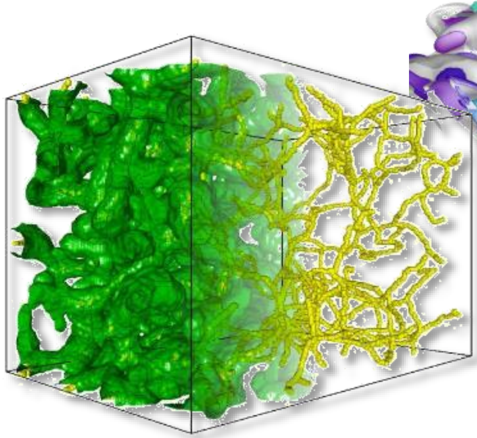
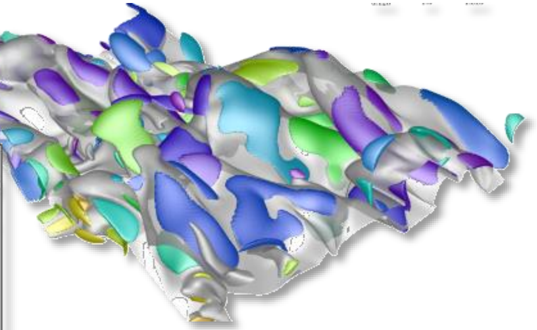
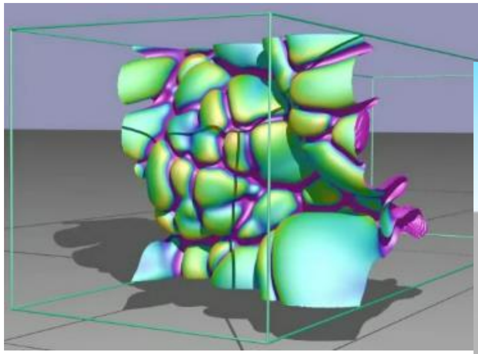
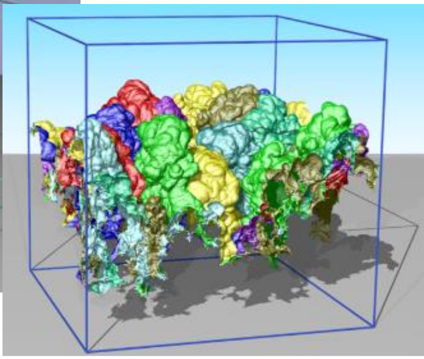
CEDMAV Center for Extreme Data Management, Analysis, and Visualization

The Big Gift of Big Data

Valerio Pascucci

Director, Center for Extreme Data Management Analysis and Visualization

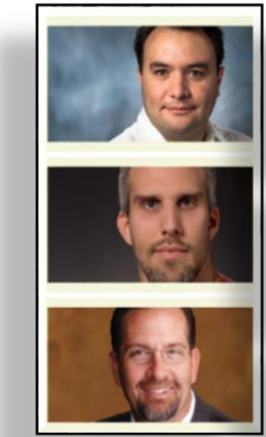
Professor, SCI institute and School of Computing, University of Utah

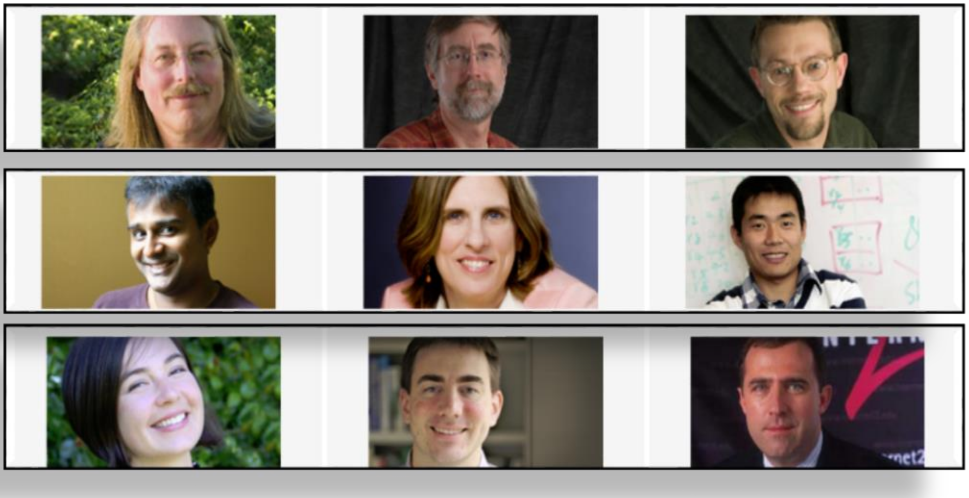
Laboratory Fellow, Pacific Northwest National Laboratory

CEDMAV

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





SDAV Scalable Data Management, Analysis, and Visualization

EXaCT

SUPER

The University of Utah Pacific Northwest National Laboratory

Lawrence Livermore National Laboratory

Sandia National Laboratories

INL Idaho National Laboratory

$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,

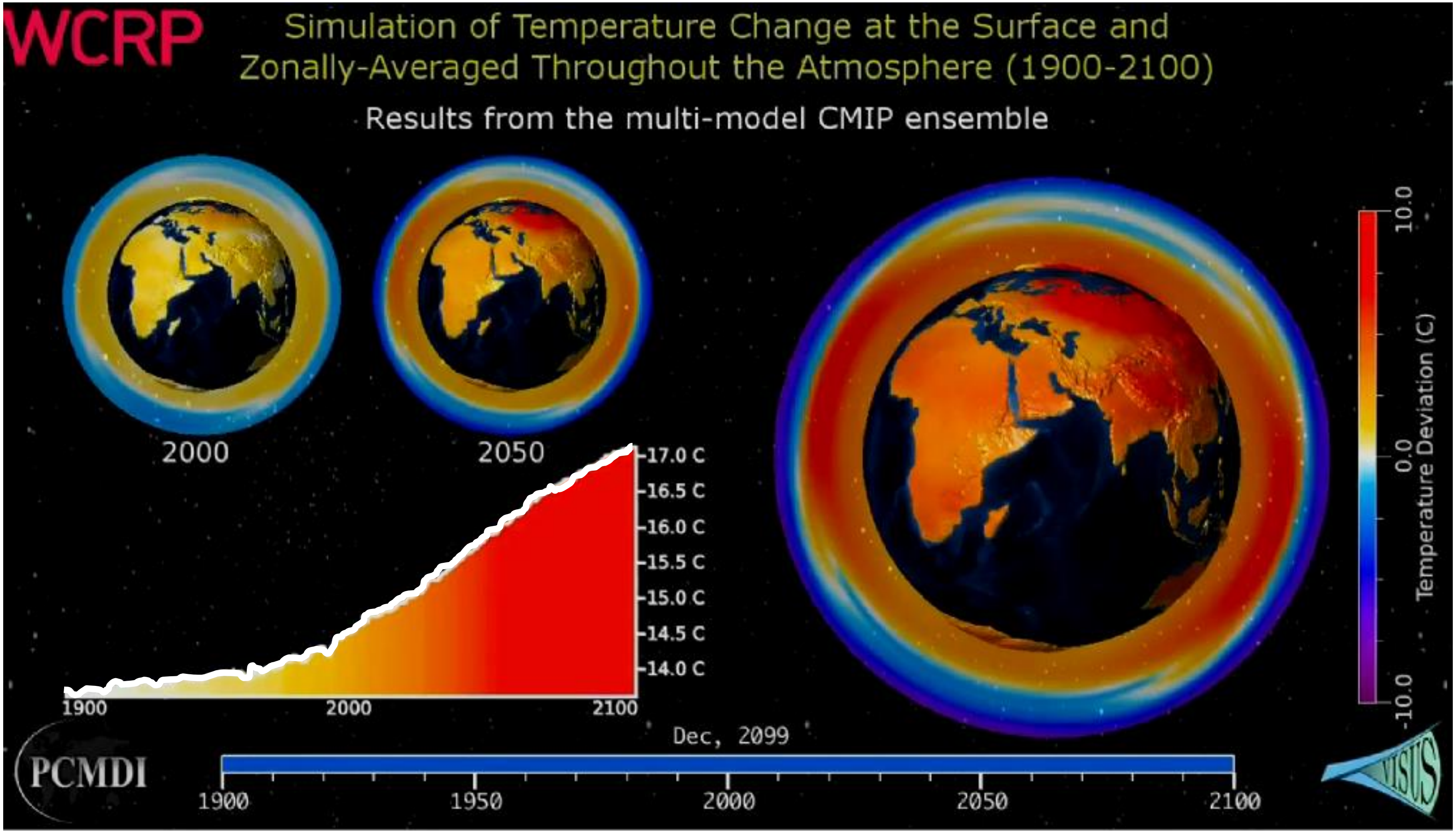
everyone thinks everyone else is doing it,

so everyone claims they are doing it ….

(Dan Ariely)

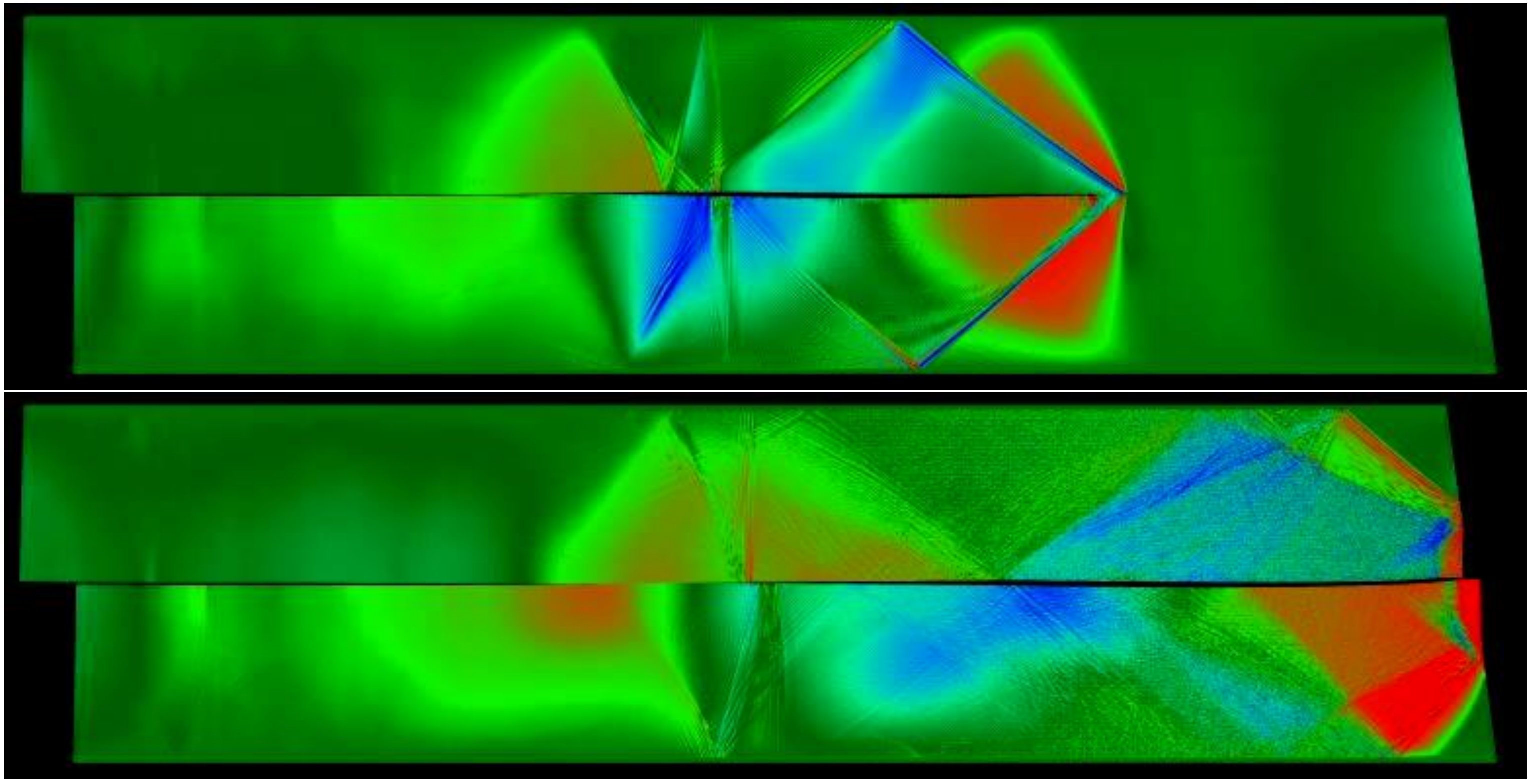
…. eventually everyone will do it!

Are Global Warming Trends Associated to Human Activities?


WCRP
Simulation of Temperature Change at the Surface and Zonally-Averaged Throughout the Atmosphere (1900-2100)
Results from the multi-model CMIP ensemble
2000
2050
17.0 C
16.5 C
16.0 C
15.5 C
15.0 C
14.5 C
14.0 C
1900
2000
2100
Dec, 2099
PCMDI
1900
1950
2000
2050
2100
10.0
0.0
-10.0
Temperature Deviation (C)
VISUS

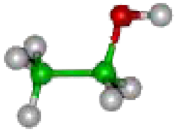
Can Devastating Material Failures Such as Earthquakes Travel at Supersonic Speed?

