

# The Moore/Sloan Data Science Environments: Advancing Data-Intensive Discovery

**Ed Lazowska**

**Bill & Melinda Gates Chair in  
Computer Science & Engineering**

**Founding Director, eScience Institute**

**University of Washington**

**AAAS**

**February 2015**

GORDON AND BETTY  
**MOORE**  
FOUNDATION



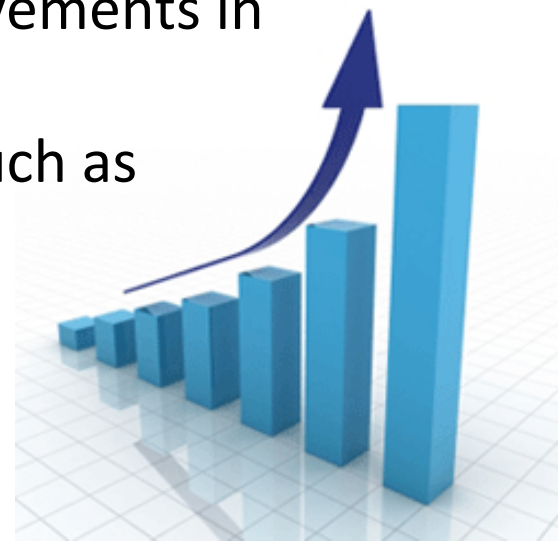
## Today

- Jim Gray's "Fourth Paradigm": Data-intensive discovery
- Background on the Moore/Sloan Data Science Environments project
- A view through the lens of the University of Washington eScience Institute

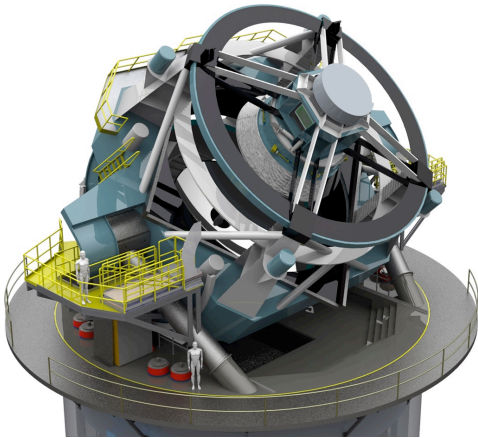


# Exponential improvements in technology and algorithms are enabling a revolution in discovery

- A proliferation of sensors
- Ever more powerful models producing data that must be analyzed
- 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



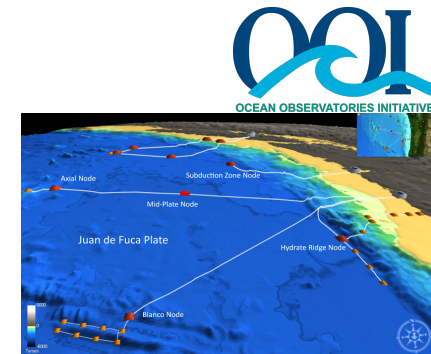
# Nearly every field of discovery is transitioning from “data poor” to “data rich”



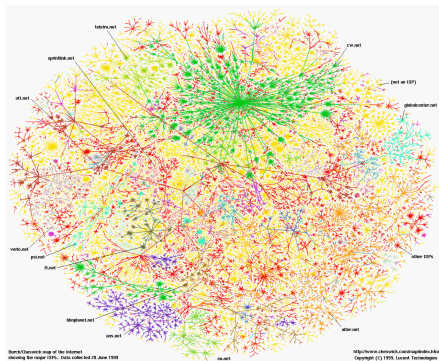
Astronomy: LSST



Physics: LHC



Oceanography: OOI



Sociology: The Web



Biology: Sequencing



Economics: POS  
terminals



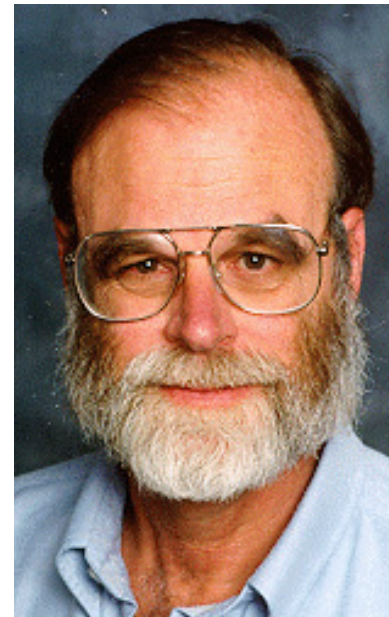
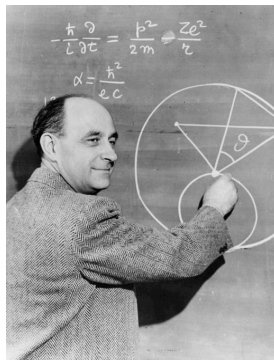
Neuroscience: EEG, fMRI

