Database-as-a-Service for Long Tail Science

Bill Howe

Garret Cole

Nodira Khoussainova
Luke Zettlemoyer
Shaminoo Kapoor
Patrick Michaud

Gordon and Betty MOORE FOUNDATION

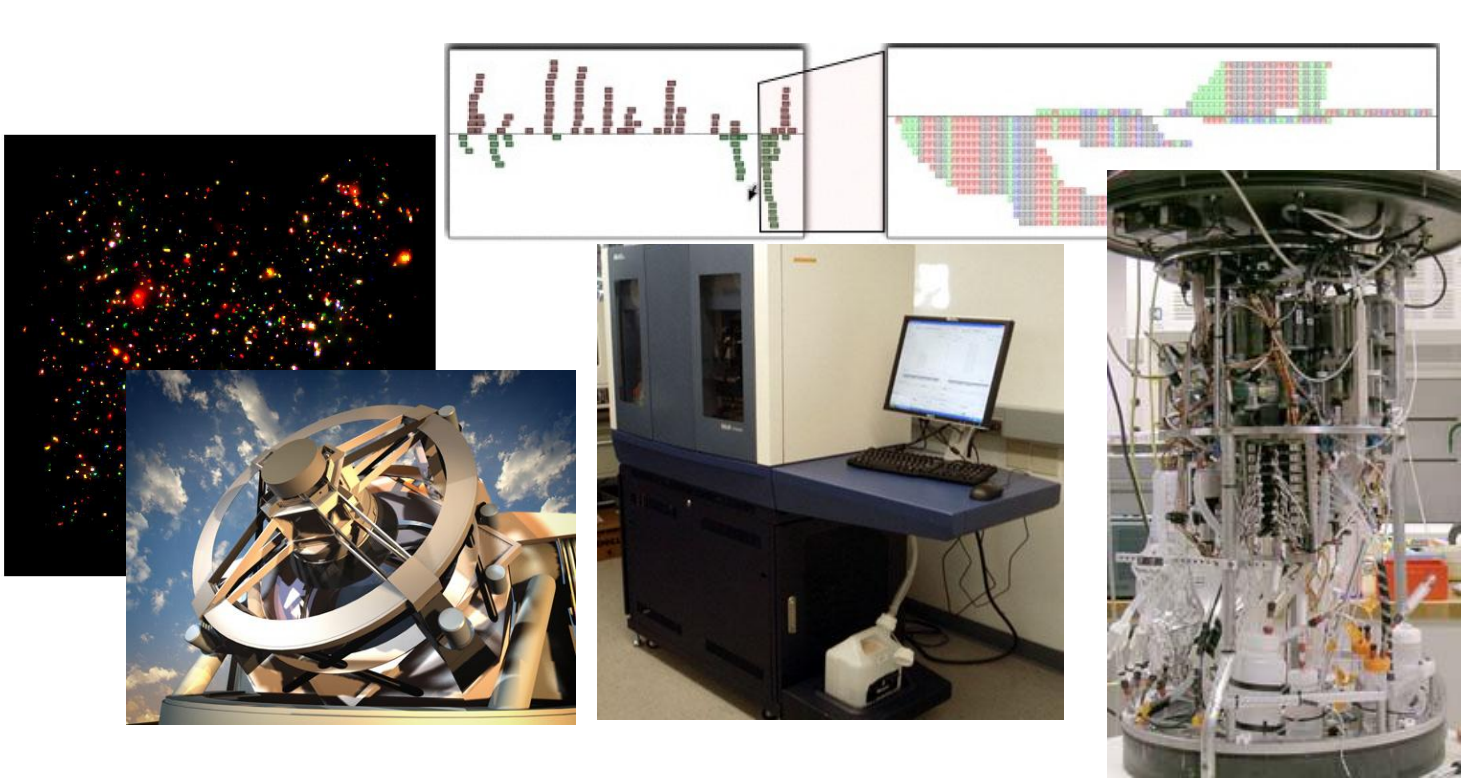
Microsoft® Research

All science is reducing to a database problem

