Emerging Themes in Data Intensive Sciences

Galen Shipman
Data Systems Architect
CSMD & OLCF

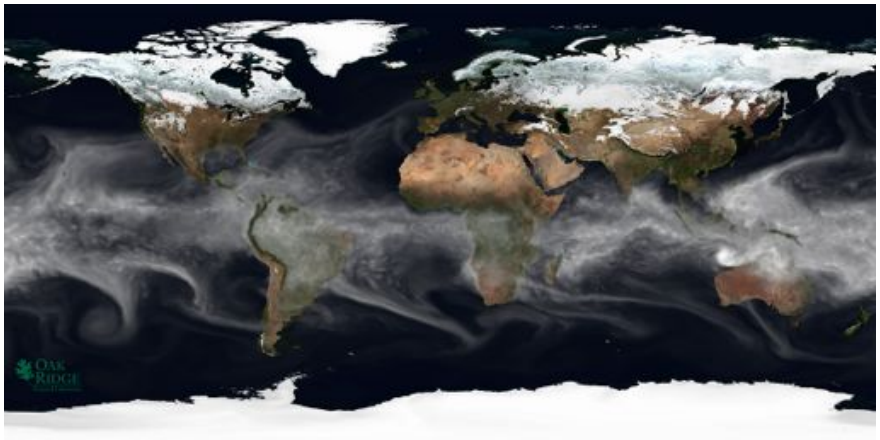
Director
Compute and Data Environment for Science

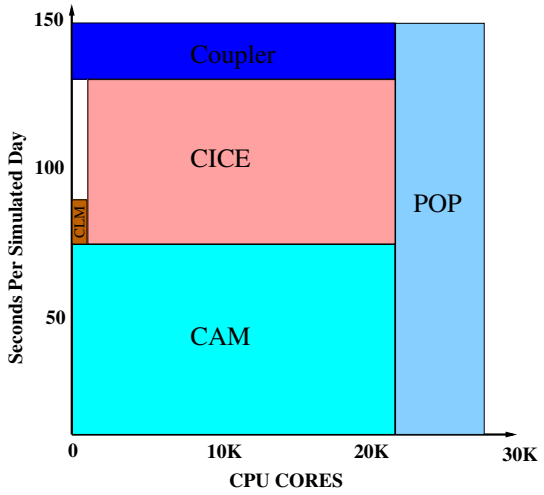
OLCF User Meeting
Wednesday, July 23, 2014

Data Science and Scientific Discovery

- The rate of scientific progress is increasingly dependent on the ability to efficiently capture, integrate, analyze, and steward large volumes of diverse data
- Increasing data volume, variety, and velocity are creating a new environment for scientific discovery
- But many facilities and research programs across the Office of Science are not prepared for this challenge

