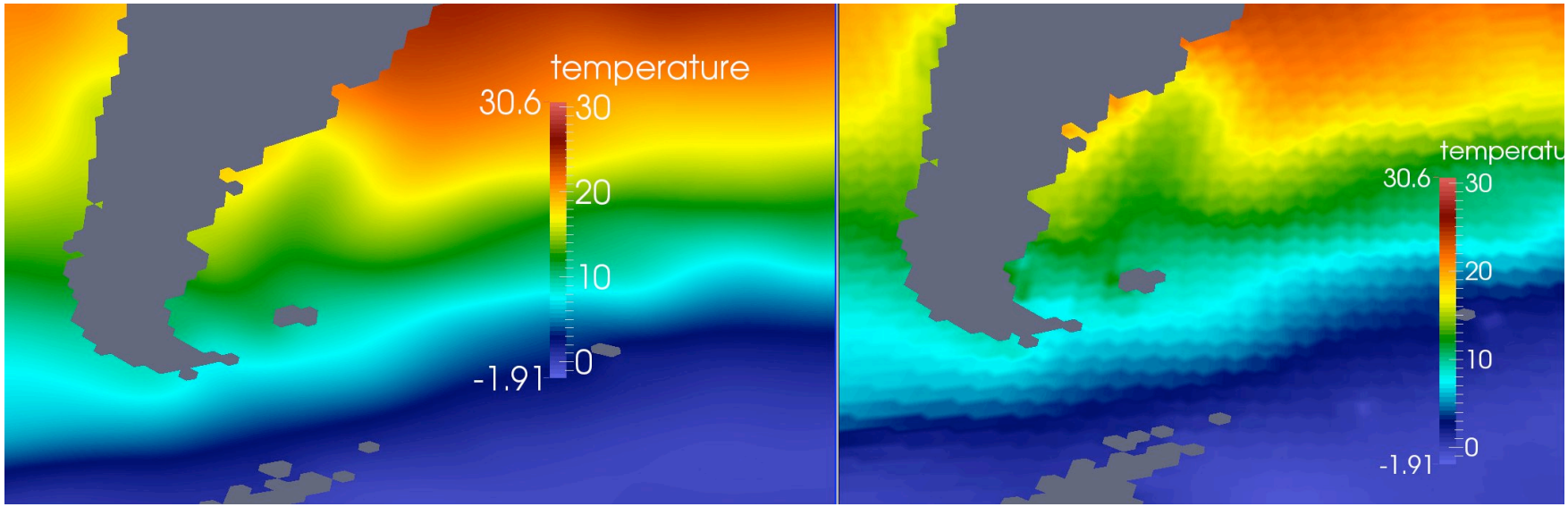
Shared Analysis for Resilience, Debugging, Verification, Validation and Discovery



James Ahrens

Los Alamos National Laboratory

April 2015 – Salishan – LA-UR-15-23284

Trends for HPC scientific visualization and analysis

<u>Relentless increase in data sizes</u>

- 3 orders of magnitude every ten years

<u>Adapting to changing infrastructure</u>

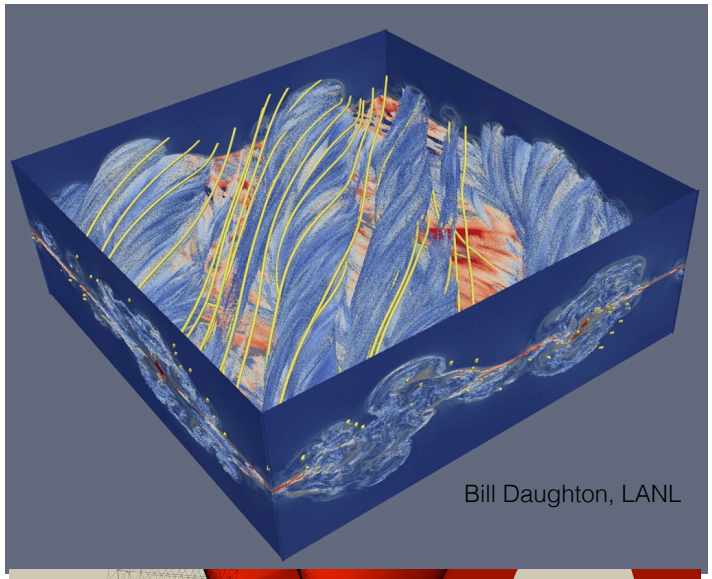
- Shared memory, clusters, threading, cloud

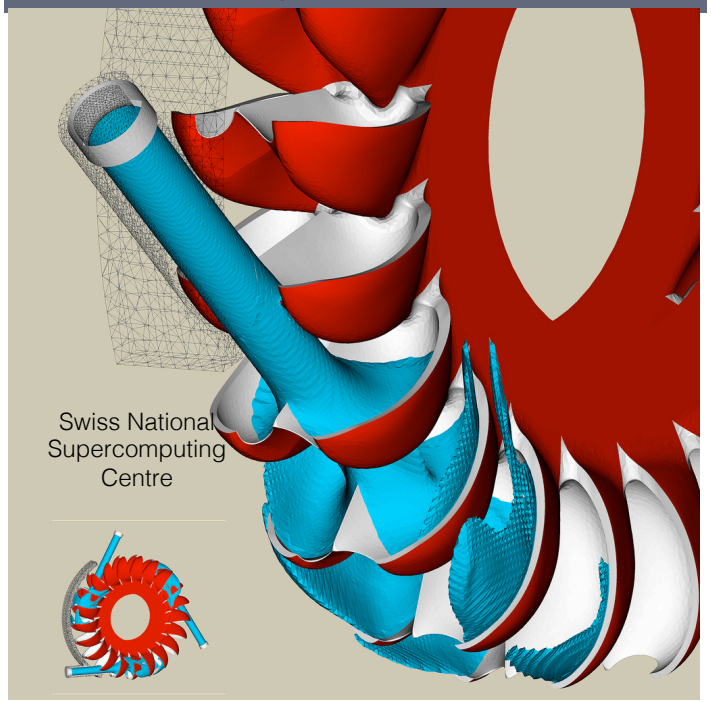
<u>Advancing the fundamentals</u>

- Improved end-to-end workflow and cognitive understanding
- <u>How about the user experience?</u>

Responding to the trends: ParaView

- An open-source, scalable, multi-platform visualization application
- Support for distributed computation models to process large data sets
  - Billions of AMR cells, Scaling test over 1 Trillion cells
- Used by academic, government and commercial institutions worldwide
  - Downloaded ~100K times per year
  - Developed by Kitware, LANL, SNL…
- <u>Originally designed to support a post processing workflow</u>
  - Simulations save data to storage and scientist interactive visualizes results





http://paraview.org

Understand our simulations

Are we solving our equations correctly?

- Debugging/verification

Is our data corrupted?

- Fault detection

How is our performance?

- Time, space, power

How do our simulation results compare to real-world experimental data?

- Validation

Have we found a new scientific phenomena or process?

- Scientific discovery

Context: In situ analysis required at exacale

The traditional post-processing approach is becoming unworkable

- Temporal simulation snapshots are saved at longer intervals
  - Full checkpoints are costly - less temporal data available for analysis
- Rate of improvement of rotating storage is not keeping pace with compute
  - Power, cost and reliability are becoming significant issues

Transition from a post-processing to an in situ focused approach (*True for all analysis problems*)

- In situ saves <u>reduced-sized data</u> products during simulation run
  - Benefits:
    - Save disk space
    - Save time in post-processing analysis
    - Produce higher temporal fidelity results

- Sampling problem

- <u>Automatic</u> analysis during the simulation run
  - Prioritized by scientist's importance metrics
  - <u>Event detection, characterization and response</u>

- Help manage cognitive and storage resource budget

In situ analysis framework

- Detection
  - Event type
  - Issues: Granularity / Data Size / Accuracy
- Characterization
  - Fault, bug or new science?
  - Interface with detection and response
- Response
  - Action
    - Examples from in situ analysis

- This talk's perspective from application level down:
  - Currently temporal granularity for in situ analysis is typically at each simulation time step

Application example of framework: Simple range constraints for MPAS ocean code

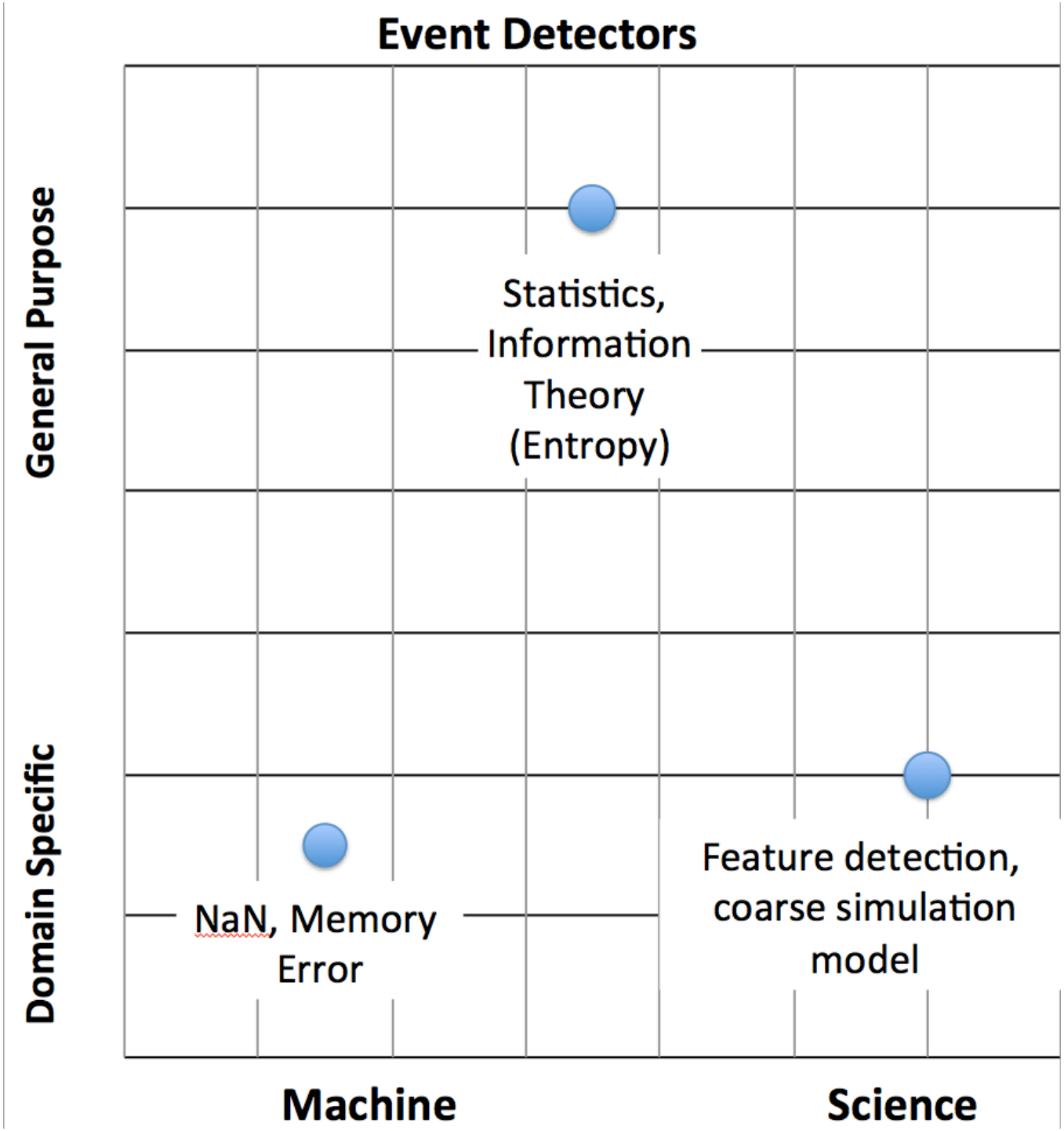
- Example: MPAS ocean code
  - Temperature variable value: -2C to 30C
    - Values are significantly more constrained spatially

