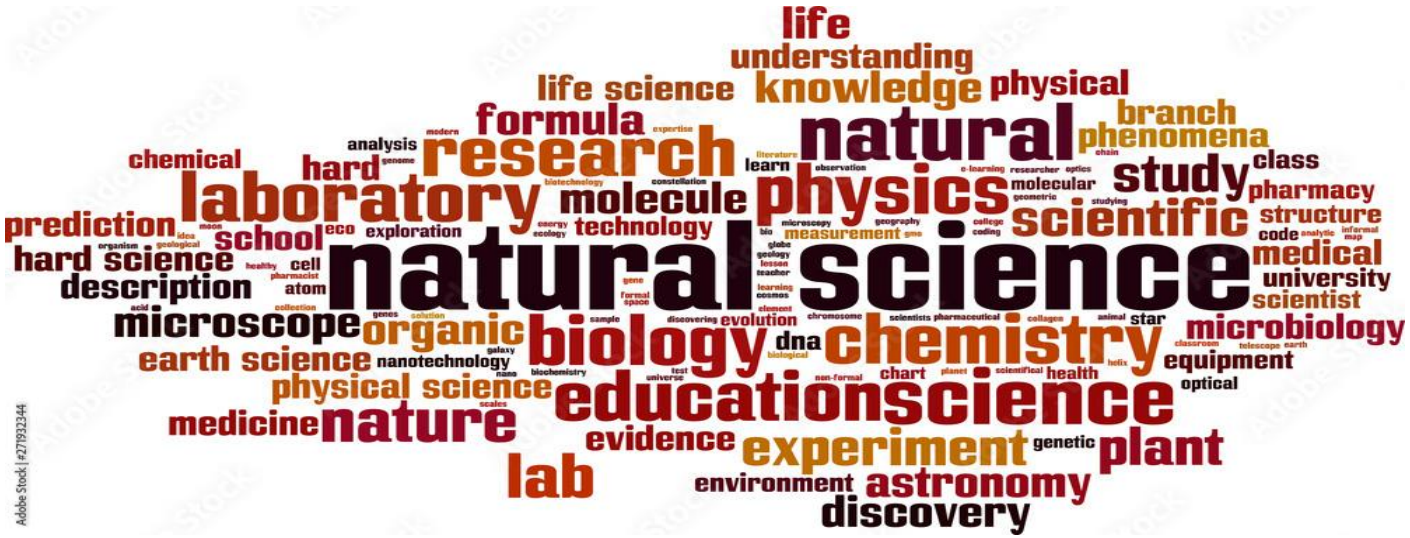
Subhasis Dasgupta, Ph.D.

Assistant Scientist, University of California San Diego

sudasgupta@ucsd.edu

Knowledge Management & Information Frameworks in the AI-Driven Scientific Ecosystem



Roadmap of the Talk

Paradigm Shift in Managing Scientific Data through AI and ML Technologies

Advancing Scientific Discovery: Ontology-Driven Knowledge Management

Storing and Retrieval of Data

Polystore Database Management Systems Research Agenda

Example Projects

Evolving Data Management in Science For AI

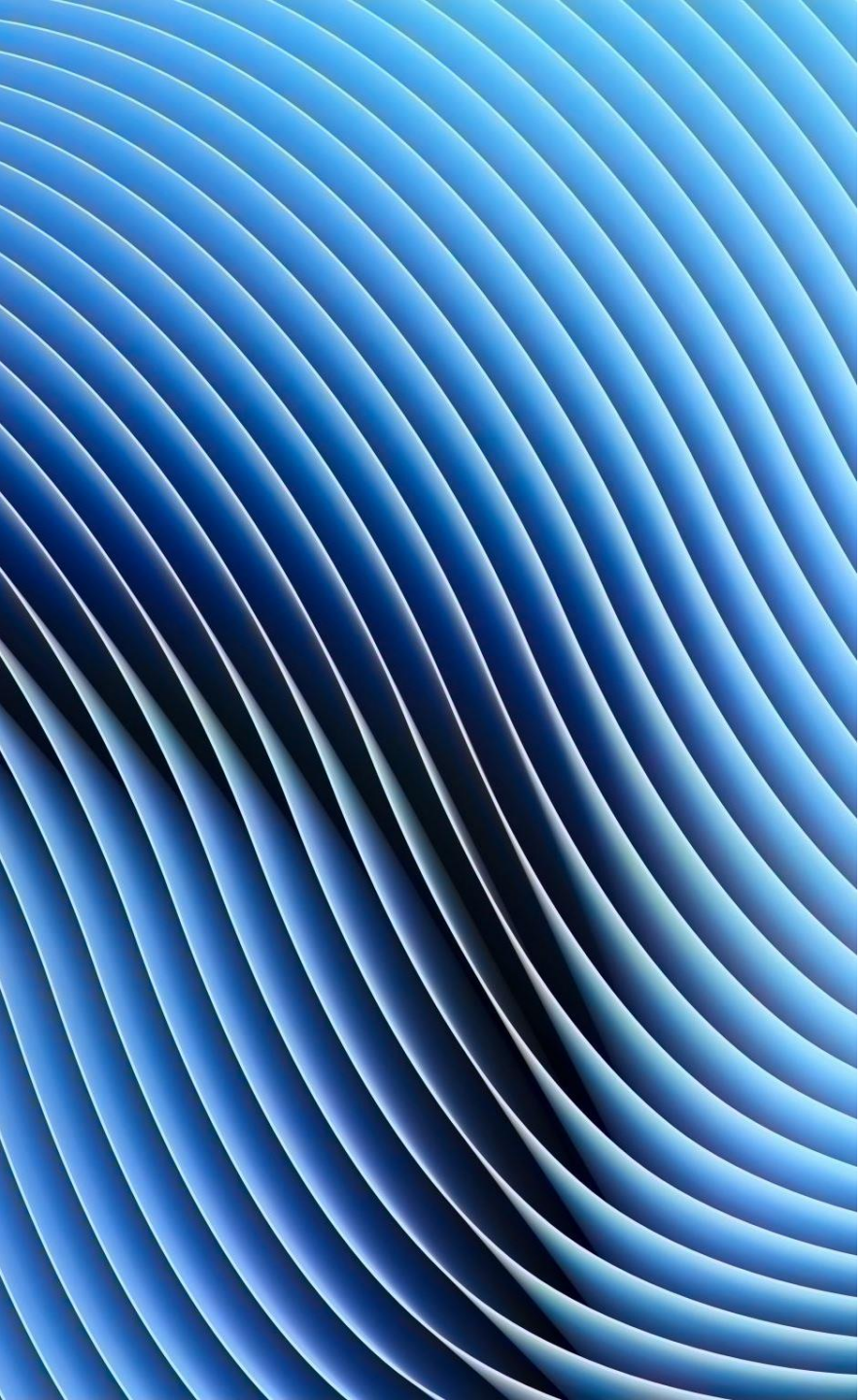
- **Unprecedented Data Growth:** Recognition of the surge in data volumes from diverse scientific disciplines. Shift in data collection methodologies to accommodate scale and diversity.
- **Heterogeneous Data Modeling:** There is a need for robust data models to support complex scientific queries. Developing sophisticated indexing mechanisms for efficient data retrieval.
- **Large Search Space and Ontological Annotation:** Integration of tools with the capability to navigate, analyze vast data sets, and annotate with Ontologies to domain-specific ontology to interpret the data.
- **Expressive Language Bindings:** Adopting programming languages and bindings that clearly define and handle intricate data requirements. Model and domain-specific language are available to explore capabilities. (CUDA, Rust, or Cipher)
- **Performance Scalability and Adaptability:** Emphasis on scalable systems that can adapt to the ever-changing landscape of scientific data. Requirement of special type hardware is required.

