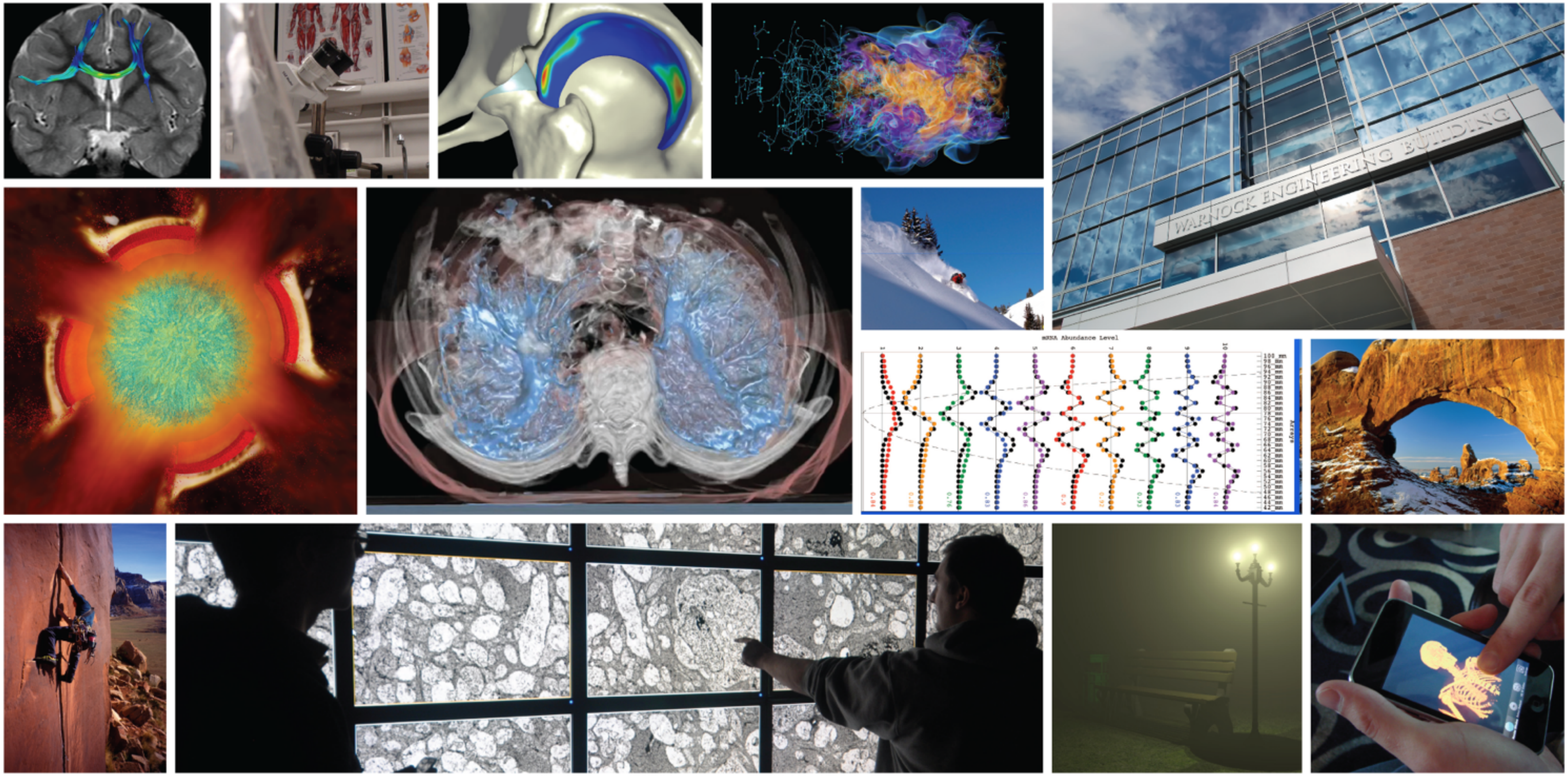
Big Data: A Scientific Visualization Perspective



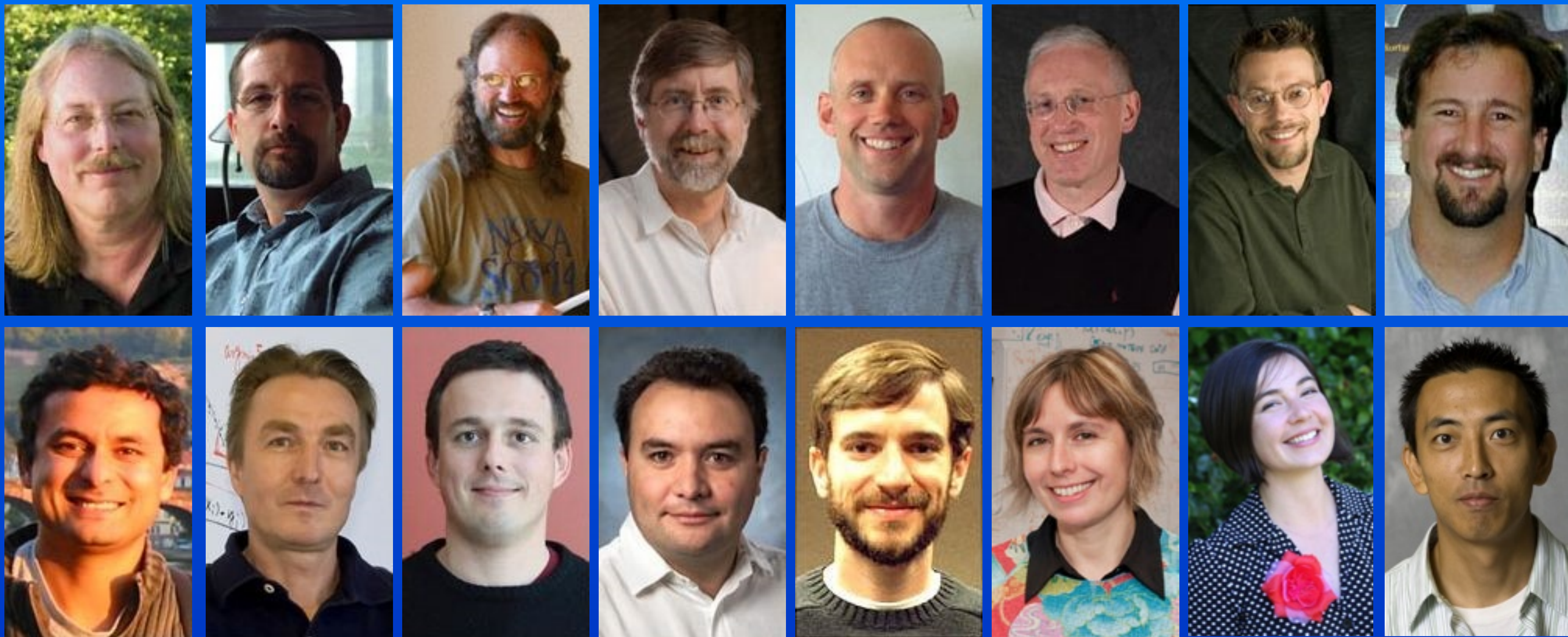
Chuck Hansen

Scientific Computing and Imaging Institute

University of Utah


COMPUTING
VISUALIZATION
Research
at SCI
IMAGING
APPLICATIONS

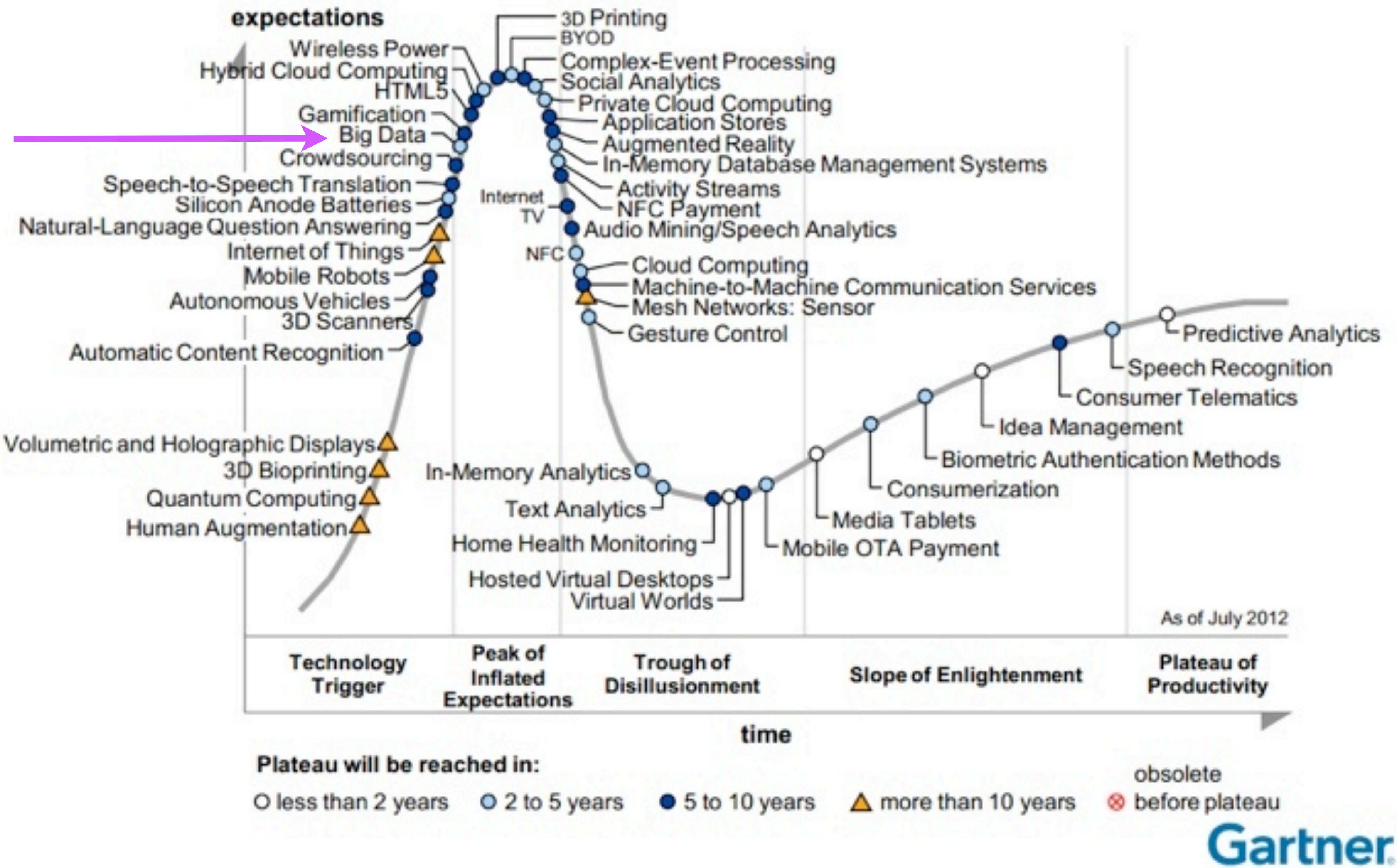
SCI Institute Faculty



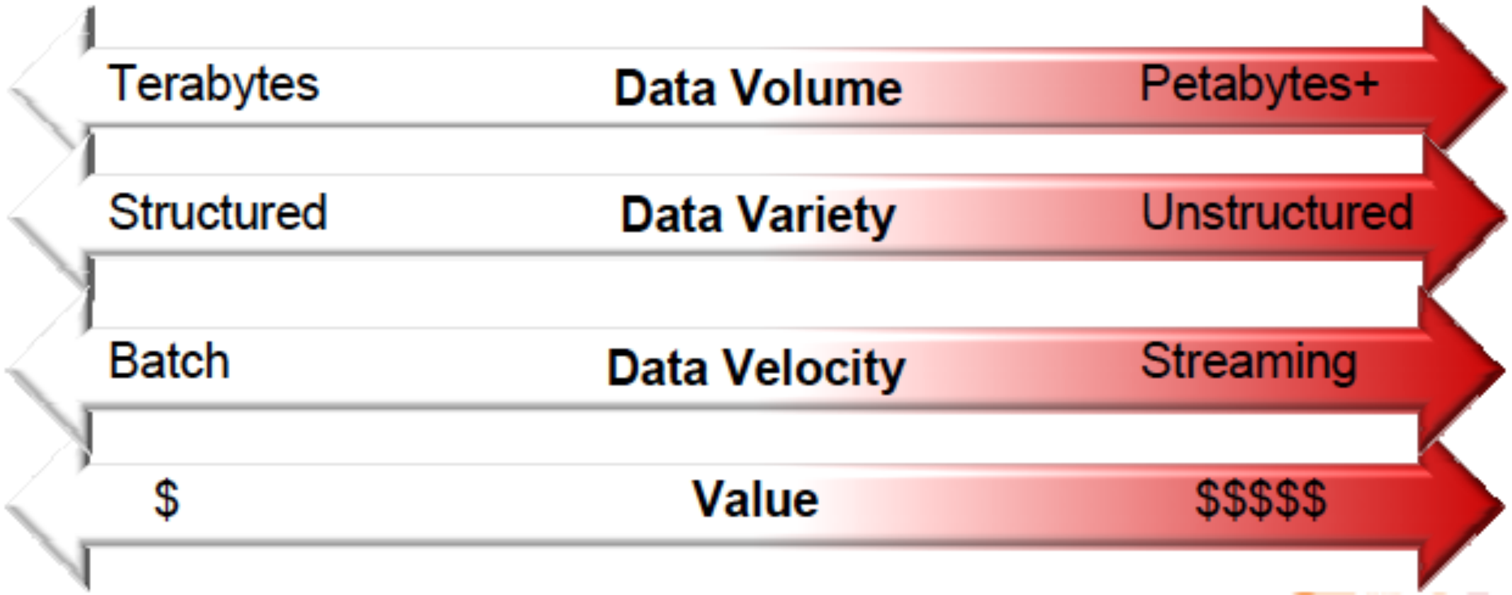
www.sci.utah.edu


VISUALIZATION
COMPUTING
Research
at SCI
IMAGING
APPLICATIONS

Emerging Technologies Hype Cycle 2012



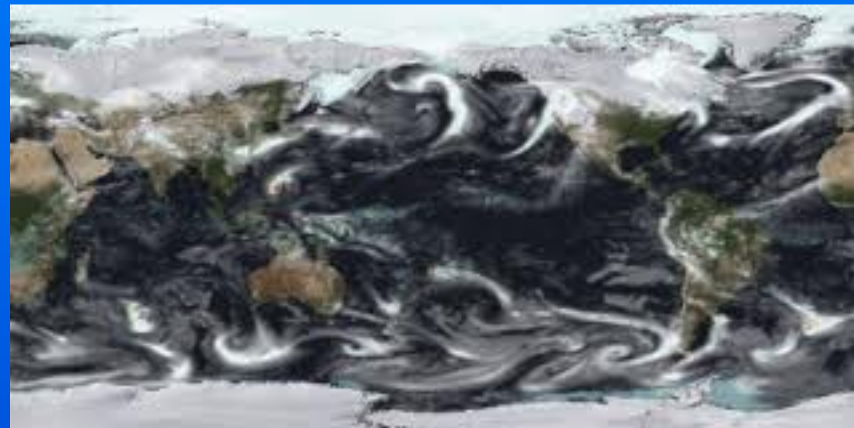
Big Data: The Vs

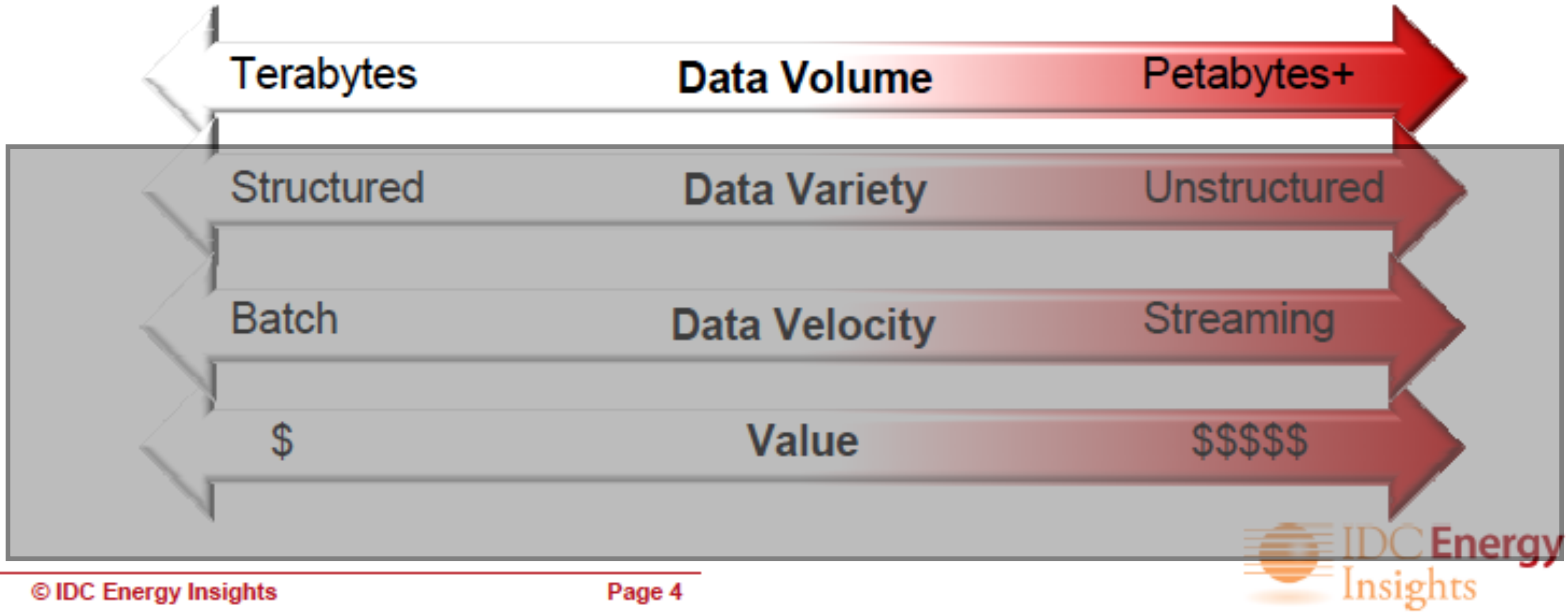


IDC Energy Insights

Big Data: The Vs

High Resolution Models

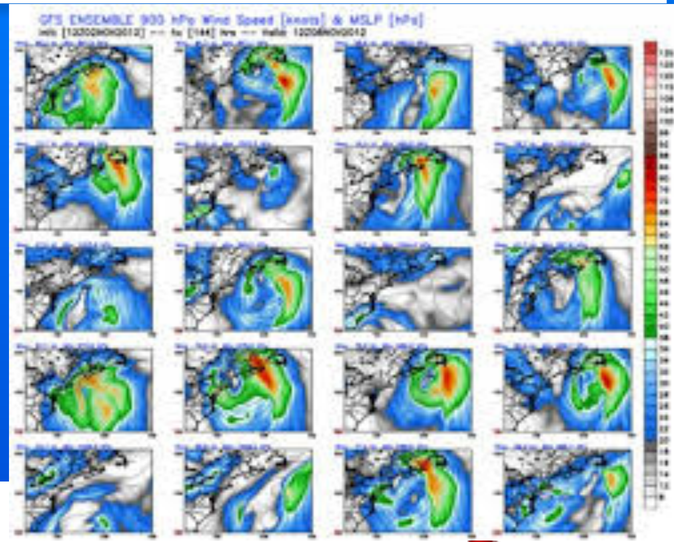


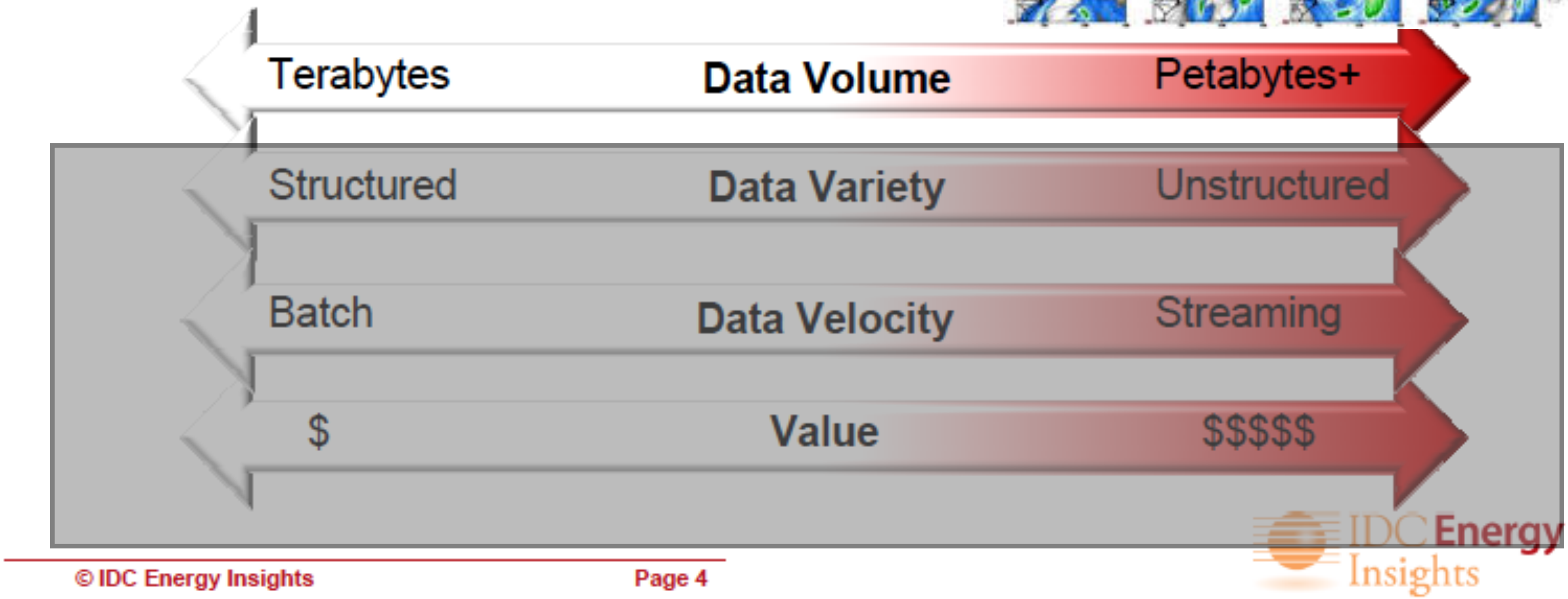


Big Data: The Vs

High Resolution Models

Ensembles





Big Data: The V

Sensor data

Coupled models

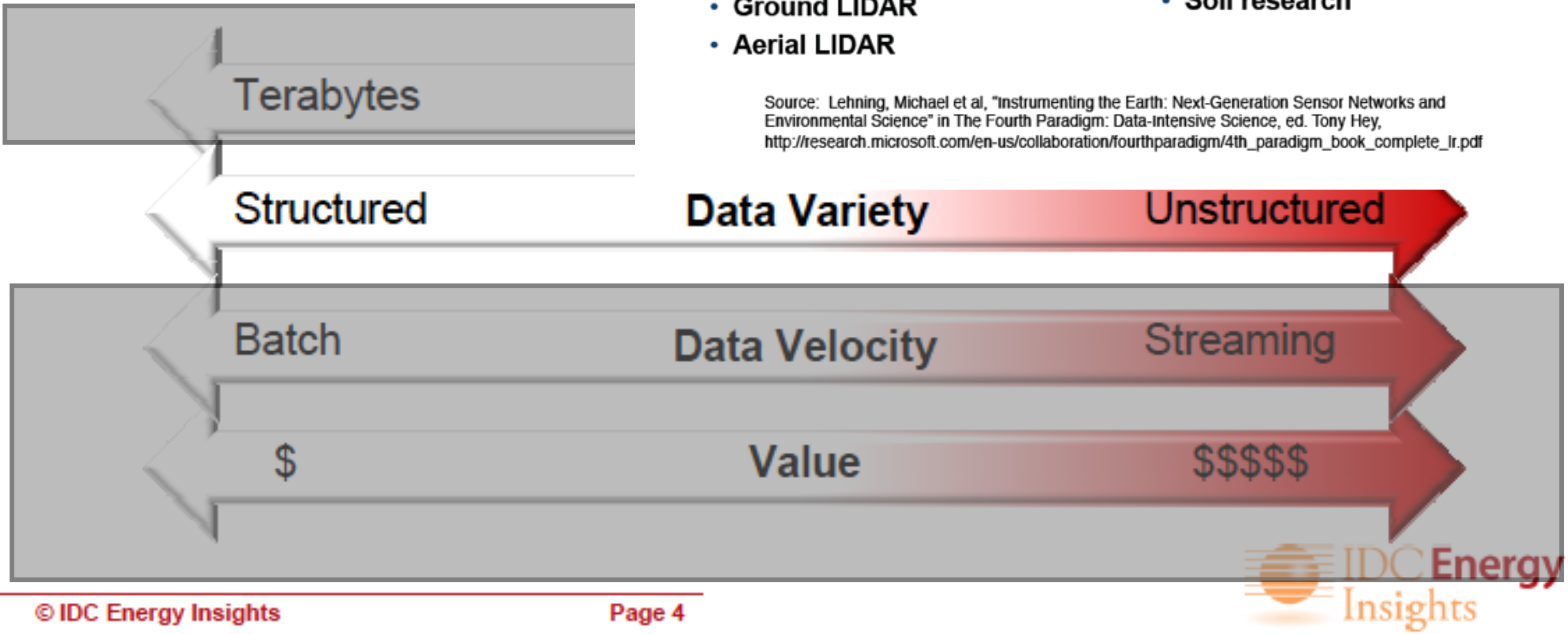
U.S. DEPARTMENT OF ENERGY

Office of Science

Heterogeneous Data: Environmental Science Example

- Mobile stations
- High-resolution weather stations
- Full-size snow/weather stations
- External weather stations
- Satellite imagery
- Weather radar
- Mobile weather radar
- Stream observations
- Citizen-supplied observations
- Ground LIDAR
- Aerial LIDAR
- Nitrogen/methane measures
- Snow hydrology & avalanche probes
- Seismic probes
- Distributed optical fiber temperature sensing
- Water quality sampling
- Stream gauging stations
- Rapid mass movements research
- Run-off stations
- Soil research

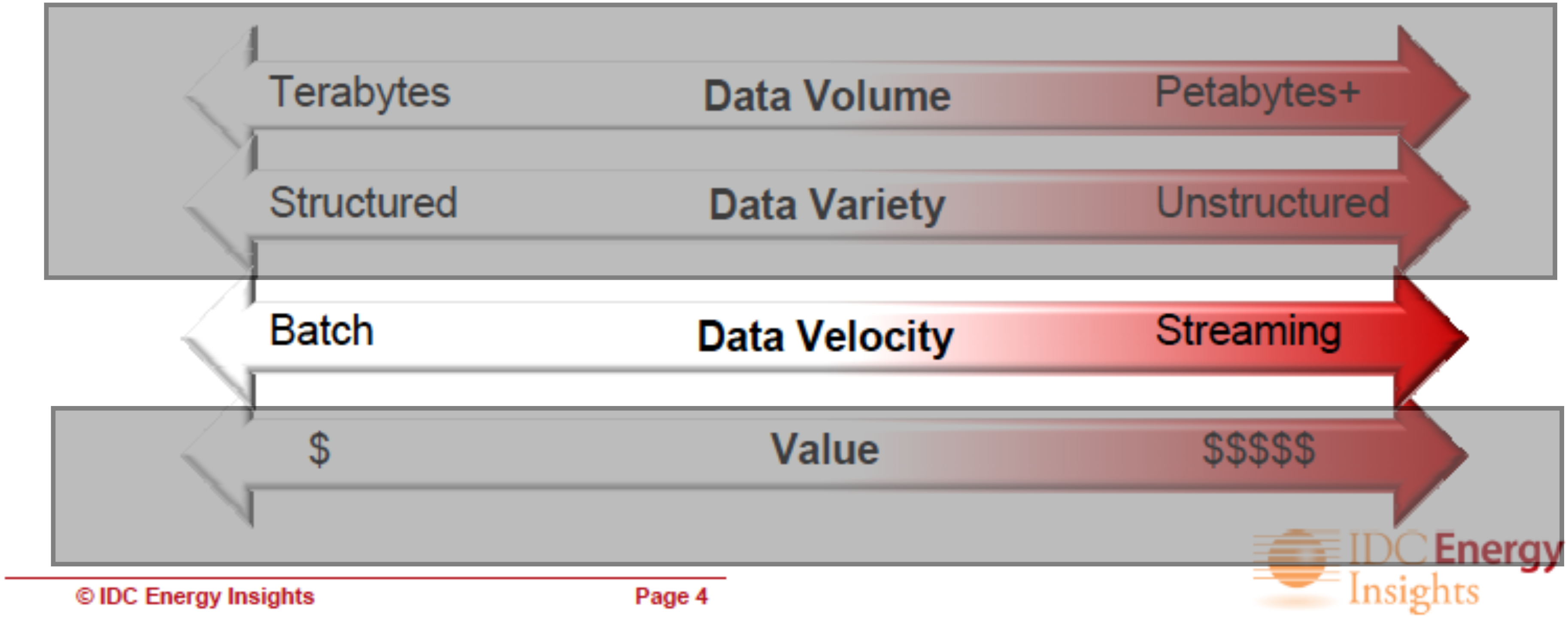
Source: Lehning, Michael et al, "Instrumenting the Earth: Next-Generation Sensor Networks and Environmental Science" in The Fourth Paradigm: Data-Intensive Science, ed. Tony Hey, http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf



Big Data: The Vs

High-throughput Analysis

Real-time predictive models

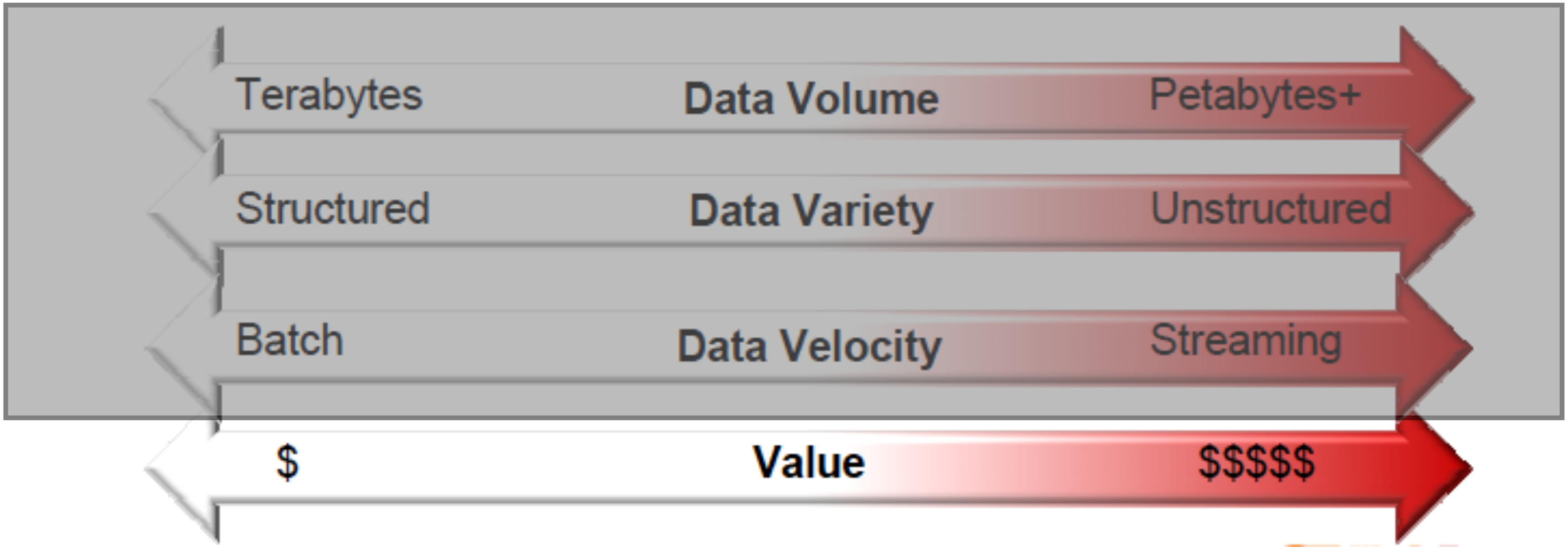


Big Data: The Vs

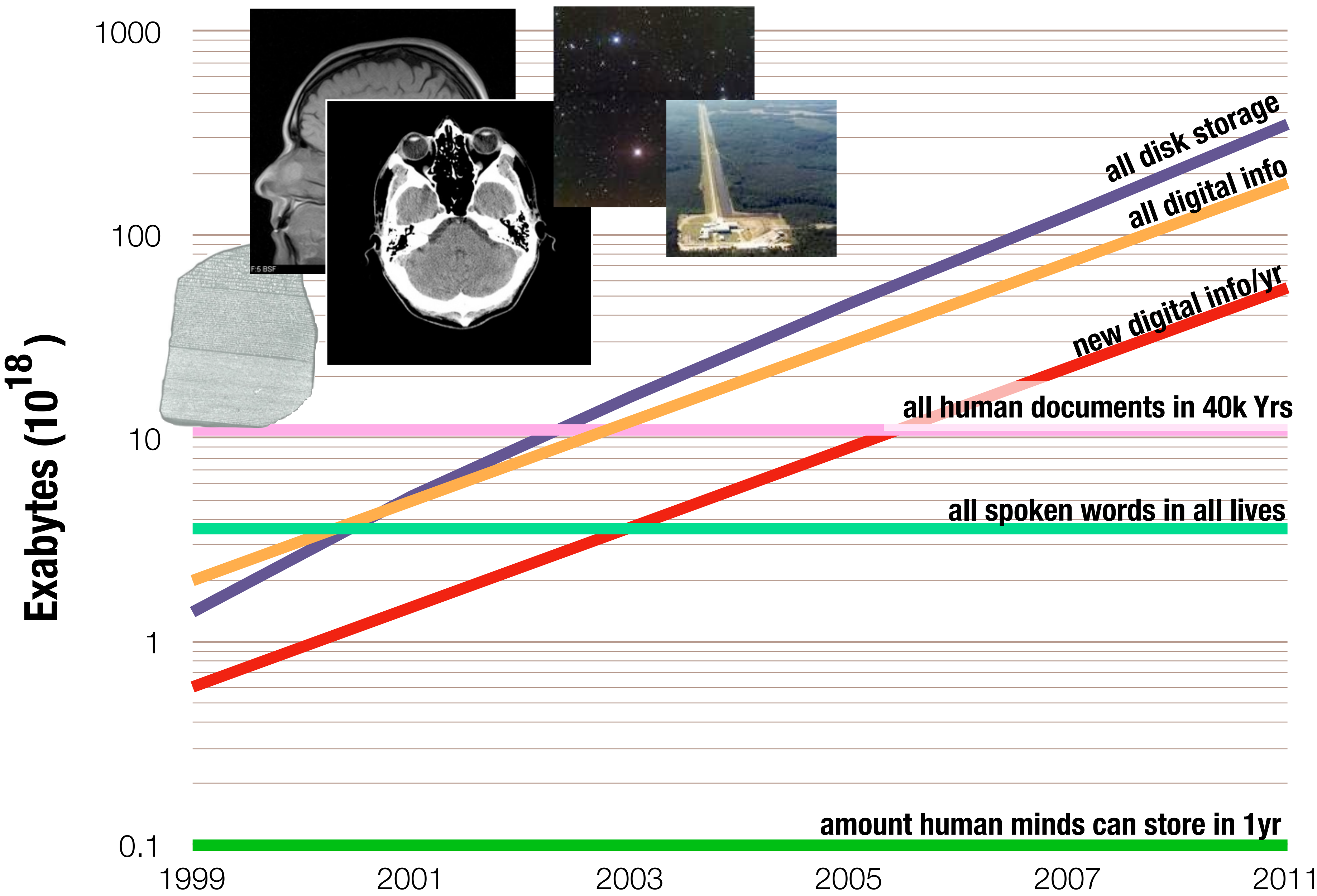
Climate Change

Clean Energy

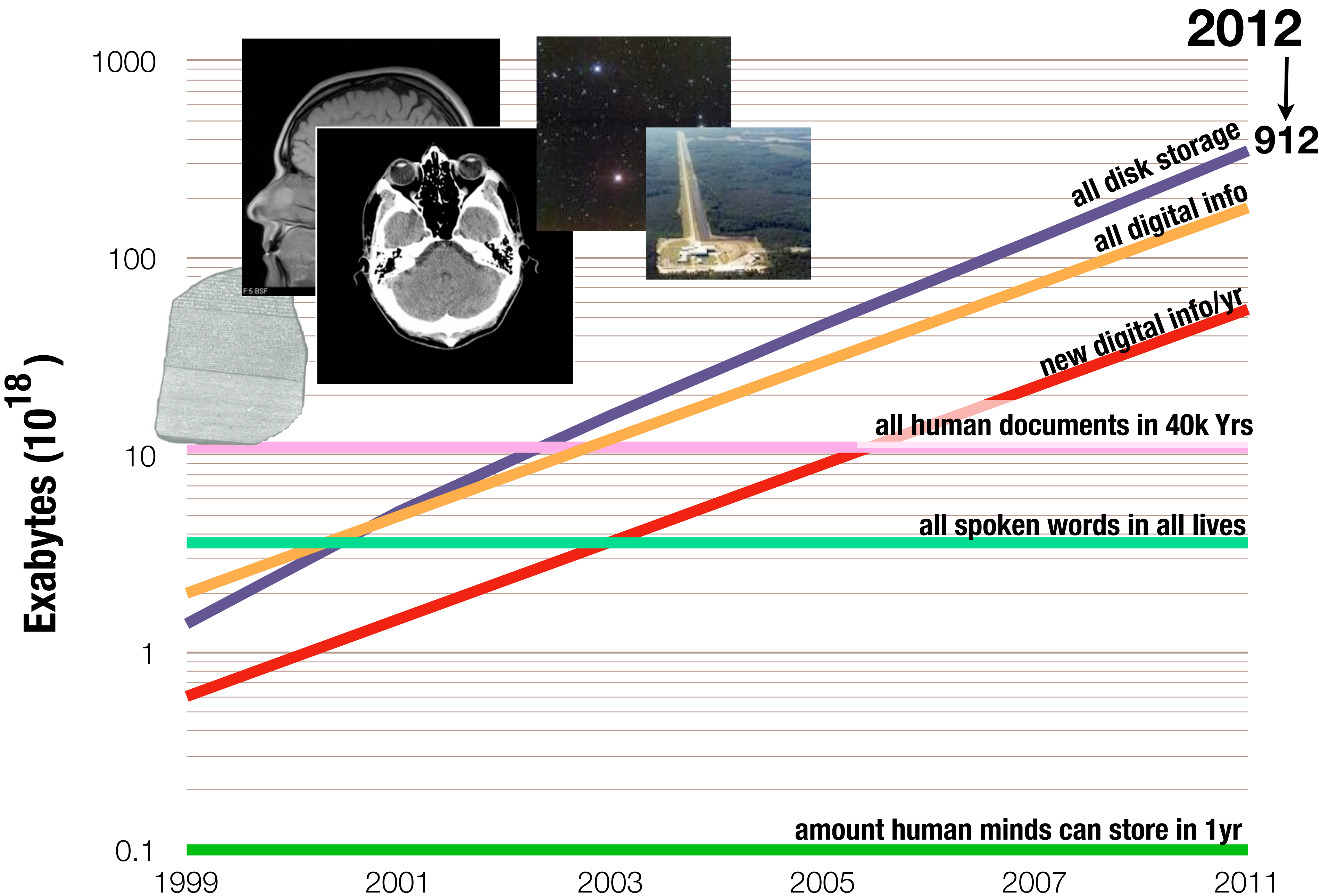
Disaster Prediction



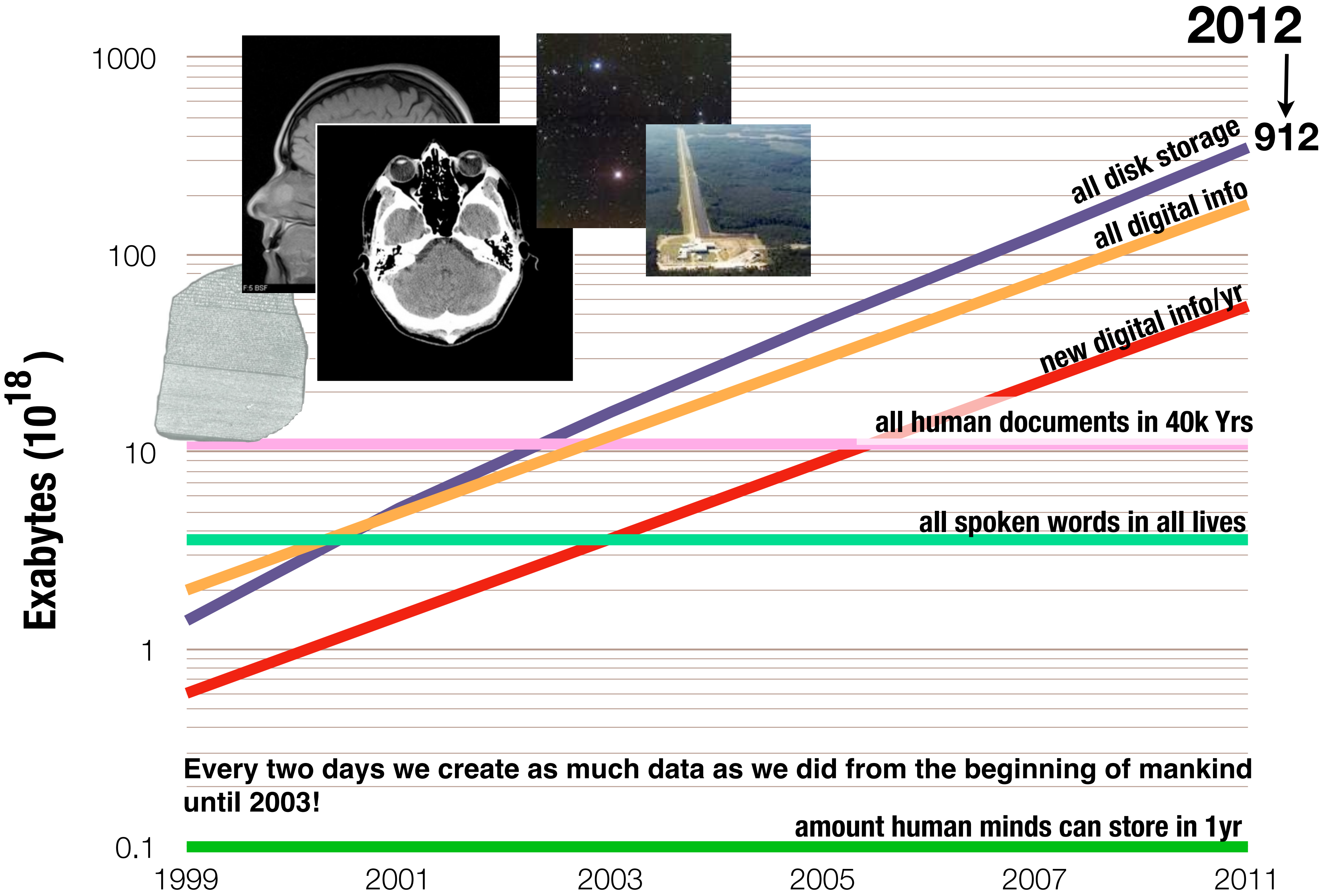
IDC Energy Insights



Sources: Lesk, Berkeley SIMS, Landauer, EMC, TechCrunch, Smart Planet



Sources: Lesk, Berkeley SIMS, Landauer, EMC, TechCrunch, Smart Planet



Sources: Lesk, Berkeley SIMS, Landauer, EMC, TechCrunch, Smart Planet

How Much is an Exabyte? (Volume)



How many trees does it take to print out an Exabyte?

How Much is an Exabyte? (Volume)

How many trees does it take to print out an Exabyte?

1 Exabyte = 1000 Petabytes = could hold approximately

How Much is an Exabyte? (Volume)



How many trees does it take to print out an Exabyte?

1 Exabyte = 1000 Petabytes = could hold approximately
500,000,000,000,000 pages of standard printed text

How Much is an Exabyte? (Volume)



How many trees does it take to print out an Exabyte?

1 Exabyte = 1000 Petabytes = could hold approximately 500,000,000,000,000 pages of standard printed text

It takes one tree to produce **94,200** pages of a book

How Much is an Exabyte? (Volume)

How many trees does it take to print out an Exabyte?

1 Exabyte = 1000 Petabytes = could hold approximately 500,000,000,000,000 pages of standard printed text

It takes one tree to produce **94,200** pages of a book

Thus it will take **530,785,562,327** trees to store an Exabyte of data

How Much is an Exabyte? (Volume)



How many trees does it take to print out an Exabyte?

1 Exabyte = 1000 Petabytes = could hold approximately 500,000,000,000,000 pages of standard printed text

It takes one tree to produce **94,200** pages of a book

Thus it will take **530,785,562,327** trees to store an Exabyte of data

In 2005, there were **400,246,300,201** trees on Earth

How Much is an Exabyte? (Volume)

How many trees does it take to print out an Exabyte?

1 Exabyte = 1000 Petabytes = could hold approximately 500,000,000,000,000 pages of standard printed text

It takes one tree to produce **94,200** pages of a book

Thus it will take **530,785,562,327** trees to store an Exabyte of data

In 2005, there were **400,246,300,201** trees on Earth

We can store **.75** Exabytes of data using all the trees on the entire planet.

How Much is an Exabyte? (Volume)



How many trees does it take to print out an Exabyte?

1 Exabyte = 1000 Petabytes = could hold approximately 500,000,000,000,000 pages of standard printed text

It takes one tree to produce **94,200** pages of a book

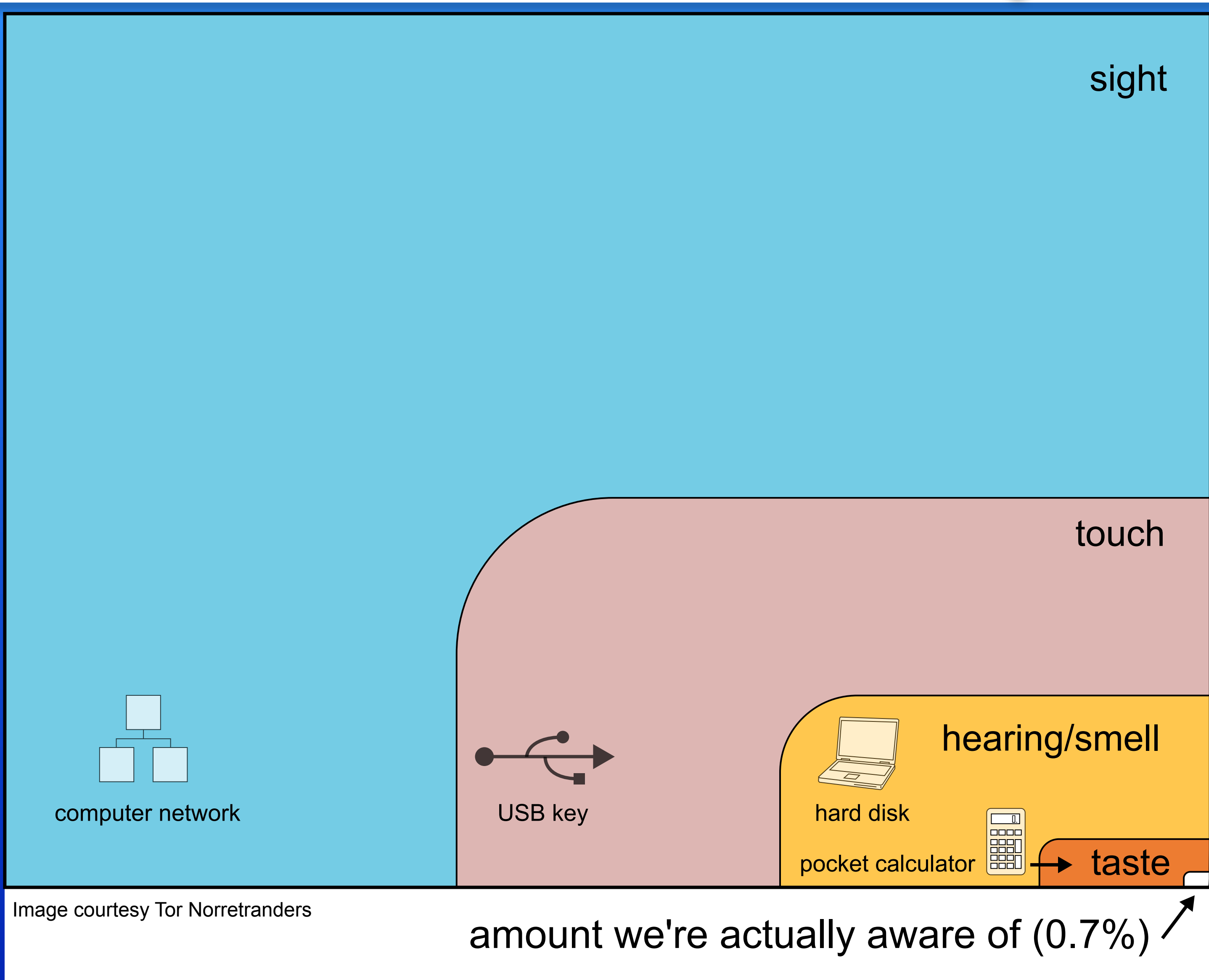
Thus it will take **530,785,562,327** trees to store an Exabyte of data

In 2005, there were **400,246,300,201** trees on Earth

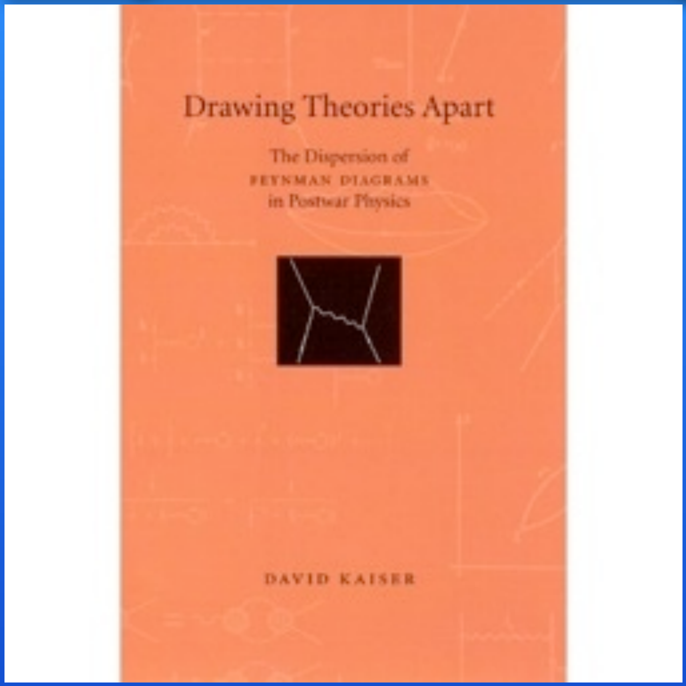
We can store **.75** Exabytes of data using all the trees on the entire planet.

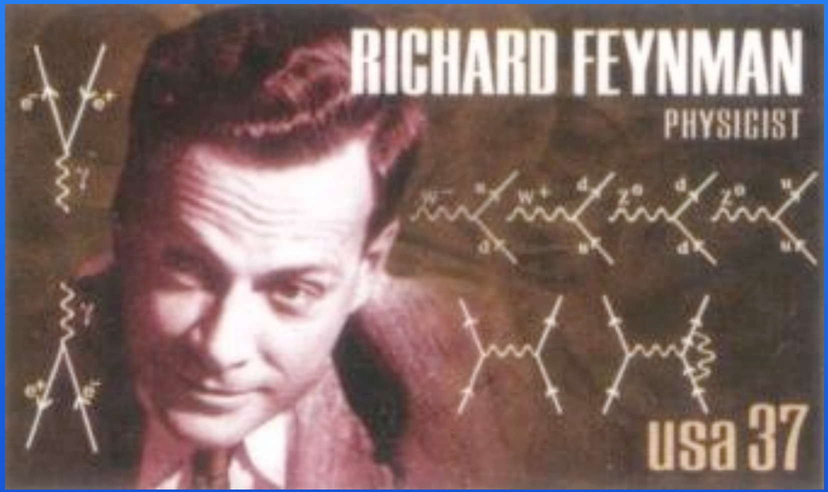
Sources: http://www.whatsabyte.com/ and http://wiki.answers.com

Brain Information Bandwidth (Velocity)



Feynman Diagrams (Data Analytics)

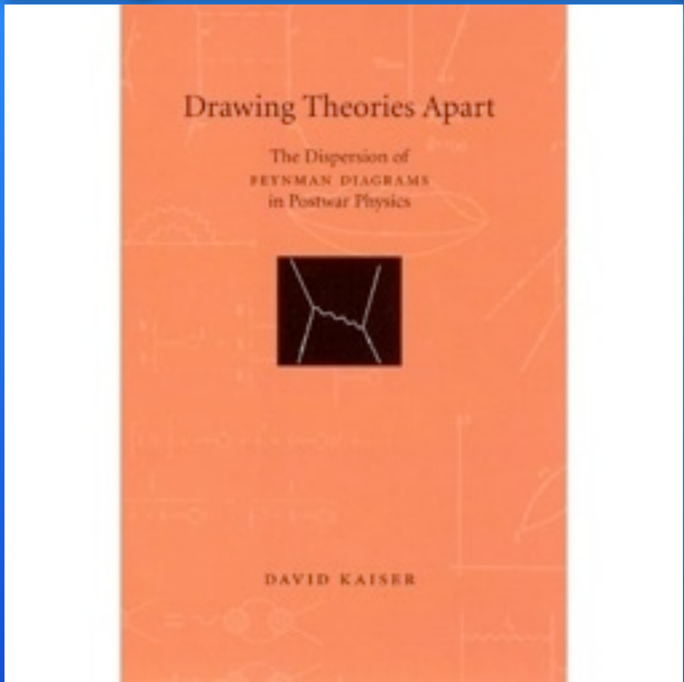


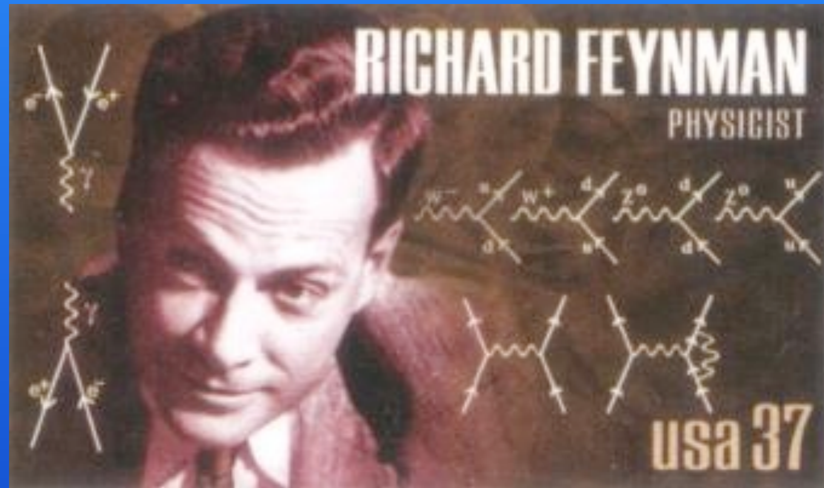


Feynman: "What I am really try to do is bring birth to clarity, which is really a half-assedly thought-out-pictorial semi-vision thing. I would see the jiggle-jiggle-jiggle or the wiggle of the path. Even now when I talk about the influence functional, I see the coupling and I take this turn - like as if there was a big bag of stuff - and try to collect it in away and to push it. It's all visual. It's hard to explain."

