

eScience:

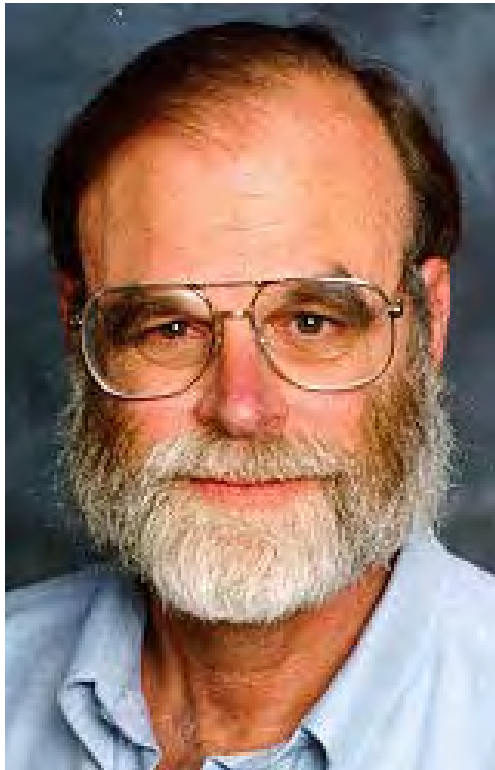
Techniques and Technologies for 21st Century Discovery

Ed Lazowska
Bill & Melinda Gates Chair in
Computer Science & Engineering
University of Washington

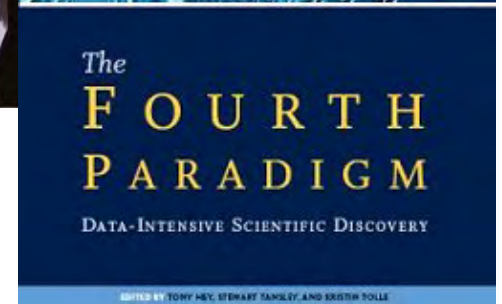


face the
challenge

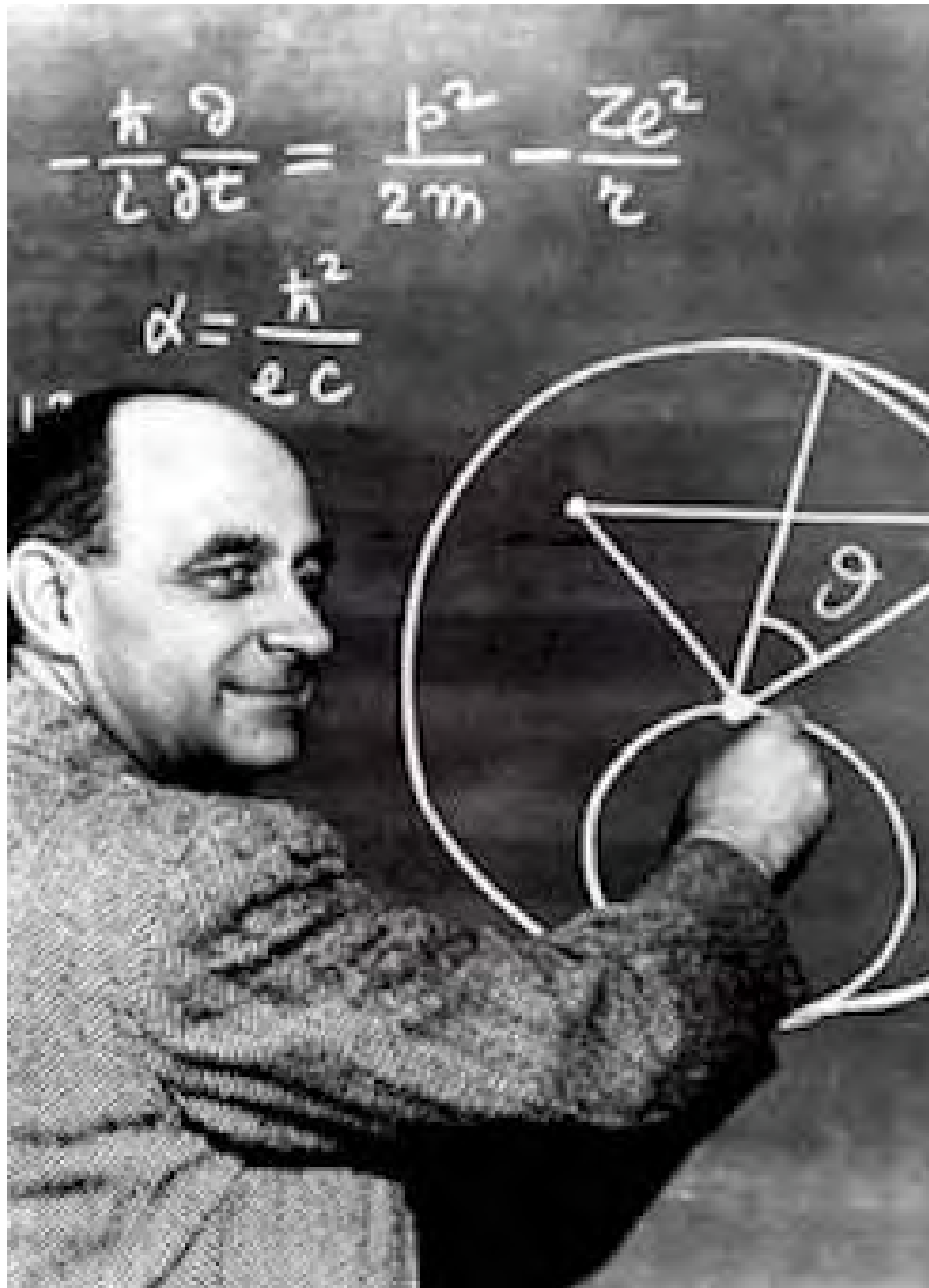
Sensor-driven (data-driven) science and engineering



