



Why Data Citation is a Computational Problem

Susan B. Davidson

University of Pennsylvania

Work partially supported by
NSF IIS 1302212, NSF ACI 1547360
NIH 3-U01-EB-020954-02S1



National Science Foundation
WHERE DISCOVERIES BEGIN

Outline

- **The power of abstraction**
 - And how it has helped with two of my favorite problems in bioinformatics
- New problem: data citation
- Bigger picture: Data Science

The power of abstraction

- The “right” abstraction is key to developing solutions to many practical problems.
 - Data Integration
 - Provenance
 -
 - Data Citation
- Developing the right abstraction requires close collaboration between end-users, systems builders, and theoreticians.



Databases meets bioinformatics

**“Genomics is the next moon landing.”
(1992)**



From left to right, Drs. Peter Buneman, Susan Davidson, and Chris Overton are trying to unravel the secrets of the DNA spiral. Photo by Derek Wong.

The map shows the University of Pennsylvania campus with a red crosshair marking a location. A portrait of a man with glasses is overlaid on the map. The University of Pennsylvania logo is in the bottom right corner.






Example 1: Data Integration

Entrez Sequence

```
>gi|2580555|gb|AF000985.1|HSAF000985 Homo sapiens dead box, Y isoform (DBY)
mRNA, alternative transcript 1, complete cds
CCAGTGAAGAGTTCCGCTATTCGGTCTCACACCTACAGTGGACTACCCGATTTTTGCTTCTCTCAGG
GATGAGTCATGTGGTGGTGAAAAATGACCCTGAACTGGACCAGCAGCTTGCTAATCTGGACCTGAACTCT
GAAAAACAGAGTGGAGGAGCAAGTACAGCGAGCAAAGGGCGCTATATACCTCCTCACTTAAGGAACAAG
AAGCATCTAAAGGATCCATGATAAAGACAGTTCAGGTTGGAGTTGCAGCAAAGATAAGGATGCATATAG
CAGTTTTGGGTCTCGAGATTCTAGAGGAAAGCCTGGTTATTTTCAGTGAACGTGGAAGTGGATCAAGGGGA
...
```

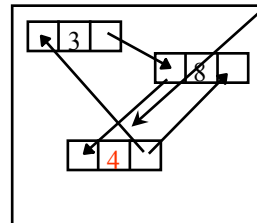
Image Data

Id	Date & Time	Image
spdfld13a	9/8/95 12:02:03	
spdfld22a	9/8/95 12:02:04	
spdfld22a	9/8/95 12:02:06	

Relational Databases

Name	P Value	Len
HT97683	0	2182
Q62167	3.1e-234	440
P16381	4.2e-230	440
P24346	4.2e-214	440
P066346	2.6e-127	423

Object-Oriented Databases



Integrating Query:

What genes are involved in bipolar schizophrenia?

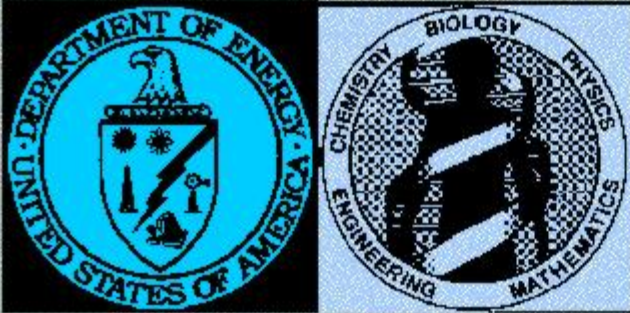
Entrez Medline

Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. *SCIENCE* Volume 282 (5396): 2012 - 2018 Issue of 11 Dec 1998
The *C. elegans* Sequencing Consortium *

The 97-megabase genomic sequence of the nematode *Caenorhabditis elegans* reveals over 19,000 genes. More than 40 percent of the predicted protein products find significant matches in other organisms. There is a variety of repeated sequences, both local and dispersed. The distinctive distribution of some repeats and highly conserved genes provides evidence for a regional organization of the chromosomes.

Array Data

1.2	3.4	5.6	7.8	9.0
3.5	6.8	9.1	2.4	5.7
8.0	7.6	5.4	3.2	1.0
1.9	2.8	3.7	4.6	5.5
7.3	8.2	9.1	0.0	1.1
6.8	9.1	2.4	5.7	8.0
7.6	5.4	3.2	1.0	9.8



Report of the Invitational DOE Workshop on Genome Informatics, 26–27 April 1993

The Office of Health and Environmental Research of the Department of Energy (OHER/DOE) sponsored a meeting in Baltimore on 26–27 April of a group of experts to assess the state of current OHER-related bioinformatics efforts and to offer advice on planning and coordinating future activities. OHER has a considerable interest in bioinformatics, in large part, because of the DOE Human Genome and the Structural Biology programs.

TOPICS DISCUSSED IN DOE INFORMATICS SUMMIT

26–27 APRIL 1993, BALTIMORE, MD

“Until a fully relationalized sequence database is available, none of the queries in this appendix can be answered.”

Why would they say that?

- Needed to pose set-oriented queries against multiple, heterogeneous databases, files, and software packages.
 - Most integration work at the time was based on the relational model
 - Embedded links in files: Clicking doesn't scale!
- Needed in-depth understanding of what data sources were available and what information they contained.

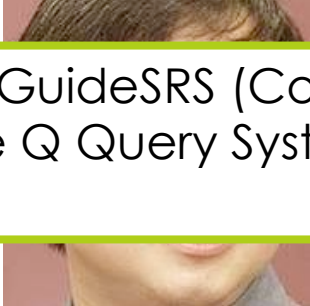
Answering the “unanswerable”

- We were able to answer the “unanswerable queries” within about a month using our data integration system, Kleisli.
- Kleisli used a complex-object model of data, language based on comprehension syntax, and optimizations that went beyond relational systems.

- **Limsoon Wong**



- Kyle Hart, Jonathan Crabtree,...
- Leonid Libkin, Dan Suciu,...