James Gleick, The Life and Science of Richard Feynman, Vintage Books, New York, 1992.

Feynman Diagrams

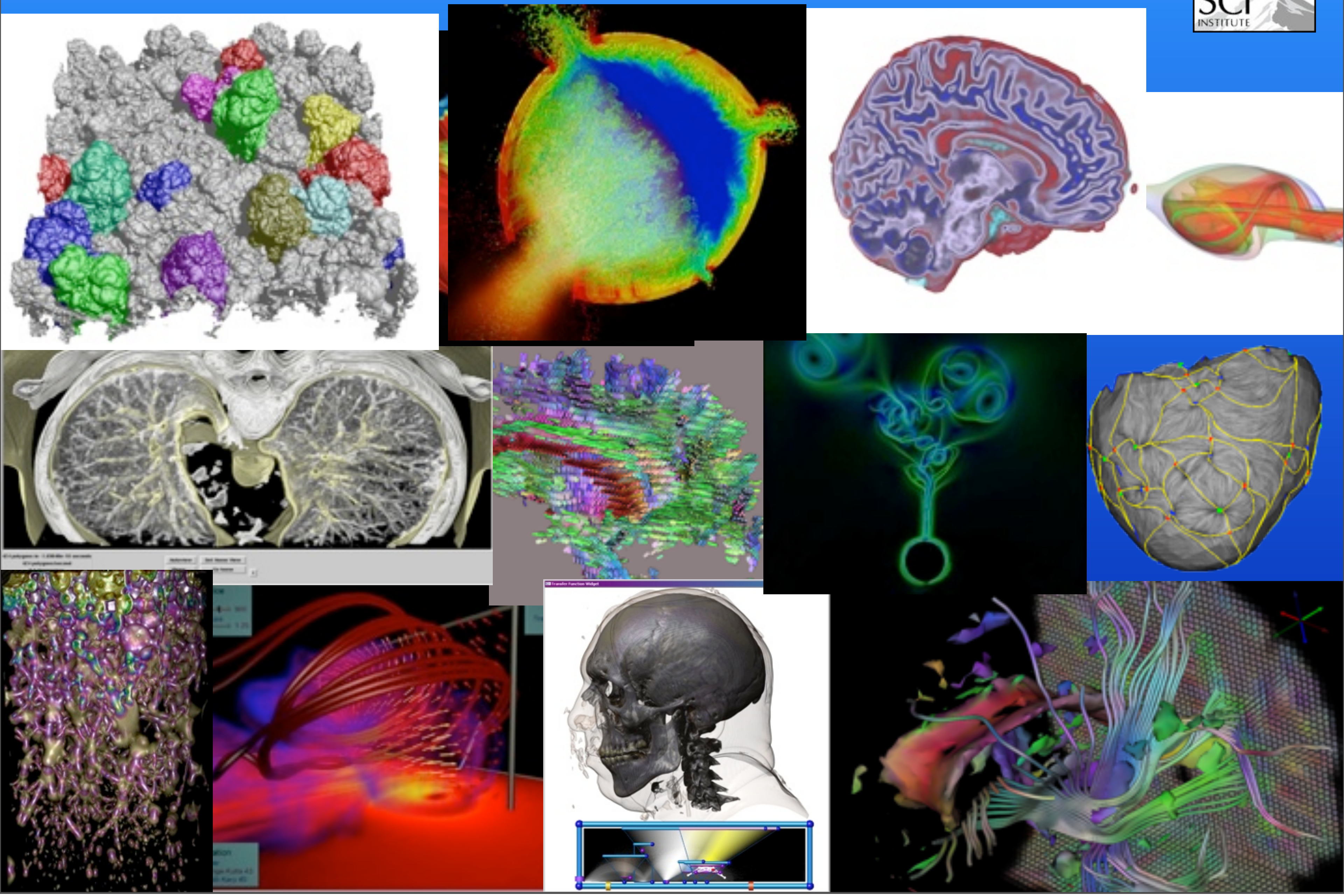




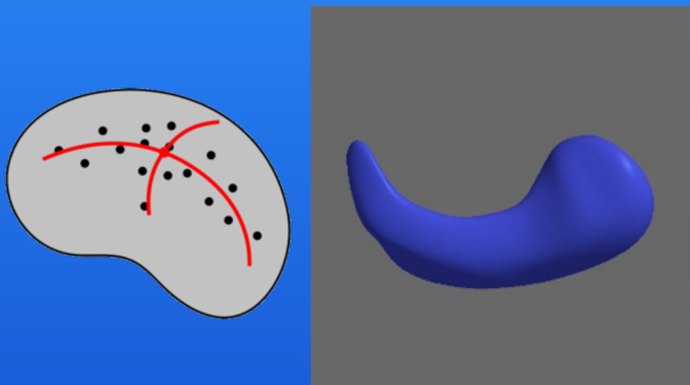
Feynman: "In certain particular problems that I have done it was necessary to continue the development of the picture as the method before the mathematics could be really done."

James Gleick, The Life and Science of Richard Feynman, Vintage Books, New York, 1992.

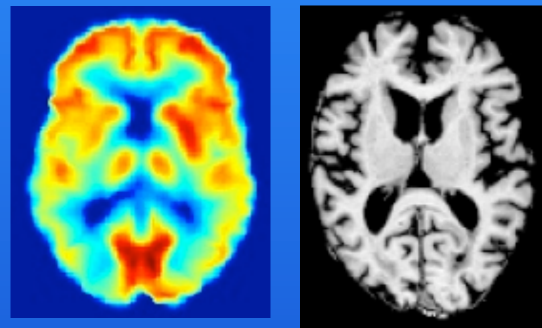
New Visual Analysis Techniques



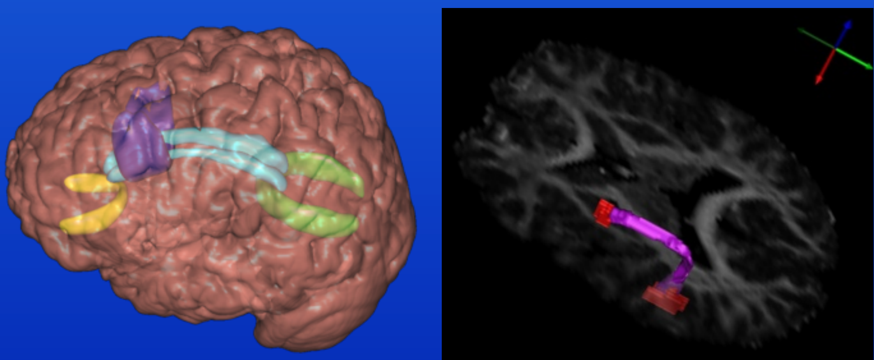
Big Data in Imaging Research



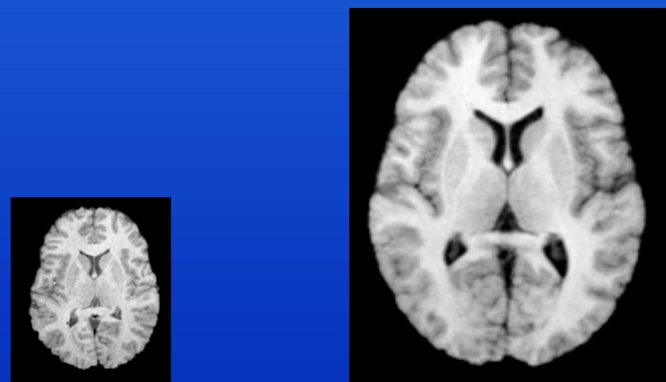
Computational Statistics
in Nonlinear Spaces



Combined PET + MRI analysis
Alzheimer's disease project

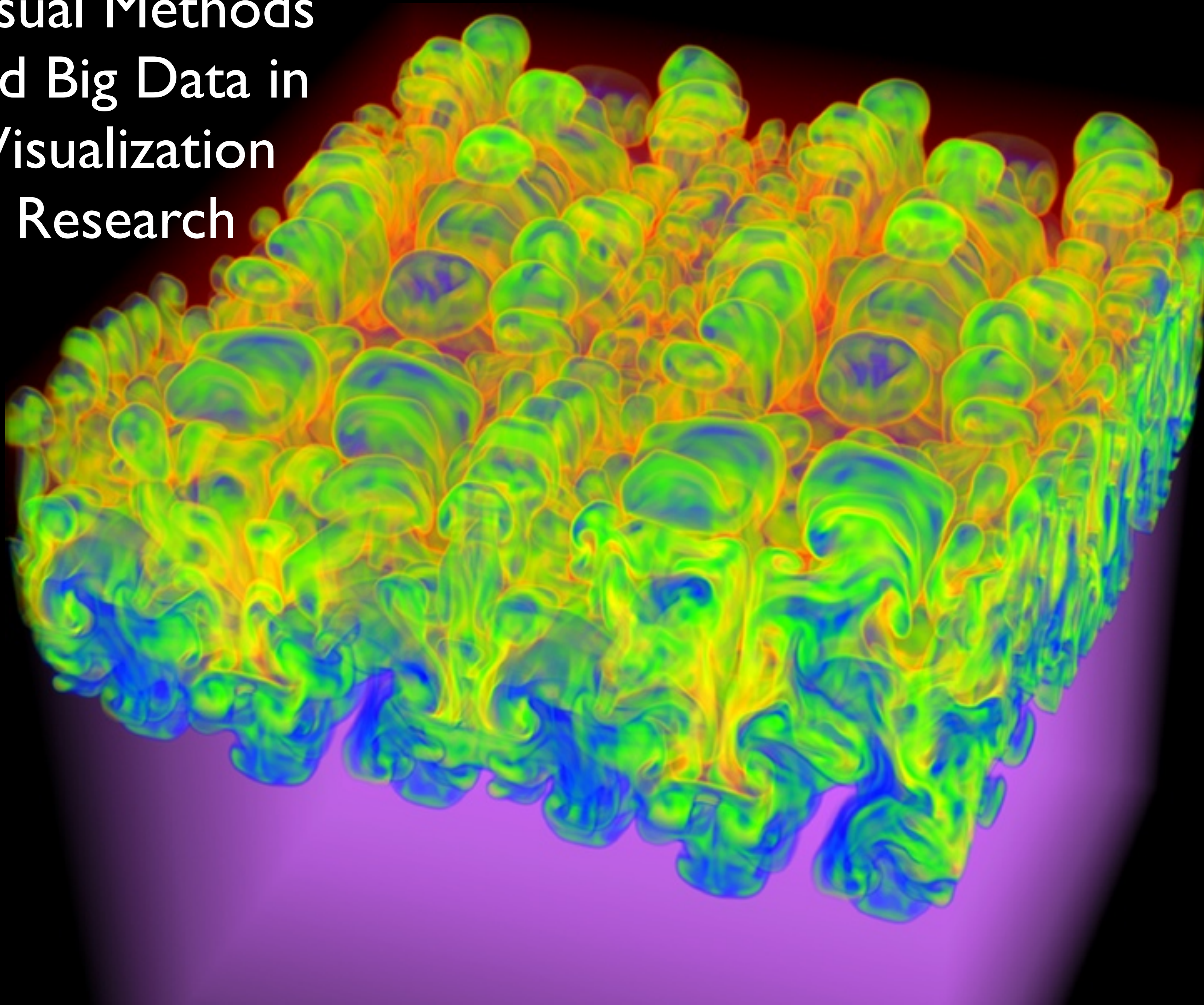


Diffusion Tensor Image Analysis
Autism project



Anatomical shape averaging
and variability

Visual Methods and Big Data in Visualization Research



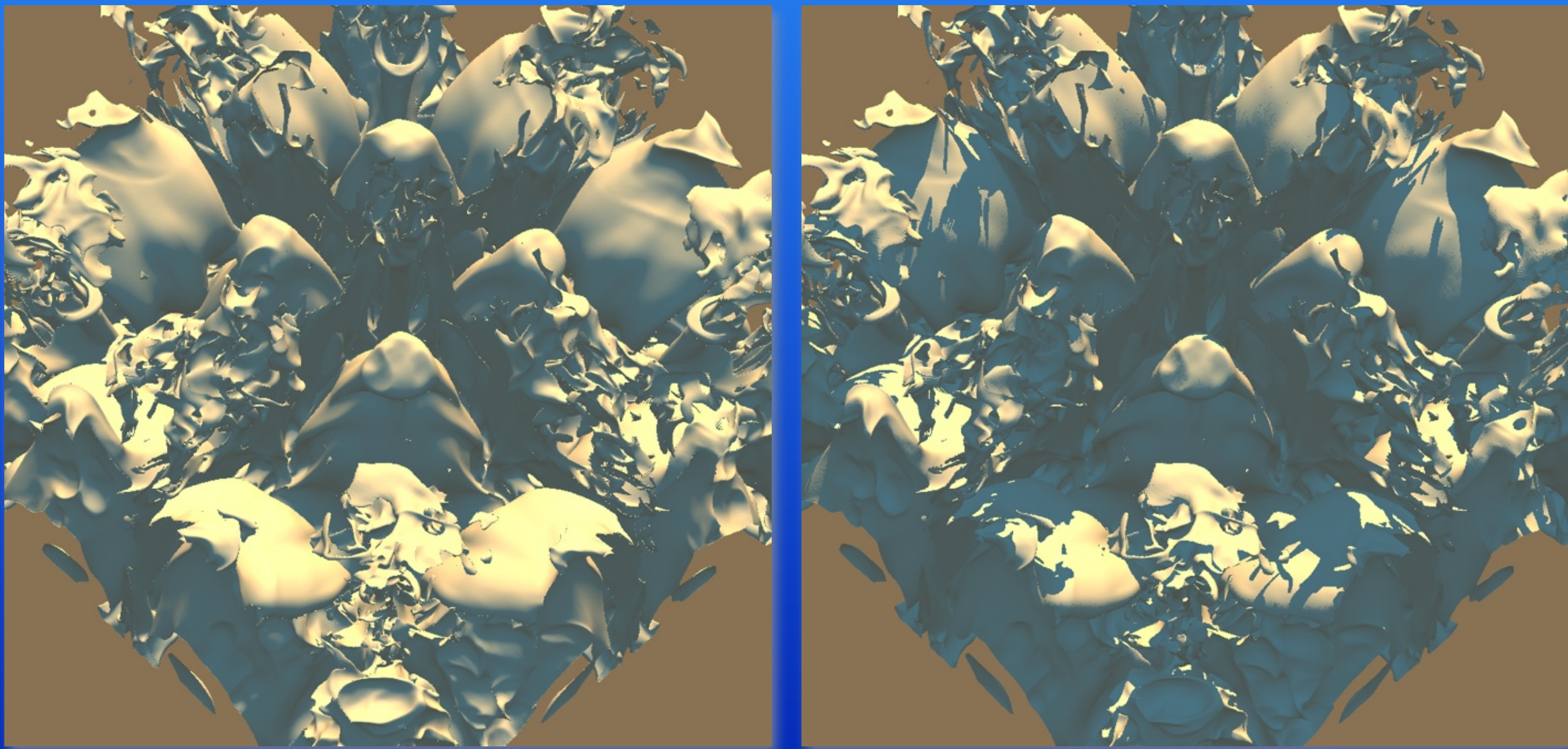
Perceptual Cues for Shading

Jim Blinn:

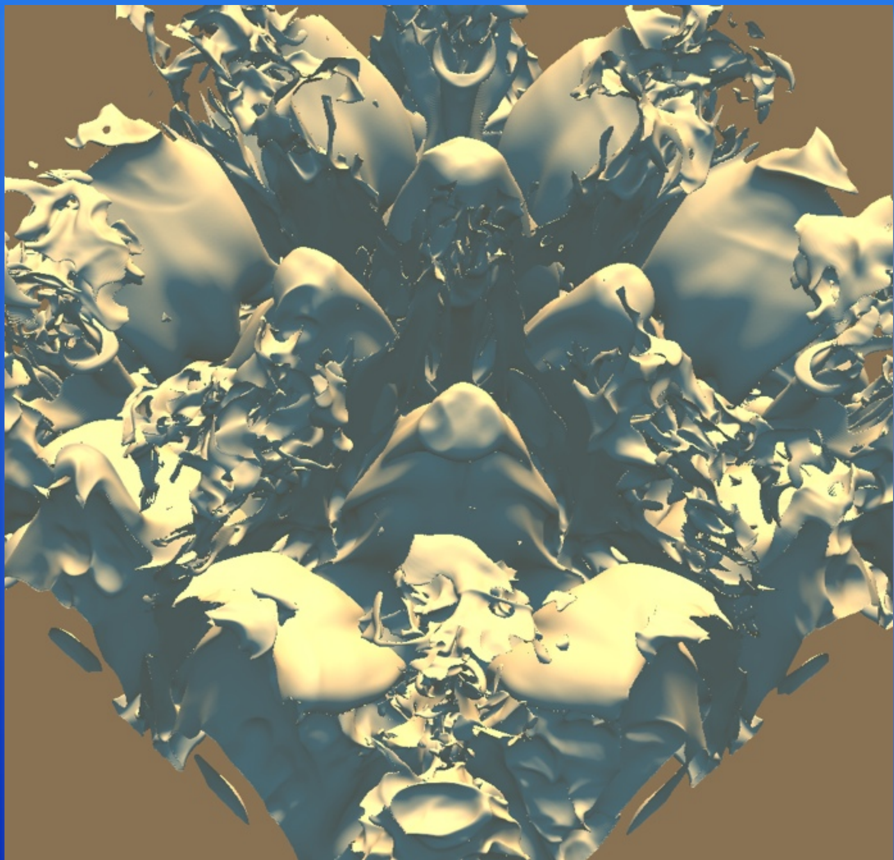
"Lighting models... there's something that always bothered me about lighting models. Bui Tuong Phong is[was] a great guy and he did wonderful work …
The thing is, this has no physical basis whatsoever ...
I'd like to see cosine power retired and better approximations being done."

