Berkeley Institute for Data Science (BIDS)

Saul Perlmutter
Department of Physics
University of California, Berkeley

Computational Astrophysics 2014-2020
LBNL
March 2014

Berkeley
UNIVERSITY OF CALIFORNIA

Data Science throughout campus

WIRED

Spark: Open Source Superstar Rewrites Future of Big Data

BY CADE METZ 06.19.13 6:30 AM

AMP Lab
Ion Stoica, CS
Michael Franklin, CS
Matei Zaharia, CS

KBase
PREDICTIVE BIOLOGY
DOE Systems Biology Knowledgebase

Adam Arkin,
Bioengineering

Moore-Sloan Data Science Initiative

Reconstructing th
in your mind

Bin Yu, Statistics
Jack Gallant, Neuroscience

Richard Allen
Earth& Plan.
Science

Charles Marshall
Rosie Gillespie
Integrative
Biology

The New York Times

Incomes Flat in Recovery, but Not for the 1%

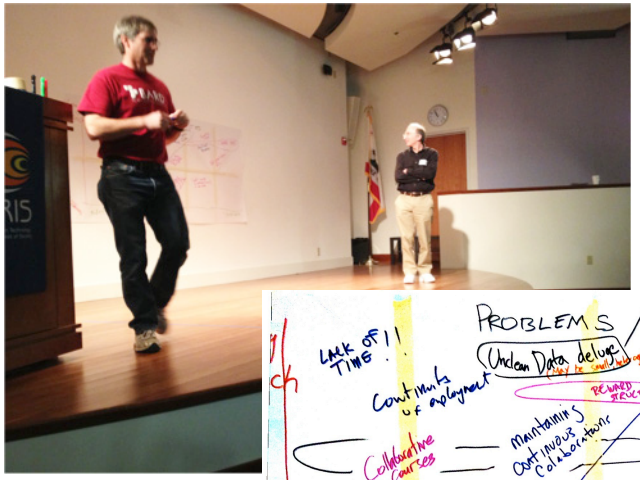
Feb 15, 2013

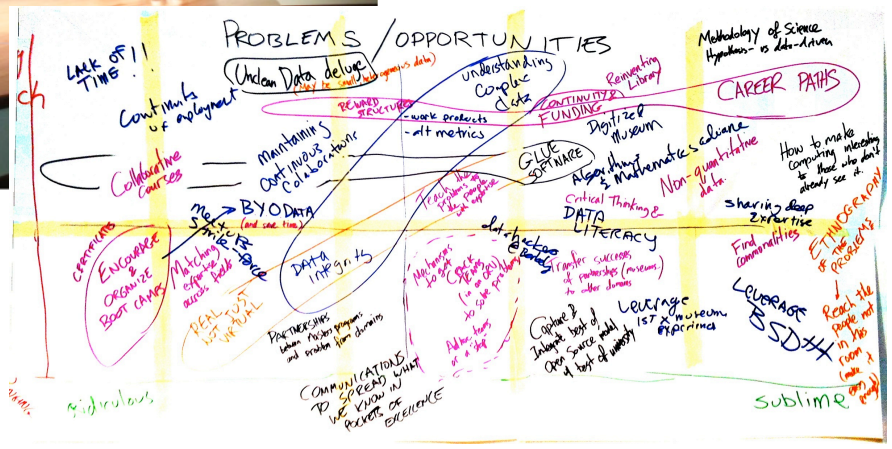
Emmanuel Saez, Economics

Berkeley
UNIVERSITY OF CALIFORNIA

Great interest from across the campus

Data Science Workshop held in February 2013 was attended by 80 researchers on three days notice; with follow-up events in May and June (to date 280+ signed up for mailing list)





