UNIVERSITY of WASHINGTON

A Plea for Greater Attention to Data-Intensive Discovery, Greater Investment in Intellectual and Software Infrastructure, and Greater Use of the Commercial Cloud

Remarks to the CSTB Committee on Future Directions for NSF Advanced Computing Infrastructure to Support US Science in 2017-2020

Ed Lazowska

Bill & Melinda Gates Chair in Computer Science & Engineering

and

Founding Director of the eScience Institute

University of Washington

December 2014

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

University of Washington Computer Science & Engineering

UNIVERSITY of WASHINGTON eScience Institute

Relevant biographical information

- A.B., 1972, Brown Univ., independent concentration in "Non-Numerical Computer Science"; M.Sc., 1974, Ph.D., 1977, Univ. of Toronto, in Computer Science
- Univ. of Washington faculty member since that time
- Relevant national roles
  - Chair of NSF CISE AC (1998-99), DARPA ISAT (2005-06)
  - Co-Chair (with Marc Benioff) of the (late) PITAC, 2003-05
    - Dan Reed was a member and chaired a study
  - Co-Chair (with David E. Shaw) of the PCAST Working Group to review the Federal NITRD Program, 2010
    - Bill Gropp was a member
  - Member of CSTB (1996-2002), DoE Pacific Northwest National Laboratory Fundamental & Computational Sciences Directorate AC (2011-), NASA AC Information Technology Infrastructure Committee (2011-12)
- Founding Director, Univ. of Washington eScience Institute, 2008

This morning

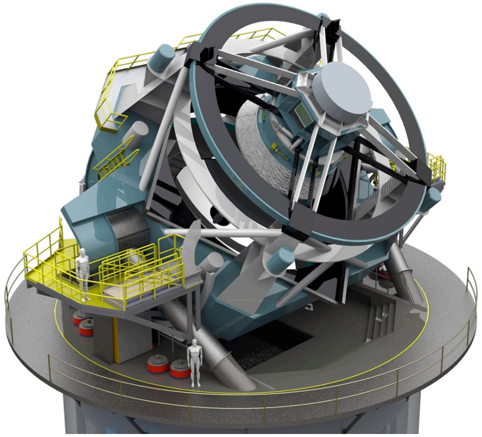
- Data-intensive discovery
- The University of Washington eScience Institute
- Implications for academia
- Implications for research policy
- The commercial cloud
- Some possible actions

[Partially an adaptation of material presented to the National Science Board in October 2013]

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

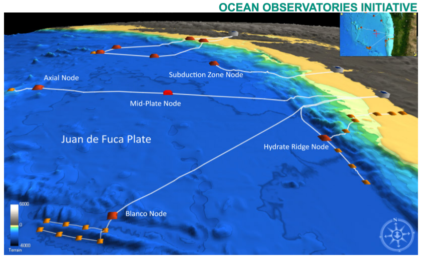
- A proliferation of sensors
- Ever more powerful models producing data that must be analyzed
- The creation of almost all information in digital form
- Dramatic cost reductions in storage
- Dramatic increases in network bandwidth
- Dramatic cost reductions and scalability improvements in computation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

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

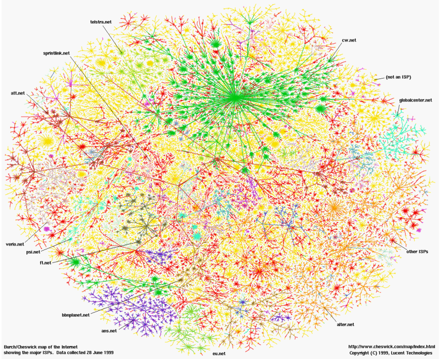


Astronomy: LSST

Physics: LHC



Oceanography: OOI



Sociology: The Web

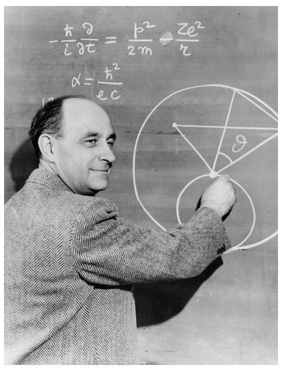
Biology: Sequencing

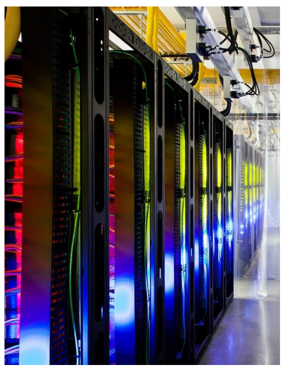
Economics: POS terminals

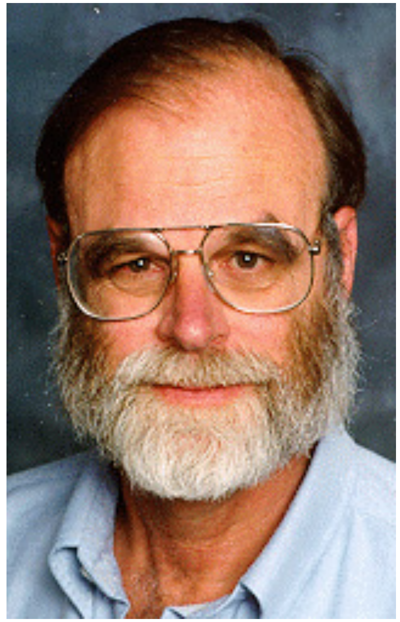
Neuroscience: EEG, fMRI

The Fourth Paradigm

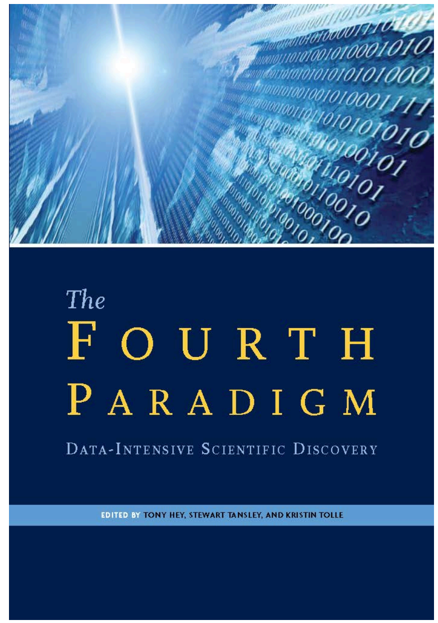
1. Empirical + experimental
2. Theoretical
3. Computational
4. Data-Intensive







Jim Gray

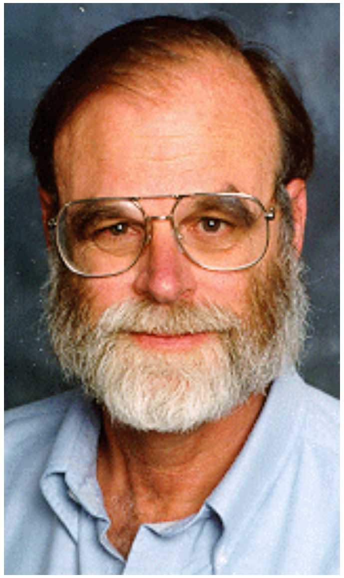


Each augments, vs. supplants, its predecessors – "another arrow in the quiver"

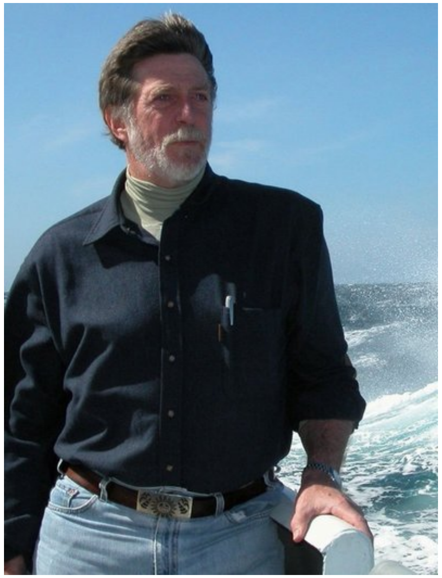
"From data to knowledge to action"

- 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

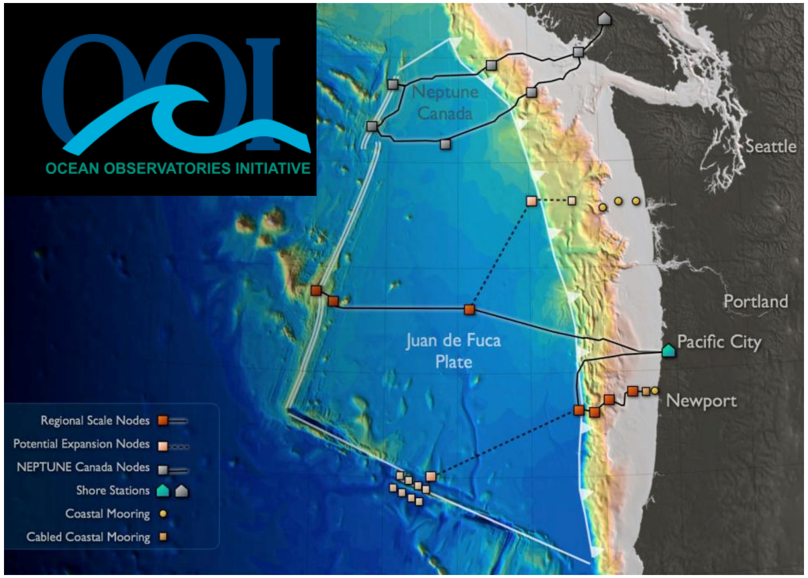
My personal story, and the story of the UW eScience Institute

