



DATA DRIVEN COLLABORATION: A SOCIAL CONSTRUCT FOR MULTIDISCIPLINARY SCIENTIFIC DISCOVERY

Carl Kesselman
carl@isi.edu

Dean's Professor, Industrial and Systems Engineering
Director: Informatics Systems Research Division, Information Sciences Institute
University of Southern California



Acknowledgments

- Many, many conversations with John Seely Brown (jsb)
 - See: Design Unbound: Designing for Emergence in a White Water World, Brown and Pendleton-Jullian, MIT Press
- Rob Schuler
- Karl Czajkowski
- Hongsuda Tangmunarunkit



What does it mean to have a scientific result

- Others have to “know” about it
 - What is the scope and scale of “knowing”
- Others have to be able to validate it
 - Reproduce the method and achieve the same result
 - Achieve the same result via a different method
 - Reuse the result in a new method

“Non-reproducible single occurrences are of no significance to science.”

Karl Popper, 1959. *The logic of scientific discovery*. Hutchinson, London, United Kingdom.



How do we create a scientific result

Any time scientists disagree, it's because we have insufficient data. Then we can agree on what kind of data to get; we get the data; and the data solves the problem. Either I'm right, or you're right, or we're both wrong. And we move on. That kind of conflict resolution does not exist in politics or religion.

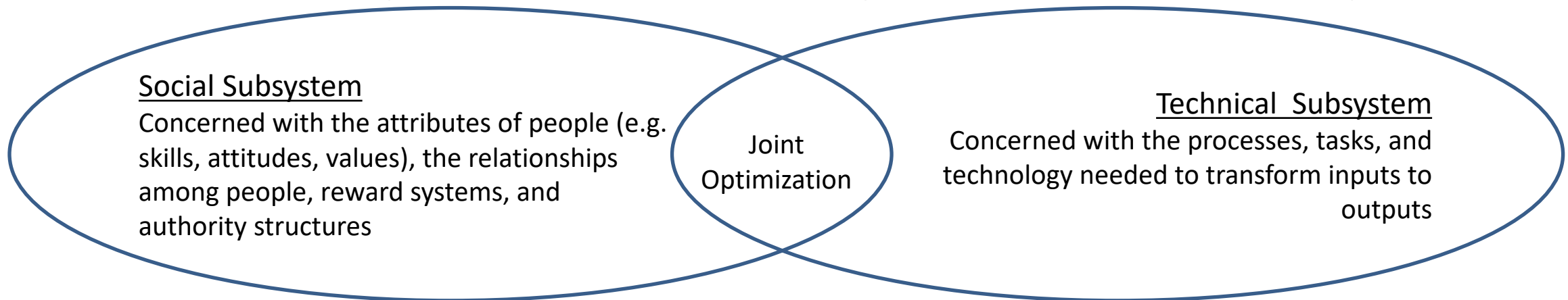
Neil deGrasse Tyson

- Science is about communities arguing over data
 - How do those communities form
 - How do communities argue: knowledge capture and communication

Data driven multidisciplinary collaboration is a adaptive socio-technical eco-system



- How do manage social and technical subsystems
- How do we optimize across the ecosystem across time and space



Man-computer symbiosis ... aims are 1) to let computers facilitate formulative thinking, and 2) to enable men and computers to cooperate in making decisions and controlling complex situations without inflexible dependence on predetermined programs.

Licklider, 1960

Militello, e. al.. (2013). Sources of variation in primary care clinical workflow: Implications for the design of cognitive support. Health informatics journal.

Top down or bottom Up: The TIMⁿ perspective



- Tribes
- Institutions
- Markets
- Networks

Ronfeldt, David, Tribes, Institutions, Markets, Networks: A Framework About Societal Evolution. Santa Monica, CA: RAND Corporation, 1996.

Traditional eScience projects look like markets built around episodic exchange of papers/data (publication)

Transformational science requires networks.



Community of practice

A group of people....

- With diverse viewpoints, role, etc.
- Engaged in joint work
- Over a significant period of time ,
- In which they build things, solve problems, learn , invent, and negotiate meaning,
- And evolve a way of talking and reading each other!

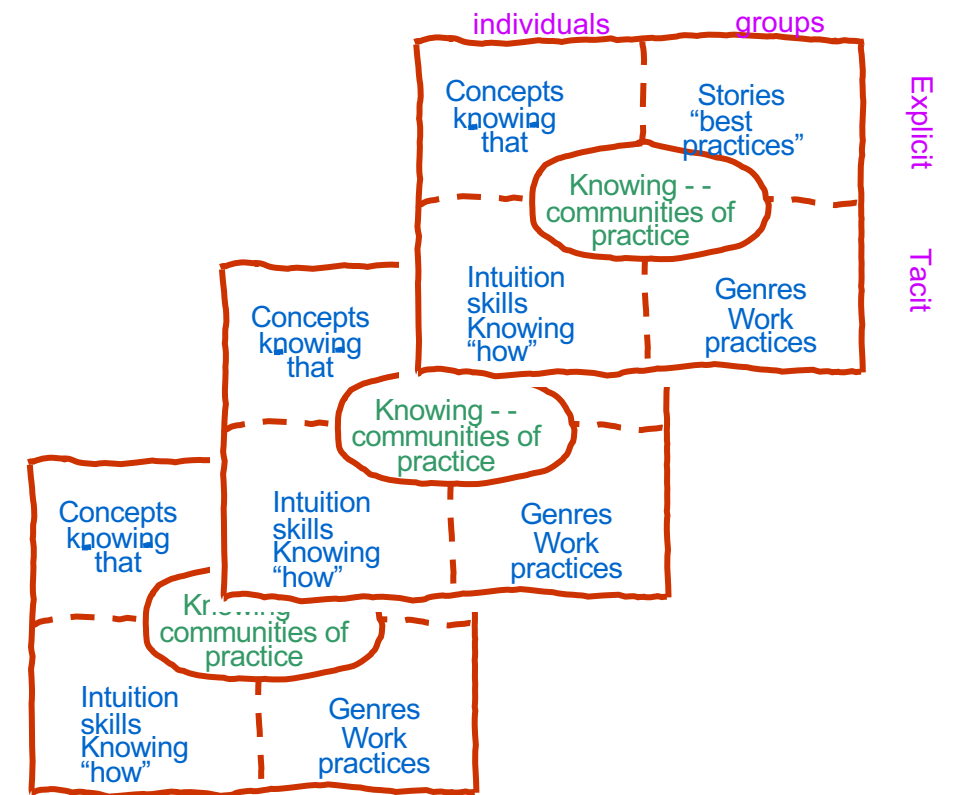


Networks of practice for transformational science

- Multiple communities working together in integrated, dynamic, adaptive and open collaboration

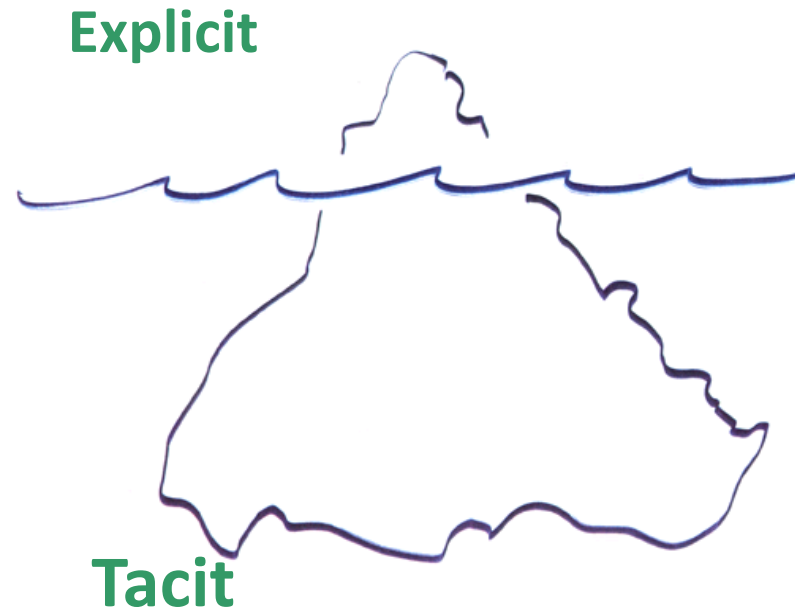


VS.





