

Knowledge Acquisition and Representation Methodology

Hande Küçük McGinty

hande@ksu.edu

April 30, 2023

Who am I?

Assistant Prof. at the Department of Computer Science at K-State



OHIO
UNIVERSITY



Agricultural
Research
Service

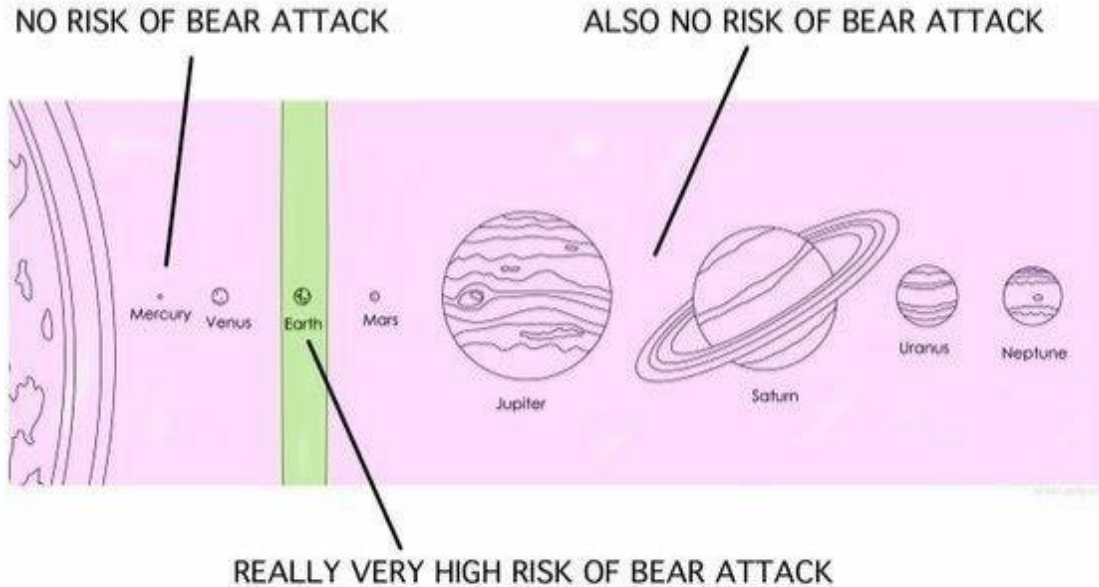


UNIVERSITY
OF MIAMI



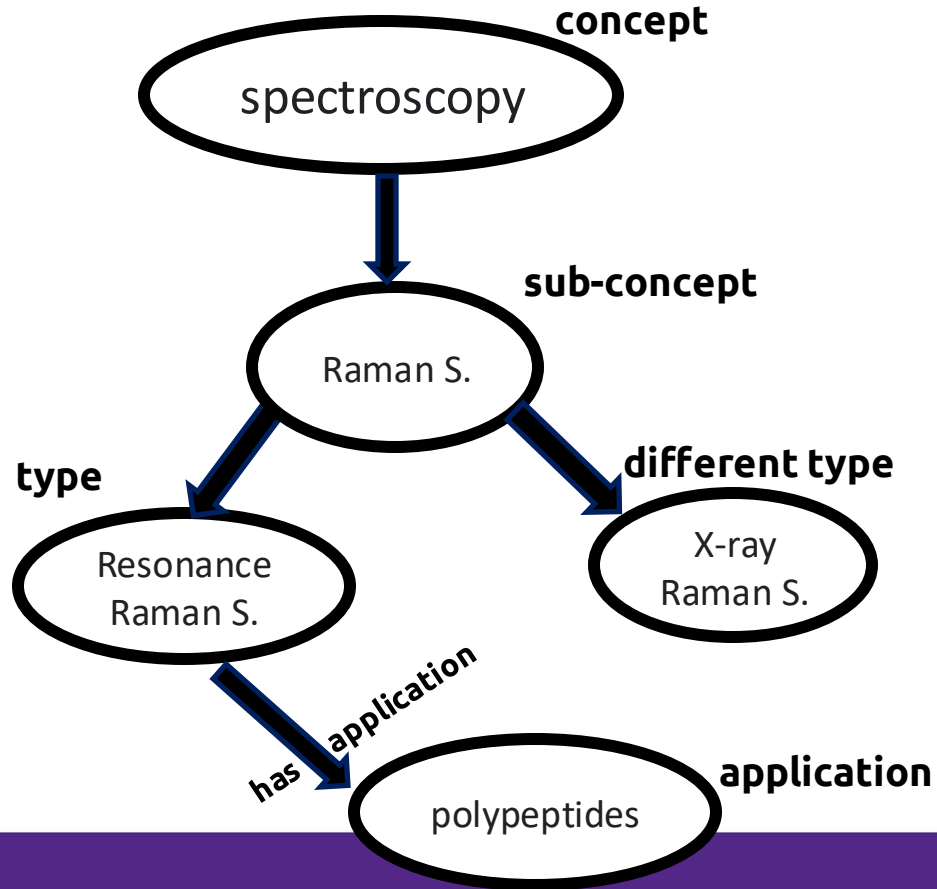
When statistical analysis doesn't tell us much..

CHART TO HELP DETERMINE RISK OF BEAR ATTACK:



What is an ontology?

- An ontology defines a set of concepts and relationships in a subject area to model it using formal logic.
- Ontologies show their concepts' properties and the relationships among them.



Why build ontologies / knowledge graphs?

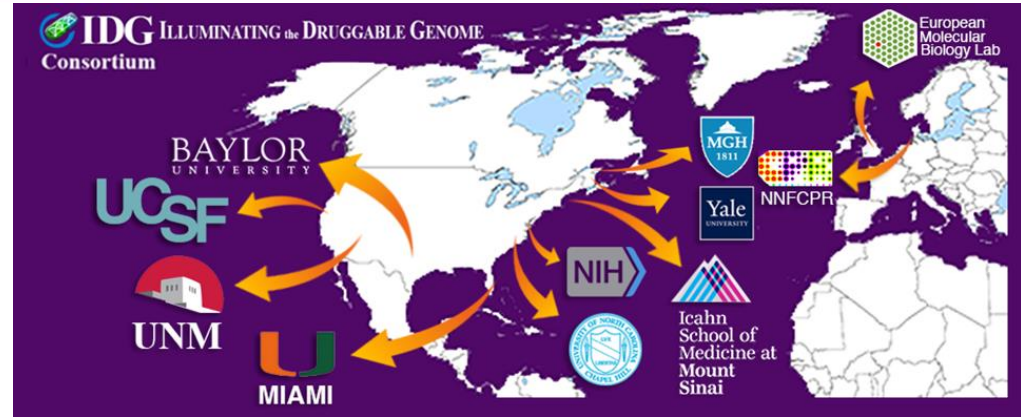
- Ontologies aim to :
 - limit complexity,
 - align domain experts' vocabulary within themselves in addition to machines',
 - organize data into information, knowledge and improve problem solving within that domain.



