

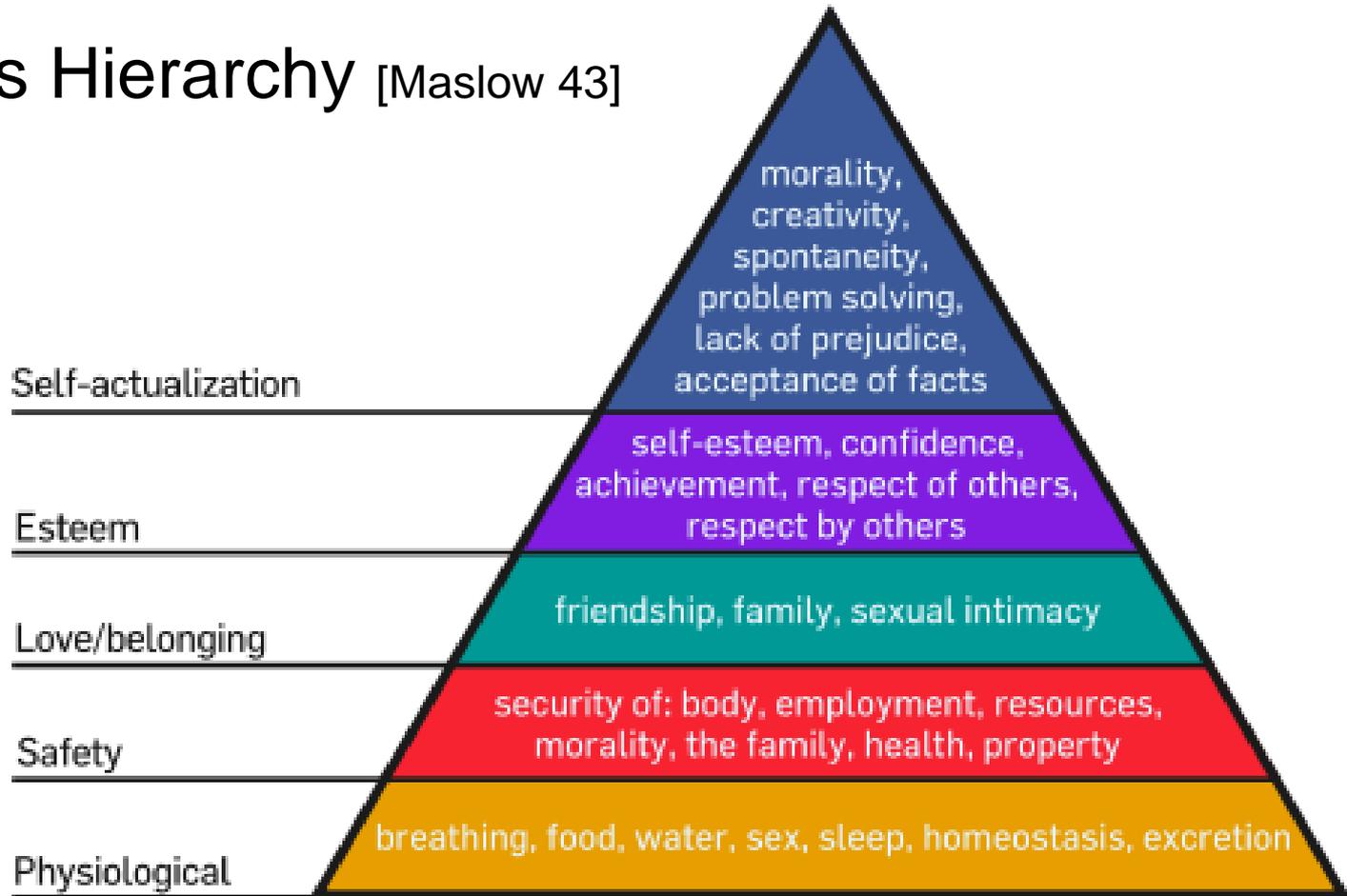


# Advancing Declarative Query for Data-Intensive Science

Bill Howe, PhD

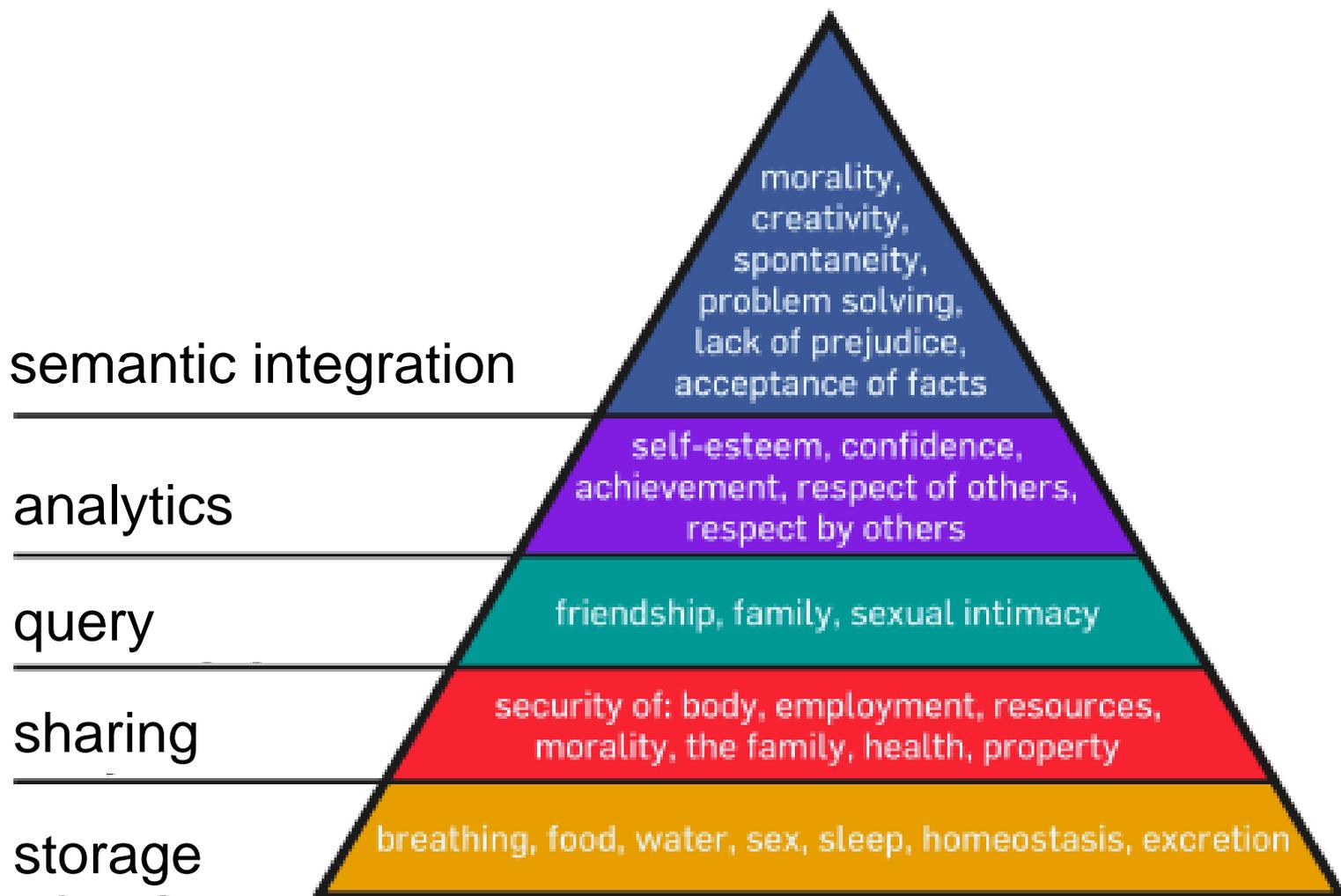
University of Washington  
CSE and eScience Institute