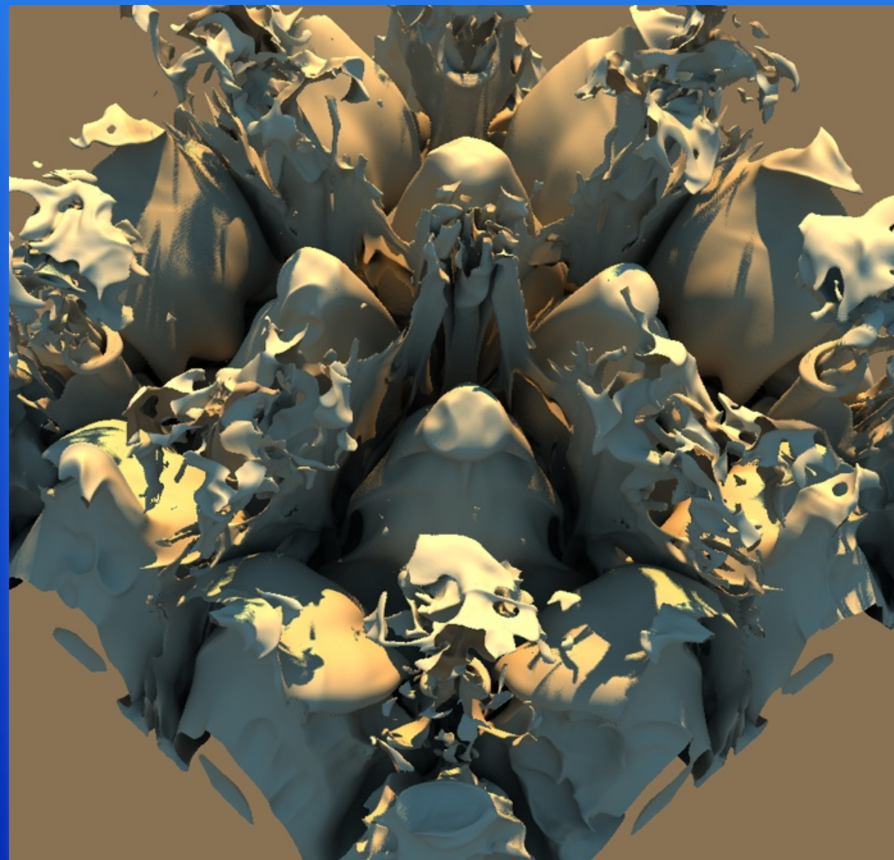
- SIGGRAPH 98 Keynote Speech

Perception - Shadows

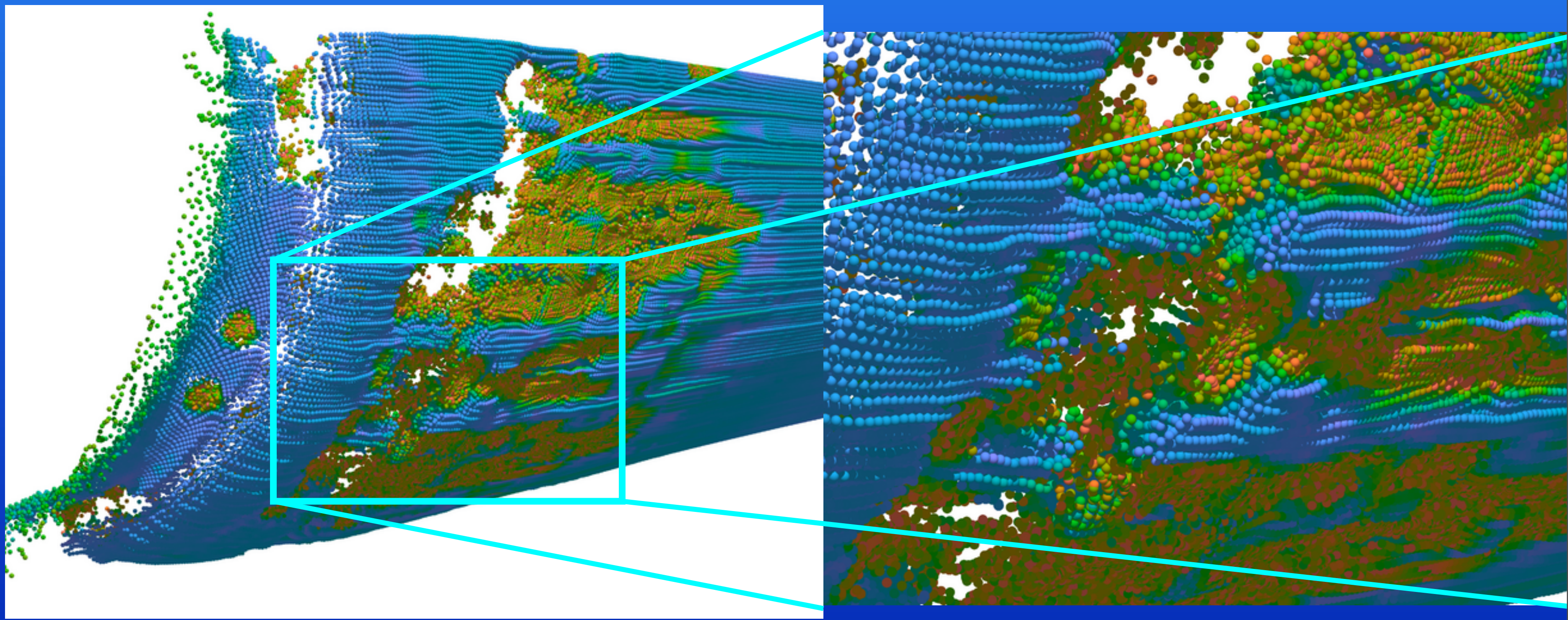


Perception - Shadows

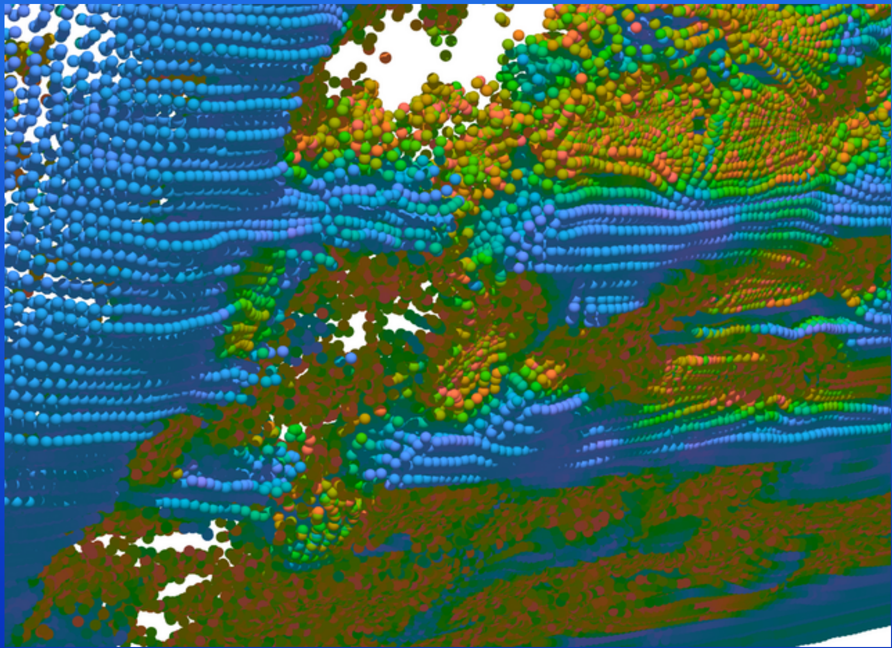


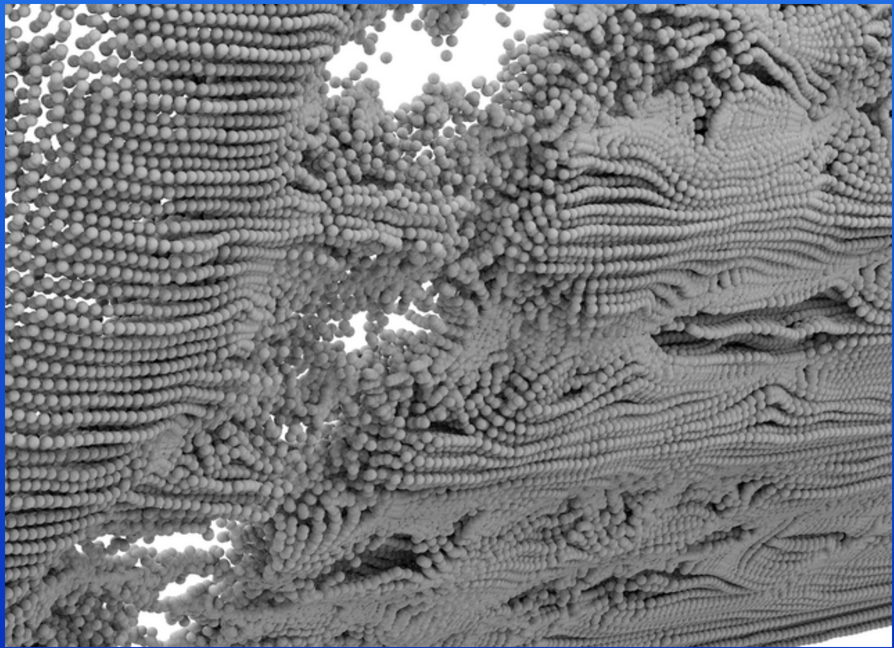


Indirect Shading of Particles

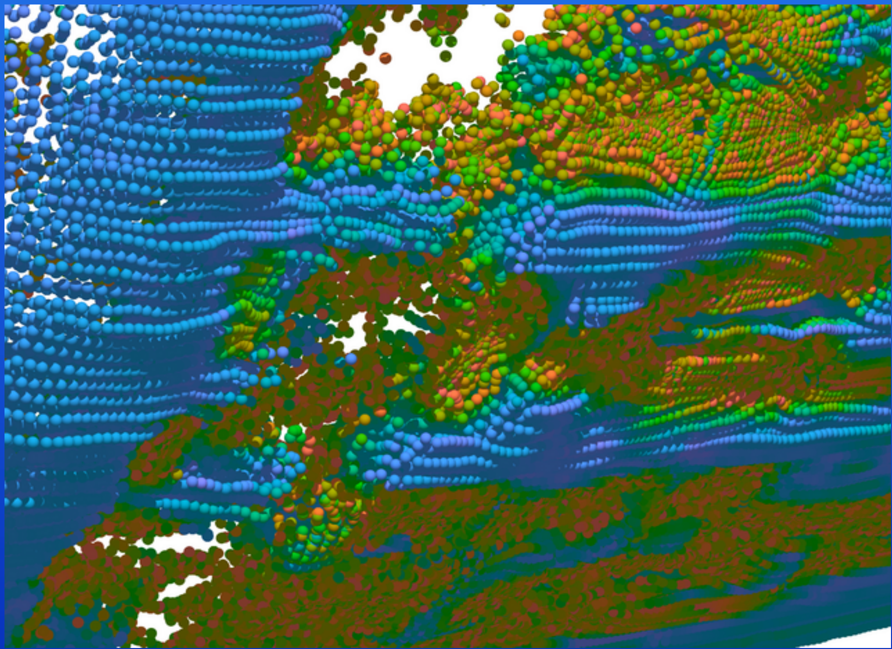


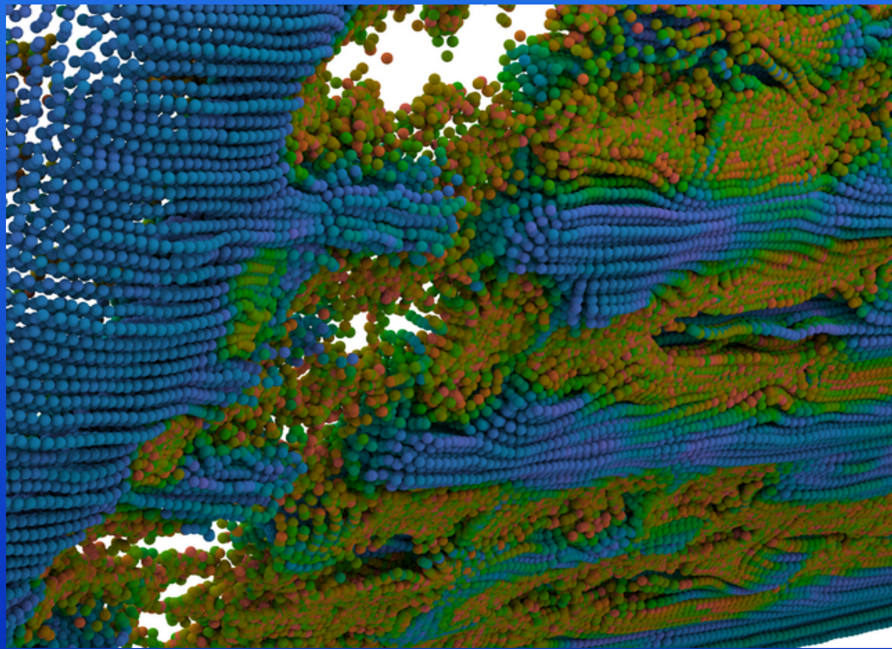
Indirect Shading of Particles



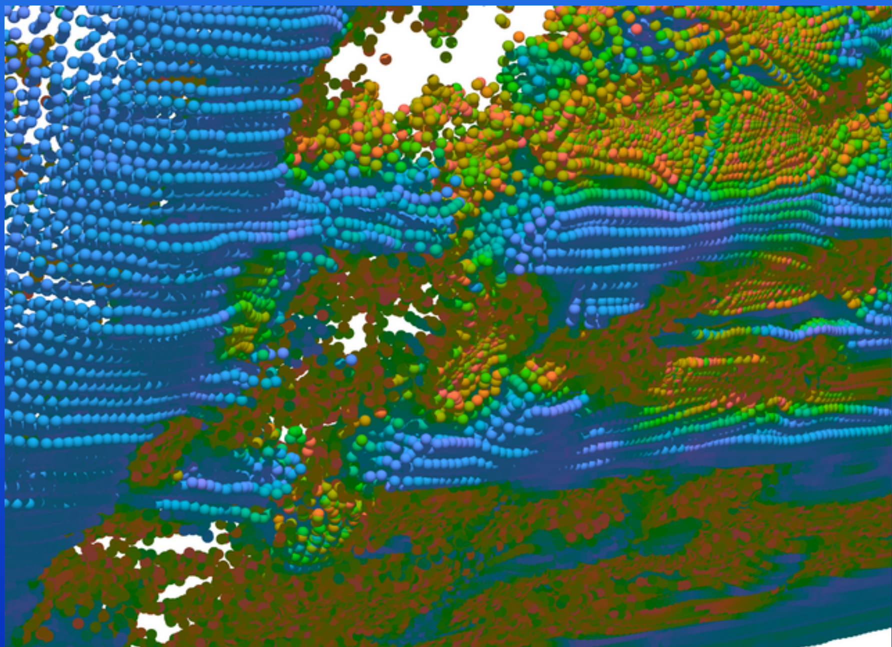


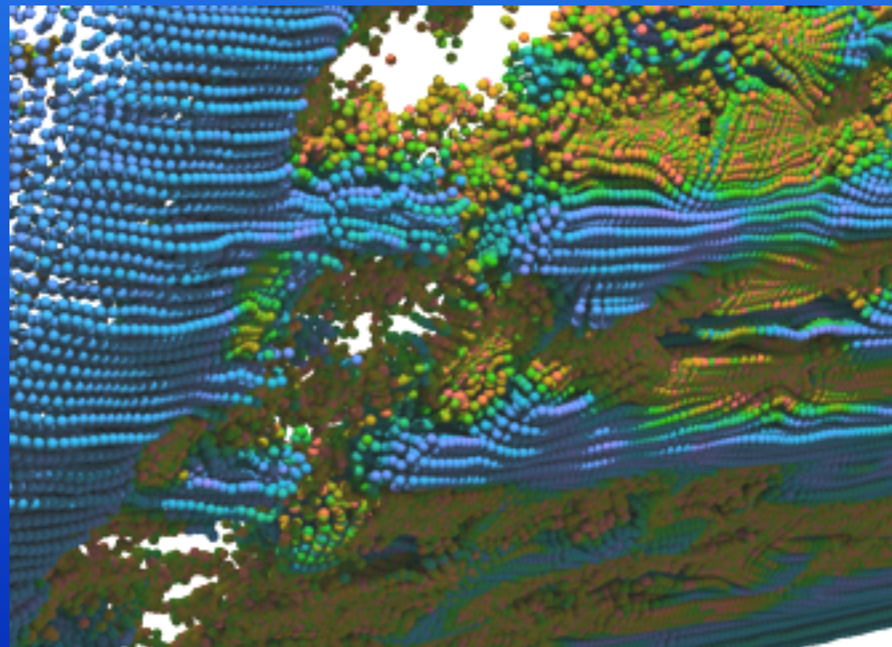
Indirect Shading of Particles



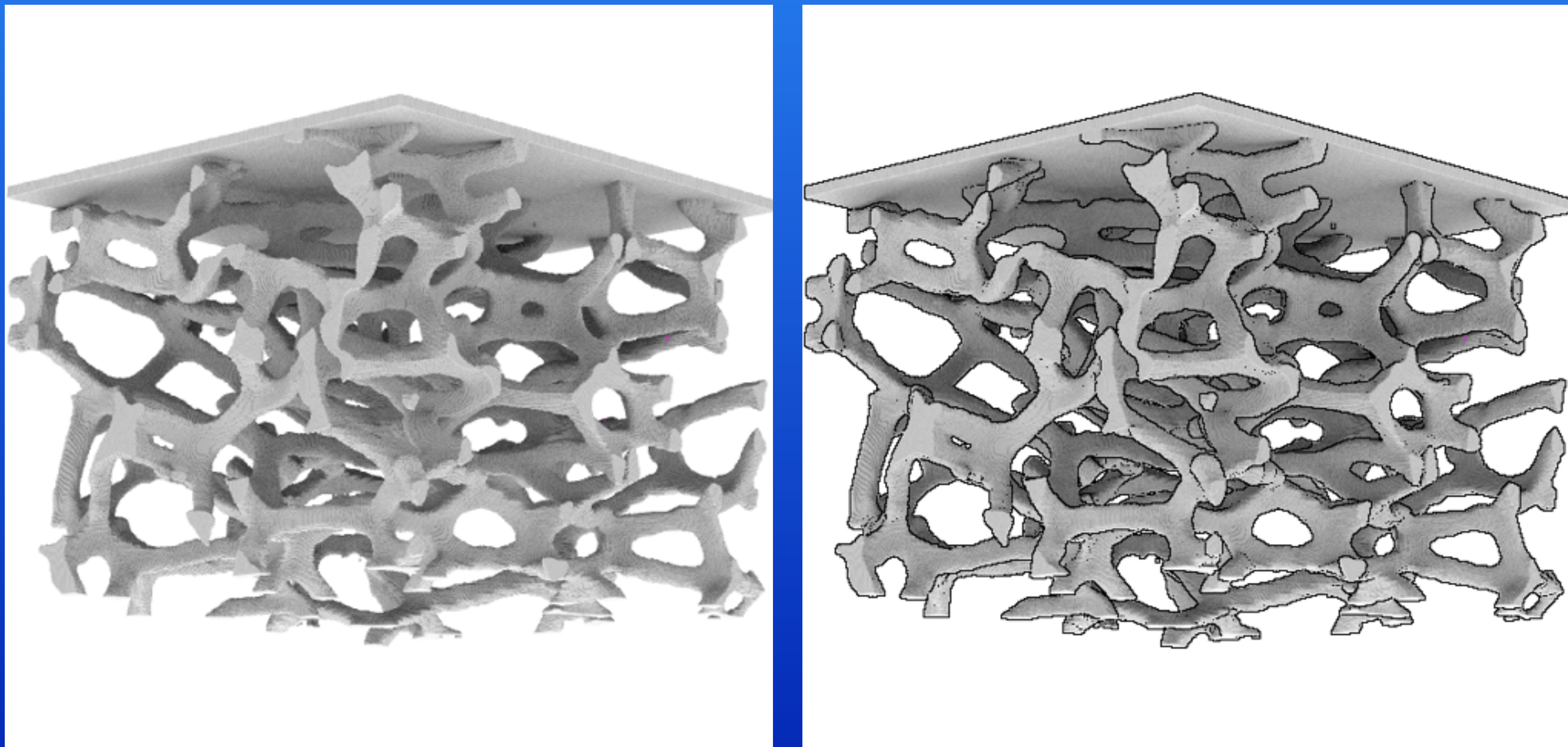


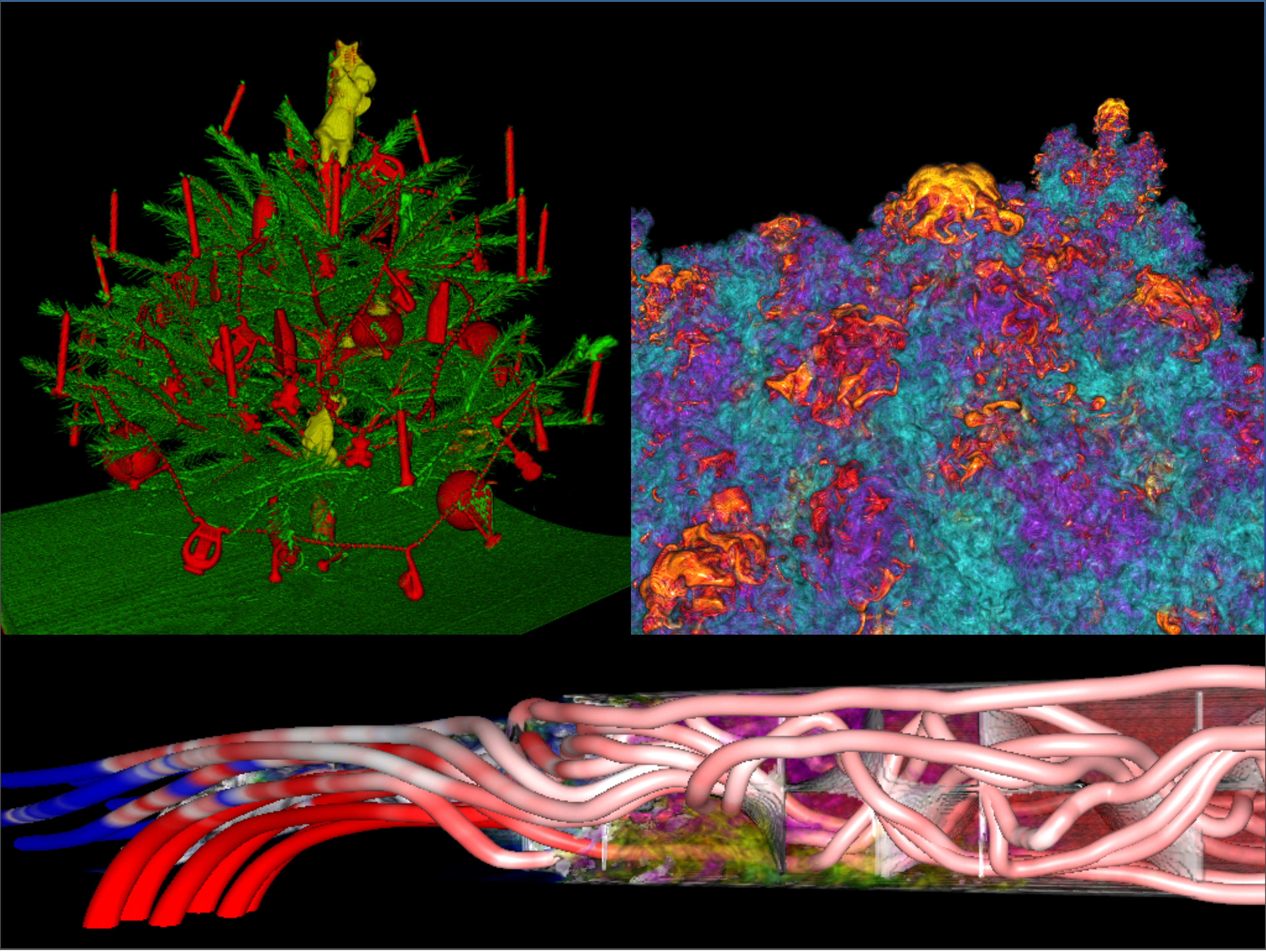
Indirect Shading of Particles

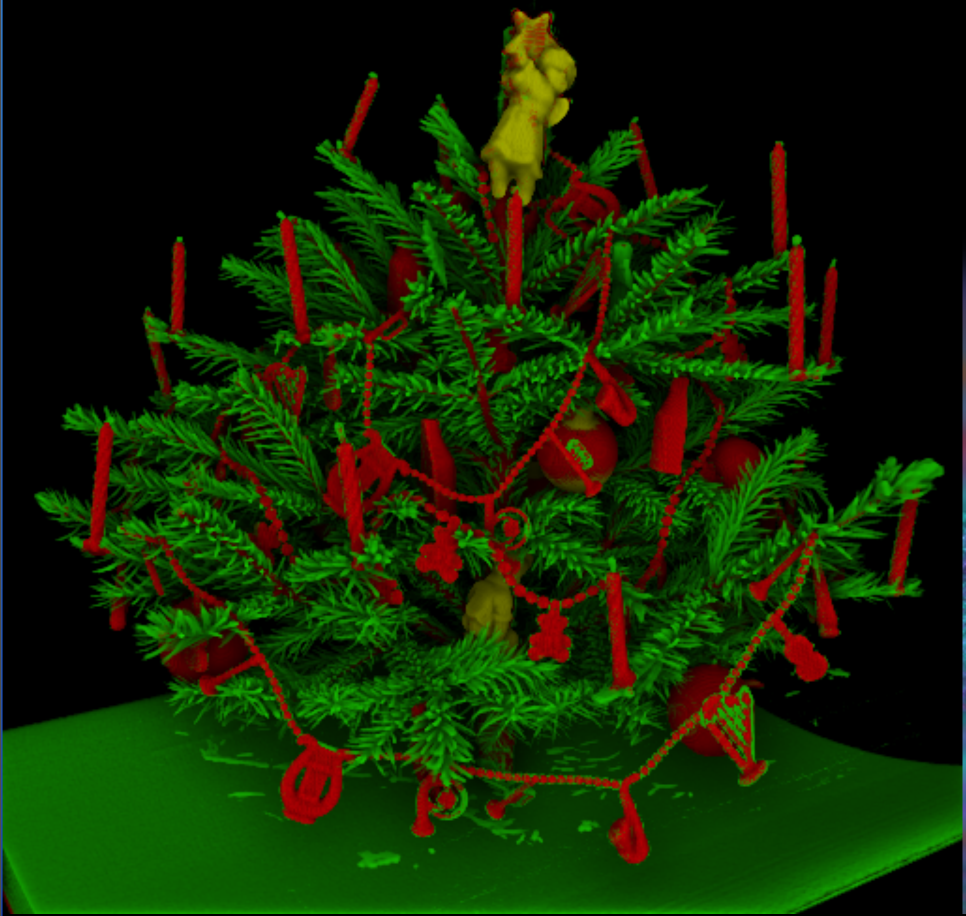
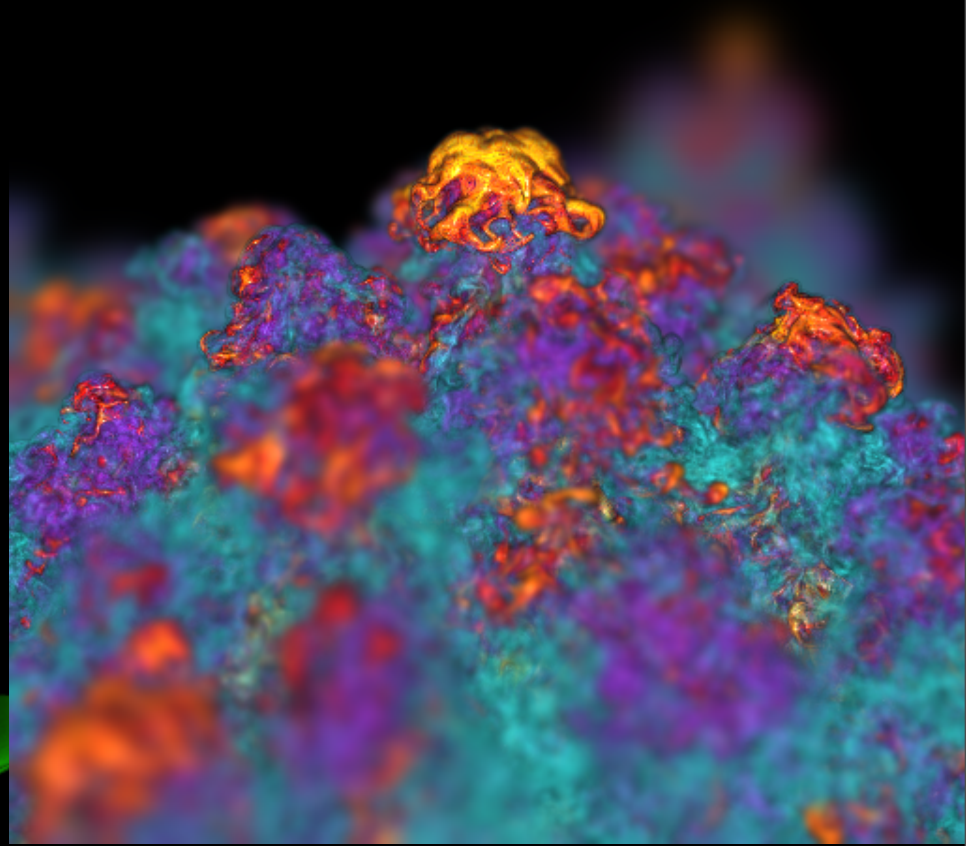
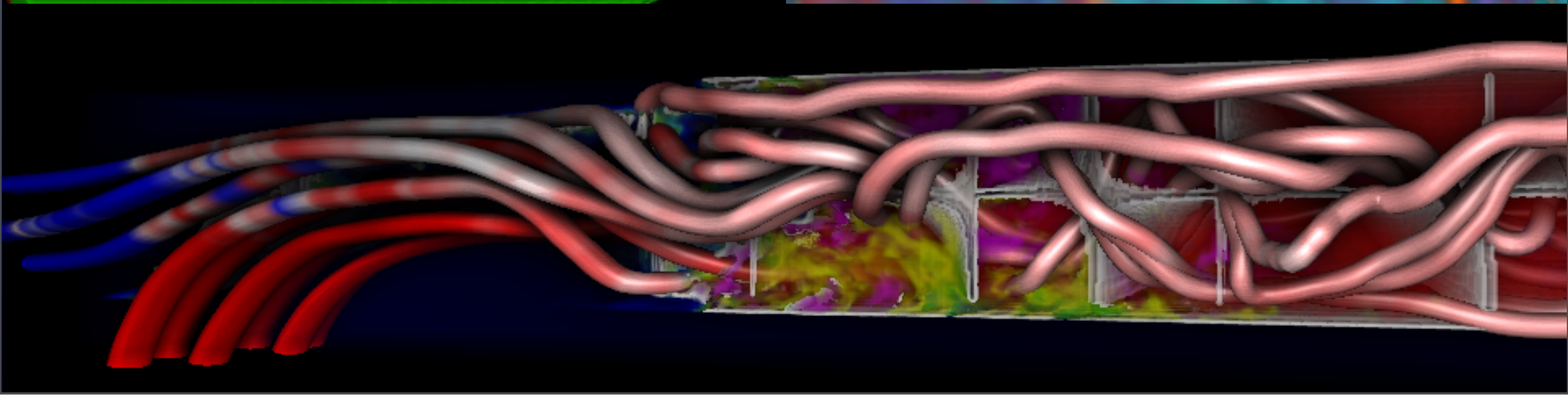




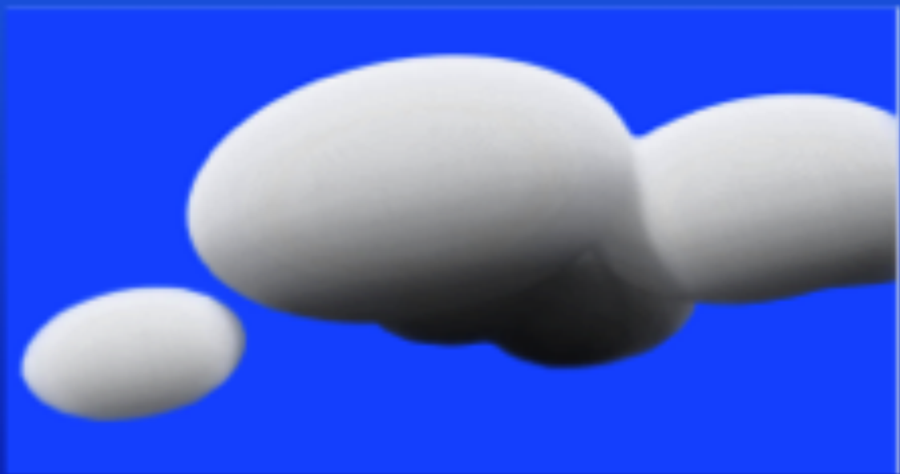
Silhouettes





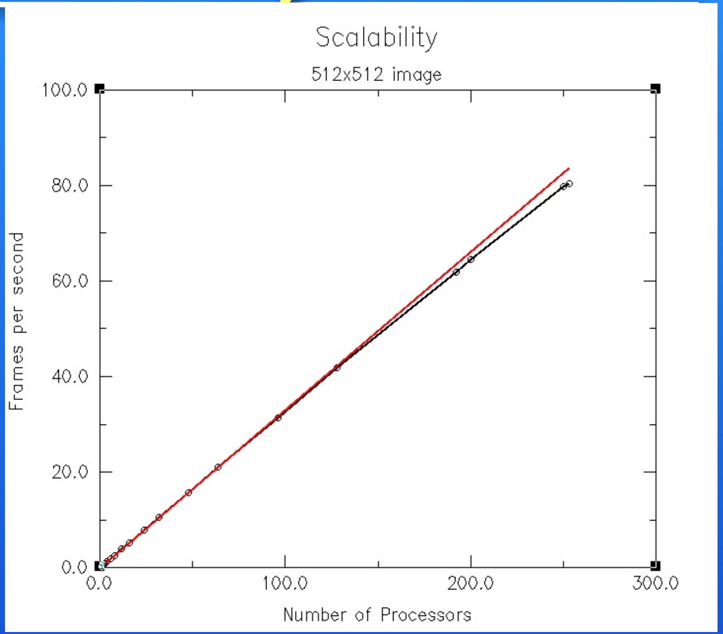




Volumetric Modeling

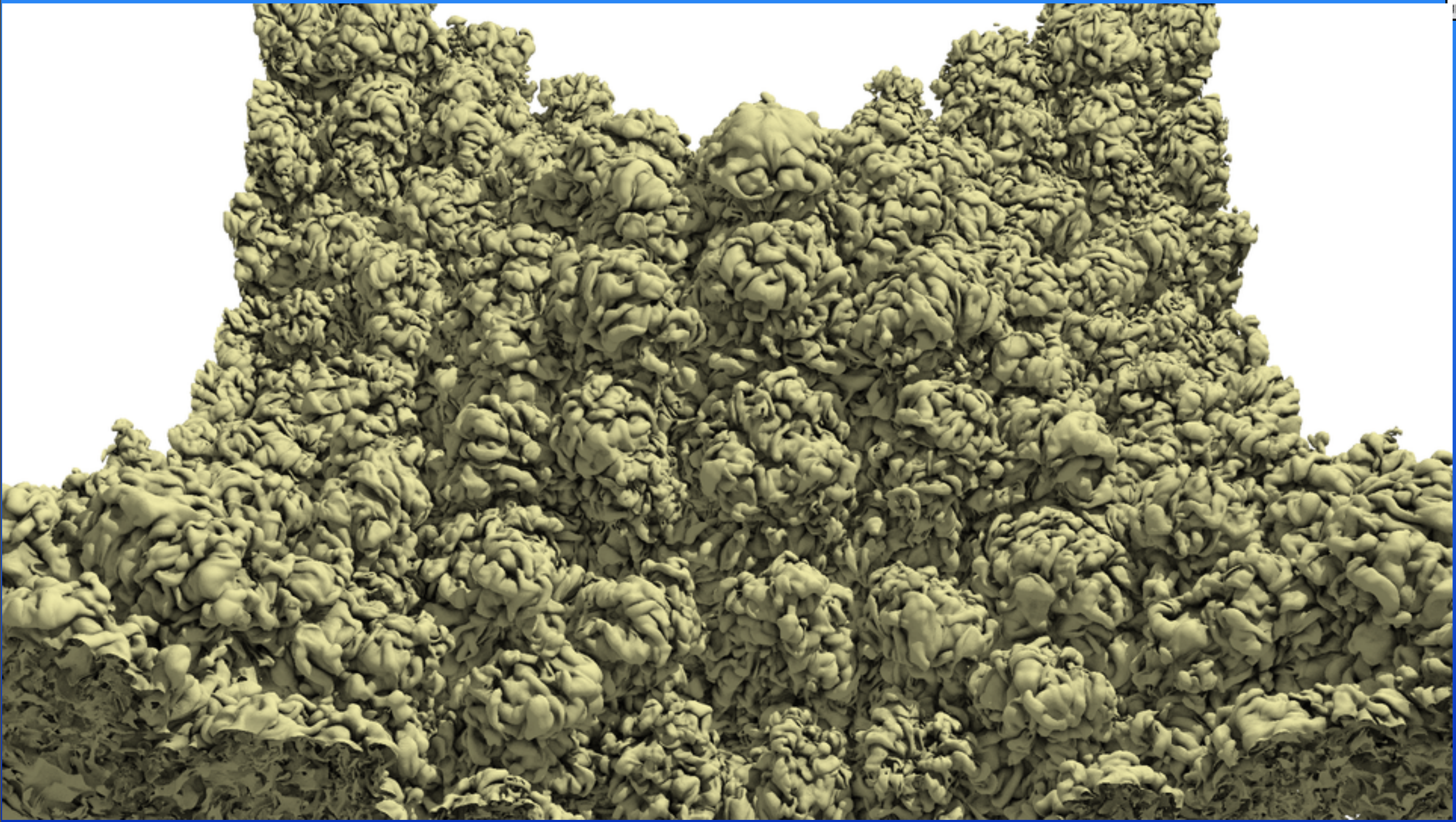


Real-Time Ray Tracer (RTRT)

