9AM

2PM in London (GMT), 11PM in Tokyo (GMT+9)

Panel: Ontologies and AI

Moderator: Bruce W. Herr II, *Indiana University*

Presenters:

- Maria-Esther Vidal, *German National Library of Science and Technology (TIB), Germany* (Knowledge Graph-driven hybrid AI)
- Yongxin (Kiki) Kong, *Indiana University & Chinese Academy of Sciences, China* (HRAIit)
- Oliver He, *University of Michigan*

Maria-Esther Vidal,
Leibniz University of Hannover, TIB-Leibniz Institute of Science and Technology, University Library Hannover, Germany

Knowledge Graph-Driven AI

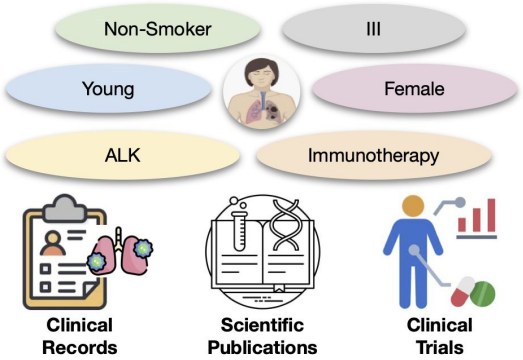
Maria-Esther Vidal

Professor Data Science Institute, Leibniz University of Hannover

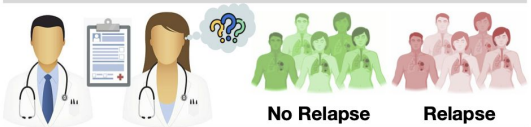
Head of the Scientific Data Management Group at TIB-Leibniz Information Center for Science and Technology, Germany

AI Models in Medicine- Scattered Data and Fragmented Knowledge Negative Impact

Multiple Data Sources and Fragmented Medical Knowledge

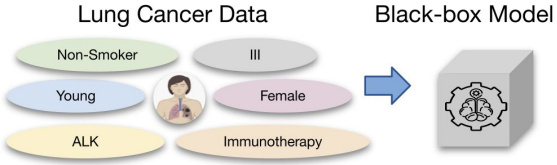


Clinical Objectives

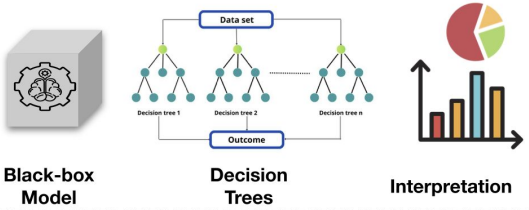


Clinical objectives:

- Maximize survival time, life quality
- Minimize toxicities, adverse events
- Avoid relapse and disease progression



Traditional Models

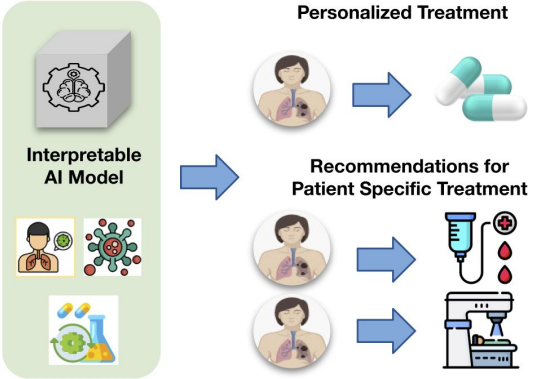


Decision-making

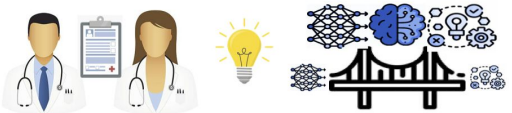
Decision-making based on:

- data (e.g., symptoms, patient history)
- similar treated patients

Interpretable Models



Hybrid-AI Systems

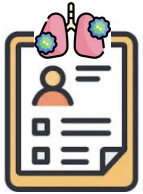


Bridging Cognitive Medical Thinking with AI

- Transparent Decision-Making
- Data Privacy and Sovereignty

Semantic Data Integration- Uniform View of Heterogeneous Data

Data Sources


Lung Cancer Registry


Breast Cancer Registry


Scientific Publications


Scientific Databases


Drug-drug interactions

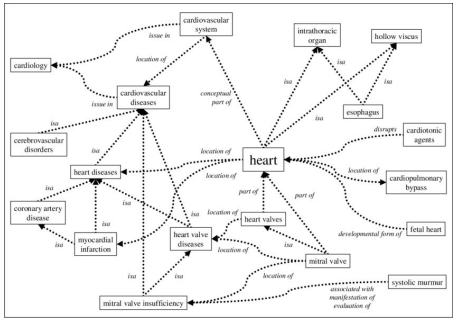

Genomics and Mutations


Medical Guidelines and Protocols

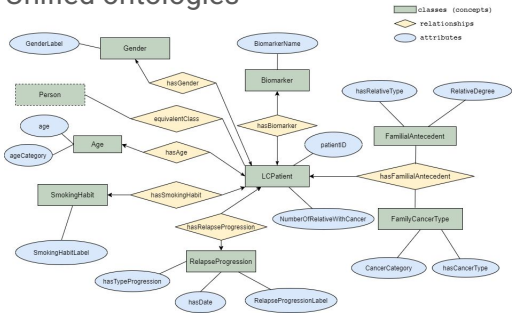

Clinical Trials

Biomedical Knowledge

Unified Medical Language System (UMLS)



Unified ontologies



Data Integration System [1]

A data integration system DIS=<O,S,M,∑>

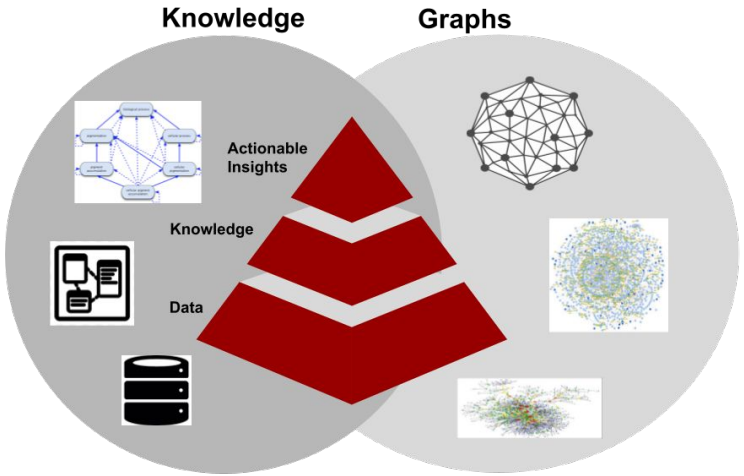
- O is an ontology or schema which provides a uniform view to the data sources in S.
- S is a set of {S1,..,Sn} of the signatures of the data sources that compose a DIS.
- M is a set of mappings between signatures of the sources in S and concepts in O.
- ∑=(φ, *S*, λ) is a shape schema over O; φ: set of shapes; *S*: set of shape labels; λ: *S*→φ *total function from labels to shapes.*

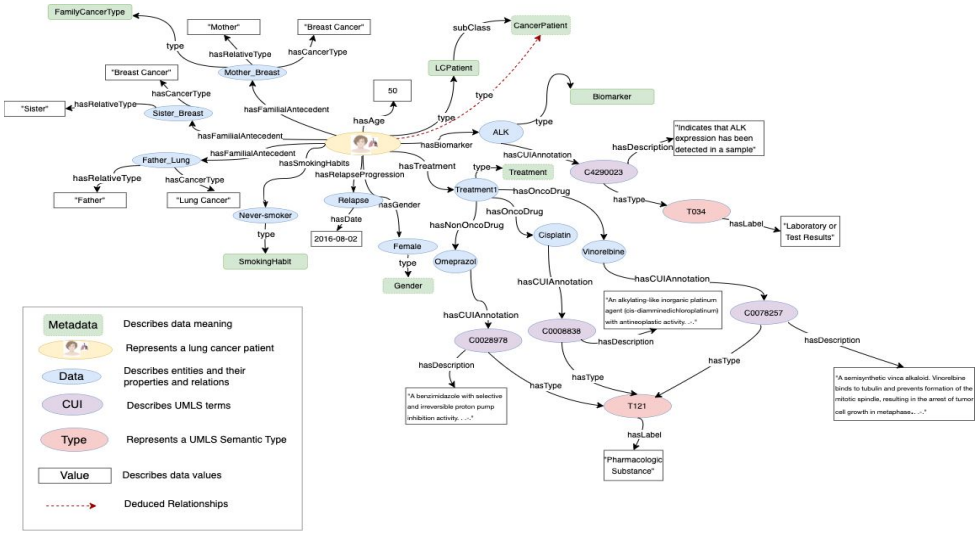
A data integration system

- provides a uniform view, and to integrate data collected from heterogeneous data sources

