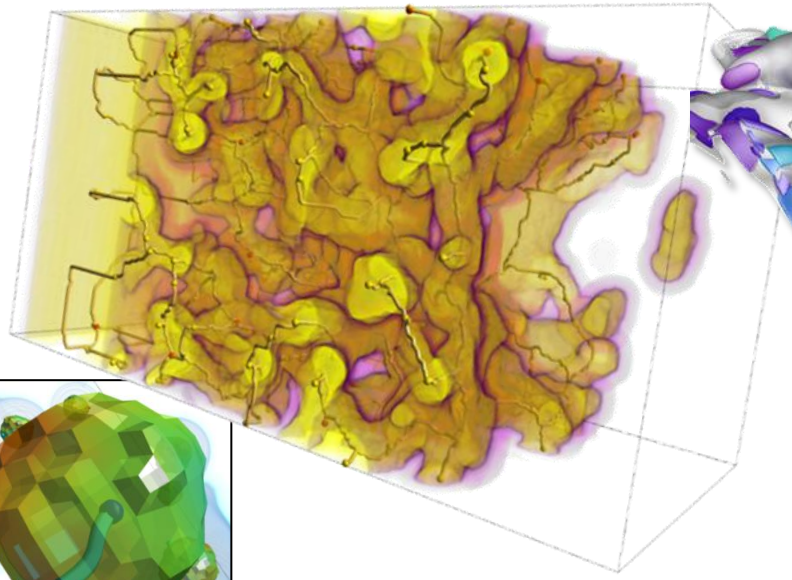
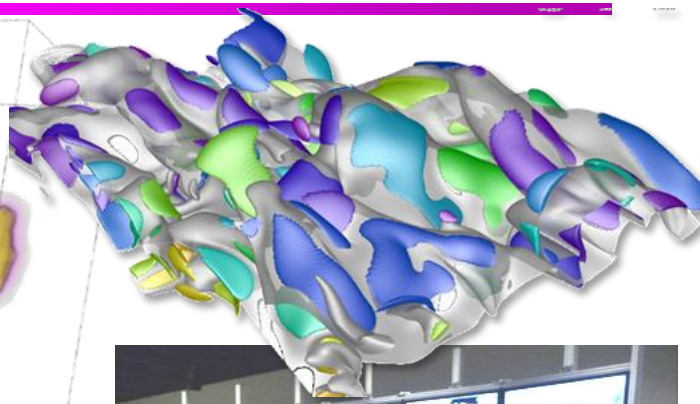
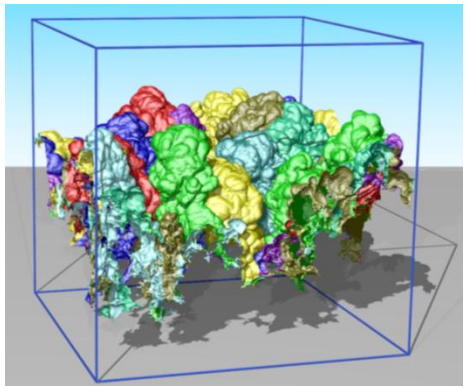
# Large Data Management, Analysis and Visualization for Science Discovery


Valerio Pascucci

Director, CEDMAV, UofU & PNNL

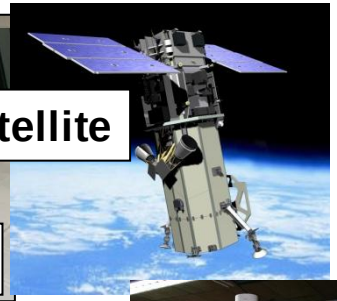
Professor, SCI Institute & School of Computing, UofU

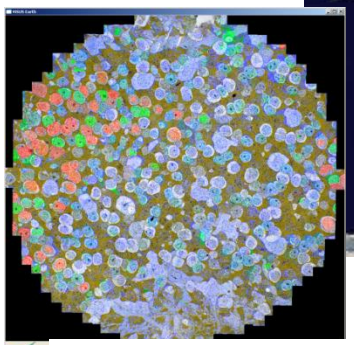
Laboratory Fellow, PNNL

CEO, ViSUS Inc.

# Massive Simulation and Sensing Devices Generate Great Challenges and Opportunities


BlueGene/L


Satellite

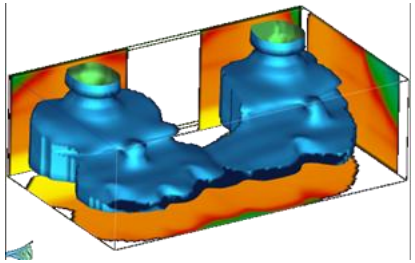

Retinal Connectome

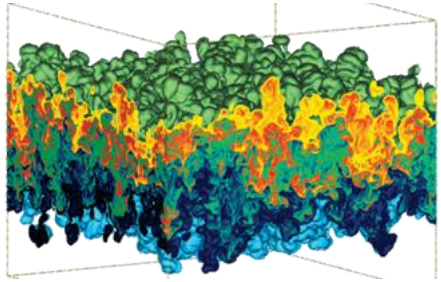

Earth Images

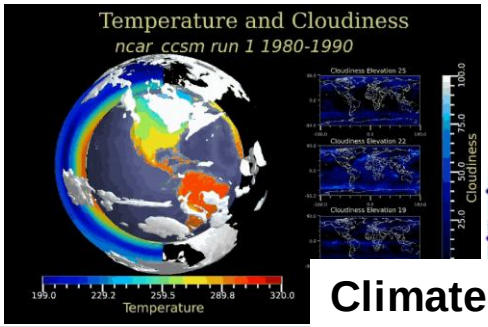

Cameras


EM

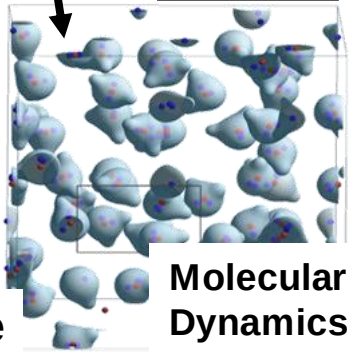
Jaguar

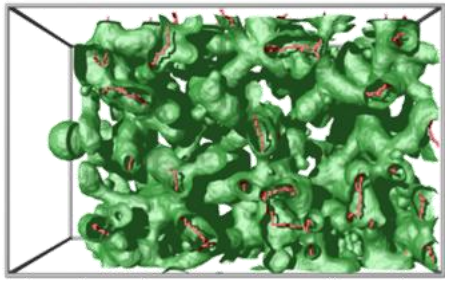

Carbon Seq. (Subsurface)

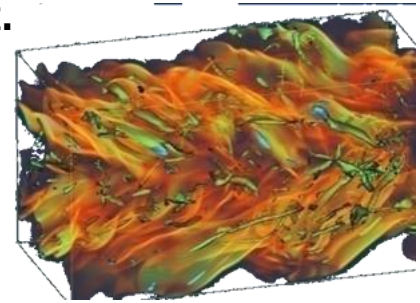

Hydrodynamic Inst.

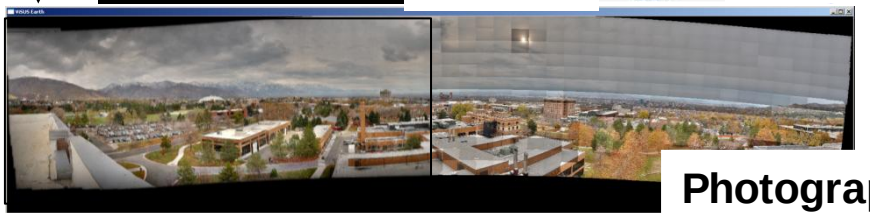


Climate


Molecular Dynamics


Porous Materials

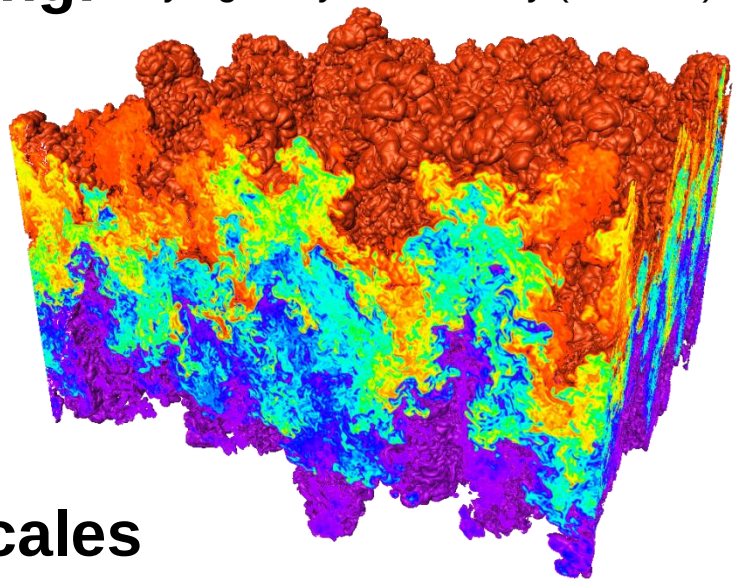

Turbulent Combustion


Photography

# Traditional Data Analysis Tools are Often Ineffective for Massive Models

- **Massive models are challenging:**
  - **Sheer volume of information**
  - **Complexity of the information represented**
  - **Complexity of presentation**

Rayleigh–Taylor instability (Miranda)

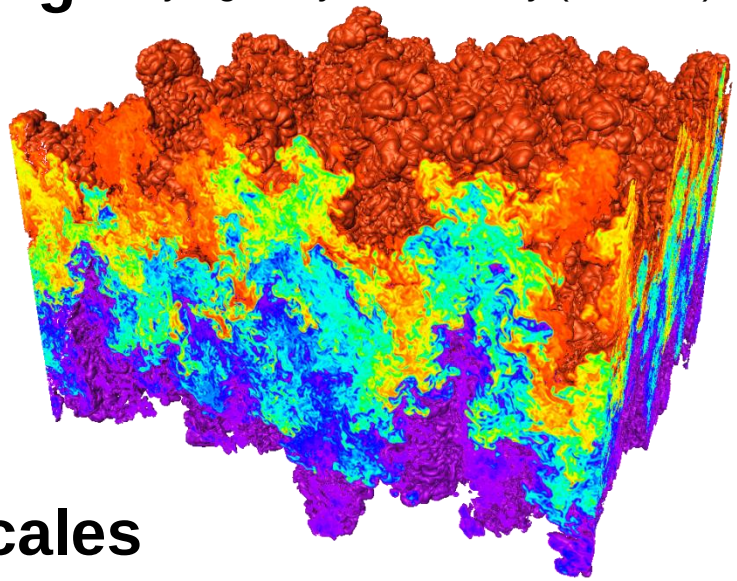


- **Tools do not scale with the data sizes**
- **Difficult to capture multiple scales**
- **Numerical methods unstable and sensitive to noise**
- **Need proper abstractions and metaphors to convey information reliably and efficiently**
- **Data Management, Analysis and Visualization are needed in a Unified Environment**

# Traditional Data Analysis Tools are Often Ineffective for Massive Models

- **Massive models are challenging:**
  - **Sheer volume of information**
  - **Complexity of the information represented**
  - **Complexity of presentation**

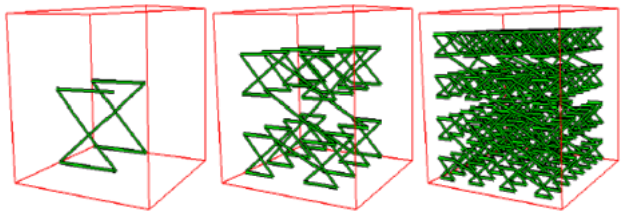
Rayleigh–Taylor instability (Miranda)



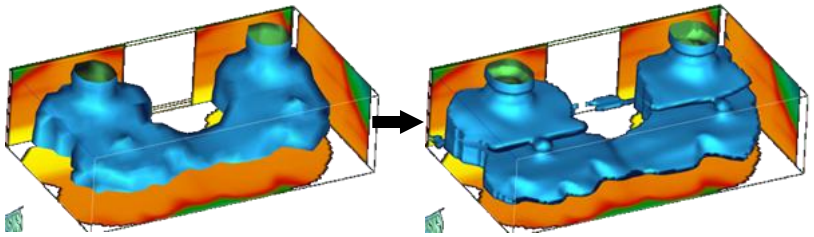
- **Tools do not scale with the data sizes**
- **Difficult to capture multiple scales**
- **Numerical methods unstable and sensitive to noise**
- **Need proper abstractions and metaphors to convey information reliably and efficiently**
- **Data Management, Analysis and Visualization are needed in a Unified Environment**

# A Cyberinfrastructure Requires Efficient Data Management and Processing

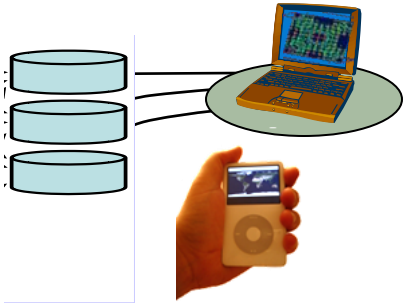
- **Advanced data storage techniques:**
  - **Data re-organization.**
  - **Compression.**



- **Advanced algorithmic techniques:**
  - **Streaming.**
  - **Progressive multi-resolution.**
  - **Out of core computations.**

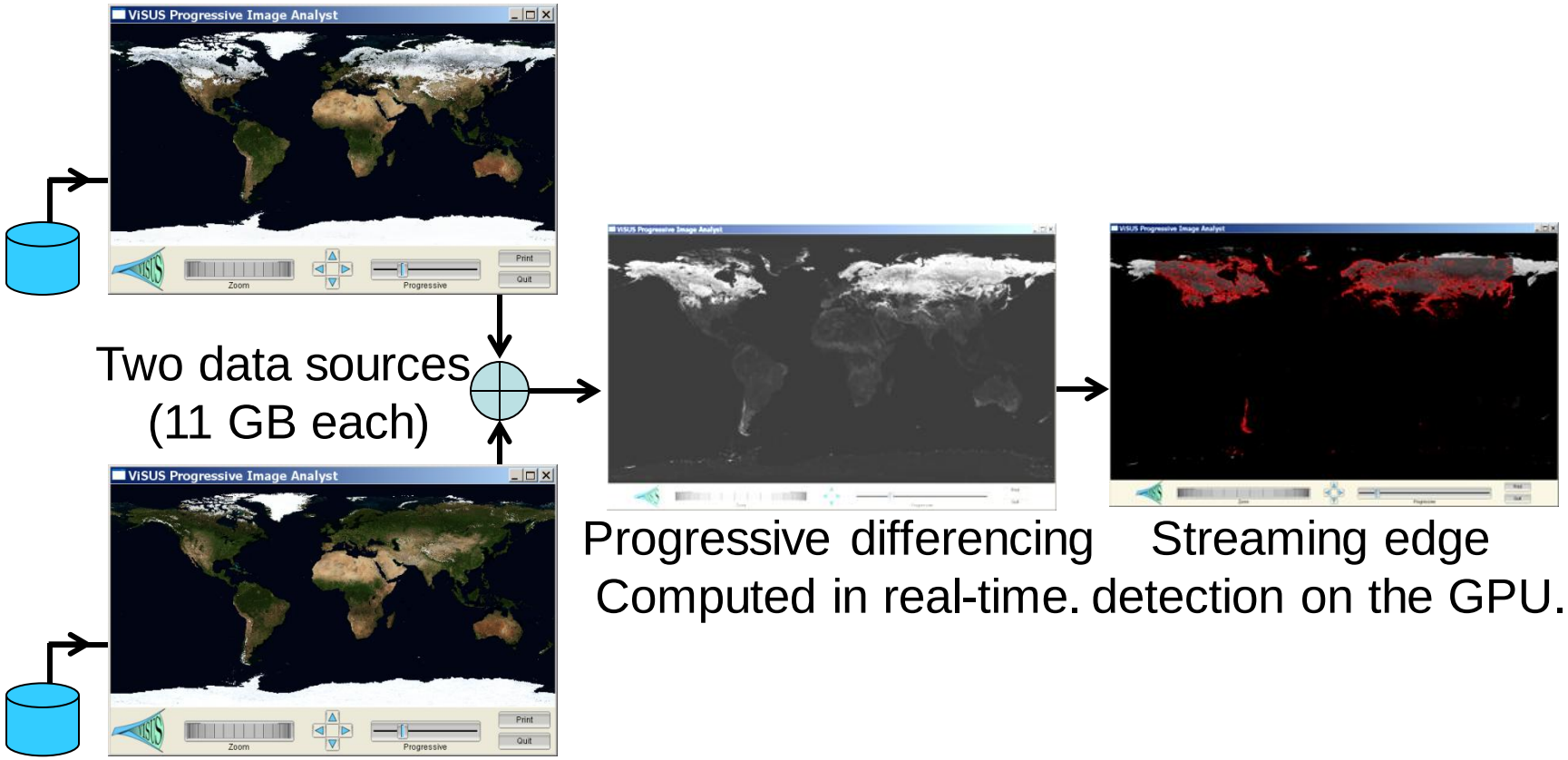


- **Scalability across a wide range of running conditions:**
  - **From laptop, to office desktop, to cluster of PC, to BG/L.**
  - **Memory, to disk, to remote data access.**



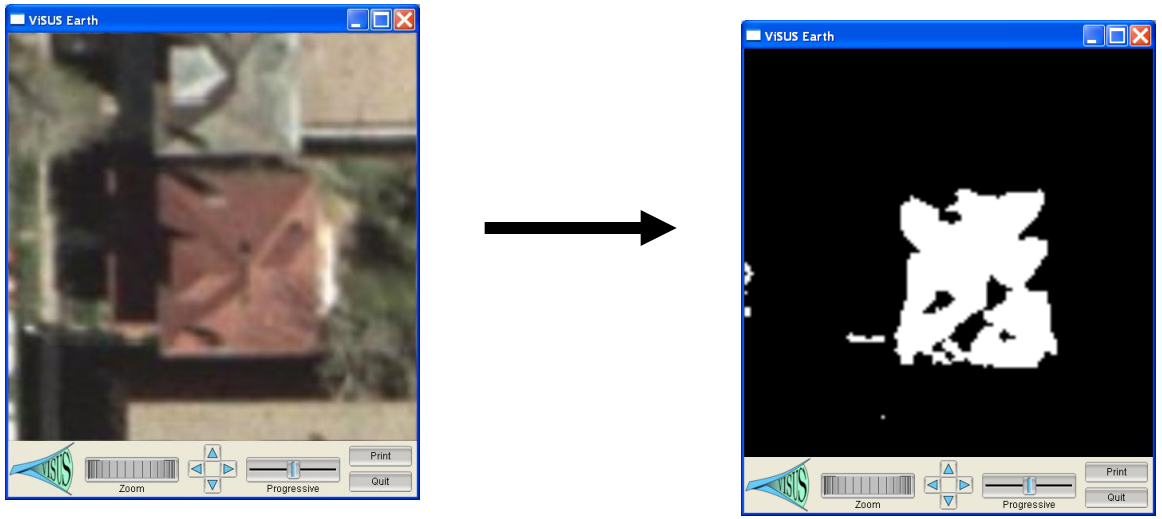
# We Allow Distributed Computations at Different Stages of the Data Stream

- Progressive Image Differencing + Editable GPU filter.