Jim Gray



Transforming science (again!)



Theory
Experiment
Observation



Theory
Experiment
Observation

Theory
Experiment
Observation



[John Delaney, University of Washington]



Theory
Experiment
Observation
**Computational
Science**



Theory
Experiment
Observation
Computational
Science
eScience



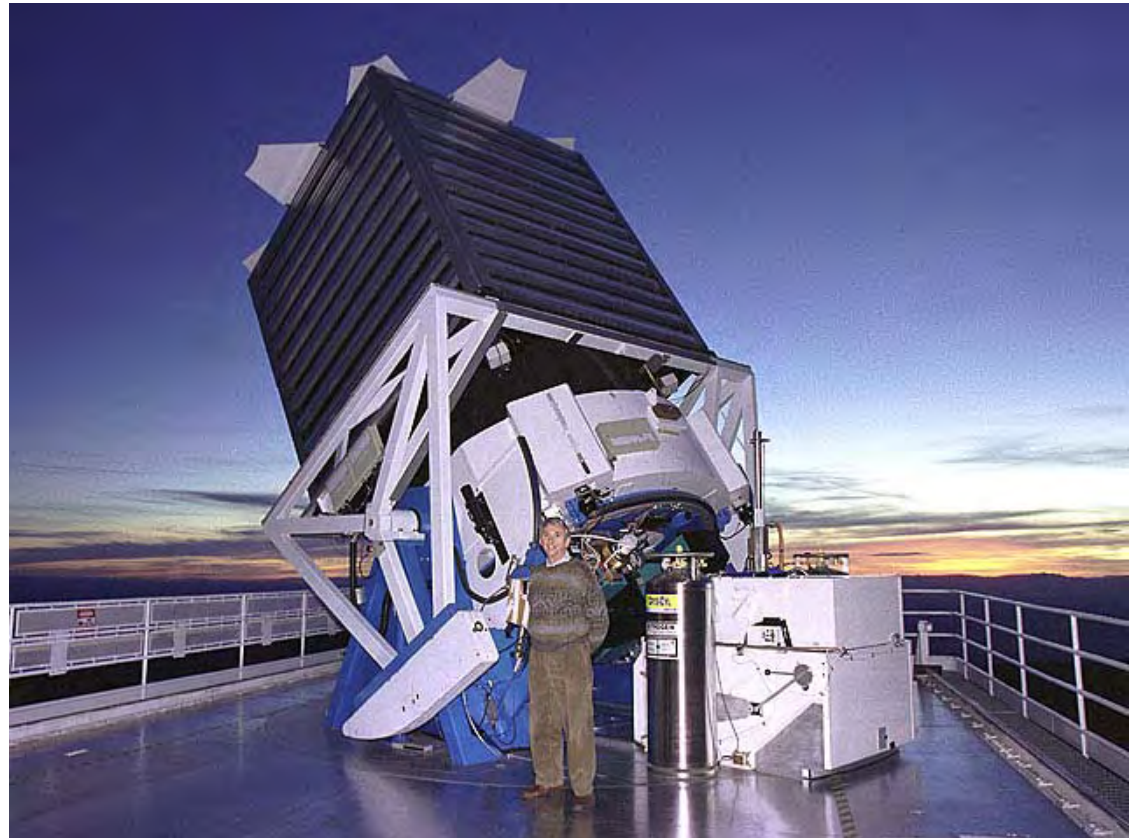
SLOAN DIGITAL SKY SURVEY

eScience is driven by *data* more than by cycles

