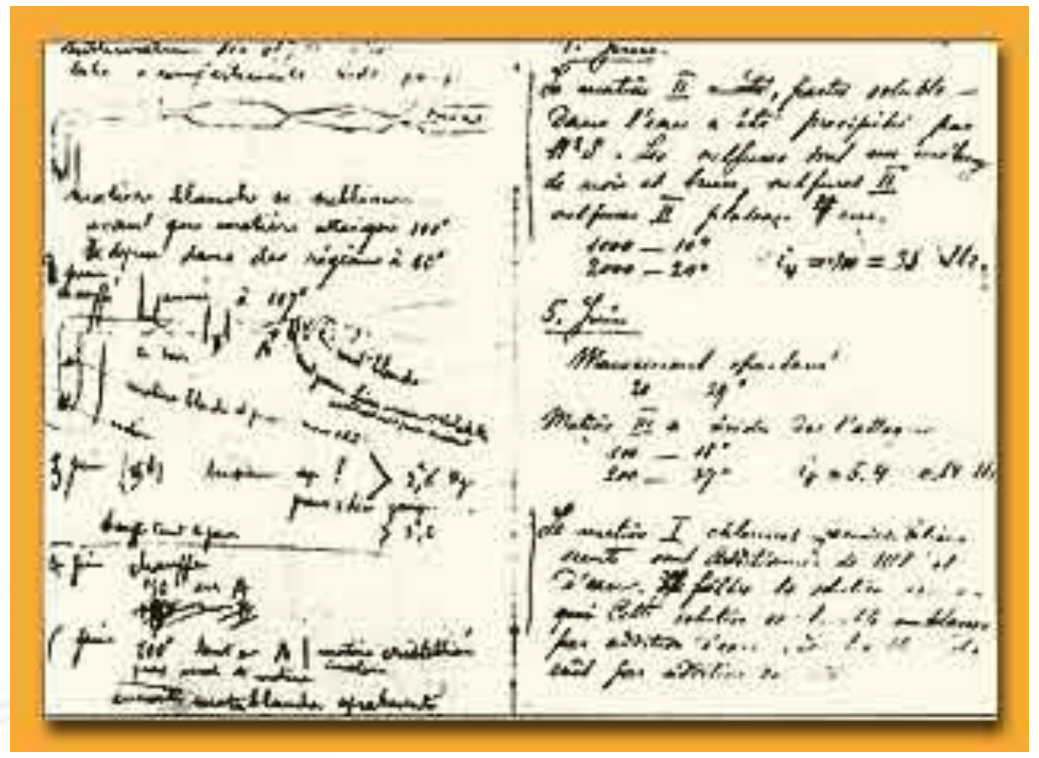
Advanced Data Management (CSCI 490/680)

Data Citation

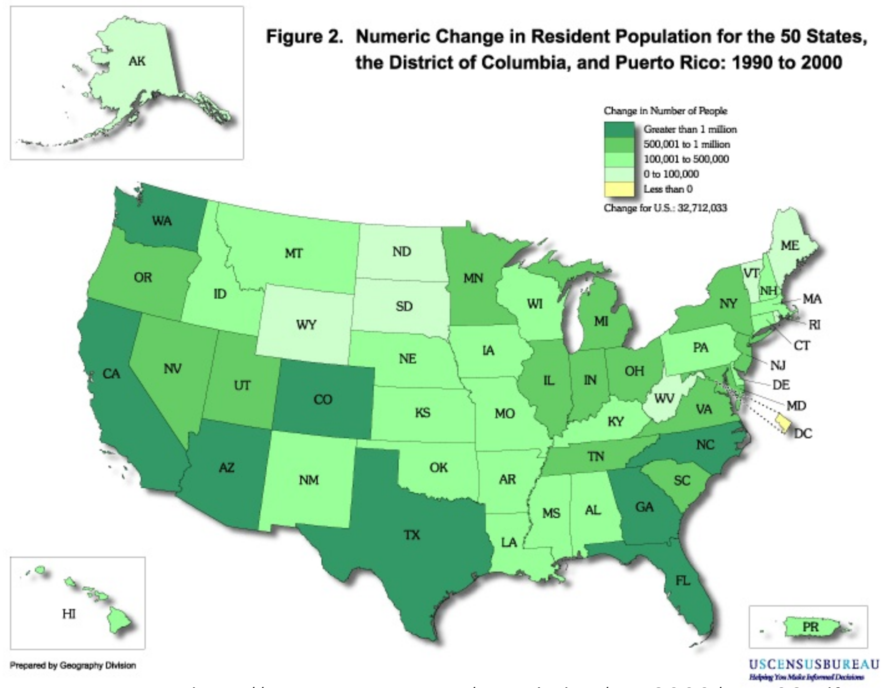
Dr. David Koop

What is Data?

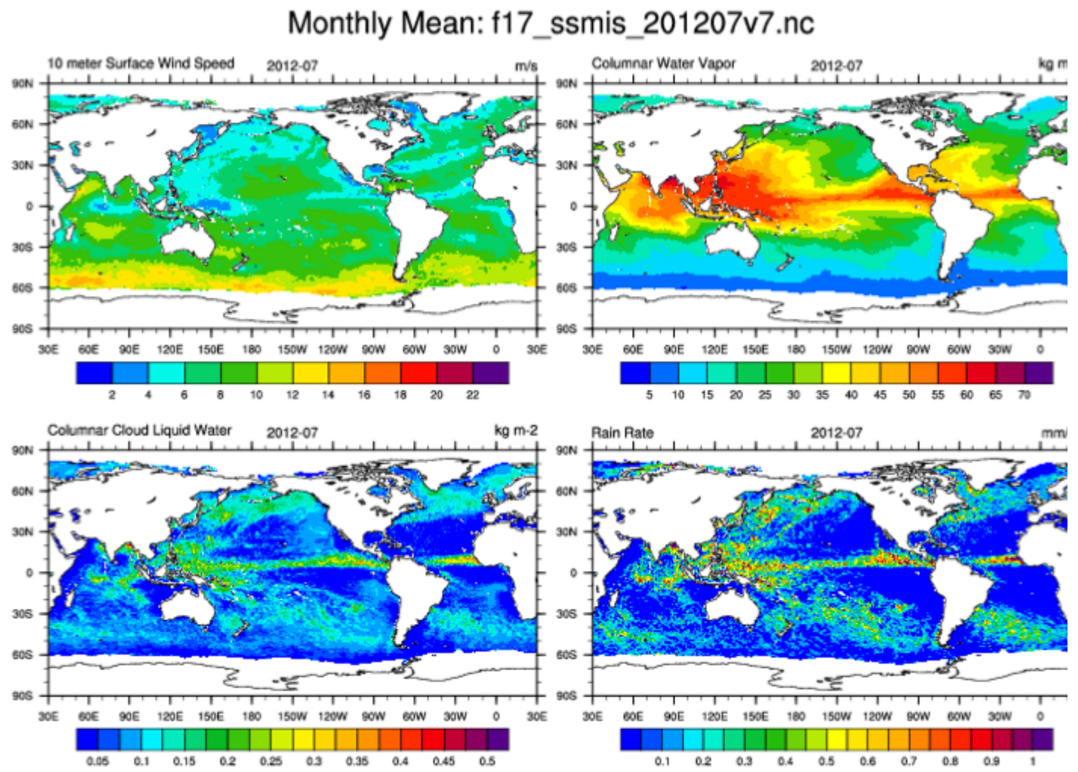
hudsonalpha.org



Marie Curie's notebook aip.org



http://www.census.gov/population/cen2000/map02.gif



ncl.ucar.edu

Date:1/2.07.75 Place:Sakaltutan
Zafor
He will grow old in his present house; new house is for sons - 5 sons. Not sure they want to live in village. He will only build another if they want him to. eS came from Germany and did the plastering. He arranged the carpentry in Kayseri. Çok para gitti. (much money went) Has a tractor.

Date:July1980 Place:Sakaltutan
Zafor:
Household now Zafor and wife; Nazif Unal and wife and youngest son, still a boy. They run two dolmuß; one with a driver from Süleymanli. Goes in and out once a day. He gets 8,000 a month. Zafor then said, keskin de©il. (not sharp - i.e.? not profitable) I said he did very well on 8,000 TL with only two journeys a day. Nazif Unal has "bought" a Durak (dolmuß stop) from Belediye and works all day in Kayseri.

http://onlineqda.hud.ac.uk/Intro_QDA/Examples_of_Qualitative_Data.php

Pisa Griffin

[C. Borgman]

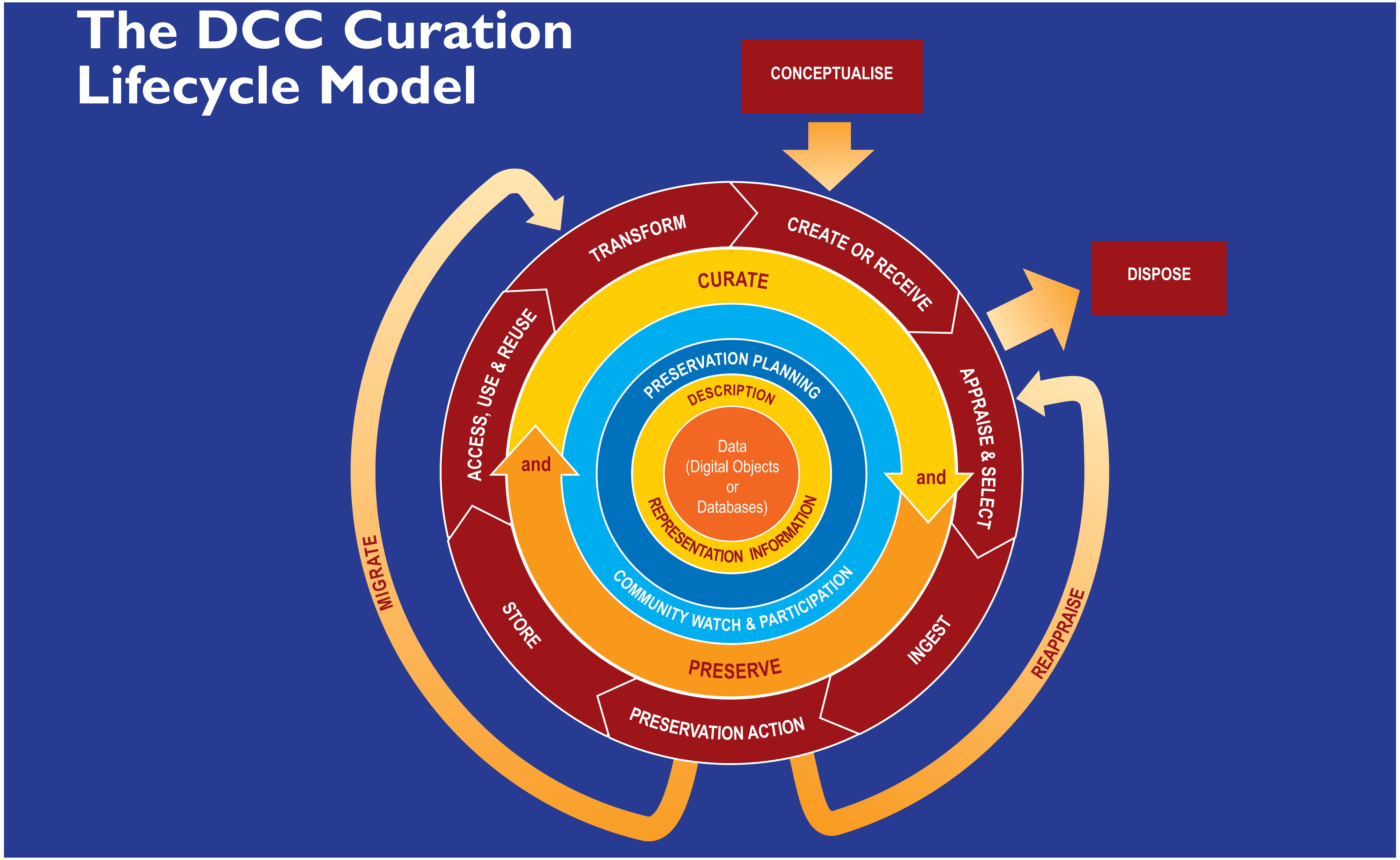
What is data?

