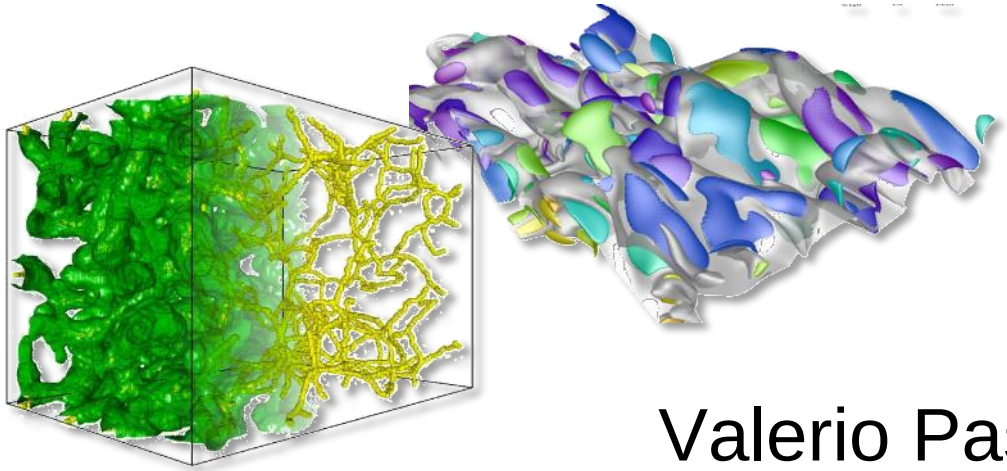
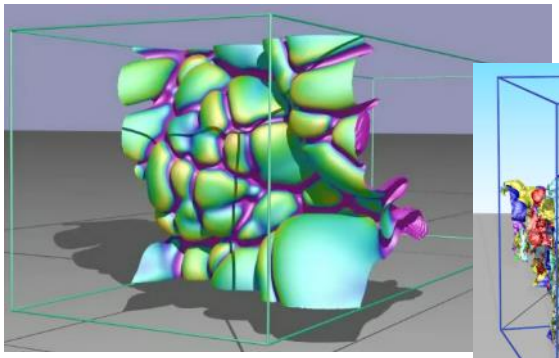
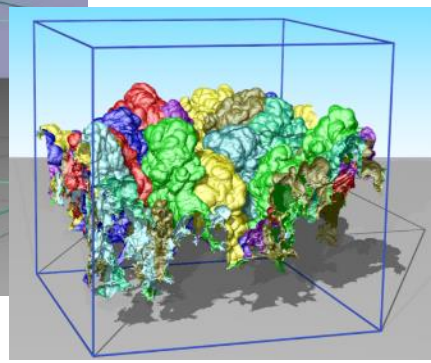
The Big Gift of Big Data

Valerio Pascucci

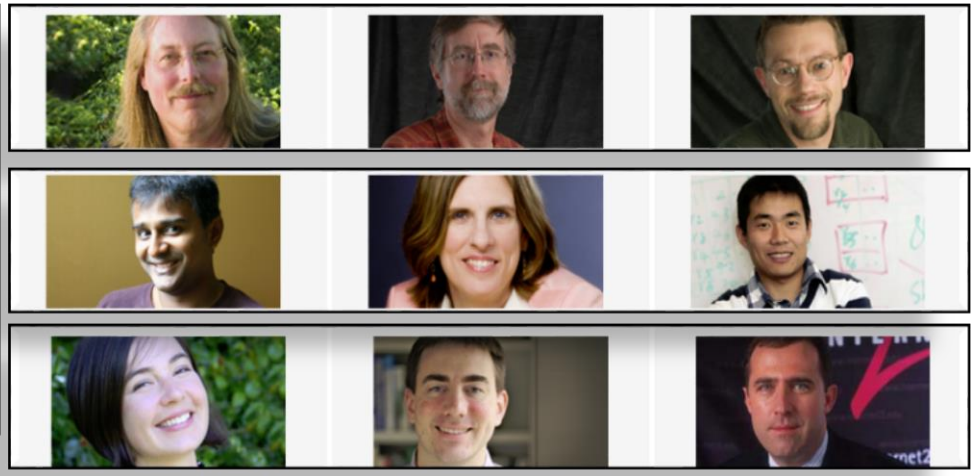
Director, Center for Extreme Data Management Analysis and Visualization

Professor, SCI institute and School of Computing, University of Utah

Laboratory Fellow, Pacific Northwest National Laboratory

Center for Extreme Data Management, Analysis, and Visualization

- 10 Faculty + scientists, developers, students, …
- Primary partners: UU, PNNL, LLNL
- Other partnerships: NSA, INL, ANL, ….
- Involvement in national Initiatives



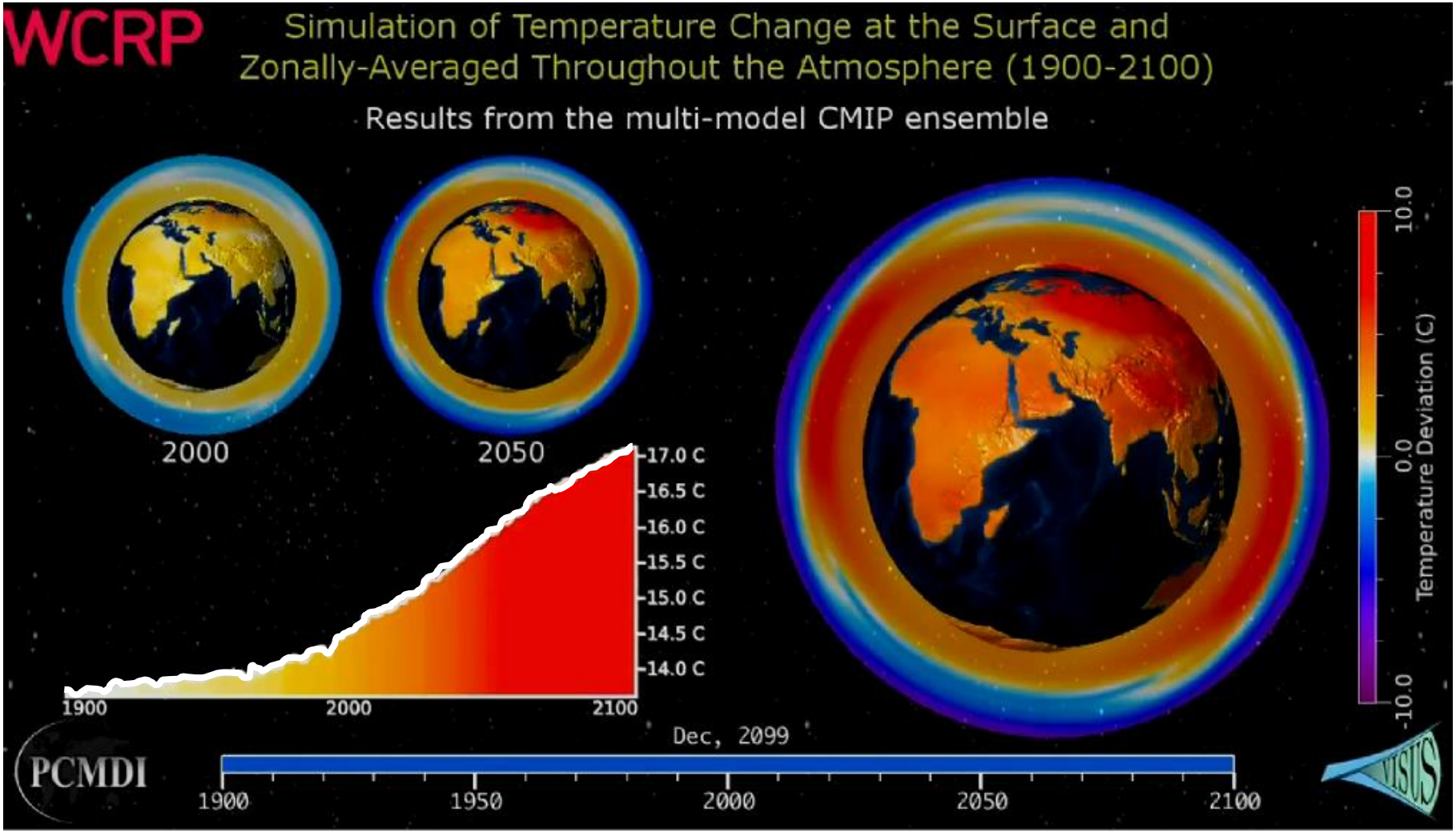
$1.6B NSA data center
(1.5 million-square-foot facility)

What is Big Data?

Big Data is like teenage sex:
everyone Talks About It,
nobody really knows how to do it,
everyone thinks everyone else is doing it,
so everyone claims they are doing it ….
(Dan Ariely)

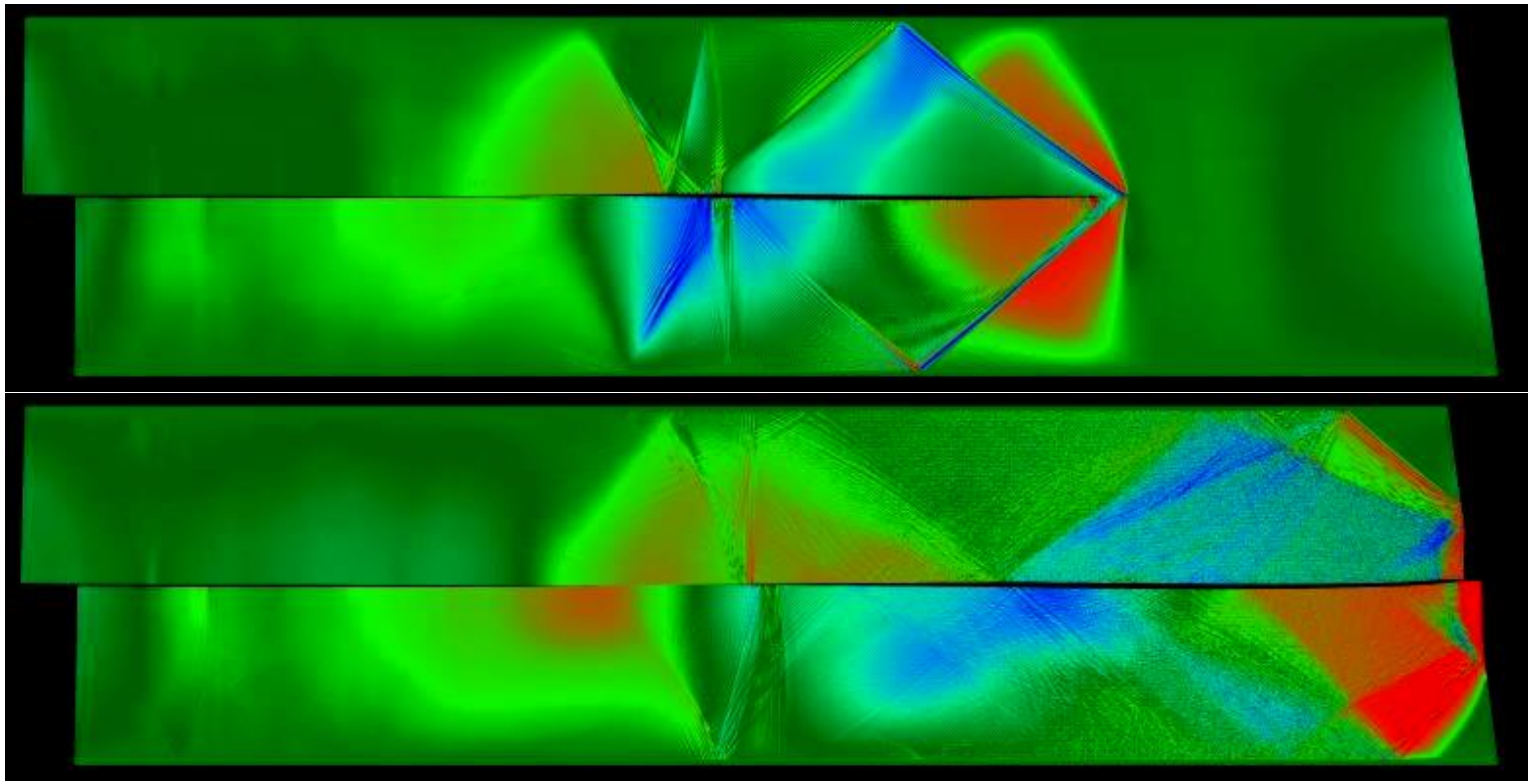
…. eventually everyone will do it!

Are Global Warming Trends Associated to Human Activities?



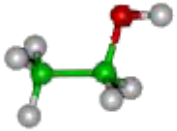
Can Devastating Material Failures Such as Earthquakes Travel at Supersonic Speed?

