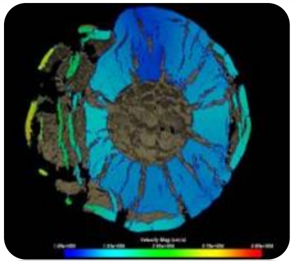
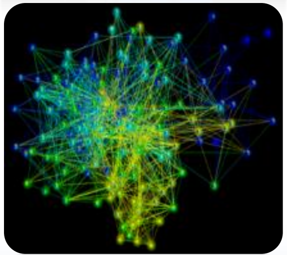
Kitware

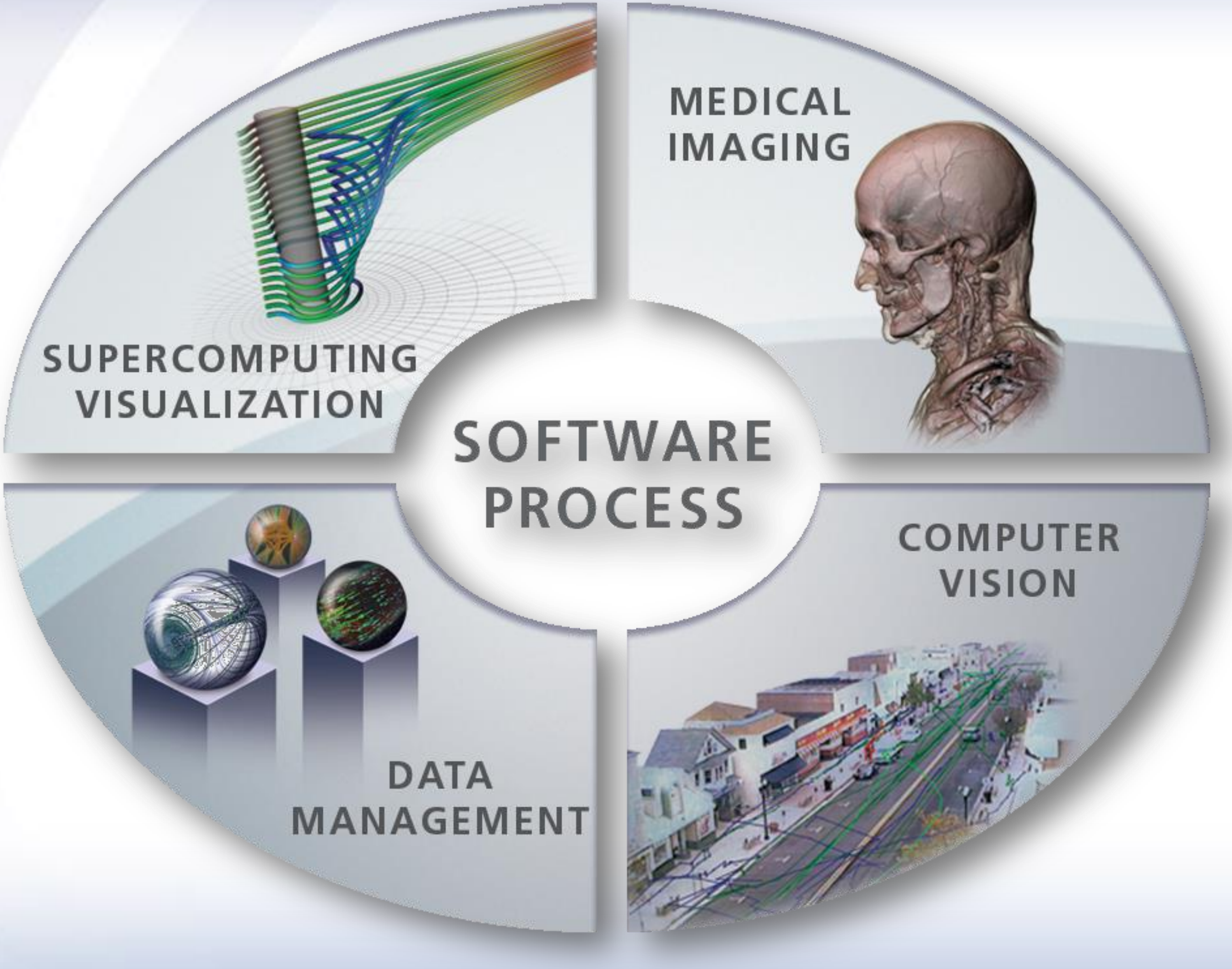
Big Data Analysis and Visualization with Open Source Tools





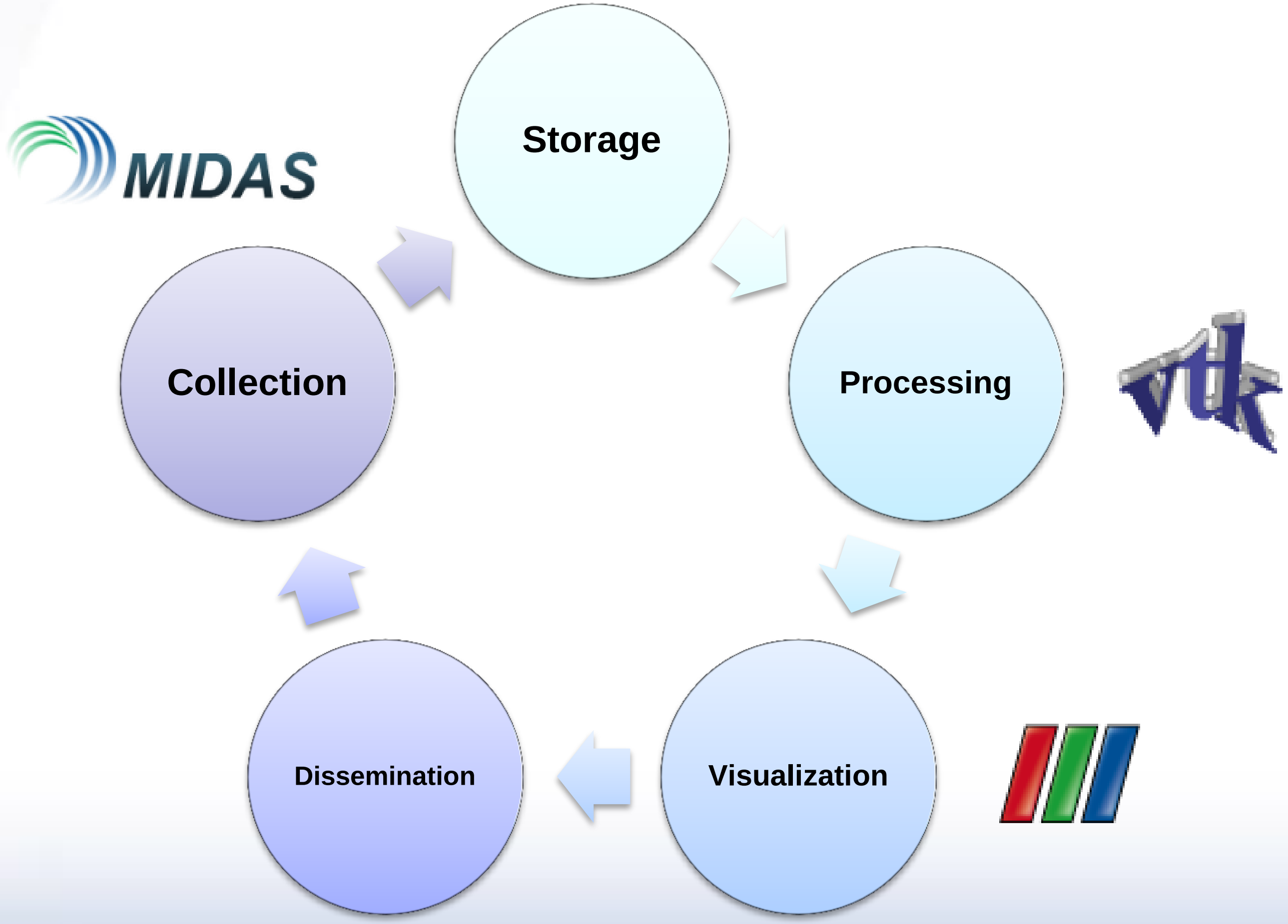
Julien Jomier, Jeff Baumes and Joachim Pouderoux
julien.jomier@kitware.com

Big data is everywhere



Big Data Challenges

- **Parallel** (large data) vs. **Distributed** (many data)



Storage

Processing

Visualization

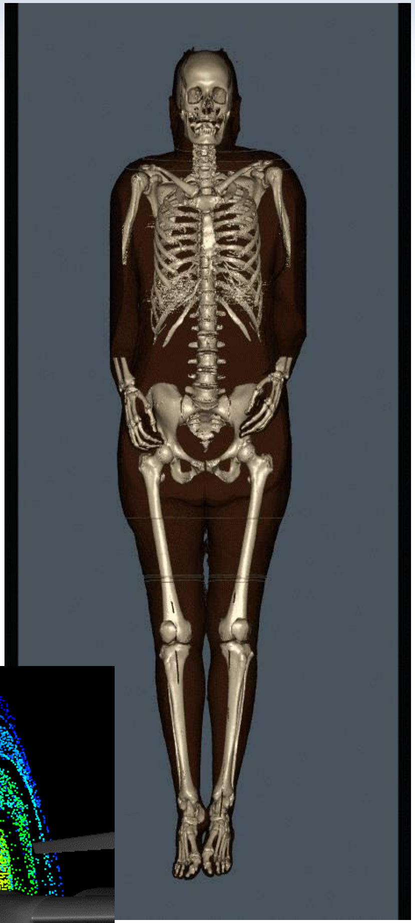
Dissemination

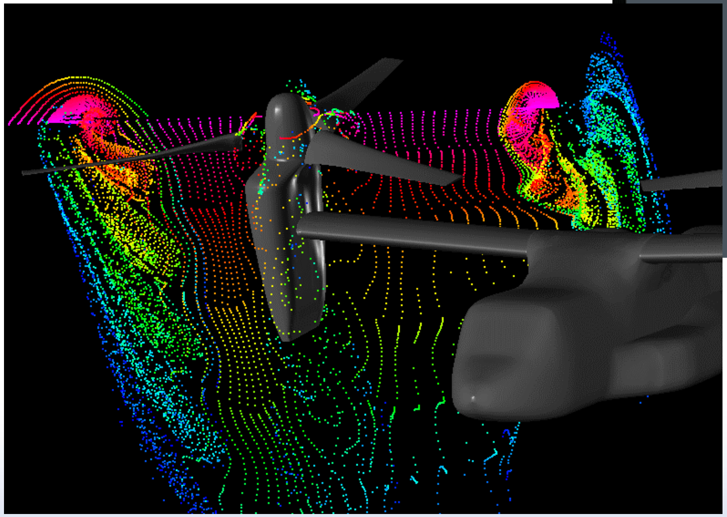
Collection

“Big” is relative

- Visible Human Project (1986)
  - Visible Woman CT Data
  - 870 MBytes 1734 Slices at 512x512x2

- Bell-Boeing V-2 2 tiltrotor
  - 140 GBytes





Big Data Facts

- **Storage** is **cheap**
  - but fast disk access remains expensive
- **Moving data** across networks is the **bottleneck**
- **Most clusters** are **CPU-based** or **GPGPU** based
  - not necessarily optimal for visualization
- **Necessity for parallel/distributed computing**
  - processor speed is getting better

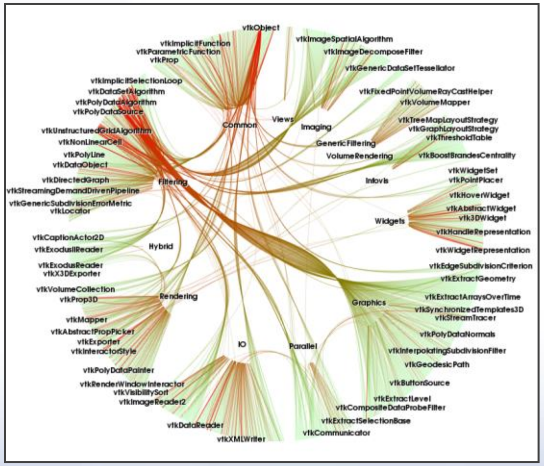
Plan

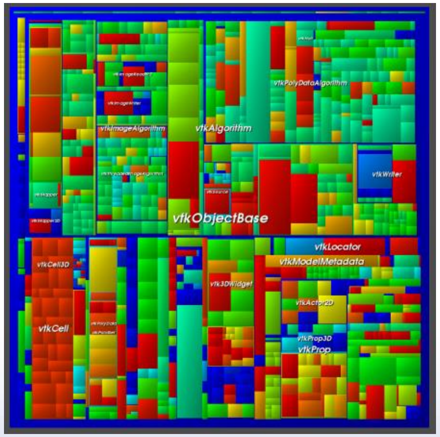
- **Post-processing**
  - VTK/ParaView
- **Co-Processing/In-Situ**
  - Catalyst
- **Informatics**
  - Tangelo, Visomics
- **Computational Chemistry**
  - Avogadro, MongChem
- **Web Visualization**
  - vtkWeb/ParaViewWeb

