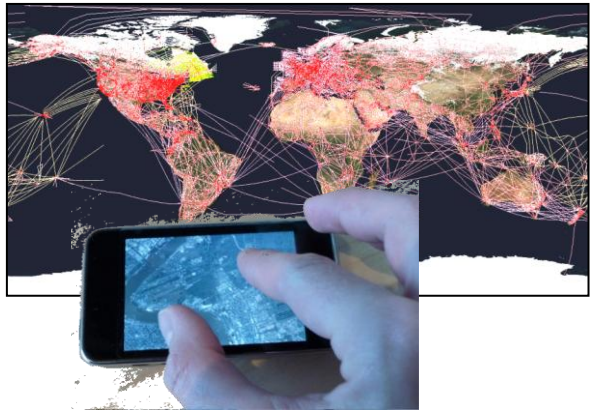
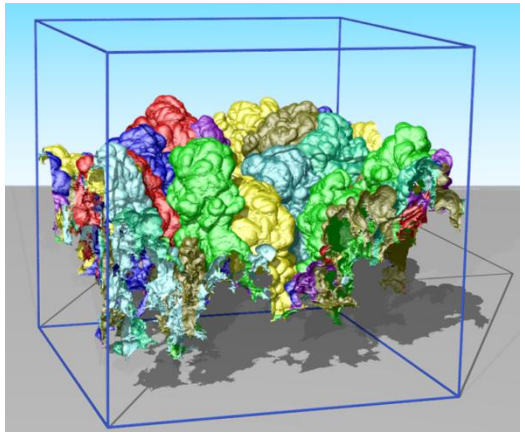
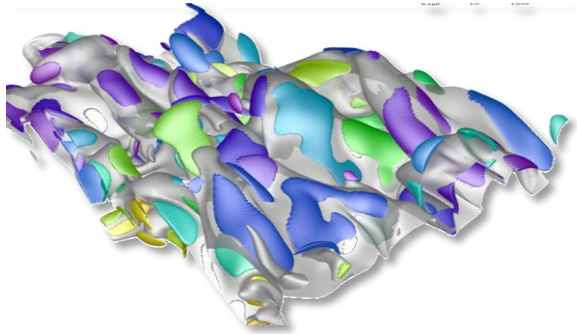
Massive Data Management, Analysis, and Visualization: Scaling From Handheld Devices to Supercomputers







Yarden Livnat

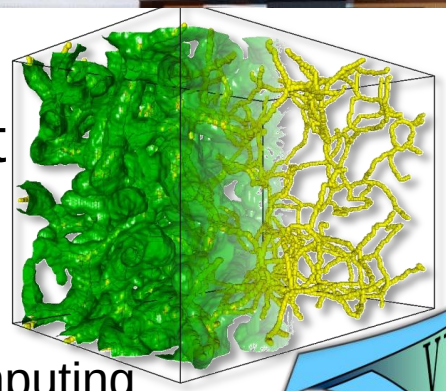
Senior Research Scientist

Valerio Pascucci

Director, CEDMAV

Professor, SCI Institute & School of Computing

Laboratory Fellow, PNNL

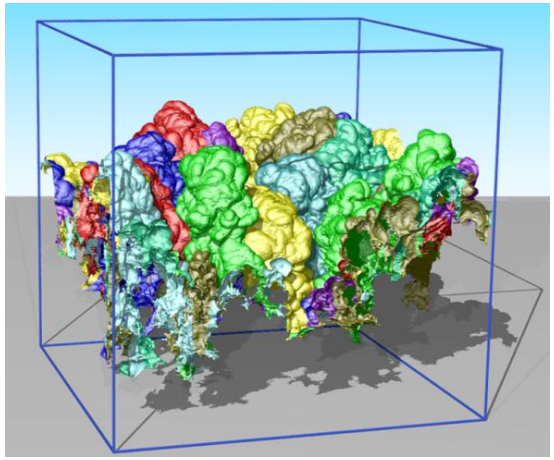


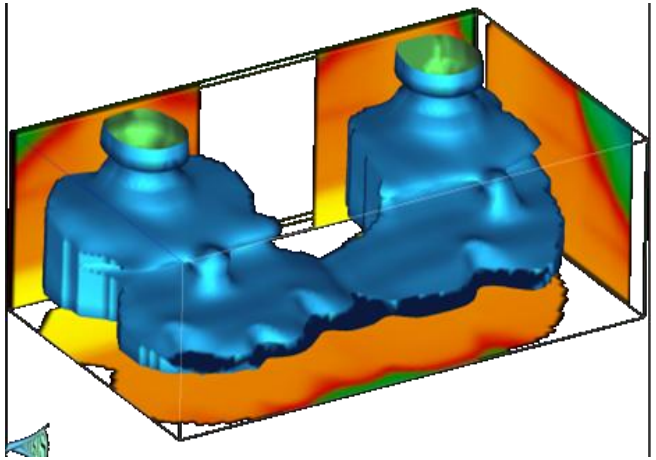
SCI INSTITUTE

VISUS

CEDMAV

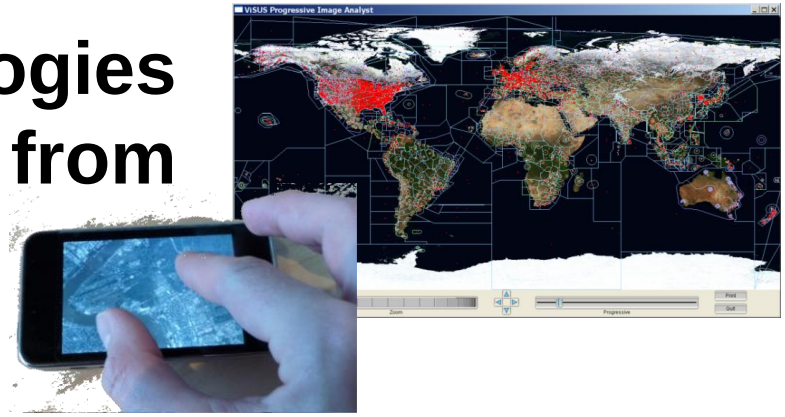
Center for Extreme Data Management, Analysis and Visualization

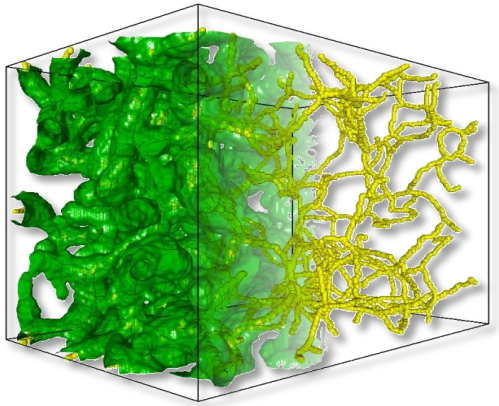




CEDMAV Mission

Research Future Technologies for Knowledge Extraction from Extreme Sized Data





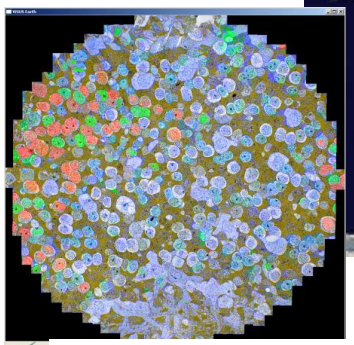
Deployment and Application of State of the Art Tools in Data Intensive Science Discovery

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

Massive Simulation and Sensing Devices Generate Great Challenges and Opportunities


BlueGene/L


Satellite

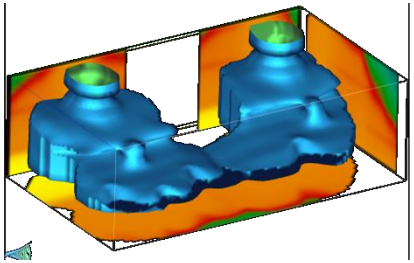

Retinal Connectome

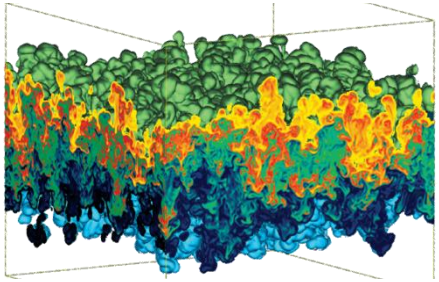

Earth Images

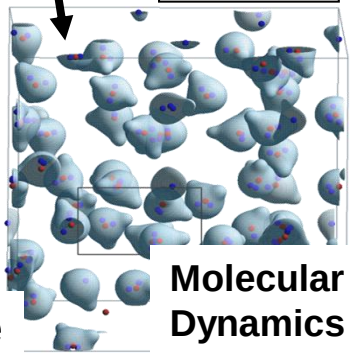

Cameras

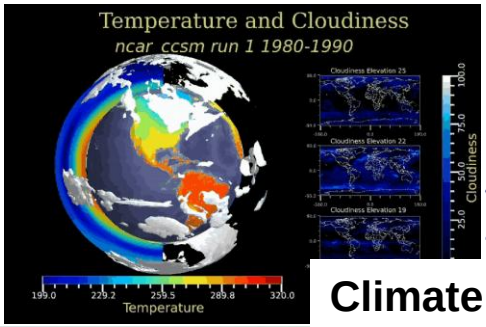

EM

Jaguar

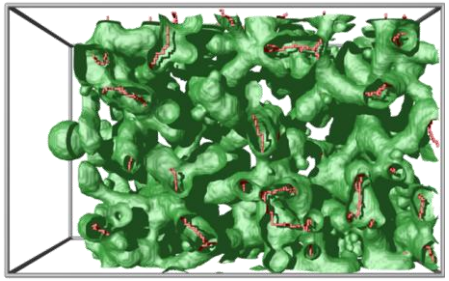

Carbon Seq. (Subsurface)

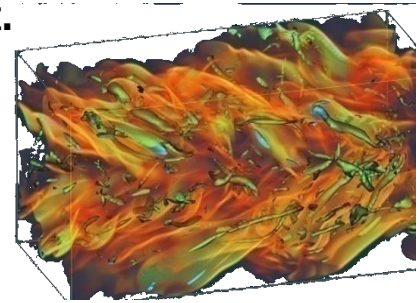

Hydrodynamic Inst.

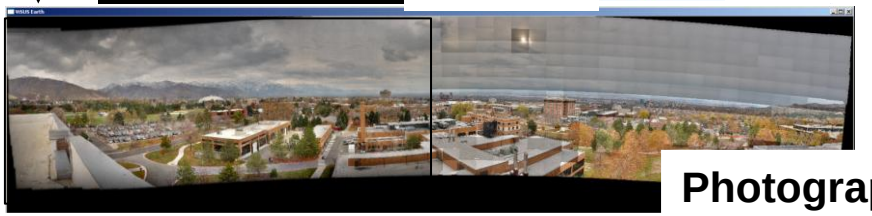

Molecular Dynamics



Climate

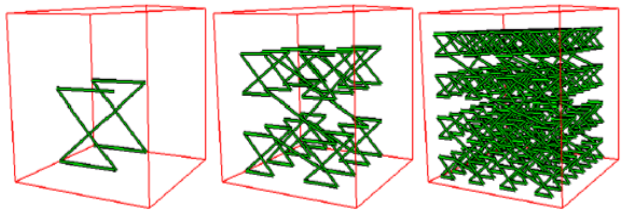

Porous Materials


Turbulent Combustion

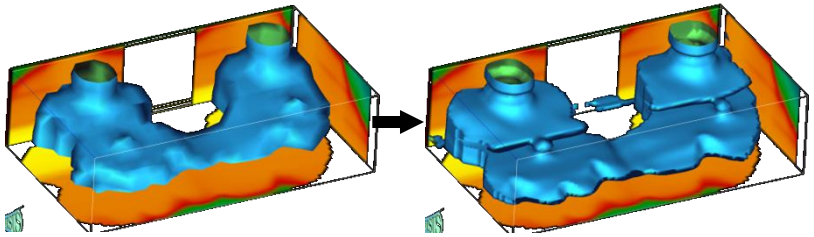

Photography

A Cyberinfrastructure Requires Efficient Data Management and Processing

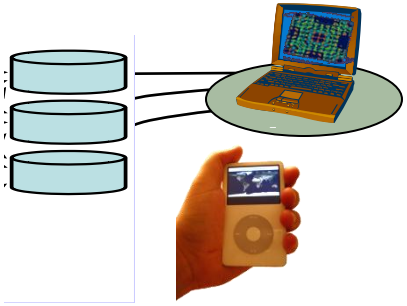
- **Advanced data storage techniques:**
  - **Data re-organization.**
  - **Compression.**



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

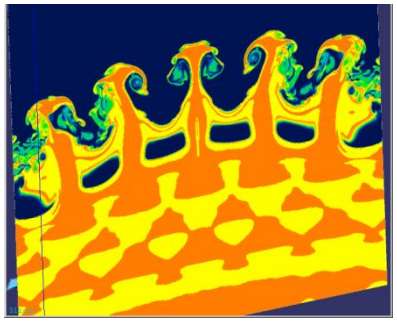


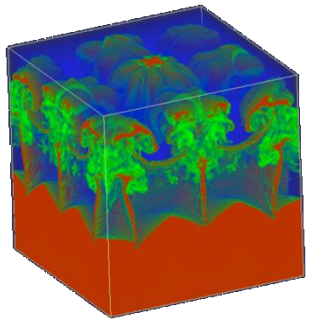
- **Scalability across a wide range of running conditions:**
  - **From laptop, to office desktop, to cluster of PC, to BG/L.**
  - **Memory, to disk, to remote data access.**

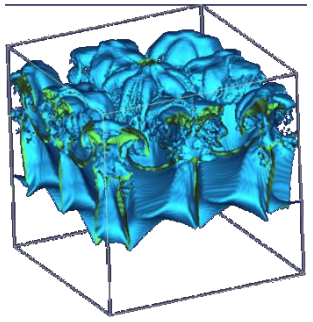


We Redesigned the Data Management and Visualization Pipeline with New Principles

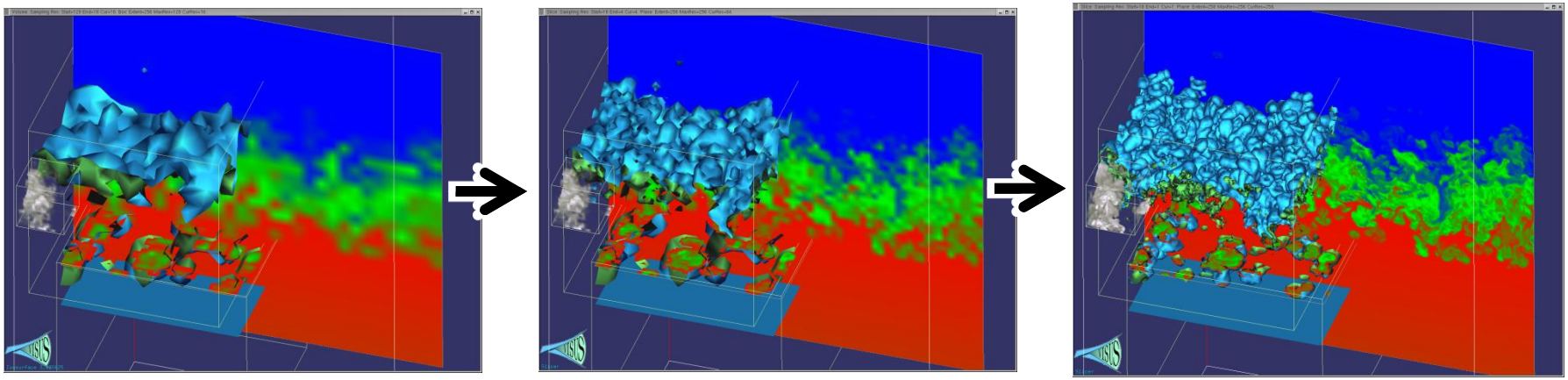
- **Basic core techniques:**
  - **Slicing**
  - **Volume rendering**
  - **Iso-surfaces**



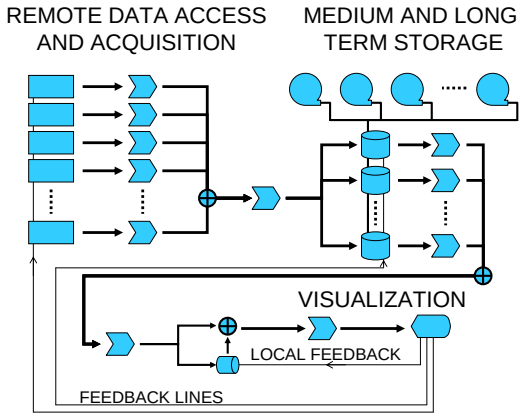


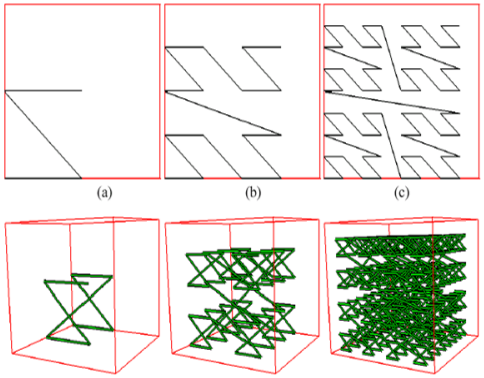


- **Cache-oblivious out-of-core processing optimizing access locality for any size of data blocks**
- **Pipelines of progressive algorithms**
- **Coarse-to-fine construction of multi-resolution models**
- **Remote data streaming**



We Consider the Three Main Components Defining a Computing Infrastructure





Processing Network
(Data Access Path)

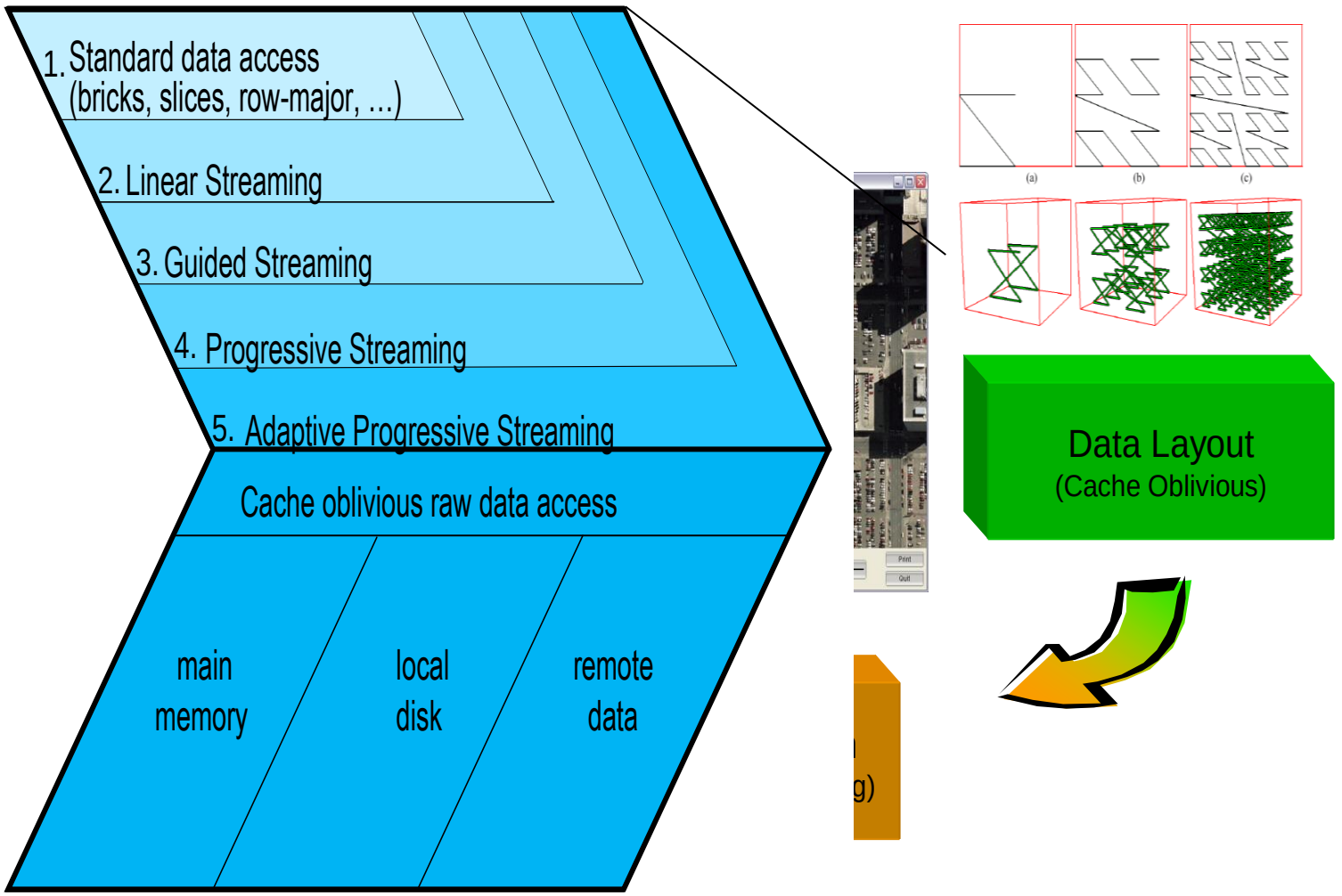
Data Layout
(Cache Oblivious)

Algorithm Design
(Progressive Processing)

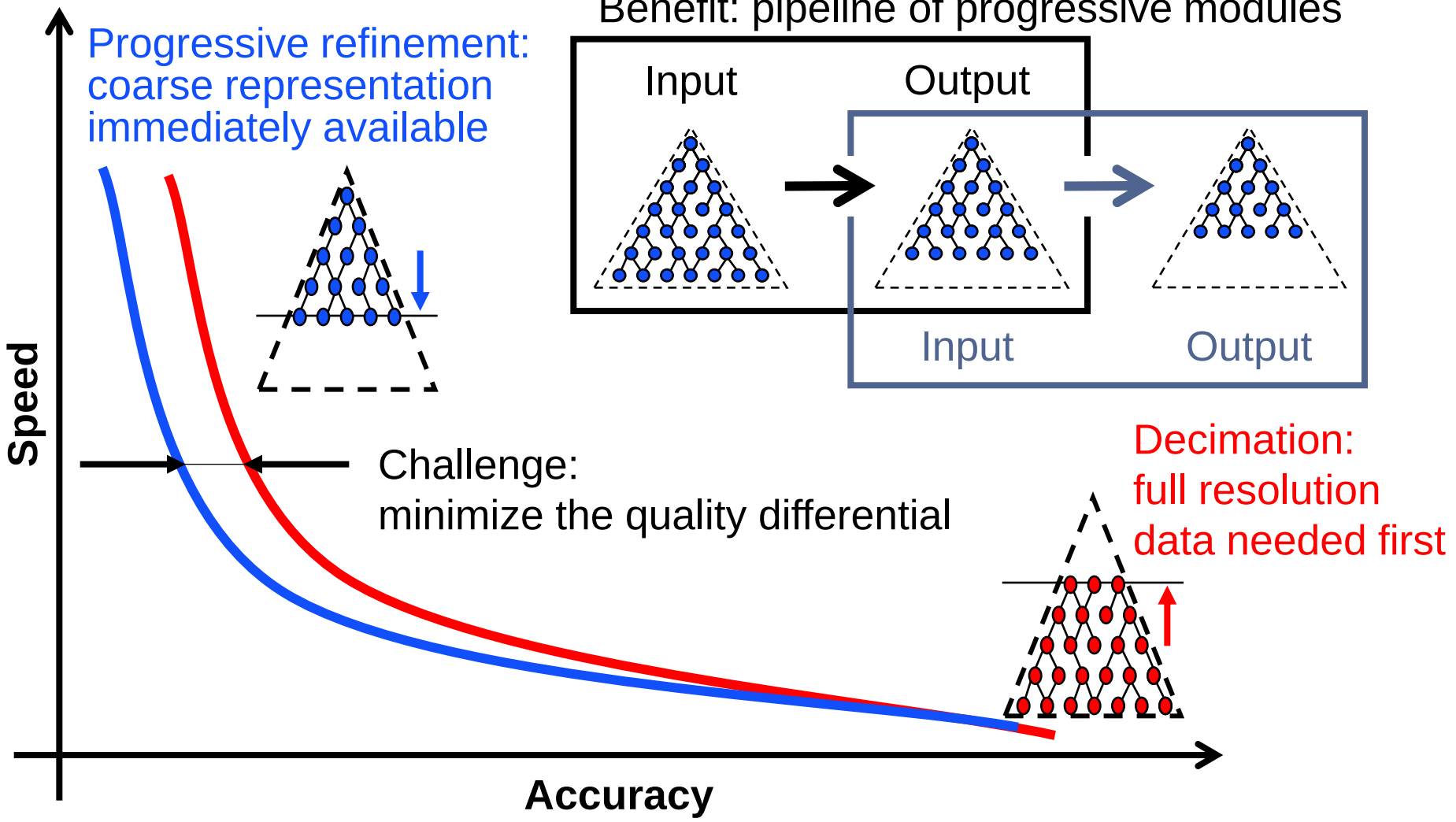

SCI
INSTITUTE


VISUS

We Characterize Algorithmic Classes Based on Effect in a Processing Network

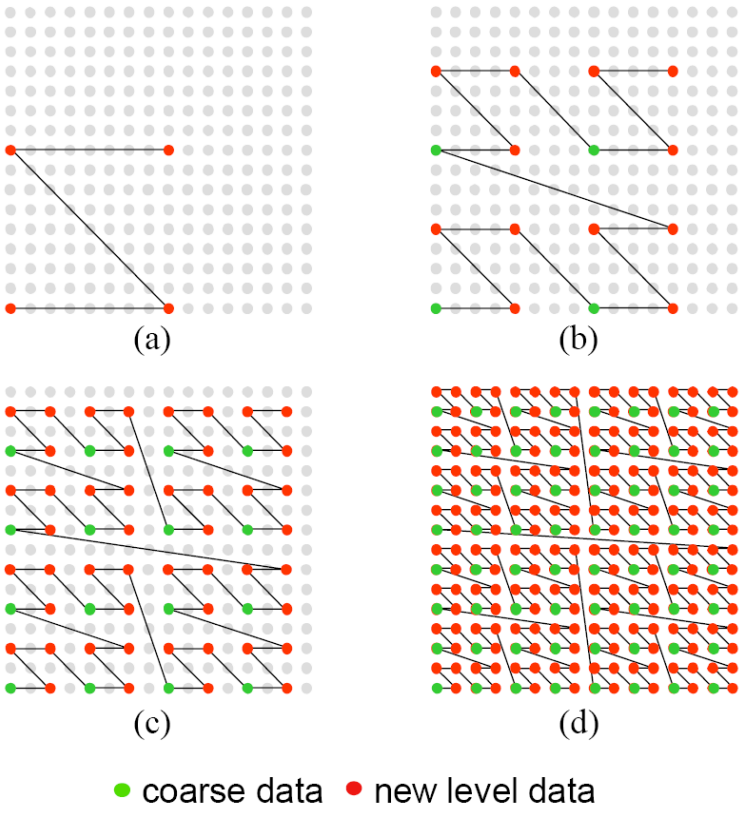


The use of top-down and bottom-up processes have a strong impact on the data stream