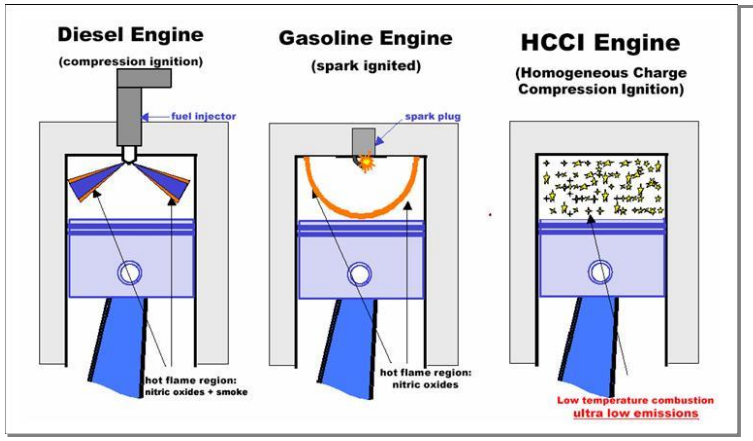
- Understanding the strength of new materials is critical in creating structures as small as microprocessors, buildings, or airplanes that withstand real-world forces

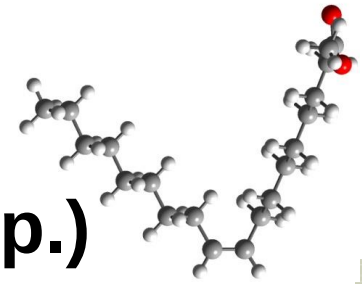


Change the World of Fuels and Engines to Increase Efficiency and Reduce Pollution

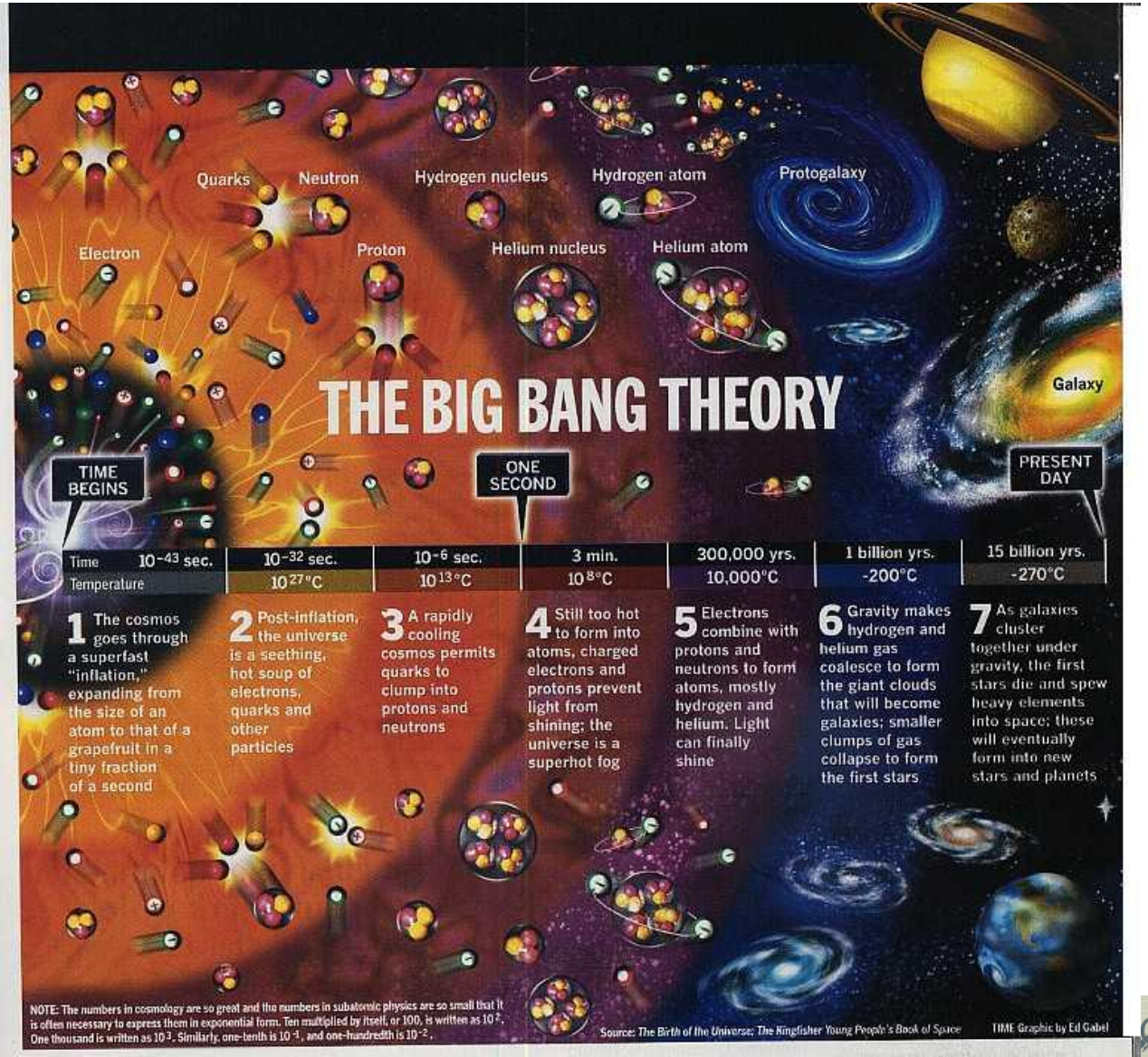
- **Fuel streams are rapidly evolving:**
  - **Heavy hydrocarbons**
    - **Oil sands**
    - **Oil shale**
    - **Coal**
  - **New renewable fuel sources**
    - **Ethanol**
    - **Biodiesel**
- **New engine technologies:**
    - **Direct Injection (DI)**
    - **Homogeneous Charge Compression Ignition (HCCI)**
    - **Low-temperature combustion**
- **Mixed modes of combustion (dilute, high-pressure, low-temp.)**



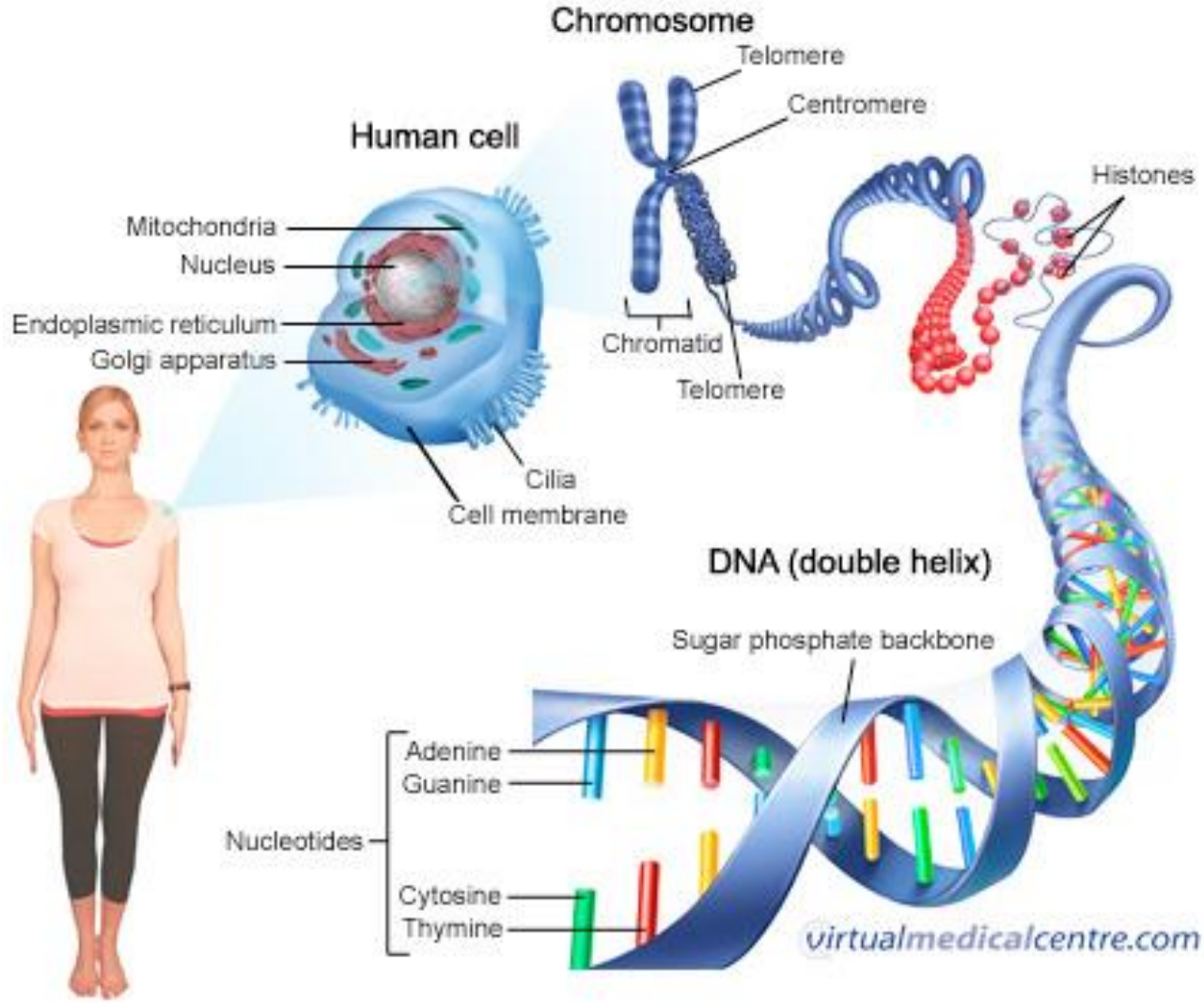




Can We Explain the Origin and the Evolution of the Universe?

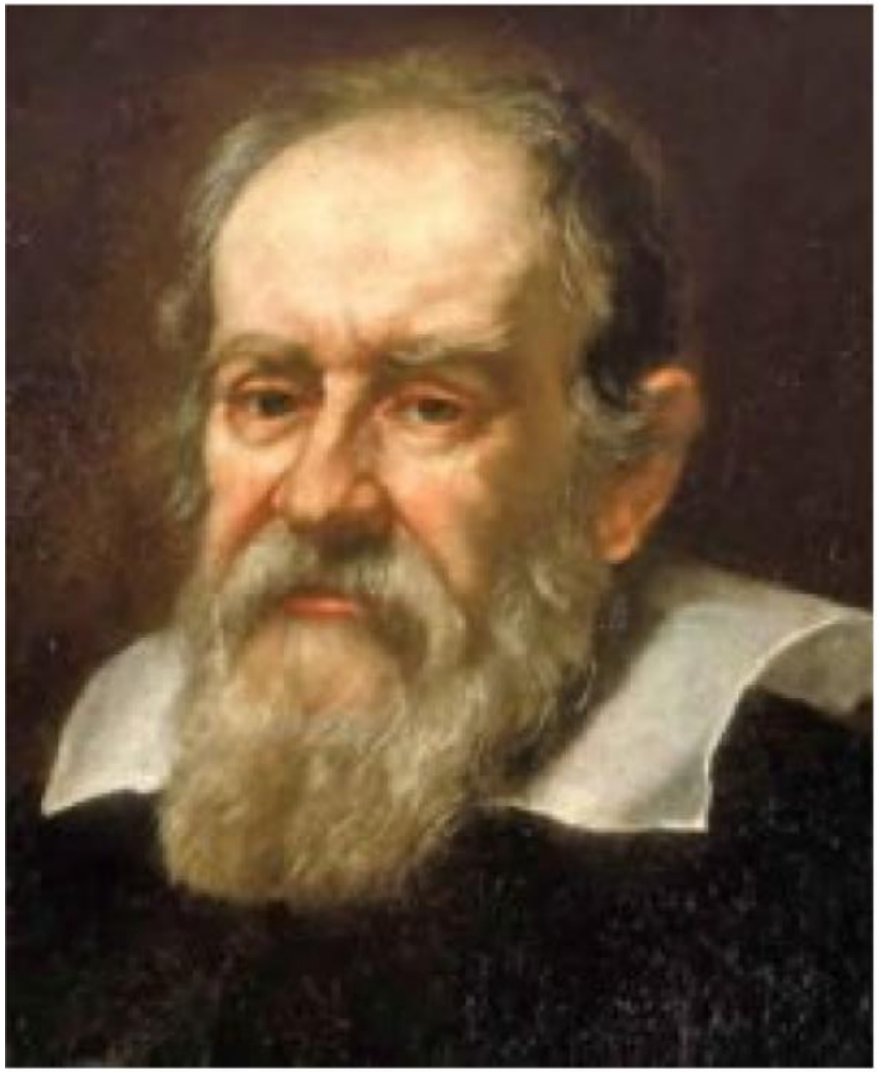


Can We Develop a New Healthcare Process that is Fully Personalized



Intermezzo to talk about a convicted felon

Galileo Galilei



Jailed in Apr 12, 1633 because

"gravely suspect of heresy“

The case was quickly reviewed by the Catholic Church

Received excuses in 1992

1564-1642

Galileo Galilei Led the Modern Revolution of the "Scientific Method"

Make observations

Propose a hypothesis

Design and perform an experiment to test the hypothesis

Analyze your data to determine whether to accept or reject the hypothesis

If necessary, iterate (propose and test a new hypothesis)

Galileo Galilei Led the Modern Revolution of the "Scientific Method"

Make observations

Propose a hypothesis

Design and perform an experiment to test the hypothesis

Analyze your data to determine whether to accept or reject the hypothesis

If necessary, iterate (propose and test a new hypothesis)

Galileo Galilei Led the Modern Revolution of the "Scientific Method"

Make observations

Propose a hypothesis

Design and perform a repetible experiment to test the hypothesis

Analyze your data to determine whether to accept or reject the hypothesis

If necessary, iterate (propose and test a new hypothesis)

Galileo Galilei Led the Modern Revolution of the "Scientific Method"

ons

Pro

repetible

ypothesis

Ar

mine whether

ypothesis

rate

hypothesis)

Galileo Galilei Led the Modern Revolution of the "Scientific Method"

Make observations

Propose a hypothesis of theoretical model

Design and perform a repetible experiment to test the hypothesis

Analyze your data to determine whether to accept or reject the hypothesis

If necessary, iterate (propose and test a new hypothesis)

Galileo Galilei Led the Modern Revolution of the "Scientific Method"