# We are Developing Progressive Scheme for Content Based Image Processing

- **Hypothesis:**



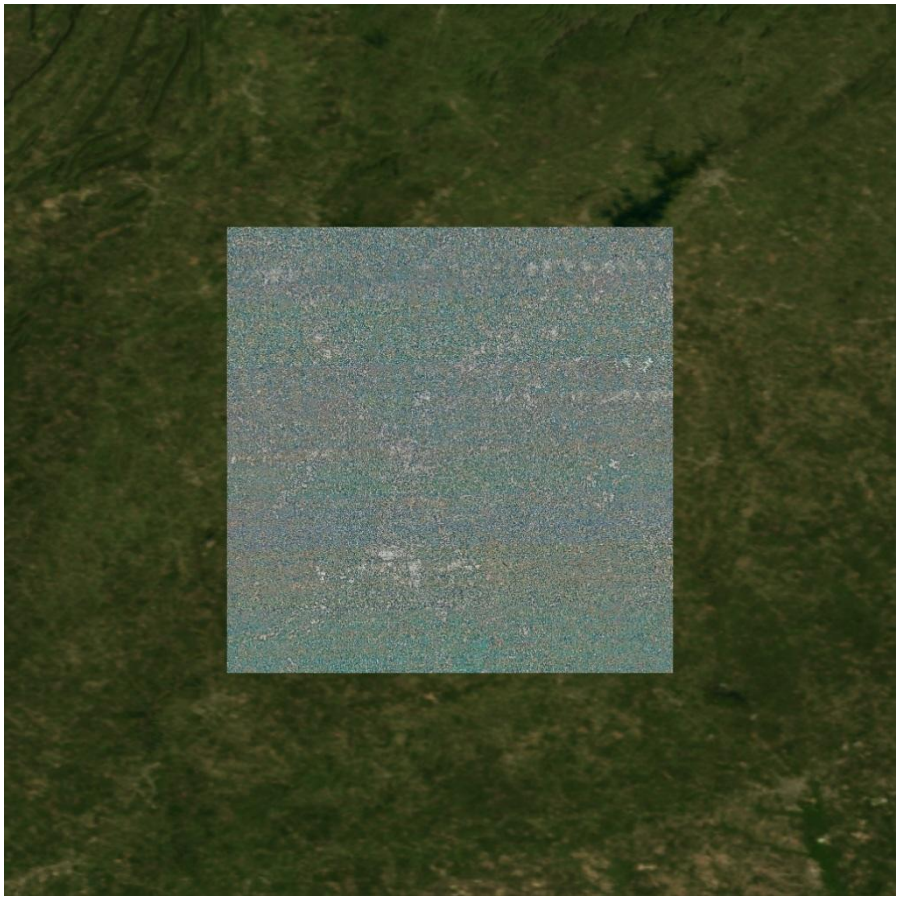

- **Progressive Analysis:**

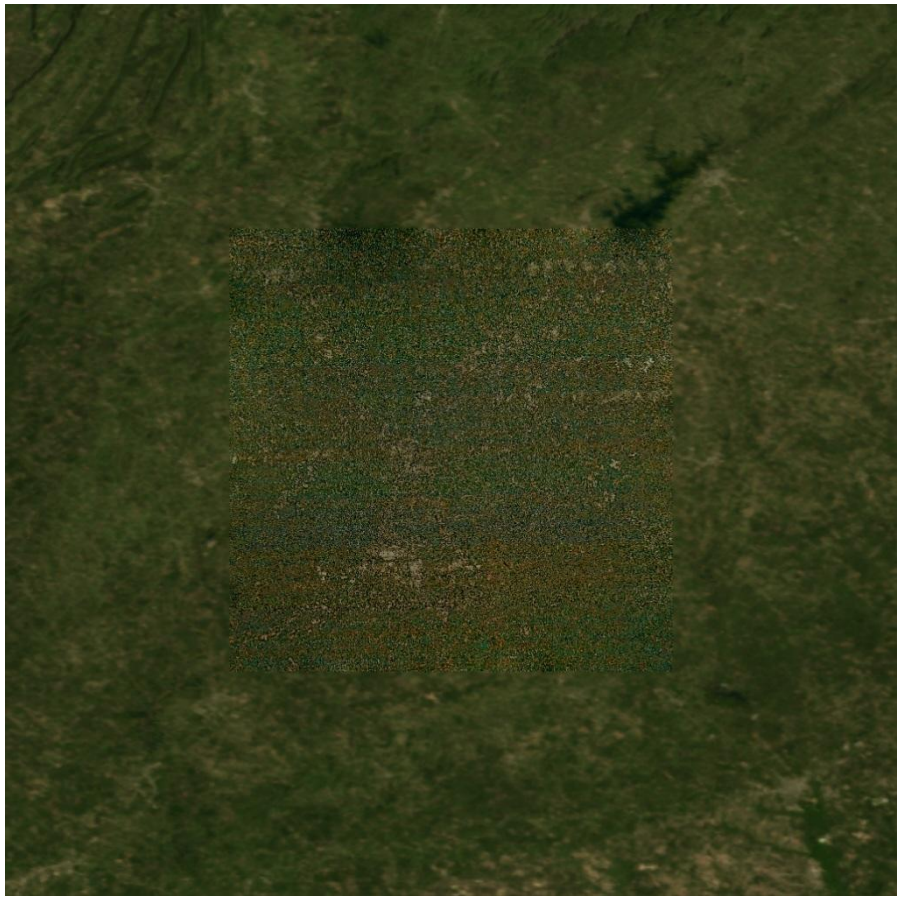

# Poisson Solver for Image Cloning in Massive Image Collections

- **Color correction of 600+ images in real time**

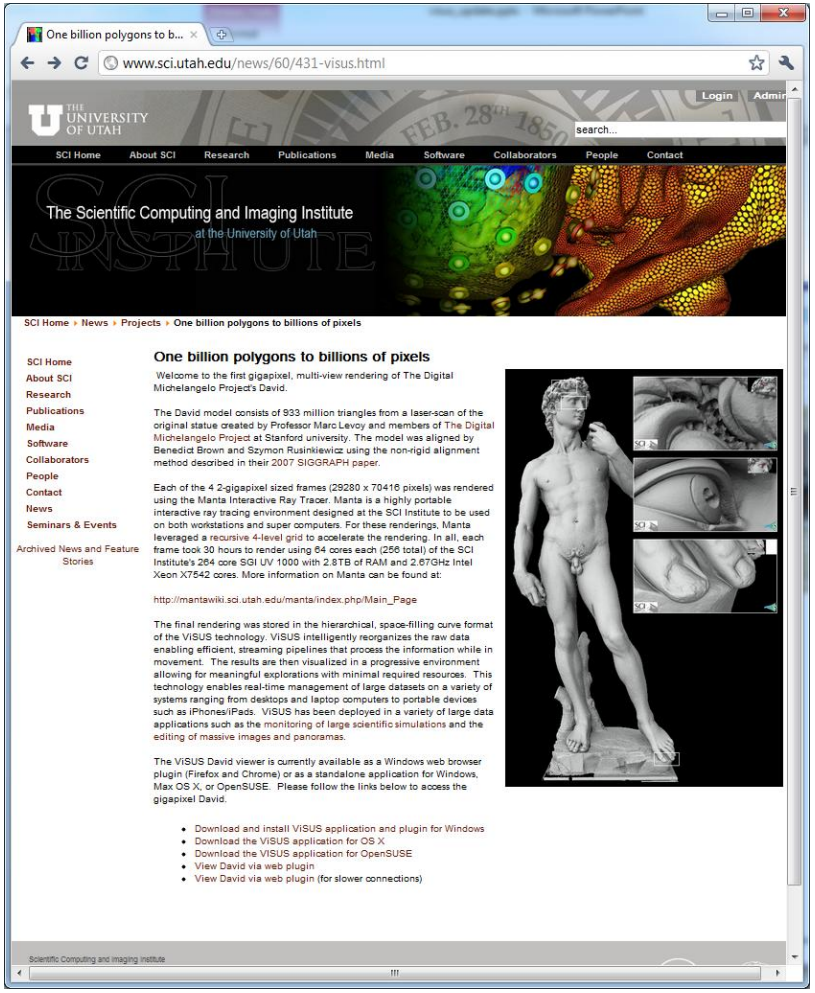
# Poisson Solver for Image Cloning in Massive Image Collections

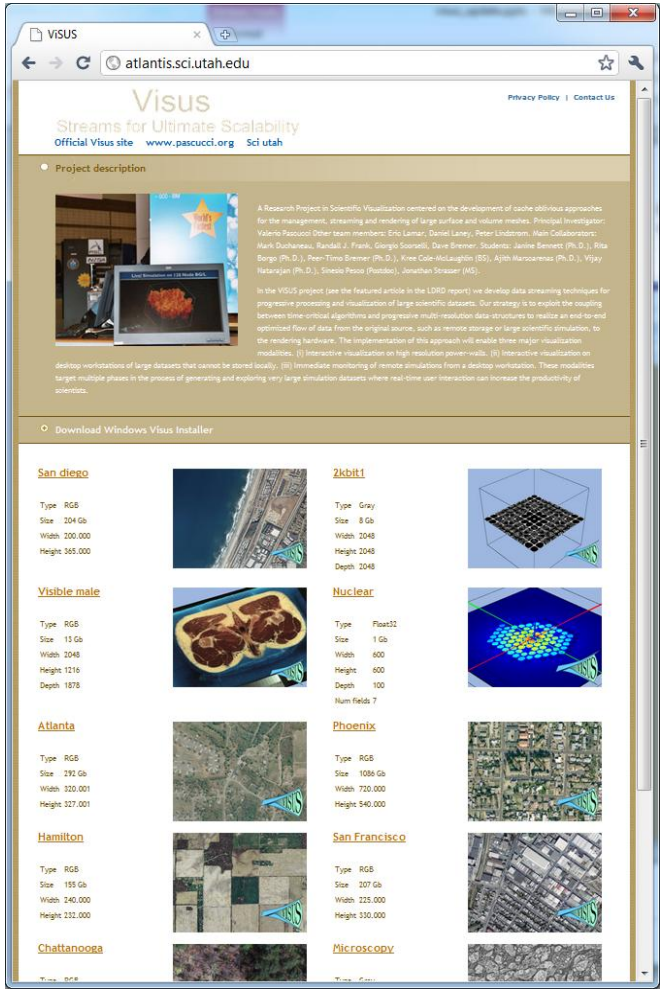
- **Pasting a 300GB satellite image of a city in background world map**





# Server can be wrapped in Apache plug-in
# Client can be run in a web browser

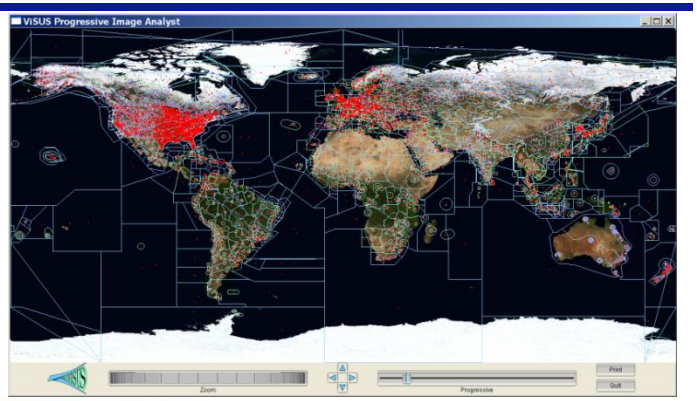


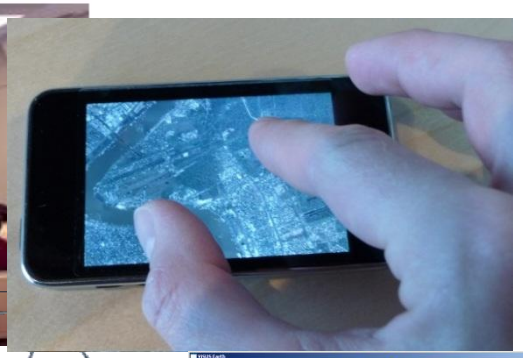
# Geospatial Data Rendering on iPad

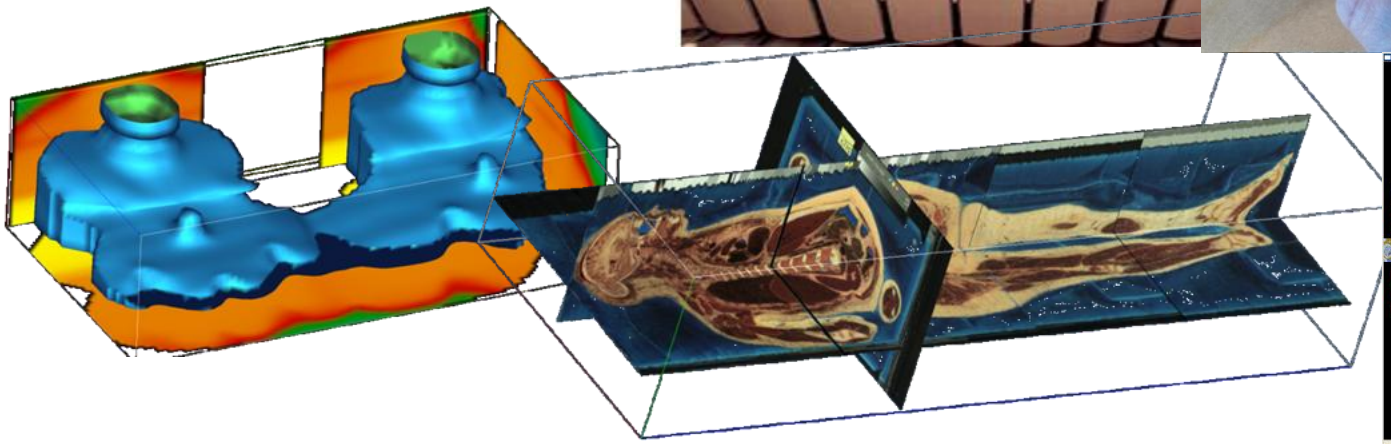
Both client and SERVER run of handheld devices, e.g. multiple iPhones can be clients and servers for each other to share information on the field

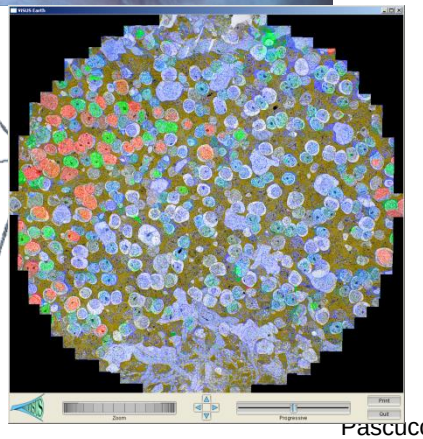
# We Demonstrated Performance and Scalability in a Variety of Applications



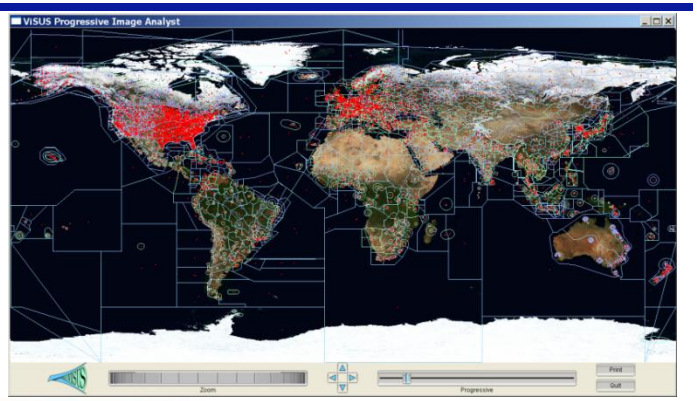


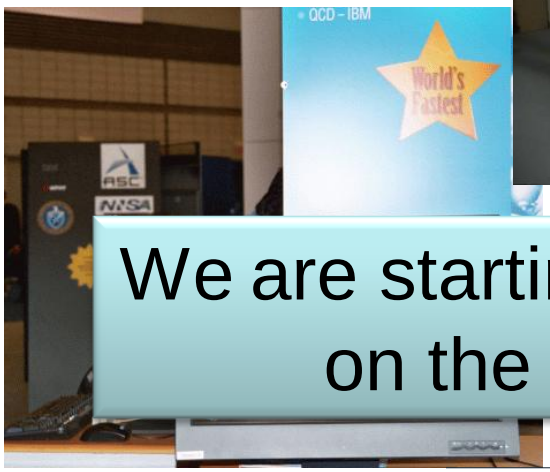




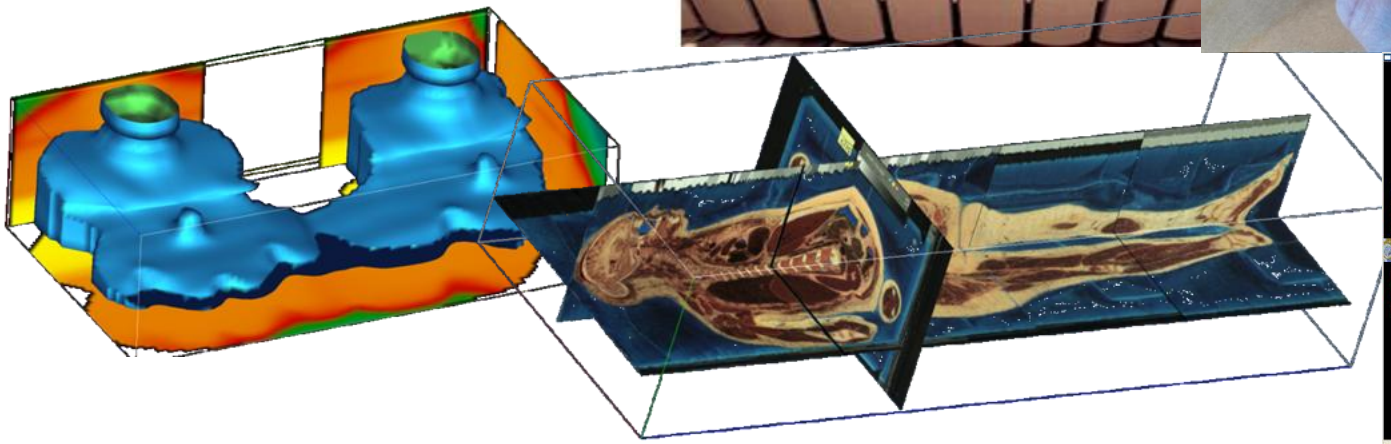


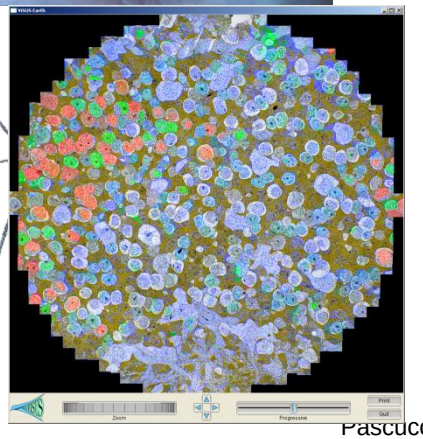
# We Demonstrated Performance and Scalability in a Variety of Applications



