About: Visualization

Mark Hereld

Argonne National Laboratory

CScADS Summer Workshops 2009

Leadership-class Machines,
Petascale Applications,
and Performance Strategies

July 27-30, Tahoe City, California

Goals of this Overview

- Presentation
  - Role of visual data analysis
  - Impediments and Issues
  - Panoramic sweep of major and minor tools
  - Mini-compendium of useful approaches
  - New directions for visual data analysis in HPC context

- Discussion (either in your own head or in the room)
  - Compare notes: needs, special concerns, goals
  - Find out what isn't working: usability, functionality, other
  - Wish list: features, process (workflow)

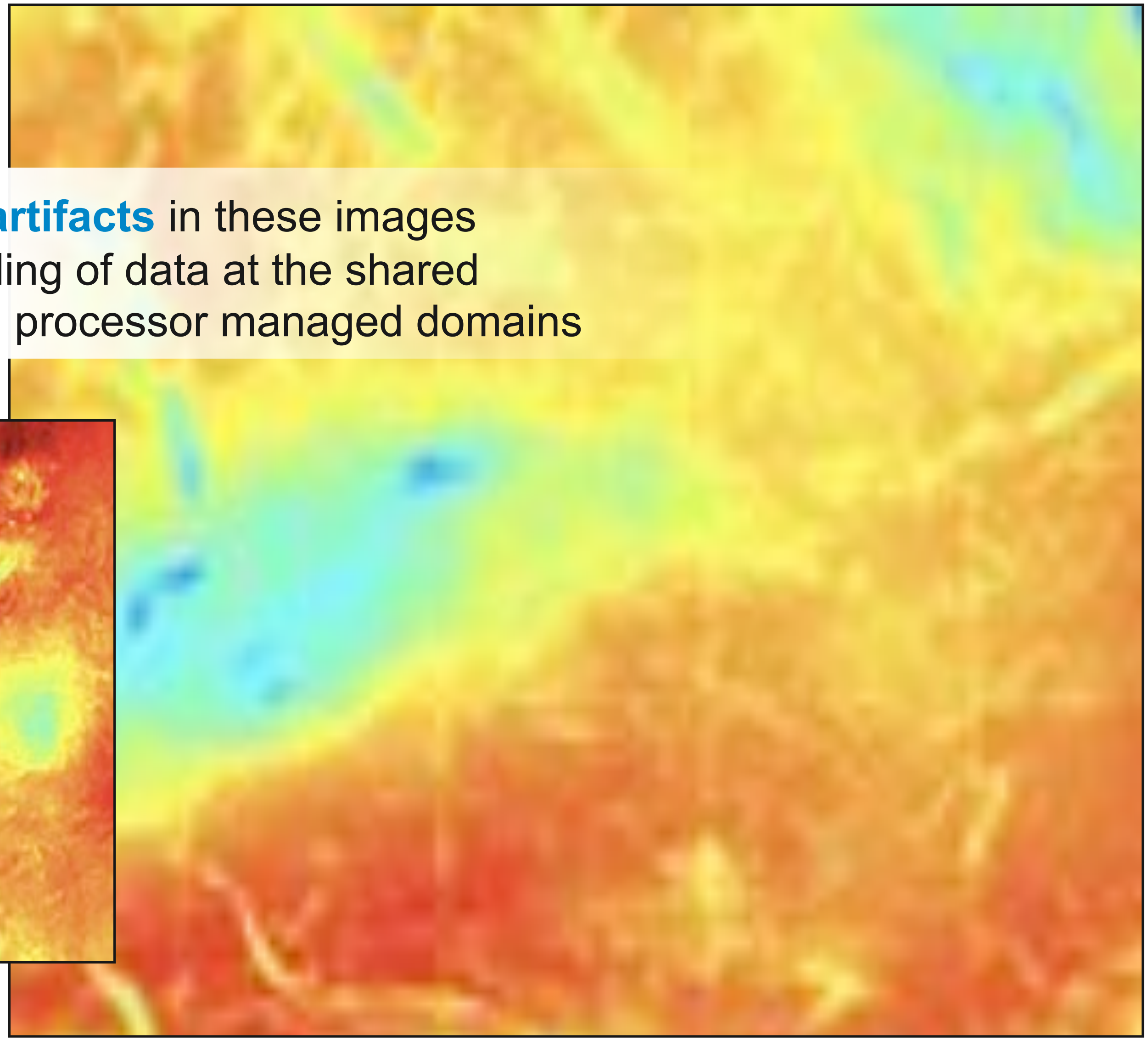
Your Job

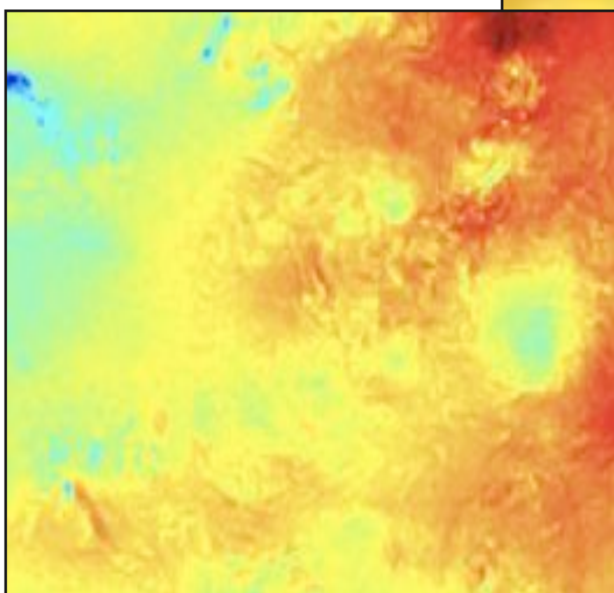
Think about…

- Practical problems?
- Annoyances?
- Common workflow / process issues and strategies?
- In situ analysis -- a boon to you?
- Data management -- what do you do today? What will you do when each run creates a million files?

Visual Data Analysis: What is it Good For?

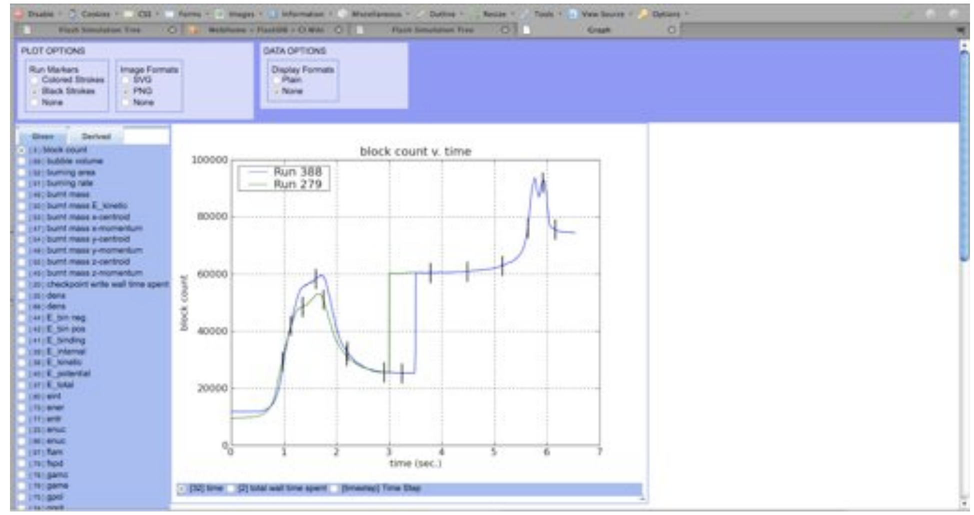
- Debugging
  - Gross and subtle **artifacts** in these images evidence mishandling of data at the shared boundary between processor managed domains

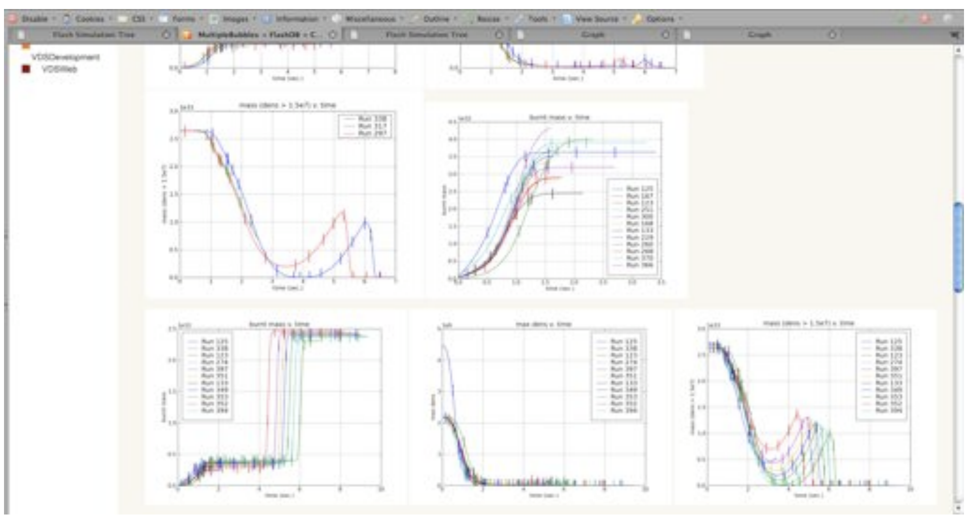


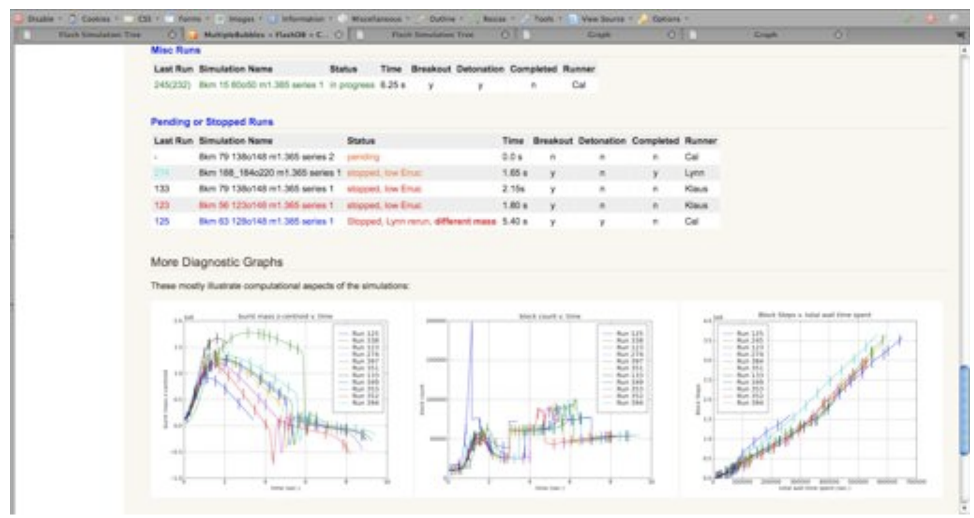


Visual Data Analysis: What is it Good For?