Dimensions of knowledge (jsb)



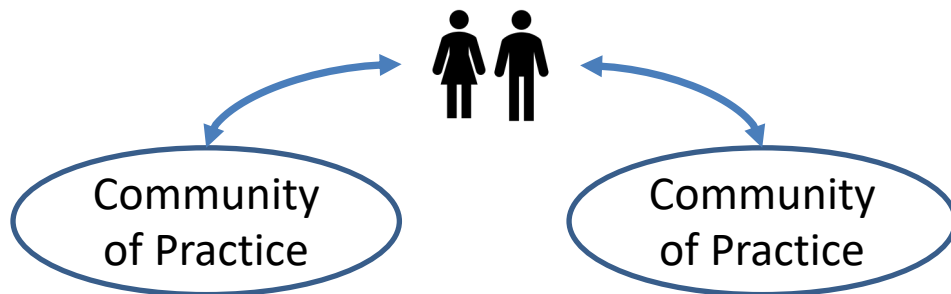
- Learning as enculturation into a practice - learning to be.
- the tacit dimension can't be completely converted to explicit
 - but some of it can - consider the doing of science.



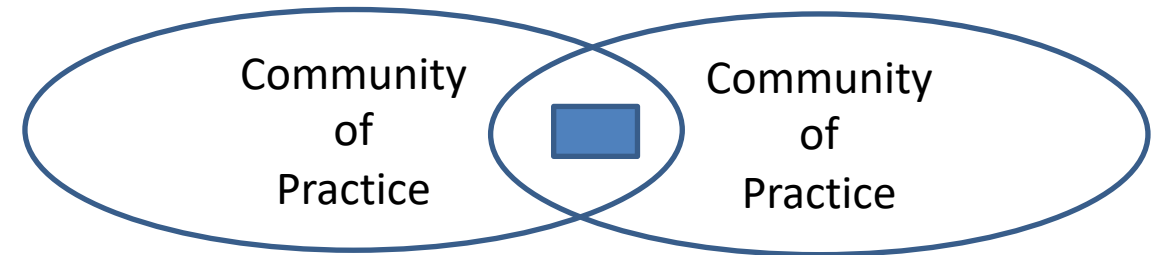
Knowledge flow across communities of practice

- Boundary objects and knowledge brokers

Translator/Knowledge Broker



Boundary object



Information used in different ways by different communities. Boundary objects are plastic, interpreted differently across communities but with enough immutable content to maintain integrity.

Star, Susan Leigh, and James R. Griesemer. "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." *Social Studies of Science* 19, no. 3



Data as a boundary object

- Must be findable by network of practice,
- Must be accessible across the network while following norms of the community
- Must be interpretable by other community (interoperable)
 - The bits
 - The terms we use to describe
 - The relationships between elements

Where have we seen these requirements before.....



FAIR data...

- Findable
 - identified by a unique identifier, characterized by rich metadata
- Accessible
 - standard protocol with access control, metadata accessible even when the data is not,
- Interoperable
 - by standardized terms to describe it
- Reusable
 - Accurate and relevant attributes.



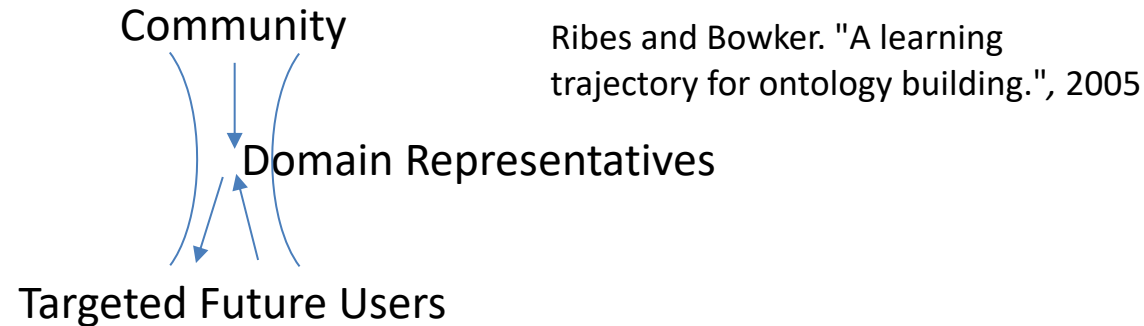
FAIR collaboration

- Data as boundary object → Create FAIR data at every part of a scientific investigation
 - Enable “long tail” science organized around data
- Requires accurate descriptions of data
 - Characteristics of data element
 - Relationships between data elements
- Question: How do we create and maintain these metadata and relationships in a collaborative environment?

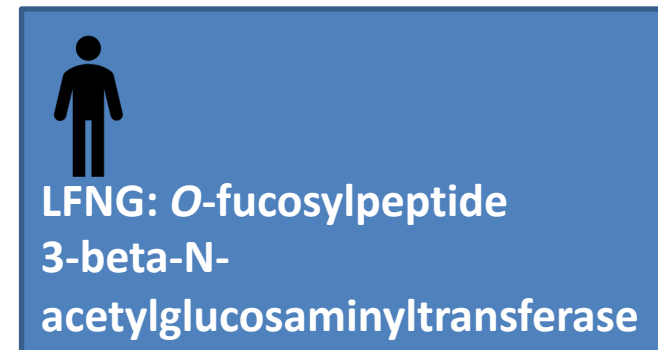


Collaboration is an evolutionary process

- Data interoperability, e.g. structure and meaning is time-dependent



- Any evaluation of FAIRness is ill-posed unless we specify community and point in time!





Tacit Knowledge in data driven collaboration

- Data captures core knowledge of community/network of practice
 - Explicit knowledge
- Tacit knowledge required as well
 - Workflow systems
 - “Web” and interaction analytics

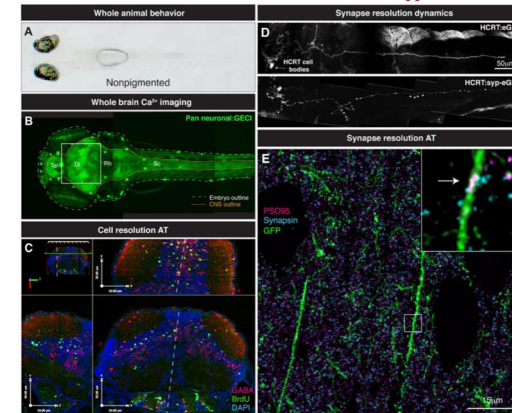
Difficulty of managing large, complex collections of data throughout research activities



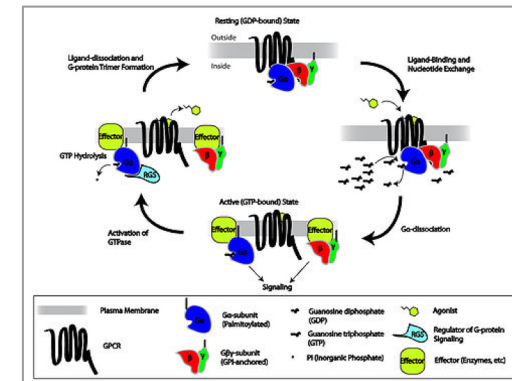
“When the Sloan Digital Sky Survey started work in 2000, its telescope in New Mexico collected more data in its first few weeks than had been amassed in the entire history of astronomy.” (Economist 2010)