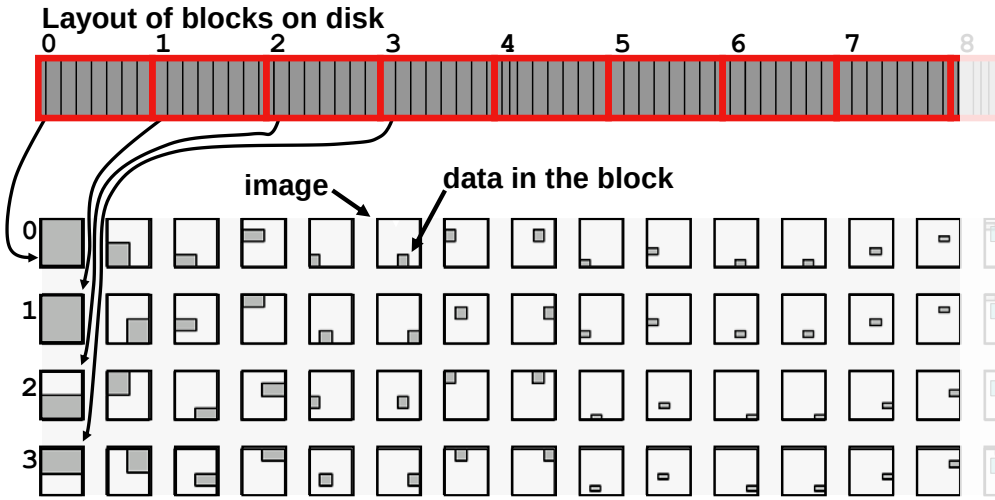


We Introduced Multi-resolution Cache Oblivious Layouts for Image Data

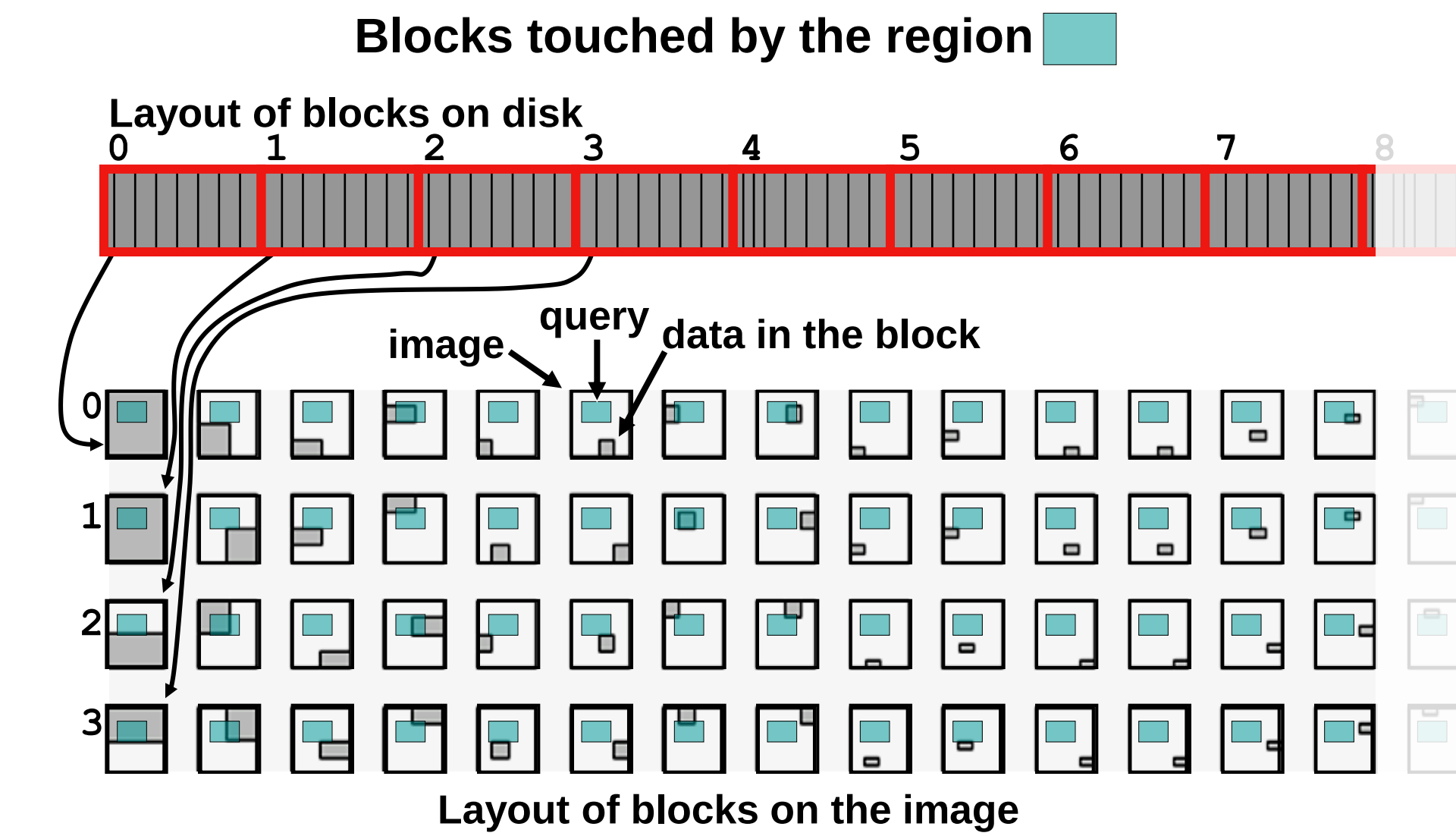
- Z-order curve used to define a hierarchical sub-sampling over a grid



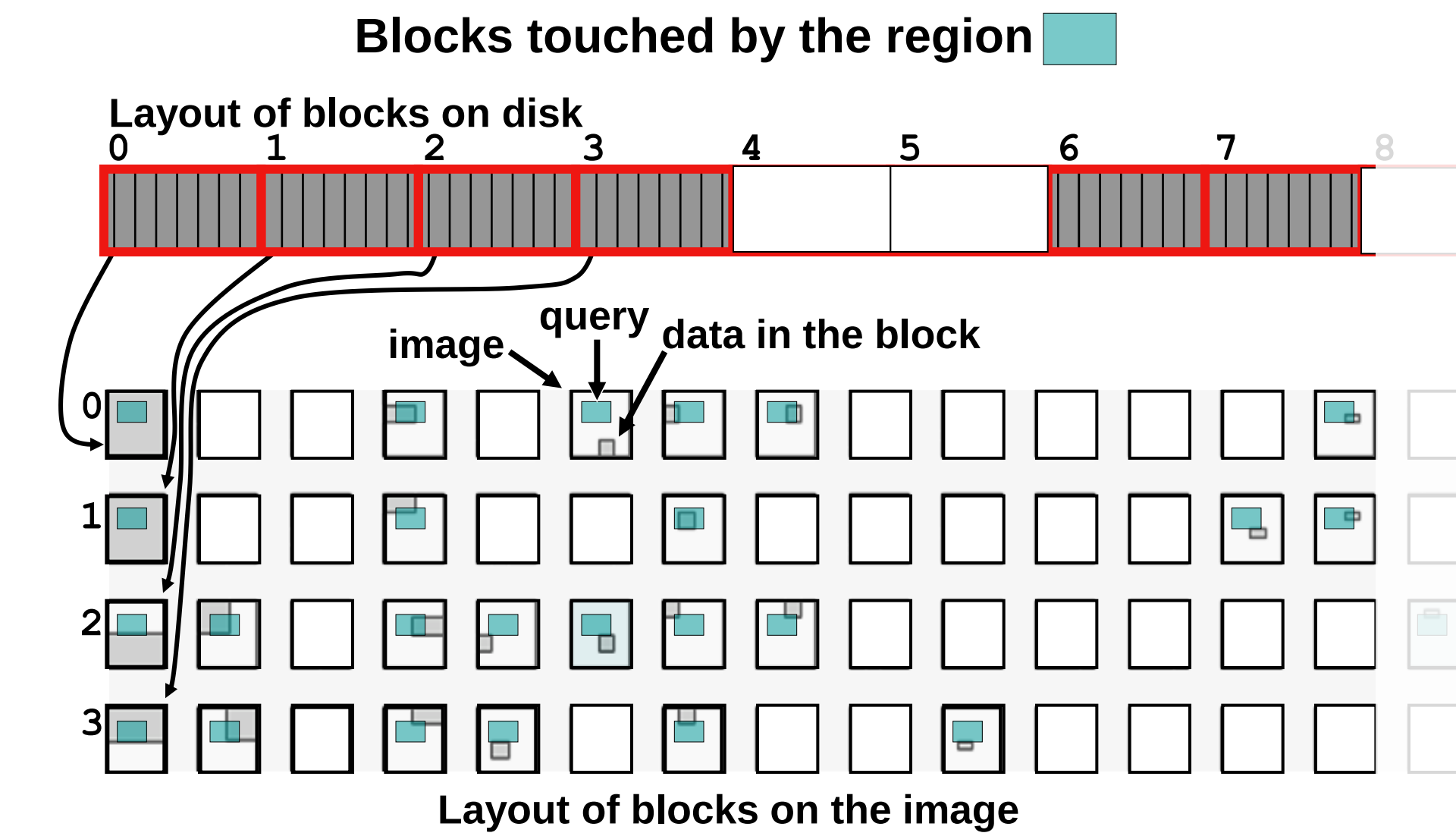
- Improve access locality:
  - **Interleaving hierarchical levels**
  - **Maintaining geometric proximity**
- Data layout is independent of the traversal of the data



We Introduced a Progressive Range Query Avoiding Unnecessary Data Access

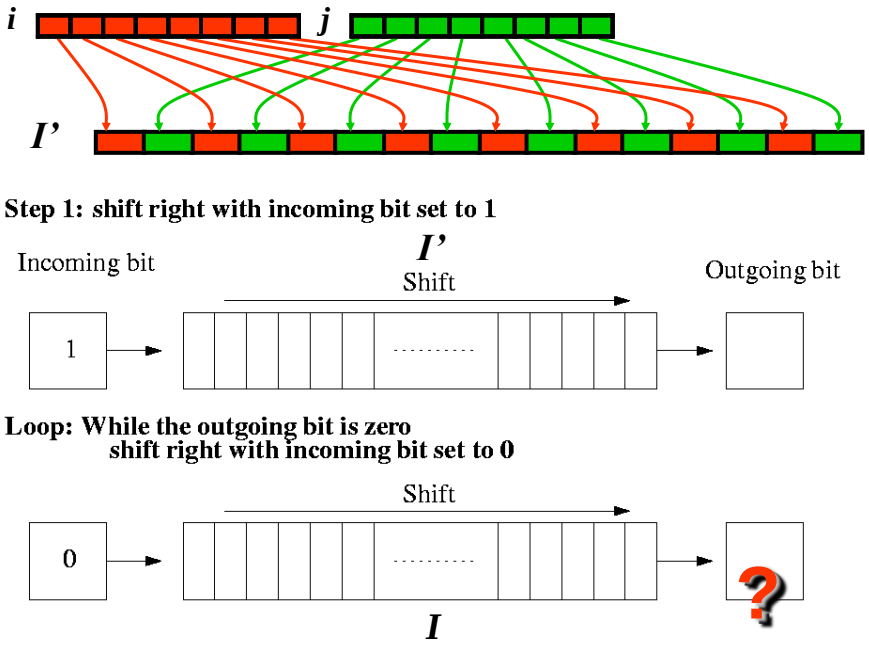


We Introduced a Progressive Range Query Avoiding Unnecessary Data Access

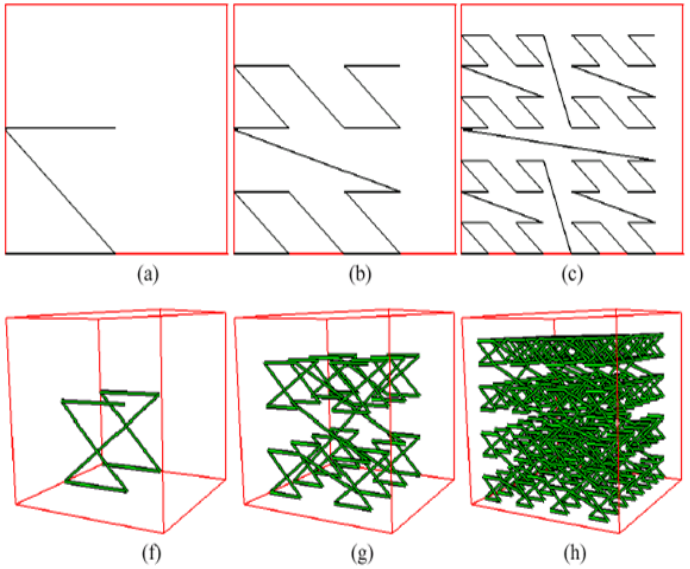


We Provided a Fast Address Computation Based on Simple Bit Manipulation

- **Simple bit manipulations to convert row major to hierarchical Z-order**



- **3D version (also nD): basic Z shape replaced by a connected pair of Z shapes**

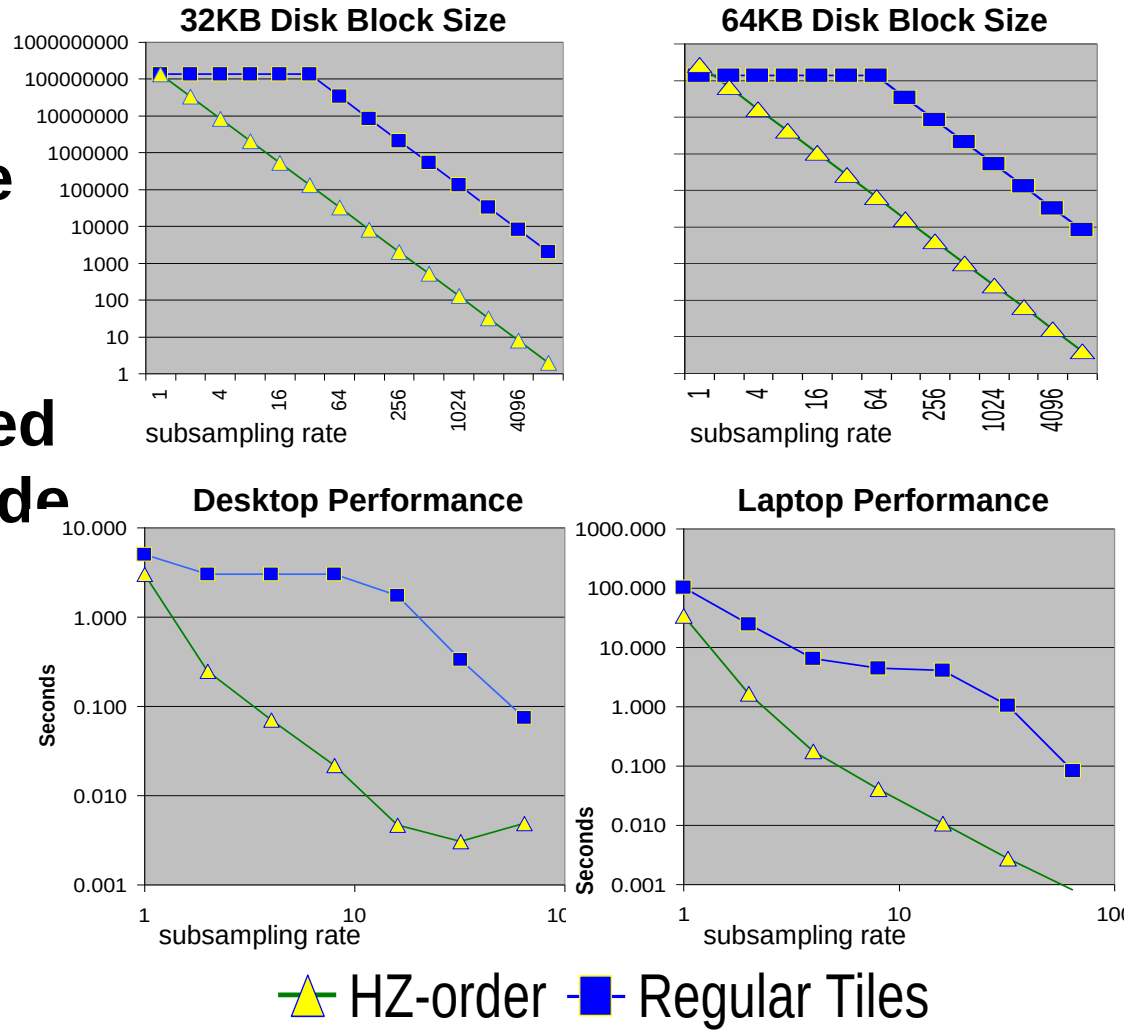


Cache-Oblivious Data Layouts Scale Well Across Different Storage Blocking Factors

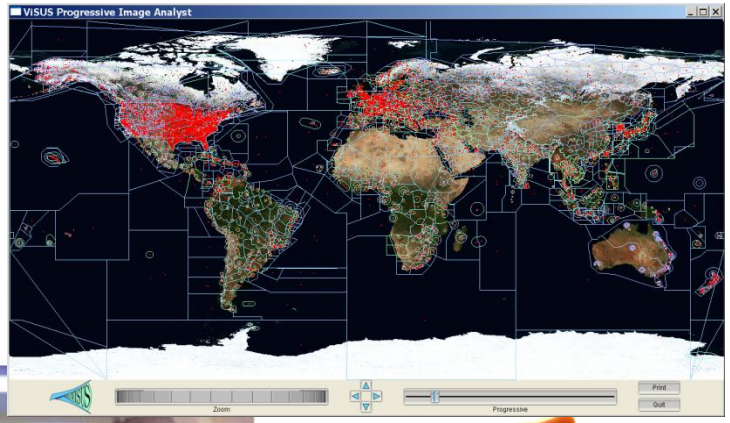
Formal analysis predicts performance and scalability

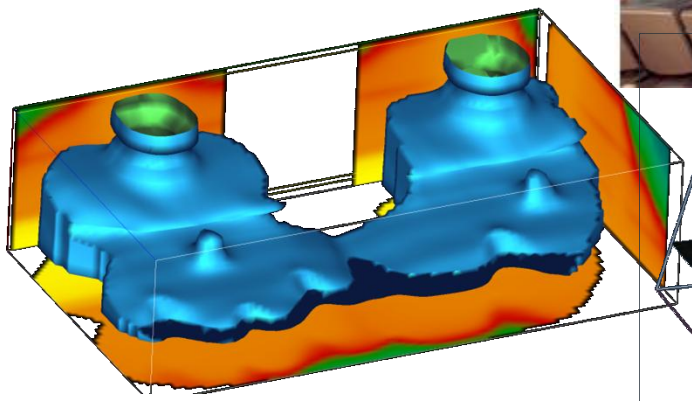
Performance improved by orders of magnitude

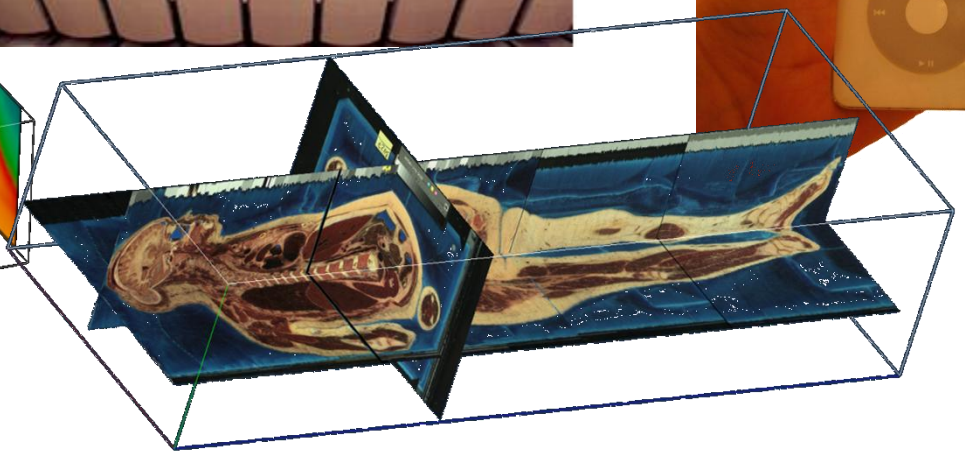
Independence of architecture and storage characteristics



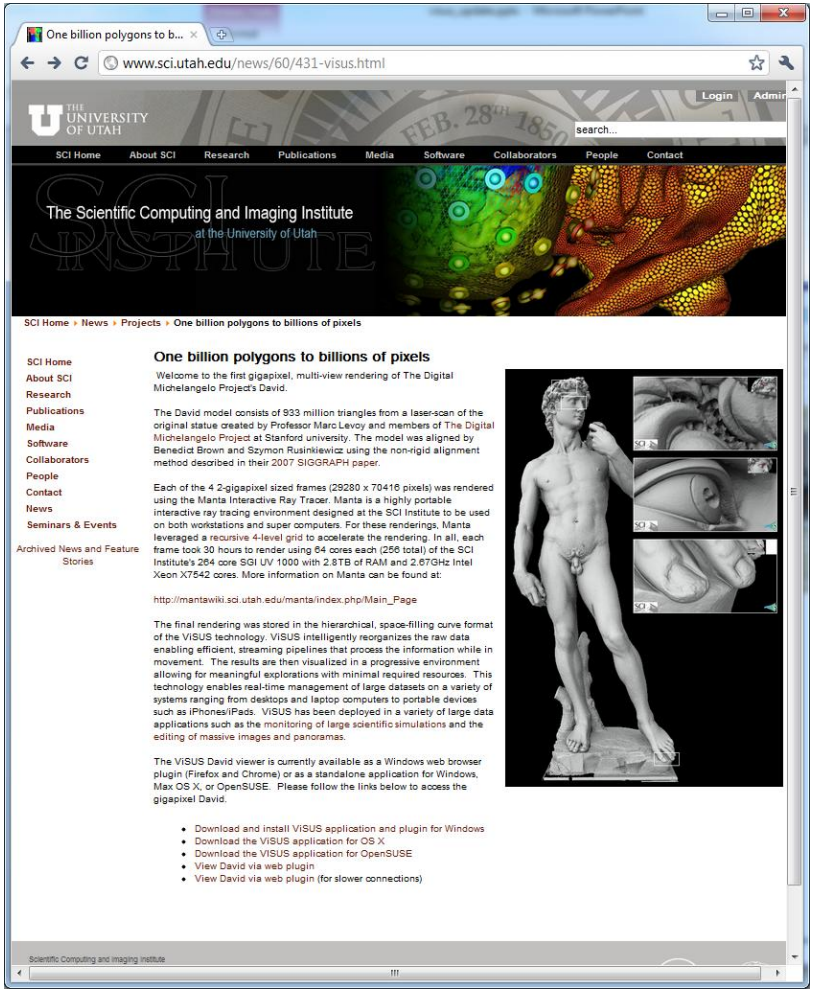
We Demonstrated Performance and Scalability in a Variety of Applications

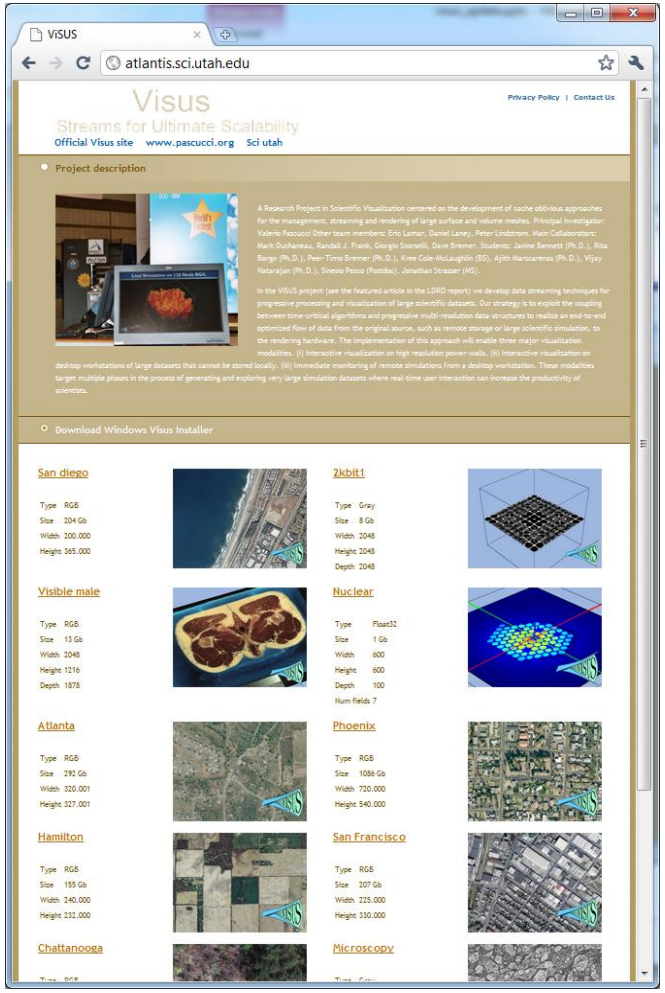






Server can be wrapped in Apache plug-in
Client can be run in a web browser





Geospatial Data Rendering on iPad

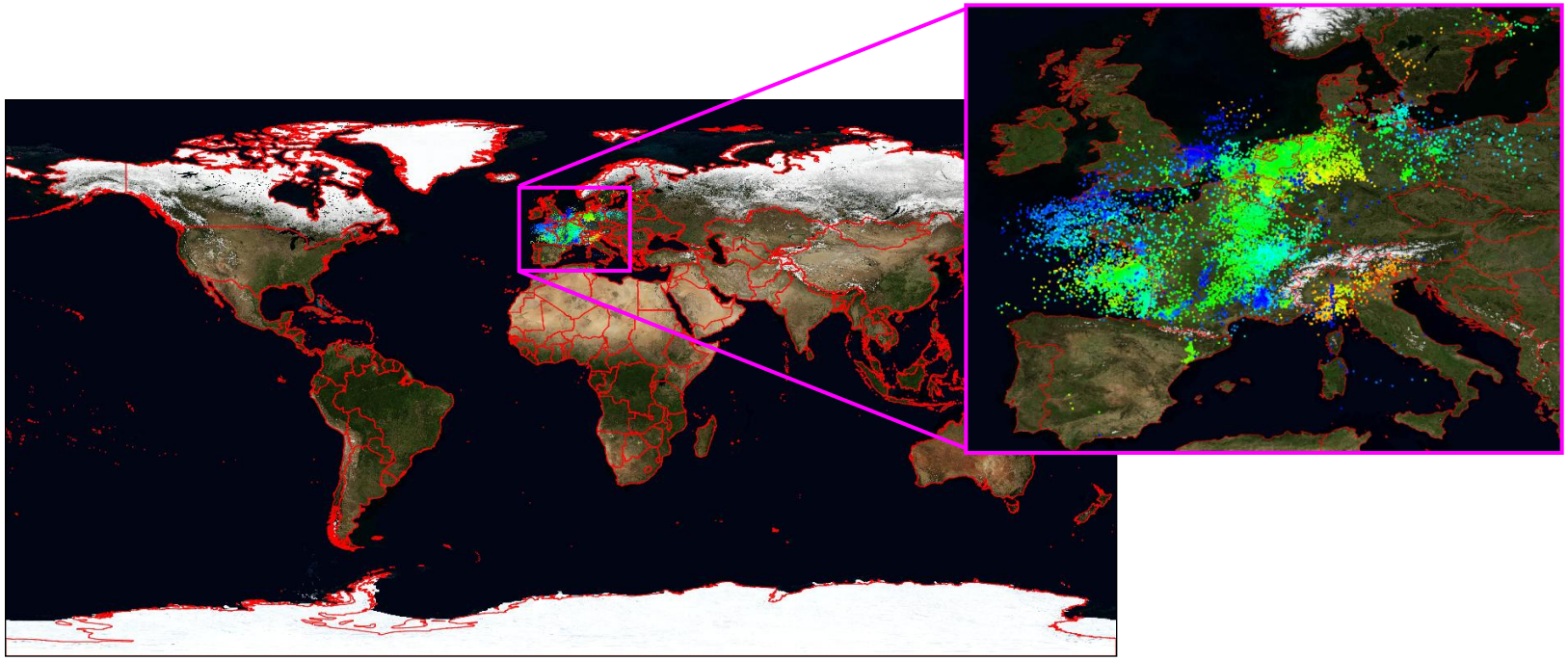
Both client and SERVER run of handheld devices, e.g. multiple iPhones can be clients and servers for each other to share information on the field

We Address the Need for Scalable Algorithms and Infrastructures

- Data formats that minimize I/O and memory transfer for most frequent operations
- New algorithms and data structures for management of large collections of time dependent information
- New theoretical models that predict the behavior of modern architectures
- New algorithms that are “intrinsically scalable” with respect to:
  - Processing capabilities
  - Diversity of hardware available
  - Locality of data
- Can benefit a variety of tools
- Scalable system infrastructures

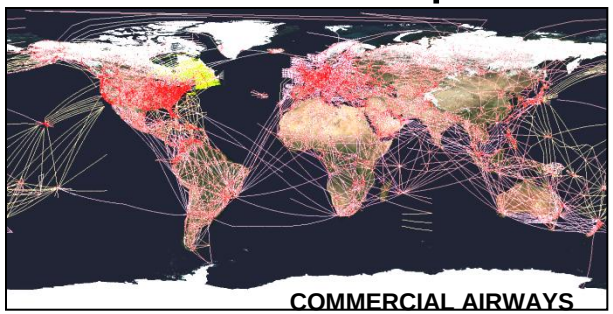
We provide real-time access to large scale time dependent data and sensory data

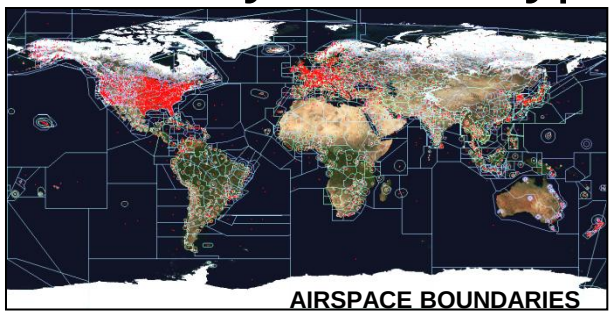
- **Blue Marble Earth (next generation) provided by NASA:**
  - **Twelve months in 2004 (11GB per month)**
  - **Resolution: 86400 x 43200 pixels (500 meters per pixel)**
- **Lightning events from distributed sensing devices over Central Europe (Blitzortung.org)**

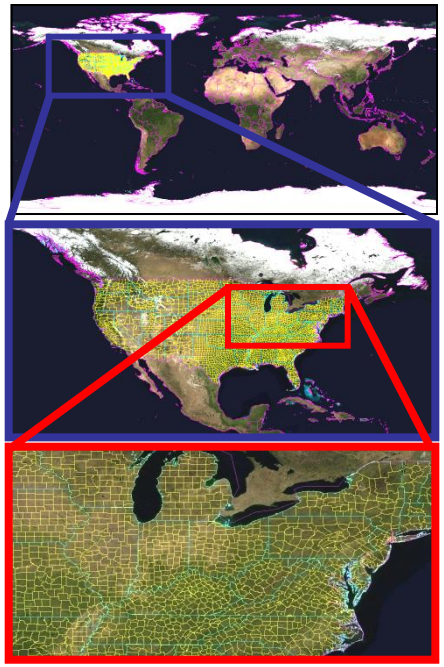


We Are Moving Towards Support of a Combination of Different Data Types

- Concurrent access to several information sources requires similar techniques for a variety of data types.