- Massive volumes of data from sensors and networks of sensors

**Apache Point telescope,
SDSS**

**80TB of raw image data
(80,000,000,000,000 bytes)
over a 7 year period**

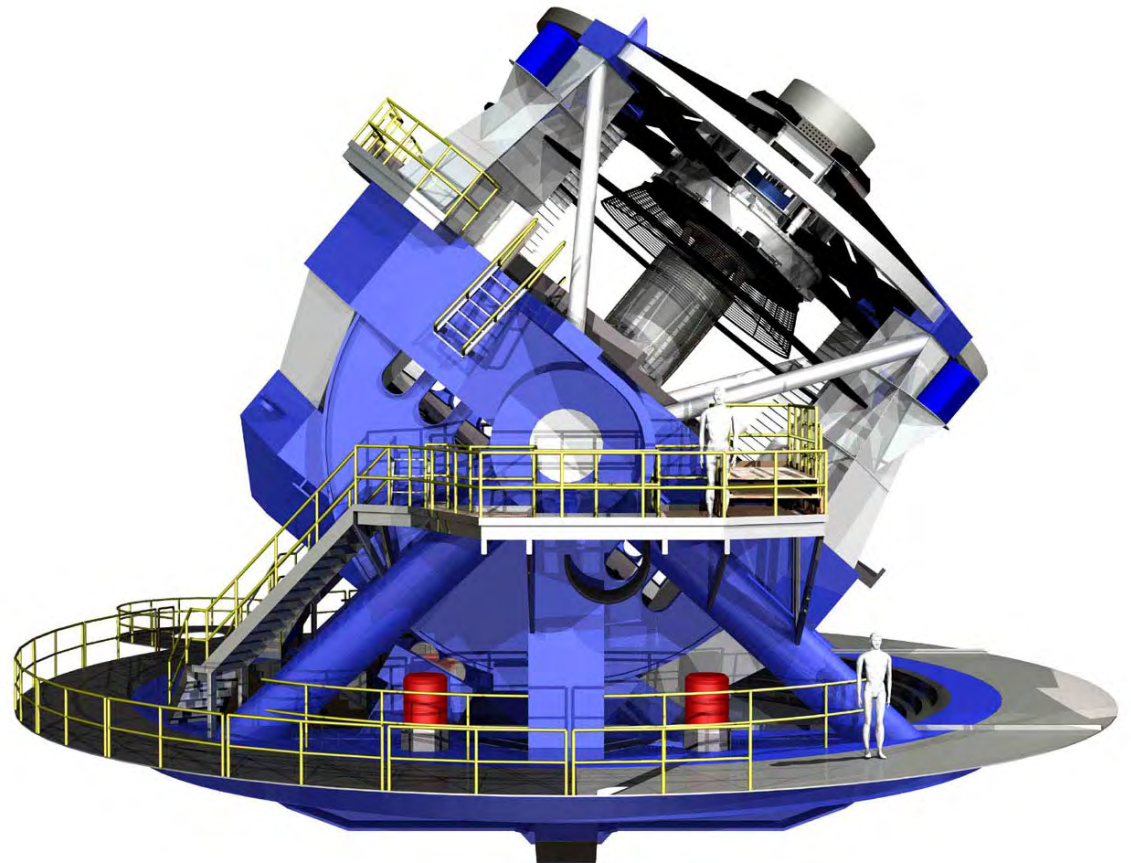




**Large Synoptic Survey
Telescope (LSST)**

**40TB/day
(an SDSS every two days),
100+PB in its 10-year
lifetime**

**400mbps sustained data
rate between
Chile and NCSA**





Large Hadron Collider

**700MB of data
per second,
60TB/day, 20PB/year**