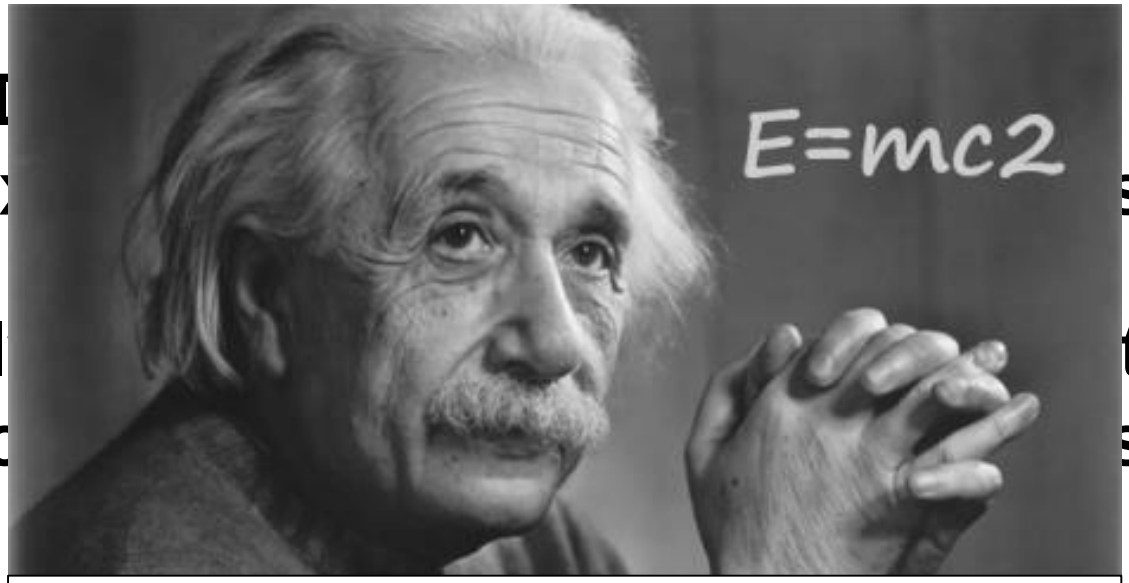
Make observations

Propose a hypothesis of theoretical model

D
ex **s**

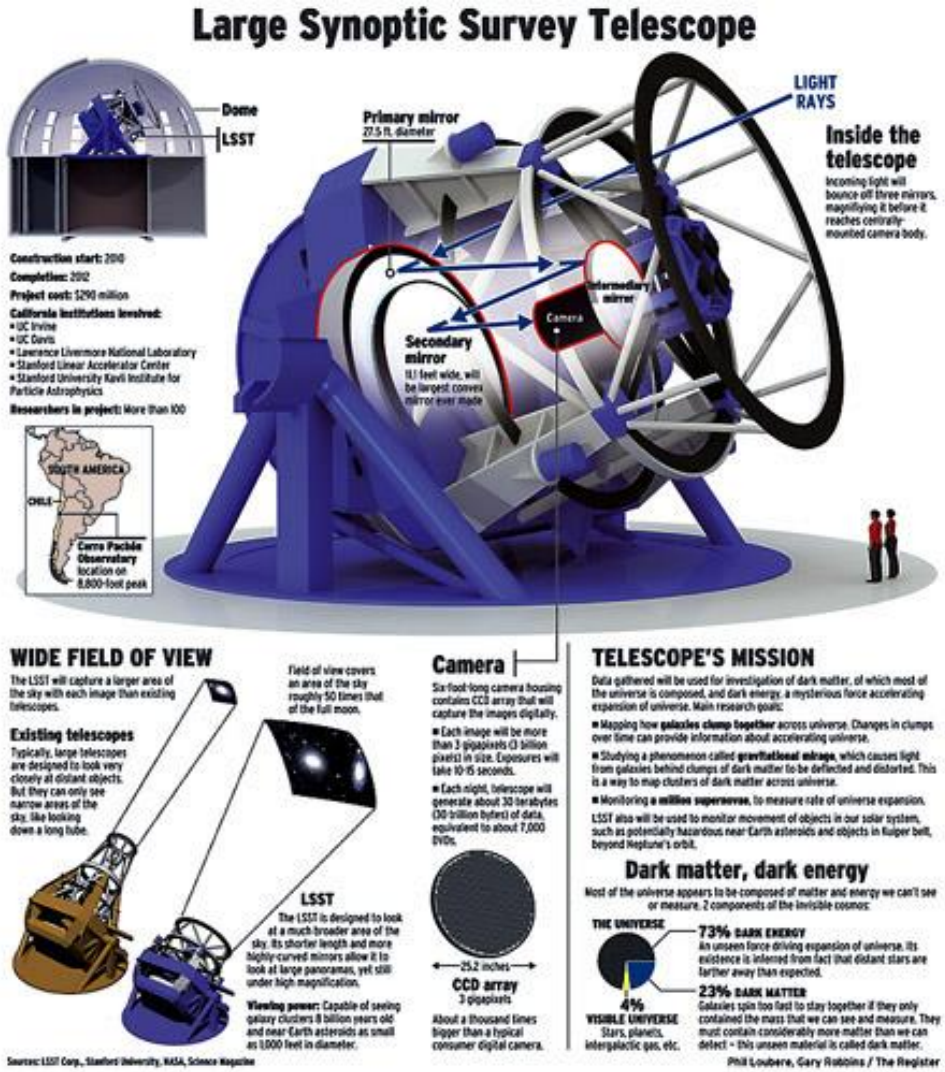
Anal **ther**
to **s**

(pr **is)**



Development of formal model description

New Data Collection and Processing Resources Challenged This Model



Large Synoptic Survey Telescope

15TB/day

100PB

in 10-year

New Data Collection and Processing Resources Challenged This Model

- **PayPal's Data Volumes**

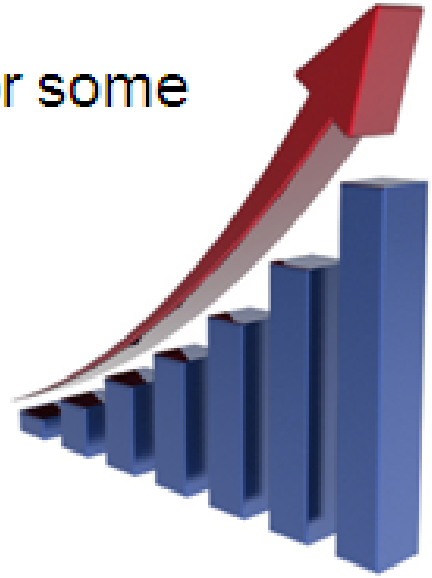
10 million+ logins / day

13 million financial transactions / day

300 variables calculated per event for some models.

~4 Billion inserts / day

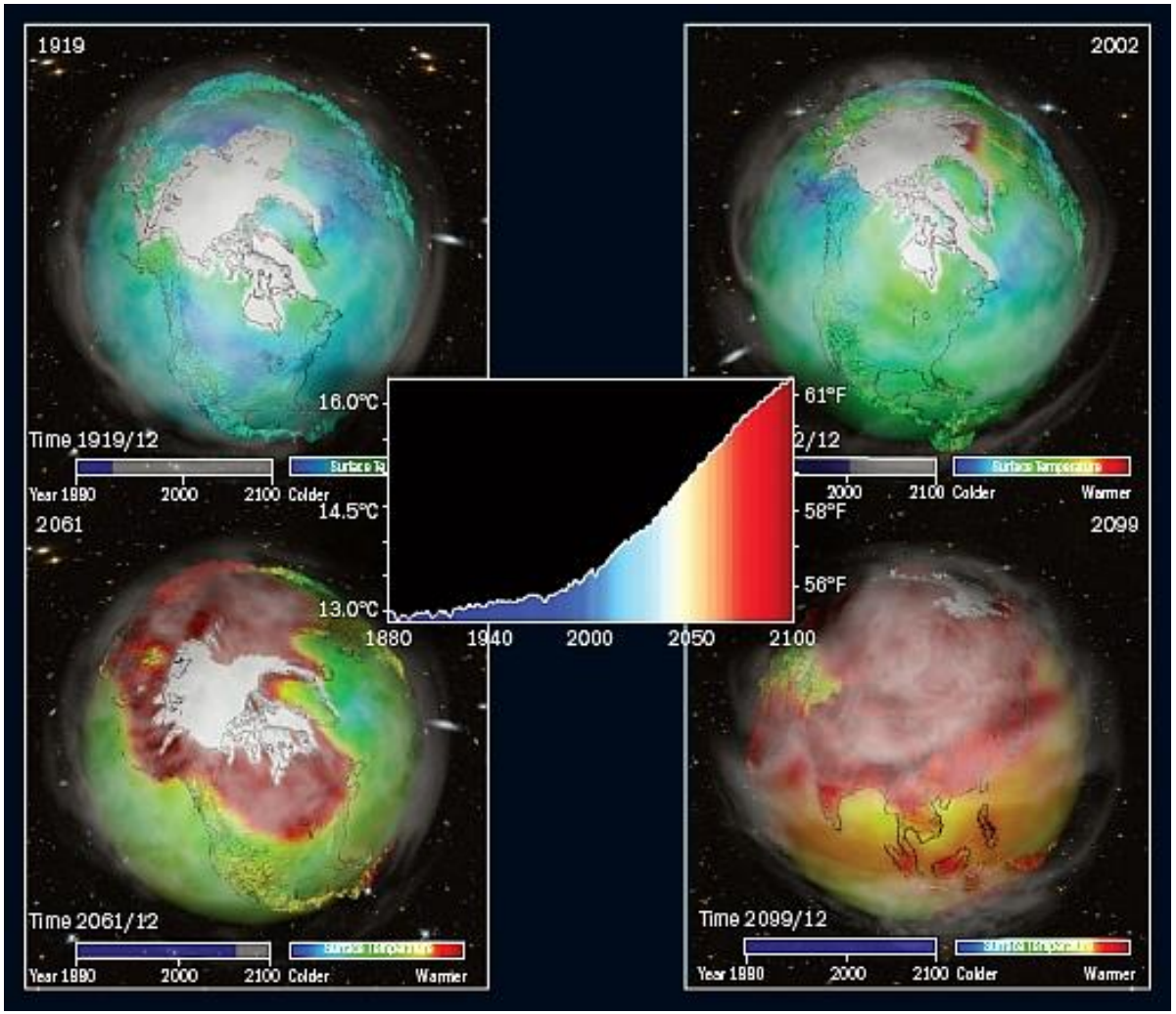
~8 Billion selects / day



New Data Collection and Processing Resources Challenged This Model

Earth System Grid

Tens of Petabytes of climate data



New Data Collection and Processing Resources Challenged This Model

Google™

Google Search | I'm Feeling Lucky

Advanced Search
Preferences
Language Tools

Year	Annual Number of Google Searches	Average Searches Per Day
2013	2,161,530,000,000	5,922,000,000
2012	1,873,910,000,000	5,134,000,000
2011	1,722,071,000,000	4,717,000,000
2010	1,324,670,000,000	3,627,000,000
2009	953,700,000,000	2,610,000,000
2008	637,200,000,000	1,745,000,000
2007	438,000,000,000	1,200,000,000
2000	22,000,000,000	60,000,000
1998	3,600,000	9,800

New Data Collection and Processing Resources Challenged This Model

- 1st paradigm, empirical science
- 2nd paradigm, model based theoretical science

New Data Collection and Processing Resources Challenged This Model

- 1st paradigm,
  empirical science

- 2nd paradigm, model
  based theoretical science

"Essentially, all models are wrong,

New Data Collection and Processing Resources Challenged This Model

- 1st paradigm,
  empirical science

- 2nd paradigm, model
  based theoretical science

"Essentially, all models are wrong, but some are useful"

New Data Collection and Processing Resources Challenged This Model

- 1st paradigm,
  empirical science

- 2nd paradigm, model
  based theoretical science

"Essentially, all models are wrong,
but some are useful"
George Edward Pelham Box (1987)

New Data Collection and Processing Resources Challenged This Model

- 1st paradigm, empirical science

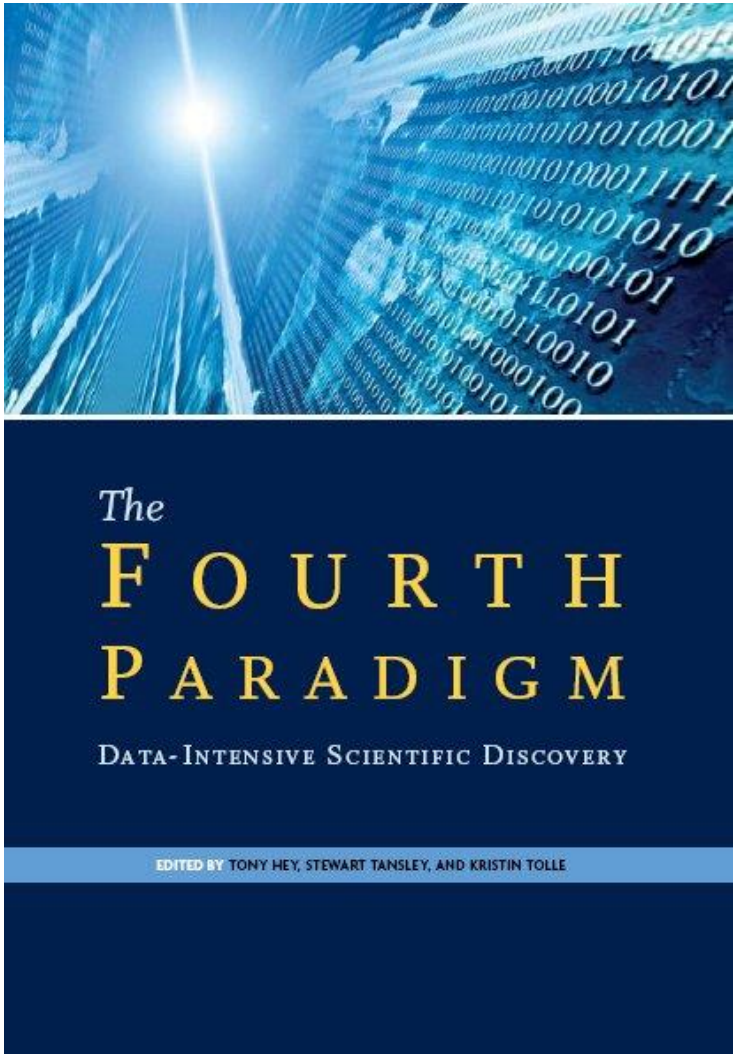
- 2nd paradigm, model based theoretical science

"Essentially, all models are wrong, but some are useful"
George Edward Pelham Box (1987)

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete
Chris Anderson (2011)

New Data Collection and Processing Resources Challenged This Model

- 1st paradigm, empirical science
- 2nd paradigm, model based theoretical science
- 3rd paradigm, computational science (simulations)
- 4th paradigm, data driven investigation (eScience)



The True Revolution is in the New Challenges that We Can Try to Tackle

Milky Way crashes into Andromeda system.

