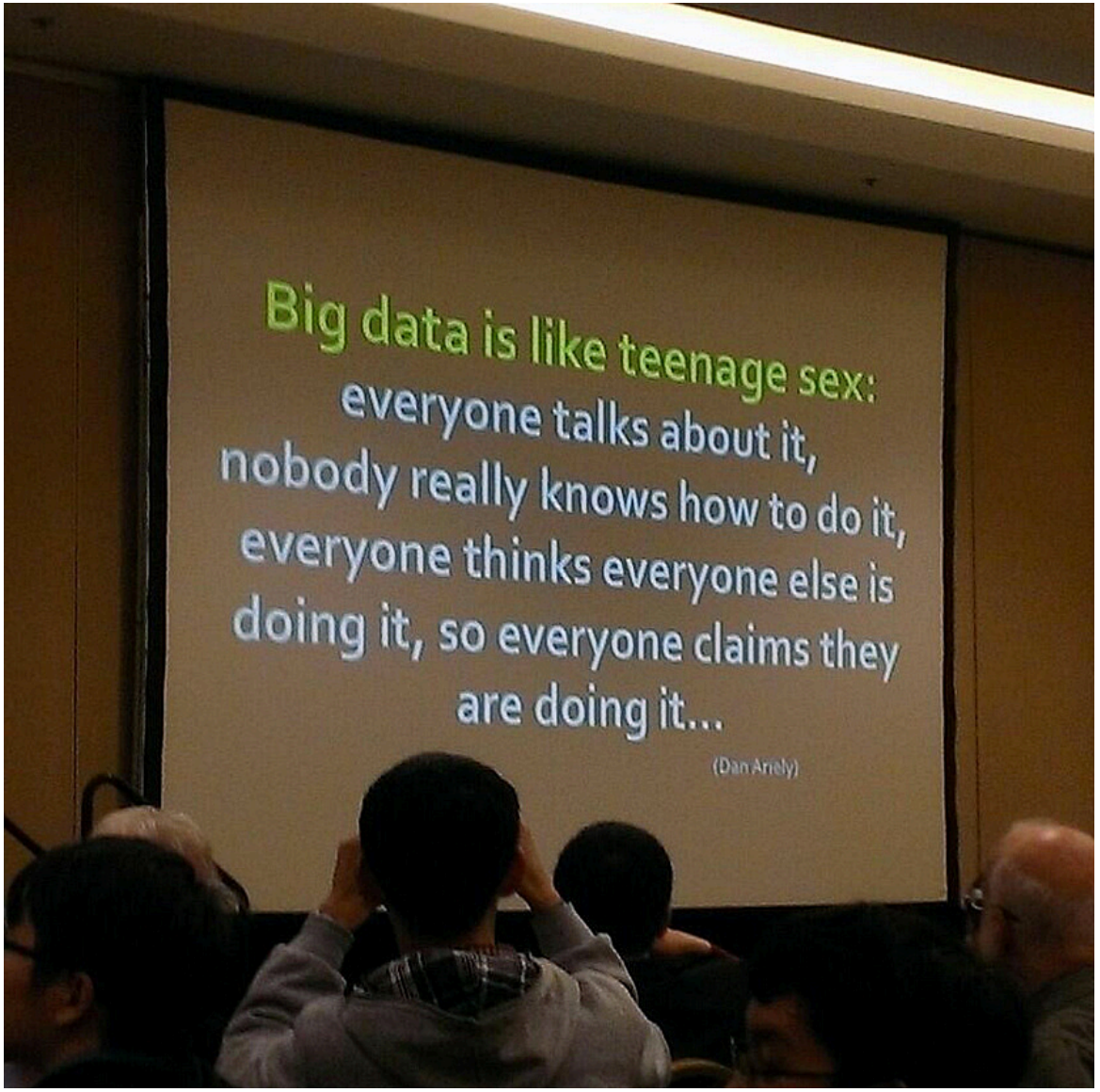
UNIVERSITY of WASHINGTON

Data Science @ UW

This afternoon

- Highlight UW's emergence as a national and international leader in data science – data-intensive discovery
  - Describe a number of related initiatives
  - Talk about how you can engage
- Particularly acknowledge the generosity of the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation
  - A 5-year, $37.8M collaborative effort to create model "data science environments" at NYU, Berkeley, and UW
- Provide an opportunity for all of you – individuals and teams at UW and in the region – to interact
  - Poster and networking event in Mary Gates Commons immediately following these presentations

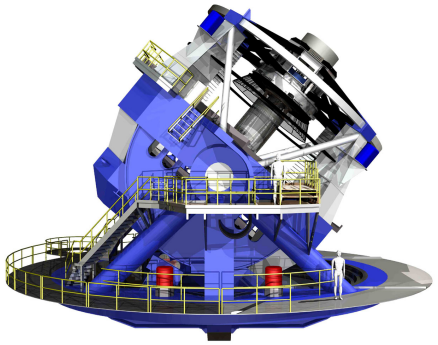
What is Data Science?