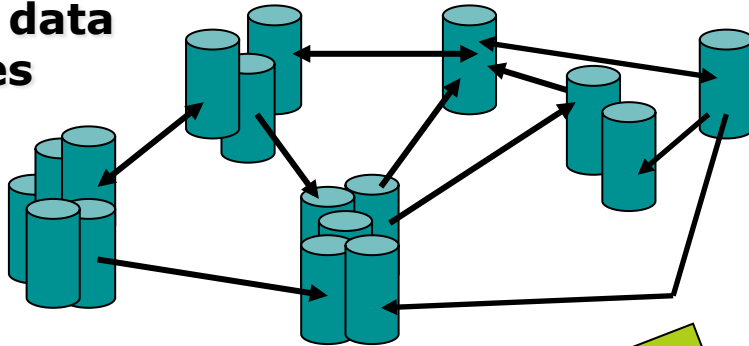


BioGuideSRS (Cohen-Boulakia)
The Q Query System (Ives)
...?



Example 2: Provenance

Public data sources



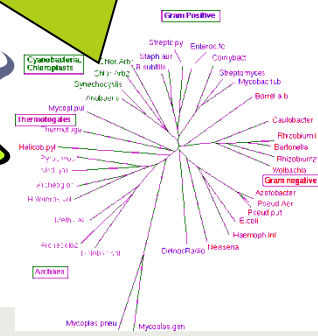
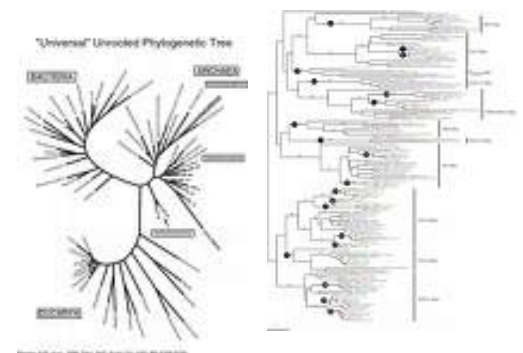
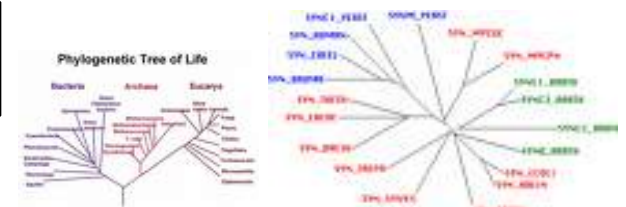
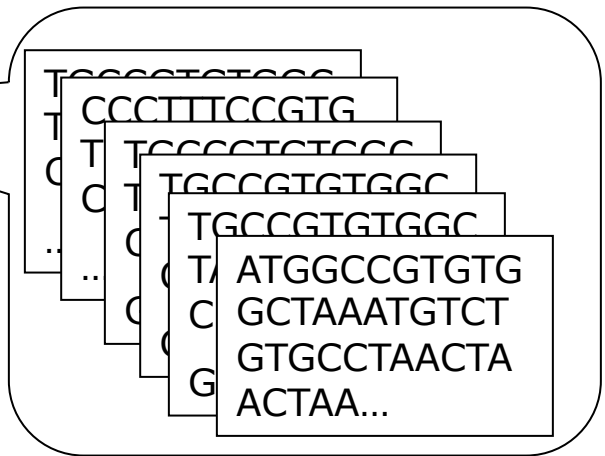
Bioinformatics protocols

Alignments
PAUPS
Bootstrap
ClustalW
Phillips
...

How this result has been generated?

Which sequences have been used to produce this result?

Which data are really important to keep?



Biologist's workspace

Different types of provenance

- “Coarse-grained” workflow provenance
 - Kepler (Ludaescher *et al.*), Pegasus (Deelman, Gil *et al.*), Taverna (Goble, Oinn *et al.*), Vistrails (Freire *et al.*),...
- “Fine-grained” database style provenance
 - Why and Where: Buneman, Khanna, Tan
 - Provenance Semirings: Tannen, Green, Karvouranakis
 - Trio: Widom, Cui, Weiner *et al.*
- Event logs (ordering, timing, causality matters)
 - Provenance-Aware Storage Systems: Seltzer *et al.*
 - Secure Network Provenance: Zhou, Loo, Haeberlen

The problem with provenance...



Continuing challenges...

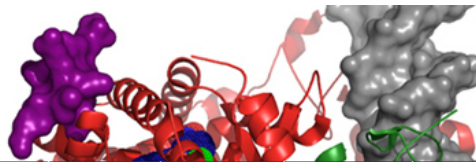
- ▣ Combining different types of provenance
- ▣ Tools to query, explore, and understand provenance
- ▣ Summarizing provenance
- ▣ Approximating provenance
- ▣ ...

Outline

- The power of abstraction
- **New problem: data citation**
 - State of the art
 - Citations for general queries
 - Building a citation system
- Bigger picture: Data Science

Publication is changing

- Information is increasing published on the web.
- Much of this information is in **curated databases** – crowd- or expert-sourced data
- These datasets are complex, structured, and evolving, and contributors need to be acknowledged



IUPHAR/BPS
Guide to PHARMACOLOGY



An expert-driven guide to pharmacological targets and the substances that act on them.

Citation: Principles and Standards

- Large number of organizations are involved, and standards are emerging: Datacite, DataONE, GEOSS, D-Lib Alliance, DCC, COPDES, Force-11, AGU, ESIP, DCMI, CODATA, ICSTI, IASSIST, ICSU...
- **Force 11:** “Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.”
- **DataCite:** “We believe that you should cite data in just the same way that you can cite other sources of information, such as articles and books.”
- **Amsterdam Manifesto:** “Data should be considered citable products of research.”

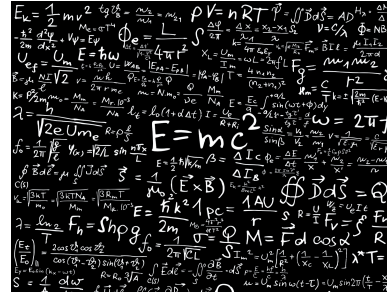
What is a (conventional) citation?