Mean climate values

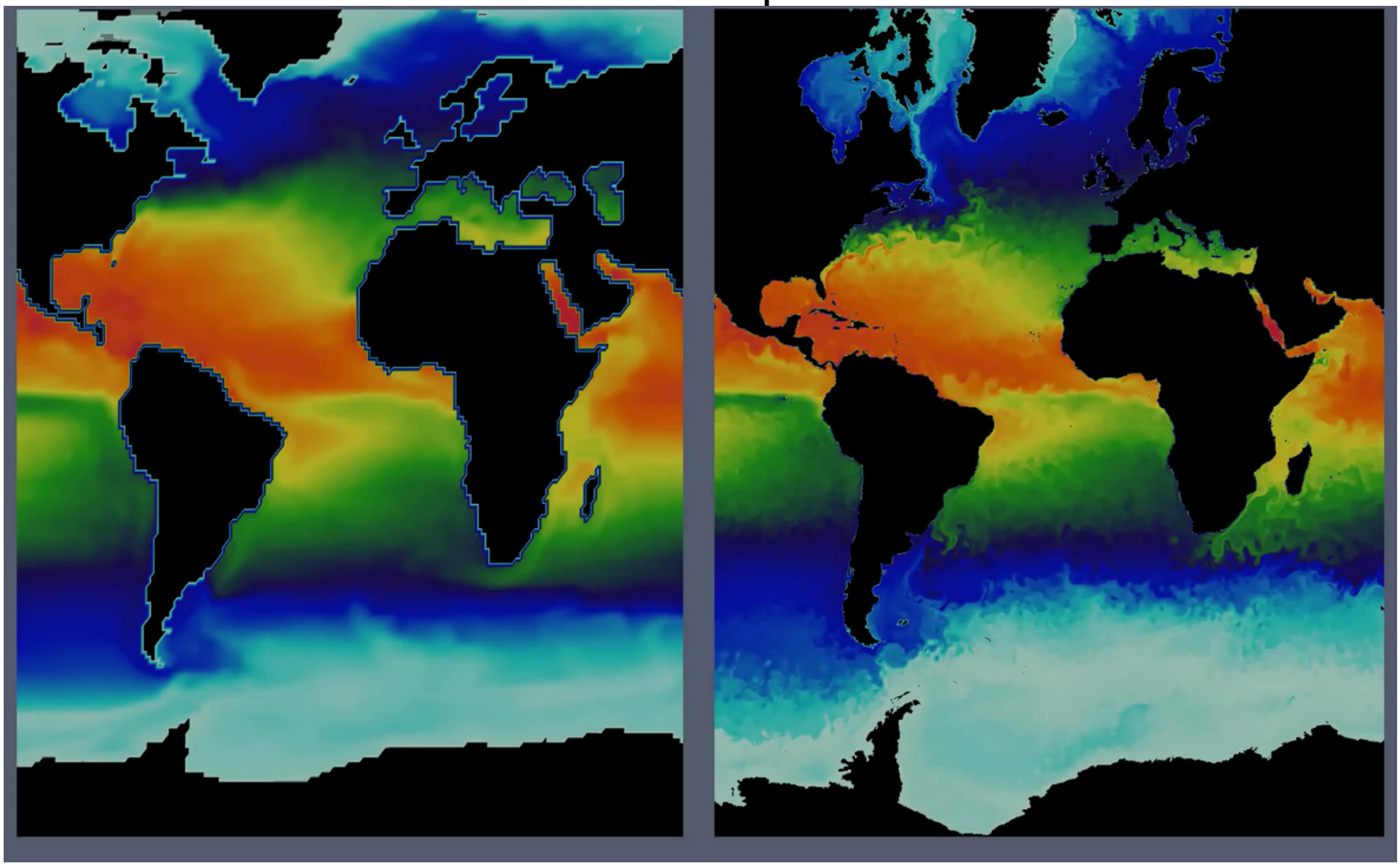
- Live spotlight dashboard for active simulation runs
  - Red > 10%
  - Yellow between 5-10%
  - Green < 5%
- Use case when coupling
  - Atmosphere, ocean

Philosophical question: What is a data value?

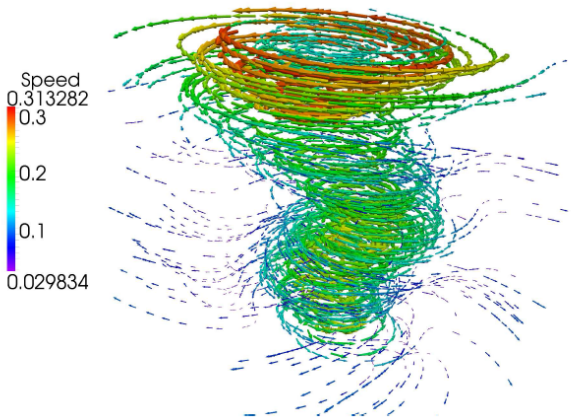
- Value at point? Spatial and temporal neighbors? Contribution to scientific feature? Model? Ensemble?

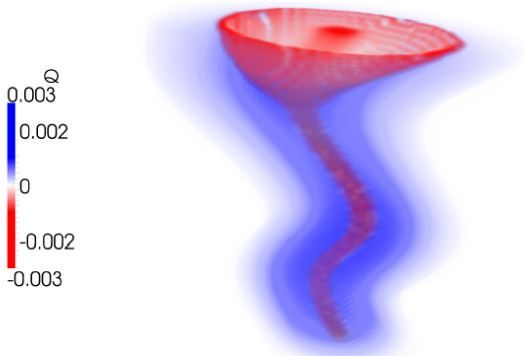

Event Detectors
General Purpose
Domain Specific
Statistics, Information Theory (Entropy)
NaN, Memory Error
Feature detection, coarse simulation model
Machine
Science

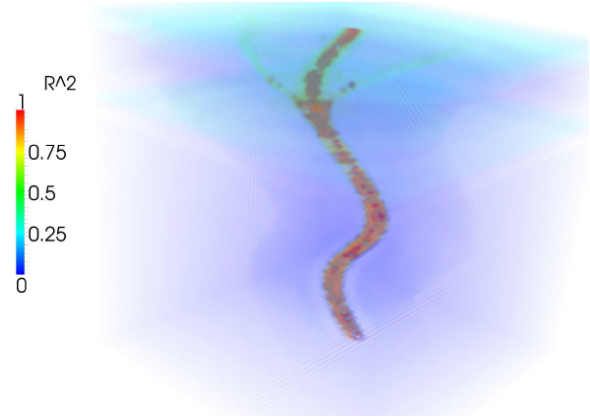
Event detection and characterization via model comparison