**Illumina
HiSeq 2000
Sequencer
~1TB/day**

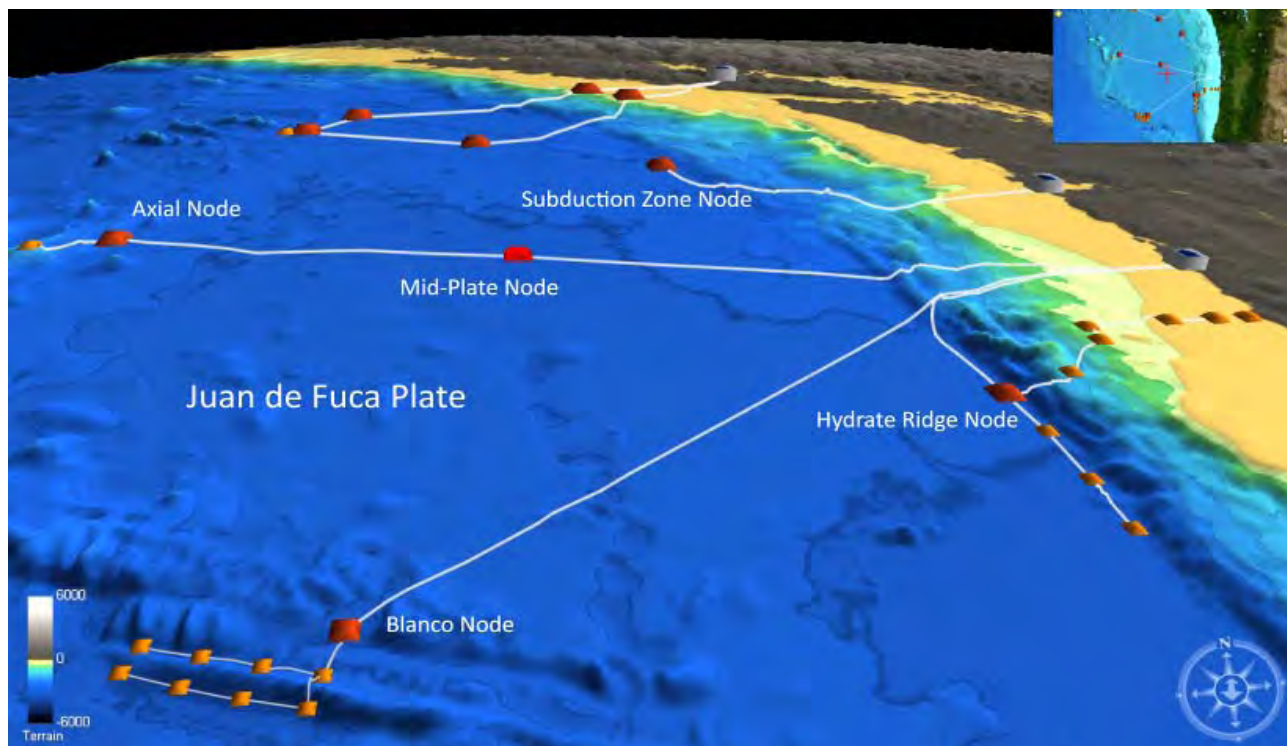


**Major labs
have 25-100
of these
machines**



**Regional Scale
Nodes of the NSF
Ocean Observatories
Initiative**

**1000 km of fiber
optic cable on the
seafloor, connecting
thousands of
chemical, physical,
and biological
sensors**

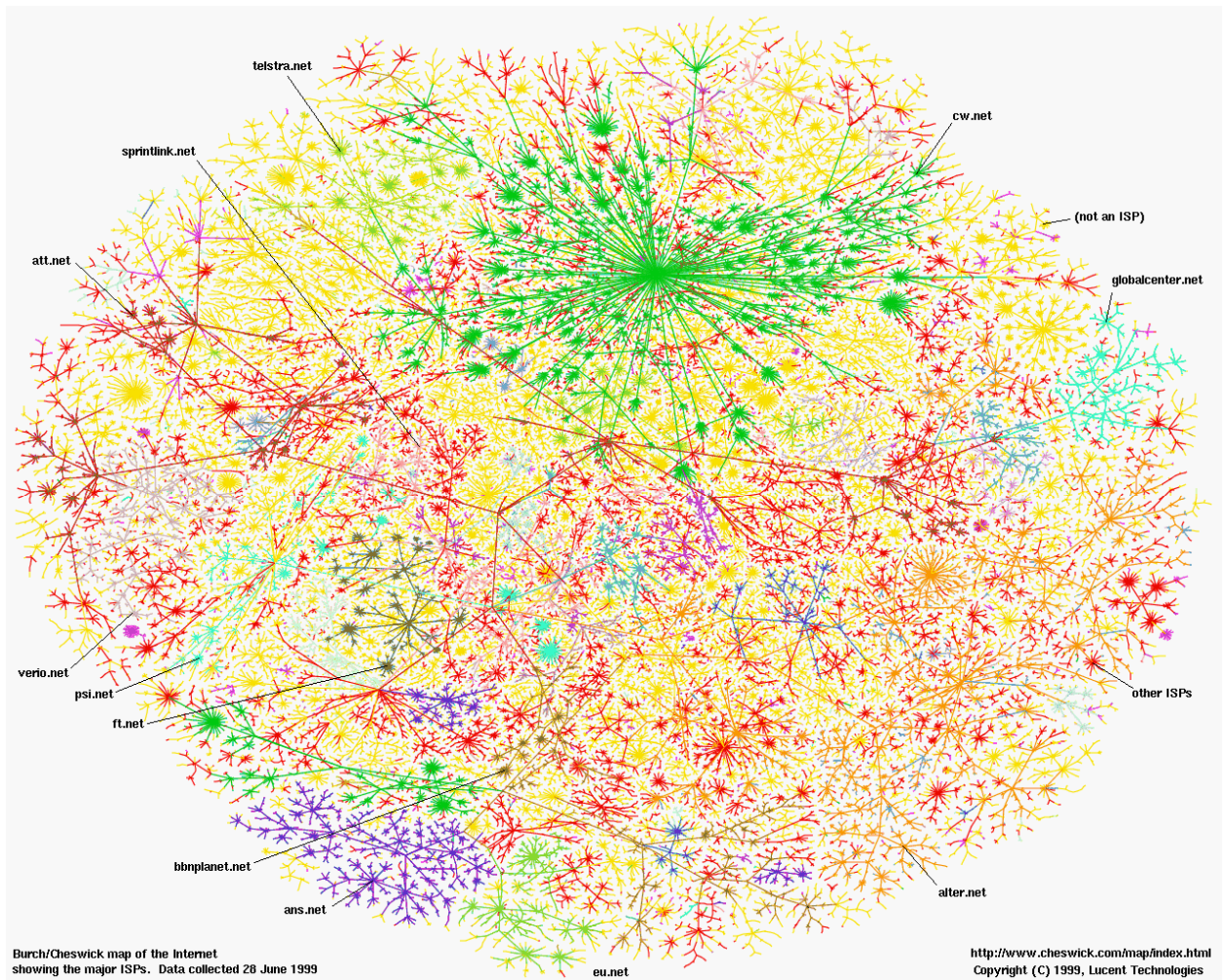




The Web

20+ billion web pages
x 20KB = 400+TB

One computer can
read 30-35 MB/sec