- A collection of “snippets” of information: authors, title, date, etc. and some kind of access mechanism
- Not exactly provenance
- Self contained, immutable (to within some choice of format)
- Needed for a variety of reasons: kudos, currency, authority, recognition, access...

Buneman, Davidson, Frew:
Why data citation is a computational problem,
Commun. ACM, 59(9): 50—57 (2016)

Citation goes beyond DOIs

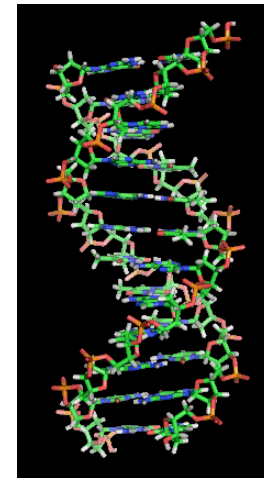
- Ann. Phys., Lpz 18 639-641



Einstein:

Does the inertia of a body depend on its energy content?

- Nature, 171,737-738



Watson and Crick:

Molecular Structure of Nucleic Acids; a structure for deoxyribose nucleic acid



State of the art in data citation ...

Example 1: eagle-i

- A “resource discovery” tool built to facilitate translational science research.
- Developed by a consortium of universities under NIH funding, headed by Harvard.
 - Penn is a member.
- End users: researchers who wish to share information about research resources (Core Facilities, iPS cell lines, software resources).
- Data is stored and distributed as RDF files (graph database).
- Resources have a “Cite this resource” button!

[Back to Search Results](#) 

Significance Tester for the Accumulation of Reads

Algorithmic software component 

[Send message to
resource contact](#)[Cite this resource](#)


University of
Pennsylvania

Software Description STAR was developed to identify regions enriched for a histone modification based on CHIP-Seq evidence, by identifying regions with a significant accumulation of reads.

Software Additional Name STAR

Used by [Computational Biology and Informatics Laboratory](#)

Contact [Grant, Gregory R., Ph.D.](#)

Related Technique [CHIP-seq assay](#) 

Software purpose DNA modification site prediction objective



Search for resources across the eagle-i Network

Go

Top Categories | Explore All

Try our new
IPS Cell Search

ABOUT GET INVOLVED NEWS + EVENTS FAQ CONTACT US HELP

[Back to Search Results](#) >

Significance Tester for the Accumulation of Reads

Algorithmic software component ⓘ

Send message to
resource contact

Cite this resource



University of
Pennsylvania

eagle-i ID for this resource:

<http://eagle-i.itmat.upenn.edu/i/0000013d-8d96-57e1-2ed7-105480000000>

Click [here](#) for citation examples and more information.

Close

Contact [Grant, Gregory R., Ph.D.](#)

Related
Technique

ChIP-seq assay ⓘ

Citing an eagle-i Resource

Citing eagle-i resources is an easy way to give credit.

The formats suggested below provide the minimum information necessary to identify and credit the resource provider, and are designed to provide a traceable, durable, and unambiguous reference for the resource being cited. These suggestions can and should be used along with those from other resource identifiers (i.e. Antibody Registry ID, Addgene, DSHB, RRID) or from the journal publishing your work.



The screenshot shows a resource page for "APP cKO x Cre ER x APLP2 KO" in *Mus musculus*. A callout box highlights the "eagle-i ID" and "eagle-i Institution" (Harvard University). The resource name and type are also highlighted. The owning organization is "Young-Pearse Laboratory".

Request this resource	Cite this resource
Organism or Virus Description	Used to study brain pathology and
Location	Young-Pearse Laboratory
Genetic alteration	APLP2 deletion APP cKO

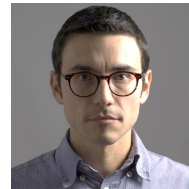
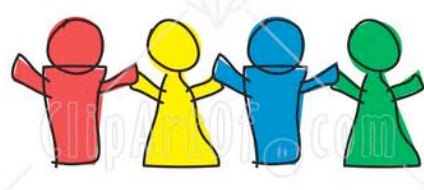
Resource Name and Type: APP cKO x Cre ER x APLP2 KO
 eagle-i ID: http://harvard.qa.eagle-i.net/i/0000012a-25bf-e274-f5ed-943080000002
 eagle-i Institution: Harvard University
 Owning Organization: Young-Pearse Laboratory

Note that for all types, the names of Core Facilities or other ambiguously named organizations should be followed by the name of the affiliated eagle-i institution in order to disambiguate them (e.g. *Flow Cytometry Core. Montana State University vs. Flow Cytometry Core. Dartmouth College*).

