Institutes for Data Science:

New York University
University of Washington
University of California, Berkeley

Advancing scientific discovery through collaboration across research domains
Data Science growing organically everywhere

AMP Lab
Ion Stoica, CS
Michael Franklin,

Fernando Perez,
Brain Imaging Center
iPython tools and community

Reconstructing the movies in your mind

Bin Yu, Statistics
Jack Gallant, Neuroscience

Richard Allen
Earth& Plan. Science
Geospatial Lab

The New York Times
Incomes Flat in Recovery, but Not for the 1%
Feb 15, 2013
Emmanuel Saez, Economics

Adam Arkin,
Bioengineering

Charles Marshall
Rosie Gillespie
Integrative Biology
Digitized Museum
Strong desire to come together and do more

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)

119 Responses to Data Science Participation
- 2-3 page description of efforts in Data Science
- opportunity and challenges

76 Data Science Fellows Proposals
Applied math / statistics: 1
Computer sciences: 8
Environmental sciences: 8
Life sciences: 20
Meta / other: 12
Neuroscience: 4
Physical sciences: 12
Social sciences: 11
Our sponsors

• Foundations
  o Moore and Sloan Foundations, $37.8 million for 5 years

• Industry partners
  o Collaborative, in-kind, financial support

• Institutional
  o Contributions from recipient universities
  o Space and in-kind contributions from libraries

• Launched fall 2013
Program 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 ecosystem of analytic 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
Working Groups

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

Our open source thrust will:
• 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 will focus on the software aspects of data science, with an emphasis on bridging the culture of academic research with that of open source software.
This working group will study the cultural, educational, legal, and technological barriers to reproducible and open research. Through example, they will 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.
Working spaces & culture

Our Data Science Environment 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 will investigate how working space and culture may be used to better engage researchers and promote cross disciplinary collaboration.
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 Institutes and the campus can use what they find to iterate and improve.
Working Groups

Working to address the major challenges facing major advances in data driven research.
Local perspectives on institutes

• University of Washington
  ○ Betsy Wilson
  ○ Stephanie Wright

• New York University
  ○ Carol Mandel

• University of California, Berkeley
  ○ Philip Stark
  ○ Erik Mitchell
UW Data Science Context

• Depth, breadth, history and velocity
  o Department of Biostatistics (1970)
  o Center for Statistics and Social Science (1999)
  o eScience Institute (2005/2008)
    ▪ Data-driven discovery for all fields

• Provost Hiring Initiative (2012)
• NSF for IGERT interdisciplinary graduate program in Data Science (2013)
• Culture of deep collaboration
• Partnerships with Amazon, Microsoft, Google, Allen Brain Institute, Tableau, etc.
Bottom Up and Top Surround

• Bottom-up, needs-based, driven by the scientists

• Top-surround support from Provost, VP for Research, Deans of Arts and Sciences, Engineering, Environment, Information School, University Libraries
CORE TEAM:
12 Departments and 5 Schools and Colleges
Start Up Focus

- New career paths for interdisciplinary fellows, post-docs, and data scientists
- Creation of the Data Science Studio and programming
- Interdisciplinary education
- Campus-wide communities
- Leverage
Interdisciplinary Community and Education

Events, seminar series, showcases, boot camps, and lunches
Incubation Program

Spring 2014 Incubation Projects

Automated Detection and Analysis of Repeating Earthquakes
Alicia Hotovec-Ellis, Kate Allstadt, Jon Connolly, and John Vidale — Earth and Space Sciences
eScience Contact: Jake Vanderplas

Using social media data to identify geographic clustering of anti-vaccination sentiments
Benjamin Brooks, Abraham Flaxman — Institute for Health Metrics and Evaluation
eScience Contact: Andrew Whitaker

Analysis of Kenya’s Routine Health Information System data
Gregoire Lurton, Abraham Flaxman, Emmanuela Gakidou — Institute for Health Metrics and Evaluation
eScience Contact: Dan Halperin

Efficient Computation on Large Spatiotemporal Network Data
Ian Kelley, Josh Blumenstock — Information School
eScience Contact: Andrew Whitaker

Scalable Manifold Learning for Large Astronomical Survey Data
Marina Meilà — Statistics
eScience Contact: Jake Vanderplas

ASPSIA: Adult Service Providers and Some Incidental Addenda
Sam Henly — Economics
eScience Contact: Andrew Whitaker

Summer 2015 Incubation—Social Good
UW Libraries Participation

• From the beginning
  o Proposal process
  o Reproducibility Working Group (even before it was the Reproducibility Working Group)
    ▪ Data Science Environment Steering Committee
    ▪ Education & Training Working Group
    ▪ Data Science Studio Governance Board
UW Libraries Participation
Data Science Kickoff Session

- 137 posters
- 30+ departments and units

http://escience.washington.edu/event/data-science-university-washington-campus-conversation
UW Libraries Participation