- "Data are representations of observations, objects, or other entities used as evidence of phenomena for the purposes of research or scholarship."
  [C. L. Borgman]
- Data can be digital but can also be physical (e.g. sculptures)
- Semantics are important (e.g. temperature to engineer and biologist)
- Grey Data: surveys, student records—think about **privacy**

Sharing Data

- Required/encouraged by universities, funding agencies, publishers
- "Publications are arguments made by authors, and **data are the evidence** used to support the arguments." [C. L. Borgman]
- Questions:
  - How is data maintained? Who is responsible?
  - What is the process for curating data?
  - How long should data be kept?
  - How should data collection and curation be acknowledged?

Data Curation Lifecycle



[DCC]

Sequential Actions in Data Curation

- Conceptualize: Plan creation of data—capture method and storage options.
- Create or Receive: Create/receive data and make sure metadata exists
- Appraise and Select: Evaluate data and select for long-term curation and preservation
- Ingest: Transfer data to an archive, repository, data centre or other custodian
- Preservation Action: Data cleaning, validation (ensure that data remains authentic, reliable and usable)
- Store: Store the data in a secure manner adhering to relevant standards Access, Use and Reuse: Make sure is accessible to users and reusers
- Transform: Create new data from the original (migrate formats, subsets, etc.)

[DCC]

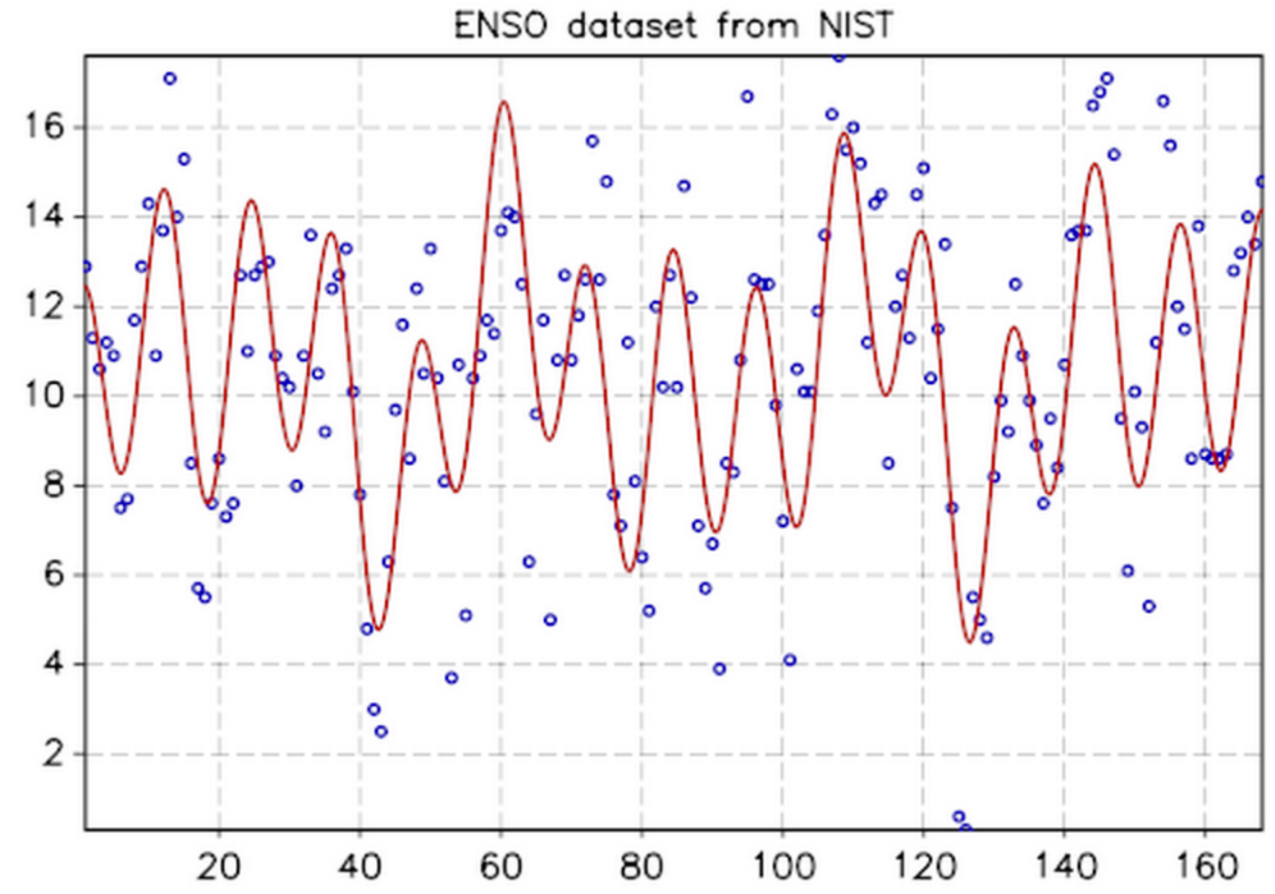
FAIR Principles

- **Findable**: Metadata and data should be easy to find for both humans and computers
- **Accessible**: Users need to know how data can be accessed, possibly including authentication and authorization
- **Interoperable**: Can be integrated with other data, and can interoperate with applications or workflows for analysis, storage, and processing
- **Reusable**: Optimize the reuse of data. Metadata and data should be well-described so they can be replicated and/or combined in different settings

[GO FAIR]