“Large amounts of data are generated using a variety of innovative technologies and the limiting step is accessing, searching and integrating this data.” (Claus, Underwood, 2002)

- 50% or more time spent on **data wrangling**; (Kandel et al 2011)
- threaten **validity** of results and (10%) **reproducibility**; (Begley, Ellis 2012)
- and sparsity of **sharing**; (Borgman 2011)



Synaptome:
complex, in-vivo,
longitudinal experiment



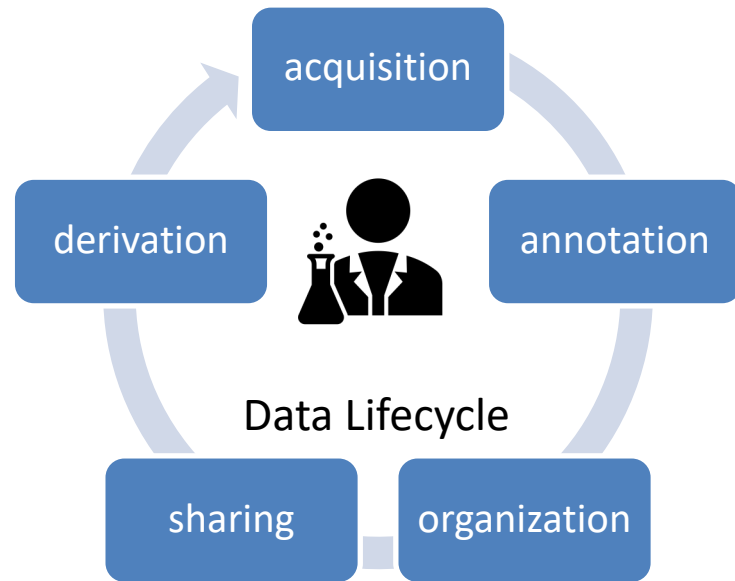
GPCR: pharma drug
discovery

Turn scientists into active participants throughout the data lifecycle



~~tossing it over the fence~~

✓ “self serve” data curation



REPLACE:

- spreadsheets,
- “meaningful” filenames,
- inefficient EAV models,...

WITH:

- well-formed relational models...

but... how to evolve model?

Driven by
deriva
deriva.isi.edu

Web Interfaces

Relational Database

(Schuler, Czajkowski, Kesselman 2014; Schuler, Czajkowski, Kesselman 2016; Bugacov et al 2017)



Scientific asset management system

Discovery Environment for Relational Information and Versioned Assets (DERIVA)

- Data Driven Collaboration
 - Think of it as a “photo manager” for scientific data
 - Maintain FAIR data from creation through publication
- Captures data and relationships between data
 - Catalog to capture relationships between data
 - Object store to hold data
- Can rapidly change to follow changes in scientific knowledge



DERIVA promotes FAIR data production

- F: providing rich metadata using an Entity-Relationship model to express relationships between diverse data elements;
- A: offering rich access control and access to metadata via standard HTTP web service interfaces;
- I: integrating with standardized terms defined by collaborators, consortium or communities; and
- R: supporting dynamic model evolution so that the data presented accurately represents the current structure and state of knowledge within an investigation.



The “20 questions” approach, gets it started, but we need more agile methods throughout formative phases

However...

“Database technology has had limited uptake [in science and data science], in part due to the overhead in designing a schema... Changing data and changing requirements make it difficult to amortize these upfront costs...”

(Jain et al 2016)

- Drive database and system design by ~20 key queries

(Gray, Szalay et al 2009)

The 20 Queries

Q1: Find all galaxies without unsaturated pixels within 1' of a given point of ra=75.327, dec=21.023

Q2: Find all galaxies with blue surface brightness between and 23 and 25 mag per square arcseconds, and -10<super galactic latitude (sgb) <10, and declination less than zero.

Q3: Find all galaxies brighter than magnitude 22, where the local extinction is >0.75.

Q4: Find galaxies with an isophotal surface brightness (SB) larger than 24 in the red band, with an ellipticity>0.5, and with the major axis of the ellipse having a declination of between 30' and 60' arc seconds.

Q5: Find all galaxies with a deVaucouleurs profile (rⁿ falloff of intensity on disk) and the photometric colors consistent with an elliptical galaxy. The deVaucouleurs profile

Q6: Find galaxies that are blended with a star, output the deblended galaxy magnitudes.

Q7: Provide a list of star-like objects that are 1% rare.

Q8: Find all objects with unclassified spectra.

Q9: Find quasars with a line width >2000 km/s and 2.5<redshift<2.7.

Q10: Find galaxies with spectra that have an equivalent width in H α >40Å (H α is the main hydrogen spectral line)

Q11: Find all elliptical anomalous

Q12: Create a grid over 60<dec<70 on a grid of same grid.

Q13: Create a color index which satisfies && r<21.75

Q14: Find stars with magnitude >15 and secondary colors

Q15: Provide a list of asteroids.

Q16: Find all objects with 5.5<redshift<6.5

Q17: Find binary colors of a galaxy

Q18: Find all objects that have velocity ratios u-g, g-r, r-i, i-z, and z-f

Two kinds of SDSS data in an SQL DB

(objects and images all in DB)

- 100M Photo Objects ~ 400 attributes

400K Spectra with ~30 lines/spectrum

Also some good queries at:
http://www.sdss.jhu.edu/ScienceArchive/sxqt/sxQT/MS3/MS3_queries.html

<http://jimgray.azurewebsites.net/talks/SciData.ppt>



- “Everyone starts with the same schema: <stuff/>. Then they refine it.”
– J. Widom

Suboptimal schema leads to suboptimal data: e.g., “messy” data!



GPCR Data Explorer

Selected by:
Clear All Filters Site **USC** Biomass Date **2012-03-01 to 2016-0...**
Current Status **completed**

Showing **1-3 of 3** results, sort by:

Biomass ID	Uniprot Name	Biomass Date	Surface Elc Percent	Total Expressing Percent	Volume
IMPT-17887	MTR1A	2015-03-24			3000
IMPT-18239	OPRK	2015-04-09			3000
IMPT-18290	ADRB2	2015-04-10			3000

Switch view:

carlkesselman@globusid.org Logout

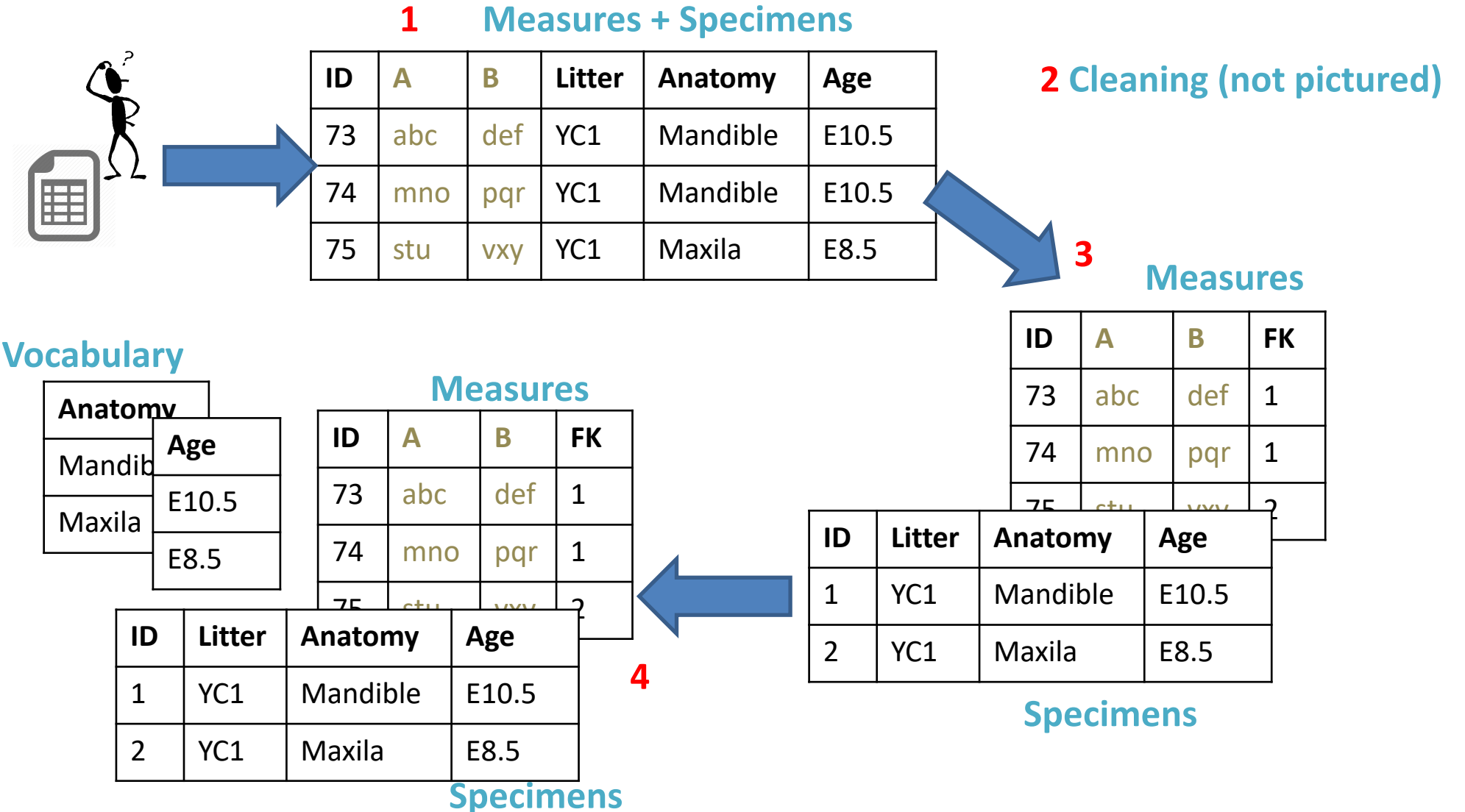
complete Complete completed completed Completed Completed Completed Completed completed, contaminated inprgress inprogress inprogress Inprogress InProgress In Progress Inrpogress requested requested Requested

This happened

- complete
- Complete
- completed
- completed
- Completed
- Completed
- Completed
- Completed
- completed, contaminated
- inprgress
- inprogress
- inprogress
- InProgress
- In Progress
- Inrpogress
- requested
- requested
- Requested



Numerous steps to evolve even “simple” database



Open challenges for schema evolution remain...



Current situation:

1. Coordination of numerous SQL operations
2. Data migration, decoupled
3. Operations outside scope of SQL

(non-trivial manual effort, human error)

What scientists need are:

1. Operations closer to the scientific domain (fewer and less complex)
2. Streamlined operations (for transforming data and schema)
3. Seamless use of general-purpose languages for specialized transforms
4. Efficient expression evaluation

Schema evolution is ... Transformation of a database schema that preserves instance data

Requirements distilled from reported and observed uses of databases in science



1. Define and alter tables of domain concepts
2. Create or change relationships between domain concepts
3. Capture categories of domain concepts (combine or separate sets)
4. Partition or merge tables of domain concepts (normalize/denormalize)
5. Express new concepts that were embedded in others (reify concepts)
6. Integrate data from external sources (external DBs, spreadsheets,...)
7. Increase the semantic coherence of data (align values, new domains,...)

Emerging “Database Evolution Languages” help, but the level of abstraction is still close to SQL



- Database Evolution Language (DEL)
 - *Schema Modification Operators* (SMOs) – the operations of the DEL
 - Recent Examples: PRISM, BIDEL
- Primary contribution
 - schema *versioning* (backward compatibility) for enterprise and web information systems
 - Do not provide significantly higher level of abstraction to users

Schema Modification Operators (SMO) Syntax

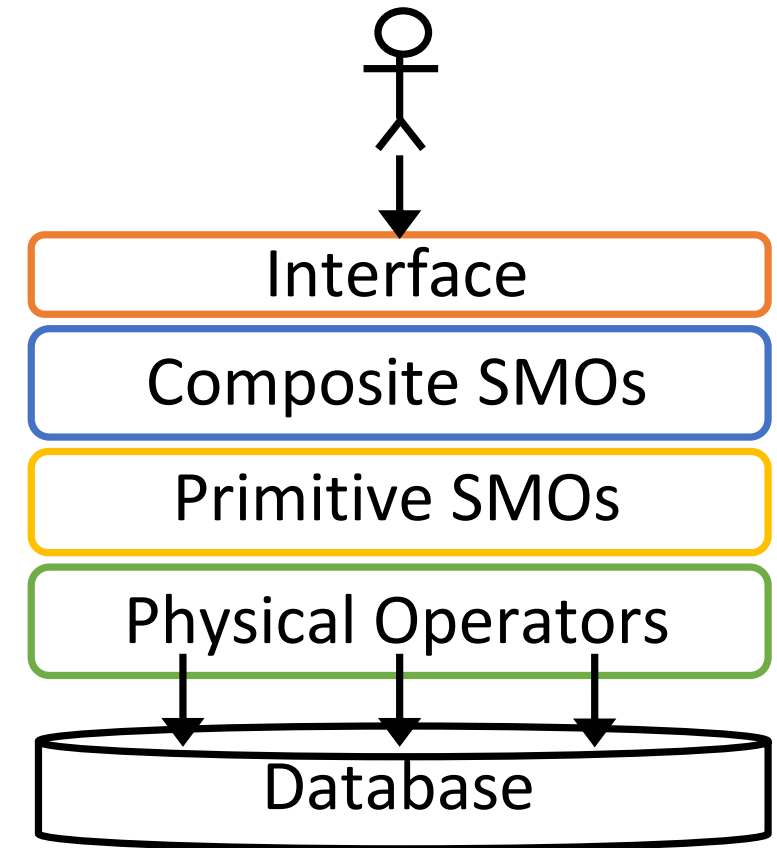
```
CREATE TABLE R (a, b, c)
DROP TABLE R
RENAME TABLE R INTO T
COPY TABLE R INTO T
MERGE TABLE R, S INTO T
PARTITION TABLE R INTO S WITH cond, T
DECOMPOSE TABLE R INTO S (a, b), T (a, c)
JOIN TABLE R, S INTO T WHERE cond
ADD COLUMN d [AS const|func(a, b, c)] INTO R
DROP COLUMN c FROM R
RENAME COLUMN b IN R TO d
```

PRISM++ (Curino, Moon, Zaniolo, Deutsch 2013)

Our approach: An open, extensible DEL based on an algebra of primitive and composite SMOs



- **Primitive SMO:** indivisible operators, extended definitions of more common relational operators
- **Composite SMO:** defined as a functional composition over other (primitive or composite) SMOs
- **Algebraic expressions** can be decomposed and *rewritten into semantically equivalent, more efficient expressions.*
- **Embed in general-purpose language** with user programs *translated into the algebra.*



Conceptual overview of the approach

Problem scenario: free text “tags” of gene names...



Before

Assays:
Table of genetic assays

ID	A	B	List_of_Genes
73	abc	def	Msx1, Foxc2, Sox9
74	mno	pqr	Tgfb3, Sox9, Bmp4, ...
...

Genes:
Table of canonical terms

ID	Term	Synonyms
1	Msx1	...
2	Tgfb3	...
3	Foxc2	...
4	Sox9	...
5	Bmp4	...
...

PROBLEM
Free text listing of gene names related to the assay records

Objective: to extract the free text “List_of_Genes” in a normalized and semantically aligned association with the canonical term set.



Problem scenario: free text “tags” of gene names...