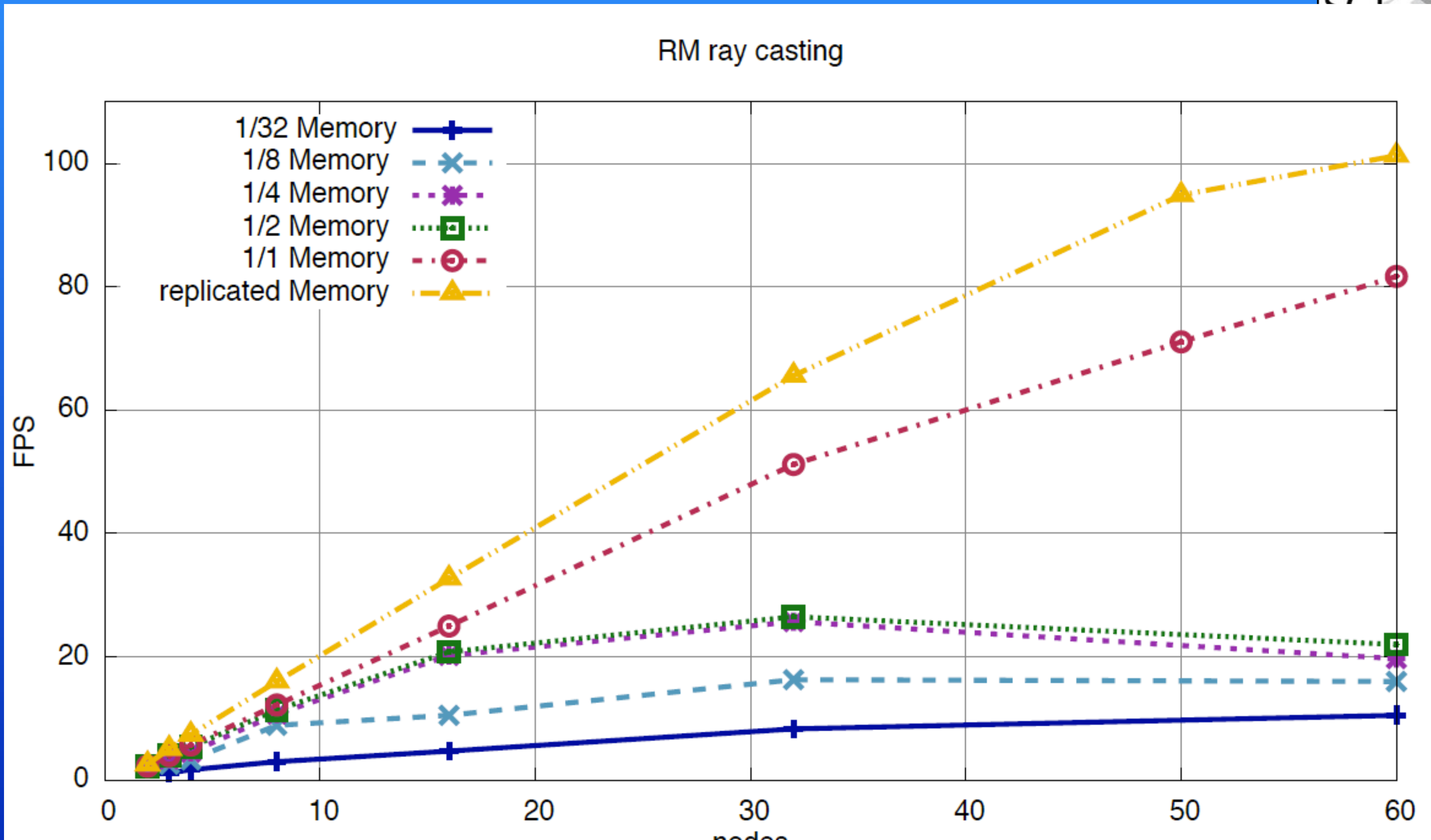
- Implemented first on an SGI architecture - up to 1024 processors, then a distributed memory version for clusters, now on other SMD machines
- Approximately linear speedup
- Load balancing and memory coherence are key to performance



Current Cluster RT Visualization



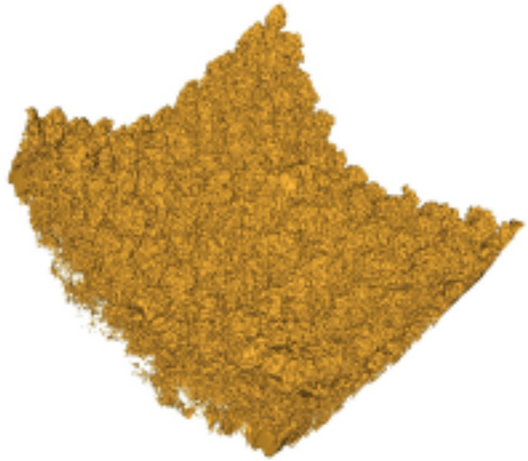
Current Cluster RT Visualization



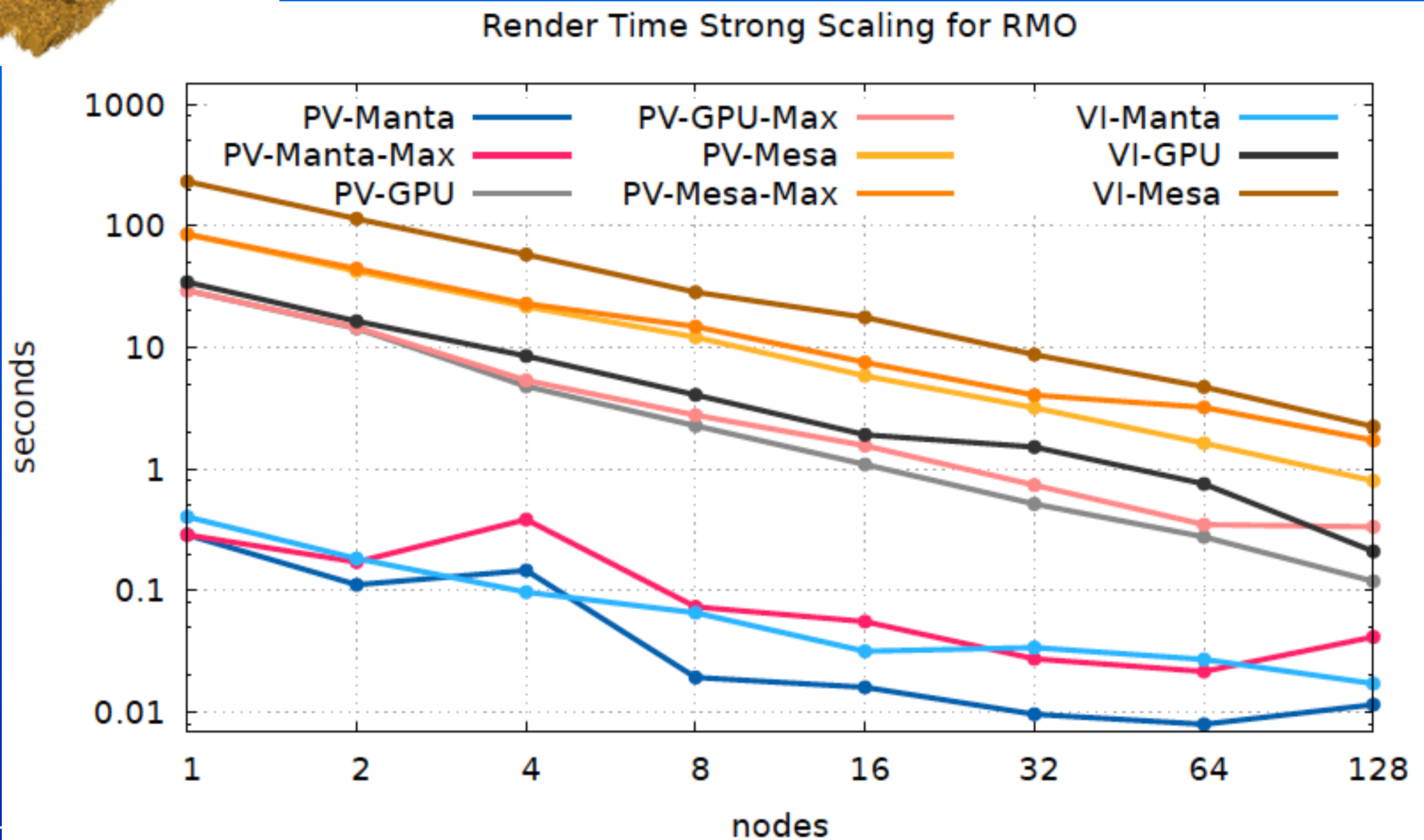
VTK integration

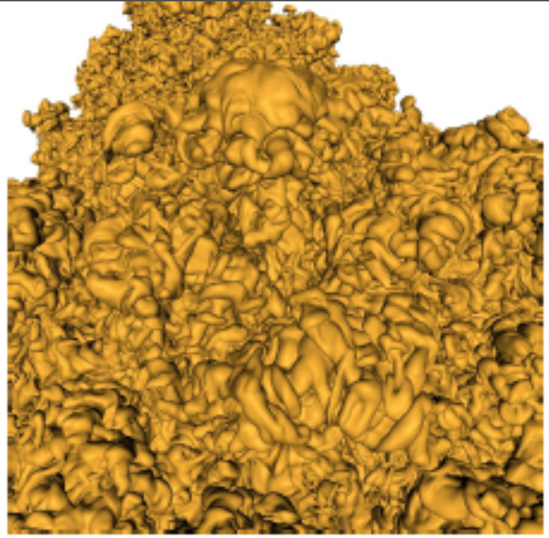
VTK integration gives a common framework for running through two common visualization packages

- ParaView
- VisIt

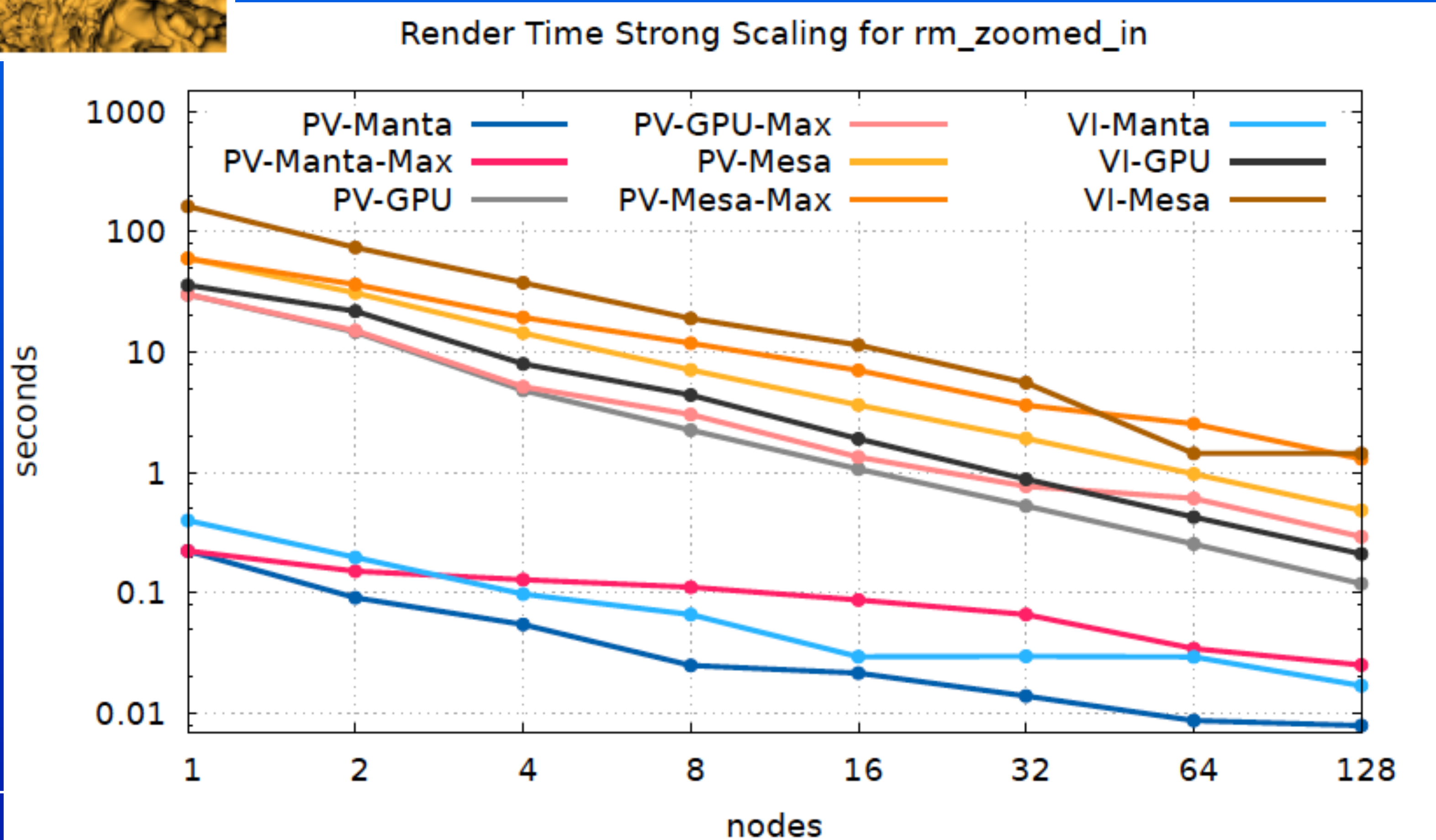


Ray-tracing over-riding VTK renderer



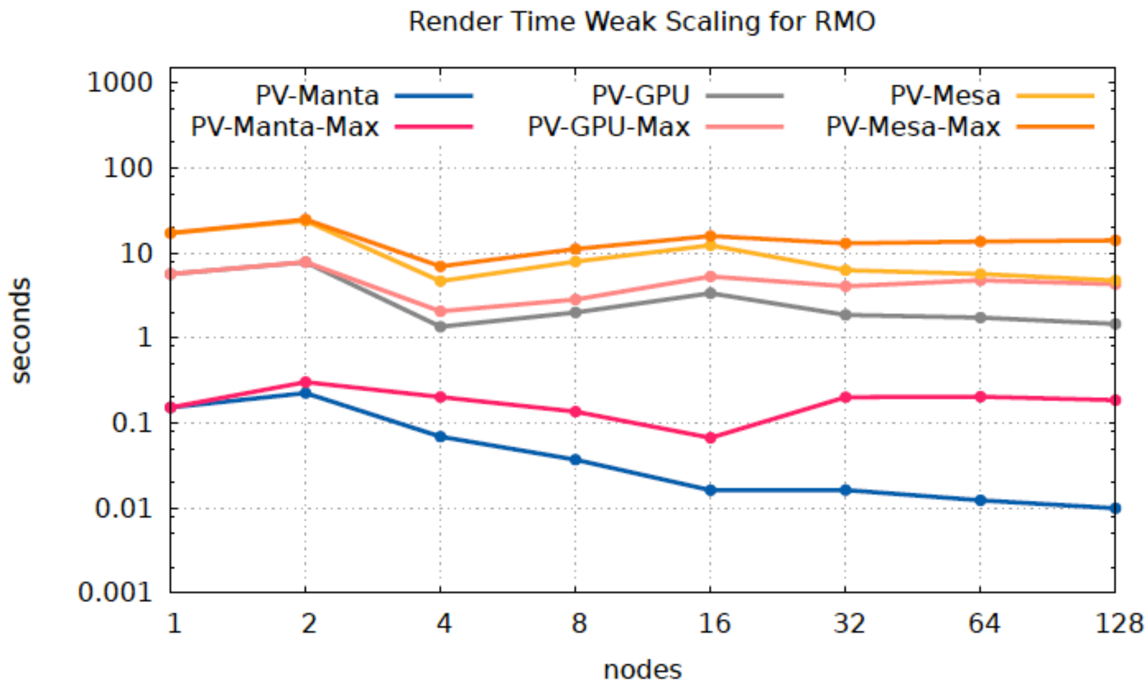


Ray-tracing over-riding VTK renderer



Ray-tracing over-riding VTK renderer

Weak Scaling



Vis Software

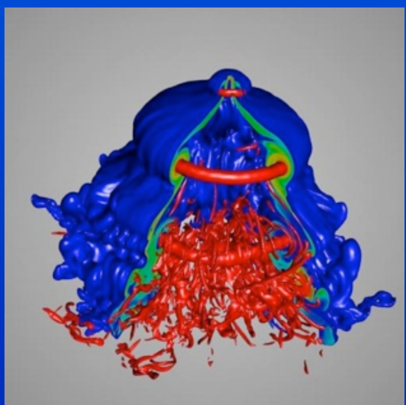
OpenGL

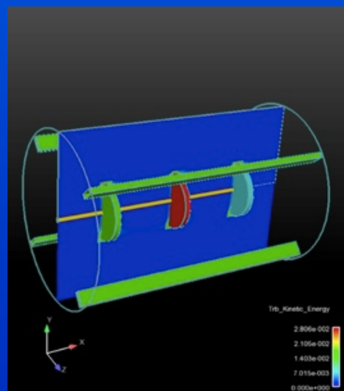
VisIt

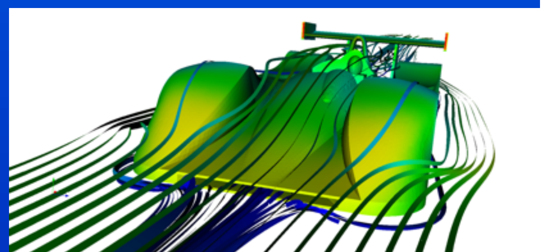
ENSIGHT
Engineering Visualization Software

