Advanced Data Management (CSCI 640/490)

Data Curation

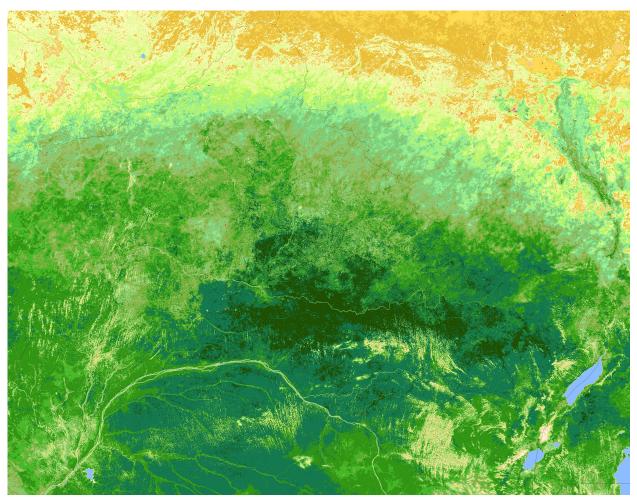
Dr. David Koop





Spatial Data

Measure vegetation density



Measure snow melt

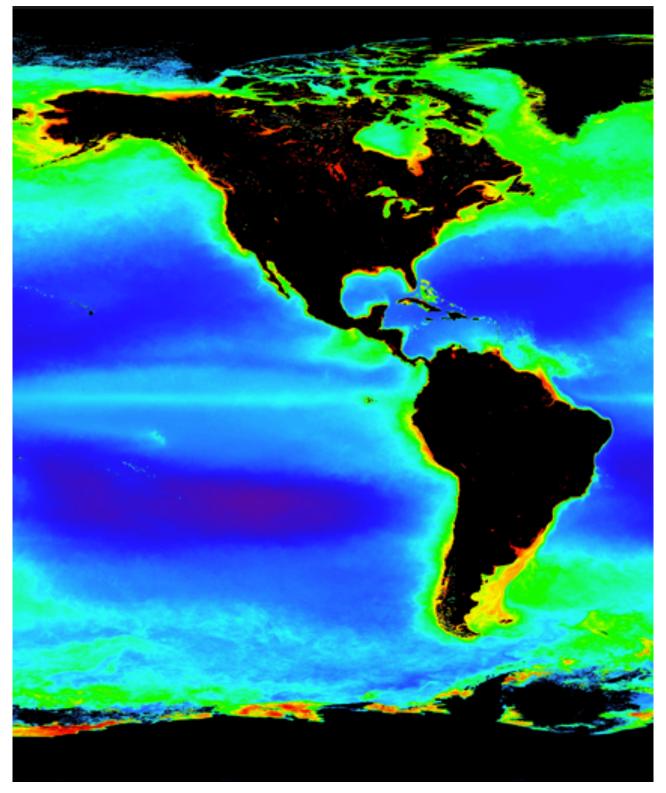


D. Koop, CSCI 640/490, Spring 2023

Track hurricanes



Track phytoplankton populations





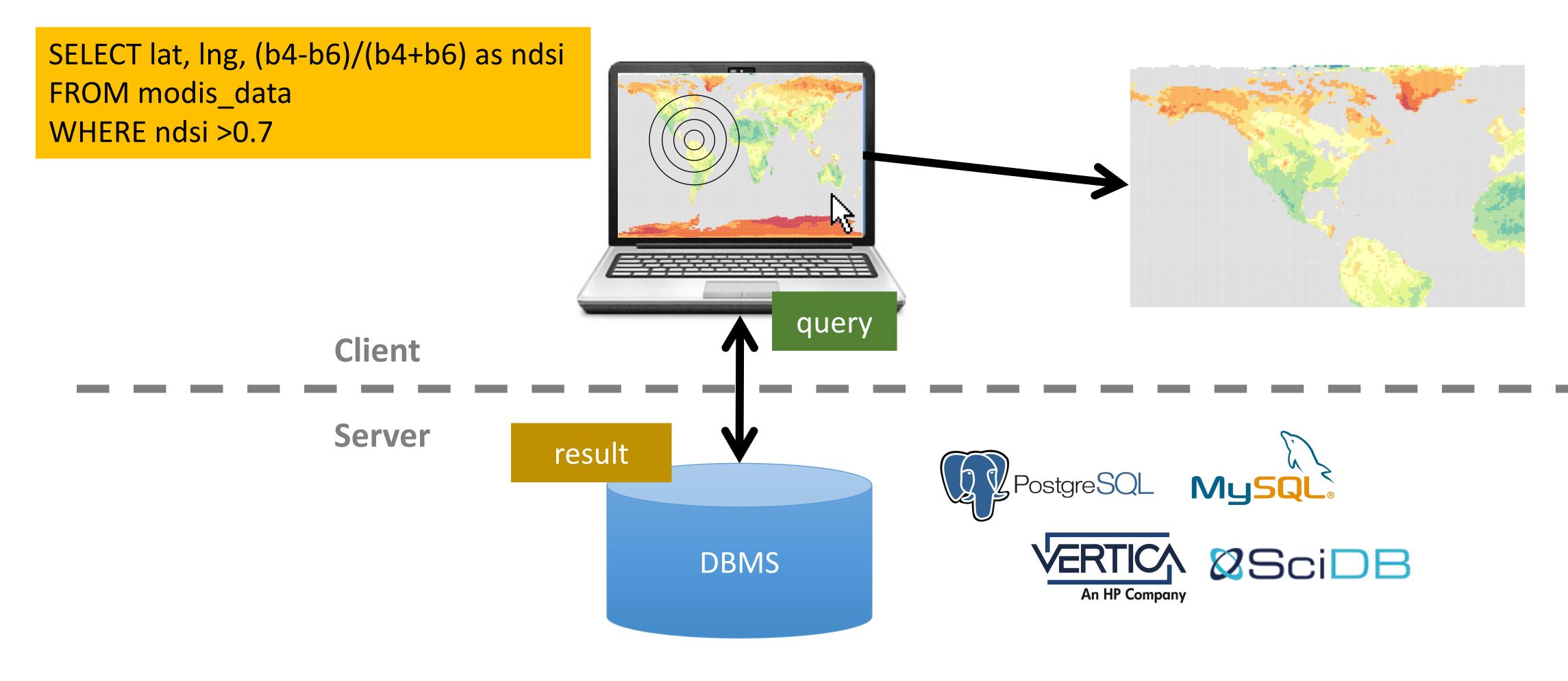








Interactive Exploration of Spatial Data





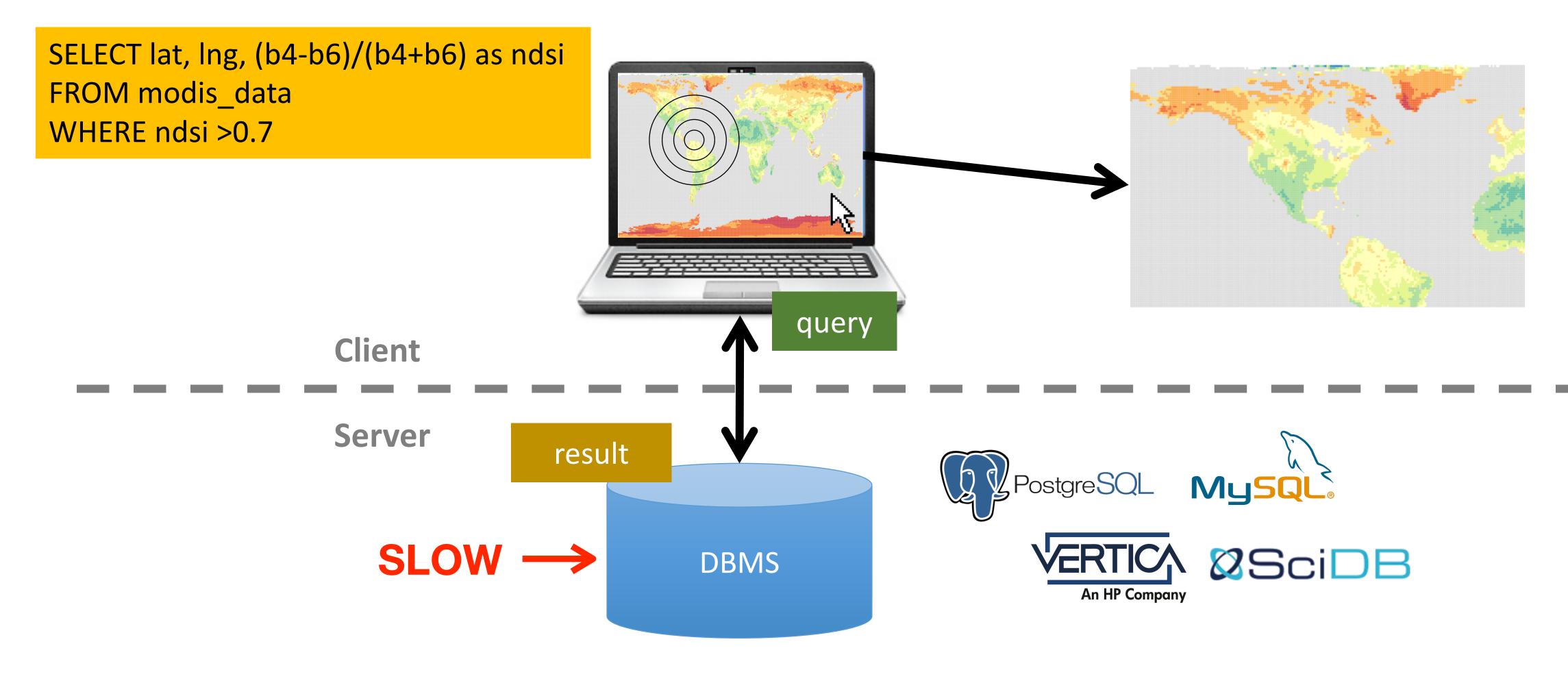








Interactive Exploration of Spatial Data





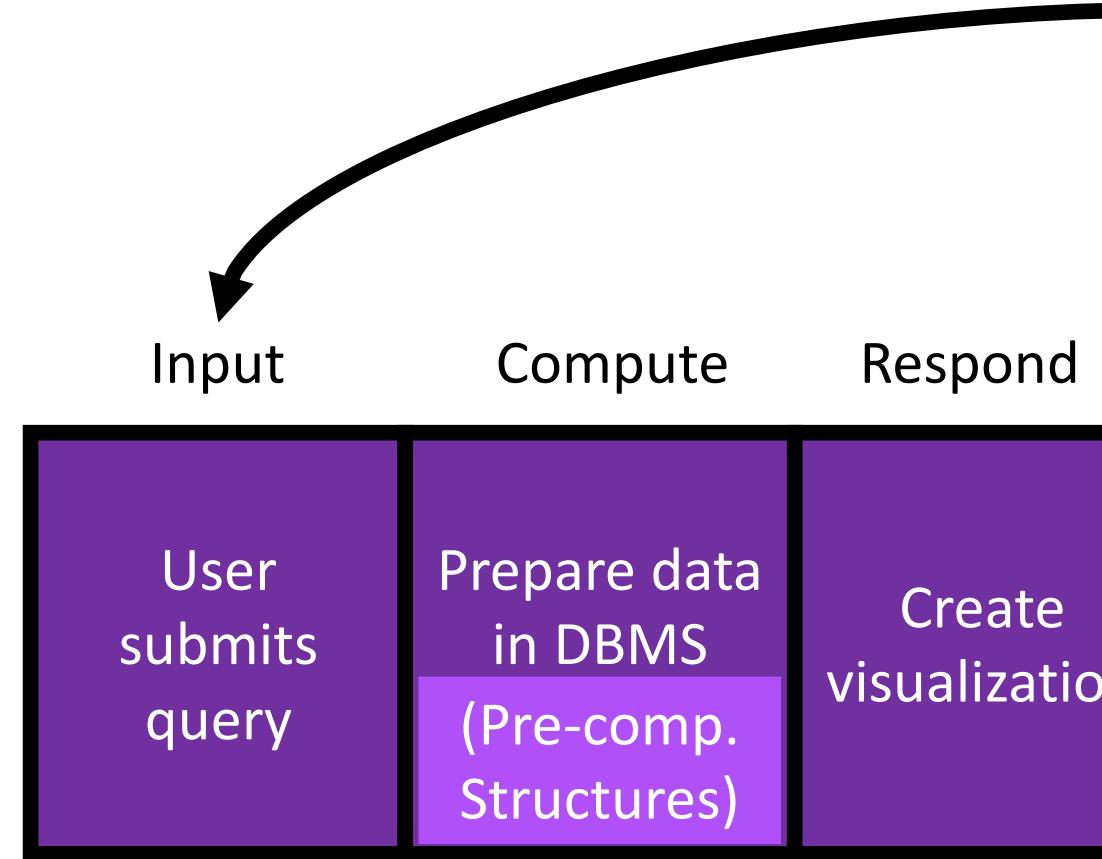








Two Inputs to Exploratory Browsing Input Compute Respond Compute Respond Input Prepare data User Fetch results User Update Create in DBMS submits from DBMS visualization visualization pan/zoom query (Pre-comp. Structures)



Cold start time

D. Koop, CSCI 640/490, Spring 2023

interaction latency < 500ms







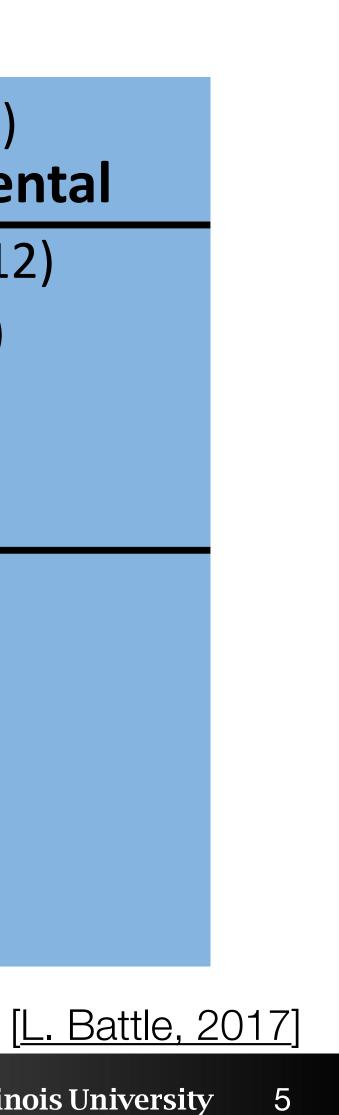


