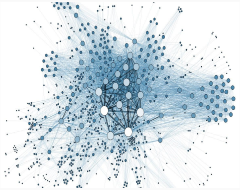
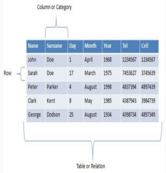
Managing Mixed Modeled Data using Polystore.

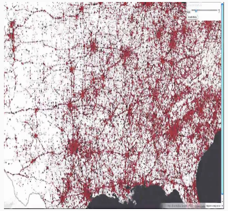
Subhasis Dasgupta, Ph.D

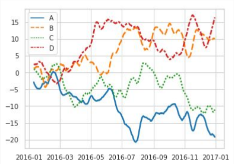
Mixed Modeled Data

- We produce, use, analyze, and store a massive amount of data for various purposes.
- We modeled them differently to enable analysis.
- Hence we need to handle different model of data in reality.



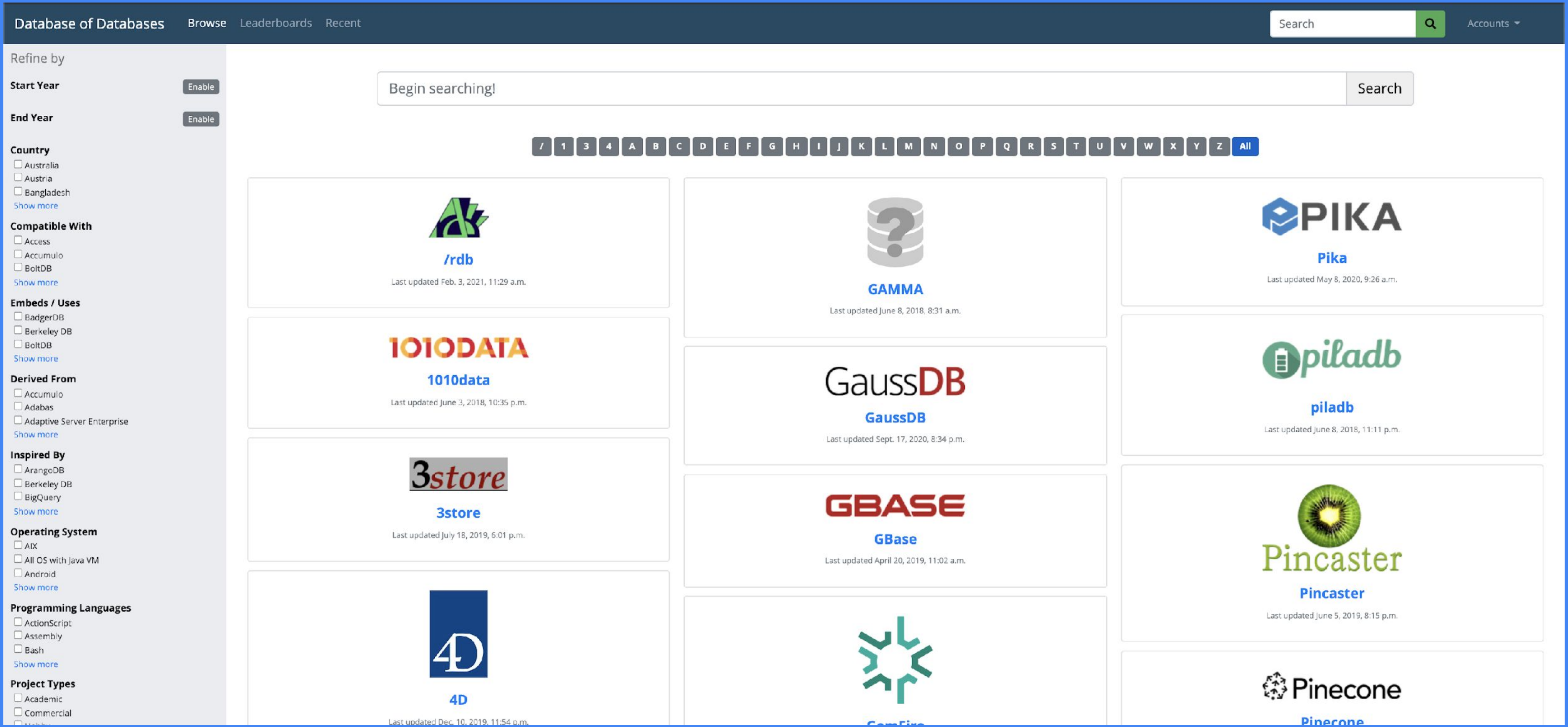






Model Specific Databases

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
1010data
Last updated June 3, 2018, 10:35 p.m.
3store
3store
Last updated July 18, 2019, 6:01 p.m.
4D
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
GBase
Last updated April 20, 2019, 11:02 a.m.
PIKA
Pika
Last updated May 8, 2020, 9:26 a.m.
piladb
piladb
Last updated June 8, 2018, 11:11 p.m.
Pincaster
Pincaster
Last updated June 5, 2019, 8:15 p.m.
Pinecone