Findable: DataCite Workflow

1. Take a dataset



2. Describe it

- Title
- Authors
- Year
- Description
- And others…

3. Assign a DOI

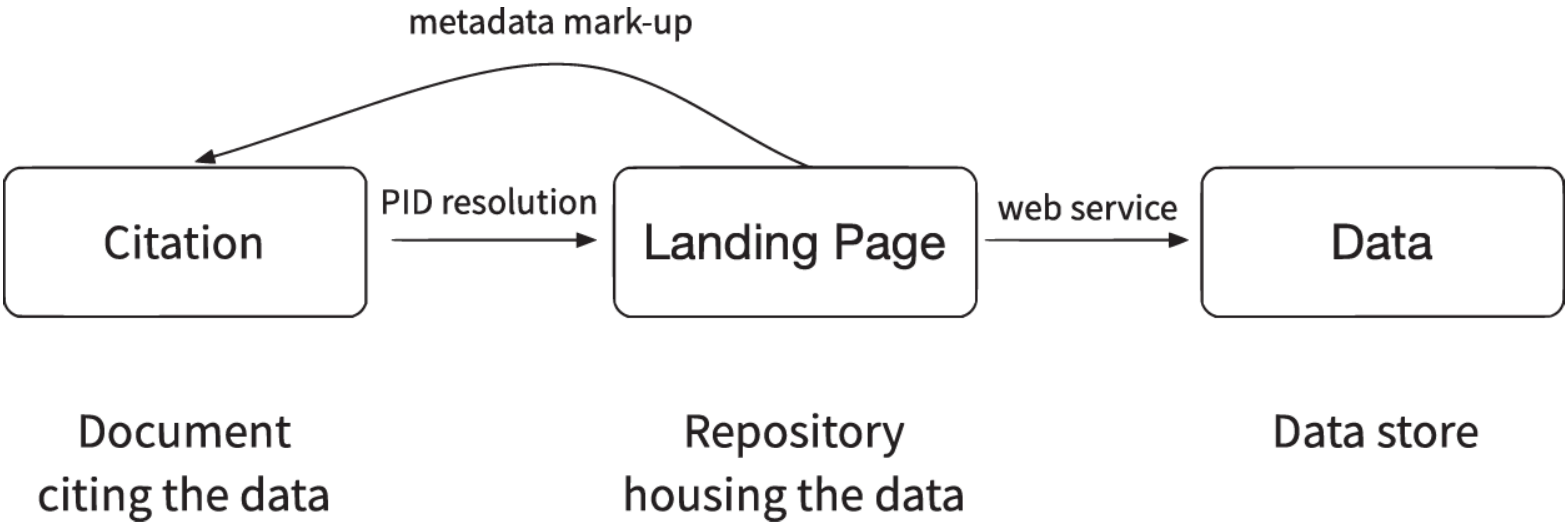
10.1234/exampledata

Proxy Prefix Suffix

https://doi.org/10.5438/n138-z3mk

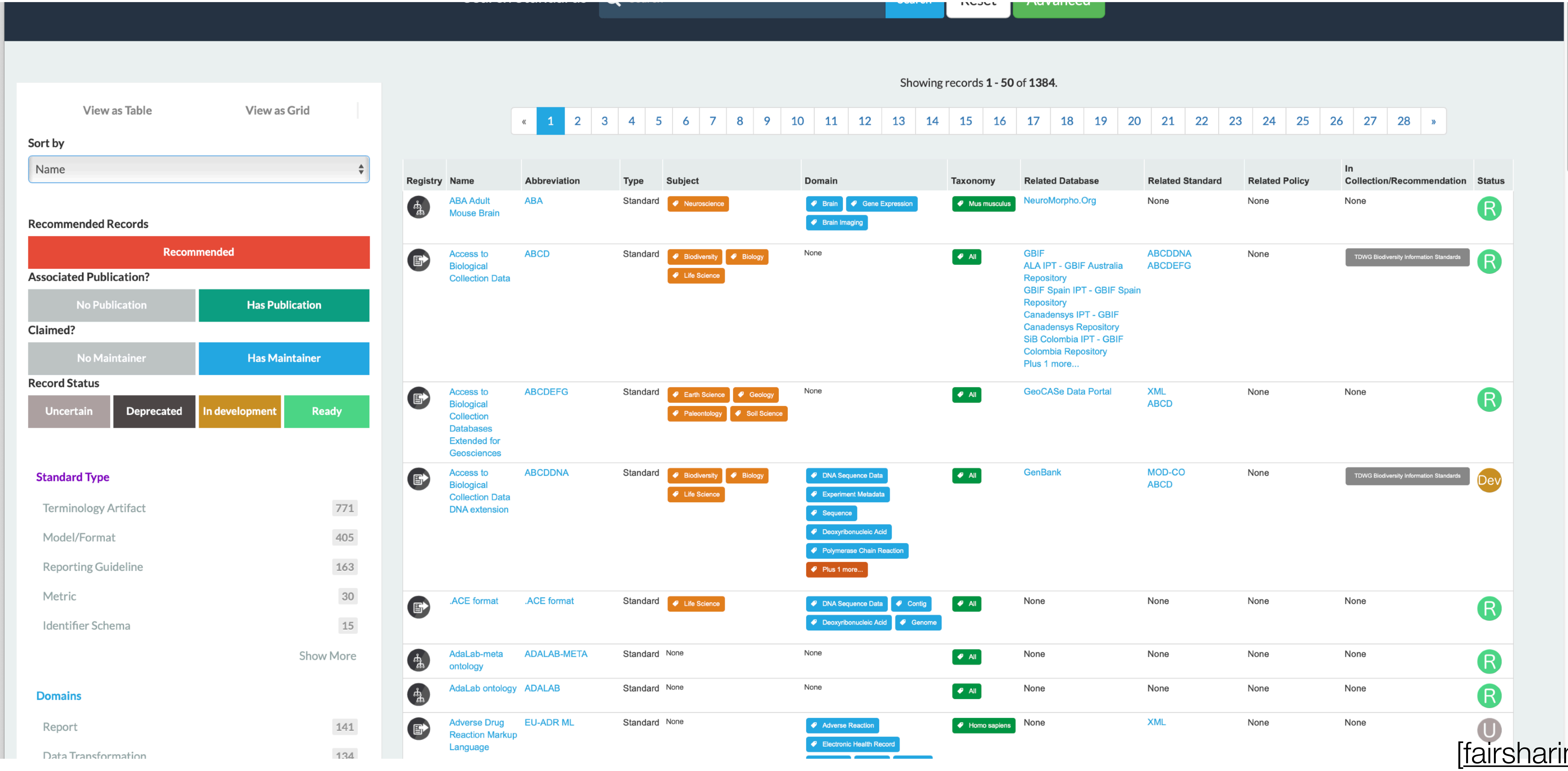
[DataCite]

Accessible: DOI to Landing Page with Metadata



[M. Fenner et al., 2019]

Interoperable: Standard vocabularies


Showing records 1 - 50 of 1384.

View as Table | View as Grid

Sort by: Name

Recommended Records: Recommended

Associated Publication? No Publication | Has Publication

Claimed? No Maintainer | Has Maintainer

Record Status: Uncertain | Deprecated | In development | Ready

Standard Type

- Terminology Artifact 771
- Model/Format 405
- Reporting Guideline 163
- Metric 30
- Identifier Schema 15

Show More

Domains

- Report 141
- Data Transformation 134

« 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 »

Registry	Name	Abbreviation	Type	Subject	Domain	Taxonomy	Related Database	Related Standard	Related Policy	In Collection/Recommendation	Status
	ABA Adult Mouse Brain	ABA	Standard	Neuroscience	Brain, Gene Expression, Brain Imaging	Mus musculus	NeuroMorpho.Org	None	None	None	R
	Access to Biological Collection Data	ABCD	Standard	Biodiversity, Biology, Life Science	None	All	GBIF, ALA IPT - GBIF Australia Repository, GBIF Spain IPT - GBIF Spain Repository, Canadensys IPT - GBIF Canadensys Repository, SiB Colombia IPT - GBIF Colombia Repository, Plus 1 more...	ABCDDNA, ABCDEFG	None	TDWG Biodiversity Information Standards	R
	Access to Biological Collection Databases Extended for Geosciences	ABCDEFG	Standard	Earth Science, Geology, Paleontology, Soil Science	None	All	GeoCASe Data Portal	XML, ABCD	None	None	R
	Access to Biological Collection Data DNA extension	ABCDDNA	Standard	Biodiversity, Biology, Life Science	DNA Sequence Data, Experiment Metadata, Sequence, Deoxyribonucleic Acid, Polymerase Chain Reaction, Plus 1 more...	All	GenBank	MOD-CO, ABCD	None	TDWG Biodiversity Information Standards	Dev
	.ACE format	.ACE format	Standard	Life Science	DNA Sequence Data, Contig, Deoxyribonucleic Acid, Genome	All	None	None	None	None	R
	AdaLab-meta ontology	ADALAB-META	Standard	None	None	All	None	None	None	None	R
	AdaLab ontology	ADALAB	Standard	None	None	All	None	None	None	None	R
	Adverse Drug Reaction Markup Language	EU-ADR ML	Standard	None	Adverse Reaction, Electronic Health Record	Homo sapiens	None	XML	None	None	U

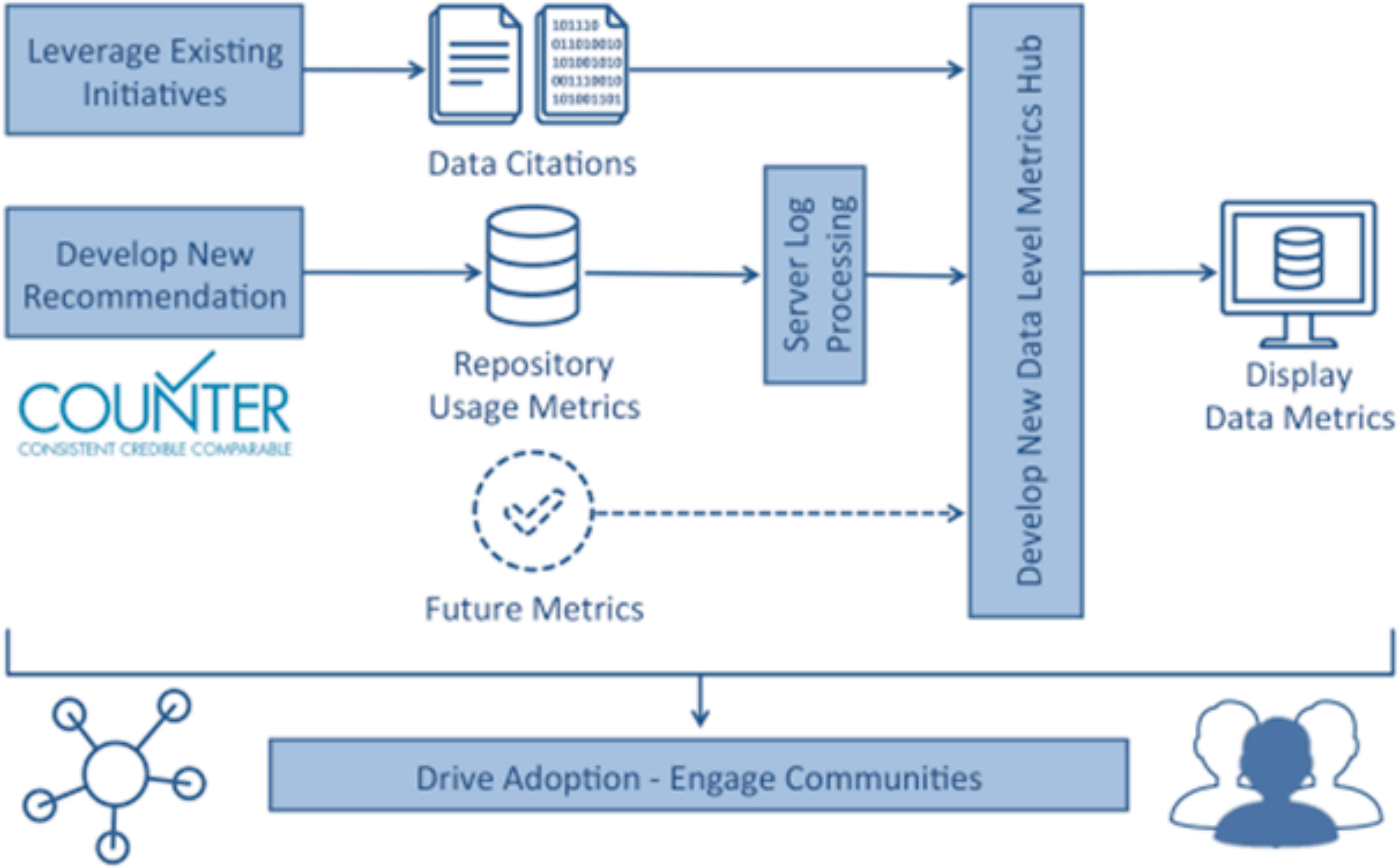
[fairsharing.org]

Reusable: Licensing

- Citation of a dataset is expected as a scholarly norm, not by law
- CC0:
  - "I hereby waive all copyright and related or neighboring rights together with all associated claims and causes of action with respect to this work to the extent possible under the law"
- CC BY: license, not a waiver as CC0
  - "You must give appropriate credit, provide a link to the license, and indicate if changes were made."
- Data Use Agreements (DUA): Used when data are restricted due to proprietary or privacy concerns.

