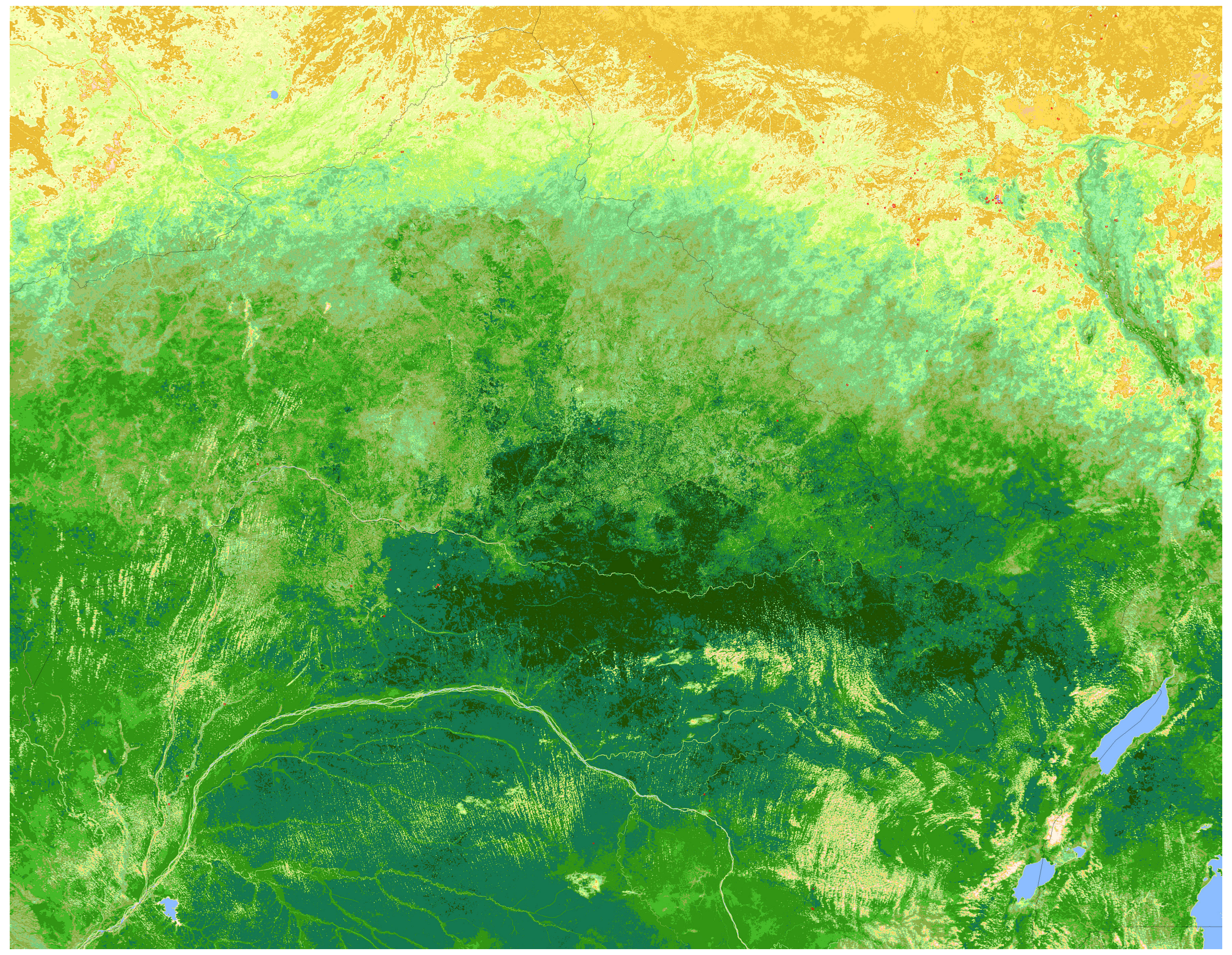
Advanced Data Management (CSCI 640/490)