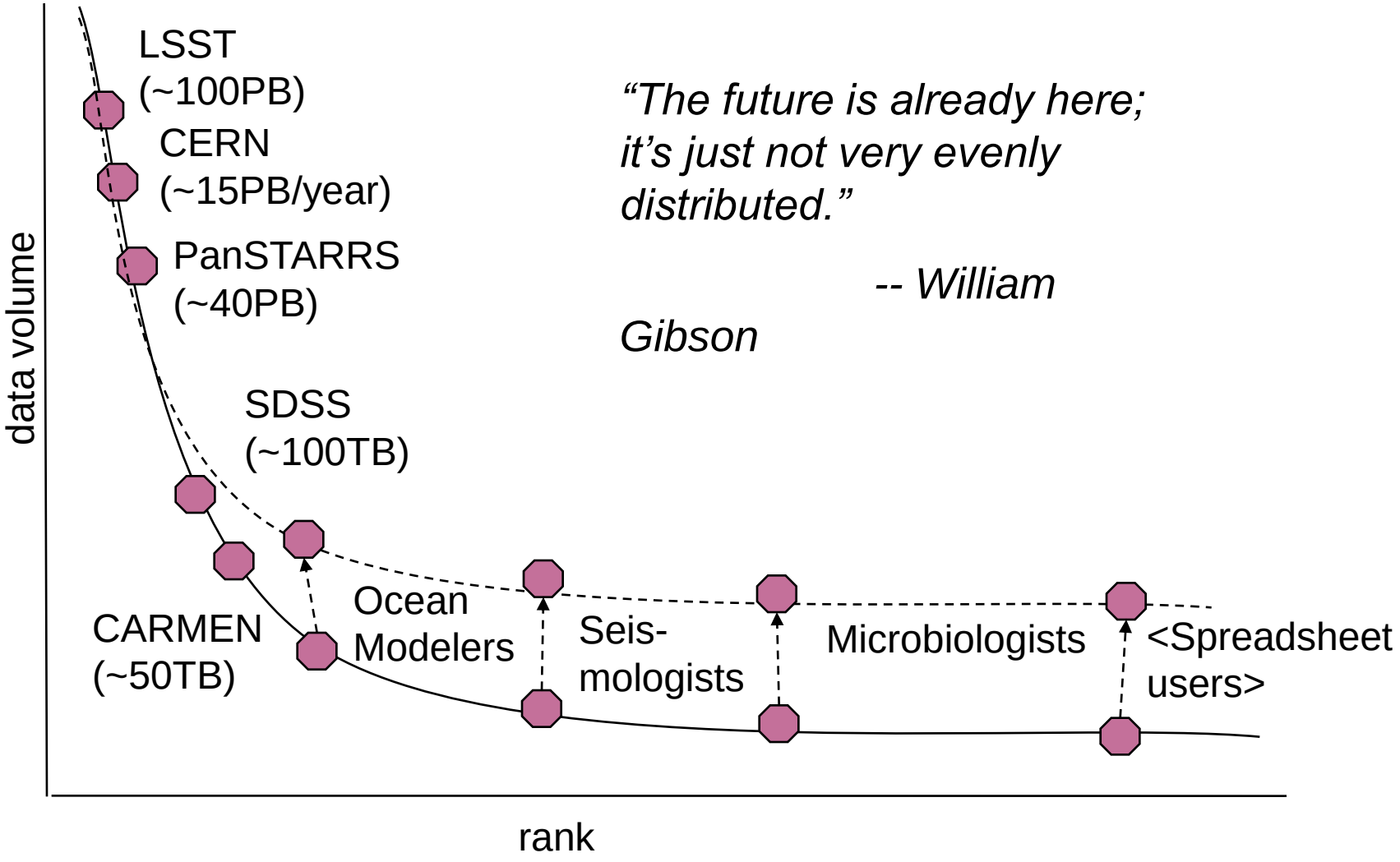
Old model: "Query the world" (Data acquisition coupled to a specific hypothesis)

New model: "Download the world" (Data acquired en masse, in support of 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,



