Advanced Data Management (CSCI 490/680)

Data Curation

Dr. David Koop

Data Fusion: What about Copying?

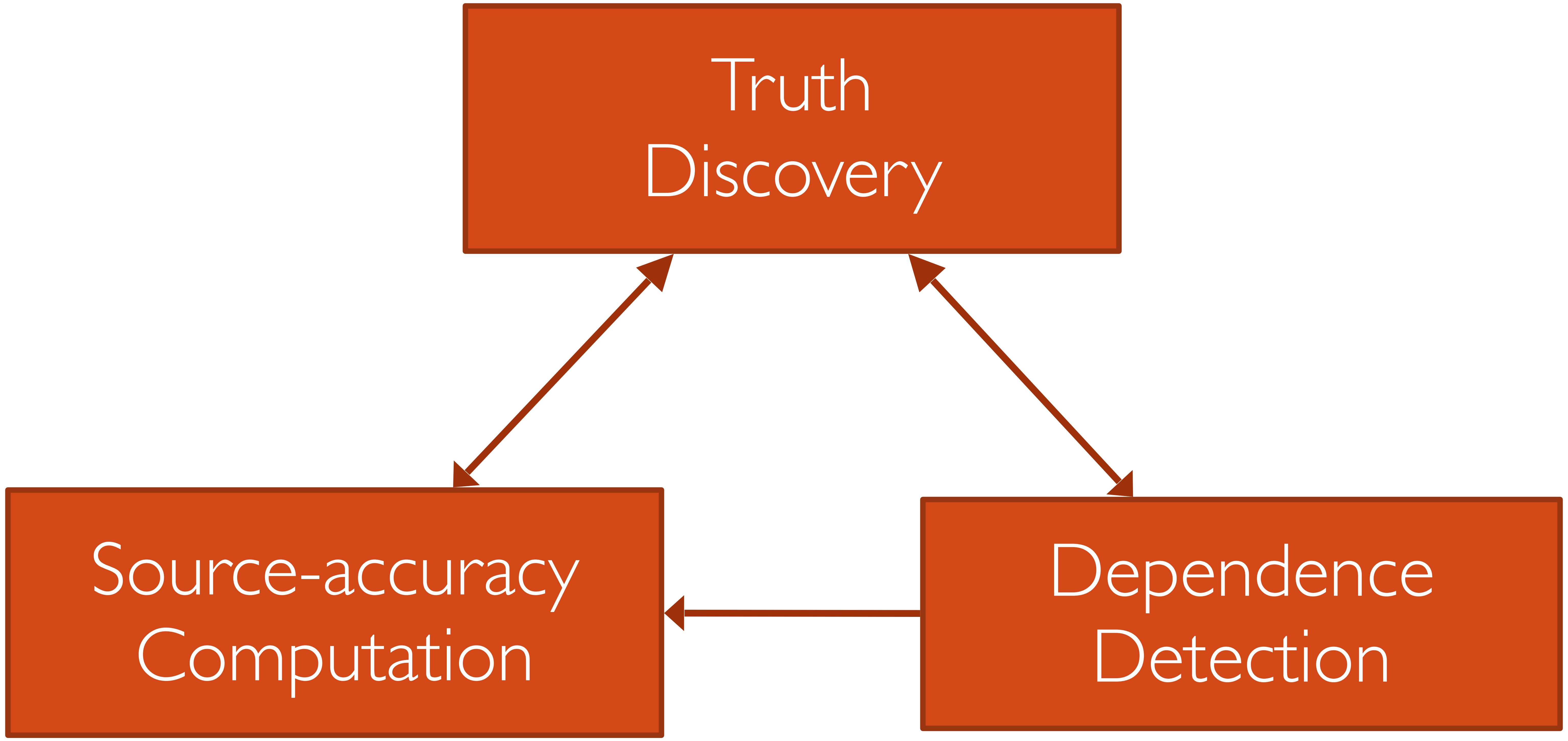
[X L Dong et al., 2009]

Data Fusion: What about Copying?

	S1	S2	S3	S4	S5
Stonebraker	MIT	Berkeley	MIT	MIT	MS
Dewitt	MSR	MSR	UWisc	UWisc	UWisc
Bernstein	MSR	MSR	MSR	MSR	MSR
Carey	UCI	AT&T	BEA	BEA	BEA
Halevy	Google	Google	UW	UW	UW

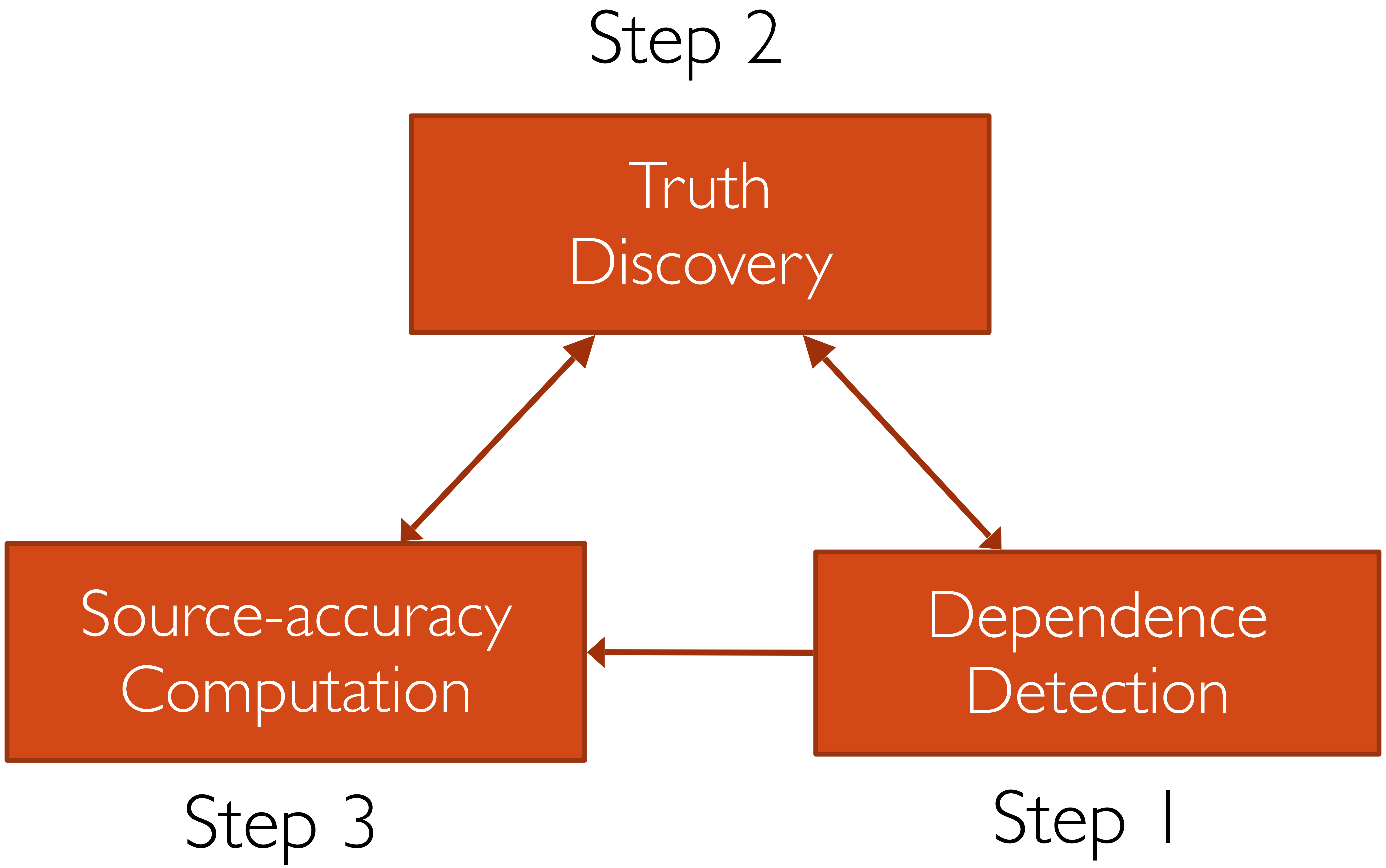
[X L Dong et al., 2009]