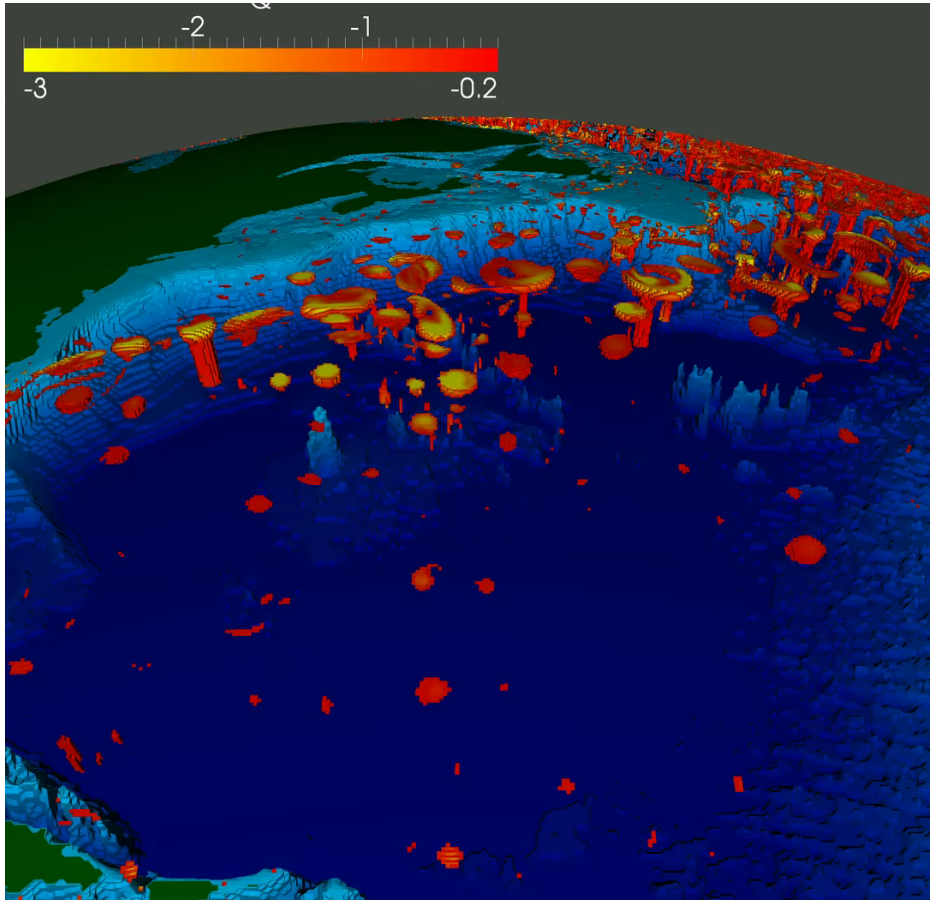


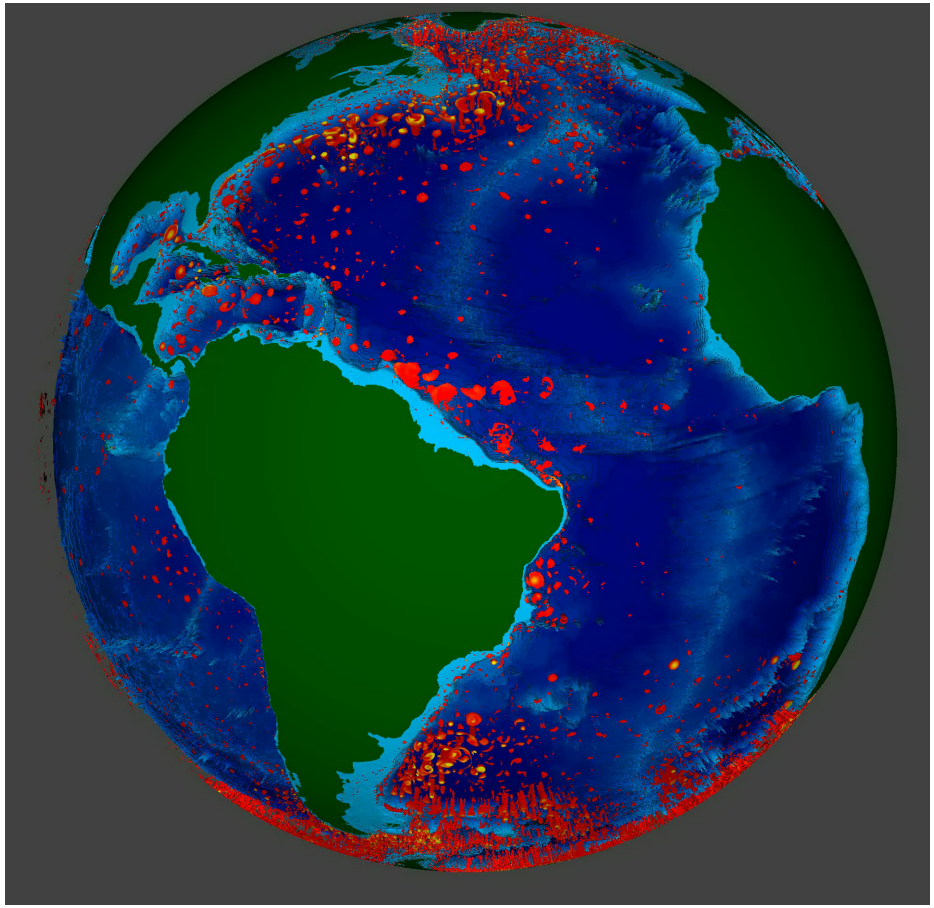
Event detection - Feature extraction and counting







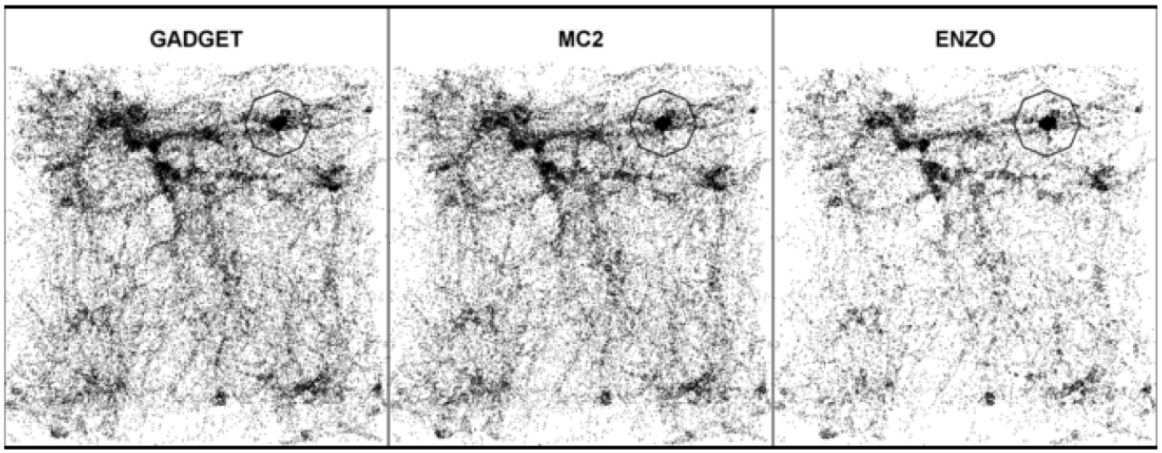




Evidence/feature-based verification process for cosmology

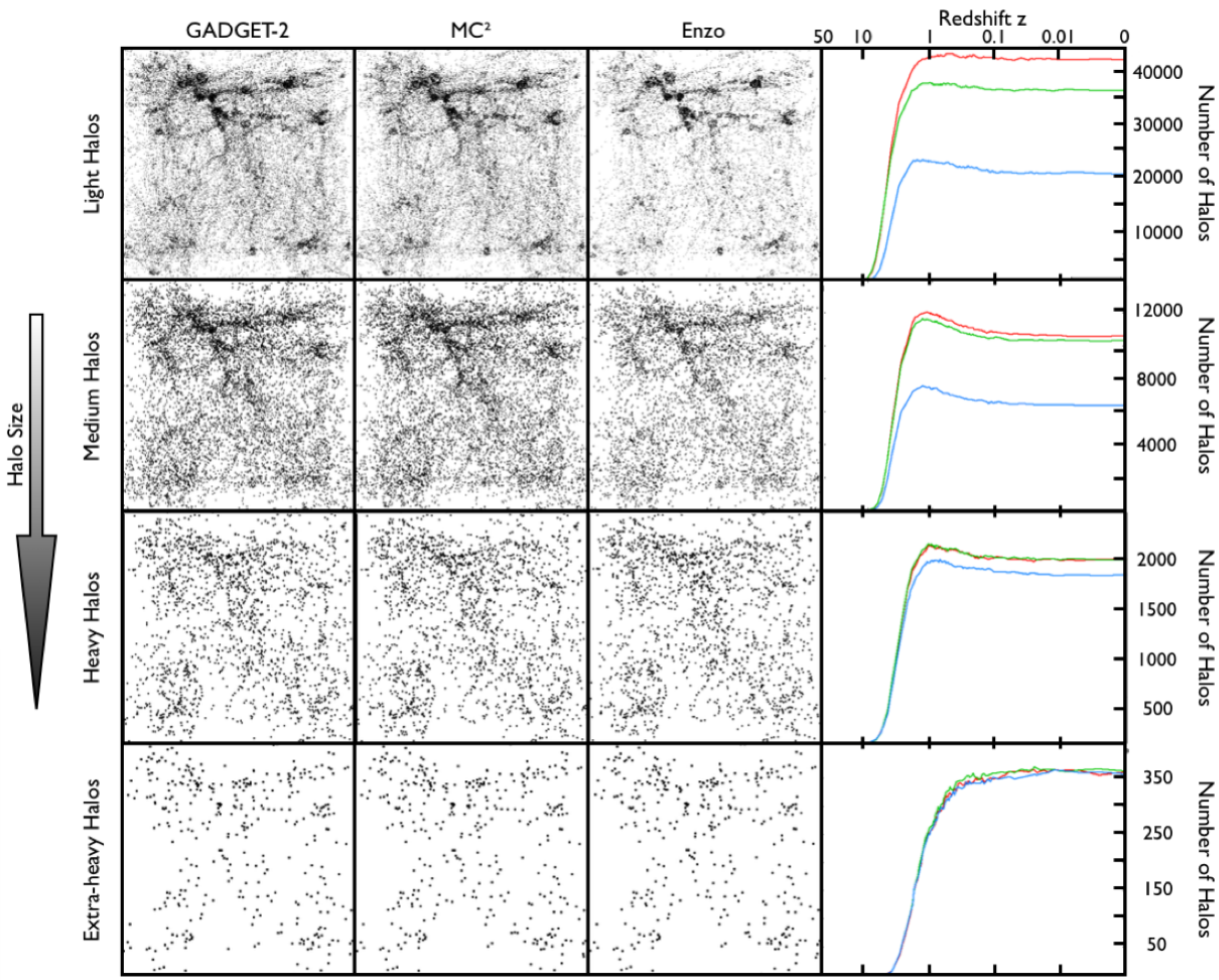
- Step 1: Define (or refine) measurable features
- Step 2: Formulate (or refine) a hypothesis about the measurable feature in the simulation codes
- Step 3: Qualitative comparative visualization
- Step 4: Quantitative comparative visualization
- Repeat starting at step 1 until the codes are verified

Hypothesis 1: An AMR code with a peak resolution equivalent to a uniform grid code should resolve all halos of interest --- find and count…



Evidence/feature-based verification process for cosmology

Hypothesis 2: --- The halos do not form at early times when the base resolution is still very low and cannot be recovered later - find, bin and count…