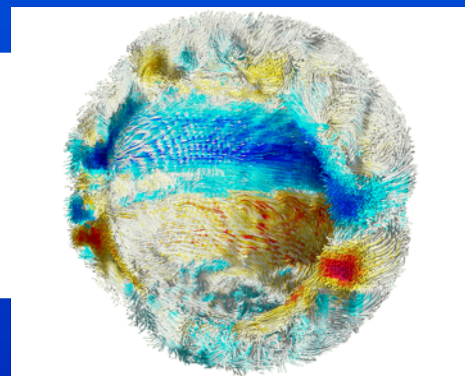
ParaView

VAPOR
Visualization & Analysis Platform

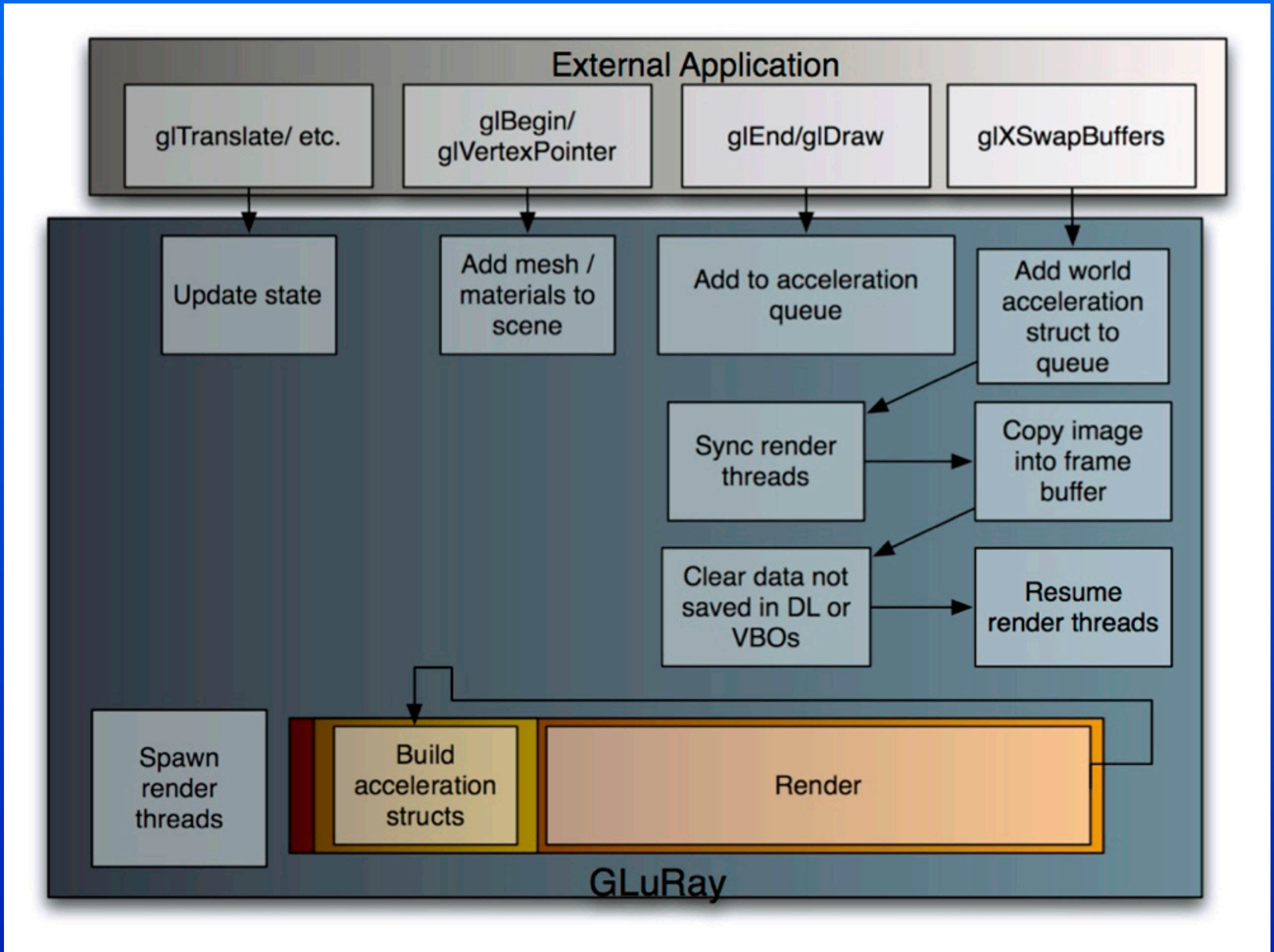






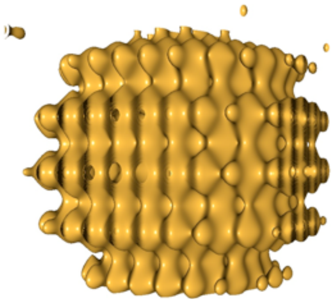


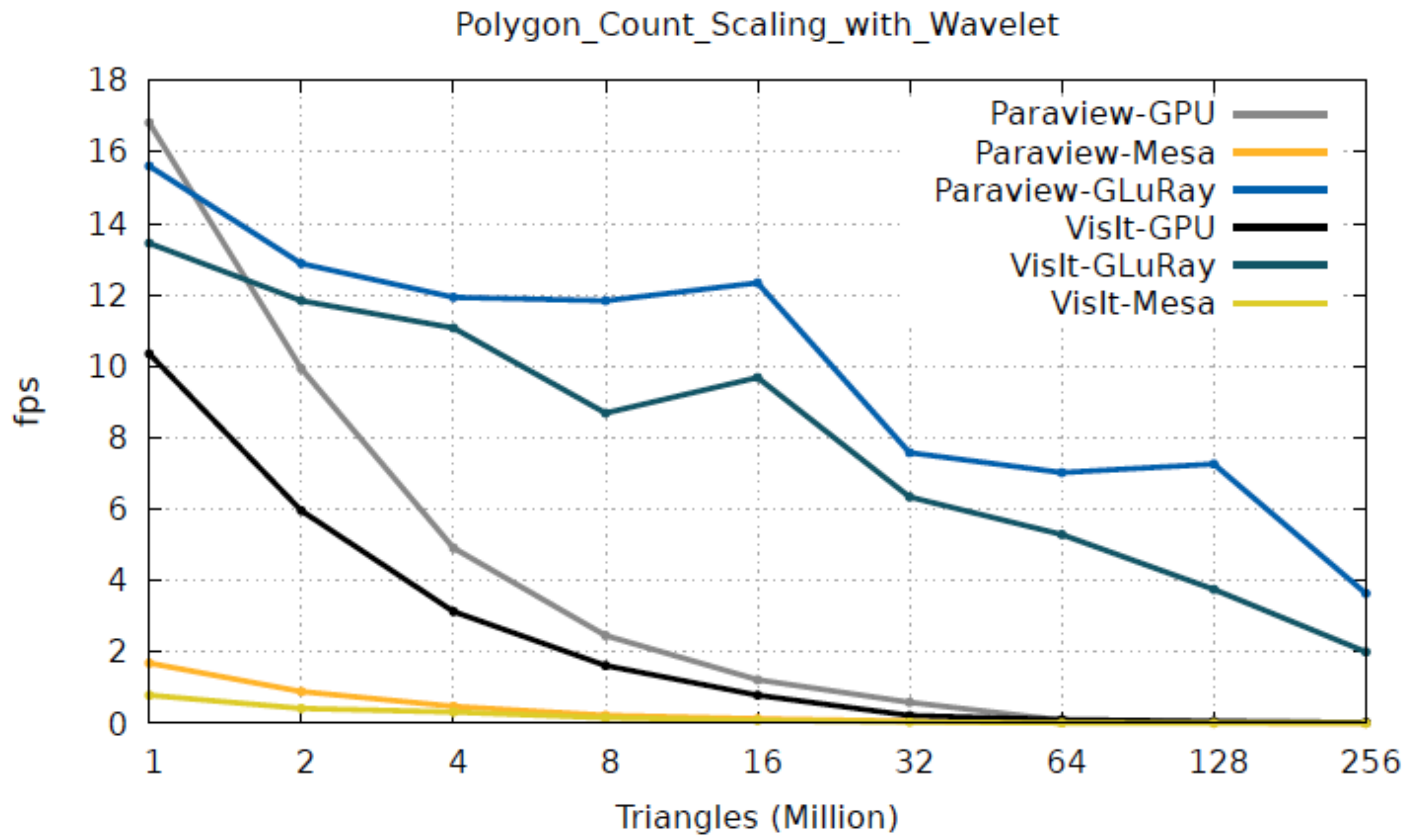
Parallel Architecture



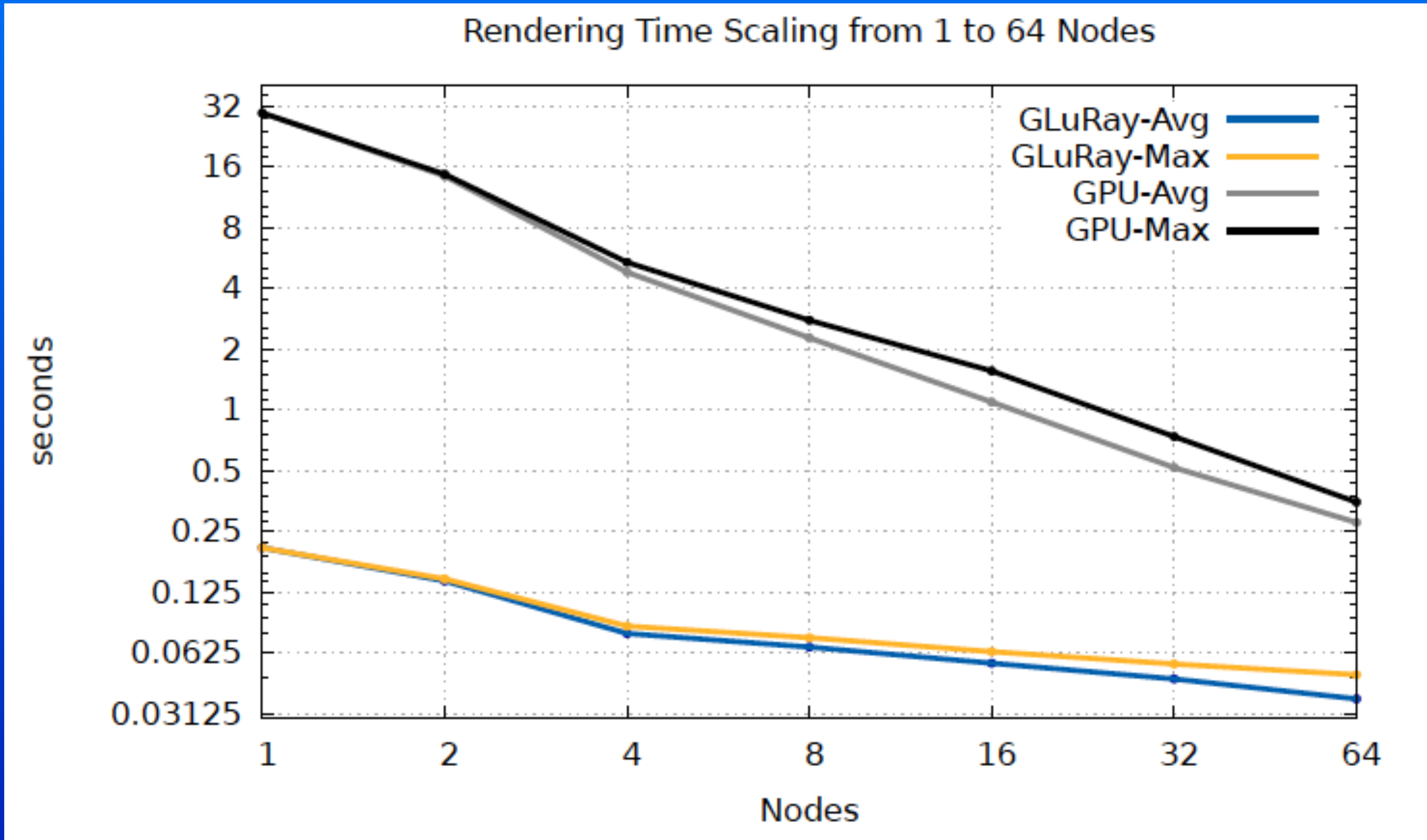
www.sci.utah.edu

Polygon Scaling



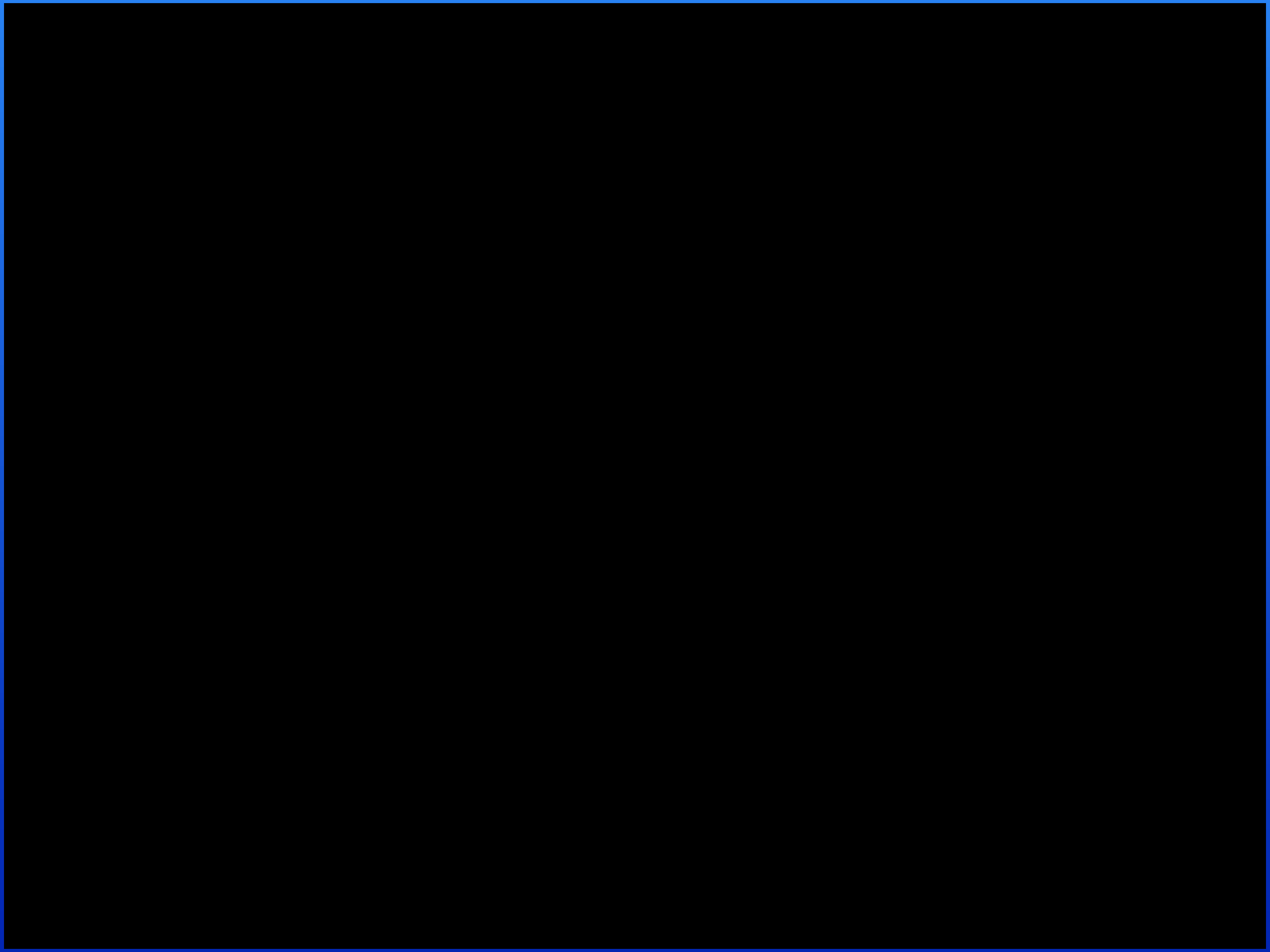


Strong Scaling



www.sci.utah.edu

Michelangelos David



Michelangelos David - Part 2

Big Data & Computational Research

Parallel computing power grows x1000 / decade

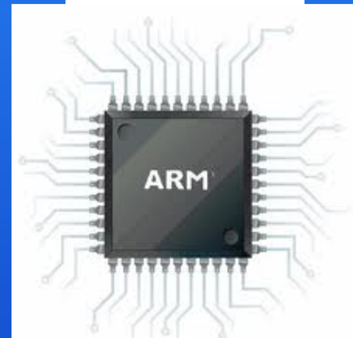
Challenges:

- Scalable multi-physics multiscale problems with possibly many millions of cores.
- Estimate the error and/or uncertainty in the solution
- Energy consumption
- Heterogeneous architectures

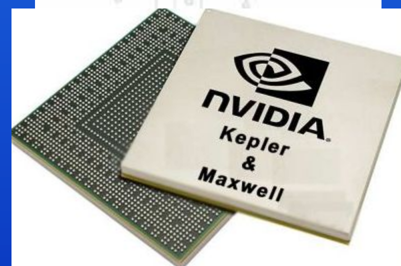
Exascale Challenge for Future Algorithms and Software?

2013 Titan: 288K cpu cores 5M gpu cores

Blue Gene Q 2 Petaflops* per MegaWatt



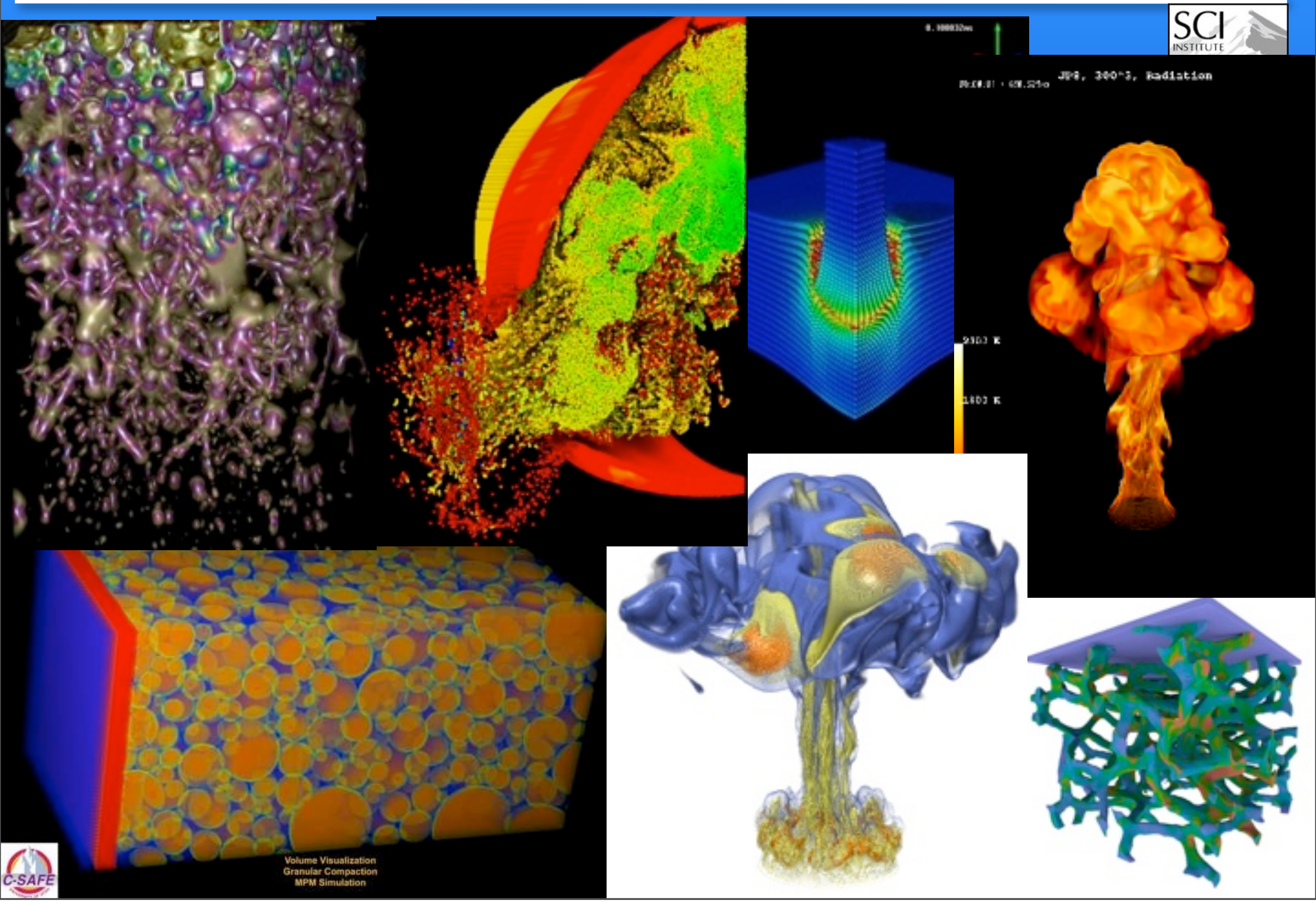
202X Exascale "goal" requires 50 Petaflops per Megawatt, 1B cores - not possible with existing hardware/software approaches



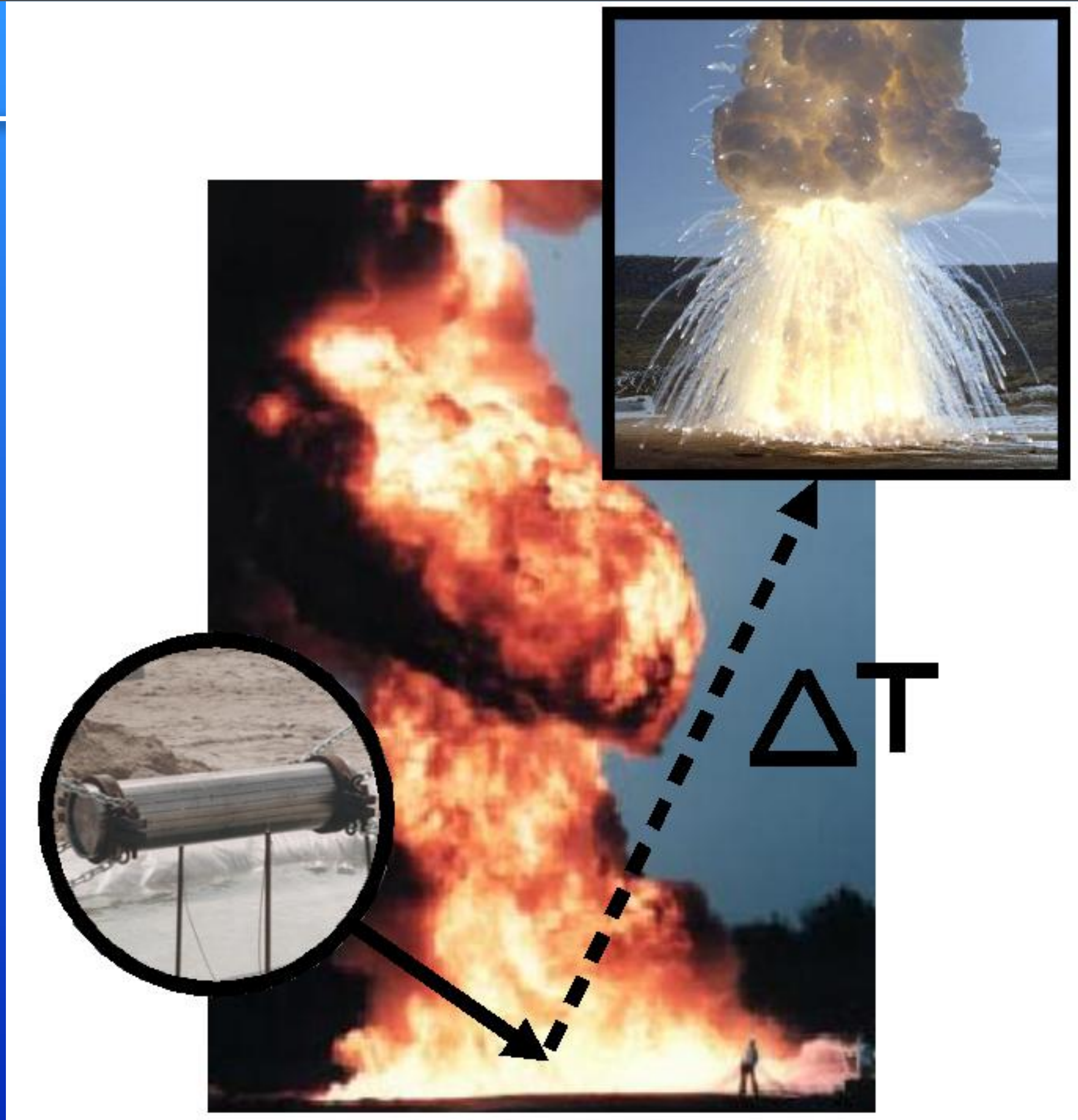
Harrod SC12: "today's bulk synchronous (BSP), distributed memory, communicating sequential processes (CSP) based execution model is approaching an efficiency, scalability, and power wall."

HPC software now has to take into account considerable uncertainty in architectures

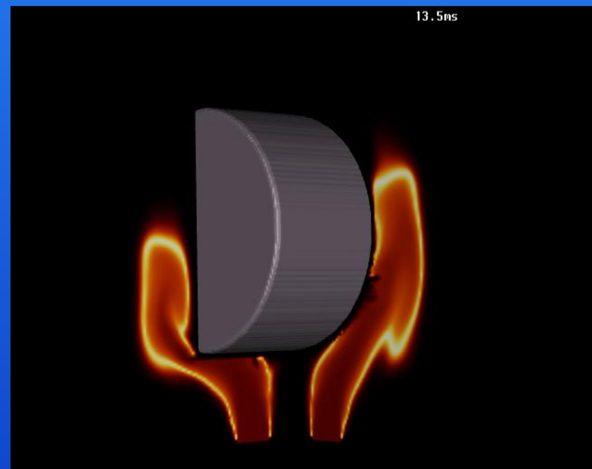
DOE ASCI Center: C-SAFE



C-SAFE



Time scales



Explosion | Fire | Heatup

10^{-6}-10^{-4} | 10^{-3}-10^{1} | 10^{3}-10^{4}

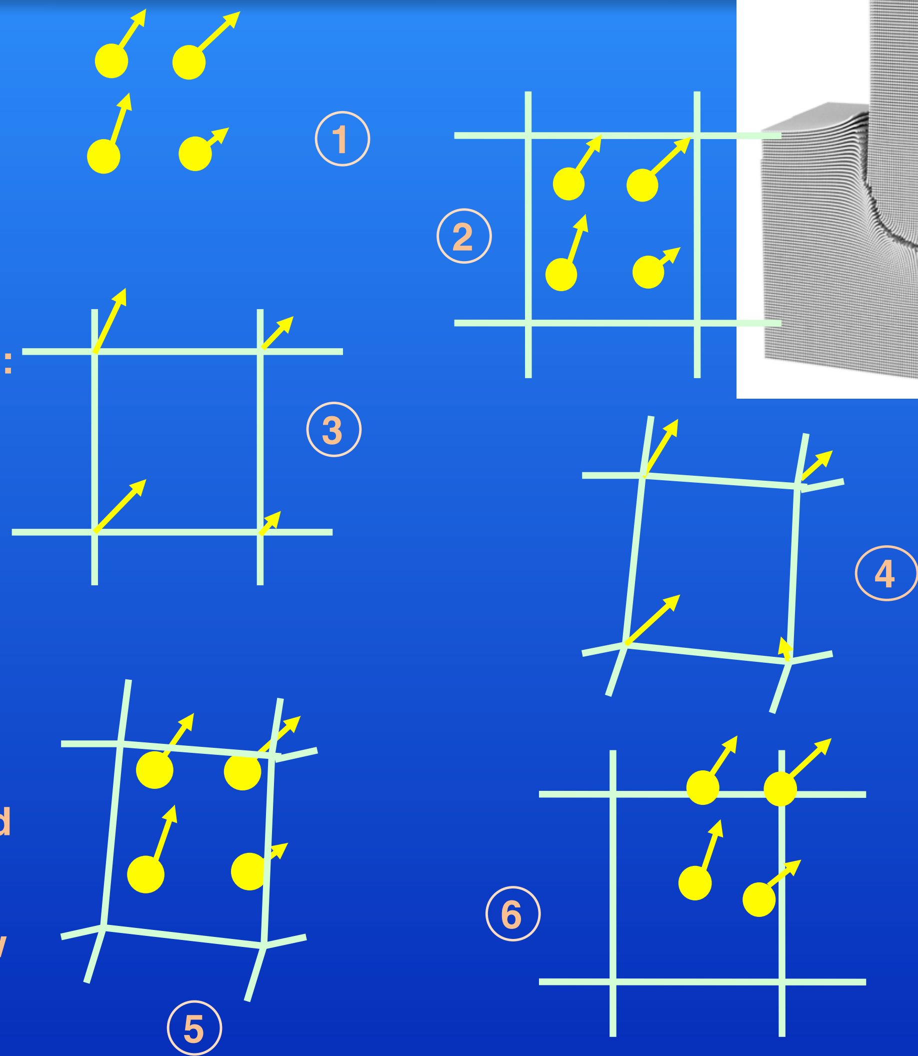
Seconds

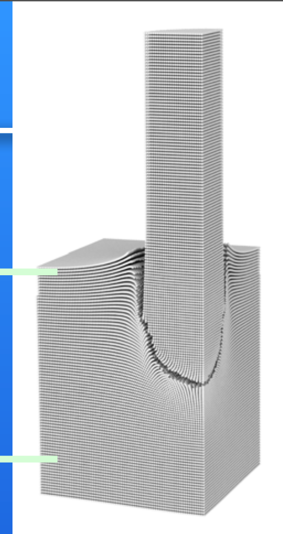
The Material Point Method (MPM)

1. Lagrangian material points carry all state data (position, velocity, stress, etc.)
2. Overlying mesh defined
3. Particle state projected to mesh, e.g.:

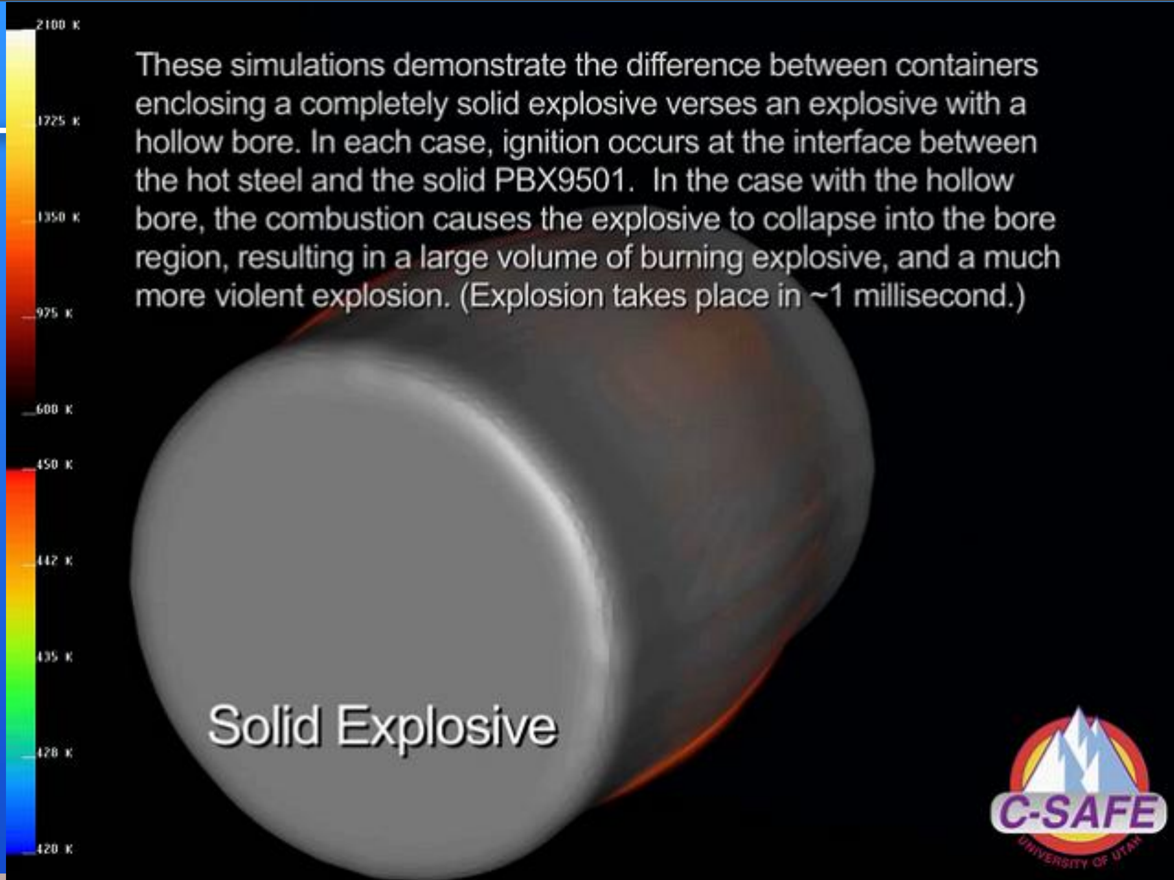
$$v_g = \sum_p S_{gp} m_p v_p \Big/ \sum_p S_{gp} m_p$$

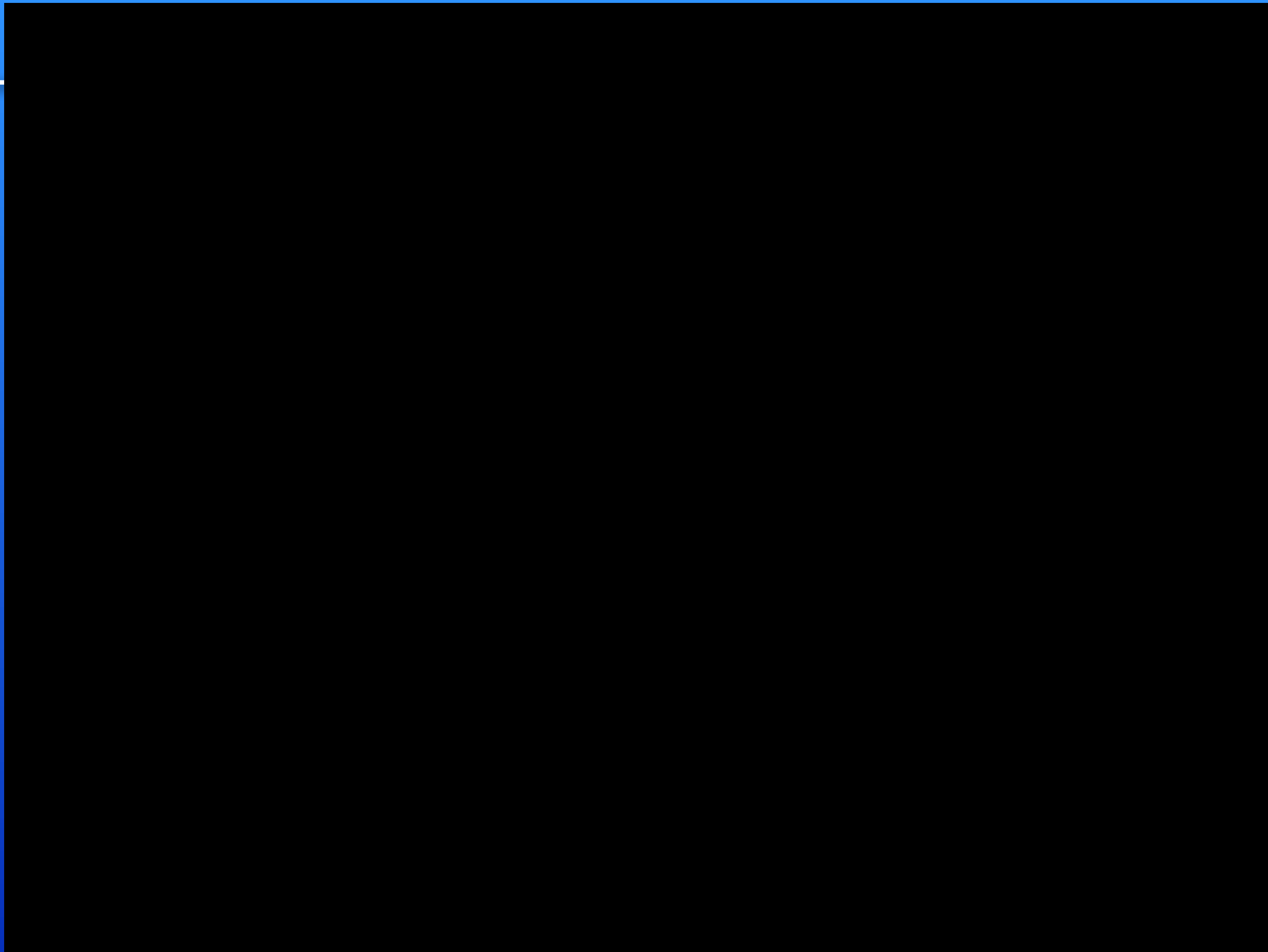
4. Conservation of momentum solved on mesh giving updated mesh velocity and (in principal) position. Stress at particles computed based on gradient of the mesh velocity.
5. Particle positions/velocities updated from mesh solution.
6. Discard deformed mesh. Define new mesh and repeat



