UNIVERSITY of WASHINGTON

Enabling Collaborative Research Data Management with SQLShare

Bill Howe, PhD

Director of Research,
Scalable Data Analytics

University of Washington
eScience Institute


eScience Institute

http://escience.washington.edu

The University of Washington eScience Institute

- Rationale
  - The exponential increase in sensors is transitioning all fields of science and engineering from data-poor to data-rich
  - Techniques and technologies include
    - Sensors and sensor networks, databases, 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 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

eScience Big Data Group

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

Staff

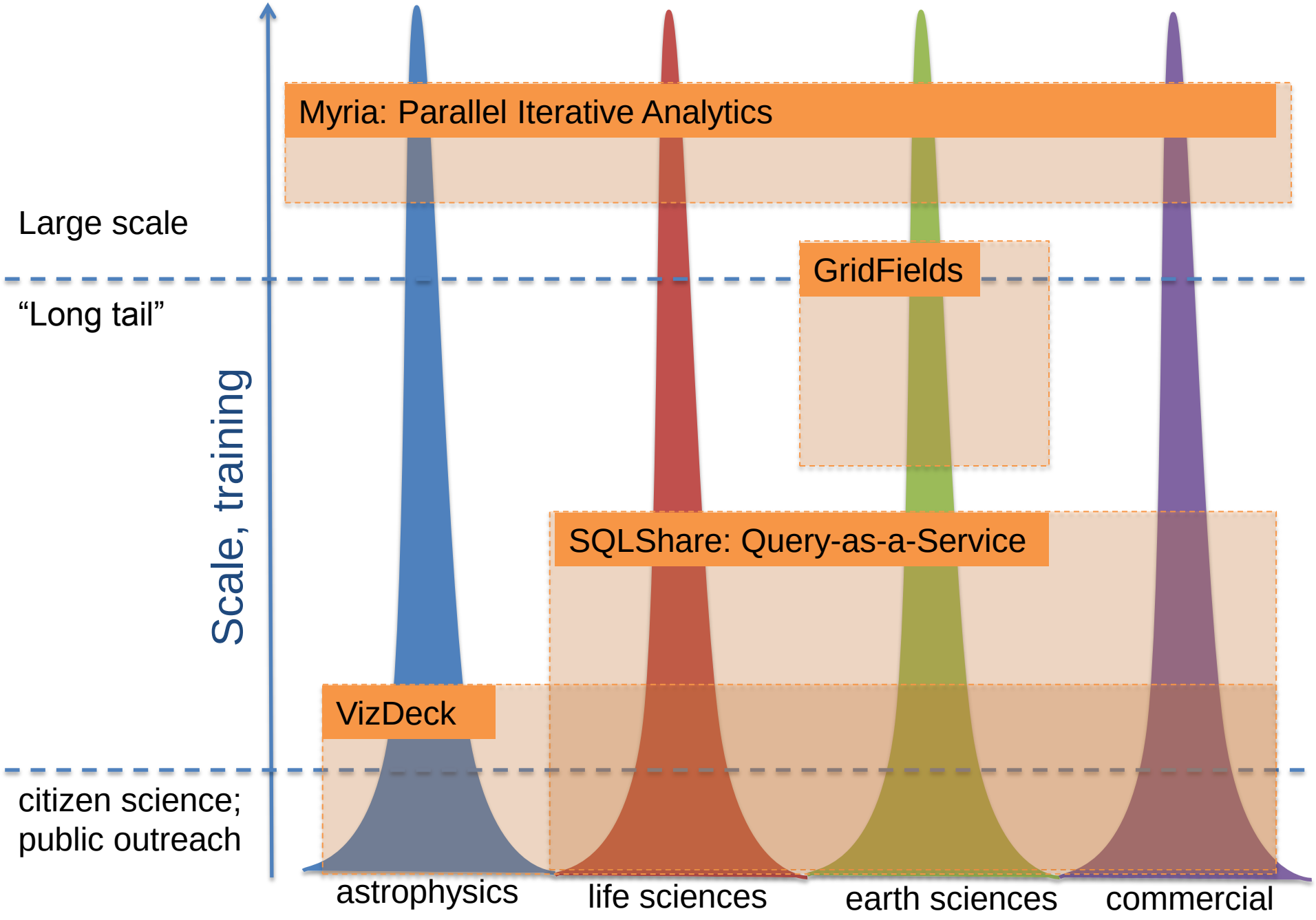
- Dan Halperin, Phd (postdoc; scalable systems)
- Seung-Hee Bae, Phd (postdoc, scalable machine learning algorithms)
- Sagar Chitnis, Research Engineer (Azure, databases, web services)
- (alumna) Marianne Shaw, Phd (hadoop, semantic graph databases)
- (alumna) Alicia Key, Research Engineer (visualization, web applications)

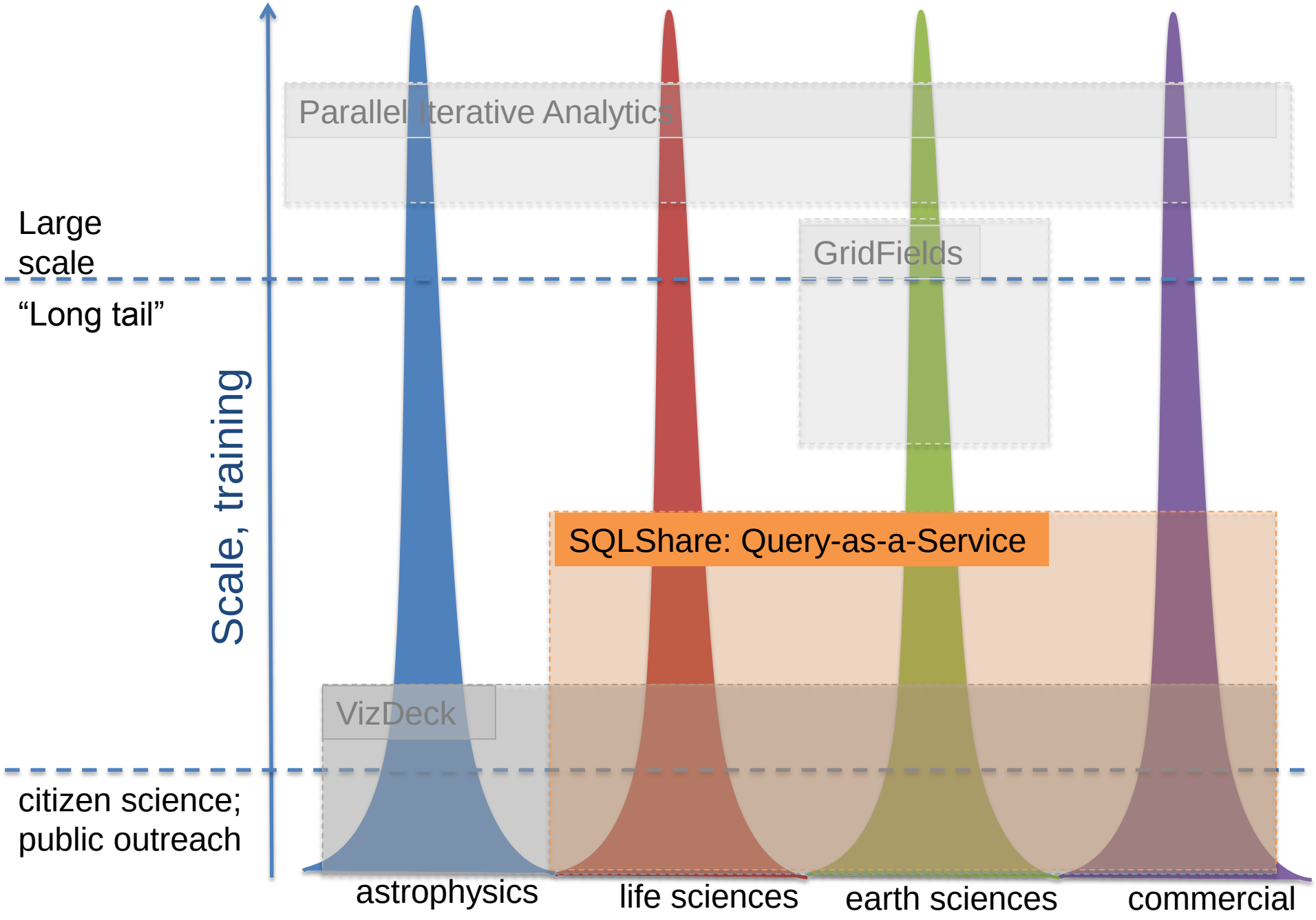
Students

- Scott Moe (2nd yr Phd, Applied Math)
- Daniel Perry (2nd yr Phd, HCDE)

Partners

- CSE DB Faculty: Magda Balazinska, Dan Suciu
- CSE students: Paris Koutris, Prasang Upadhyaya,
- UW-IT (web applications, QA/support)
- Cecilia Aragon, Phd, Associate Professor, HCDE (visualization, scientific applications)


Myria: Parallel Iterative Analytics
Large scale
GridFields
"Long tail"
Scale, training
SQLShare: Query-as-a-Service
VizDeck
citizen science;
public outreach
astrophysics
life sciences
earth sciences
commercial

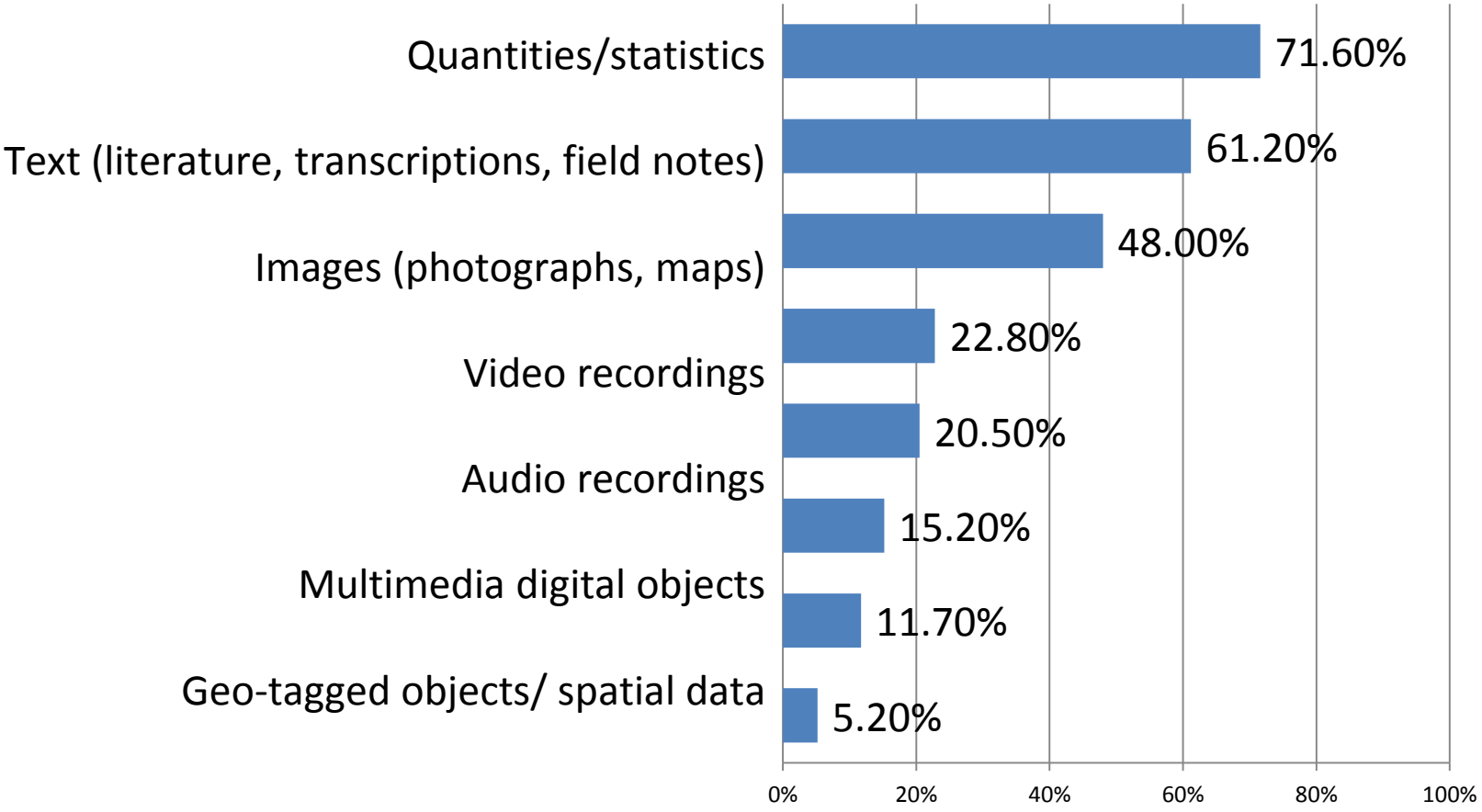

Parallel Iterative Analytics
GridFields
SQLShare: Query-as-a-Service
VizDeck
Large scale
"Long tail"
citizen science; public outreach
Scale, training
astrophysics
life sciences
earth sciences
commercial

Assessing UW Researchers' Data Management Needs

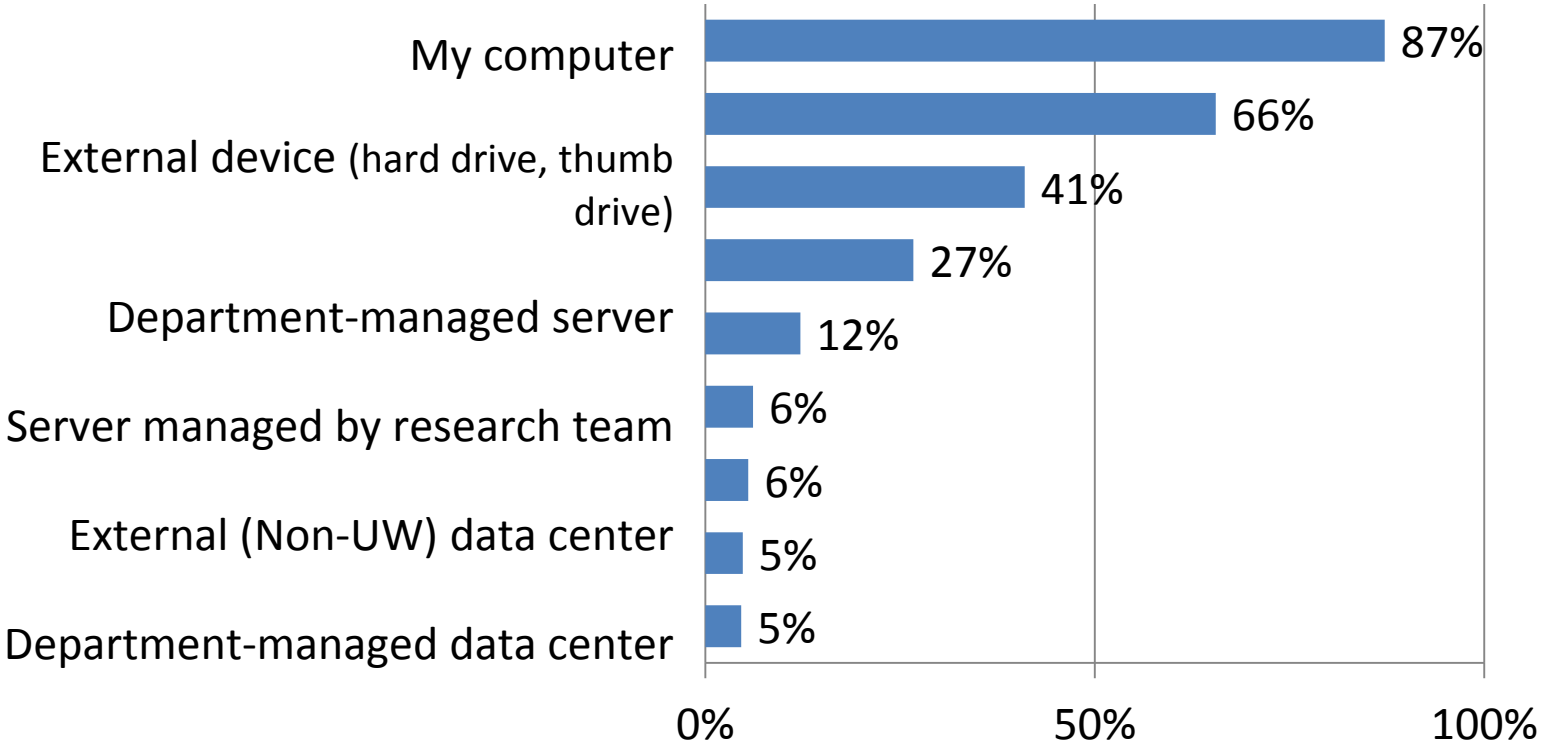
1. Conversations with Research Leaders (2008)
   - First large-scale assessment of researchers' needs
   - 124 Interviews with top researchers

2. Faculty Technology Survey (2011)
   - Use of teaching and research technologies
   - Paired with student and TA surveys
   - Reached all disciplines, levels of research
   - 689 instructors responded

Types of Data Stored



Where do you store your data?