Basic Data Techniques

Data Selection

Data Generation

Data Refinements



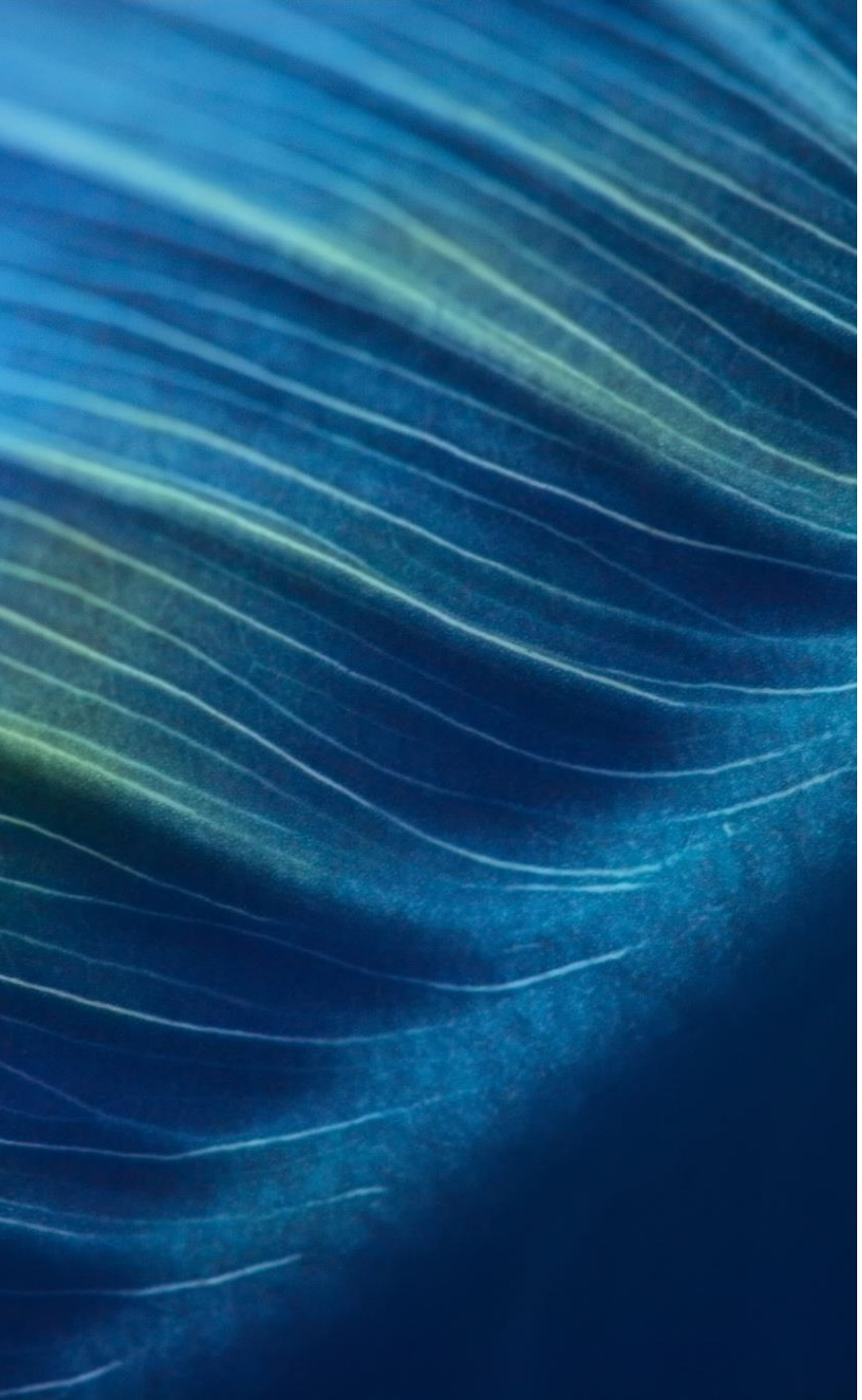
Data Selection

In science experiments, processing vast amounts of data is challenging.

- *For example, a single nuclear physics experiment can generate terabytes of data per second, which must be analyzed in real time. This is crucial in fields like physics and materials science, where immediate analysis is necessary for informed decision-making.*

Therefore, selecting relevant variables carefully becomes essential to manage and interpret these datasets effectively.

(Example Distributed Energy Resource Grid, Material Science Synthesis)



Data Selection

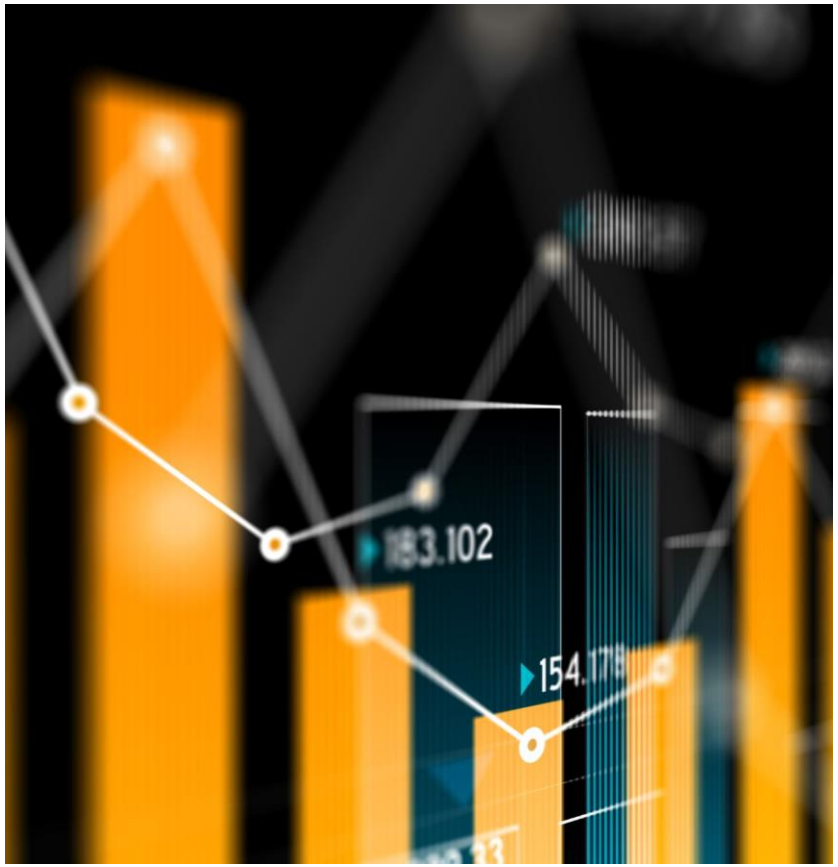
Scientists conduct experiments with the aim of discovering significant findings across various groups of subjects. Medical and biomedical sciences often create a demographic cohort to validate results across a large population, which is a good variation of their findings.

Advancements in deep learning and software-centric methods are eclipsing traditional hardware-focused technologies, employing autoencoders and unsupervised anomaly detection to identify novel patterns in data.

- Appropriate data curation enhances performance and mitigates hardware demands in these technological applications.

Data Generation

- A large and diverse dataset is required for efficient, fast, and robust AI/ML models. One way of getting data is by using synthetic data generation to increase the data's quality, diversity, and scale. Here are a few pathways for data generation:
  - **Generative Models:** Estimate a probability distribution of the underlying data and can then generate new samples from that distribution. Examples include variational autoencoders, generative adversarial networks, normalizing flows, diffusion models and generative pretrained transformers.
  - **Representation Learning:** Techniques automatically generate representations of data such as images, documents, sequences, or graphs. These representations are typically dense, compact vectors, referred to as embeddings or latent vectors, optimized to capture essential features of input data.
  - **Reinforcement Learning:** involves sequential decision-making and is represented as a Markov decision process comprising an agent, a set of states, a space of actions, an environment (which determines shows the state changes with actions) and a reward function.
  - **Deep Generative Learning, Physics-informed AI/ML etc.**



Data Refinement

Precision instruments such as ultrahigh-resolution lasers, biomedical sensors, or noninvasive microscopic enables direct measurement of physical quality or indirect measurement of the objects. AI/ML has significantly improved measurement by reducing noise, eliminating errors in measuring the roundness, and increasing efficacy and performance.

Such as Visualizing astrophysics data, understanding and predicting Black holes, Capturing Particle Collision, and understanding liver cell images to detect. Deep Convolution Methods variational autoencoder have been used for

The data requirements for those methods are very high volume and require special hardware and methodology needs to develop. Often people refer such pipeline as end to end learning.

End-to-end Learning: End-to-end learning refers to a method in machine learning where differentiable components, like neural network modules, are used to link raw inputs to outputs directly. This approach bypasses the necessity for manually crafted input features, allowing for the direct generation of predictions from inputs.

A Few Examples of Learning Scientific Data

- **Geometric priors:** Incorporating geometric priors into learned models is beneficial, as geometry and structure are key elements in scientific fields. The concept of symmetry, which is extensively examined in geometry, plays a significant role. It is defined through the principles of invariance and equivariance, which describe how a mathematical function, like a neural feature encoder, behaves when subjected to a set of transformations.

- **Geometric deep learning:** Graph neural networks have emerged as a leading approach for deep learning on datasets with underlying geometric and relational structures.

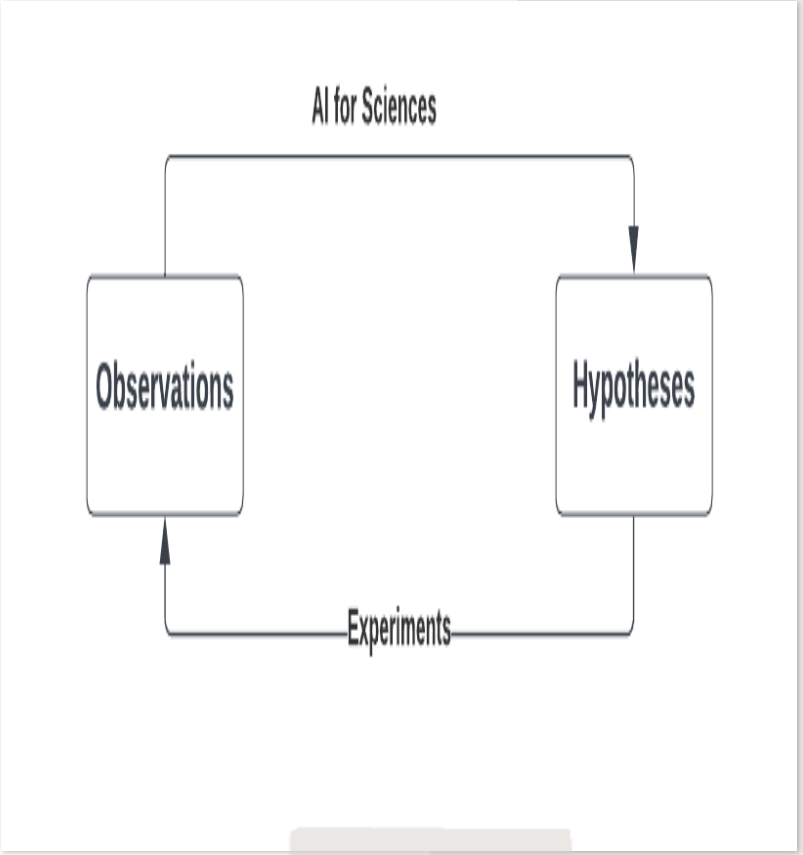
- **Self-Supervised Learning:** Self-supervised learning enables models to learn the general features of a dataset without relying on explicit labels.