[1] M. Lenzerini. Managing Data through the Lens of an Ontology. AI Mag. 39(2): 65-74 (2018)

Knowledge-Graphs- structures to integrate heterogeneous data, capture domain knowledge, and enable explainable AI through symbolic reasoning





Knowledge Graphs

- data **structures** representing the **convergence** of **knowledge** and **data** as **factual** statements
- using a **graph data model**

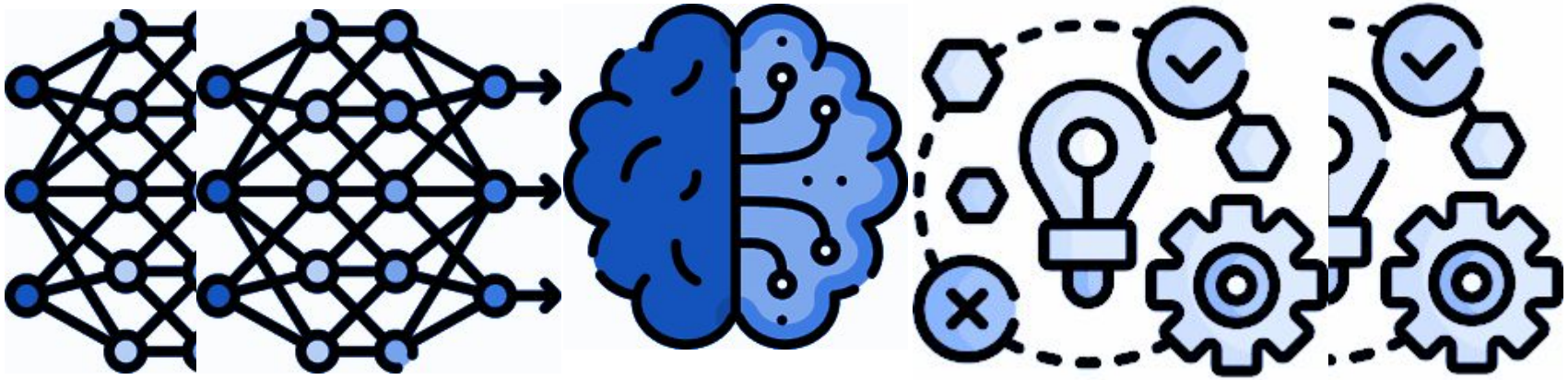
- Entities and their relationships are represented as first-class citizens.
- Metadata (via ontologies) describing and providing information about other data.
- Metadata and data can be empowered with inference to deduce new facts.