COMMERCIAL AIRWAYS

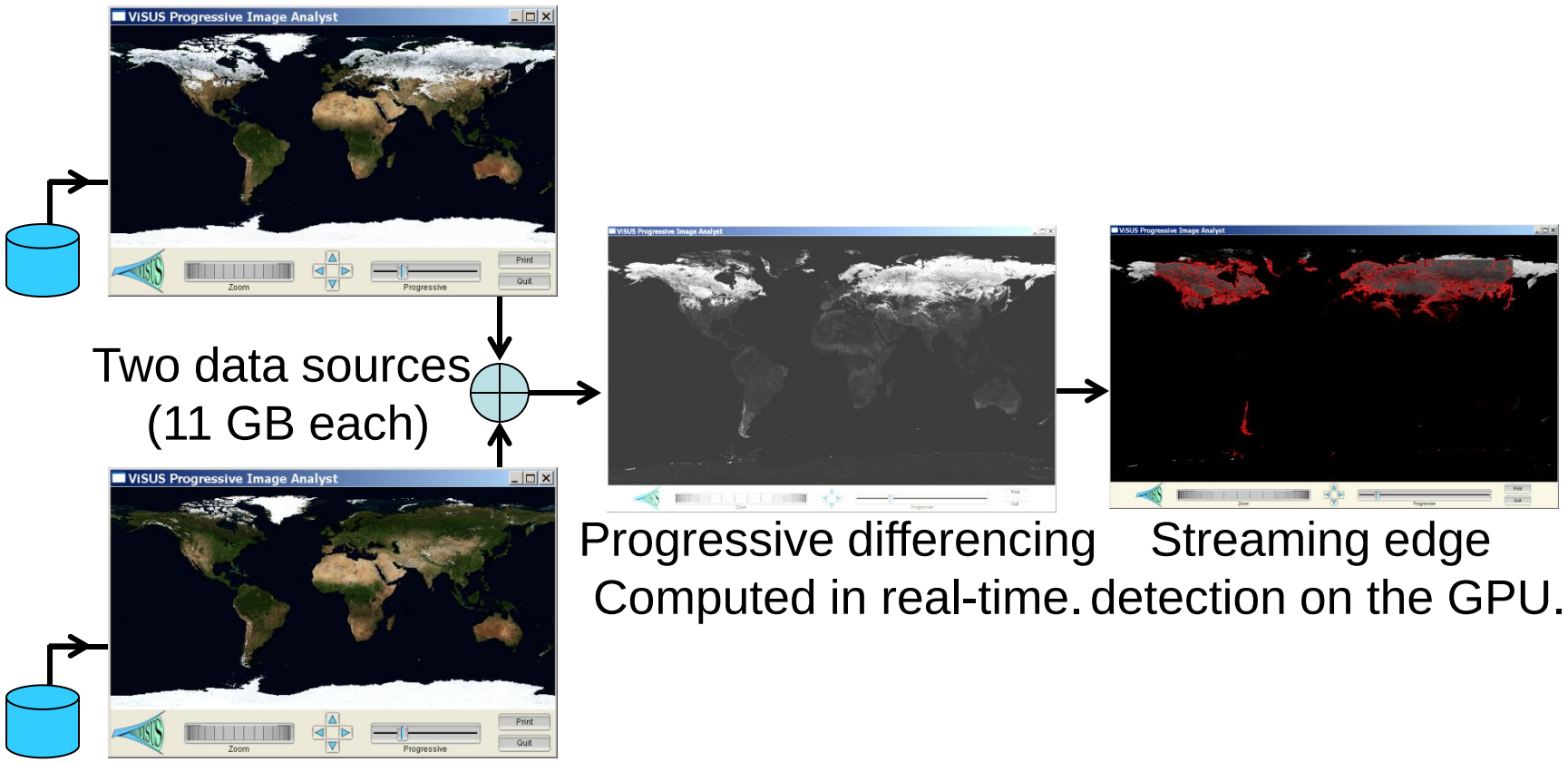

AIRSPACE BOUNDARIES


BOUNDARIES OF US COUNTIES

- **Algorithmic directions:**
  - **integration with other frameworks**
  - **progressive streaming infrastructure for "vector data" (points, lines, polygons);**
  - **uncertainty/incompleteness in the data**
  - **progressive resolution of queries with quantification and visualization of error/confidence;**
  - **on-line update of internal data structures to render new data immediately available.**

We Allow Distributed Computations at Different Stages of the Data Stream

- Progressive Image Differencing + Editable GPU filter.

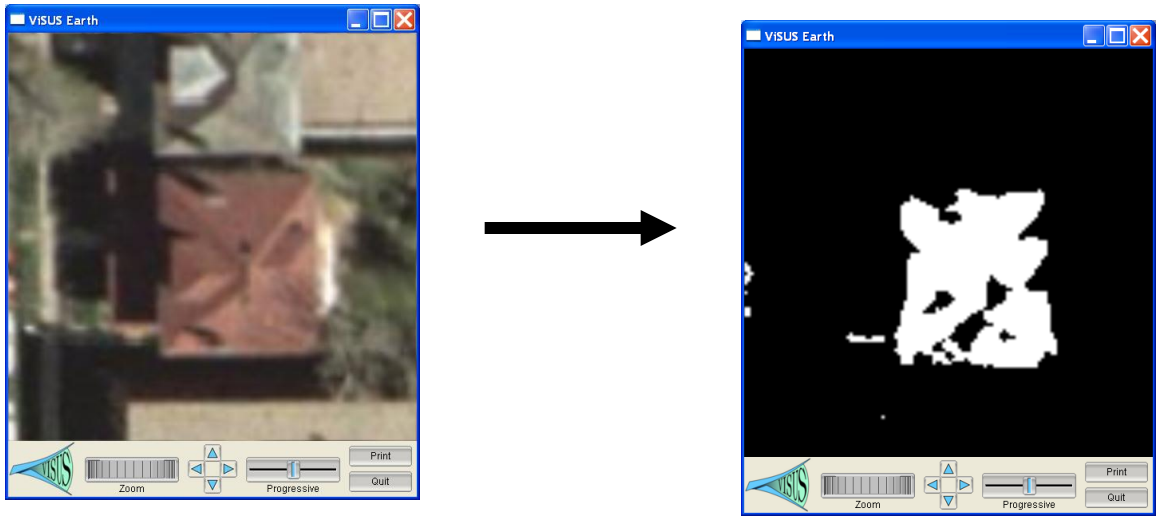


Progressive differencing Computed in real-time.

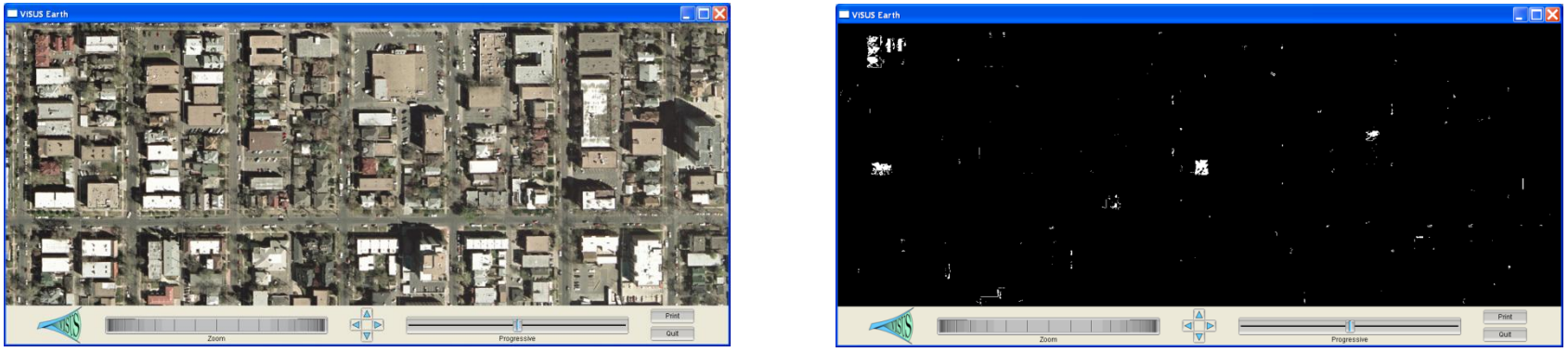
Streaming edge detection on the GPU.

We are Developing Progressive Scheme for Content Based Image Processing

- **Hypothesis:**



- **Progressive Analysis:**

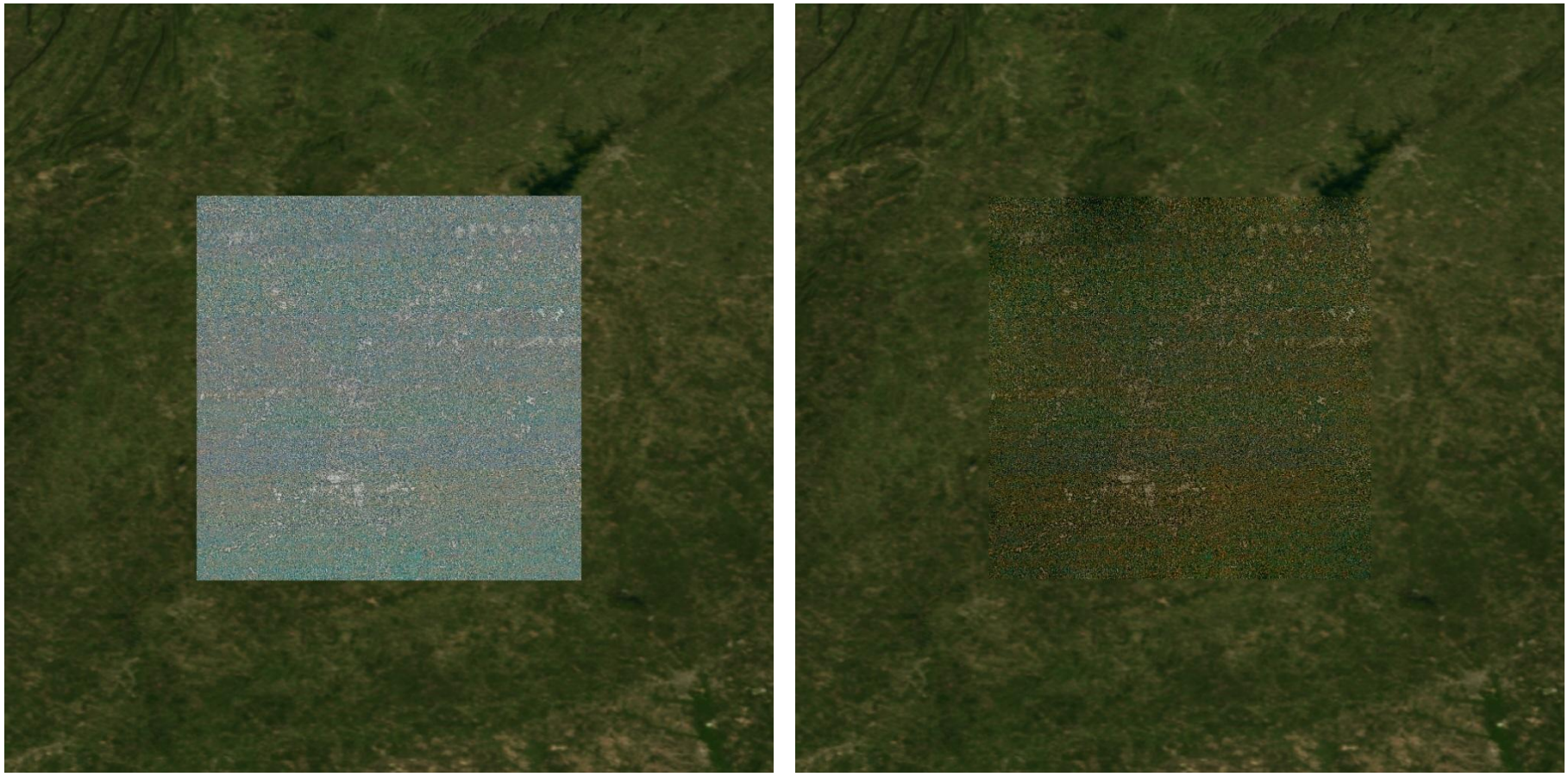


Poisson Solver for Image Cloning in Massive Image Collections

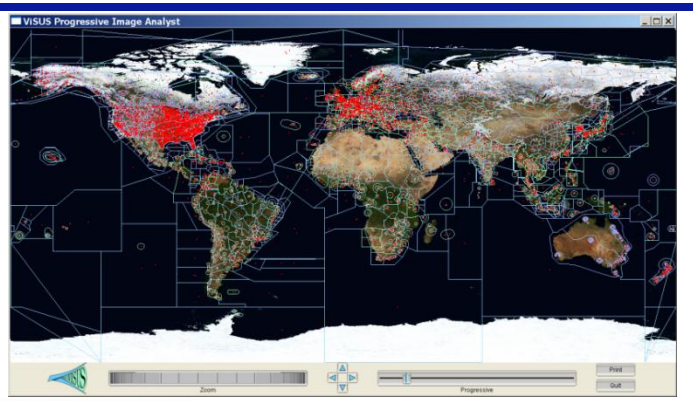
- **Color correction of 600+ images in real time**

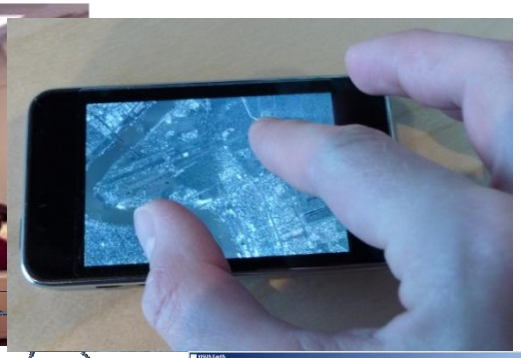
Poisson Solver for Image Cloning in Massive Image Collections

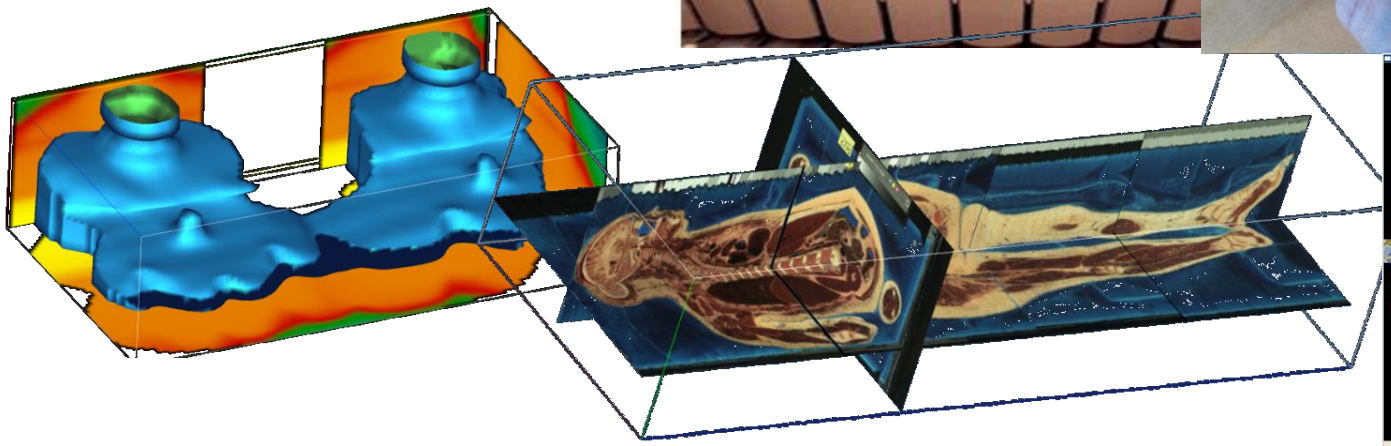
- **Pasting a 300GB satellite image of a city in background world map merged in real time**

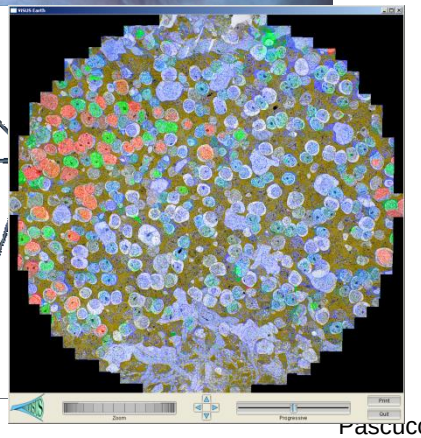


We Demonstrated Performance and Scalability in a Variety of Applications

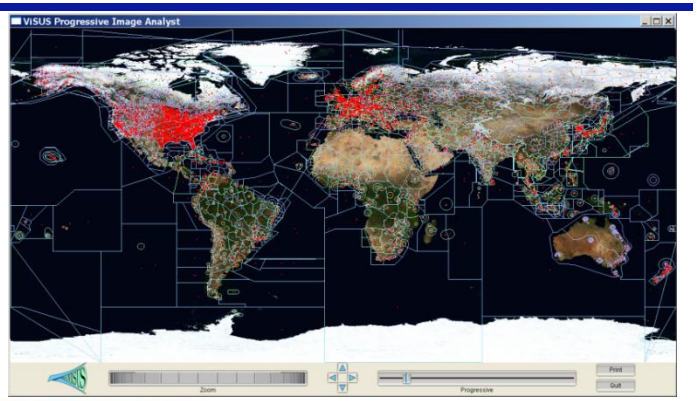


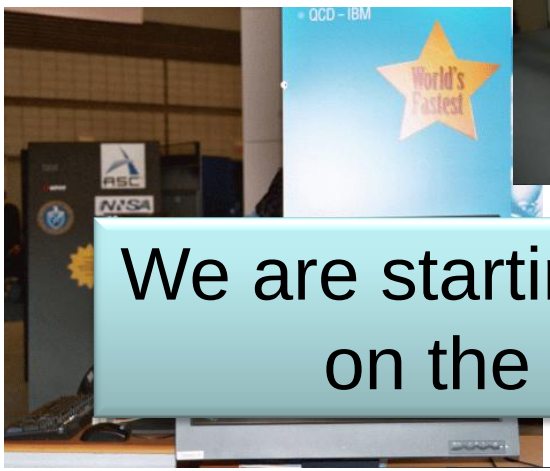




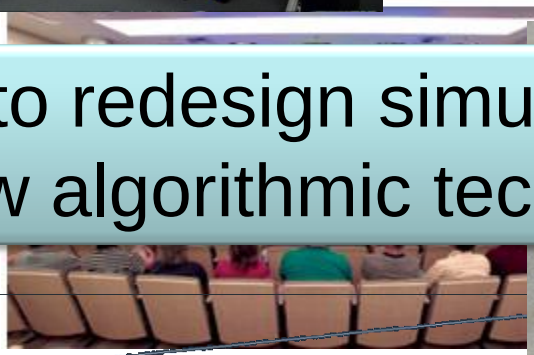


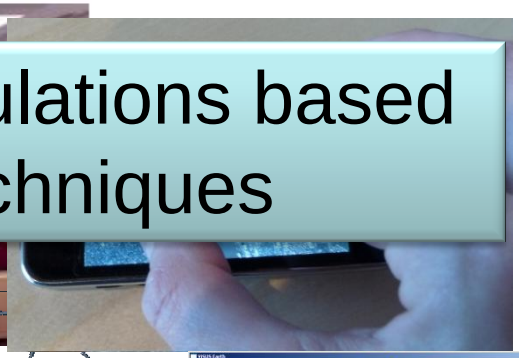
We Demonstrated Performance and Scalability in a Variety of Applications

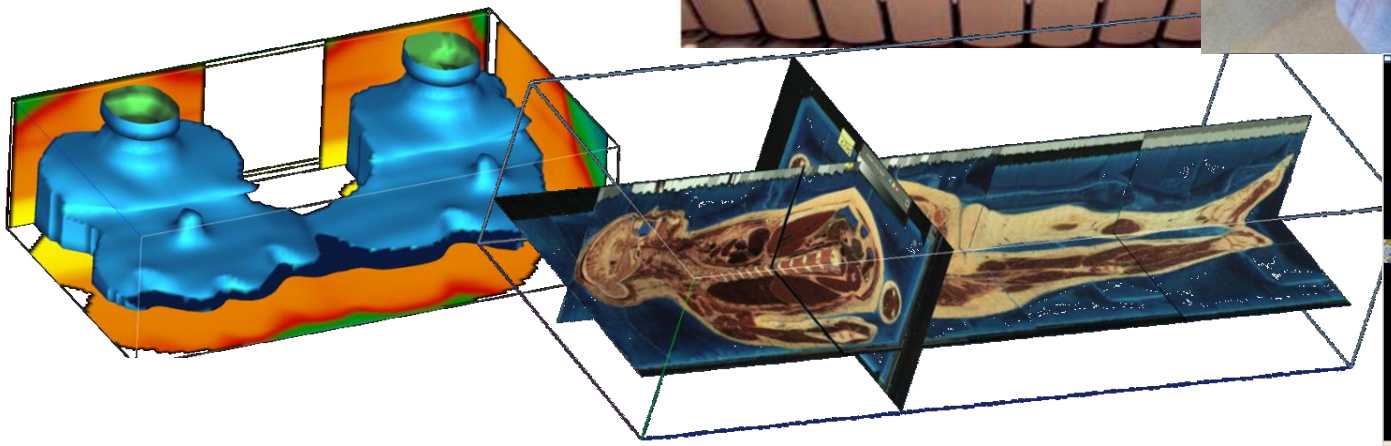


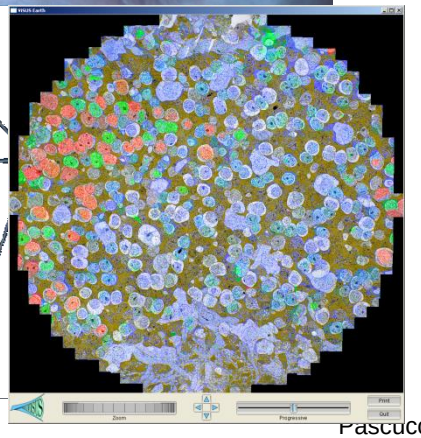


We are starting to redesign simulations based on the new algorithmic techniques



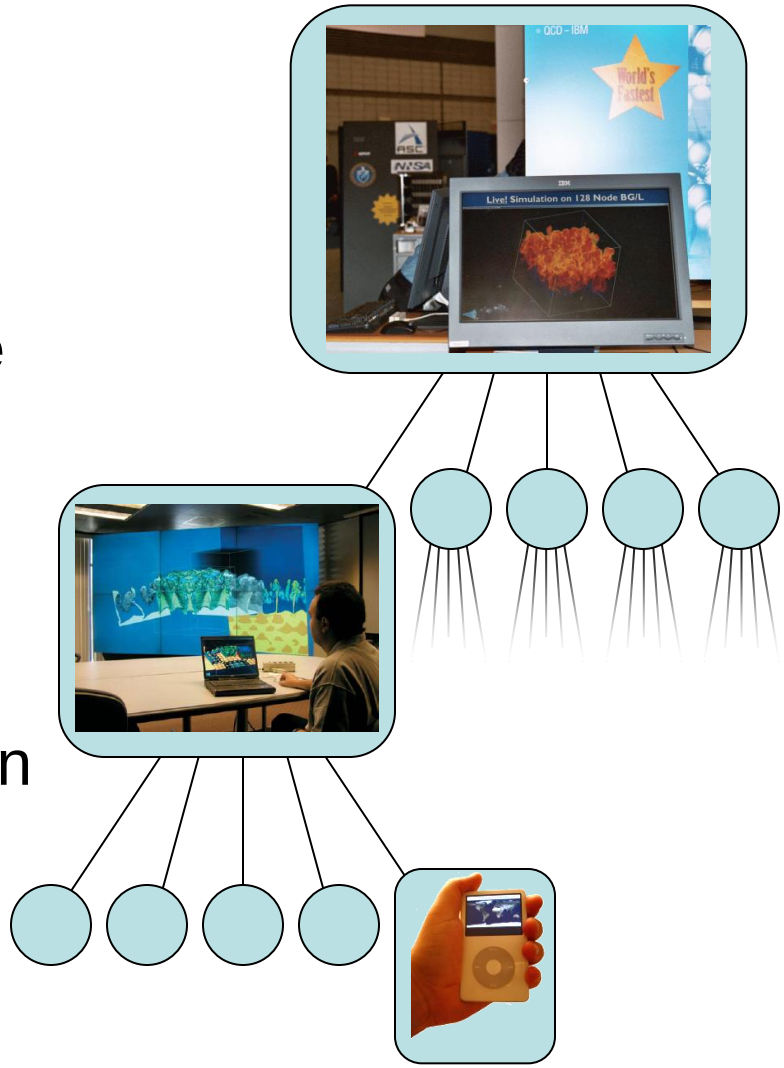




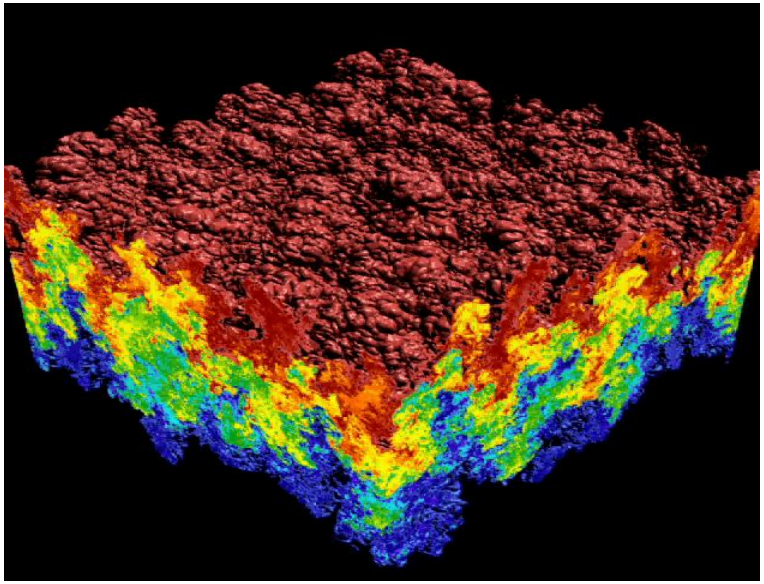


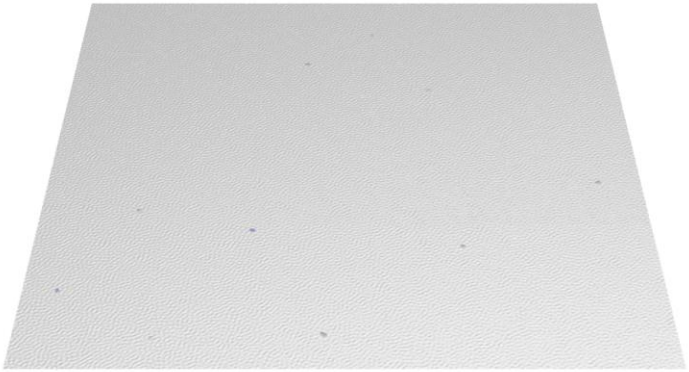
We Are Moving Towards a Distributed Storage and Processing Environment

- **Distributed storage**
- **Data redundancy**
- **Security**
- **Heterogeneous collaborative infrastructure**
- **Multi-scale collaborative interfaces accessing shared data sources:**
  - data collection and validation
  - interactive analytics
  - decision making



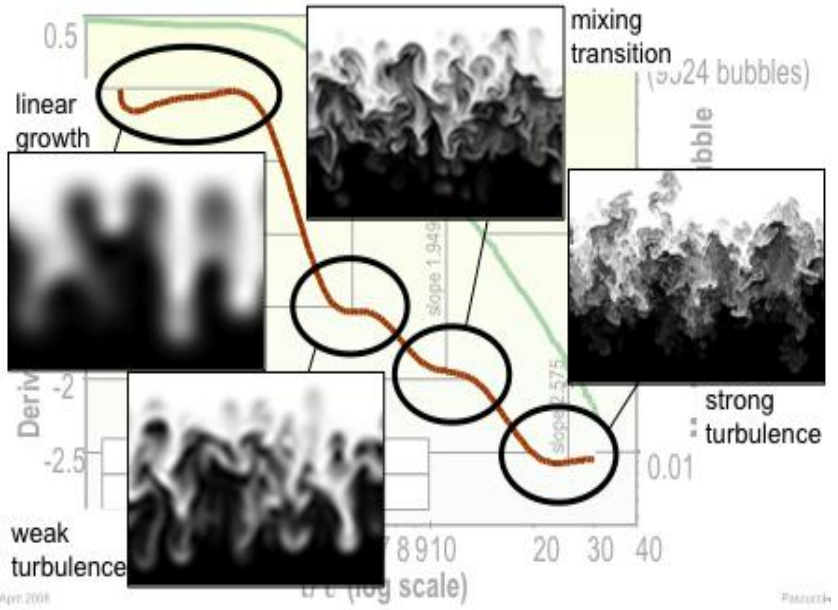
We Focus on Quantitative Analysis for Answering Science Questions



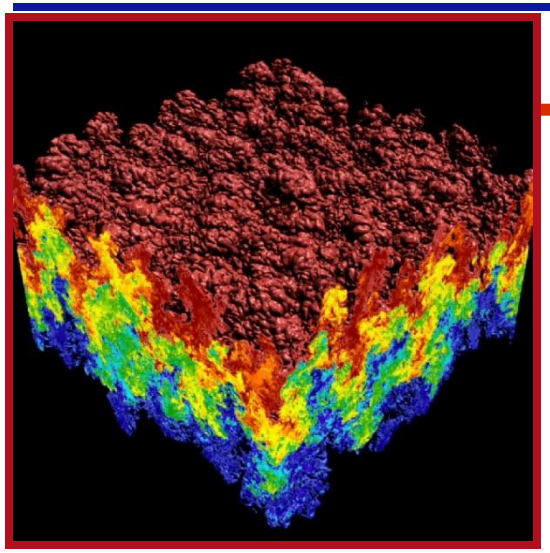


Rayleigh–Taylor Instability (fusion, super-novae, …).