## “From data to knowledge to action”

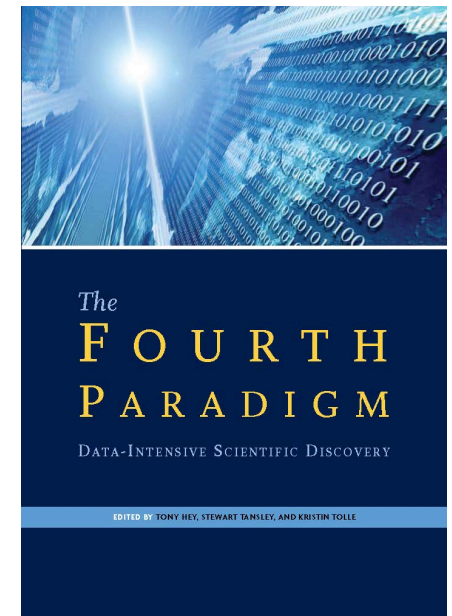
- The ability to extract knowledge from large, heterogeneous, noisy datasets – to move “from data to knowledge to action” – lies at the heart of 21st century discovery
- To remain at the forefront, researchers *in all fields* will need access to state-of-the-art data science methodologies and tools
- These methodologies and tools will need to advance rapidly, driven by the requirements of discovery
- Data science is driven more by *intellectual infrastructure* (human capital) and *software infrastructure* (shared tools and services – digital capital) than by hardware
- Data science is inextricably linked to the commercial cloud: cost-effective scalable computing and storage for everyone

# The Fourth Paradigm

1. Empirical + experimental
2. Theoretical
3. Computational
4. Data-Intensive



Jim Gray



*Each augments, vs. supplanting, its predecessors – “another arrow in the quiver”*



# Genesis of the Moore/Sloan Data Science Environments project

- The Foundations have a focus on novel advances in the physical, life, environmental, and social sciences
- They recognized the emergence of data-intensive discovery as an important new approach that would lead to new advances
- They perceived a number of impediments to success
- They sought partners who were prepared to work together in a distributed collaborative experiment focused on tackling these impediments