[M. Crosas]

Reusable: Data Citation & Metrics



[H. Cousijn et al., 2019]

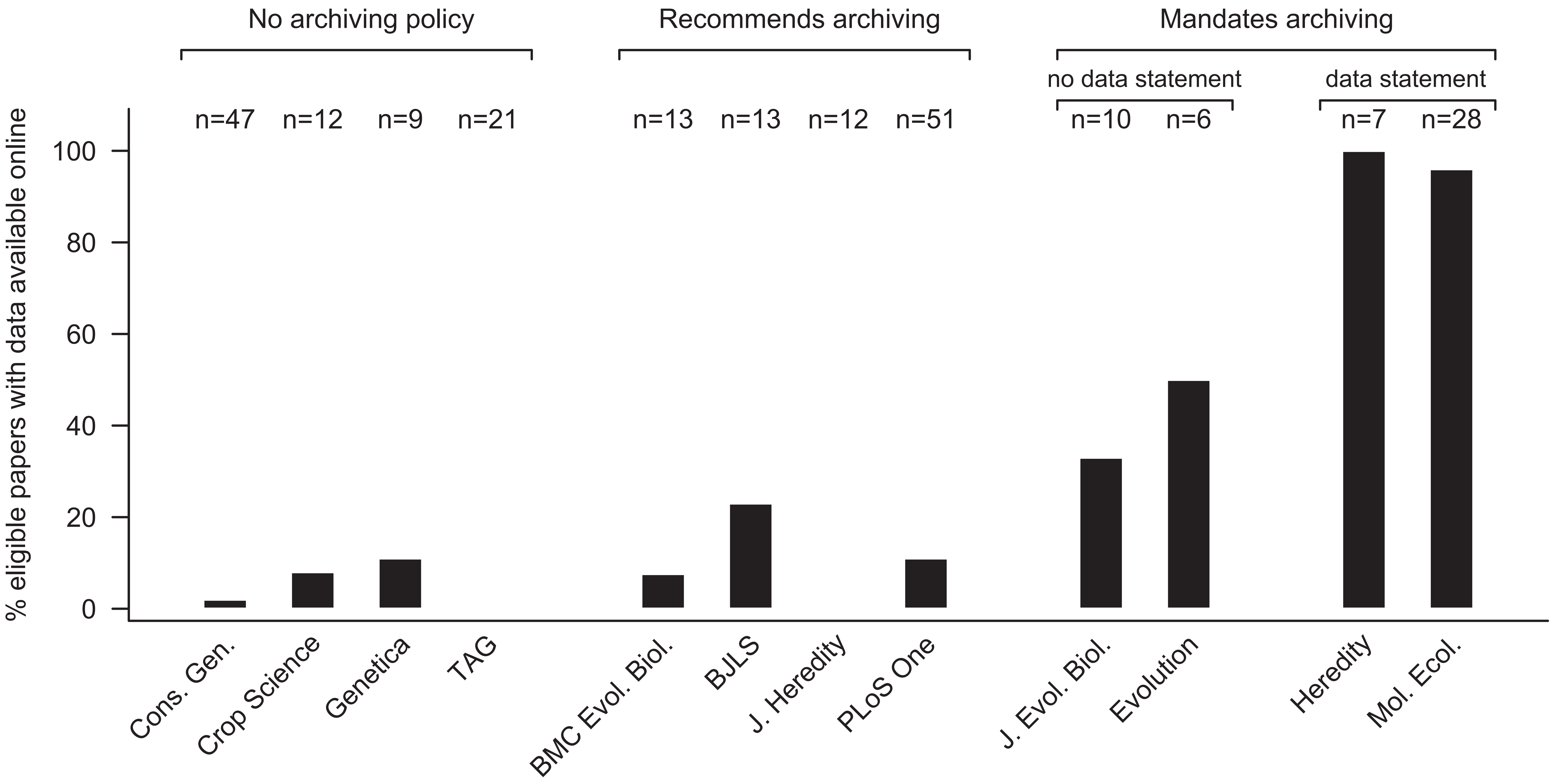
Assignment 4

- World Education Data
- Collected/collated by UNESCO, World Bank, and OECD
- Transform World Bank Data
- Impute missing year data
- Integrate teacher and student numbers
- Fuse three datasets

Studying Data Availability

- Who **mandates** data sharing, and what is the impact?
  - Government
  - Funding agencies
  - Institutions
  - **Journals**
- How does the **age** of a publication/data item affect availability?
  - If not curated, how to locate?
  - What factors influence this?

Data Availability by Journal Policy



[T. Vines et al., 2013]

Data Availability by Year

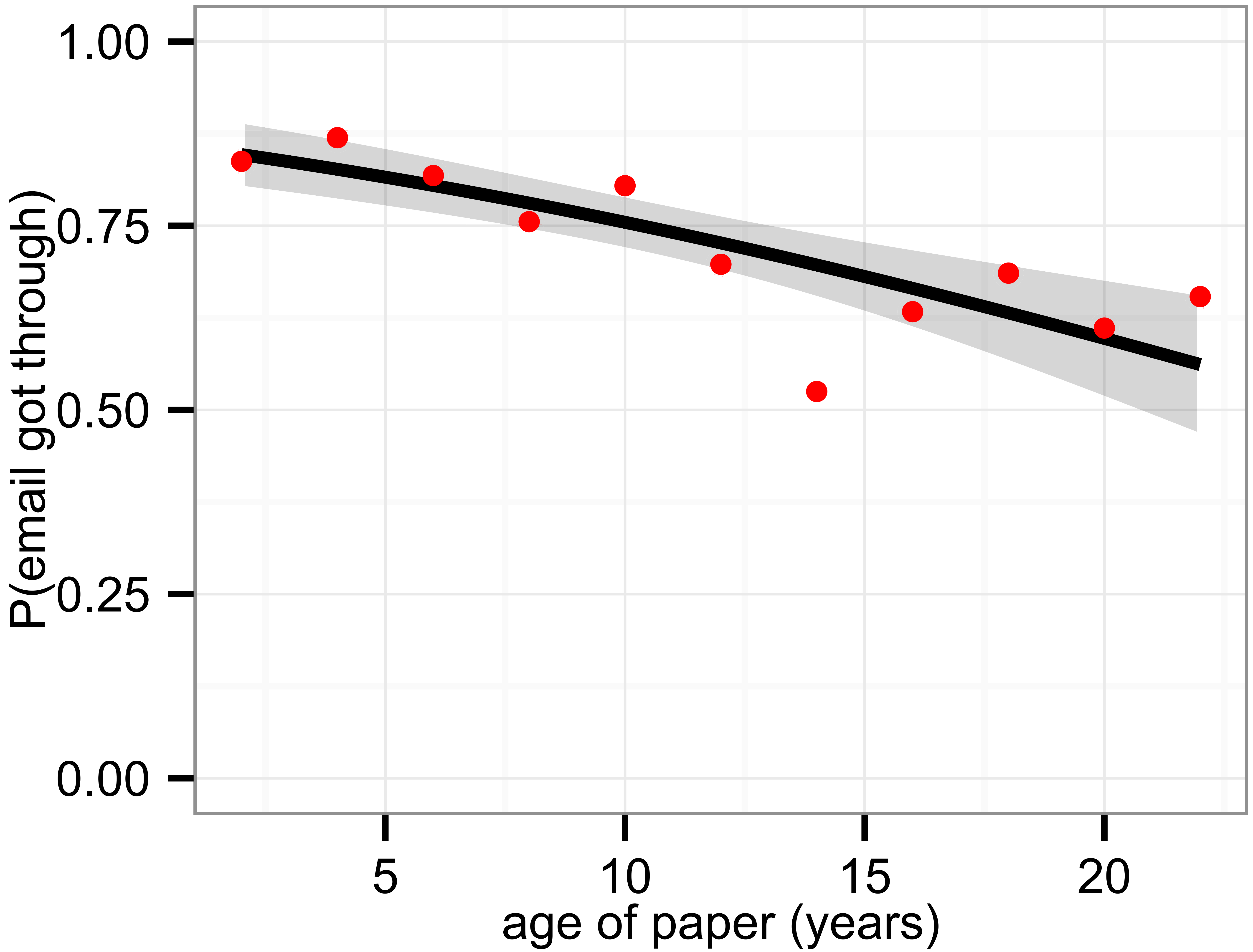
Table 1. Breakdown of Data Availability by Year of Publication

Year	No Working E-Mail	No Response to E-Mail	Response Did Not Give Status of Data	Data Lost	Data Exist, Unwilling to Share	Data Received	Data Extant (Unwilling to Share + Received)	Number of Papers
1991	9 (35%)	9 (35%)	2 (8%)	4 (15%)	1 (4%)	1 (4%)	2 (8%)	26
1993	14 (39%)	11 (31%)	3 (8%)	7 (19%)	0 (0%)	1 (3%)	1 (3%)	36
1995	11 (31%)	9 (26%)	0 (0%)	7 (20%)	2 (6%)	6 (17%)	8 (23%)	35
1997	11 (37%)	9 (30%)	1 (3%)	2 (7%)	3 (10%)	4 (13%)	7 (23%)	30
1999	19 (48%)	13 (32%)	1 (2%)	1 (2%)	0 (0%)	6 (15%)	6 (15%)	40
2001	13 (30%)	15 (35%)	3 (7%)	4 (9%)	0 (0%)	8 (19%)	8 (19%)	43
2003	9 (20%)	20 (43%)	4 (9%)	2 (4%)	0 (0%)	11 (24%)	11 (24%)	46
2005	11 (24%)	14 (31%)	6 (13%)	1 (2%)	0 (0%)	13 (29%)	13 (29%)	45
2007	12 (18%)	31 (47%)	2 (3%)	4 (6%)	1 (2%)	16 (24%)	17 (26%)	66
2009	9 (13%)	34 (49%)	3 (4%)	5 (7%)	6 (9%)	12 (17%)	18 (26%)	69
2011	13 (16%)	29 (36%)	8 (10%)	0 (0%)	7 (9%)	23 (29%)	30 (38%)	80
Totals	131 (25%)	194 (38%)	33 (6%)	37 (7%)	20 (4%)	101 (19%)	121 (23%)	516

Data are displayed as n (%); the percentages are calculated by rows.

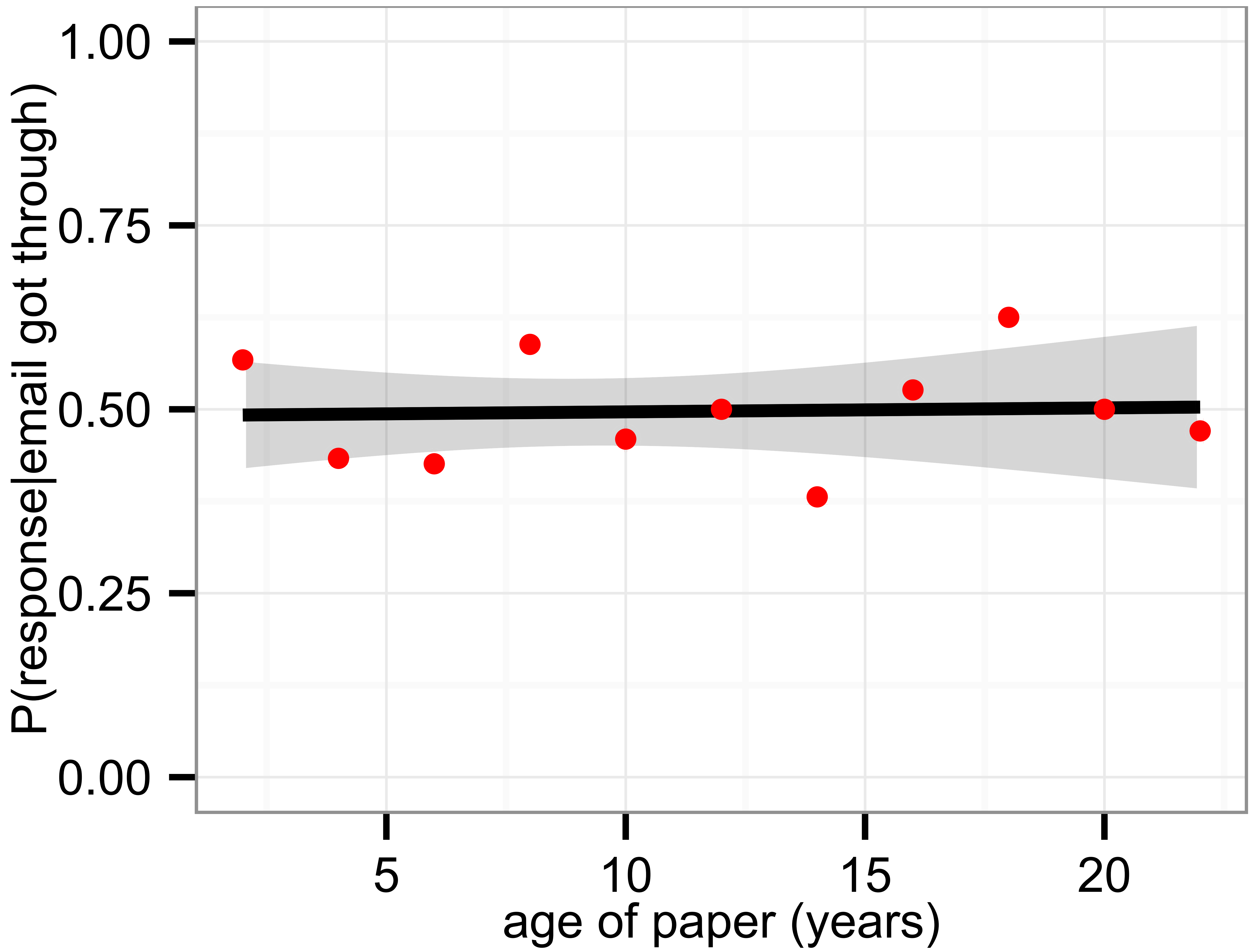
[T. Vines et al., 2014]