Data Curation

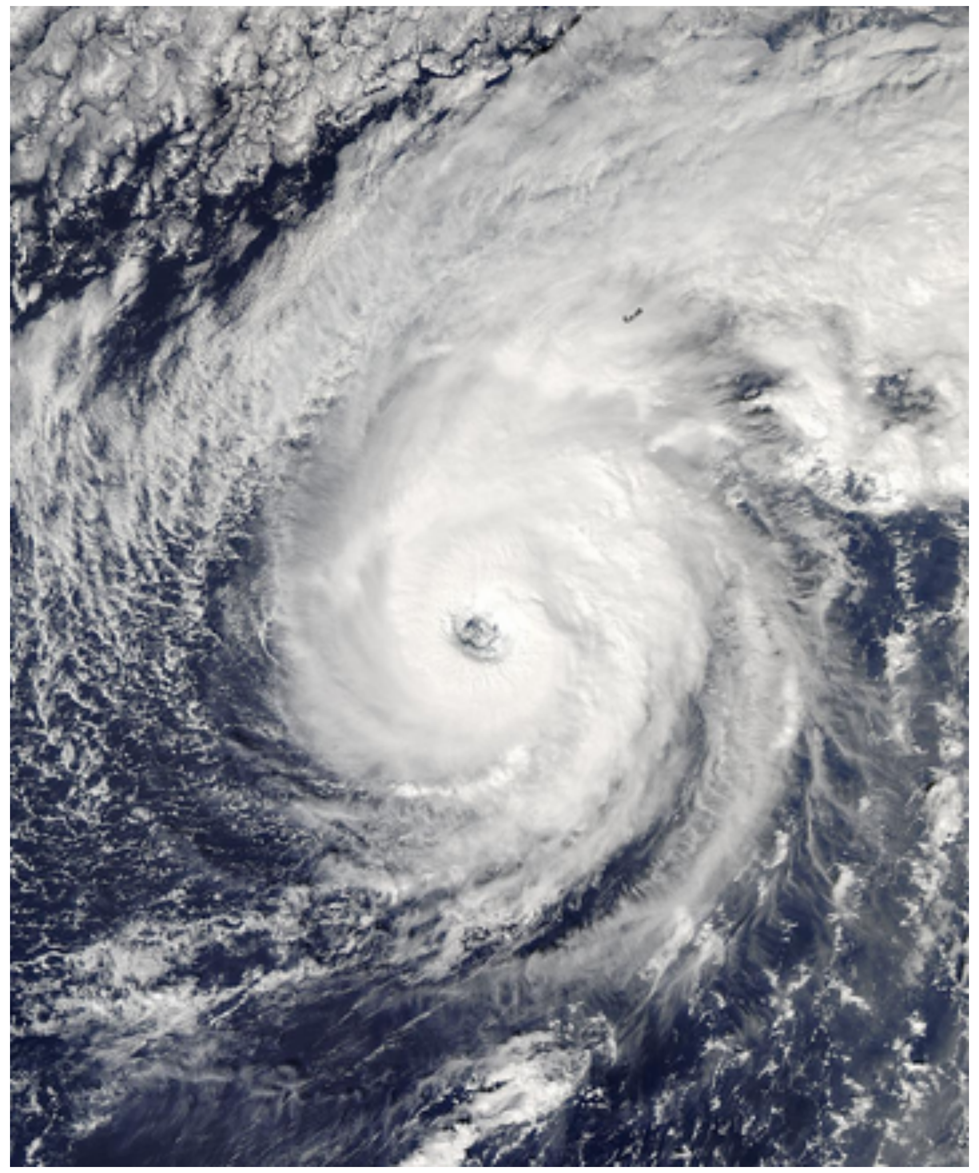
Dr. David Koop

Spatial Data

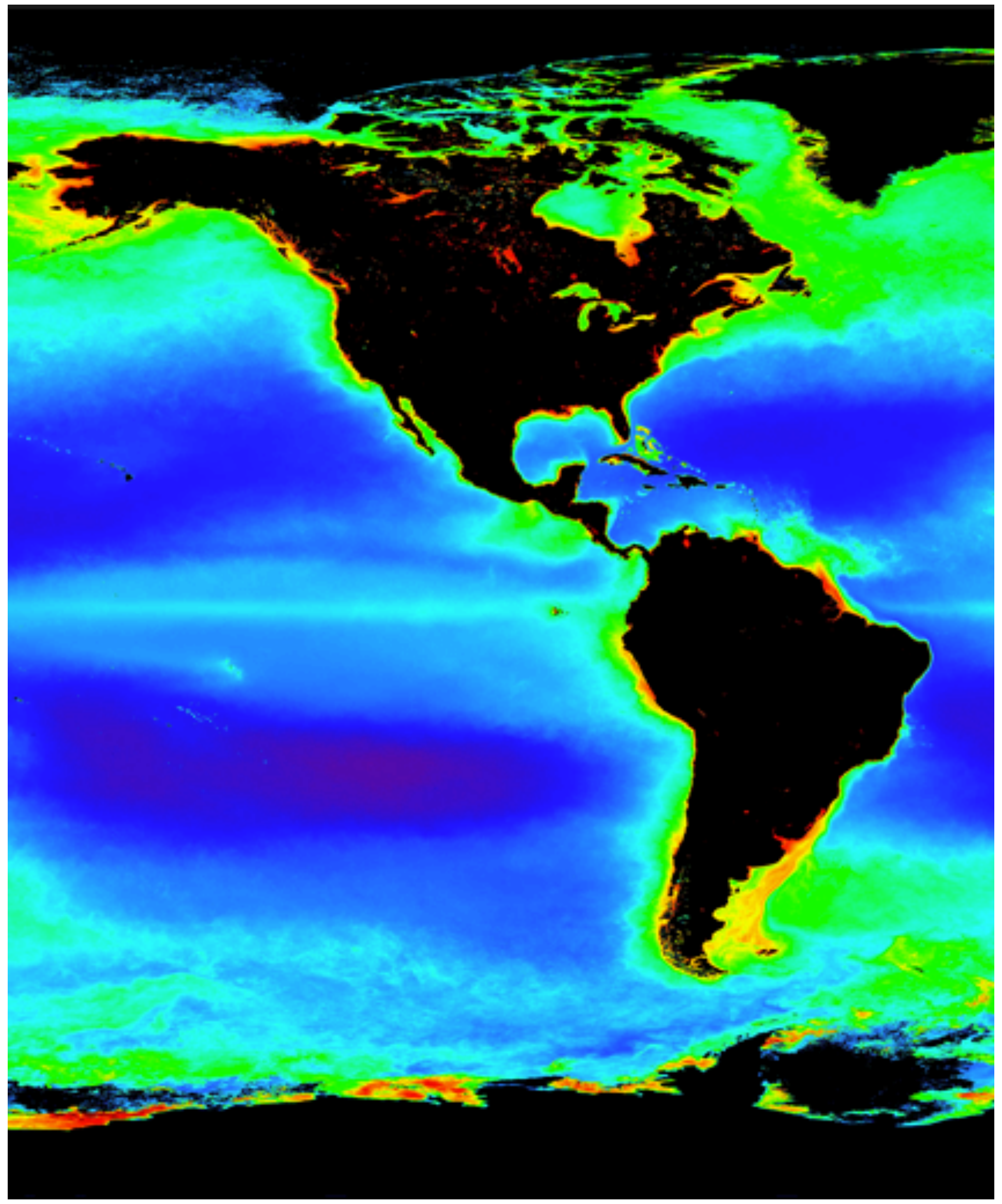
Measure vegetation density



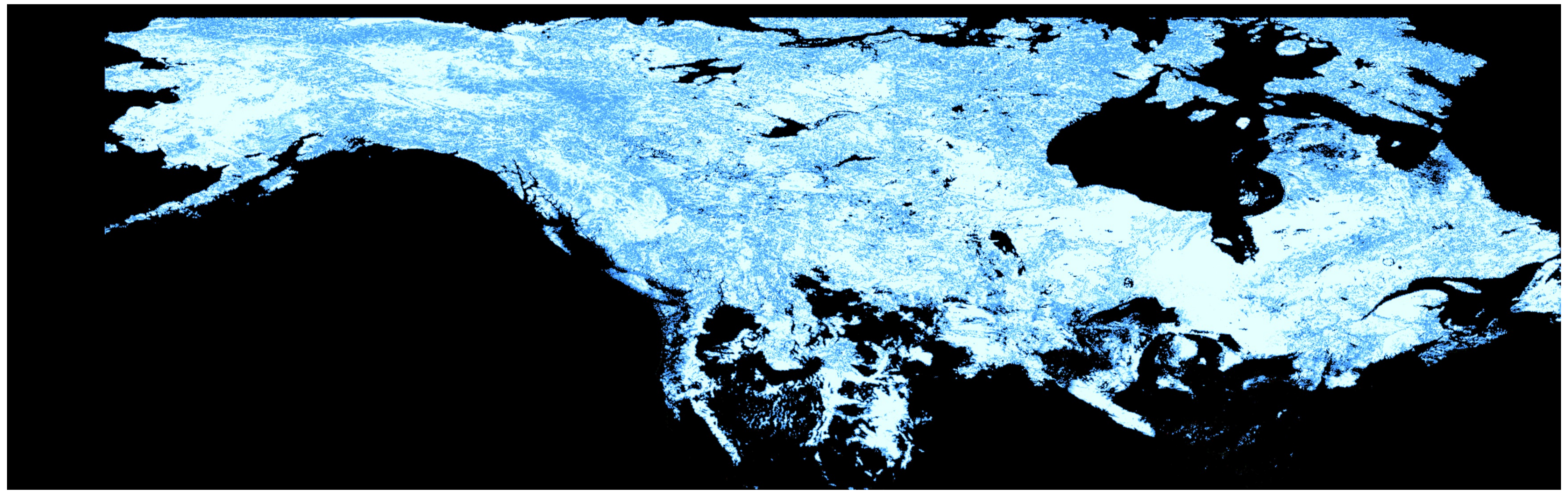
Track hurricanes



Track phytoplankton populations



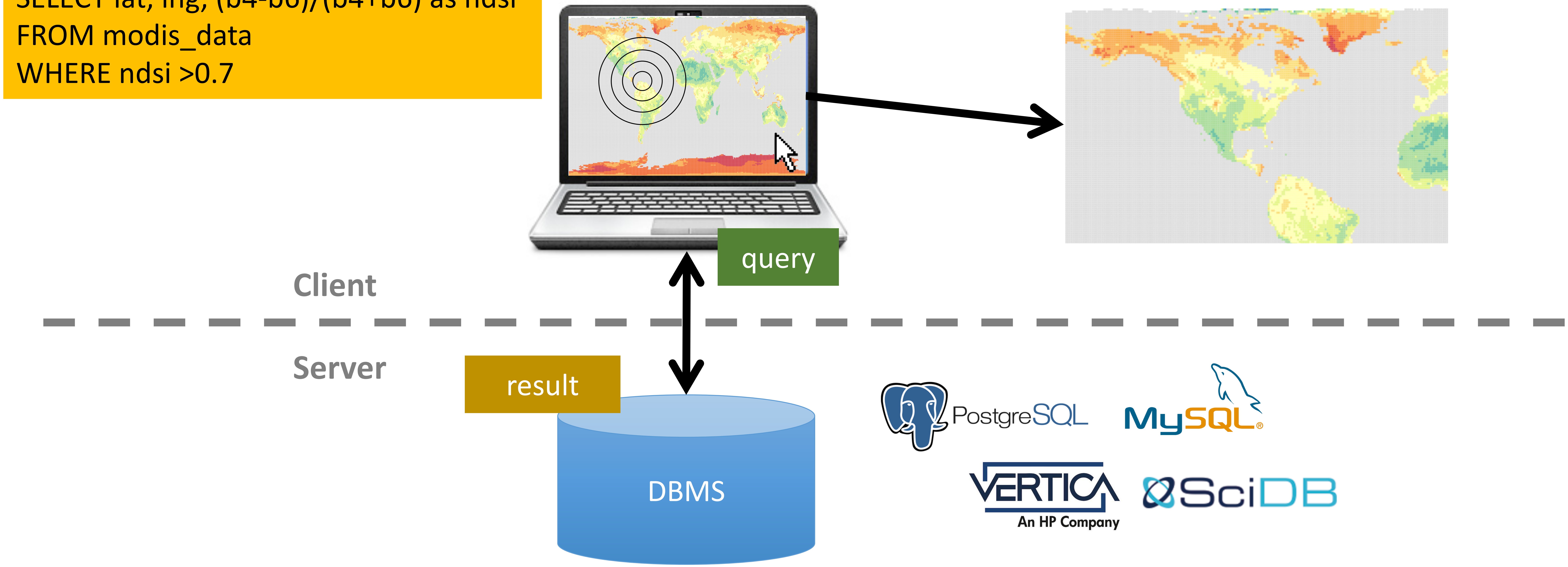
Measure snow melt



[L. Battle, 2017]

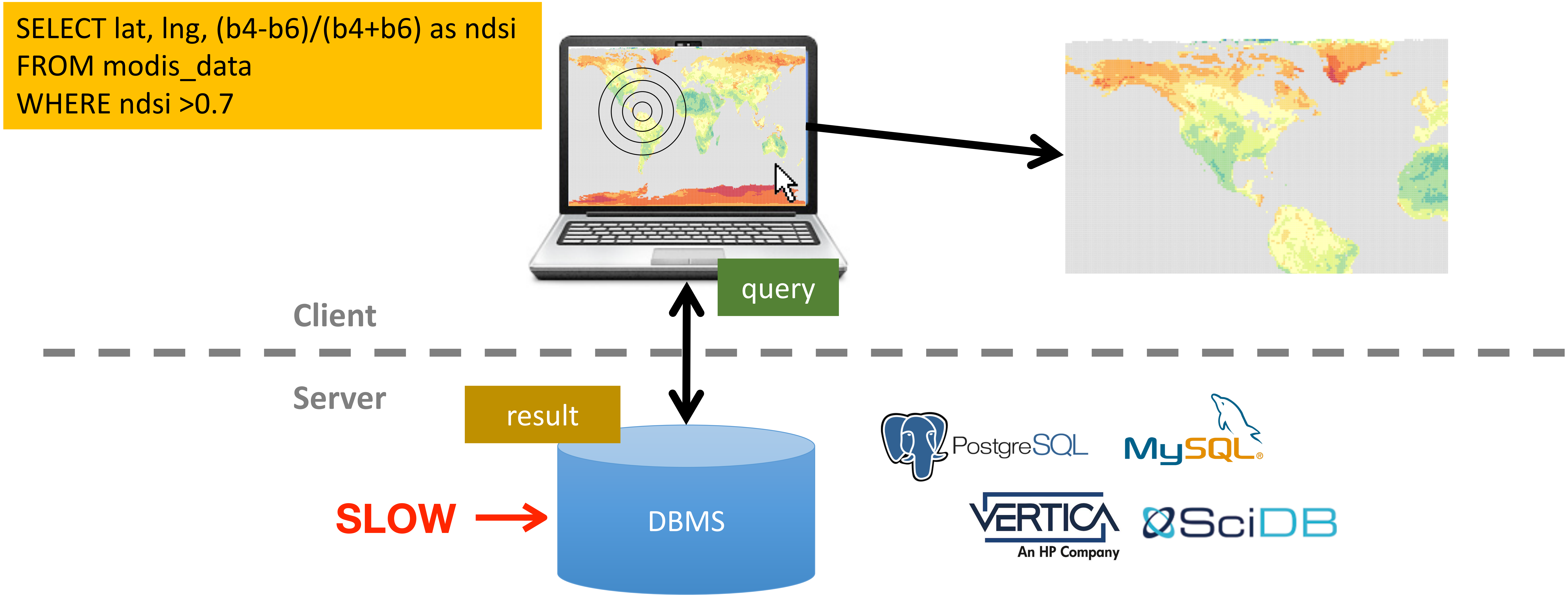
Interactive Exploration of Spatial Data

```
SELECT lat, lng, (b4-b6)/(b4+b6) as ndsi
FROM modis_data
WHERE ndsi >0.7
```



[L. Battle, 2017]

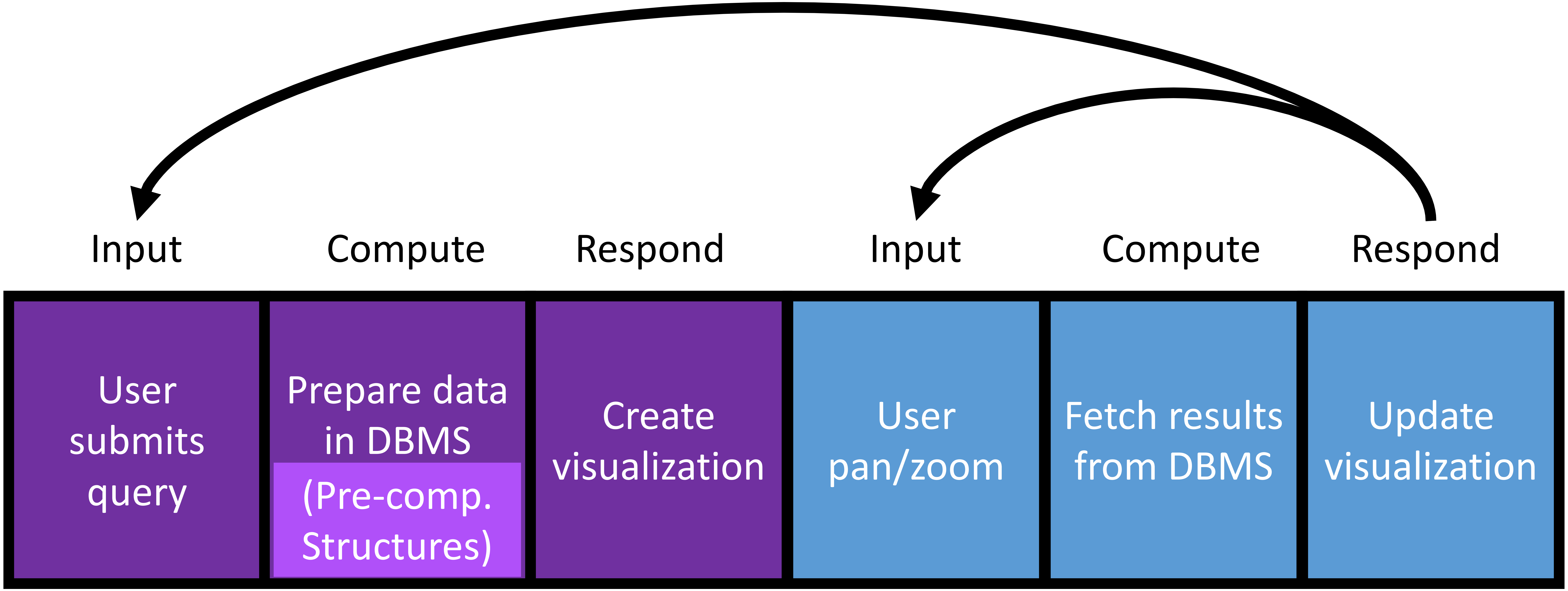
Interactive Exploration of Spatial Data



```
SELECT lat, lng, (b4-b6)/(b4+b6) as ndsi
FROM modis_data
WHERE ndsi >0.7
```

[L. Battle, 2017]

Two Inputs to Exploratory Browsing



Input

Compute

Respond

Input

Compute

Respond

User submits query

Prepare data in DBMS

(Pre-comp. Structures)