- Monitoring
  - Sampling the Scientific Pipeline
  - Dynamically Generated Notebooks
  - Collaborative Annotation

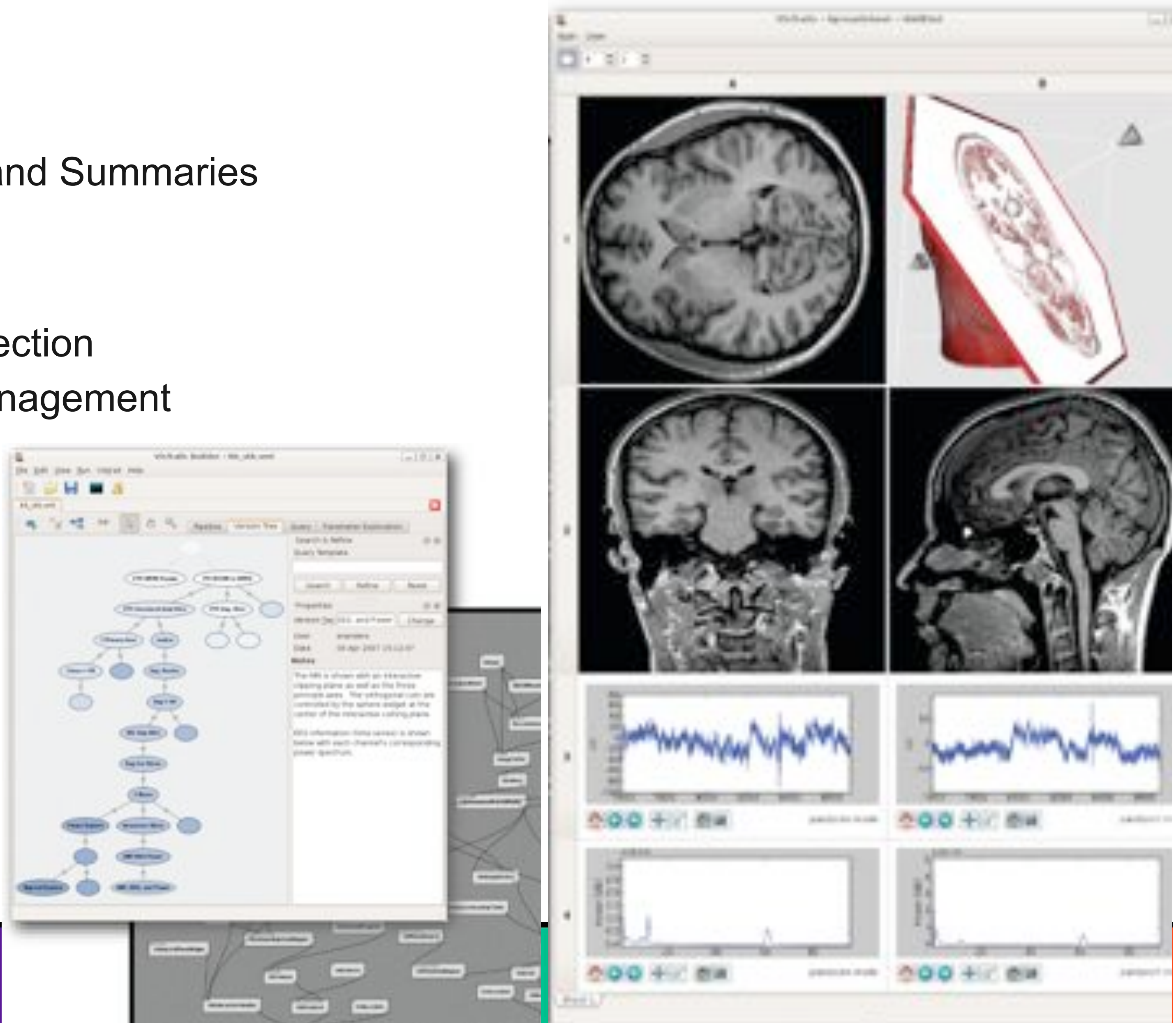






Visual Data Analysis: What is it Good For?

- Exploration
  - Overviews and Summaries
  - Deep dive
  - Interactivity
  - Feature detection
  - Process management
  - Provenance



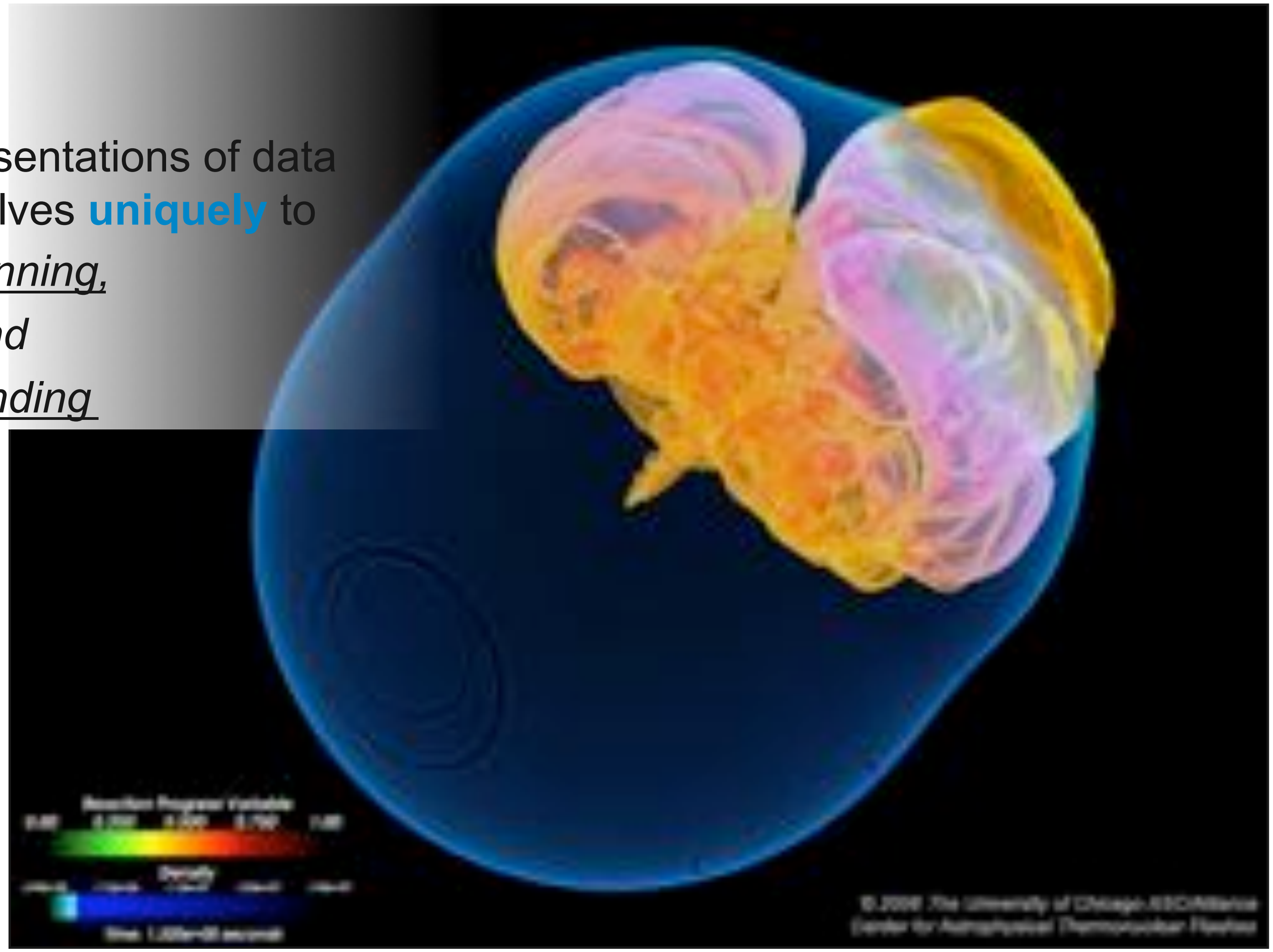
Argonne NATIONAL LABORATORY

Visual Data Analysis: What is it Good For?

- Collaboration
  - Group decision-making
  - Communicating ideas and results
  - Collaborative data exploration

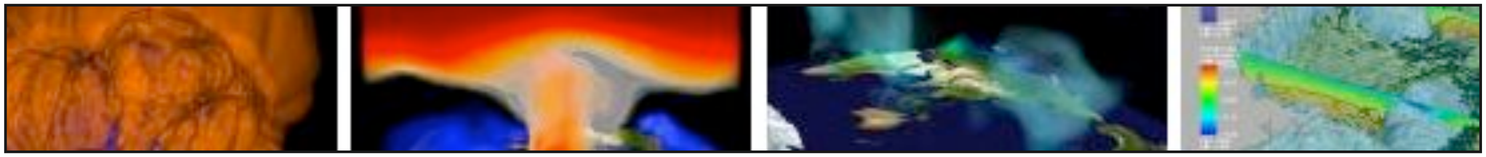
Visual Data Analysis: What is it Good For?

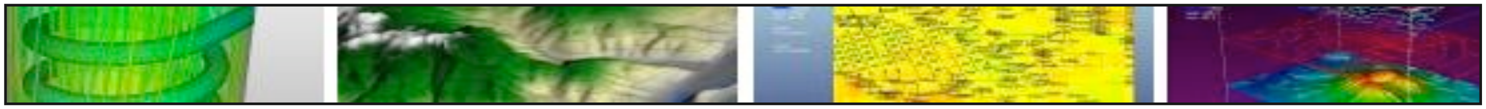
- Discovery
  - Visual representations of data lend themselves **uniquely** to
    - *rapid scanning,*
    - *ingest, and*
    - *understanding*

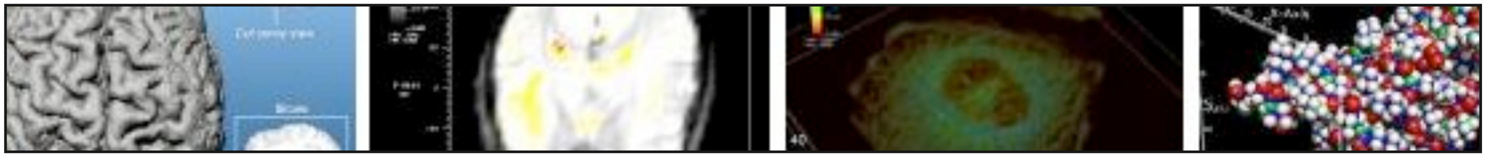


Visual Data Analysis: What is it Good For?

- Dissemination
  - Undeniably **powerful** and useful
    - *Publication of results*
    - *Outreach and education*
    - *Icons / labels*
    - *Impact >> promotion*







Ping

- Debug – monitor – explore – collaborate – discover – disseminate
- Relative weights of these applications of visualization in your mind?
- Other important roles / refinements for your work flow?

Under the Circumstances

- Confounding circumstances
  - Compute is far away, expensive, batch
  - Storage is distant as well
    - *Datasets are very large*
    - *Disk speed and network bandwidth are constraining*
  - Workstation and display pixels are local
    - *And these are limited in capacity*
  - Data management
    - *Large number of data files*
    - *Large number of runs*
    - *Large number of collaborators*
- Exploring results is challenging

32K procs
* 29x29x29 cells
928x928x928 cells
751 time steps
21 variable
30 GB / time step
22 TB total sim

4K x 4K x 4K
4 byte singles
1 variable
65 GB / time step

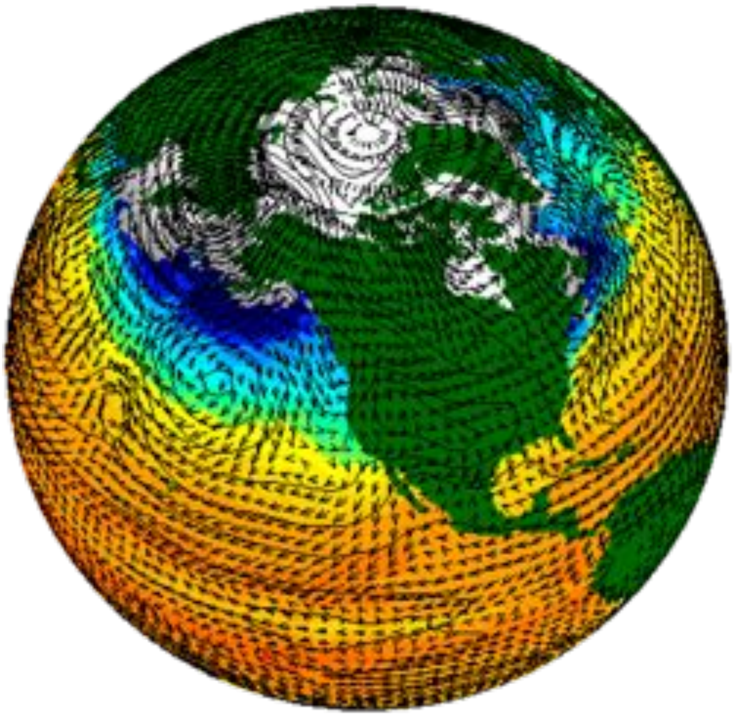
Nuclear Reactor Simulation

- Preliminary studies
  - 4.5 million elements
  - 7 variables per element
  - 20 K timesteps
  - Total data produced 2.5 TB
- Science runs
  - 3 – 4 runs with 120 million elements
  - Several runs at ½ and ¼ resolution
  - 90 K timesteps
  - Total data produced 900TB – 1.2 PB

Climate Modeling

- Preliminary studies
  - 50-100 with 3 million grid points (1 M atmosphere, 2 M ocean)
  - 100 variables per grid point (30 vectors, 70 scalars)
  - Simulating 5 - 10 years of climate
  - Total data produced 30 -124 TB