Assays:

Table of genetic assays

ID	A	B	List_of_Genes
73	abc	def	Msx1, Foxc2, Sox9
74	mno	pqr	Tgfb3, Sox9, Bmp4, ...
...

Objective: to extract the free text “List_of_Genes” in a normalized and semantically aligned association with the canonical term set.

Genes:
Table of canonical terms

ID	Term	Synonyms
1	Msx1	...
2	Tgfb3	...
3	Foxc2	...
4	Sox9	...
5	Bmp4	...
...

Assay_Genes:
Table of semantically aligned assay to gene name association

Assay	Gene
73	Msx1
73	Foxc2
73	Sox9
74	Tgfb3
74	Sox9
74	Bmp4
...	...

Tagify: normalize and semantically align free form tags...



d Tagify _{a} r

Iteratively apply composite SMO definitions until we have an expression of only primitive SMOs

Input:

d : Genes

r : Assays

Parameters:

a : List_of_Genes

(others omitted for brevity)

Result:

Computes association table

Tagify: normalize and semantically align free form tags...



$$d \text{ Tagify}_a r \mapsto d \text{ Align}_a (\text{Atomize}_a r)$$

(Tagify definition)

Input:

d: Genes

r: Assays

Parameters:

a: List_of_Genes

(others omitted for brevity)

Result:

Computes association table

Tagify: normalize and semantically align free form tags...



$d \text{ Tagify}_a r \mapsto d \text{ Align}_a (\text{Atomize}_a r)$

(Tagify definition)

$\mapsto d \text{ Align}_a (\mu (\text{Reify}^{\text{Sub}}_a r))$

(Atomize definition)

Input:

d : Genes

r : Assays

Parameters:

a : List_of_Genes

(others omitted for brevity)

Result:

Computes association table

Tagify: normalize and semantically align free form tags...



$$\begin{aligned}d \text{ Tagify}_a r &\mapsto d \text{ Align}_a (\text{Atomize}_a r) && \text{(Tagify definition)} \\ &\mapsto d \text{ Align}_a (\mu (\text{Reify}^{\text{Sub}}_a r)) && \text{(Atomize definition)} \\ &\mapsto d \text{ Align}_a (\mu (\pi_{\text{key}(r),a} r)) && \text{(ReifySub definition)}\end{aligned}$$

Input:

d : Genes

r : Assays

Parameters:

a : List_of_Genes

(others omitted for brevity)

Result:

Computes association table

Tagify: normalize and semantically align free form tags...



$$\begin{aligned}d \text{ Tagify}_a r &\mapsto d \text{ Align}_a (\text{Atomize}_a r) && \text{(Tagify definition)} \\ &\mapsto d \text{ Align}_a (\mu (\text{Reify}^{\text{Sub}}_a r)) && \text{(Atomize definition)} \\ &\mapsto d \text{ Align}_a (\mu (\pi_{\text{key}(r),a} r)) && \text{(ReifySub definition)} \\ &\mapsto Q_{t/a} (\pi_{-a,-s} (\mu (\pi_{\text{key}(r),a} r)) \bowtie_{\equiv} d) && \text{(Align definition)}\end{aligned}$$

Input:

d : Genes

r : Assays

Parameters:

a : List_of_Genes

(others omitted for brevity)

Result:

Computes association table



Collaborative Metadata Management

- ERMRest: A web service for managing metadata and relationships between data assets
 - Everything is a resource with a URI and interface
- Entity/Relationship modeling



Data API

- All operations mapped into URL:
`ermrest/catalog/1@Y-1Z4B/entity/foo/bar/x::gt::7`
 - Queries can be viewed as naming data
 - Tables exchanges as CSV or JSON
- Operate on whole or partial entity, projections, aggregates
- Joins, filter-based row selection, column projection, configurable sort order, and pagination.
- Insertion, update, or deletion of data in one table at a time.



Fine grain access control

- ACL = (resource, permission, roles)
 - schema, table, column, or reference
 - access permission (e.g., insert, update, etc.)
 - a set of user or group identity
- ACL is associated with a resource in the hierarchy:
 - catalog, schema, table, column, and constraint
- ACLs are inherited simplifying management
- Static and dynamic policy specifications



Snapshots and history

- Logical snapshots at every catalog mutation
 - Prior values are stored in the catalog
 - Snapshot ID created for every snapshot
- Catalog URI can include snapshot ID
 - `ermrest/catalog/1@Y-1Z4B/entity/foo/bar/x`
- History Management
 - Range Discovery, Truncation, Policy Amendment, Data Redaction



Persistent identifiers

- Every entity ERMRest has unique ID (RID)
 - Assigned by catalog, can be used as foreign-key
- Every version of entire catalog has a unique ID.
- Query with snapshot ID uniquely names a entity set
 - ERMRest URI with snapshot ID
- RID and Snapshot ID uniquely name an specific version of entity
- Support for identifiers from community controlled vocabulary
 - E.g. Uberon, Schema.org, RRIDs



Evolving applications with schema

- Hard problem in general
 - In practice, schema are broken to accommodate applications.
- Pragmatic solutions:
 - Loose coupling around interchange format
 - Model introspection and dynamic interfaces



Big data bags

- A packaging format for encapsulating
 - Payload: arbitrary content
 - Tags: metadata describing the payload
 - Checksums: supports verification of content

Chard, Kyle, et al. "I'll take that to go: Big data bags and minimal identifiers for exchange of large, complex datasets." *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016

```
Bio_data_bag/  
|-- data  
| \-- genomic  
|   \-- 2a673.fastq  
|     \-- 2a673.fastq  
|   -- manifest-md5.txt  
|       afbfa23123bfa data/genomic/2a673.fasta  
|   -- bagit.txt  
|       Contact-Name: John Smith
```



Making Bags big

- Manifest lists all content and checksums
- Available content contained in “data” directory
- Content may be “missing”
 - Missing content must be listed in “fetch.txt”
 - Fetch entries list local name in data directory, and URL of where to fetch data
- Have created tool for creating, validating and materializing bags



Rich metadata for Bags

Open Knowledge Foundation Table Schema

```
{
  "name": "Method",
  "schema": {
    "fields": [
      {
        "name": "id",
        "title": "A globally unique ID",
        "type": "string",
        "constraints": { "required": true, "unique": true }
      }
    ],
    "missingValues": [ "" ], "primaryKey": "id"
  }
},
```

Research Objects

```
{
  "@context": {
    "@vocab": "http://purl.org/dc/terms/",
    "dcmi":
      "http://purl.org/dc/dcmitype/Dataset"
  },
  "@id": "../..../data/numbers.csv",
  "@type": "dcmi:Dataset",
  "title": "CSV files of beverage
consumption",
  "description": "A CSV file listing the number
of cups consumed per person."
}
```

Chaise – An adaptive model-driven Web user interface



- How little can we assume?
 - discovery, analysis, visualization, editing, sharing and collaboration over tabular data (ERMRest).
- Makes almost no assumptions about data model
 - Introspect the data model from [ERMrest](#).
 - Use heuristics, for instance, how to flatten a hierarchical structure into a simplified presentation for searching and viewing.
 - Schema annotations are used to modify or override its rendering heuristics, for instance, to hide a column of a table or to use a specific display name.
 - Apply user preferences to override, for instance, to present a nested table of data in a transposed layout.



Management of all scientific assets

▼ Protocol

Search

- PrcDsy20160101A
- PrcDsy20171030A
- PrcDsy20171030B
- PrcDsy20170613A
- PrcDsy20170613B
- PrcDsy20170615A

Show Details ...

► Nucleic Classifier

► Behavior (Learned?)

► Std. Len.

► Subject Issue Date

▼ Subject

Search

- ZIDsy20180117A
- ZIDsy20180119A
- ZIDsy20180119B
- ZIZdu20180116A
- ZIZdu20180116B
- ZIZdu20180117A
- ZIDsy20180103A
- ZIDsy20180103B
- ZIZdu20180102A
- ZIZdu20180102B

Show Details ...