The Long Tail

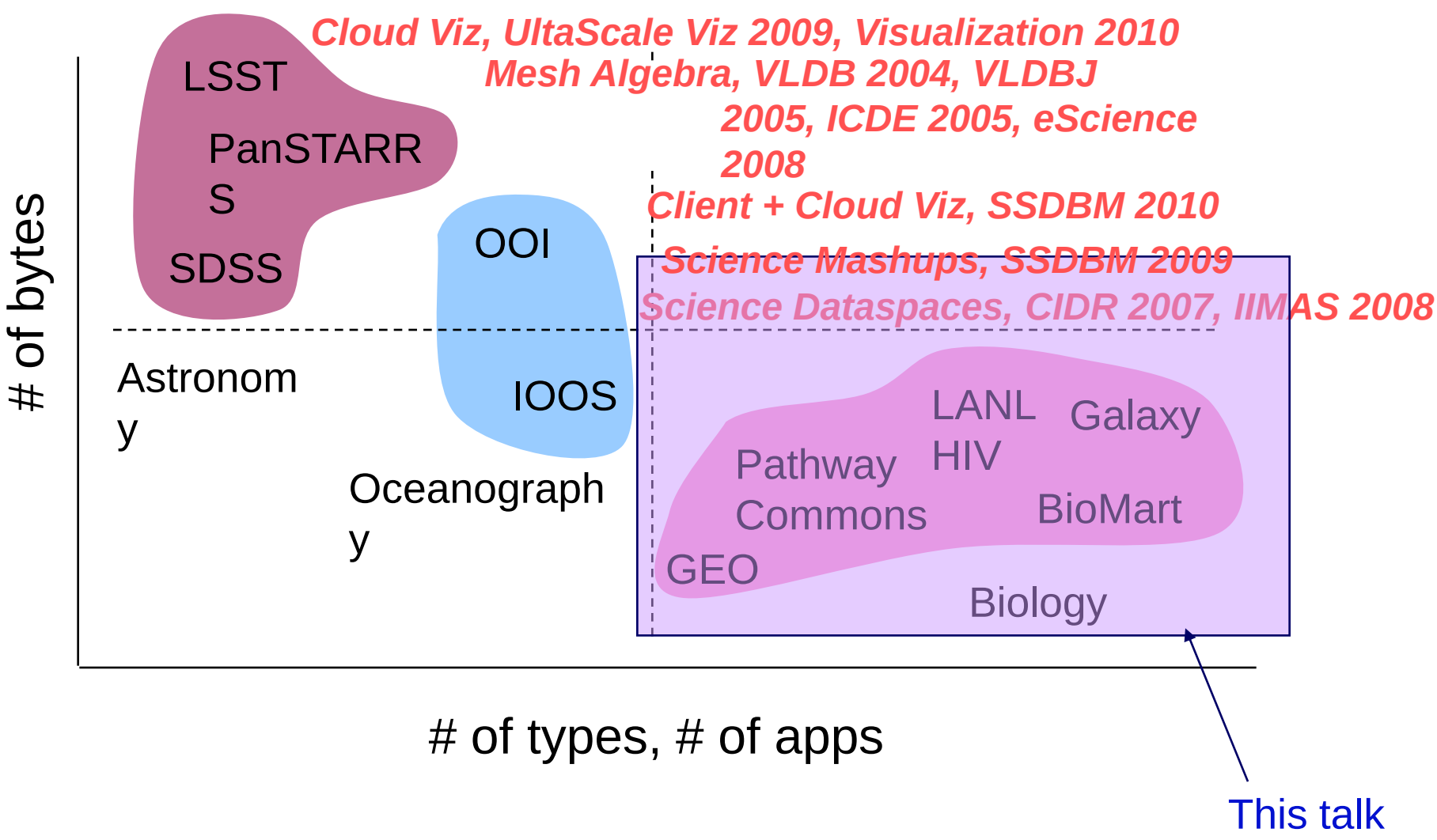


"The future is already here; it's just not very evenly distributed."

-- William Gibson

The other “Large Scale”

see also:
Skew handling, SOCC 2010
Clustering, SSDBM 2010



Ad Hoc Research Data

Fasta format

```
;LCBO - Prolactin precursor - Bovine
; a sample sequence in FASTA format
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*
```

```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*
```

Spread sheets

F16

	A	B	C	D	E	F
1						
2						
3	Date	Start time	End time	Pause	Sum	Comment
4	2007-05-07	9,25	10,25	0	1	Task 1
5	2007-05-07	10,75	12,50	0	1,75	Task 1
6	2007-05-07	18,00	19,00	0	1	Task 2
7	2007-05-08	9,25	10,25	0	1	Task 2
8	2007-05-08	14,50	15,50	0	1	Task 3
9	2007-05-08	8,75	9,25	0	0,5	Task 3
10	2007-05-14	21,75	22,25	0	0,5	Task 3
11	2007-05-14	22,50	23,00	0	0,5	Task 3
12	2007-05-15	11,75	12,75	0	1	Task 3
13						
14						
15						
16						
17						
18						

Delimited ASCII

```
#query  GO  reference DB    reference family    e-value description
lcl|10082_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.6 GO:0006412  TIGRFAM TIGR00001   6e-08   translation
lcl|10082_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.6 GO:0022625  TIGRFAM TIGR00001   6e-08   cytosolic larg
lcl|10082_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.6 GO:0000315  TIGRFAM TIGR00001   6e-08   organellar lar
lcl|10082_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.6 GO:0003735  TIGRFAM TIGR00001   6e-08   structural cor
lcl|9019_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.1  GO:0005507  TIGRFAM TIGR00003   5.5e-06 copper ion bir
lcl|9019_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.1  GO:0006825  TIGRFAM TIGR00003   5.5e-06 copper ion tra
lcl|5439_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.1  GO:0006402  TIGRFAM TIGR00004   5.9e-67 mRNA catabolic
lcl|5439_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.1  GO:0004521  TIGRFAM TIGR00004   5.9e-67 endoribonucles
lcl|813_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.2   GO:0009451  TIGRFAM TIGR00005   2.1e-29 RNA modificati
lcl|813_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.2   GO:0001522  TIGRFAM TIGR00005   2.1e-29 pseudouridine
lcl|813_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.2   GO:0009982  TIGRFAM TIGR00005   2.1e-29 pseudouridine
lcl|6708_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.4  GO:0009451  TIGRFAM TIGR00005   1.2e-18 RNA modificati
lcl|6708_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.4  GO:0001522  TIGRFAM TIGR00005   1.2e-18 pseudouridine
lcl|6708_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.4  GO:0009982  TIGRFAM TIGR00005   1.2e-18 pseudouridine
lcl|4_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.1 GO:0009451  TIGRFAM TIGR00005   1.4e-16 RNA modification
lcl|4_1_CCCI_CCOA_CCOB_CCOC_CFAP_CFAS.1 GO:0001522  TIGRFAM TIGR00005   1.4e-16 pseudouridine synt
```