Create visualization

User pan/zoom

Fetch results from DBMS

Update visualization

Cold start time

interaction latency < 500ms

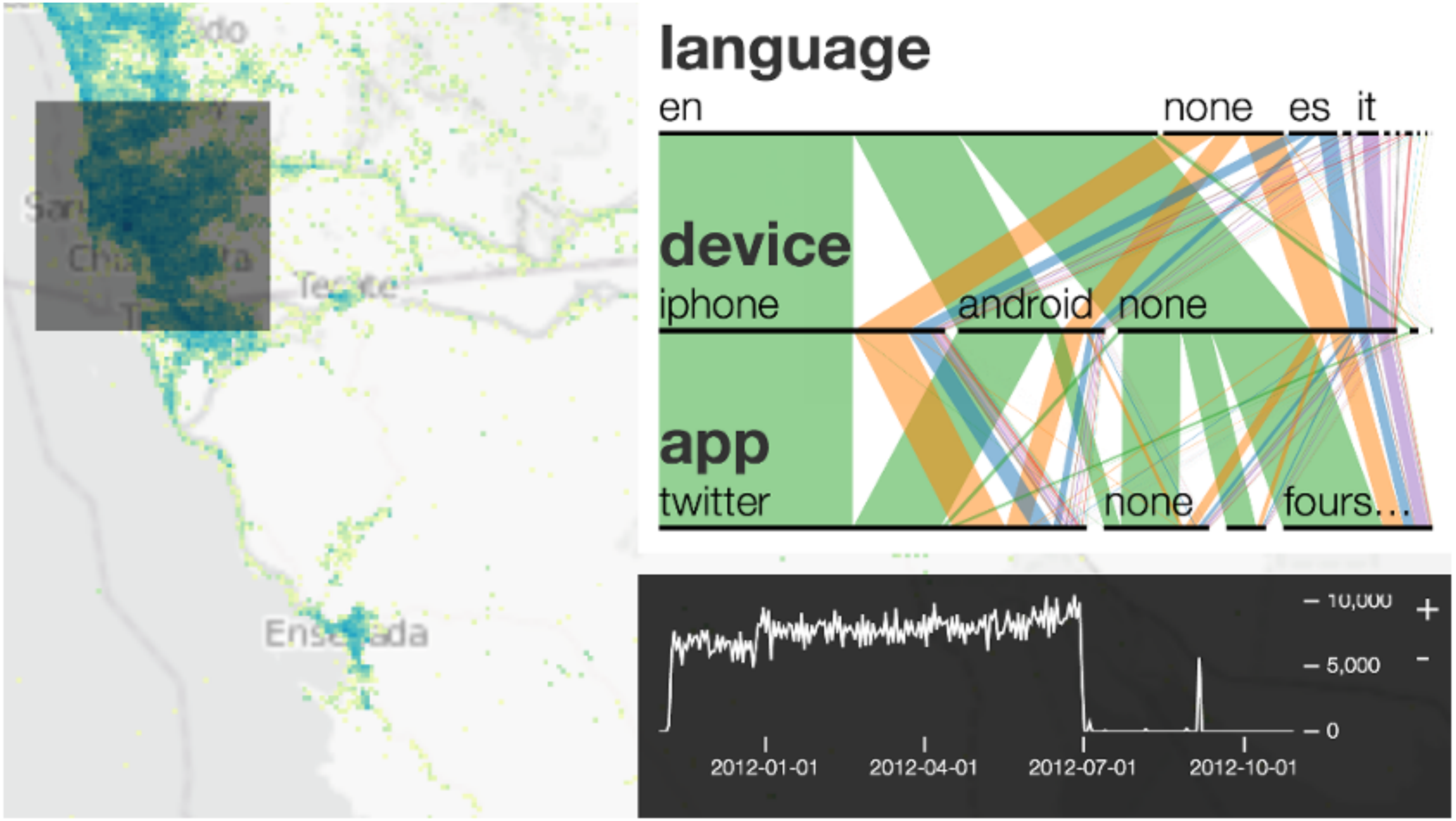
[L. Battle, 2017]

Systems for Interactive Exploration

<table>
<tr><td></td><td></td><th colspan="3">Time</th></tr>
<tr><td></td><td></td><th>(Offline)
Pre-computed structures</th><th>(Before interaction)
Predictive</th><th>(After interaction)
Progressive/Incremental</th></tr>
<tr><th rowspan="4">Output format</th><th rowspan="2">Sampling</th><td></td><td></td><td>SampleAction (CHI 2012)
Vizdom (VLDB 2015)</td></tr>
<tr><td></td><td colspan="2">DICE (ICDE 2014)
A-WARE (HILDA 2016)</td></tr>
<tr><th rowspan="2">Aggregation</th><td>Nanocubes (Infovis 2013)
imMens (Eurovis 2013)</td><td>ATLAS (VAST 2008)
XmdvTool (*DASFAA* 2003)</td><td></td></tr>
<tr><td colspan="2">ForeCache</td><td></td></tr>
</table>

[L. Battle, 2017]

Nanocubes



Linked view of tweets in San Diego, US

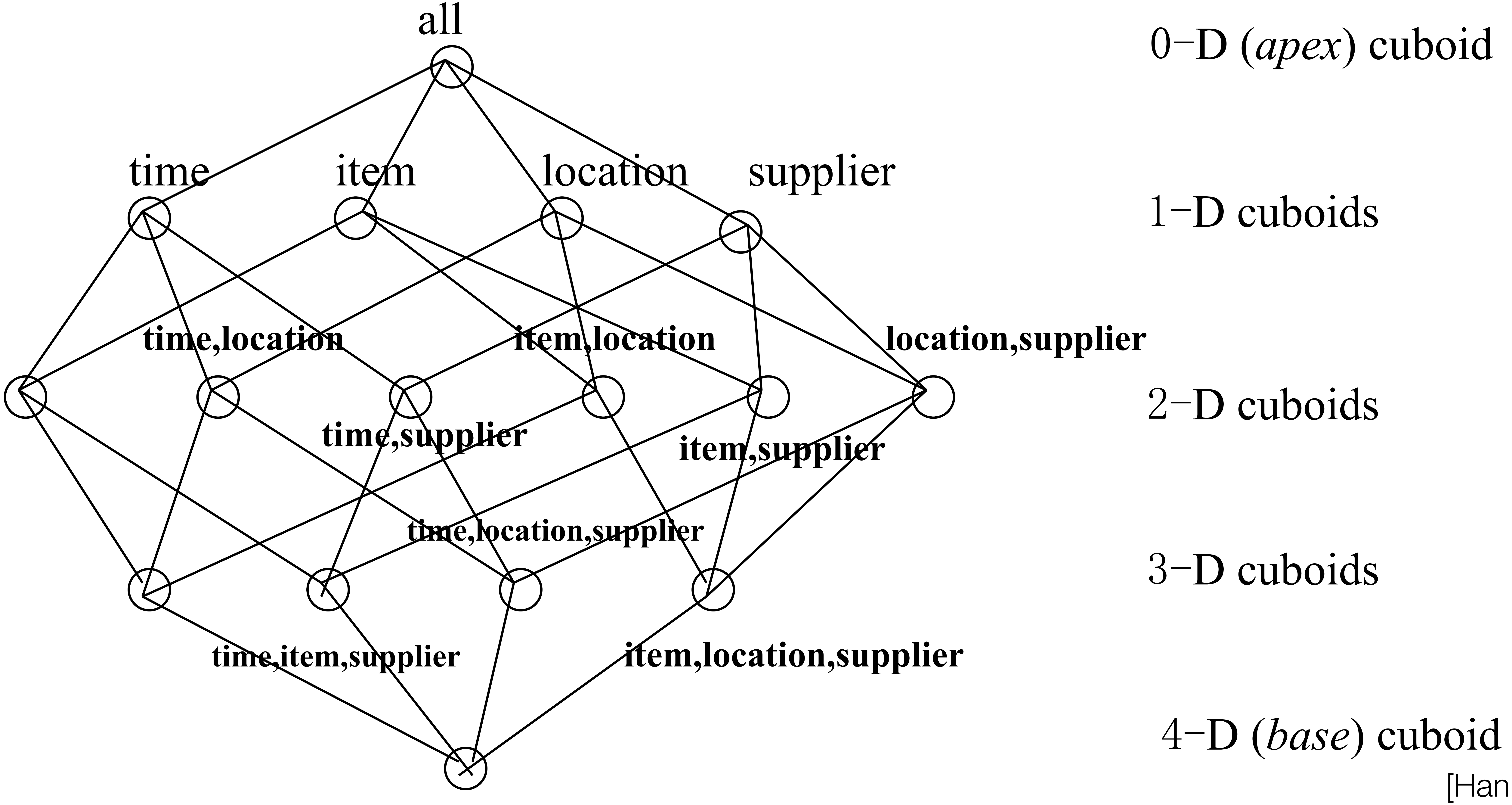
[Lins et. al, 2013]

From Tables and Spreadsheets to Data Cubes

- A **data warehouse** is based on a multidimensional data model which views data in the form of a data cube
- A **data cube**, such as sales, allows data to be modeled and viewed in multiple dimensions
  - **Dimension tables**, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - **Fact table** contains **measures** (such as dollars_sold) and keys to each of the related dimension tables
- In data warehousing literature, an n-D base cube is called a **base cuboid**. The top most 0-D cuboid, which holds the highest-level of summarization, is called the **apex cuboid**. The lattice of cuboids forms a **data cube**.

[Han et al., 2011]

Data Cube: A Lattice of Cuboids



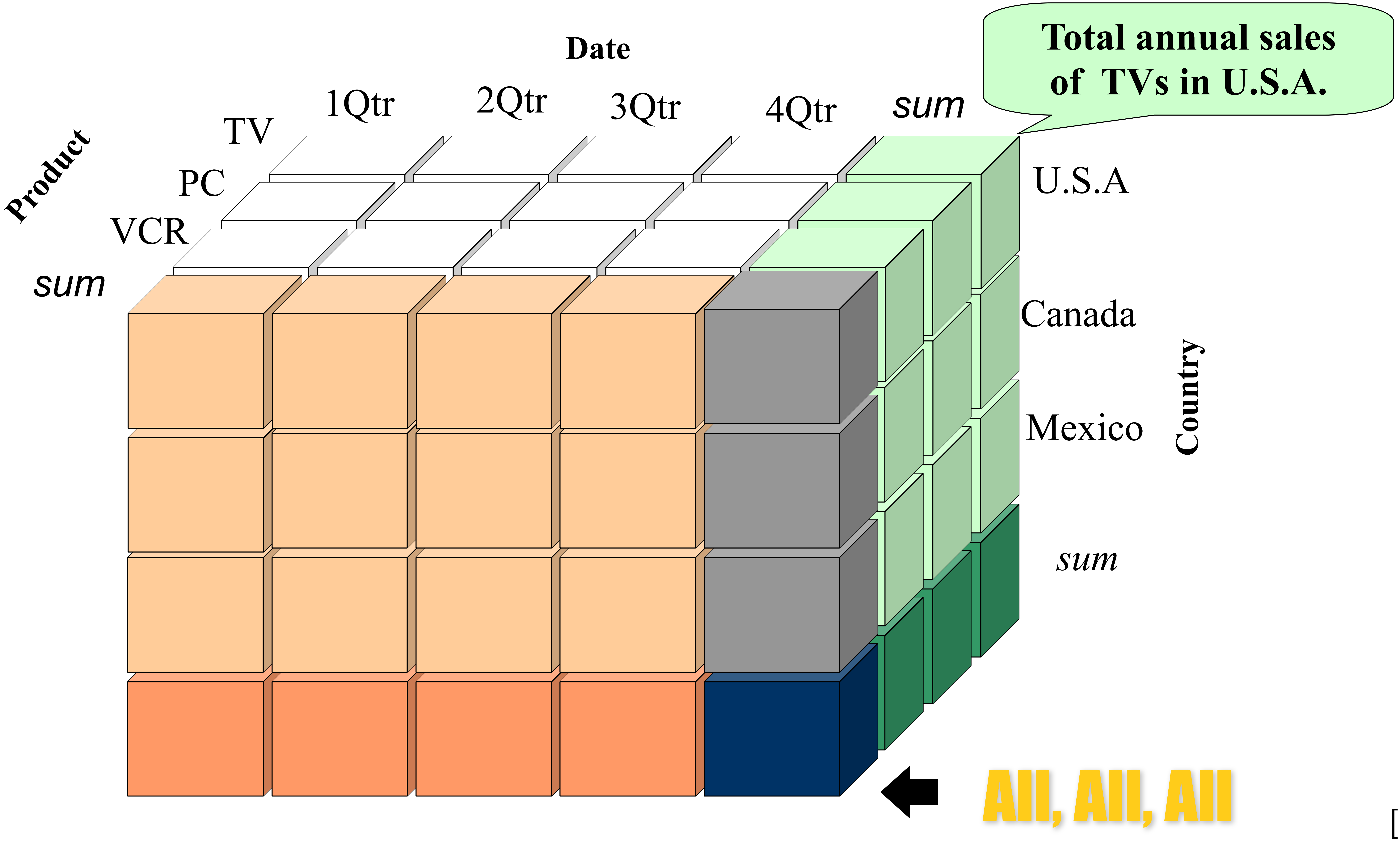
[Han et al., 2011]