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

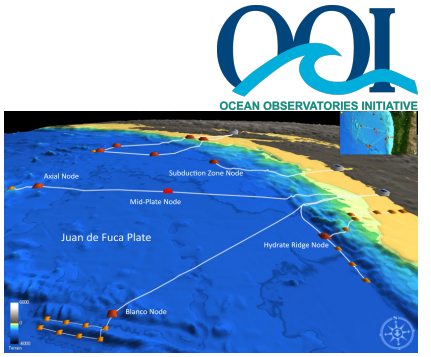
- A proliferation of sensors
- Ever more powerful models producing data that must be analyzed
- The creation of almost all information in digital form
- Dramatic cost reductions in storage
- Dramatic increases in network bandwidth
- Dramatic cost reductions and scalability improvements in computation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

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

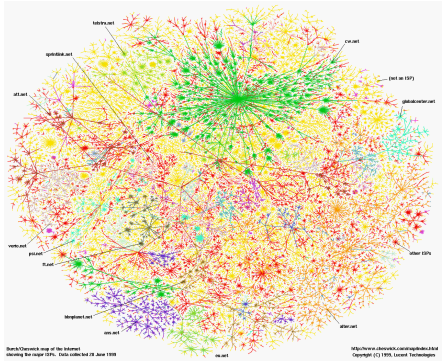


Astronomy: LSST

Physics: LHC



Oceanography: OOI



Sociology: The Web

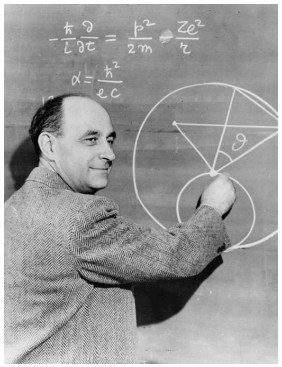
Biology: Sequencing

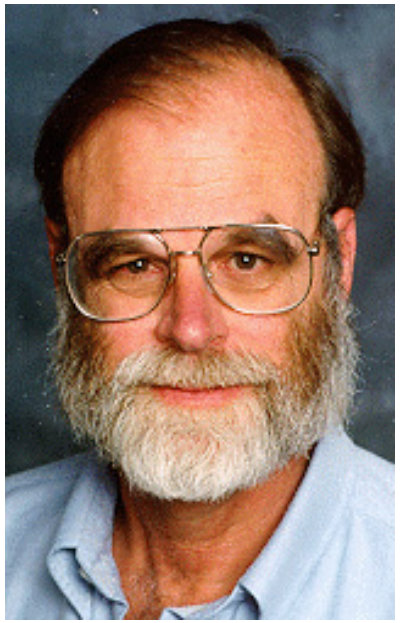
Economics: POS terminals

Neuroscience: EEG, fMRI

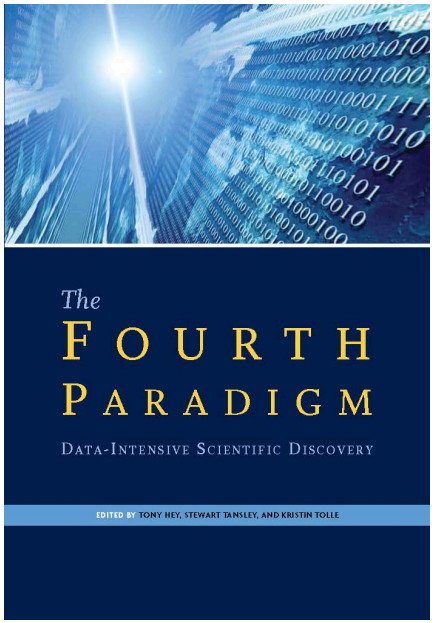
The Fourth Paradigm

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





Jim Gray

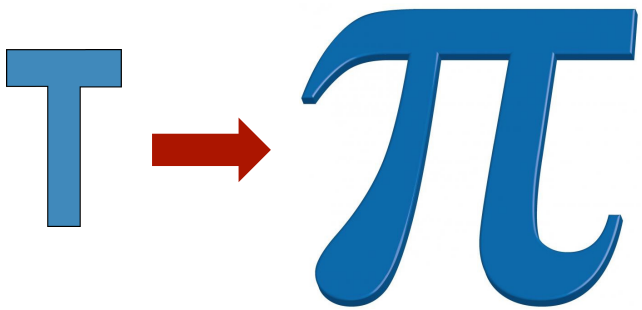


"From data to knowledge to action"