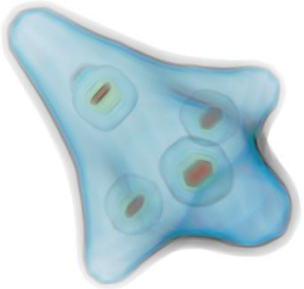

We are starting to redesign simulations based on the new algorithmic techniques

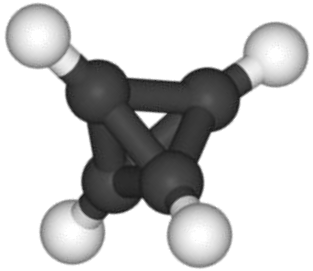


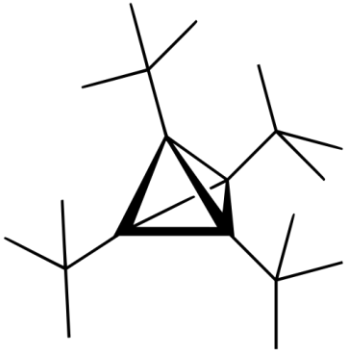


# Abstractions Have Been Used in Science Discovery/Communication for a Long Time

# Abstraction



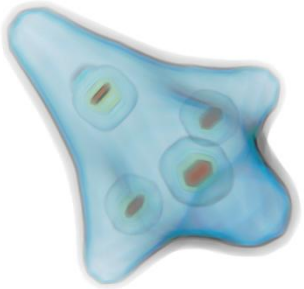


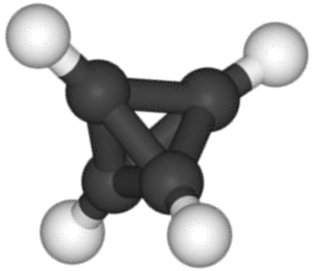


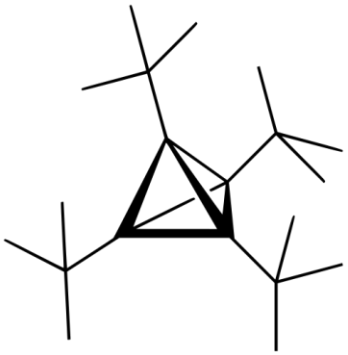
$C_4H_4$ Tetrahedrane

# Abstractions Have Been Used in Science Discovery/Communication for a Long Time

# Abstraction



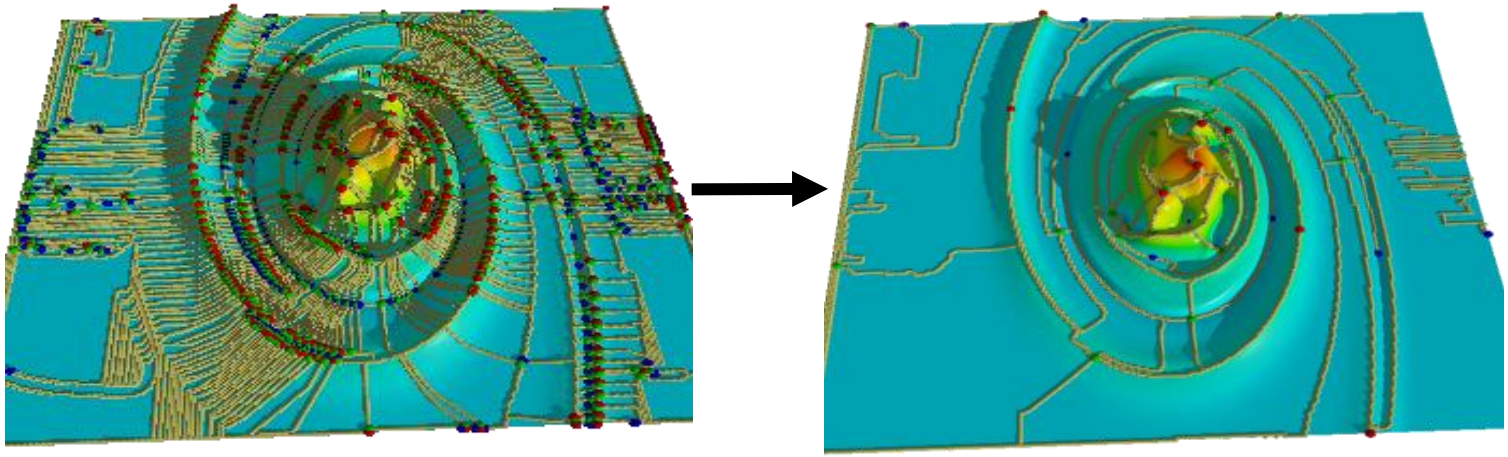




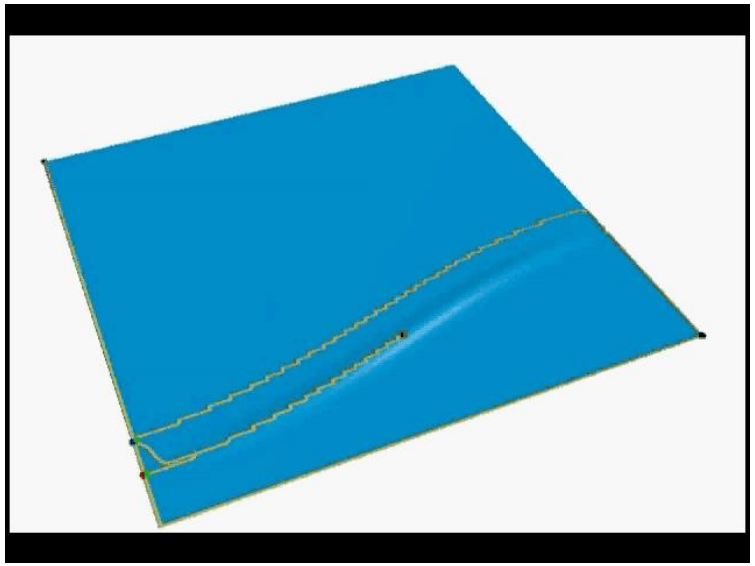
$C_4H_4$ Tetrahedrane

Language

# Topology is and Effective Language to Describe Abstractions of Features from Raw Data

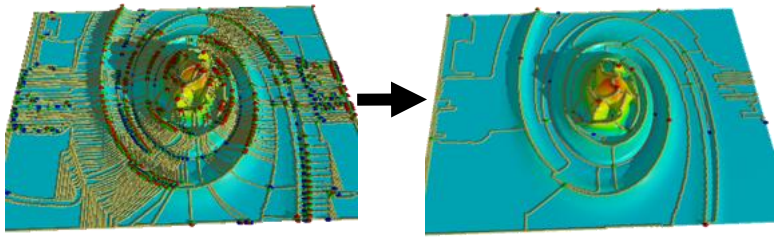


**Hierarchical topology of a 2D Miranda vorticity field**

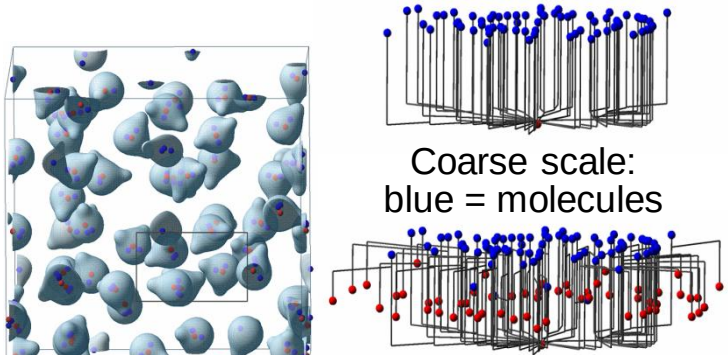


# We Adopt Robust Topological Methods to Abstract Features from Raw Data

- **Provably robust computation**
- **Provably complete feature extraction and quantification**
- **Hierarchical structures used to capture multiple scales**
- **Error-bounded approximations associated with each scale**
- **Formal definition associated with each analysis**
- **Streaming techniques to achieve scalable performance**



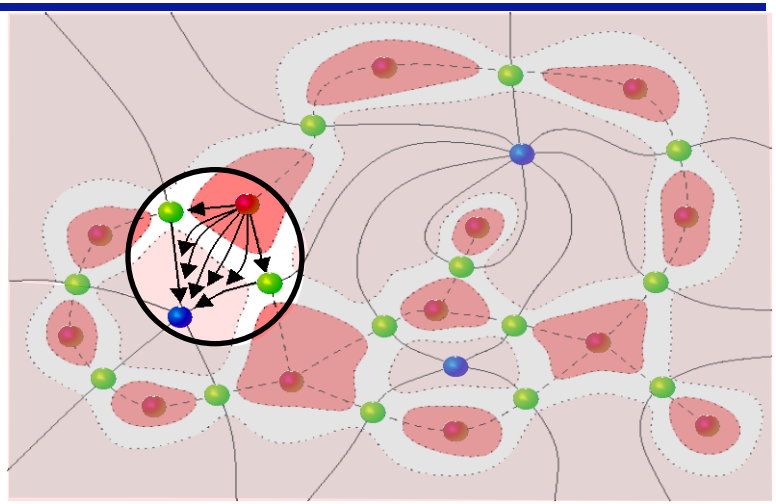
Hierarchical topology of a 2D Miranda vorticity field

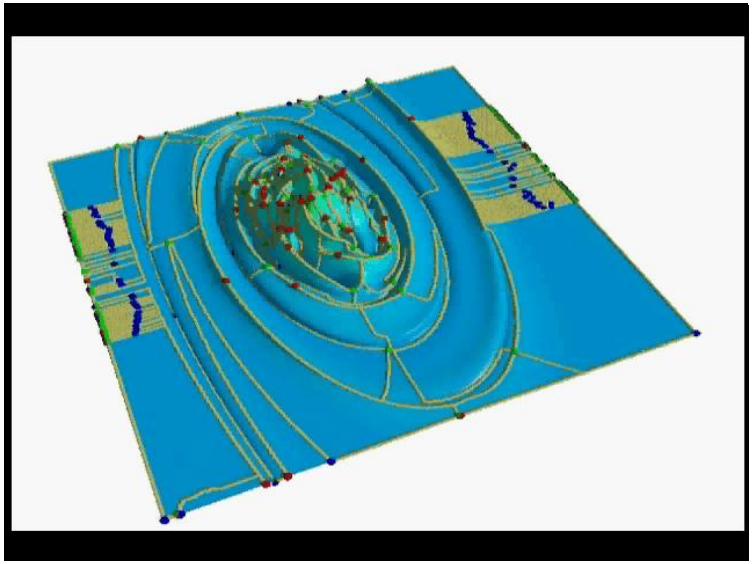


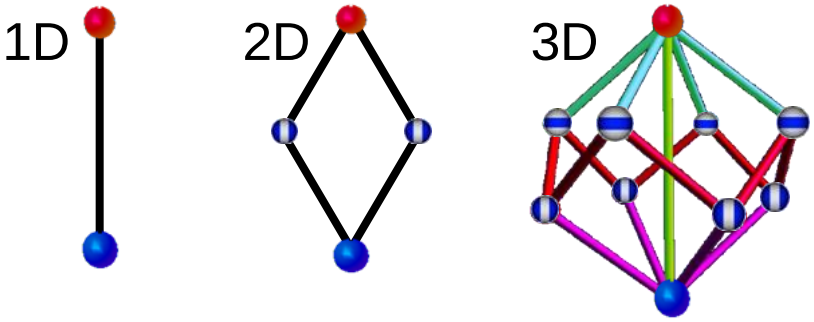

Molecular dynamics simulation (left) with abstract graph representation of its features at two scales (right)

# The Morse–Smale Complex Provides a Complete Data Segmentation for Analysis

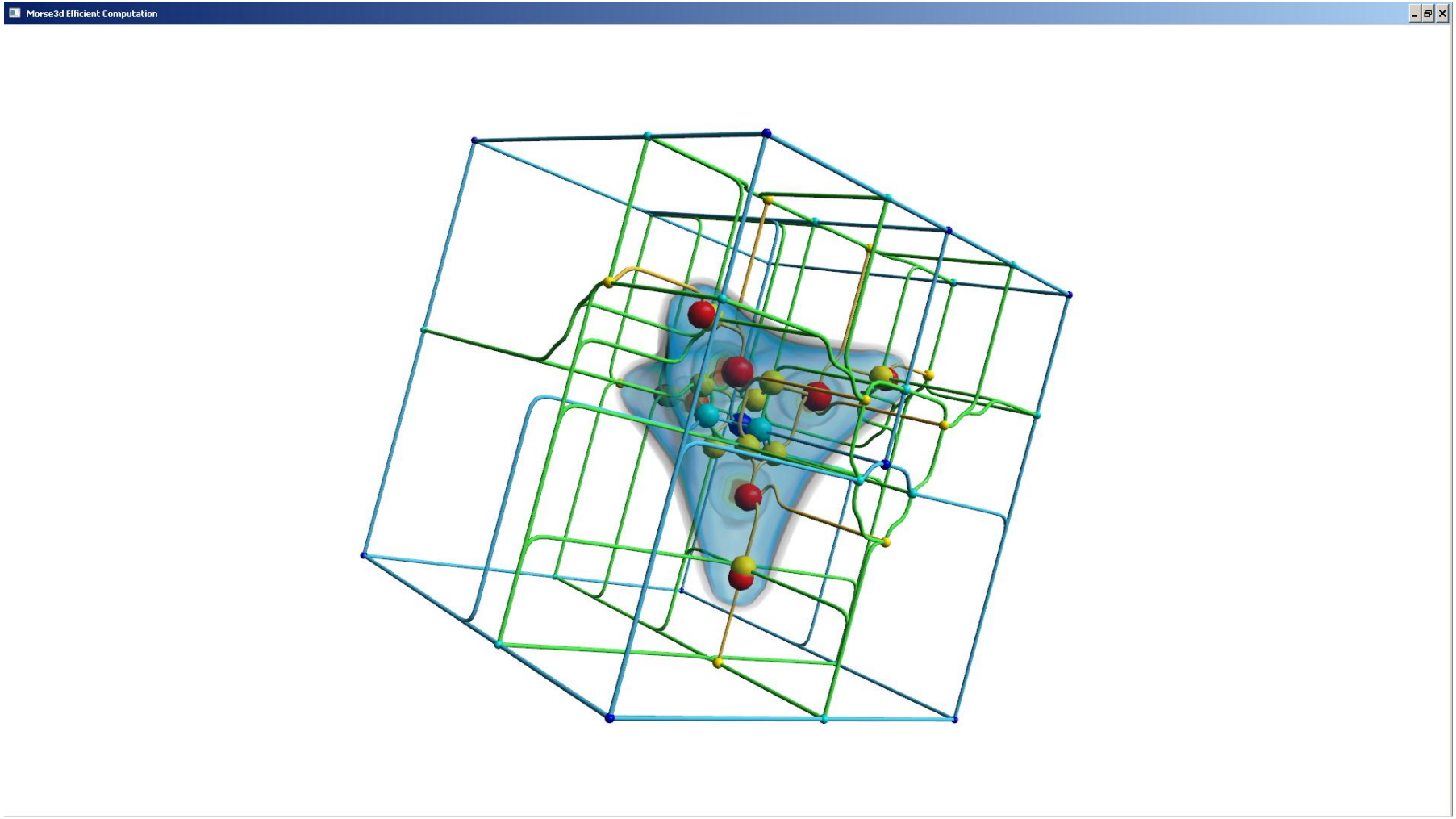
- **The Morse–Smale complex partitions the domain of $f$ in regions of uniform gradient**
- **Generalizes the notion of monotonic interval**
- **Dimension of a region equal index difference of source and destination**
- **Remove inconsistency of local gradient evaluations**





# 3D Morse Smale Complex for $C_4H_4$

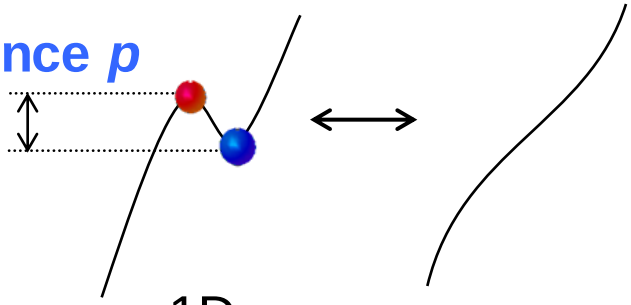


# We Build Hierarchical Representations to Capture the Data at Multiple Scales

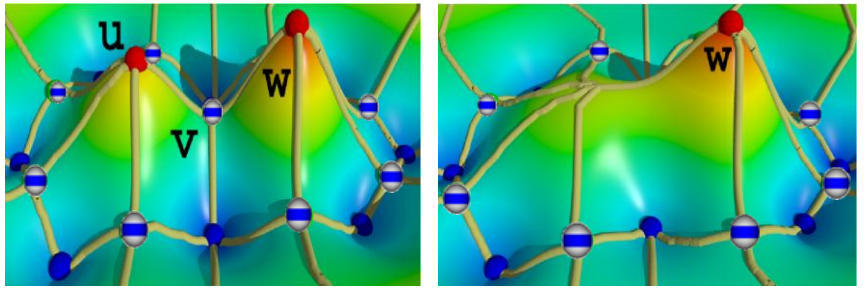
**Simplification:** critical points can be created or destroyed in pairs connected by arcs in the MS-complex

**Approximation**: error ≤ persistence

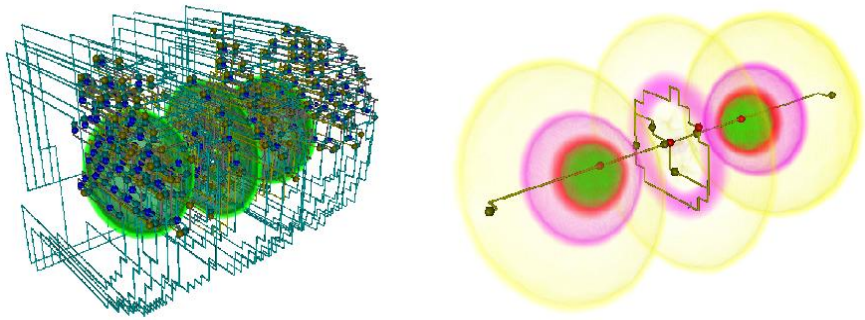
**Hierarchical Model**: consistent gradient segmentation at all scales



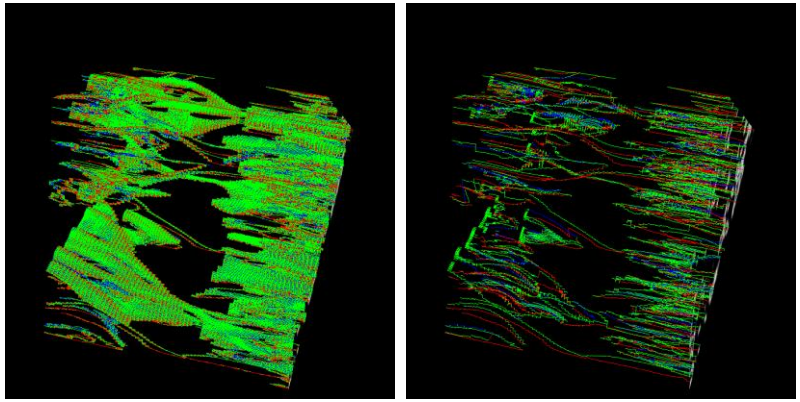

1D




2D

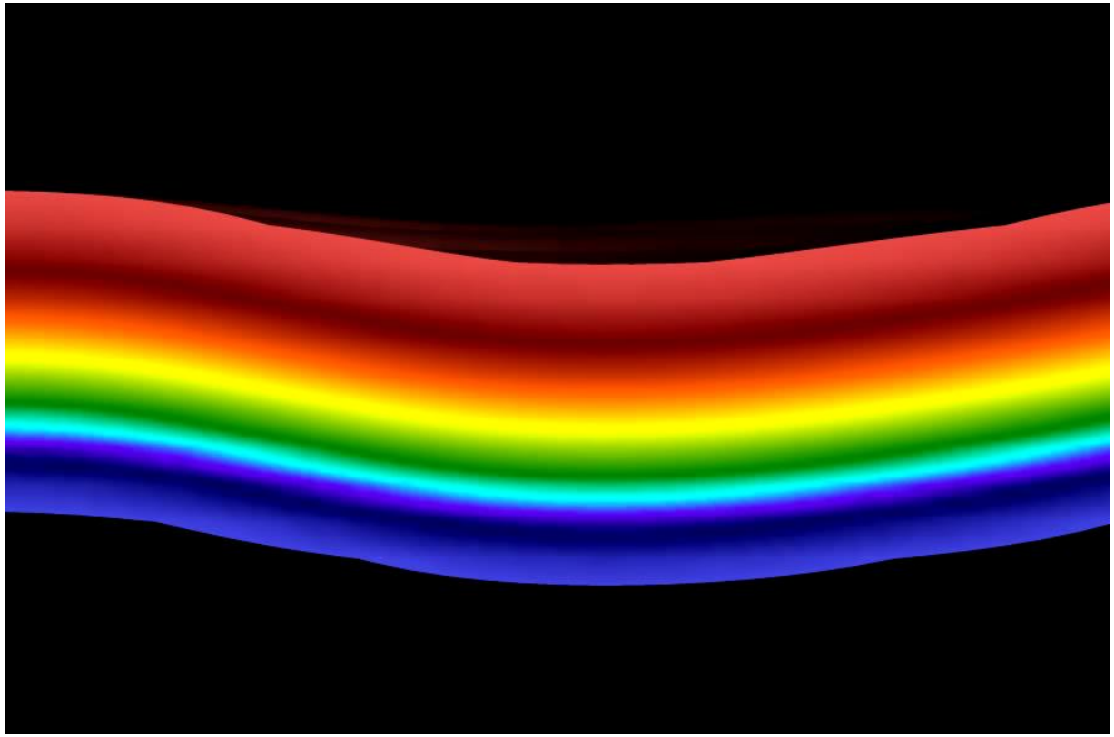


3D



Time

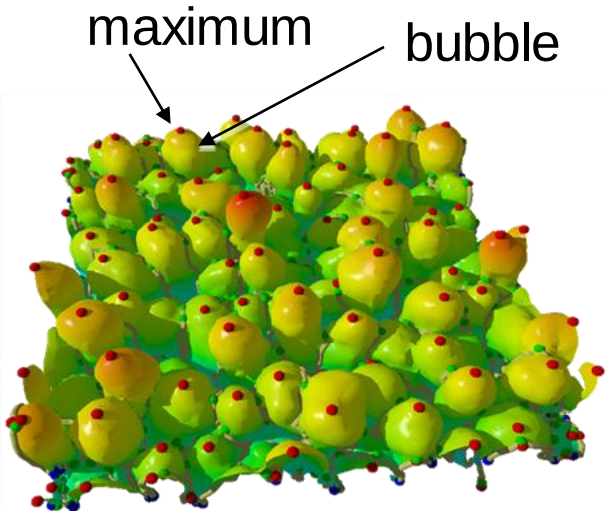
# Understanding the Dynamics of Rayleigh-Taylor instabilities