“No One Size Fits All”

LARGE VOLUME AND CAPABILITIES

EXAMPLE OF LARGE QUERIES

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

```
8 rows                                                     Comma-...d (CSV)
  QUERY PLAN
1 Gather  (cost=1000.00..1307233.58 rows=600 width=979) (actual time=293360.623..1907816.219 rows=42 loops=1)
2   Workers Planned: 4
3   Workers Launched: 4
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)
5         Filter: (content ~~ '%華婦%'::text)
6         Rows Removed by Filter: 1311854
7 Planning time: 4.014 ms
8 Execution time: 1907863.092 ms
```

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

```
   QUERY PLAN
1  Custom Scan (Citus Real-Time)  (cost=0.00..0.00 rows=0 width=0) (actual time=3105325.446..3188792.688 rows=21048916 loops=1)
2    Task Count: 32
3    Tasks Shown: One of 32
4    ->  Task
5          Node: host=10.128.22.143 port=5432 dbname=postgres
6          ->  Gather  (cost=1000.00..4457975.83 rows=59729 width=885) (actual time=5.433..774096.735 rows=656963 loops=1)
7                Workers Planned: 2
8                Workers Launched: 2
9                ->  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…
10                     Filter: (lower((text)::text) ~~ '%trump%'::text)
11                     Rows Removed by Filter: 12222626
12             Planning time: 474.305 ms
13             Execution time: 774193.441 ms
14 Planning time: 3.717 ms
15 Execution time: 3191286.075 ms
```

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

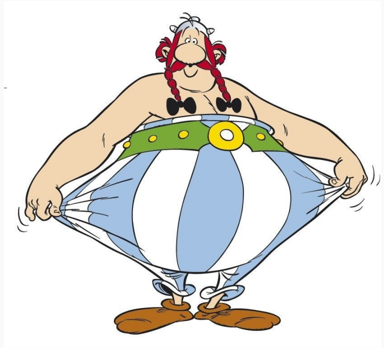
APACHE Spark

Flink 1.0

AWS RDS

DASK

And Many...



Why so many models and apps?

- In recent times, the database community has built many apps and techniques to handle various models and capabilities.(Like relational, semi-structured, or networks models and the capabilities like inverted index, data cube (Group by, CUBE by), centrality computation)
- These developments focused on a targeted vertical or solved domain-specific problems.
- Each of the apps developed and tuned best for its model and capabilities but incapable of the others.

PostgreSQL

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

neo4j

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

Solr

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

Polystore Database Systems

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 provided by its underlying storage engine.
- **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

1. BigDAWG, MIT
2. CloudMdsQL, Inria
3. Estocada, UCSD and Inria
4. Polypheny-DB, University of Basel, Switzerland
5. Awesome, UCSD(*)
6. Polystore++ , Stanford University
7. Polybase, Microsoft.
8. OoX, HP-Labs