- **Large Language Model:** Many LLM modes are prevalent, but the efficacy of such systems for the scientific world is not yet established.

Apart from the above, many methodologies are available, like Transformers, Neural Networks, etc.

Scientific Hypothesis and AI/ML

- **Testable hypotheses** are central to scientific discovery. It could be a symbolic expression in mathematics to molecules in chemistry or a genetic variant in biology.
- AI methods can be helpful at several stages of this process. They can **generate hypotheses by identifying candidate symbolic expressions from noisy observations**.
- Common hypothesis Spaces are :
    - **Black-box predictors of scientific hypotheses**
    - **Navigating combinatorial hypothesis spaces**
- Proper knowledge management instruments can optimize hypothesis spaces by providing better visibility of knowledge required for data analysis.



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.

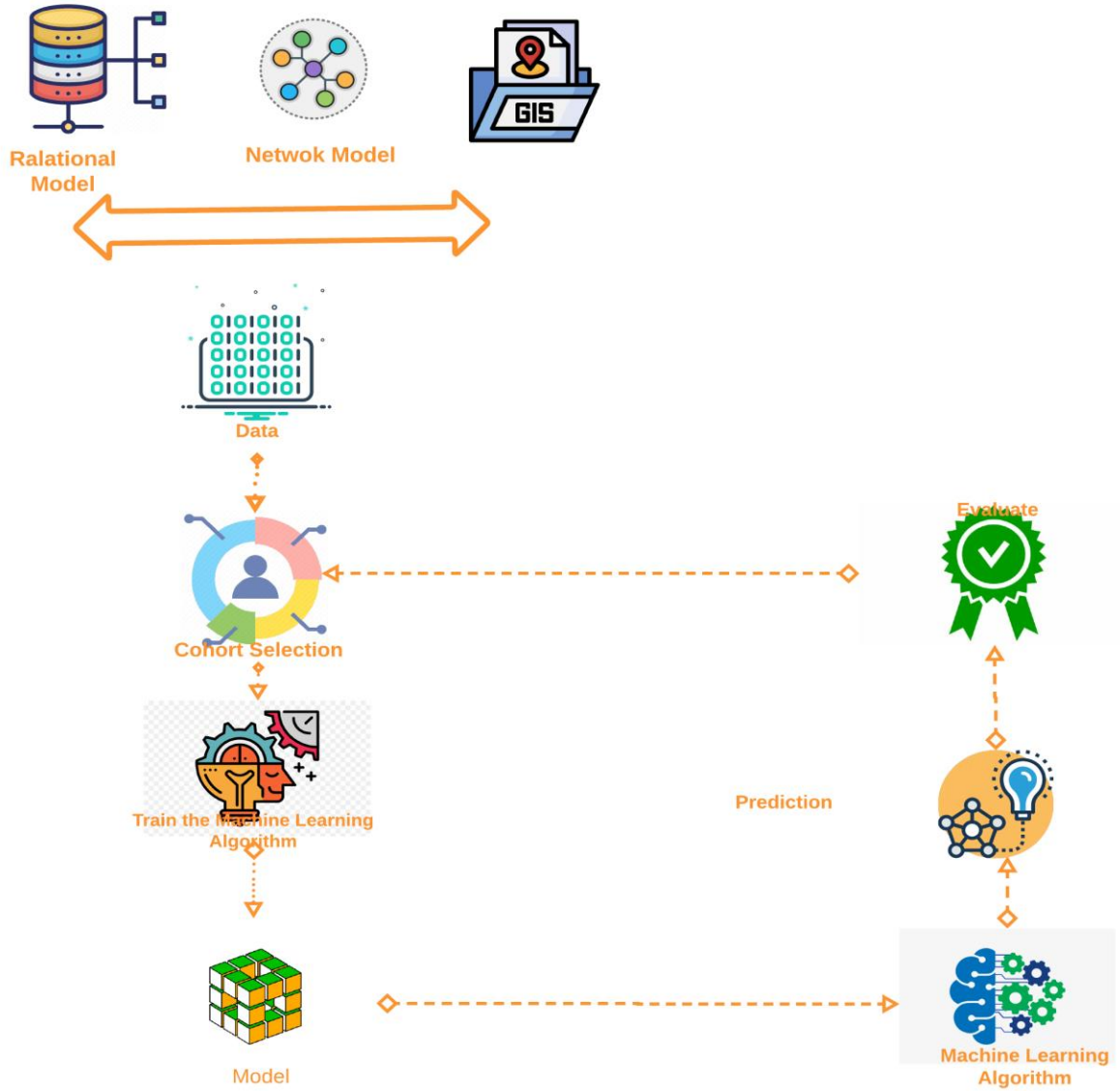
While using a database for data retrieval can be effective, traditional databases may not offer the level of flexibility that some situations require.

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.



What is the process for effectively managing knowledge?

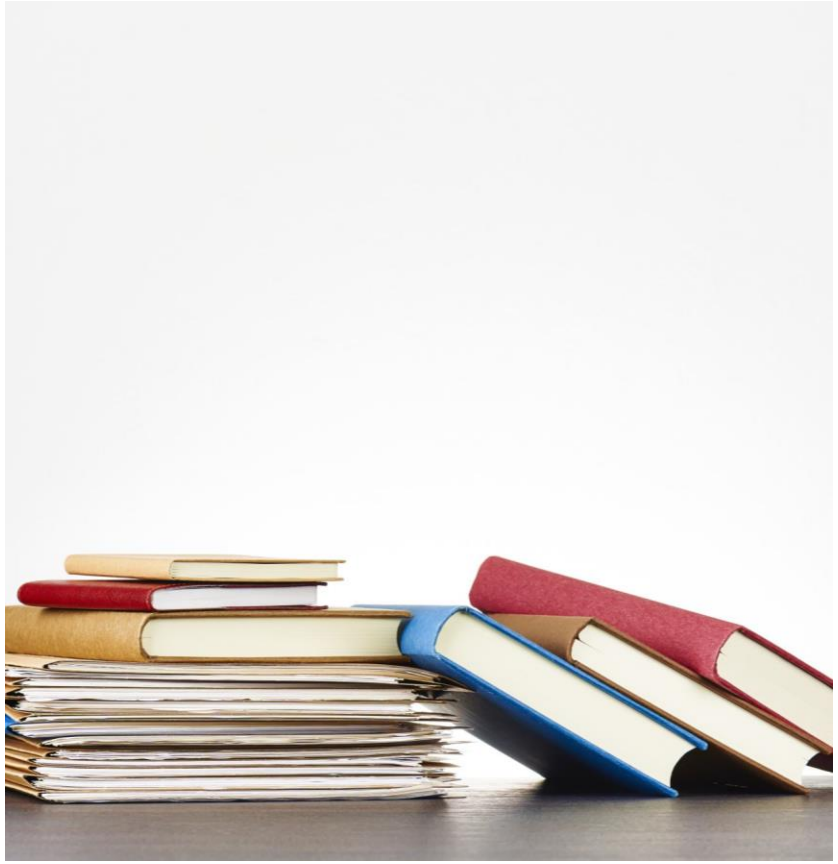
The roots of knowledge management can be traced back to a significant period in history.

- **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

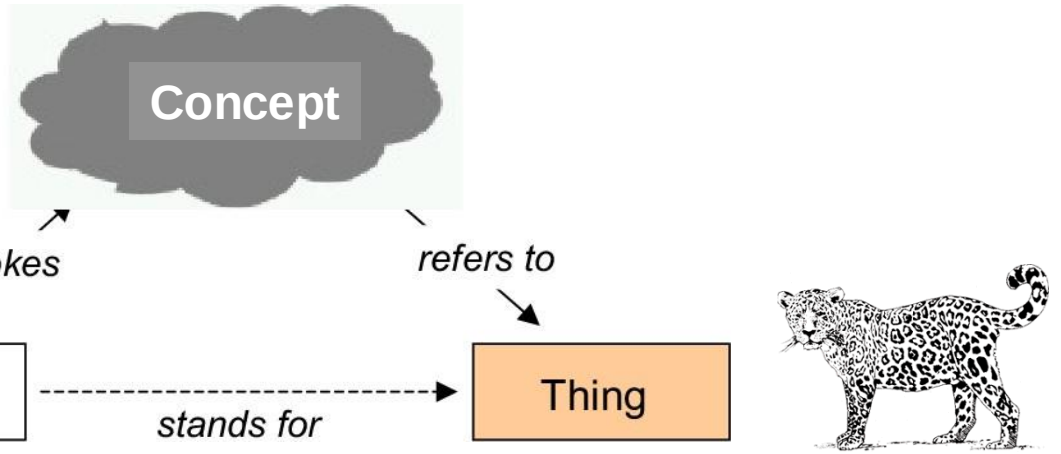
- Documentalist
- Technologist
- Communicator and Curator
- Lerner and Scholar

