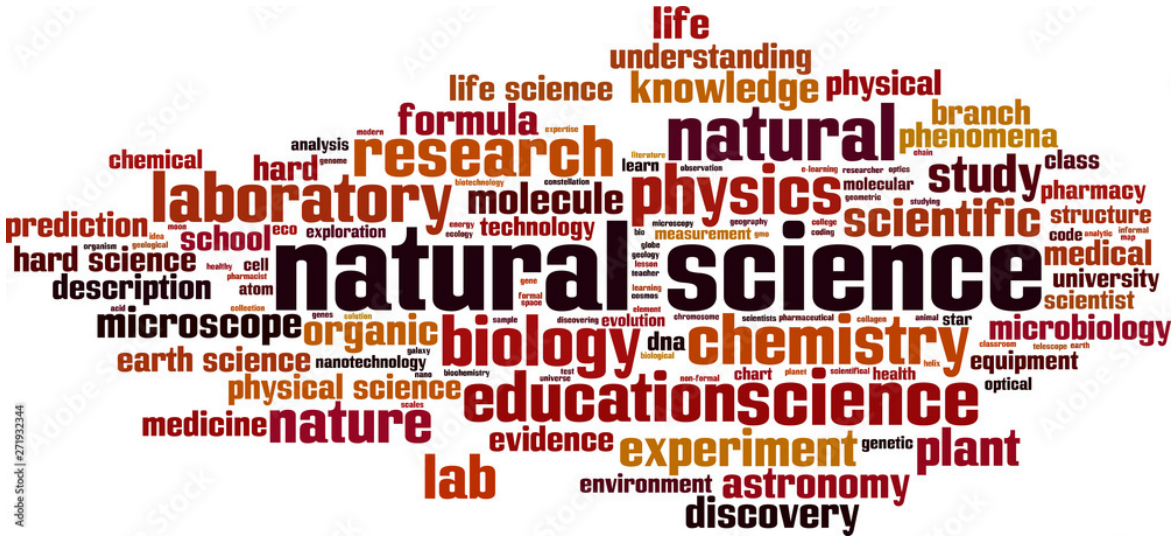
Knowledge Management for Scientific Application Using Polystore



Subhasis Dasgupta, Ph.D.

Assistant Scientist, University of California San Diego

sudasgupta@ucsd.edu

Knowledge Management Goals

There is a vast amount of data available from diverse sources, including various cohorts, demographics, and features.

In a perfect world, the data is expertly ingested, meticulously modeled, seamlessly indexed, and efficiently processed to guarantee effortless searching.

Efficient exploration is an absolute necessity for achieving faster innovation.

Developing effective knowledge management systems for your research group is essential and should be prioritized to prevent any potential delays or setbacks.

Choosing the appropriate management stack is necessary.

A typical ML workload

Working with machine learning data and selecting cohorts can be complex when using a file-based system.

In-memory DataFrame technologies such as Pandas or Dask can support cohort selection, but for large volumes of data, they may be inefficient regarding resource utilization.

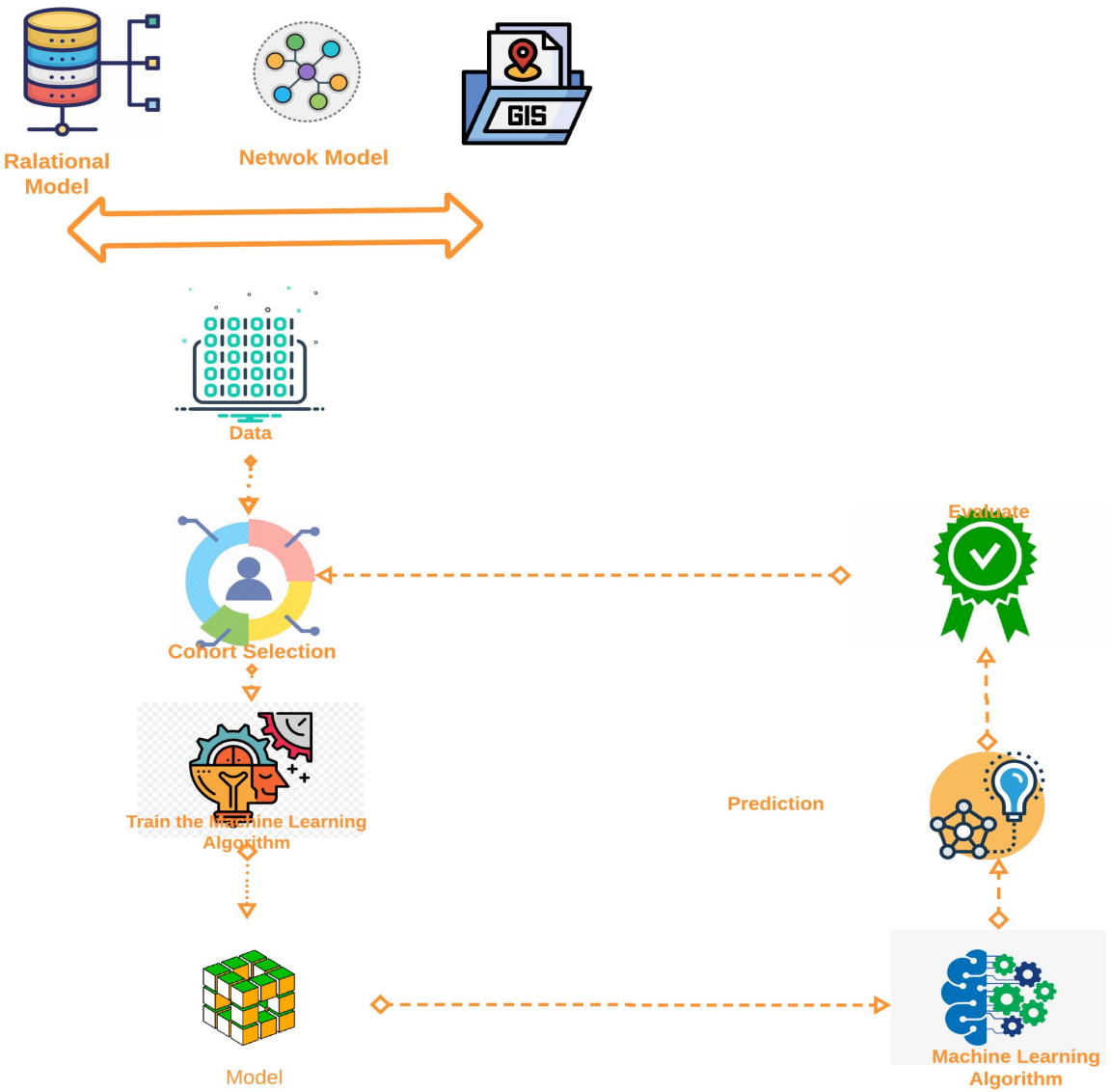
While using a database for data retrieval can be effective, traditional databases may not offer the level of flexibility that some situations require.

A typical ML workload

Data can come from different sources with varying structures, such as relational tables, temporal data, or network data. Creating a unified model can be very expensive or impossible.

It is essential to capture both direct and indirect relationships between entities. To validate data, researchers frequently rely on analytical and logical operations.

Establishing relationships and resolving entities require the use of semantic and schematic mapping. To ensure efficient and swift retrieval, it is crucial to implement appropriate indexing.



Why do you need knowledge?

- **Source Selection**
- **Schema alignment**
- **Entity Resolution**
- **Data Fusion**

StoreID	City	zip	productId	price
001	San Diego	92122	003	100
001	San Diego	92093	004	100
003	Ramona	92113	003	100

pid	batch	producer	date	zip	city
001	110100	AA Inc	02/20/2002	95370	Sonora
003	110102	AA Inc	02/20/2005	95370	Sonora
004	110100	BB Inc	02/20/2012	76950	Sonora

What is the process for effectively managing knowledge?

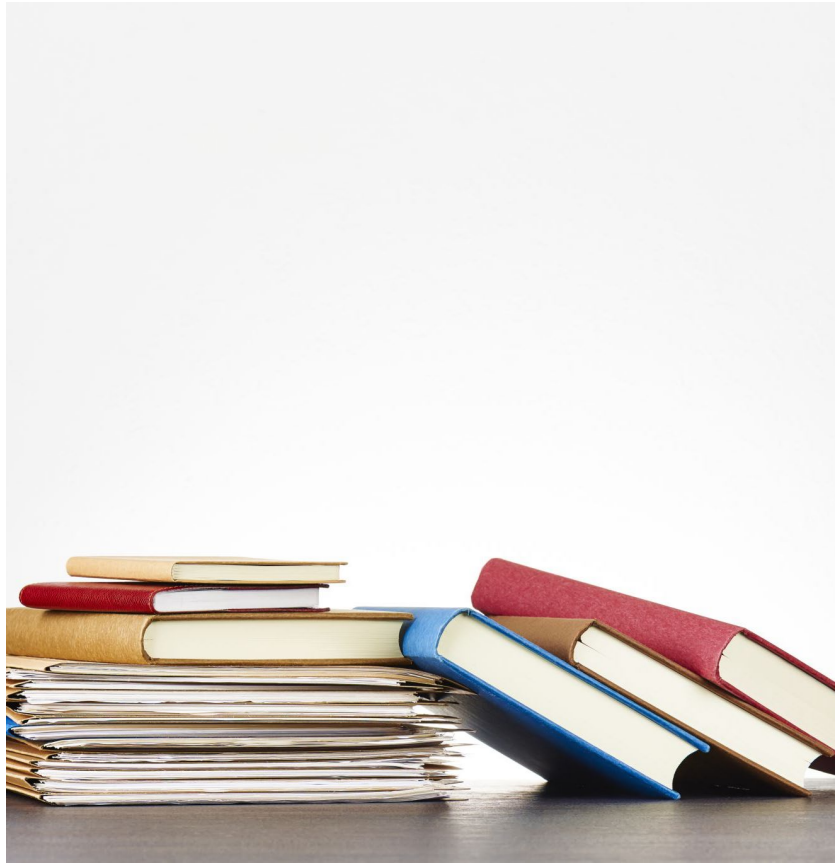
The roots of knowledge management can be traced back to a significant period in history.