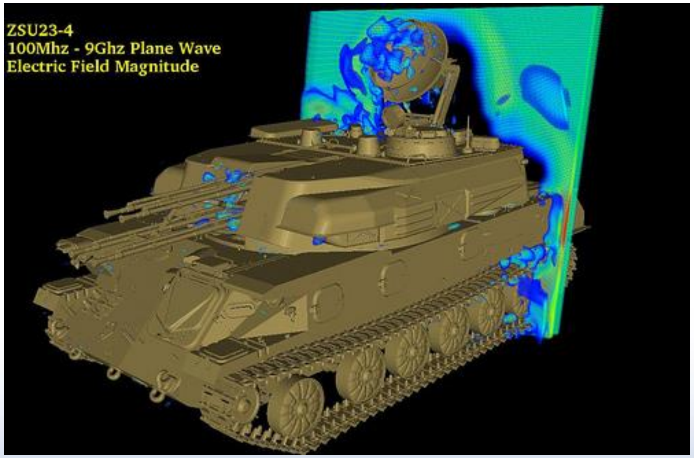
VTK/PARAVIEW

The Visualization Toolkit (VTK)

- www.vtk.org
- Started in 1993 at GE
- Visualization Library
    - Written in C++ (+5.5 million LOC)
    - Automatic binding for Java, TCL, Python
    - Portable by design: Linux, Windows, Mac OSX, Solaris…
- Very active community: 4000+ users

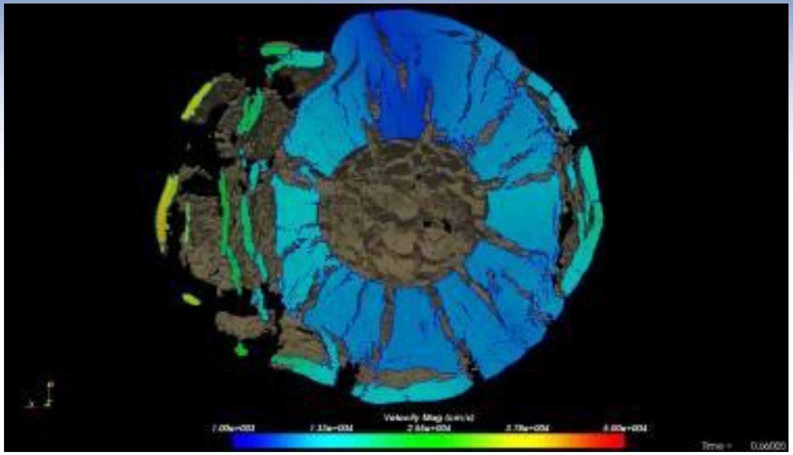




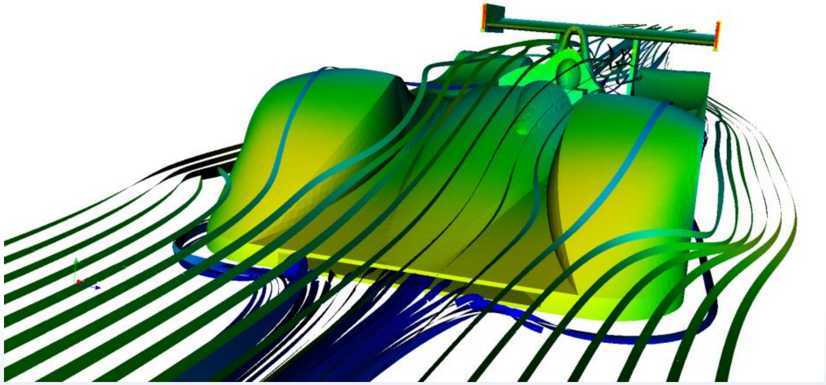


ParaView

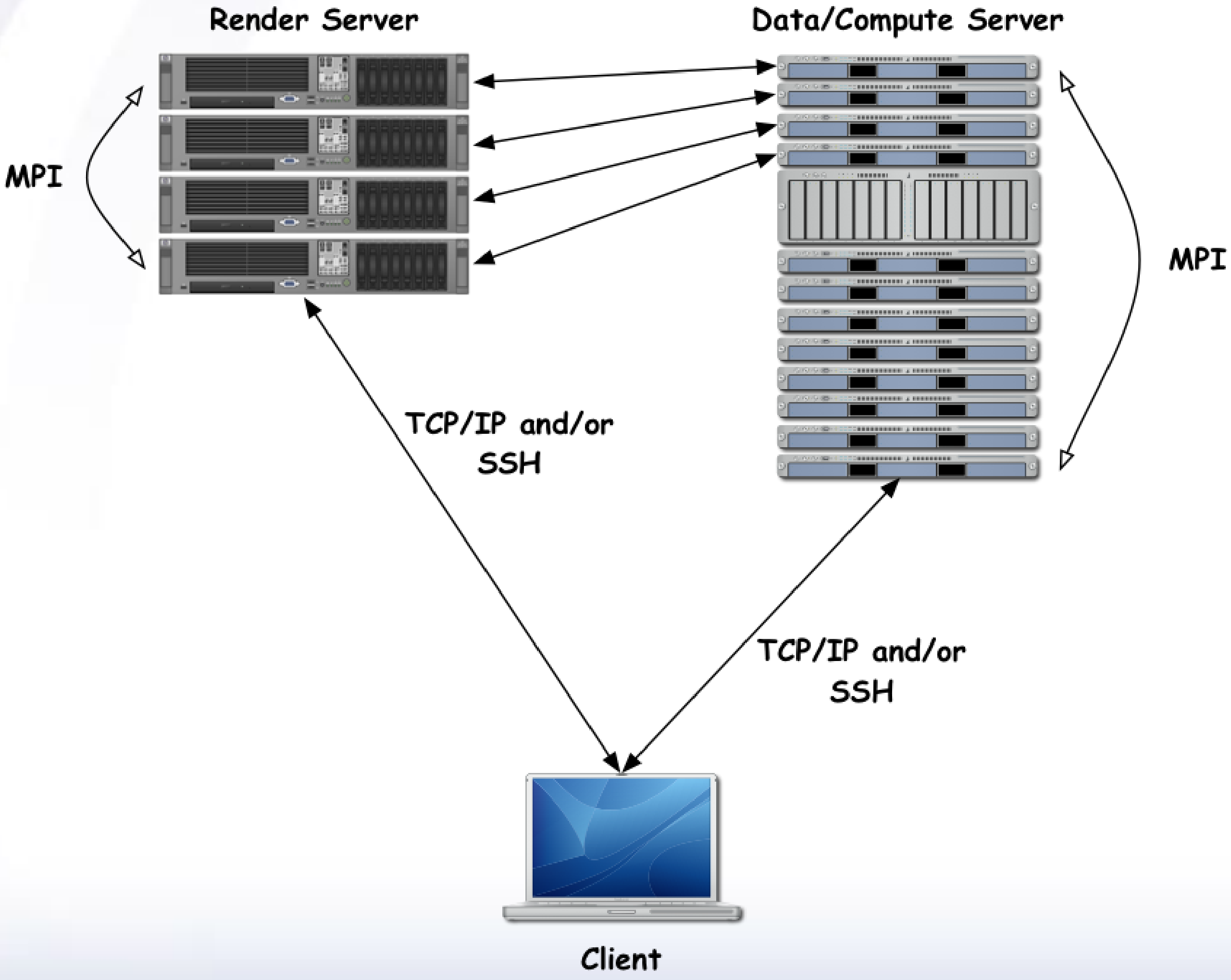
- www.paraview.org
- OpenSource (BSD)
- Based on VTK
- C++/Qt
- Python support
- Very active community (HPC wire award)
- Multi-core support (MPI)
- Co-Processing (in-situ)
- More than 50 news readers
- Visit plugins are supported
- User’s guide online



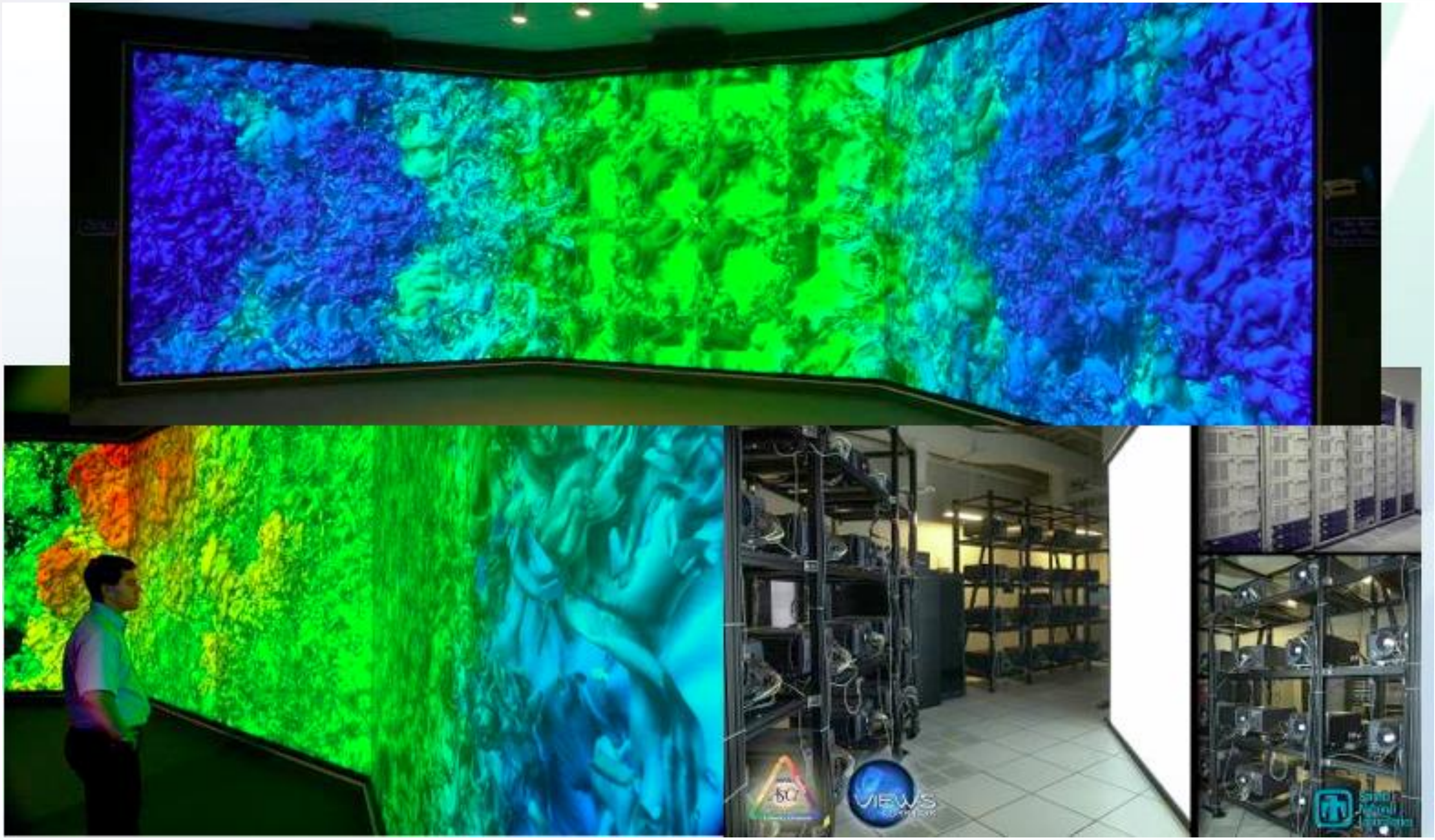
1 billion cell asteroid
detonation simulation



ParaView Architecture



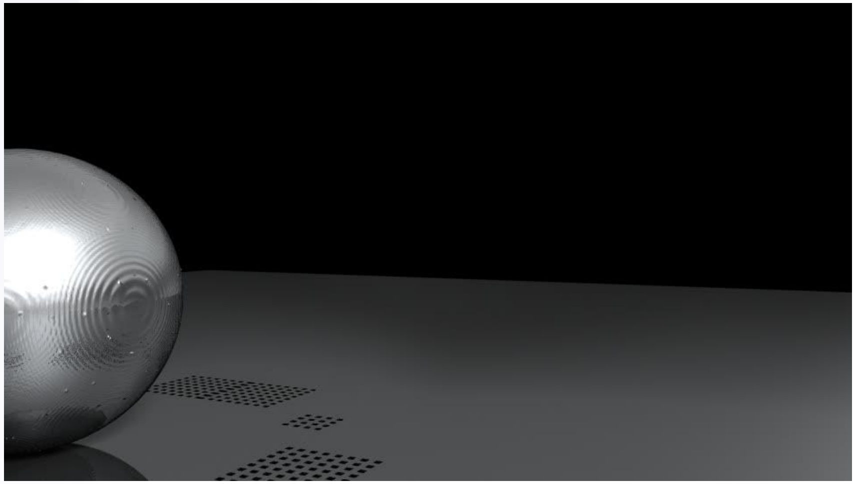
Support for Large Displays



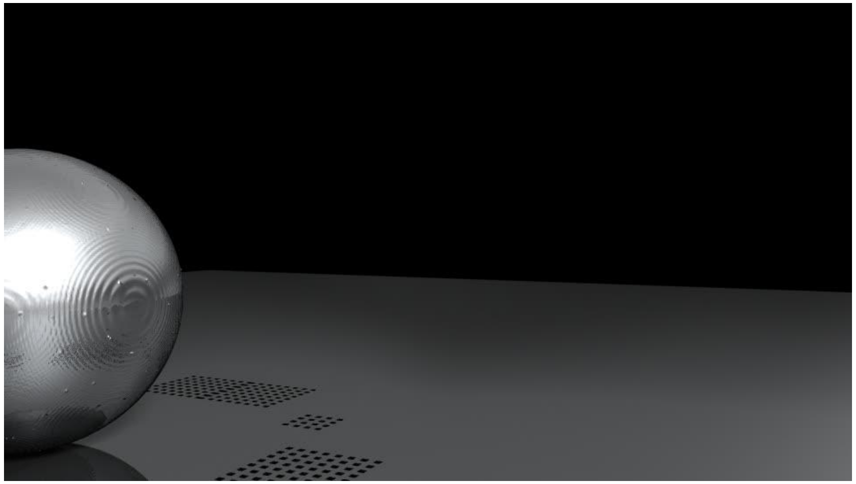
CATALYST

In-Situ : Access to More/Richer Data

Post-Processing
(every 100 time steps)