Berkeley
UNIVERSITY OF CALIFORNIA

A 5-year, $37.8 million cross-institutional collaboration

NYU — Berkeley UNIVERSITY OF CALIFORNIA — UNIVERSITY of WASHINGTON

GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION

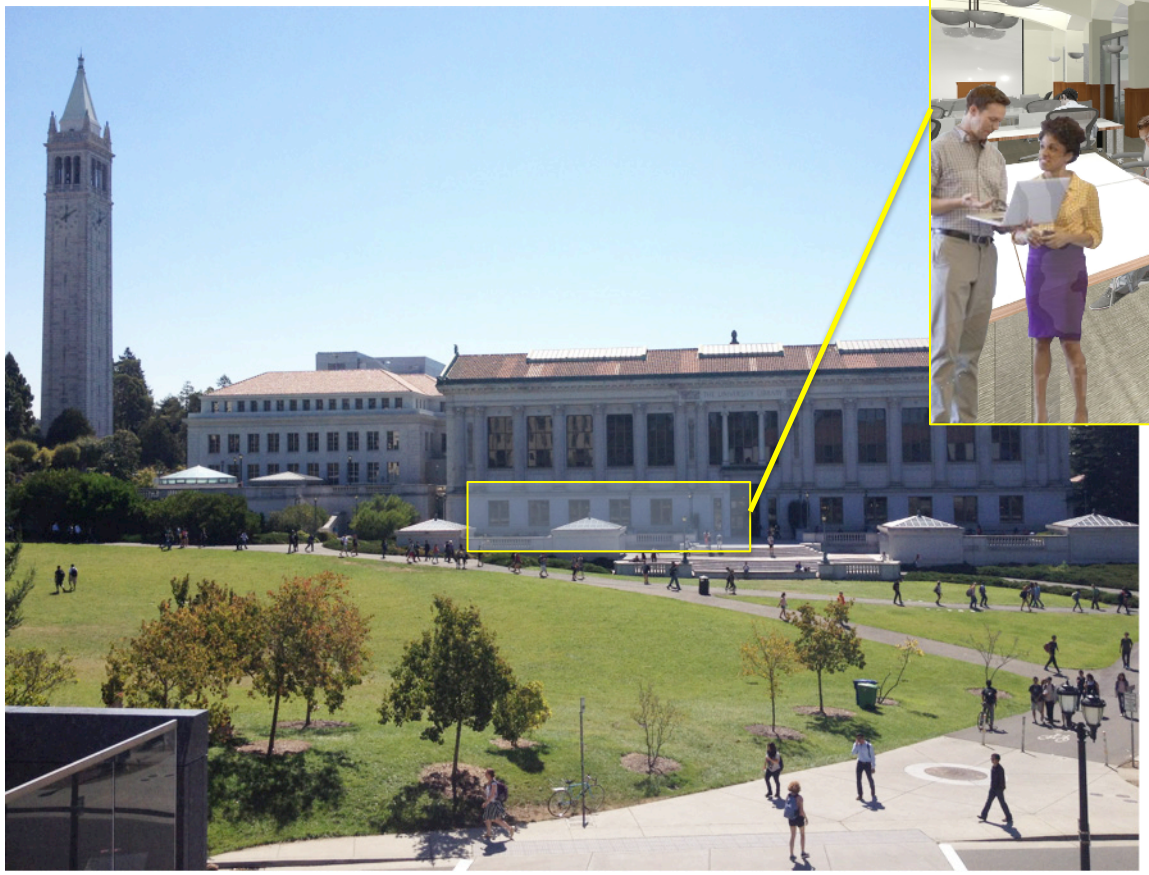
Berkeley Institute for Data Science (BIDS)

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

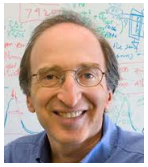
Enhancing strengths of
- Simons Institute for the Theory of Computing
- AMP Lab
- SDAV Institute
- CITRIS
- etc.

