



Prototype Open Knowledge Network

Proto-OKN Workshop @ KGC 2024

9:00 AM ET on Tuesday May 7, 2024

Chaitan Baru & Jemin George
Directorate for Technology, Innovation and Partnerships (TIP)
U.S. National Science Foundation

Workshop Agenda

□ Introduction

- *Chaitan Baru & Jemin George, TIP Directorate, National Science Foundation*

□ Presentation by Theme 1 Groups focusing on

○ Environment

- *Lilit Yeghiazarian, University of Cincinnati*

○ Biology & Health

- *Sergio, Baranzini, University of California, San Francisco (UCSF)*

○ Justice

- *Adam Pah, Georgia State University (GSU)*

○ Technology & Manufacturing

- *Farhad Ameri, Arizona State University (ASU)*

□ Presentation by Theme 2: Proto-OKN Fabric

- *Chris Bizon, University of North Carolina at Chapel Hill (UNC) & Patrick Grinaway, Onai*

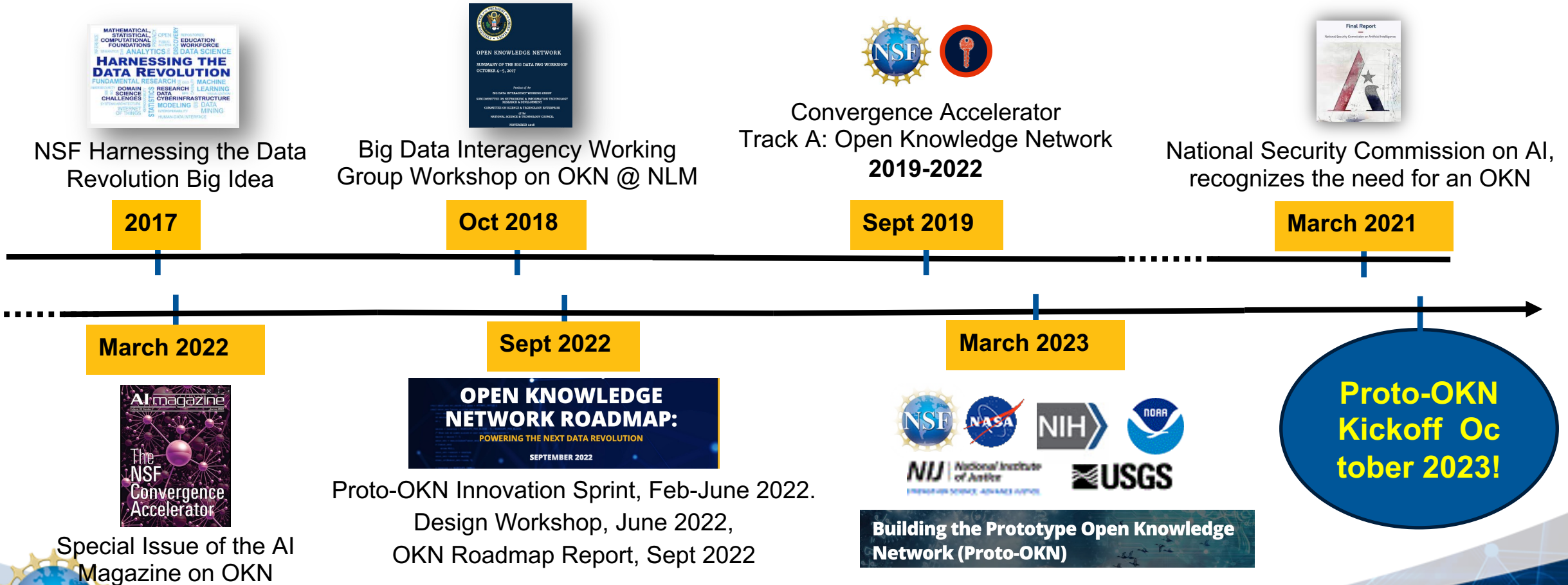
□ Presentation by Theme 3: Proto-OKN Education and Public Engagement

- *Cogan Shimizu, Wright State University*



The OKN Vision: Where we came from

An interconnected network of knowledge graphs build on public data to address various societal challenges.

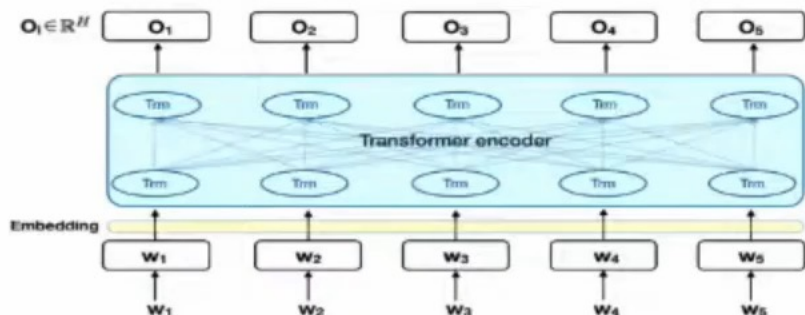


Three Current Revolutions in AI



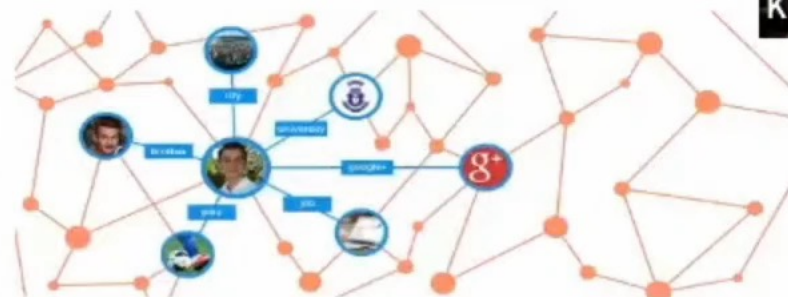
Kenneth D For...

Deep Learning



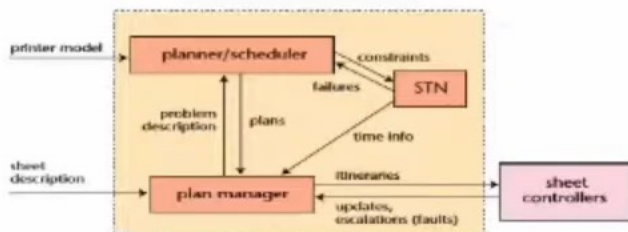
- With massive data, can do classification and other tasks in low-risk settings
- Uses: Speech recognition, facial recognition, etc.

Knowledge Graphs



- Massive (10^6 - 10^{10} fact) symbolic relational representations used by Google, Microsoft, Facebook, Spotify, etc.
- Uses: Higher-precision web search, web-scale question answering, better recommendations

Reasoning



- Finds solutions to large problems, finds flaws in engineered systems, real-time control, diagnosis, and workarounds

**Each revolution is:
Built on 4+ decades of AI research**

**Fueled by massive increases in
computational power and available data
over the last two decades**



Proto-OKN Program Goals

- Build a prototype version of an integrated data and knowledge infrastructure called the Open Knowledge Network (OKN)
- Create a platform that would empower government and non-government users — fueling evidence-based policymaking, continued strong economic growth, and game-changing scientific breakthroughs, while addressing complex societal challenges from climate change to social equity.

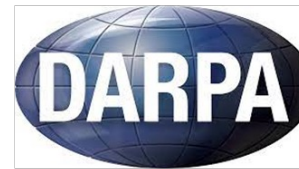


Proto-OKN Team: Multiagency Effort

- Core partner agencies



- Other stakeholder agencies



- AFRL
- DARPA
- CDO - DOC
- NCATS
- NIEHS
- NIJ
- NIST
- NREL
- Joint Staff J6
- USAFRICOM
- USDA
- VA

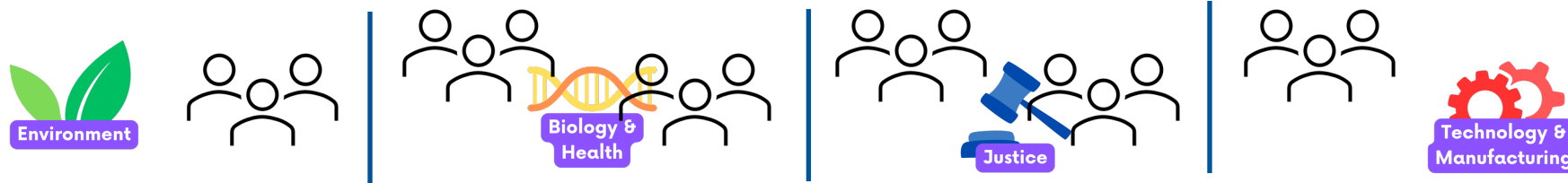
- Monthly sync-ups and coordinated outreach activities & sustainability talks



Proto-OKN Program Structure

Theme 1: Use Case Projects (15)

Focus on the creation of knowledge graphs to address **specific societal challenges**



Theme 2: Proto-OKN Fabric Projects (2)

Develop and deploy a platform to interconnect KGs and create a cloud-based infrastructure



Theme 3: Education and Outreach Project (1)

Develop education and outreach resources for a variety of stakeholders

All 18 projects are required to work together as a *single cohort*



Workshop Agenda

□ Introduction

- *Chaitan Baru & Jemin George, TIP Directorate, National Science Foundation*

□ Presentation by Theme 1 Groups focusing on

○ Environment

- *Lilit Yeghiazarian, University of Cincinnati*

○ Biology & Health

- *Sergio, Baranzini, University of California, San Francisco (UCSF)*

○ Justice

- *Adam Pah, Georgia State University (GSU)*

○ Technology & Manufacturing

- *Farhad Ameri, Arizona State University (ASU)*

□ Presentation by Theme 2: Proto-OKN Fabric

- *Chris Bizon, University of North Carolina at Chapel Hill (UNC) & Patrick Grinaway, Onai*

□ Presentation by Theme 3: Proto-OKN Education and Public Engagement

- *Cogan Shimizu, Wright State University*



Water-Energy Nexus OKN (WEN-OKN)

Lilit Yeghiazarian, U. Cincinnati

Wildlife Management (KN-Wildlife)

Xiangliang Zhang, Notre-Dame U.

Safe Agricultural Products and Water Graph (SAWGraph)

Torsten Hahmann, U. Maine

Soil Carbon Data Modeling (SOCKG)

Chengkai Li, U. Texas - Arlington

Knowledge Graph to Support Evaluation and Development of Climate Models

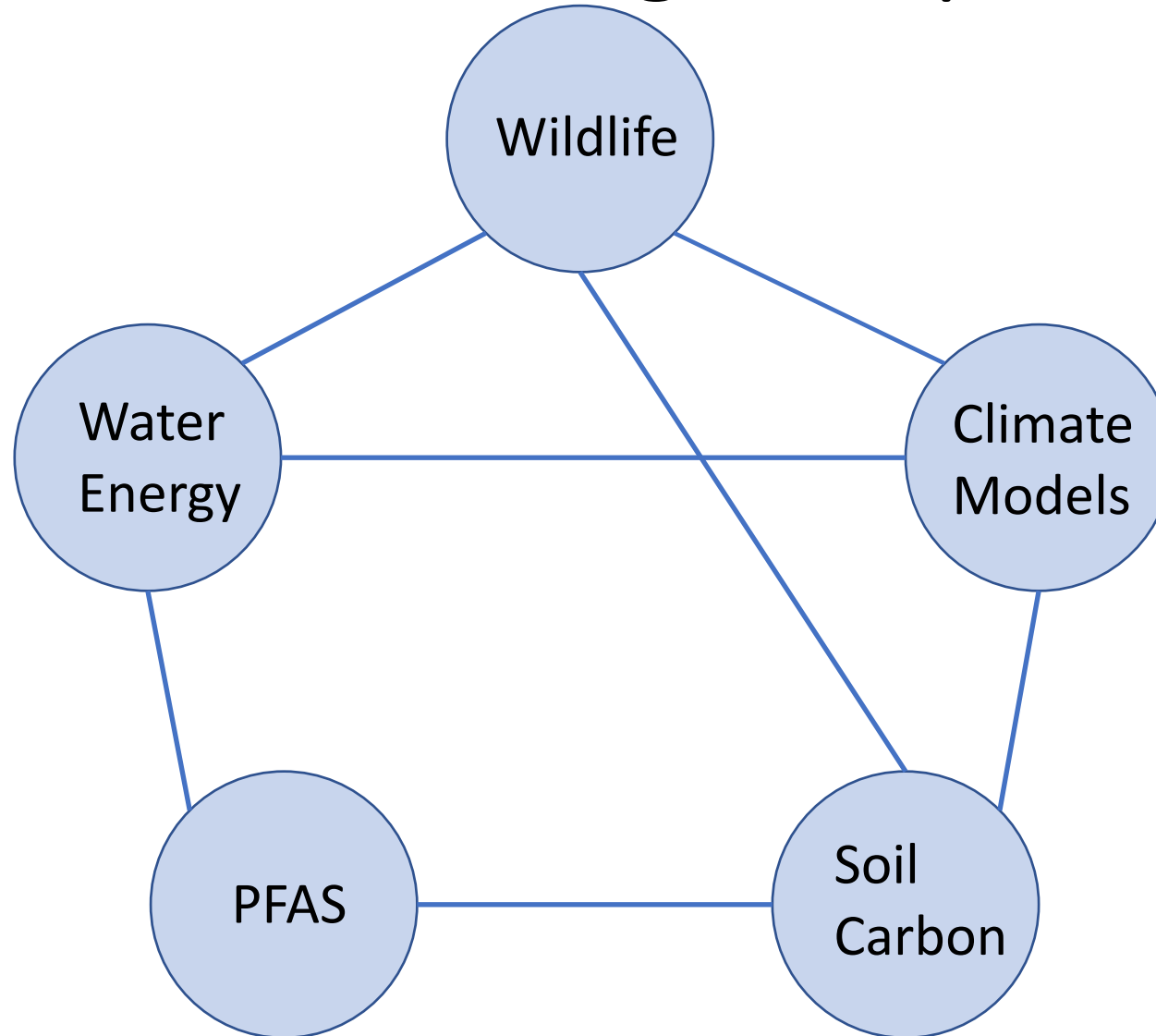
Eduard Dragut, Temple U.

Supported by:

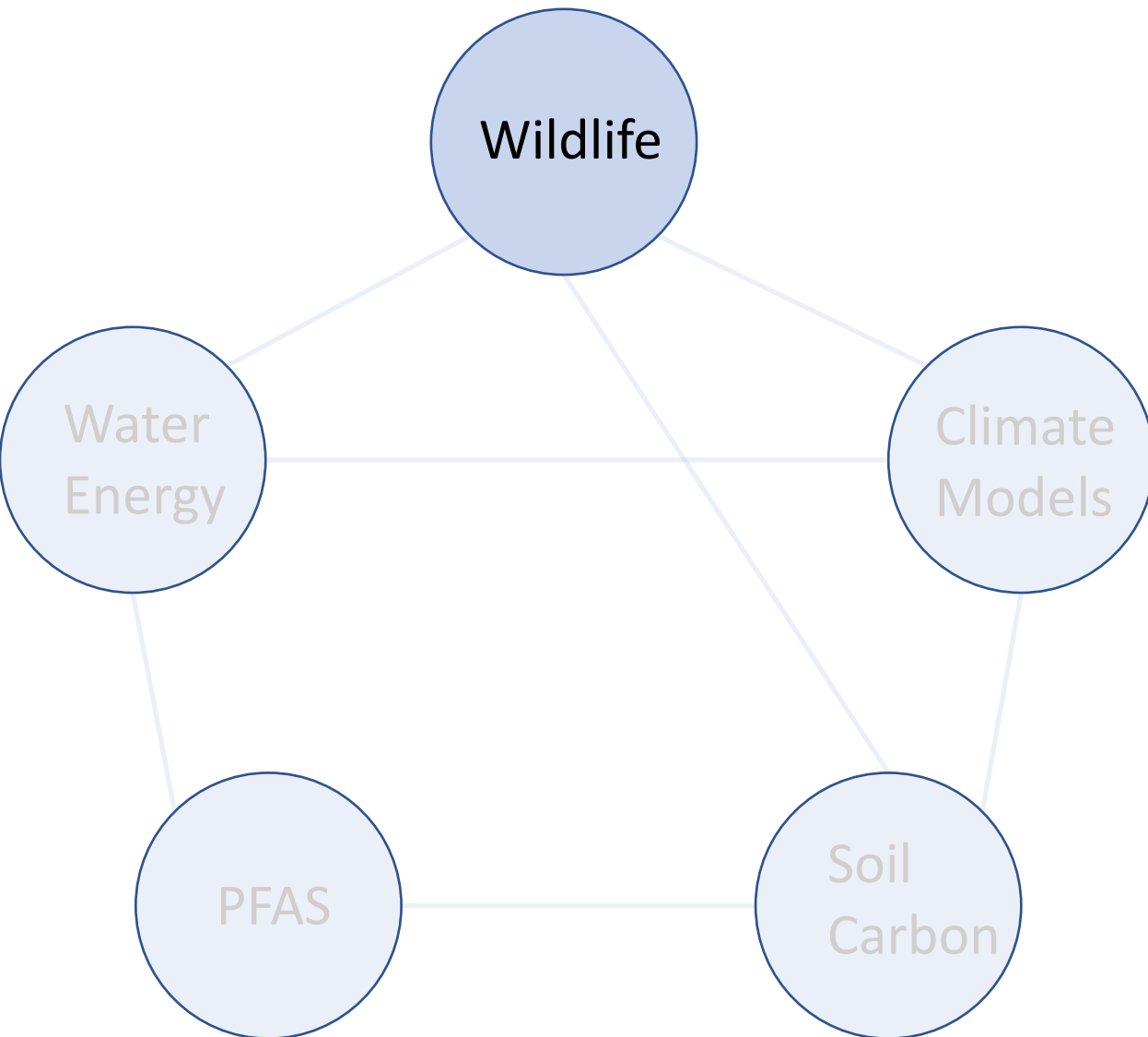


Proto-OKN

The Environment Working Group

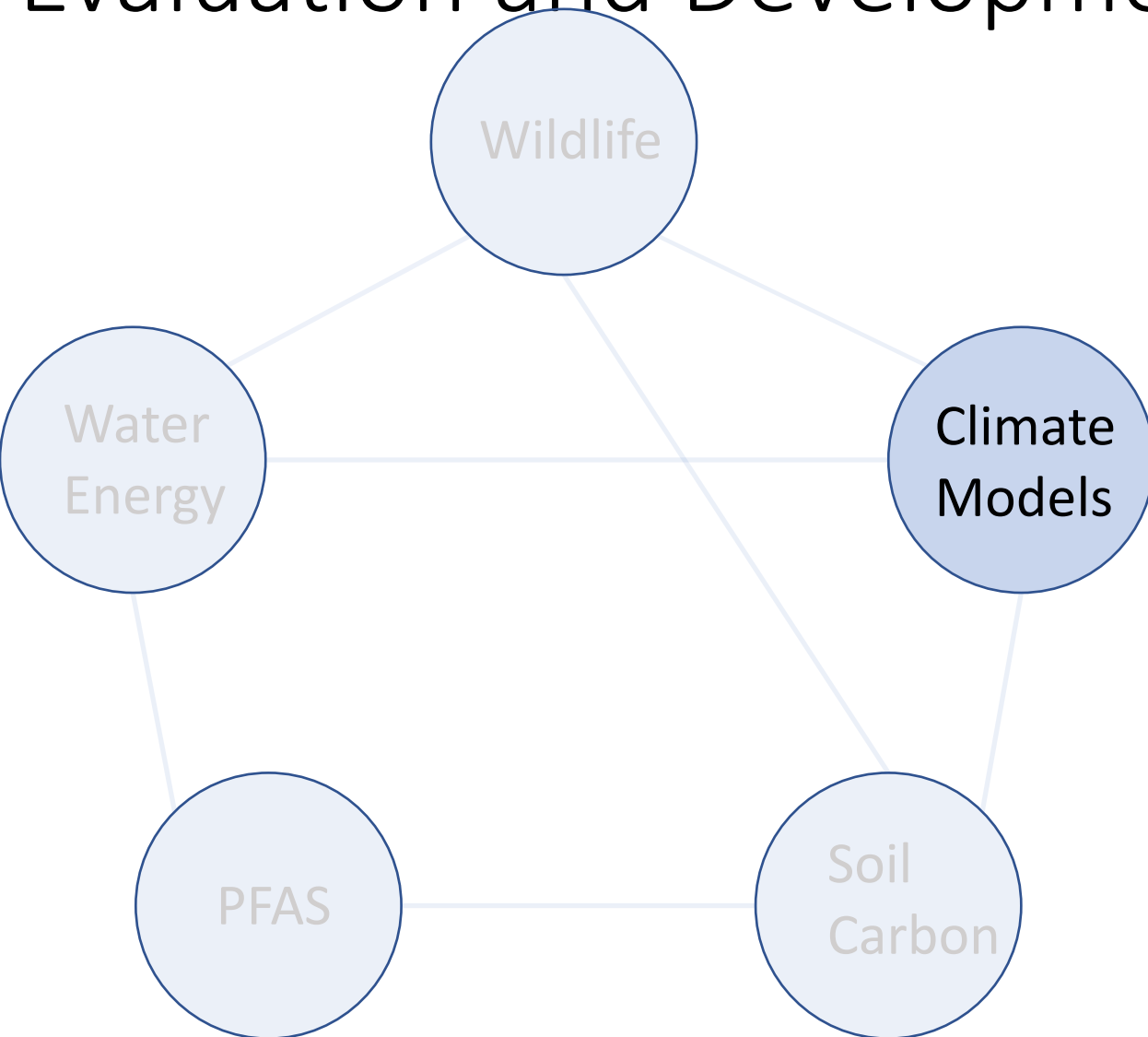


Wildlife Management OKN



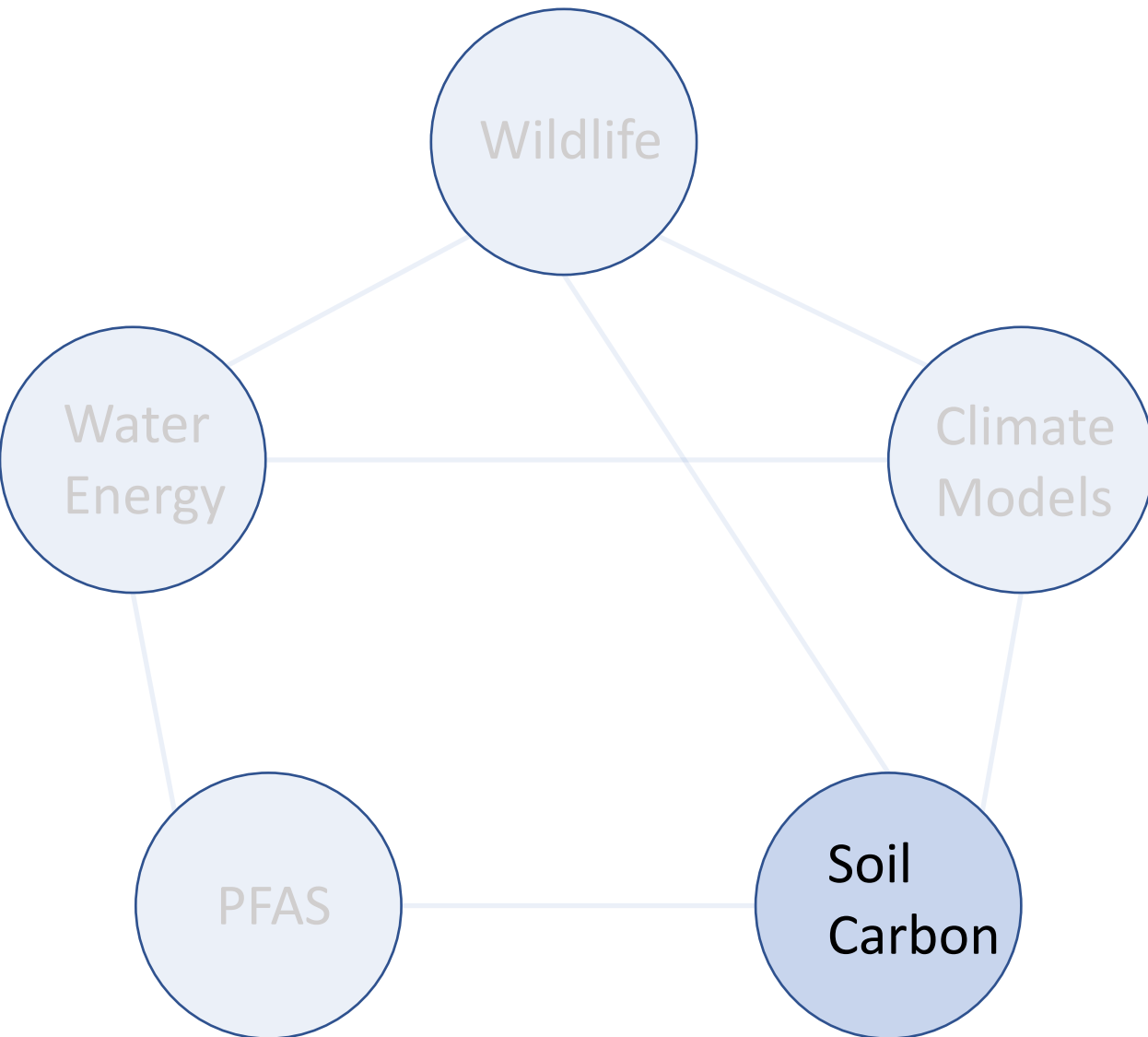
- **Objective:**
 - Support wildlife management in the context of climate change
 - Focus on invasive, threatened, and human-health-related species
- **Queries:**
 - Which threatened species in Florida are impacted by rising temperatures?
 - Which invasive species are relevant to the transmission of West Nile virus?
 - Predict the distribution of West Nile virus
- **Users:** Stakeholders for wildlife management, USGS, K-12 education
- **Data:** GBIF, IUCN Red list, NEON, USGS.