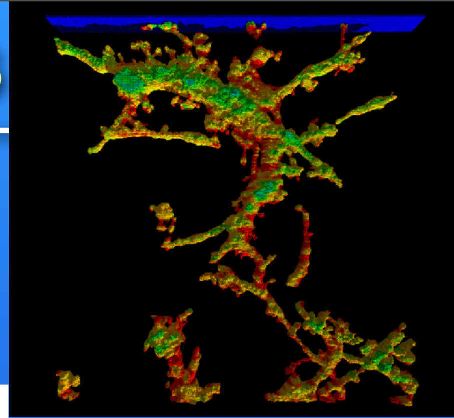


SCI INSTITUTE

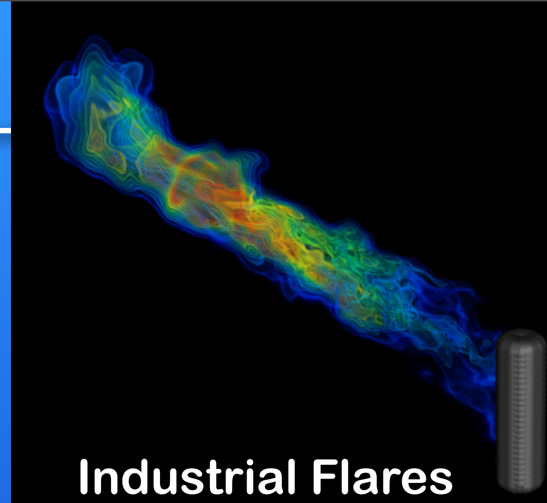


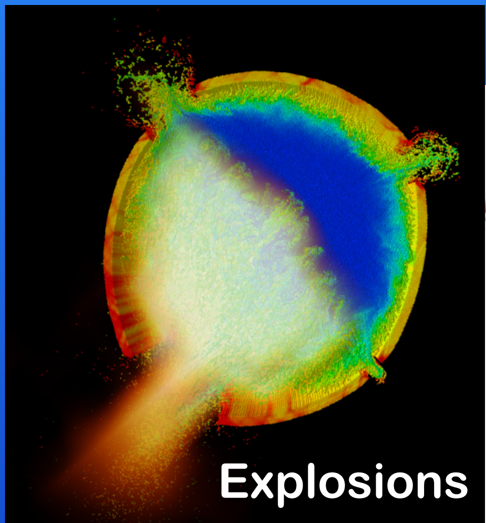


Uintah Applications

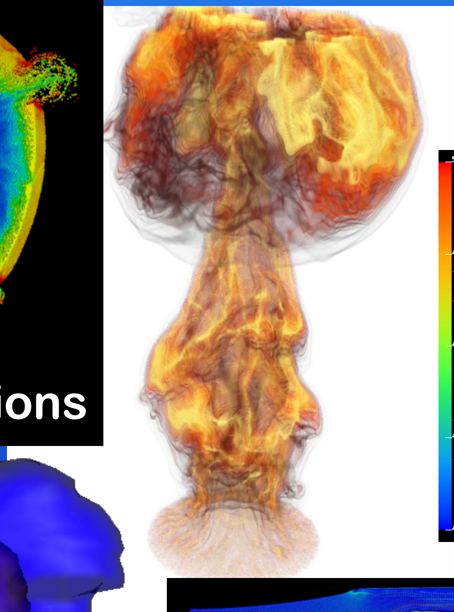


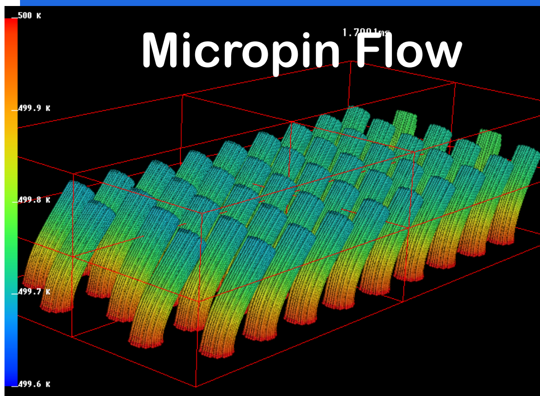
Angiogenesis

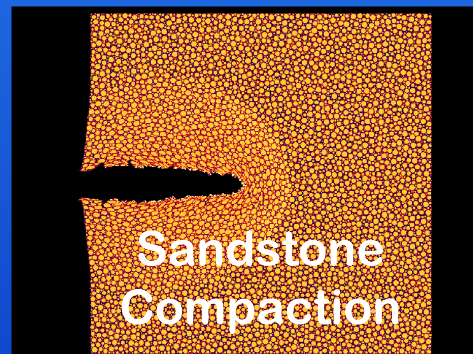


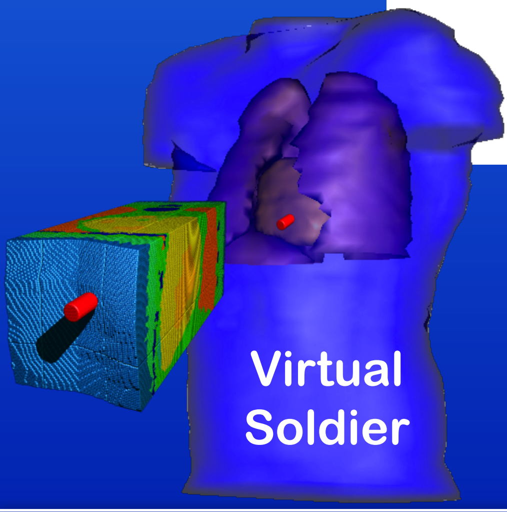


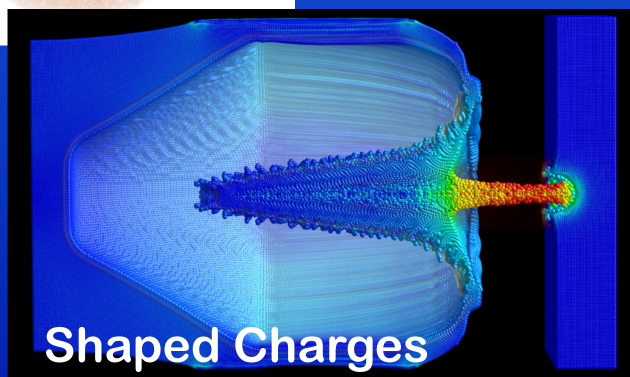
Plume Fires

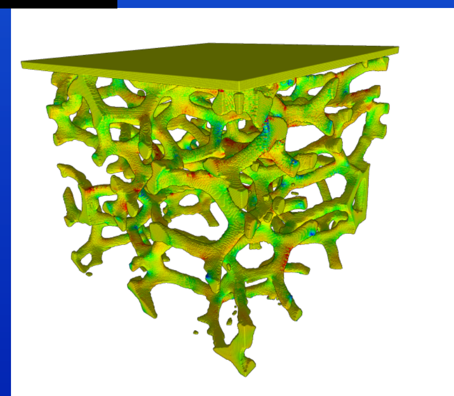










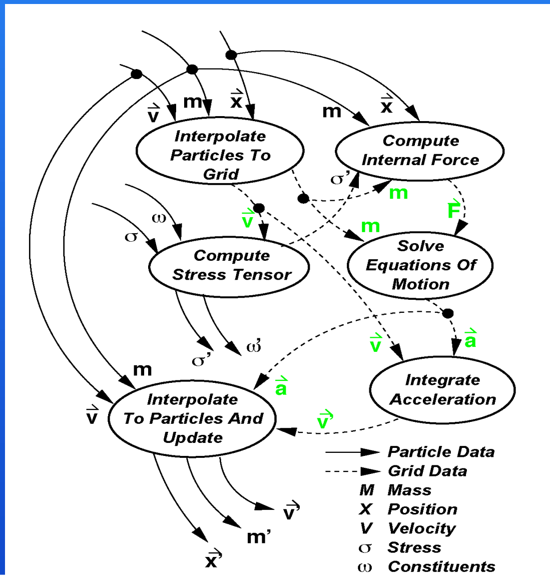


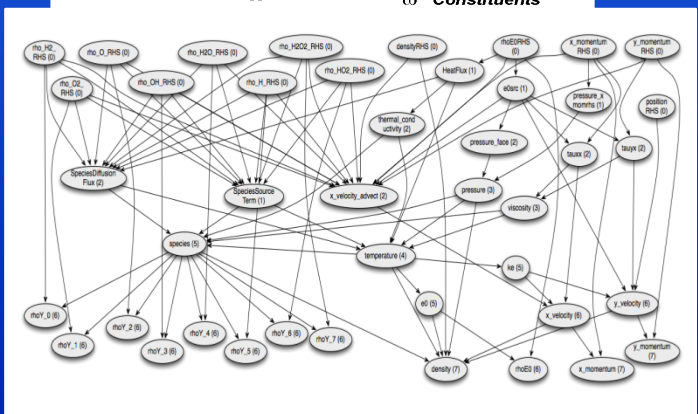
Foam Compaction

Utah Uintah Software Parallelism

SCI INSTITUTE

Uintah uses both data parallelism and task parallelism

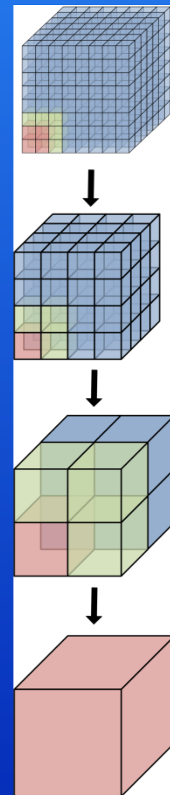




- Structured Grid + Unstructured Points
  - Patch-based Domain Decomposition
  - Adaptive Mesh Refinement

- Dynamic Load Balancing
  - Profiling + Forecasting Model
  - Parallel Space Filling Curves
  - Works on MPI and/or thread level

- Uses **asynchronous task directed graph** approach to scale to 200K cores even for adaptive mesh refinement and fluid-structure interaction



Spanish Fork Accident

Images from KUTV and Deseret News

Highway 6 Explosion

Use Utah Uintah Software

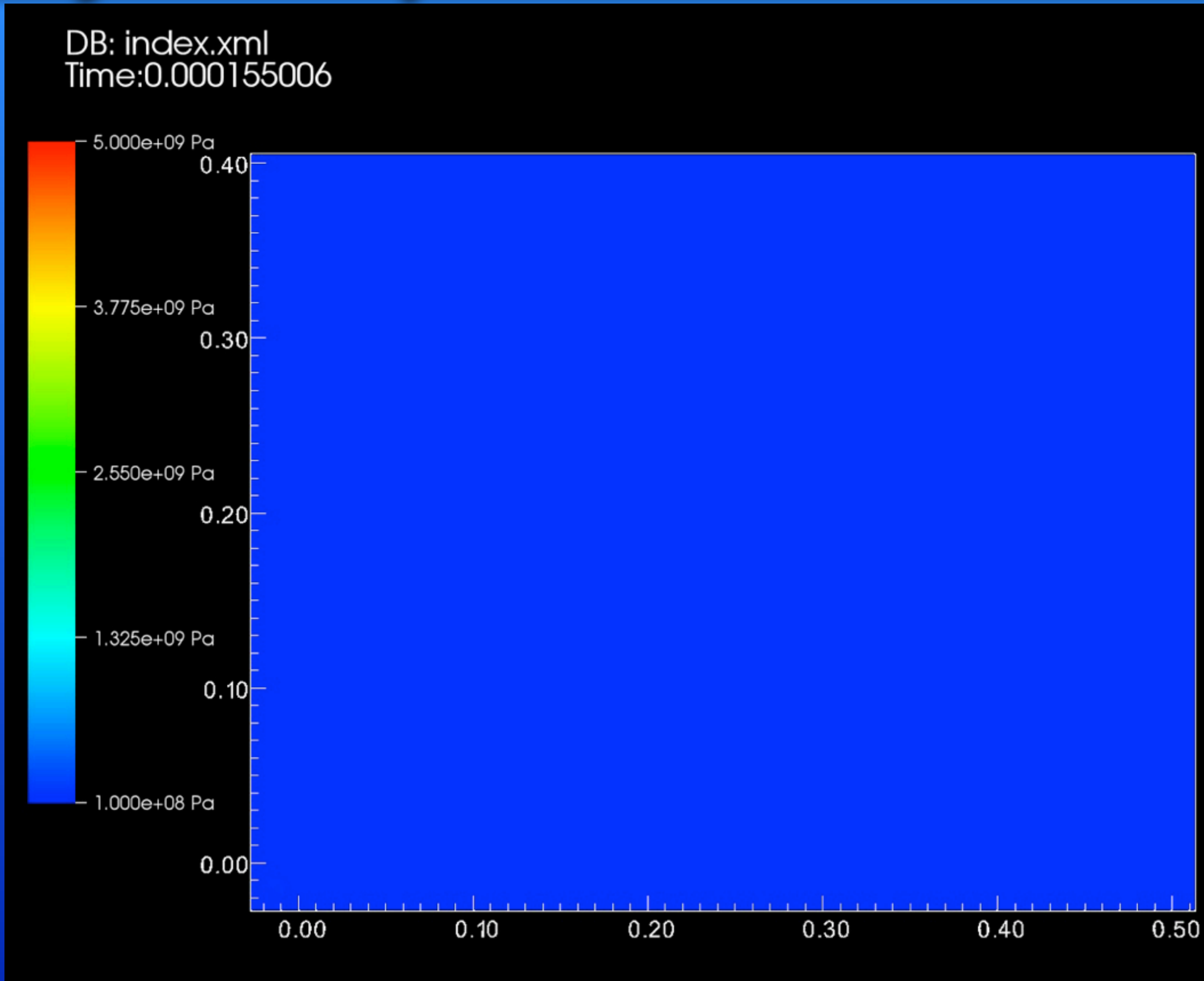
To explain why this happened

And ensure it never happens again

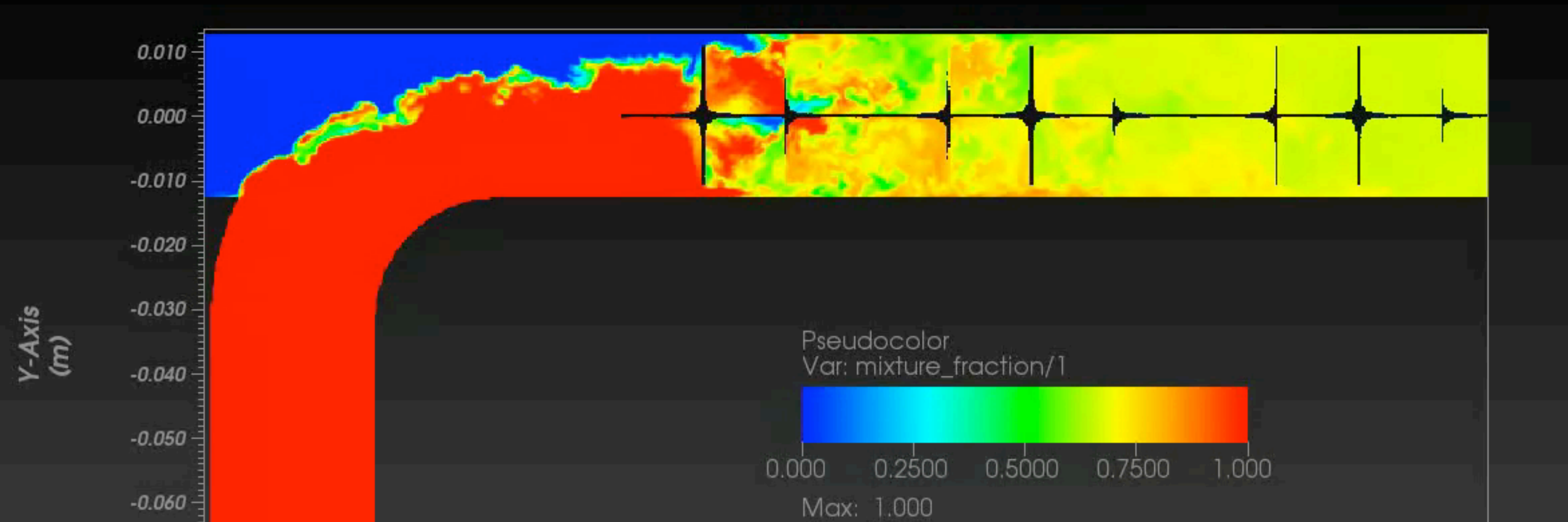
Results by Jacqueline Beckvermit

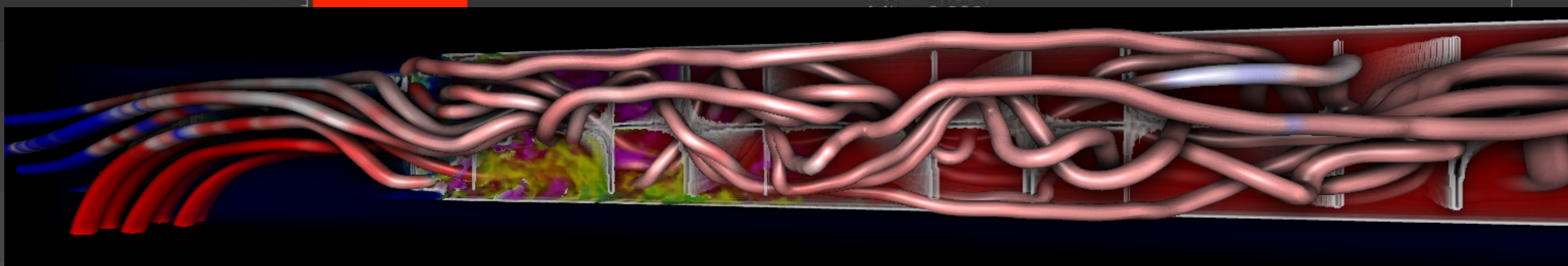
DOE Titan (5M cores)

Highway 6 Explosion



ICSE Carbon Dioxide Cleanup – Red is C02



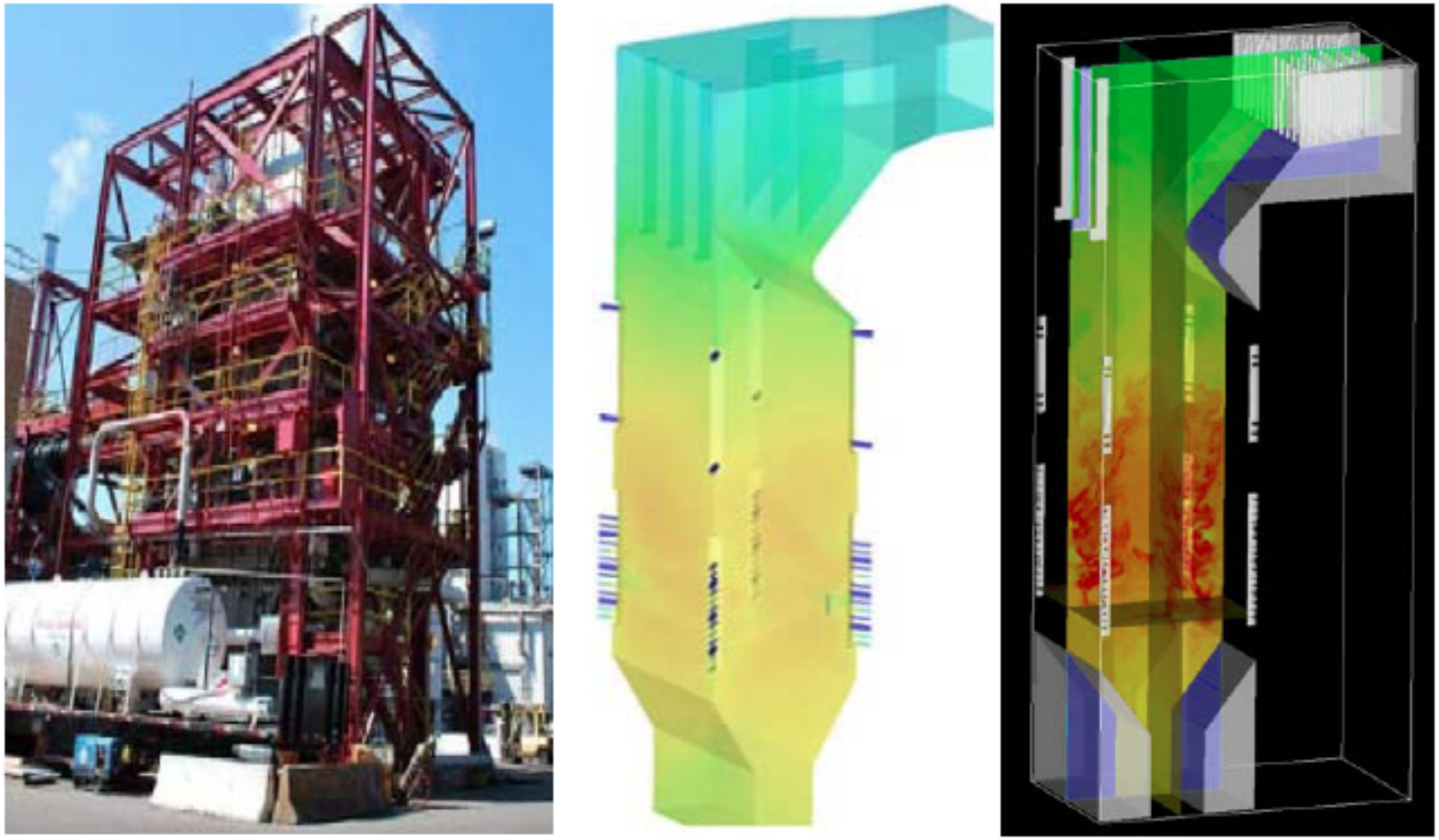


Turbulent flow problem - need to quantify the uncertainty in the Simulation to estimate how much CO2 is removed.

Need at least 100K cores to resolve the problem scales.

DOE PSAAP-II Center

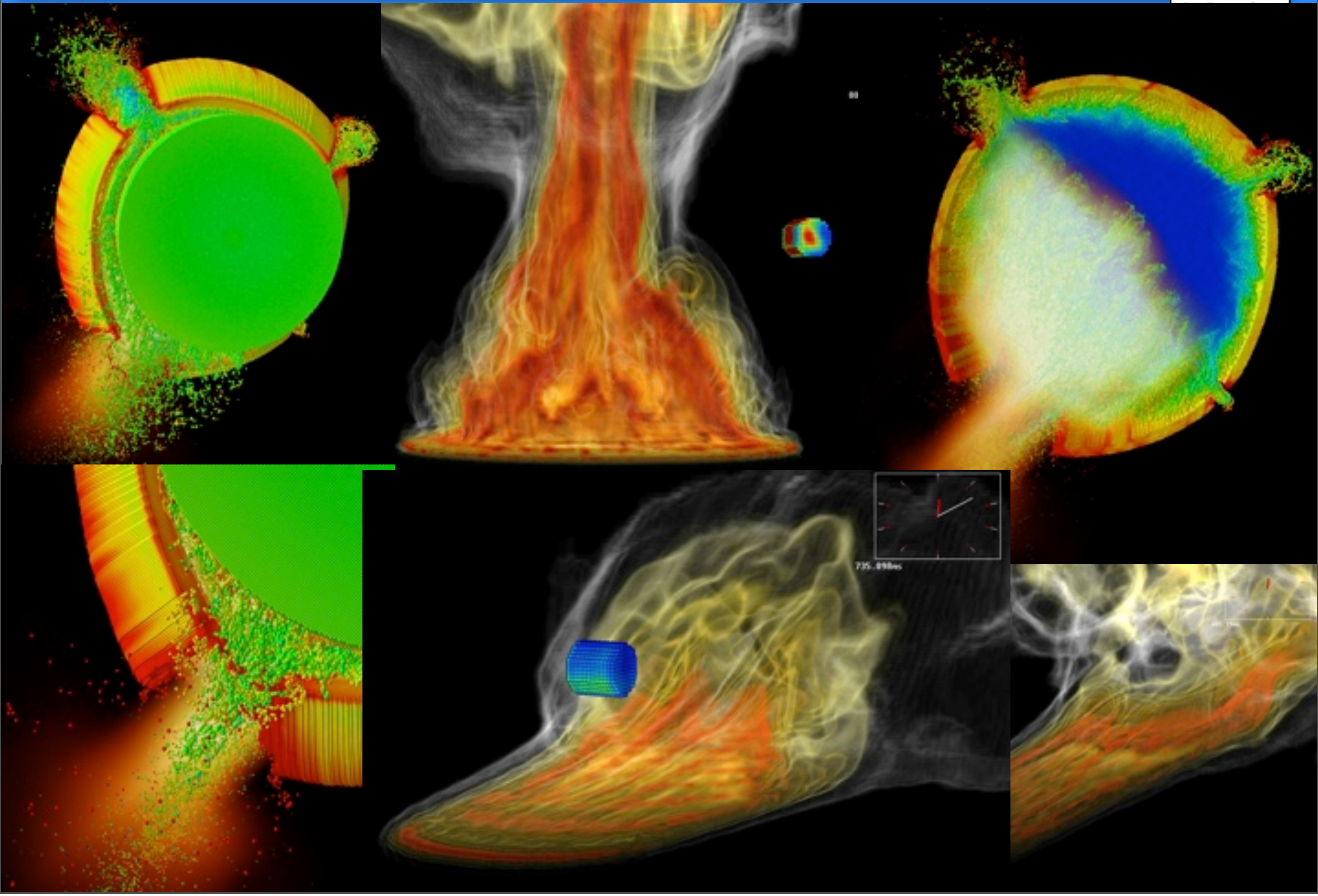
SCI INSTITUTE



(a) BSF test bed (b) RANS Simulation domain (c) LES Simulation domain

Figure 45. LES Simulation domain and grid of BSF

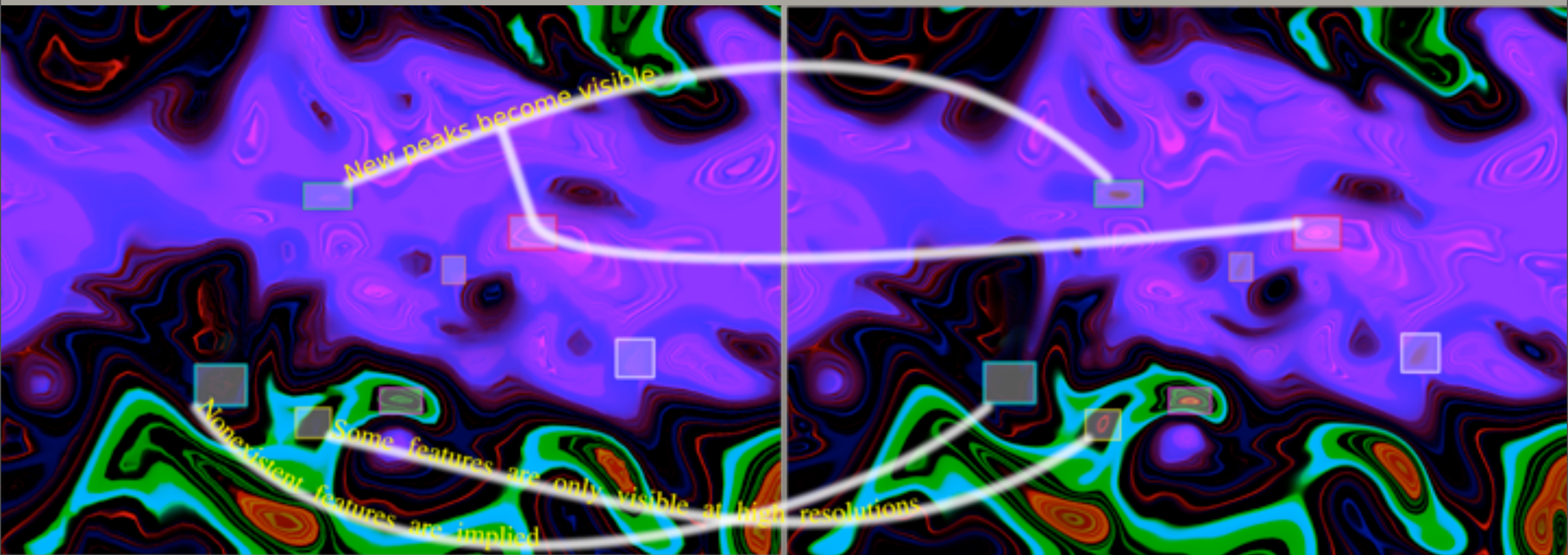
Manta - Real Time Ray Tracer



VACET

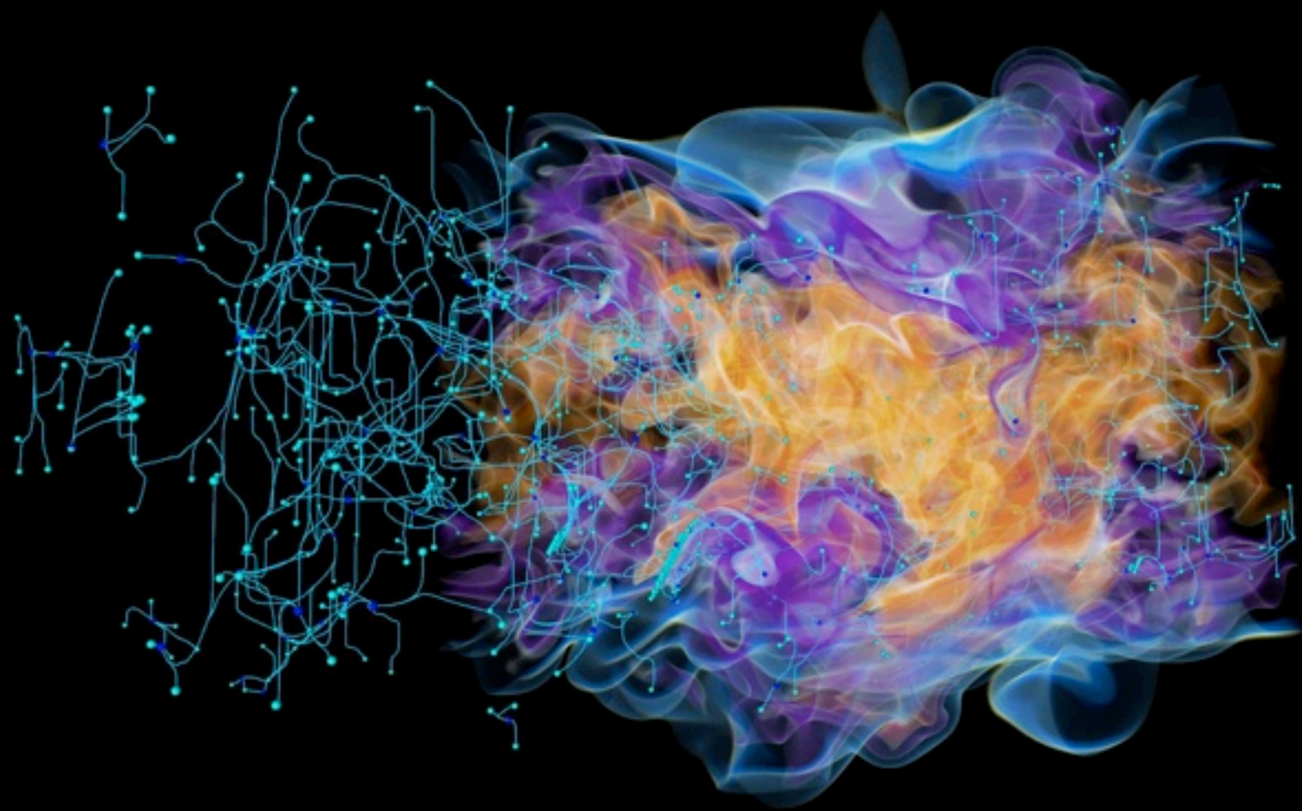
The Need for High Resolution Visualization