Data Fusion: Source Dependence and Accuracy



[X L Dong et al., 2009]

Data Fusion: Source Dependence and Accuracy



Step 2

Truth Discovery

Source-accuracy Computation

Dependence Detection

Step 3

Step 1

[X L Dong et al., 2009]

Data Fusion Example

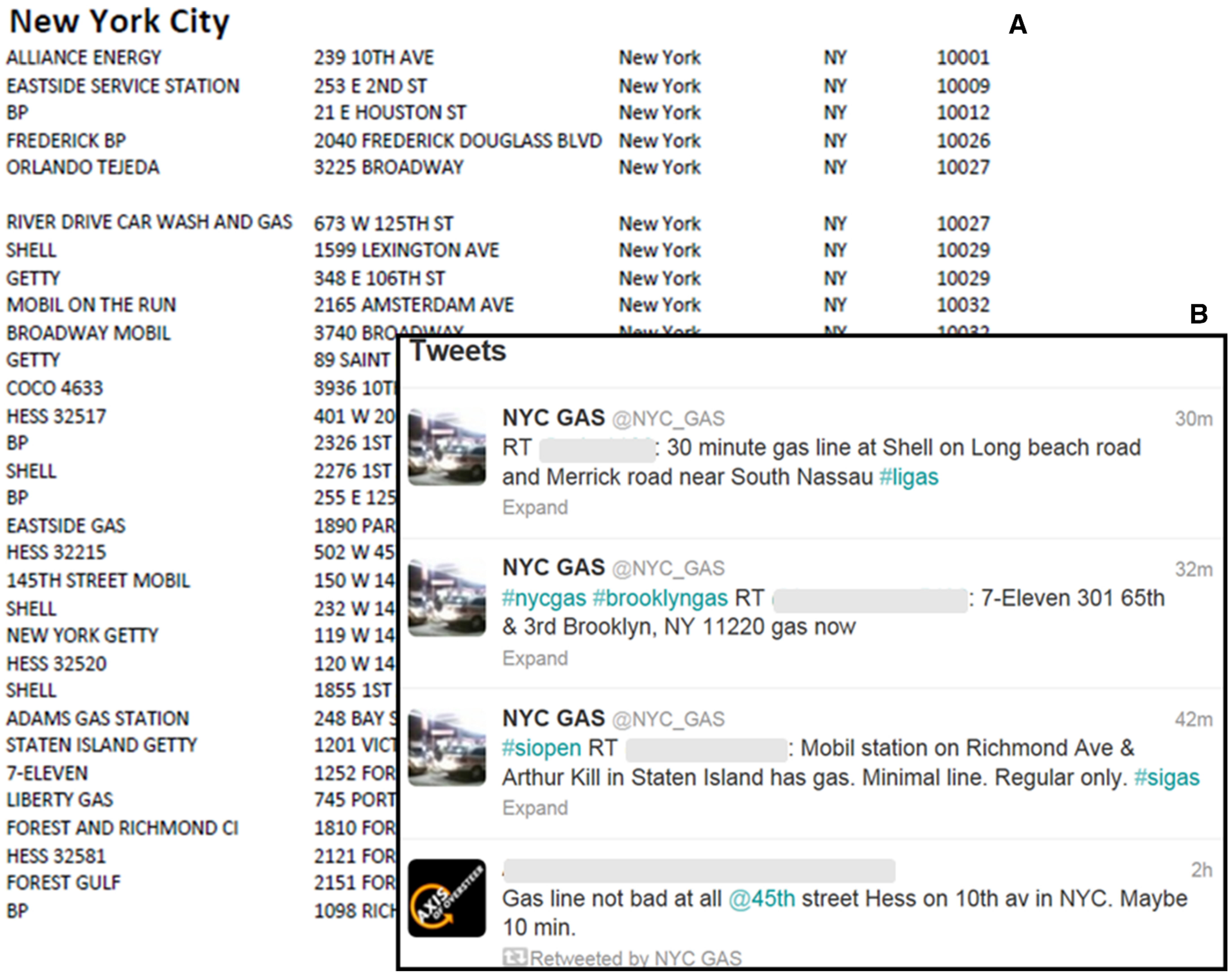
Accuracy	S1	S2	S3	S4	S5
Round 1	.52	.42	.53	.53	.53
Round 2	.63	.46	.55	.55	.55
Round 3	.71	.52	.53	.53	.37
Round 4	.79	.57	.48	.48	.31
…	…	…	…	…	…
Round 11	.97	.61	.40	.40	.21

Value Confidence	Carey			Halevy	
	UCI	AT&T	BEA	**Google**	UW
Round 1	1.61	1.61	2.0	2.1	2.0
Round 2	1.68	1.3	2.12	2.74	2.12
Round 3	2.12	1.47	2.24	3.59	2.24
Round 4	2.51	1.68	2.14	4.01	2.14
…	…	…	…	…	…
Round 11	4.73	2.08	1.47	6.67	1.47

[X L Dong et al., 2009]

How do we find datasets?