from disk => 4 months
just to read the web





Point-of-sale terminals

eScience is about the *analysis* of data



- The automated or semi-automated extraction of knowledge from massive volumes of data
 - There's simply too much of it to look at
- It's not just a matter of volume
 - Volume
 - Rate
 - Complexity / dimensionality

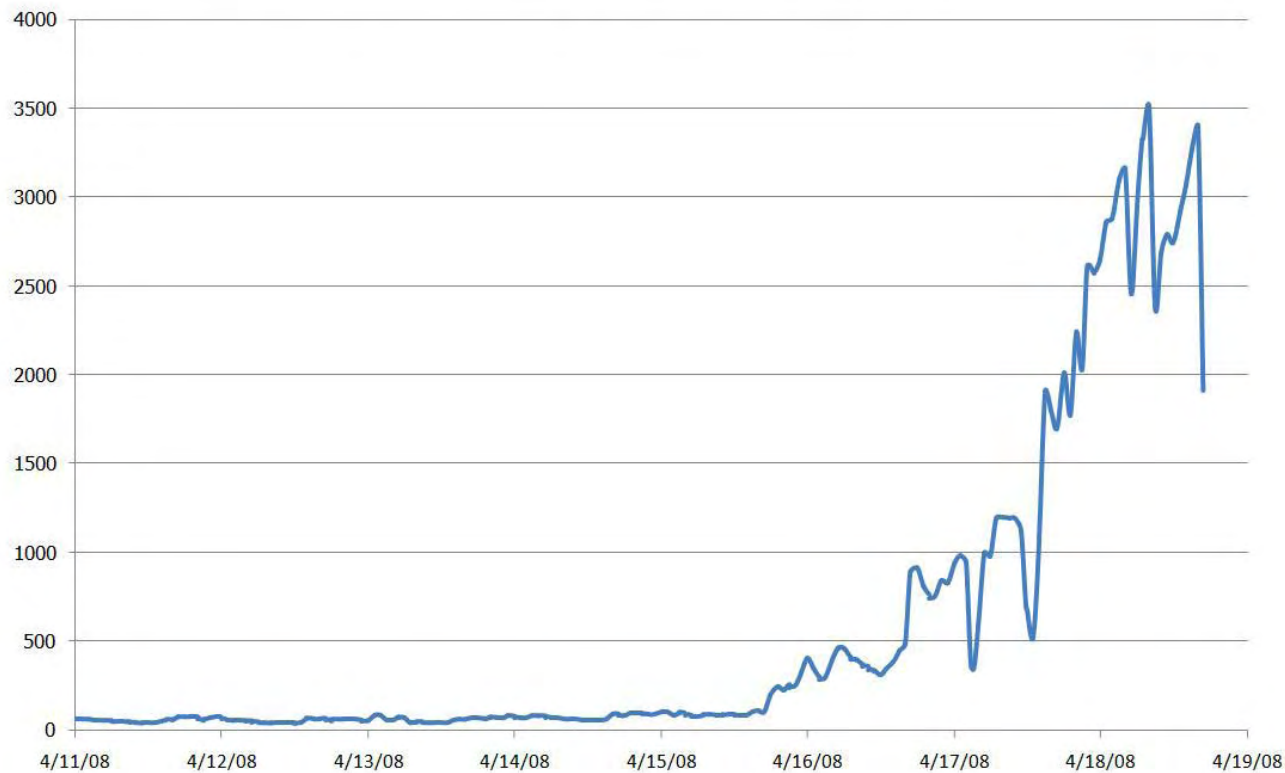
eScience utilizes a spectrum of computer science techniques and technologies

