Virtual Appliances, Cloud Computing, and Reproducible Research

Bill Howe, Phd
eScience Institute, UW

eScience Institute

http://escience.washington.edu

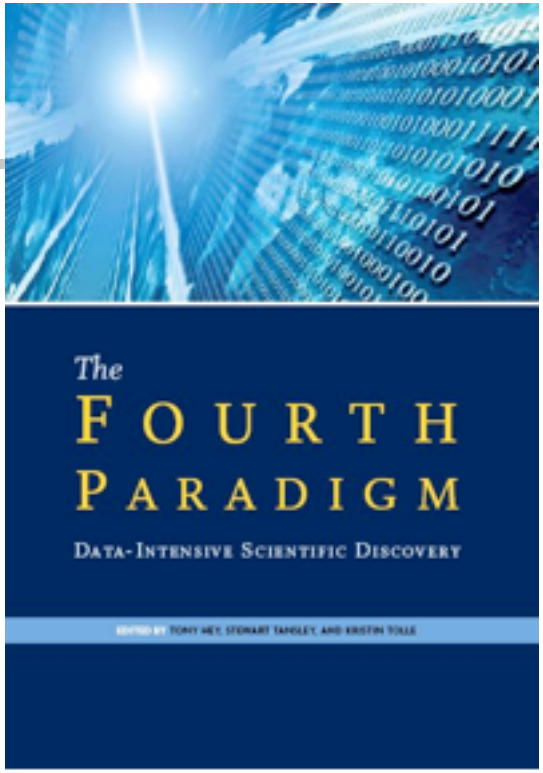
An Observation

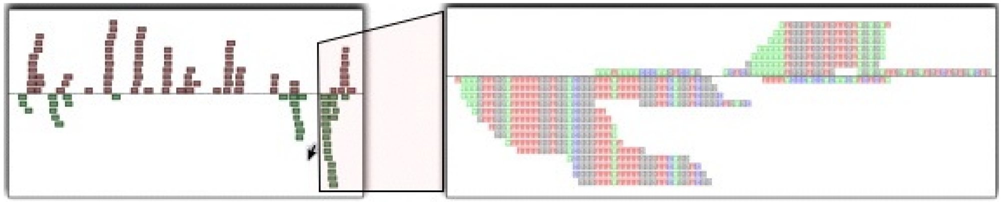
- There will always be experiments data housed outside of a managed environments
  - “Free” experimentation is a beautiful property of software
  - We should be conservative about constraining the process
- There is no difference between debugging, testing, and experiments.
  - When it works, it's an experiment.
  - When it doesn't, it's debugging.
- Conclusion: We need post hoc approaches
  - that can tolerate messy, heterogeneous code and data

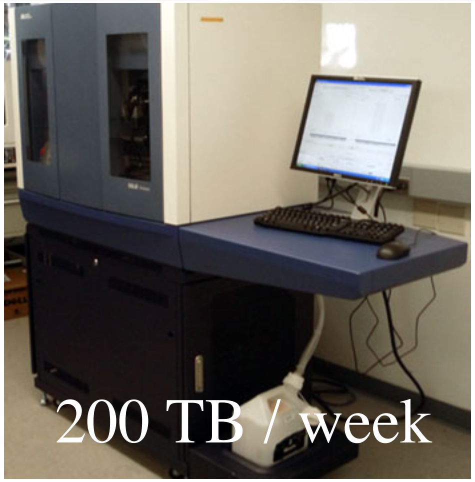
eScience is about data

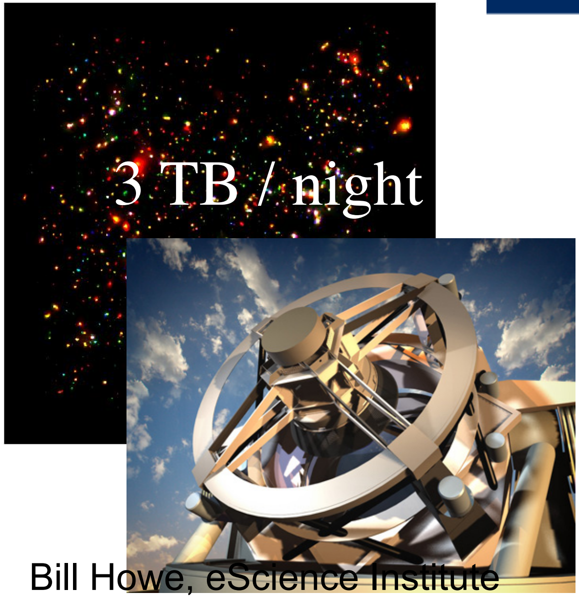
“Fourth Paradigm”

- Theory, Experiment, Computational Science
- Data-driven discovery


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE




200 TB / week

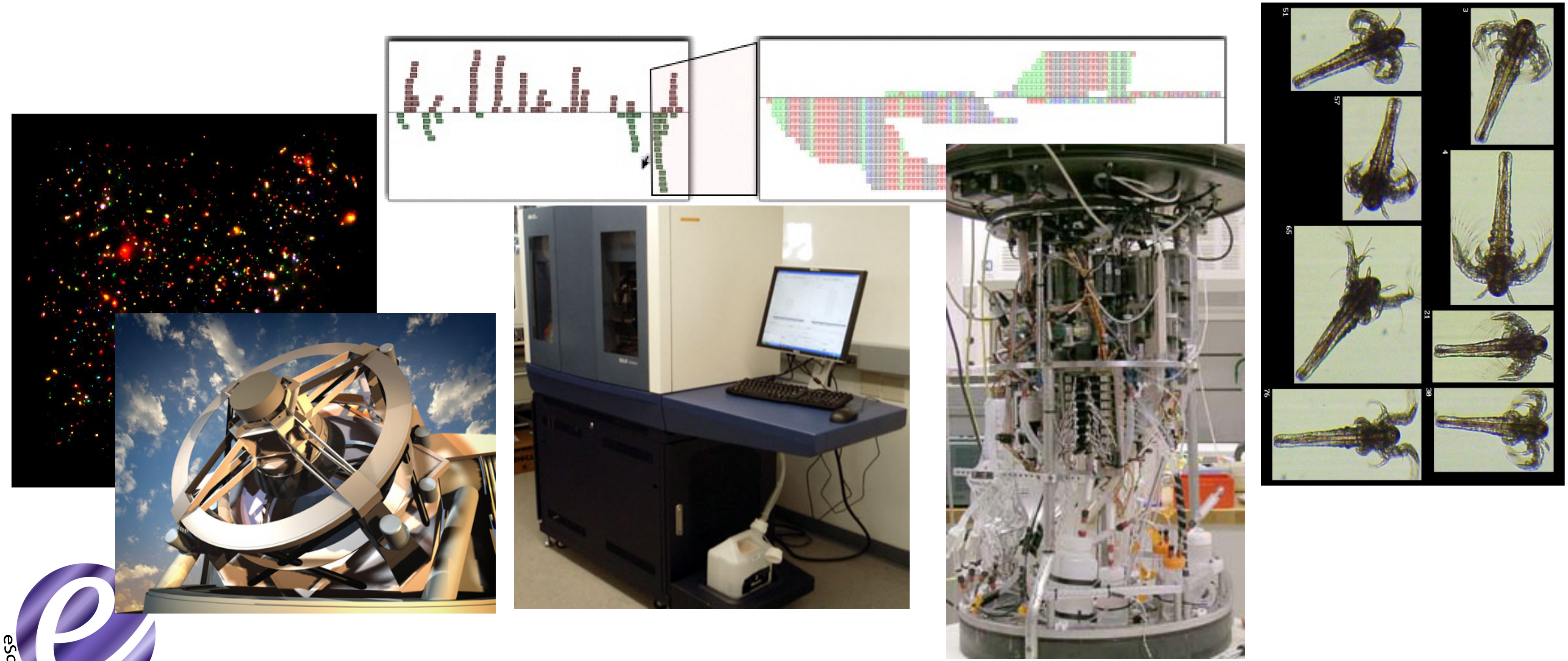

3 TB / night

eScience is about data

Old model: "Query the world" (Data acquisition coupled to a specific hypothesis)

New model: "Download the world, query the DB" (Data acquired en masse, to support many hypotheses)

- Astronomy: High-resolution, high-frequency sky surveys (SDSS, LSST, PanSTARRS)
- Oceanography: high-resolution models, cheap sensors, satellites
- Biology: lab automation, high-throughput sequencing,



Some projects

🟧 visualization + cloud
🟦 scientific data integration
🟩 scalable query processing

Analytics and Visualization with Hadoop (with Juliana Freire)

- $380k ($190k), 2/2009 - 2/2011, NSF Cluster Exploratory 2009 (joint with University of Utah)

eScience and Data-intensive computing (lead: Lazowska)

- $750k, 10/2009 – 10/2011 Gordon and Betty Moore Foundation

Cloud prototypes for the Ocean Observatories Initiative

- $107k, 9/2009 - 12/2009, Subcontract from SDSC/Woods Hole, NSF OOI

Microsoft Research Jim Gray Seed Grant, 2008 and 2010

- $25k, $40k

3D Visualization in the Cloud

- $117k, 9/10 – 09/12, NSF EAGER through Computing in the Cloud (CiC)

Hybrid Query Language for a Graph Databases

- $150k, 9/10 - 9/12, PNNL XMT project

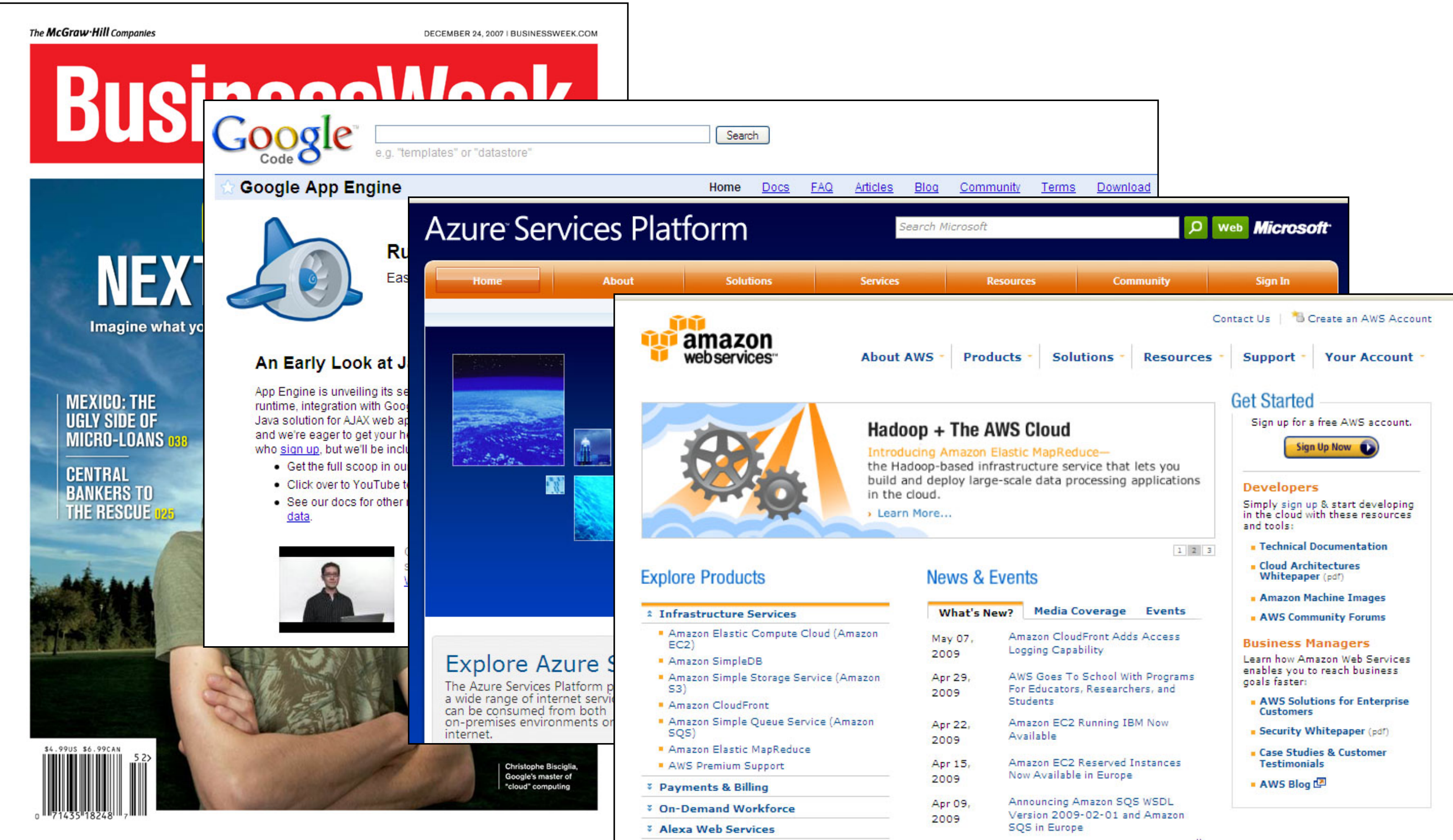
SQLShare: Database as a Service with Long-Tail Science

- $800k, 3 institutions, NSF

Data Markets (lead: Balazinska)

- ~$300k, 4/11 – 4/13, NSF Computing in the Cloud

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



MEXICO: THE UGLY SIDE OF MICRO-LOANS 038

CENTRAL BANKERS TO THE RESCUE 025

Christophe Bisciglia, Google's master of "cloud" computing

Google App Engine

Ru
Eas

An Early Look at J

App Engine is unveiling its se runtime, integration with Goo Java solution for AJAX web ap and we're eager to get your h who sign up, but we'll be inclu

- Get the full scoop in ou
- Click over to YouTube t
- See our docs for other data.

Explore Azure S

The Azure Services Platform p a wide range of internet servi can be consumed from both on-premises environments or internet.

Hadoop + The AWS Cloud

Introducing Amazon Elastic MapReduce—the Hadoop-based infrastructure service that lets you build and deploy large-scale data processing applications in the cloud.

› Learn More...

Explore Products

Infrastructure Services

- Amazon Elastic Compute Cloud (Amazon EC2)
- Amazon SimpleDB
- Amazon Simple Storage Service (Amazon S3)
- Amazon CloudFront
- Amazon Simple Queue Service (Amazon SQS)
- Amazon Elastic MapReduce
- AWS Premium Support

Payments & Billing

On-Demand Workforce

Alexa Web Services

News & Events

May 07, 2009	Amazon CloudFront Adds Access Logging Capability
Apr 29, 2009	AWS Goes To School With Programs For Educators, Researchers, and Students
Apr 22, 2009	Amazon EC2 Running IBM Now Available
Apr 15, 2009	Amazon EC2 Reserved Instances Now Available in Europe
Apr 09, 2009	Announcing Amazon SQS WSDL Version 2009-02-01 and Amazon SQS in Europe

Get Started

Sign up for a free AWS account.

Developers

Simply sign up & start developing in the cloud with these resources and tools:

- Technical Documentation
- Cloud Architectures Whitepaper (pdf)
- Amazon Machine Images
- AWS Community Forums

Business Managers

Learn how Amazon Web Services enables you to reach business goals faster:

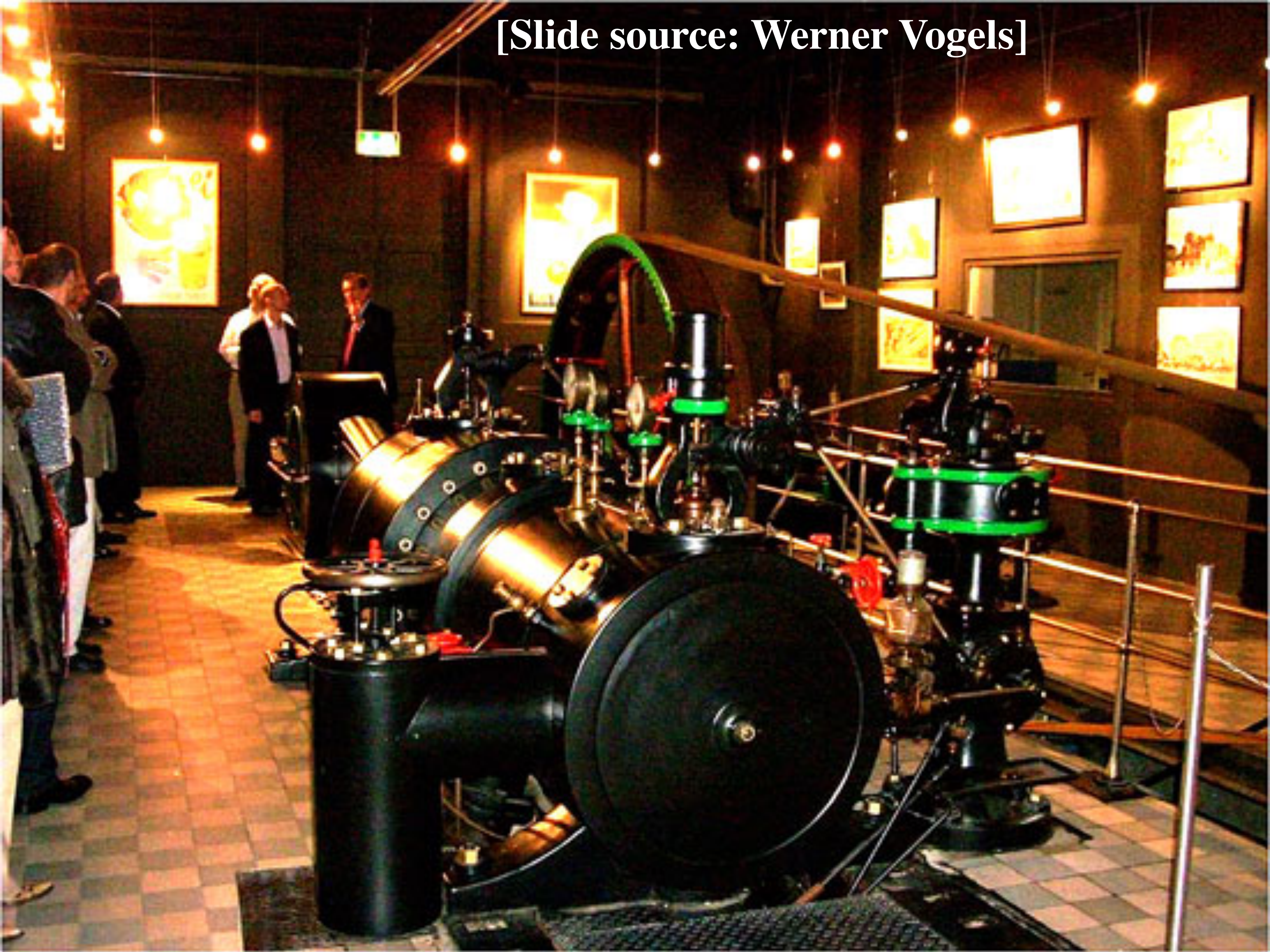
- AWS Solutions for Enterprise Customers
- Security Whitepaper (pdf)
- Case Studies & Customer Testimonials
- AWS Blog

The point of this talk

Explore the roles the cloud can play in reproducible research

“What if *everything* was in the cloud?”

