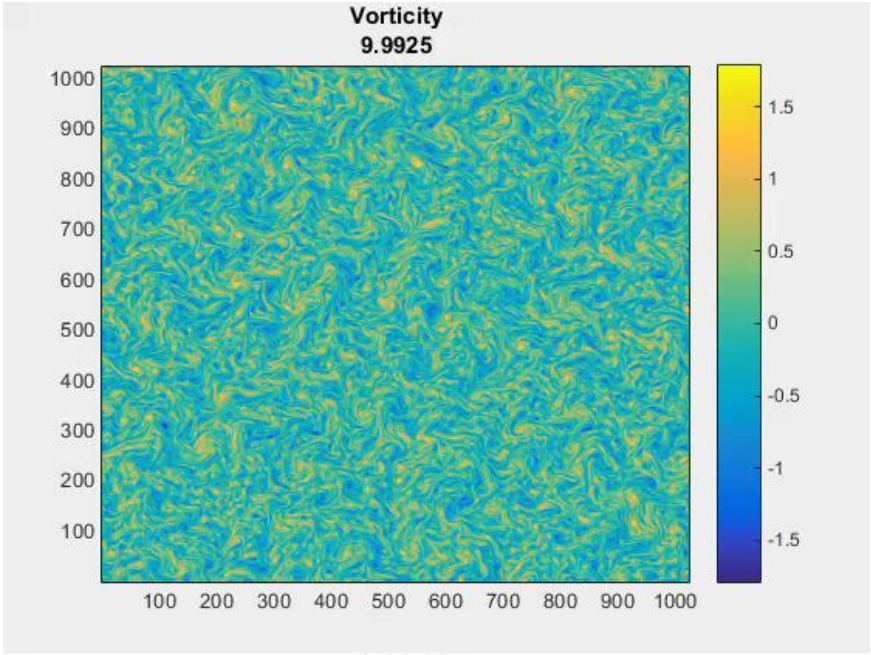
Use integration scheme to maximize Δt

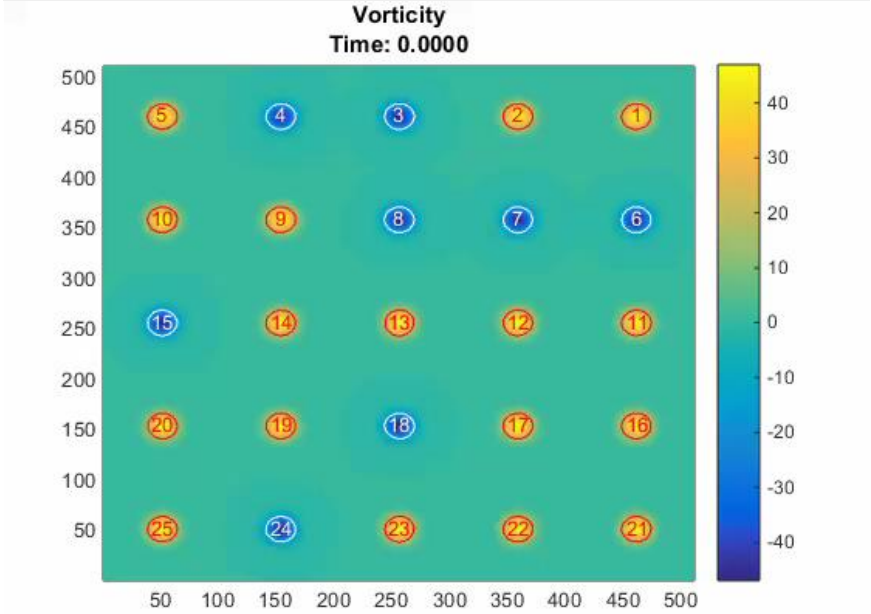
Computational Problem

Higher Re → refined mesh → smaller Δt → Increased cost

Turbulence is chaotic – many sims needed for reliable stats

Even with simplified physics simulations take a week



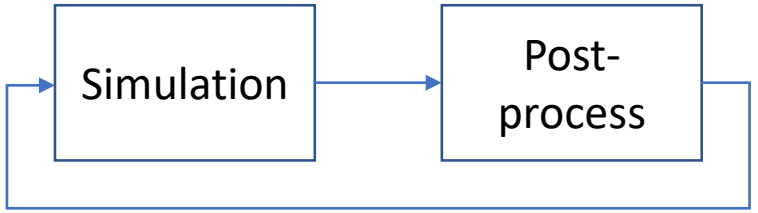


Direct Numeric Simulations: Turbulent Flows

Short term work
Code Optimization
Parallel Programming
Post-process feedback to simulation
GPU Computing
Build cluster

Long term work
Model additional physics
3D turbulence
Heat transfer
Complex geometry

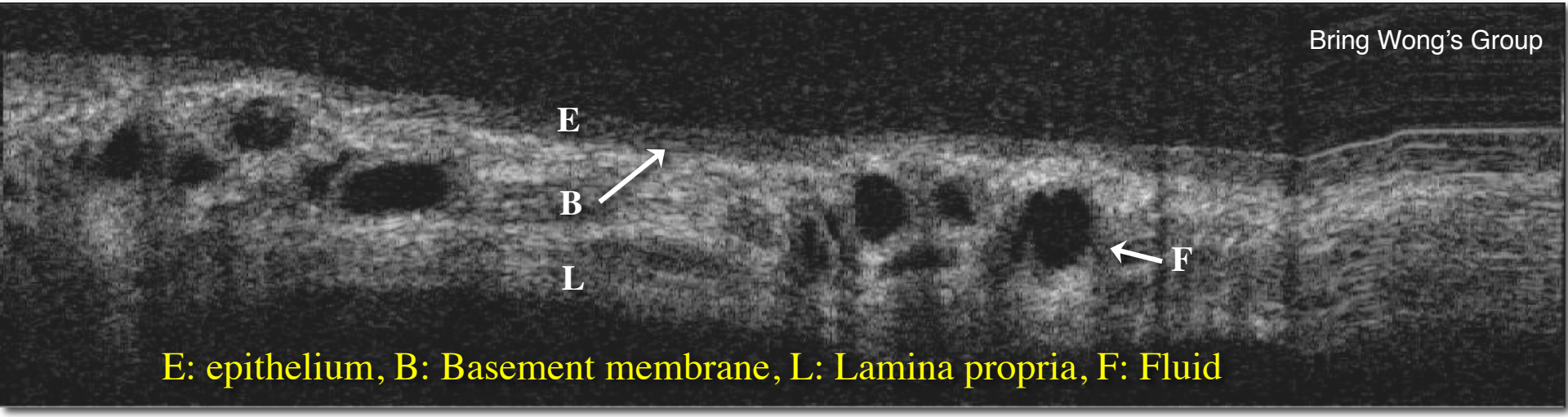
```
for k = 1:Num_Vortices
    Update_Perimeter()
    Update_Properties()
    Verify_Structure()
end
```

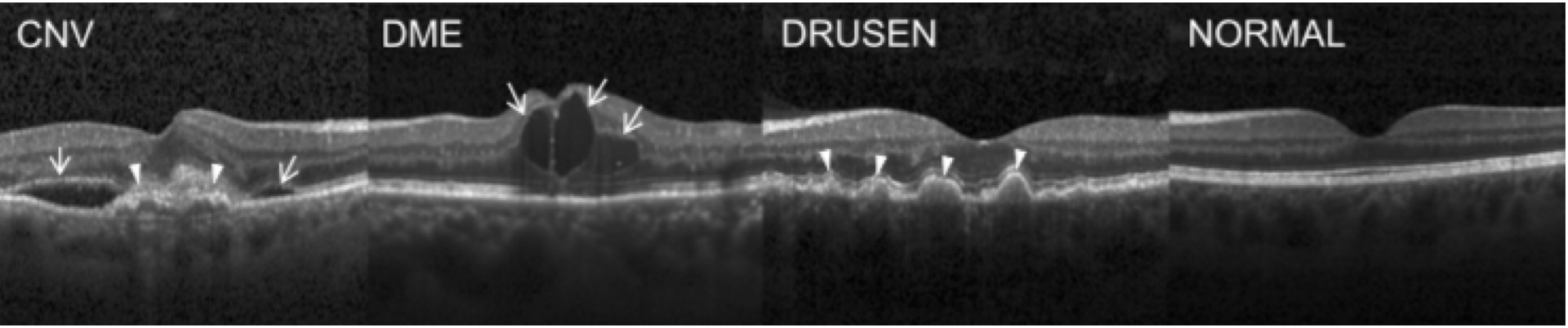


Update Δt
Update N
End loop

Jason Chen
University of California, Irvine

Optical Coherence Tomography

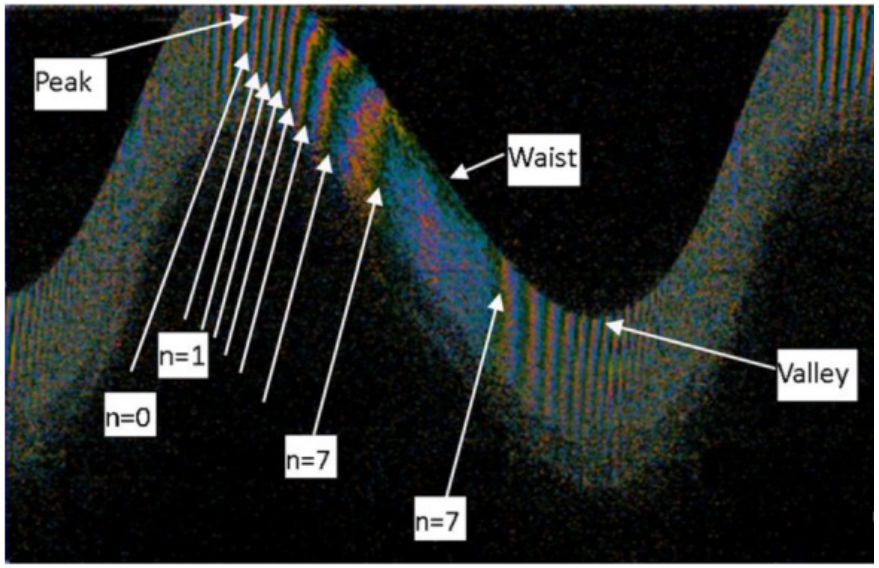




Jason Chen
University of California, Irvine

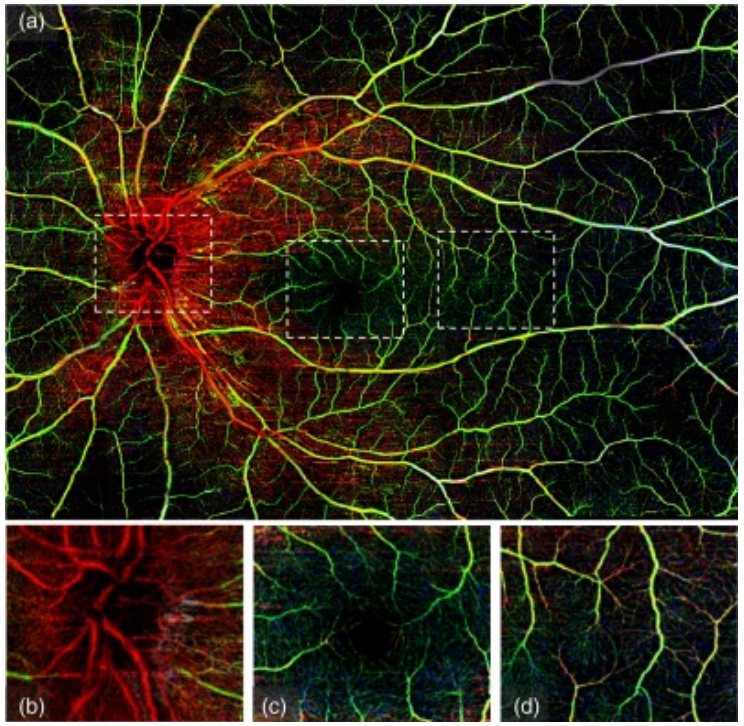
Functional Imaging

Doppler



Zhongping Chen's Group

Angiography



Ricky Wang's Group

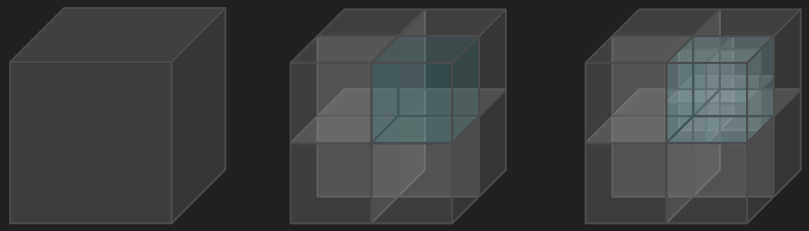
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Yi-Hao Chen

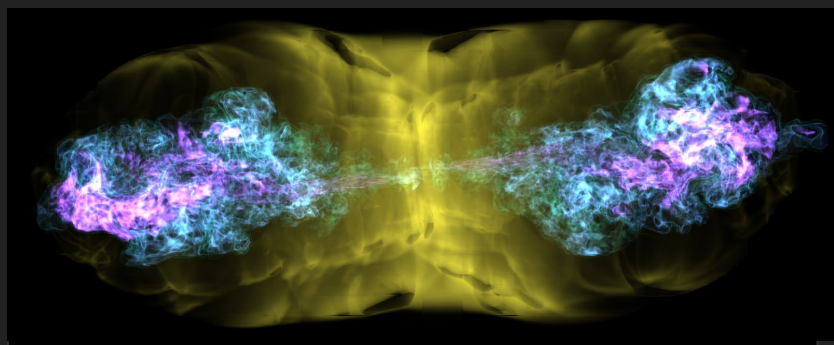
Simulating magnetized jets

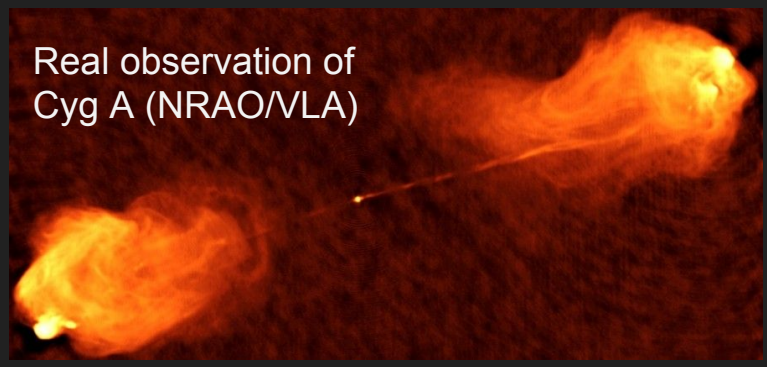
We use HPC to solve magnetohydrodynamic equations to achieve high resolution that is necessary for the delicate magnetic field structure in the jets.

We use Adaptive Mesh Refinement (AMR) to save computing time



I implement the magnetization and develop particle tracers to synthesize radio images and spectra



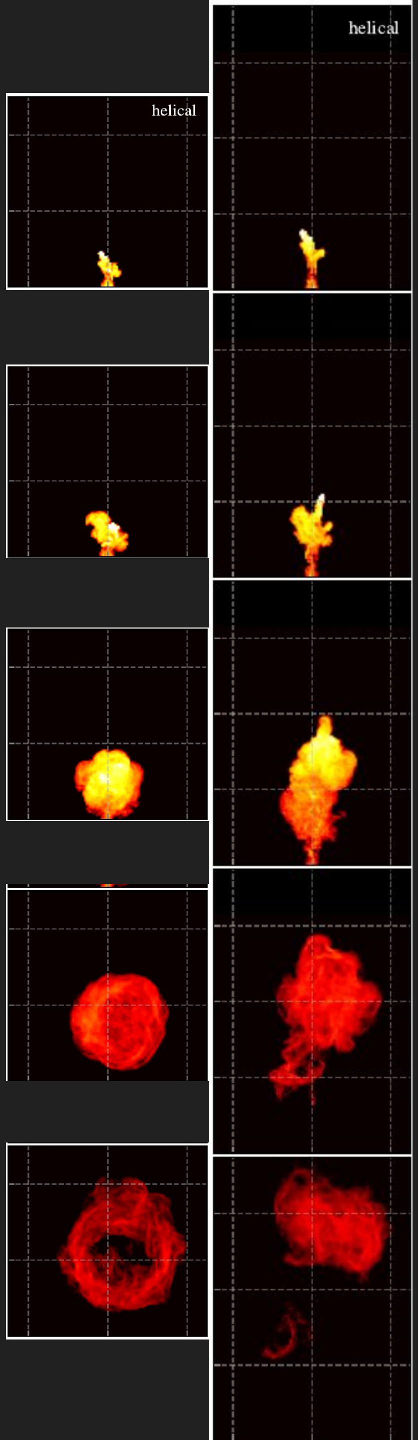


WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Yi-Hao Chen

Computational challenges and goals

- How to map the information from particles to mesh grids in a meaningful and computationally efficient way?
- How to use the simulations to determine the properties of the jets from the observations
- Better understand the parallel algorithms of the code (debugging…)
- Learn new ideas and cool tools!