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*:



"Jaguar"

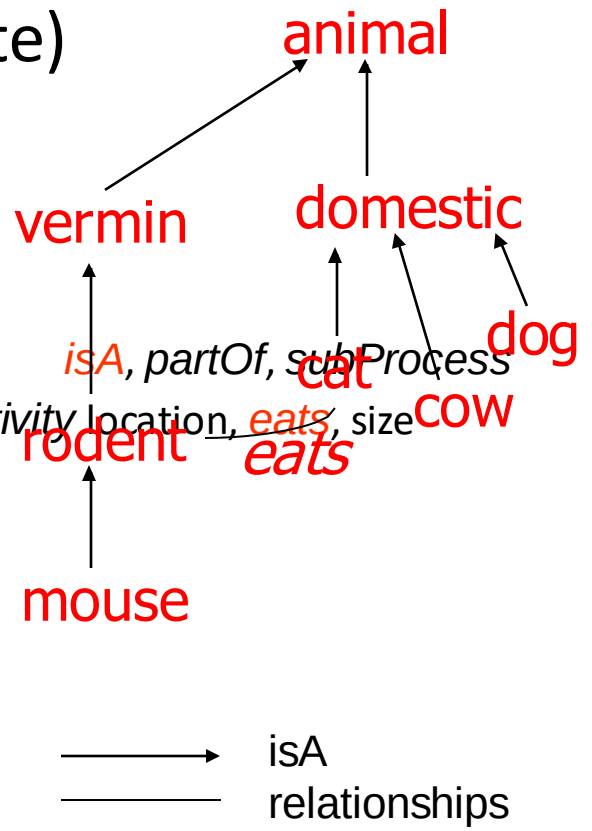
Ogden, C. K. & Richards, I. A. 1923. "The Meaning of Meaning." 8th Ed. New York, Harcourt, Brace & World, Inc

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]

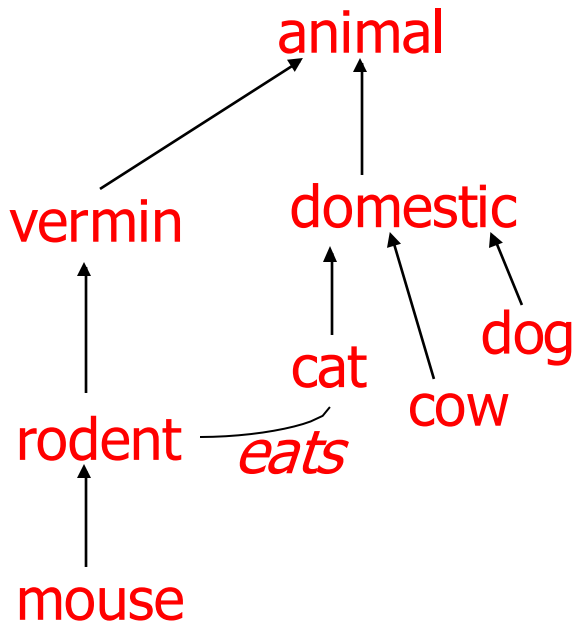
An explicit description of a domain

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]