Three Nationwide Projects



NIH LINCS
PROGRAM

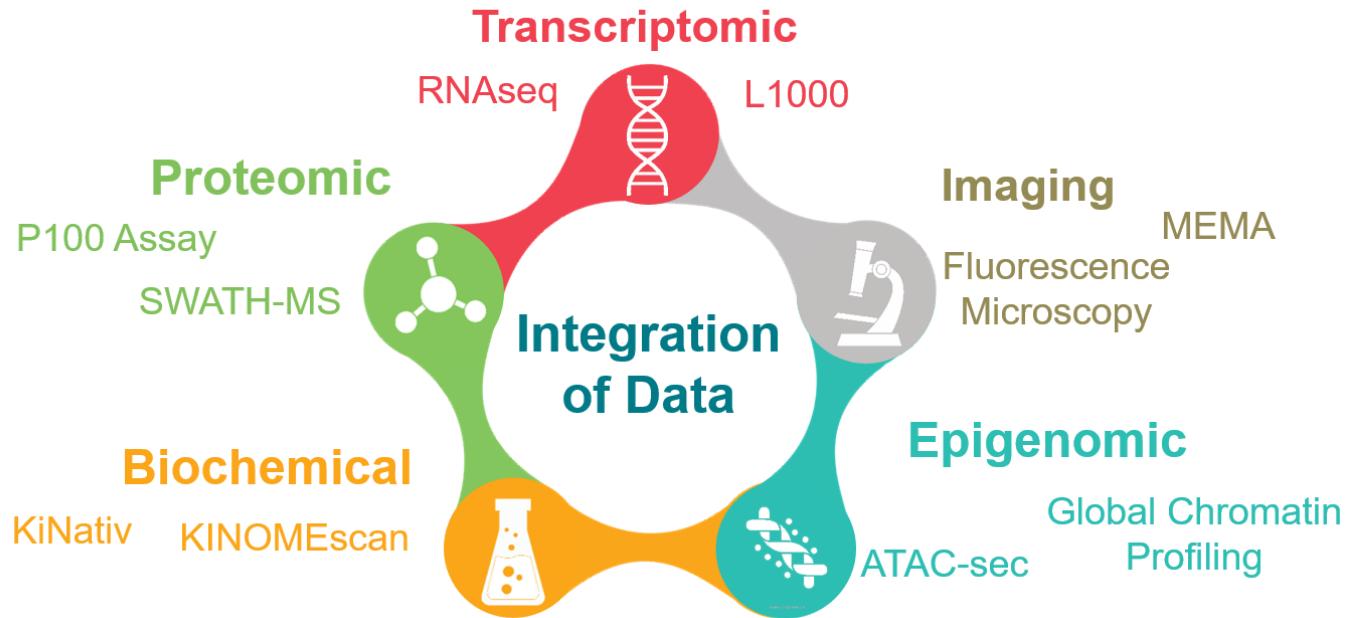


NIH grants

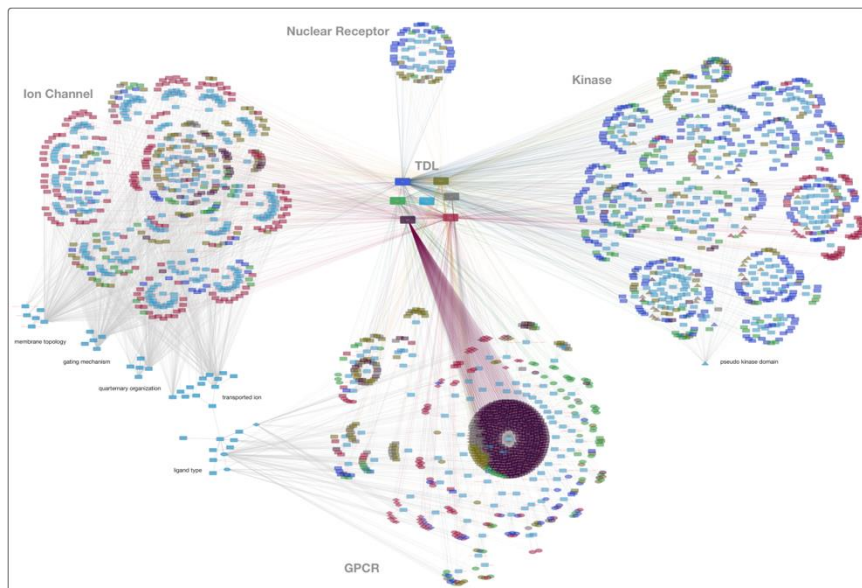
- 5U01HL111561-02 (LINCS Information Framework (LIFE) to Integrate and Analyze Diverse Data Sets)
- U54CA189205 (Illuminating the Druggable Genome Knowledge Management Center, IDG-KMC)
- U54HL127624 (BD2K LINCS Data Coordination and Integration Center, DCIC).

Bioassays in LINCS and Big Data To Knowledge Proj.

LINCS generates diverse multidimensional signatures



More data that should work concordantly..

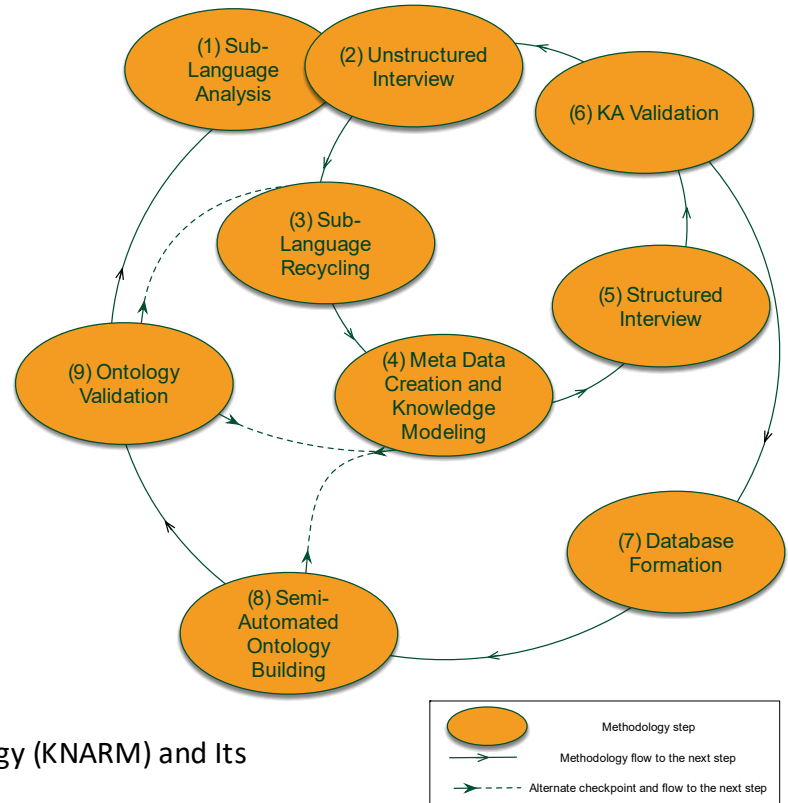


Family	Tbio	Tchem	Tclin	Tdark
Kinase	197	354	50	33
GPCR	131	127	96	52
Ion channels	106	84	128	25
NR	7	23	18	—

IDG Project Drug Target Proteins Target
Development Levels

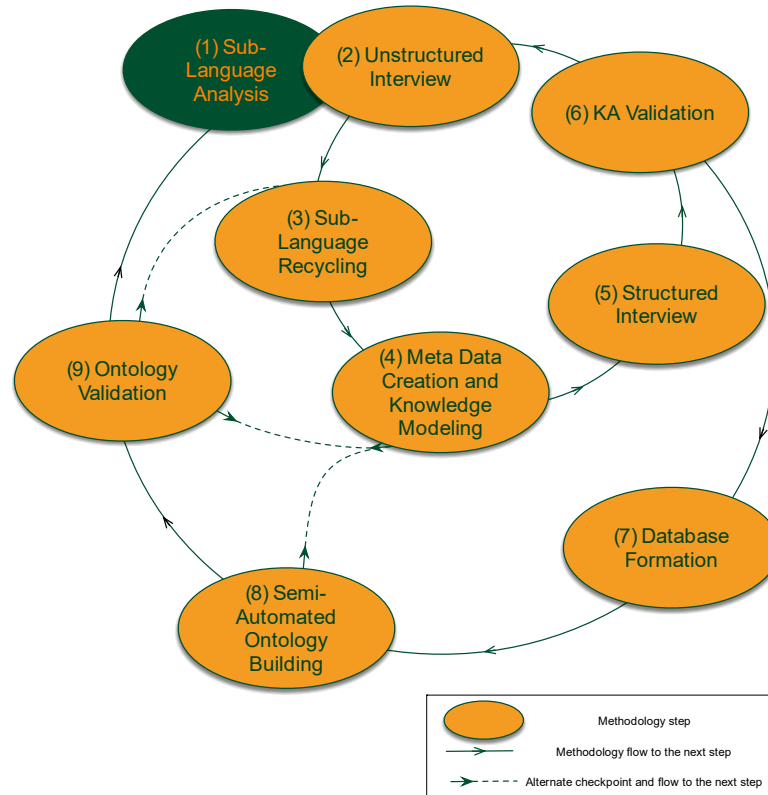
Hande Küçük McGinty, Lin, Yu, Saurabh Mehta, John Paul Turner, Dusica Vidovic, Michele Forlin, Amar Koleti et al. "Drug target ontology to classify and integrate drug discovery data." *Journal of biomedical semantics* 8, no. 1 (2017): 1-16.

KNnowledge Acquisition and Representation Methodology (KNARM)



McGinty, Hande Küçük. "Knowledge Acquisition and Representation Methodology (KNARM) and Its Applications." PhD diss., University of Miami, 2018.

Sub-language Analysis



Literature review



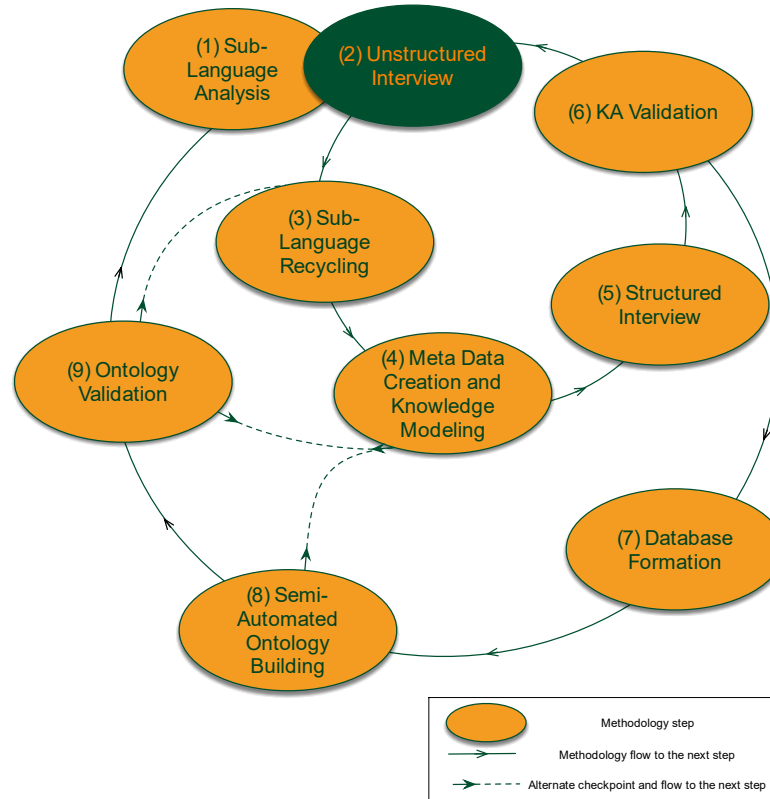
- * CDC Water Contamination Definitions
(<https://www.cdc.gov/healthywater/drinking/contamination.html>)
- * FDA Total Diet Study (<https://www.fda.gov/food/fda-total-diet-study-tds/>)
- * Agency for Toxic Substances and Disease Registry (ATSDR)
(<https://www.atsdr.cdc.gov/pfas/PFAS-health-effects.html>)
- * ATSDR Report "Toxicological Profile for Perfluoroalkyls", May 2021,
(<https://www.atsdr.cdc.gov/toxprofiles/tp200.pdf>)

What are the different datasets?



