The University of Washington eScience Institute

Ed Lazowska

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

Director
University of Washington
eScience Institute

Cloud Futures 2010

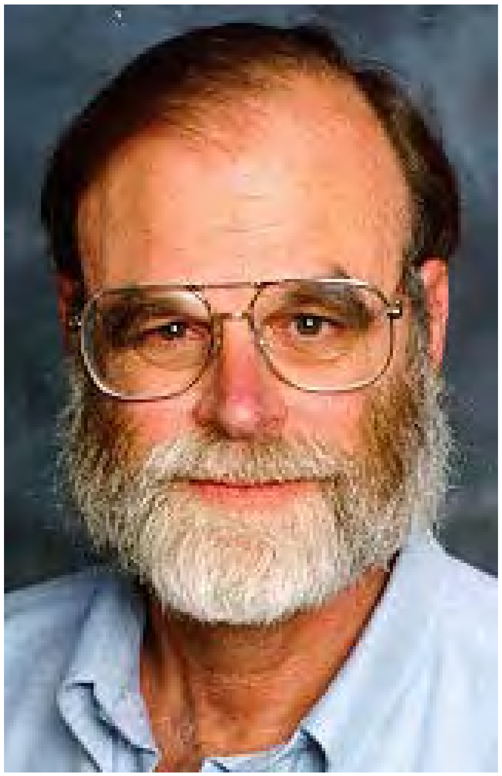
April 2010

http://lazowska.cs.washington.edu/cloud2010.pdf

This morning

- The nature of eScience
- A bit of history
- The University of Washington eScience Institute
- Some example activities
- A few observations
- A plug for computing research

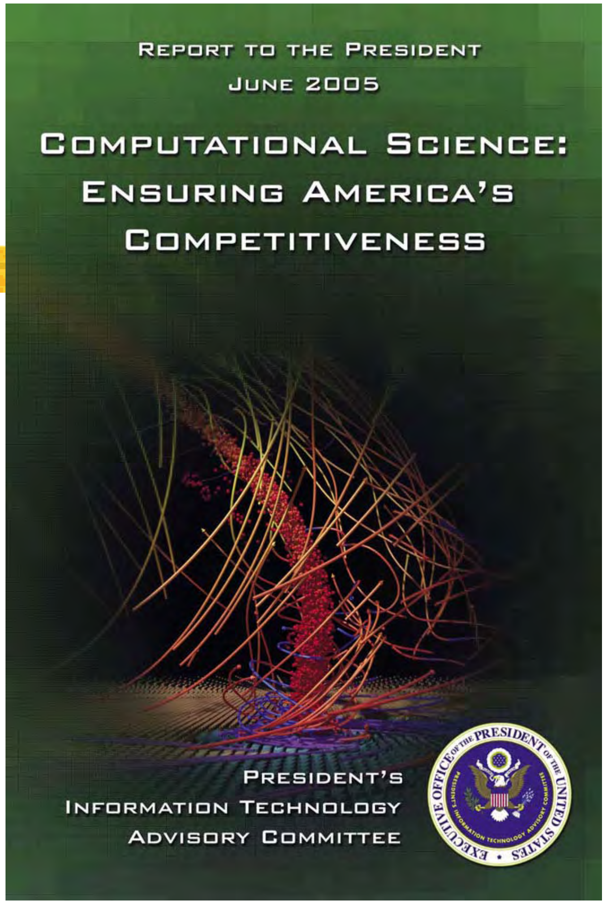
eScience: Sensor-driven (data-driven) science and engineering



Jim Gray

Transforming science (again!)

Dan Reed

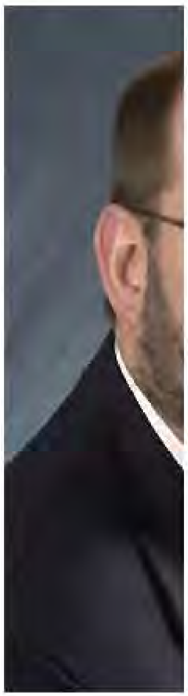


RECOMMENDATION

The Federal government must rebalance its R&D investments to: (a) create a new generation of well-engineered, scalable, easy-to-use software suitable for computational science that can reduce the complexity and time to solution for today's challenging scientific applications and can create accurate simulations that answer new questions; (b) design, prototype, and evaluate new hardware architectures that can deliver larger fractions of peak hardware performance on scientific applications; and (c) focus on sensor- and data-intensive computational science applications in light of the explosive growth of data.

Report to the President
June 2005

Computational Science:



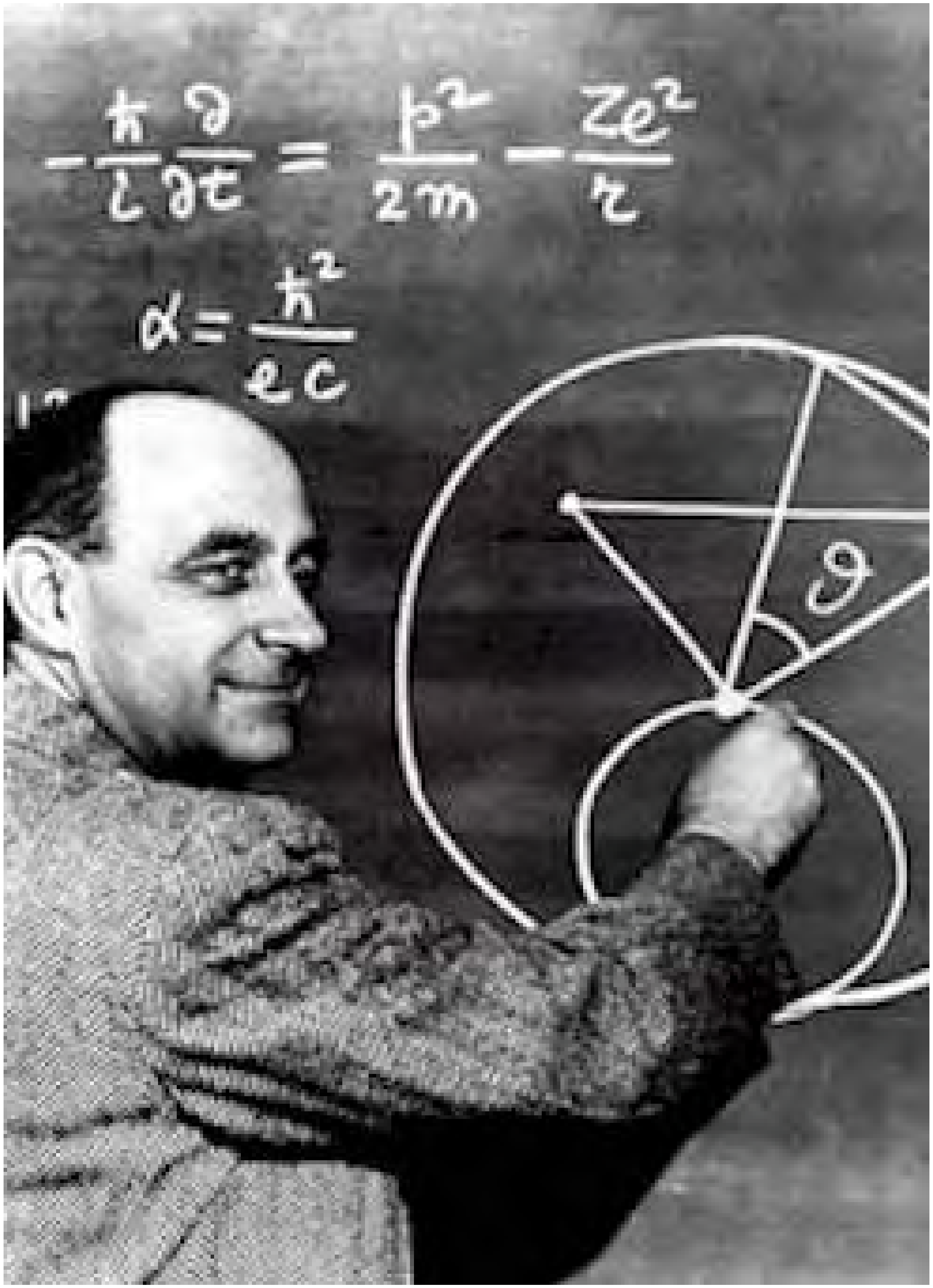
Sidebar 2

Repeating History: Lessons Not Learned

During the past two decades, the national science community has produced a plethora of reports, each recommending sustained, long-term investment in the underlying technologies (algorithms, software, architectures, hardware, and networks) and applications needed to realize the benefits of computational science. These reports have stressed the now essential role that computational science plays in supporting, stimulating, catalyzing, and transforming the conduct of science and engineering.

The reports have also emphasized how computing can address applications of significantly greater complexity, scope, and scale, including problems and issues of national importance that cannot be otherwise addressed. Many of the reports generated responses, but they were often short-lived. In general, short-term investment and limited strategic planning have led to excessive focus on incremental research rather than on long-term, sustained research with lasting impact that can solve important problems. These reports and their messages are summarized in Appendix B.

A report card of national performance might record a grade of C–, with an accompanying teacher's note that says, "This student has great potential, but struggles to maintain focus and complete work on time. This student sometimes has difficulty sharing and playing well with others."



Theory

Experiment

Observation

Theory

Experiment

Observation


Theory
Experiment
Observation
[John Delaney, University of Washington]

Theory
Experiment
Observation
Computational Science

Theory

Experiment

Observation

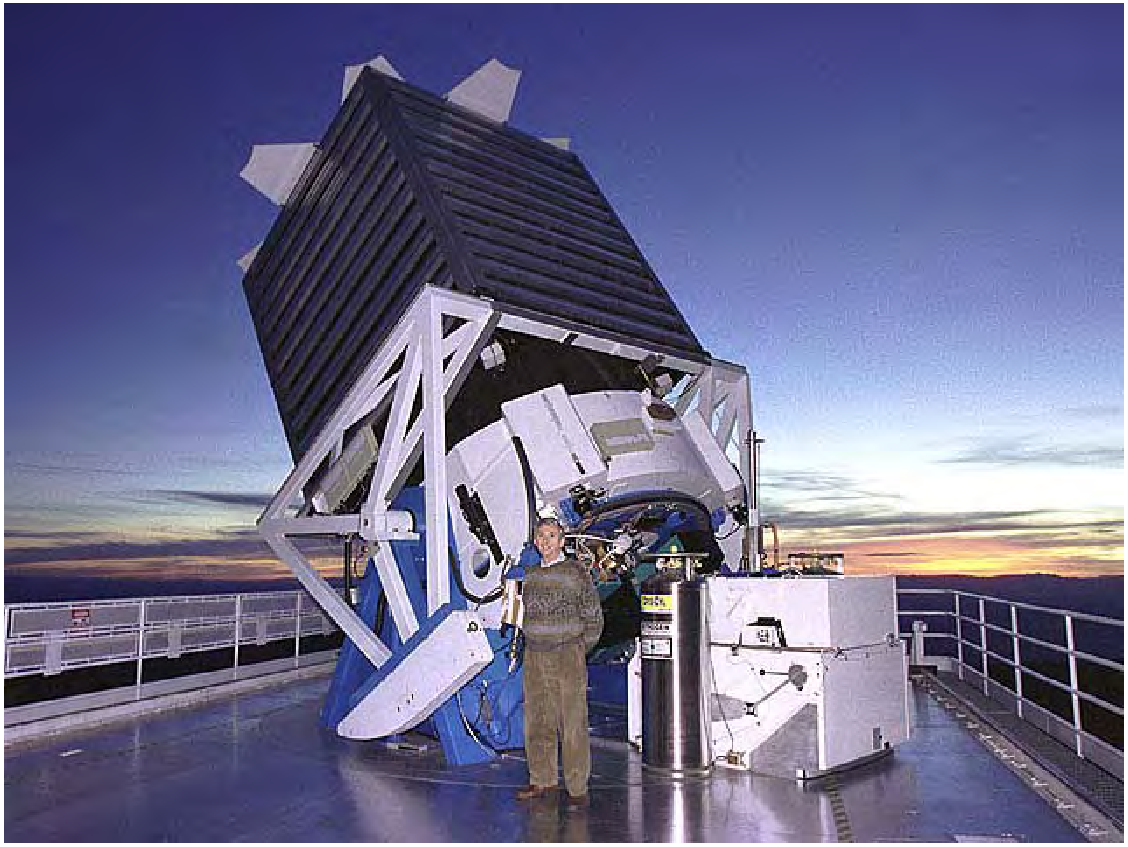
Computational Science

eScience

SLOAN DIGITAL SKY SURVEY

eScience is driven by *data* more than by cycles

- Massive volumes of data from sensors and networks of sensors



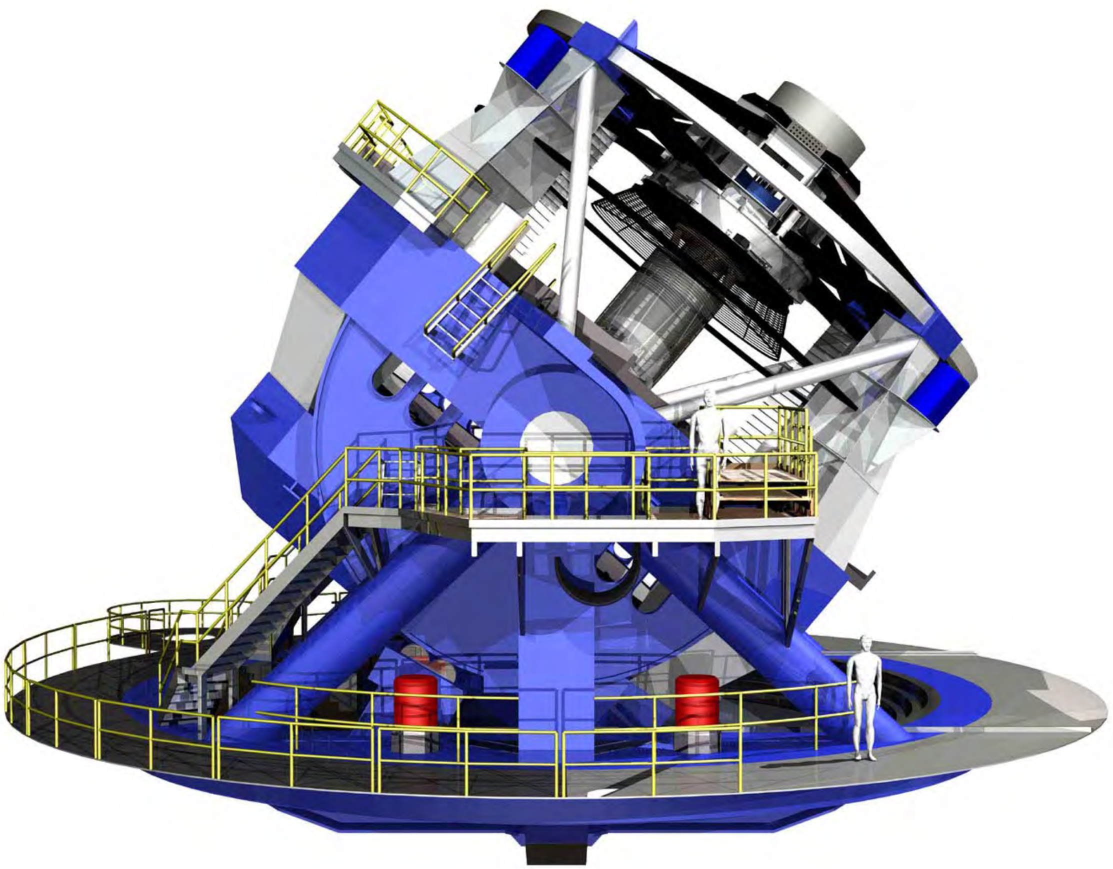
Apache Point telescope, SDSS

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

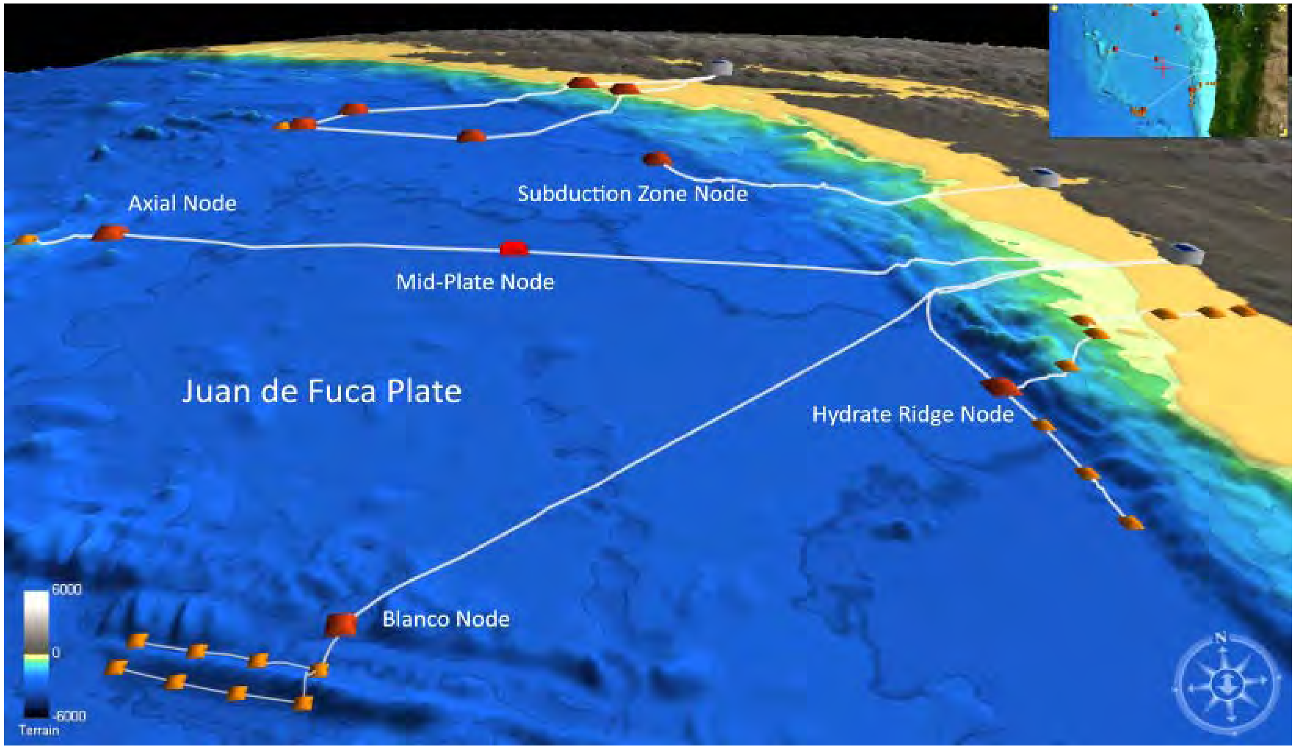
Illumina HiSeq 2000 Sequencer

~1TB/day

Major labs have 25-100 of these machines

Regional Scale Nodes of the NSF Ocean Observatories Initiative

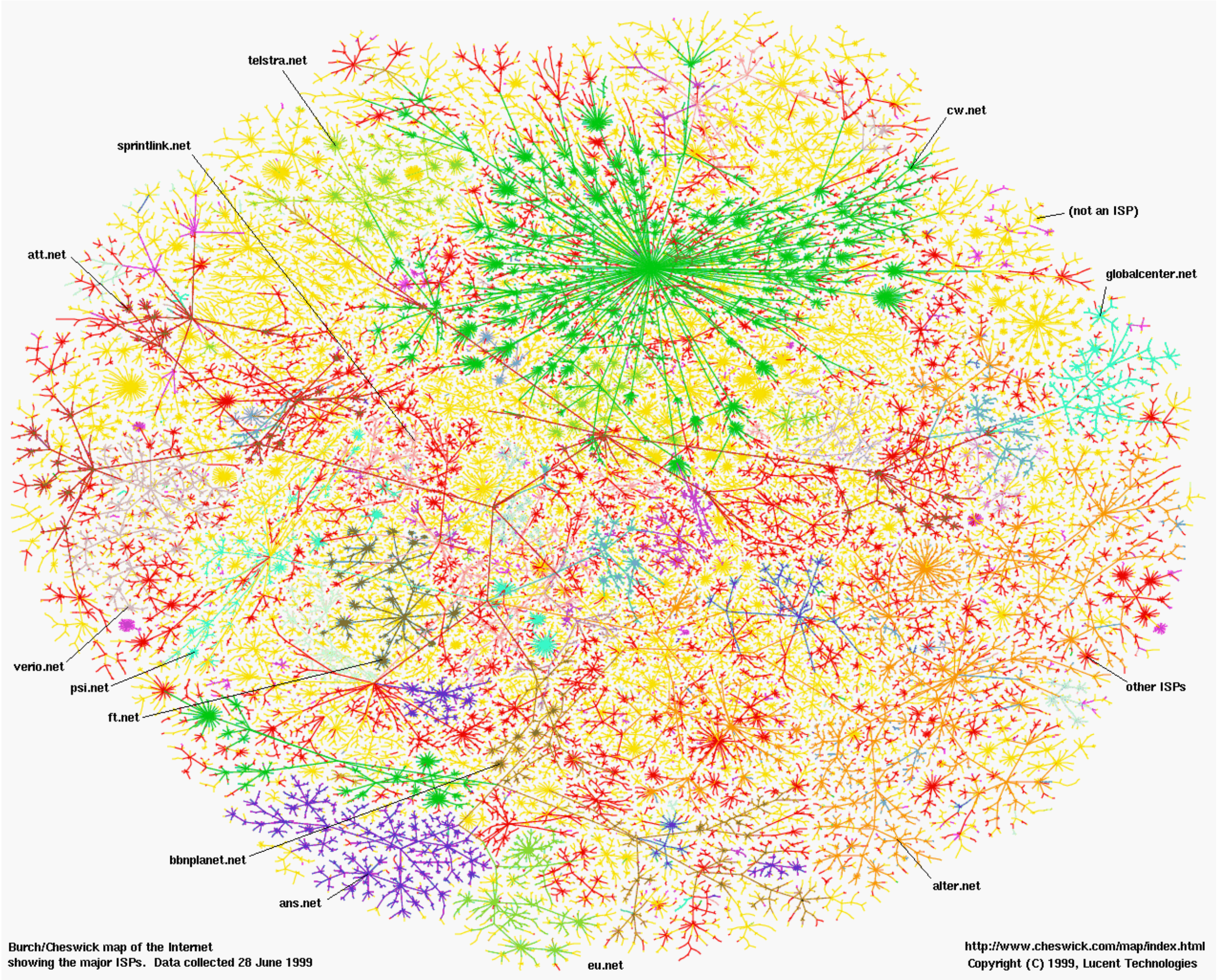
1000 km of fiber optic cable on the seafloor, connecting thousands of chemical, physical, and biological sensors