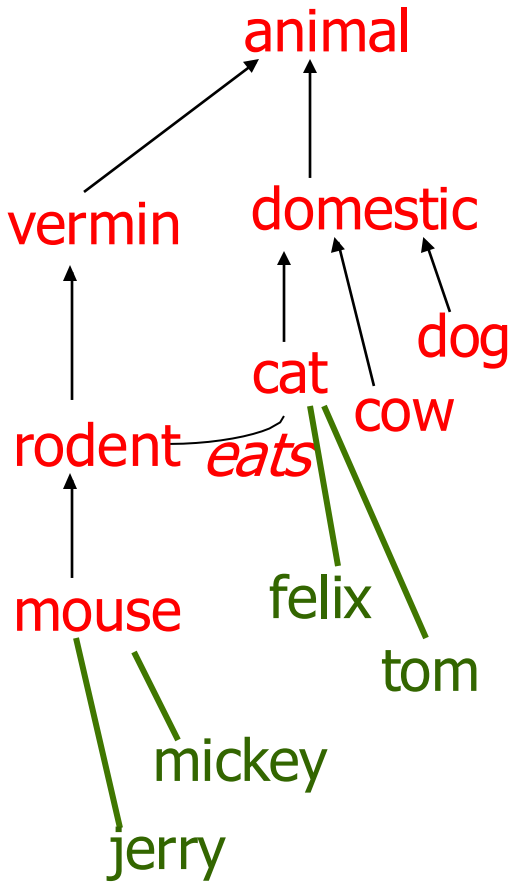
An explicit description of a domain

Individuals or *Instances*

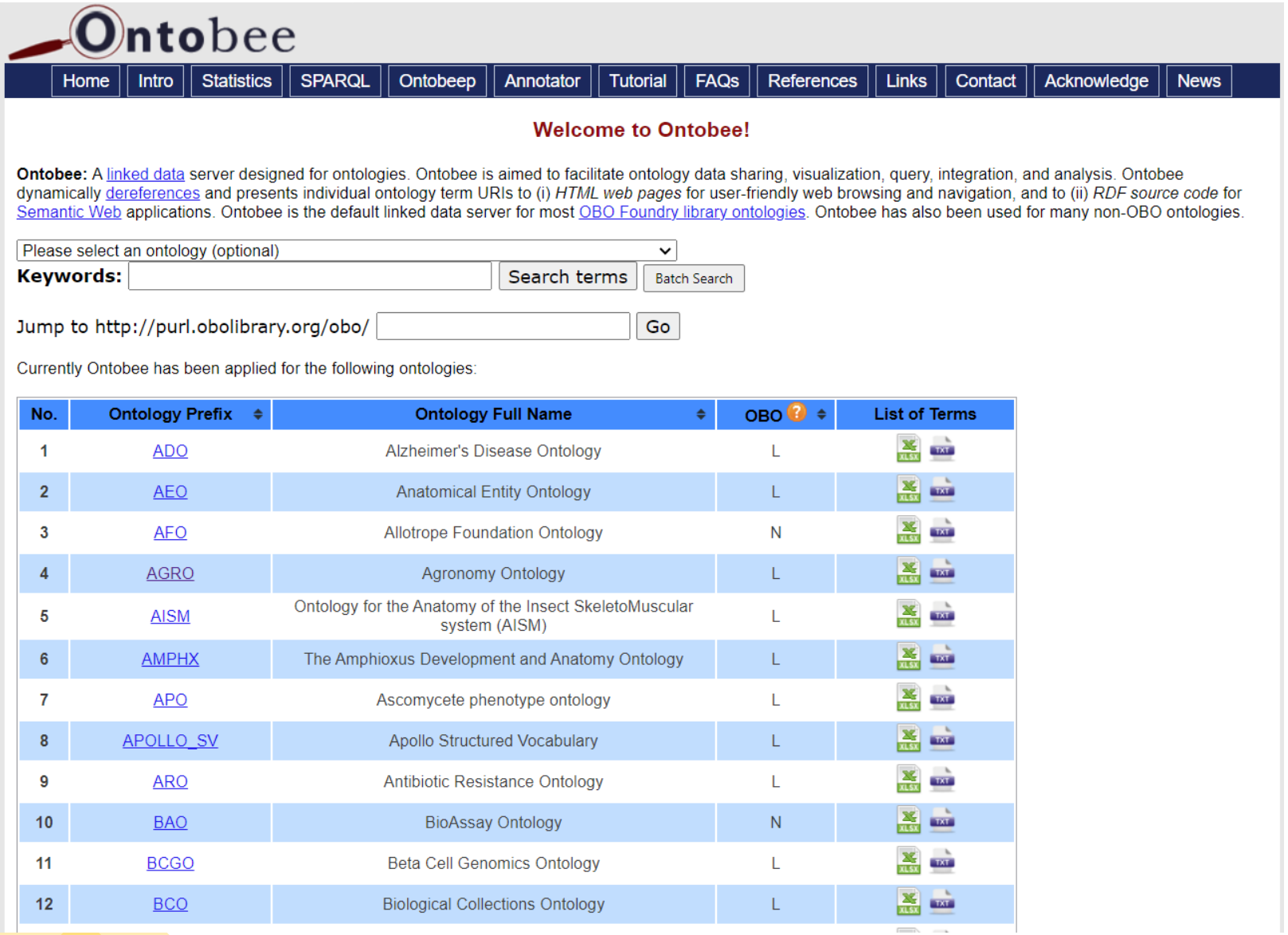
- sulphur, trpA Gene, **felix**

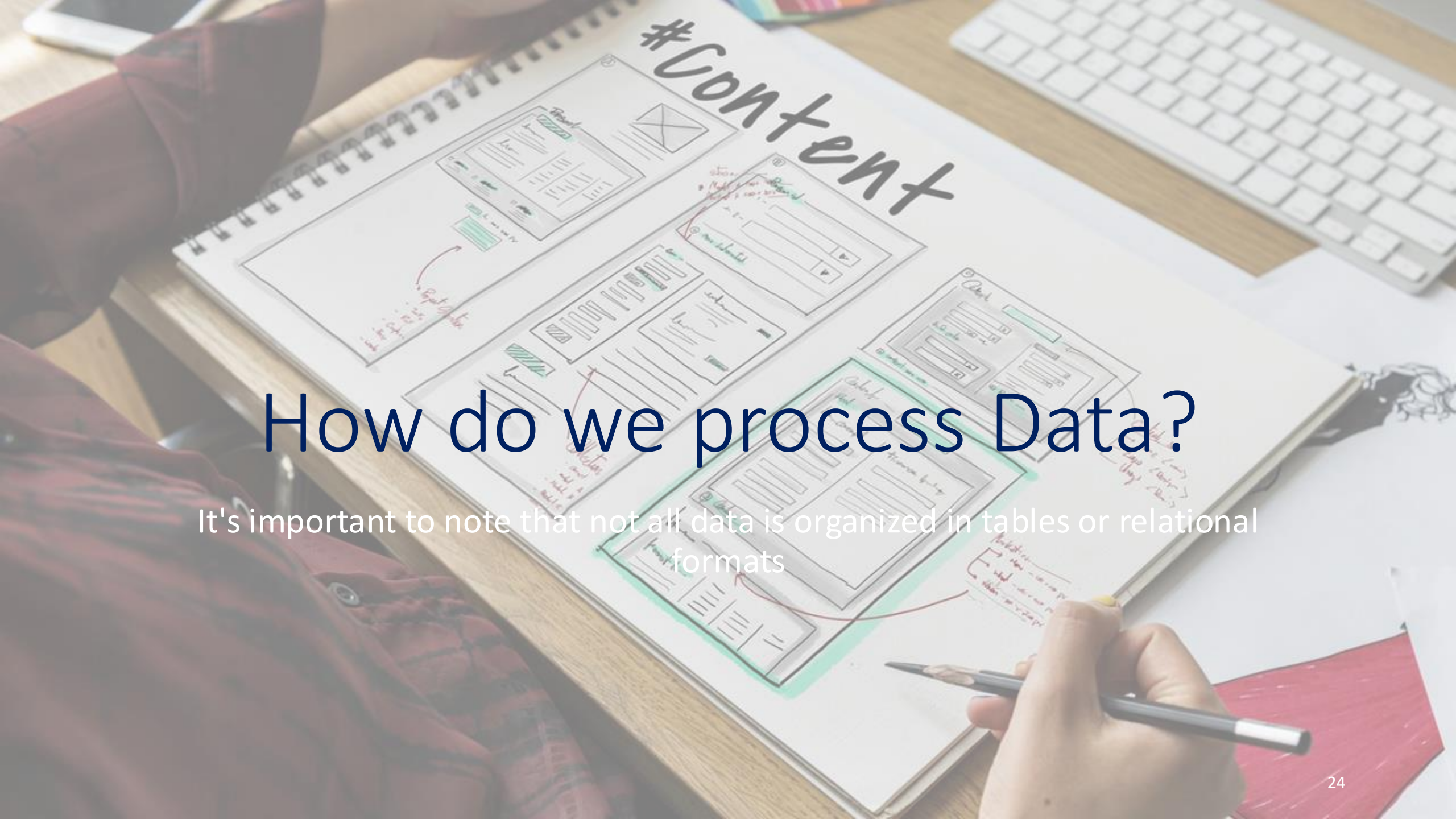
Ontology versus Knowledge Base

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



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

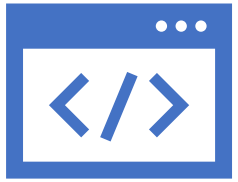

Ontobee
Home
Intro
Statistics
SPARQL
Ontobeep
Annotator
Tutorial
FAQs
References
Links
Contact
Acknowledge
News
Welcome to Ontobee!
Ontobee: A linked data server designed for ontologies. Ontobee is aimed to facilitate ontology data sharing, visualization, query, integration, and analysis. Ontobee dynamically dereferences and presents individual ontology term URIs to (i) HTML web pages for user-friendly web browsing and navigation, and to (ii) RDF source code for Semantic Web applications. Ontobee is the default linked data server for most OBO Foundry library ontologies. Ontobee has also been used for many non-OBO ontologies.
Please select an ontology (optional)
Keywords:
Search terms
Batch Search
Jump to http://purl.obolibrary.org/obo/
Go
Currently Ontobee has been applied for the following ontologies:
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



How do we process Data?

It's important to note that not all data is organized in tables or relational formats

Comma Separated Value (CSV) File



Each value is separated by a comma (except plain text)

No Specified field lengths.

The total number of CSV datasets on Kaggle is 115,936, but 21,181 JSON datasets are available.

CSV format Problem

The CSV file format lacks complete standardization, with the only standardized rule being the use of the comma.

Representing a dynamic schema can be quite challenging due to its large and complex nature.

Advantages of Semi Structure Representation

01

Structure Data: Organized according to a formal data model(i.e., relational)

02

Semi-structured Data: No formal data model, but contains symbols to separate and label data element

03

Unstructured Data: No data model no pre-defined organization

Data Example


Relational: CSV, Relational Databases


Semi-structured: XML, JSON


Unstructured: Text, Document, Image

Why is semi-structured data important even though CSV is the most popular format?

- *Heterogeneous data integration is achieved through input and output sources that can be of various types.*
- *The most accepted format for web services or machine learning libraries is used.*

XML VS JSON

Here is a comparison between the XML and JSON data outputted by the GridLab-D Simulator.

```
▼<property>
   <object>trip_meter13</object>
   <name>measured_real_energy</name>
   <value>+169411 Wh</value>
   <type>double</type>
 </property>
```

```
{       "object" : "trip_meter13",
        "name" : "measured_real_energy",
        "type" : "double",
        "value" : "+996682 Wh"
}
```

XML (Extensible Markup Language)

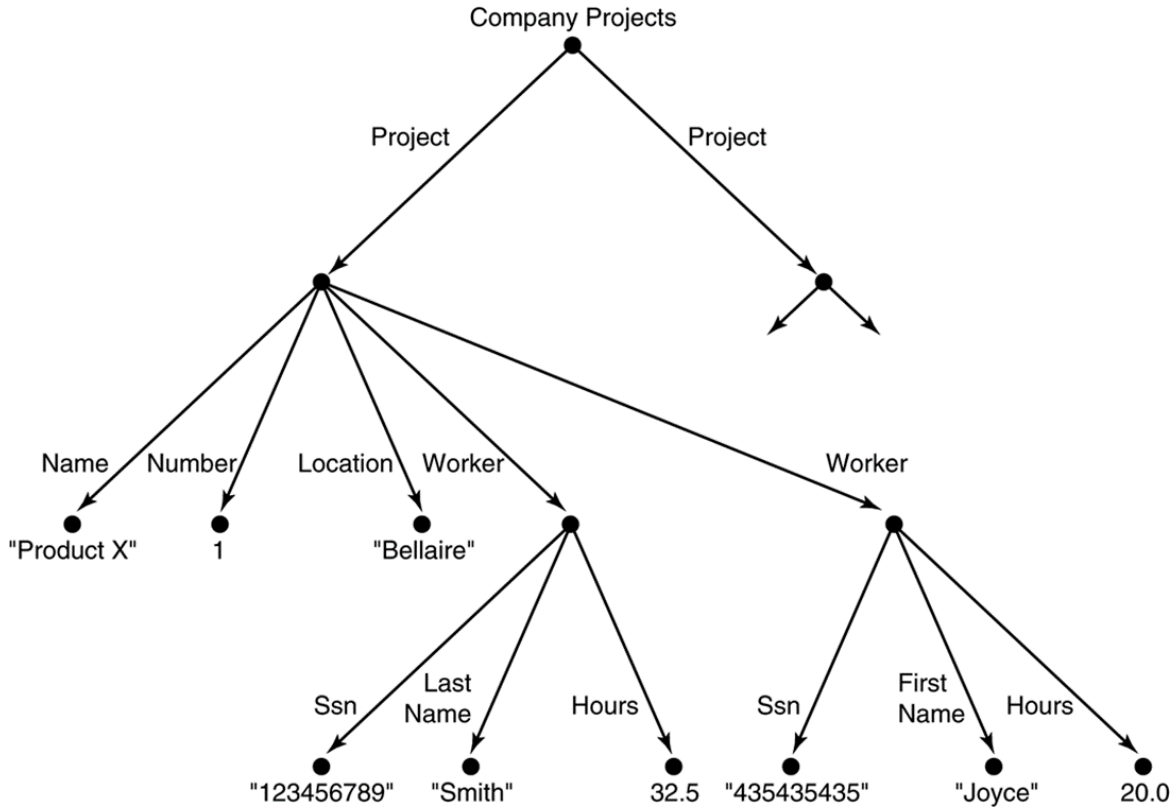
- Plain Text
- User text for values between tags for labels
- Value can be any length
- Commas and Quotes are valid
- Field can be skipped or create a hierarchy

```
▼<property>
   <object>trip_meter13</object>
   <name>measured_real_energy</name>
   <value>+169411 Wh</value>
   <type>double</type>
 </property>
```

JavaScript Object Notation(JSON)

- Plain Text
- Organized as objects within braces {}
- Uses key-value pairs
  - Keys are field names, and values are data strings, numbers, Boolean
- You have the option to skip the "Field" section.
- Is it possible to use a datatype such as datetime or a specific type of index like r-index?

