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 Colloquium on

Future Cyberinfrastructure for Scientific Discovery

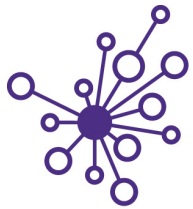
Ed Lazowska

Bill & Melinda Gates Chair in
Computer Science & Engineering

and

Founding Director of the
eScience Institute

University of Washington

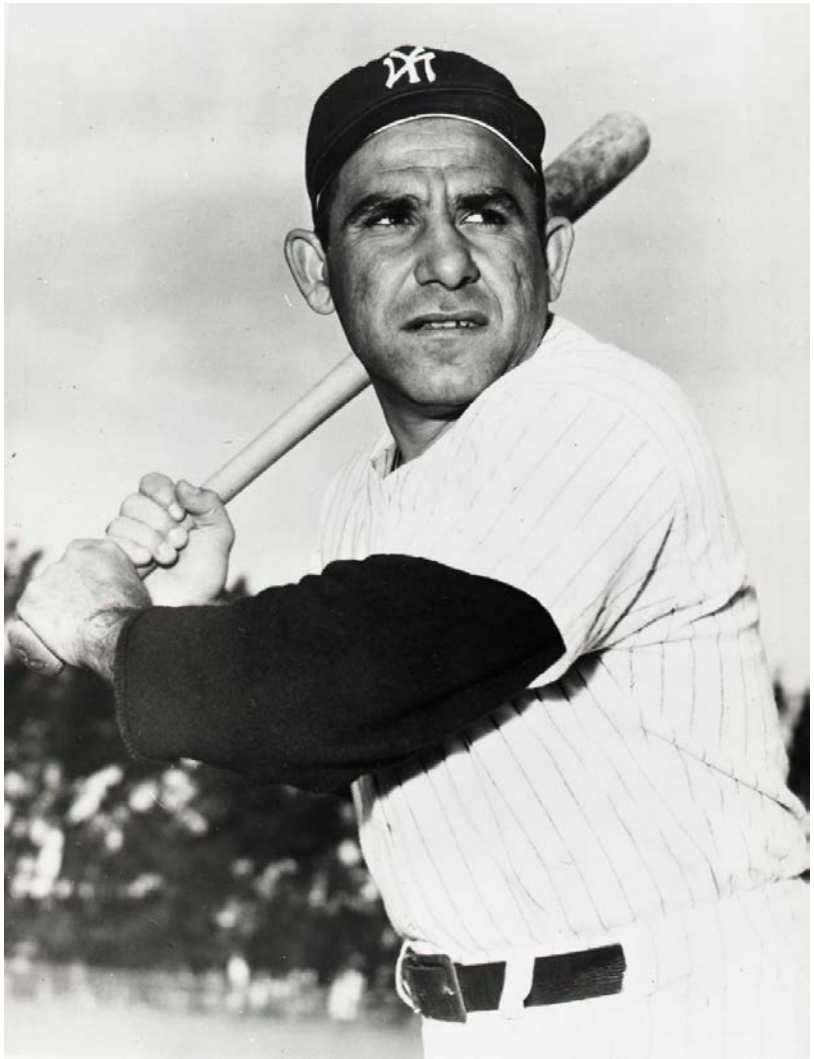


UNIVERSITY of WASHINGTON
eScience Institute

September 2016

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

“It’s déjà vu all over again”



Thoughts on "The Future of Advanced Cyberinfrastructure for Science and Engineering Research and Education"

Ed Lazowska

Bill & Melinda Gates Chair in
Computer Science & Engineering

Founding Director, eScience Institute

University of Washington

National Science Board

September 2013

http://lazowska.cs.washington.edu/NSB.pdf

This morning …

- Why must America remain the world leader in computer science?
- How did we gain the lead, and how can we retain it?
- How should our competitiveness be defined?
- The coming decade: Dramatic improvements in technology and algorithms enable "smart everything"
- Cyberinfrastructure to support 21st century "smart discovery"
  - Implications for academia
  - Implications for research policy
  - Implications for K-12 education

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

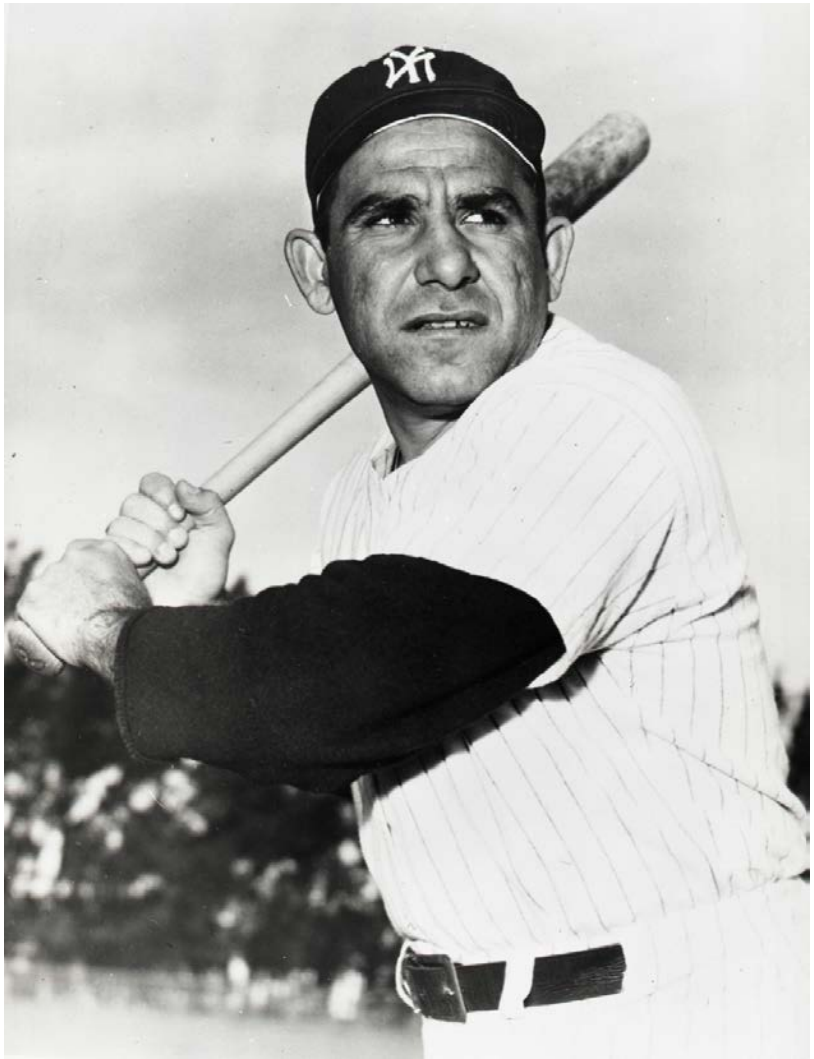
This morning

- 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]