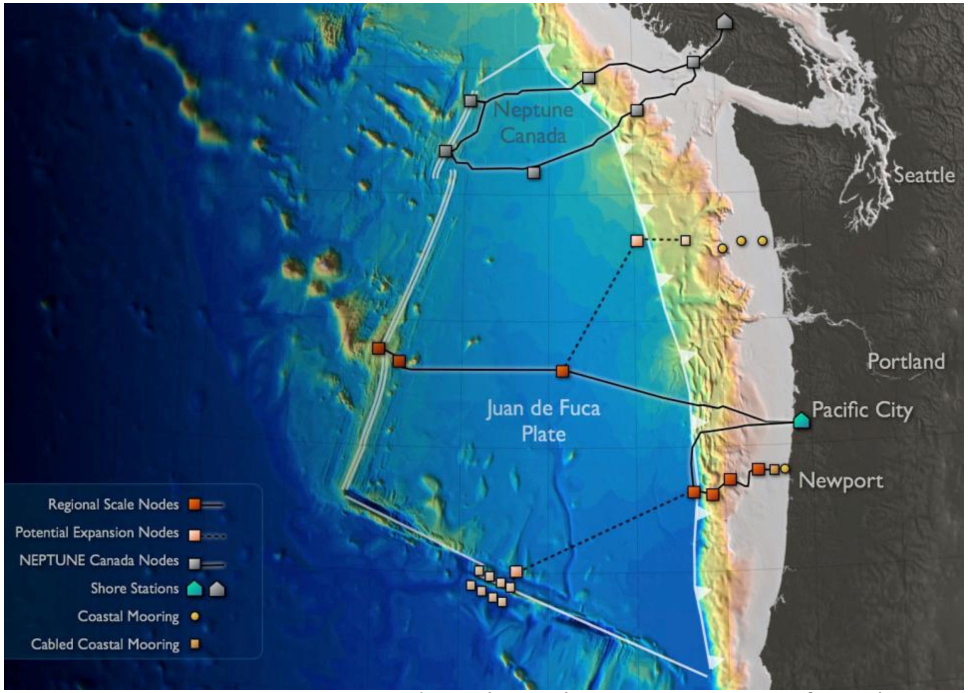


Early 1980s



Late 1990s





Credit: John Delaney, University of Washington

Mark Emmert

LSU

2004

"When I was at LSU I porked me a supercomputer center. I was thinking I'd do that here."

Ed Lazowska, Computer Science & Engineering

Tom Daniel, Biology

Werner Stuetzle, Statistics

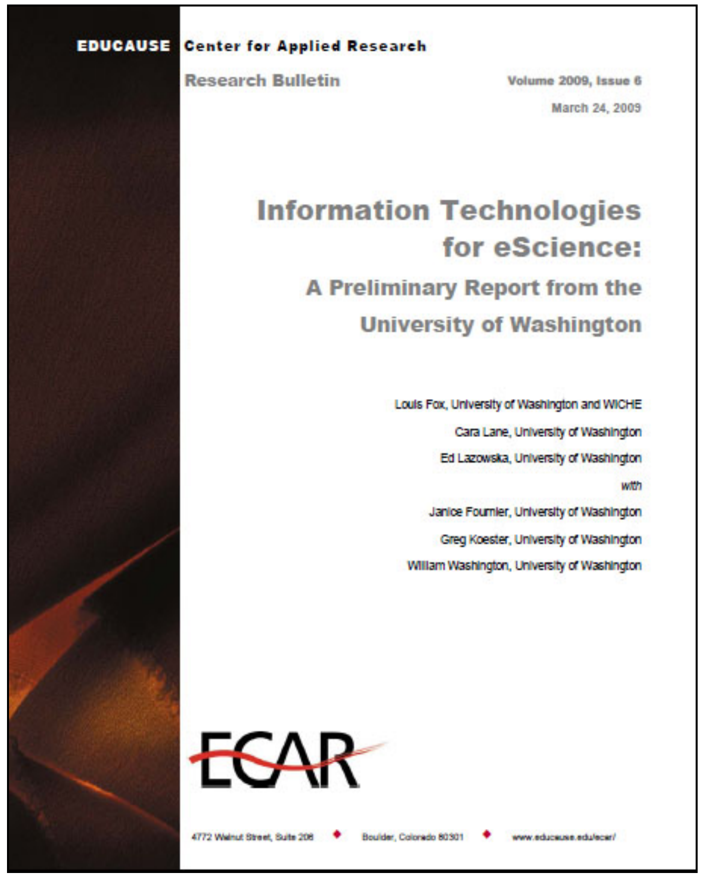
UW eScience Institute

- *"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."*

UNIVERSITY of WASHINGTON
eScience Institute

This was not as obvious ~2006 as it is today

- But we asked UW's leading faculty – across all ages and fields, and regardless of "label" – and they confirmed this view of the future
  - From its inception, this effort has been bottom-up, needs-based, grass-roots, driven by the scientists
- There was vociferous national knuckle-dragging until several years after the 2010 PCAST report
- Low-level University of Washington knuckle-dragging continues to this day



EDUCAUSE Center for Applied Research

Research Bulletin

Volume 2009, Issue 6

March 24, 2009

Information Technologies for eScience:

A Preliminary Report from the University of Washington

Louis Fox, University of Washington and WICHE

Cara Lane, University of Washington

Ed Lazowska, University of Washington

with

Janice Fournier, University of Washington

Greg Koester, University of Washington

William Washington, University of Washington

ECAR

4772 Walnut Street, Suite 206 ♦ Boulder, Colorado 80301 ♦ www.educause.edu/ecar/

UNIVERSITY of WASHINGTON

eScience Institute

ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

- 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)

UNIVERSITY of WASHINGTON

NSF

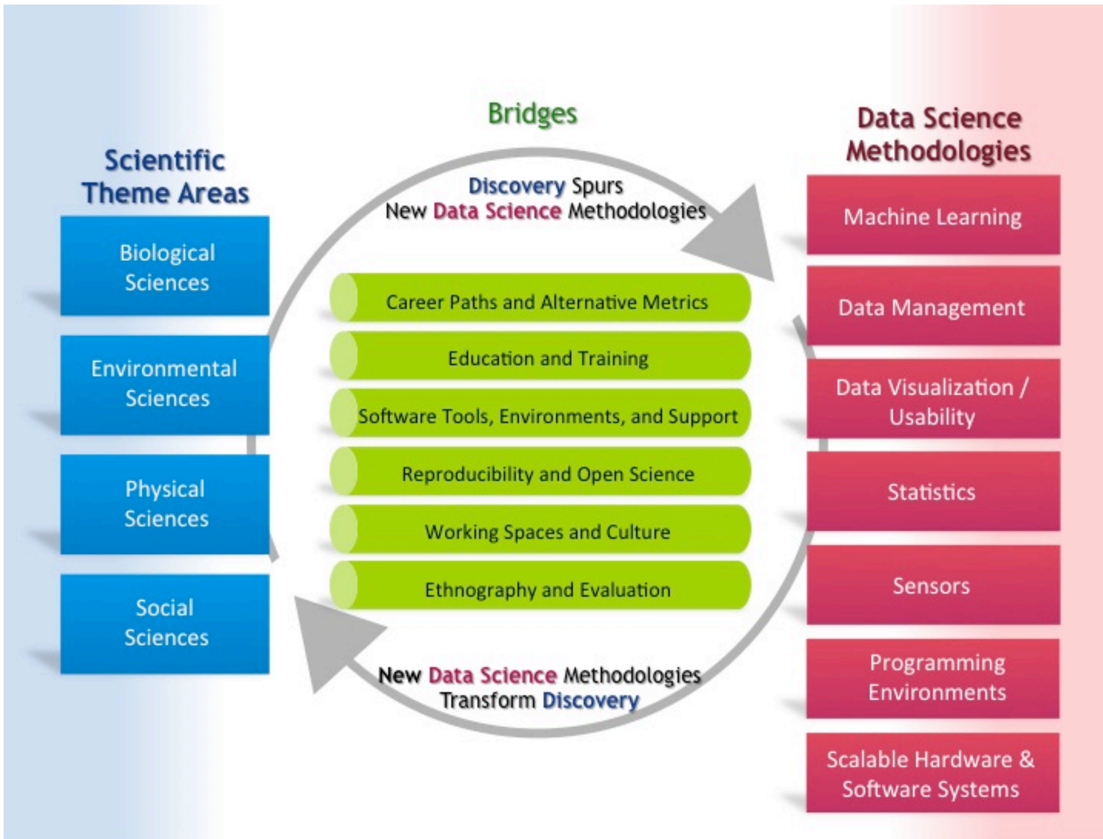
GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION 1934

Washington Research FOUNDATION

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



UNIVERSITY of WASHINGTON

Berkeley UNIVERSITY OF CALIFORNIA

NYU

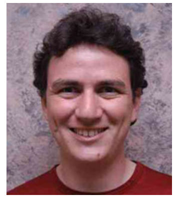
Original core faculty team

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


Thom. Richardson
Statistics, CSSS


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

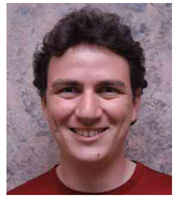
Original core faculty team

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


Thom. Richardson
Statistics, CSSS

Biological sciences


David Beck
Chemical Engr.