CLOUD IN 2 SLIDES



[Slide source: Werner Vogels]

Growth

“Every day, Amazon buys enough computing resources to run the entire Amazon.com infrastructure as of 2001”

-- James Hamilton, Amazon, Inc., SIGMOD 2011 keynote

VIRTUALIZATION ANECDOTE

2007: The Ocean Appliance

Software

- Linux Fedora Core 6
- web server (Apache)
- database (PostgreSQL)
- ingest/QC system (Python)
- telemetry system (Python)
- web-based visualization (Drupal, Python)

Hardware

- 2.6GHz Dual
- 2GB RAM
- 250 GB SATA
- 4 serial ports
- ~$500
- ~1’x1’x1.5’

Responsibilities: Shipboard computing

- –Data Acquisition
- –Database Ingest
- –Telemetry with Shore
- –Visualization
- –App Server

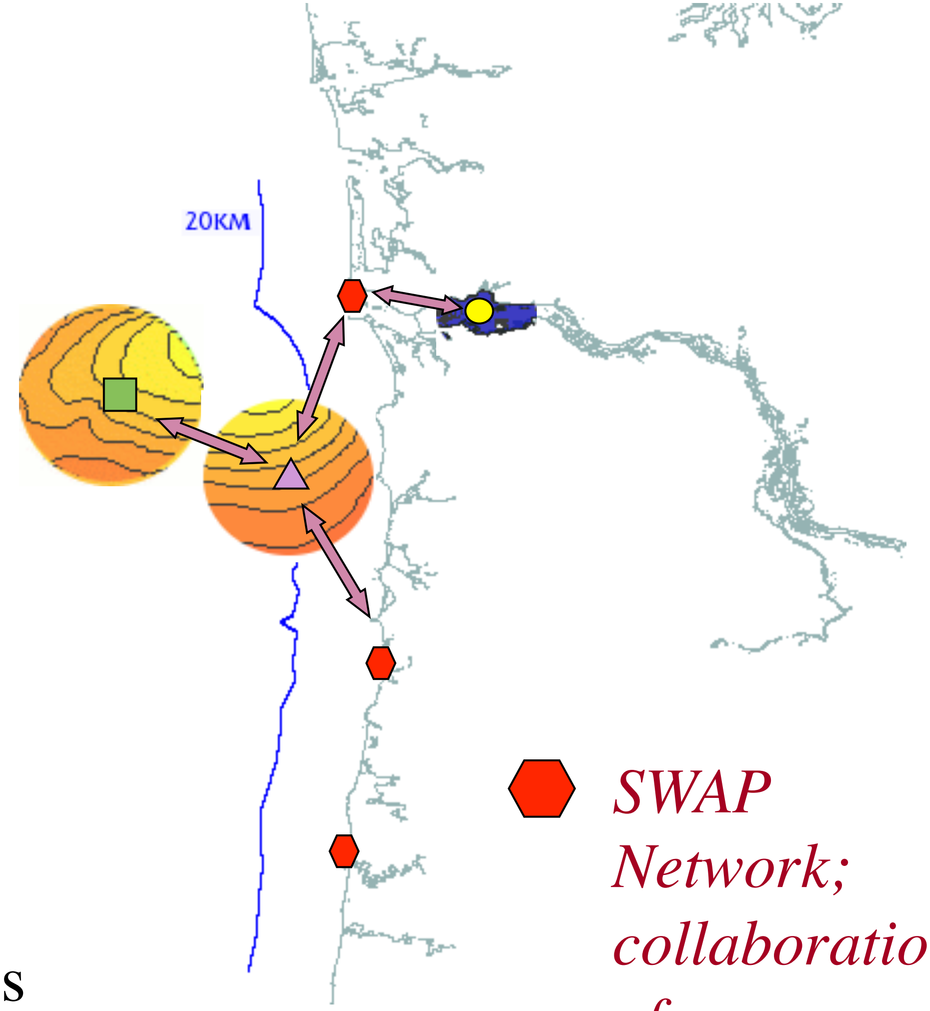
Deployment on R/V Barnes

Ship-to-Ship and Ship-to-Shore Telemetry

🟩 Wecoma

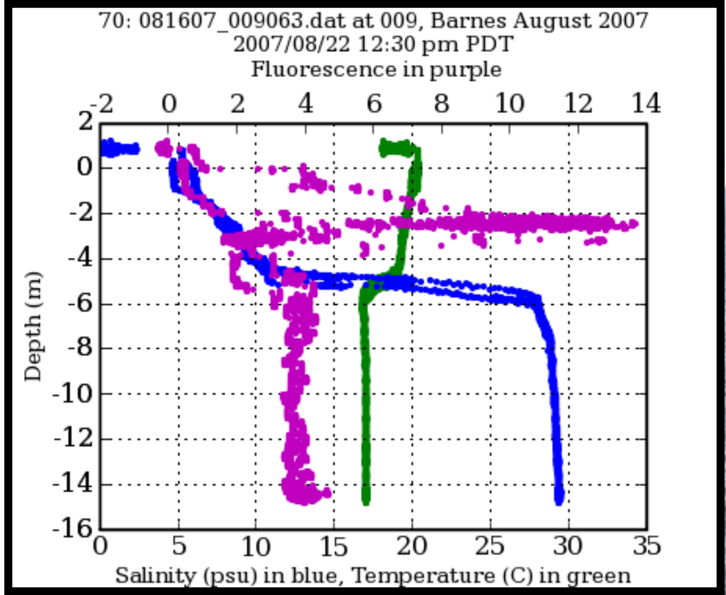
● Forerunner

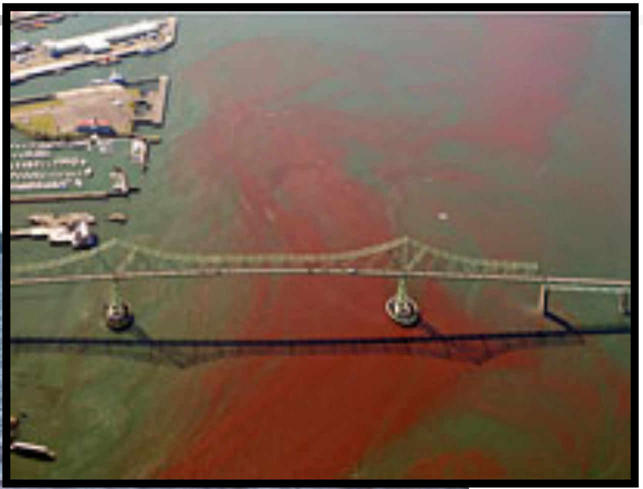
▲ Barnes

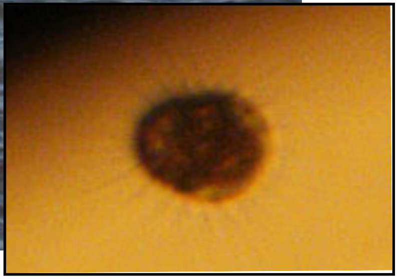


⬢ *SWAP Network; collaboration of:*
- OSU
- OHSU

Event Detection: Red Water







myrionecta rubra

Code + Data + Environment

- Easier, cheaper, and safer to build the box in the lab and hand it out for free than to work with the ships' admin to get our software running.

- Modern analog: Easier to build and distribute a virtual appliance than it is to support installation of your software.

Cloud + RR Overview

- Virtualization = Code + Data + Environment
  - Virtualization enables cross-platform, generalized, reliable ad hoc (and post hoc) environment capture

- Cloud = Virtualization + Resources + Services
  - any code, any data (more structure -> more services)
  - scalable storage and compute for everyone
  - services for processing big data, various data models
  - services for managing VMs
  - secure, reliable, available

Challenges

- Costs and cost-sharing
- Data-intensive science

- Offline discussion
  - Security / Privacy
  - Long-term Preservation
  - Cultural roadblocks

OBSERVATIONS ABOUT CLOUD, VIRTUALIZATION, RR

An Observation

- There will always be experiments data housed outside of a managed environments
  - “Free” experimentation is a beautiful property of software
  - We should be conservative about constraining the process
- There is no difference between debugging, testing, and experiments.
  - When it works, it's an experiment.
  - When it doesn't, it's debugging.
- Conclusion: We need post hoc approaches
  - that can tolerate messy, heterogeneous code and data

An Observation (2)

- Code + Data + Environment + Platform

- “Download it to my laptop” is insufficient
- Ex: de novo assembly
  - 64 GB RAM, 12 cores

- So we need more than VMs – we need a place to run them

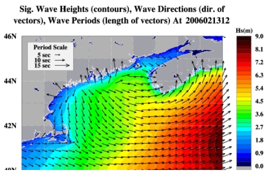
An Observation (3)

- Experiment environments span multiple machines
- Databases, models, web server

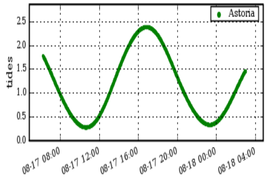
- 1 VM may not be enough

CMOP: Observation and Forecasting

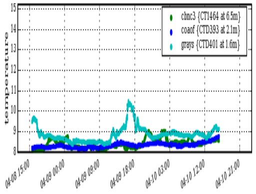
Atmospheric models


Sig. Wave Heights (contours), Wave Directions (dir. of vectors), Wave Periods (length of vectors) At 2006021312
Period Scale
5 sec
10 sec
15 sec
46N
44N
42N
Hs(m)
9.0
8.1
7.2
6.3
5.4
4.5
3.6
2.7
1.8
0.9
0.0

Tides

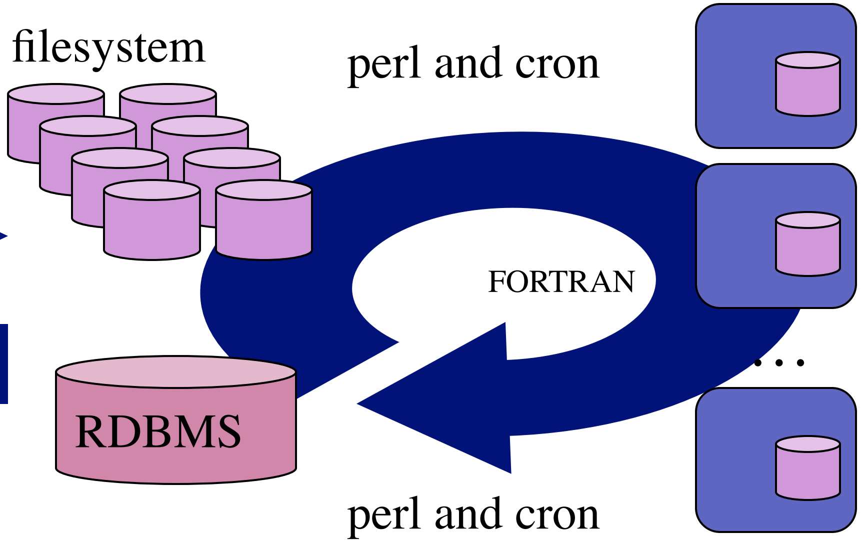


River discharge

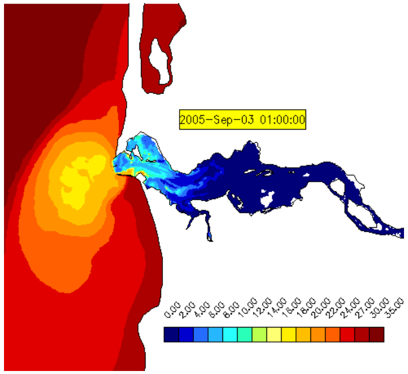


forcings (i.e., inputs)

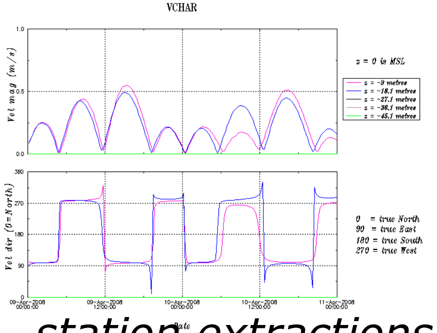
products via the web



cluster



salinity isolines



station extractions

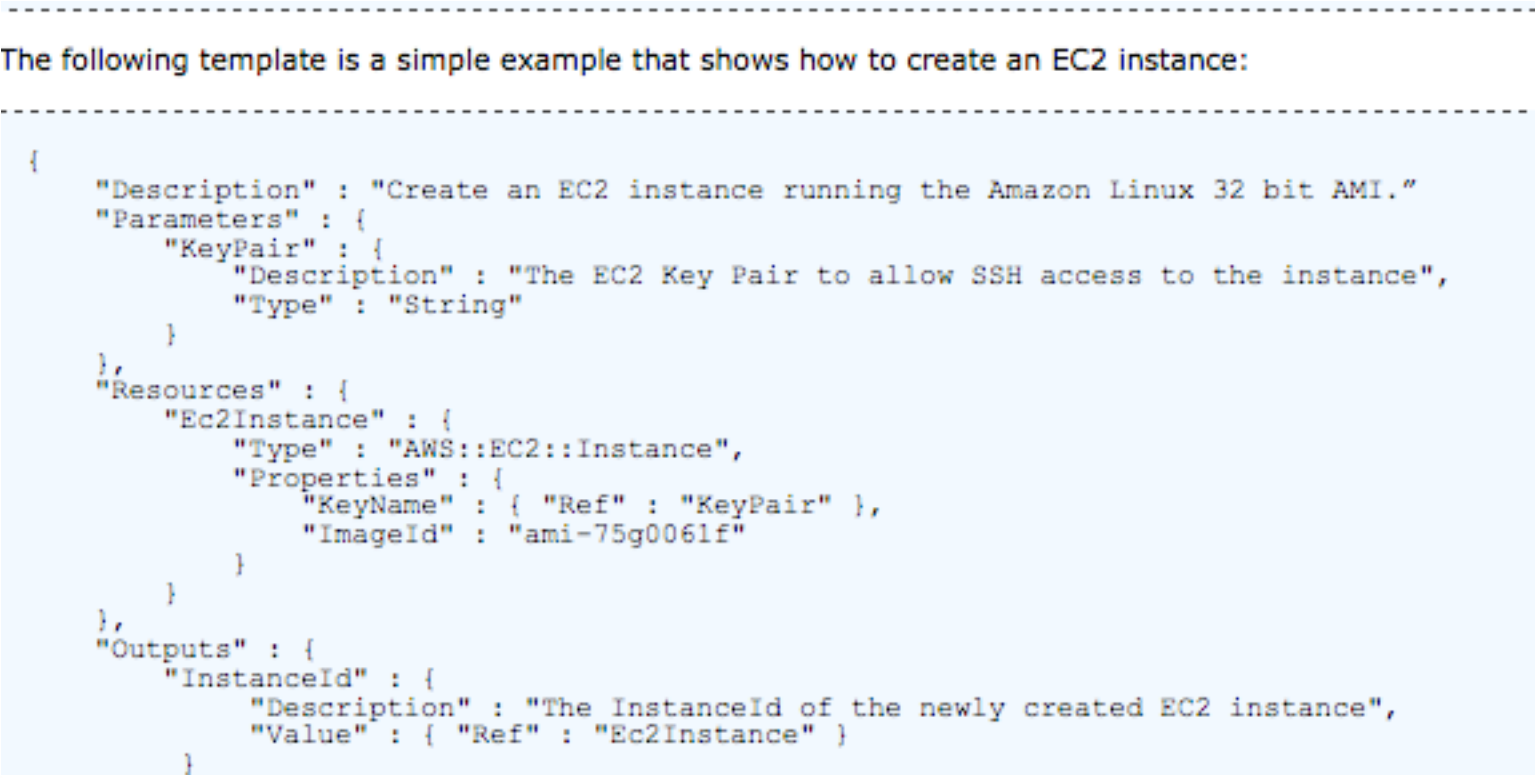
model-data comparisons

Simulation results
Config and log files
Intermediate files
Annotations
Data Products
Relations

CMOP
Center for Coastal Margin Observation & Prediction

Amazon CloudFormation

- Ensembles of Virtual Machines
- Launch and configure as a unit



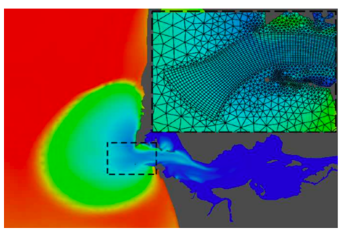
The following template is a simple example that shows how to create an EC2 instance:

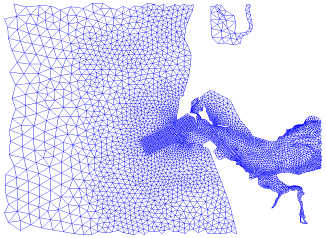
```
{
    "Description" : "Create an EC2 instance running the Amazon Linux 32 bit AMI."
    "Parameters" : {
        "KeyPair" : {
            "Description" : "The EC2 Key Pair to allow SSH access to the instance",
            "Type" : "String"
        }
    },
    "Resources" : {
        "Ec2Instance" : {
            "Type" : "AWS::EC2::Instance",
            "Properties" : {
                "KeyName" : { "Ref" : "KeyPair" },
                "ImageId" : "ami-75g0061f"
            }
        }
    },
    "Outputs" : {
        "InstanceId" : {
             "Description" : "The InstanceId of the newly created EC2 instance",
             "Value" : { "Ref" : "Ec2Instance" }
         }
```

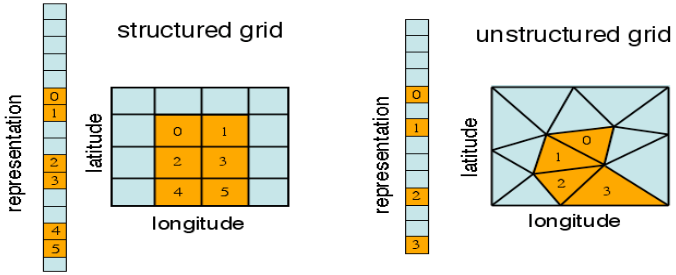
Observation (3): "Google Docs for developers"

- The cloud offers a "demilitarized zone" for temporary, low-overhead collaboration
  - A temporary, shared development environment outside of the jurisdiction of over-zealous sysadmins
  - No bugs closed as "can't replicate"
- Example: New software for serving oceanographic model results, requiring collaboration between UW, OPeNDAP.org, and OOI

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







Similarly, Seattle's Institute for Systems Biology uses EC2/S3 for collaborative development of computational pipelines

INSTITUTE FOR Systems Biology

COSTS AND COST-SHARING

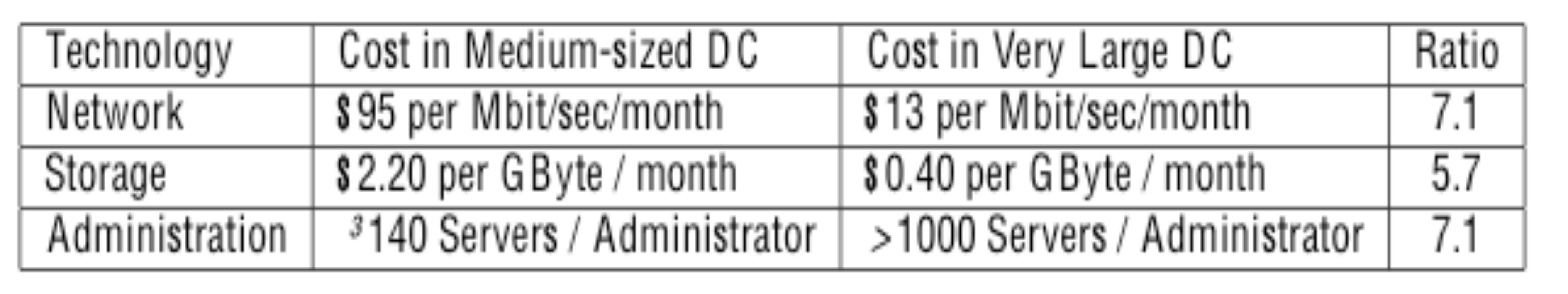
Who pays for reproducibility?