Goal of Dataset Search: Accurate (A) vs. Timely (B)



New York City

ALLIANCE ENERGY	239 10TH AVE	New York	NY	10001
EASTSIDE SERVICE STATION	253 E 2ND ST	New York	NY	10009
BP	21 E HOUSTON ST	New York	NY	10012
FREDERICK BP	2040 FREDERICK DOUGLASS BLVD	New York	NY	10026
ORLANDO TEJEDA	3225 BROADWAY	New York	NY	10027
RIVER DRIVE CAR WASH AND GAS	673 W 125TH ST	New York	NY	10027
SHELL	1599 LEXINGTON AVE	New York	NY	10029
GETTY	348 E 106TH ST	New York	NY	10029
MOBIL ON THE RUN	2165 AMSTERDAM AVE	New York	NY	10032
BROADWAY MOBIL	3740 BROADWAY	New York	NY	10032
GETTY	89 SAINT			
COCO 4633	3936 10T			
HESS 32517	401 W 20			
BP	2326 1ST			
SHELL	2276 1ST			
BP	255 E 125			
EASTSIDE GAS	1890 PAR			
HESS 32215	502 W 45			
145TH STREET MOBIL	150 W 14			
SHELL	232 W 14			
NEW YORK GETTY	119 W 14			
HESS 32520	120 W 14			
SHELL	1855 1ST			
ADAMS GAS STATION	248 BAY S			
STATEN ISLAND GETTY	1201 VICT			
7-ELEVEN	1252 FOR			
LIBERTY GAS	745 PORT			
FOREST AND RICHMOND CI	1810 FOR			
HESS 32581	2121 FOR			
FOREST GULF	2151 FOR			
BP	1098 RICH			

Tweets

NYC GAS @NYC_GAS 30m
RT [redacted]: 30 minute gas line at Shell on Long beach road and Merrick road near South Nassau #ligas
Expand

NYC GAS @NYC_GAS 32m
#nycgas #brooklyngas RT [redacted]: 7-Eleven 301 65th & 3rd Brooklyn, NY 11220 gas now
Expand

NYC GAS @NYC_GAS 42m
#siopen RT [redacted]: Mobil station on Richmond Ave & Arthur Kill in Staten Island has gas. Minimal line. Regular only. #sigas
Expand

[redacted] 2h
Gas line not bad at all @45th street Hess on 10th av in NYC. Maybe 10 min.
Retweeted by NYC GAS

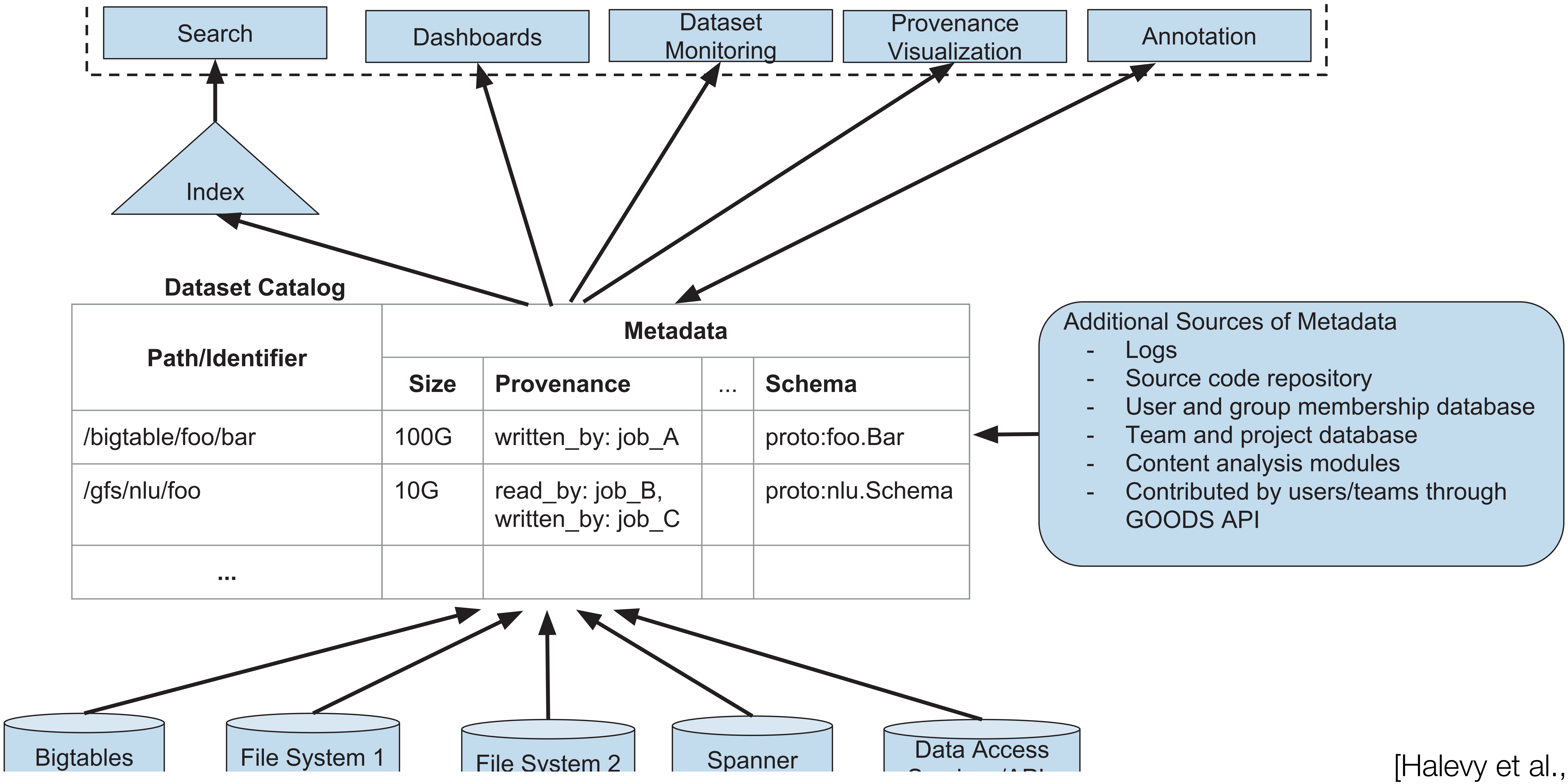
[Chapman et al., 2020]

Goods: Organizing Google's Datasets

- Tool for Google to help its employees find internal data
- Keep data where it is, how it is, but extract metadata to aid search
- Challenges:
    - Dataset size and scale: >26 billion datasets
    - Variety: formats (text, csv, Bigtable), storage (GoogleFS, db server)
    - Churn: ~5% of datasets deleted each day
    - Metadata uncertainty: protocol buffers, primary key identification
    - Computing importance: need to understand users
    - Recovering semantics: understanding the data aids metadata extraction

[Halevy et al., 2016]