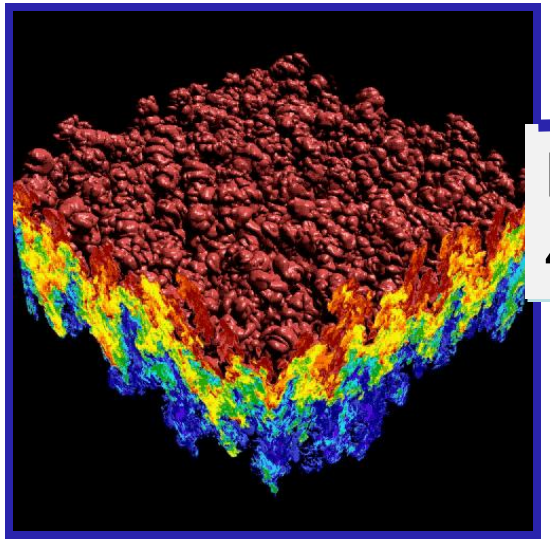
- What are the stages of the turbulent mixing process?
- Over 40 TB



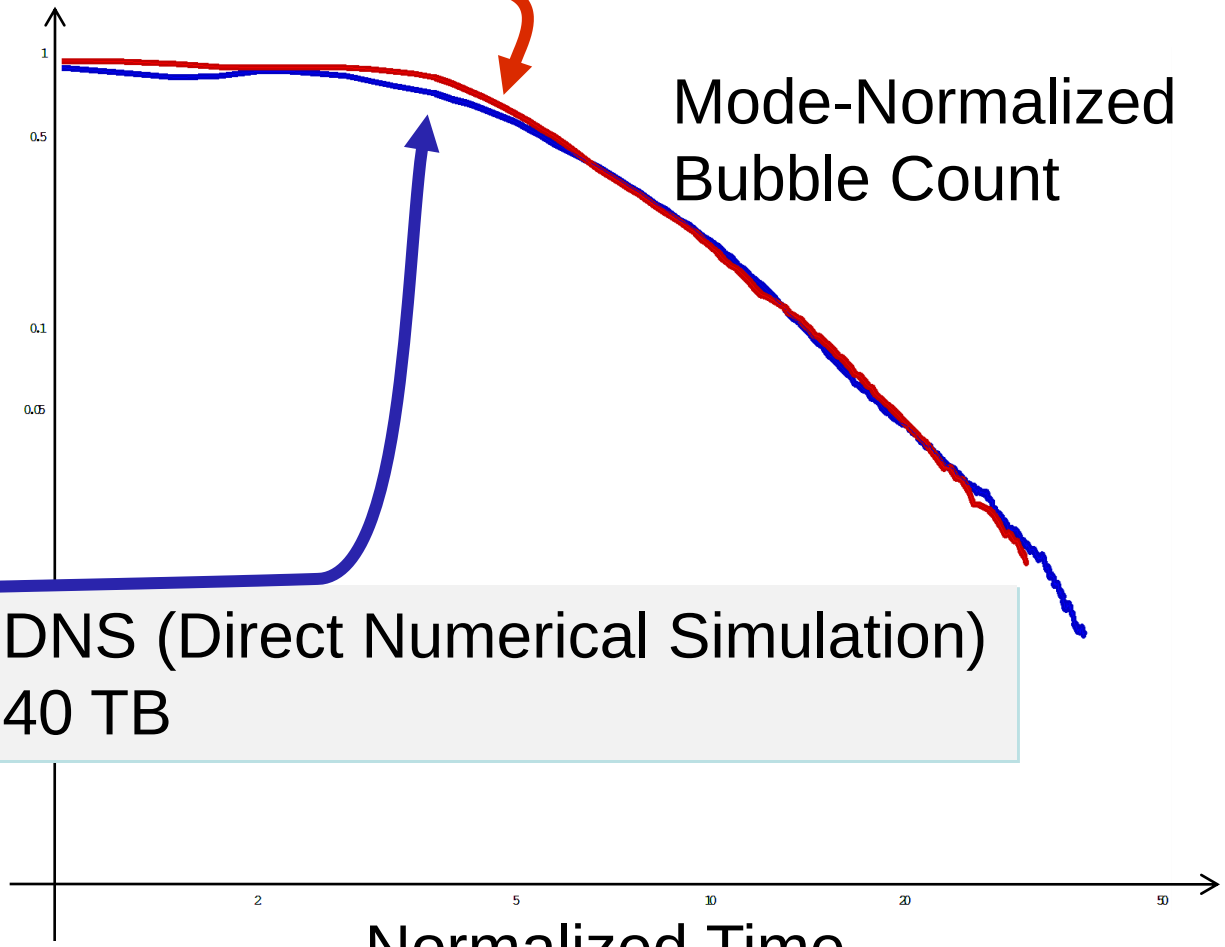
We Provided the First Feature-Based Validation of a LES with Respect to a DNS



LES (Large Eddy Simulation) 25TB

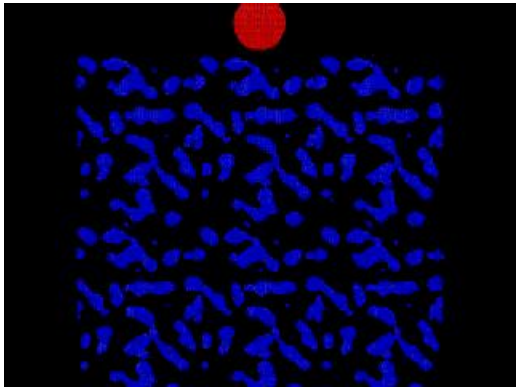


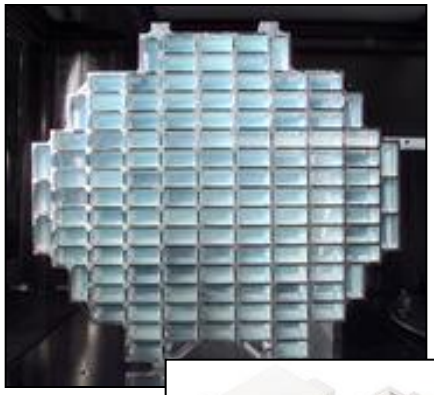
DNS (Direct Numerical Simulation) 40 TB

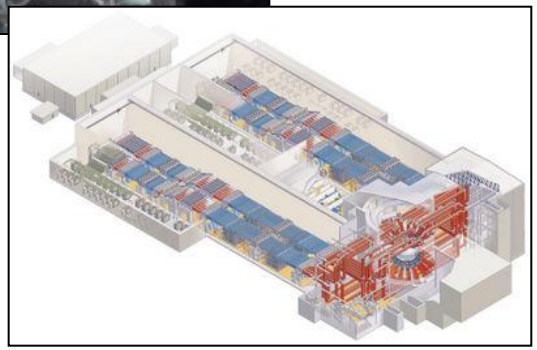


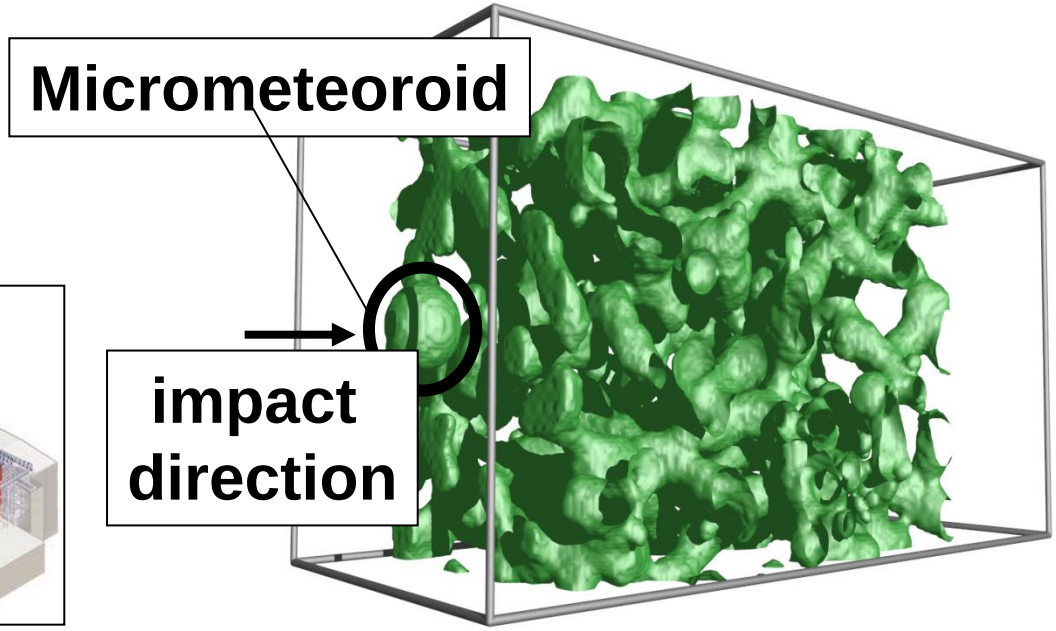
Quantitative Analysis of the Impact of a Micrometeoroid in a Porous Medium

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

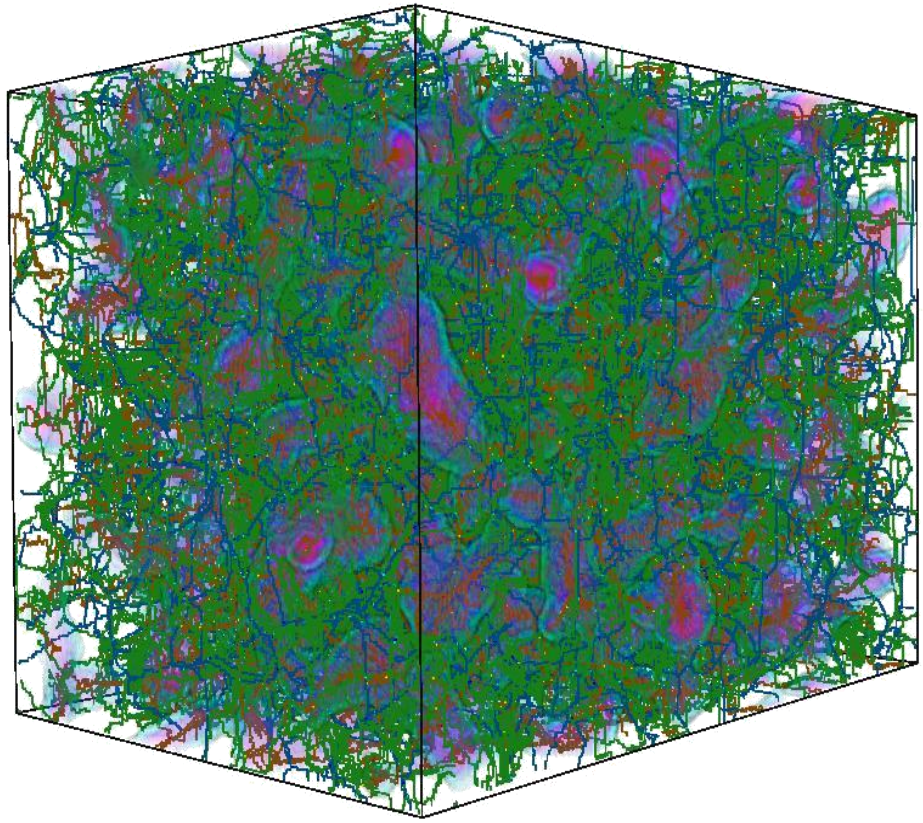


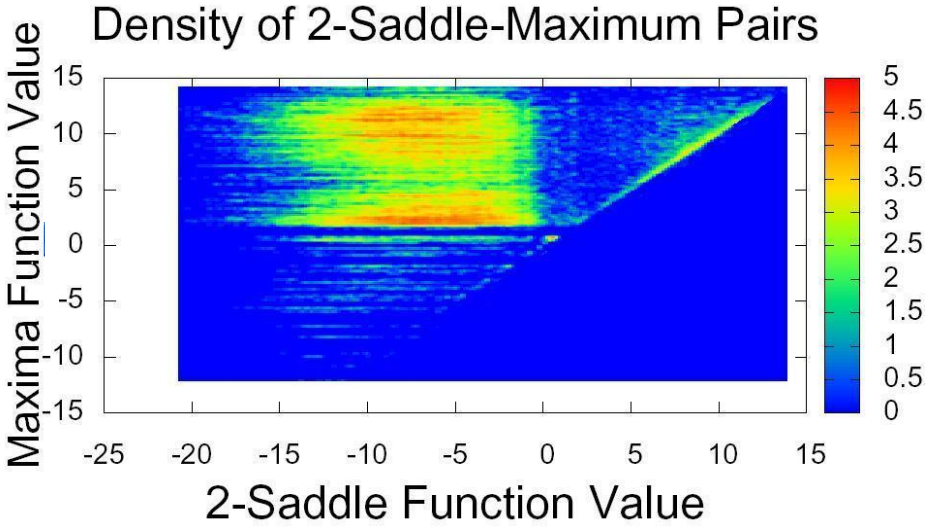




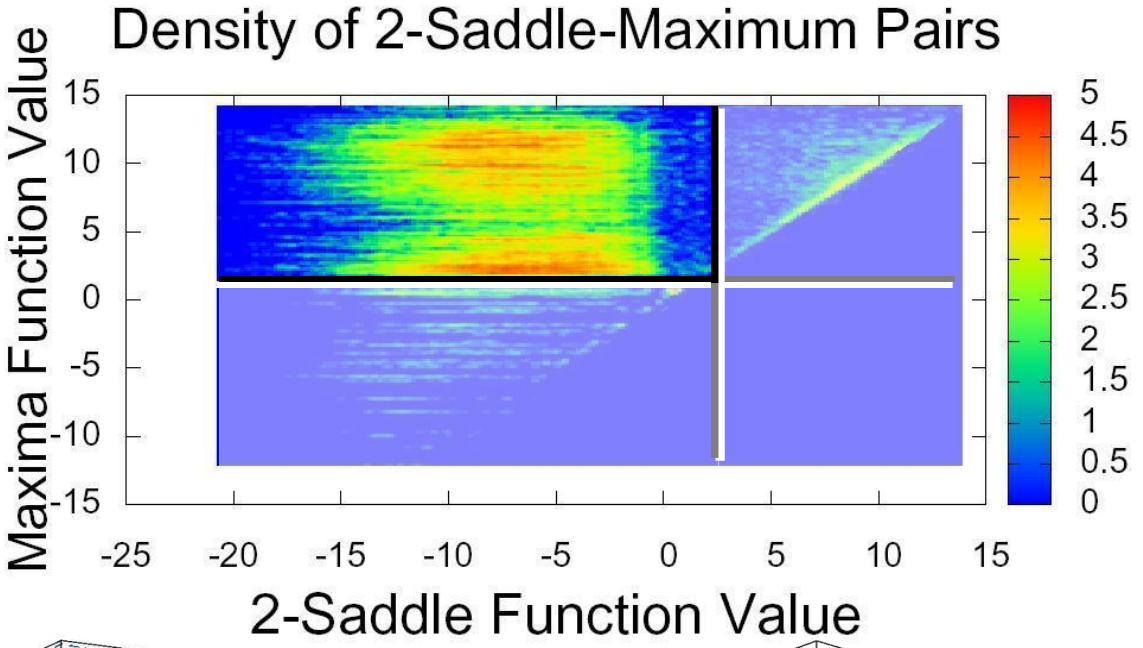


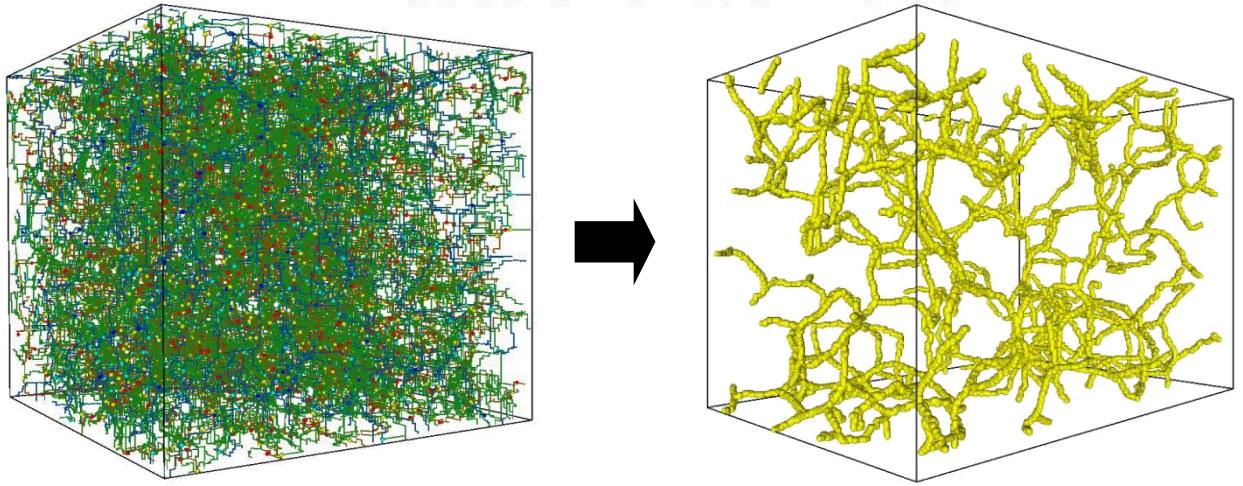
We Use Topological Methods to Describe all Possible Reconstruct. of the Porous Medium



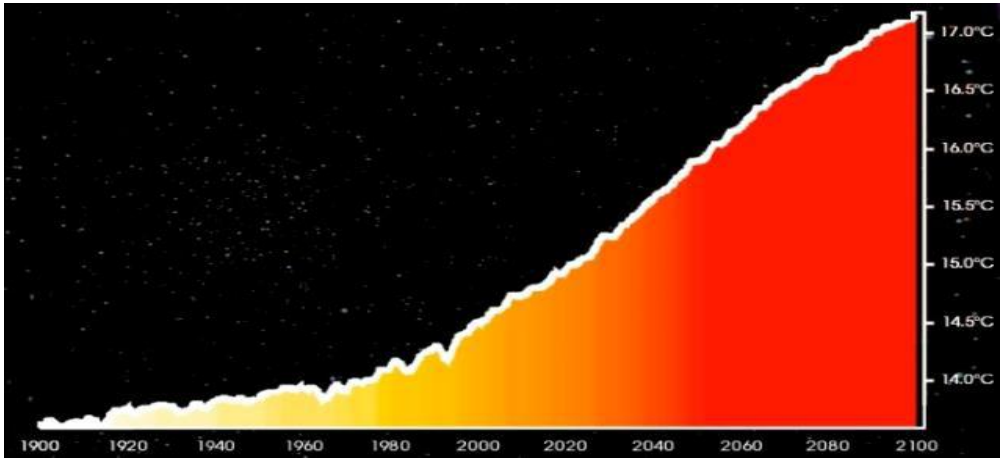


We Obtain a Robust Reconstruction of the Filament Structures in the Material

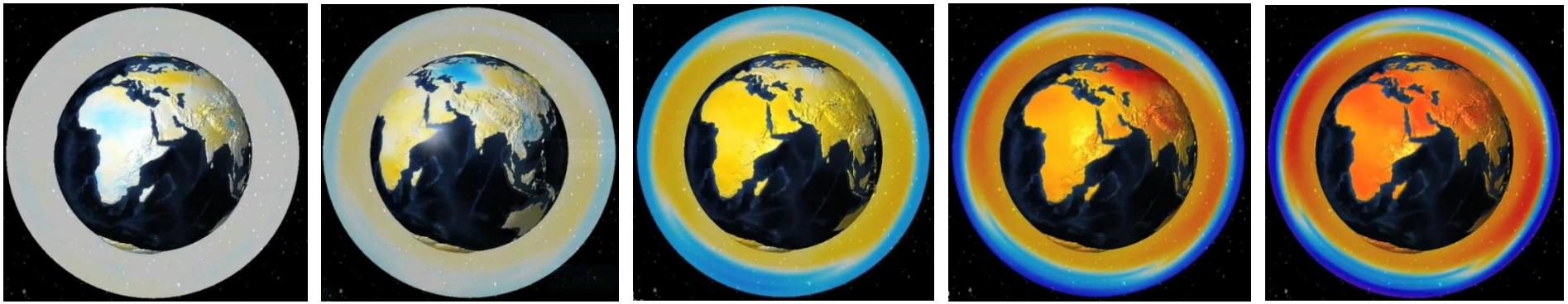




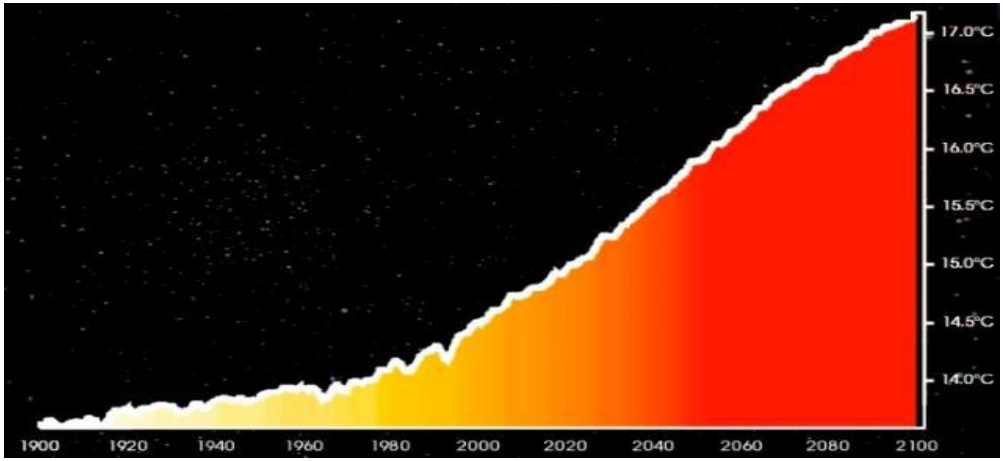
What is the Long Term Impact of Human Activities on the Global Climate?



What is the Long Term Impact of Human Activities on the Global Climate?

