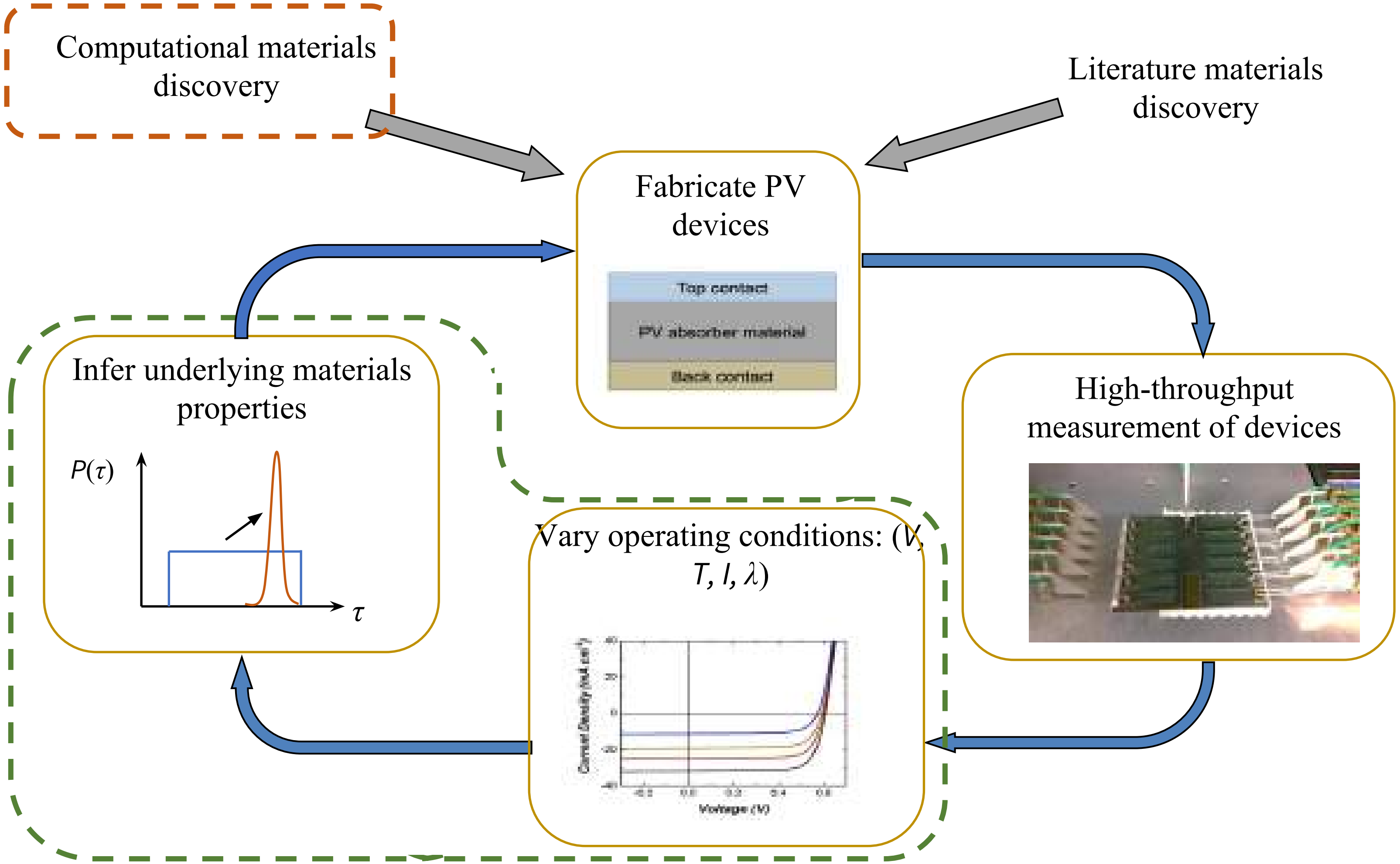
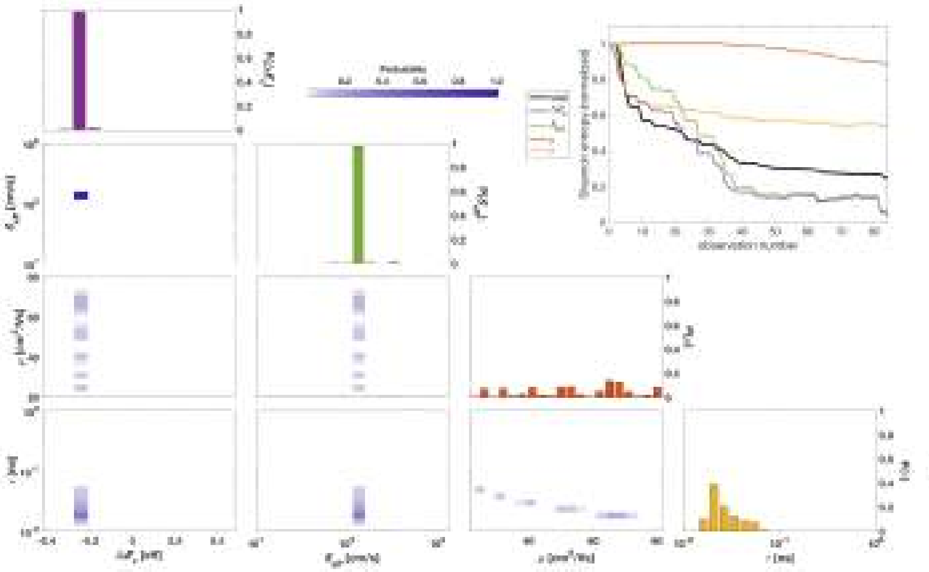
User Support Services **shelley.knuth@colorado.edu**
Shelley Knuth
Research Computing, University of Colorado-Boulder

- Infrastructure is only as good as how well the users understand how to use it
- Manage and provide many services
  - Trainings
    - “Basics of Supercomputing” Bootcamp
    - “New User Seminar”
    - “Parallelization in Scripted Languages”
    - Software Carpentry
    - Python Short Course
    - Bokeh
  - Consulting
    - Student cohort
    - Office hours – data, HPC
    - Help desk
    - Online documentation
    - Videos

Rachel Kurchin: Discovery of Next-Generation Photovoltaic Materials

- Density functional theory (VASP)
- Seeking to understand the physics of point defects (structure, energetics, etc.) in order to design materials less susceptible to their ill effects on device performance





- Run PV device simulators (pre-existing packages) to produce predicted current-voltage sweeps at a variety of values for input parameters
- Use Bayesian parameter estimation to find best values for those parameters

- **Fast algorithms** for fluid-structure and fluid-particle interactions.

Figure: A suspension of colloidal boomerangs (B. Sprinkle *et al.*)

- $\mathbf{Q} = \{\mathbf{q}_\beta, \boldsymbol{\theta}_\beta\}_{\beta=1}^{N}$: **positions** and **orientations**
- The Ito stochastic equation of **Brownian Dynamics** (BD) is

$$\frac{d\mathbf{Q}}{dt} = \mathcal{N}\mathbf{F} + (2k_B T\mathcal{N})^{\frac{1}{2}}\mathcal{W}(t) + (k_B T)\,\partial_{\mathbf{Q}} \cdot \mathcal{N}$$

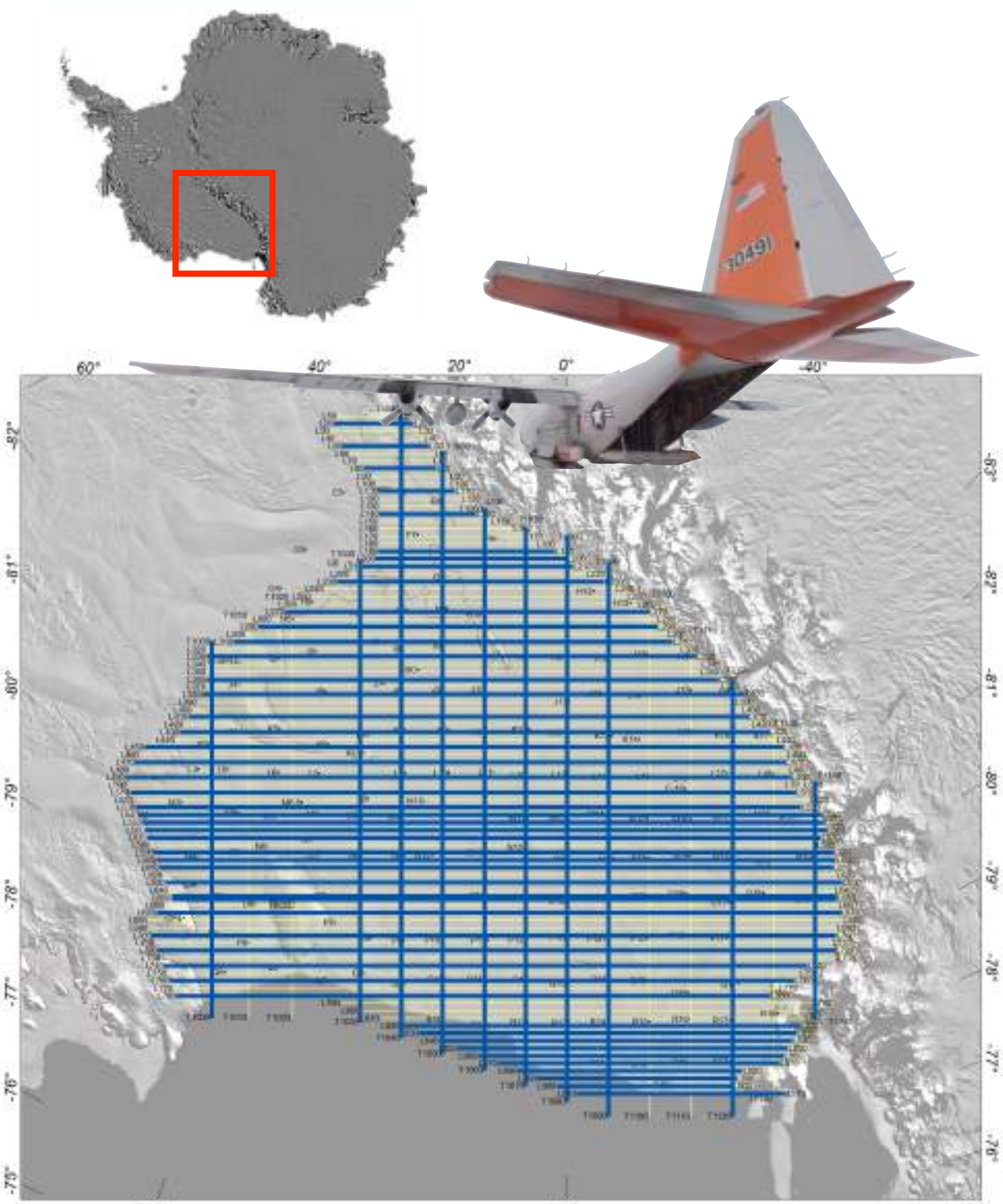
  Hydrodynamic interations, Brownian displacements, stochastic drift.
- $\mathcal{N}(\mathbf{Q})$ is a $6N \times 6N$ *dense* matrix.
- $\mathcal{N}^{\frac{1}{2}}$ is a matrix "square root", defined by $\mathcal{N}^{\frac{1}{2}}\left(\mathcal{N}^{\frac{1}{2}}\right)^{\top} = \mathcal{N}$.
- Goal: to develop methods that can scale for $N \sim O(10^4) - O(10^5)$.

$$\frac{d\mathbf{Q}}{dt} = \underbrace{\mathcal{N}\mathbf{F}}_{O(N^2)} + \sqrt{2k_B T}\,\underbrace{\mathcal{N}^{\frac{1}{2}}}_{O(N^3)}\,\mathcal{W}(t) + (k_B T)\,\partial_{\mathbf{Q}} \cdot \mathcal{N}$$

- Naively, computing the right-hand-side is $O(N^3)$ for every time step (even if you knew $\mathcal{N}$ already)!
- Demands both **fast algorithm** and **parallel implementation**.
- We now have a fast method that scales $O(N)$ and a serial implementation.
- Heavily builds on top of fast linear solvers (GMRES, Lanczos, FFT).
- Things to try: Python/Cython, Numba, PyCUDA/CUDA.
- Code optimization and performance analysis.

Mapping Ross Ice Shelf with Airborne Laser Altimetry

Maya Becker, Scripps Institution of Oceanography, UC San Diego
Advisor: Dr. Helen Amanda Fricker



Raw GPS/ attitude data

↓ Dask?

Processed GPS/ attitude data + Raw laser data (proprietary format)

↓ Singularity container

Georeferenced elevations (LAS/ASCII format)

↓

Application of geophysical corrections

↓ MPI/OpenMP

Corrected height values

Tracking oceanic eddies

Suyash Bire, Christopher Wolfe

August 4, 2017

Eddy Tracking

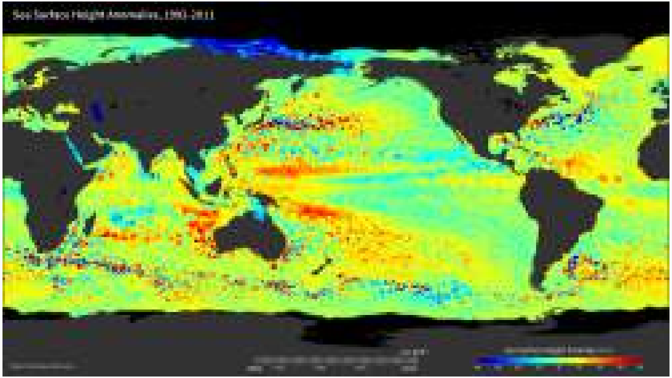


Figure 1: Eddies in the ocean

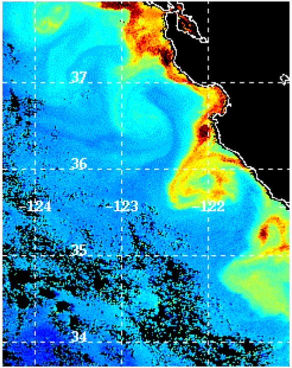
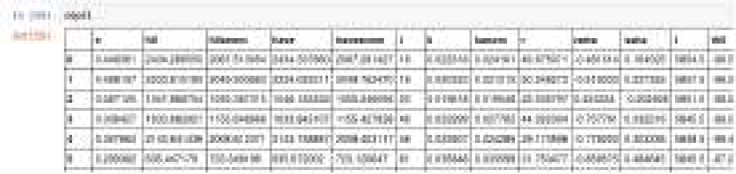
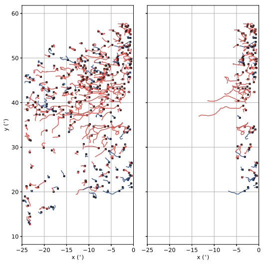


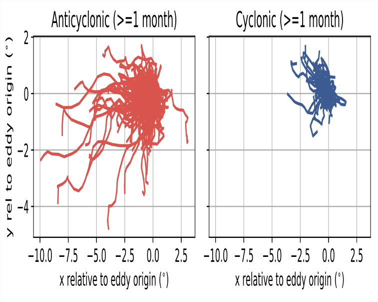
Figure 2: California Current eddies

Animation!