Accelerated Climate Model for Energy

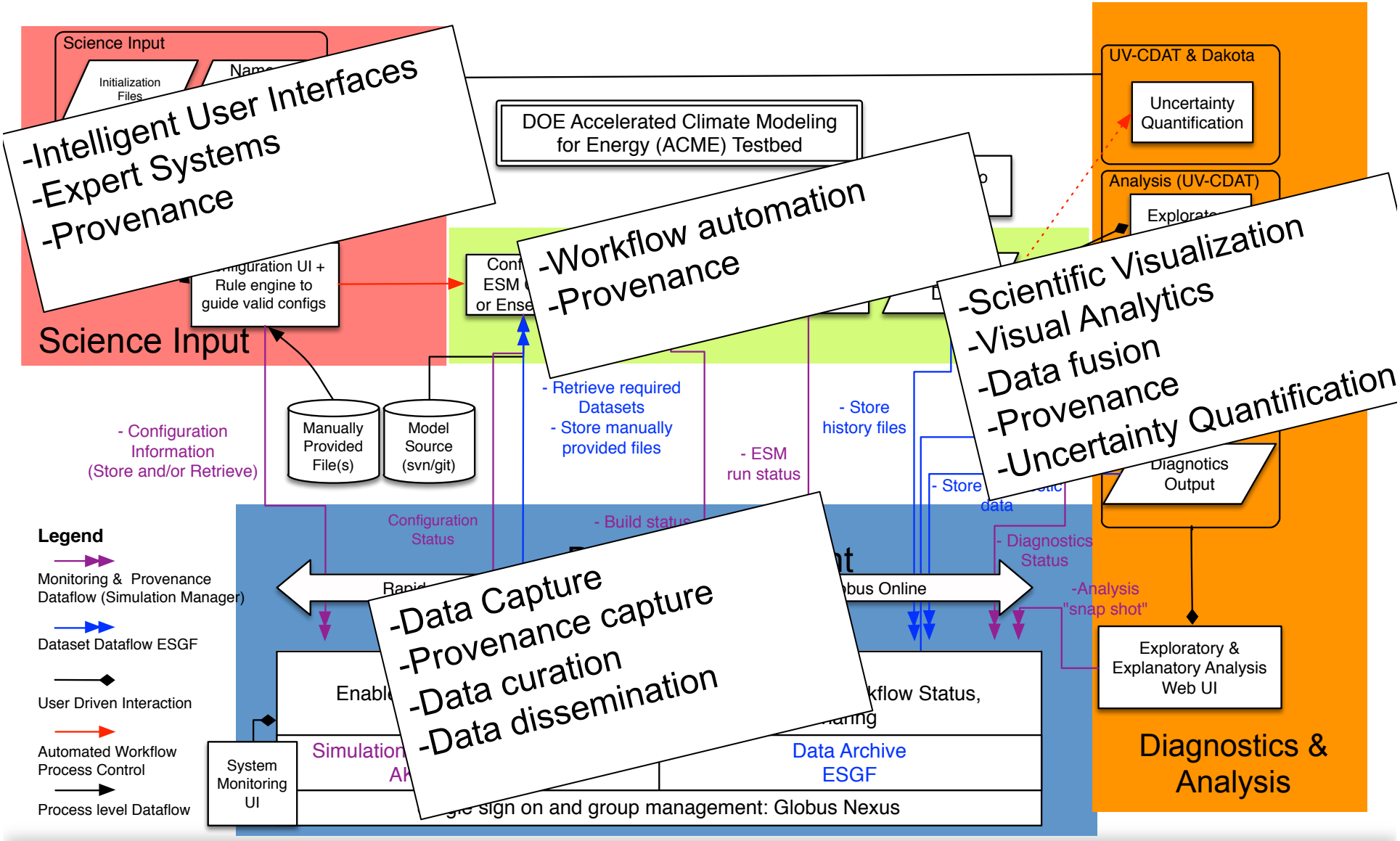




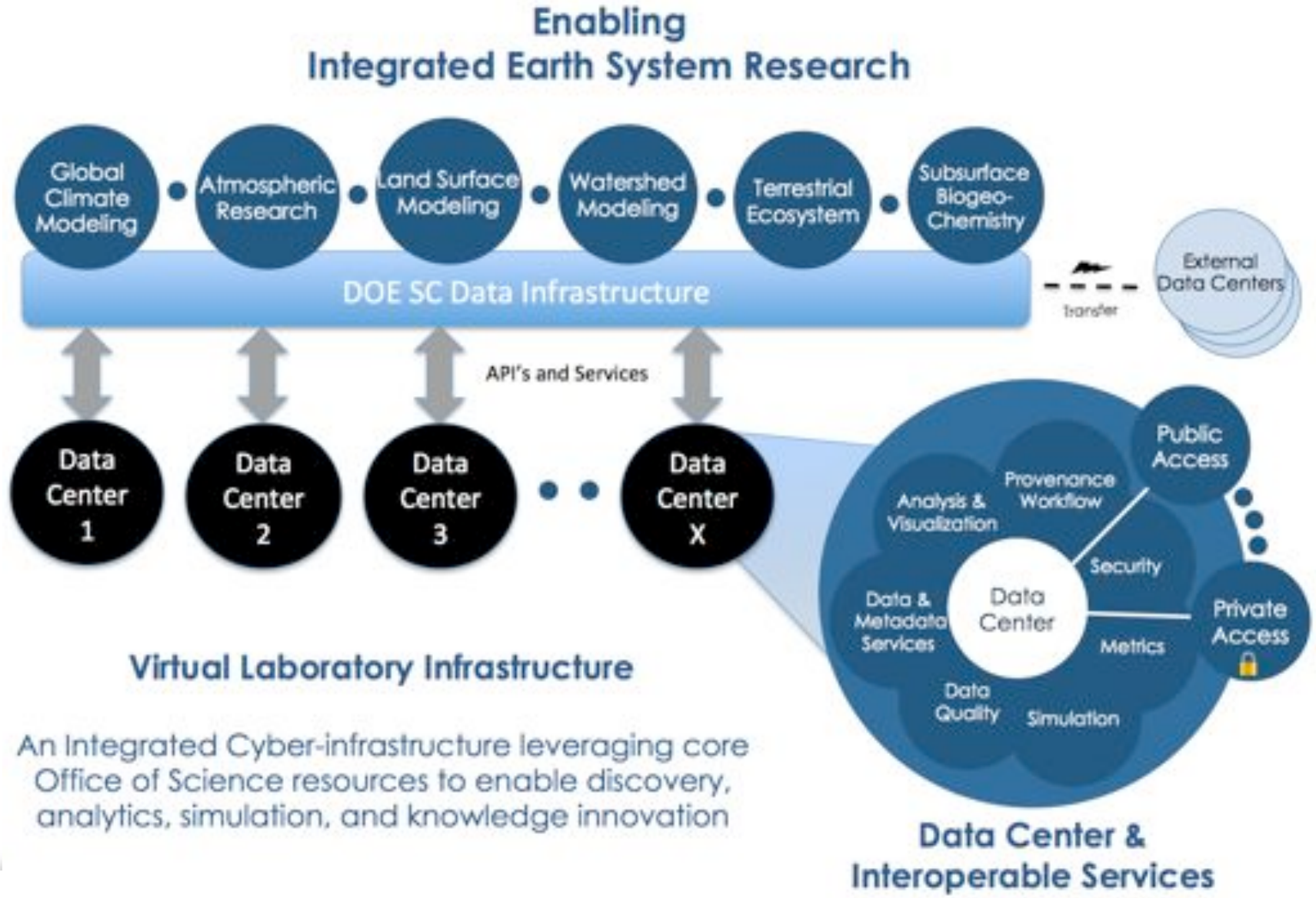
Snapshot of water vapor from a coupled simulation with DOE/NCAR CESM (Jamison Daniel, NCCS). Current processor layout of CESM on Titan (Pat Worley, CSMD)

- Hypothesis-driven development of a global coupled Earth system model
- Tailored for DOE Office of Science needs for high-resolution coupled simulation
- Enhanced evaluation of the coupled system using coordinated workflows and metrics
- ORNL is leading tasks related to workflow, land model development, and computational performance

ACME – Scientific infrastructure



A vision for an integrated data ecosystem for climate science



SNS Data Life Cycle

Acquisition

-Intelligent User Interfaces
-Live feedback
-Provenance

Reduction

- Corrected reduced data (hi
-
-
-

-Workflow automation
-On-demand computation
-Provenance

Analysis

- Multi dimensional fitting
- Advanced visualization
- Comparison to simulation / feedback
- Field dependent

-On-demand computational (potentially large scale)
-Provenance

Simulation Modeling

- Mul
- Adv
- 'Refi
- mental data
- Multiple experiments / probes

User Facility

Variety of experiments, topics, methods and 'computer literacy' of users are significant challenges.