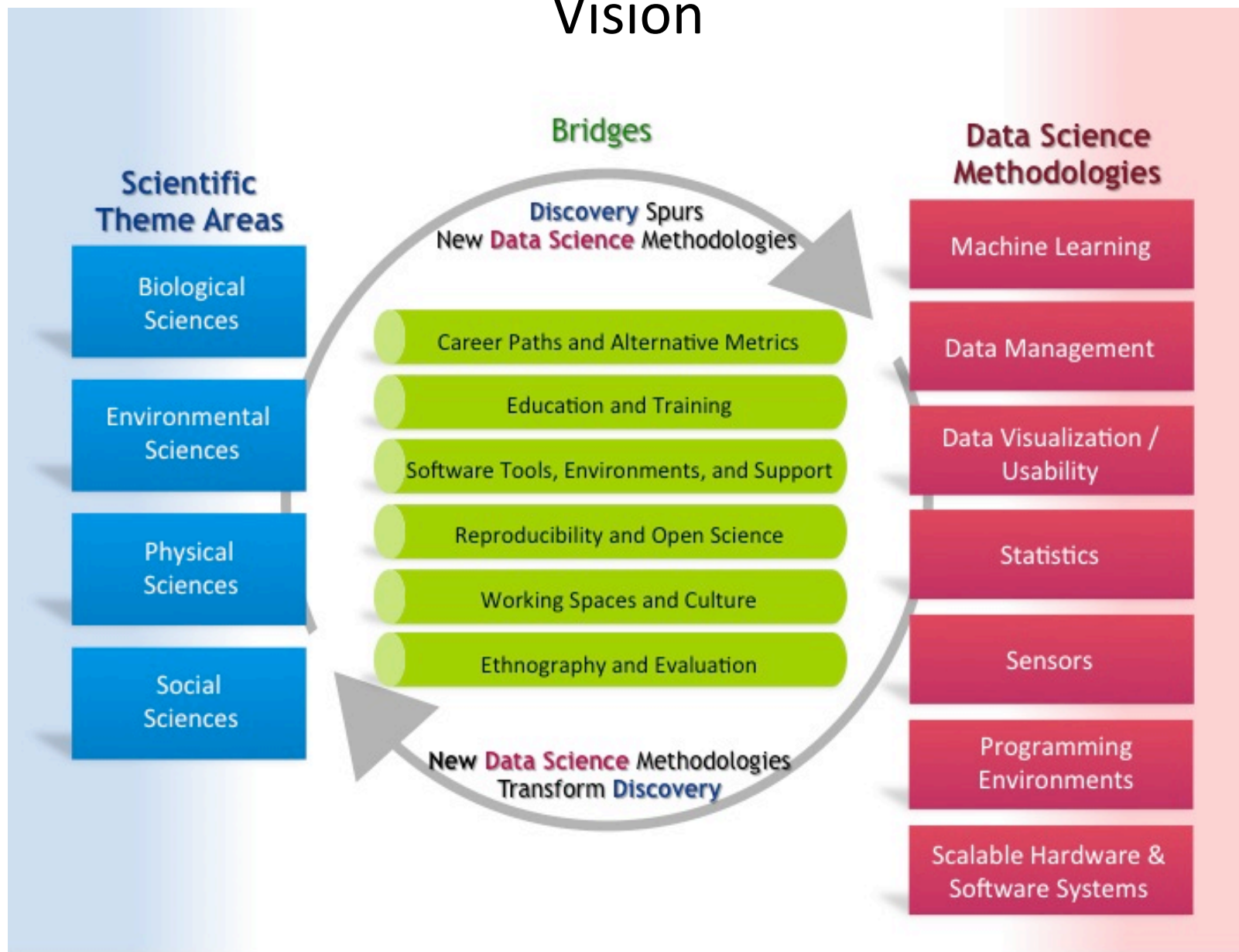


## Chronology of the project

- December 2012: Letters of Interest solicited
- April 2013: Site visits conducted
- May 2013: NYU, UC Berkeley, and UW selected as partners
- June 2013: Multi-day facilitated retreat (~8 attendees from each university, plus Foundation staff, plus facilitators); Working Groups established
- August 2013: Linked proposals submitted
- October 2013: Foundations' Boards act
- November 2013: Formal project launch
- October 2014: First Annual Moore/Sloan Data Science Summit (~30 attendees from each university, plus Foundation staff)



# Vision



## Approach

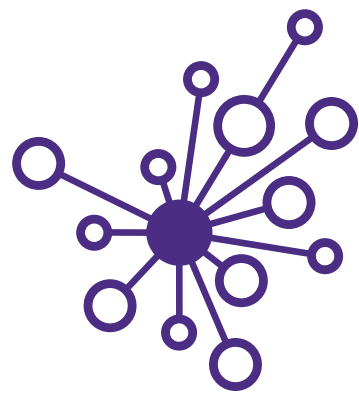
- Carry out a distributed collaborative experiment in creating university environments in which data-intensive discovery flourishes
  - Conduct breakthrough research, in
    - Methodology areas
    - Scientific theme areas
  - Enable breakthrough research, by tackling issues including
    - Career paths
    - Education and training
    - Tools
    - Reproducible research
    - Working spaces & culture
    - Ethnography
- While the balance varies among individuals, all participants are “in this” for their institutions and their fields, in addition to themselves











UNIVERSITY *of* WASHINGTON

# eScience Institute

ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS



<http://escience.washington.edu/>



<http://cds.nyu.edu/>



## A view through the lens of the University of Washington eScience Institute

- Center for Statistics and the Social Sciences
  - Established in 1999
  - Multiple faculty joint-appointed between Statistics and a social science department
  - State-of-the-art statistical methodology tracks in the Ph.D. programs of 10 social science departments
- eScience Institute
  - Established in 2008 (genesis in 2005)
  - Deep collaborations between computer scientists on one hand, and life scientists (environmental oceanography, neuroscience) and physical scientists (survey astronomy) on the other