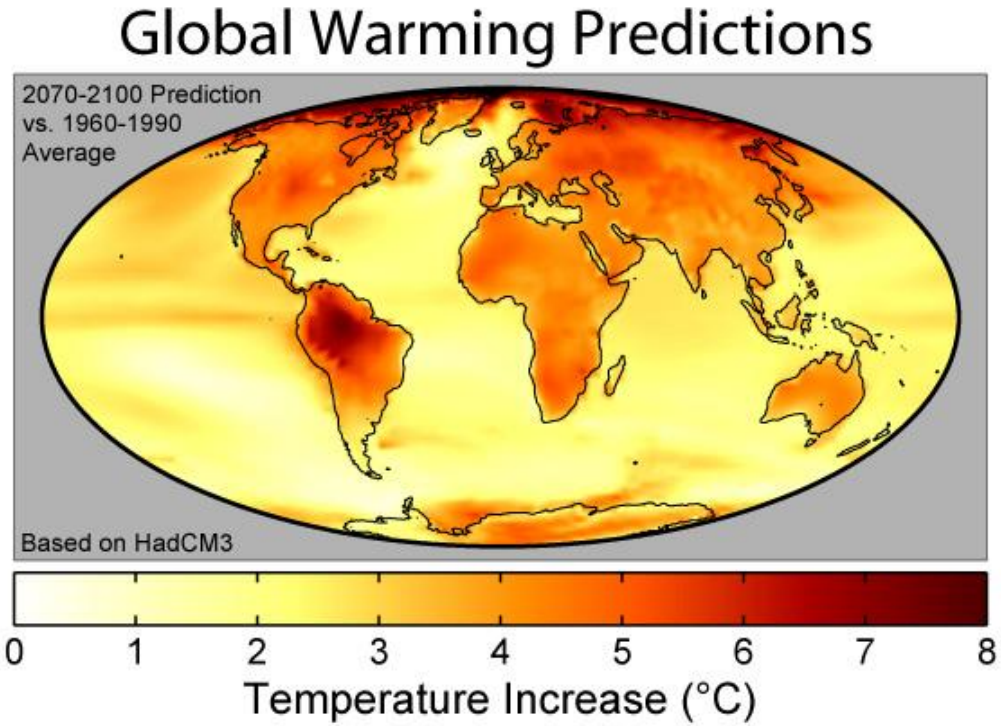
Can we test Empirically multiple scenarios?

In four billion years

The True Revolution is in the New Challenges that We Can Try to Tackle

Global warming expectation

Can we test Empirically multiple scenarios?

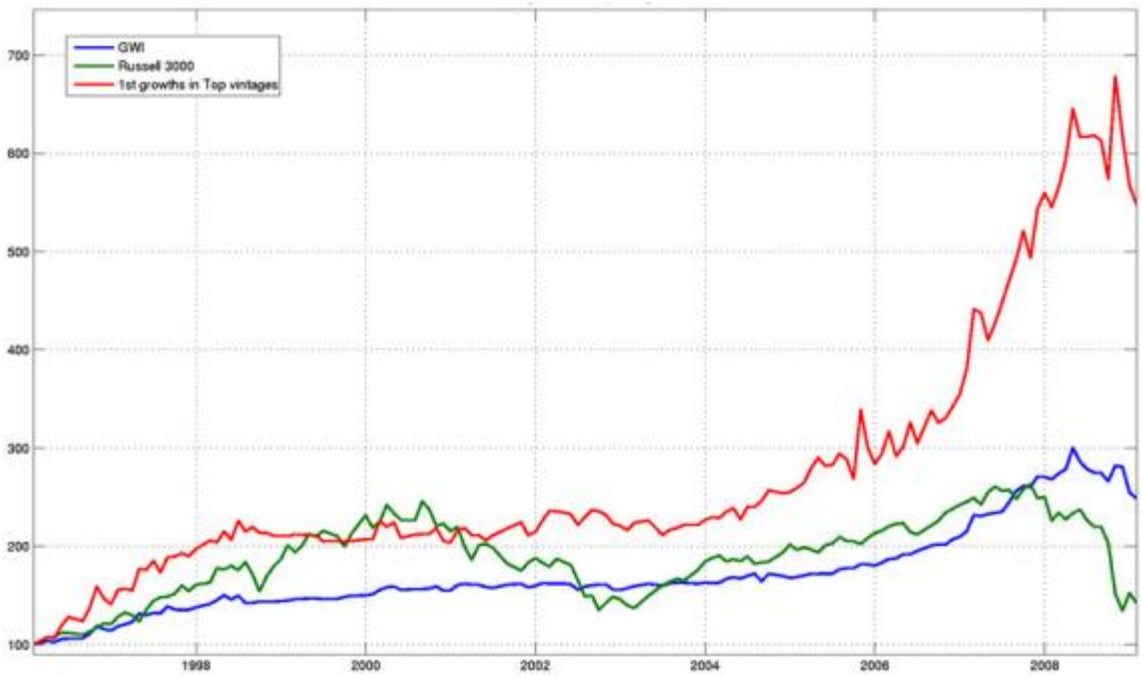


Predict the outcome in 2100

The True Revolution is in the New Challenges that We Can Try to Tackle

Evolution of the stock market

Can we test Empirically multiple scenarios?

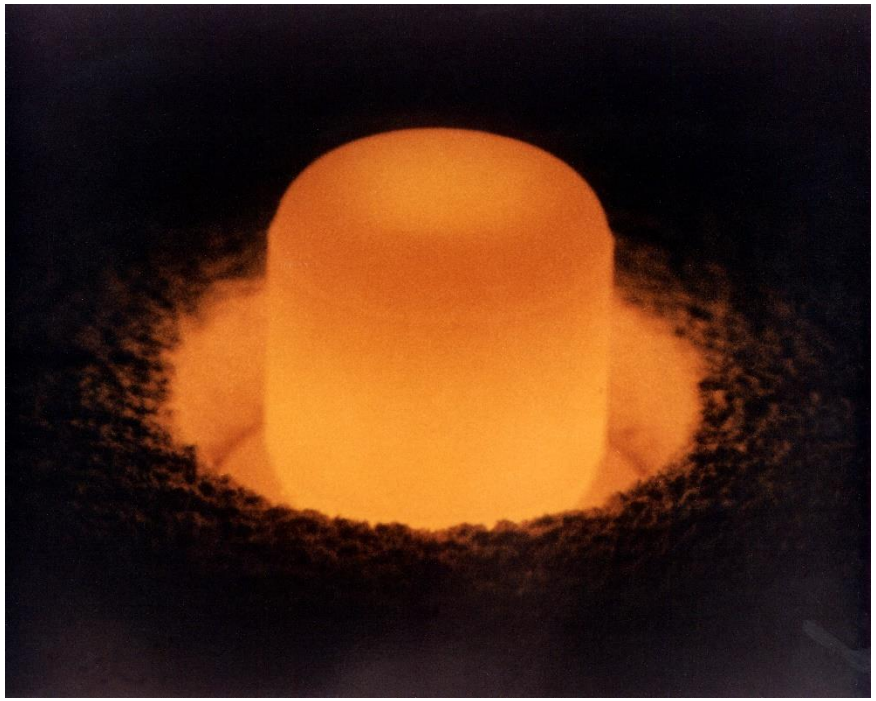


Value of investments in 10 years

The True Revolution is in the New Challenges that We Can Try to Tackle

Predicting 100 year aging effects on “new” elements

Can we test Empirically multiple scenarios?



Plutonium discovered in 1940

Would you go to space with a vehicle that has been developed only based on simulations?

- Ariane 5's first flight (Flight 501) on 4 June 1996
- $370 million in 37 seconds
- Software bug

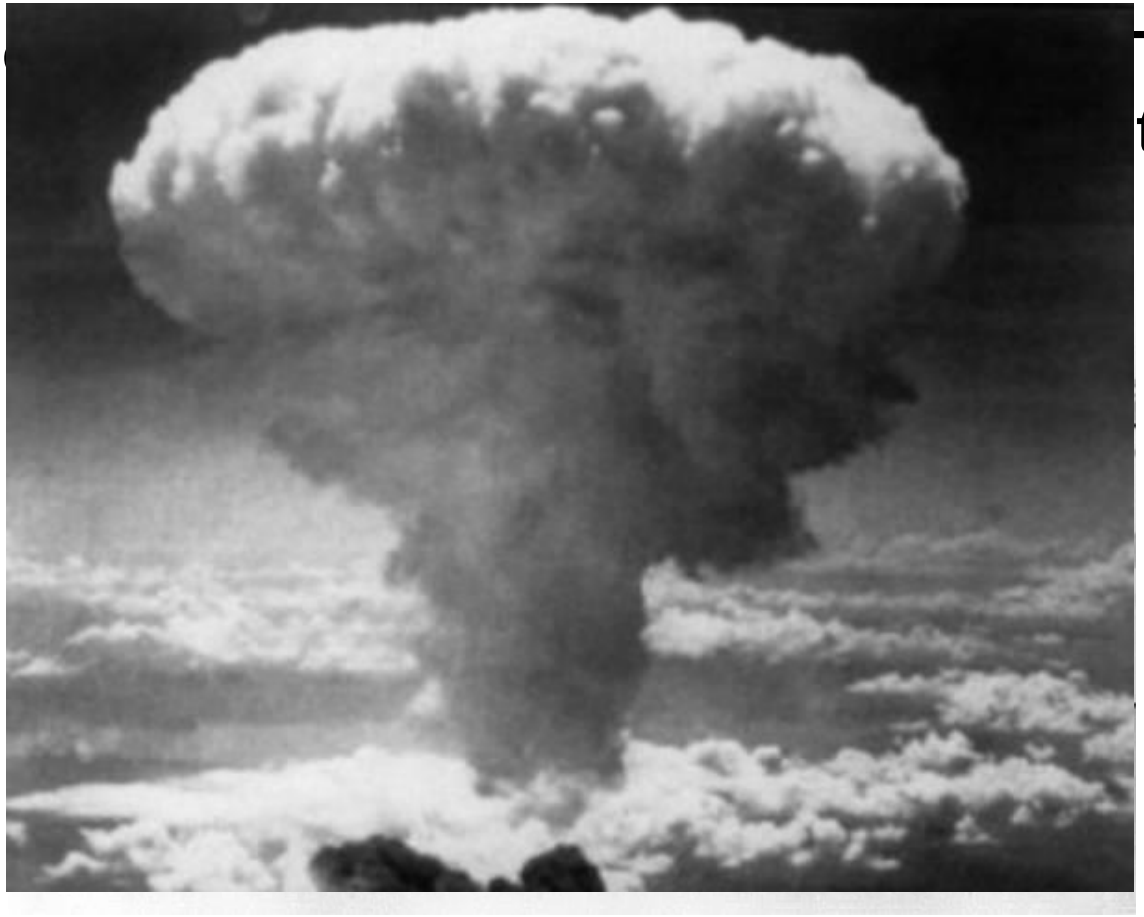
How Would You Maintain an Arsenal of nuclear Bombs that are not tested?

Comprehensive Test-Ban Treaty (CTBT)
United Nations General Assembly on 10 Sept. 1996

Is the nuclear arsenal aging properly or is it becoming dangerous and ineffective?

How Would You Maintain an Arsenal of nuclear Bombs that are not tested?

C)
United t. 1996



Is the nuclear arsenal aging properly or is it becoming dangerous and ineffective?

Unprecedented Evolution of High Performance Computing Resources

M-5: Los Alamos National Lab
No 1. system in June 1993C

Unprecedented Evolution of High Performance Computing Resources

Num. Wind Tunnel: National Aerospace Laboratory of Japan No. 1 system in November 1993 and November 1994

