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

Chesapeake Large Scale Analytics Conference

October 2015

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

Today

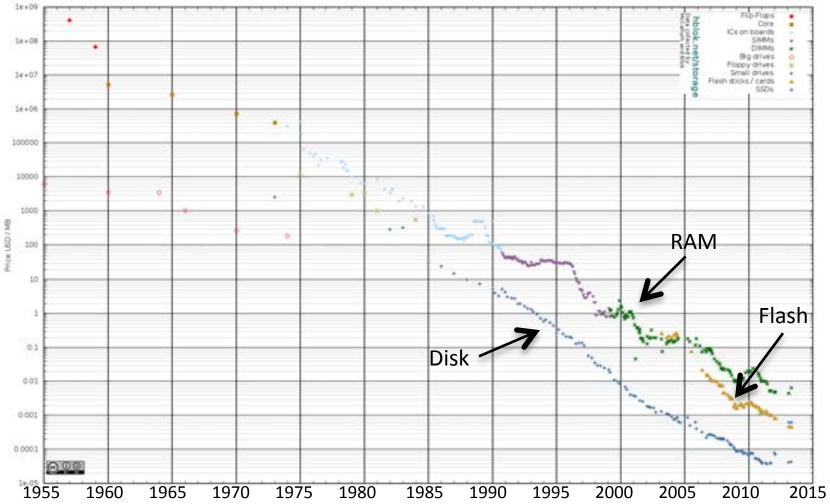
- A reminder of the extraordinary progress that Computer Science has achieved
- “Big Data” and “Smart Everything”
- Jim Gray’s “Fourth Paradigm”: smart discovery / data-intensive discovery / eScience
- The University of Washington eScience Institute, and the Moore/Sloan Data Science Environments
- A 21st century view of Computer Science
- Recommendations for the support of 21st century cyberinfrastructure

Every aspect of computing has experienced exponential improvement

- Processing capacity
- Storage capacity
- Network bandwidth
- Sensors
- Astonishingly, even algorithms in some cases!

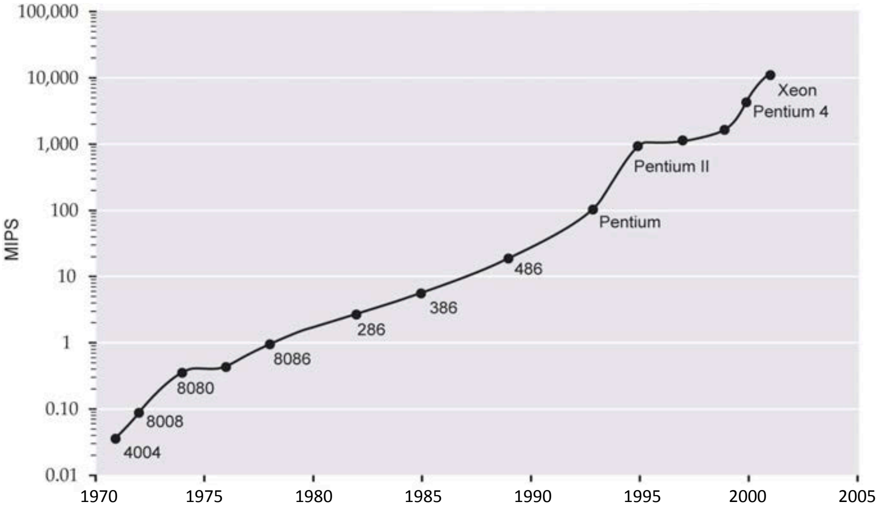
You can exploit these improvements in two ways

- Constant capability at exponentially decreasing cost
- Exponentially increasing capability at constant cost



Storage Price / MB, USD
(semi-log plot)

John McCallum / Havard Blok



Microprocessor Performance, MIPS
(semi-log plot)

Ray Kurzweil

Today, these exponential improvements in technology and algorithms are enabling a "big data" revolution

- A proliferation of sensors
  - Think about the sensors on your phone
- More generally, the creation of almost all information in digital form
  - It doesn't need to be transcribed in order to be processed
- Dramatic cost reductions in storage
  - You can afford to keep all the data
- Dramatic increases in network bandwidth
  - You can move the data to where it's needed

- Dramatic cost reductions and scalability improvements in computation
  - With Amazon Web Services, 1000 computers for 1 day costs the same as 1 computer for 1000 days
- Dramatic algorithmic breakthroughs
  - Machine learning, data mining – fundamental advances in computer science and statistics
- Ever more powerful models producing ever-increasing volumes of data that must be analyzed

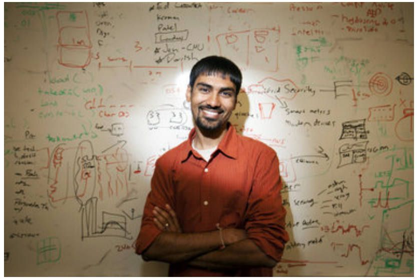
“Big Data” is enabling computer scientists to put the “smarts” into everything

Xconomy

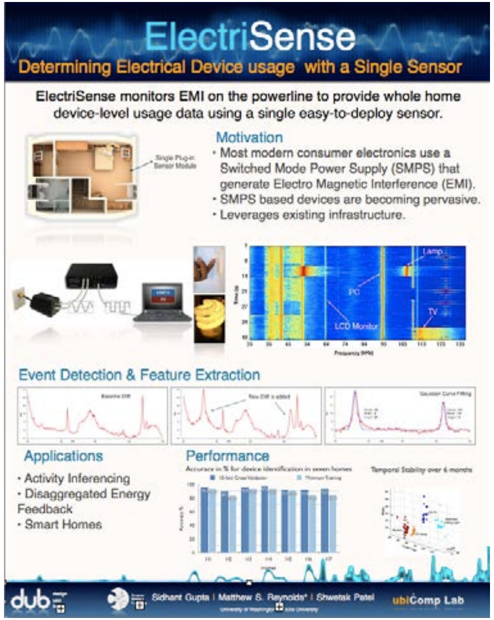
Business + Technology in the Exponential Economy

- Smart homes
- Smart cars
- Smart health
- Smart robots
- Smart crowds and human-computer systems
- Smart education
- Smart interaction (virtual and augmented reality)
- Smart cities
- Smart discovery

Smart homes (the leaf nodes of the smart grid)



Shwetak Patel,
University of Washington
2011 MacArthur Fellow



ElectriSense

Determining Electrical Device usage with a Single Sensor

ElectriSense monitors EMI on the powerline to provide whole home device-level usage data using a single easy-to-deploy sensor.

Motivation

- Most modern consumer electronics use a Switched Mode Power Supply (SMPS) that generate Electro Magnetic Interference (EMI).
- SMPS based devices are becoming pervasive.
- Leverages existing infrastructure.

Event Detection & Feature Extraction

Applications

- Activity Inferencing
- Disaggregated Energy Feedback
- Smart Homes

Performance

Accuracy in % for device identification in seven homes

Temporal Stability over 6 months

Sidhant Gupta | Matthew S. Reynolds* | Shwetak Patel

belkin echo

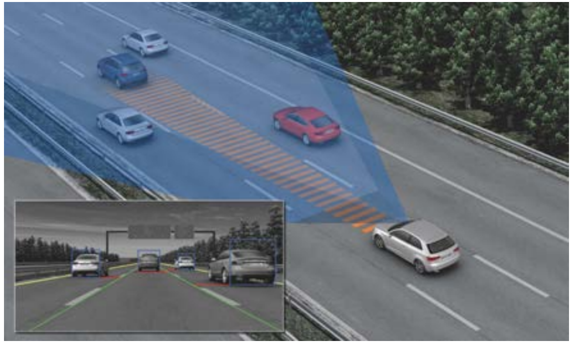
Smart cars

DARPA Grand Challenge

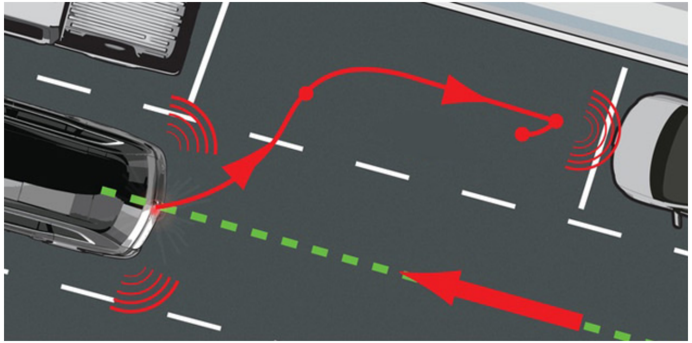
DARPA Urban Challenge

Google Self-Driving Car

Adaptive cruise control



Self-parking



Smart health

Larry Smarr – “quantified self”