Systems for Interactive Exploration

billing SampleAction (CHI 2012) Vizdom (VLDB 2015) DICE (ICDE 2014) A-WARE (HILDA 2016) A-WARE (HILDA 2016)											
Vizdom (VLDB 2015) DICE (ICDE 2014) A-WARE (HILDA 2016)				· · · ·	(After interaction) Progressive/Incremental						
	ormat	Sampling			Vizdom (VLDB 2015) DE 2014)						
	put	egation	Nanocubes (Infovis 2013) imMens (Eurovis 2013)	ATLAS (VAST 2008) XmdvTool (<i>DASFAA</i> 2003)							
ForeCache	JO	188	ForeCa	che							

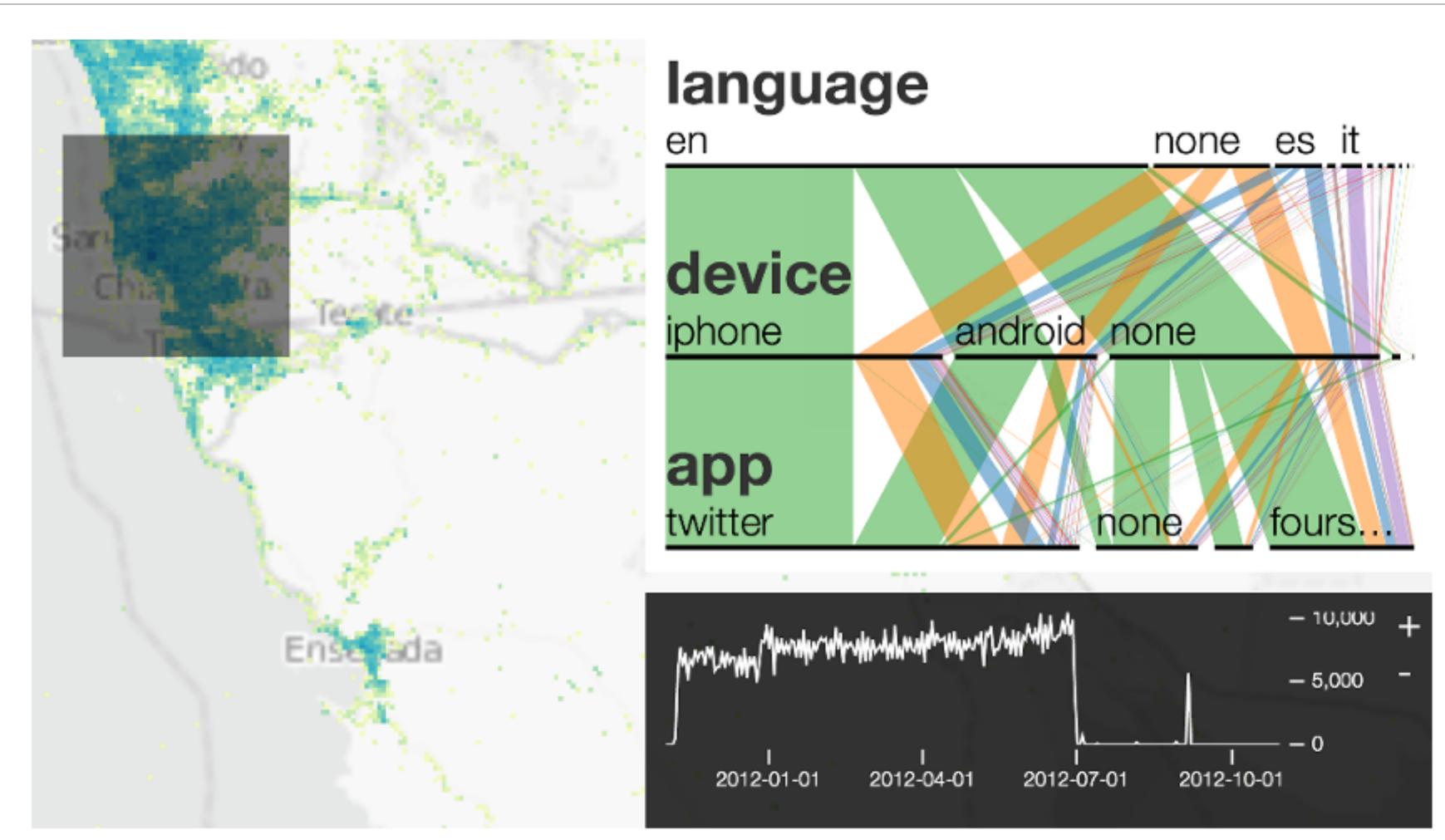
D. Koop, CSCI 640/490, Spring 2023

Time



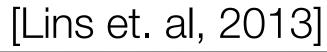


Nanocubes



Linked view of tweets in San Diego, US







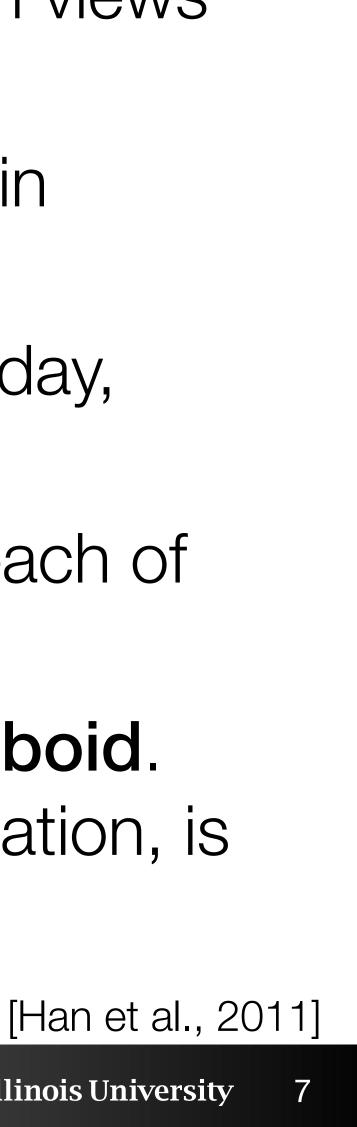
From Tables and Spreadsheets to Data Cubes

- 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
- called the apex cuboid. The lattice of cuboids forms a data cube.

