Advanced Data Management (CSCI 490/680)

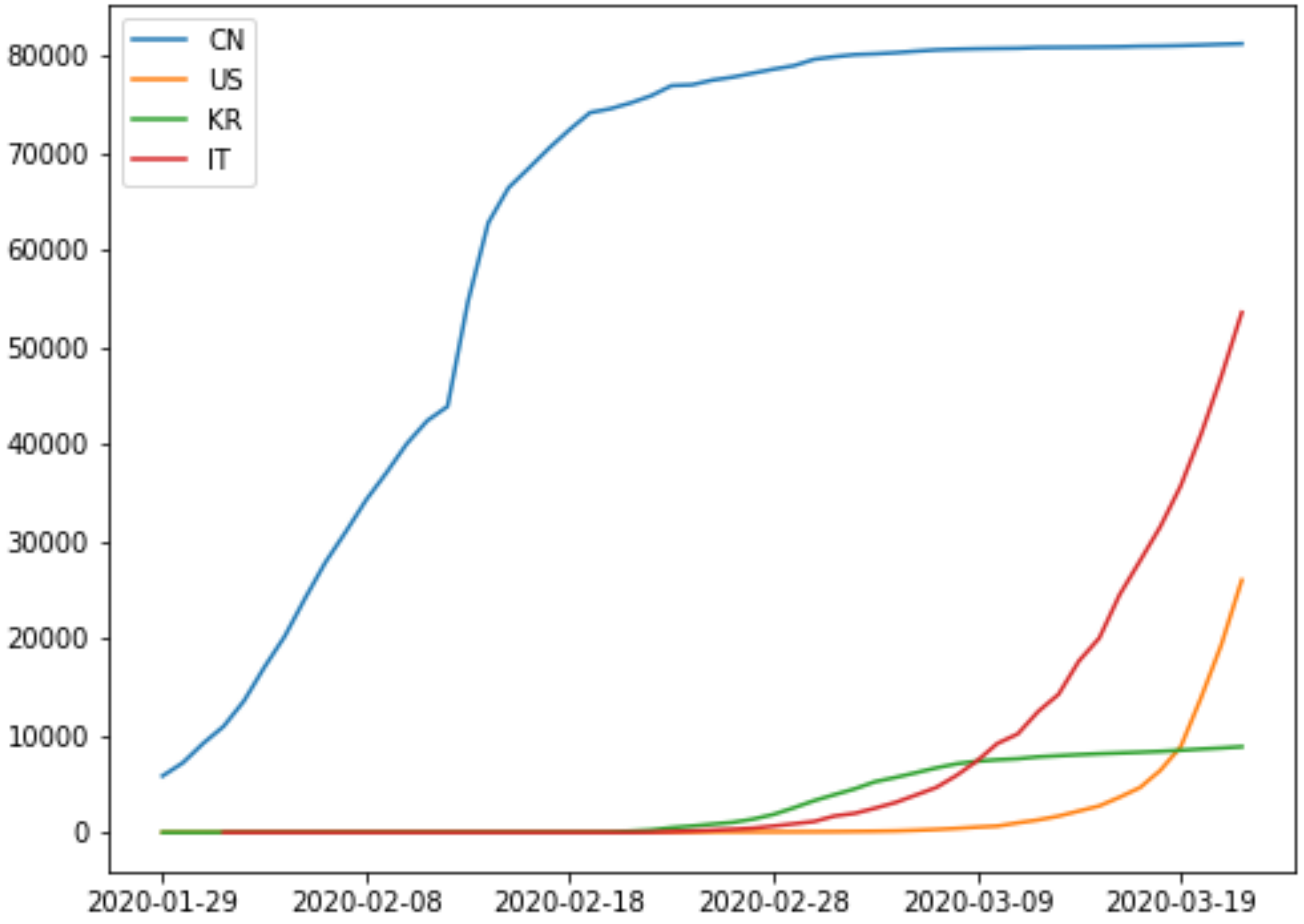
Data Curation

Dr. David Koop

Discussions

- Please post at least once on the discussion board in Blackboard with a question, answer, or discussion point about one of the lectures
- Feedback is useful—I hope the lectures are clear, but I am pretty sure there are still places where I can clarify things better
- You may also post questions about the assignment there if you believe they are relevant to all students

Assignment 4



- COVID-19 data
- Data Integration
  - Population
  - Temperature
- Data Fusion:
  - Our World in Data
  - Johns Hopkins
  - Wikipedia
- Questions?

Test 2

- Online on Blackboard (webcourses.niu.edu)
- Thursday, April 9 from 3:30-4:45pm
- If you have conflicts, **let me know as soon as possible**
- Format:
  - Some multiple choice
  - More short answer/free response
- Focus on topics since the first test
- More details this week

Recent History in Databases

- Early 2000s: Commercial DBs dominated, Open-source DBs missing features
- Mid 2000s: MySQL adopted by web companies
- Late 2000s: NoSQL dos scale horizontally out of the box
- Early 2010s: New DBMSs that can scale across multiple machines natively and provide ACID guarantees

[A. Pavlo]