Goods: Organizing Google's Datasets



Dataset Catalog

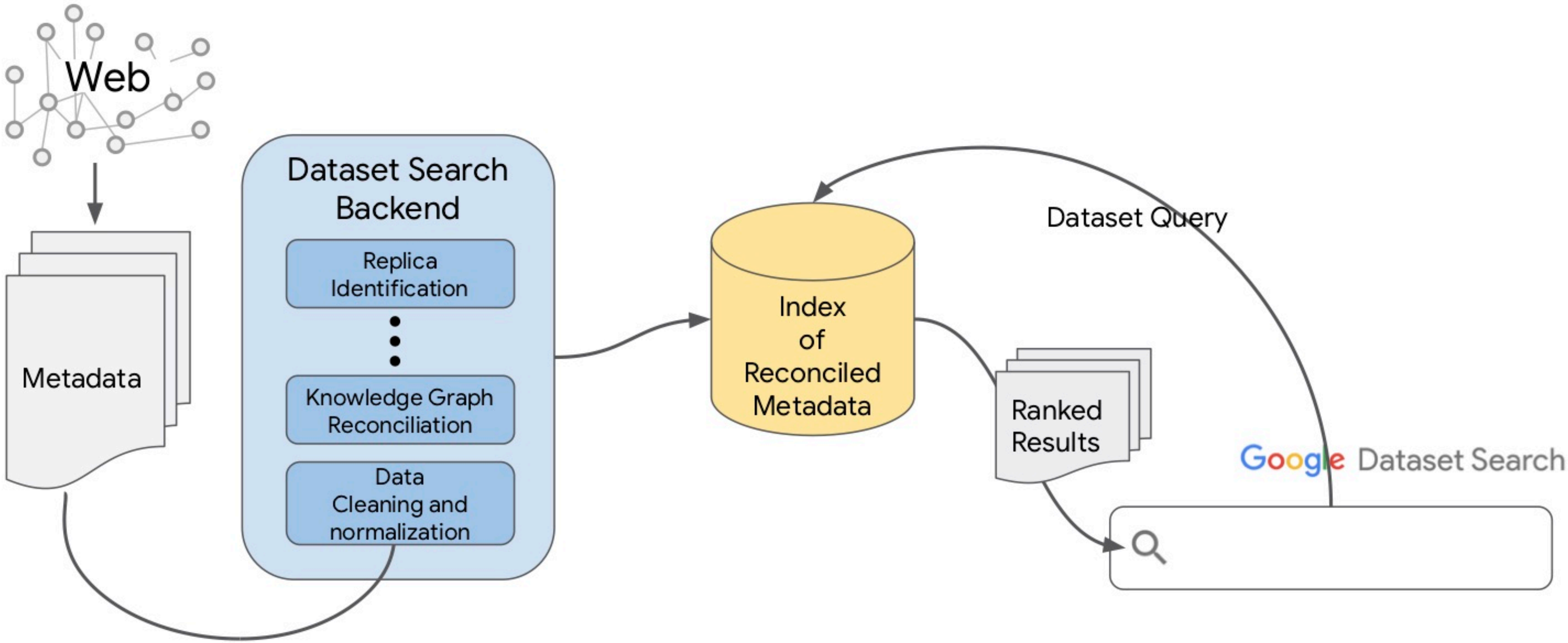
Path/Identifier	**Metadata**			
	Size	**Provenance**	...	**Schema**
/bigtable/foo/bar	100G	written_by: job_A		proto:foo.Bar
/gfs/nlu/foo	10G	read_by: job_B, written_by: job_C		proto:nlu.Schema
...				

Additional Sources of Metadata

- Logs
- Source code repository
- User and group membership database
- Team and project database
- Content analysis modules
- Contributed by users/teams through GOODS API

[Halevy et al., 2016]

Google Dataset Search Overview



[N. Noy et al., 2019]

Requirements

- System must be **open** so new providers can add their own datasets
- Search is over **metadata** (a provider may require users to pay/create account)
- Metadata must be published by the data publishers themselves, adhering to a **standard**

[N. Noy et al., 2019]