Shane Coffield

UC Irvine, Dept of Earth System Science

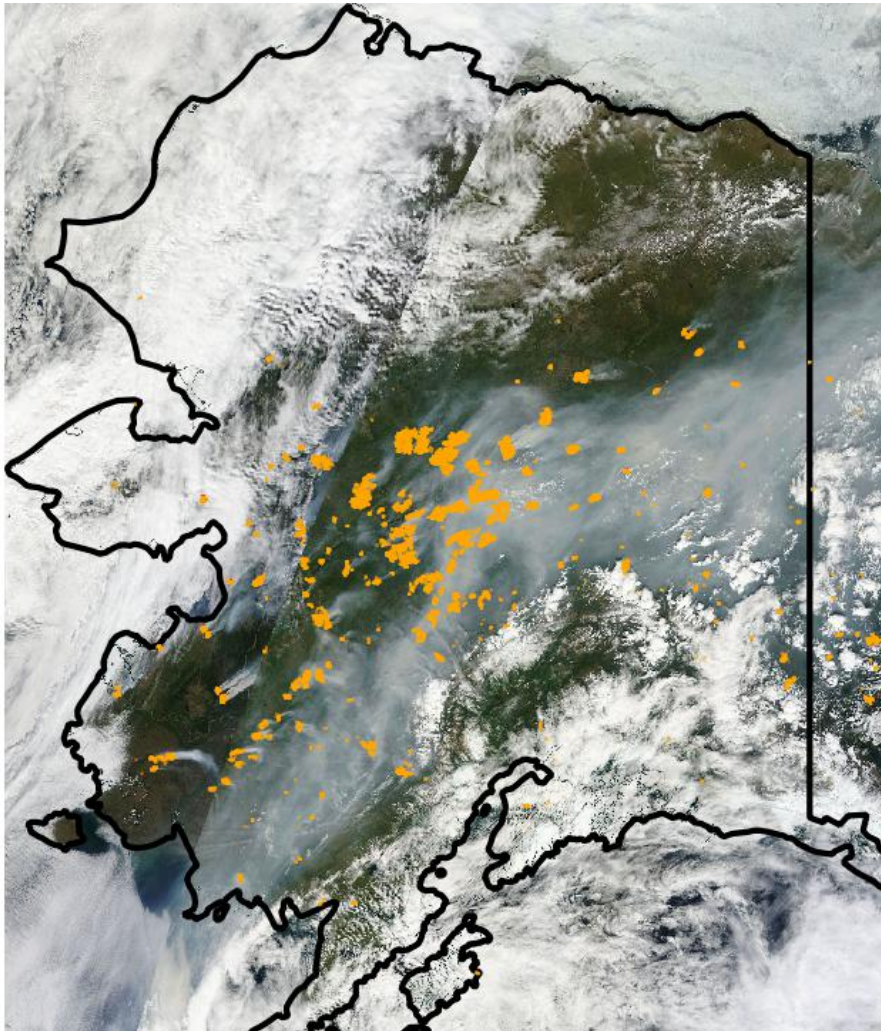
Statistical fire prediction in Alaska

Research questions

- How much variability in final fire size can be explained by information available at the time of ignition?
- How accurately can we forecast active fire counts?

Data – large-scale geospatial datasets for weather and land cover

Methods – machine learning (regression and classification)



Shane Coffield

UC Irvine, Dept of Earth System Science

Wildfire research involves:

Large-scale geospatial data streams

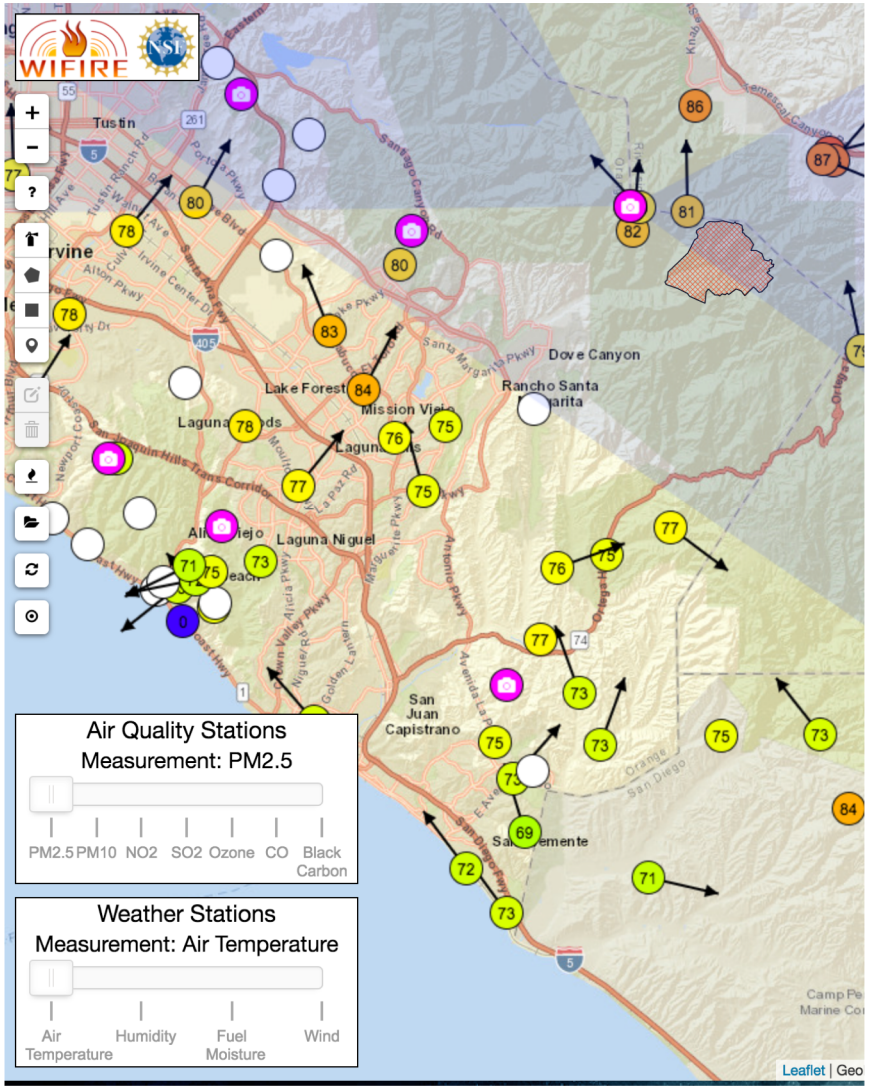
- Remote sensing (MODIS, Landsat)
- Meteorological reanalysis (GFS, ERA)

Data visualization

Machine learning and statistics

Scaling to globe

WIFIRE at SDSC →



cosine
communication
signal processing
networking

Yousef El-Laham

Ph.D. Student, Electrical Engineering

STONY BROOK UNIVERSITY

- Bayesian signal processing for complex systems
- Application: Bayesian parameter estimation of demographic rates for age-structured population models

State-Space Model

$$\mathbf{x}_t = f_t(\mathbf{x}_{t-1}, \boldsymbol{u}_t)$$

$$\mathbf{y}_t = g_t(\mathbf{x}_t, \boldsymbol{v}_t)$$

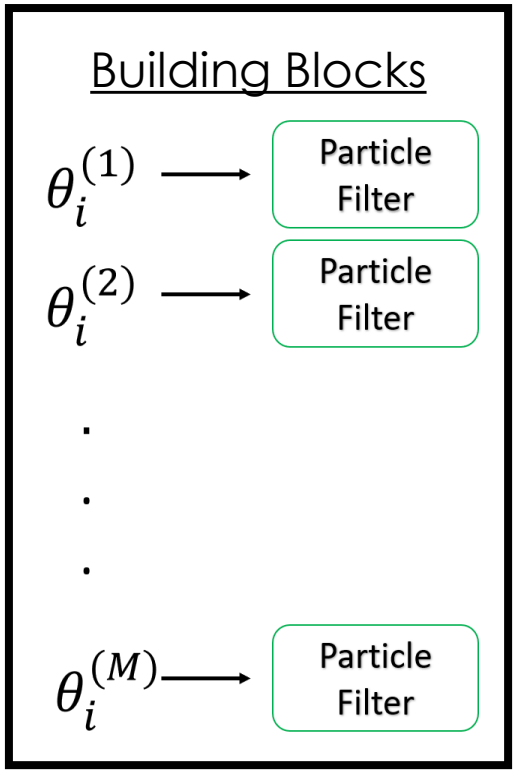
Learn the posterior distribution of latent states and model parameters

Adélie Penguin

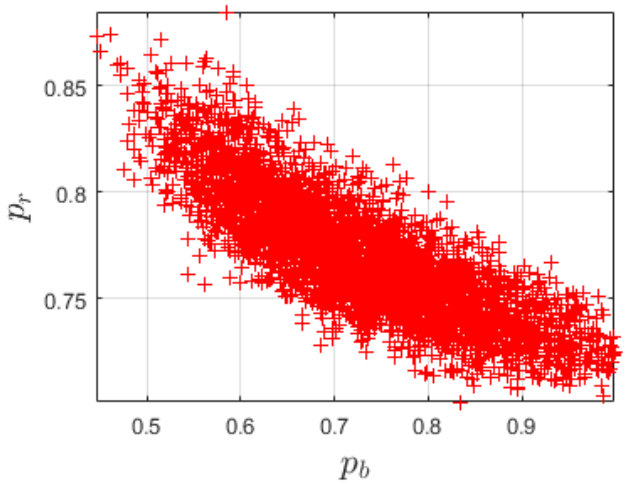
http://www.penguinmap.com/

Why HPC?

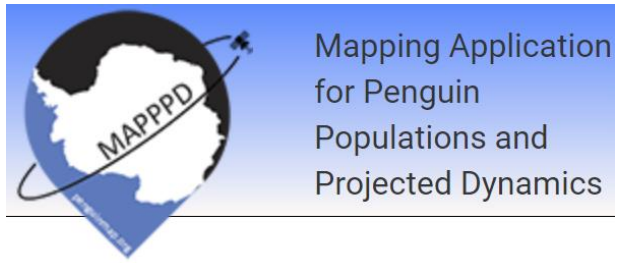
- Parallelizable algorithms
- Monte Carlo methods
- Lots of data (for a Bayesian)



"Embarrassingly" Parallel



Sampling methods!



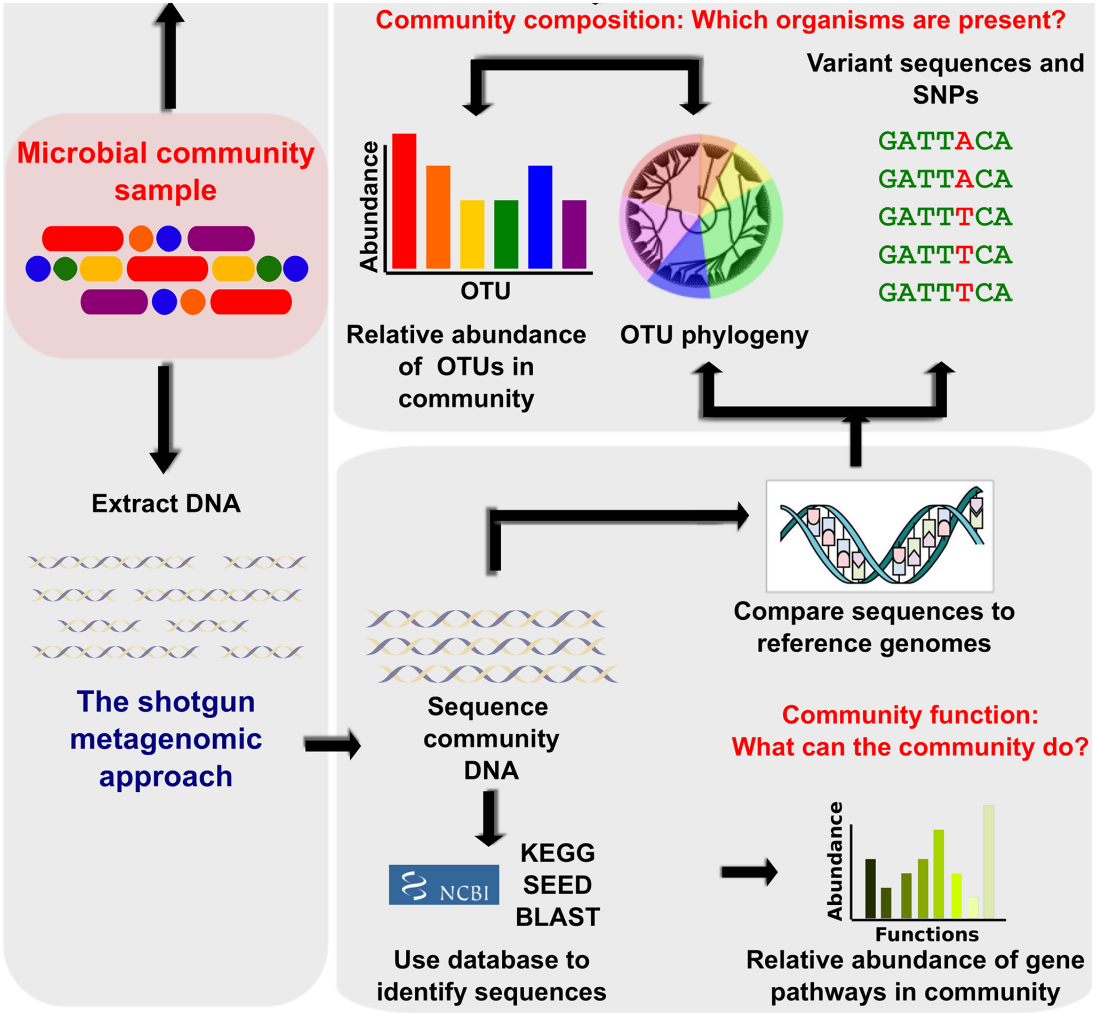
Goals for SDSC institute?

- Take advantage of parallelization
  - Python - Dask, Numba
- Scale methods for large datasets and complex models
- Become familiar with HPC concepts
- Meet some nice people ☺

Xin Fang

UC San Diego

Research Focus: Gut microbiome in Inflammatory Bowel Disease



- Investigate the taxonomy, function, and dynamics of gut microbiome in IBD patients
- Use bioinformatics tools to analysis sequencing data:
  - Metagenomics
  - Metatranscriptomics
  - Genomics

Xin Fang

UC San Diego

Challenge: I am a relatively new COMET user with limited experience in parallel computing…

Goals:

- Learn more about COMET
- Improve the performance of my code
- Explore other exciting techniques that could be helpful to my research

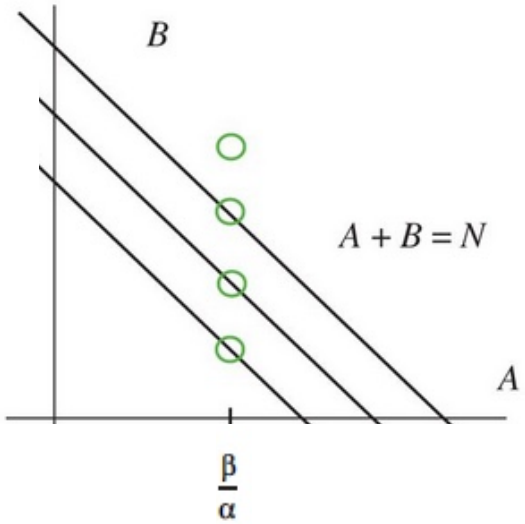
Alvaro Fletcher
University of California, Irvine

Chemical Reaction Networks, Absolute Concentration Robustness, and bifunctional enzymes.

$$A + B \xrightarrow{\alpha} 2B$$
$$B \xrightarrow{\beta} A$$

$$\dot{c}_A = -\alpha c_A c_B + \beta c_B$$
$$\dot{c}_B = \alpha c_A c_B - \beta c_B$$

$$c_A = \frac{\beta}{\alpha}$$

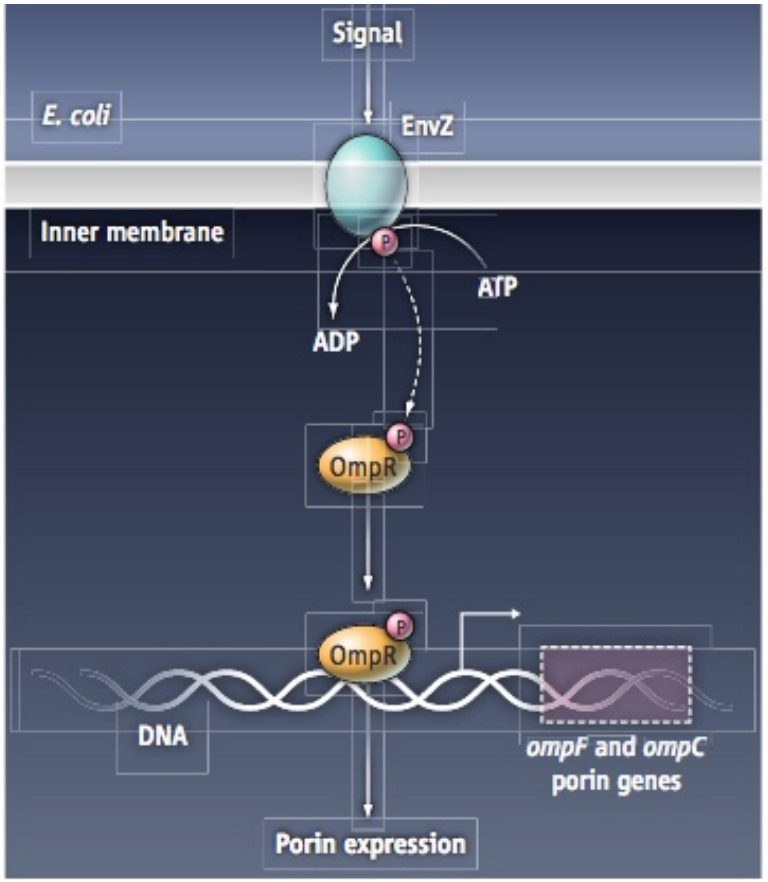


Fructose 1, 6-bisphosphate aldolase/phosphatase may be an **ancestral gluconeogenic enzyme**

RF Say, G Fuchs - Nature, 2010 - nature.com

Most archaeal groups and deeply branching bacterial lineages harbour thermophilic organisms with a chemolithoautotrophic metabolism. They live at high temperatures in volcanic habitats at the expense of inorganic substances, often under anoxic conditions 1. These autotrophic organisms use diverse carbon dioxide fixation mechanisms generating acetyl-coenzyme A, from which gluconeogenesis must start 2, 3, 4. Here we show that virtually all archaeal groups as well as the deeply branching bacterial lineages contain a …

☆ 99 Cited by 137 Related articles All 13 versions Web of Science: 91



Alvaro Fletcher
University of California, Irvine

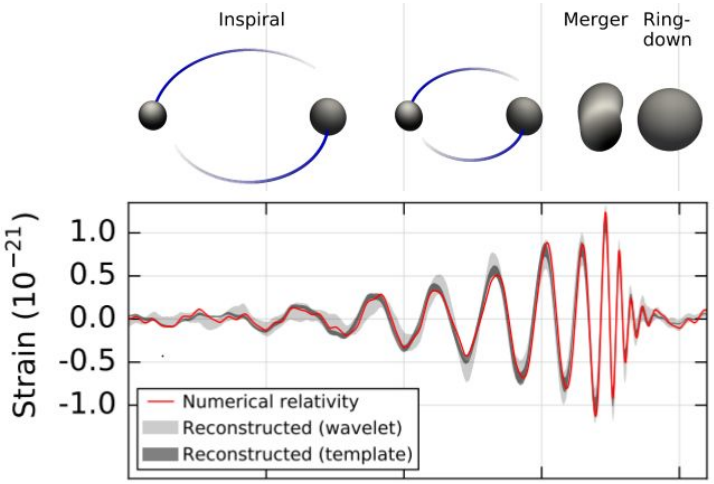
Challenges:

- For most networks of interest, analytical solutions to their mass-action equations are intractable or non-existent.
- Conditions for the existence of absolute concentration robustness (ACR) remain mostly unknown.
- However, we have some heuristics of where to look for networks exhibiting ACR.
- Parallelizing the generation and stability analysis of chemical reaction networks would alleviate the combinatorial explosion.
- Clustering these networks based on their respective topologies could provide us with intuition for discovering necessary and sufficient conditions for ACR.