NewSQL

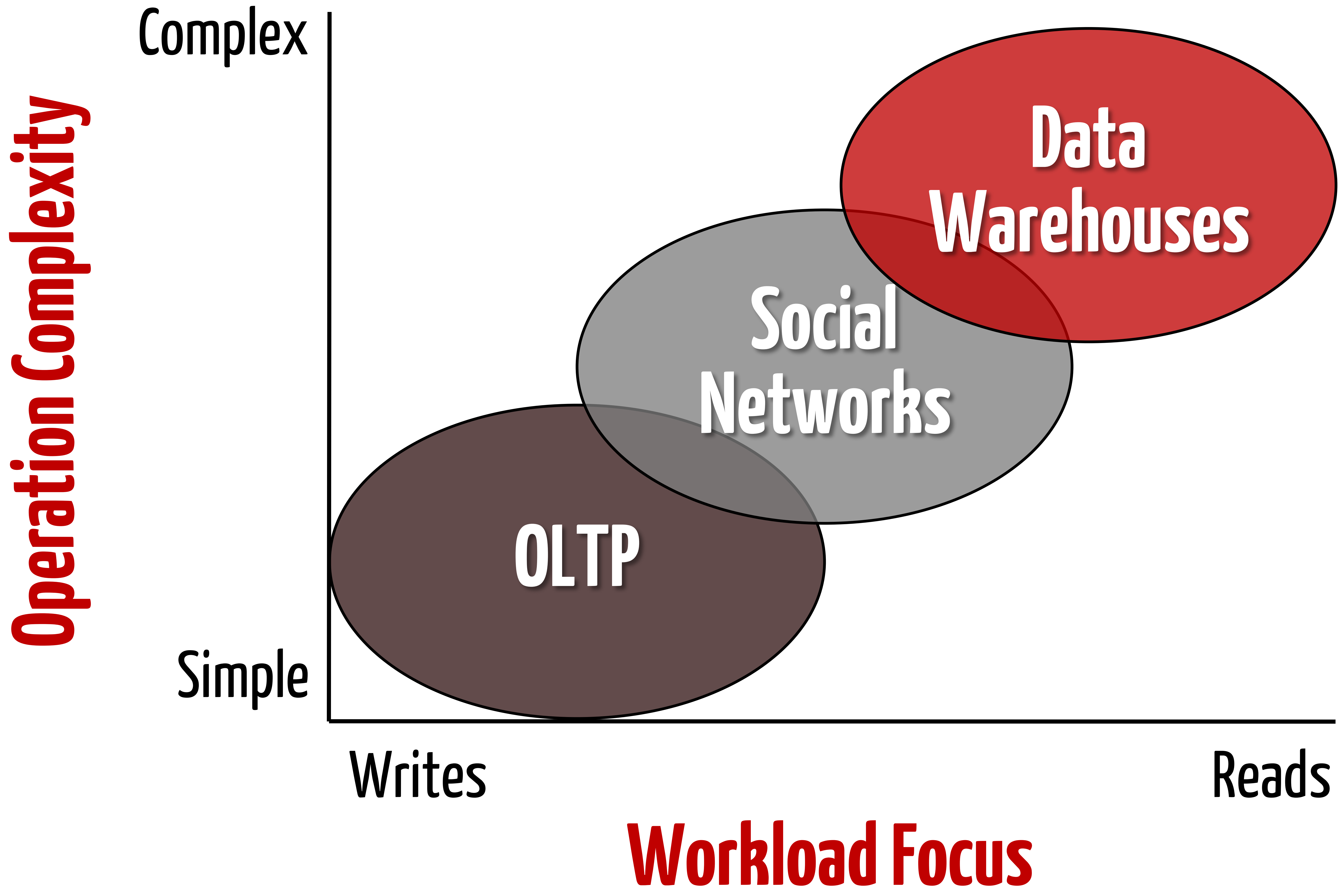
[A. Pavlo]

NewSQL

- 451 Group's Definition:
  - A DBMS that delivers the scalability and flexibility promised by NoSQL while retaining the support for SQL queries and/or ACID, or to improve performance for appropriate workloads.
- Stonebraker's Definition:
  - SQL as the primary interface
  - ACID support for transactions
  - Non-locking concurrency control
  - High per-node performance
  - Parallel, shared-nothing architecture

[A. Pavlo]

OLTP Workload



[A. Pavlo]

