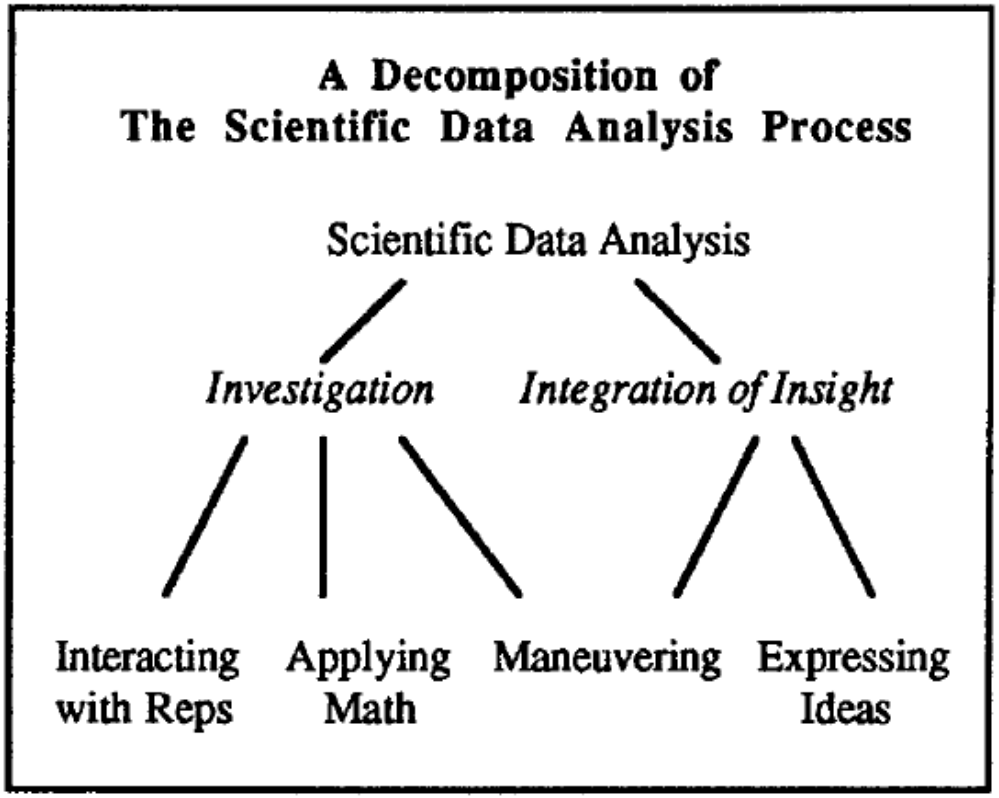
Lecture 1

Large Data Analysis and Visualization (LDAV)

30th July 2024

Data Analysis Process

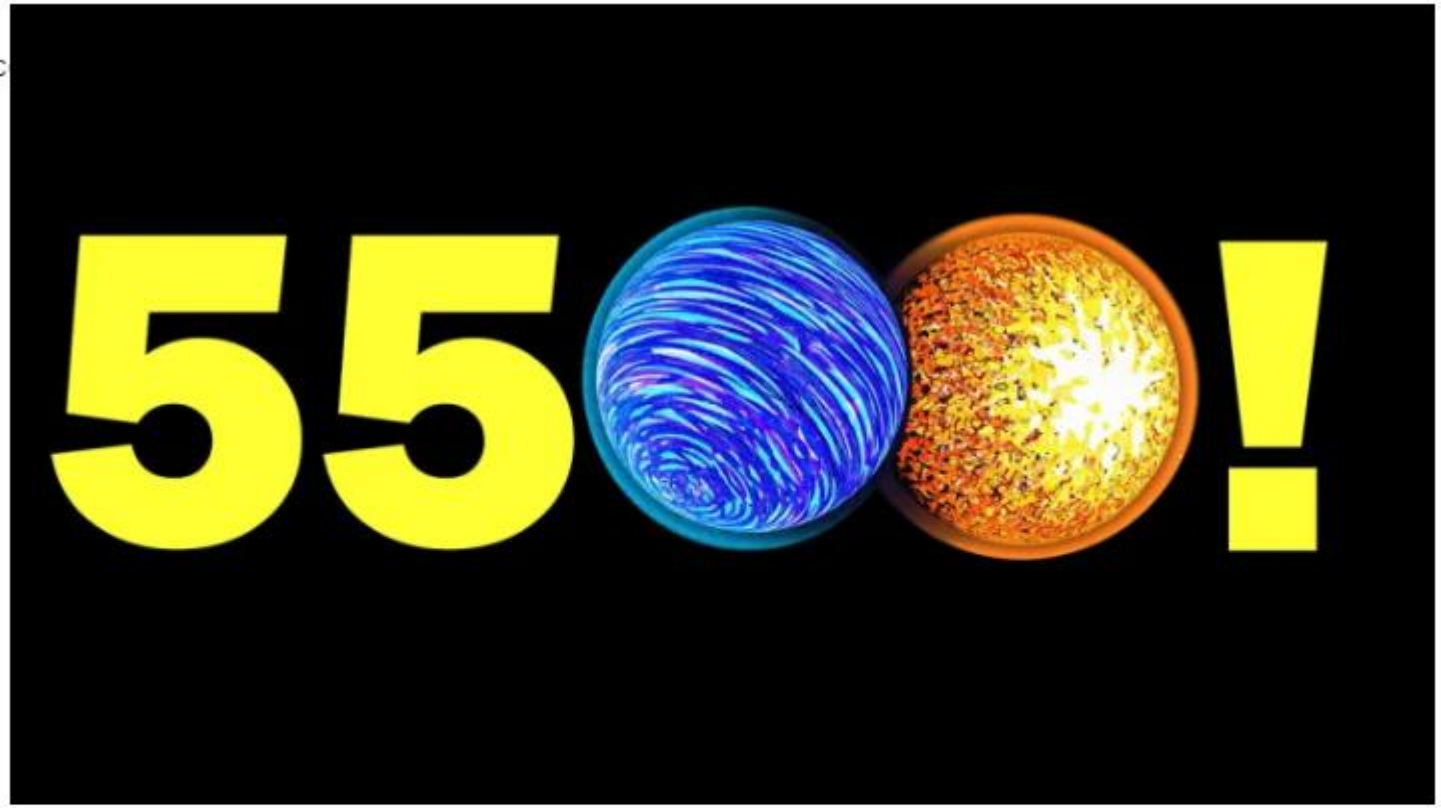


"Scientific data analysis is the process of distilling potentially large amounts of measured or calculated data into a few simple rules or parameters which characterize the phenomenon under study. This may include quantifying known analytical relations between data parameters and inferring previously unknown relations."

A characterization of the scientific data analysis process, Springmeyer et al., IEEE VIS, 1992

Sources of Large Data?

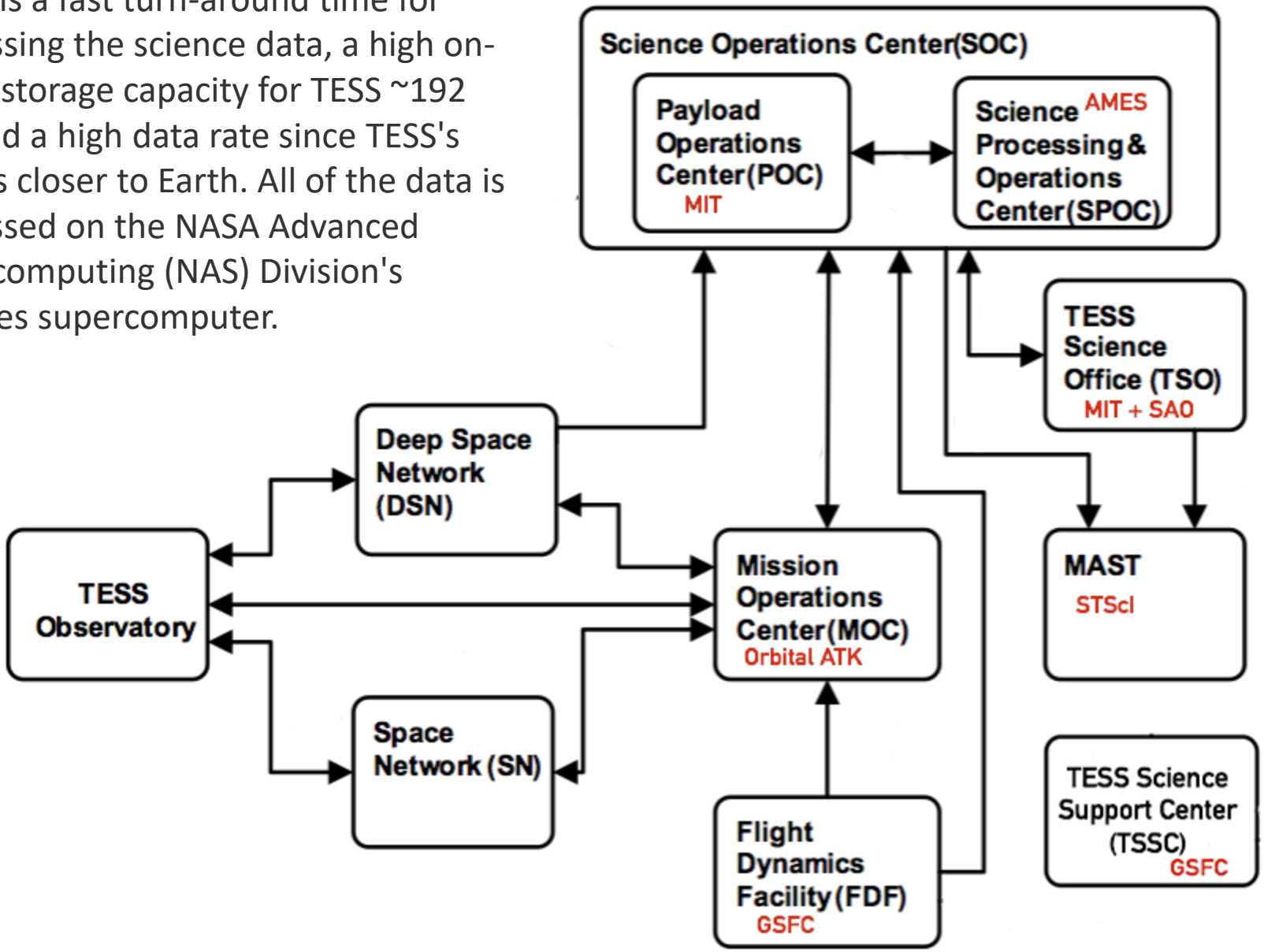

ISRO
INDIA



NASA's Exoplanet Archive confirmed four new worlds, bringing the total past 5,500.
NASA/JPL-Caltech

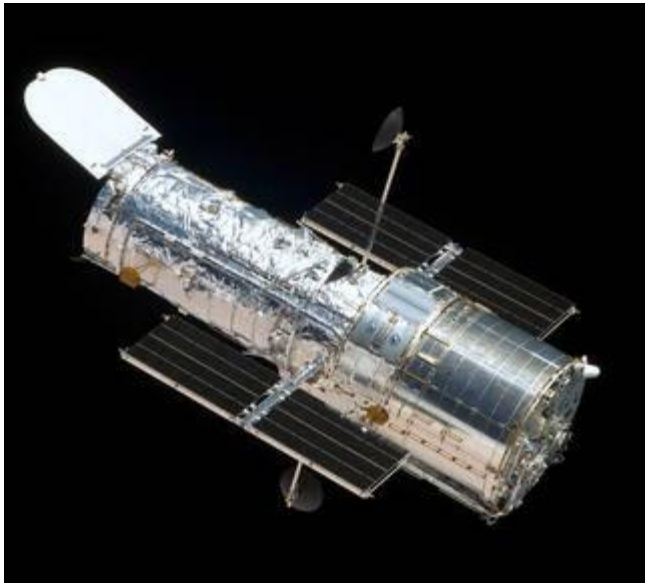
On Aug. 24, 2023, more than three decades after the first confirmation of planets beyond our own solar system, scientists announced the discovery of six new exoplanets, stretching that number to 5,502. From zero exoplanet confirmations to over 5,500 in just a few decades, this new milestone marks another major step in the journey to understand the worlds beyond our solar system.