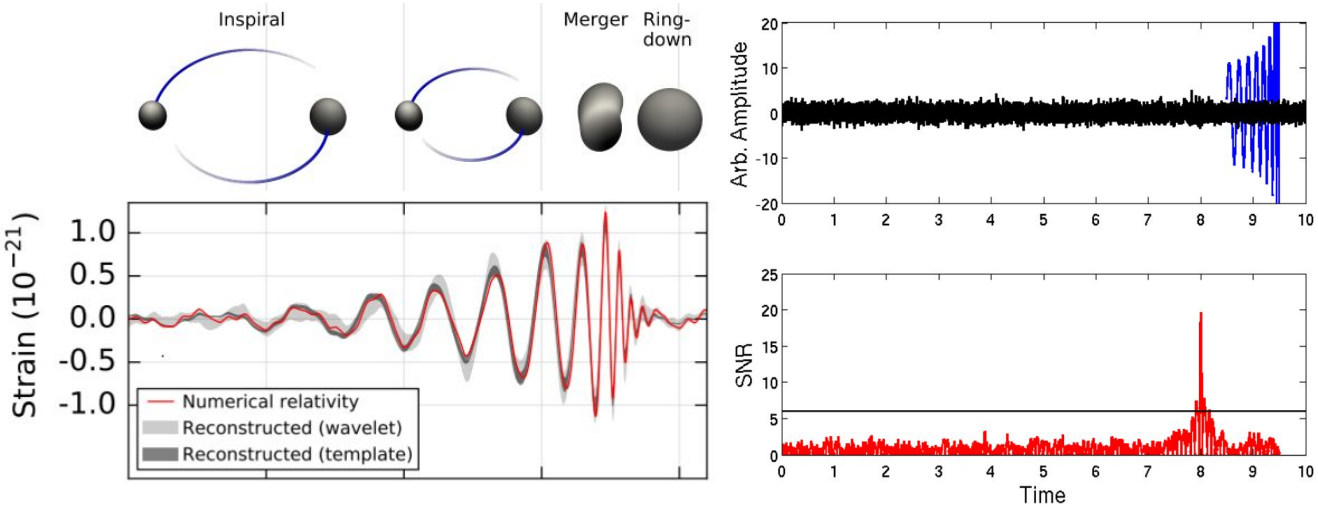
n	The number of Sudoku grids of order n (boxes are size $\sqrt{n} \times \sqrt{n}$)
1	1
4	288 [4][5]
9	6,670,903,752,021,072,936,960 [4][6]
16	5.96×10^{98} **(estimated)** [7]
25	4.36×10^{308} **(estimated)** [8]

References:
1. German A. Enciso, J. R. Soc. Interface 2016 13 20160475; DOI: 10.1098/rsif.2016.0475. Published 31 August 2016
2. Jeremy Gunawardena, *Science* 30 Apr 2010: Vol. 328, Issue 5978, pp. 581-582
3. R. F. Say, G. Fuchs,Nature, *464, 1077 (2010)*.
4. Guy Shinar, Martin Feinberg, Science. 2010 Mar 12;327(5971):1389-91.
5. https://en.wikipedia.org/wiki/Combinatorial_explosion

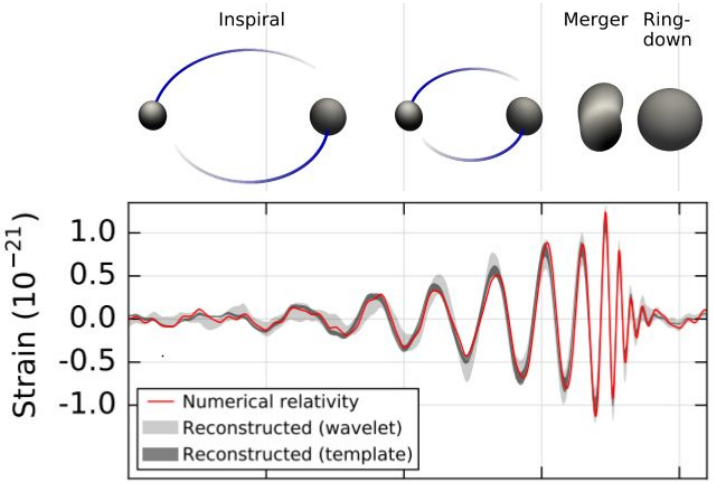
Patrick Godwin
Pennsylvania State University

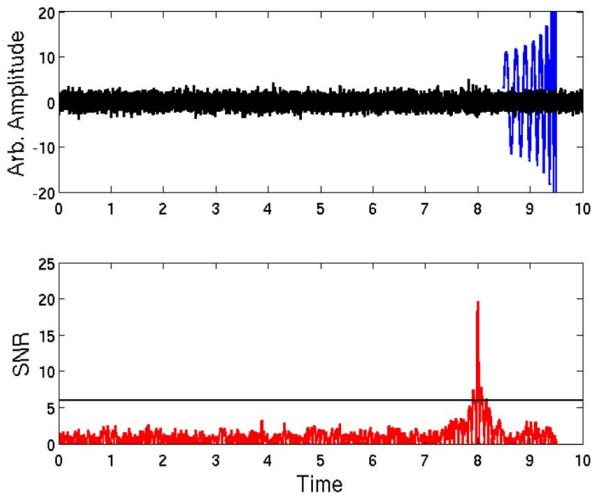


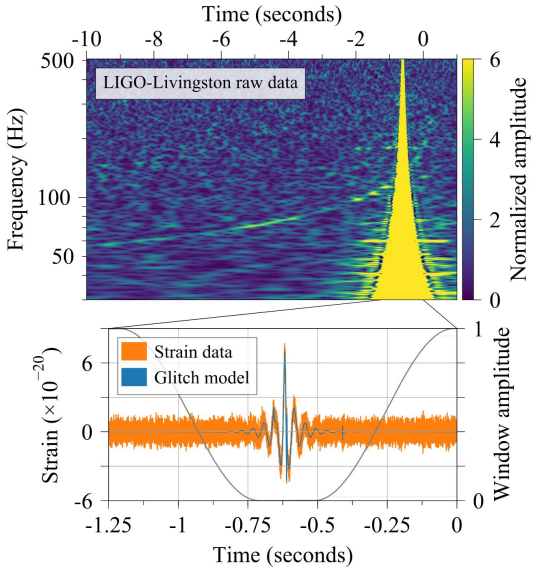
Patrick Godwin
Pennsylvania State University



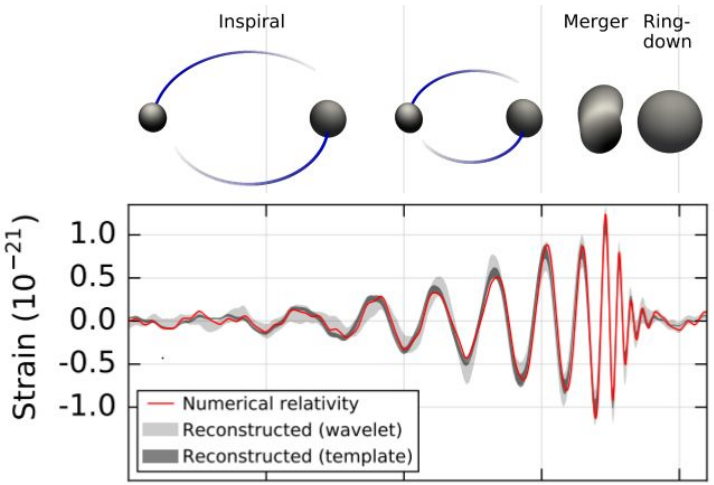
Patrick Godwin
Pennsylvania State University

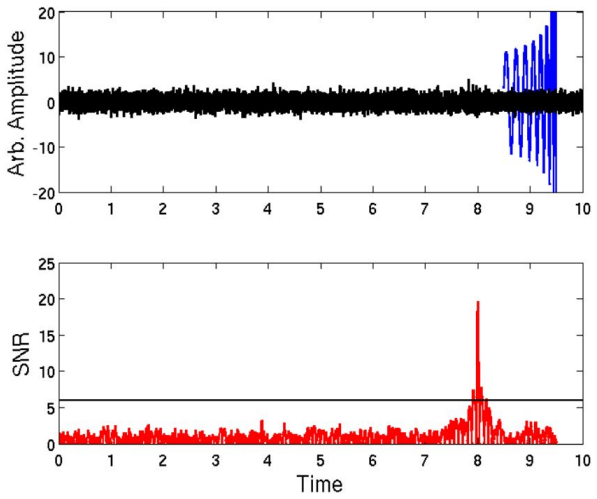


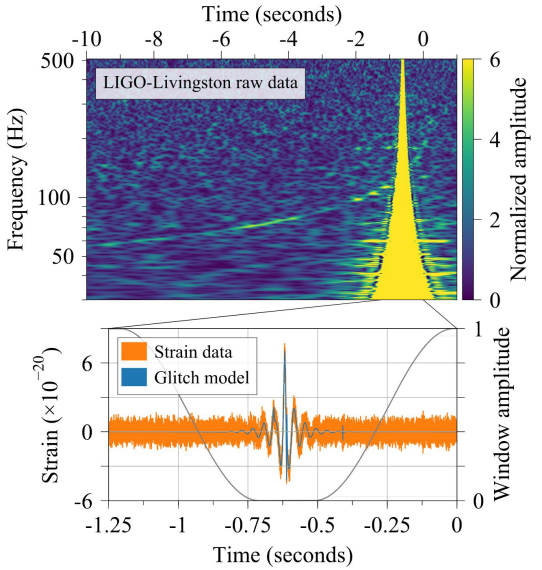


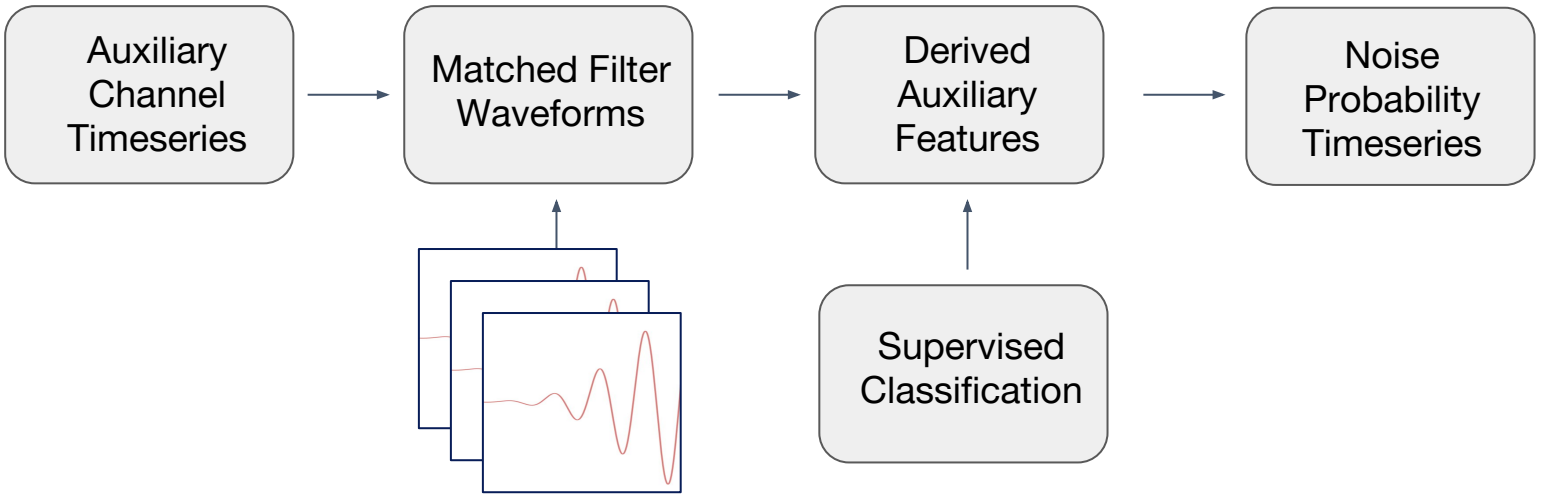


Patrick Godwin
Pennsylvania State University









Patrick Godwin
Pennsylvania State University

Challenges:

- Computational:
  - I/O constraints
  - CPU constraints
  - Latency constraints

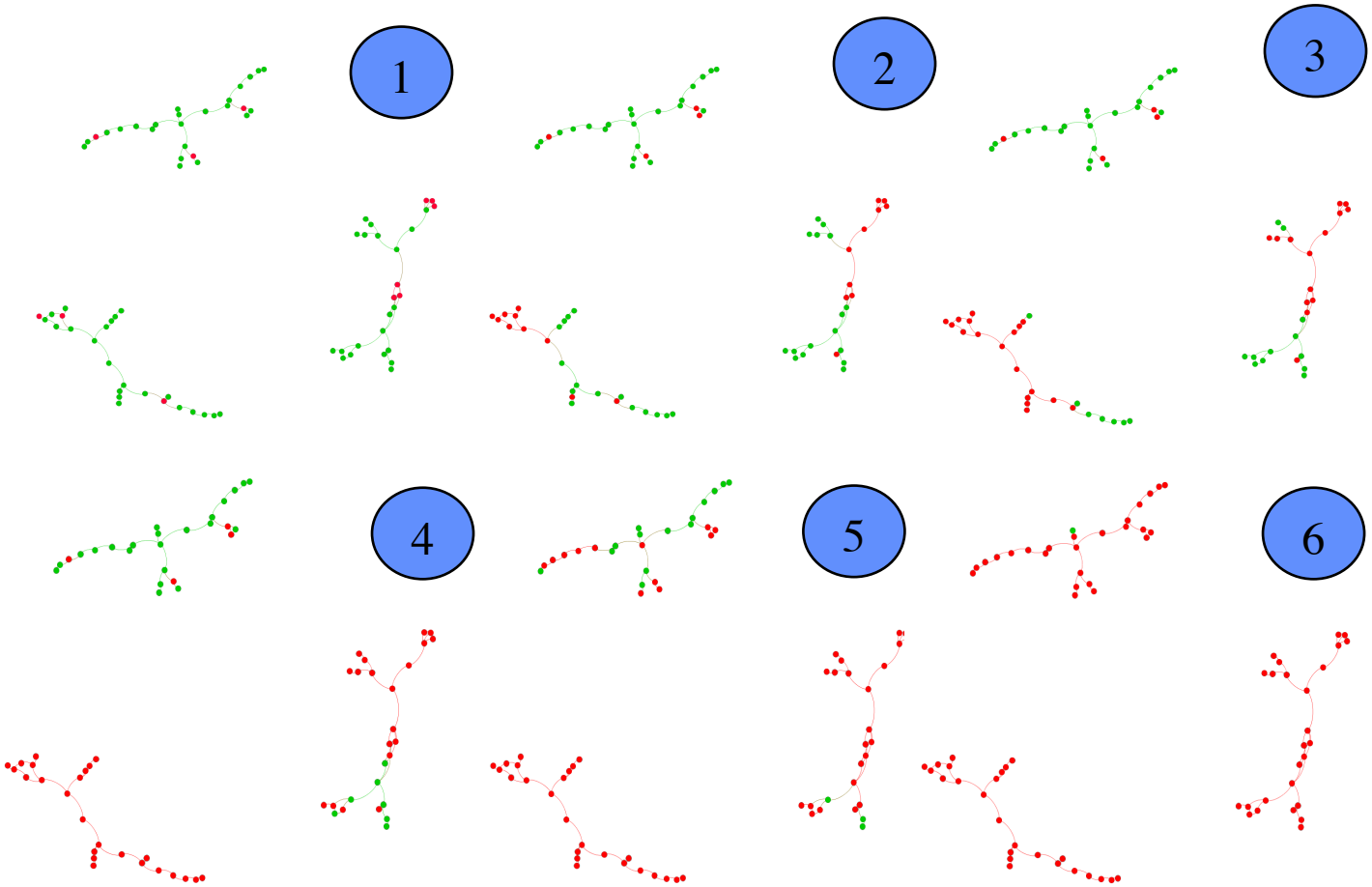
Patrick Godwin
Pennsylvania State University

Challenges:

- Computational:
  - I/O constraints
  - CPU constraints
  - Latency constraints
- Machine learning:
  - Hyperparameter Tuning
  - Feature Engineering
  - Event Labeling
  - Dealing with Drifts in Stationarity