Berkeley
UNIVERSITY OF CALIFORNIA

Doe Memorial Library @ the heart of UC Berkeley



Initial Data Science Faculty Group



Faculty Lead/PI: **Saul Perlmutter**, Physics, Berkeley Center for Cosmological Physics

Joshua Bloom, Professor, Astronomy; Director, Center for Time Domain Informatics

Henry Brady, Dean, Goldman School of Public Policy

Cathryn Carson, Associate Dean, Social Sciences; Acting Director of Social Sciences Data Laboratory "D-Lab"

David Culler, Chair, EECS

Michael Franklin, Professor; EECS, Co-Director, AMP Lab

Erik Mitchell, Associate University Librarian

Fernando Perez, Researcher, Henry H. Wheeler Jr. Brain Imaging Center

Jasjeet Sekhon, Professor, Political Science and Statistics; Center for Causal Inference and Program Evaluation

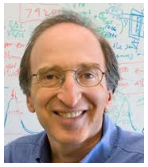
Jamie Sethian, Professor, Mathematics

Kimmen Sjölander, Professor, Bioengineering, Plant and Microbial Biology

Philip Stark, Chair, Statistics

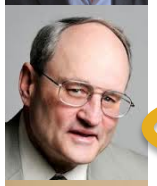
Ion Stoica, Professor, EECS; Co-Director, AMP Lab

Initial Data Science Faculty Group



Faculty Lead/PI: **Saul Perlmutter**, Physics; Berkeley Center for Cosmological Physics

Joshua Bloom, Professor, Astronomy; Director, Center for Time Domain Informatics



Henry Brady, Dean, Goldman School of Public Policy

Cathryn Carson, Associate Dean, Social Sciences; Acting Director of Social Sciences Data Laboratory "D-Lab"

David Culler, Chair, EECS

Michael Franklin, Professor; EECS, Co-Director, AMP Lab

Erik Mitchell, Associate University Librarian

Fernando Perez, Researcher, Henry H. Wheeler Jr. Brain Imaging Center

Jasjeet Sekhon, Professor, Political Science and Statistics; Center for Causal Inference and Program Evaluation

Jamie Sethian, Professor, Mathematics

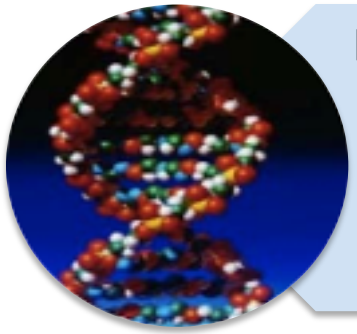
Kimmen Sjölander, Professor, Bioengineering, Plant and Microbial Biology

Philip Stark, Chair, Statistics

Ion Stoica, Professor, EECS; Co-Director, AMP Lab

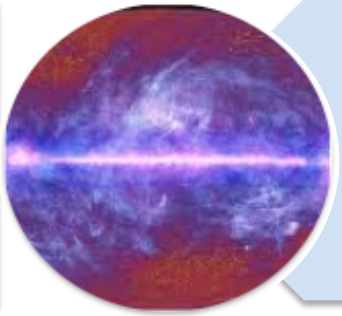
DOE "Big Data" Challenges

Volume, velocity, variety, and veracity



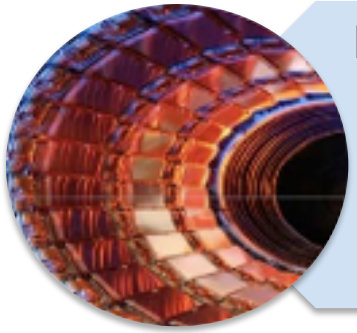
Biology

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



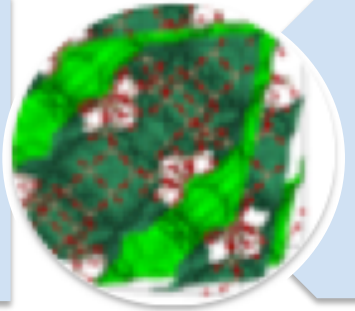
Cosmology & Astronomy:

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



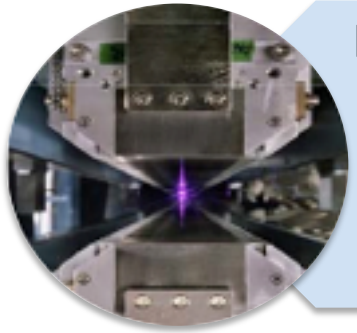
High Energy Physics

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



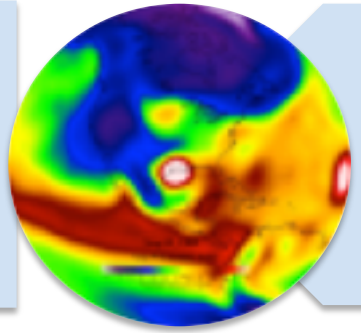
Materials:

- *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, 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**.