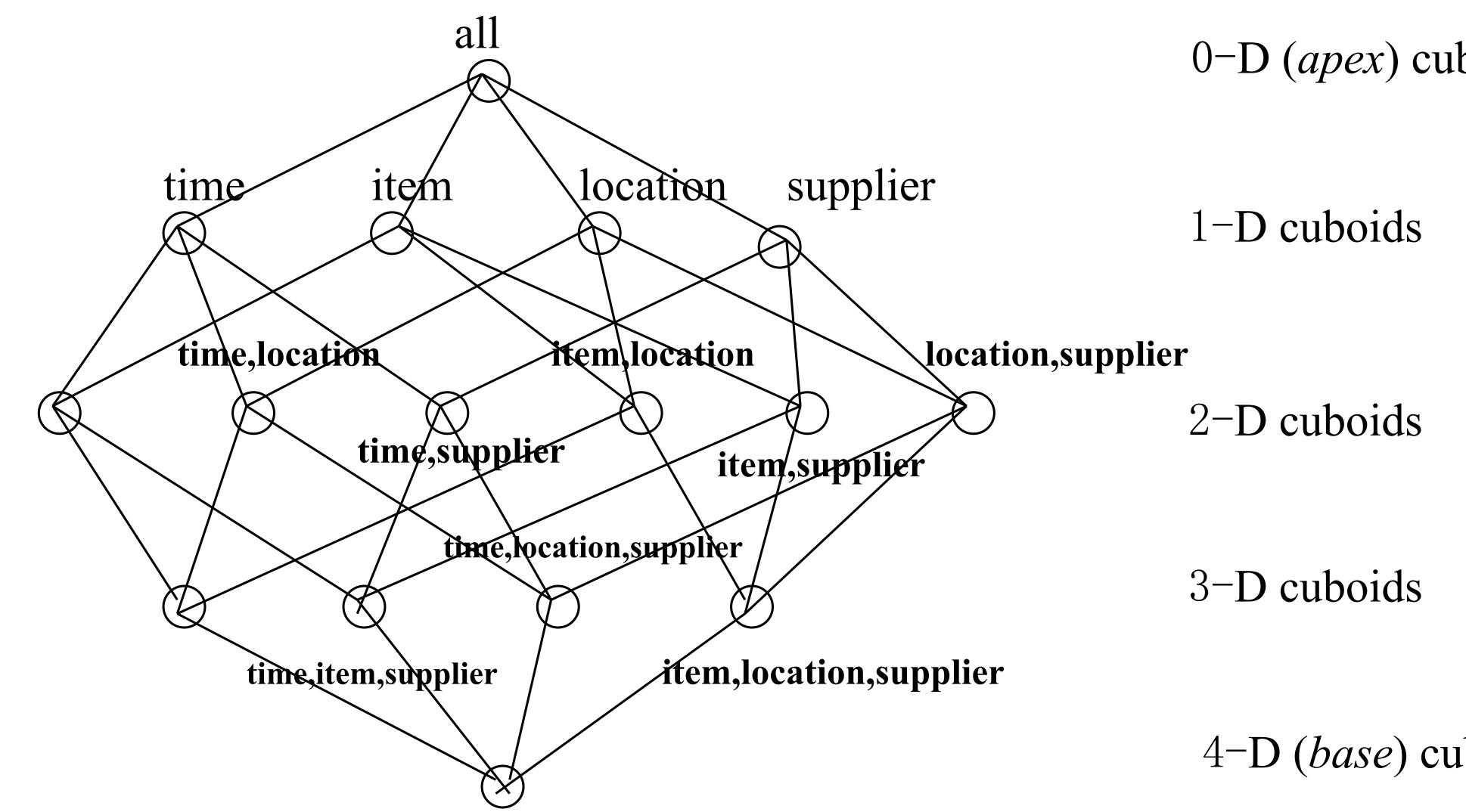
• A data warehouse is based on a multidimensional data model which views

 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





Data Cube: A Lattice of Cuboids

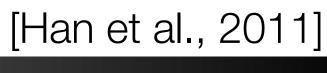


D. Koop, CSCI 640/490, Spring 2023

0-D (*apex*) cuboid

4-D (base) cuboid





8

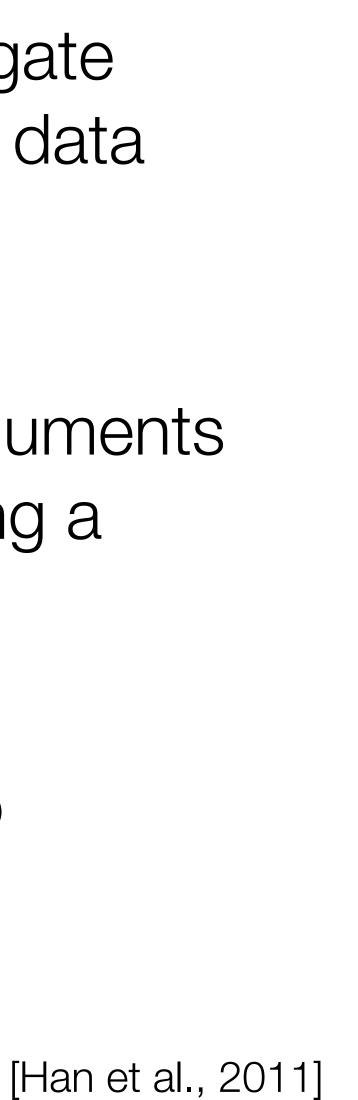
Data Cube Measures: Three Categories

- without partitioning
 - E.g., count(), sum(), min(), max()
- 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()

• **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

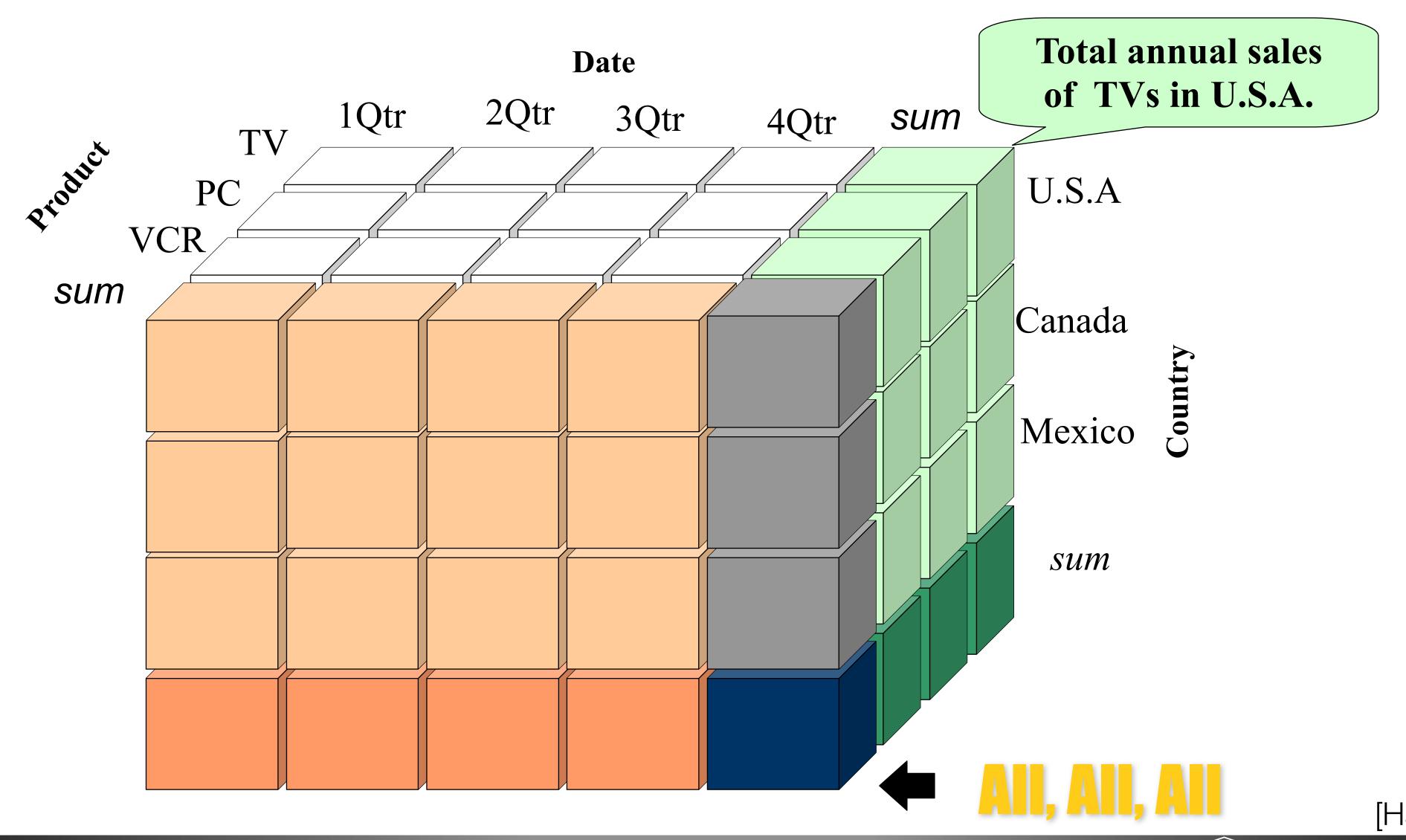
• Algebraic: if it can be computed by an algebraic function with M arguments (where M is a bounded integer), each of which is obtained by applying a