- Costs of hosting code?
- Costs of hosting data?
- Costs of executing code?

- *Answer: you, you, them*

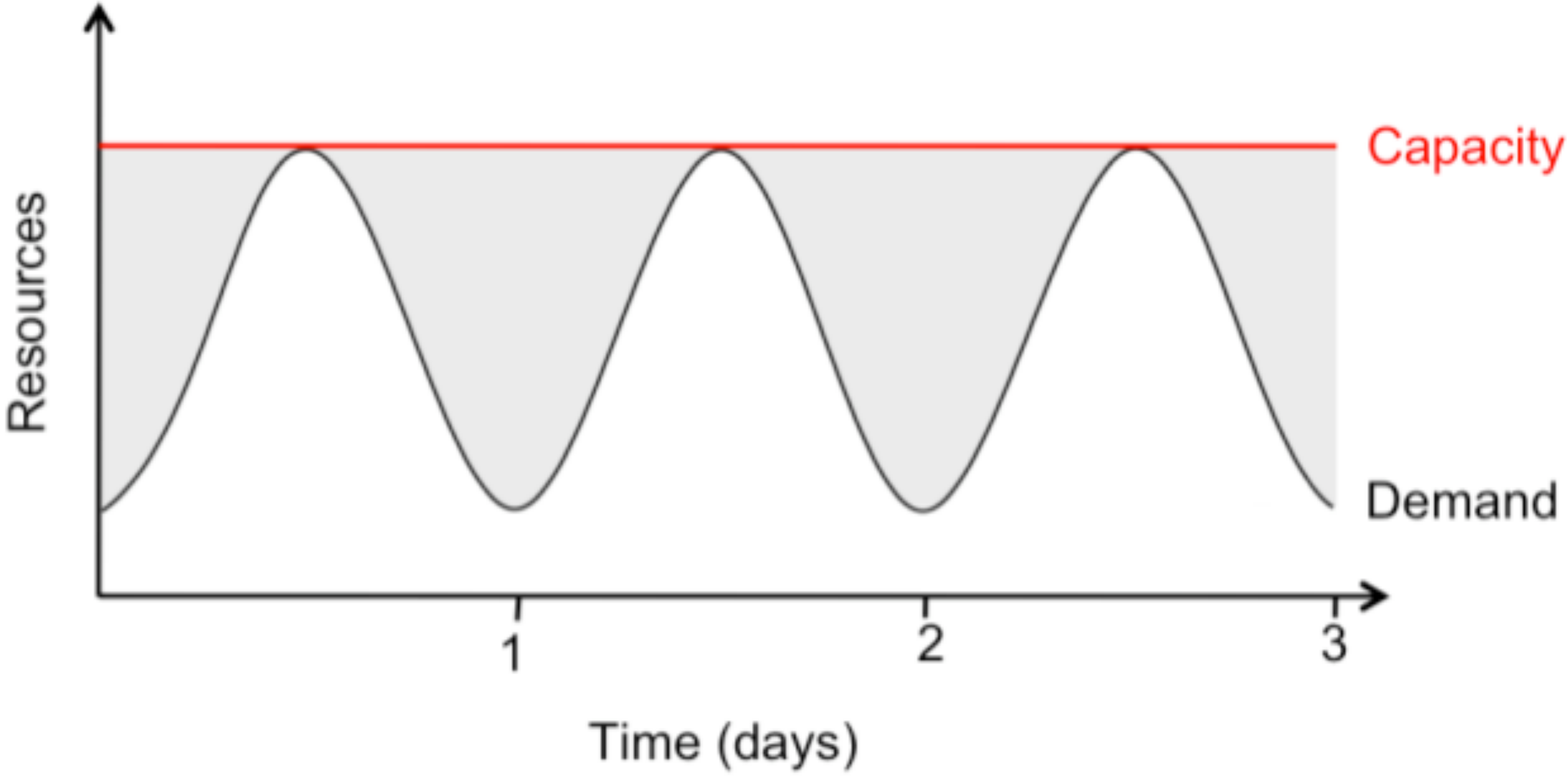
- Is this affordable?

Economies of Scale



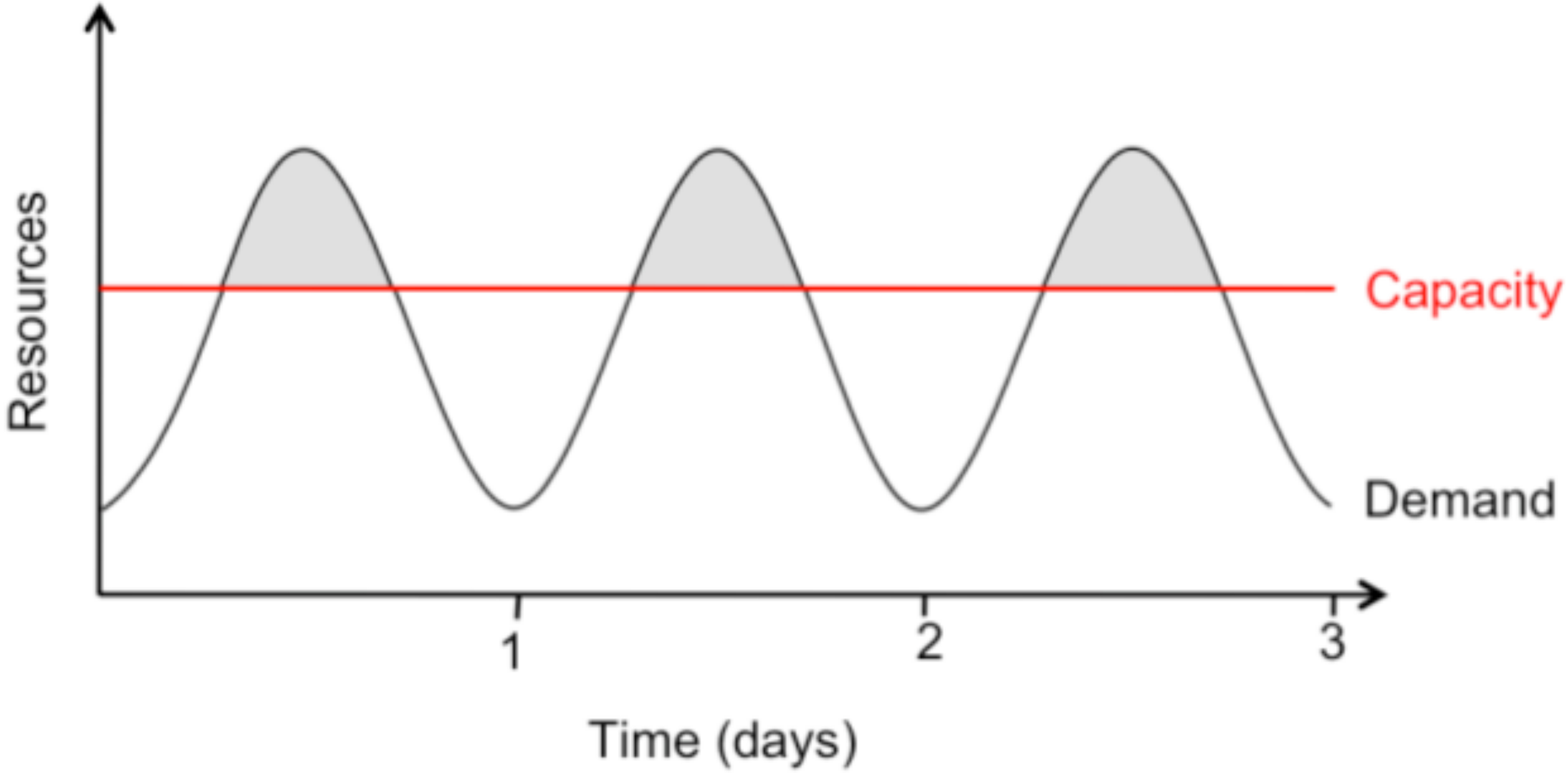
Technology	Cost in Medium-sized DC	Cost in Very Large DC	Ratio
Network	$95 per Mbit/sec/month	$13 per Mbit/sec/month	7.1
Storage	$2.20 per GByte / month	$0.40 per GByte / month	5.7
Administration	^{3}140 Servers / Administrator	>1000 Servers / Administrator	7.1

src: Armbrust et al., Above the Clouds: A Berkeley View of Cloud Computing, 2009



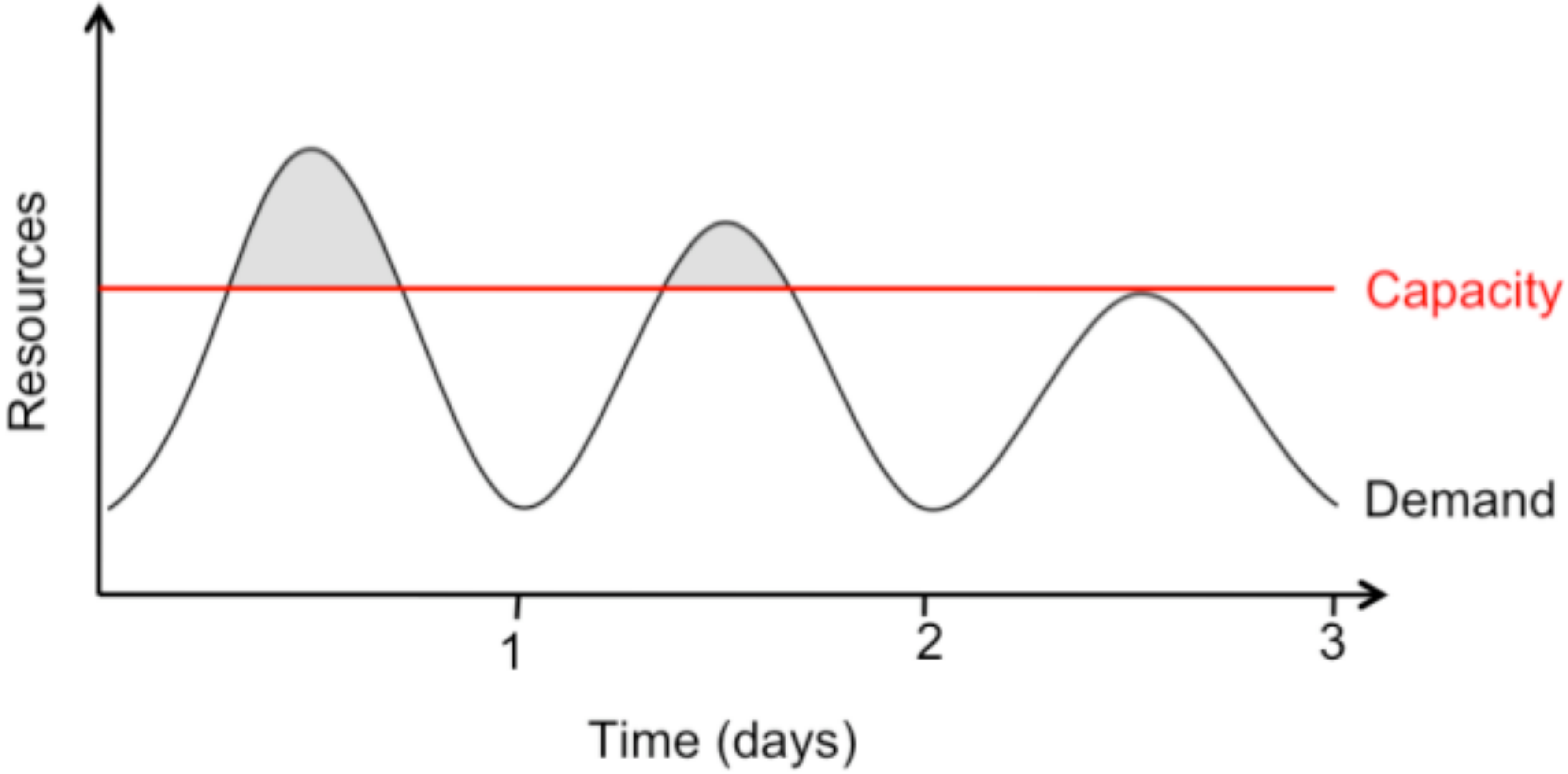
Provisioning for peak load

src: Armbrust et al., Above the Clouds: A Berkeley View of Cloud Computing, 2009



Underprovisioning

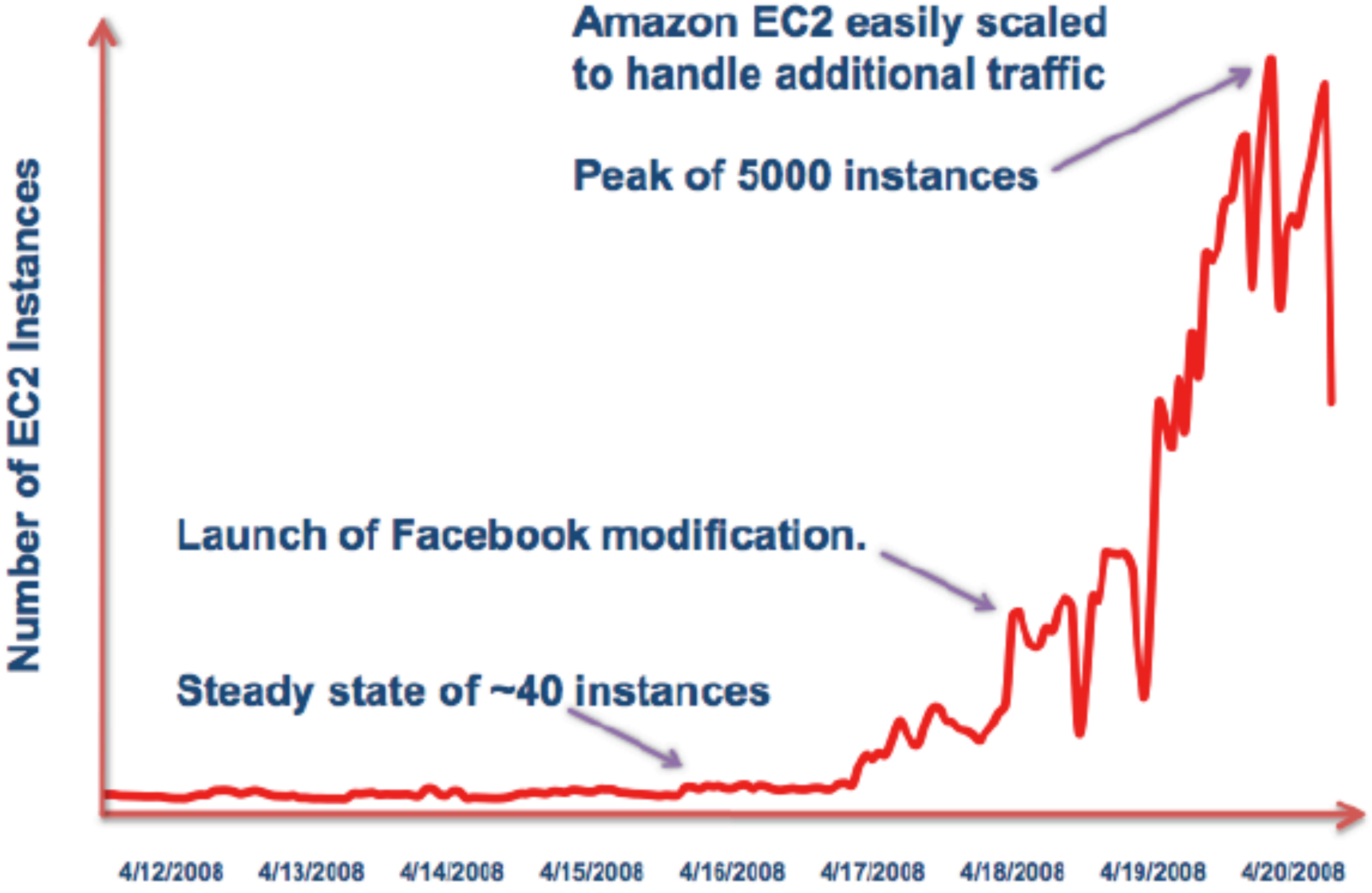
src: Armbrust et al., Above the Clouds: A Berkeley View of Cloud Computing, 2009



Underprovisioning, more realistic

src: Armbrust et al., Above the Clouds: A Berkeley View of Cloud Computing, 2009

animoto



Amazon EC2 easily scaled to handle additional traffic

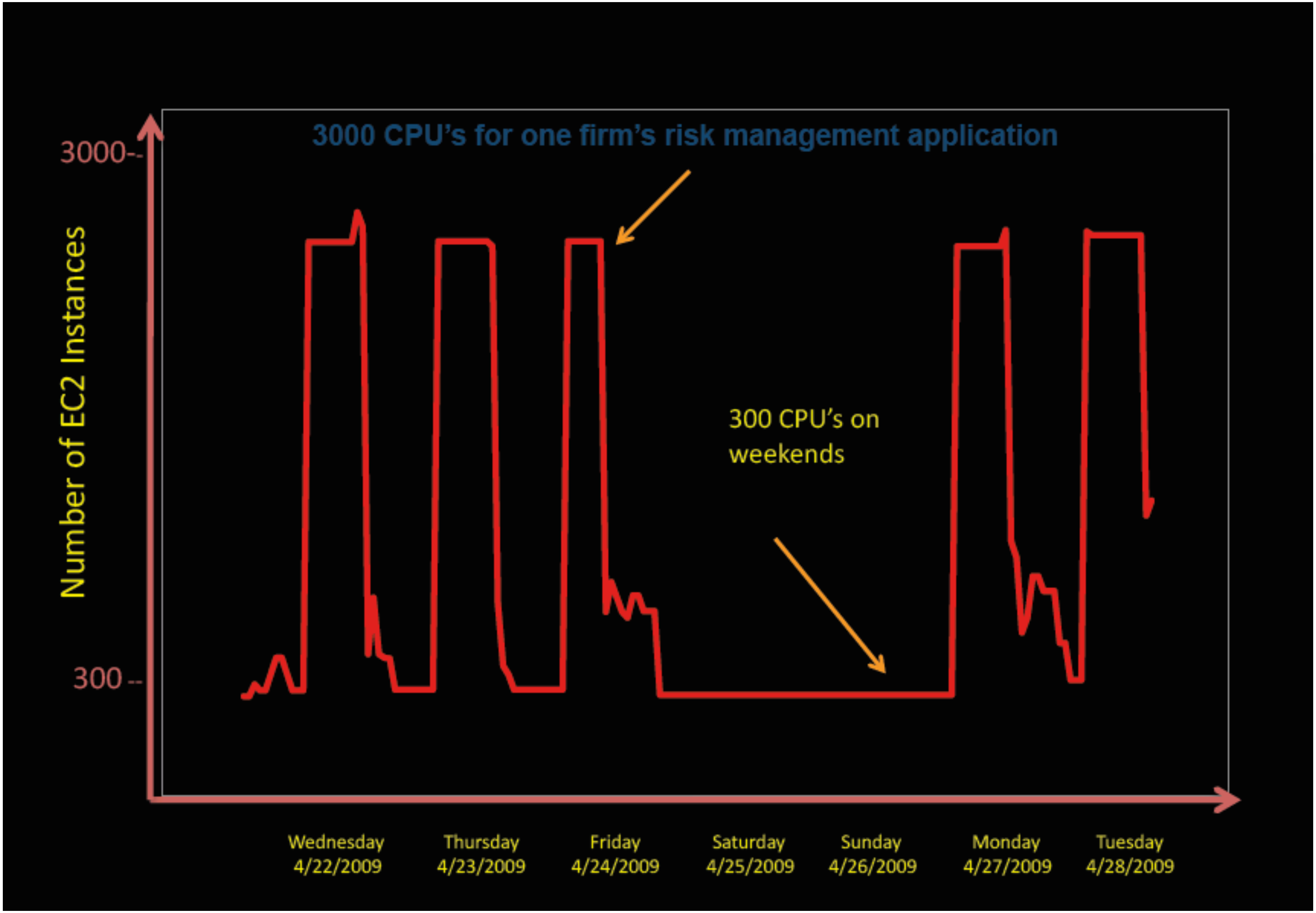
Peak of 5000 instances

Launch of Facebook modification.