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 protei
chr_9[320001-420000].548	2833	COG5406	2.00E-04	38	43.9	1001	Nucleosome binding factor SP
chr_27[320001-404298].20	3991	COG4547	5.00E-05	18	46.2	620	Cobalamin biosynthesis protei
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 hy
chr_12[720001-820000].86	2895	COG5032	2.00E-09	30	60.5	2105	Phosphatidylinositol kinase an
chr_12[800001-900000].109	1463	COG5032	1.00E-09	30	60.1	2105	Phosphatidylinositol kinase an
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
...										

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 protei
chr_9[320001-420000].548	2833	COG5406	2.00E-04	38	43.9	1001	Nucleosome binding factor SP
chr_27[320001-404298].20	3991	COG4547	5.00E-05	18	46.2	620	Cobalamin biosynthesis protei
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-		077	1.00E-25	26	114	1089	Ubiquitin carboxyl-terminal hy
chr_12[72000							
chr_12[80000							
chr_11[1-1000							
chr_11[80001-							

coastal sample

id	query	hit	e_value	query_start	query_end	hit_start	hit_end	hit_length
6409	FHJ7DRN01BYA61.1	TIGR00149	2.20E-21	1	84	43	125	134
6410	FHJ7DRN01BDTEA.1	TIGR00149	3.40E-09	3	42	30	69	134
6411	FHJ7DRN02HEUGQ.1	TIGR00149	1.70E-05	4	46	1	46	134
6412	FHJ7DRN01CA4BO.1	TIGR00149	5.30E-05	4	45	1	45	134
6413	FHJ7DRN01DM2FK.3	TIGR01651	5.70E-64	1	76	511	586	606
6414	FHJ7DRN01B8BPS.1	TIGR01651	1.20E-36	1	52	500	551	606
6415	FHJ7DRN02JM54P.1	TIGR01651	2.20E-24	15	80	301	366	606
6416	FHJ7DRN02FK6C5.2	TIGR00039	2.70E-16	1	45	37	85	153
6417	FHJ7DRN01D019A.1	TIGR00039	8.90E-12	5	65	48	118	153
6418	FHJ7DRN02FYAFO.1	TIGR00039	1.60E-11	1	76	67	153	153

SwissProt

Search in

Protein

1 result for **CC0672** in Uni

Reduce sequence redun

› Did you mean cc067?

PIRSF

SMART

TIGRFAMs

PROSITE

ProtoNet

Search...

	Accession	Entry name	Status	Protein names
☐	Q9AAD0	Q9AAD0_CAUCR	★	**Cobalamin biosynthesis protein**

TIGR01650	GO:0051116	contributes_to
TIGR01651	GO:0009236	NULL
TIGR01651	GO:0051116	NULL
TIGR01660	GO:0008940	NULL
TIGR01660	GO:0009061	NULL
TIGR01660	GO:0009325	NULL
TIGR01663	GO:0000012	NULL
TIGR01663	GO:0046403	NULL

azole

uf family,

3c YPR042c

C4G9.05

PBP35G2.14 SPCC1682.08

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
```

An observation about “handling data”

- 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]
```