There is a fast turn-around time for processing the science data, a high on-board storage capacity for TESS ~192 GB, and a high data rate since TESS's orbit is closer to Earth. All of the data is processed on the NASA Advanced Supercomputing (NAS) Division's Pleiades supercomputer.



https://heasarc.gsfc.nasa.gov/docs/tess/data-handling.html

Astrophysics

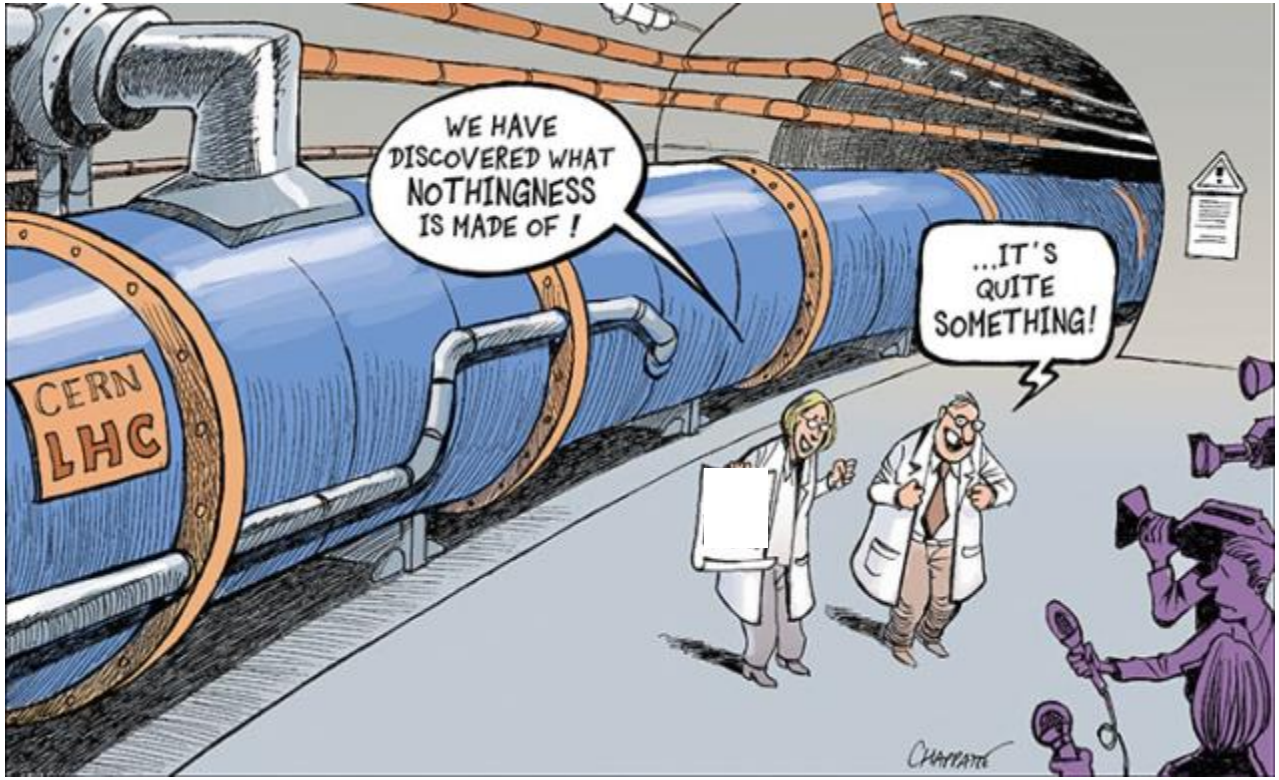


Source: NASA

Hubble transmits about 150 GBs of raw science data every week.

Rubin is on track to revolutionize understanding of dark matter, dark energy, the changing Universe, and other big questions of astronomy and astrophysics.

Discovery at CERN



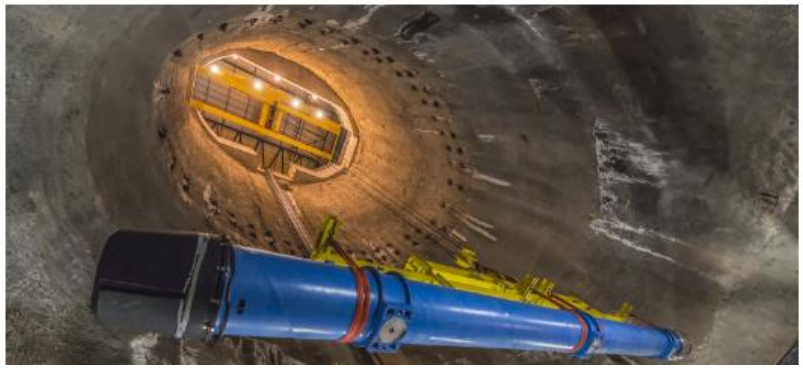
Source: cerncourier.com

LHC@CERN

(Image: Anna Pantelia/CERN)

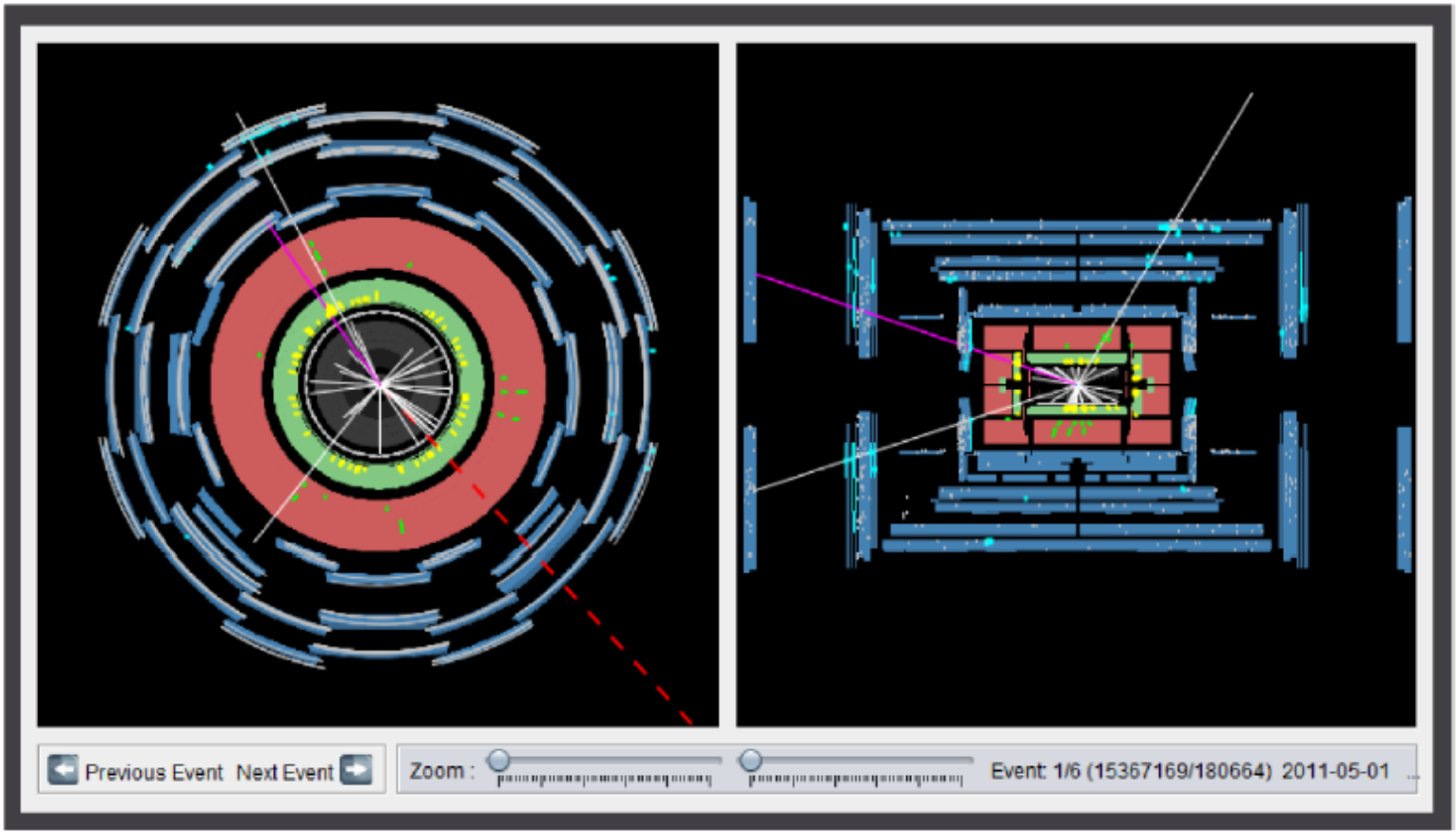
Thousands of magnets of different varieties and sizes are used to direct the beams around the accelerator. These include 1232 dipole magnets, 15 metres in length, which bend the beams, and 392 quadrupole magnets, each 5–7 metres long, which focus the beams. Just prior to collision, another type of magnet is used to "squeeze" the particles

Inside the accelerator, two high-energy particle beams travel at close to the speed of light before they are made to collide. The beams travel in opposite directions in separate beam pipes – two tubes kept at ultrahigh vacuum. They are guided around the accelerator ring by a strong magnetic field maintained by superconducting electromagnets. The electromagnets are built from coils of special electric cable that operates in a superconducting state, efficiently conducting electricity without resistance or loss of energy. This requires chilling the magnets to -271.3°C – a temperature colder than outer space. For this reason, much of the accelerator is connected to a distribution system of liquid helium, which cools the magnets, as well as to other supply services.



600 petabytes of data was stored and analyzed

Event Visualization



HYPATIA tool
ATLAS detects billions of collisions every day
Source: Physics Education, 2014

