W UNIVERSITY of WASHINGTON

UNIVERSITY of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

Data Science @ UW

GORDON AND BETTY MOORE FOUNDATION

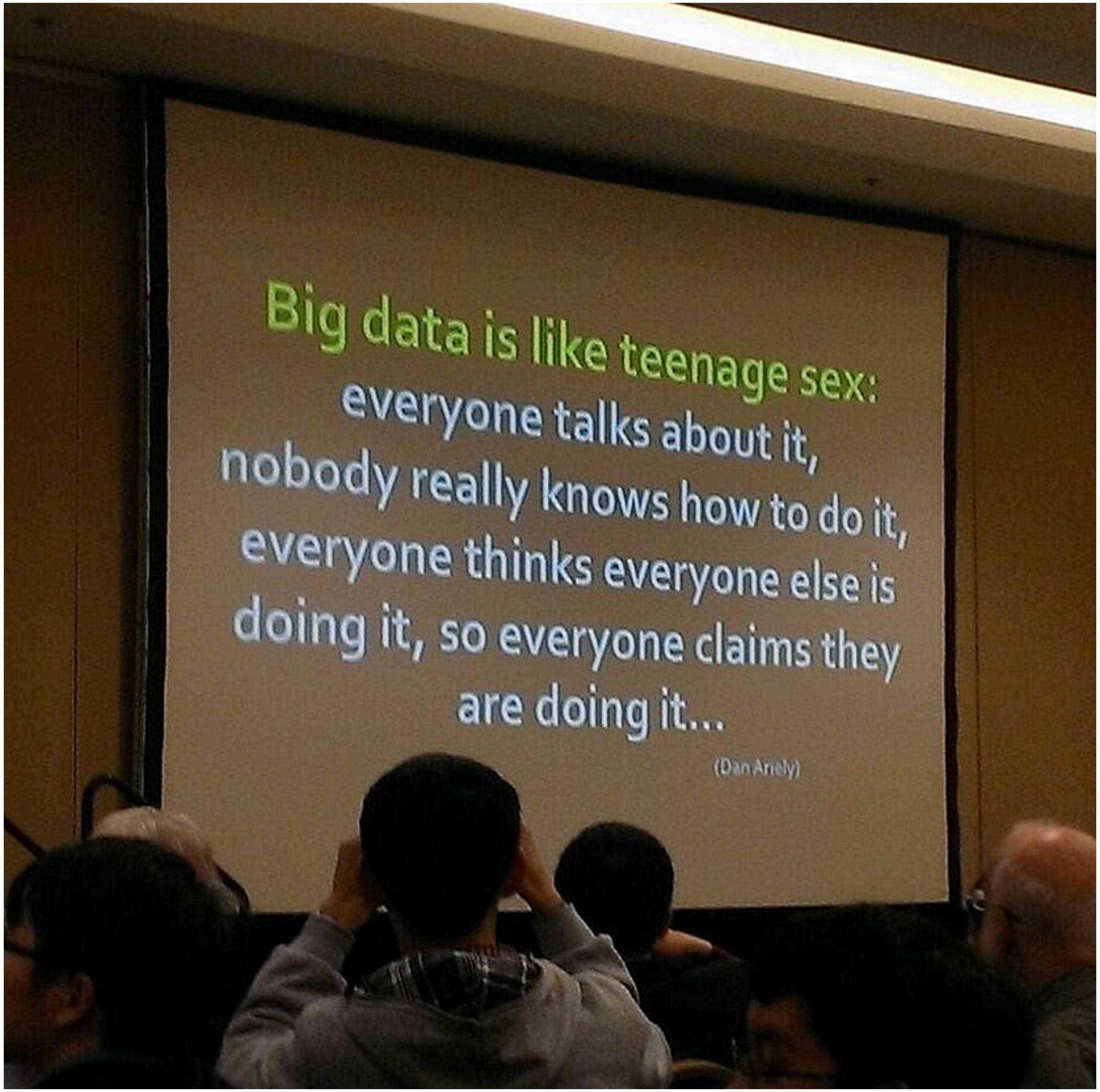
ALFRED P. SLOAN FOUNDATION 1934

Washington Research FOUNDATION

NSF

W UNIVERSITY of WASHINGTON

What is data science?



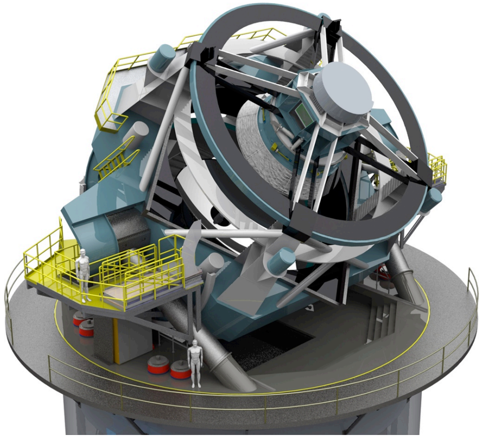
Big data is like teenage sex:
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is
doing it, so everyone claims they
are doing it...