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



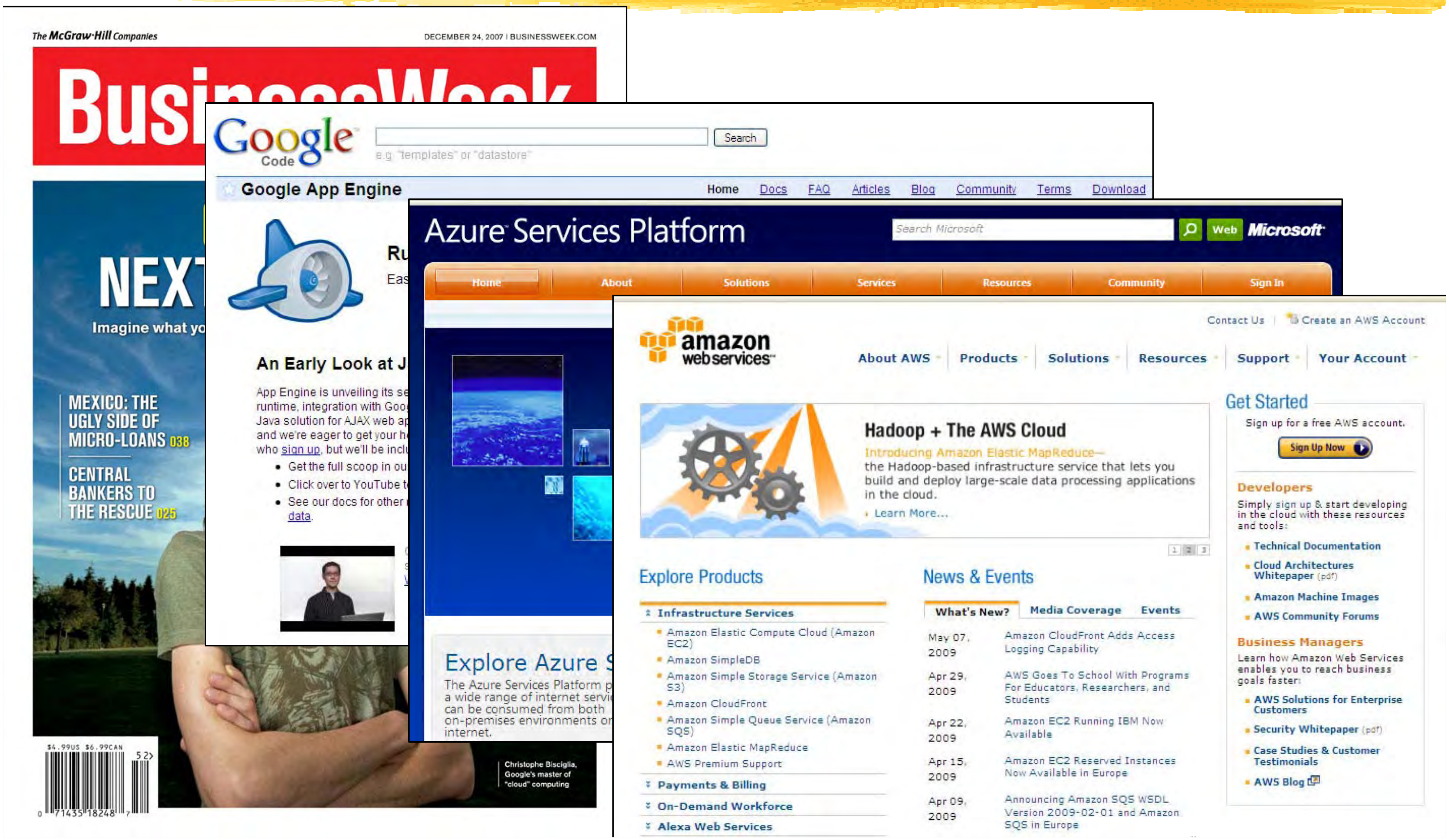
eScience is about the *analysis* of data

- 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



eScience will be pervasive

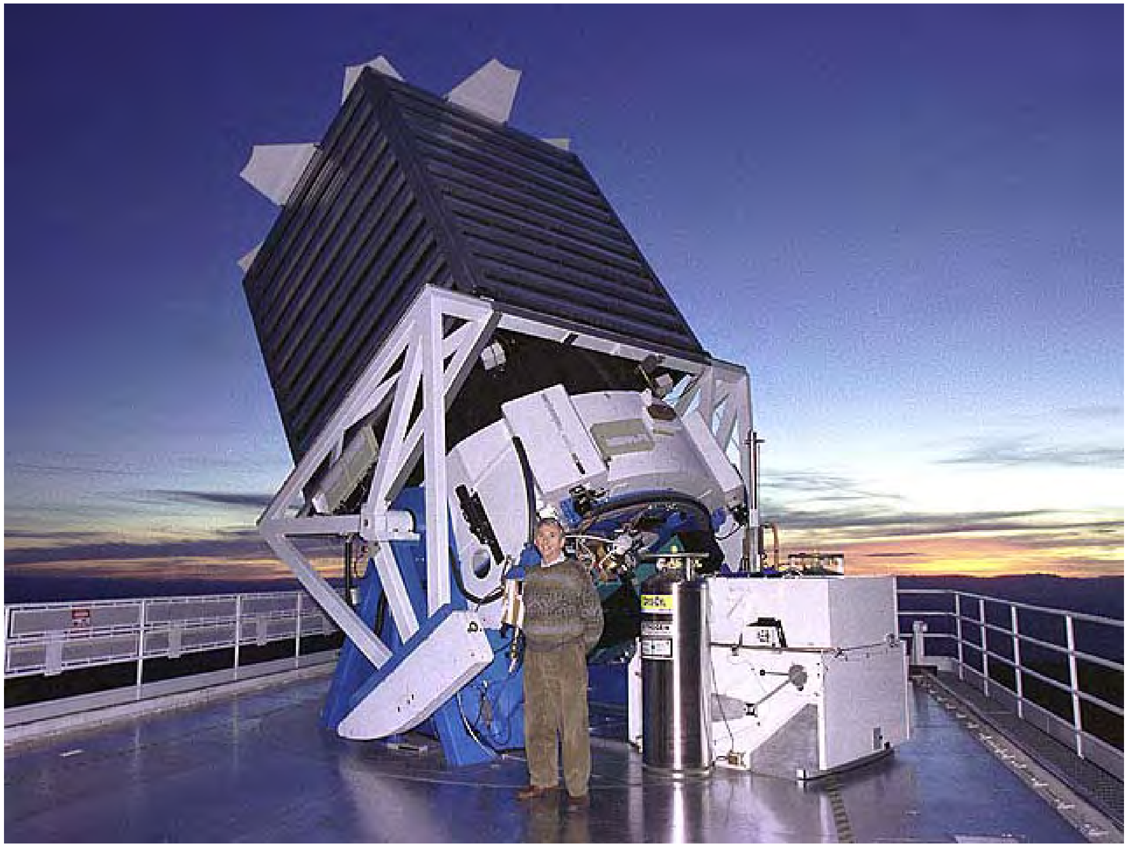
- Simulation-oriented computational science has been transformational, but it has been a niche
  - As an institution (e.g., a university), you didn't need to excel in order to be competitive
- eScience capabilities must be broadly available in any institution
  - If not, the institution will simply cease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

Some history, from astronomy

SLOAN DIGITAL SKY SURVEY

Apache Point telescope, SDSS

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

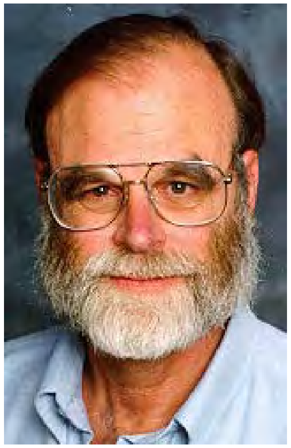


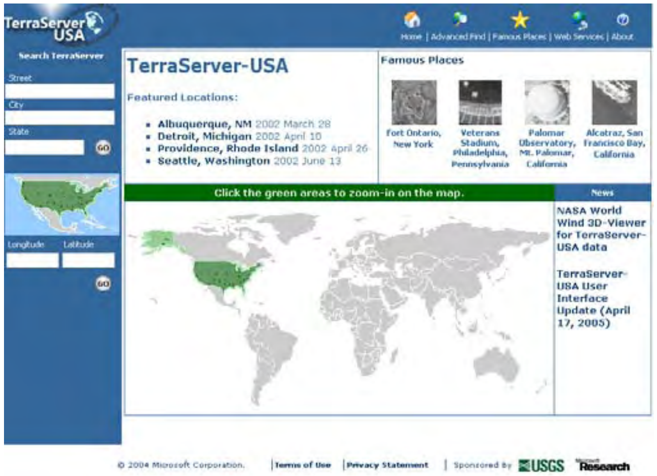
- Project plan
  - SDSS was budgeted as a $16 million project
  - The software was to be written by astronomy faculty during the summers, when they weren't teaching
  - Use Objectivity as the data store
    - Developed by Motorola for the Iridium satellite project

- Project reality
  - $80 million
  - 30% spent on software, *plus* Microsoft's enormous contributions through Jim Gray and his collaborators
  - Research impact: "If it weren't for Jim Gray's contributions, SDSS would have been more likely to yield 100 research papers than the 5,000 that actually resulted."

    \- Andy Connolly, University of Washington

How'd it come to be?



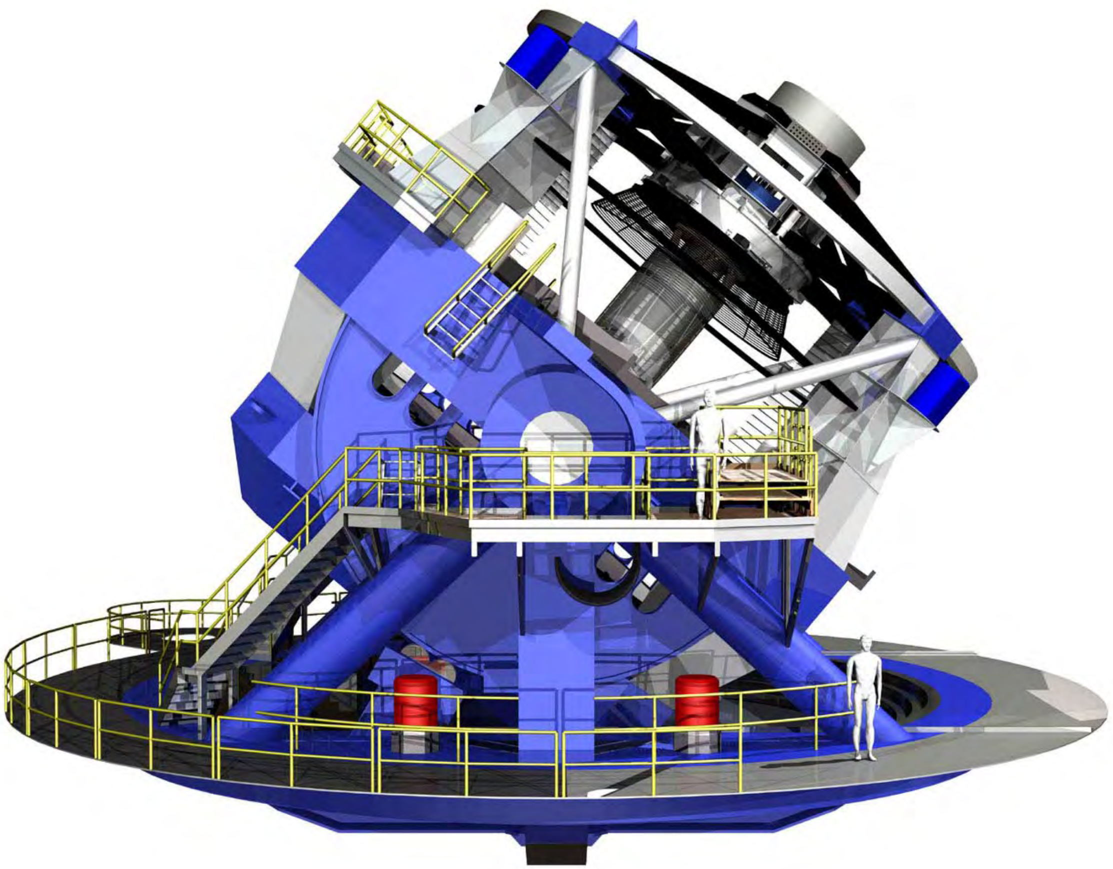


SLOAN DIGITAL SKY SURVEY

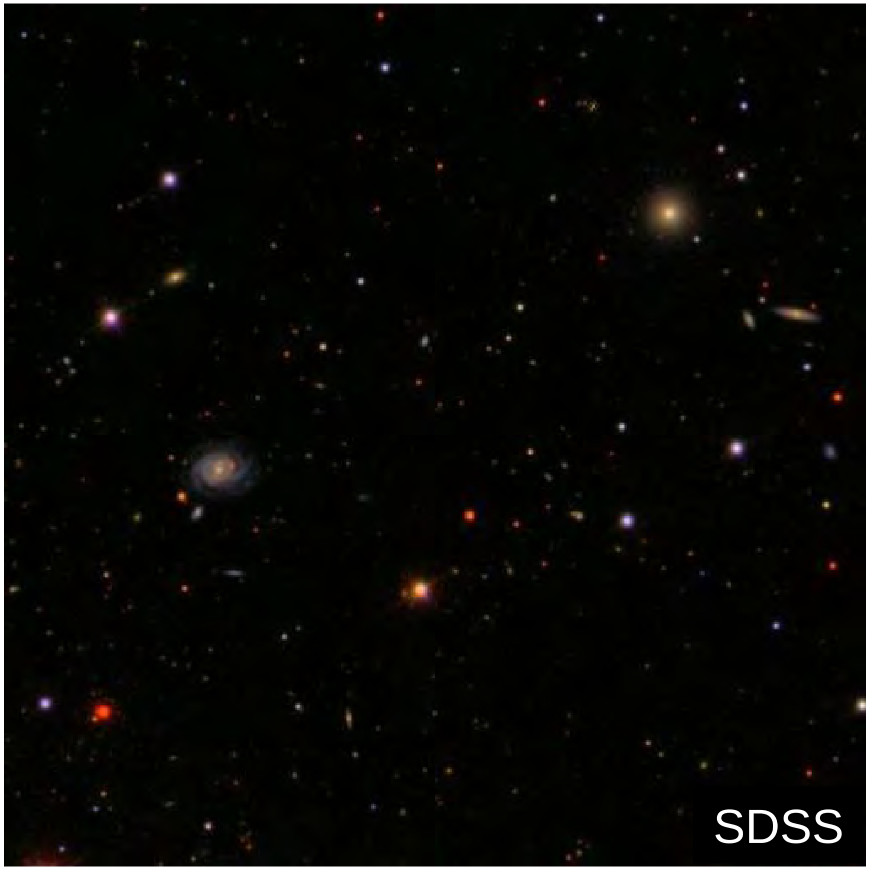
Large Synoptic Survey Telescope (LSST)

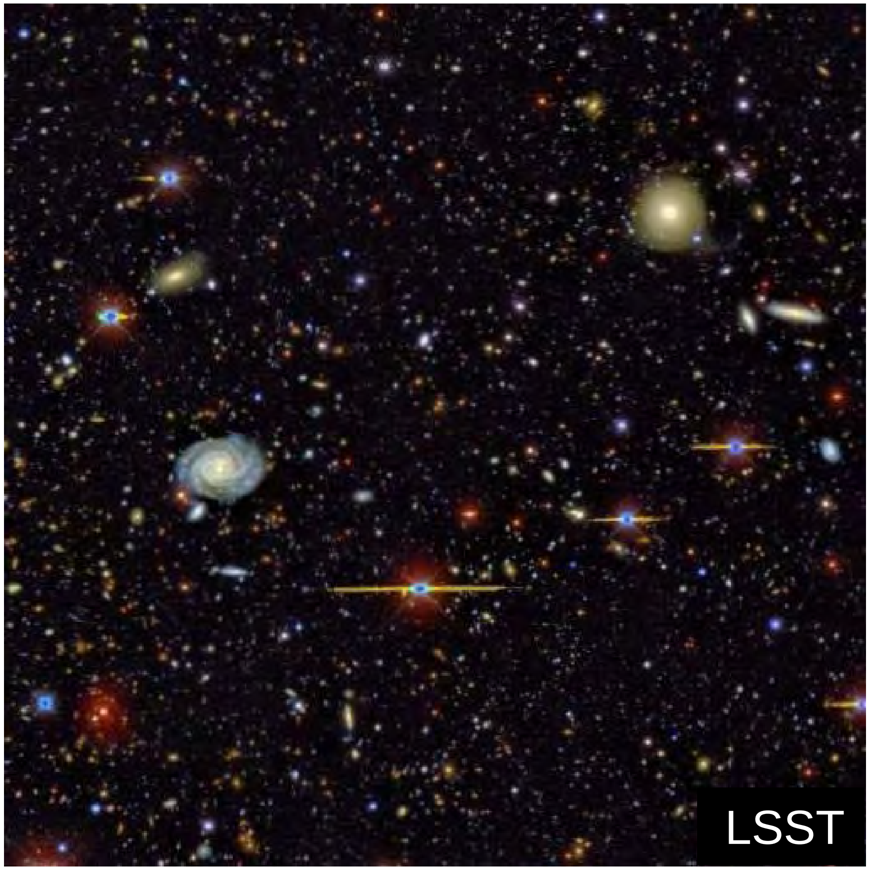
40TB/day (an SDSS every two days), 100+PB in its 10-year lifetime

400mbps sustained data rate between Chile and NCSA



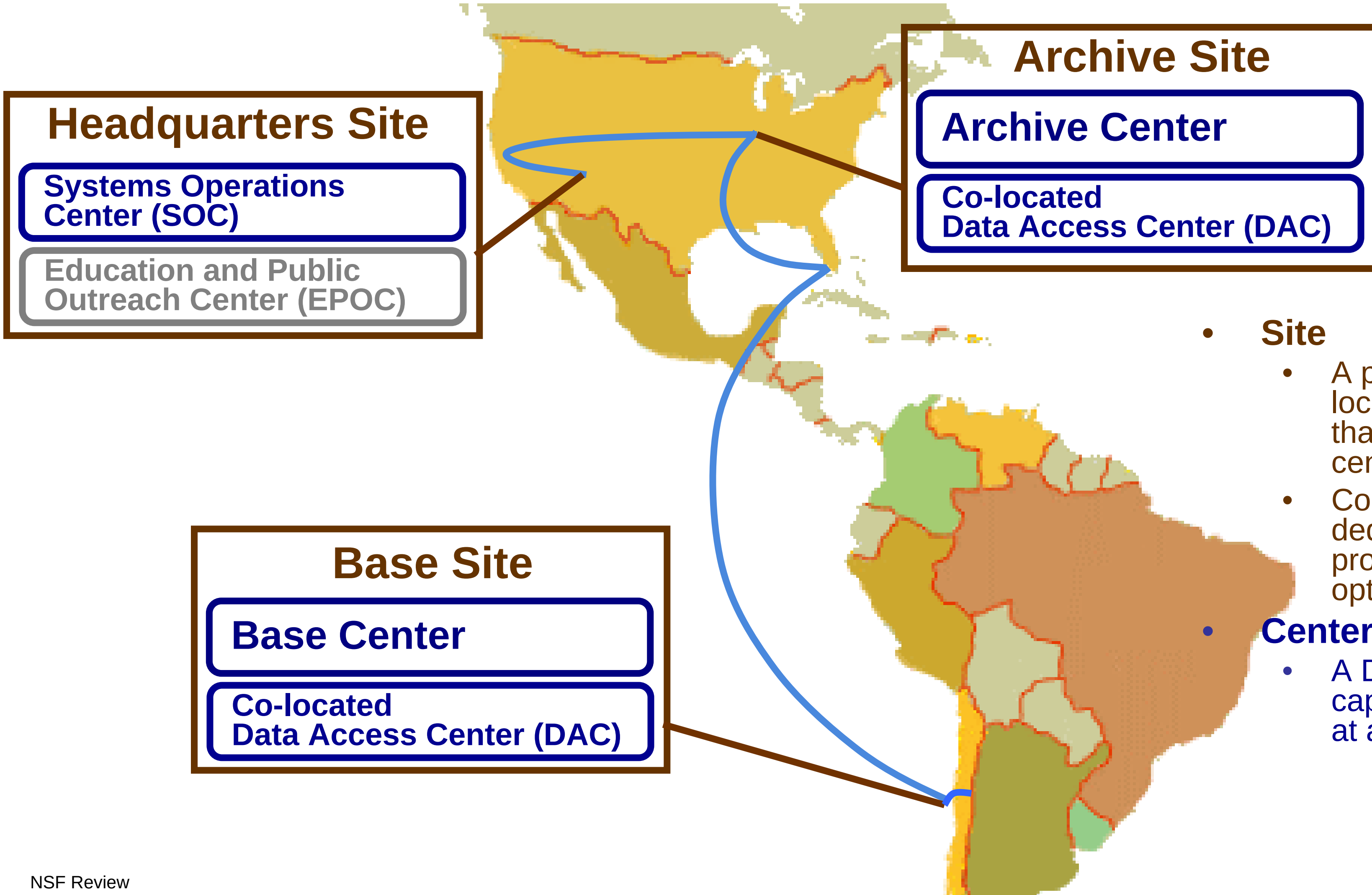
- Why?





