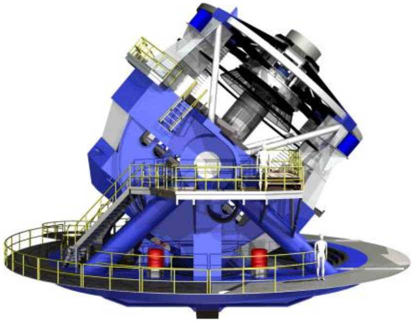
BERKELEY INSTITUTE FOR DATA SCIENCE

Advancing scientific discovery through collaboration across research domains

BIDS
BERKELEY INSTITUTE
FOR DATA SCIENCE

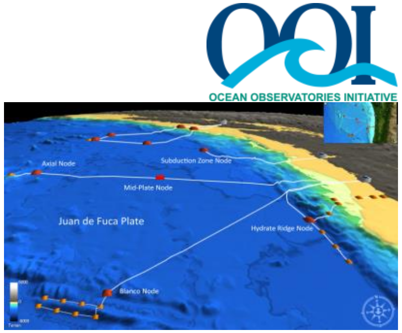
Berkeley
UNIVERSITY OF CALIFORNIA

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

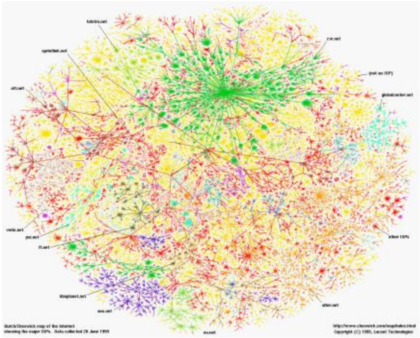


Astronomy: LSST

Physics: LHC



Oceanography: OOI



Sociology: The Web

Biology: Sequencing

Economics: POS terminals

Neuroscience: EEG, fMRI

Data Science growing organically everywhere

WIRED

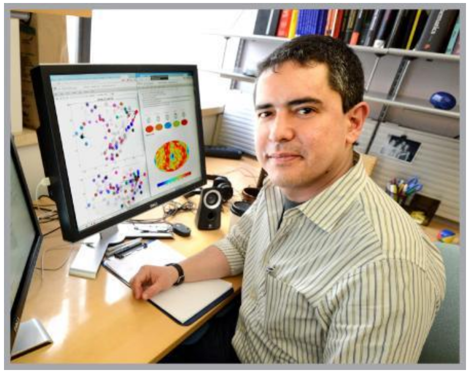
Spark: Open Source Superstar Rewrites Future of Big Data

BY CADE METZ 06.19.13 6:30 AM

AMP Lab

Ion Stoica, CS

Michael Franklin, CS



Fernando Perez,

Brain Imaging Center

iPython tools and community

KBase

PREDICTIVE BIOLOGY

DOE Systems Biology Knowledgebase

Adam Arkin,

Bioengineering

Charles Marshall

Rosie Gillespie

Integrative Biology

Digitized Museum

Reconstructing the movies in your mind

Bin Yu, Statistics

Jack Gallant, Neuroscience

Richard Allen

Earth& Plan. Science

Seismology Lab

The New York Times

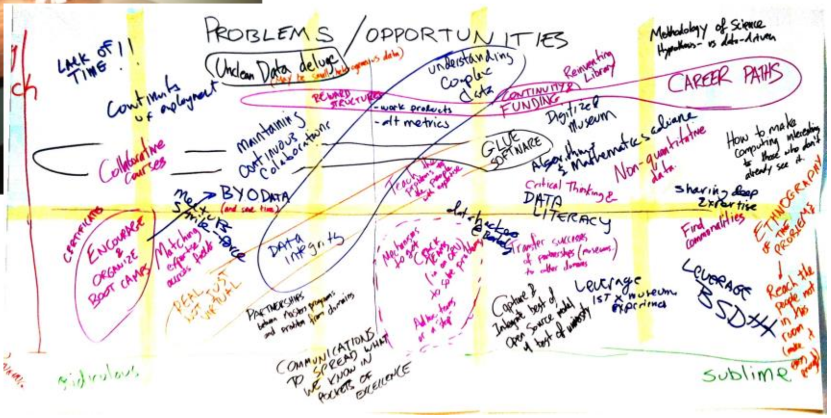
Incomes Flat in Recovery, but Not for the 1%

Feb 15, 2013

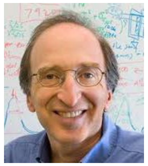
Emmanuel Saez, Economics

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)