Long Tail Science DaaS Requirements

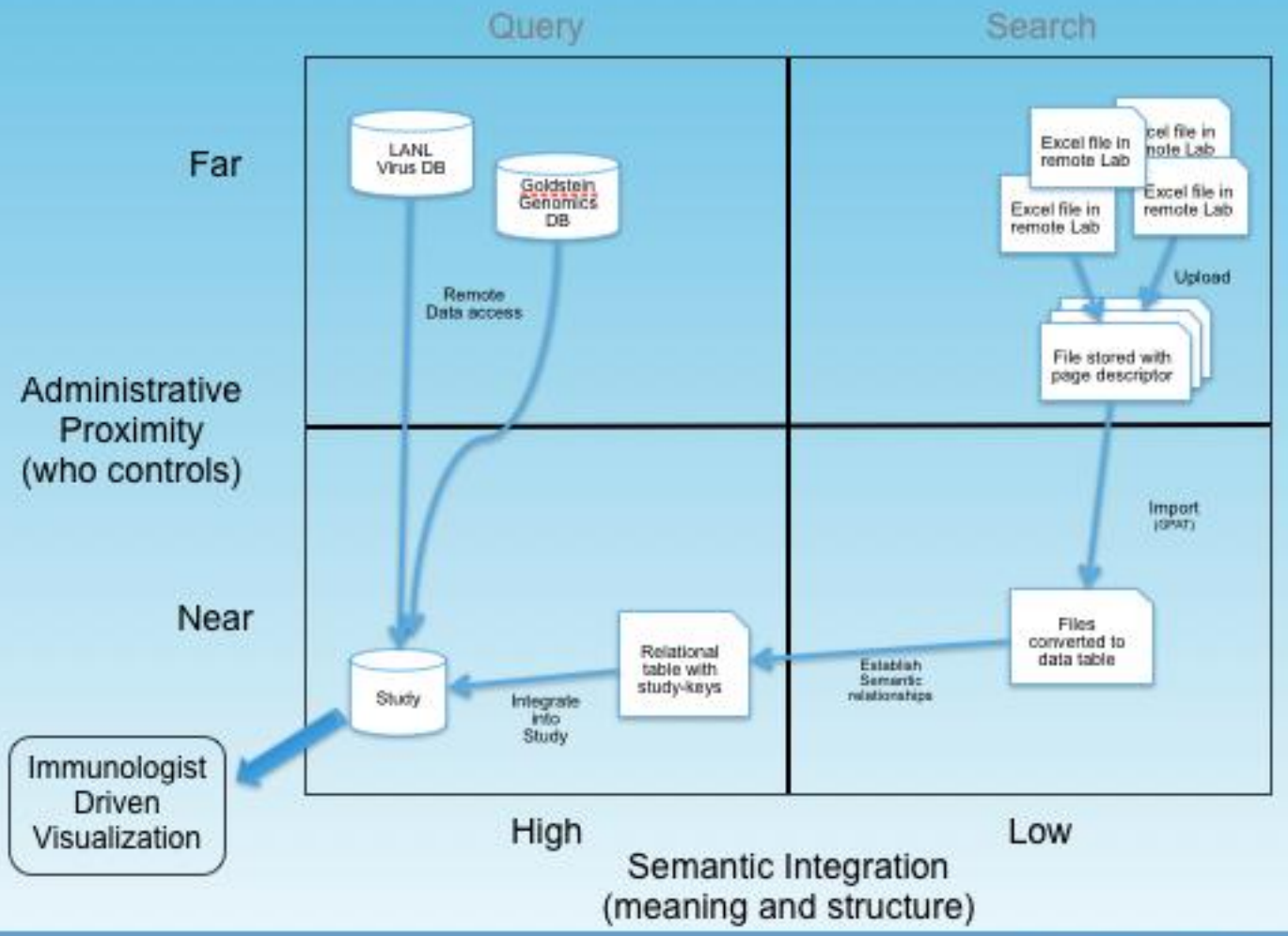
- Schema-Later or Schema-Free
  - Schema represents a shared consensus on structure, semantics, data model, usage modalities
  - By definition, no such consensus exists at the frontier of research
  - By definition, lots of schema churn
  - By definition, dirty data
- Consistency?
  - Read mostly, appends, versioning/batch replace
- Scale?
  - Relatively small (<100GB)
- Dataspace abstraction attractive [Halevy, Maier, Franklin 2005]
  - anecdotally well-received

Some Science DaaS Motivations

- Chronic IT poverty + exploding data volumes
  - especially in the long tail
- Data sharing is the whole point
  - mandated by funding agencies
  - in the cloud, sharing reduces to policy
- Public reference databases
  - Globally accessible in the cloud

CHAVI-CAVD Dataspace Concept

Towards an HIV Enterprise Dataspace



More Examples

What is the location of the E.Coli glycerol stock(s) for gene X promoter fusion?

What is the -80 freezer and liquid nitrogen location of worm strain for gene
x promoter fusion and/or protein fusion?

Show me all worm strains currently in storage?

Show me all worm strains for gene X?

Show me all worm strains for gene X promoter fusion?

Show me all worm strains for gene X protein fusion?

Show me a table of all worm strains with early embryonic expression?

Show me the location of the imaging data for gene x?

What strains have been shipped to Yale, Stanford etc, and when were they shipped?