“There you go again”

“It ain’t over ‘til it’s over”



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
  - Co-Chair (with David E. Shaw) of the PCAST Working Group to review the Federal NITRD Program, 2010
  - Member of CSTB (1996-2002), DoE Pacific Northwest National Laboratory Fundamental & Computational Sciences Directorate AC (2009-15), NASA AC Information Technology Infrastructure Committee (2012-13)
- Founding Director, Univ. of Washington eScience Institute, 2008

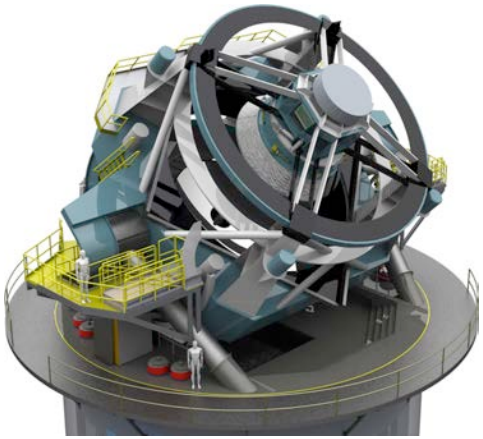
This morning

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

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

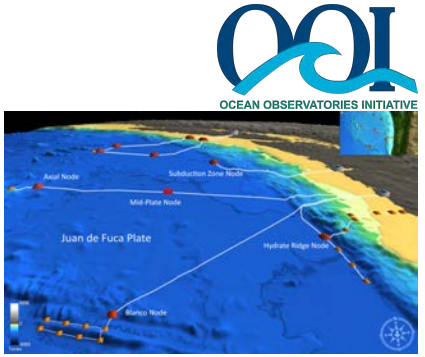
- A proliferation of sensors
- Ever more powerful models producing data that must be analyzed
- The creation of almost all information in digital form
- Dramatic cost reductions in storage
- Dramatic increases in network bandwidth
- Dramatic cost reductions and scalability improvements in computation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

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

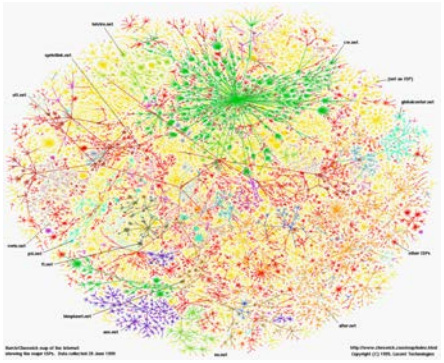


Astronomy: LSST

Physics: LHC



Oceanography: OOI



Sociology: The Web

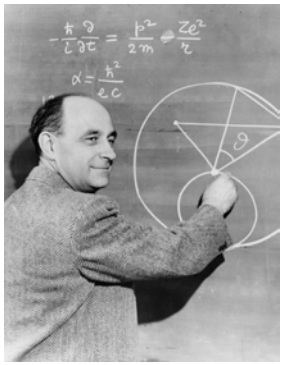
Biology: Sequencing

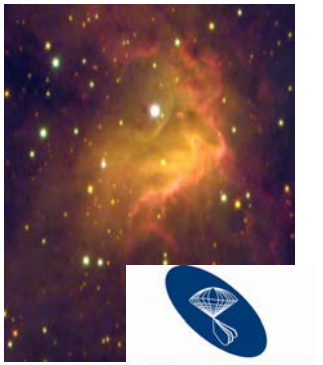
Economics: POS terminals

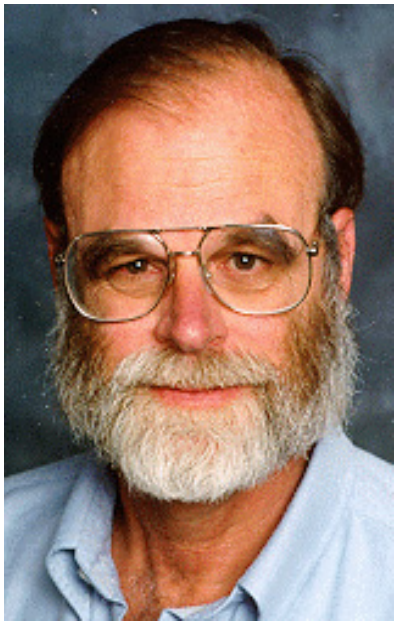
Neuroscience: EEG, fMRI

The Fourth Paradigm

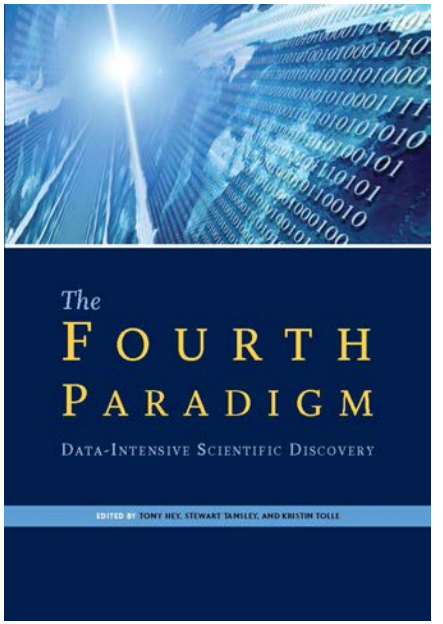
1. Empirical + experimental
2. Theoretical
3. Computational
4. Data-Intensive







Jim Gray

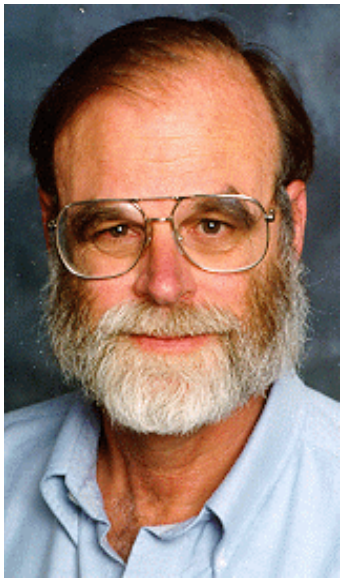


Each augments, vs. supplants, its predecessors – "another arrow in the quiver"