PROBLEMS / OPPORTUNITIES
LACK OF TIME !!
Unclean Data deluge
(May be small heterogeneous data)
Continuity of employment
REWARD STRUCTURES
- work products
- alt metrics
understanding complex data
Reinventing Library
CONTINUITY & FUNDING
CAREER PATHS
Methodology of Science
Hypothesis- vs data-driven
Collaborative courses
maintaining continuous collaborations
Digitized Museum
GLUE SOFTWARE
Algorithms & Mathematics alliance
Non-quantitative data
How to make computing interesting to those who don't already see it.
Critical Thinking &
DATA LITERACY
sharing deep expertise
Teach the problems to the people with expertise
BYODATA
(and see time)
Strike force
CERTIFICATES
ENCOURAGE & ORGANIZE BOOT CAMPS
Matching expertise across fields
DATA integrity
REAL NOT JUST VIRTUAL
PARTNERSHIPS between Masters programs and problems from domains
Mechanisms to get crack TEAMS to solve problems
Ad hoc teams
data hackers
Transfer success of partnerships (museums) to other domains
Find commonalities
ETHNOGRAPHY OF THE PROBLEMS
Reach the people not in this room
Capture & Integrate best of Open Source model & best of university
Leverage 1ST x museum experience
LEVERAGE BSDxx
COMMUNICATIONS TO SPREAD WHAT WE KNOW IN POCKETS OF EXCELLENCE
ridiculous
sublime

Initial 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 Science; Acting Director of Social Sciences Data Laboratory "D-Lab"

David Culler, Professor, EECS

Michael Franklin, Chair, 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

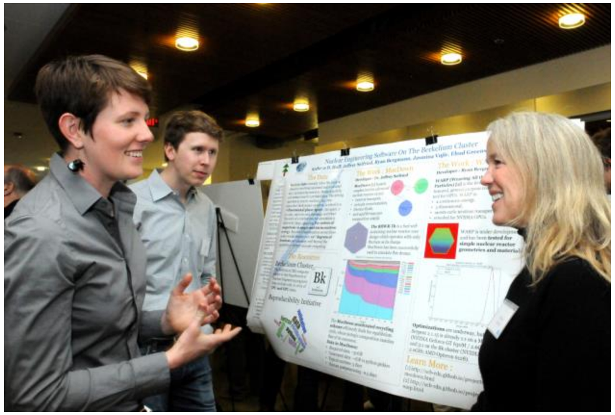
A 5-year, $37.8 million cross-institutional collaboration

NYU — Berkeley UNIVERSITY OF CALIFORNIA — W UNIVERSITY of WASHINGTON

GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION

Launched December 2013



Nuclear Engineering Software On The Berkeley Cluster
Bk
Reproducibility Initiative
Learn More

Our sponsors

- Foundations
  - Moore and Sloan Foundations $12.5 million
- Industry
  - Siemens
  - State Street
- Institutional
  - UC Berkeley

GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION 1934

SIEMENS

STATE STREET

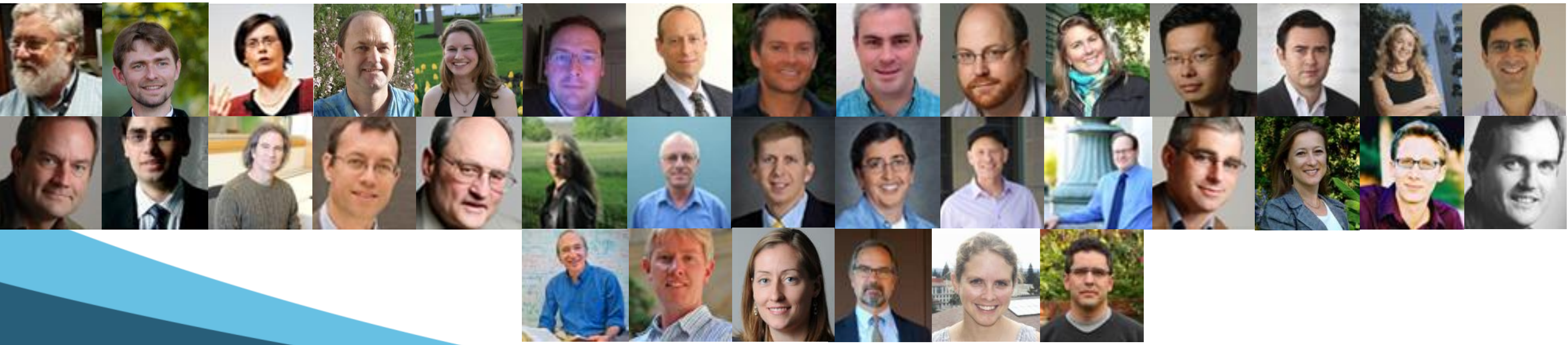
Berkeley UNIVERSITY OF CALIFORNIA

BIDS Goals

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

People are at the heart of BIDS

We are **building a community** that represents some of the brightest researchers across our campus that are **leading the data science revolution** in their own disciplines.