**Rayleigh-Taylor instabilities arise in fusion, super-novae, and other fundamental phenomena:**

- start: heavy fluid above, light fluid below
- gravity drives the mixing process
- the mixing region lies between the upper envelope surface (red) and the lower envelope surface (blue)
- 25 to 40 TB of data from simulations

# We Analyze High-Resolution Rayleigh–Taylor Instability Simulations

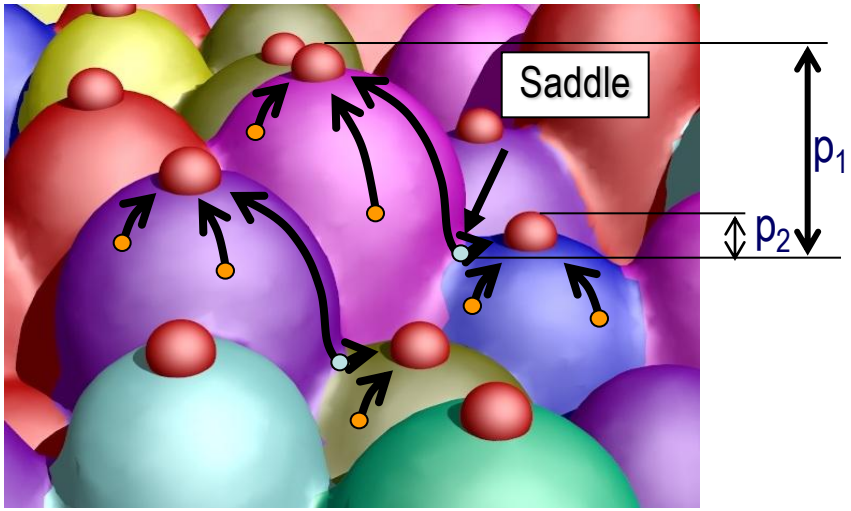


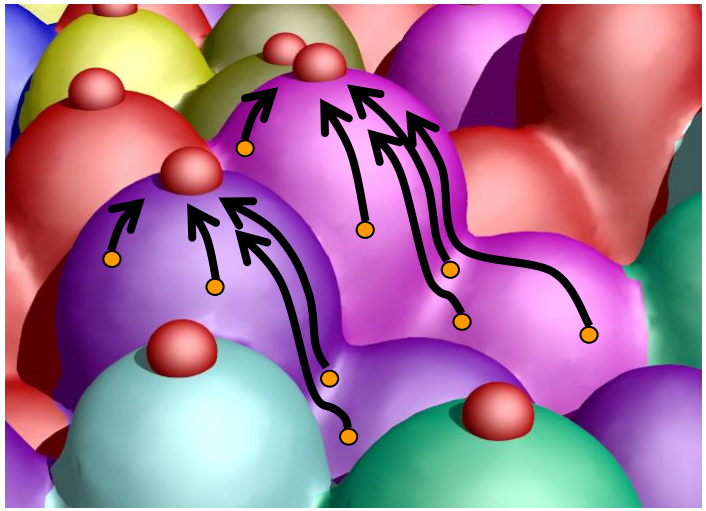

- **Large eddy simulation run on Linux cluster: 1152 x 1152 x 1152**
  - **~ 40 G / dump**
  - **759 dumps, about 25 TB**
- **Direct numerical simulation run on BlueGene/L: 3072 x 3072 x Z**
  - **Z depends on width of mixing layer**
  - **More than 40 TB**
- **Bubble-like structures are observed in laboratory and simulations**
- **Bubble dynamics are considered an important way to characterize the mixing process**
  - **Mixing rate =** $\partial(\#bubbles)/\partial t.$
- **There is no prevalent formal definition of bubbles**

# A Hierarchal Model is Generated by Simplification of Critical Points

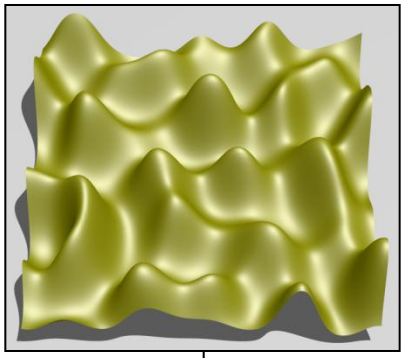
- **Persistence is varied to annihilate pairs of critical points and produce coarser segmentations**

- **Critical points with higher persistence are preserved at the coarser scales**



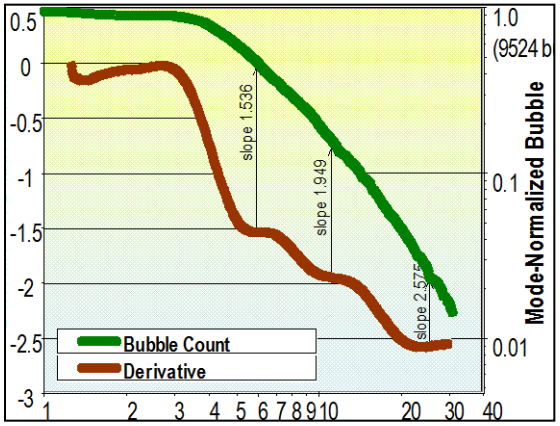



# Our workflow utilizes streaming data management and analysis tools



isosurface extraction

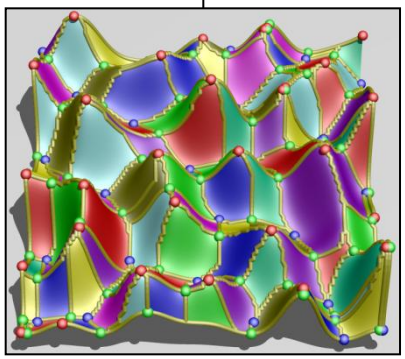
multi scale Morse analysis

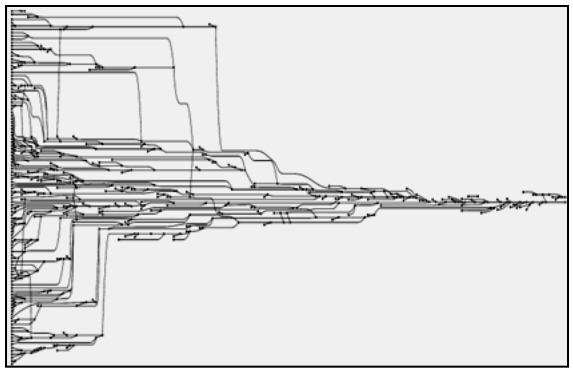



quantitative analysis

RAW DATA

4.5TB

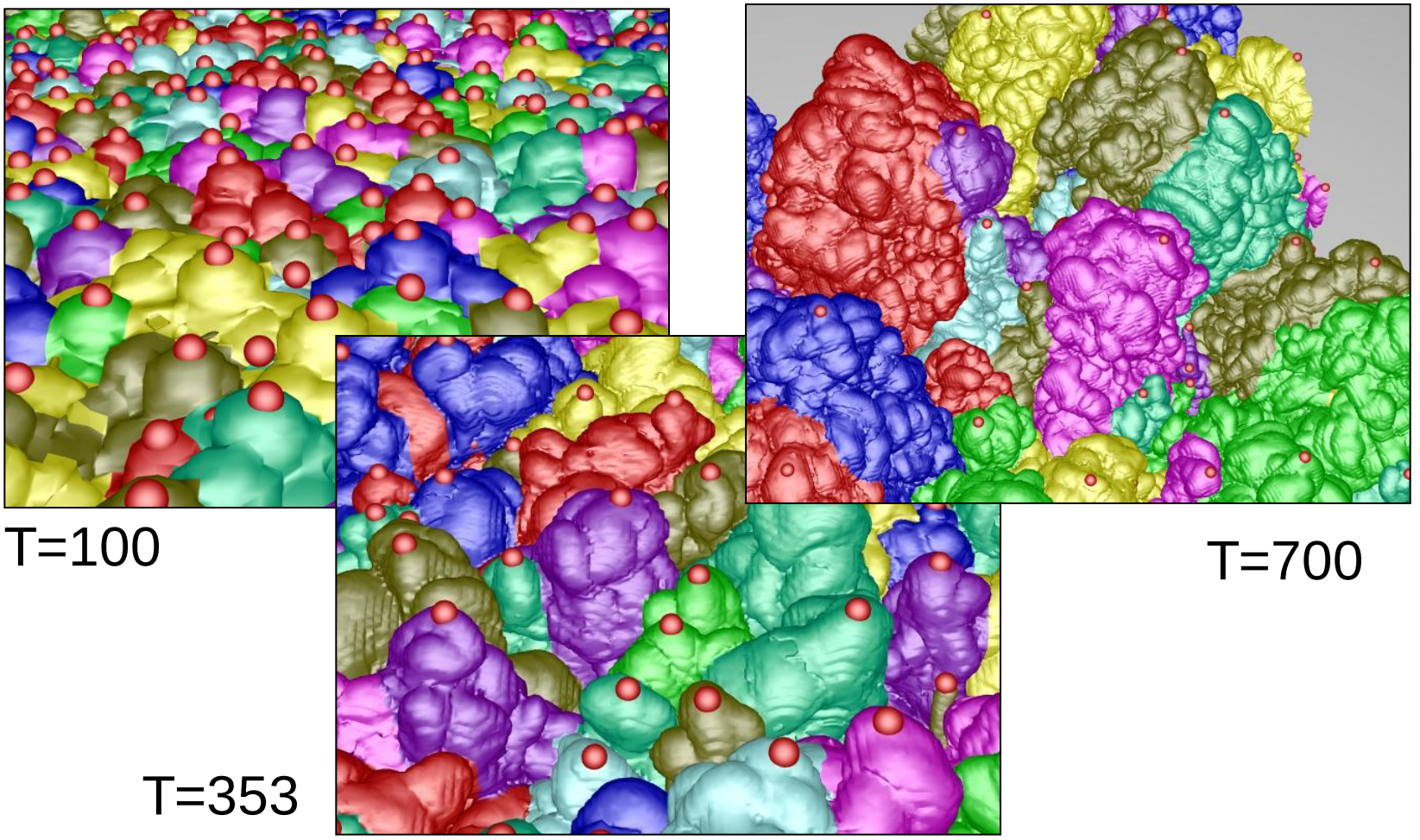




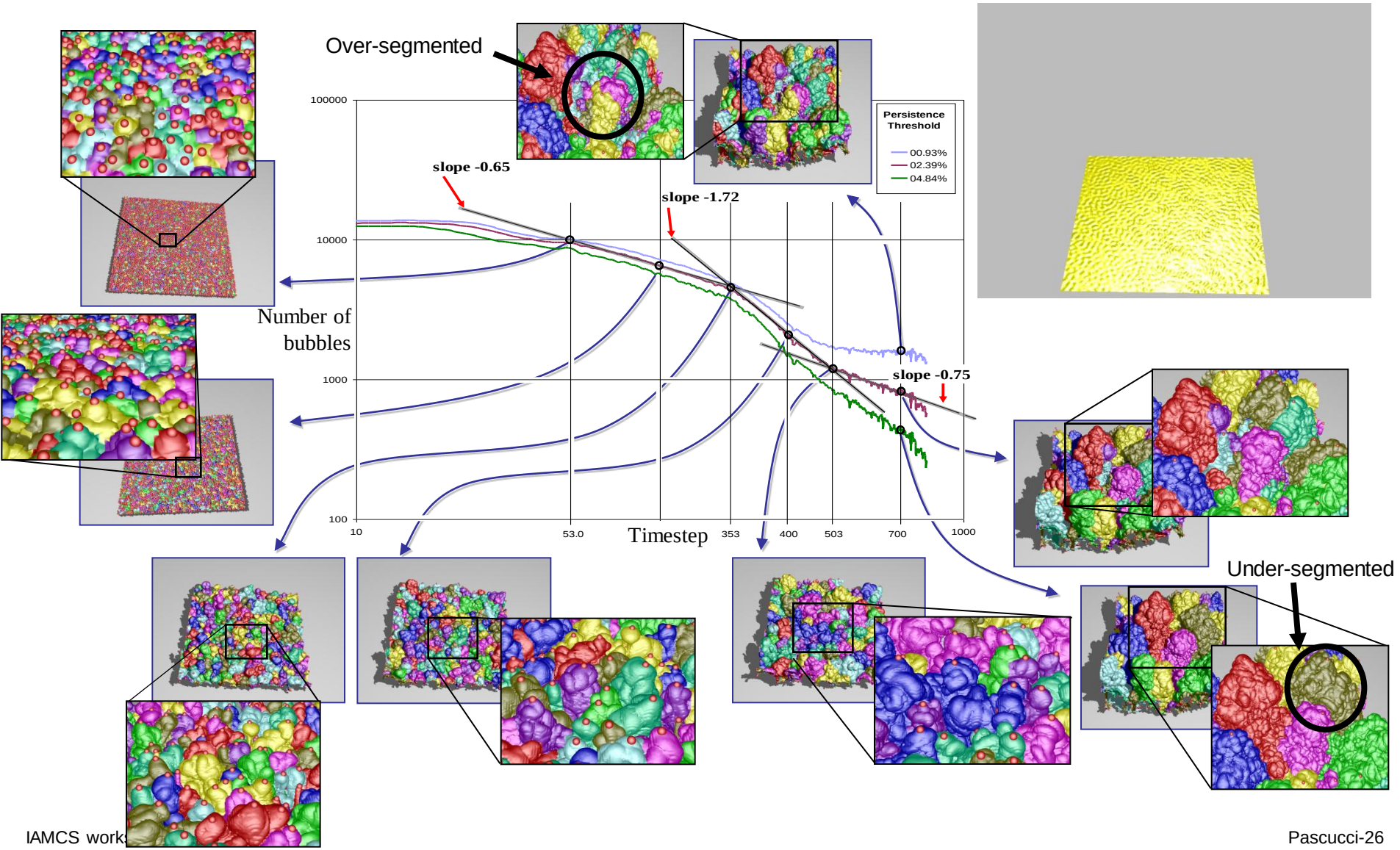
geometric tracking

feedback for adjustment and revaluation

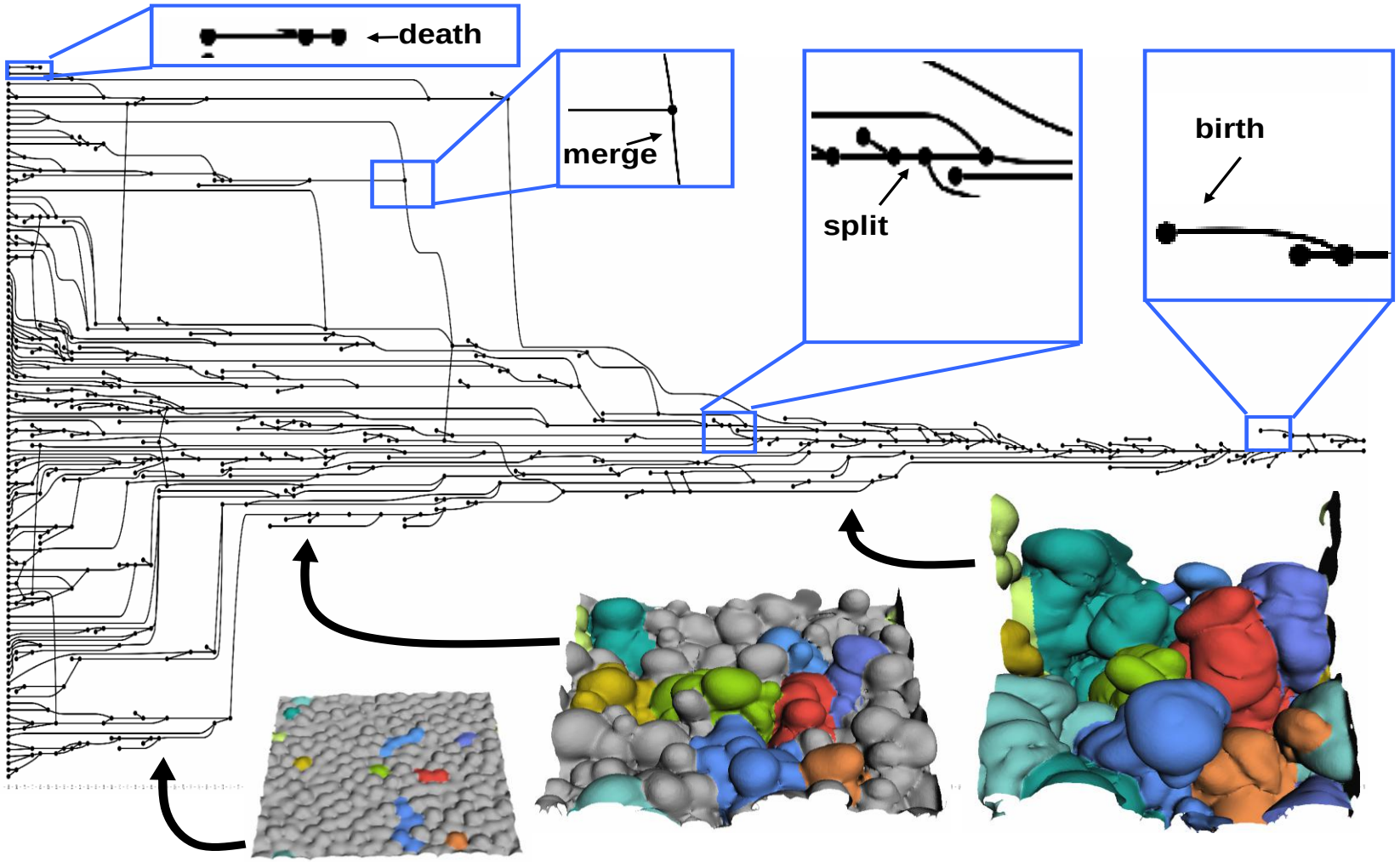
# The Segmentation Method is Robust From Early Mixing to Late Turbulence