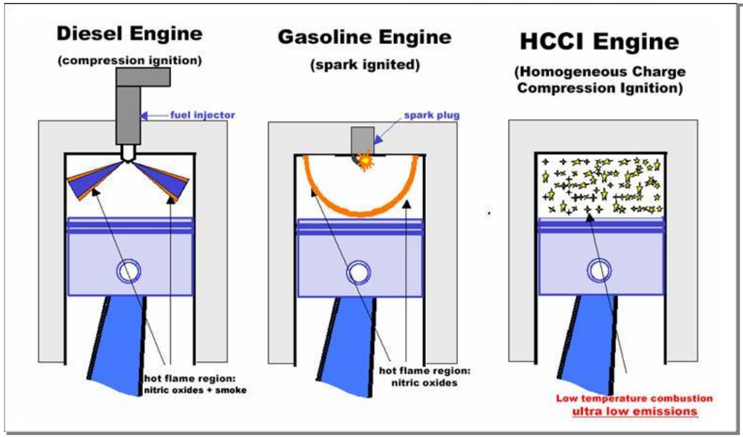
- Understanding the strength of new materials is critical in creating structures as small as microprocessors, buildings, or airplanes that withstand real-world forces

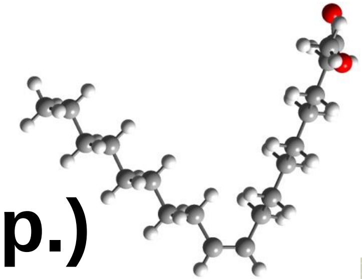


Change the World of Fuels and Engines to Increase Efficiency and Reduce Pollution

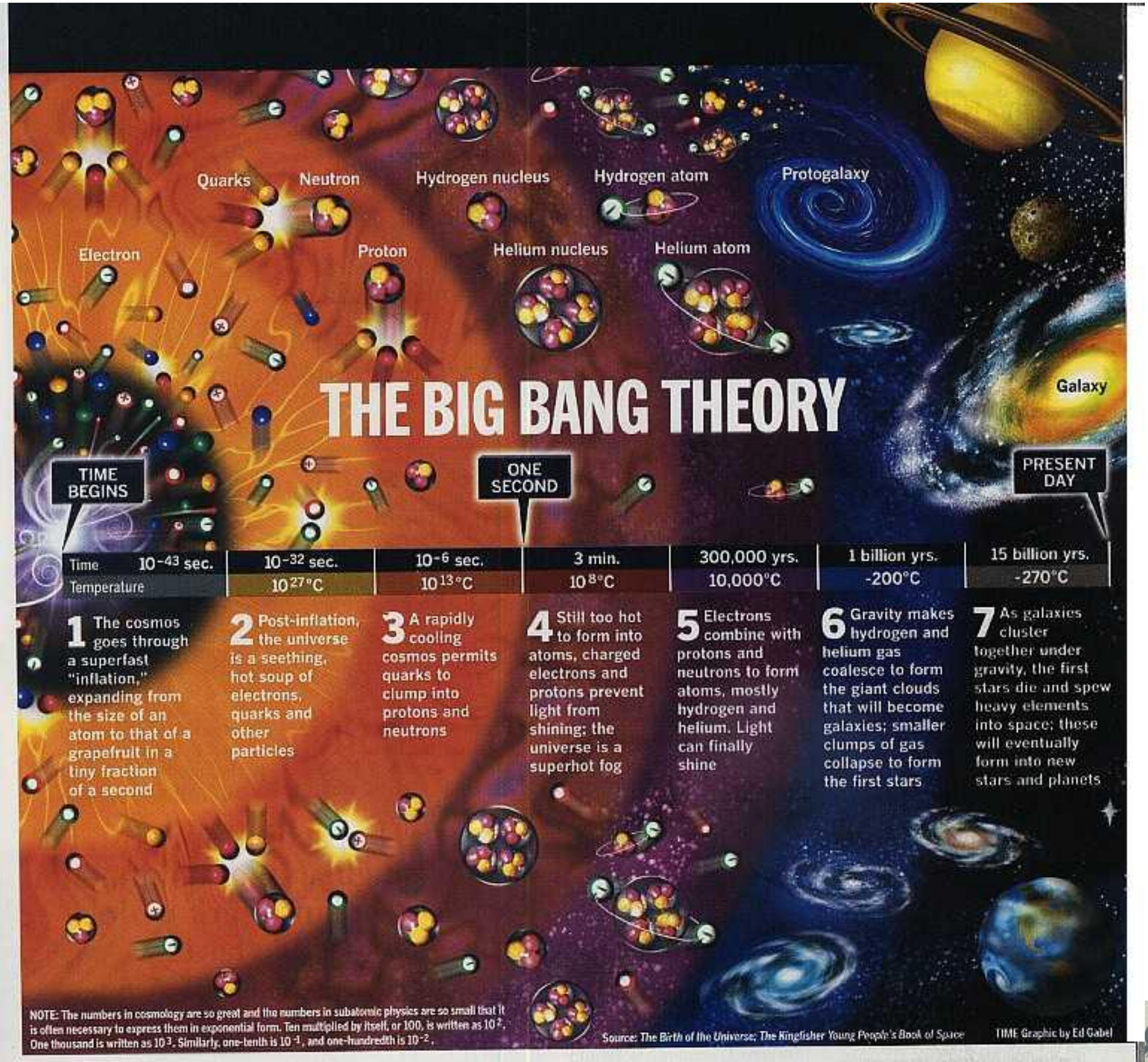
- Fuel streams are rapidly evolving:
  - Heavy hydrocarbons
    - Oil sands
    - Oil shale
    - Coal
  - New renewable fuel sources
    - Ethanol
    - Biodiesel
- New engine technologies:
    - Direct Injection (DI)
    - Homogeneous Charge Compression Ignition (HCCI)
    - Low-temperature combustion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




Diesel Engine
(compression ignition)
fuel injector
hot flame region: nitric oxides + smoke
Gasoline Engine
(spark ignited)
spark plug
hot flame region: nitric oxides
HCCI Engine
(Homogeneous Charge Compression Ignition)
Low temperature combustion
ultra low emissions



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



Time	10^{-43} sec.	10^{-32} sec.	10^{-6} sec.	3 min.	300,000 yrs.	1 billion yrs.	15 billion yrs.
Temperature		10^{27}°C	10^{13}°C	10^{8}°C	10,000°C	-200°C	-270°C

1 The cosmos goes through a superfast "inflation," expanding from the size of an atom to that of a grapefruit in a tiny fraction of a second

2 Post-inflation, the universe is a seething, hot soup of electrons, quarks and other particles

3 A rapidly cooling cosmos permits quarks to clump into protons and neutrons

4 Still too hot to form into atoms, charged electrons and protons prevent light from shining; the universe is a superhot fog

5 Electrons combine with protons and neutrons to form atoms, mostly hydrogen and helium. Light can finally shine

6 Gravity makes hydrogen and helium gas coalesce to form the giant clouds that will become galaxies; smaller clumps of gas collapse to form the first stars

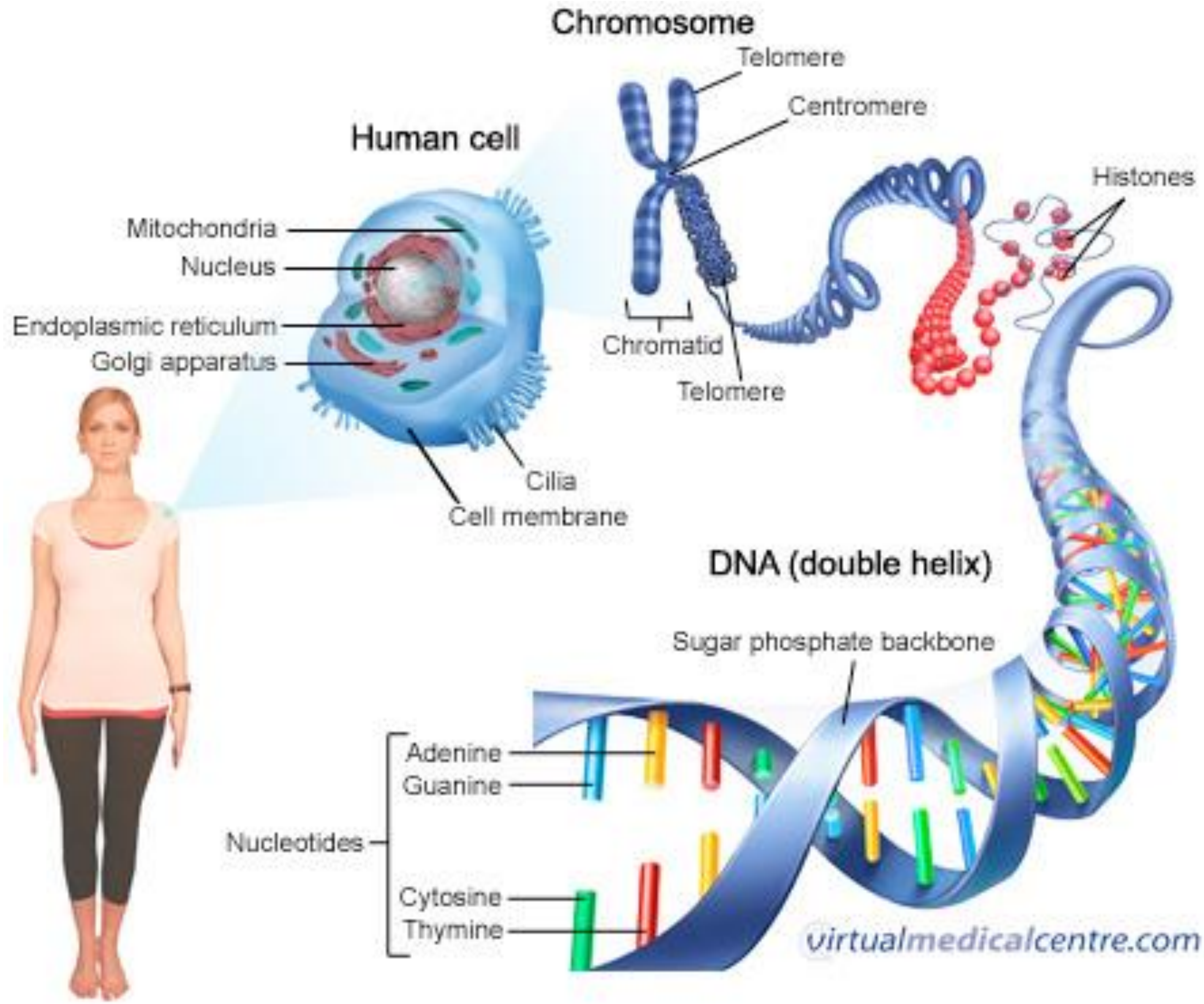
7 As galaxies cluster together under gravity, the first stars die and spew heavy elements into space; these will eventually form into new stars and planets

NOTE: The numbers in cosmology are so great and the numbers in subatomic physics are so small that it is often necessary to express them in exponential form. Ten multiplied by itself, or 100, is written as 10^{2}. One thousand is written as 10^{3}. Similarly, one-tenth is 10^{-1}, and one-hundredth is 10^{-2}.