(Dan Ariely)

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

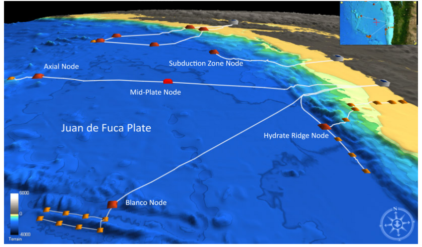
- A proliferation of sensors
- Ever more powerful models producing data that must be analyzed
- The creation of almost all information in digital form
- Dramatic cost reductions in storage
- Dramatic increases in network bandwidth
- Dramatic cost reductions and scalability improvements in computation
- Dramatic algorithmic breakthroughs in areas such as machine learning

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

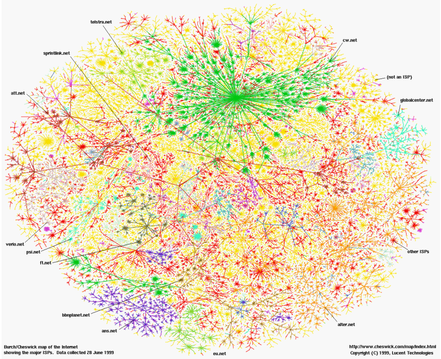


Astronomy: LSST

Physics: LHC



Oceanography: OOI



Sociology: The Web

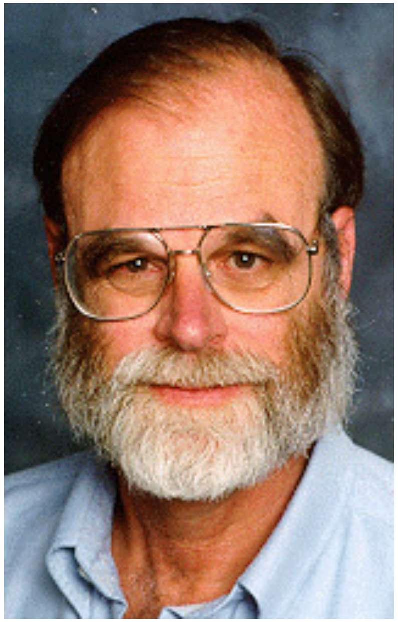
Biology: Sequencing

Economics: POS terminals

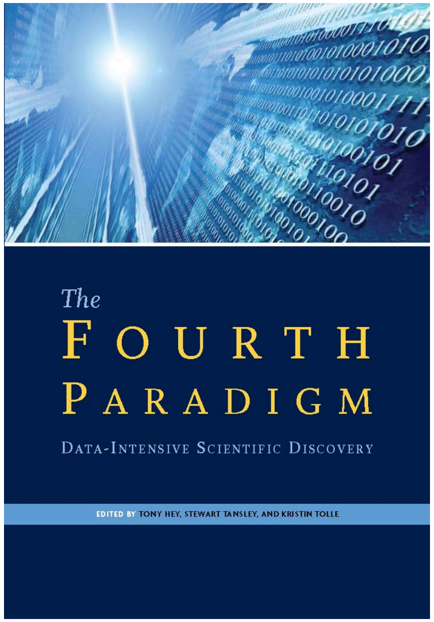
Neuroscience: EEG, fMRI

The Fourth Paradigm

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



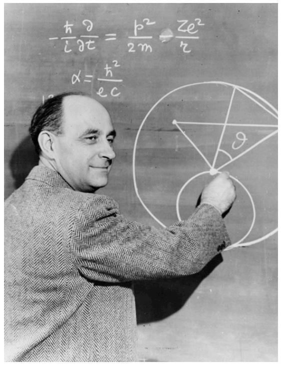
Jim Gray

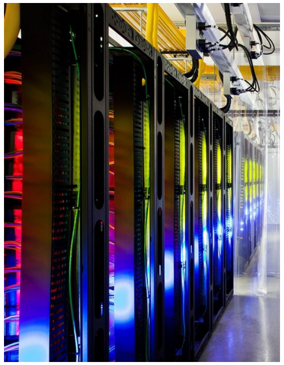


The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE





Each augments, vs. supplants, its predecessors – "another arrow in the quiver"

"From data to knowledge to action"