Working Email



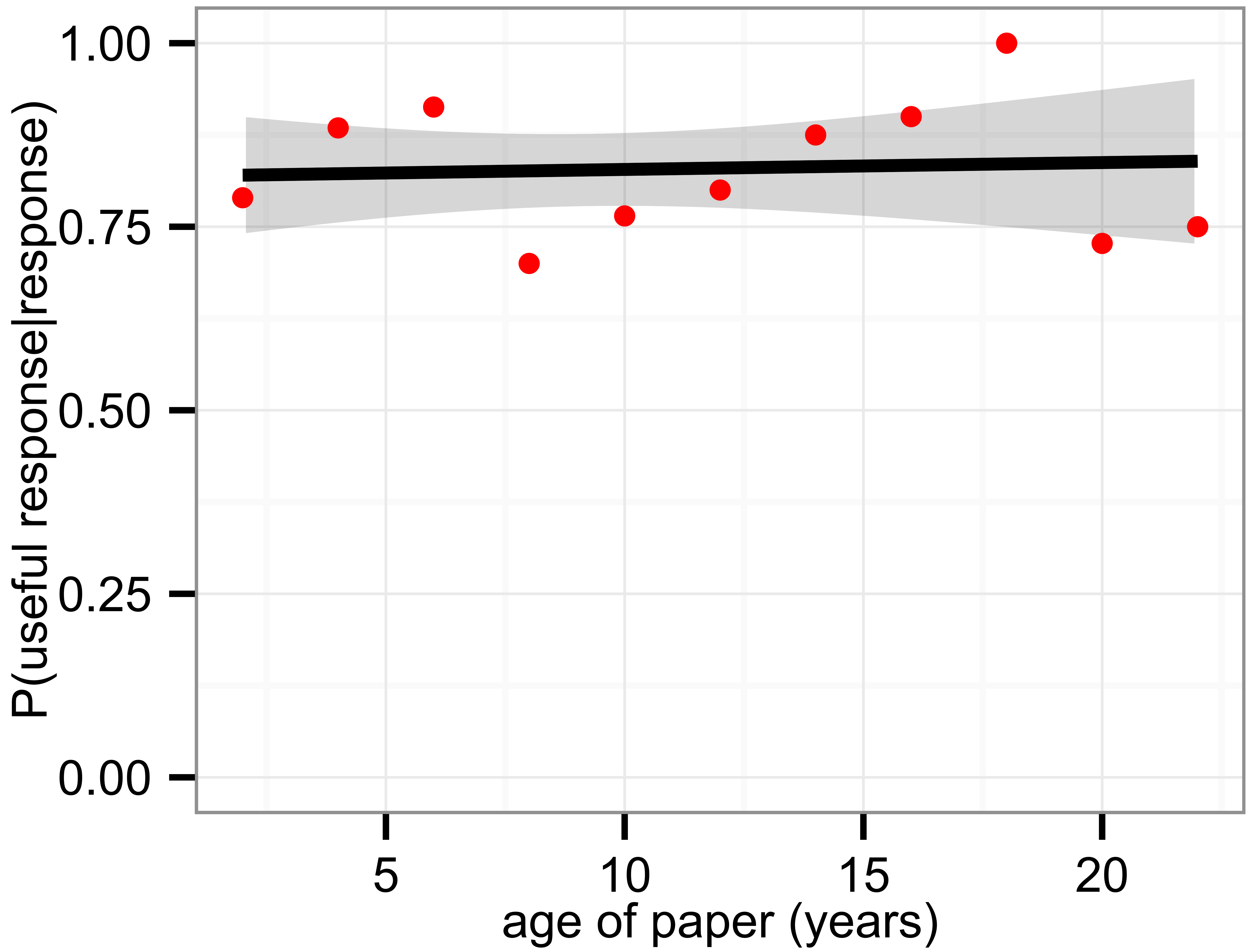
[T. Vines et al., 2014]

Received Response



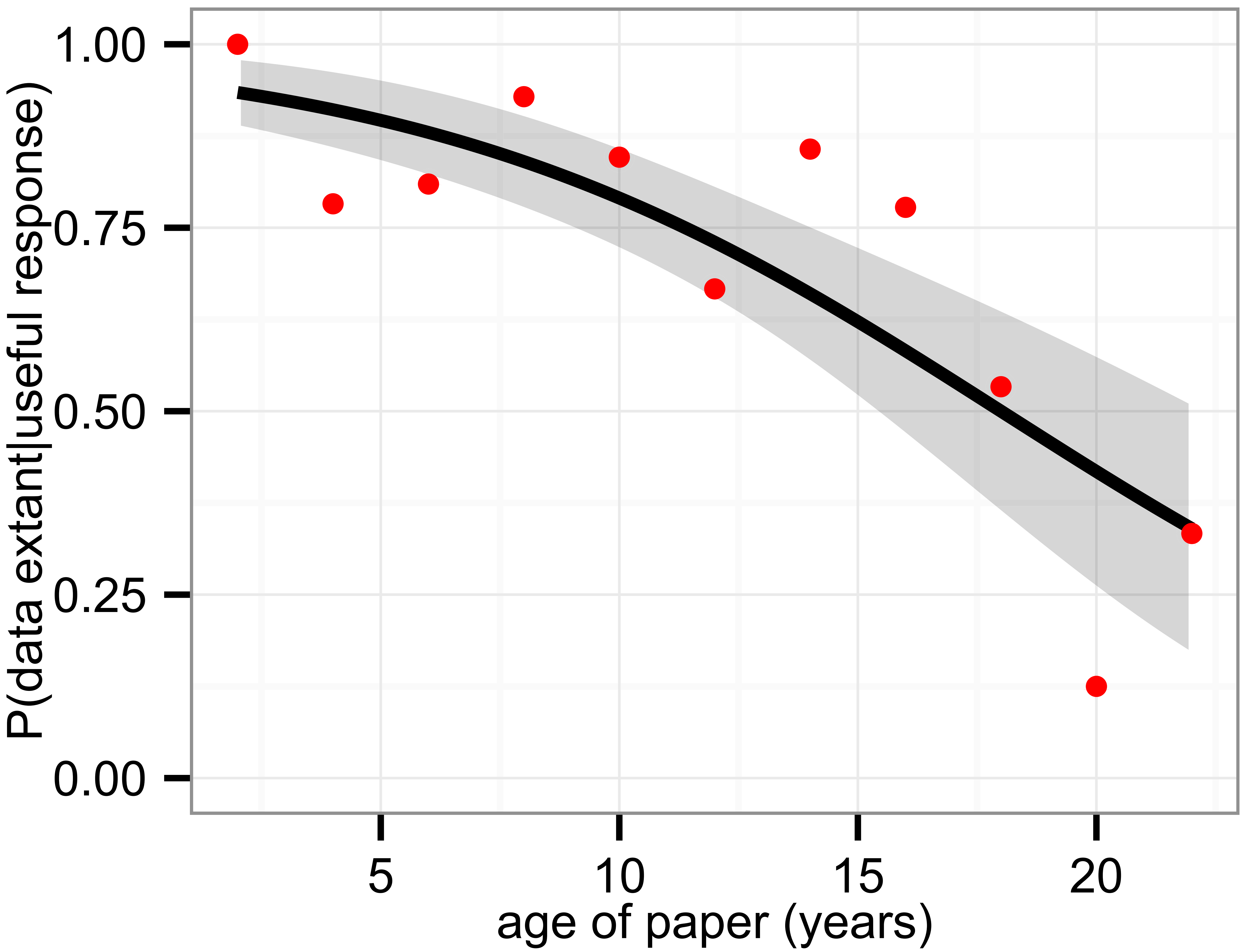
[T. Vines et al., 2014]

Status of Data



[T. Vines et al., 2014]