Example Semi-Structured Data

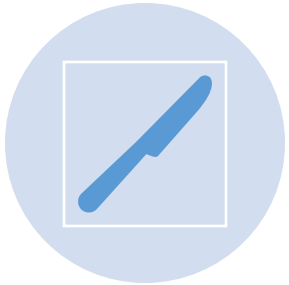


A complex XML element called <projects>

```
<?xml version="1.0" standalone="yes"?>
<projects>

 <project>
  <Name>ProductX</Name>
  <Number>1</Number>
  <Location>Bellaire</Location>
  <DeptNo>5</DeptNo>
  <Worker>
   <SSN>123456789</SSN>
   <LastName>Smith</LastName>
   <hours>32.5</hours>
  </Worker>
  <Worker>
   <SSN>453453453</SSN>
   <FirstName>Joyce</FirstName>
   <hours>20.0</hours>
  </Worker>
 </project>
 </project>
  <Name>ProductY</Name>
  <Number>2</Number>
  <Location>Sugarland</Location>
  <DeptNo >5</DeptNo >
  <Worker>
   <SSN>123456789</SSN>
   <hours>7.5</hours>
  </Worker>
  <Worker>
   <SSN>453453453</SSN>
   <hours>20.0</hours>
  </Worker>
  <Worker>
   <SSN>333445555</SSN>
   <hours>10.0</hours>
  </Worker>
 </project>

...

</projects>
```

Example API Services

How to handle JSON Data?



File System with Python library? (You could have similar for R, Java, Go, rust, etc.)

Accessing the Data using dataframe: I will use Panda

Can we use Relational Data?

Can we have a Semistructured Database?


OPENLINK®
VIRTUOSO
UNIVERSAL SERVER

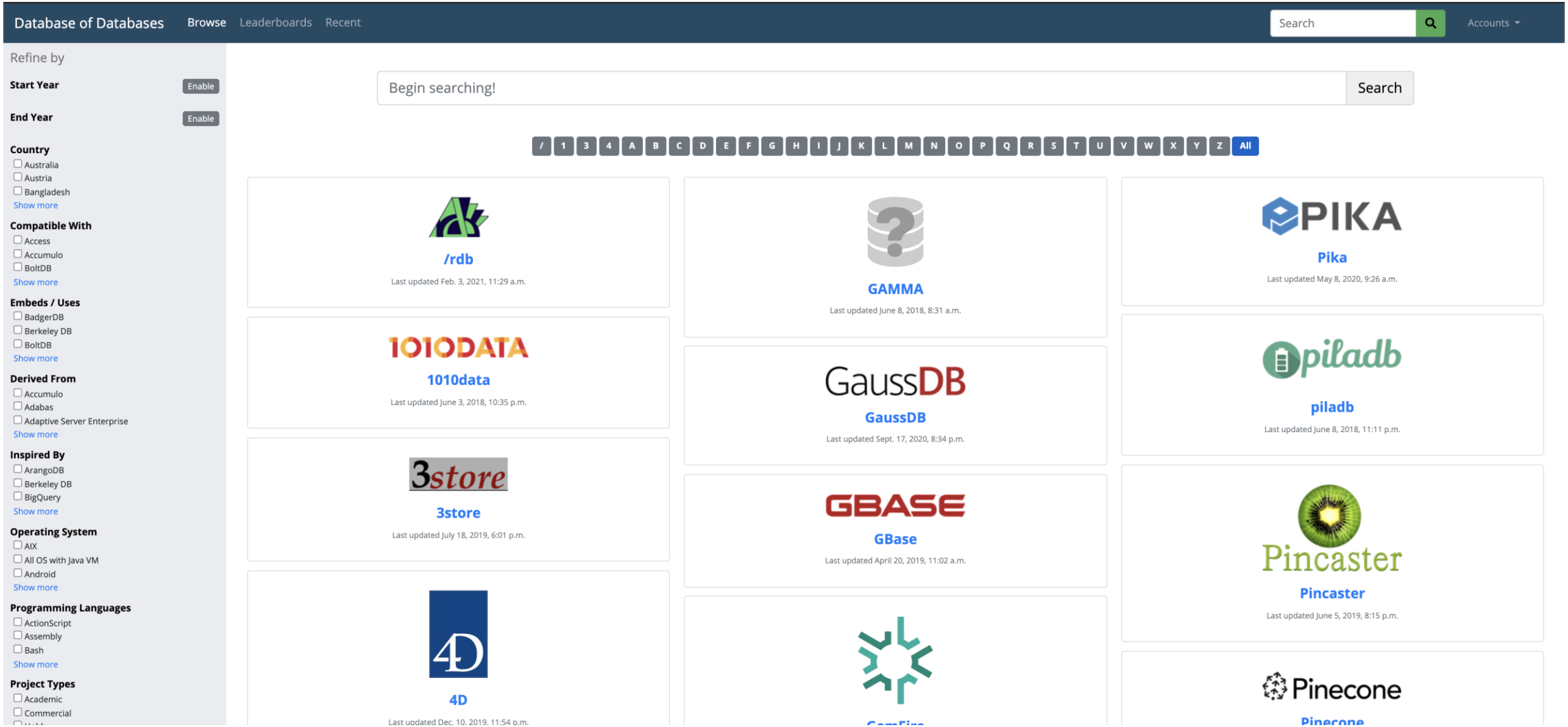

neo4j


protégé

“No One Size Fits All”


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
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

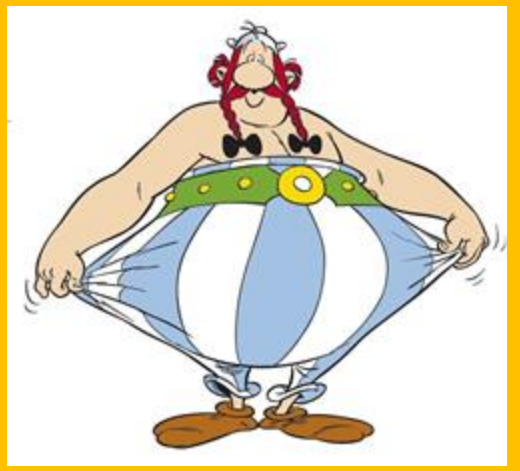
APACHE Spark

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

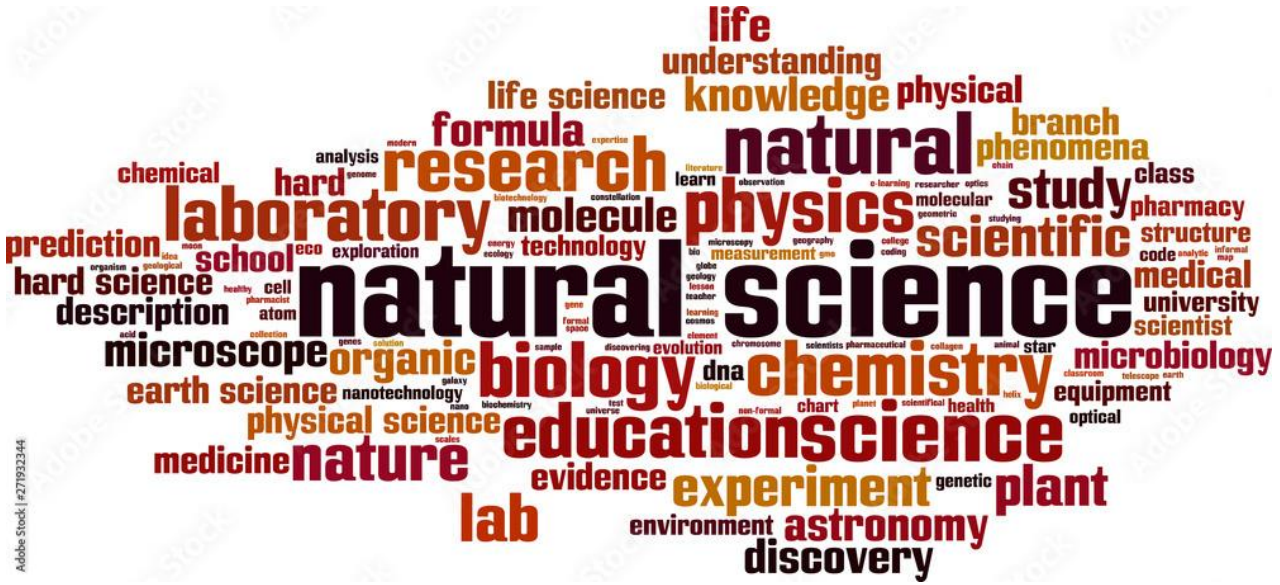
1. Use a common interface to interact among stores, like a standard data frame or data structure for the whole polystore.
2. Local storage has less control over the data, and the central controller decides everything.
3. Transforming or rewriting queries from one store to another store is complex and 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



