Data, Computation, and the Fate of the Universe

Saul Perlmutter

University of California, Berkeley
Lawrence Berkeley National Laboratory

NERSC Lunchtime Nobel Keynote Lecture
June 2014
Atomic Matter 4%

Dark Matter 24%

Dark Energy 72%

or a modification of Einstein’s Theory of General Relativity?
Supernova (SN):

Large quantities of data need to be analyzed in near-real-time.

Current: 1.5 TB/night processed
LSST era: ~50 TB/night processed

Machine Learning, Boosted Decision Trees to find transient SNe, which are needles in haystack of 1 M candidates/night.

SN observations compared to supercomputer-based simulations.

Statistical analyses of cosmological parameters need Markov Chain Monte Carlo (MCMC).
Baryon Acoustic Oscillations (BAO):

Large quantities of data need to be analyzed.

Imaging survey in 2005: 20 TB
in 2025 60 PB

Statistical analyses need MCMC for cross-correlation of the millions of galaxies -- collapsing the problem to just 2-point statistics.

All data analysis dependent on comparisons to supercomputer-based N-body simulations of the evolution of matter in the universe.

Current state of art: $2048^3 - 4096^3$ “particles.”
Need an order of magnitude more.
Cosmic Microwave Background (CMB):

Exponentially growing data chasing fainter echoes:

- BOOMERanG: $10^9$ samples in 2000
- Planck: $10^{12}$ samples in 2013 (0.5 PB)
- CMBpol: $10^{15}$ samples in 2025

Uncertainty quantification through Monte Carlos

- Simulate $10^4$ realizations of the entire mission
- Control both systematics and statistics

Mission-class science relies on HPC evolution.
DOE “Big Data” Challenges

Volume, velocity, variety, and veracity

**Biology**
- **Volume:** Petabytes now; computation-limited
- **Variety:** multi-modal analysis on bioimages

**Cosmology & Astronomy:**
- **Volume:** 1000x increase every 15 years
- **Variety:** combine data sources for accuracy

**High Energy Physics**
- **Volume:** 3-5x in 5 years
- **Velocity:** real-time filtering adapts to intended observation

**Materials:**
- **Variety:** multiple models and experimental data
- **Veracity:** quality and resolution of simulations

**Light Sources**
- **Velocity:** CCDs outpacing Moore’s Law
- **Veracity:** noisy data for 3D reconstruction

**Climate**
- **Volume:** Hundreds of exabytes by 2020
- **Veracity:** Reanalysis of 100-year-old sparse data
We have computing power, we have applied math techniques, we have database approaches, so...

What’s missing?
Data Science for academic scientists: What’s still needed?

*Make it “progressive”:*

Today, for each project, a new set of students/post-docs writes code that often re-invents previous solutions, and then they graduate, leaving little that can be built on since the code was written to reach a conference/paper/thesis as rapidly as possible.

We must make it **easy to**

1. **find the best code/algorithm/approach/tutorial** for a given purpose, within your own group, your own discipline, another discipline, industry, ...

2. **contribute and maintain code** that could be useful for a larger community
DS for academic scientists: What’s still needed?

Easy to see:

3. **Long term career paths** for crucial members of our science teams who become engaged in the data science side of the work.

4. **Data science training** for undergraduates, graduate students, and post-docs to quickly come up to speed in research.
DS for academic scientists: What’s still needed?

*Less obvious:*

5. Our programming languages and **programming environments should not distract from the science.**

6. **Bridge the current gaps** between the interests/needs of domain scientists and the interests/needs of data science methodologists.

7. Use ethnography to rigorously **study what slows the scientists down in their use of data.**
Potential gains:

• **Remove/reduce barriers for those who are less data-science savvy** than those in this room.

• **Data science as a bridge between disciplines** and a magnet for *in-person human interaction*. 
First, a 6-year gift to Berkeley for data science in cosmology

Then, a 5-year, $37.8 million cross-institutional collaboration
Berkeley Institute for Data Science (BIDS)

Relevance across the campus suggests need for central location that will serve as home for data science efforts

Enhancing strengths of
• Simons Institute for the Theory of Computing
• AMP Lab
• SDAV Institute
• CITRIS
• etc.
Scientific Data Initiative at Berkeley Lab

Leverage Berkeley Lab talent in math, computer science, interdisciplinary team science, networking, software engineering and our new infrastructure to enable new modes of inquiry and discovery from scientific data sets.
An Extreme Data Scientific Facility (XDSF) would bring together diverse data sets for learning and discovery.
XDSF will bring scientists together with data researchers and software engineers.
XDSF will bring scientists together with data researchers and software engineers.
This is an exciting time for the field of cosmology: We are now ready to collect, simulate, and analyze the next level of precision data.

Cosmology is a young field: we haven’t yet taken a big step in precision without surprise(s).
This is an exciting time for Data Science for science: We are now ready to explore new approaches.

There’s more to high performance scientific computing than we have yet accomplished. We can find ways to make scientists’ use of data fluent and fluid.
Exploring Scientific-Computational Collaboration: NERSC and the Supernova Cosmology Project

Computational Innovations to Measure the Parameters of the Universe

Using Cosmologically Distant Supernovae

S. Perlmutter  G. Goldhaber

LDRD FY96 "Cosmology"

Using astrophysics techniques developed by the Cosmology Group, we will be measuring in the next few years the parameters that shape our current understanding of the Universe. This requires unusual computational environments beyond these current capabilities and will necessitate new computational access to large data sets, tools, and collaborative multi-site experimental environments. The purpose of NERSC, and we hope over the next few years, is to support these research fronts. As a first step towards these goals we begin with one part of this collaboration: a comparison of computer calculations to bear on the near-real-time data obtained at the Keck Telescope.
“...For a while, the whole NASA USA net this morning was screwed up, but it's back. Shortly after it came back, the transfers to Chile were back up to their normal 4kbyte/sec speed....”
Happy 40th Anniversary NERSC!