Scientific Simulations

Simulation and Visualization

Simulation

- Process of exercising a mathematical
  ic
  nd

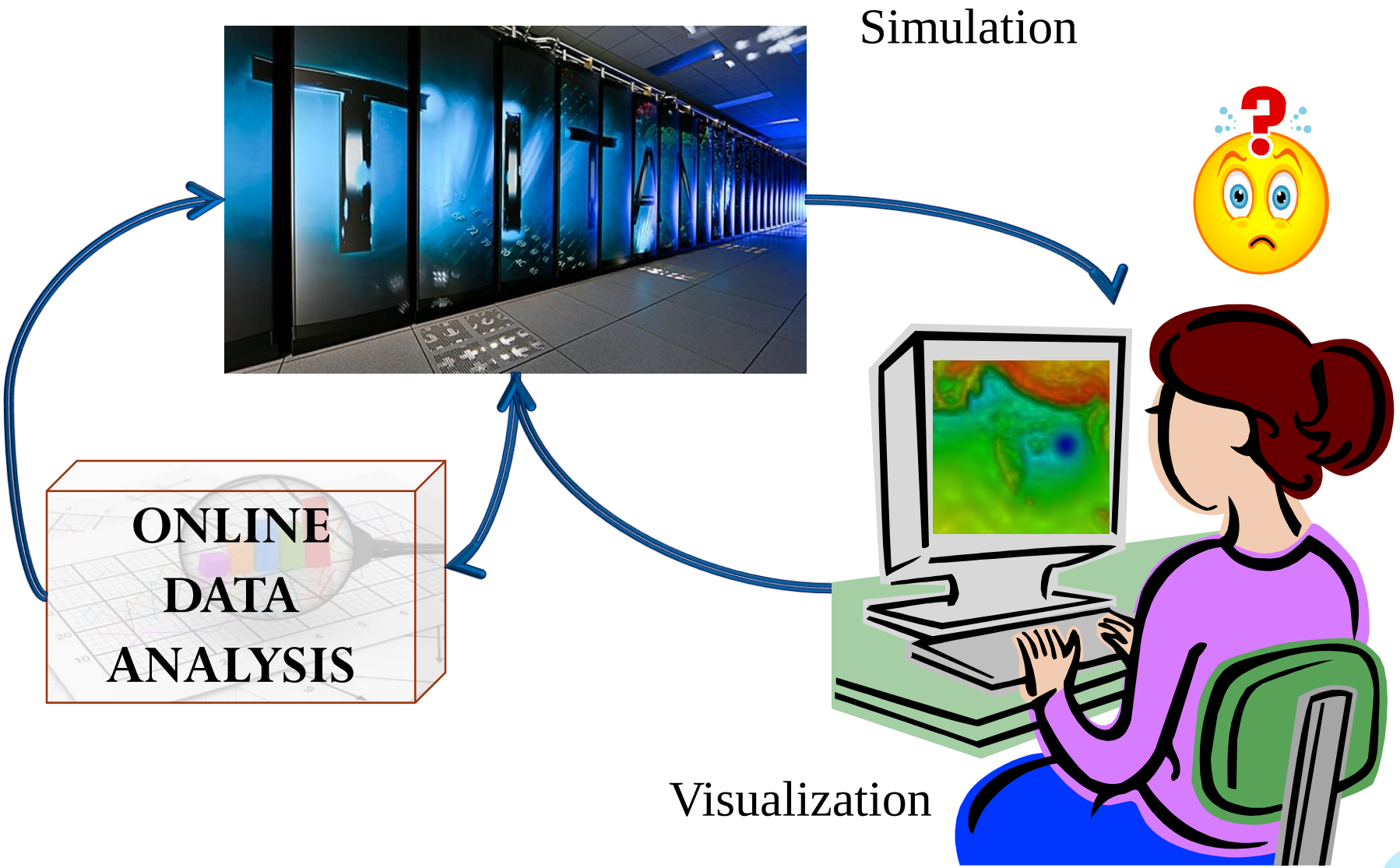
Visual

- Cor
  perc

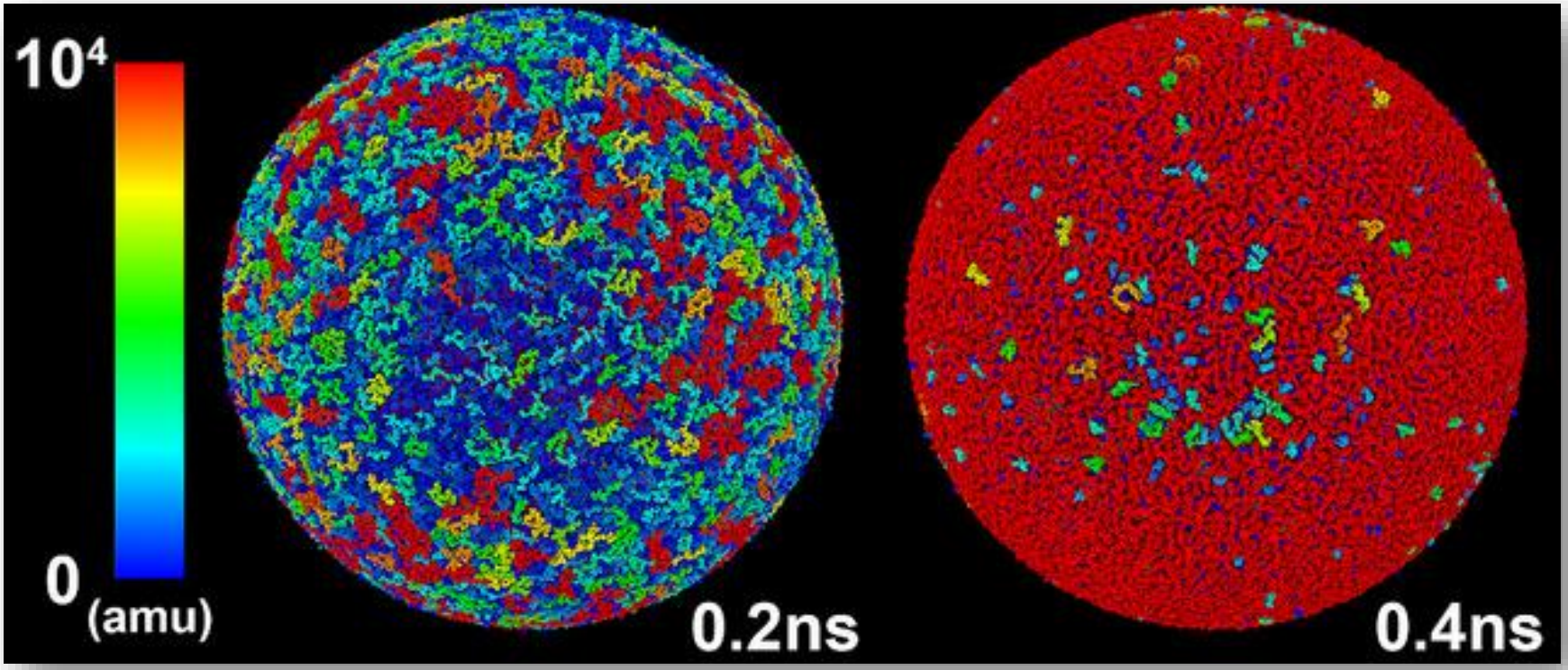
Source: Physical Review E, 2005

Source: Millenn
simulation, MPA

Simulation, Data Analysis, Visualization



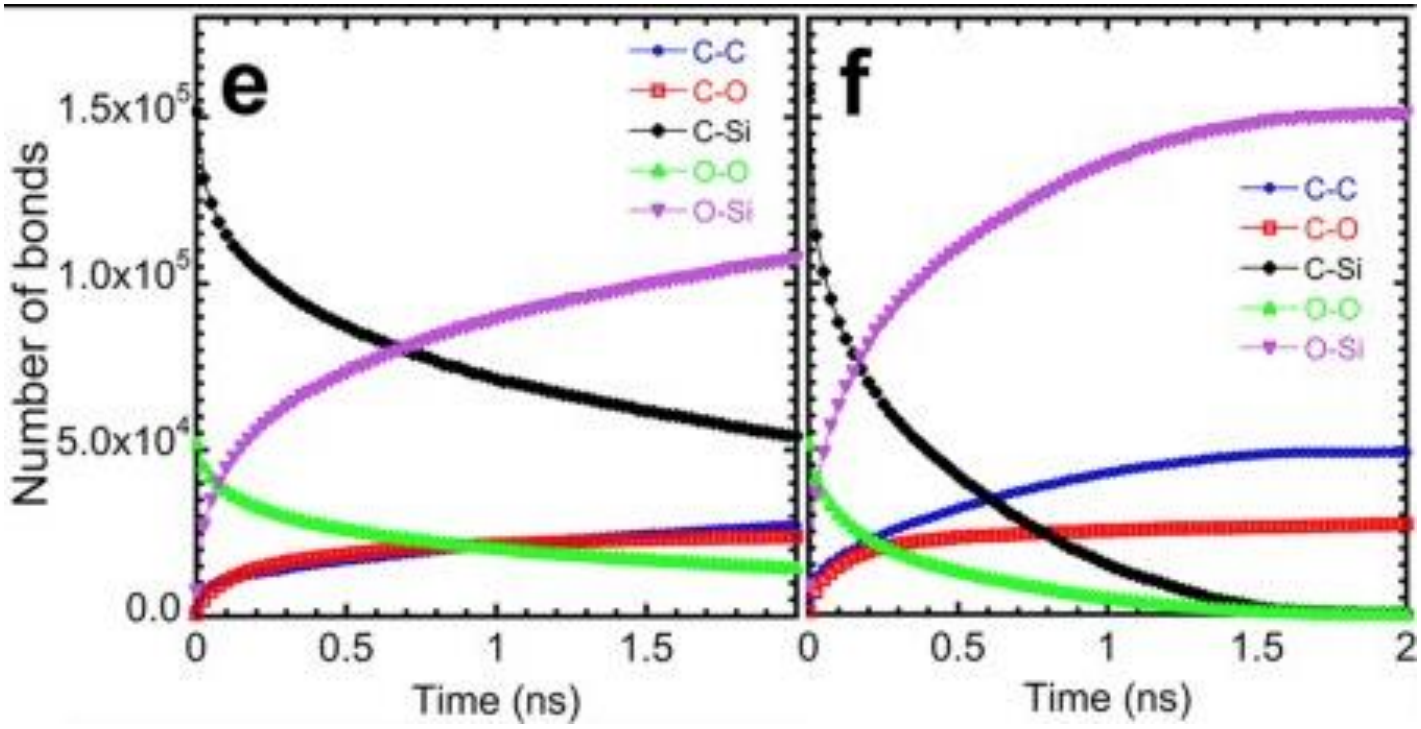
Massively Parallel Material Simulations



Self-healing material simulation

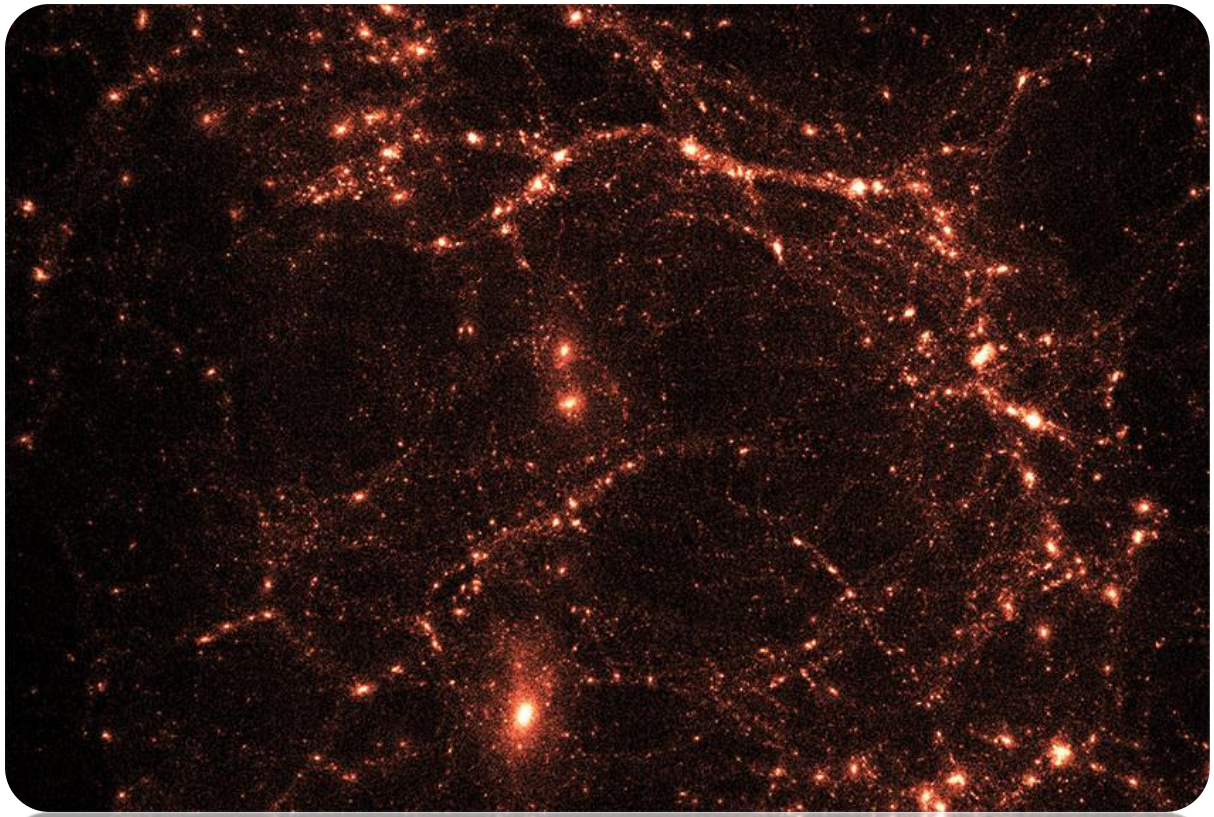
[Nomura et al., "Nanocarbon synthesis by high-temperature oxidation of nanoparticles", Scientific Reports, 2016]

Massively Parallel Analysis



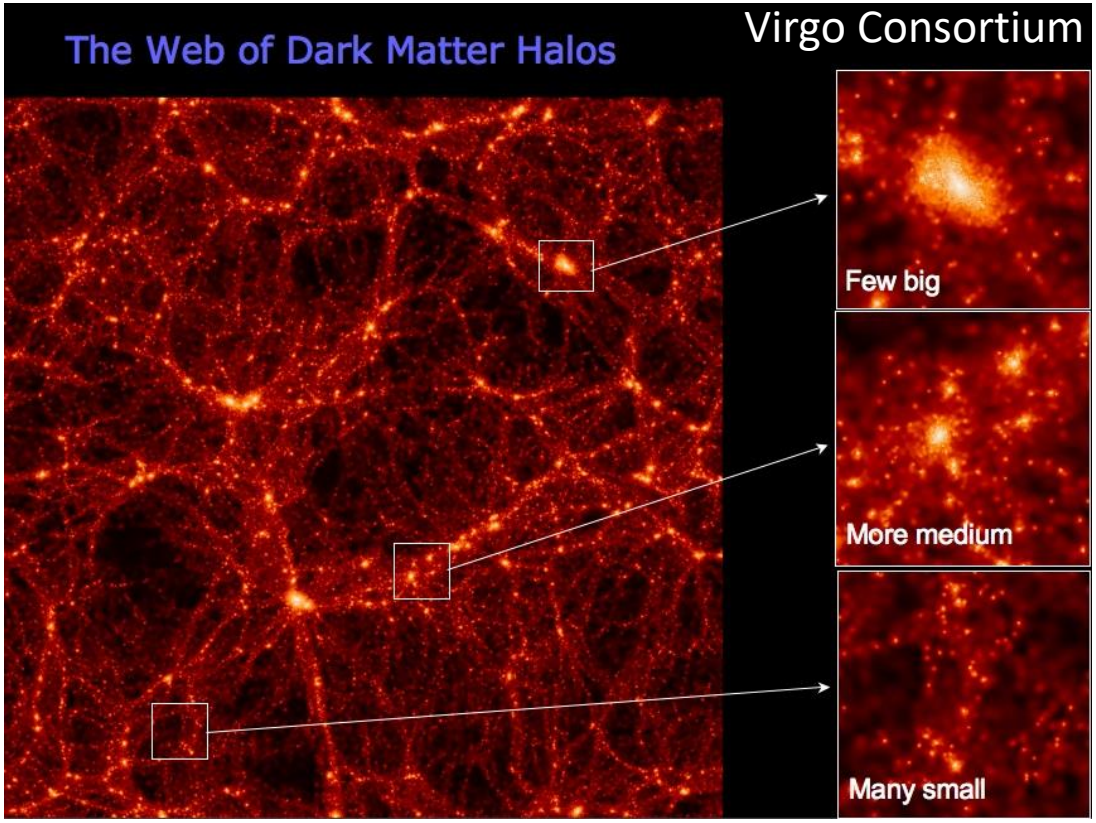
[Nomura et al., "Nanocarbon synthesis by high-temperature oxidation of nanoparticles", Scientific Reports, 2016]

Massively Parallel Codes - Cosmology



Cosmological simulation [Credit: ANL]

Massively Parallel Analysis



Millennium simulation (MPA)

> 10 billion particles traced over 2 billion light years

Weather and Climate Simulations

Area of US = 3.8×10^6 sq. mi.