<https://riversideca.gov/press/understanding-pfas>

Unstructured Interview



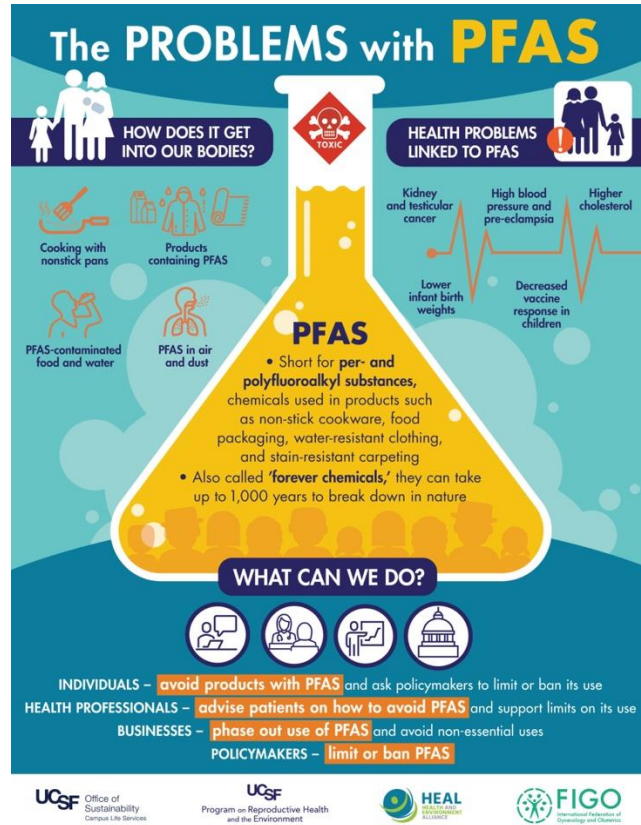
Understanding the questions



What may be some of our use-case problems:

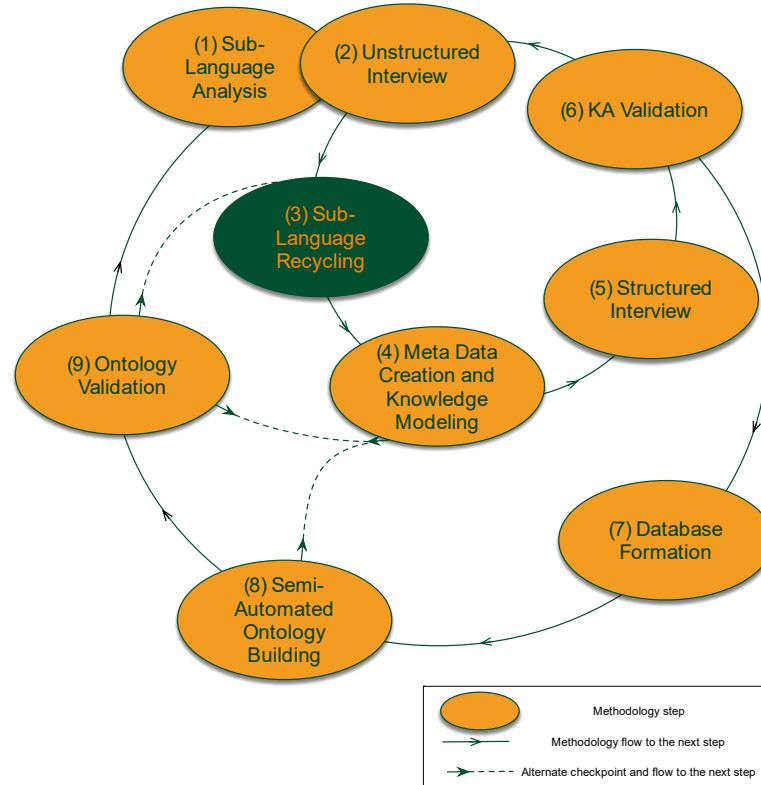
- * Understand the size of the problem,
- * Where and how to prioritize testing for PFAS,
- * How PFAS contamination is affecting the environment and health of their communities,
- * How to ensure safe food and water for their communities, and
- * Where there are serious gaps in the knowledge about PFAS in the state/country.

Understanding the use cases and stakeholders..



<https://www.env-health.org/how-pfas-chemicals-affect-women-pregnancy-and-human-development-health-actors-call-for-urgent-action-to-phase-them-out/>

Sub-Language Recycling



What are the surrounding efforts?

OBOFoundry/
OBOFoundry.github.io

Metadata and website for the Open Bio Ontologies
Foundry Ontology Registry



131

Contributors

181

Issues

5

Discussions

124

Stars

194

Forks



Environmental
ontology Ontology

Drug Target Ontology

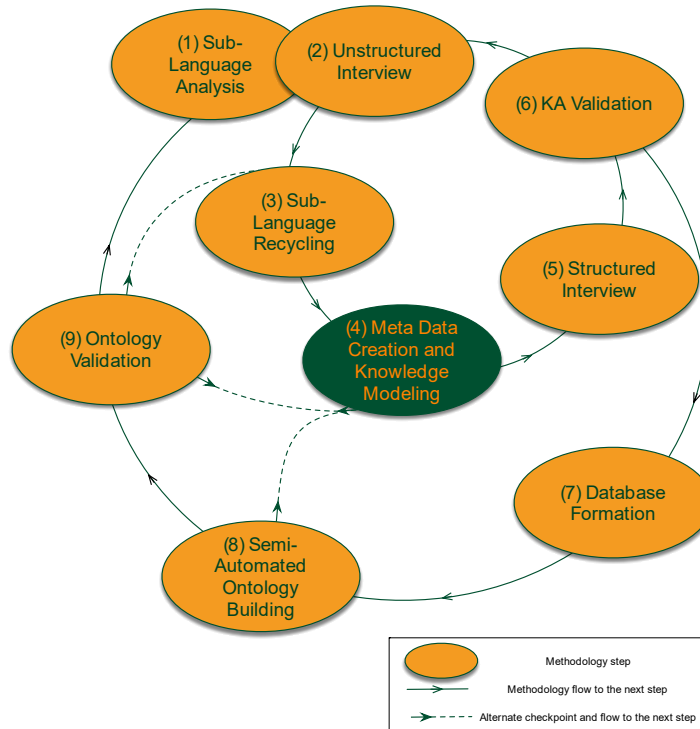
Chemical Entities
of Biological
Interest Ontology