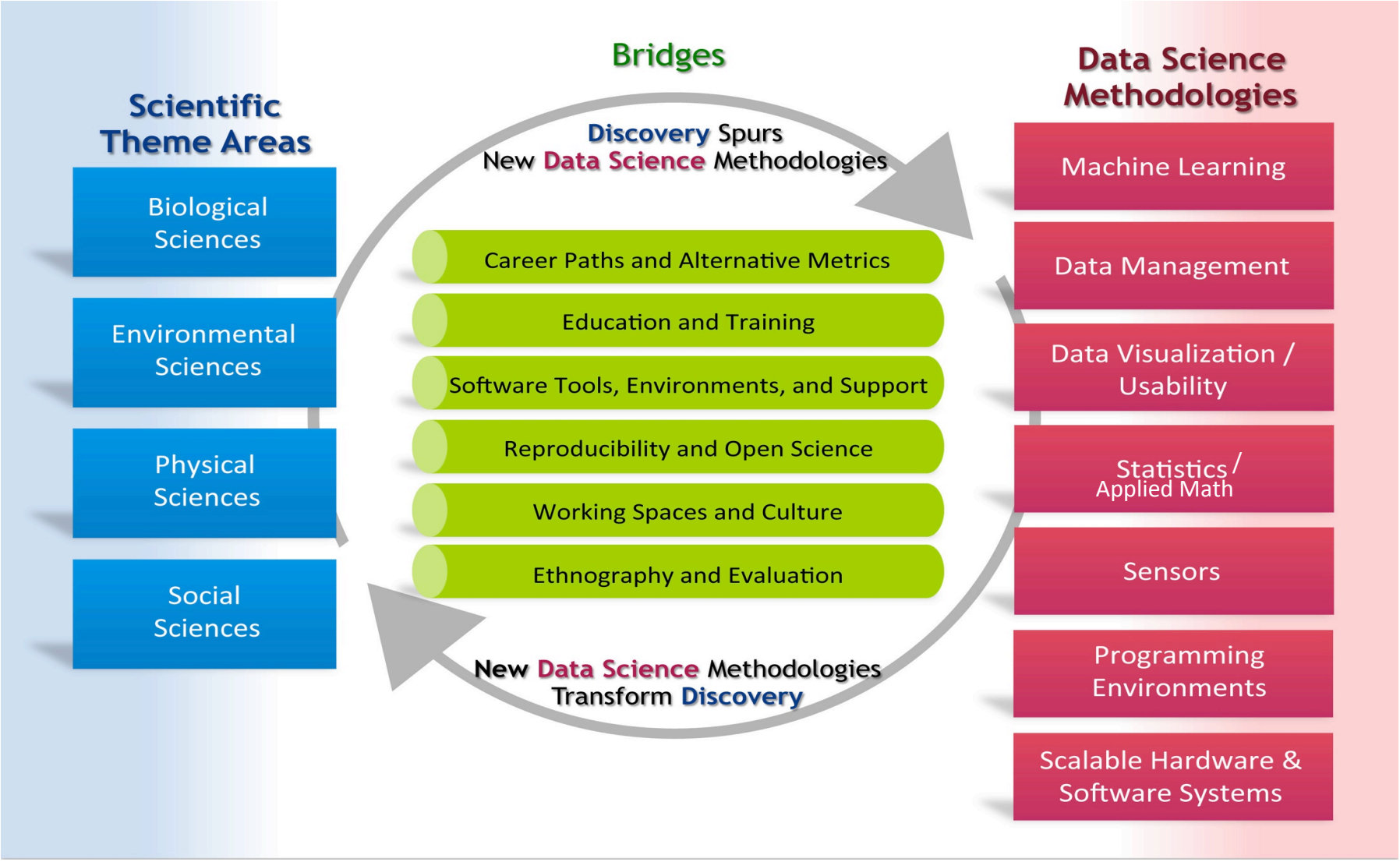
DS for academic scientists: What's still needed?