Properties of individual eddies

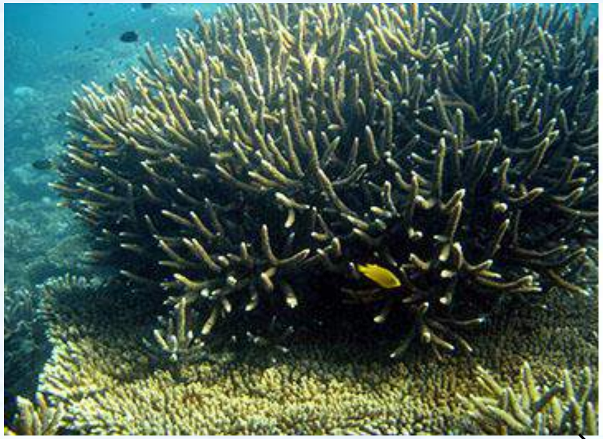


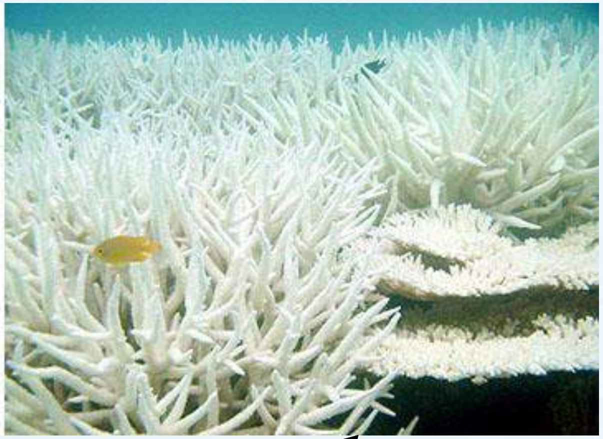




Tracking code makes use of python’s multiprocessing module! https://github.com/suyashbire1/eddy_tracking

Problem: Coral Reef Bleaching





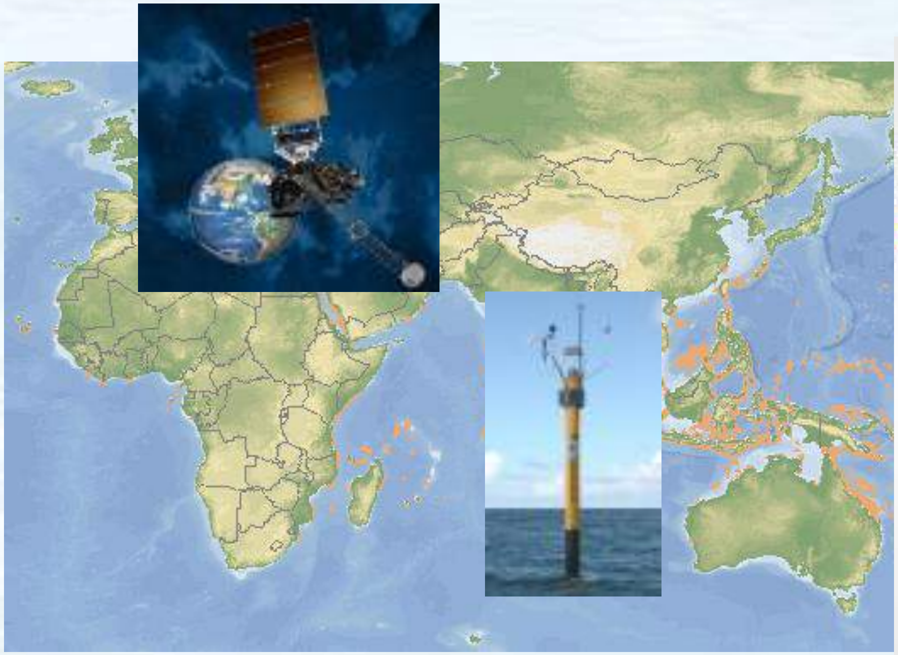
- Under natural and anthropogenic stress the zooxanthellae algae leaves the coral
- The coral is likely to die after several weeks

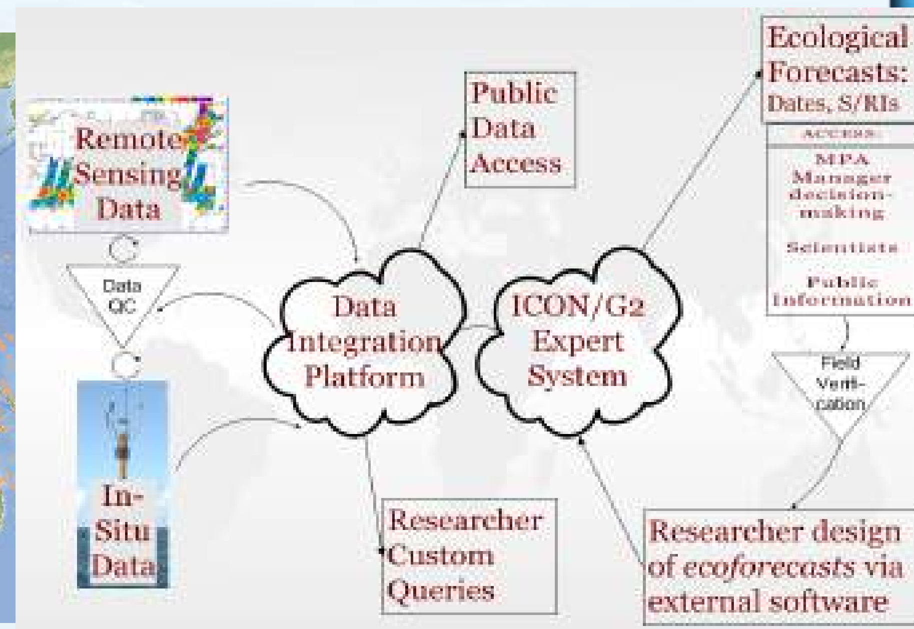
According to a study published by Scientific American last October 2014

"If tropical reefs and other ecosystems are destroyed, the oceans could lose $1 trillion in economic value "by the end of the century,"

Sources:
http://www.ucar.edu/communications/staffnotes/0803/images/coral.jpg:
http://www.nbcnews.com/business/economy/caribbean-economies-face-peril-coral-reefs-decline-n347731

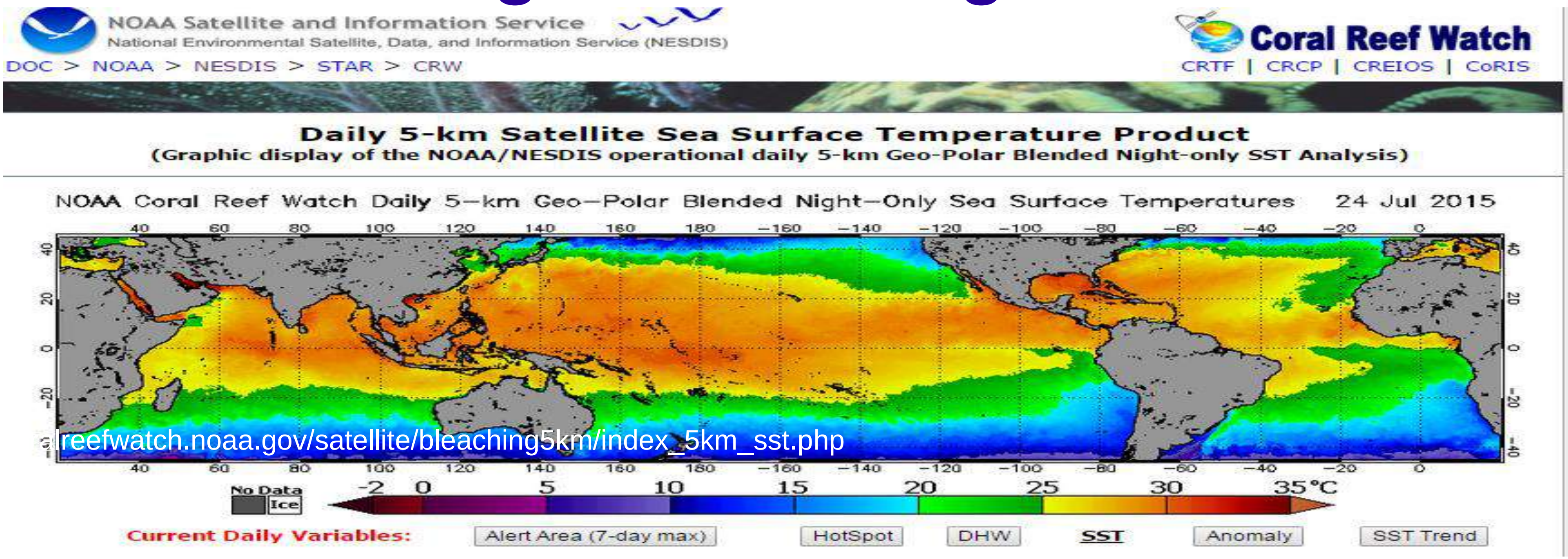
NOAA's Integrated Ocean Observing System



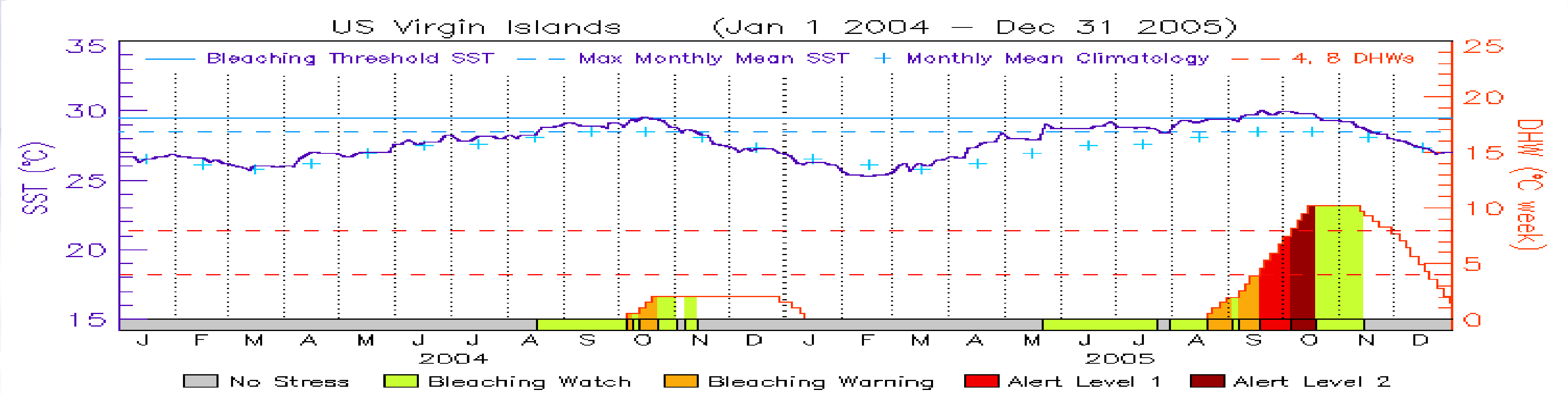


Satellites, radars, weather stations, weather balloons, sounders, buoys, along with expertise and the data management infrastructure needed for
monitoring and analyzing the system=> BIG DATA SETS

Background:Sea Surface Temperature and Degree Heating Week

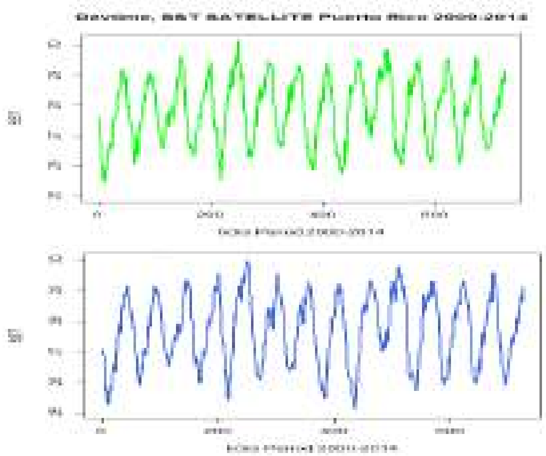


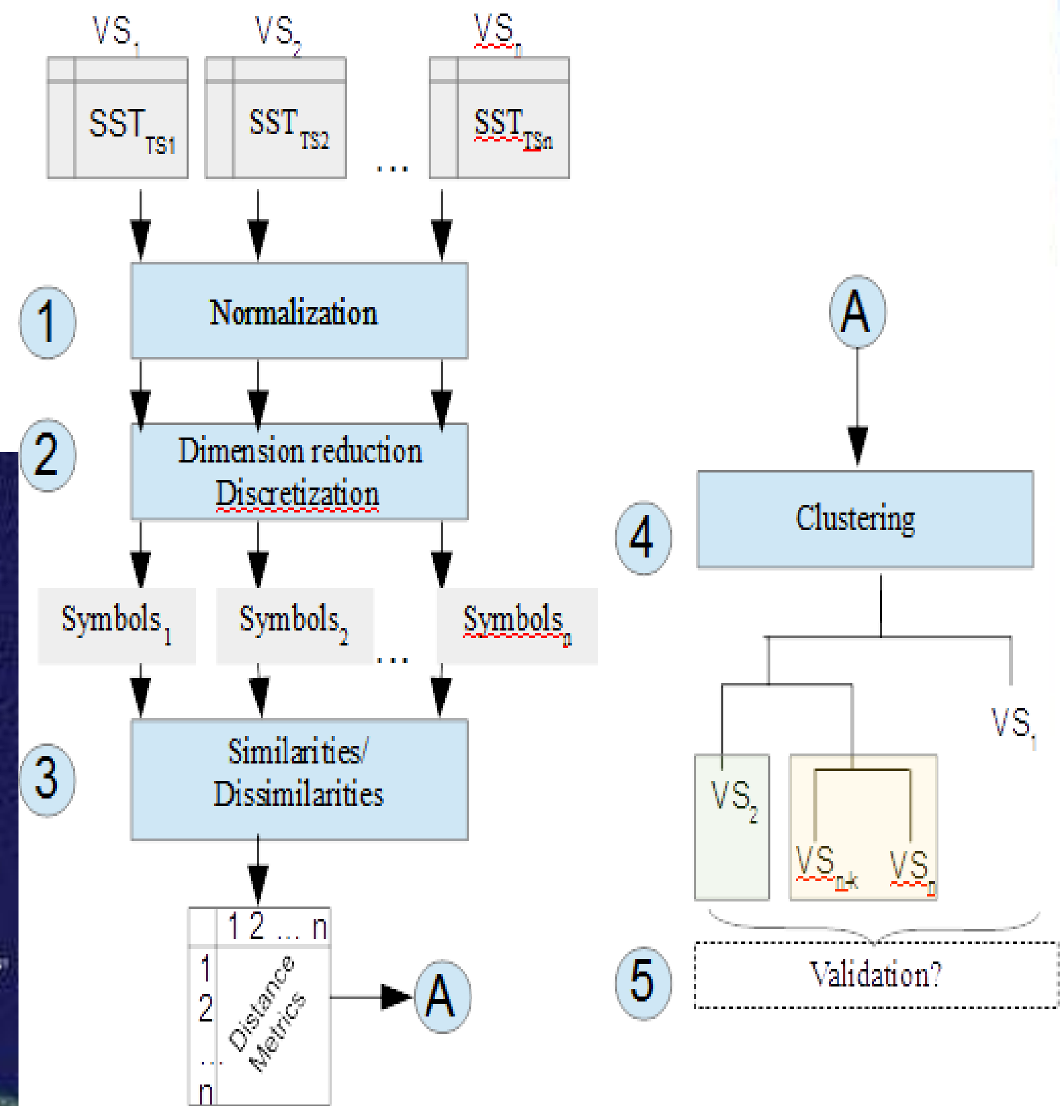
http alreefwatch.noaa.gov/satellite/bleaching5km/index_5km_sst.php



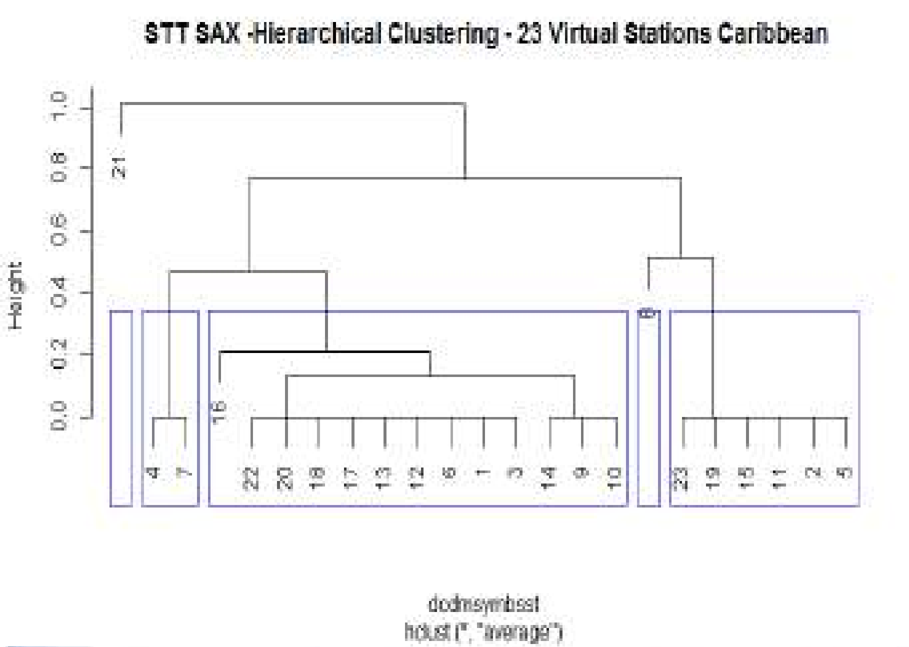
http://coralreefwatch.noaa.gov/satellite/bleaching5km/index_5km_baa_max_r07d.php

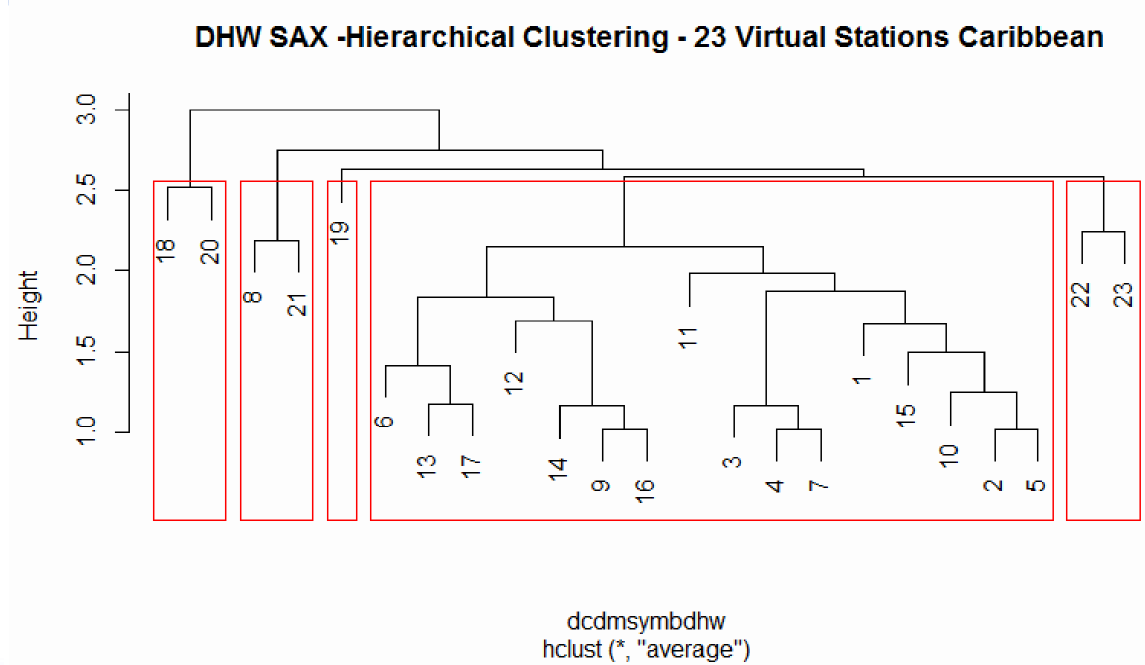
Overall Method





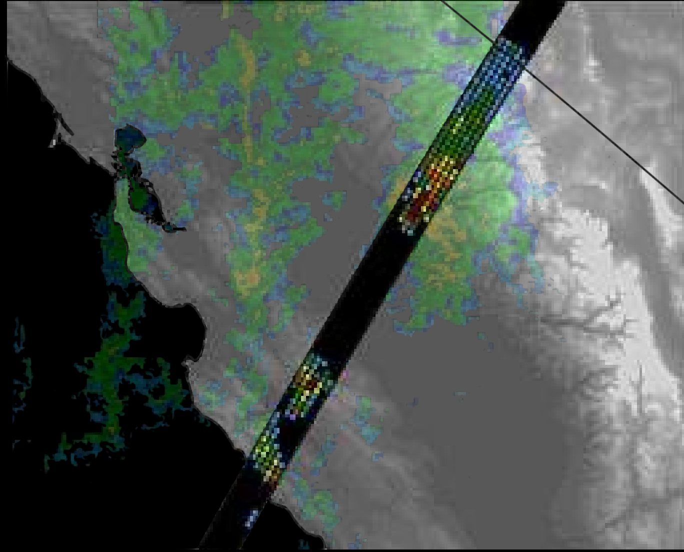
Sea Surface Temperatures vs. Degree Heating Week