- **Understanding Knowledge Management involves comprehending the process of knowing, the difference between knowledge and information, and information management.**

- **Documentalists in the early 20th century pioneered techniques that became the basis of today's knowledge management.**

- **In Europe and America in the first part of the twentieth century, documentalists had grand visions of collecting, codifying, and organizing the world's knowledge for world peace.**

- ***- Claire McInerney, Seminar in Information Studies, 2020 SCILS-Rutgers***



Knowledge Management Models

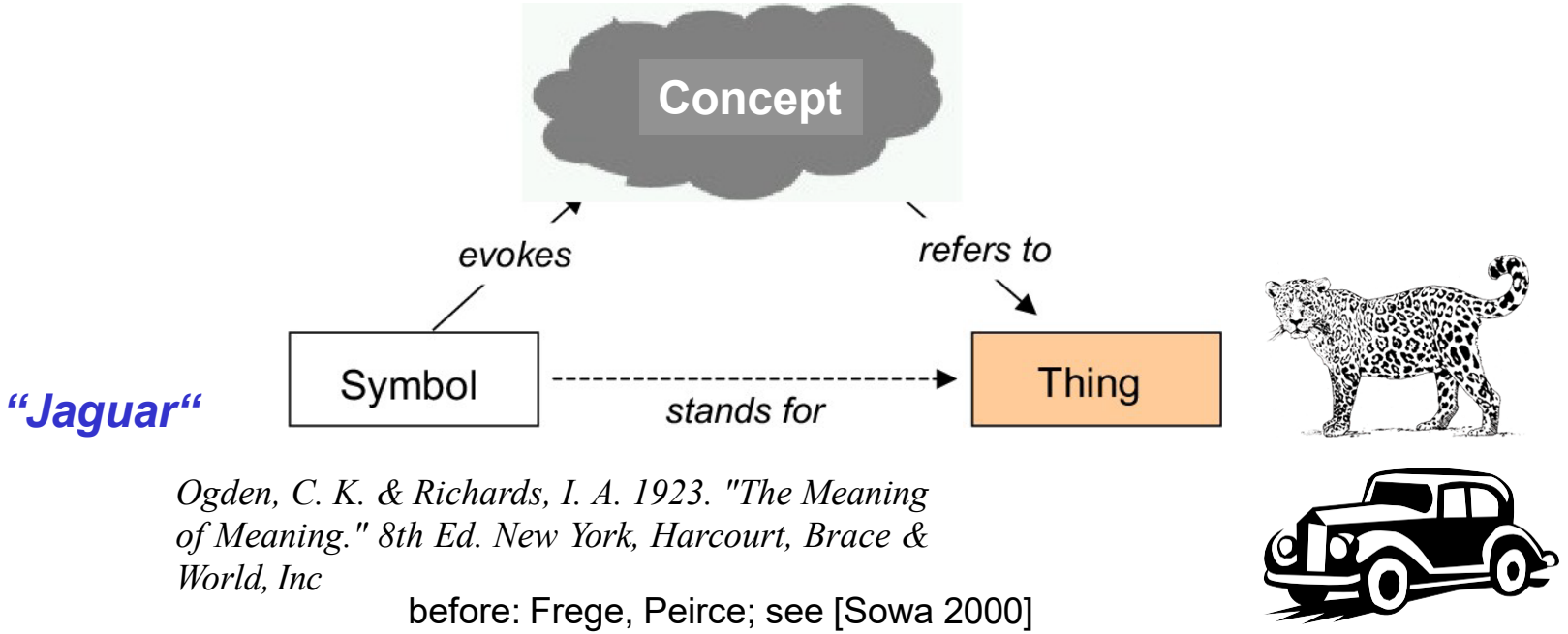
- Documentalist
- Technologist
- Communicator and Curator
- Lerner and Scholar

Ontology or Semantic Approach

- Problem:
  - Information retrieval and knowledge organization is a semantic and conceptual challenge because the formal record of scientific research is domain-specific and complex.
- Solution:
  - Apply ontologies to increase the efficiency of information retrieval and prioritization
  - Expand controlled vocabulary to normalize information extracted using systematic review methodology
  - Standardized data extraction formats for enhanced interoperability between systematic review tools and databases
  - Develop knowledge organization systems to enhance data curation and evidence integration frameworks

The Meaning Triangle

- Humans require words (or at least symbols) to communicate efficiently. The mapping of words to things is indirect. We do it by creating *concepts* that refer to things.
- The relation between symbols and things has been described in the form of the *meaning triangle*:



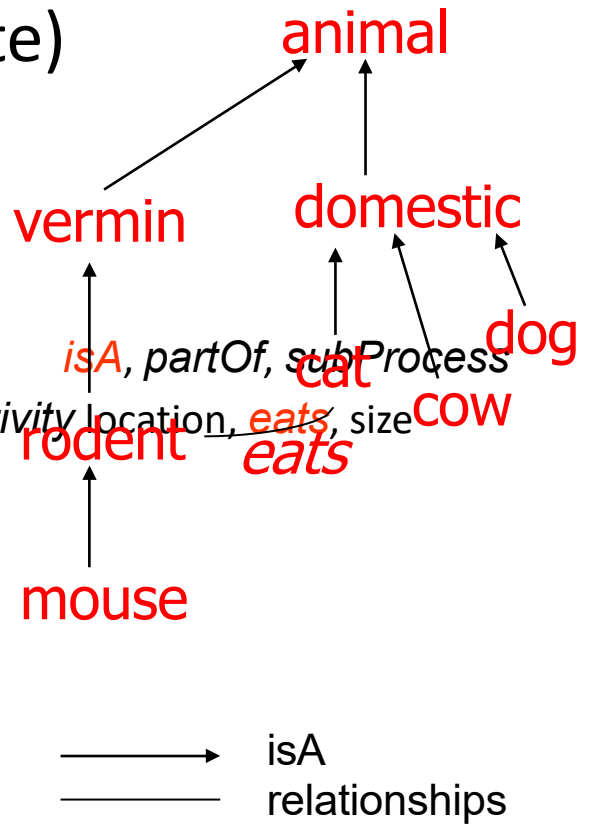
Ogden, C. K. & Richards, I. A. 1923. "The Meaning of Meaning." 8th Ed. New York, Harcourt, Brace & World, Inc

before: Frege, Peirce; see [Sowa 2000]

[Carole Goble, Nigel Shadbolt, Ontologies and the Grid Tutorial]

Ontology Modeling
An explicit description of a domain

- *Concepts* (**class**, set, type, predicate)
  - event, gene, gammaBurst, atrium, molecule, cat
- *Properties* of concepts and
- *relationships* between them (**slot**)
  - *Taxonomy*: generalisation ordering among concepts
  - *Relationship*, *Role* or *Attribute*: *functionOf*, *hasActivity* location, *eats*, size



[Carole Goble, Nigel Shadbolt, Ontologies and the Grid Tutorial]

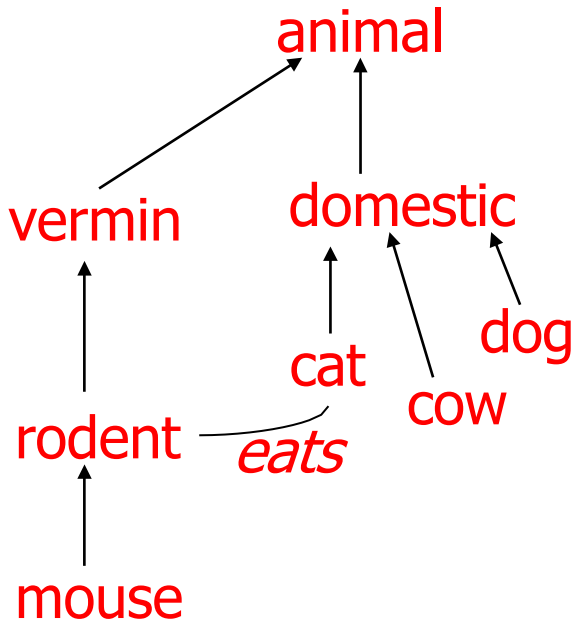
An explicit description of a domain

Constraints or *axioms* on properties and concepts:

- value: integer
- domain: cat
- cardinality: at most 1
- range: 0 <= X <= 100
- oligonucleiotides < 20 base pairs
- cows are larger than dogs
- cats cannot eat only vegetation
- cats and dogs are disjoint

Values or *concrete domains*

- integer, strings
- 20, trypotoplan-synthetase



[Carole Goble, Nigel Shadbolt, Ontologies and the Grid Tutorial]

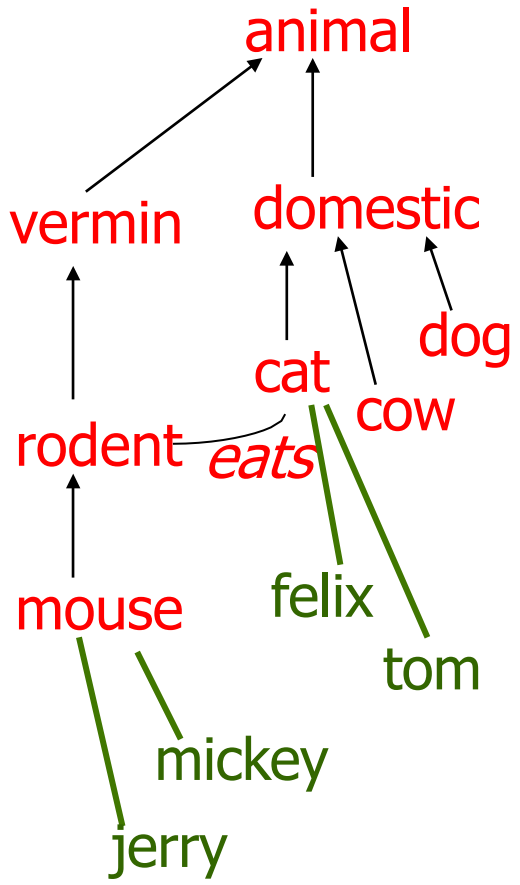
An explicit description of a domain

Individuals or *Instances*

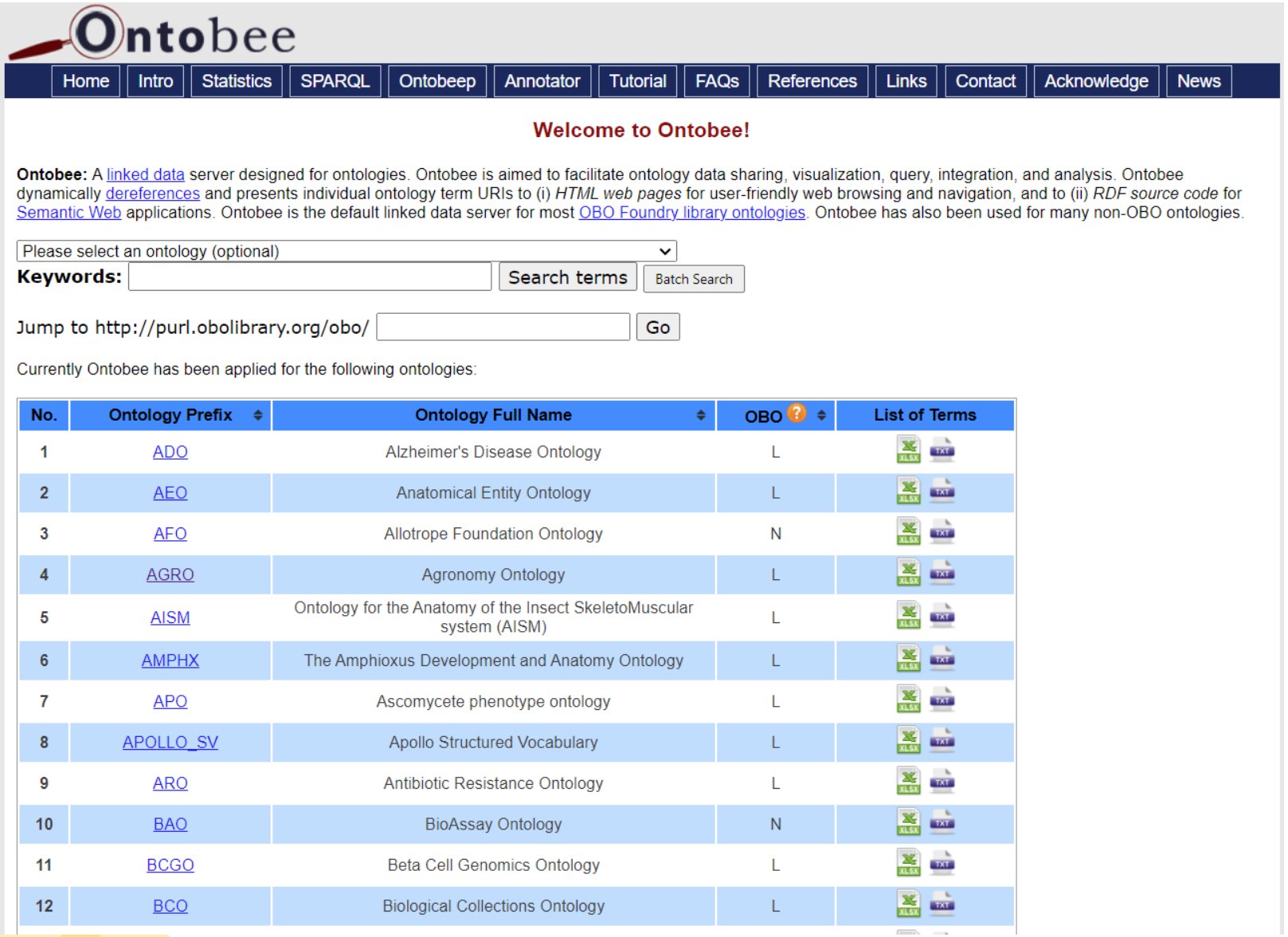
- sulphur, trpA Gene, **felix**

Ontology versus Knowledge Base

- An *ontology* = concepts+properties+axioms +values
- A *knowledge base* = ontology+instances



[Carole Goble, Nigel Shadbolt, Ontologies and the Grid Tutorial]



No.	Ontology Prefix	Ontology Full Name	OBO	List of Terms
1	ADO	Alzheimer's Disease Ontology	L	
2	AEO	Anatomical Entity Ontology	L	
3	AFO	Allotrope Foundation Ontology	N	
4	AGRO	Agronomy Ontology	L	
5	AISM	Ontology for the Anatomy of the Insect SkeletoMuscular system (AISM)	L	
6	AMPHX	The Amphioxus Development and Anatomy Ontology	L	
7	APO	Ascomycete phenotype ontology	L	
8	APOLLO_SV	Apollo Structured Vocabulary	L	
9	ARO	Antibiotic Resistance Ontology	L	
10	BAO	BioAssay Ontology	N	
11	BCGO	Beta Cell Genomics Ontology	L	
12	BCO	Biological Collections Ontology	L	


OPENLINK®
VIRTUOSO
UNIVERSAL SERVER

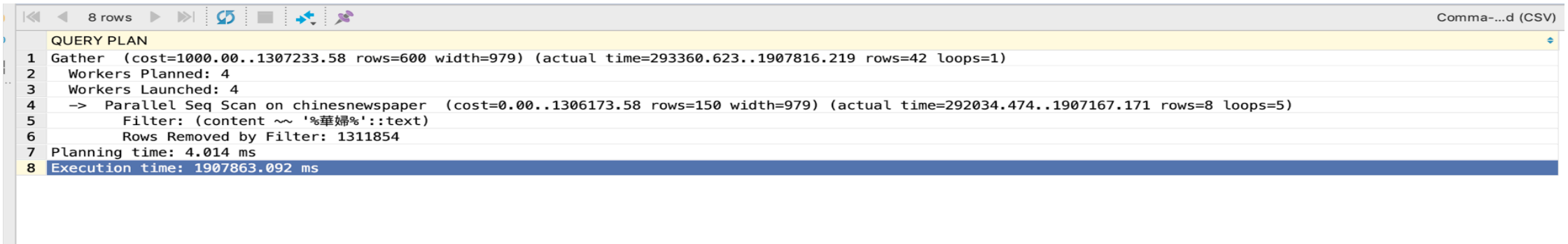

neo4j


protégé

“No One Size Fits All”