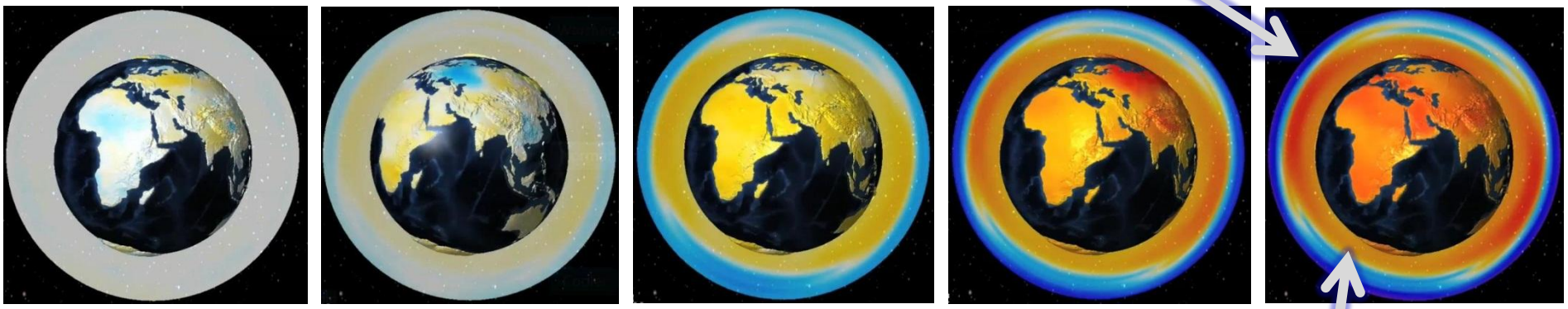


1900 1950 2000 2050 2100



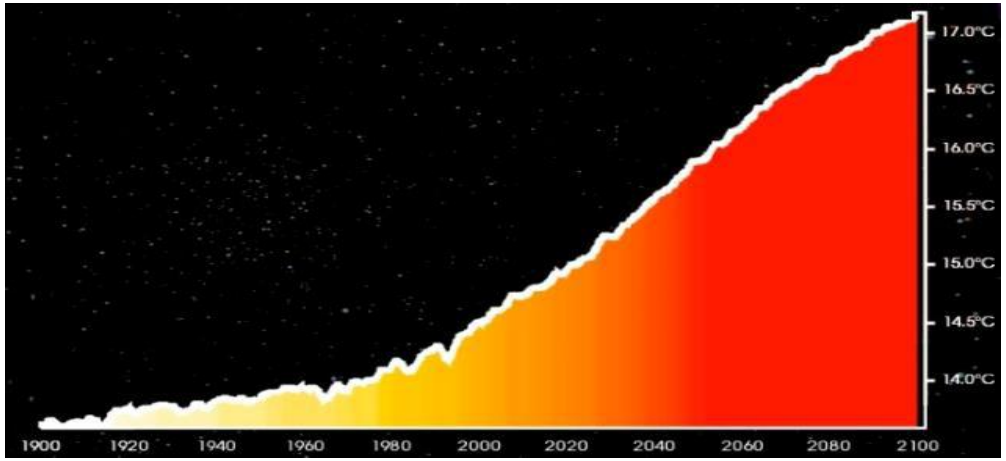
What is the Long Term Impact of Human Activities on the Global Climate?

cooling



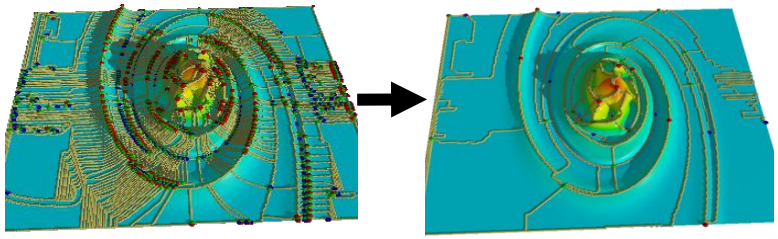
1900 1950 2000 2050 2100

warming

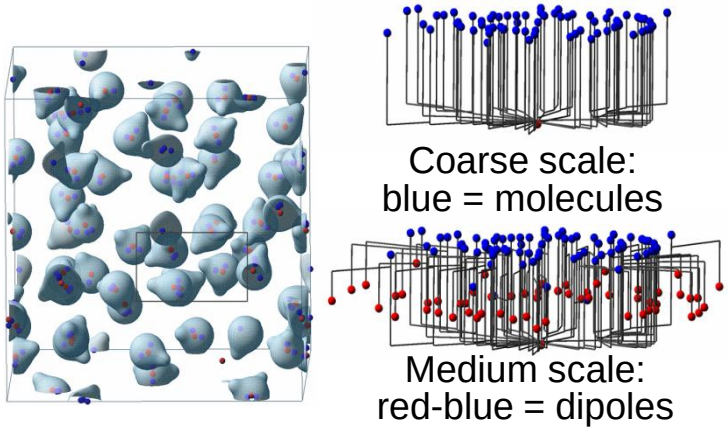


We Introduced Robust Topological Methods for Quantitative Data Analysis

- **Provably robust computation**
- **Provably complete feature extraction and quantification**
- **Hierarchical topological structures used to capture multiple scales**
- **Error-bounded approximations associated with each scale**
- **Formal mathematical definition associated with each analysis**
- **Scalable performance in association with streaming techniques**



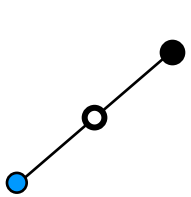
Hierarchical topology of a 2D Miranda vorticity field



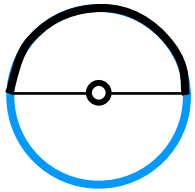
Molecular dynamics simulation (left) with abstract graph representation of its features at two scales (right)

We Rewrote Morse Theory for Provably Robust and Correct Computations

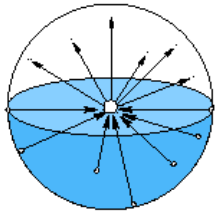
	$f(x): D \to \Re$	$F(x): S \to \Re$
	Classical mathematical definitions	Simulation of differentiability
domain	D smooth manifold	S simplicial complex
function	f infinitely differentiable	$F(x)$ PL-extension of $f(x_i)$
critical point	$\nabla f(p) = 0$ **numerical**	$LowerLink(p) \neq B^{d-1}$ **combinatorial**



1D



2D



3D

Independent local computation yield globally consistent results

We Introduced New Techniques for Critical Point Classification

type	index
Minimum	0
⋮	⋮
Saddle	d-1
Maximum	*d*

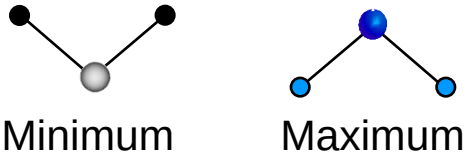
$$\nabla f(p) = 0$$

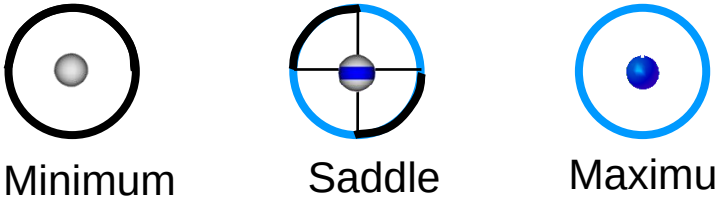
$$f(x)|_p = f(p) + \sum_{i=1}^{d-k} x_i^2 - \sum_{j=d-k+1}^{d} x_j^2$$

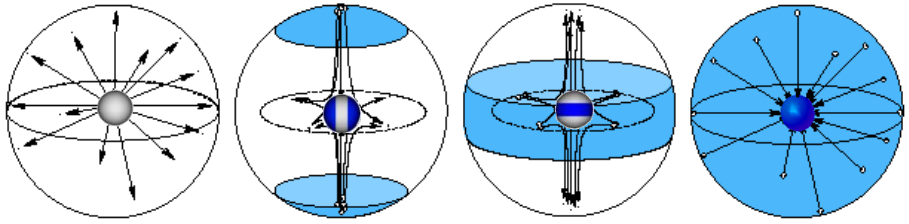
numerical

The Morse Lemma

There are *d+1* types of critical points



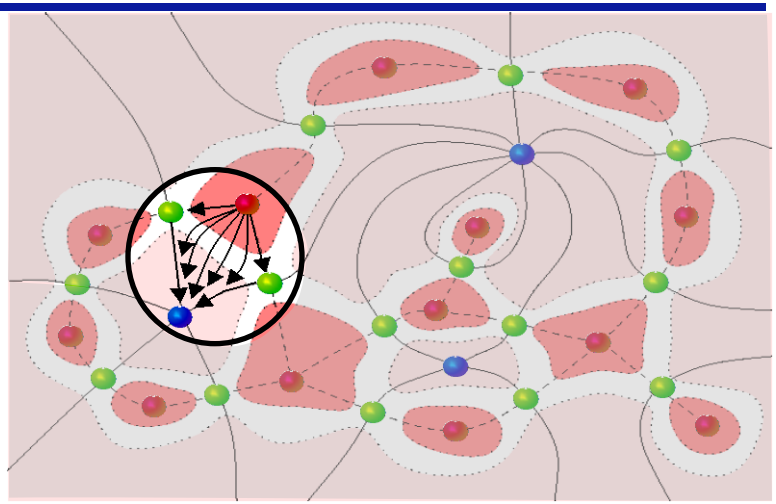


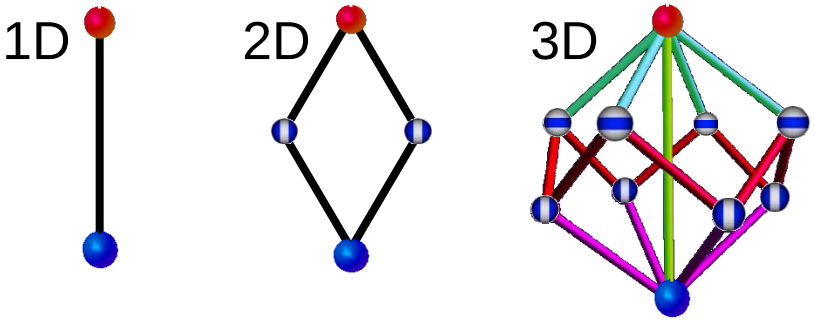


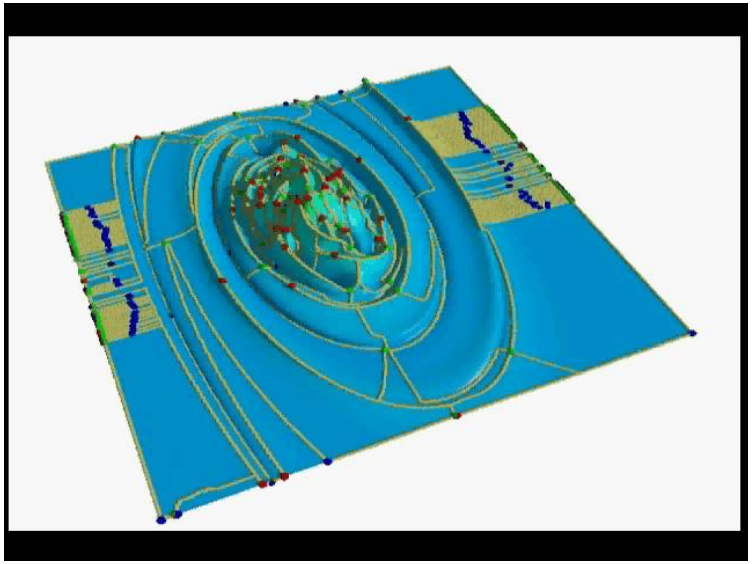
combinatorial

We Introduced the Morse–Smale Complex for Complete Data 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**





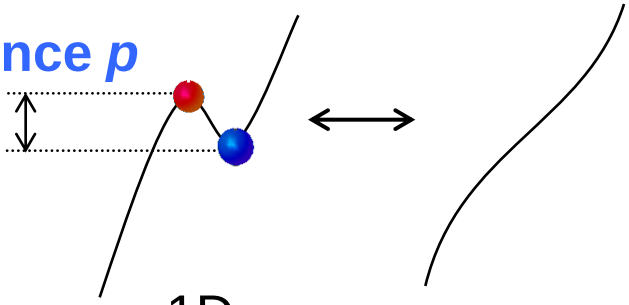


We Mapped the Index Lemma to a Morse–Smale Complex Simplification

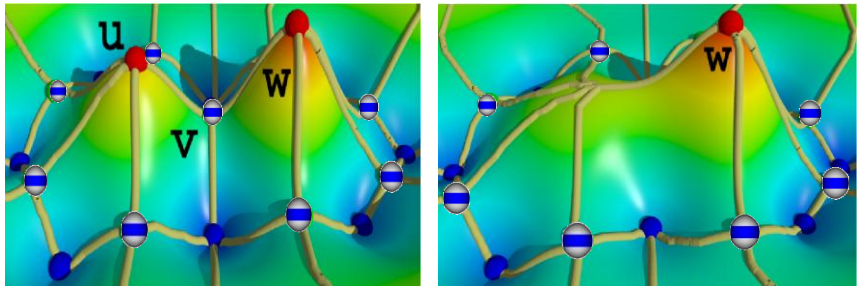
Index Lemma: critical points can be created or destroyed in pairs that differ by one in index

Approximation: error = persistence/2 (proven lower bound)

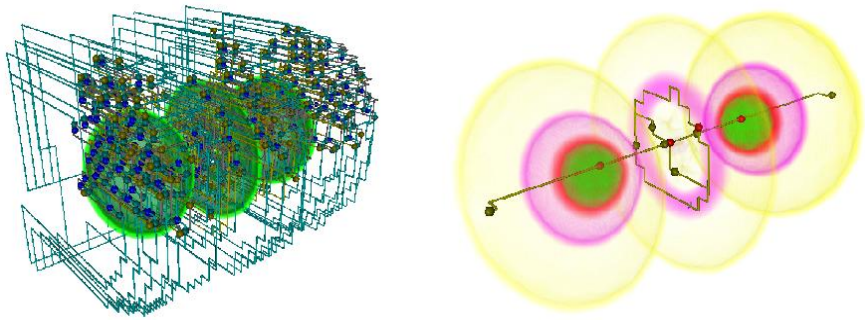
Multiscale: consistent gradient segmentation at all scales



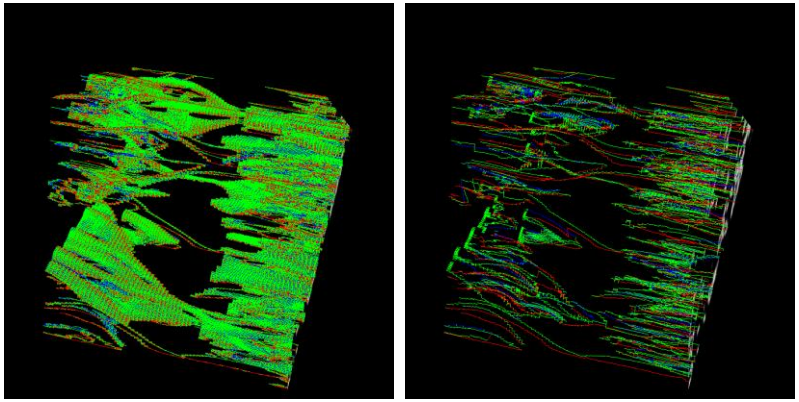
1D



2D



3D

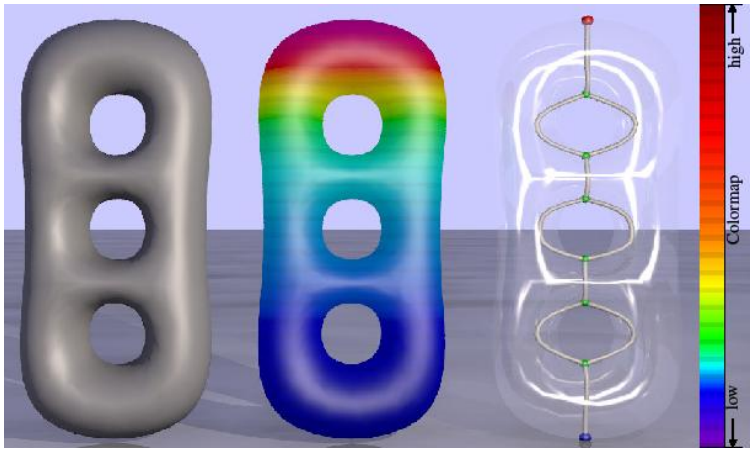


Time

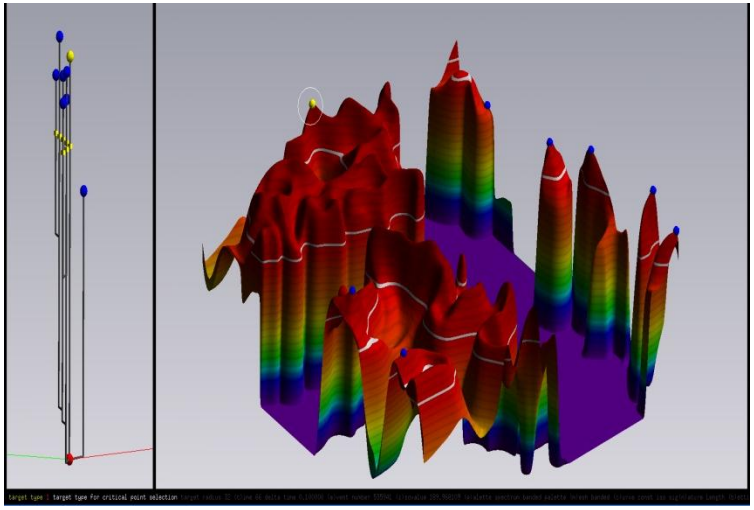
We Achieve a Theory of Provably-Correct Topological Computations

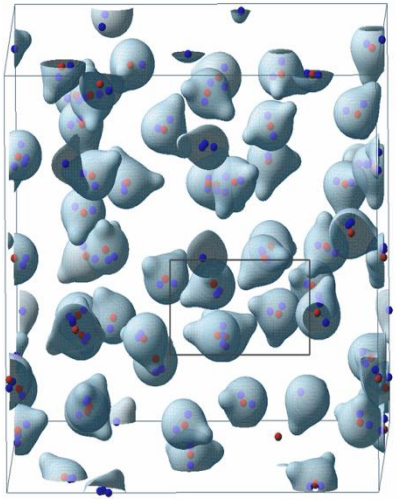
- No approximation introduced when translating mathematical definitions to algorithms
- The quality of the analysis does not deteriorate when the data size increases
- New topological concepts that allow complete data segmentations
- Multi-scale representation of the input data
- Explicit error bounds for any approximation

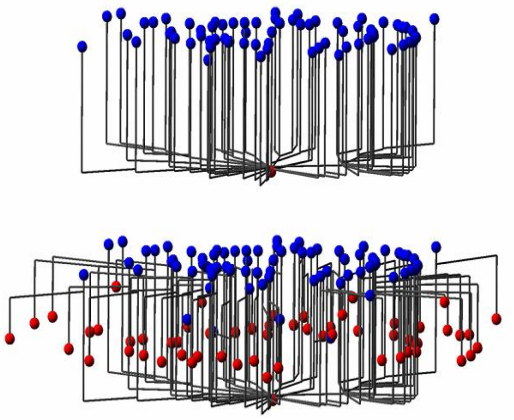
Topological Constructs Allow Building Effective and Succinct Shape Descriptors



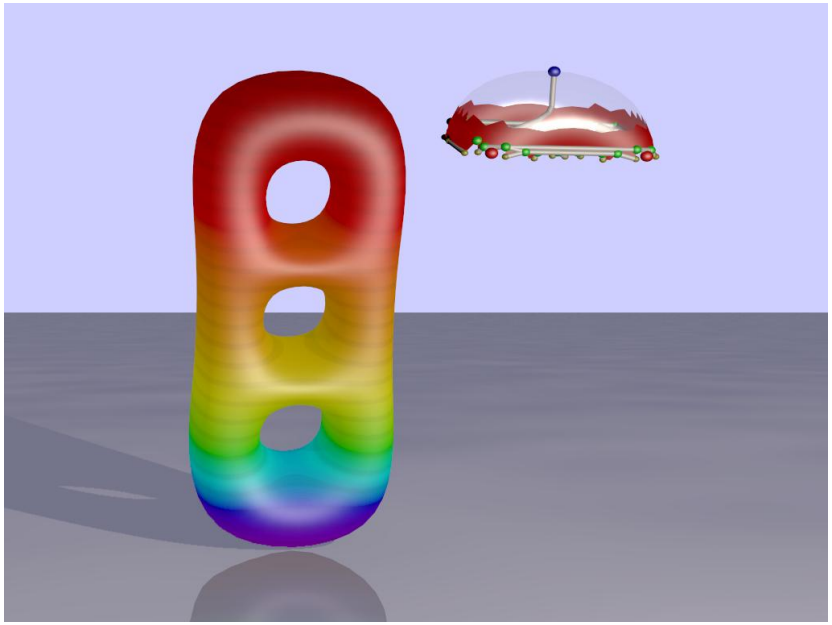
The Reeb graph is the graph obtained by continuous contraction of all the contours in a scalar field, where each contour is collapsed to a distinct point.

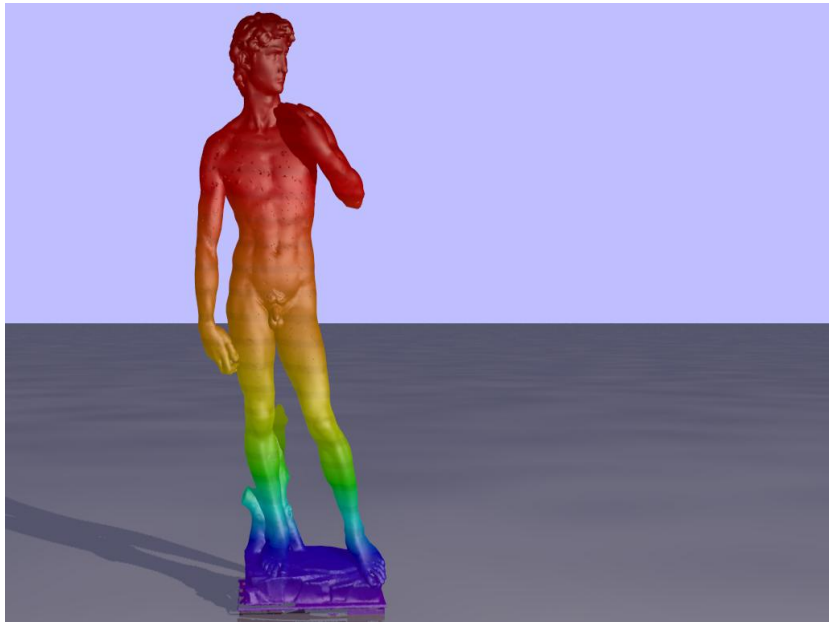




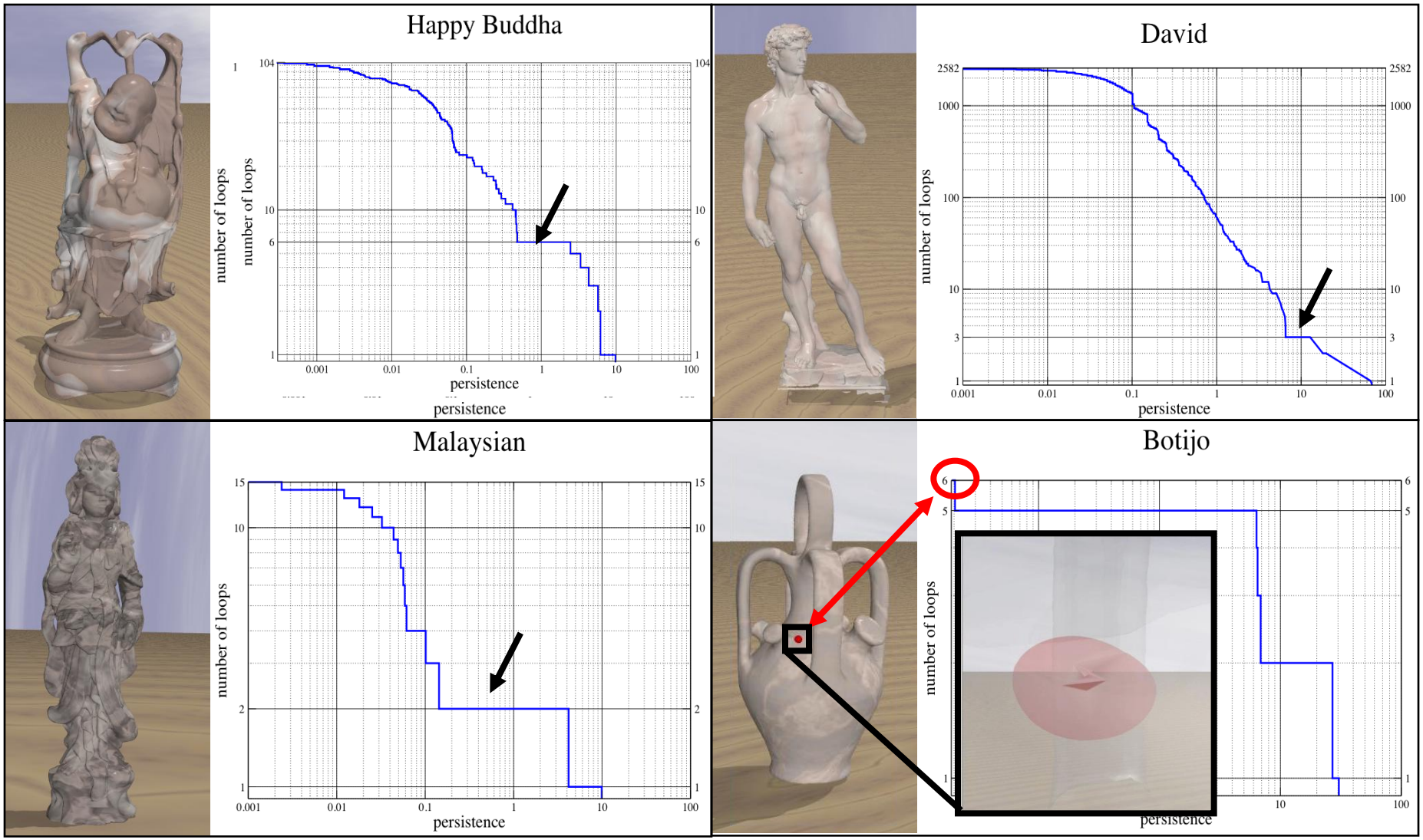


We Use Streaming Techniques to Achieve High Performance Analysis of Shapes

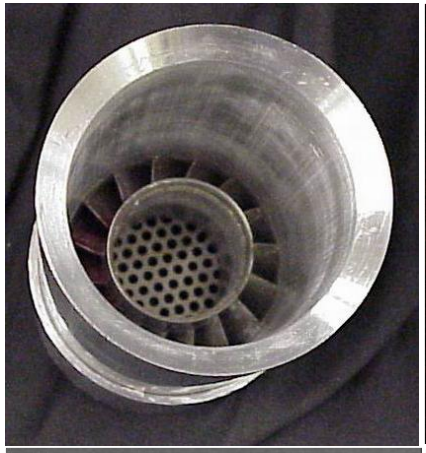


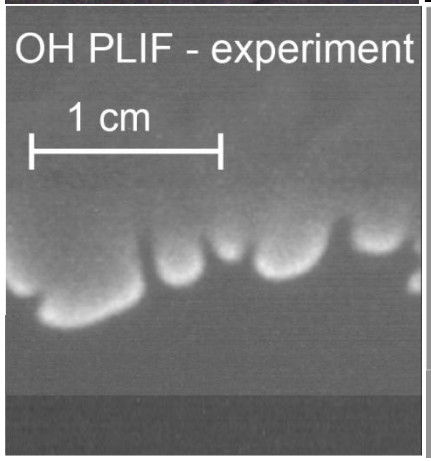


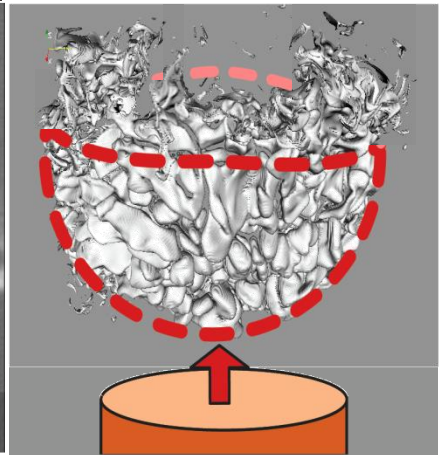
We Develop Shape Signatures to Find Defects in Large Scale Geometric Models



Understanding Turbulence for Low Emission, High Efficiency Combustion







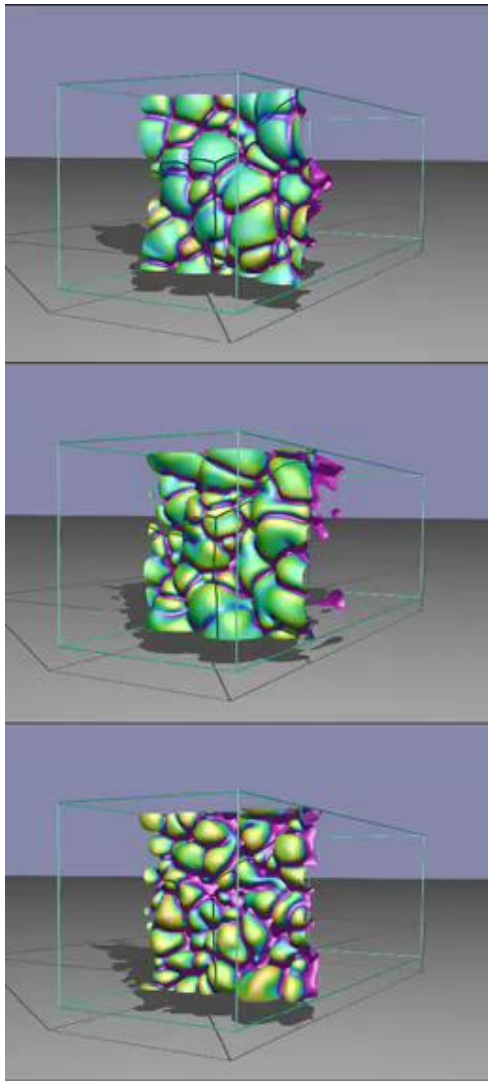
Experiment Simulation

- 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)