# Needs Hierarchy [Maslow 43]

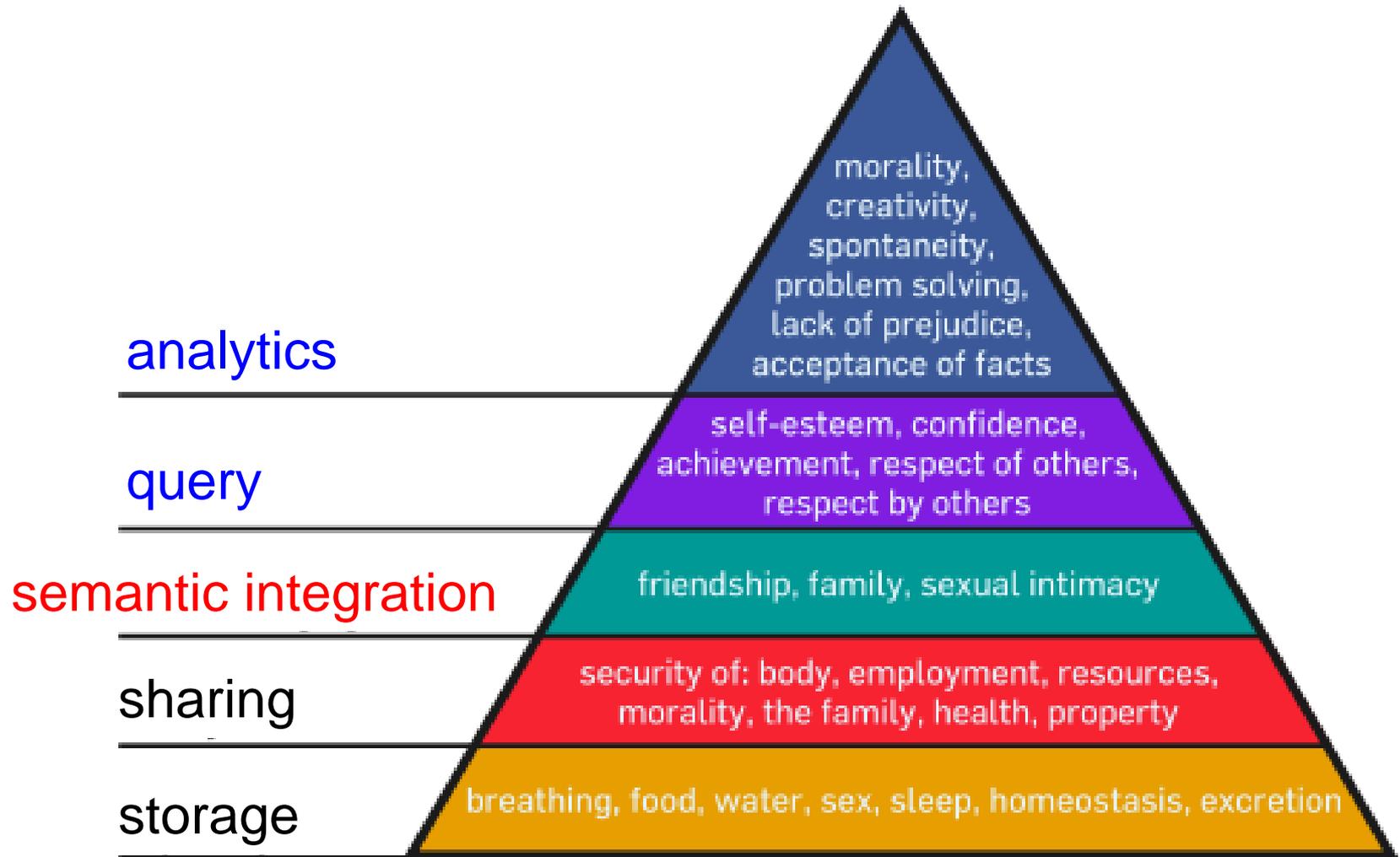


*“As each need is satisfied, the next higher level in the