Evidence-based medicine

P4 medicine

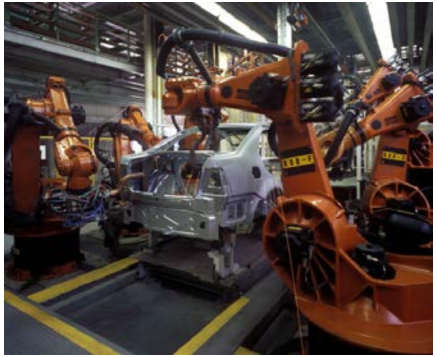
→

PIONEER 100
HUNDRED PERSON WELLNESS PROJECT
Institute for Systems Biology

→

Arivale
Your scientific path to wellness

Smart robots



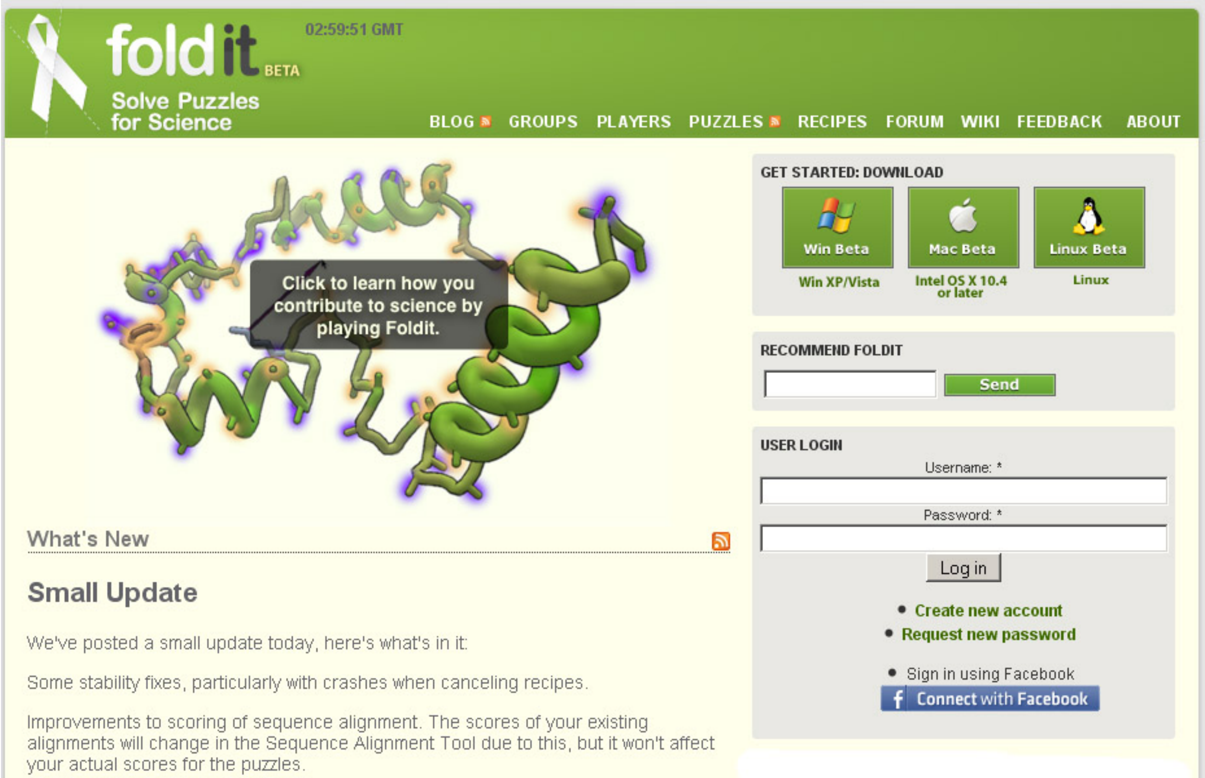
iRobot

rethink robotics

Smart crowds and human-computer systems

Zoran Popovic
UW Computer Science & Engineering

David Baker
UW Biochemistry



What's New

Small Update

We've posted a small update today, here's what's in it:

Some stability fixes, particularly with crashes when canceling recipes.

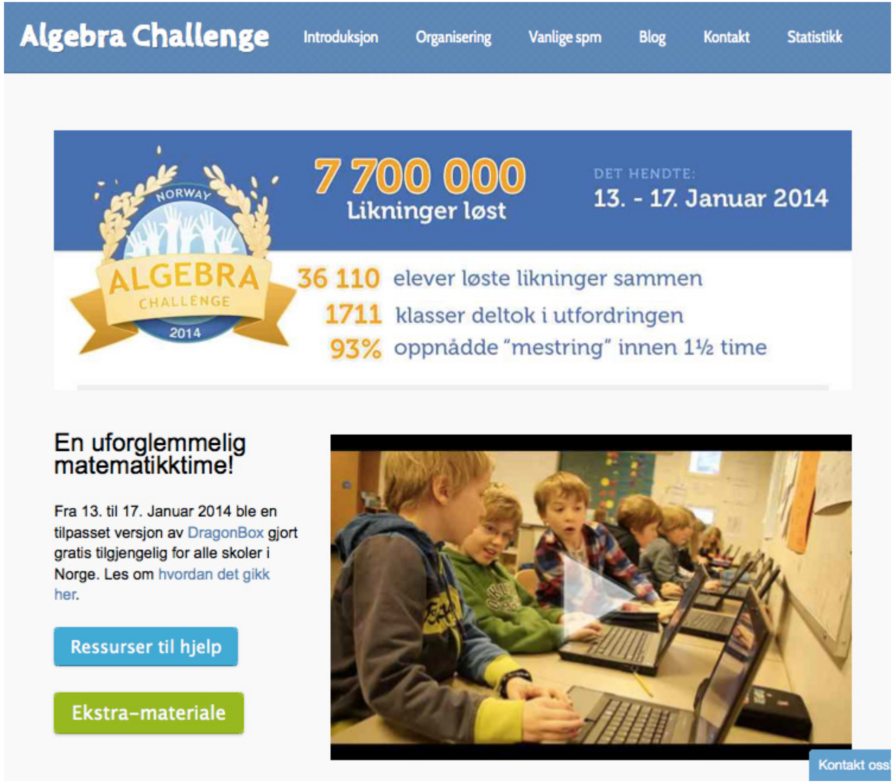
Improvements to scoring of sequence alignment. The scores of your existing alignments will change in the Sequence Alignment Tool due to this, but it won't affect your actual scores for the puzzles.

Smart education

Zoran Popovic
UW Computer Science & Engineering

CGS Center for Game Science

Enlearn



7 700 000 Likninger løst

DET HENDTE:
13. - 17. Januar 2014

36 110 elever løste likninger sammen
1711 klasser deltok i utfordringen
93% oppnådde "mestring" innen 1½ time

En uforglemmelig matematikktime!

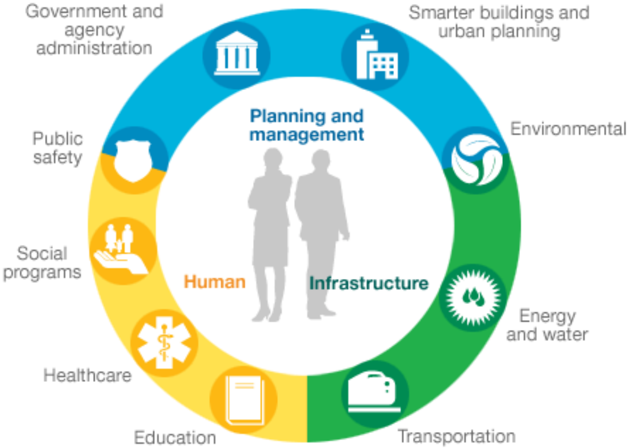
Fra 13. til 17. Januar 2014 ble en tilpasset versjon av DragonBox gjort gratis tilgjengelig for alle skoler i Norge. Les om hvordan det gikk her.