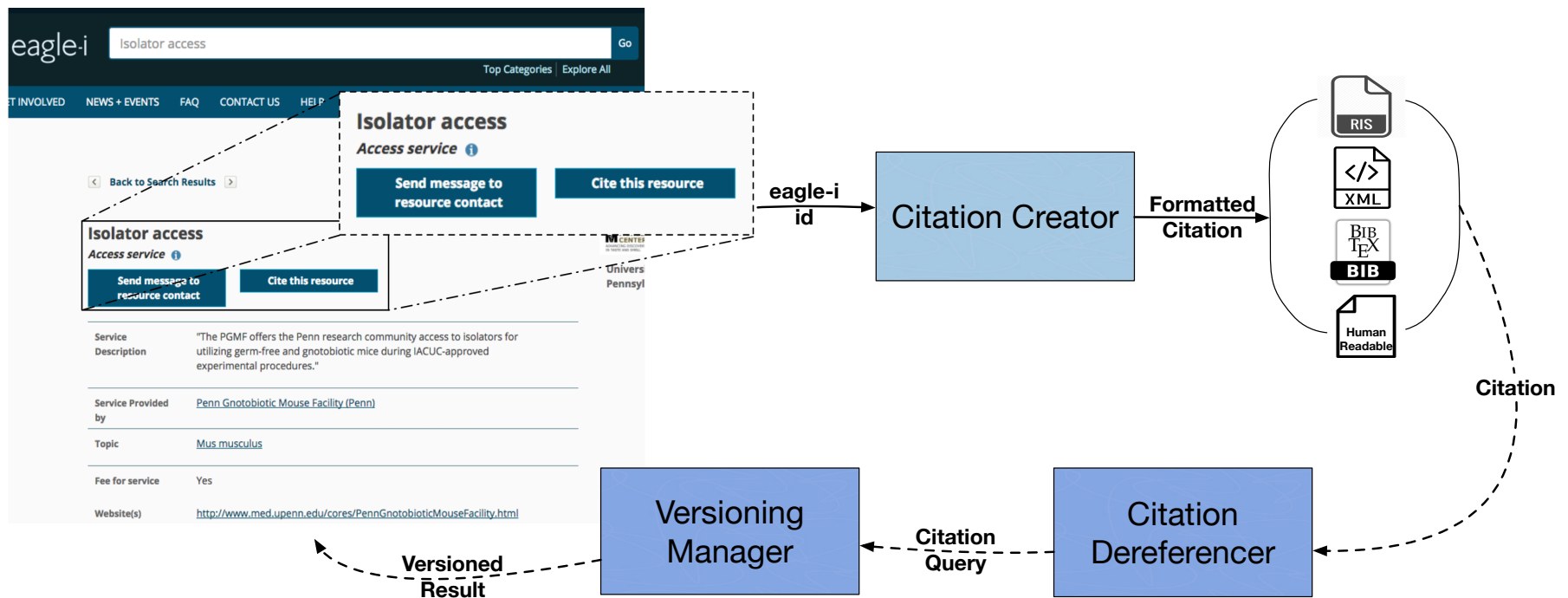
Citation Guidelines

Although only the most commonly cited types are listed below, the same rules can be used to cite any eagle-i resource.

Automating citation in eagle-i

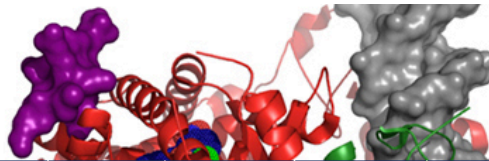


A. Alawini, L. Chen, S. B. Davidson, N. Da Silva, G. Silvello:
“Automating data citation: the eagle-I experience.” JCDL 2017.



Example 2: IUPHAR

- IUPHAR Guide to Pharmacology is a database of information about drug targets, and the prescription medicines and experimental drugs that act on them.
- Information is presented to users through a hierarchy of **web views**, with an **underlying relational implementation**.
 - Targets are arranged into groups called “families”
- Contents of the database are generated by hundreds of experts who, in small groups, contribute to portions of the database. Thus the authorship depends on what part of the database is being cited.

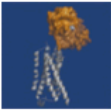

 Search database

IUPHAR/BPS Guide to PHARMACOLOGY

- Home
- About ▾
- Targets ▾
- Ligands ▾
- Resources ▾
- Advanced search ▾

An expert-driven guide to pharmacological targets and the substances that act on them.

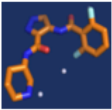
Targets



- ▶ G protein-coupled receptors
- ▶ Ion channels
- ▶ Nuclear hormone receptors
- ▶ Kinases
- ▶ Catalytic receptors
- ▶ Transporters
- ▶ Enzymes
- ▶ Other protein targets

Search for targets [GO](#)

Ligands



- ▶ Approved drugs
- ▶ Synthetic organics
- ▶ Metabolites
- ▶ Natural products
- ▶ Endogenous peptides
- ▶ Other peptides
- ▶ Inorganics
- ▶ Antibodies
- ▶ Labelled ligands

Search for ligands [GO](#)

Get email updates

Email address

Email format html text

[Subscribe](#)

What's new to Guide to PHARMACOLOGY

New version (2015.3) released 19th Oct 2015!

Target updates:

- ▶ GPCR updates:
 - Thyrotropin-releasing hormone receptors
 - NOP receptor
- ▶ VGIC updates:
 - Members of the Transient Receptor Potential superfamily of channels
- ▶ NHR updates:
 - Retinoic acid-related orphans introduction
 - Liver X receptor- α and β
 - COUP-TF-like receptors
 - 3-Ketosteroid receptors
- ▶ Enzyme updates:
 - Enzymes involved in hydrogen sulphide synthesis

BLAST search

Latest News

From our blog

GtoPdb database release 2015.3
by guidetopharmacology - Oct 19, 2015
Our total number of curated interactions now stands at 13859. New BLAST tool for sequence-based searching of targets available ...

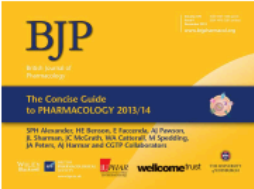
New project to develop "The Guide to Immunopharmacology"
by guidetopharmacology - Sep 25, 2015
We are very pleased to announce a new initiative (from 1st Nov 2015) to establish "The Guide to Immunopharmacology": ...

Latest news from NC-IUPHAR

Hot topic: GPR3 may be a target for AD drug development
Oct 20, 2015
New paper describes how loss of GPR3 reduces the amyloid plaque burden and improves memory in Alzheimer's disease mouse models. ...

Database update: version 2015.3 released
Oct 20, 2015
The latest release includes new ligands, updates to GPCRs


The Concise Guide to PHARMACOLOGY 2013/14



A publication snapshot created from the database summary pages.

[Access the table of contents](#) [GO](#)

IUPHAR Database



The IUPHAR/BPS Guide to PHARMACOLOGY builds upon and replaces the original IUPHAR Committee on Receptor Nomenclature and Drug Classification Database (IUPHAR-DB)

Citations in IUPHAR

- Citation to the IUPHAR database as a whole (the root) is a traditional paper written by the main curators (**owners**) of the database.
- Each IUPHAR Family and Family Introduction page has an independent citation.
 - Information about a Family is managed by a set of **curators**, which may be different for each family.
 - The detailed Family Introduction page is written by a set of **contributors**, which may be different from the curators of the Family.