hierarchy dominates conscious functioning.”*

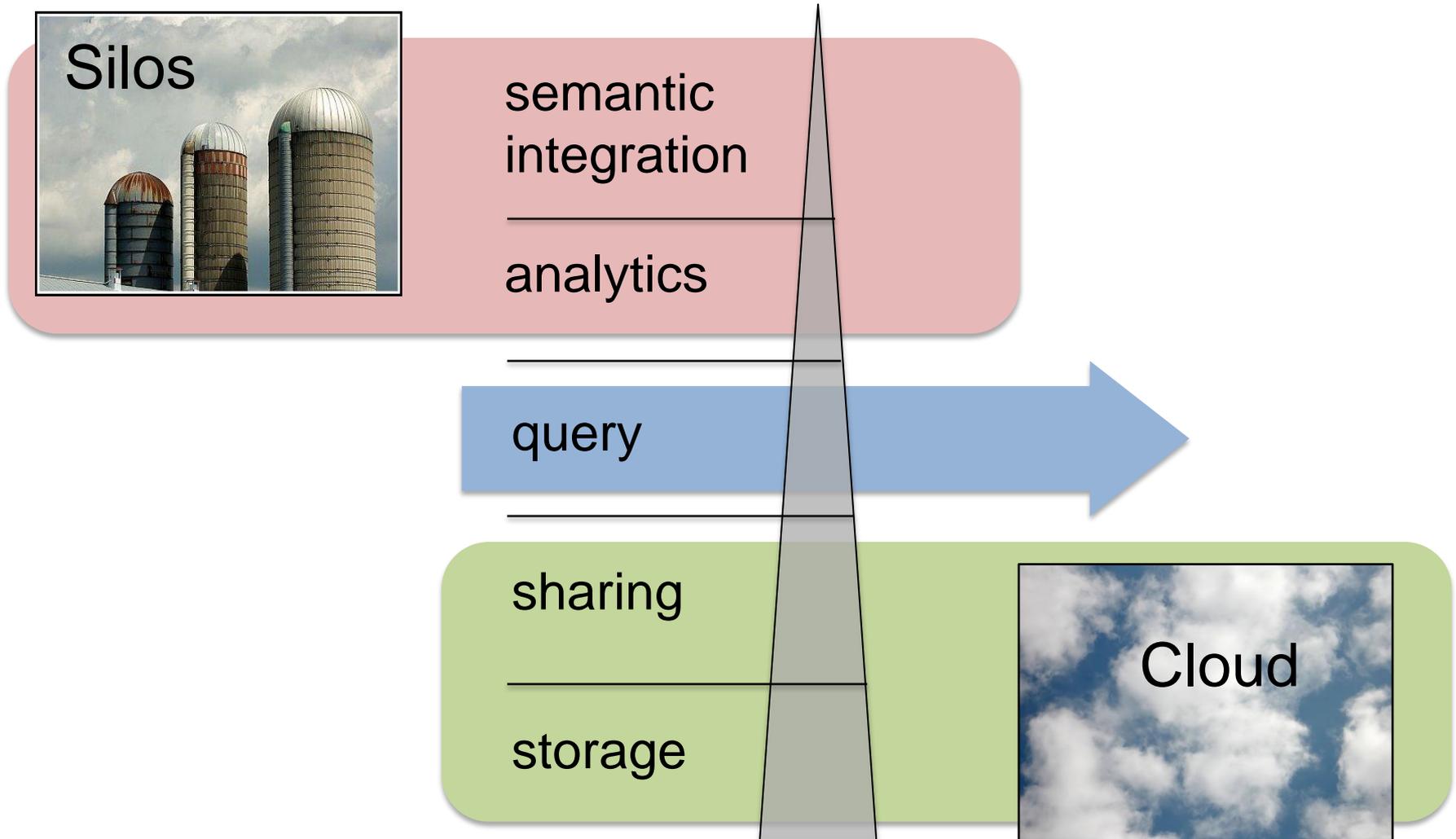
## A “Needs Hierarchy” of Science Data Management