Sameera Horawalavithana
University of South Florida

Goal: Modeling Information Diffusion in Online Social Networks

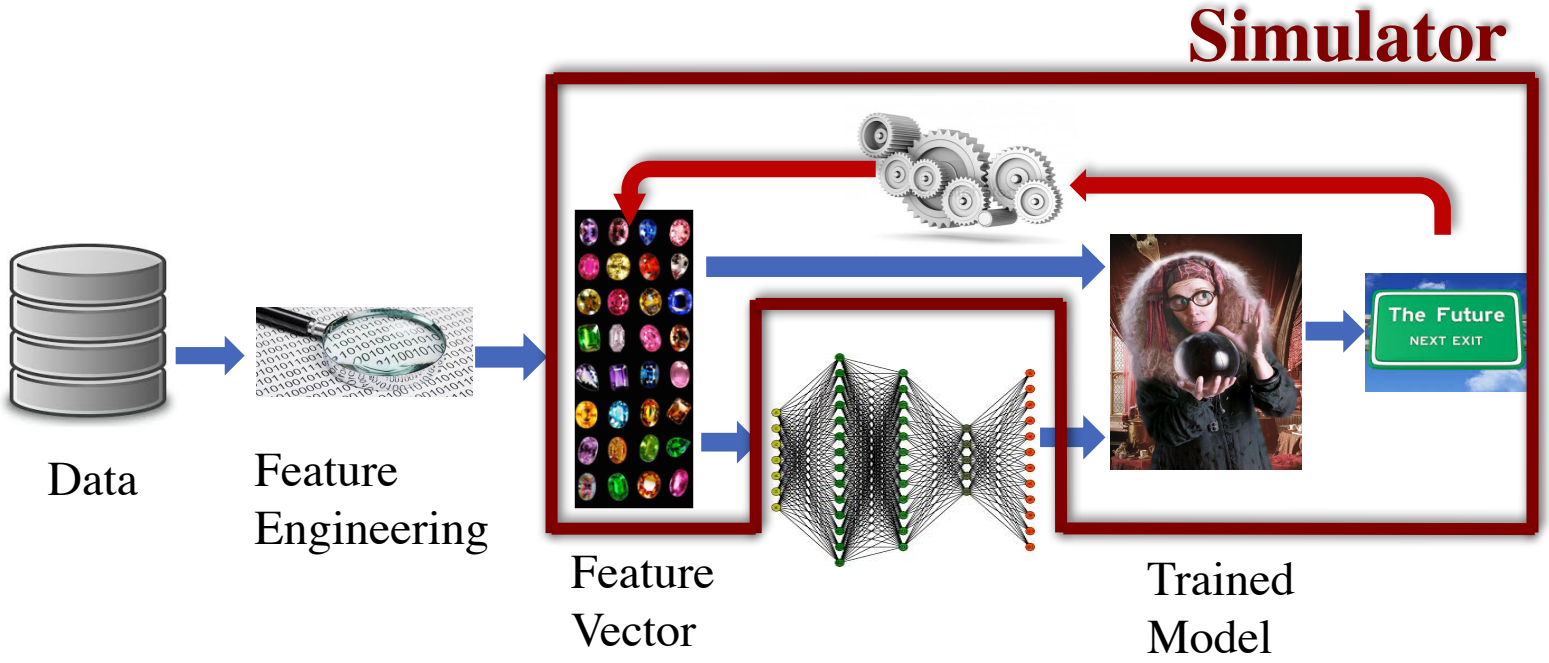


Dataset:
Steam Online Gaming Community

Cheaters

Non-cheaters

Sameera Horawalavithana
University of South Florida



- ***Task:*** *Design a fast and scalable simulator engine*
  - **Context:** Heterogeneous Dynamic Graphs
  - **Architecture:** Multi-CPU/Multi-GPU Architecture
  - **Tools:** Python with tensorflow, dask & multiprocessing Etc.
  - **Database:** Elasticsearch

Ping Hou, Ph.D. Candidate

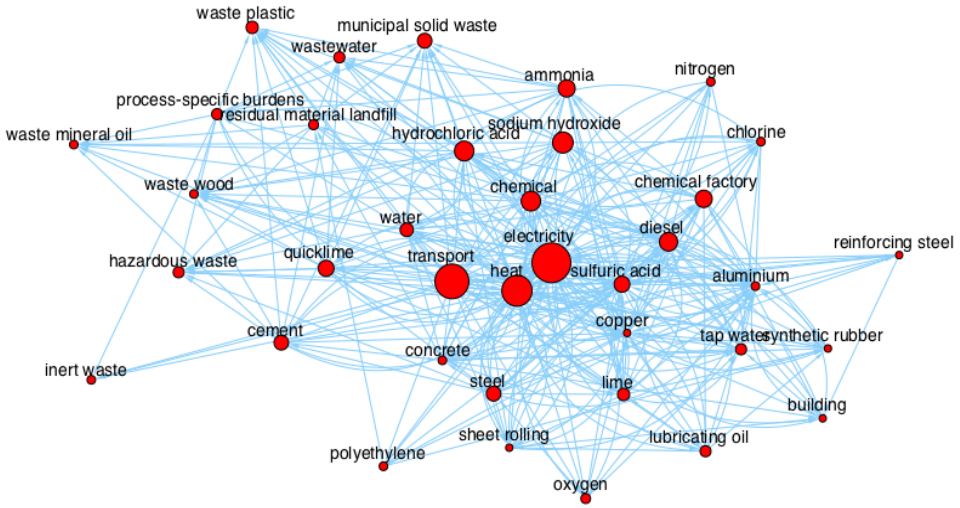
School for Environment and Sustainability, University of Michigan

Research topic: Computational approach for estimating missing data in life cycle assessment (LCA)

Data:

1) resource consumption and environmental emissions (e.g., 1.07 kg CO_2 emissions per kWh electricity produced)
2) impact assessment data (e.g., potentially disappeared fraction of species/kg chemical emitted)

1. link prediction for estimating missing consumption and emission data in the LCA database of the industrial system.

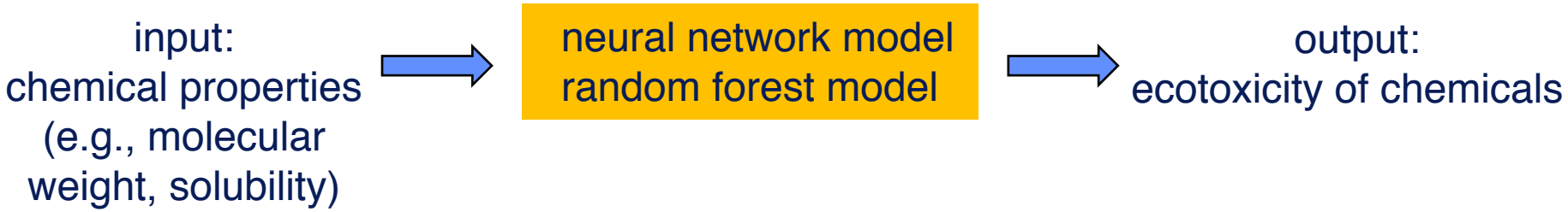


Code: Matlab and R, parallel code run on Flux (HPC cluster at the UM)

2. machine learning for estimating ecotoxicity of chemicals.
   - neural network model
   - random forest model
   - Code: Matlab and Python

Ping Hou, Ph.D. Candidate

School for Environment and Sustainability, University of Michigan



Current data source: USEtox, 3,077 chemicals with 11 chemical properties and their ecotoxicity.

Challenges:

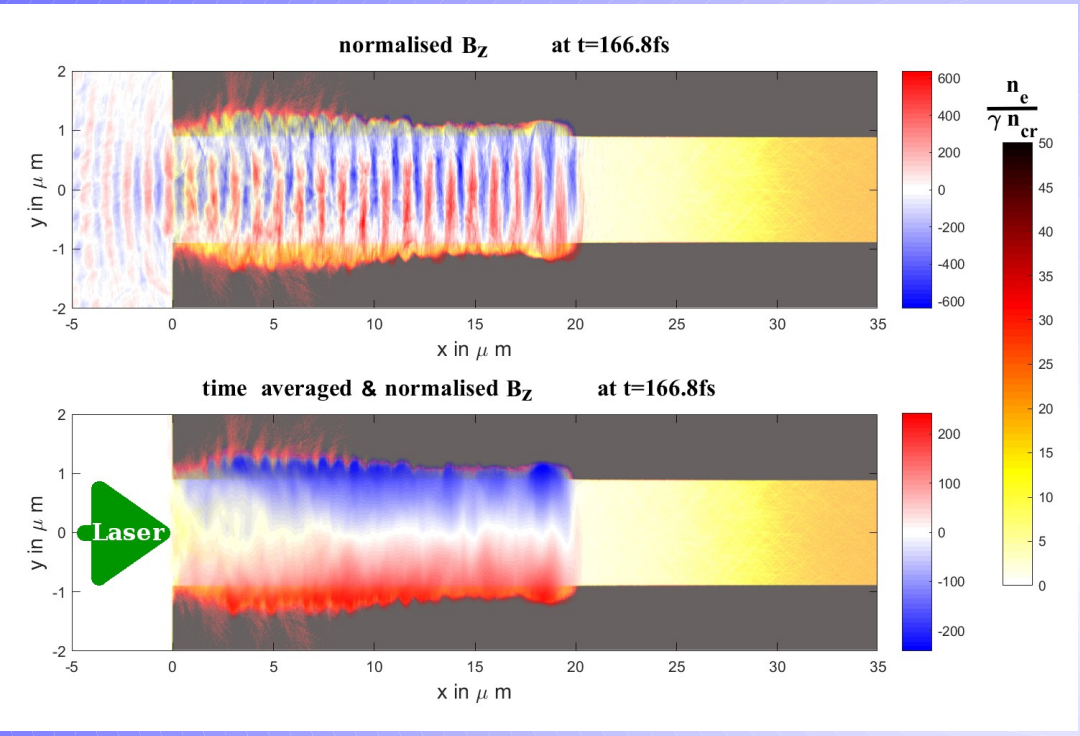
1) how to parallel the code when include more chemical properties (e.g., Dragon has more than 5,000 chemical properties)
2) how to parallel the code when include more chemicals (e.g., 85,000 chemicals listed under US EPA Toxic Substances Control Act (TSCA)

Hope to accomplish:

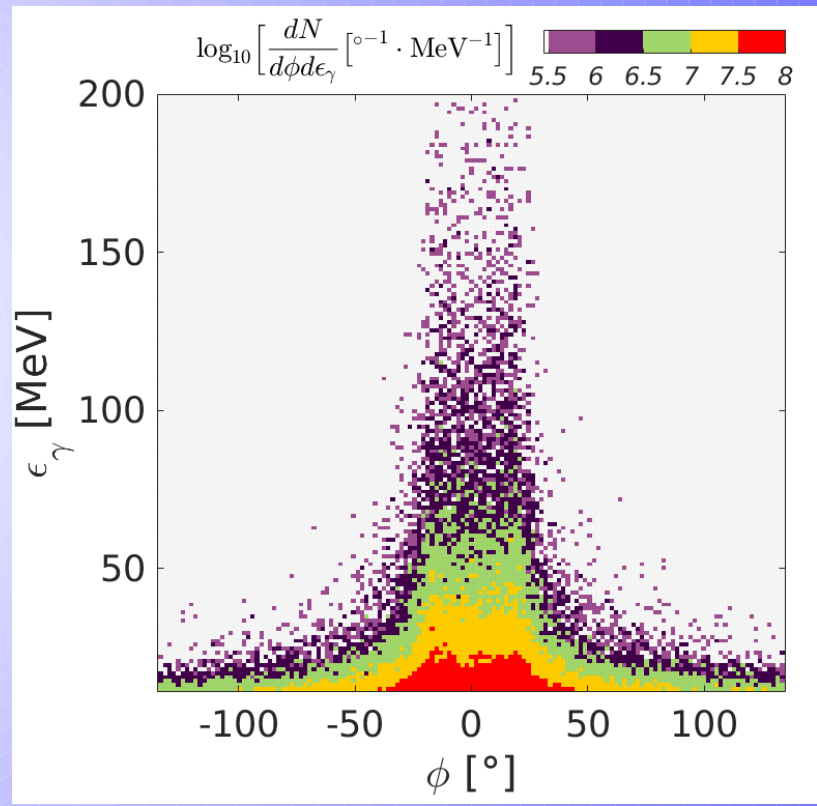
1) learn the HPC skills to parallel the code and improve the code performance;
2) learn data and code management skills.

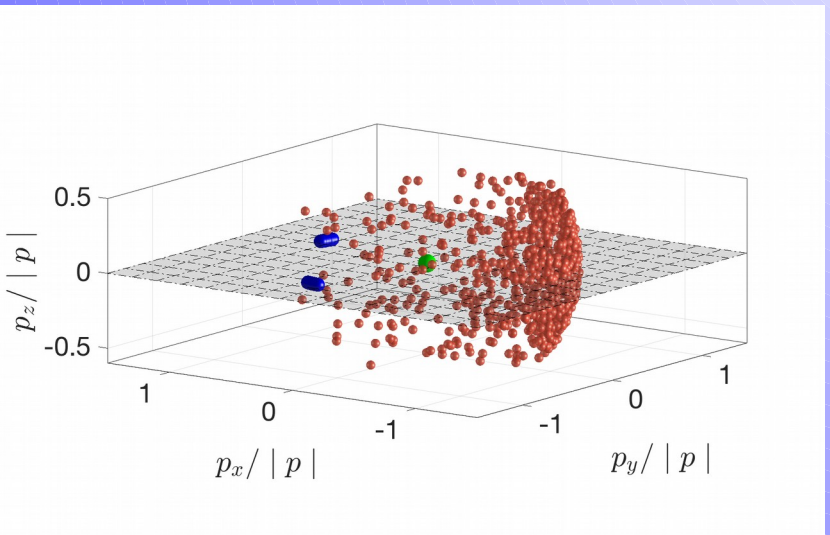
Oliver Jansen
UCSD - MAE

A laser pulse inside a plasma channel leads to the formation of a strong quasi-static magnetic field



Electrons accelerated in the channel emit high-energy photons in the magnetic field of the channel



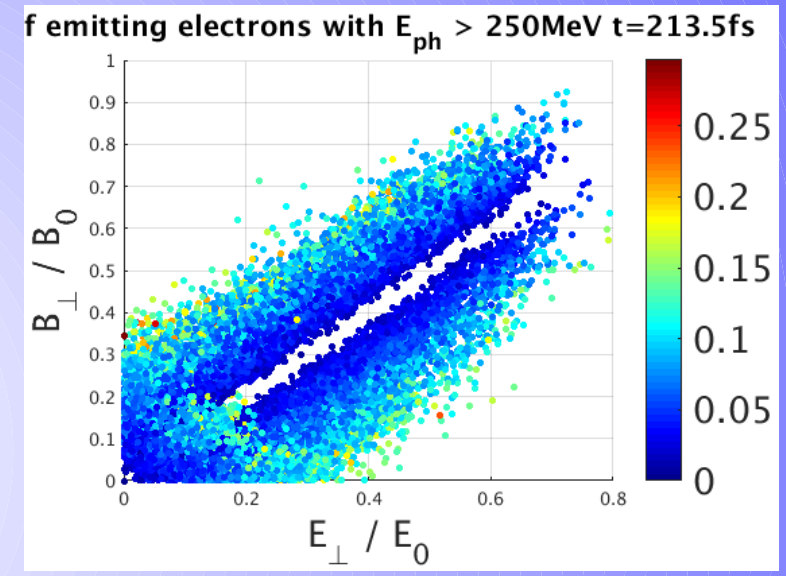


Those photons can be used for several applications. Here, they are collided in order to create electron-positron pairs
(The inverse to the annihilation of an electron and one positron).

Oliver Jansen
UCSD - MAE

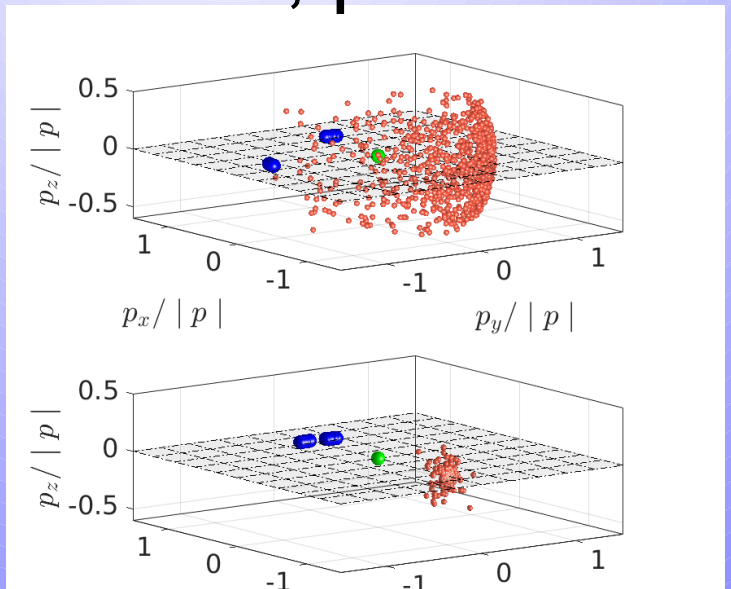
Challenges:

- Post-processing of large data



- Visualization (of large data)

- Development of efficient, parallelized simulations



He Jiang

UC San Diego Dept Mathematics

Research focus:

Resampling methods and hypothesis testing

Importance of simulations:

(1) To derive patterns that are extremely difficult to find mathematically

(2) To confirm and visually illustrate results proved mathematically

A Comparison of Permutation and Bootstrap Tests

(1)Does not need to know the underlying distribution of the sample

(2)Testing for equality of *means* (mainly in small sample sizes)

(3)Compare *Type I Error Rates* and *Power* in order to find out which test works best under different circumstances

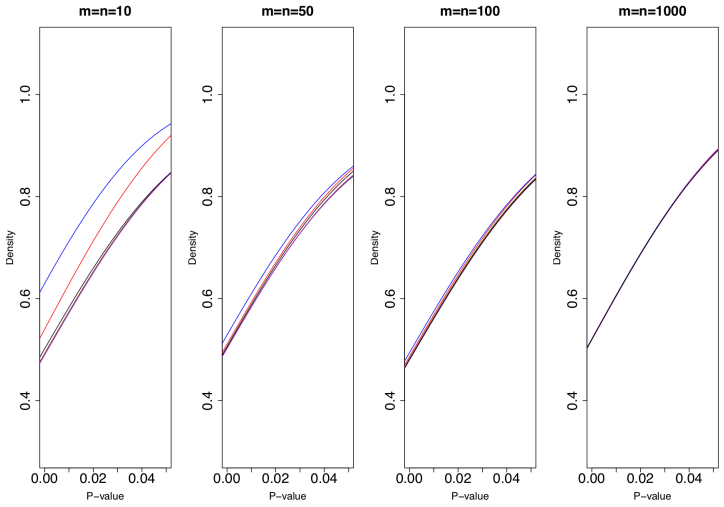
(4)Each situation needs *5,000* samples(of *X* and *Y*), then each resampling method is carried out *B=10,000* times

(5)Four situations for Normal and Exponential cases respectively in finding *Type I Error Rate*, and four more situations for Normal and Exponential cases in finding *Power*

He Jiang

UC San Diego Dept Mathematics

Problem: The simulation takes too long on a personal computer, and cannot be accomplished in a timely manner