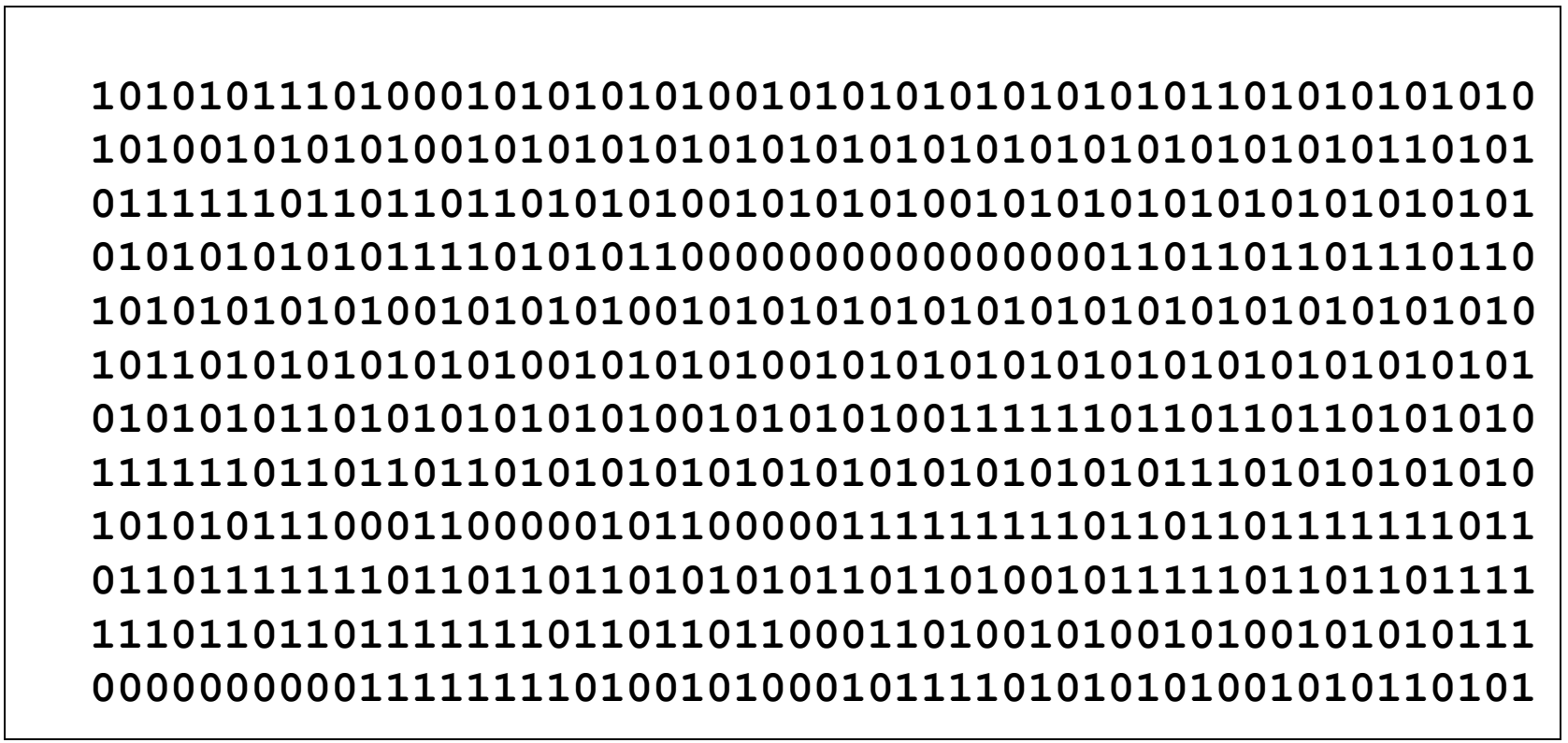
We Take on the Challenge of Developing a Quantitative Analysis Detecting Turbulence



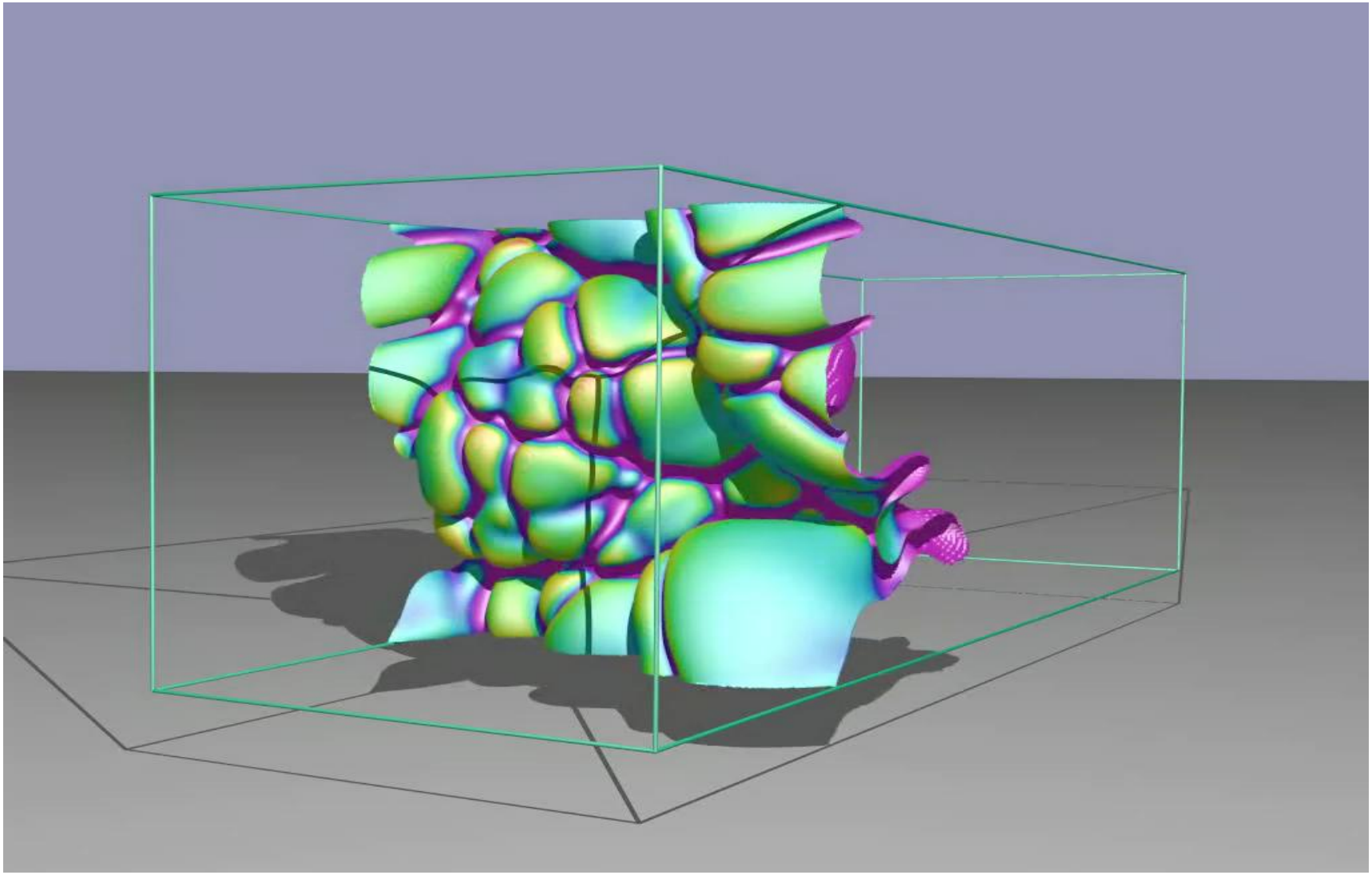
Understanding combustion processes over a broad range of burning conditions is an important problem for designing engines and power plants.

- **Simulation with AMR mesh.**
- **Simulations of lean premixed hydrogen flames with three degrees of turbulence.**
- **Can we identify precisely and track in time burning regions?**
- **Can we discriminate the degree of turbulence from a quantitative analysis?**

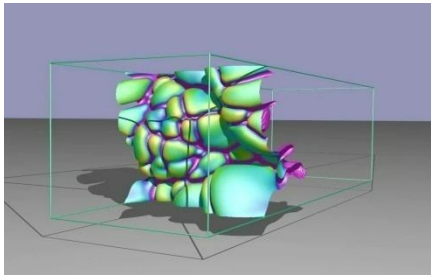
We Use Topology To Build Abstractions From Raw Data

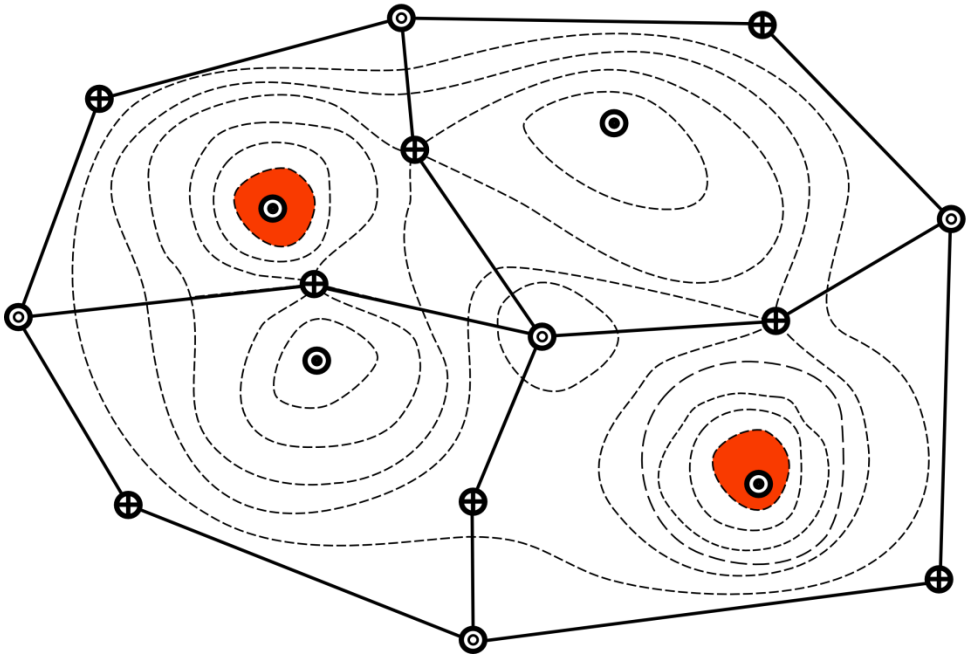


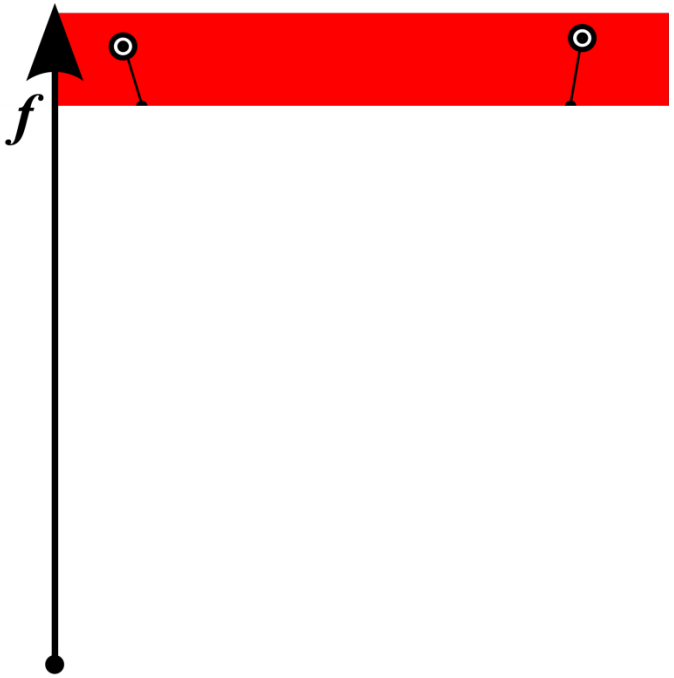
We build a hierarchical Morse complex of H2 consumption on an isotermal surface