Tom Daniel
Biology


Bill Noble
Genome Sciences

Environmental sciences


Ginger Armbrust
Oceanography


Werner Stuetzle
Statistics

13 Departments

5 Schools / Colleges

Social sciences


Josh Blumenstock
iSchool


Mark Ellis
Geography


Tyler McCormick
Sociology, CSSS

Physical sciences

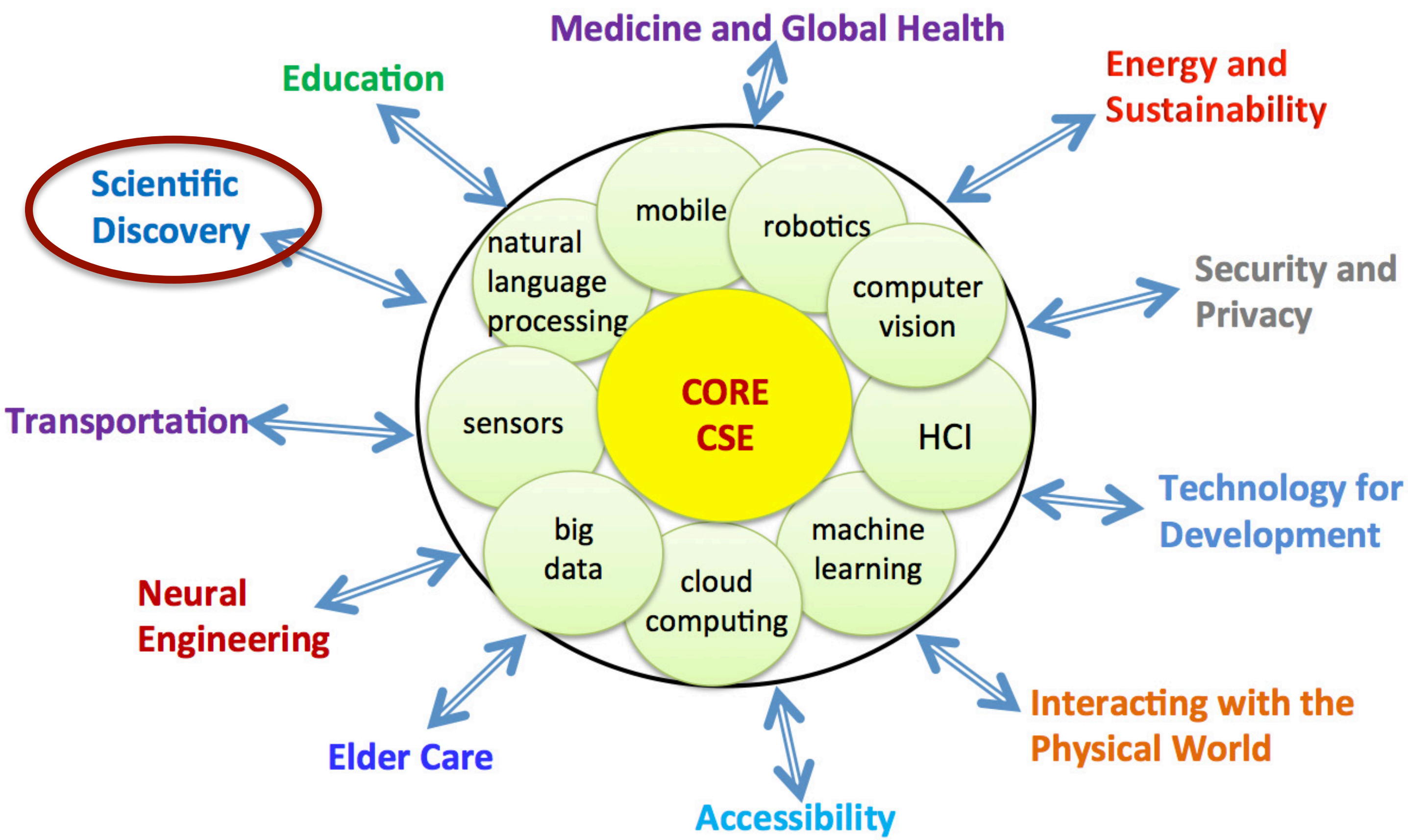

Andy Connolly
Astronomy


John Vidale
Earth & Space Sciences

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

Implications for academia

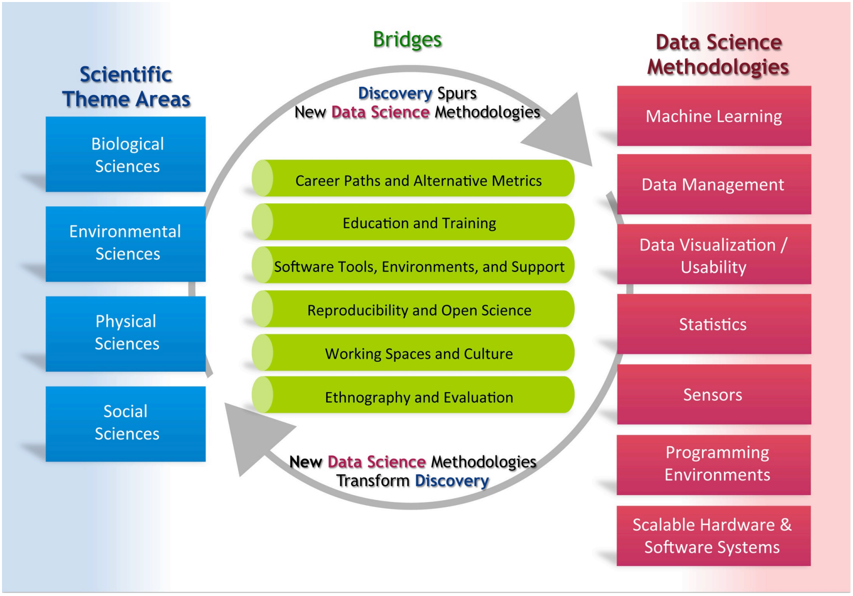
- Computer Science is a field that is unique in its societal impact



Implications for research policy

- NSF has a unique role in driving advances in Computer Science
  - Computer Science does not have an NIH or a Department of Energy
  - NSF provides 75% of Federal support for academic Computer Science research
- Other fields 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 later we do different/smarter/better"*
- Meeting evolving cyberinfrastructure needs requires research, not merely procurement
  - This is true for HPC … and for data-intensive discovery … and for cyber-enabled advances in education and assessment

- 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!
- Advancing data-intensive discovery requires broad-based programs that strive to create a "virtuous cycle" – and that drive institutional change



- Nationally and institutionally, there are various policies that distort behavior – *and that should be changed*
  - One example: Use of commercial cloud resources is discouraged by
    - Indirect cost on outsourced services (and *not* on equipment purchases)
      - *This is totally nuts!*
    - 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!