➔ Moore/Sloan team and today's eScience Institute

## Major sources of support for our “core effort”

- University of Washington
  - \$550,000/year for staff support
  - \$600,000/year for faculty support
- National Science Foundation
  - \$2.8 million over 5 years for graduate program development and Ph.D. student funding (IGERT)
- Gordon and Betty Moore Foundation and Alfred P. Sloan Foundation
  - \$37.8 million over 5 years to UW, Berkeley, NYU
- Washington Research Foundation
  - \$9.3 million over 5 years for faculty recruiting packages, postdocs
    - Also \$7.1 million to the closely-aligned Institute for Neuroengineering (Tom Daniel and Adrienne Fairhall)



# Original core faculty team

## Data science methodology



Cecilia Aragon  
Human Centered  
Design & Engr.



Magda Balazinska  
Computer Science  
& Engineering



Emily Fox  
Statistics



Carlos Guestrin  
CSE



Bill Howe  
CSE



Jeff Heer  
CSE



Ed Lazowska  
CSE

## Life sciences



David Beck  
Chemical Engr.



Tom Daniel  
Biology



Bill Noble  
Genome Sciences

## Environmental sciences



Ginger Armbrust  
Oceanography



Randy LeVeque  
Applied  
Mathematics



Thom. Richardson  
Statistics, CSSS



Werner Stuetzle  
Statistics

## Social sciences



Josh Blumenstock  
iSchool



Mark Ellis  
Geography



Tyler McCormick  
Sociology, CSSS

## Physical sciences



Andy Connolly  
Astronomy



John Vidale  
Earth & Space Sciences

## Original core faculty team

### Data science methodology



Cecilia Aragon  
Human Centered  
Design & Engr.



Magda Balazinska  
Computer Science  
& Engineering



Emily Fox  
Statistics



Carlos Guestrin  
CSE



Bill Howe  
CSE



Jeff Heer  
CSE



Ed Lazowska  
CSE

### Life sciences



David Beck  
Chemical Engr.



Tom Daniel  
Biology



Bill Noble  
Genome Sciences



Ginger Armbrus  
Oceanography



Randy LeVeque  
Applied  
Mathematics



Thom. Richardson  
Statistics, CSSS



Werner Stuetzle  
Statistics

### Social sciences



Josh Blumenstock  
iSchool



Mark Ellis  
Geography



Tyler McCormick  
Sociology, CSSS

### Physical sciences



Andy Connolly  
Astronomy



John Vidale  
Earth & Space Sciences

13 Departments  
5 Schools / Colleges

## Career paths and alternative metrics

*Flagship activity: Establish two new roles on campus: “Data Science Fellows” and “Data Scientists”*

- Recruited / recruiting data scientists – and put processes into place
  - Typically Ph.D.-educated; fully supported by DSE; research position with emphasis on taking responsibility for core activities (e.g., incubator projects)
- Recruited / recruiting research scientists – and put processes into place
  - Typically Ph.D.-educated; partially supported by DSE; research position with emphasis on specific science goals
- Designated 33 faculty and staff as Data Science Fellows – ditto
  - We cribbed Berkeley’s excellent idea
- Recruited four “Provost’s Initiative” faculty members – ditto
  - Provost provide 6 faculty “half-positions”
  - Truly “pi-shaped” – strength and commitment both to advancing data science methodology and to applying it at the forefront of a specific field
  - Astronomy, Biology, Mechanical Engineering, Sociology
- Recruited an initial cohort of 6 Data Science Postdoctoral Fellows – ditto
  - Each is co-mentored by “methodology” and “applications” faculty



## Education and training

*Flagship activity: Establish new graduate program tracks in data science*

- IGERT Ph.D. program in Big Data / Data Science
  - Seven departments have put in place *Big Data Tracks*
    - Data science classes count toward Ph.D. degree (no extra work)
    - Departments: Astronomy, Biology, Chemical Engineering, Computer Science & Engineering, Genome Sciences, Oceanography, and Statistics
  - Started IGERT seminar as the eScience Community Seminar
  - Put in place detailed program evaluation plan with Data2Insight
  - First cohort of 6 students from a variety of departments
    - Each is co-mentored by “methodology” and “applications” faculty
- Workshops and Bootcamps
  - Multiple Software Carpentry Bootcamps (Python, R, etc.)
  - AstroData Hack Week
  - Many others



- Two vibrant seminar series
  - eScience Community Seminar (weekly, centered on IGERT students and Data Science Postdoctoral Fellows)
  - Data Science Seminar (external “distinguished lectures” targeting the campus at large)
- Education working group is actively tracking *all* relevant curricular activities campus-wide

## UW Data Science Seminar

ANALYSIS, VISUALIZATION & DISCOVERY







The Data Science Seminar is a university-wide effort bringing together thought-leading speakers and researchers across campus to discuss topics related to data analysis, visualization and applications to domain sciences. The seminar is typically held on Wednesdays 3:30-4:30pm. Locations for Winter Quarter 2015 will be announced for each speaker individually.