Geographical area of simulation (US land area)

1 day of simulation on 512 cores takes 12 minutes
~ 25 hours on your laptop!

Analyzing Weather Simulations

Pressure

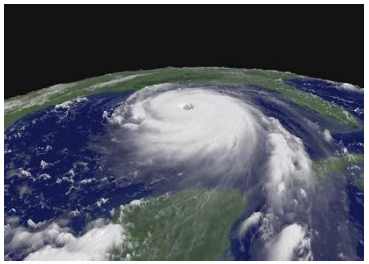
Wind velocity

Precipitation

Temperature

Water vapor

Snow

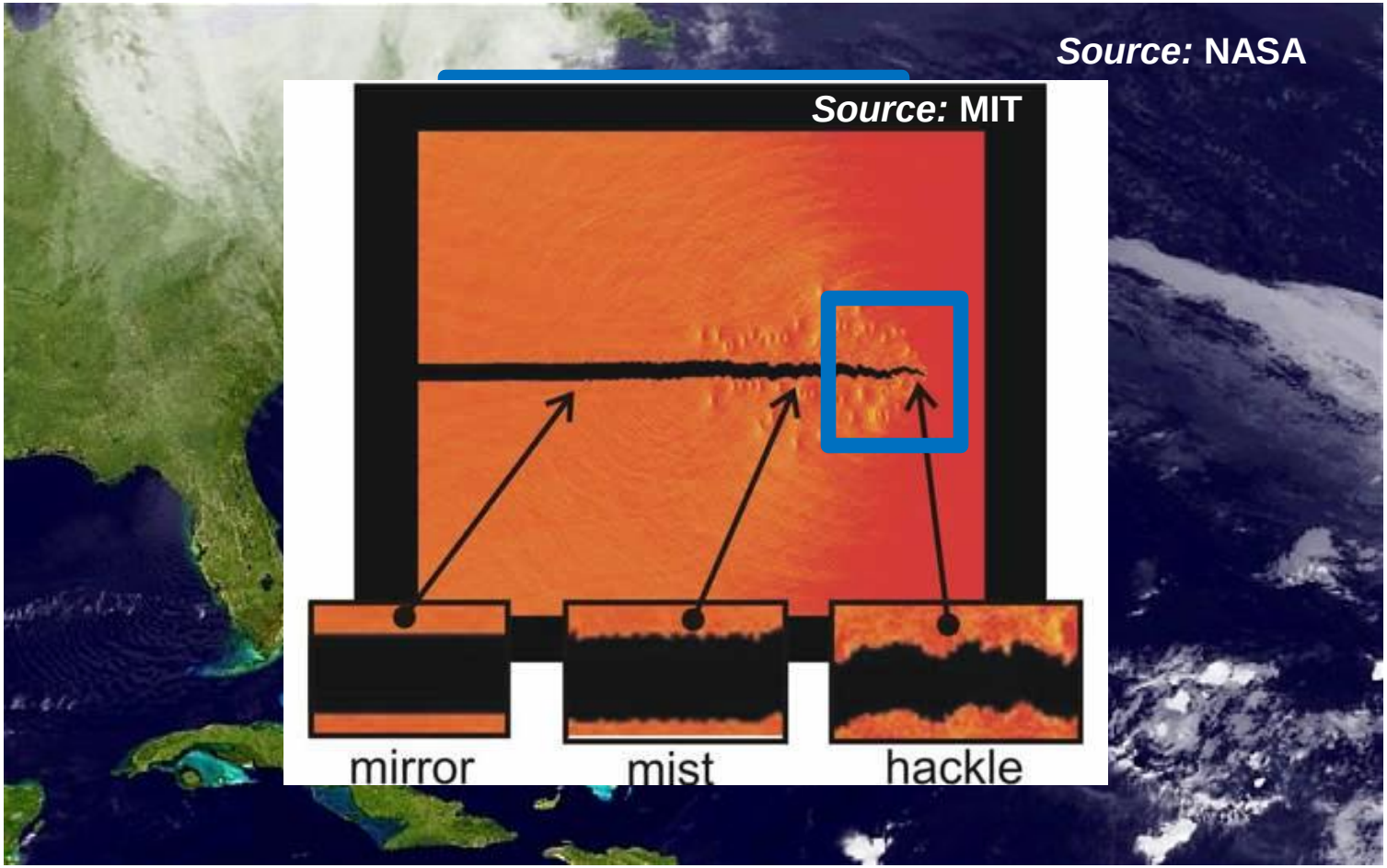


Cloud cover

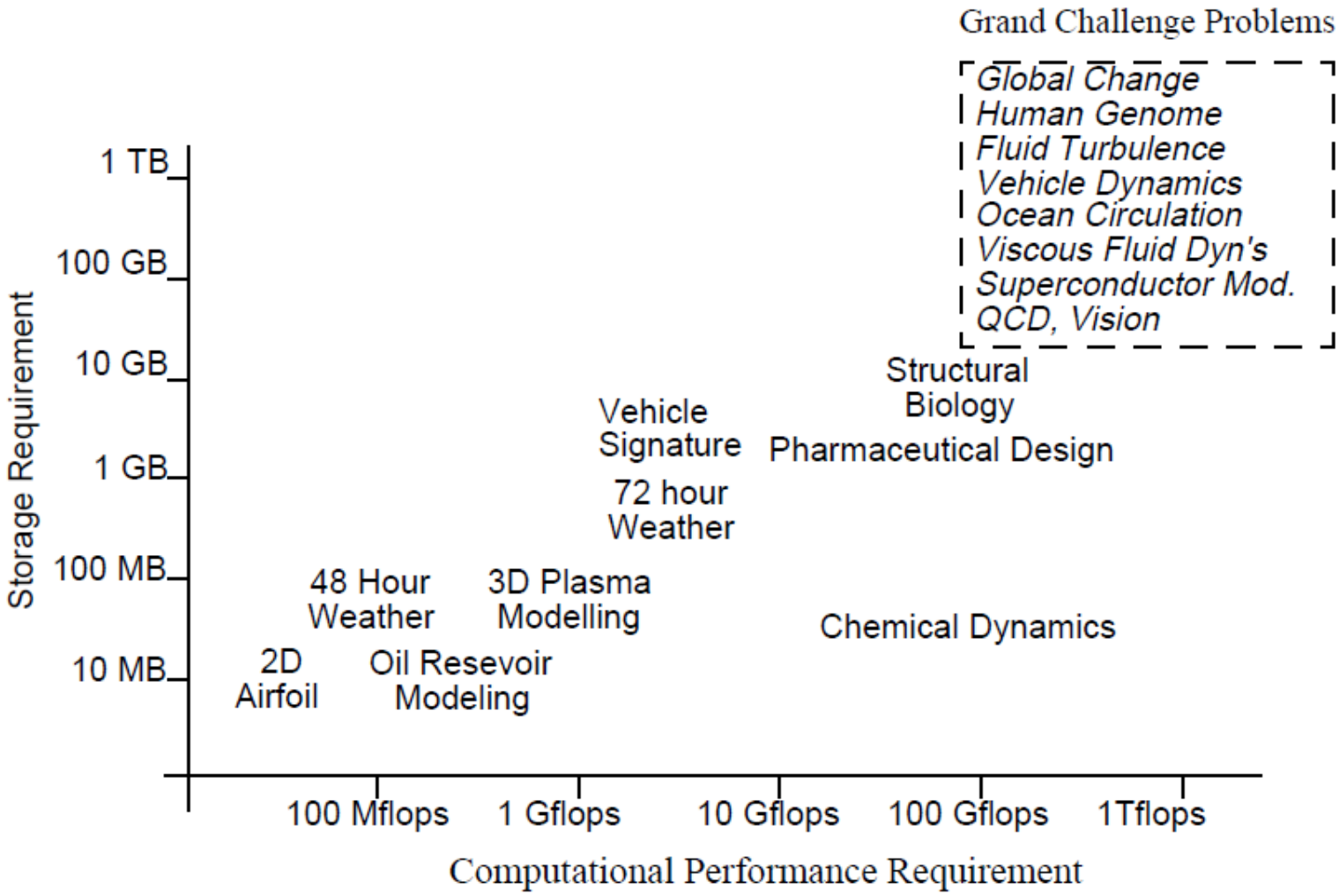
Sea ice

Weather simulation
Source: NOAA

Region of Interest (ROI)



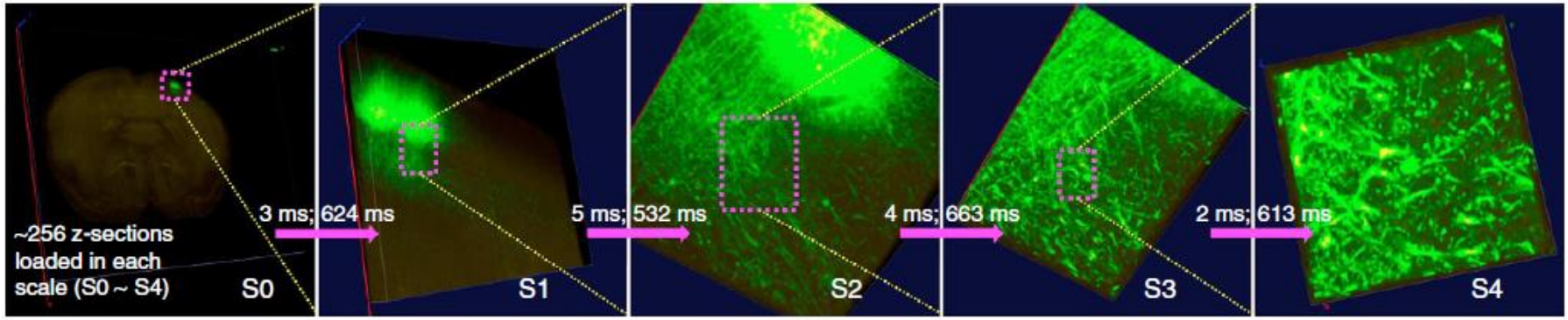
Computational Science



[Source: Culler, Singh and Gupta]