# A “Needs Hierarchy” of Science Data Management

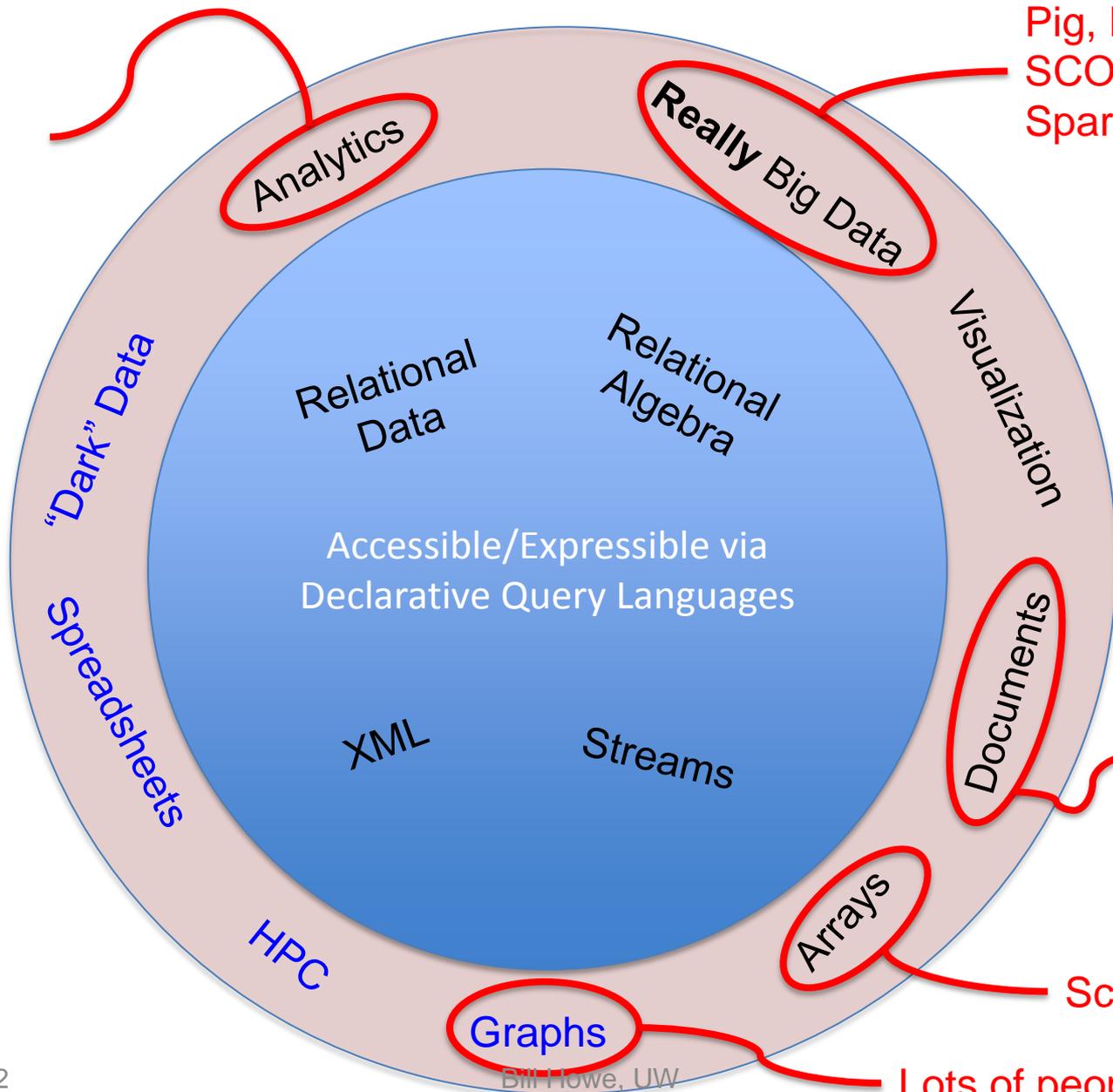


# A “Needs Hierarchy” of Science Data Management



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

MADLib,  
Chris Re,  
etc.



Pig, HIVE,  
SCOPE, ...,  
Spark, etc.

AI, IR folks

SciDB

Lots of people

## Projects in this Space

- Database-as-a-Service for the 99%  
<http://sqlshare.escience.washington.edu>
- Parallel Datalog for Declarative, Scalable Analytics for the 1%  
<http://clue.cs.washington.edu>
- Exploratory Visual Analytics  
<http://vizdeck.com>
- Data Pricing: Incentives for Data Sharing  
(lead: Magda Balazinska)





# **DATABASE-AS-A-SERVICE FOR THE 99%**

# 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	1462	COG5032	1.00E-09	30	60.1	2105	Phosphatidylinositol kinase and
chr_11[1-100000].70	2586						
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**

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]
```

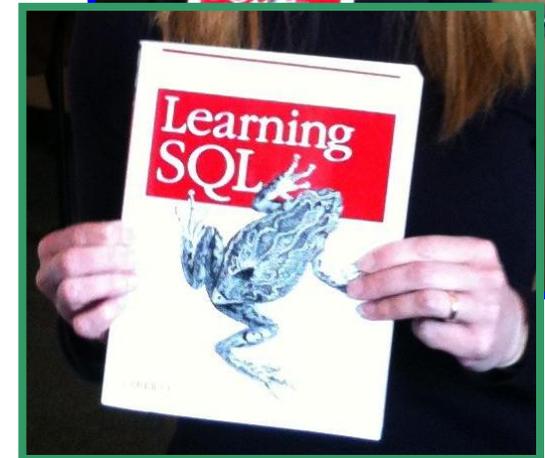
# Saved Queries = Views

- 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, 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)

# What's the point?

- Databases are underused by the 99%
- Conventional wisdom says “Scientists won’t write SQL”
  - This is nonsense
  - We corroborate findings by SDSS, etc.
- Instead, we implicate difficulty in
  - installation
  - configuration
  - **schema design**
  - performance tuning
  - **data loading**
  - app-building

*We ask instead, “How can we deliver declarative query in support of ad hoc scientific Q&A?”*



**Your datasets**

All datasets

Shared datasets

Recent activity... 2

Recently viewed »

---

Upload dataset

New query

---

YOUR TOP VIEWED

csv2.csv 115

csv2.csv 115

blackhole 16

Vizlet Scores 14

vizlets\_23nov... 14

---

POPULAR TAGS

biomed 138

ht\_screening\_r... 81

seqvalidation 52

protein 47

oceanography 23

tsg 16

suna 16

Your Datasets

Filter dataset by keyword:

Name	Sharing / Owner	Modified
topic.csv research topics <a href="#">simpleschema</a>	billhowe@washington.edu	Feb 24, 2012 9:04 AM