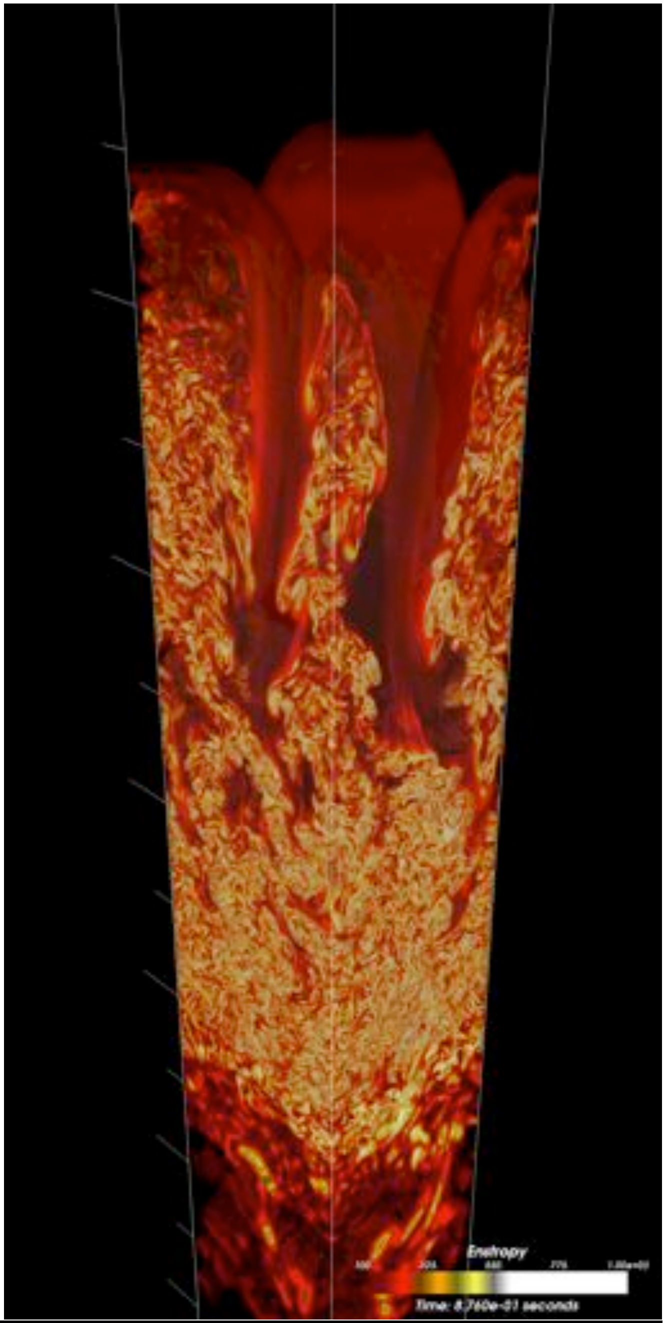
- Science runs
  - 50 runs with 6 million grid points
  - Simulating 100 years of climate
  - Total data produced 1.2 PB



Astrophysics

- Preliminary studies
  - ~80 with 67 M grid points
  - ~5 with 536 M grid points
  - 6 variables (1 vector, 3 scalars)
  - ~1800 time steps
  - Total data produced 78 TB
- Science run*
  - 1024^2 x 4096 grid points
  - 6 variables (1 vector, 3 scalars)
  - ~1800 time steps
  - Total data produced 48 TB

*3-5 times bigger allocation is needed



All Sorts of Tools

- Visualization Applications
  - VisIt
  - ParaView
  - EnSight
- Domain Specific
  - PyMol, RasMol
- APIs
  - VTK: visualization
  - ITK: segmentation & registration
- GPU performance
  - Scout: GPGPU acceleration
  - vl3: shader-based vol ren
- Analysis Environments
  - Matlab
  - Parallel R (ORNL)
- Utilities
  - GnuPlot
  - ImageMagick
- Visualization Workflow
  - VisTrails

All Sorts of Concerns

- Data dimensionality: 1D, 2D, 3D, .. high-D
- Structure of your data
- Fusion of multi-modal data
- Multi-scale data
- Interactivity needs speed and low latency

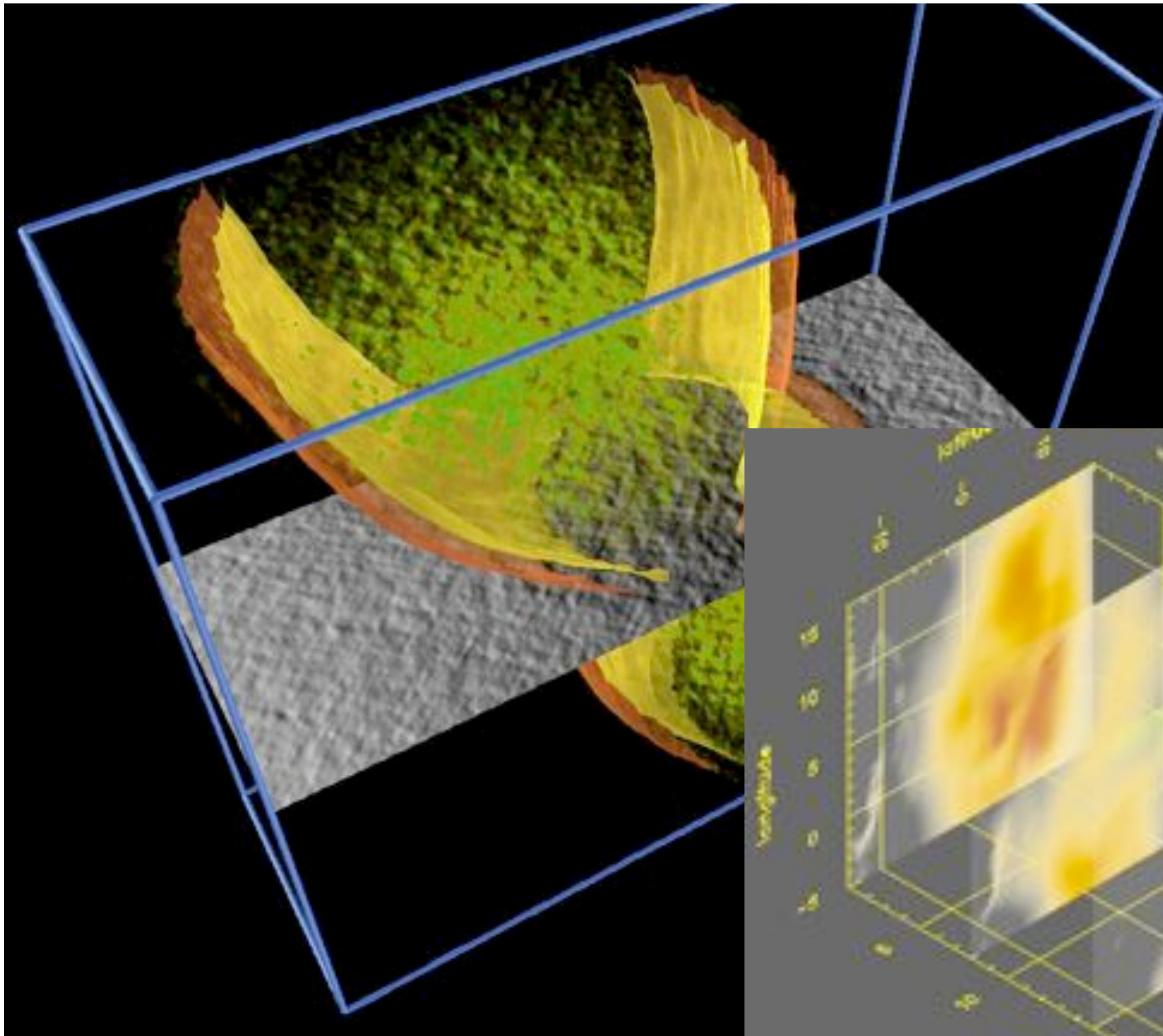
- Perceptual issues

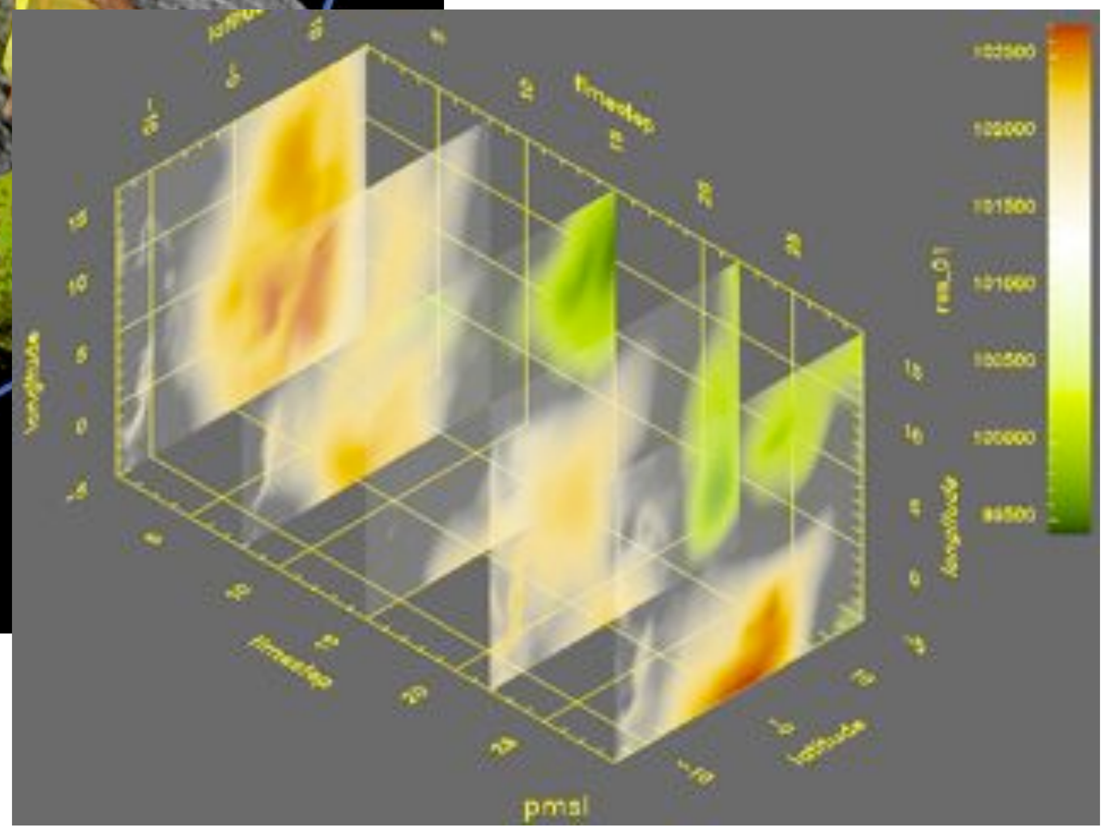
All Sorts of Visual Representations

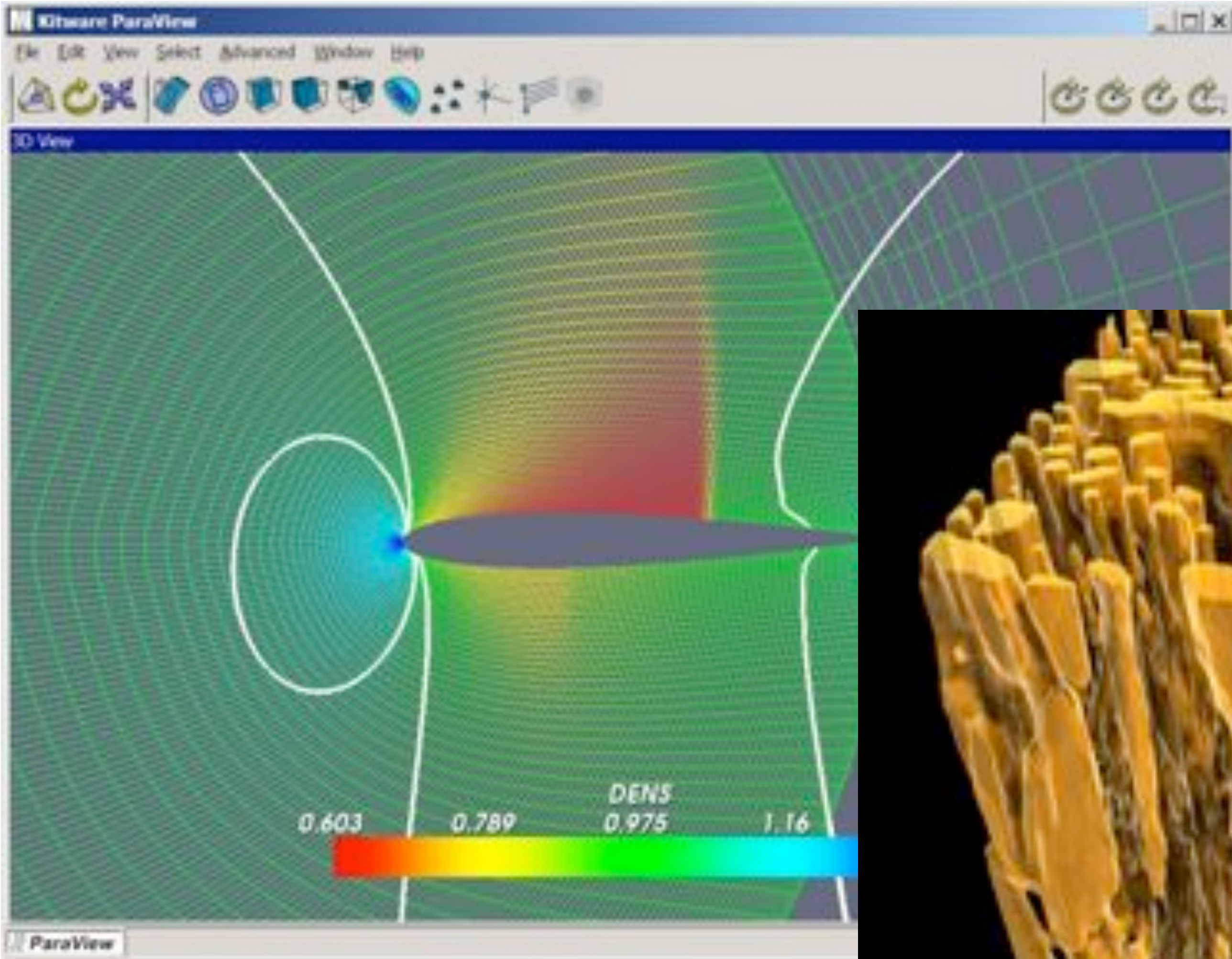
- Graphs
- Volume visualization
  - Transparency
  - Feature extraction
- Particles
- Streamlines
- Isoclines and Isosurfaces