Other Large Data Sources

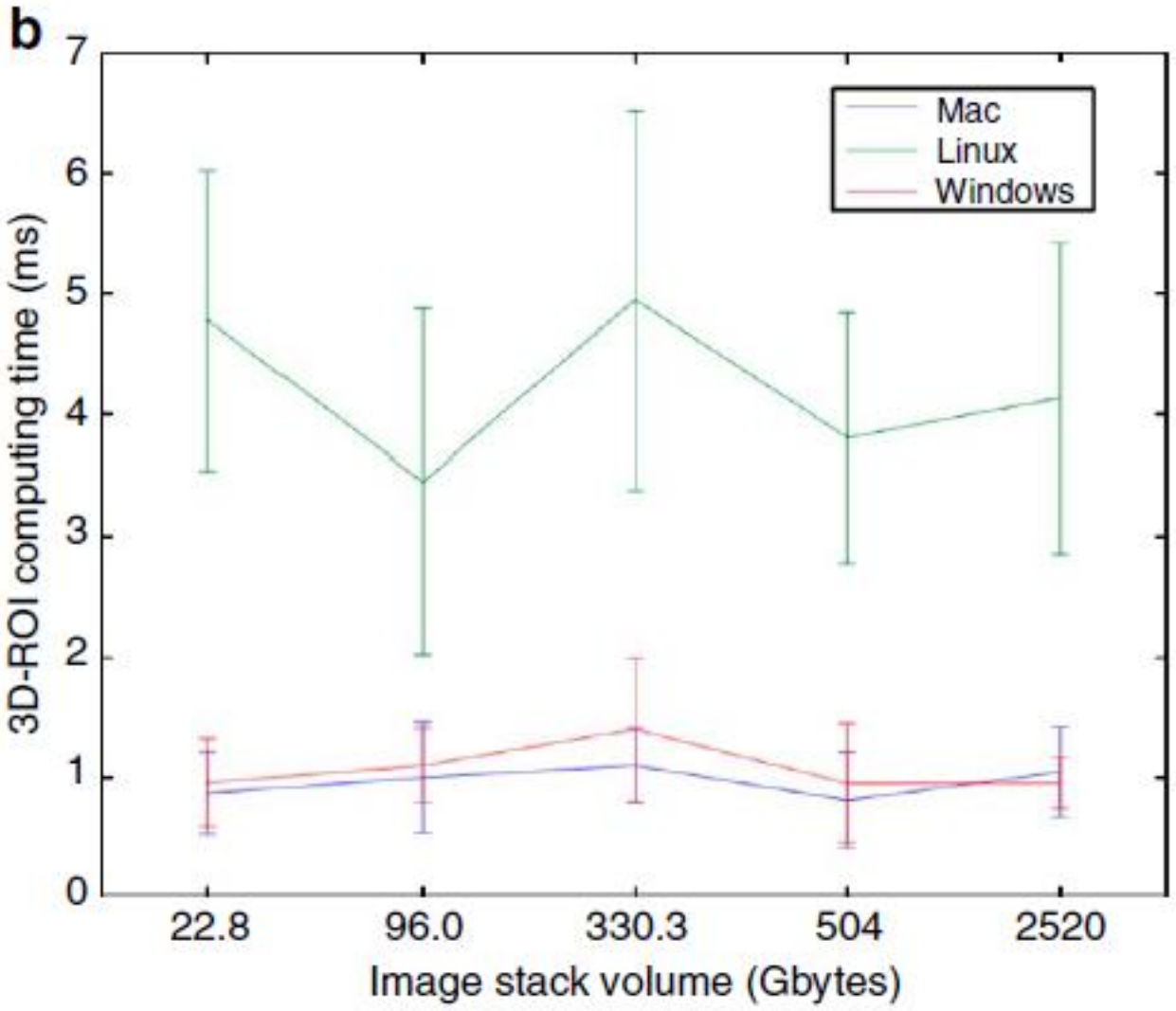
Imaging Technology



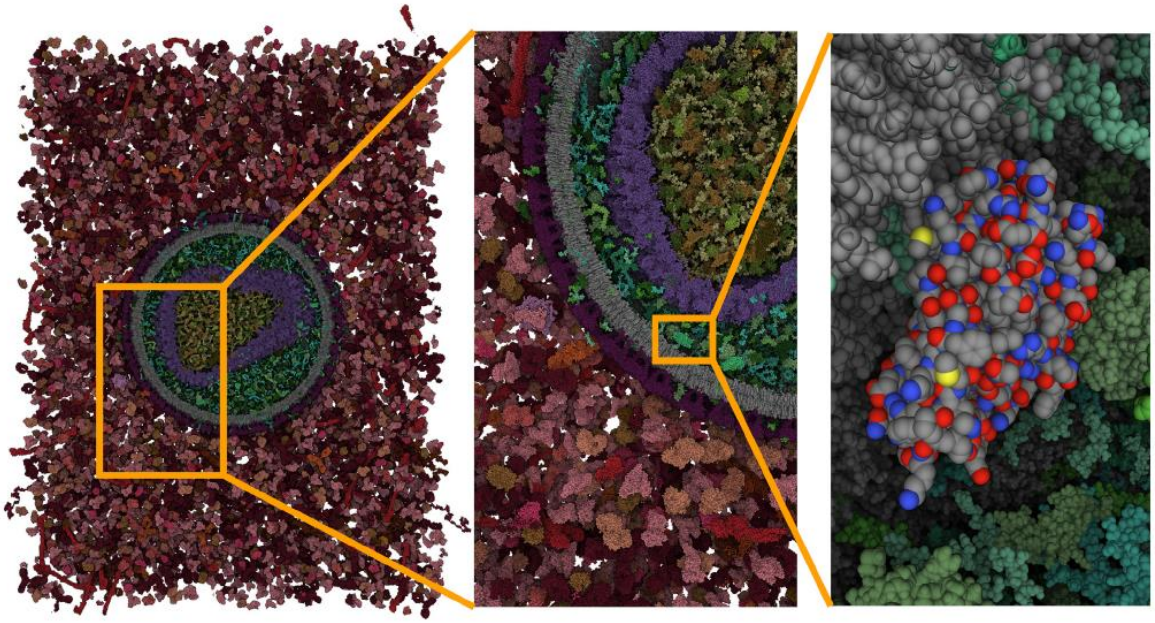
Visualization of a 2.52 Tb whole-mouse brain image stack

Source: Virtual finger boosts three-dimensional imaging and microsurgery as well as terabyte volume image visualization and analysis, Peng et al., Nature Communications, 2014

Analyze Yourself



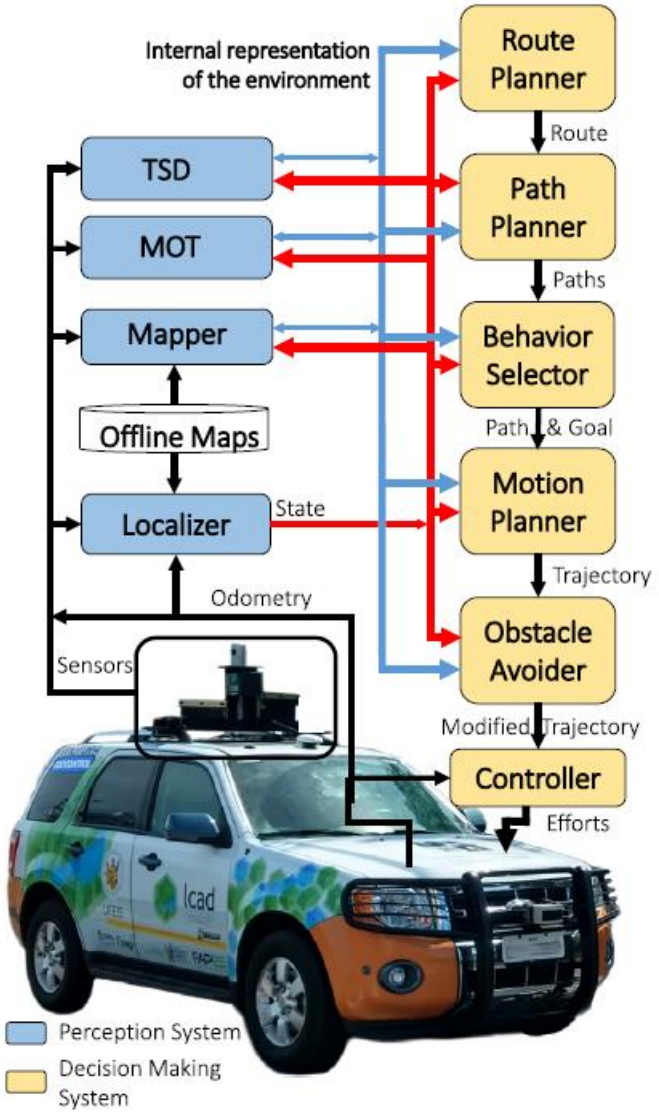
Molecular Dynamics



Visualization of hundreds of millions of atoms

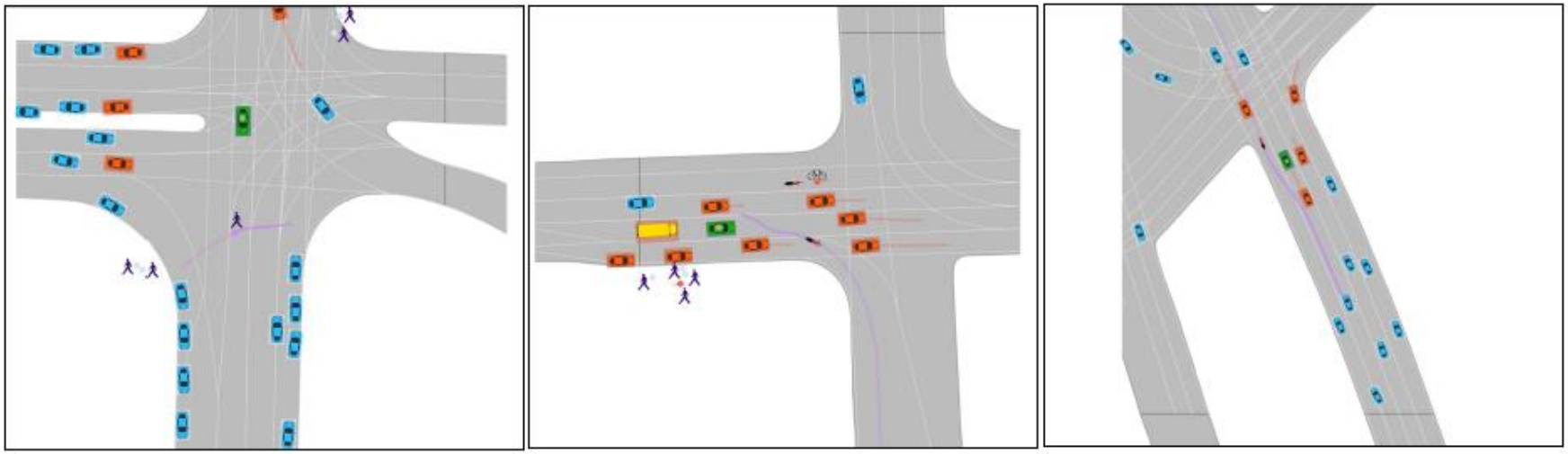
Source: Nucleic Acids Research, 2021

Self-driving Cars



Source: Expert Systems with Applications, 2021

Self-driving Cars



Visualization of interesting scenarios (colours represent different types of vehicles)

Source: NeurIPS 2021

Self-driving Cars

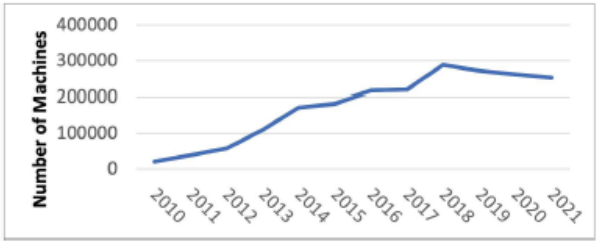
	Argoverse [6]	Inter [52]	Lyft [22]	Waymo [12]	NuScenes [4]	Yandex [34]	Ours
# Scenarios	324k	-	170k	104k	41k	600k	250k
# Unique Tracks	11.7M	40k	53.4M	7.6M	-	17.4M	13.9M
Average Track Length	2.48 s	19.8 s	1.8 s	7.04 s	-	-	5.16 s
Total Time	320 h	16.5 h	1118 h	574 h	5.5 h	1667 h	763 h
Scenario Duration	5 s	-	25 s	9.1 s	8 s	10 s	11 s
Test Forecast Horizon	3 s	3 s	5 s	8 s	6 s	5 s	6 s
Sampling Rate	10 Hz	10 Hz	10 Hz	10 Hz	2 Hz	5 Hz	10 Hz
# Cities	2	6	1	6	2	6	6
Unique Roadways	290 km	2 km	10 km	1750 km	-	-	2220 km
Avg. # tracks per scenario	50	-	79	-	75	29	73
# Evaluated object categories	1	1	3	3	1	2	5
Multi-agent evaluation	×	✓	✓	✓	×	✓	✓
Mined for Interestingness	✓	×	-	✓	×	×	✓
Vector Map	✓	×	×	✓	✓	×	✓
Download Size	4.8 GB	-	22 GB	1.4 TB	48 GB	120 GB	58 GB
# Public Leaderboard Entries†	194	-	935	23	18	3	-

Different motion forecasting datasets

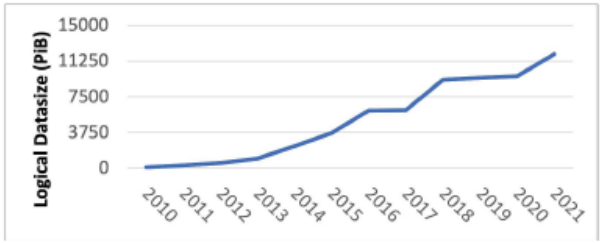
Source: NeurIPS 2021

Cosmos

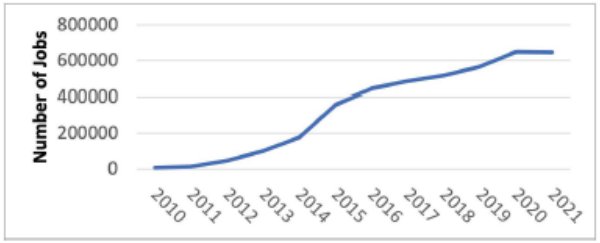
- Microsoft's data store, used by their product teams
- Compute and storage ecosystem
- Multiple exabytes of data



(a) Number of servers in Cosmos clusters.