ClimatePub4KG: Knowledge Graph to Support Evaluation and Development of Climate Models



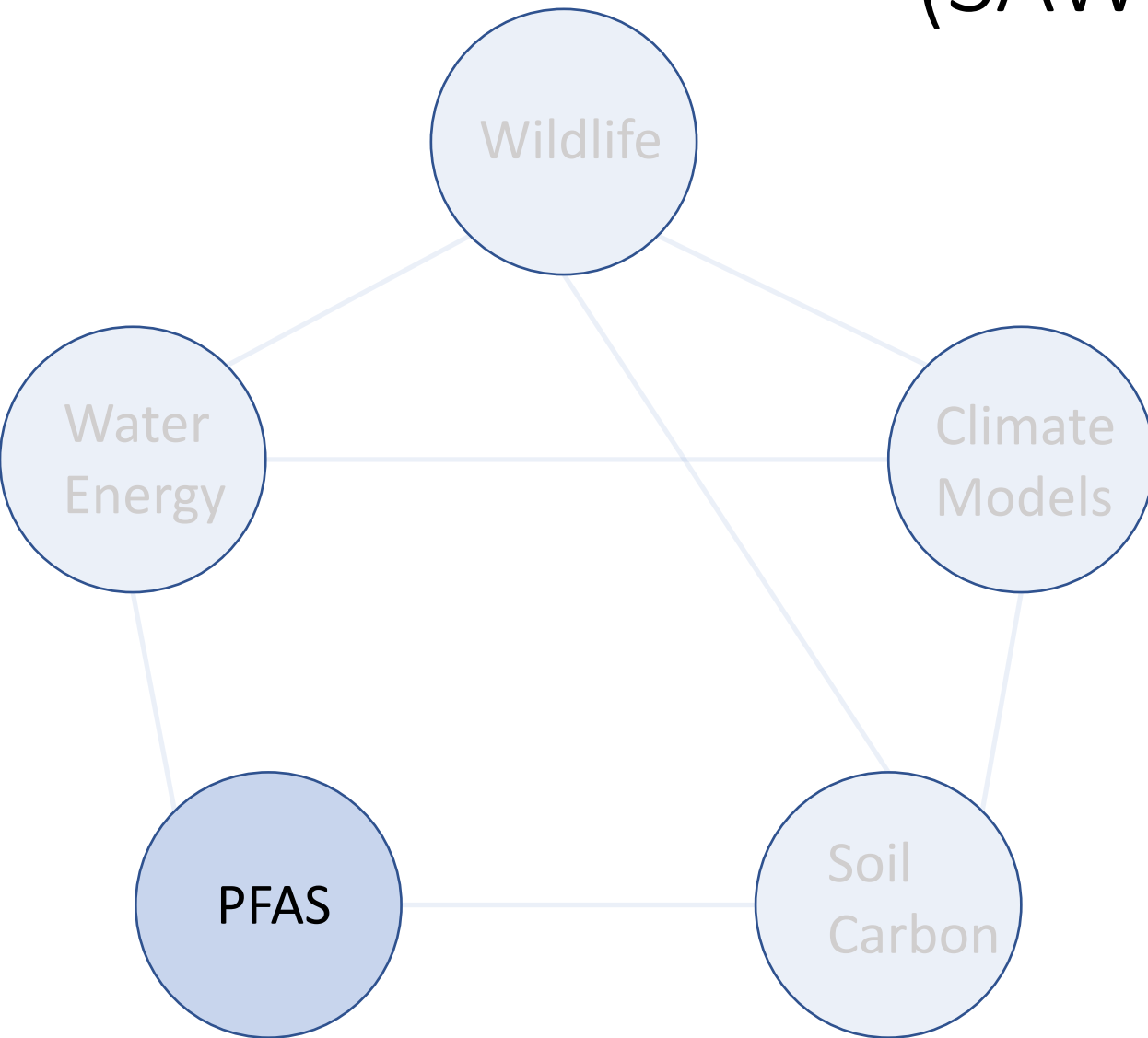
- **Objective:** Multimodal KG of salient aspects of climate modeling (data, mechanistic and AI climate models).
- **Queries:**
 - How is volcanic ash related to hydrology?
 - Which climate models are used to predict the ocean temperature on the east coast?
 - Which climate models are used for long-term rainfall predictions?
 - Which projects/tasks use a specific data and model?
- **Users:**
 - NOAA (National Oceanic and Atmospheric Administration)
 - CMIP (Coupled Model Intercomparison Project)
- **Data:** Climate scientific literature, GCMD (Global Change Master Directory)

Soil Carbon Data Modeling (SOCKG)



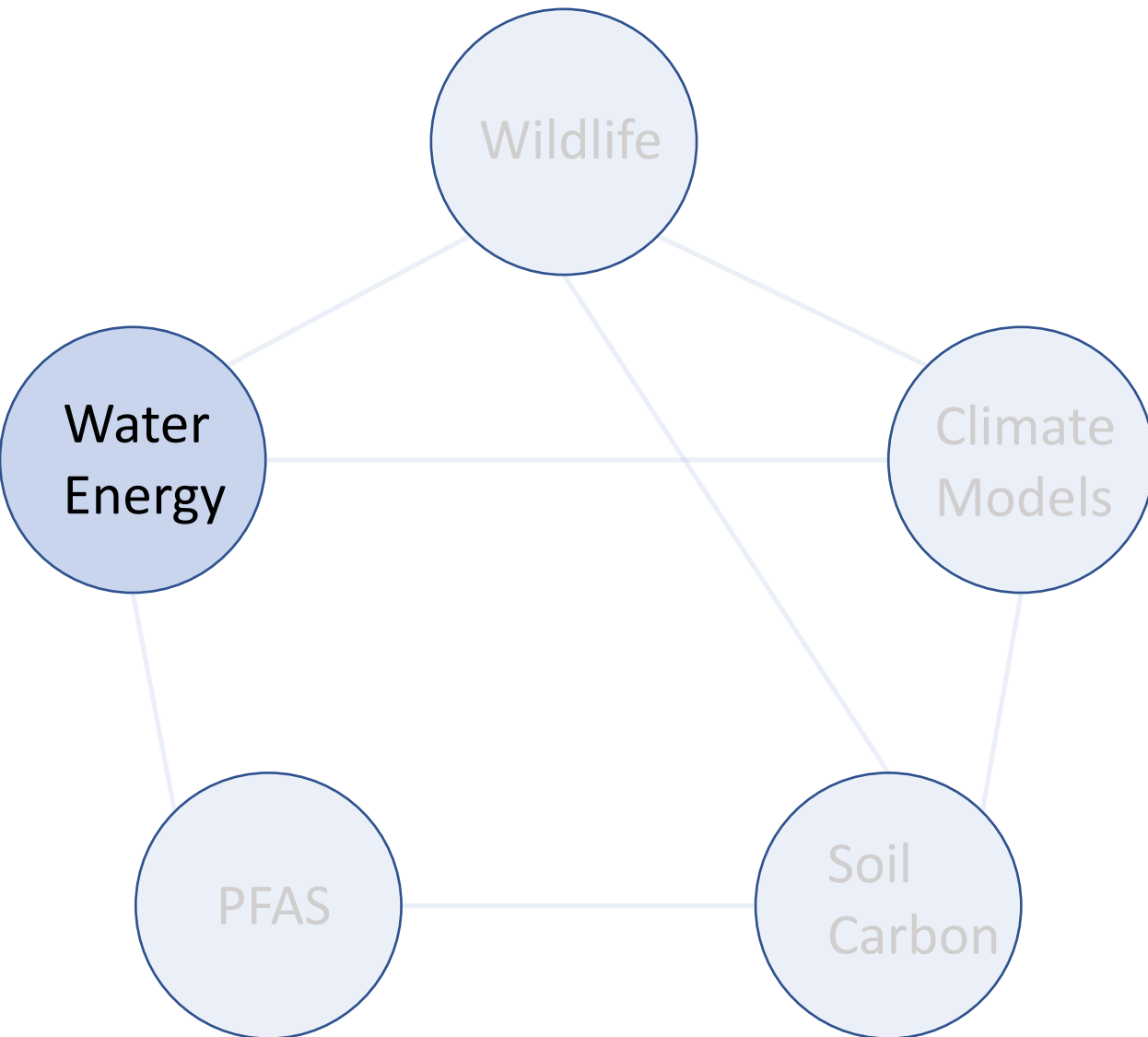
- **Objective:**
 - Provide access to soil organic carbon (SOC) data
 - Predict SOC change
 - Attribute that change to agricultural practices
- **Queries:**
 - Which management treatment results in the greatest amount of SOC storage?
 - What management combinations maximize SOC storage?
 - Is there a relationship between crop yield and SOC storage?
- **Users:** USDA scientists, policy makers, and modelers of soil carbon
- **Data:** Soil property measurements, rotation and management information from USDA experiments

Safe Agricultural Products and Water Graph (SAWGraph)



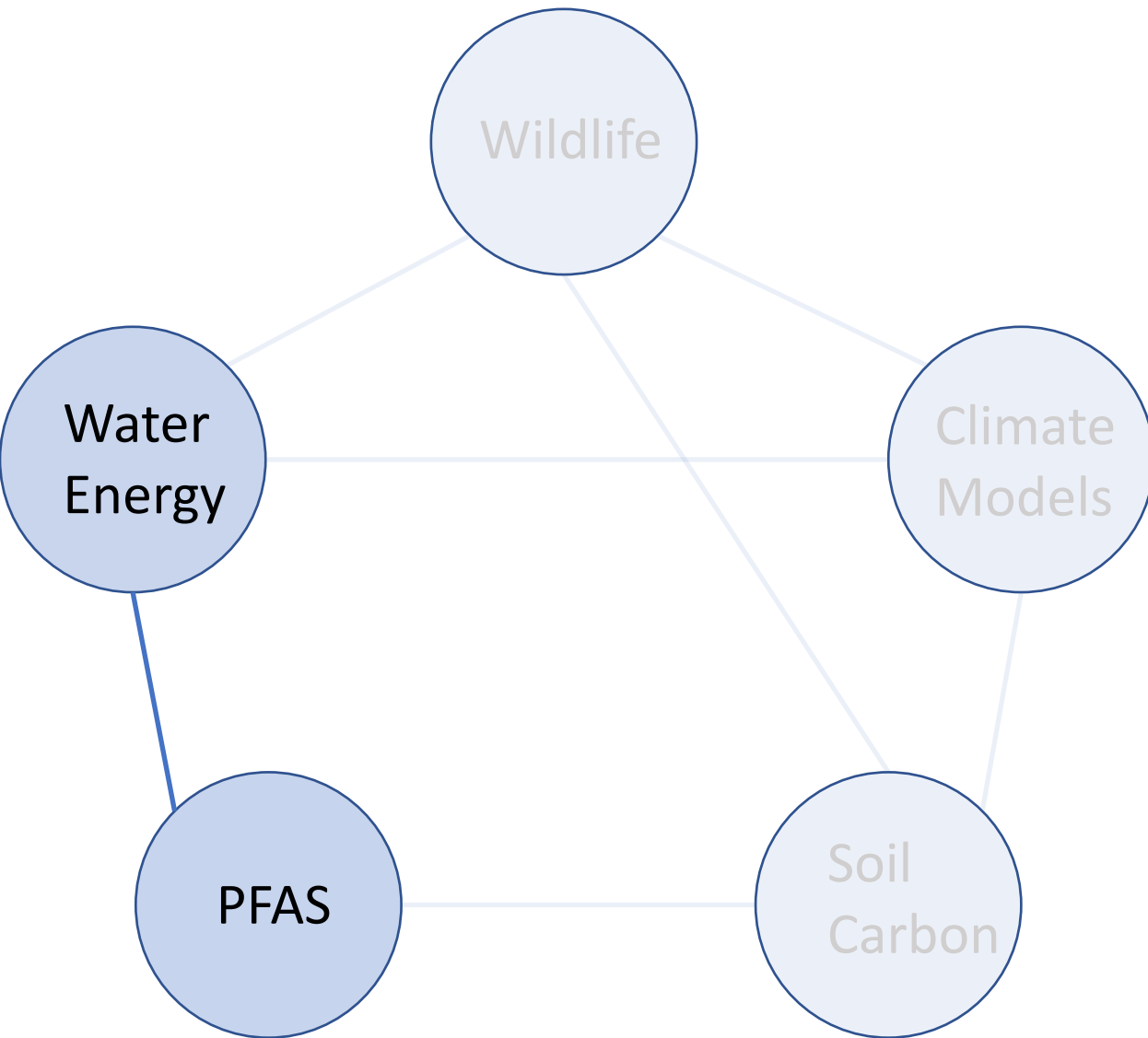
- **Objective:** One-stop information hub to answer questions around PFAS contamination
 - Connect test results, contamination impacts
 - Facilitate decision making around testing, regulations and remediation
- **Queries:**
 - Where have we tested?
 - Where are gaps in testing?
 - Who is impacted the most?
 - Where are known sources concentrated?
 - Where may the contamination in this well originate from?
- **Users:** Federal and state environmental protection and other agencies (USDA, FDA, USGS)
- **Data:** Contamination test results, locations from EPA, state agencies, USGS, FDA

Water-Energy Nexus OKN



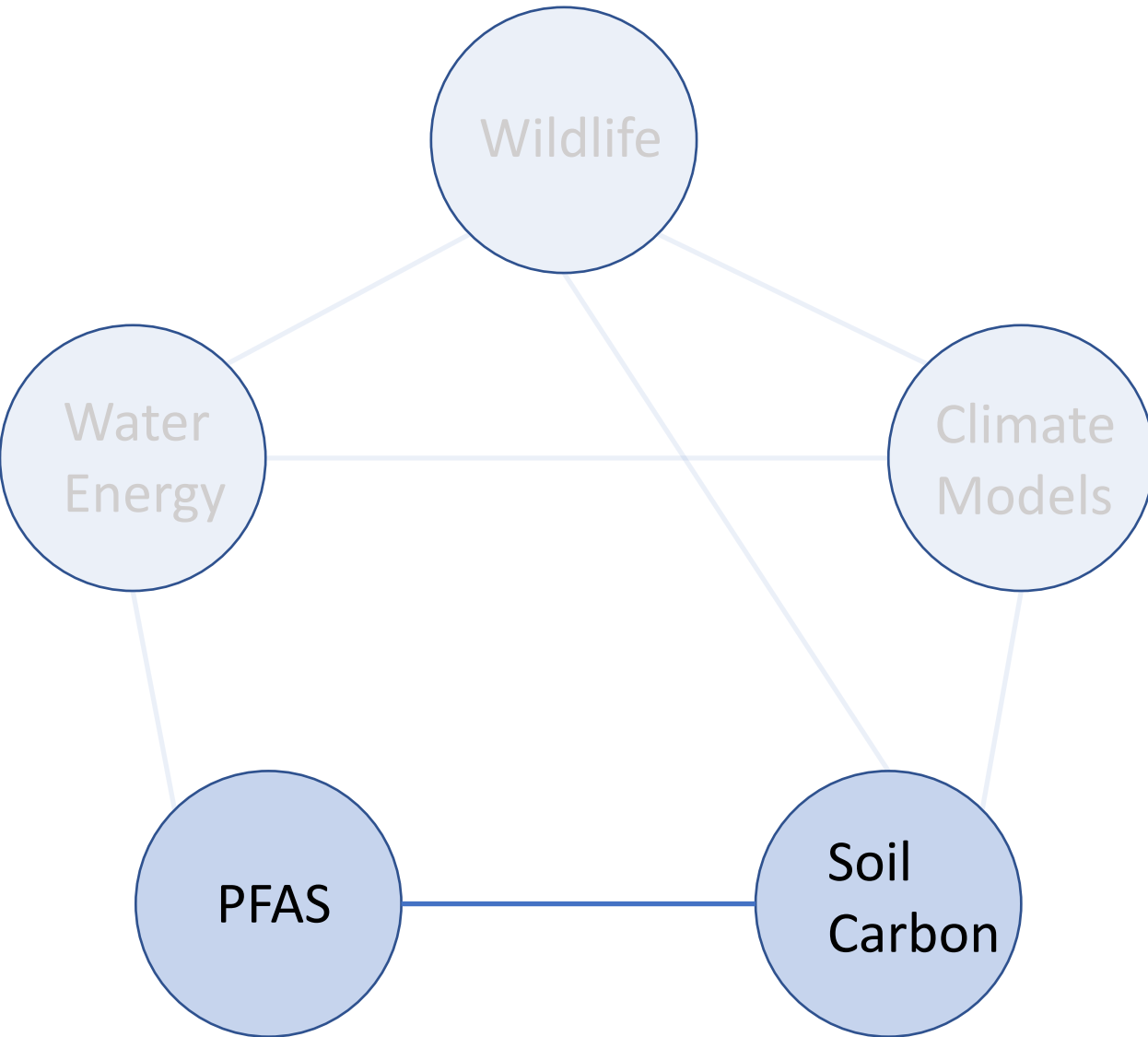
- **Objective:** Connect water-energy data to
 - Answer questions at the Nexus
 - Help align policies, rules, regulations
- **Queries:**
 - What are current and future threats to water availability for water and energy needs?
 - How much water and energy is being used, where and for what purpose?
 - What are the current and future threats to water and energy infrastructure?
 - Which rules/regulations support integrated management of water and energy nexus?
- **Users:** USGS, USACE, USEPA, DoE
- **Data:** USGS, NOAA, DoE

Water Energy and PFAS joint use case



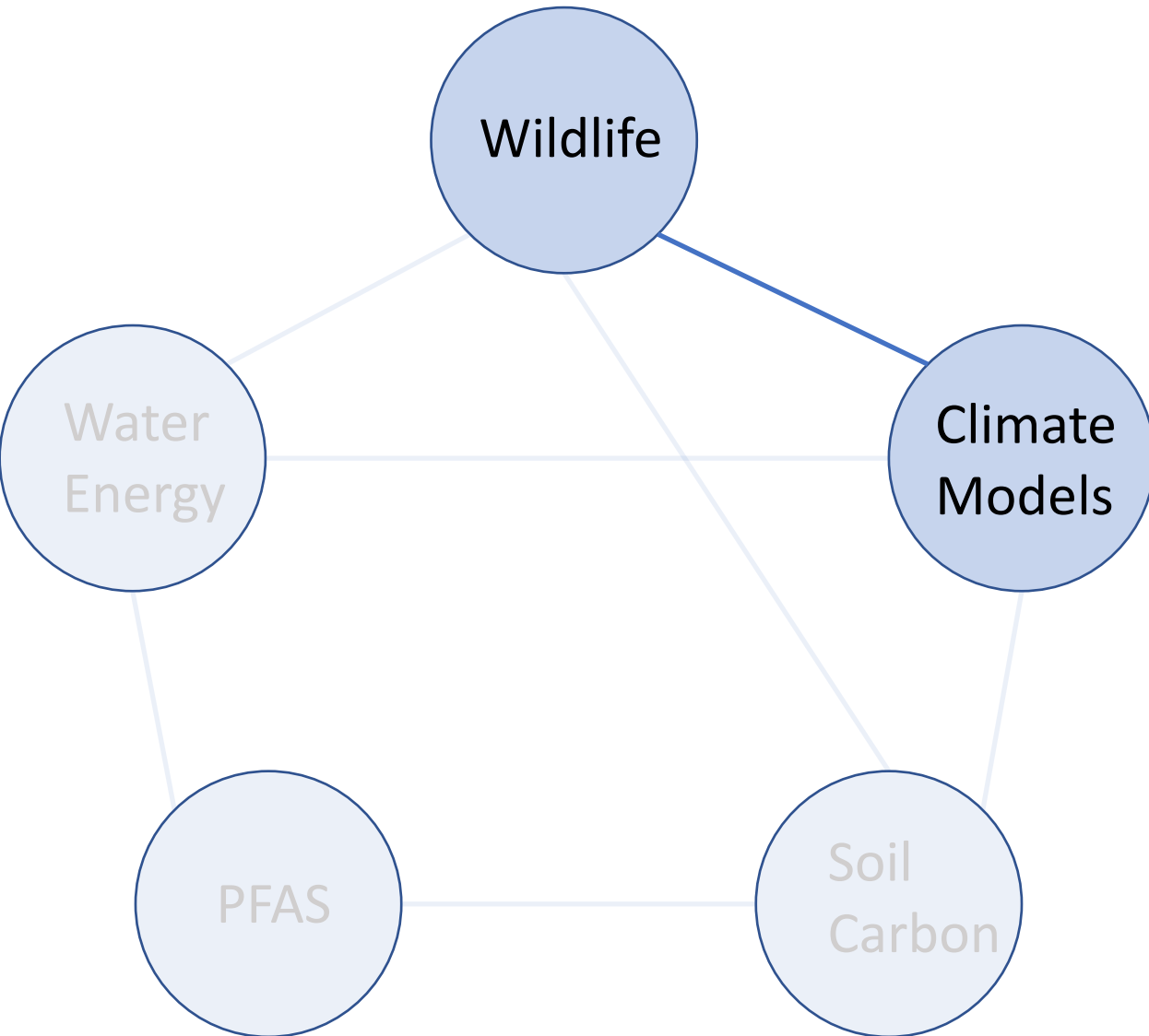
- Environmental Justice in Clean Water Availability (quantity and quality)
- Competency Questions
 - How much water is available to community X for specific use?
 - Is water contaminated in community X?
 - Is there *potential* for water contamination in community X?
 - Is remediation necessary? Where?
- Rank communities according to threat level and vulnerability
- Shared data and concepts
 - Federal facilities and industries
 - Hydrologic upstream-downstream relationships

Soil Carbon and PFAS joint use case



- Relationship between bioavailability of PFAS and soil carbon
 - PFAS binds to soil carbon, which reduces
 - PFAS uptake by crops
 - PFAS percolation into groundwater
- Shared Data and Concepts: Samples and measurements
 - Representation of spatial and temporal aspects
 - What is sampled (chemicals)
 - Metadata such as testing methods or lab information

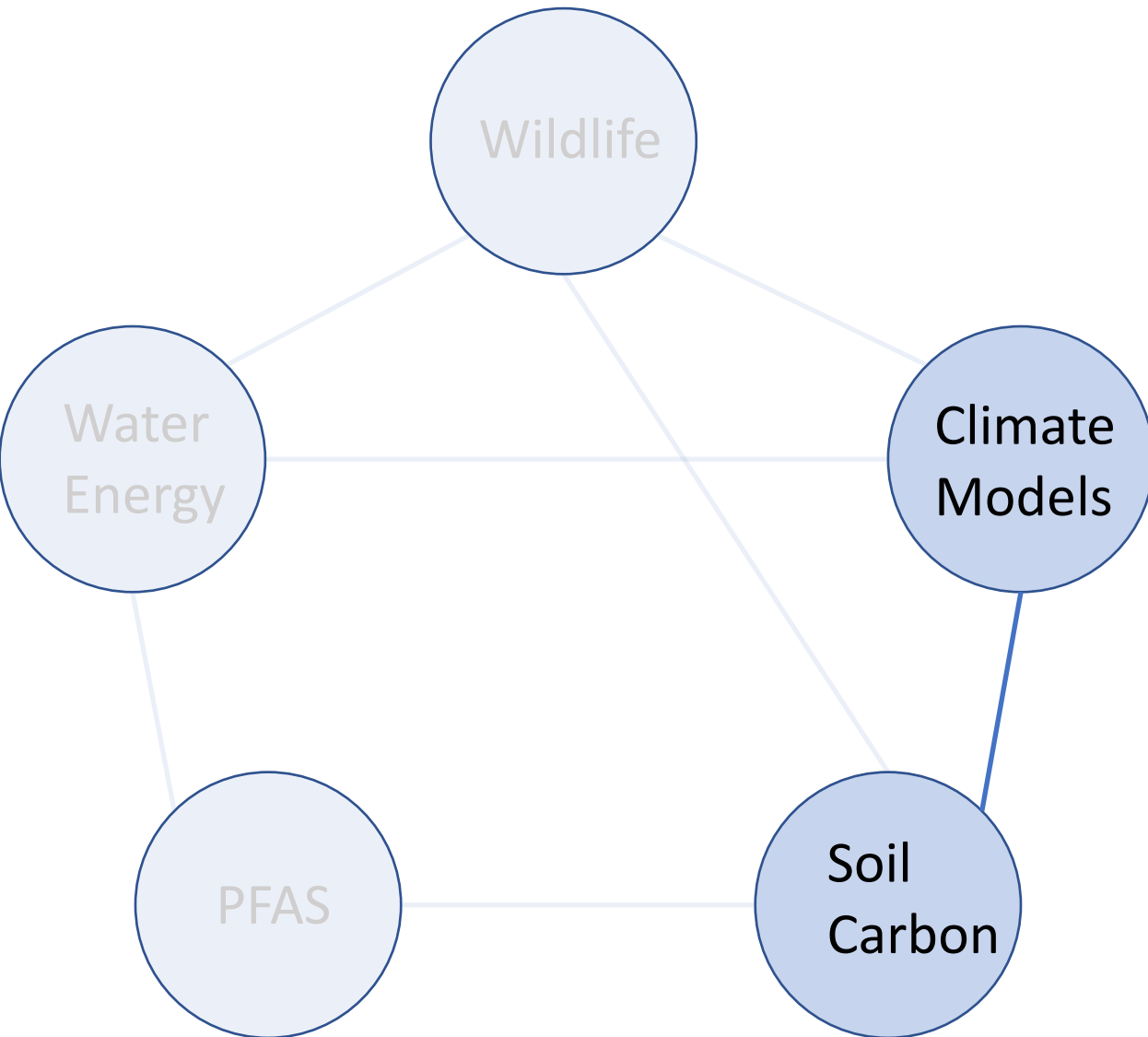
Climate models and Wildlife joint use case



○ Connection points

- How do changes in precipitation and temperature affect survival and distribution of species?
- What are the projected impacts of climate change on wildlife habitats and biodiversity hotspots?