Steady state of ~40 instances

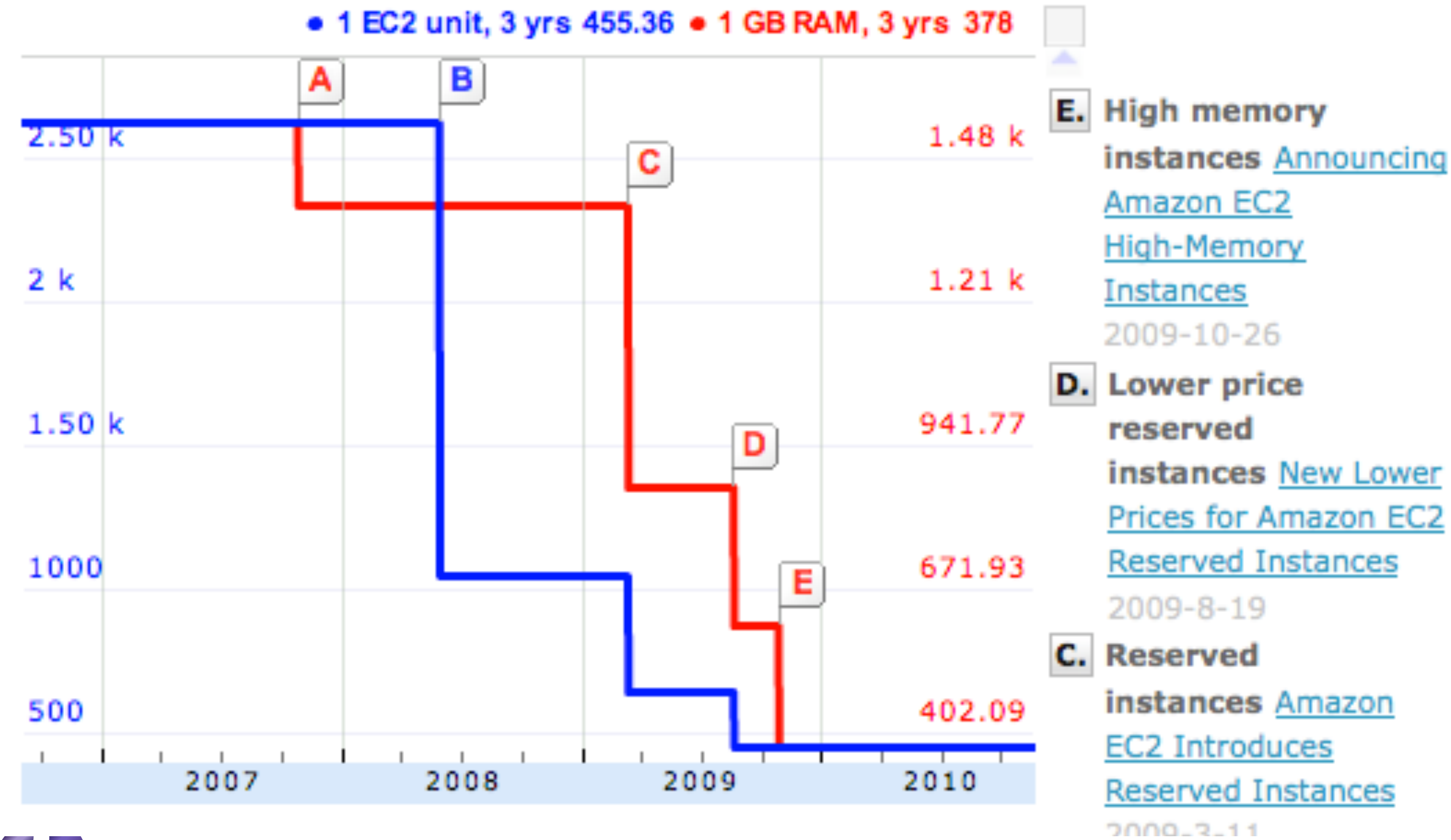
[Werner Vogels, Amazon.com]

eScience Institute



[Deepak Singh, Amazon.com]

Change in Price: compute and RAM

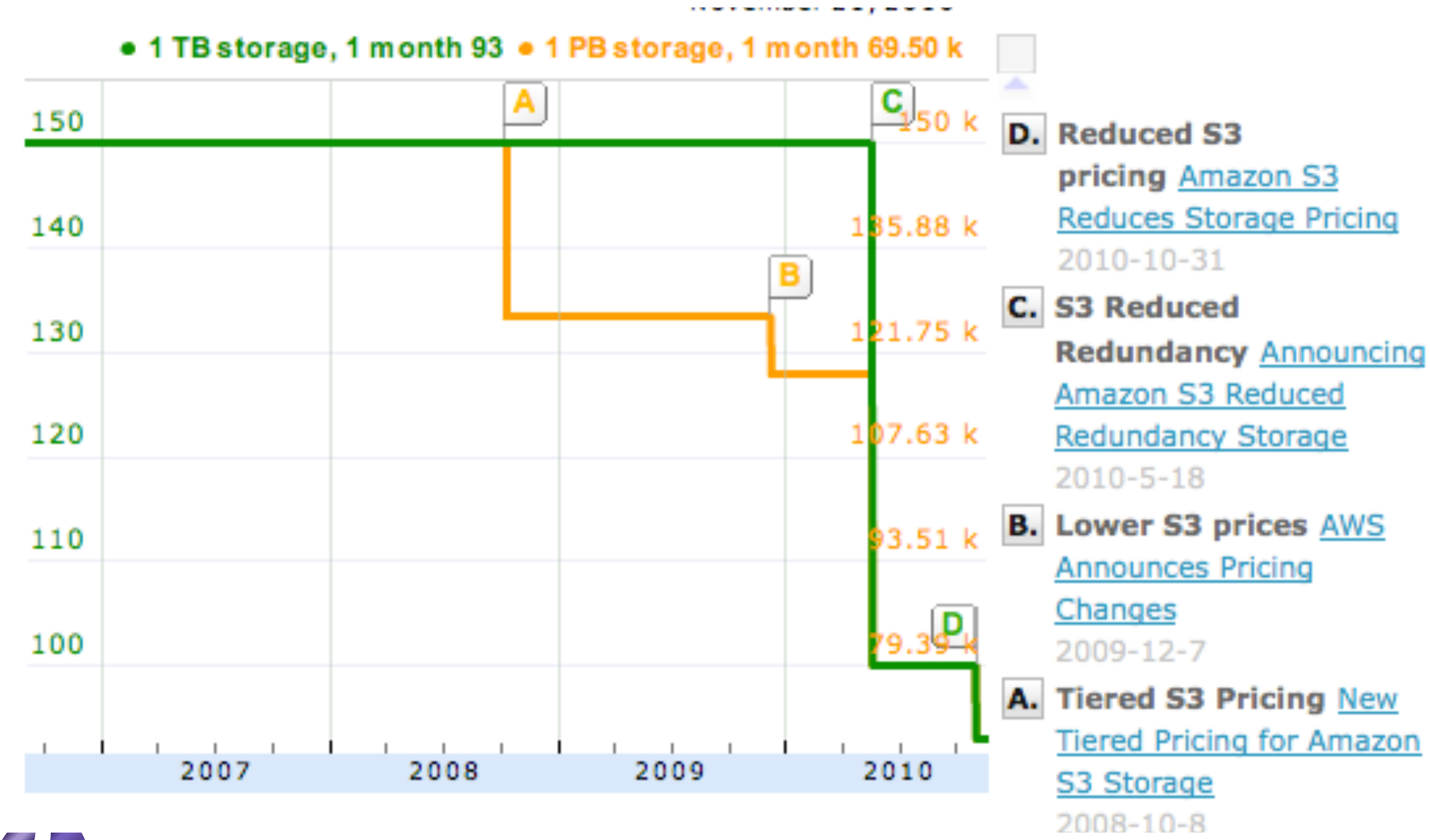


E. High memory instances Announcing Amazon EC2 High-Memory Instances 2009-10-26

D. Lower price reserved instances New Lower Prices for Amazon EC2 Reserved Instances 2009-8-19

C. Reserved instances Amazon EC2 Introduces Reserved Instances 2009-3-11

Change in price: Storage (1TB, 1PB)



D. **Reduced S3 pricing** Amazon S3 Reduces Storage Pricing
2010-10-31

C. **S3 Reduced Redundancy** Announcing Amazon S3 Reduced Redundancy Storage
2010-5-18

B. **Lower S3 prices** AWS Announces Pricing Changes
2009-12-7

A. **Tiered S3 Pricing** New Tiered Pricing for Amazon S3 Storage
2008-10-8

Aside: Fix the funny money

- Computing equipment incurs no indirect costs
  - "Capital Expenditures"
  - Power, cooling, administration?
- "Services" are charged full indirect cost load
  - Ex: 54% at UW; 100% at Stanford

- So every dollar spent on Amazon costs the PI $1.54
- Every dollar spent on equipment costs the PI $1.00, but also costs the university ~$1.00

Bottom line?

- Buy the equipment if
  - Utilization over 90%
  - You need big archival storage (“data cemetery”)
- Otherwise, you probably shouldn’t
- Check the pricing calculator

http://calculator.s3.amazonaws.com/calc5.html

Aside: Quantifying the Value of Data

- Ex: Azure marketplace http://www.microsoft.com/windowsazure/marketplace/
- New NSF grant to study data pricing
  - Early results: proof that there is no non-trivial pricing function that can prevent arbitrage and respects monotonicity
- Unpopular idea: Can we sell access to data to fund its preservation?
  - Might be required – it's becoming clear we can't keep everything
  - Important data (heavily used data) is "worth more." Which means: easier to amortize the cost of storage.
- Beyond money: Value models may be useful to formalize attribution requirements.
  - If I use your data in my research, I am "charged."
  - Minimal usage is free
  - At some threshold, citation is expected
  - At some theshold, acknowledgement is expected
  - At some threshold, co-authorship is expected

DATA-INTENSIVE EXPERIMENTS

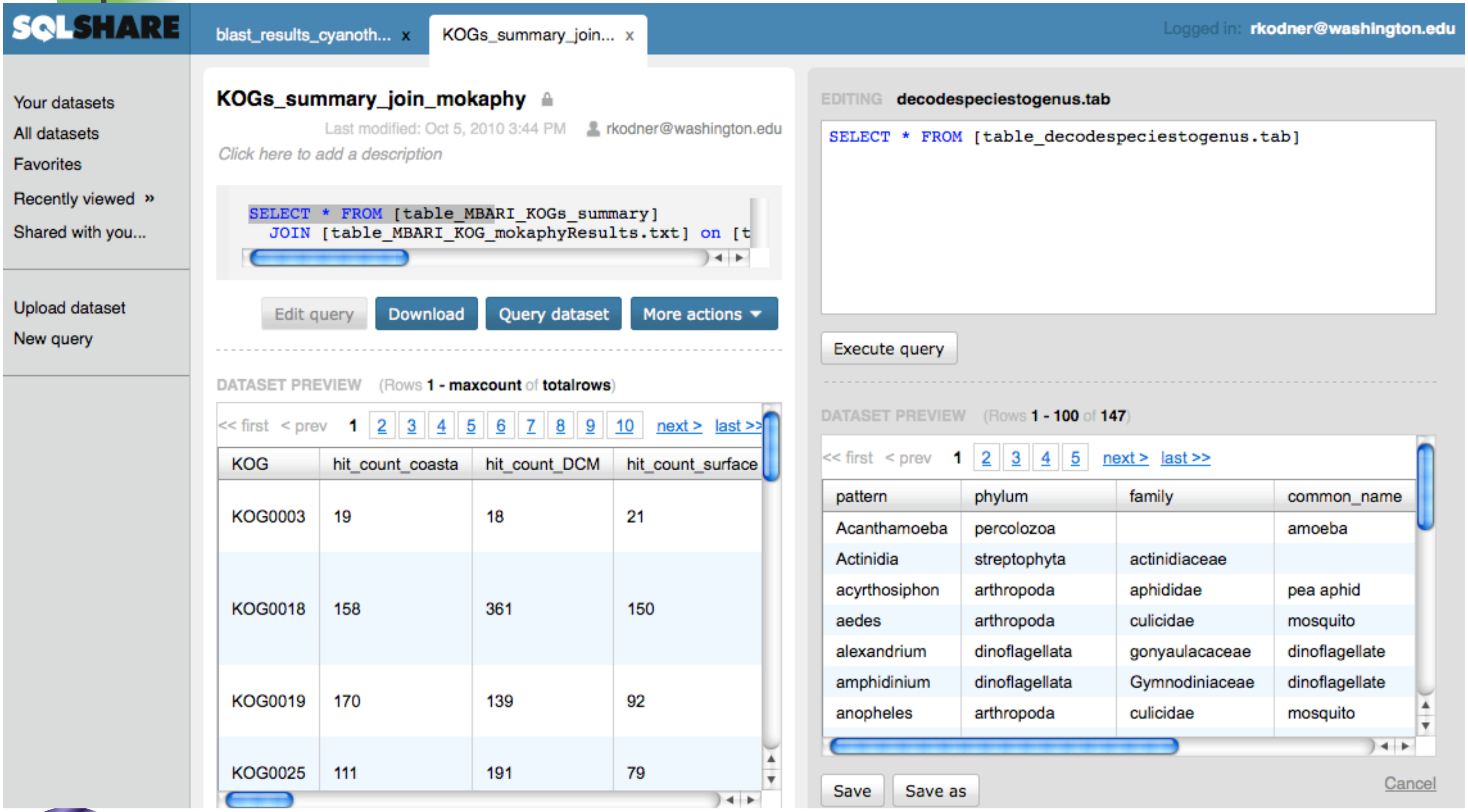
An Observation on Big Data

- The days of FTP are over
  - It takes days to transfer 1TB over the Internet, and it isn't likely to succeed.
  - Copying a petabyte is operationally impossible
- The only solution: Push the computation to the data, rather than push the data to the computation
  - Upload your code rather than download the data

Another Observation

- RR tends to emphasize computation rather than data
- Re-executing “canned” experiments is not enough
- Need to support ad hoc, exploratory Q&A, which means:
- Queries, not programs
- Databases, not files

Database-as-a-Service for Science



KOGs_summary_join_mokaphy

Last modified: Oct 5, 2010 3:44 PM rkodner@washington.edu

Click here to add a description

```
SELECT * FROM [table_MBARI_KOGs_summary]
  JOIN [table_MBARI_KOG_mokaphyResults.txt] on [t
```

DATASET PREVIEW (Rows 1 - maxcount of totalrows)

KOG	hit_count_coasta	hit_count_DCM	hit_count_surface
KOG0003	19	18	21
KOG0018	158	361	150
KOG0019	170	139	92
KOG0025	111	191	79

EDITING **decodespeciestogenus.tab**

```
SELECT * FROM [table_decodespeciestogenus.tab]
```

DATASET PREVIEW (Rows 1 - 100 of 147)

pattern	phylum	family	common_name
Acanthamoeba	percolozoa		amoeba
Actinidia	streptophyta	actinidiaceae	
acyrthosiphon	arthropoda	aphididae	pea aphid
aedes	arthropoda	culicidae	mosquito
alexandrium	dinoflagellata	gonyaulacaceae	dinoflagellate
amphidinium	dinoflagellata	Gymnodiniaceae	dinoflagellate
anopheles	arthropoda	culicidae	mosquito

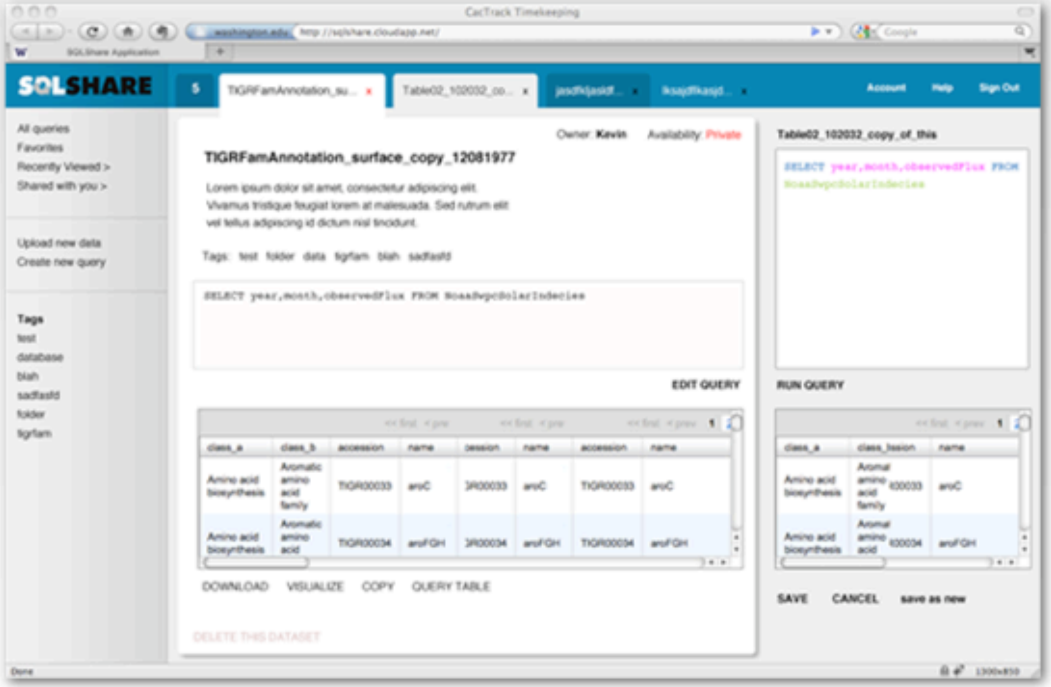
eScience Institute

http://escience.washington.edu/sqlshare

Microsoft® SQL Azure

SQLSHARE

SQLShare is an easier way to store and share your data. Get answers to your research questions right now.