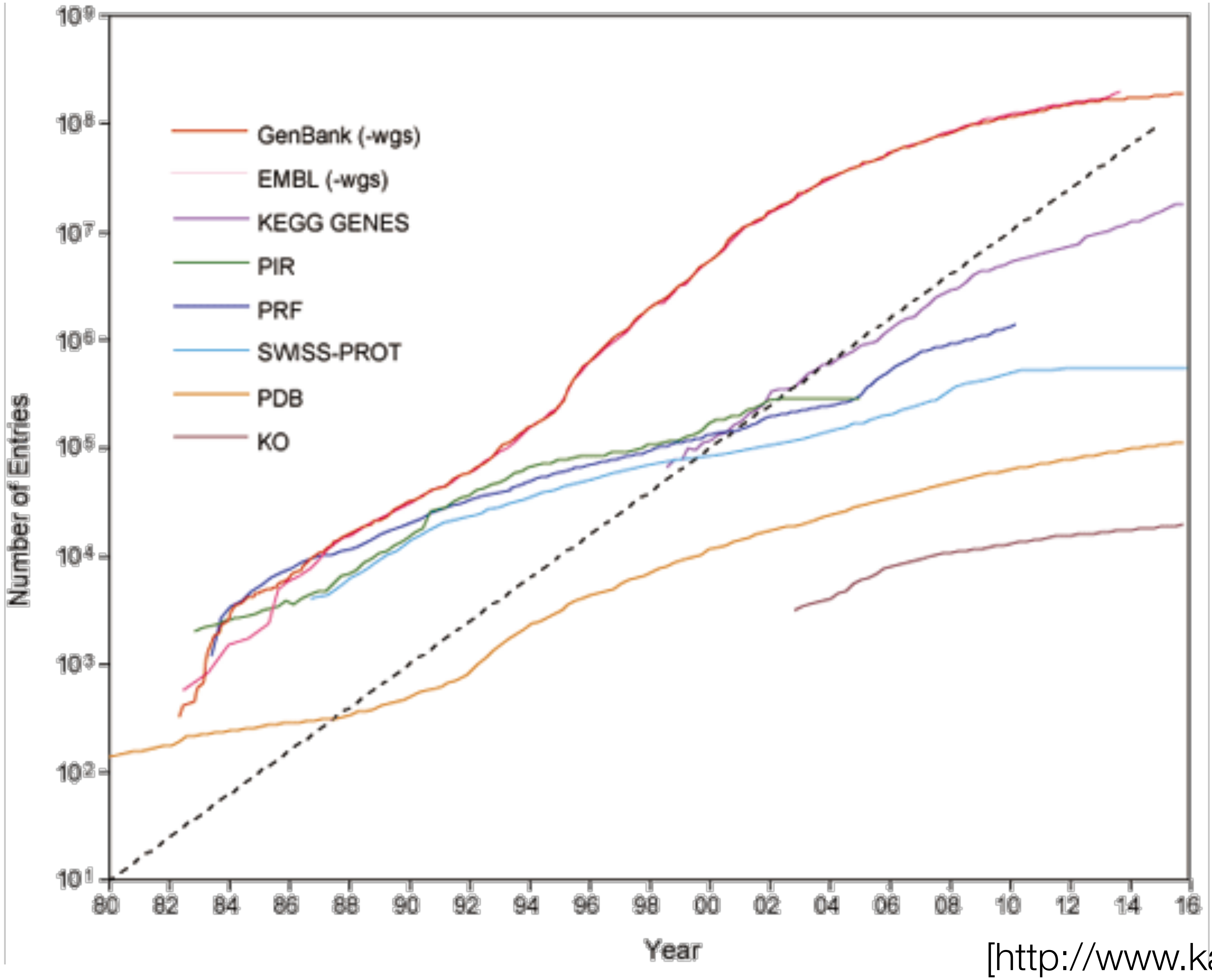
Data Extant (Shared or Exists)



[T. Vines et al., 2014]

Lots of Data is Shared…

Genome Sequence and Structure Data



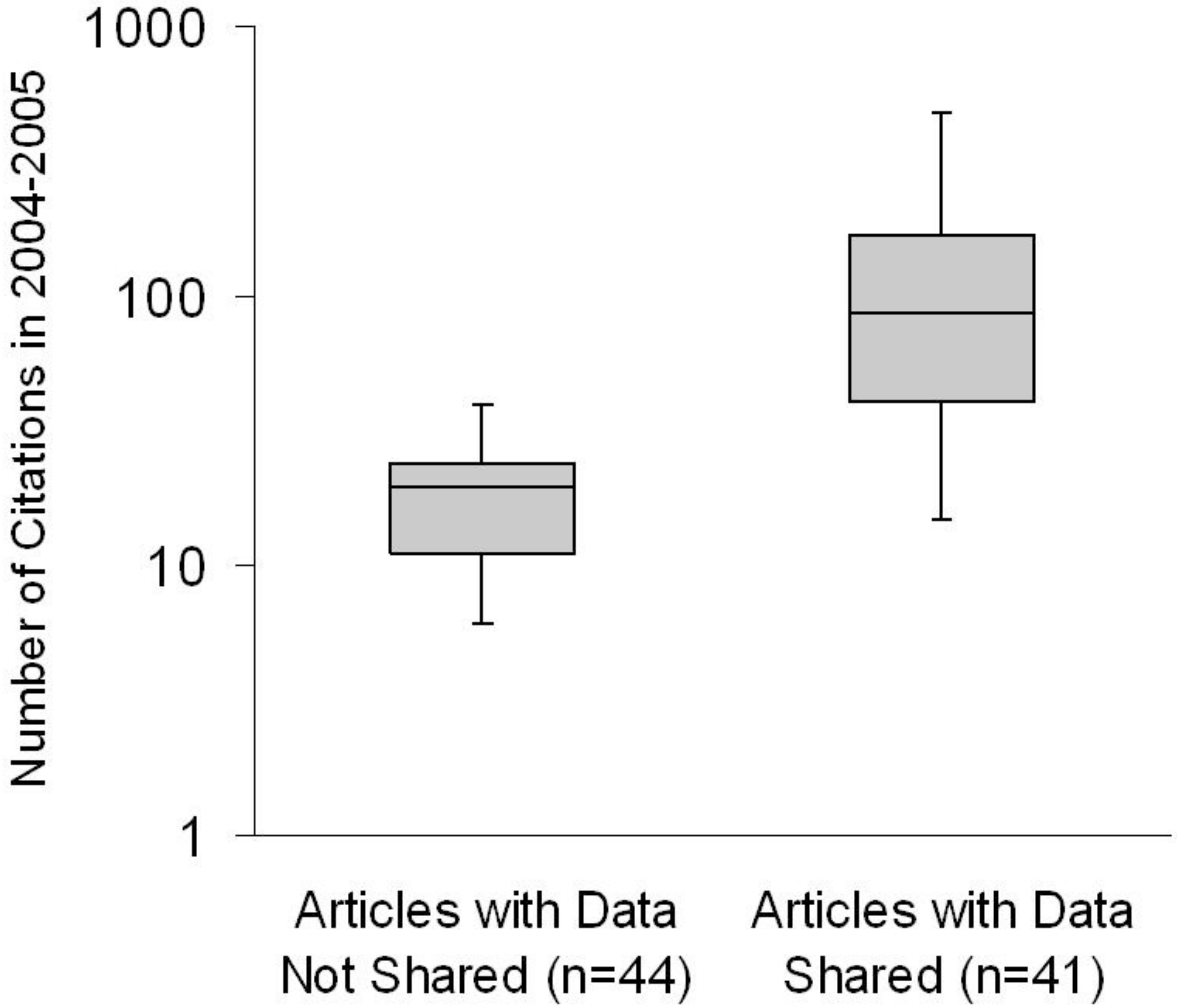
[http://www.kanehisa.jp/en/db_growth.html]

…but how much isn't shared?

- What isn't shared?
- Who isn't sharing?
- Why not?
- How much does it matter?
- What can be done about it?

[H. Piwowar, 2013]

Why Share Data? Increased Citations



[H. Piwowar, 2013]

What Factors Impact Sharing?

Funder

- funded by NIH?
- size of grant
- sharing plan req'd?
- funded by non-NIH?

Journal

- impact factor
- strength of policy
- open access?
- number of microarray studies published

Investigator

- years since first paper
- # pubs
- # citations
- previously shared?
- previously reused?
- gender

Institution

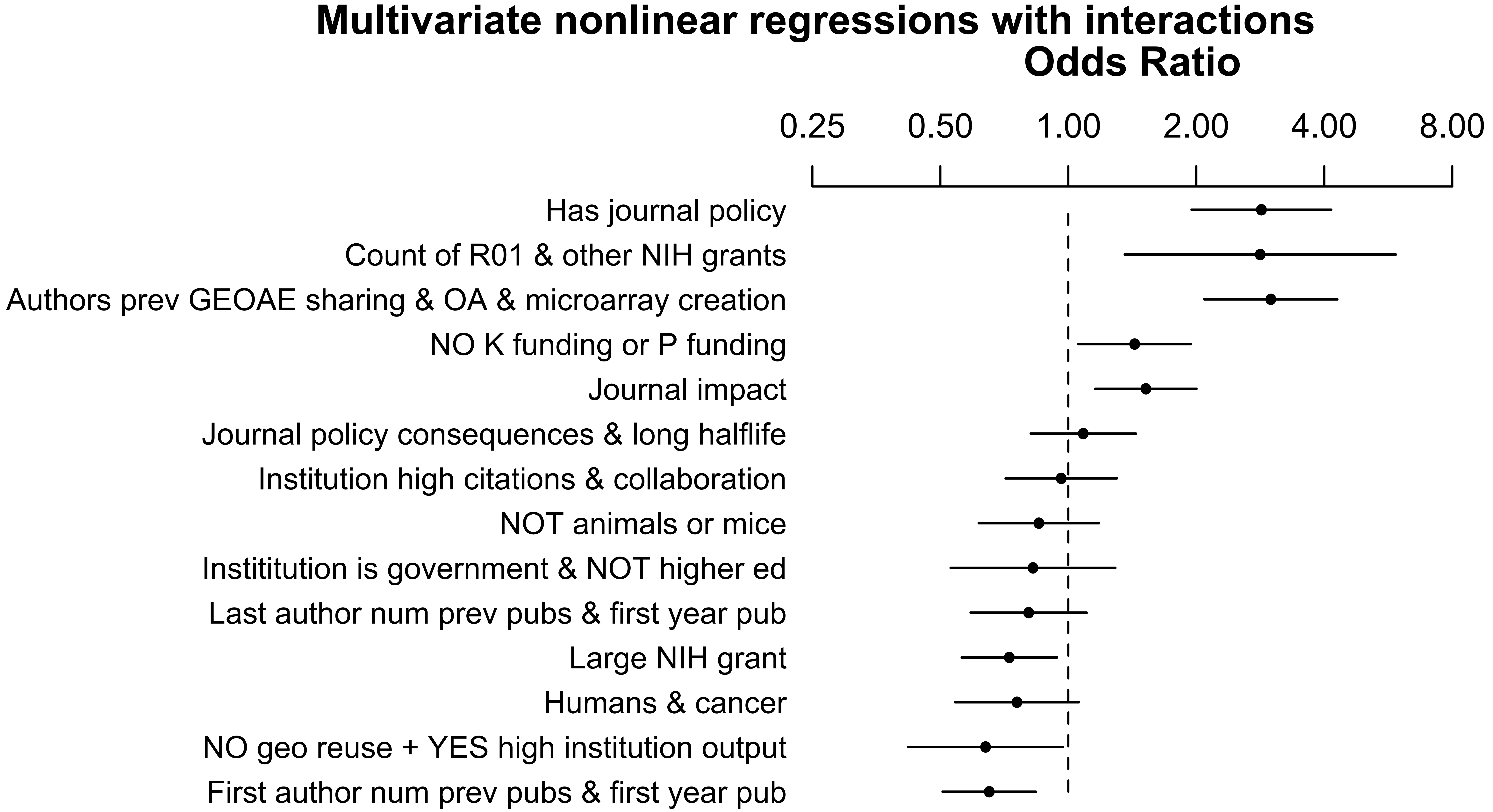
- sector
- size
- impact rank
- country

Study

- humans?
- mice?
- plants?
- cancer?
- clinical trial?
- number of authors
- year