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**.

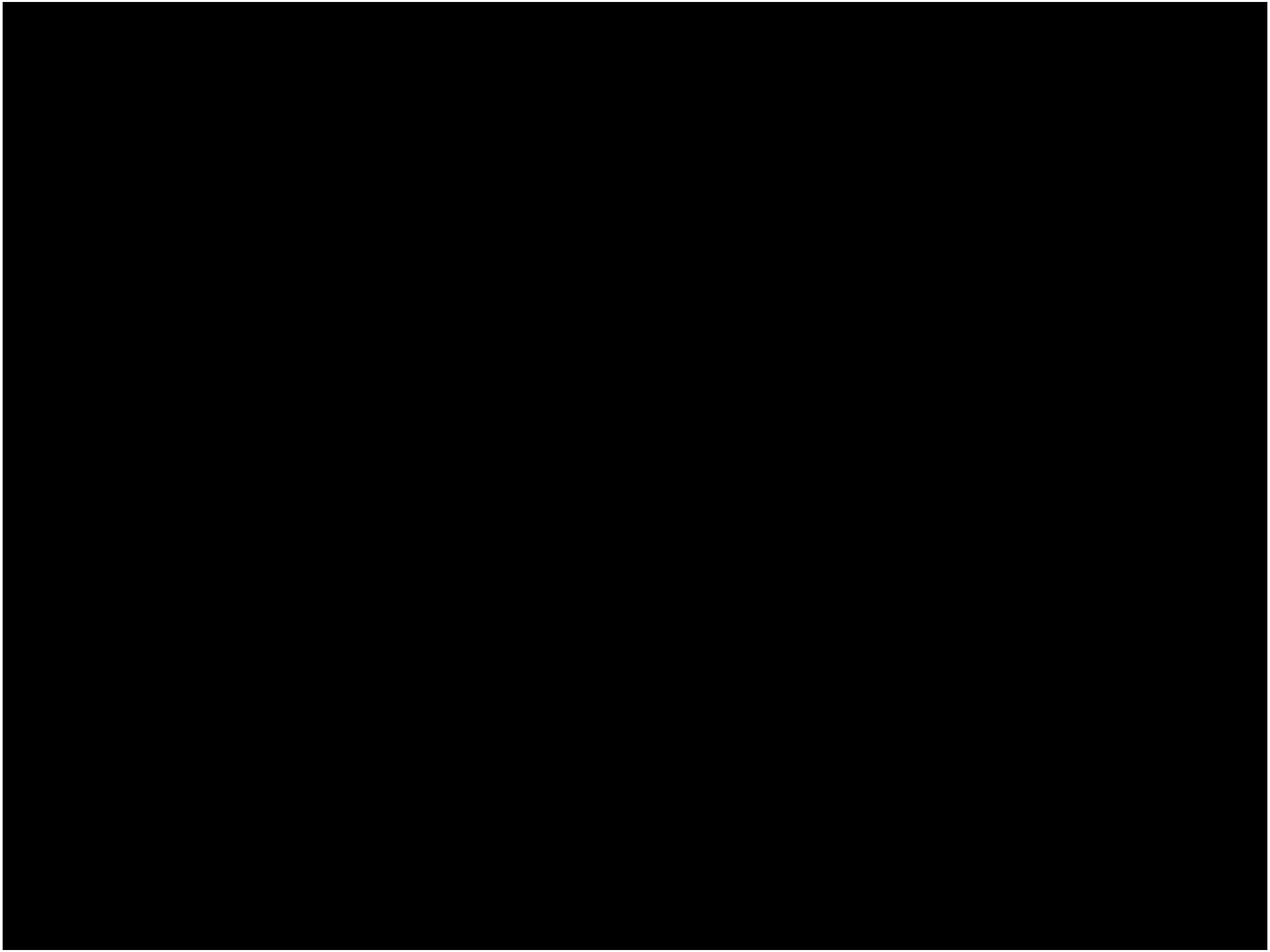
Working Groups as Bridges

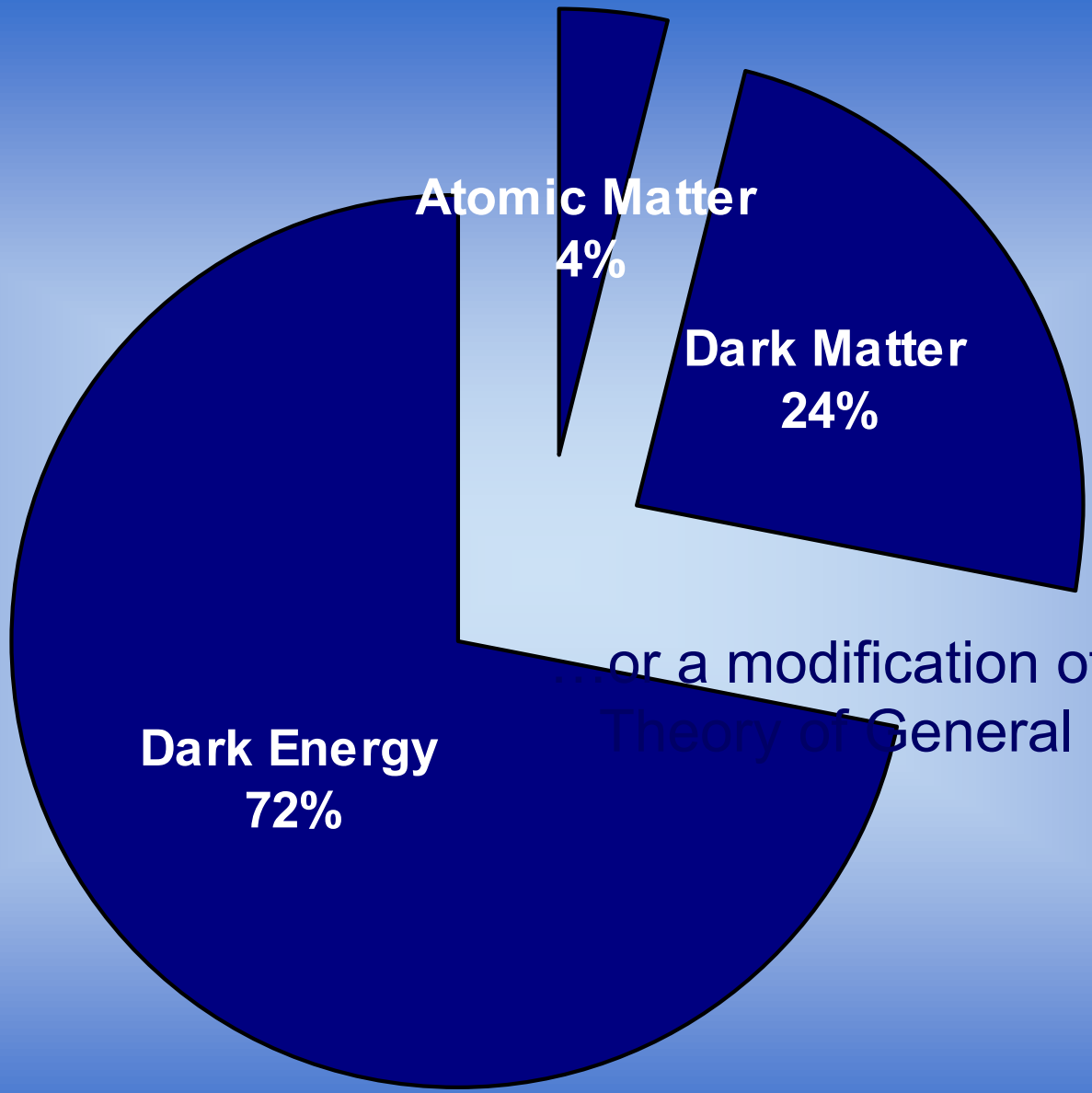


BIDS goals

- **Support meaningful and sustained interactions and collaborations** between
  - Methodology fields: computer science, statistics, applied mathematics
  - Science domains: physical, environmental, biological, neural, social