Smart interaction

KINECT™
for XBOX 360

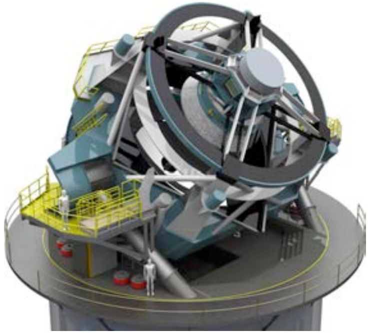
Microsoft HoloLens

Smart cities

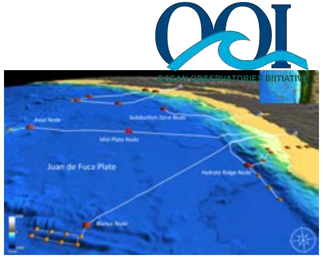


Smart discovery (data-intensive discovery, or eScience)

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



Astronomy: LSST



Oceanography: OOI

Physics: LHC

Sociology: The Web

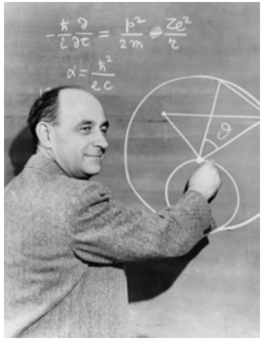
Biology: Sequencing

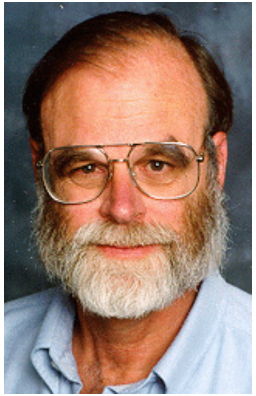
Economics: POS terminals

Neuroscience: EEG, fMRI

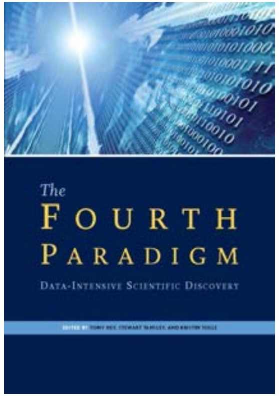
The Fourth Paradigm

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





Jim Gray



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

UW eScience Institute

- *"All across our campus, the process of discovery will increasingly rely on researchers' ability to extract knowledge from vast amounts of data... In order to remain at the forefront, UW must be a leader in advancing these techniques and technologies, and in making [them] accessible to researchers in the broadest imaginable range of fields."* (2007)

UNIVERSITY of WASHINGTON

eScience Institute

ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

Major sources of support for our “core effort”

- University of Washington
  - $725,000/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

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

GORDON AND BETTY MOORE FOUNDATION

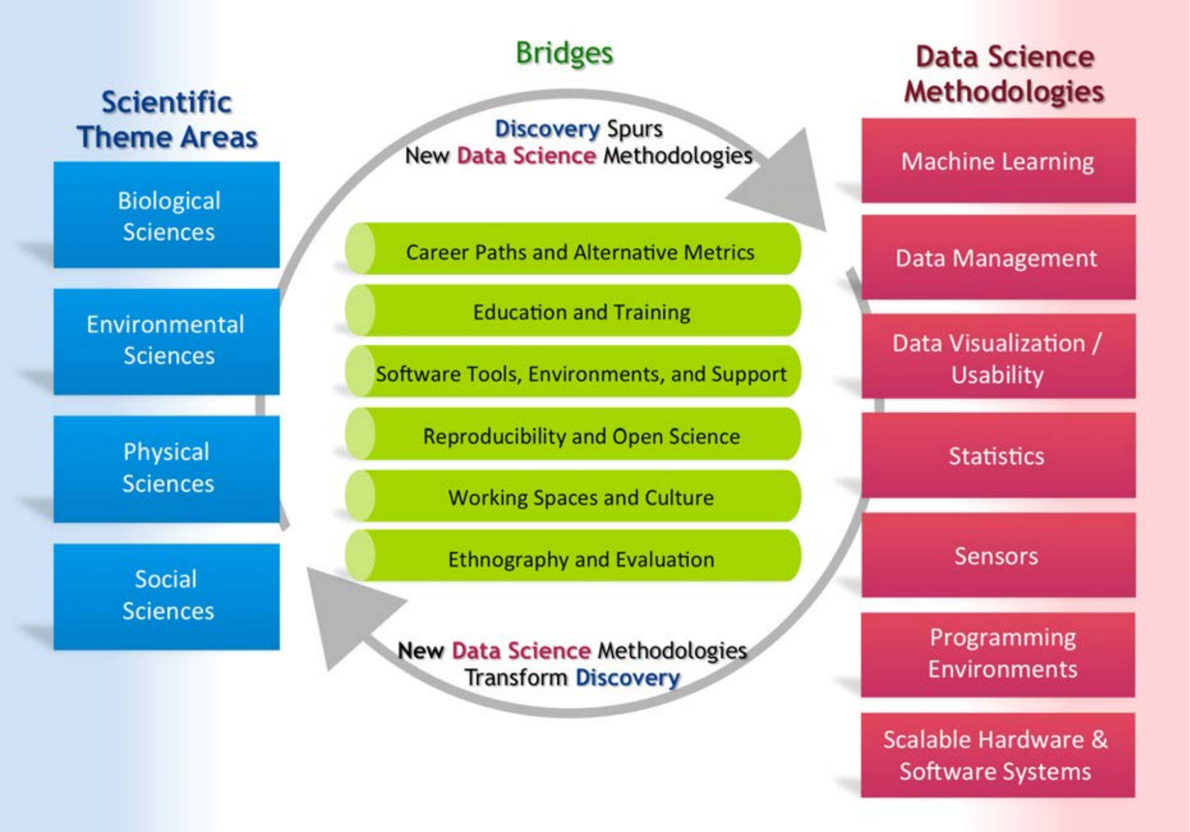
ALFRED P. SLOAN FOUNDATION 1934

NYU

Berkeley UNIVERSITY OF CALIFORNIA

W UNIVERSITY of WASHINGTON

Vision



Scientific Theme Areas

- Biological Sciences
- Environmental Sciences
- Physical Sciences
- Social Sciences

Bridges

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

UW’s 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, CSSS

Werner Stuetzle
Statistics

Life 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

UW's original core faculty team

13 Departments

5 Schools / Colleges

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

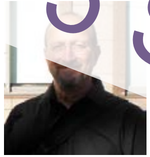
Randy LeVeque
Applied Mathematics

Thomas Richardson
Statistics, CSSS

Werner Stuetzle
Statistics

Life sciences

David Beck
Chemical Engr.


Tom Daniel
Biology

Bill Noble
Genome Sciences

Environmental sciences

Ginger Armbrust
Oceanography

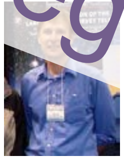
Social sciences

Josh Blumenstock
iSchool

Mark Ellis
Geography

Tyler McCormick
Sociology, CSSS

Physical sciences


Andy Connolly
Astronomy

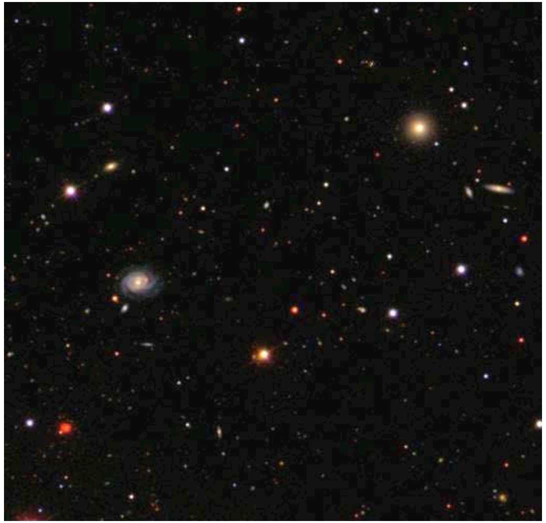
John Vidale
Earth & Space Sciences

Science example: AstroDB – Cosmology at Scale

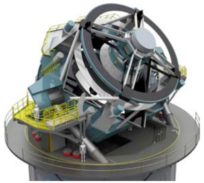
Andrew Connolly (Astronomy), Magda Balazinska (Computer Science & Engineering)