We build a hierarchical Morse complex of H2 consumption on an isotermal surface



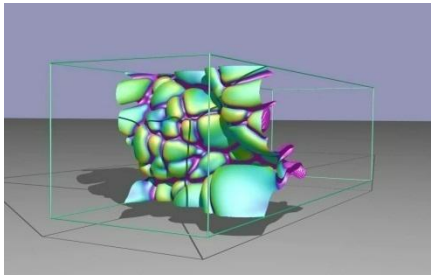


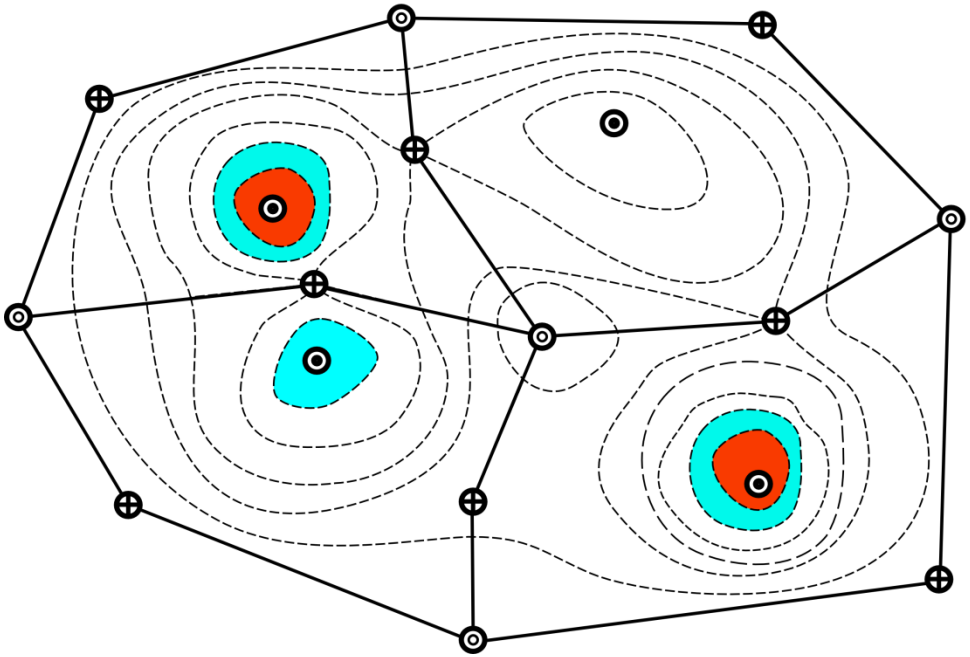


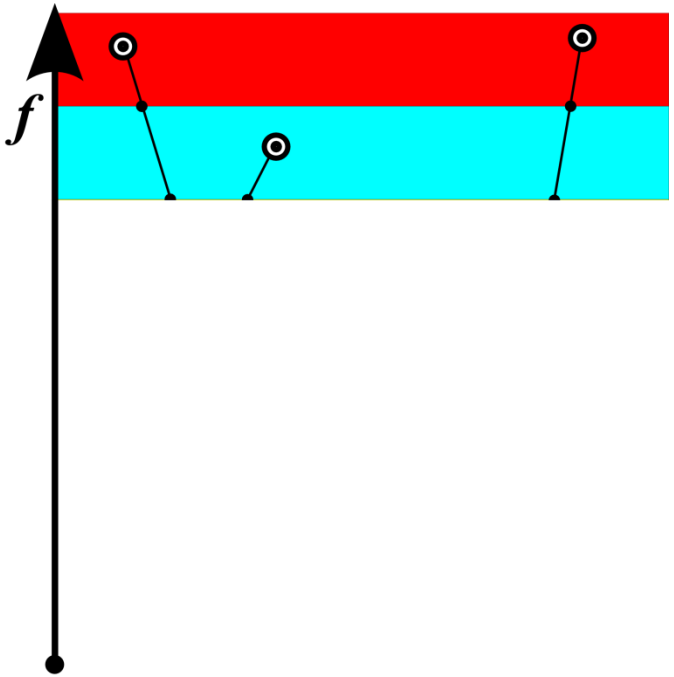
domain

hierarchy

We build a hierarchical Morse complex of H2 consumption on an isotermal surface



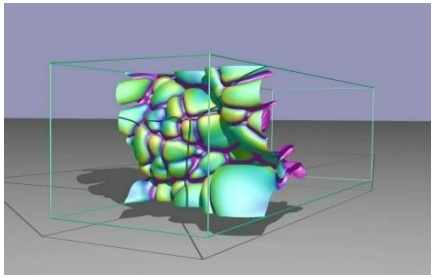


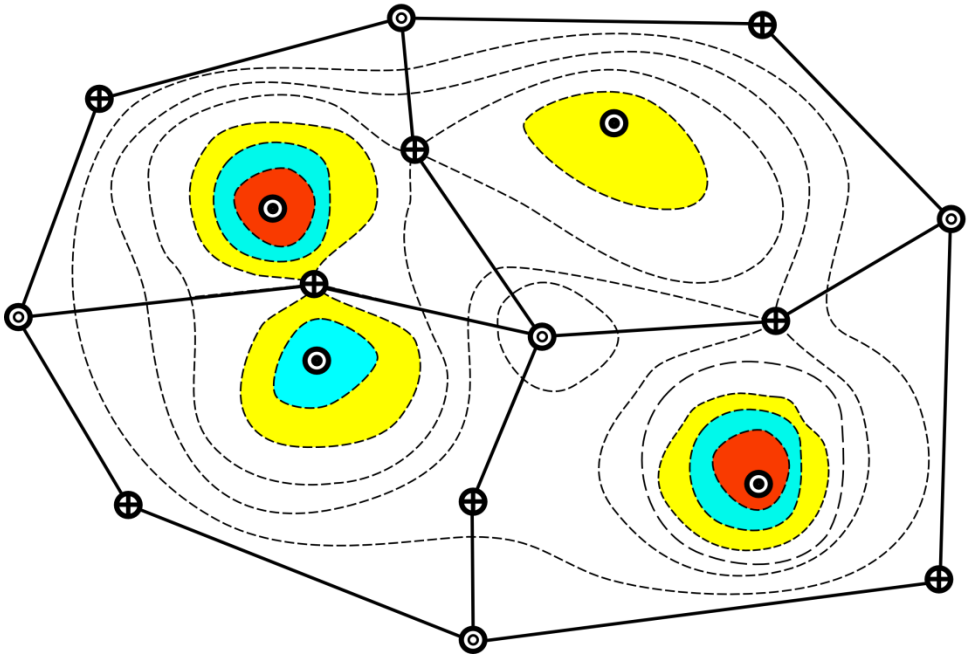


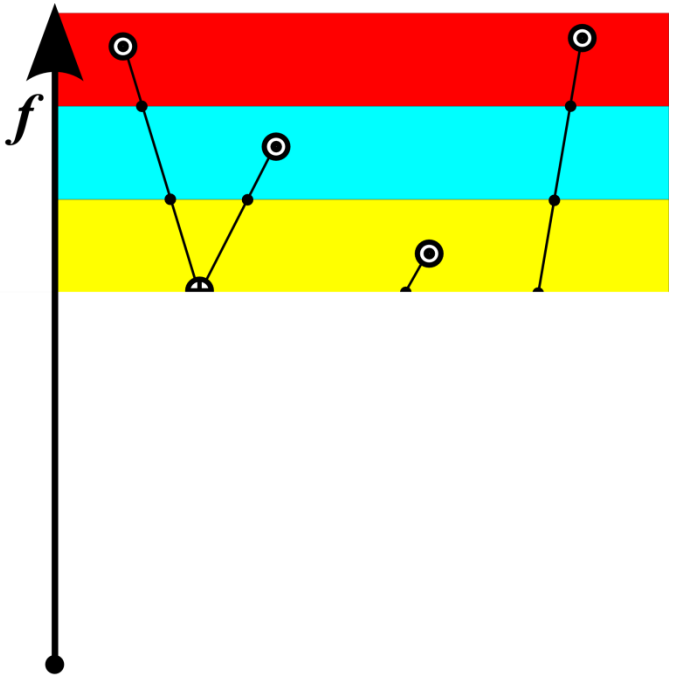
domain

hierarchy

We build a hierarchical Morse complex of H2 consumption on an isotermal surface



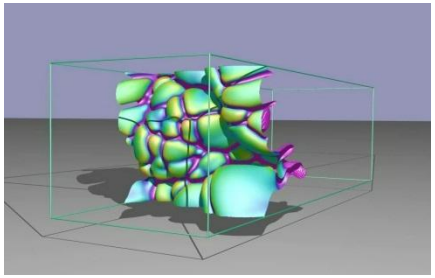


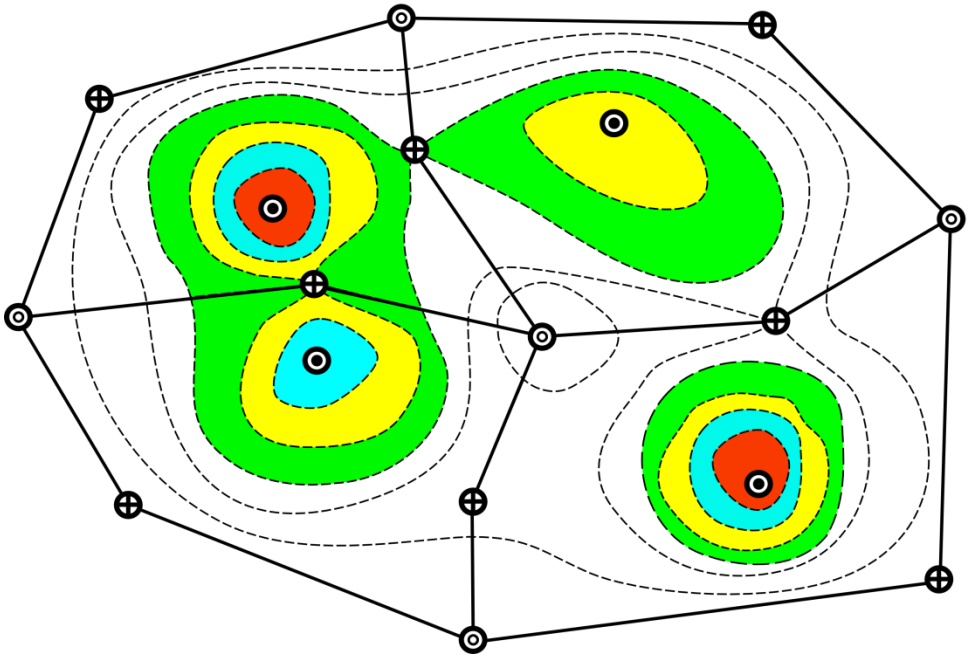


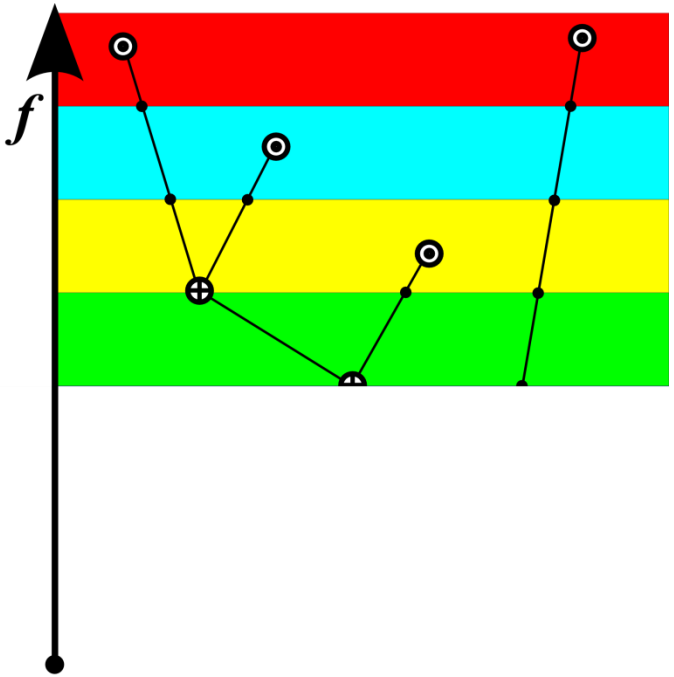
domain

hierarchy

We build a hierarchical Morse complex of H2 consumption on an isotermal surface



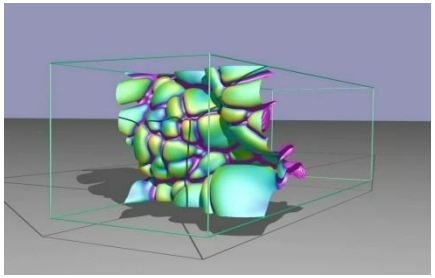


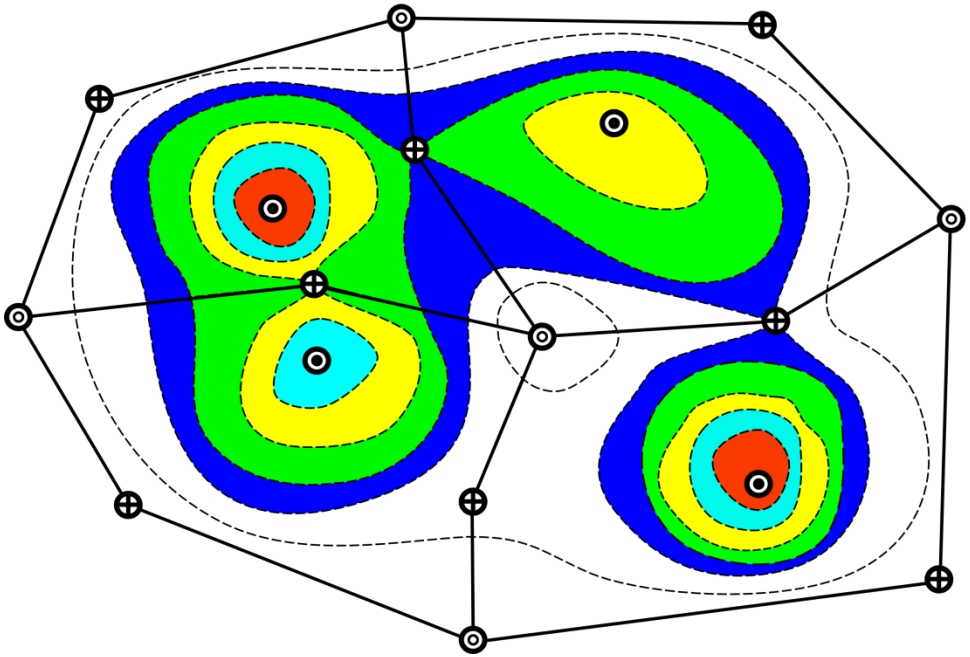


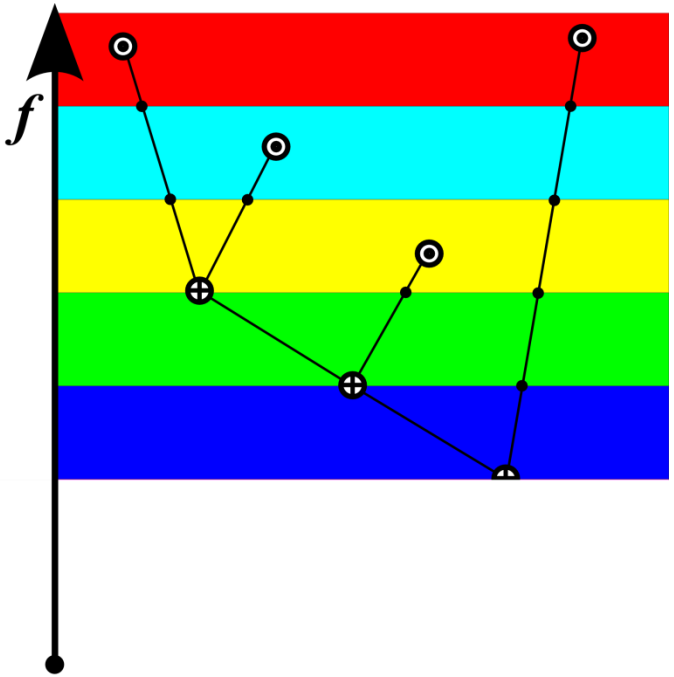
domain

hierarchy

We build a hierarchical Morse complex of H2 consumption on an isotermal surface



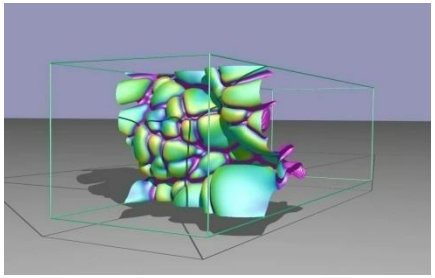


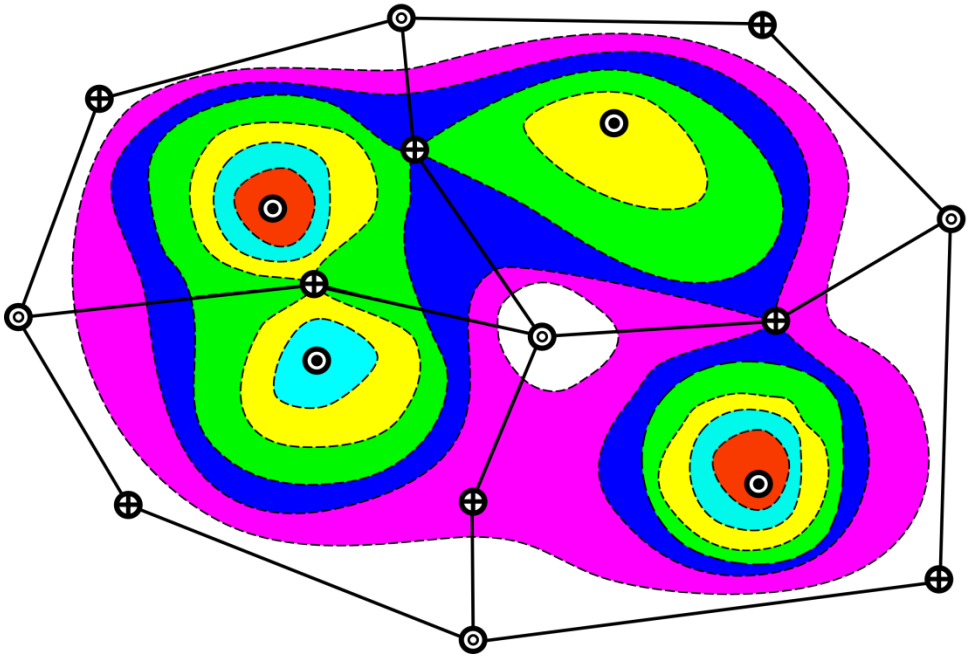


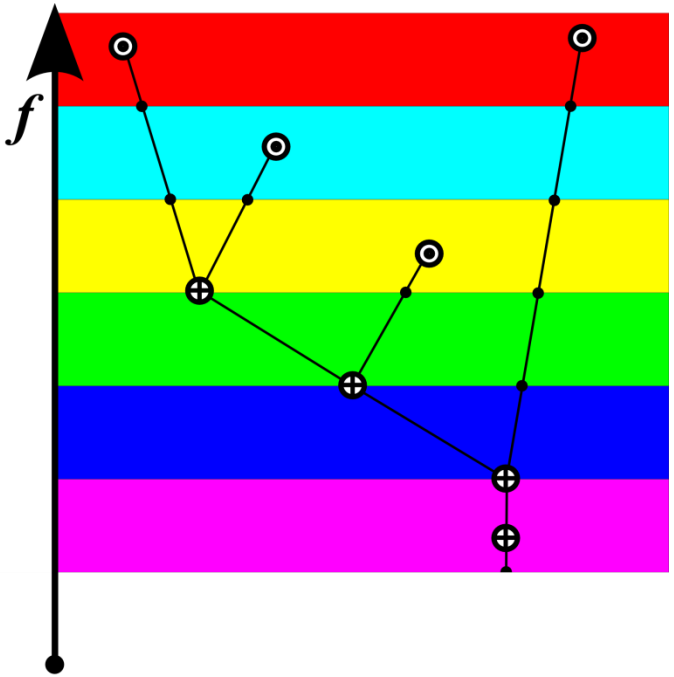
domain

hierarchy

We build a hierarchical Morse complex of H2 consumption on an isotermal surface



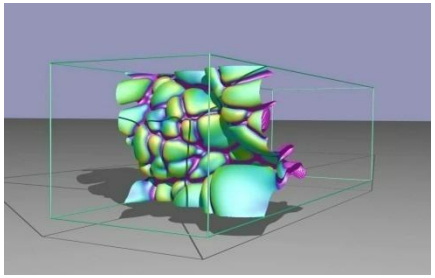


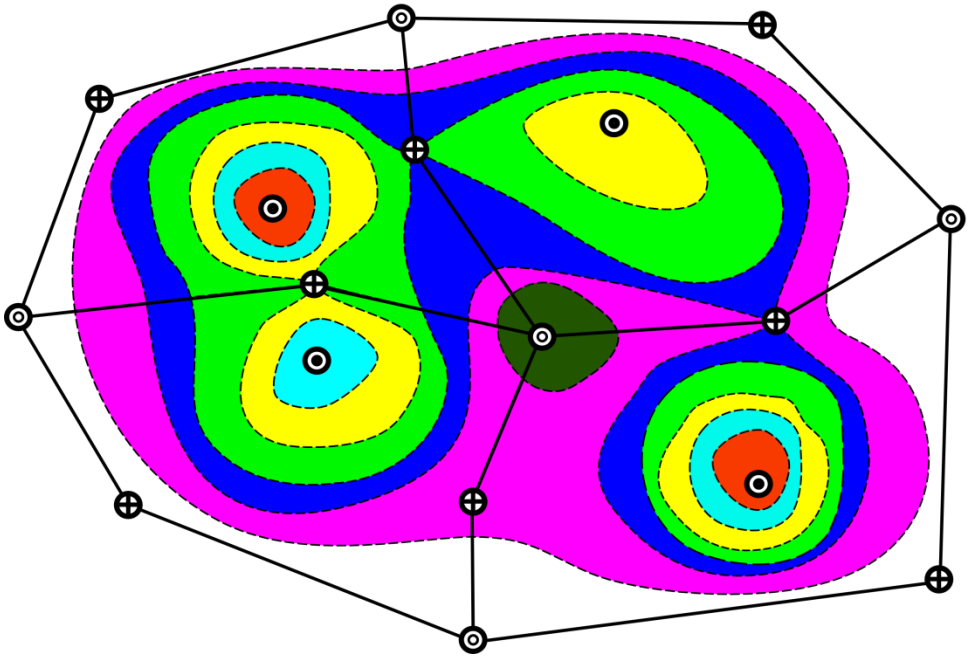


domain

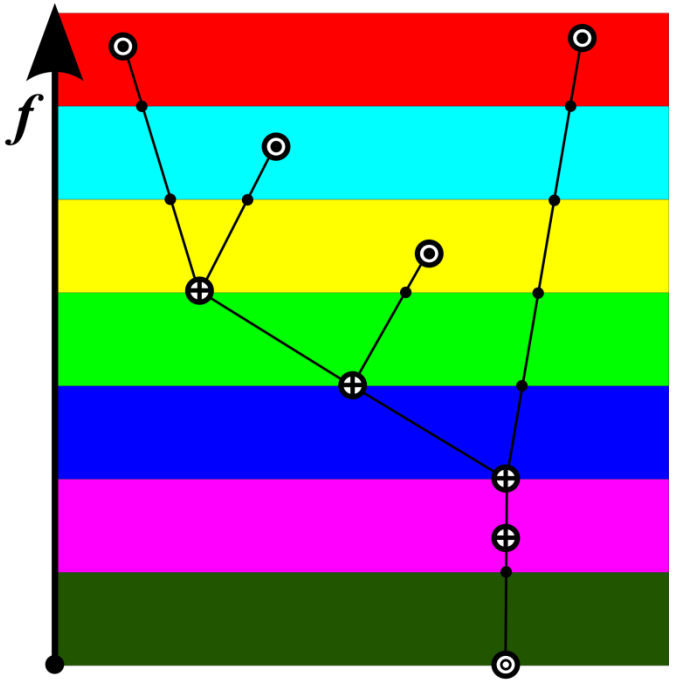
hierarchy

We build a hierarchical Morse complex of H2 consumption on an isotermal surface



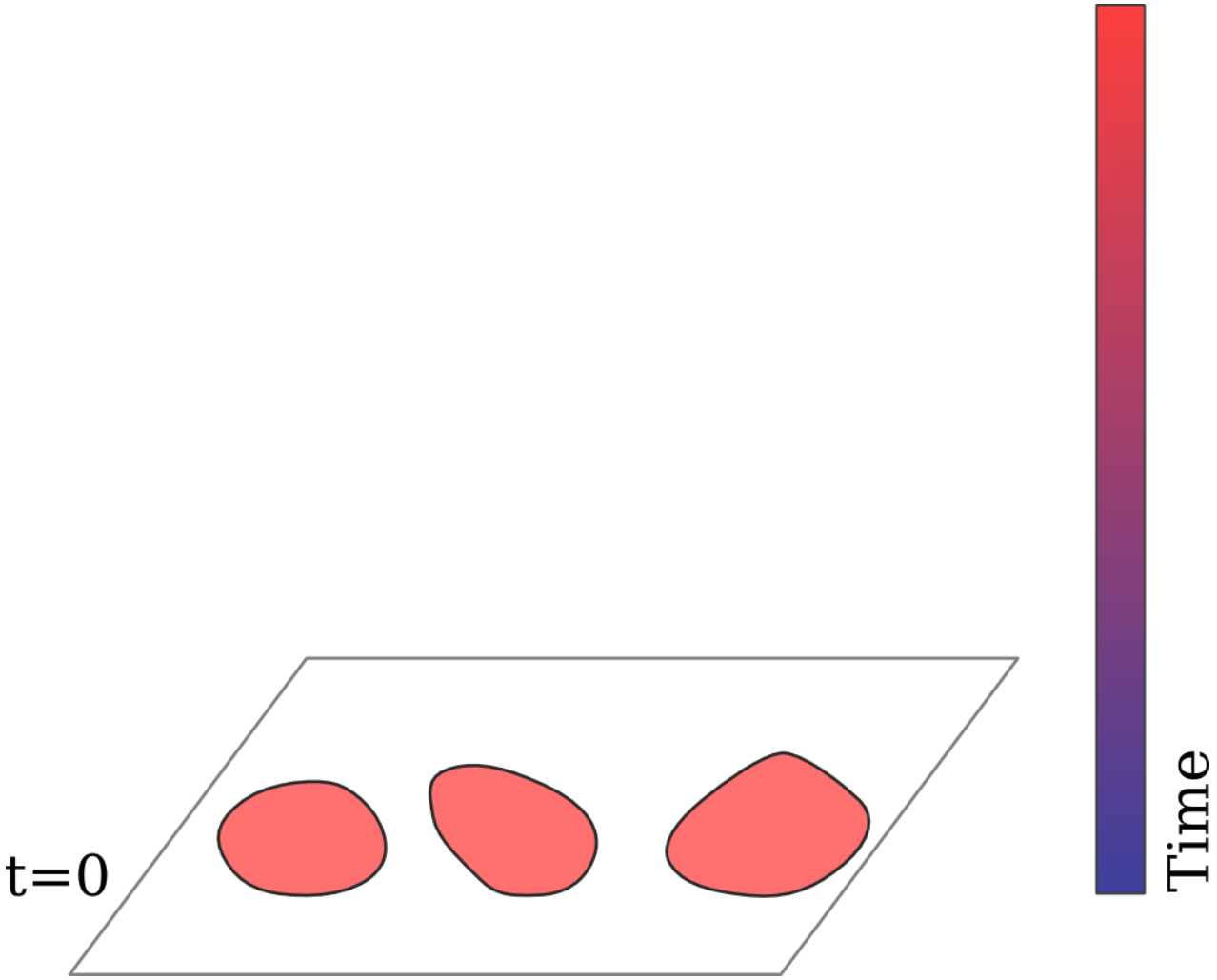


domain

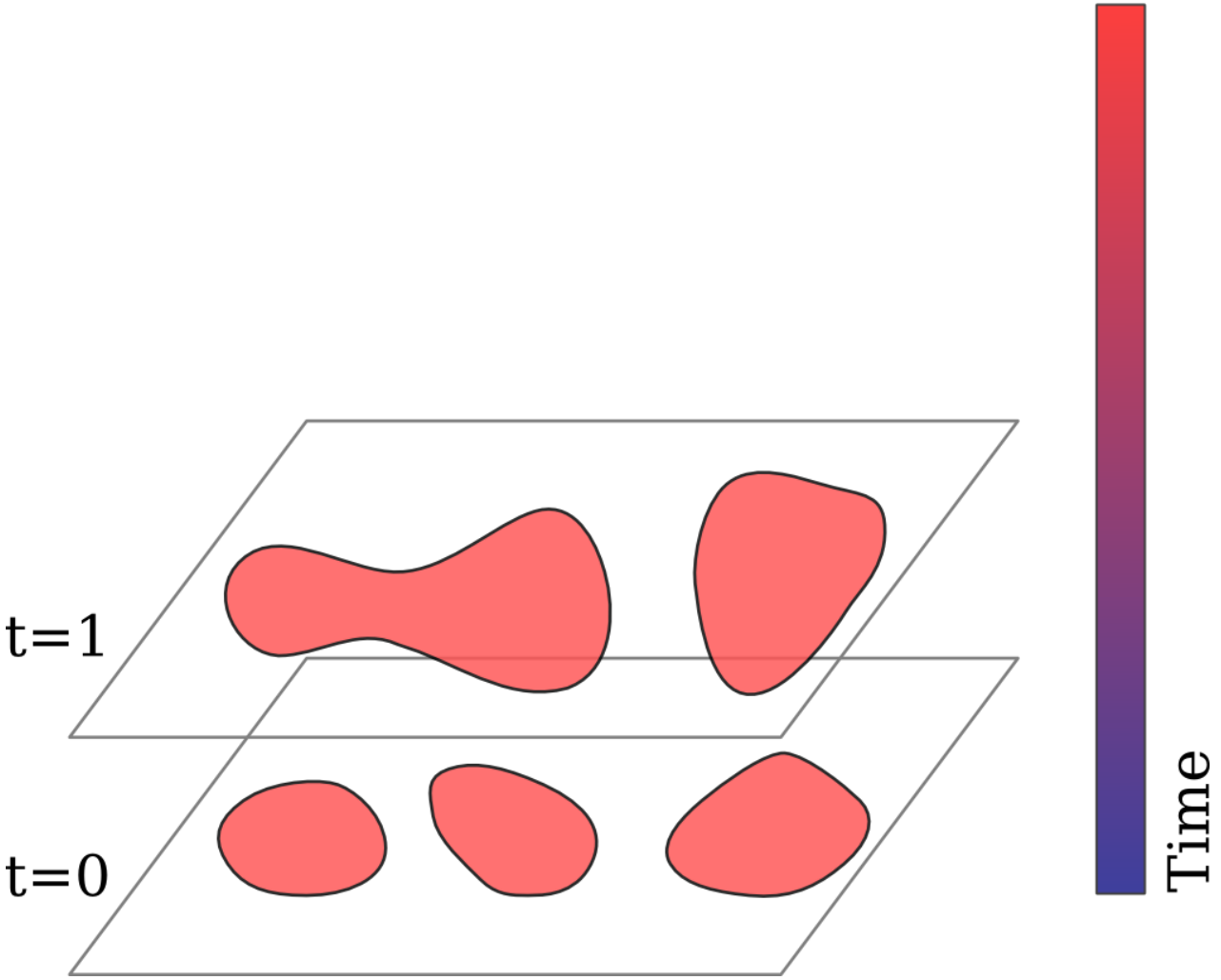


hierarchy

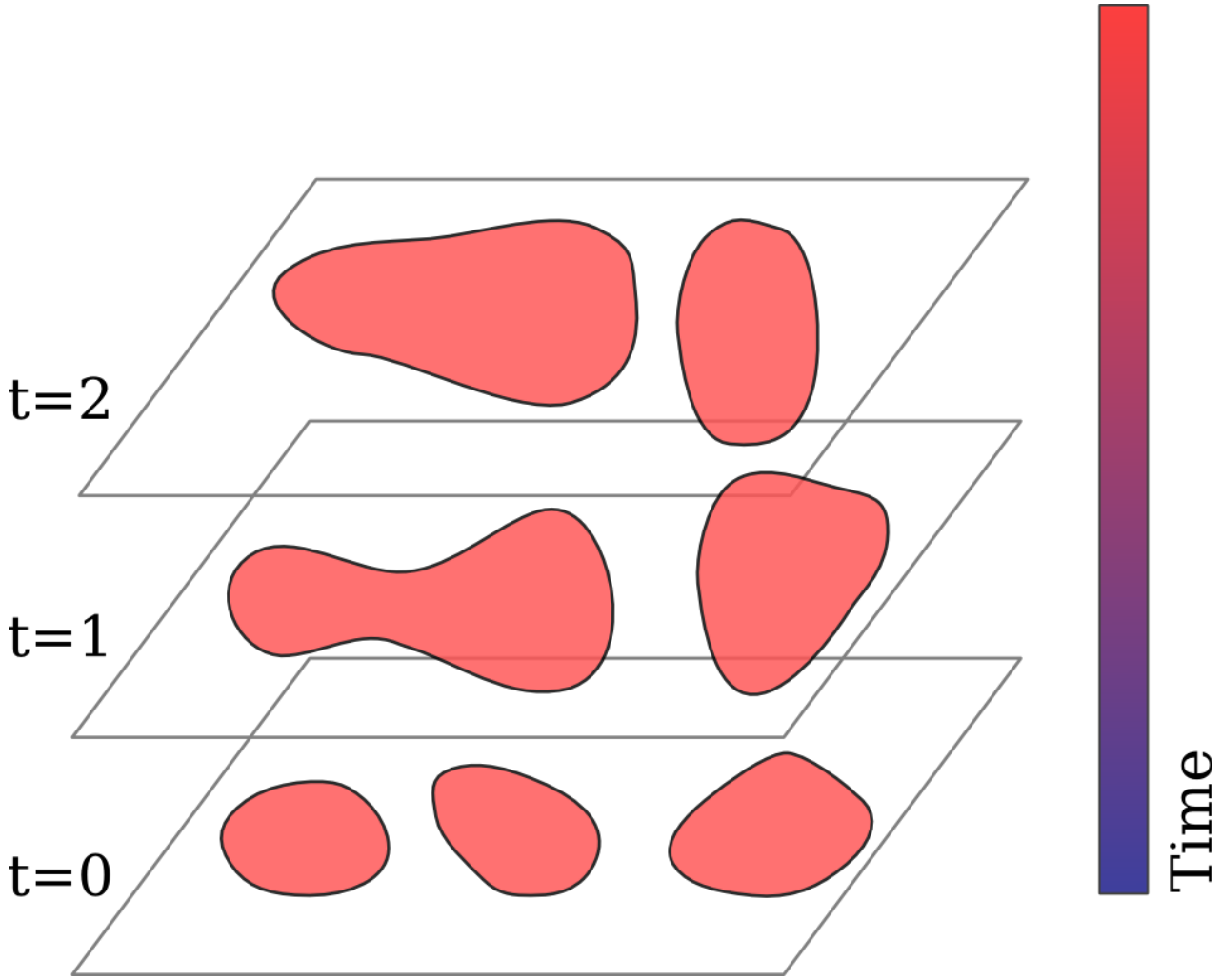
We track in time by interpolation in 4D and contraction to a graph (Reeb Graph)



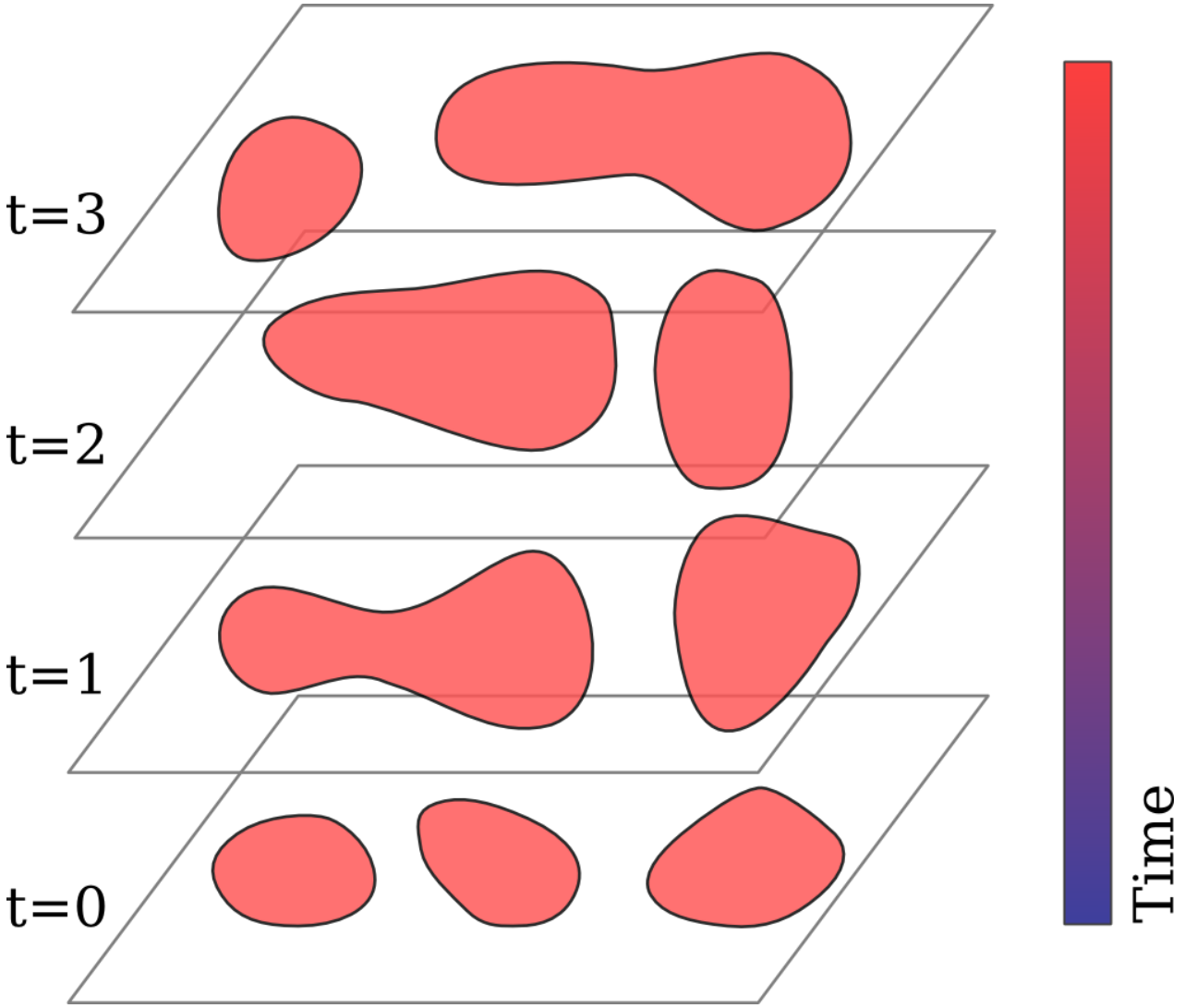
We track in time by interpolation in 4D and contraction to a graph (Reeb Graph)



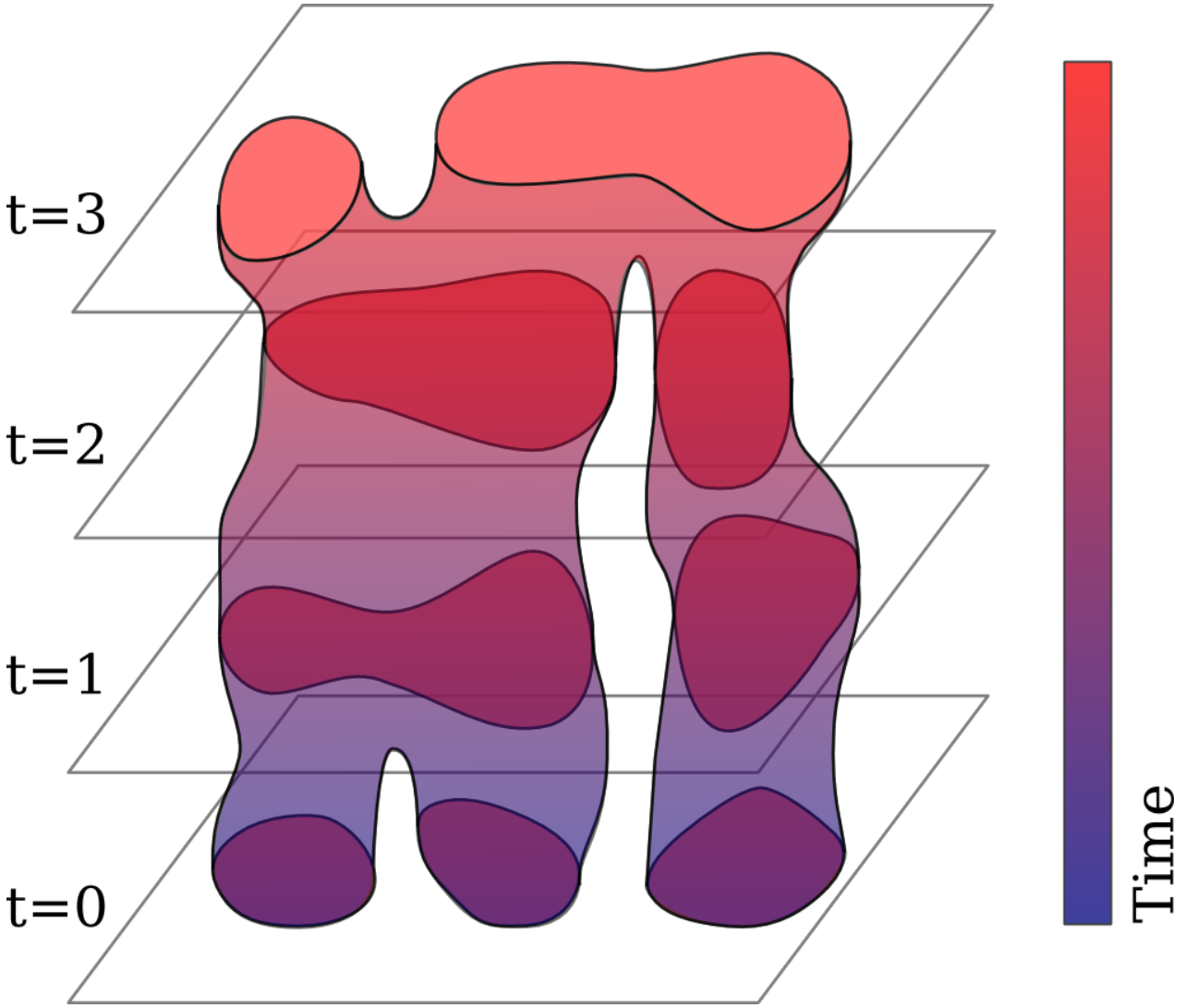
We track in time by interpolation in 4D and contraction to a graph (Reeb Graph)



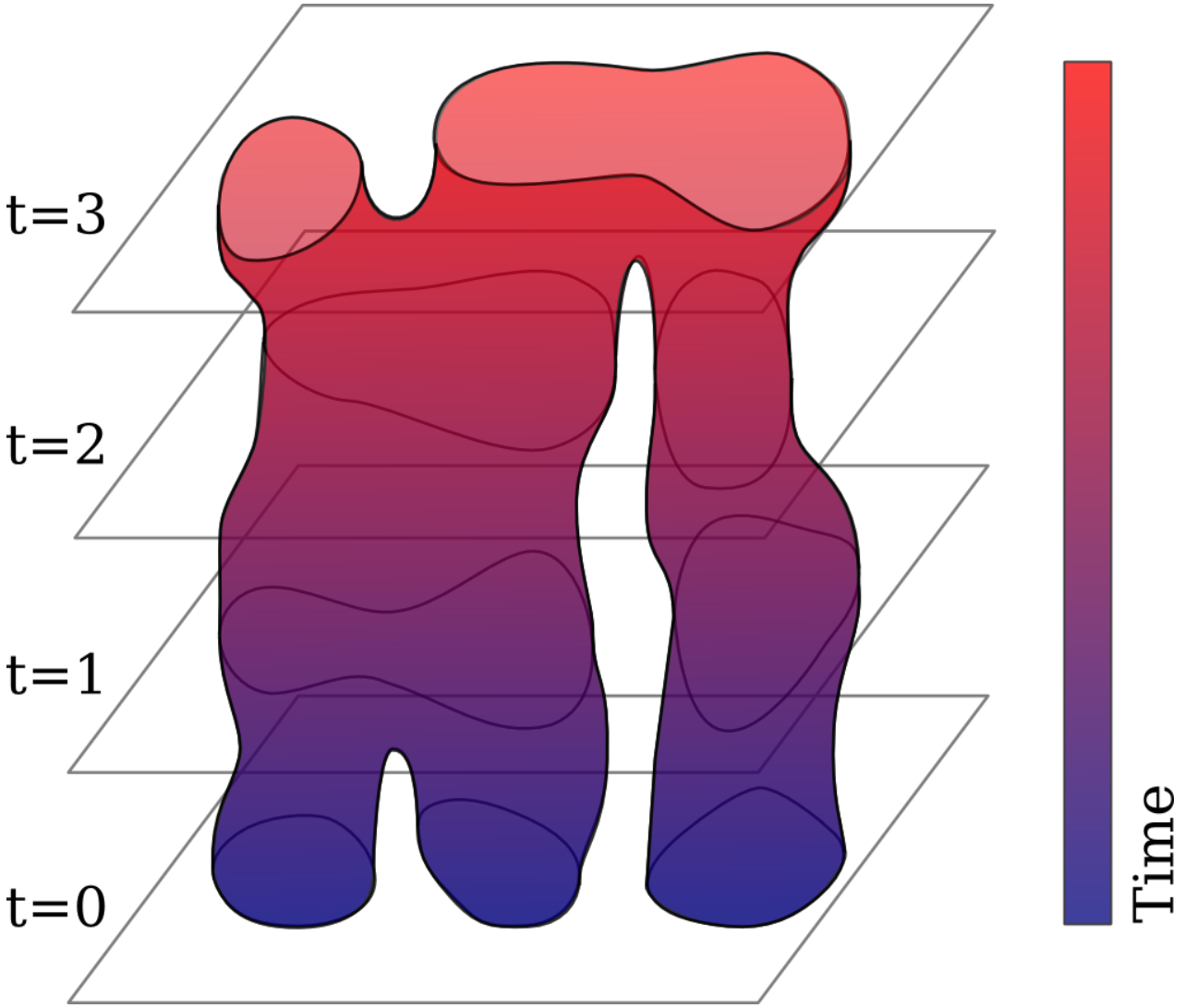
We track in time by interpolation in 4D and contraction to a graph (Reeb Graph)