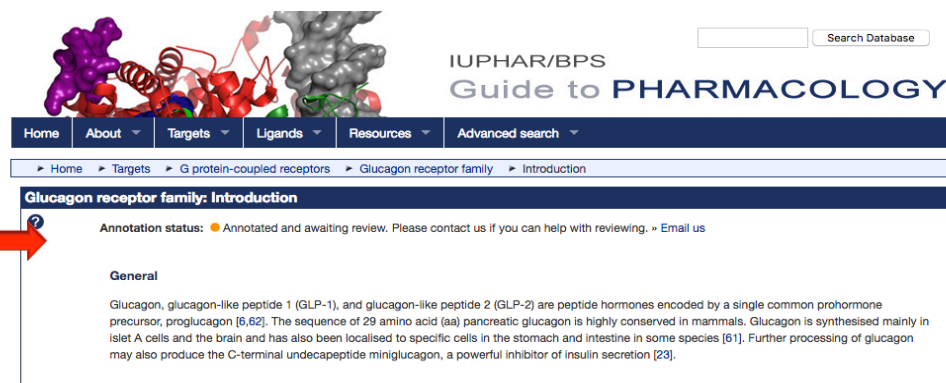
Citations in IUPHAR, cont.

Family page



The screenshot shows the 'Family page' for the Glucagon receptor family. At the top, there is a 3D molecular model of the receptor. Below it, the IUPHAR/BPS Guide to PHARMACOLOGY logo is visible. A navigation menu includes Home, About, Targets, Ligands, Resources, and Advanced search. The breadcrumb trail is: Home > Targets > G protein-coupled receptors > Glucagon receptor family. The main heading is 'Glucagon receptor family'. Below this, a note states: 'Unless otherwise stated all data on this page refer to the human proteins. Gene information is provided for human (Hs), mouse (Mm) and rat (Rn)'. There are buttons for 'Expand all sections' and 'Collapse all sections'. An 'Overview' section is partially visible, with a red box highlighting a link that says 'More detailed introduction [6]'. An arrow points from this link to the right.

Family Introduction page



The screenshot shows the 'Family Introduction page' for the Glucagon receptor family. It features the same 3D molecular model and IUPHAR/BPS Guide to PHARMACOLOGY logo as the family page. The navigation menu is identical. The breadcrumb trail is: Home > Targets > G protein-coupled receptors > Glucagon receptor family > Introduction. The main heading is 'Glucagon receptor family: Introduction'. Below this, the 'Annotation status' is shown as 'Annotated and awaiting review. Please contact us if you can help with reviewing. » Email us'. The 'General' section contains the following text: 'Glucagon, glucagon-like peptide 1 (GLP-1), and glucagon-like peptide 2 (GLP-2) are peptide hormones encoded by a single common prohormone precursor, proglucagon [6,62]. The sequence of 29 amino acid (aa) pancreatic glucagon is highly conserved in mammals. Glucagon is synthesised mainly in islet A cells and the brain and has also been localised to specific cells in the stomach and intestine in some species [61]. Further processing of glucagon may also produce the C-terminal undecapeptide miniglucagon, a powerful inhibitor of insulin secretion [23].'

Database page citation:

Miller, Drucker, Bataille, Chan, Delagrangé, Göke, Mayo, Thorens, Hills.
Glucagon receptor family.
Accessed on 08/05/2017.
IUPHAR/BPS Guide to PHARMACOLOGY,
<http://www.guidetopharmacology.org/GRAC/FamilyDisplayForward?FamilyId=1>.

To cite this family introduction, please use:

Miller, Drucker, Bataille, Chan, Delagrangé, Göke, Mayo, Thorens, Hills.
Glucagon receptor family, introduction.
Accessed on 08/05/2017.
IUPHAR/BPS Guide to PHARMACOLOGY,
<http://www.guidetopharmacology.org/GRAC/FamilyIntroductionForward?FamilyId=1>.

Why not just hard code citations?

- Citations vary with what part of the database is being cited.
 - There are a very large number of “parts” of a database.
- In the future, IUPHAR would like to enable general queries
 - Queries may combine “parts” in different ways.
- We cannot expect to create a citation for each possible query result.
- **Citations should be lifted up to schema-level specifications so they can be reasoned about.**

**Given a database D and a query Q ,
generate an appropriate citation.**

Citations for general queries

The citation generation problem

- It is common for database owners to supply citations for some parts (**views**) of the database, $V_1 \dots V_n$.
- So the problem becomes: Given a query Q , can it be rewritten using the views? That is, is there a Q_i such that

$$\forall D \in \mathcal{S}. Q(D) = Q_i(V_{i1}(D), \dots, V_{ik}(D))$$

- If so, the citations for V_{i1}, \dots, V_{ik} could be used to create a citation for Q .

Answering queries using views

- The problem of answering queries using views has been well studied and is generally hard – but in our context may be tractable.
- A. Halevy. Answering queries using views: A survey. VLDB J., 10(4):270–294, 2001.
- Lenzerini. Data Integration: A Theoretical Perspective: PODS, 2002.
- A. Deutsch, L. Popa, and V. Tannen. Query reformulation with constraints. SIGMOD Record, 35(1): 65–73, 2006.
- F. Afrati, C. Li and J. Ullman. Using views to generate efficient evaluation plans for queries. JCSS 73(5): 703 - 724, 2007.

Effect of parameters

- Parameterized views define a family of views, one for each value of the parameter.