[Andy Connolly, University of Washington]

LSST Data Management System is widely distributed

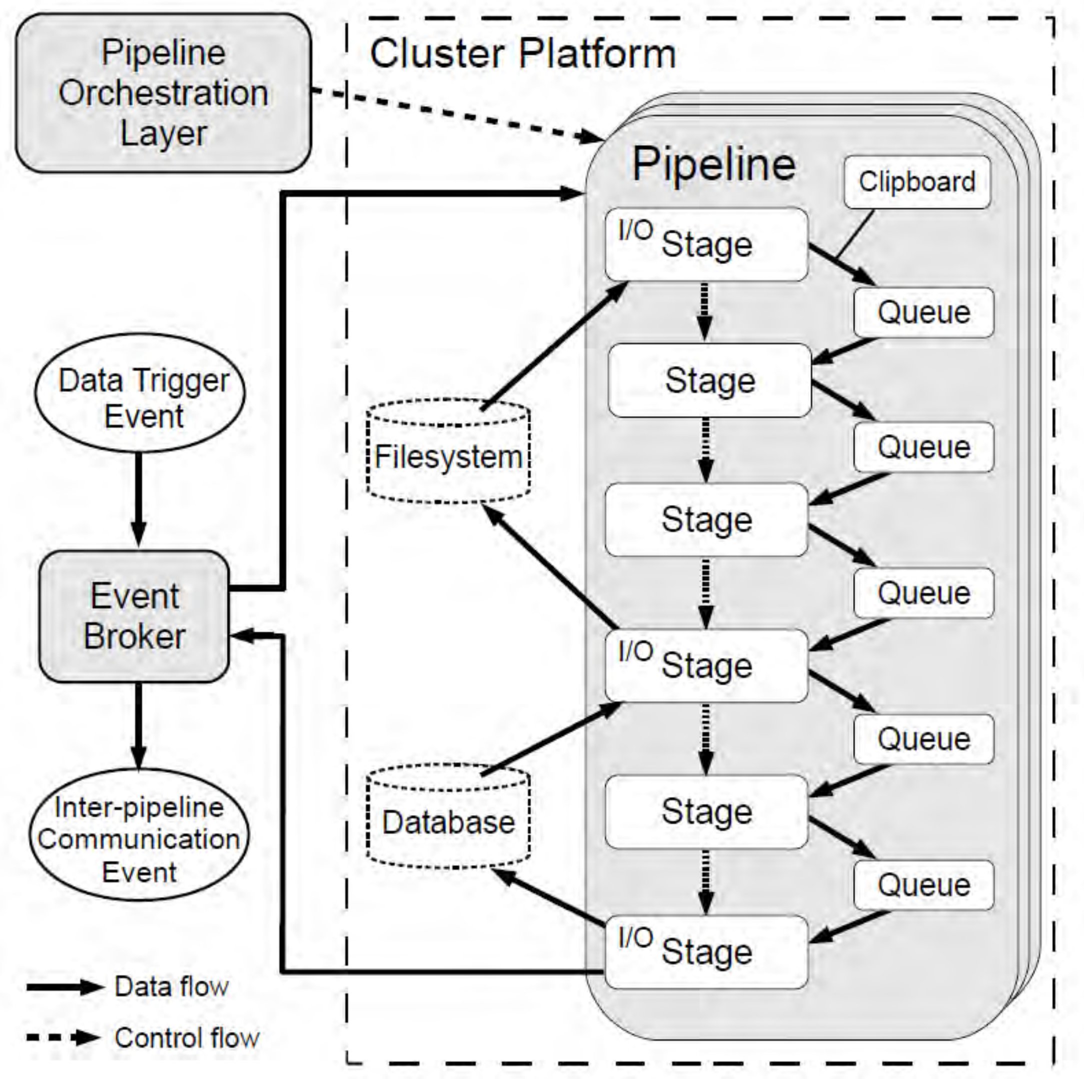


- **Site**
  - A physical location/space that hosts DM centers
  - Connected via dedicated, protected fiber optic circuits
- **Center**
  - A DM functional capability hosted at a Site

[Andy Connolly, University of Washington, and LSST]

LSST Data Management System relies on large-scale parallelism

- **With few exceptions, LSST pipeline processing is “embarrassingly parallel”**
  - **3024 parallel image readouts**
  - **O(10^8) sky tiles**
  - **O(10^9) objects**
- **Computational clusters are well matched to the available parallelism**
  - **5000 cores at Base**
  - **12000 (yr1) – 33000 (yr10) cores at Archive**
- **Middleware implements flexible pipeline/ production model of parallelism**



[Andy Connolly, University of Washington, and LSST]

- Project plan
  - Fully 30% of project budget is allocated to software

But astronomy is substantially ahead of most other fields

- Data management in computational astrophysics
  - `fopen()`
  - `fread()`
  - `fwrite()`
  - `fclose()`
  - `scp`

– Jeff Gardner, UW eScience Institute

Each simulation generates a sequence of snapshots; each snapshot is a single flat file; analysis is via C or Fortran programs

Data management in biology

ANNOTATIONSUMMARY-COMBINEDORFANNOTATION16_Phaeo_genome

###query	length	COG hit #1	e-value #1	identity #1	score #1	hit length #1	description #1
chr_4[480001-580000].287	4500						
chr_4[560001-660000].1	3556						
chr_9[400001-500000].503	4211	COG4547	2.00E-04	19	44.6	620	Cobalamin biosynthesis proteir
chr_9[320001-420000].548	2833	COG5406	2.00E-04	38	43.9	1001	Nucleosome binding factor SPN
chr_27[320001-404298].20	3991	COG4547	5.00E-05	18	46.2	620	Cobalamin biosynthesis proteir
chr_26[320001-420000].378	3963	COG5099	5.00E-05	17	46.2	777	RNA-binding protein of the Puf
chr_26[400001-441226].196	2949	COG5099	2.00E-04	17	43.9	777	RNA-binding protein of the Puf
chr_24[160001-260000].65	3542						
chr_5[720001-820000].339	3141	COG5099	4.00E-09	20	59.3	777	RNA-binding protein of the Puf
chr_9[160001-260000].243	3002	COG5077	1.00E-25	26	114	1089	Ubiquitin carboxyl-terminal hyd
chr_12[720001-820000].86	2895	COG5032	2.00E-09	30	60.5	2105	Phosphatidylinositol kinase and
chr_12[800001-900000].109	1463	COG5032	1.00E-09	30	60.1	2105	Phosphatidylinositol kinase and
chr_11[1-100000].70	2886						
chr_11[80001-180000].100	1523						

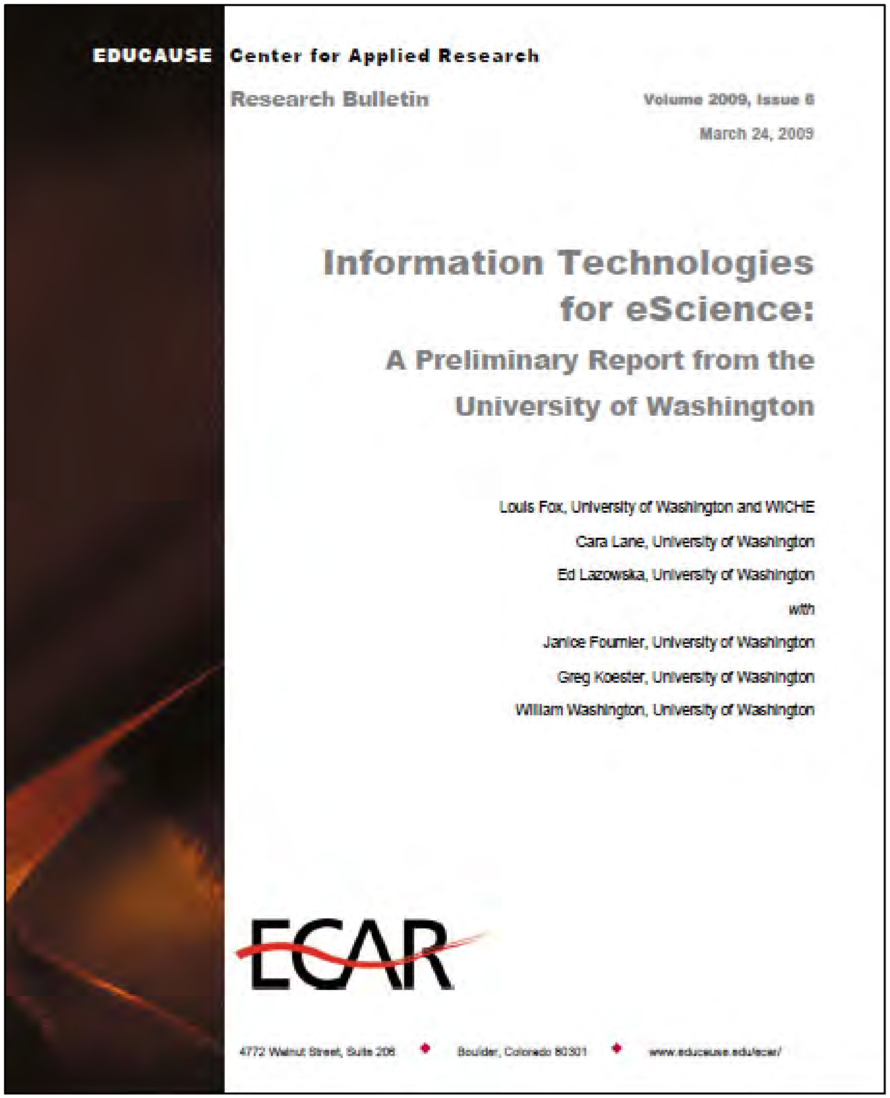
COGAnnotation_coastal_sample.txt

id	query	hit	e_value	identity_	score	query_start	query_end	hit_start	hit_end	hit_length
1	FHJ7DRN01A0TND.1	COG0414	1.00E-08	28	51	1	74	180	257	285
2	FHJ7DRN01A1AD2.2	COG0092	3.00E-20	47	89.9	6	85	41	120	233
3	FHJ7DRN01A2HWZ.4	COG3889	0.0006	26	35.8	9	94	758	845	872
	...									
2853	FHJ7DRN02HXTBY.5	COG5077	7.00E-09	37	52.3	3	77	313	388	1089
2854	FHJ7DRN02HZO4J.2	COG0444	2.00E-31	67	127	1	73	135	207	316
	...									
3566	FHJ7DRN02FUJW3.1	COG5032	1.00E-09	32	54.7	1	75	1965	2038	2105
	...									

90% of all business data is maintained in spreadsheets

– Enrique Godreau, Voyager Capital

Top faculty across all disciplines understand and fear the coming data tsunami



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/

- 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
  - Help position the University of Washington 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

- Technical staff


David Beck


Jeff Gardner


Bill Howe

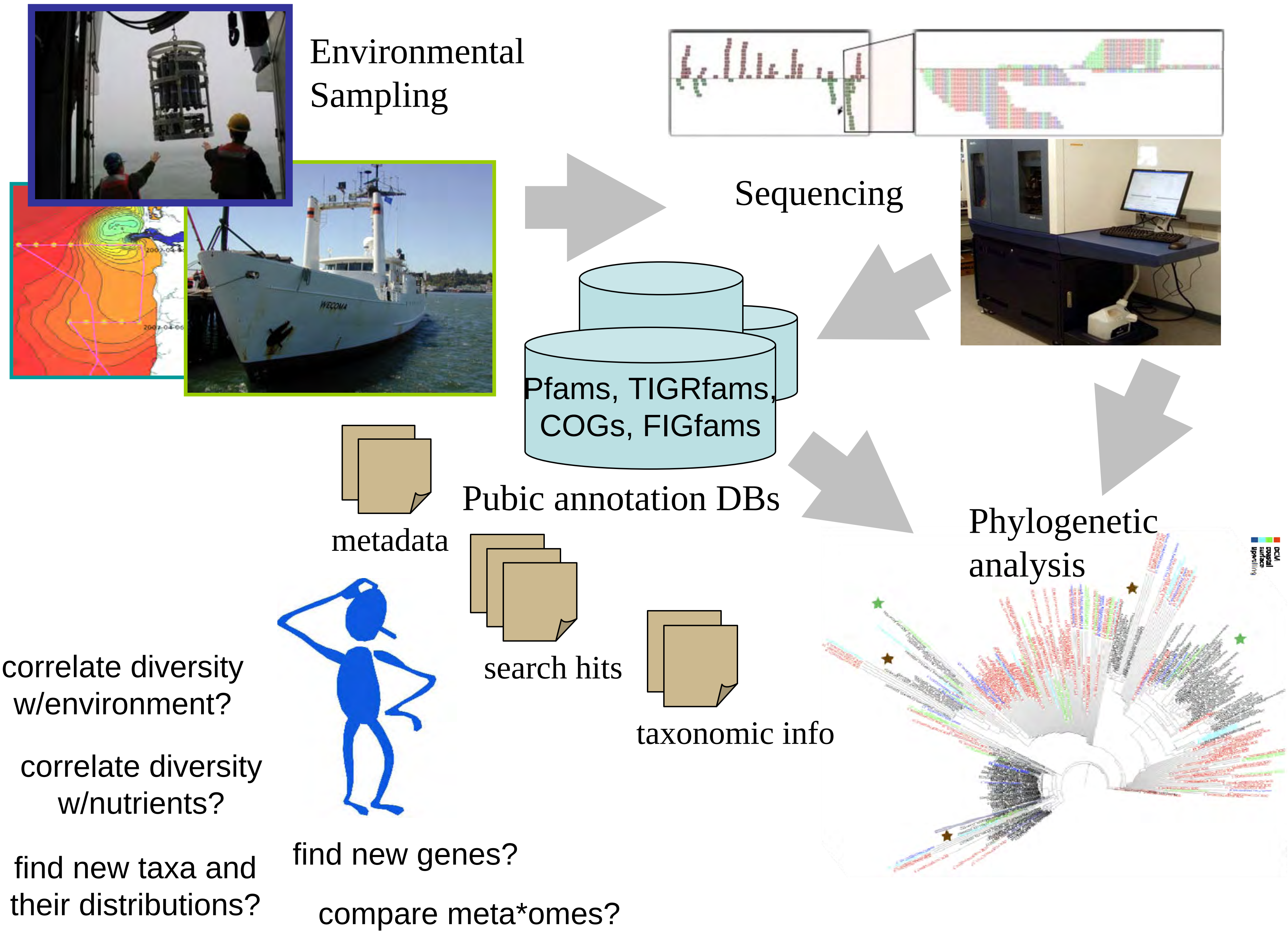

Erik Lundberg

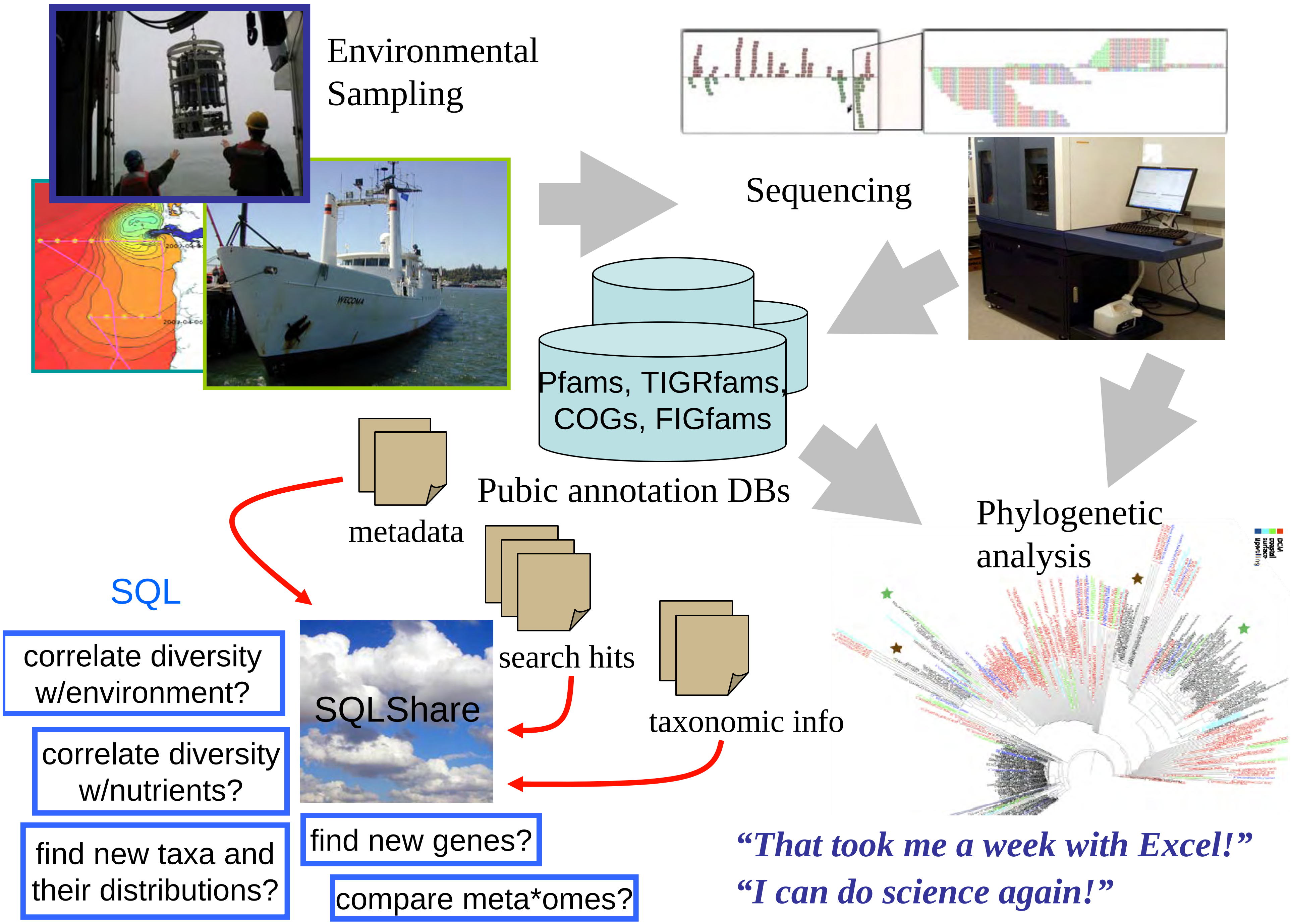

Chance Reschke

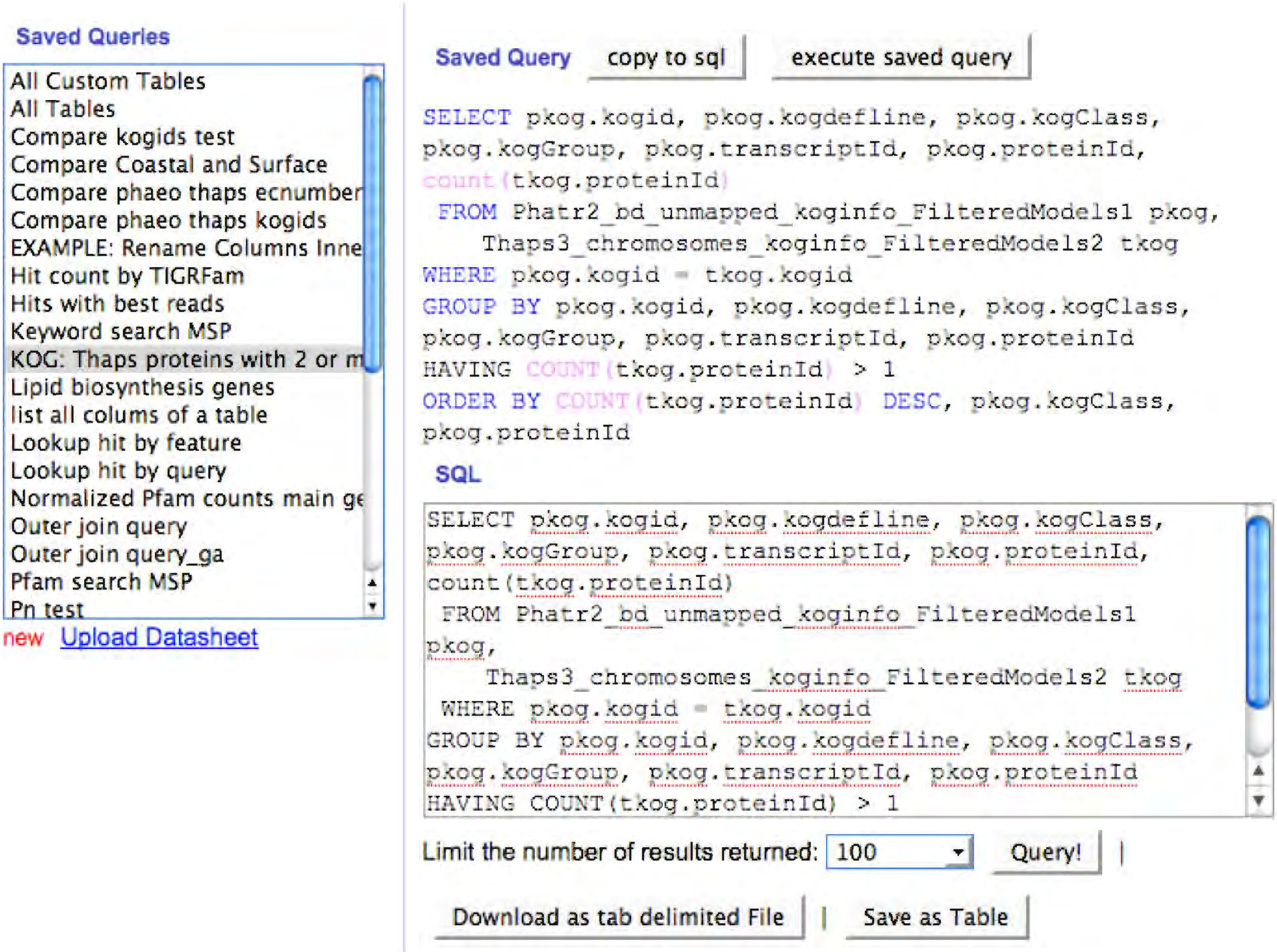
Environmental metagenomics / metatranscriptomics / metaproteomics

Ginger Armbrust

- Study microbial populations sampled from the environment instead of individual organisms
  - Who is there?
    - Which organisms make up the population?
  - What are they doing?
    - Which metabolic pathways are present and active (and who is doing what)?
  - Compare datasets
    - Across a transect (nearshore vs. deep ocean)
    - Before/after some event (e.g., Spring flooding)
    - Across salinity/temperature gradients
    - Diurnal cycles (day/night)


Environmental Sampling
Sequencing
Pfams, TIGRfams, COGs, FIGfams
Pubic annotation DBs
metadata
search hits
taxonomic info
Phylogenetic analysis
correlate diversity w/environment?
correlate diversity w/nutrients?
find new taxa and their distributions?
find new genes?
compare meta*omes?