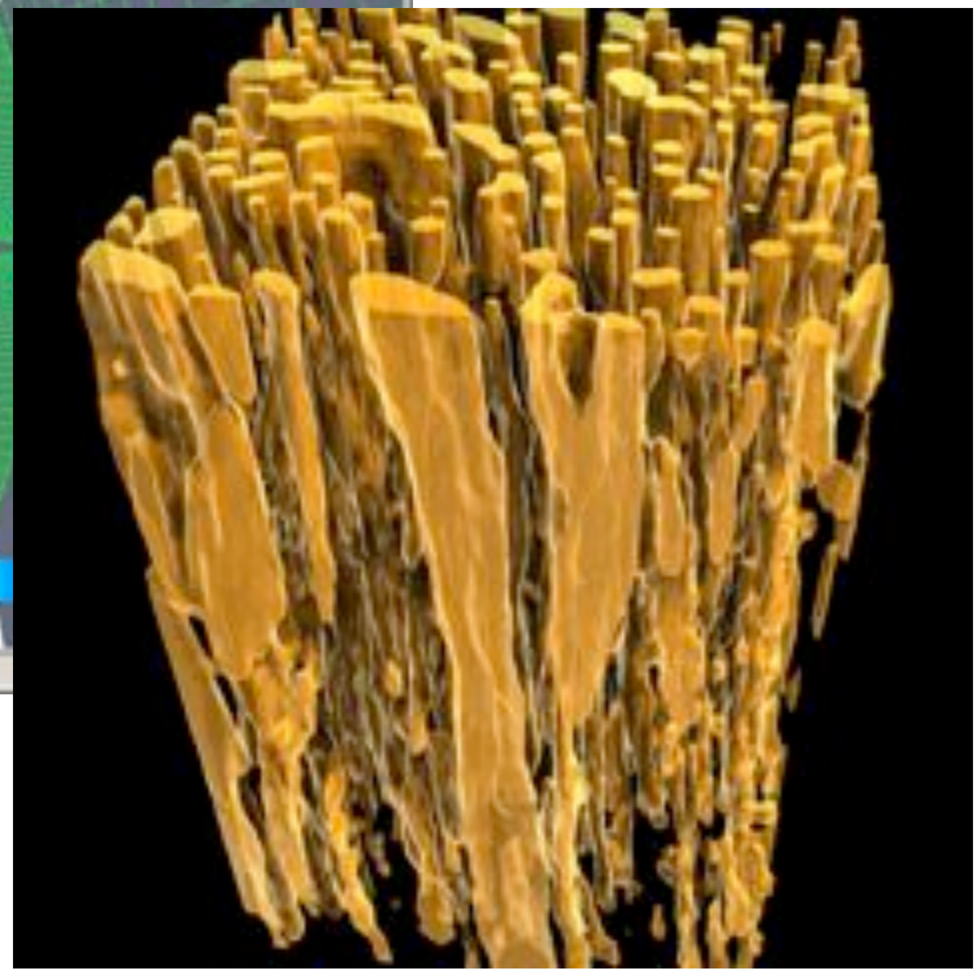
- Slices
- Boxes
- Brushes
- Calipers

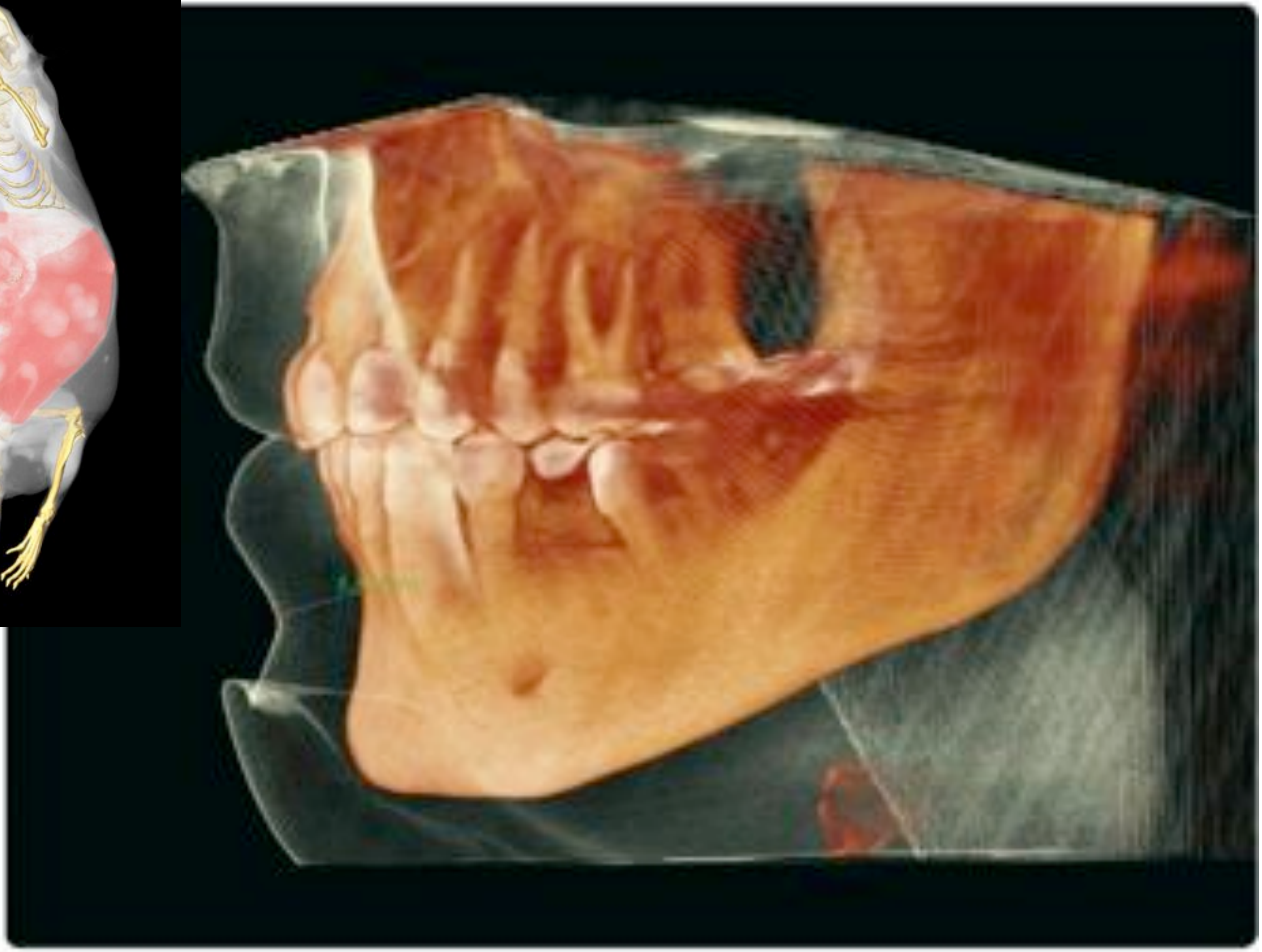
- and many Combinations

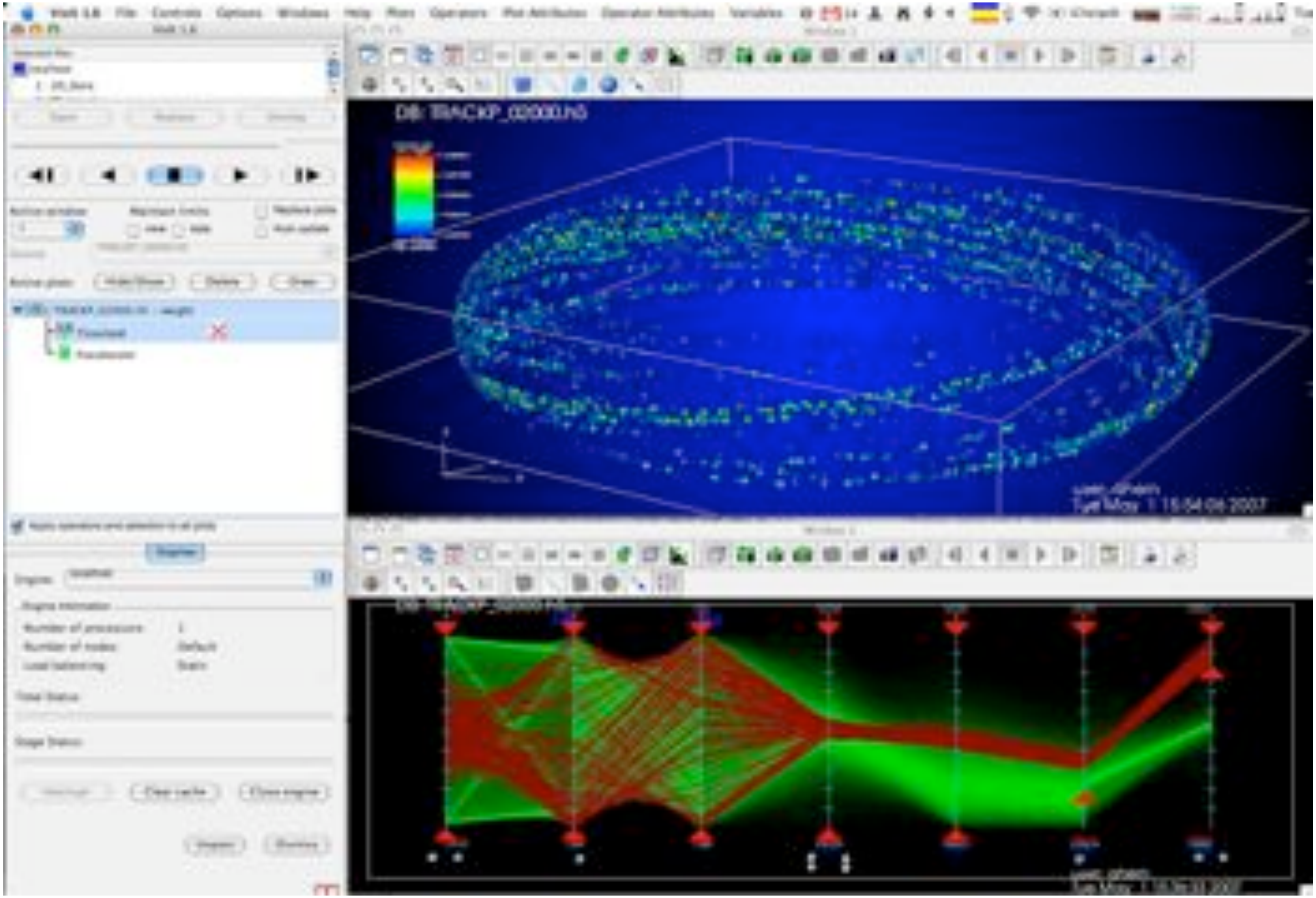



longitude
timestep
latitude
pmsl


Kitware ParaView
File Edit View Select Advanced Window Help
3D View
DENS
0.603
0.789
0.975
1.16
ParaView



