Harnessing the Potential of Data Scientists and Big Data for Scientific Discovery

Ed Lazowska, University of Washington

Saul Perlmutter, UC Berkeley

Yann LeCun, New York University

Josh Greenberg, Alfred P. Sloan Foundation

Chris Mentzel, Gordon and Betty Moore Foundation

November 12, 2013

NEW YORK UNIVERSITY

Berkeley
UNIVERSITY OF CALIFORNIA

W UNIVERSITY of WASHINGTON

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
1934
ALFRED P. SLOAN FOUNDATION

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

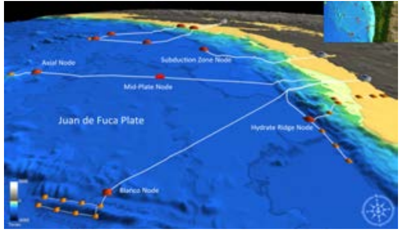
- A proliferation of sensors
- The creation of almost all information in digital form
- Dramatic cost reductions in storage
- Dramatic increases in network bandwidth
- Dramatic cost reductions and scalability improvements in computation
- Dramatic algorithmic breakthroughs in areas such as machine learning

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

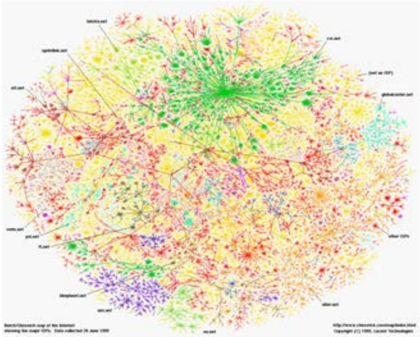
Astronomy: LSST

Physics: LHC

OCEAN OBSERVATORIES INITIATIVE



Oceanography: OOI



Sociology: The Web

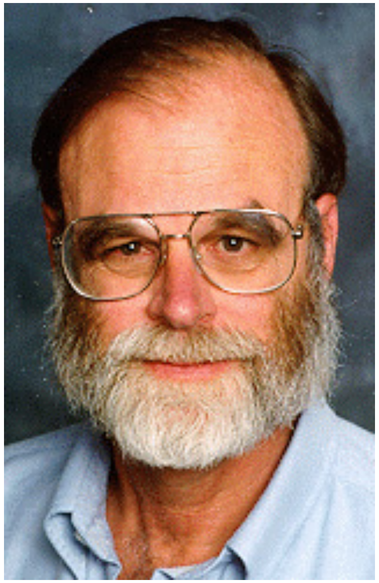
Biology: Sequencing

Economics: POS terminals

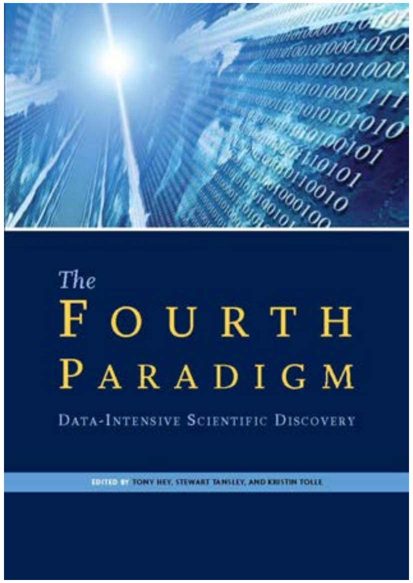
Neuroscience: EEG, fMRI

The Fourth Paradigm

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



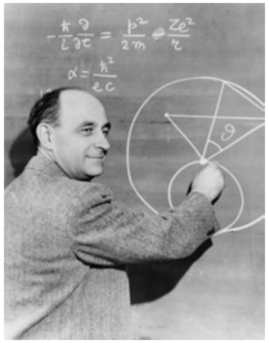
Jim Gray



The Fourth Paradigm

Data-Intensive Scientific Discovery

Edited by: Tony Hey, Stewart Tansley, and Kristin Tolle



“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 scientific discovery
- Data science is driven more by *intellectual infrastructure* (human capital) and *software infrastructure* (shared tools and services – digital capital) than by hardware

Capturing the full potential of the data-rich world has become a daunting challenge

- Both for data scientists, and for those who use data science to advance their research
- Success ultimately will belong to the individuals, teams, and institutions that effectively integrate domain expertise with computational, statistical and mathematical skills

Data Science for academic scientists: What's still needed?