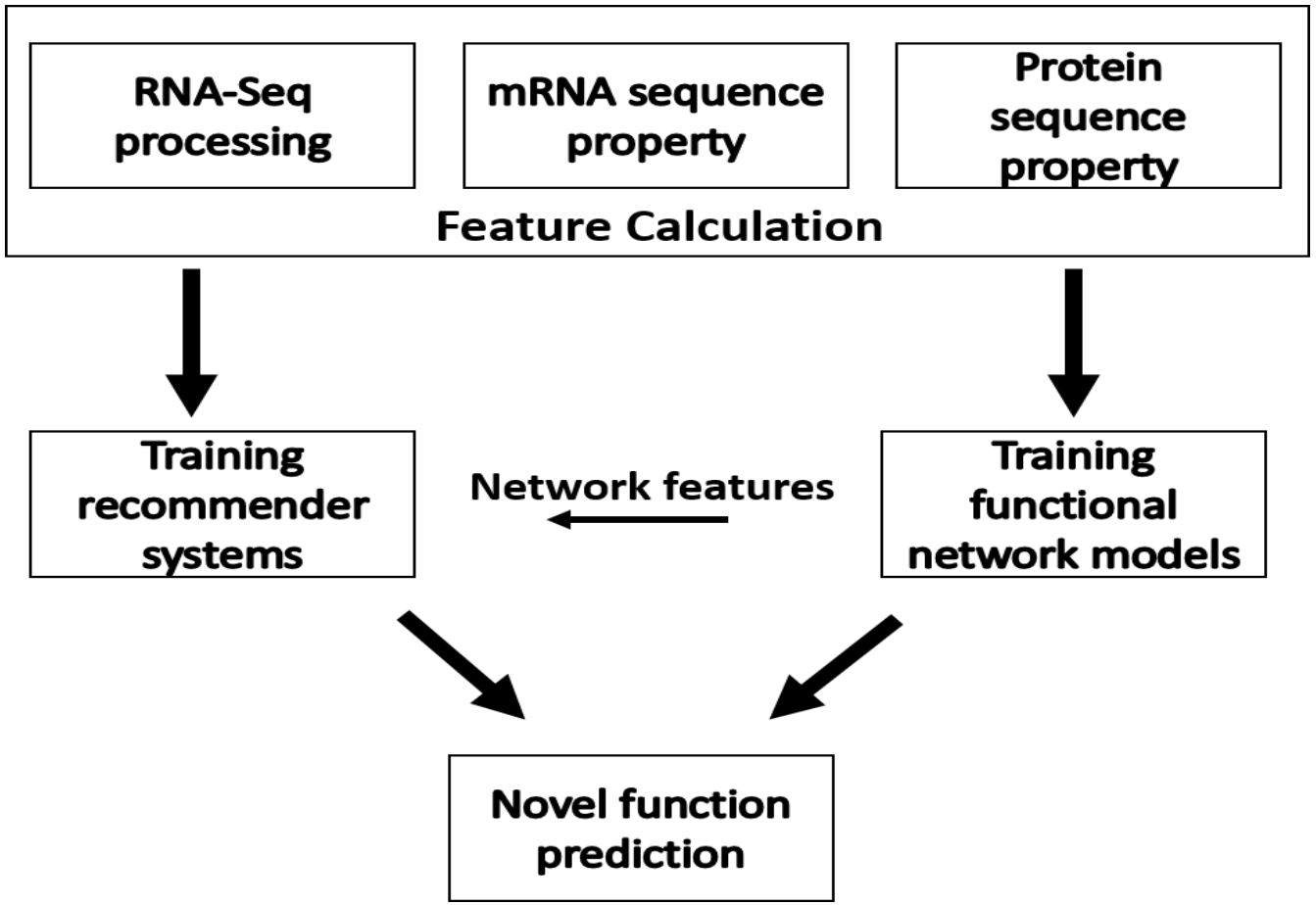
For example, this “simple” graph on the left illustrating *Type I Error Rates*, takes four sets of *12* hours to draw, which is around *48* hours total

Note that this is only around *1/4* of the total simulations in the entire project

Goal: Speeding up simulation process would greatly benefit further statistical researches in their computational part

IOWA STATE UNIVERSITY
Bioinformatics and Computational Biology
Graduate Program
Gaurav Kandoi

Function prediction methods for mRNA isoforms



Challenges

- Scalability and optimization: Several steps limited to single node or even single core!
- Large datasets → Large memory and disk space requirements

Expectations

- Code performance and optimization
- GPU computing
- Scalability and Parallel computing

KU

Dr. Boryana Koseva
University of Kansas

K-INBRE

1. Access to KU HPC
2. Education
   - *Individualized training*
   - *Workshops*
3. Bioinformatics support
   - *Consult on experimental design*
   - *Carry out data analysis*
     - *Differential gene expression*
     - *Genome assembly*
     - *Variant calling and population genomics*

Our services are free to Kansas researchers

Dr. Boryana Koseva
University of Kansas

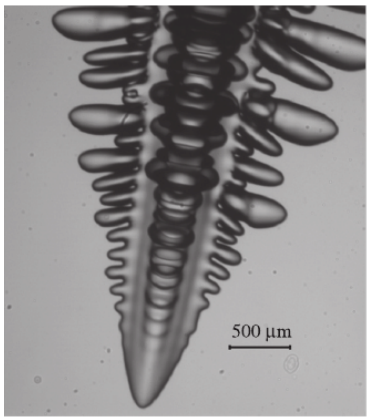
Challenges	SDSC Summer Institute
1. Software Installation	• Singularity containers
2. Efficient use of HPC resources	• Python Parallelization • Job Arrays
3. Generating good reports	• Scientific visualization
4. Reproducible Research	• Talk to others

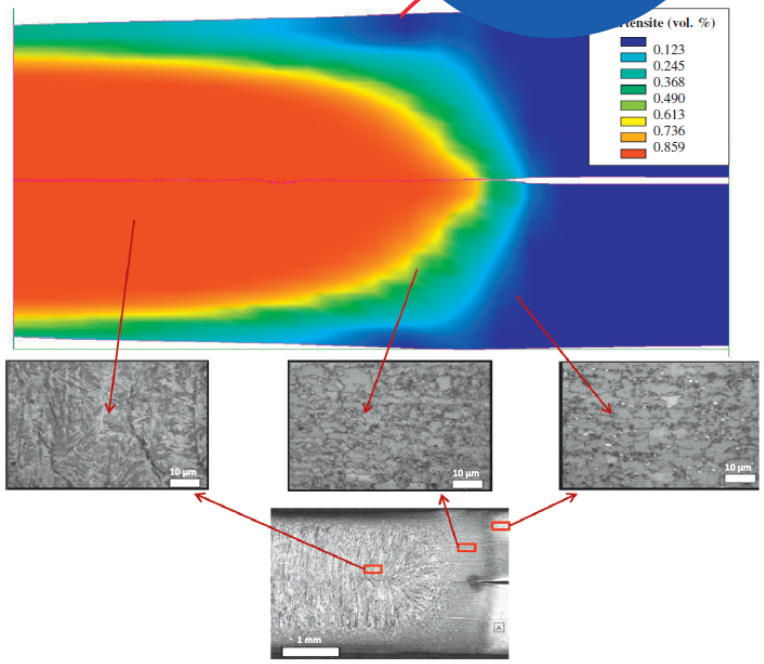
Ryan Lenart

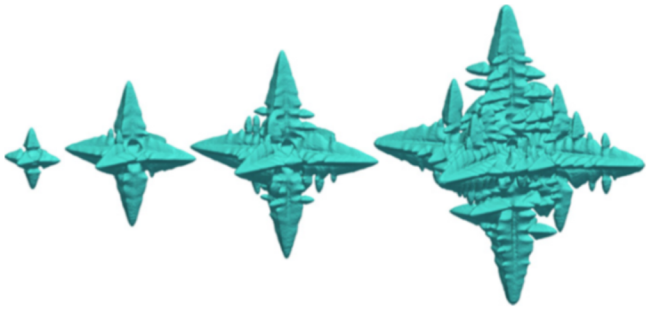
California State University, Los Angeles

- ***Modeling Dendritic Solidification Using Lattice Boltzmann and Phase Field Methods***
  - *Prediction of Solidification Microstructure for Powder Bed Fusion Additive Manufacturing*
  - *Modeling Dendritic Solidification in Microgravity and Terrestrial Conditions*

- Phase Field Method, Lattice Boltzmann Method, Discrete Element Method
- Some parallel code, most is still serial.

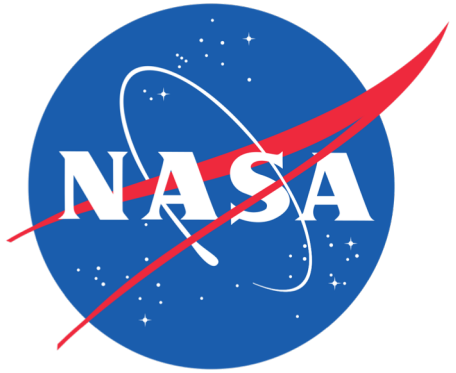






Ryan Lenart

California State University, Los Angeles



- My code works. Now I need to make it faster.
- Modeling a significant domain takes too much time.

- I want to
  - Optimize my code as much as possible
  - Take full advantage of our computing resources
    - New graphics cards
    - Many cores

Haiqing Li
City of Hope

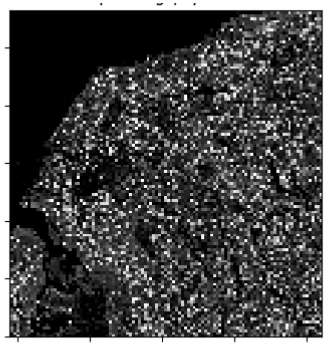
- Bioinformatics Data Analysis
- Bio-Image Analysis
- Nature Language process

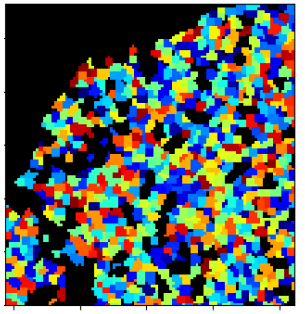
the MIRACLE of SCIENCE with SOUL™

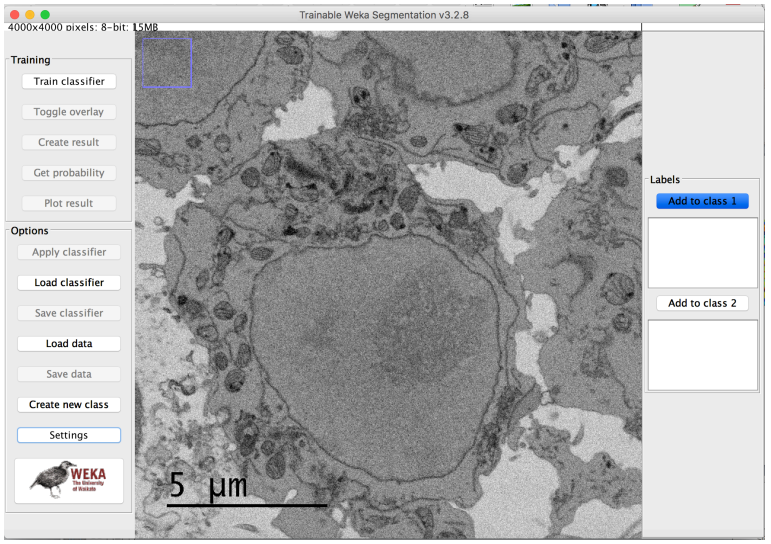
City of Hope®

computational challenges

- Image Analysis
  - Segmentation using machine learning method
  - Whole slide images registration
  - Atlas-free registration







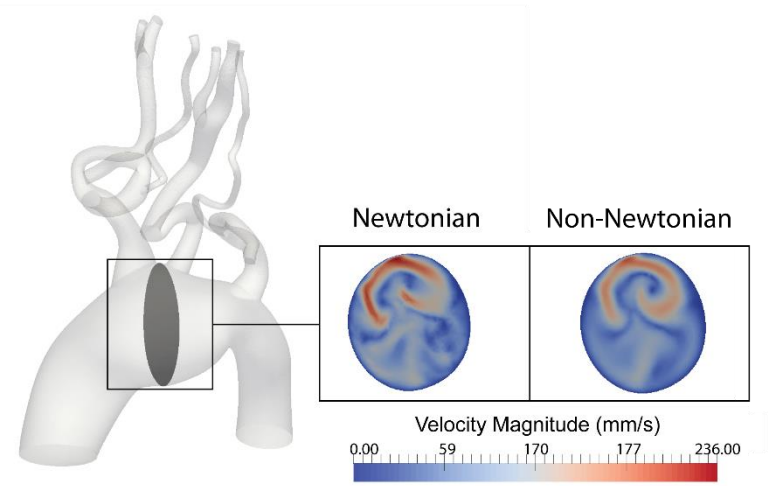
Sabrina Lynch
University of Michigan, Ann Arbor

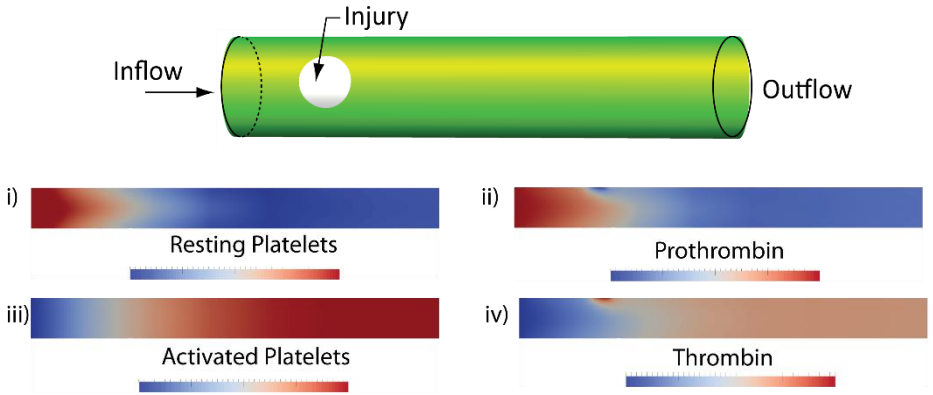
Parallel Image-Based FEM flowsolver

CRIMSON

$$\rho\left(\frac{\partial v_i}{\partial t} + \boldsymbol{v}\cdot\nabla\boldsymbol{v}\right) = -\nabla P + \mu\nabla^2\boldsymbol{v} + \rho\boldsymbol{f}, \nabla\boldsymbol{v} = 0$$

$$\frac{\partial c_i}{\partial t} + \nabla\cdot(-D_i\nabla c_i) = R_i - \boldsymbol{v}\cdot\nabla c_i$$





Sabrina Lynch

University of Michigan, Ann Arbor

Challenges

- Coupling Python interface to Fortran subroutines
- Computation efficiency
- Data Visualization

$$\rho\left(\frac{\partial v_i}{\partial t} + \boldsymbol{v}\cdot\nabla\boldsymbol{v}\right) = -\nabla P + \mu\nabla^2\boldsymbol{v} + \rho\boldsymbol{f}, \nabla\boldsymbol{v} = 0$$

$$\frac{\partial c_i}{\partial t} + \nabla\cdot(-D_i\nabla c_i) = R_i - \boldsymbol{v}\cdot\nabla c_i$$

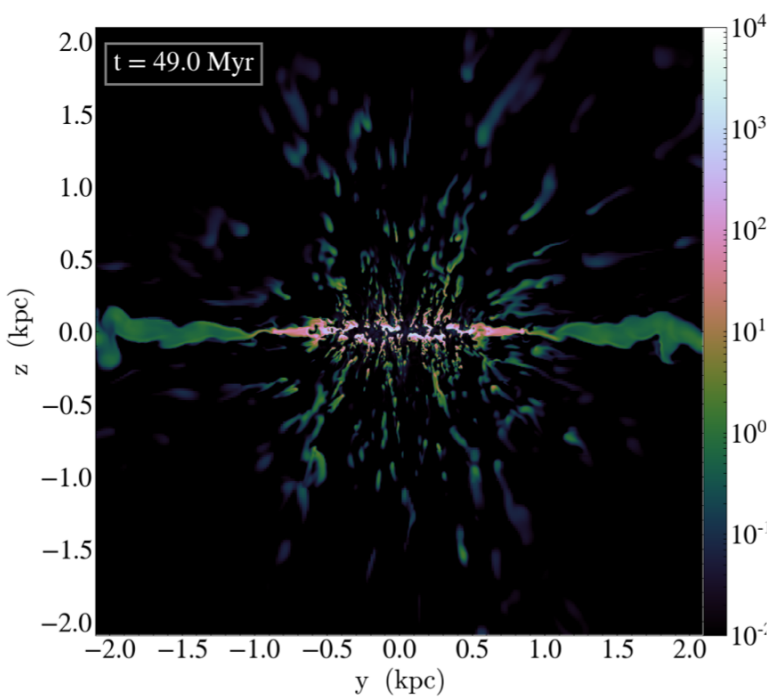
$$R_i = k_i c_i + k_{mi} c_m c_i + k_{ii} c_{i2}$$

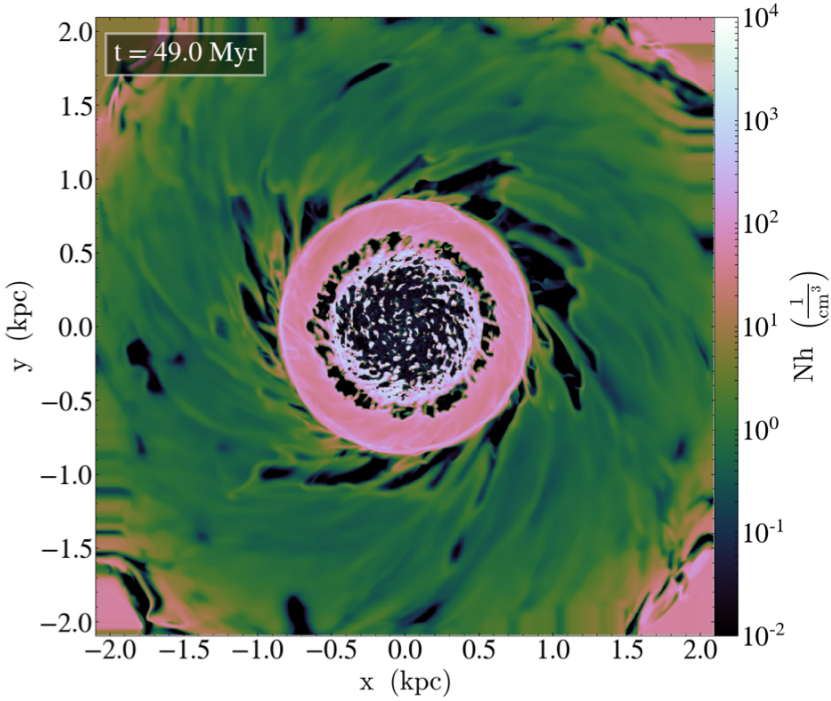
Computational Galaxy Evolution and Cosmology

University of California, Santa Cruz
University of Copenhagen

Davide Martizzi

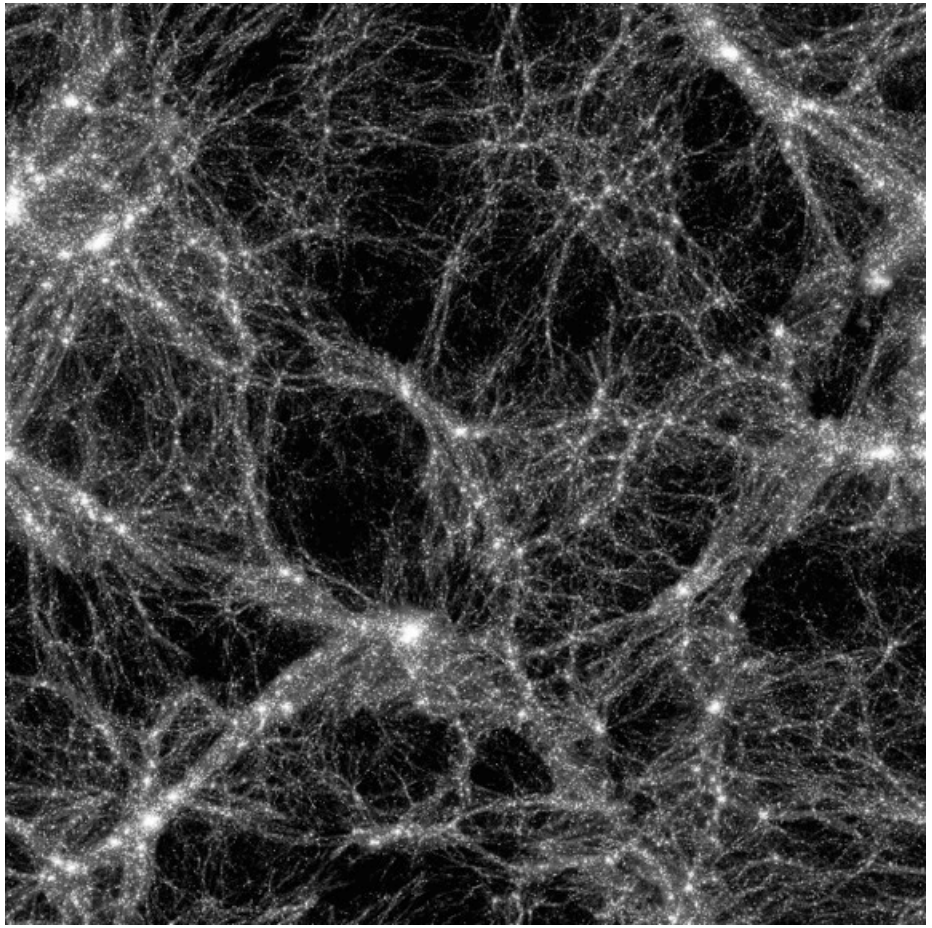
M82: a real galaxy with blowing out gas.
These outflows are called Galactic Winds.