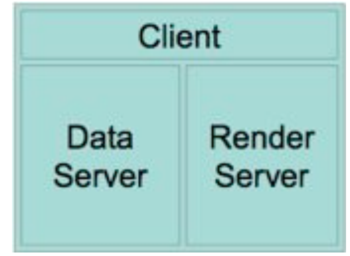




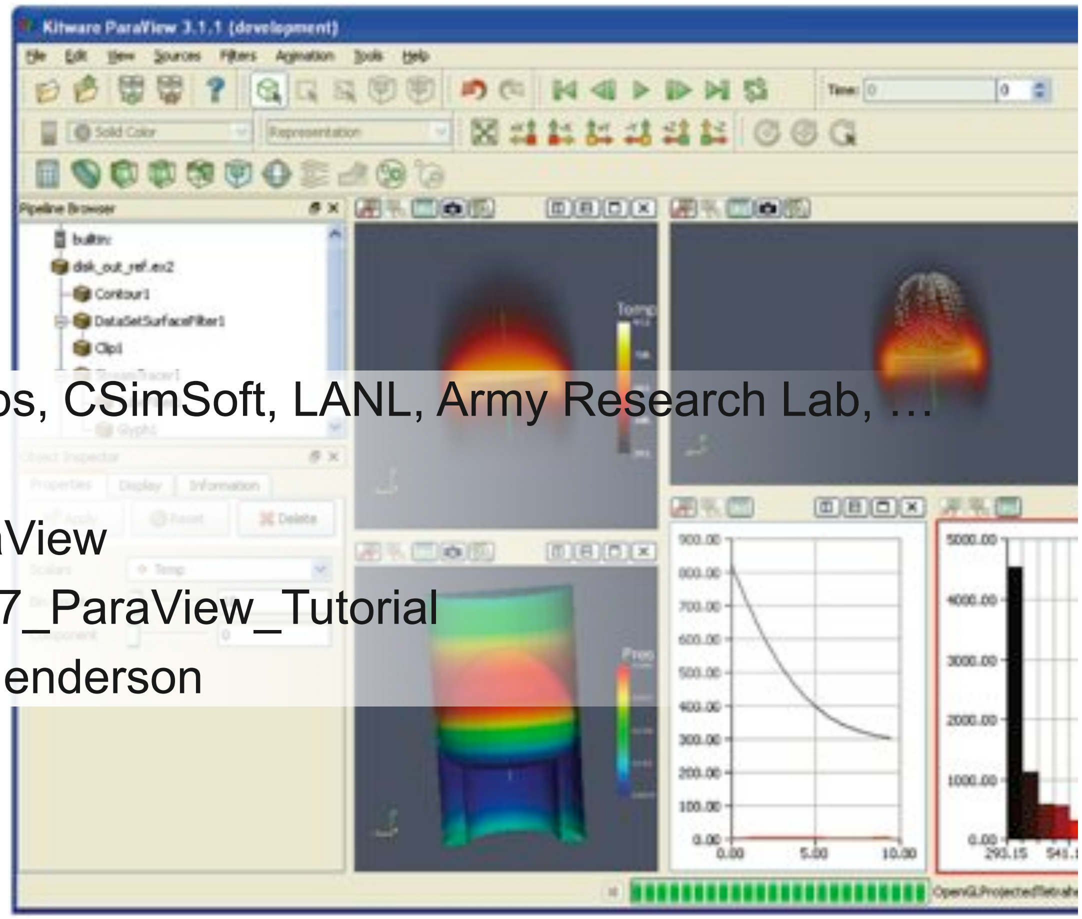
Ping

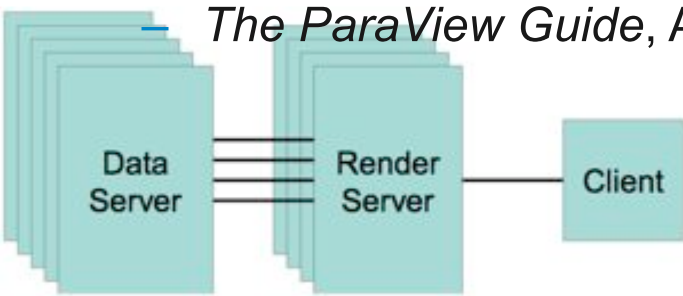
- Any interest in interacting with your simulation?
- How complicated is your setup / config?

ParaView Overview



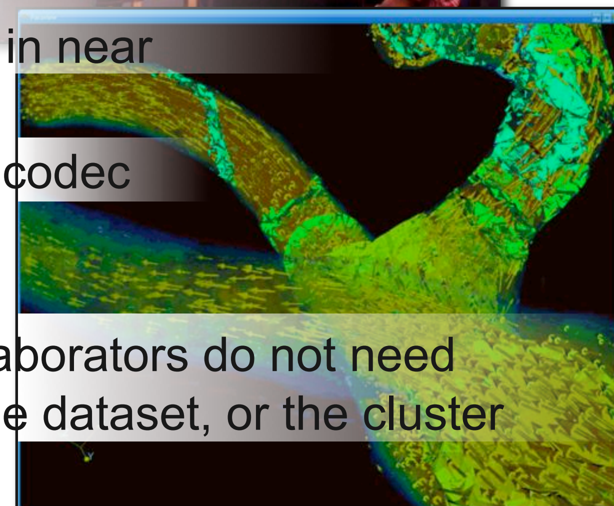
- Parallel Visualization Application
- Open source
- VTK + Tcl
- Python scripting
- Interactive and batch
- About
  - Kitware, Sandia National Labs, CSimSoft, LANL, Army Research Lab, …
  - http://www.paraview.org
  - http://paraview.org/Wiki/ParaView
  - http://paraview.org/Wiki/SC07_ParaView_Tutorial
  - *The ParaView Guide*, Amy Henderson



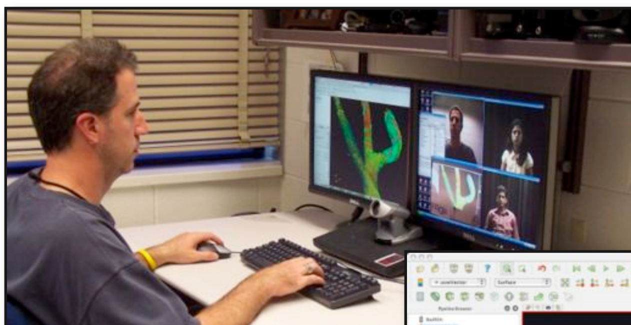


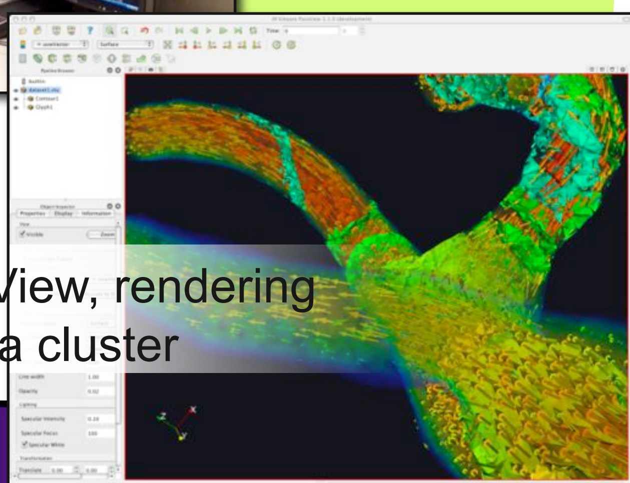
Augmenting ParaView

- Extended to stream visualization using video codecs
  - Simple, first-level sharing of visualization results in near real-time, at native resolution
  - Optimized for bandwidth and efficiency by video codec
- Plugin available for ParaView on all platforms



Remote collaborators do not need ParaView, the dataset, or the cluster

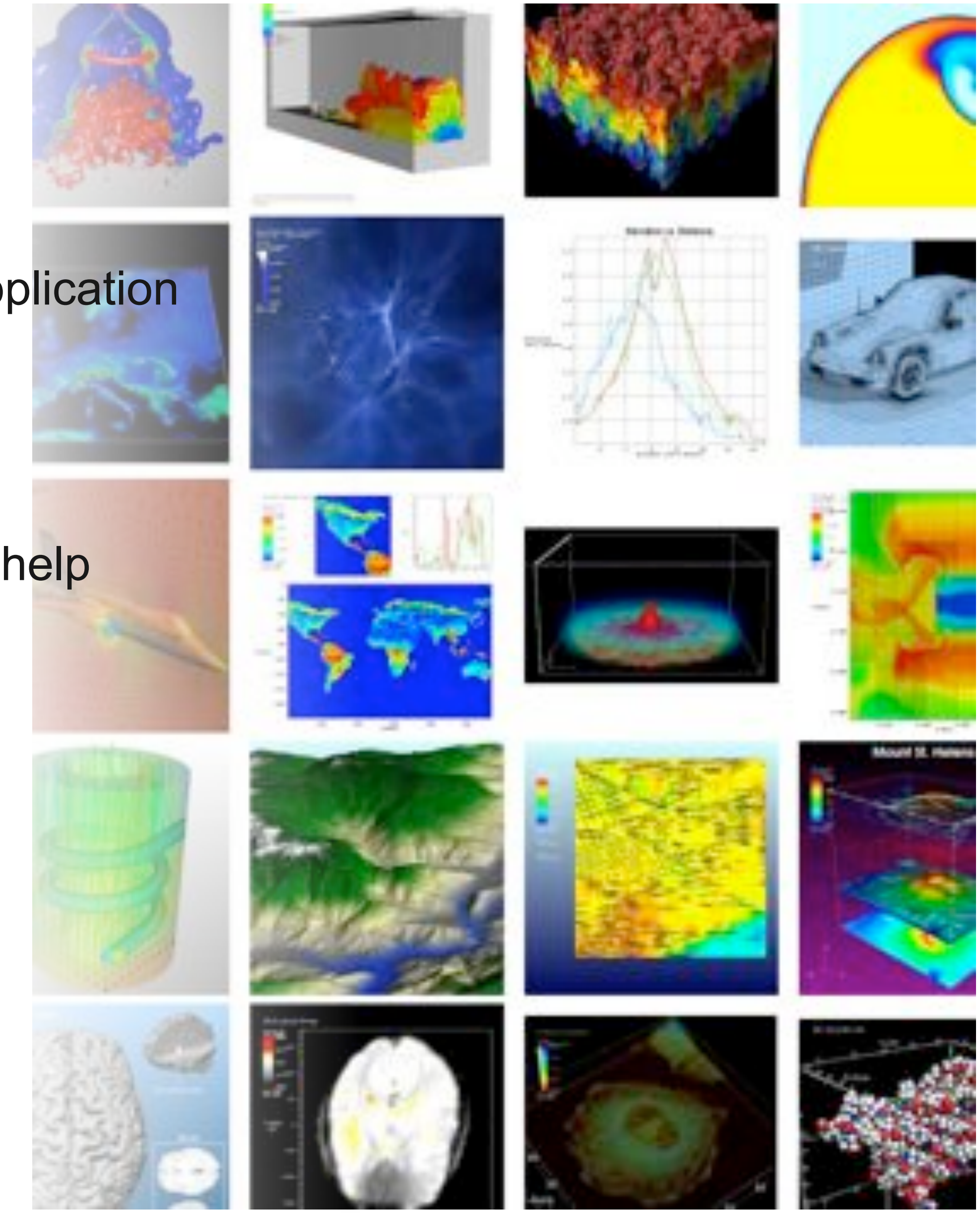




User running ParaView, rendering remote dataset on a cluster

VisIt Overview

- Parallel interactive visualization application
- About
  - DOE ASCI
  - https://www.llnl.gov/visit
  - Manuals, tutorials, application help