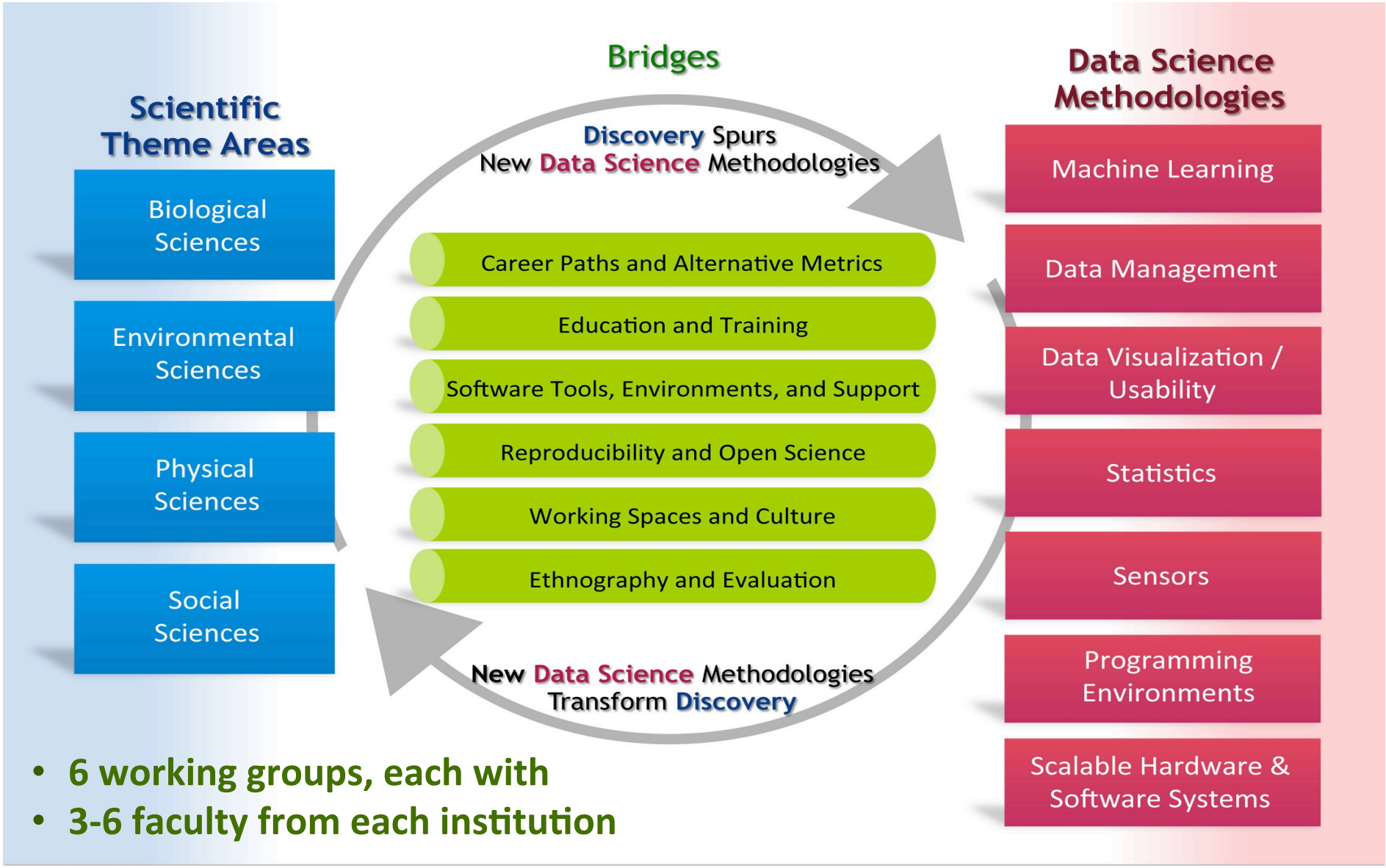
- The ability to extract knowledge from large, heterogeneous, noisy datasets – to move "from data to knowledge to action" – lies at the heart of 21st century discovery
- To remain at the forefront, researchers *in all fields* will need access to state-of-the-art data science methodologies and tools
-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

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

- Both for those who advance the methodologies of data science, 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



Establish a virtuous cycle



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

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

GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION

Berkeley
UNIVERSITY OF CALIFORNIA

UNIVERSITY of WASHINGTON

UW has been striving to support data-intensive discovery since long before it became cool!