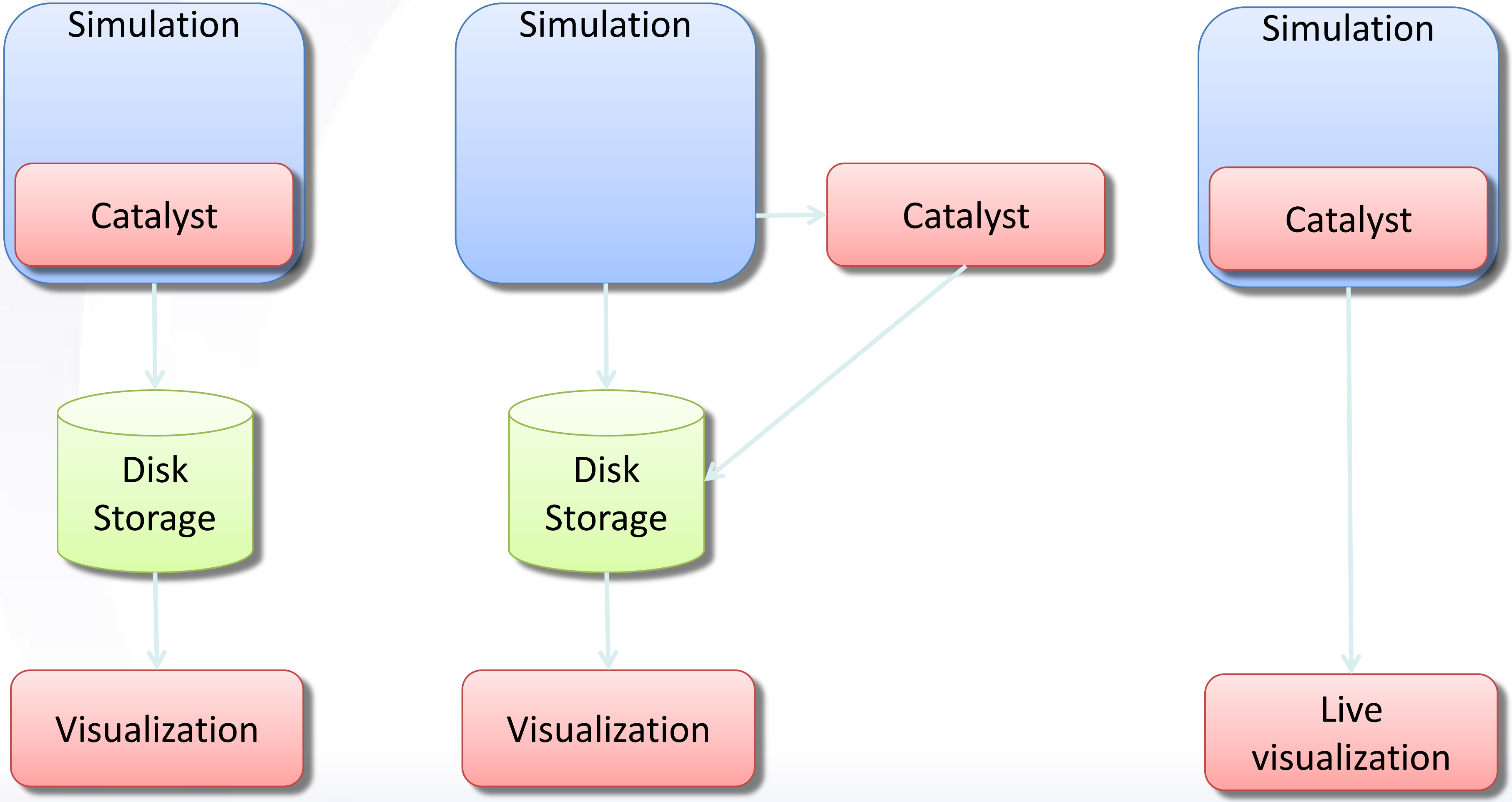


In-situ
(every time step)



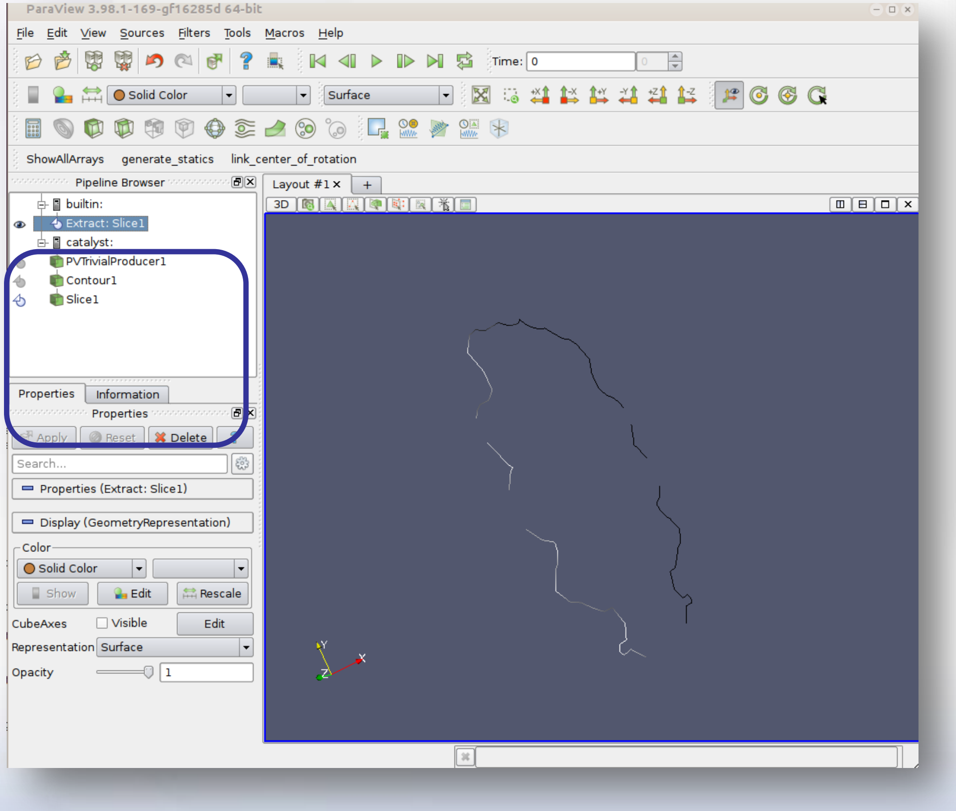
Note: Reflections and shadows added in post-processing

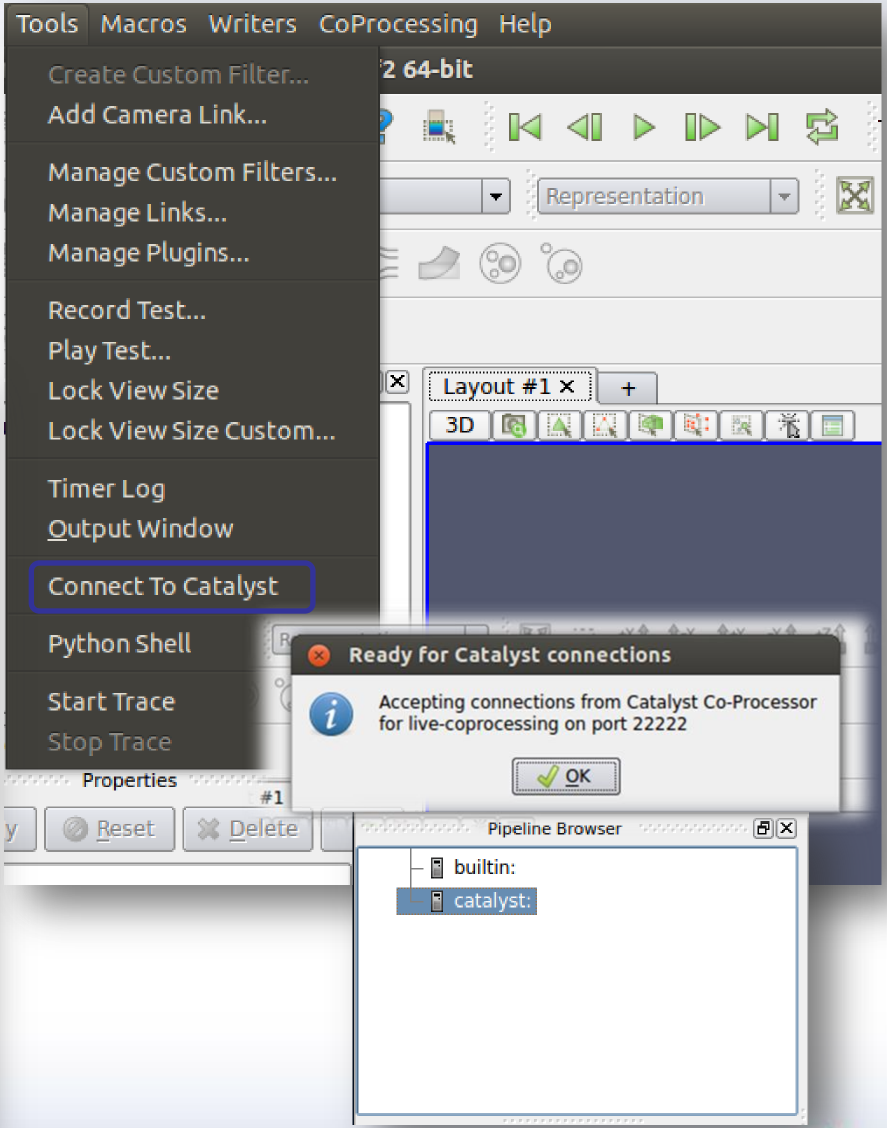
In situ Analysis & Visualization



Connecting to Live InSitu Visualization

- ParaView client connects to Catalyst (acting as a server to receive live data)

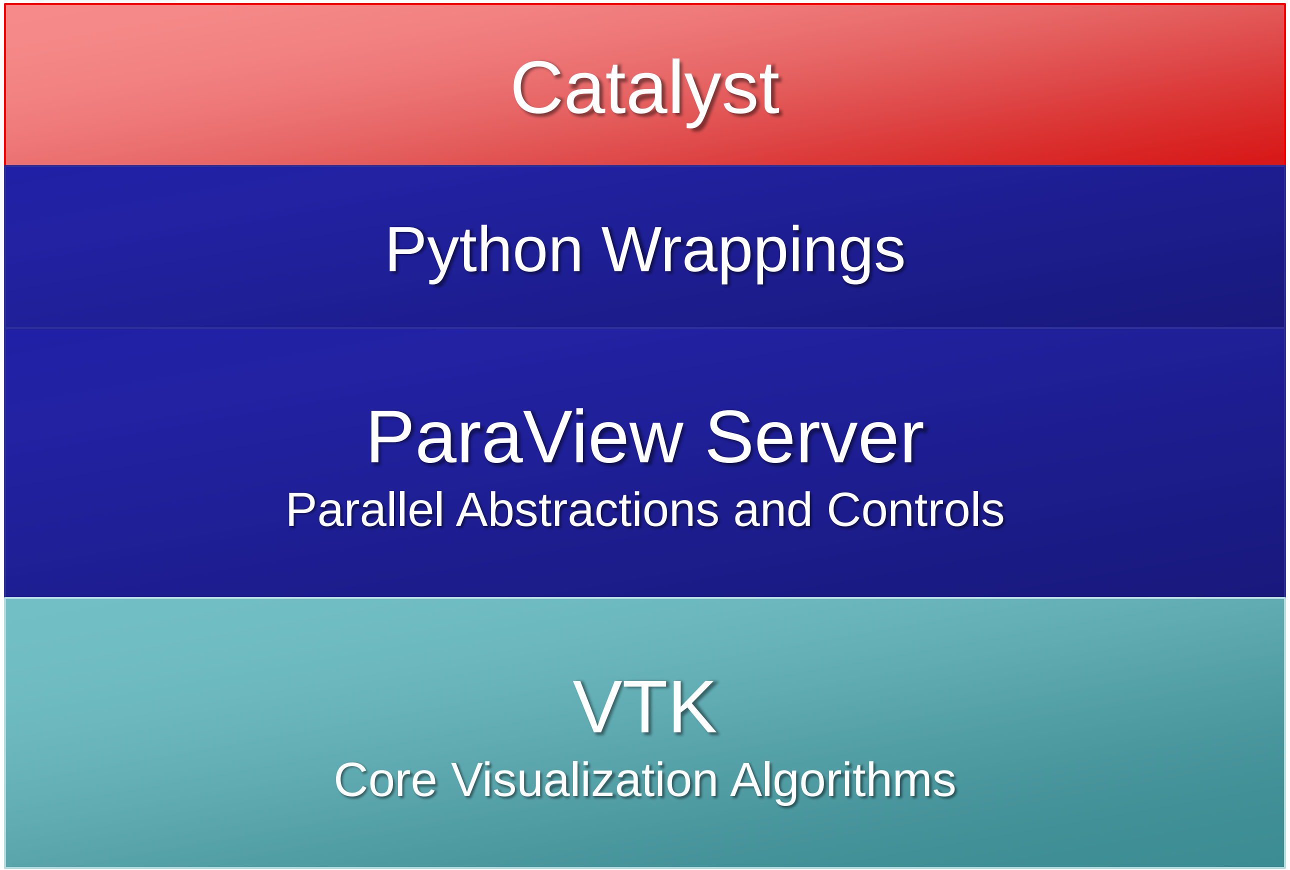




Ready for Catalyst connections

Accepting connections from Catalyst Co-Processor for live-coprocessing on port 22222