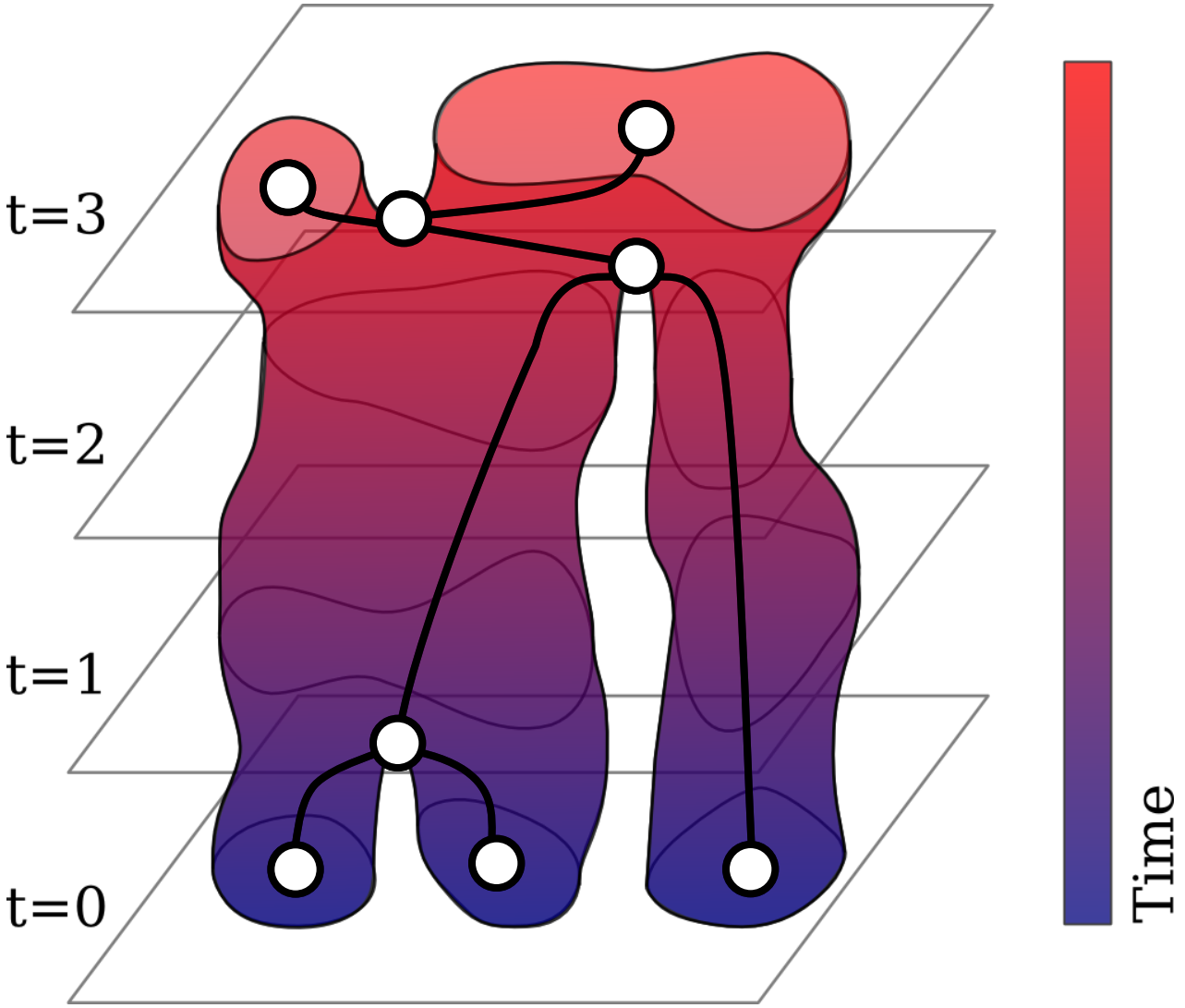
We track in time by interpolation in 4D and contraction to a graph (Reeb Graph)



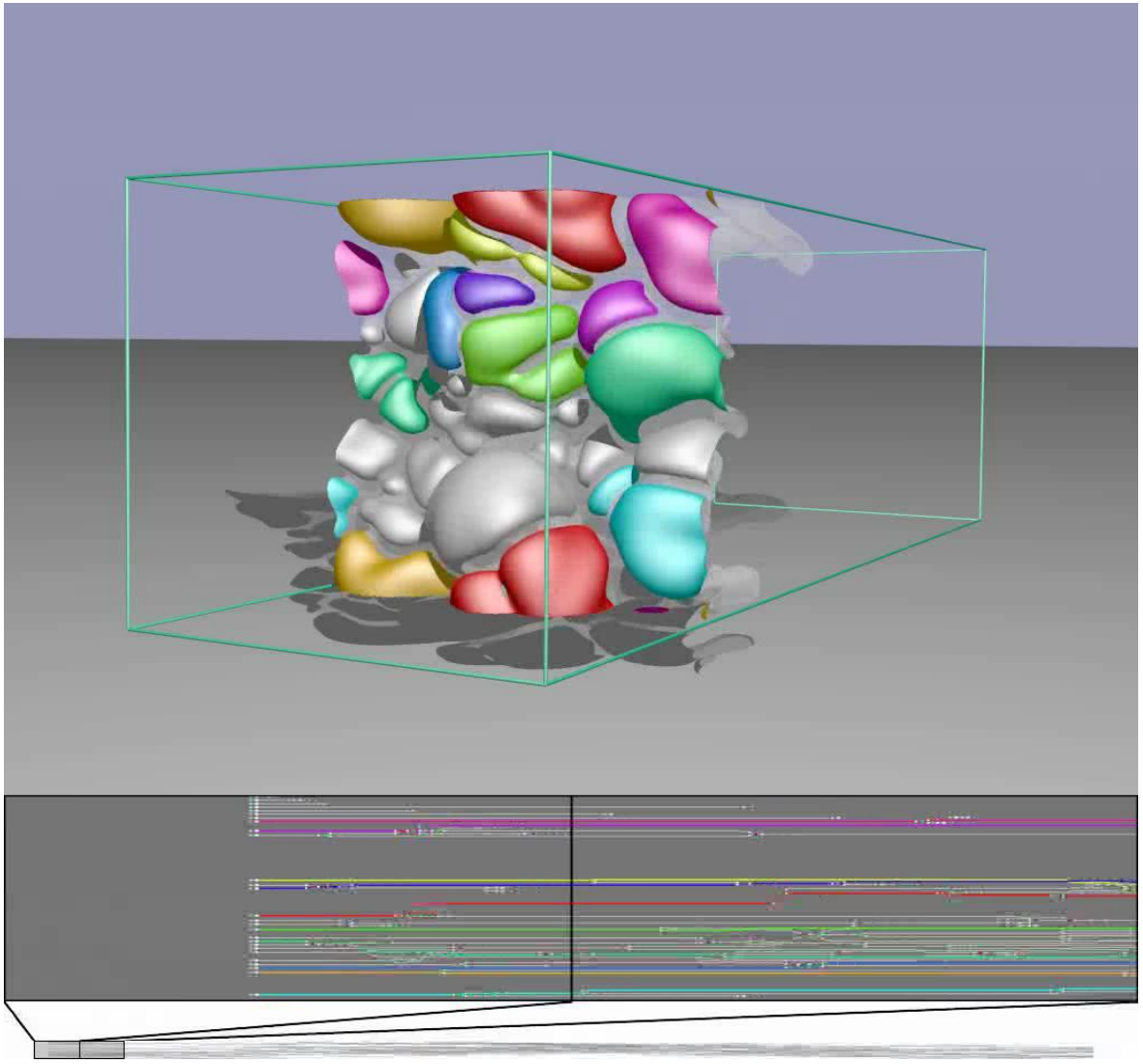
We track in time by interpolation in 4D and contraction to a graph (Reeb Graph)



We track in time by interpolation in 4D and contraction to a graph (Reeb Graph)

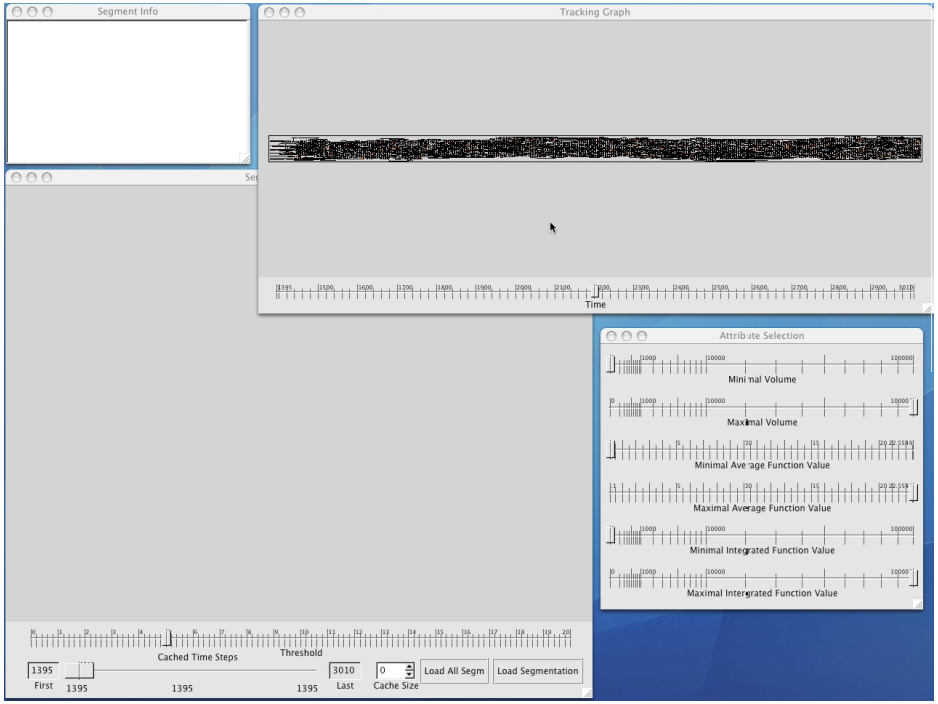


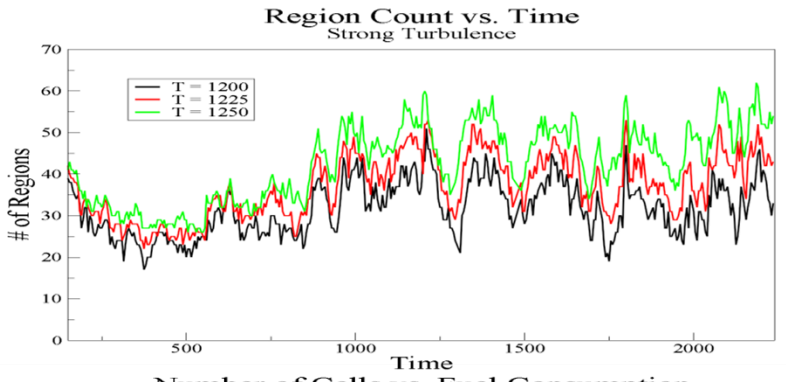
Each Set of Parameters Results in a Robust Segmentation and Tracking of Burning Cells

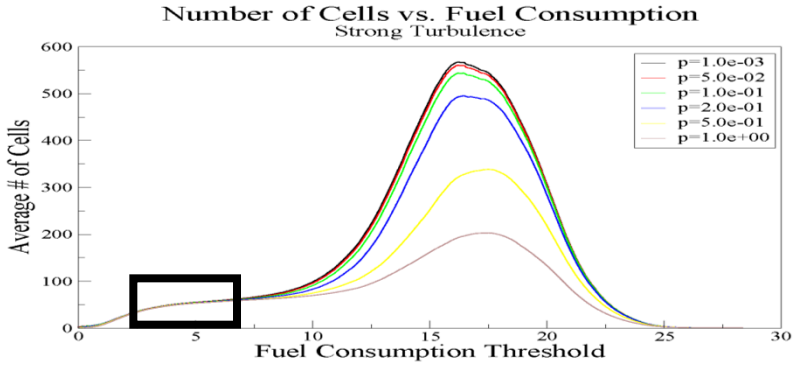


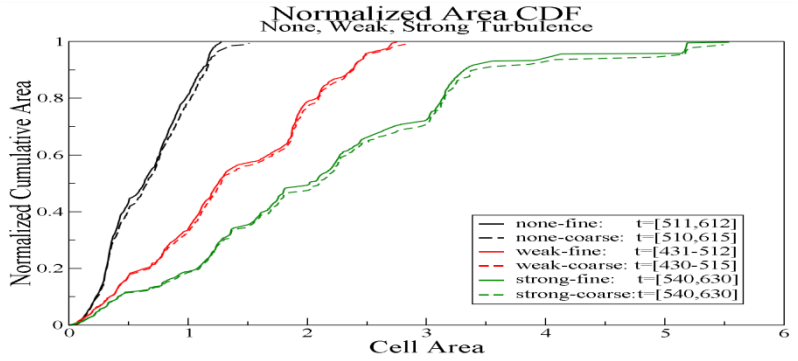
We Allow Exploration of the Full Space of Parameters Defining the Features

- interactive exploration
- comparison of statistics

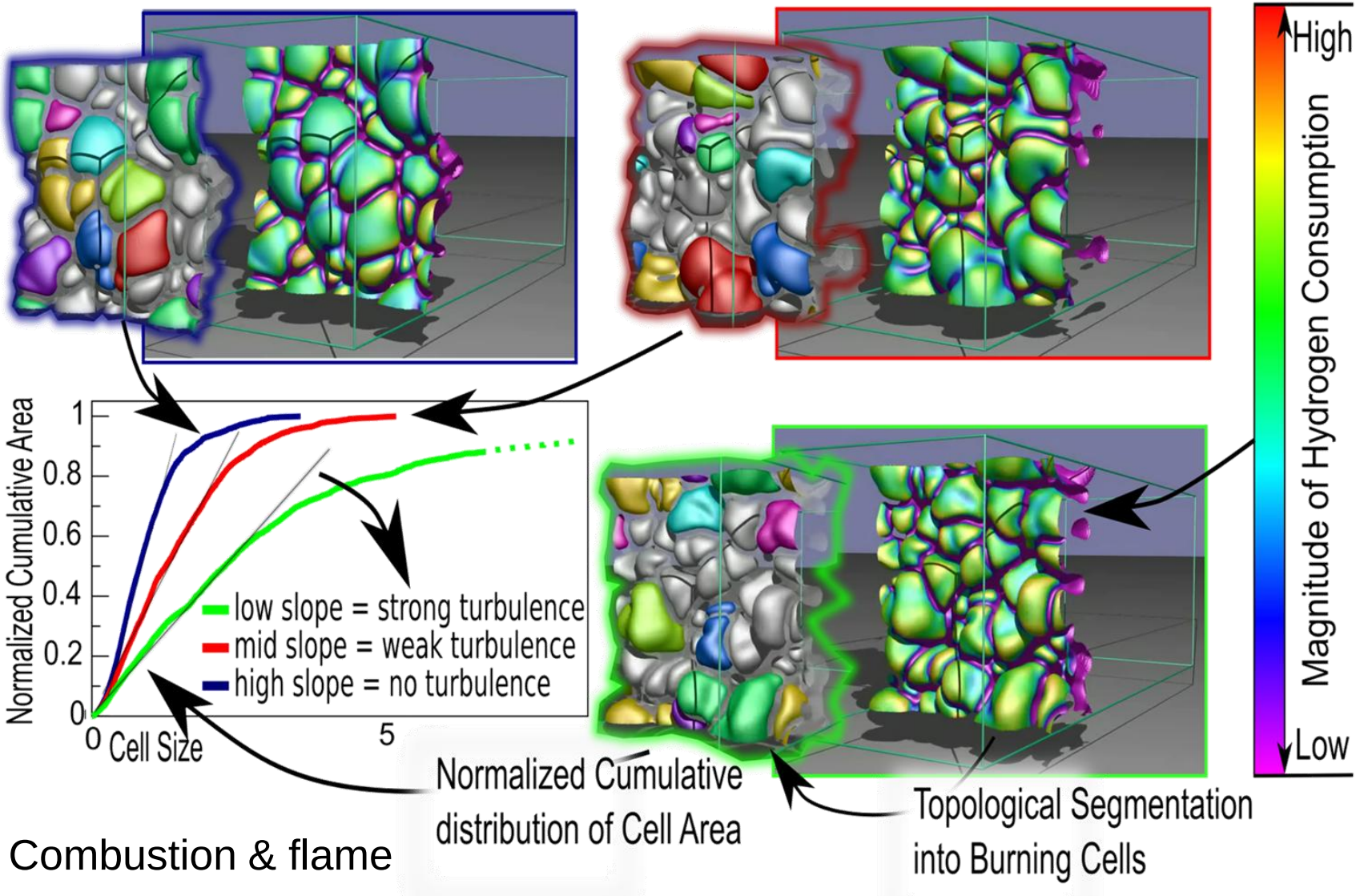








Topological Segmentation Allows to Quantify Turbulence as Slope of the Area Distributions



- Combustion & flame

Visual Analytics, Decision Support and Situational Awareness

(or *how to get the software out of the way)*

Yarden Livnat, Ph.D.
Scientific Computing and Imaging (SCI) Institute
Rocky Mountain Center of Excellence in Public Health Informatics
Intermountain Healthcare
VA Salt Lake City Health Care System

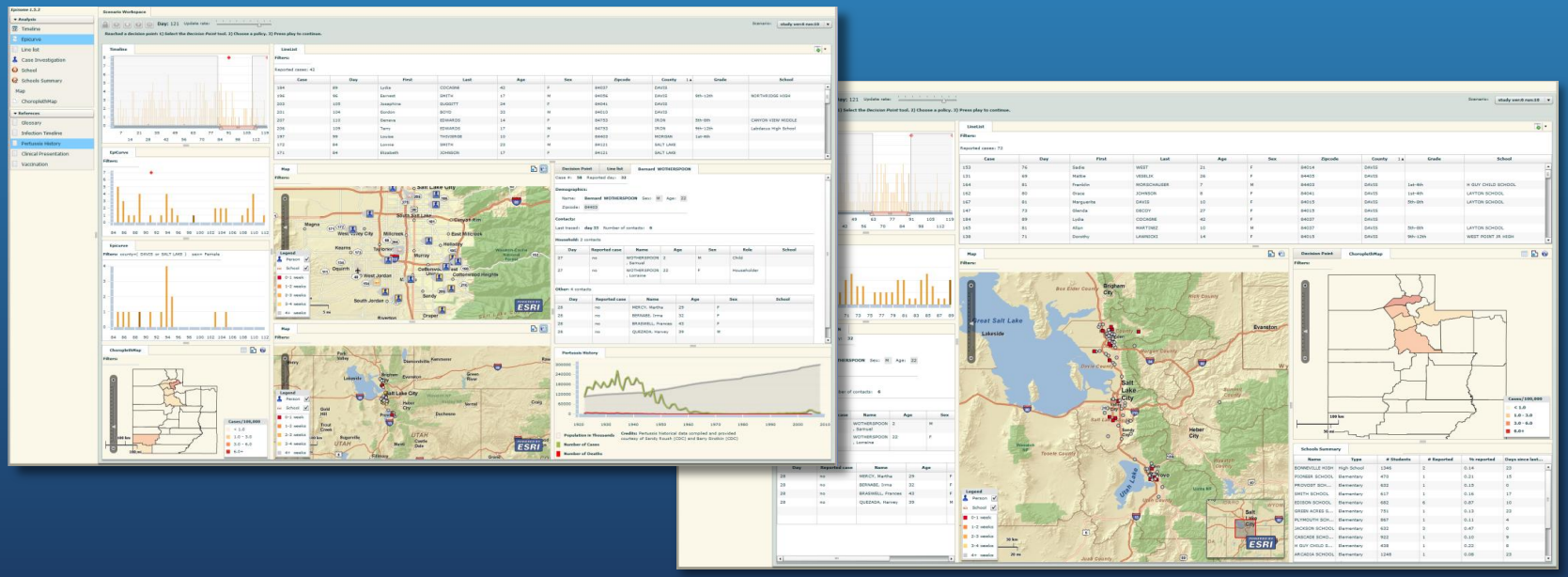
Methodologies and Principles

- Understand the stake holders
- Focus on what the user need to know rather than on what data is available
- Separate between the user and the incidental form the data is stored in
- Empower the user to explore the data
- Simple, direct and intuitive interactions
- Reduce information overload

Epinome:
An interactive web-based workbench for epidemic investigation

Epinome

- Facilitate investigation infectious disease outbreaks
  - High-fidelity and large scale simulations
  - Multiple scenarios with multiple decision points
- Display evolves as the user focus changes



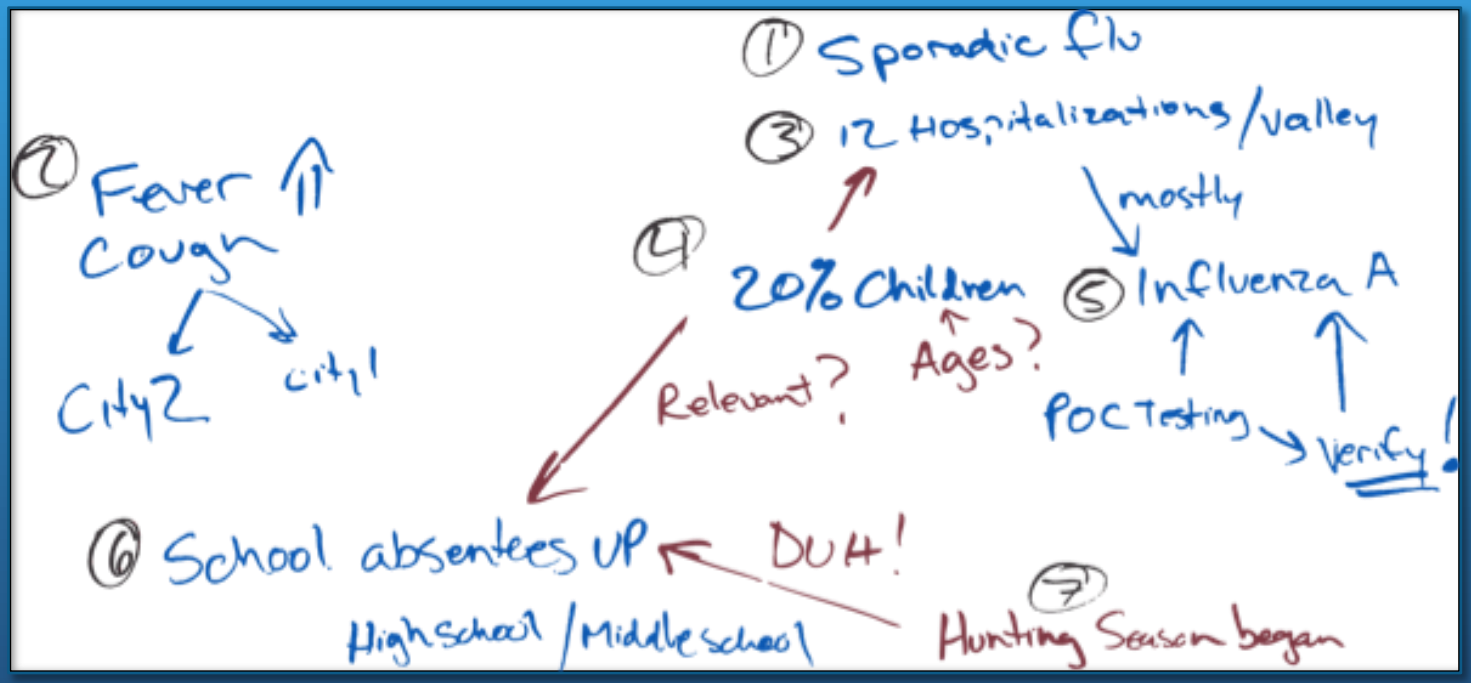
Epinome

- Coordinated multiple views
- Filter by example via drag-and-drop of displayed data (textual or graphical)
- Per-view (local) and global filters
- Nested workspaces
- Facilitate multiple lines of thought

CommonGround:
Infectious Disease Weather Map
An Interactive Visual Exploration of Temporal Correlations

How do users Capture Information?

Whiteboard illustration of mental model map of influenza activity



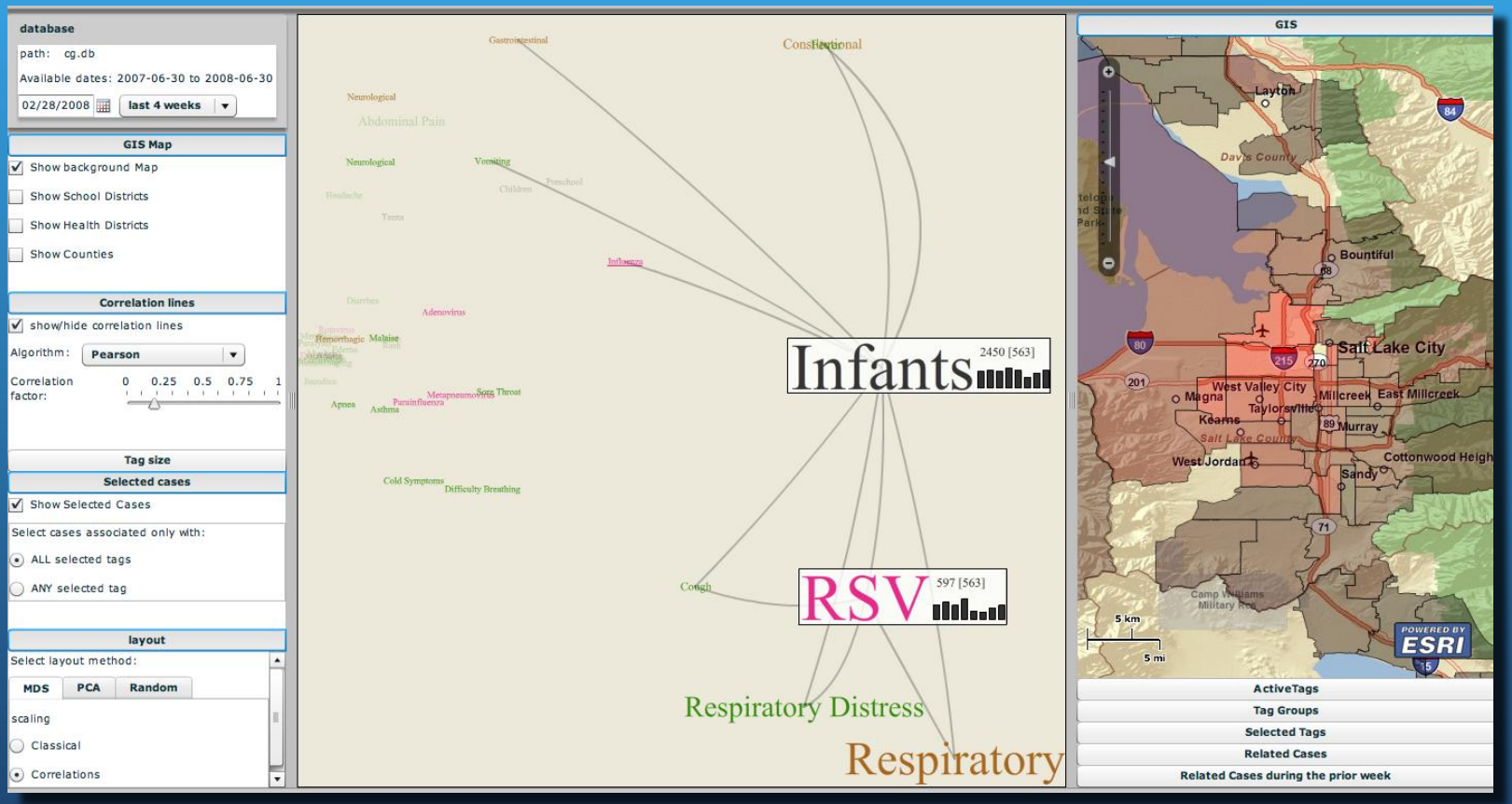
Graphical cues:

Glyphs, size, clustering, relations, correlations, annotations

The Conceptual Model

- An abstract representation
- Visualization of meta-data (tags) not raw data
- Number of tags is much smaller than number of data items
- Automatic layout based on temporal relations
- Suggests correlations between tags
- Dynamic evolving queries
- Employ information scent to reduce information overload

The CommonGround Prototype



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**