- Sensors and sensor networks
- Backbone networks
- Databases
- Data mining
- Machine learning
- Data visualization
- Cluster computing at enormous scale



eScience is married to the Cloud: Scalable computing and storage for everyone

Animoto: EC2 Instance Usage



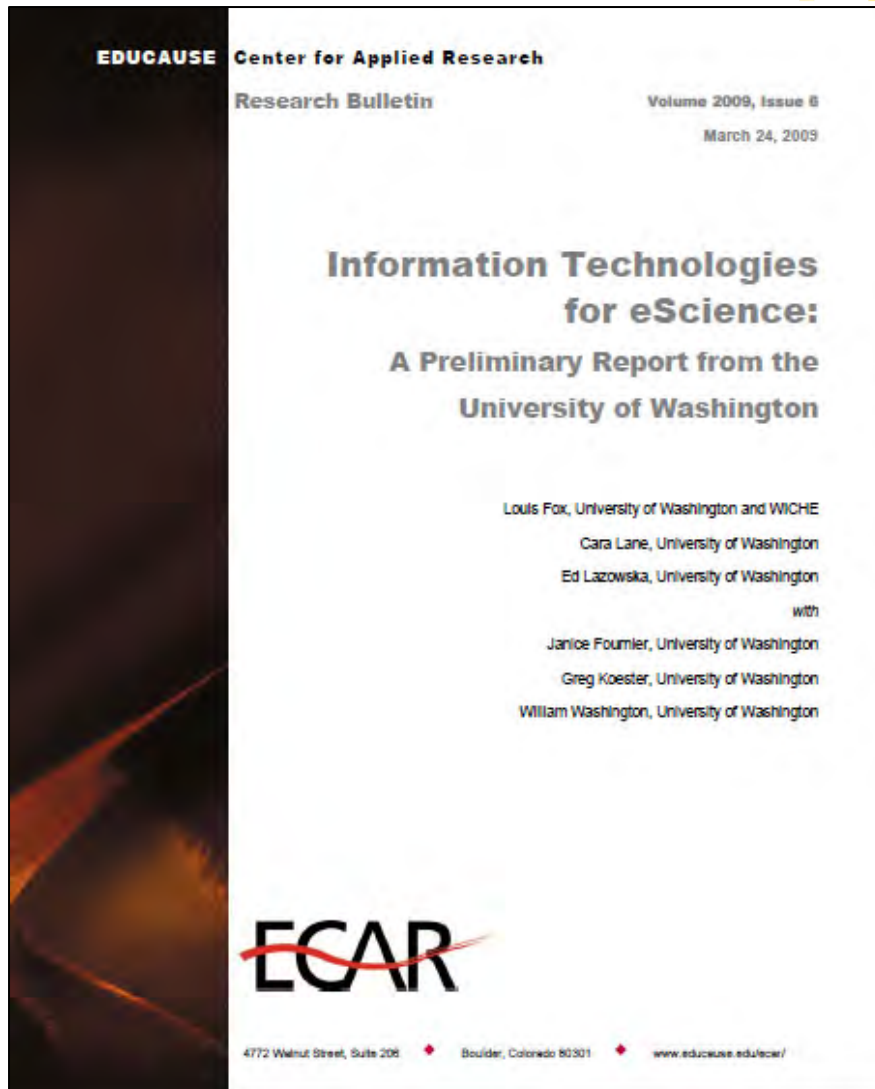
[Werner Vogels, Amazon.com]

eScience will be pervasive



- Simulation-oriented computational science has been transformational, but it has been a niche
 - As an institution (e.g., a university), you didn't need to excel in order to be competitive
- eScience capabilities must be broadly available in any institution
 - If not, the institution will simply cease to be competitive