- Department of Biostatistics: 1970
  - Ranked #3 among the nation's Statistics programs
- Department of Statistics: 1979
  - Ranked #6 among the nation's Statistics programs
- Center for Statistics and the Social Sciences: 1999
  - Bringing state-of-the-art statistical methodology to the social sciences
- eScience Institute: 2008
  - Genesis in 2005
- Aggressively expanding our faculty – both those who contribute advances in key methodology areas, and those who apply these methodologies in discovery

The goal of everything that you will hear about and see this afternoon

- Ensure that UW is a national and international leader in data-intensive discovery
  - In *advancing the methodologies*
  - In *putting them to work* by the broadest possible range of investigators
  - In *preparing students* for 21st century discovery
- Truly a campus-wide partnership

eScience Institute

Center for Statistics and the Social Sciences

UNIVERSITY of WASHINGTON

Reflected in eScience Institute leadership

- Executive Committee
  - Cecilia Aragon, Human Centered Design & Engineering
  - Ginger Armbrust, Oceanography
  - Magda Balazinska, Computer Science & Engineering
  - Andy Connolly, Astronomy
  - Tom Daniel, Biology
  - Bill Howe, Computer Science & Engineering (Associate Director)
  - Ed Lazowska, Computer Science & Engineering (Director)
  - Randy LeVeque, Applied Mathematics
  - Tyler McCormick, Statistics + Sociology

8 units, 3 Schools & Colleges

- Additional members of the Steering Committee
  - David Beck, research scientist liaison
  - Josh Blumenstock, Information School
  - Mark Ellis, Geography
  - Emily Fox, Statistics
  - Terry Gray, UW Information Technology liaison
  - Carlos Guestrin, Computer Science & Engineering
  - Dan Halperin (data scientist liaison)
  - Thomas Richardson, Statistics + CSSS
  - Mani Soma, Office of Research liaison
  - Kari Stephens, Psychiatry & Behavioral Sciences
  - Werner Stuetzle, Statistics + A&S
  - Jake Vanderplas, postdoc liaison
  - John Vidale, Earth & Space Sciences

9 additional units, 2 additional Schools & Colleges

More than 50 affiliated faculty

During the rest of this hour you'll learn about …

- A 5-year $2.8M NSF IGERT award
  - Supports the creation of an interdisciplinary graduate program in data science
- The Provost's Initiative in Data-Intensive Discovery
  - 50% support to incent the hiring of faculty in any field whose scholarship has a significant data science methodology component
    - 50% obligation to broad-interest teaching and engagement
- The establishment of a campus-wide "Data Science Studio"
  - A center for collaborative activities
  - The heart of an "Incubation" program in which we invite your participation

And a number of other activities that I'll simply mention

- Data Science Lunches
  - Data science is the "unifier and connector"
  - E.g., "Big Social Data Researchers" (2 lunches), "Text Analytics," "Urban Science" (planned)
- Data Science Seminar Series
  - Monthly for the past few years, moving to bi-weekly and then weekly
  - Advertised on the web and through an email listserv
  - Joint with iSchool beginning in spring
- Technology "boot camps"
  - Thus far Python (4 sessions thus far; next session in February), version control, cloud computing on AWS, cloud computing on Azure

- Special events
  - E.g., UW/MSR Machine Learning day last quarter (~300 participants), and a forthcoming campus workshop on Reproducible Research
- A program of small “pilot grants” joint with the Institute of Translational Health Sciences
  - A new bridge between lower campus and upper campus

A tiny sample of *many* closely related activities

- The establishment of NIAC, the Northwest Institute for Advanced Computing
  - A joint initiative of PNNL and UW; co-directed Thom Dunning and Vikram Jandhyala
  - Primary foci
    - Advanced and future computing systems
    - Scalable modeling, simulation, and design
    - Data-intensive science and discovery