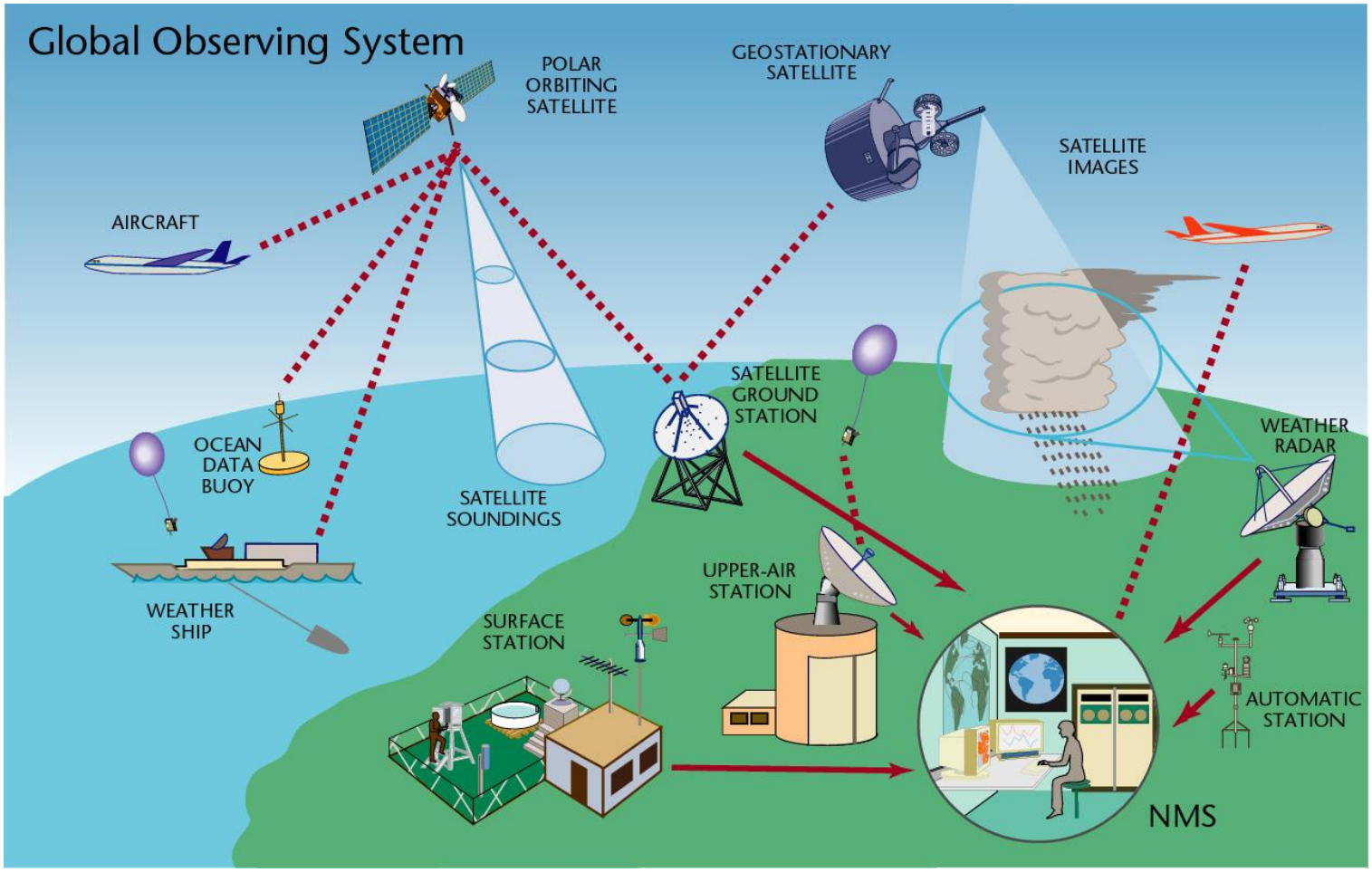
(b) Data size before compression/replication.



(c) Number of batch SCOPE jobs run per day.

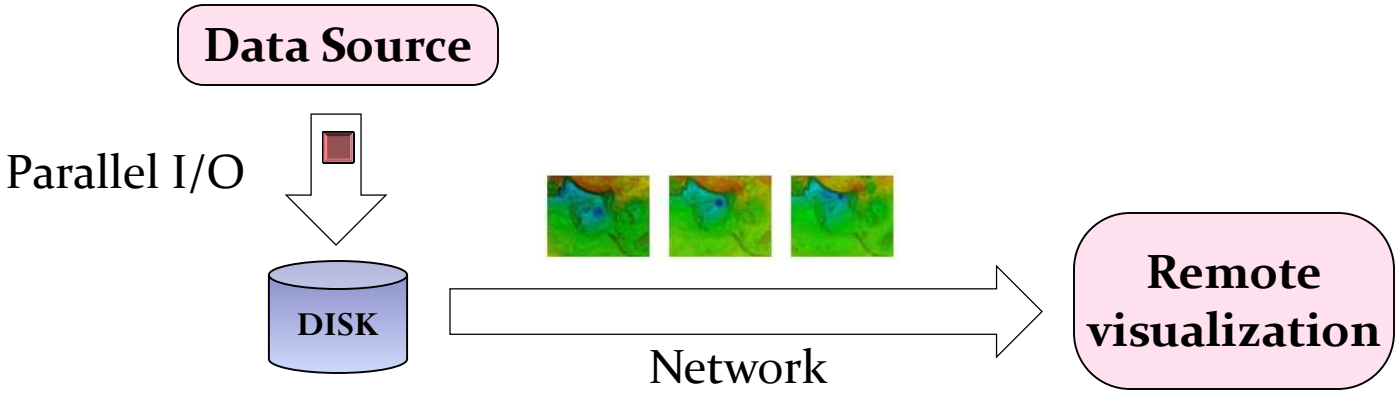
Source: VLDB 2021

Input Data



[Credit: World Meteorological Organization]

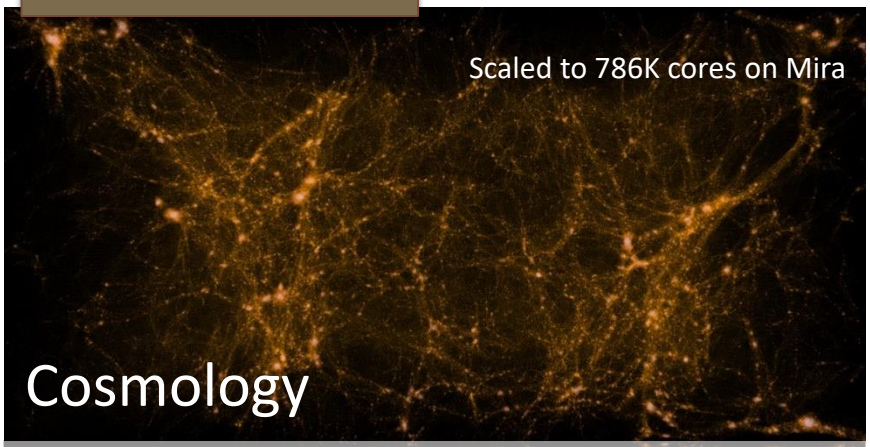
Analysis and Visualization Pipeline



Large Data and Storage

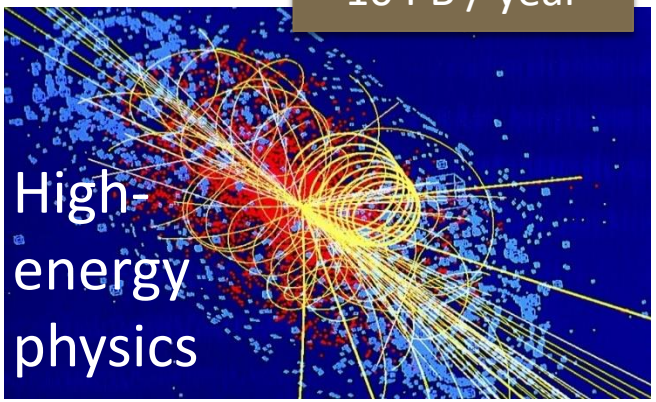
Output Data of Simulations

2 PB / simulation



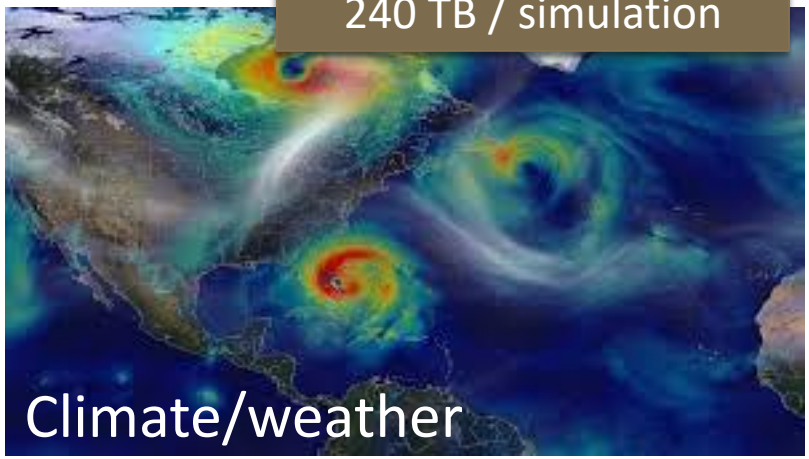
Q Continuum simulation
Source: Salman Habib et al.