BERKELEY
Institute for
Data Science

Diverse expertise

- Sociology
- Phylogenomics
- Cosmological Physics
- Nuclear Science
- Neuroscience
- Energy and Resources
- System software
- High-performance computing
- Global Change Biology

- Geospatial
- Statistics
- Environmental science
- Computer Vision
- Distributed computing
- Seismology
- Computer Science
- Astronomy
- Public Policy

- Social Sciences
- Psychology
- Library science
- Molecular & Cell Biology
- Political Science
- Mathematics
- Bioengineering
- City & Regional Planning
- …

Diverse Software Development

http://bids.berkeley.edu/research

BIDS Fellows engage in a range of projects that address the ongoing needs of effectively advancing data-intensive research.

IP[y]:
IPython
scikit-image
image processing in python
HOLOS
BERKELEY ECOINFORMATICS ENGINE
The DECIDING FORCE Project
NIMBLE
CLUS
txt_thrshr
arrested 75 people;
another 100 people
delay their start
NumPy
astropy
A Community Python Library for Astronomy
R OpenSci

Project Jupyter

”Jupyter is like IPython, but language agnostic”

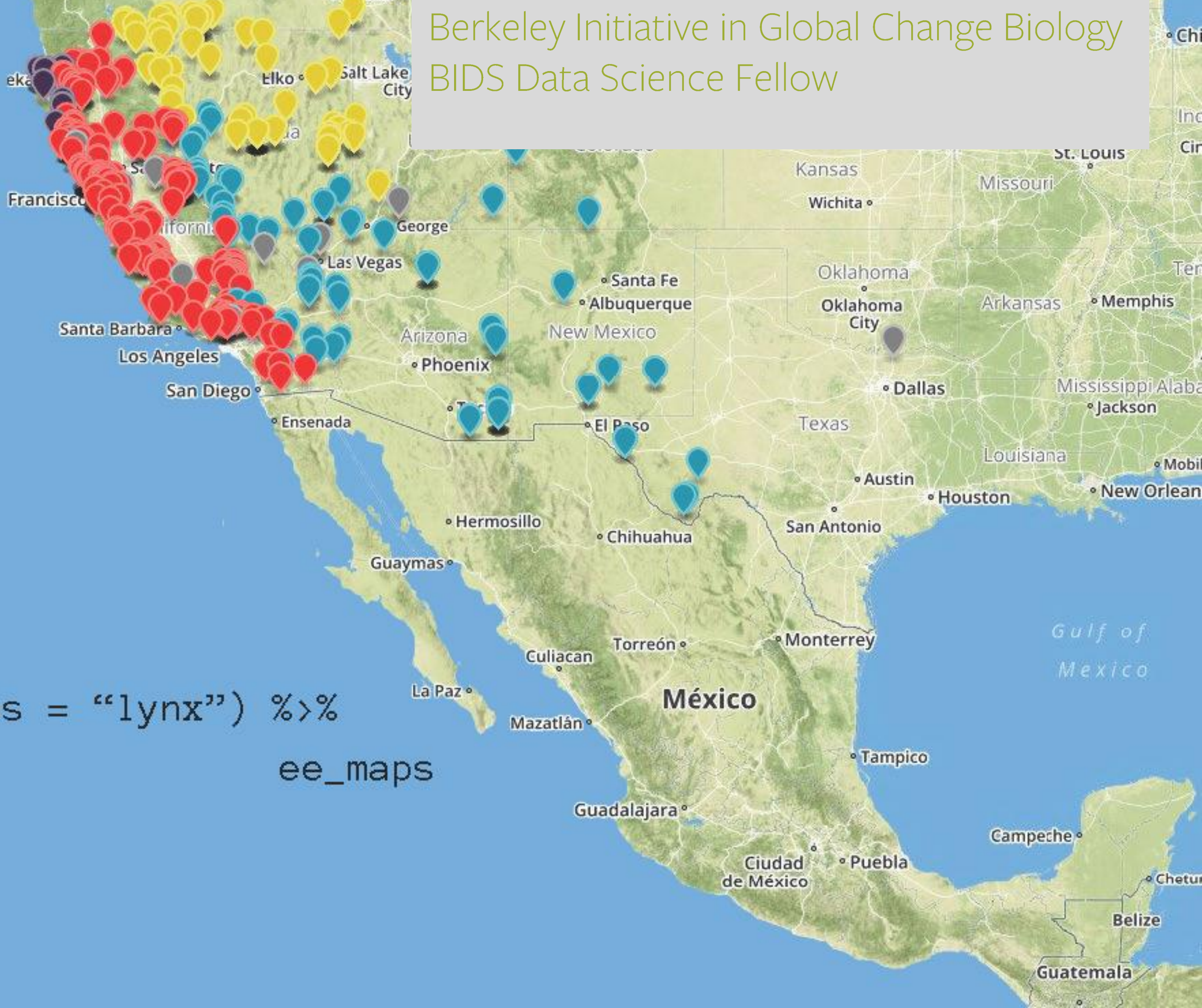
IP[y]:
IPython

BERKELEY
Institute for
Data Science

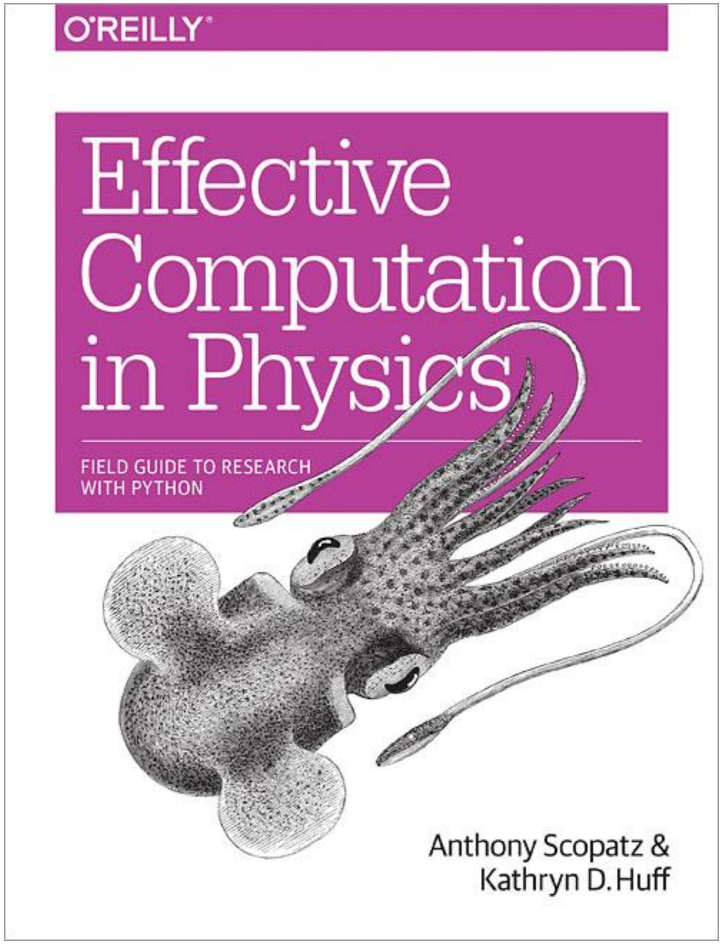
rOpenSci

Karthik Ram

Assistant Researcher
Berkeley Initiative in Global Change Biology
BIDS Data Science Fellow



```
ee_observations(genus = “lynx”) %>%
                          ee_maps
```