Ideal OLTP System

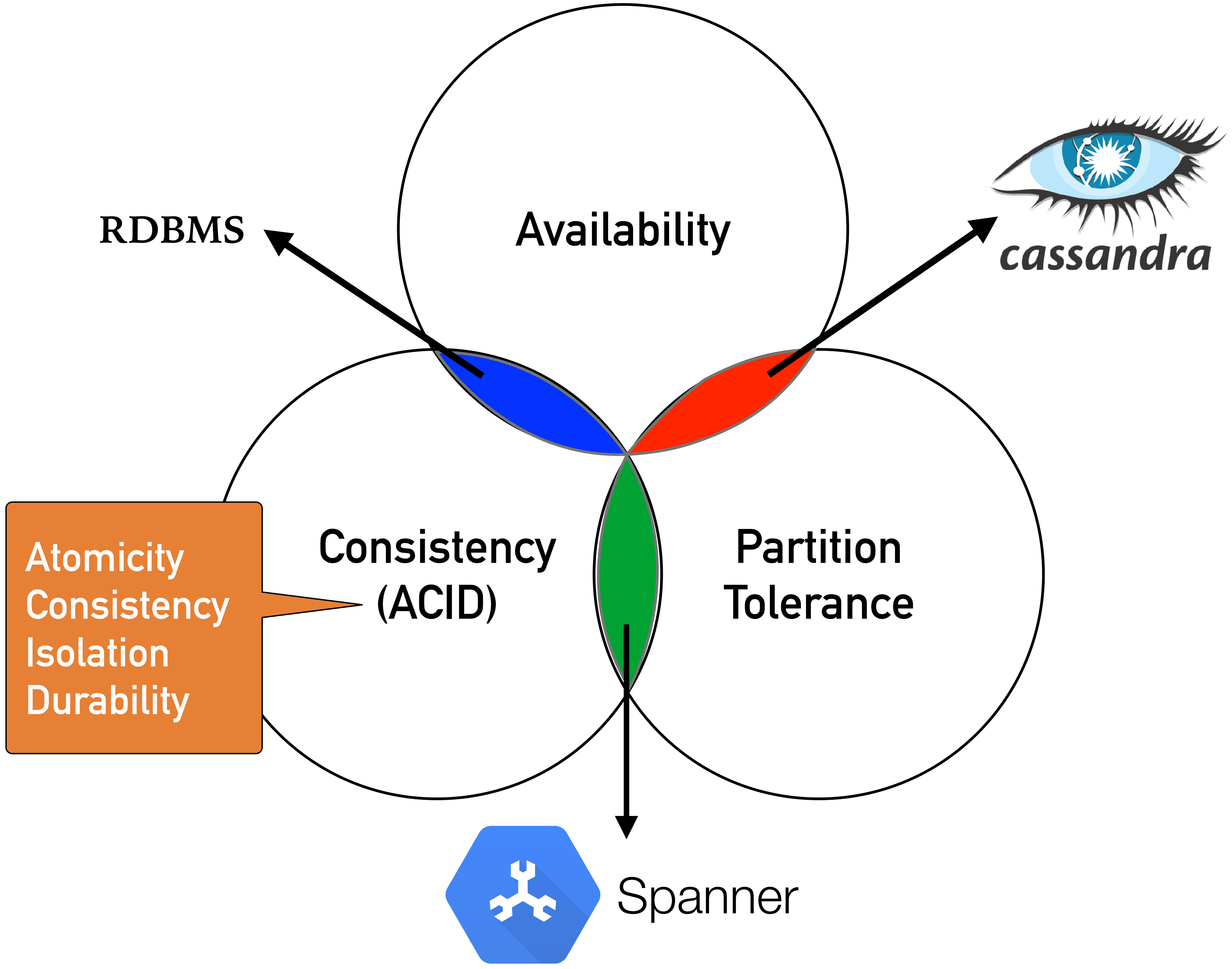
- Main Memory Only
- No Multi-processor Overhead
- High Scalability
- High Availability
- Autonomic Configuration

[A. Pavlo]

Spanner Overview

- Focus on scaling databases focused on OLTP (not OLAP)
- Since OLTP, focus is on sharding **rows**
- Tries to satisfy CAP (which is impossible per CAP Theorem) by not worrying about 100% availability
- External consistency using multi-version concurrency control through timestamps
- ACID is important
- Structured: universe with zones with zone masters and then spans with span masters
- SQL-like (updates allow SQL to be used with Spanner)

Spanner and the CAP Theorem



- Which type of system is Spanner?
  - C: consistency, which implies a single value for shared data
  - A: 100% availability, for both reads and updates
  - P: tolerance to network partitions
- Which two?
  - CA: close, but not totally available
  - So actually **CP**

External Consistency

- Traditional DB solution: **two-phase locking**—no writes while client reads
- "The system behaves as if all transactions were executed sequentially, even though Spanner actually runs them across multiple servers (and possibly in multiple datacenters) for higher performance and availability" [Google]
- Semantically indistinguishable from a single-machine database
- Uses multi-version concurrency control (MVCC) using **timestamps**
- Spanner uses **TrueTime** to generate monotonically increasing timestamps across all nodes of the system

Google Cloud Spanner: NewSQL

	CLOUD SPANNER	TRADITIONAL RELATIONAL	TRADITIONAL NON-RELATIONAL
Schema	✓ **Yes**	✓ Yes	✗ No
SQL	✓ **Yes**	✓ Yes	✗ No
Consistency	✓ **Strong**	✓ Strong	✗ Eventual
Availability	✓ **High**	✗ Failover	✓ High
Scalability	✓ **Horizontal**	✗ Vertical	✓ Horizontal
Replication	✓ **Automatic**	Configurable	Configurable