► Image 1

► Image 2

► Nucleic Region 1

► Nucleic Region 2

Search

4 Items per page ▼

Displaying 4 of 71 Records

Actions	ID ↑↓	Plot	Download ↑↓
	ID StdV04 Issued 2018-01-16 Region₁ NuclmgZIZdu20180116B3B Region₂ NuclmgZIZdu20180116B6B Step₁ PrcDsy20170615A-tp1 Std.Len.₁ 0.0041 Step₂ PrcDsy20170615A-tp2 Std.Len.₂ 0.0041 Status "aligned"		<ul style="list-style-type: none">• StdV04-n1-registered.csv• StdV04-n2-registered.csv
	ID StdTZW Issued 2017-12-08 Region₁ NuclmgZIZdu20171208A3B Region₂ NuclmgZIZdu20171208A6B Step₁ PrcDsy20170613A-tp1 Std.Len.₁ 0.0049 Step₂ PrcDsy20170613A-tp2 Std.Len.₂ 0.0049 Status "aligned"		<ul style="list-style-type: none">• StdTZW-n1-registered.csv• StdTZW-n2-registered.csv
	ID StdTZM Issued 2017-12-08 Region₁ NuclmgZIDsy20171208A3B Region₂ NuclmgZIDsy20171208A6B Step₁ PrcDsy20170613A-tp1 Std.Len.₁ 0.0044 Step₂ PrcDsy20170613A-tp2 Std.Len.₂ 0.0044 Status "aligned"		<ul style="list-style-type: none">• StdTZM-n1-registered.csv• StdTZM-n2-registered.csv
	ID StdTZC Issued 2017-12-07 Region₁ NuclmgZIDsy20171207B3B Region₂ NuclmgZIDsy20171207B6B Status "aligned"		<ul style="list-style-type: none">• StdTZC-n1-registered.csv• StdTZC-n2-registered.csv



Dashboards to track progress

Synapse Browse Create carlkesseleman@globusid.org

Classifier Status

[Download CSV](#) [Permalink](#)

Search 20 Items per page

Displaying 20 of 24 Records

Actions	Classifier	Modified	Fish	Regions	In Progress	Segmented
	William Dempsey	2018-02-05 16:47:34	85	261	5/2614S1N	173 173N
	Serina Applebaum	2018-02-02 16:09:19	30	65	2/652S	63 63S
	Kaori Watanabe	2018-01-08 17:57:51	19	46	1/461S	40 40S
		2018-02-02 16:24:06	76	232	0/232	39 36S 4N
	Phillip Richards	2018-02-02 15:12:20	25	53	2/532S	28 28S
	Austin Nguyen	2018-02-06 15:36:02	13	28	2/282S	25 25S
	Nicki Karimi-Mostowfi	2018-02-06 16:48:42	11	24	1/241S	23 23S
	Yasmin Davis	2018-01-08 17:43:14	9	25	0/25	20 16S 4N
	Emily Yang	2018-02-06 14:38:22	10	20	0/20	18 18S
	William Liu	2018-02-02 16:26:31	5	12	0/12	12 12S
	Weiguang Weng	2018-01-08 16:34:12	4	10	1/101S	8 8S
	Emma Factor	2018-01-08 16:34:12	3	6	0/6	5 5S
	Helen Jin	2018-02-06 15:21:53	2	6	1/61S	4 4S
	Matt Jones	2017-10-26 19:44:11	3	8	0/8	4 4S
	Lilit Oganessian	2018-02-05 12:01:23	2	4	0/4	2 2S
	Zhuowei Du	2017-10-02 18:05:25	4	8	2/82N	2 2N
	Benjamin Shapero	2018-01-08 17:52:38	5	9	1/91S	1 1S
	Porshad Elie	2018-01-08 17:18:41	1	2	0/2	1 1S
	Donald Arnold	2017-10-02 18:05:25	2	2	0/2	0
	Karl Czajkowski	2017-12-14 13:22:35	1	1	0/1	0

< >

Dynamic data collections with global IDs



Synapse Development Browse Create carfkesselman@globusid.org

Collection + Create Edit Download CSV Permalink

▼ Row ID

Search

16A6
VTP
W26
W96
WJA
WJE
[Show Details ...](#)

> Created

> Modified

> #Bytes

> Orig. Basename

Search 25 Items per page

Displaying 6 of 6 Records

Actions	Row ID ↑↓	Description ↓↑
	ID 16A6	This will get an ensemble plot if I had one available to upload...
	ID VTP #Bytes 2,205 MD5 a6d62e7067272de8d1ef47b2a8faf82e Download Data_VTP	This is a test that doesn't have the optional "Orig. Basename" column populated.
	ID W26 Orig. Name SynlmgZfZdu20171012B5E.synapses-only.csv #Bytes 109,371 MD5 04f299d729d4a3c5a4e79ca6e1448255 Download Data_W26 SynlmgZfZdu20171012B5E.synapses-only.csv	This is another test including original basename.
	ID W96	this is a test with no uploaded asset.
	ID WJA Orig. Name synapse-studies.zip #Bytes 5,160,401 MD5 58d7a0ef1cb5e6e30b6e7f229691aa18 Download Data_WJA synapse-studies.zip	This is a test by Carl
	ID WJE	another test

◀ ▶

Model driven data entry...



Synapse Development Browse Create carikesselman@globusid.org

Create Behavior Record

* indicates required field

Submit Data + ▼

Record Number	1
Behavior	Automatically generated
* Subject	Select a value ▼
Experimentalist	Select a value ▼
Room Date	Date: YYYY-MM-DD Time: HH:MM:SS AM Now Clear
Image Date	Date: YYYY-MM-DD Time: HH:MM:SS AM Now Clear
FScope	Select a value ▼
Image Step	Select a value ▼
Std. Len.	
Volume	
Trial Counts	
Notes	

With controlled vocabulary.



Synapse Development

Create Behavior

* indicates required fields

Record Number

Behavior

* Subject

Experimentalist

Room Date

Image Date

FScope

Image Step

Std. Len.