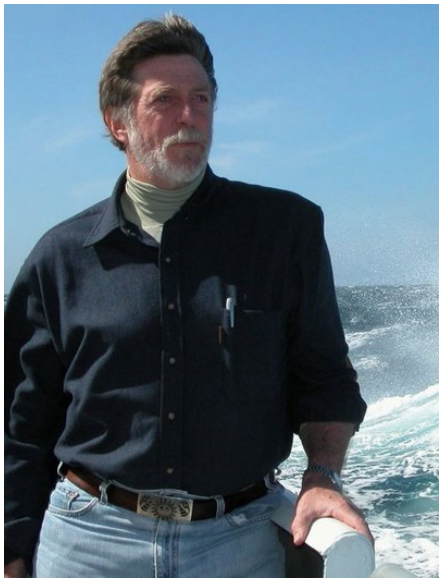
"From data to knowledge to action"

- 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

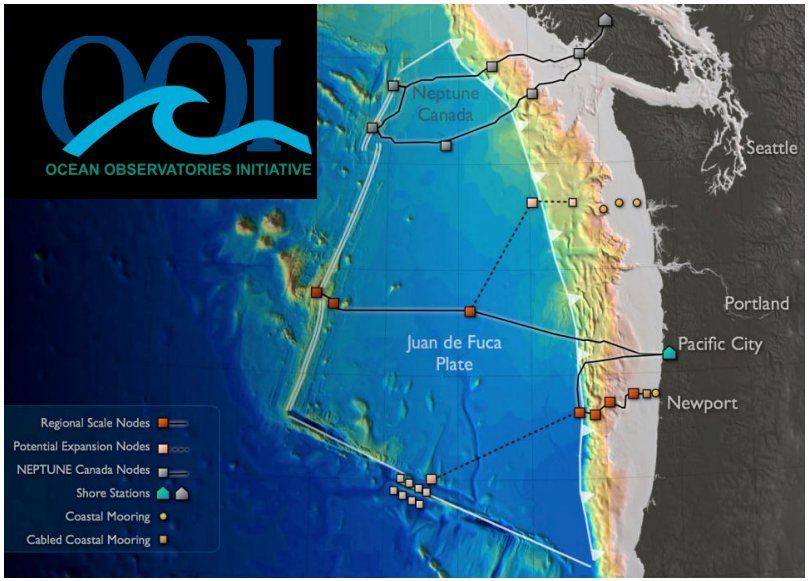
My personal story, and the story of the UW eScience Institute

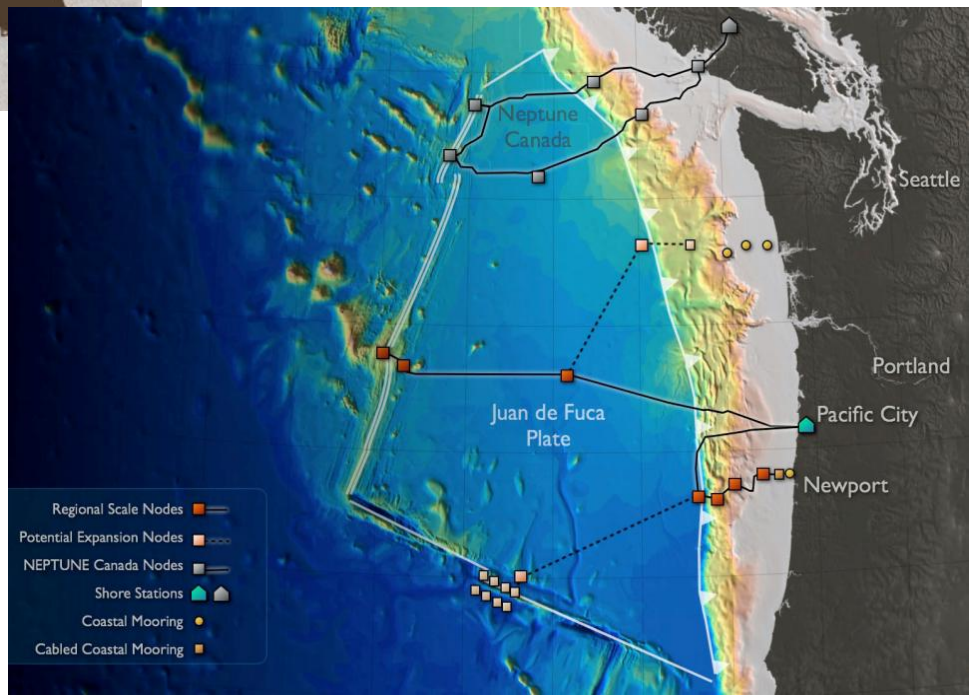


Early 1980s



Late 1990s





Credit: John Delaney, University of Washington

Mark Emmert

LSU

2004

UNIVERSITY OF WASHINGTON 1861 LVX SIT

"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

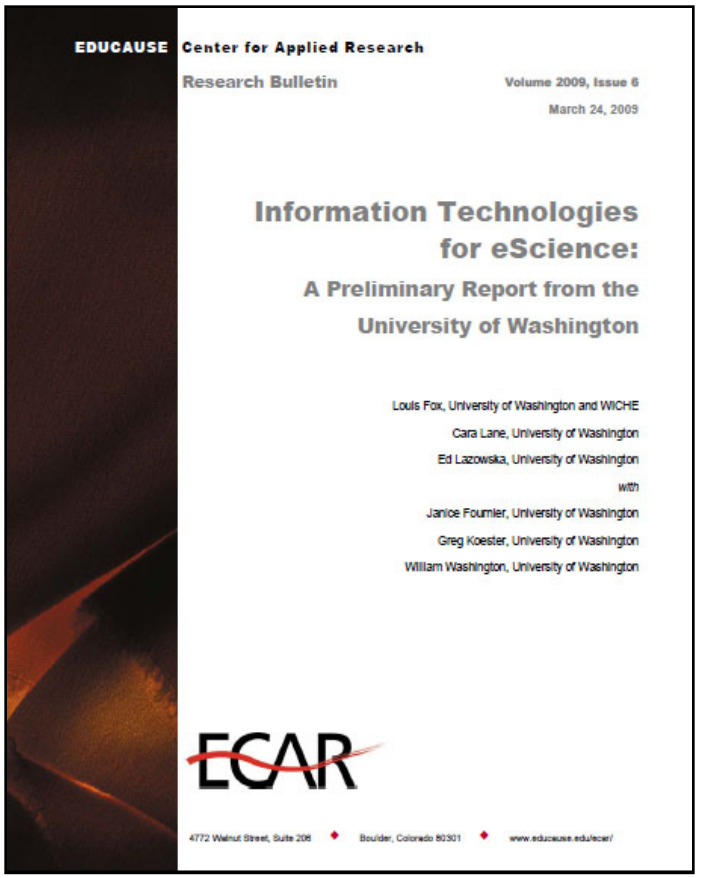
UW eScience Institute

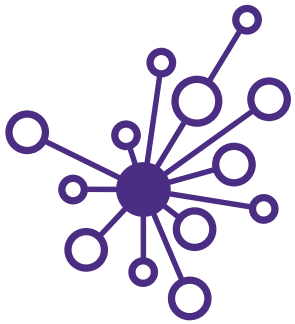
- *"All across our campus, the process of discovery will increasingly rely on researchers' ability to extract knowledge from vast amounts of data... In order to remain at the forefront, UW must be a leader in advancing these techniques and technologies, and in making [them] accessible to researchers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