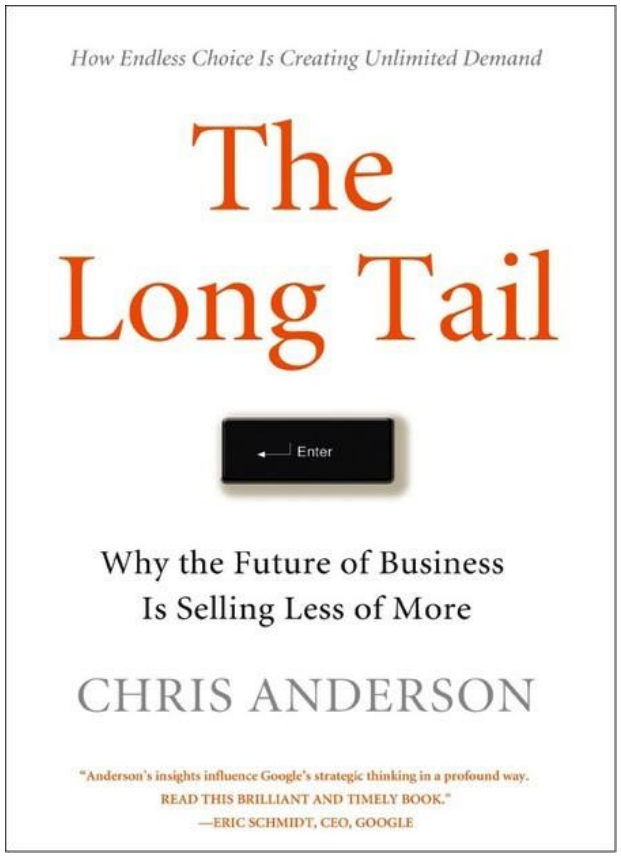


src: Conversations with Research Leaders (2008)

src: Faculty Technology Survey (2011)

- Power law
- 80:20 rule

[src: Carol Goble]



First published May 2007, Wired Magazine article 2004

Long Tail of Research Data

[src: Carol Goble]

GenBank

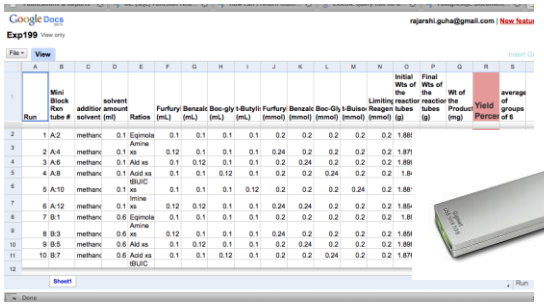
PDB

UniProt

Pfam

ChemSpider

CATH, SCOP (Protein Structure Classification)



Spreadsheets, Notebooks
Local, Lost

Problem

How much time do you spend "handling data" as opposed to "doing science"?

Mode answer: "90%"

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 protein
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 protein
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						

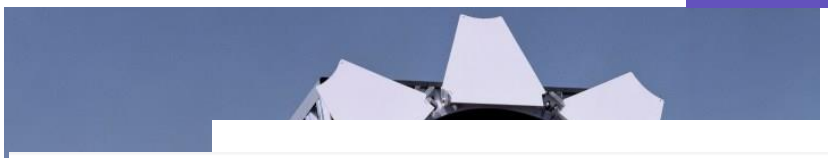
COGAnnotation_coastal_sample.txt

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
	...									

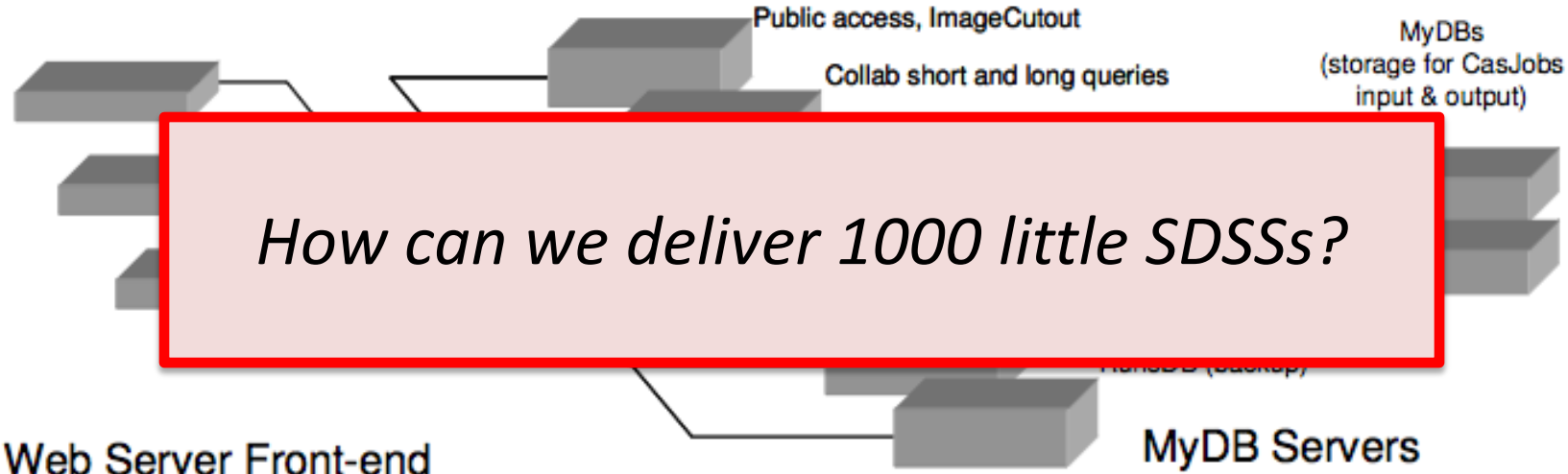
SELECT * FROM Phaeo_genome p, coastal_sample c WHERE p.COG_hit = c.hit

Why not build a database?

- Not a perfect fit – it's hard to design a "permanent" database for a fast-moving research target
  - A schema/ontology/standard is some shared consensus about a model of the world
  - Does not exist at the frontier of research, *by definition!*
- Requires specialized skills and huge up-front effort
- Researchers have little interest in operating and maintaining a data system – they just want to organize, manipulate, and share data
- *But this doesn't mean we need to punt and go back to scripts and files*



BestDR8 (2.9 TB); TargDR8 (2.6 TB)



How can we deliver 1000 little SDSSs?

Web Server Front-end

Load-balancing configuration
managing web and soap access to all data

3 servers

specs:
2U rack mount Dual Xeon2.8 GHz Server
2GB memory
(2) 250GB SATA drives (~250GB as RAID10)
MS Windows (implementing s/w load-balancing)
Est. cost: $3.6K ea. (Jun-05 pricing)

CAS Database Servers

6 servers (RAID10)

specs:
4U rack mount Dual Opteron 2.2GHz Server
4 GB memory
(24) 500GB SATA drives (~5.45 TB as RAID10)
MS Windows + SQL Server 2005
Est., cost: $13.7K ea. (Mar-06 pricing)

MyDB Servers

2 servers (primary & mirror)
Storage for 630 users @ 500 MB each; expandable to 1 GB

specs:
2U rack mount Dual Opteron Server
2.2GHz; 4 GB memory
(4) 500GB SATA drives
(~1.28 TB as RAID5)
MS Windows + SQL Server 2005
Est., cost: $3.6 ea.

NoSchema (not NoSQL)

- A schema* is a shared consensus about some universe of discourse
- At the frontier of research this does not exist, *by definition*
- Any schema that does emerge will change frequently, *by definition*
- Data “from the wild” will not conform to your schema, *by definition*

The database community needs to modularize its contributions

* ontology/metadata standard/controlled vocabulary/etc.

Digression: Relational Database History

Pre-Relational: if your data changed, your application broke.

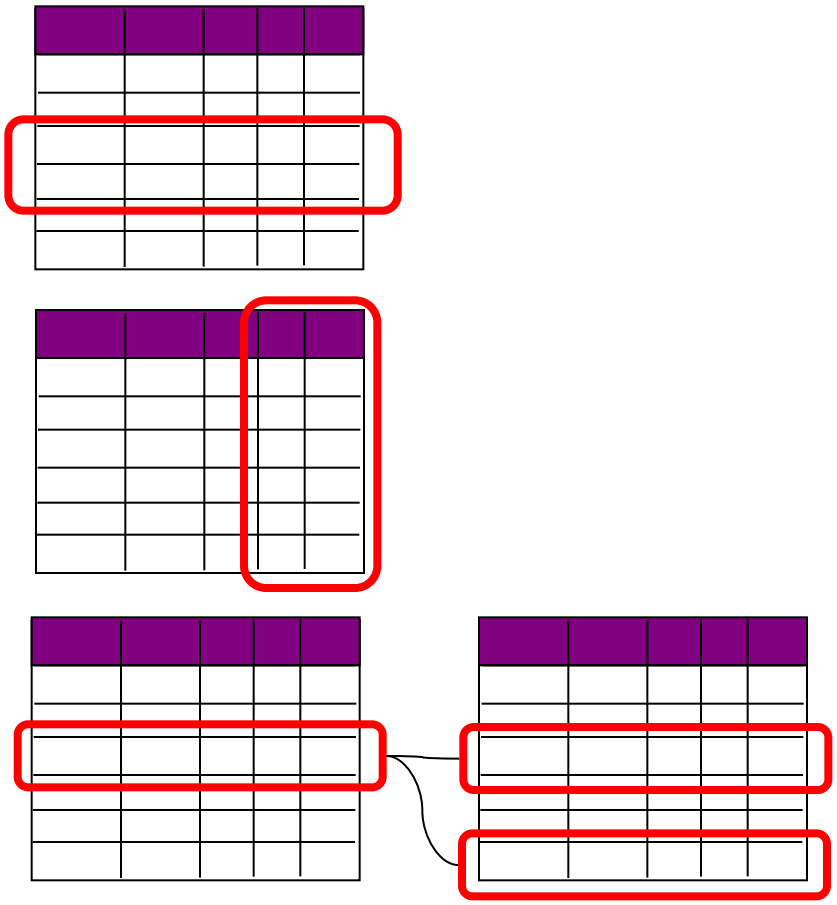
Early RDBMS were buggy and slow (and often reviled), but required only 5% of the application code.

"Activities of users at terminals and most application programs should remain unaffected when the internal representation of data is changed and even when some aspects of the external representation are changed."

-- Codd 1979

Key Ideas: Programs that manipulate tabular data exhibit an algebraic structure allowing reasoning and manipulation independently of physical data representation

Key Idea: An *Algebra of Tables*



select

project

join

Other operators: aggregate, union, difference, cross product

Algebraic Optimization

N = ((4*2)+((4*3)+0))/1

Algebraic Laws:

1. (+) identity: x+0 = x
2. (/) identity: x/1 = x
3. (*) distributes: (n*x+n*y) = n*(x+y)
4. (*) commutes: x*y = y*x

Apply rules 1, 3, 4, 2: N = (2+3)*4

two operations instead of five, no division operator

Same idea works with very large tables / graphs, but the payoff is much higher

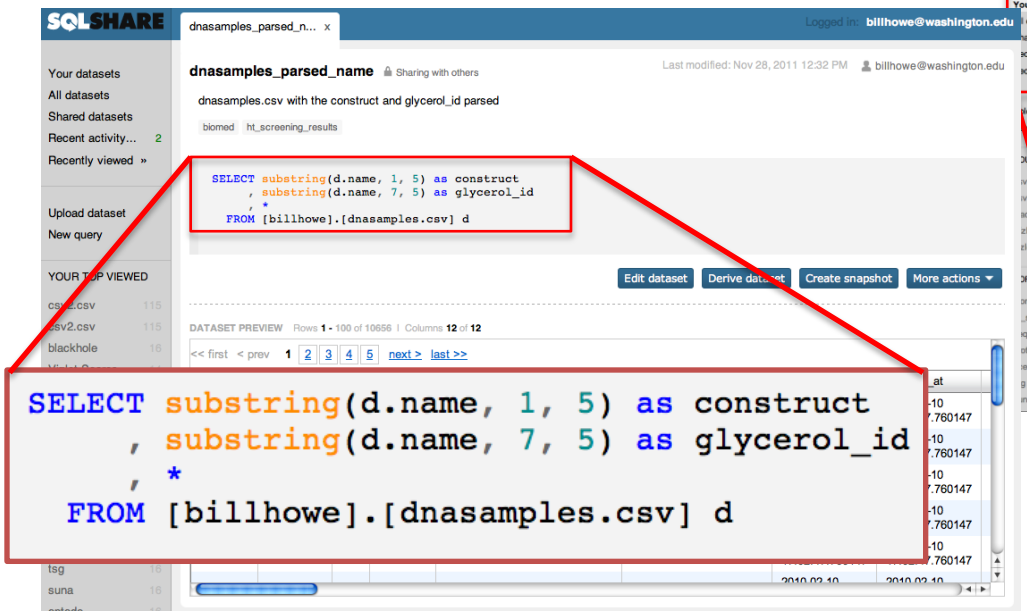
~~Database~~ Query-as-a-Service for the 99%

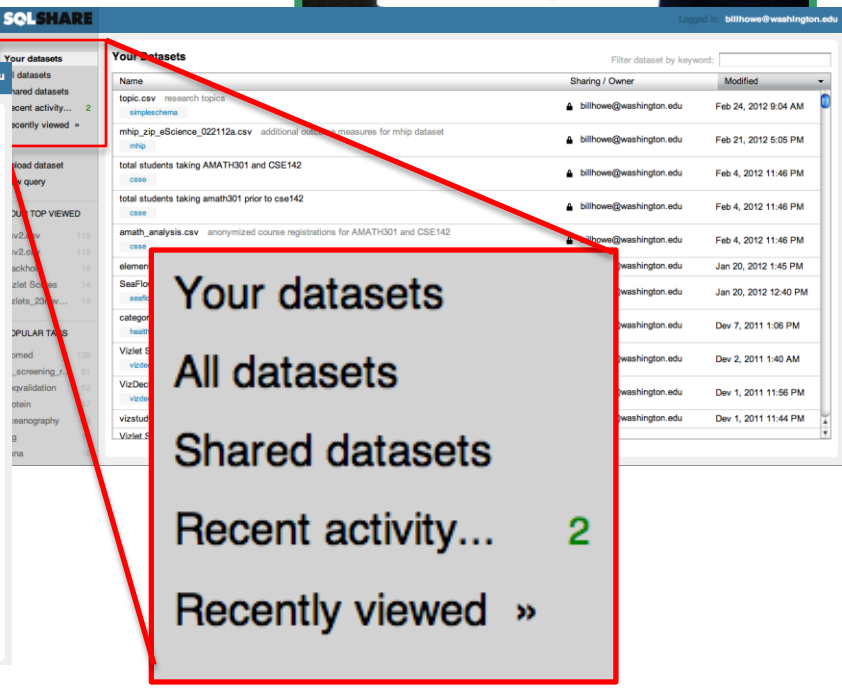
Approach: Strip down databases to bare essentials