Volume

Trial Counts

Notes

Submit Data

+ -

AM

AM

Choose FScope

Search

25 Items per page

Displaying 15 of 15 Records

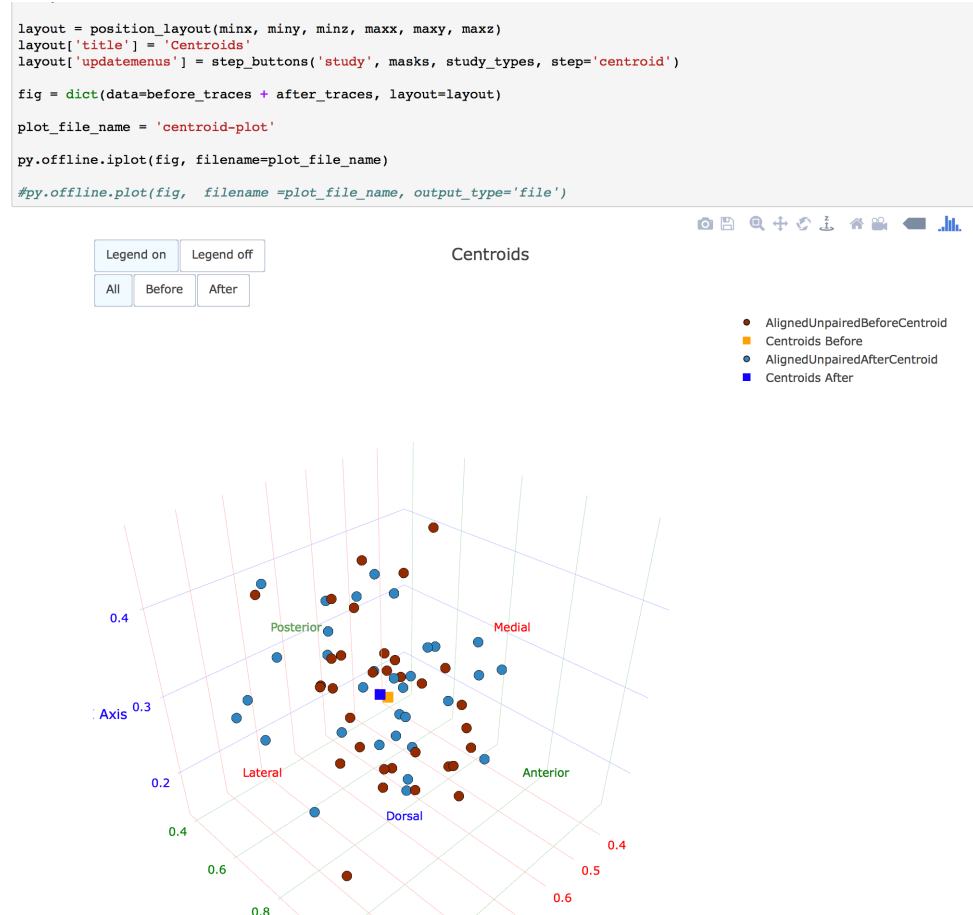
Select	ID	Term
<input checked="" type="checkbox"/>	Dummy-123	Dummy-123
<input checked="" type="checkbox"/>	FScope-1	Fish Scope 1 in ...
<input checked="" type="checkbox"/>	FScope-2	Fish Scope 2 in ...
<input checked="" type="checkbox"/>	TEMP_2017-06-22_FScope1	background level light was adjusted for "3 level" illumination (background level, level when CS is on, level when US is on); probably will need to be adjusted again
<input checked="" type="checkbox"/>	20170125_FScope1	FScope1 was re-initialized after the camera failed in January; restarted as of 2017-01-25
<input checked="" type="checkbox"/>	Temp_FScope1_newCaddy	This is the first attempt at finding parameters for FScope1 with an imaging caddy. Note that the new (and hopefully final) version of the caddy is coming soon and will differ from this one somewhat, which is why I say Temp...
<input checked="" type="checkbox"/>	Temp_FScope2_newCaddy	This is the first attempt at finding parameters for FScope2 with an imaging caddy. Note that the new (and hopefully final) version of the caddy is coming soon and will differ from this one somewhat, which is why I say Temp...
<input checked="" type="checkbox"/>	2016-06-22_FScope2	Second adjustment of background level light for "3 level" illumination (background level, level when CS is on, level when US is on)
<input checked="" type="checkbox"/>	20170125_FScope1_2x	Like 20170125_FScope1 but without the normal binning, so video is 2x2 as many pixels
<input checked="" type="checkbox"/>	20170315_FScope2	FScope2 lost its original camera and computer in January; the heating laser also turned out to be unreliable at maintaining a consistent power level throughout the training rounds so it needed to be replaced; the computer, camera, and laser were replaced; Took until March for the FScope to be operational again
<input checked="" type="checkbox"/>	2017-06-28_3LvL_FScope2	Yet another adjustment of background level light to get "3 level" illumination for better contrast (background level, level when CS is on, level when US is on)
<input checked="" type="checkbox"/>	2017-06-28_3LvL_FScope1	Adjustment of background level light to get "3 level" illumination (background level, level when CS is on, level when US is on)
<input checked="" type="checkbox"/>	20160116_FScope1	Early iteration of FScope1 collecting 800x600 movies with wide framing.
<input checked="" type="checkbox"/>	20161122_FScope2	Early iteration of FScope2 with small 284x228 movies and tighter framing.
<input checked="" type="checkbox"/>	20170620_FScope2	FScope2 was operating with a low level of background NIR light; Now, we have increased the background level at rest ("Dark" periods) from -25 intensity units to -100



RESTORED 33 STUDIES

```
In [14]: sp.get_studies()  
'AlignP2': {'x': 79, 'y': 904, 'z': 59},  
'Aligned': True,  
'Alignment': <synspy.analyze.pair.ImageGrossAlignment at 0x811a07550>,  
'AlignmentPts': {},  
'BeforeImageID': 'ImgZfDsy20160909A3',  
'BeforeURL': '/hatrac/Zf/ZfDsy20160909A/SynStd6473-s1-registered.csv:3D7ACQZ2FZIZMRC3KCTLK3PWEU',  
'Learner': False,  
'Paired': False,  
'Protocol': 'PrcDsy20160101A',  
'Provenence': {'CatlogVersion': '2P9-5Z0X-E12P',  
'GITHash': 'c877aa13709b12d701ae2afe95faeaf4944c81f3'},  
'Region1': 'SynImgZfDsy20160909A3C',  
'Region2': 'SynImgZfDsy20160909A6B',  
'Study': 'SynStd6473',  
'StudyAlignmentPts':  
      x           y           z  
0 -6.393723e-09  8.331937e-08 -3.033229e-08  
1  1.554701e+00  1.438417e+00 -4.437405e-08  
2 -4.365282e-09  1.000000e+00 -4.630136e-08,  
'Subject': 'ZfDsy20160909A',  
'Type': 'nonlearner'},
```

Web services interface can be used to pull data directly from catalog. Notice that all data assets are versioned (BeforeURL), and that we have the exact snapshot version of the data catalog and code, allowing us to perform RDA compliant persistent search.



Interaction with the notebook allows us to plot data. Plots can be placed into catalog as an asset that is part of the study.



NEW RESEARCH IN

Physical Sciences

Social Sciences

Biological Sciences

In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells



Article Alerts

Share

Email Article

Tweet

Citation Tools

Like 0

Request Permissions

Mendeley

Diya B. Joseph, Anoop S. Chandrashekar, Lisa L. Abler, Li-Fang Chu, James A. Thomson, Cathy Mendelsohn, and Chad M. Vezina

PNAS August 14, 2018 115 (33) 8394-8399; published ahead of print July 30, 2018
<https://doi.org/10.1073/pnas.1802966115>

Edited by Marianne Bronner, California Institute of Technology, Pasadena, CA, and approved July 2, 2018 (received for review February 19, 2018)

[View Full Text](#)

Article

Figures & SI

Info & Metrics

PDF

Significance

When the bladder's specialized epithelial lining is damaged by infection or injury, its own basal and intermediate cell progenitors are called upon to restore a functional barrier. Here we show that when these progenitor cells are depleted in conditional *Dnmt1* mutant mice,

More Articles of This Classification

Biological Sciences

Single-molecule force spectroscopy reveals folding steps associated with hormone binding and activation of the glucocorticoid receptor

Specific recognition of two MAX effectors by integrated HMA domains in plant immune receptors involves distinct binding surfaces

Phospholipid flippases enable precursor B cells to flee engulfment by macrophages



○ Referen

- Figures & SI
- Info & Metrics
- PDF

Methods

Data Dissemination.

To increase rigor, reproducibility, and transparency, raw image files and other data generated as part of this study were deposited into the GUDMAP consortium database and are fully accessible at: <https://doi.org/10.25548/W-QXXC> (25).

Conditional *Dnmt1* Mutants.

Mice were housed as previously described (26). All procedures performed on mice were approved by the University of Wisconsin–Madison Animal Care and Use Committee and were carried out in accordance with the Guide for the Care and Use of Laboratory Animals. *Shh*^{cre} alleles (B6.Cg*Shh*^{tm1(EGFP/cre)Cjt/J}) (11) were used to conditionally inactivate *Dnmt1* using *Dnmt1**flox* alleles (B6.129S4-*Dnmt1*^{tm2Jae/Mmucd}) in *Shh* lineage cells marked

In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells

COLLECTION

[↗ Show All Related Records](#)
[↗ Export ▾](#)
[↗ Share](#)

Contents

[▶ Main](#)
[He Slide Collection \(13\)](#)
[Specimen Collection \(25\)](#)


RID

W-QXXC

Title

In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells

Description

Figures and data relating to the PNAS 2018 paper titled “In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells” by Joseph et al.