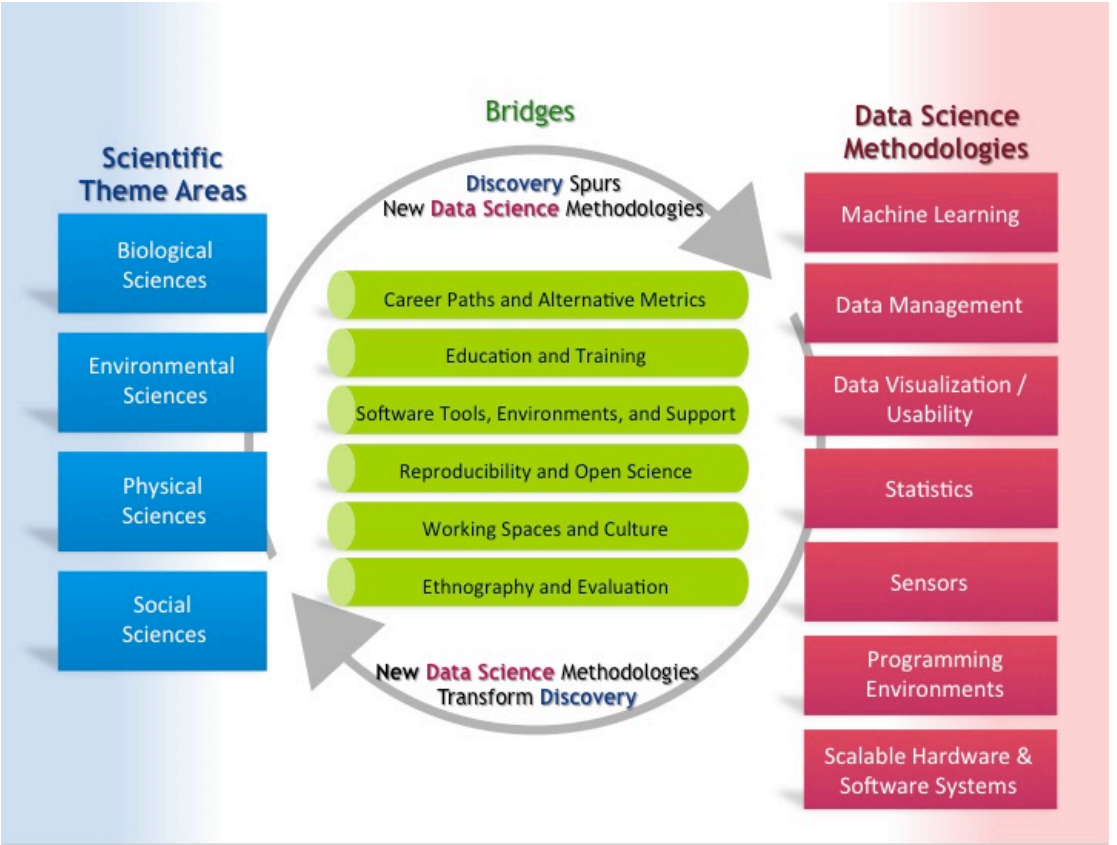


UNIVERSITY of WASHINGTON
eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

- University of Washington
  - $720,000/year for staff support
  - $750,000/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

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

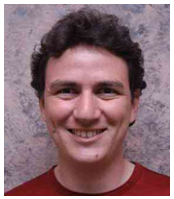
Original core faculty team

Data science methodology


Cecilia Aragon
Human Centered Design & Engr.


Magda Balazinska
Computer Science & Engineering


Emily Fox
Statistics


Carlos Guestrin
CSE


Bill Howe
iSchool


Jeff Heer
CSE


Ed Lazowska
CSE

Biological sciences


David Beck
Chemical Engr.

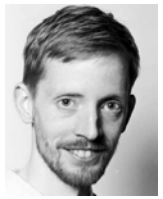

Tom Daniel
Biology


Bill Noble
Genome Sciences

Environmental sciences


Ginger Armbrust
Oceanography


Randy LeVeque
Applied Mathematics


Thom. Richardson
Statistics, CSSS


Werner Stuetzle
Statistics

Social sciences


Josh Blumenstock
iSchool


Mark Ellis
Geography


Tyler McCormick
Sociology, Statistics, CSSS

Physical sciences


Andy Connolly
Astronomy


John Vidale
Earth & Space Sciences

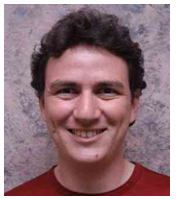
Original core faculty team

Data science methodology


Cecilia Aragon
Human Centered Design & Engr.


Magda Balazinska
Computer Science & Engineering


Emily Fox
Statistics


Carlos Guestrin
CSE


Bill Howe
iSchool


Jeff Heer
CSE


Ed Lazowska
CSE

13 Departments
5 Schools / Colleges

Biological sciences


David Beck
Chemical Engr.