Catalyst Architecture



Catalyst

Python Wrappings

ParaView Server

Parallel Abstractions and Controls

VTK

Core Visualization Algorithms

TANGELO/VISOMICS

Tangelo

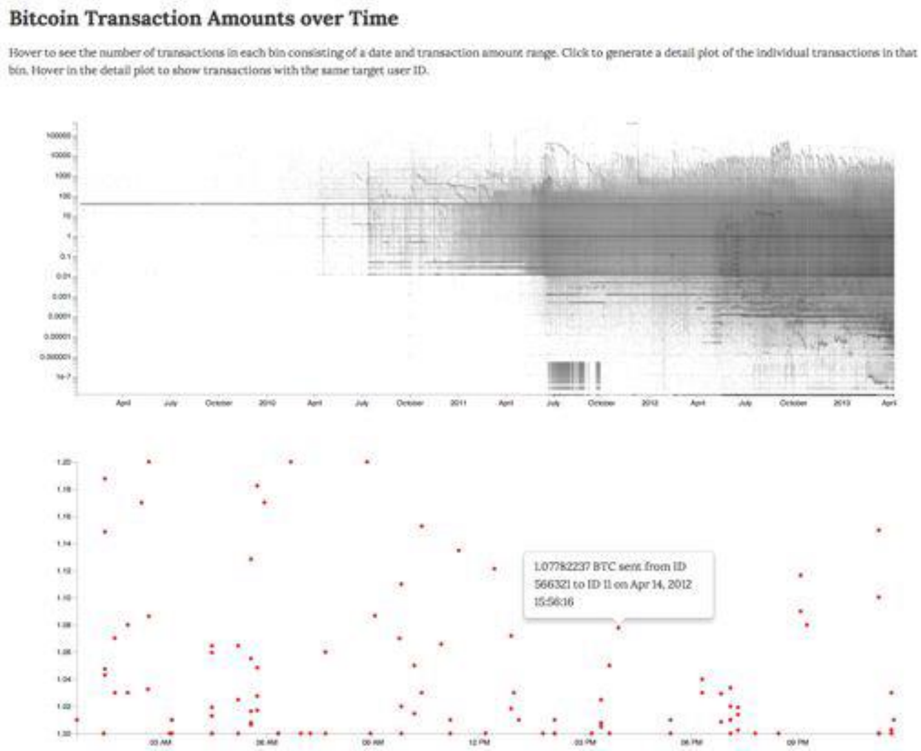
- Python web framework built on CherryPy
- Flexible HTML5 web server architecture
- Developed with a clean separation
  - Application in HTML, JavaScript, CSS
  - Service in pure Python (+ wrapped C/C++)
- Packages several other frameworks too
  - Bootstrap, D3, Vega,MongoDB
- Making web apps easier to develop/deploy
- htttp://tangelo.kitware.com

Tangelo

- Python for server side, native web clients
- Easily add new services (single .py file)
  - Use RESTful API
  - JSON delivery of data
  - Full power of Python
- Rapid prototyping

Bitcoin Analysis

- Uses bitcoin blockchain
  – Individual transactions
- Intensity histogram with transaction volume in date/amount ranges
- Detail plot with individual transactions
- Anomaly search
  - – Theft detection
- Study large scale behavior over time



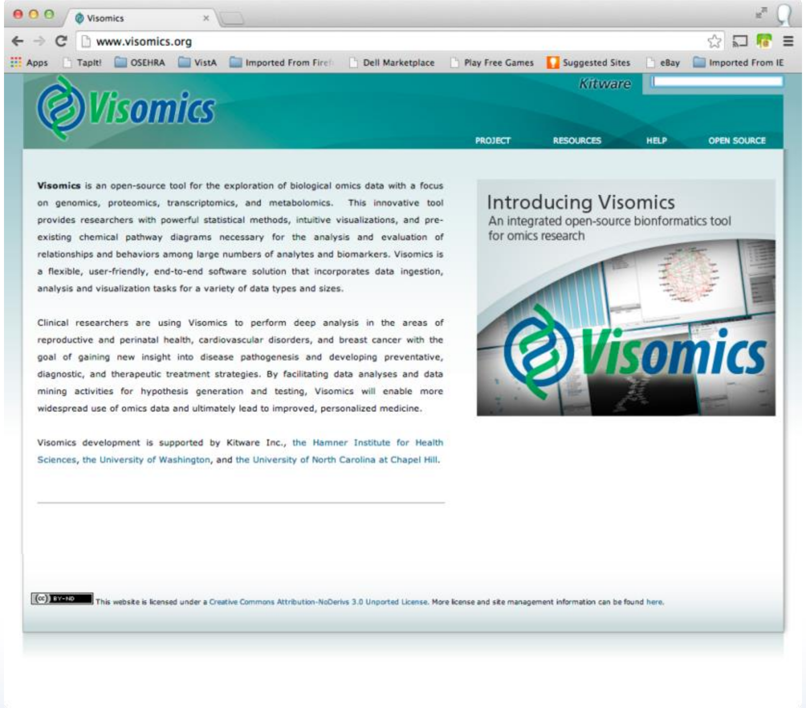
Bitcoin Transaction Amounts over Time

Hover to see the number of transactions in each bin consisting of a date and transaction amount range. Click to generate a detail plot of the individual transactions in that bin. Hover in the detail plot to show transactions with the same target user ID.

1.07782237 BTC sent from ID 566321 to ID 11 on Apr 14, 2012 15:56:16

Visomics

- Visomics is an extensible, open source application for research in Omics applications.
- http://www.visomics.org/



Visomics is an open-source tool for the exploration of biological omics data with a focus on genomics, proteomics, transcriptomics, and metabolomics. This innovative tool provides researchers with powerful statistical methods, intuitive visualizations, and pre-existing chemical pathway diagrams necessary for the analysis and evaluation of relationships and behaviors among large numbers of analytes and biomarkers. Visomics is a flexible, user-friendly, end-to-end software solution that incorporates data ingestion, analysis and visualization tasks for a variety of data types and sizes.

Clinical researchers are using Visomics to perform deep analysis in the areas of reproductive and perinatal health, cardiovascular disorders, and breast cancer with the goal of gaining new insight into disease pathogenesis and developing preventative, diagnostic, and therapeutic treatment strategies. By facilitating data analyses and data mining activities for hypothesis generation and testing, Visomics will enable more widespread use of omics data and ultimately lead to improved, personalized medicine.

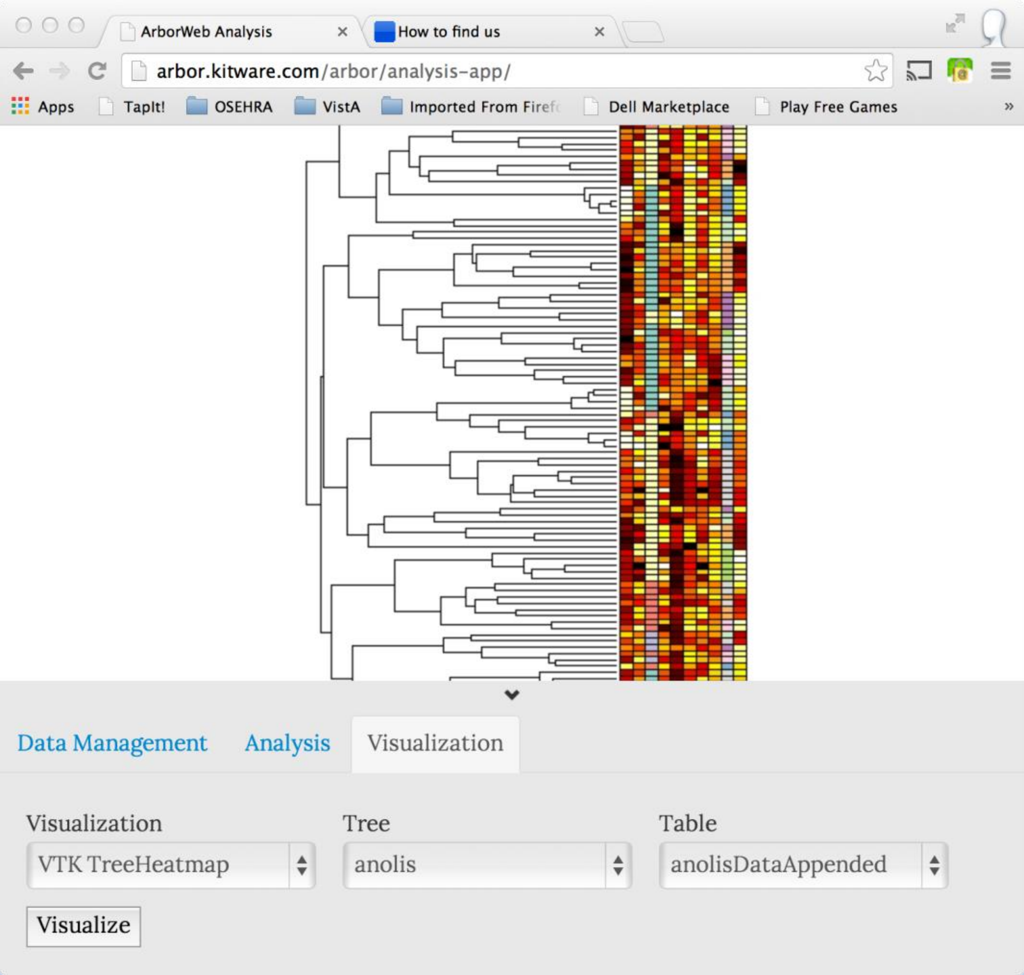
Visomics development is supported by Kitware Inc., the Hamner Institute for Health Sciences, the University of Washington, and the University of North Carolina at Chapel Hill.