O'REILLY

Effective Computation in Physics

FIELD GUIDE TO RESEARCH WITH PYTHON

Anthony Scopatz &
Kathryn D. Huff

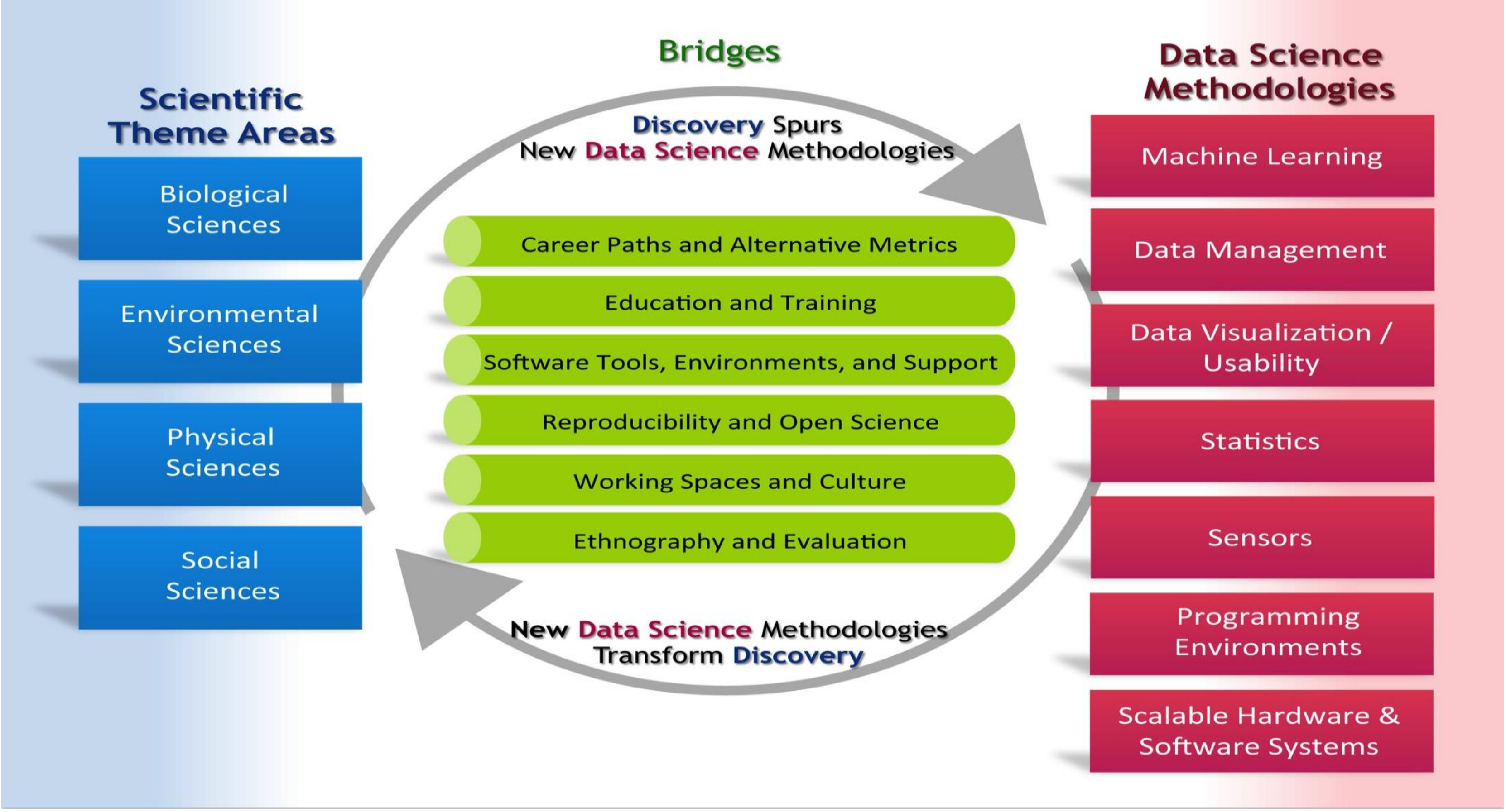
Katy Huff

Nuclear Engineering Postdoc
BIDS Data Science Fellow

- physics.codes
- github.com/physics.codes/examples
- shop.oreilly.com/product/0636920033424.do

Working Groups

Working to address the major challenges facing major advances in data driven research.



Scientific Theme Areas

- Biological Sciences
- Environmental Sciences
- Physical Sciences
- Social Sciences

Bridges

Discovery Spurs New Data Science Methodologies

- Career Paths and Alternative Metrics
- Education and Training
- Software Tools, Environments, and Support
- Reproducibility and Open Science
- Working Spaces and Culture
- Ethnography and Evaluation

New Data Science Methodologies Transform Discovery

Data Science Methodologies

- Machine Learning
- Data Management
- Data Visualization / Usability
- Statistics
- Sensors
- Programming Environments
- Scalable Hardware & Software Systems

Career paths & alternative metrics

Working group aims to identify and promote alternative metrics and career paths that lead to opportunities for growth and advancement for scientists that do not fit the typical academic mold, but are critical to its success.

Education & training

Investigating the requirements for successful adoption of data science approaches.

- Domain scientists need training in the foundations of data science including
  - Programming
  - Statistics
  - Reproducible computational science
- Methodological scientists need training to work productively in domain areas.
- Activities including workshops and bootcamps.

Software tools & environments

This working groups open source emphasis to:

- lead the development of novel, open, high-impact computational tools for data science
- train the next generation of researchers so they can wield computational tools rigorously and effectively

This working group focuses on the software aspects of data science, with an emphasis on bridging the culture of academic research with that of open source software.