10 PB / year



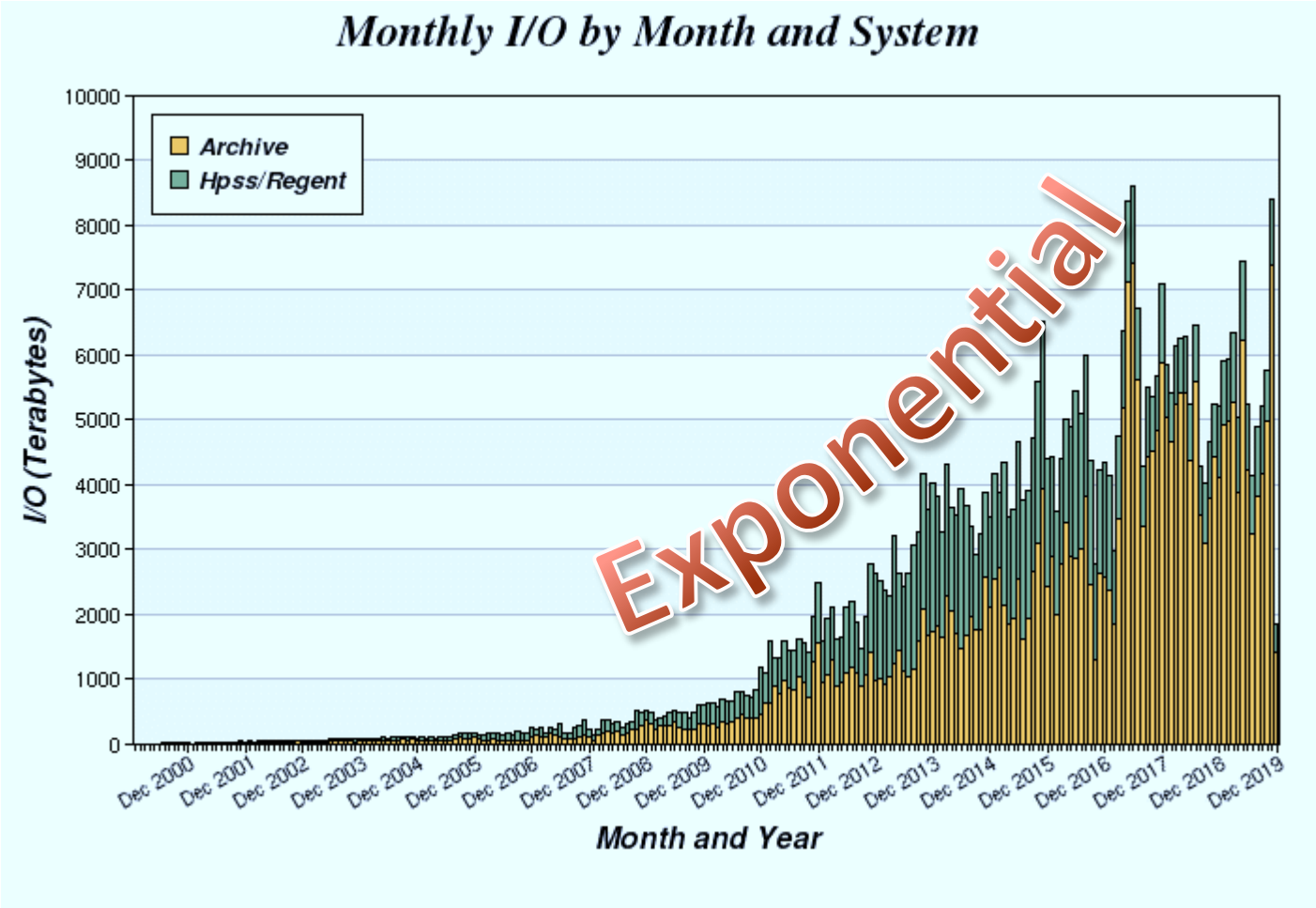
Higgs boson simulation
Source: CERN

240 TB / simulation



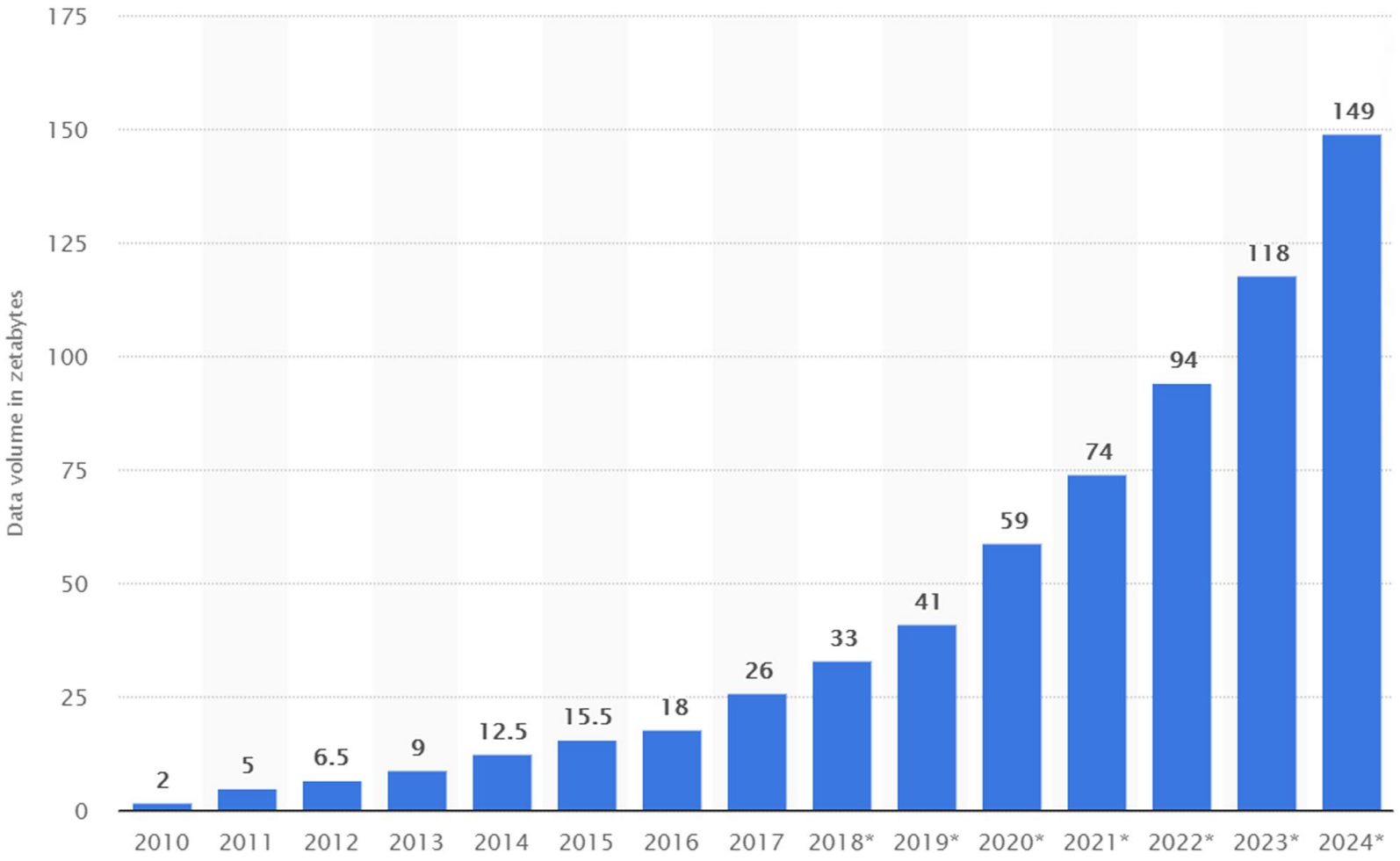
Hurricane simulation
Source: NASA

I/O trends



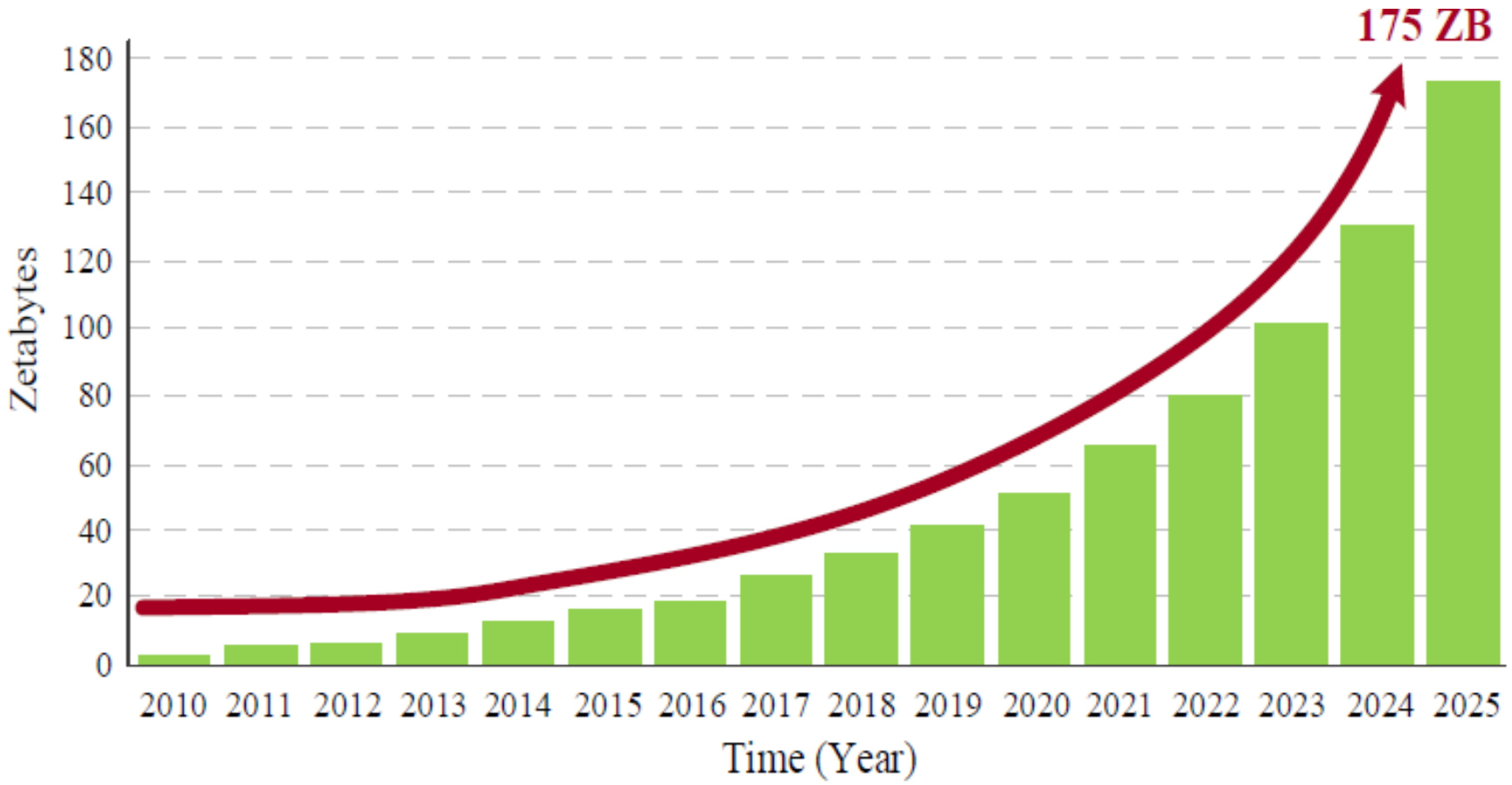
NERSC I/O trends [Credit: www.nersc.gov]

Data trends



Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2024 (estimated). Source: Statista