Introducing Visomics

An integrated open-source bionformatics tool for omics research

Arbor

- Brining Visomics to the web



Visualization: VTK TreeHeatmap

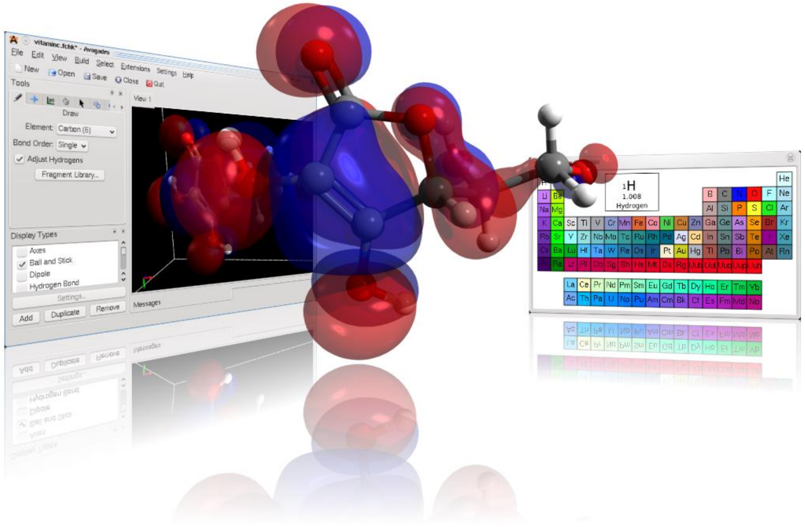
Tree: anolis

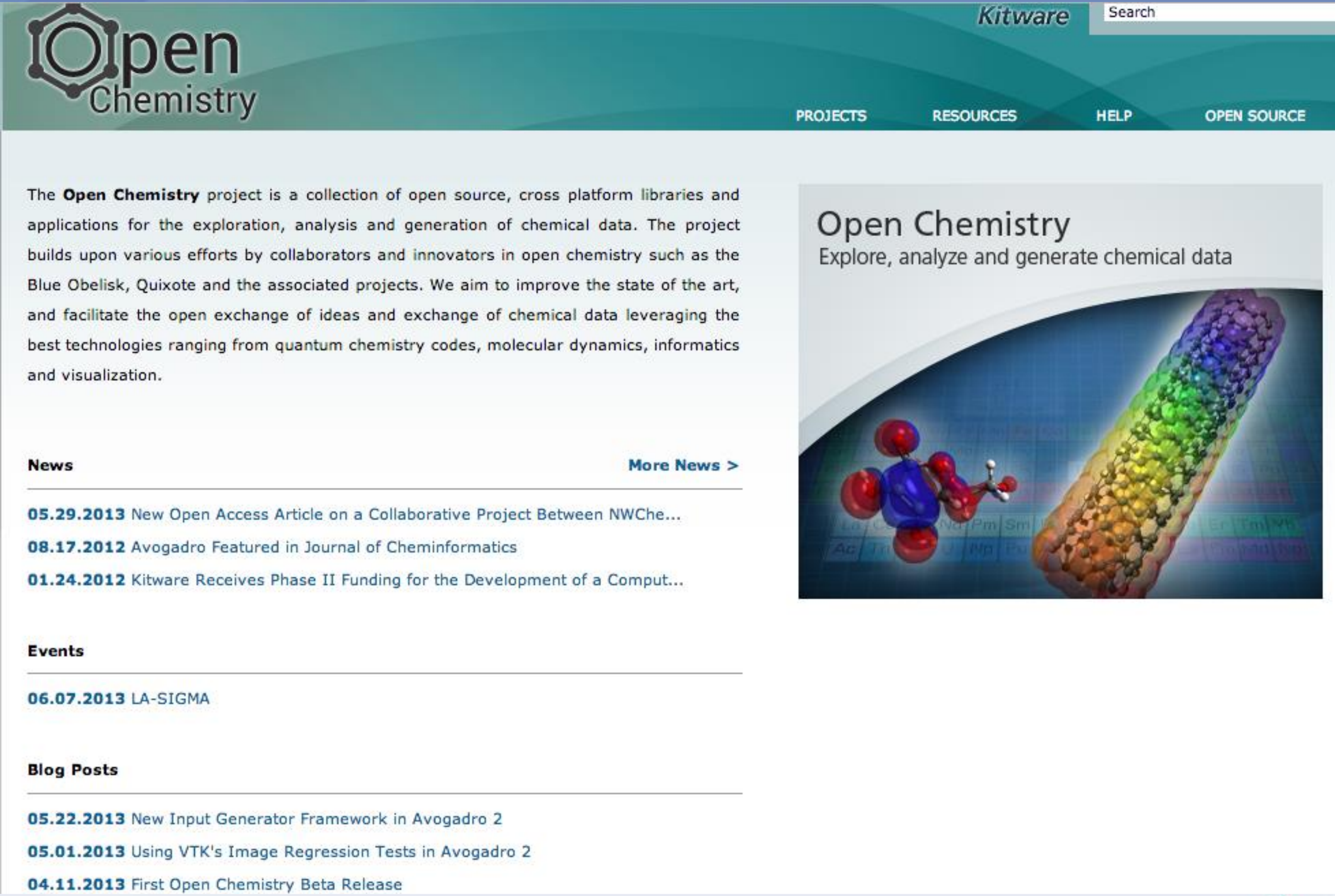
Table: anolisDataAppended

OPEN CHEMISTRY

Vision for Open Chemistry

- Advancing the state-of-the-art
- Tight integration is needed
  - Computational codes
  - Clusters/supercomputers
  - Data repositories
  - Reduce, reuse, and recycle!
- Facilitate sharing and searching of data
- Embracing data-centric workflows





The **Open Chemistry** project is a collection of open source, cross platform libraries and applications for the exploration, analysis and generation of chemical data. The project builds upon various efforts by collaborators and innovators in open chemistry such as the Blue Obelisk, Quixote and the associated projects. We aim to improve the state of the art, and facilitate the open exchange of ideas and exchange of chemical data leveraging the best technologies ranging from quantum chemistry codes, molecular dynamics, informatics and visualization.

Open Chemistry

Explore, analyze and generate chemical data

News **More News >**

05.29.2013 New Open Access Article on a Collaborative Project Between NWChe...

08.17.2012 Avogadro Featured in Journal of Cheminformatics

01.24.2012 Kitware Receives Phase II Funding for the Development of a Comput...

Events

06.07.2013 LA-SIGMA

Blog Posts

05.22.2013 New Input Generator Framework in Avogadro 2

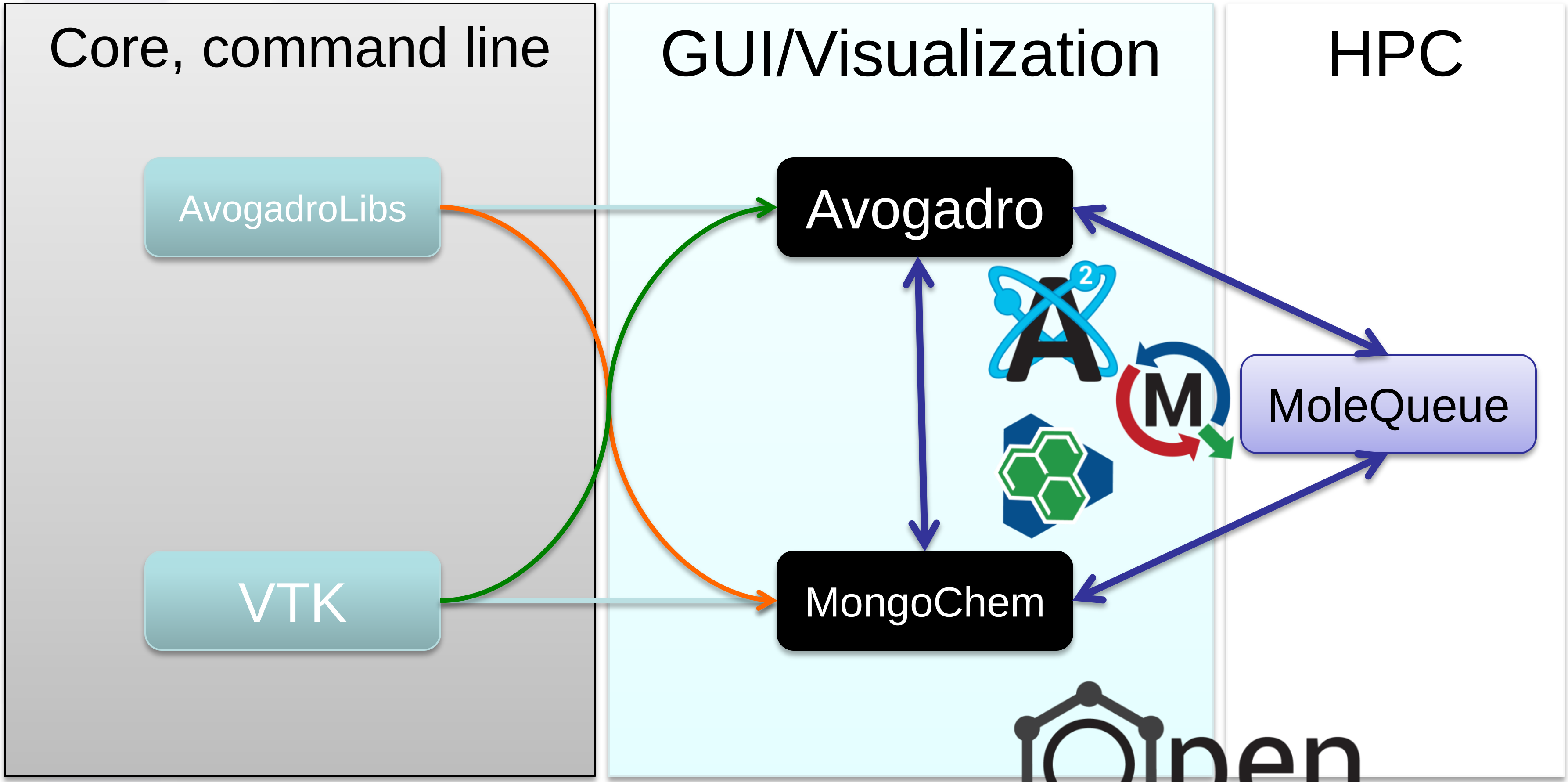
05.01.2013 Using VTK's Image Regression Tests in Avogadro 2

04.11.2013 First Open Chemistry Beta Release

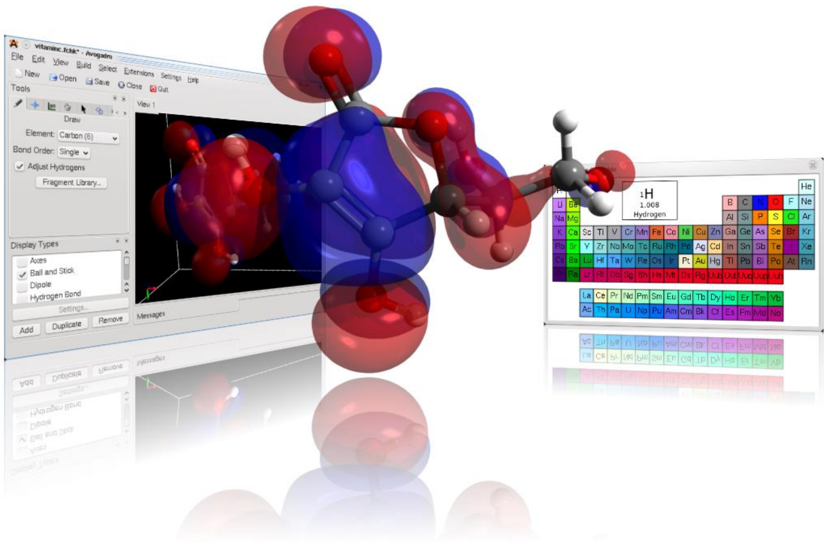
Applications Being Developed

- Three independent applications
- Communication handled with local sockets
- **Avogadro 2:** Structure editing, input generation, output viewing, and analysis
- **MoleQueue:** Running local and remote jobs in standalone programs, and management
- **MongoChem:** Storage of data, searching, entry, and annotation

Project Diagram: Libraries/Apps



Avogadro^2



Element: Carbon (6)

Bond Order: Single

☑ Adjust Hydrogens

□ Axes

☑ Ball and Stick

□ Dipole

□ Hydrogen Bond

- Rewrite of Avogadro
- Split into libraries & application (plugin-based)
- Still one of very few open source **editors**
- Still using Qt, C++, Eigen, OpenGL, and CMake
- Use AvogadroLibs for core data
- Introduces client-server dataflow/patterns
- New, efficient rendering code
- More liberally-licensed: from GPL to BSD

MoleQueue: Job Management

- Tight integration with remote queues
- Integration with databases
  - Retains full log of computational jobs
  - Triggers actions on completion
- Plugin-based system
  - Easy addition of new codes
  - Easy addition of new queue systems
- Provides client API for applications

MongoChem

- A desktop cheminformatics tool
  - Chemical data exploration and analysis
  - Interactive, editable, and searchable database
- Leverages several open-source projects
  - Qt, VTK, MongoDB, Avogadro 2, Open Babel
- Designed to look at many molecules
- Spots patterns, outliers; runs many jobs
- Scales to studies with ~3 million structures

MongoChem Architecture

- Native, cross-platform C++ application built with Qt and Avogadro 2
- Stores chemical data in a NoSQL MongoDB database
- Uses VTK for 2D and 3D dataset visualization

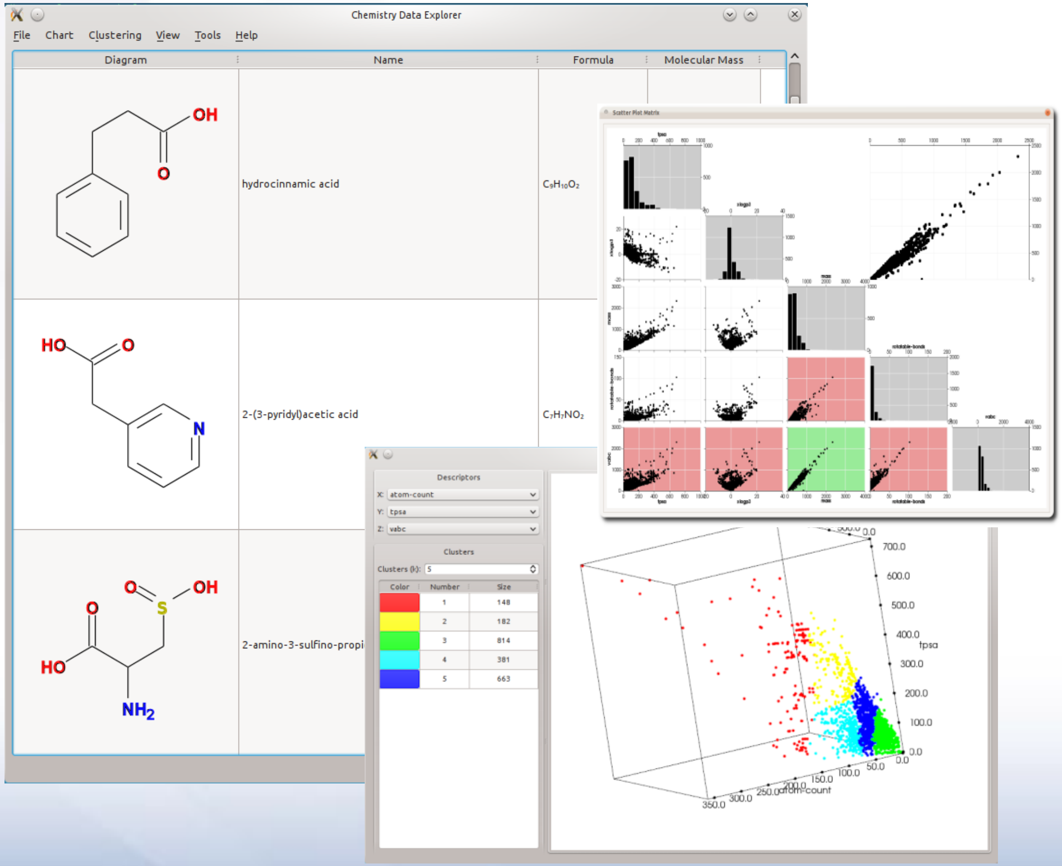


Diagram	Name	Formula	Molecular Mass
	hydrocinnamic acid	$C_9H_{10}O_2$	
	2-(3-pyridyl)acetic acid	$C_7H_7NO_2$	
	2-amino-3-sulfino-propi		