9

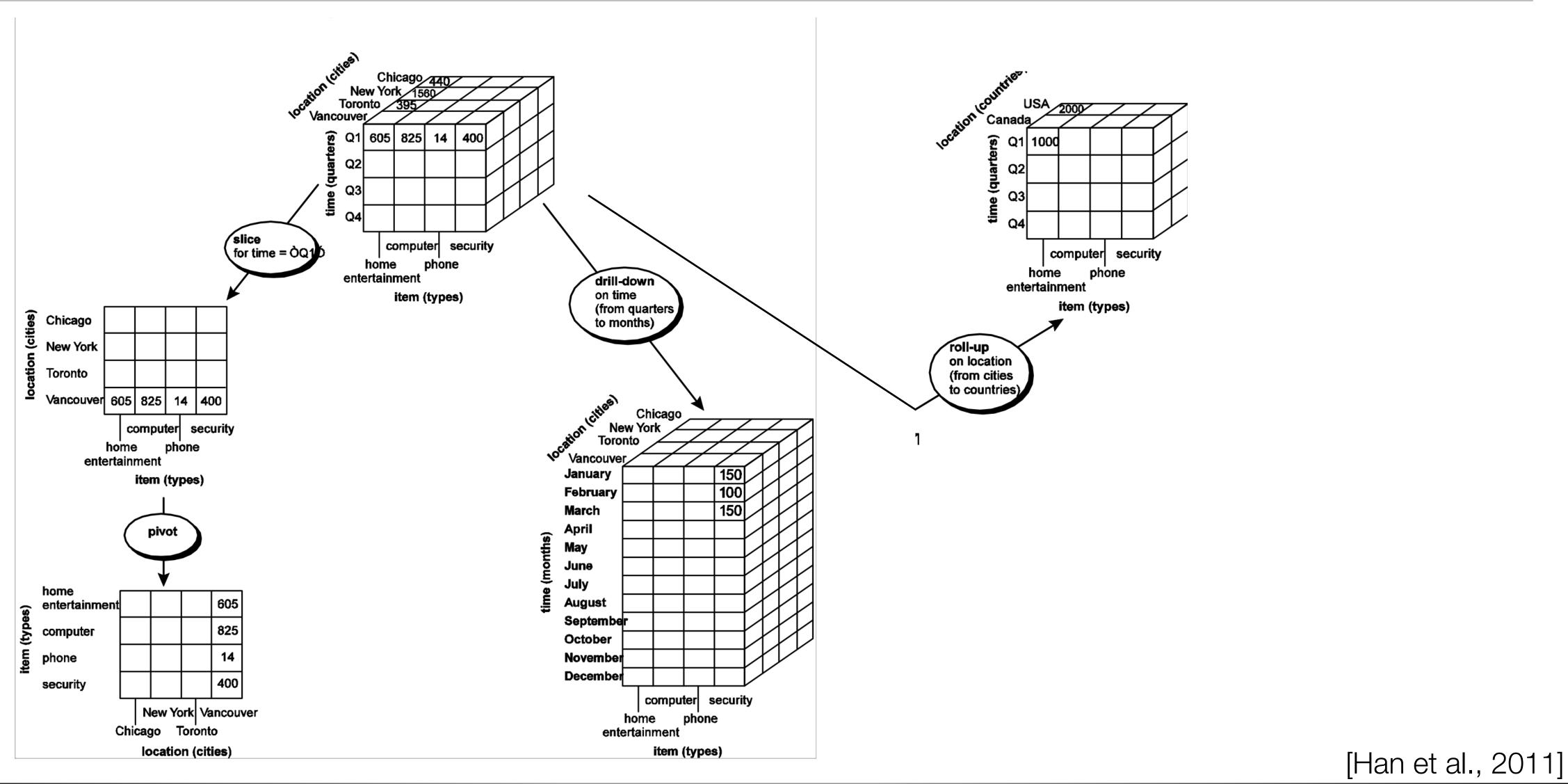
A Sample Data Cube







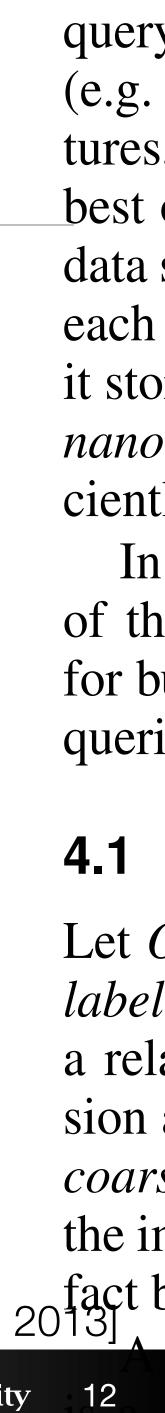
OLAP Operations





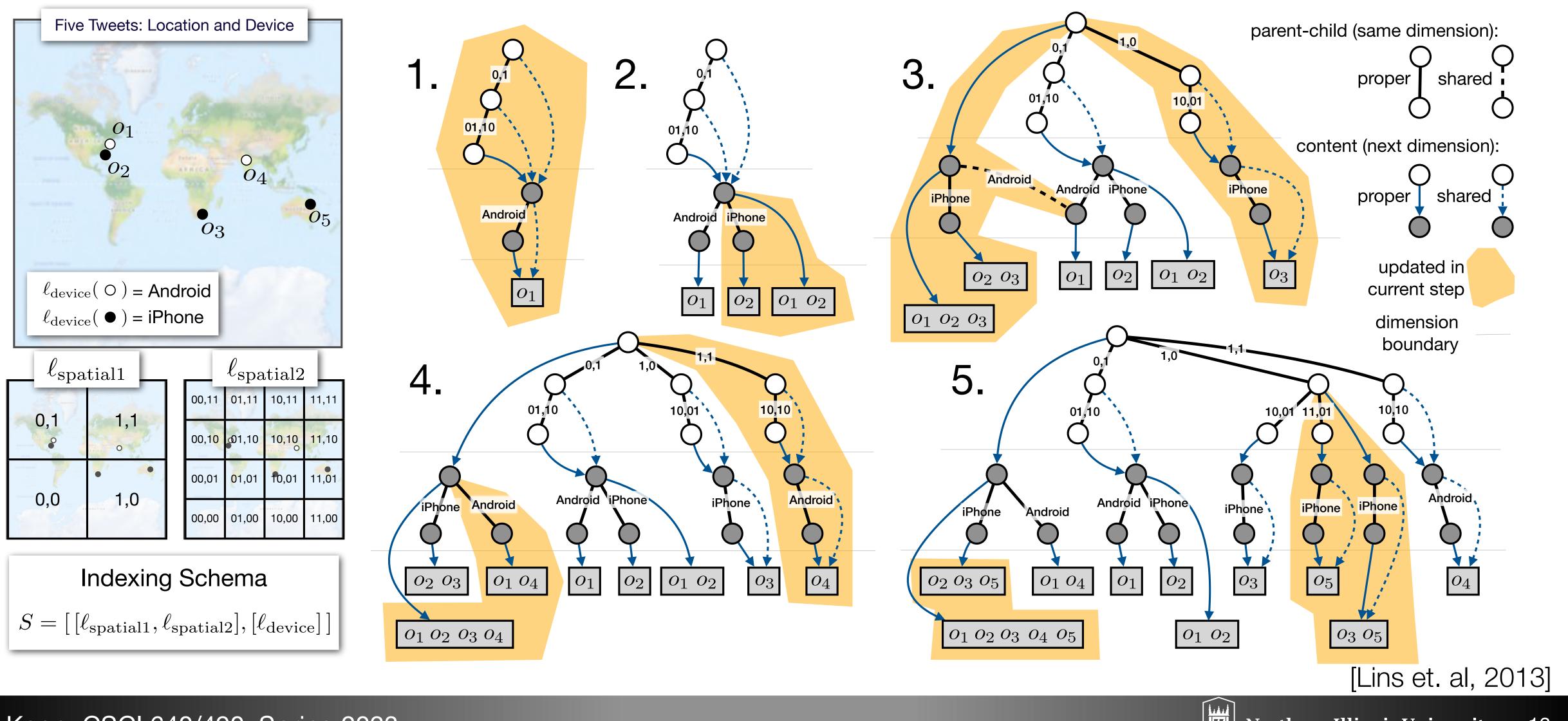


Data Cube According the second of the second	ult. It	ve anow entri	es in the recording	-			U	
Data Cube Arroretoretions For even heref and a breezen of bergen all breezen of bergen all breezen of breezen	d rep	resent this age	greation as the faille	Country	Device	Language	Count	
Interference Interference <td< td=""><td></td><td></td><td>tion For ever</td><td>All</td><td>All</td><td>All</td><td>5</td><td></td></td<>			tion For ever	All	All	All	5	
is e using counted midths case, five way and an analysis of a second and a second an a second and a second and a second and a second an a seco			Language Config to	All	Android	All	2	
All A	A / / .			All	iPhone	All	3	
All iPhone ru 1 of describing a degrado for a given d for a grad by by the first of a given d for a grad by the first of a grad		ותי יו		All	All	eu	4	
in descripting		welation hate	lon thin the special	All	All	ru	1	
<i>Primer group by</i> speed attributes a <i>Primer group by</i> speed attributes attributes a <i>Primer group by</i> speed attributes attrbibutes attributes attributes attr				All	iPhone	ru	1	
<i>Country Device Construction of the information of </i>	of desc	ribing a reg	ations for a given	All	Android	en	2	
se felation given a list of attributes a se felation given a list of attributes a <i>Country Device Language</i> C <i>All Android for attributes and the offel</i> C <i>All Android for attributes and the offel</i> C <i>All iPhone group by on Litter to Group By on up by's on: (1) e. Note that the As the results and group by's strengthere and the and the As the results attributes is an an accurate to group by's strengthere and the and the attributes is an an accurate the anothere and the attributes is an anothere and the attributes of group by's strengthere and the attributes is an another and the attributes attributes and the attributes of group by's strengthere attributes and a strengthere and a strengthere attributes and the attributes attributes and the attributes </i>	Sount	ryroubebiçeope	Lationas on Eantwhi	All	iPhone	en	2	
Country Device Language Language Country Device Language Country Device Language Language Country Equivalent to Group By on all possible subsets of all possible subsets of all possible subsets of all possible subsets of group by on Linguity. {Device, Language} Equivalent to Group By on all possible subsets of all possible subsets of all possible subsets of all possible subsets of group by on Linguity. {Device, Language} As the results All All iPhone group by on Linguity. {Device, Language} As the results All iPhone group by on Linguity. {Device, Language} As the results All iPhone group by on Linguity. {Device, Language} As the results All iPhone group by on Linguity. {Device, Language} As the results All iPhone group by on Linguity. {Device, Language} As the results All iPhone group by on Linguity. {Device, Language} As the results All iPhone group by on Linguity. {Device, Language} As the results All iPhone group by on Linguity. {Device, Language} As the results All iPhone group by on Linguity. {Device, Language} As we will describe nanocubes is and query cubes of roll ups. All iPhone </td <td></td> <td>ation given a</td> <td>list of attributes a</td> <td></td> <td></td> <td></td> <td></td> <td></td>		ation given a	list of attributes a					
<i>Country Device Language Langu</i>					4			
<i>Country Device</i> <i>All</i> Android <i>All</i> Android <i>All</i> iPhone <i>All</i> iPho								peration where
All Android All Android All iPhone All			the order . A	TAPA				ll up on Device
All iPhone and attributes; (2502) all possible subsets of e. Note that the All iPhone group by on Linear to be an accude a results of group by short a remover and a second as results of group by short a remover and a second as results and an accude a result of group by short a remover and a second as results and a second as results and an accude a result of group by short a remover and a second as results and an accude a result of group by short a remover and a second as results and an accude a result of the second as results and an accude a result of group by short a remover a result of the second as results and the results and the second as the second as results and the second as the second as the second as results and the second as the sec		~			quivalent to	Group By	on	up by's on: (1)
All iPhone group by on Line (Device, Language) As the results of group by's field framework of second as relations, we can additional and the second as relations, we can additional and the second as relations of second as relations, we can additional and the second as relations of second as relations, we can additional and the second as relations of second as relations, we can additional and the second as relations of second as relations, we can additional and the second as relations of second as relations, we can additional and the second as relations of second as relations, we can additional and the second as relations of second as relations, we can additional and the second as relations of the second as relations, we can additional and the second as relations of the second as relations, we can additional and the second as relations of the second as relations, and the second as relations of the second as relations of the second as relations, we can additional and the second as relations of the second as			no attributes; (2		-	1 5		e. Note that the
And Informed of group by's terrations approach of second as relations, we can apprend to the second of group by's terrations approach of second of the secon					•			As the results
As we will describe nanocubes is a first of the first of	All	1Phone		There				
A Star of the star		Patei Hubiction	results tatthe relation	CIPA TETA	three pations		ll descri	he nanocubas is
with we here the records in a state of the factor of the sector of the s			Central Central Val					f voll upg
es. stributes midbalbases and an and an and an an an and the set of the subsets of the subset of the subsets of the subset of the			An sense to the statest			and query	cubes of	rou ups.
es. gregation records. Using this house and the construction of the subsets of les ventional from a base schation allowed list of attributes and an assigned to the which all from a base schation allowed list of attributes and an assigned to the why its fundation the forward edition allowed in the structure are necessarily bounded by display.	ANTION PROMINENCE STATE							
Wentional ways of describing ageregations for a given relation. Synthetic are necessarily bounded by display, 2	tes. bribubes and holbase and		LONG KORCHYRKIEL CO	landladi	to set to be subled s		_	
be, its funded by the base relation and a second of the dataset that would end up in the same pixel for a great of the dataset that would end up in the same pixel for the formula of the provident of the dataset that would end up in the same pixel for the formula of the dataset that would end up in the same pixel for the formula of the dataset that would end up in the same pixel for the formula of the dataset that would end up in the same pixel for the formula of the dataset that would end up in the same pixel for the formula of the dataset that would end up in the same pixel for the formula of the dataset that would end up in the same pixel for the formula of the formula of the dataset that would end up in the same pixel for the formula of the dataset that would end up in the same pixel for the formula of the dataset that would end up in the same pixel for the formula of the same pixel for the dataset that would end up in the same pixel for the formula of the same pixel for the same pixel for the formula of the same pixel for the same pixel f	es derived from a base relati	on since a lis	t of wattributest and a	naggie	evice, Lang	guage}		
Dy, cupe, and rough is a lation and asked structure and up in the same nivel. In the same nivel with the correct of the base is lation and asked structure and up in the same nivel. In the same nivel with the same nivel of the same nivel of the same nivel.	the its fundation the force ways of description	Terent Gombh	ation a great rela	HOID gr	const uter a	re necessa	rily bour	nded by display
D. Koop, CSCI 640/490, Spring 2023 were a rist or attributes and up in the same nivel in the screen University	the by cupe, and for the base to	alfonandages	CITAS CITC IN WINCH		the to-be at	ole to quic	kly colle	ct subspaces of
	D. Koop, CSCI 640/490, Spring 2	2023 IVCH a HSt	orithe dataget that		nd up in the	e same nix	el on th	e screen Huniversity