Question: what drives Galactic Winds?
Eulerian hydro-dynamical simulations of a disc galaxy blowing out gas. Supernovae can drive outflows.

Large Scale Structure of the Universe



Cosmic Web: matter distribution on the scale of 10^8 light years.

Questions:

1. How can we detect if a point in the Universe is in a knot, filament, sheet or void?

2. What are the properties of matter and galaxies in each region of the Cosmic Web?

3. What does that tell us about the evolution of the Universe after the Big Bang?

Answers require large simulations, simulation data exploration, feature detection, feature classification, time series analysis, etc..

Mike L Morgan

BAE Systems – Electronic Systems

My work spans Geospatial eXploitation Products to Activity-Based Intelligence to Cloud Computing (specifically AWS).

Algorithms for GXP revolve around registration, ortho-rectification, and image balance. We have been concentrating on image recognition and object tracking, generally requiring parallel stream computing of large data objects.

Algorithms for ABI revolve around tracking multiple events and activities to anticipate nefarious activities. This requires parallelized graph algorithms and semantic technologies.

Cloud Computing, properly implemented, provides unparalleled performance, reliability, and security to enterprise applications. My goal is to find turnkey methods to move algorithms and existing technology into this environment.

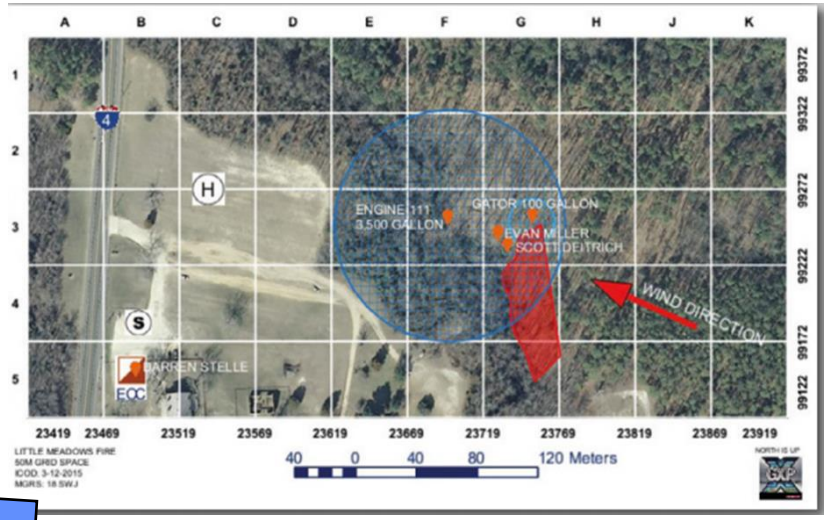
Mike L Morgan

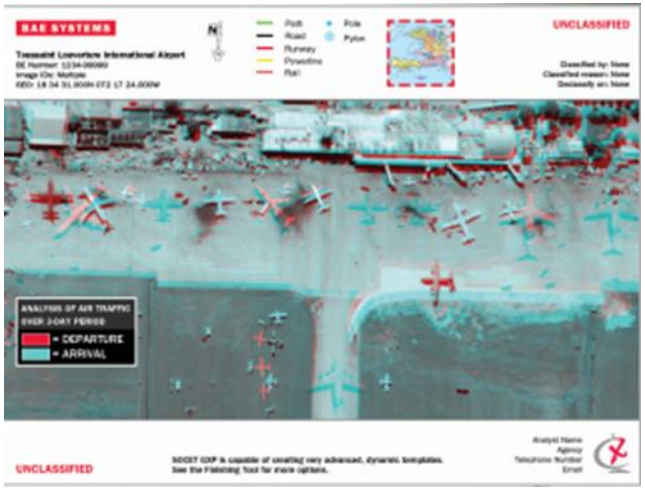
BAE Systems – Electronic Systems

Computational Challenges:

1. Investigate methods to compute at the tactical edge.
2. Find ways to make algorithms loosely-coupled from the computational resources. Essentially, can we develop methods to make algorithm implementation turnkey on different platforms.
3. Look for opportunities to team with academia on various projects.

https://www.geospatialexploitationproducts.com/





Michael Pazzani

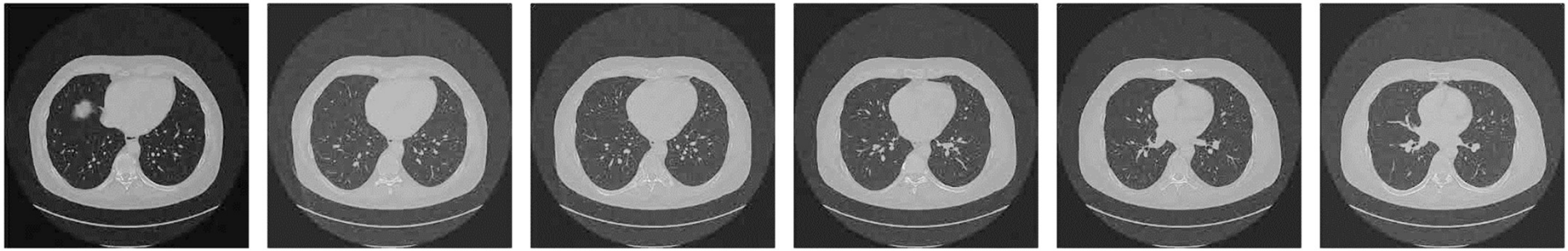
University of California, Riverside

- **Machine Learning**
  - UCI Repository: Medium Data
  - Explanation-based Learning
  - Naïve Bayes
  - Personalization
  - Recommendation Systems
  - Text Classification
  - Contrast Sets
  - Medical Informatics
  - Deep Learning
  - Explainable AI (XAI)

s=-180+10runs+5hit+0.9obp+15hrs+14rbi-0.8ave-18dbl-39trpl
s=-207+15runs+0.8hit+11hr+11rbi+0.33ave+5dbl

Michael Pazzani
University of California, Riverside

1. Deep Learning on Big Data: 3D Radiology, fMRI
- Data Size
- Is ImageNet or AlexNet the right starting point?
- XAI



2. Scaling Contrast Sets to Gene Discover

Find x_1 & … x_n

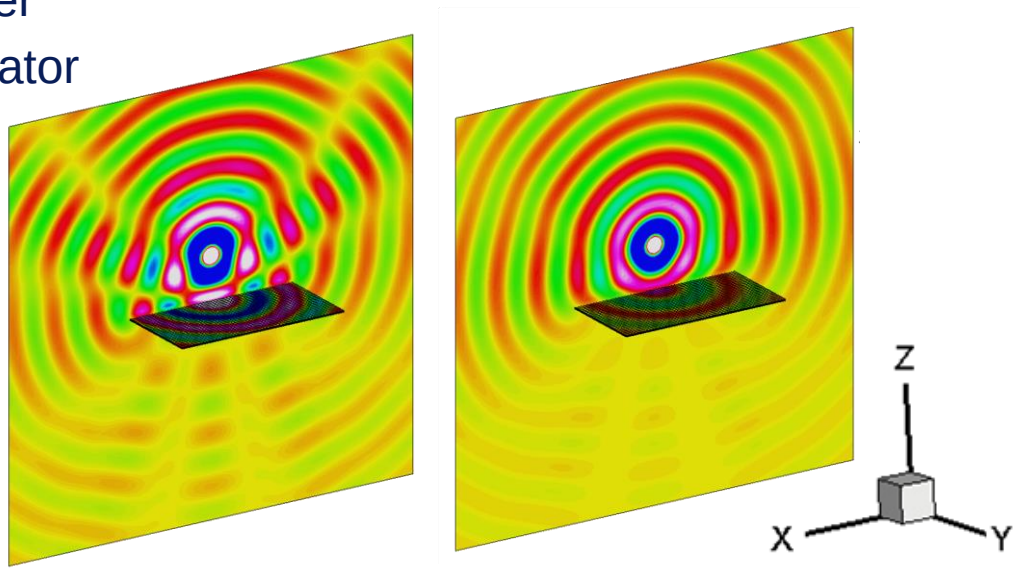
$P(c \mid x_1 \& \ldots \& x_n) > P(c)$

Michelle Pizzo

Old Dominion University // NASA Langley Research Center

- Embry-Riddle Aeronautical
  - BS Aerospace Engineering
  - MS Mechanical Engineering
- Old Dominion University
  - MS Applied Mathematics, Mod/Sim Certification
  - PhD Candidate Applied Mathematics, Computational Aeroacoustics
- NASA Langley Research Center
  - Civil Servant, HPC Incubator

Contour plots of the frequency-domain solutions for a rigid body [left] and soft body [right].



Michelle Pizzo

Old Dominion University // NASA Langley Research Center

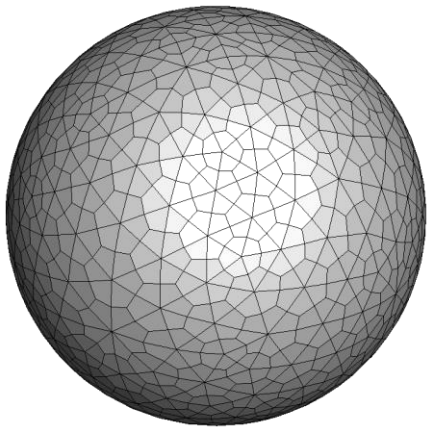
PhD Work: Time-Domain Fast Acoustics Solver

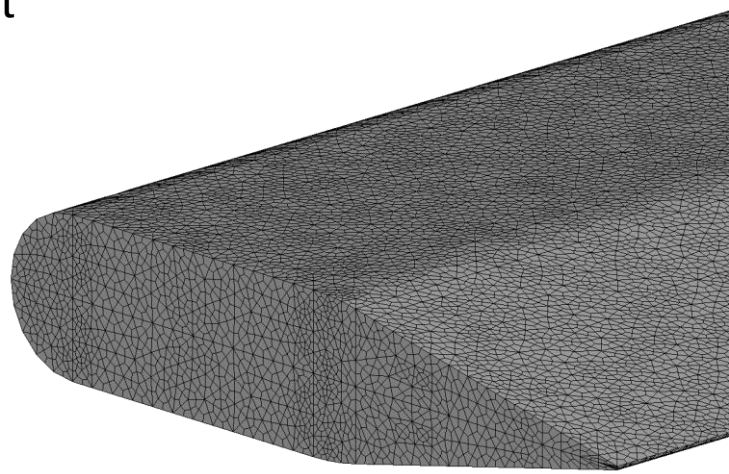
Time-Domain Solvers

- Allow for simulation & study of broadband sources & time-domain transient signals
- Allow for scattering solutions at all frequencies to be obtained within a single computation
- Coupled with nonlinear computational fluid dynamics simulation of noise sources more naturally
- Time-domain boundary integral equations have an intrinsic numerical instability & carry a high computational cost

HPCI

High Performance Computing Incubator





Majid Rasouli
University of Utah

RESEARCH: SOLVING LARGE LINEAR SYSTEMS

$$Ax = b$$

SAENA – SCALABLE ALGEBRAIC MULTIGRID SOLVER
HYBRID: MPI + OPENMP
FUNCTIONALITIES:

- MATVEC
- MATRIX-MATRIX PRODUCT
- ITERATIVE SOLVERS
- EIGEN SOLVERS
- …

Majid Rasouli
University of Utah

WHY THIS SUMMER INSTITUTE:

- COMET
- ADVANCED GIT
- OPTIMIZATION
- VISUALIZATION
- CUDA

Patrick Strassmann

UC San Diego | Salk Institute

Behavioral neuroscience

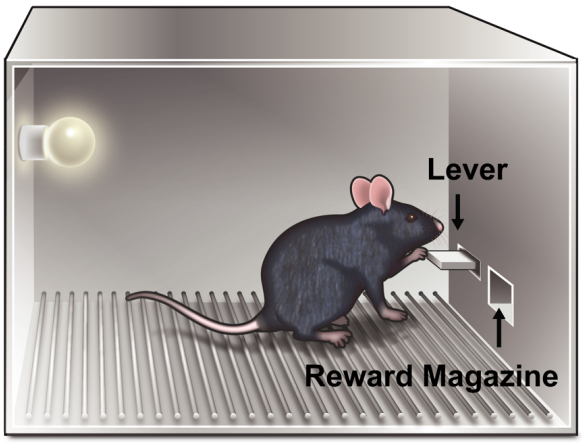
Understand the relationship between brain and complex behavior

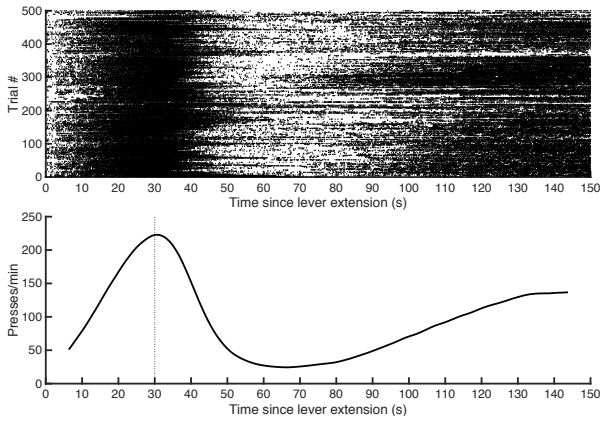
Training and performance of behavior

Record brain activity

Manipulate brain activity

DATA





Patrick Strassmann
UC San Diego | Salk Institute

Challenges:

Highly dynamic brain activity, highly dynamic behavior
Lots of data, obtuse relationships

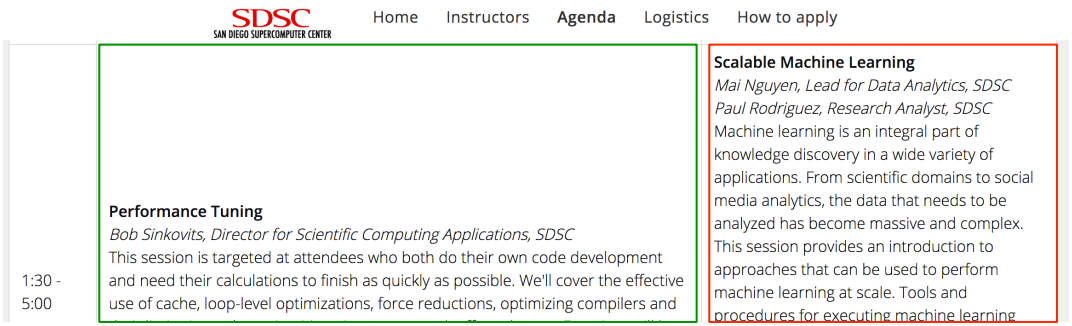
- Scale depth and resolution of ongoing analyses
- Reveal relationships that we don't know to look for
- Generate hypotheses, inform experiments, predict results

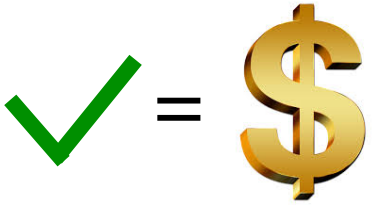
Peng Sun, UC Irvine

1. Human decision making

Or

Left? Right?





2. Human + Machine Intelligence

Better feature engineering; To derive a optimal combination rule

Peng Sun

Human Information Processing Lab, UC Irvine

Computational challenges:

Past: Data size used to be delt with: < 100 M

Now: Tens of millions of online game plays (in hundreds of gigabytes); Medical imaging videos (in TB).