BRENDA Cell Line and
Tissue Ontology

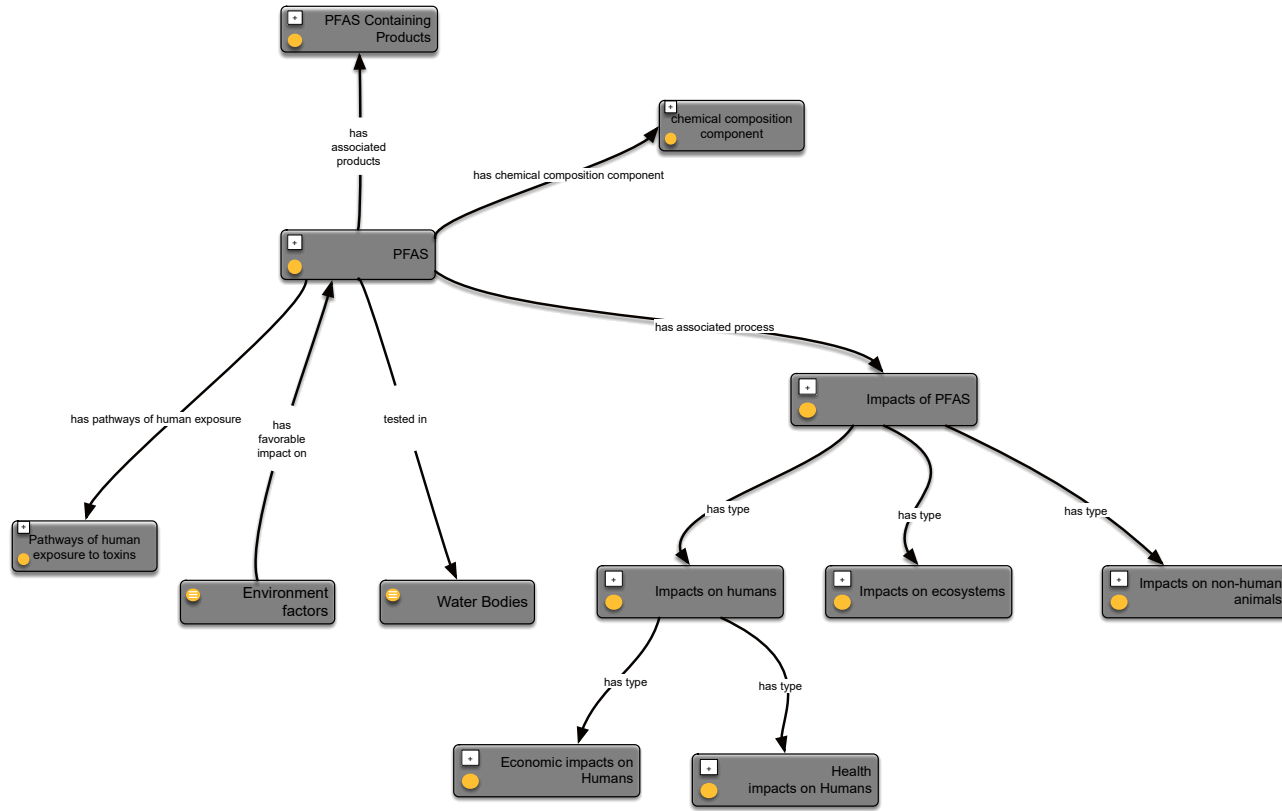
Disease Ontology

FDC Ontology

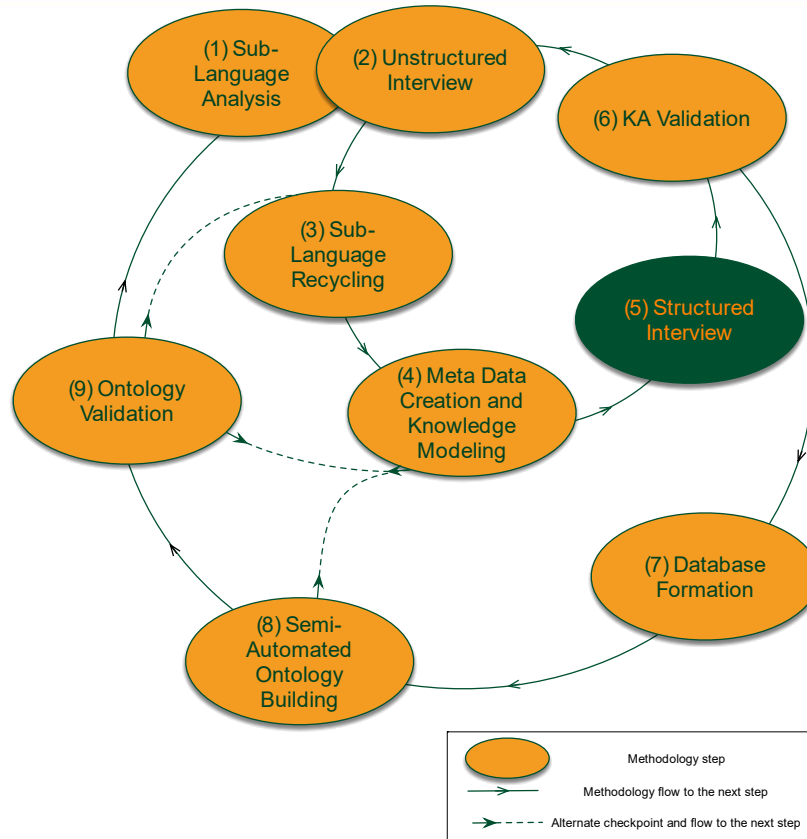
Adapted from Kai Blumberg



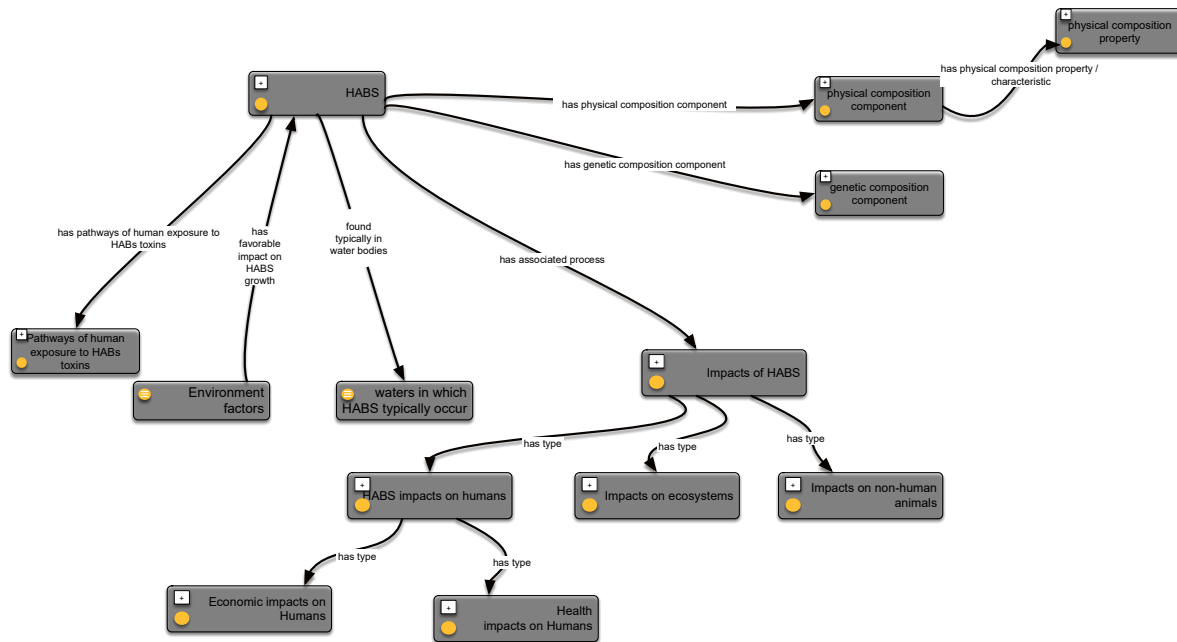
Modeling Example



Structured Interview

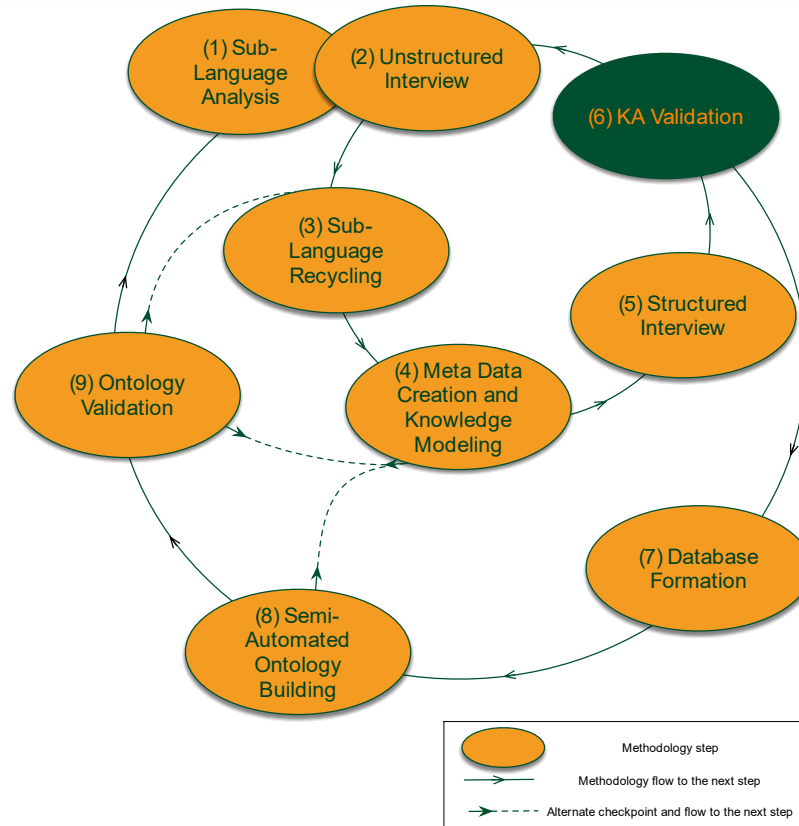


Because of the time restrictions – reused research..



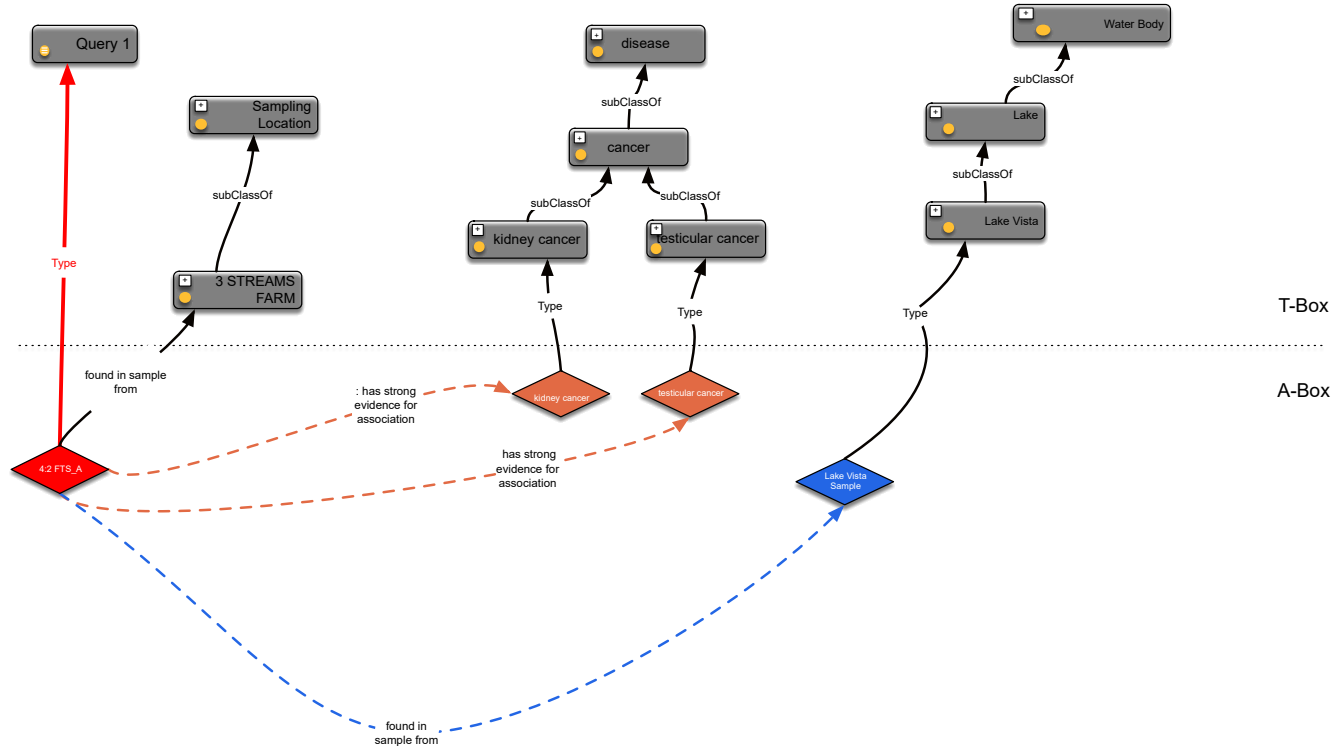
Leveraging Existing Literature on the Web and Deep Neural Models to Construct a Knowledge Graph Focused on Water Quality and Health Risks, Nikita Gautam, David Shumway, Megan Kowalczyk, Sarthak Khanal, Doina Caragea, Cornelia Caragea, **Hande McGinty** and Samuel Dorevitch, 2023, theWebConf, 11:50 AM – 12:00 PM – May 3rd, @AT&T Hotel and Conference Center Classroom #115