Large Synoptic Survey Telescope

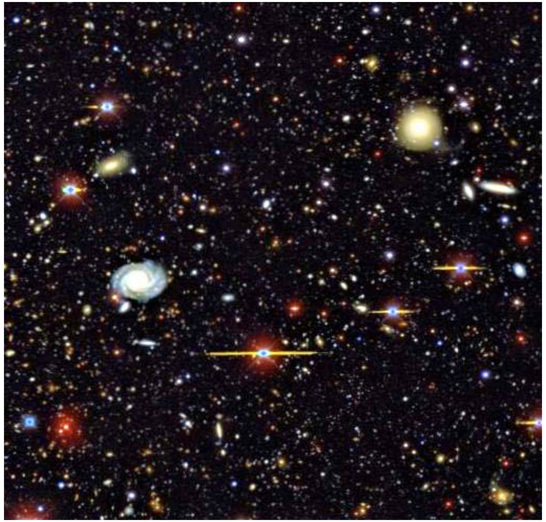
- Survey half the sky every 3 nights (1000-fold increase in data vs. Sloan Digital Sky Survey)
- Enabled by a 3.2 Gigapixel camera with a 3.5 degree field
- 15 TB/night (100 PB over 10 years), 20 billion objects, and 20 trillion measurements
- Will enable dramatically improved resolution, time-series analysis



SDSS



LSST



Credit: Andy Connolly, University of Washington

How do we do science at petabyte scale?

Science questions …

- Finding the unusual
  - Supernova, GRBs
  - Probes of Dark Energy
- Finding moving sources
  - Asteroids and comets
  - Origins of the solar system
- Mapping the Milky Way
  - Tidal streams
  - Probes of Dark Matter
- Measuring shapes of galaxies
  - Gravitational lensing
  - The nature of Dark Energy

Credit: Andy Connolly, University of Washington

How do we do science at petabyte scale?

Science questions … map to computational questions

- Finding the unusual
  - Supernova, GRBs
  - Probes of Dark Energy
- Finding moving sources
  - Asteroids and comets
  - Origins of the solar system
- Mapping the Milky Way
  - Tidal streams
  - Probes of Dark Matter
- Measuring shapes of galaxies
  - Gravitational lensing
  - The nature of Dark Energy

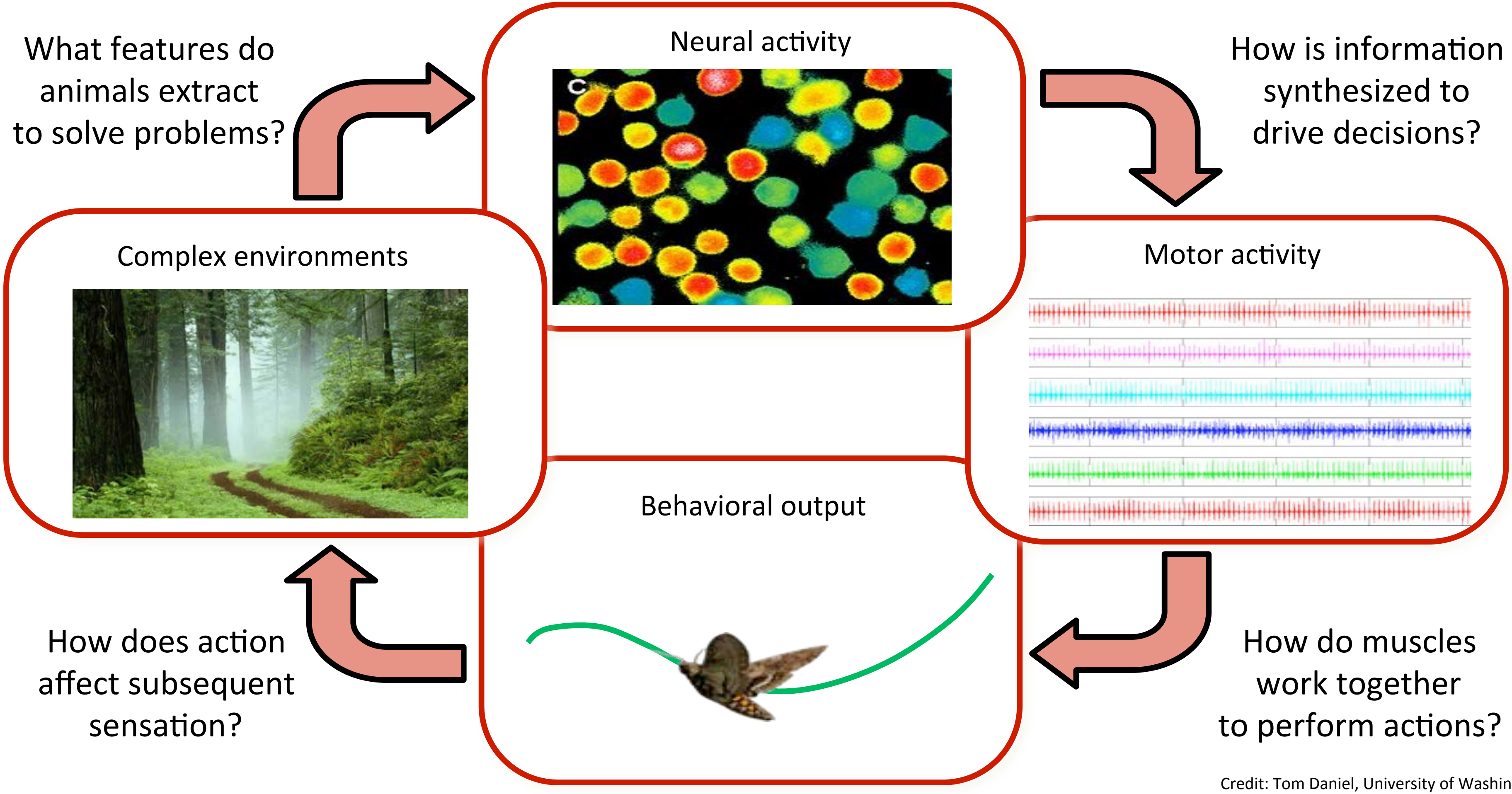
- Finding the unusual
  - Anomaly detection
  - Density estimations
- Finding moving sources
  - Tracking algorithms
  - Kalman filters
- Mapping the Milky Way
  - Clustering techniques
  - Correlation functions
- Measuring shapes of galaxies
  - Image processing
  - Data intensive analysis

Credit: Andy Connolly, University of Washington

Science example: Devices + Neuroscience + Data Science

Tom Daniel & Bing Brunton (Biology), Adrienne Fairhall (Physiology & Biophysics)

Credit: Tom Daniel, University of Washington



What features do animals extract to solve problems?

Neural activity

How is information synthesized to drive decisions?

Motor activity

How do muscles work together to perform actions?

Behavioral output

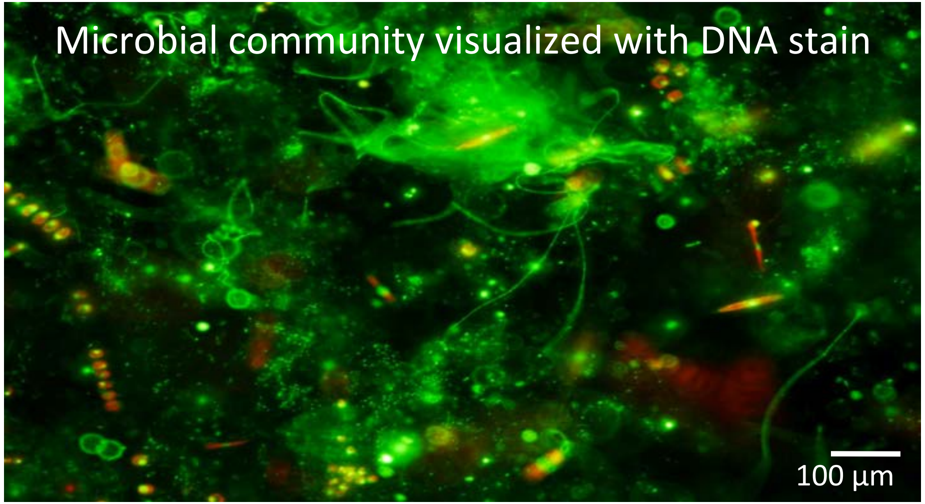
How does action affect subsequent sensation?