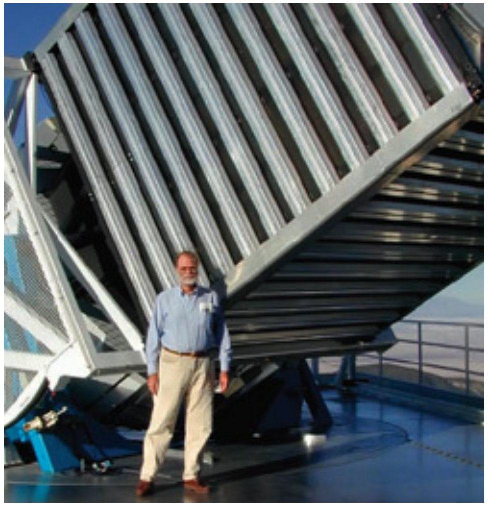
[1] According to Ed Seidel when he was at NSF

The commercial cloud

amazon web services™

Google™

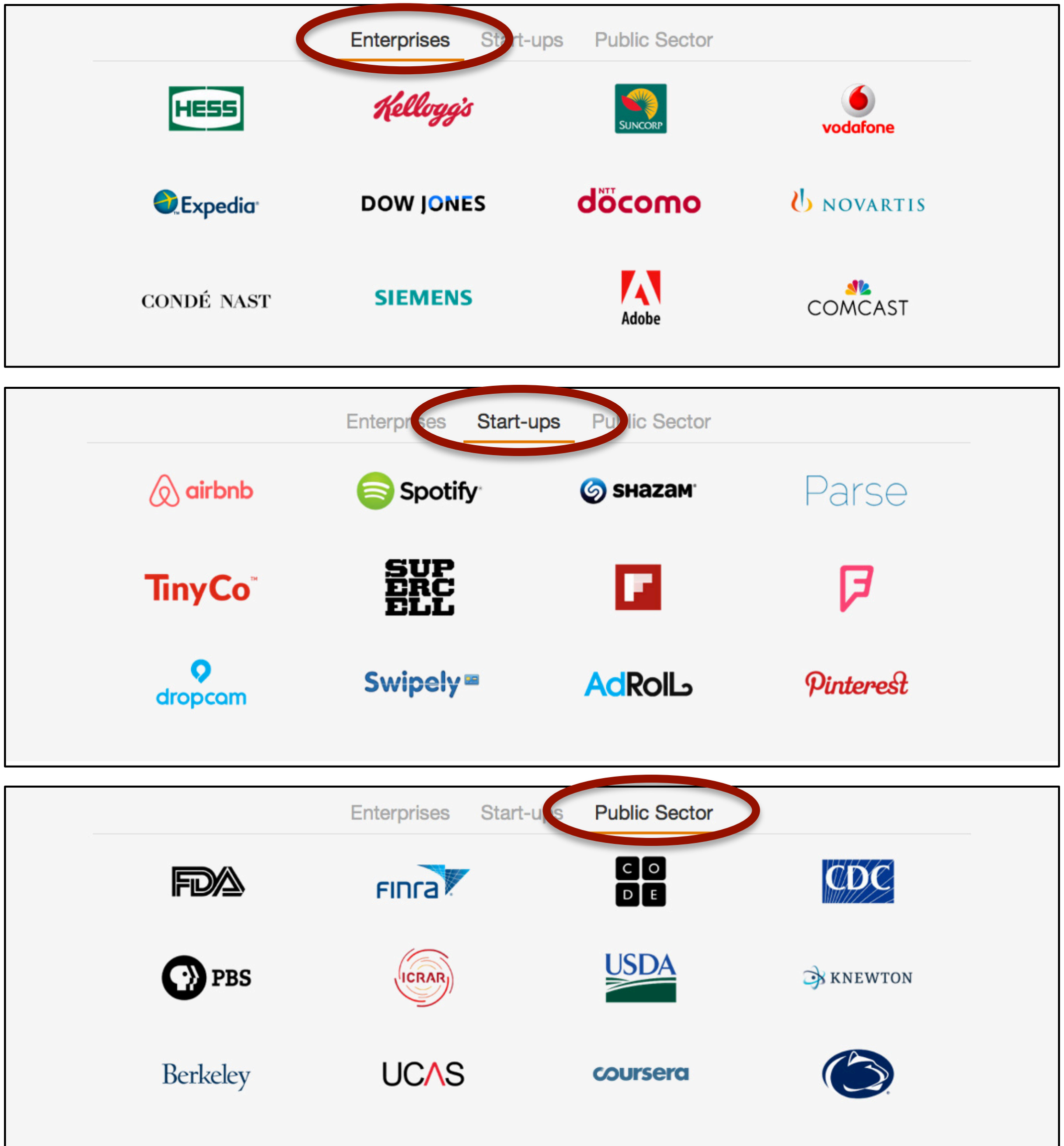
We have a dogged resistance to utilizing commercial software, services, and systems



Can a commercial RDBMS host large-scale science data?

- 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

Some Amazon customers

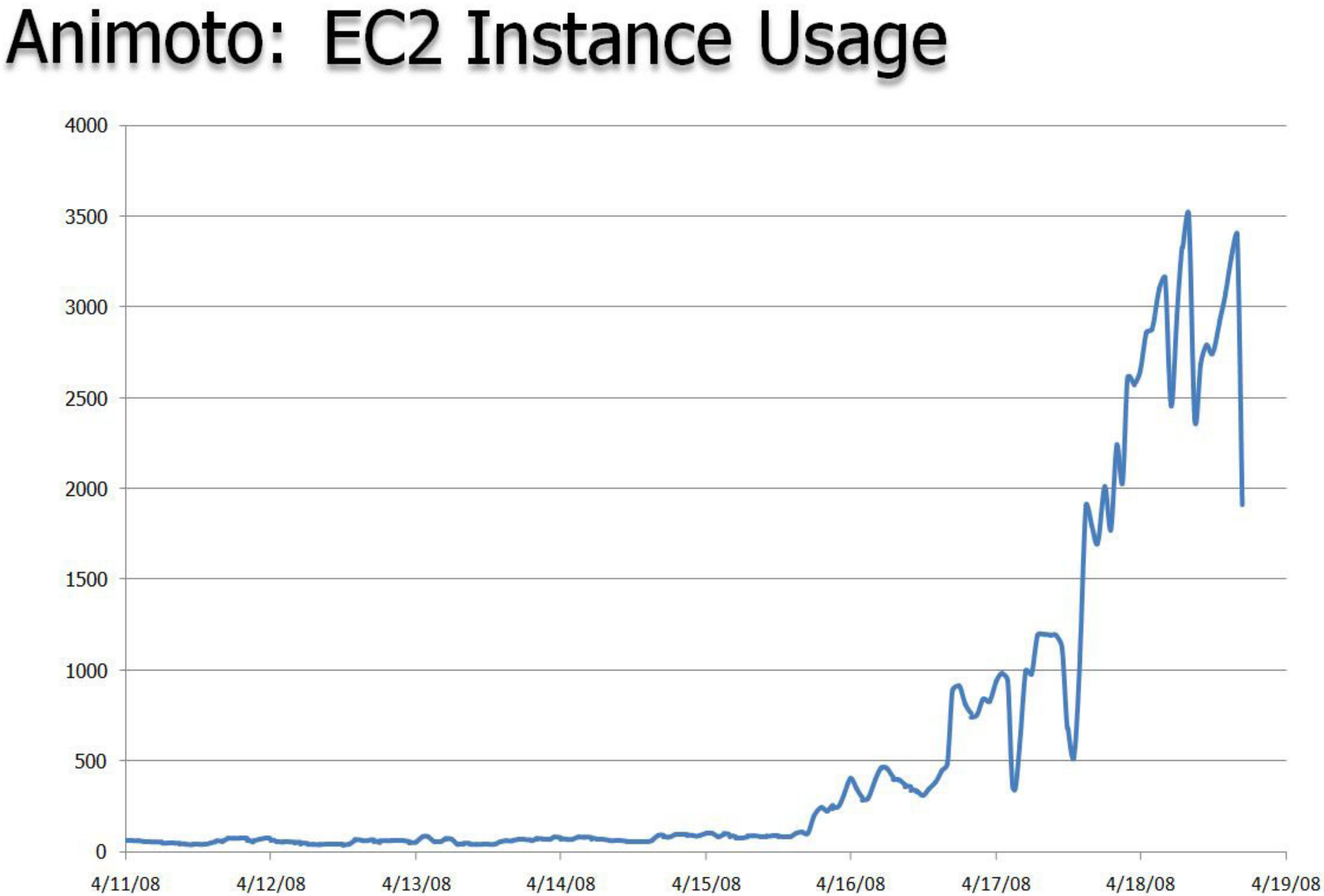


What's so special about our requirements, compared to theirs, that causes us to doggedly adhere to the old world?