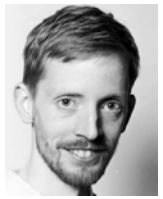

Tom Daniel
Biology


Bill Noble
Genome Sciences

Environmental science


Ginger Armbrust
Oceanography


Randy LeVeque
Applied Mathematics


Thom. Richardson
Statistics, CSSS


Werner Stuetzle
Statistics

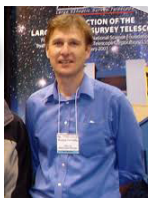
Social sciences

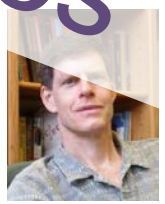

Josh Blumenstock
iSchool


Mark Ellis
Geography


Tyler McCormick
Sociology, Statistics, CSSS

Physical sciences

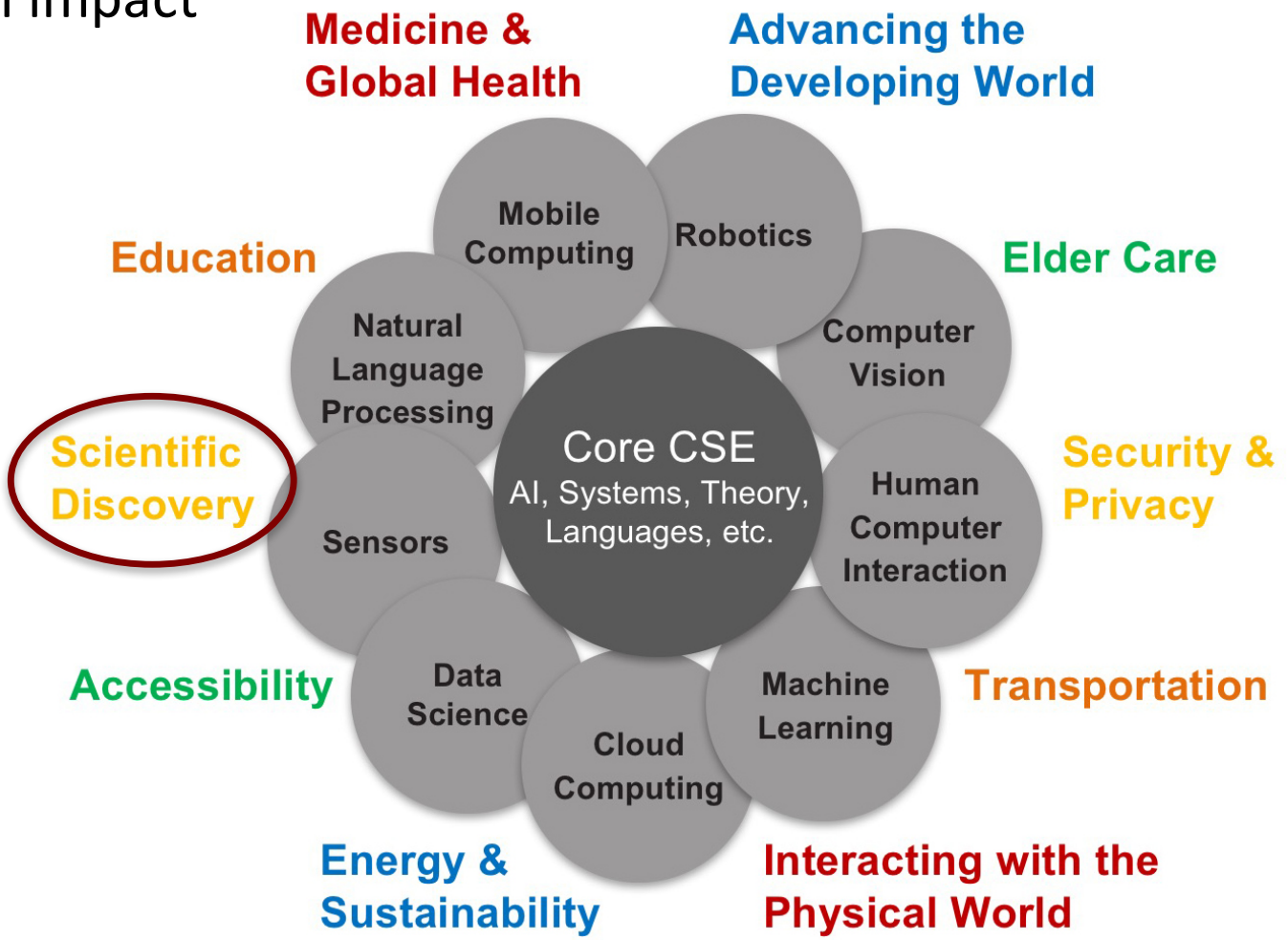

Andy Connolly
Astronomy


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 and institutional impact

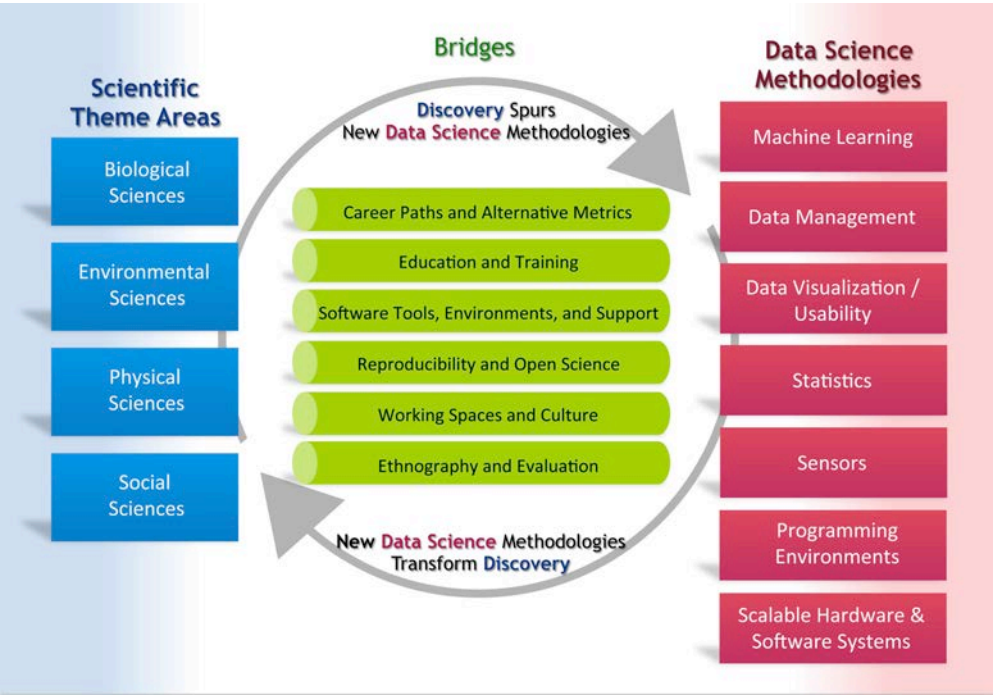


Implications for research policy

- 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 82% of Federal support for basic research in Computer Science in academia
    - 53% of Federal support for all research in Computer Science in academia
- 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