Integrating Semantics and Learning: The Role of Hybrid AI Systems

Neural Components

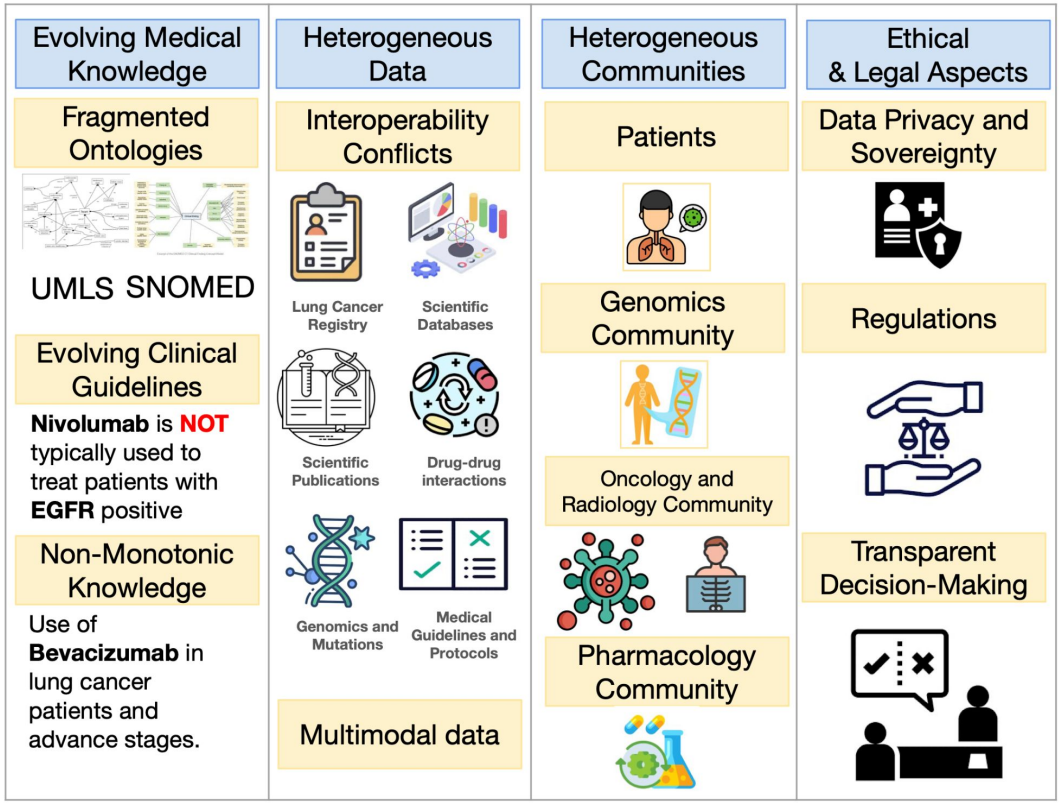
Hybrid AI Systems

Symbolic Components



Data-driven, black-box AI systems that approximate a given function (e.g., neural networks)	AI systems that can learn from data while also incorporating logical, interpretable reasoning.	Logic-based reasoning modules with explicit constraints or specifications.		

Challenges in Knowledge Graph-driven Hybrid AI Systems



Evolving Medical Knowledge:

- Fragmented ontologies like UMLS and SNOMED limit data consistency.
- Frequent updates to clinical guidelines, e.g., Nivolumab not for EGFR-positive cases.
- Non-monotonic knowledge, e.g., Bevacizumab use in advanced lung cancer.

Heterogeneous Data:

- Multimodal sources: genomic data, drug interactions, and medical guidelines.
- Interoperability issues across registries, databases, and publications.

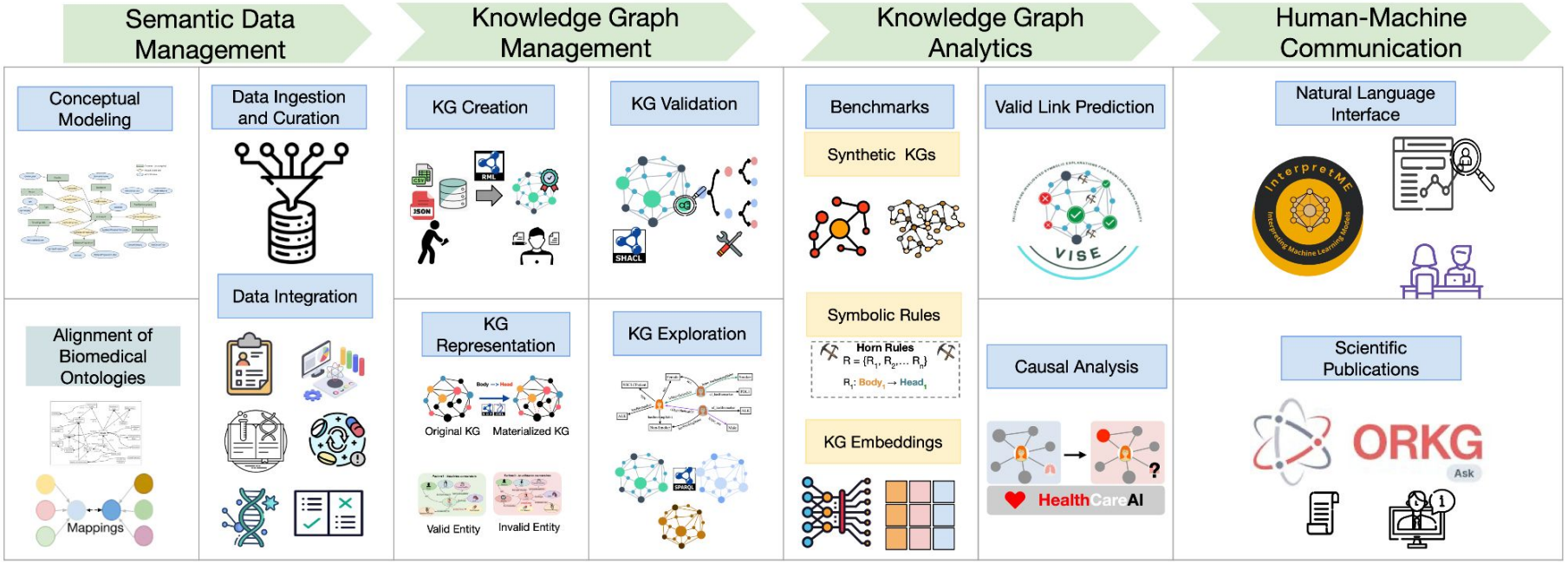
Heterogeneous Communities:

- Stakeholders include patients, genomics experts, oncologists, and pharmacologists.
- Conflicting perspectives complicate shared data modeling.

Ethical and Legal Aspects:

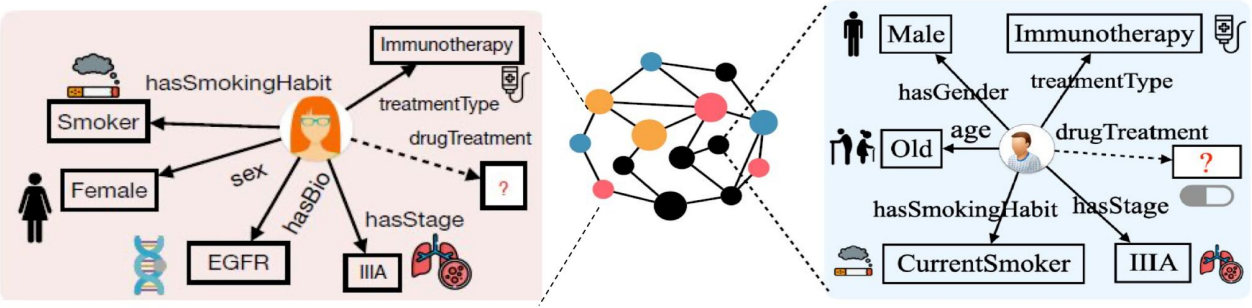
- Address data privacy, sovereignty, and regulations (e.g., GDPR, HIPAA).
- Require transparent frameworks for decision-making and accountability.

TrustKG-Hybrid AI framework to bridge symbolic reasoning and inductive learning and deliver interpretable and user-centric recommendations

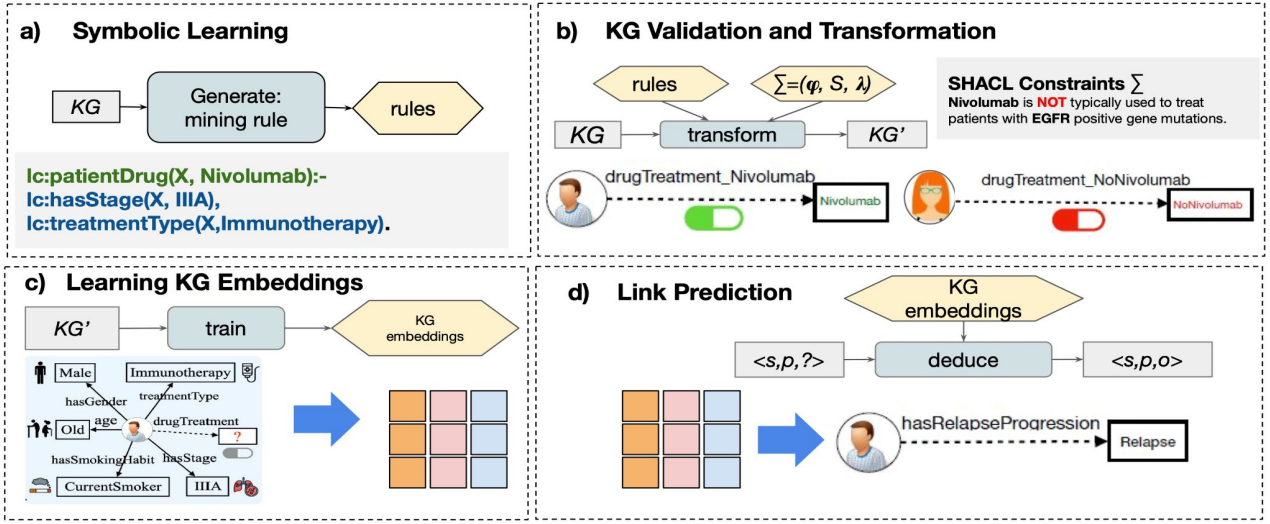


- Semantic Data Integration: Combines biomedical data using modeling, ontology alignment, and semantic reconciliation.
- Knowledge Graph Analytics: Enables link prediction and causal analysis for transparency and interpretability.
- User-Centric Tools: Tools like ORKG Ask provide actionable insights and foster trust.