Key attributes of the commercial cloud

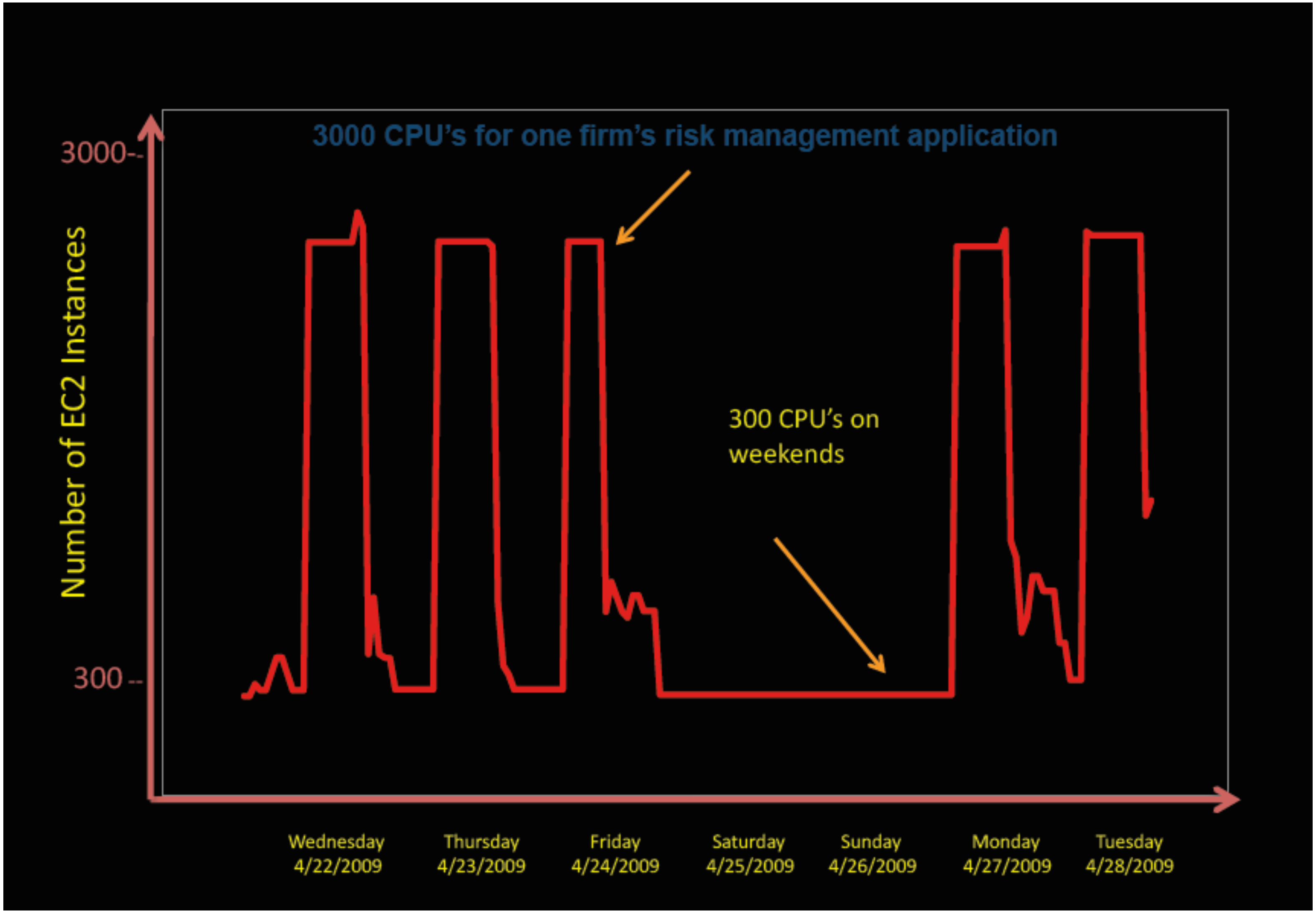
- Essentially infinite capacity
- You pay for *exactly* what you use (instantaneous expansion *and* contraction)
- *Zero* capital cost
- 1,000 processors for 1 day costs the same (or less) as 1 processor for 1,000 days *(totally revolutionary!)*
- 7x24x365 operations support, auxiliary power, redundant 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

Instantaneous expansion, effectively without limit



Credit: Werner Vogels, Amazon

Instantaneous contraction, too



Credit: Deepak Singh, Amazon, via Bill Howe, University of Washington

One-time use on a massive scale for special projects

SECTIONS HOME SEARCH The New York Times

OPEN
All the Code That's Fit to `printf()`

Self-Service, Prorated Supercomputing Fun!

By DEREK GOTTFRID NOVEMBER 1, 2007 5:30 PM

Magic EC2 / S3 Button

As part of eliminating TimesSelect, The New York Times has decided to make all the public domain articles from 1851–1922 available free of charge. These articles are all in the form of images scanned from the original paper. In fact from 1851–1980, all 11 million articles are available as images in PDF format. To generate a PDF version of the article takes quite a bit of work — each article is actually composed of numerous smaller TIFF images that need to be scaled and glued together in a coherent fashion.

Generating pdfs of 11 million articles – 2007

SECTIONS HOME SEARCH The New York Times

OPEN
All the Code That's Fit to `printf()`

The New York Times Archives + Amazon Web Services = TimesMachine

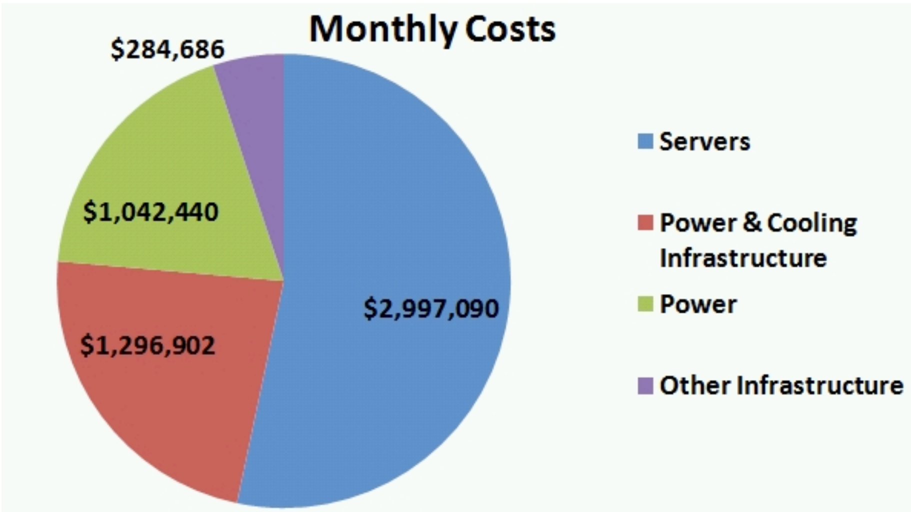
By DEREK GOTTFRID MAY 21, 2008 5:29 PM

Not long ago I posted about a little trip we took through the world of self-service prorated super computing in order to scale out our digital archives. Well, we had so much fun the first time around, we decided to have another go. This time, we've generated a browsable interface to the mounds and mounds of interesting data that lay in our archives.

Generating a browsable interface to all archived data – 2008

Much research computing has similar characteristics

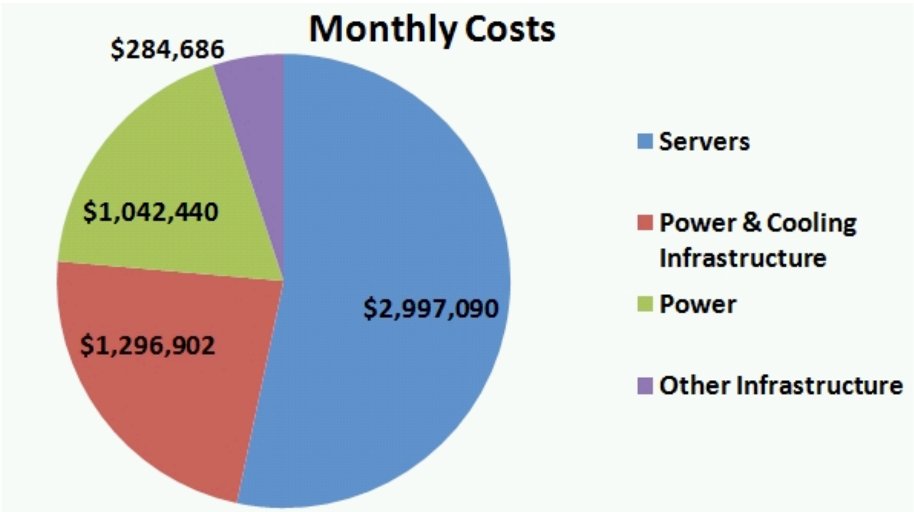
- Bursts, with intervening lulls
- Massive commercial cloud services can accommodate this enormous variability
  - Tremendous scale, plus auction approaches such as AWS Spot Instances
- Keeping the infrastructure busy is important because it's the predominate cost



[3 yr amortization for servers, 15 yr for power, cooling, and other physical infrastructure]

Credit: James Hamilton, Amazon, via Bill Howe, University of Washington

[An aside: What does the pie chart say about the sense of charging overhead on outsourced cloud services but not on equipment purchases?]