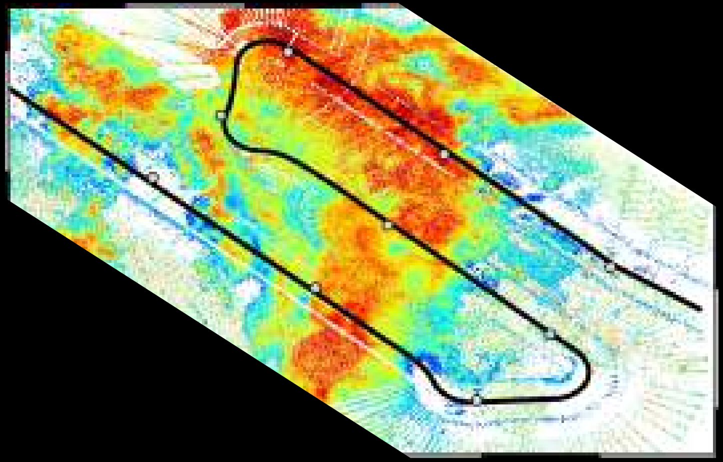




Radar Validation of Weather Forecasts in California

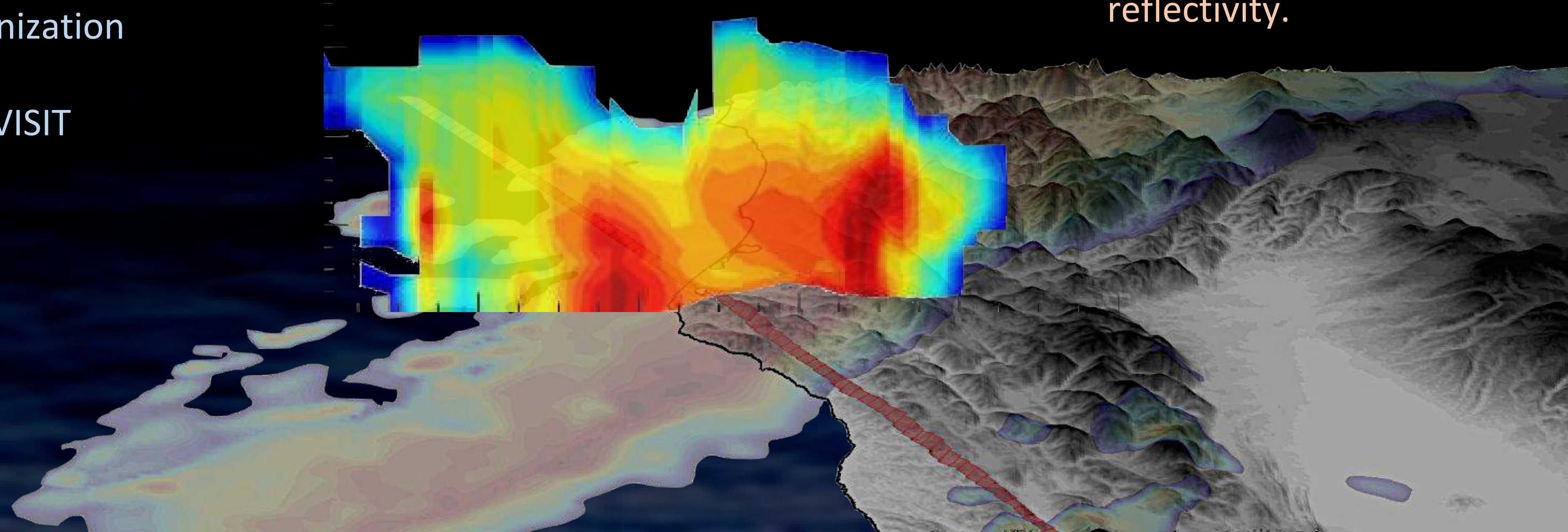
Forest Cannon
Center for Western Weather and Water Extremes
Scripps Institution of Oceanography, UCSD





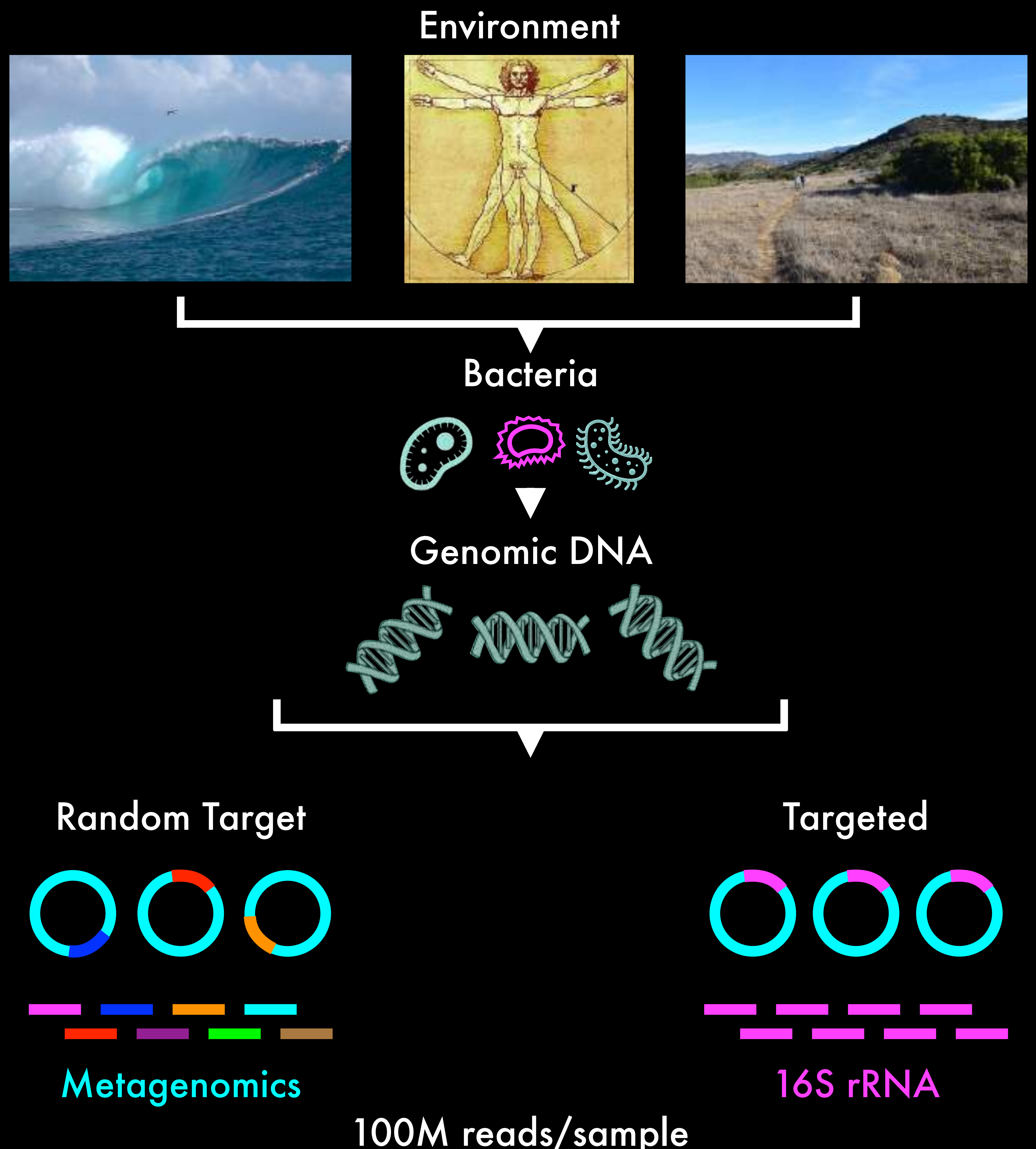
- Weather forecasts are an original HPC problem
- Additional need for NRT assimilation of observations & rapidly verified forecasts
- Complex geometry problem with multiple dependencies
- Well-suited for workflow organization
- Could benefit from improved VISIT

Various aircraft, satellite, and ground-based radars to compare with simulated reflectivity.



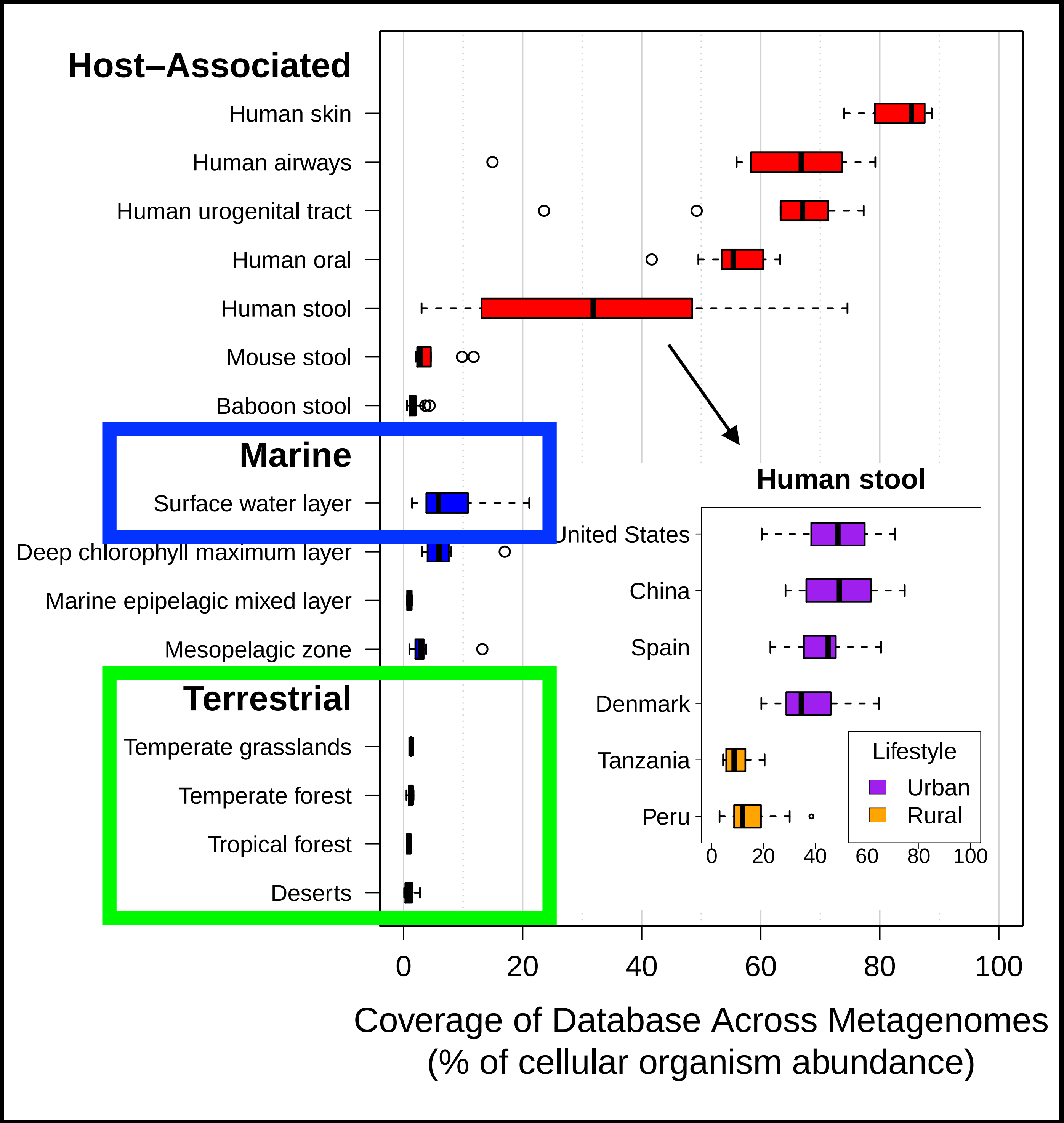
microbial ecology

Alexander B. Chase
UC Irvine
Dr. Jennifer Martiny

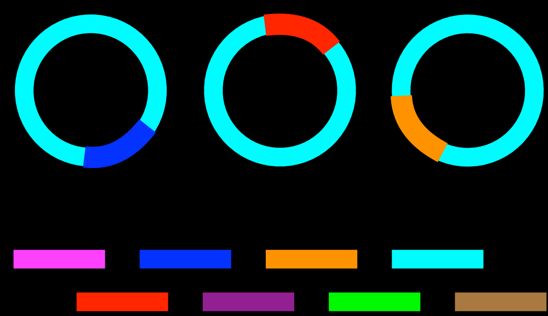

Environment
Bacteria
Genomic DNA
Random Target
Targeted
Metagenomics
16S rRNA
100M reads/sample

Microbial "Species" - ???; can be defined at OTU level, >90% AAI or >95% ANI at whole genome level

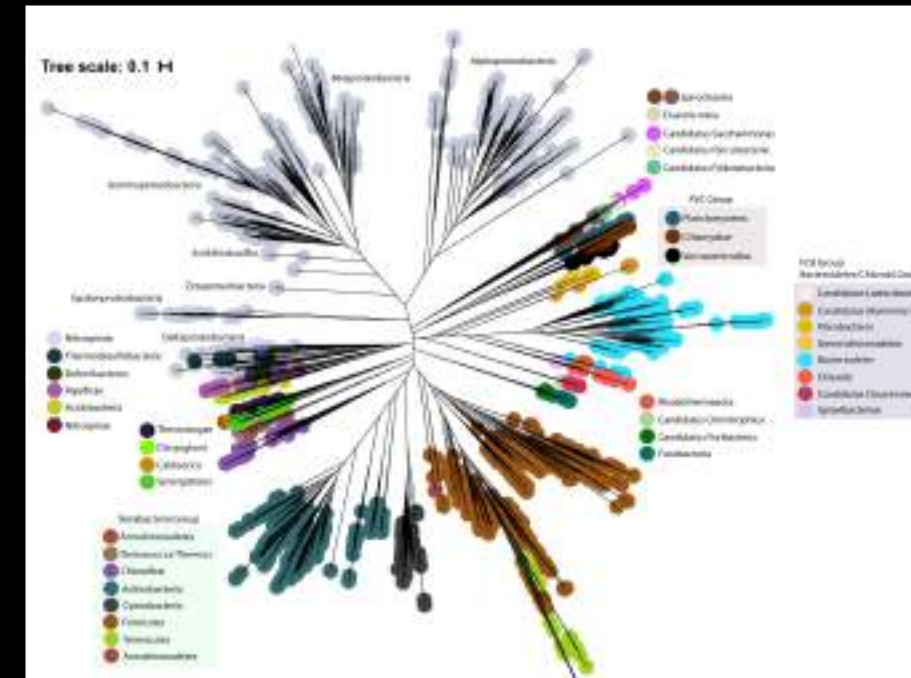
Microdiversity - closely-related (>97% similar 16S rRNA; same OTU) with distinct sub-taxonomic groups