Data Cube Measures: Three Categories

- **Distributive**: if the result derived by applying the function to n aggregate values is the same as that derived by applying the function on all the data without partitioning
  - E.g., `count(), sum(), min(), max()`
- **Algebraic**: if it can be computed by an algebraic function with M arguments (where M is a bounded integer), each of which is obtained by applying a distributive aggregate function
  - E.g., `avg(), min_N(), standard_deviation()`
- **Holistic**: if there is no constant bound on the storage size needed to describe a subaggregate.
  - E.g., `median(), mode(), rank()`

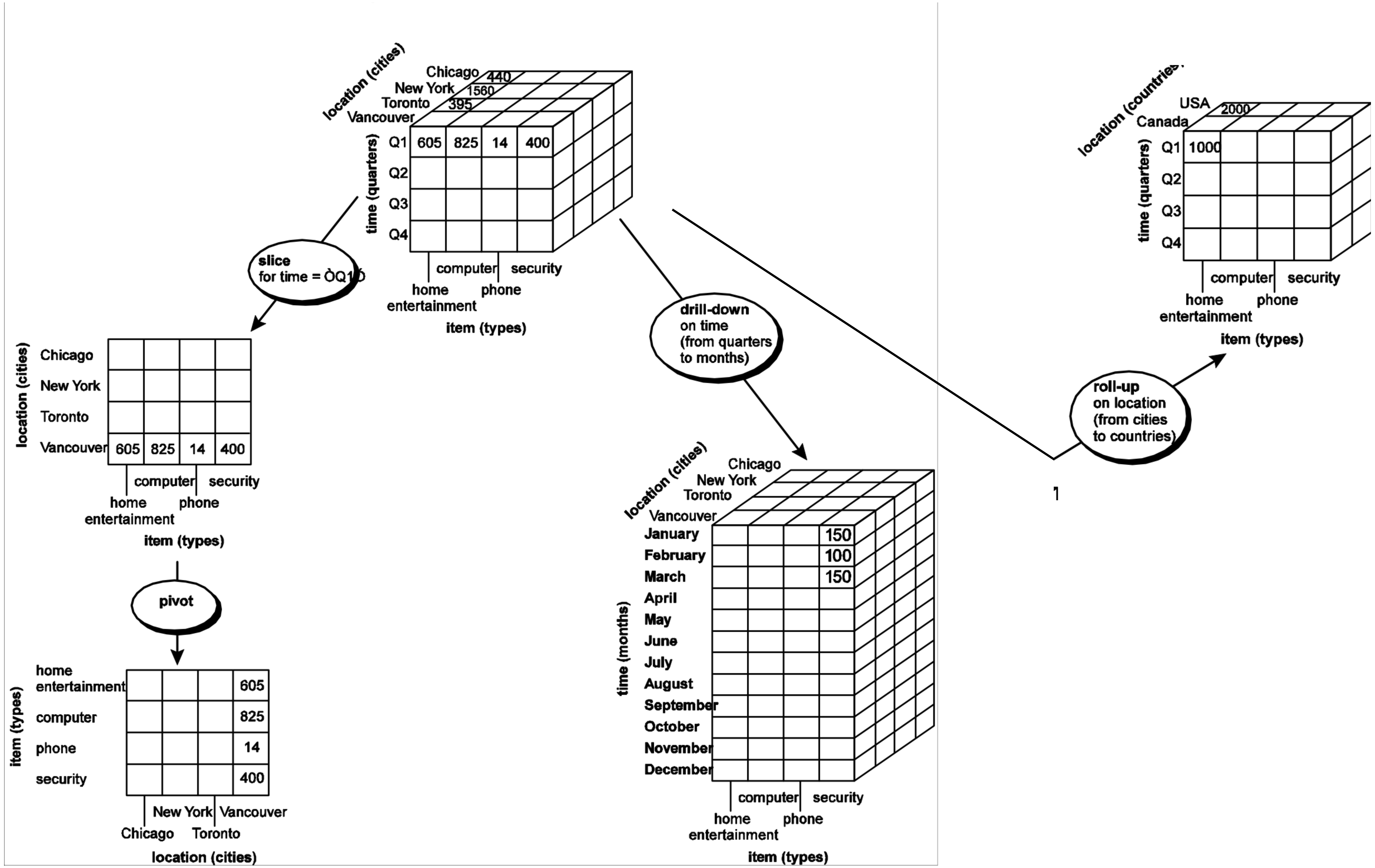
[Han et al., 2011]

A Sample Data Cube



[Han et al., 2011]

OLAP Operations



[Han et al., 2011]

Data Cube Aggregations

Relation A

Country	*Device*	*Language*
US	Android	en
US	iPhone	ru
South Africa	iPhone	en
India	Android	en
Australia	iPhone	en

Aggregation B

Country	*Device*	*Language*	*Count*
All	*All*	*All*	5

Group By on *Device, Language* C

Country	*Device*	*Language*	*Count*
All	Android	en	2
All	iPhone	en	2
All	iPhone	ru	1

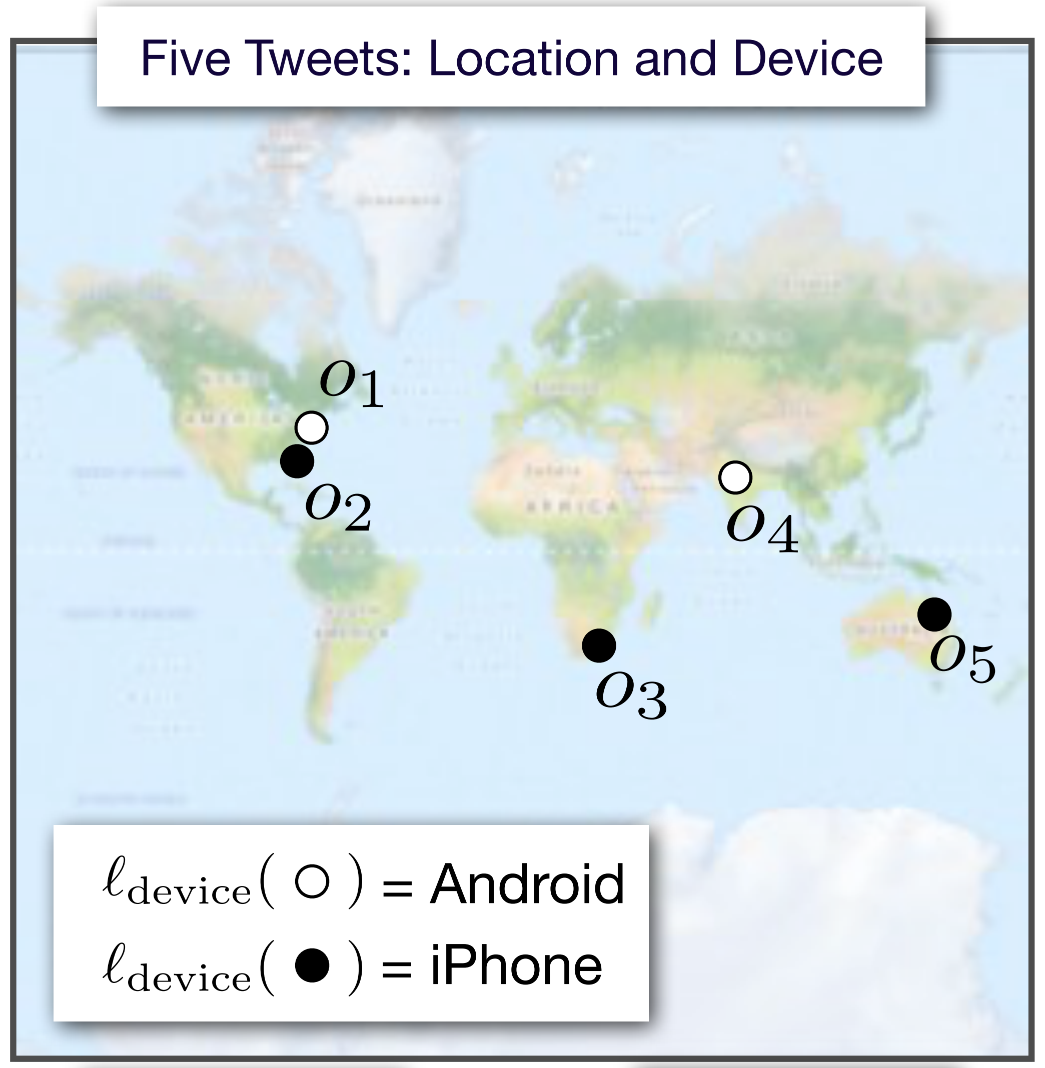
Cube on *Device, Language* D

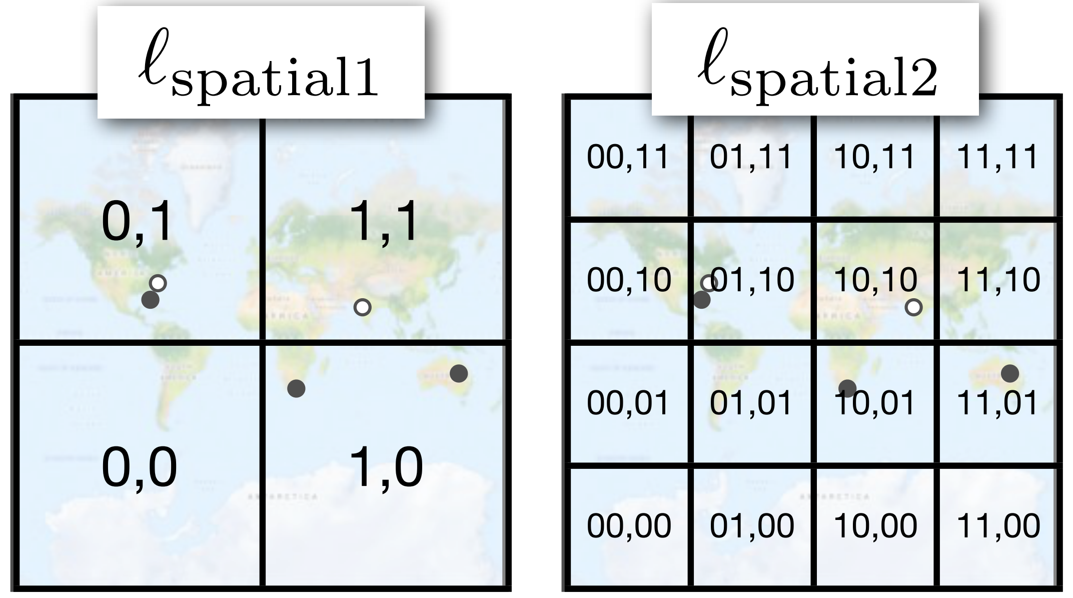
Country	*Device*	*Language*	*Count*
All	*All*	*All*	5
All	Android	*All*	2
All	iPhone	*All*	3
All	*All*	en	4
All	*All*	ru	1
All	iPhone	ru	1
All	Android	en	2
All	iPhone	en	2

Equivalent to Group By on all possible subsets of {*Device, Language*}

[Lins et. al, 2013]

Building a Nanocube


Five Tweets: Location and Device
o_1
o_2
o_3
o_4
o_5
$\ell_{\text{device}}(\circ)$ = Android
$\ell_{\text{device}}(\bullet)$ = iPhone