The Awesome Polystore

Data Variety

Social Network

Newspaper

Contract and Legal

Clinical Trial Data

Organizational Data

SEMISTRUCTURED, VIDEO, TEXT

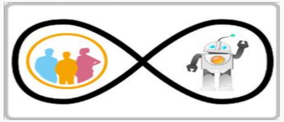
STRUCTURED TEXT, TIME, PLACES

TEXT

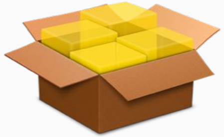
TEXT, IMAGE

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

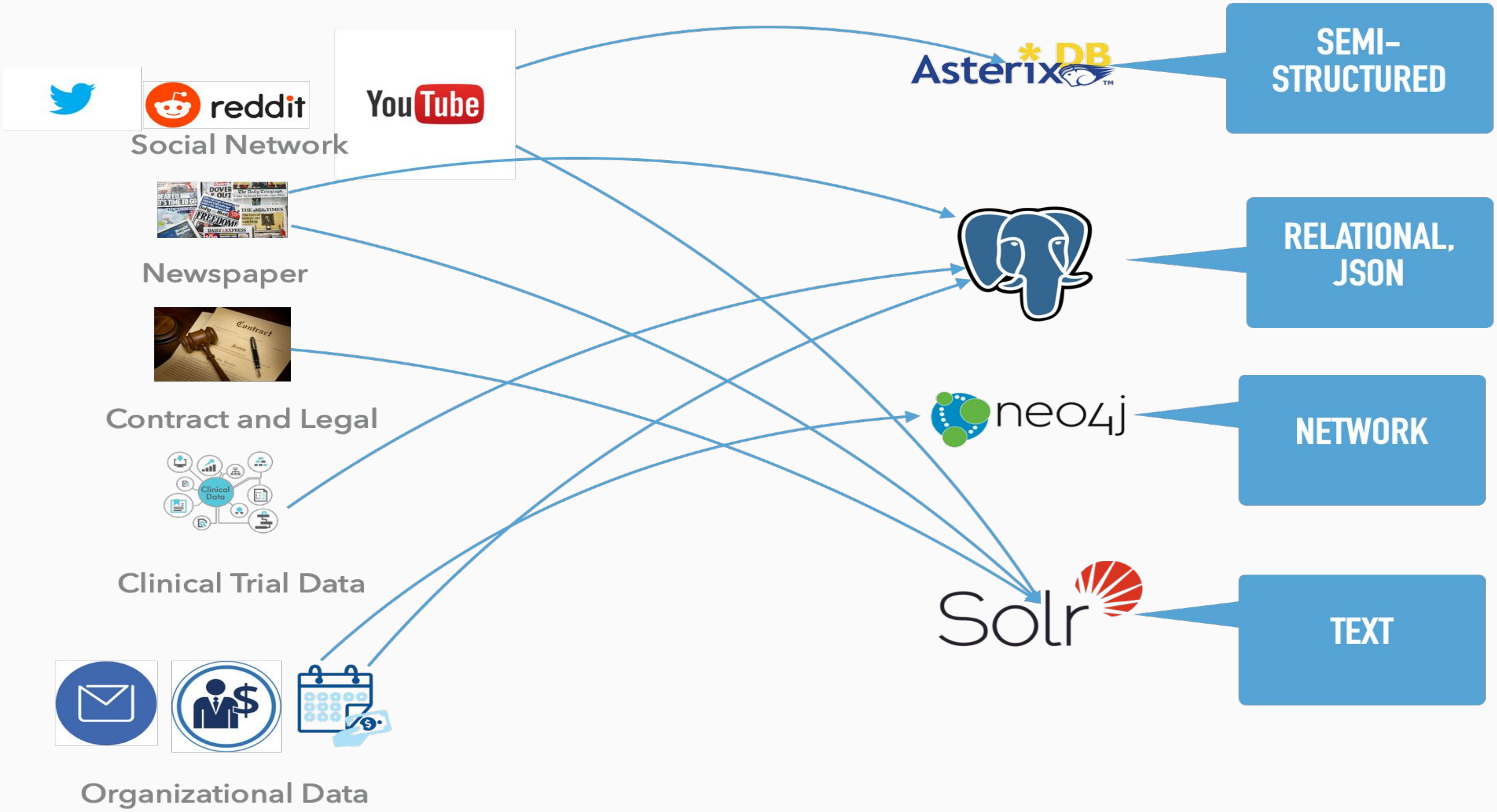
Dictionaries



HITL



Packages



Social Network
Newspaper
Contract and Legal
Clinical Trial Data
Organizational Data
SEMI-STRUCTURED
RELATIONAL, JSON
NETWORK
TEXT

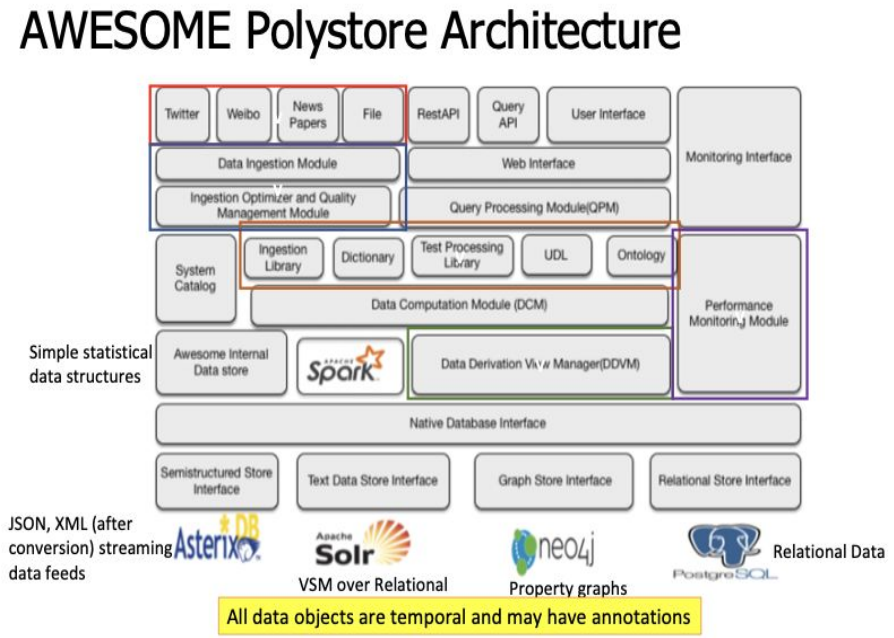
Social Sciences Questionnaire:

1. List all the accounts mentioned “Elizabeth Warren ” and also talking about “American DOS Movements.”
2. Find top 100 influential users those are co-spiking and talking about racial terms from the “Elizabeth Warren’s” network.
3. Find out top-k topics from the newspaper, those are also discussed in the “Elizabeth Warren’s” network.
4. 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 Case Study

Building a Knowledge Graph on the top of a polystore

The Problem

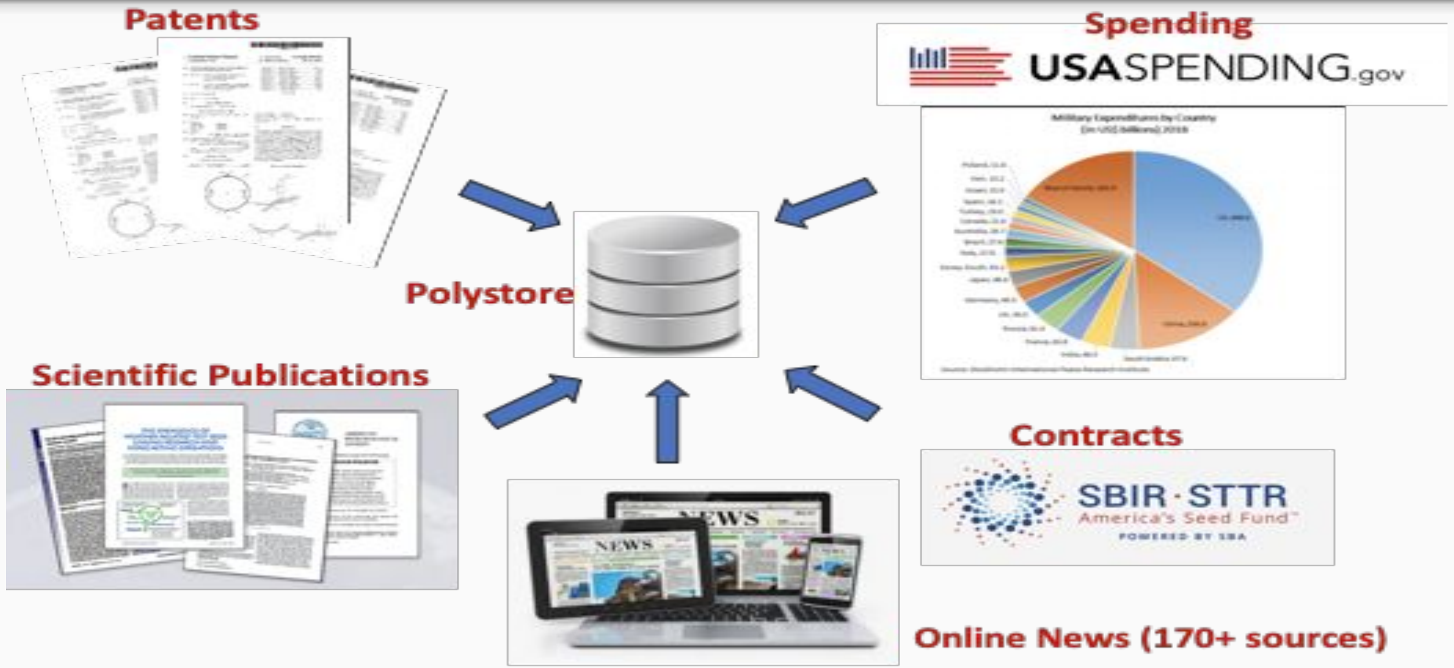
"Discover technology gaps in some domain and who can bridge the gaps?"

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..



Query on Knowledge Graph

Search

1. Expand the search term automatically using the ontology.
2. Return a data cube with various distributions.

Gap Analysis

1. Find the technological associations by looking at network structure
2. Find the differences or gap of those structural associations.

Opportunity Analysis

Find collaboration opportunities from the network associations and other data.

Demo

Thanks!

Contact me:

sudasgupta@ucsd.edu