Source: *The Birth of the Universe; The Kingfisher Young People's Book of Space*

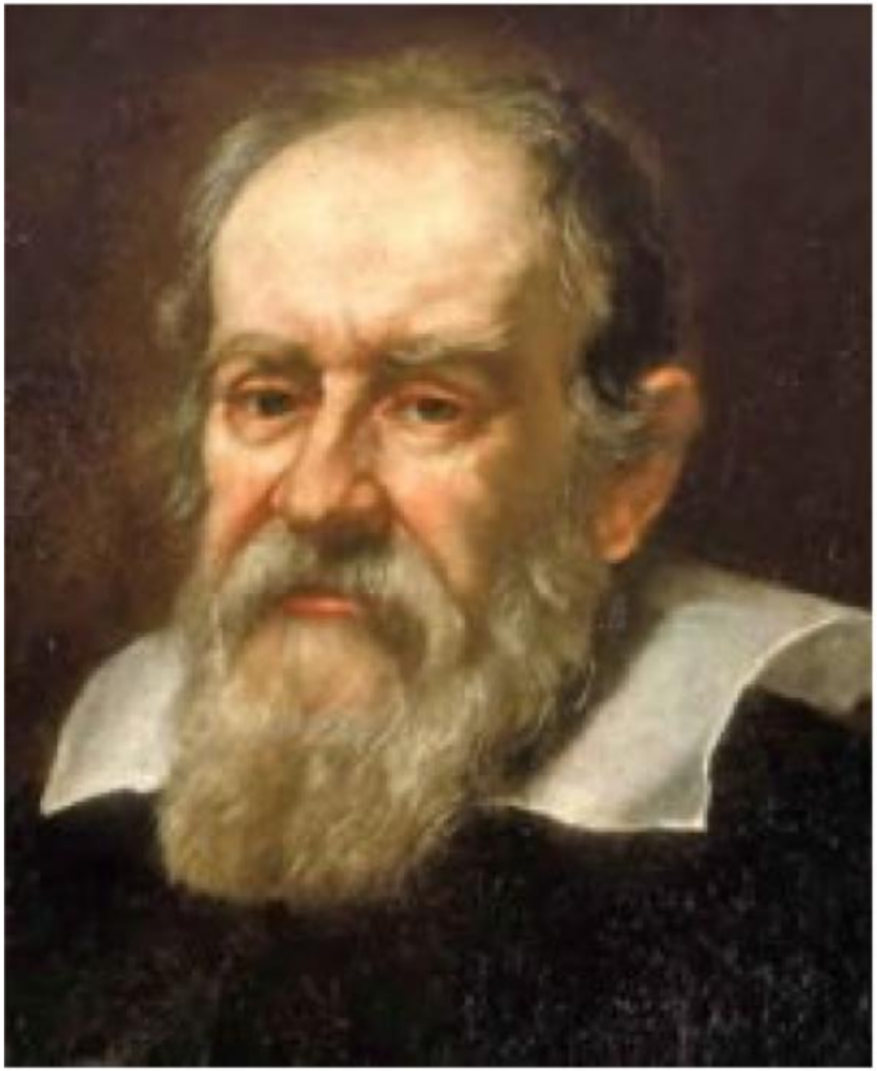
TIME Graphic by Ed Gabel

Can We Develop a New Healthcare Process that is Fully Personalized



Intermezzo to talk about a convicted felon

Galileo Galilei



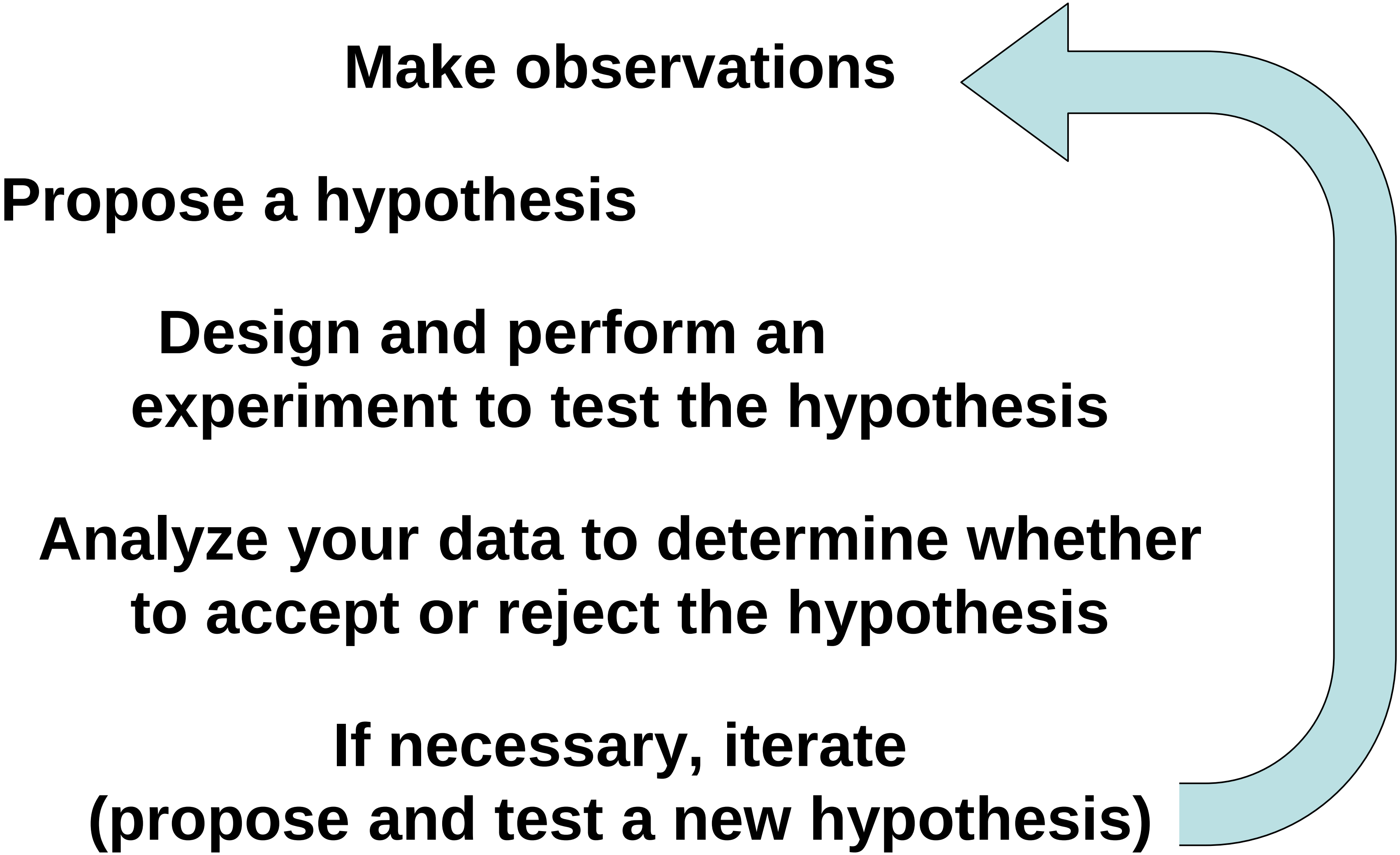
Jailed in
Apr 12, 1633
because
"gravely suspect
of heresy“

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

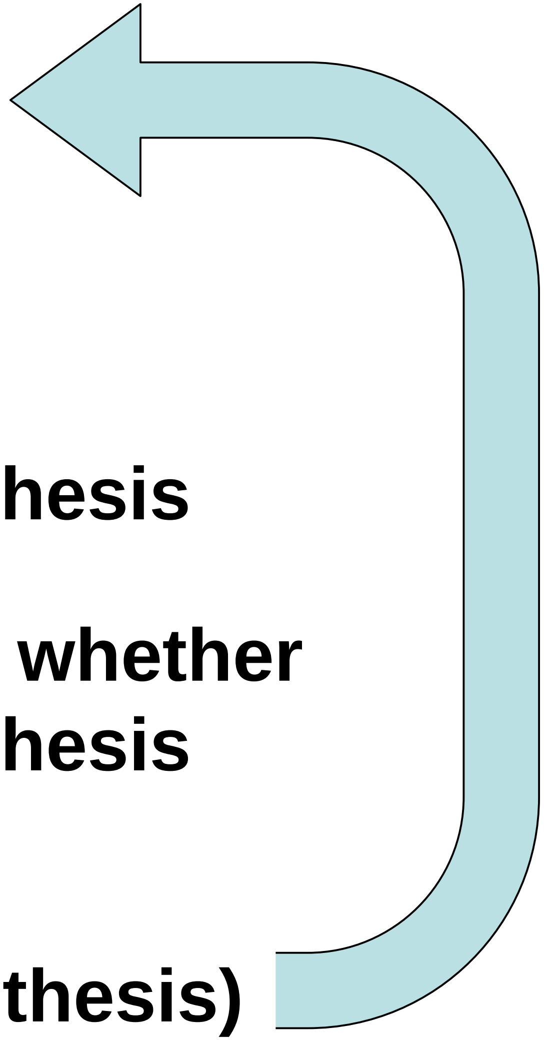
Propose a hypothesis

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

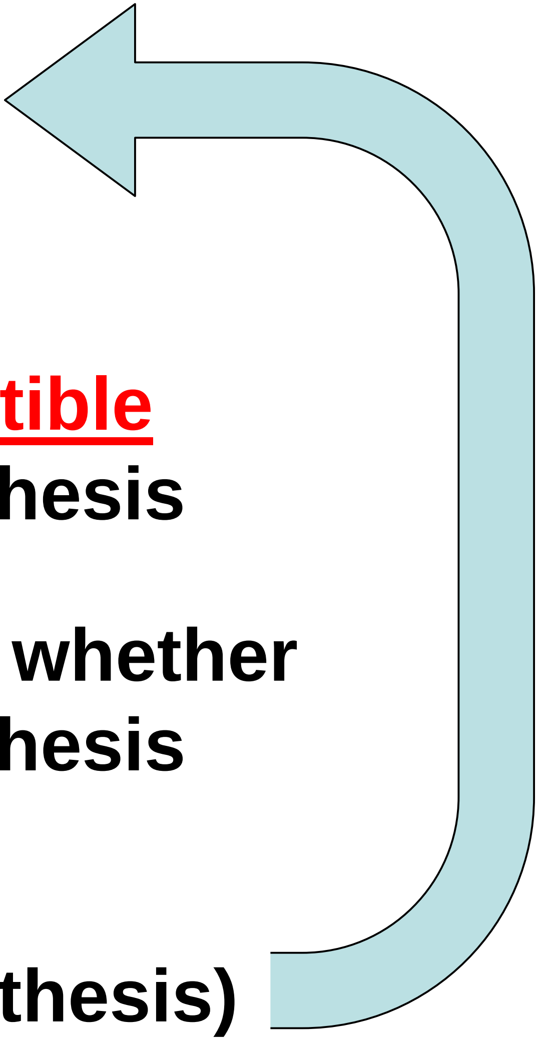
Propose a hypothesis

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

<u>repetible</u>

ypothesis

Ar

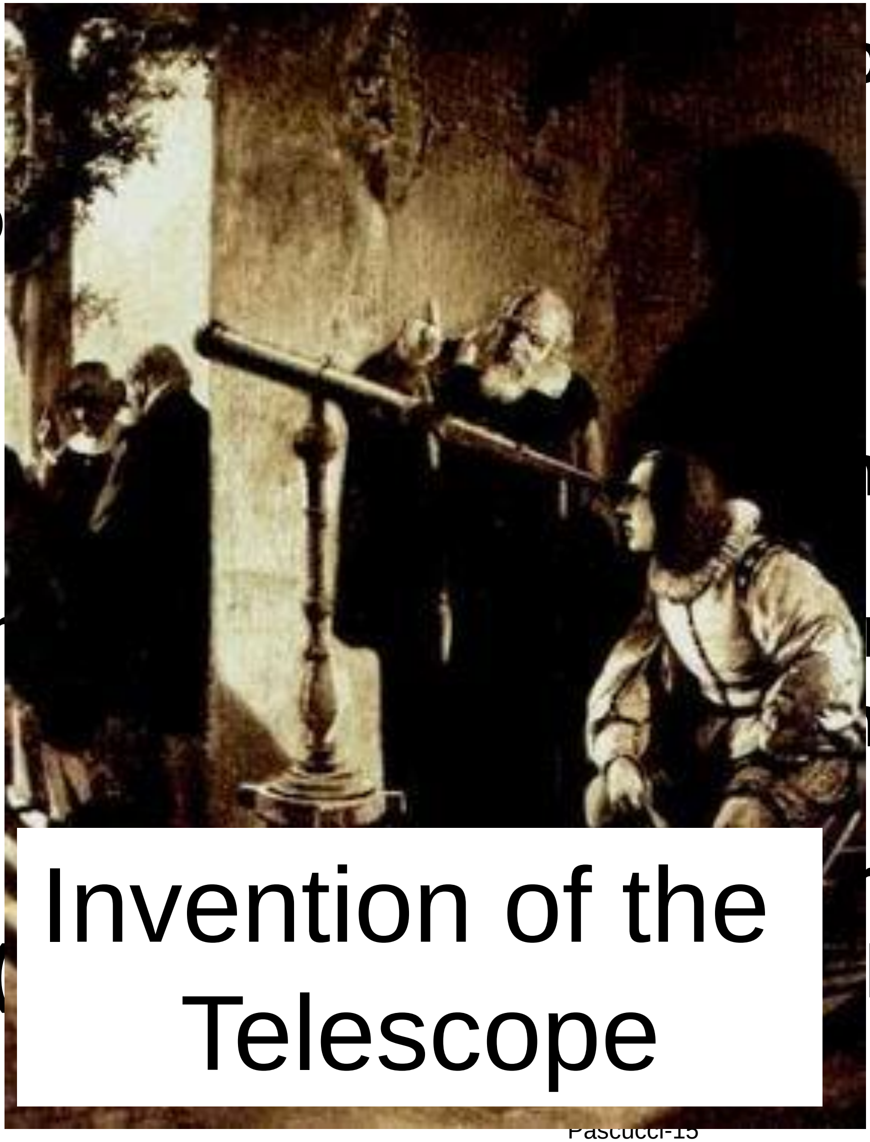
mine whether

ypothesis

rate

(

hypothesis)



Invention of the Telescope

Galileo Galilei Led the Modern Revolution of the “Scientific Method”

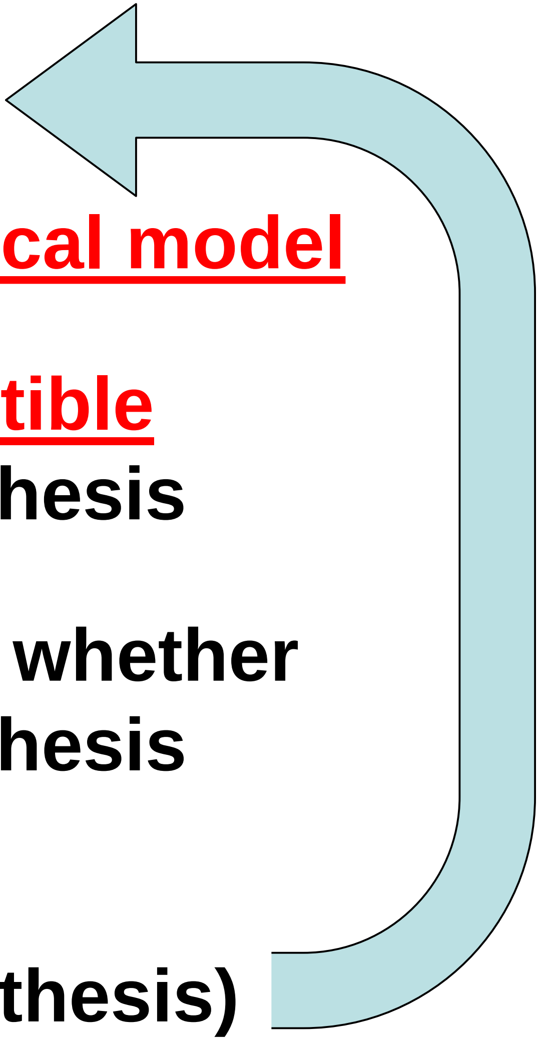
Make observations

Propose a hypothesis of <u>theoretical model</u>

Design and perform a <u>repetible</u> 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 <u>theoretical model</u>