Measures and mappings

Data and compute representation

Libraries

Run time

HW Model

Component

Simulation

Multi-physics

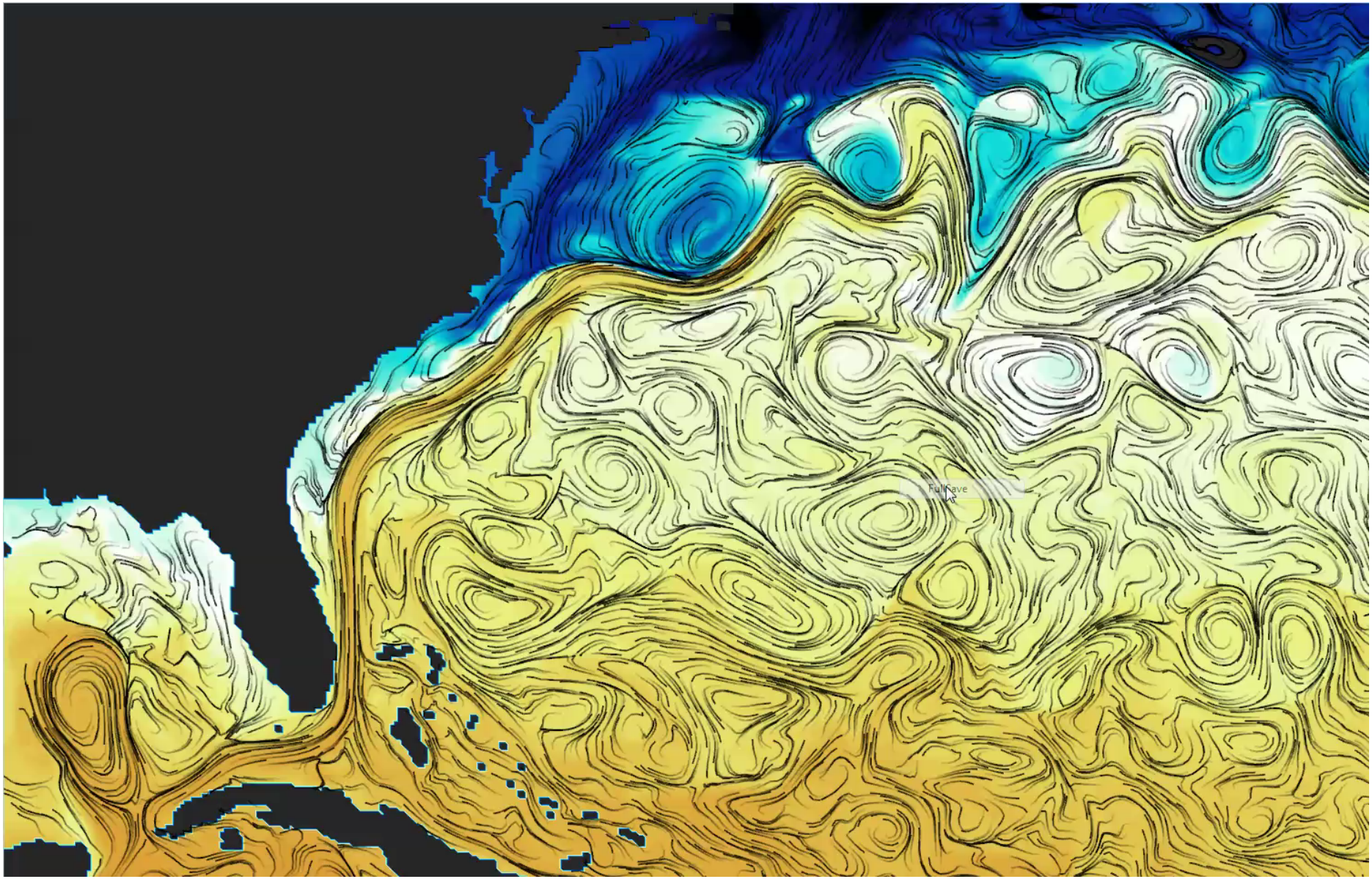
V&V

Machine

Science

Example: McCormick – Debugging a domain-specific language

Issue - Spatial and temporal granularity



Issue - Data size
Significant in situ data reduction

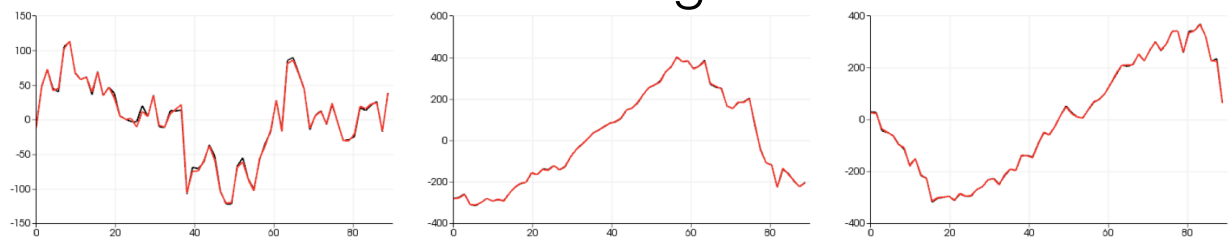
Algorithm	Reduction
Data parallelism	Handle large datasets Make reduction possible
Multi-resolution	Make focused exploration possible
Analysis operators	A dimension reduction
Statistical sampling	1-2 orders of magnitude
Compression	1 order of magnitude
Feature extraction	2 orders of magnitude

Issue - How accurate do we need to be? data reduction / quality

- Random sampling provides a data representation that is unbiased for statistical estimators, e.g., mean and others

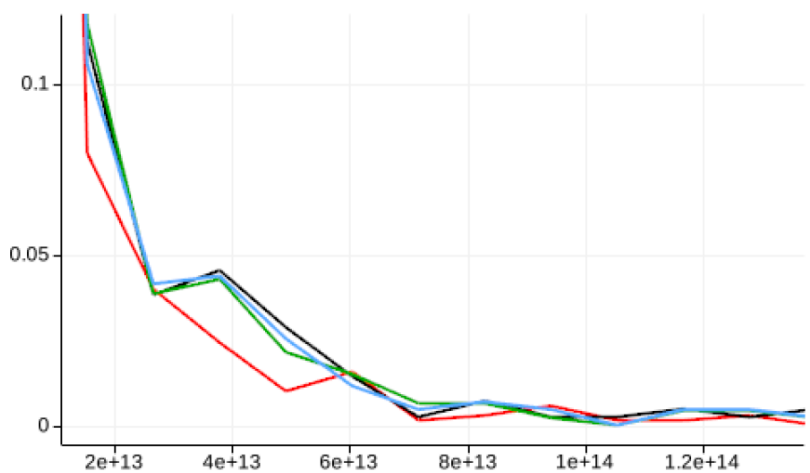
- Since the sampling algorithm is in situ: accuracy metric(simulation data, sampled representation)

Plotting



Red is 0.19% sample data, black is original simulation data

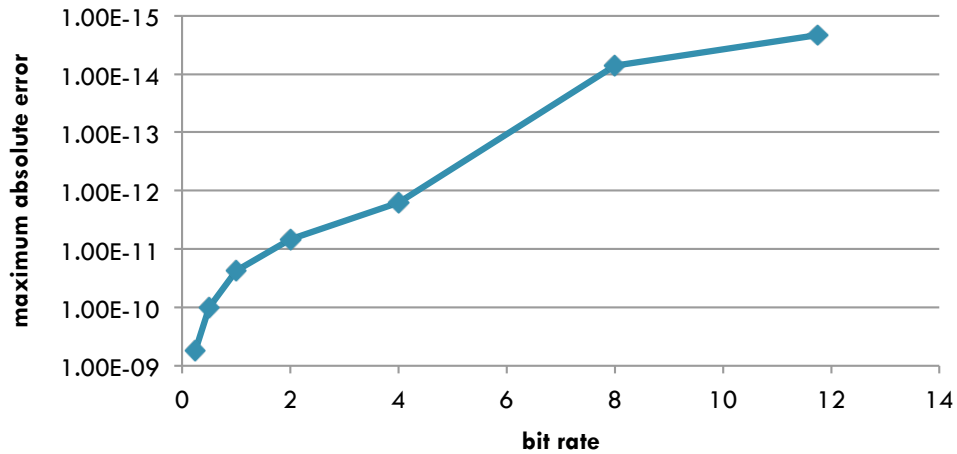
Feature Extraction: Halo Finding



. The red, green, and blue curves are 0.19%, 1.6%, and 12.5% samples. . The black curve is the original data. Calculate the halo mass function for different sample sizes of 256^3 particles

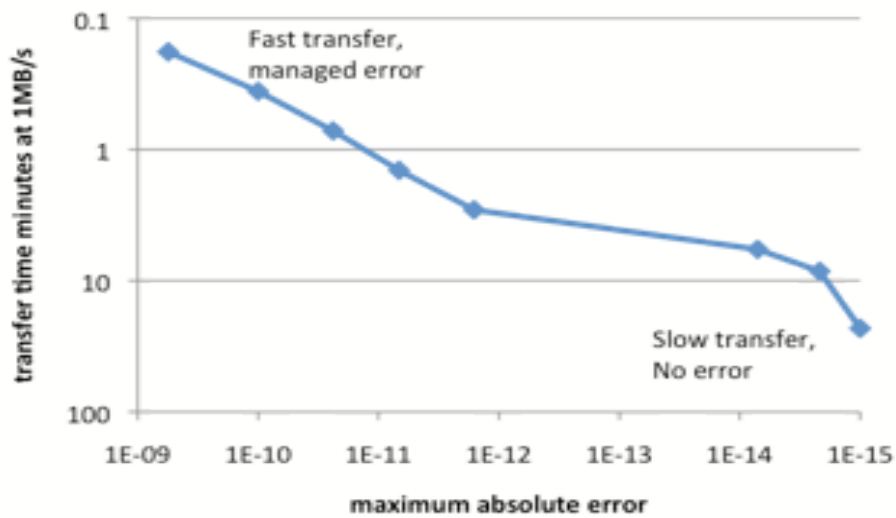
How accurate do we need to be?
In situ compression with quantified accuracy

- *In situ* compression of simulation data
  - Use JPEG 2000 to compress data
  - Quantify the maximum/L-infinity norm) data quality for scientific analysis



- Measure the maximum point error
  - Guarantee accuracy to x decimal places
  - Accuracy Metric

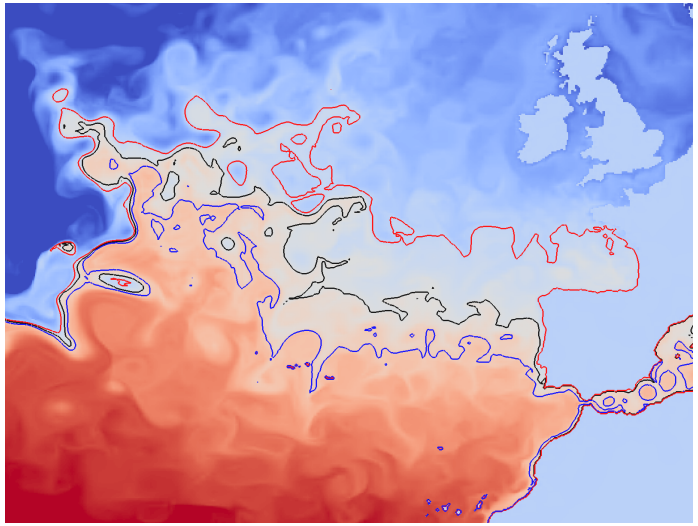
  (Simulation data – Compressed representation)