Reproducibility & open science

This working group studies the cultural, educational, legal, and technological barriers to reproducible and open research.

Through example, they document and demonstrate the advantages reproducibility has for:

- The scientific process
- How individuals and teams can improve their productivity by adopting tools and workflow that support reproducibility, such as revision-controlled environments.

Ethnography & evaluation

Leveraging faculty expertise in Science and Technology Studies, ethnography, quantitative social scientific research design, and evaluation.

Providing generalizable insights that will inform data science environments at large so BIDS and the campus can use what they find to iterate and improve.

Working spaces & culture

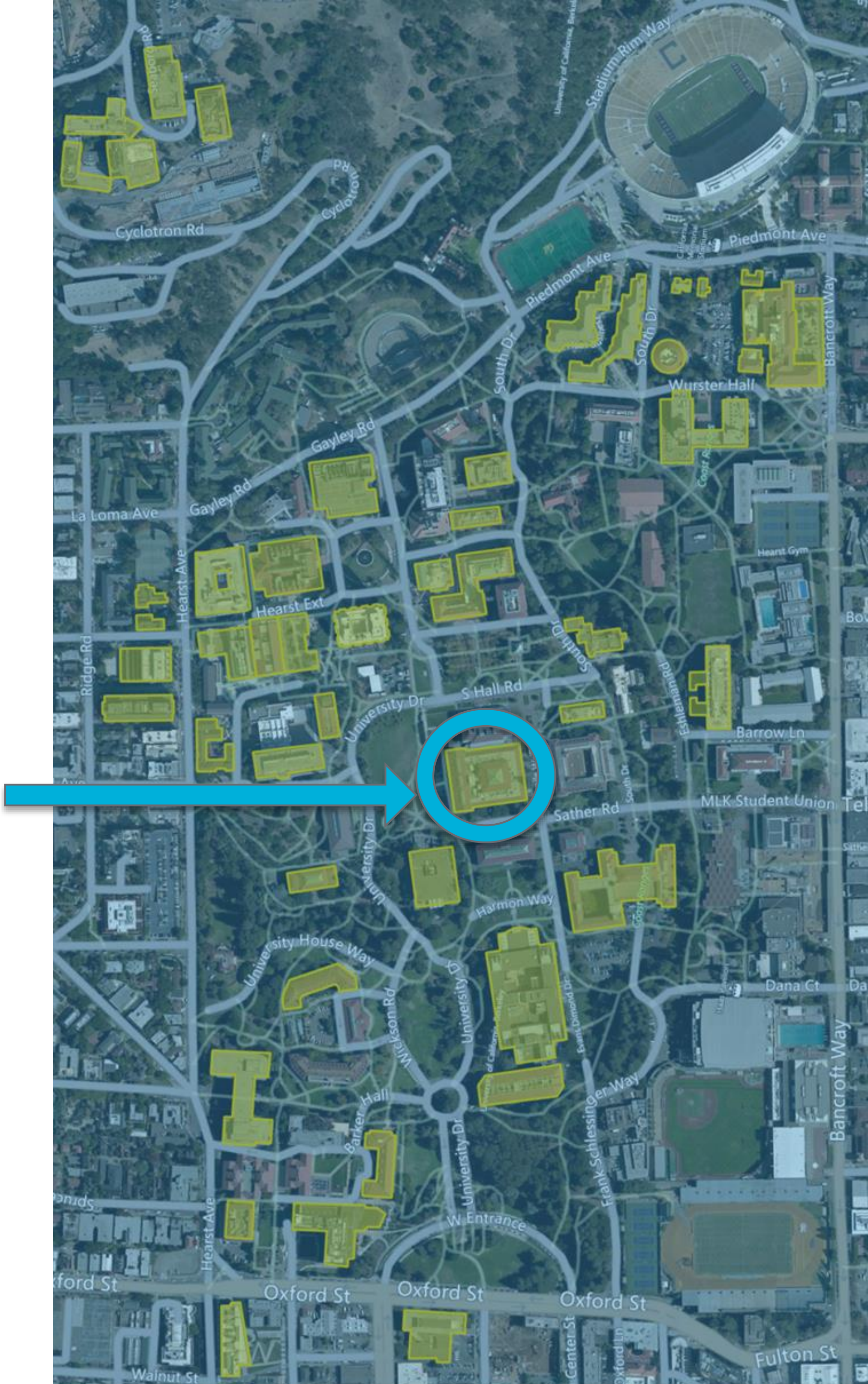
BIDS brings people who are developing data science opportunities to work together in an environment where daily collaboration, through targeted activities and shared physical space, will help grow a real community of practice.