- Computational Neuroscience Training Grant
  - Graduate and undergraduate students from Medicine, Arts & Sciences, and Engineering, advancing the theory and data-driven discovery of neural system function; directed by Adrienne Fairhall
- Air Force Center of Excellence for Nature Inspired Technologies
  - Partnership of UW, Case Western, and Maryland, supporting students and postdocs studying sensory information processing by natural systems in flight control, as inspiration for control of man-made dynamical systems; directed by Tom Daniel

GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION

NYU

Berkeley UNIVERSITY OF CALIFORNIA

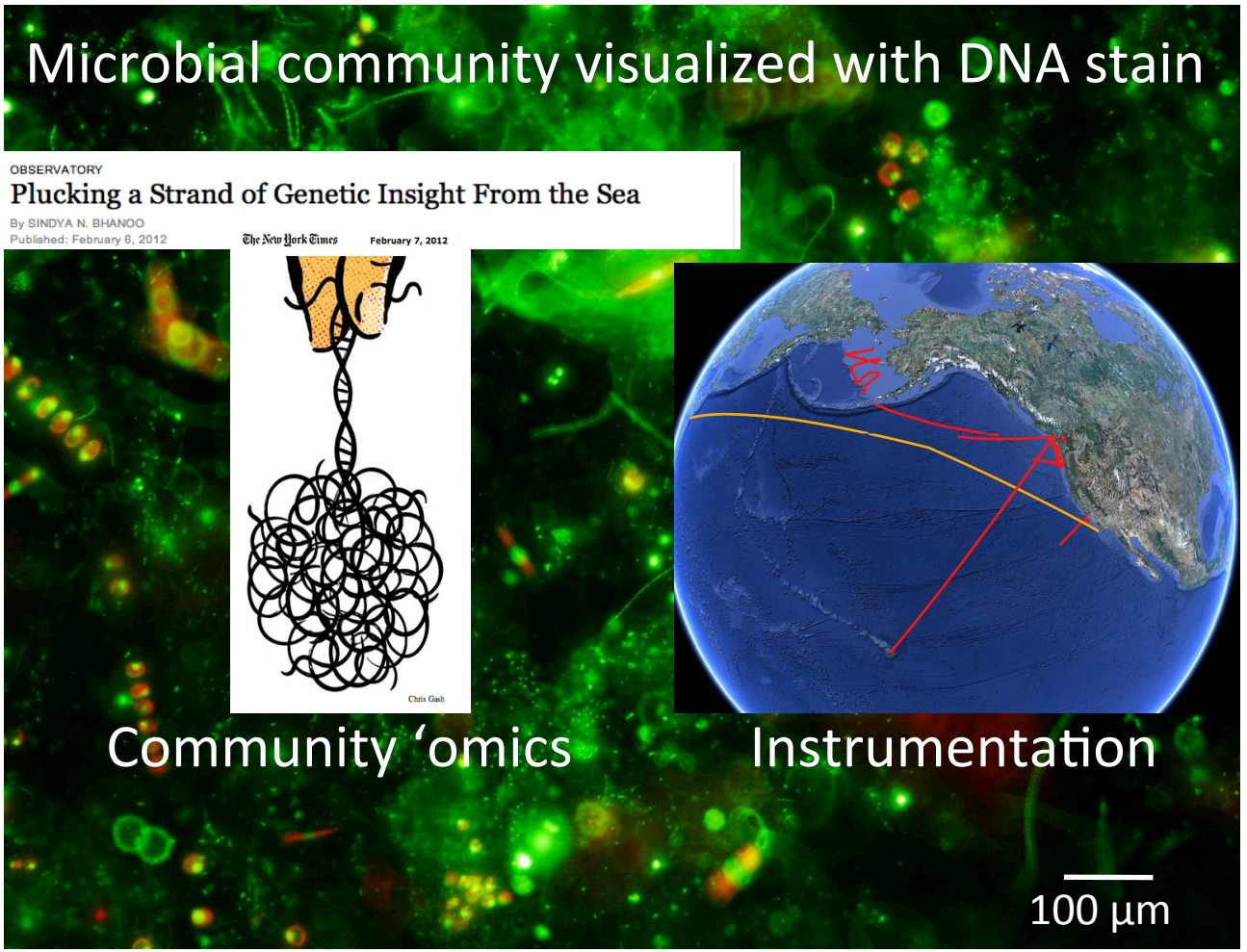
UNIVERSITY of WASHINGTON

The poster and networking session spans the campus and the region

- Allen Institute for Brain Science
- Anthropology
- Applied Mathematics
- Aquatic and Fishery Sciences
- Astronomy
- Biology
- Biomedical Informatics and Medical Education
- Biostatistics
- Center for Statistics and the Social Sciences
- Civil and Environmental Engineering
- Climate Impacts Group
- Computer Science & Engineering
- Earth and Space Sciences