Owner: Kevin Availability: Private

TIGRFamAnnotation_surface_copy_12081977

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus tristique feugiat lorem at malesuada. Sed rutrum elit vel tellus adipiscing id dictum nisl tincidunt.

Tags: test folder data tigrfam blah sadfasfd

```
SELECT year,month,observedFlux FROM NoaaSwpcSolarIndecies
```

class_a	class_b	accession	name	session	name	accession	name
Amino acid biosynthesis	Aromatic amino acid family	TIGR00033	aroC	3R00033	aroC	TIGR00033	aroC
Amino acid biosynthesis	Aromatic amino acid	TIGR00034	aroFGH	3R00034	aroFGH	TIGR00034	aroFGH

Table02_102032_copy_of_this

```
SELECT year,month,observedFlux FROM
NoaaSwpcSolarIndecies
```

class_a	class_lsion	name
Amino acid biosynthesis	Aromal amino acid family I00033	aroC
Amino acid biosynthesis	Aromal amino acid I00034	aroFGH

Log in using your account:

W UNIVERSITY *of* WASHINGTON

Google

Don't have an account?

Create a Google Account and start using SQLShare quickly.

Upload

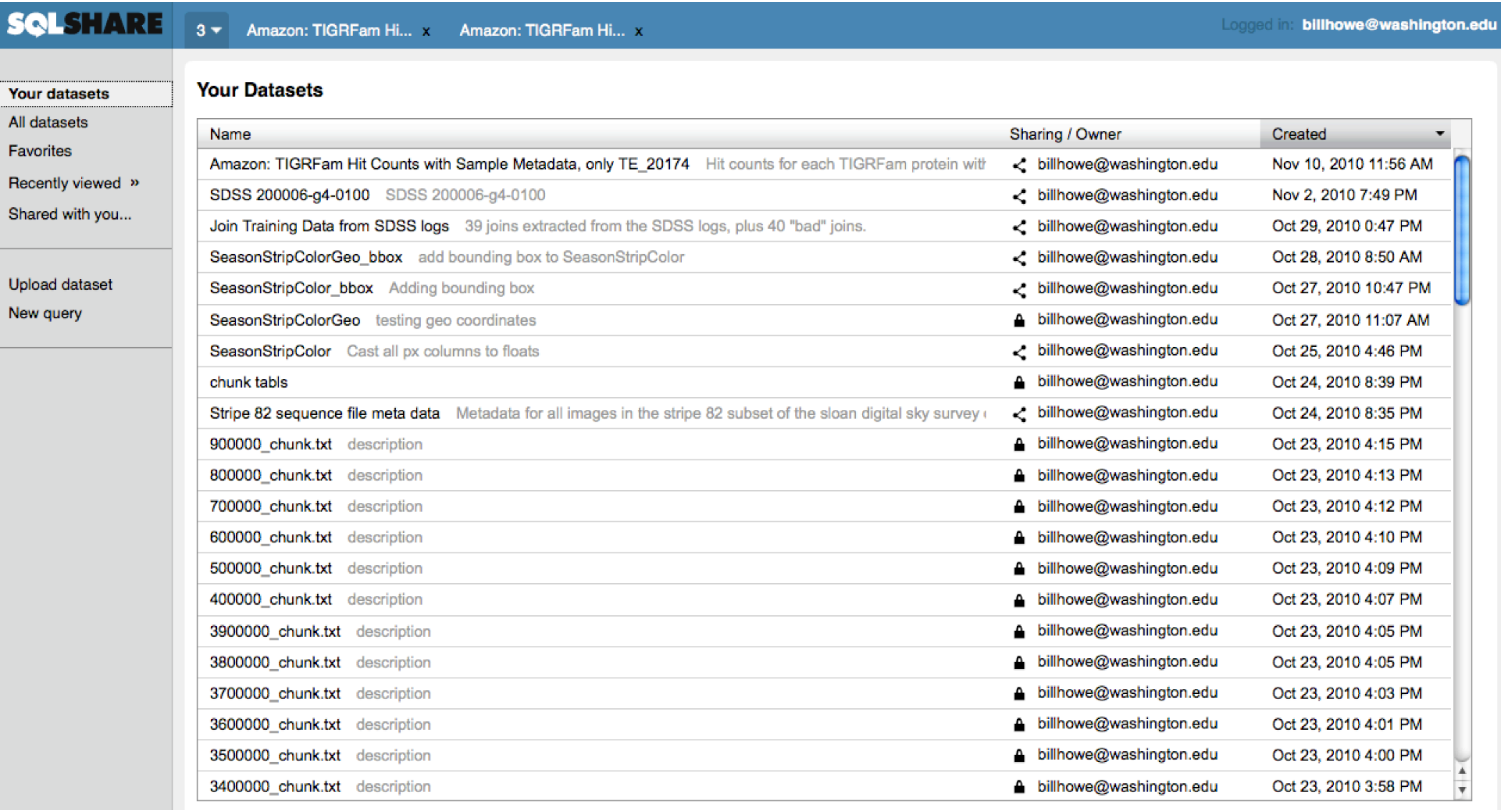
Upload any tabular data and start analyzing instantly. No need to install, configure, or design a database.

Modify

Exercise the full power of SQL even with zero programming experience: joins, subqueries, set operations.

Share

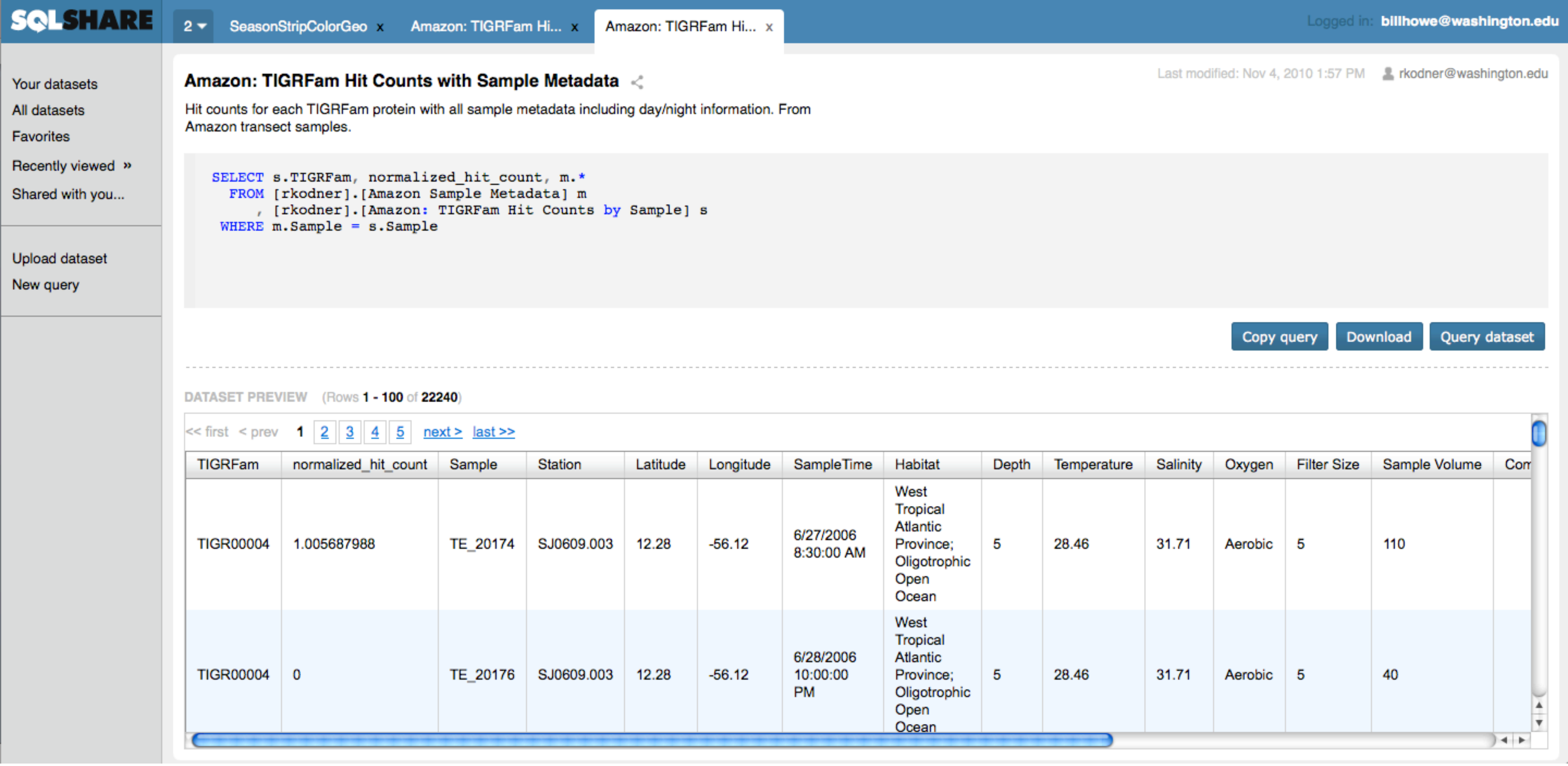
Analyze and compare your data collaboratively. Derive new datasets and share them with your colleagues.



Your Datasets

Name	Sharing / Owner	Created
Amazon: TIGRFam Hit Counts with Sample Metadata, only TE_20174 Hit counts for each TIGRFam protein wit	< billhowe@washington.edu	Nov 10, 2010 11:56 AM
SDSS 200006-g4-0100 SDSS 200006-g4-0100	< billhowe@washington.edu	Nov 2, 2010 7:49 PM
Join Training Data from SDSS logs 39 joins extracted from the SDSS logs, plus 40 "bad" joins.	< billhowe@washington.edu	Oct 29, 2010 0:47 PM
SeasonStripColorGeo_bbox add bounding box to SeasonStripColor	< billhowe@washington.edu	Oct 28, 2010 8:50 AM
SeasonStripColor_bbox Adding bounding box	< billhowe@washington.edu	Oct 27, 2010 10:47 PM
SeasonStripColorGeo testing geo coordinates	🔒 billhowe@washington.edu	Oct 27, 2010 11:07 AM
SeasonStripColor Cast all px columns to floats	< billhowe@washington.edu	Oct 25, 2010 4:46 PM
chunk tabls	🔒 billhowe@washington.edu	Oct 24, 2010 8:39 PM
Stripe 82 sequence file meta data Metadata for all images in the stripe 82 subset of the sloan digital sky survey (	< billhowe@washington.edu	Oct 24, 2010 8:35 PM
900000_chunk.txt description	🔒 billhowe@washington.edu	Oct 23, 2010 4:15 PM
800000_chunk.txt description	🔒 billhowe@washington.edu	Oct 23, 2010 4:13 PM
700000_chunk.txt description	🔒 billhowe@washington.edu	Oct 23, 2010 4:12 PM
600000_chunk.txt description	🔒 billhowe@washington.edu	Oct 23, 2010 4:10 PM
500000_chunk.txt description	🔒 billhowe@washington.edu	Oct 23, 2010 4:09 PM
400000_chunk.txt description	🔒 billhowe@washington.edu	Oct 23, 2010 4:07 PM
3900000_chunk.txt description	🔒 billhowe@washington.edu	Oct 23, 2010 4:05 PM
3800000_chunk.txt description	🔒 billhowe@washington.edu	Oct 23, 2010 4:05 PM
3700000_chunk.txt description	🔒 billhowe@washington.edu	Oct 23, 2010 4:03 PM
3600000_chunk.txt description	🔒 billhowe@washington.edu	Oct 23, 2010 4:01 PM
3500000_chunk.txt description	🔒 billhowe@washington.edu	Oct 23, 2010 4:00 PM
3400000_chunk.txt description	🔒 billhowe@washington.edu	Oct 23, 2010 3:58 PM

eScience Institute



Amazon: TIGRFam Hit Counts with Sample Metadata

Hit counts for each TIGRFam protein with all sample metadata including day/night information. From Amazon transect samples.

```
SELECT s.TIGRFam, normalized_hit_count, m.*
  FROM [rkodner].[Amazon Sample Metadata] m
     , [rkodner].[Amazon: TIGRFam Hit Counts by Sample] s
 WHERE m.Sample = s.Sample
```

DATASET PREVIEW (Rows 1 - 100 of 22240)

TIGRFam	normalized_hit_count	Sample	Station	Latitude	Longitude	SampleTime	Habitat	Depth	Temperature	Salinity	Oxygen	Filter Size	Sample Volume	Com
TIGR00004	1.005687988	TE_20174	SJ0609.003	12.28	-56.12	6/27/2006 8:30:00 AM	West Tropical Atlantic Province; Oligotrophic Open Ocean	5	28.46	31.71	Aerobic	5	110	
TIGR00004	0	TE_20176	SJ0609.003	12.28	-56.12	6/28/2006 10:00:00 PM	West Tropical Atlantic Province; Oligotrophic Open Ocean	5	28.46	31.71	Aerobic	5	40	

Why SQL?

- Find all TIGRFam ids (proteins least one of three samples (rel

```
SELECT col0 FROM [refseq_hma
UNION
SELECT col0 FROM [est_hma_fa
UNION
SELECT col0 FROM [combo_hm

EXCEPT

SELECT col0 FROM [refseq_hma
INTERSECT
SELECT col0 FROM [est_hma_fasta_TGIRfam_refs]
INTERSECT
SELECT col0 FROM [combo_hma_fasta_TGIRfam_refs]
```

SQLShare Extension Projects

- SQL Autocomplete
  - (Nodira Khoussainova, YongChul Kwon, Magda Balazinska)
- English to SQL
  - (Bill Howe, Luke Zettlemoyer, Emad Soroush, Paras Koutris)
- Automatic “Starter” Queries
  - (Bill Howe, Garret Cole, Nodira Khoussainova, Leilani Battle)

SSDBM 2011
SIGMOD 2011 (demo)

- VizDeck: Automatic Mashups and Visualization
  - (Bill Howe, Alicia Key)

SSDBM 2011

- Personalized Query Recommendation
  - (Yuan Zhou, Bill Howe)
- Crowdsourced SQL authoring
  - (nobody)
- Info Extraction from Spreadsheets
  - (Mike Cafarella, Dave Maier, Bill Howe)
- Data P

eScience Institute

NSF

Gordon and Betty MOORE FOUNDATION

Usage

- About 8 months old, essentially zero advertising
- 8-10 labs around UW campus and externally
- 51 unique users (UW and external)
- ~1200 tables (~400 are public)
- ~900 views (~300 are public)
- ~5000 queries executed.
- ~40 GB (these are SMALL datasets!)
- largest table: 1.1M rows
- smallest table: 1 row

Big Data (2)

- Distributed computation is hard
  - VMs aren’t enough
- Need native services for big data, not (just) storage
- Elastic MapReduce
  - Integrated with S3 – any data in S3 can be processed with MapReduce
- Languages over MapReduce
  - Pig (Relational Algebra, from Yahoo)
  - HIVE (SQL, from Facebook)

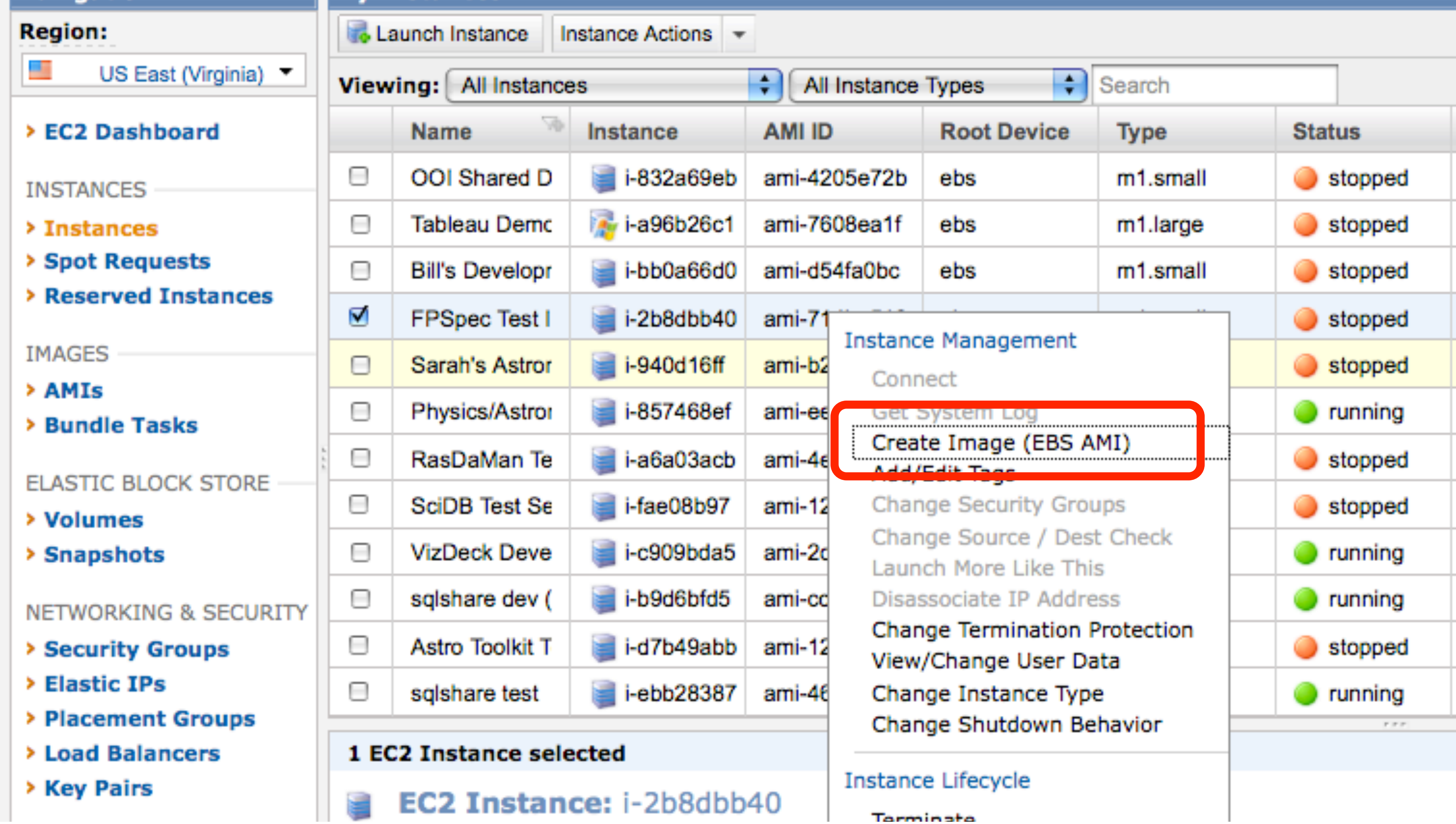
Cloud Services for Big Data

Product	Provider	Prog. Model	Storage Cost	Compute Cost	IO Cost
Megastore	Google	Filter	$0.15 / GB / mo.	$0.10 / corehour	$.12 / GB out
BigQuery	Google	SQL-like	Closed beta	Closed beta	Closed beta
Microsoft Table	Microsoft	Lookup	$0.15 / GB / mo.	$0.12 / hour and up	$.15 / GB out
Elastic MapReduce	Amazon	MR, RA-like, SQL	$0.093 / GB / mo.	$0.10 / hour and up	$0.15 / GB out (1st GB free)
SimpleDB	Amazon	Filter	$0.093 / GB / mo.	1st 25 hours free, $0.14 after that	$0.15 / GB out (1st GB free)

http://escience.washington.edu/blog

Recommendations (last slide)