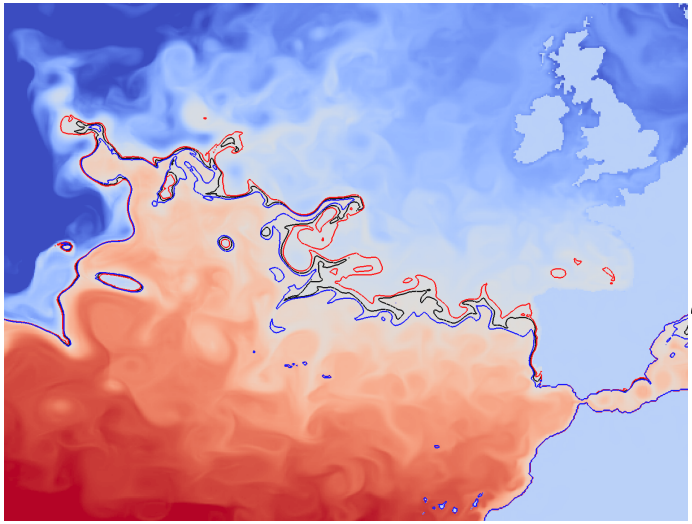


- User can trade read I/O time vs. data accuracy in a quantifiable manner

Isovalues on compressed simulation data with bounding error - (32 bits, 3200x2400x42, 1.4 GB)

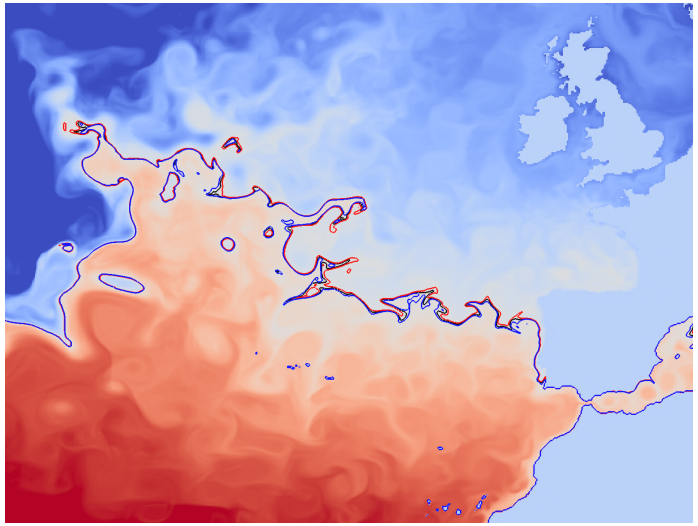
0.25 bits
10.8 MB

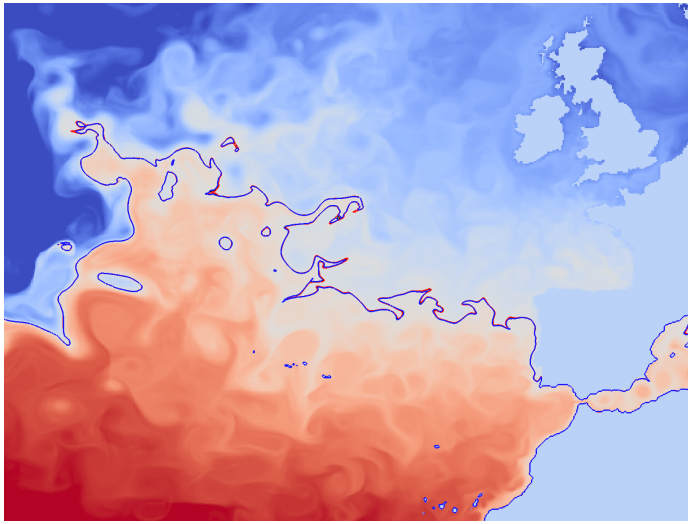




0.5 bits
21.6 MB

1.0 bits
43.3 MB





2.0 bits
86.5 MB

Event characterization

Event detected – What type?

<u>Fault</u>

Spatial locality? Memory based?
Not reproducible? – SDC?

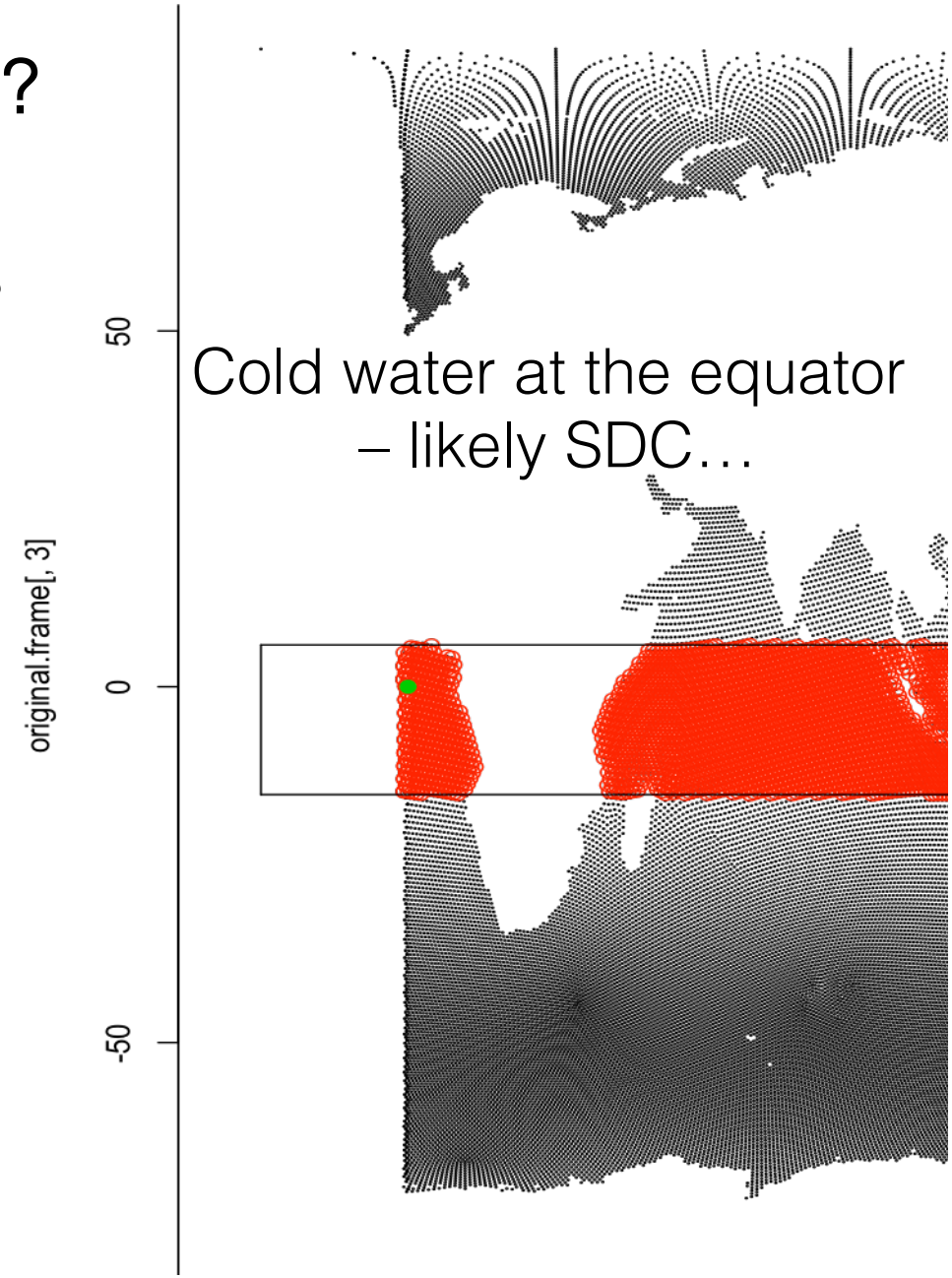
<u>Bug</u>

Reproducible? Violate data and compute representations?

<u>Science Result</u>

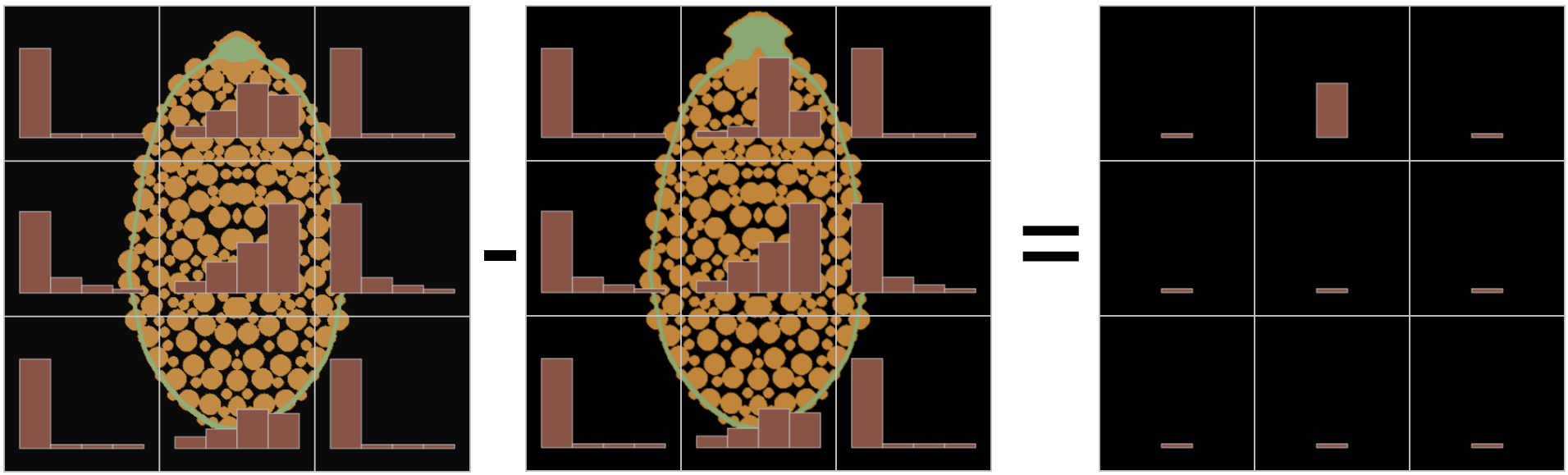
Reproducible? V&V?