Climate models and Soil carbon joint use case

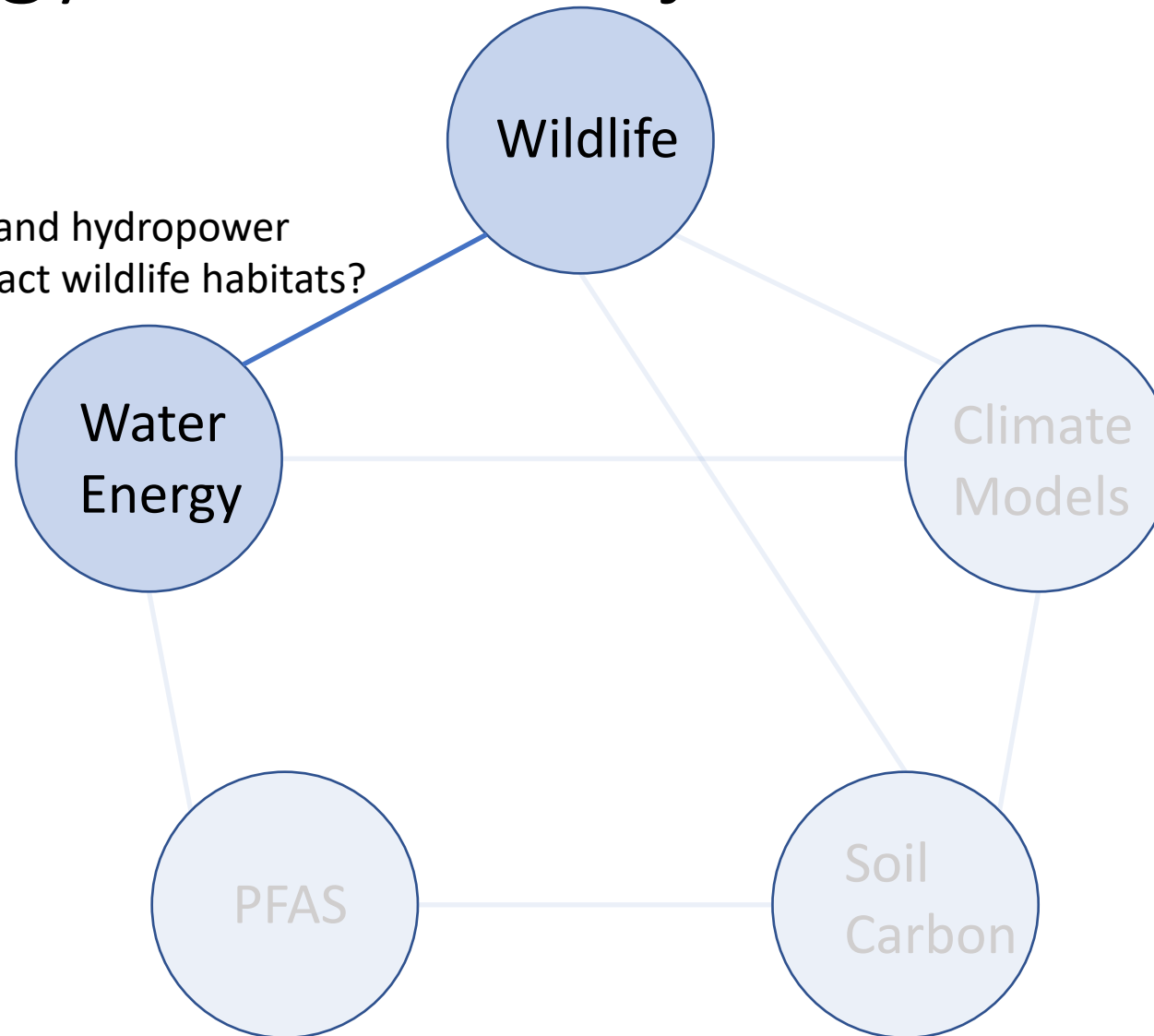


○ Connection points

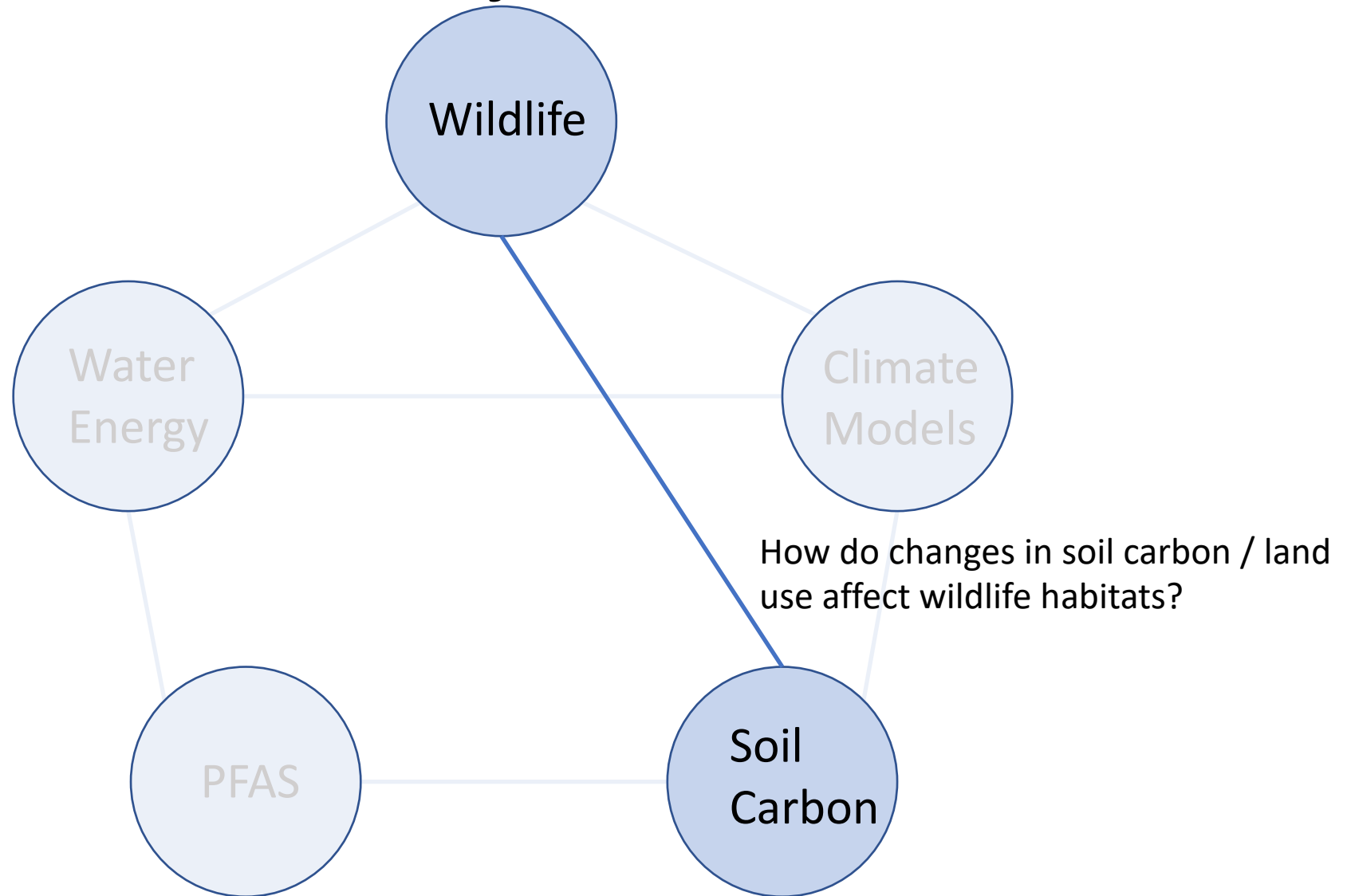
- How do climate variables and parameters from CMIP models correlate with soil carbon stocks and fluxes?
- Are there specific CMIP experiments or scenarios that directly impact soil carbon dynamics?
- What are the policy implications derived from the integration of climate model projections and soil carbon data, especially in terms of carbon credits quantification and sustainability farming practices?

Water Energy and Wildlife joint use case

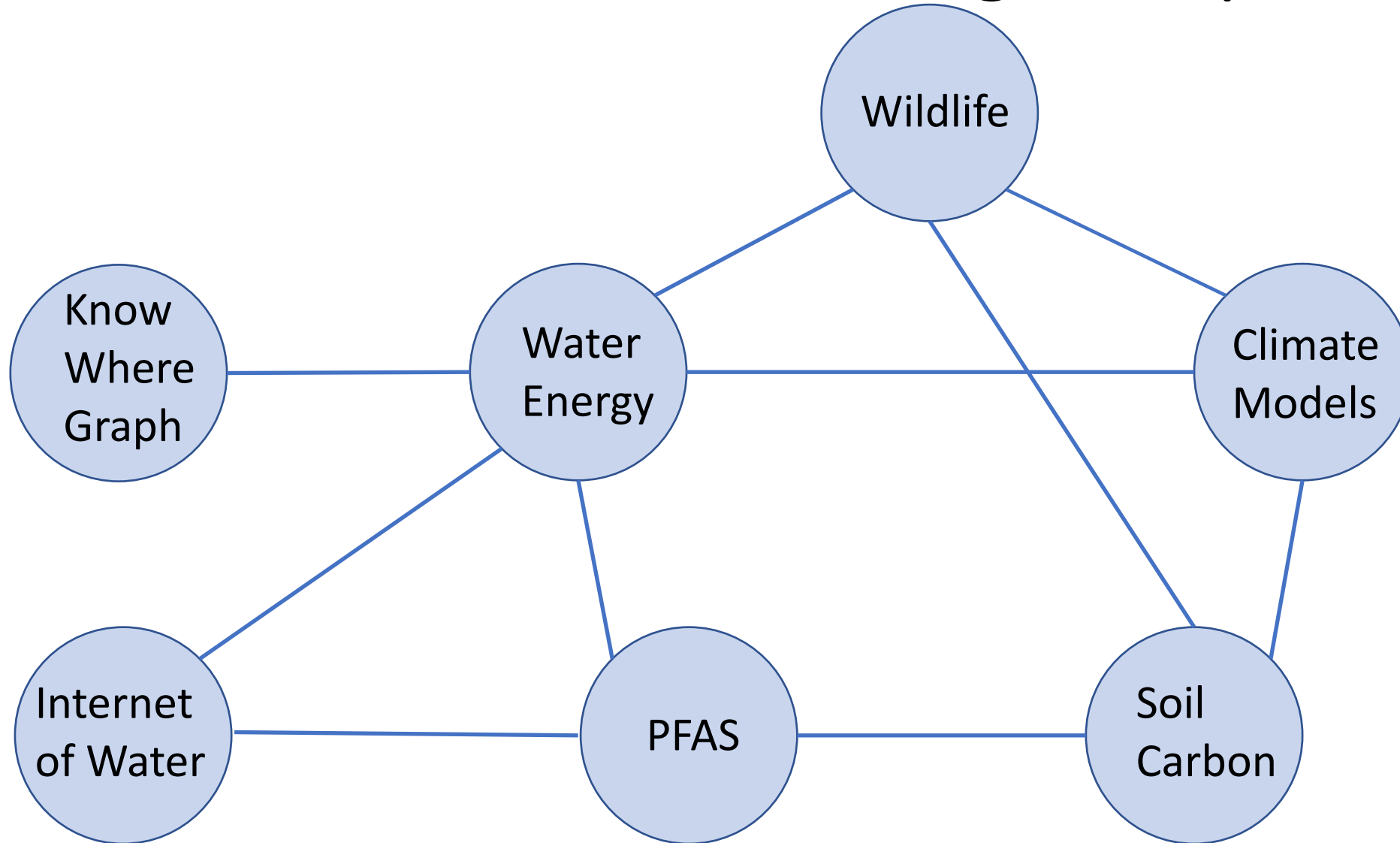
How do floods and hydropower generation impact wildlife habitats?



Soil Carbon and Wildlife joint use case



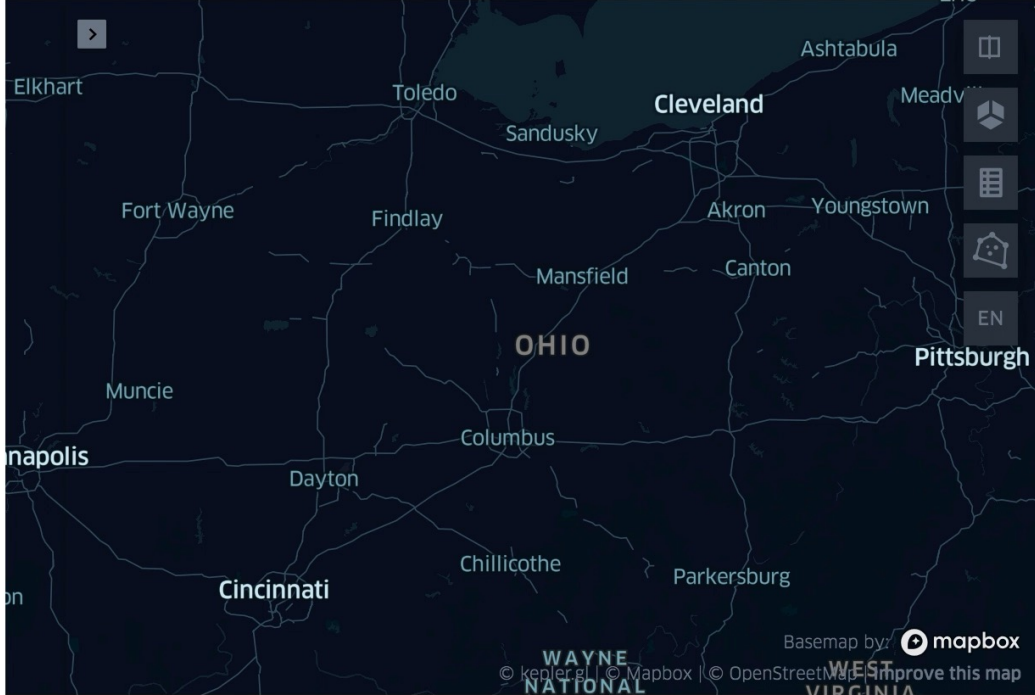
The Environment Working Group +



Demonstration of federated GeoSparql queries

Share ☆ 🔄 ⋮

WEN-OKN: Dive into Data, Never Easier



The map displays the state of Ohio with various cities labeled, including Toledo, Cleveland, Columbus, Cincinnati, and Pittsburgh. A search bar on the right contains the text "What can I help you with?". The map is powered by Mapbox, as indicated by the "Basemap by mapbox" logo at the bottom.

What can I help you with? ➤

Manage app

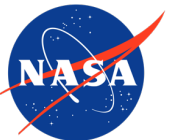
Biology and Health

Project 1: SPOKE for space (S. Baranzini)

Project 2: Bio-Health-OKN (A. Zhang)

Project 3: Biobricks (T. Luechtefeld)

Supported by:



Biology &
Health



Proto-OKN

Project 1: Connecting Biomedical information on Earth and in Space via the SPOKE knowledge graph

Sergio Baranzini, Ph.D.
Professor of Neurology
University of California, San Francisco

Supported by:



Biology &
Health



Proto-OKN



Team Composition

PI: Sergio Baranzini **UCSF**

Peter Rose	<u>UC San Diego</u>
Sui Huang	 ISB
Karthik Soman	UCSF
Ebru Akbas	UCSF
Scooter Morris	UCSF
Sylvain Costes	
Lauren Sanders	
Sam Gebre	 UCSF
Aenor Sawyer	
Charlotte Nelson	MATE  BIOSERVICES





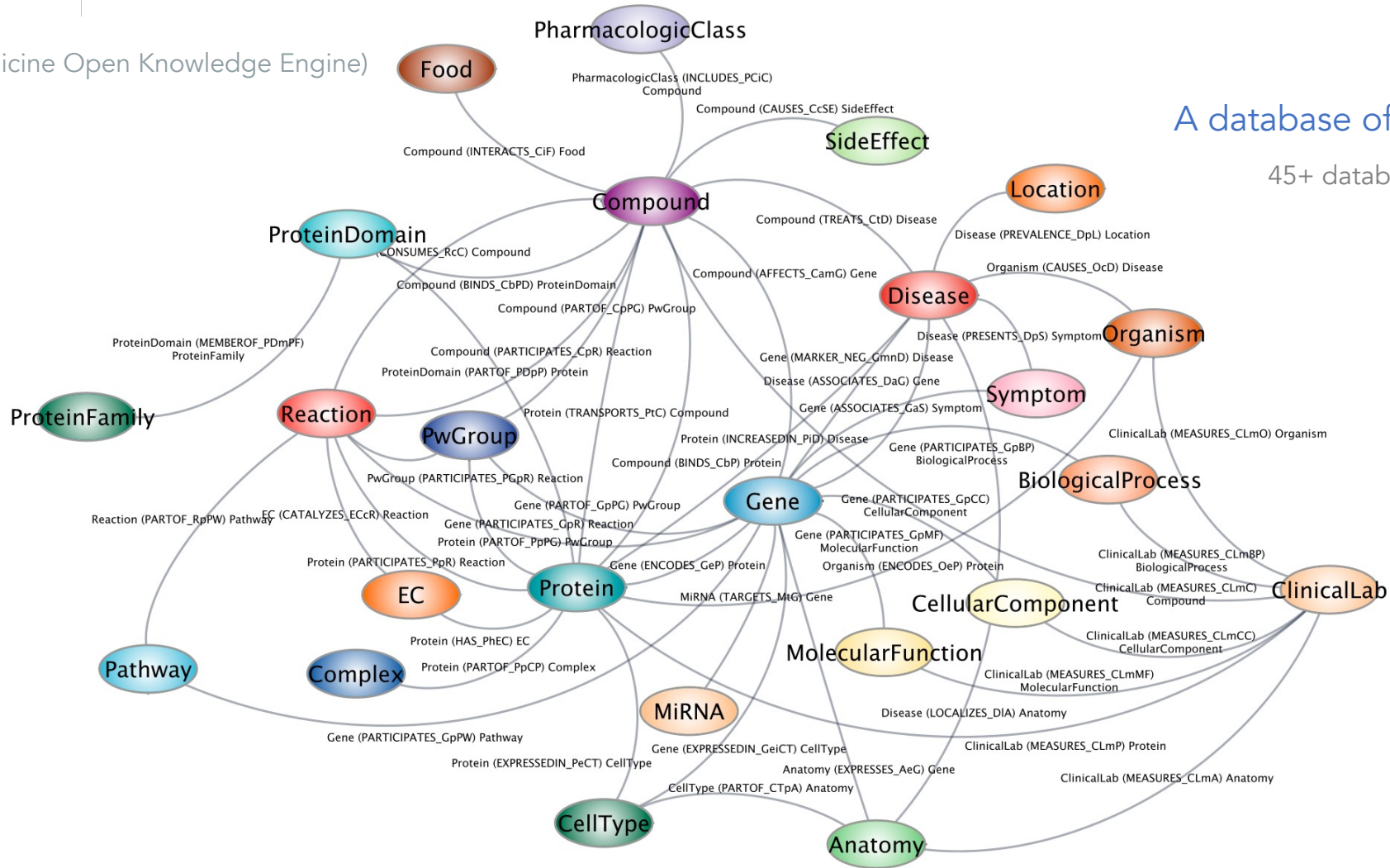
SPOKE : A biomedical knowledge graph



(Scalable Precision Medicine Open Knowledge Engine)

A database of databases

45+ databases



Proto-OKN

Current SPOKE Ecosystem





Early detection of Parkinson's disease through enriching the electronic health record using a biomedical knowledge graph

Karthik Soman¹, Charlotte A. Nelson¹, Gabriel Ceroni¹, Samuel M. Goldman², Sergio E. Baranzini¹ and Ethan G. Brown^{3*}

¹Department of Neurology, Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, United States, ²Division of Occupational and Environmental Medicine, University of California, San Francisco, San Francisco, CA, United States

OPEN ACCESS

EDITED BY
Surapaneni Krishna Mohan,
Panimalar Medical College Hospital and
Research Institute, India

REVIEWED BY
Nurlan Dauletbayev,
McGill University, Canada
Nuno Jorge Lamas,
University of Minho, Portugal

*CORRESPONDENCE
Ethan G. Brown
✉ ethan.brown@ucsf.edu

Journal of the American Medical Informatics Association, 29(3), 2022, 424–434
<https://doi.org/10.1093/jamia/ocab270>
Advance Access Publication Date: 16 December 2021
Research and Applications



OXFORD

Research and Applications

Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis

Charlotte A. Nelson^{1,2}, Riley Bove³, Atul J. Butte^{2,4}, and Sergio E. Baranzini^{1,2,3}

¹Integrated Program in Quantitative Biology, University of California San Francisco, San Francisco, California, USA, ²Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, California, USA, ³Department of Neurology, UCSF Weill Institute for Neurosciences, University of California San Francisco, San Francisco, California, USA, and ⁴Department of Pediatrics, University of California San Francisco, San Francisco, California, USA



Article

Knowledge Network Embedding of Transcriptomic Data from Spaceflown Mice Uncovers Signs and Symptoms Associated with Terrestrial Diseases

Charlotte A. Nelson¹, Ana Uriarte Acuna^{2,3}, Amber M. Paul^{2,4}, Ryan T. Scott^{2,3}, Atul J. Butte^{5,6}, Egle Cekanaviciute², Sergio E. Baranzini^{1,5,7,*} and Sylvain V. Costes^{2,*}



Proto-OKN Project overview

GeneLab



SDOH



Proto-OKN



Open Science Data Repository



Open Science for Life in Space

[Home](#)

[About](#) ▾

[Data & Tools](#) ▾

[Research & Resources](#) ▾

[Working Groups](#) ▾

[Help](#) ▾

General Search Filters

Data Source

- GeneLab
- ALSDA
- NIH GEO
- EBI PRIDE
- ANL MG-RAST

Data Type

- Study
- Experiment
- Subject
- Biospecimen
- Payload
- Mission
- Hardware
- Vehicle

Study Search Filters

Project Type

- Ground
- Spaceflight
- High Altitude

Open Science Data Repository Search



Sort By: [Accession \(Ascending\)](#) ▾

Items per page: [25](#) ▾

1 - 25 of 449



Expression data from drosophila melanogaster

Study
OSD-1

Organisms	Factors	Assay Types	Release Date	Description
Drosophila melanogaster	Spaceflight Infection	transcription profiling	11-Dec-2013	Space travel presents unlimited opportunities for exploration and discovery, but requires a more complete understanding of the immunological consequences of long-term exposure to the conditions of spa...