  to recognize what it takes to move all of these fields forward
- **Establish new Data Science career paths that are long-term and sustainable**
  - A generation of multi-disciplinary scientists in data-intensive science
  - A generation of data scientists focused on tool development
- **Build an ecosystem of analytical tools and research practices**
  - Sustainable, reusable, extensible, easy to learn and to translate across research domains
  - Enables scientists to spend more time focusing on their science



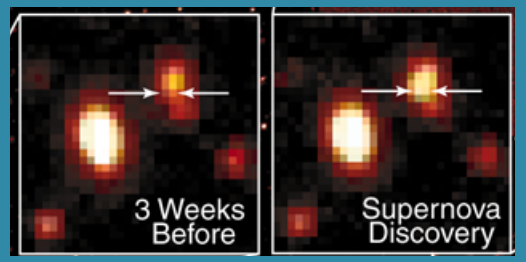


or a modification of Einstein's
Theory of General Relativity?

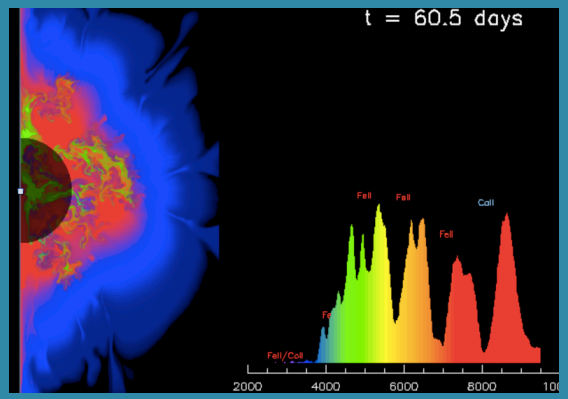
Supernova (SN):

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

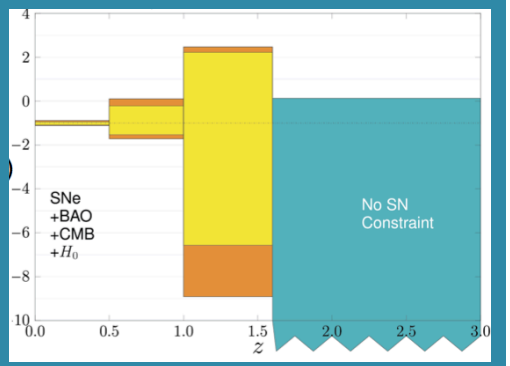
In 1982 (first generation CCDs): 150 MB/night
Current: 1.5 TB/night
LSST era (2020): ~50 TB/night processed



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



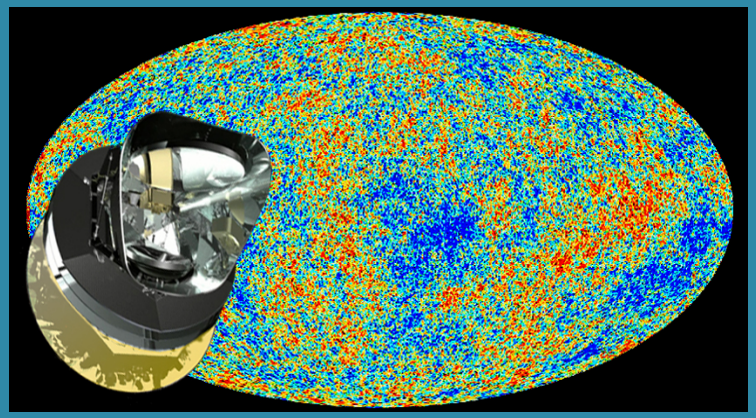
SN observations compared to (supercomputer-based) simulations.