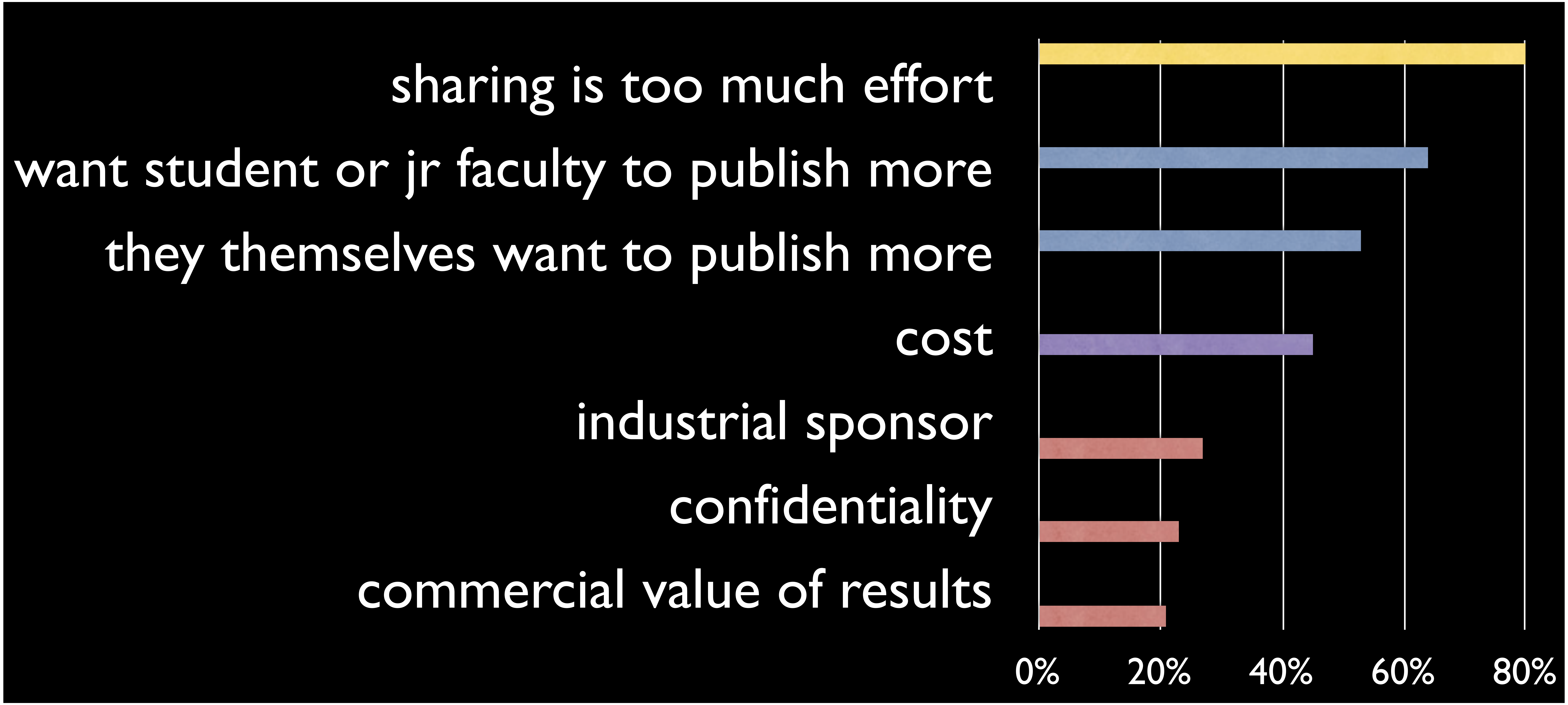
[H. Piwowar, 2013]

Factors



[H. Piwowar, 2013]

Why not data sharing? (self-reported)



[Campbell et al., 2002 via Piwowar, 2013]

Nature data availability and data citations

- Policy as of July 2016
- http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

The Evolution of Data Citation: From Principles to Implementation

M. Altman and M. Crosas

Data Sharing Policies

- *Science*:
  - "all data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of *Science*"
  - "**citations to unpublished data** and personal communications cannot be used to support claims in a published paper"
- Often this is only used as reason to retract work when issues arise
- Need:
  - Recognition of data authorship
  - Robust citation practices and infrastructure

Chronology of Data Citation

	Exemplar Systems	Core Principles	Key Work
1977-1998	*ICPR* *Archive* *MARC catalog systems.*	- Facilitate description & information retrieval - Describe data in archives - Describe as works not media - Provide author, title, version.	[Avram 1975] [Dodd 1979] [ISBD 1990] [ISO 1997]
1999-2003	*NESSTAR* *Virtual Data Center*	- Facilitate access & persistence - Cite research data in all publications that use it. - Provide actionable URI's - Provide persistent identifiers - Use persistent institutions	[Altman, et al. 2001] [Ryssevik & Musgrave 2001]
2004-2009	*TIB DOI Service* *Dataverse Network*	- Facilitate verification & reproducibility - Provide bit- or semantic- fixity - Provide granularity	[Brase 2004] [Buneman 2006] [Altman & King 2007]
2009-	*Dataverse Network* *DataCite* *Data Dryad* *FigShare* *Data Citation Index*	- Facilitate integration - Include data citations in standard locations in text - Index data citations in existing catalogs - Integrate data citation with	[Uhlir (ed.) 2012] [CODATA 2013] [Data Synthesis Group 2014]

Phases of Data Citation (1977-2009)

1. Support description and information retrieval: what should be included in a citation? (Libraries)
2. Support data access and persistence: if citations to data in publications, need methods to discover information about data
3. Support verification and **reproducibility**: allow verification of claims based on the data (wider integration into publishing

Joint Declaration of Data Citation Principles

1. **Importance**. Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.
2. **Credit and Attribution**. Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.
3. **Evidence**. In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.
4. **Unique Identification**. A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community.

Joint Declaration of Data Citation Principles

5. **Access**. Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data.
6. **Persistence**. Unique identifiers, and metadata describing the data, and its disposition, should persist -- even beyond the lifespan of the data they describe.

Joint Declaration of Data Citation Principles

7. **Specificity and Verifiability**. Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity sufficient to facilitate verifying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.
8. **Interoperability and flexibility**. Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities.

Generic Data Citation

- Author(s), Year, Dataset Title, Global Persistent Identifier, Data Repository or Archive, version or subset
- Authors, repository → Principle 2
- Year and title → not related to principle but consistent with other citations
- Global Persistent Identifier: Principle 4 and 6

More Information

- Provide via the web
  - Metadata
  - Fixity and provenance information
- Community Indices:
  - CrossRef
  - DataCite
- Structured Identifiers (ORCID, ISNI) preferred over unstructured metadata

Example Repositories with Citations

- Dryad, Dataverse, Figshare
- Dataverse:
  - Draft citation **automatically** generated
  - Includes versioning information

Remaining Challenges

- Provenance: chain of ownership
- Identity: equivalence and derivation relationships
  - Equivalence: if not bitwise equal, can data still be interchangeable?
  - Versioning: if data is updated, how to find updated version?
  - Granularity: How to describe subsets of data (deep citation)
- Attribution: ensure that the correct people and institutions receive credit

DataCite

www.datacite.org

Why Data Citation is a Computational Problem

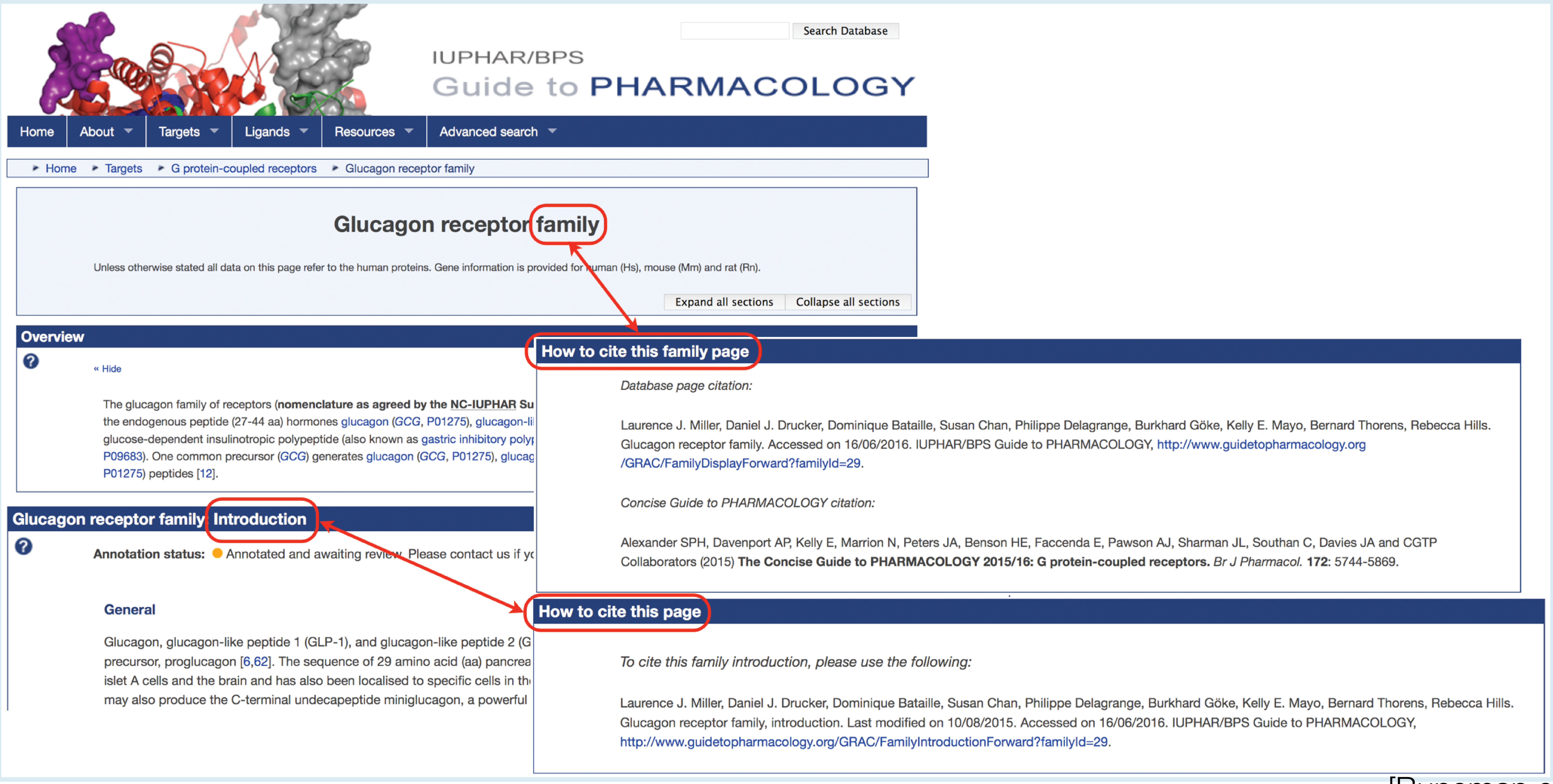
P. Buneman, S. Davidson, and J. Frew

Computational Data Citation

- Given a database D and a query Q, generate an appropriate citation.
- Automatic Citation requires the answers to two questions:
  - Does the citation depend on both Q and D or just on the data Q(D) extracted by Q from D?
  - If we have appropriate citations for some queries, can we use them to construct citations for other queries?
- If the data is an image or numbers, cannot expect the citation to live in that data
- If the query returns an empty dataset, we still may wish to cite that
- People know how to cite certain parts of a dataset but not all…