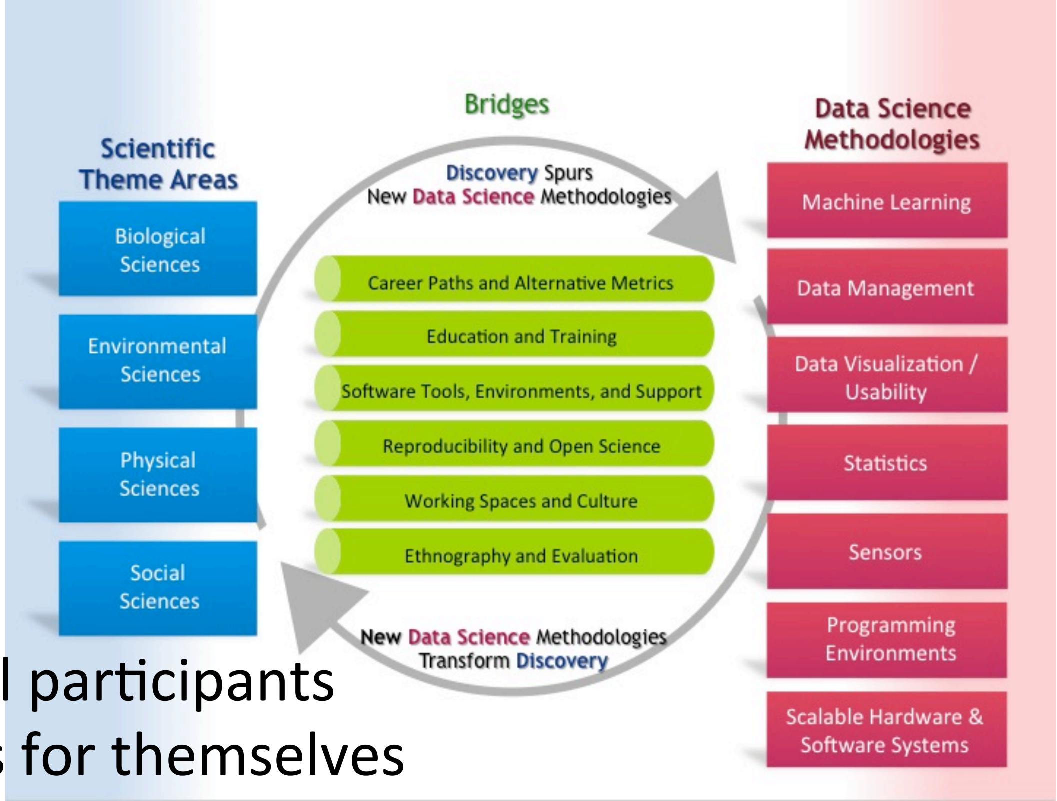
- The ability to extract knowledge from large, heterogeneous, noisy datasets – to move "from data to knowledge to action" – lies at the heart of 21st century discovery
- To remain at the forefront, researchers *in all fields* will need access to state-of-the-art data science methodologies and tools
- 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

Major sources of funding 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)

Over-arching objective

- Work with our Berkeley, NYU, and Foundation partners to carry out a distributed collaborative experiment in creating university environments in which data-intensive discovery flourishes
  - Doing research
    - Methodology areas
    - Scientific theme areas
  - Enabling research
    - Career paths
    - Education and training
    - Tools
    - Reproducible research
    - Working spaces & culture
    - Ethnography


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

- While the balance varies, all participants are in this for UW as well as for themselves

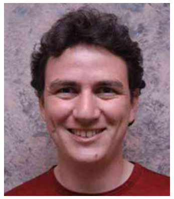
People: Original core faculty team

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

Randy LeVeque
Applied Mathematics

Thomas Richardson
Statistics

Werner Stuetzle
Statistics

Biological sciences

David Beck
Chemical Engr.

Tom Daniel
Biology

Bill Noble
Genome Sciences

Environmental sciences

Ginger Armbrust
Oceanography

Social sciences

Josh Blumenstock
iSchool

Mark Ellis
Geography

Tyler McCormick
Sociology, CSSS

Physical sciences

Andy Connolly
Astronomy

John Vidale
Earth & Space Sciences

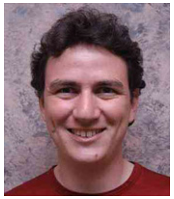
People: Original core faculty team

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

Randy LeVeque
Applied Mathematics

Thomas Richardson
Statistics

13 Departments

5 Schools / Colleges

Biological sciences

David Beck
Chemical Engr.

Tom Daniel
Biology

Bill Noble
Genome Sciences

Environmental sciences

Ginger Armbrust
Oceanography

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

People: Current participants

- 9-person Executive Committee
- 24-person Steering Committee
- 33 Data Science Fellows (faculty and research staff who are "all in")
- 73 Affiliates
- An outstanding, expanding staff
- Provost's Initiative hires
- Postdocs
- IGERT Ph.D. students

People: Research staff

- Director, Associate Director

Ed Lazowska

Bill Howe

- Co-Program Managers

Micaela Parker
Ph.D., Oceanography

Sarah Stone
Ph.D., Oceanography

- Data scientists / research scientists / research faculty

Dave Beck
Director of Research, Life Sciences
Ph.D. Medicinal Chemistry, Biomolecular Structure & Design

Dan Halperin
Director of Research, Scalable Data Analytics
Ph.D., Computer Science

Joe Hellerstein
Senior Data Science Fellow
IBM Research, Microsoft Research, Google (ret.)