FID	FName	Type
1	Glucagen receptor...	GPCR
2	CLR (calcitonin receptor-like receptor) ...	GPCR
3	Peptidases and proteinases...	Kinase
4	A multifunctional molecule, adenosine...	Kinase
5	Chromatin modifying enzymes...	Kinase

$\lambda F. V1(F, N, Ty) :- \text{Family}(F, N, Ty)$

$V4(F, N, Ty) :- \text{Family}(F, N, Ty)$

“Instantiated views”: $V1(F, N, Ty)(1), V1(F, N, Ty)(2), \dots, V1(F, N, Ty)(5)$

Citation views

- To specify a citation, there are three components:

- **View definition:** specifies

- **Citation query:** specifies information to include

“Universal” across different types of databases (e.g. relational, XML, RDF...)

Simplifies reasoning over queries and views

specifies how to construct the snippets of information

- We call this triple a **citation view**.

- For now, we will focus on the **view definition**, which is expressed in Datalog.

IUPHAR: Citation views

Schema:

Family(FID, FName, Type)
FamilyIntro(FID, Text)
Person(PID, PName, Affiliation)
FC(FID, PID) FIC (FID, PID)

View definitions:

$\lambda F. V1(F, N, Ty) :- \text{Family}(F, N, Ty)$
 $\lambda F. V2(F, Tx) :- \text{FamilyIntro}(F, Tx)$

Citation queries:

$\lambda F. C_{V1}(F, PN) :- \text{Family}(F, N, Ty), \text{FC}(F, P), \text{Person}(P, PN)$
 $\lambda F. C_{V2}(F, PN) :- \text{FamilyIntro}(F, Tx), \text{FIC}(F, P), \text{Person}(P, PN)$

Generating citations

- If the query matches a view definition, we can use the associated citation query and function.
- But what if it doesn't?
 - Nothing matches the query
 - A set of view definitions are used to rewrite the query
 - More than one set of view definitions can be used to rewrite the query

What is a “good” citation?

- Contains appropriate snippets of information
 - E.g. as suggested by DataCite Schema
- Allows the data as it appeared at time of citation to be retrieved
 - Query and timestamp
 - Proll and Rauber: Scalable data citation in dynamic, large databases: Model and reference implementation (IEEE Big Data 2013).
- Concise
- Specific
- ◆ **Our approach enables the DBA to specify the tradeoff between conciseness and specificity.**

IUPHAR: Generating the citation (1)

Schema:

Family(FID, FName, Type)
FamilyIntro(FID, Text)

View definitions:

$\lambda F. V1(F, N, Ty) :- \text{Family}(F, N, Ty)$
 $\lambda F. V2(F, Tx) :- \text{FamilyIntro}(F, Tx)$

- A query is another Datalog expression (unparameterized).

$Q_1(F, N, Ty) :- \text{Family}(F, N, Ty), F = 1$

- This can be rewritten using V1

$Q_1'(F, N, Ty) :- V1(F, N, Ty)(1)$

- We can then construct a citation to Q in terms of the citation for V1(F, N, Ty) ("1").

IUPHAR: Generating the citation (2)

Schema:

Family(FID, FName, Type)
FamilyIntro(FID, Text)

View definitions:

$\lambda F. V1(F, N, Ty) :- \text{Family}(F, N, Ty)$
 $\lambda F. V2(F, Tx) :- \text{FamilyIntro}(F, Tx)$

- Consider another input query

$Q_2(F, N, Y) :- \text{Family}(F, N, Ty)$

- This can be rewritten using V1

$Q_2'(F, N, Y) :- V1(F, N, Ty)$

- Now we must use *all instantiations* of V1 to construct a citation to Q

- $V1(F, N, Ty)(1), V1(F, N, Ty)(2), \dots, V1(F, N, Ty)(5)$

IUPHAR: Generating the citation (3)

Schema:

Family(FID, FName, Type)
FamilyIntro(FID, Text)

View definitions:

$\lambda F. V1(F, N, Ty) :- \text{Family}(F, N, Ty)$
...
 $V4(F, N, Ty) :- \text{Family}(F, N, Ty)$

- Consider the following query, with another view V4

$Q_2(F, N, Ty) :- \text{Family}(F, N, Ty), Ty = \text{"GPCR"}$

- This can be rewritten using V1 or V4 (**alternate** use)

$Q_2'(F, N, Ty) :- V1(F, N, Ty), Ty = \text{"GPCR"}$
 $Q_2''(F, N, Ty) :- V4(F, N, Ty), Ty = \text{"GPCR"}$

- We can then construct a citation to Q in terms of the citations $V1(F, N, Ty)(1)$, $V1(F, N, Ty)(2)$ **or** $V4(F, N, Ty)$

IUPHAR: Generating the citation (4)

Schema:

Family(FID, FName, Type)
FamilyIntro(FID, Text)

View definitions:

$\lambda F. V1(F, N, Ty) :- \text{Family}(F, N, Ty)$
 $\lambda F. V2(F, Tx) :- \text{FamilyIntro}(F, Tx)$

- Another query:

$Q_1(F, N, Ty, Tx) :- \text{Family}(F, N, Ty), \text{FamilyIntro}(F, Tx), F = 1$

- This can be rewritten using V1 and V2 (**joint** use)

$Q_1'(F, N, Ty, Tx) :- V1(F, N, Ty)(1), V2(F, Tx)(1)$

- We can then construct a citation to Q in terms of the citations for V1(F, N, Ty)(1) and V2(F, Tx)(1).

Citation views as annotation

- Citation views are a type of annotation on tuples.
- Provenance is a form of annotation on tuples, which is well understood while being carried through queries.
 - Green, Karvounarakis, Tannen: Provenance Semirings, PODS 2007:
 - *Joint use*: joins of tuples
 - *Alternate use*: unions and projections of tuples
- **Can we use these ideas to understand how citation “annotations” on tuples are combined in general queries?**



2017 PODS Test of Time

Citation “semiring”?