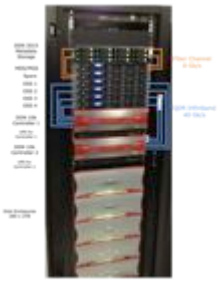
ADARA

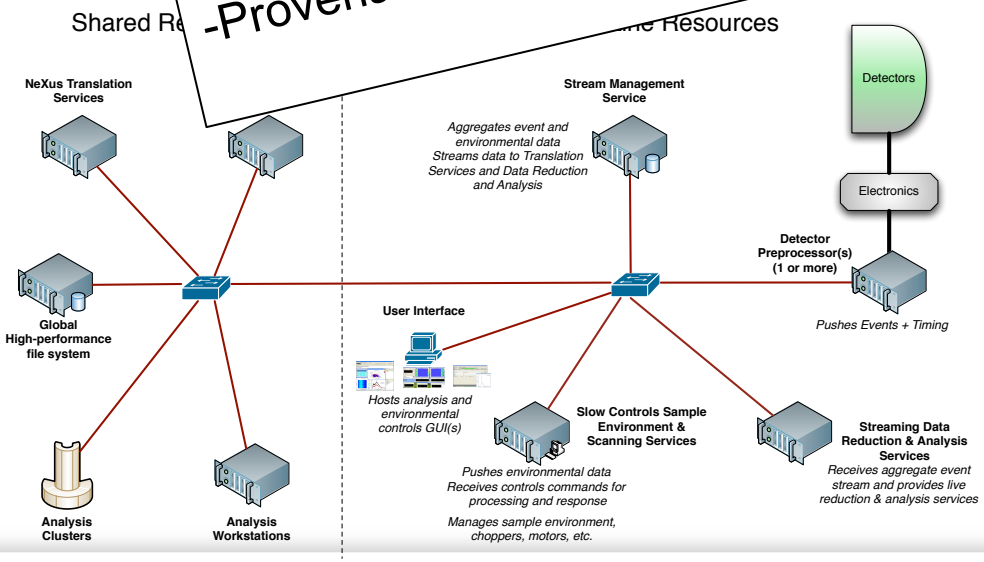
Accelerating Data Acquisition, Reduction, and Analysis at the SNS

Funded by Laboratory Directed Research & Development at ORNL

- We stream data (neutron and SE) from the DAS to a publish subscribe system
  - ***Stream Management Service (SMS)***
- We re-configure the data translation (file creation) to read the dat… from SMS and create the files while the run is taking place… end of run = close file [file…
  - ***Streaming Translation Service (S…***
- We modify MANTID (data reduc… SMS
  - ***Streaming Reduction Servic…***
- Files are created on an HPC infras… …d data reduction

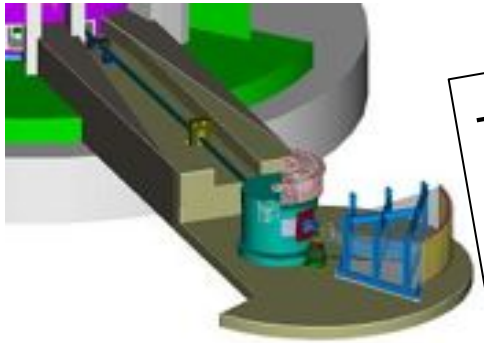
-In situ data reduction
-Live feedback
-On demand computation
-Workflow automation
-Provenance



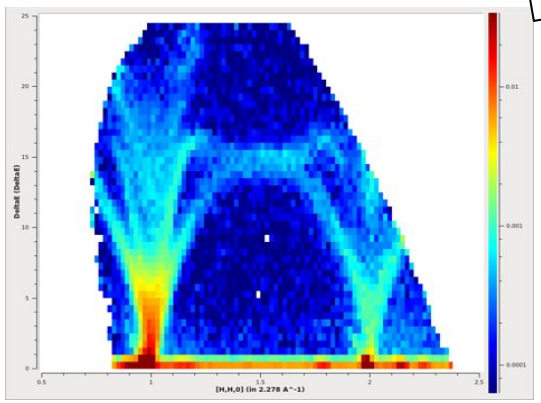


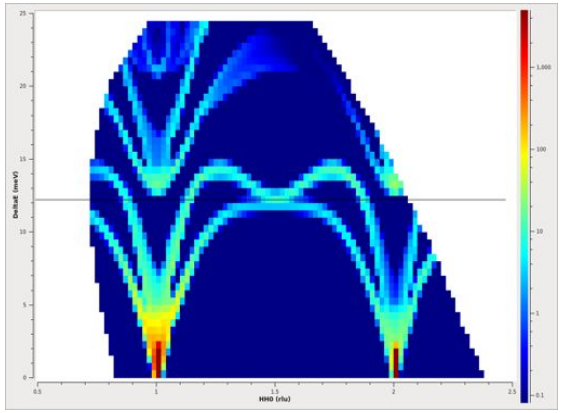
Center for Accelerating Materials Modeling