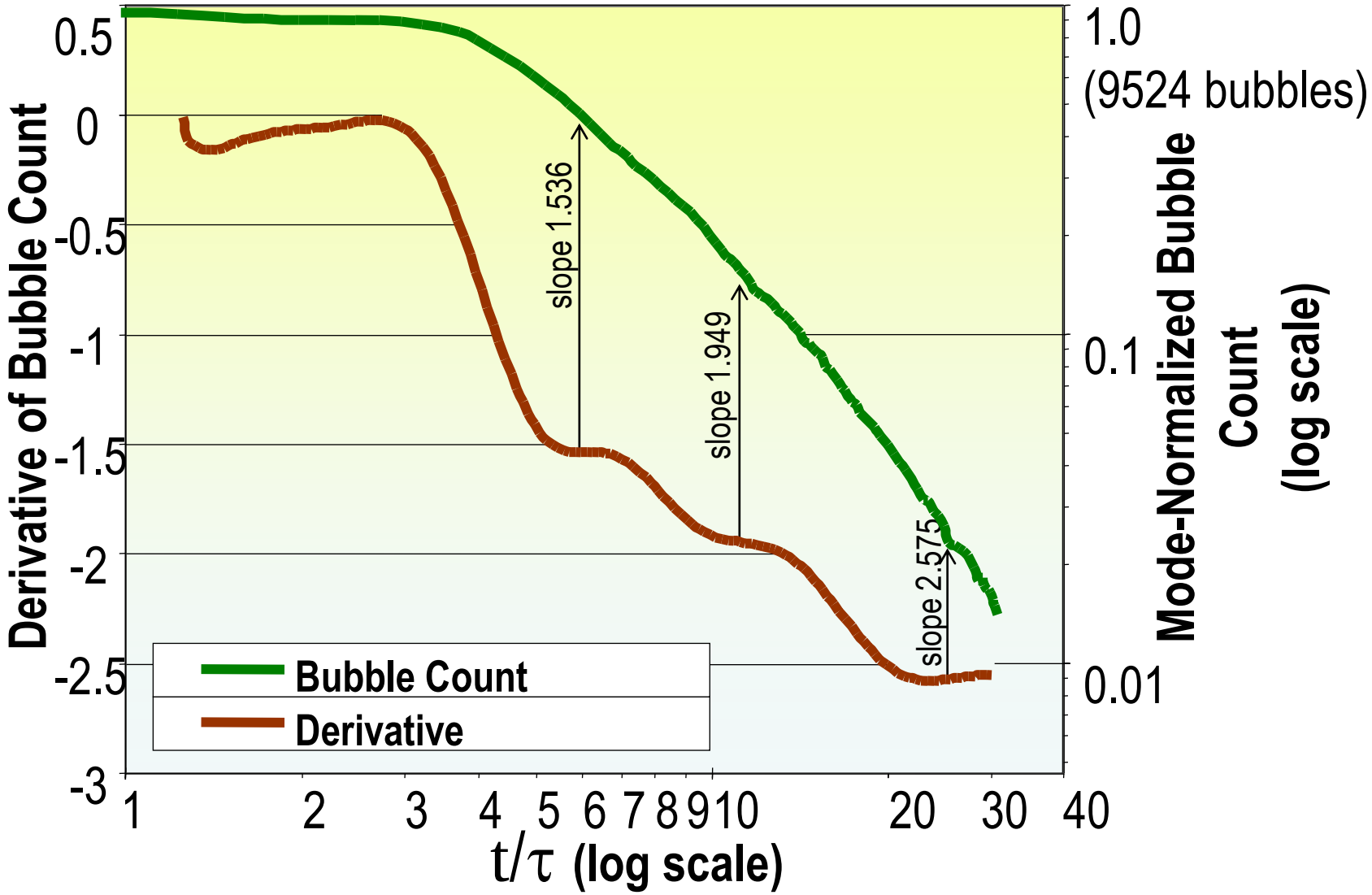
# We Evaluated Our Quantitative Analysis at Multiple Scales



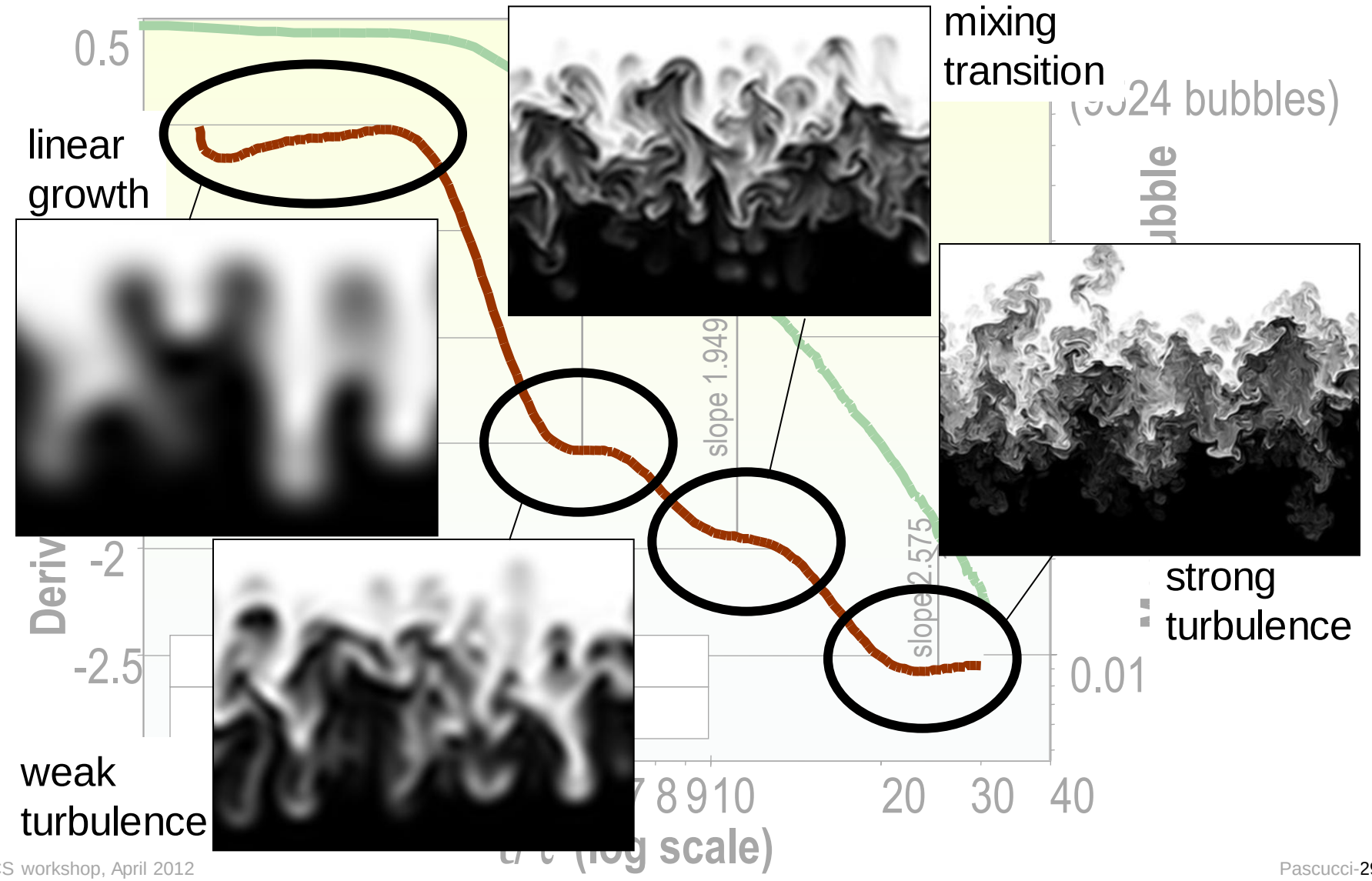
# We Characterize Events that Occur in the Mixing Process



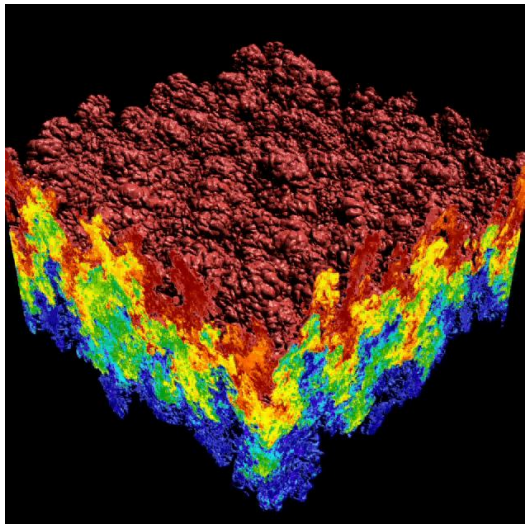
# First Time Scientists Can Quantify Robustly Mixing Rates by Bubble Count

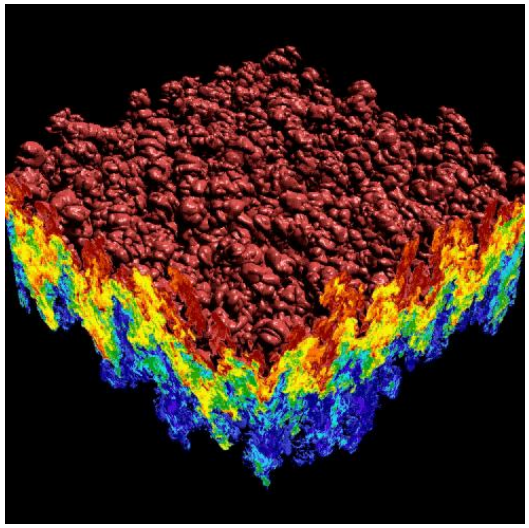


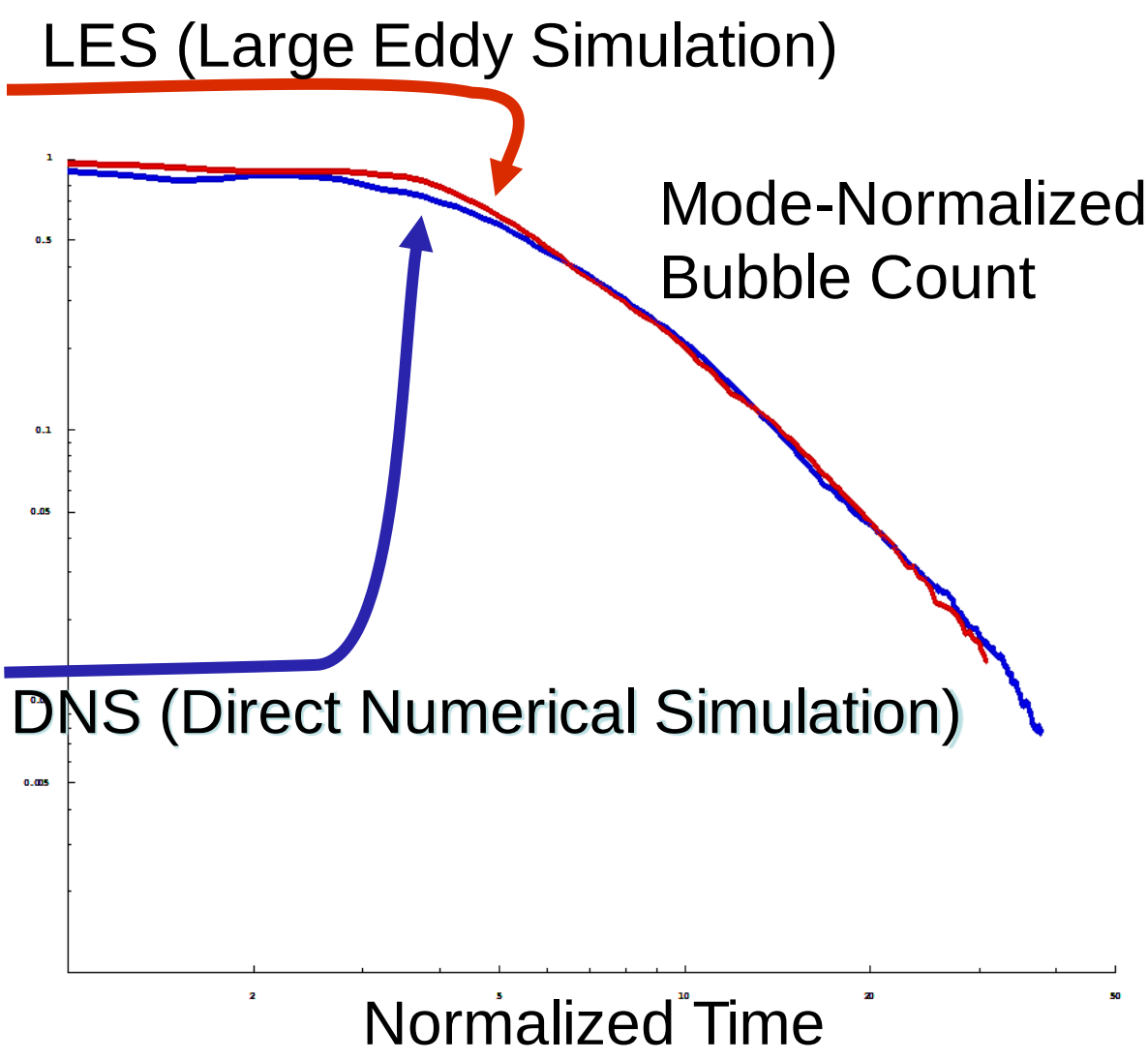
# We Provide the First Quantification of Known Stages of the Mixing Process



# We Provided the First Feature-Based Validation of a LES with Respect to a DNS

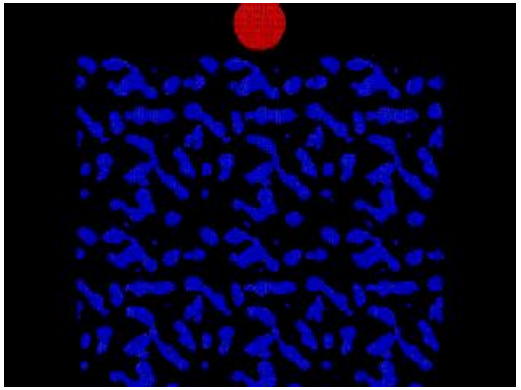


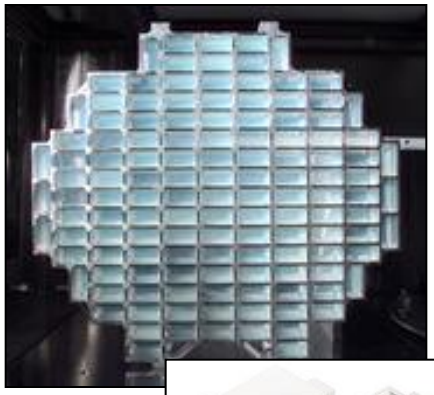


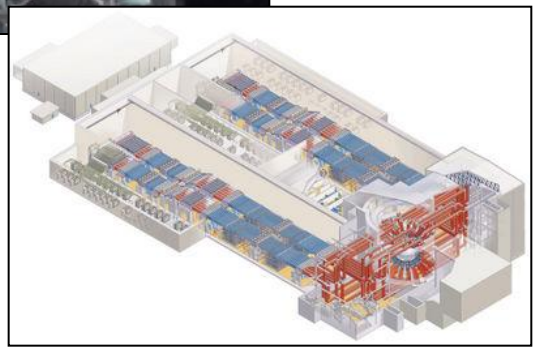


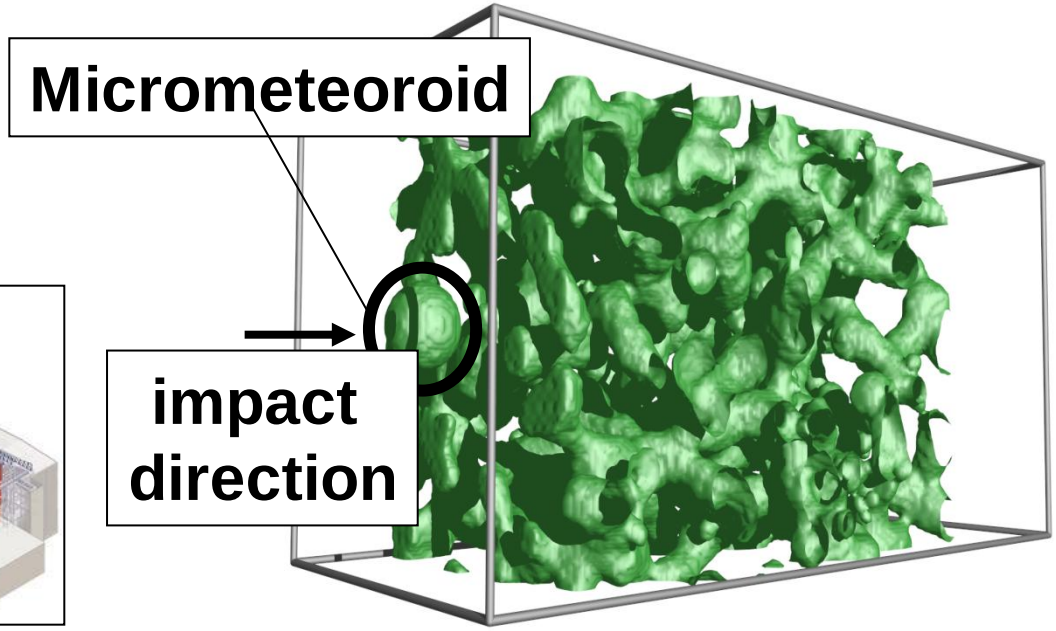
# Quantitative Analysis of the Impact of a Micrometeoroid in a Porous Medium

- **Many possible applications:**
  - **NASA's Stardust Spacecraft**
  - **National Ignition Facility Targets**
  - **Light and Robust Materials**
  - **many more...**