ℓ_{spatial1}

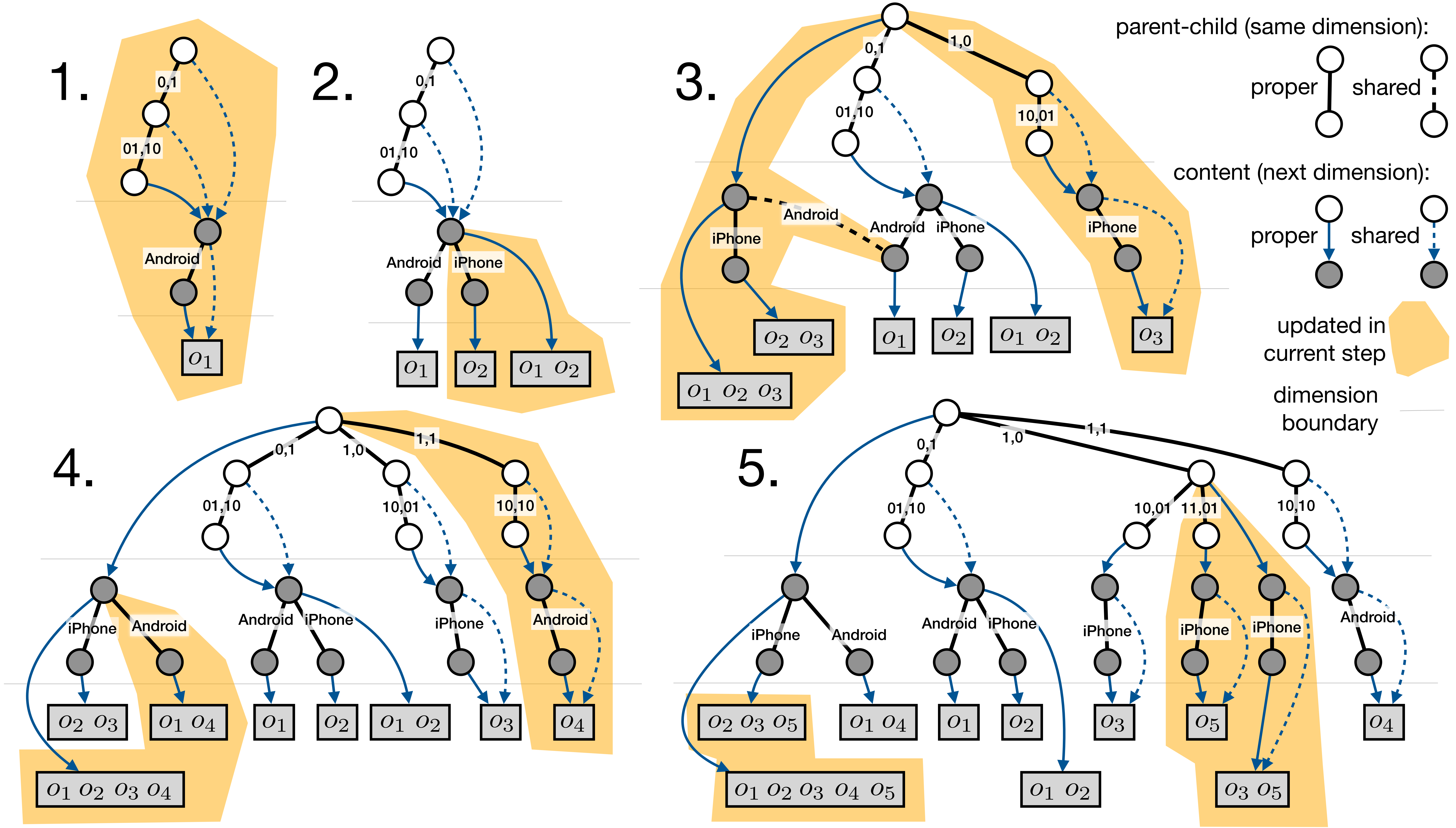
0,1	1,1
0,0	1,0

ℓ_{spatial2}

00,11	01,11	10,11	11,11
00,10	01,10	10,10	11,10
00,01	01,01	10,01	11,01
00,00	01,00	10,00	11,00

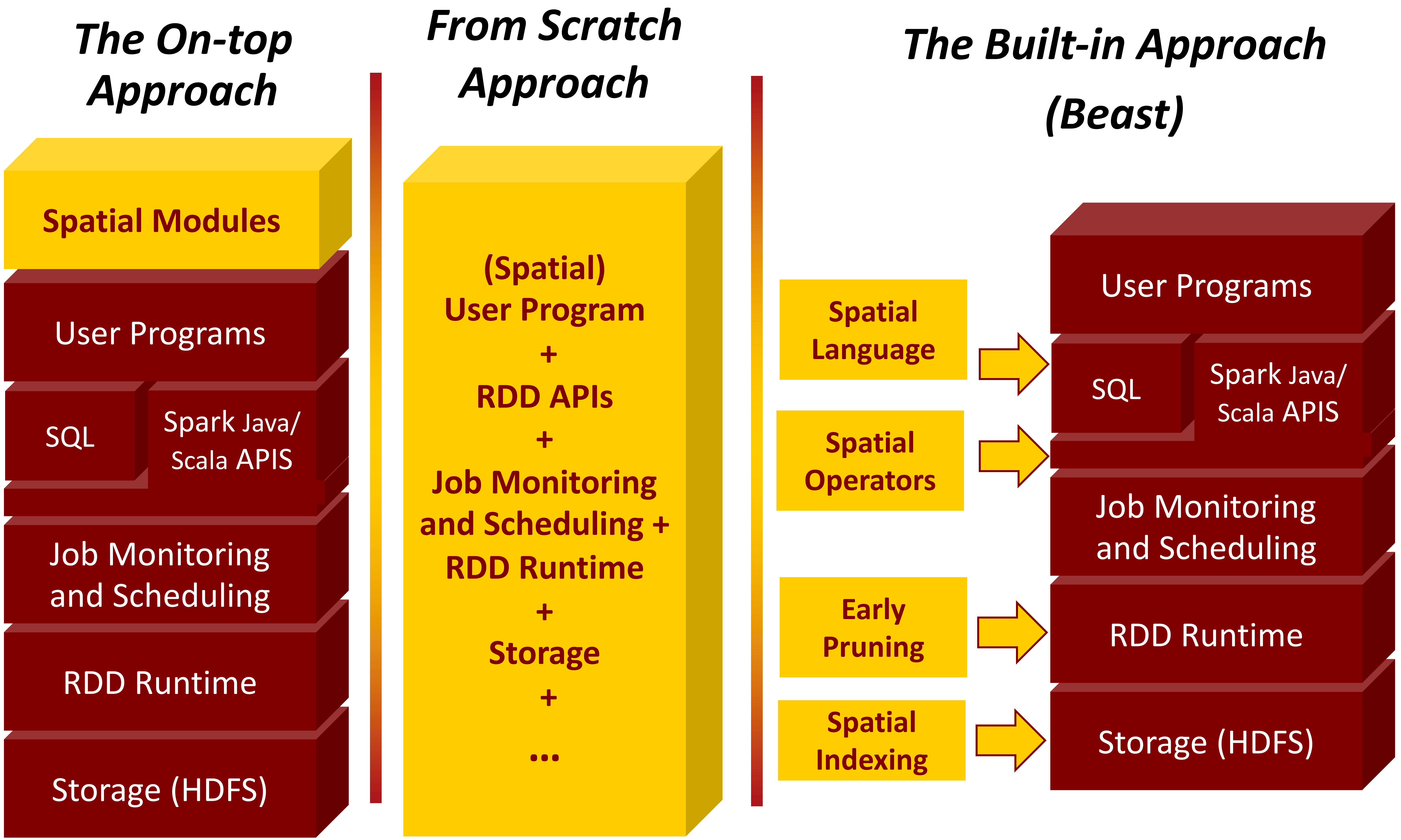
Indexing Schema

$$S = [\,[\ell_{\text{spatial1}}, \ell_{\text{spatial2}}], [\ell_{\text{device}}]\,]$$



[Lins et. al, 2013]

Beast Architecture



The On-top Approach

Spatial Modules

User Programs

SQL

Spark Java/ Scala APIS

Job Monitoring and Scheduling

RDD Runtime

Storage (HDFS)

From Scratch Approach

(Spatial) User Program + RDD APIs + Job Monitoring and Scheduling + RDD Runtime + Storage + …

The Built-in Approach (Beast)

Spatial Language

Spatial Operators

Early Pruning

Spatial Indexing

User Programs

SQL

Spark Java/ Scala APIS

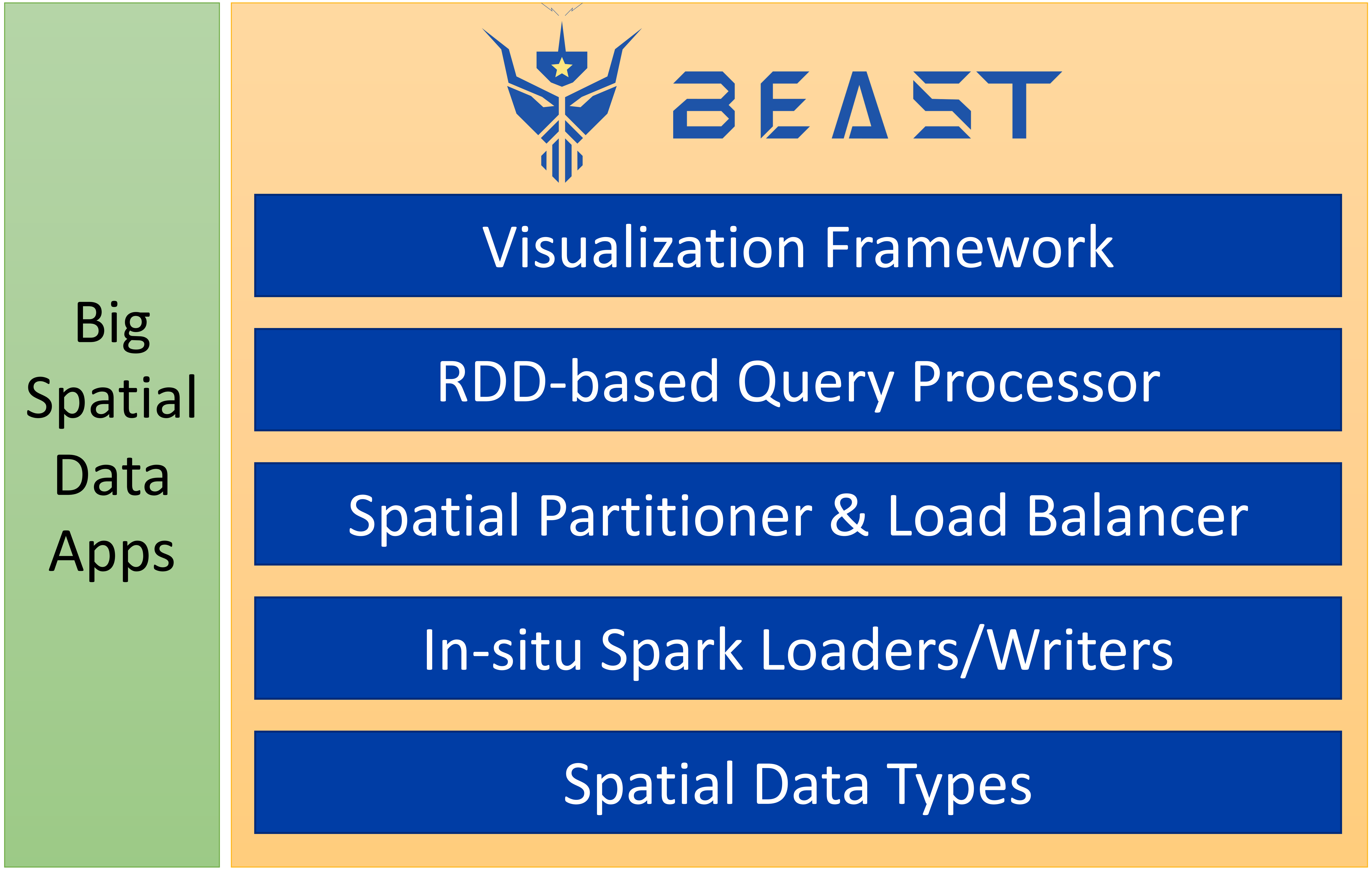
Job Monitoring and Scheduling

RDD Runtime

Storage (HDFS)

[A. Eldawy, 2021]