Environmental Sampling
Sequencing
Pfams, TIGRfams, COGs, FIGfams
Pubic annotation DBs
Phylogenetic analysis
metadata
search hits
taxonomic info
SQL
SQLShare
correlate diversity w/environment?
correlate diversity w/nutrients?
find new taxa and their distributions?
find new genes?
compare meta*omes?
"That took me a week with Excel!"
"I can do science again!"



Saved Queries
All Custom Tables
All Tables
Compare kogids test
Compare Coastal and Surface
Compare phaeo thaps ecnumber
Compare phaeo thaps kogids
EXAMPLE: Rename Columns Inne
Hit count by TIGRFam
Hits with best reads
Keyword search MSP
KOG: Thaps proteins with 2 or m
Lipid biosynthesis genes
list all colums of a table
Lookup hit by feature
Lookup hit by query
Normalized Pfam counts main ge
Outer join query
Outer join query_ga
Pfam search MSP
Pn test
new
Upload Datasheet
Saved Query
copy to sql
execute saved query
SELECT pkog.kogid, pkog.kogdefline, pkog.kogClass, pkog.kogGroup, pkog.transcriptId, pkog.proteinId, count(tkog.proteinId)
FROM Phatr2_bd_unmapped_koginfo_FilteredModels1 pkog, Thaps3_chromosomes_koginfo_FilteredModels2 tkog
WHERE pkog.kogid = tkog.kogid
GROUP BY pkog.kogid, pkog.kogdefline, pkog.kogClass, pkog.kogGroup, pkog.transcriptId, pkog.proteinId
HAVING COUNT(tkog.proteinId) > 1
ORDER BY COUNT(tkog.proteinId) DESC, pkog.kogClass, pkog.proteinId
SQL
SELECT pkog.kogid, pkog.kogdefline, pkog.kogClass, pkog.kogGroup, pkog.transcriptId, pkog.proteinId, count(tkog.proteinId)
FROM Phatr2_bd_unmapped_koginfo_FilteredModels1 pkog,
Thaps3_chromosomes_koginfo_FilteredModels2 tkog
WHERE pkog.kogid = tkog.kogid
GROUP BY pkog.kogid, pkog.kogdefline, pkog.kogClass, pkog.kogGroup, pkog.transcriptId, pkog.proteinId
HAVING COUNT(tkog.proteinId) > 1
Limit the number of results returned:
100
Query!
Download as tab delimited File
Save as Table

Your query generated 100 result(s)

kogid	kogdefline	kogClass	kogGroup	transcriptId	proteinId	Column1
KOG2992	Nucleolar GTPase/ATPase p130	Nuclear structure	CELLULAR PROCESSES AND SIGNALING	1437	1437	302
KOG2992	Nucleolar GTPase/ATPase p130	Nuclear structure	CELLULAR PROCESSES AND SIGNALING	1553	1553	302
KOG1216	von Willebrand factor and related coagulation proteins	Defense mechanisms	CELLULAR PROCESSES AND SIGNALING	1435	1435	202
KOG1216	von Willebrand factor and related coagulation proteins	Defense mechanisms	CELLULAR PROCESSES AND SIGNALING	1718	1718	202
KOG1216	von Willebrand factor and related coagulation proteins	Defense mechanisms	CELLULAR PROCESSES AND SIGNALING	1760	1760	202
KOG1216	von Willebrand factor and related coagulation proteins	Extracellular structures	CELLULAR PROCESSES AND SIGNALING	1435	1435	202
KOG1216	von Willebrand factor and related coagulation proteins	Extracellular structures	CELLULAR PROCESSES AND SIGNALING	1718	1718	202
KOG1216	von Willebrand factor and related coagulation proteins	Extracellular structures	CELLULAR PROCESSES AND SIGNALING	1760	1760	202
KOG2806	Chitinase	Carbohydrate transport and metabolism	METABOLISM	1438	1438	191
KOG2806	Chitinase	Carbohydrate transport and metabolism	METABOLISM	1686	1686	191
			INFORMATION			

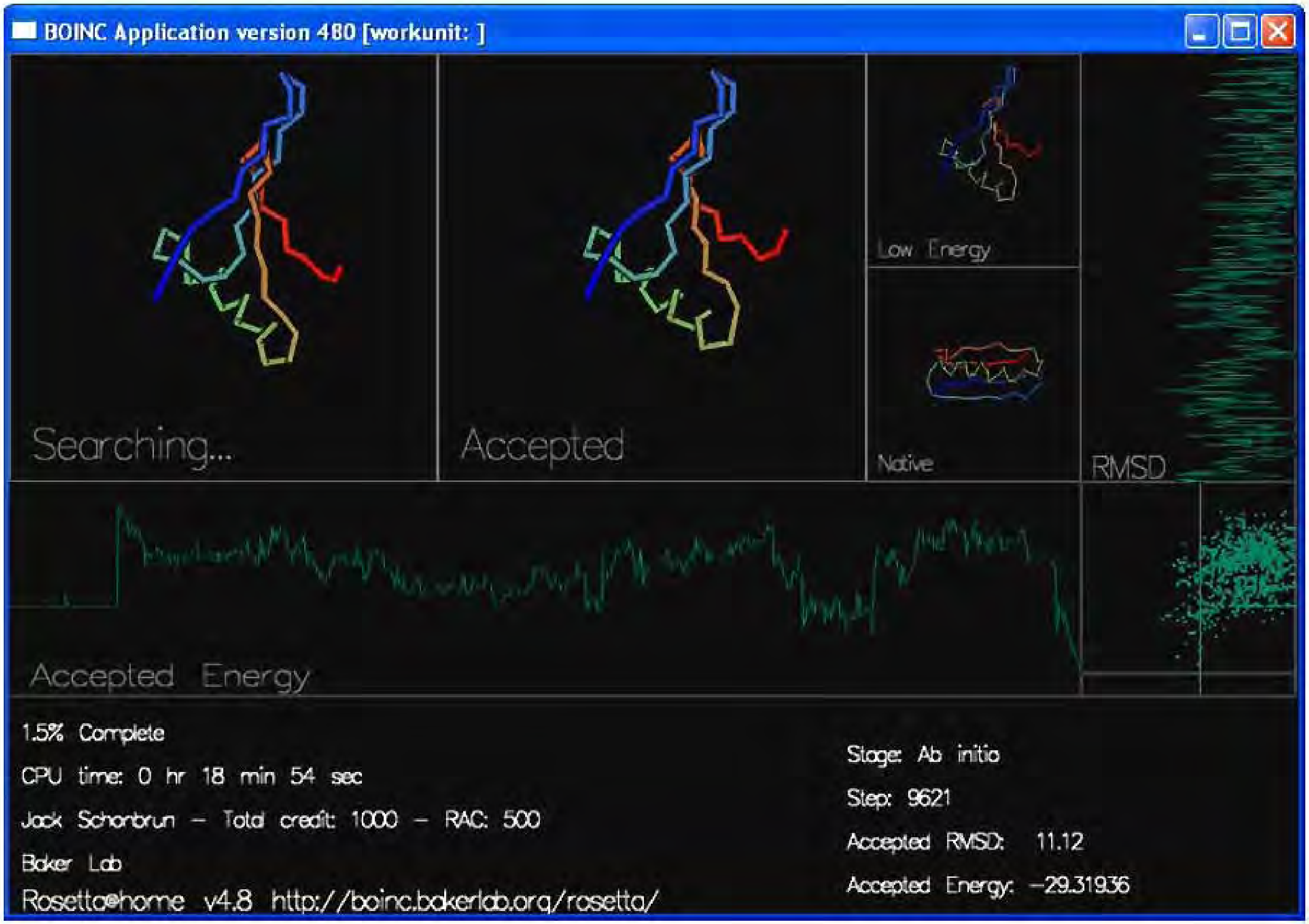
Protein structure prediction and design

David Baker

DARPA
HHMI HOWARD HUGHES MEDICAL INSTITUTE
NSF
NATIONAL INSTITUTES OF HEALTH
Microsoft
Adobe

Rosetta@home

Protein Folding, Design, and Docking




foldit BETA
02:59:51 GMT
Solve Puzzles for Science
BLOG
GROUPS
PLAYERS
PUZZLES
RECIPES
FORUM
WIKI
FEEDBACK
ABOUT
Click to learn how you contribute to science by playing Foldit.
GET STARTED: DOWNLOAD
Win Beta
Win XP/Vista
Mac Beta
Intel OS X 10.4 or later
Linux Beta
Linux
RECOMMEND FOLDIT
Send
USER LOGIN
Username: *
Password: *
Log in
Create new account
Request new password
Sign in using Facebook
Connect with Facebook
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.

fold it BETA

20:46:49 GMT

Solve Puzzles for Science

BLOG GROUPS PLAYERS PUZZLE

BootsMcGraw

Global Soloist Rank: #6

Global Soloist Score: 3784

Cases

Profile

Name: BootsMcGraw

Location: Dallas, Texas USA

Started Folding: 12/06/08

About me: An educated redneck here, from Dallas, Texas.

When I was in grad school in 1985 at the State University of New York at Buffalo, my master's thesis was to construct and present a computer program that predicted the secondary structures (helix, sheet, loop) of proteins based on their amino acid sequences. Tertiary structure (i.e. folding) prediction was a pie-in-the-sky fantasy.

Imagine my delight, a quarter century later, to find out that not only are people determining tertiary structures of proteins, but they've made a *game* of it.

Hobbies: Licensed Massage Therapist; also a photographer, videographer, and webmaster. I have studied health and nutrition for over twenty years. Ask me my opinions about the subject.

Group: Contenders

Arzeda Corporation

New enzymes to drive the industrial biotech revolution

Arzeda.

- ✓ ***Spin-out from UW research group of David Baker from the Dept. of Biochemistry***
- ✓ ***At the convergence of digital biology and green chemistry***
- ✓ ***World leader in the computational design and commercialization of novel, proprietary enzymes***

March 08 Apr. 08 Apr. 09 June 09 Feb. 10

Computational enzyme design methodology in Rosetta

Incorporated in Seattle, WA

Seed investment, WRF Capital

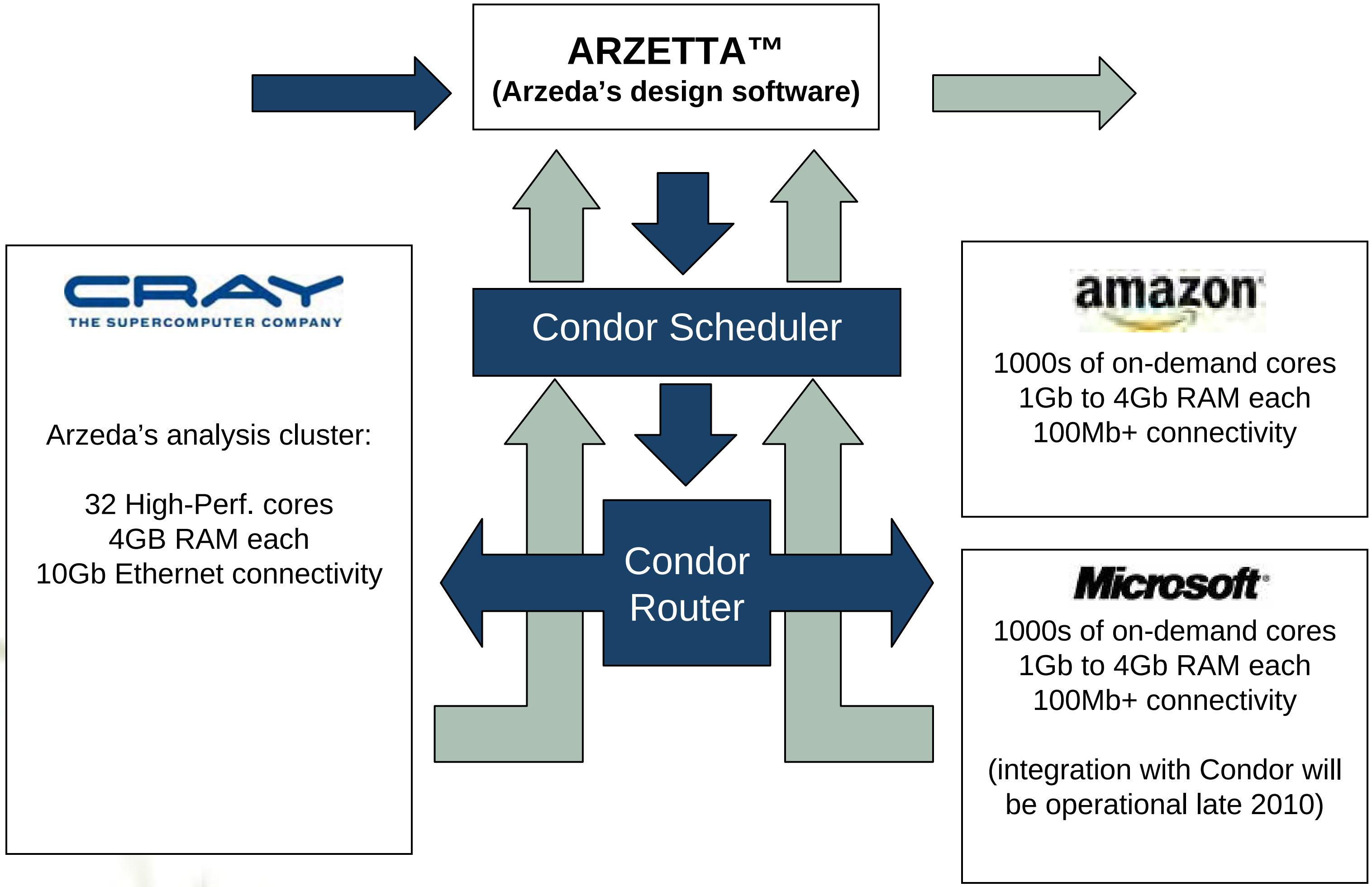
2y research partnership w/ Dupont

IP licenses with UW

NSF and USDA SBIR awards

Arzeda's Platform: The Infrastructure Layer

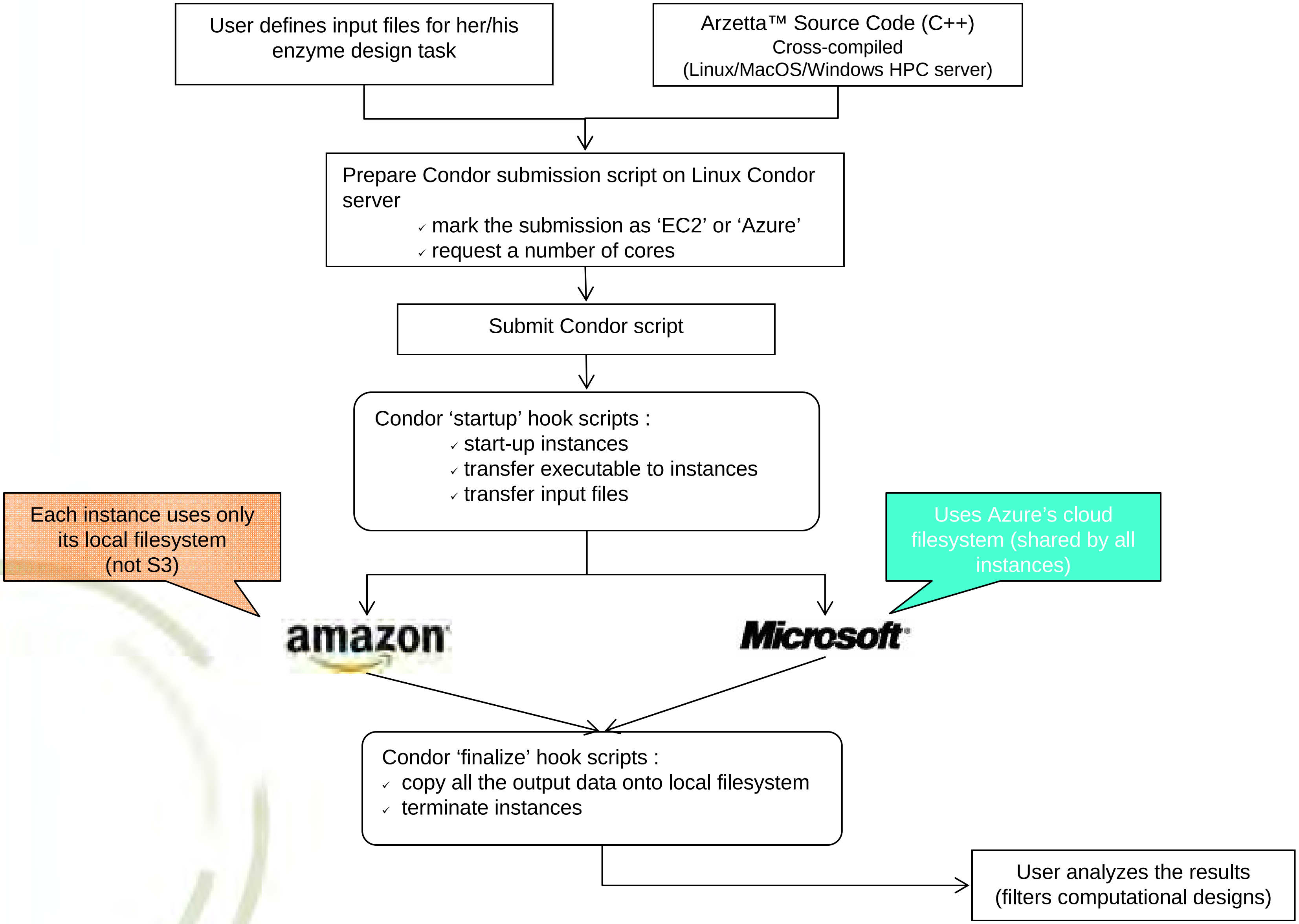
Achieving Scalability through Cloud Computing



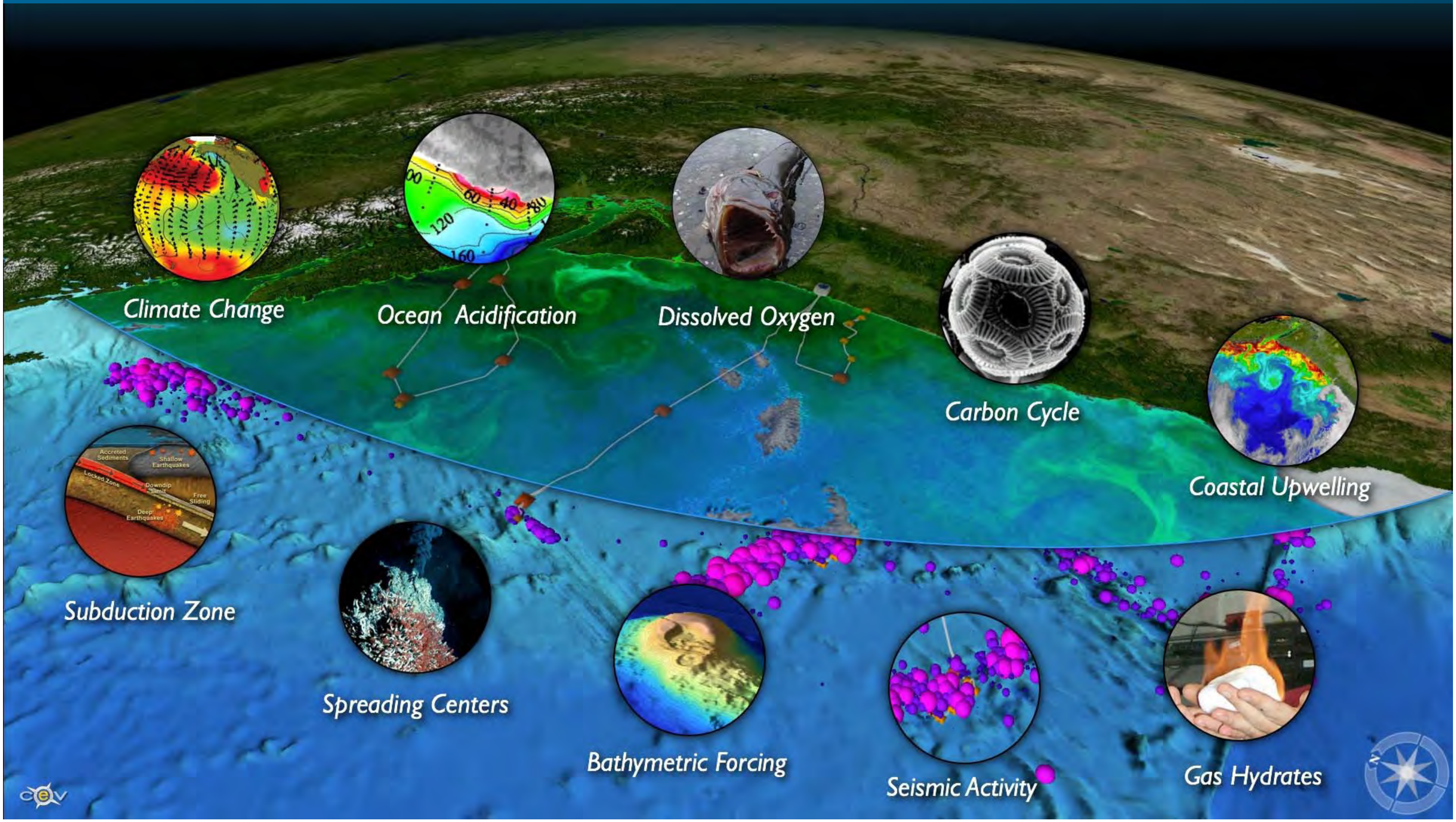
- ✓ *Scalability: immediate scaling to 1000s of cores; only OpEx.*
- ✓ *Price performance: currently $0.08 per hour, going down*
- ✓ *More info on condor: http://www.cs.wisc.edu/condor/*