- Cloud is absolutely mainstream
- Try it. Get your computing out of the closet.
- Create VMs. Cite them. (If cost is the issue, contact me)
- For data-intensive experiments, data hosting is still expensive, but you're not likely to do better yourself.
- Prices are dropping, new services are released literally monthly
- Tell your university to stop charging overhead on cloud services
- My opinion: In 10 years, everything will be in the cloud
- *"I think there is a world market for maybe 5 computers"*

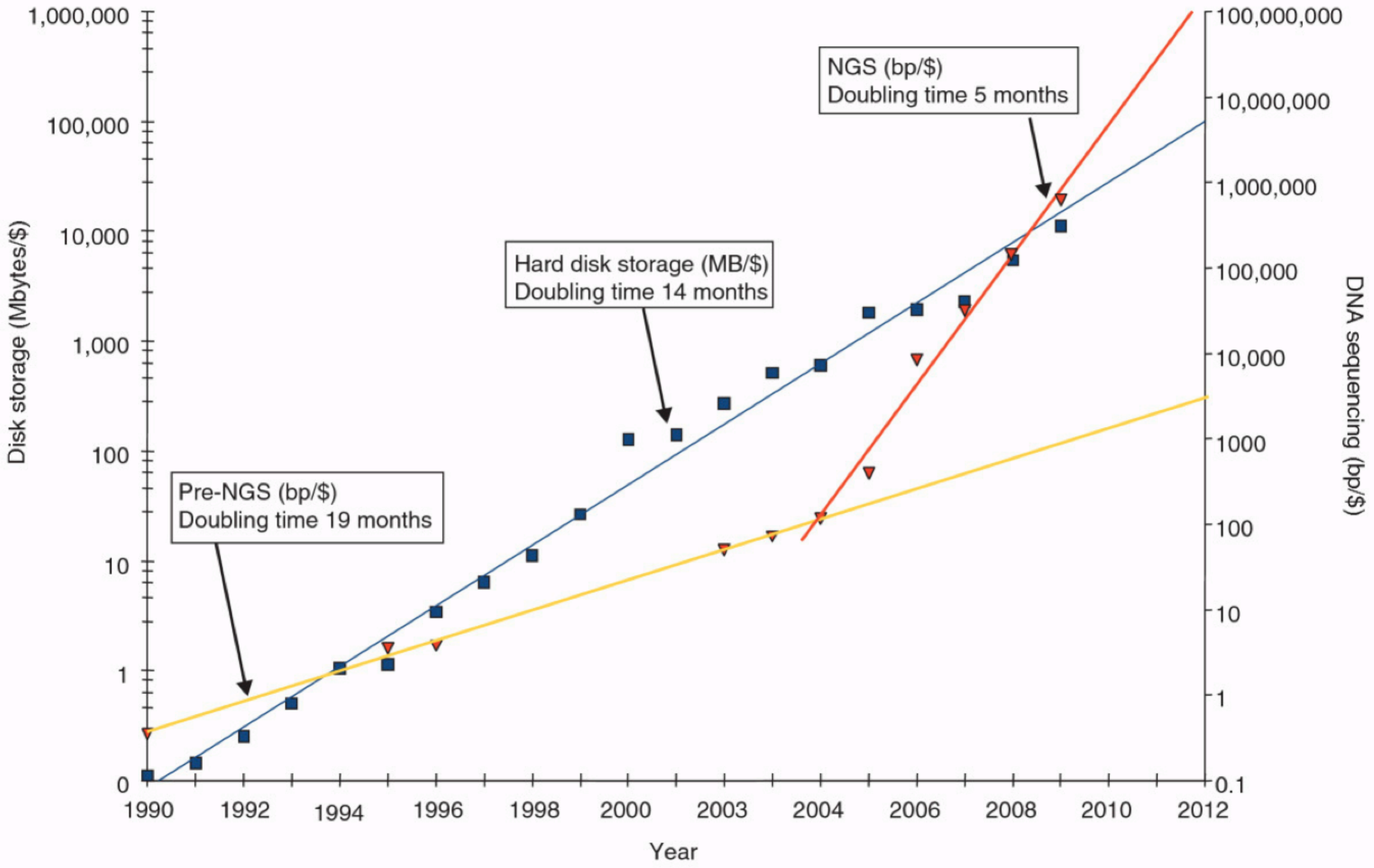


	Name	Instance	AMI ID	Root Device	Type	Status
☐	OOI Shared D	i-832a69eb	ami-4205e72b	ebs	m1.small	● stopped
☐	Tableau Demc	i-a96b26c1	ami-7608ea1f	ebs	m1.large	● stopped
☐	Bill's Developr	i-bb0a66d0	ami-d54fa0bc	ebs	m1.small	● stopped
☑	FPSpec Test I	i-2b8dbb40	ami-71			● stopped
☐	Sarah's Astror	i-940d16ff	ami-b2			● stopped
☐	Physics/Astror	i-857468ef	ami-ee			● running
☐	RasDaMan Te	i-a6a03acb	ami-4e			● stopped
☐	SciDB Test Se	i-fae08b97	ami-12			● stopped
☐	VizDeck Deve	i-c909bda5	ami-2c			● running
☐	sqlshare dev (	i-b9d6bfd5	ami-cc			● running
☐	Astro Toolkit T	i-d7b49abb	ami-12			● stopped
☐	sqlshare test	i-ebb28387	ami-46			● running

1 EC2 Instance selected

EC2 Instance: i-2b8dbb40

NextGen Sequencing a Game-Changer

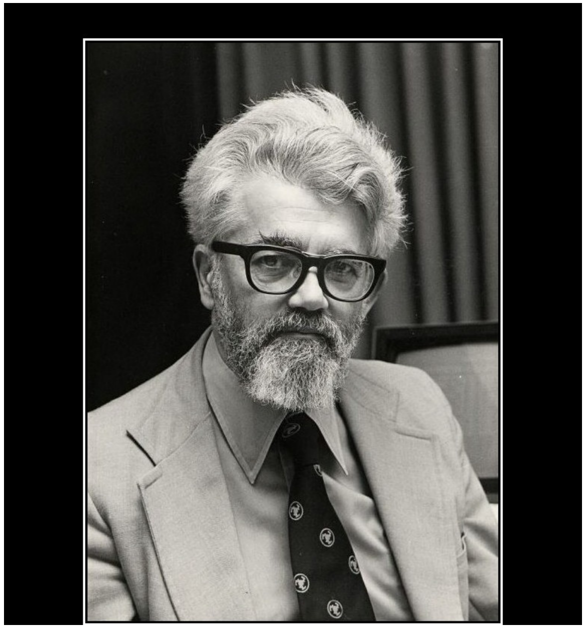


src: Lincoln Stein

Exemplars

- Software as a Service — Google docs
- Platform as a Service — Windows Azure
- Infrastructure as a Service — amazon web services™

"... computing may someday be organized as a public utility just as the telephone system is a public utility... The computer utility could become the basis of a new and important industry.”



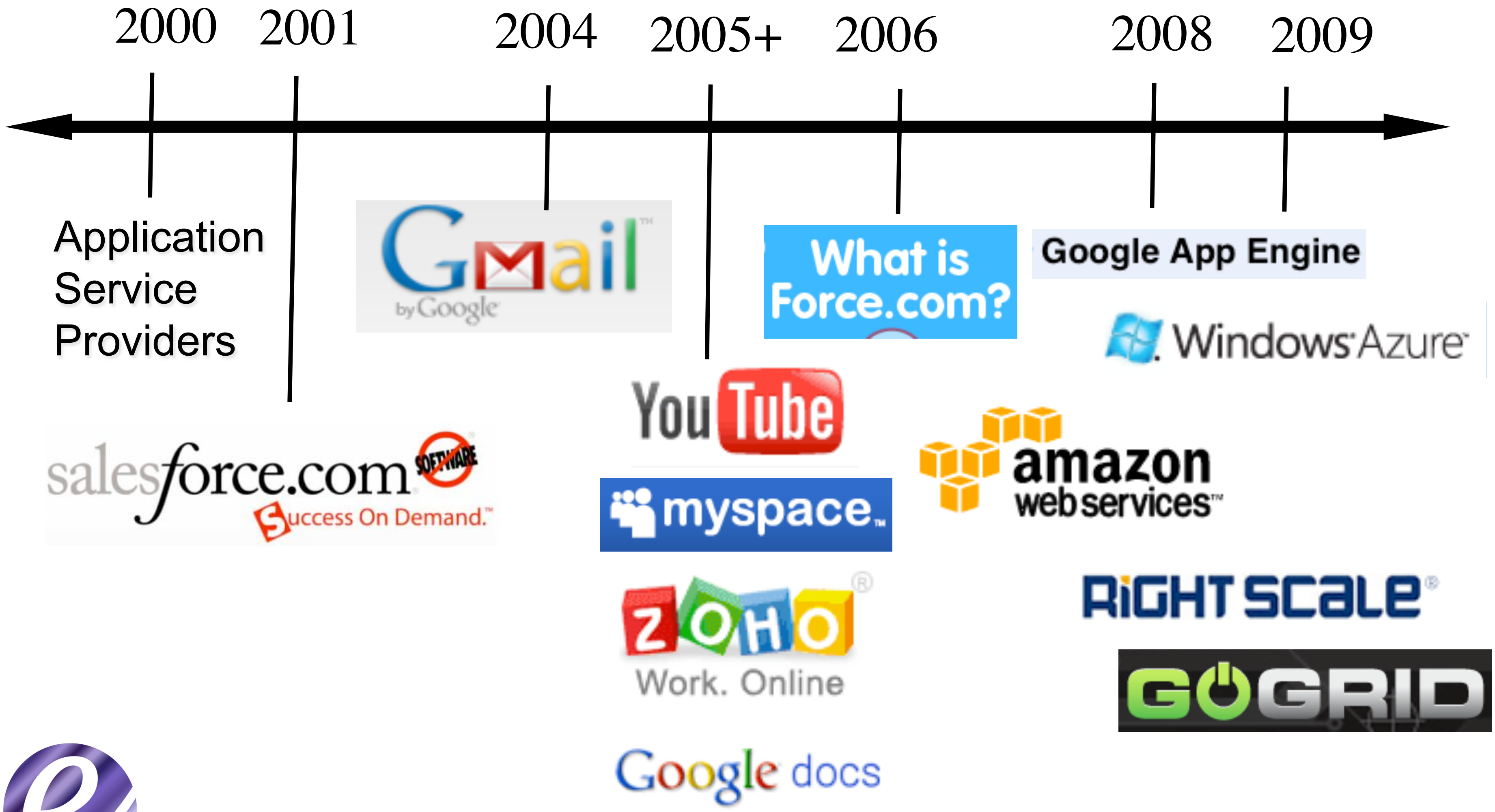
-- *John McCarthy*

Emeritus at Stanford

Inventor of LISP

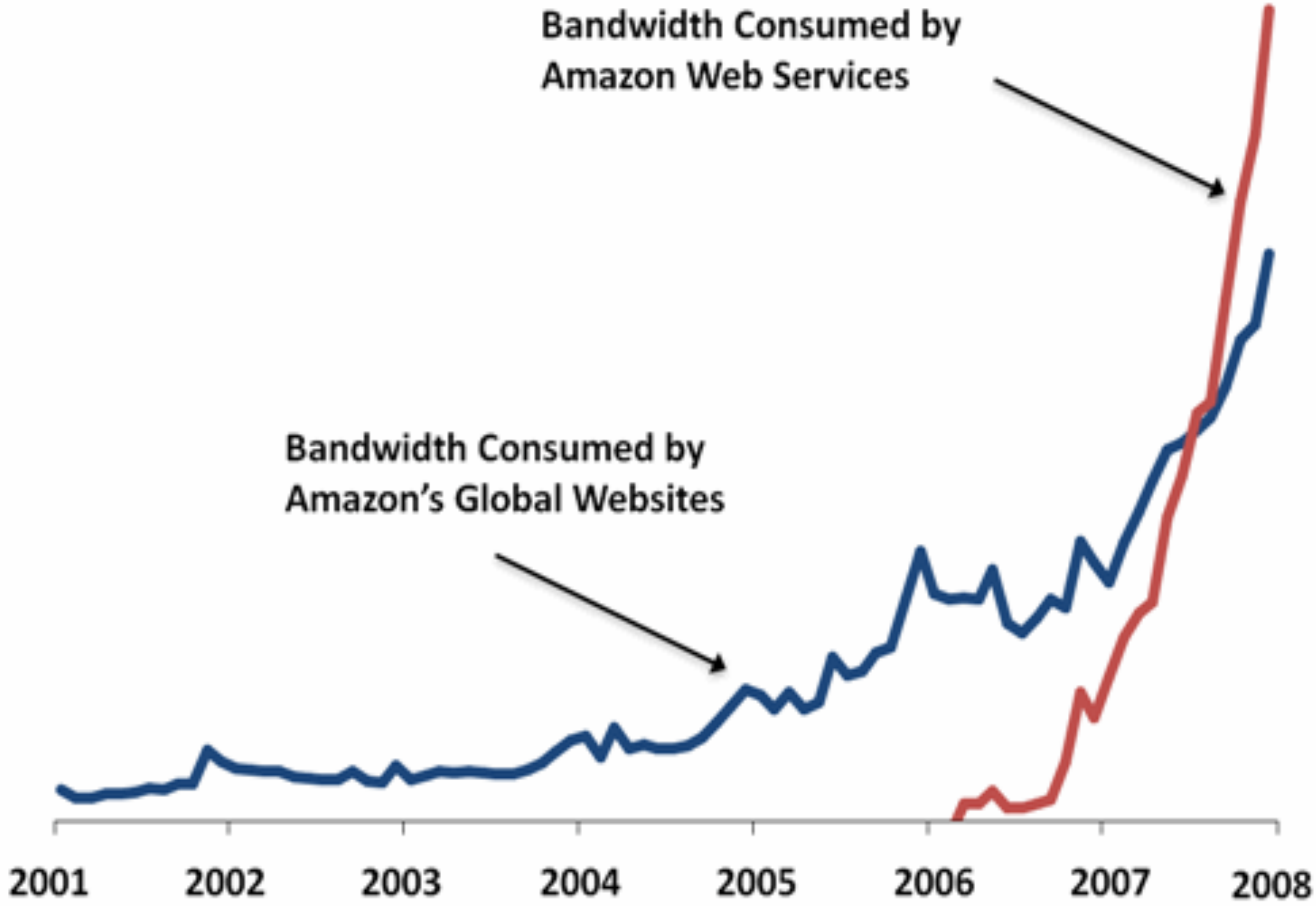
1961

Timeline



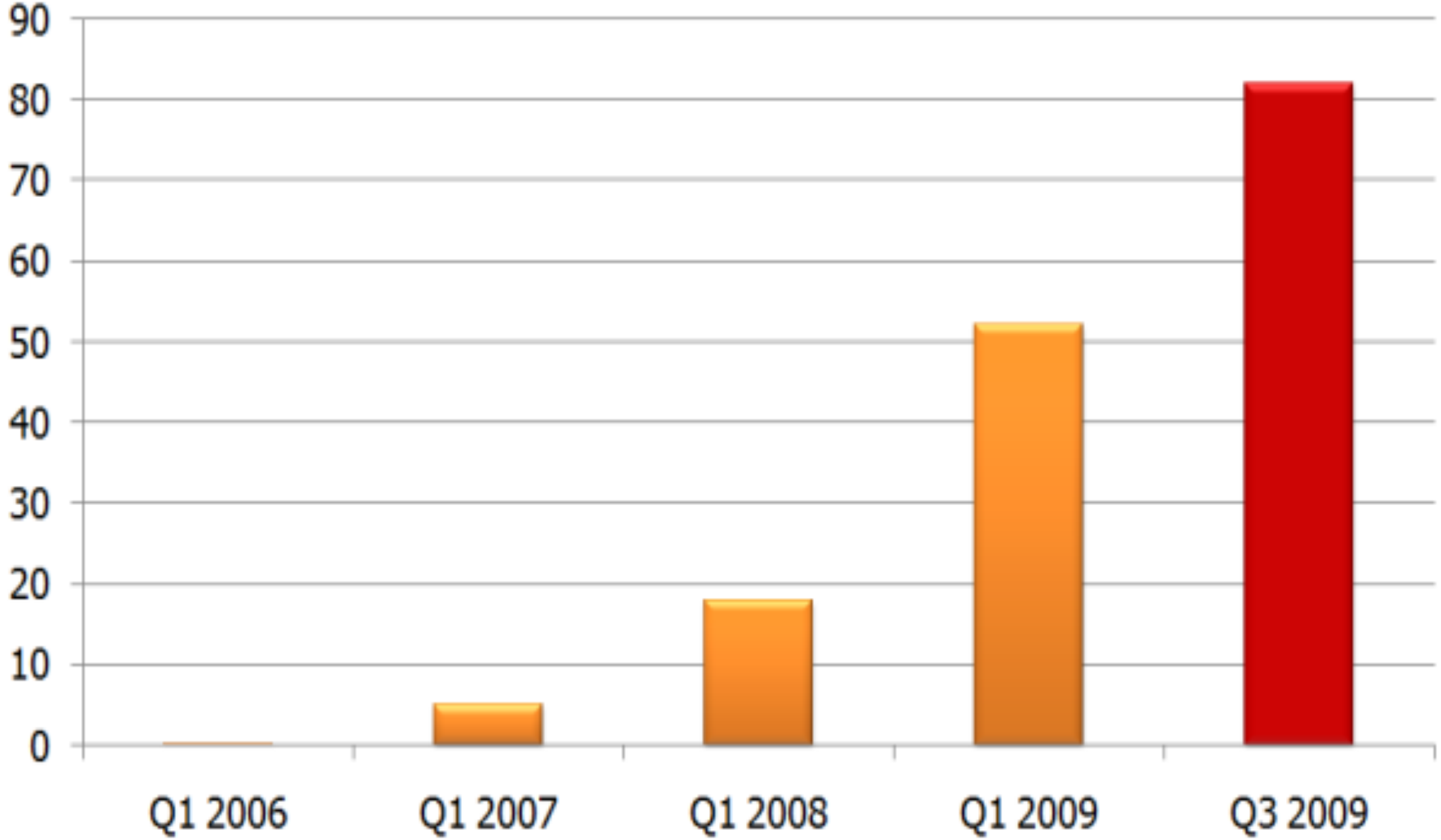
2000 2001 2004 2005+ 2006 2008 2009

Application Service Providers



[Werner Vogels, Amazon.com]