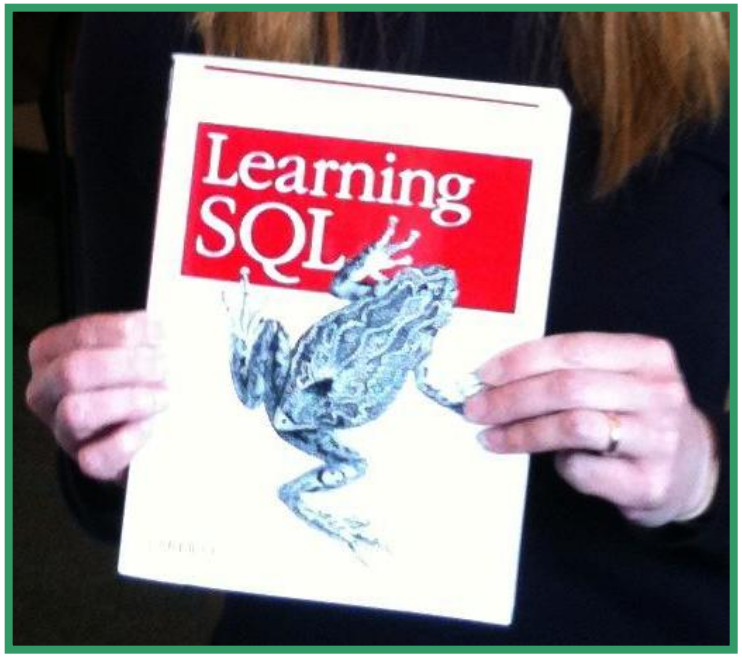
Show me a list of all primers with PCR failure?

What genes have midiprep stocks but no worm strains?

Discovery: SQL Does not Terrify Scientists

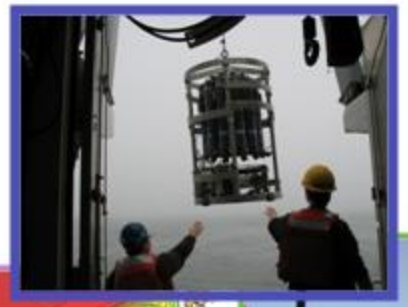
What's the point?

- Databases are underused in (long tail) science
- Conventional wisdom says "Scientists won't write SQL"
  - This is utter horseshit
  - witness SDSS if you don't trust us
- Instead, we implicate difficulty in
  - installation
  - configuration
  - schema design
  - performance tuning
  - data ingest
  - app-building (over-reliance on GUIs)



So we ask "What kind of platform can support ad hoc scientific Q&A?"

Example Workflow: Environmental Metagenomics



Environmental Sampling

Sequencing

Sequence data

Public annotation DBs

Questions?

correlate diversity w/environment?

correlate diversity w/nutrients?

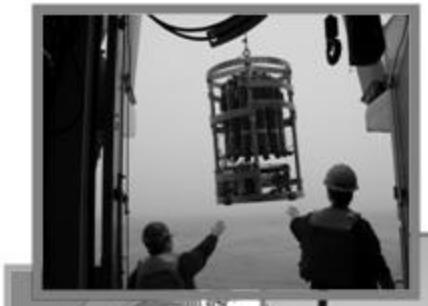
find new taxa and their distributions?

find new genes?

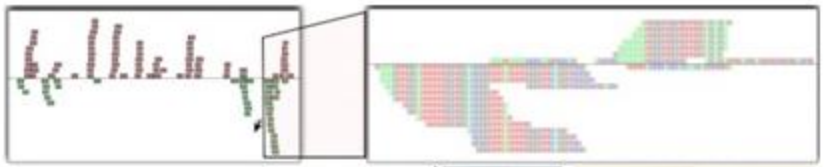
compare meta'omes?

search hits

metadata



Environmental Sampling



Sequencing

Sequence data

reads/kits

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?

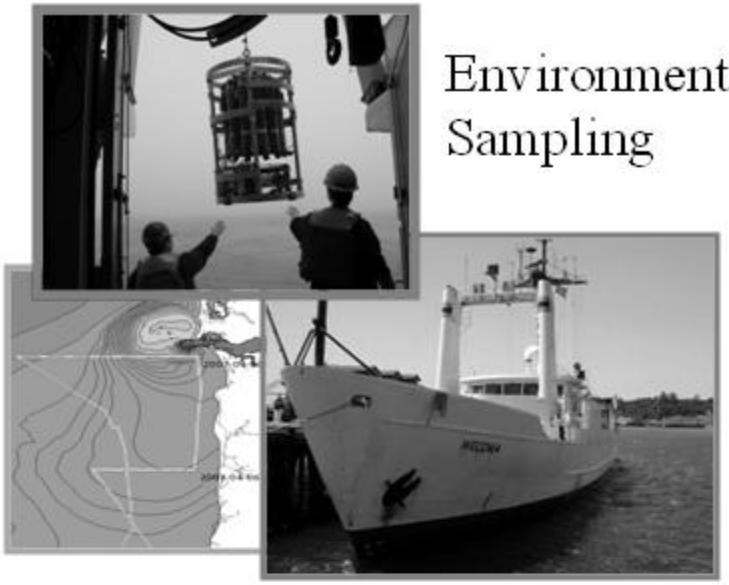
correlate diversity w/environment?