1. We need to be creative in providing **long term career paths** for crucial members of our science teams who become engaged in the data science side of the work, but could be seen as neither conventional (tenurable) faculty in the science domain, nor in methodology domains
2. We should make it easy for scientists to find students (undergraduate and graduate) and post-docs with sufficient **data science training** to quickly come up to speed in research – and write reasonably well-written, de-bug-able, maintainable code when they do

DS for academic scientists: What's still needed?

We must **stand on each other's shoulders, not on each other's toes**. 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 can and must do better!

3. We must make it **easy to find the best code/algorithm/ approach/tutorial** for a current purpose (or to determine that nothing yet exists) within your own discipline, and even possible to find them if they exists in another discipline, despite the likely language barriers
4. We also must make it **easy to contribute and maintain code** that could be useful for a larger community (particularly if it is to dovetail with other projects), and to generate/maintain reproducible results, with ever-changing libraries, hardware, etc.

DS for academic scientists: What's still needed?

5. Our programming languages and **programming environments should not distract from the science** – they should be designed to keep you focused on the question you should be focused on
6. Let's find ways to take advantage of the possibility of using data science as a bridge between disciplines and a magnet for **in-person human interaction** – the data science work we are doing does not need to be isolated or isolating, even though so much of it takes place in the on-line world
7. We must **bridge the current gaps between the needs of domain scientists and the contributions of data science methodologists**, so that the process of discovery is truly transformed

DS for academic scientists: What's still needed?

8. Let's address the challenges we have described in ways that **remove/reduce barriers for those who are less data-science savvy** than those in this room. This potentially could bring in researchers from many disciplines. Moreover, perhaps we can regain some of the lost ground in making this an attractive entryway to the sciences for women and underrepresented minorities
9. These are just opportunities that we have noticed and identified through introspection – we may be missing some of the (potentially more important) hidden problems. We need to study this rigorously – that's what **ethnography** is for – not just today, but **continually** as the inter-relationship between data science and the sciences evolves