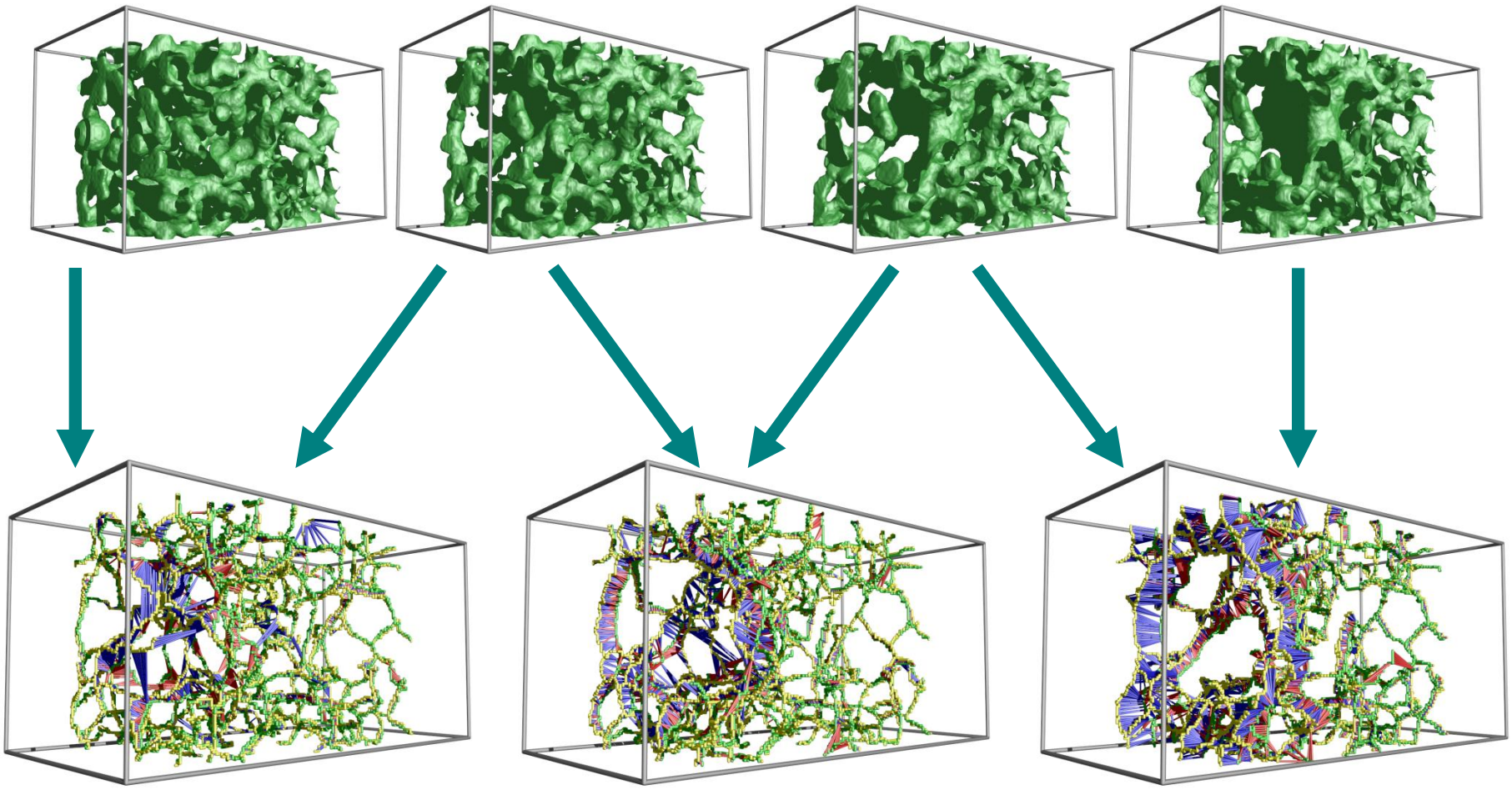








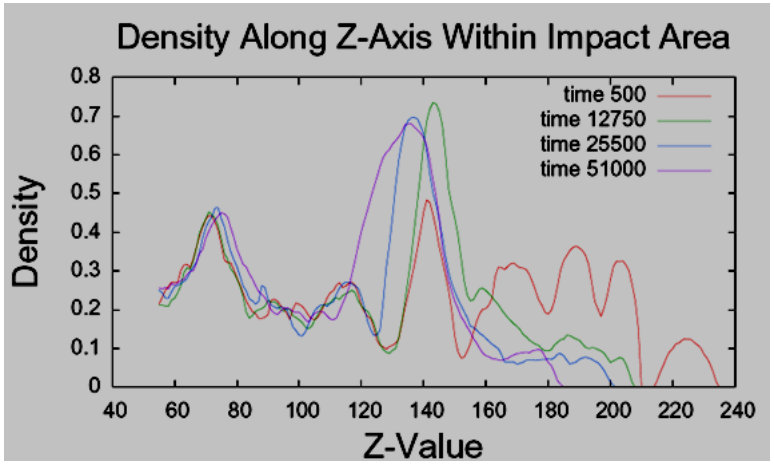
# We Track the Evolution of the Filament Structure of the Material Under Impact

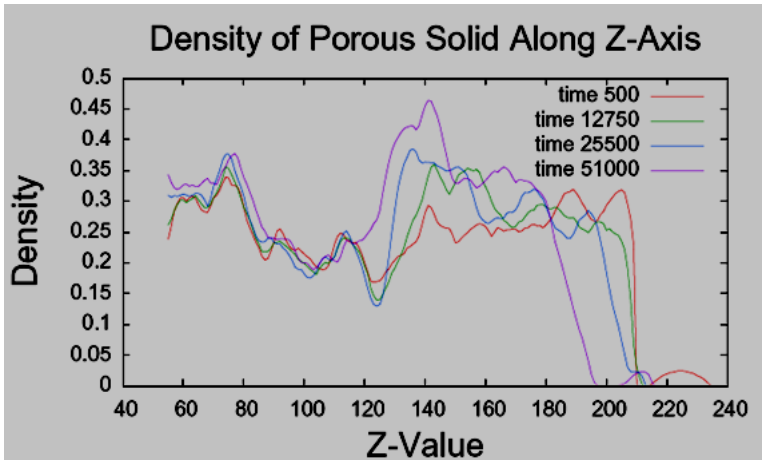


Time comparison of the reconstructions

# The Extracted Structures Allow to Quantify the Change in Porosity of the Material

## Density profiles

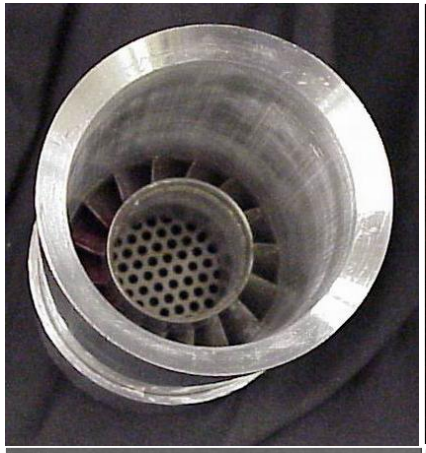


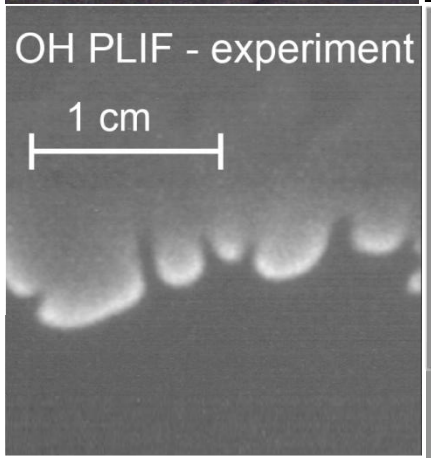



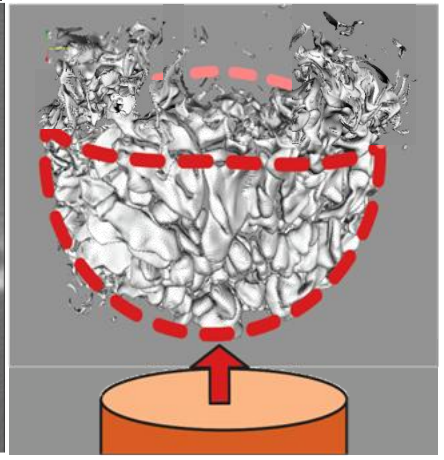

## Decay in porosity of the material

| Metric | t=500 | t=12750 | t=25500 | t=51000 |
|---|---|---|---|---|
| # Cycles | 762 | 340 | 372 | 256 |
| Total Length | 34756 | 24316 | 23798 | 18912 |

# Understanding Turbulence for Low Emission, High Efficiency Combustion





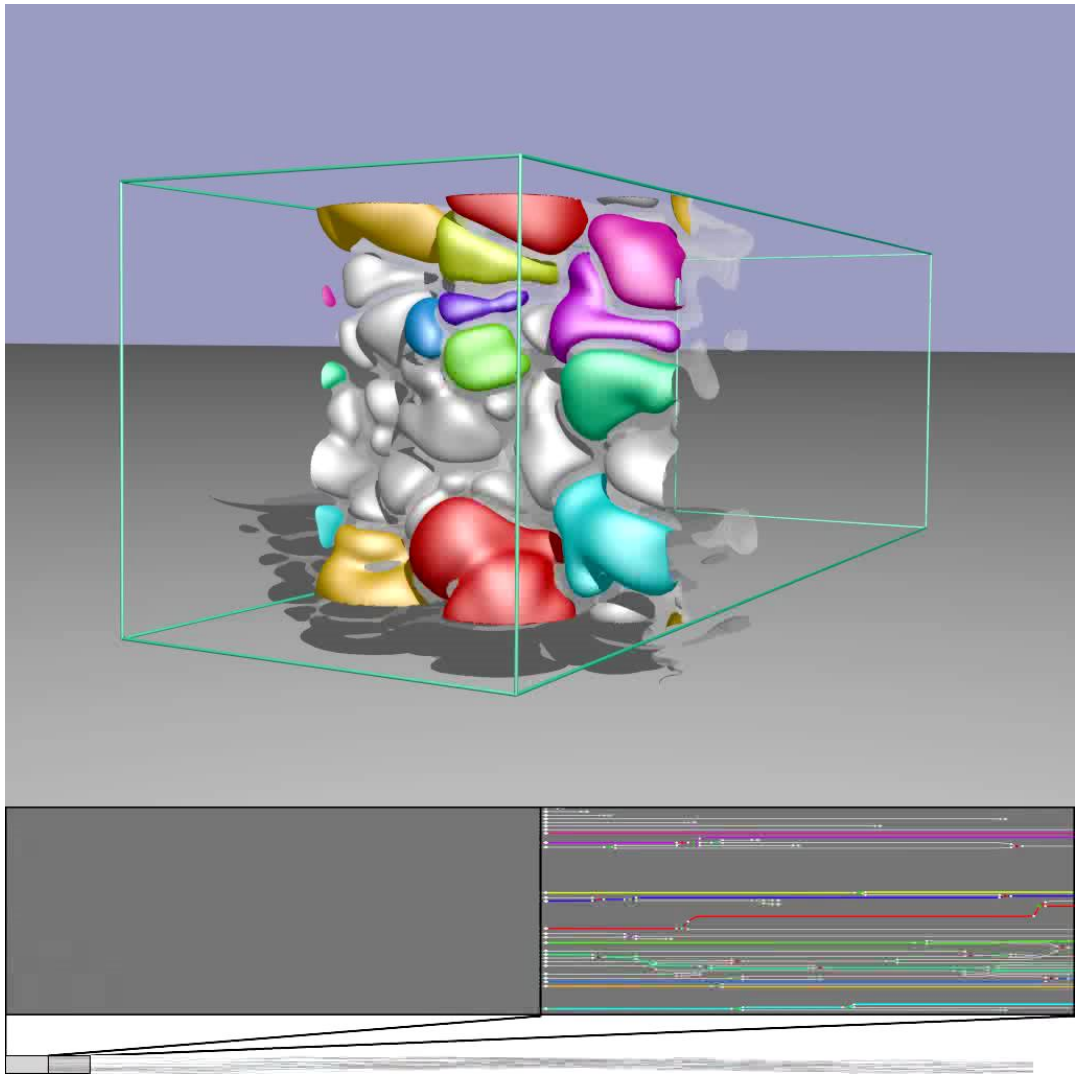
**Experiment** **Simulation**

- Lean premixed $H_2$ flames
- Low Swirl Combustion (LSC) Burners
- Low pollution in energy production
- High Efficiency in fuel consumption
- Scalable from residential to industrial use
- Each variable 3.9-4.5 TB

1" burner (5 kW, 17 KBtu/hr)

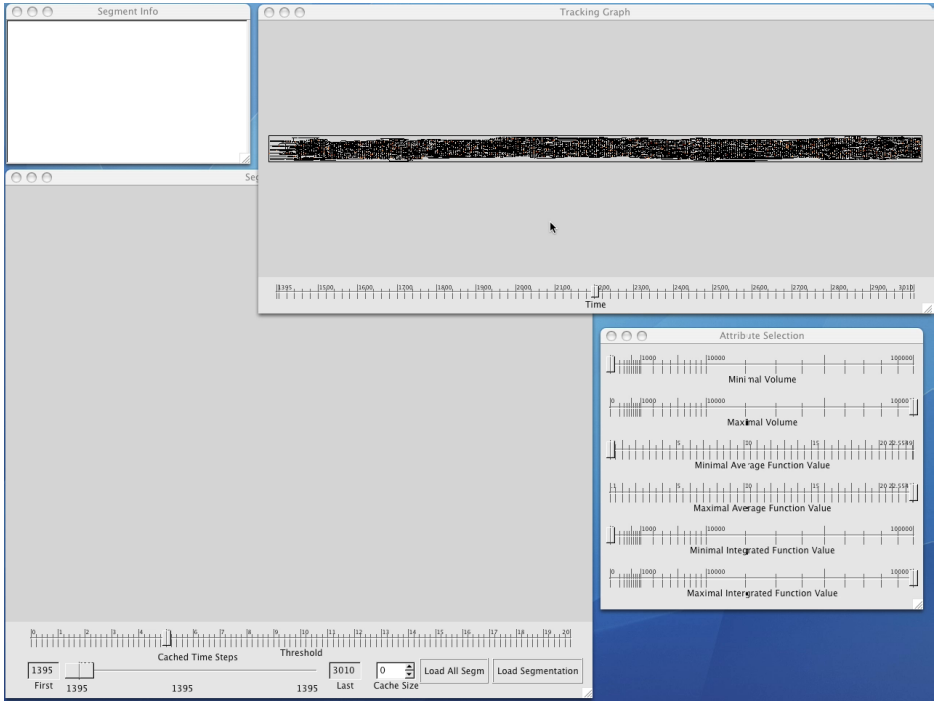
28" burner (44 MW, 150 MBtu/hr)

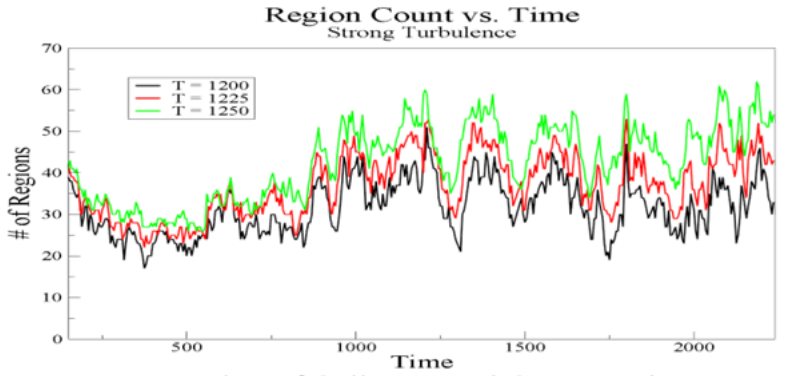
# Each Set of Parameters Results in a Robust Segmentation and Tracking of Burning Cells

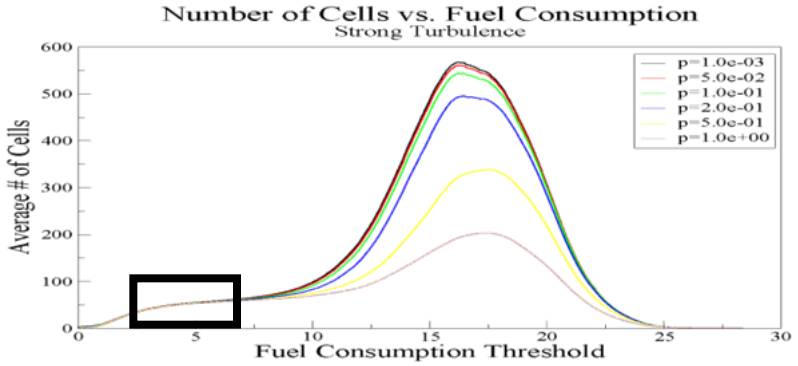


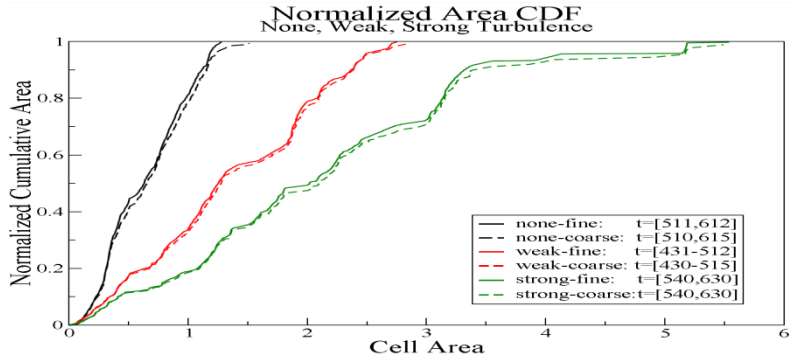
# We Allow Exploration of the Full Space of Parameters Defining the Features