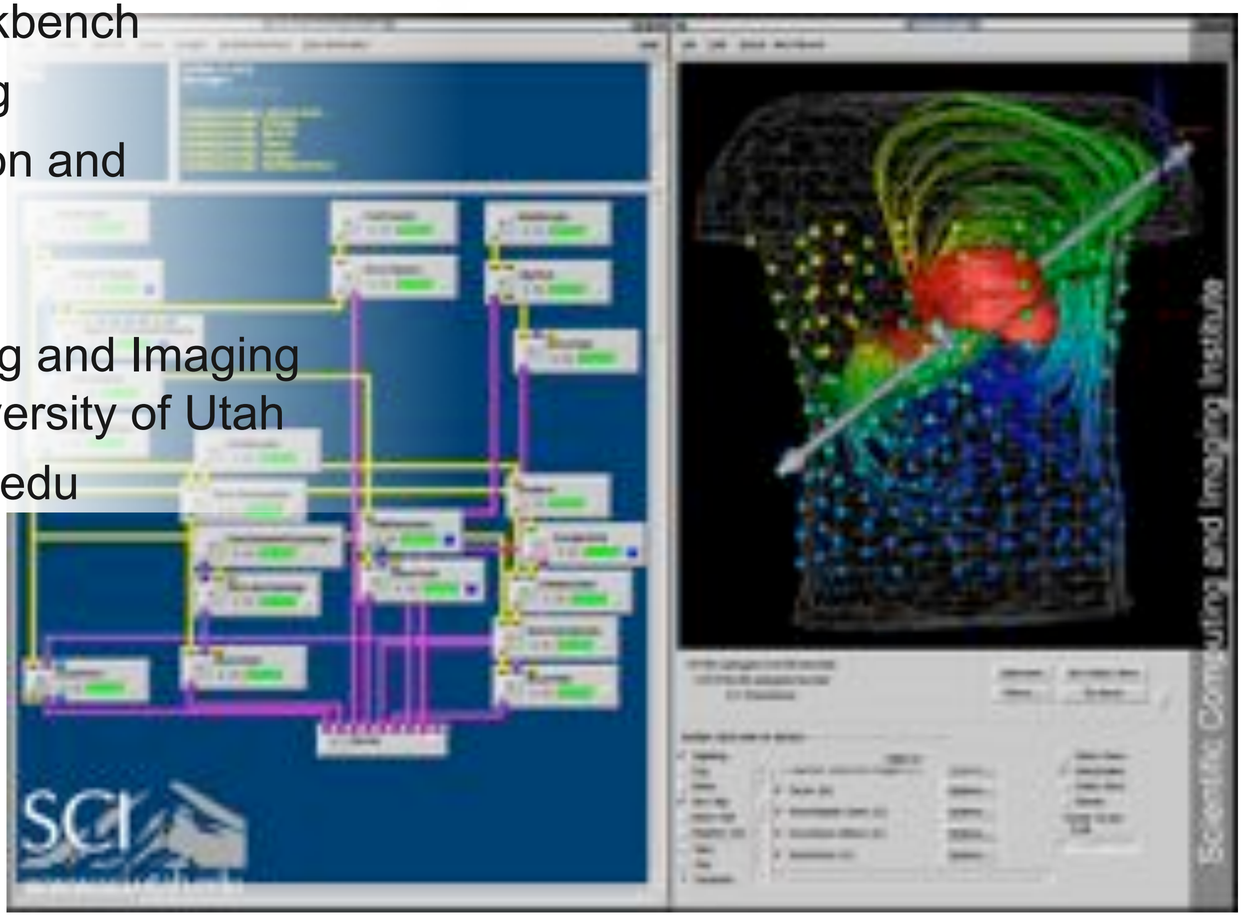


EnSight Overview

- Comprehensive visualization application
- Parallel and distributed rendering
- Flexible use of mixed CPU and GPU resources
- Tiled display support
- Interactive VR support
- About
  - CEI
  - http://www.ensight.com

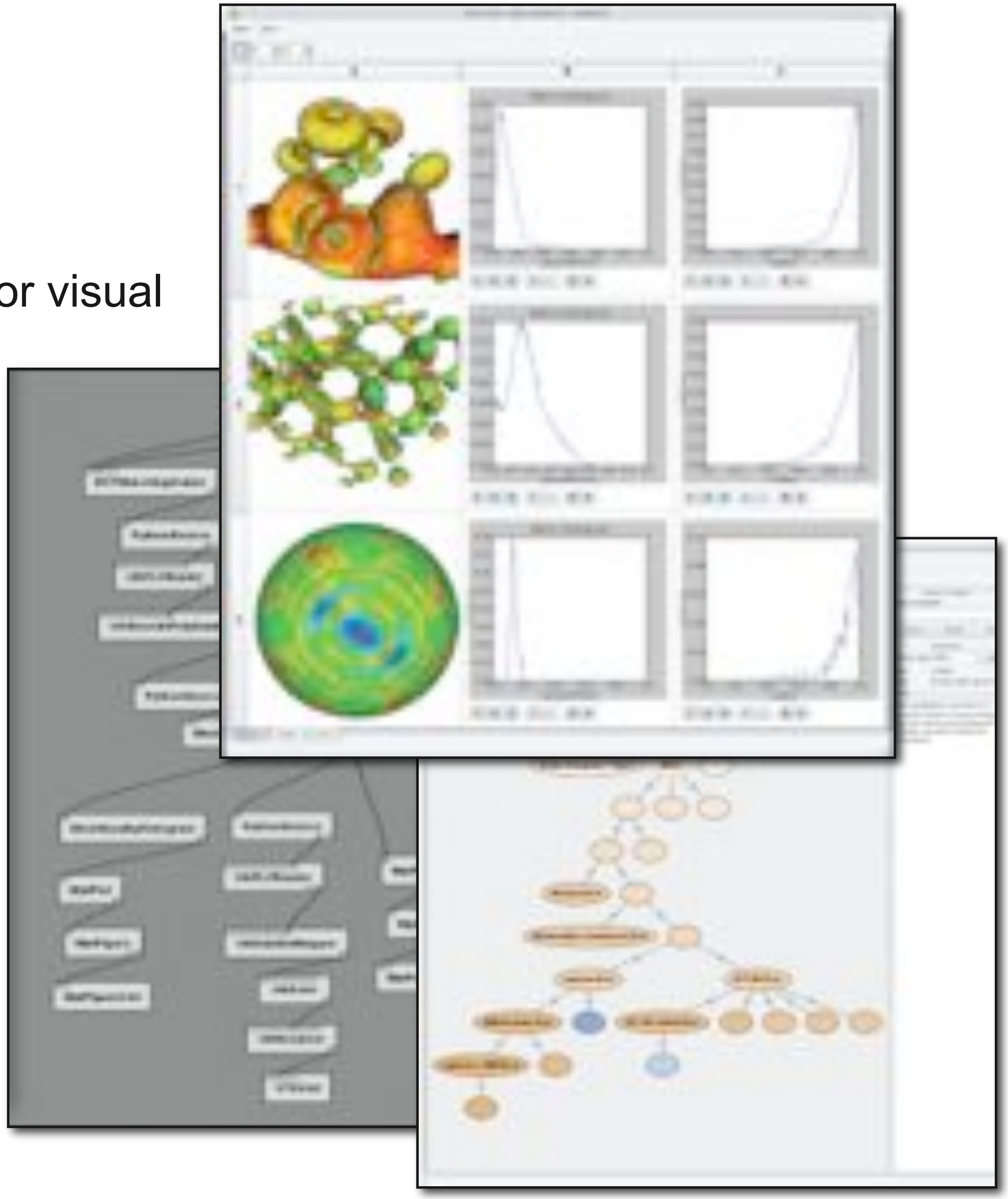
SCIRun Problem Solving Environment

- Extended suite of tools
  - Computational workbench
  - Visual programming
  - Modelling, simulation and visualization
- About
  - Scientific Computing and Imaging (SCI) Institute, University of Utah
  - http://www.sci.utah.edu



VisTrails

- Scientific workflow management for visual data analysis
- Visual programming
- Construct and execute pipelines
  - VTK, ITK, and Matplotlib
- History tree captures provenance
- Visualization spreadsheet
- About
  - http://www.vistrails.org



Full-featured visualization environments

- Interactive
- Parallel and distributed rendering
- Remote visualization
- Large library of algorithms
- Scriptable
- Rich format library
- Computation

VTK: a Visualization API

- Open source
- multi-platform
  - Unix, Windows, Mac OS X
- object oriented
- Tcl/Tk, Java, Python, C++
  - On top of extensible C++ class library
- About
  - Kitware, Inc.
  - http://www.vtk.org
  - *The Visualization Toolkit*, Will Schroeder, Ken Martin, Bill Lorensen
  - *The VTK User's Guide*

ITK: a Data Segmentation & Registration API

- Open source
- Multi-platform
- Multi-D (2,3,…)
- Tcl, Java, Python, C++
  - On top of C++
- About
  - Visible Human Project
  - http://www.itk.org
  - *ITK Software Guide*, Luis Ibanez, William Schroeder (book and pdf available)
  - *Insight into Images*, Terry Yoo (Editor)

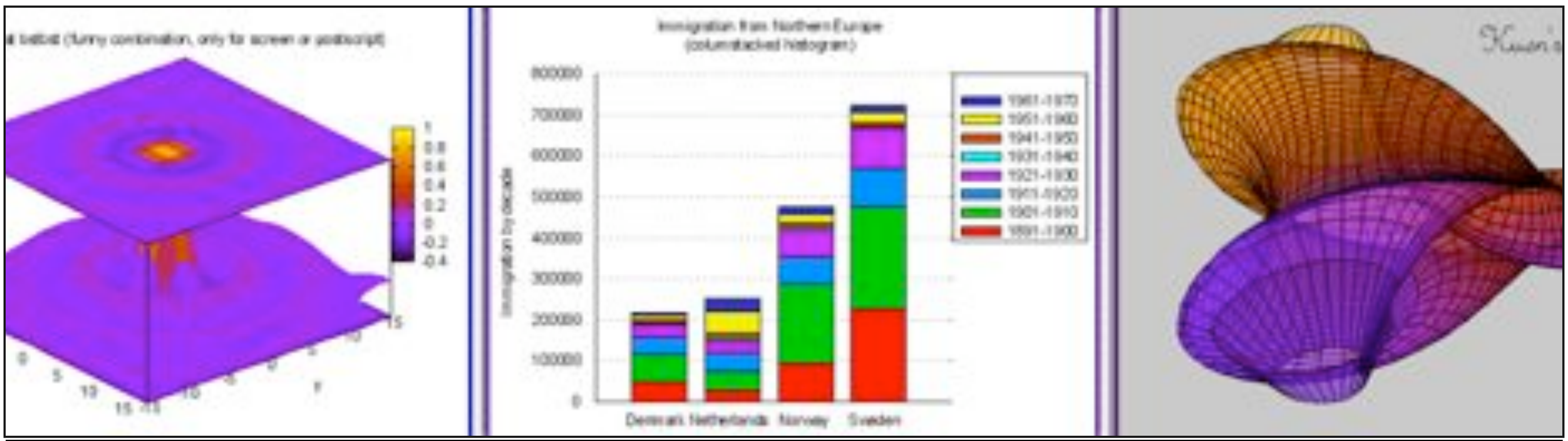
Matlab

- Analysis -- matrix, objects
- Basic visualization
- Command line interactive
- Programming language
- GUI tools
- Rich set of toolkits
  - Image, simulations, signal, optimization, statistics
- Support for HDF5
- Parallel extensions (*)

GnuPlot

- General purpose 2-D and 3-D scientific data plotting
- Command line interactive, scriptable
- Multi-platform
- LaTex integration
- About
  – http://www.gnuplot.info

```
#
# $Id: mgr.dem,v 1.8 2004/01/13 07:01:10 sfeam Exp $
#
print "Watch some  cubic splines"
set samples 50
set xlabel "Angle (deg)"
set ylabel "Amplitude"
set key box
set title "Bragg reflection -- Peak only"
plot "big_peak.dat" title "Rate" with errorbars, \
        "" smooth csplines t "Rate"
```



ImageMagick

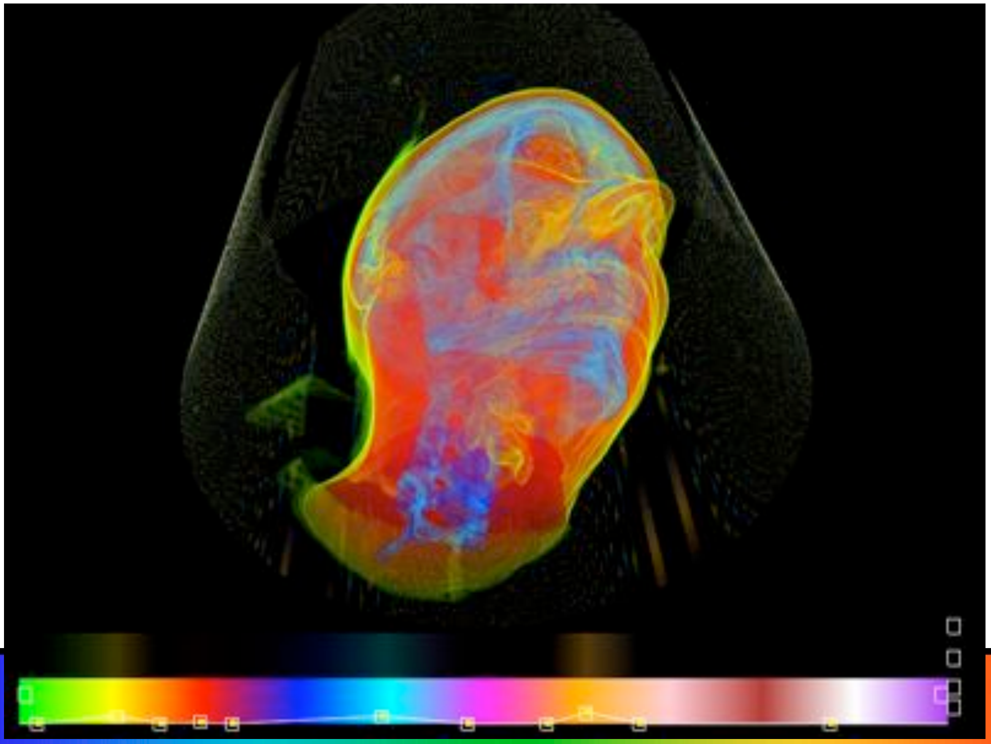
- Image manipulation, creation, and format conversion utility
  - Montage, annotate, size, filter, crop, color modifications
- Command line interface
- (Programming interface)
- Unix, Mac OS X, Windows
- About
  - http://www.imagemagick.org

```
% convert \( font_1.gif font_2.gif font_3.gif +append \) \
%         \( font_4.gif font_5.gif font_6.gif +append \) \
%         \( font_7.gif font_8.gif font_9.gif +append \) \
%         \( -size 32x32 xc:none  font_0.gif +append \) \
%         -background none -append   append_array.gif
```