Jake VanderPlas
Director of Research, Physical Sciences
Ph.D., Astronomy

Andrew Whitaker
Data Science Fellow
Ph.D., Computer Science

  - Arriving this spring …

Ariel Rokem
Data Scientist
Ph.D., Neuroscience

Valentina Staneva
Data Scientist
Ph.D., Applied Mathematics and Statistics

People: First-year Provost's Initiative hires

- Bing Brunton, Biology
- Steve Brunton, Mechanical Engineering
- Mario Juric, Astronomy
- Emilio Zagheni, Sociology

People: First-year Postdocs

- Rahul Biswas, cosmology
  - Mentors: Andy Connolly (Astronomy), Magda Balazinska (CSE)
- Thiago Costa, computational social sciences
  - Mentors: Tyler McCormick (Statistics), Josh Blumenstock (iSchool)
- Brittany Fiore-Silfvast, ethnography
  - Mentors: Cecilia Aragon (HCDE), Gina Neff (Communication)
- Jie Liu, computational genetics
  - Mentors: Bill Noble (Genome Sciences), Jeff Bilmes (EE)
- Allison Smith, oceanography
  - Mentors: Curtis Deutsch (Oceanography), Jeff Heer (CSE)
- Dave Williams, biophysics
  - Mentors: Tom Daniel (Biology), Magda Balazinska (CSE)

People: First-year IGERT Ph.D. students

- Will Gagne-Maynard
  - Oceanography & Microsoft Research
- Ryan Maas
  - Astronomy & CSE
- Matt Murbach
  - Chemical Engineering & machine learning
- Cecilia Noecker
  - Genome Sciences & machine learning
- Alex Tank
  - Statistics & Allen Institute for Brain Science
- Grace Telford
  - Astronomy & Statistics

Education and training

Flagship activity: Establish a new graduate program 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
    - Centered around IGERT students (required to attend)
    - Moore/Sloan postdocs also are expected to attend. Others encouraged
    - Seminar topics include reproducibility, ethics, science, etc.
  - Put in place detailed program evaluation plan with Data2Insight
  - First cohort of 6 students from a variety of departments
    - All students have co-advisors in methods and science
    - Some have co-advisors in research labs or industry

Education and training (cont'd)

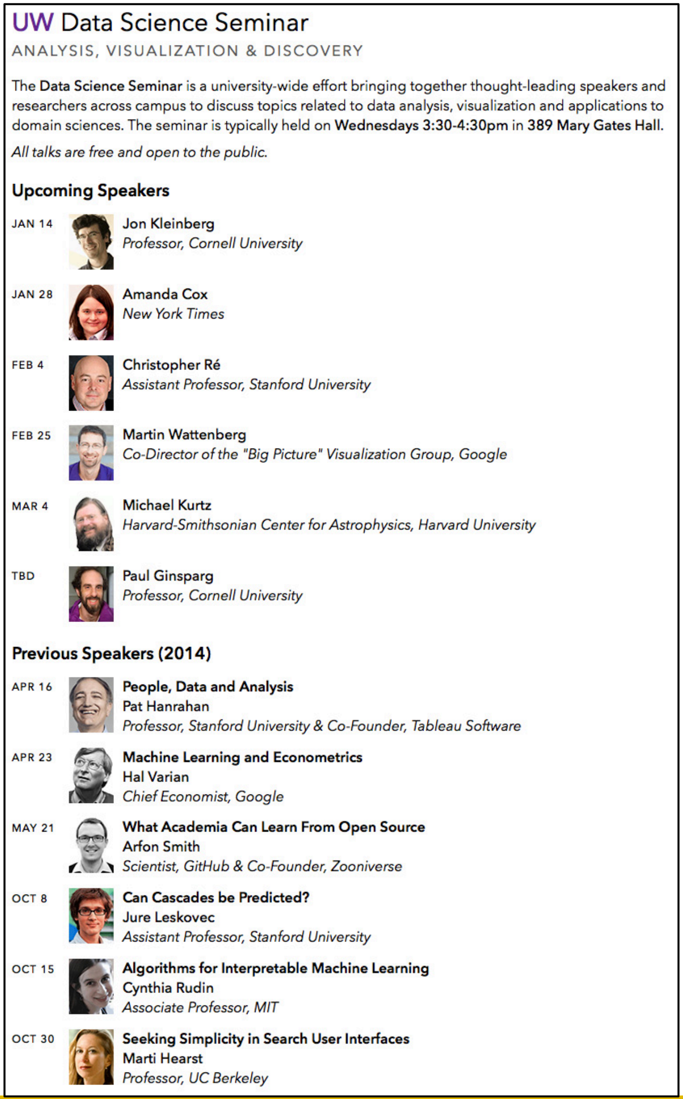
Flagship activity: Establish a new graduate program in data science