<u>Create a log of events, characterizations and actions</u>



Cold water at the equator – likely SDC…

Event detection: automated algorithms
Adaptive focus based on selected events

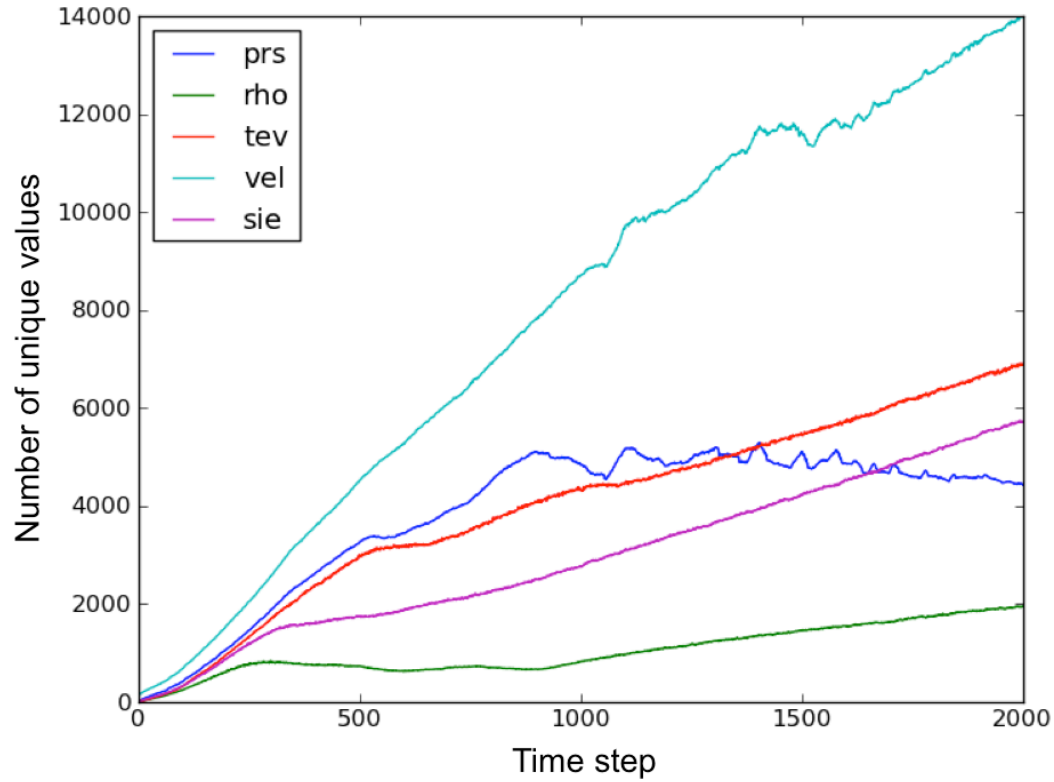


timestep$_n$ timestep$_{n+1}$ difference histogram

- Create adaptive analysis-based grid
  - Histogram at each grid element
    - Across all axises (spatial, value, multivariate)
- Use for spatial, temporal selection
  - Cameras, storage, feature identification

Event Detection - Statistical Measures

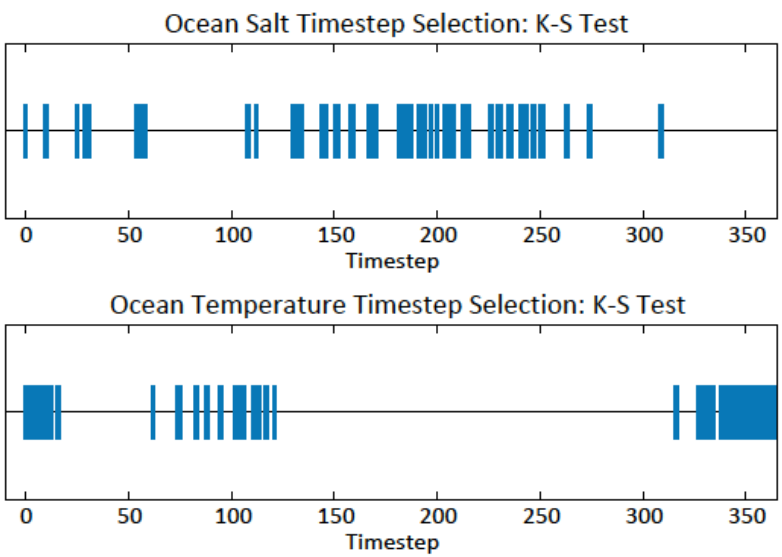
- Based on the number of unique data values
  - *u* - *N*umber of unique histogram values
    - Histogram across all axises (spatial, value, multivariate)
  - If the difference exceeds a threshold, detect event
- Currently exploring other more sophisticated statistical metrics
  - Kolmogorov–Smirnov distance metric



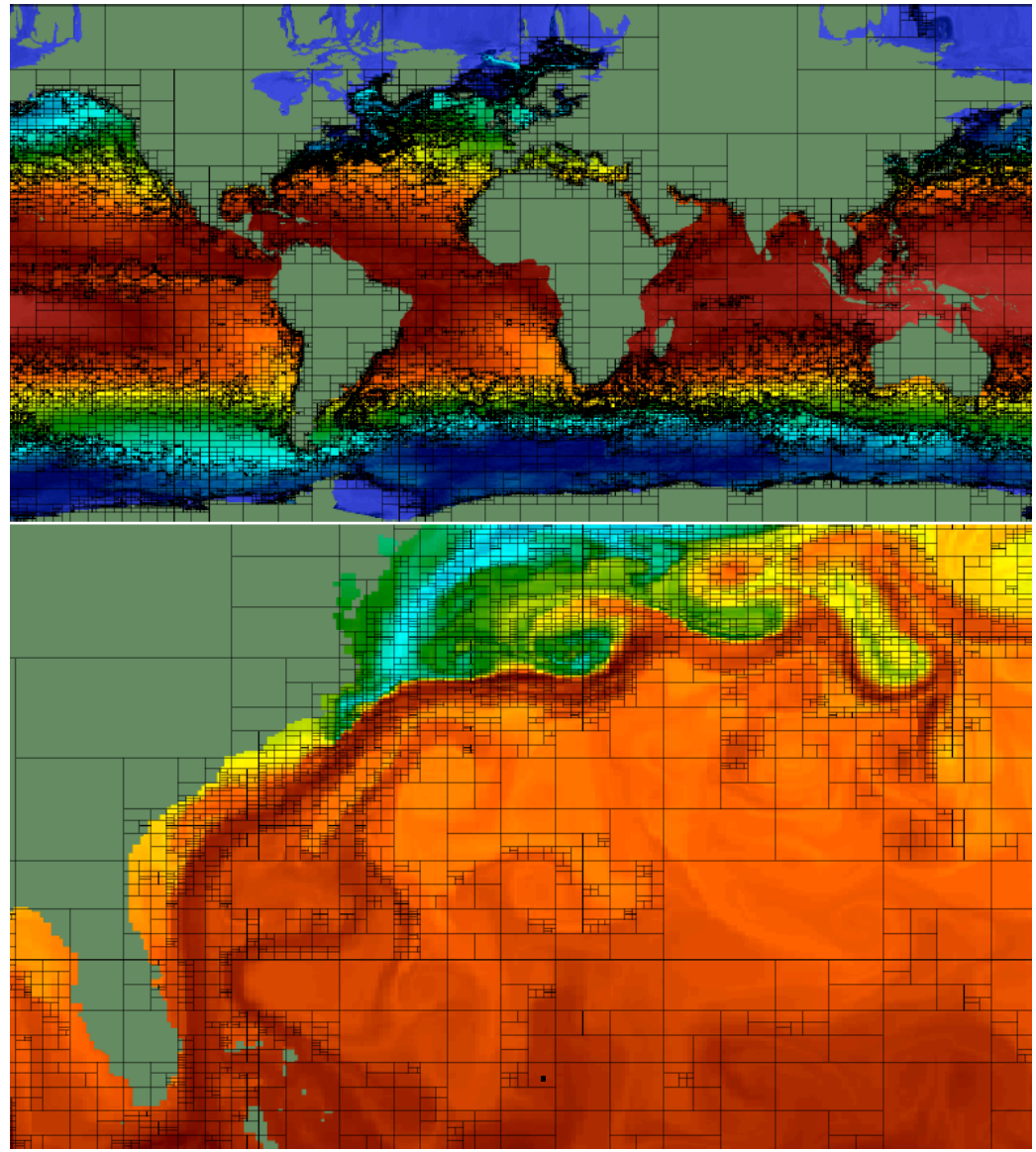
Event Detection: Sampling Using Analysis Driven Refinement (ADR)