When to consider custom solutions

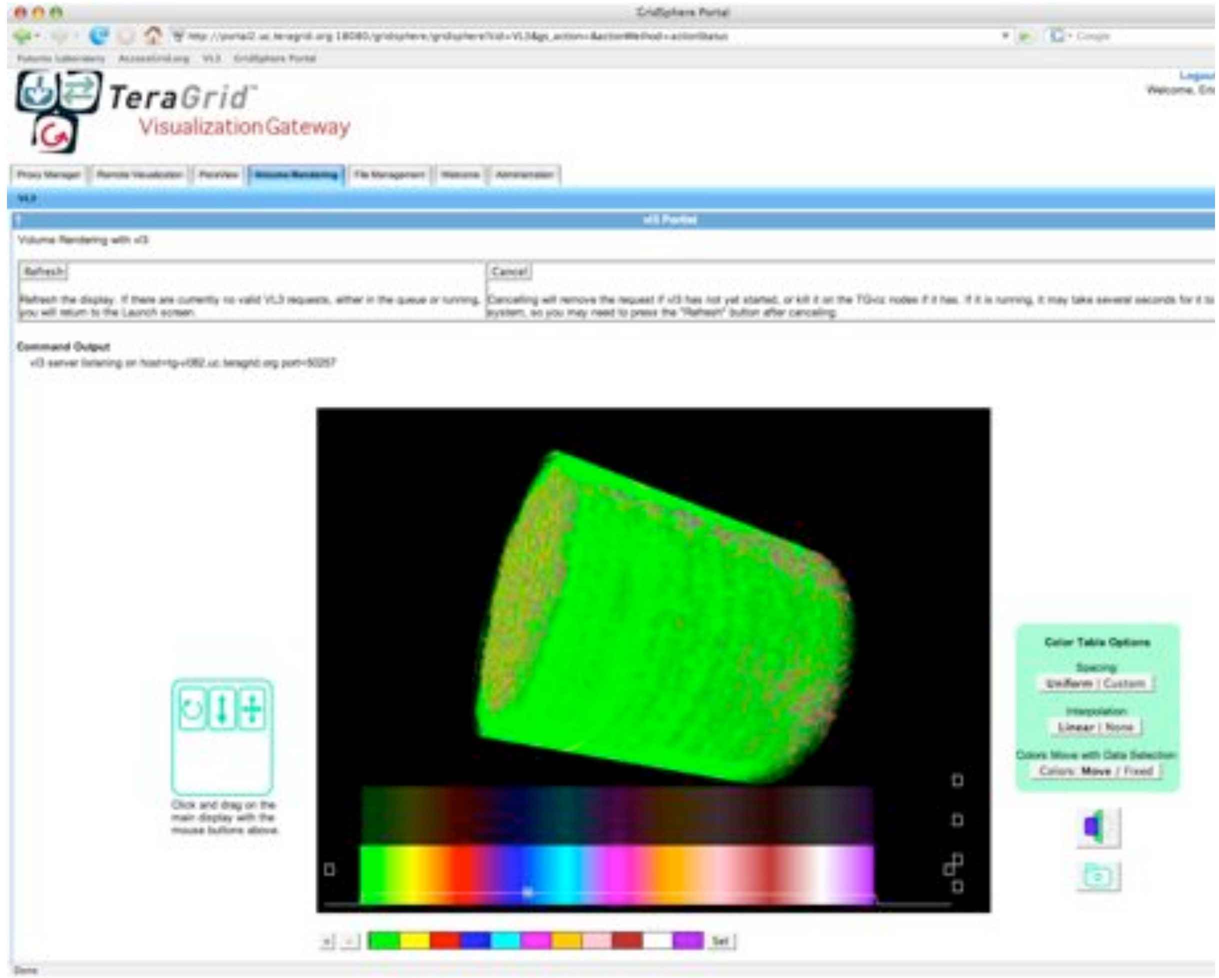
- Issues
  - Performance
  - Algorithms
  - Formats
  - Hybrids
- Considerations
  - Cost-benefit
  - Range of available solutions
  - Available tools
  - Portability
- Tool-chain

- Parallel Volume Renderer (vl3)
  - Hardware acceleration support
  - Composited volume can be streamed to remote users
  - Remote mouse and keyboard interaction
  - Can publish itself to an Access Grid venue
    - *Collaborators double-click to join the session.*



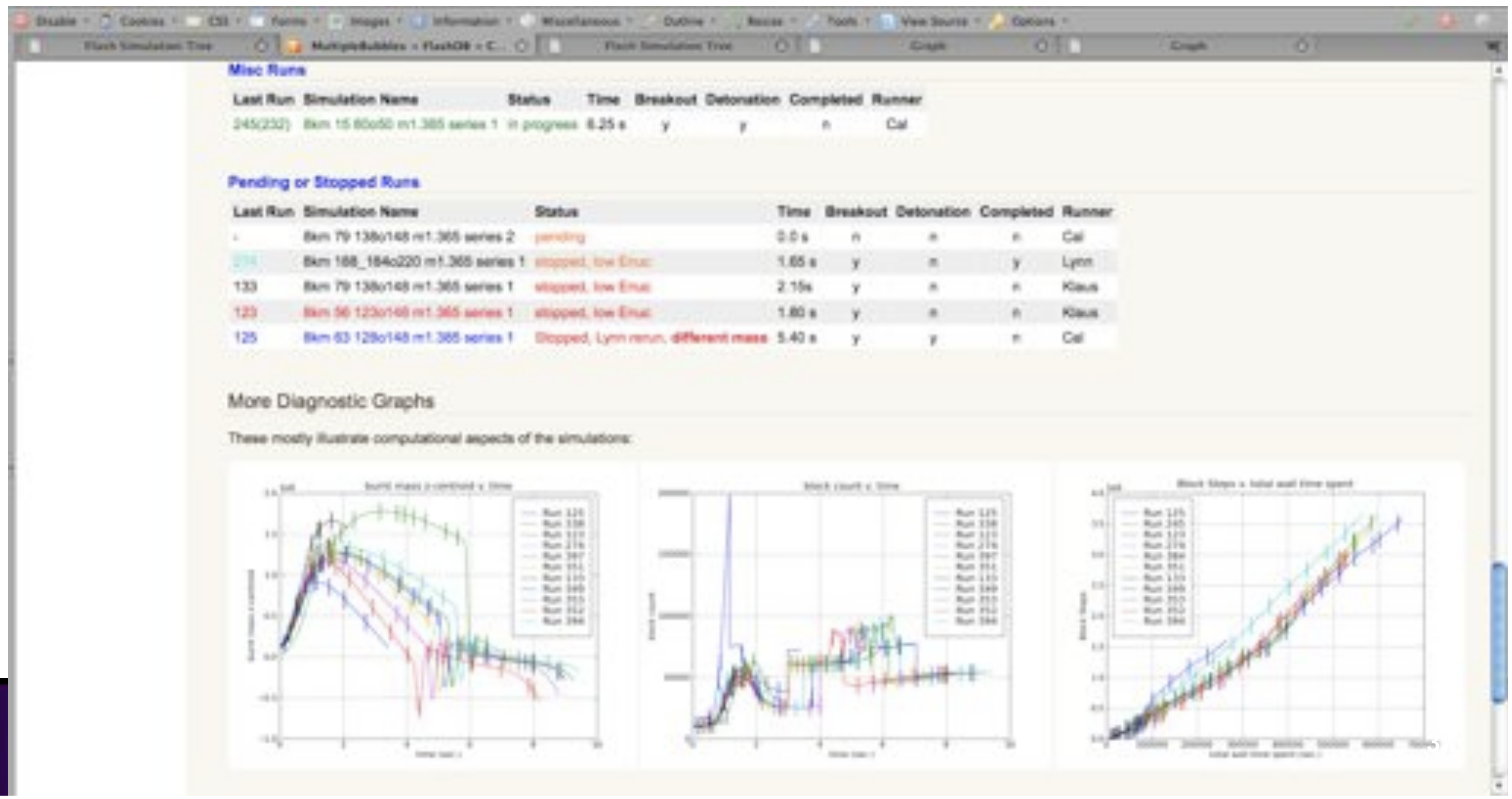
Collaboration

- Portals
  - Shared infrastructure
  - Domain tailored
  - Web-based
  - Community-based



Collaboration

- Dynamically Generated Notebooks
  - Agents pull from pipeline
  - Users interact through collaborative annotations
  - Users steer (configure) agents



Collaboration

Access Grid

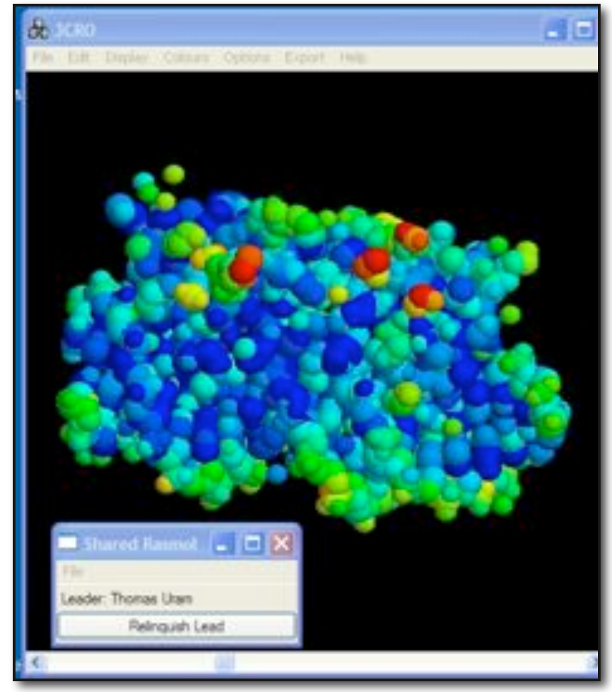
- Model natural collaboration space
  – immersive conferencing environment
- Support collaborative work
  – familiar working environment
  – cooperative use of domain-specific applicatio
- Scalable solution
  – collaboration between groups
  – participation from a laptop or desktop
- Secure
  – confidentiality of collaboration content
  – restrict access to select participants

Collaboration

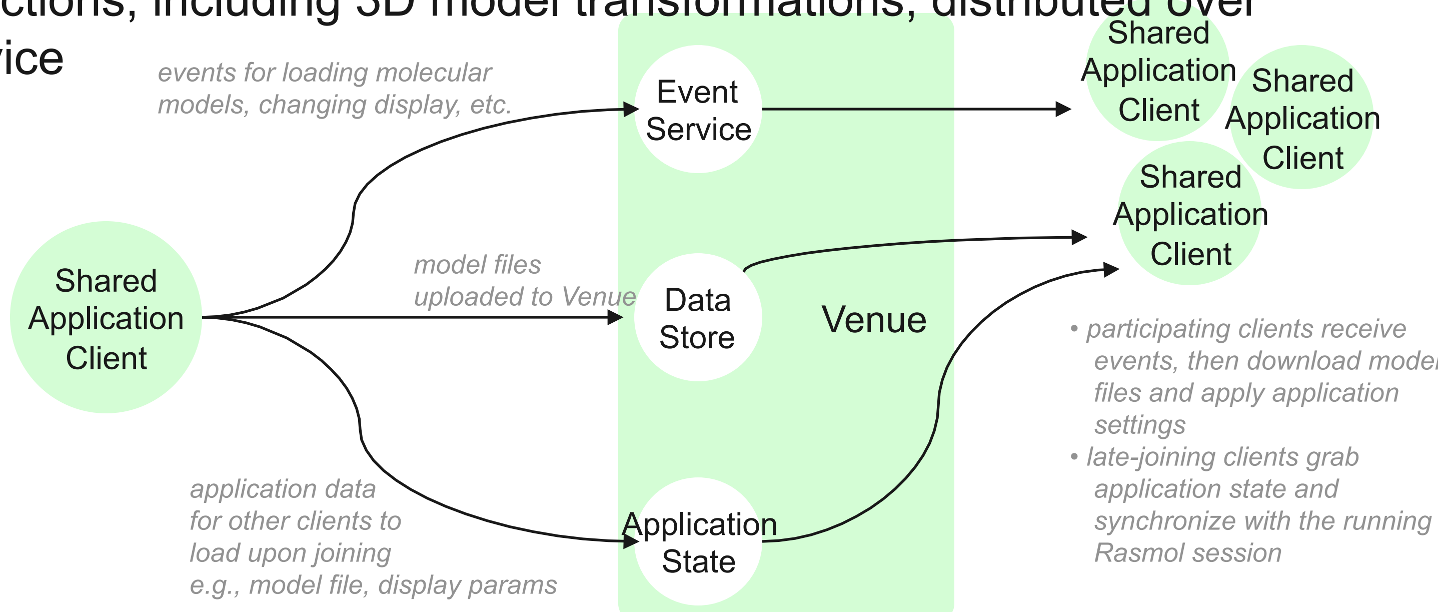
Access Grid Environment

- Venue Server
- Venue Client
- Development Toolkit

Collaboration



- Use Case: Shared RasMol
  - Distributed inspection of 3D molecular models by
  - biologists
- Implementation
  - Model file stored in AG Venue
  - Model filename, display parameters, and
  - transformation matrix stored in Venue application state
  - App interactions, including 3D model transformations, distributed over event service