- Upload -> Query -> Share
- Try to eliminate installation, configuration, schema design, data loading, tuning, app-building

easy, thanks to the cloud

harder, requires some research





3) Share the results

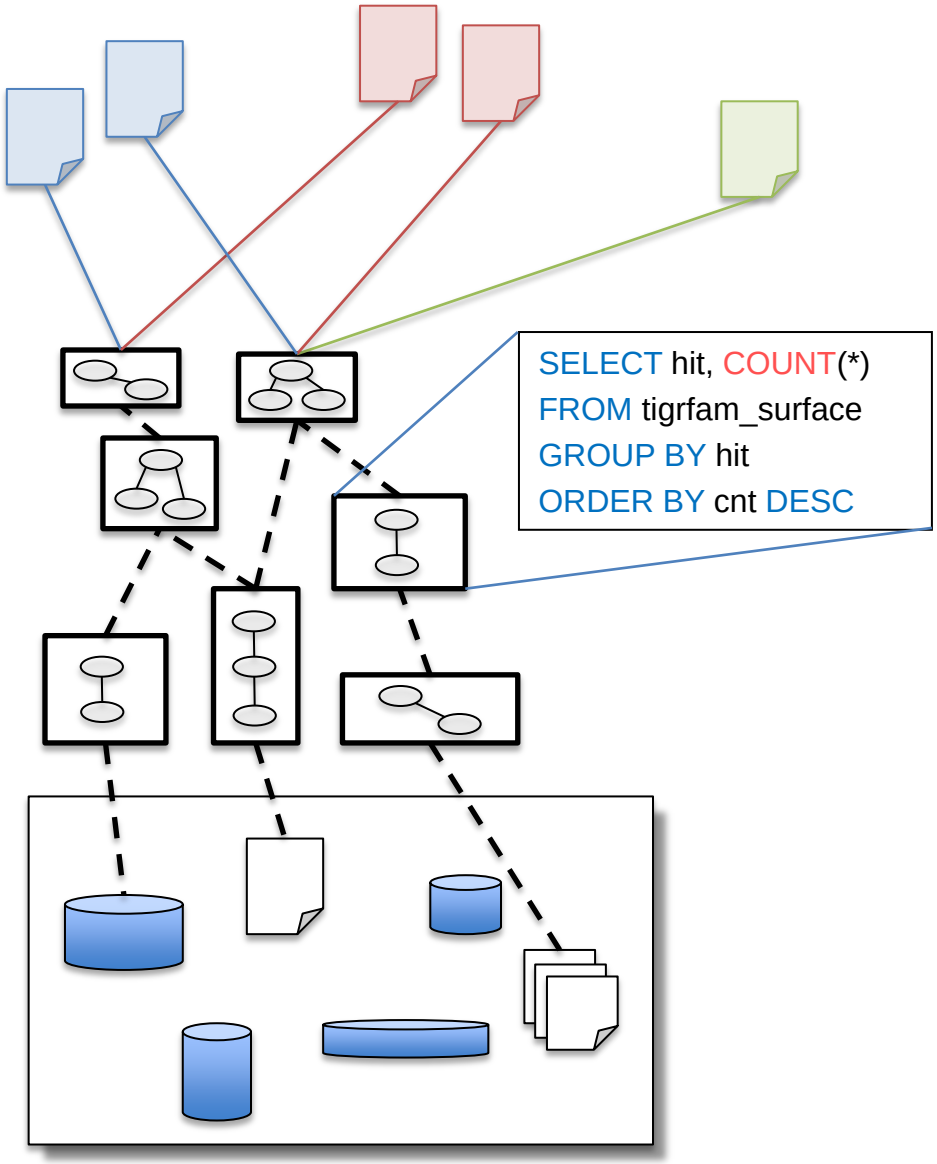
Click on the science question, see the SQL that answers it

2) Analyze data with SQL

Right in your browser, writing queries on top of queries on top of queries ...

1) Upload data "as is"

Cloud-hosted; no need to install or design a database; no pre-defined schema



Scientific data management reduces to sharing views

- Integrate data from multiple sources?
  - *joins and unions with views*
- Standardize on units, apply naming conventions?
  - rename columns, apply functions with views
- Attach metadata?
  - add new tables with descriptive names, add new columns with views
- Data cleaning, quality control?
  - hide bad values with views
- Maintain provenance?
  - inspect view dependencies
- Propagate updates?
  - view maintenance
- Protect sensitive data?
  - expose subsets with views (assuming views carry permissions)

Select from a list of English descriptions

Edit a Query

```
SELECT interpro_description, count(DISTINCT gene_id) as num_genes
FROM [koesterj@washington.edu].[809_interpro2GO0005515_sort.csv]
GROUP BY interpro_description
order by num_genes DESC
```

Your datasets
All datasets
Shared datasets
Recent activity... 0
Recently viewed »

Upload dataset
New query

YOUR TOP VIEWED
exon.feature.o... 36
Lengths_hotsp... 26
noncoding_po... 25
hotspot_desert... 17
HD_NC_overla... 16

POPULAR TAGS
protein 49
oceanography 23
fluorimeter 16
suna 16

GO0005515_domains Only you can view this

Last modified: Apr 21, 2011 2:34 PM koesterj@washington.edu

How many genes have which domain

Click here to add a tag

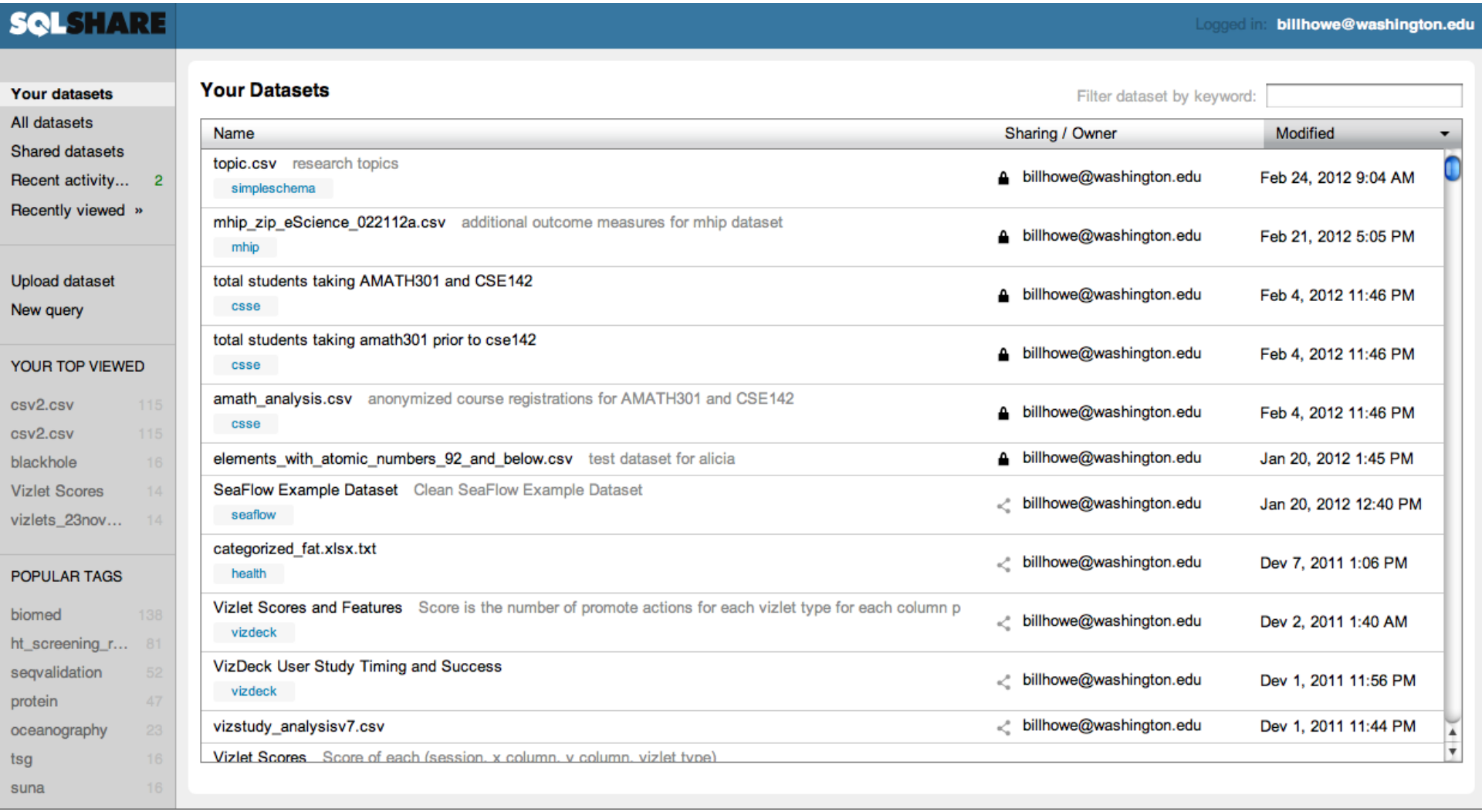
```
SELECT interpro_descripti
FROM [koesterj@washington
GROUP BY interpro_descrip
order by num_genes DESC
```

DATASET PREVIEW Rows 1 - 100 of 148 | Columns 2 of 2

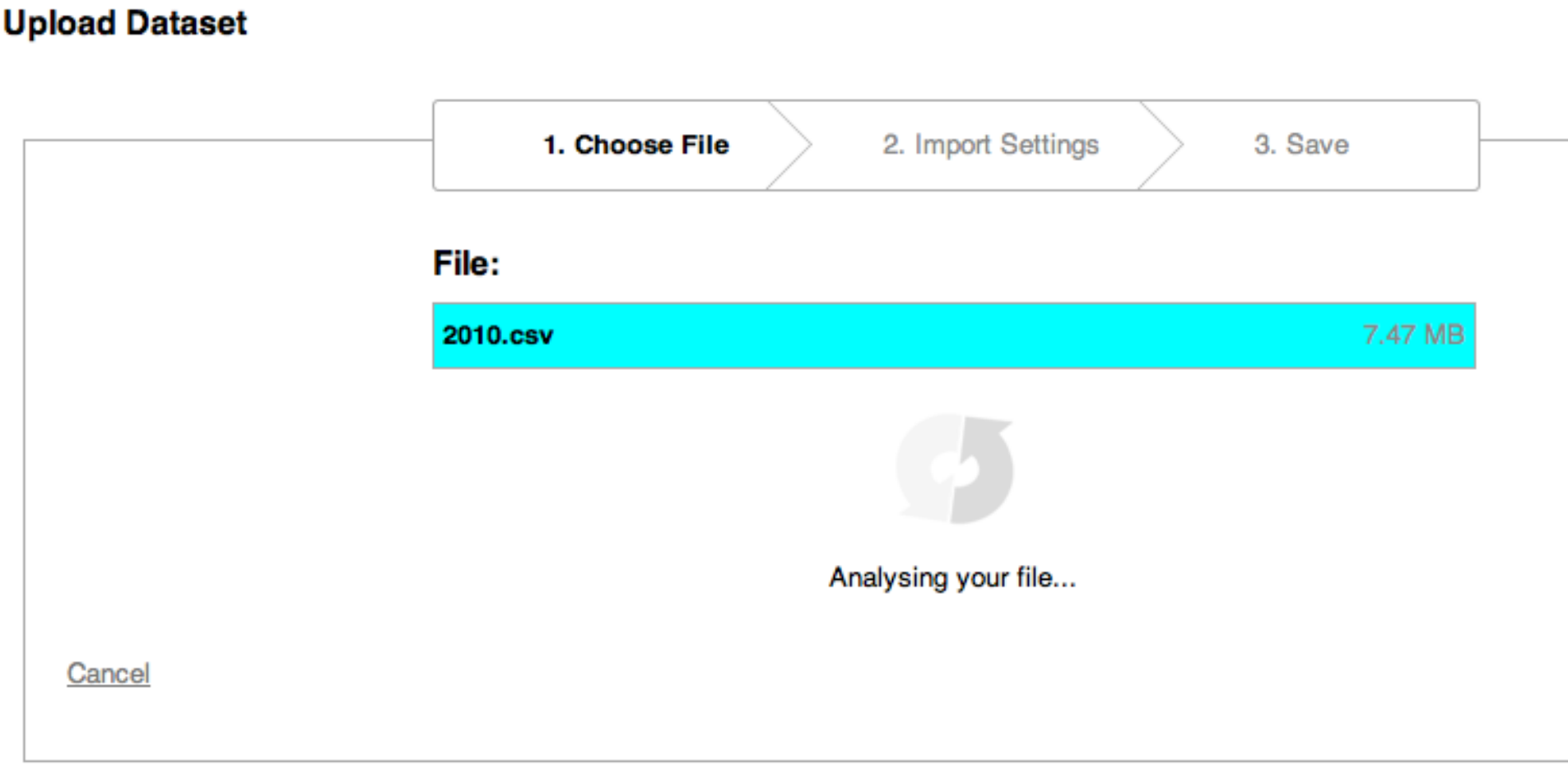
<< first < prev 1 2 3 4 5 next > last >>

interpro_description	num_genes
NULL	112
WD40/YVTN repeat-like-containing domain	18
WD40 repeat-like-containing domain	14
Zinc finger RING/FYVE/PHD-type	13
WD40 repeat subgroup	13
Tetratricopeptide-like helical	13
WD40 repeat	12
Tetratricopeptide repeat	11
Tetratricopeptide repeat-containing	9
WD40-repeat-containing domain	9

Save the results, share them with others



http://sqlshare.escience.washington.edu



http://sqlshare.escience.washington.edu

Upload Dataset

1. Choose File | **2. Import Settings** | 3. Save

Dataset was imported with the following settings:

You can change the parser options if your data was not properly imported.

☑ Contains column header | Values are separated by ,

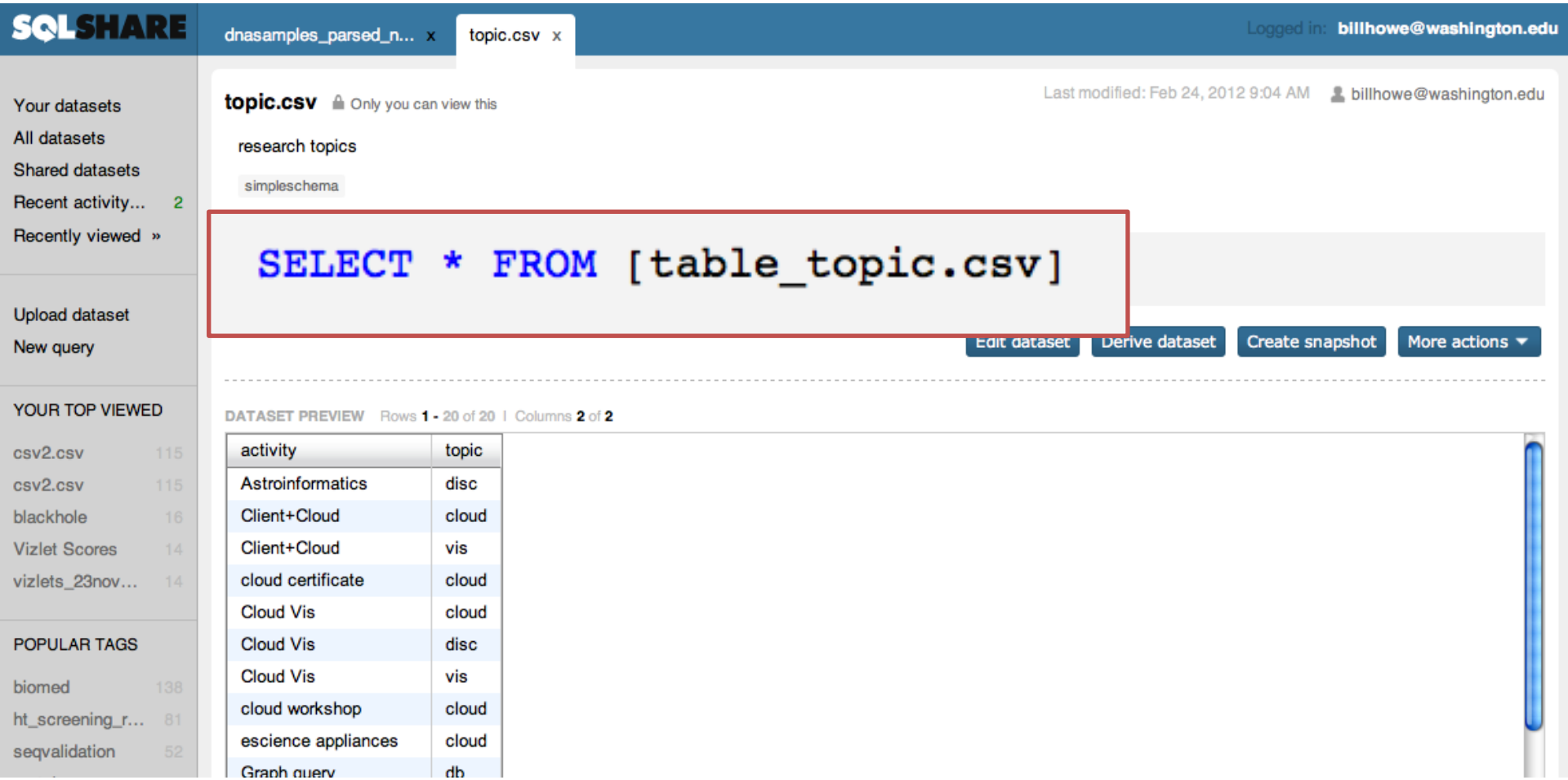
DATASET PREVIEW (Imported table with **3 columns**)

activity	thrust	time in past 12 months
SQLShare Engineering	long-tail	1
SQLShare research	long-tail	1
Client+Cloud	long-tail	1
HaLoop	scalable analytics	1.5
Cloud Vis	scalable analytics	1