- Workshops and Boot Camps
  - Software Carpentry Bootcamp (Jake VanderPlas, March 17-18 2014)
  - Community Data Science Workshops (Benjamin Mako Hill, 3 Saturdays in April and May 2014)
  - Astro Hack Week (Jake VanderPlas, Sept 15-19 2014)
  - ASTR 599/AMATH 500 boot camp (Jake VanderPlas, Sept 22-23, 2014)
  - Software Carpentry Instructor Course (Ben Marwick, Nov. 12-14, 2014)
  - UW Libraries Scholars' Studio (quarterly, 2014-2015)

Education and training (cont'd)

Flagship activity: Establish a new graduate program in data science

- 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)
- Self-sustaining Masters in Data Science under active development
- Education working group is actively tracking *all* curricular activities



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** in **389 Mary Gates Hall.**

All talks are free and open to the public.

Upcoming Speakers

JAN 14 Jon Kleinberg
Professor, Cornell University

JAN 28 Amanda Cox
New York Times

FEB 4 Christopher Ré
Assistant Professor, Stanford University

FEB 25 Martin Wattenberg
Co-Director of the "Big Picture" Visualization Group, Google

MAR 4 Michael Kurtz
Harvard-Smithsonian Center for Astrophysics, Harvard University

TBD Paul Ginsparg
Professor, Cornell University

Previous Speakers (2014)

APR 16 **People, Data and Analysis**
Pat Hanrahan
Professor, Stanford University & Co-Founder, Tableau Software

APR 23 **Machine Learning and Econometrics**
Hal Varian
Chief Economist, Google

MAY 21 **What Academia Can Learn From Open Source**
Arfon Smith
Scientist, GitHub & Co-Founder, Zooniverse

OCT 8 **Can Cascades be Predicted?**
Jure Leskovec
Assistant Professor, Stanford University

OCT 15 **Algorithms for Interpretable Machine Learning**
Cynthia Rudin
Associate Professor, MIT

OCT 30 **Seeking Simplicity in Search User Interfaces**
Marti Hearst
Professor, UC Berkeley

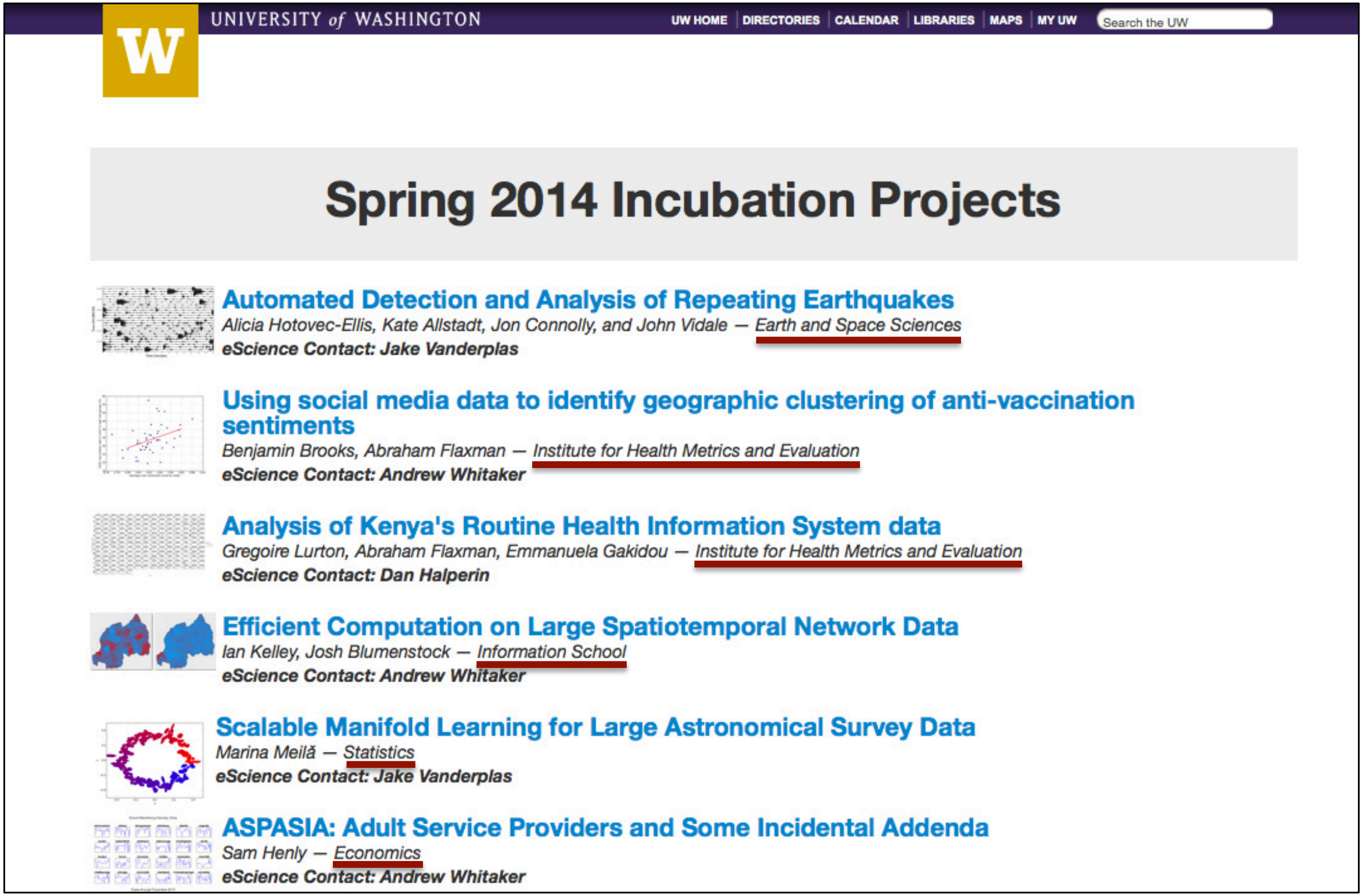
Software tools, environments, and support