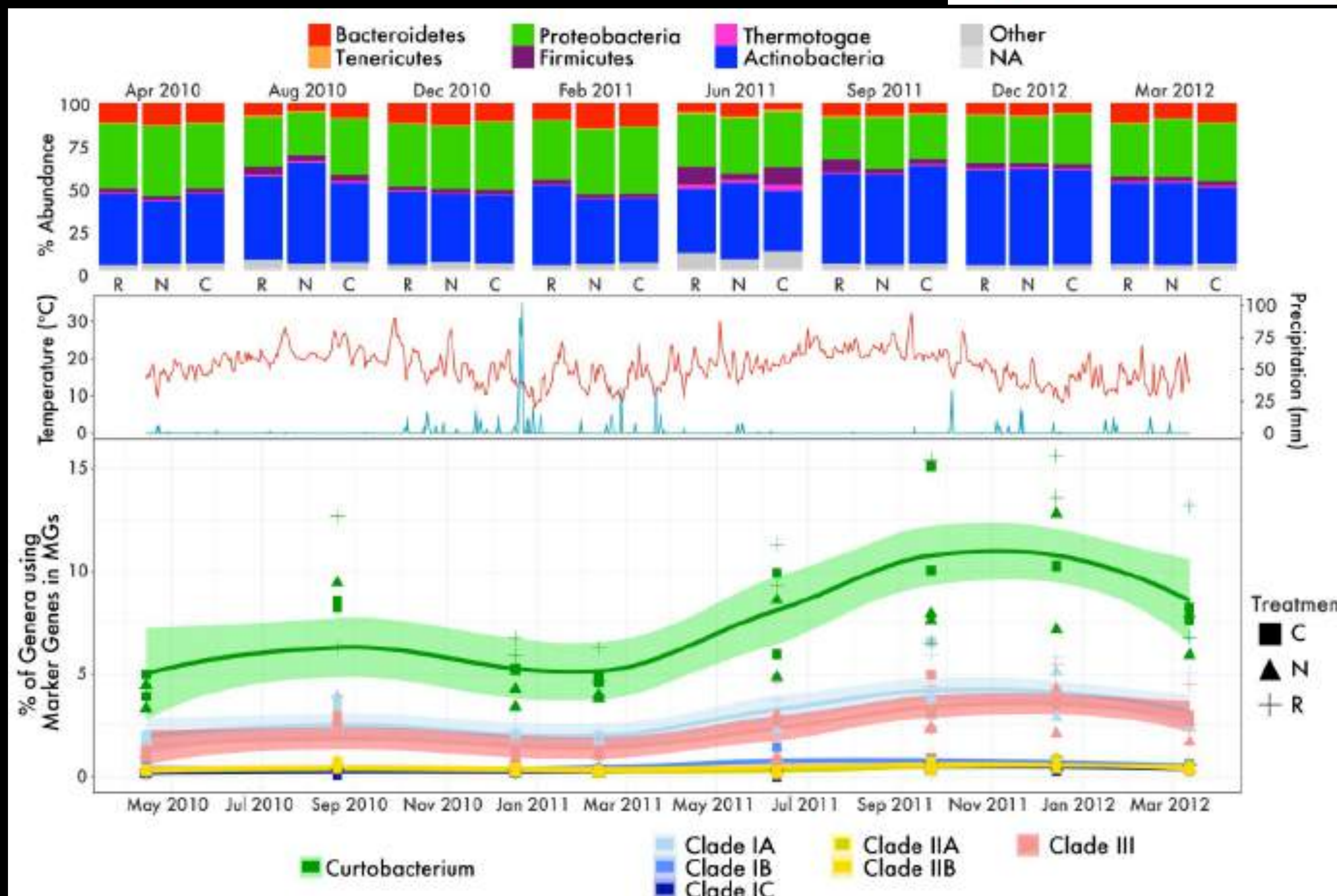


Nayfach et al. Genome Research. 2016

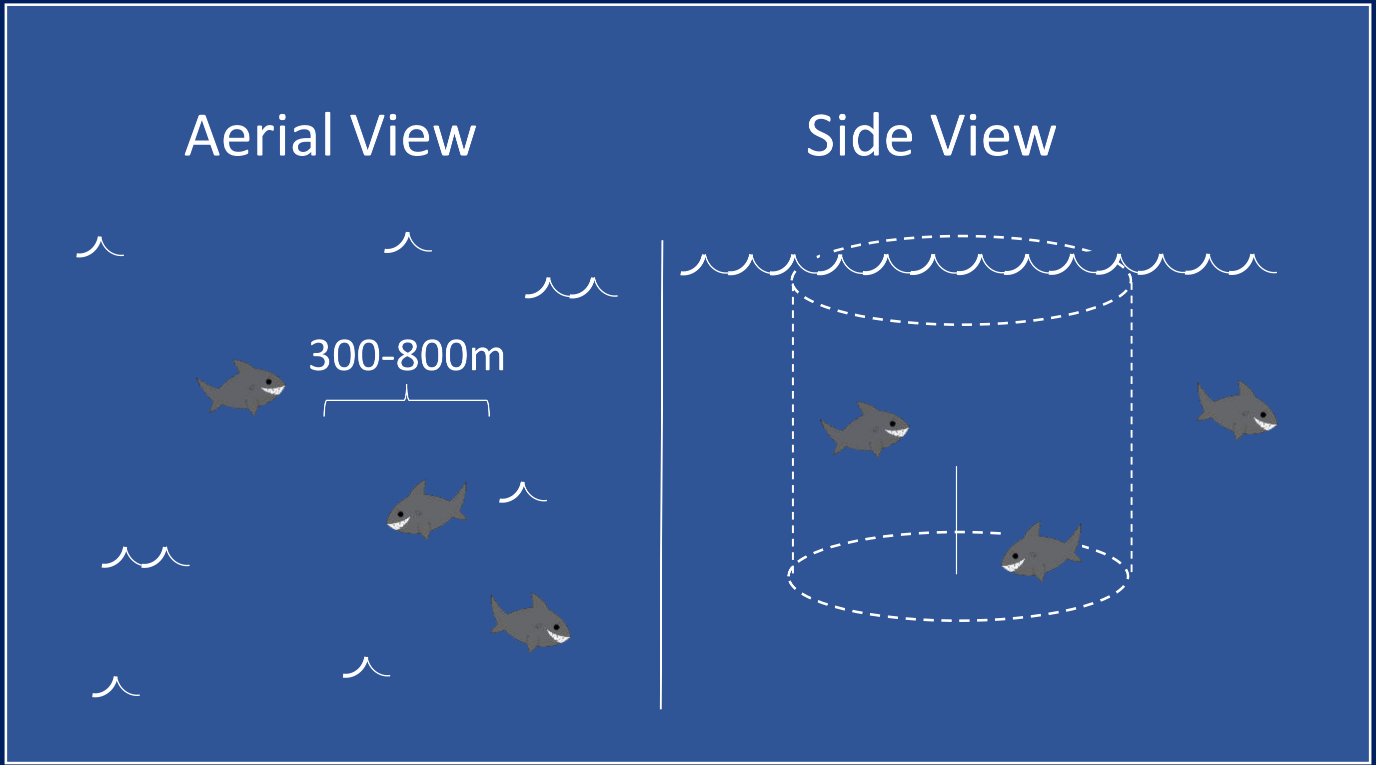


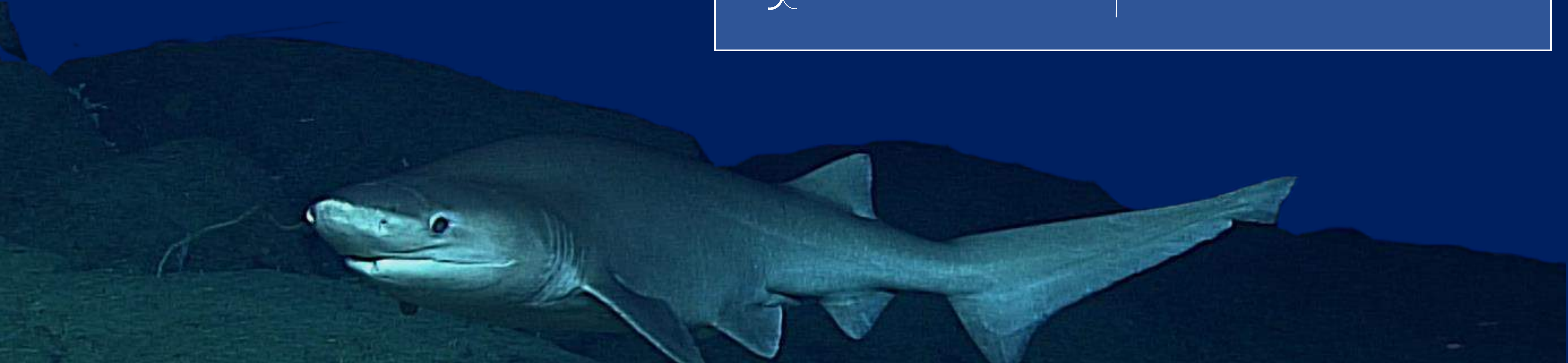
Metagenomics



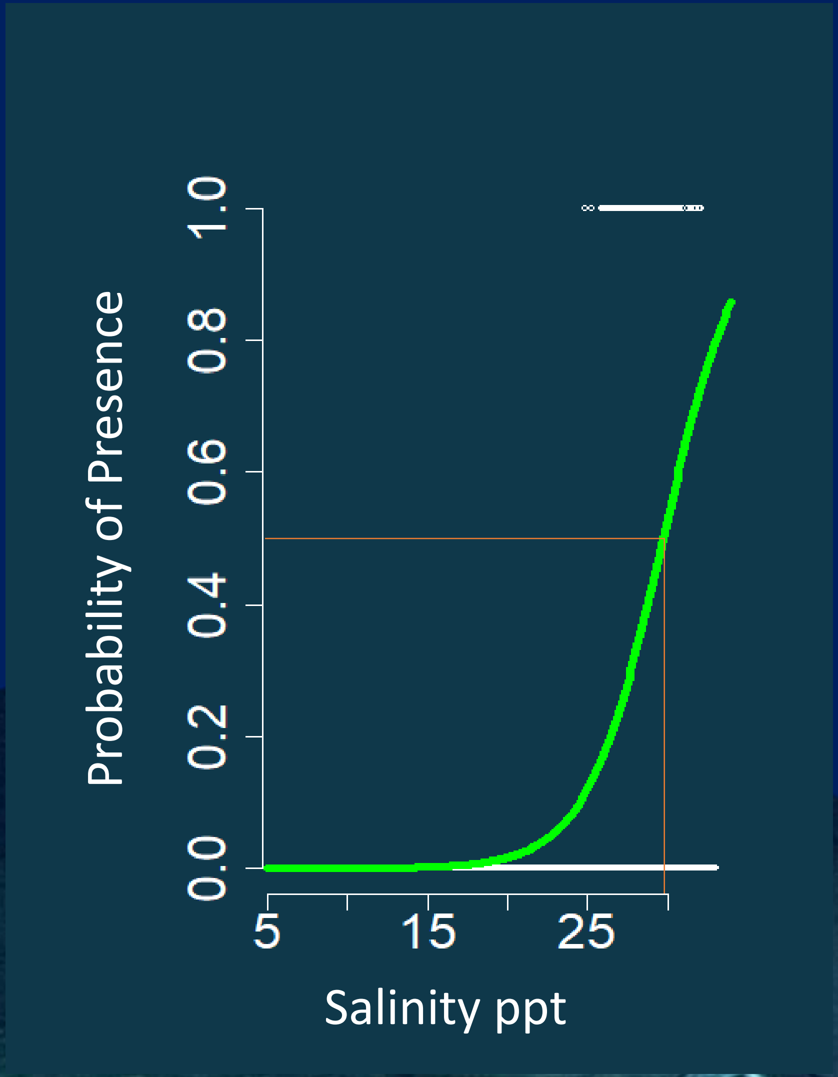


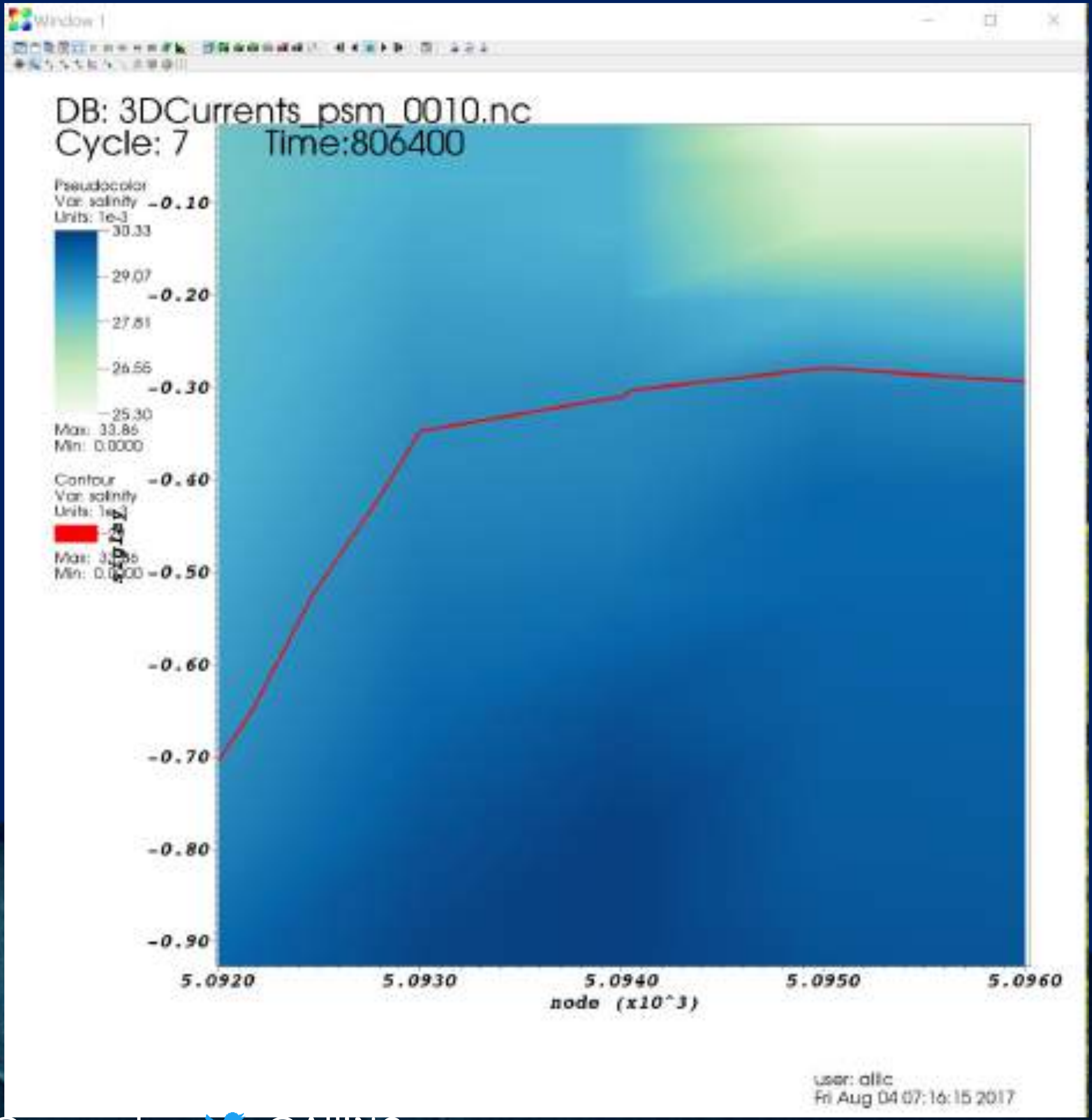
Chase et al. mBio. submitted





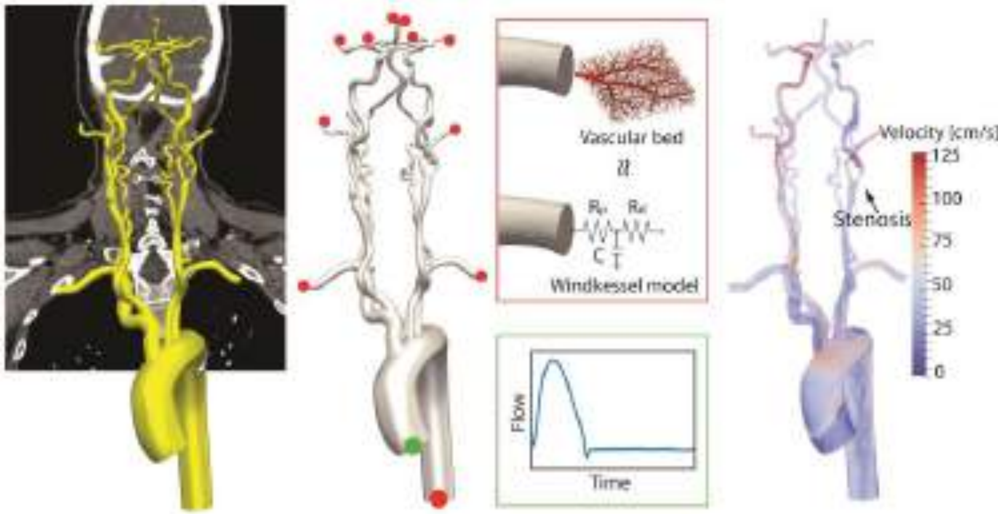
Alli N. Cramer, Washington State University, allison.cramer@wsu.edu, @AlliNCramer





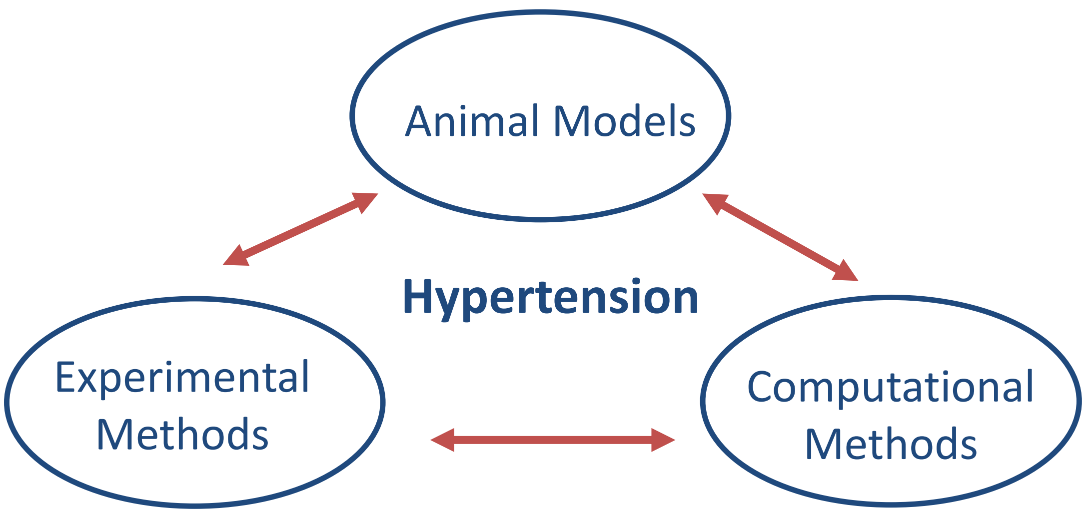
Alli N. Cramer, Washington State University, allison.cramer@wsu.edu, @AlliNCramer