- interactive exploration
- comparison of statistics

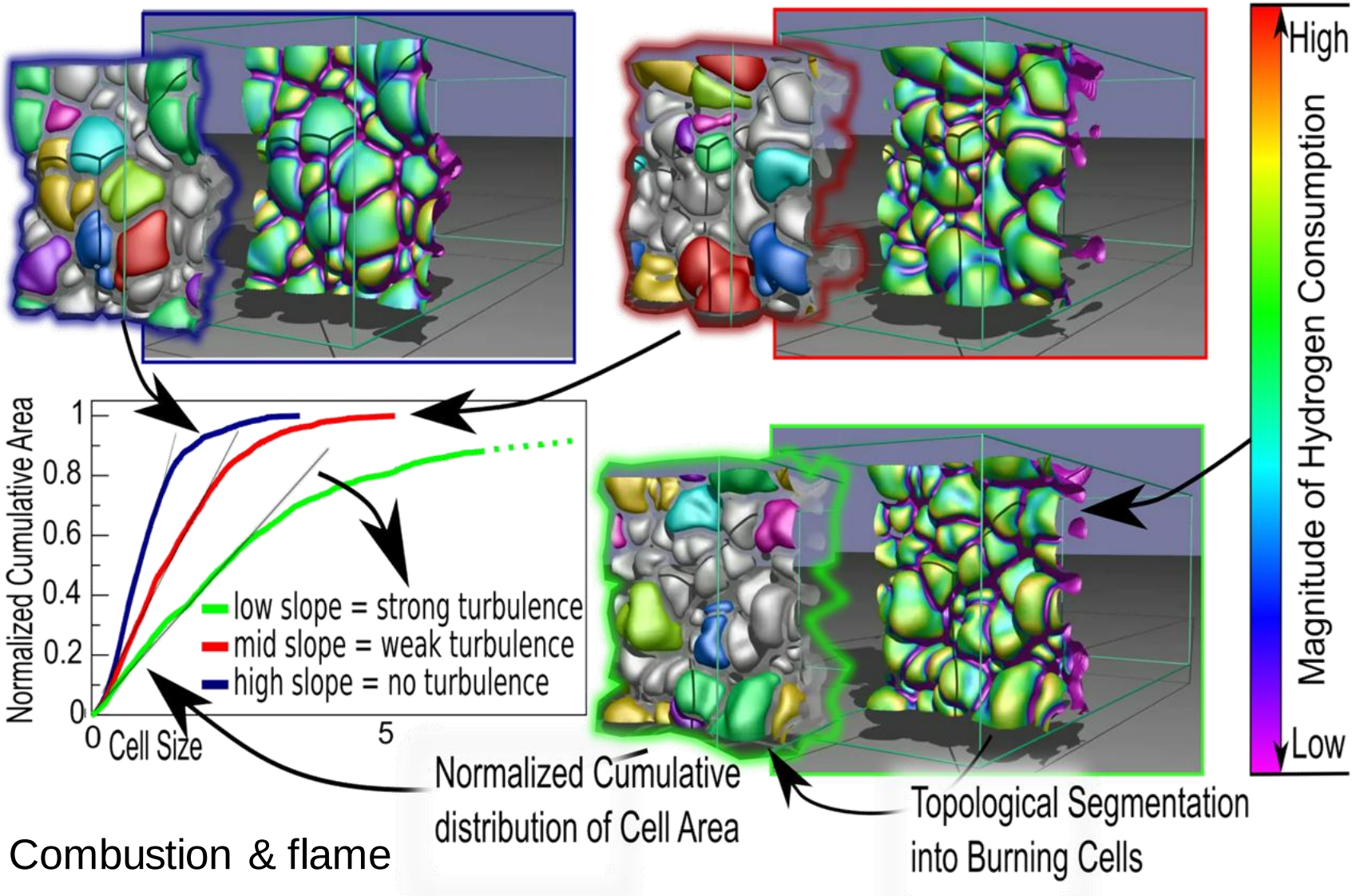


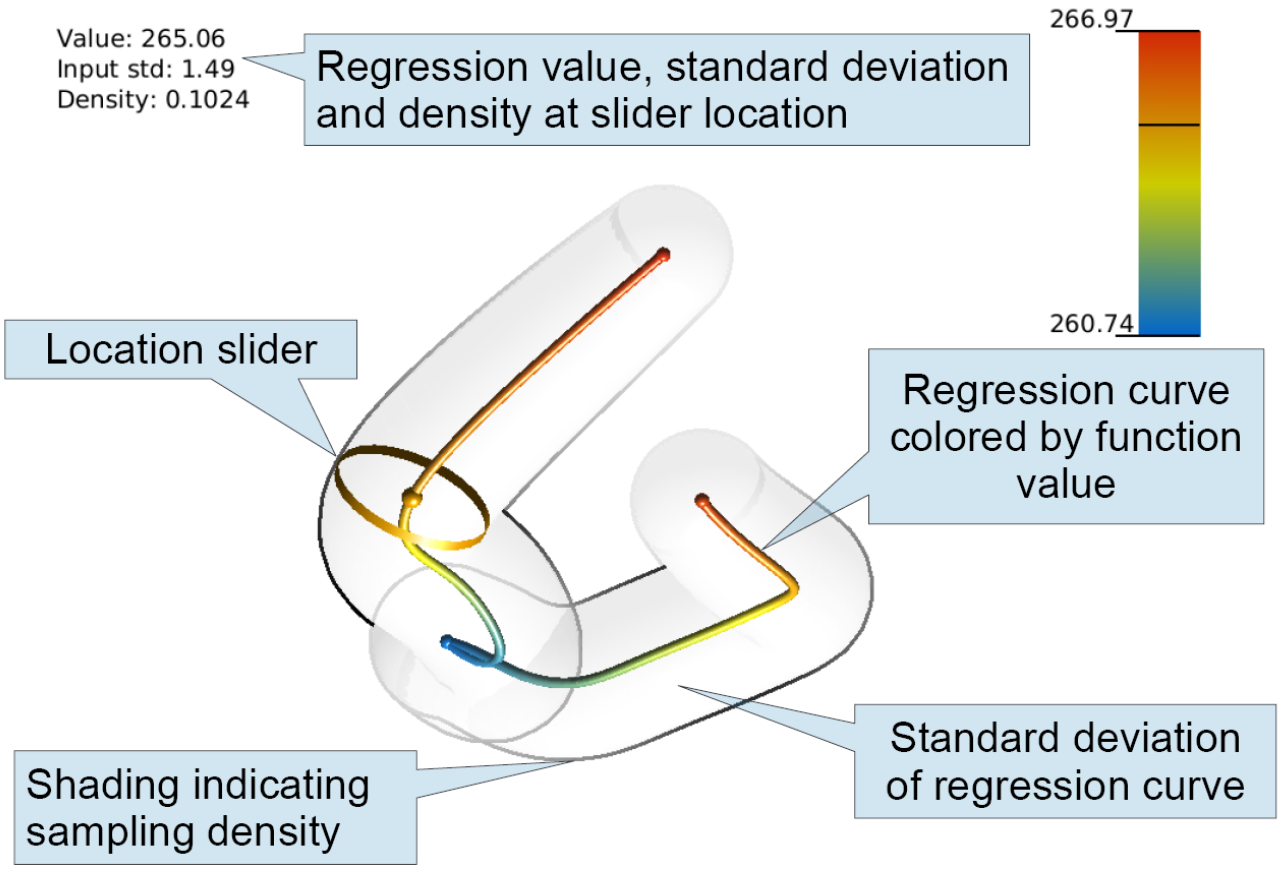
# Topological Segmentation Allows to Quantify Turbulence as Slope of the Area Distributions




- Combustion & flame

# Exploration of High Dimensional Functions for Sensitivity Analysis

## Integrated presentation of statistics and topology



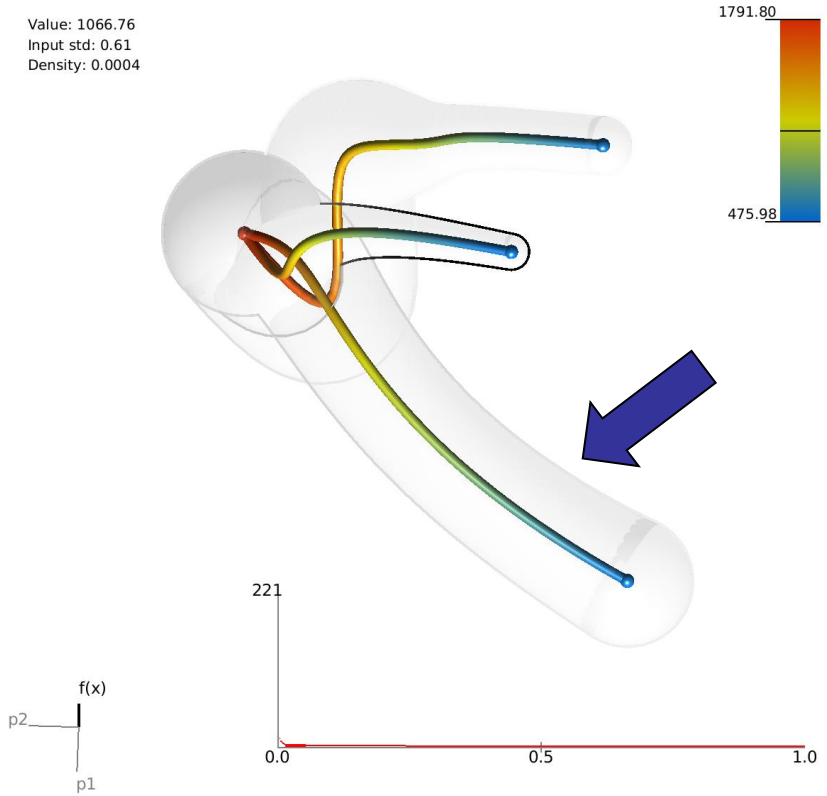
# Analysis of Combustion Simulations

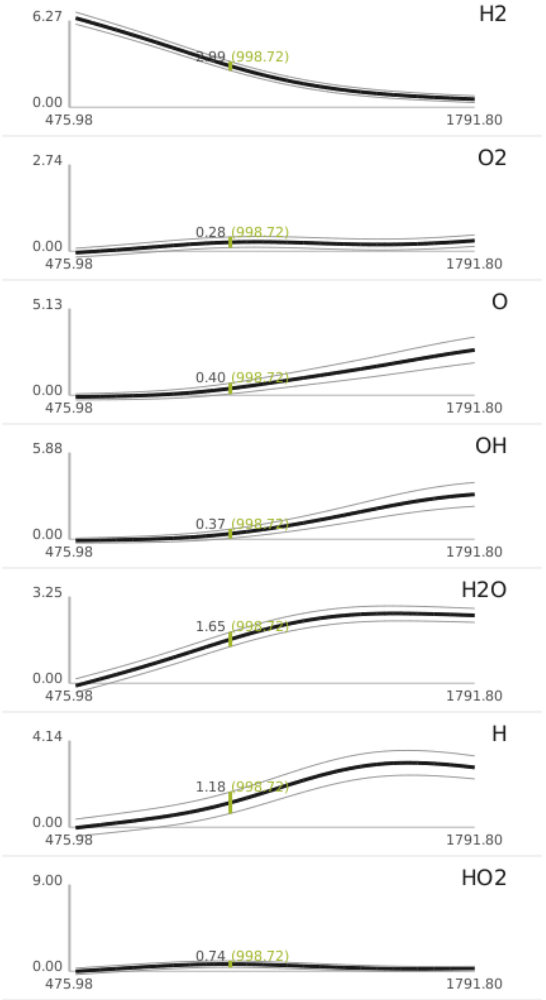
## Combustion Simulation of Jet CO/H2-Air Flames

**Input:** Composition of 10 chemical species

**Output:** Temperature

# The Framework Allows Detailed Visualization and Analysis of High Dimensional Functions

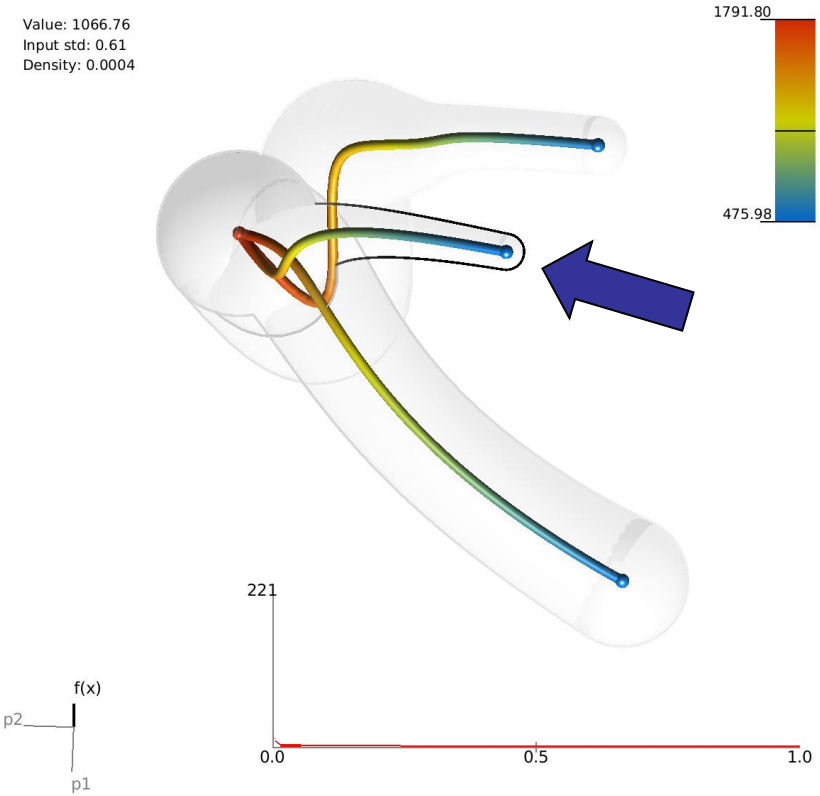


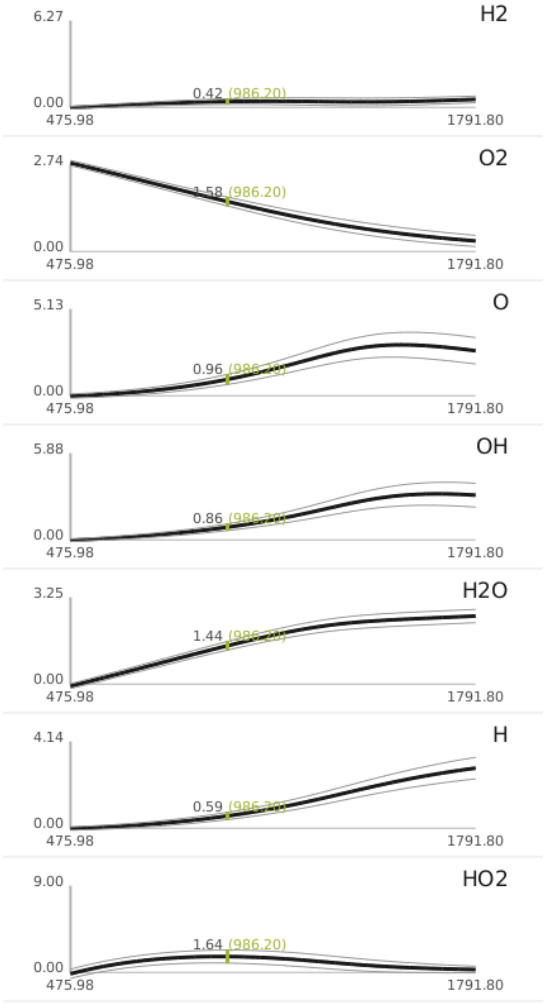




Pure fuel

10 dimensional data set describing the heat release wrt. to various chemical species in a combustion simulation

# The Framework Allows Detailed Visualization and Analysis of High Dimensional Functions

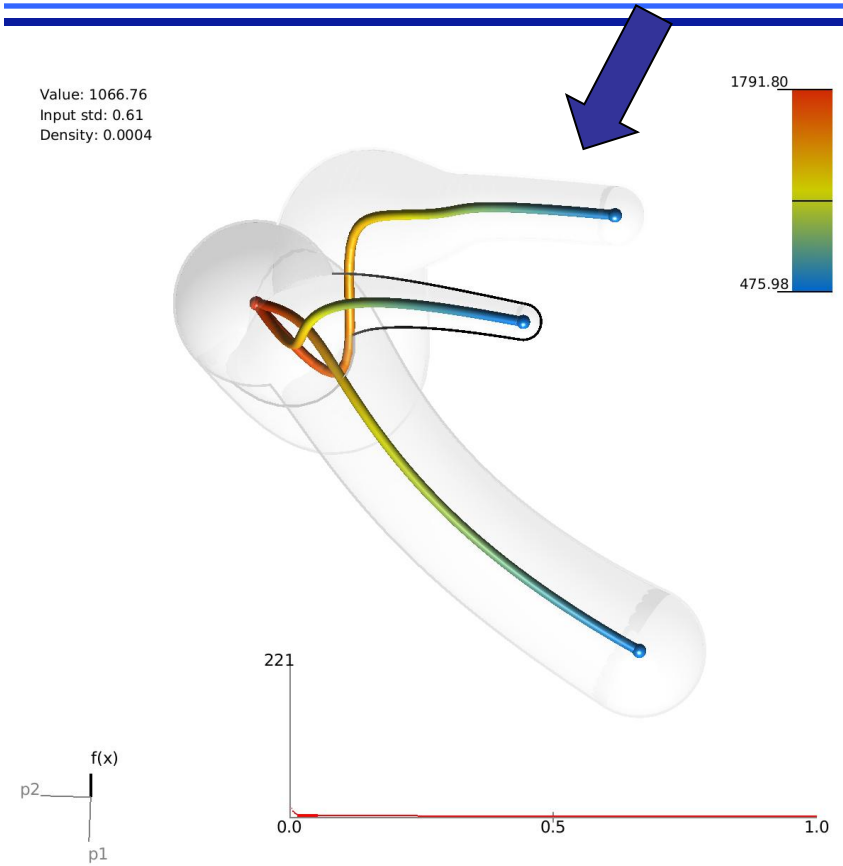


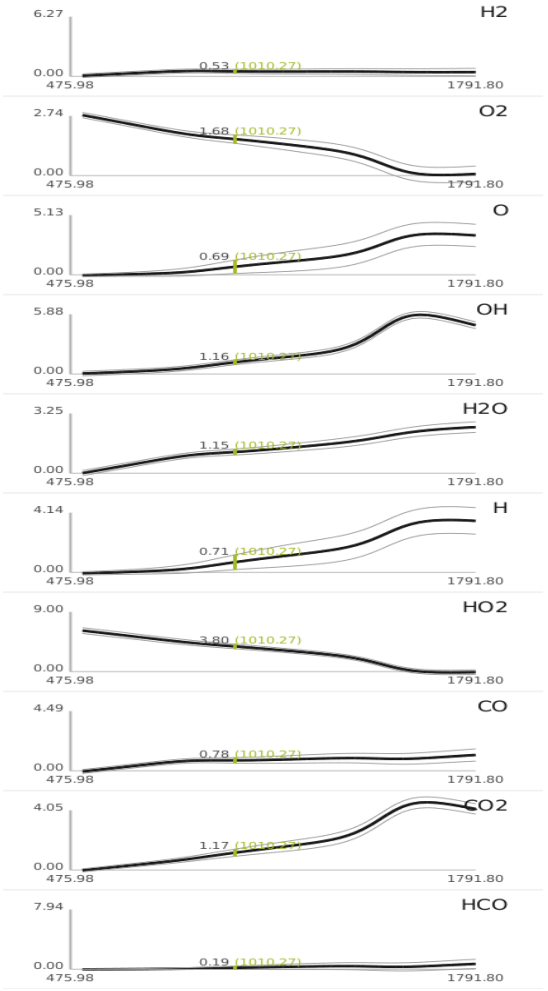




Pure oxidizer

10 dimensional data set describing the heat release wrt. to various chemical species in a combustion simulation

# The Framework Allows Detailed Visualization and Analysis of High Dimensional Functions

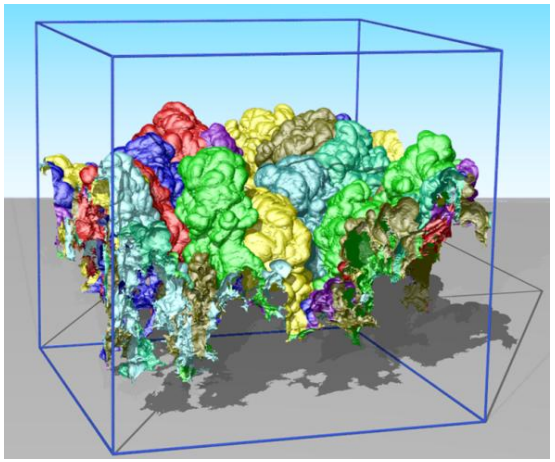


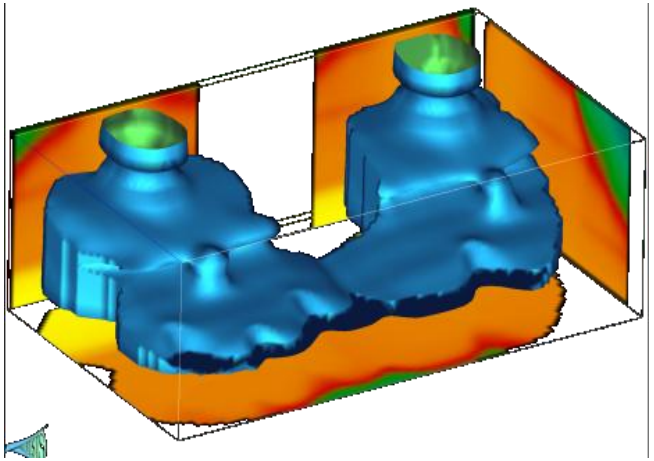




Local extinction

10 dimensional data set describing the heat release wrt. to various chemical species in a combustion simulation

# Center for Extreme Data Management, Analysis and Visualization

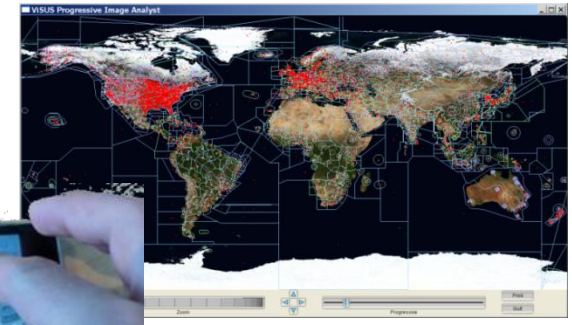



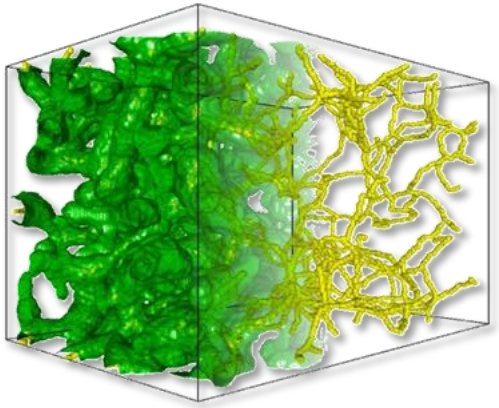


**A permanent center of the University of Utah within the Scientific Computing and Imaging Institute**

# Mission

**Research** Future Technologies for Knowledge Extraction from Extreme Sized Data



**Deployment and Application** of State of the Art Tools in Data Intensive Science Discovery

**Education** of the Next Generation Workforce Supporting Data Intensive Science and Engineering

# Center Structure

**Valerio Pascucci**
Director CEDMAV
Professor, SCI Institute,
School of Computing, University of Utah
Laboratory Fellow, Pacic Northwest National Laboratory