Knowledge Acquisition Validation

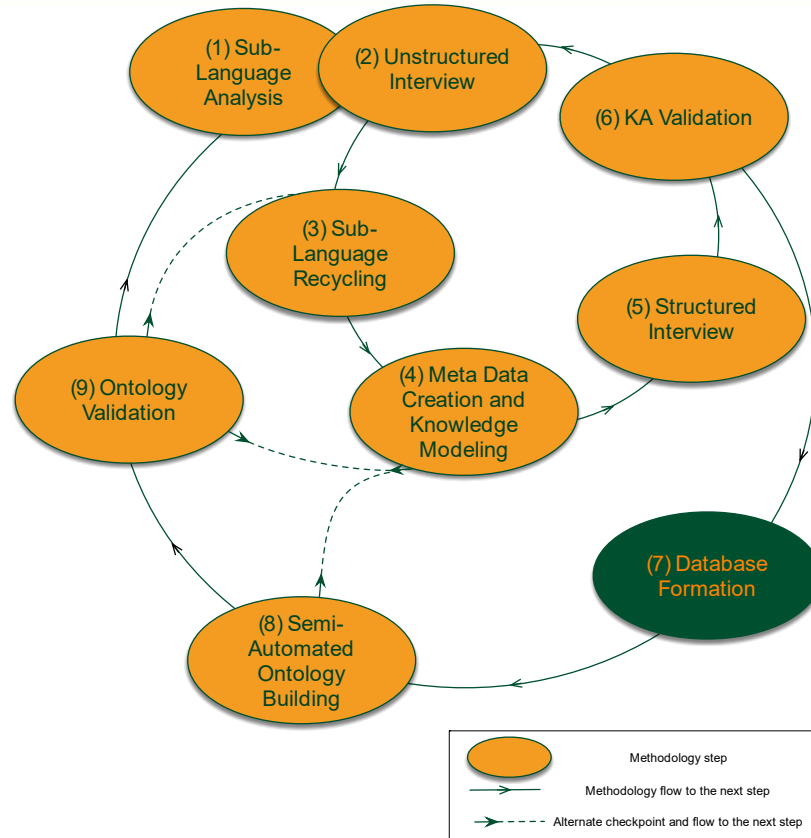


Ensure that we can answer questions of interest

What are the chemicals that are found near 3 Streams Farm that may cause kidney cancer?

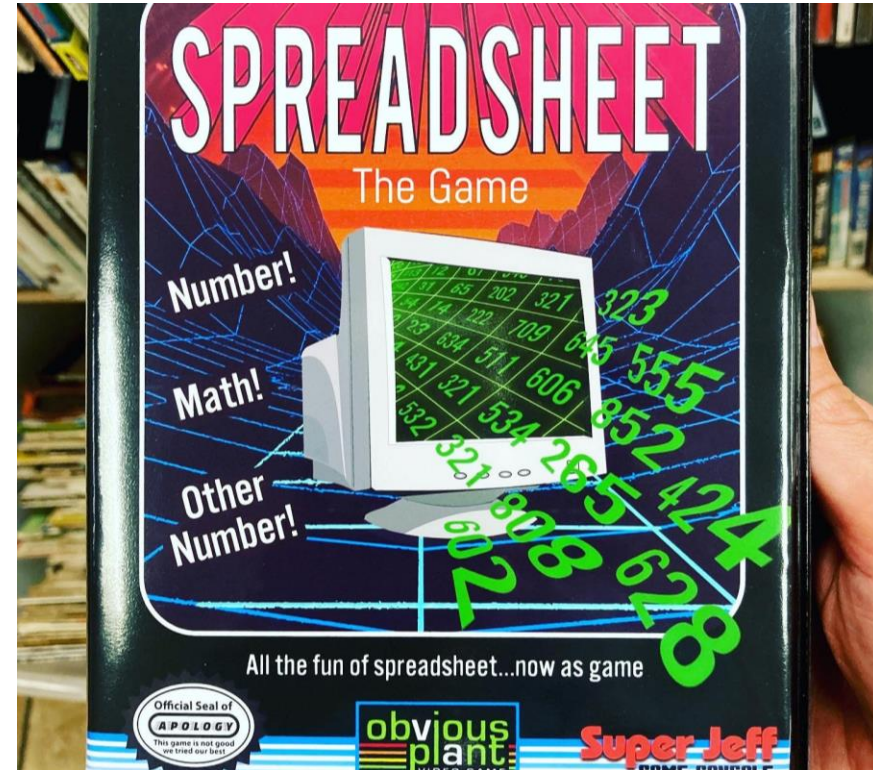


Database Formation



Sustainable Data Storage

CURRENT SITE NAME	Town	SAMPLE POINT SEQ	SAMPLE DATE	SAMPLE TYPE	PARAMETER	CONCENTRATION	UNITS	LAB QUALIFIER	RL	TS
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	HFPO-DA A		ng/L	U	43	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFDOA A		ng/L	U	344	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFOS A	281	ng/L		1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFUNDA A		ng/L	U	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	N-MeFOSAA		ng/L	U	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFHxA A	32.6	ng/L		1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFPS A		ng/L	U	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	6:2 FTS A		ng/L	U	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	N-EFOSAA	2.89	ng/L	J	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFHxA A	34.8	ng/L		1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFDOA A		ng/L	U	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFDA A	113	ng/L		1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFDA A	5.9	ng/L		1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFOS A		ng/L	U	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFHxA A	6.27	ng/L		1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFBA A	12.7	ng/L		1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFBS A		ng/L	U	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFHxA A	26.9	ng/L		1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFPS A	3.37	ng/L		1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFNA A	12.3	ng/L		1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFTEA A		ng/L	U	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	6:2 FTS A		ng/L	U	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFHxA A		ng/L	U	3.44	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFNS A		ng/L	U	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFHxA A		ng/L	U	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFDA A		ng/L	U	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	6:2 FTS A		ng/L	U	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	ADONA A		ng/L	U	1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	PFDA + PFOS	394	ng/L		1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39173	12/30/21	L	SUM OF 6 PFAS	445	ng/L		1.72	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39569	12/30/21	L	HFPO-DA A		ng/L	U	62	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39569	12/30/21	L	PFDOA A		ng/L	U	4.96	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39569	12/30/21	L	PFUNDA A	0.953	ng/L	J	2.48	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39569	12/30/21	L	N-MeFOSAA	8.16	ng/L	J	2.48	N
PIXELLE ANDROSOGGIN JAY LANDFILL	JAY	39569	12/30/21	L	PFHxA A	428	ng/L	J	2.48	N



DB view of Samples, Chemicals and Concentration Values

LOAD CSV WITH HEADERS FROM 'file:///Book.csv'
AS row

MERGE (loc:Location {name: row.Location})

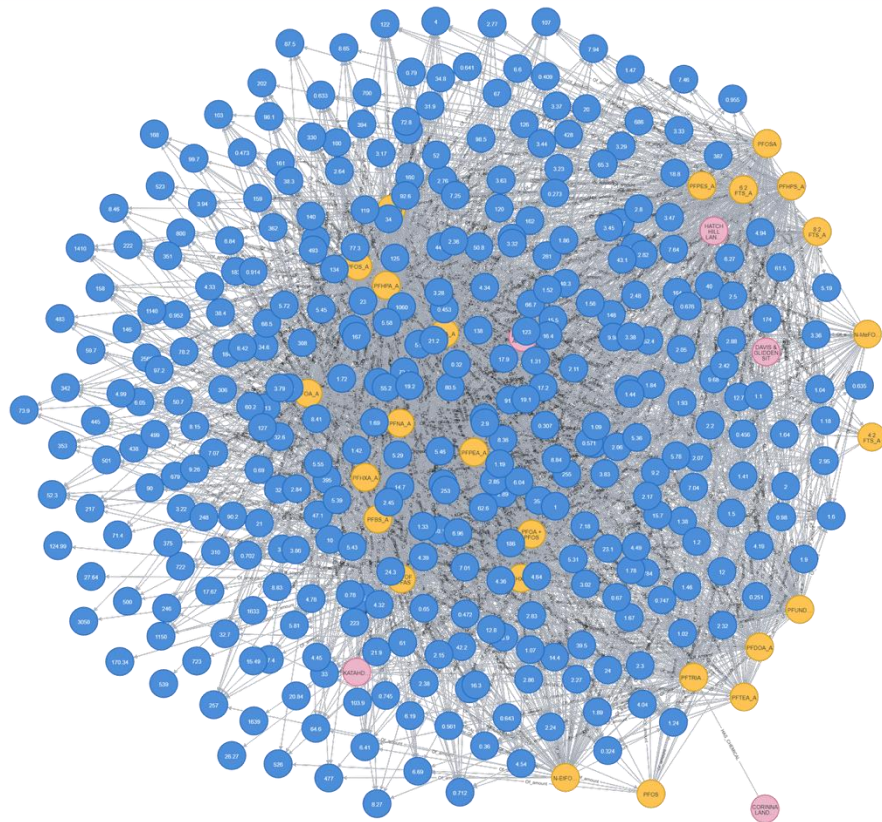
MERGE (Chem:Chemical {name: row.Chemical})