Case Study : The Awesome Polystore

Data Variety

Social Network

Newspaper

Contract and Legal

Clinical Trial Data

Organizational Data

SEMISTRUCTURED, VIDEO, TEXT

STRUCTURED TEXT, TIME, PLACES

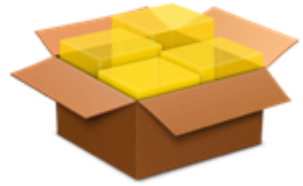
TEXT

TEXT, IMAGE

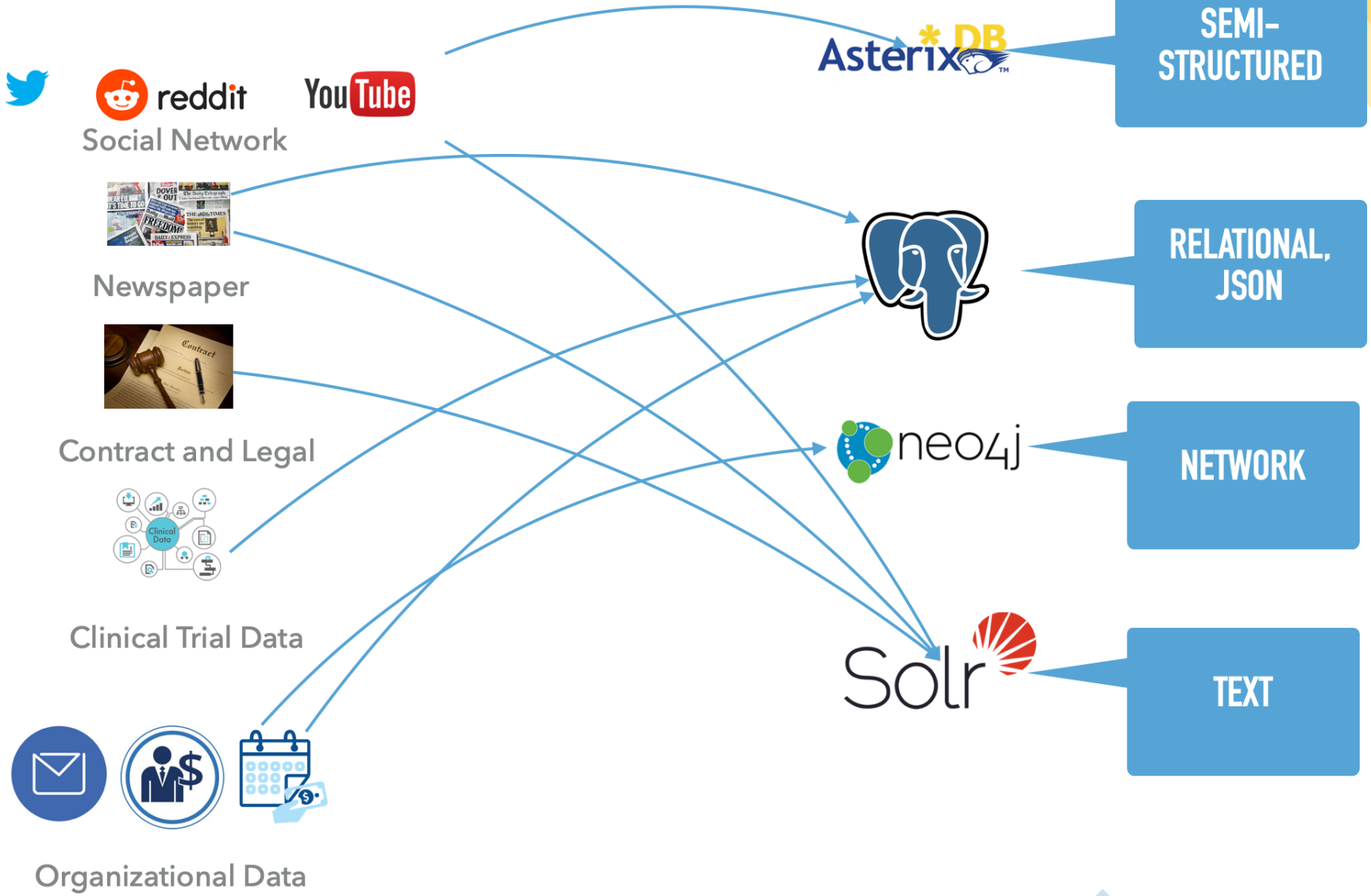
EMAIL, RELATIONAL TABLES, SAP DATA, TEXT, CUSTOMER REVIEW DATA

Dictionaries

HITL



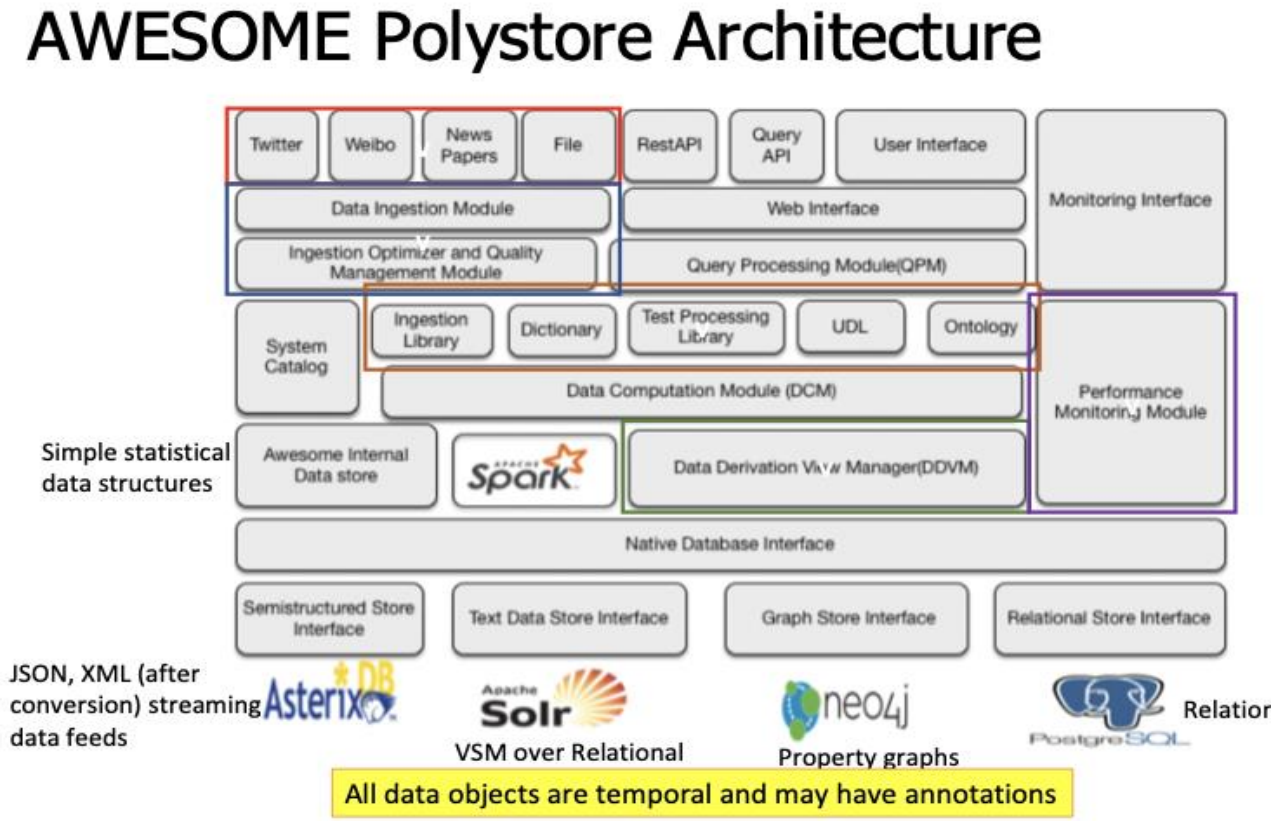
Packages


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

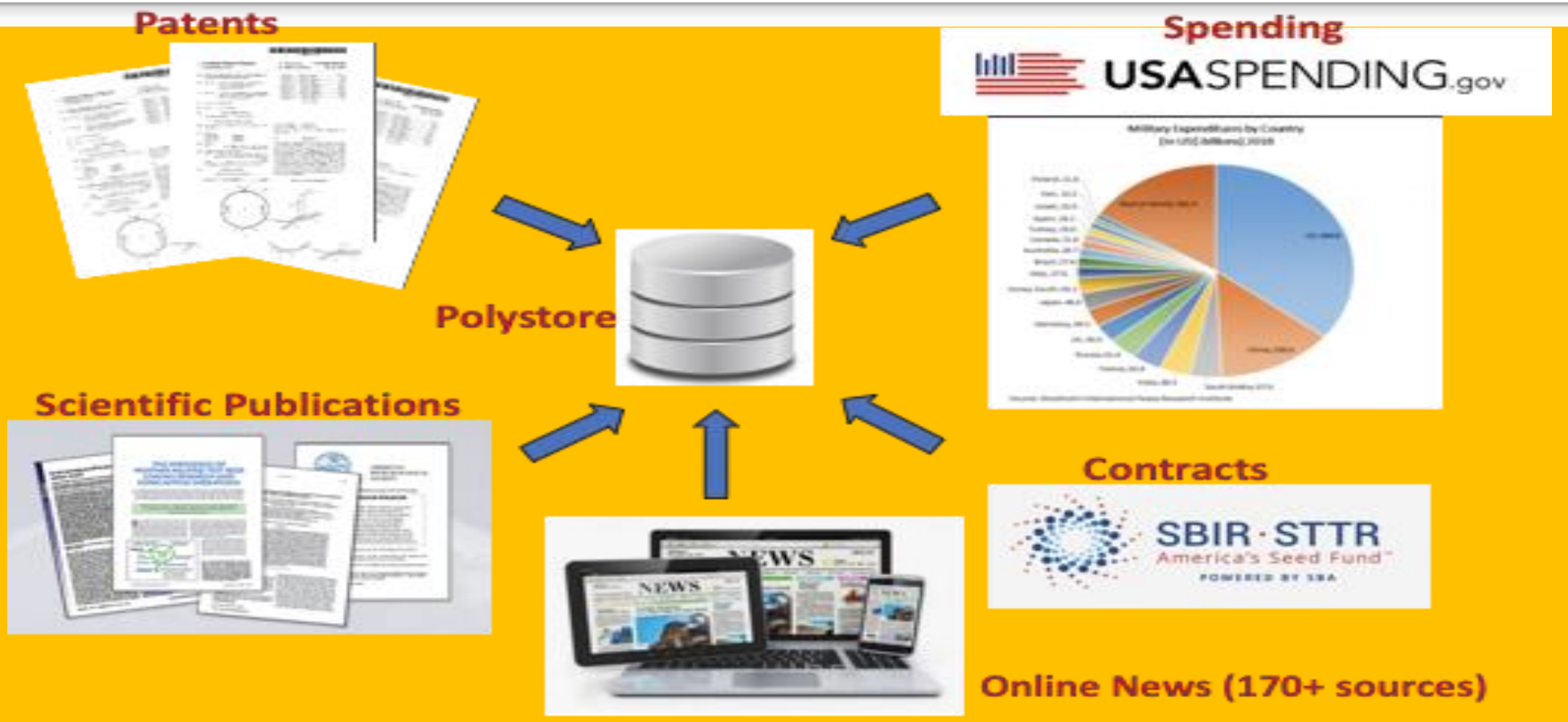
- Social Sciences Questionnaire:
  - List all accounts talking about "Elizabeth Warren" and "American DOS Movements.“
  - Find the top 100 influential co-spiking users and discuss racial terms from the “Elizabeth Warren’s” network.
  - Find out top-k topics from the newspaper that are also discussed 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