**Peer-Timo Bremmer**
Associate Director for Research

**Greg Jones**
Associate Director for Business and Commercialization

**Chris Johnson**
Distinguished Professor
SCI Institute
School of Computing

**Charles Hansen**
Professor
SCI Institute
School of Computing

**Mike Kirby**
Associate Professor
SCI Institute
School of Computing

**Suresh Venkatasubramanian**
Assistant Professor
School of Computing

**Mary Hall**
Associate Professor
School of Computing

**Feifei Li**
Assistant Professor
School of Computing

**Miriah Meyer**
Assistant Professor
School of Computing

**Jeff Phillips**
Assistant Professor
School of Computing

**Steve Corbató**
Executive Director, Cyberinfrastructure,
University Information Technology
Adjunct Faculty, School of Computing

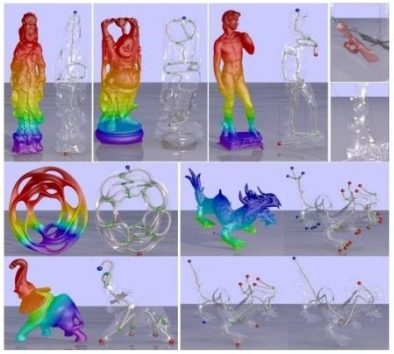
# Partnerships

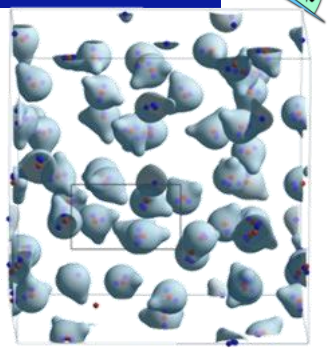
$1.6B NSA data center
(1.5 million-square-foot facility)

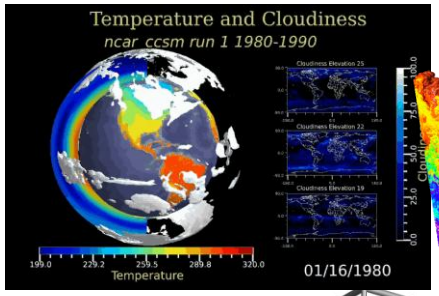
**IRTG** International Research Training Group

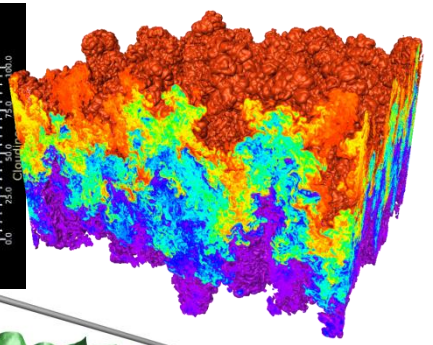
# A Data Analysis and Visualization Center Can be a Catalyst for a Virtuous Cycle of Collaborative Activities

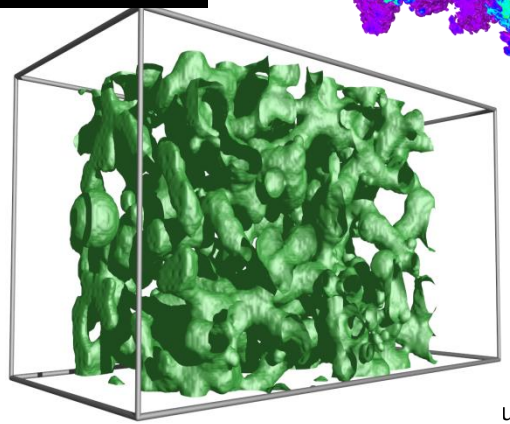
- **Tight cycle of :**
  - **basic research,**
  - **software deployment**
  - **user support**
- **Coordination among eight projects:**
  - **unified techniques for several applications**
- **Strong University-Lab-Industry collaboration**
- **Focused technical approach:**
  - **performance tools for fast data access**
  - **general purpose data exploration**
  - **error bounded quantitative analysis**
  - **feature extraction and tracking**
- **Interdisciplinary collaboration with domain scientists (from math to physics):**
  - **motivating the work**
  - **formal theoretical approaches**
  - **feedback to specific disciplines**





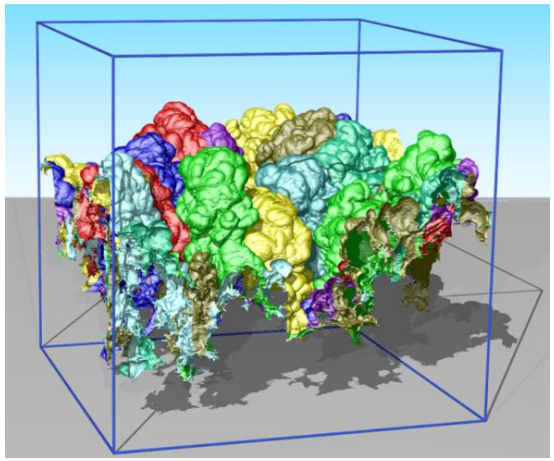


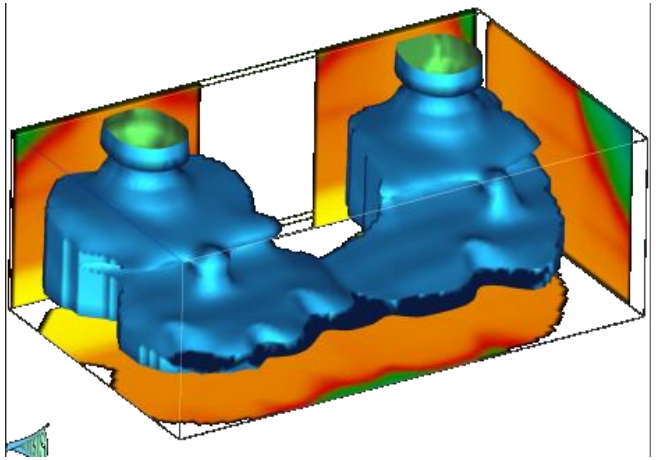





CEDMAV

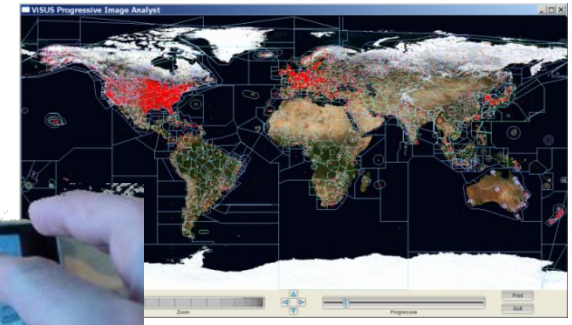
# Center for Extreme Data Management, Analysis and Visualization

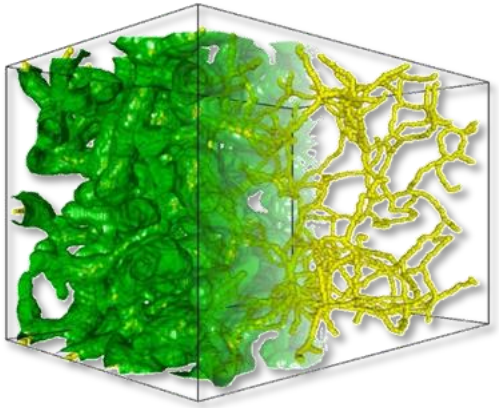





# Mission

**Research** Future Technologies for Knowledge Extraction from Extreme Sized Data



**Deployment and Application** of State of the Art Tools in Data Intensive Science Discovery

**Education** of the Next Generation Workforce Supporting Data Intensive Science and Engineering

# Center Structure

Valerio Pascucci
Director,
CEDMAV

Advisory Board
David Pershing, SVP UofU
Chris Johnson, SCI
Additional members

CEDMAV Administration
Greg Jones
Brenda Peterson

Chris Johnson
Distinguished Professor
SCI Institute
School of Computing

Charles Hansen
Professor
SCI Institute
School of Computing

Mike Kirby
Associate Professor
SCI Institute
School of Computing

Suresh Venkatasubramanian
Assistant Professor
School of Computing

Mary Hall
Associate Professor
School of Computing

$1.6B NSA data center
(1.5 million-square-foot facility)

Pacific Northwest NATIONAL LABORATORY

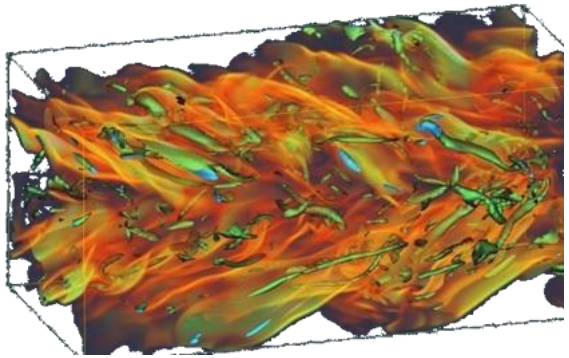
Sandia National Laboratories

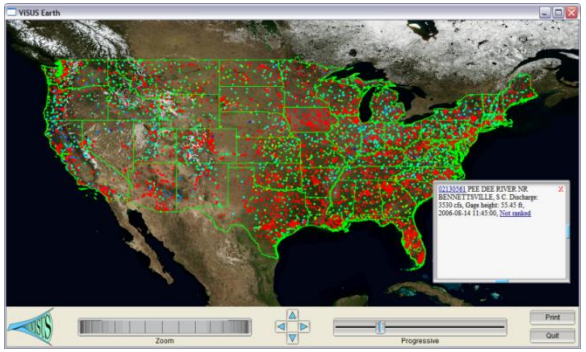
Lawrence Livermore National Laboratory

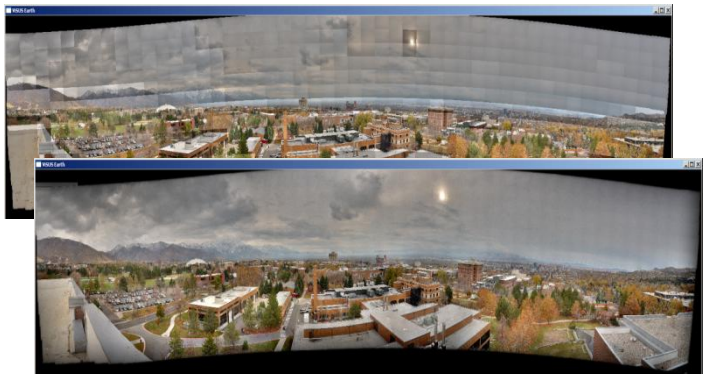
INL Idaho National Laboratory

# Research Scope

- **Data management , analysis and visualization for exploring and extracting knowledge from massive amounts of data (images, graphs, vector fields, text, geospatial, …..)**
- **Enable science discovery driven by large scale simulations and high resolution sensing devices**
- **Increase awareness of events and patterns underlying massive and complex data feeds**
- **Develop theoretical foundations for unified cross disciplinary technology**

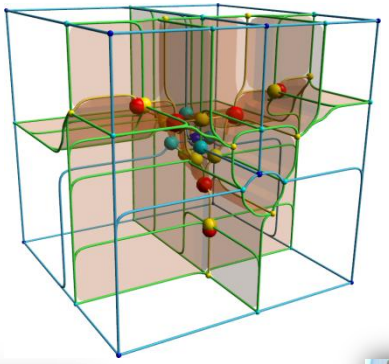


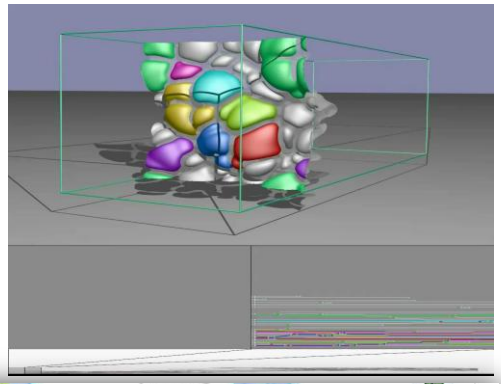


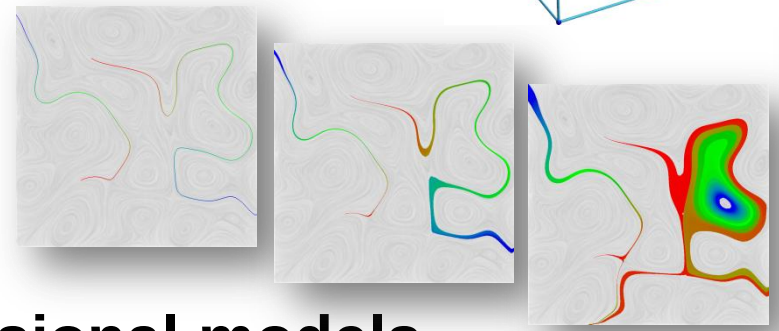



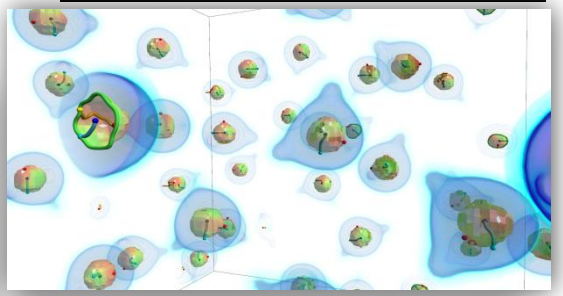
# Research Areas

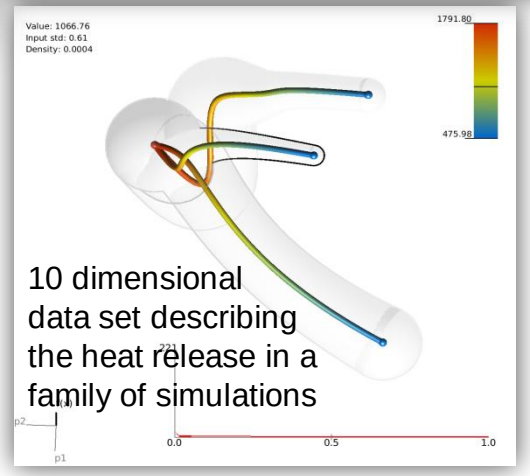
- Mathematics of scalar, vector, and tensor fields
- Discrete methods for graphs and text
- Uncertainty
- Statistics
- Topology
- High dimensional models
- Infrastructure for data movements
- Parallel computing for data analysis
- Scalable Visualization techniques
- Building data abstractions







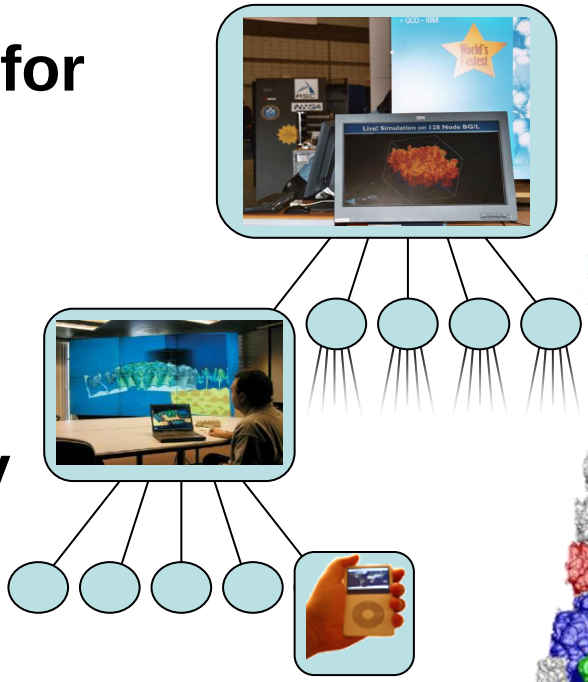


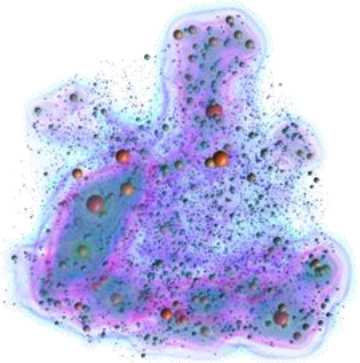


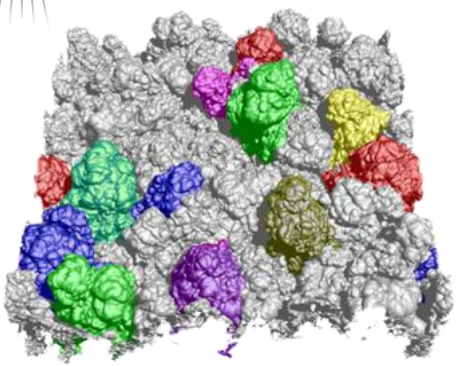

10 dimensional data set describing the heat release in a family of simulations

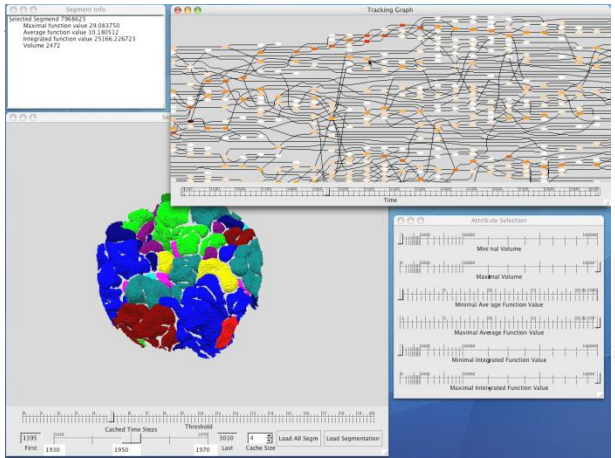
# Deployment

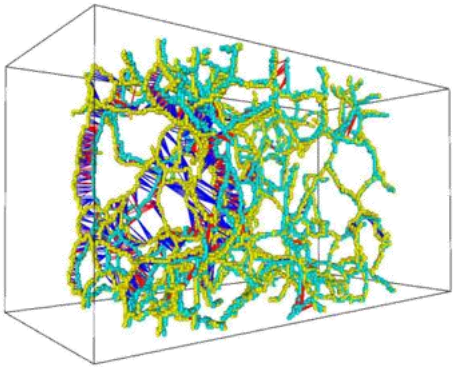
- **Software infrastructure for distributed data access and visualization**
- **Collaborative tools for large, distributed teams solving problems jointly**
- **Tracking of Information Decision making**
- **Scalable data analysi tools exploiting HPC resources**
- **Ad hoc interfaces for specialized tasks**











# Applications

- **Embedded collaboration with science teams to best design solutions**
- **Specialize software tools to address specific science challenges**
- **Develop focus areas such as:**
  - **Large scale image processing**
  - **Climate Modeling**
  - **Combustion Chemistry**
  - **Molecular Dynamics**
  - **Physics**
  - **Earth Sciences**
  - **Medicine**
  - **Power grid**
  - **…..**