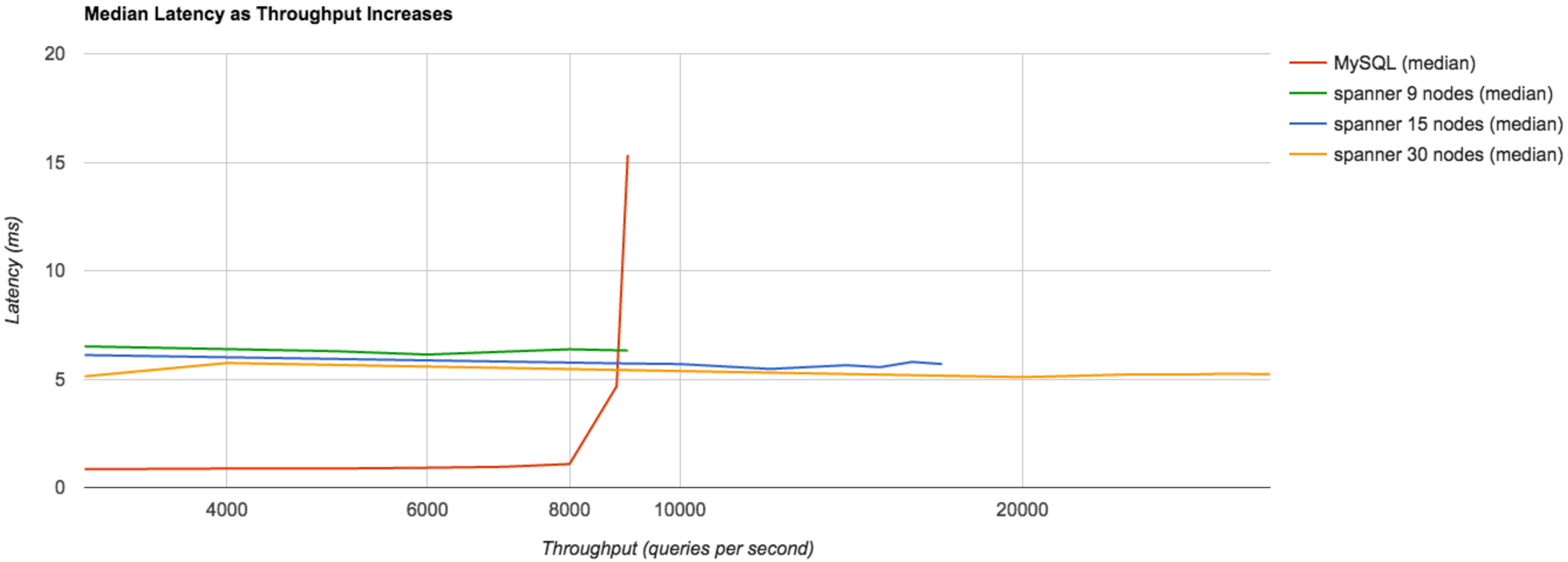
[https://cloud.google.com/spanner/]

Spanner as "Effectively CA"

- Criteria for being "effectively CA"
  1. At a minimum it must have very high availability in practice (so that users can ignore exceptions), and
  2. As this is about partitions it should also have a low fraction of those outages due to partitions.
- Spanner meets both of these criteria
- Spanner relies on Google's **network** (private links between data centers)
- TrueTime helps create **consistent snapshots**, sometimes have a commit wait

[E. Brewer, 2017]