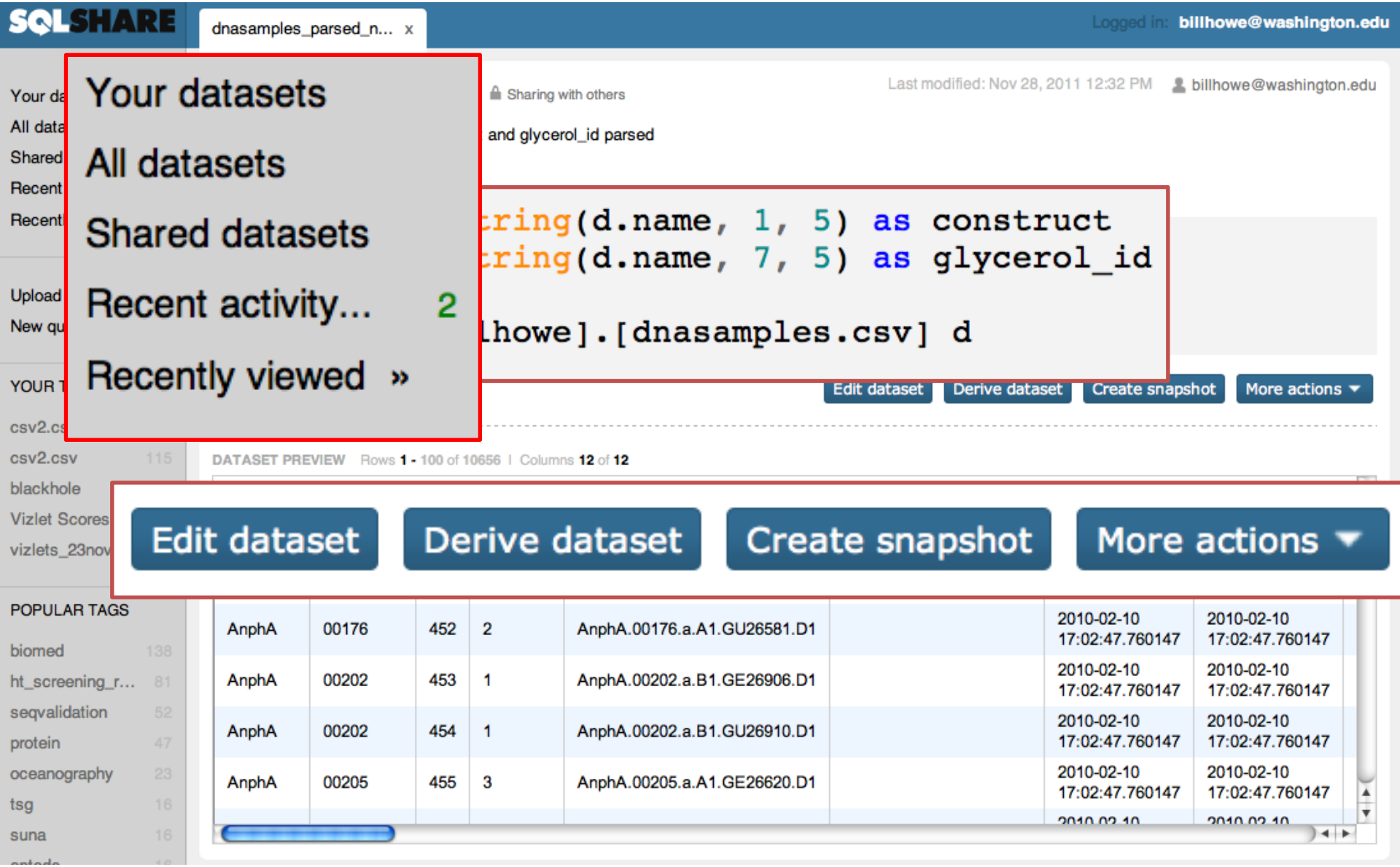
Cancel

Back Next

http://sqlshare.escience.washington.edu

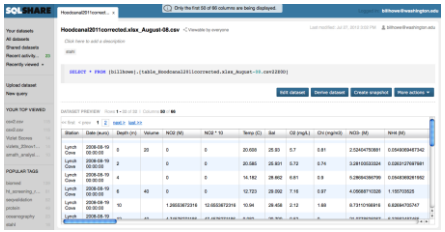


http://sqlshare.escience.washington.edu



http://sqlshare.escience.washington.edu

amazon web services™ | EC2



VizDeck

"Flagship" SQLShare App (Python) on EC2

Spreadsheet Crawler

Excel Addin

Python Client

R Client

ASP.NET

SQLShare REST API

OAuth2

WCF

Windows Azure™

Microsoft® SQL Azure™

USAGE

Environmental Sampling

Sequencing

Sequence data

Questions?

correlate diversity w/environment?

correlate diversity w/nutrients?

find new taxa and their distributions?

find new genes?

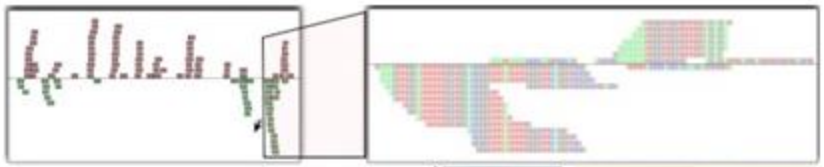
compare meta'omes?

search hits

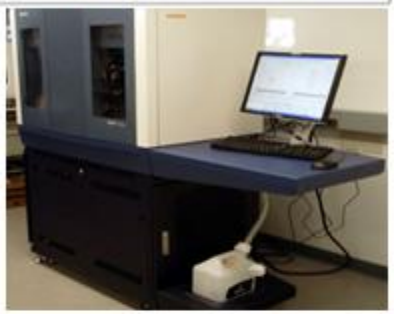
metadata

Public sequence DBs

Environmental Sampling



Sequencing



Sequence data

search hits

metadata

Pfams, TIGRfams, COGs, FIGfams

Public annotation DBs

Questions?

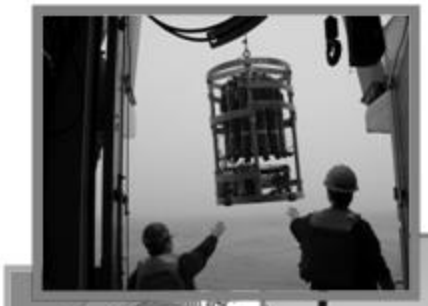
correlate diversity w/environment?

correlate diversity w/nutrients?

find new taxa and their distributions?

find new genes?

compare meta'omes?



Environmental Sampling

Sequencing

Sequence data

search lists

metadata

Pfams, TIGRfams,
COGs, FIGfams

Public annotation DBs

Questions?

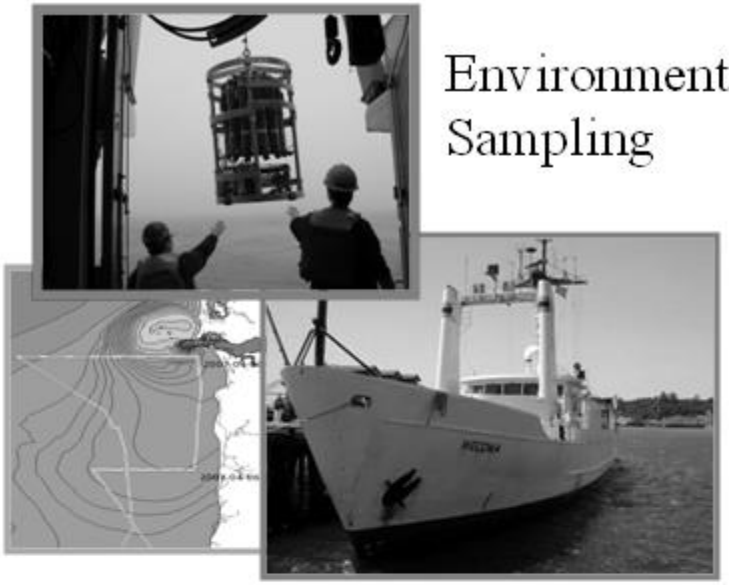
correlate diversity w/environment?

correlate diversity w/nutrients?

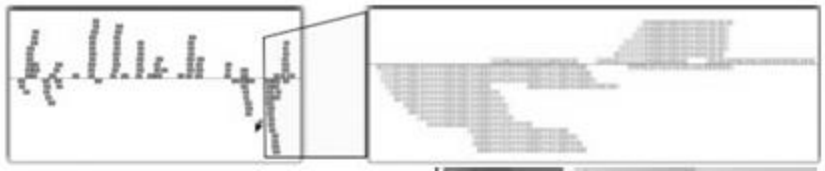
find new taxa and their distributions?

find new genes?

compare meta'omes?



Environmental Sampling



Sequencing

metadata

sequence data

search results

Pfams, TIGRfams, COGs, FIGfams

Public annotation DBs

Questions?

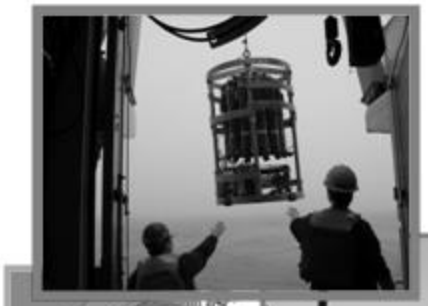
correlate diversity w/environment?

correlate diversity w/nutrients?

find new taxa and their distributions?

find new genes?

compare meta*omes?



Environmental Sampling

Sequencing

Sequence data

search hits

metadata

SQLShare

Pfams, TIGRfams, COGs, FIGfams

Public annotation DBs

SQL

correlate diversity w/environment?

correlate diversity w/nutrients?

find new taxa and their distributions?

find new genes?

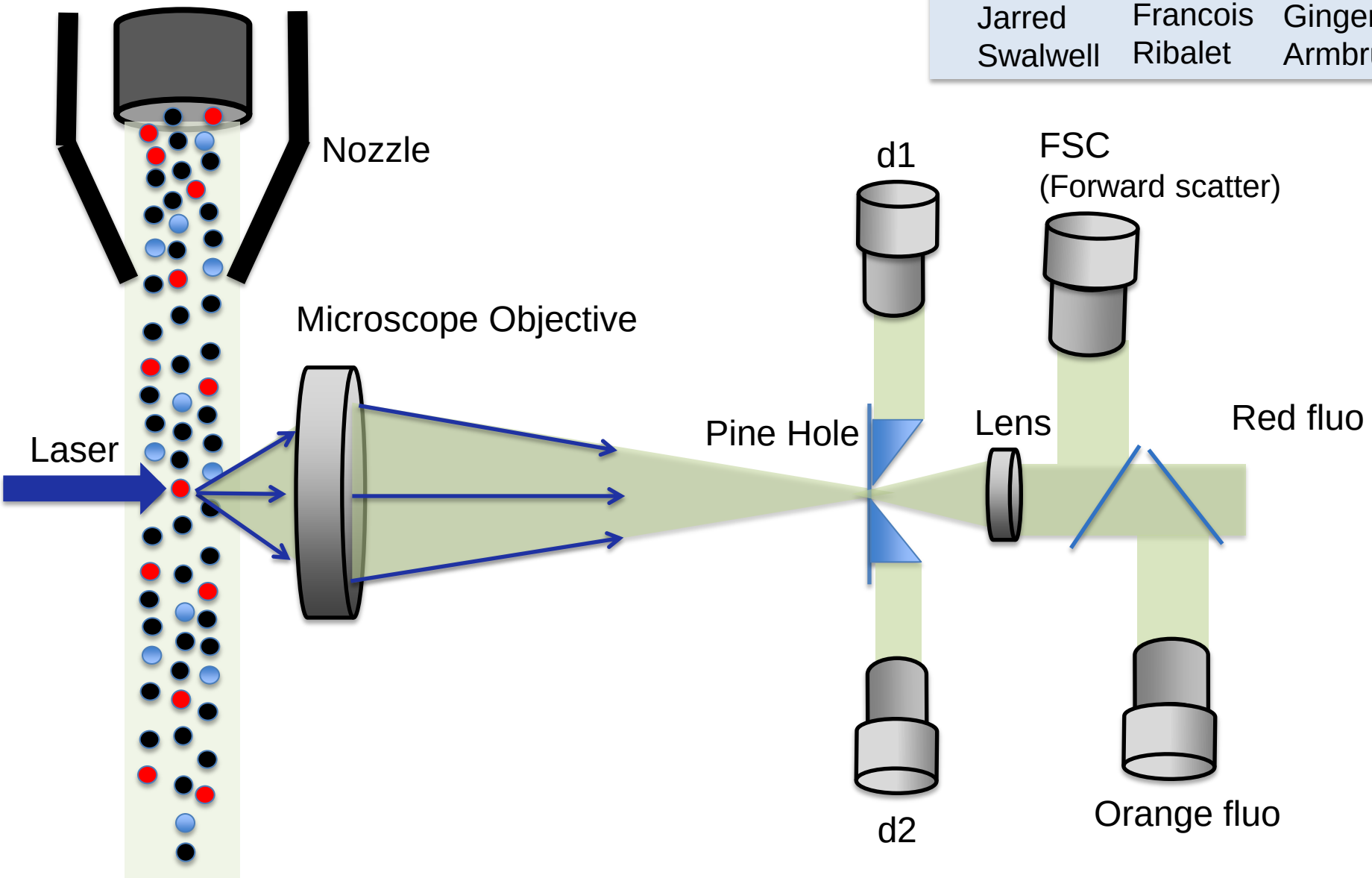
compare meta*omes?