MERGE (cont:Content {name: row.Content})

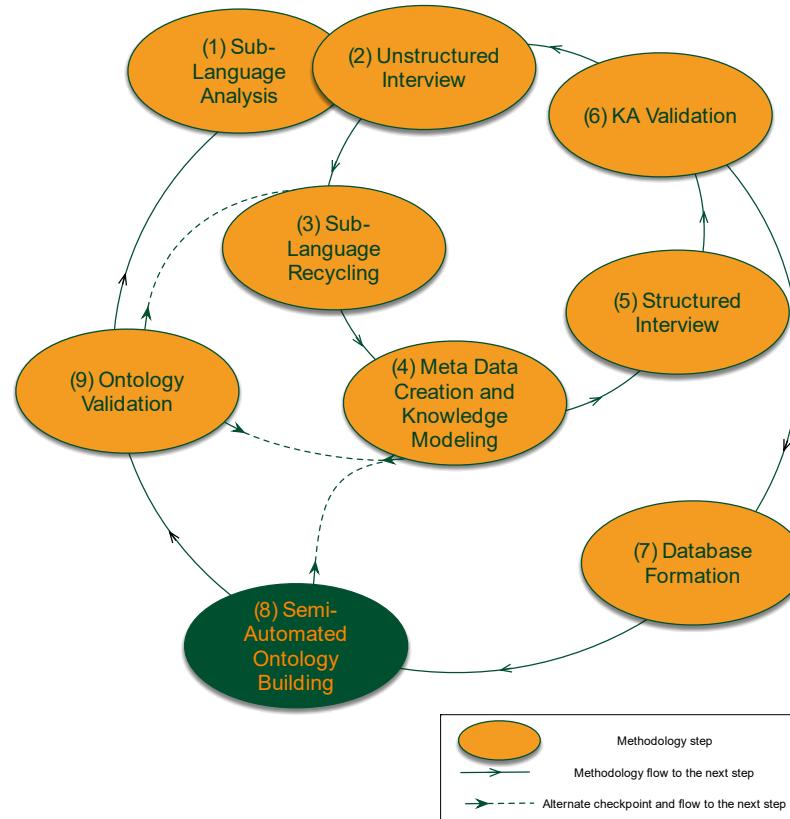
MERGE (loc)-[:HAS_CHEMICAL]->(Chem)

MERGE (Chem)-[:Of_amount]->(cont)