What I want to learn and achieve:

To improve computation efficiency of my program

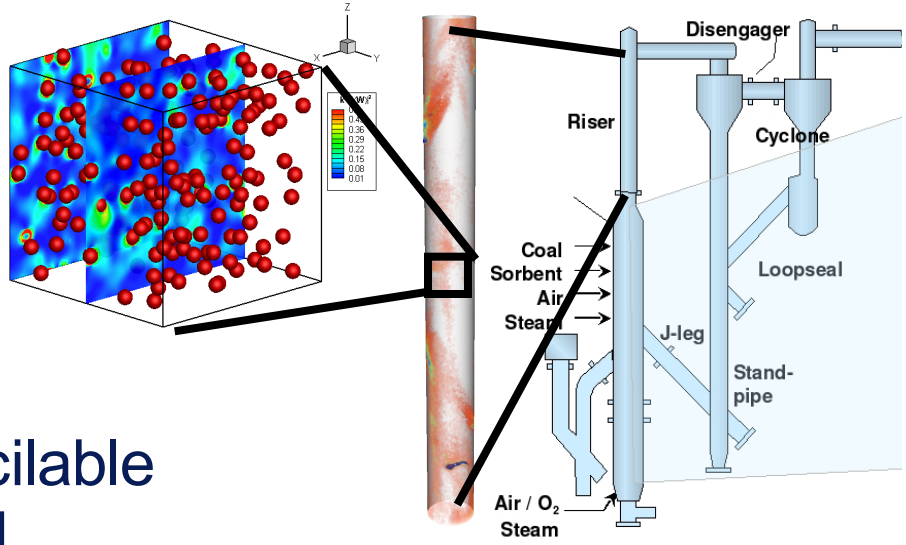
To process and store large amount of data more efficiently

Vahid Tavanashad (Ph.D. Student)

Iowa State University

Analysis of dispersed multiphase flow using particle-resolved direct numerical simulation: flow physics and modeling

- ✓ Governing equations
  - ✓ capture all the details
- ✓ Solution method
  - ✓ PUReIBM: particle-resolved uncontaminated fluid reconcilable immersed boundary method



Wilsonville Coal Gasifier, *Guenther et al.*, 2002, www.mfix.org

- ✓ Challenge: Computational resources
- ✓ Solution: Simplifying equations with mathematical modeling
  - ✓ capture important information
- ✓ How to develop a model?

Vahid Tavanashad (Ph.D. Student)
Iowa State University

High Performance Computing → Generating Data

- ✓ Larger domain is needed for some flow physics
  - ✓ Clustering of dispersed phase (solid particles)
- ✓ More equations should be solved for realistic problems
  - ✓ Heat equation
  - ✓ Species transport equation (Reacting Flows)

Data Science (Machine Learning) → Developing Model

- ✓ Discretization of equations on millions of grid
- ✓ Several quantities at each grid point

Tsogbayar Tsednee (PhD)
California State University, Northridge

My previous research work: theoretical investigations on scattering for few-electron atomic and molecular systems in the field of atomic, molecular and optical physics. Main numerical/mathematical method is a *pseudospectral* method by which a differential equations describing a physical natures of a nature are solved numerically.

- did not use any supercomputing, parallel codes etc.,
- use mostly own written codes (*Fortran (77,90)*; matlab, maple, etc.,)
- no experiences on data analysis;

Current research work: in general, lab's research work is mostly focused on simulations of (drug) biomolecules, such as, DNA, etc.,

- research uses an *integral equation (IE) theory* in study of classical liquids.
- previously own developed a codes based on 1D and 3D reference interaction site model (3D-RISM) (*Fortran (90,95), C++;* Python, etc.,)
- has in-house small cluster (metropolis.csun.edu)

My role since June 2017: theoretical study on IE theory; solve the equation by developing (new) different mathematical approach and apply it for atomic and molecular fluids using own written codes (*Fortran* & matlab).

Tsogbayar Tsednee (PhD)
California State University, Northridge

What I hope to learn at the Summer Institute:

- how to use super computing clusters, such as, Comet etc.,
- understand how it works
- understand what a high-performance computation, CPU and GPU are.
- understand what a code parallelization, MPI and OpenMP, etc., are.
- get acquainted with data science

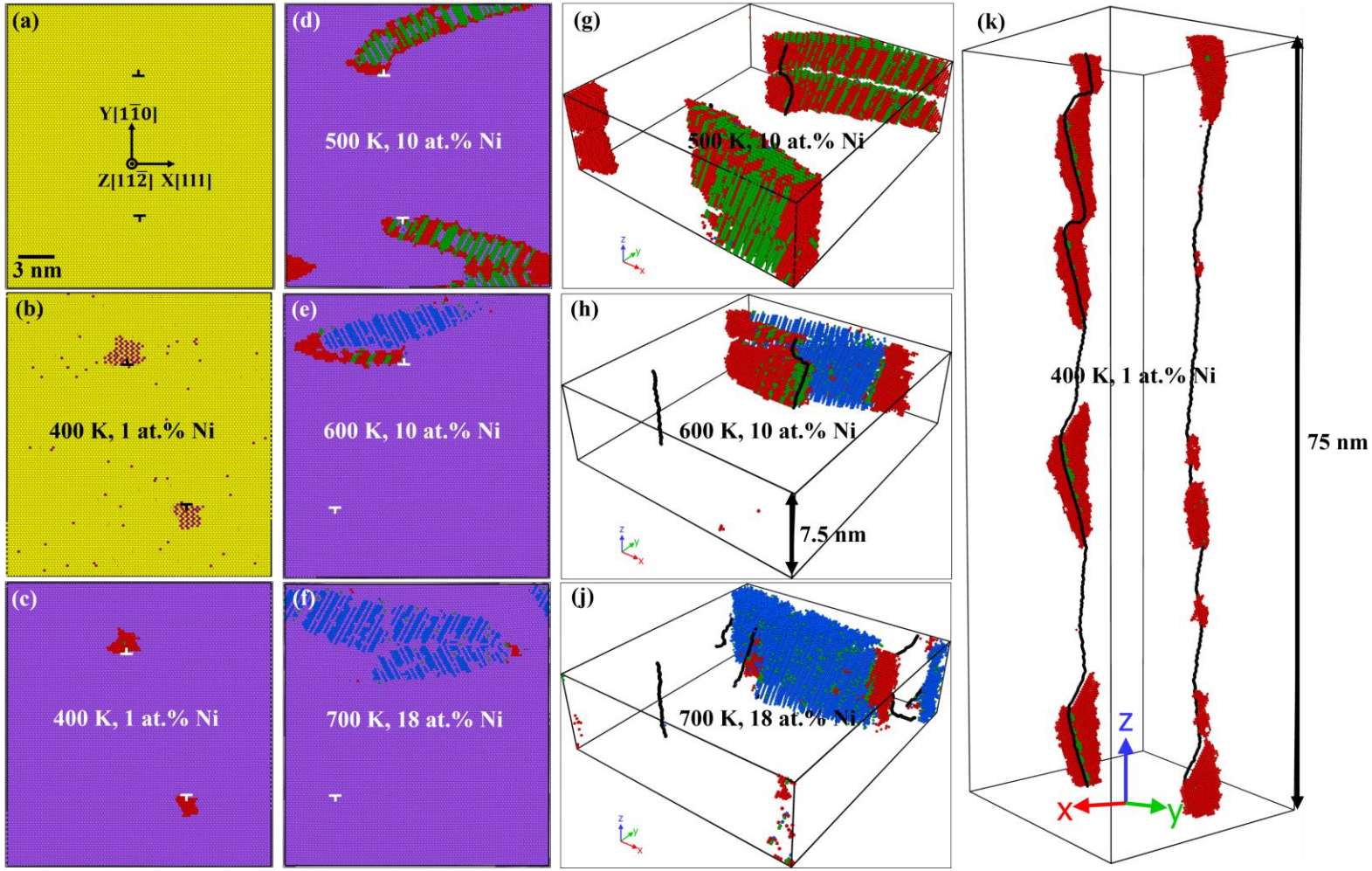
In future:

- I will employ such a kind of computational cluster, such as, Comet in my research of simulation of large biomolecules.
- I will try to use a parallelization of codes if applicable

Vladyslav Turlo

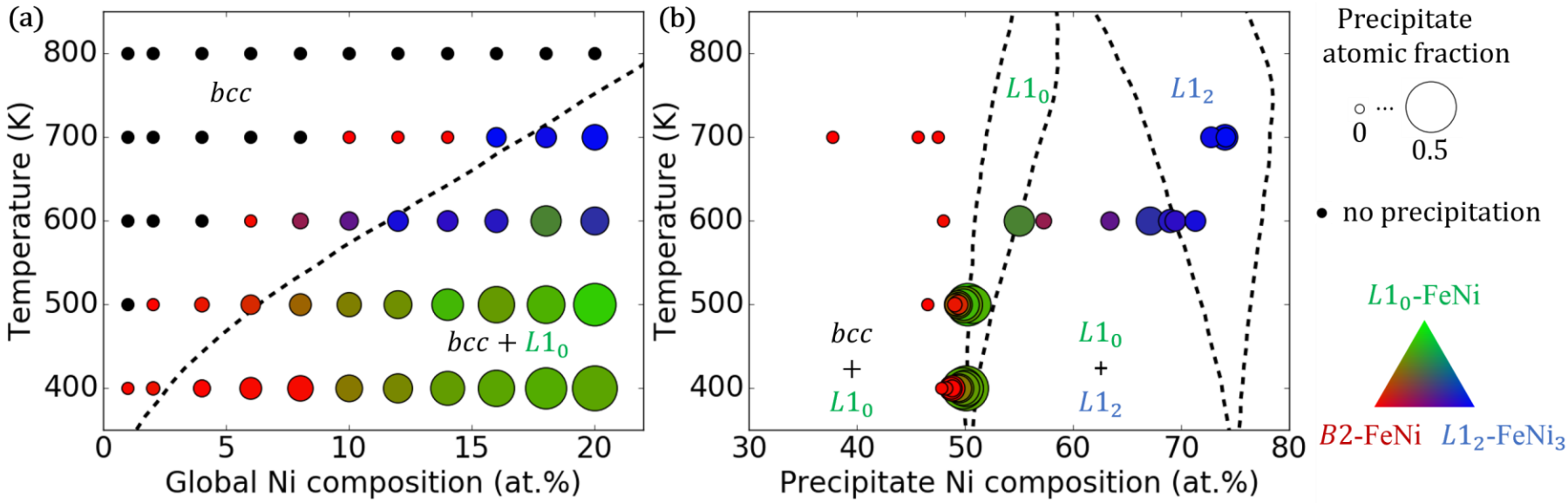
University of California, Irvine

Atomistic simulations of metals and alloys



Vladyslav Turlo
University of California, Irvine

- Exploring complexion (phase) diagrams is challenging:
  each point requires 1 node - 64 cores - 1 week



- Using optimized parallelization and machine learning techniques, we hope to build complexion diagrams based on material's properties and limited number of simulations

Manuel Valera

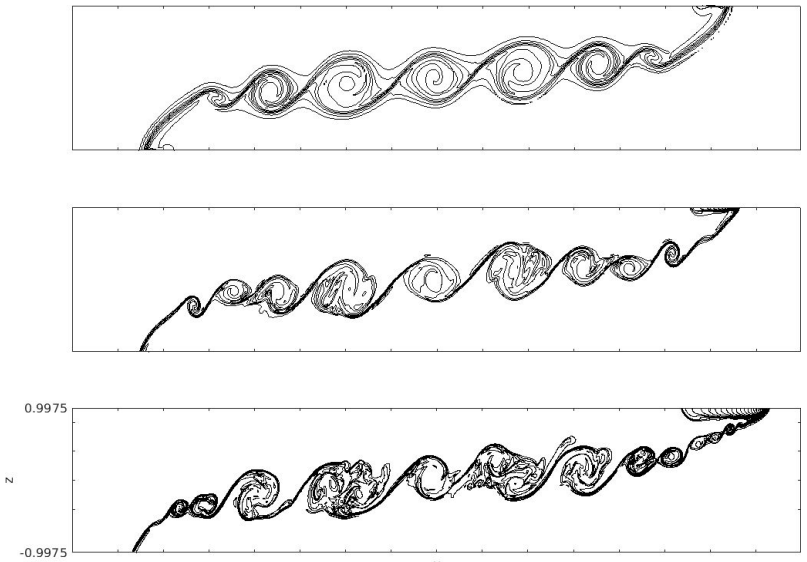
San Diego State University // Claremont Graduate University

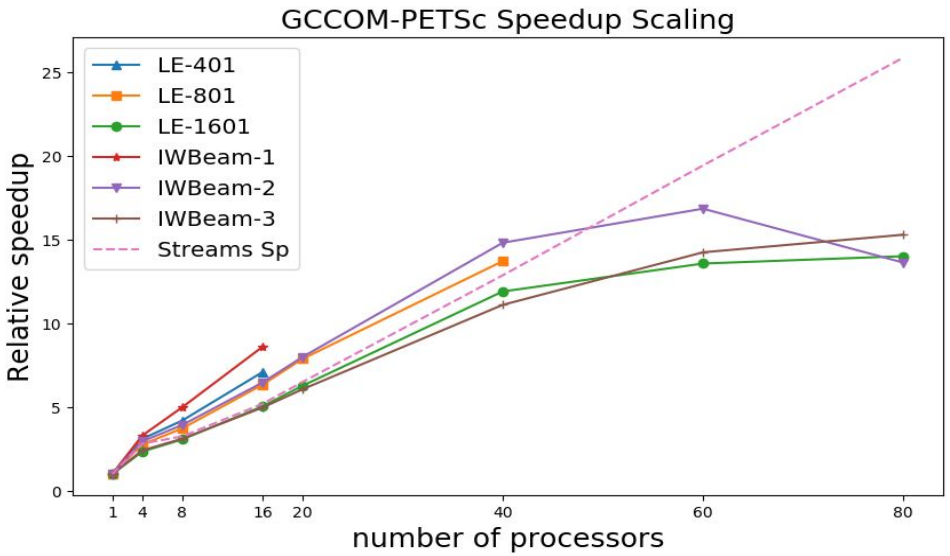
Main Research Task:

- Parallelize a Coastal Ocean Dynamics model (GCCOM) with PETSc:
  - Data arrays distribution and Linear System solver
  - Validate against serial version
  - Carry out performance tests

Goal:

- Be able to run field-scale experiments on the model:
  - Internal waves and Bores in underwater canyons on the West Coast (Monterey Bay (CalPoly) and La Jolla Canyon (Scripps / UCSD))

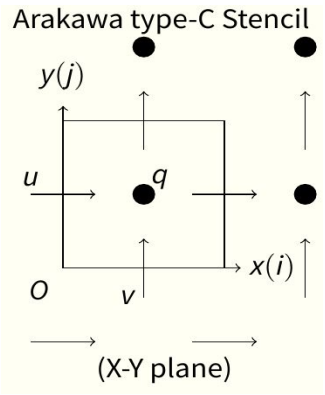


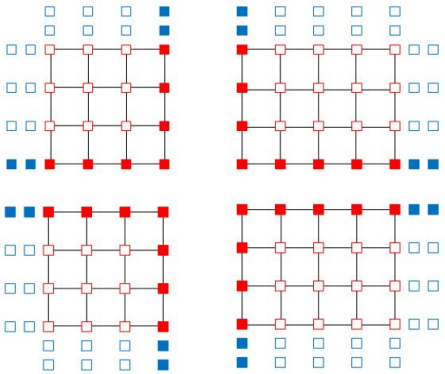


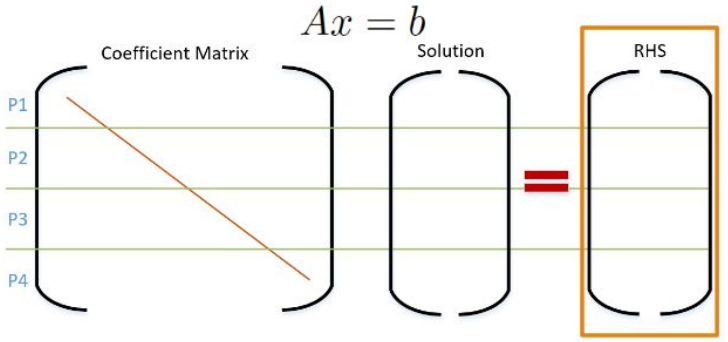
M. Valera, *et al*, "Parallel Implementation of a PETSc-based Framework for the General Curvilinear Coastal Ocean Model", Journal of Computational Geosciences. *Submitted.*

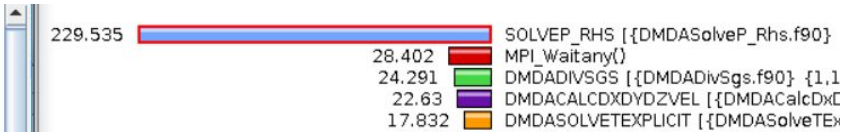
Manuel Valera

San Diego State University // Claremont Graduate University





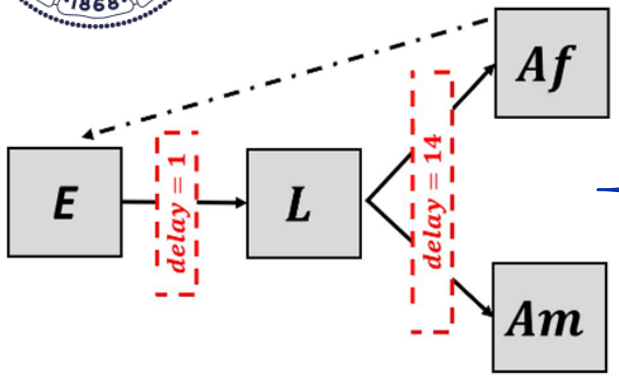




Challenges:

- No previous knowledge or residing expert in PETSc
- Arakawa-C type grid
- Full 3D curvilinear transformation meshes
- Unique transformed Laplacian (56% of time)
- Modules for advection, dispersion, turbulence, thermodynamics, and so on.
- Moving into MPI/GPU Hybrid implementation.

Valeri Vasquez
University of California, Berkeley



Dynamic population model developed to study the **optimal release strategy** for genetically modified mosquitoes, with applications for eradicating malaria.

Fig. 1. A visual representation of the population dynamics model. The entire system is solved in each timestep.

Model Predictive Control (MPC) is a dynamic programming approach based on an **iterative, deterministic, finite-horizon** optimization.