This initiative

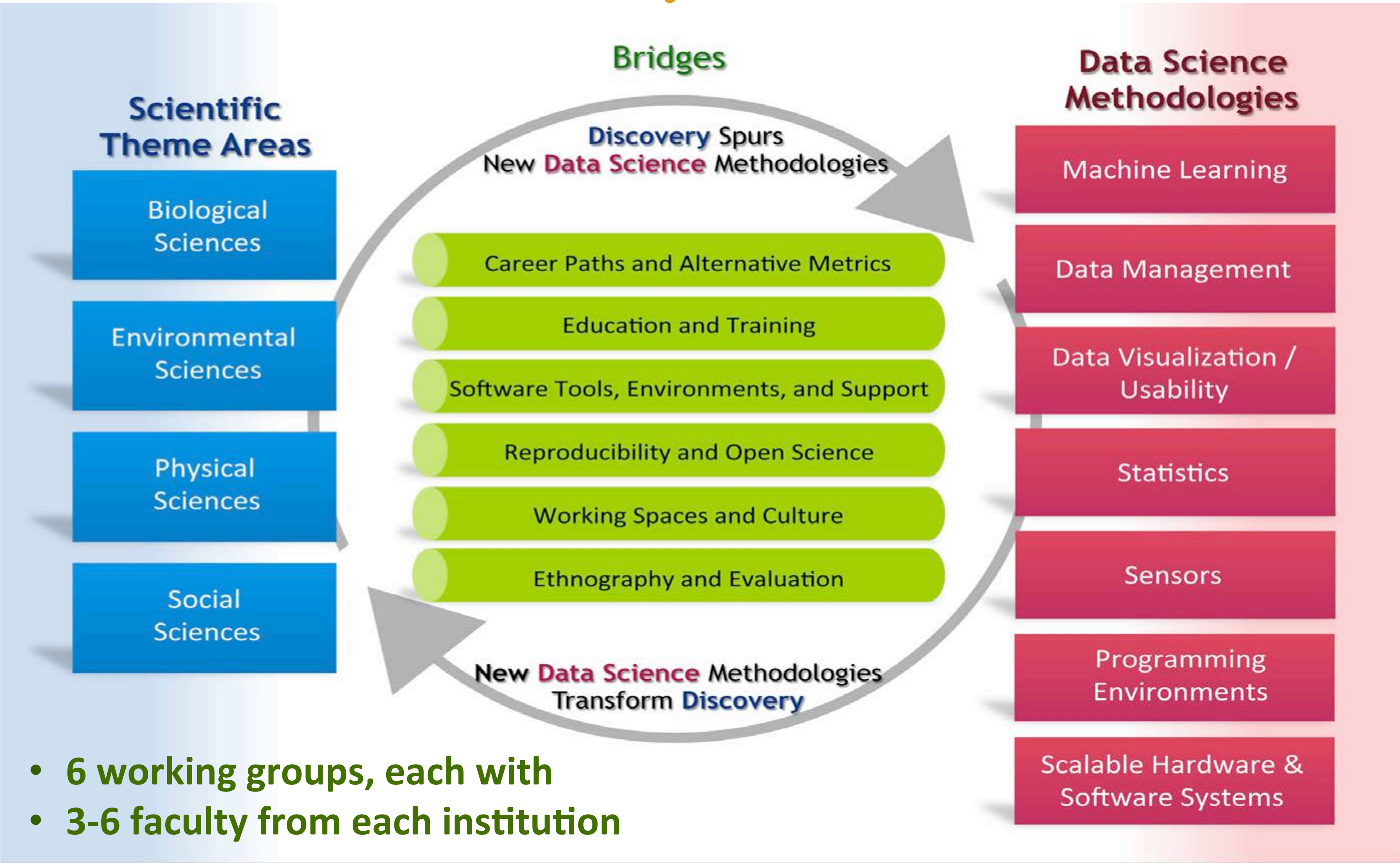
- **Deep collaborations between top researchers in Science and Methods**
  - Physical science, environmental science, life science, neural science, social science
  - Computer science, statistics, applied mathematics
- At each of our institutions we have experienced **the power of data science to transform discovery**
- But there are **significant challenges to realizing the potential of data-intensive discovery**
- This initiative focuses on approaches to **address these challenges, allowing data science to truly flourish**

Core 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 new generation of multi-disciplinary scientists in data-intensive science
  - A new 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

Establish a virtuous cycle


Bridges
Scientific Theme Areas
Biological Sciences
Environmental Sciences
Physical Sciences
Social Sciences
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