Arzeda's Cloud Computing Workflow

A Unified Interface to the Cloud based upon Open-Source Tools



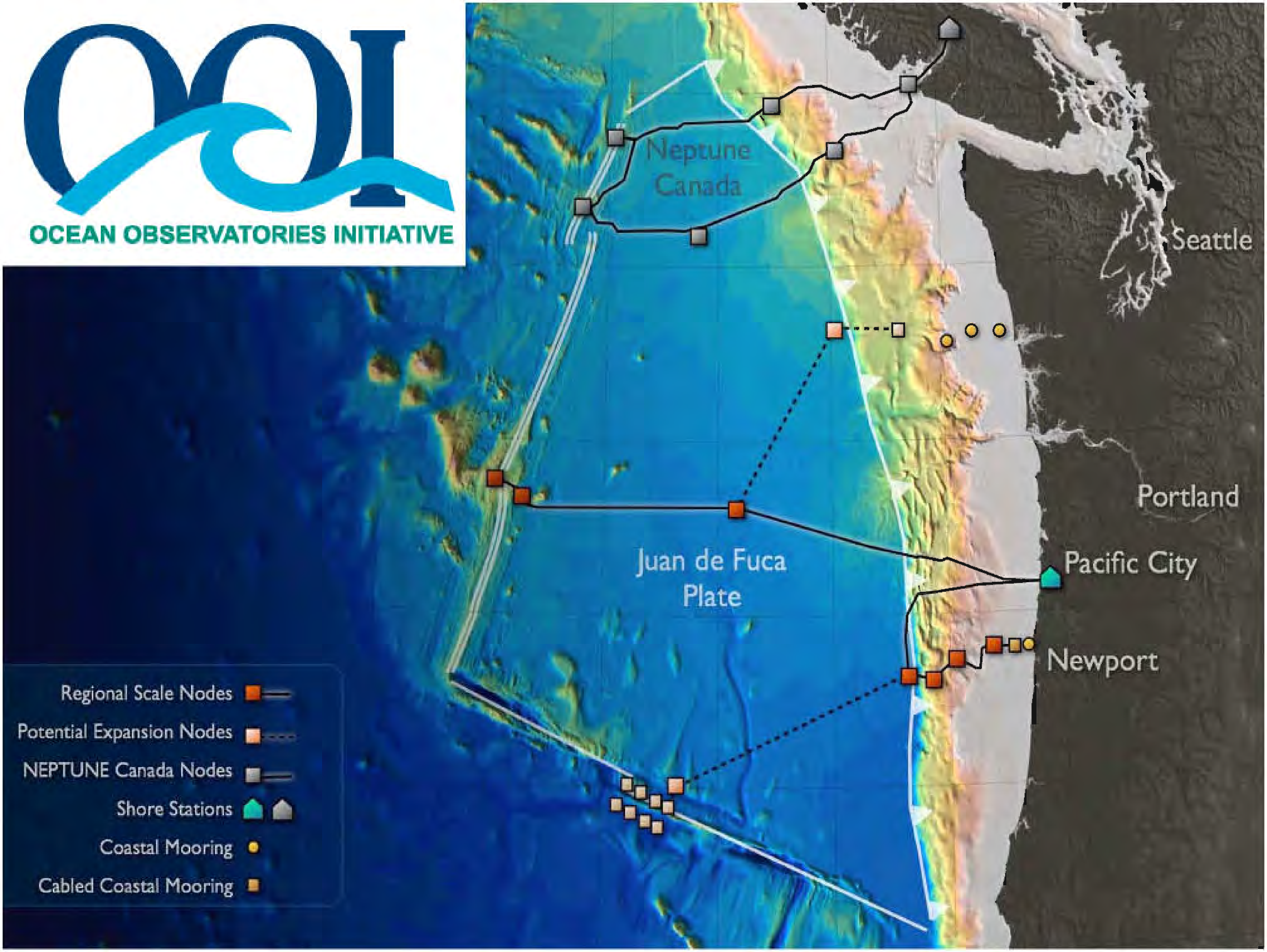
Azure Ocean:
Visualization and Workflow for Ocean Science



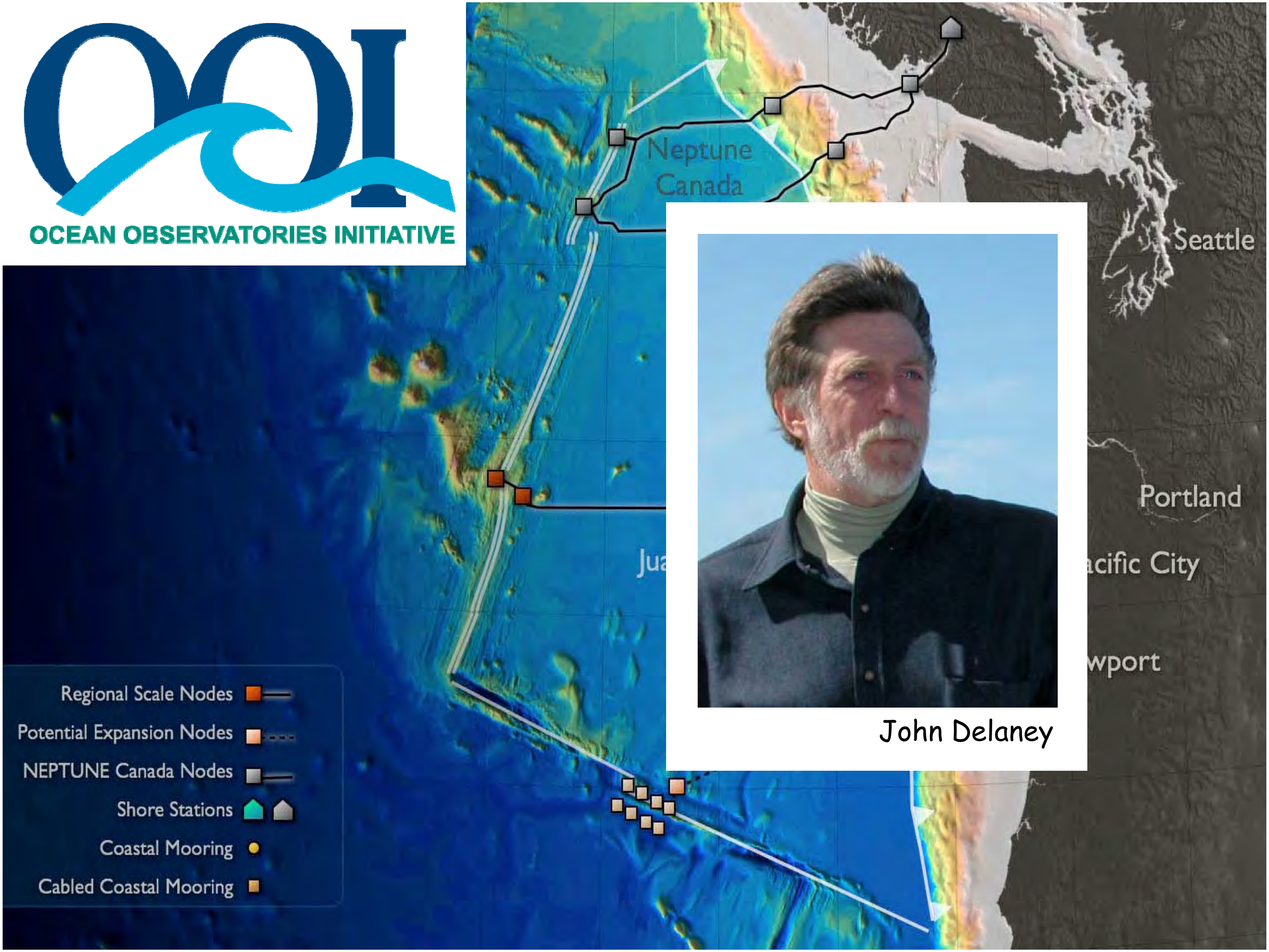

[John Delaney, University of Washington]

OOI

OCEAN OBSERVATORIES INITIATIVE

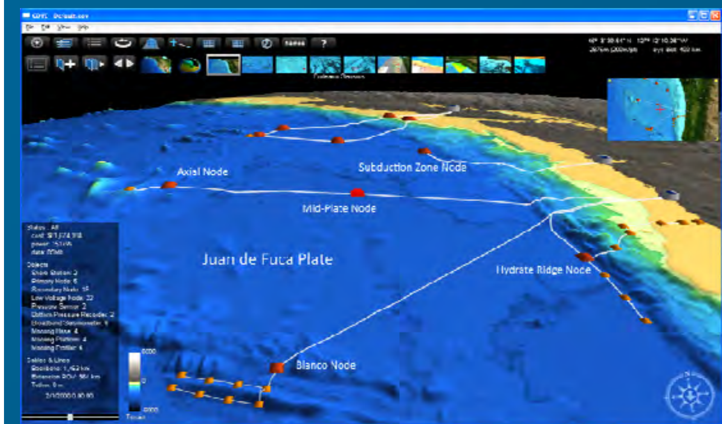


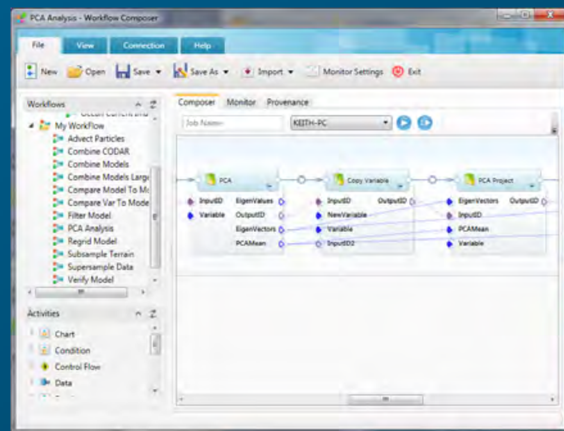
OCEAN OBSERVATORIES INITIATIVE



John Delaney

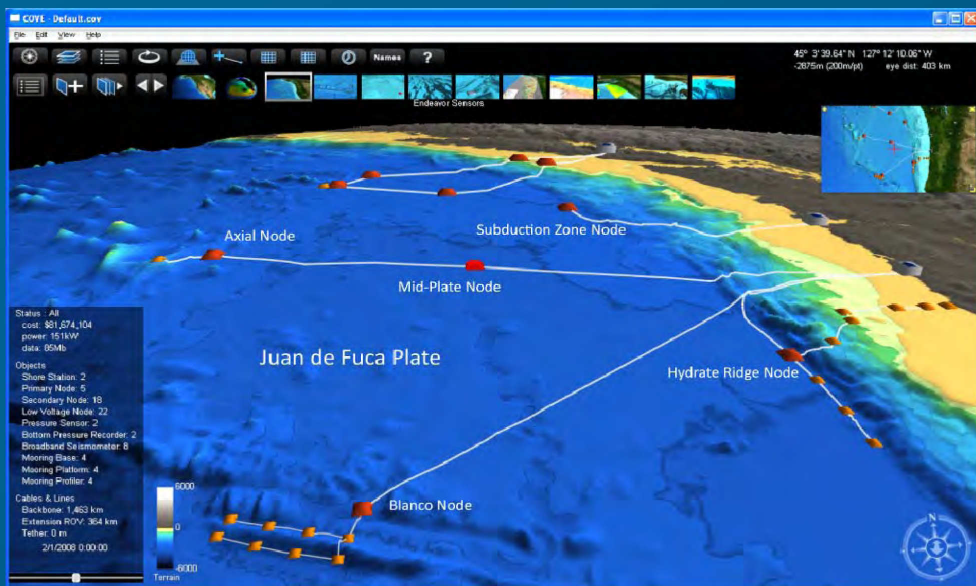
Azure Ocean

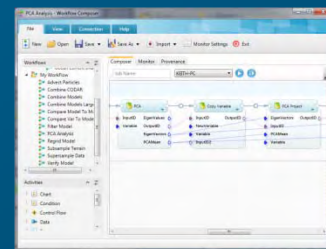




COVE for Visualization + Trident for Processing + Azure for Data

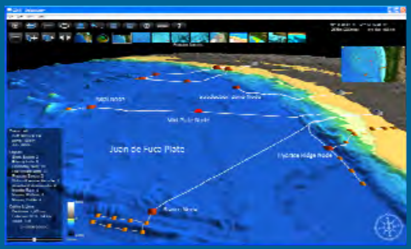
COVE

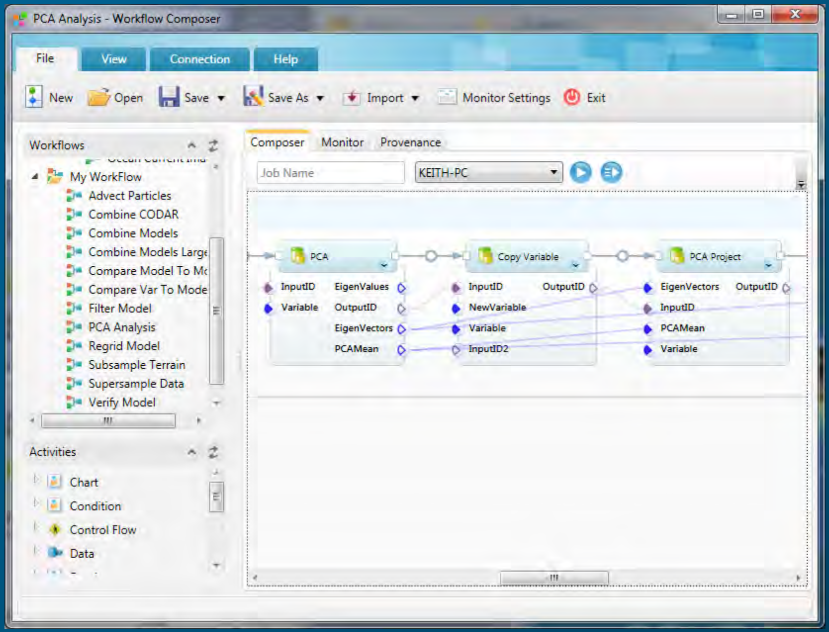




- Research into new interfaces for cross-disciplinary ocean science
- Extensive instrument and cable layout for creating experiments
- Flexible terrain and image engine for visualizing site
- True 3D/4D science dataset visualization
- Field tested in RSN observatory layout and on ocean expeditions
- Cross platform and extensible with python and workflow systems

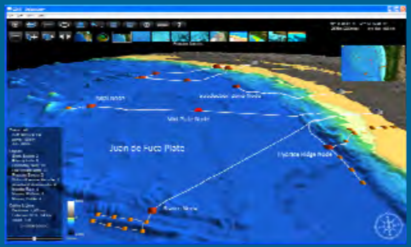
Trident

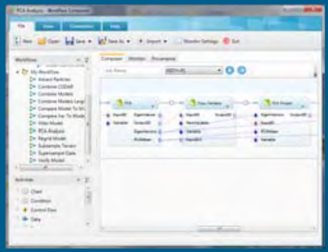




- Microsoft Research scientific workflow system
- Visual programming environment for connecting tasks
- Science-specific task libraries including one for ocean sciences
- Automated provenance capture, monitoring, and fault tolerance
- Runs on local system, Windows server, or HPC Cluster
- Cross platform with Silverlight and web service interface

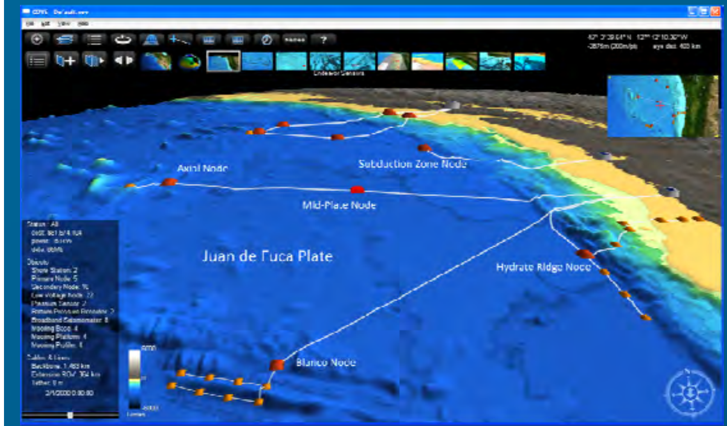
Azure

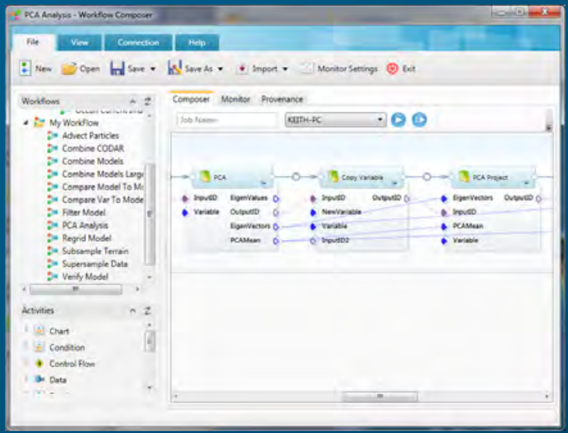




- Microsoft’s cloud computing platform
- Provides storage and computing as pay-as-you-go services
- From development standpoint, system looks like provisioned VM’s
- SQL, table, and blob (file system) storage models are included
- Access to storage via RESTful HTTP interface

Azure Ocean

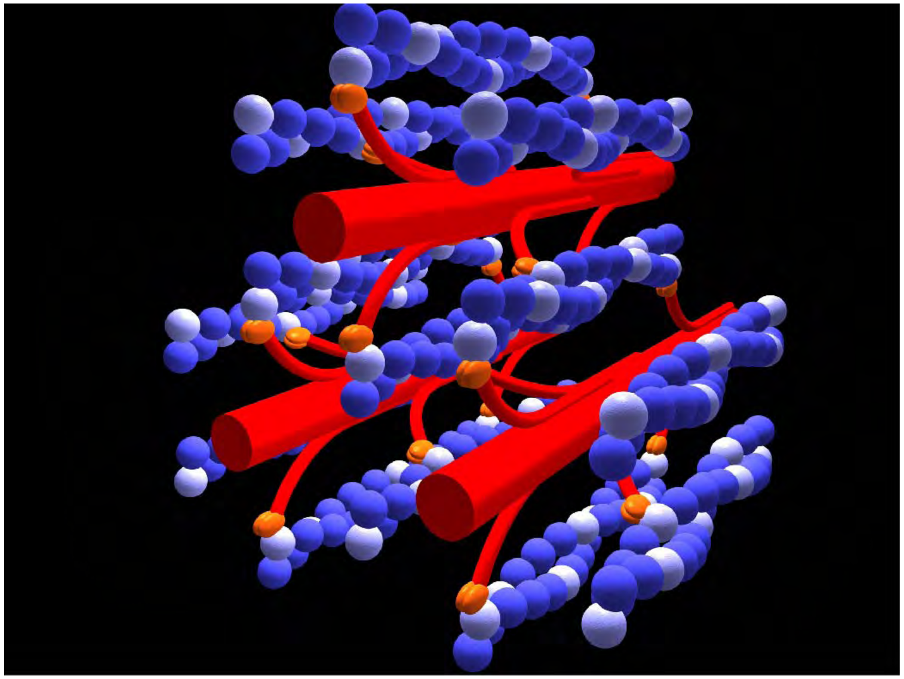




- COVE + Trident + Azure provides visual analytics to scientists
- Any component – *Visualization*, *Computing*, or *Data* – can be provisioned locally, on a server, or in the cloud
- When on same machine, system APIs are leveraged for speed
- When distributed, communication is through HTTP and RESTful APIs
- Flexible platform for the diverse ocean science needs

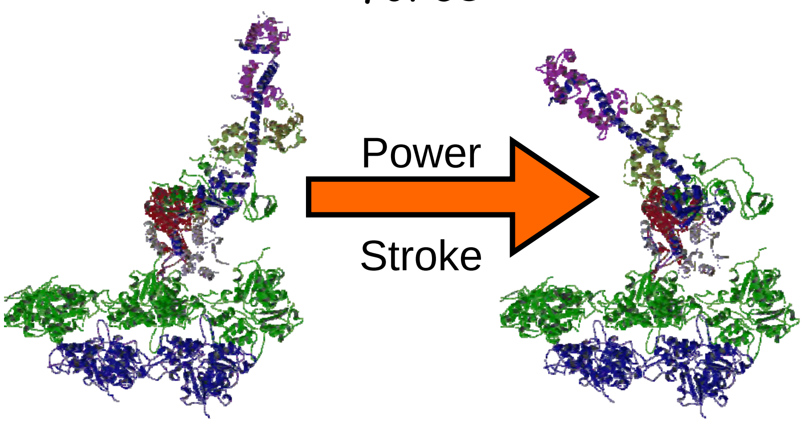
Modeling protein interactions in striated muscles