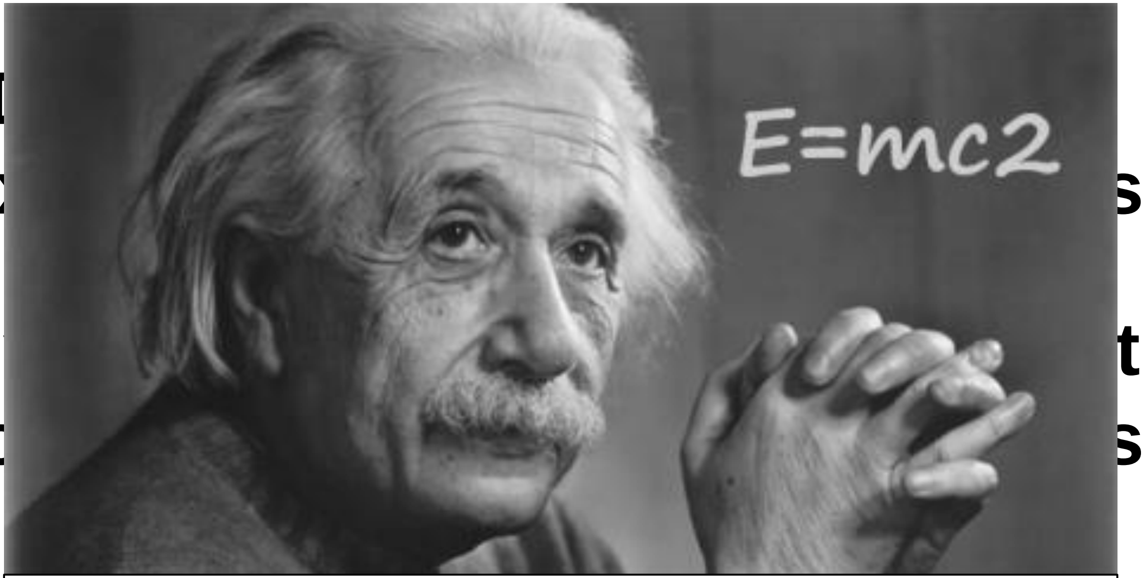
D

ex s

Anal ther

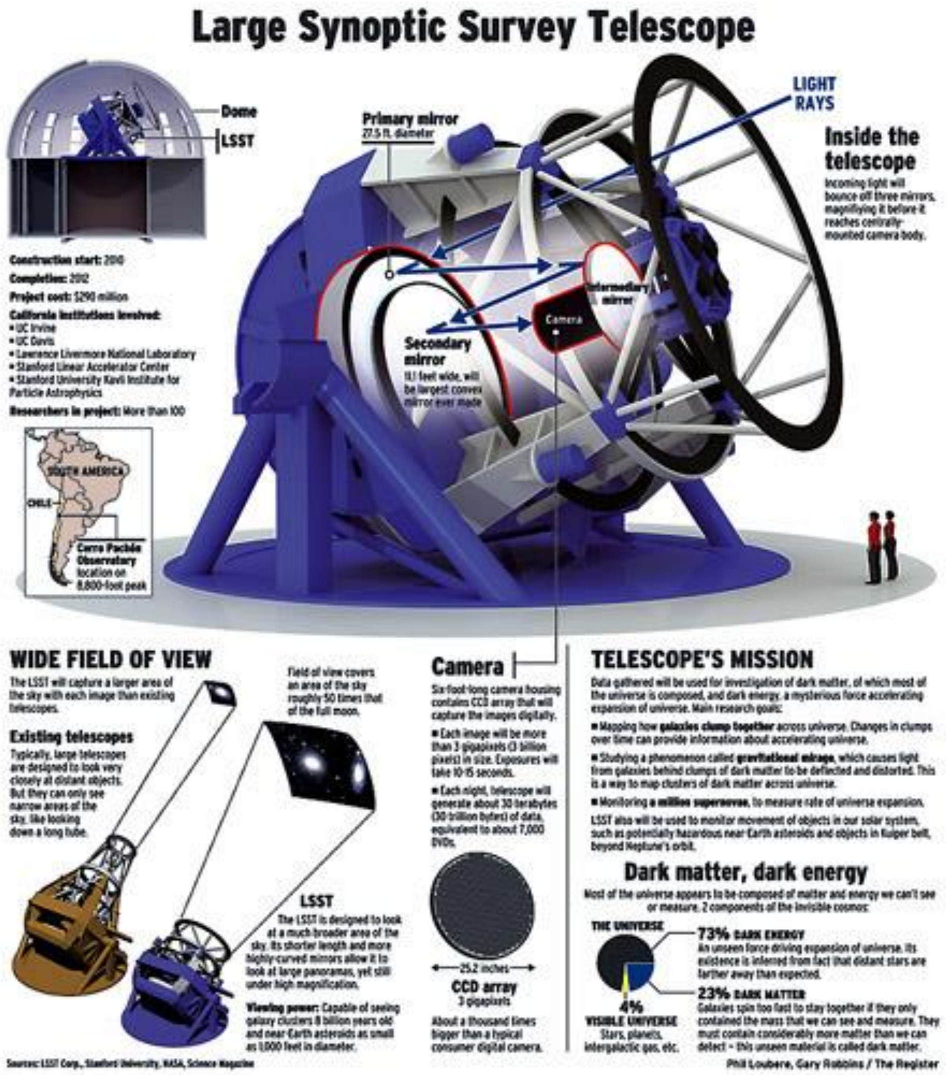
tc s

(pr is)



Development of formal model description

New Data Collection and Processing Resources Challenged This Model



Inside the telescope

Incoming light will bounce off three mirrors, magnifying it before it reaches centrally-mounted camera body.

Secondary mirror

11.1 feet wide, will be largest convex mirror ever made

Construction start: 2010
Completion: 2012
Project cost: $290 million
California institutions involved:
- UC Irvine
- UC Davis
- Lawrence Livermore National Laboratory
- Stanford Linear Accelerator Center
- Stanford University Kavli Institute for Particle Astrophysics

Researchers in project: More than 100

WIDE FIELD OF VIEW

The LSST will capture a larger area of the sky with each image than existing telescopes.

Existing telescopes

Typically, large telescopes are designed to look very closely at distant objects. But they can only see narrow areas of the sky, like looking down a long tube.

Field of view covers an area of the sky roughly 50 times that of the full moon.

LSST

The LSST is designed to look at a much broader area of the sky. Its shorter length and more highly-curved mirrors allow it to look at large panoramas, yet still under high magnification.

Viewing power: Capable of seeing galaxy clusters 8 billion years old and near-Earth asteroids as small as 1,000 feet in diameter.

Sources: LSST Corp., Stanford University, NASA, Science Magazine

Camera

Six-foot-long camera housing contains CCD array that will capture the images digitally.

- Each image will be more than 3 gigapixels (3 billion pixels) in size. Exposures will take 10-15 seconds.
- Each night, telescope will generate about 30 terabytes (30 trillion bytes) of data, equivalent to about 7,000 DVDs.

About a thousand times bigger than a typical consumer digital camera.

TELESCOPE'S MISSION

Data gathered will be used for investigation of dark matter, of which most of the universe is composed, and dark energy, a mysterious force accelerating expansion of universe. Main research goals:

- Mapping how **galaxies clump together** across universe. Changes in clumps over time can provide information about accelerating universe.
- Studying a phenomenon called **gravitational mirage**, which causes light from galaxies behind clumps of dark matter to be deflected and distorted. This is a way to map clusters of dark matter across universe.
- Monitoring **a million supernovae**, to measure rate of universe expansion.

LSST also will be used to monitor movement of objects in our solar system, such as potentially hazardous near Earth asteroids and objects in Kuiper belt, beyond Neptune's orbit.

Dark matter, dark energy

Most of the universe appears to be composed of matter and energy we can't see or measure. 2 components of the invisible cosmos:

73% DARK ENERGY: An unseen force driving expansion of universe. Its existence is inferred from fact that distant stars are farther away than expected.

23% DARK MATTER: Galaxies spin too fast to stay together if they only contained the mass that we can see and measure. They must contain considerably more matter than we can detect – this unseen material is called dark matter.

Phil Loubere, Gary Robbins / The Register

Large Synoptic Survey Telescope

15TB/day

100PB

in 10-year

New Data Collection and Processing Resources Challenged This Model

- **PayPal’s Data Volumes**

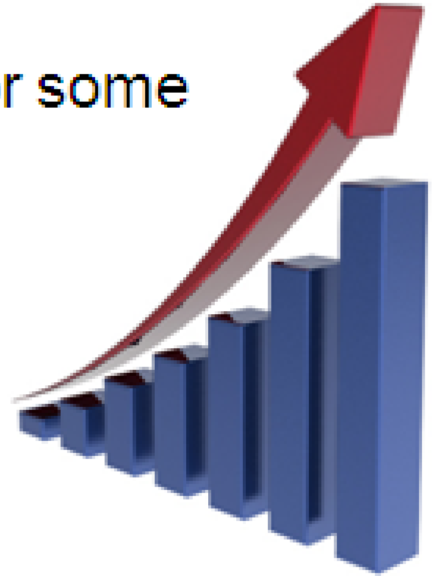
10 million+ logins / day

13 million financial transactions / day

300 variables calculated per event for some models.

~4 Billion inserts / day

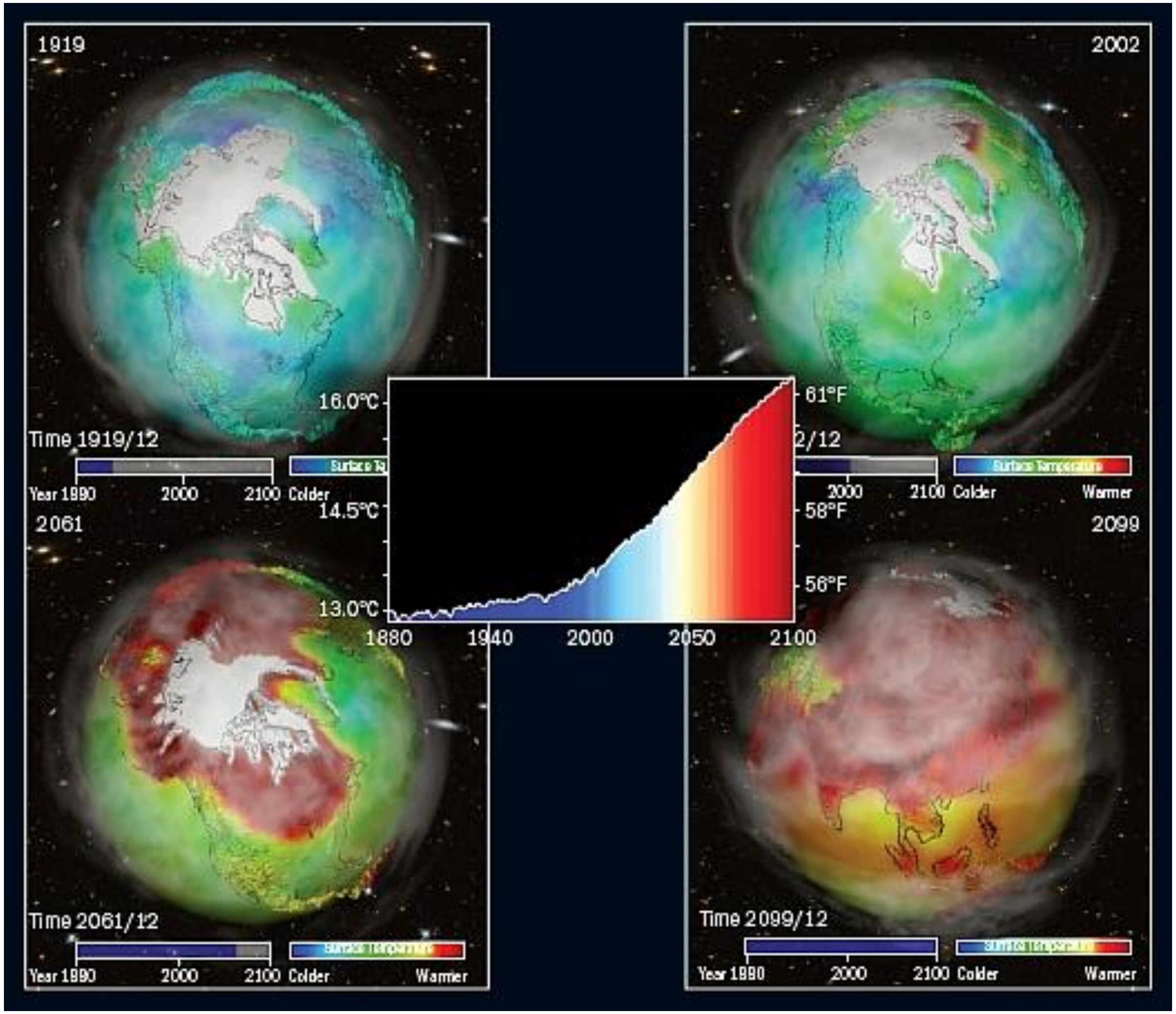
~8 Billion selects / day



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

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

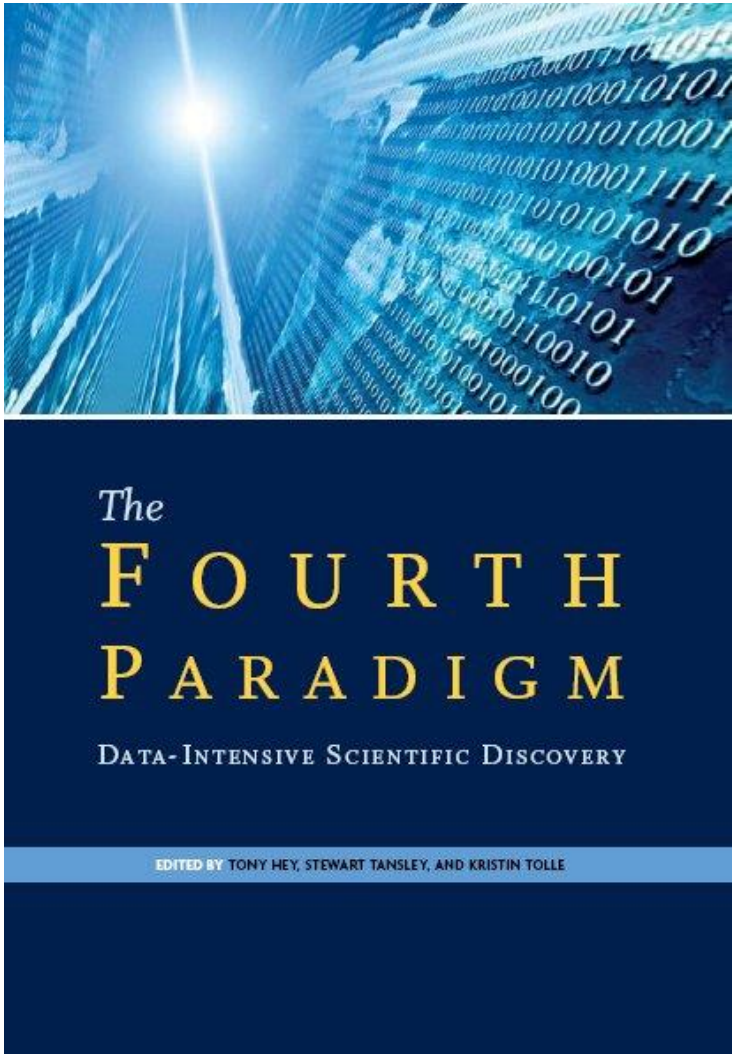
New Data Collection and Processing Resources Challenged This Model

- 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
- 3rd paradigm, computational science (simulations)
- 4th paradigm, data driven investigation (eScience)



The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

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

Milky Way crashes into Andromeda system.

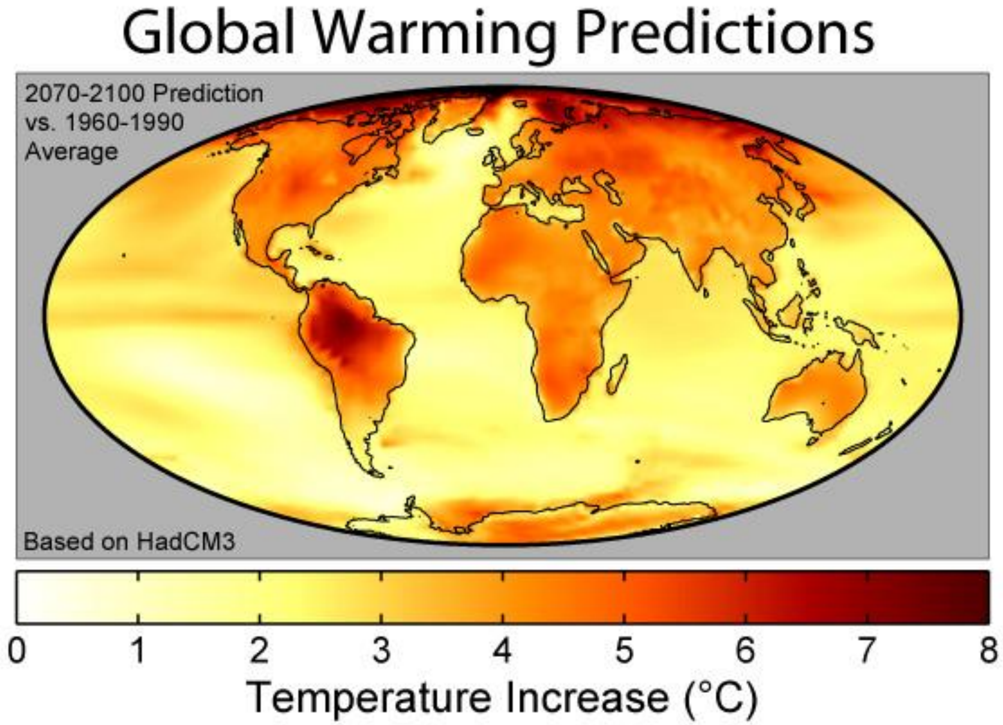
Can we test Empirically multiple scenarios?