Descriptors

X: atom-count

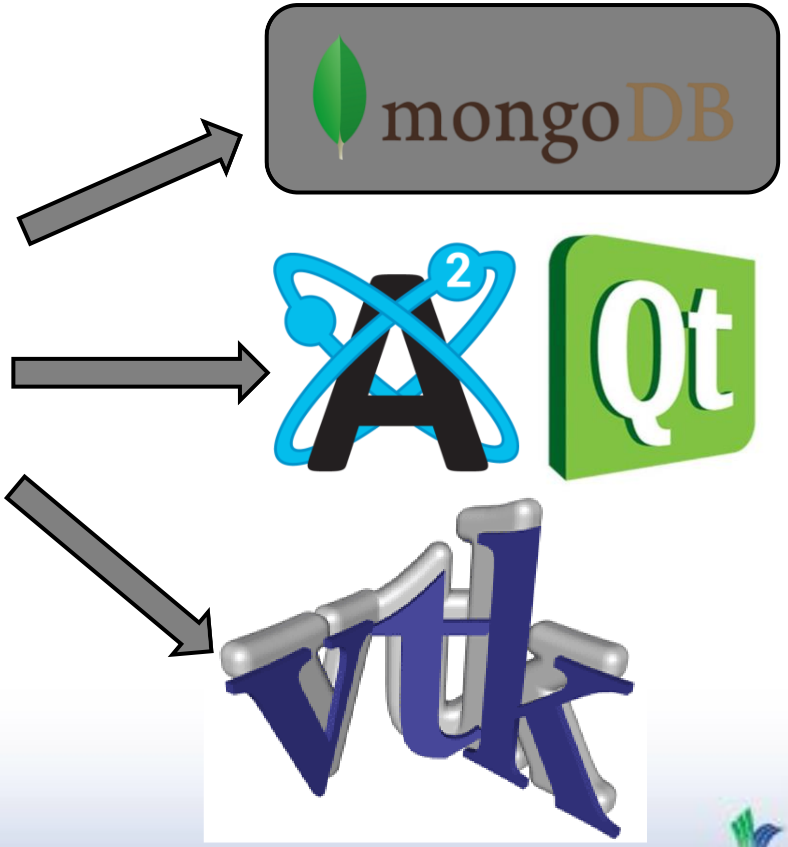
Y: tpsa

Z: vabc

Clusters

Clusters (k): 5

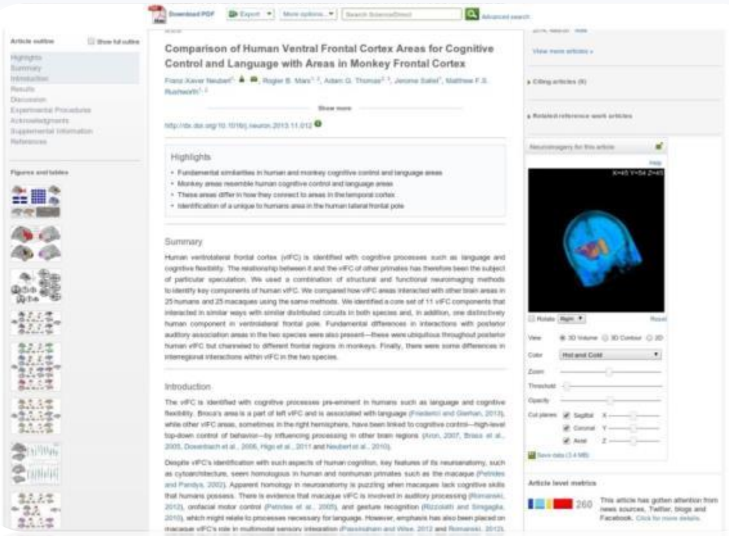
Color	Number	Size
🟥	1	148
🟨	2	182
🟩	3	814
🟦	4	381
🟦	5	663



WEBVIZ

Online Visualization – vtkWeb/ParaViewWeb

- No plugin
- Works on all devices and browsers
- Instant Visualization (fast loading)
- Fully interactive visualization
- **http://www.webviz.org**



Download PDF · Export · More options... · Search ScienceDirect · Advanced search

Article outline · Show full outline

Highlights
Summary
Introduction
Results
Discussion
Experimental Procedures
Acknowledgments
Supplemental Information
References

Figures and tables

Comparison of Human Ventral Frontal Cortex Areas for Cognitive Control and Language with Areas in Monkey Frontal Cortex

Franz Xaver Neubert[1], Rogier B. Mars[1, 2], Adam G. Thomas[2, 3], Jerome Sallet[1], Matthew F.S. Rushworth[1, 2]

Show more

http://dx.doi.org/10.1016/j.neuron.2013.11.012

Highlights

- Fundamental similarities in human and monkey cognitive control and language areas
- Monkey areas resemble human cognitive control and language areas
- These areas differ in how they connect to areas in the temporal cortex
- Identification of a unique to humans area in the human lateral frontal pole

Summary

Human ventrolateral frontal cortex (vlFC) is identified with cognitive processes such as language and cognitive flexibility. The relationship between it and the vlFC of other primates has therefore been the subject of particular speculation. We used a combination of structural and functional neuroimaging methods to identify key components of human vlFC. We compared how vlFC areas interacted with other brain areas in 25 humans and 25 macaques using the same methods. We identified a core set of 11 vlFC components that interacted in similar ways with similar distributed circuits in both species and, in addition, one distinctively human component in ventrolateral frontal pole. Fundamental differences in interactions with posterior auditory association areas in the two species were also present—these were ubiquitous throughout posterior human vlFC but channeled to different frontal regions in monkeys. Finally, there were some differences in interregional interactions within vlFC in the two species.

Introduction

The vlFC is identified with cognitive processes pre-eminent in humans such as language and cognitive flexibility. Broca's area is a part of left vlFC and is associated with language (Friederici and Gierhan, 2013), while other vlFC areas, sometimes in the right hemisphere, have been linked to cognitive control—high-level top-down control of behavior—by influencing processing in other brain regions (Aron, 2007, Brass et al., 2005, Dosenbach et al., 2006, Higo et al., 2011 and Neubert et al., 2010).

Despite vlFC's identification with such aspects of human cognition, key features of its neuroanatomy, such as cytoarchitecture, seem homologous in human and nonhuman primates such as the macaque (Petrides and Pandya, 2002). Apparent homology in neuroanatomy is puzzling when macaques lack cognitive skills that humans possess. There is evidence that macaque vlFC is involved in auditory processing (Romanski, 2012), orofacial motor control (Petrides et al., 2005), and gesture recognition (Rizzolatti and Sinigaglia, 2010), which might relate to processes necessary for language. However, emphasis has also been placed on macaque vlFC's role in multimodal sensory integration (Passingham and Wise, 2012 and Romanski, 2012).

View more articles »

Citing articles (0)

Related reference work articles

Neuroimagery for this article · Help

X=45 Y=54 Z=45

□ Rotate Right · Reset

View ● 3D Volume ○ 3D Contour ○ 2D

Color Hot and Cold

Zoom

Threshold

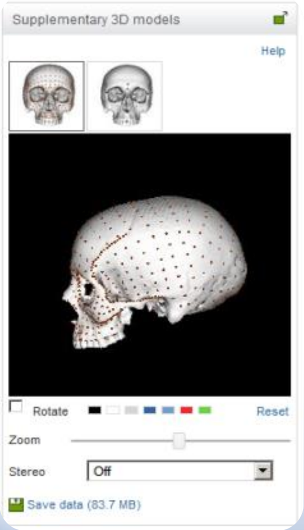
Opacity

Cut planes ☑ Sagittal X ☑ Coronal Y ☑ Axial Z

Save data (3.4 MB)

Article level metrics

260 This article has gotten attention from news sources, Twitter, blogs and Facebook. Click for more details.



Supplementary 3D models

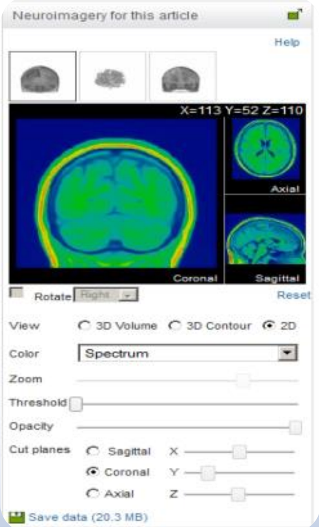
Help

□ Rotate · Reset

Zoom

Stereo Off

Save data (83.7 MB)



Neuroimagery for this article

Help

X=113 Y=52 Z=110

□ Rotate Right · Reset

View ○ 3D Volume ○ 3D Contour ● 2D

Color Spectrum

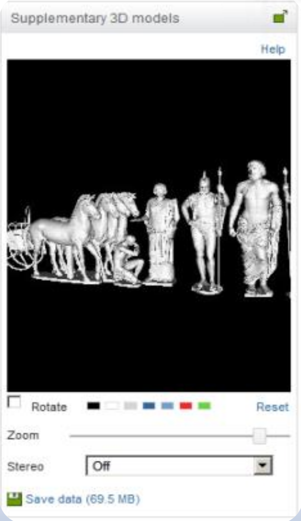
Zoom

Threshold

Opacity

Cut planes ○ Sagittal X ● Coronal Y ○ Axial Z

Save data (20.3 MB)