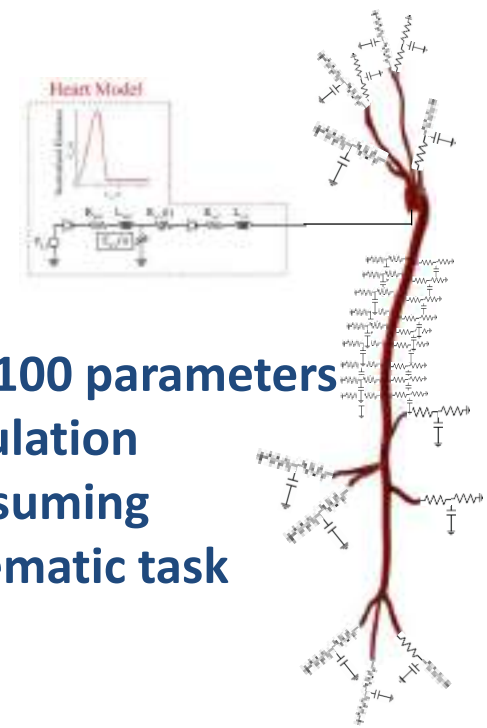
Image based blood flow simulation



Applications:

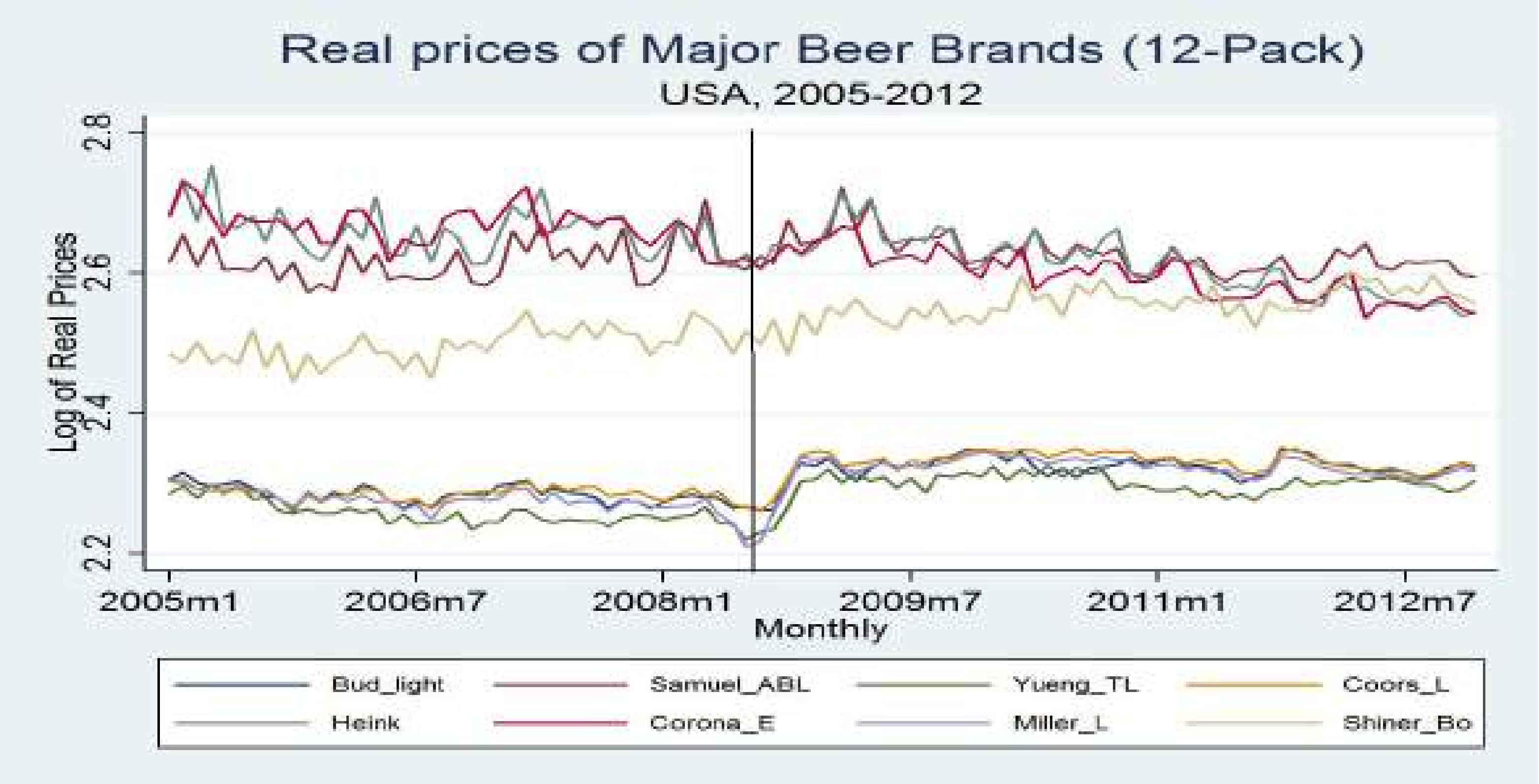
- Disease research
- Medical device design and evaluation
- Surgical planning





- **Estimate 100 parameters each simulation**
- **Time consuming**
- **Non systematic task**

Does a recent merger leads to collusive behavior in the US beer market?



Approach………HPC!!!

- Date sets: Retail scanner, Demographic, brand features data
- Estimation of (RCLM) demand/Supply models to find the best fit.
- Optimization techniques require solving hessian matrix (numerically) –
- Result – Reject collusive behavior – support brewers take all and control prices
- Way forward – Nielson Data set

KANSAS STATE UNIVERSITY

My research: Biology – related

Type of computing:

- Thousands of gigabytes of sequencing/text files downloaded from databases
- Lots of intermediate steps
- Mostly parallel until later stages
- Associations studies (logistic regression)
- Image classification

Huan Fan

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

What I've learnt

- San Diego is a very nice place
- Sea lions are more common than seals
- Singularity
- Performance of code and optimization
- Python3
  - Compiler: numba
  - Outofcore problem: dask, joblib, concurrant.future
  - Spark: big record data
  - Never leave your jupyter notebook
- Image classification: scikit-learn + deep learning

BioComplex
bio-inspired computing and
complex networks lab

Social Networks and Social Influence

Image Source: CASIS https://casis.llnl.gov/

Social Influence in organizations

Why the need for supercomputing?

- Social influence algorithms have high complexity but can run in parallel
- Availability of large scale data of social connectivity
- Analysis of Social Influence over time requires dealing with even larger datasets

References:

Albert, Réka, and Albert-László Barabási. "Statistical mechanics of complex networks." *Reviews of modern physics* 74.1 (2002): 47.
Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, *98*(2), 404-409.
Newman, Mark EJ, and Juyong Park. "Why social networks are different from other types of networks." *Physical Review E* 68.3 (2003): 036122.
Wasserman, Stanley, and Katherine Faust. *Social network analysis: Methods and applications*. Vol. 8. Cambridge university press, 1994.

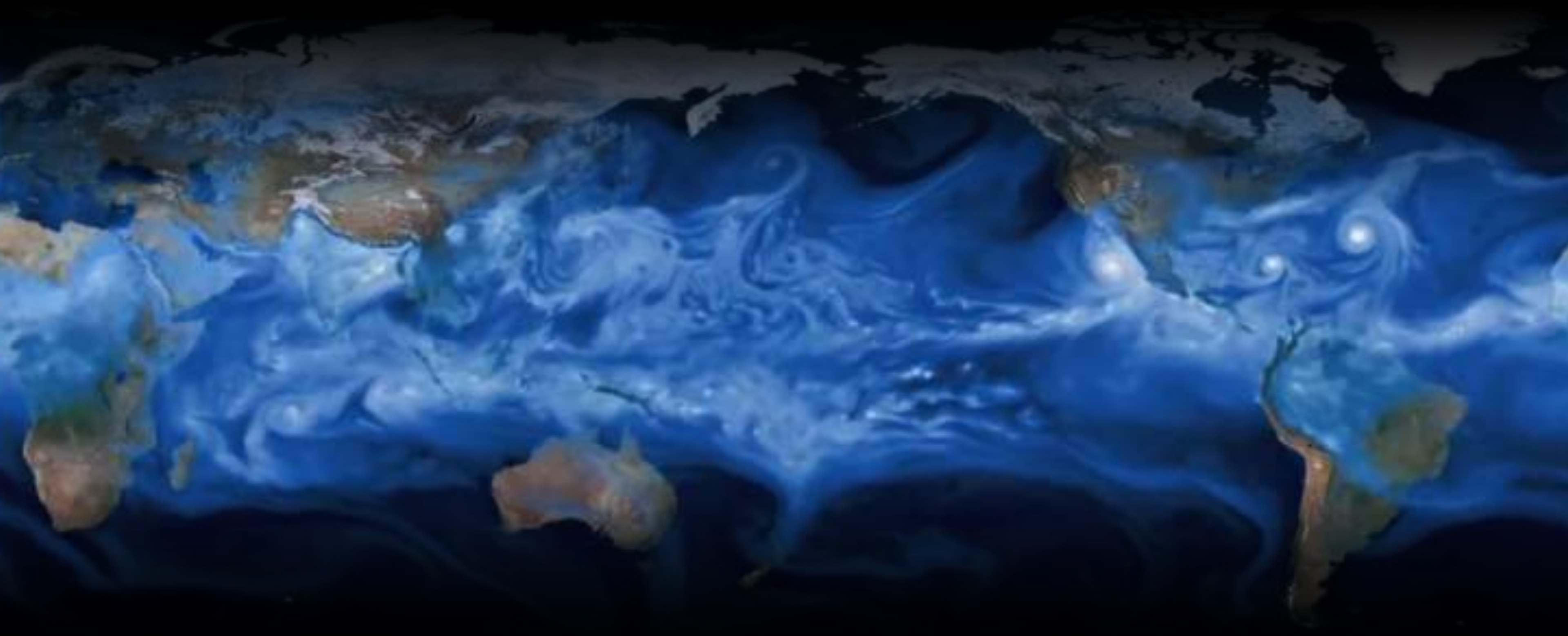
@josemarfaustino cruzj2012@my.fit.edu

Benjamin Fildier
UCB

Full movie at https://www.youtube.com/watch?v=cNyftYdjt-Q

Credentials: Michael Wehner, LBNL
Resolution: 25km

Atmospheric content of "rainable" water



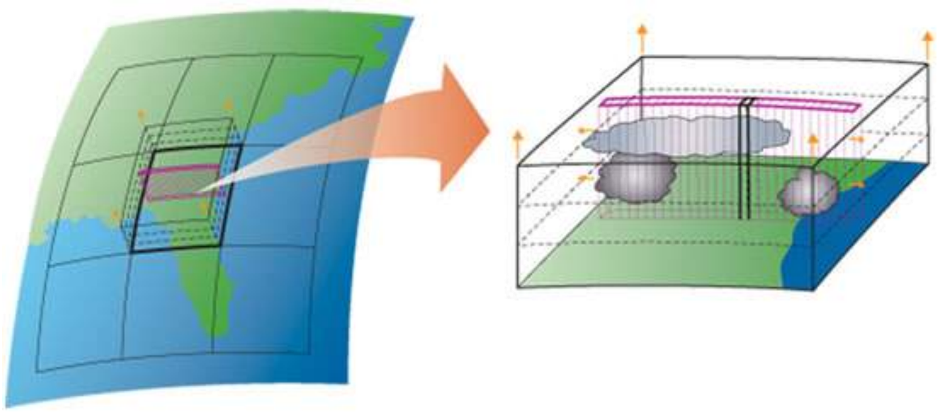
(Very) crude representation of rain and clouds in climate models

- Main source of uncertainty in climates projections
- Collateral damages on the energy balance, temperature changes, etc.
- Misunderstood dynamics and future impacts of heavy rainfall events

Latest project: rainfall extremes in a multiscale modeling framework

Simulations (needed some notions of OpenMP&MPI and optimization)

Analysis (will definitely use VisIt, Python-dask or Spark, and workflows)



Each 3D variable on smaller grid at each time slice is ≈ 10MB.

To study the dynamics of extremes, I have to constantly operate on hourly-averaged outputs. Full dataset is currently **2TB** (1-year runs only).

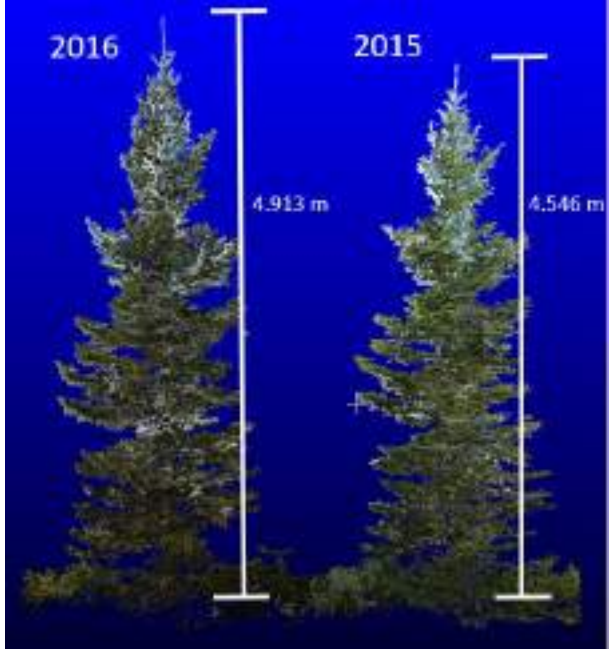
# of CPU*hours/simulated_year is $\mathcal{O}(10^4)$
(for Haswell nodes on Cori with 32 cores/node)
For 100 cores ≈ 4 days/simulated_year

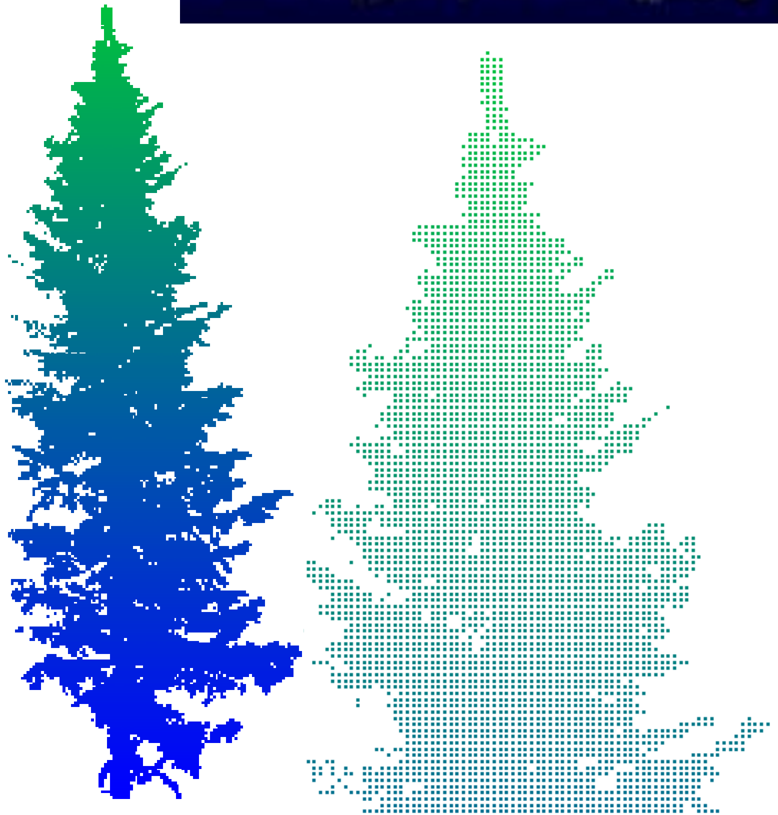
Now I want to upscale…
(finer time/space resolutions and longer runs)

$\mathcal{O}$(100TB)-$\mathcal{O}$(1PB) ?

TLS at SPRUCE

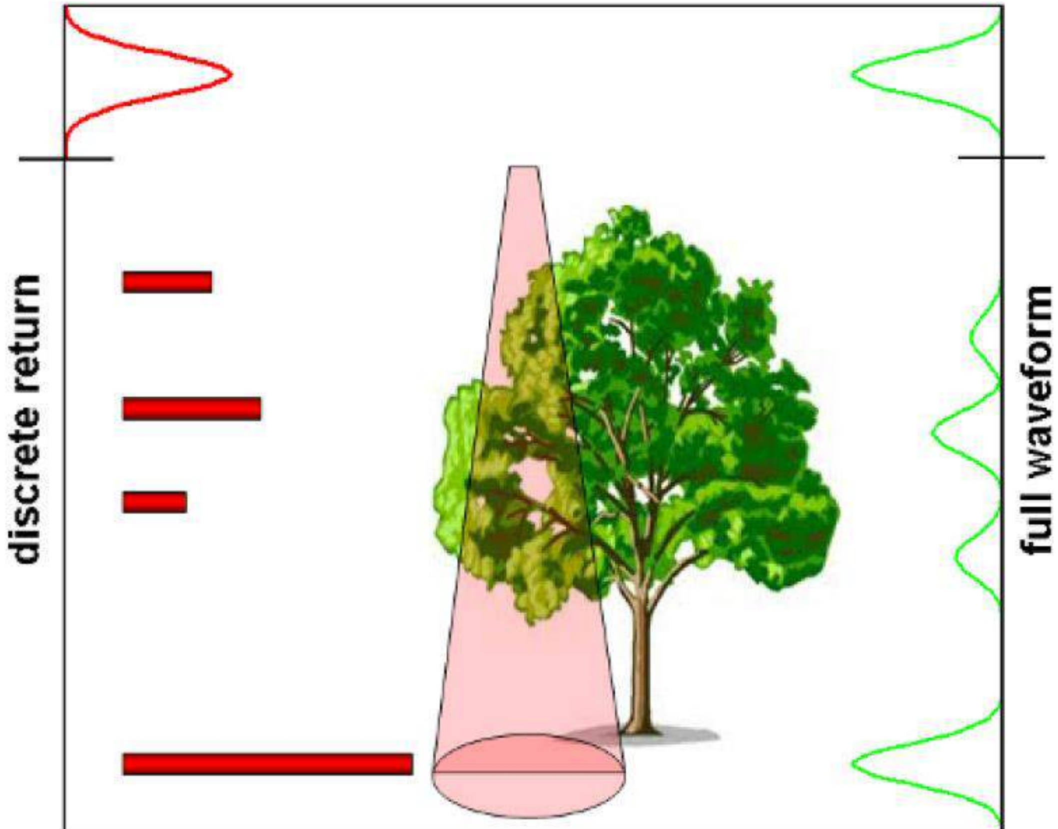
- Elevated CO_2 and gradient of elevated temperatures in chambers





Data Integration & Processing

- Ex. Reynolds Creek Experimental Watershed
- Hyperspectral imagery & full waveform lidar
- Terabytes of data to process



Full-3D Tomography of Central Mexico

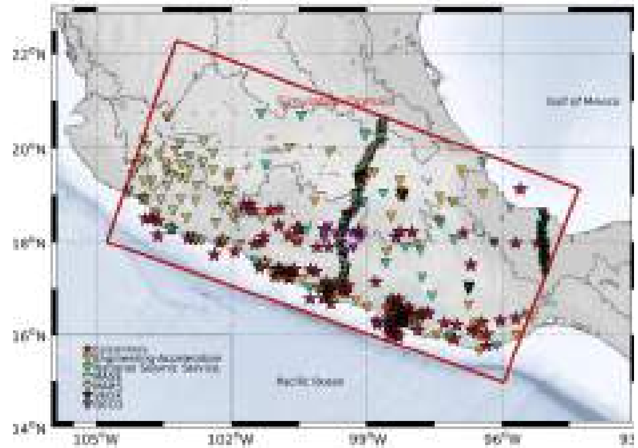
Alan Juarez*, Thomas Jordan*, and Leonardo Ramírez-Guzmán+

*University of Southern California, Los Angeles CA. +National Autonomous University of Mexico, Mexico City.