- Economics
- Electrical Engineering
- eScience Institute
- Fred Hutchinson Cancer Research Center
- Genome Sciences
- Global Health
- Google
- GraphLab
- Human Centered Design and Engineering
- Industrial and Systems Engineering
- Information School
- Institute for Systems Biology
- Institute of Technology, UW-Tacoma
- Institute of Translational Health Sciences
- Jackson School of International Studies
- Joint Institute for the Study of the Atmosphere and Ocean
- Madrona Venture Group
- Mathematics
- Microbiology
- Microsoft
- Northwest Institute for Advanced Computing
- Oceanography
- Pacific Northwest National Laboratory
- Physics
- Political Science
- Psychiatry & Behavioral Sciences
- Sage Bionetworks
- Sociology
- Statistics
- Tableau Software
- Urban Design and Planning
- UW Information Technology
- UW Libraries

Data Science Example
Ginger Armbrust, Oceanography

Role of microbes in marine ecosystems



Data Science Example

Data Science Example

Data Science Example

Integrating across physics, biology, and chemistry

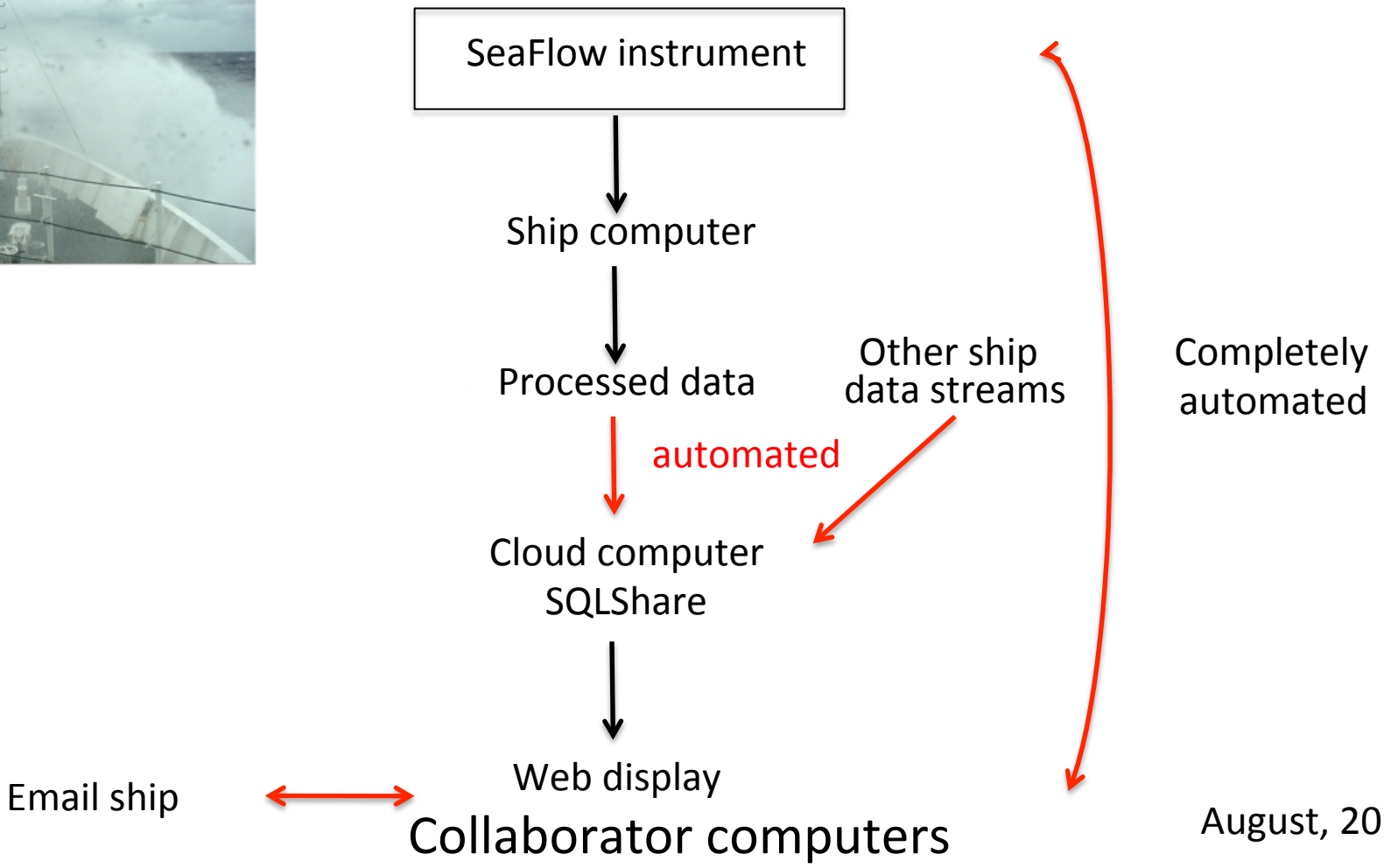
Query across data sets in real-time
"not just faster…different!"