*All talks are free and open to the public.*

### 2015 Speakers

JAN 14		<b>Algorithms for Analyzing On-Line Social Network Data</b> Jon Kleinberg <i>Professor, Cornell University</i>
JAN 28		<b>Data Visualization at the New York Times</b> Amanda Cox <i>New York Times</i>
FEB 4		<b>Christopher Ré</b> <i>Assistant Professor, Stanford University</i>
FEB 18		<b>Prediction in Social Science</b> Sendhil Mullainathan <i>Professor, Harvard University</i>
FEB 25		<b>Simplicity, Complexity, and Duplicity in Visualizations</b> Martin Wattenberg <i>Co-Director of the "Big Picture" Visualization Group, Google</i>
MAR 4		<b>The Emerging Scholarly Brain (with Applications)</b> Michael Kurtz <i>Harvard-Smithsonian Center for Astrophysics, Harvard University</i>
APR 22		<b>Lada Adamic</b> <i>Computational Social Scientist, Facebook</i>
TBD		<b>Paul Ginsparg</b> <i>Professor, Cornell University</i>

### Previous Speakers (2014)

APR 16		<b>People, Data and Analysis</b> Pat Hanrahan <i>Professor, Stanford University &amp; Co-Founder, Tableau Software</i>
APR 23		<b>Machine Learning and Econometrics</b> Hal Varian <i>Chief Economist, Google</i>
MAY 21		<b>What Academia Can Learn From Open Source</b> Arfon Smith <i>Scientist, GitHub &amp; Co-Founder, Zooniverse</i>
OCT 8		<b>Can Cascades be Predicted?</b> Jure Leskovec <i>Assistant Professor, Stanford University</i>
OCT 15		<b>Algorithms for Interpretable Machine Learning</b> Cynthia Rudin <i>Associate Professor, MIT</i>
OCT 30		<b>Seeking Simplicity in Search User Interfaces</b> Marti Hearst <i>Professor, UC Berkeley</i>

# Software tools, environments, and support

*Flagship activity: Establish an “incubator” seed grant program*

- “Incubator” program
  - Our experiment at achieving scalability
  - A lightweight 2-page proposal process several times each year
    - I have an interesting science problem
    - I’m stumped by the data science aspects
    - If you cracked it, others would benefit
    - I’m going to send you the following person half-time for 3-6 months to provide the labor; you provide the guidance
  - Preceded by an information session to clarify expectations and commitments
  - Activities take place in the Data Science Studio, staffed by our Data Scientists
  - We coach software hygiene as well as methodology
  - Running two cohorts annually

- Drop-in “Office Hours”
- Specific broadly applicable tools – democratize access to big data and big data infrastructure



- SQLShare: Database-as-a-Service for scientists and engineers



- Myria: Easy Scalable-Analytics-as-a-Service with database DNA

# Reproducibility and open science

*Flagship activity: Establish a campus-wide community around reproducible research*

- UW campus wide monthly meetings
- May 2014 Workshop
  - More than 80 participants
  - Attendees from NYU, Berkeley, Fred Hutchinson Cancer Research Center, Allen Institute for Brain Science, Sage Bionetworks, Google, ...
  - Report available: <http://uwescience.github.io/reproducible/>
- Draft guidelines for reproducible research
  - <http://uwescience.github.io/reproducible/>
- Weekly tutorials on “research hygiene” topics
  - E.g. GitHub, KnitR, iPython Notebook