Supplementary 3D models

Help

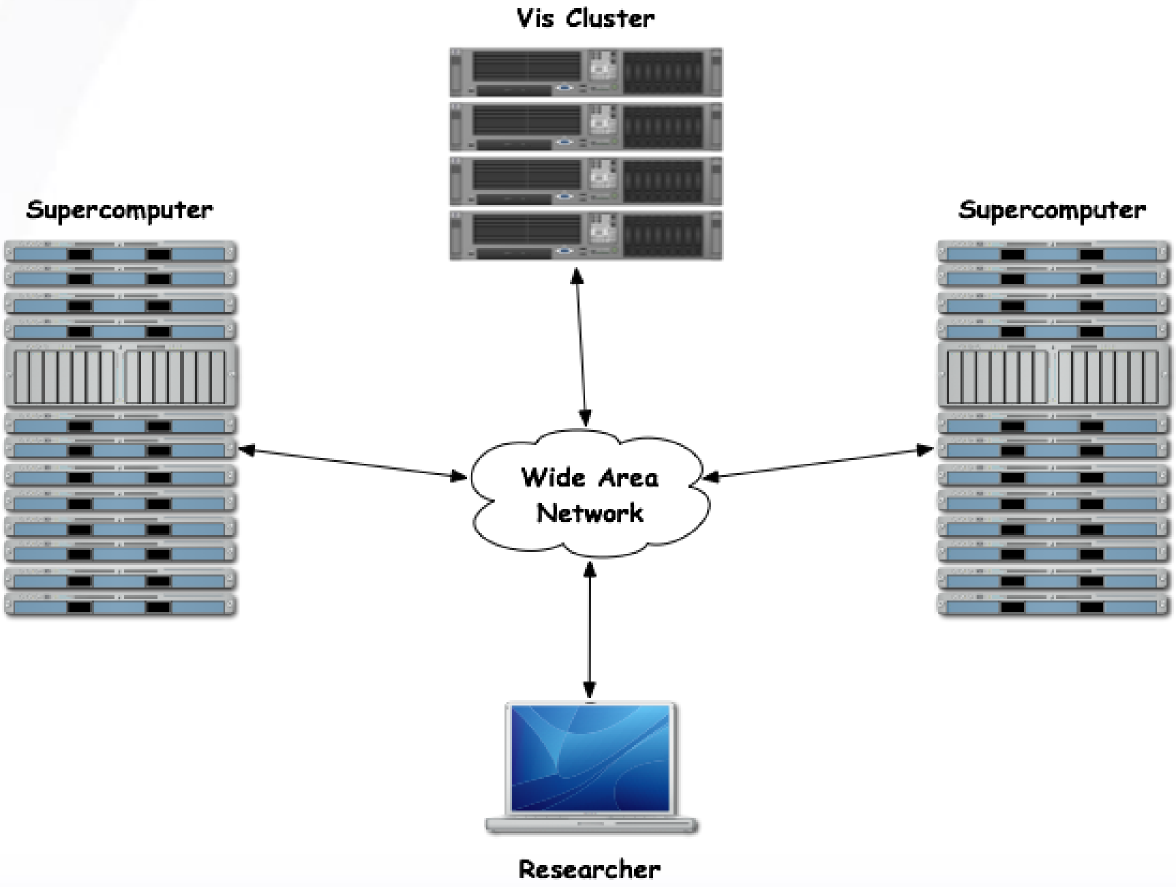
□ Rotate · Reset

Zoom

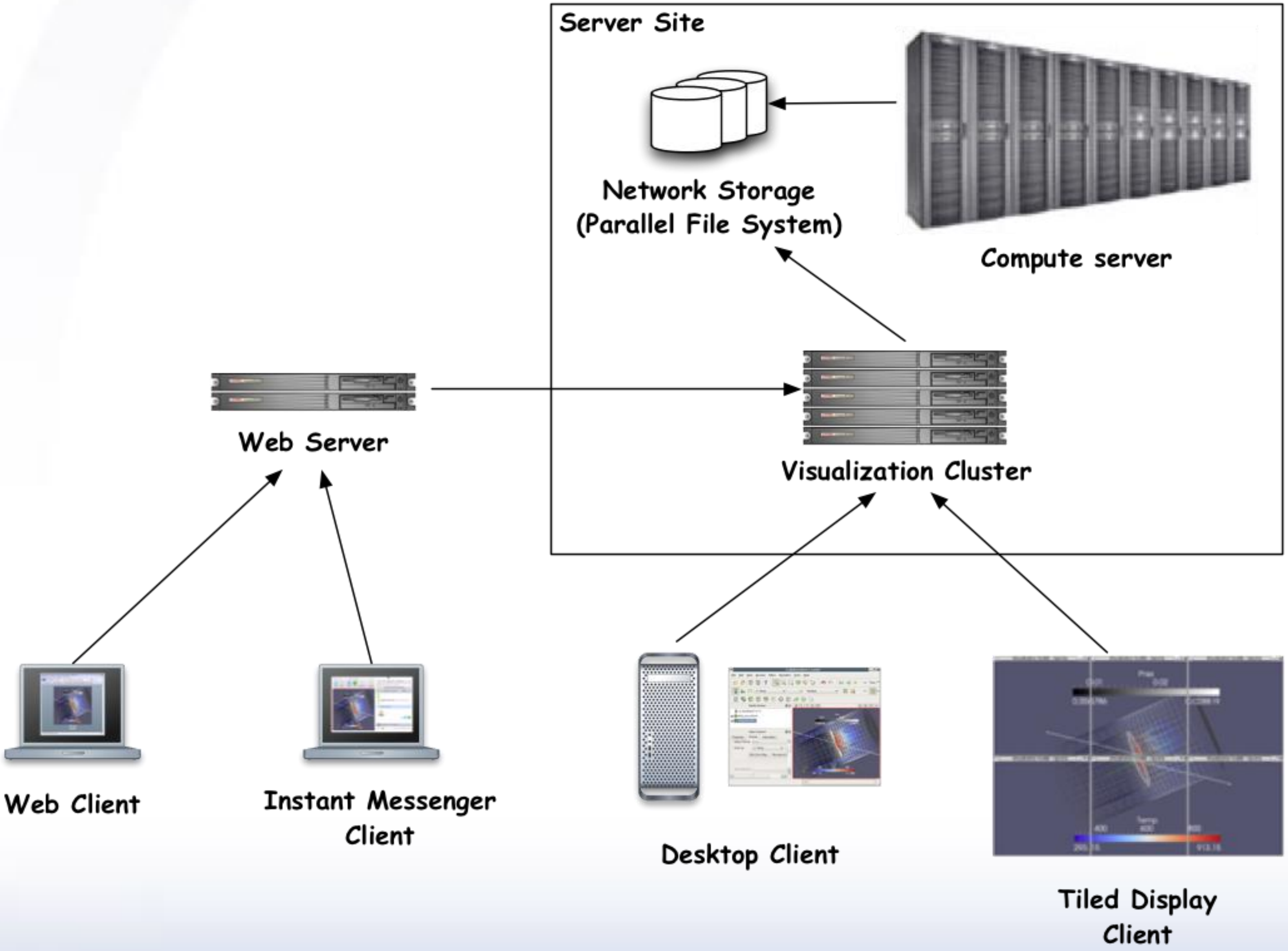
Stereo Off

Save data (69.5 MB)

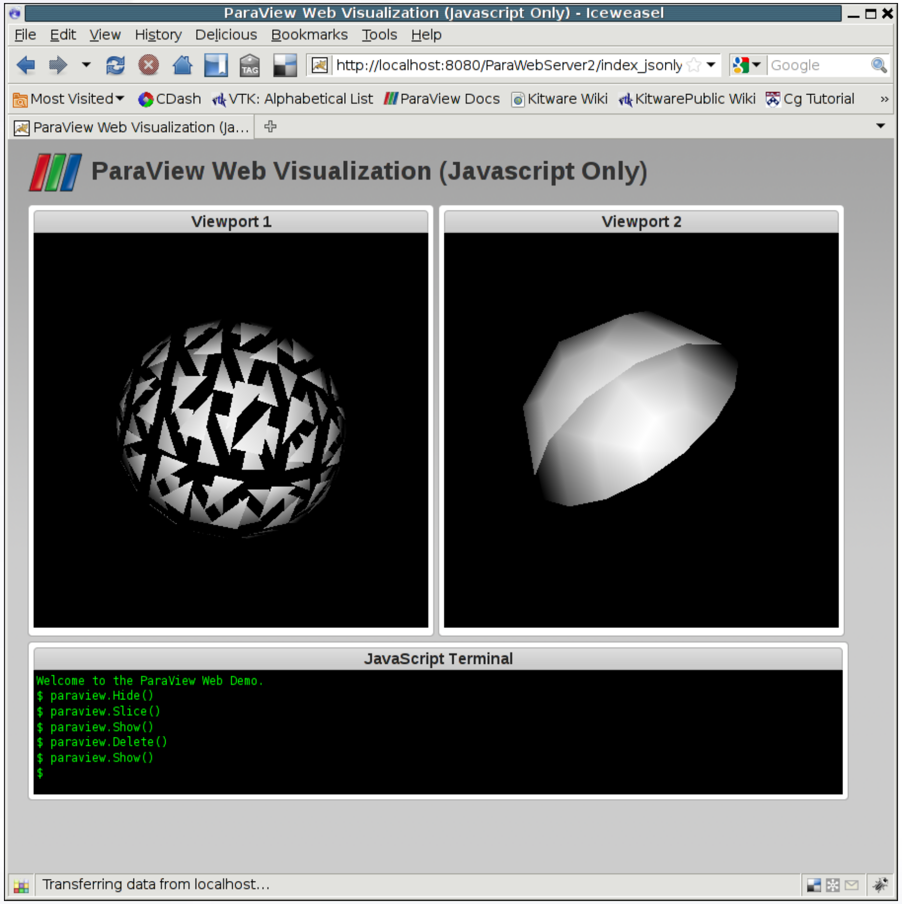
ParaViewWeb - Remote Access



ParaViewWeb - Collaboration

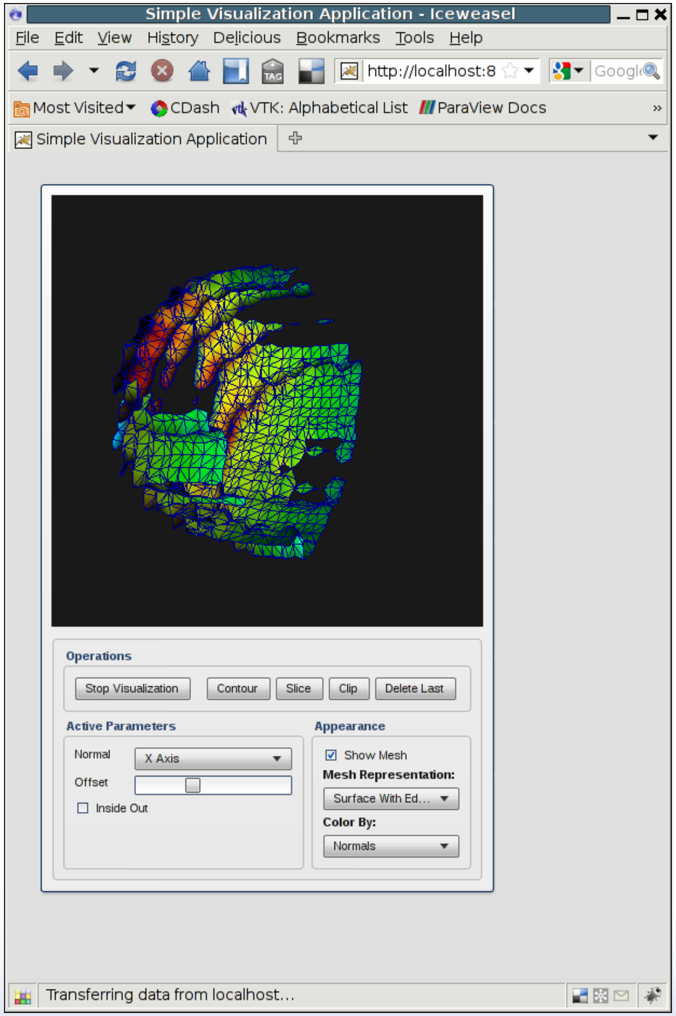


ParaViewWeb - Visualization


ParaView Web Visualization (Javascript Only) - Iceweasel
File Edit View History Delicious Bookmarks Tools Help
http://localhost:8080/ParaWebServer2/index_jsonly
Google
Most Visited CDash VTK: Alphabetical List ParaView Docs Kitware Wiki KitwarePublic Wiki Cg Tutorial
ParaView Web Visualization (Ja…
ParaView Web Visualization (Javascript Only)
Viewport 1
Viewport 2
JavaScript Terminal

```
Welcome to the ParaView Web Demo.
$ paraview.Hide()
$ paraview.Slice()
$ paraview.Show()
$ paraview.Delete()
$ paraview.Show()
$
```

Transferring data from localhost…


Simple Visualization Application - Iceweasel
File Edit View History Delicious Bookmarks Tools Help
http://localhost:8
Google
Most Visited CDash VTK: Alphabetical List ParaView Docs
Simple Visualization Application
Operations
Stop Visualization
Contour
Slice
Clip
Delete Last
Active Parameters
Appearance