SeaFlow


Jarred Swalwell


Francois Ribalet


Ginger Armbrust

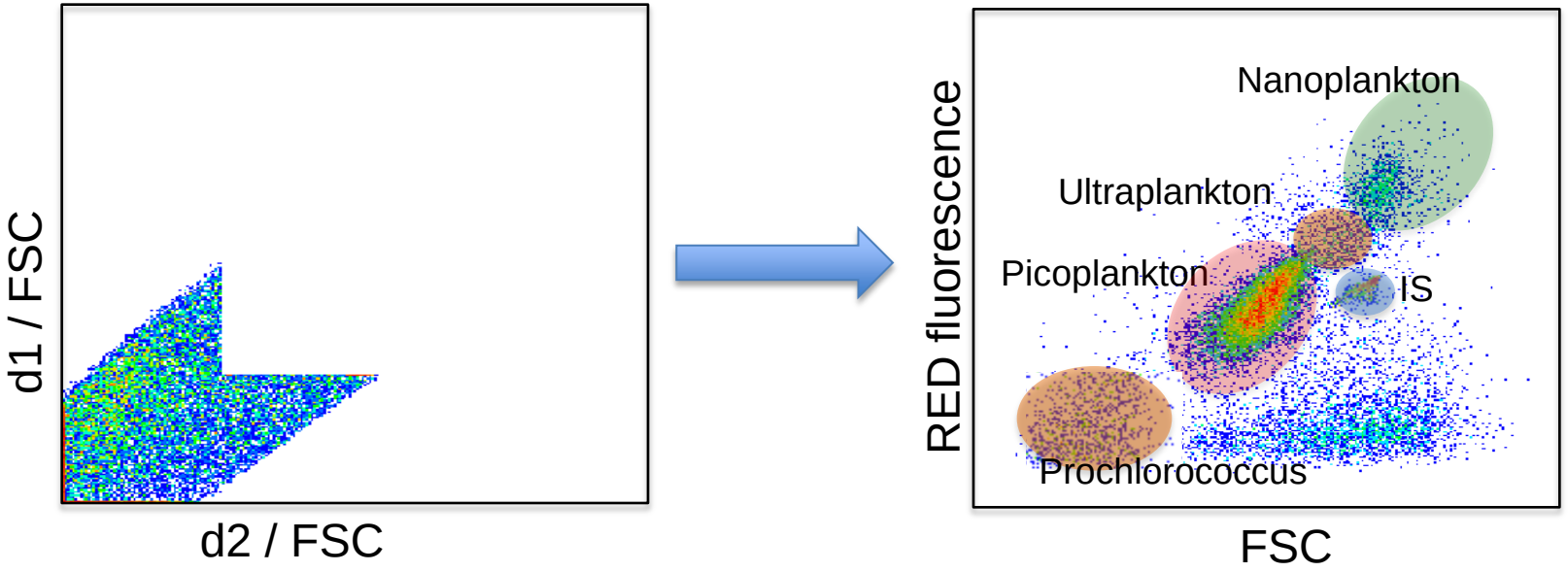


SeaFlow


Jarred Swalwell


Francois Ribalet


Ginger Armbrust



- Continuous observations of various phytoplankton groups from 1-20 μm in size
  - Based on RED fluo: *Prochlorococcus*, Pico-, Ultra- and Nanoplankton
  - Based on ORANGE fluo: *Synechococcus,* Cryptophytes
  - Based on FSC: Coccolithophores

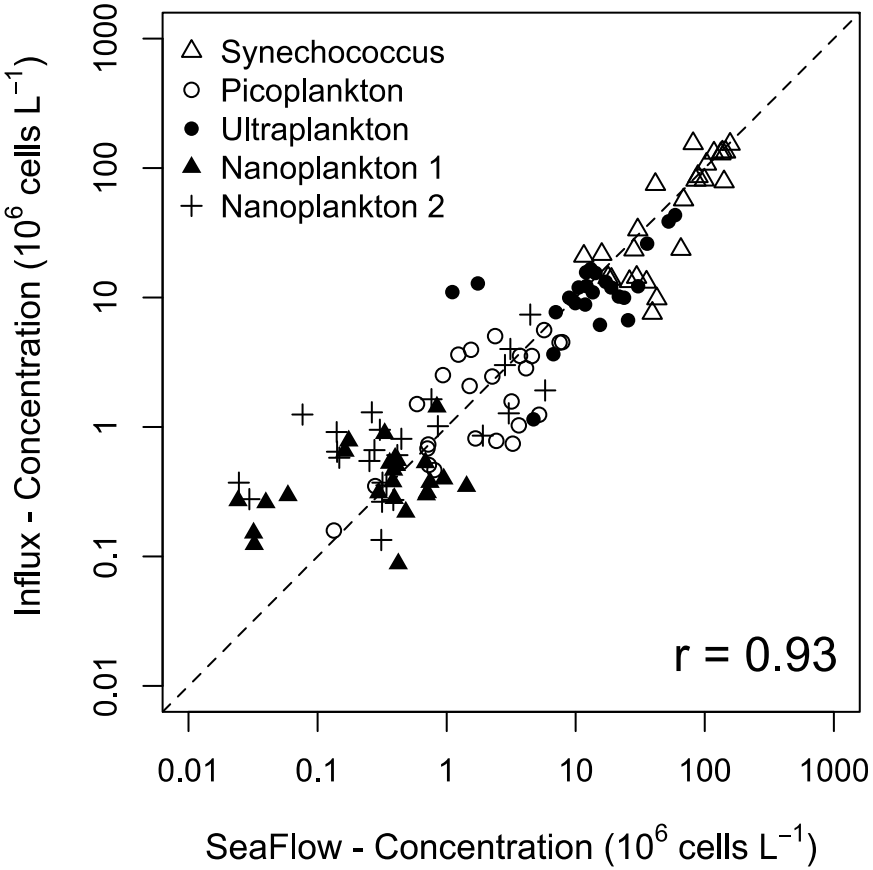
UNIVERSITY of WASHINGTON

SeaFlow


Jarred Swalwell


Francois Ribalet


Ginger Armbrust

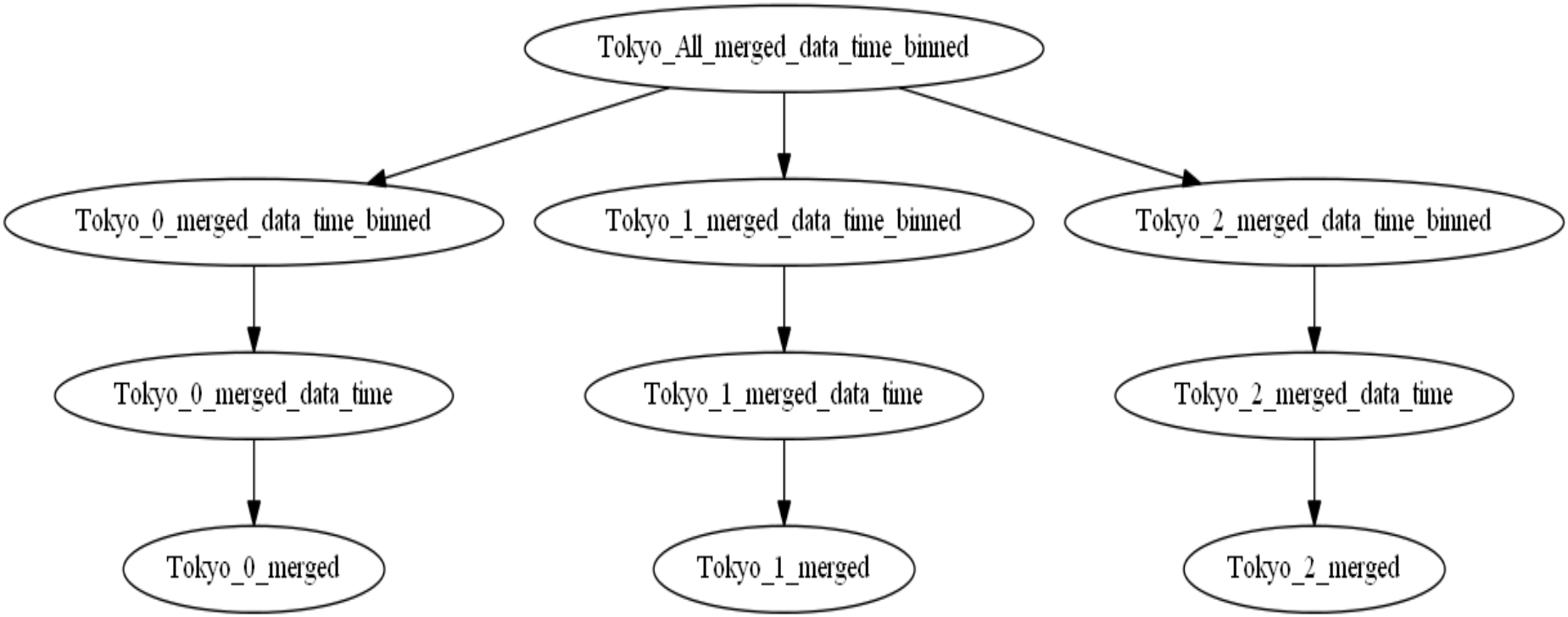


Distributed Collaboration, Status Quo

- Scripts (typically in R) must be pre-shared with all collaborators
- When the data changes, everybody has to re-run all the scripts
- When the scripts change, everybody has to re-run all the scripts.
- Implicit assumption that all data fits in main memory
- No provenance.
- Pipeline of scripts dependent on intricate file formats and file naming schemes

[redacted]

- [redacted]tion that all data fits in main memory
- [redacted]
- Pipeline of scripts dependent on intricate file formats and/or file naming schemes



Deeply nested hierarchies of views

Provenance

Controlled Sharing

Implicit re-execution

Data management and statistics for biologists

Robin Kodner

Fall 2012

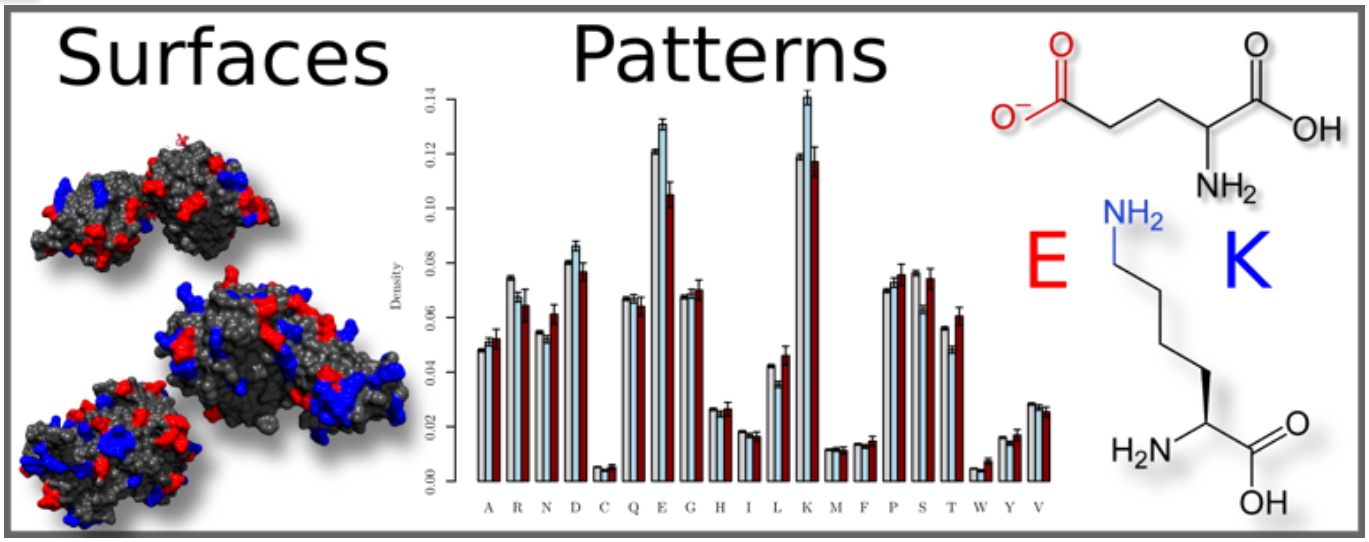
- Home
- Announcements
- Assignments
- Discussions
- Grades
- People
- Files
- Syllabus
- Modules
- Conferences

Course Modules

Data and Summary Statistics

- Introduction to Bio 340
- minute paper 1 — Sep 28 — 3 pts
- Data (Sept 28)
- minute paper 2 — Oct 1 — 3 pts
- ct 1
- t 3 — 5 pts
- t 8 — 5 pts
- Describing data and Summary Statistics (Oct 5)
- Calculating Spread and plotting Multiple Variables (Oct 8)
- lecture 8 example problems for homework — Oct 10 — 1 pts
- Building Comprehensive Data Sets and Intro to Databases (Oct 10)
- Homework Problem Set 1 — Oct 17 — 30 pts
- Intro to SQL (Oct 12)
- More SQL (Oct 15)
- DATA MODULE READINGS

"I have had two students who are struggling with R come up and tell me how much more they like working in SQLShare."



Andrew White,
UW Chemistry

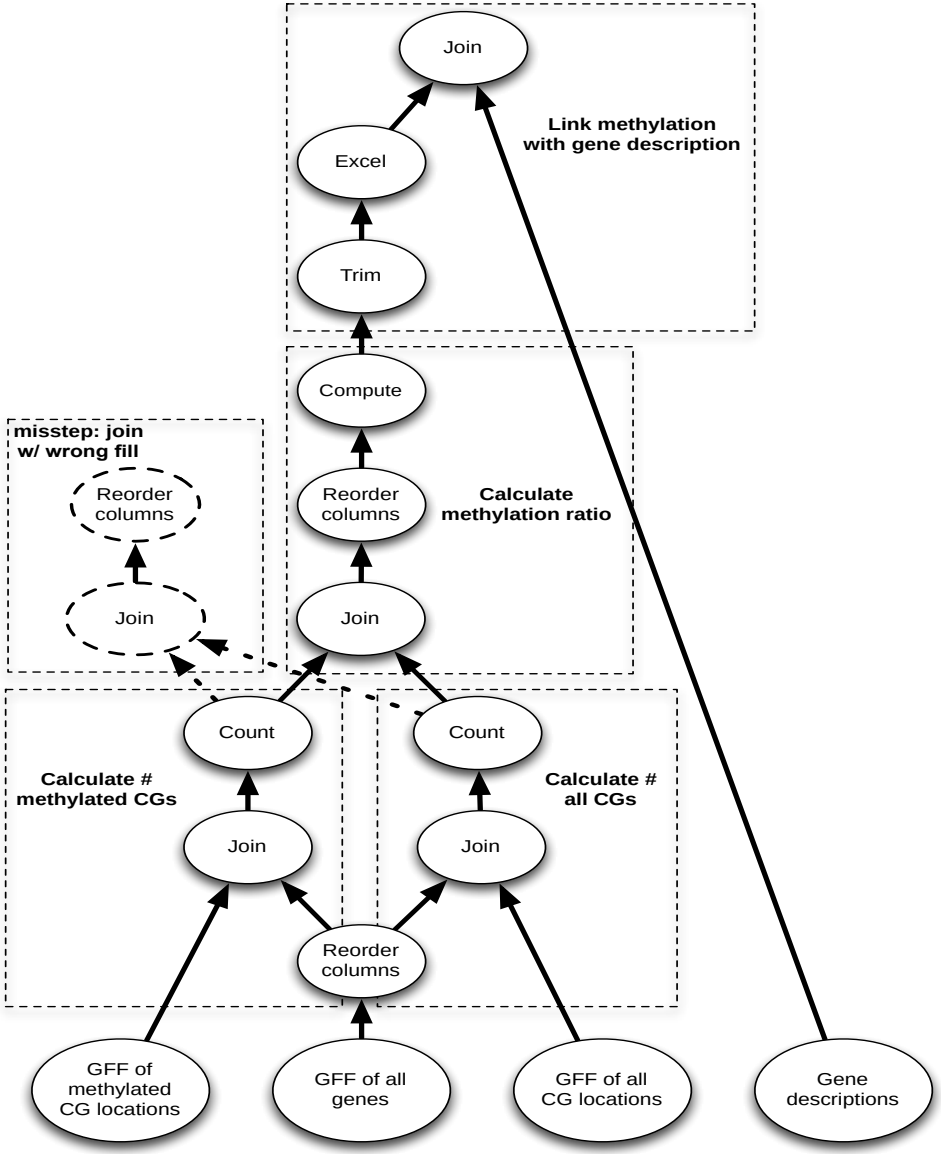
"An undergraduate student and I are working with gigabytes of tabular data derived from analysis of protein surfaces.

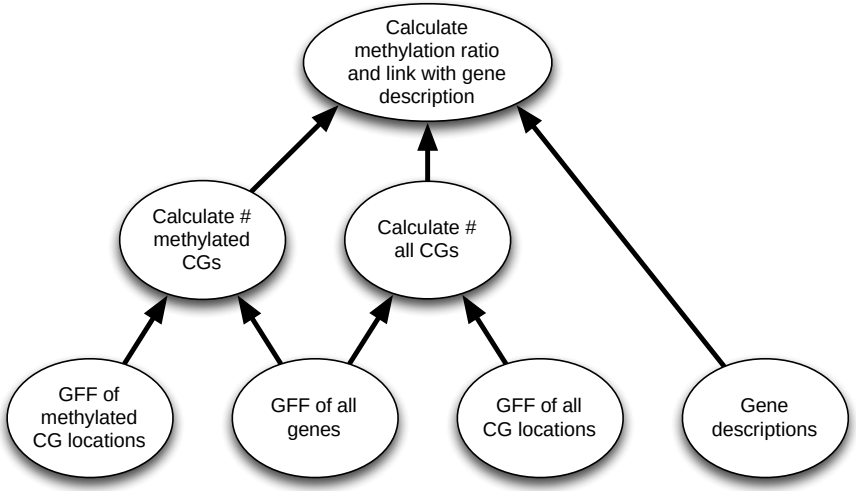
Previously, we were using huge directory trees and plain text files.

Now we can accomplish a 10 minute 100 line script in 1 line of SQL."

-- Andrew D White

Decoding nonspecific interactions from nature. A. White, A. Nowinski, W. Huang, A. Keefe, F. Sun, S. Jiang. (2012) Chemical Science. Accepted





Steven Roberts

SQL as a lab notebook:
http://bit.ly/16Xj2JP

Galaxy

Popular service for Bioinformatics Workflows

SQLSHARE

WHY SQL?