ty

Building a Nanocube



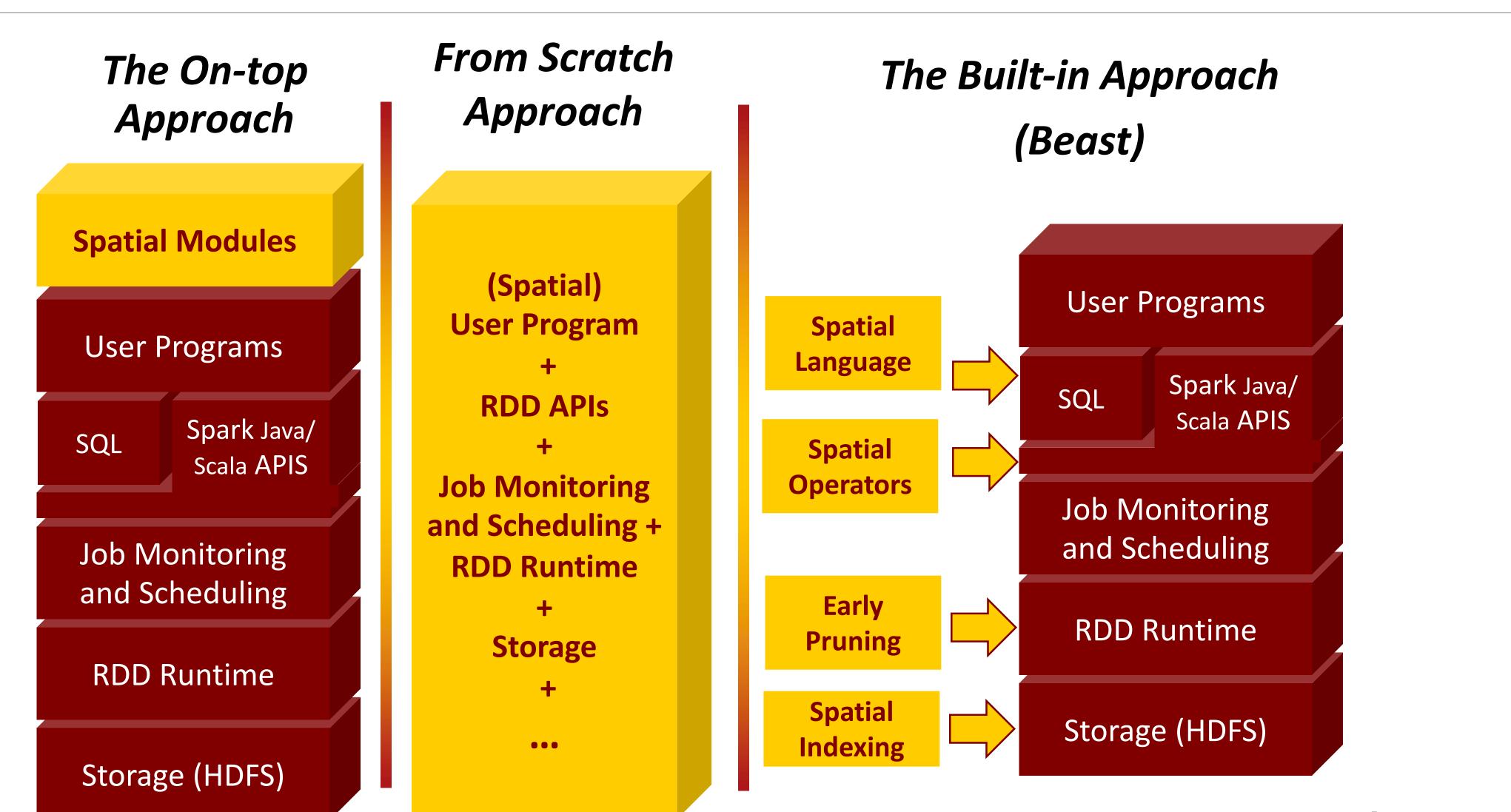








Beast Architecture



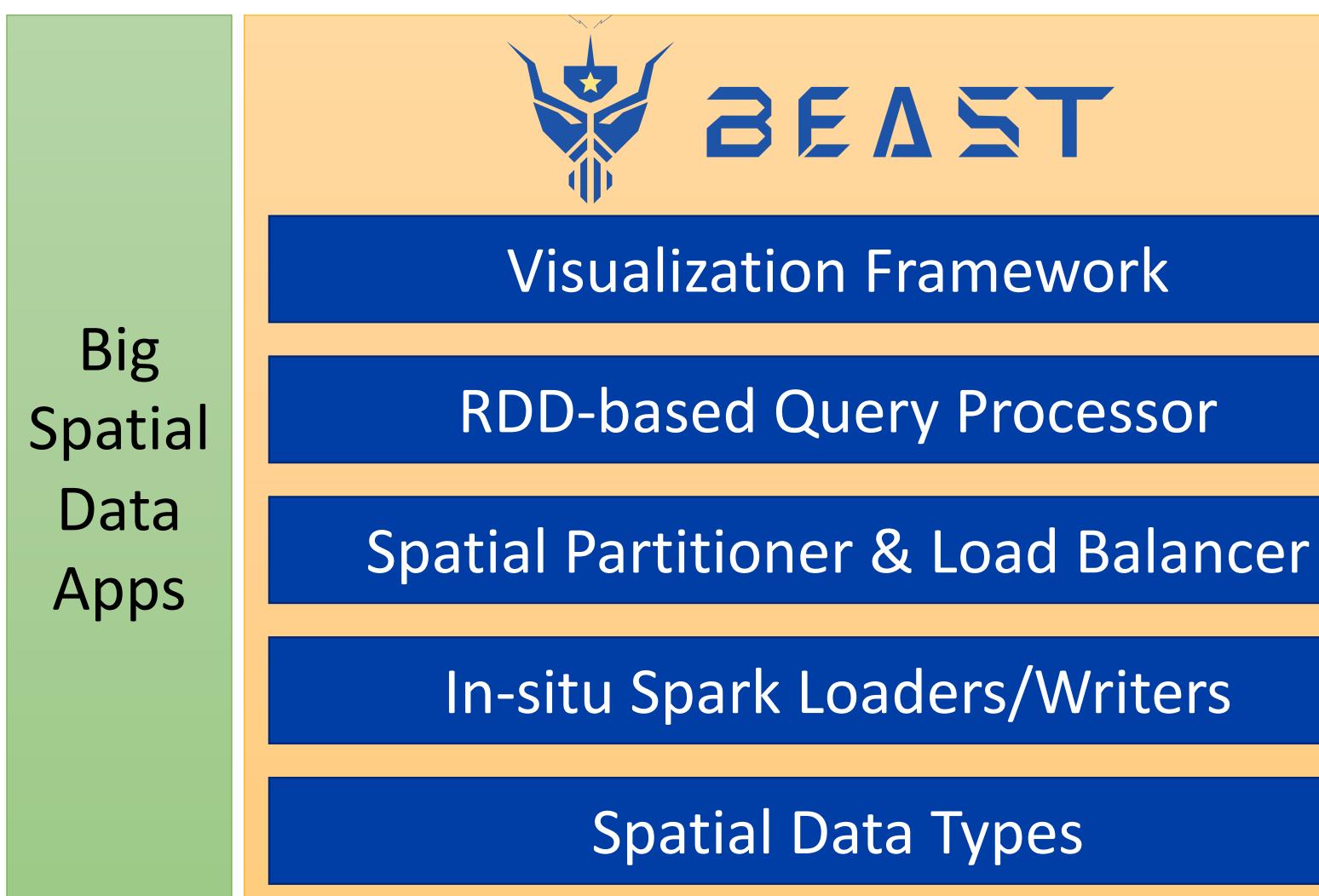
D. Koop, CSCI 640/490, Spring 2023





14

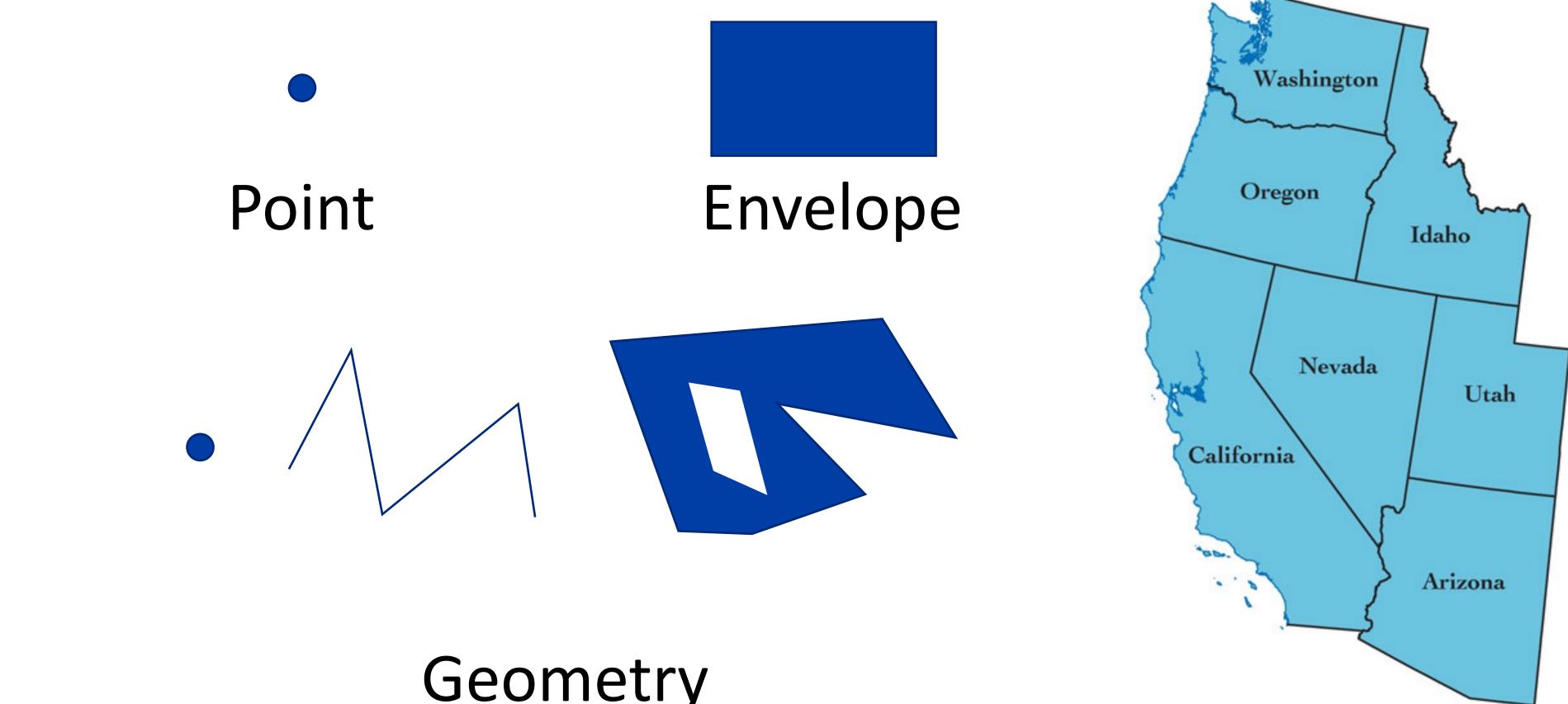
Beast Architecture







Beast Spatial Data Types



Geometry

D. Koop, CSCI 640/490, Spring 2023

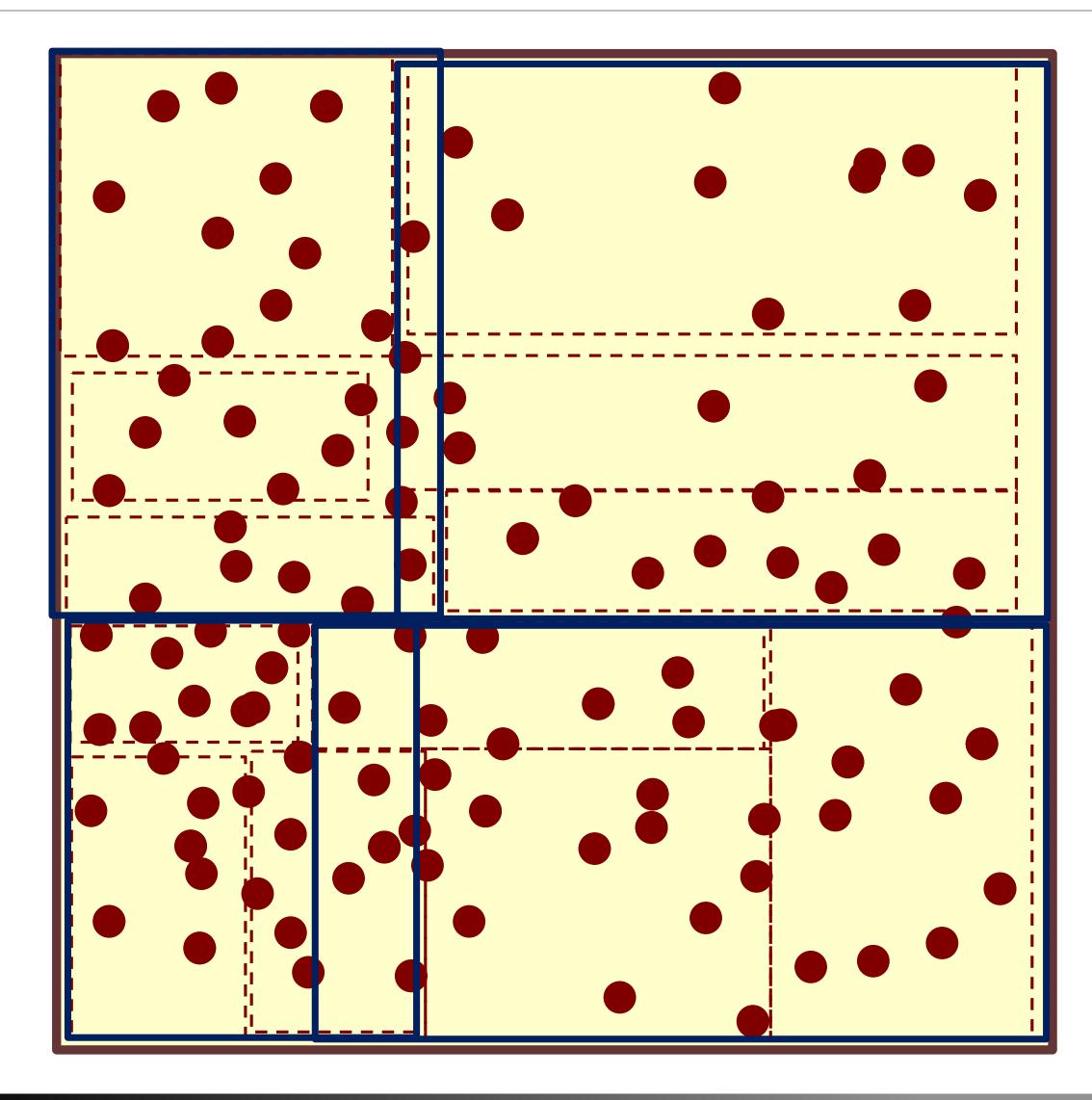
Feature

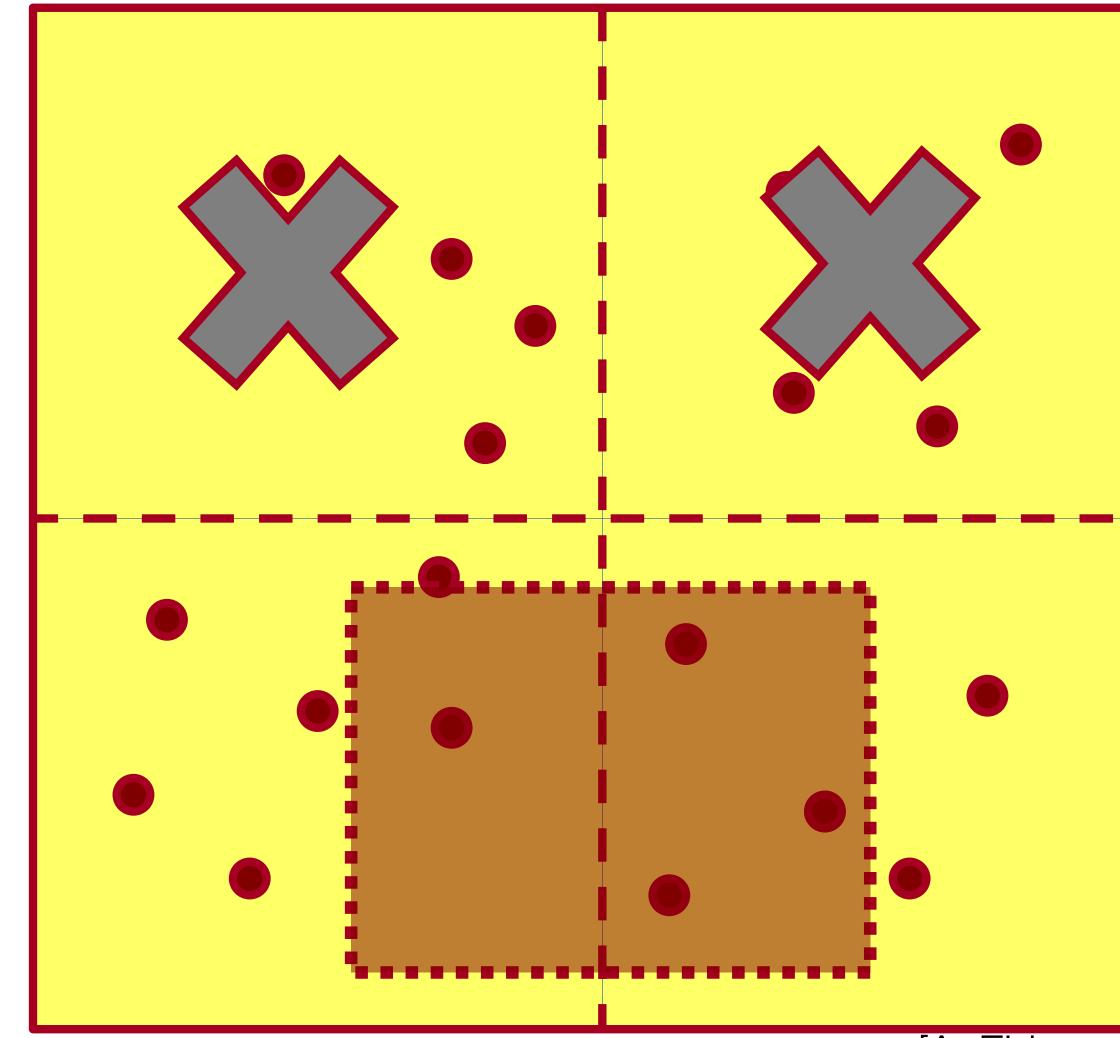




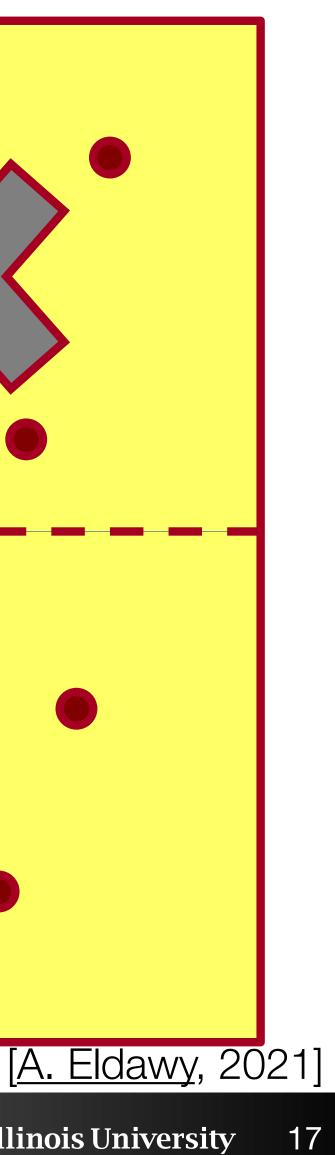
16

Beast Partitioning/Indexing & Range Query



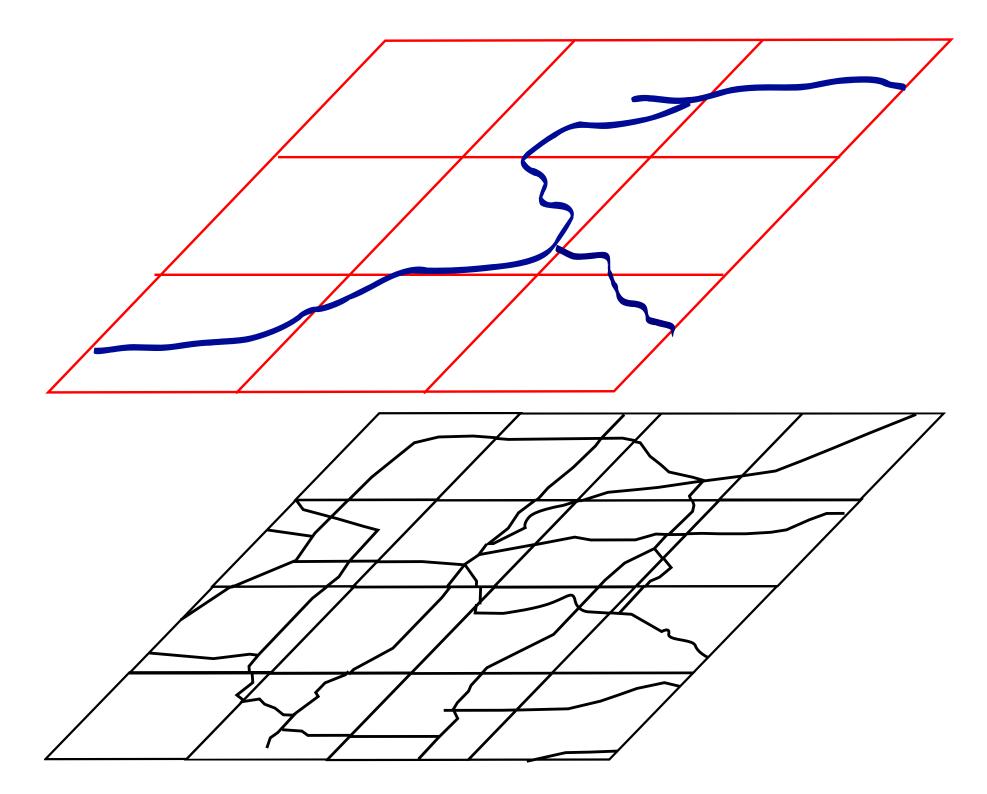






Beast Spatial Join

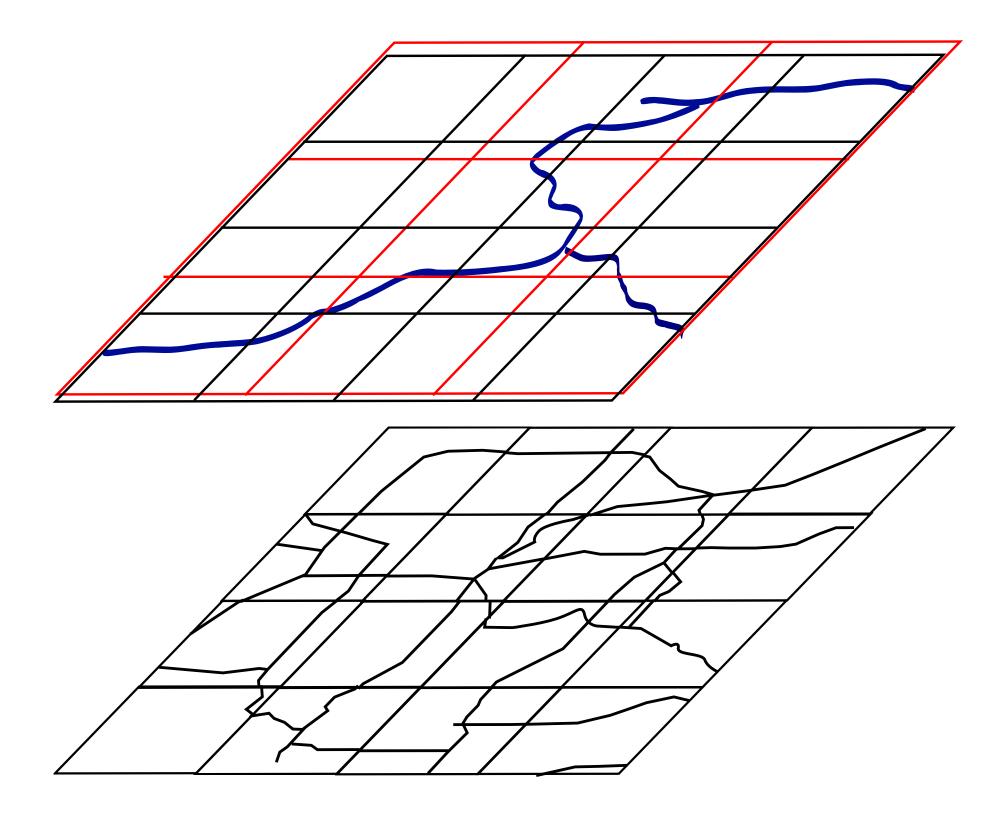
Join Directly



Total of 36 overlapping pairs

D. Koop, CSCI 640/490, Spring 2023

Partition – Join



Only 16 overlapping pairs





18

<u>Assignment 5</u>

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

D. Koop, CSCI 640/490, Spring 2023

Processing odin for extra credit)





D. Koop, CSCI 640/490, Spring 2023

Data Curation