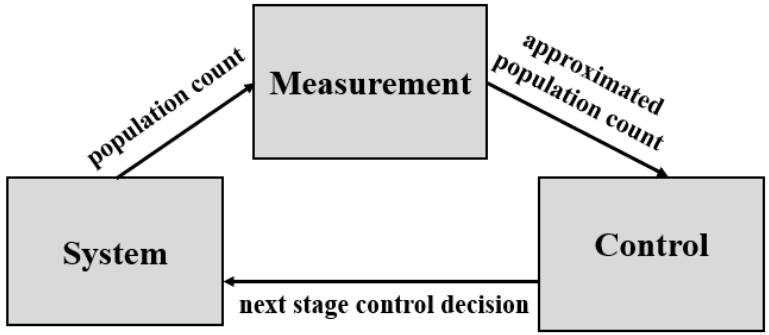
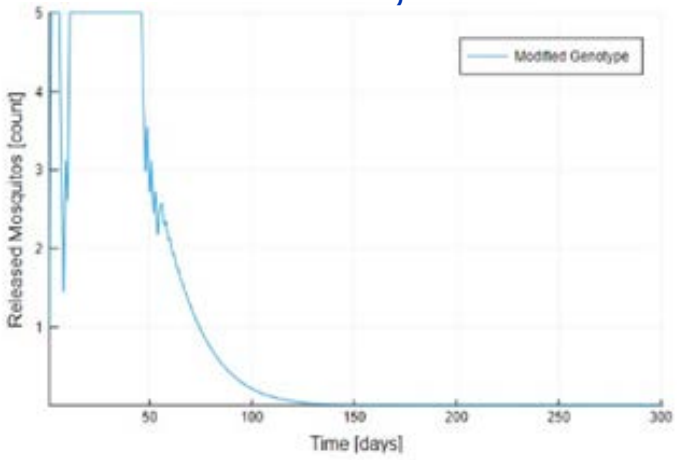


Fig. 2. A conceptual diagram of the MPC process.



The objective function seeks to **reduce the difference** between the wildtype and modified adult male populations as much as possible in each timestep.

Fig. 3. The control trajectory in the 5-step lookahead MPC problem.

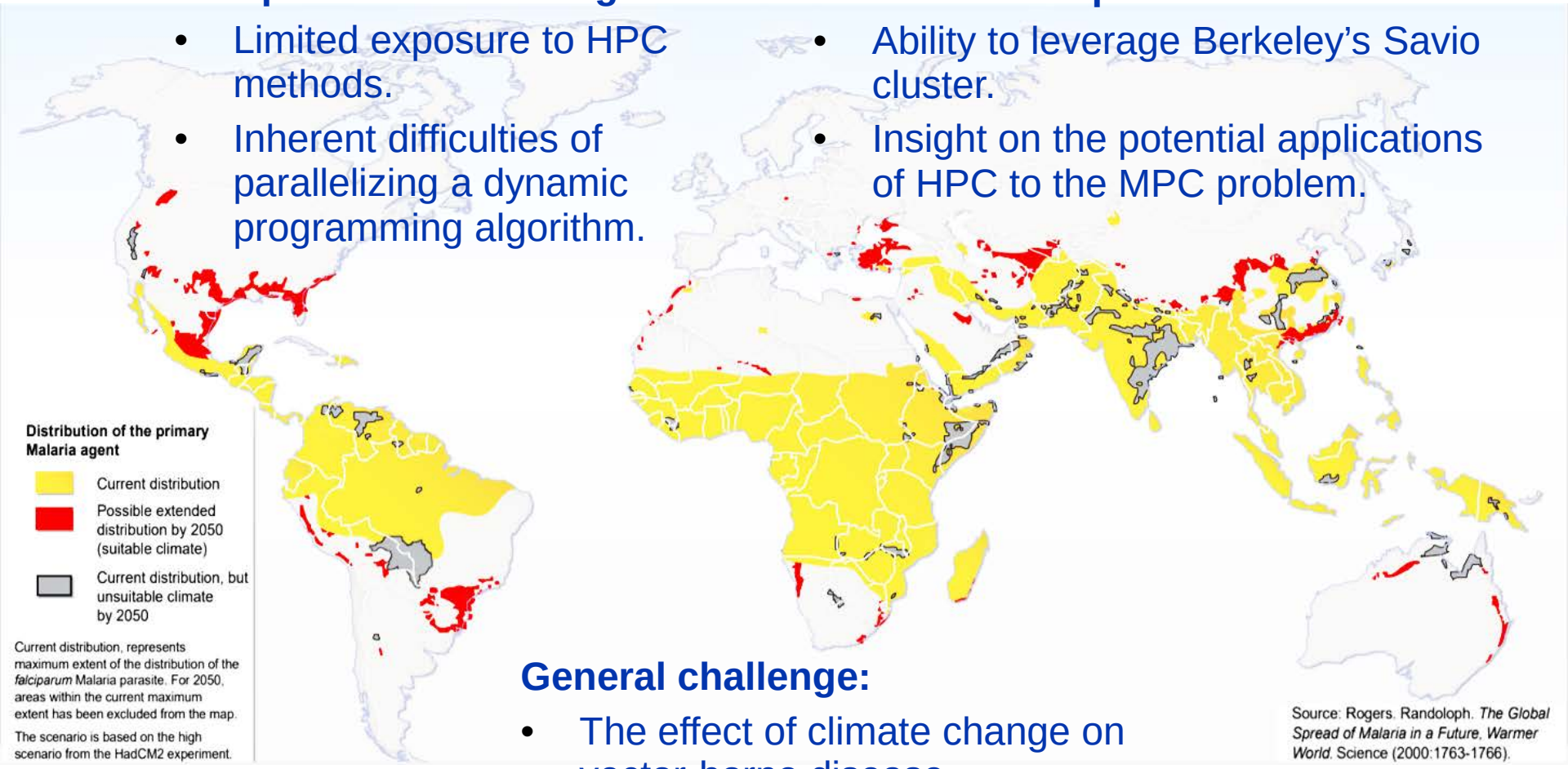
Valeri Vasquez
University of California, Berkeley

Computational challenges:

- Limited exposure to HPC methods.
- Inherent difficulties of parallelizing a dynamic programming algorithm.

Desired accomplishments:

- Ability to leverage Berkeley's Savio cluster.
- Insight on the potential applications of HPC to the MPC problem.



General challenge:

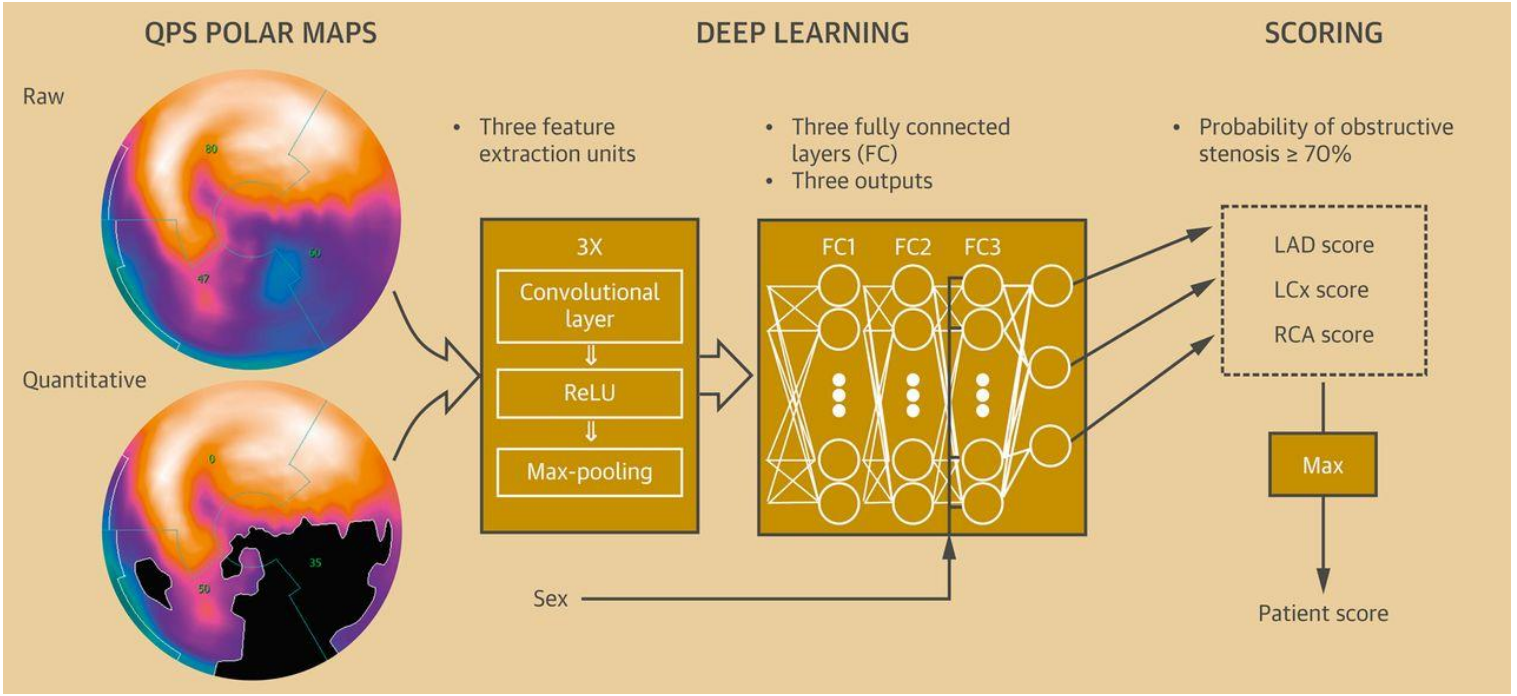
- The effect of climate change on vector-borne disease.

Frances Wang

Case Western Reserve University

Biostatistician

Research Focus: Cardiovascular Imaging, Machine Learning for Assessing Risk of CVD



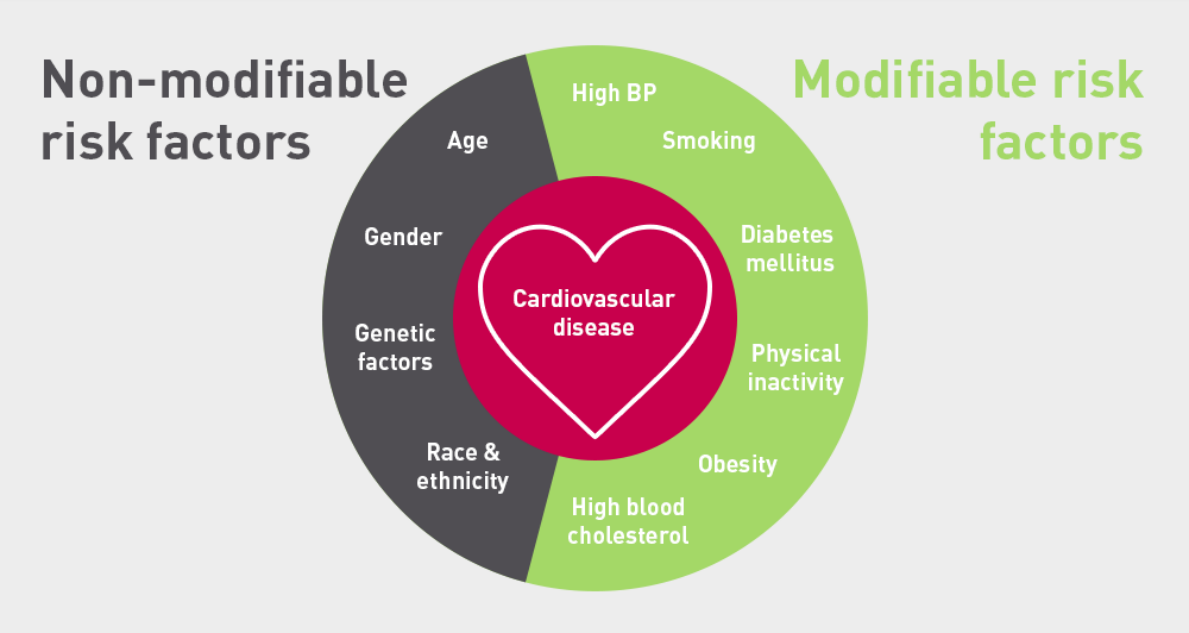
Betancur, 2018; JACC:Img

Frances Wang

Case Western Reserve University

Future Research Direction: Electronic Medical Records; Machine Learning for Identifying Risk Factors to Better Predict Patient Prognosis

Goals for SDSC: Learn more about the usage of HPC resources for machine learning and big data processing



Towards data-driven quantum-accurate Neural Network Interatomic Force Fields

Ji Wei Yoon (UC Berkeley)

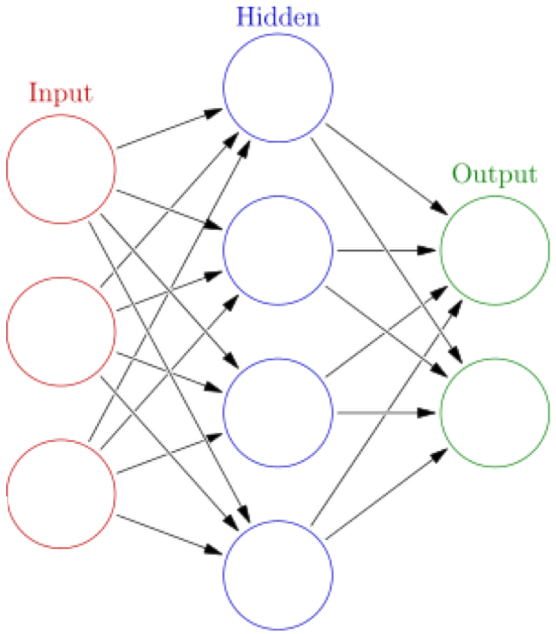
Specifics

- **Methodology**

1. Flexible Neural Network functional form
2. Behler-like descriptors to describe Atomic Environments
3. Pytorch for dynamic computation graphs
   (as opposed to tensorflow that deals with static graphs)

- **Data**

1. Mo dataset from UCSD Ong's group (MSE) in the form of ~100 mb of JSON files.



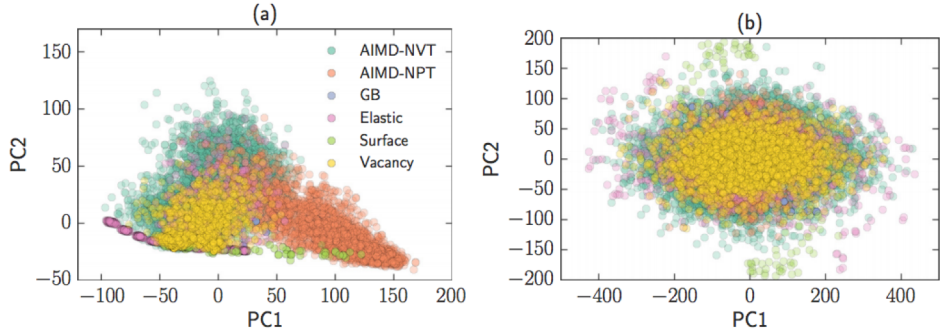


FIG. 2. Two-dimensional projection of the principal components of (a) the atomic bispectrum coefficients and (b) their first derivatives.

Berkeley
UNIVERSITY OF CALIFORNIA

Resources

- Berkeley Research Computing Savio Cluster

- Jupyter notebooks on Jupyterhub installation on the cluster

Partition	Nodes	Node List	CPU Model	# Cores/Node	Memory/Node	Infiniband	Speciality	Scheduler Allocation
savio	164	n0[000-095].savio1 n0[100-167].savio1	Intel Xeon E5-2670 v2	20	64 GB	FDR	-	By Node
savio_bigmem	4	n0[096-099].savio1	Intel Xeon E5-2670 v2	20	512 GB	FDR	BIGMEM	By Node
savio2	136	n0[027-162].savio2	Intel Xeon E5-2670 v3	24	64 GB	FDR	-	By Node
savio2	4	n0[183-186].savio2	Intel Xeon E5-2680 v3	24	64 GB	FDR	-	By Node
savio2	4	n0[290-293].savio2 n0[230-240].savio2	Intel Xeon E5-2650 v4	24	64 GB	FDR	-	By Node
savio2	35	n0[187-210].savio2 n0[230-240].savio2	Intel Xeon E5-2680 v4	28	64 GB	FDR	-	By Node
savio2_bigmem	20	n0[163-182].savio2	Intel Xeon E5-2670 v3	24	128 GB	FDR	-	By Node
savio2_bigmem	8	n0[282-289].savio2	Intel Xeon E5-2650 v3	24	128 GB	FDR	-	By Node
savio2_htc	20	n0[000-011].savio2 n0[215-222].savio2	Intel Xeon E5-2643 v3	12	128 GB	FDR	HTC	By Core
savio2_gpu	17	n0[012-026].savio2 n0[223-224].savio2	Intel Xeon E5-2623 v3	8	64 GB	FDR	4x Nvidia K80	By Core
savio2_1080ti	3	n0[227-229].savio2	Intel Xeon E5-2623 v3	8	64 GB	FDR	4x Nvidia 1080ti	By Core
savio2_knl	28	n0[254-281].savio2	Intel Xeon Phi 7210	64	188 GB	FDR	Intel Phi	By Node

Challenges

- Generated training data(~150 Gb) fills up the memory of a node and cause the execution to crash:

  Find ways to implement multiple node generation of training data and training

- Matrices are small in size so does not achieve good parallelism during training:

  Rewrite code to introduce batching. Then, use GPU (get allocation).

- If GPU be used, need to generate training data on CPU and then run on GPU. Lack of space on home directory storage (10 Gb allocated) so cant write it out. Could explore the use of global scratch but need to know scratch policy.

Berkeley
UNIVERSITY OF CALIFORNIA

Unsupervised Clustering for Mass Cytometry Data

Xuhong Zhang
School of Public Health
CU Anshutz

Flow Cytometry / Mass Cytometry

- In biotechnology, **flow cytometry** is a laser- or impedance-based, biophysical technology employed in cell counting, cell sorting, biomarker detection and protein engineering, by suspending cells in a stream of fluid and passing them through an electronic detection apparatus. A flow cytometer allows simultaneous **multiparametric** analysis of the physical and chemical characteristics of up to thousands of particles per second.
- Flow cytometry is routinely used in the diagnosis of health disorders, especially blood cancers, but has many other applications in basic research, clinical practice and clinical trials.
- More recently, **mass cytometry** is a mass spectrometry technique based on inductively coupled plasma mass spectrometry and time of flight mass spectrometry used for the determination of the properties of cells (cytometry).
  - The practical flow rate is around 500 cells per second versus several thousand in flow cytometry.
  - Current chemical methods limits cytometer use to around 40 parameters per cell

Unsupervised Clustering

- We have hundreds of samples (people) for disease/control groups. Each sample, we have millions of single cells. For each cell, we have 38 parameters/bio makers.
- Unsupervised clustering will help us to learn and identify unknown cell subpopulations.
- BUT, we tried:
  - Classical unsupervised clustering, like K-means ---- not satisfying our requirement
  - MCMC ---- never converge
  - Bayesian Variational Inference ---- out of memory problem