VISE-Hybrid AI for Accurate and Interpretable Link Prediction in KGs



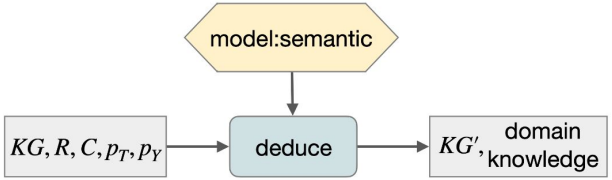
Approaches	Results for KG_3					
	Model	Hits@1	Hits@3	Hits@5	Hits@10	MRR
Baseline 1	TransE	0.000	0.560	0.795	0.943	0.324
	TransD	0.002	0.551	0.690	0.872	0.310
	TransH	0.622	0.864	0.943	0.983	0.756
	RotatE	**0.696**	**0.933**	**0.969**	**0.987**	**0.820**
Baseline 2	TransE	0.000	0.713	0.840	0.931	0.376
	TransD	0.008	0.694	0.824	0.935	0.379
	TransH	**0.882**	**0.969**	**0.997**	**1.000**	**0.929**
	RotatE	0.864	0.987	0.995	1.000	0.924
Baseline 3	TransE	0.000	0.519	0.747	0.923	0.310
	TransD	0.011	0.551	0.716	0.884	0.322
	TransH	0.596	0.876	0.925	0.977	0.740
	RotatE	**0.714**	**0.941**	**0.969**	**0.990**	**0.829**
Baseline 4	TransE	0.000	0.536	0.735	0.931	0.311
	TransD	0.002	0.551	0.733	0.870	0.318
	TransH	0.542	0.849	0.908	0.974	0.702
	RotatE	**0.700**	**0.945**	**0.972**	**0.992**	**0.818**
VISE	TransE	0.000	0.760	0.878	0.948	0.388
	TransD	0.013	0.684	0.762	0.884	0.368
	TransH	0.868	0.980	0.994	**1.000**	0.924
	RotatE	**0.887**	**0.986**	**0.996**	0.998	**0.936**



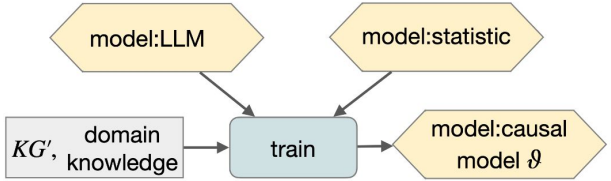
Existing KG embedding models: -) impacted by the representation of factual statements in KGs
-) enhanced by explicitly expressing valid and invalid links according domain-specific integrity constraints.