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?

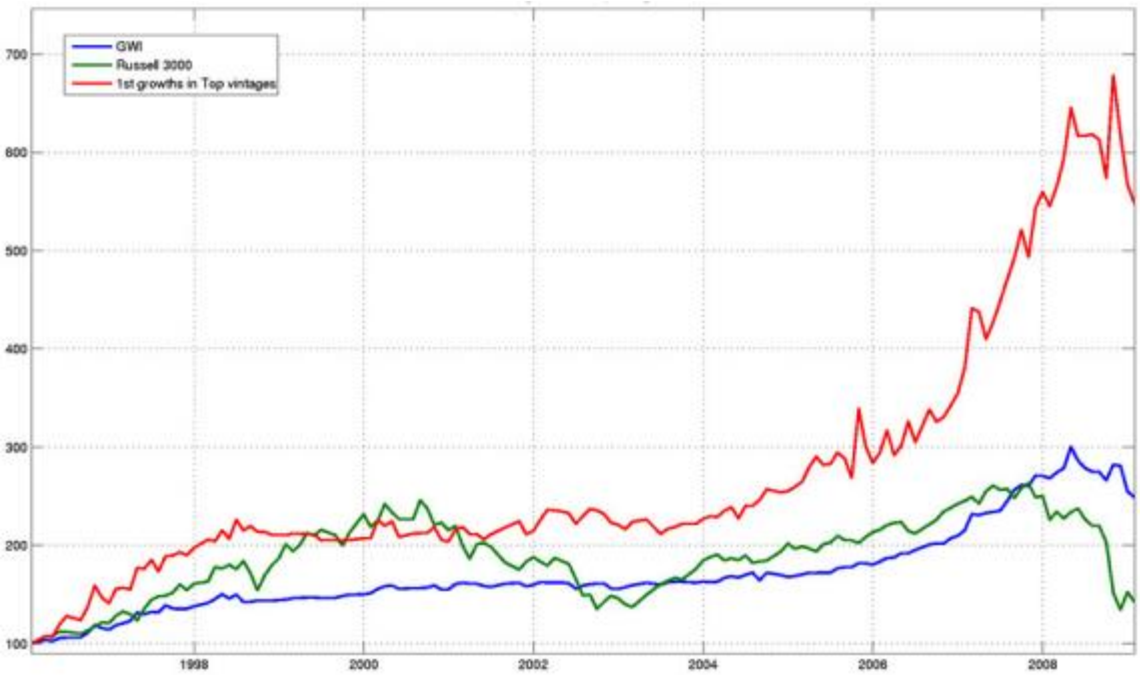


Predict the outcome in 2010

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?

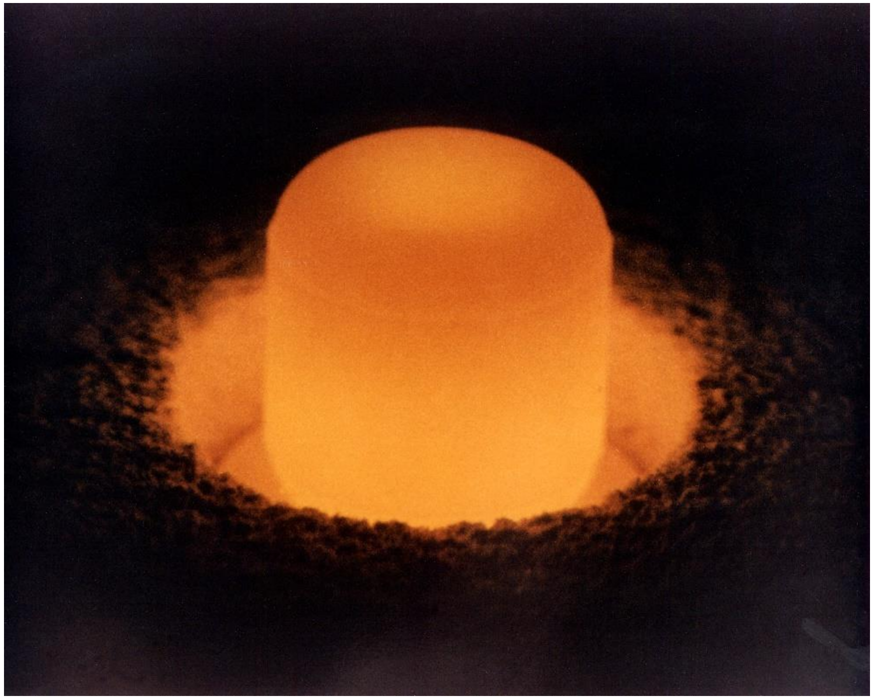


Value of investments in 10 years

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 T)

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

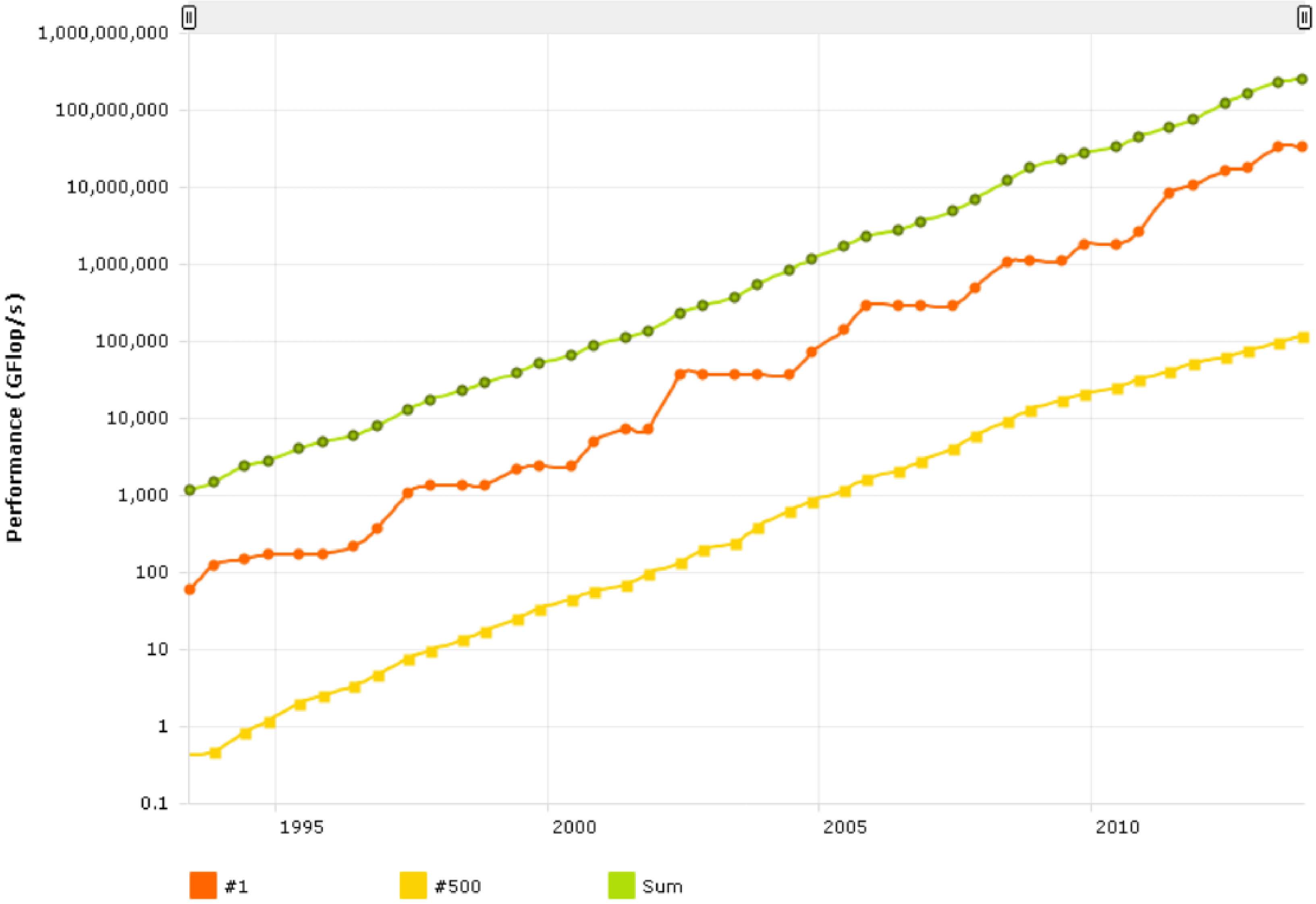
Titan: Oak Ridge National Laboratory
No. 1 in November 2012

Unprecedented Evolution of High Performance Computing Resources

Tianhe-2 (MilkyWay-2)

No. 1 system since June 2013

Unprecedented Evolution of High Performance Computing Resources

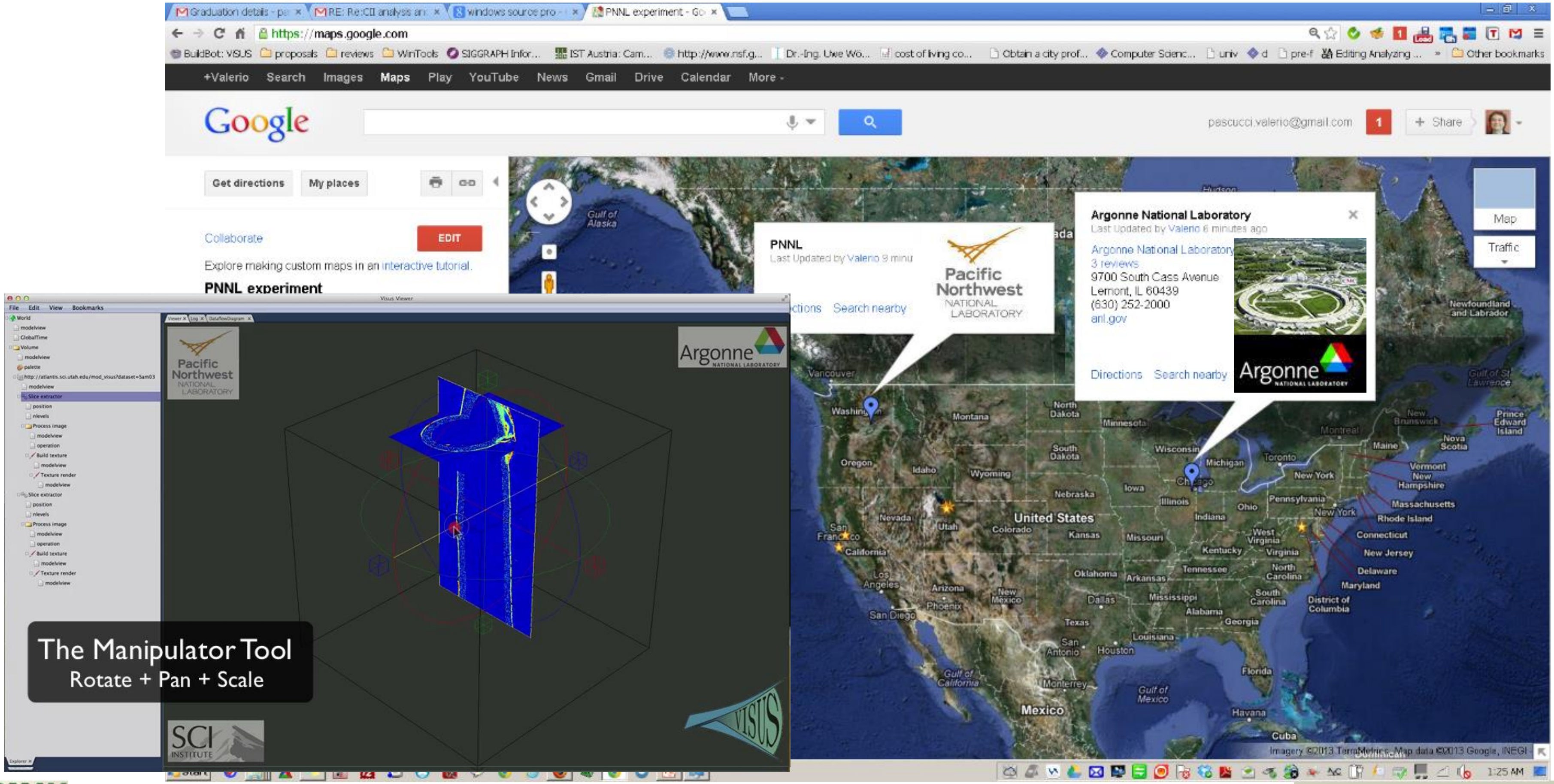


The Fundamental Paradigm Shift of eScience

- The data is the 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
- 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



https://maps.google.com

Collaborate EDIT

Explore making custom maps in an interactive tutorial.