Details

The following table shows the mapping of figures and image record IDs (RID) presented in the paper.

Figure	Reference	Additional Images
1A	W-QXXW	W-QXZ4 , W-QY2C
1B	W-QY34	W-QY38 , W-QY3C
1C	W-QY66	W-QY6T , W-QY86
1D	W-QY8Y	W-QY9A , W-QYA6
1E	W-QYB6	W-QYBP , W-QZ6T
1F	W-QYC2	W-QYCE , W-QYCT
1G	W-QYDP	W-QYDY , W-QYEA
1H	W-QYEP	W-QYF2 , W-QYFE
1I	W-QYGP	W-QYH2 , W-QYHE
1J	W-QYHT	W-QYJ6 , W-QYJJ
1K	W-QYKP	W-QYM2 , W-QYME

In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells

[↗ Show All Related Records](#)
[↗ Export ▾](#)
[↗ Share](#)

COLLECTION

Require DOI?	Yes
Persistent ID	https://doi.org/10.25548/W-QXXC
Principal Investigator	Chad Vezina
Data Provider	University of Wisconsin
Consortium	GUDMAP
Creation Time	2018-05-17 21:19:14
Last Modified Time	2018-07-09 22:13:01

Contents

- ▶ [Main](#)
- [He Slide Collection \(13\)](#)
- [Specimen Collection \(25\)](#)

▼ He Slide Collection (showing all 13 results)

[View More](#)

View	RID ↓↑	Thumbnail ↓↑	Name ↓↑	Species ↓↑	Tissue ↓↑	Age Stage ↓↑	Gender ↓↑	Image File ↓↑
	W-R01Y		Urogenital sinus from Control embryo (Shhcre/+; Dnmt1flox/+) (1 of 3)	Mus musculus	urogenital sinus	18.5 embryonic days	Male	20160826ShhcreDnmt1LOFME18.5U
	W-R02A		Urogenital sinus from Control embryo (Shhcre/+; Dnmt1flox/+) (2 of 3)	Mus musculus	urogenital sinus	18.5 embryonic days	Male	20160826ShhcreDnmt1LOFME18.5U
	W-R02P		Urogenital sinus from Control embryo (Shhcre/+; Dnmt1flox/+) (3 of 3)	Mus musculus	urogenital sinus	18.5 embryonic days	Male	20160826ShhcreDnmt1LOFME18.5U
	W-R02Y		Urogenital sinus from Conditional Dnmt1 embryo (Shhcre/+; Dnmt1flox/flox) (1 of 3)	Mus musculus	urogenital sinus	18.5 embryonic days	Male	20160826ShhcreDnmt1LOFME18.5U
	W-R036		Urogenital sinus from Conditional Dnmt1 embryo (Shhcre/+; Dnmt1flox/flox) (2 of 3)	Mus musculus	urogenital sinus	18.5 embryonic	Male	20160826ShhcreDnmt1LOFM

Whole-mount 3D views of the hum... COLLECTION

RID	Q-3K5A
Title	Whole-mount 3D views of the
Description	A collection of human embry... Related to JASN https://doi.org/10.25548/BURB-6P44
Require DOI?	Yes
Persistent ID	https://doi.org/10.25548/BURB-6P44

Record Status Detail	Complete
Curation Status	Release
Principal Investigator	Andrew McMahon, USC
Data Provider	University of Southern California
Consortium	GUDMAP
Creation Time	2017-09-23 02:18:02
Last Modified Time	2018-05-23 03:39:13

He Slide Collection (no results found)

[Add](#) | [View More](#)

Actions	RID ↑↓	Thumbnail ↑↓	Name ↑↓	Species ↑↓	Tissue ↑↓	Age Stage ↑↓	Gender ↑↓	Image File ↑↓	Record Stat
---------	--------	--------------	---------	------------	-----------	--------------	-----------	---------------	-------------

No Results Found

Specimen Collection (no results found)

[Add](#) | [View More](#)

Share X

Share Link
<https://dev.gudmap.org/chaise/record/#2/Common:Collection/RID=Q-3K5A>

Citation
 Andrew McMahon *GUDMAP Consortium* <https://doi.org/10.25548/BURB-6P44> (2017).

Download Citation:
[BibTex](#)

[Hide Empty Related Records](#) [Export](#) [Share](#)

Contents

- Main**
- He Slide Collection (0)**
- Specimen Collection (0)**
- IF Slide Collection (13)**
- IF Video Collection (0)**
- Sequencing Study Collection (0)**

In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells

COLLECTION

[+ Create](#) [✎ Edit](#) [📄 Copy](#) [🗑 Delete](#) [🔍 Hide Empty Related Records](#) [📄 Export ▾](#) [🔗 Share](#)

RID	W-QXXC
Title	In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells
Description	Figures and data relating to the PNAS 2018 paper titled "In vivo replacement of damaged bladder urothelium by Wolffian duct epithelial cells" by Joseph et al.
Details	The following table shows the mapping of figures and image record IDs (RID) presented in the paper.

Figure	Reference	Additional Images
1A	W-QXXW	W-QXZ4 , W-QY2C
1B	W-QY34	W-QY38 , W-QY3C
1C	W-QY66	W-QY6T , W-QY86
1D	W-QY8Y	W-QY9A , W-QYA6
1E	W-QYB6	W-QYBP , W-QZ6T
1F	W-QYC2	W-QYCE , W-QYCT
1G	W-QYDP	W-QYDY , W-QYEA
1H	W-QYEP	W-QYF2 , W-QYFE
1I	W-QYGP	W-QYH2 , W-QYHE
1J	W-QYHT	W-QYJ6 , W-QYJJ

Con

CSV

BAG

Main

[He Slide Collection \(13\)](#)[Specimen Collection \(25\)](#)[IF Slide Collection \(0\)](#)[IF Video Collection \(0\)](#)[Sequencing Study
Collection \(0\)](#)



Name	Size	Kind	Date Added
▼ Collection	--	Folder	Today at 12:35 AM
▶ metadata	--	Folder	Today at 12:35 AM
▼ data	--	Folder	Today at 12:35 AM
Sequenci...llection.csv	324 bytes	Comm...et (.csv)	Today at 12:35 AM
IF Video C...ection.csv	291 bytes	Comm...et (.csv)	Today at 12:35 AM
IF Slide Collection.csv	706 bytes	Comm...et (.csv)	Today at 12:35 AM
Specimen...ection.csv	176 KB	Comm...et (.csv)	Today at 12:35 AM
He Slide...llection.csv	9 KB	Comm...et (.csv)	Today at 12:35 AM
Collection.csv	9 KB	Comm...et (.csv)	Today at 12:35 AM
fetch.txt	3 KB	Plain Text	Today at 12:35 AM
manifest-md5.txt	2 KB	Plain Text	Today at 12:35 AM
tagmanifest-md5.txt	238 bytes	Plain Text	Today at 12:35 AM
bag-info.txt	299 bytes	Plain Text	Today at 12:35 AM
bagit.txt	55 bytes	Plain Text	Today at 12:35 AM



Summary

- Think of eScience infrastructure as part of socio-technical system
 - Design for the interface across communities of practice
- Validated this approach with Deriva across many domains and scales
 - Craniofacial dysmorphia, protein structure database, molecular atlas, kidney reconstruction, dynamic synapse mapping, optimization models, pancreatic beta cell modeling, developmental biology, ...
- Other platform approaches possible
 - E.g. L. Trani, M. Atkinson, D. Bailo, R. Paciello, R. Filgueira, Establishing core concepts for information-powered collaborations, FGCS 89 (2018) 421–437.
- For more information: www.derivacloud.org