82 Billion Objects in Amazon S3



[Werner Vogels, Amazon.com]

The University of Washington eScience Institute

- Rationale
  - The exponential increase in physical and virtual sensing tech is transitioning all fields of science and engineering from *data-poor to data-rich*
  - Techniques and technologies include
    - Sensors and sensor networks, data management, data mining, machine learning, visualization, cluster/cloud computing
  - If these techniques and technologies are not widely available and widely practiced, UW will cease to be competitive
- Mission
  - Help position the University of Washington and partners at the forefront of research both in modern eScience techniques and technologies, and in the fields that depend upon them.
- Strategy
  - Bootstrap a cadre of Research Scientists
  - Add faculty in key fields
  - Build out a “consultancy” of students and non-research staff
- Funding
  - $650/year direct appropriation from WA State Legislature
  - augmented with soft money from NSF, DOE, Gordon and Betty Moore Foundation

eScience Data Management Group

**Bill Howe, Phd (databases, visualization, data-intensive scalable computing, cloud)

Staff and Post Docs

- Keith Grochow (Visualization, HCI, GIS)
- **Garret Cole (cloud computing (Azure, EC2), databases, web services)
- Marianne Shaw, Phd (health informatics, semantic web, RDF, graph databases)
- Alicia Key (visualization, user-centered design, web applications)

Students

- Nodira Khoussainova (4th yr Phd), databases, machine learning
- Leilani Battle (undergrad), databases, performance evaluation
- Yuan Zhou (masters, Applied Math), machine learning, ranking, recommender systems
- YongChul Kwon (4th yr Phd), databases, DISC, scientific applications
- Meg Whitman (undergrad)

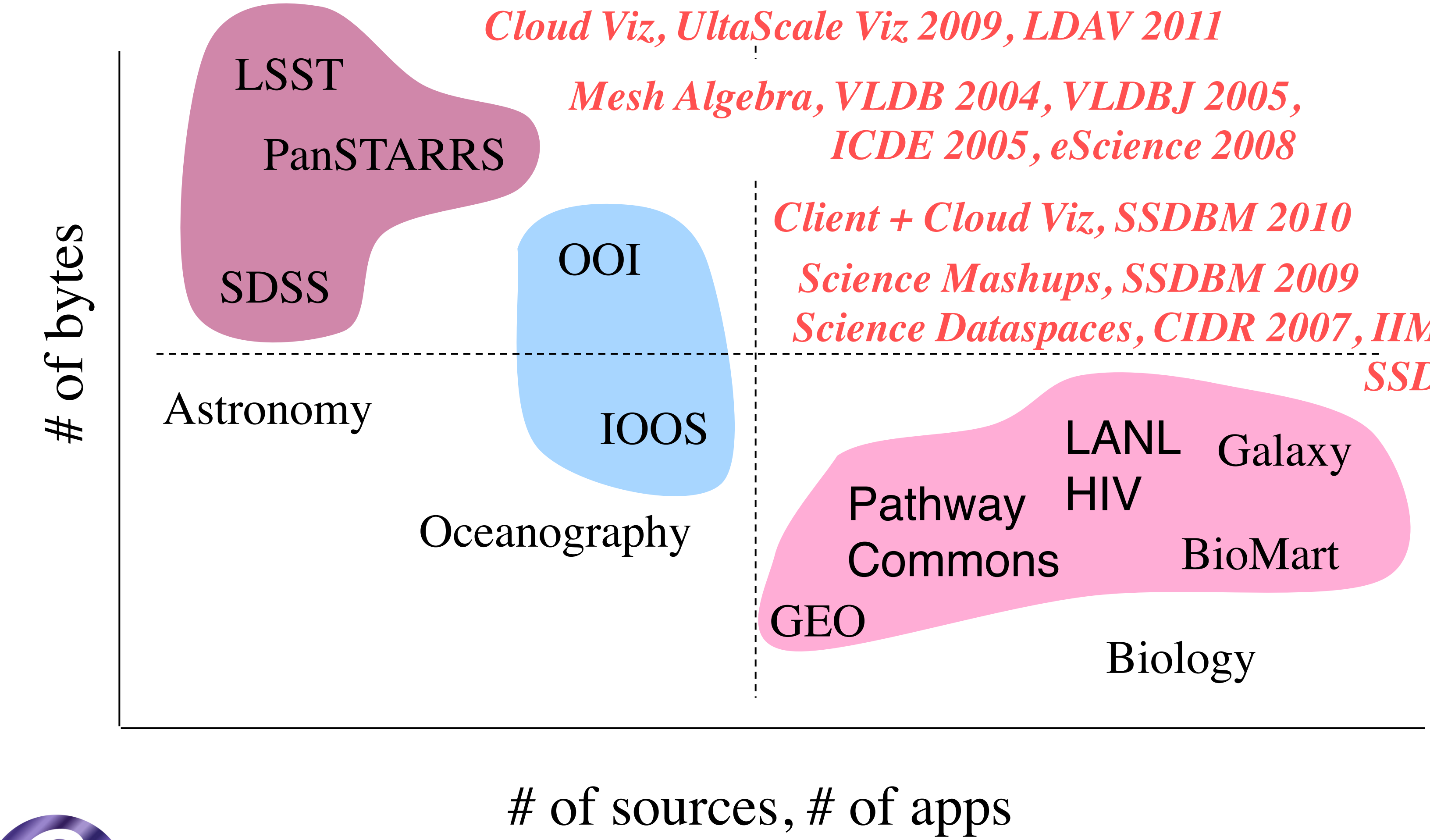
Partners

- **UW Learning and Scholarly Technologies (web applications, QA/support, release mgmt)
- **Cecilia Aragon, Phd, Associate Professor, HCDE (visualization, scientific applications)
- Magda Balazinska, Phd, Assistant Professor, CSE (databases, cloud, DISC)
- Dan Suciu, Phd, Professor, CSE, (probabilistic databases, theory, languages)

** funded in part by eScience core budget

Science Data Management

Led by Balazinska:
Skew handling, SOCC 2010
Clustering, SSDBM 2010



HaLoop, VLDB 2010

Cloud Viz, UltaScale Viz 2009, LDAV 2011

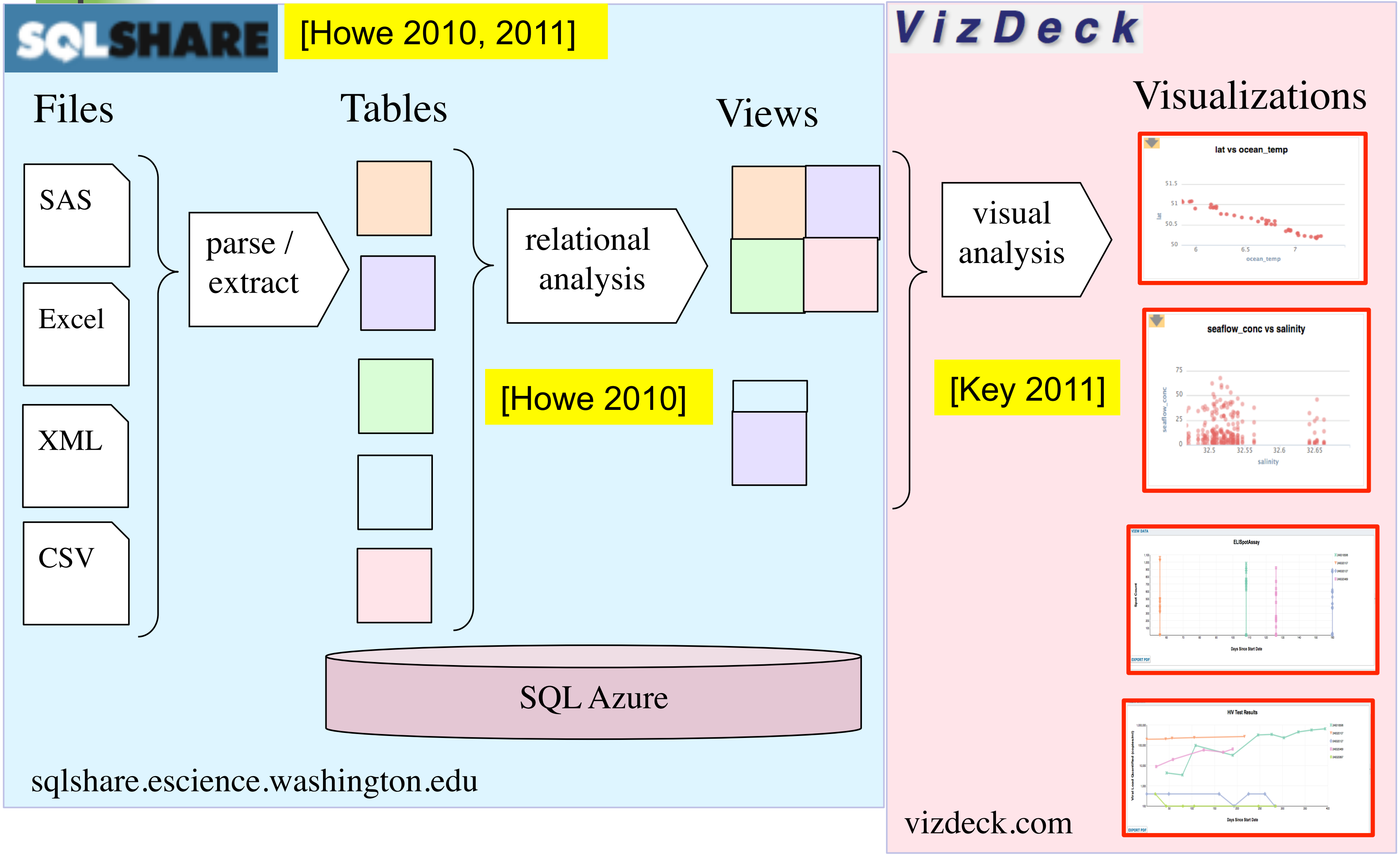
Mesh Algebra, VLDB 2004, VLDBJ 2005, ICDE 2005, eScience 2008

Client + Cloud Viz, SSDBM 2010

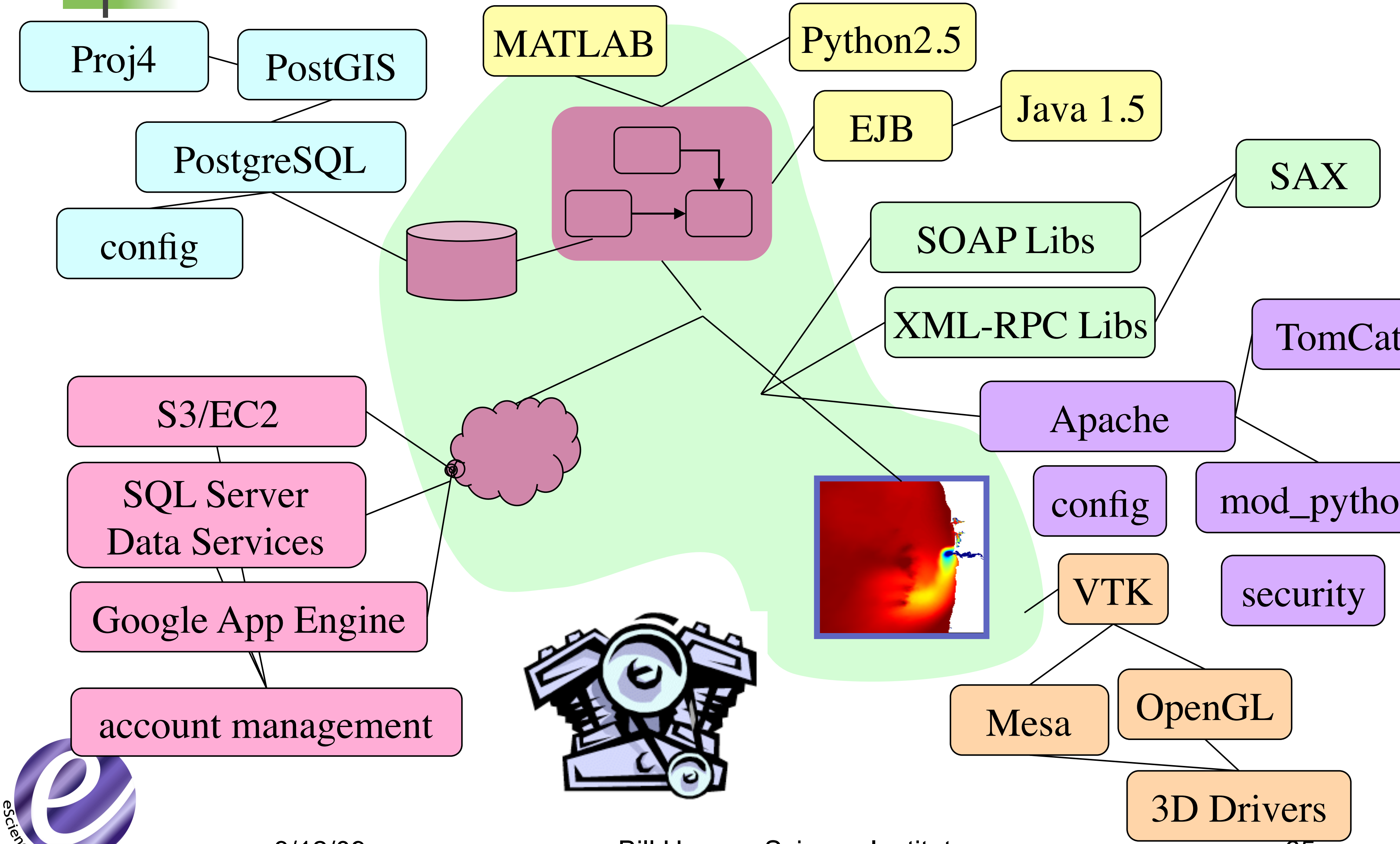
Science Mashups, SSDBM 2009

Science Dataspaces, CIDR 2007, IIMAS 2008, SSDBM 2011

Integrative Analysis



Why Virtualization? (1)



Proj4

PostGIS

PostgreSQL

config

MATLAB

Python2.5

EJB

Java 1.5

SAX

SOAP Libs

XML-RPC Libs

TomCat

Apache

config

mod_pytho

security

VTK

Mesa

OpenGL

3D Drivers

S3/EC2

SQL Server Data Services

Google App Engine

account management

Division of Responsibility

Q: Where should we place the division of responsibility between developers and users?

Need to consider skillsets

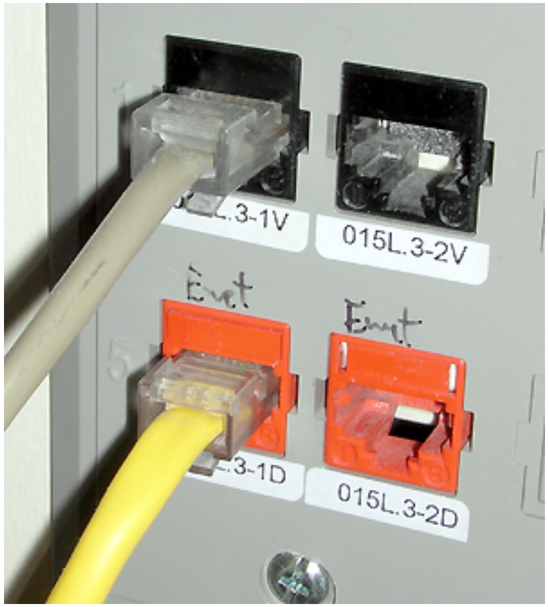
- Can they install packages?
- Can they compile code?
- Can they write DDL statements?
- Can they configure a web server?
- Can they troubleshoot network problems?
- Can they troubleshoot permissions problems?

Frequently the answer ***is*** *"No"*

Plus: Tech support is hard. Usually easier to "fix it yourself."