Statistical analyses of cosmological parameters need Markov Chain Monte Carlo (MCMC).

Cosmic Microwave Background (CMB):

Exponentially growing data chasing fainter echos:



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

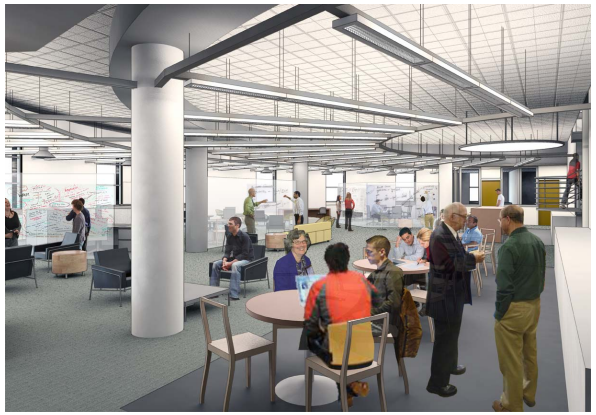
- Simulate 10^4 realizations of the entire mission
- Control both systematic and statistical uncertainties

In 2012/13 alone…

- **Simons Institute for the Theory of Computing**

$60M investment from the Simons Foundation

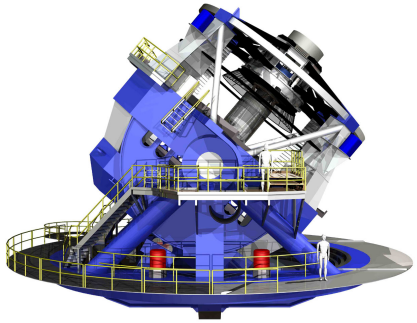
Opened in newly renovated Calvin Hall in Sep 2013



- **"Algorithms, Machines and People (AMP) Expedition"** led by [illegible]in/Ion Stoica secured $10 million NSF grant. The AMPLab also s[illegible] XDATA contract, and has raised $10M to date f[illegible]
- NSF also awarded **$25M f[illegible]** **[illegible]agement, Analysis, and Visualization (SDAV) Institu[illegible]** project with five other National Labs and seven universiti[illegible]
- **New degree program** launched by School of Information
- Significant data science efforts **across many domains**, incl. astrophysics, ecoinformatics, seismology, neuroscience, computational biology, social science

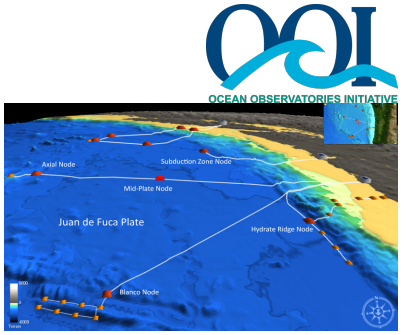
Moore-Sloan Data Science Initiative

Nearly every field of discovery is transitioning from “data poor” to “data rich”

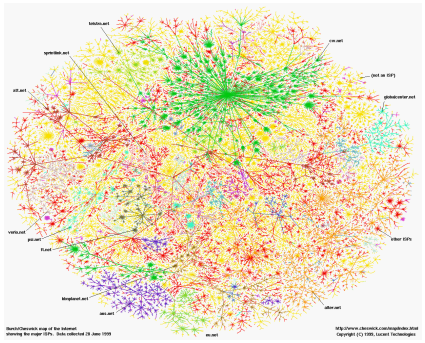


Astronomy: LSST

Physics: LHC



Oceanography: OOI



Sociology: The Web

Biology: Sequencing

Economics: POS terminals

Neuroscience: EEG, fMRI

Exponential improvements in technology and algorithms are enabling a revolution in discovery

- A proliferation of sensors
- The creation of almost all information in digital form
- Dramatic cost reductions in storage
- Dramatic increases in network bandwidth
- Dramatic cost reductions and scalability improvements in computation
- Dramatic algorithmic breakthroughs in areas, such as machine learning