- Recursive metric-based refinement
- Multidimensional

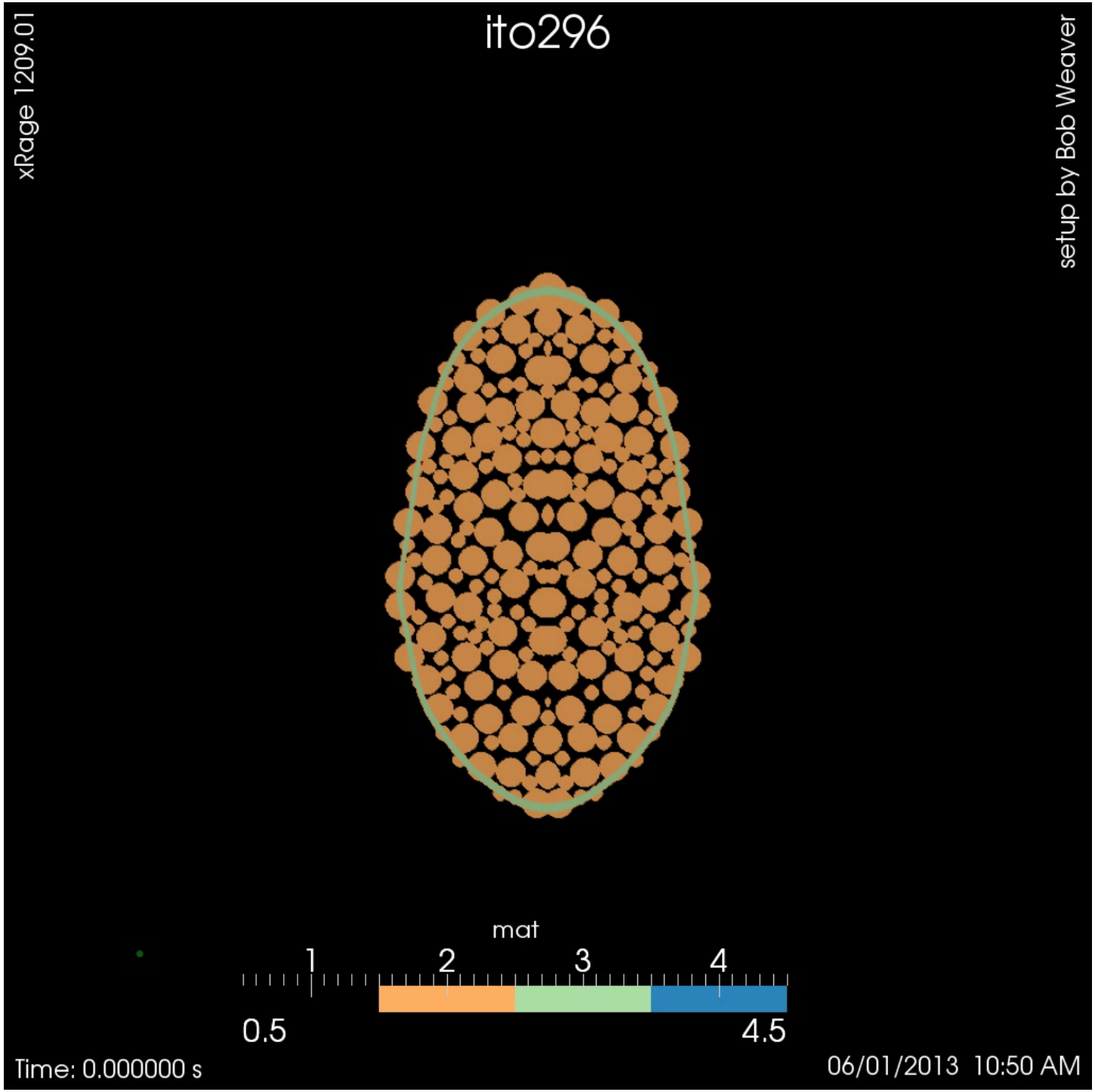
Entropy metric used on right, unique value below



Sampling in Time



Sampling in Space


ito296
xRage 1209.01
setup by Bob Weaver
mat
1
2
3
4
0.5
4.5
Time: 0.000000 s
06/01/2013 10:50 AM

Data reduction and event response: Image database approach - Cinema

<u>Challenge</u>

In situ is a batch process

Concern that exploratory aspect of analysis will be lost

<u>Idea</u>

Store *many* images that sample the visualization parameter space

In less than the space needed for a single scientific data dump

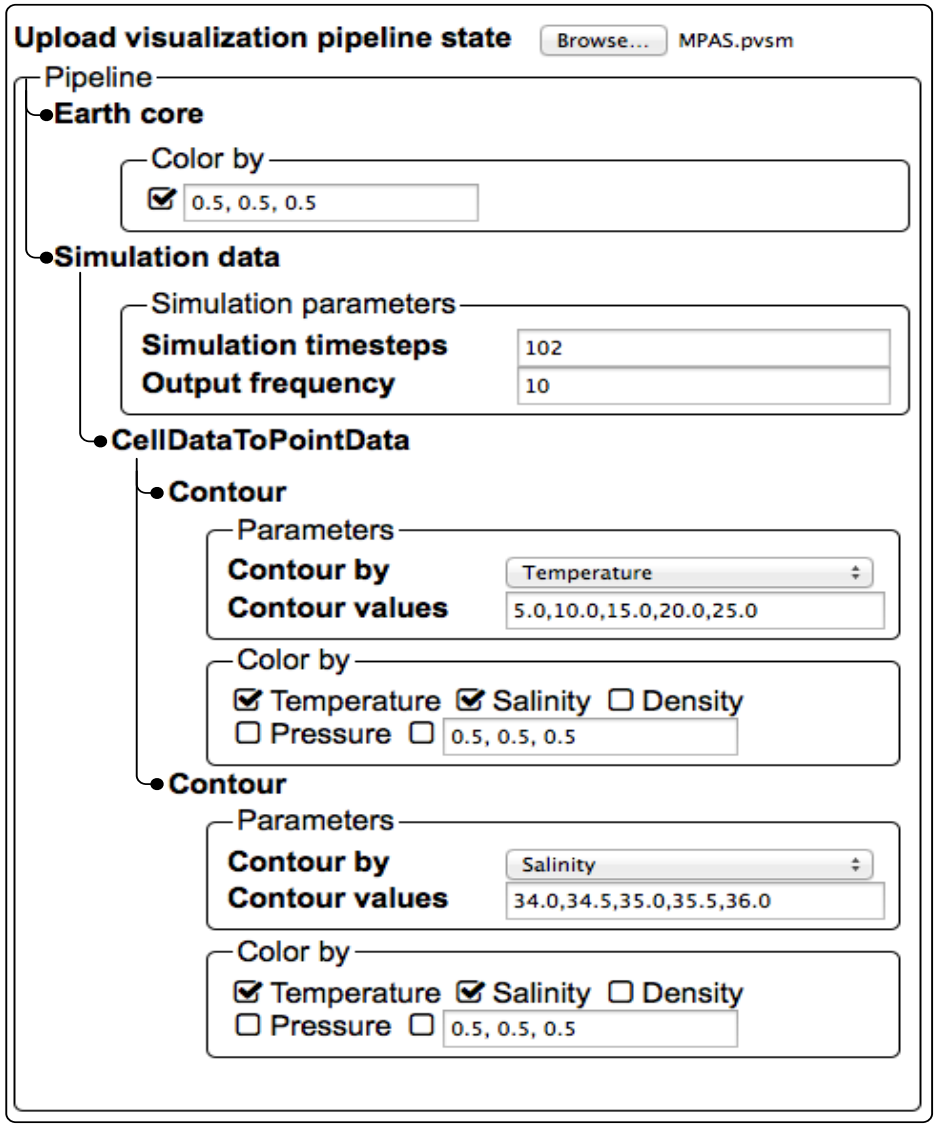
Ex: Cameras, operations, parameters

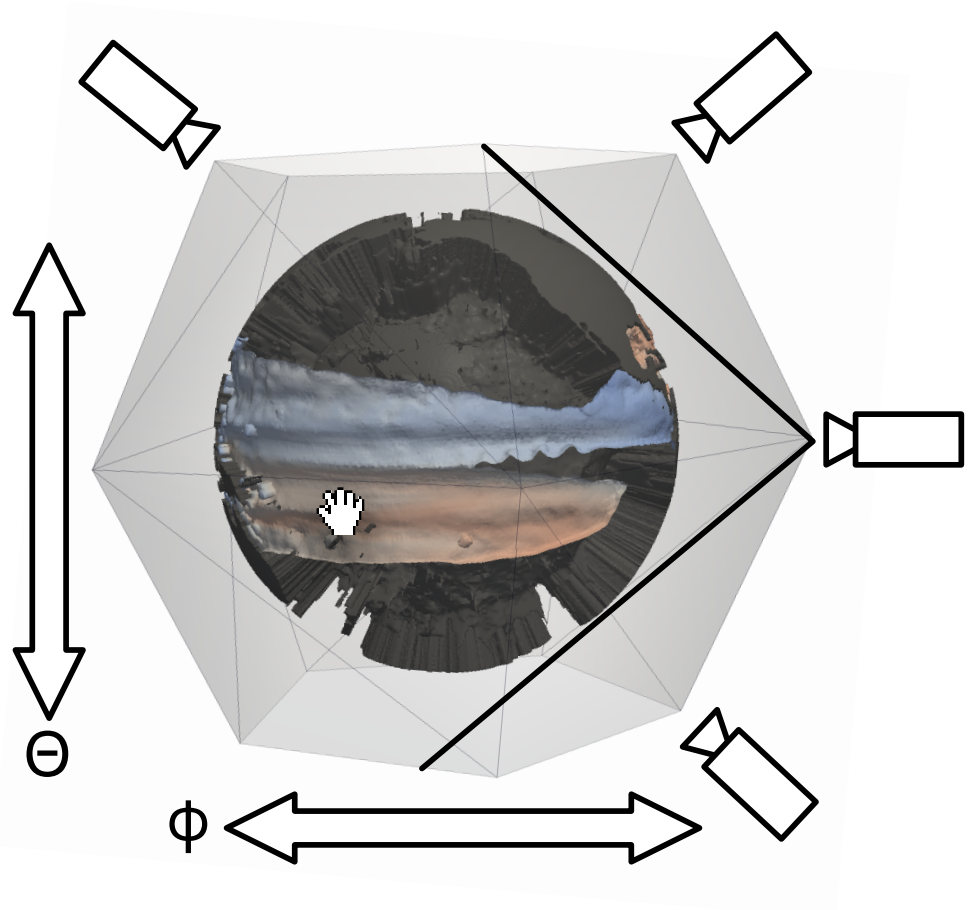
Create an image database from in situ analysis

Post-processing exploration of image database

Mega	Giga	Tera	Peta	Exa
10^6	10^9	10^12	10^15	10^18
Image speed	Storage & network speed	Operations speed	Operations speed	Operations speed