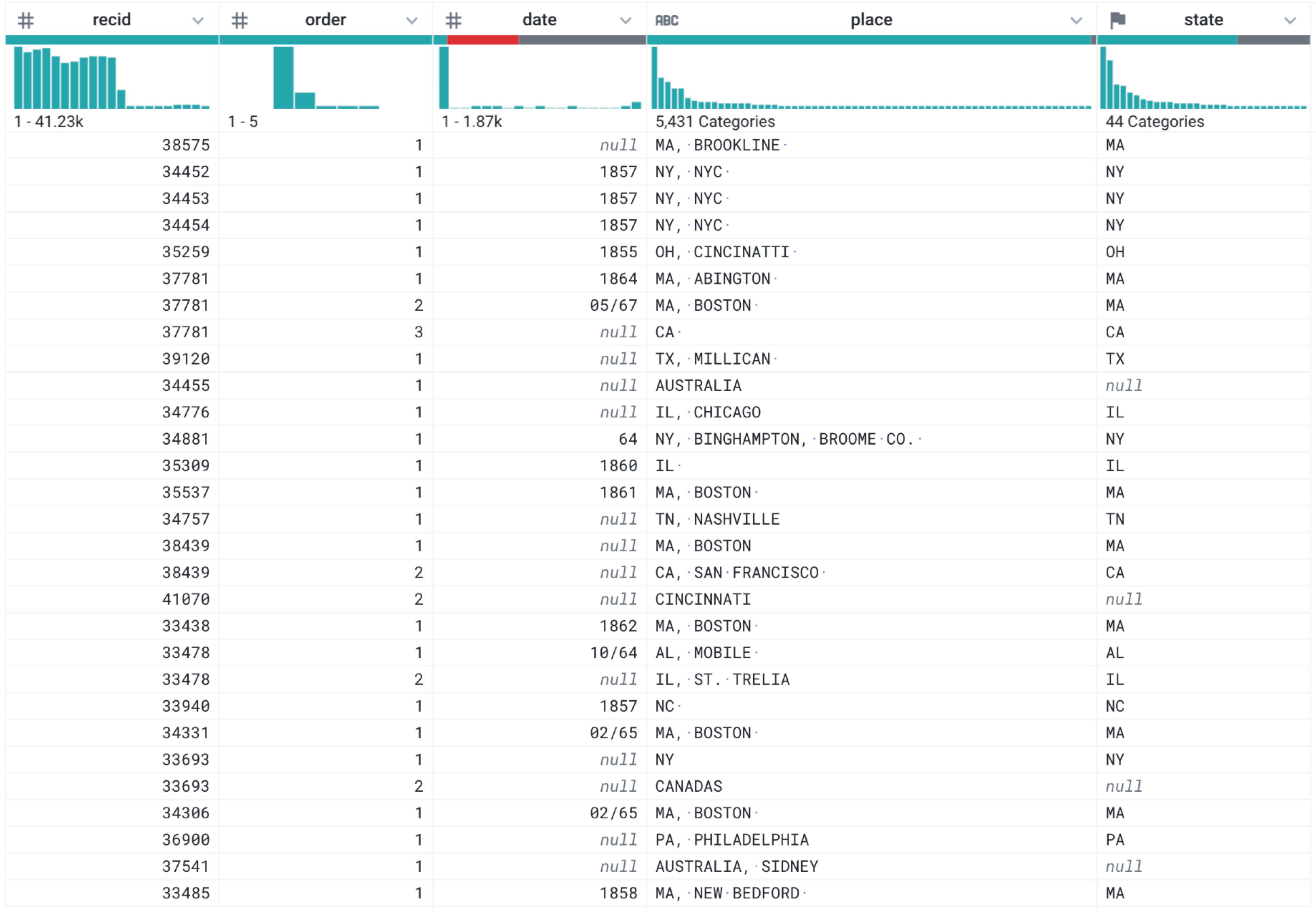
Challenges

- Metadata Quality: providers don't adhere to the specs
- Metadata Duplication in Search Results: search results vs. profile pages
- Dataset Replication and Provenance: identify replicas across providers
- Churn and Stale Sites:
  - 3% deleted, 7-10% added per day
  - standard web crawlers check high-traffic sites more often
- Ranking/Relevance: data citation might help
- Multiple Dataset-Metadata Standards: schema.org vs DCAT

[N. Noy et al., 2019]

Assignment 3

- Due Today
- Same Info Wanted data
- Data wrangling with
  - Trifacta Wrangler
  - pandas
- For place, date extraction: 2 regexs, don't try to standardize anything, CS680 need to extract place details, date is EC



recid	order	date	place	state
1 - 41.23k	1 - 5	1 - 1.87k	5,431 Categories	44 Categories
38575	1	null	MA, BROOKLINE	MA
34452	1	1857	NY, NYC	NY
34453	1	1857	NY, NYC	NY
34454	1	1857	NY, NYC	NY
35259	1	1855	OH, CINCINATTI	OH
37781	1	1864	MA, ABINGTON	MA
37781	2	05/67	MA, BOSTON	MA
37781	3	null	CA	CA
39120	1	null	TX, MILLICAN	TX
34455	1	null	AUSTRALIA	null
34776	1	null	IL, CHICAGO	IL
34881	1	64	NY, BINGHAMPTON, BROOME CO.	NY
35309	1	1860	IL	IL
35537	1	1861	MA, BOSTON	MA
34757	1	null	TN, NASHVILLE	TN
38439	1	null	MA, BOSTON	MA
38439	2	null	CA, SAN FRANCISCO	CA
41070	2	null	CINCINNATI	null
33438	1	1862	MA, BOSTON	MA
33478	1	10/64	AL, MOBILE	AL
33478	2	null	IL, ST. TRELIA	IL
33940	1	1857	NC	NC
34331	1	02/65	MA, BOSTON	MA
33693	1	null	NY	NY
33693	2	null	CANADAS	null
34306	1	02/65	MA, BOSTON	MA
36900	1	null	PA, PHILADELPHIA	PA
37541	1	null	AUSTRALIA, SIDNEY	null
33485	1	1858	MA, NEW BEDFORD	MA

Data Curation

Why?

Big Data, Little Data, or No Data?

C. L. Borgman

What is data and why share it?

- "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]

Data attribution and citation

- Publications are counted, authorship is negotiated
- For data:
  - Often compound
  - Ownership is rarely clear
  - Attribution?
  - What about derived data?
- Bibliometrics and Altmetrics

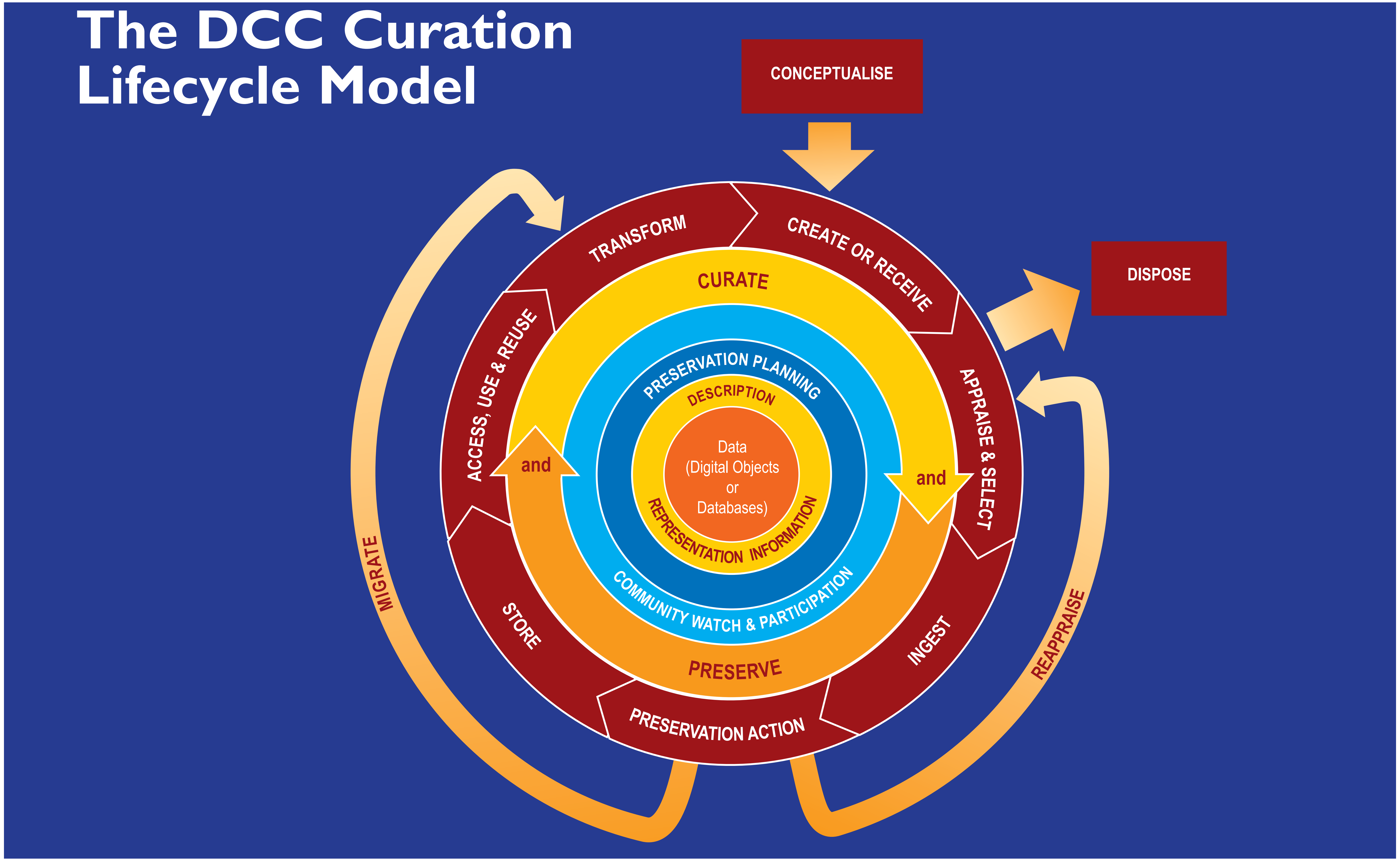
Data Identity

- Identifiers: DOIs, URIs
- Naming and namespaces: ORCID, KEGG Identifier
- Description: Metadata, Self-describing