D. Koop, CSCI 640/490, Spring 2023

Why?







Research Data Infrastructure

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]





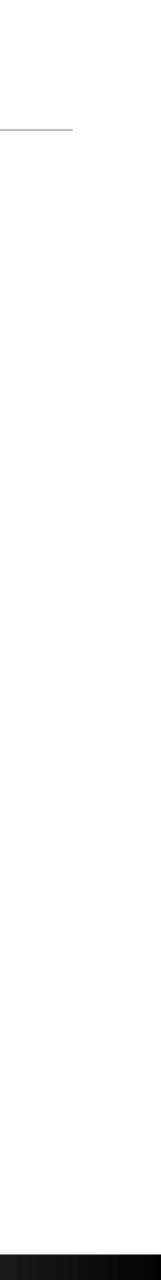




Data attribution and citation

- 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





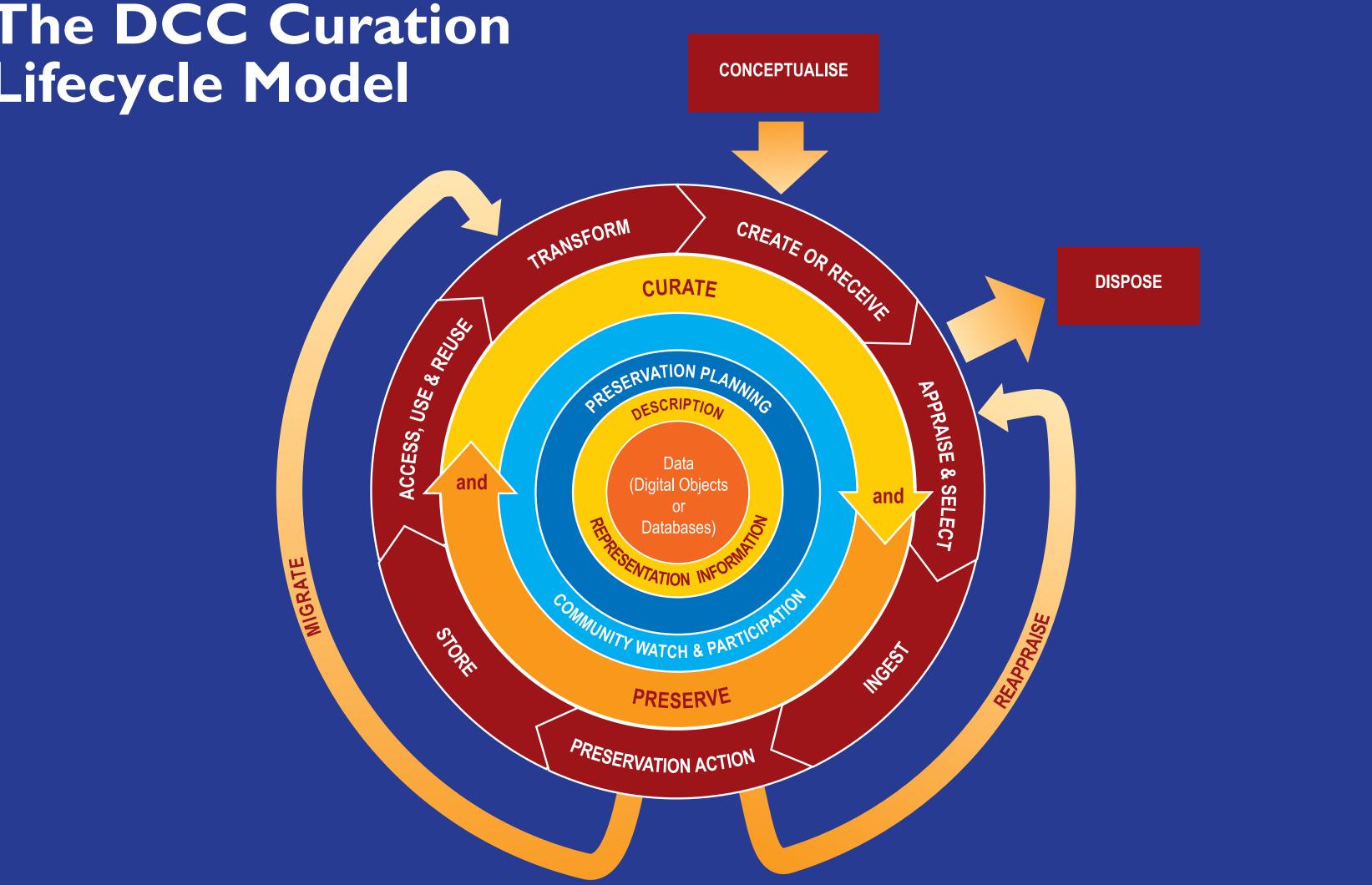






Data Curation Lifecycle

The DCC Curation Lifecycle Model













Data (Digital Objects or Databases)

- 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.

D. Koop, CSCI 640/490, Spring 2023

• Data, any information in binary digital form, is at the centre of the Curation









Full Lifecycle Actions

- Description and Representation Information: Assign metadata, using appropriate standards, to ensure adequate description and control
- of digital material
- Community Watch and Participation: Watch standards, tools, software.
- curation lifecycle

• Preservation Planning: Plan for preservation throughout the curation lifecycle

Curate and Preserve: Promote curation and preservation throughout the









Sequential Actions

- Create or Receive: Create/receive data and make sure metadata exists
- preservation
- 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

• Conceptualize: Plan creation of data—capture method and storage options. Appraise and Select: Evaluate data and select for long-term curation and

• Ingest: Transfer data to an archive, repository, data centre or other custodian

Transform: Create new data from the original (migrate formats, subsets, etc.)











Occasional Actions

- Dispose: Transfer to another archive or perhaps destroy data
- Reappraise: Return data which fails validation procedures for further appraisal and reelection
- hardware or software obsolescence

• Migrate: Migrate data to a different format—ensure the data's immunity from









The FAIR Guiding Principles for Scientific Data Management and Stewardship

M. D. Wilkinson et al.





Who and Why?