Highlights: *cgene*



Rodent Research-1 (RR1) NASA Validation Flight: Mouse eye transcriptomic and epigenomic data

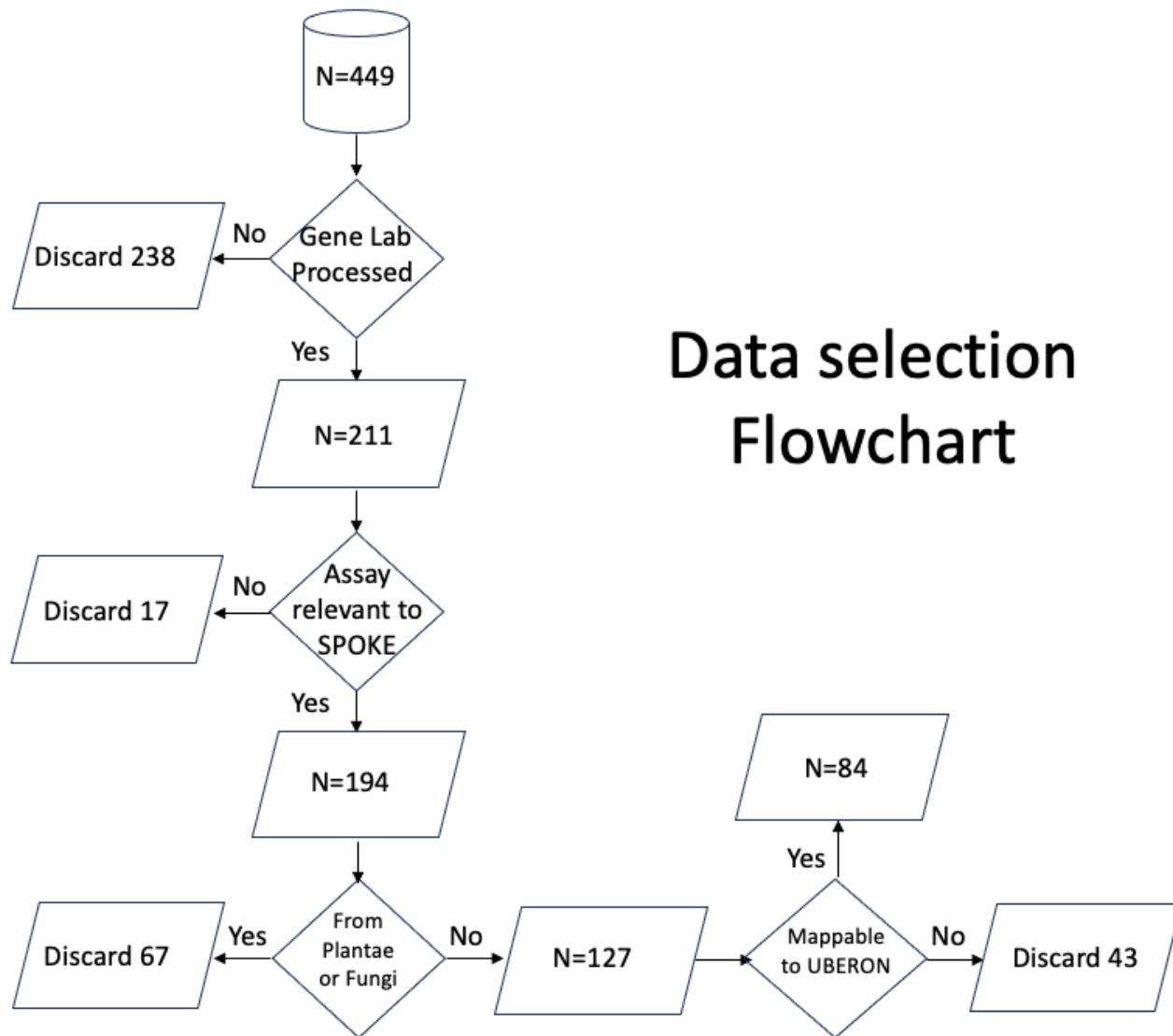
Study
OSD-100

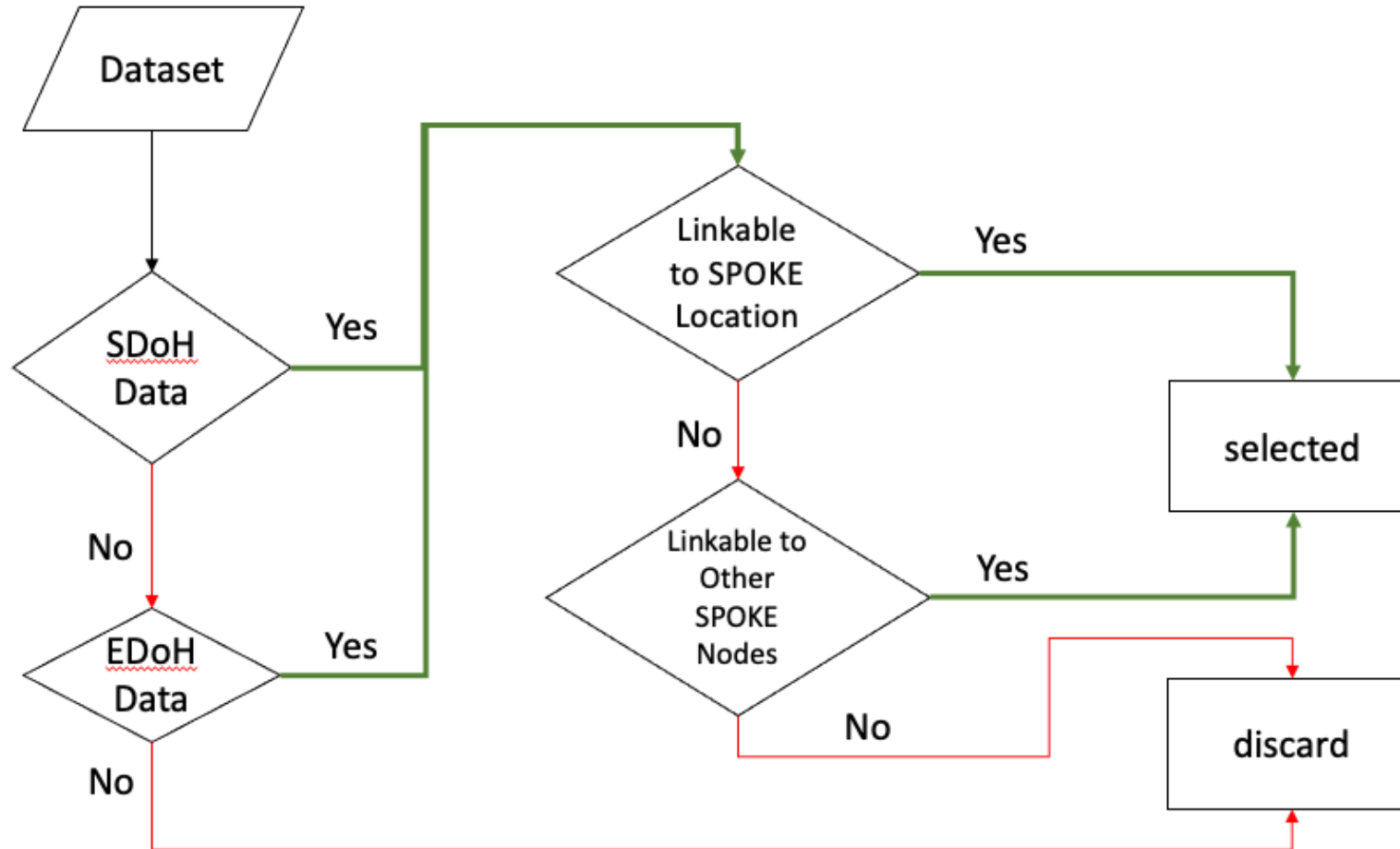
Organisms	Factors	Assay Types	Release Date	Description
Mus musculus	Spaceflight	DNA methylation profiling transcription profiling	28-Feb-2017	NASA's Rodent Research (RR) project is playing a critical role in advancing biomedical research on the physiological effects of space environments. Due to the limited resources for conducting biologic...

Highlights: *cgene*



Proto-OKN

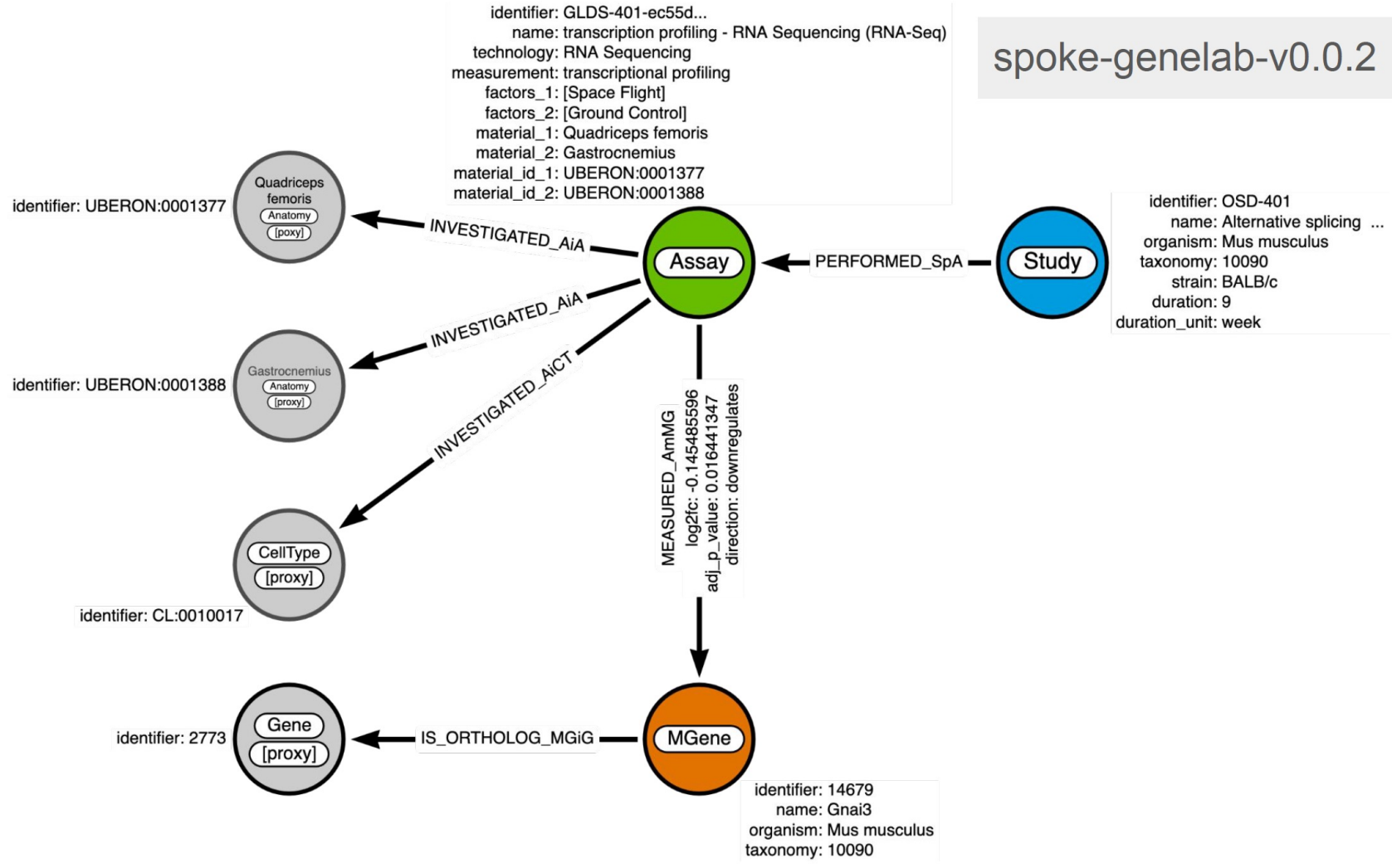




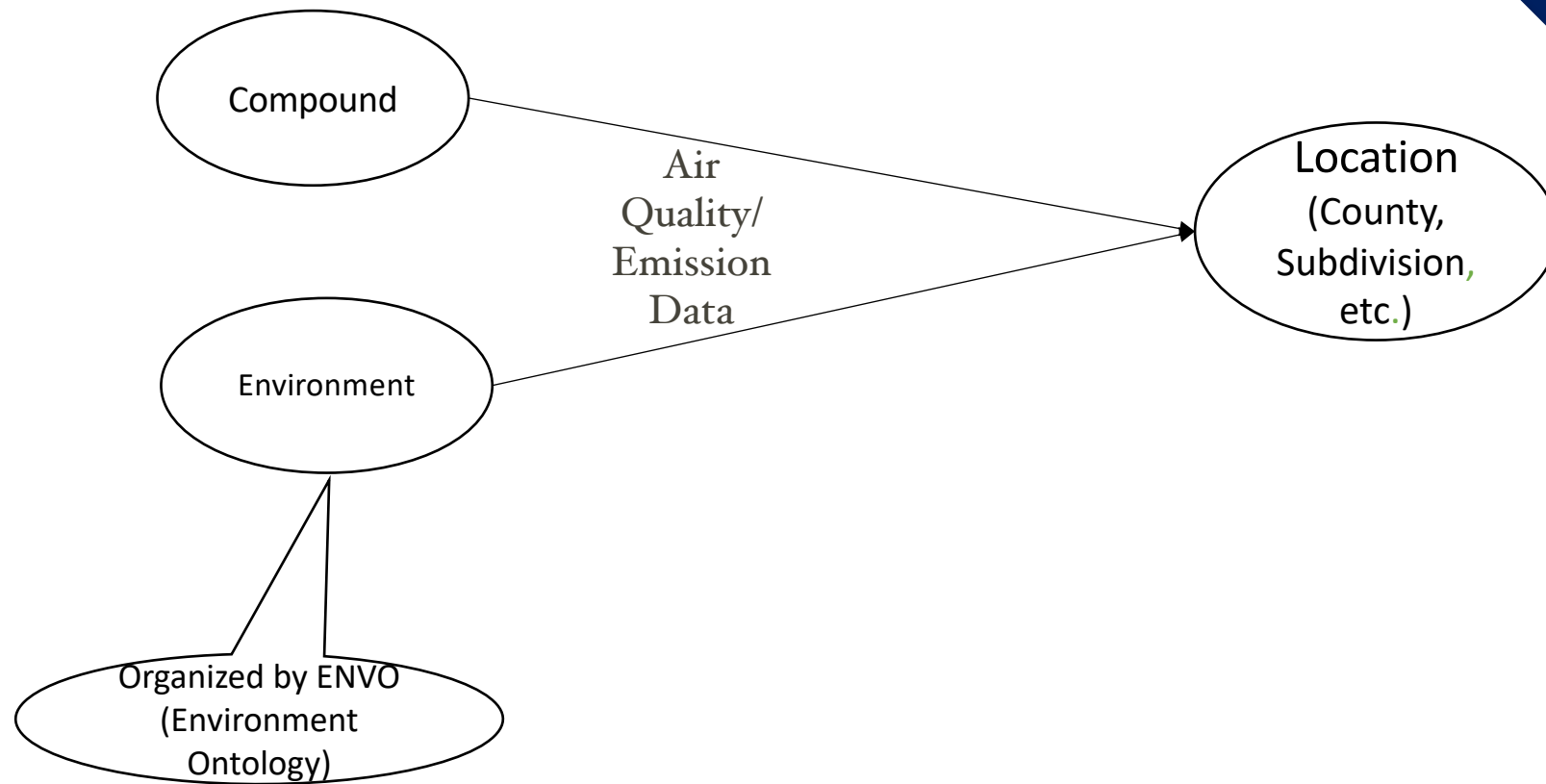
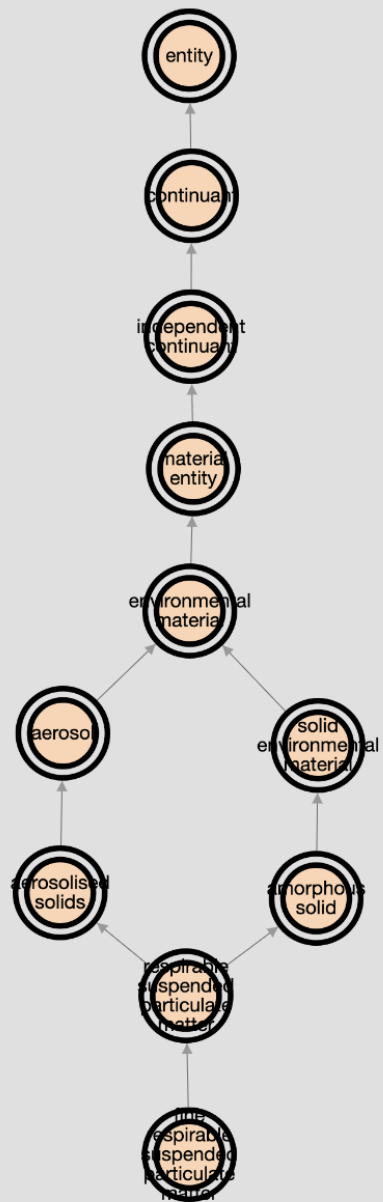
spoke-genelab-v0.0.2

ontology_mapper

ortholog_mapper



Modeling

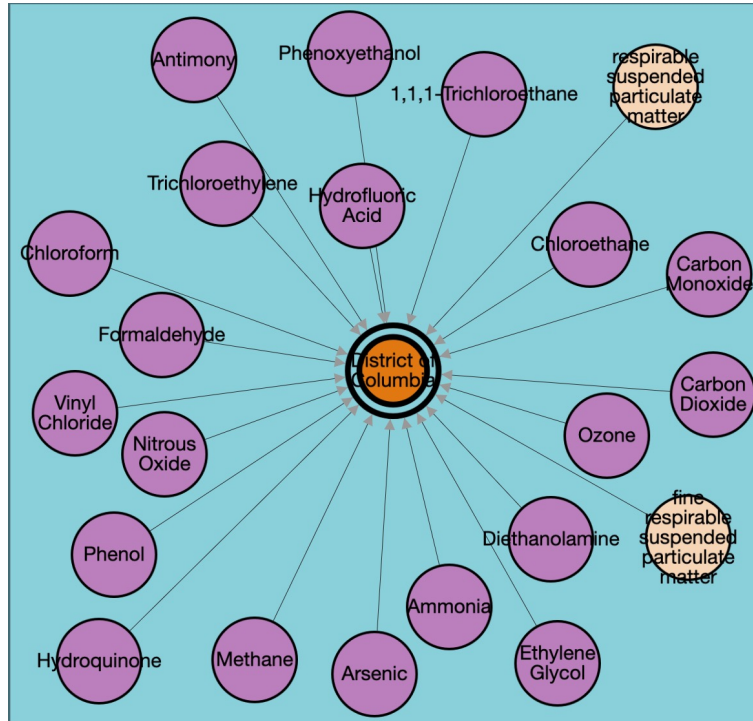
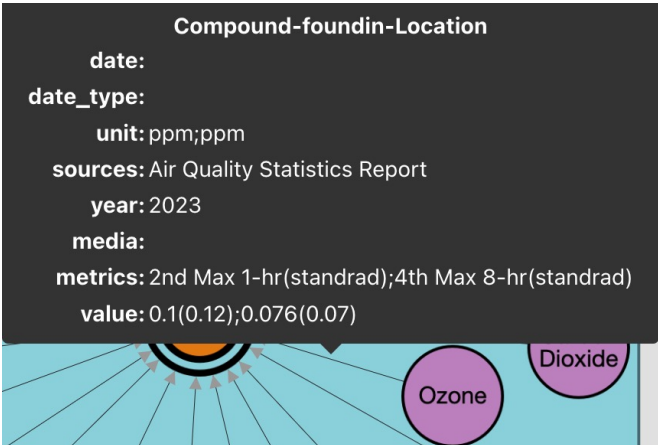


Air Quality and Emission Data in US

- Two data sources from EPA
 - Air Quality Statistics Report
 - 2020 National Emissions Inventory (NEI) data
- Connect Compounds/Environments to US counties

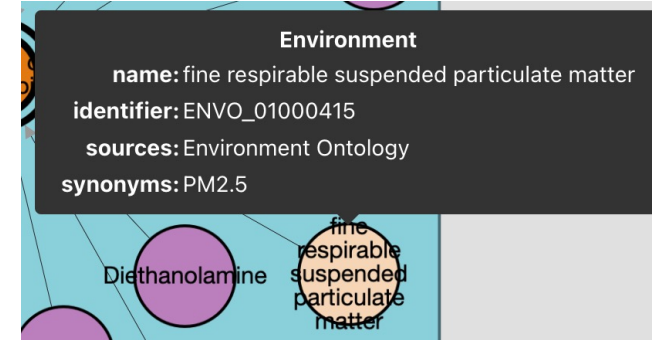
Compound-foundin-Location

date:
date_type:
unit: ppm;ppm
sources: Air Quality Statistics Report
year: 2023
media:
metrics: 2nd Max 1-hr(standrad);4th Max 8-hr(standrad)
value: 0.1(0.12);0.076(0.07)



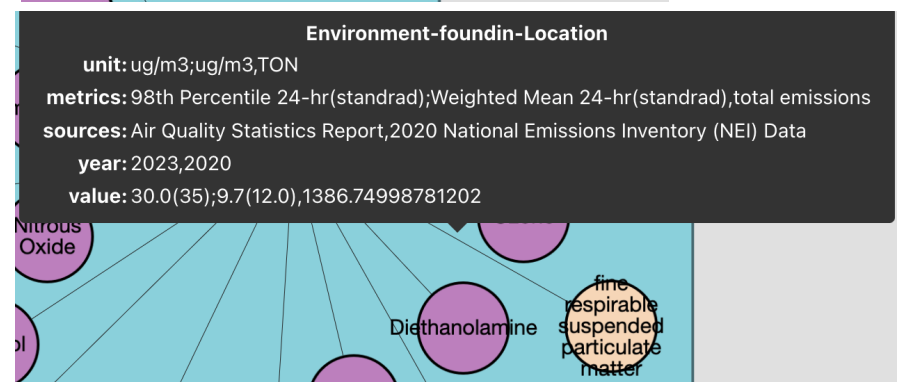
Environment

name: fine respirable suspended particulate matter
identifier: ENVO_01000415
sources: Environment Ontology
synonyms: PM2.5



Environment-foundin-Location

unit: ug/m3;ug/m3,TON
metrics: 98th Percentile 24-hr(standrad);Weighted Mean 24-hr(standrad),total emissions
sources: Air Quality Statistics Report,2020 National Emissions Inventory (NEI) Data
year: 2023,2020
value: 30.0(35);9.7(12.0),1386.74998781202



Project 2: A Dynamically-Updated Open Knowledge Network for Health: Integrating Biomedical Insights with Social Determinants of Health (Bio-Health-OKN)

PI: Aidong Zhang (UVA)

Co-PIs: Cathy Wu (UD), Kishlay Jha (Ulowa)

Stefan Bekiranov (UVA), Rahmat Beheshti (UD)

Senior Personnel: Melissa Little (UVA), Tom Powers (UD), Chuming Chen (UD)

Agency Partners: Suzanne Milbourne (VA), Tim Burke (VA)

Supported by:

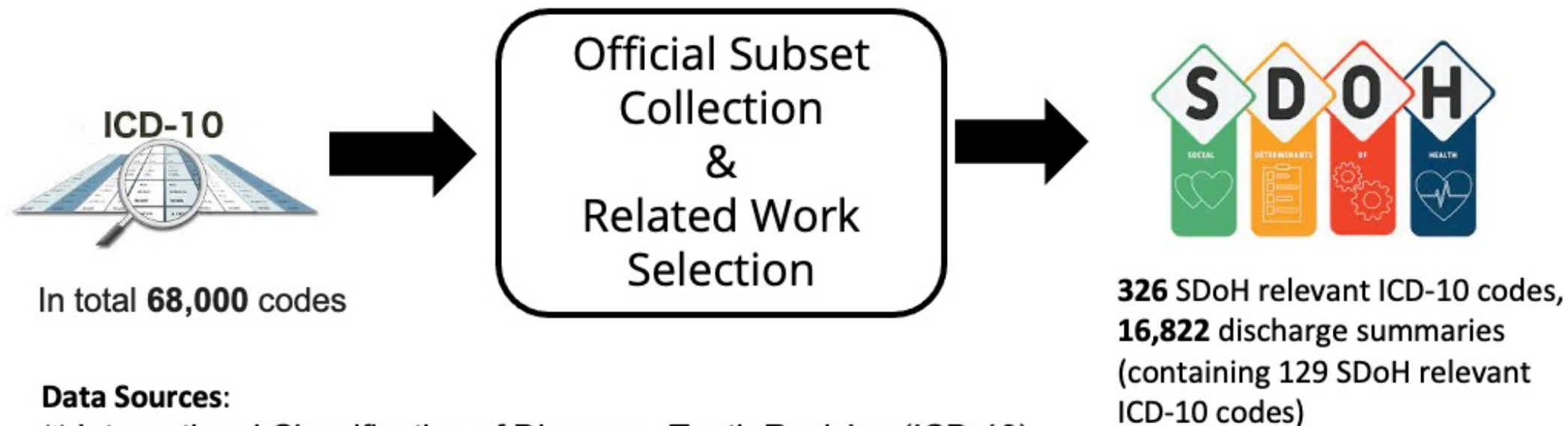


**Biology &
Health**



Proto-OKN

SDoH Labels & Data Collection Pipeline



Data Sources:

- (i) International Classification of Diseases, Tenth Revision (ICD-10),
- (ii) Medical Information Mart for Intensive Care (MIMIC) – IV
- (iii) Assessment of Use of *ICD-9* and *ICD-10* Codes for Social Determinants of Health in the US, 2011-2021



Vocabularies Extension through UMLS



F43.9: Reaction to
severe stress,
unspecified



Concept
Unique
Identifier

C4042925



Trauma and Stressor
Related Disorders



SDoH Extraction Modeling

❑ Tool Used:

- ❑ MetaMap, a biomedical entity extraction model developed by UMLS.