Tom Daniel

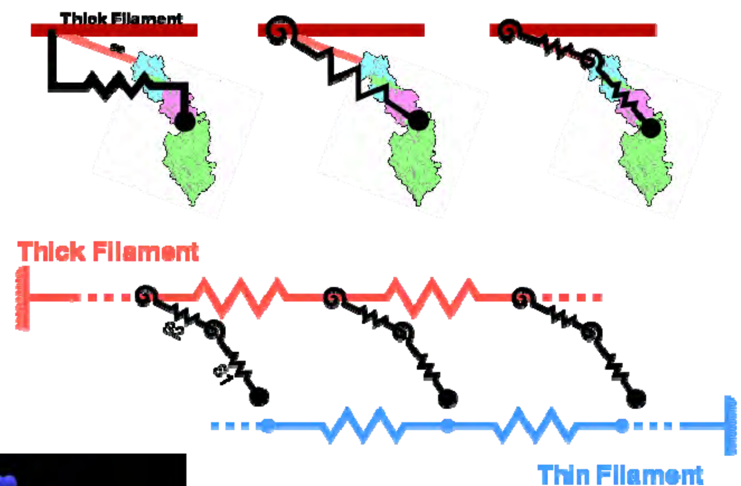


NATIONAL INSTITUTES OF HEALTH

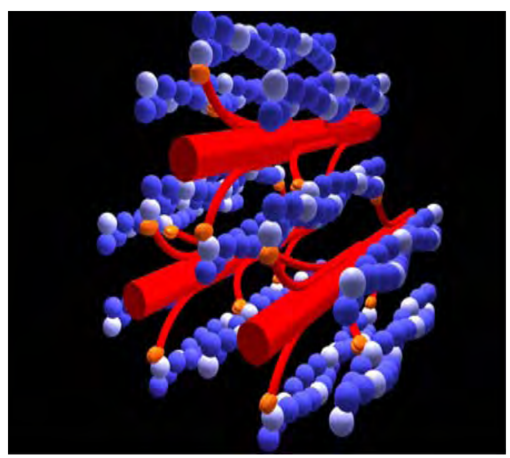
Myosin's lever-arm generates force

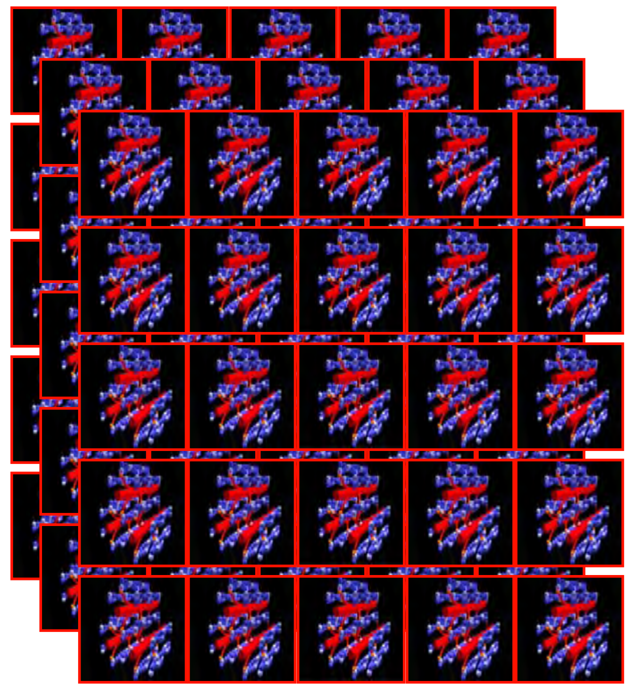


Model the lever arm with multiple springs



Incorporate into a multi-filament model (an embarrassingly parallel Monte Carlo simulation)

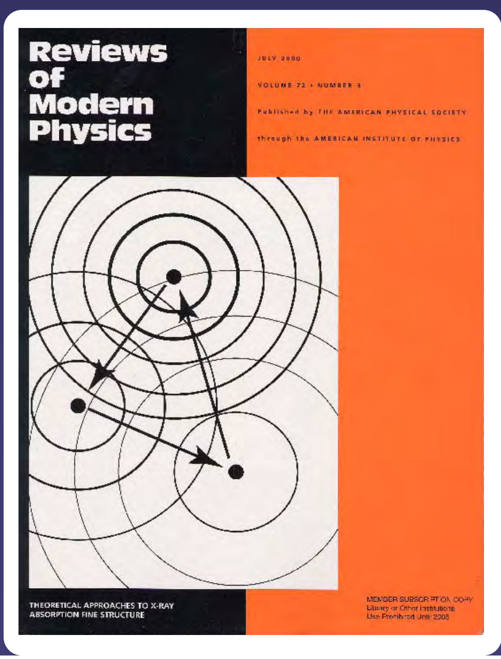




Simple Python scripts automate the management of 1000s of simultaneous experiments using EC2 API

FEFF: Real-space Green's function code for electronic structure, x-ray spectra, …

John Rehr



A "cluster to cloud" story:

Naturally parallel

Each CPU calculates a few points in the energy grid

Loosely coupled

Very little communication between processes

J. J. Rehr & R.C. Albers
Rev. Mod. Phys. **72**, 621 (2000)

http://leonardo.phys.washington.edu/feff/

Challenge of NSF Grant

- Is Cloud Computing feasible for on-demand, High-Performance Computing (HPC) for scientific research in the face of declining budgets?
- Who is interested?
- Is it for everybody?
- What kind of code could benefit from it?
- How do we make it possible?

Disadvantages of Current HPC Approach

- Expensive infrastructure:

  Big clusters = ~1000$/node + capital costs + power + cooling + …
- Expensive HPC staff & maintenance
- Need expertise in HPC to use efficiently

Advantages of CC for Scientific Computing

- For “casual” HPC users:
  - On-demand access without the need to purchase, maintain, or even understand HPCs
  - Lease *vs.* buy: lease as many as needed at ~10¢/cpu-hr
  - Plug & Play HPC scientific codes
- For developers:
  - Scientific codes can be optimized and pre-installed
- For administrators & funding agencies:
  - HPC access to a wider class of scientists at lower costs

Development Strategy

1. Develop AMI (Amazon Machine Image) customized for HPC scientific applications
2. Test single-instance performance
3. Develop shell-scripts that make the EC2 look and run like a local HPC cluster ("virtual supercomputer on a laptop")
4. Test parallel performance

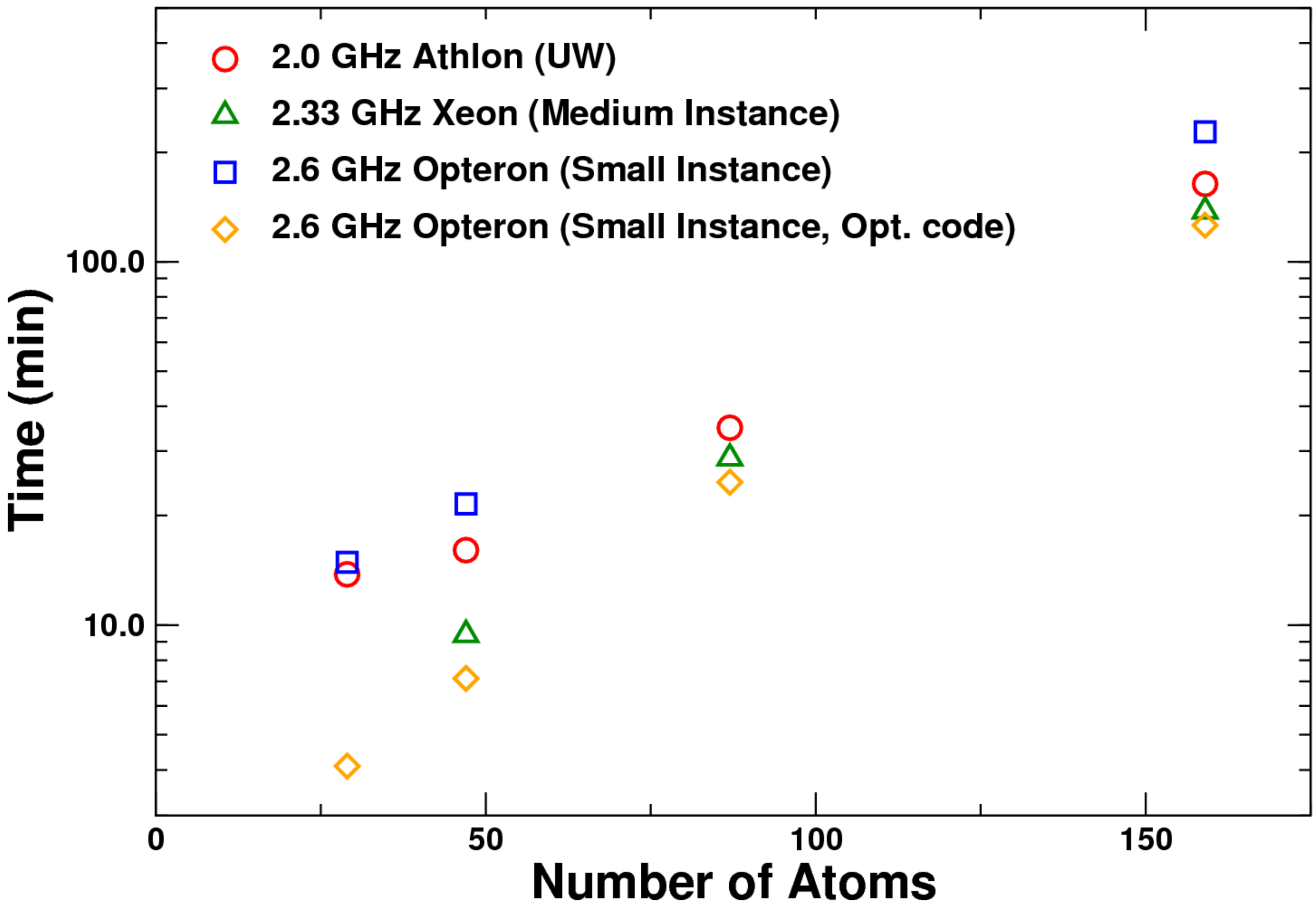
FEFFMPI EC2 AMI

Custom Linux distribution replicated on each instance in cluster

- Standard Linux AMI:

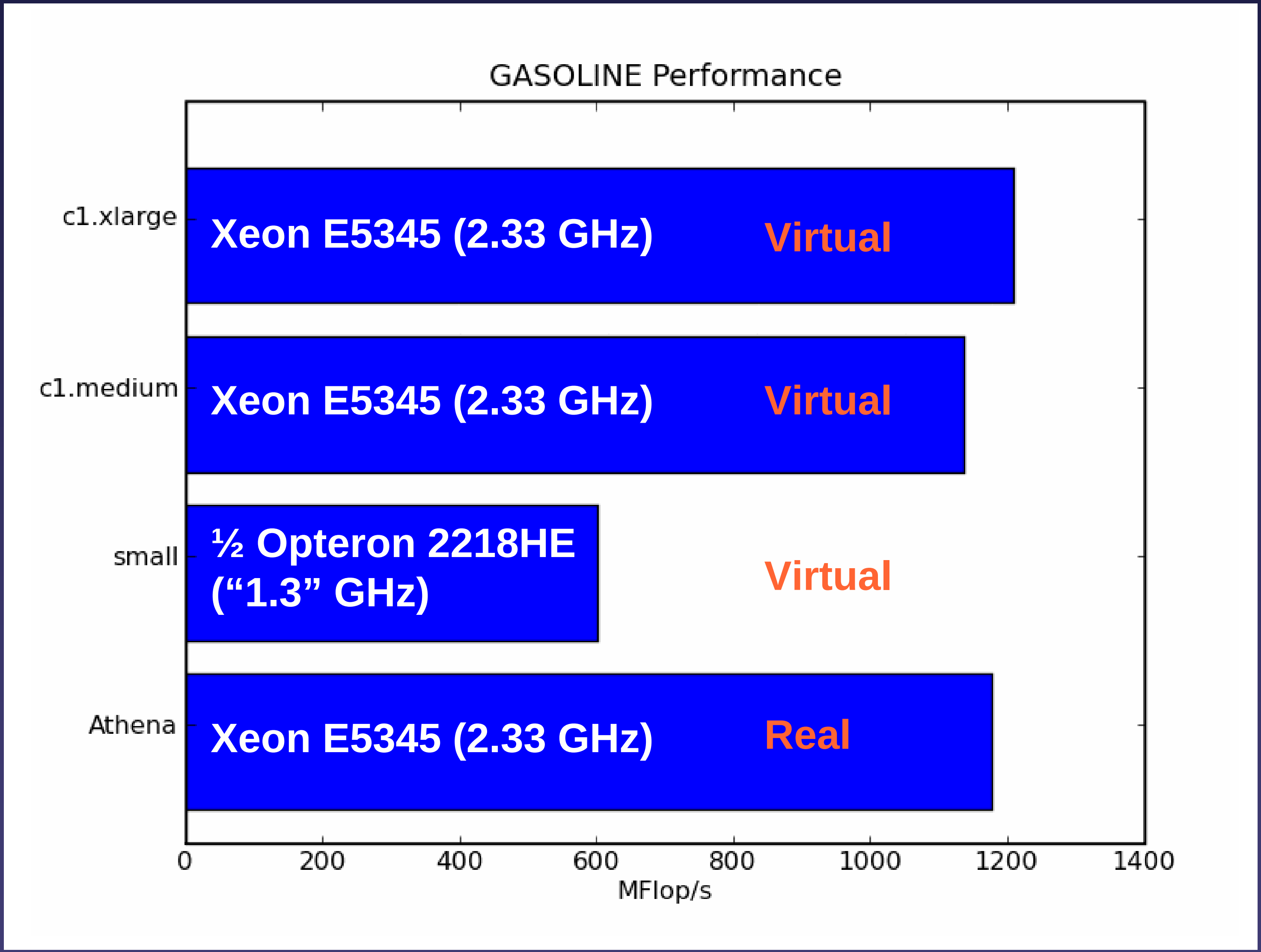
  Fedora 8 32-bit distribution with Gnu FORTRAN compilers (gfortran and g77)
- AWS tools for the EC2: AMI, API and S3 tools
- LAM 7.1.4 for parallel MPI codes
- Java Runtime Environment 6
- Java Development Kit 1.6
- EC2 Cluster tools
- FEFF8.4 serial and parallel versions
- JFEFF graphical interface for FEFF8.4

Serial Performance of FEFF on EC2



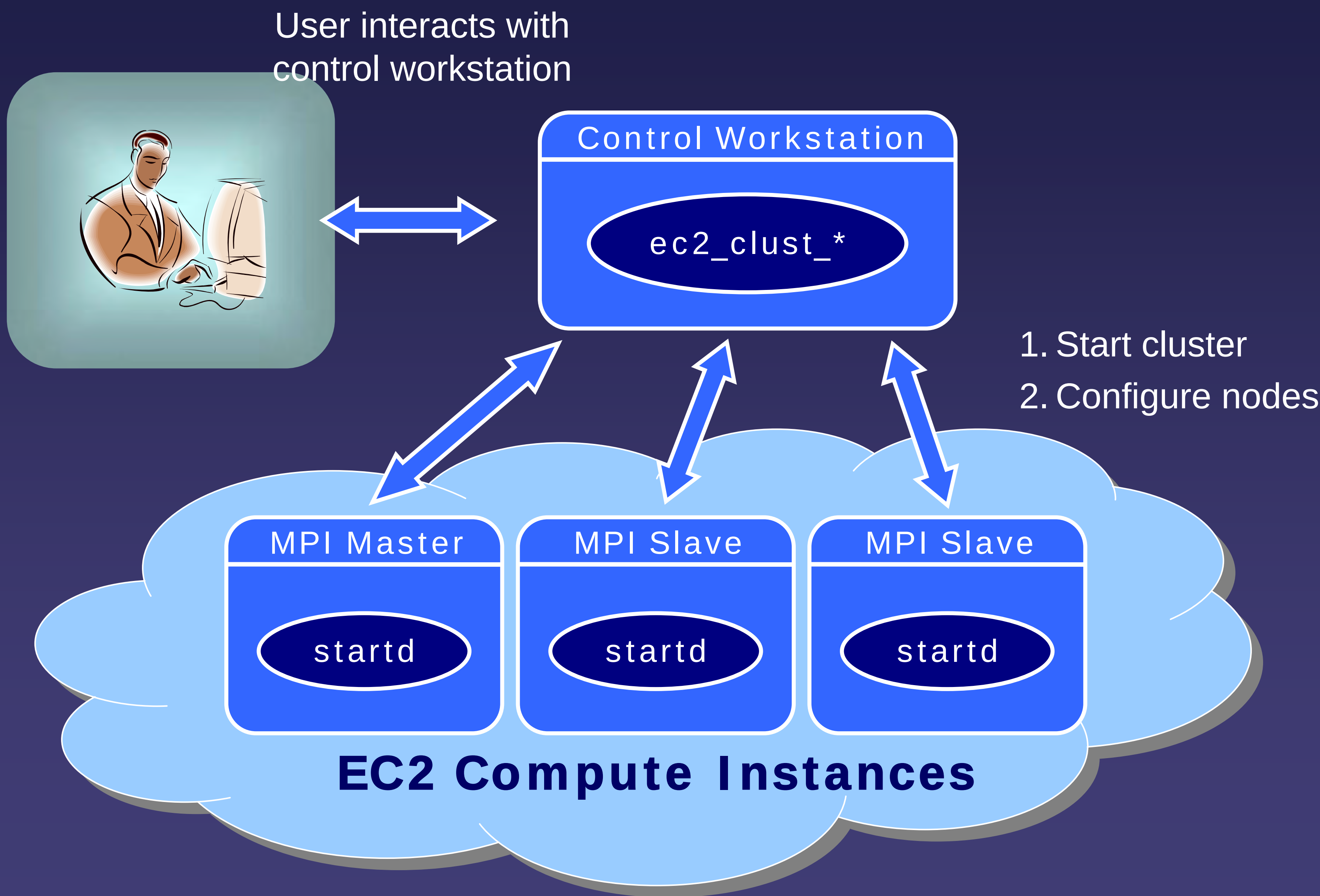
Virtual machine performance similar to “real”

Serial Performance of Gasoline on EC2



No penalty from virtualization

Current MPI Scenario



UW EC2 Cluster Tools

Tools in the local control machine

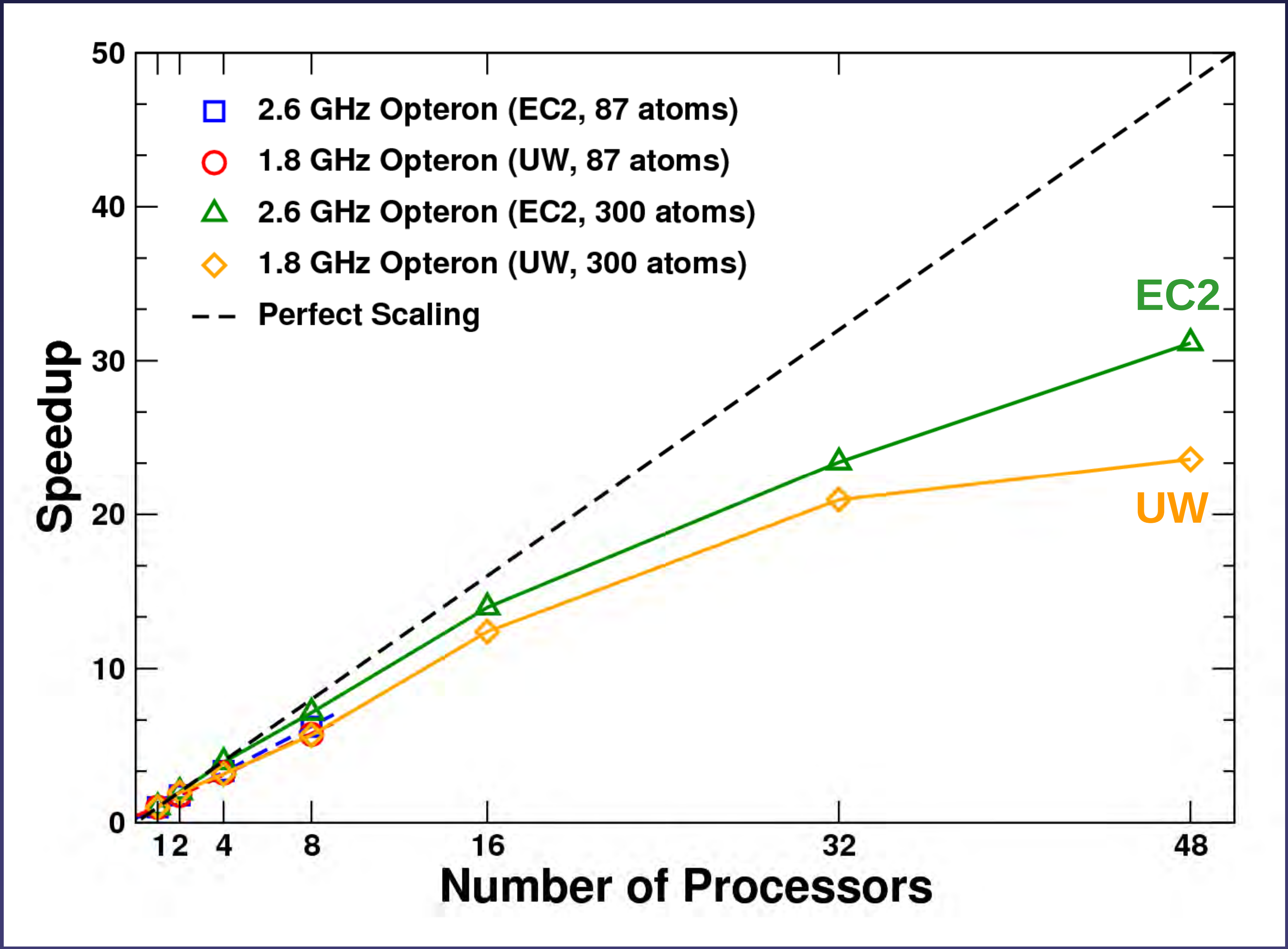
Name	Function	Analog
`ec2_clust_launch N`	Launches cluster with N instances	boot
`ec2_clust_connect`	Connect to a cluster	ssh
`ec2_clust_put`	Transfer data to EC2 cluster	scp
`ec2_clust_get`	Transfer data from EC2 cluster	scp
`ec2_clust_list`	List running clusters	
`ec2_clust_terminate`	Terminate a running cluster	shutdown

The tools hide a lot of the "ugliness":

ec2_clust_connect

ssh -i /home/fer/.ec2_clust/.ec2_clust_info.7729.r-de70cdb7/key_pair _fdv.pem root@ec2-72-44-53-27.compute-1.amazonaws.com