- **6 working groups, each with**
- **3-6 faculty from each institution**

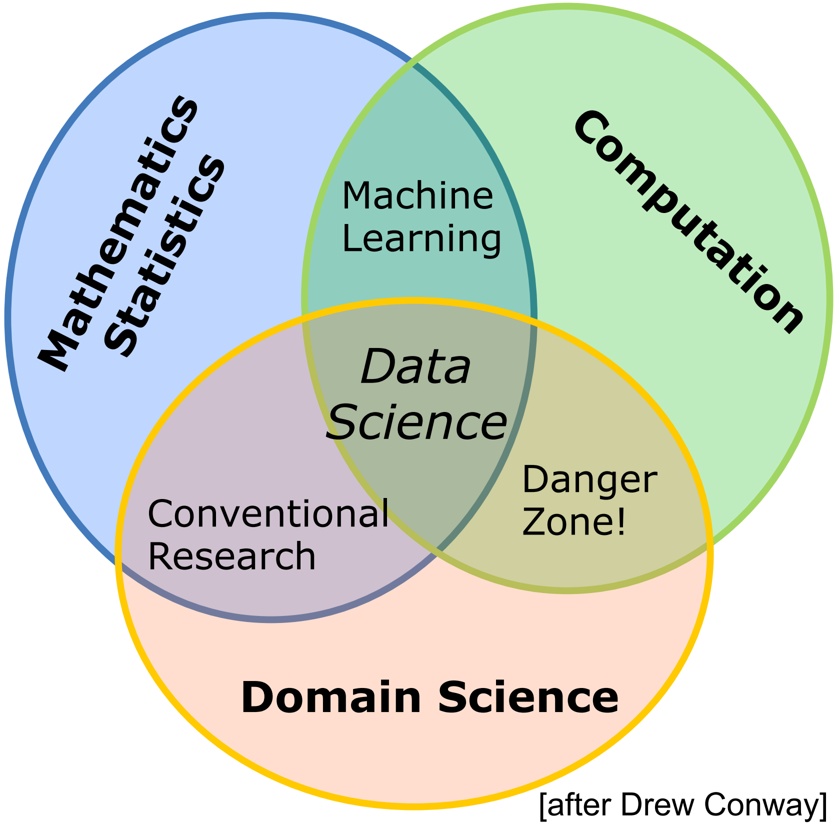
Career Paths and Alternative Metrics

- **Create and sustain long-term career trajectories for:**
  - Scientists focused on computer-assisted analysis of massive/noisy/complex data
  - Researchers focused on creating tools used by others to derive new science
- **These career paths are complicated by:**
  - Traditional measures of scientific impact (e.g., publications) that don't reflect the full contribution (e.g., software tools)
  - Strong competition from industry
  - Difficulty of providing a supportive environment and culture

Example approach: We will establish "Data Scientist" career tracks and "Data Science Fellow" positions at our institutions

Education and Training

- **Training in data science:** undergraduate, M.S., Ph.D., postdocs, research staff, and faculty
- **Methods:** computer science, statistics, applied mathematics, machine learning, signal processing
- **Tools:** databases, data management, scientific computing, information visualization
- **Domains:** Data Science in X



[after Drew Conway]

Example approach: M.S. and Ph.D. programs in Data Science, undergraduate course sequences, boot camps and summer schools

Software Tools, Environments, and Support

- **Software environments and tools are crucial**
  - Organic, sustainable, reusable, extensible
  - Easy to translate across problem domains
- The creation and usage of today's tools and software environments are distracting from the science
- **Today's academic environments do not reward tool builders**
- How can the development, hardening, sustaining, sharing, and integration of techniques into a reusable software infrastructure be recognized and incentivized?

Example approach: Teams of software architects, engineers, and researchers who will produce data science tools and will be evaluated on the impact of these tools

Reproducibility and Open Science

- Maximizing the ability to build upon each other's work accelerates the rate of progress of science
- In data-intensive scientific discovery, this means:
  - Sharing software tools, data and practices
  - Developing tools that support the sharing, preservation, provenance, and reproducibility of data, and scientific workflows
- We have the opportunity to spend more time standing on each other's shoulders and less time standing on each other's toes

Example approach: Development of "best practices" for reproducibility and of tools that support these best practices; advocacy of open source; exploration of a certification process

Working Spaces and Culture

- **Physical and virtual environments are essential**
  - The "water cooler" – make sure people run into each other
  - The "video wormhole" – even if they are far away
- **Data science will be "the great unifier"**
- **We must:**
  - Facilitate sustained collaborations that cross disciplinary boundaries
  - Reach the broader constituency and increase diversity

Example approach: Co-locate domain scientists, data science methodology researchers, and tool-builders: NYU "Center for Data Science," Berkeley "Institute for Data Science," UW "Data Science Studio"

Ethnography and Evaluation

- **The world of data science is changing rapidly**! How do we measure our impact on the scientific community, not just with science but with
  - New methodologies
  - New tools
  - New organizational models
  - New models for knowledge dissemination
- We need to understand the complexity and evolution of the landscape in a quantitative way ("data science for data science")

Example approach: Ethnography teams will study the complex ecology of Data Science in different problem domains, to understand the diverse training and support needs, and to identify opportunities for high-impact interventions

We are at the dawn of a revolutionary new era of discovery and learning

- Soon, much of the knowledge in the world will be extracted from data by machines
- Data-intensive discovery will help us answer the big scientific questions of our time …
  - What is the Universe made of?
  - What are the mechanisms of life?
  - How does the brain work?
  - How can we understand human behavior and societies?

We are at the dawn of a revolutionary new era of discovery and learning

NYU

Berkeley
UNIVERSITY OF CALIFORNIA

W
UNIVERSITY *of* WASHINGTON

GORDON AND BETTY
MOORE
FOUNDATION

ALFRED P. SLOAN FOUNDATION
1934

ALFRED P. SLOAN
FOUNDATION