Complex environments

Credit: Tom Daniel, University of Washington

Science example: Role of microbes in marine ecosystems

Ginger Armbrust (Oceanography), Bill Howe (CSE + eScience Institute)

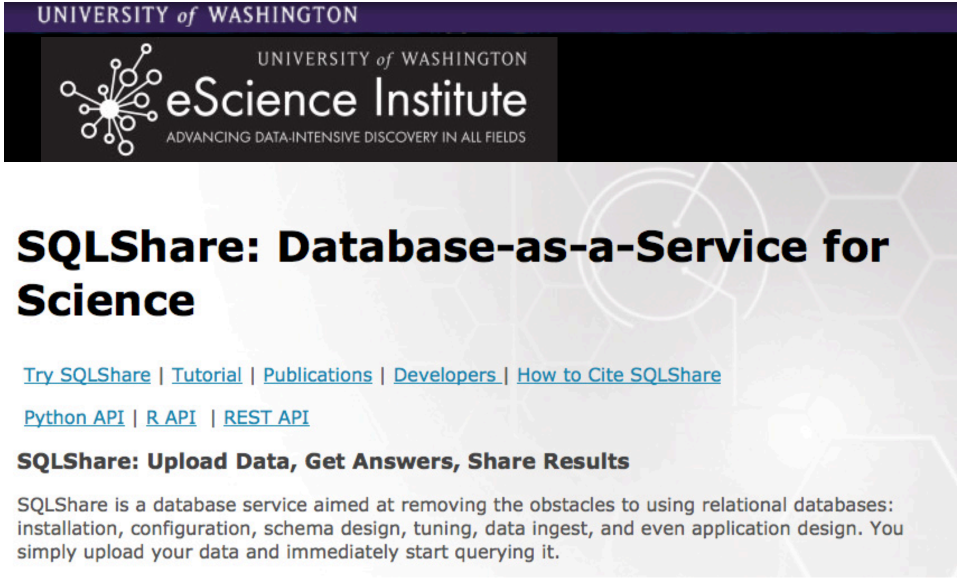


Microbial community visualized with DNA stain

Challenges:

- Integration across different data types
- Distributed and remote labs

Credit: Ginger Armbrust, University of Washington



SQLShare: Database-as-a-Service for Science

Try SQLShare | Tutorial | Publications | Developers | How to Cite SQLShare

Python API | R API | REST API

SQLShare: Upload Data, Get Answers, Share Results

SQLShare is a database service aimed at removing the obstacles to using relational databases: installation, configuration, schema design, tuning, data ingest, and even application design. You simply upload your data and immediately start querying it.

Credit: Ginger Armbrust, University of Washington

Integrating across physics, biology, and chemistry

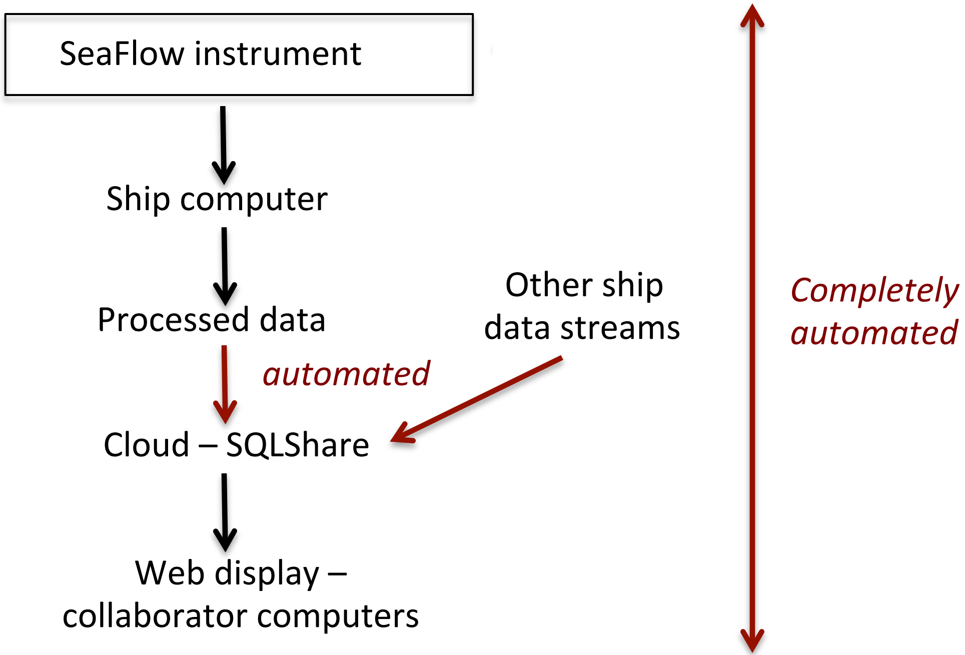
Query across data sets in real-time: “not just faster…different!”

Dan Halperin,
Research Scientist, eScience Institute

Konstantin Weitz
Graduate student, CSE

Credit: Ginger Armbrust, University of Washington

Connecting across distributed labs



SeaFlow instrument

Ship computer

Processed data

automated

Cloud – SQLShare

Web display – collaborator computers

Other ship data streams

Completely automated

Credit: Ginger Armbrust, University of Washington

Science Example: Data Science for Social Good / Urban Science

Summer 2015

- 4 projects (from among 11 proposals):
  - Optimizing Paratransit Routing
    - In collaboration with King County Metro and UW's Taskar Center for Accessible Technology
  - Assessing Community Well-Being through Open Data & Social Media
    - In collaboration with Third Place Technologies
  - Open Sidewalks – Sidewalk Maps for Low-Mobility Citizens
    - In collaboration with UW's Taskar Center for Accessible Technology
  - Predictors of Permanent Housing for Homeless Families
    - In collaboration with the Bill & Melinda Gates Foundation, Building Changes, and King, Pierce, and Snohomish Counties WA
- 16 undergraduate and graduate students (from among 144 applicants)
- 6 ALVA socioeconomically disadvantaged high school students
- 8 eScience Institute Data Scientists

THE UNIVERSITY OF CHICAGO

Georgia Institute of Technology

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

Data Science for Social Good

Predictors of Permanent Housing for Homeless Families

The Bill & Melinda Gates Foundation and Building Changes have partnered with King, Pierce, and Snohomish Counties WA to make homelessness in these counties rare, brief, and one-time

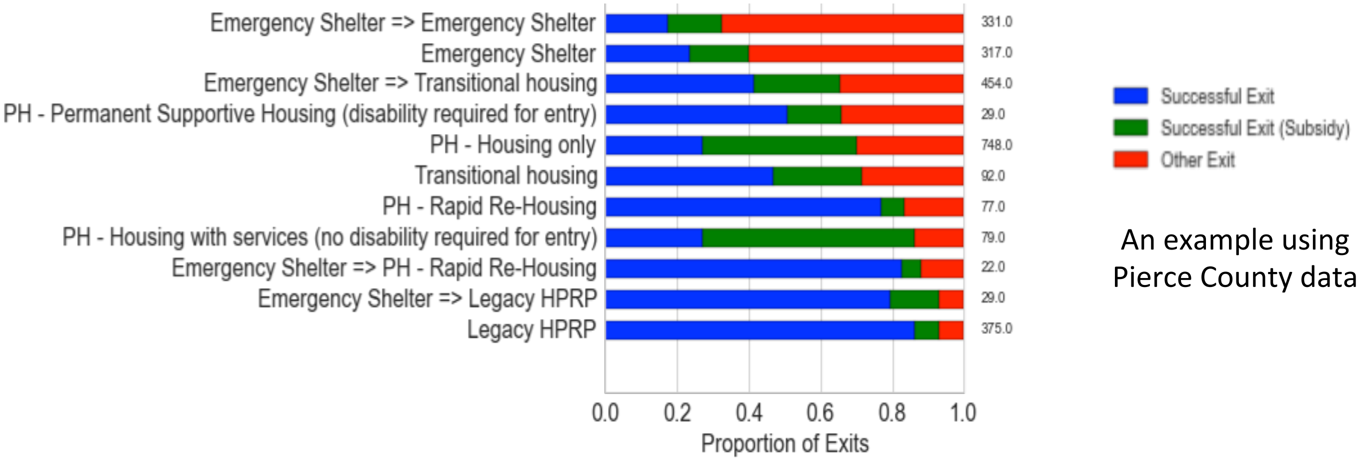
When homeless families engage in services and programs, what factors are most likely to lead to a successful exit?