Throughput: Spanner vs. MySQL

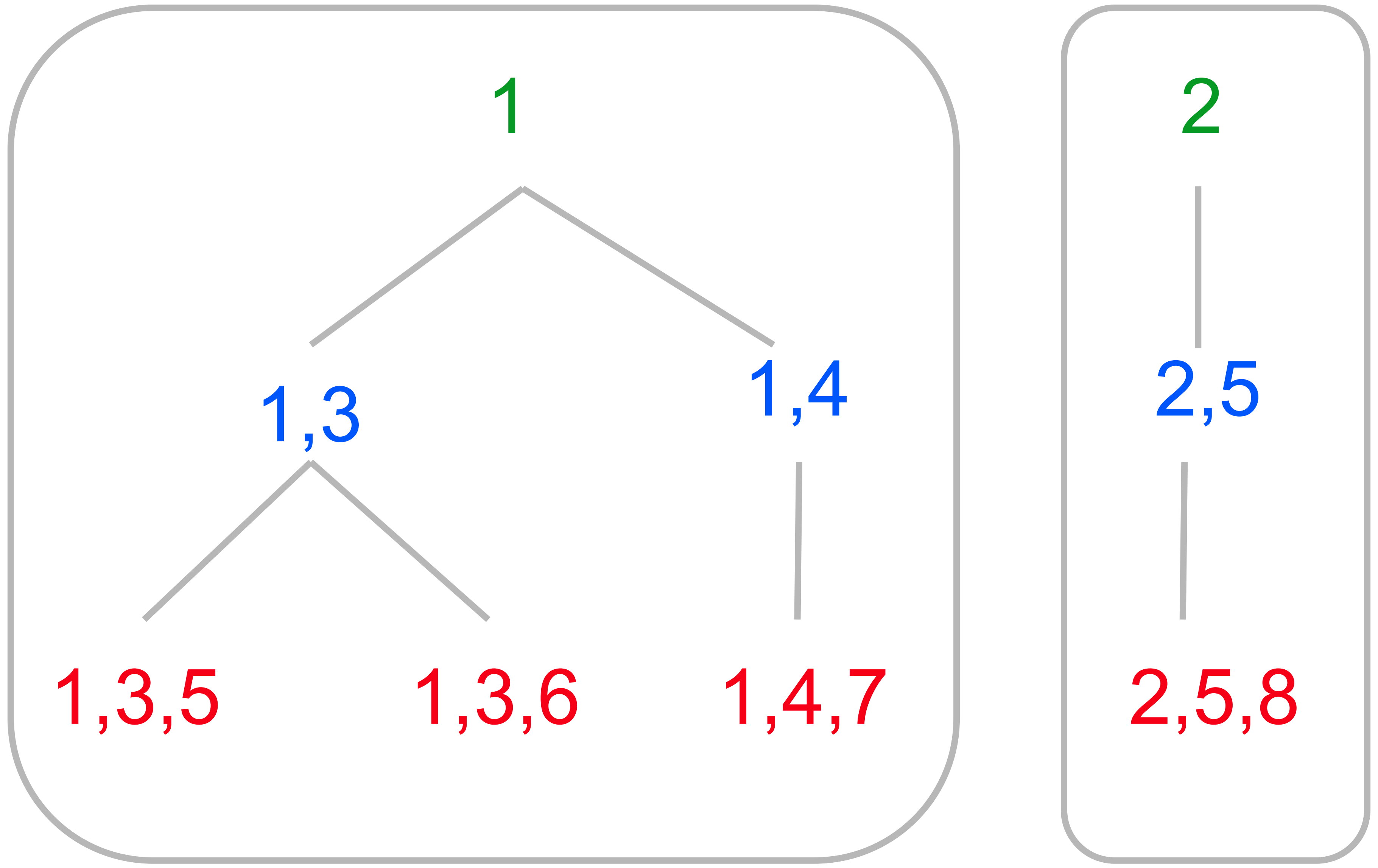


[P. Bakkum and D. Cepeda, 2017]

F1's Hierarchical Schema and Clustered Storage

- Child rows under one root row form a **cluster**
- Cluster stored on one machine (unless huge)
- Transactions within one cluster are most efficient
- Very efficient joins inside clusters (can merge with no sorting)

Rows and PKs



Storage Layout

```
Customer (1)
Campaign (1,3)
AdGroup  (1,3,5)
AdGroup  (1,3,6)
Campaign (1,4)
AdGroup  (1,4,7)
```

```
Customer (2)
Campaign (2,5)
AdGroup  (2,5,8)
```

[Shute et al., 2012]

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

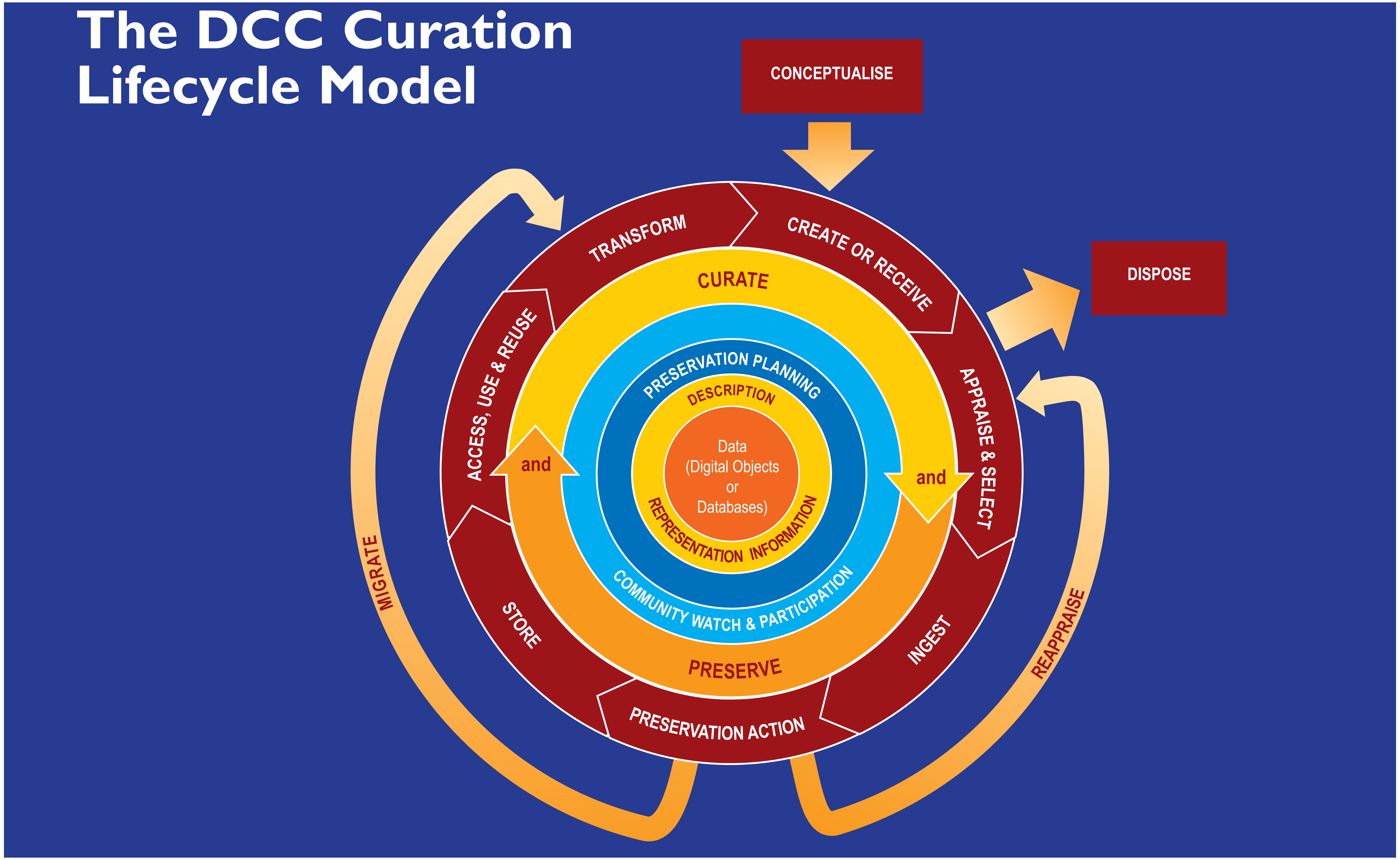
Data Identity

- 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]