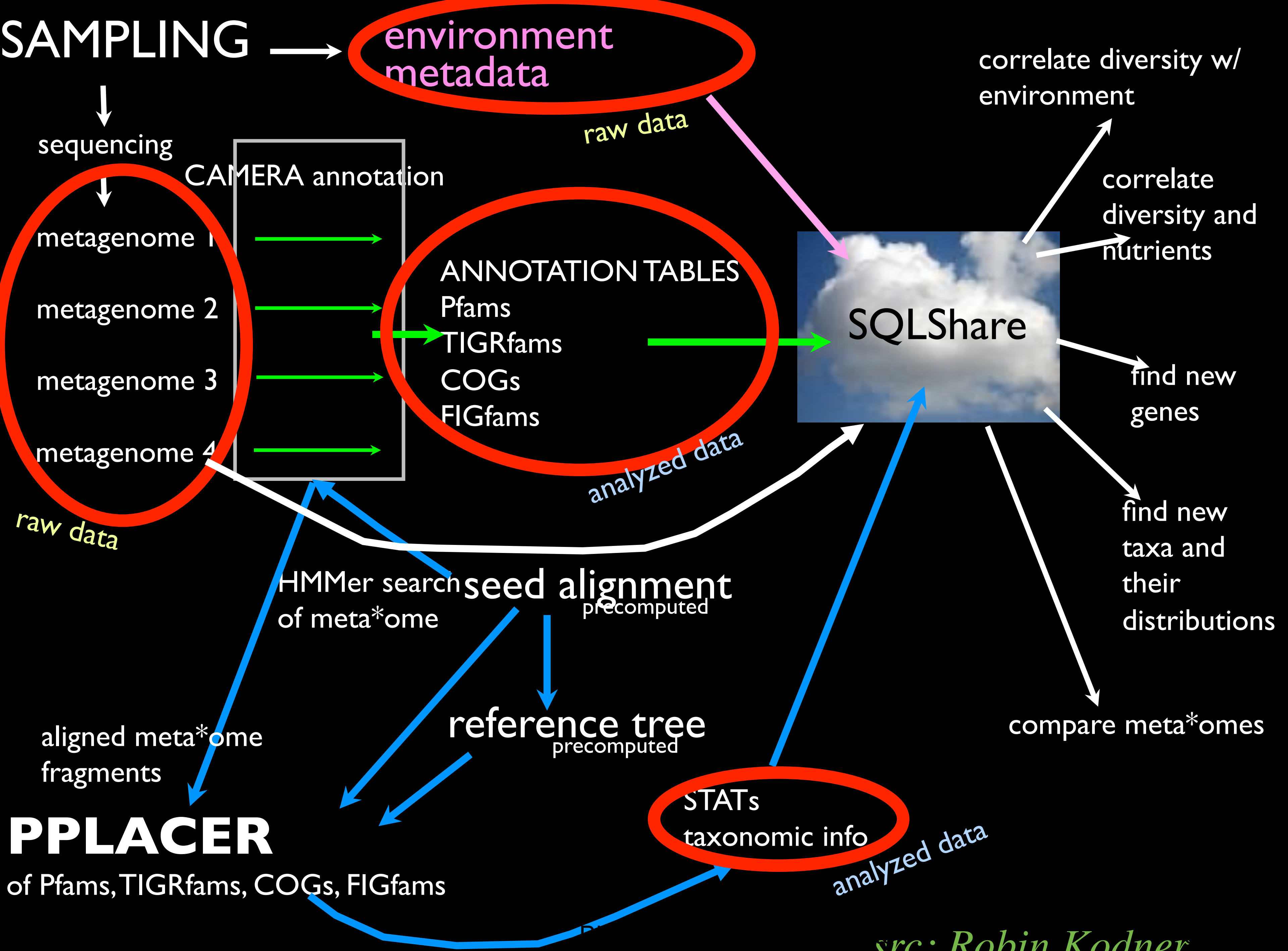
Division of Responsibility

Is there anything your peers **are** willing to do to get your software working?



Gold standard

- Your experimental procedures are completely unaffected.
- Others use your exact environment as it was at the time of the experiment.



SAMPLING

environment metadata

raw data

correlate diversity w/ environment

sequencing

CAMERA annotation

metagenome 1

metagenome 2

metagenome 3

metagenome 4

raw data

ANNOTATION TABLES
Pfams
TIGRfams
COGs
FIGfams

SQLShare

correlate diversity and nutrients

find new genes

find new taxa and their distributions

analyzed data

HMMer search of meta*ome

seed alignment precomputed

reference tree precomputed

aligned meta*ome fragments

compare meta*omes

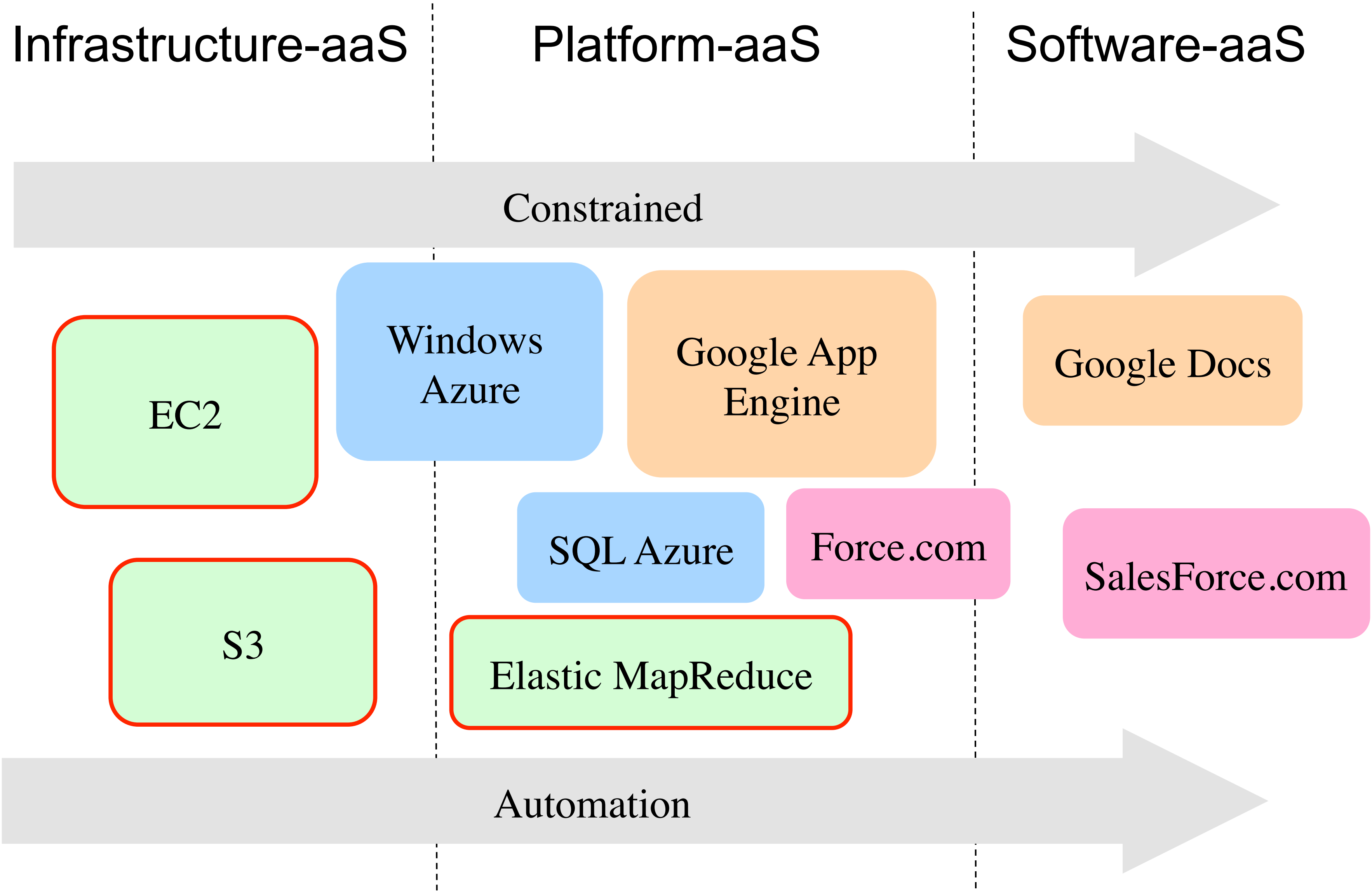
PPLACER

of Pfams, TIGRfams, COGs, FIGfams

STATs
taxonomic info

analyzed data

src: Robin Kodner



Infrastructure-aaS

Platform-aaS

Software-aaS

Constrained

EC2

Windows Azure

Google App Engine

Google Docs

SQL Azure

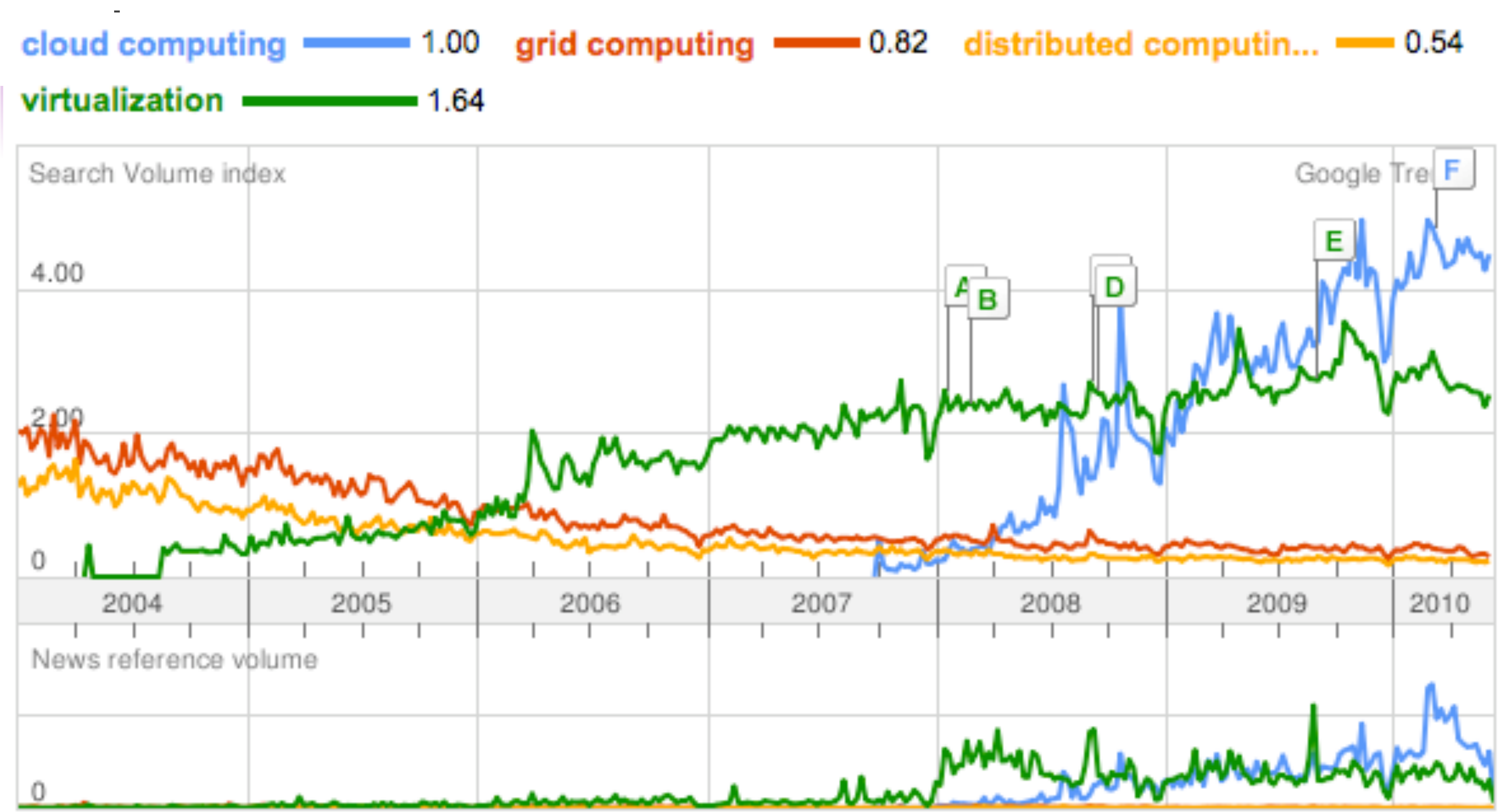
Force.com

SalesForce.com

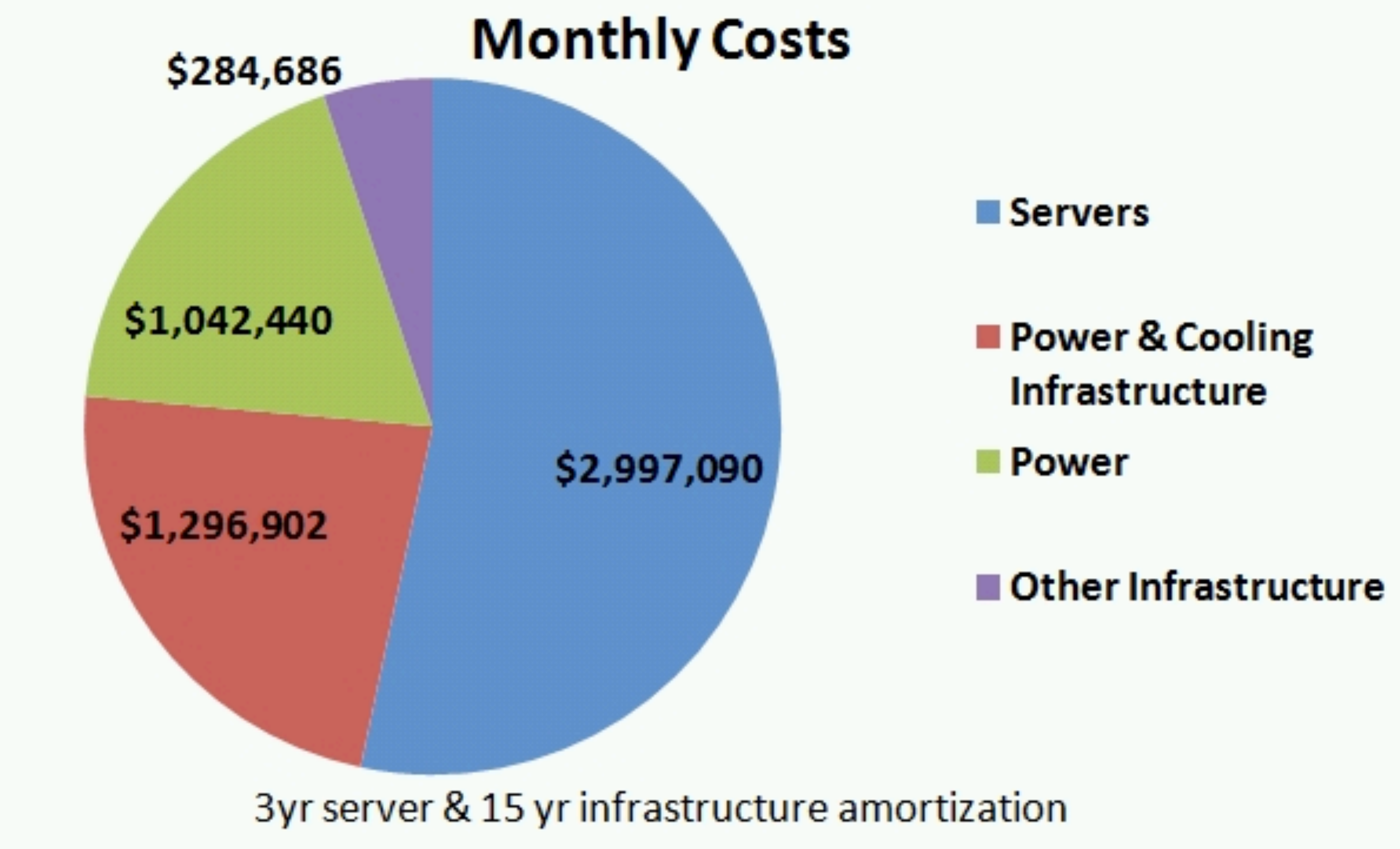
S3

Elastic MapReduce

Automation

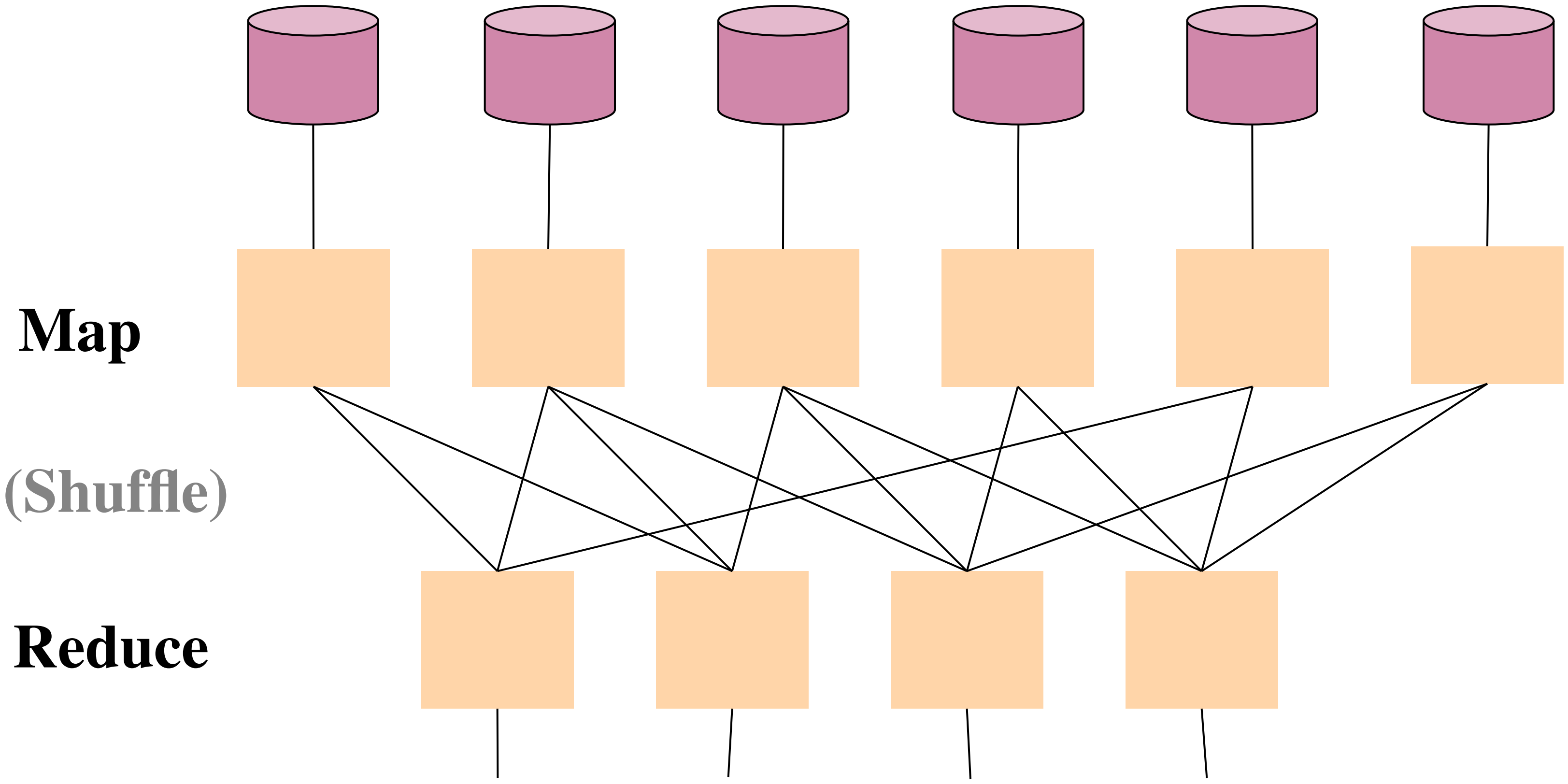

cloud computing 1.00
grid computing 0.82
distributed computin... 0.54
virtualization 1.64
Search Volume index
Google Tre
4.00
2.00
0
A
B
D
E
F
2004
2005
2006
2007
2008
2009
2010
News reference volume
0

Economies of Scale



src: James Hamilton, Amazon.com

Map Reduce


Map
(Shuffle)
Reduce