Beast Architecture



Big Spatial Data Apps

BEAST

Visualization Framework

RDD-based Query Processor

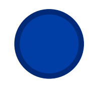
Spatial Partitioner & Load Balancer

In-situ Spark Loaders/Writers

Spatial Data Types

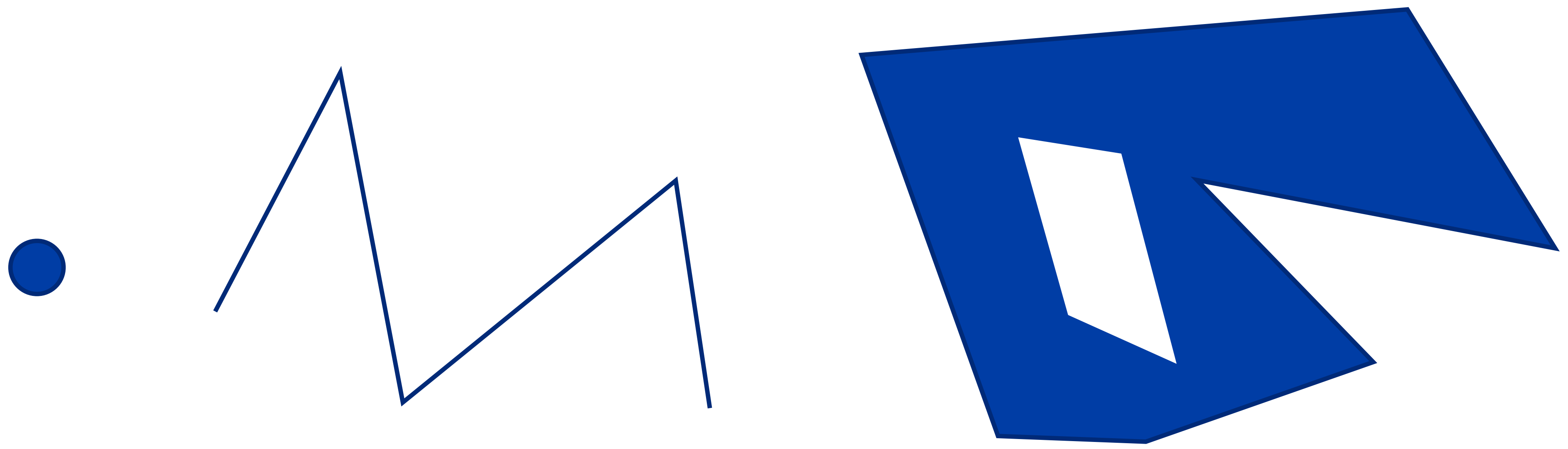
[A. Eldawy, 2021]

Beast Spatial Data Types



Point

Envelope



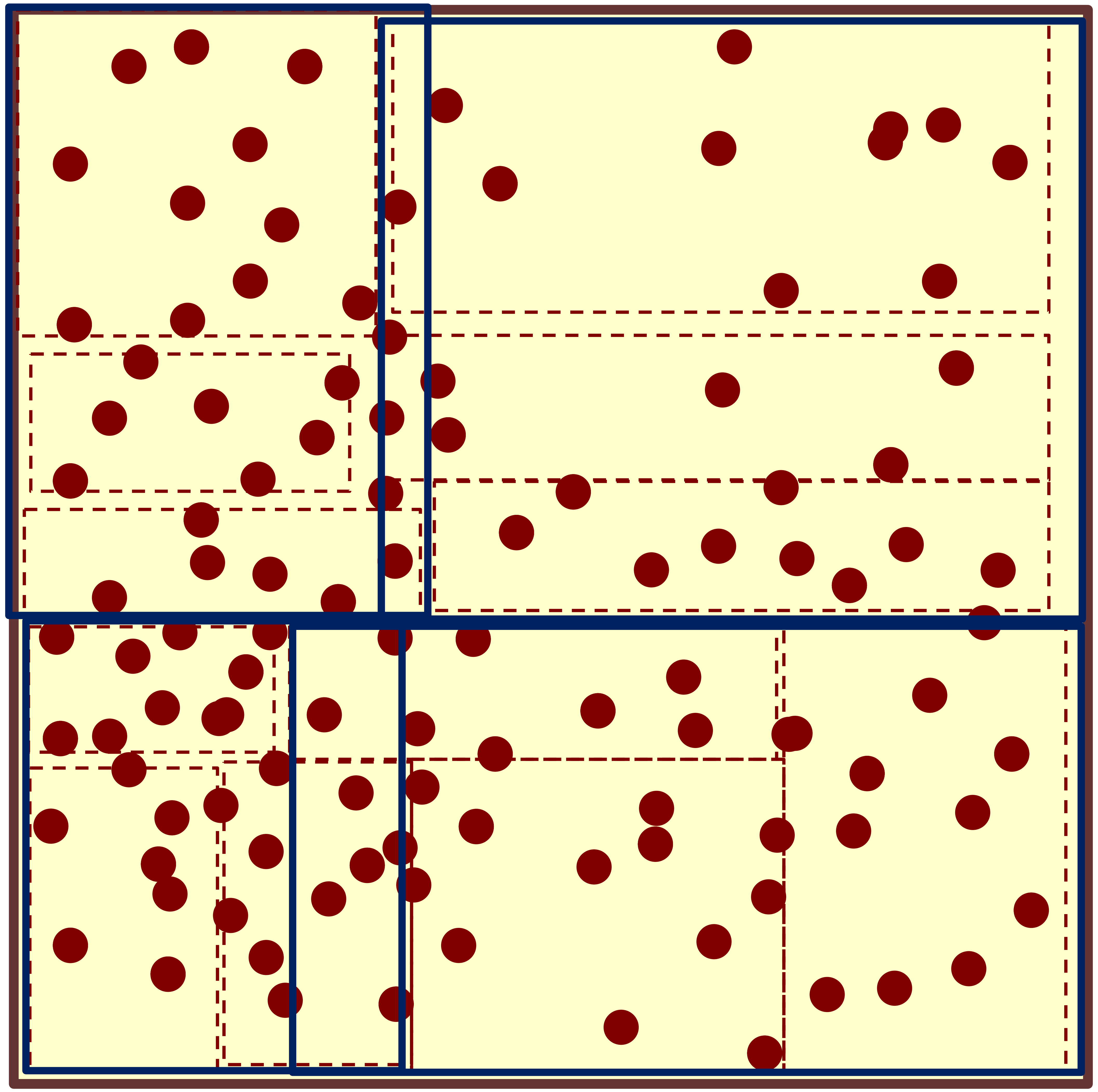
Geometry

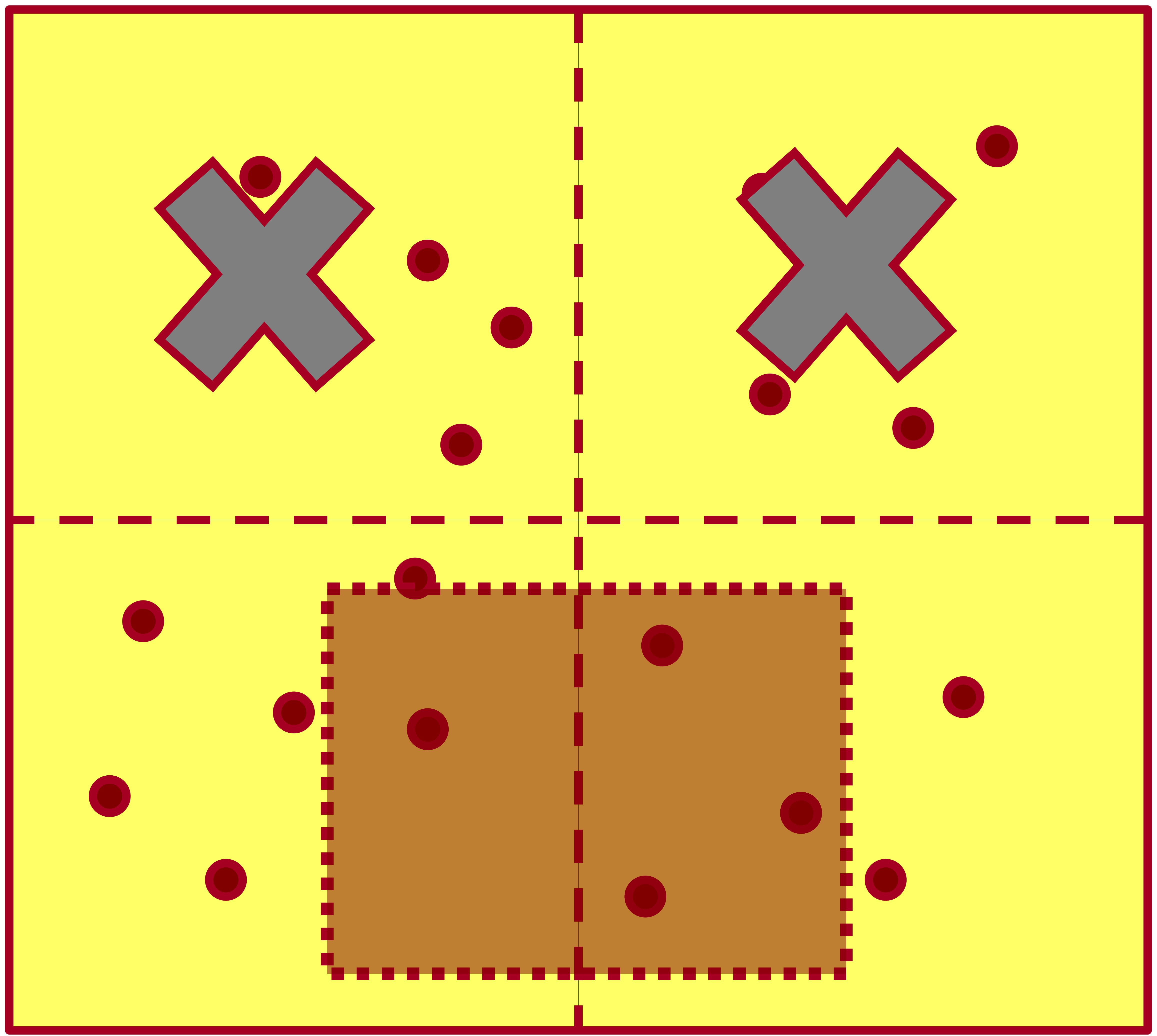
Washington
Oregon
Idaho
Nevada
Utah
California
Arizona

Feature

[A. Eldawy, 2021]

Beast Partitioning/Indexing & Range Query

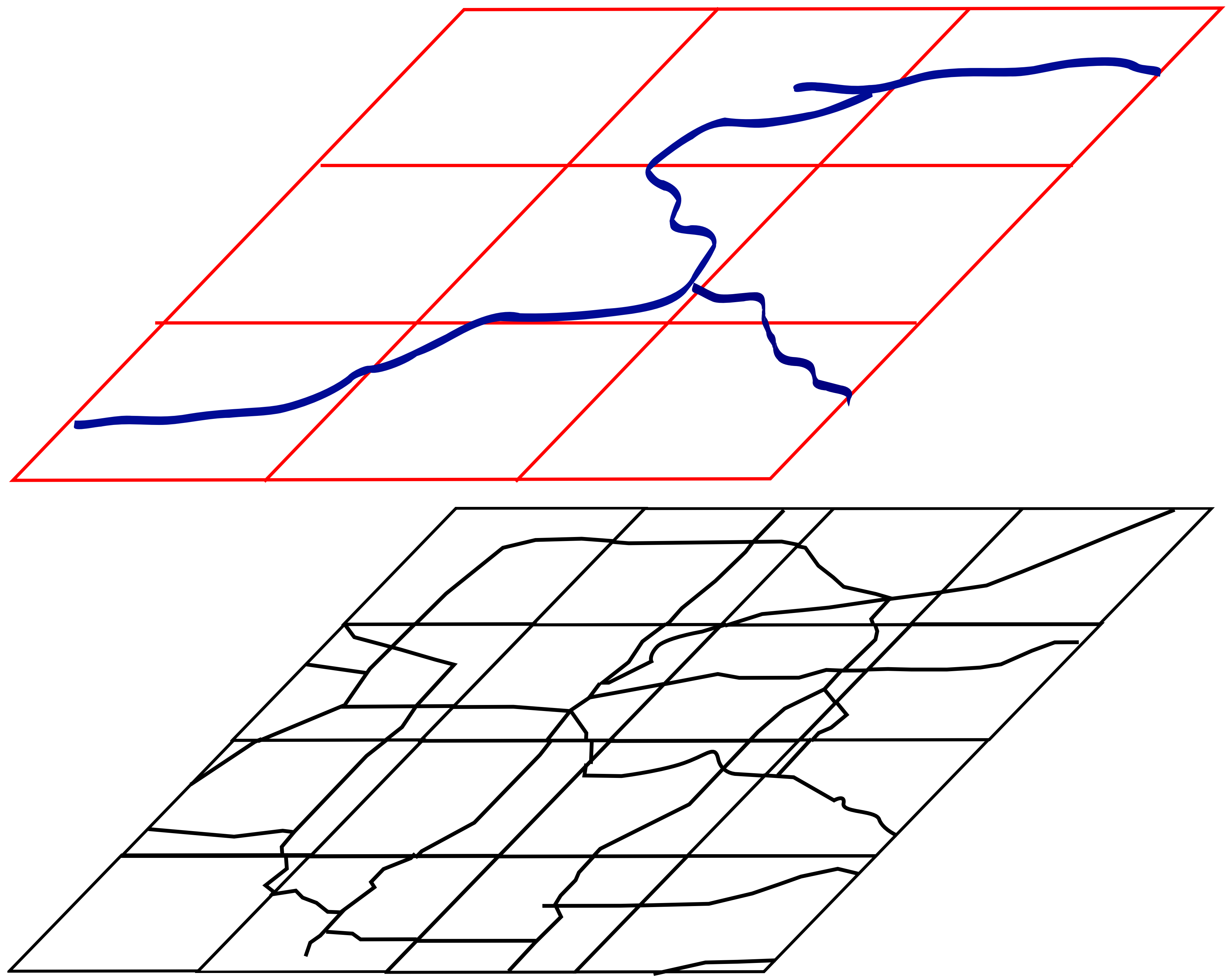




[A. Eldawy, 2021]