PNNL experiment

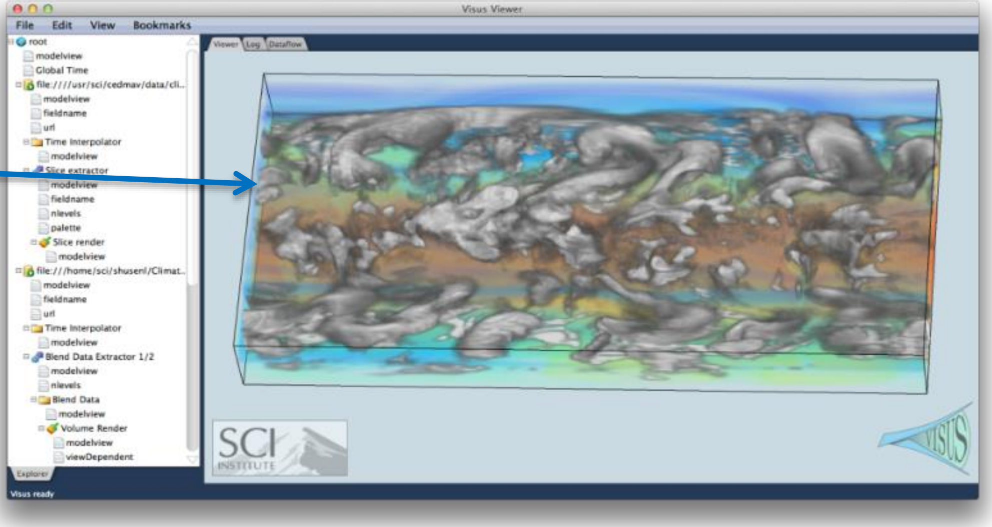
PNNL
Last Updated by Valerio 9 minu

Argonne National Laboratory
Last Updated by Valerio 6 minutes ago
Argonne National Laboratory
3 reviews
9700 South Cass Avenue
Lemont, IL 60439
(630) 252-2000
anl.gov
Directions Search nearby

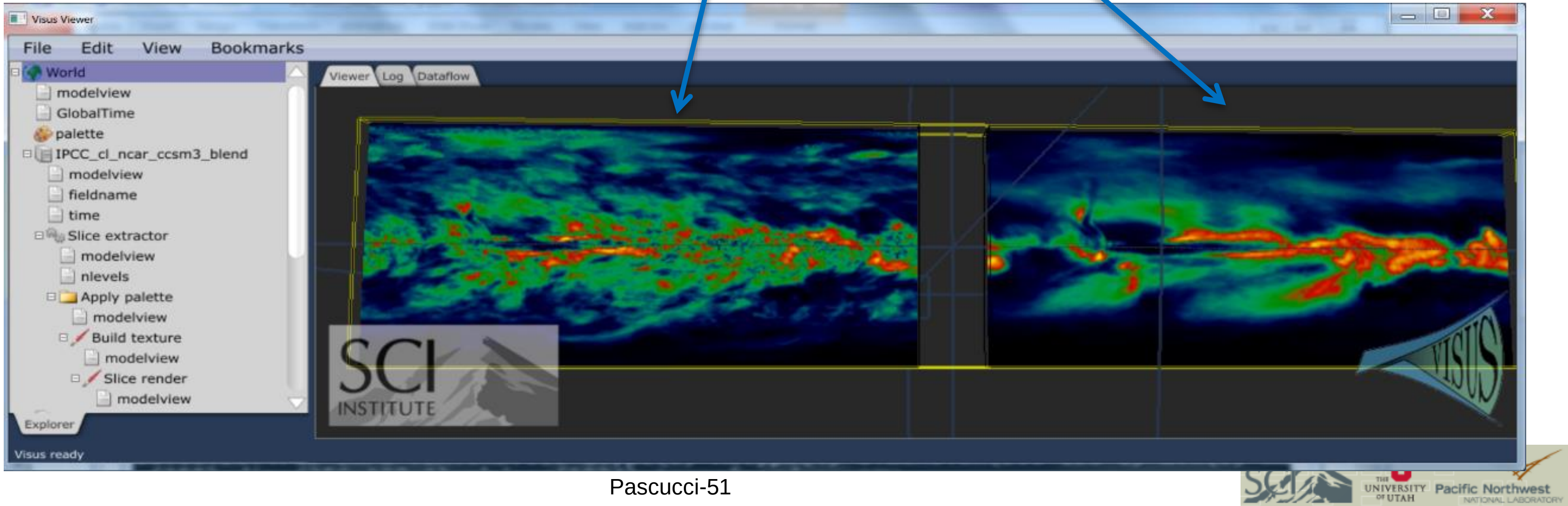
The Manipulator Tool
Rotate + Pan + Scale

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, …….

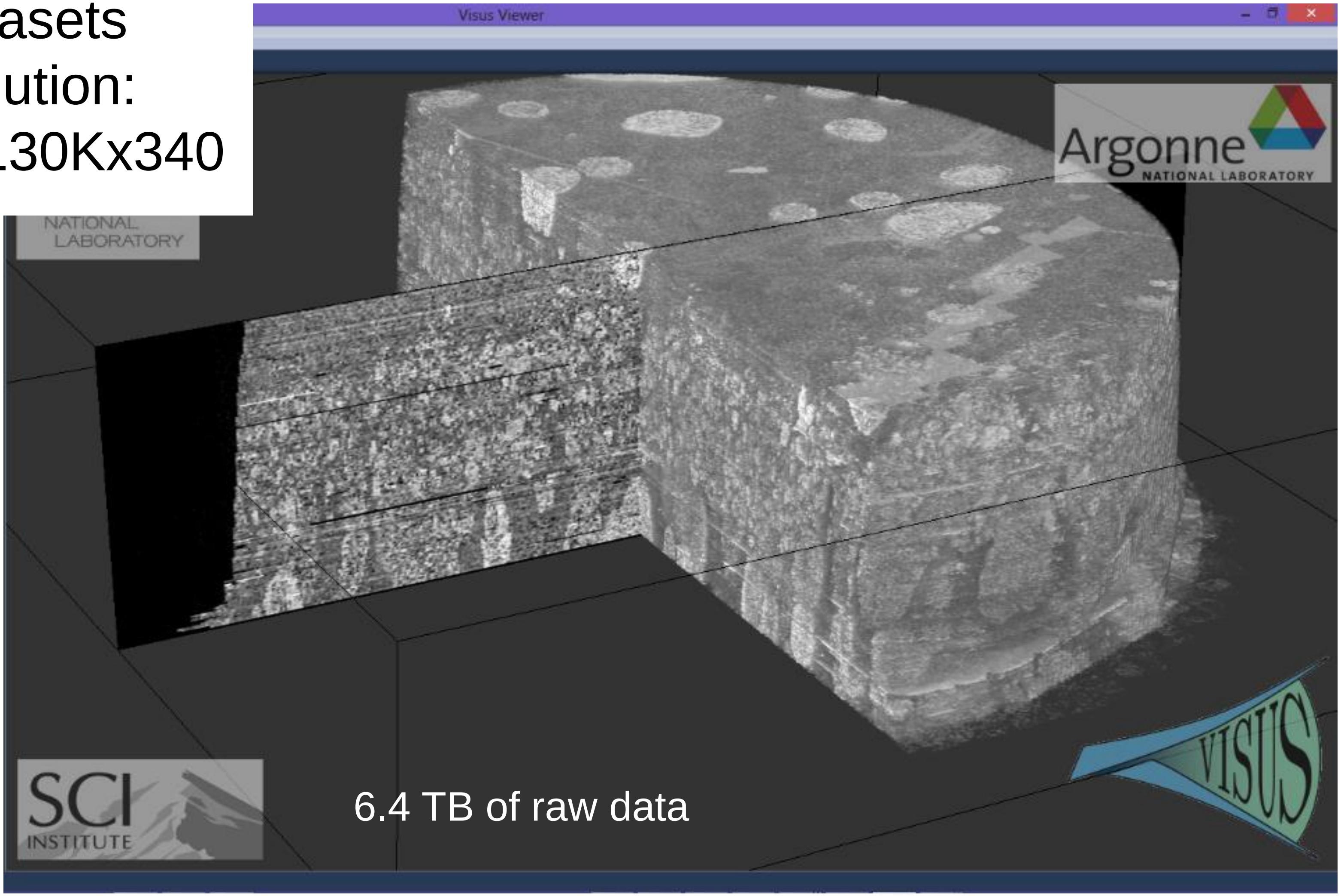


Standard Deviation and Average of ten climate models



Interactive Remote Analysis and Visualization of 6TB Imaging Data

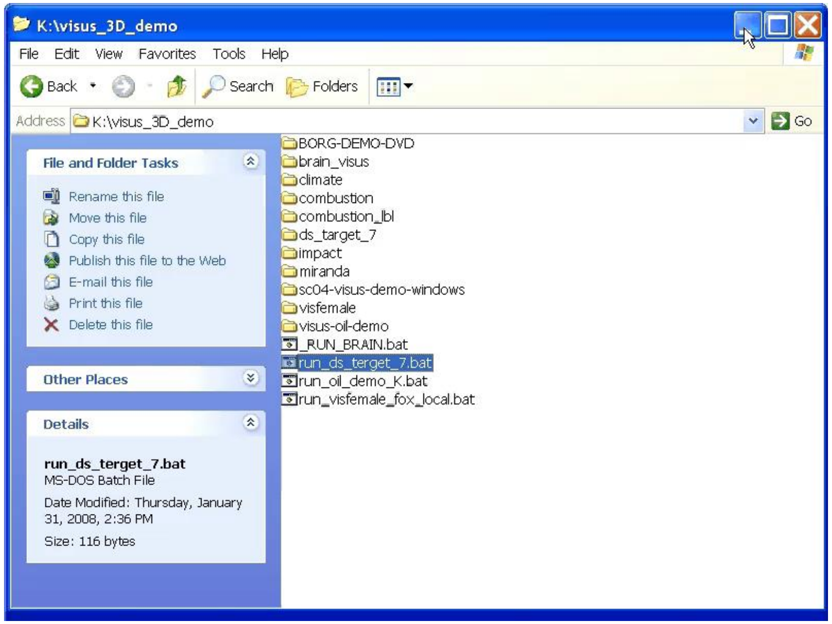
- EM datasets of resolution: 130Kx130Kx340



6.4 TB of raw data

Web Server

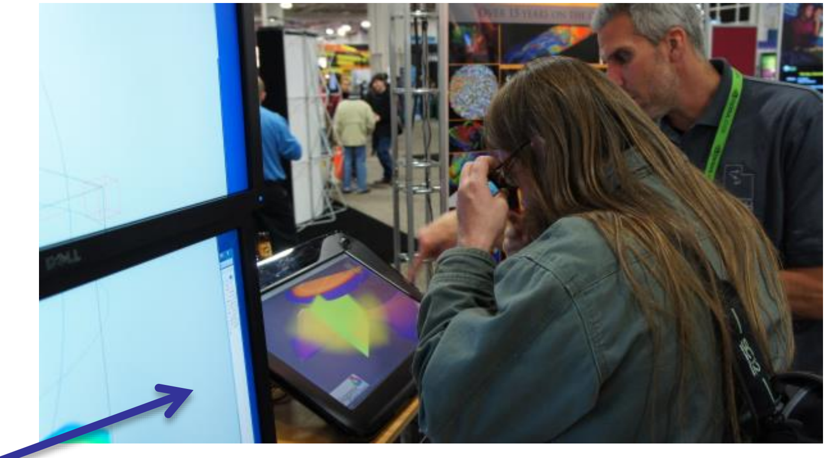
SC12/13 Demonstration of data streaming analytics and visualization



K:\visus_3D_demo

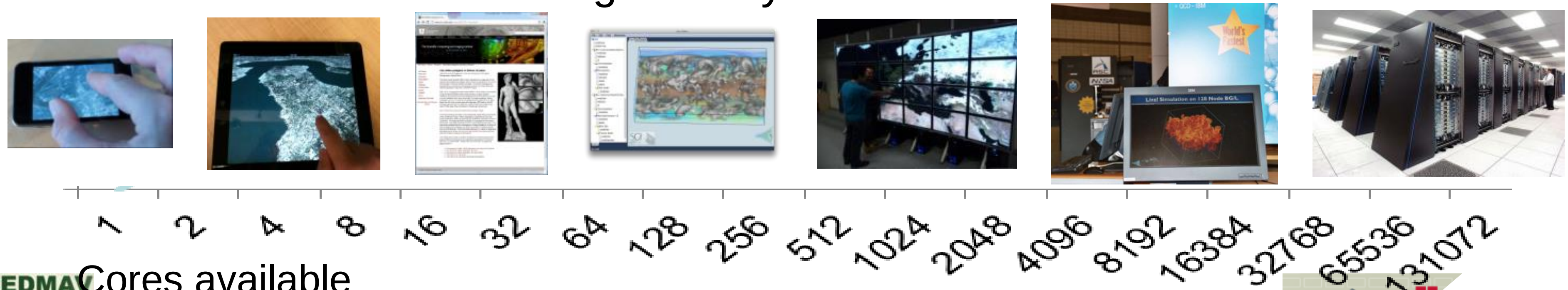
- BORG-DEMO-DVD
- brain_visus
- climate
- combustion
- combustion_lbl
- ds_target_7
- impact
- miranda
- sc04-visus-demo-windows
- visfemale
- visus-oil-demo
- _RUN_BRAIN.bat
- run_ds_terget_7.bat
- run_oil_demo_K.bat
- run_visfemale_fox_local.bat

run_ds_terget_7.bat
MS-DOS Batch File
Date Modified: Thursday, January 31, 2008, 2:36 PM
Size: 116 bytes



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 every American adult and household (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

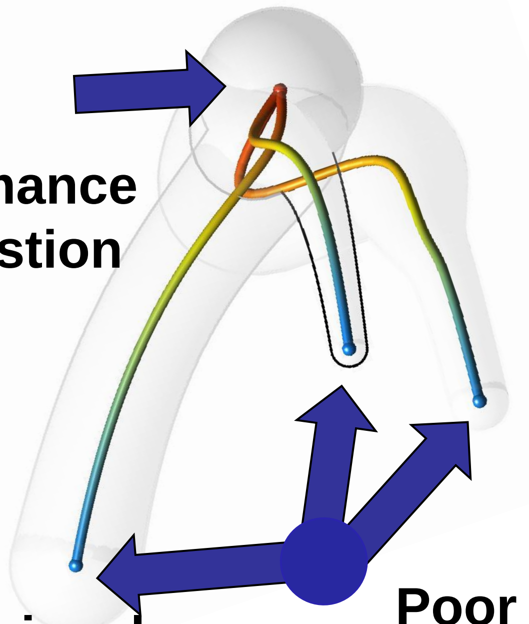
Need to develop scientific knowledge and actionable information that cannot be tested empirically