[3 yr amortization for servers, 15 yr for power, cooling, and other physical infrastructure]

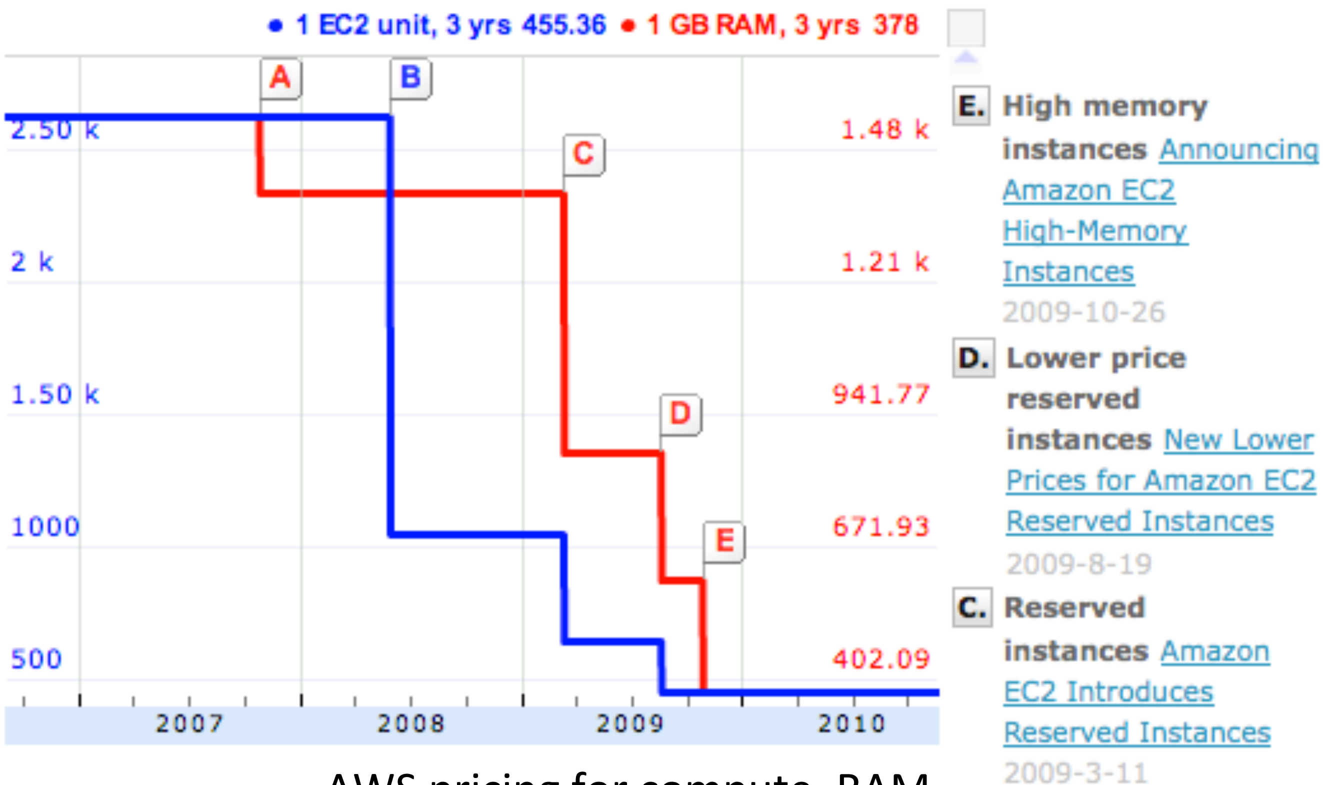
Credit: James Hamilton, Amazon, via Bill Howe, University of Washington

- Additionally, there are tremendous economies of scale to be had:

Technology	Cost in Medium-sized DC	Cost in Very Large DC	Ratio
Network	$95 per Mbit/sec/month	$13 per Mbit/sec/month	7.1
Storage	$2.20 per GByte / month	$0.40 per GByte / month	5.7
Administration	^{3}140 Servers / Administrator	>1000 Servers / Administrator	7.1

Credit: Armbrust, et al., *Above the Clouds: A Berkeley View of Cloud Computing*, 2009, via Bill Howe, University of Washington

Commercial cloud costs drop continuously, without the customer lifting a finger (or a wallet) – in stark contrast to purchased equipment



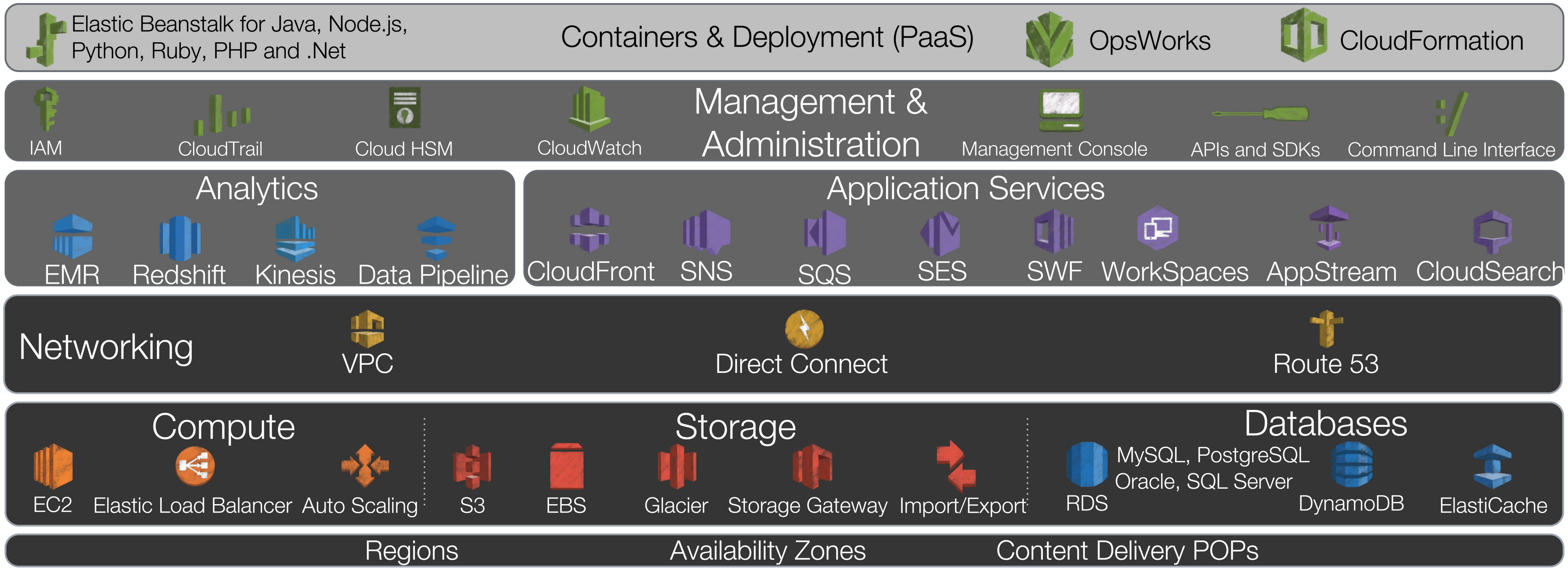
AWS pricing for compute, RAM
[old data – for illustrative purposes only!]

AWS pricing for storage (1TB, 1PB)

[old data – for illustrative purposes only!]