- SNS + HFIR collect a lot of materials spectra – if we can validate/refine simulation models against SNS/HFIR data then models "predict" measured atomistic properties. (Same for – APS, ALS, NSLS-I/II, LCLS, SSRL)
- Bring materials modeling/simulation directly into the chain for neutron scattering data analysis



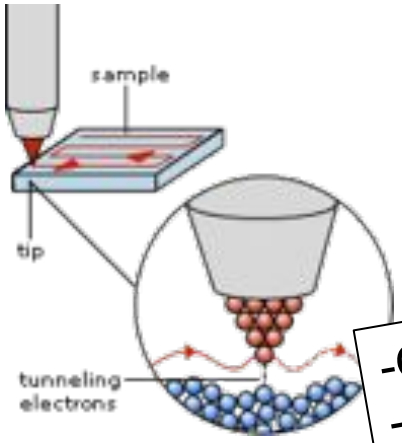
-On demand computation
-Workflow automation
-Intelligent user interfaces
-Provenance
-Uncertainty Quantification

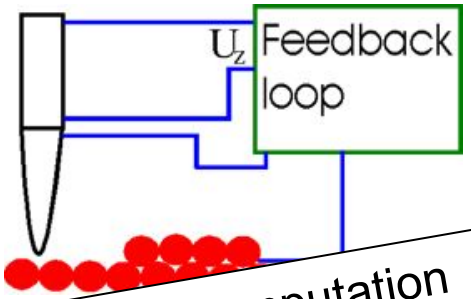


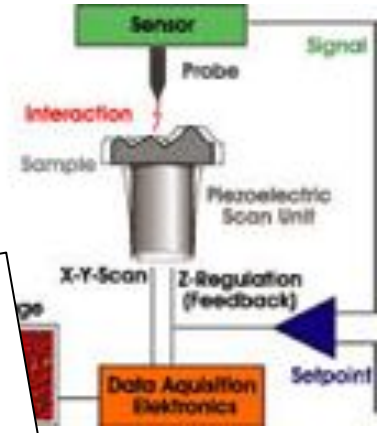


From Measurement to Knowledge

Atomic imaging







-On demand computation
-Workflow automation
-Intelligent user interfaces
-Provenance
-Image Processing
-Machine Learning
-Graph Analytics
-Uncertainty Quantification

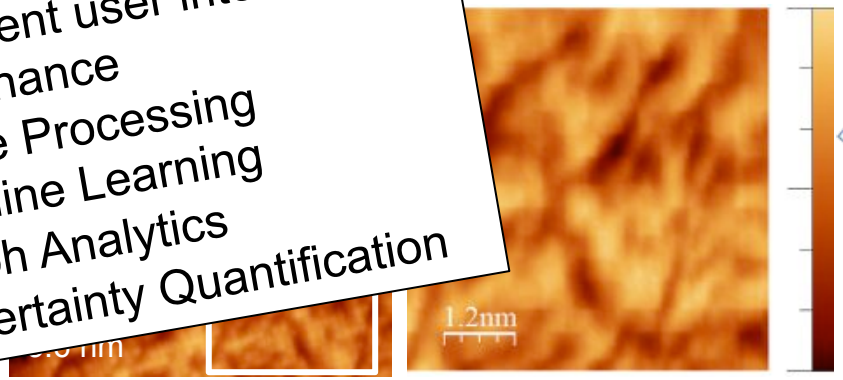
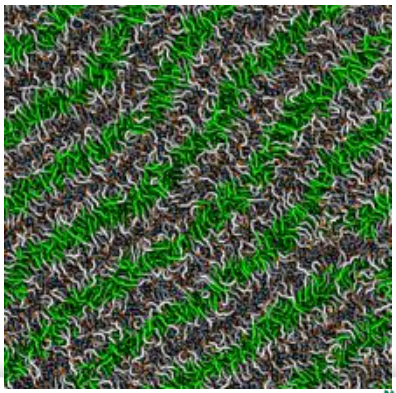