This working group investigates how working space and culture may be used to better engage researchers and promote cross disciplinary collaboration.

Our collaborative space

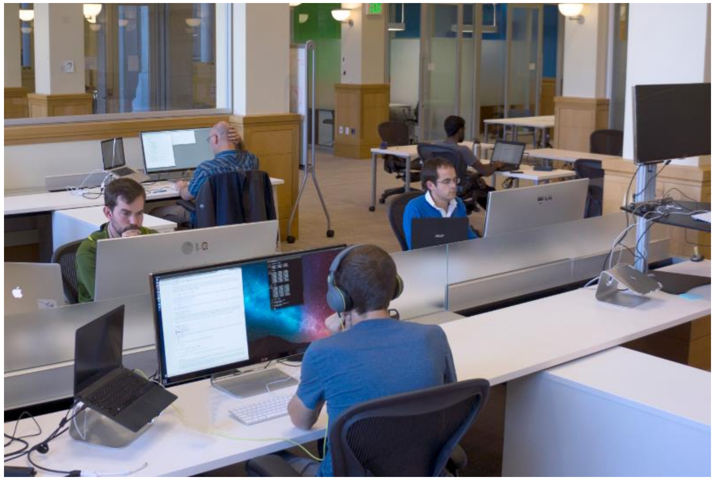
190 Doe Library

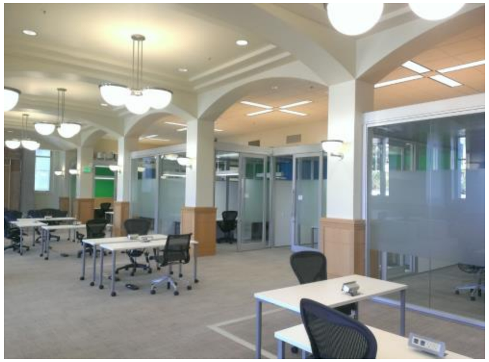
Central location that serves as home for data science efforts



Our collaborative space

190 Doe Library





BIDS Tea

Monday’s, 3:30-4:30pm

Time for networking and discussion

Lightning talk by invited guest

BIDS
BERKELEY INSTITUTE FOR DATA SCIENCE

Berkeley
UNIVERSITY OF CALIFORNIA

The Hacker Within

Wednesdays, 4-6pm

Peers at all levels of experience share topics useful in our scientific software development workflows.

Recent topics:

- Parallel Programming
- Advanced Git
- IPython
- Matplotlib

BERKELEY
Institute for
Data Science

Data Science Collaborative

http://datasci.berkeley.edu/

Interdisciplinary teams working with real-word projects for semester- and year-long commitments.

Current projects involving

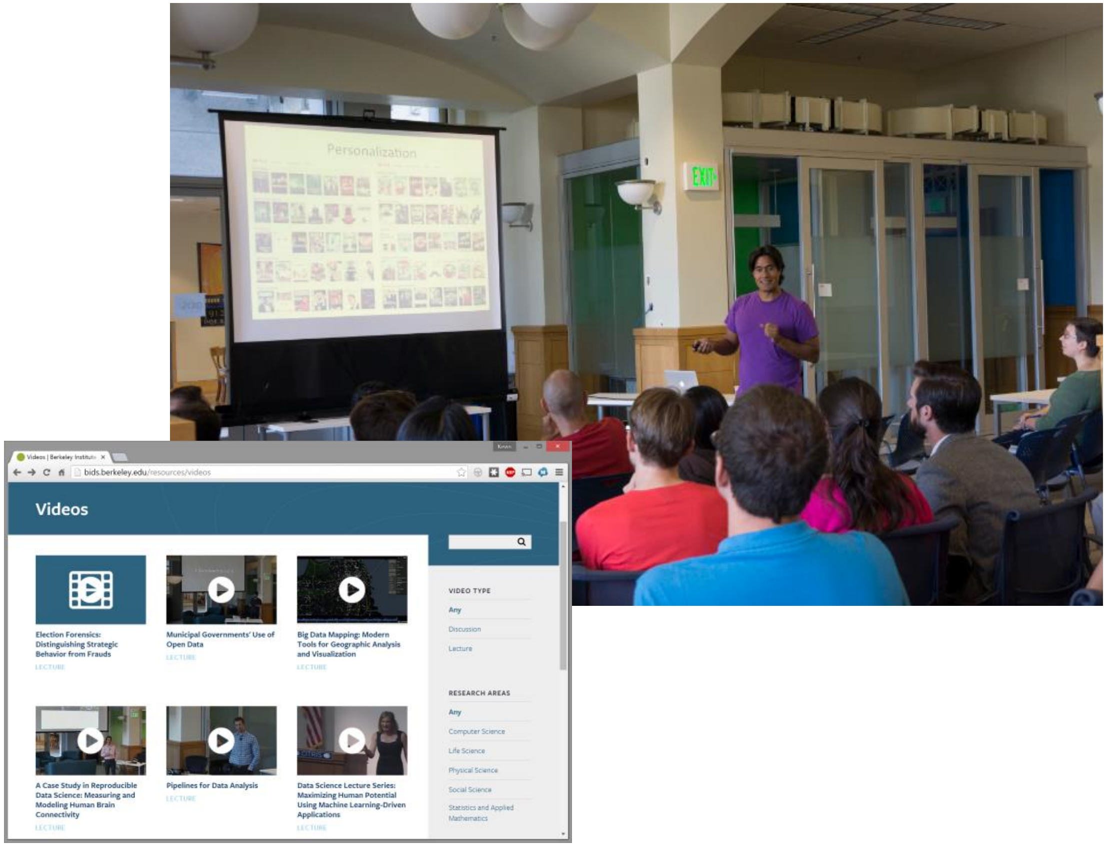
- Government
- Startup
- Finance
- Scientific Research