EXAMPLE OF LARGE QUERIES

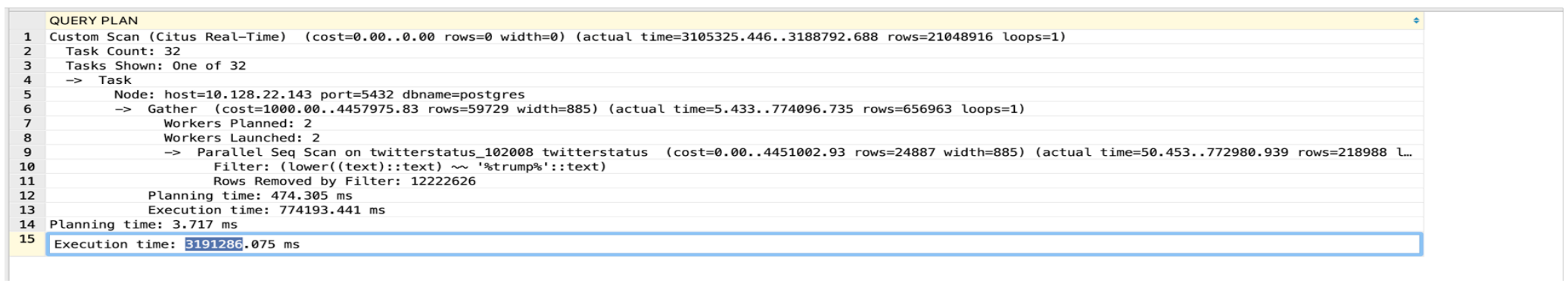
```
EXPLAIN ANALYZE SELECT * from chinesnewspaper WHERE content LIKE '%華婦%'
```



8 rows — Comma-...d (CSV)

```
QUERY PLAN
Gather  (cost=1000.00..1307233.58 rows=600 width=979) (actual time=293360.623..1907816.219 rows=42 loops=1)
  Workers Planned: 4
  Workers Launched: 4
  ->  Parallel Seq Scan on chinesnewspaper  (cost=0.00..1306173.58 rows=150 width=979) (actual time=292034.474..1907167.171 rows=8 loops=5)
        Filter: (content ~~ '%華婦%'::text)
        Rows Removed by Filter: 1311854
Planning time: 4.014 ms
Execution time: 1907863.092 ms
```

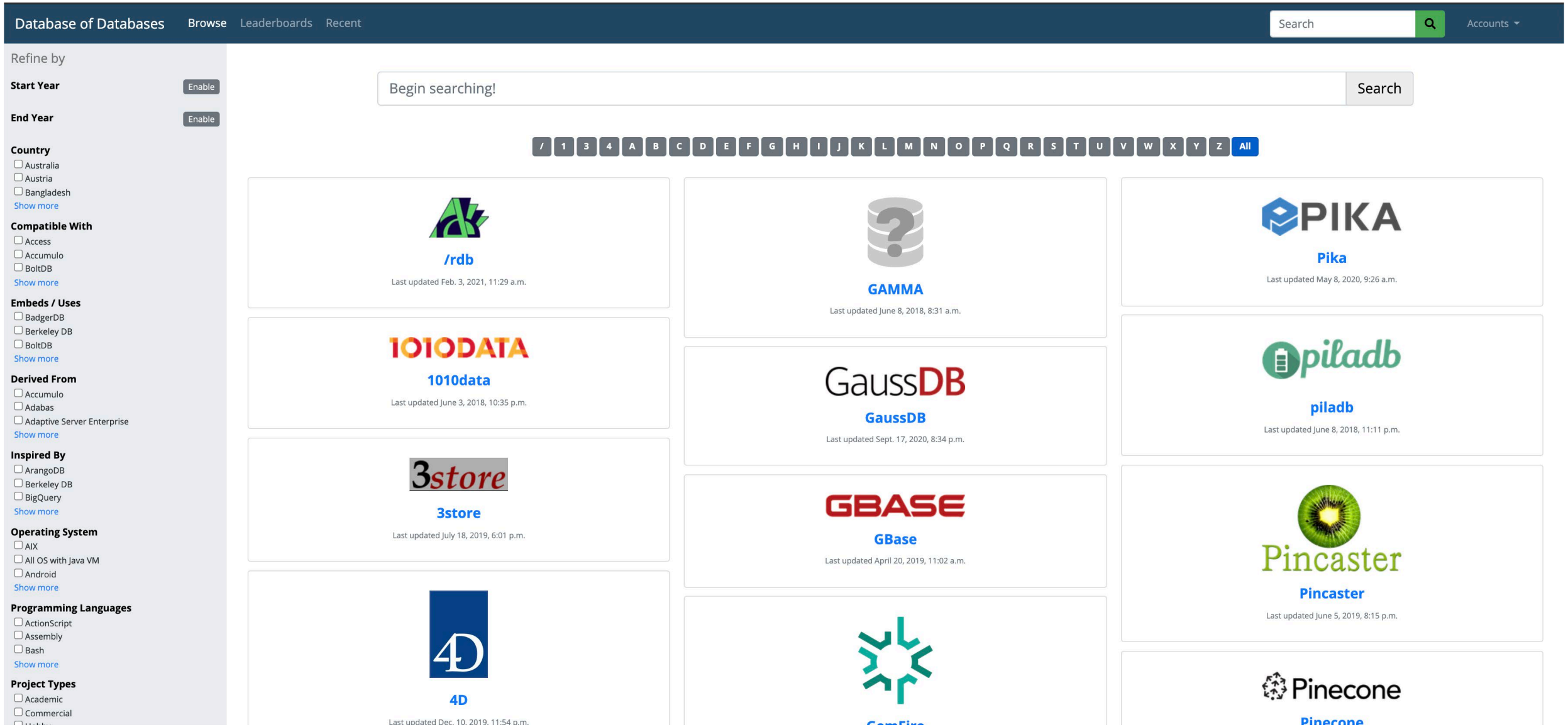
```
EXPLAIN ANALYZE SELECT * FROM twitterstatus where lower(text) Like '%trump%';
```



```
QUERY PLAN
Custom Scan (Citus Real-Time)  (cost=0.00..0.00 rows=0 width=0) (actual time=3105325.446..3188792.688 rows=21048916 loops=1)
  Task Count: 32
  Tasks Shown: One of 32
  ->  Task
        Node: host=10.128.22.143 port=5432 dbname=postgres
        ->  Gather  (cost=1000.00..4457975.83 rows=59729 width=885) (actual time=5.433..774096.735 rows=656963 loops=1)
              Workers Planned: 2
              Workers Launched: 2
              ->  Parallel Seq Scan on twitterstatus_102008 twitterstatus  (cost=0.00..4451002.93 rows=24887 width=885) (actual time=50.453..772980.939 rows=218988 l…
                    Filter: (lower((text)::text) ~~ '%trump%'::text)
                    Rows Removed by Filter: 12222626
            Planning time: 474.305 ms
            Execution time: 774193.441 ms
Planning time: 3.717 ms
Execution time: 3191286.075 ms
```


Where do I find a database of databases?

www.dbdb.io


Database of Databases
Browse
Leaderboards
Recent
Search
Accounts
Refine by
Start Year
Enable
End Year
Enable
Country
Australia
Austria
Bangladesh
Show more
Compatible With
Access
Accumulo
BoltDB
Show more
Embeds / Uses
BadgerDB
Berkeley DB
BoltDB
Show more
Derived From
Accumulo
Adabas
Adaptive Server Enterprise
Show more
Inspired By
ArangoDB
Berkeley DB
BigQuery
Show more
Operating System
AIX
All OS with Java VM
Android
Show more
Programming Languages
ActionScript
Assembly
Bash
Show more
Project Types
Academic
Commercial
Begin searching!
Search
/ 1 3 4 A B C D E F G H I J K L M N O P Q R S T U V W X Y Z All
/rdb
Last updated Feb. 3, 2021, 11:29 a.m.
1010data
Last updated June 3, 2018, 10:35 p.m.
3store
Last updated July 18, 2019, 6:01 p.m.
4D
Last updated Dec. 10, 2019, 11:54 p.m.
GAMMA
Last updated June 8, 2018, 8:31 a.m.
GaussDB
Last updated Sept. 17, 2020, 8:34 p.m.
GBase
Last updated April 20, 2019, 11:02 a.m.
Pika
Last updated May 8, 2020, 9:26 a.m.
piladb
Last updated June 8, 2018, 11:11 p.m.
Pincaster
Last updated June 5, 2019, 8:15 p.m.
Pinecone

Model Specific Databases

A Few Model Specific Databases

Relational DB

Graph DB

Search DB

Semi-structured DB

Timeserise DB

Spatial DB

Model Specific Databases

Key Value Storage

RocksDB

redis

Cassandra

Data Processing Platform and dataframe technology

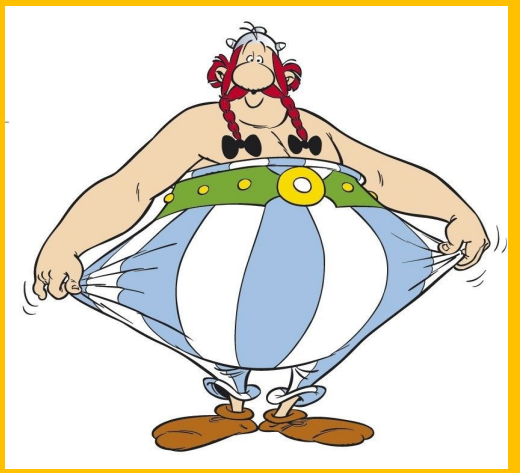
APACHE Spark

Flink 1.0

AWS RDS

DASK

And Many...