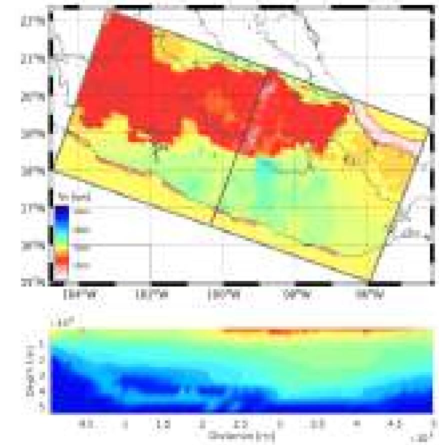
Database:

- 100 ($4.5<M_w<5.5$) earthquakes from 2005-2015
- 65 Green Functions Sources = 165 Sources
- 262 three-component velocity and acceleration stations (> 16,000 seismograms)

Simulations in Octree-based 3D Finite Element Method using the Hercules toolchain (Tu *et al.*, 2006) for elastic wave propagation modeling.



Database: Stations and epicenters.

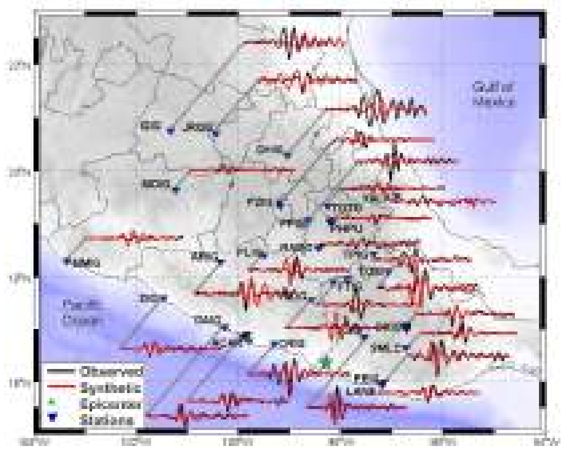


Vs velocity model.

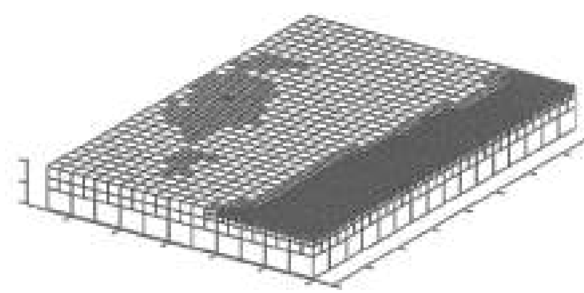
Computing (1 iteration):

- 330 simulations
- 0.25 Hz of resolution, 10 ppw
- 1000x500x75 km domain
- 36,441,778 elements/grid
- 84,480 pe units (CPU x hr)
- 90,338 x 9 nodes stored (deformations)
- 300s of simulations, 0.25 s sampling
- 3.8Tb of memory

Data processing and optimization problem.



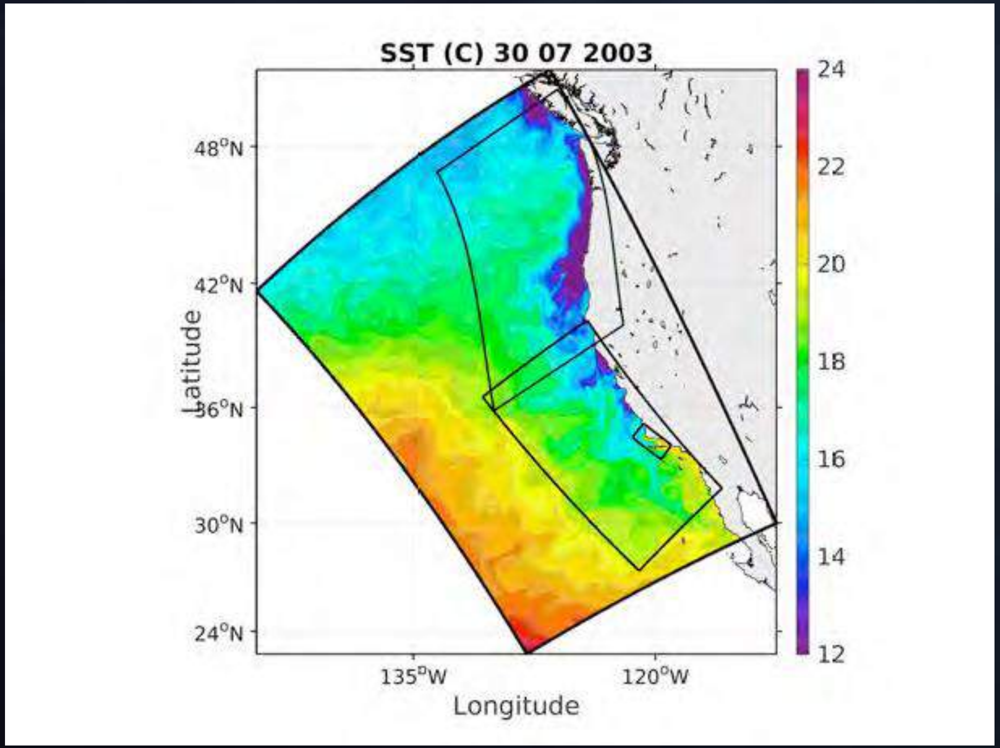
08/13/2014 - 06:47:30, Oaxaca, Mexico.
Mw 4.8, Lon -98.2 Lat 16.4 Depth 7



Octree mesh of central Mexico.

Faycal Kessouri, PhD

Field : Ocean Modeling – Physics & Biogeochemistry
Project : Modeling the Ocean US West Coast (Pacific Side) Acidification and Hypoxia
Affiliation (postdoc): University of California Los Angeles (LA)
Southern California Coast Water Research Project (Costa Mesa)



My simulations on Comet: > 30 000 SUs / year
Grid: 3D irregular sigma grid : 770 x 1440 x 60 cells
Frequency outputs: Daily

Objectives:

Management:
- Anthropogenic effect on acidification and hypoxia along the US Pacific coast
- Helping managers on getting decisions on the inputs

Science :
- From the grid downscaling: Sub mesoscale impact on ecological behavior of the plankton
- Ocean current-wind parametrization improves the representation of primary production and offshore matter transport

Methods:
ocean physical biogeochemical coupling modeling: Downscaling, refinement and improvement of the model

SDSC: SI2017

Makefile →
Workflow manager

Inputs (organization: open boundaries and atmospheric forcing, preparation of the parallel sub-domains)
Parallel jobs
Output joining + other operations …

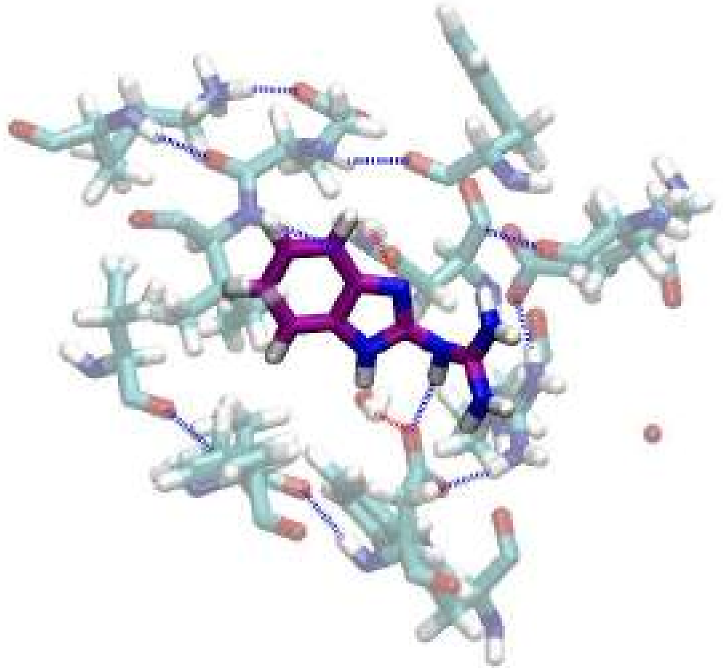
Machine learning class

Posttreatment of the outputs :
- organization, preparation of the outputs for analysis
- Validation of the model (using cruise data, moorings, satellite…)

GitHub workflows &
Visualization &
Globus transfer
classes

Transfer and sharing model outputs
Science applications (model analysis, visualization, statistics…)
Sharing scripts
Sharing visual outputs (movies, graphics)

The Hv1 proton channel has therapeutic potential



- There is much to discover on Hv1:
    - What is its structure?
    - How does it work?
    - Can we inhibit it selectively?
- Applying ideas from SDSC SI 2017 -- developing a workflow using Kepler for conducting **alchemical free energy perturbation calculations**

Predicting glass transition temperature from small-time simulation.

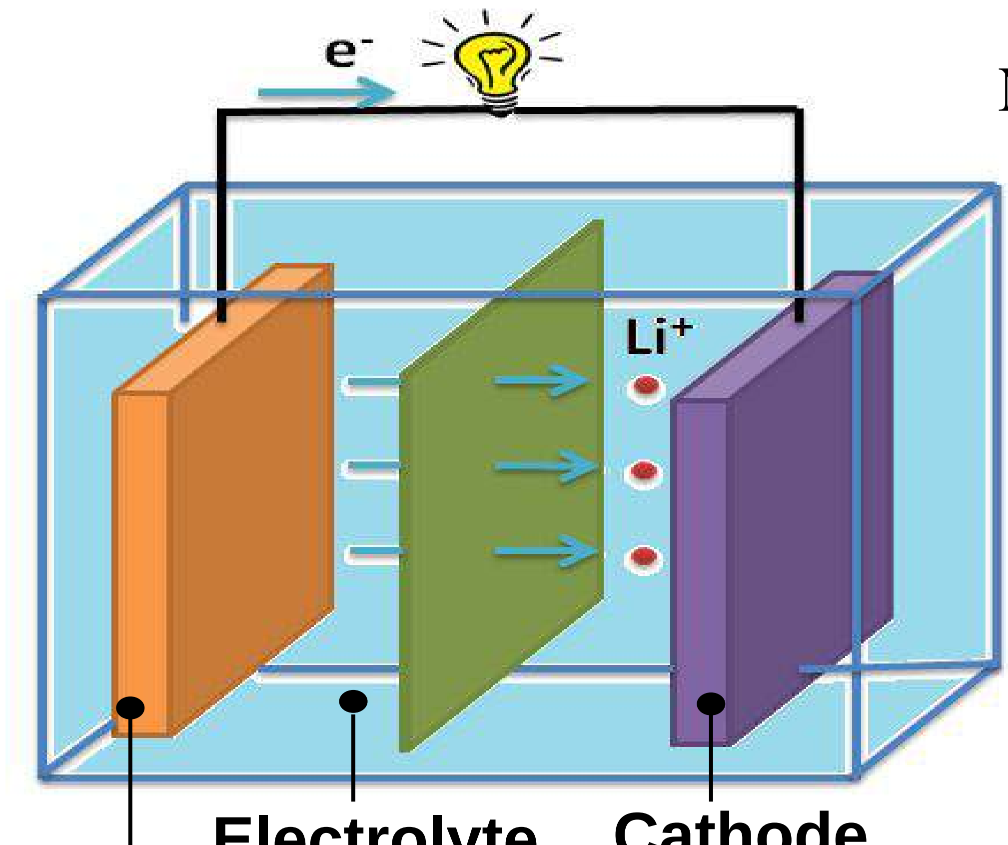
Zijun Lu

Department of Physics

04/08/2017

How I benefit from this workshop

- 1. Machine learning with Spark.
  Speed up the data analysis with Spark.
- 2. GPU use in both data analysis and simulation stage.
  Speed up my simulation and data analysis with GPU use.



Energy density is limited by the cathode material

New material & new chemistry

$$Fe^{2+}F_2 + Li^+ \rightarrow LiF + Fe^0$$

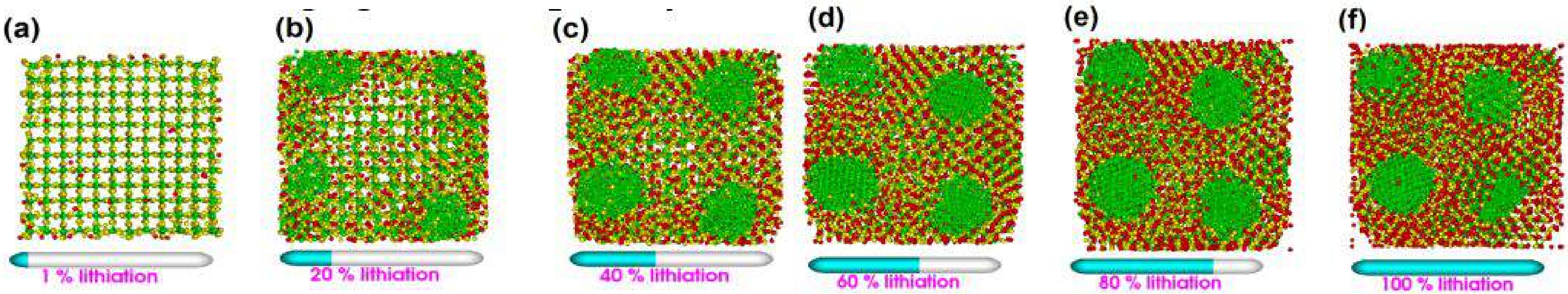
Structure

Charges

Atomic interactions

New structure

Computationall
y
very expensive!



Performance profiling & optimization

Introducing OpenMP/OpenACC

Maybe CUDA

GitHub version control

MSE 451 Computational Materials Science (4 crs)

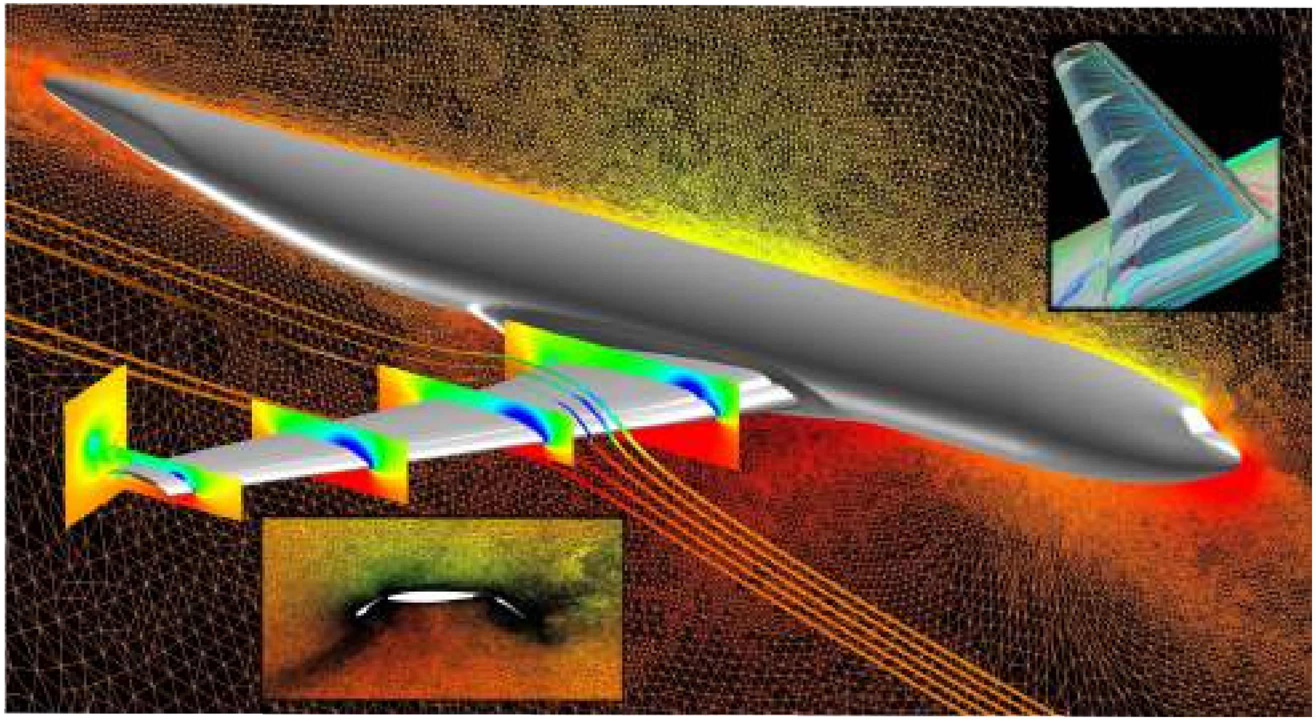
Prerequisite: MSE 350 or PHYS 333 or CHEM 434.