Unprecedented Evolution of High Performance Computing Resources

Intel XP/S 140 Paragon: Sandia National Labs No. 1 system in June 1994

Unprecedented Evolution of High Performance Computing Resources

Hitachi SR2201: University of Tokyo
No. 1 system in June 19

Unprecedented Evolution of High Performance Computing Resources

CP-PACS: University of Tsukuba

Unprecedented Evolution of High Performance Computing Resources

ASCI Red: Sandia National Laboratory
No. 1 system from June 1997 to June 2000

Unprecedented Evolution of High Performance Computing Resources

ASCI White: LLNL

No. 1 system from Nov. 2000 to Nov. 2001

Unprecedented Evolution of High Performance Computing Resources

The Earth Simulator

No. 1 from June 2002 to June 2004

Unprecedented Evolution of High Performance Computing Resources

BlueGene/L: LLNL

No. 1 from November 2004 to November 2007

Unprecedented Evolution of High Performance Computing Resources

Roadrunner: Los Alamos National Laboratory No. 1 from June 2008 to June 2009

Unprecedented Evolution of High Performance Computing Resources

Jaguar: Oak ridge National Laboratory
No. 1 from November 2009 to June 2010

Unprecedented Evolution of High Performance Computing Resources

Tianhe-1A: National SC in Tianjin
No. 1 in November 2010

Unprecedented Evolution of High Performance Computing Resources

K Computer: RIKEN Institute for
No. 1 June 2011 to November 2011

Unprecedented Evolution of High Performance Computing Resources

Sequoia: LLNL
No. 1 in June 2012

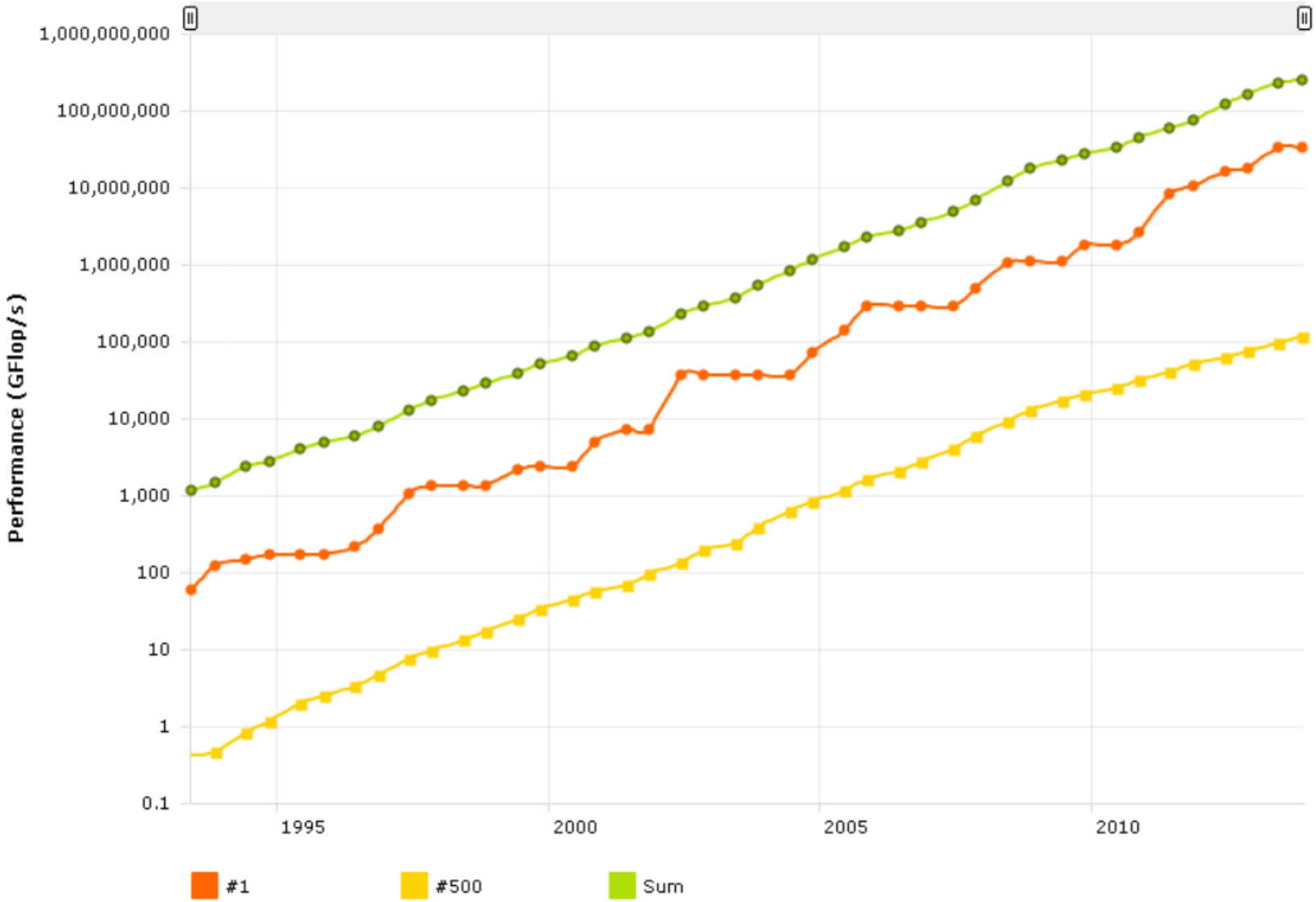
Unprecedented Evolution of High Performance Computing Resources

Titan: Oak Ridge National Laboratory
No. 1 in November 2012

Unprecedented Evolution of High Performance Computing Resources

Tianhe-2 (MilkyWay-2)
No. 1 system since June 2013

Unprecedented Evolution of High Performance Computing Resources

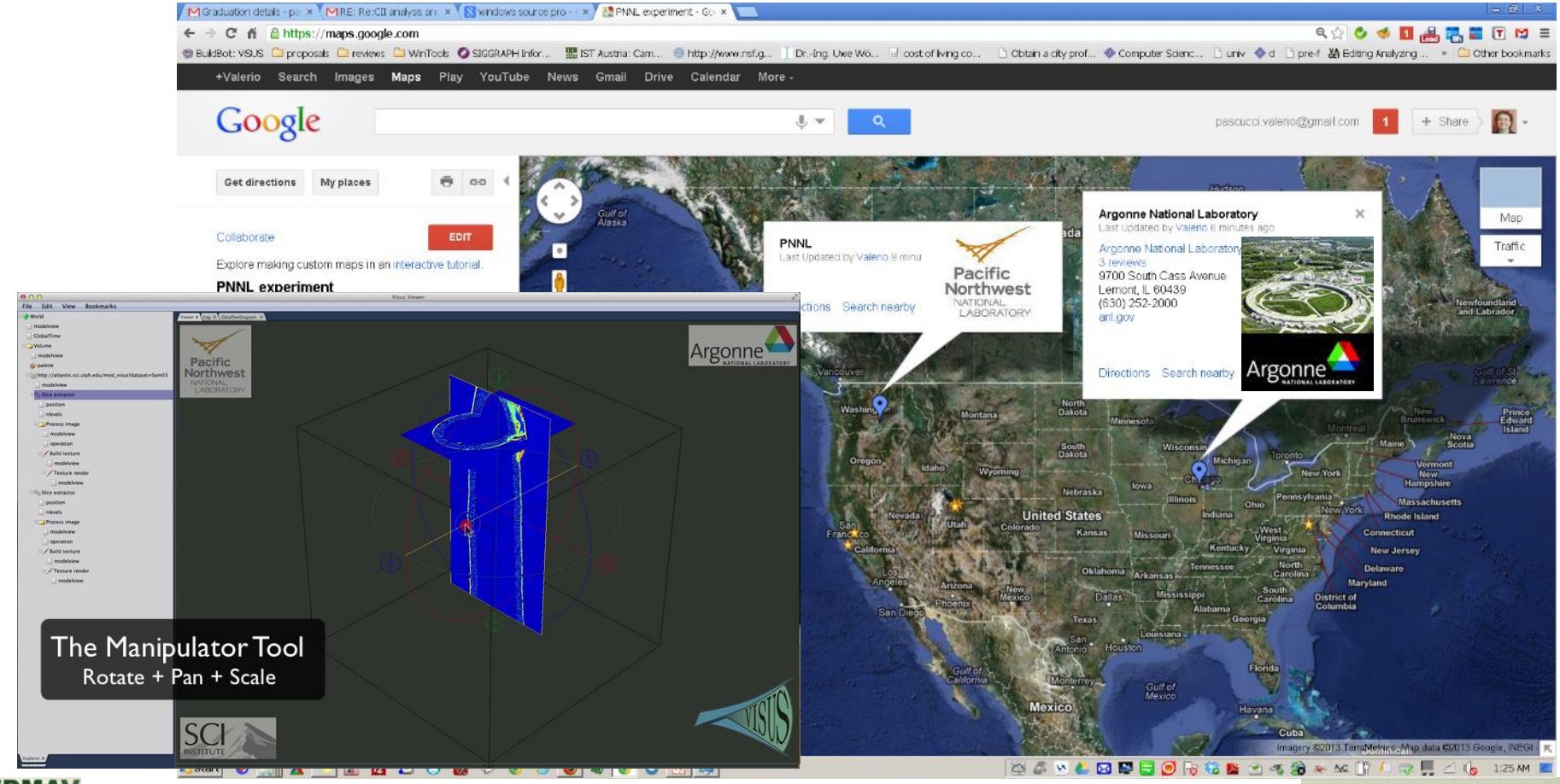


The Fundamental Paradigm Shift of eScience

- **The data is a major driver of the investigations**
- **Having the data does NOT guarantee to have the right questions and answers**
- **NOT having the data guarantees that you CANNOT develop the right questions and answers**
- **HPC, scientific computing, machine learning, data analysis, mining, exploration and visualization are critical activities to knowledge discovery.**

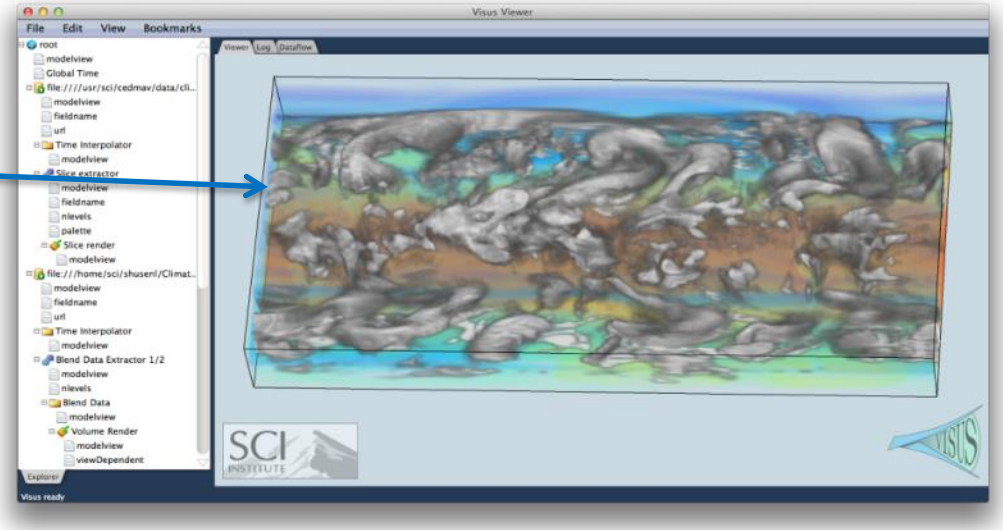
High Performance Data Movements for Real-Time Access to Large Scale Experimental Data

- **Experiment run at Advance Photon Source at ANL**
- **Scientists located at PNNL**

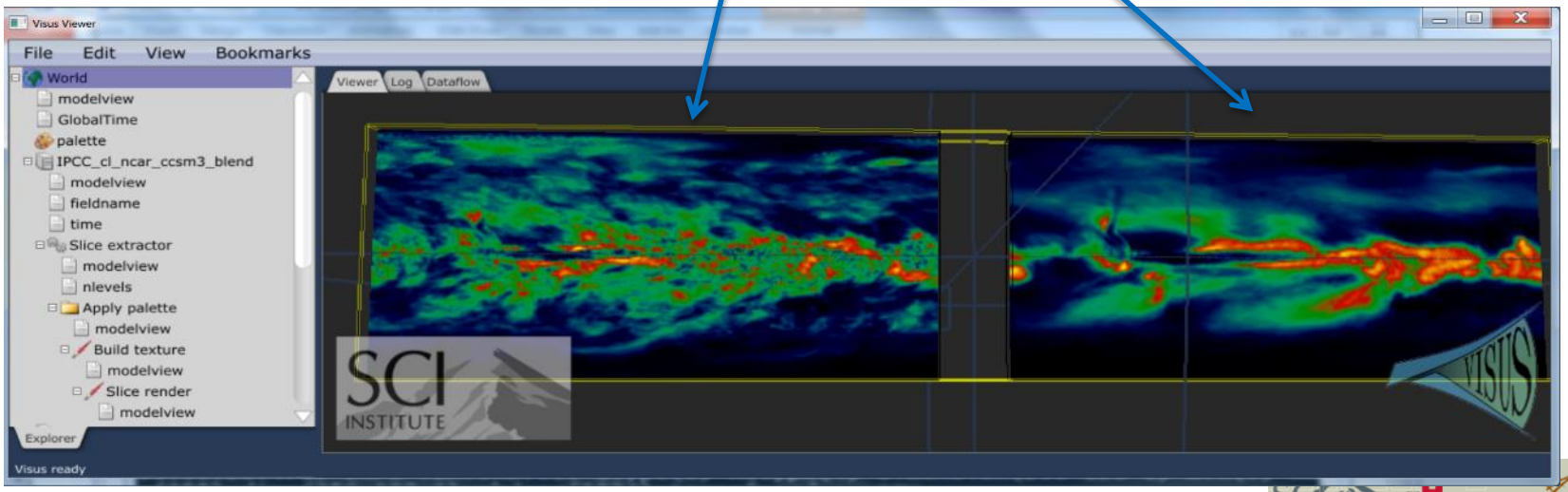


High Performance Data Movements for Real-Time Access to Large Scale Simulation Data

- Data streams that allow merging multiple datasets in real time
- Time interpolation of and concurrent visualization of climate data ensembles defined on different time scales
- Server side and client side computation of statistical functions such as median, average, standard deviation, …….