Credit: Bill Howe, University of Washington

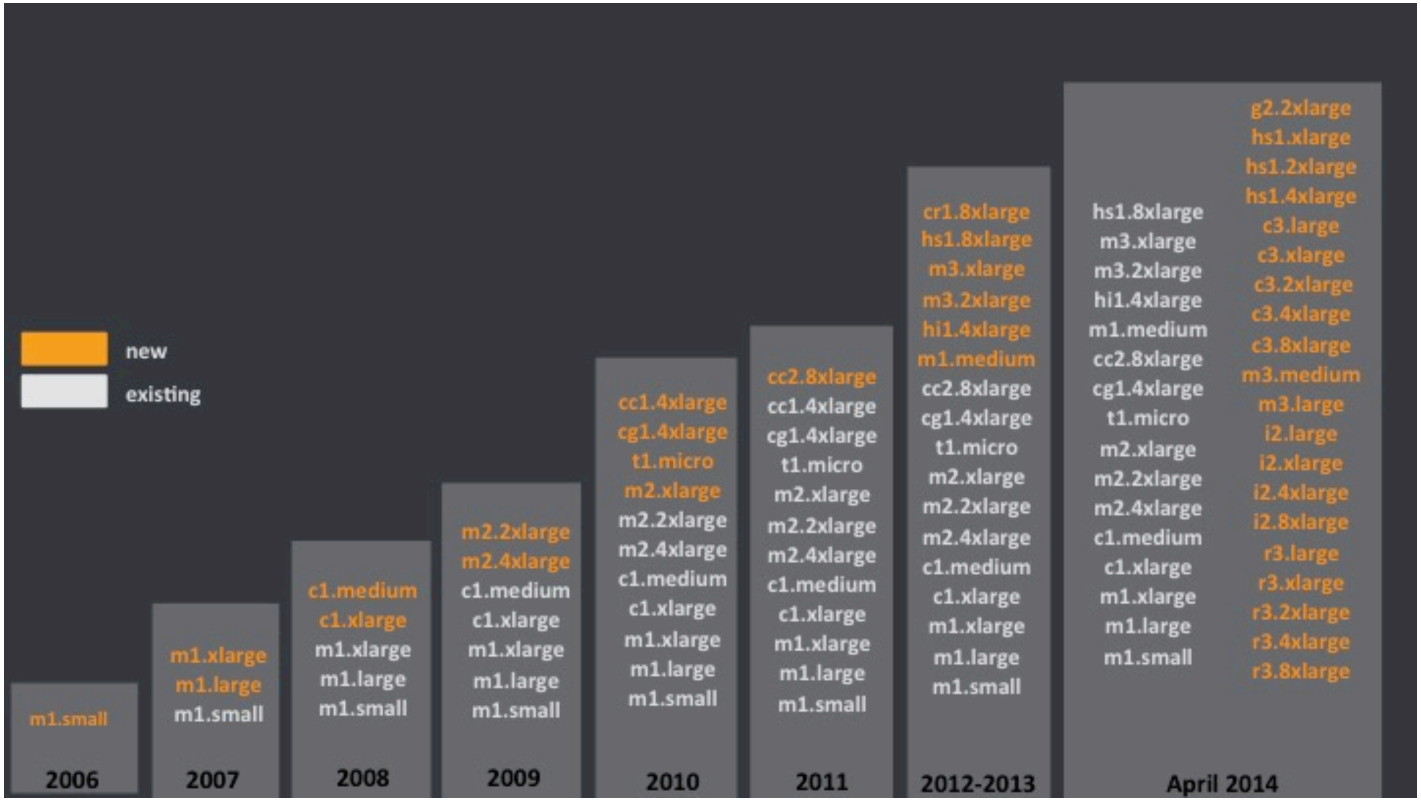
Capabilities evolve at a rapid pace



AWS Stack

Credit: Jamie Kinney, Amazon

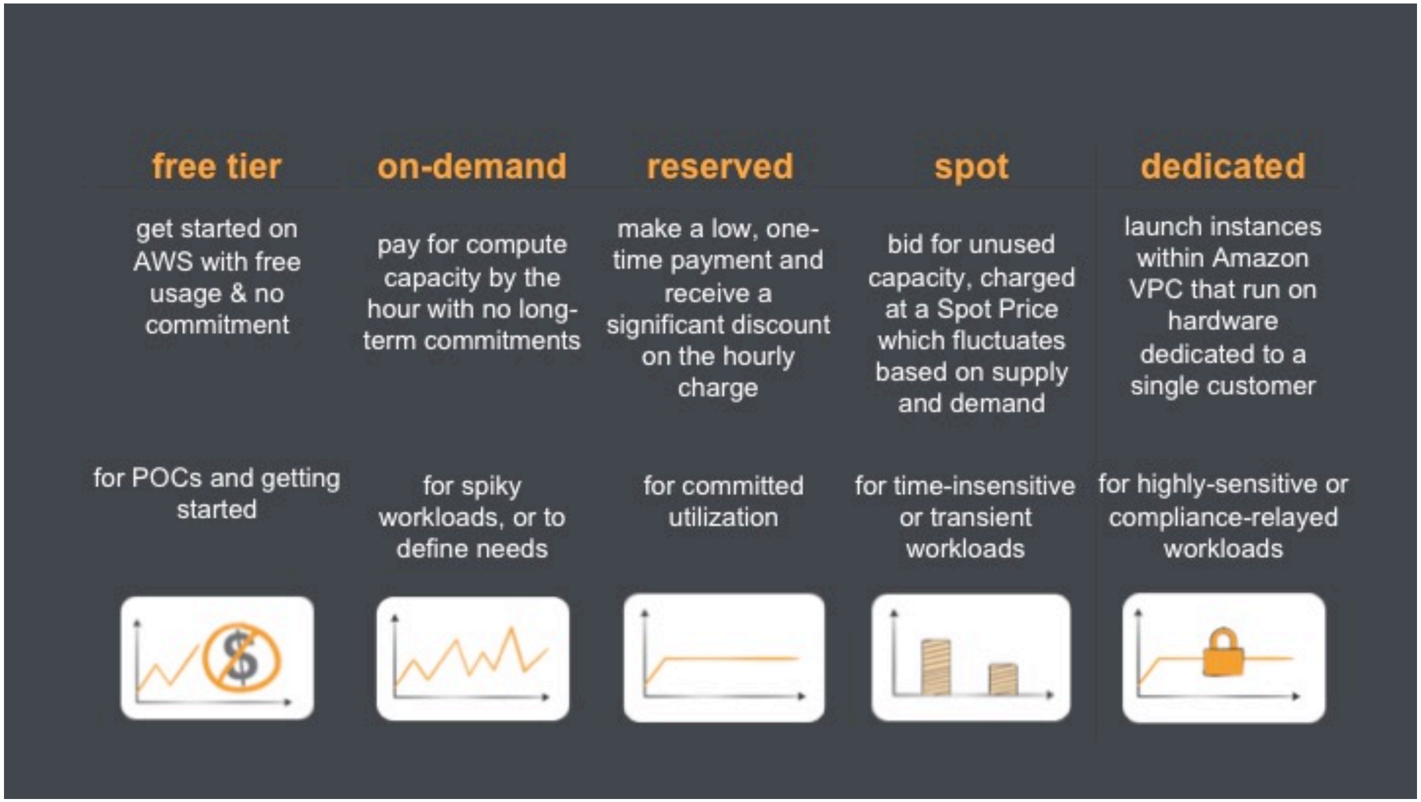
Choices evolve at a rapid pace



AWS EC2 Instance Type History

Credit: Jamie Kinney, Amazon

Purchase models evolve at a rapid pace



AWS Purchase Models

Credit: Jamie Kinney, Amazon

Competition is growing at a rapid pace

amazon web services™

Google™

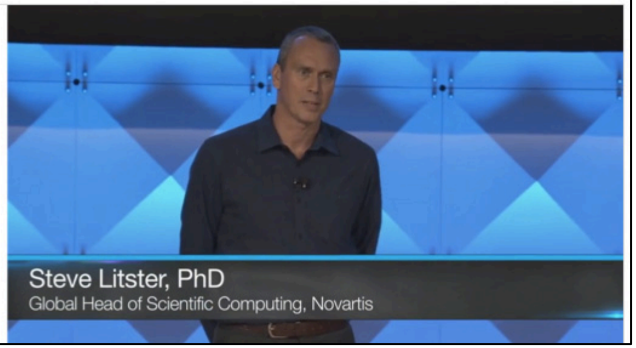
Including competition for academic and commercial science workloads and datasets!

"We completed the equivalent of 39 years of computational chemistry in just under 9 hours."

Steve Litster, Ph.D., Global Head of Scientific Computing, Novartis

NOVARTIS

Watch the video »



"Pfizer did not have to invest in additional hardware and software, which is only used during peak loads; that savings allowed for investments in other [R&D] activities."

Dr. Michael Miller, Head of High Performance Computing for R&D, Pfizer

Pfizer

Read the case study »

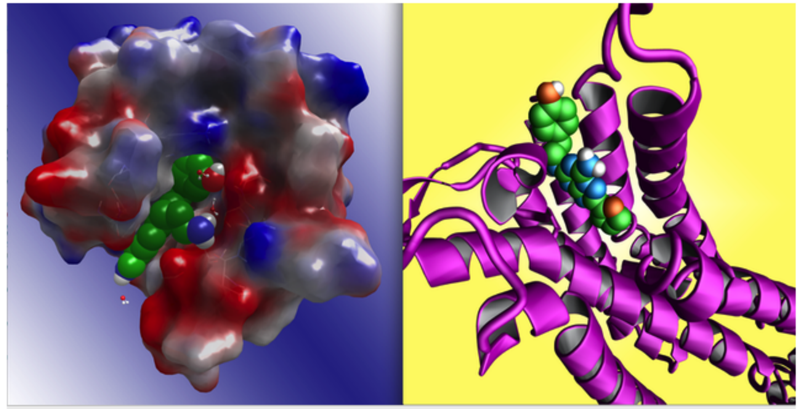
IT PRODUCTIVITY INCREASE: 52%

AVERAGE SAVINGS PER APPLICATION: $518,990

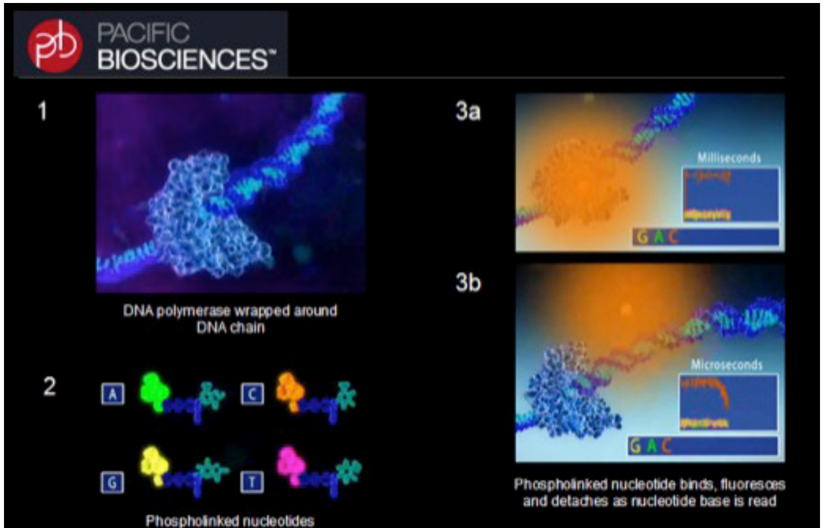
$4,829-per-hour supercomputer built on Amazon cloud to fuel cancer research

By Jon Brodkin | Published 2 days ago

[~50K cores, ~6.7K EC2 instances]