Data Persistence

- How long should this data be kept?
  - Perishable
  - Long-lived
  - Permanent
- Who is responsible for keeping the data?
  - Scientists/investigators?
  - Publishers?
  - Librarians?
- Privacy should be considered from the beginning

Data Curation Lifecycle



[DCC]

Data (Digital Objects or Databases)

- Data, any information in binary digital form, is at the centre of the Curation Lifecycle. This includes:
  - **Digital Objects**
    - Simple Digital Objects are discrete digital items; such as textual files, images or sound files, along with their related identifiers and metadata.
    - Complex Digital Objects are discrete digital objects, made by combining a number of other digital objects, such as websites.
  - **Databases**: Structured collections of records or data stored in a computer system.

[DCC]

Full Lifecycle Actions

- Description and Representation Information: Assign metadata, using appropriate standards, to ensure adequate description and control
- Preservation Planning: Plan for preservation throughout the curation lifecycle of digital material
- Community Watch and Participation: Watch standards, tools, software.
- Curate and Preserve: Promote curation and preservation throughout the curation lifecycle

[DCC]

Sequential Actions

- 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]

Occasional Actions

- Dispose: Transfer to another archive or perhaps destroy data
- Reappraise: Return data which fails validation procedures for further appraisal and reelection
- Migrate: Migrate data to a different format—ensure the data's immunity from hardware or software obsolescence

[DCC]

The FAIR Guiding Principles for Scientific Data Management and Stewardship

M. D. Wilkinson et al.

Who and Why?

- Who: People from academia, industry, funding agencies, & scholarly publishers
- Why?
  - Data management leads to knowledge discovery, innovation, and reuse
  - Existing digital ecosystem **prevents** maximum benefit
  - Need to specify what "good" data management/curation/stewardship is
  - Enhance the ability of machines to automatically find and use the data
  - Principles should also apply to **tools**

[M. D. Wilkinson et al., 2016]

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]

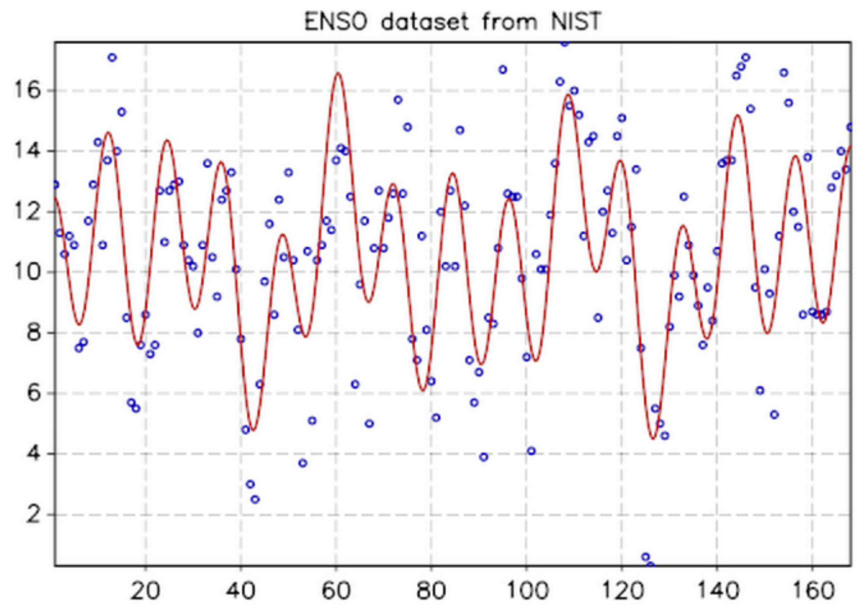
To be Findable

- F1. (Meta)data are assigned a **globally unique and persistent identifier**
- F2. Data are described with **rich metadata** (defined by R1)
- F3. Metadata clearly and explicitly include the **identifier** of the data it describes
- F4. (Meta)data are **registered or indexed** in a searchable resource

[M. D. Wilkinson et al., 2016]

DataCite Workflow

1. Take a dataset



2. Describe it

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

3. Assign a DOI

10.1234/exampledata

4. Reuse and reference!

ATLAS Collaboration, "Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC: $H \to \gamma\gamma$," http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44

☑ Unique ☑ Persistent

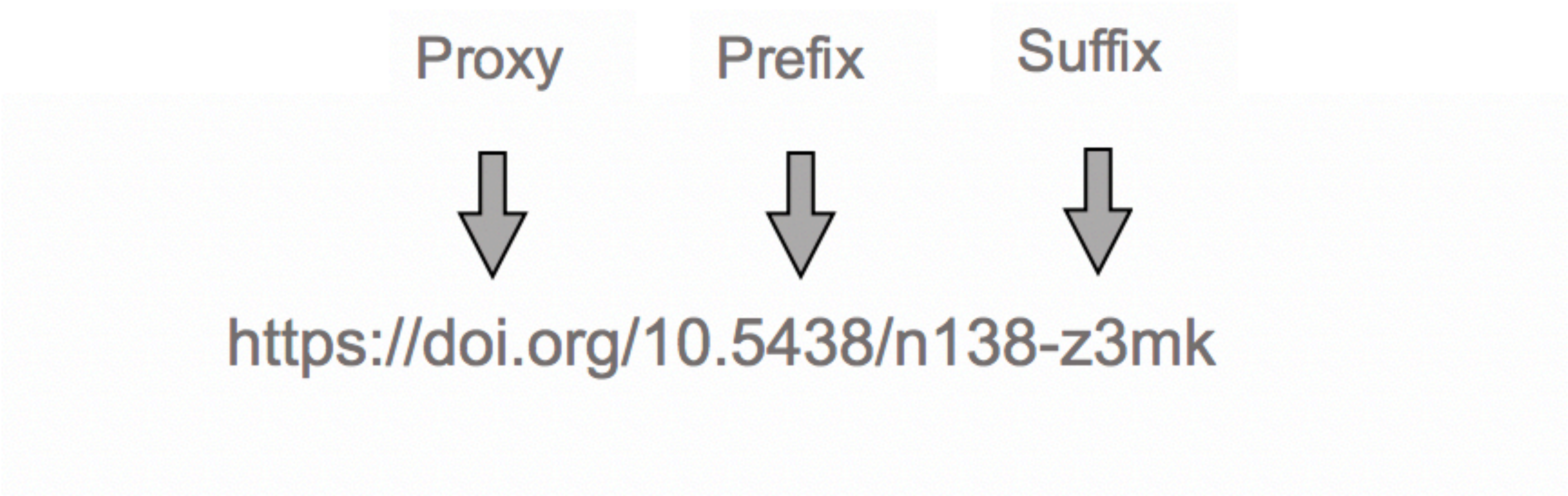
5. Enjoy the benefits

Findability	Track citations
Reusability	Measure impact

[DataCite]