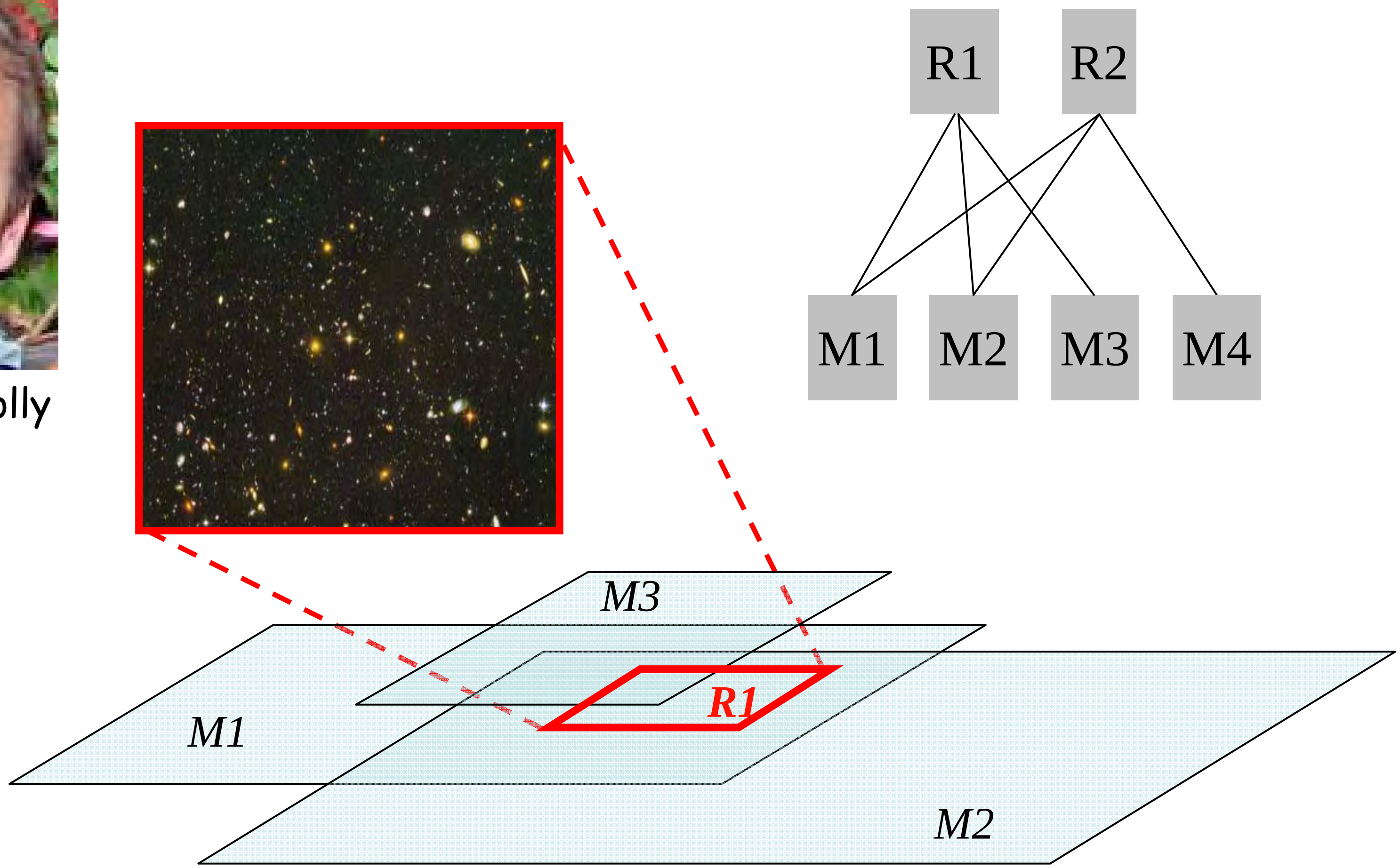
FEFFMPI on EC2



EC2 works well for highly parallelized applications like FEFF

SkyScraper: Scalable Image Registration and Query in the Cloud with MapReduce

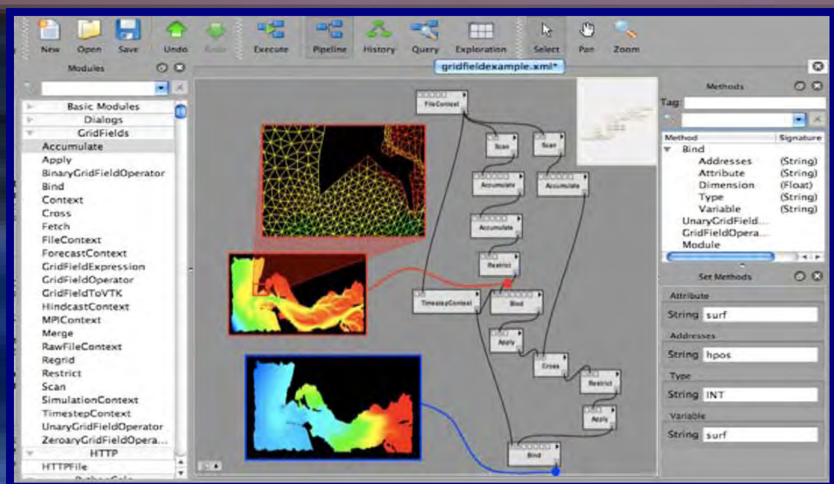
Andy Connolly



Horizon: Where the Ocean meets the Cloud

- Need interactive "climatologies": Decade-scale averages under different assumptions
- Must manipulate 40 terabytes the same way you manipulate 40 megabytes: efficiently, interactively, visually

- Client + Cloud: VisTrails, GridFields, 400-node Hadoop Cluster (NSF CluE program)




Bill Howe


Claudio Silva


Juliana Freire

NSF Google IBM CMOP Center for Coastal Margin Observation & Prediction

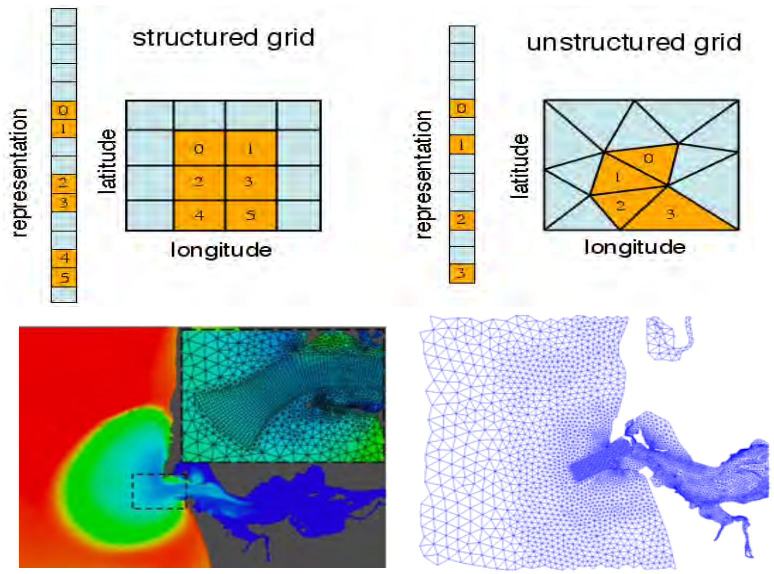
http://clue.cs.washington.edu/

"EC2 is Google Docs for developers"

- The cloud is the ultimate collaborative development environment
  - A shared environment outside of the jurisdiction of over-protective (or otherwise non-responsive) sysadmins
  - No bugs closed as "can't replicate"
- Example: New software for serving oceanographic model results, requiring collaboration between UW, OPeNDAP.org, and OOI

Bill Howe

- Waited two weeks for credentials to be established
- Gave up, spun up an EC2 instance, were rolling within an hour



- Similarly, Seattle's Institute for Systems Biology uses EC2/S3 for sharing computational pipelines

Observations

- Flat files and Excel spreadsheets are the most common data management tools for scientists
  - Data management workflows are choking science
- Even superb scientists are doing things you wouldn't believe
  - Such as manual joins on huge spreadsheets, exemplified by Ginger Armbrust's environmental metagenomics lab
- Simple tools can change their lives
  - E.g., the spreadsheet->SQLShare and web SQL query interface for Armbrust's lab
- Many of these tools have broad applicability
  - E.g., the above, and the Condor-to-cloud interface designed for Arzeda

- Workflow management is hugely important; building on commercial workflow engines is the smart approach
  - Trident has been widely adopted
- Flexible client+cloud architectures are winners – there is no "one size fits all"
  - COVE + Trident + Azure, Horizon
- A huge proportion of interesting science is, or can be made, embarrassingly parallel – many "HPC" researchers can thrive in the cloud
  - Tom Daniel's Monte Carlo muscle simulations
  - John Rehr's FEFF and Gasoline

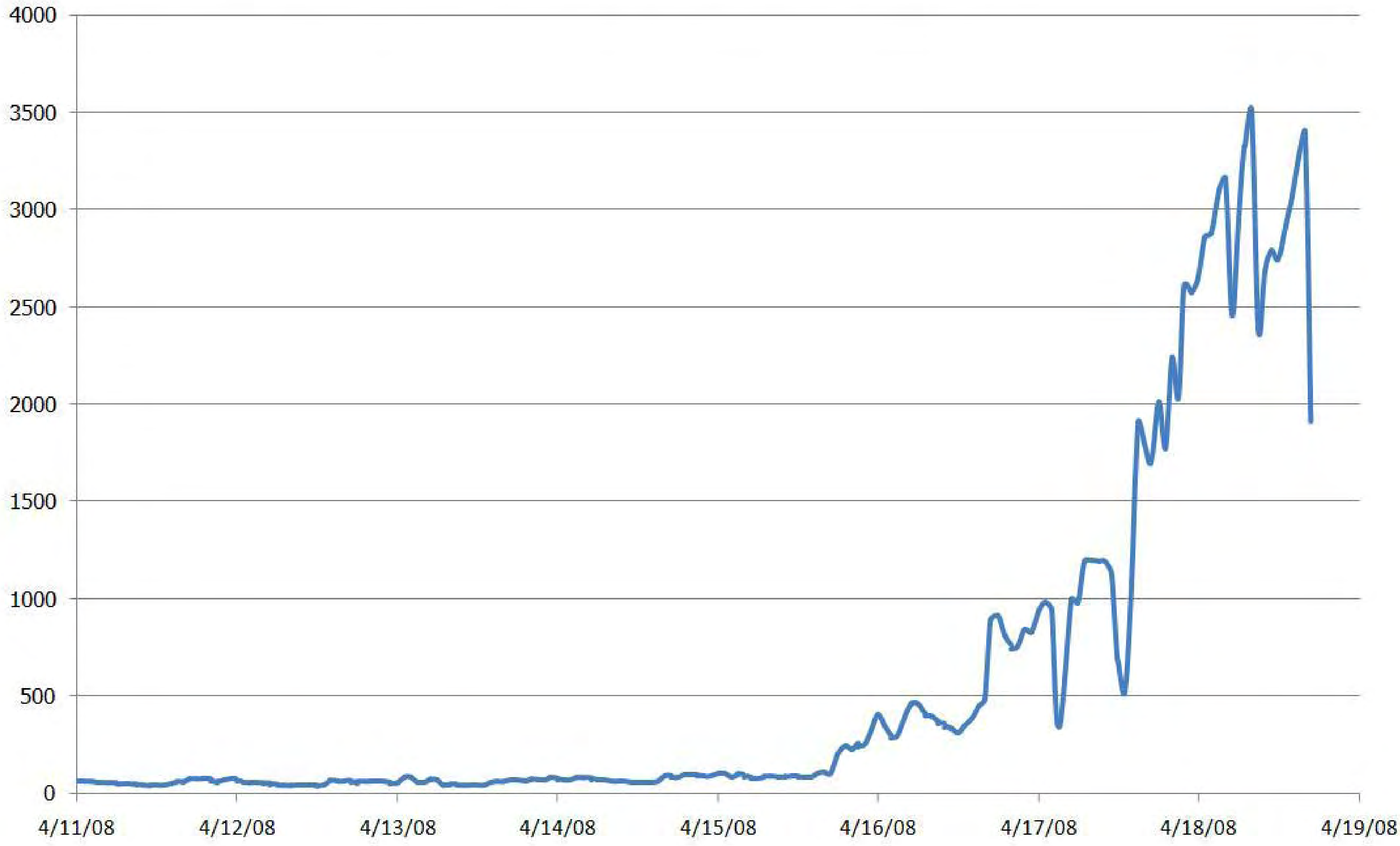
- Many science apps lend themselves to MapReduce / Dryad – style computation
  - Andy Connolly's SkyScraper
  - Bill Howe's Horizon
- "EC2 is Google Docs for developers"
  - UW / OPeNDAP.org / OOI
  - Institute for Systems Biology

eScience Institute

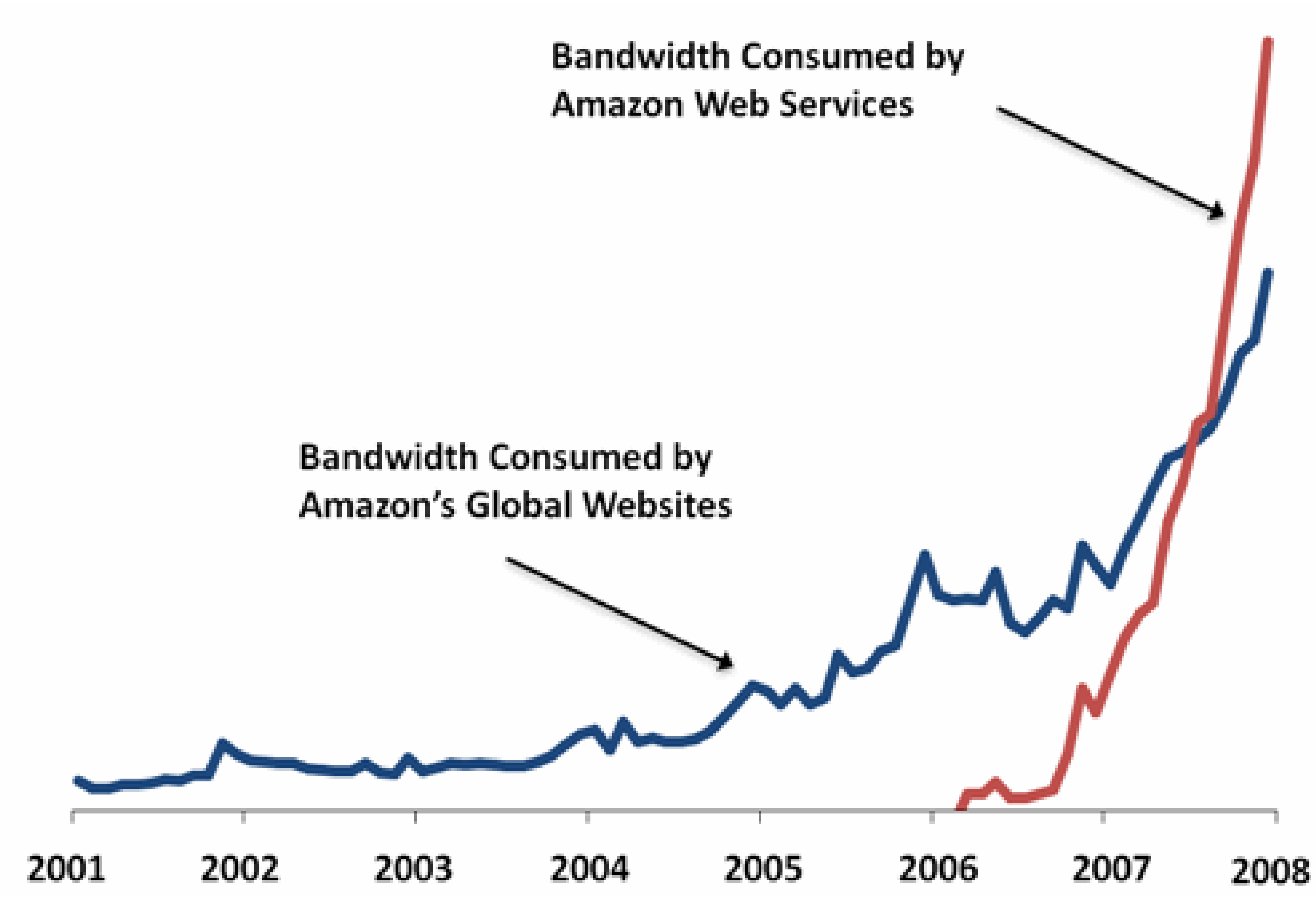
Gordon and Betty MOORE FOUNDATION

Microsoft®

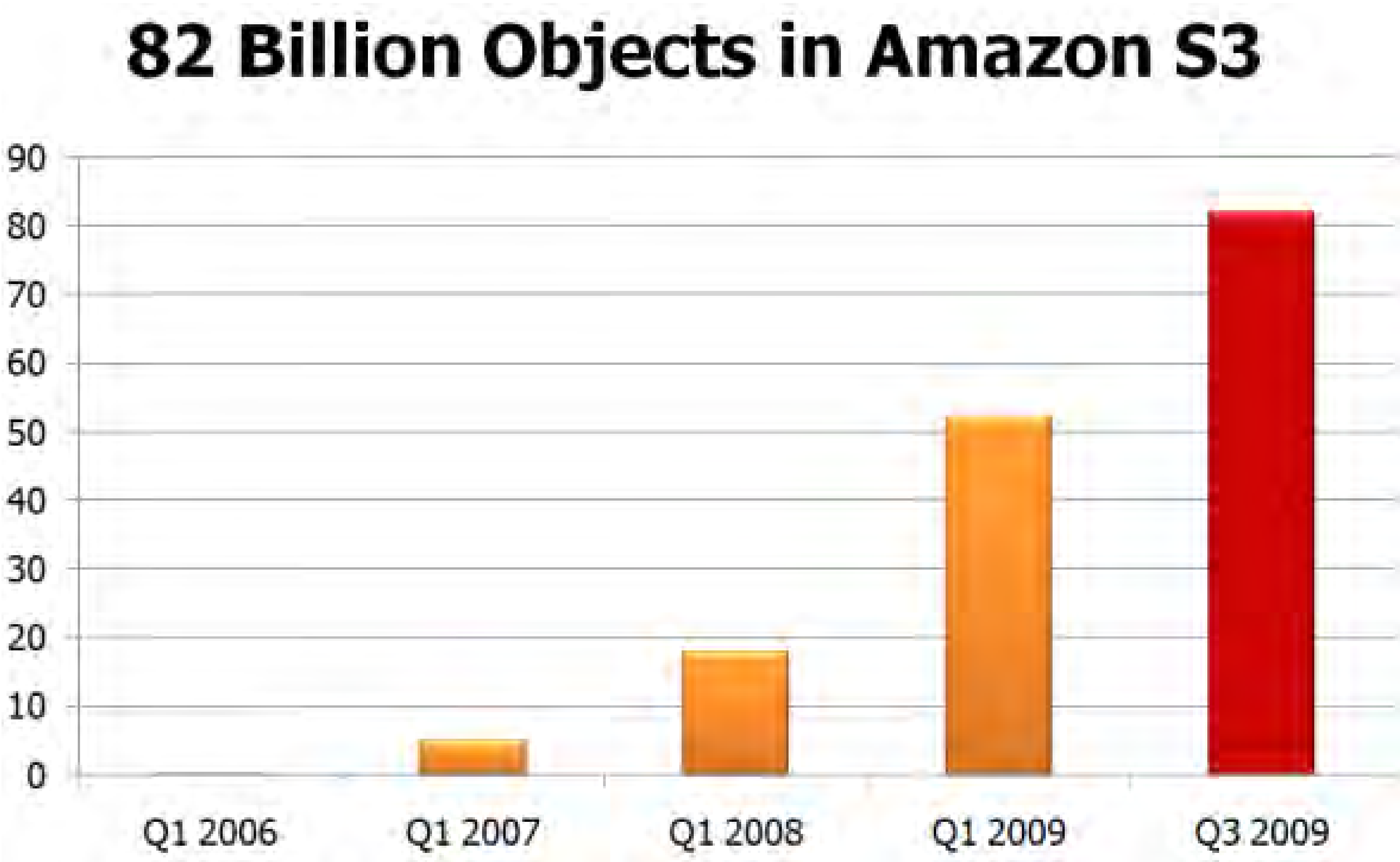
Animoto: EC2 Instance Usage



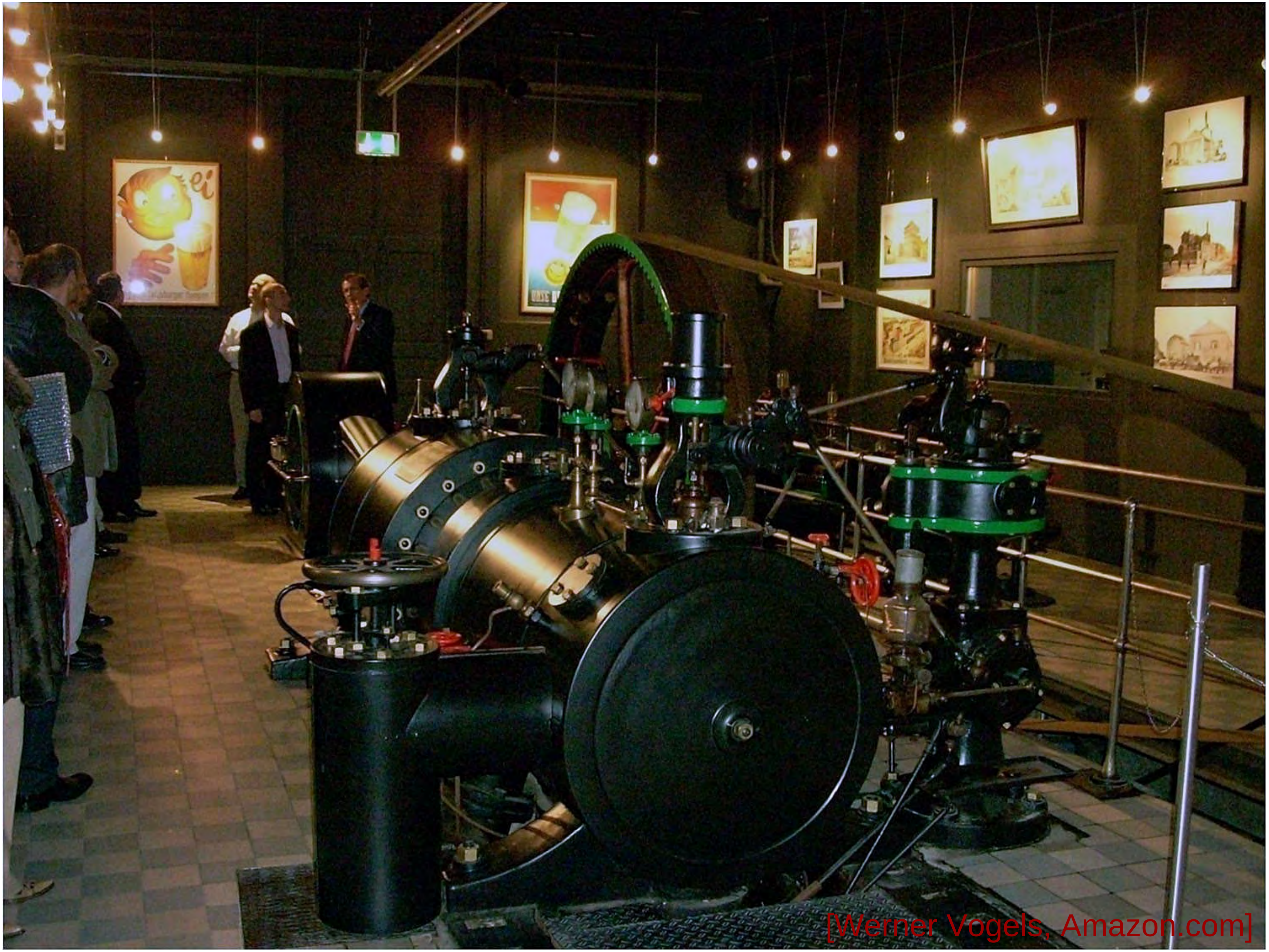
[Werner Vogels, Amazon.com]