mhip_zip_eScience_022112a.csv additional outcome measures for mhip dataset <a href="#">mhip</a>	billhowe@washington.edu	Feb 21, 2012 5:05 PM
total students taking AMATH301 and CSE142 <a href="#">csse</a>	billhowe@washington.edu	Feb 4, 2012 11:46 PM
total students taking amath301 prior to cse142 <a href="#">csse</a>	billhowe@washington.edu	Feb 4, 2012 11:46 PM
amath_analysis.csv anonymized course registrations for AMATH301 and CSE142 <a href="#">csse</a>	billhowe@washington.edu	Feb 4, 2012 11:46 PM
elements_with_atomic_numbers_92_and_below.csv test dataset for alicia	billhowe@washington.edu	Jan 20, 2012 1:45 PM
SeaFlow Example Dataset Clean SeaFlow Example Dataset <a href="#">seaflow</a>	billhowe@washington.edu	Jan 20, 2012 12:40 PM
categorized_fat.xlsx.txt <a href="#">health</a>	billhowe@washington.edu	Dev 7, 2011 1:06 PM
Vizlet Scores and Features Score is the number of promote actions for each vizlet type for each column p <a href="#">vizdeck</a>	billhowe@washington.edu	Dev 2, 2011 1:40 AM
VizDeck User Study Timing and Success <a href="#">vizdeck</a>	billhowe@washington.edu	Dev 1, 2011 11:56 PM
vizstudy_analysisv7.csv	billhowe@washington.edu	Dev 1, 2011 11:44 PM
Vizlet Scores Score of each (session_x column_v column_vizlet type)		

<http://sqlshare.escience.washington.edu>

# Upload Dataset



**File:**

<b>2010.csv</b>	<b>7.47 MB</b>
-----------------	----------------



Analysing your file...

Cancel

<http://sqlshare.escience.washington.edu>

# Upload Dataset



### 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](#)

<http://sqlshare.escience.washington.edu>

Your datasets  
All datasets  
Shared datasets  
Recent activity... 2  
Recently viewed »

Upload dataset  
New query

## YOUR TOP VIEWED

csv2.csv 115  
csv2.csv 115  
blackhole 16  
Vizlet Scores 14  
vizlets\_23nov... 14

## POPULAR TAGS

biomed 138  
ht\_screening\_r... 81  
seqvalidation 52

**topic.csv** Only you can view this

Last modified: Feb 24, 2012 9:04 AM billhowe@washington.edu

research topics

simpleschema