Find all TIGRFam ids (proteins) that are missing from at least one of three samples (relations)

```
SELECT col0 FROM [refseq_hma_fasta_TGIRfam_refs]
UNION
SELECT col0 FROM [est_hma_fasta_TGIRfam_refs]
UNION
SELECT col0 FROM [combo_hma_fasta_TGIRfam_refs]

EXCEPT

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

Why SQL?

- Covers 80% of what we need
  - Ex: Sloan Digital Sky Survey
  - Ex: Hybrid Hash Join algorithm published in BMC bioinformatics
- Empower a new class of data-savvy scientist who isn't forced to be trained as an IT professional
- Automatic optimization, parallelization, scalability, fault-tolerance
  - [Ask me about this if you're interested]
- Views: Logical and physical data Independence
  - Reason about the problem independently of the data representation
  - No re-execution of "workflows"
  - No file format incompatibilities
  - No version mismatches
  - Data and code tightly coupled and (logically) centralize

Ex: Interval arithmetic


```
SELECT x.strain, x.chr, x.region as snp_region, x.start_bp as snp_start_bp
    , x.end_bp as snp_end_bp, w.start_bp as nc_start_bp, w.end_bp as nc_end_bp
    , w.category as nc_category
    , CASE WHEN (x.start_bp >= w.start_bp AND x.end_bp <= w.end_bp)
  THEN x.end_bp - x.start_bp + 1
  WHEN (x.start_bp <= w.start_bp AND w.start_bp <= x.end_bp)
  THEN x.end_bp - w.start_bp + 1
  WHEN (x.start_bp <= w.end_bp AND w.end_bp <= x.end_bp)
  THEN w.end_bp - x.start_bp + 1
END  AS len_overlap

FROM [koesterj@washington.edu].[hotspots_deserts.tab] x
INNER JOIN [koesterj@washington.edu].[table_noncoding_positions.tab] w
ON x.chr = w.chr
WHERE (x.start_bp >= w.start_bp AND x.end_bp <= w.end_bp)
OR (x.start_bp <= w.start_bp AND w.start_bp <= x.end_bp)
OR (x.start_bp <= w.end_bp AND w.end_bp <= x.end_bp)
ORDER BY x.strain, x.chr ASC, x.start_bp ASC
```

We see thousands of queries written by non-programmers

"Am I allowed to do outer joins in sqlshare?"

"I am trying to use the CASE WHEN structure..."

SQLShare as a CS Research Platform

- 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, Daniel Perry, Cecilia Aragon)

SSDBM 2011
CHI 2012
SIGMOD 2012 (demo)

- Info Extraction from Spreadsheets
  - (Mike Cafarella, Dave Maier, Bill Howe, Sagar Chitnis, Abdu Alwani)
- Scalable Analytics-as-a-Service
  - (Dan Suciu, Magda Balazinska, Bill Howe)
- Optimizing Iterative Queries for Machine Learning
  - (Dan Suciu, Magda Balazinska, Bill Howe)

VLDB 2010
Datalog2.0 2012
CIDR 2013

- Case Studies in Metagenomics, Chemistry, more

Data engineering 2012
CiSE 2012

Microsoft Research

eScience Institute

NSF

Gordon and Betty MOORE FOUNDATION

SQLSHARE

http://sqlshare.escience.washington.edu

billhowe@cs.washington.edu

http://escience.washington.edu

Microsoft Research

eScience Institute

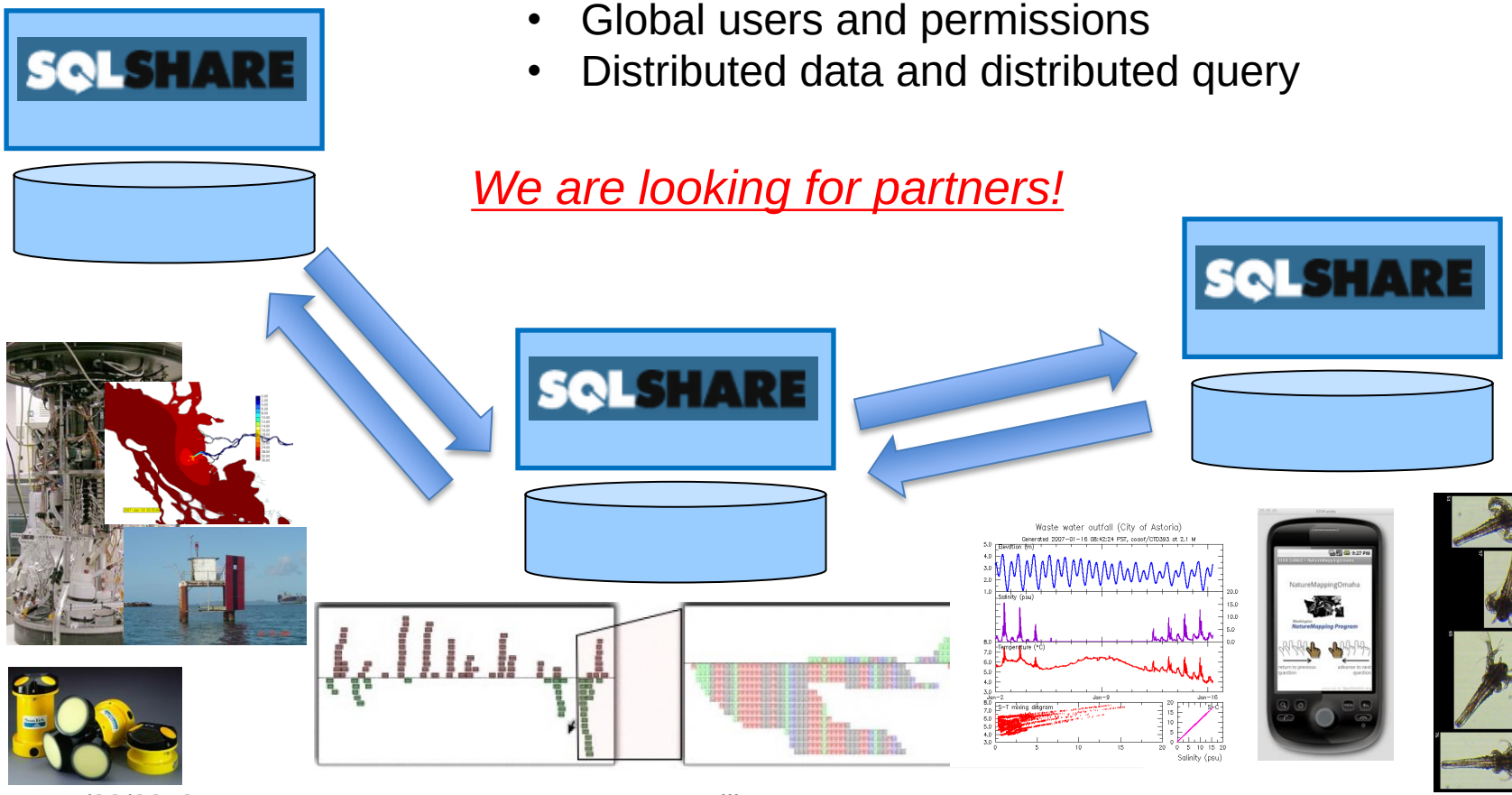
NSF

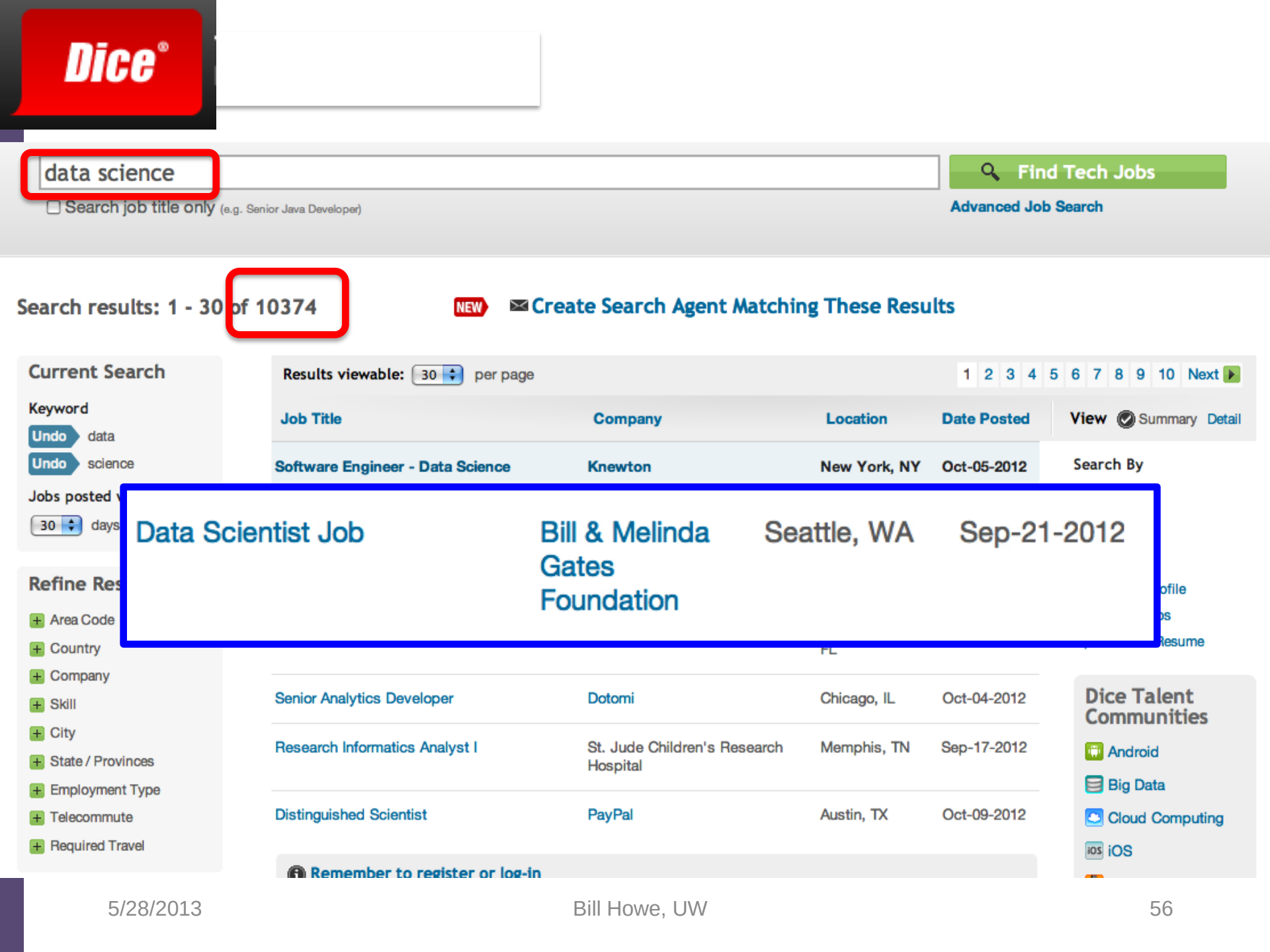
Gordon and Betty MOORE FOUNDATION

Where we're headed:

- *Local or cloud-hosted deployments* ***done!***
- Multi-institution sharing
- Global users and permissions
- Distributed data and distributed query

We are looking for partners!



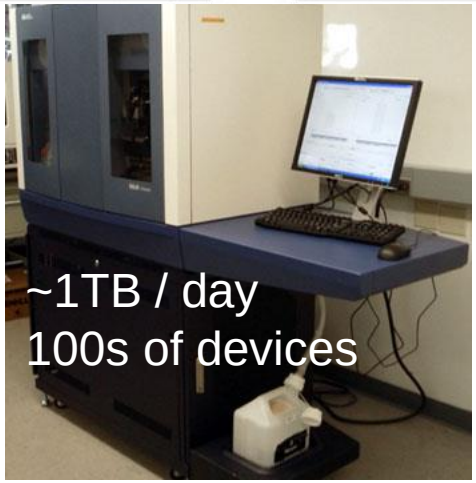

Dice
data science
Find Tech Jobs
Search job title only (e.g. Senior Java Developer)
Advanced Job Search
Search results: 1 - 30 of 10374
NEW
Create Search Agent Matching These Results
Current Search
Keyword
Undo data
Undo science
Jobs posted
30 days
Refine Res
Area Code
Country
Company
Skill
City
State / Provinces
Employment Type
Telecommute
Required Travel
Results viewable: 30 per page
1 2 3 4 5 6 7 8 9 10 Next
Job Title
Company
Location
Date Posted
View Summary Detail
Search By
Software Engineer - Data Science
Knewton
New York, NY
Oct-05-2012
Data Scientist Job
Bill & Melinda Gates Foundation
Seattle, WA
Sep-21-2012
Senior Analytics Developer
Dotomi
Chicago, IL
Oct-04-2012
Research Informatics Analyst I
St. Jude Children's Research Hospital
Memphis, TN
Sep-17-2012
Distinguished Scientist
PayPal
Austin, TX
Oct-09-2012
Remember to register or log-in
Dice Talent Communities
Android
Big Data
Cloud Computing
iOS

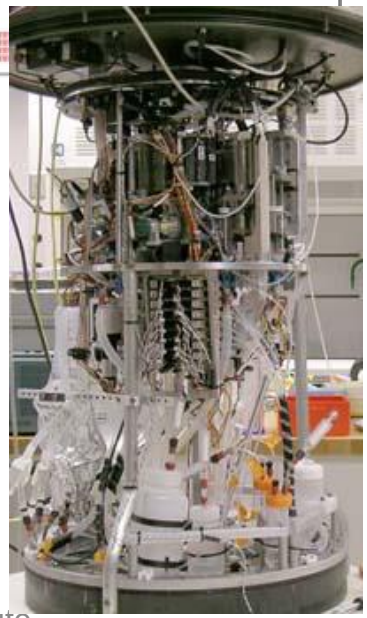
Science is reducing to a database query problem

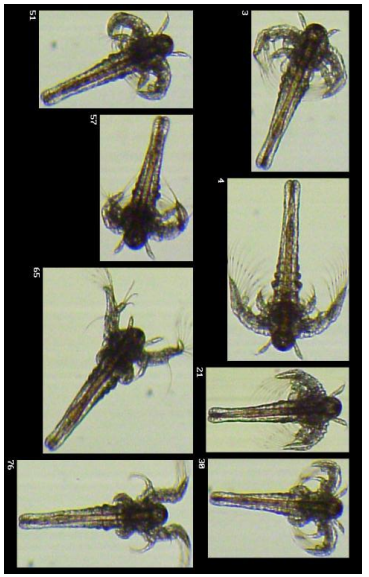
Old model: "Query the world" (Data acquisition coupled to a specific hypothesis)

New model: "Download the world" (Data acquisition supports many hypotheses)

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



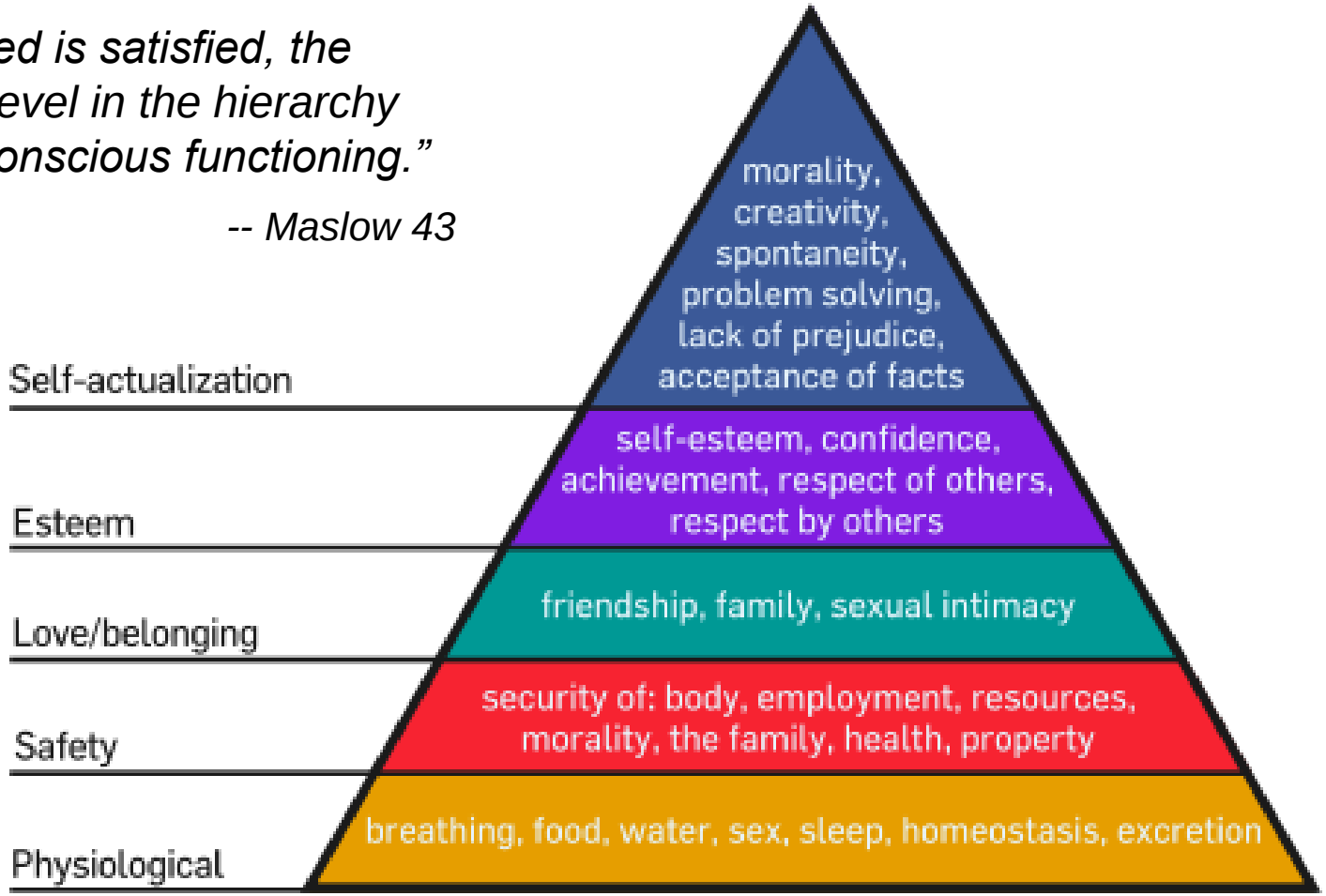




Needs Hierarchy

"As each need is satisfied, the next higher level in the hierarchy dominates conscious functioning."

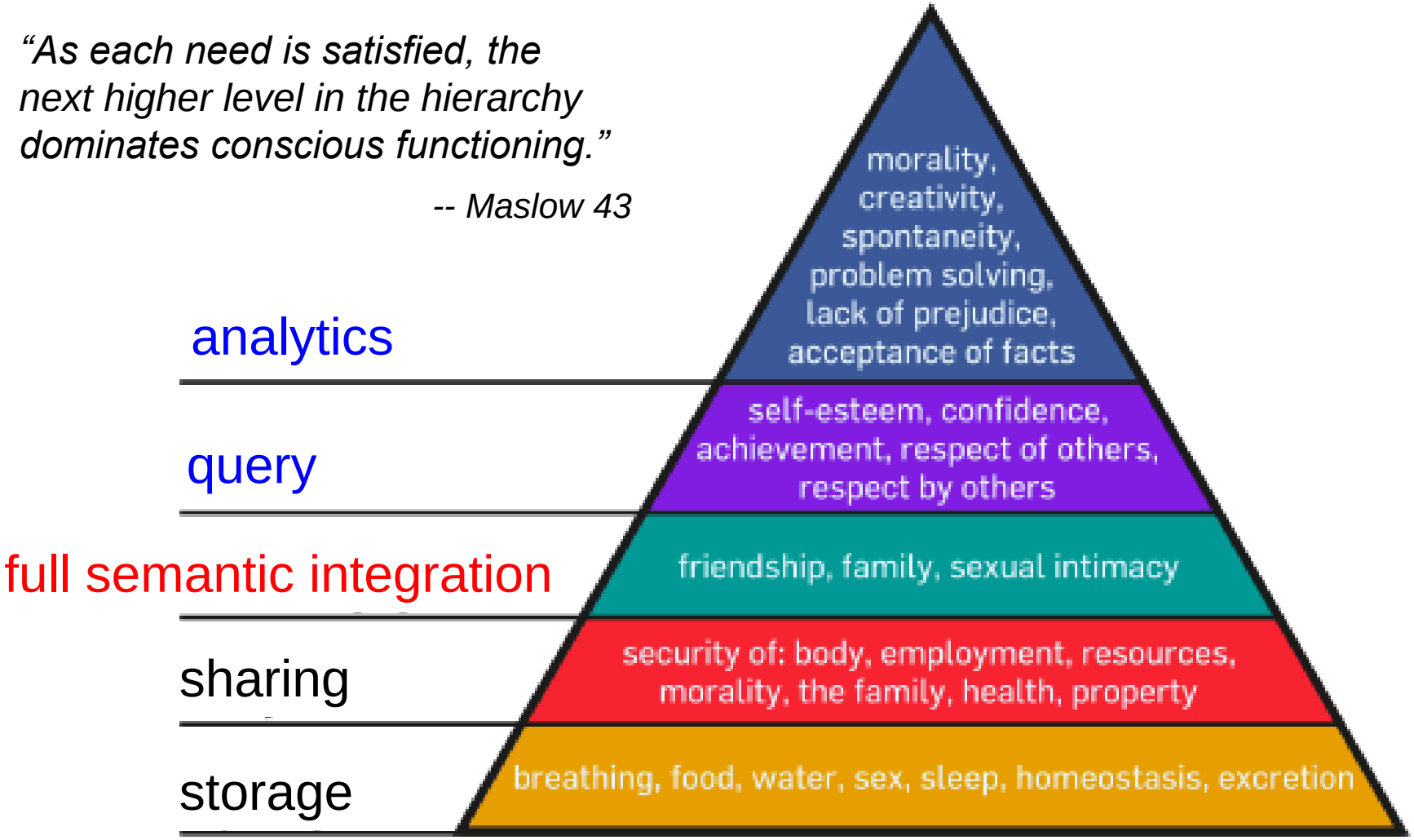
-- Maslow 43



A "Needs Hierarchy" of Science Data Management

"As each need is satisfied, the next higher level in the hierarchy dominates conscious functioning."

-- Maslow 43



Goal: Expose all the world's science data through declarative query interfaces

An Observation about NoSQL

- 2004 Dean et al. MapReduce
- 2008 Hadoop 0.17 release
- 2008 Olston et al. Pig: Relational Algebra on Hadoop
- 2008 DryadLINQ: Relational Algebra in a Hadoop-like system
- 2009 Thusoo et al. HIVE: SQL on Hadoop

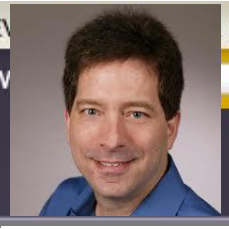
NoSQL is a misnomer

- NoMySQL?
- NoSchema?
- NoLoading?
- NoLicenseFees!

UW Data Science Education Efforts

	Students				Non-Students	
	CS/Informatics		Non-Major			
	undergrads	grads	undergrads	grads	professionals	researchers
UWEO Data Science Certificate					■	
Graduate Certificate in Big Data		■		■		
CS Data Management Courses	■	■	■			
eScience workshops						■
Intro to data programming	■		■			
eScience Masters (planned)		■		■	■	
Coursera Course: Intro to Data Science	■	■	■	■	■	■

Previous courses:
Scientific Data Management, Graduate CS, Summer 2006, Portland State University
Scientific Data Management, Graduate CS, Spring 2010, University of Washington



100 YEARS W

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

I EV ... I CONTACT US

Explore Programs | Online Learning | Student Resources

Bill Howe Richard Sharp Roger Barga

Approved by the UW Department of Computer Science & Engineering.

Certificates » Data Science » Winter 2013 Details » **Introduction to Data Science**

Winter 2013 Certificate

Details
Admissions
Apply Now

Courses

Introduction to Data Science
Methods for Data Analysis
Deriving Knowle...

Course Description

Introduction to Data Science

Bellevue, Classroom, Winter 2013

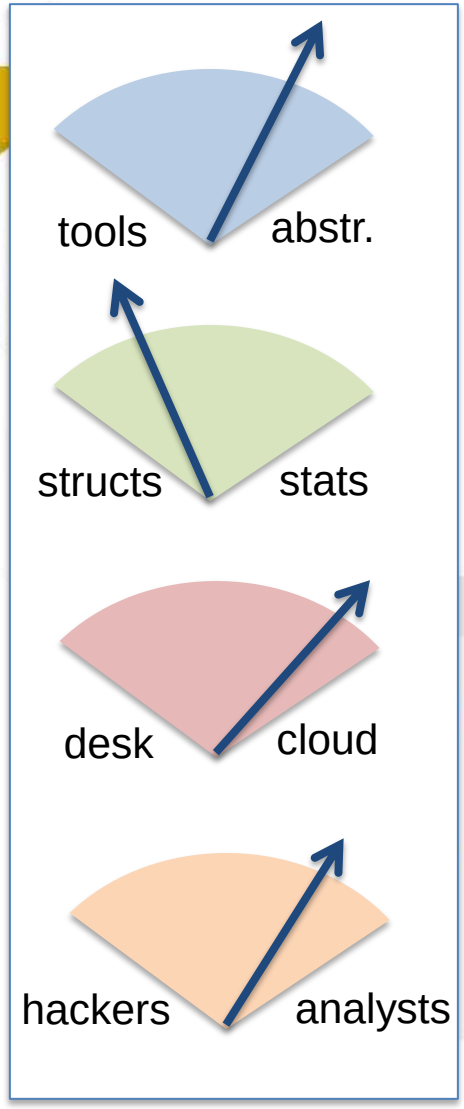
Instructor: Ernst Henle

Th, 1/10 - 3/14, 2013, 6-9 p.m.

Cost: $999 | 3 CEUs

This course is designed to introduce students to the data management, storage and manipulation tools common in data science and will apply those tools to real scenarios. An overview of different SQL and No-SQL database technologies is presented and the course finishes with a discussion of choosing the appropriate tool to get the job done.

Topics include:



Bill Howe, UW

Bill Howe

coursera COURSES UNIVERSITIES ABOUT ▼

Course Dashboard

Users

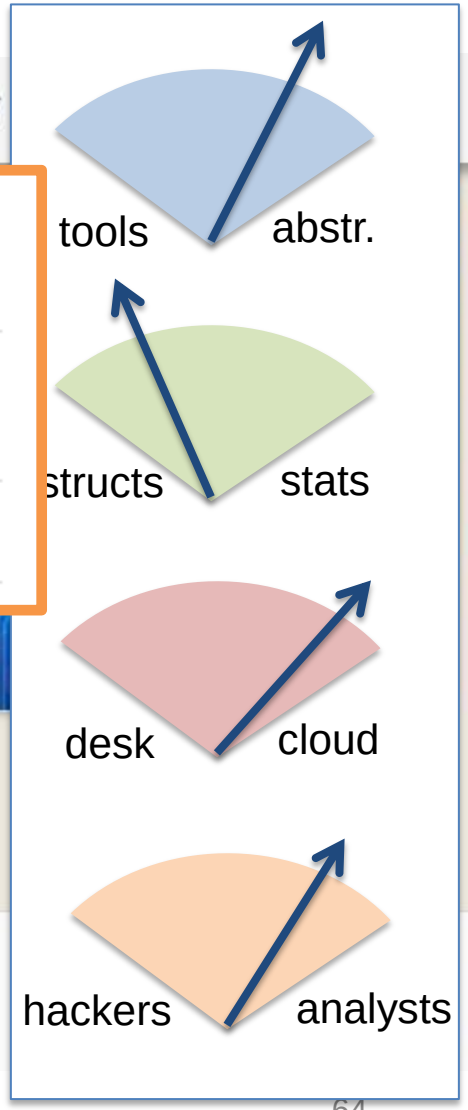
Total Registered Users	17834

easy to obtain through conventional curricula. Introduce yourself to the basics of data science and leave armed with practical experience programming massive databases.

You are signed up

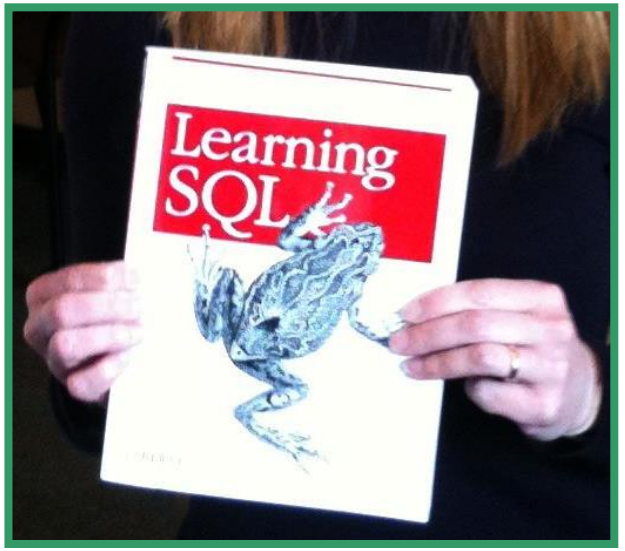
Next session: April 2013 (10 weeks long)

Statistics, Data Analysis, and Scientific Computing



What's the point?

- Conventional wisdom says "Science data isn't relational"
  - This is utter nonsense
- Conventional wisdom says "Scientists won't write SQL"
  - This is utter nonsense
- So why aren't databases being used more often?
  - They're a PITA
- We implicate difficulty in
  - installation, configuration
  - schema design, data loading
  - performance tuning
  - app-building (NoGUI?)



We ask instead, "What kind of platform can support ad hoc scientific Q&A with SQL?"

An observation about “handling data”

- How many plasmids were bombarded in July and have a rescue and expression?