❑ Experiments:

- ❑ Concatenated the semantic types extracted by MetaMap to the actual text of the data.

❑ Findings:

- ❑ The addition of semantic types can significantly supplement the textual information in short texts.
- ❑ Submitted findings to the AMIA 2024 Annual Symposium for peer review and discussion.



Project 3: BioBricks

PI: Thomas Luechtefeld (InSilica)

Supported by:



Biology &
Health



Proto-OKN

NSF Award #2333836

Creating a Cross-Domain Knowledge Graph to Integrate Health and Justice for Rural Resilience

PI: Jiaqi Gong, University of Alabama
Co-PIs: James Geyer, Xiaoyan Hong, Matthew Hudnall, Hee Yun Lee

NSF Award #2333803

IJP: An Integrated Justice Platform to Connect Criminal Justice Data Across Data Silos

PI: Adam Pah, Georgia State University;
Charlotte Alexander

Website

NSF Award #2333790

A Knowledge Graph Warehouse for Neighborhood Information

PI: Jing Gao, Purdue University
Co-PIs: Fenglong Ma, Jingbo Shang, Daniel Semenza

NSF Award #2333703

DREAM-KG: Develop Dynamic, Responsive, Adaptive, and Multifaceted Knowledge Graphs to Address Homelessness With Explainable AI

PI: Yuzhou Chen, Temple University
Co-PIs: Ying Ding, Chiu Tan, Huanmei Wu

Supported by:



Proto-OKN

The problem:

A fragmented data system with human costs

Police

Courts

Corrections

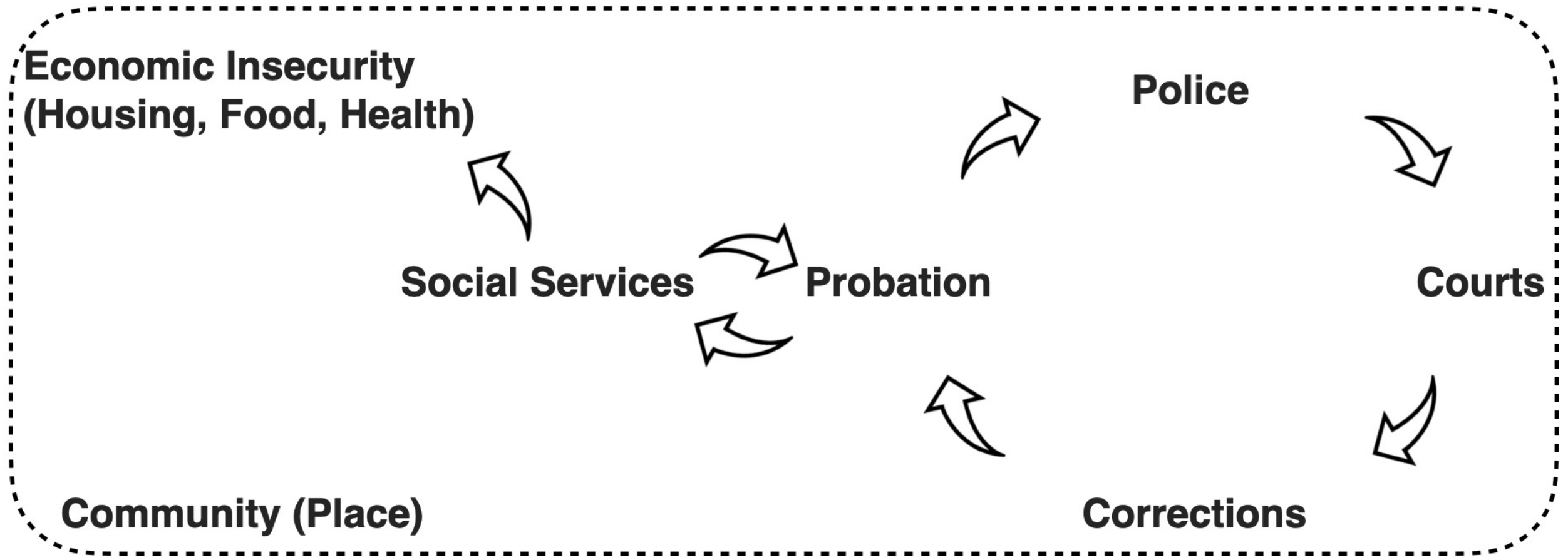
Probation

Social Services



The problem:

A fragmented data system with human costs



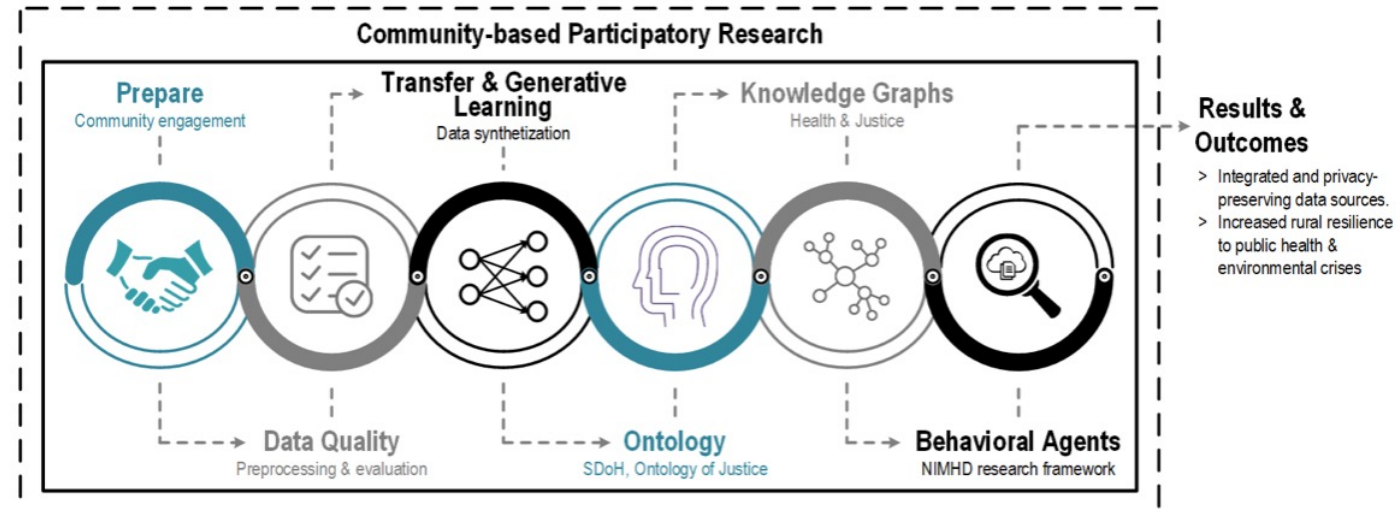
NSF Award #2333836

Creating a Cross-Domain Knowledge Graph to Integrate Health and Justice for Rural Resilience

PI: Jiaqi Gong, University of Alabama
Co-PIs: James Geyer, Xiaoyan Hong,
Matthew Hudnall, Hee Yun Lee



- Integrate, depict, and link previously isolated health and justice datasets, offering a robust resource for researchers, practitioners, and educators to enhance their insights into risk landscapes in rural areas and strengthen their resilience.



Proto-OKN

Current Project Status:

Datasets

- What we accessed:
 - **SchARE**: Diseases and Conditions, Economic Stability, Education Access and Quality, Health Behaviors, Health Care Access and Quality, Multiple Categories, Neighborhood and Built Environment, and Social and Community Context
 - **PLACES/500 cities**: 29 measures, including: 13 for health outcomes, 9 for preventive services use, 4 for chronic disease-related health risk behaviors, and 3 for health status.
 - **Social determinants of health (SDOH)**: social context (e.g., age, race/ethnicity, veteran status), economic context (e.g., income, unemployment rate), education, physical infrastructure (e.g, housing, crime, transportation), and healthcare context (e.g., health insurance).
 - **Behavioral Risk Factor Surveillance System (BRFSS)**: health-related risk behaviors, chronic health conditions, health-care access, and use of preventive services from the noninstitutionalized adult population.
 - **National Incident-Based Reporting System (NIBRS)**: administrative, offense, property, victim, offender, and arrestee.
 - **National Survey on Drug Use and Health (NSDUH)**: use of tobacco, alcohol, and drugs; substance use disorders; mental health issues; and receipt of substance use and mental health treatment among the civilian.
- What we have integrated thus far:
 - **SchARE**: Economic Stability: American Housing Survey (AHS) National Public Use File (PUF) (2015, 2017, 2019, 2021)
 - **National Incident-Based Reporting System (NIBRS)**: administrative, offense, property, victim, offender, and arrestee.



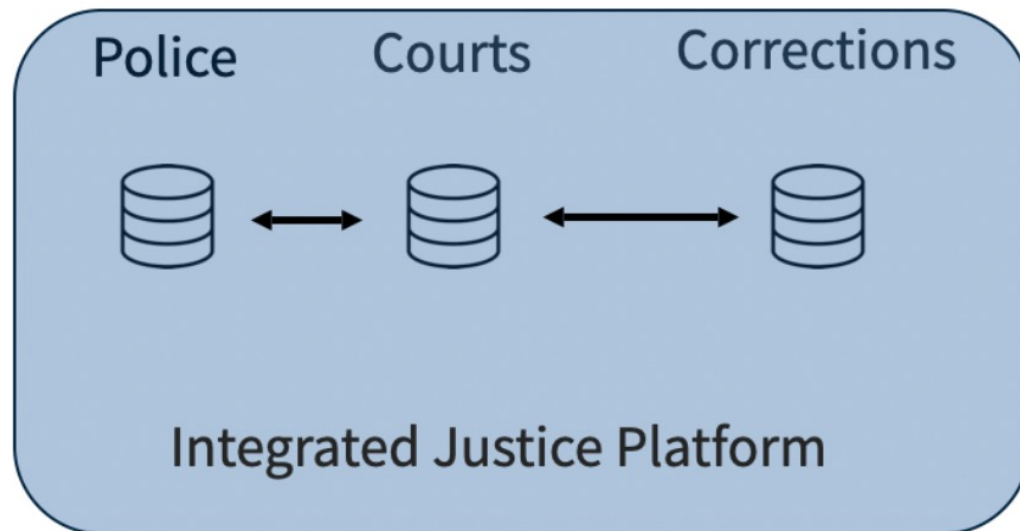
NSF Award #2333803

**IJP: An Integrated Justice Platform to
Connect Criminal Justice Data Across
Data Silos**

PI: Adam Pah, Georgia State University;
Charlotte Alexander

Website

The Integrated Justice Platform harmonizes criminal justice data and makes it public and analyzable



Probation



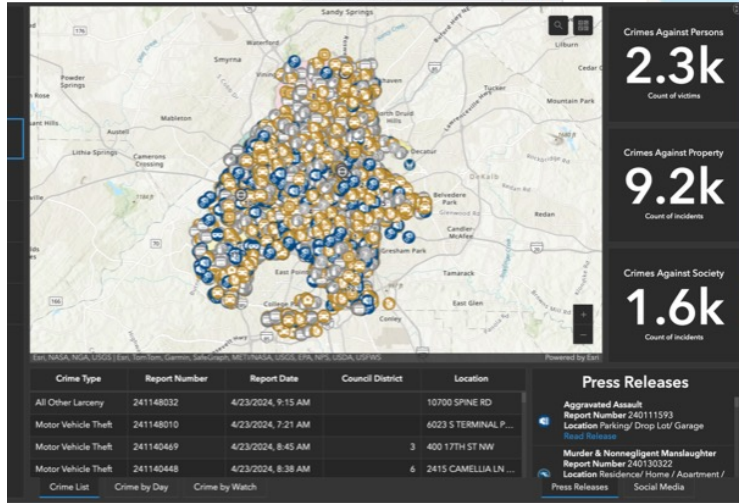
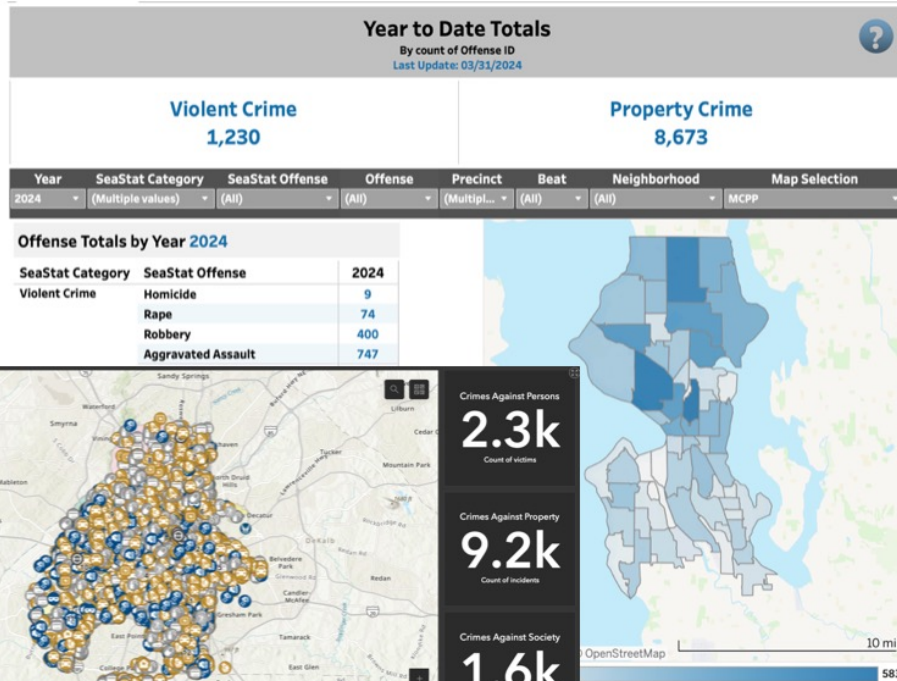
Community Reintegration



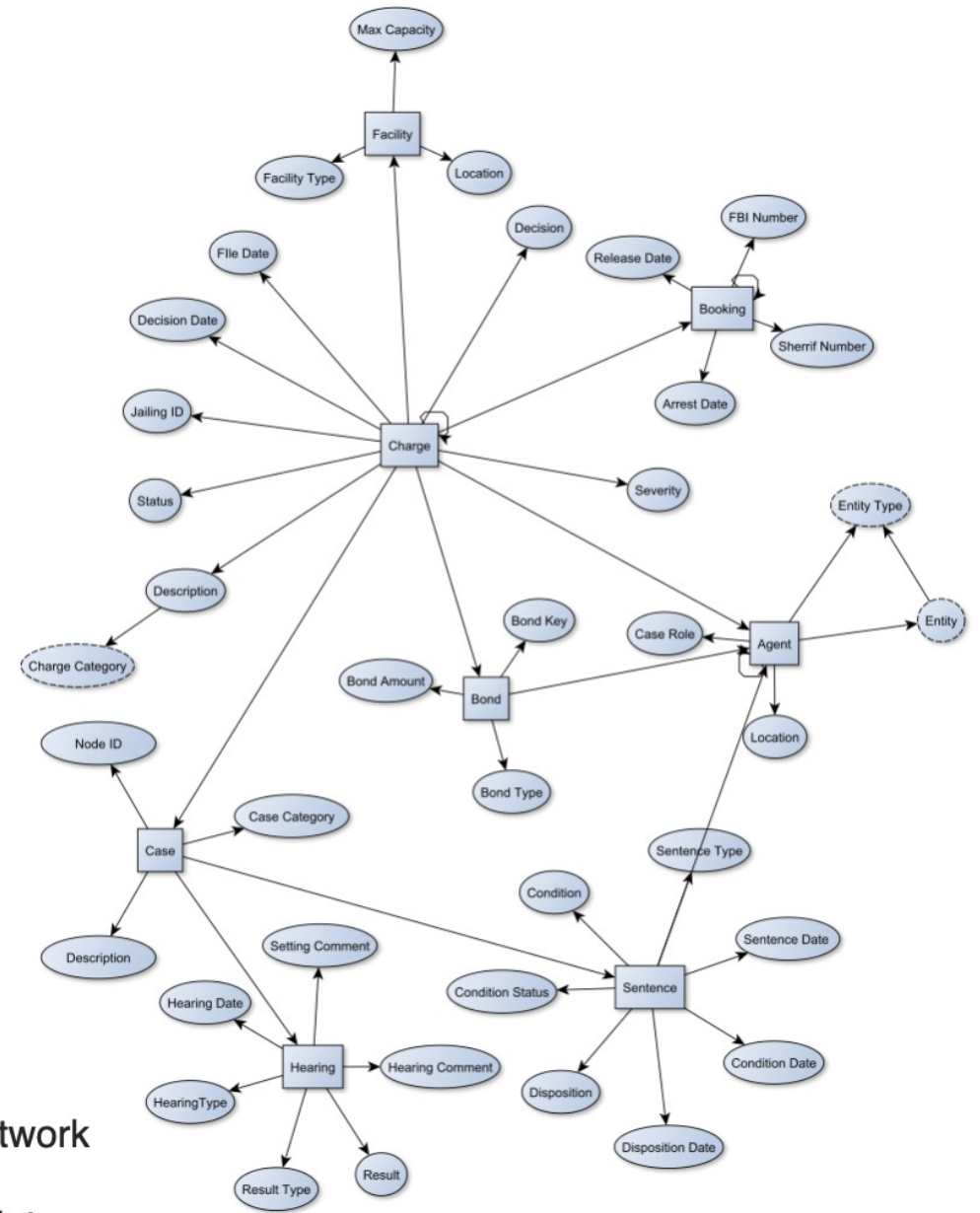
Case study approach
to **build infrastructure**
and **pilot pipelines**



Seattle, WA



Atlanta, GA



Alpha Knowledge Network
 • 1,044,296 URIs
 • 6,600,961 RDF Triplets



Proto-OKN

NSF Award #2333790

**A Knowledge Graph Warehouse for
Neighborhood Information**

PI: Jing Gao, Purdue University

Co-PIs: Fenglong Ma, Jingbo Shang,

Daniel Semenza

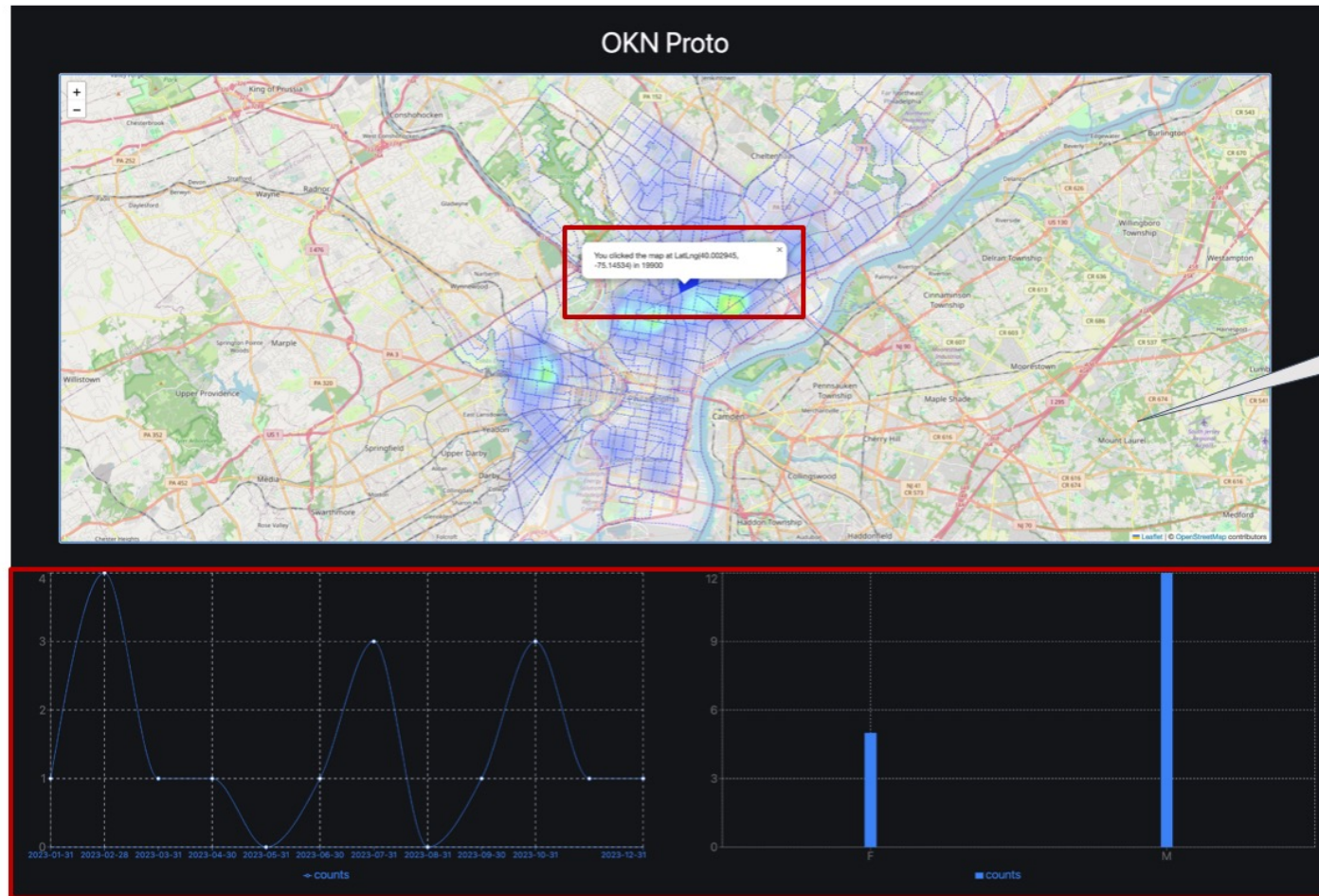
- shooting.csv: <https://opendataphilly.org/datasets/shooting-victims/>

Details of the shootings,
including the date, exact
time, age, sex, etc.