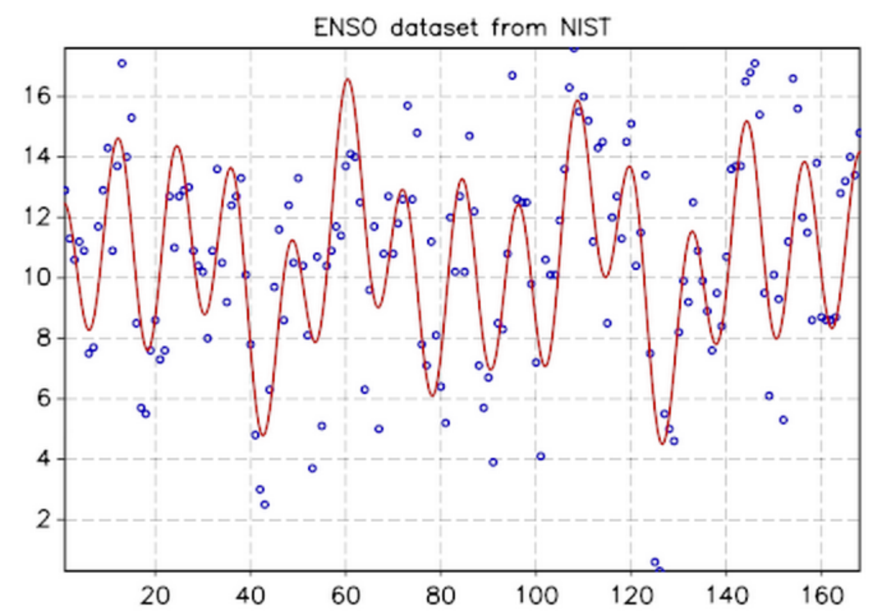
To be Findable

- 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

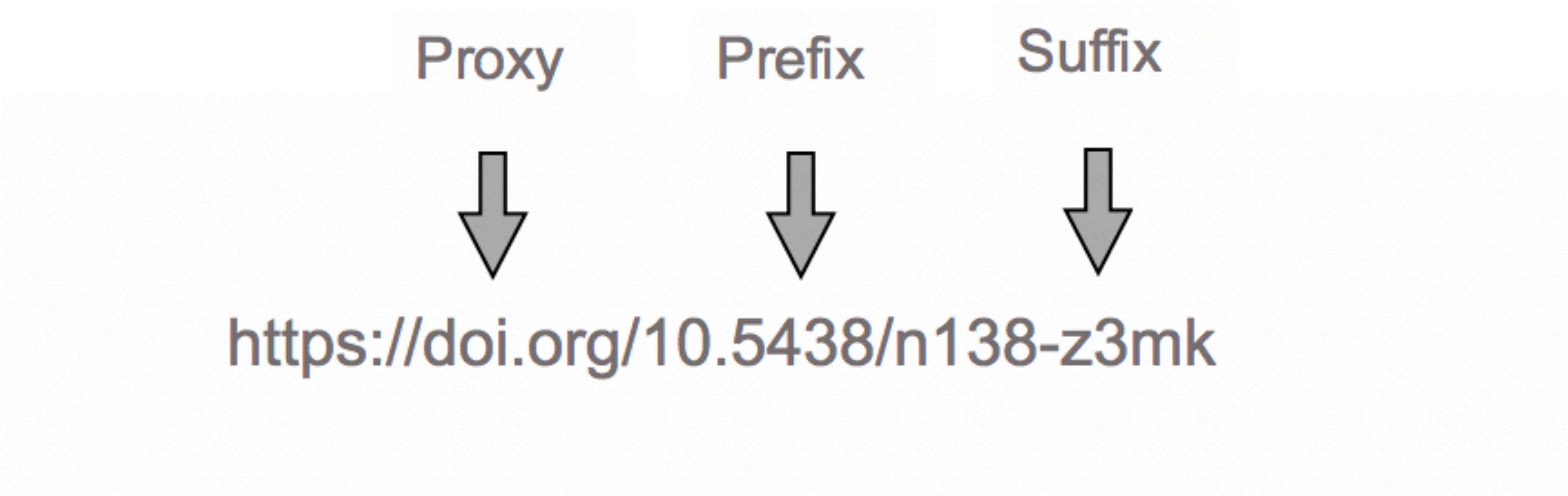
5. Enjoy the benefits