```
SELECT count(*)
FROM [bombardment_log]
WHERE bomb_date BETWEEN ’7/1/2010' AND ’7/31/2010'
 AND rescue clone IS NOT NULL
 AND [expression?] = 'yes'
```

An observation about “handling data”

- Which samples have not been cloned?

```
SELECT *
FROM plasmiddb
WHERE NOT (ISDATE(cloned) OR cloned = ‘yes’ )
```

An observation about “handling data”

- How often does each RNA hit appear inside the annotated surface group?

```
SELECT hit, COUNT(*) as cnt
  FROM tigrfamannotation_surface
GROUP BY hit
ORDER BY cnt DESC
```

An observation about “handling data”

For a given promoter (or protein fusion), how many expressing line have been generated (they would all have different strain designations)

```
SELECT strain, count(distinct line)
FROM glycerol_stocks
GROUP BY strain
```

Find all TIGRFam ids (proteins) that are missing from at least one of three samples (relations)

```
SELECT col0 FROM [refseq_hma_fasta_TGIRfam_refs]
UNION
SELECT col0 FROM [est_hma_fasta_TGIRfam_refs]
UNION
SELECT col0 FROM [combo_hma_fasta_TGIRfam_refs]

EXCEPT

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

On NoSQL

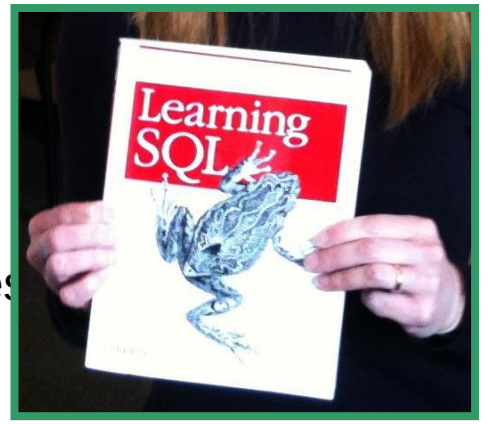

by @jrecursive
So, how do I query the database?
It's not a database. It's a key-value store!
Ok, it's not a database. How do I query it?
You write a distributed map reduce function in Erlang!
Did you just tell me to go fuck myself?
I believe I did, Bob.

An Observation on NoSQL

- 2004 Dean et al. MapReduce
- 2008 Hadoop 0.17 release
- 2008 Olston et al. Pig: Relational Algebra on Hadoop
- 2008 DryadLINQ: Relational Algebra in a Hadoop-like system
- 2009 Thusoo et al. HIVE: SQL on Hadoop

NoSQL is a misnomer

- NoMySQL?
- NoSchema?
- NoLoading?
- NoLicenseFees!



Problem

- Research data is captured and manipulated in spreadsheets
- This perhaps made sense five years ago; the data volumes were manageable
- But now: 50k rows, 100s of files, “mega-collabs”

Why not put everything into a database?