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?

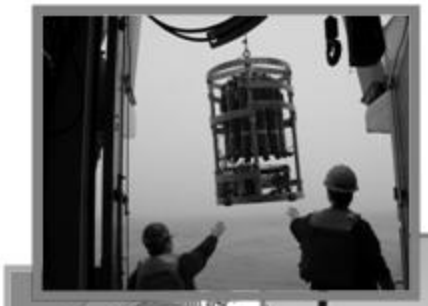
correlate diversity w/environment?

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

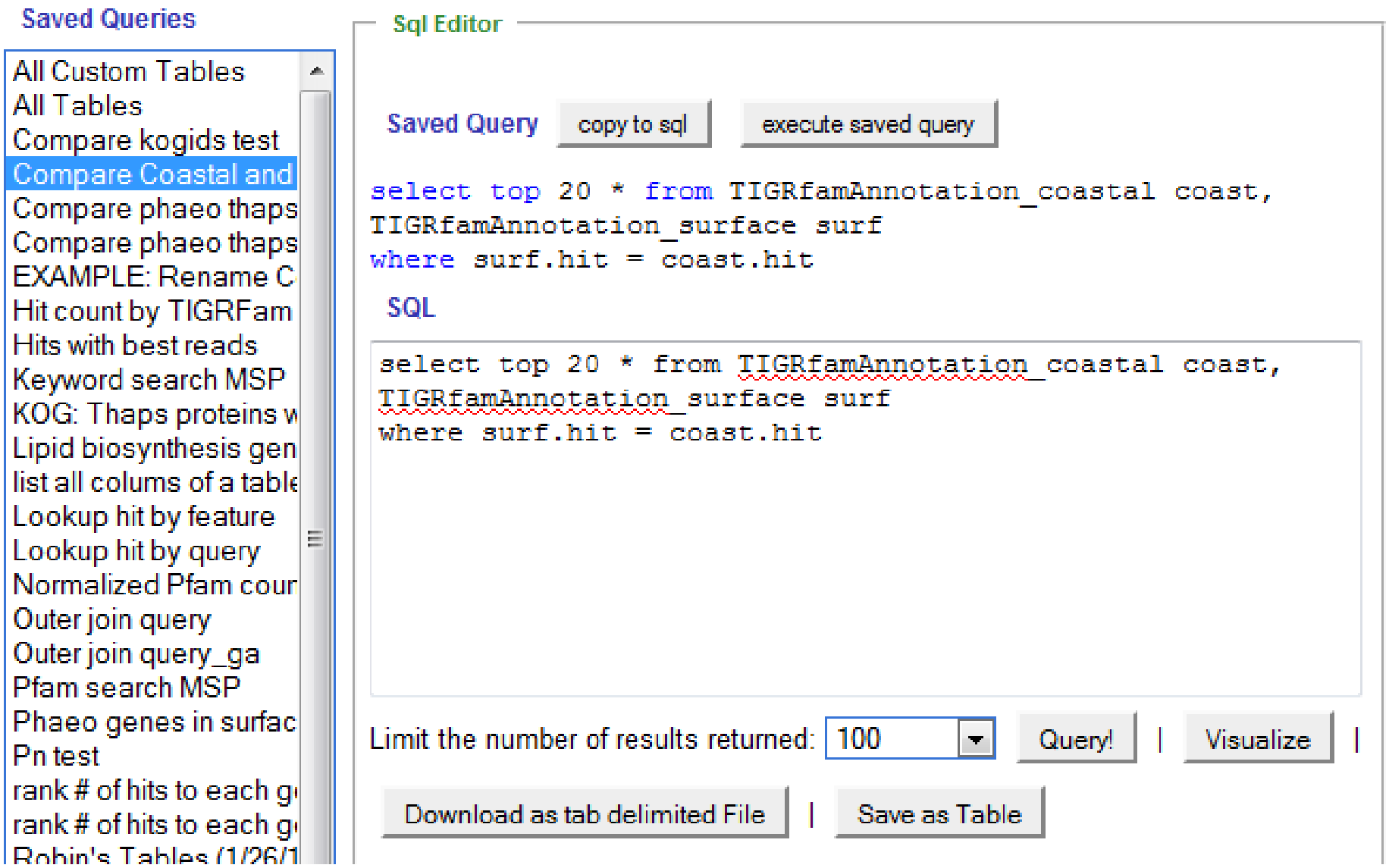
correlate diversity w/environment?

correlate diversity w/nutrients?

find new taxa and their distributions?

find new genes?

compare meta*omes?