Why so many models and apps?

- In recent years, the database community has developed numerous applications and techniques to handle different models and capabilities, including relational, semi-structured, and network models, inverted index, data cube (group by, cube by), and centrality computation.
- These technological advancements were developed to tackle particular challenges in various industries or fields.
- Each app is developed and tuned best for its model and capabilities but incapable of the others.

- Relational Structure
- SQL
- Cube and Group queries
- Text search (Gin, GIST)
- Network queries
- Centrality computation

- Relational Structure
- Cypher
- Cube and Group queries
- Text search (Gin, GIST)
- Network queries
- Network analytics

- Text search
- Network queries
- Network analytics(Centrality, cluster, pagerank)

Design Goals of a Polystore Systems

- **Polystore should support location transparency** like federated databases (i.e., common query language).
- **Semantic Completeness:** The user will not lose any capabilities its underlying storage engine provides.
- **Object Version Consistency:** The same version of the object should be available in multiple models.
- **Capability-based Optimization:** Optimize the analytical computation depending on the capabilities.

Architectural Variations

Loosely Coupled

1. Cross model mediator-based design, each provider will have a dedicated mediator to communicate with other providers.
2. Local storage has more control over the data, and the global controller maintains consistency and transparency.
3. Local operations are efficient, but model transformation cost is high.
4. Challenging to optimize analytical operations and create cross-model materialized view and cross-model index.

Tightly Coupled

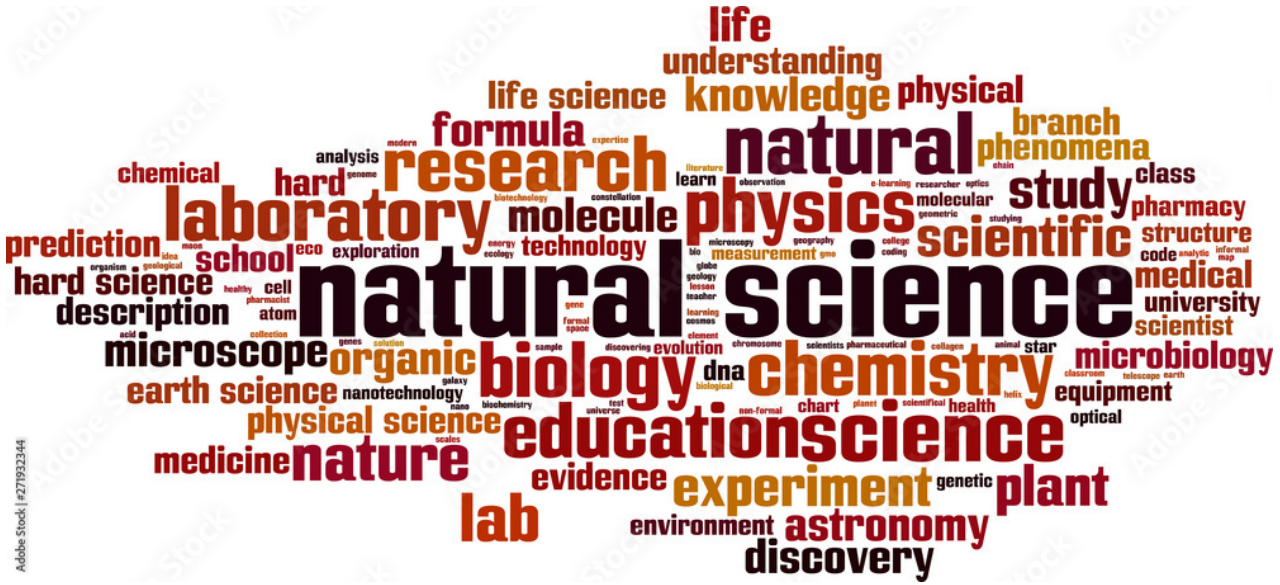
1. Use a common interface to interact among stores, like a standard data frame or data structure for the whole polystore.
2. Local storage has less control over the data, and the central controller decides everything.
3. Transforming or rewriting queries from one store to another store is complex and ultimately boils down to a multi-query optimization problem.
4. It is hard to optimize the best plan for each store. The optimization cost is very high.
5. Easy to build a materialized view and index.

Hybrid

1. Trade-off between global control and local control
2. Very efficient for optimizing queries for each local storage.
3. Easy and efficient use of materialized view is possible.
4. Very much domain or vertical specific.

Example Polystore Systems

BigDAWG, MIT	CloudMdsQL, Inria	Estocada, UCSD and Inria	Polypheny-DB, University of Basel, Switzerland
Awesome, UCSD(*)	Polystore++ , Stanford University	Polybase, Microsoft.	OoX, HP-Labs