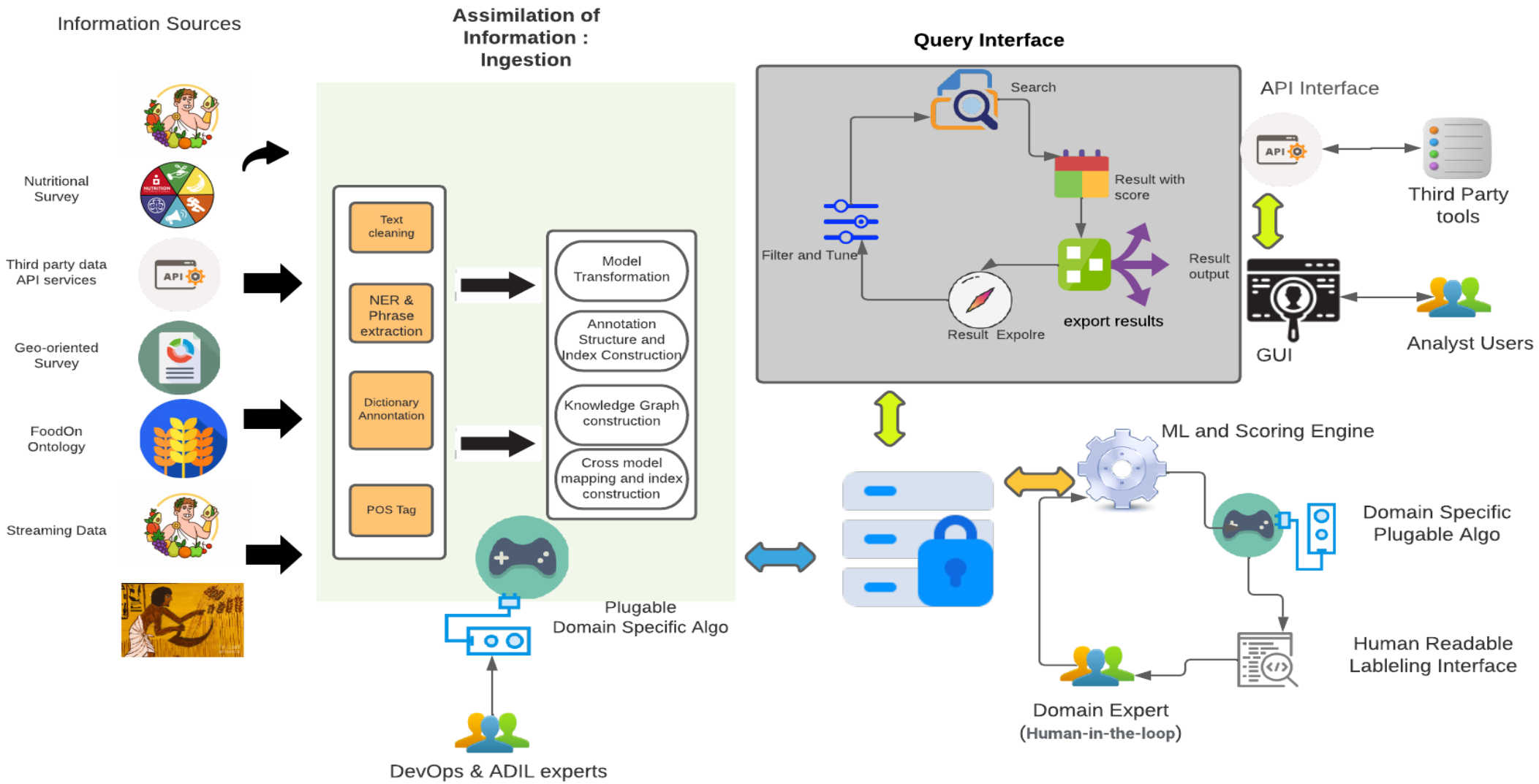
Building a Knowledge Graph



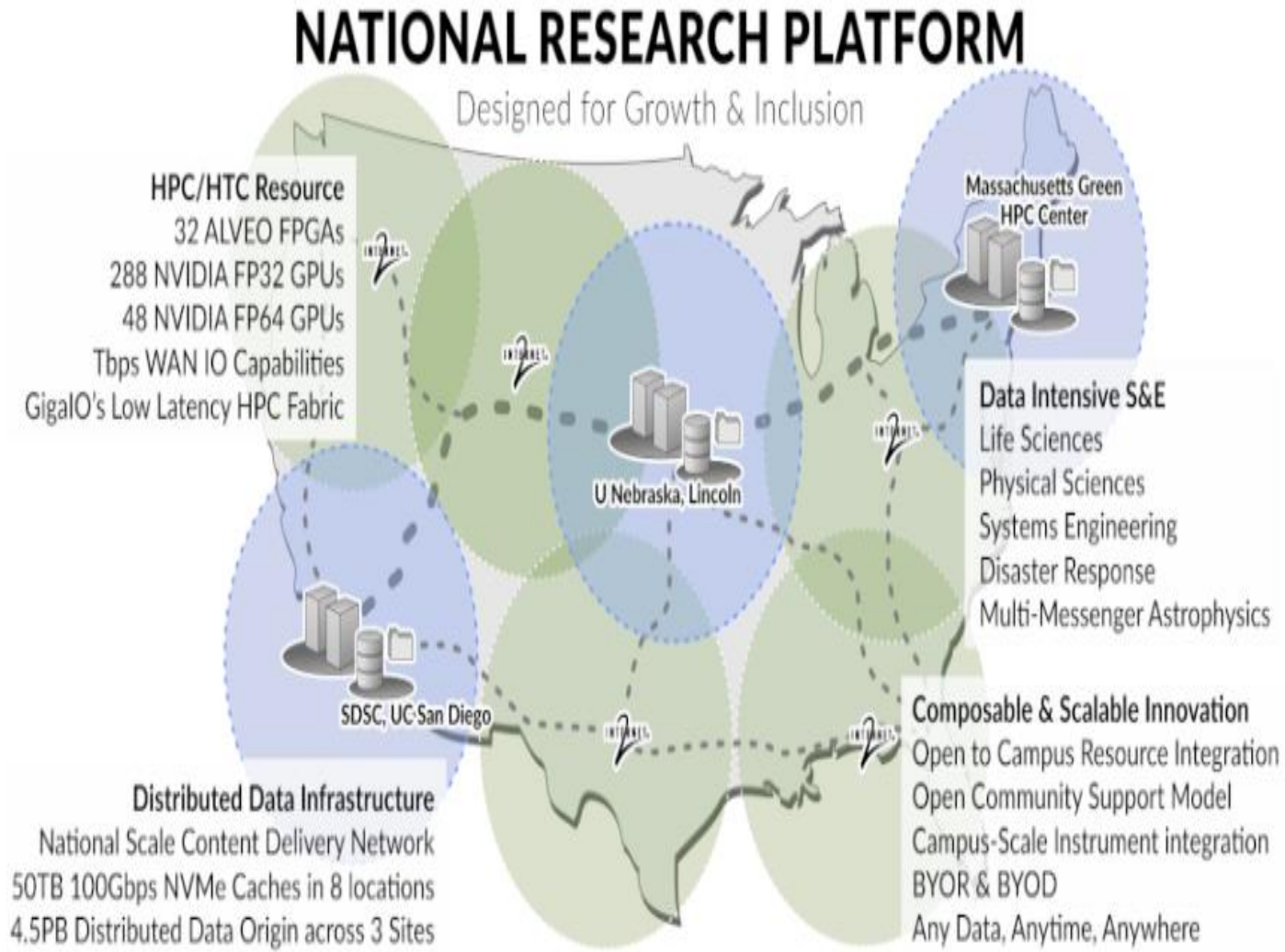
A Sample Project : Nourish

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.