Uncertainty Quantification

Verification and Validation

Assessing the Uncertainty in Fuel Design For New Clean Burning Devices

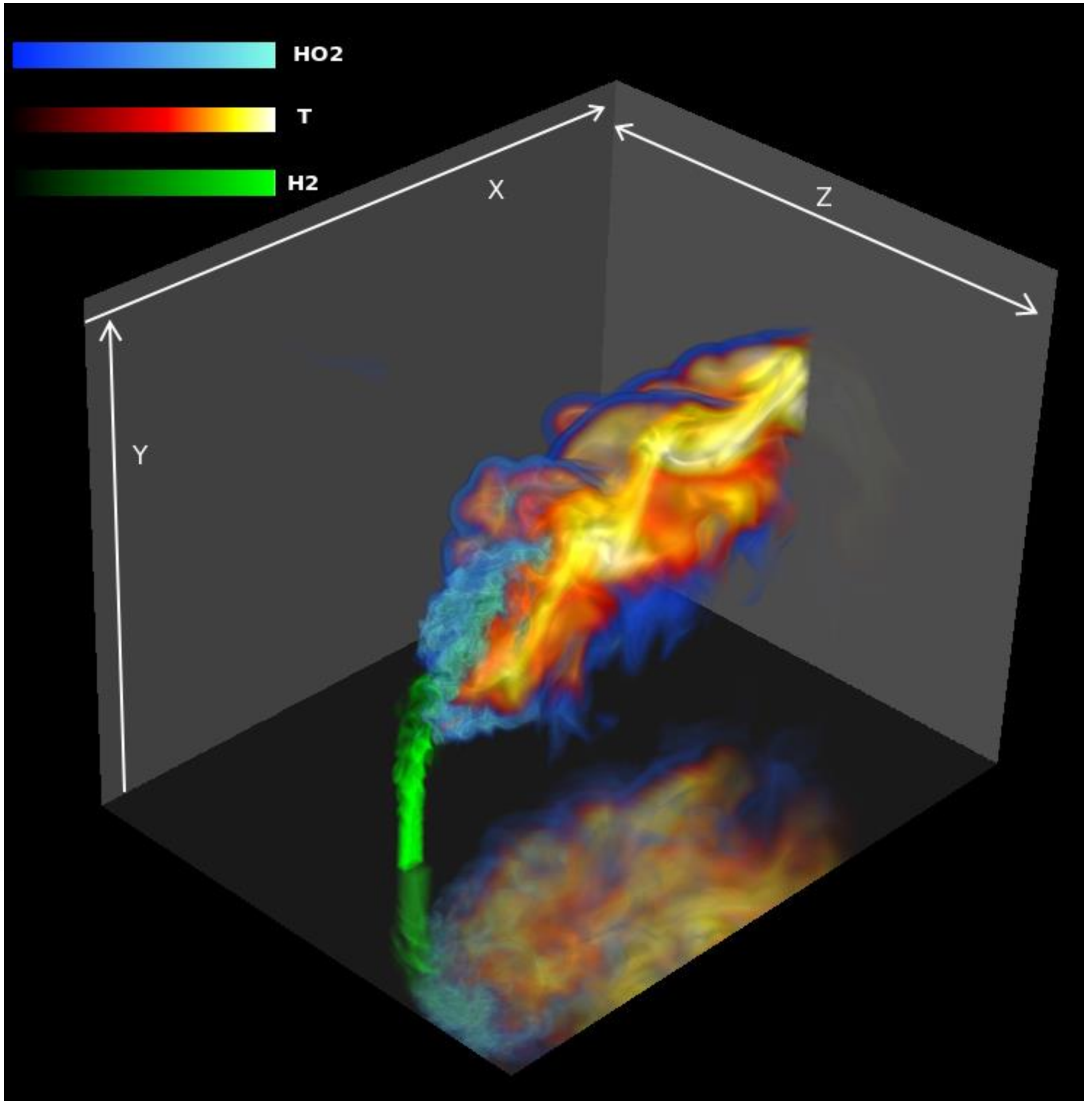
Exploration of high dimensional space of possible “configurations.”



High Performance Combustion

10 dimensional landscape

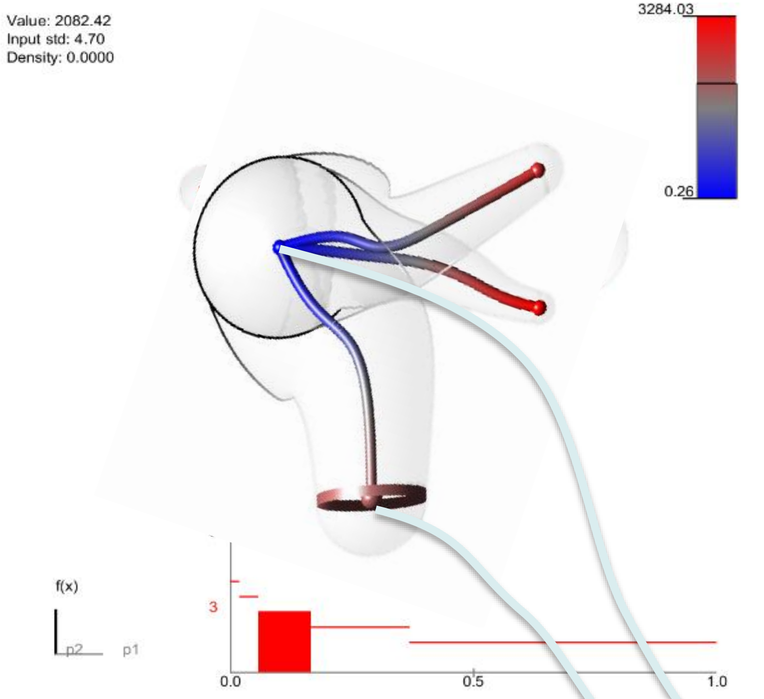
Poor Combustion to be avoided

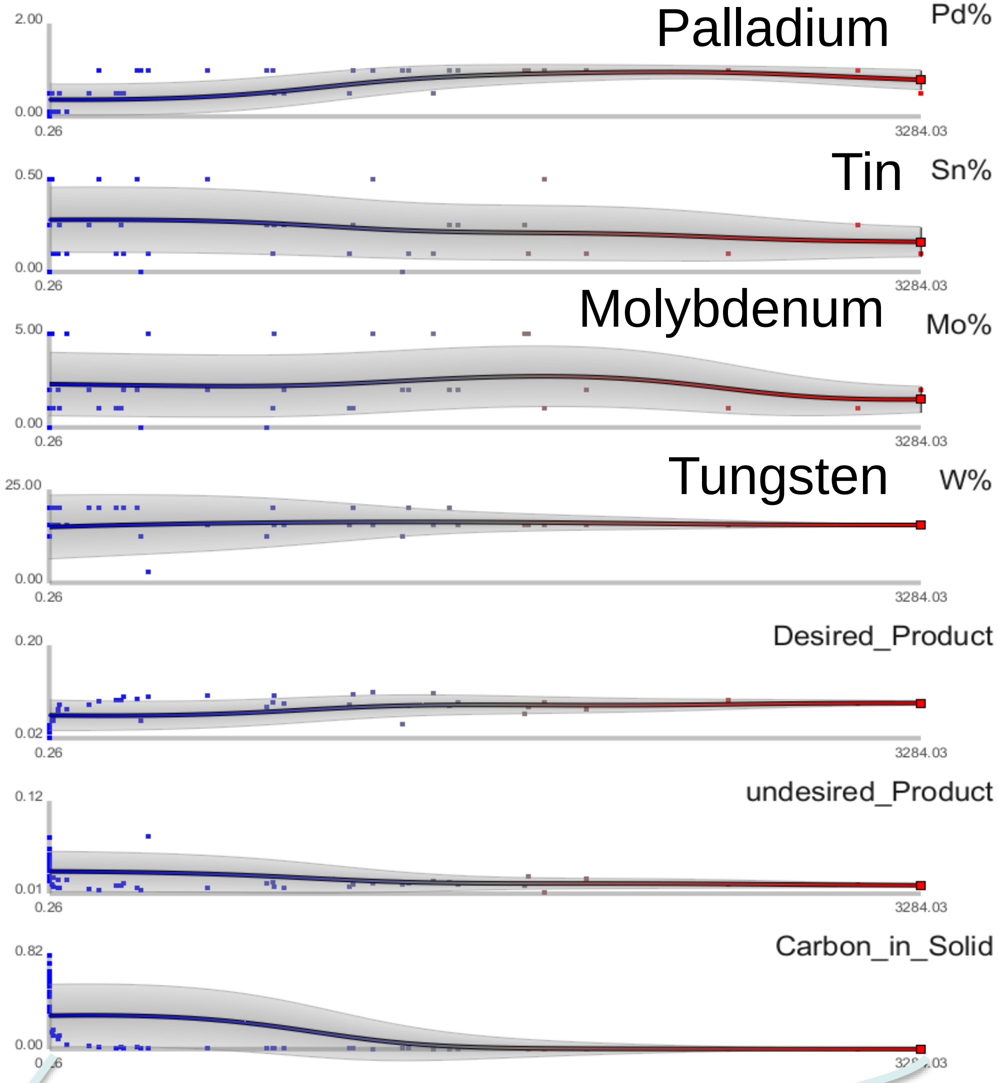


Topological Analysis of the Space of Composite Materials of a Given Class

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

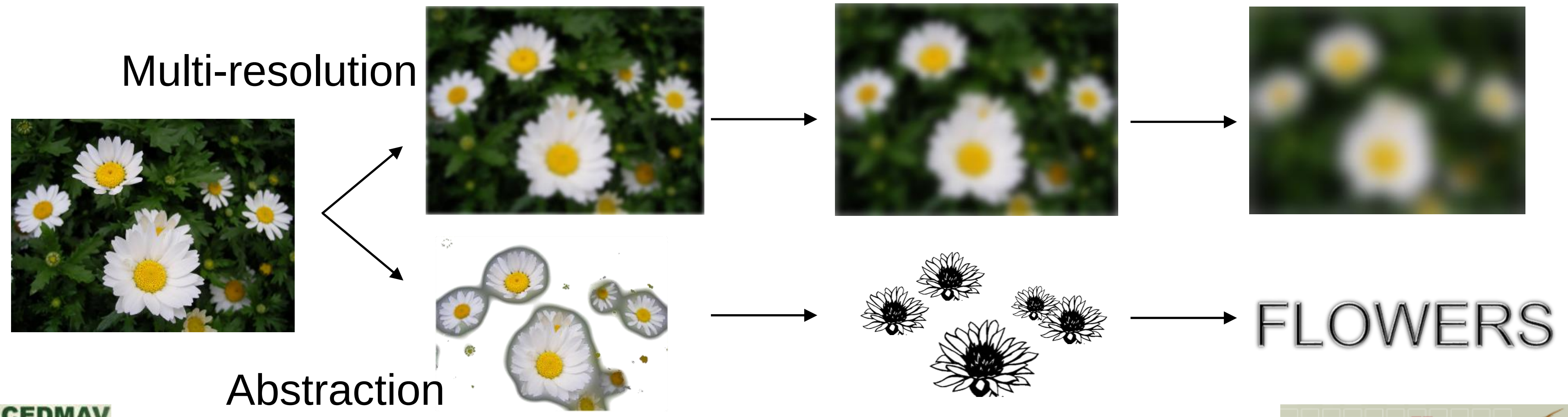
  Stakeholder: A. Karim, PNNL.



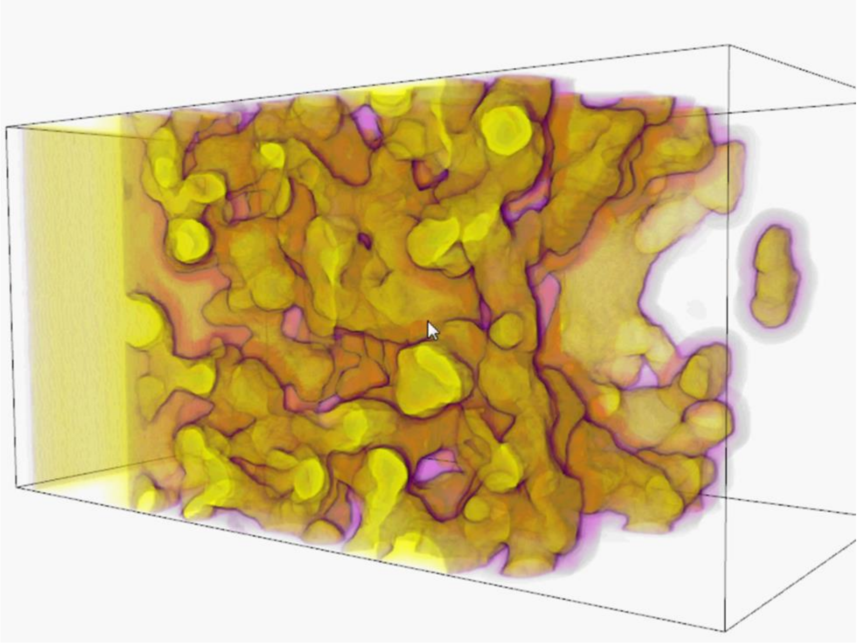


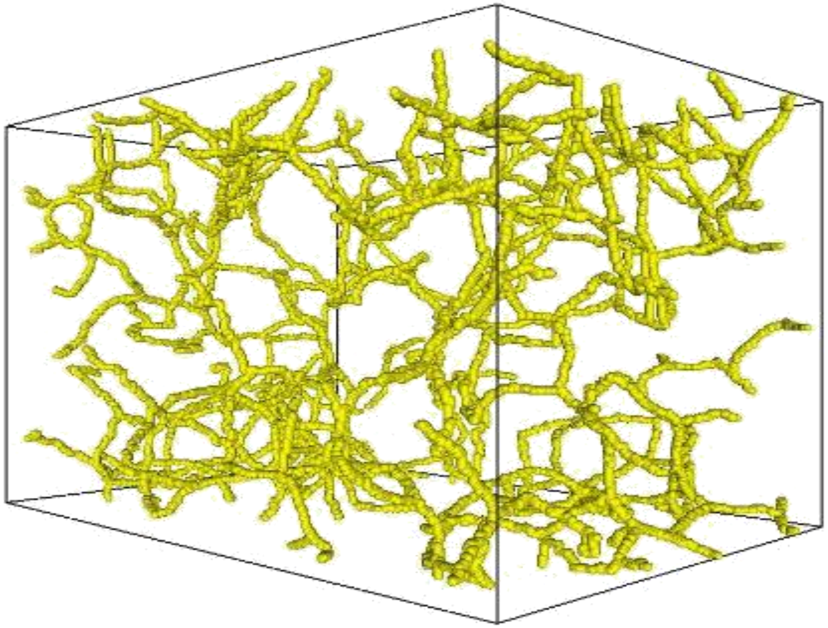
Rethinking Multi-Scale Representation of Massive Data Models

- **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**



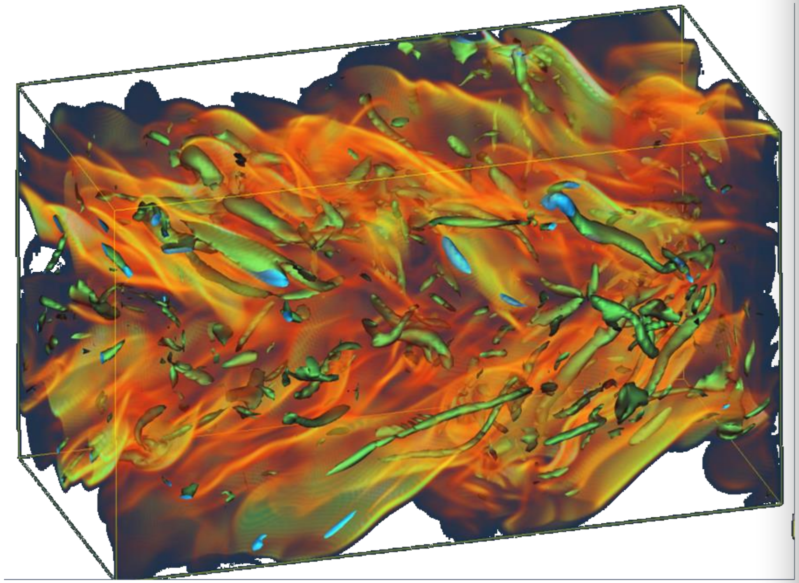
Rethinking Multi-Scale Representation of Massive Data Models