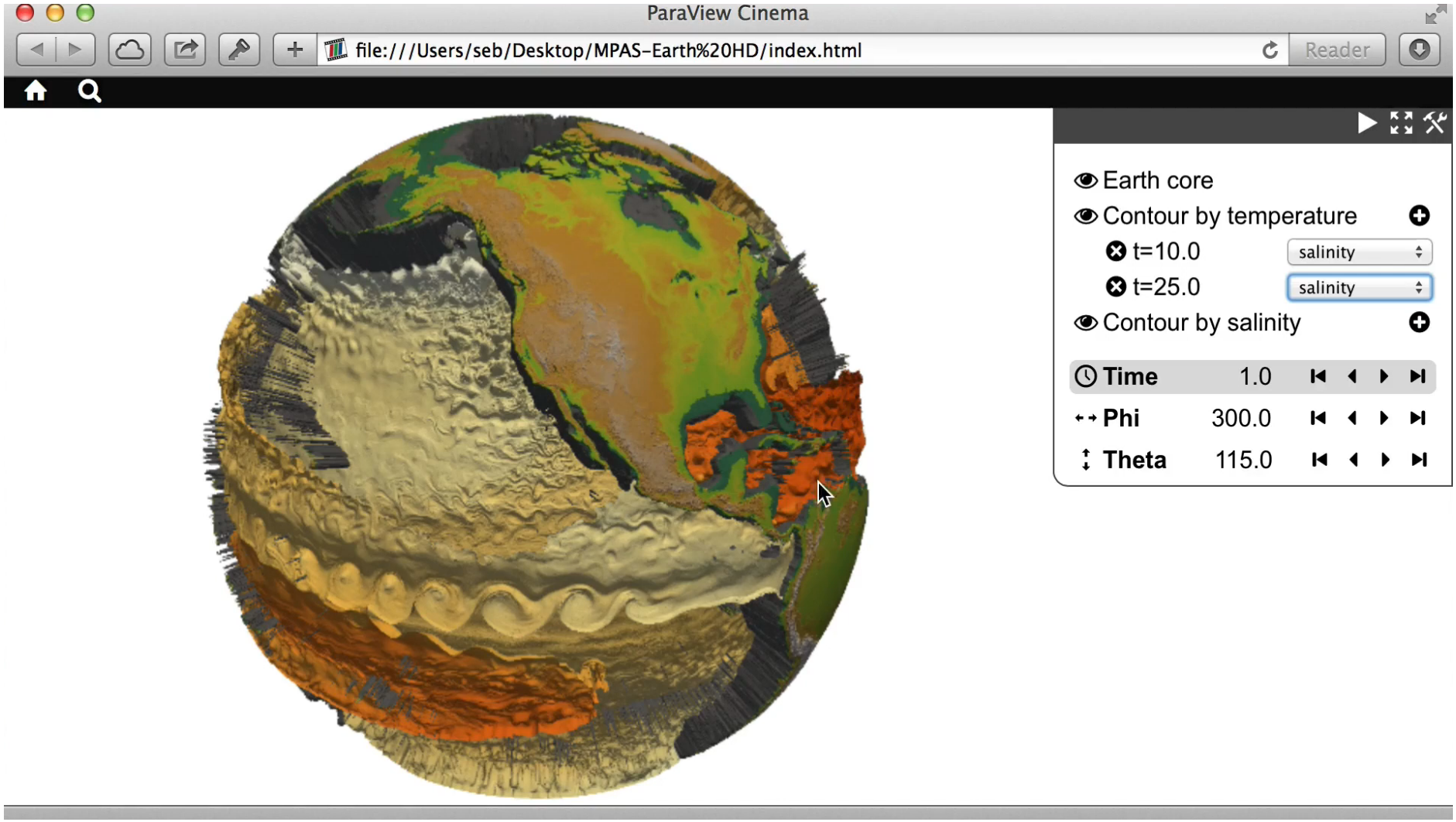
Event responses





Set camera and operator parameters to visualize

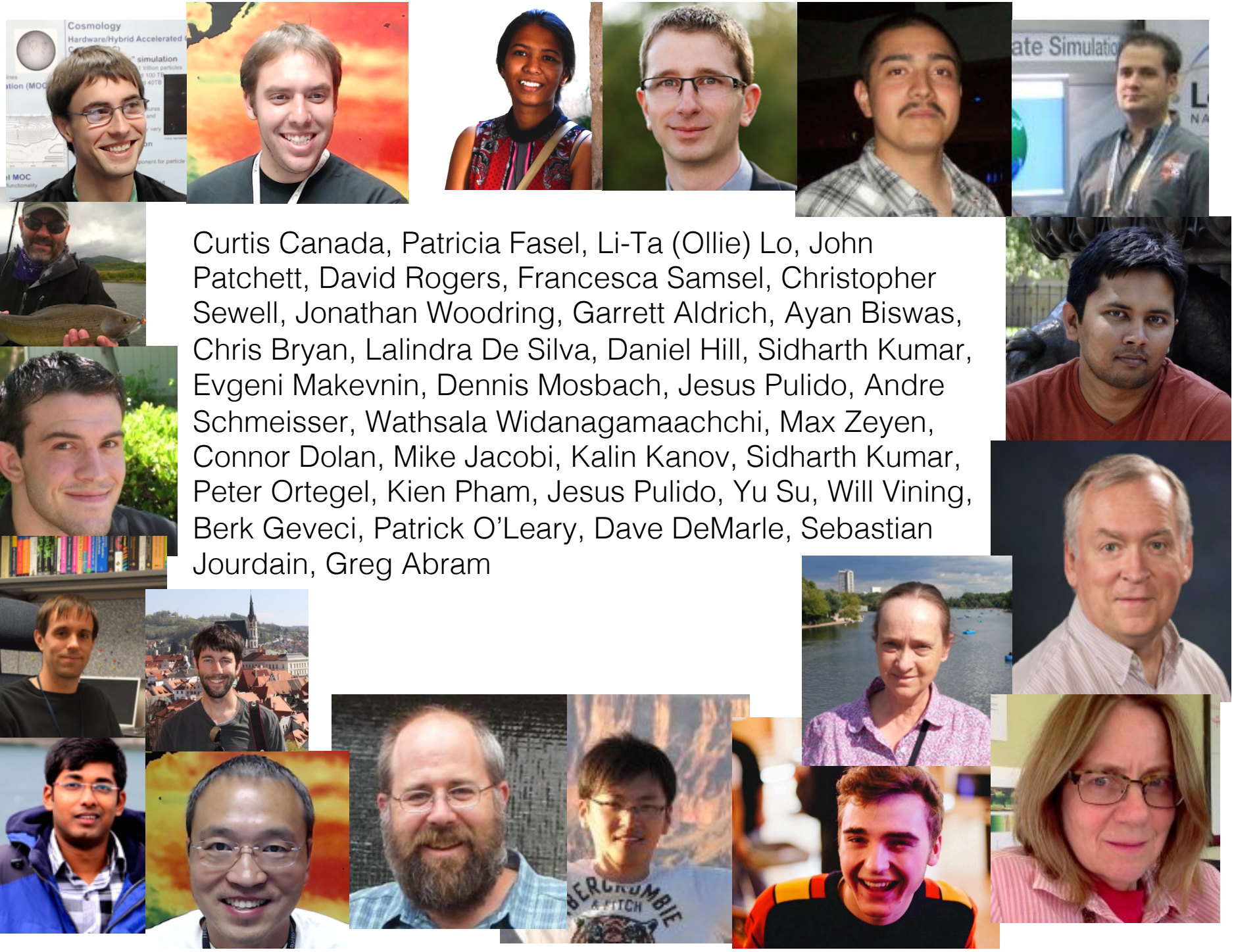
Use Case – Traditional interactive exploration



In this video:
Processing, combining and showing images from the image database
No raw scientific data is read, no geometry is created during viewing

Conclusions

- Joint effort between CS areas and science communities
- Tools to detection, characterize and respond
  - Support mapping up and down the stack
- http://datascience.lanl.gov



Curtis Canada, Patricia Fasel, Li-Ta (Ollie) Lo, John Patchett, David Rogers, Francesca Samsel, Christopher Sewell, Jonathan Woodring, Garrett Aldrich, Ayan Biswas, Chris Bryan, Lalindra De Silva, Daniel Hill, Sidharth Kumar, Evgeni Makevnin, Dennis Mosbach, Jesus Pulido, Andre Schmeisser, Wathsala Widanagamaachchi, Max Zeyen, Connor Dolan, Mike Jacobi, Kalin Kanov, Sidharth Kumar, Peter Ortegel, Kien Pham, Jesus Pulido, Yu Su, Will Vining, Berk Geveci, Patrick O'Leary, Dave DeMarle, Sebastian Jourdain, Greg Abram

Publications

- B. Nouanesengsy, J. Woodring, K. Myers, J. Patchett, and J. Ahrens, “ADR Visualization: A Generalized Framework for Ranking Large-Scale Scientific Data using Analysis-Driven Refinement”, LDAV 2014, November 2014, Paris, France.
- K. Myers, E. Lawrence, M. Fugate, J. Woodring, J. Wendelberger, and J. Ahrens, “An In Situ Approach for Approximating Complex Computer Simulations and Identifying Important Time Steps”, in submission, arXiv: 1409.0909.
- A. Biswas, S. Dutta, H.-W. Shen, J. Woodring. “An Information-Aware Framework for Exploring Multivariate Data Sets.” IEEE Visualization 2013, Atlanta, GA, November, 2013.
- Y. Su, G. Agrawal, J. Woodring, K. Myers, J. Wendelberger and J. Ahrens, "Effective and Efficient Data Sampling Using Bitmap Indices", Cluster Computing, March 2014.
- Y. Su, G. Agrawal, J. Woodring, A. Biswas and H.-W. Shen, "Supporting Correlation Analysis on Scientific Datasets in Parallel and Distributed Settings", in Proceedings of the International ACM Symposium on High-Performance Parallel and Distribued Computing (HPDC'14), June 2014, Vancouver, Canada.
- Y. Su, G. Agrawal, J. Woodring, K. Myers, J. Wendelberger and J. Ahrens. “Taming Massive Distributed Datasets: Data Sampling Using Bitmap Indices.” In Proceedings of the International ACM Symposium on High-Performance Parallel and Distributed Computing (HPDC’13), New York, NY, USA, June 2013.
- Y. Su, G. Agrawal, and J. Woodring, “Indexing and Parallel Query Processing Support for Visualizing Climate Datasets”, Proceedings of the 41st International Conference on Parallel Processing, Pittsburgh, PA, Sept. 2012.
- J. Ahrens, S. Jourdain, P. O'Leary, J. Patchett, D. H. Rogers, M. Petersen, “An Image-based Approach to Extreme Scale In Situ Visualization and Analysis”, Supercomputing 2014, New Orleans.

MPAS Ocean initial conditions without smoothing