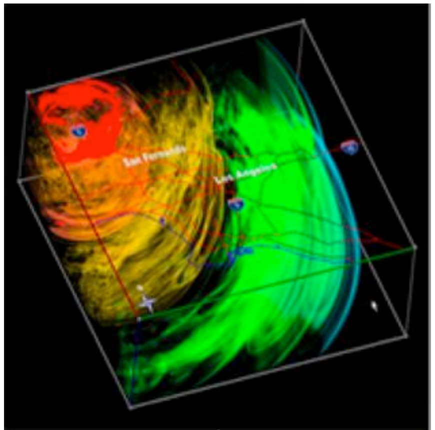
Argonne NATIONAL LABORATORY

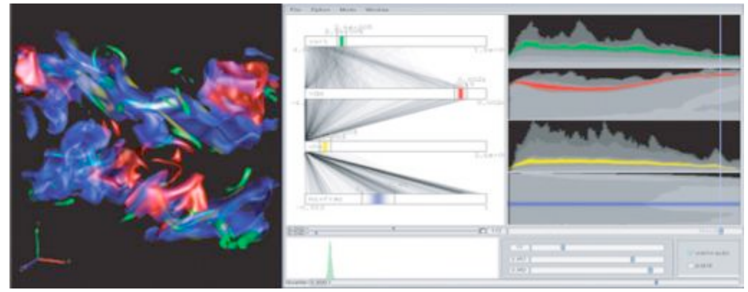
Collaboration

Matlab	VisIt	Paraview
vmd	Vis5D	vapor
Rasmol	VisTrails	R

Ping

- Think about your application and the steps you take from start to finish
- What steps to you take to guide design / setup / config of your NEXT simulation?
- Provenance?
  - Do you need to return to data, rerun simulations, track conditions and state of code?
  - Is this a big or little problem for you?
  - What about hypothesis tracking?
- Data management
  - Connecting: code, setup / config, input, output
  - Moving & tracking: results
  - Comparing runs





Interactivity, Interfaces, and Tools

Remote Visualization

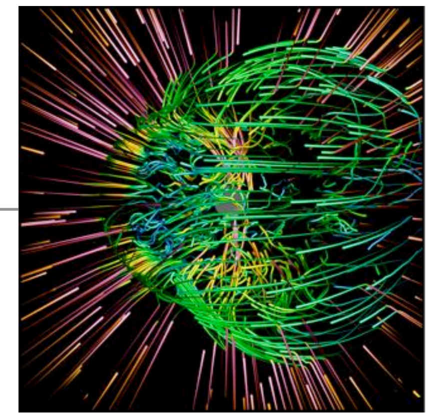
User Interfaces for Multi-Resolution Volumes

Fundamental Algorithms

In Situ Visualization

Multivariate and Multidim. Visualization

Vector Field Visualization



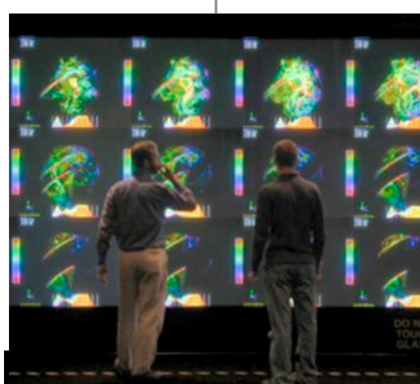
Architectures for Visualization

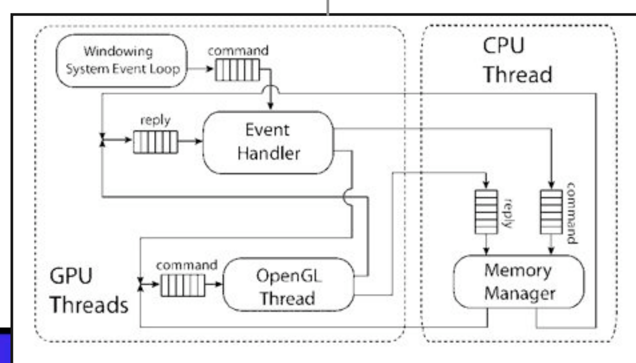
Distributed Visualization

Hardware Acceleration

Extreme-Scale Visualization

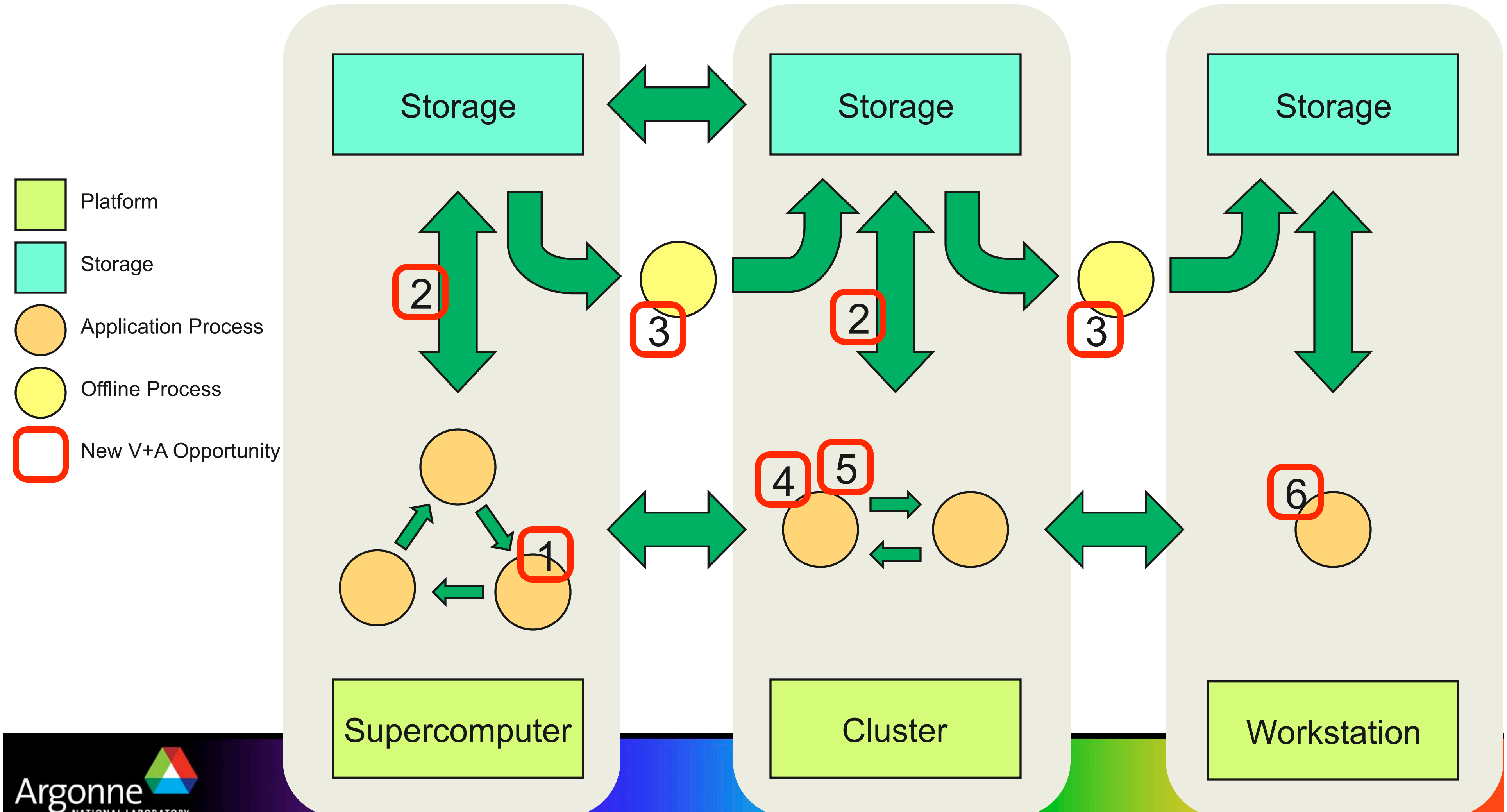
Integration with Production Tools





Possible Visualization and Analysis Points

1. In situ analysis, filtering, reduction
2. Embedded in readers & writers
3. Modified movers
4. Real-time co-analysis
5. Cluster post-processing
6. Real-time multi-stage co-analysis

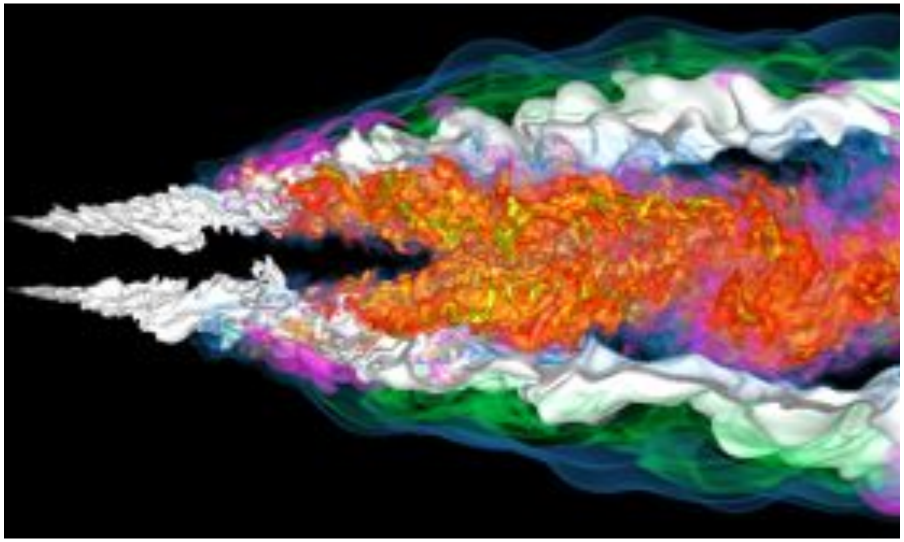


In situ analysis and data reduction

- Incorporate analysis routines into the simulation code
  - operate on data while it is still in memory
- Potential for significant reduction the I/O demands
  - application scientist identifies features of interest
  - compress data of less interest

Here, the feature of interest is the mixture fraction with an iso-value of 0.2 (white surface). Colored regions are a volume rendering of the HO2 variable (data courtesy J. Chen).

By compressing data more aggressively the further it is from this surface, we can attain a compression ratio of 20-30x while still retaining full fidelity in the vicinity of the surface.



C. Wang, H. Yu, and K.-L. Ma, “Application-driven compression for visualizing large-scale time-varying volume data”, IEEE Computer Graphics and Applications, accepted for publication.

Ping

- In situ: what would you put into the analysis slot?
  - Existing synergies: overlap, partial analysis?
- How much time would you willingly trade into analysis -- perhaps losing from simulation time?
- Salience:
  - Could you design / select a filter that robustly identifies something of startling interest?
  - Do you primarily look for things you are expecting -- or does serendipity play a significant role in your process?

Summary

- Visual data analysis fills many roles in the scientific pipeline
- Many sophisticated tools are available to you now
- Many clueful visual representations and efficient algorithms
  - Explore!
- Confused about which to use?
  - Try VisIt and ParaView
  - Keep your eye on VisTrails
- Continually consider and assess your needs in terms of
  - Collaboration support tools and environments
  - Data management support tools
  - Your workflow
  - The end-to-end computational science pipeline
  - Consider these requirements in your proposal budgets!

The End

- Thanks to
  - Mike Papka
  - Rick Stevens
  - Tom Uram
  - Joe Insley
  - Rob Ross
  - Pete Beckman
  - Katherine Riley