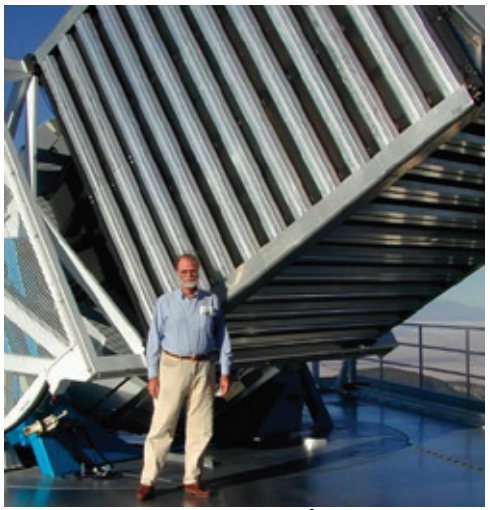
- In 1984, through the establishment of the Office of Scientific Computing and the launch of the Supercomputer Centers Program, NSF leadership catalyzed the widespread adoption of numerical computational science
  - Although the focus was far too great on hardware, far too small on software and on computer science research
- NSF should be exercising the same sort of leadership and catalysis for data-intensive discovery, but it is largely whiffing
  - **IMPORTANT: This is not a suggestion that "national centers" are called for!**

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

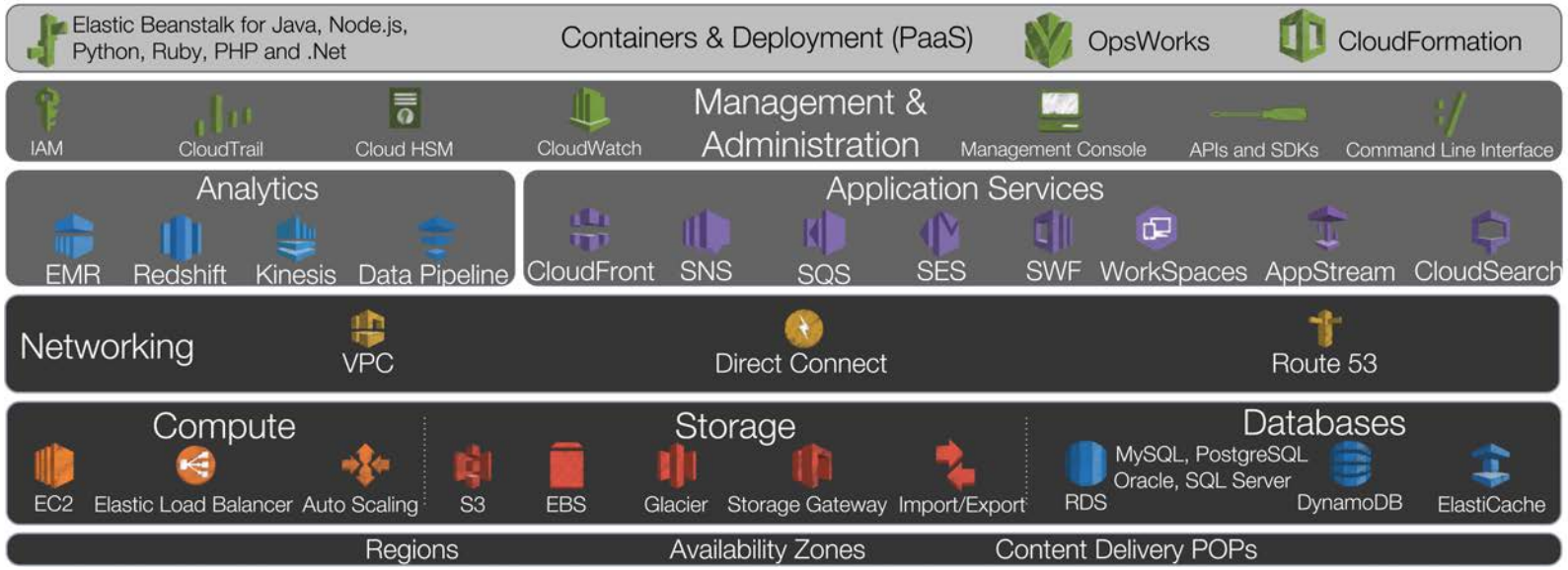
Some Amazon customers

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

1. Essentially infinite capacity
2. You pay for *exactly* what you use: Instantaneous expansion *and* contraction
3. *Zero* capital cost: The user avoids investment in infrastructure that's redundant, under-utilized, and has a short lifetime
4. Burst capacity: 1,000 processors for 1 day costs the same (or less) as 1 processor for 1,000 days – *totally revolutionary!*
5. 7x24x365 operations support
6. Reliability: Auxiliary power, redundant network connections, geographical diversity
7. For many services, someone else handles backup, someone else handles software updates