- 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

• Who: People from academia, industry, funding agencies, & scholarly publishers















FAIR Principles

- 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 welldescribed so they can be replicated and/or combined in different settings

• Findable: Metadata and data should be easy to find for both humans and











To be Findable

- 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

F1. (Meta)data are assigned a globally unique and persistent identifier









DataCite Workflow



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



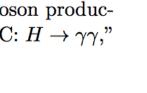










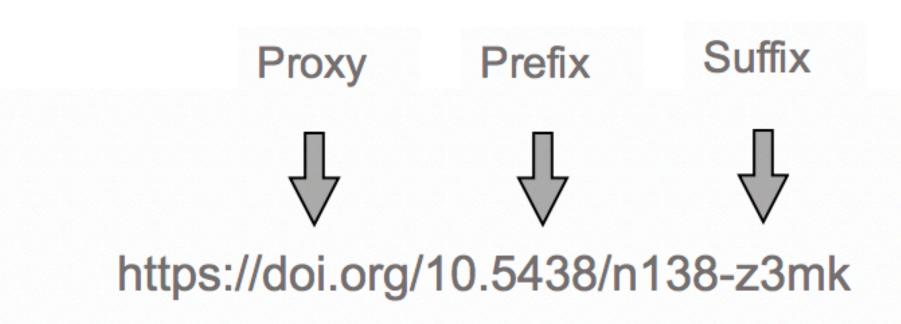






Digital Object Identifier

• Name: Proxy + Prefix + Suffix



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









DataCite Metadata

Mandatory Properties	Details
Identifier	with mandatory type sub-p
Creator	with optional name identifie
Title	with optional type sub-prop
Publisher	
PublicationYear	
ResourceType	with mandatory general typ

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

D. Koop, CSCI 640/490, Spring 2023

property

ier and affiliation sub-properties perties

pe description sub-property

and affiliation sub-properties

e sub-properties

gon sub-properties

Optional Properties

Language

AlternateIdentifier

Size

Format

Version

Rights

FundingReference









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**



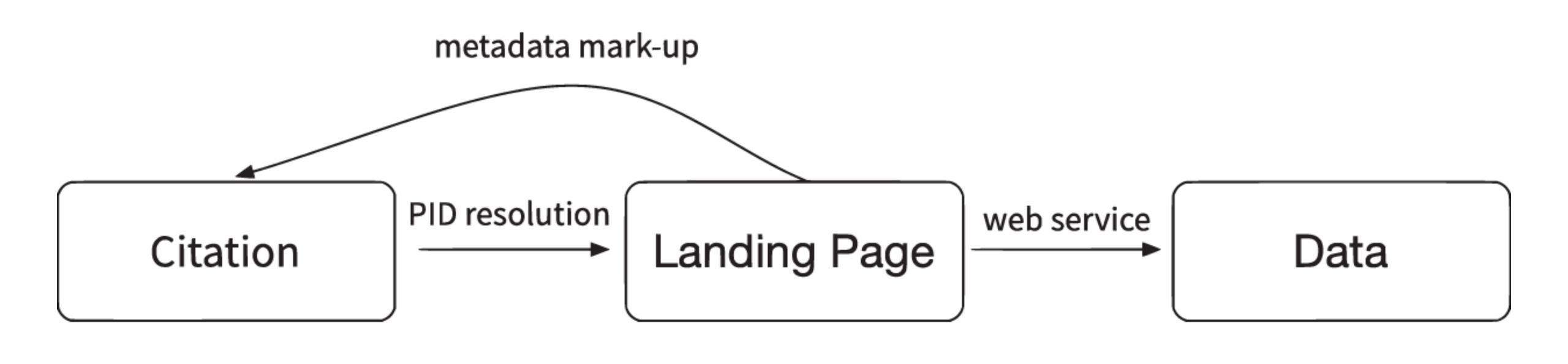








How data accessibility might work within publications



Document citing the data

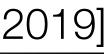
D. Koop, CSCI 640/490, Spring 2023

Repository housing the data

Data store







40

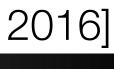
To be Interoperable

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

D. Koop, CSCI 640/490, Spring 2023







41

Standard vocabularies

View as Table	View as Grid				« 1 2	3	4 5	6	7		
ort by											
Name		\$	Registry	Name	Abbreviation		Туре	Subject	ł		
			њ.	ABA Adult Mouse Brain	ABA		Standard	 Neur 			
ecommended Records			888	Mouse Brain							
Recomme	ended			Access to Biological	ABCD		Standard	🗣 Biodi			
ssociated Publication?		_		Collection Data				🧳 Life S	Science		
No Publication	Has Publication										
Claimed?		_									
No Maintainer Record Status	Has Maintainer										
	n development Rea	ady		Access to Biological Collection Databases Extended for Geosciences	ABCDEFG		Standard	EarthPaleo			
Standard Type				Access to Biological	ABCDDNA		Standard	🕜 Biodi			
Terminology Artifact		771		Collection Data DNA extension					Science		
Model/Format		405									
Reporting Guideline		163									
Metric		30	E	.ACE format	.ACE format		Standard	🖌 Life S	Science		
Identifier Schema		15									
	Show	More	க்	AdaLab-meta ontology	ADALAB-MET	A	Standard	None			
Domains			Å	AdaLab ontology	ADALAB		Standard	None			
Report		141		Adverse Drug Reaction Markup	EU-ADR ML		Standard	None			
Data Transformation		134		Language							

					ocar		Rea	~	7 (61	vance												
	Showing records 1 - 50 of 1384 .																					
8	3 9 10 11 12 13 14 15 16						17	18	19	20	21 22 23			3 24 25 2			26 27 28 »					
Domain T							Taxono	my	Related Database					Standard	1	Related I	Policy		In Collection/Recommendation			
Brain Gene Expression Mus musculus Brain Imaging										None			None			None						
Ø Bio	Biology None			All		Repos GBIF Repos Canad Canad SiB C Colorr	sitory Spain IP sitory densys II densys F		⁻ Spain F Ƴ	ABCDDNA ABCDEFG			None			TDWG Biodiversity Information Standards						
_	Geology oil Science	No	ne				III		GeoC	ASe Dat	a Portal		XML ABCD		I	None		٢	Vone			R
	 Biology DNA Sequence Data Experiment Metadata Sequence Deoxyribonucleic Acid Polymerase Chain Reaction Plus 1 more 					I All		GenB	ank			MOD-CO ABCD			None			TDWG Biodiversity Information S			rds Dev	
	 DNA Sequence Data Contig Deoxyribonucleic Acid Genome 					III	✓ AII None					None			None			None			R	
		No	ine				🗣 All		None				None		l	None		٢	lone			R
		No	one				🖋 All		None				None		I	None		١	lone			R
			AdverseElectron	Reaction	ecord		Hom	o sapiens	None				XML			None		٢	lone			(<u>fair</u> :







To be Reusable

- 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

D. Koop, CSCI 640/490, Spring 2023

R1. (Meta)data are richly described with a plurality of accurate and relevant







43

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.

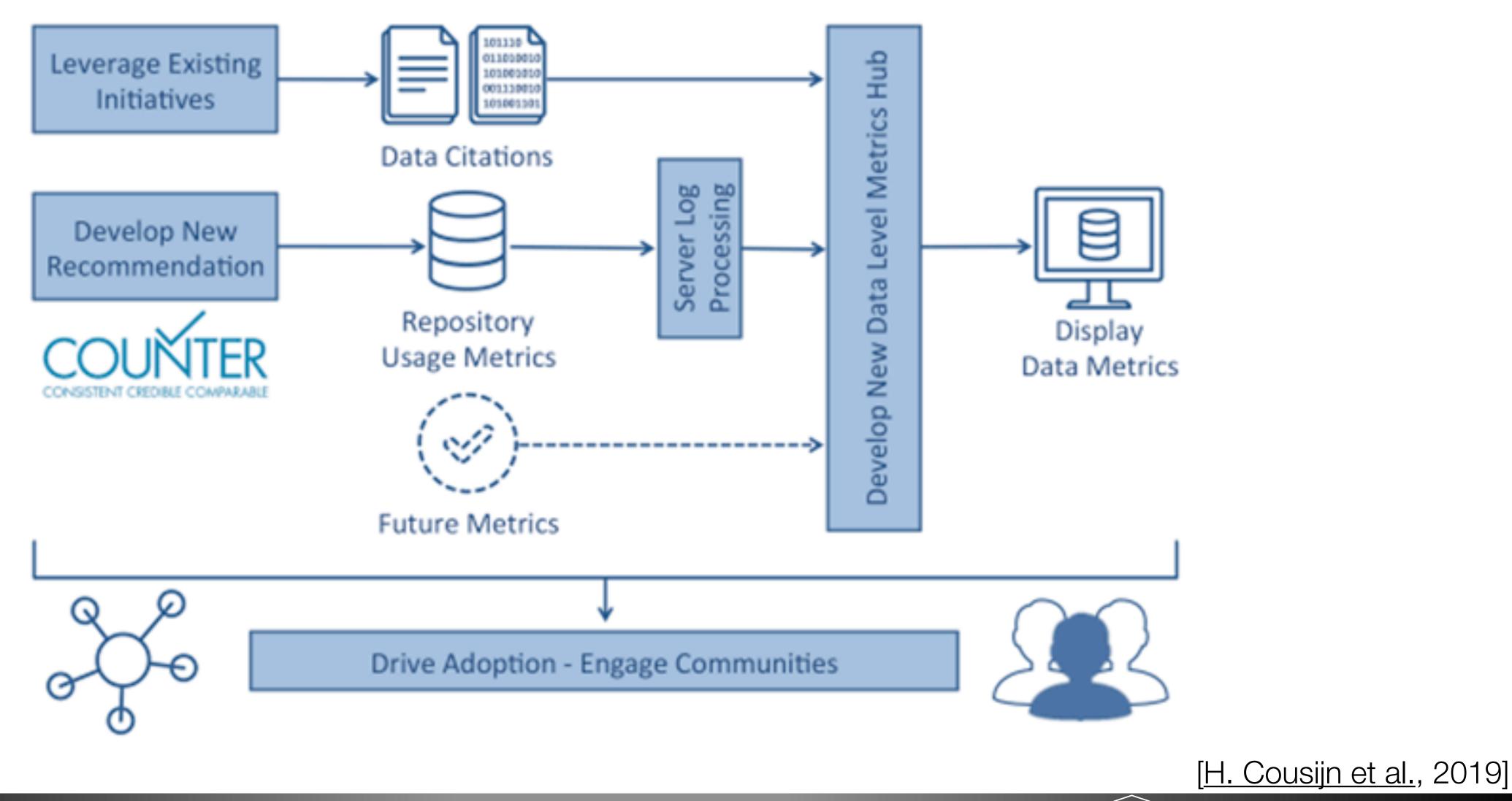








Make Data Count



D. Koop, CSCI 640/490, Spring 2023





45