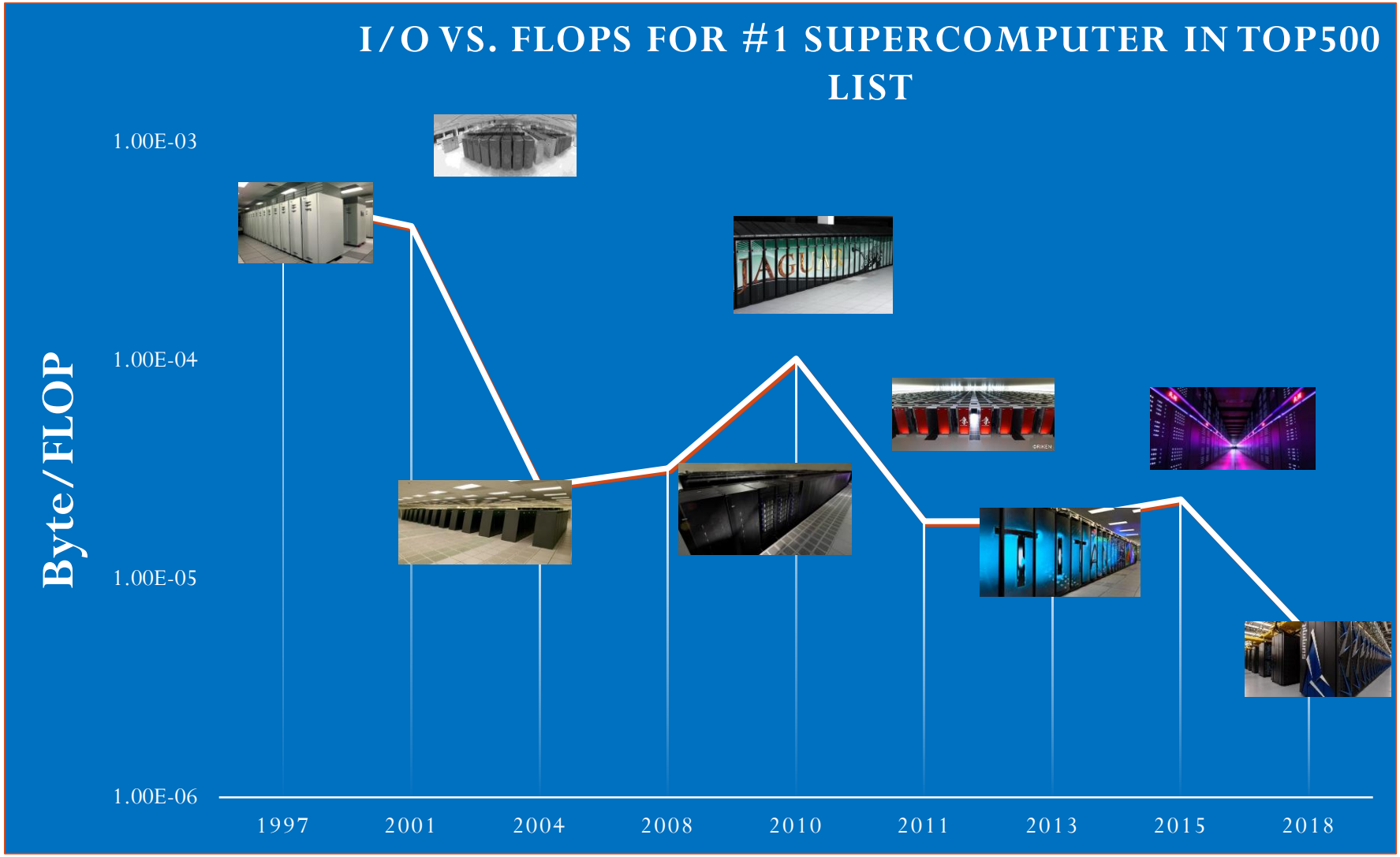
Data Trends



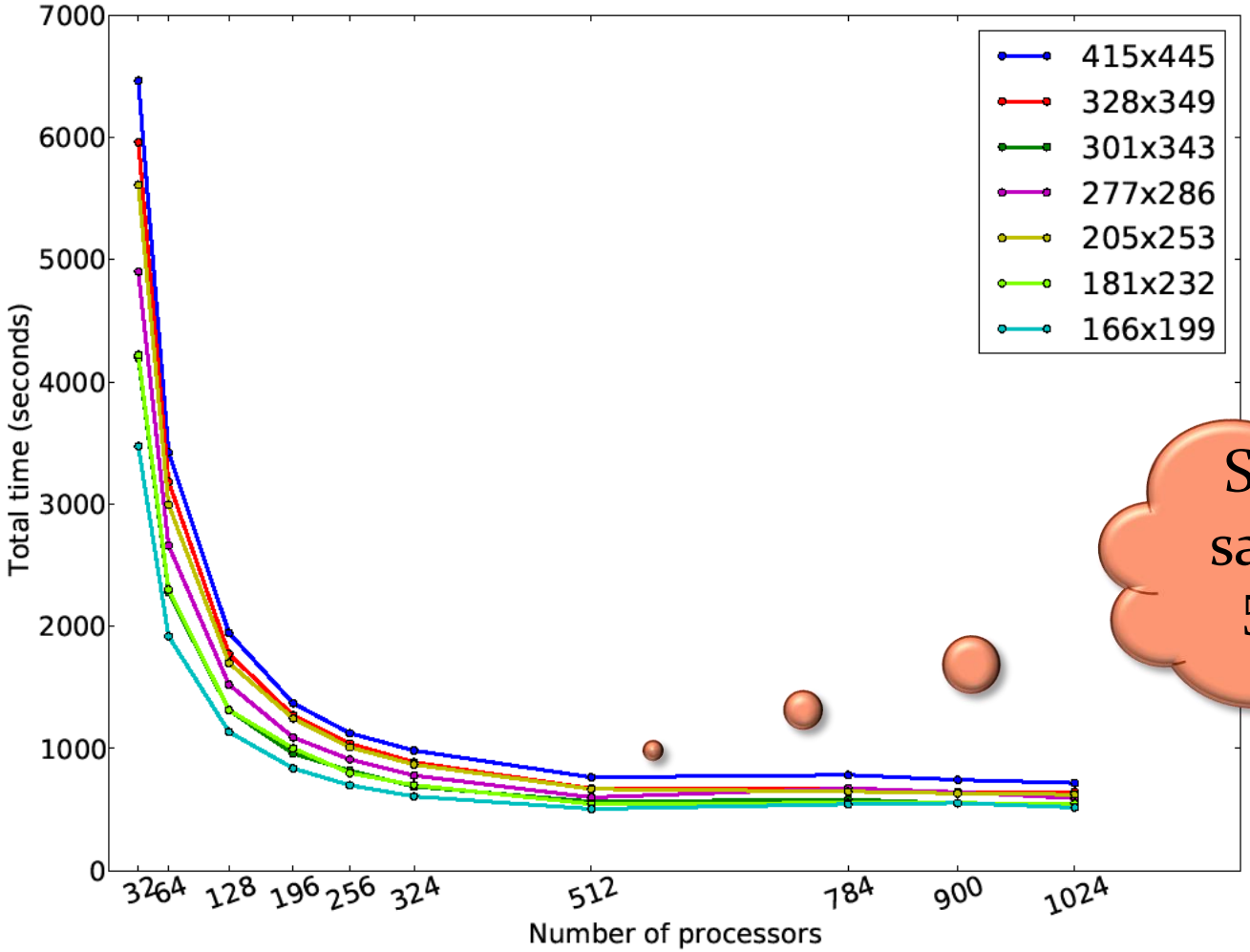
Annual Growth Rate of Global Data

Source: IDC

Compute vs. I/O trends



Scalability



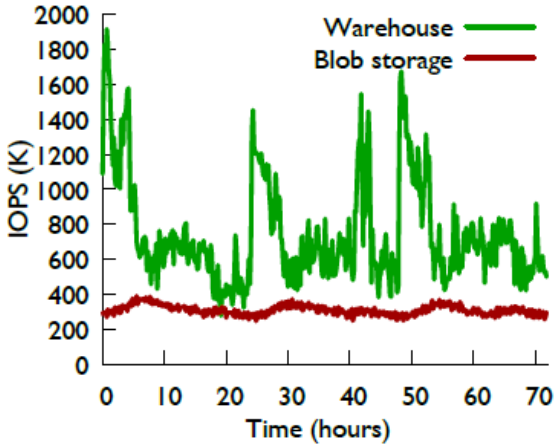
Scalability saturates at 512 cores

Execution time of a weather simulation over a 10^7 sq. km. domain on up to 1024 processors

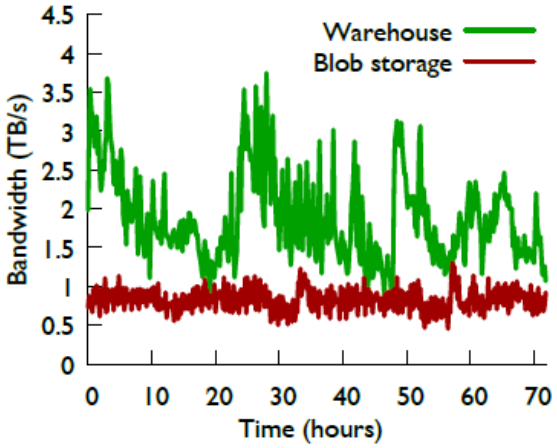
How to handle large data?

Tectonic

- Facebook's exabyte-scale distributed filesystem (FAST'21)

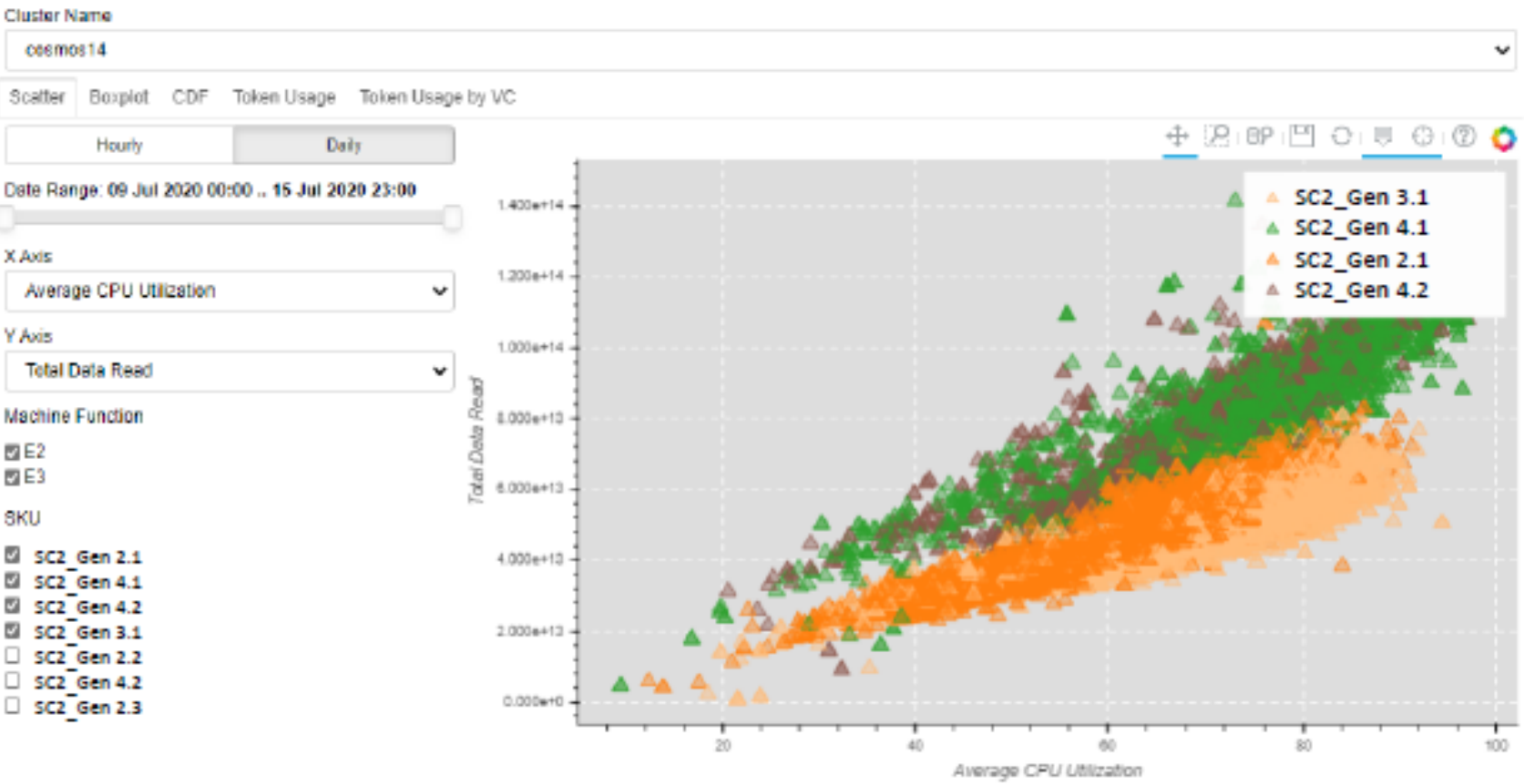


(a) Aggregate cluster IOPS



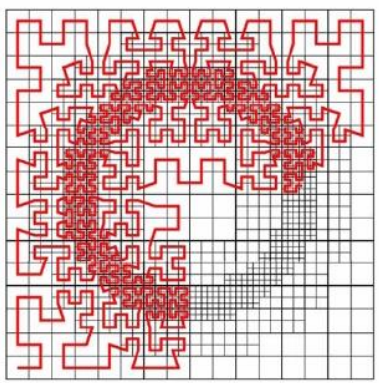
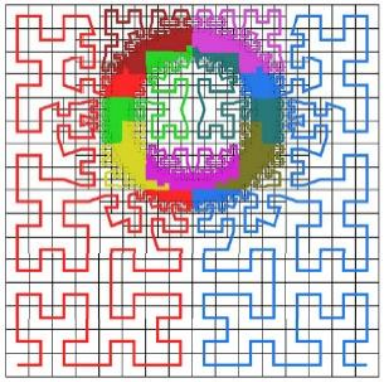
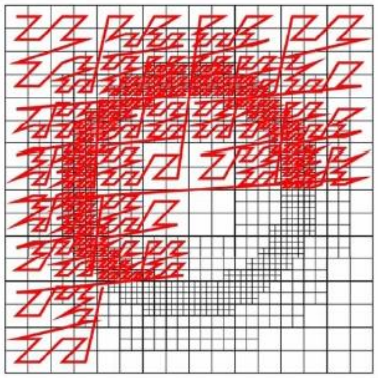
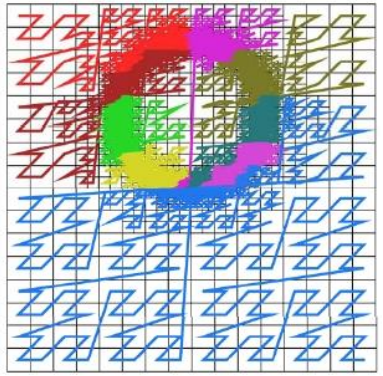
(b) Aggregate cluster bandwidth

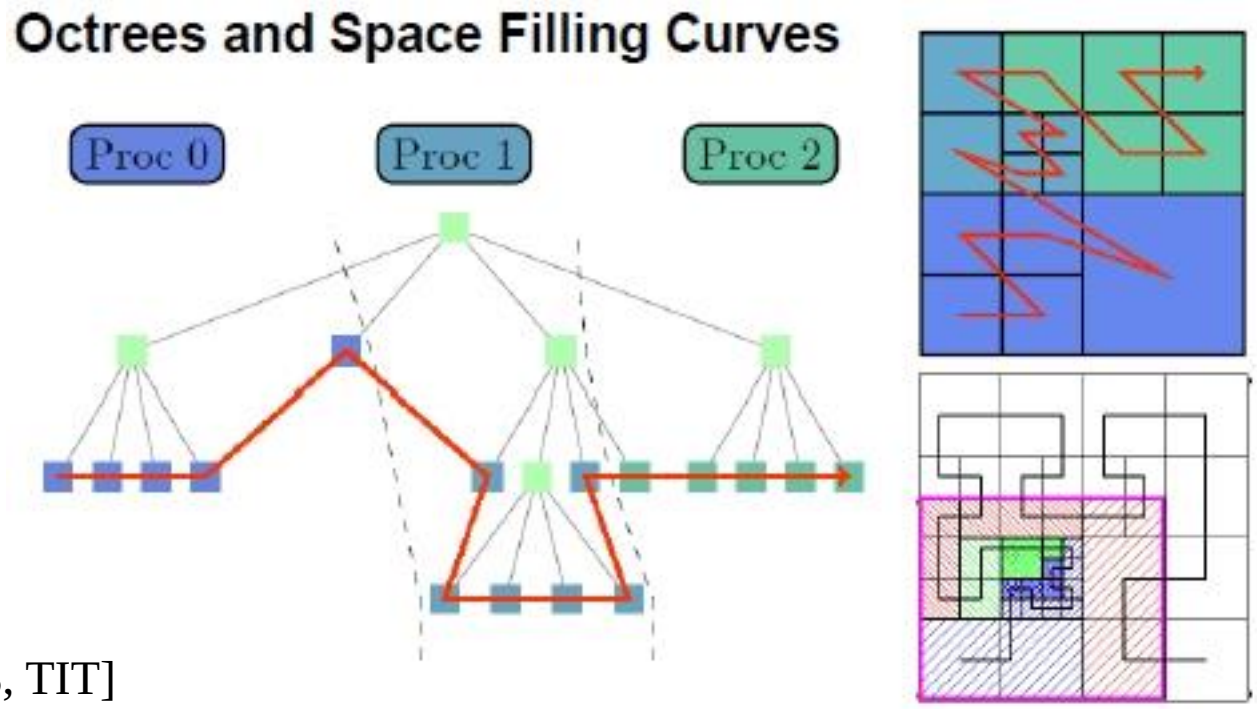
KEA – Tuning Resources (Microsoft)



300K machines, 0.6M daily jobs [Source: SIGMOD 2021]

Data Representation



[Source: SC13, TIT]

Large Data Challenges – Brief Summary

[Source: IEEE GPECOM 2023]