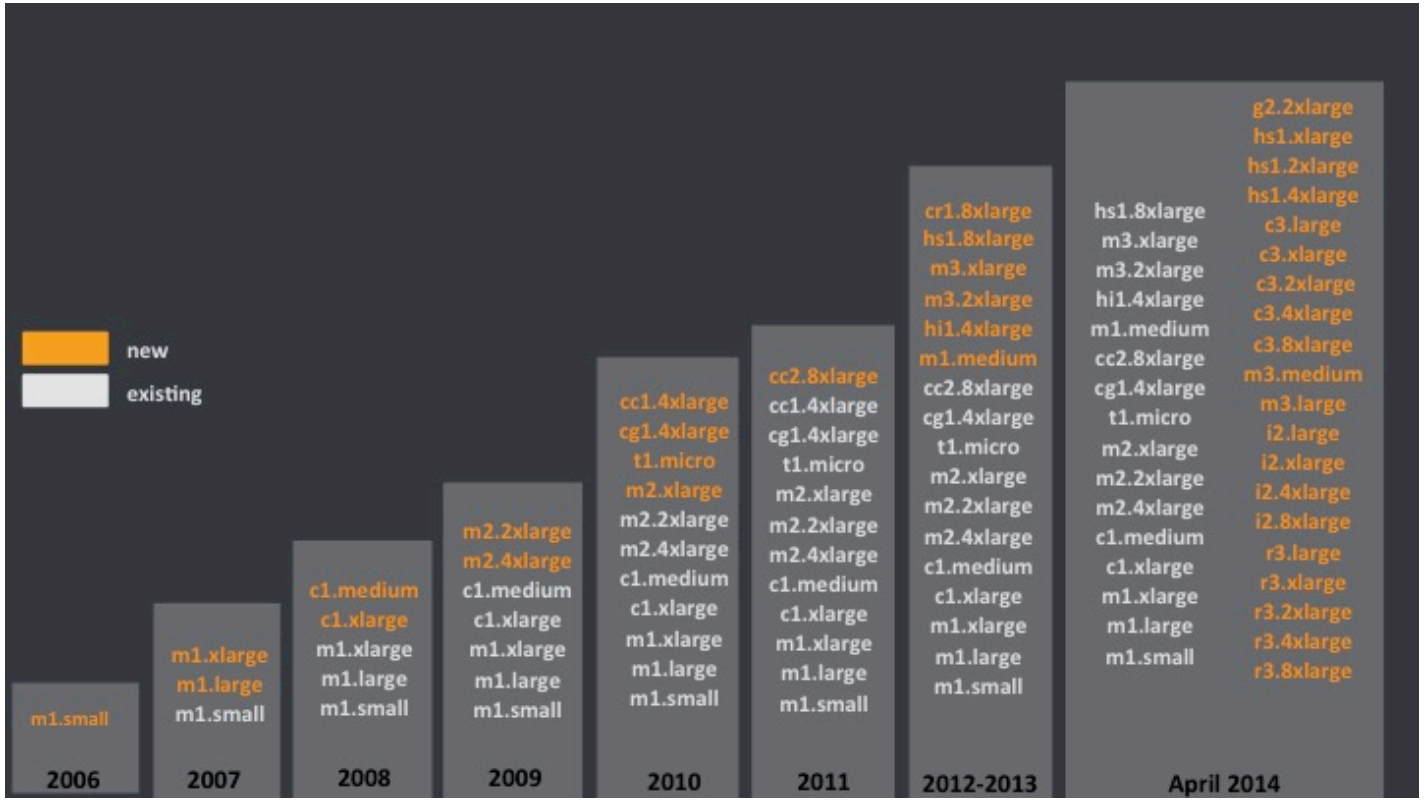
8. Sharing and collaboration are easy
9. This enhances reproducibility – investigators use the same tools and data … exactly the same computational environment
10. It continuously gets bigger, faster, less expensive
11. Capabilities evolve at a rapid pace