Theory and application of computational methods to model, understand and predict the behavior of materials. Labs provide hands-on experience in solving real materials problems using computational approaches.

Thank you SDSC!

CFD Laboratory

NASA Vision 2030



Large eddy simulation of aircraft envelope (billions of years using current methods)

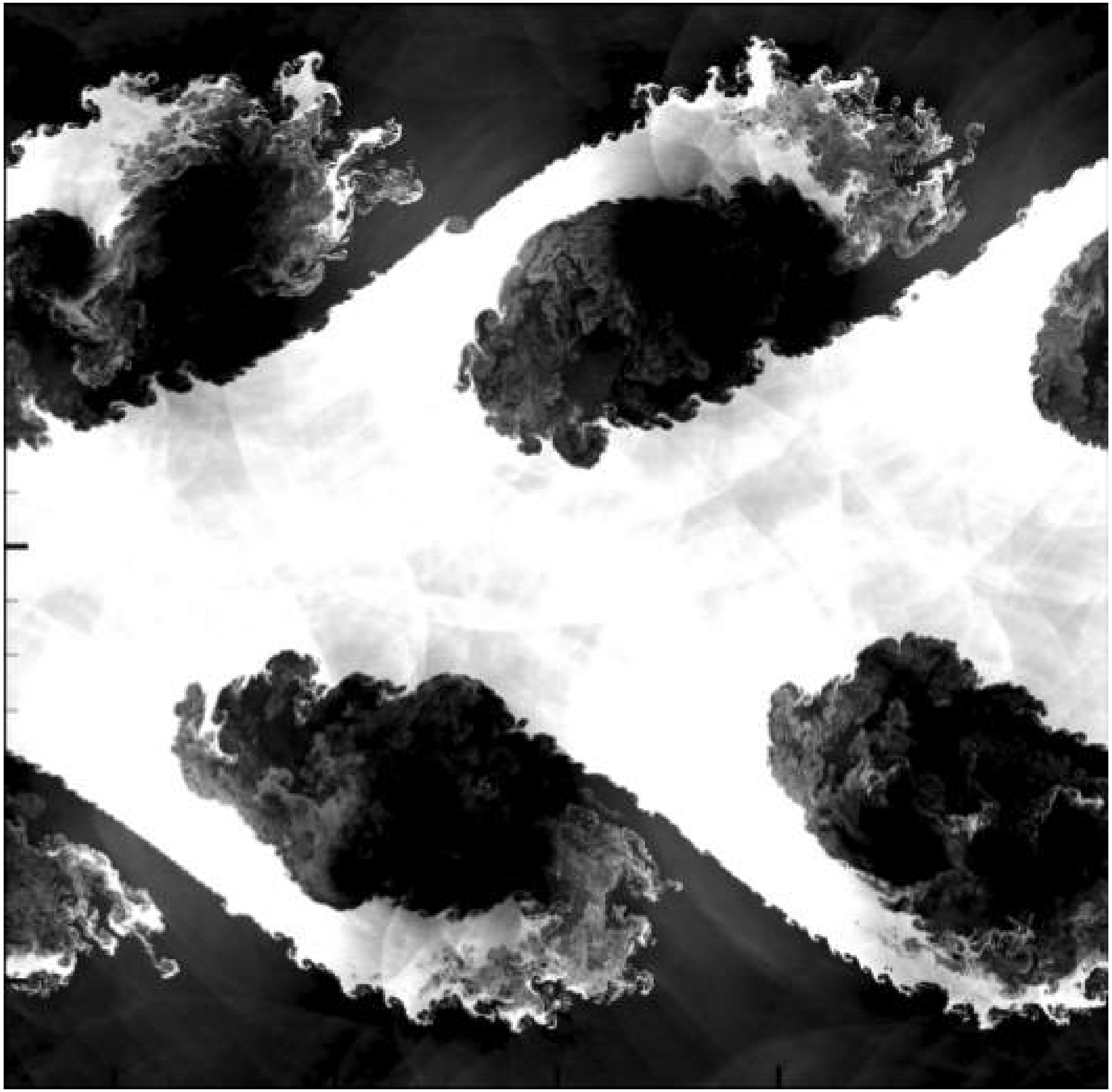
"Turbulence is the most important unsolved problem in classical physics"
– Richard Feynman

EVIL! (Needs modeling of higher spatial frequencies)

$$
\begin{array}{lll}
\dfrac{\partial \mathbf{u}}{\partial t} - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f}(\mathbf{x}, t) & \text{in} & Q = \Omega \times (0, T), \\
\nabla \cdot \mathbf{u} = 0 & \text{in} & \Omega \times (0, T), \\
\mathbf{u}(\mathbf{x}, t) = \mathbf{0} & \text{on} & \partial\Omega \times (0, T), \\
\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(x) & \text{in} & \Omega.
\end{array}
$$

Key focus areas

1. High Performance Computing
2. Numerical Algorithms
3. Knowledge Extraction



A ***simple*** flow (two-dimensional) required the solution of 273X4 million ODES (MPI-Fortran, 192 processes, 410 wall clock hours). One snapshot
of solution field (4 variables) - 21GB

Interdisciplinary Research

HPC – Primarily MPI
Signal Processing – Spectral, Wavelet, Compressed sensing
Image Processing – Deconvolution, Edge detection
Optimization – Krylov Subspace Methods (for Pressure Poisson Eq)
Data driven – PCA, Transformed space methods, inverse problems
ML – ANNs, Computer Vision

Kelvin-Helmholtz Instability – First studied in 1871
Observed in engineering, geophysical, astrophysical phenomena

*Showed **1810** results on Google Scholar for 2017!*

Energy Flow in the Diffusive Regime; relating contact dynamics to energy flow.

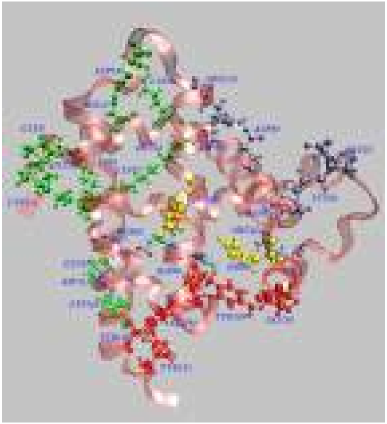


Figure: Long-range and long-lived, 150ps, residue-residue contacts.

Aims

1. Relate energy flow through contacts with contact dynamics by a master equation.
2. Investigate bound water's contribution to energy flow with emphasis on interface waters
3. Develop easily to distribute software to extend this analysis to multi-protein environments

How SDSC Summer Institute 2017 Will apply to my work.

- Utilize workflow management to streamline and create easily reproducible amber simulations.
- Implement dask everywhere! Time to get things done with multicore support!!!
- Implement CUDA programming in place of mpi code to accelerate energy flow calculations.
- Take advantage of the OpenMP framework and dive deeper into mpi.
- Most of all, use github so I can easily find changes and look at prior commits.

Bayesian Adaptive Design for Clinical Trials

- Large cardiovascular outcome trials (CVOTs) are commonly used in the evaluation of cardiovascular risk for new therapeutic agents intended for the treatment of Type 2 diabetes mellitus per US FDA guidance.
- These trials are large in size and can take years to complete.
- Would like to use information from already completed CVOTs in designing future CVOTs to provide a reasonable means to decrease the number of subjects required and/or time to complete future CVOTs.
- Such an approach requires halting the new CVOT and performing an interim analysis of whether the study needs to be continued or can be stopped.
- Determining when to halt and perform the interim analysis is a parameter that is calibrated using simulated datasets.

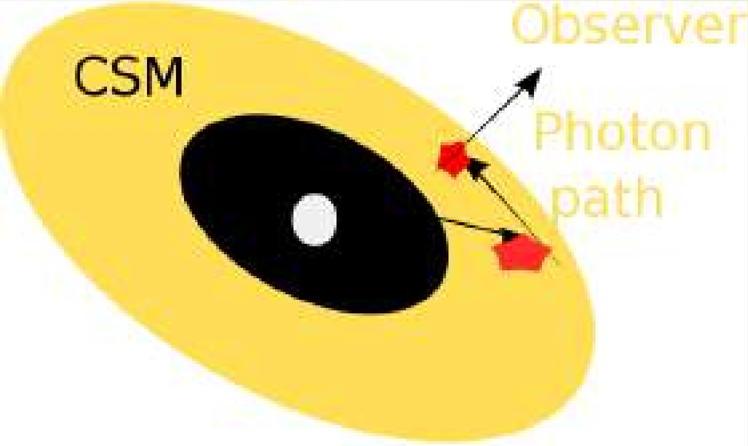
Computational aspect

- A simulated dataset is hundreds of thousands of simulated observations of patient data.
- Patient data includes things like an enrollement time, treatment indicator, covariates, etc.
- Simulating an observation of such patient data is independent of simulating any other observation and so can be easily parallelized.

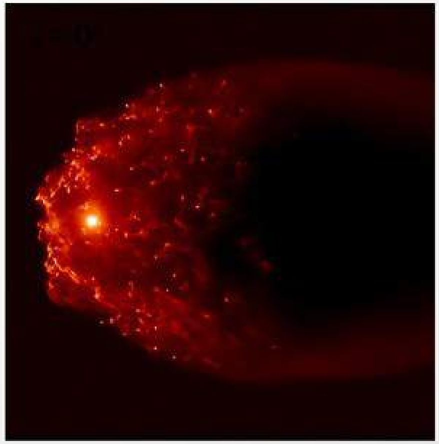
Supernovae LIne Polarization (*SLIP*)

—Manisha Shrestha

- Monte Carlo based radiative transfer code
- We get polarized flux as output.
- Need large number of photons for good signal to noise ratio.
- Going from analytic to SPH is computationally expensive.



Schematic of *SLIP*



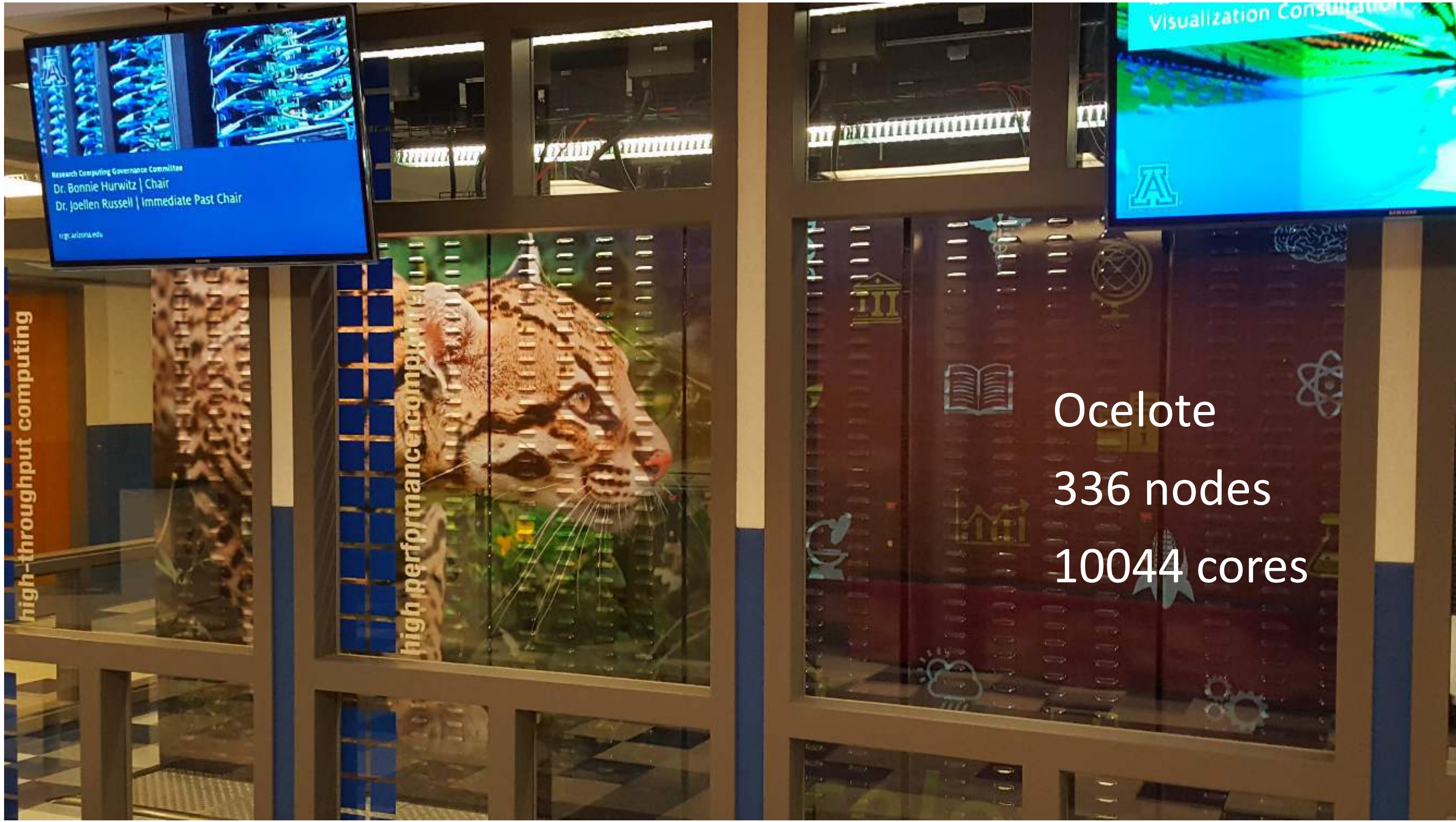
Density distribution from SPH

Dima Shyshlov

HPC Consultant
University of Arizona

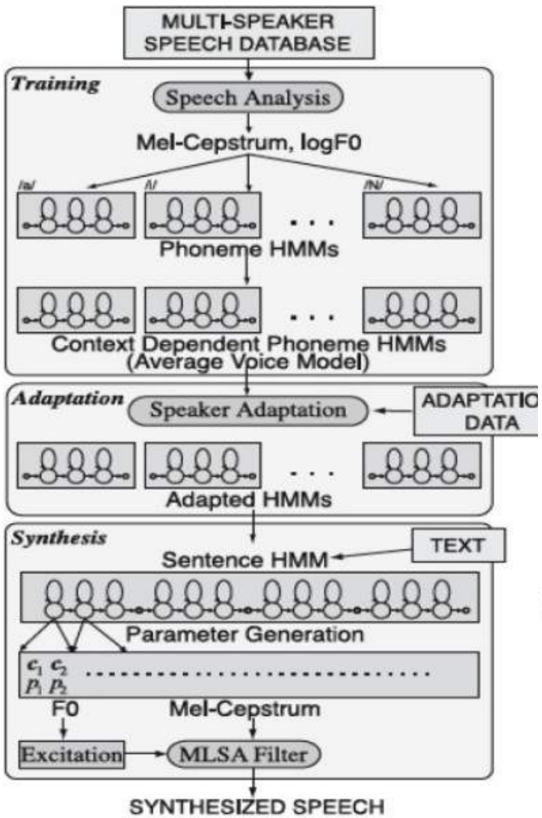
Trends in UA HPC:

- Interest in GPU’s
- Machine Learning
- Singularity (TensorFlow)
- Workshops

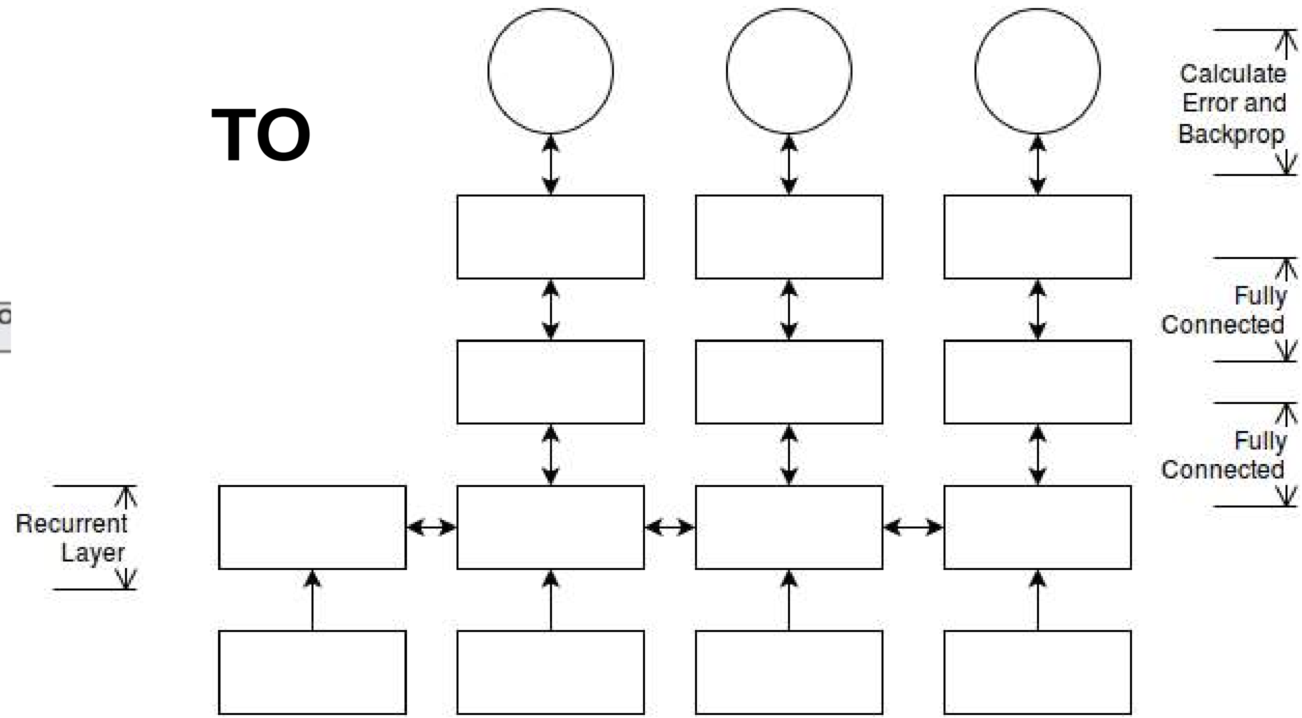


Towards Hyper Efficient End-to-End RNN Speech Synthesis with Speaker Adaptation

Research by Marcelo Siero



TO

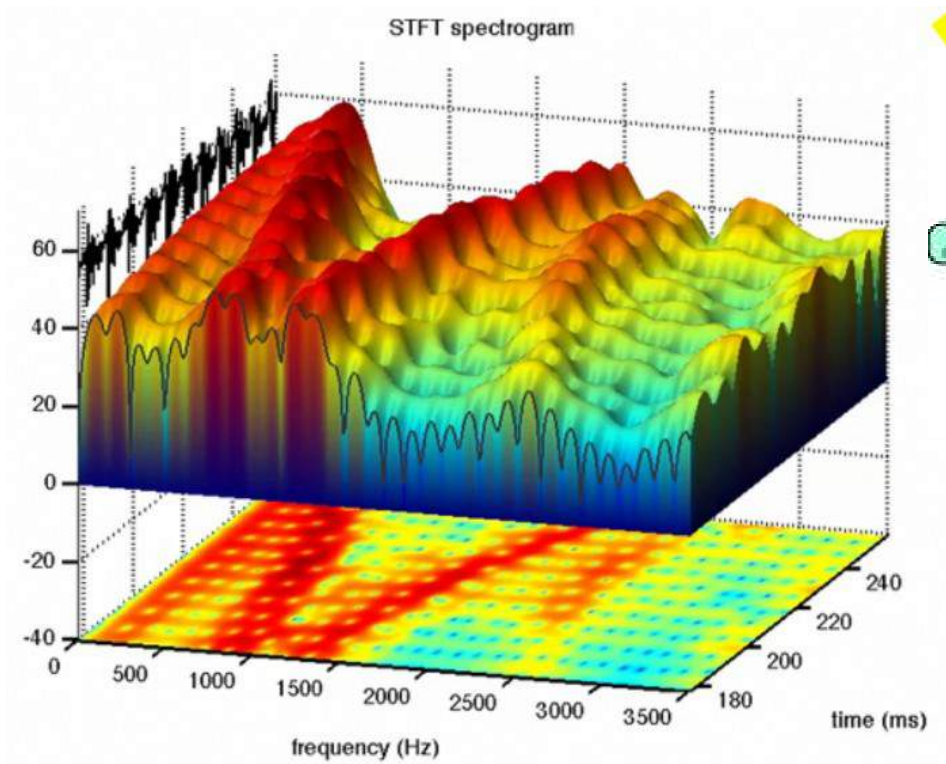


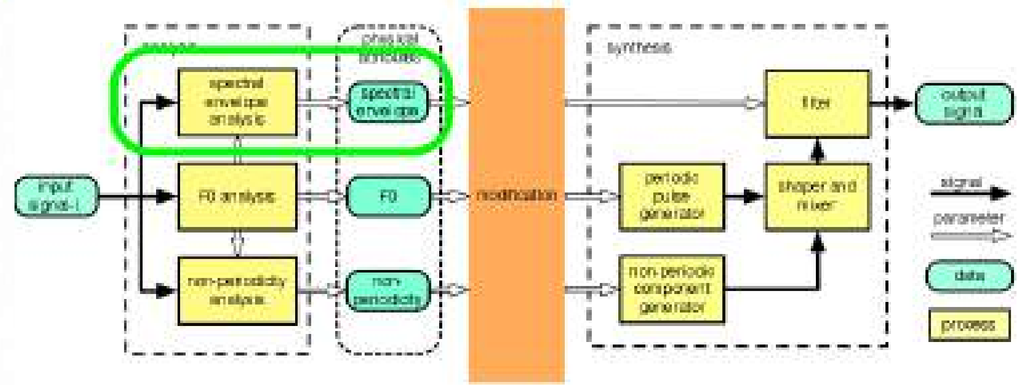
P100 GPU Comet nodes could allow fast training on many hours of training speech corpus by providing paralellism while using cuFFT, cuDNN libraries and more to convert HMMs technology to end-to-end RNNs. Use of Singularity containers may aid the task. On the right is a diagram of HTS, and HMM-based synthesizer.

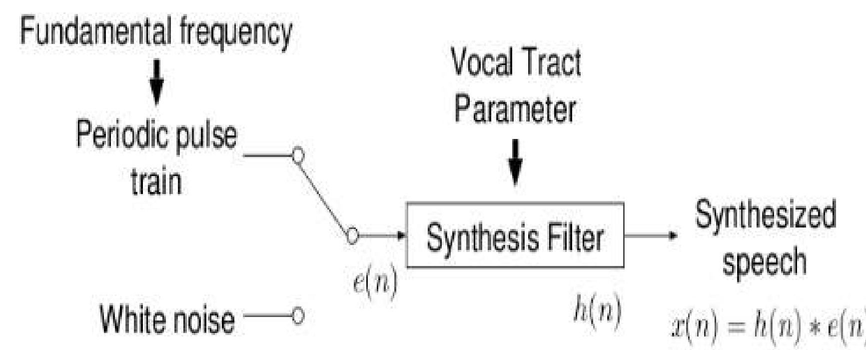
See Demo of Speech Conversion and Speaker Adaptation at:
http://FrankenThespian.com/demo

Towards Hyper Efficient End to End RNN Speech Synthesis with Speaker Adaptation

Research by Marcelo Siero







Better Visualization using **Visit** might create charts like the one on the left. Above right is a visualization describing the use of the STRAIGHT vocoder (H. Kawahara). It clearly shows quantization noise which the STRAIGHT vocoder is known to remove (smooth out). The right shows a block diagram of the STRAIGHT vocoder and the more simplified source / vocal tract filter model.

Acceleration of Scientific Computing (collaborate with UCI Earth Science Dept.)

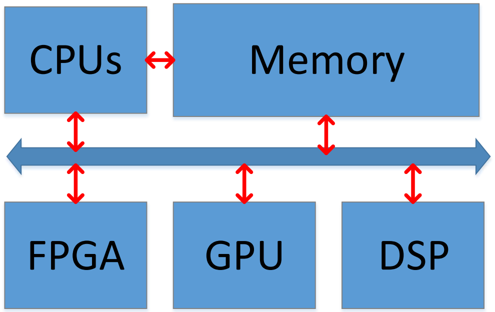
- Application
  - Climate Model in Weather Prediction (FastJ)
    - Radiative Transfer Equation
  - …
- Performance Optimization
  - MPI, OpenMP, …
  - Cache, loop-level optimizations, …
  - GPU computing
    - CUDA, OpenACC programming

Know more from SDSC sessions

Hsin-Yu Ting, Ph.D. student
Advisor: Eli Bozorgzadeh

Acceleration of Scientific Computing (collaborate with UCI Earth Science Dept.)

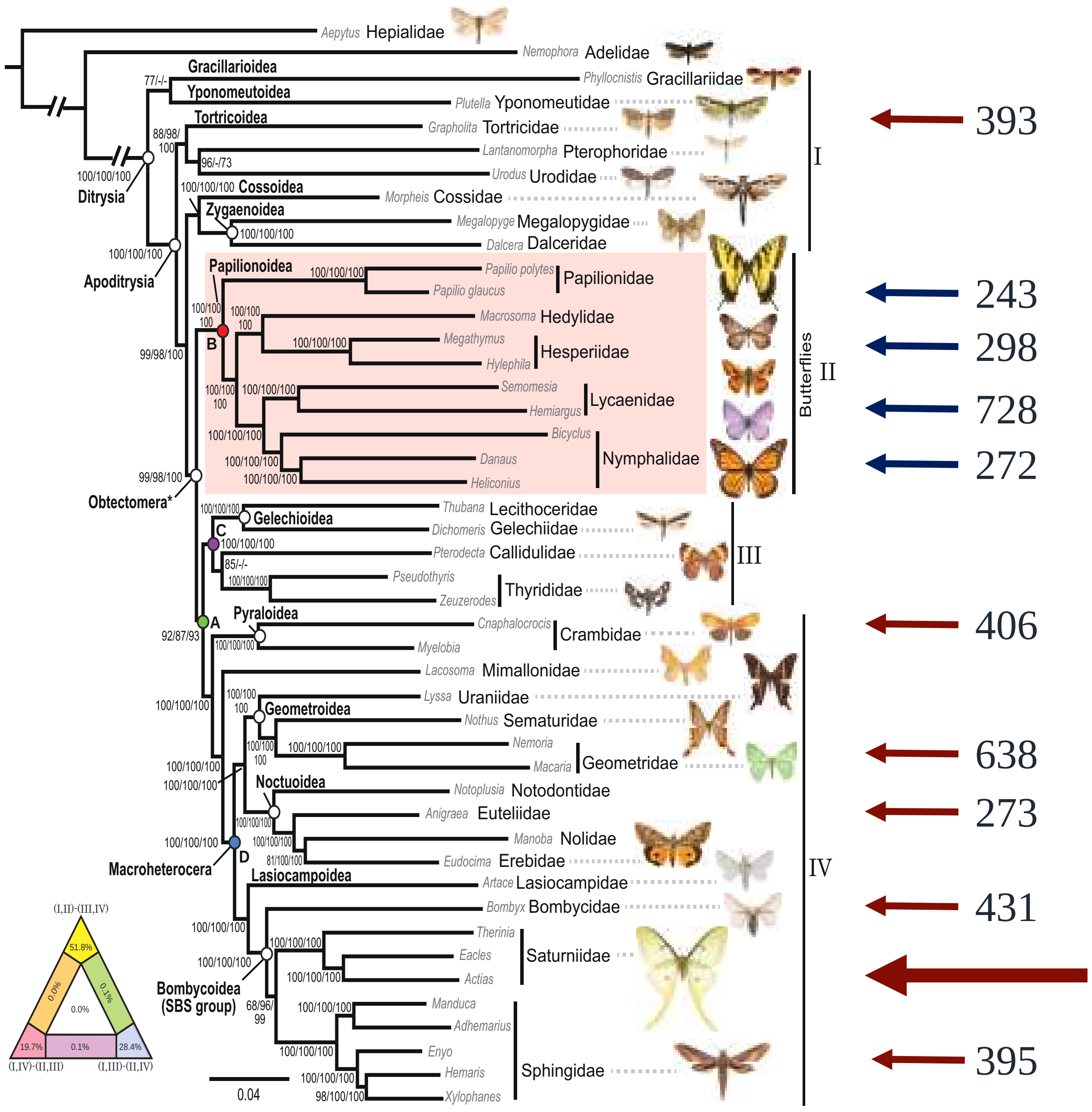
- Heterogeneous computing
  - Central processing unit (CPU)
  - Graphics Processing Unit (GPU)
  - Field-Programmable Gate Array (FPGA)
  - Digital Signal Processor (DSP)



- Design space exploration
  - Tons of combinations
    - Memory hierarchy/bandwidth
    - Resource utilization/power/energy consumption/ performance
  - Adopt machine learning to find the optimal solution

Thx. SDSC

Hsin-Yu Ting, Ph.D. student
Advisor: Eli Bozorgzadeh



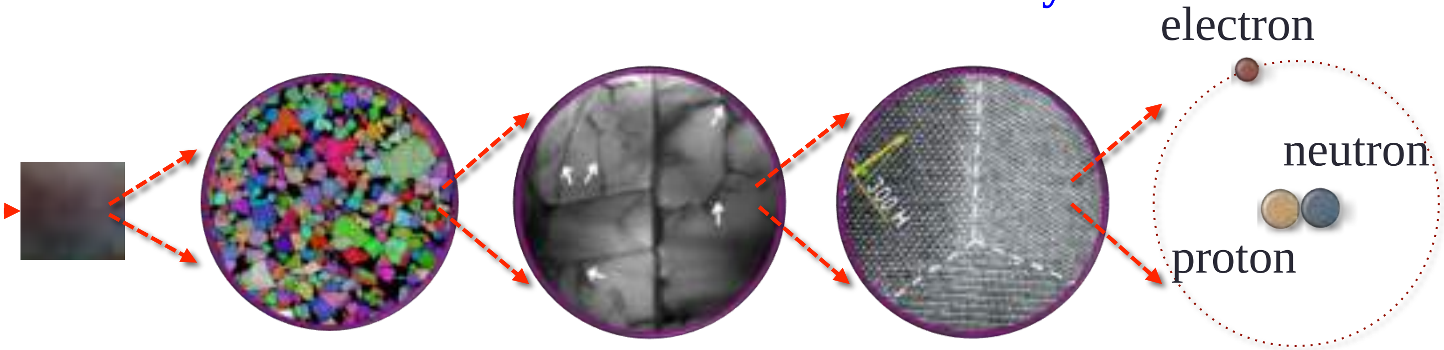
Lepidoptera Genomes

- ~25 assemblies (LepBase, NCBI)
  - 12 families; 8 moth species
  - 275 - 725 Mb

Kawahara and Breinholt (2014): *Proc. R. Soc. B*

Triton Statue @ UCSD

Materials with Structural Hierarchy



Macroscale (10^{-3} m) | Microscale (10^{-6} m) | Nanoscale (10^{-9} m) | Atomistic (10^{-10} m) | Electronic (10^{-15} m)

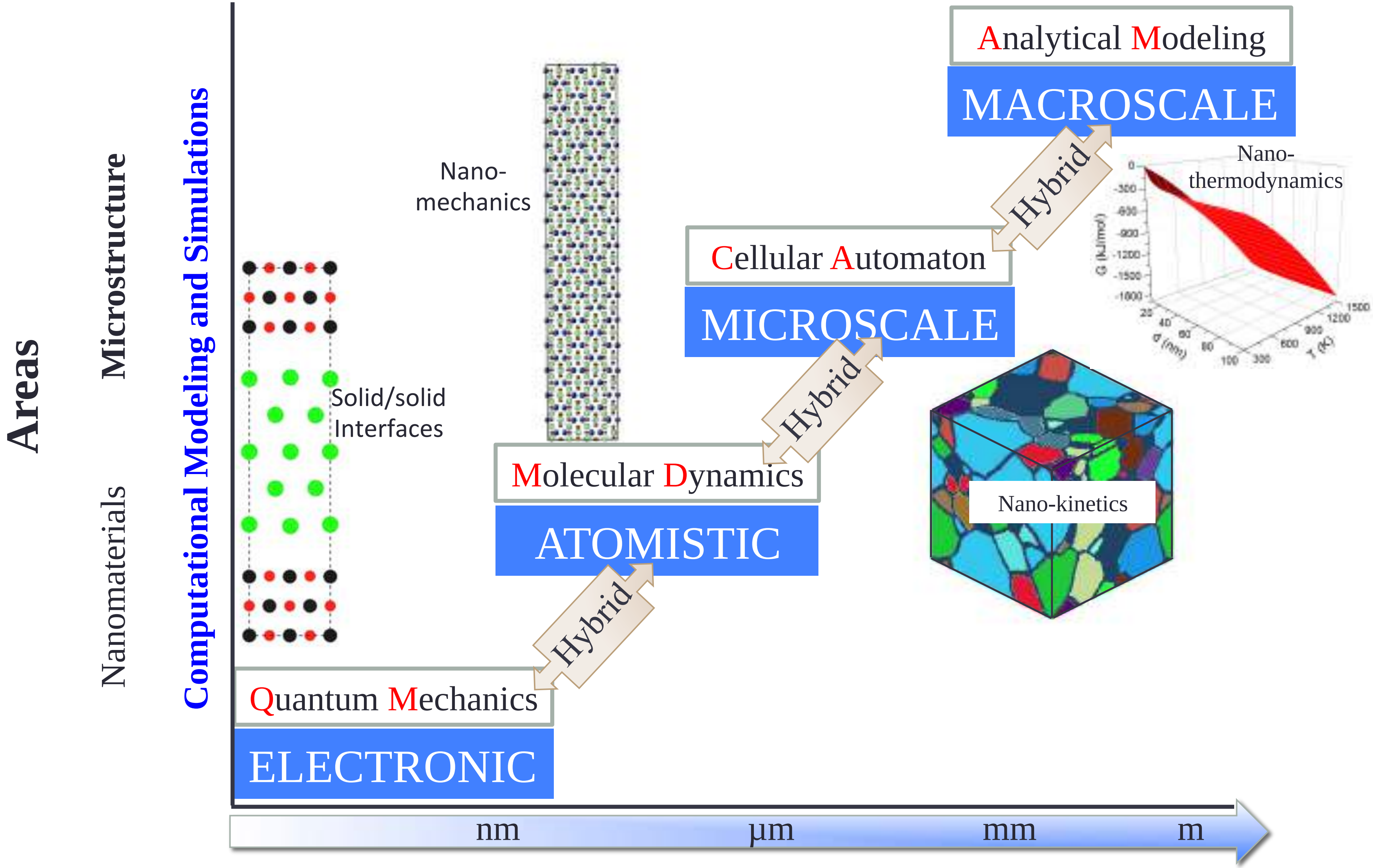
Experiments

www.123rf.com

Computer Modeling

www.unc.edu

Understanding the effects of hierarchical structure can guide the synthesis of new materials tailored for specific applications.



Multiscale Modeling of Materials using High Performance Computing

The M^3 group: www.m3sdsu.com

How many clinics at least does a study have to recruit in oder to detect a treatment effect with controlled level of error statistical significance?

Approximate the Distribution of Test Statistic with Simulation

Parallelization! Performance Optimization, OpenMP, MPI, Spark ...