Top scientists across all fields grasp the implications of the looming data tsunami



- Survey of 125 top investigators
 - "Data, data, data"
- Flat files and Excel are the most common data management tools
 - Great for Microsoft ... lousy for science!
- Typical science workflow:
 - 2 years ago: 1/2 day/week
 - Now: 1 FTE
 - In 2 years: 10 FTE
- Need tools, tools, tools!

The University of Washington eScience Institute



■ Motivating observations

- Like simulation-oriented computational science, data-intensive science will be transformational
- Unlike simulation-oriented computational science, data-intensive science will be pervasive
- Even more broadly than simulation-oriented computational science, data-intensive science draws on new techniques and technologies from computer science, statistics, and other fields
- Cloud services are essential - "get computing out of the closet"
- If we don't lead in the *development* and *application* of these techniques and technologies, we're going to lose



■ Mission

- Ensure the University of Washington's position at the forefront of research both in modern eScience techniques and technologies, and in the fields that depend upon these techniques and technologies

■ Strategy

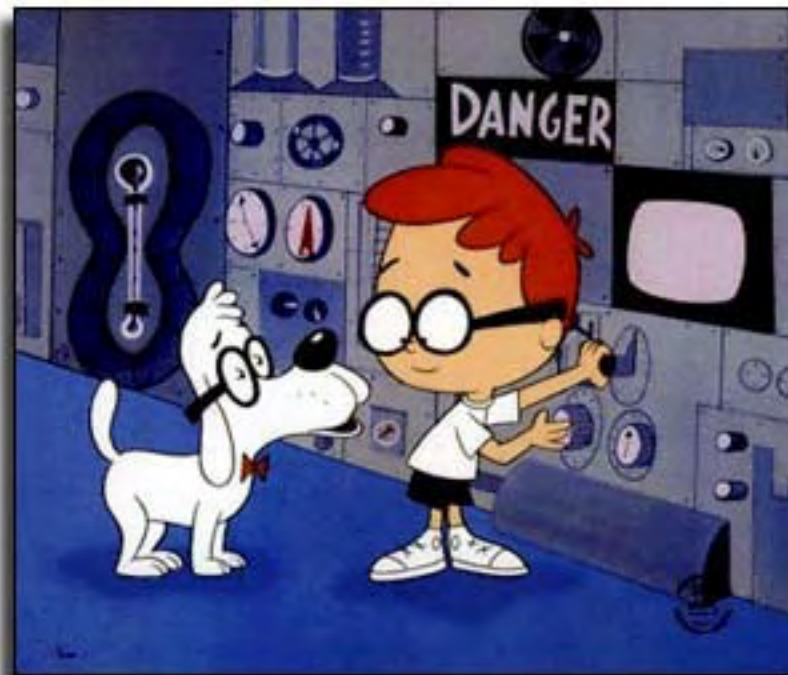
- Bootstrap a cadre of Research Scientists
 - Help leading faculty become exemplars and advocates
 - Broaden impact by aggressive community-building and sharing of expertise and facilities
 - Add faculty in key fields
- ## ■ Launched in July 2008 with \$1 million in permanent funding from the Washington State Legislature
- Many grants received since then