- A huge amount of up-front effort
- Hard to design for a moving target
- Running a database system is huge drain

Approach: SQLShare

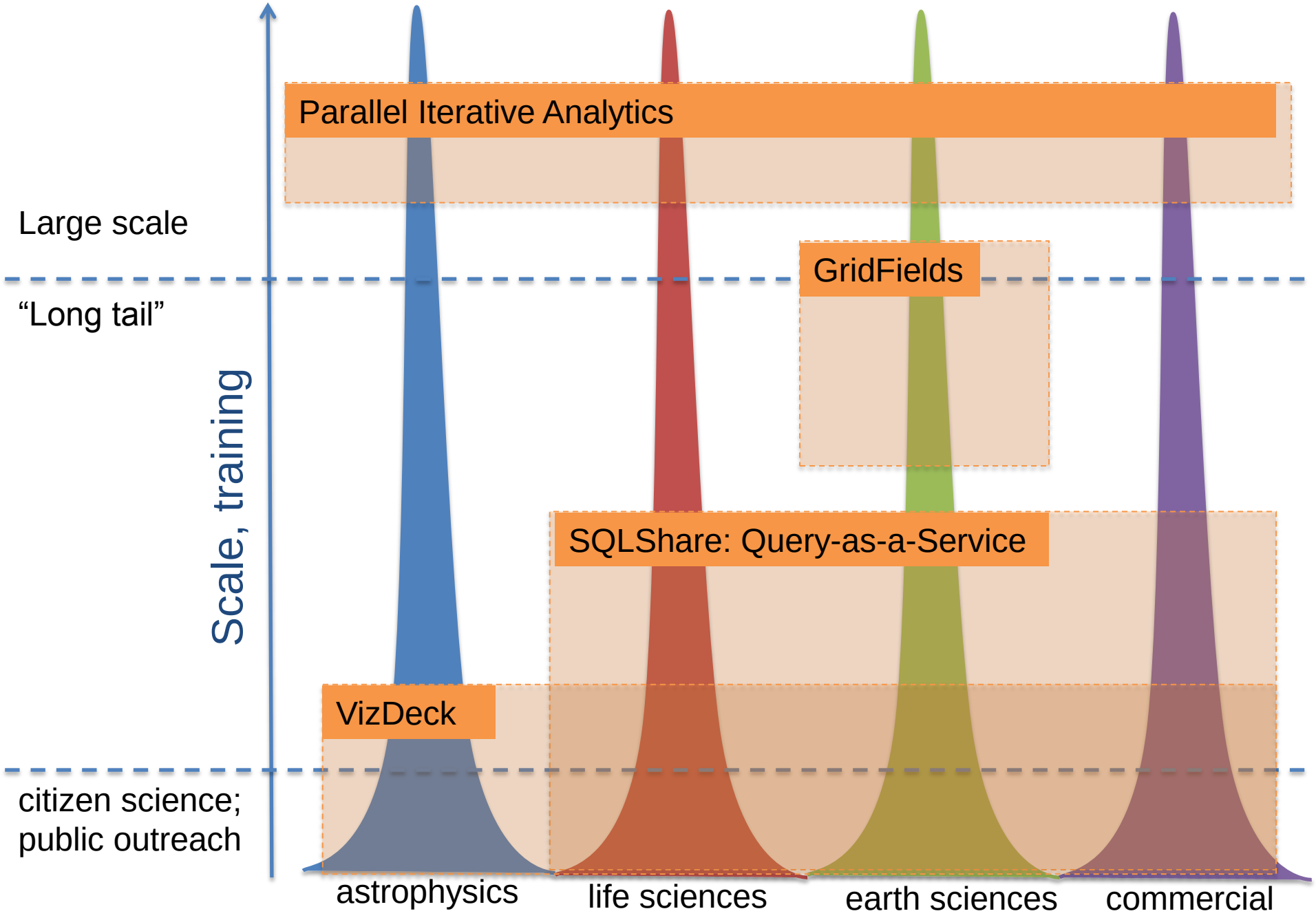
- Upload data through your browser: no setup, no install
- Login to browse science questions in English
- Click a question, see the SQL to answer it question
- Edit the SQL to answer an "adjacent" question, even if you wouldn’t know how to write it from scratch

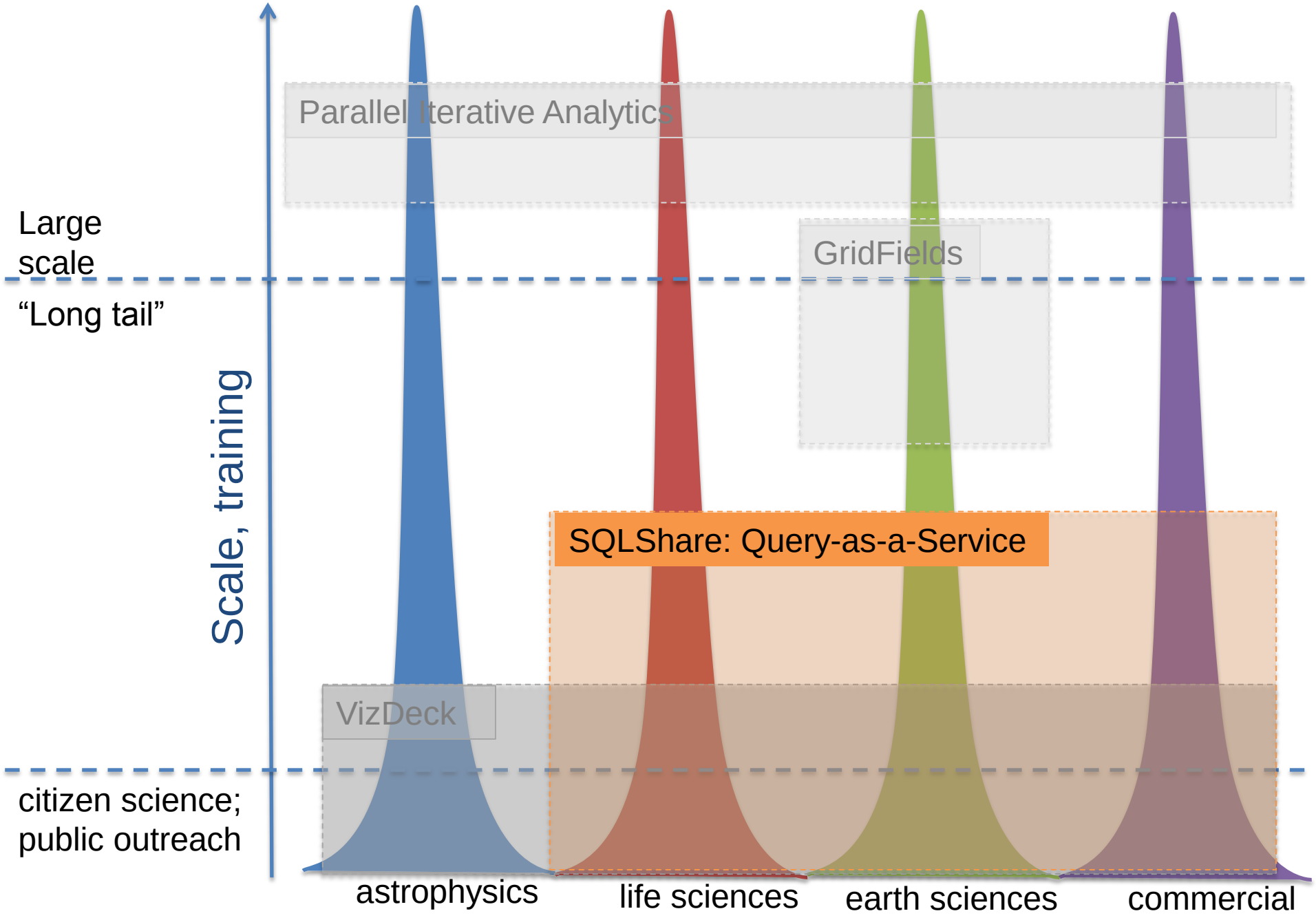
https://sqlshare.escience.washington.edu/

Gordon and Betty MOORE FOUNDATION

eScience Institute

NSF


Parallel Iterative Analytics
Large scale
GridFields
"Long tail"
Scale, training
SQLShare: Query-as-a-Service
VizDeck
citizen science;
public outreach
astrophysics
life sciences
earth sciences
commercial


Parallel Iterative Analytics
GridFields
SQLShare: Query-as-a-Service
VizDeck
Large scale
"Long tail"
citizen science; public outreach
Scale, training
astrophysics
life sciences
earth sciences
commercial

Four Conjectures about Declarative Query for Science

- Most science data manipulation tasks can be expressed in relational algebra
- Most science analytics task can be expressed in relational algebra + recursion

  Hellerstein 09, Re 12
- These expressions can be efficiently and scalably executed in the cloud
- Researchers are willing and able to program using relational algebra languages

  c.f. SDSS

Environmental Sampling

Sequencing

Sequence data

Questions?

correlate diversity w/environment?

correlate diversity w/nutrients?

find new taxa and their distributions?

find new genes?

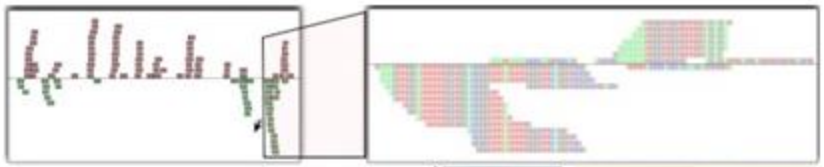
compare meta'omes?

search hits

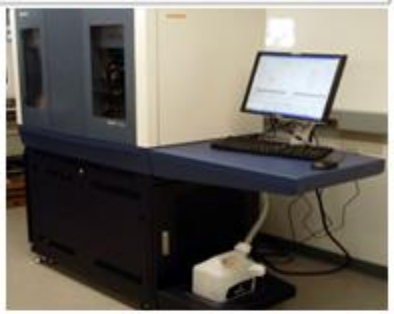
metadata

Public sequence DBs

Environmental Sampling



Sequencing



Sequence data

search hits

metadata

Pfams, TIGRfams, COGs, FIGfams

Public annotation DBs

Questions?

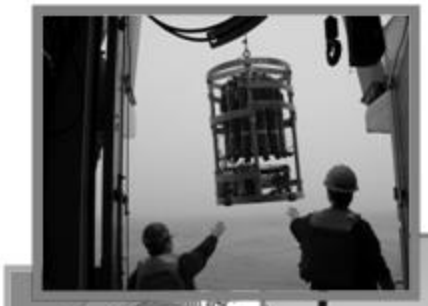
correlate diversity w/environment?

correlate diversity w/nutrients?

find new taxa and their distributions?

find new genes?

compare meta'omes?



Environmental Sampling

Sequencing

Sequence data

search lists

metadata

Pfams, TIGRfams,
COGs, FIGfams

Public annotation DBs

Questions?

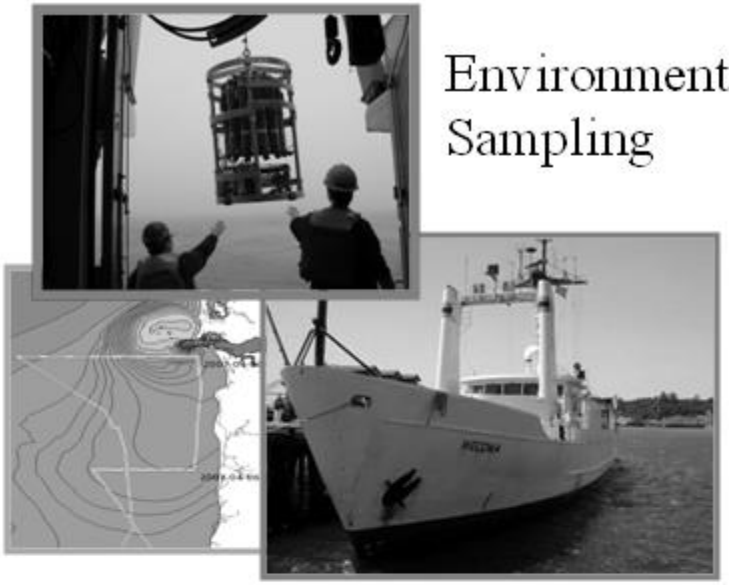
correlate diversity w/environment?

correlate diversity w/nutrients?

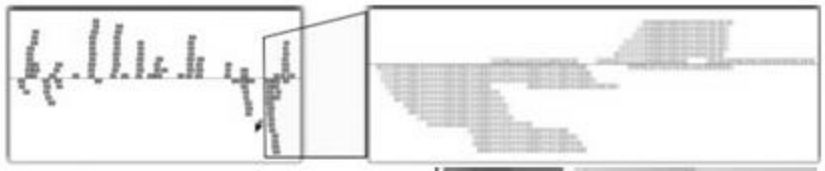
find new taxa and their distributions?

find new genes?

compare meta'omes?



Environmental Sampling



Sequencing

metadata

sequence data

search results

Pfams, TIGRfams, COGs, FIGfams

Public annotation DBs

Questions?

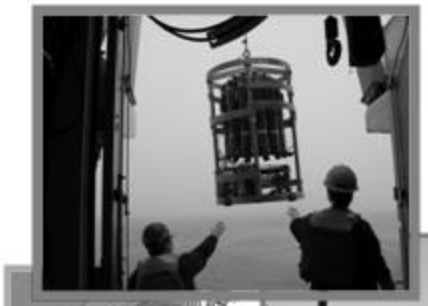
correlate diversity w/environment?

correlate diversity w/nutrients?

find new taxa and their distributions?

find new genes?

compare meta*omes?



Environmental Sampling

Sequencing

Sequence data

search hits

metadata

SQLShare

Pfams, TIGRfams, COGs, FIGfams

Public annotation DBs

SQL

correlate diversity w/environment?

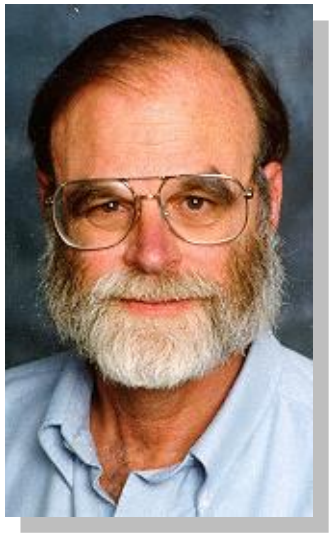
correlate diversity w/nutrients?

find new taxa and their distributions?

find new genes?

compare meta*omes?

Experimental Engagement Algorithm for the Long Tail



A stripped-down version of Jim Gray's "20 questions" methodology

1. Get the data
2. Load the data "as is" – no schema design
3. Get ~20 questions (in English)
4. Translate the questions into SQL (when possible)
5. Provide these "starter queries" to the researchers

Q: Can researchers questions be expressed in SQL?
Q: Are a few examples sufficient for novices to self-train with SQL?
Q: Can we scale this process up?
Q: If so, will the use of SQL reduce their data handling overhead?

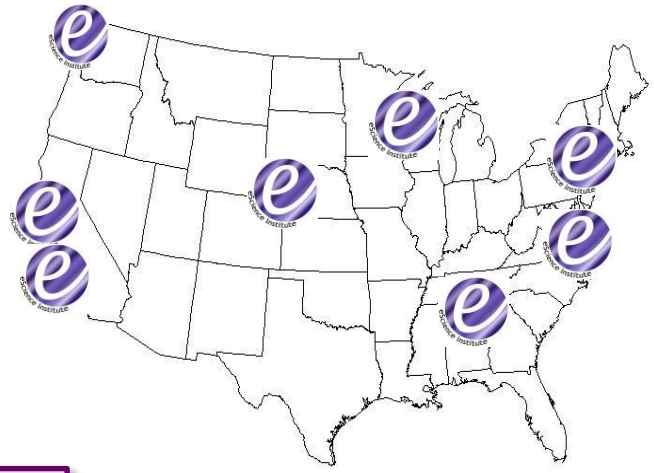
- Which samples have not been cloned?

```
SELECT *
FROM plasmiddb
WHERE NOT (ISDATE(cloned) OR cloned = 'yes')
```

Incubator

- Seed grants to students and postdocs
- Rotating staff from science **and** industry
- An evolving portfolio of reusable services
- A network of cross-boundary **partnerships**
- *Produce digital capital **and** human capital*

2018



2013

Data Science Incubator

2008

eScience Institute

Some local observations:

- Big data work exposes common ground
- Every job is becoming “data scientist”
- More TT-shaped people!
- Democratization to the long tail is key
- *Industry and research aren’t too*


eScience Institute