- Given a (conjunctive) query, we rewrite it to a set of minimal equivalent queries that contain at least one citation view.
 - Let the set of queries obtained in this way be $\{Q_1, \dots, Q_n\}$
- Each Q_i contains a set of citation views $\{V_{i1}, \dots, V_{imi}\}$. The **joint** use ($*$) of their citations constructs a citation for Q_i , $C(Q_i)$.
 - $C(Q_i) = C(V_{i1}) * \dots * C(V_{imi})$
- The **alternate** use ($+$) of each $C(Q_i)$ constructs a citation for Q , $C(Q)$.
 - $C(Q) = C(Q_1) + \dots + C(Q_n)$

*“Model for Fine-Grained Data Citation”, CIDR 2017
S. Davidson, D. Deutch, T. Milo, and G. Silvello.*

Interpreting * and +

- **Joint** use of citations: $C(V_{i1}) * \dots * C(V_{imi})$
 - * could be union or some sort of join
 - E.g. in example 4, V1 and V2 were jointly used:
 $V1(F, N, Ty) ("F123") * V2(F, Tx) ("F123")$
- **Alternate** use of citations: $C(Q_1) + \dots + C(Q_n)$
 - + could be union or min (wrt some ordering on views)
 - E.g. in example 3, both the parameterized and unparameterized views on Family matched
 $(V1(F, N, Ty)(1), V1(F, N, Ty)(2)) + V4$
- ◆ **Joint and alternate use are “policies” specified by the DBA**

Example of output citation

View definition:

$\lambda F. V1(F, N, Ty) :- \text{Family}(F, N, Ty)$

Citation query:

$\lambda F. C_{V1}(F, PN) :- \text{Family}(F, N, Ty), \text{FC}(F, P), \text{Person}(P, PN)$

$Q_1(F, N, Ty) :- \text{Family}(F, N, Ty), F = 1$

$Q_1'(F, N, Ty) :- V1(F, N, Ty)(1)$

FID	FName	Type
1	Glucagen ...	GPCR

Citation:

Miller, Drucker, Bataille, Chan, Delagrangé,
Göke, Mayo, Thorens, Hills.
Glucagon receptor family.
Accessed on 08/05/2017.
IUPHAR/BPS Guide to PHARMACOLOGY,
Family(F, N, Ty), F = 1

Example, with * as “join”

View definitions:

$\lambda F. V1(F, N, Ty) :- \text{Family}(F, N, Ty)$

$\lambda F. V2(F, Tx) :- \text{FamilyIntro}(F, Tx)$

Citation queries:

$\lambda F. C_{V1}(F, PN) :- \text{Family}(F, N, Ty), FC(F, P), \text{Person}(P, PN)$

$\lambda F. C_{V2}(F, PN) :- \text{FamilyIntro}(F, Tx), FIC(F, P), \text{Person}(P, PN)$

$Q_1(F, N, Ty, Tx) :- \text{Family}(F, N, Ty), \text{FamilyIntro}(F, Tx), F=1$

$Q_1'(F, N, Ty, Tx) :- V1(F, N, Ty)(1), V2(F, Tx)(1)$

FID	FName	Type	Text
1	Glucagen ...	GPCR	Glucagon regulates ...

Citation:

Miller, Drucker, Bataille, Chan, Delagrangé,
Göke, Mayo, Thorens, Hills.
Glucagon receptor family, introduction.

Miller, Drucker, Bataille, Chan, Delagrangé,
Göke, Mayo, Thorens, Hills.
Glucagon receptor family.

Accessed on 08/05/2017.
IUPHAR/BPS Guide to PHARMACOLOGY,
Family(F, N, Ty), FamilyIntro(F, Tx), F=1

Reaction of the DBA...



“Partitioning” views

- In current practice, citation views are simple
 - Project-select views of a single relation
- It is easily shown that if the views “partition” a relation then there is a single maximal rewriting using the views.
- And the implementation is much simpler...