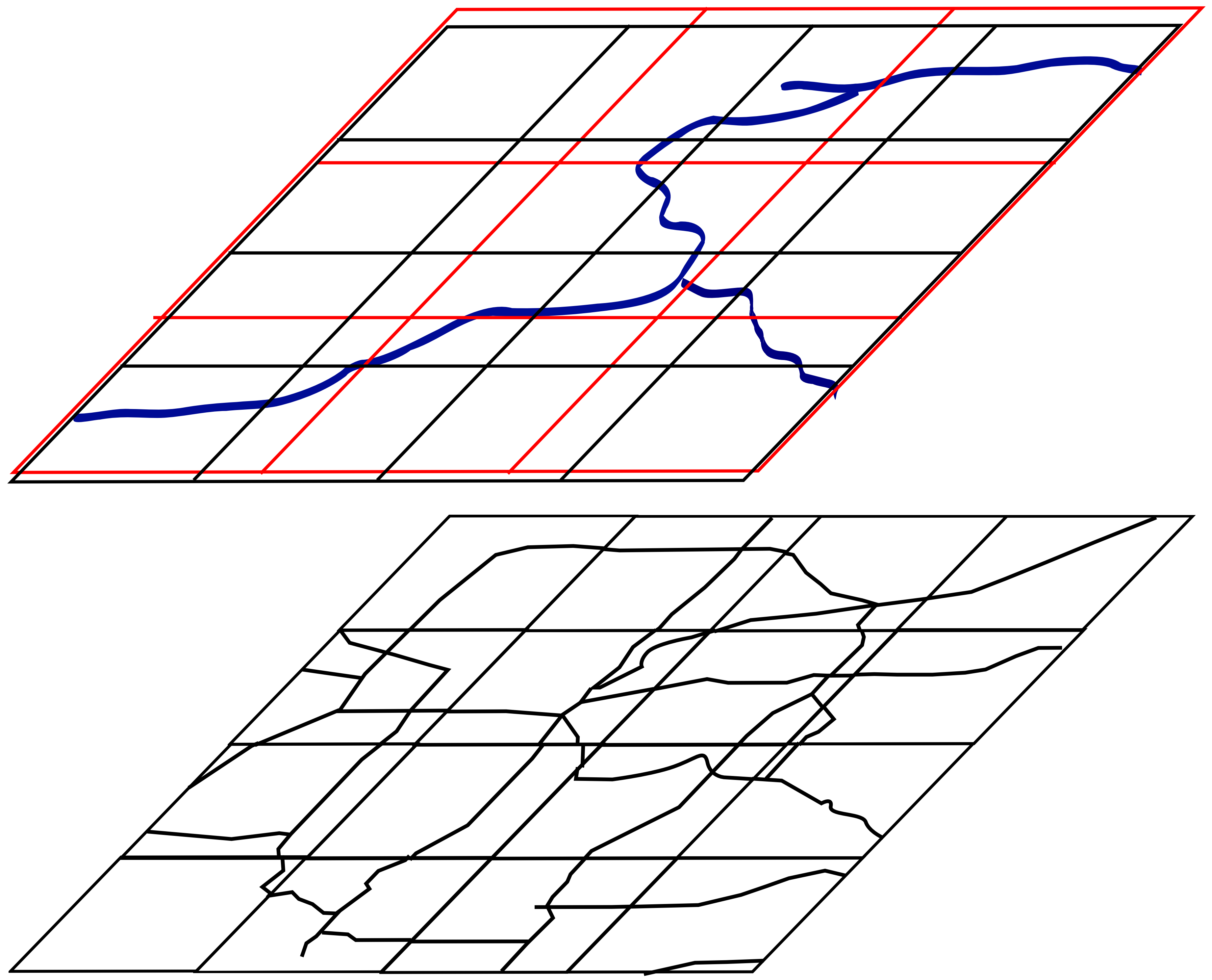
Beast Spatial Join

Join Directly



Total of 36 overlapping pairs

Partition – Join



Only 16 overlapping pairs

[A. Eldawy, 2021]

Assignment 5

- Divvy Bikes Data
- Spatial, Graph, and Temporal Data Processing
- Use pandas, geopandas, neo4j, (modin for extra credit)
- Due May 1

Data Curation

Why?

Big Data, Little Data, or No Data?

Why Human Interaction with Data is a Hard Problem

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

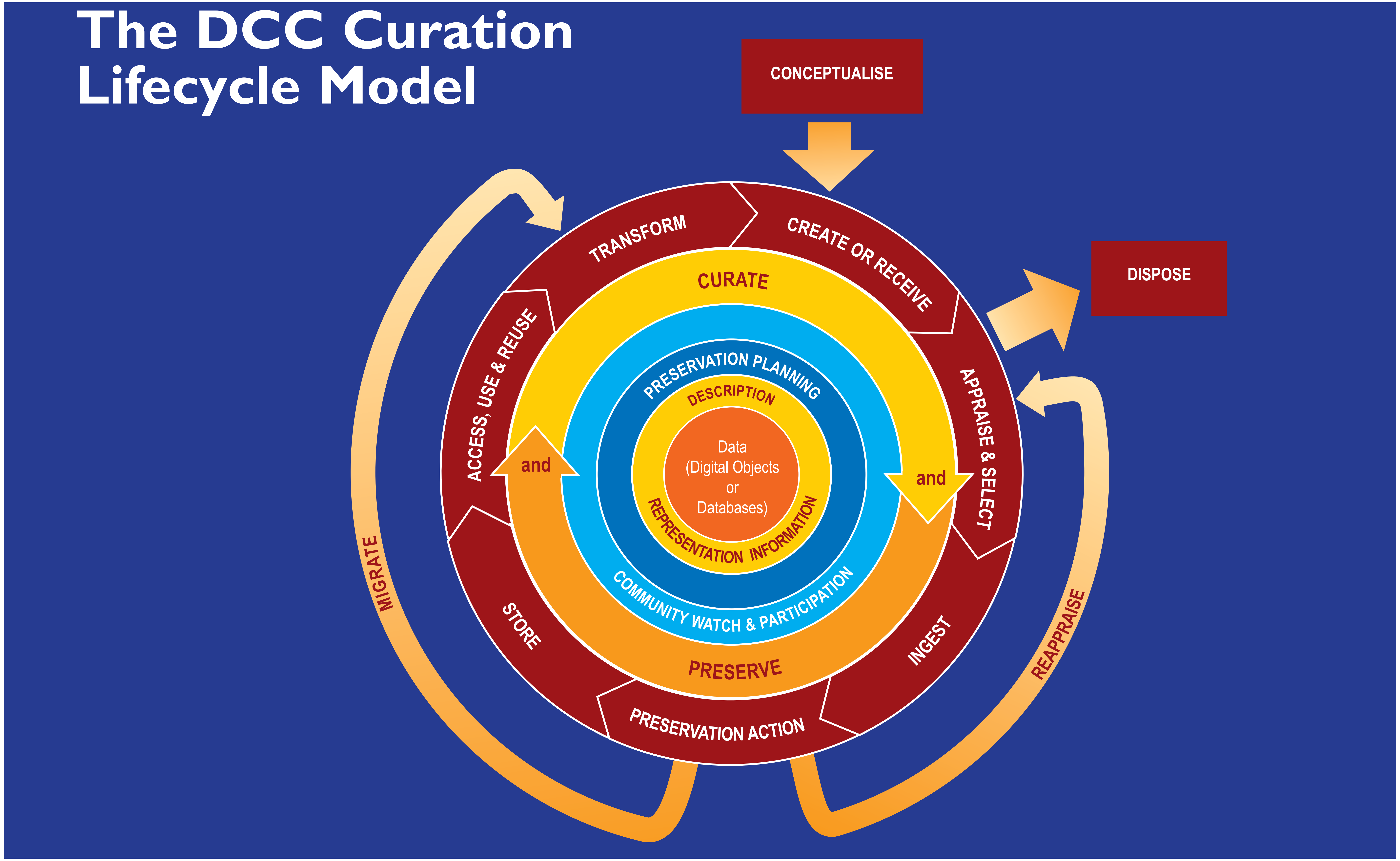
Data Identity

- 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]

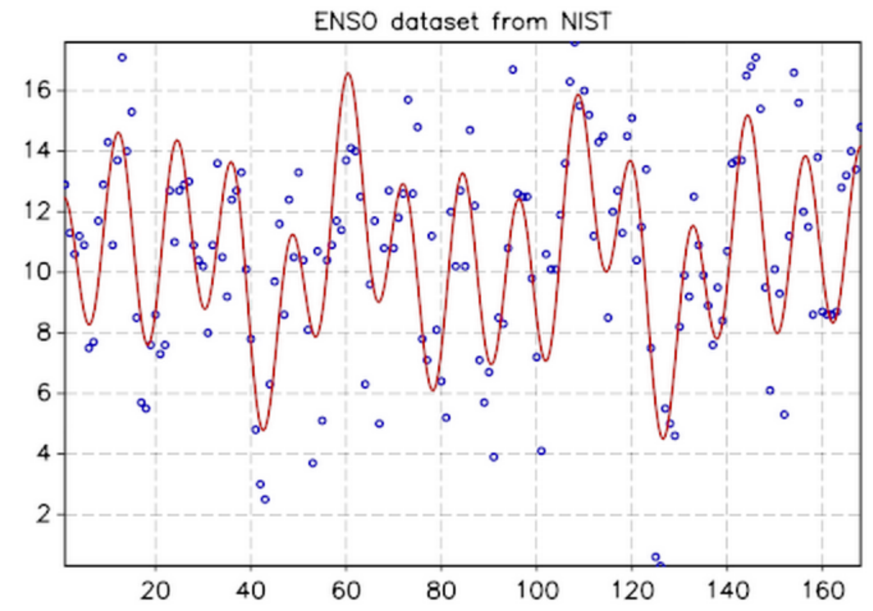
To be Findable

- 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: $H \rightarrow \gamma\gamma$," http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44