Data Science Lecture Series

Fridays, 1-2:30pm

Recent speakers from:

- Netflix
- UW
- LinkedIn
- RStudio
- Stanford
- DreamWorks
- Bayes Impact
- Gild
- Code for America



bids.berkeley.edu/resources/videos

Videos

Election Forensics: Distinguishing Strategic Behavior from Frauds
LECTURE

Municipal Governments' Use of Open Data
LECTURE

Big Data Mapping: Modern Tools for Geographic Analysis and Visualization
LECTURE

A Case Study in Reproducible Data Science: Measuring and Modeling Human Brain Connectivity
LECTURE

Pipelines for Data Analysis
LECTURE

Data Science Lecture Series: Maximizing Human Potential Using Machine Learning-Driven Applications
LECTURE

VIDEO TYPE
Any
Discussion
Lecture

RESEARCH AREAS
Any
Computer Science
Life Science
Physical Science
Social Science
Statistics and Applied Mathematics

http://bids.berkeley.edu/resources/videos

BERKELEY
Institute for
Data Science

Data Science Faire

May 5, 2015

- Lightning talks
- Demos
- Posters
- Discussion

Distributed analytics and machine learning with Apache Spark

January 12-14, 2015

Hosted AMPLab workshop teaching researchers how to tackle their big data with Apache Spark

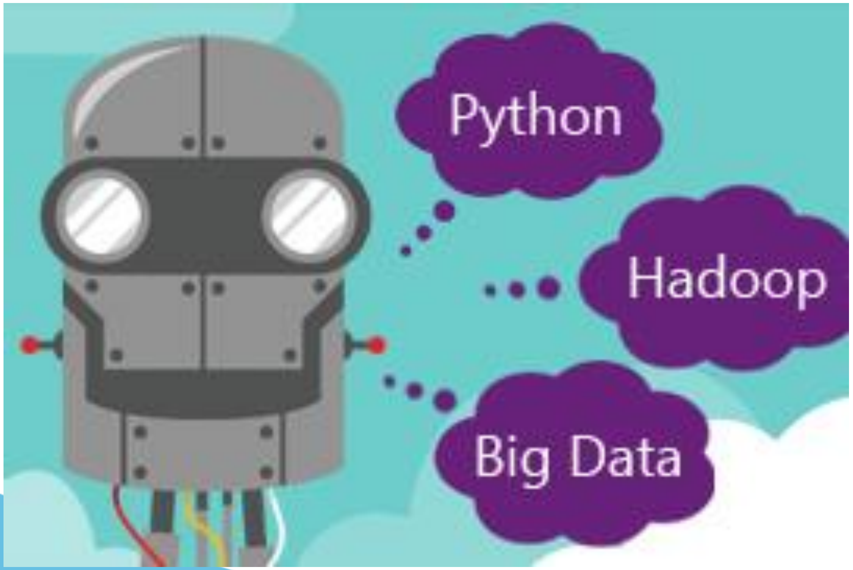
amplab

BERKELEY
Institute for
Data Science

Microsoft Azure for Research Training

February 11, 2015

Acquiring hands-on experience in the major design patterns for successful cloud applications



Microsoft® Research

BERKELEY Institute for Data Science