Building a citation system

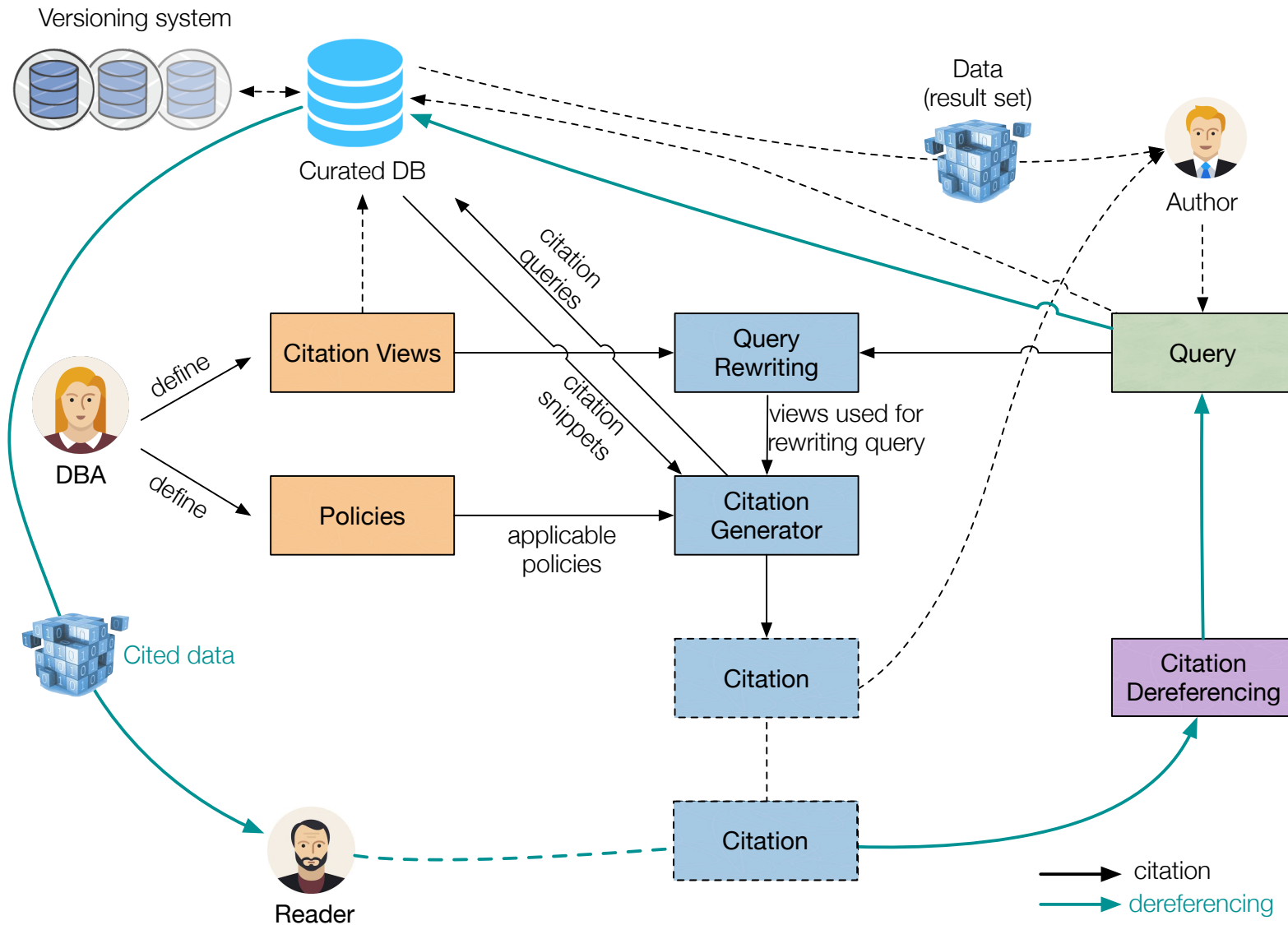


The big picture

- **Database owners** need to be able to specify citation views for the database – schema level information.
- **Database users** (“authors”) need to have citations “served up” as they extract data through queries.
- **Dereferencing** the citation should bring back the data to “readers” as of the time it was cited.

Alawini, Davidson, Hu, Wu: Automating Data Citation in CiteDB. VLDB 2017 (Demo paper).

Citation architecture



Computational challenges

- Schema-level versus instance level?
 - Should we store the citations as annotations on tuples, or should we reason at the schema level and then calculate the citation?
- Given an expected query workload, what are the “best” citation views?
 - And are the necessary snippets of citation information in the schema?
- The number of alternative uses of citation views can be large.
 - Are there efficient algorithms to find the “best” according to some metric of quality (e.g. involving the number of views, the specificity of views, or related to a view hierarchy)?

Take home message

- If we want people to cite the data they use, we need to make it easy for them to do so.
- We must also make it easy for people who publish data. Data should be easy to cite.
- For many applications, the creation of “parameterized queries” in which citations can be attached.
- Joint and alternate use semantics are “policies” to be specified by the DBA

And there are many other interesting computational challenges with data citation!

Outline

- The power of abstraction
- New problem: data citation
- **Bigger picture: Data Science**



The Tsunami of Data Science..

Role of computing research in DS



- Report by CRA's Committee on Data Science
 - **Lise Getoor**(Chair), David Culler, Eric de Sturler, David Ebert, Mike Franklin, and H. V. Jagadish

- Topics:
 - Models for data representation, acquisition, storage and access.
 - Large scale system and algorithms.
 - Learning with biased, incomplete and heterogeneous data.
 - User interaction: with data and models.
 - Ethical use: privacy, fairness, transparency

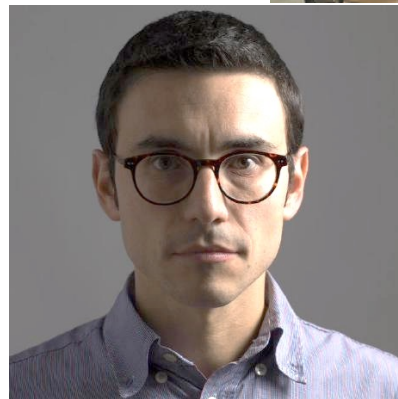
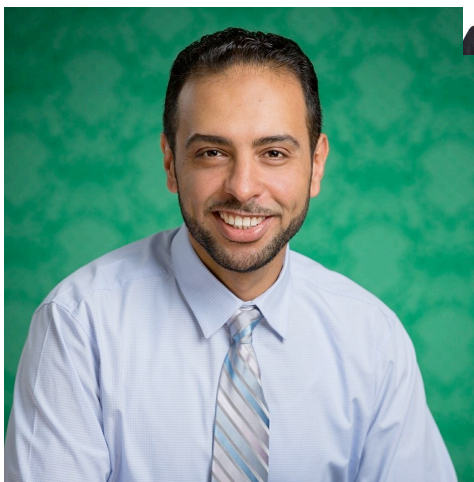
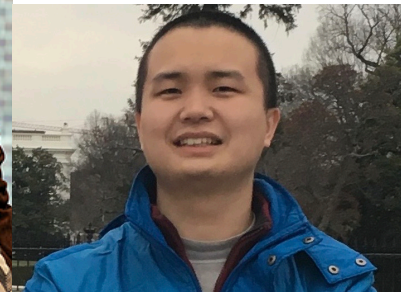
My personal perspective...

Data Science=
(CS+Stat+Math) \cap (Science | Economics | Sociology | Business | Law | ...)



- (Data Management) \cap (Machine Learning)
- “Data Engineering” akin to “Software Engineering”
 - Collecting, cleaning and organizing data sets is reported to take nearly 80% of a data scientist's time yet is the least enjoyable part of their job
- “Why Analysis” of Algorithms
- Ethical data management

Thanks to my collaborators



And to our funders...



National Science Foundation
WHERE DISCOVERIES BEGIN

NSF IIS 1302212,
NSF ACI 1547360,

...



National Institutes of Health
Turning Discovery Into Health

NIH 3-U01-EB-020954-02S1



Research at Google



U.S. DEPARTMENT OF
ENERGY



DEFENSE ADVANCED
RESEARCH PROJECTS AGENCY



WWW.ARL.ARMY.MIL
UNITED STATES ARMY RESEARCH LABORATORY



Questions?