Flagship activity: Establish an "incubator" seed grant program

- We began with deep (but not scalable) engagements
  - Survey astronomy
  - Environmental oceanography
- Our experiment at achieving scalability: "Incubator" program
  - 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
  - Ran a cohort of 6 in Spring 2014, and another in Autumn 2014

Software tools, environments, and support (cont'd)

Flagship activity: Establish an "incubator" seed grant 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

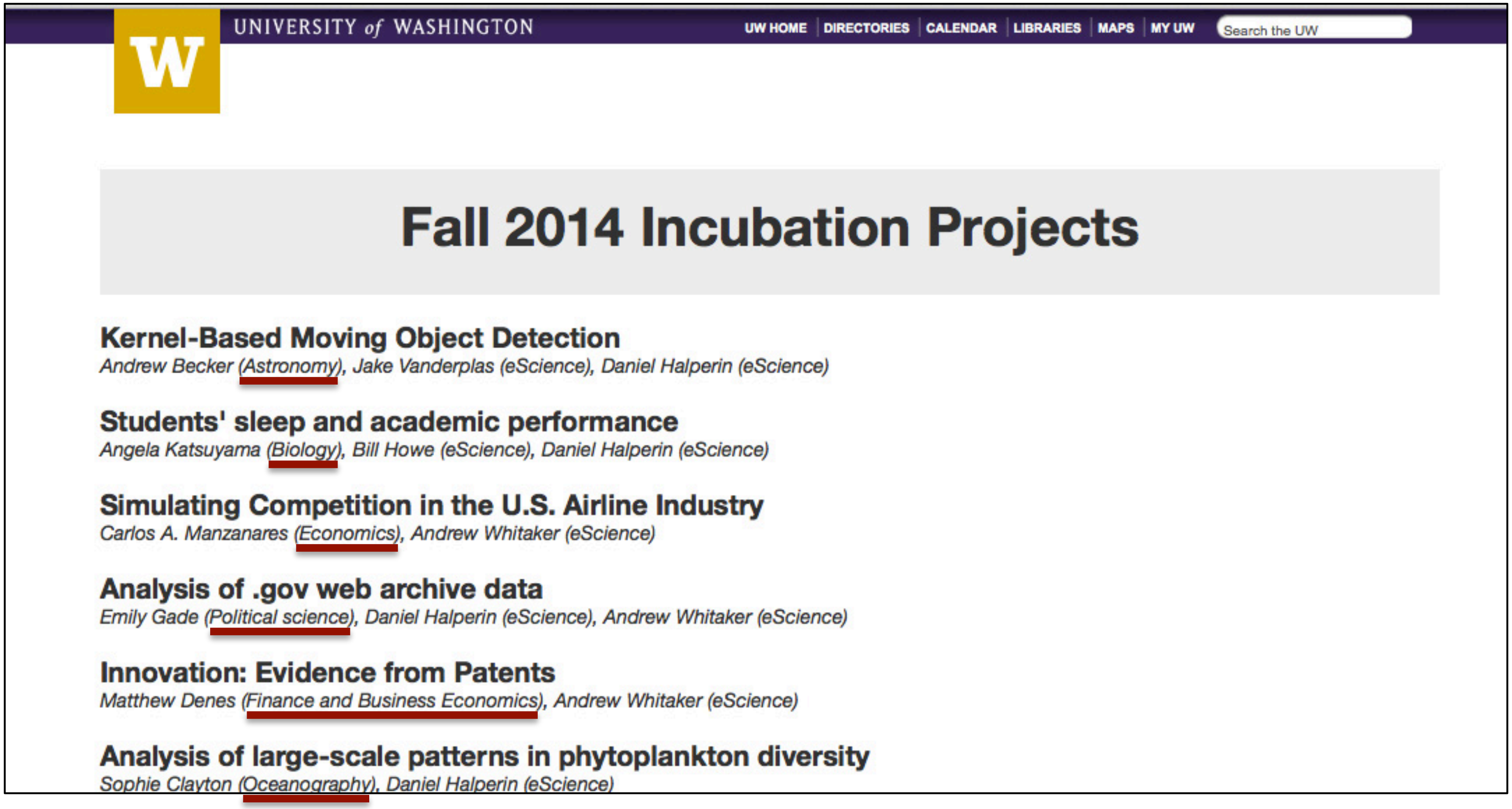
Efficient Computation on Large Spatiotemporal Network Data
Ian Kelley, Josh Blumenstock — Information School
eScience Contact: Andrew Whitaker