```
SELECT * FROM [table_topic.csv]
```

Edit dataset

Derive dataset

Create snapshot

More actions ▾

DATASET PREVIEW Rows 1 - 20 of 20 | Columns 2 of 2

activity	topic
Astroinformatics	disc
Client+Cloud	cloud
Client+Cloud	vis
cloud certificate	cloud
Cloud Vis	cloud
Cloud Vis	disc
Cloud Vis	vis
cloud workshop	cloud
escience appliances	cloud
Graph query	db

<http://sqlshare.escience.washington.edu>

Your datasets

All datasets

Shared datasets

Recent activity... 2

Recently viewed »

Sharing with others

Last modified: Nov 28, 2011 12:32 PM billhowe@washington.edu

and glycerol\_id parsed

```

string(d.name, 1, 5) as construct
string(d.name, 7, 5) as glycerol_id
[howe].[dnasamples.csv] d

```

Edit dataset Derive dataset Create snapshot More actions

DATASET PREVIEW Rows 1 - 100 of 10656 | Columns 12 of 12

Edit dataset

Derive dataset

Create snapshot

More actions

AnphA	00176	452	2	AnphA.00176.a.A1.GU26581.D1	2010-02-10 17:02:47.760147	2010-02-10 17:02:47.760147
AnphA	00202	453	1	AnphA.00202.a.B1.GE26906.D1	2010-02-10 17:02:47.760147	2010-02-10 17:02:47.760147
AnphA	00202	454	1	AnphA.00202.a.B1.GU26910.D1	2010-02-10 17:02:47.760147	2010-02-10 17:02:47.760147
AnphA	00205	455	3	AnphA.00205.a.A1.GE26620.D1	2010-02-10 17:02:47.760147	2010-02-10 17:02:47.760147

# 2010 Pilot- Outreach and Education-based sampling: Schooner Adventuress

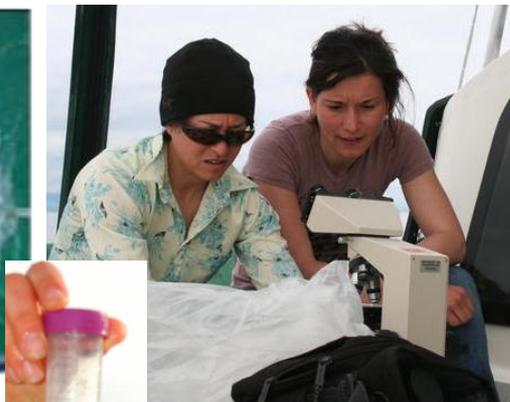


Robin Kodner



5/23/2012

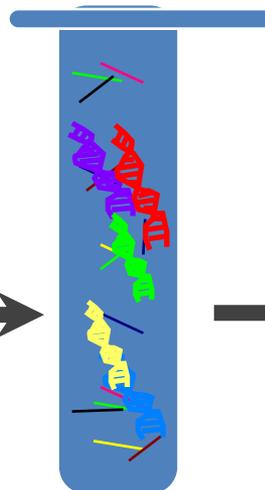
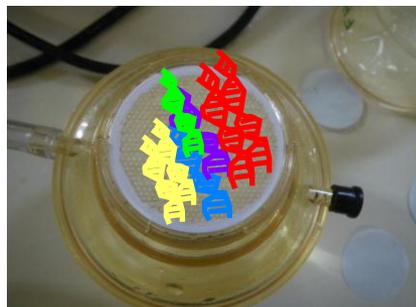
# plankton tows



	A	B	C	D	E	F	G	H	I	J	K
1	Sample: 13 - Stuart Island										
2	Date: 1 Sep 2010, counted on 21 Sep 2010										
3	Dilution Factor: 1										
4	Notes: 55 mL of sample										
5	Total volume counted (uL):										10
6	<i>Chaetosceros</i>	35	34	20	28	67	20	47	13	10	16
7	<i>Asterionellopsis</i>	19	21	5	15	1	14	6		1	4
8	<i>Thalassiosira</i>	6			4	6	1	1			4
9	<i>Ditylum</i>	2	1	1	2	2	2	1	1	6	3
10	<i>Cylindrotheca</i>	2			1		1				1
11	<i>Dictyocha</i>	2									1
12	<i>Pseudo-nitzschia</i>	2	6		2	3			5	4	2
13	<i>Odontella</i>	1	2	2				2	3		
14	<i>Leptocylindrus</i>	1			4		5	1	14	5	1
15	<i>Skeletonema</i>		17		13	10	2		22		
16	<i>Rhizosolenia</i>		2	2	4	1		1	2	1	2
17	<i>Tropidoneis</i>		1	1	1				1		1
18	<i>Pyrophacus</i>					2					
19	<i>Protoperdinium</i>								1	1	
20	<i>Ceratium</i>									1	
21											
22											
23											
24											
25											
26											
27											
28	Total Cell Count:	84	31	74	92	45	60	63	31	33	
29	Total Cells:										513
30	Cells/mL:										51300

taxa counts

# whole water



DNA/RNA sequencing