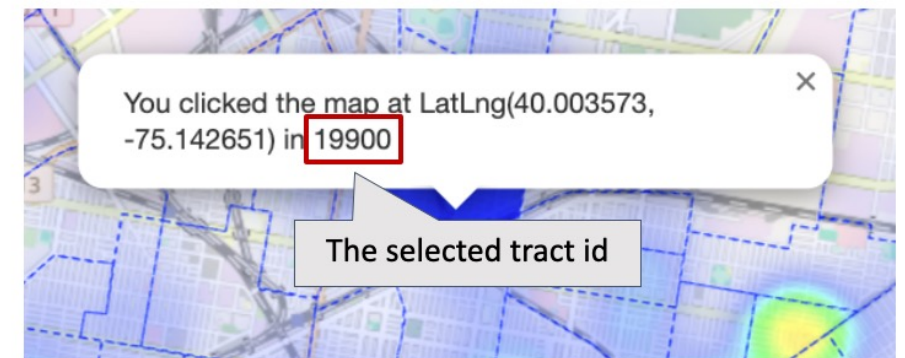
date_	time	race	sex	age	wound	officer_invol	offender_inj	offender_de	location	latino	point_x	point_y	dist	inside	outside	fatal	lat	lng	Block Group	Census tract
) 2021-03-18	1:59:00	B	M	20	Multiple	N	N	N	8100 BLOCK	0	-75.02049	40.0319523	8	0	1	1	40.0319523	-75.02049	2	32900
) 2021-03-29	17:05:00	B	M	21	Chest	N	N	N	1400 BLOCK	0	-74.959364	40.0866439	8	1	0	1	40.0866439	-74.959364	1	36301
) 2021-05-15	8:48:00	B	F	33	Buttocks	N	N	N	8600 BLOCK	0	-75.013539	40.0457315	8	0	1	0	40.0457315	-75.013539	2	34900
) 2021-06-02	1:30:00	B	M	22	Back	N	N	N	2800 BLOCK	0	-75.033909	40.0572187	8	0	1	0	40.0572187	-75.033909	3	34702
) 2021-05-31	12:44:00	B	M	22	Multiple/Hea	N	N	N	500 BLOCK N	0	-75.164533	39.963546	9	0	1	1	39.963546	-75.164533	1	13300
) 2021-07-15	1:03:00	B	M	29	Abdomen	N	N	N	BROAD & CH	0	-75.16386	39.9508656	9	0	1	0	39.9508656	-75.16386	2	901
) 2021-11-01	2:25:00	B	M	22	Multiple	N	N	N	1400 BLOCK	0	-75.160042	39.9689999	9	0	1	0	39.9689999	-75.160042	2	13300
) 2021-11-01	2:25:00	B	M	22	Back	N	N	N	1400 BLOCK	0	-75.160042	39.9689999	9	0	1	1	39.9689999	-75.160042	2	13300
) 2021-11-01	2:25:00	B	M	28	Multiple	N	N	N	1400 BLOCK	0	-75.160042	39.9689999	9	0	1	0	39.9689999	-75.160042	2	13300
) 2021-10-30	11:24:00	B	M	29	Leg	N	N	N	4200 BLOCK	0	-75.024766	40.0409227	8	0	1	0	40.0409227	-75.024766	1	32900
) 2021-12-09	5:45:00	W	M	31	Buttocks	N	N	N	4600 BLOCK	1	-75.012655	40.0436776	8	0	1	0	40.0436776	-75.012655	2	34900
) 2021-02-18	4:10:00	B	M	16	Chest	N	N	N	2900 BLOCK	0	-75.233291	39.9135905	12	0	1	1	39.9135905	-75.233291	3	6100
) 2021-01-01	1:48:00	B	F	22	Arm	N	N	N	2100 BLOCK	0	-75.177121	39.9511359	9	0	1	0	39.9511359	-75.177121	1	401
) 2021-01-24	19:58:00	W	M	18	Multiple	N	N	N	1400 BLOCK	0	-75.164482	39.9535539	9	0	1	0	39.9535539	-75.164482	2	402
) 2021-03-10	22:25:00	B	M	34	Leg	N	N	N	1400 BLOCK	0	-75.160849	39.97032	9	0	1	0	39.97032	-75.160849	2	14000



Implementation details - UI design (now)



When clicking a tract, the diagrams will be updated to be the information about the selected tract.



NSF Award #2333703

DREAM-KG: Develop Dynamic, Responsive, Adaptive, and Multifaceted Knowledge Graphs to Address Homelessness With Explainable AI

PI: Yuzhou Chen, Temple University

Co-PIs: Ying Ding, Chiu Tan, Huanmei Wu

- **Objective:** to create a knowledge graph (KG) system (i.e., **D**ynamic, **R**Esponsive, **A**daptive, and **M**ultifaceted **K**nowledge **G**raph (**DREAM-KG**)) that will
 - provide a comprehensive understanding of the social, economic, and political factors that contribute to homelessness
 - triage existing services and resources to support people experience homelessness (PEH)
 - provide an automated end-to-end knowledge graph/graph AI pipelines that simplifies and standardizes the process of data loading, experimental setup, and model evaluation

• Homelessness-Related Service Data Collection and Preprocessing

Findhelp (Signed Data Use Agreement):

1. Main service
2. Other service
3. Serving
4. Phone number
5. Official website
6. Eligibility
7. Availability
8. Description
9. Languages
10. Cost
11. Facebook/Twitter URLs
12. Coverage
13. Opening hours
14. Geolocation
15. Physical address
16. Zip code



Google Maps (Open Source):

1. Reviewer ID
2. Review summary (rate)
3. Review text
4. Review link
5. Review likes
6. Review timestamp
7. Satellite image
8. Resolution (1200 × 800)
9. Zoom level (20)



Google Maps

OpenDataPhilly (Open Source):

1. Crime incidents from the Philadelphia Police Department, including violent offenses such as aggravated assault, rape, arson, simple assault, prostitution, gambling, fraud, and other non-violent offenses, etc.
2. Dispatch date
3. Police service area (PSA)
4. Geolocation
5. Physical address

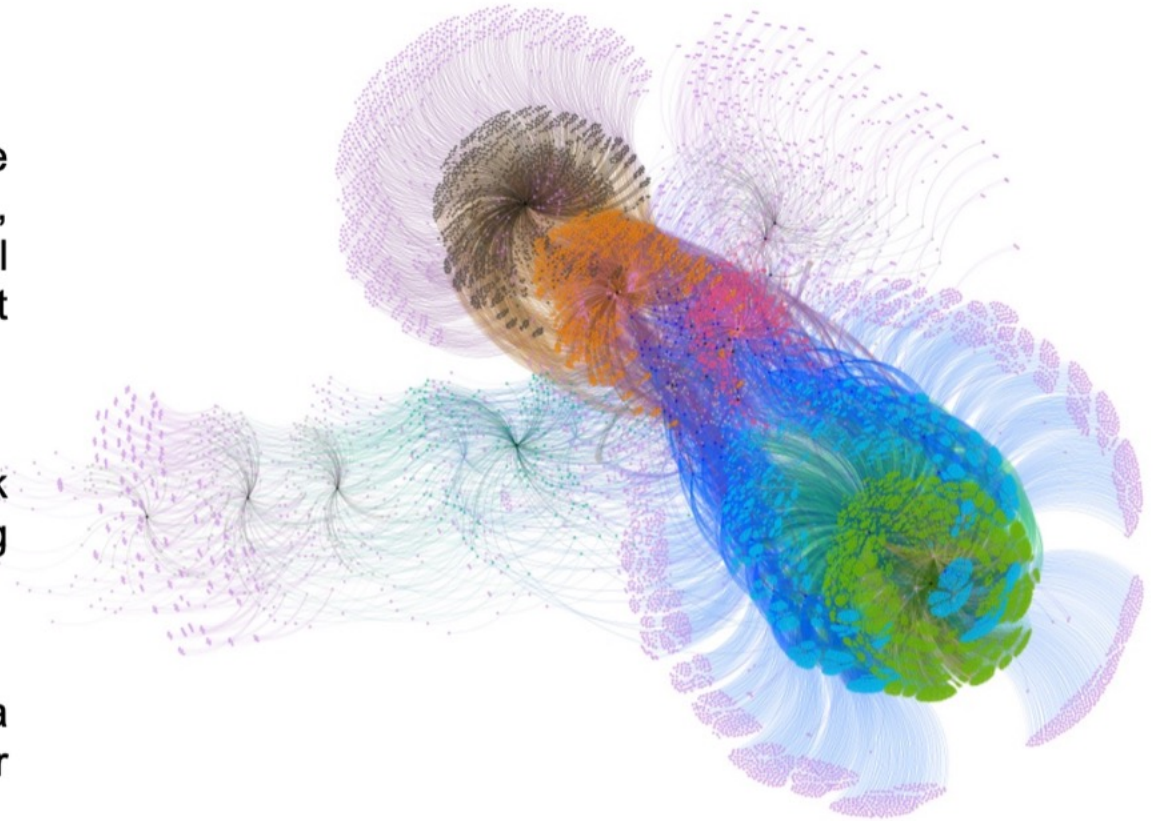


Proto-OKN

• DREAM-KG Construction

Service types: **emergency food, temporal shelter, mental health**

1. Create a spreadsheet containing information about the individual homelessness service providers, such as the name, services provided, contact information, address, geospatial coordinates, opening/closing hours, languages, target audiences, eligibility criteria, and google reviews.
2. Converted this spreadsheet to resource description framework (RDF) triples using PostgreSQL, Ontop VKG, and Schema.org ontology.
3. Once the triples are generated, we also convert the RDF into a tab-separated value (TSV) format for use in the further downstream tasks.
4. Provide SPARQL for query and Gephi tool for visualizations.
5. DREAM-KG: **16,286** entities, **35** relations, and **41,633** RDF triples.



Overview of health-related homelessness service knowledge graph. **Green and light blue represent user reviews**, **dark blue and red represent service information**, and **orange and black represent questions and answers about the services**.

Proto-OKN Technology & Manufacturing Working Group

CollabNext

PI: Lew Lefton, Georgia Tech

Software Supply Chain

PI: Tianyi Zhang, Purdue University

Supply and Demand Open Knowledge Network

PI: Farhad Ameri, Arizona State University

Supported by:



Technology & Manufacturing

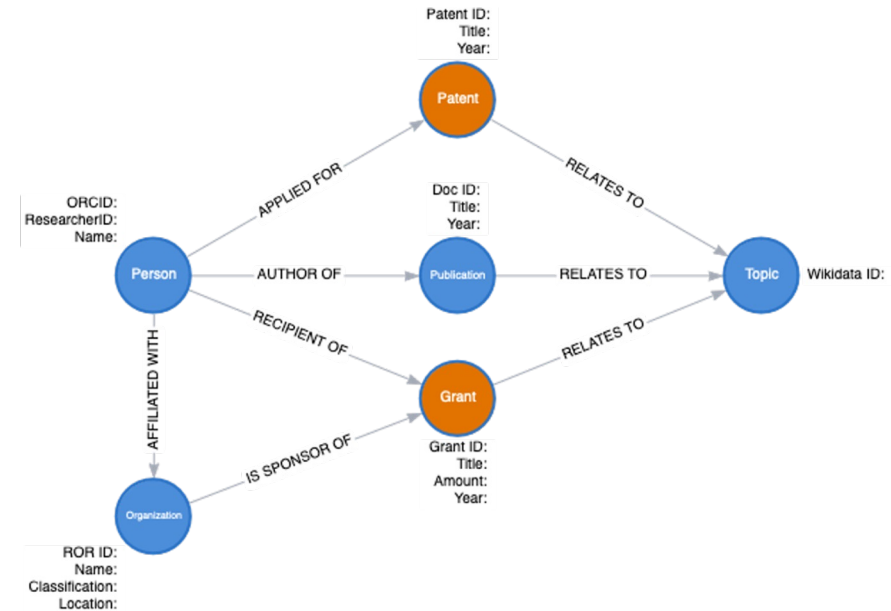


Proto-OKN

CollabNext: A Person-Focused Metafabric for Open Knowledge Networks

Objective: We will develop a knowledge graph of **People, Organizations and Research Topics**. We are adopting an intentional design approach which initially prioritizes HBCUs and emerging researchers in a deliberate effort to counterbalance the Matthew effect: an accumulated advantage of well-resourced research organizations.

Data: We use open science data sources. Current proof-of-concept CollabNext tool uses OpenAlex (formerly Microsoft Academic Graph) and is integrating institutional data from MUP: Measuring University Performance. Additional data sources will be added.



CollabNext: A Person-Focused Metafabric for Open Knowledge Networks

Users: Researchers, Sponsored Program Officers, Journal Editors, University Administration and Leadership, Students, Librarians, Industry Professionals

Example Use Cases:

As a principal investigator (or industry partner), I want to **identify and contact colleagues (at HBCUs/MSIs or at well-resourced institutions) with interest and expertise in a specific research areas**, so that I can **build a stronger research team and increase collaborative research efforts**.

As a sponsoring agency program officer (or journal editor), I want to **identify faculty researchers who specialize in certain areas** to serve as reviewers so I can **expose more researchers to what successful submissions** look like.

As a program or conference organizer I want to **curate a diverse panel of knowledgeable experts** so that my event will engage a broader audience.



CollabNext: A Person-Focused Metafabric for Open Knowledge Networks

Collaborators: Fisk University, Georgia Tech, Morehouse College, Texas Southern University, University at Buffalo

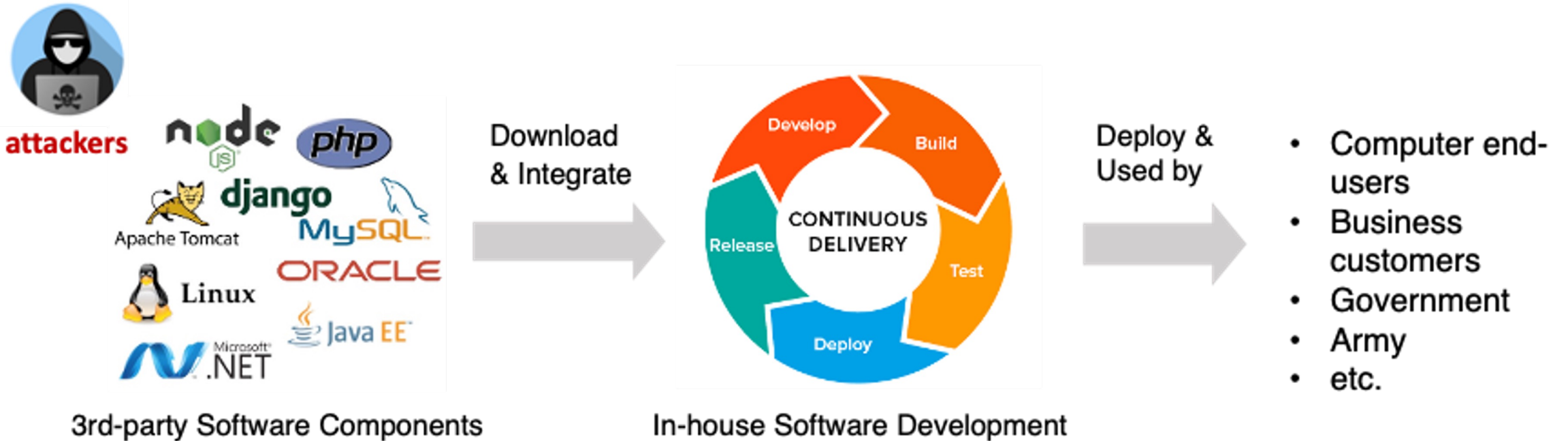
Schema will connect **Person** and **Topic** entities using OpenAlex and other data including: **Publications** (Articles, Books, Preprints, Theses, Conference Proceedings), **Grants/Awards**, **Patents**.

Leverage consistent state-of-the-art algorithms for entity resolution (eg name disambiguation), topic classification, LLM integration for Natural Language Interface



Knowledge Graph Construction for Resilient, Trustworthy, and Secure Software Supply Chains

Objective: Develop a knowledge graph to continually track and assess vulnerability propagation in software supply chains



Knowledge Graph Construction for Resilient, Trustworthy, and Secure Software Supply Chains

Objective: Develop a knowledge graph to continually track and assess vulnerability propagation in software supply chains



Supply Chain Attack: Major Linux Distributions Impacted by XZ Utils Backdoor

Urgent security alerts issued as malicious code was found embedded in the XZ Utils data compression library used in many Linux distributions.

WIRED

The Untold Story of the Boldest Supply-Chain Hack Ever

The attackers were in thousands of corporate and government networks. They might still be there now. Behind the scenes of the SolarWinds investigation.

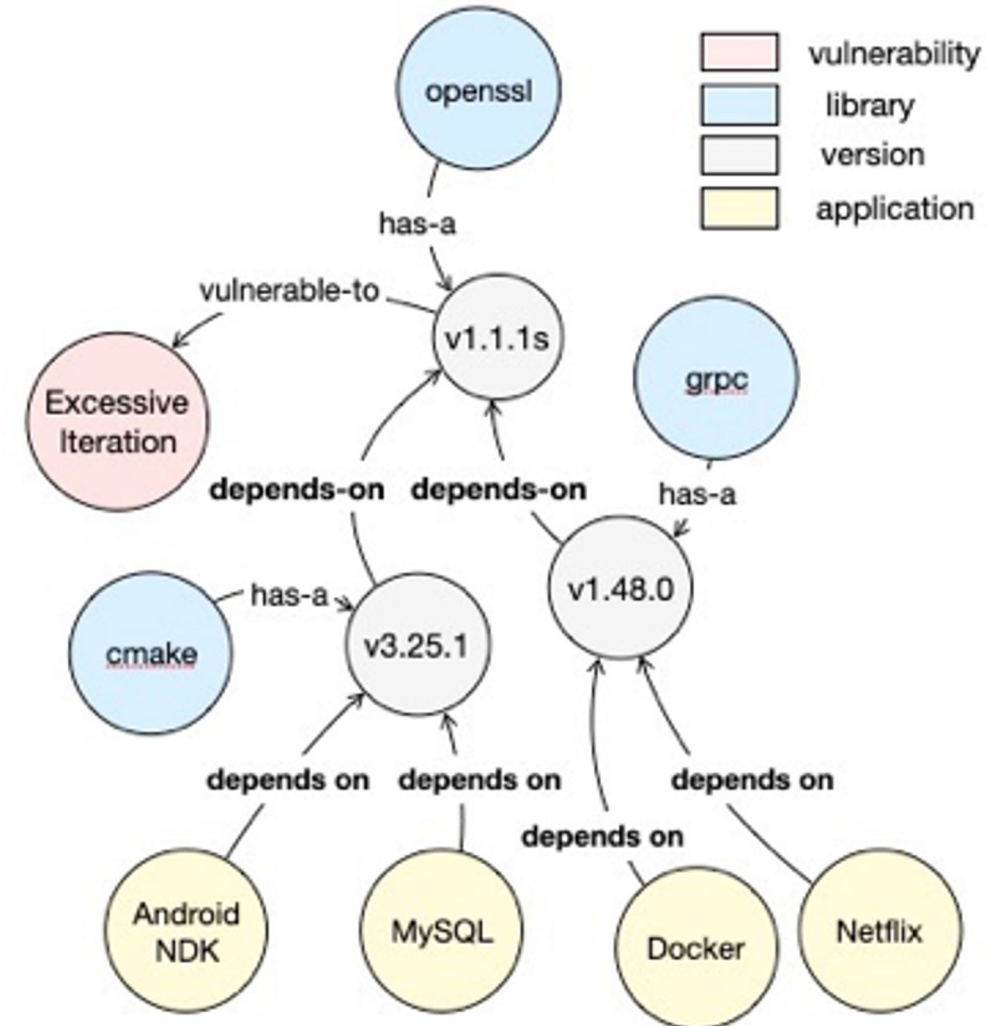


Knowledge Graph Construction for Resilient, Trustworthy, and Secure Software Supply Chains

Users: IT managers, security analysts, software developers, computer end-users

Use Cases:

- IT managers and security analysts can query the KG to check whether any of the software in their companies depends on a vulnerable 3rd party component
- End-users can check whether any of the software on their computers are affected by any recently discovered vulnerabilities
- Developers can check the security of the libraries they use to build the software and also rely on the KG to generate a comprehensive and precise Software Bill of Materials



Knowledge Graph Construction for Resilient, Trustworthy, and Secure Software Supply Chains

Data:

- The Libraries.io dataset: open-source software libraries and applications from 33 package managers and 3 source code hosting platforms.
- GitHub repositories
- CVE List: publicly disclosed cybersecurity vulnerabilities.
- The Hacker News: the latest vulnerability reports from many sources



Supply and Demand Open Knowledge Network(SUDOKN)

Challenge:

- There are about 400,000 small and medium-sized manufacturers in the U.S
- Due to the lack of comprehensive and accurate information about their capabilities, supplier discovery is a lengthy and resource-intensive process.

Objective:

- Prototype an integrated data and knowledge infrastructure representing the capabilities of US manufacturers.

Impact:

- Improving the findability and visibility of SMMs.
- Improving the resilience of manufacturing supply chains.



Use Case: Supplier Discovery



Use Case: Capacity scale-up and capability extension



Proto-OKN

Supply and Demand Open Knowledge Network(SUDOKN)

Users: Supply chain managers, manufacturers, OEMs, MEP Centers, Economic developers

Example Use Cases:

- As a supply chain manager, I need to find a company for pipe bending that is certified for 3D bending for oxygen pipes in a submarine.
- As a precision machine shop in northern California, I need to know what unique capabilities I have that differentiate my company from the peer companies in this area.
- As an economic developer, I need to identify technological gaps in Nevada in the semiconductor manufacturing.

0847976000005



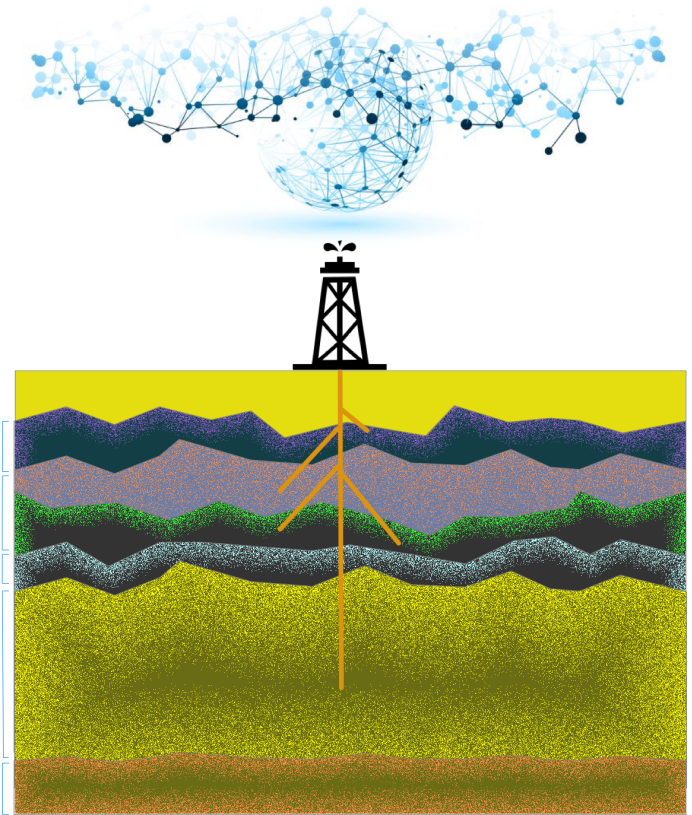
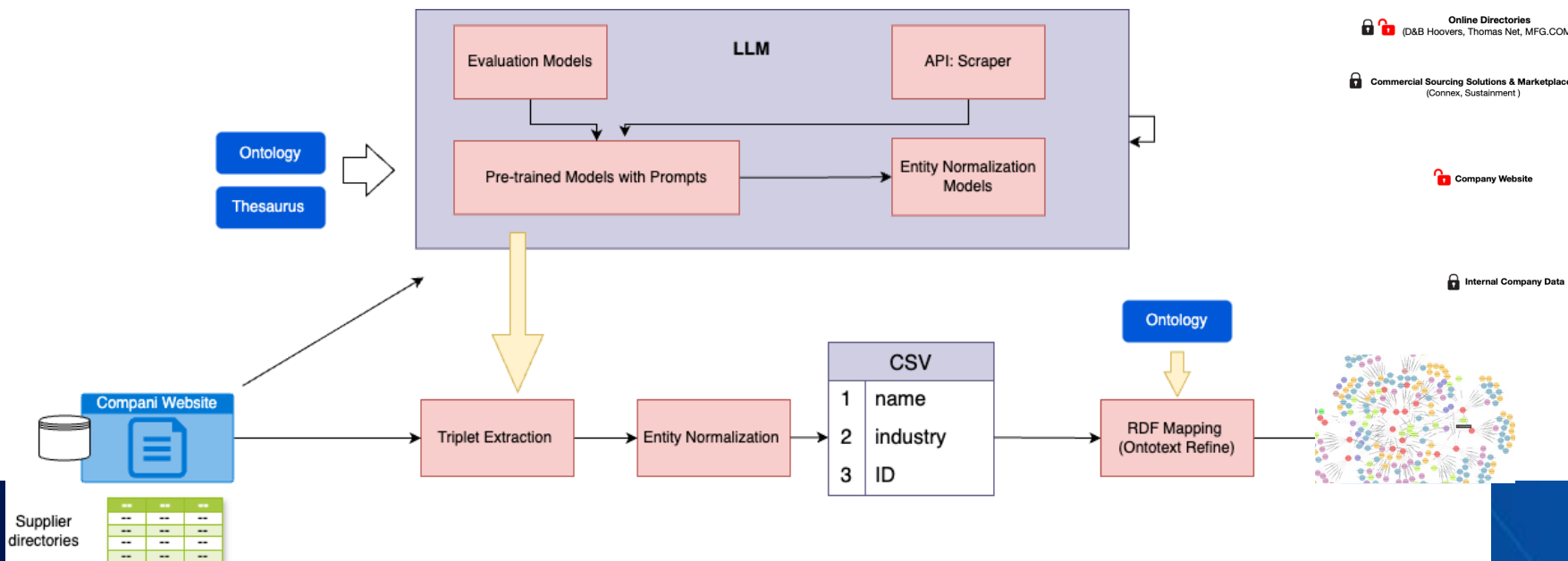
Supply and Demand Open Knowledge Network(SUDOKN)

○ Data Sources:

- Company websites
- Online Directories
- Manufacturing Marketplaces
- DHS Geospatial Management Office

○ Data Ingestion Pipeline:

- Formal ontologies (BFO-complaint), SKOS Thesaurus, LLM



Natural Language Queries

Your Query:

Give me a list of companies that can provide precision machining services for aerospace-grade materials.