AWS Stack

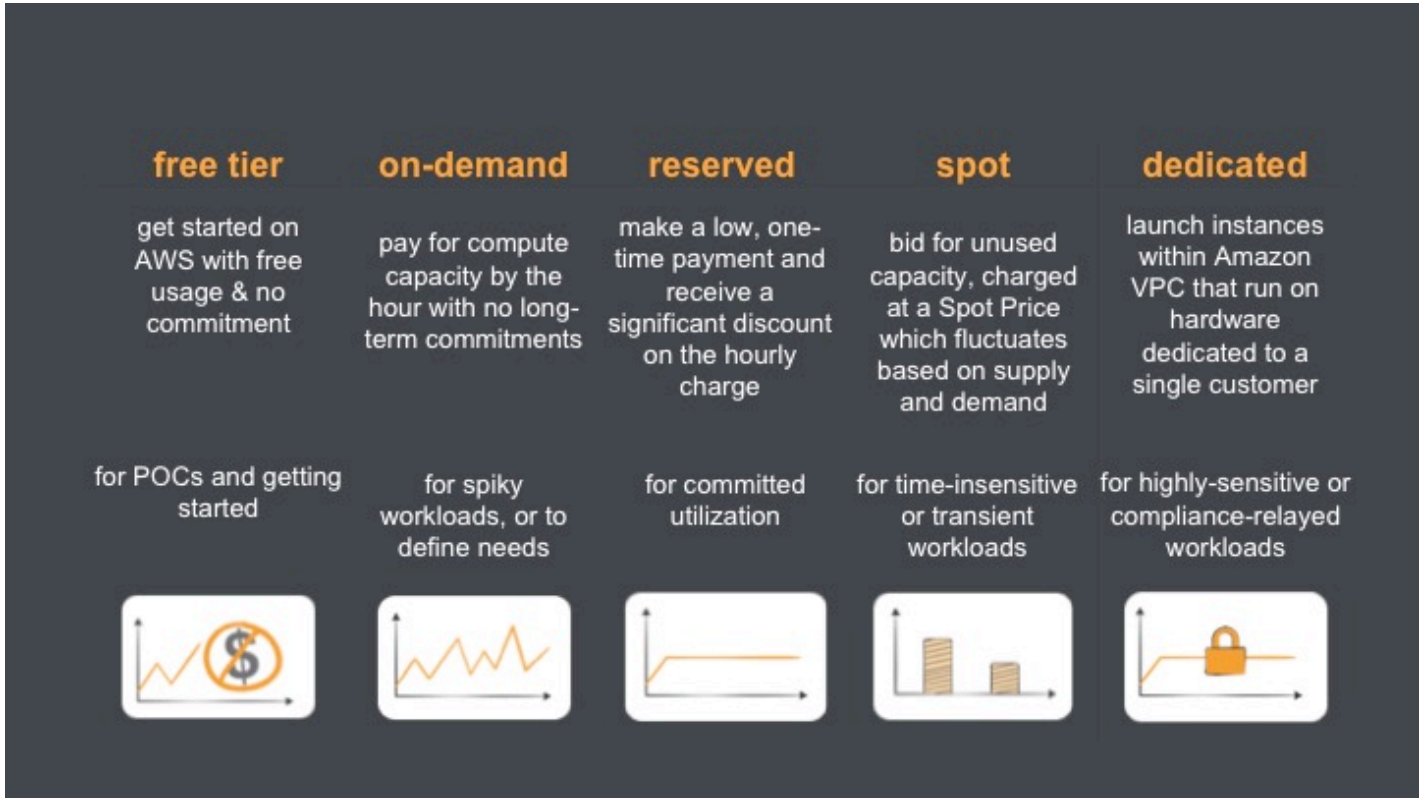
Credit: Jamie Kinney, Amazon

12. Configuration choices evolve at a rapid pace



AWS Instance Type History Credit: Jamie Kinney, Amazon

13. Purchase models evolve at a rapid pace

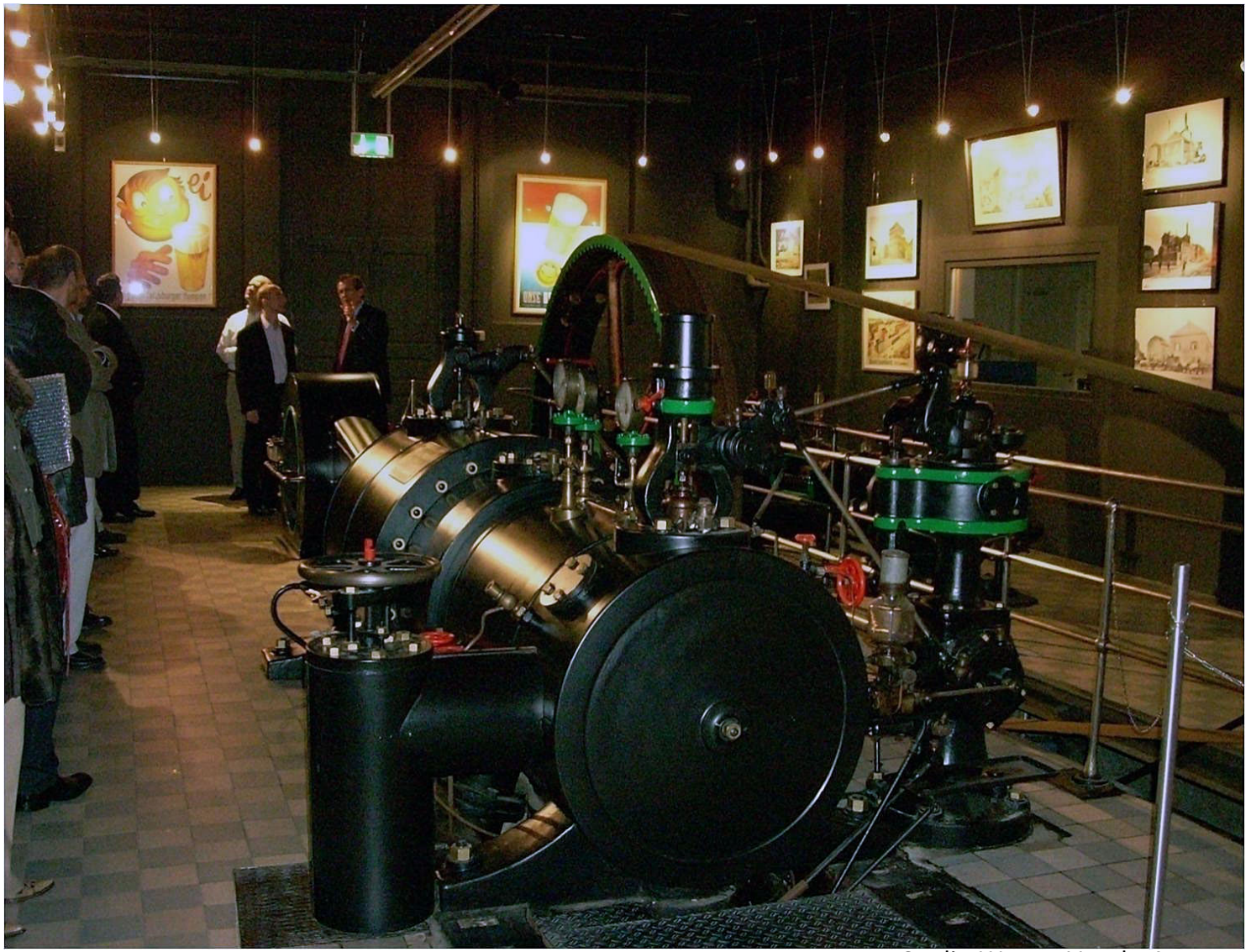


AWS Purchase Models

Credit: Jamie Kinney, Amazon

14. Competition is growing at a rapid pace

– *Including competition for academic and commercial science workloads and datasets!*



Credit: Werner Vogels, Amazon

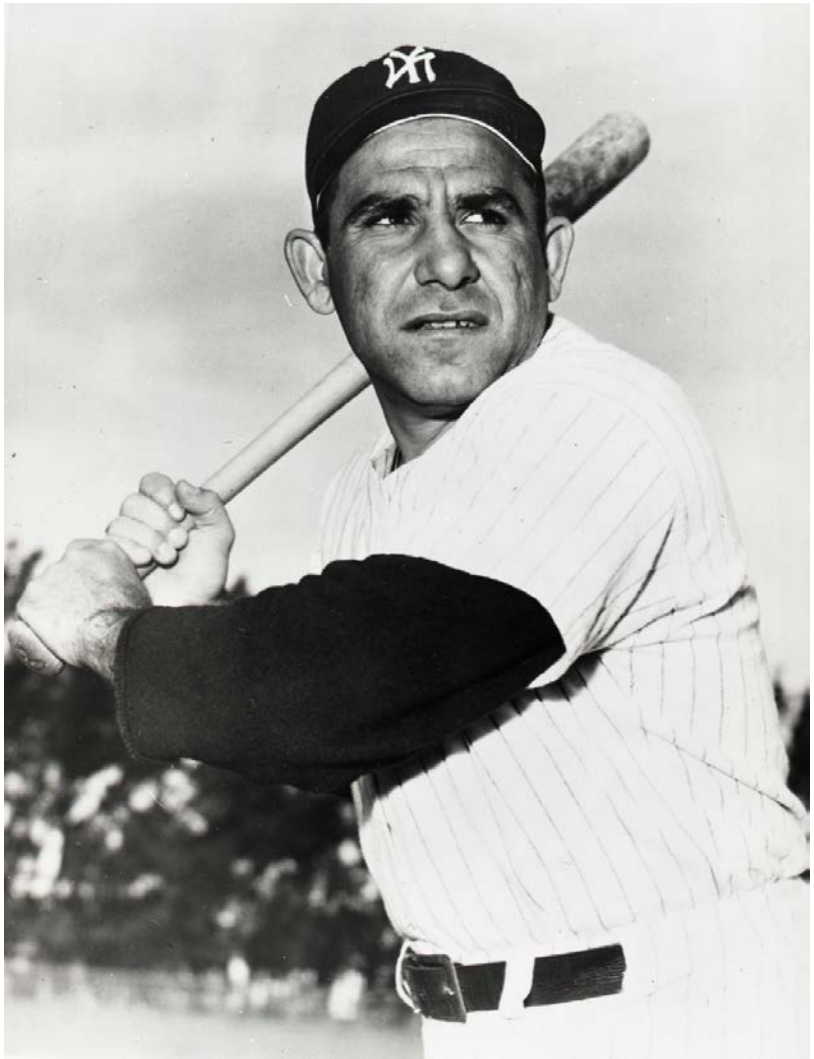
Some possible actions

- *Eliminate overhead* on outsourced cloud services
  - *The University of Washington has unilaterally done this!*
- *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?

~~"It ain't over 'til it's over"~~

"Three strikes and you're out"



Thanks for inviting me!

UNIVERSITY of WASHINGTON
eScience Institute

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