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

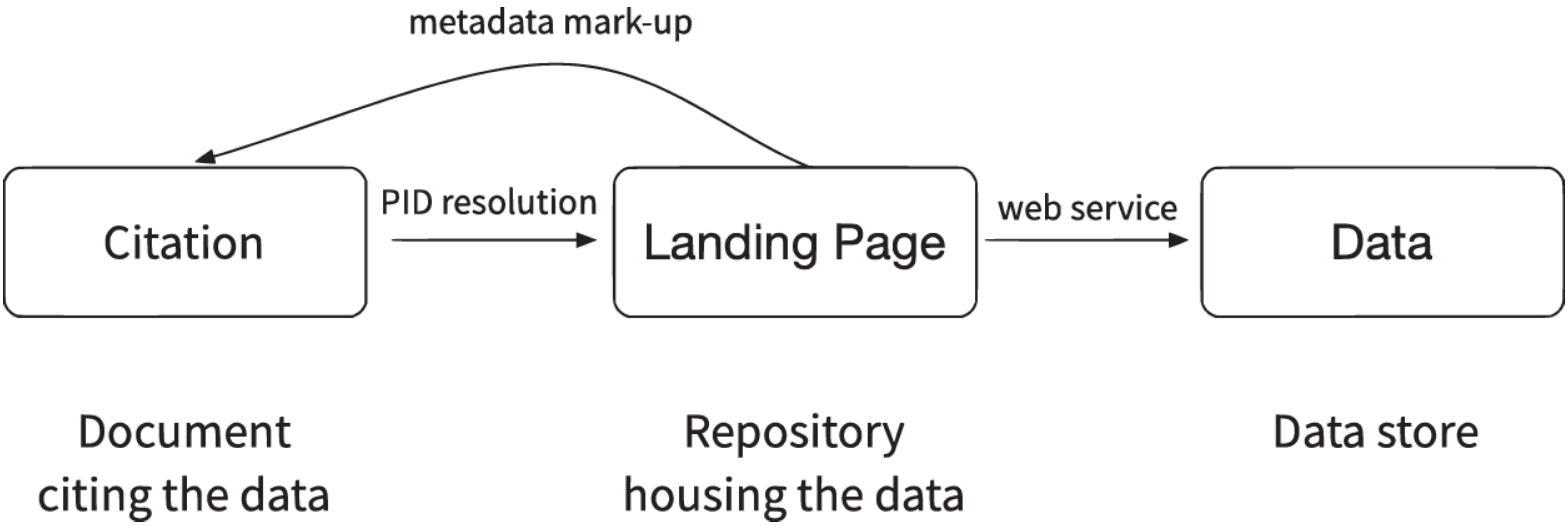
[DataCite]

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



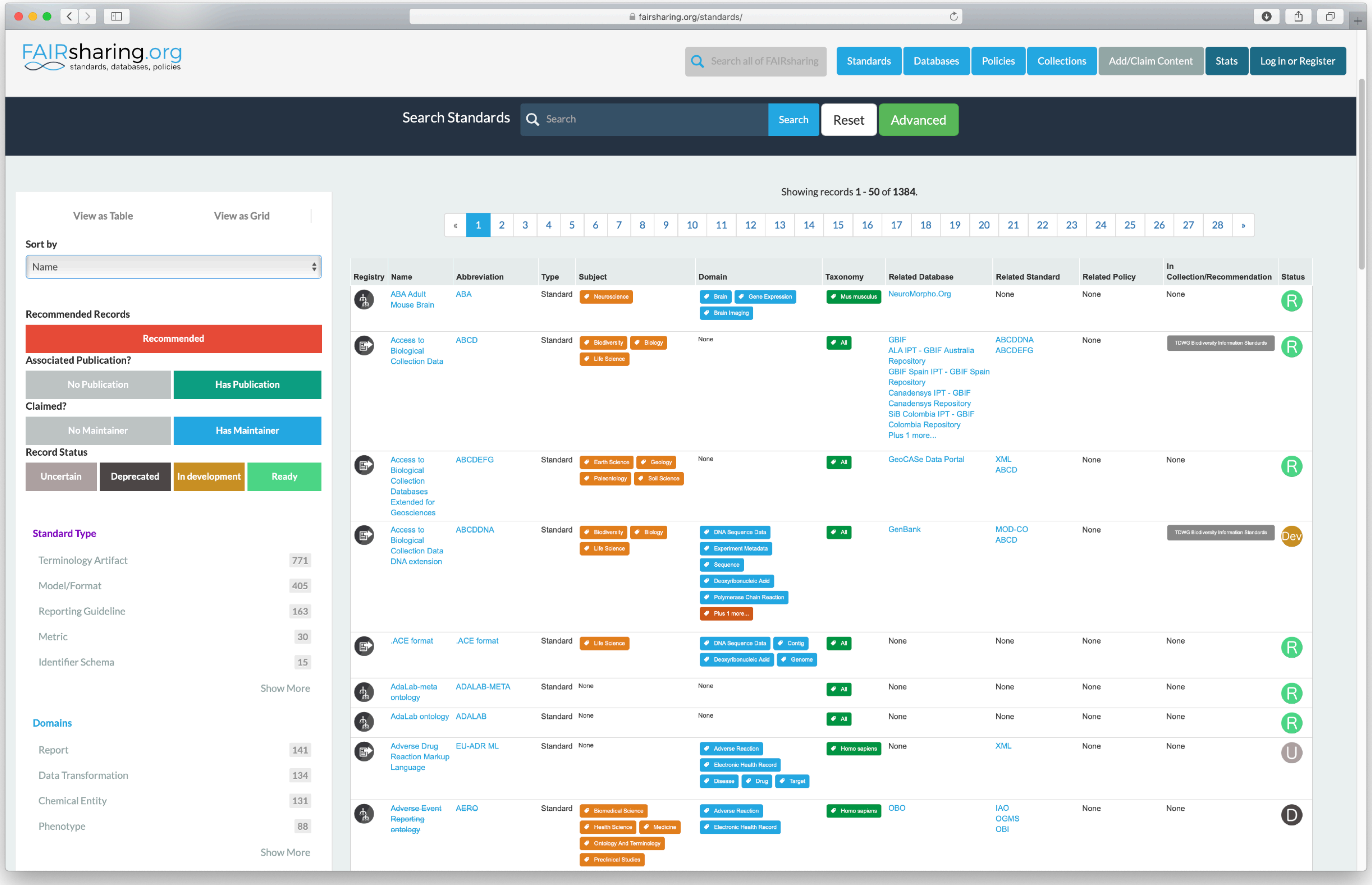
[M. Fenner et al., 2019]