The DSSG team:

- Developed algorithms to identify “families”
- Developed algorithms to identify “episodes” of homelessness including back-to-back or overlapping enrollments in individual programs
- Devised innovative ways to visualize and analyze the ways families transition between programs

Project Leads: Neil Roche & Anjana Sundaram, Bill & Melinda Gates Foundation
DSSG Fellows: Joan Wang, Jason Portenoy, Fabliha Ibnat, Chris Suberlak
ALVA High School Students: Cameron Holt, Xilalit Sanchez
eScience Institute Data Scientist Mentors: Ariel Rokem, Bryna Hazelton

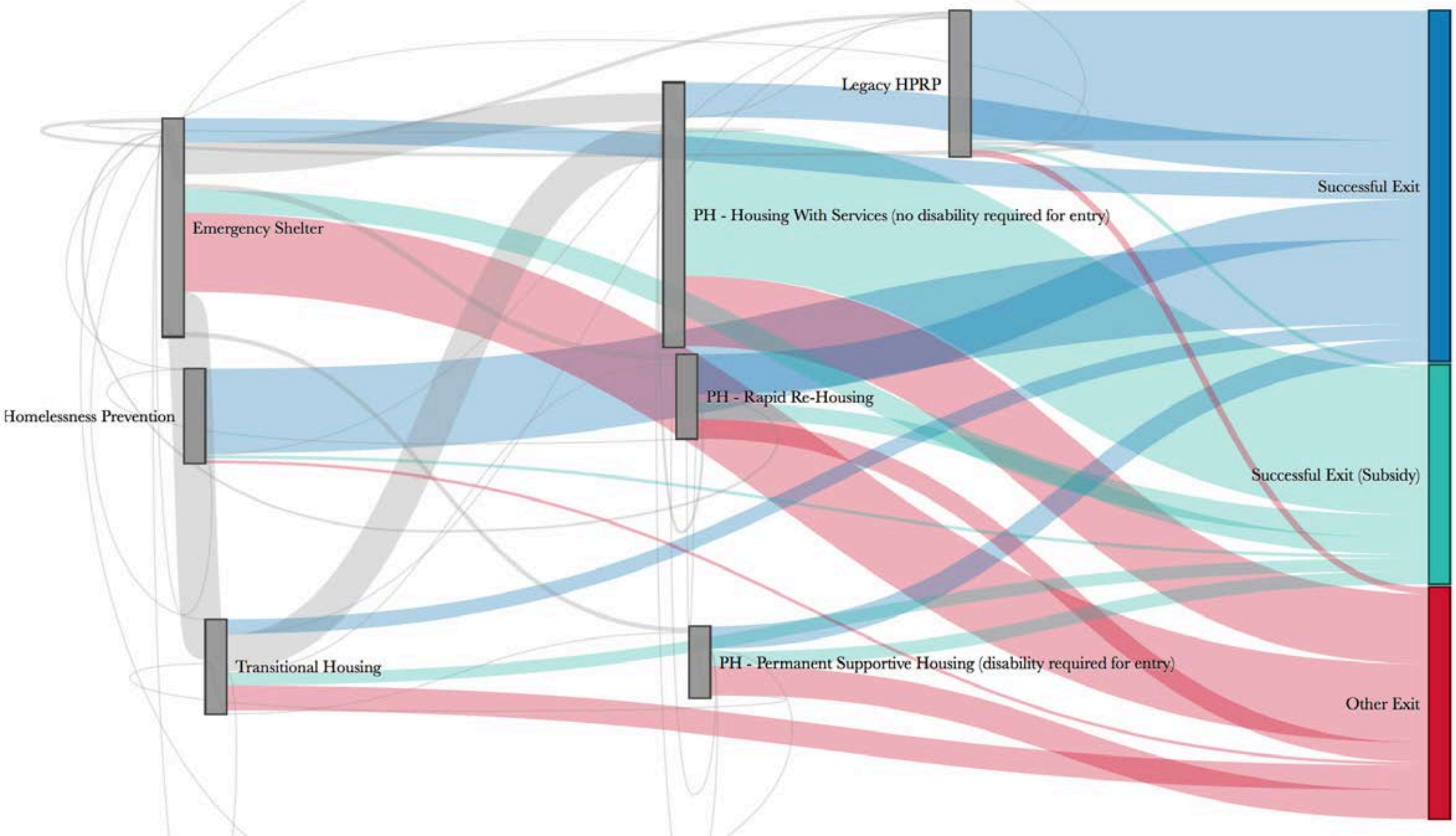
Novel Analyses of Family Trajectories through Programs



An example using Pierce County data

Common trajectories lead to different outcomes:

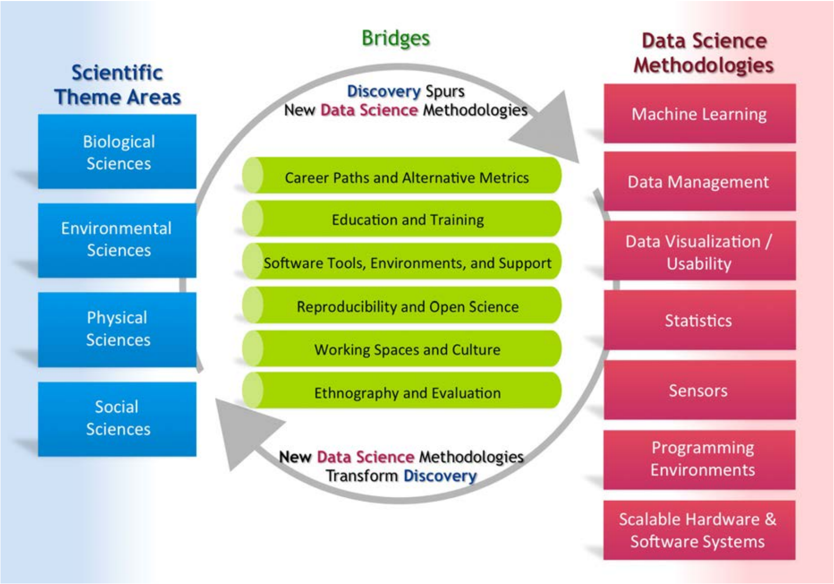
- A highly successful exit from an episode would mean that the family found a permanent housing solution
- Another successful exit involves continued receipt of government subsidies
- Other exits are exits back into homelessness, or to other, unknown destinations


Legacy HPRP
PH - Housing With Services (no disability required for entry)
Emergency Shelter
Homelessness Prevention
PH - Rapid Re-Housing
Transitional Housing
PH - Permanent Supportive Housing (disability required for entry)
Successful Exit
Successful Exit (Subsidy)
Other Exit

Using the D3 technology developed in Jeff Heer's group, the DSSG team created interactive Sankey diagrams and other visualizations to facilitate exploration of the data by stakeholders. (This diagram shows the proportional flow from one program to another, as well as the eventual outcome.)

A closer look at the Moore/Sloan Data Science Environments

Launched late fall 2013



Scientific Theme Areas

- Biological Sciences
- Environmental Sciences
- Physical Sciences
- Social Sciences

Bridges

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

GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION 1934

NYU

Berkeley UNIVERSITY OF CALIFORNIA

UNIVERSITY of WASHINGTON

Career paths and alternative metrics

UW flagship activity: Establish two new roles on campus: "Data Science Fellows" and "Data Scientists"