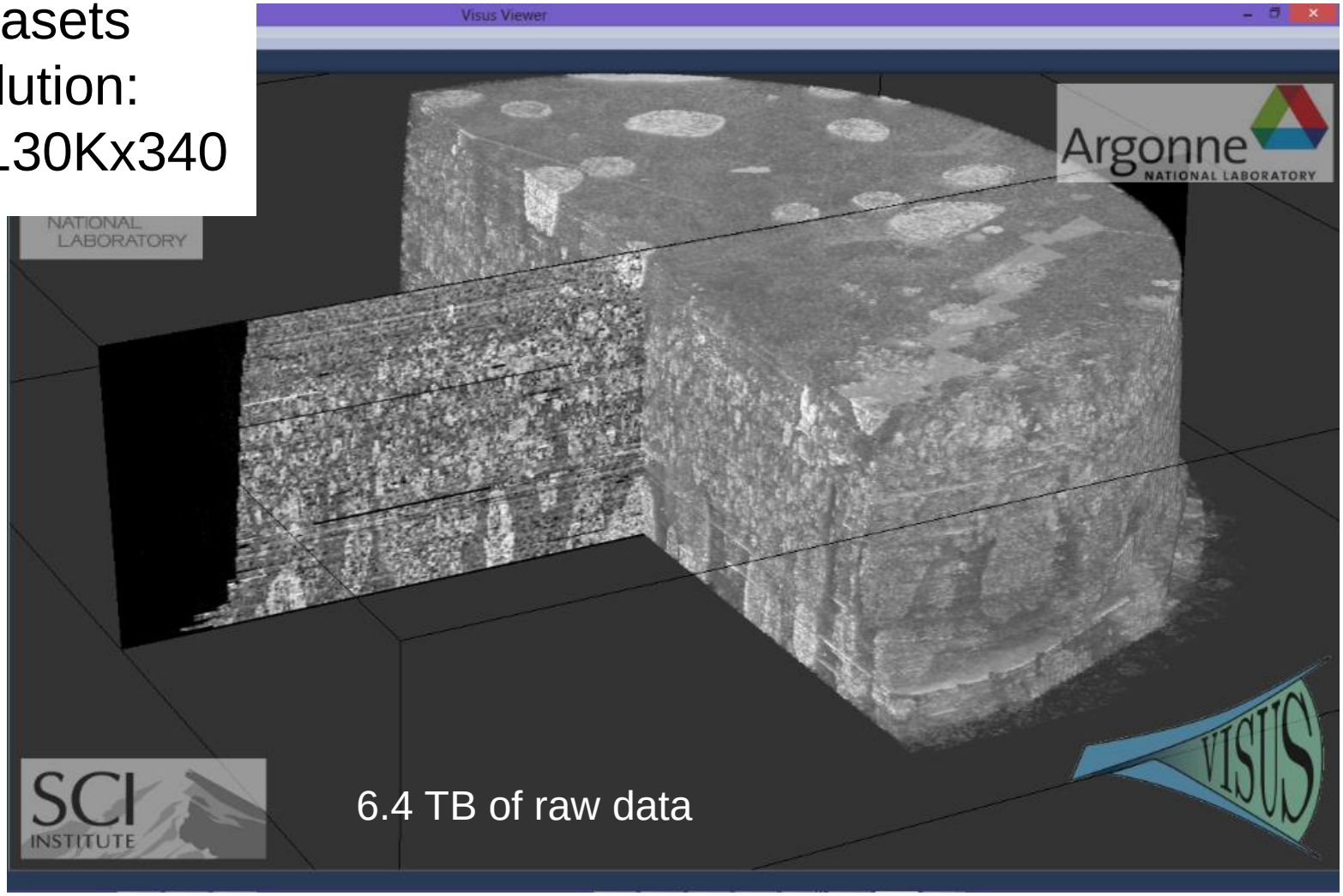


Standard Deviation and Average of ten climate models



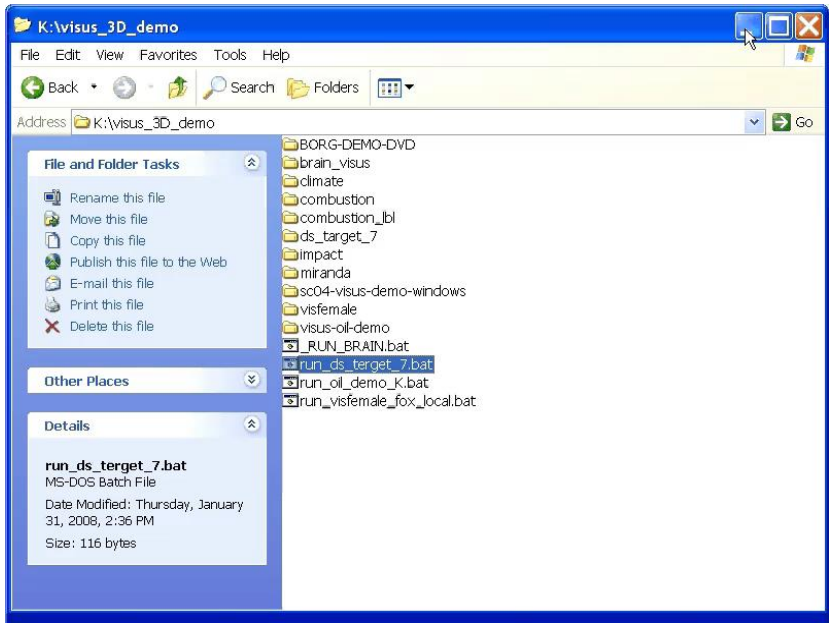
Interactive Remote Analysis and Visualization of 6TB Imaging Data

- EM datasets of resolution: 130Kx130Kx340



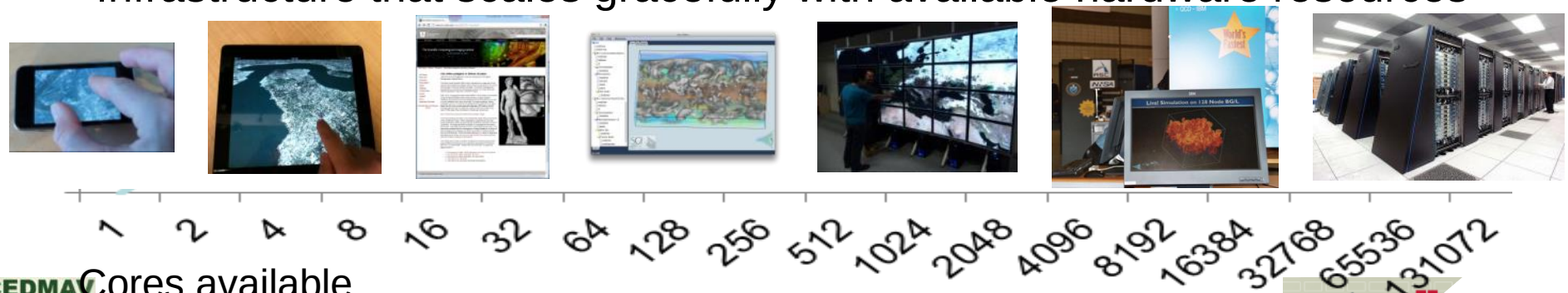
Web Server

SC12/13 Demonstration of data streaming analytics and visualization



Live demonstration from Argonne National Laboratory to Supercomputing exhibit floor

Infrastructure that scales gracefully with available hardware resources



Cores available

Massive Precomputations Can Avoid the Need for Real Time Processing

- **Problem:** Need accurate automated phone quotes in 100ms. They couldn't do these calculations nearly fast enough on the fly.
- **Solution:** Each weekend, use a new HPC cluster to pre-calculate quotes for <u>every American adult and household</u> (60 hour run time)

The Fundamental Paradigm Shift of eScience

Development and curation of massive data collections from simulations and sensing will be crucial to any scientific progress

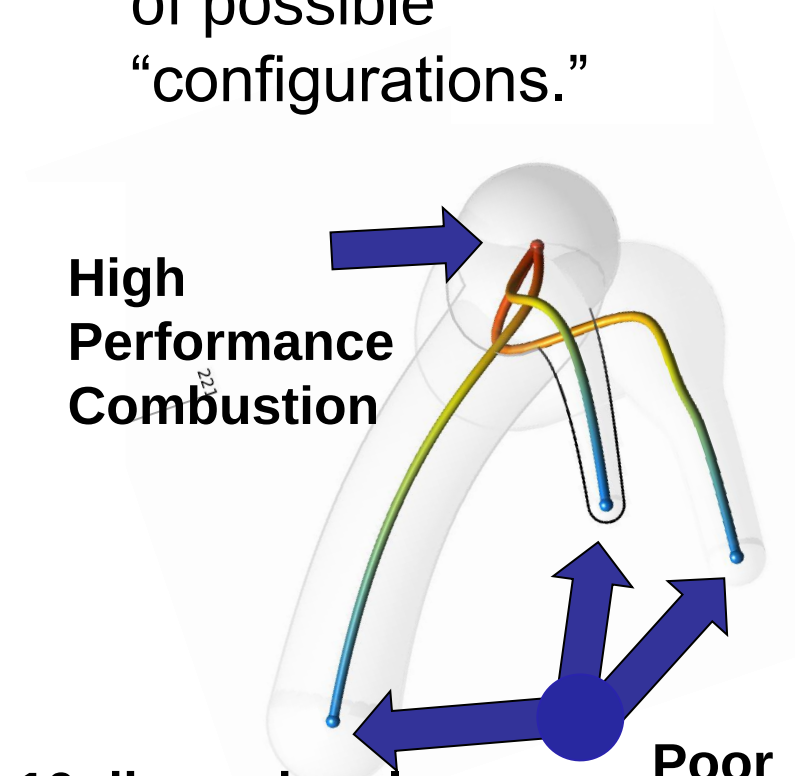
Need to develop scientific knowledge and actionable information that cannot be tested empirically

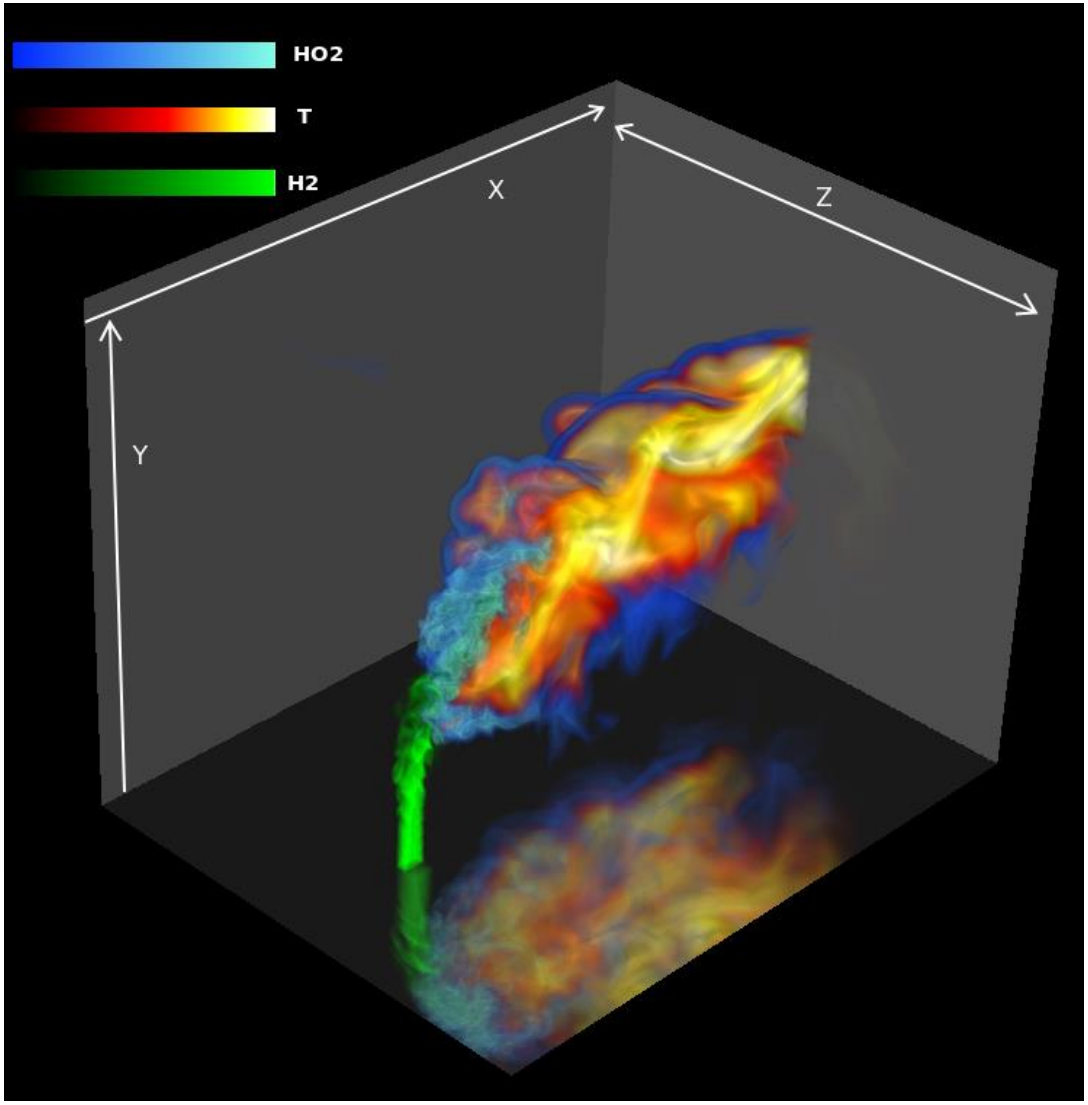
Uncertainty Quantification

Verification and Validation

Assessing the Uncertainty in Fuel Design For New Clean Burning Devices

Exploration of high dimensional space of possible "configurations."