Reproducibility Workshop

Participants

• NYU & Berkeley
• Allen Inst for Brain Sci
• Google
• Sage Bionetworks
• Fred Hutchinson Cancer Research Ctr

http://escience.washington.edu/event/first-reproducibility-workshop
UW Libraries Participation

WRF Data Science Studio

http://escience.washington.edu/dss
UW Libraries Participation

Studio Hours

• eScience Data Scientists
• UW-IT
• Center for Statistics & the Social Sciences
• Amazon Web Services Scientific Computing Team

http://escience.washington.edu/dss-hours
UW Libraries Participation

**DRUW – Data Repository @ UW**

- Fedora 4 / Hydra
- Self-deposit
- Collaboration with UW-IT
- Tentative rollout in Fall 2015

[Link to Blog Post](http://data.blogspot.com/2015/01/druw-glance-under-hood.html)
Local perspectives on institutes

- University of Washington
  - Betsy Wilson
  - Stephanie Wright

- New York University
  - Carol Mandel

- University of California, Berkeley
  - Philip Stark
  - Erik Mitchell
NYU M/S Data Science Environment

- NYU Center for Data Science (CDS) is a university initiative to establish NYU as a leader in Data Science training and research
- CDS builds on strengths in Courant Institute of Mathematics and partnerships with “Silicon Alley” (e.g., Facebook, Google, Medidata)
- University is investing in space, faculty, and new academic programs
- M/S Data Science Environment is a key part of CDS, and is located in CDS space
NYU M/S Data Science Environment

- 6 inter-institutional working groups, plus a 7th on Methods, aimed at joining domain scientists with new methodologies
- Start up focus has been on:
  - hiring fellows, post-docs, and a lead engineer
  - interdisciplinary events, seminars, showcases
  - tools development
  - design of working group projects
  - space development at CDS
NYU/MS Data Science Environment and the Libraries

• Library participation has been “light” but is becoming more active in: Space, Ethnography, Reproducibility/Open Science

• Importance of space may be illustrated by our lack of co-location – now need to overcome

• Planning a library position to be shared by and dedicated to Reproducibility and Open Science work
NYU M/S Data Science Environment and the Libraries

• Projects may include, e.g.:
  o Tracking (and improving) data lifecycles in selected labs
  o Creating and applying a “reproducibility badge”
  o Identifying and exposing model work
  o Developing a tools registry
  o Exploring deep preservation issues for a data set(s) dependent on software/tools
  o Outreach, promotion, events
Center for Urban Science and Progress (CUSP) cusp.nyu.edu

• CUSP Mission:
  o Use NYC as a laboratory and classroom to help cities become more productive, livable, equitable and resilient

• CUSP Method:
  o Urban Informatics, integrating and analyzing large, heterogeneous data resources
Urban Data includes:

- Environmental sensing (including new Urban Observatory)
- Transportation and security camera images
- Soundscape
- Mass transit, taxis, traffic
- Real estate, buildings
- Agency data: NYC government creates a terabyte of raw data every day, from parking tickets to electricity
CUSP Data User Facility

- Secure data warehouse to ingest and process data and a computing environment to make it available and useful for research across disciplines and perspectives
- Requires expertise across computer science, social science, legal policy, security – and, of course, data curation
CUSP Data User Facility

• CUSP/Libraries partnership: make more generally available an improved, integrated version of NYC OpenData (a subset of CUSP data warehouse)

• Libraries’ role: additional curation and metadata, tools for teaching, outreach within NYU and regionally
Local perspectives on institutes

- University of Washington
  - Betsy Wilson
  - Stephanie Wright
- New York University
  - Carol Mandel
- University of California, Berkeley
  - Philip Stark
  - Erik Mitchell
Our collaborative space

190 Doe Library

Central location that serves as home for data science efforts
UCB Perspective: How we work

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.
UCB perspective: 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
UCB perspective: Office Hours

http://bids.berkeley.edu/resources/office-hours

BIDS fellows are available throughout the week for office hours to discuss data-intensive research.
Focus: Reproducibility @ UCB

http://datasci.berkeley.edu/

Interdisciplinary team working to understand reproducibility across M/S scientists.

Current projects include:

• Assessing reproducibility of work/workflows
• Building frameworks to evaluate & improve reproducibility
• Software: BCE, IPython, ...
• Weaving reproducibility into the curriculum foundationally
• Reproducibility case studies
• Upcoming reproducibility workshop 5/21-22
Events: Data Science Faire

May 5, 2015
Interactive event featuring:
• Lightning talks
• Demos
• Posters
• Discussion
Panel discussion

• University of Washington
  ○ Betsy Wilson (betsyw@u.washington.edu)
  ○ Stephanie Wright (swright@uw.edu)

• New York University
  ○ Carol Mandel (carol.mandel@nyu.edu)

• University of California, Berkeley
  ○ Philip Stark (pbstark@berkeley.edu)
  ○ Erik Mitchell (erik@berkeley.edu)