Dan Halperin,
Research Scientist, eScience Institute

Konstantin Weitz
Graduate student, CSE

Data Science Example

Connecting across distributed labs



August, 2013

Big Data U: A Program for
Integrated Multidisciplinary Education & Research for
Big Data Science

PhD → πhD

NSF IGERT Grant

What is IGERT?

Integrative **G**raduate **E**ducation and **R**esearch **T**raineeship

NSF's flagship interdisciplinary training program

Special IGERT Track:

"A mechanism to address the ***training and education*** needs in computational and ***data enabled science and engineering*** (CDS&E) and ***cyberinfrastructure research***."

Big Goals

- Multidisciplinary Big Data Education
  - Ultimate goal: A new PhD program
  - Initial goal: A new certificate
  - Stepping stone: Big Data tracks in individual depts
- End-to-End Research Agenda
  - Big Data mgmt, analytics, modeling, & collaboration
- Cyberinfrastructure Development
  - All working toward a ***Big Data analysis service***

Small Budget

- $2.8 million over 5 years, nearly all for students
- Two year fellowships
- Four cohorts of students: 5, 6, 6, and 5

Participating Departments

- Six departments included in original proposal
  - Astronomy
  - Chemical Engineering
  - Computer Science & Engineering
  - Genome Sciences
  - Oceanography
  - Statistics

- All have a Big Data track or are putting one in place as we speak

The Sad Equation Today

departments >= # budgeted students

- Constrains IGERT fellowships to original six departments
- Departments must put a Big Data Track in place

How Can Others Get Involved?

- Come and talk to us about:
  - Creating a Big Data track in your department
  - Making your Big Data students part of IGERT cohorts even if not funded through IGERT

- Need to plan for sustainability beyond the IGERT grant
  - Let's start now!

IGERT Details: Curriculum

- Need to take 3 out of 4 core courses in methods
  - Statistics
  - Machine learning
  - Data management
  - Data visualization

- We also plan to require one applied course

IGERT Details: PhD Thesis

- Multidisciplinary Supervision
  - Primary advisor from participating department
  - Secondary advisor in a complementary field
- Interdisciplinary project
  - Two-quarter long project in complementary field
- Big Data Tools
  - Either use or build new tools
- Industrial Practical Training

IGERT Details: Research Focus

Streaming Data

Big Data access and management

Big Data modeling

Big Data analytics

Collaborative Big Data science

Big Data as a Service

Other Data Science Educational Initiatives on Campus

- Online courses
  - Computational methods of data analysis (Nathan Kutz)
  - High performance scientific computing (Randy LeVeque)
  - Introduction to Data Science (Bill Howe)
    - 7,000 certificates of completion granted!
- Traditional courses
  - Many! Example: "Biochemistry for Computer Scientists" (Joe Hellerstein)
  - We try to list relevant courses on eScience website
- UWEO certificate program in Data Science
- Workshops and Bootcamps
  - Software Carpentry (Winter & Summer '13, Winter, Spring, & Summer '14)
  - Cosmology and Machine Learning (Fall '14)

More Information Online

http://data.washington.edu/

And click on “Big Data U”

More Information: IGERT Contact People (1/3)

- IGERT Program Manager:
  - Jennifer Worrell – Comp. Science & Eng.
- Steering Committee:
  - Ginger Armbrust – Oceanography
  - Magda Balazinska – Comp. Science & Eng. (IGERT director)
  - Andrew Connolly – Astronomy
  - Emily Fox – Statistics
  - Carlos Guestrin – Comp. Science & Eng.
- Additional Department Leads:
  - David Beck – Chemical Engineering
  - Bill Noble – Genome Sciences