Digital Object Identifier

- Name: Proxy + Prefix + Suffix



- Metadata: description of the object
- URL: resolves to a digital location, which contains object's details

[DataCite]

DataCite Metadata

Mandatory Properties	Details
Identifier	with mandatory type sub-property
Creator	with optional name identifier and affiliation sub-properties
Title	with optional type sub-properties
Publisher	
PublicationYear	
ResourceType	with mandatory general type description sub-property

Recommended Properties	Details
Subject	with scheme sub-property
Contributor	with type, name identifier, and affiliation sub-properties
Date	with type sub-property
RelatedIdentifier	with type and relation type sub-properties
Description	with type sub-property
GeoLocation	with point, box, and polygon sub-properties

Optional Properties
Language
AlternateIdentifier
Size
Format
Version
Rights
FundingReference

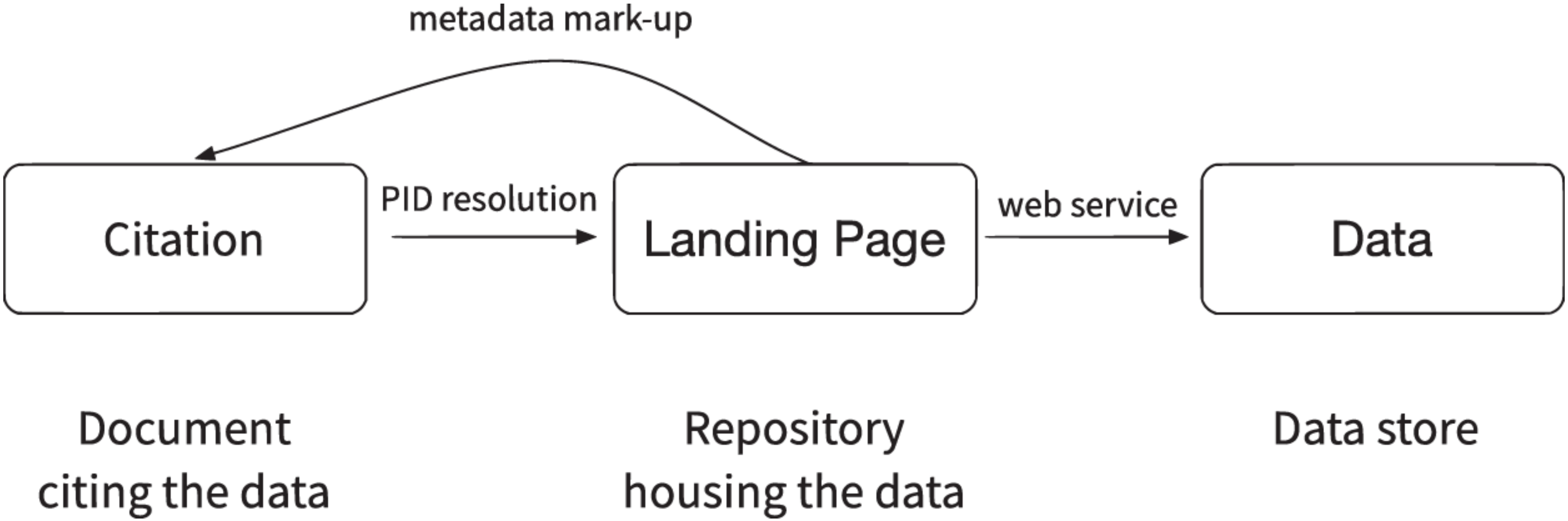
[DataCite]

To be Accessible

- A1. (Meta)data are **retrievable** by their identifier using a standardized communications protocol
  - A1.1. The protocol is **open**, free, and universally implementable
  - A1.2. The protocol allows for an **authentication** and authorization procedure, where necessary
- A2. Metadata are accessible, even when the data are **no longer available**

[M. D. Wilkinson et al., 2016]

How data accessibility might work within publications



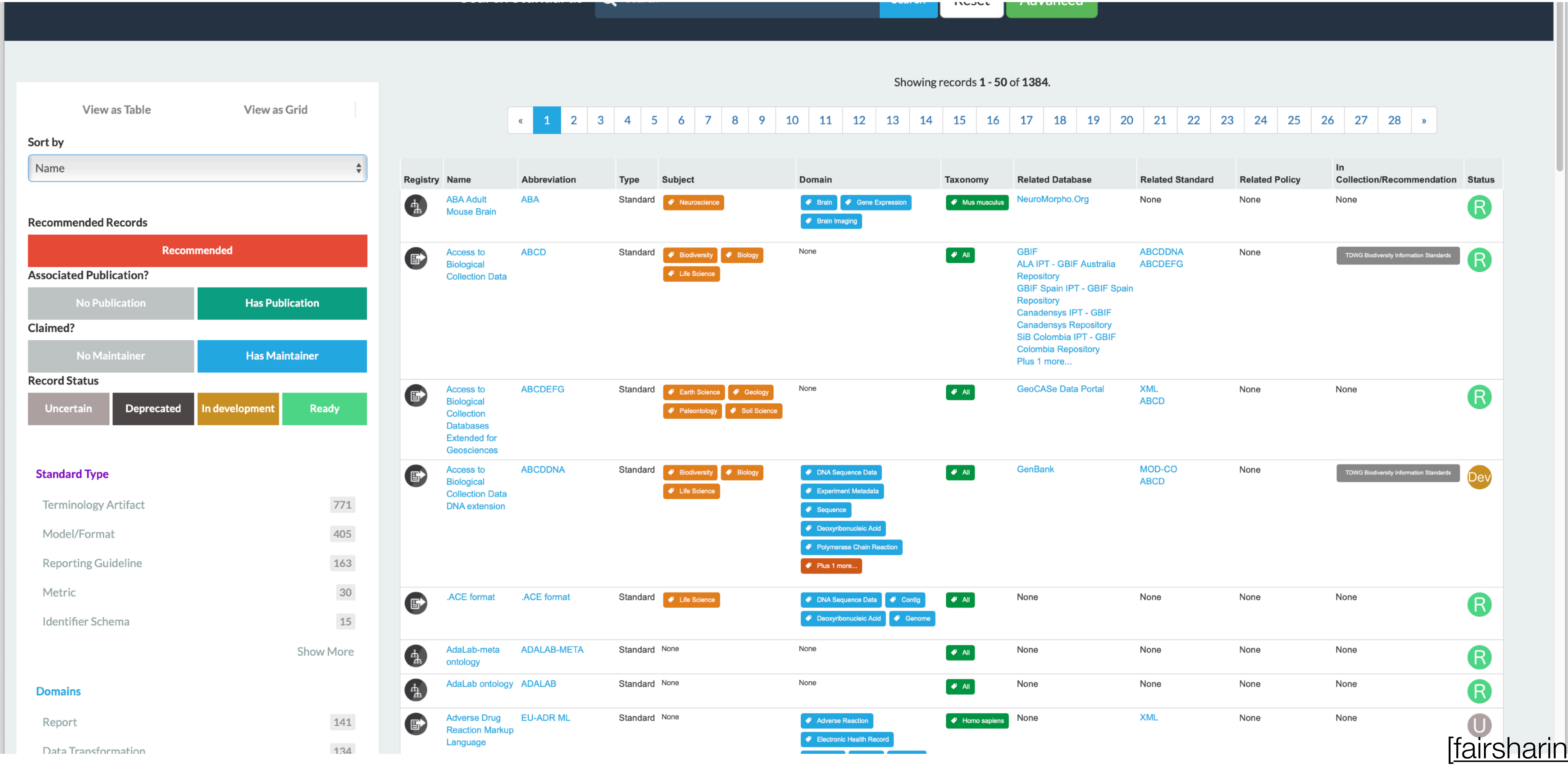
[M. Fenner et al., 2019]