Computer Science: From data to insight to action

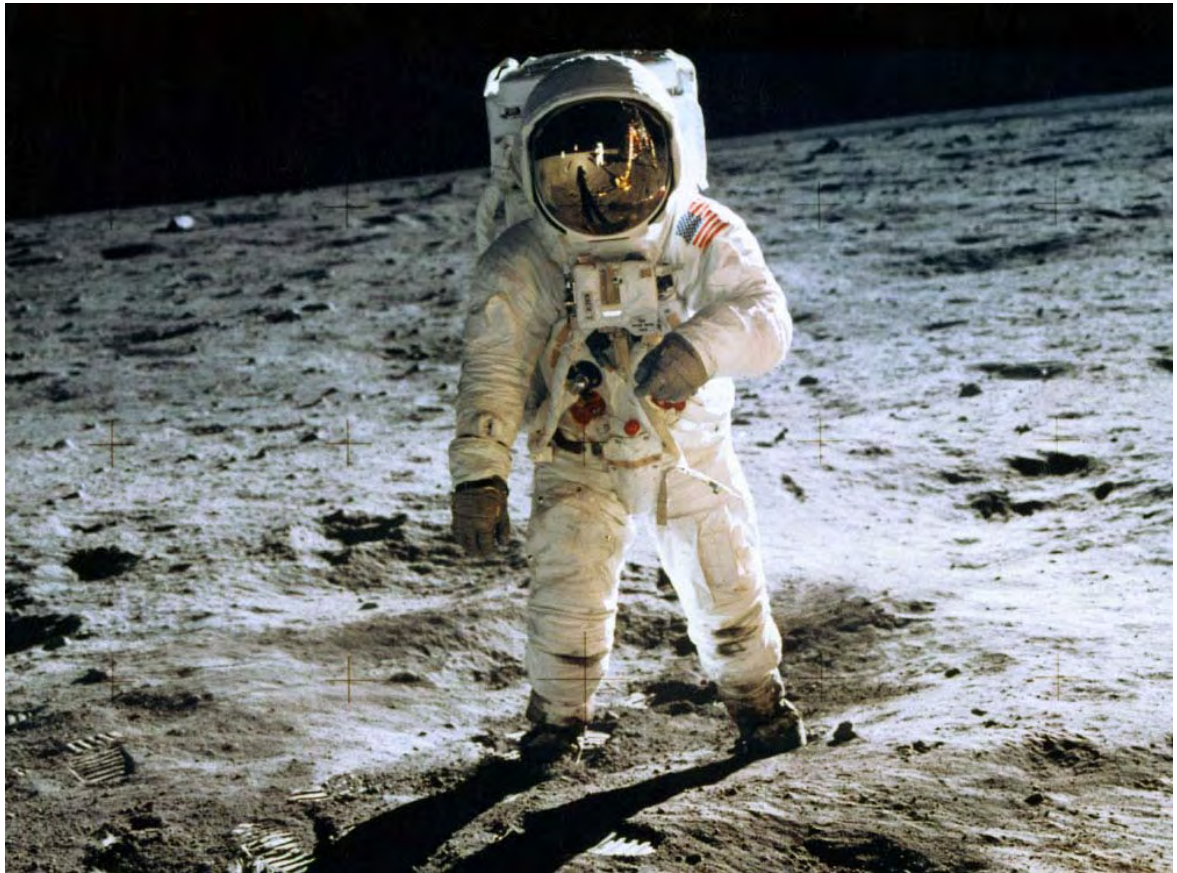


- Sensors and sensor networks
- Backbone networks
- Databases
- Data mining
- Machine learning
- Data visualization
- Cluster computing at enormous scale
- Enabling 21st Century Discovery in Science and Engineering
- Enabling Evidence-Based Healthcare
- Enabling the New Biology
- Enabling Advanced Intelligence and Decision Making for America's Security
- Enabling a Revolution in Transportation
- Enabling a Transformation of American Education
- Enabling the Smart Grid

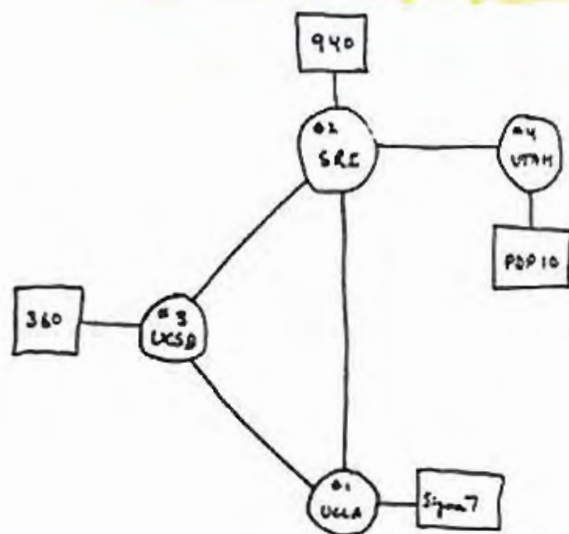
Forty years ago ...







[Peter Lee, DARPA, and Pat Lincoln, SRI]




THE ARPA NETWORK
DEC 1969
4 NODES

29 OCT 69	2100	LOADED OP. PROGRAM	SK
		EDIC BEN BARKER	
		BBV	
	22:30	Talked to SRI	SK
		Host to Host	
		Left op prog running	SK
		after sending	
		a host dead message	
		to imp.	



With forty years of hindsight, which had the greatest impact?



EXPONENTIALS  **US**

Is this a great time, or what?!?!



face the
challenge