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



[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**

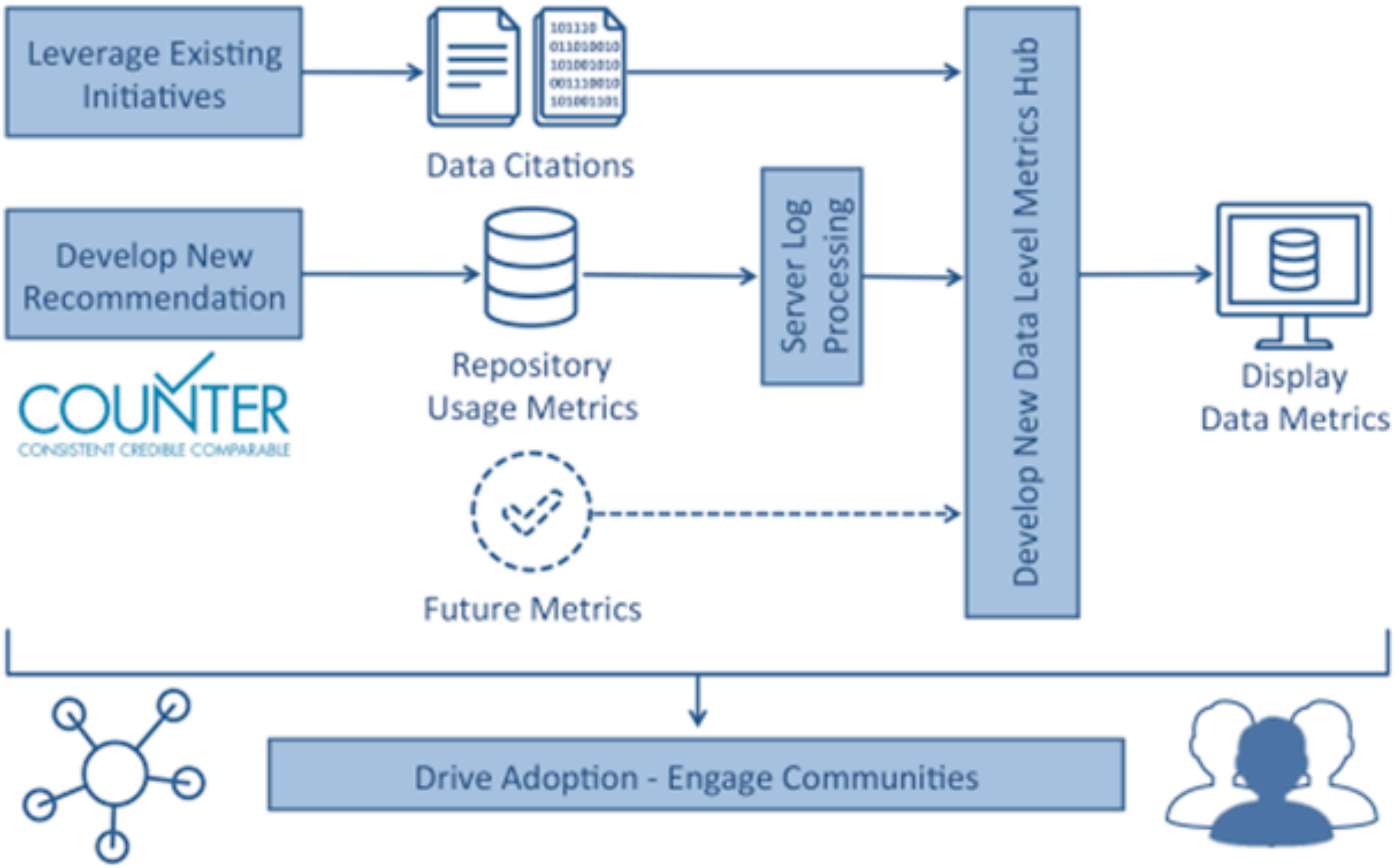
[M. D. Wilkinson et al., 2016]

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



[H. Cousijn et al., 2019]

Reminder: Discussions

- Please post at least once on the discussion board in Blackboard with a question, answer, or discussion point about one of the lectures
- Feedback is useful—I hope the lectures are clear, but I am pretty sure there are still places where I can clarify things better
- You may also post questions about the assignment there if you believe they are relevant to all students