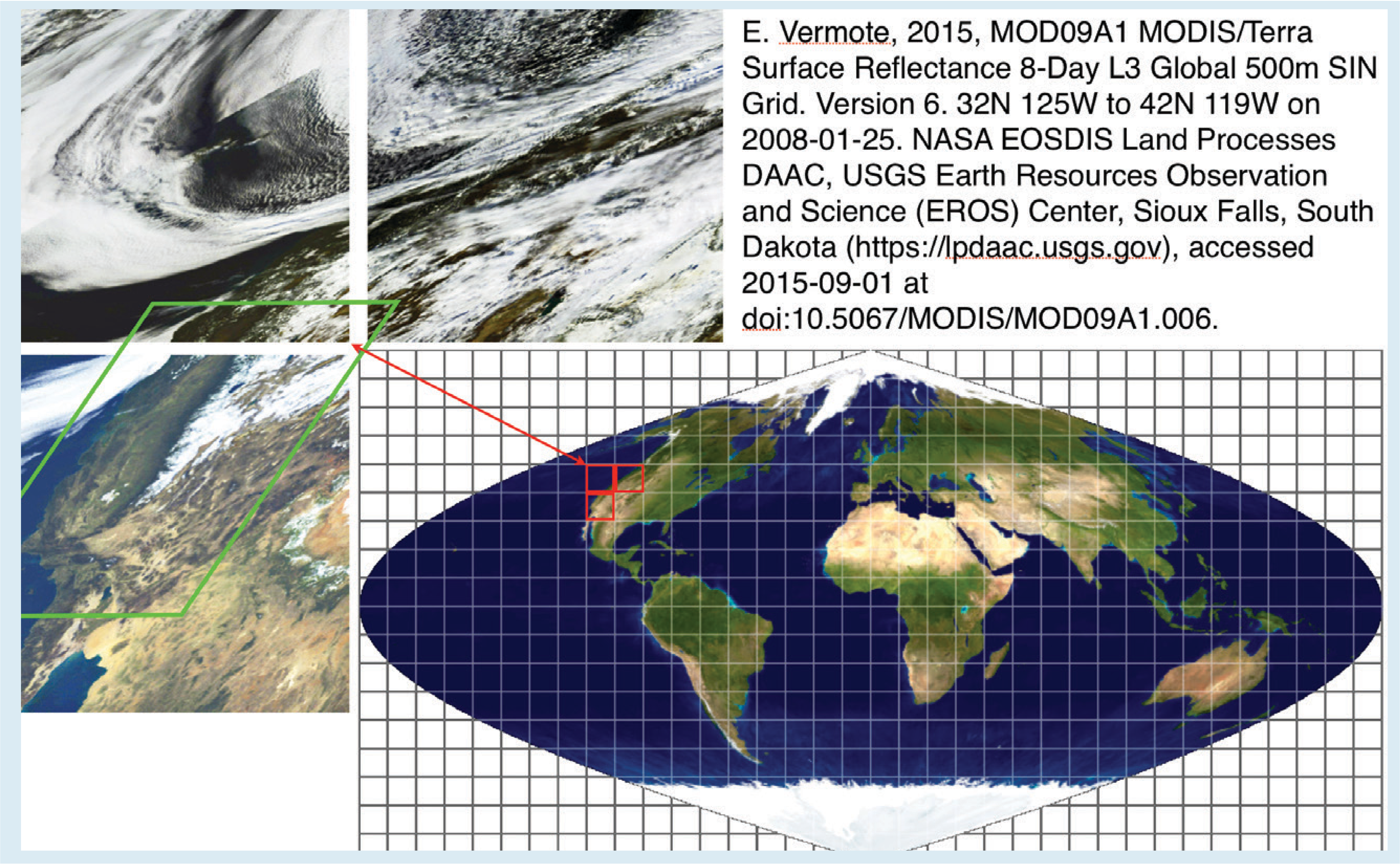
[Buneman et al., 2016]

Computational Data Citation (GtoPdb)



[Buneman et al., 2016]

Computational Data Citation (MODIS)



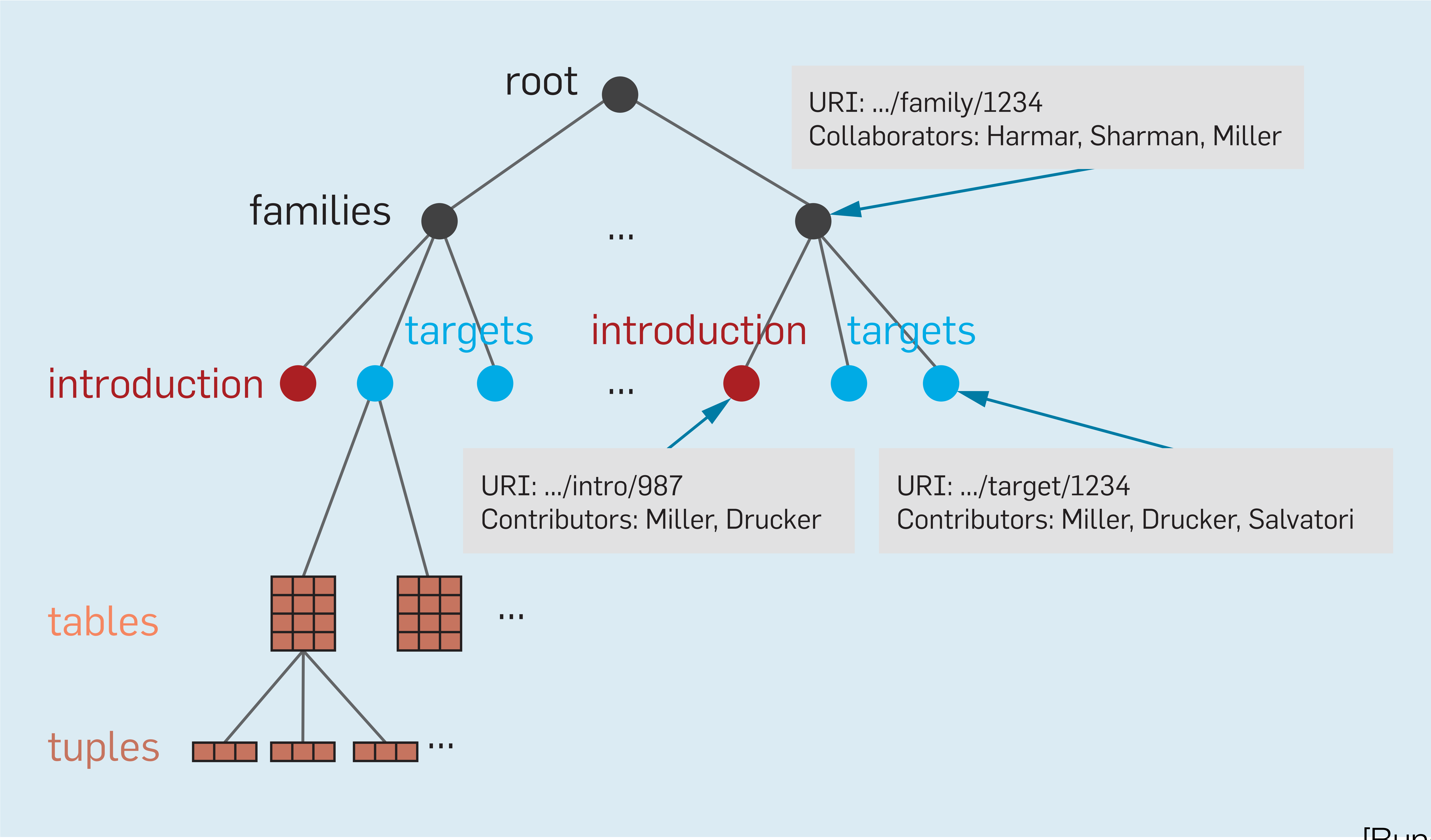
[Buneman et al., 2016]

Views and Citable Units

- Views describe "areas of responsibility" for parts of a database
- Use views to create "citable units"
- Determine which view V answers a particular query Q and generate a citation for the view
- What happens if two different views can answer the same query?

[Buneman et al., 2016]

Citable Views and Partial Citations



[Buneman et al., 2016]

Hierarchies of Views

- In GtoPdb, three classes of views
- Family view:
  - /Root/Family[FamilyName=$$f]
- Introduction view:
  - /Root/Family[FamilyName=$$f]/ Introduction
- Target view:
  - /Root/Family[FamilyName=$$f]/ Target[TargetName=$$t]

Citation Rule and Partial Result (GtoPdb)

- Rule:
  - { Title: "IUPHAR/BPS Guide to Pharmacology", Version: $v,
    Family: $$f, Contributors: $a, URI: "www.iuphar.org" }
    ←
    /Root[VersionNumber: $v]/Family[FamilyName: $$f]/Introduction[Contributor-list: $a]
- Citation:
  - { Title: "IUPHAR/BPS Guide to Pharmacology", Version: 26, Family: "Calcitonin",
    Contributors: ["Debbie Hay", "David R. Poyner"], URI: "www.iuphar.org" }

[Buneman et al., 2016]

Citation Rule and Sample Result (MODIS)

- { author: m_auth($p,$$v), m_year:($p,$$v), title: m_title($p), version: $v,
  bounding-box : [$$minlong, $$minlat, $$maxlong, $$maxlat],
  interval: [$$mint, $$maxt], organization: m_org($p), url: m_url($p),
  accessed: DATE(), doi = m_doi($p,$$v) }
  ←
  /root/product[ProdName=$p]/version[vnum=$$v]
  /file[Lat ≥ $$minlat and Lat ≤ $$maxlat and
  Lon ≥ $$minlon and Lon ≤ $$maxlon and
  Time ≥ $$mint and Time ≤ $$maxt]
- { author: “E. Vermote”, title: “MOD09A1 ... SIN Grid”, version: 6,
  bounding-box: [-125, 32, -119, 42],
  interval: [2008-01-25, 2008-01-25],
  organization: “NASA EOSDIS ... South Dakota”, URL: “https://lpdaac.usgs.gov”,
  accessed: “2015-09-01”, doi: “10.5067/MODIS/MOD09A1.006” }

[Buneman et al., 2016]