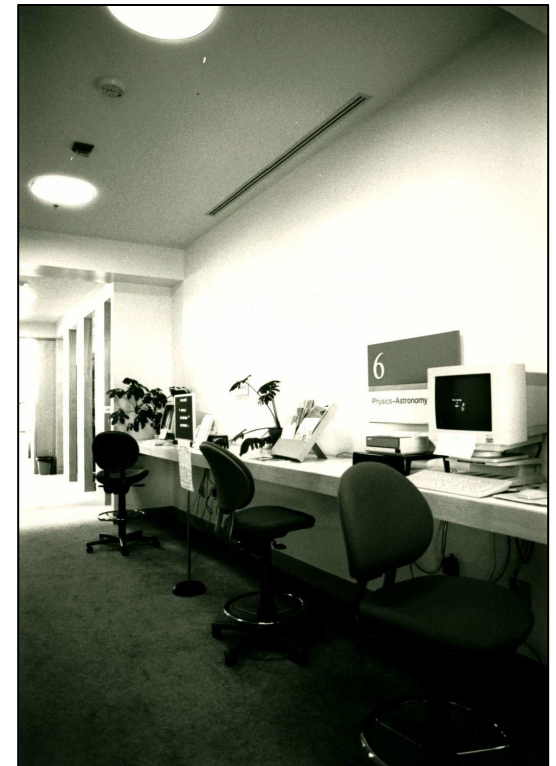
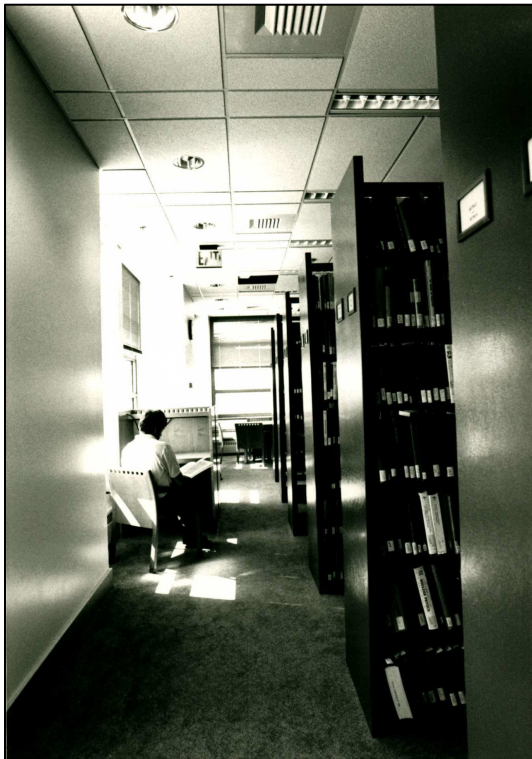
# Working spaces and culture

*Flagship activity: Establish a “Data Science Studio”*

- WRF Data Science Studio



- Before (the Physics/Astronomy Reading Room)





- After





## Ethnography and evaluation

*Flagship activity: Establish a research program in “the data science of data science”*

- Have integrated ethnography and evaluation into a wide range of Data Science Environment activities
  - Project overall (beginning with in-depth baseline interviews with participants from grad students through faculty)
  - IGERT
  - AstroData Hack Week
  - Incubator projects
- Developed ethnography research questions
  - E.g., who does data science, how are they networked, forms of social interaction and organization, intellectual groupings, career reward structures, collaborative tool use in scientific workflows, data science values and ethics, etc.
- Established baseline for evaluation, and determined evaluation questions

## Similarly at NYU and UC Berkeley

- Pursuing the same goals
- Exploring a variety of approaches
- *Interacting extensively*
  - Bi-weekly one-hour teleconferences of the universities' project leadership teams and Foundation staff
  - Frequent interaction among each Working Group's members from the three universities
  - Joint events (AstroData Hack Week, Moore/Sloan Data Science Summit, others)
  - Visits
  - Open sharing of successes and – importantly – failures
  - Wormhole (always-on videoconferencing link) – coming soon

GORDON AND BETTY  
**MOORE**  
FOUNDATION



**NYU**

**Berkeley**  
UNIVERSITY OF CALIFORNIA



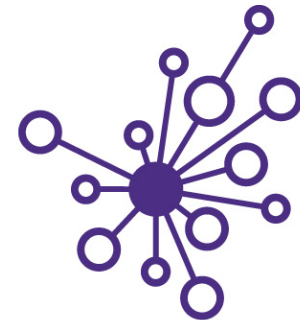
UNIVERSITY *of* WASHINGTON



<http://cds.nyu.edu/>



<http://bids.berkeley.edu/>



UNIVERSITY *of* WASHINGTON  
**eScience Institute**

<http://escience.washington.edu/>

<http://lazowska.cs.washington.edu/AAAS.pdf>, pptx