Image Processing to allow intercomparision of data from multiple modalities

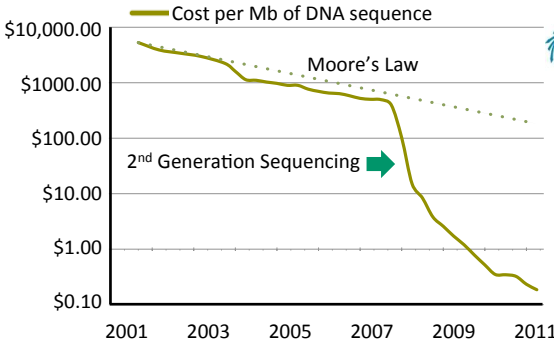
Machine learning for feature detection from these improved images

Molecular Dynamics full atomistic level simulation based on structural assemblies



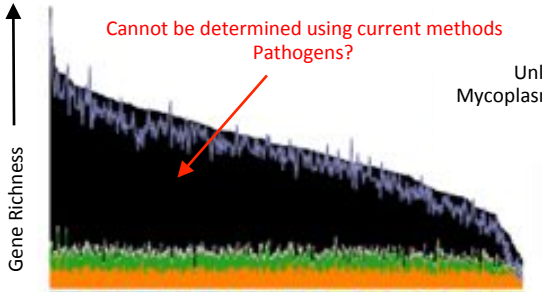
The Million Genome Project

Opportunity: Use Third Generation Sequencing for:
- Rapid Disease Identification
- Rapid Community Profiling—human microbiome



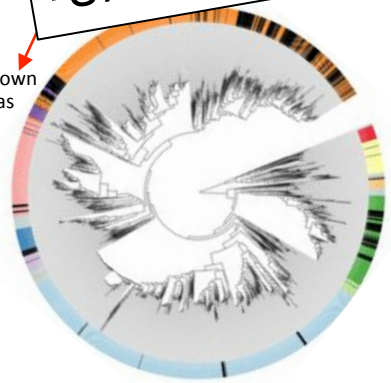
Infectious disease diagnosis:
- Currently requires culturing and can take
- By 2020, will be **entirely sequence-based**

Goal 1: Identify critical pathogens



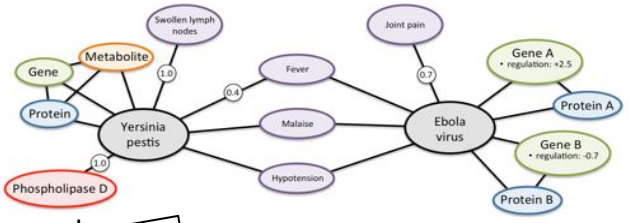
400 Human Intestinal Flora Samples, 10Tb, 9M genes

Approach:
- Build tree of **all known genomes**
- Sequence metagenomic sample
- Map DNA sequences to tree **in real time** using *k*-mer (short sequence) approaches



Tree of 2700 Genomes
GenBank currently has 46,000 genomes
several thousand new genomes each week
Need fast algorithms to update trees

Goal 2: Determine treatment



Approach:

-On demand computation
-Workflow automation
-Intelligent user interfaces
-Provenance
-Graph Analytics
-Uncertainty Quantification

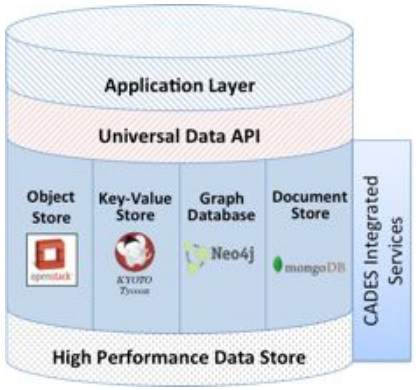
…otate genomes using ORNL tools, such as Prodigal
…ase to store characteristics of pathogenic bacterial strains,
…toxin genes, incubation periods, and host responses

…reaks

…s in the

…on to
…ation and
…meta-data
- Identify genomic differences with related pathogens

Key-Value Stores will be used for characteristic short sequences or distances between markers
Graph Databases will be used for genome trees and bacterial strain characteristics



Storage Infrastructure
for dealing with sequence data and genomes on a massive and unprecedented scale
*Need **million-fold increase** in speed*
Tightly integrated suite of storage stacks using the **CADES Platform** as a Service (PaaS) model

Need fast update algorithms for highly organized storage
*Need **reduced data models** for efficient storage and analysis of closely related genomes*