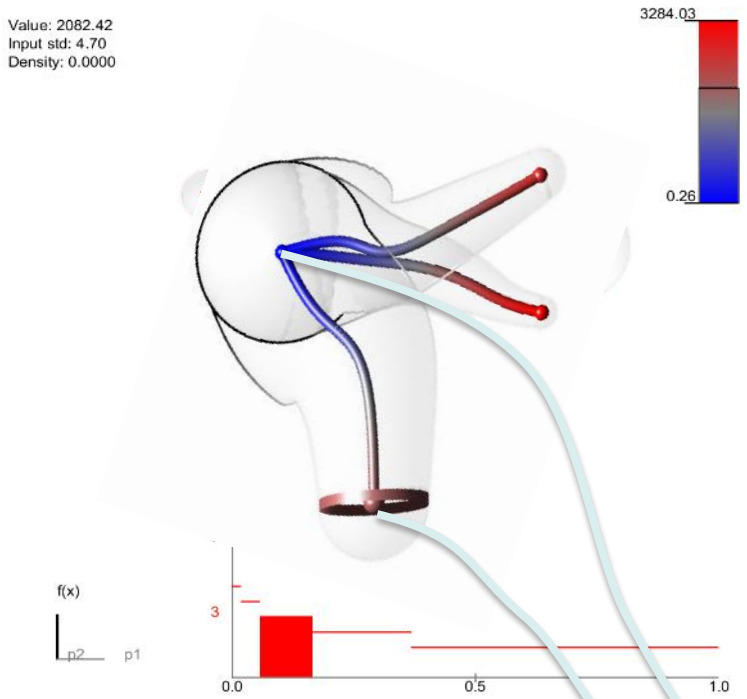


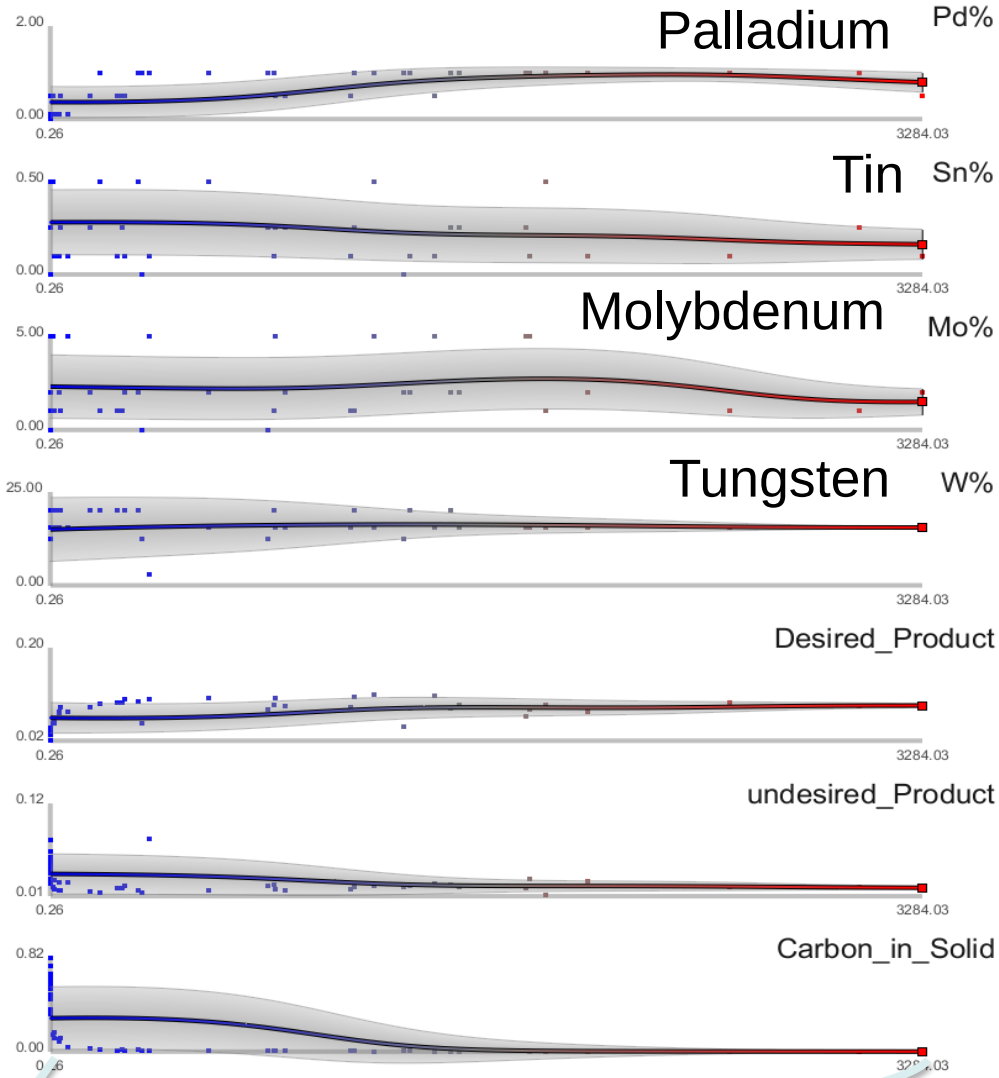


Topological Analysis of the Space of Composite Materials of a Given Class

- Features in experimental data show unexpected structures and are used to plan future experiments.

Stakeholder: A. Karim, PNNL.





How Can We Achieve True Verification and Validation?

Verification:

Are we building the software right?

How Can We Achieve True Verification and Validation?

Verification:

Are we building the software right?

Is my software bug free?

How Can We Achieve True Verification and Validation?

Verification:

Are we building the software right?

Is my software bug free?

How Can We Achieve True Verification and Validation?

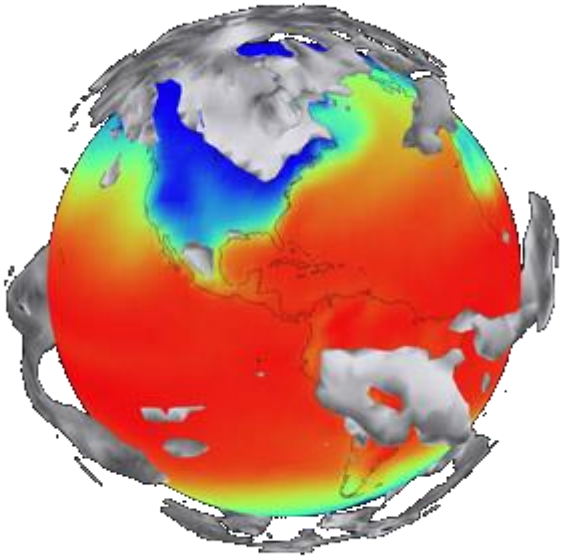
Validation:

Is the process providing accurate predictions?

How Can We Achieve True Verification and Validation?

Validation:

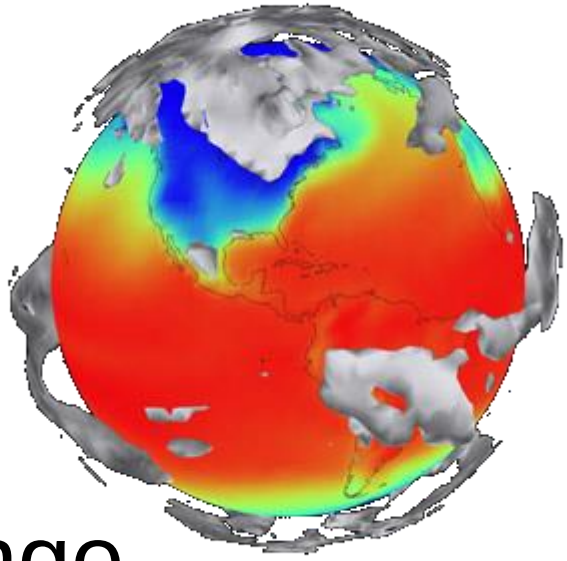
Is the process providing accurate predictions?



How Can We Achieve True Verification and Validation?

Validation:

Is the process providing accurate predictions?



U.S. Securities and Exchange Commission (SEC) requires to tell: investors that a fund's:

"past performance does not necessarily predict future results"

Example:

Sidharth Kumar

Aaditya Landge

“past performance does not necessarily predict future results”

Example:

Sidharth Kumar

ATPESC summer school 2013

Aaditya Landge

“past performance does not necessarily predict future results”

Example:

Sidharth Kumar

ATPESC summer school 2013

Aaditya Landge

SC14

Efficient I/O and Storage of Adaptive Resolution Data

SC14

In-Situ Feature Extraction of Large Scale Combustion Simulations Using Segmented Merge Trees

"past performance does not necessarily predict future results"

Example:

Sidharth Kumar

ATPESC summer school 2013

Aaditya Landge

SC14

Efficient I/O and Stora
Adaptive Resolution

SC14

Feature Extraction of Large
Combustion Simulations
g Segmented Merge Trees

“past performance does not necessarily predict future results”

Example:

Sidharth Kumar

ATPESC summer school 2013

Aaditya Landge

SC14

Efficient I/O and Stora
Adaptive Resolution

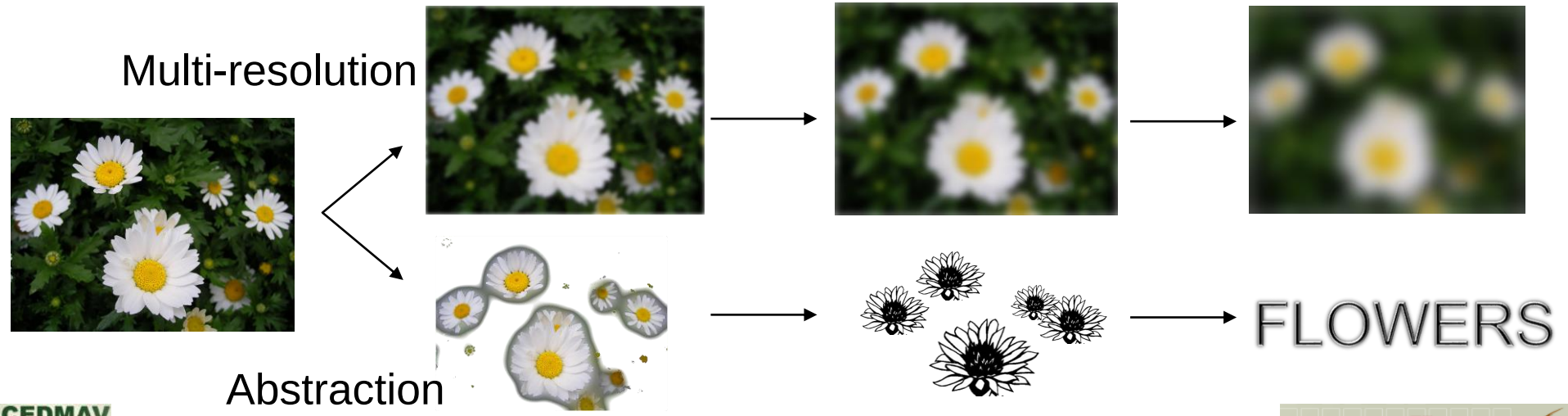
SC14

Feature Extraction of Large
Combustion Simulations
Segmented Merge Trees

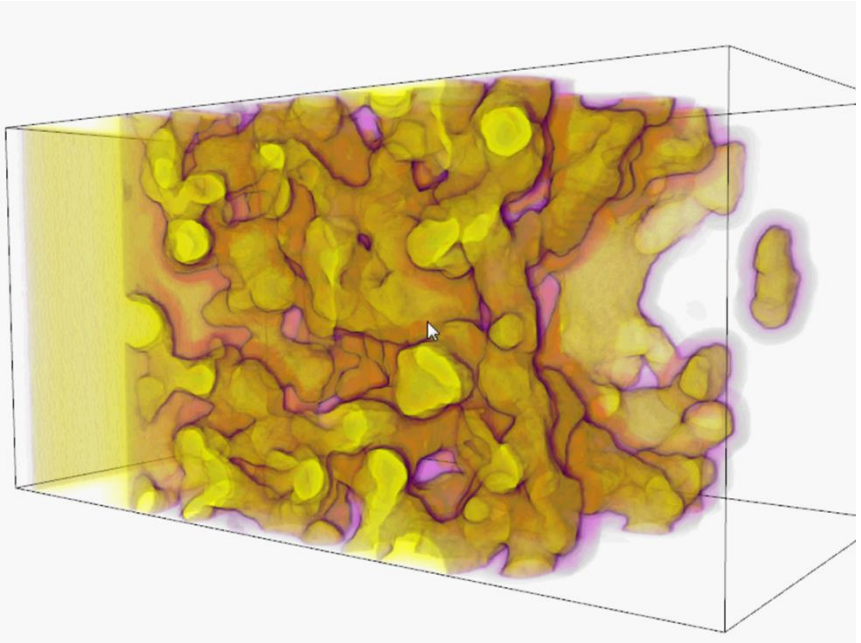
“past performance does not necessarily predict future results”

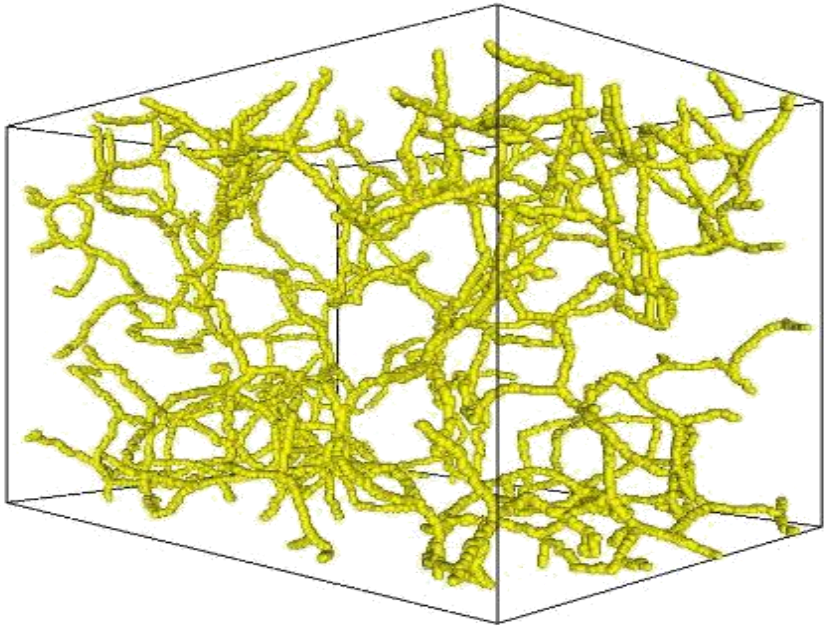
Rethinking Multi-Scale Representation of Massive Data Models

- **Multi-resolution representations are insufficient to deal with big data:**
  - **Data preprocessing is typically too long**
  - **Wavelet-like averaging looses information**
  - **Data analysis results often do not represent well important trends (e.g. multi-modal distributions)**
- **New data "abstractions" are needed for Big Data**



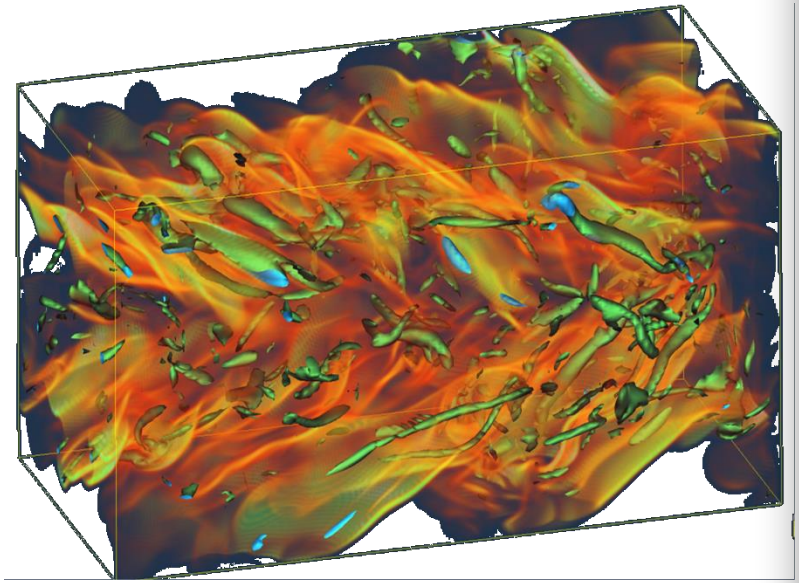
Rethinking Multi-Scale Representation of Massive Data Models