The Awesome Polystore

Data Variety

Social Network

Newspaper

Contract and Legal

Clinical Trial Data

Organizational Data

SEMISTRUCTURED, VIDEO, TEXT

STRUCTURED TEXT, TIME, PLACES

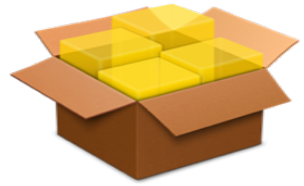
TEXT

TEXT, IMAGE

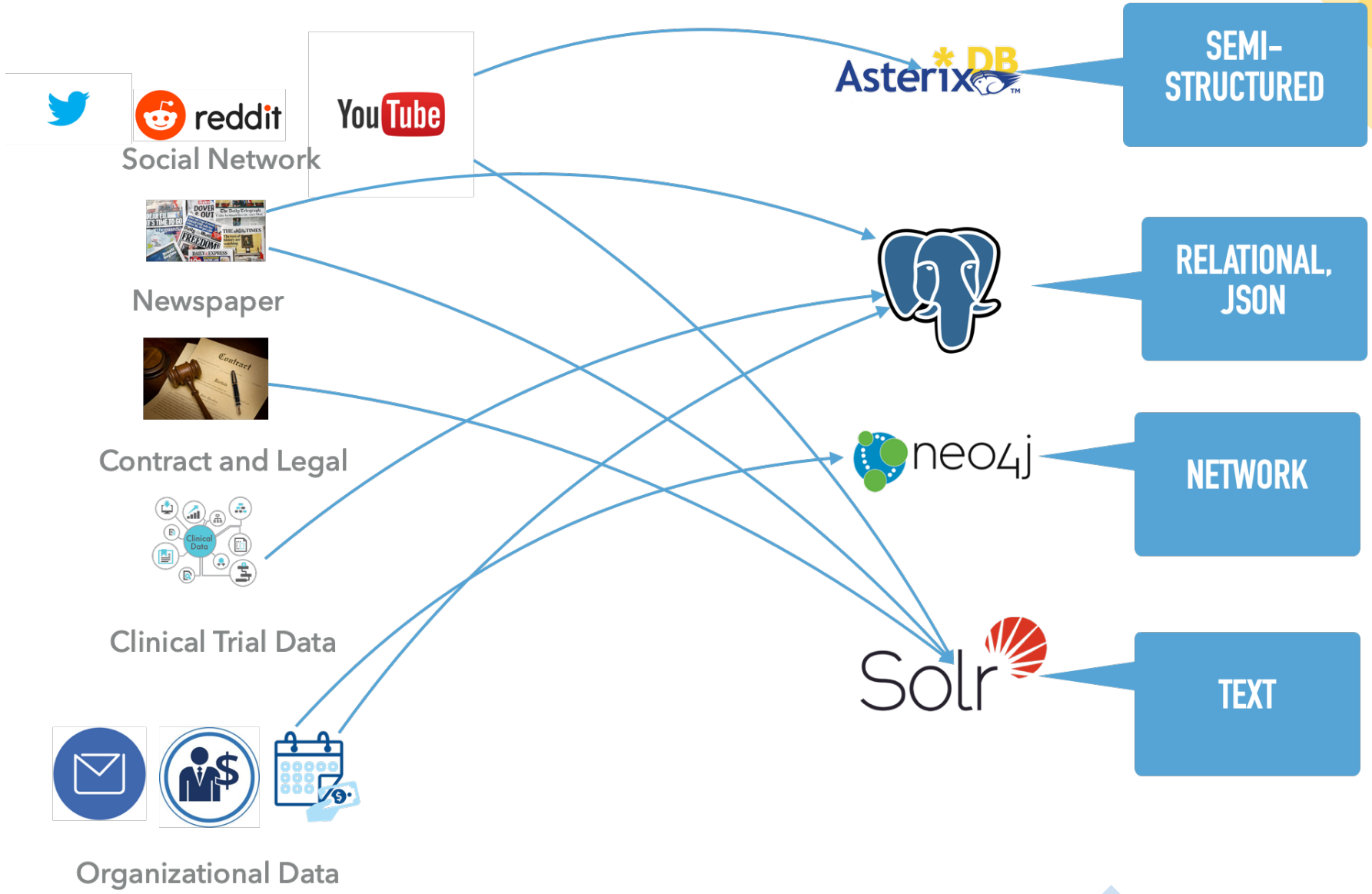
EMAIL, RELATIONAL TABLES, SAP DATA, TEXT, CUSTOMER REVIEW DATA

Dictionaries

HITL



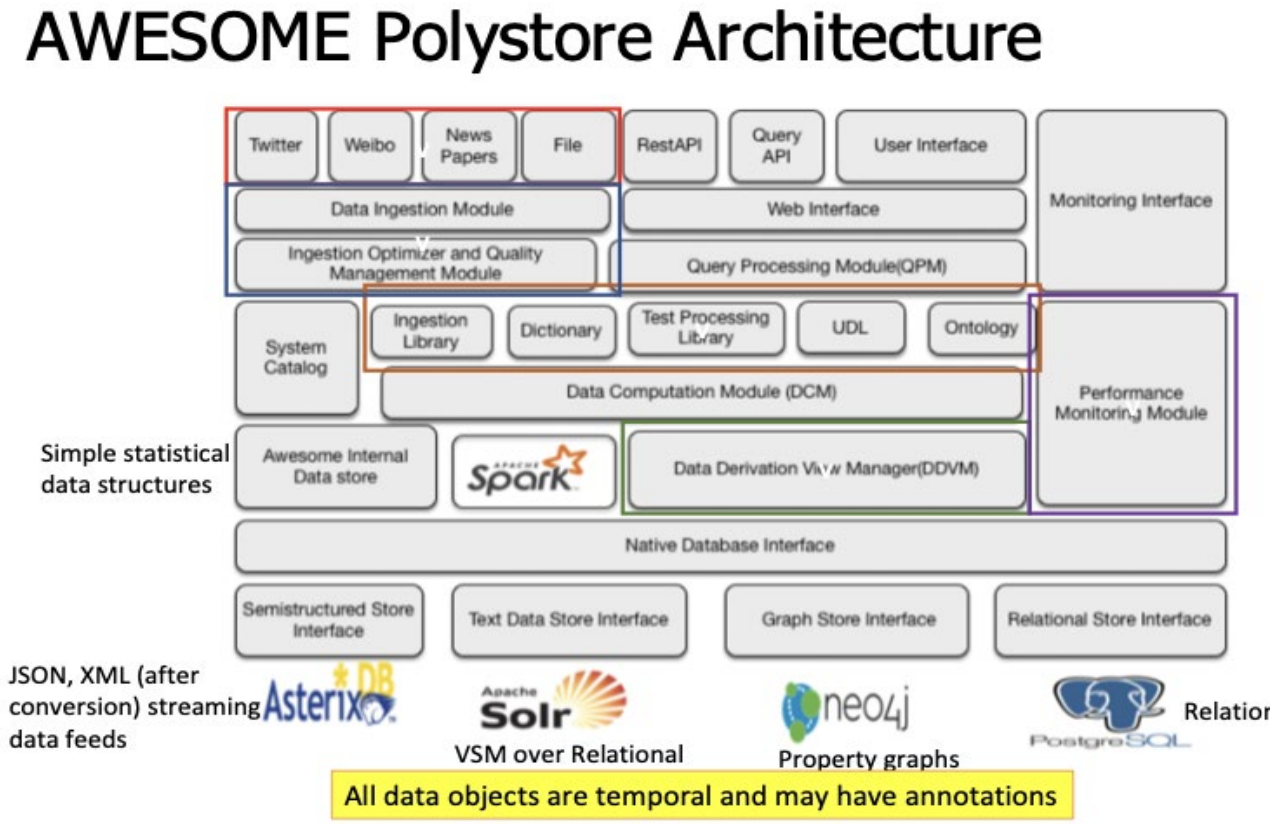
Packages


reddit
YouTube
Social Network
Newspaper
Contract and Legal
Clinical Trial Data
Organizational Data
AsterixDB
SEMI-STRUCTURED
RELATIONAL, JSON
neo4j
NETWORK
Solr
TEXT

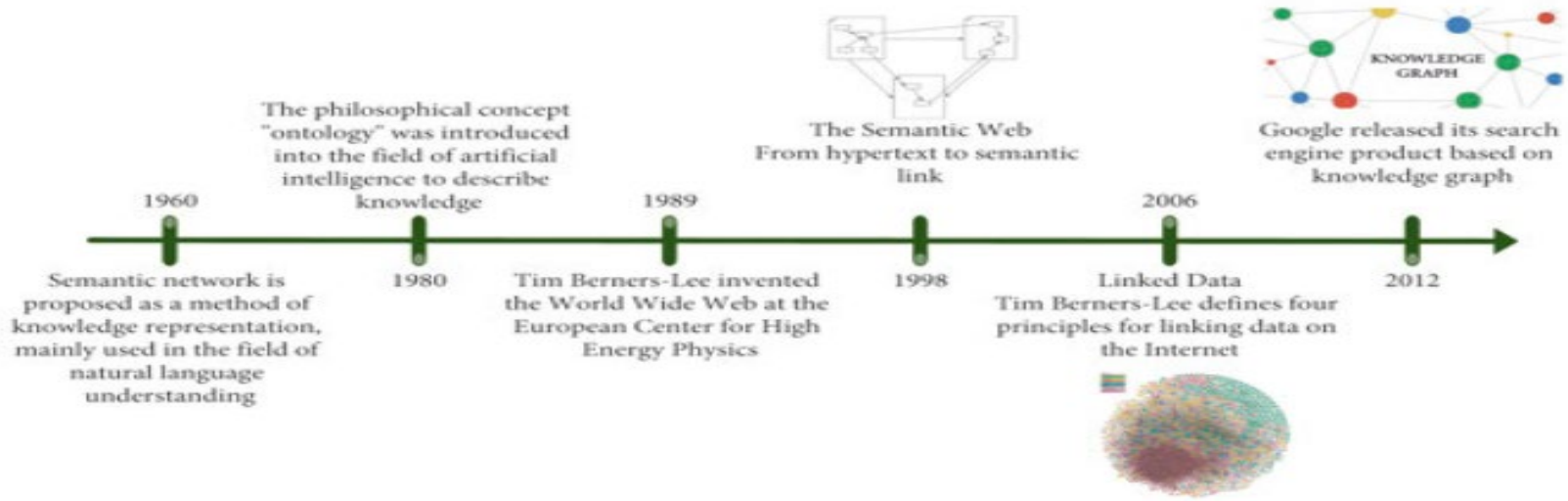
- Social Sciences Questionnaire:
  - List all accounts talking about "Elizabeth Warren" and "American DOS Movements.“
  - Find the top 100 influential co-spiking users and discuss racial terms from the “Elizabeth Warren’s” network.
  - Find out top-k topics from the newspaper that are also discussed in the “Elizabeth Warren’s” network.
  - Top K-topics discussed in the network but not covered by the newspaper.

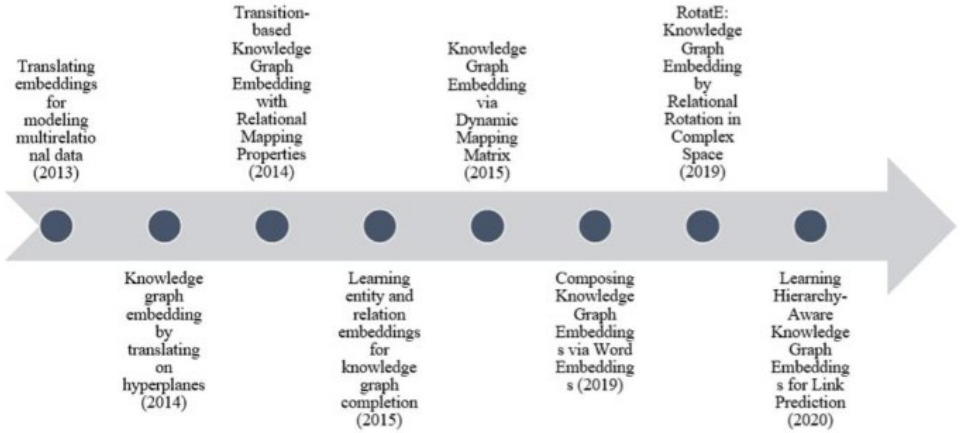
Summary of Awesome Architecture

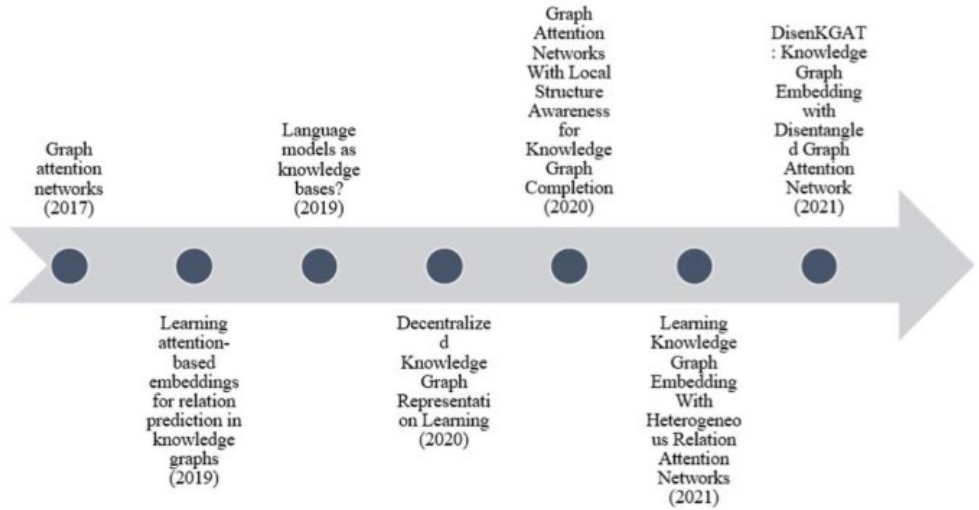
- AWESOME integrates information over heterogeneous data
  - A relational DBMS
  - A graph DBMS
  - A document/semi-structured DBMS
  - A text search engine
  - Vector and Matrix data from Analytics engines