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

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

Software tools, environments, and support (cont'd)

Flagship activity: Establish an "incubator" seed grant program



Fall 2014 Incubation Projects

Kernel-Based Moving Object Detection
Andrew Becker (Astronomy), Jake Vanderplas (eScience), Daniel Halperin (eScience)

Students' sleep and academic performance
Angela Katsuyama (Biology), Bill Howe (eScience), Daniel Halperin (eScience)

Simulating Competition in the U.S. Airline Industry
Carlos A. Manzanares (Economics), Andrew Whitaker (eScience)

Analysis of .gov web archive data
Emily Gade (Political science), Daniel Halperin (eScience), Andrew Whitaker (eScience)

Innovation: Evidence from Patents
Matthew Denes (Finance and Business Economics), Andrew Whitaker (eScience)

Analysis of large-scale patterns in phytoplankton diversity
Sophie Clayton (Oceanography), Daniel Halperin (eScience)

Software tools, environments, and support (cont'd)

Flagship activity: Establish an "incubator" seed grant program

- Specific broadly applicable tools – democratize access to big data and big data infrastructure

SQLSHARE

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

Myria

  - 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
  - Average 10 – 15 participants
  - Working group: LeVeque, Beck, Hellerstein, Howe, Wright, & others
- May 2014 Workshop
  - More than 80 participants
  - Participants from NYU, UCB, Fred Hutch CC, Allen Institute for Brain Science, Sage Bionetworks, Google
  - Mix of talks and breakout groups
  - Report available: http://uwescience.github.io/reproducible/
- State of reproducibility on campus part of Ethnography survey

- Draft guidelines for reproducible research
  - http://uwescience.github.io/reproducible/
  - Presented to post-docs and IGERT students; lots of discussion
- Weekly tutorials on “research hygiene” topics
  - E.g. GitHub, KnitR, iPython Notebook
  - To begin when Data Science Studio is online
- Template for recording & categorizing research publications on reproducibility spectrum
- Self-certification & badging of research groups for reproducibility
- Shared web presence between UW, UCB, & NYU in discussion

Working spaces and culture

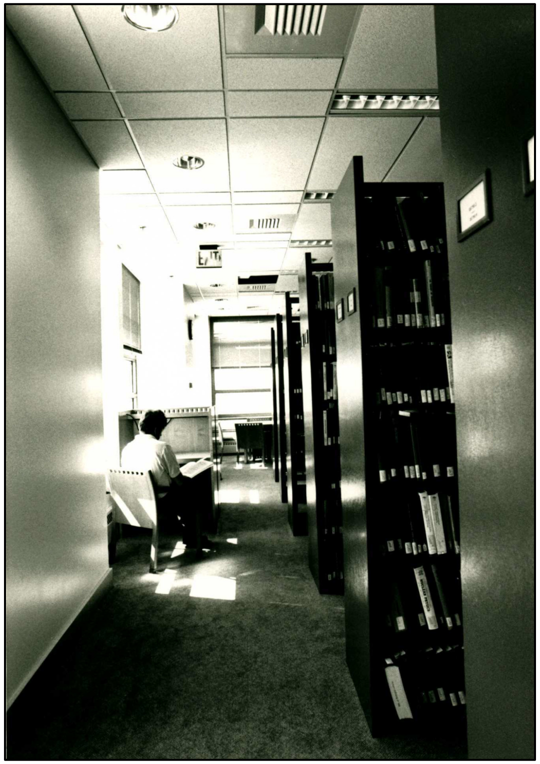
Flagship activity: Establish a "Data Science Studio"

- WRF Data Science Studio – a campus-wide collaboration space

©2008 Eric Haley

Working spaces and culture (cont'd)

Flagship activity: Establish a "Data Science Studio"