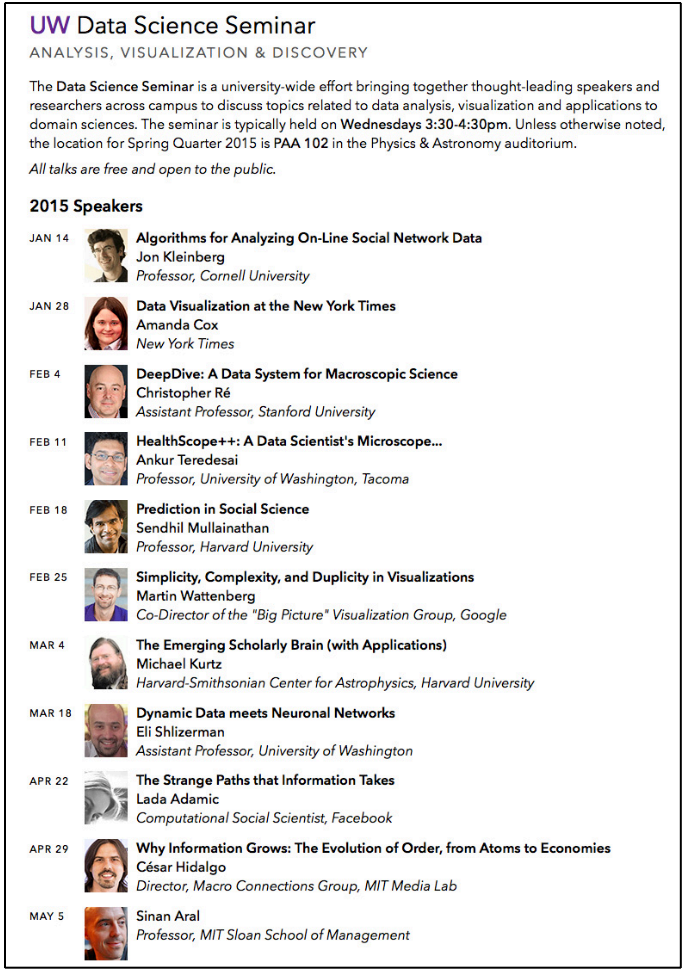
- Recruited / recruiting data scientists – and put processes into place
  - Typically Ph.D.-educated; fully supported by DSE; research position with emphasis on taking responsibility for core activities (e.g., incubator projects)
- Recruited / recruiting research scientists – and put processes into place
  - 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 6 "Provost's Initiative" faculty members – ditto
  - Provost provided 6 faculty "half-positions"
  - Individuals who are truly "π-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, Applied Mathematics, Statistics + Computer Science & Engineering
- Recruited 2 cohorts of 6 Data Science Postdoctoral Fellows – ditto
  - Each is co-mentored by "methodology" and "applications" faculty

Education and training

UW flagship activity: Establish new graduate program tracks in data science

- IGERT Ph.D. program in Big Data / Data Science
  - 6 departments have added a transcript-recognized *Advanced Data Science Option* to their Ph.D. programs
    - Data science classes count toward Ph.D. (no extra work)
  - "Regular" Data Science Option coming soon
    - Prepares students to use advanced data science tools, vs. creating them
  - Started IGERT seminar as the eScience Community Seminar
  - Put in place a detailed program evaluation plan with Data2Insight
  - 3rd cohort of IGERT Ph.D. students, from a variety of departments, arriving this fall
    - Each student is co-mentored by "methodology" and "applications" faculty
- Undergraduate "transcriptable option" starting this fall
- Fall 2016 launch of a Data Science Masters degree

- 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**. Unless otherwise noted, the location for Spring Quarter 2015 is **PAA 102** in the Physics & Astronomy auditorium.

All talks are free and open to the public.

2015 Speakers

JAN 14 — **Algorithms for Analyzing On-Line Social Network Data**
Jon Kleinberg
Professor, Cornell University

JAN 28 — **Data Visualization at the New York Times**
Amanda Cox
New York Times

FEB 4 — **DeepDive: A Data System for Macroscopic Science**
Christopher Ré
Assistant Professor, Stanford University

FEB 11 — **HealthScope++: A Data Scientist's Microscope...**
Ankur Teredesai
Professor, University of Washington, Tacoma

FEB 18 — **Prediction in Social Science**
Sendhil Mullainathan
Professor, Harvard University

FEB 25 — **Simplicity, Complexity, and Duplicity in Visualizations**
Martin Wattenberg
Co-Director of the "Big Picture" Visualization Group, Google

MAR 4 — **The Emerging Scholarly Brain (with Applications)**
Michael Kurtz
Harvard-Smithsonian Center for Astrophysics, Harvard University

MAR 18 — **Dynamic Data meets Neuronal Networks**
Eli Shlizerman
Assistant Professor, University of Washington

APR 22 — **The Strange Paths that Information Takes**
Lada Adamic
Computational Social Scientist, Facebook

APR 29 — **Why Information Grows: The Evolution of Order, from Atoms to Economies**
César Hidalgo
Director, Macro Connections Group, MIT Media Lab

MAY 5 — Sinan Aral
Professor, MIT Sloan School of Management

Software tools, environments, and support

UW 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 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
    - Data Science for Social Good was a "special case" Incubator cohort
- Weekly code reviews
- Leadership in the open source science community
  - Keynotes at PyData
  - Contributions to mainstream projects (e.g., scikit-learn (machine learning in Python))

- Drop-in “Office Hours”
  - eScience Institute Data Scientists
  - UW-IT Academic & Collaborative Applications Team, Research Computing Team, Network Design & Architecture Team
  - AWS Scientific Computing Team
  - Center for Statistics and the Social Sciences Statistical Consulting Service
  - UW Libraries Research Data Management Team
  - Google Cloud Platform Team
- 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

SQLSHARE

Myria

Reproducibility and open science

UW flagship activity: Establish a campus-wide community around reproducible research

- UW campus-wide monthly reproducibility seminars and working group meetings
- National workshops at UW (2014), Berkeley (2015), NYU (2016)
  - Broad involvement from academia, industry, non-profits
- Draft guidelines for reproducible research
- Weekly tutorials on "research hygiene" topics
  - E.g. GitHub, KnitR, iPython Notebook
- Template for recording & categorizing research publications on reproducibility spectrum
- Self-certification & badging of research groups for reproducibility

Working spaces and culture

UW flagship activity: Establish a "Data Science Studio"

- Washington Research Foundation Data Science Studio

Washington Research
FOUNDATION

Ethnography and evaluation

UW flagship activity: Establish a research program in "the data science of data science"

- Ethnography and evaluation integrated 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 (Data Science tracks in multiple Ph.D. programs)
  - Workshops (e.g. Software Carpentry, NSF-sponsored Data Science Workshop, M9 Interdisciplinary Workshop), Bootcamps (e.g. Python, R) , Hack Weeks (e.g. AstroData Hack)
  - Incubator projects ("regular" + Data Science for Social Good)
  - Case studies across Astronomy and Oceanography
- 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

General role as a catalyst

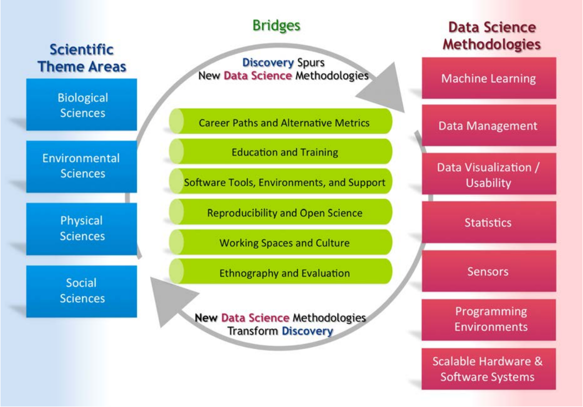
- Annual campus-wide “all call” data science research poster sessions

- Various “special interest group” lunches held periodically to build community (e.g., “Big Social Data”)
- Played a central role in launching Urban@UW