The Knowledge Graph Innovation Timeline




Translating embeddings for modeling multirelational data (2013)
Knowledge graph embedding by translating on hyperplanes (2014)
Transition-based Knowledge Graph Embedding with Relational Mapping Properties (2014)
Learning entity and relation embeddings for knowledge graph completion (2015)
Knowledge Graph Embedding via Dynamic Mapping Matrix (2015)
Composing Knowledge Graph Embeddings via Word Embeddings (2019)
RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space (2019)
Learning Hierarchy-Aware Knowledge Graph Embeddings for Link Prediction (2020)


Graph attention networks (2017)
Learning attention-based embeddings for relation prediction in knowledge graphs (2019)
Language models as knowledge bases? (2019)
Decentralized Knowledge Graph Representation Learning (2020)
Graph Attention Networks With Local Structure Awareness for Knowledge Graph Completion (2020)
Learning Knowledge Graph Embedding With Heterogeneous Relation Attention Networks (2021)
DisenKGAT: Knowledge Graph Embedding with Disentangled Graph Attention Network (2021)

Building a Knowledge Graph on the top of a polystore

The Problem

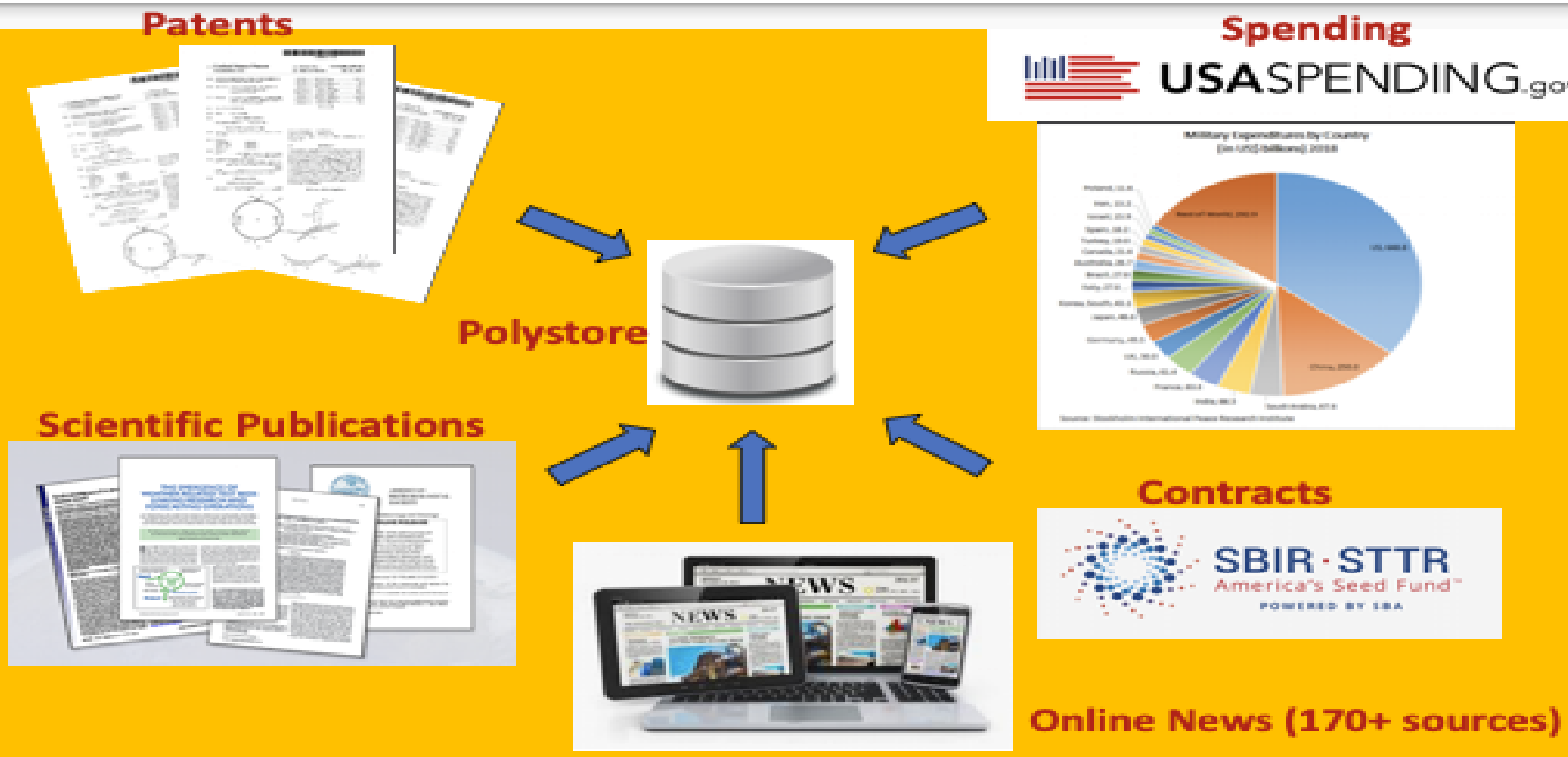
"Discover technology gaps in some domain and who can bridge the gaps?"

Data Set : all publicly available data

Solution Approach

- Create a Knowledge graph by assimilating information from multiple data sources.
- Search term expansion and association mining using KG
- Gap discovery using network query
- Potential partnership determination using Cube query

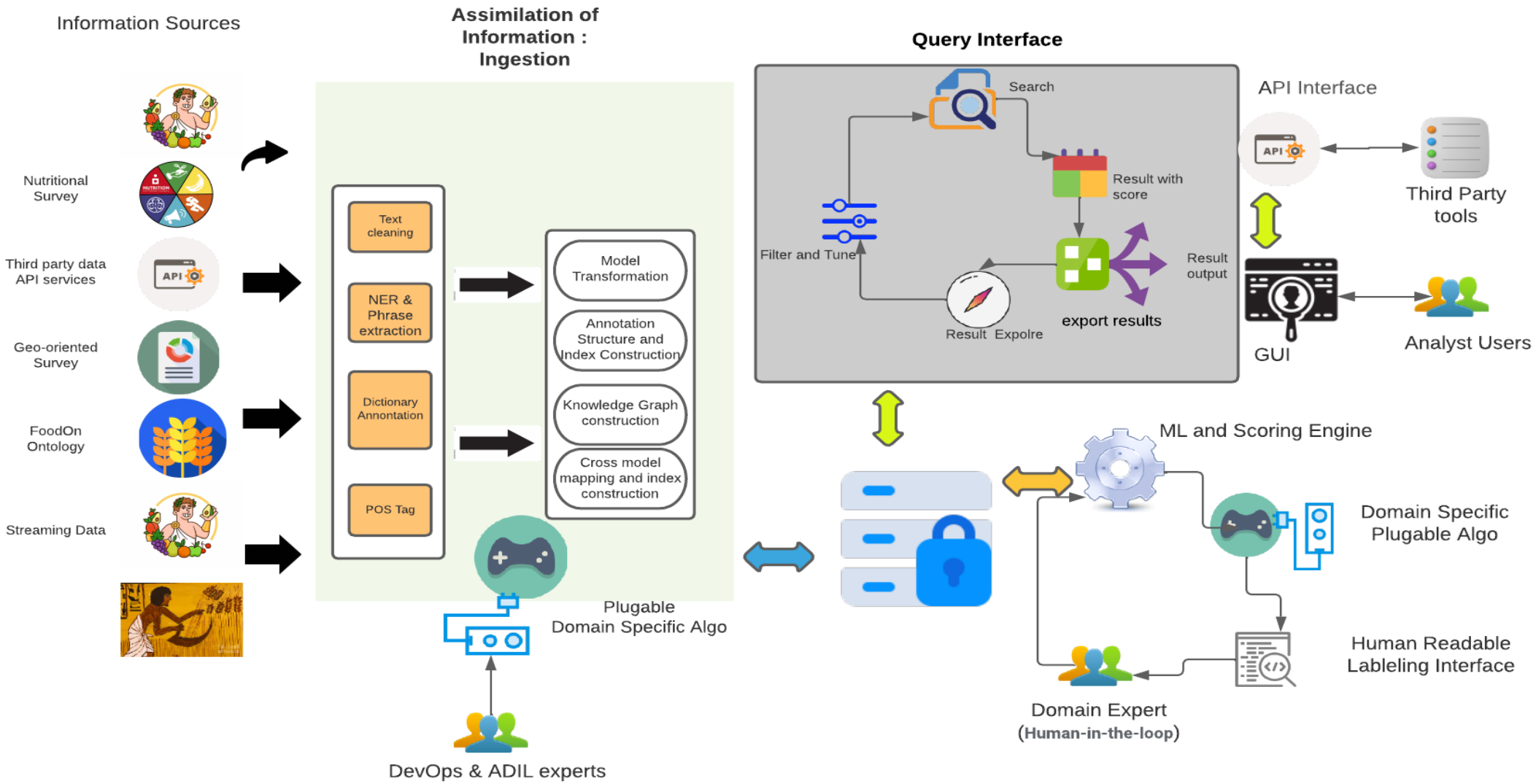
Building a Knowledge Graph Contd..