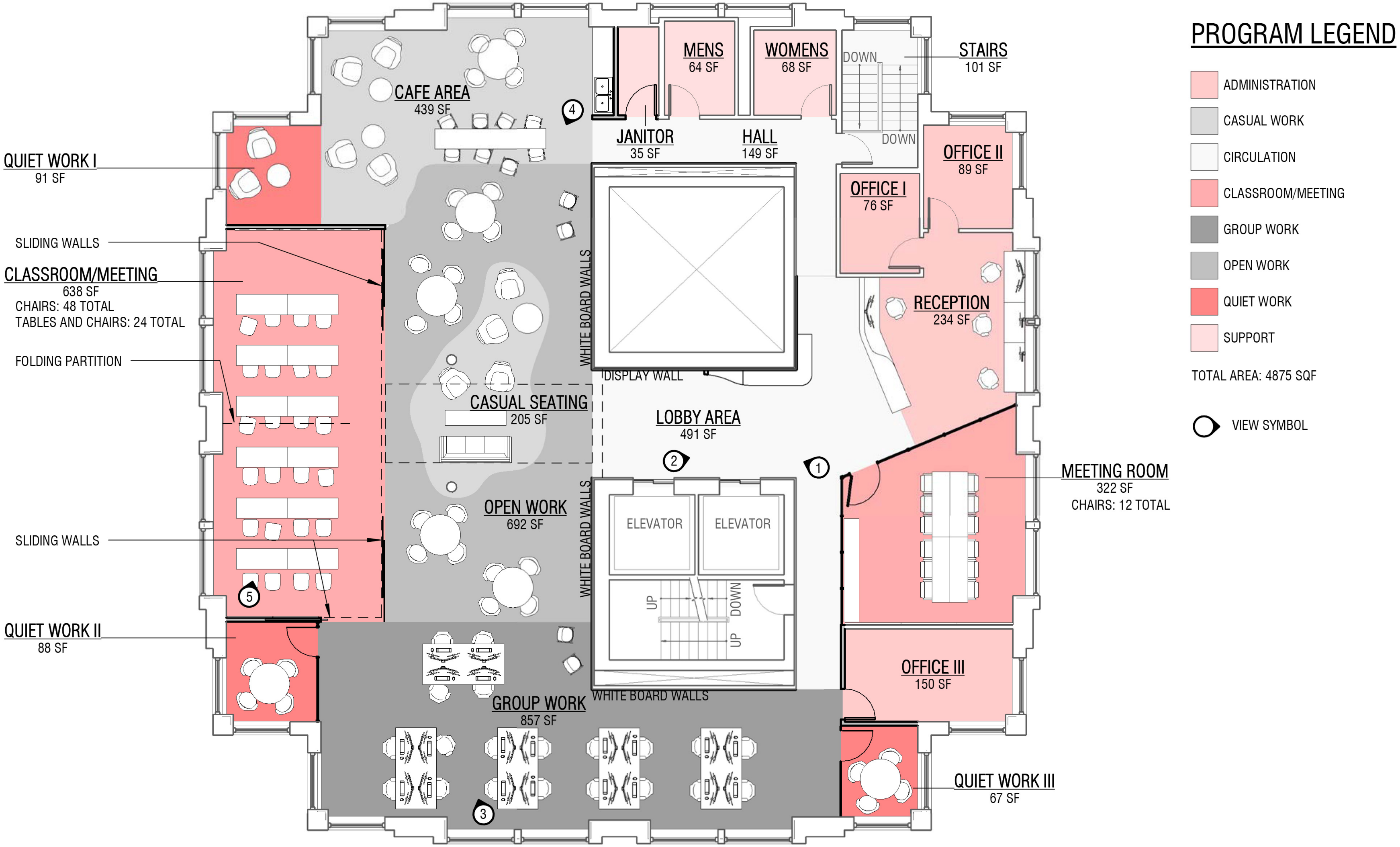


UNIVERSITY of WASHINGTON

UNIVERSITY of WASHINGTON eScience Institute ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

Working spaces and culture (cont'd)

Flagship activity: Establish a "Data Science Studio"

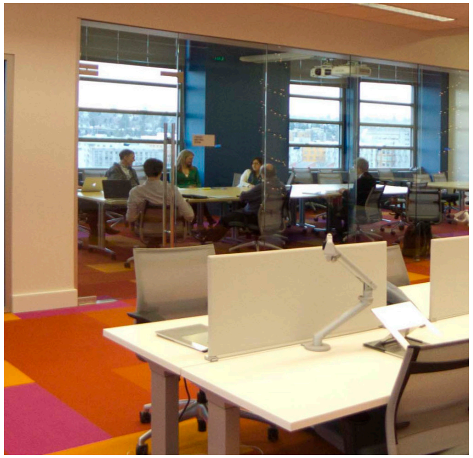


Working spaces and culture (cont'd)

Flagship activity: Establish a "Data Science Studio"

Working spaces and culture (cont'd)

Flagship activity: Establish a "Data Science Studio"



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

http://lazowska.cs.washington.edu/RAB.pdf, pptx