More Information: IGERT Contact People (2/3)

- Advisory Committee
  - Terrie Klinger - School of Marine and Env. Affairs
  - Ed Lazowska - Comp. Science & Eng. and eScience Institute
  - Hank Levy - Comp. Science & Eng.
  - Vikki Meadows - Astronomy
  - Daniel Schwartz - Chemical Engineering
  - Werner Stuetzle - Statistics

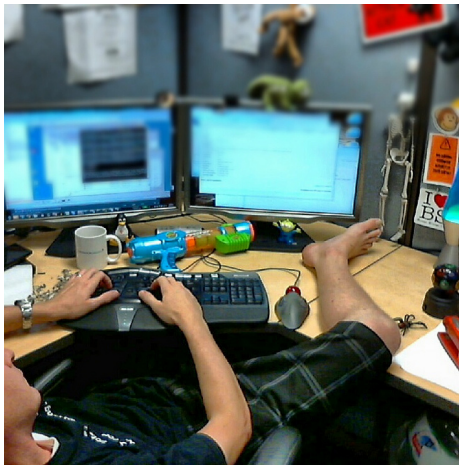
More Information: IGERT Contact People (3/3)

- Other faculty participants
  - Dan Grossman, Comp. Science & Eng.
  - Jeff Heer, Comp. Science & Eng.
  - Bill Howe, Comp. Science & Eng. and eScience Institute
  - Željko Ivezić, Astronomy
  - Marina Meila, Statistics
  - LuAnne Thompson, Oceanography

Provost's Initiative in Data-Intensive Discovery

- A pool of funds to support the hiring of extraordinarily capable faculty who
  - do cutting-edge research on new methodologies for data-intensive discovery
  - are committed to putting these methodologies into the hands of UW's broad base of outstanding researchers
- Provides up to a maximum of 50% of salary and benefits, not to exceed $75,000
- A corresponding proportion of teaching and service are devoted to campus-wide activities (e.g., teaching IGERT courses)
- See the eScience Institute website or the Provost's Initiatives website

The Data Science Studio: Resurrecting the water cooler



A partnership among …

- Provost
- UW Libraries
- Physics, Astronomy, Arts & Sciences
- eScience Institute

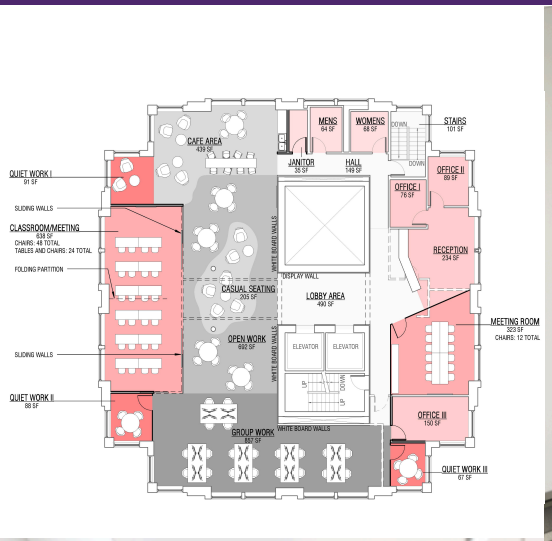
6th floor Physics Astronomy Building

The Data Science Studio

- An open research space where YOU will come to collaborate
- A resident data science team
  - Permanent staff of ~5 *data scientists* – applied research and development
  - ~15-20 data science fellows (research scientists, visitors, postdocs, students)

- How to Engage:
  - Drop-in open workspace
  - Studio "Office Hours"
  - Pilot Incubation Program

    (Ramp-up only; full launch Fall 2014)
  - …plus seminars, sponsored lunches, workshops, bootcamps, joint proposals...



Estimated Timeline:

- Design Phase Jan-June
- Construction June – Sep
- *Target: October 1, 2014*

Data Science Incubation Program

- *Goal:* Create watercooler opportunities and scale our efforts by co-locating collaborations from different fields in the studio
- *Protocol:* ~1-page proposals for short-term, intensive data science collaborations with the studio staff
- *What we're looking for:* Projects where fruitful collaboration is possible, with potential for significant impact, and that have sustained engagement
- *Next steps*: Join us for an **information session February 20, 10-12, EE 303**

Pilot underway now, with full launch in Fall 2014

http://data.uw.edu/incubator

UNIVERSITY of WASHINGTON

Data Science @ UW

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