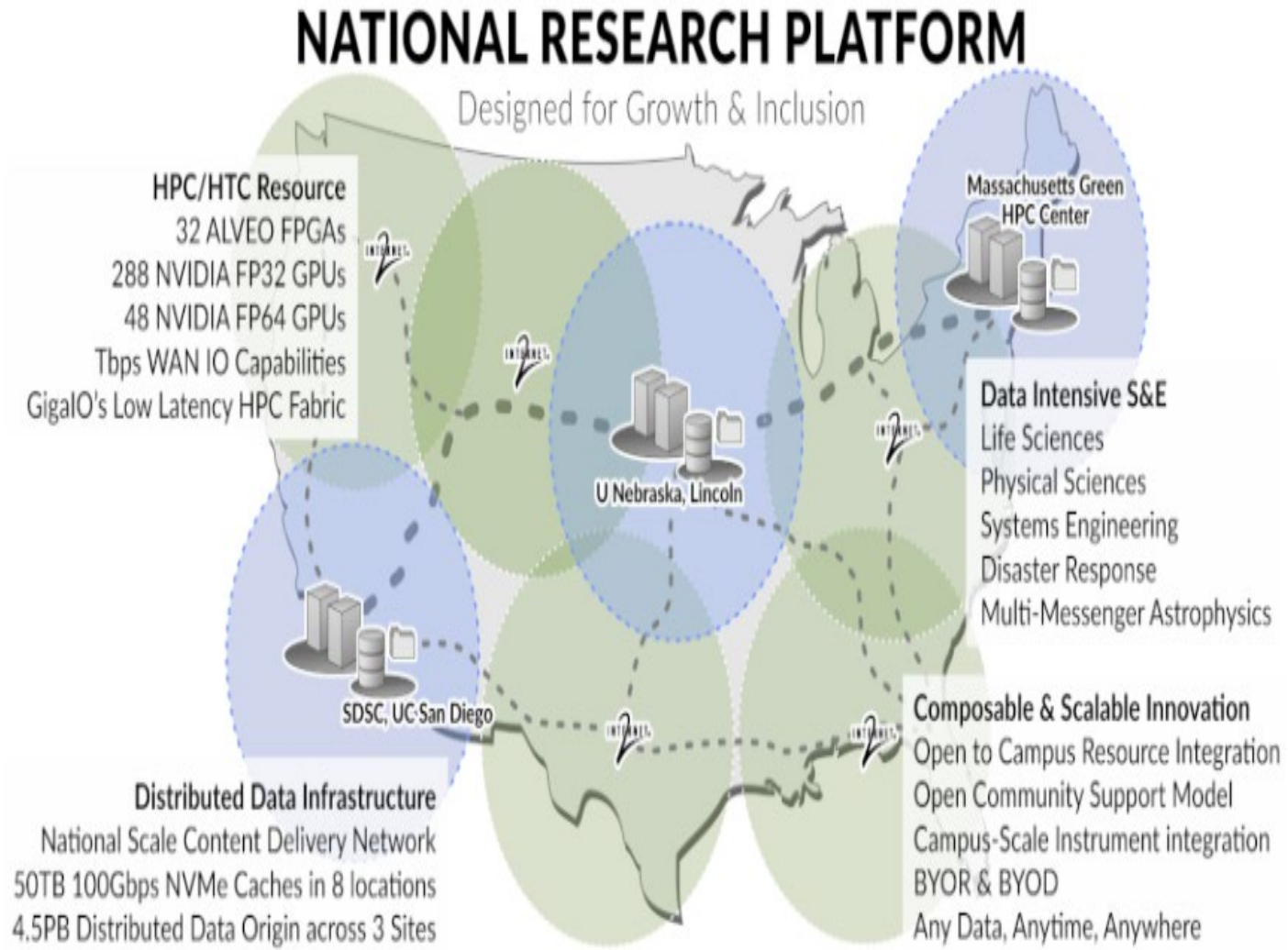
A Sample Project : Nourish

Network Of User-engaged Researchers building Interdisciplinary Scientific infrastructures for Healthy food (NOURISH)

Create technological solutions that aid individuals in converting food swamps into nutritious food systems.


Information Sources
Assimilation of Information : Ingestion
Query Interface
Nutritional Survey
Third party data API services
Geo-oriented Survey
FoodOn Ontology
Streaming Data
Text cleaning
NER & Phrase extraction
Dictionary Annontation
POS Tag
Model Transformation
Annotation Structure and Index Construction
Knowledge Graph construction
Cross model mapping and index construction
Plugable Domain Specific Algo
DevOps & ADIL experts
Search
Result with score
Filter and Tune
Result output
export results
Result Expolre
API Interface
Third Party tools
GUI
Analyst Users
ML and Scoring Engine
Domain Specific Plugable Algo
Human Readable Labeling Interface
Domain Expert
(Human-in-the-loop)


NRP
NATIONAL RESEARCH
PLATFORM


NATIONAL RESEARCH PLATFORM
Designed for Growth & Inclusion
HPC/HTC Resource
32 ALVEO FPGAs
288 NVIDIA FP32 GPUs
48 NVIDIA FP64 GPUs
Tbps WAN IO Capabilities
GigaIO's Low Latency HPC Fabric
Massachusetts Green HPC Center
U Nebraska, Lincoln
SDSC, UC San Diego
Data Intensive S&E
Life Sciences
Physical Sciences
Systems Engineering
Disaster Response
Multi-Messenger Astrophysics
Composable & Scalable Innovation
Open to Campus Resource Integration
Open Community Support Model
Campus-Scale Instrument integration
BYOR & BYOD
Any Data, Anytime, Anywhere
Distributed Data Infrastructure
National Scale Content Delivery Network
50TB 100Gbps NVMe Caches in 8 locations
4.5PB Distributed Data Origin across 3 Sites

Tech Publication and Patents

Dasgupta, S., K. Coakley, and A. Gupta. 2016. “Analytics-Driven Data Ingestion and Derivation in the AWESOME Polystore.” *2016 IEEE International*. https://ieeexplore.ieee.org/abstract/document/7840897/.

Dasgupta, S., C. McKay, and A. Gupta. 2017. “Generating Polystore Ingestion plans—A Demonstration with the AWESOME System.” *2017 IEEE International*. https://ieeexplore.ieee.org/abstract/document/8258297/.

Zheng, Xiuwen, Subhasis Dasgupta, Arun Kumar, and Amarnath Gupta. 2023. “An Optimized Tri-Store System for Multi-Model Data Analytics.” *arXiv [cs.DB]*. arXiv. http://arxiv.org/abs/2305.14391.

Gupta, A., and S. Dasgupta. (4th July,) 2023, Query processing in a polystore. *US Patent 11,693,856*, issued 2023. https://patents.google.com/patent/US20220083552A1/en.

Gupta, A., S. Dasgupta, and M. Roberts. 2022. Data ingestion into a polystore. *US Patent 11,288,261*, issued 2022. https://patents.google.com/patent/US11288261B2/en.

Significant Other Publications

1. Dasgupta, S., and A. Gupta. 2020. “Discovering Interesting Subgraphs in Social Media Networks.” In Social Networks Analysis and Mining …. https://ieeexplore.ieee.org/abstract/document/9381293/.

2. Mason, Ashley E., Frederick M. Hecht, Shakti K. Davis, Joseph L. Natale, Wendy Hartogensis, Natalie Damaso, Kajal T. Claypool, et al. 2022. “Author Correction: Detection of COVID-19 Using Multimodal Data from a Wearable Device: Results from the First TemPredict Study.” Scientific Reports 12 (1): 4568.

3. Purawat, Shweta, Subhasis Dasgupta, Luke Burbidge, Julia L. Zuo, Stephen D. Wilson, Amarnath Gupta, and Ilkay Altintas. 2021. “Quantum Data Hub: A Collaborative Data and Analysis Platform for Quantum Material Science.” In Computational Science – ICCS 2021, 656–70. Springer International Publishing.