Submit

Cancel

Filter

- veteran-owned
- woman-owned
- minority-owned
- disadvantaged

Apply

Cancel



Workshop Agenda

□ Introduction

- *Chaitan Baru & Jemin George, TIP Directorate, National Science Foundation*

□ Presentation by Theme 1 Groups focusing on

○ Environment

- *Lilit Yeghiazarian, University of Cincinnati*

○ Biology & Health

- *Sergio, Baranzini, University of California, San Francisco (UCSF)*

○ Justice

- *Adam Pah, Georgia State University (GSU)*

○ Technology & Manufacturing

- *Farhad Ameri, Arizona State University (ASU)*

□ Presentation by Theme 2: Proto-OKN Fabric

- *Chris Bizon, University of North Carolina at Chapel Hill (UNC) & Patrick Grinaway, Onai*

□ Presentation by Theme 3: Proto-OKN Education and Public Engagement

- *Cogan Shimizu, Wright State University*



FRINK: FabRic Integrating Networked Knowledge

Supported by:



Proto-OKN

Many Theme 1s will create Knowledge Graphs



The Open Knowledge Network is formed by the integration of these graphs.



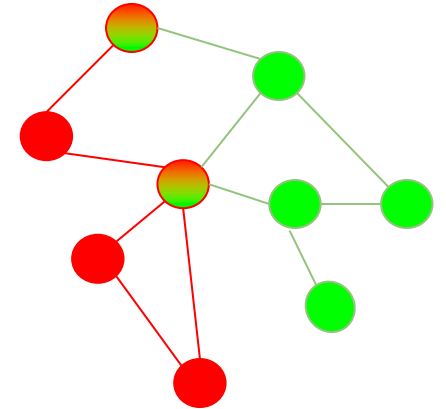
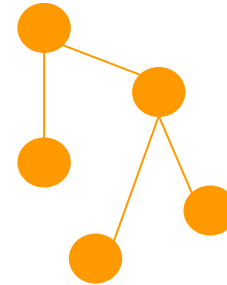
Themes 2 and 3 provide the fabrics supporting this integration

- A **Social Fabric**: Building a community across members to establish a shared vision for the Proto-OKN.
- A **Knowledge Fabric**: Creating the standards and methods for graph interoperability.
- A **Technical Fabric**: The common cloud-based infrastructure in which the Proto-OKN will exist.



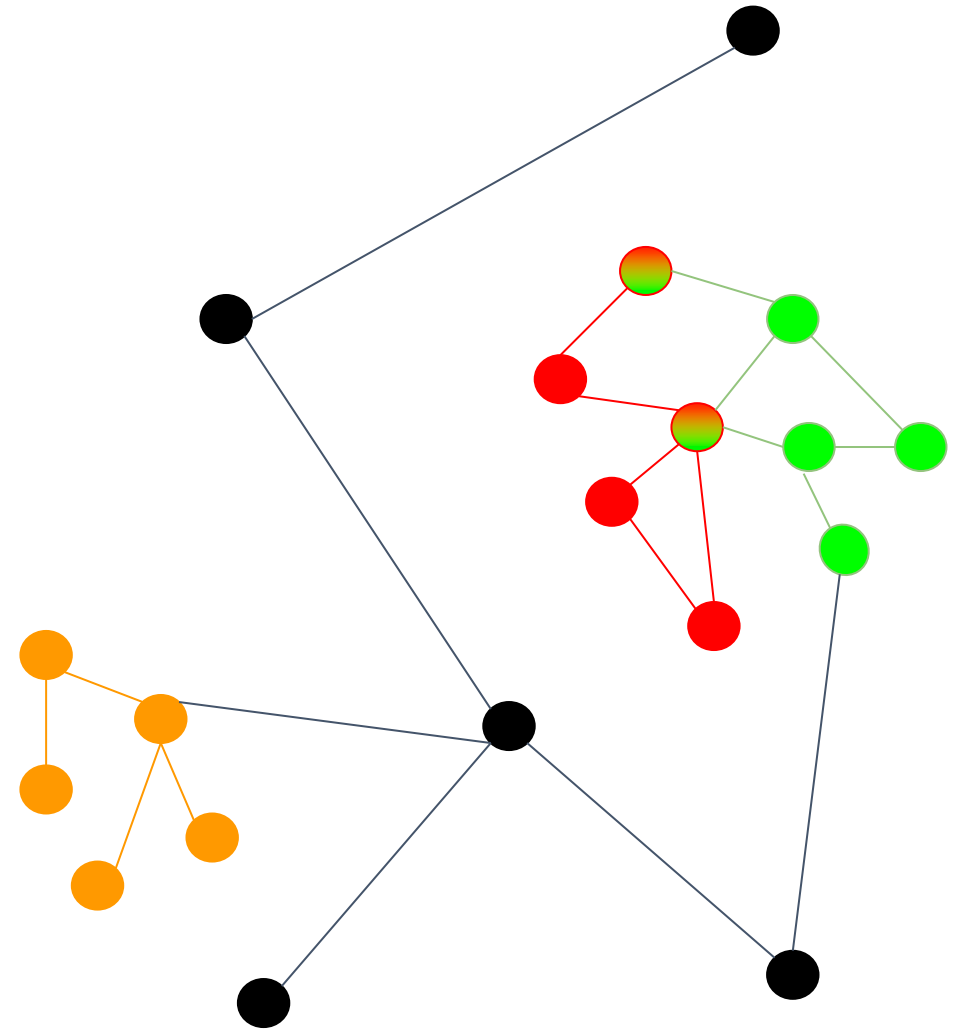
Knowledge Fabric: Interoperability

- Graphs will be created using a variety of
 - Formats
 - Data modeling styles
 - Naming conventions
- We will establish standards, drawing on **pre-existing community approaches** to allow these graphs to interoperate
- Graphs will sometimes connect naturally, but not always.

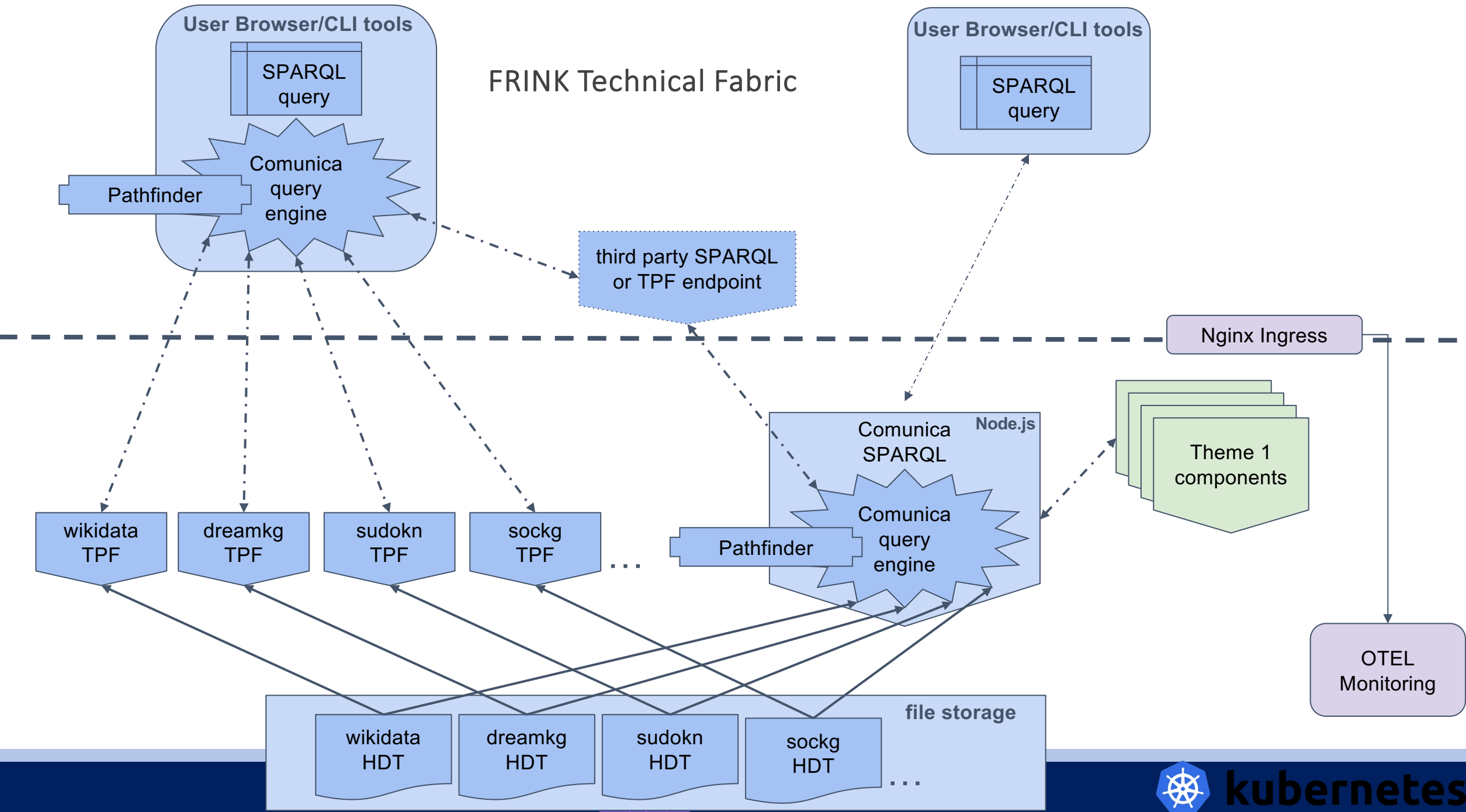


Knowledge Fabric: Wikidata

- We will rely on the pre-existing open-source Wikidata as a knowledge substrate.
- Wikidata covers a wide range of knowledge at a lower density than the specific graphs created for Proto-OKN



FRINK Technical Fabric



Scalable Public Infrastructure for Distributed Entity Relationships

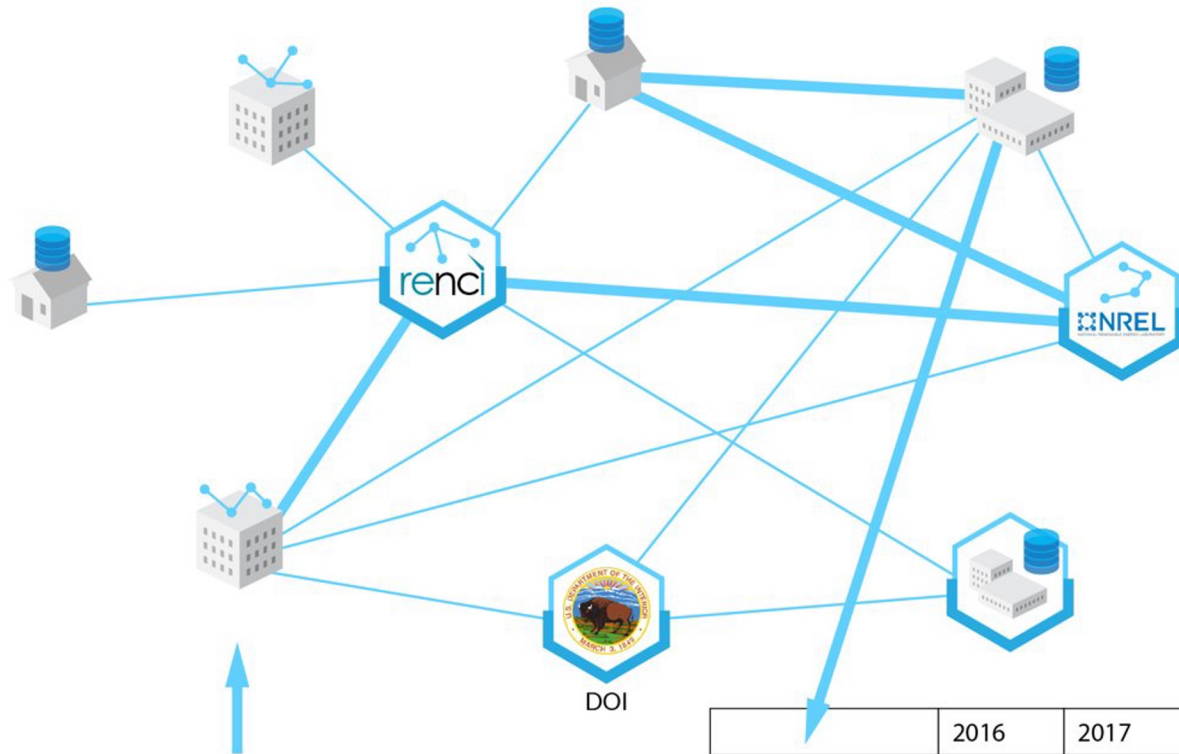
info@onai.com

Supported by:



Proto-OKN

SPIDER Project Overview



Objective:
Scalable federated infrastructure for distributed knowledge graphs, enabling powerful queries, artificial intelligence models, and more.

🔍 # of unique cases grouped by state where the filing year is 2016 or 2017 and Nature of Suit contains "Freedom of Information Act"

	2016	2017
Alabama	██████	██████
Alaska	██████	██████
Arizona	██████	██████
Arkansas	██████	██████



Desired Attributes

- accessible open participation,
- rapid response time,
- automated interoperability across domains, formats, and nomenclature,
- ever-improving support for cutting-edge AI,
- individualized confidence scores on results,
- scalability to large numbers of (sub-)graphs and large sizes for individual subgraphs,
- ability to leverage sensitive data while provably maintaining confidentiality,
- decentralization across premises and clouds,
- ease of use for nonexpert users via a natural language interface, and
- an emphasis on transparent ethics and governance.



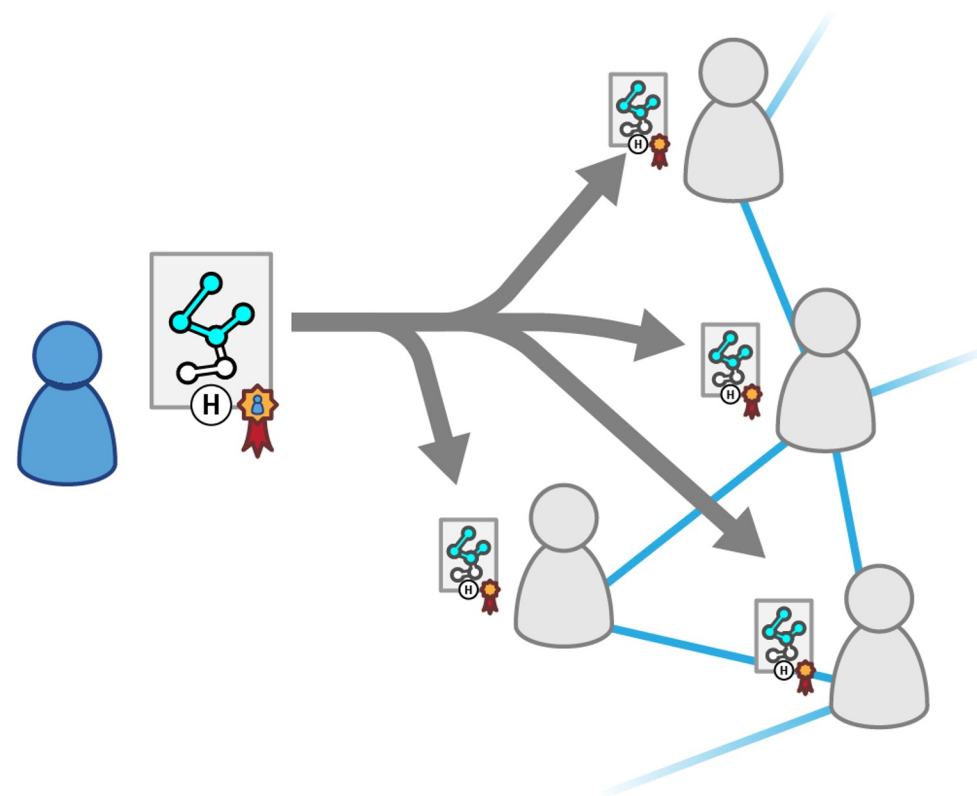
Distributed Architecture

Each location participating in the distributed network as a subgraph storage site runs an instance of the software agent, which listens for queries and participates in responding to queries.

Graphs produced by the Theme 1 participants—and in future others—can reside at disparate locations. Support for decentralized hosting facilitates ease of onboarding of new datasets and redundancy.

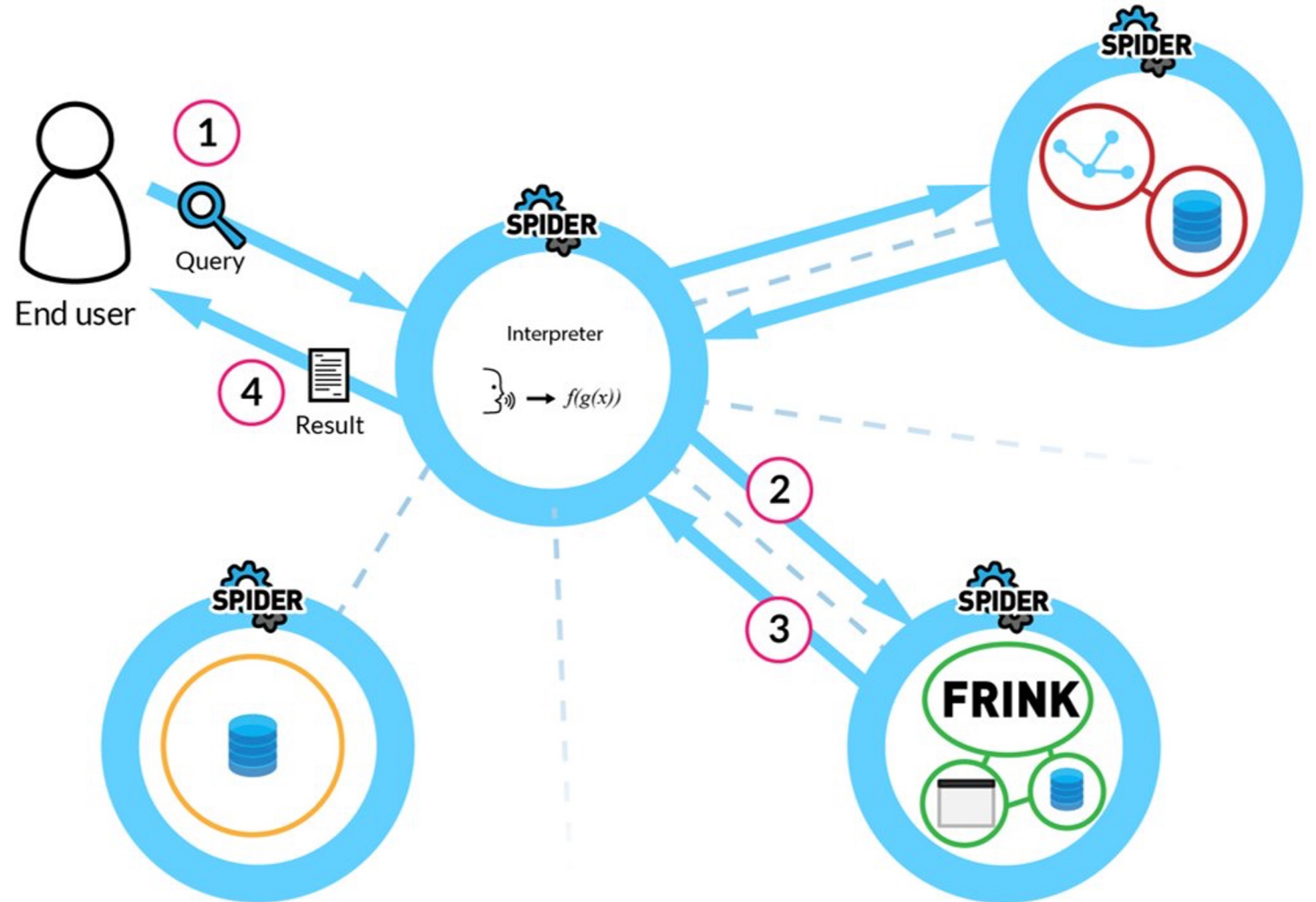
Popular subgraphs will be cached in the primary cloud-based instance (FRINK) to ensure there are always live instances with those subgraphs.

Computation occurs locally at each instance.

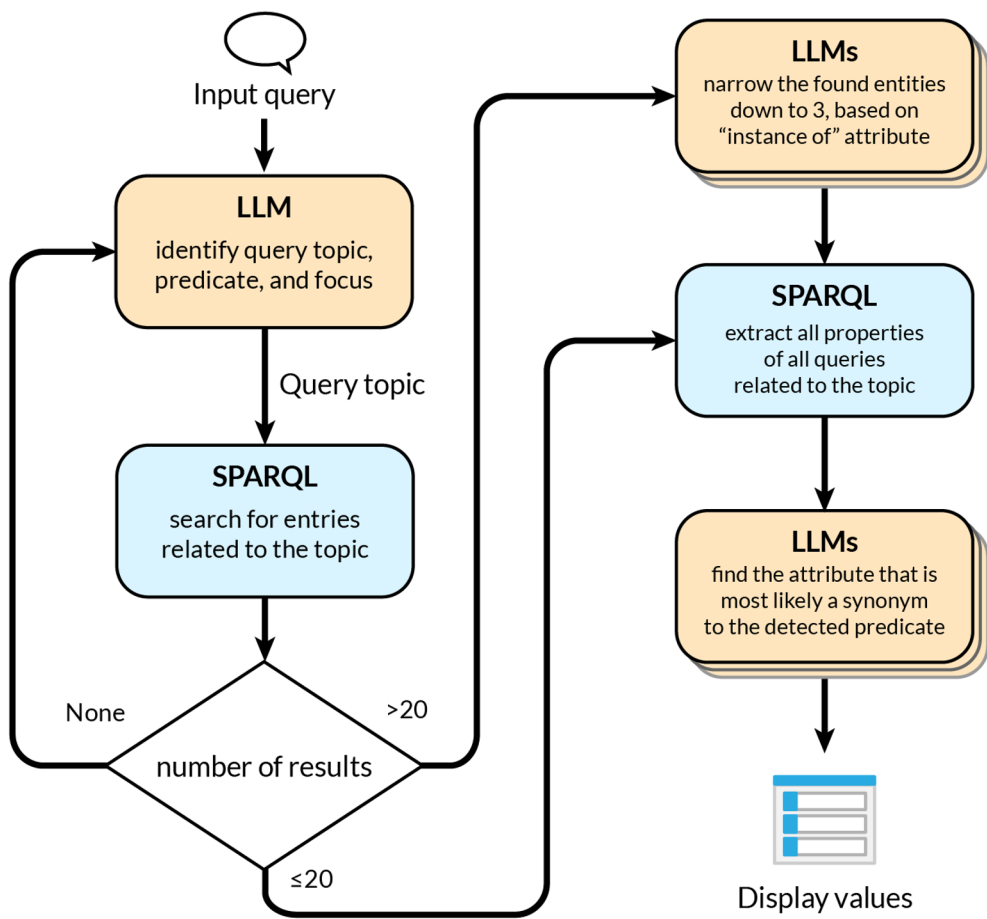


Query Interpretation

1. Query sent to an "Interpreter" node, which leverages an LLM to convert the query to specific actions
2. Interpreter invokes functionality at relevant instances
3. Response is returned for presentation preparation
4. Result presented to user



KG Query Translation



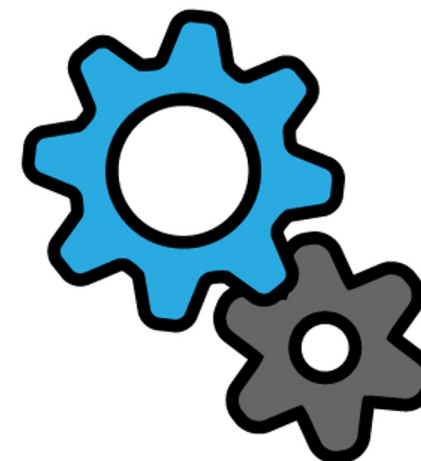
Go from query to SPARQL to correct answer from KG

Don't let the LLM use its own knowledge!