Saved Queries
All Custom Tables
All Tables
Compare kogids test
Compare Coastal and
Compare phaeo thaps
Compare phaeo thaps
EXAMPLE: Rename C
Hit count by TIGRFam
Hits with best reads
Keyword search MSP
KOG: Thaps proteins w
Lipid biosynthesis gen
list all colums of a table
Lookup hit by feature
Lookup hit by query
Normalized Pfam cour
Outer join query
Outer join query_ga
Pfam search MSP
Phaeo genes in surfac
Pn test
rank # of hits to each g
rank # of hits to each g
Robin's Tables (1/26/1
Sql Editor
Saved Query
copy to sql
execute saved query
select top 20 * from TIGRfamAnnotation_coastal coast,
TIGRfamAnnotation_surface surf
where surf.hit = coast.hit
SQL
select top 20 * from TIGRfamAnnotation_coastal coast,
TIGRfamAnnotation_surface surf
where surf.hit = coast.hit
Limit the number of results returned:
100
Query!
|
Visualize
|
Download as tab delimited File
|
Save as Table

Your query generated 100 result(s)

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

TABLES
All tables
Favorites
Recently viewed
Shared with you

TAGS
testing
Folder10
TIGR
customers

TIGRfamAnnotation_surface

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse vitae felis vel lorem dapibus laoreet nec a leo. Nam fermentum rhoncus laoreet. Donec sit amet turpis nisl, non sagittis eros. Aliquam erat volutpat. Morbi porttitor faucibus urna quis faucibus. Suspendisse volutpat viverra elementum.

Query

```
INSERT INTO customers(customer_id, customer_name)
SELECT cus_key, cus_name
FROM jimmyscustomers WHERE customer_name LIKE 'B%'
```

AVAILABILITY: Private TAGS: Testing DATA_TIGR

DOWNLOAD VISUALIZE DELETE SAVE NEW FROM QUERY

Previous < 1 2 3 4 5 6 7 8 9 10 > Next

Rows **1 - 20** of **45694**

name	object_id	principal_id	schema_id	parent_object_id	type	type_desc	create_date	modify_date	is_ms_shipped	is_published	is_schenna_published
new.csv	32719169		19	0	U	USER_TABLE	7/8/2010 10:58:35 PM	7/8/2010 10:58:35 PM	False	False	False
test.csv	96719397		19	0	U	USER_TABLE	7/13/2010 12:21:44 AM	7/13/2010 12:21:44 AM	False	False	False
test_upload1_txt	165575628		6	0	U	USER_TABLE	4/2/2010 11:39:54 PM	4/2/2010 11:39:54 PM	False	False	False

Usage

- about 5 months old
- 8 labs around UW campus
- ~200 tables
- ~400 views

Implementation

- Windows Azure app serves GUI and RESTful API for uploading data, saving queries
- SQL Azure Database
- SQL Server on AWS to spill over 50GB and manage distributed query
- shared database, separate schemas per account
- Accounts 1:1 with DB roles

View Semantics and Features

- "Saved query" = View with attached metadata
- Unify views and tables as "datasets"
  - **table = "select * from [raw_table]"**
- Replacement semantics for name conflicts
  - **old versions materialized and archived**
- Materialize downstream views
  - **when dependencies deleted**
  - **when dependencies become incompatible**
- Permissions
  - **public vs. private vs. ACLs vs. groups**
- Sharing, social querying, CQMS*
  - **search, recent queries, friends' queries, favorites, ratings**
  - **facilitate sharing and recommendations of not just whole queries, but common predicates, join patterns, etc.**
- Discover and expose implicit relationships between datasets
  - **View synthesis [Garcia-molina, Widom, ICDT 2010]**
  - **Proactively create views for potential joins, unions, filters**

* [Khoussainova, CIDR 2009]

SQLShare as a Research Platform

- SQL Autocomplete — Logs -> Snippets
  - (Nodira Khoussainova, YongChul Kwon, Magda Balazinska)
- English to SQL — English -> Snippets
  - (Bill Howe, Luke Zettlemoyer, Shaminoo Kapoor)
- Automatic Mashups and Visualization
  - (Bill Howe, Alicia Key)
- Semi-Automatic Logical Design — Schema, Data -> Snippets
  - Join, Union Recommendations (Bill Howe, Garret Cole)
  - View Synthesis: Find Q given result R and database D s.t. R = Q(D)
- Crowdsourced SQL authoring — Crowd -> Snippets
- Information Extraction — Raw Data -> Snippets