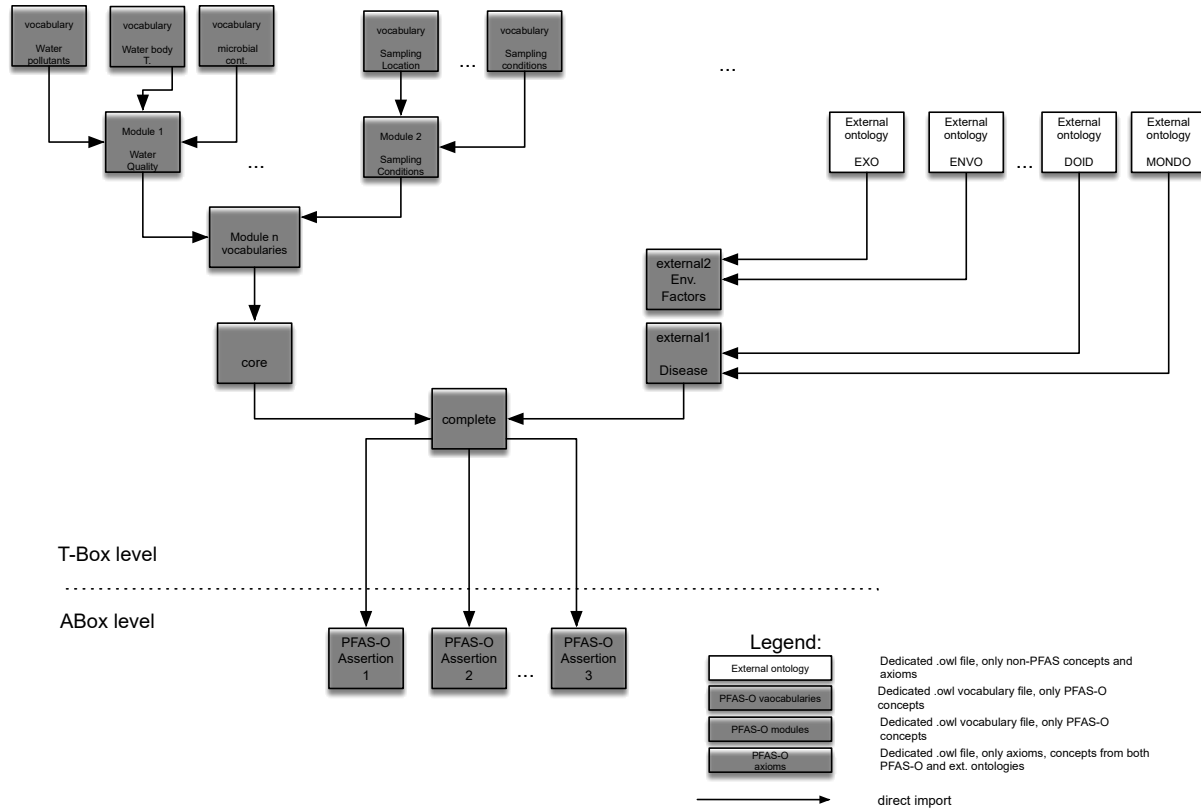
RETURN loc, cont, Chem



Semi-Automated Ontology Building



Semi-Automated Modular Architecture

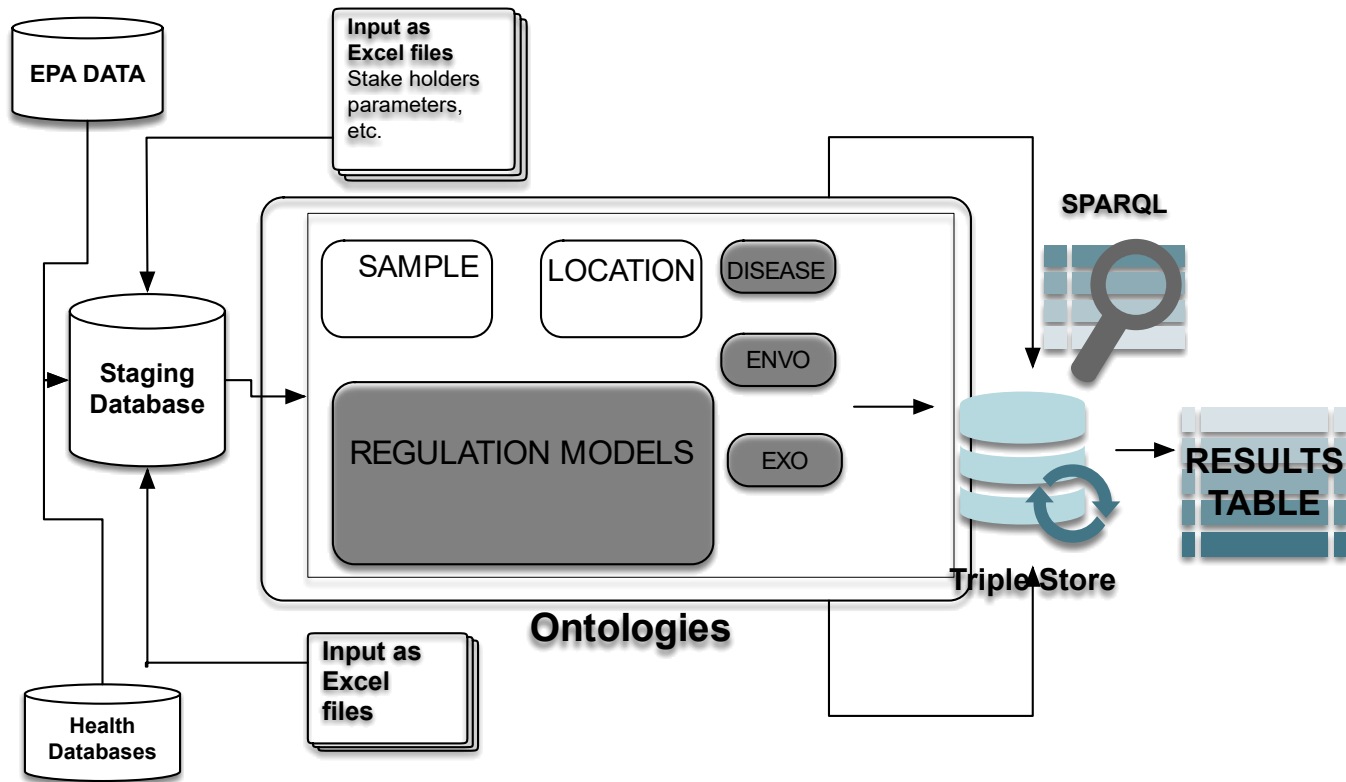


Semi-Automated Ontology Output

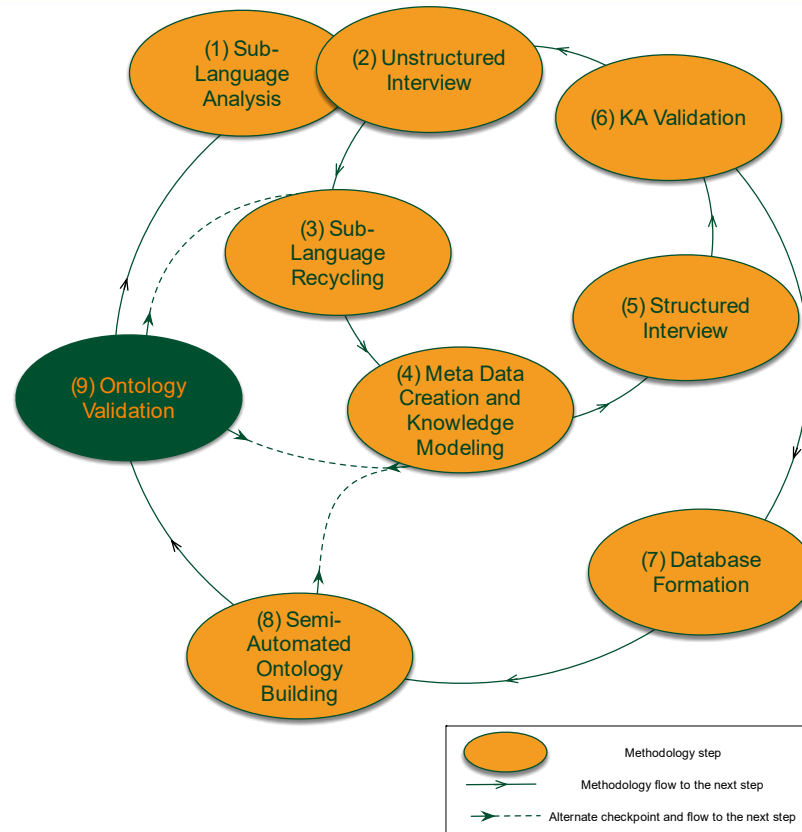
The screenshot displays a web-based OWL ontology editor. The browser's address bar shows the URL: `sample_axiom_module (http://www.semanticweb.org/handemcinty/ontologies/2023/sample_axiom_module.owl)`. The left sidebar shows a class hierarchy for `4:2 FTS_A` under `owl:Thing`. The main area displays an 'OntoGraf' visualization, which is a directed graph showing relationships between various entities, including `'SINCLAIR POTW'`, `'MARCAL PAPER SLUDGE DUMP FIELDS'`, `'MAINE RESOURCES'`, and `'A C LAWRENCE CO LANDFILL'`. The graph is rendered with orange lines connecting nodes.

```
robot template --input /Users/yz/Desktop/csv_to_owl/output/parameter.owl \
--template /Users/yz/Desktop/csv_to_owl/axiom_edit.csv \
--prefix "id:https://fdc.nal.usda.gov/fdc/FDC" \
--output /Users/yz/Desktop/csv_to_owl/output/axiom_edit.owl
```


These Systems are complex, so we need a systematic approach to build them!



Ontology Validation



Quality Control

Active ontology x Entities x Individuals by class x DL Query x Individual Hierarchy Tab x

Annotation properties Datatypes Individuals

Classes Object properties Data properties

Annotations Usage OntoGraf

Class hierarchy: AROOSTOOK WASTE SOLUTIONS

Asserted

- PFTRDA_A
- PFTRIA_A
- PFUNDA
- PFUNDA_A
- SUM OF 5 PFAS
- SUM OF 6 PFAS
- TOTAL PFCA
- disease
- location
 - 104 JENNIFER DRIVE-AUBURN
 - 11 EAST PITSTON RD
 - 1969 BANGOR ROAD - CLINTON
 - 198 RANGE ROAD-CUMBERLAND
 - 3 STREAMS FARM
 - 36 MOOSE HILL ROAD - WEST GARDINER
 - 788 US RT1-MADAWASKA
 - 9 RIVER STREET - ROXBURY
 - A C LAWRENCE CO LANDFILL
 - A FITZPAT SITE HOME FARM
 - ABBOT LANDFILL
 - AFFF SPILL - RICHMOND
 - AIM LANDFILL BUCKSPORT RT 15
 - AIRPORT RD-MACHIAS
 - ALEXANDER ROAD-VAN BUREN
 - ALFRED WATER DISTRICT
 - ALNA LANDFILL
 - AMERICARB, INC
 - ANDROSCOGGIN RIVER - AGI
 - ANDROSCOGGIN RIVER - ALS
 - ANDROSCOGGIN RIVER - ALV
 - ANDROSCOGGIN RIVER - ARF
 - ANDROSCOGGIN RIVER - ARP
 - ANDROSCOGGIN RIVER - ARY
 - ANDROSCOGGIN RIVER-FW08ME027-NRSA
 - AQUA ME-CAMDEN & ROCKLAND
 - ARNDT-JACKSON SITE FIELDS
 - AROOSTOOK RIVER - ACB
 - AROOSTOOK WASTE SOLUTIONS**
 - ARUNDEL LANDFILL

Search: contains Search Clear

OntoGraf:

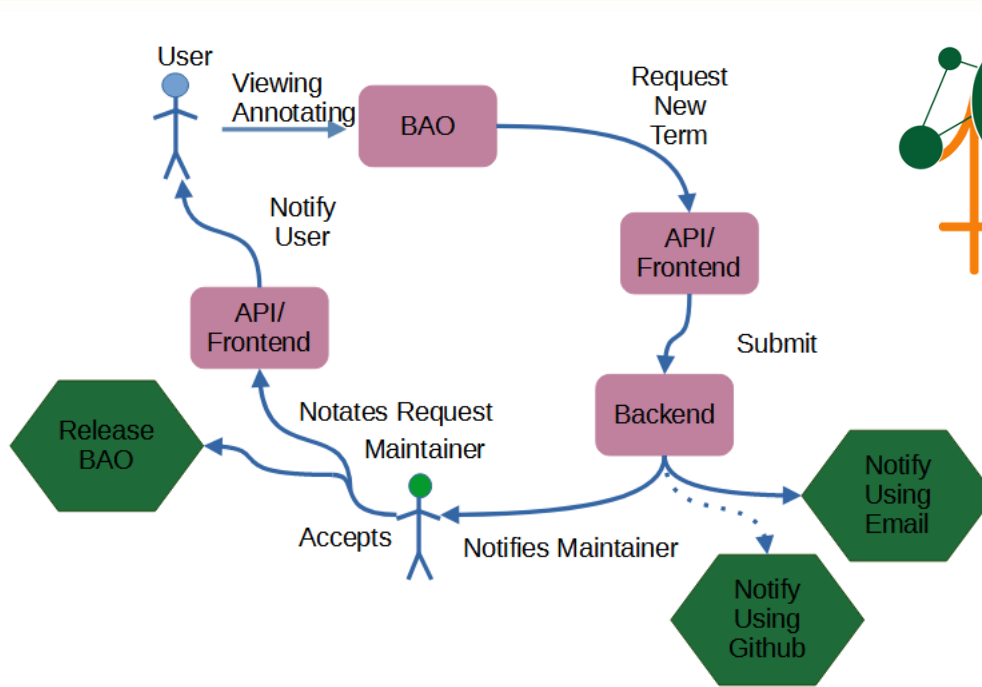
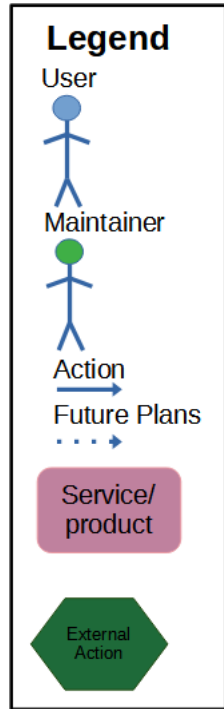
Description: AROOSTOOK WASTE SOLUTIONS

Equivalent To +

SubClass Of +

- 'has chemical found in sample' some '4:2 FTS_A'
- 'has chemical found in sample' some '6:2 FTS_A'
- 'has chemical found in sample' some '8:2 FTS_A'
- 'has chemical found in sample' some 'PFOA + PFOS'

OntoloBridge



UNIVERSITY
OF MIAMI



Hande Küçük McGinty, John Paul Turner, Alex M. Clark, Peter Gedeck, John Graybeal, Michael Dorf, Caty Chung, Mark Musen, Barry A. Bunin and Stephan Schürer;
OntoloBridge – A FAIR Semi-Automated Ontology Update Request System; (In preparation)

NIH - 1U01LM012630-01 (Unifying Templates, Ontologies and Tools to Achieve Effective Annotation of Bioassay Protocols (OntoloBridge))

Use it as a part of backend for Applications using Machine Operable Data

Assay

Similar Assays

relation between CCR6 and colorectal liver metastasis. The association between CCR6 expression levels in 64 primary tumor specimens in primary CRC and synchronous liver metastases suggests that CCR6 and CCL20 are involved in the metastatic spread to the liver. A small molecule tool would address a key hypothesis: Modulation of the CCR6/CCL20 axis will regulate pathogenic activities of B cells in a variety of diseases including hematopoietic malignancy and cancer metastasis.

The project goal is to identify a chemical probe of CCR6 receptor that can specifically act as 'chemical modulator' of CCR6 through inhibition (antagonism) of functional response. An important objective of this research program is to provide new insight into the regulation of cancer metastasis modulated by the CCR6/CCL20 (MIP-3 alpha) axis.

Request Suggestions Stop

Autogenerated Text

This is a confirmatory assay to identify potential treatments for hematologic cancer, by investigating the biological process of chemokine-mediated signaling pathway, regulation of C-C chemokine binding and regulation of C-C chemokine receptor CCR7 signaling pathway, in Homo sapiens, specifically targeting C-C chemokine receptor type 6, C-C chemokine receptor type 6 (human) and C-C chemokine receptor type 6 [Homo sapiens].