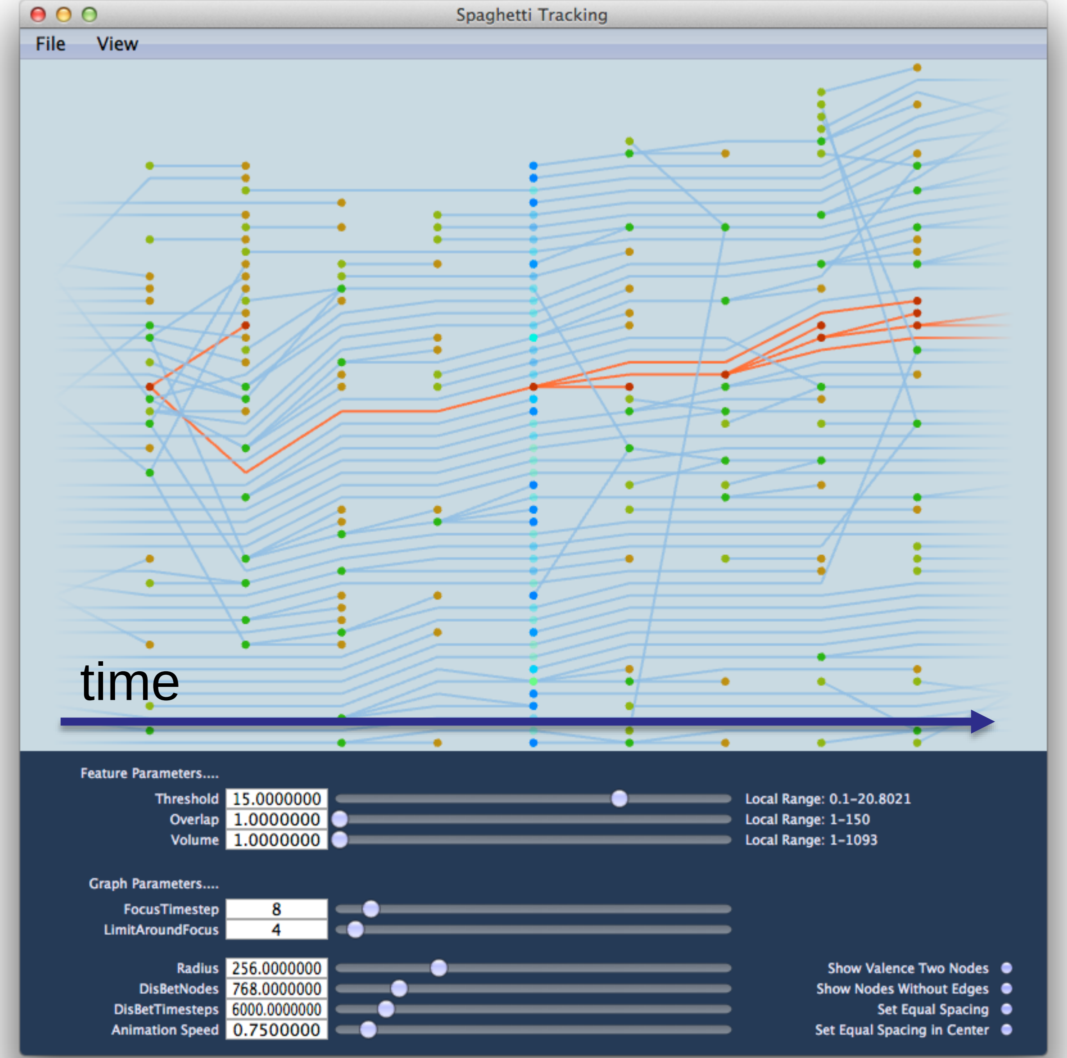




Topological Analysis of Massive Combustion Simulations

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





Feature Parameters....

Threshold 15.0000000 Local Range: 0.1-20.8021

Overlap 1.0000000 Local Range: 1-150

Volume 1.0000000 Local Range: 1-1093

Graph Parameters....

FocusTimestep 8

LimitAroundFocus 4

Radius 256.0000000

DisBetNodes 768.0000000

DisBetTimesteps 6000.0000000

Animation Speed 0.7500000

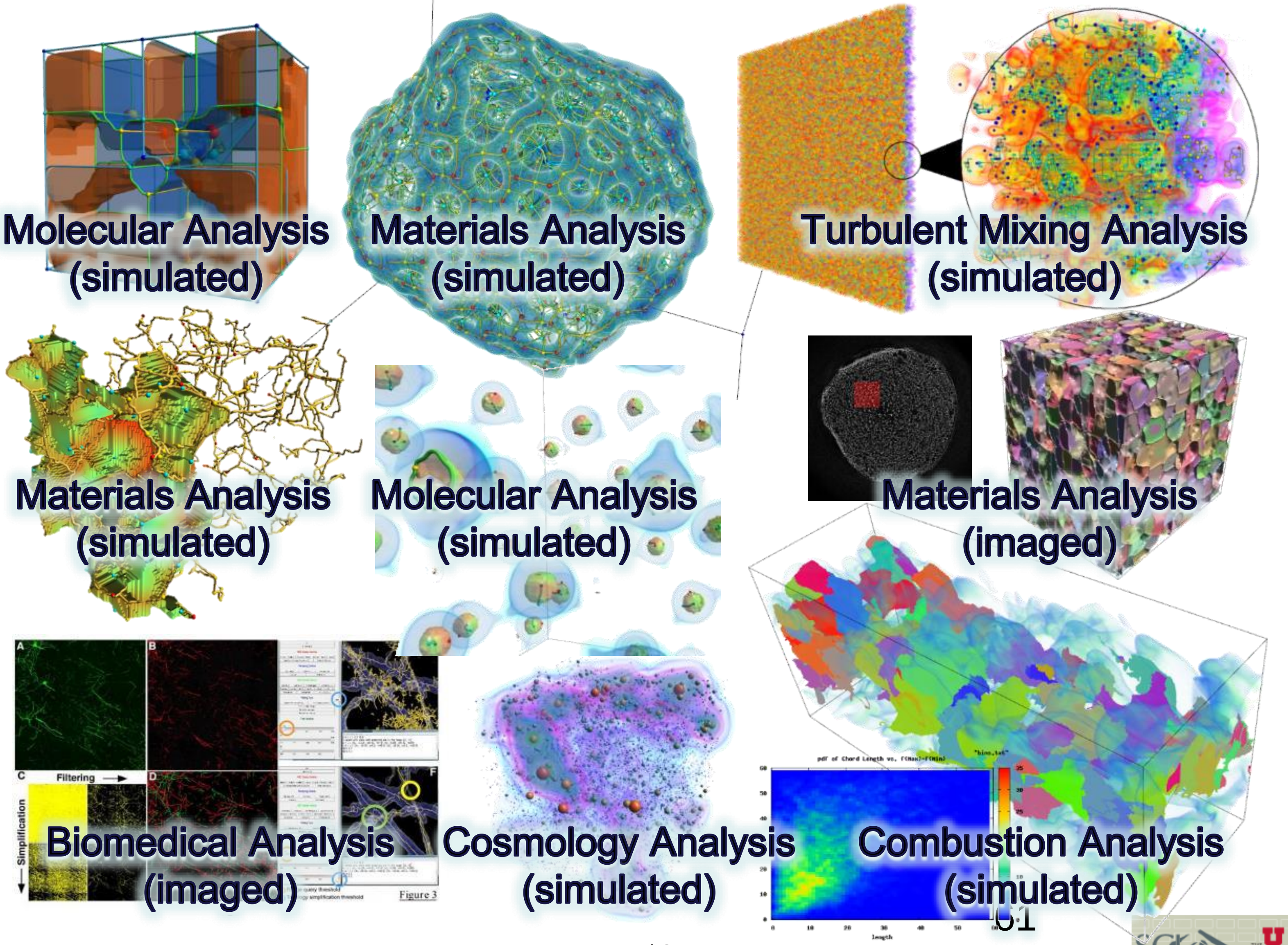
Show Valence Two Nodes ●

Show Nodes Without Edges ●

Set Equal Spacing ●

Set Equal Spacing in Center ●

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



Molecular Analysis (simulated)

Materials Analysis (simulated)

Turbulent Mixing Analysis (simulated)

Materials Analysis (simulated)

Molecular Analysis (simulated)

Materials Analysis (imaged)

Biomedical Analysis (imaged)

Cosmology Analysis (simulated)

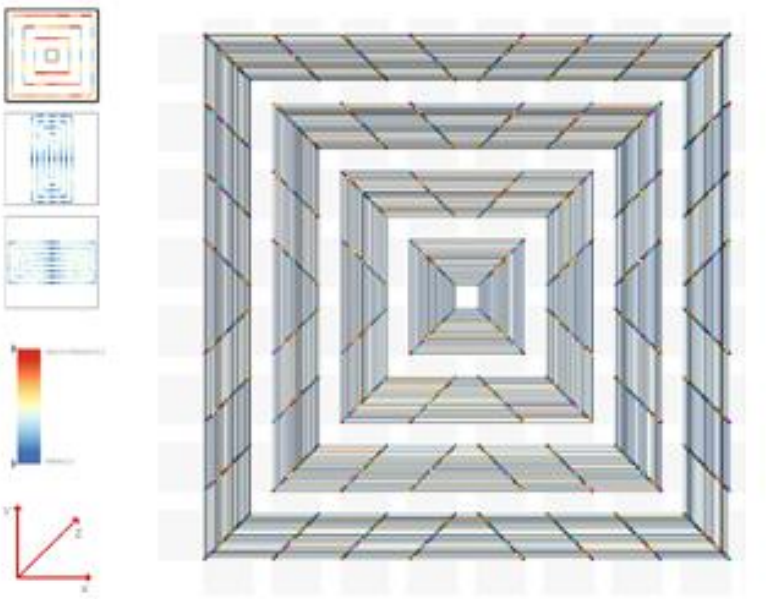
Combustion Analysis (simulated)

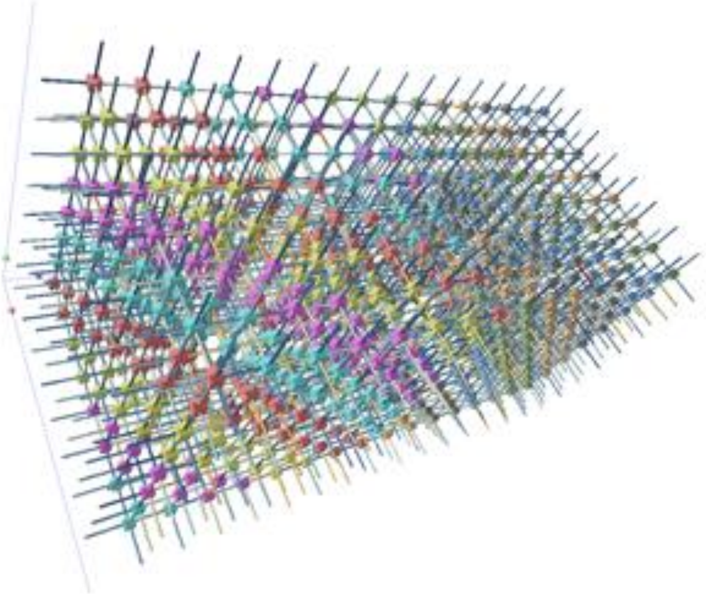
Running Efficiently Big Data Computations is a Big Data Problem

- 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

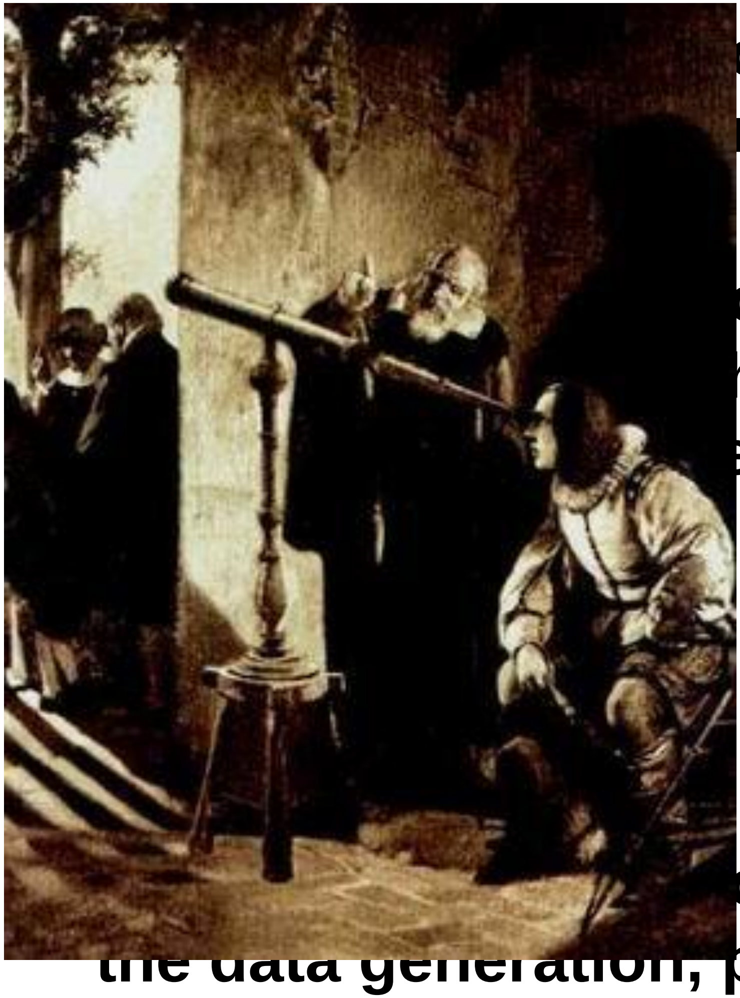
- **A great *opportunity* to achieve new scientific discoveries and engineering innovations**

- **A great *opportunity* for the Computer Science community to become a central player in the development 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



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

END