[Werner Vogels, Amazon.com]



[Werner Vogels, Amazon.com]

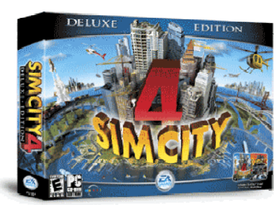


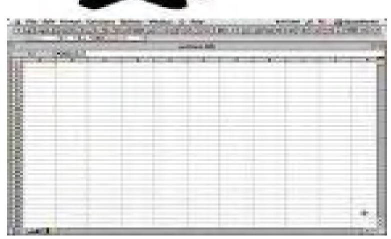
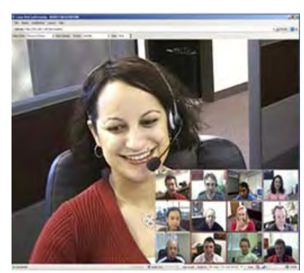
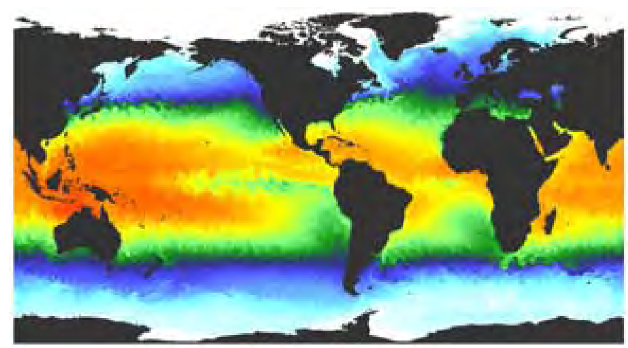
[Werner Vogels, Amazon.com]

Computer science: Changing the world

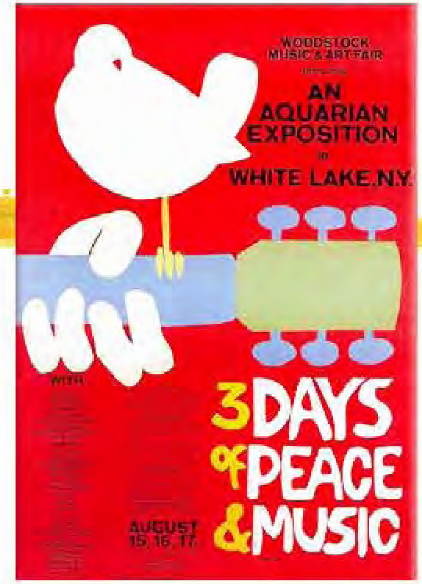
- Advances in computing change the way we live, work, learn, and communicate
- Advances in computing drive advances in nearly all other fields
- Advances in computing power our economy
  - Not just through the growth of the IT industry – through productivity growth across the entire economy



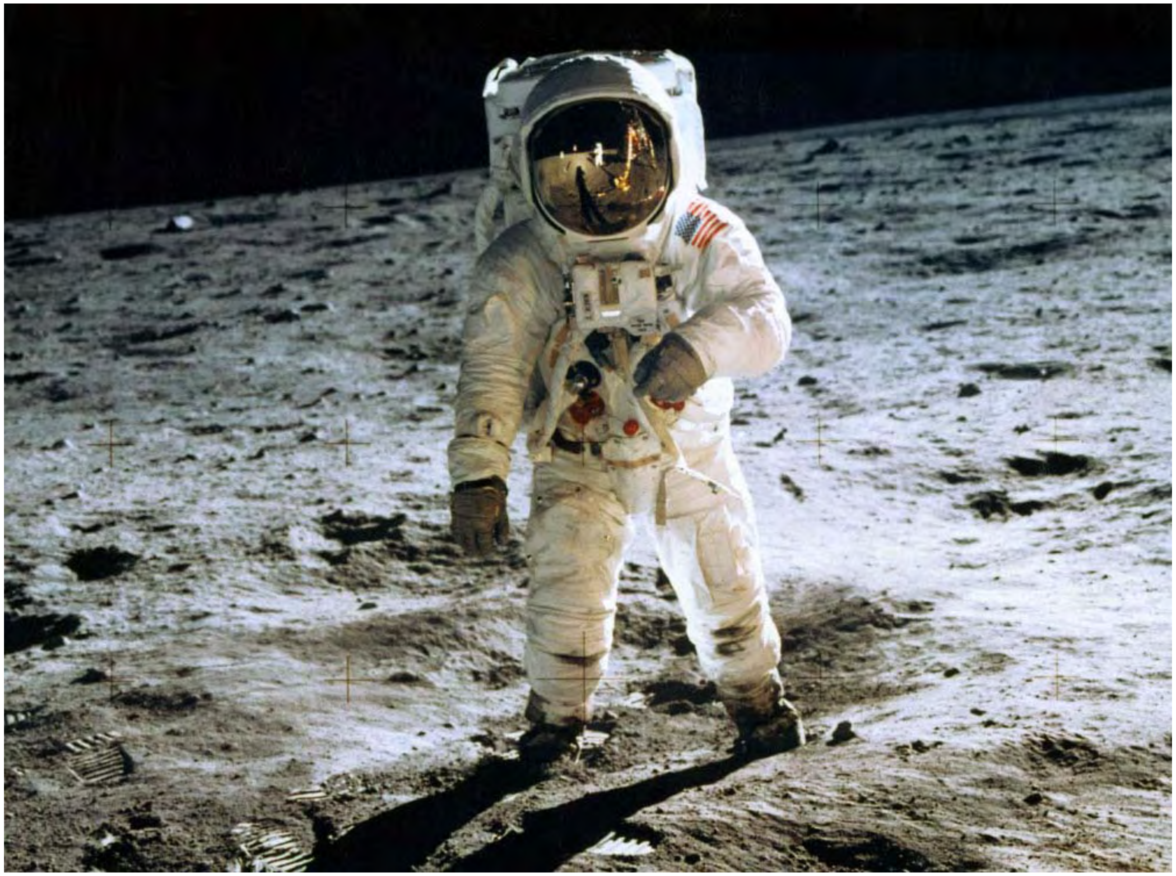


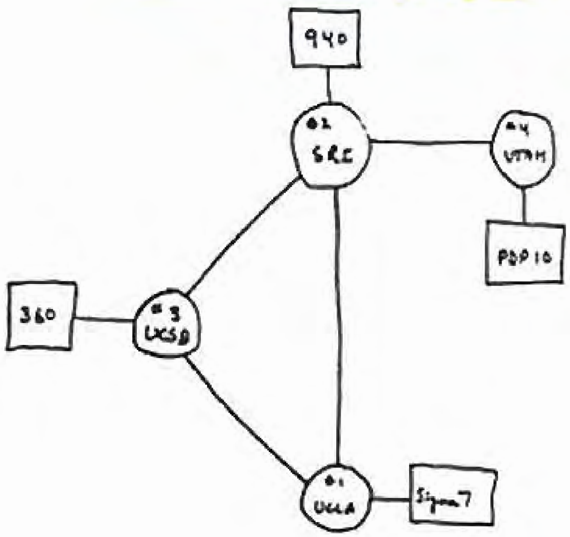
Forty years ago …


WOODSTOCK
MUSIC & ART FAIR
AN
AQUARIAN
EXPOSITION
WHITE LAKE, N.Y.
WITH
3 DAYS
of PEACE
& MUSIC
AUGUST
15, 16, 17.

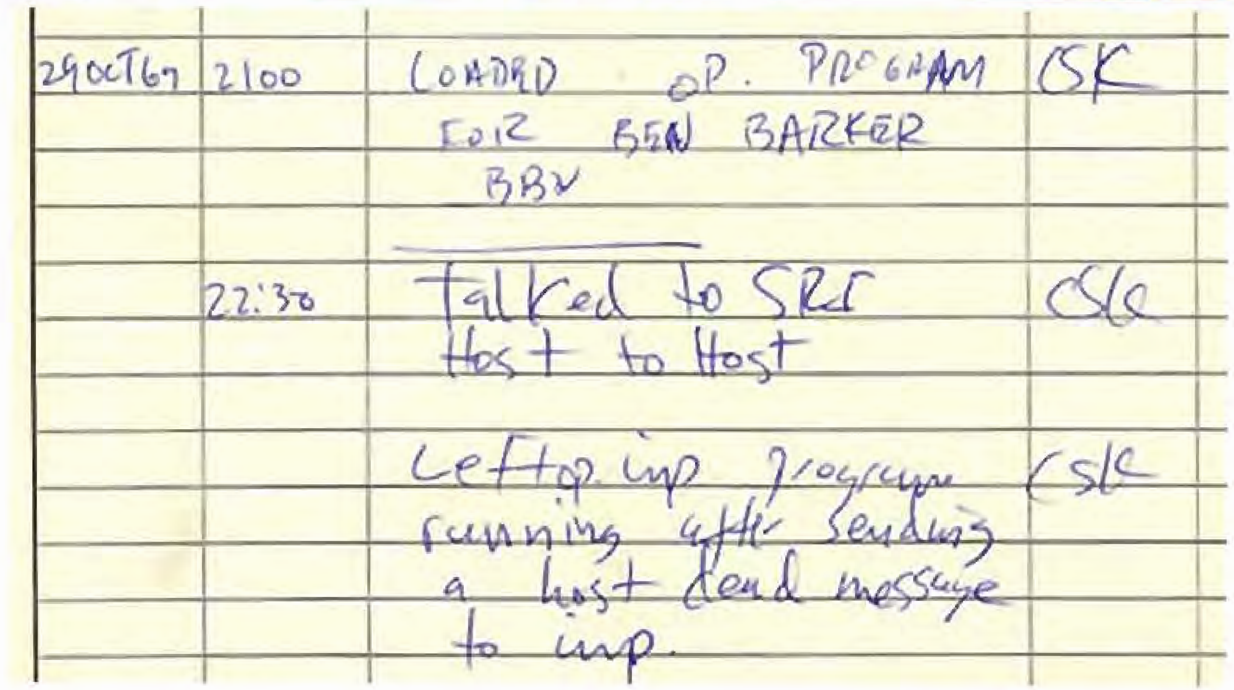

1969



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



THE ARPA NETWORK
DEC 1969
4 NODES



With forty years hindsight, which had the greatest impact?

- Unless you're big into Tang and Velcro (or sex and drugs), the answer is clear …

- And so is the reason …

EXPONENTIALS Я US

The past thirty years …

THE COUNT

Internet, Mobile Phones Named Most Important Inventions

By PHYLLIS KORKKI
Published: March 7, 2009

In response to the shouted-out question, "What are some of the greatest inventions of all time?," nearby office workers in a recent informal survey gave the following answers: the wheel, the engine, the ballpoint pen, diapers and the cheese Danish.

A panel of eight judges from the Wharton School of the University of Pennsylvania was required to go back only 30 years — not to the dawn of history — when asked a similar question. So its answers, of course, were very different.

In the survey, the Internet was voted the biggest innovation of the last three decades, followed by computers, mobile phones and e-mail. The survey was sponsored by Knowledge@Wharton, the school's business publication, and PBS's "Nightly Business Report."

Good, important choices all, but for classic, long-lasting appeal, they still can't beat the wheel. **PHYLLIS KORKKI**

Life Changers

The top innovations of the last 30 years, according to judges at the Wharton School of the University of Pennsylvania.

1. Internet, broadband
2. PC and laptop computers
3. Mobile phones
4. E-mail
5. DNA testing and sequencing
6. Magnetic resonance imaging
7. Microprocessors
8. Fiber optics
9. Office software
10. Laser/robotic surgery
11. Open-source software
12. Light-emitting diodes
13. Liquid crystal display
14. GPS devices
15. E-commerce and auctions
16. Media file compression
17. Microfinance
18. Photovoltaic solar energy
19. Large-scale wind turbines
20. Internet social networking

THE NEW YORK TIMES

Next Article in Business (22 of 29) »

Next Article in Business (22 of 29) »

Business

Life Changers

The top innovations of the last 30 years, according to judges at the Wharton School of the University of Pennsylvania.

...amed Most Important Inventions

In response to the shouted-out question, "What are some of the greatest inventions of all time?," nearby office workers in a recent informal survey gave the following answers: the wheel, the engine, the ballpoint pen, diapers and the cheese Danish.

Life Changers

The top innovations of the last 30 years, according to judges at the Wharton School of the University of Pennsylvania.

1. Internet, broadband
2. PC and laptop computers
3. Mobile phones
4. E-mail
5. DNA testing and sequencing
6. Magnetic resonance imaging
7. Microprocessors
8. Fiber optics
9. Office software
10. Laser/robotic surgery
11. Open-source software
12. Light-emitting diodes
13. Liquid crystal display
14. GPS devices
15. E-commerce and auctions
16. Media file compression
17. Microfinance
18. Photovoltaic solar energy
19. Large-scale wind turbines
20. Internet social networking

THE NEW YORK TIMES

A panel of eight judges from the Wharton School of the University of Pennsylvania was required to go back only 30 years — not to the dawn of history — when asked a similar question. So its answers, of course, were very different.

In the survey, the Internet was voted the biggest innovation of the last three decades, followed by computers, mobile phones and e-mail. The survey was sponsored by Knowledge@Wharton, the school's business publication, and PBS's "Nightly Business Report."

Good, important choices all, but for classic, long-lasting appeal, they still can't beat the wheel. PHYLLIS KORKKI

Next Article in Business (22 of 29) »

Next Article in Business (22 of 29) »

Life Changers

The top innovations of the last 30 years, according to judges at the Wharton School of the University of Pennsylvania.

1. Internet, broadband
2. PC and laptop computers
3. Mobile phones
4. E-mail
5. DNA testing and sequencing
6. Magnetic resonance imaging
7. Microprocessors
8. Fiber optics
9. Office software
10. Laser/robotic surgery
11. Open-source software
12. Light-emitting diodes
13. Liquid crystal display
14. GPS devices
15. E-commerce and auctions
16. Media file compression
17. Microfinance
18. Photovoltaic solar energy
19. Large-scale wind turbines
20. Internet social networking

THE NEW YORK TIMES

THE COUNT

Internet, Mobile Phones Named Most Important Inventions

In response to the shouted-out question, "What are some of the greatest inventions of all time?," nearby office workers in a recent informal survey gave the following answers: the wheel, the engine, the ballpoint pen, diapers and the cheese Danish.

A panel of eight judges from the Wharton School of the University of Pennsylvania was required to go back only 30 years — not to the dawn of history — when asked a similar question. So its answers, of course, were very different.

In the survey, the Internet was voted the biggest innovation of the last three decades, followed by computers, mobile phones and e-mail. The survey was sponsored by Knowledge@Wharton, the school's business publication, and PBS's "Nightly Business Report."

Good, important choices all, but for classic, long-lasting appeal, they still can't beat the wheel. PHYLLIS KORKKI

Life Changers
The top innovations of the last 30 years, according to judges at the Wharton School of the University of Pennsylvania.
1. Internet, broadband
2. PC and laptop computers
3. Mobile phones
4. E-mail
5. DNA testing and sequencing
6. Magnetic resonance imaging
7. Microprocessors
8. Fiber optics
9. Office software
10. Laser/robotic surgery
11. Open-source software
12. Light-emitting diodes
13. Liquid crystal display
14. GPS devices
15. E-commerce and auctions
16. Media file compression
17. Microfinance
18. Photovoltaic solar energy
19. Large-scale wind turbines
20. Internet social networking

E-MAIL | PRINT | REPRINTS | SHARE

Next Article in Business (22 of 29) »

News for Education Professionals
FROM NYTIMES.COM

- Colleges Sweat Out Admissions This Year
- Schumer Says Schools and State Will Get Some Stimulus Money This Month
- Districts Pursue School-Closing Plans to Save Money
- Parents Sue Trustees Over Prep School's Shutdown
- Doctoral Candidates Anticipate Hard Times

Next Article in Business (22 of 29) »

Life Changers

The top innovations of the last 30 years, according to judges at the Wharton School of the University of Pennsylvania.

- Internet, broadband
- PC and laptop computers
- Mobile phones
- E-mail
- DNA testing and sequencing
- Magnetic resonance imaging
- Microprocessors
- 8. Fiber optics
- Office software
- Laser/robotic surgery
- Open-source software
- 12. Light-emitting diodes
- 13. Liquid crystal display
- GPS devices
- E-commerce and auctions
- Media file compression
- 17. Microfinance
- 18. Photovoltaic solar energy
- 19. Large-scale wind turbines
- Internet social networking

THE NEW YORK TIMES

THE COUNT

Internet, Mobile Phones Named Most Important Inventions

By PHYLLIS KORKKI
Published: March 7, 2009

In response to the shouted-out question, "What are some of the greatest inventions of all time?," nearby office workers in a recent informal survey gave the following answers: the wheel, the engine, the ballpoint pen, diapers and the cheese Danish.

Life Changers

The top innovations of the last 30 years, according to judges at the Wharton School of the University of Pennsylvania.

1. Internet, broadband
2. PC and laptop computers
3. Mobile phones
4. E-mail
5. DNA testing and sequencing
6. Magnetic resonance imaging
7. Microprocessors
8. Fiber optics
9. Office software
10. Laser/robotic surgery
11. Open-source software
12. Light-emitting diodes
13. Liquid crystal display
14. GPS devices
15. E-commerce and auctions
16. Media file compression
17. Microfinance
18. Photovoltaic solar energy
19. Large-scale wind turbines
20. Internet social networking

THE NEW YORK TIMES

A panel of eight judges from the Wharton School of the University of Pennsylvania was required to go back only 30 years — not to the dawn of history — when asked a similar question. So its answers, of course, were very different.

In the survey, the Internet was voted the biggest innovation of the last three decades, followed by computers, mobile phones and e-mail. The survey was sponsored by Knowledge@Wharton, the school's business publication, and PBS's "Nightly Business Report."

Good, important choices all, but for classic, long-lasting appeal, they still can't beat the wheel. PHYLLIS KORKKI

E-MAIL
PRINT
REPRINTS
SHARE

Next Article in Business (22 of 29) »

News for Education Professionals What's This?

FROM NYTIMES.COM

- Colleges Sweat Out Admissions This Year
- Schumer Says Schools and State Will Get Some Stimulus Money This Month
- Districts Pursue School-Closing Plans to Save Money
- Parents Sue Trustees Over Prep School's Shutdown
- Doctoral Candidates Anticipate Hard Times

LinkedIn

Next Article in Business (22 of 29) »

The most recent ten years …

- Search
- Scalability
- Digital media
- Mobility
- eCommerce
- The Cloud
- Social networking and crowd-sourcing

The cloud: A triumph of computing research

- Enormous volumes of data
- Extreme parallelism
- The cheapest imaginable components
  - Failures occur all the time
  - You couldn't afford to prevent this in hardware
- Software makes it
  - Fault-Tolerant
  - Highly Available
  - Recoverable
  - Consistent
  - Scalable
  - Predictable
  - Secure

AlphaServer 1200 product brief

Leadership

"To support our rapid growth, we had to find a highly upgradable and scaleable Internet server. The AlphaServer platform provides the upgrade path we need."

Jeff Bezos
CEO and Founder
Amazon.com

NATIONAL ACADEMY OF ENGINEERING
OF THE NATIONAL ACADEMIES

CHALLENGES | IDEAS | NEXT STEPS | COMMITTEE

GRAND CHALLENGES FOR ENGINEERING


Make solar energy economical


Provide energy from fusion

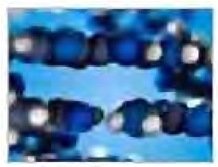

Develop carbon sequestration methods


Manage the nitrogen cycle


Provide access to clean water


Restore and improve urban infrastructure


Advance health informatics


Engineer better medicines


Reverse-engineer the brain


Prevent nuclear terror


Secure cyberspace


Enhance virtual reality


Advance personalized learning


Engineer the tools of scientific discovery


NATIONAL ACADEMY OF ENGINEERING
OF THE NATIONAL ACADEMIES
CHALLENGES
IDEAS
NEXT STEPS
COMMITTEE
GRAND CHALLENGES
FOR ENGINEERING
Make solar energy economical
Provide energy from fusion
Develop carbon sequestration methods
Manage the nitrogen cycle
Provide access to clean water
Restore and improve urban infrastructure
Advance health informatics
Engineer better medicines
Reverse-engineer the brain
Prevent nuclear terror
Secure cyberspace
Enhance virtual reality
Advance personalized learning
Engineer the tools of scientific discovery
Predominant CS component
Significant CS component

We put the "smarts" in …

- Smart homes
- Smart cars
- Smart bodies
- Smart robots
- The data deluge (smart science)
- Virtual and augmented reality
- Smart crowds and human-computer systems

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