"…the data show for the first time how detailed transport and chemistry effects can influence the mixing of reactive scalars. It may be advantageous to incorporate these effects within molecular mixing models. It is worth noting that at present it is impossible to obtain this type of information any other way than by using the type of highly resolved simulation performed here."
Jacqueline Chen, Sandia National Laboratories



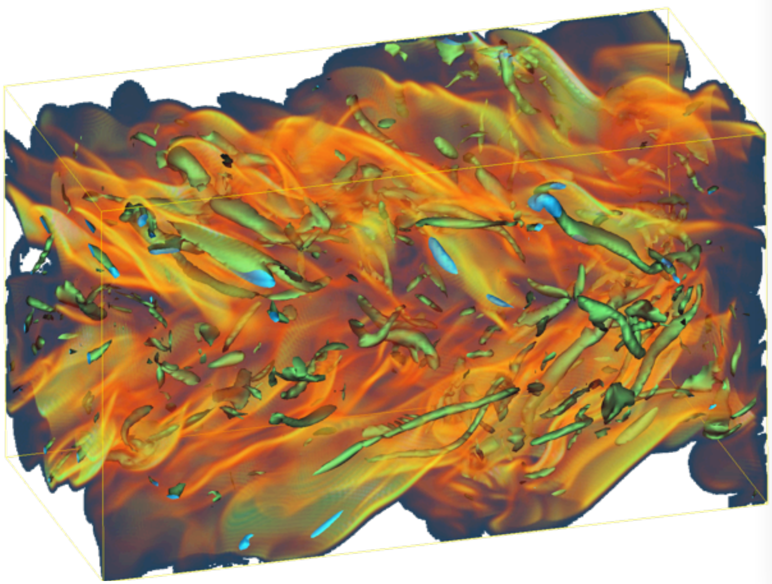
Lower Resolution

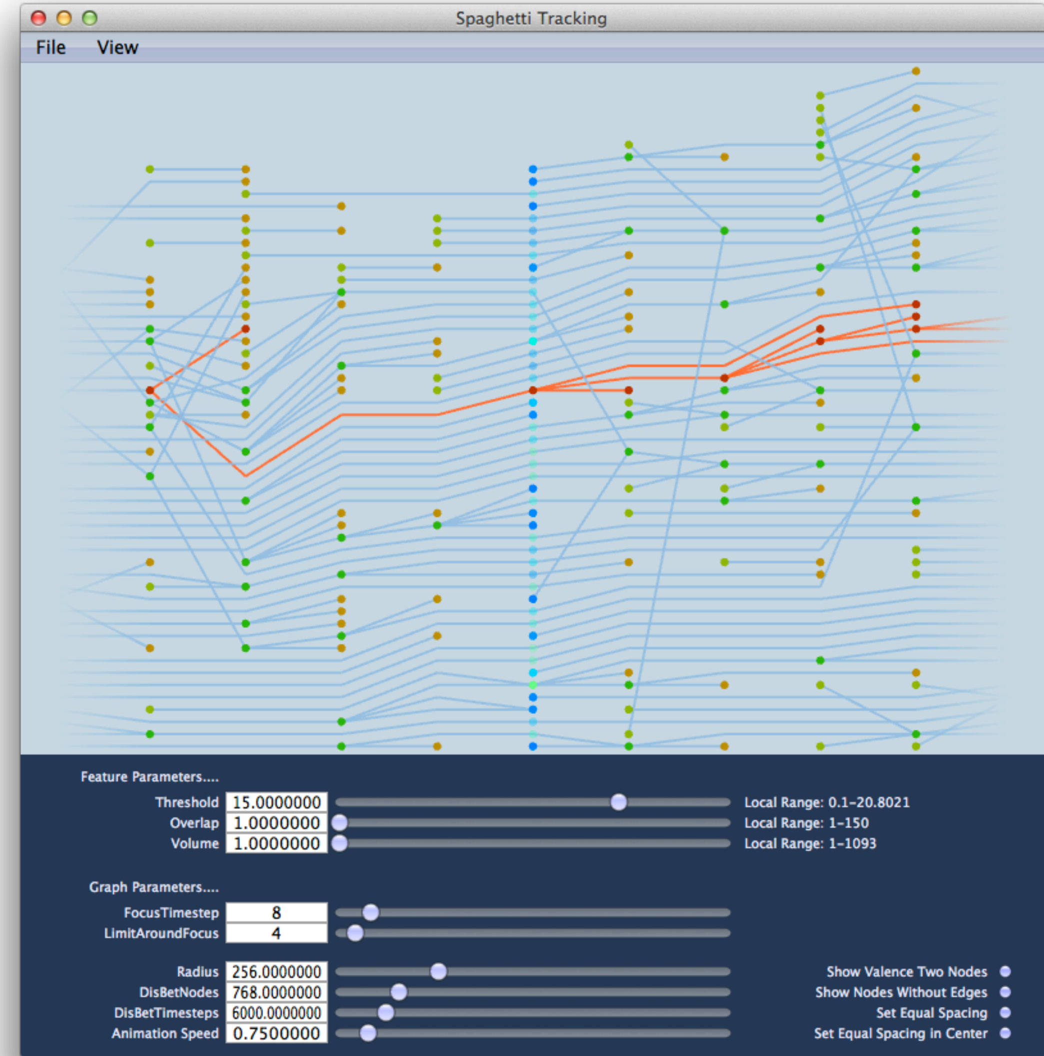
High Resolution



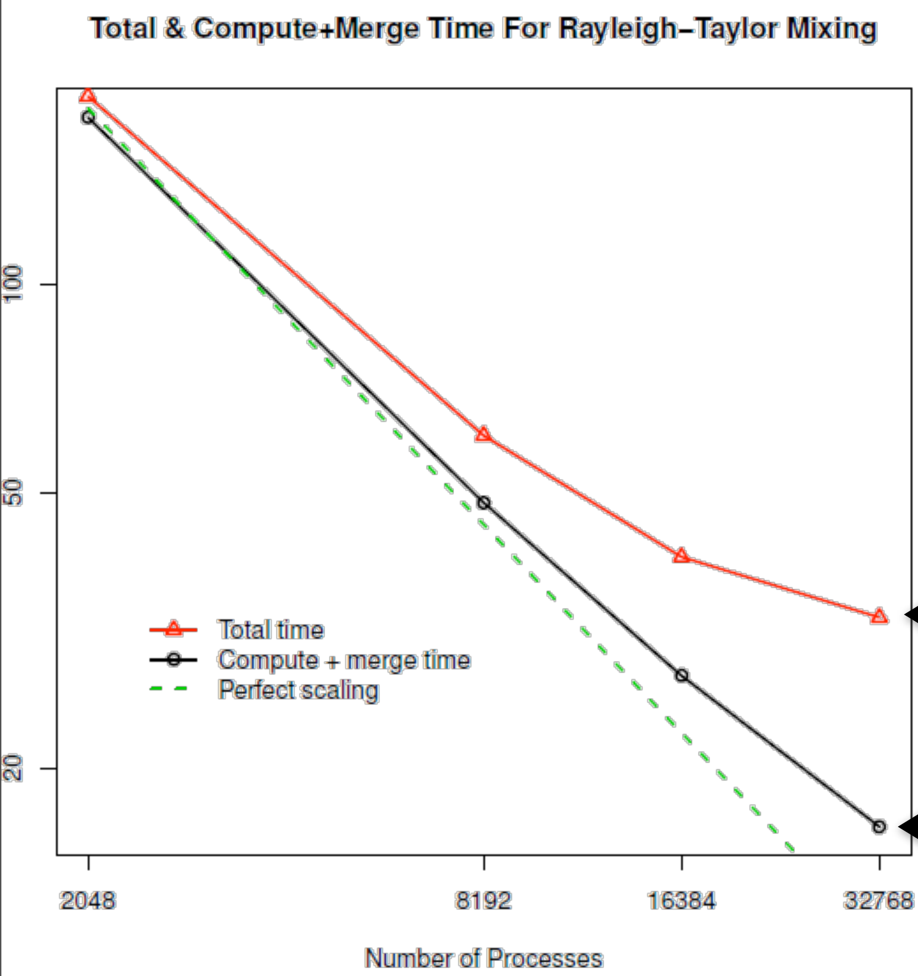
Topological Analysis of Massive Combustion Simulations

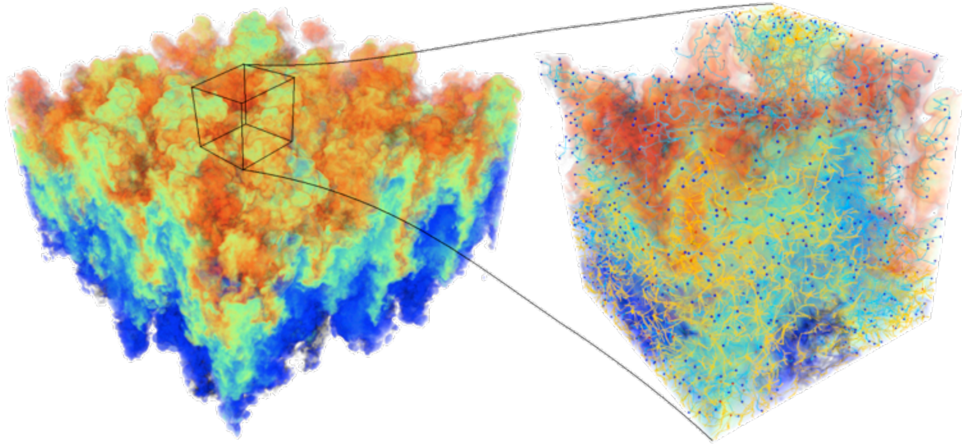
- Non-premixed DNS combustion (J. Chen, SNL): Analysis of the time evolution of extinction and reignition regions for the design of better fuels





New Parallel Topological Computations Achieve High Performance at Scale



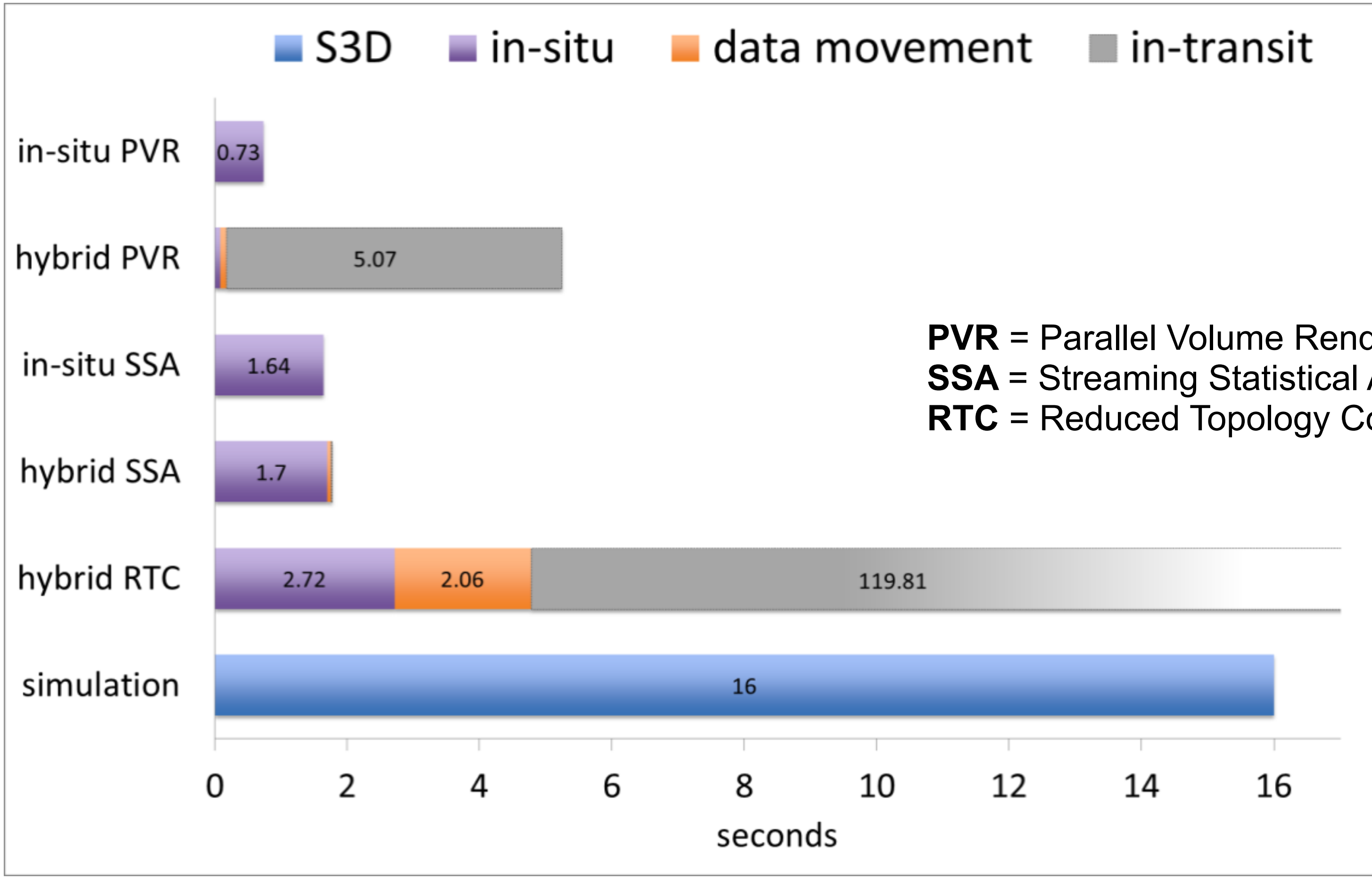


Computation + I/O

Pure Computation

Exploring algorithm design and task allocation

in-situ+in-transit workflows enable matching algorithms with architectures

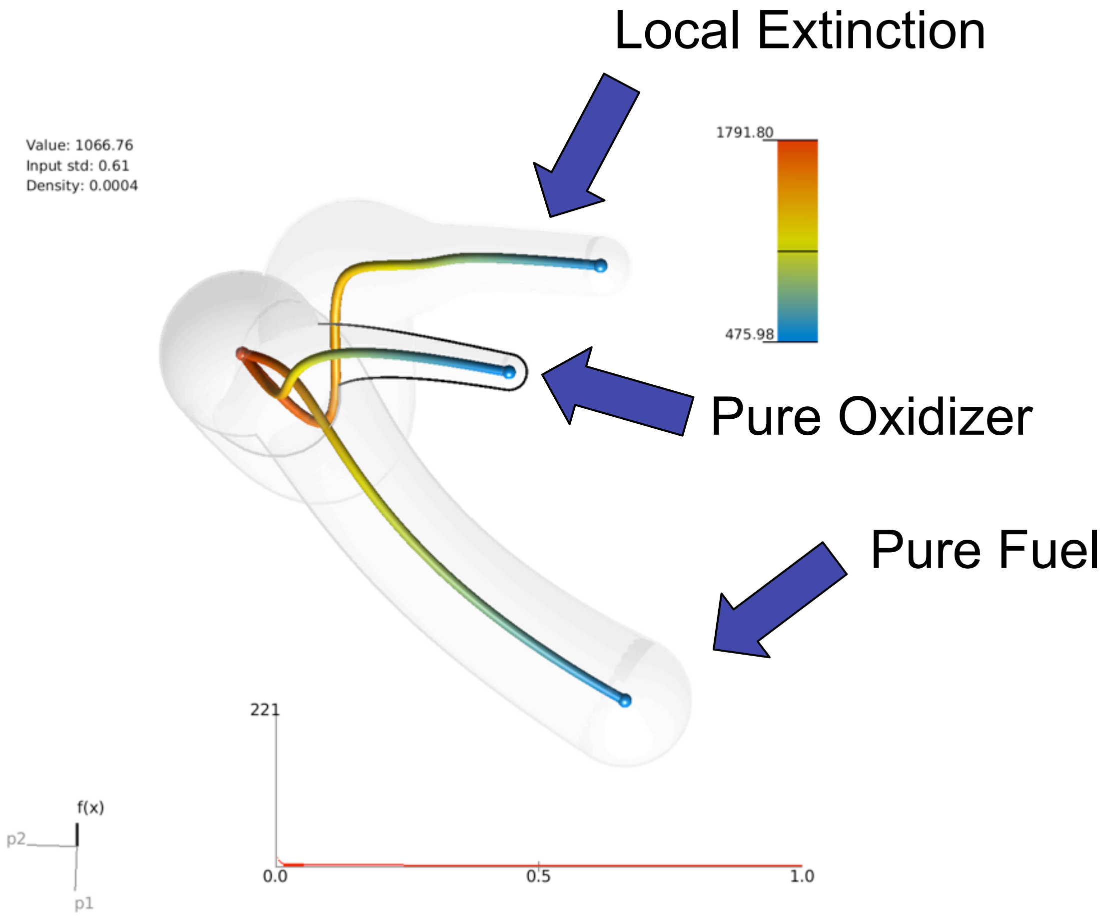


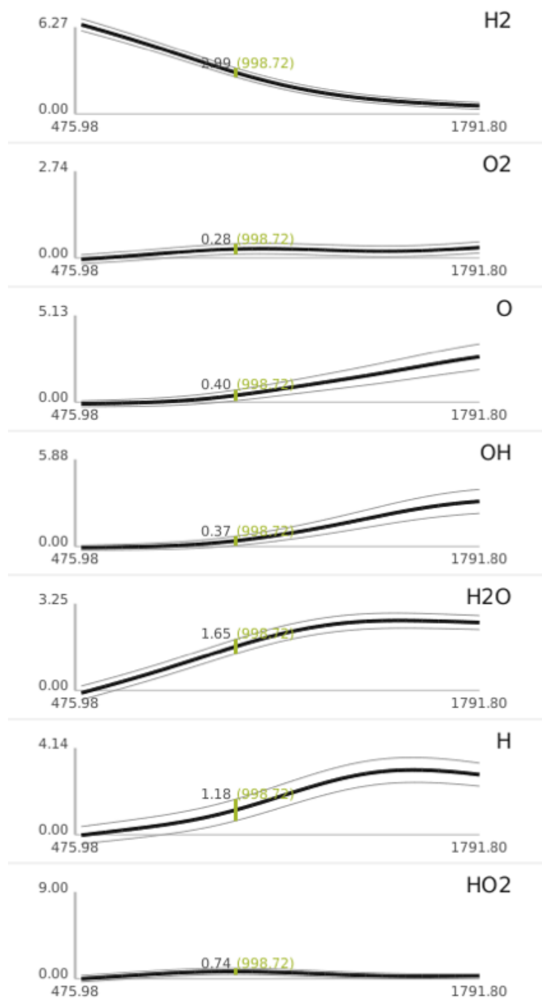
PVR = Parallel Volume Rendering
SSA = Streaming Statistical Analysis
RTC = Reduced Topology Computation

- 4896 cores total (4480 simulation/in situ; 256 in transit; 160 task scheduling/data movement)
- Simulation size: 1600x1372x430 ; All measurements are per simulation time step

[SC12a] **Combining In-Situ and In-Transit Processing to Enable Extreme-Scale Scientific Analysis**

Visualization of 10D Combustion Simulation of Jet CO/H2-Air Flames

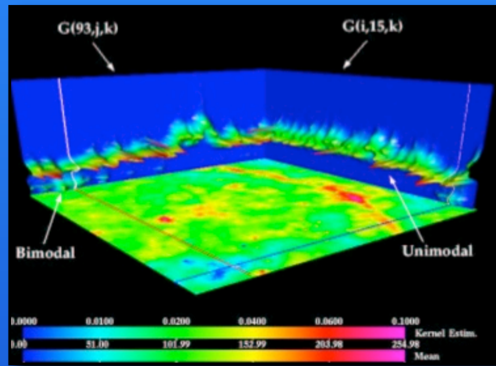




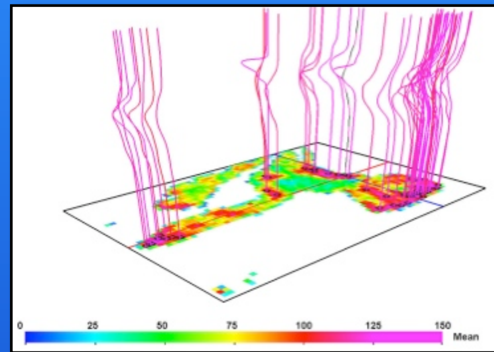
10 dimensional data set describing the heat release wrt. to various chemical species in a combustion simulation

Ensembles

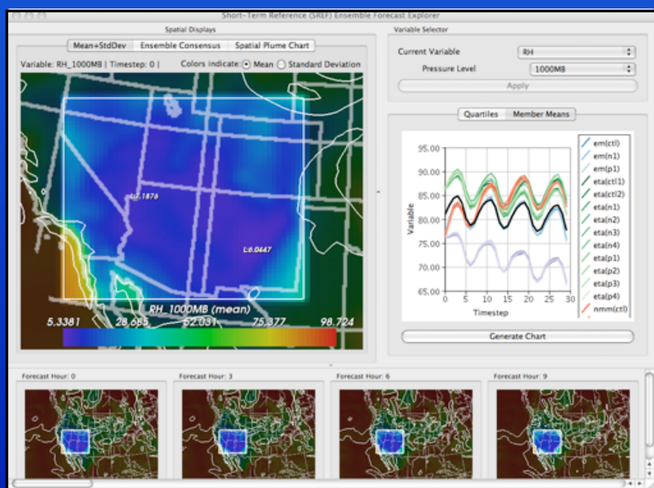
- Multi-run/model simulations
- Distribution of data at every point
- Mean/std dev may not be appropriate



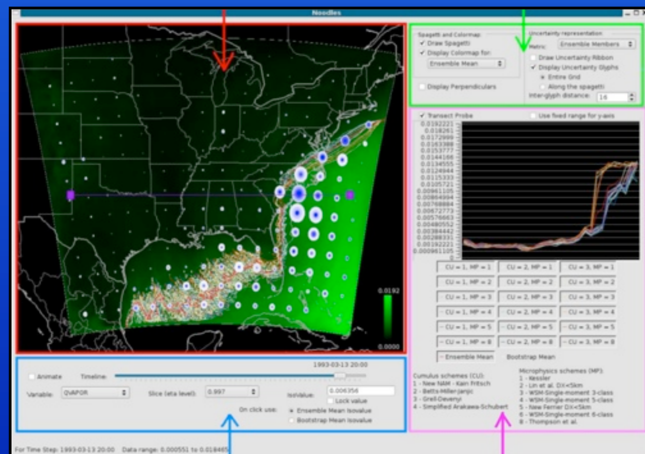
D. Kao, A. Luo, J. Dungan, A. Pang.
Visualizing Spatially Varying Distribution Data.
In Proc *Information Visualization*, 2002.



D. Kao, M. Kramer, A. Luo, J. Dungan, A. Pang.
Visualizing Distributions from Multi-Return Lidar Data to Understand Forest Structure.



K. Potter, et al.
Ensemble-Vis: A Framework for the Statistical Visualization of Ensemble Data.
In *IEEE ICDM Workshop on Knowledge Discovery from Climate Data: Prediction*, 2009.



J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn.
Noodles: A Tool for Visualization of Numerical Weather Model Ensemble Uncertainty
In Proc *IEEE Vis*, 2010.

Big Data Challenges in Sci. Vis.

Scalable methods

In-situ / in-transit methods

Feature extraction / tracking

Power aware algorithms

Reliability / resiliency

Uncertainty quantification

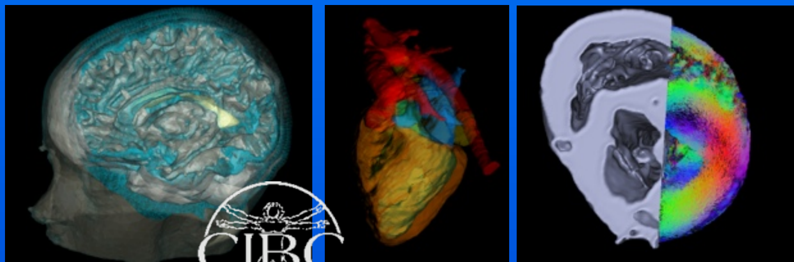
Visual comparisons

The SCI Institute

www.sci.utah.edu

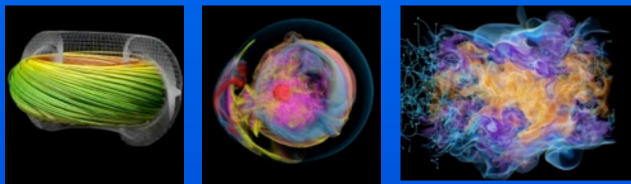
Acknowledgments

NIH/NIGMS Center for Integrative Biomedical Computing



SDAV

Scalable Data Management, Analysis and Visualization



UTAH Center for Computational Earth Sciences

Center for Extreme Data Management, Analysis, and Visualization

CEDMAV

NIH NAMIC

NIH NAMIC

KAUST

IAMCS
Institute for Applied Mathematics and Computational Science

National Science Foundation
Directorate for Computer & Information Science & Engineering (CISE)

SCI
www.sci.utah.edu

More Information

www.sci.utah.edu

hansen@sci.utah.edu