To be Interoperable

- I1. (Meta)data use a formal, accessible, shared, and broadly applicable **language** for knowledge representation.
- I2. (Meta)data use **vocabularies** that follow FAIR principles
- I3. (Meta)data include **qualified references** to other (meta)data

[M. D. Wilkinson et al., 2016]

Standard vocabularies



Showing records **1 - 50** of **1384**.

« 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 »

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

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]

To be Reusable

- R1. (Meta)data are richly described with a plurality of accurate and relevant attributes
  - R1.1. (Meta)data are released with a clear and accessible data usage **license**
  - R1.2. (Meta)data are associated with detailed **provenance**
  - R1.3. (Meta)data meet domain-relevant **community standards**

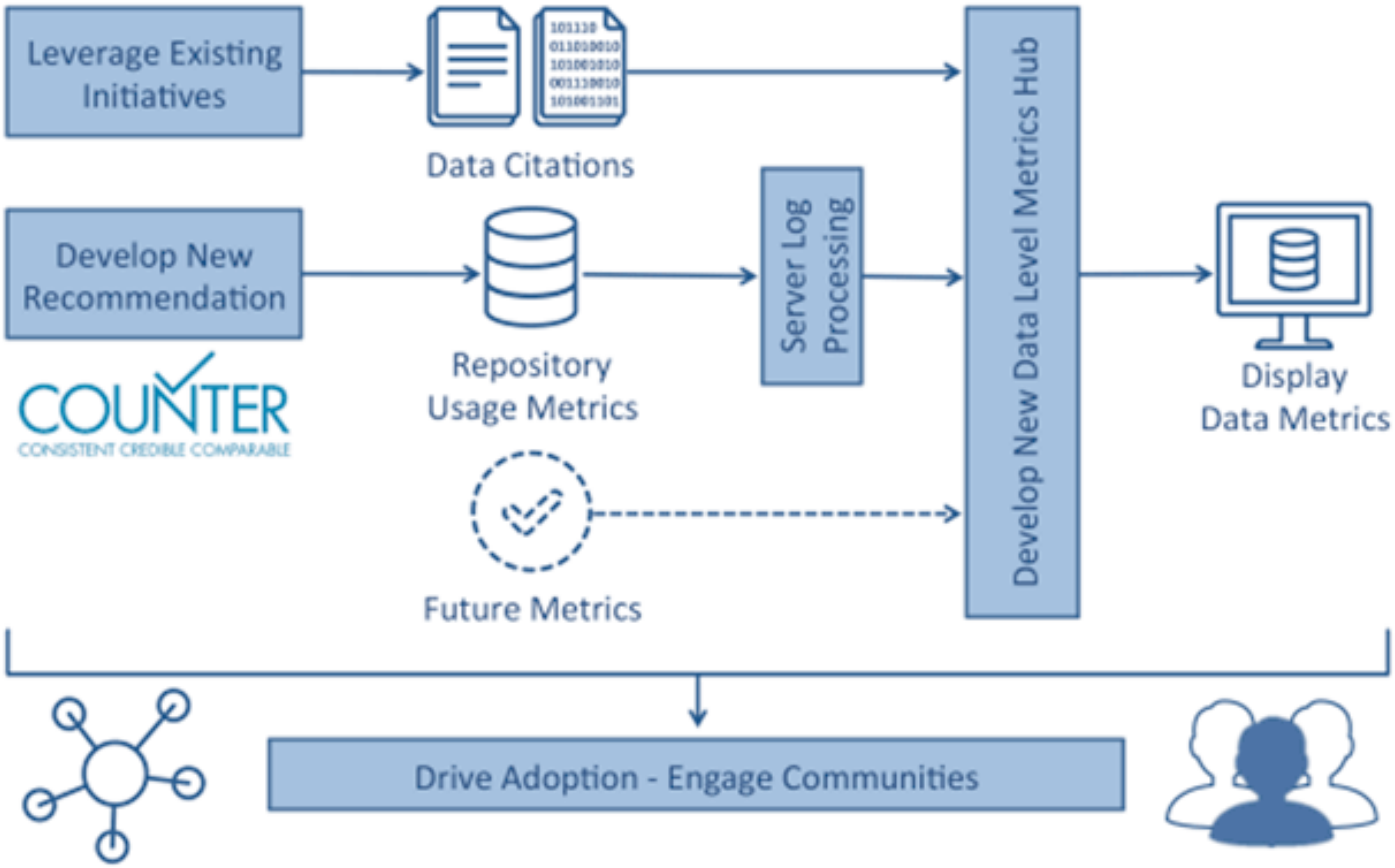
[M. D. Wilkinson et al., 2016]

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]

Make Data Count


Leverage Existing Initiatives
Data Citations
Develop New Recommendation
COUNTER
CONSISTENT CREDIBLE COMPARABLE
Repository Usage Metrics
Server Log Processing
Future Metrics
Develop New Data Level Metrics Hub
Display Data Metrics
Drive Adoption - Engage Communities

[H. Cousijn et al., 2019]