Simulated images of compounds studied in pharmaceutical research

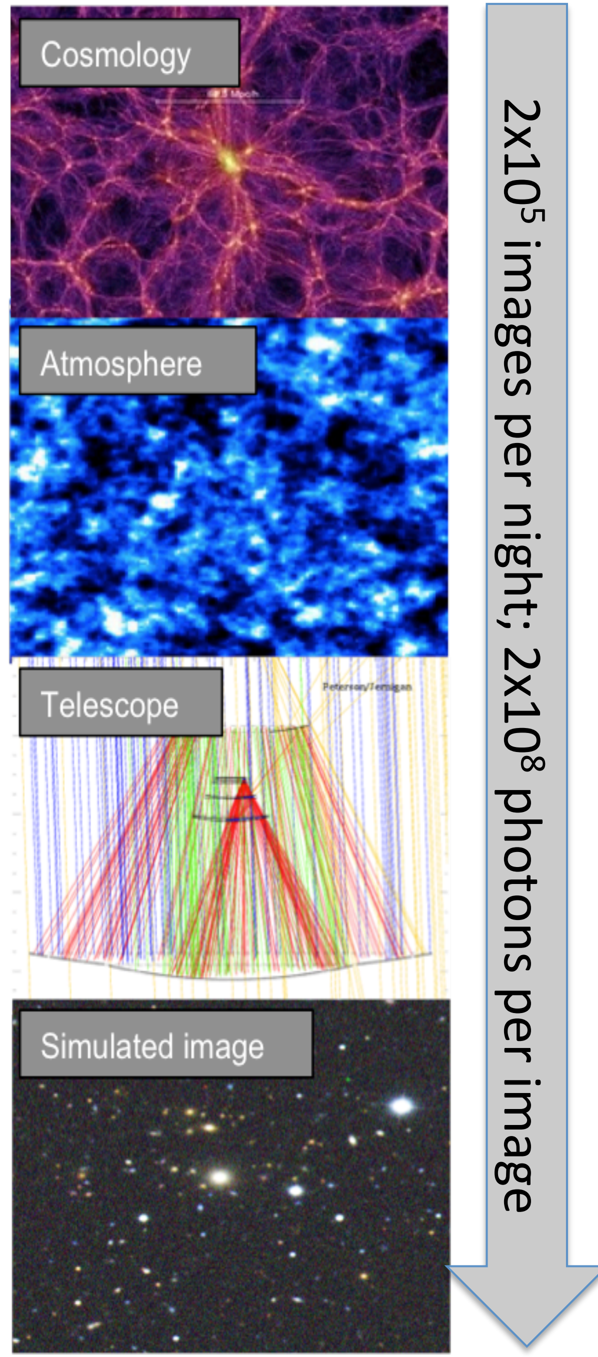


The Google Cloud Platform delivered 405,000 core-hours in a single day to align the long reads to each other

Note: Many of these workloads are compute-intensive, not data-intensive!

UW, LSST, and Google's Exacycle Program

- Harvesting spare cycles on the Google commercial infrastructure
- Designed for small footprint Monte Carlo simulations (<40 min runtime) that scale to 300,000+ processors
- Projects: antibiotic drug resistance, G protein-coupled receptors, simulating astronomical instrumentation
- Impact: *elastic nature of the service* more than total CPU hrs (scale from thousands to hundreds of thousands of processors).



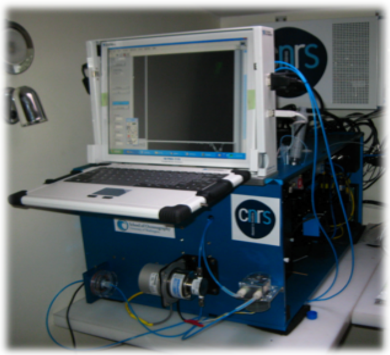
Ability to simulate LSST data (15TB/night) in real time

UW Database-as-a-Service for Science using Microsoft's Azure SQL Database

- Databases are underused in science – hard to design a "permanent" database for a fast-moving research target
- Approach: Wrap a cloud-hosted DB with an easy web interface – focus on getting work done rather schema design pedantry
- We're seeing changes in behavior:
  - In biology, non-programmers are writing 40-line SQL queries with zero training
  - In oceanography, 100-line R scripts are replaced with 10 lines of SQL – for a 10x speedup over 10x as much data
  - … many more

SQLSHARE

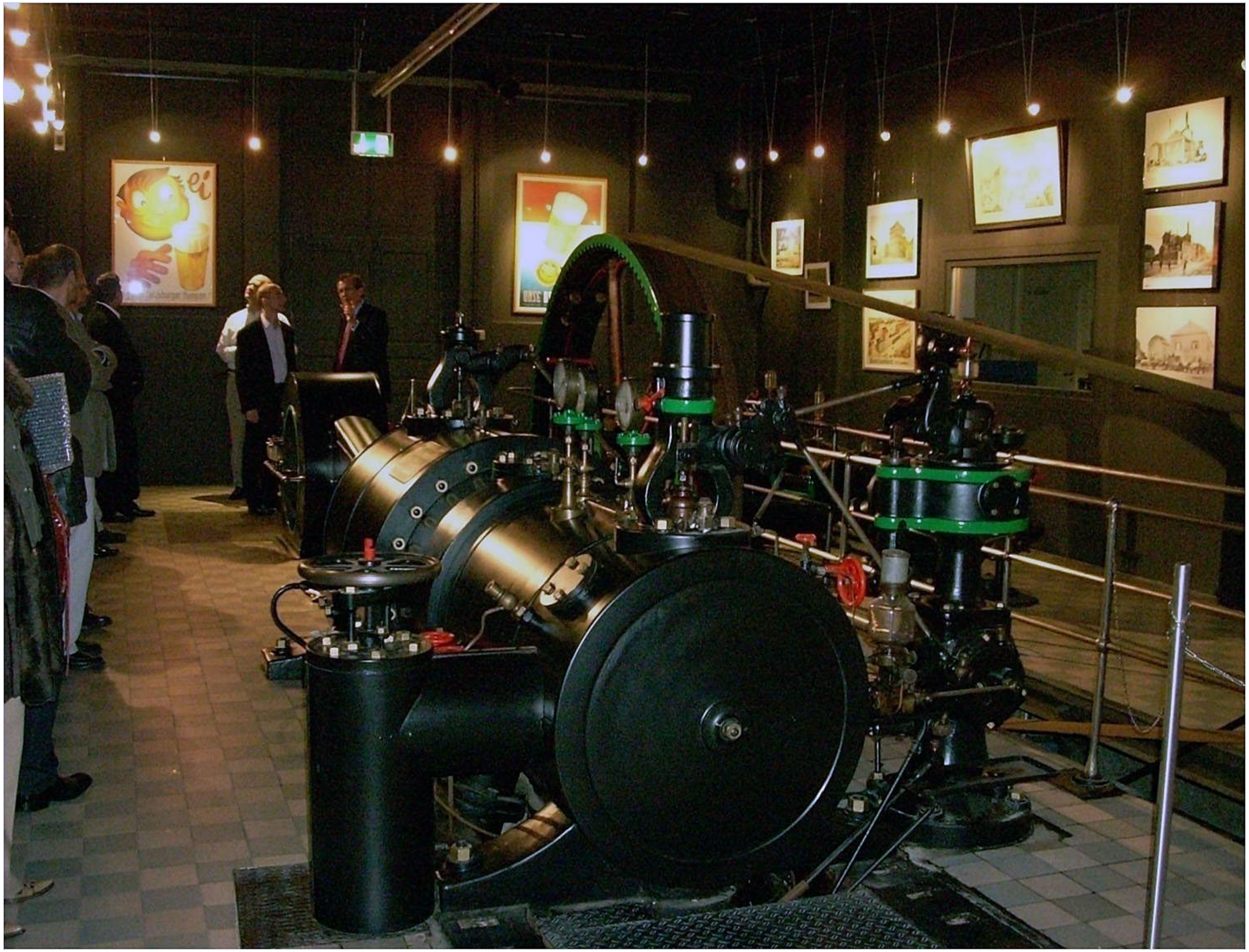
GeoMICS Project: Real-time integration across chemical, physical, and biological oceanography



SeaFlow Project: Continuous Environmental Flow Cytometry

Specific advantages, for science, of the commercial cloud

- Burst capacity
  - Access to many thousands of cores: *"1,000 processors for 1 day costs the same (or less) as 1 processor for 1,000 days"*
- Reproducibility
  - Investigators use the same tools and data – the same exact computational environment
- Sharing and collaboration
  - With zero overhead
- Efficient use of scarce research dollars
  - Avoid investments in infrastructure that's redundant, under-utilized, difficult to share, and has a short lifetime



Credit: Werner Vogels, Amazon

Some possible actions

- *Eliminate overhead* on outsourced cloud services
- *Attribute 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!*

- *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?

UNIVERSITY *of* WASHINGTON

Thanks for inviting me!

University of Washington Computer Science & Engineering

UNIVERSITY *of* WASHINGTON eScience Institute

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