Normal: X Axis

Offset

□ Inside Out

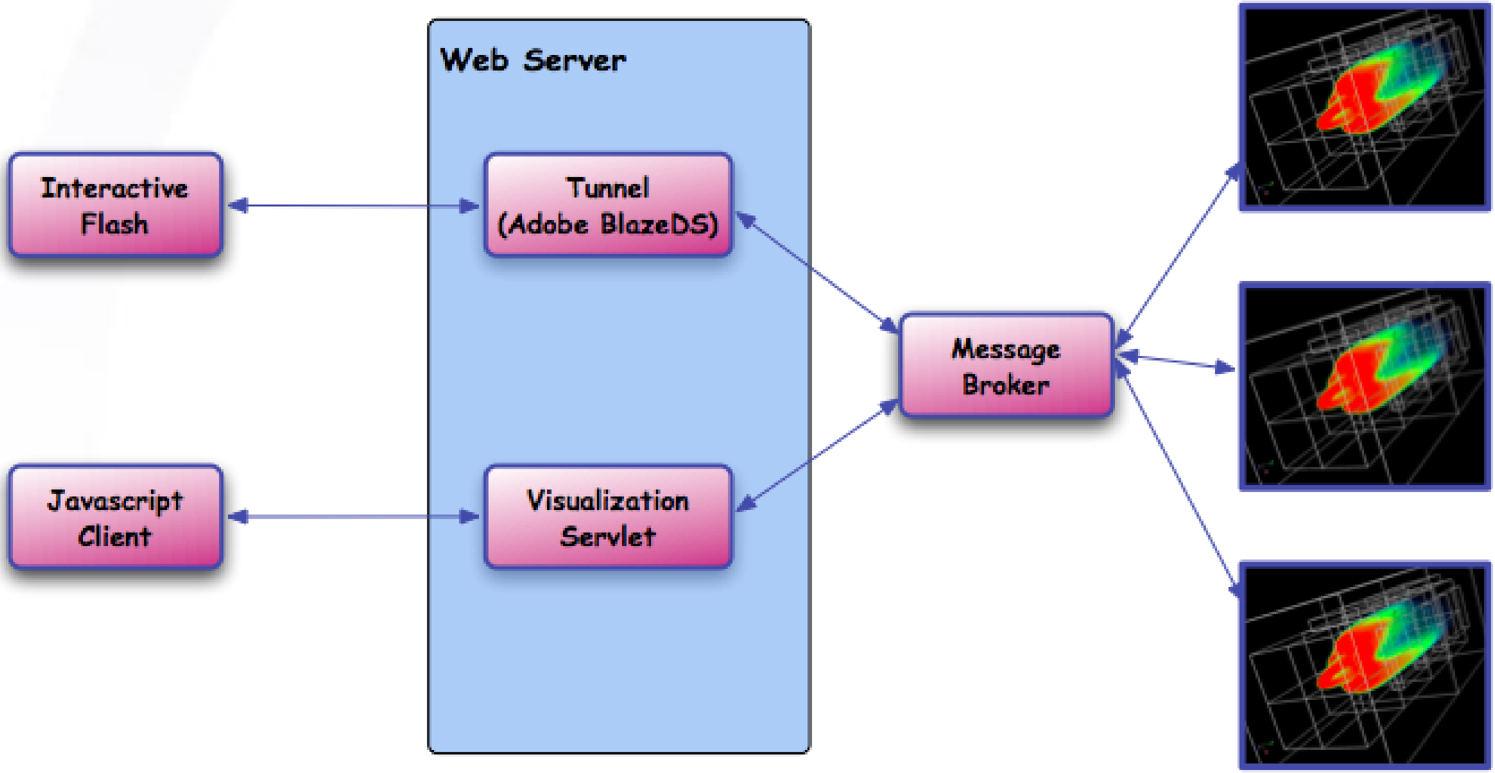
☑ Show Mesh

Mesh Representation: Surface With Ed…

Color By: Normals

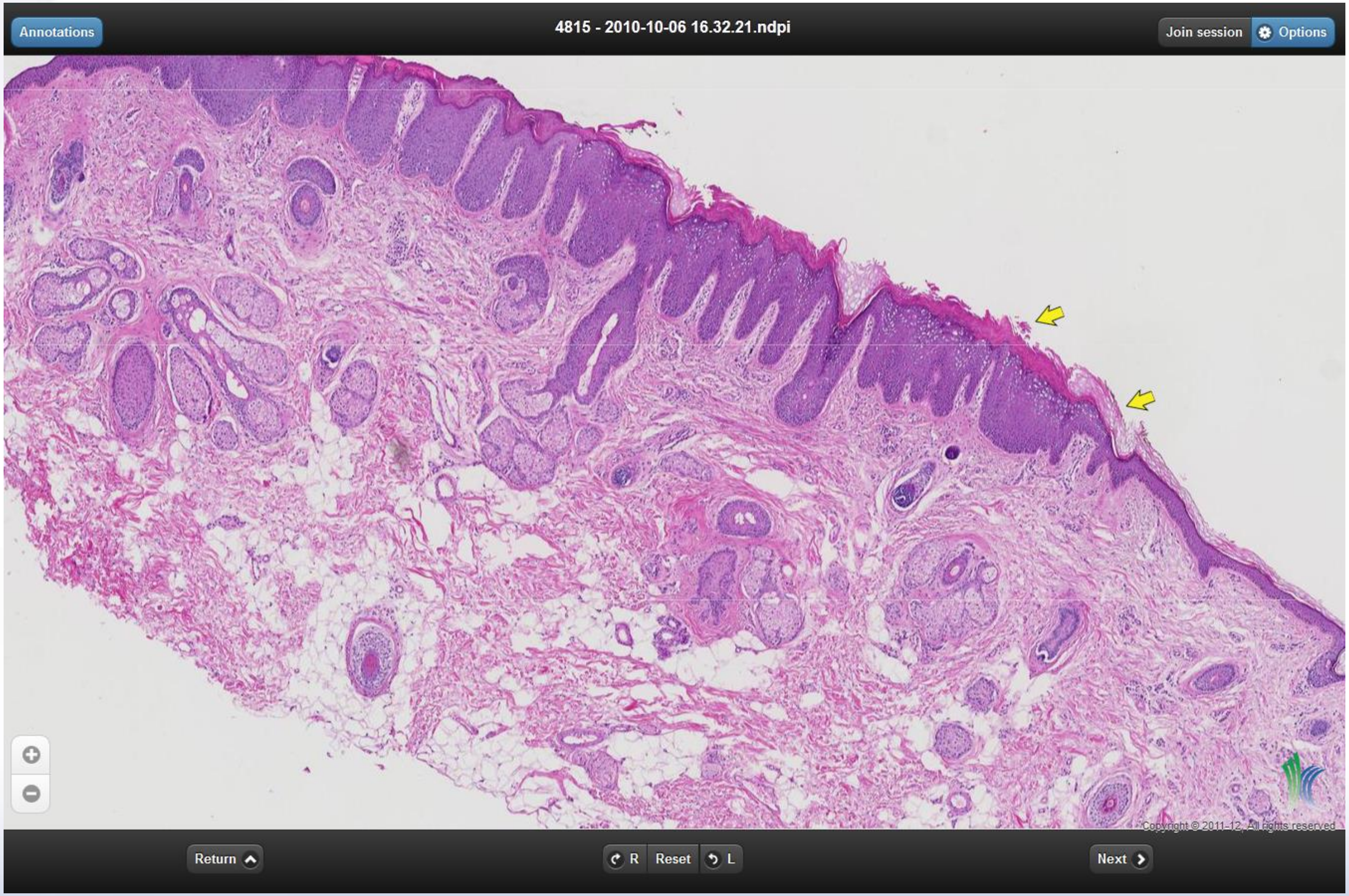
Transferring data from localhost…

ParaViewWeb - Architecture



Digital Pathology

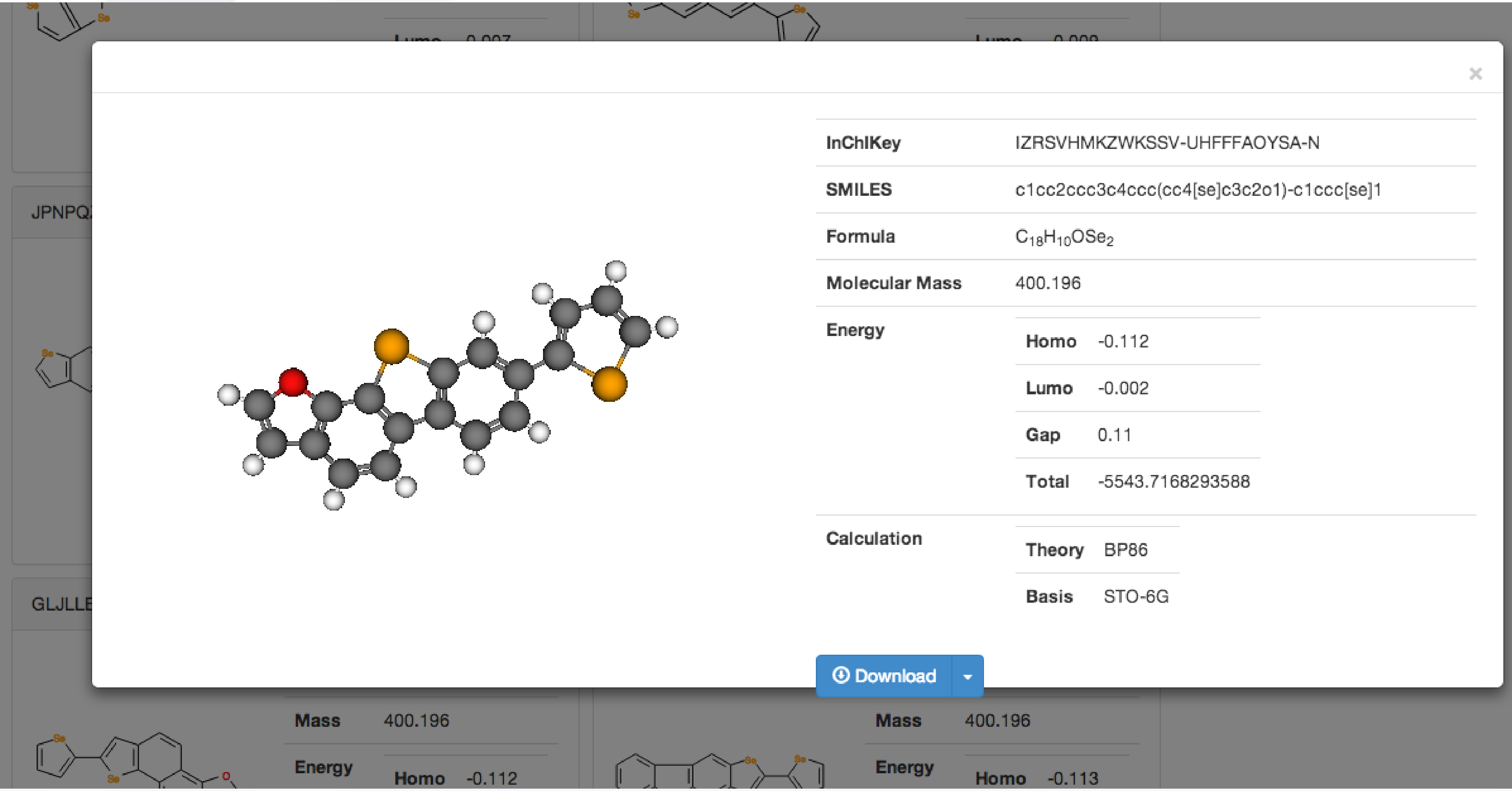
- https://slide-atlas.org/


Annotations
4815 - 2010-10-06 16.32.21.ndpi
Join session
Options
Copyright © 2011-12, All rights reserved
Return
R
Reset
L
Next

VTKWeb and MongoChem

- Uses VTK’s web architecture
- Performs interactive 3D rendering
- Runs in any modern web browser
- Same MongoDB server as MongoChem
- Moves more to the client JavaScript code
- Using a simple, Python-based server
  - Easy to add new APIs
  - Easy to deploy/integrate into other solutions

VTKWeb and Open Chemistry



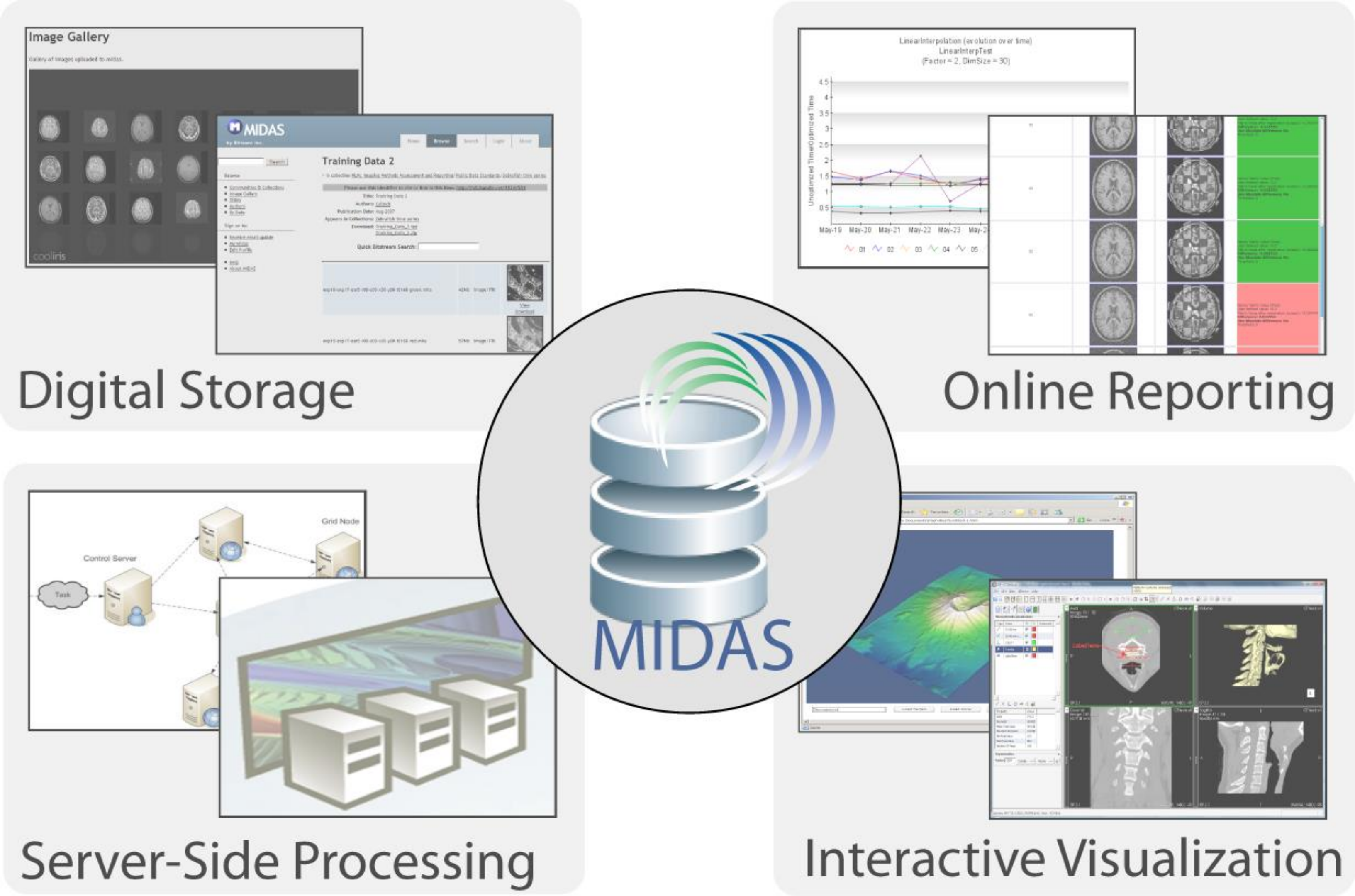
InChIKey	IZRSVHMKZWKSSV-UHFFFAOYSA-N
SMILES	c1cc2ccc3c4ccc(cc4[se]c3c2o1)-c1ccc[se]1
Formula	$C_{18}H_{10}OSe_2$
Molecular Mass	400.196
Energy	Homo -0.112 Lumo -0.002 Gap 0.11 Total -5543.7168293588
Calculation	Theory BP86 Basis STO-6G

http://data.openchemistry.org/

VES

- www.kiwiviewer.org
- Mobile visualization toolkit based on VTK
- Native applications
- Open Source
- Works on iOS and Android
- High performance

Complex Data Management



Where are we heading?

- **Computing infrastructure** is available
  - Cloud computer
  - Large clusters
- **Data management**
  - Interoperability
- **Pre-processing tools**
- Coupling **computing** with **visualization**
  - In-situ
- **Visualization nodes** on large cluster is still an issue
- **Easier/Better interfaces** for the users

Thank You!

Julien Jomier, Jeff Baumes and Joachim Pouderoux
julien.jomier@kitware.com