This is a beta galactosidase enzyme activity assay.

assay cell line ?

CHO-K1 细胞

organism ?

Homo sapiens

biological process ?

regulation of C-C chemokine receptor CCR7 signaling pathway

chemokine-mediated signaling pathway

regulation of C-C chemokine binding

target ?

C-C chemokine receptor type 6

C-C chemokine receptor type 6 (human)

androgen receptor

coagulation factor V

DnaJ homolog subfamily C member 15

1,2-phenylacetyl-CoA epoxidase, subunit C

[Fe-S]-dependent transcriptional repressor FecC

anaerobic glycerol-3-

androgen receptor (m

androgen receptor (ra

Matching assays: 100

All None Show Compounds

Assay Grid

Predict

Bulk Remap

Copy

Download

Query

Similarity Compounds

Summary

Title

Develop

Disease

Models

ID

<input type="checkbox"/>	100%	0	Compound toxicity assay for inhibition of an unknown target using a cytotoxicity assay, in a cell-free format with an ATP quantitation using luciferase and viability measurement method and bioluminescence-based detection to determine EC50 as units of micromolar, threshold unknown.	View	AID 2253
<input type="checkbox"/>	100%	0	Primary assay for inhibition of the target oxidoreductase using an enzyme activity assay, in a cell-free format with an enzyme activity measurement method and absorbance-based detection to determine percent inhibition as units of percent, threshold greater than (>) 50.0.	View	AID 2304
<input type="checkbox"/>	100%	0	Lead optimization assay for inhibition and modulation of the target protein using a phosphorylation assay, in a cell-free format with a protein abundance method and fluorescence intensity-based detection to determine raw activity as units of unknown, threshold unknown.	View	AID 504478
<input type="checkbox"/>	100%	0	Counter screening assay and selectivity assay for agonism of the target Nuclear hormone receptor, PPARG, Peroxisome proliferator-activated receptors and see Gene ID using a reporter gene assay, in a cell-free format with a luciferase induction and reporter gene method and luminescence method-based detection to determine EC50 as units of micromolar, threshold less than (<) 2.0.	View	AID 504735
<input type="checkbox"/>	100%	0	Confirmatory assay and counter screening assay of an unknown target, in a cell based format and cell-free format and fluorescence intensity-based detection to determine raw activity as units of unknown, threshold unknown.	View	AID 720681
<input type="checkbox"/>	100%	0	Confirmatory assay and counter screening assay of an unknown target, in a cell-free format and fluorescence intensity-based detection to determine raw activity as units of unknown, threshold unknown.	View	AID 720682
<input type="checkbox"/>	100%	0	Confirmatory assay and counter screening assay of an unknown target, in a cell-free format and fluorescence intensity-based detection to determine raw activity as units of unknown, threshold unknown.	View	AID 720683
<input type="checkbox"/>	100%	0	Confirmatory assay and counter screening assay of an unknown target, in a cell-free format and fluorescence intensity-based detection to determine raw activity as units of unknown, threshold unknown.	View	AID 720684
<input type="checkbox"/>	100%	0	Confirmatory assay of an unknown target, in a cell-free format and fluorescence intensity-based detection to determine raw activity as units of unknown, threshold unknown.	View	AID 720685
<input type="checkbox"/>	100%	0	Unknown stage of an unknown target, in a cell-free format and fluorescence intensity and fluorescence method-based detection to determine raw activity as units of unknown, threshold unknown.	View	AID 720686
<input type="checkbox"/>	67%	0	Counter screening assay for inhibition of the target unknown using a protease activity assay and protein-protein interaction assay, in a protein complex format and subcellular format with a fluorescent ligand binding method and substrate coupled enzyme activity measurement method and fluorescence intensity-based detection to determine percent inhibition as units of percent, threshold unknown.	View	AID 2132
			Alternate assay format for inhibition of the target unknown using a transferase activity assay, in a whole cell lysate format with a DNA or RNA abundance		

** Text \Rightarrow Annotation

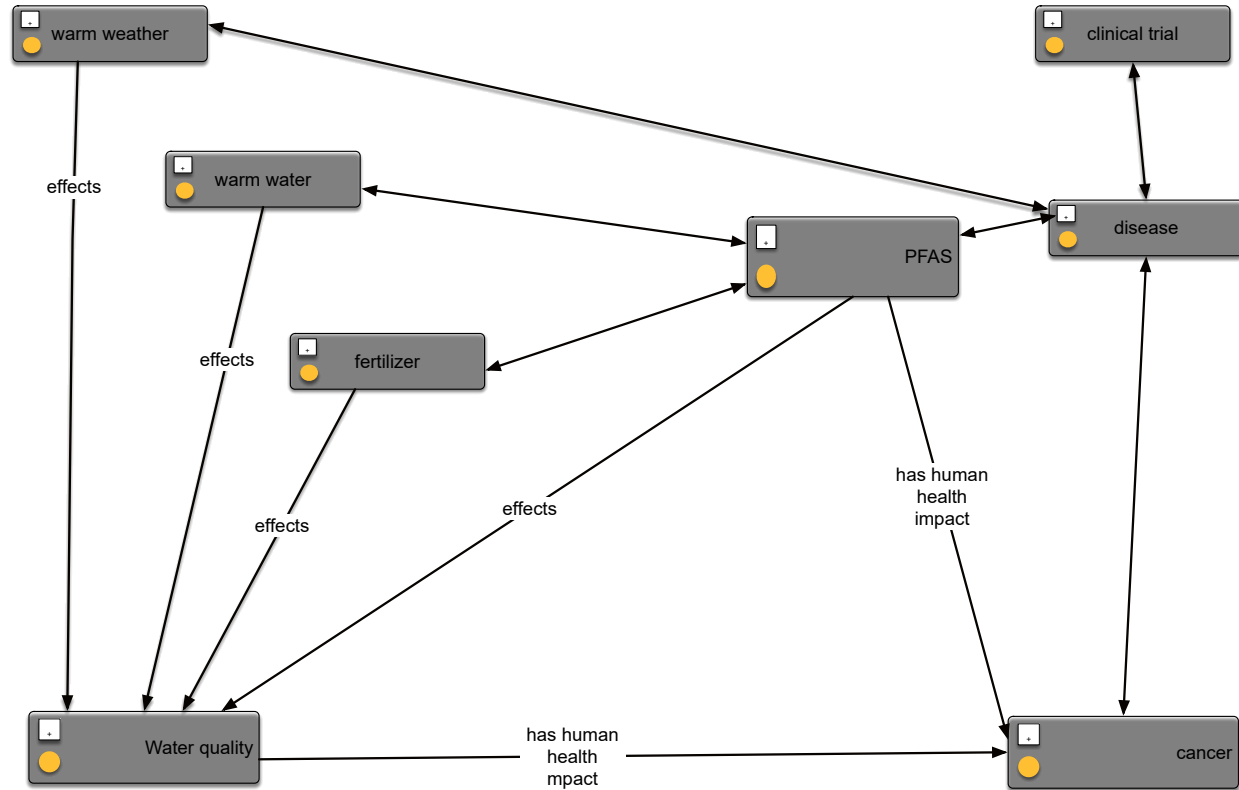
○ Dictionary

○ Naïve Bayes classifier

** Annotations \Rightarrow Annotation

○ Associative rules mining

Scalable Knowledge Graph Generation



Acknowledgements



NIH, USDA and FDA funded projects

- RC2HG005668 (BioAssay Ontology and Software Tools to Integrate and Analyze Diverse Data Sets)
- 5U01HL111561-02 (LINCS Information FramEwork (LIFE) to Integrate and Analyze Diverse Data Sets)
- U54CA189205 (Illuminating the Druggable Genome Knowledge Management Center, IDG-KMC)
- U54HL127624 (BD2K LINCS Data Coordination and Integration Center, DCIC).
- 1U01LM012630-01 (Unifying Templates, Ontologies and Tools to Achieve Effective Annotation of Bioassay Protocols (OntoloBridge))
- 2R44TR000185-04 (BioAssay Express Phase 2)
- 3R43TR002528-01S1 (NIH I-CORPS Grant Supplement to SBIR Phase 1 Grant related with Chemical Mixtures)
- 2R44TR002528-02 (Digital Representation of Chemical Mixtures to Aid Drug Discovery and Formulation)

Teammates

- My students: Yinglung Zhang and Aryan Dalal
- Harrington Lab – Ohio U
- Musen Lab & BioPortal Team – Stanford
- Collaborative Drug Discovery Team
- Schurer Lab – Uni. Of Miami, Coral Gables, FL

Check us out again!

Monday, May 1st

Workshop: Knowledge Graphs for Sustainability - KG4S

Workshop

🕒 9:00 AM – 12:30 PM

📍 University of Texas Gates Dell Complex - Room # 6.202

11:05-11:25 Short paper presentation:

Yinglun Zhang, Antonina Broyaka, Jude Kasten, Allen Featherstone, Cogan Shimizu, Pascal Hitzler and Hande Küçük McGinty. *Sustainable Grain Transportation in Ukraine Amidst War Utilizing KNARM and KnowWhereGraph*

Web4Good: Health



11:00 AM – 12:30 PM at AT&T Hotel and Conference Center Classroom #115

Web4Good

Wednesday, May 3rd

8420 Leveraging Existing Literature on the Web and Deep Neural Models to Construct a Knowledge Graph Focused on Water Quality and Health Risks

🕒 11:50 AM – 12:00 PM

Description

Authors: Nikita Gautam, David Shumway, Megan Kowalczyk, Sarthak Khanal, Doina Caragea, Cornelia Caragea, Hande McGinty and Samuel Dorevitch

Thank you



QUESTIONS?



HANDE@KSU.EDU