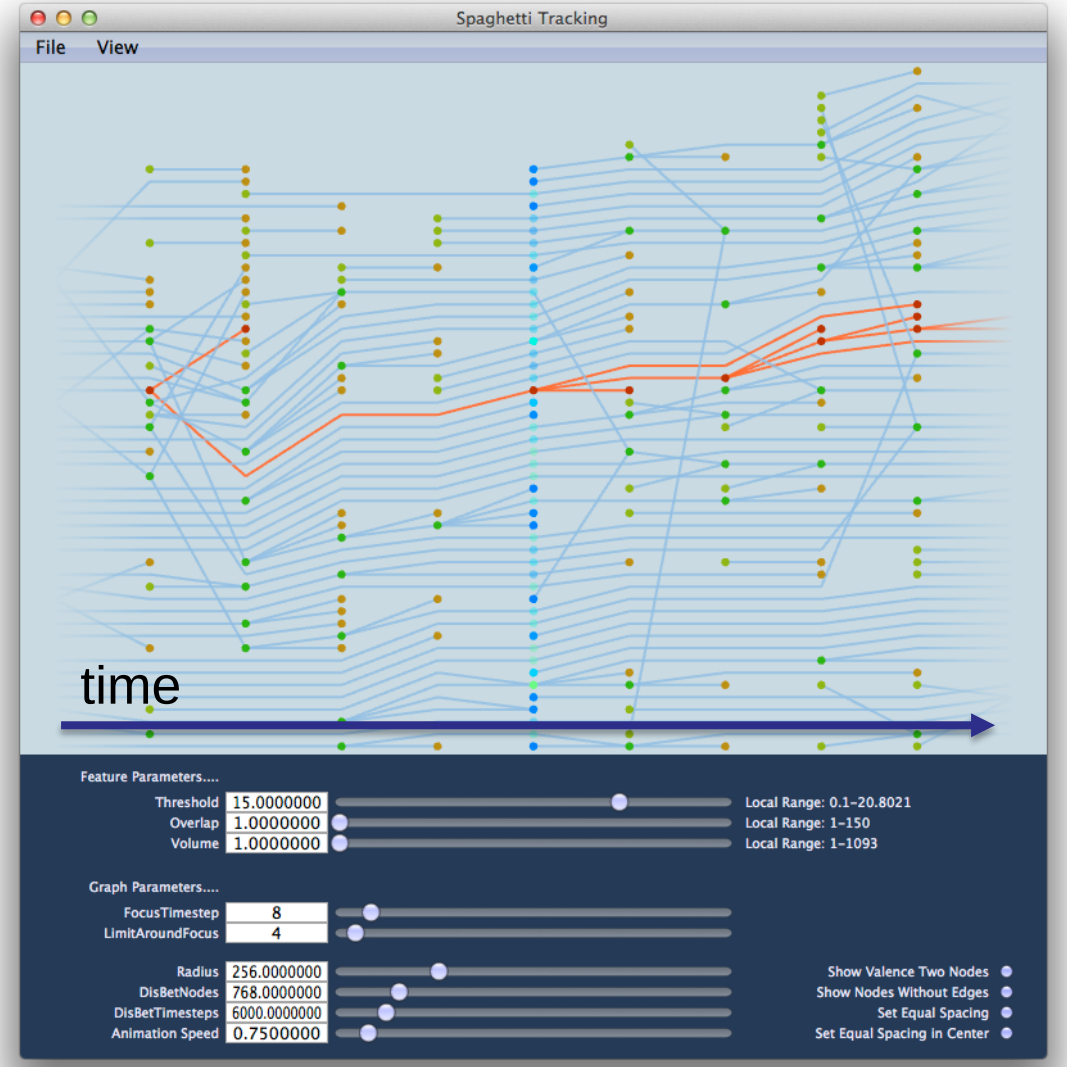




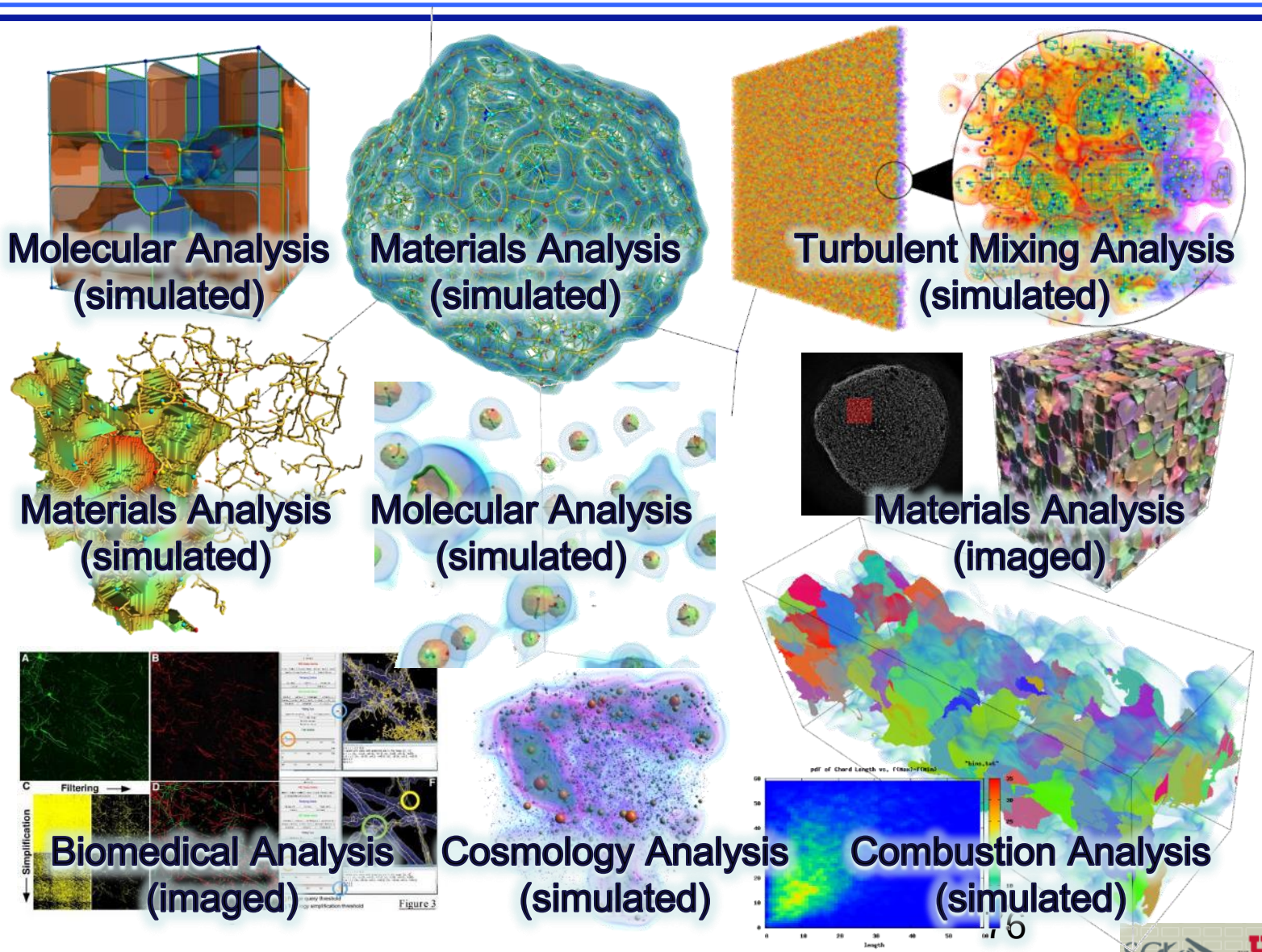
Topological Analysis of Massive Combustion Simulations

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





Topology Has Been Successful for Analysis and Visualization of Massive Scientific Data

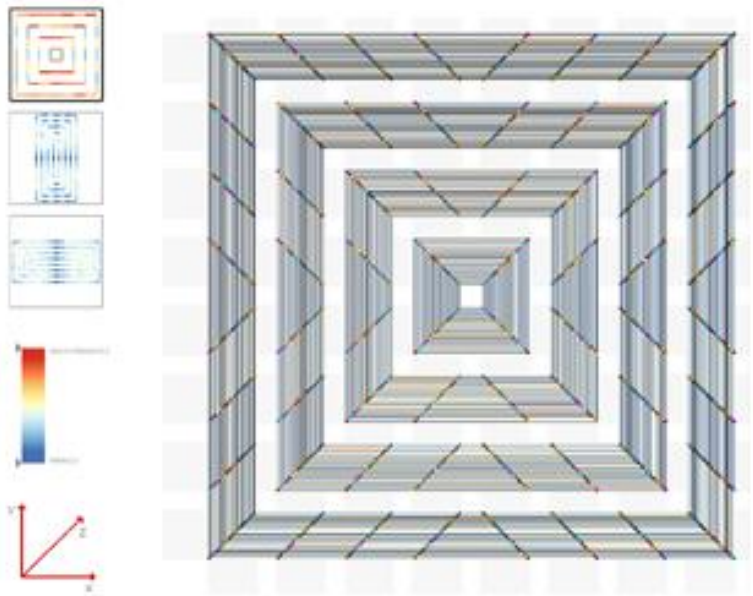


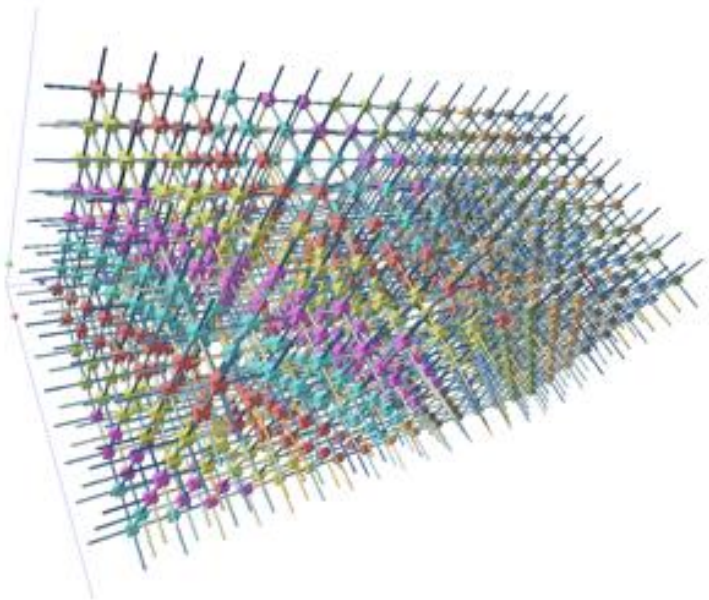
Running Efficiently Big Data Computations is a Big Data Problem

- **Massive logs**
- **Complex Memory Hierarchies**
- **Complex and Diverse Interconnects**
  - **3D, 4D, 5D tori**
  - **Fat tree**
- **Complex I/O pathways**
- **Cost of Power Dominated by Data Movements**

Running Efficiently Big Data Computations is a Big Data Problem

- **Growing Community involving Performance analysis and vis community**
  - **Workshop at last IEEE VIS**
  - **Dagstuhl Perspective Workshop**
  - **Workshop at SC14 Conference**





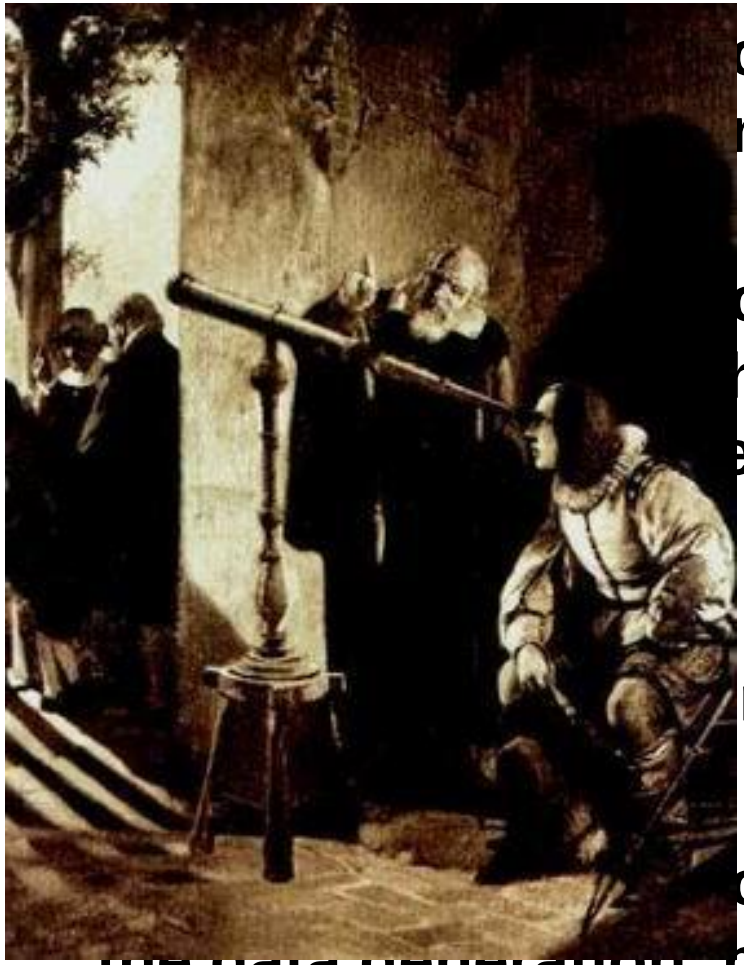
Large Team Requiring Multi-disciplinary and Multi-institutional Collaboration

- Challenging collaborations among:
  - Government laboratories
  - Industry
  - Academia
- Close collaborations with domain scientists requiring to cross language and cultural barriers:
  - New education needed
  - Communicate problems not tasks!!!!!!

The Big Gift of Big Data

- A great *opportunity* to achieve new scientific discoveries and engineering innovations
- A great *opportunity* for the Computer Science community to become a central player in the development of modern science
- A great *challenge* for all communities to become strongly engaged in interdisciplinary collaboration
- A great *opportunity* for our community to become the data generation, processing and exploration "telescope" of modern science and engineering

The Big Gift of Big Data



o
n

o
n
e

a
r

o

the data generation, processing and exploration
“telescope” of modern science and engineering

END