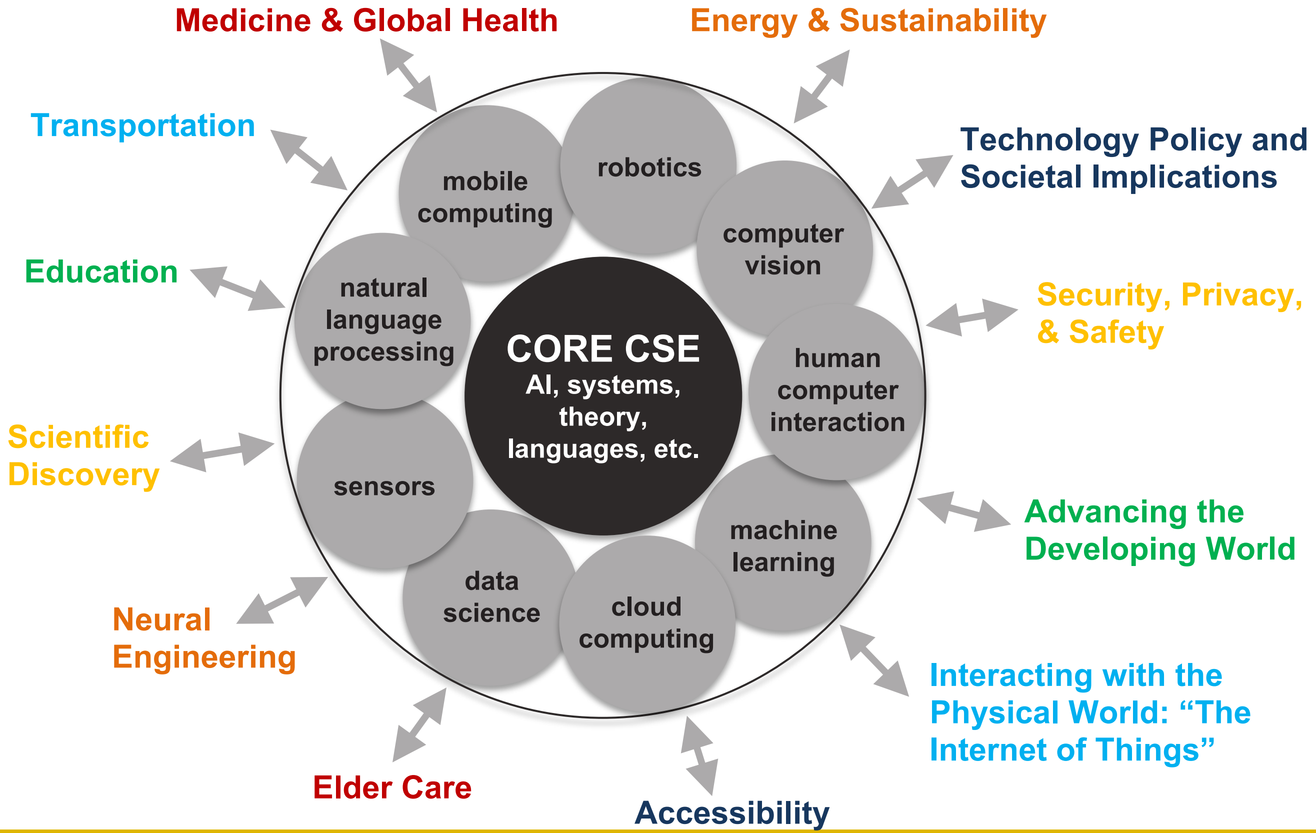
- “A Switzerland” to thwart attempts at data science “land grabs”

Similarly at NYU and UC Berkeley

- Pursuing the same goals
  - Lead in advancing data science methodologies
  - Lead in putting these methodologies to work in discovery
  - Lead in creating environments where data science can flourish
- 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, annual Moore/Sloan Data Science Summit, …)
  - Visits
  - Open sharing of successes and – importantly – failures


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

A 21st century view of Computer Science: A field unique in its societal impact


CORE CSE
AI, systems, theory, languages, etc.
mobile computing
robotics
computer vision
human computer interaction
machine learning
cloud computing
data science
sensors
natural language processing
Medicine & Global Health
Energy & Sustainability
Technology Policy and Societal Implications
Security, Privacy, & Safety
Advancing the Developing World
Interacting with the Physical World: "The Internet of Things"
Accessibility
Elder Care
Neural Engineering
Scientific Discovery
Education
Transportation

Is this stuff computer science?

Medicine & Global Health

Energy & Sustainability

Transportation

Technology Policy and Societal Implications

Education

Security, Privacy, & Safety

Scientific Discovery

Advancing the Developing World

Neural Engineering

Interacting with the Physical World: “The Internet of Things”

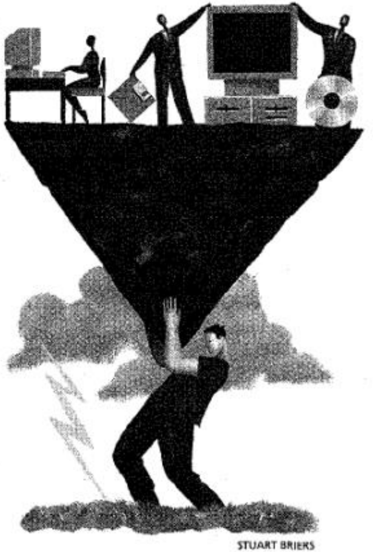
Elder Care

Accessibility

“The last electrical engineer”

“I am worried about the future of our profession. … I see the world as an inverted pyramid. It balances precariously on the narrow point at the bottom. … This point is being impressed into the ground by the heavy weight at the wide top of the inverted pyramid where all the applications reside. … Electrical engineering will be in danger of shrinking into a neutron star of infinite weight and importance, but invisible to the known universe. … Somewhere in the basement of Intel or its successor … the last electrical engineer will sit.”

Bob Lucky
IEEE Spectrum
May 1998



“Computer Science: The ever-expanding sphere”

Credit: Alfred Spector, Google (ret.)

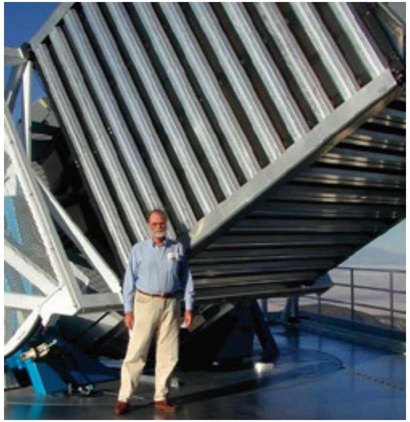
Support for 21st century cyberinfrastructure

- Many fields of discovery are becoming *information* fields, not just computational fields
  - The *intellectual approaches* of Computer Science are as important to advances as is cyberinfrastructure
  - *New approaches* will enable *new discoveries*
  - *"First we do faster … then we do different/smarter/better"*
- Meeting evolving cyberinfrastructure needs requires investment in *intellectual* as well as physical infrastructure
  - We have a crazy obsession with buying shiny objects – the bigger and more expensive, the better

- Nationally and institutionally, there are various policies that distort behavior – *and that should be changed*
  - One example: Use of commercial cloud resources – *essential to cost-effectiveness and scalability* – is discouraged by
    - Indirect cost on outsourced services (and *not* on equipment purchases)
      - *This is totally nuts!*
    - NSF MRI viewed as a pot separate from Directorates/Divisions
    - Institutional subsidies (power, cooling, space)
- We're investing 9:1 in hardware over software[1] – it ought to be the reverse!

[1] According to Ed Seidel when he was at NSF

- We have a dogged resistance to utilizing commercial software, services, and systems
  - We purchase our own
  - We operate our own
  - We roll our own
  - Often with amateurs
  - Why?
    - Outmoded policies
    - Subsidies
    - Defense of turf
    - Politics
    - People whose paychecks depend on convincing you that your needs are so special that no commercial offering could possibly be suitable
    - Failure to do hard-nosed cost-benefit analyses



Can a commercial RDBMS host large-scale science data?

Credit: Werner Vogels, Amazon

- Key attributes of the commercial cloud:
  - Essentially infinite capacity
  - You pay for *exactly* what you use (instantaneous expansion *and* contraction)
  - *Zero* capital cost
  - 1,000 processors for 1 day costs the same (or less) as 1 processor for 1,000 days *(totally revolutionary!)*
  - 7x24x365 operations support, auxiliary power, redundant network connections, geographical diversity
  - For many services, someone else handles backup, someone else handles software updates
  - Sharing and collaboration are easy
  - It continuously gets bigger, faster, less expensive, more capable

Some possible actions

- *Eliminate subsidies (or at least be transparent about them)!*
  - Space, power, cooling, backup, upgrades
- *Eliminate overhead* on outsourced cloud services
- *Attribute NSF MRIs* to Directorates/Divisions
- Take steps to encourage and evolve data-intensive discovery that are *at least as aggressive* as the steps taken decades ago to encourage numerical computational science
- Establish the use of commercial cloud services as *the strong default for science at all scales*. Every request to purchase computing equipment that won't fit on a desktop should be rigorously justified. *Invest in intellectual infrastructure, software infrastructure, and outsourced services, not big shiny objects!*

UW has done this, unilaterally

- *Do not allow* a group without a rock-solid track record to be responsible for the creation of complex mission-critical software infrastructure (e.g., for MREFCs)
- Major national facilities – to the extent that these are necessary at all – should be used *only by applications that truly require them*
- Take additional steps to *encourage reproducible research and the useful/usable sharing of code and data*
- Recognize that *data has both value and cost*. How should the costs be covered?

Is this a great time or what?

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