☑ Unique ☑ Persistent

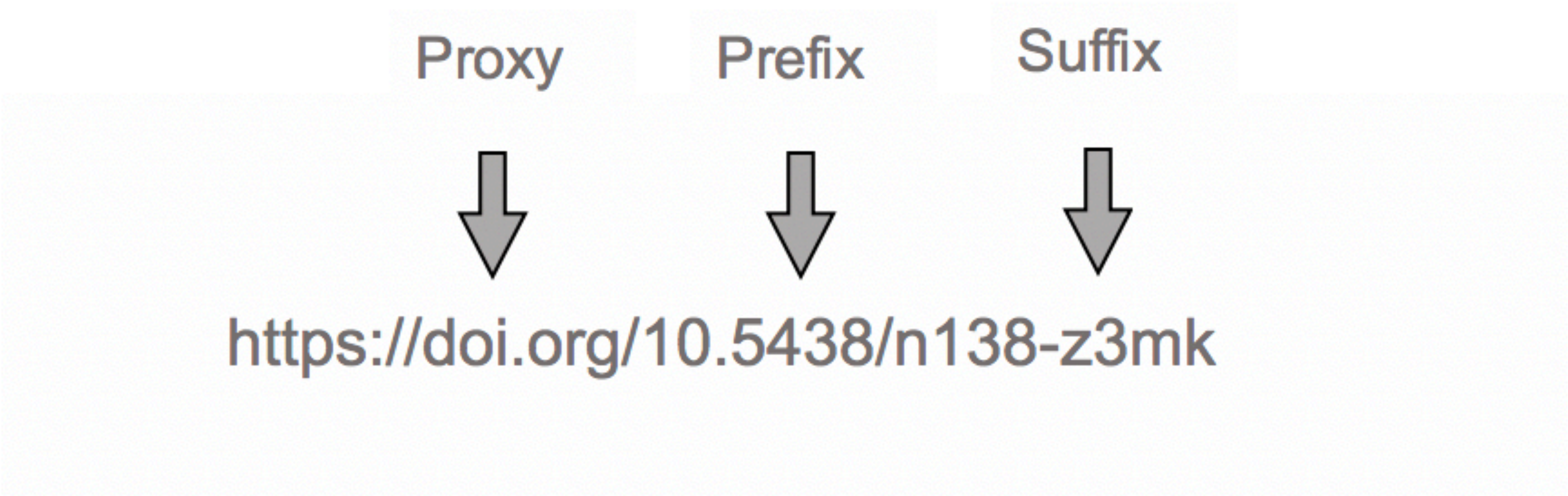
5. Enjoy the benefits

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

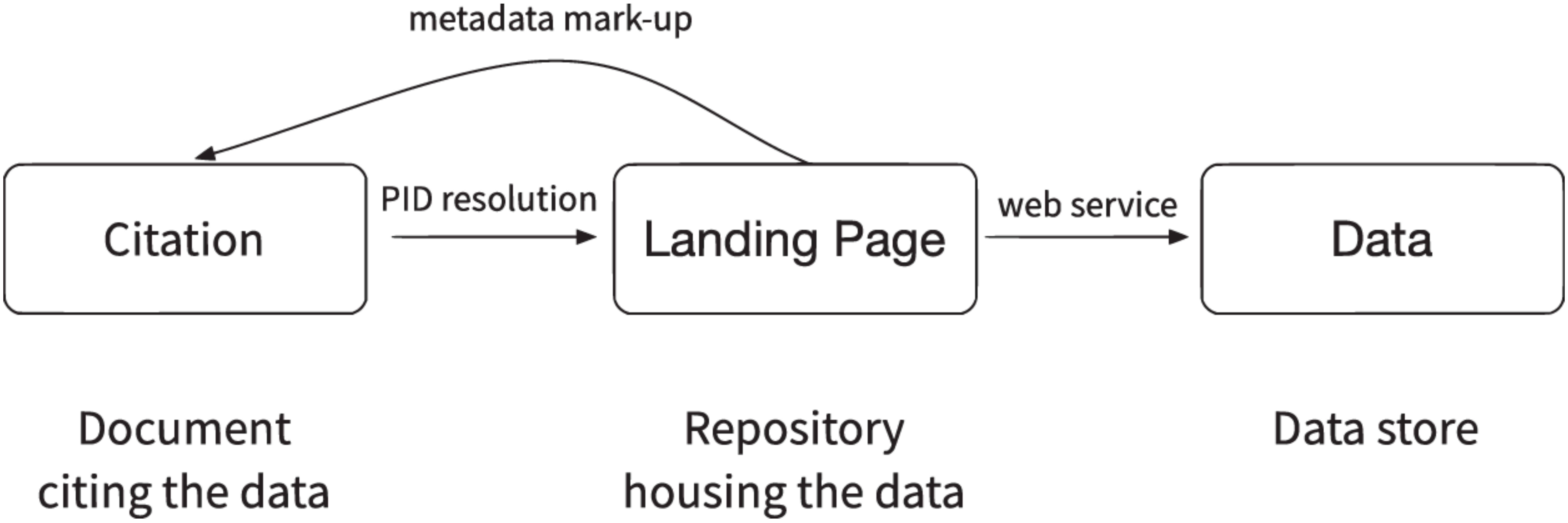
[DataCite]

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



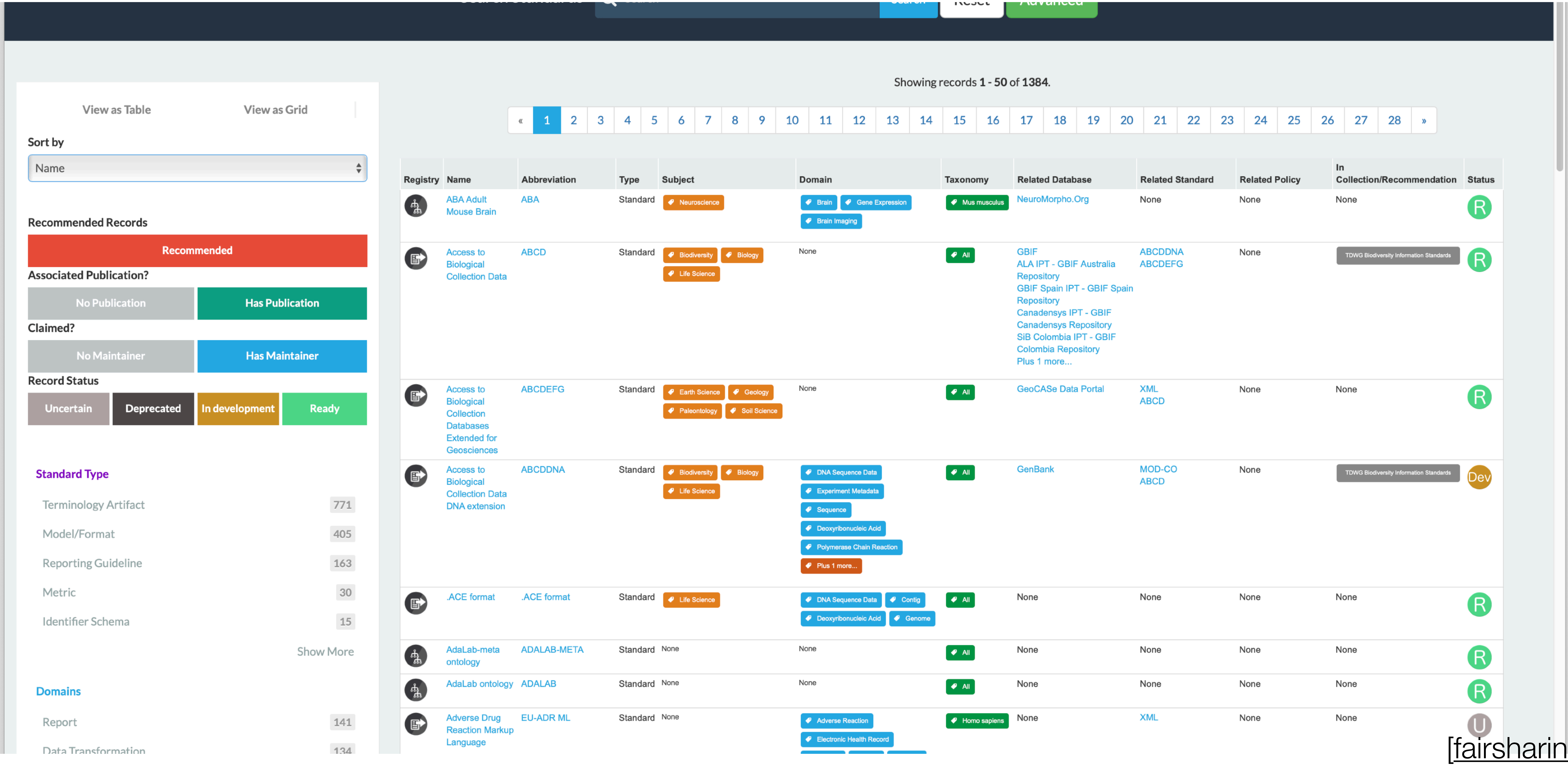
[M. Fenner et al., 2019]

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



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

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 »

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**

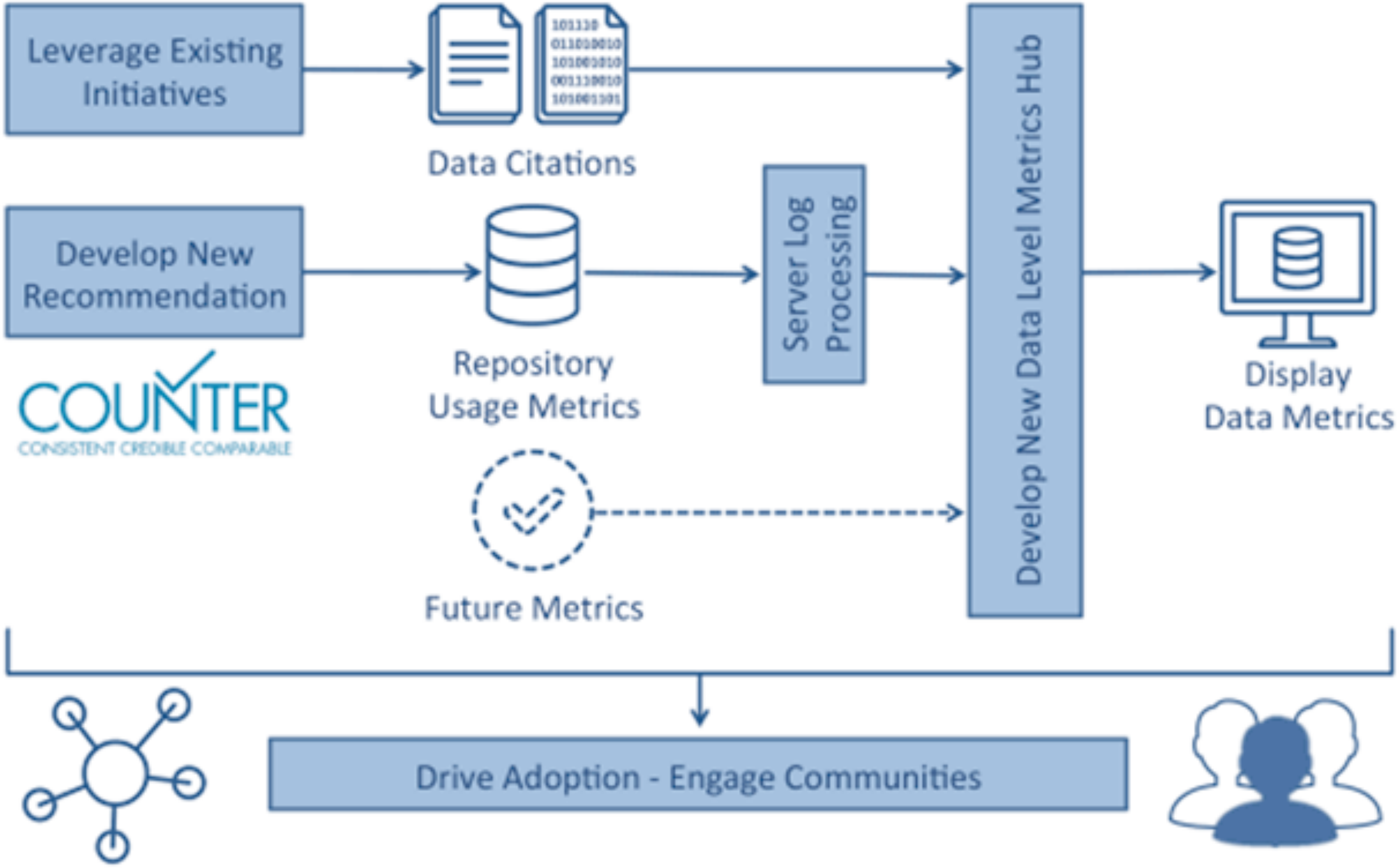
[M. D. Wilkinson et al., 2016]

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
101110
011010010
101001010
001110010
101001101
Data Citations
Develop New Recommendation
COUNTER
CONSISTENT CREDIBLE COMPARABLE
Repository Usage Metrics
Future Metrics
Server Log Processing
Develop New Data Level Metrics Hub
Display Data Metrics
Drive Adoption - Engage Communities

[H. Cousijn et al., 2019]