Benefits of rapid identification:
- Track outbreaks in real-time
- Eliminate inappropriate antibiotic prescriptions, slowing the development of antibiotic resistant bacteria

Emerging themes to support these initiatives

Scientific Domain Data Specialists

- Biology
- Climate Science
- Fusion Energy
- Healthcare
- High Energy Physics
- Materials Science
- Neutron Scattering
- Nuclear Energy
- Nuclear Physics
- Urban Environments

Visualization and Human Computer Interfaces

- Scientific Visualization
- Visual Analytics
- Interface technology and perception
- Visualization environments

Analytic Services

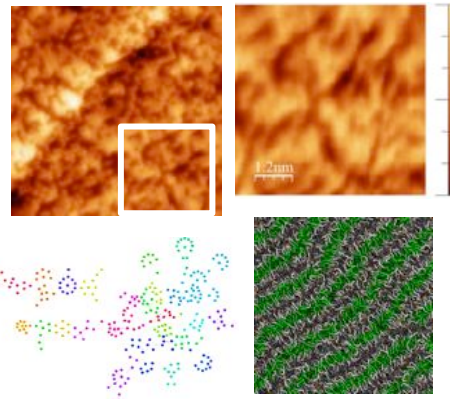
- Data Mining
- Mathematics
- Image processing
- Data fusion

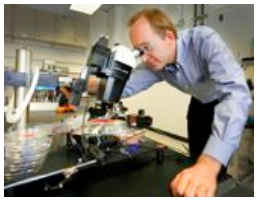
Data Management

- Data quality
- Data curation
- Data security
- Data access

Multiple Data Demonstration Projects Lined up for SC 2014

Multi-Modal Analysis of Ferroic Materials

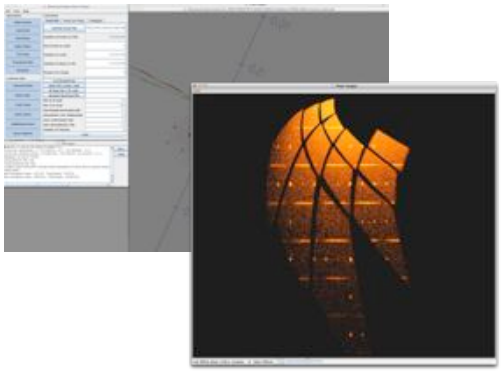




ORNL Lead:
Sergei Kalinin

IFIM Director

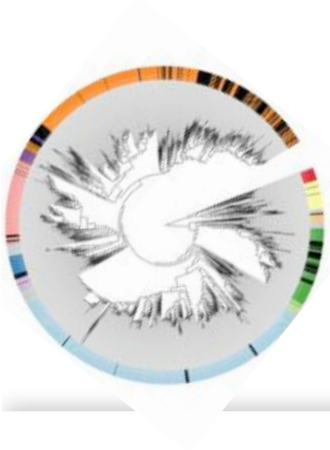
Feature Detection in X-Ray and Neutron Data



ORNL Lead:
Thomas Proffen

NDAV Director

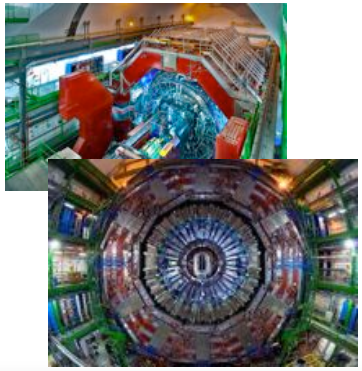
Towards a Million Genomes



ORNL Lead:
David Ussery

Comparative Genomics Group Leader

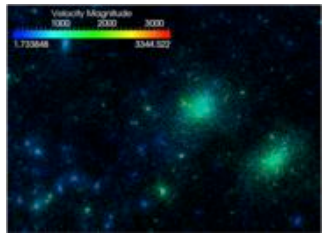
Scalable Analysis in High Energy & Nuclear Physics



ORNL Lead:
Kenneth Read

Experimental Nuclear Physics (UTK/ORNL)

Extreme Data Analysis for Cosmology



ORNL Lead:
Bronson Messer

Scientific Computing