Not just an algorithmic/AI challenge, but also engineering

Computational Extensibility

- Containerized capabilities can be disseminated to all network participants.
- Executed locally in a sandboxed environment.
- Easily create new format translation, querying, inference, or other functionality.
- Secondary goal is to help drive rapid prototyping.

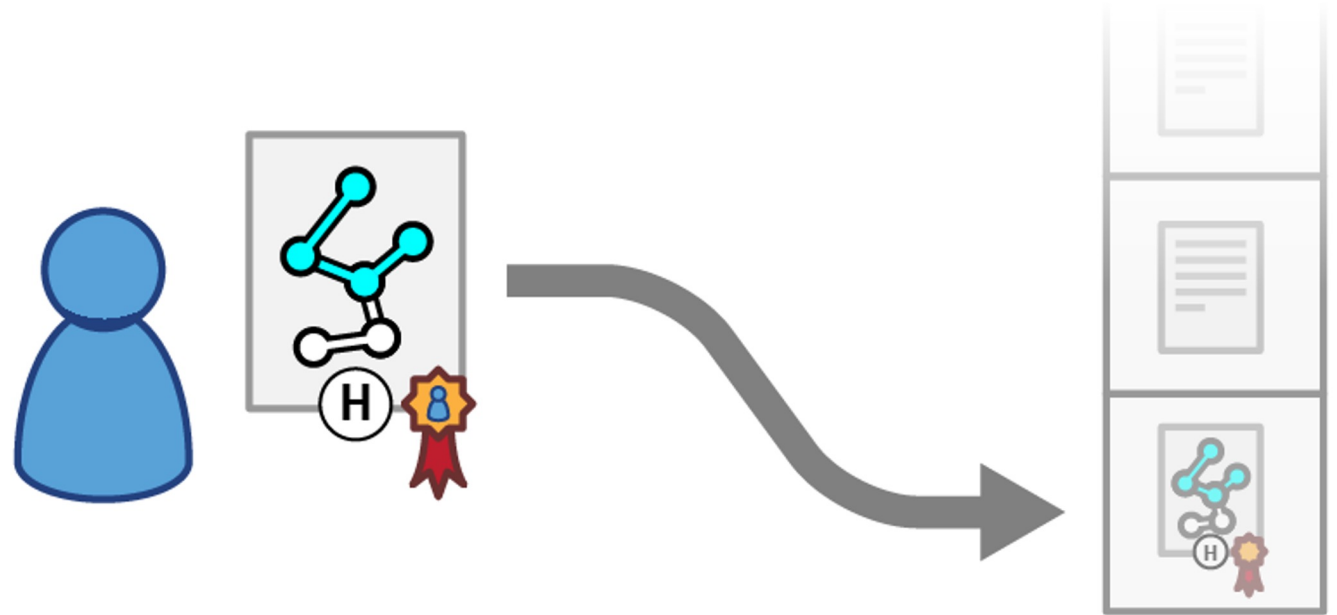


Commitments / Provenance

A signed cryptographic commitment (hash) of a subgraph is committed to a blockchain / distributed ledger.

This allows anyone to verify, and prove, they are operating with authentic data (even if using a cache) and allows for versioning.

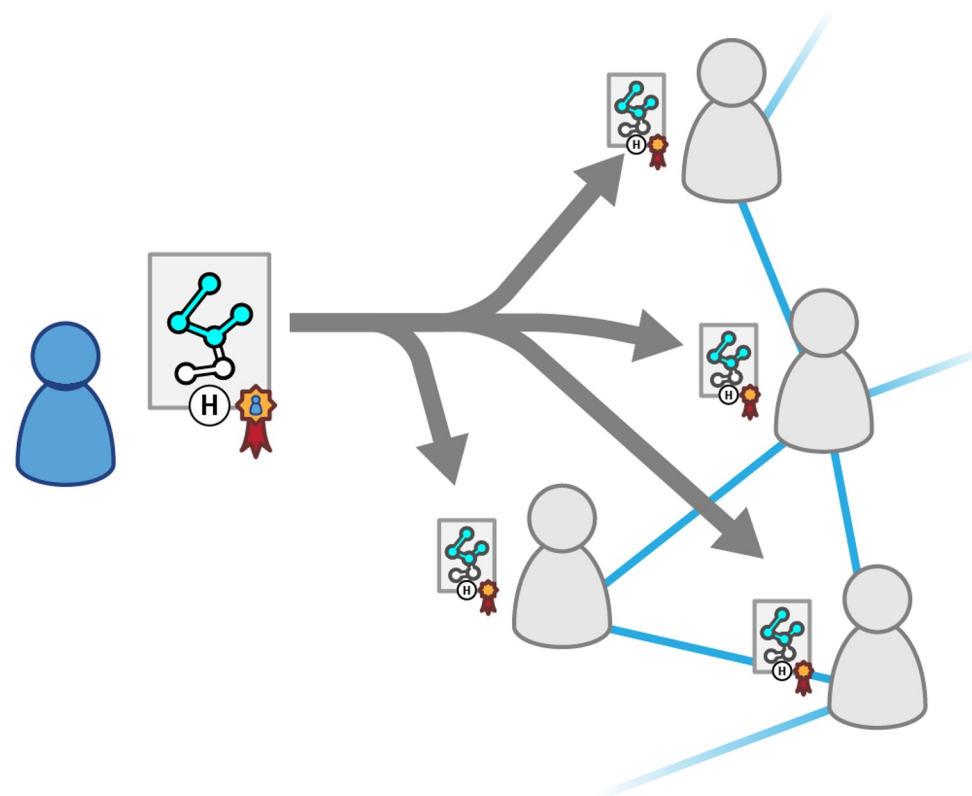
Commitments can also be made to processed outputs of data.



Zero-Knowledge Proofs

Zero knowledge (ZK) proofs are a cryptographic technique through which a verifier can rapidly confirm that a computational result was produced by given code on given inputs, without needing to repeat the entire computation.

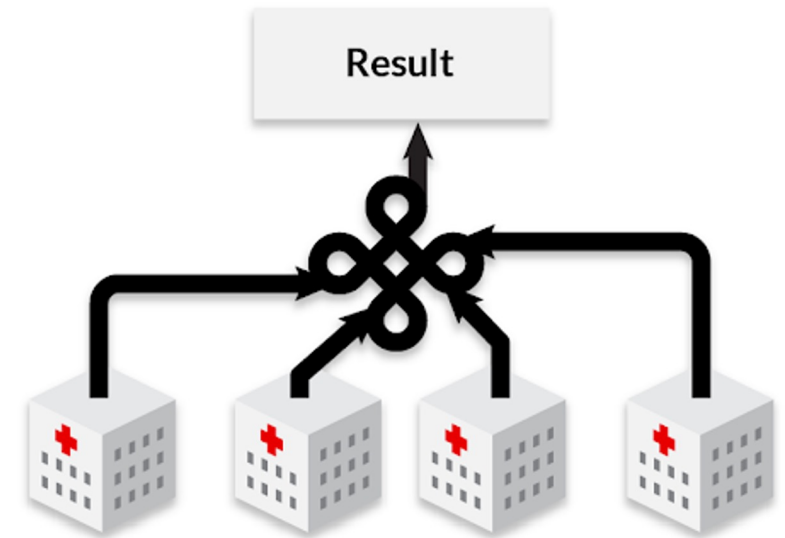
ZK proofs can allow efficient verification that a given graph query result is the correct result obtained from performing the requisite computation on authentic registered subgraphs that were signed and committed by their creators.



Private/Confidential Data

For sensitive data, utilize cryptographic techniques that allows multiple parties to jointly perform computations without revealing their respective inputs to other participants or third parties.

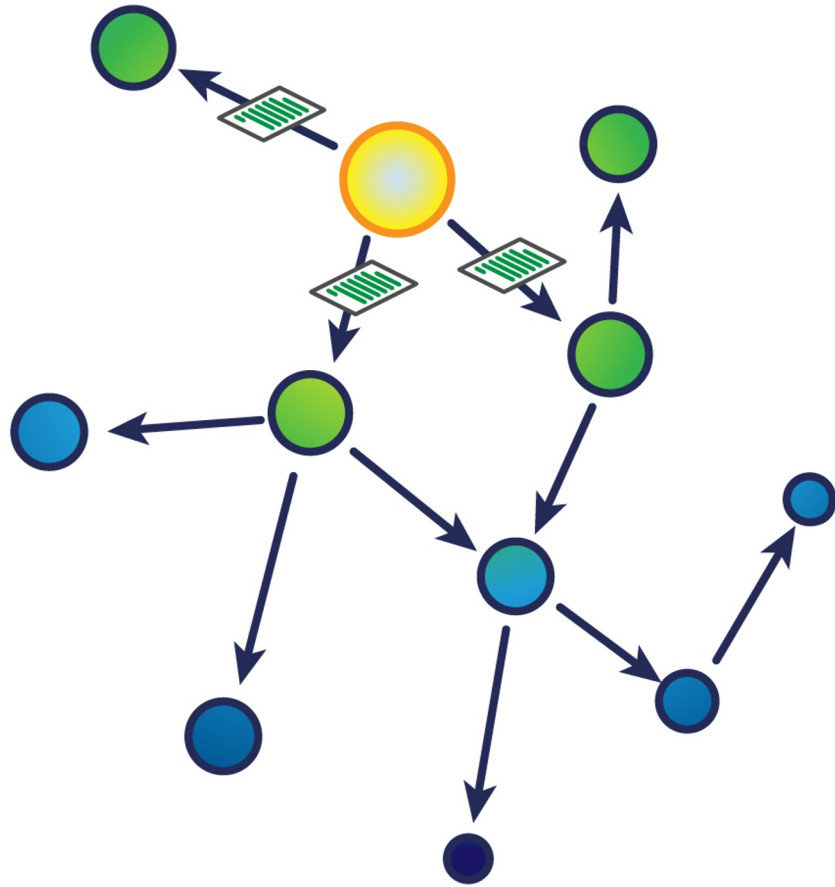
We can even enable graph neural network training such that confidential subgraphs may contribute to a result while residing in a secure location, and without leaking sensitive information.



We have previously utilized MPC for healthcare data.



Trust



Instead of having a central authority judge the reliability of each piece of information introduced to the system, instead focus on transparency and empowering users to judge source reliability.

Minimize the barrier to new participants providing data to the system, as we do not require a rigid gatekeeping process in order to prevent the new data from polluting overall results.

Workshop Agenda

□ Introduction

- *Chaitan Baru & Jemin George, TIP Directorate, National Science Foundation*

□ Presentation by Theme 1 Groups focusing on

○ Environment

- *Lilit Yeghiazarian, University of Cincinnati*

○ Biology & Health

- *Sergio Baranzini, University of California, San Francisco (UCSF)*

○ Justice

- *Adam Pah, Georgia State University (GSU)*

○ Technology & Manufacturing

- *Farhad Ameri, Arizona State University (ASU)*

□ Presentation by Theme 2: Proto-OKN Fabric

- *Chris Bizon, University of North Carolina at Chapel Hill (UNC) & Patrick Grinaway, Onai*

□ Presentation by Theme 3: Proto-OKN Education and Public Engagement

- *Cogan Shimizu, Wright State University*



EduGate: The Education Gateway to the Proto-OKN

Supported by:



Proto-OKN

Our Team



Cogan Shimizu
Antrea Christou
Brandon Dave



Pascal Hitzler
Hande Küçük
McGinty
Joseph Zalewski



Florence Hudson
Lauren Close
Emily Rothenberg



François Scharffe
Thomas Deeley
Hugo Sureau
Maru Wilson



Proto-OKN

Overview

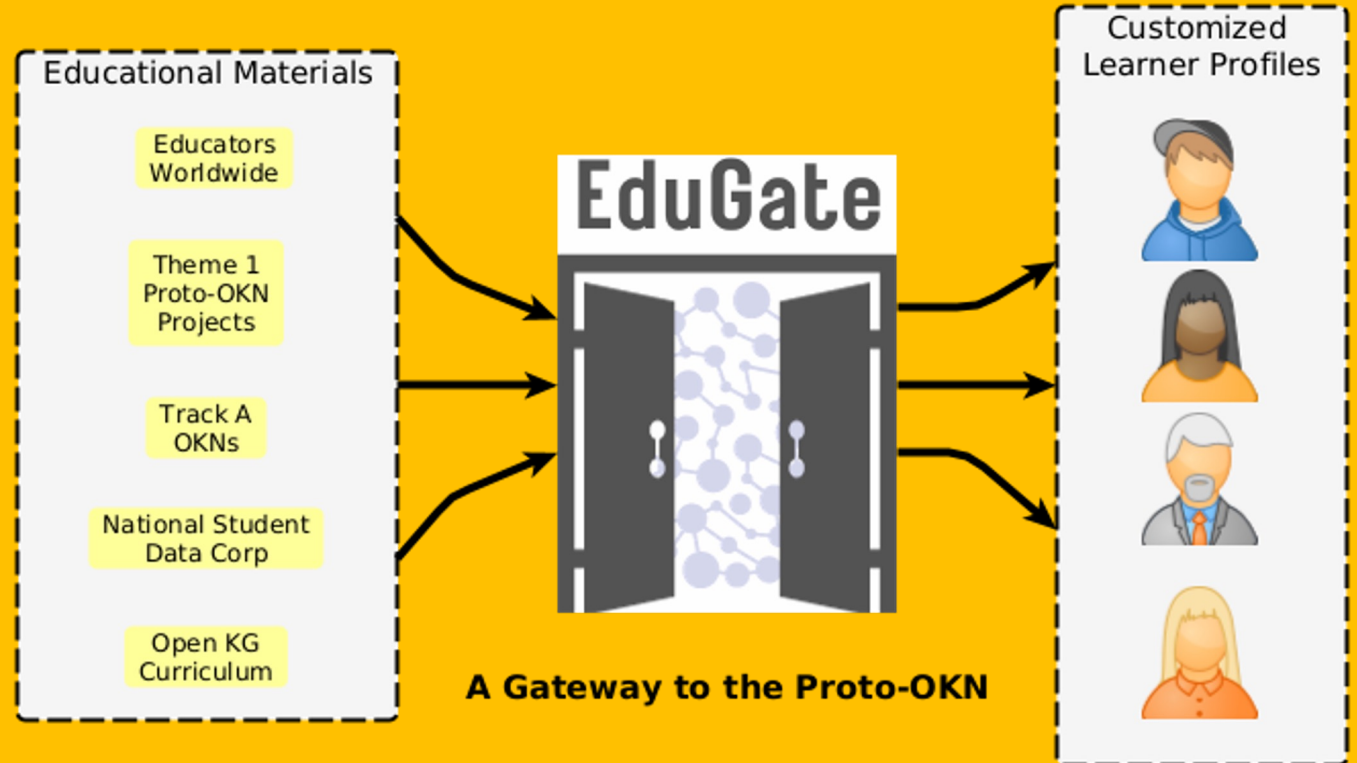
- Act as the entry point to the Proto-OKN
- Provide uniform access to richly integrated educational material, spanning
 - Technical concepts
 - Technology stacks (e.g., tutorials)
 - Individual use case documentation
- Engage stakeholders with detailed learner profiles that are optimized for different needs



Overview

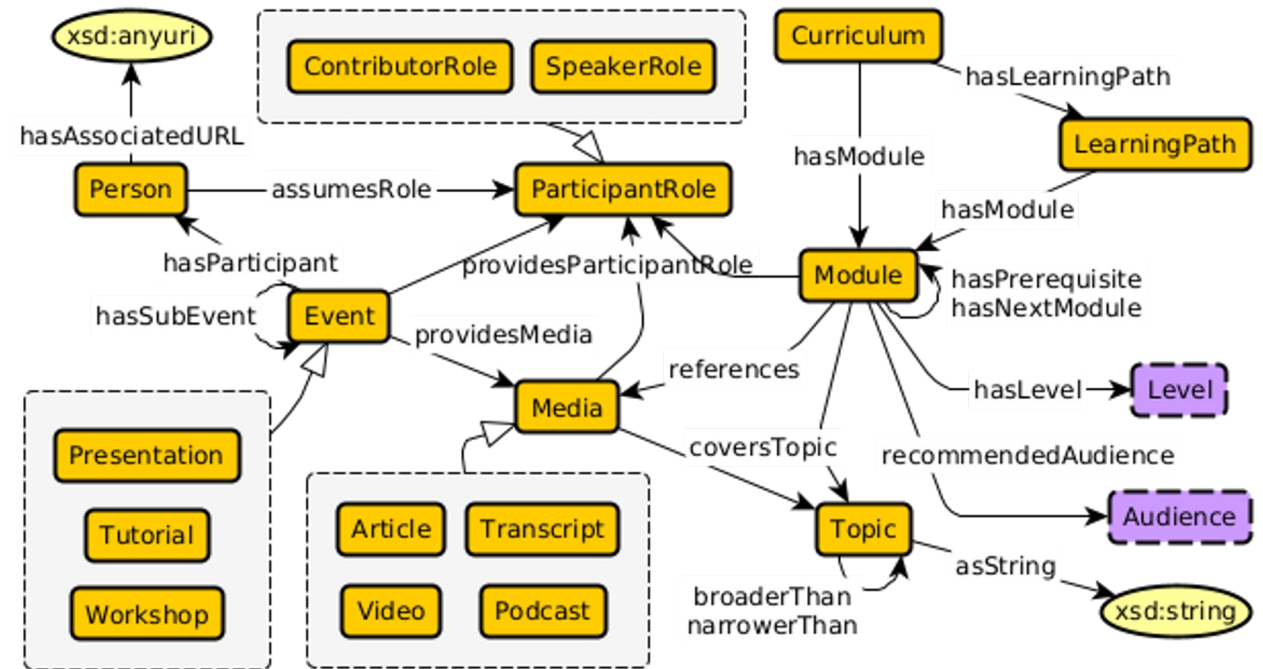
EduGate is collecting education, training, outreach, and learning materials from:

- educators around the world,
- our (Proto-)OKN colleagues,
- internal initiatives,
- our agency partners.



Overview

- The Education Gateway will be organized through its own KG
 - That connects people and teams
 - To their educational material
 - And retains proper licensing and provenance
- Syllabi to videos to slidedecks are all easily incorporated



Learner Profiles

<u>Enthusiast</u>	<u>Contributor</u>	<u>Developer</u>	<u>Executive</u>
<p>I am a graduate student pursuing my thesis. My research is in biomedical engineering, but my research path requires the use of Knowledge Graphs! However, I currently lack the background needed to implement the KG part of my research. I need something that can teach me the basics, but also fit into my busy research schedule.</p>	<p>At my organization, I am tasked with training junior data scientists; however, my organization uses a KG. I need to know how to standardize integration in the organization so junior data scientists can continue to grow out an existing knowledge graph as new data is gathered.</p>	<p>I am a Data Scientist/Engineer currently taking a new position where I need to know about KGs! As a developer, I am already proficient in the basics. While taking this new role in this project, I must add to my skills and preliminary knowledge to familiarize myself with KGs both in theory and in practice.</p>	<p>I am the CTO for my organization, in charge of the overall technical direction of the applications and any R&D. I need rapid insights, e.g., regarding functionality and tradeoffs, into whether or not this technology is worth pursuing for our applications, and therefore worth investing time and funds for an optimal outcome.</p>



Current & Future Working Groups

WGs consist of a chair and liaison Most meet biweekly Additional Groups will be founded on an as-needed basis	<u>Census Data</u>	<u>Geospatial</u>	<u>LLMs</u>
	<ul style="list-style-type: none">Common representation and integration of demographic data	<ul style="list-style-type: none">Common representation and integration of geospatial dataIncorporating KWG data	<ul style="list-style-type: none">Using LLMs with KGsCommon set of UI/UX Patterns and Library
	<u>Agency Bridge</u>	<u>Data & AI Ethics</u>	<u>Data Privacy</u>
	<ul style="list-style-type: none">Pulling together Agency initiatives and ensuring we have a common understanding of KGs	<ul style="list-style-type: none">Provide ethical guidelines for data use, protection, etc.Engage with NAIRR	<ul style="list-style-type: none">How to engage with multiple KGs where some of the data might be confidential, proprietary, or private



Thanks!

Please direct offline questions to edugate@wright.edu

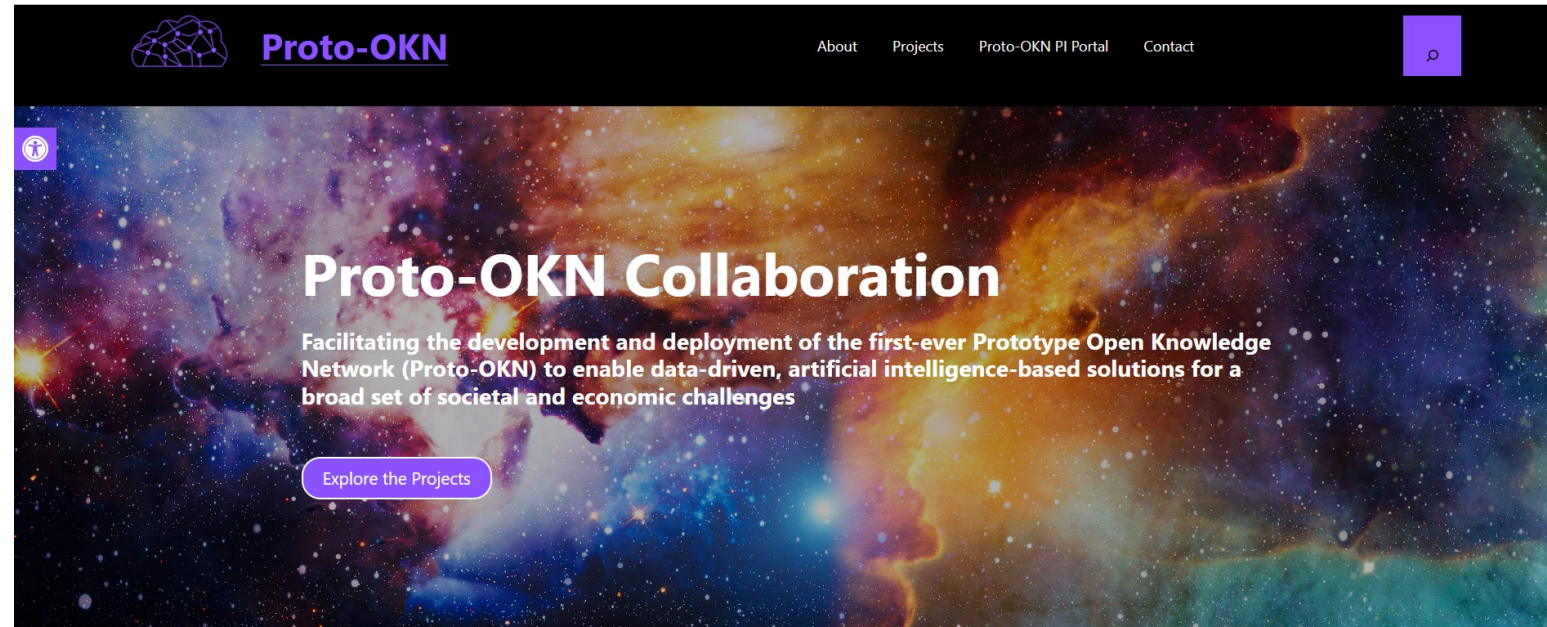


Proto-OKN

Program Site: [proto-okn.net](https://www.proto-okn.net)

How to get involved?

- **Attend any of the following:**
 - Outreach activities such as conferences/workshops
- **Join one of the Working Groups**
- **Join Slack channel**



NSF invests \$26.7 million in building the first-ever prototype open knowledge network

[Read the full NSF Press Release](#)



<https://www.proto-okn.net>



okn@nsf.gov