```
1          10        20        30        40        50
ACATCAGCCCATCCATCATTCCAAAAATAATCTAATCATGCTTTTCAAGACTCT
90          100       110       120       130
GACTCTCCGTGGCGCGAAACAACTCGACCTCGAGATCGTCTCGGAGTCGGGCG
180         190       200       210       220
AGTGCCCTACCGGCAACGCTTTCCGCCTTCGCAACCACTCCACCAGAGCTGCC
260         270       280       290       300       310
AAGACTCCCCCGCCGCAACTCAACTGGAAACGAACTCAACAACCTCAA
350         360       370       380       390
CCGCTCGGCCATCCAGGGCCCTCTTTCTGCTGCGGTTCTCCGTCGCCG
440         450       460       470       480
AGGACGTCAAGATGGGATCCGACTCTGGACAGCTCGTTTTCGTTCCCTCCAA
520         530       540       550       560
ACCAAACAAGGGGAGGACCCCAACAATGTTGTCTTTCGATGAAGATGCCATCC
```

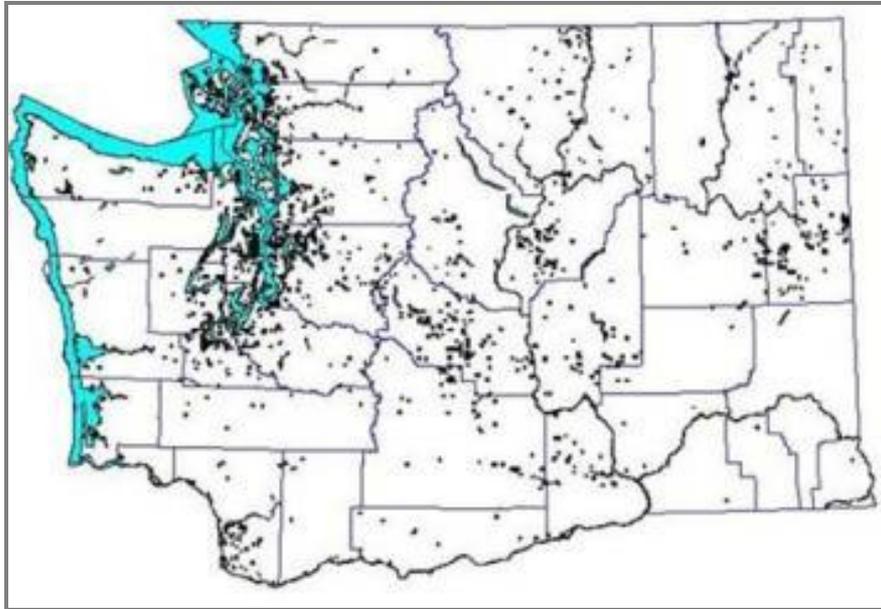
nutrient analysis + toxin analysis

# NatureMapping Program

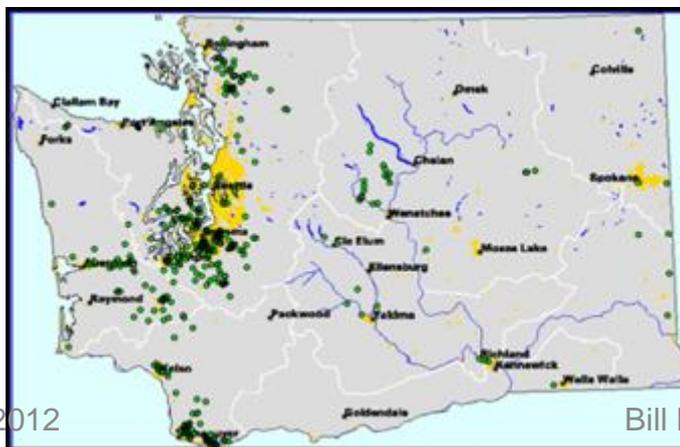


Karen  
Dvornich

## Wildlife Observations (1902- )

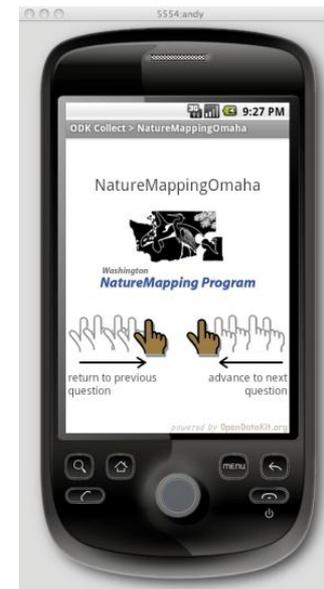
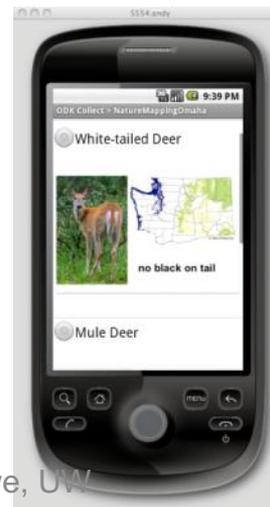


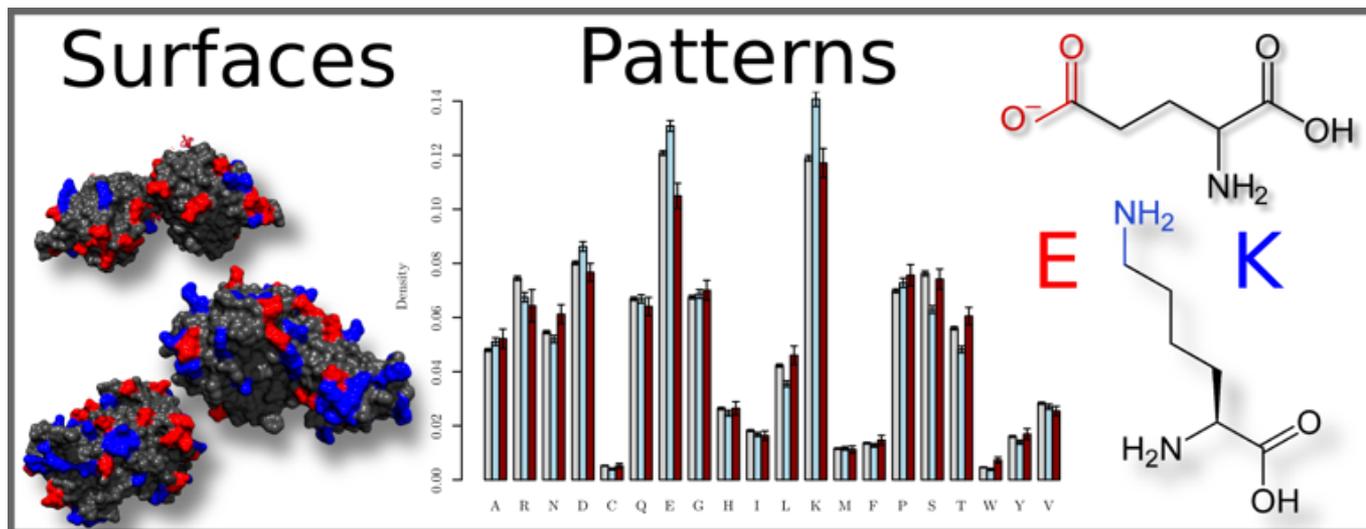
## Water Quality Monitoring Sites (2003 - )



## Data collection and submission options:

1. Download/upload spreadsheet
2. Online data entry
3. NatureTracker on handheld/GPS
4. Android ODK (Open Data Kit)





*“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.”*

# SQLShare as a CS Research Platform

- 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)
- VizDeck: Automatic Mashups and Visualization
  - (Bill Howe, Alicia Key)
- Personalized Query Recommendation
  - (Yuan Zhou, Bill Howe)
- Info Extraction from Spreadsheets
  - (Mike Cafarella, Dave Maier, Bill Howe)
- Differential Privacy in Multi-tenant DBs
  - (Dan Suciu, Bill Howe)

SSDBM 2011

SIGMOD 2011 (demo)

SSDBM 2011

CHI 2012

SIGMOD 2012 (demo)



## 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