HealthCareAI-Hybrid AI for Accurate Counterfactual Predictions

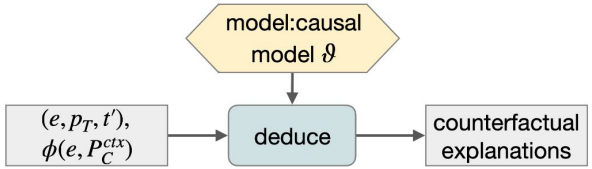
a) Symbolic Reasoning

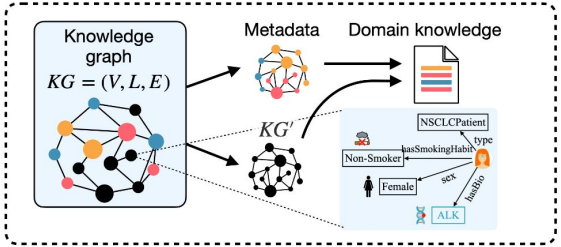


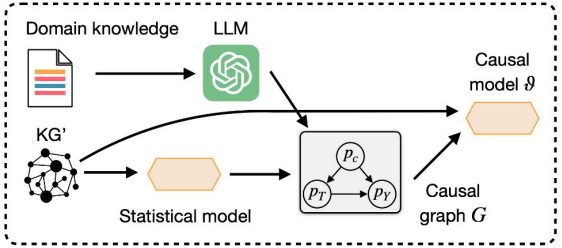
b) Causal Model Learning

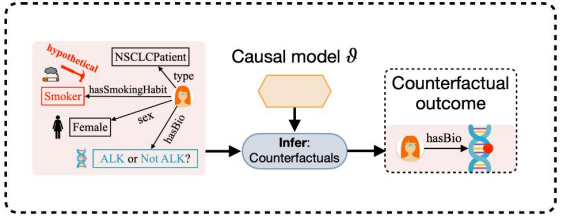


c) Counterfactual Prediction









Expert Causal Graph

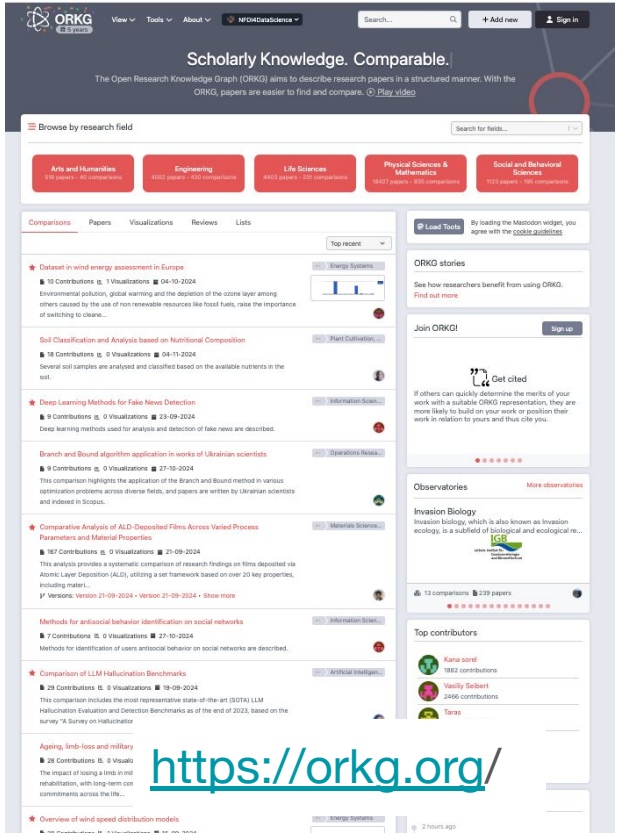
Approaches	Results (%) for KG_{real}				
	Model	Jaccard Index	Precision	Recall	F1-Score
Baseline 1	PC	13.3	40.0	16.7	23.5
	FCI	16.0	23.5	33.3	27.6
	GES	6.7	25.0	8.3	12.5
Baseline 2	GPT4 wo domain knowledge	0.0	0.0	0.0	0.0
Baseline 3	GPT4 w domain knowledge	23.5	**44.4**	33.3	38.1
HealthCareAI	Baseline 3 + PC	**30.0**	42.9	50.0	**46.2**
	Baseline 3 + FCI	29.6	34.8	**66.7**	45.7
	Baseline 3 + GES	25.0	38.5	41.7	40.0

Counterfactual Prediction

Approaches	Results (%) for KG_{10k}				
	Model	Jaccard Index	Precision	Recall	F1-Score
Baseline 1	PC	75.0	**100.0**	75.0	85.7
	FCI	75.0	**100.0**	75.0	85.7
	GES	62.5	**100.0**	62.5	76.9
Baseline 2	GPT4 wo domain knowledge	40.0	66.7	50.0	57.1
Baseline 3	GPT4 w domain knowledge	62.5	**100.0**	62.5	76.9
HealthCareAI	Baseline 2 + PC	70.0	77.8	87.5	82.4
	Baseline 3 + PC	**100.0**	**100.0**	**100.0**	**100.0**

-) Compared against statistical methods (PC, FIC, GES) and LLM-based approaches with/without domain knowledge.
-) Achieved higher accuracy by integrating semantics and domain-specific knowledge.
-) Semantic enrichment improved interpretability and aligned predictions with human reasoning.

Open Research Knowledge Graph (ORKG)- represents, organizes, and shares scholarly knowledge in a structured and machine-readable format



https://orkg.org/

Prof. Dr. Sören Auer

Purpose and Vision:

Aims to transform scientific knowledge into a connected, interactive, and navigable structure that enhances understanding and accessibility.

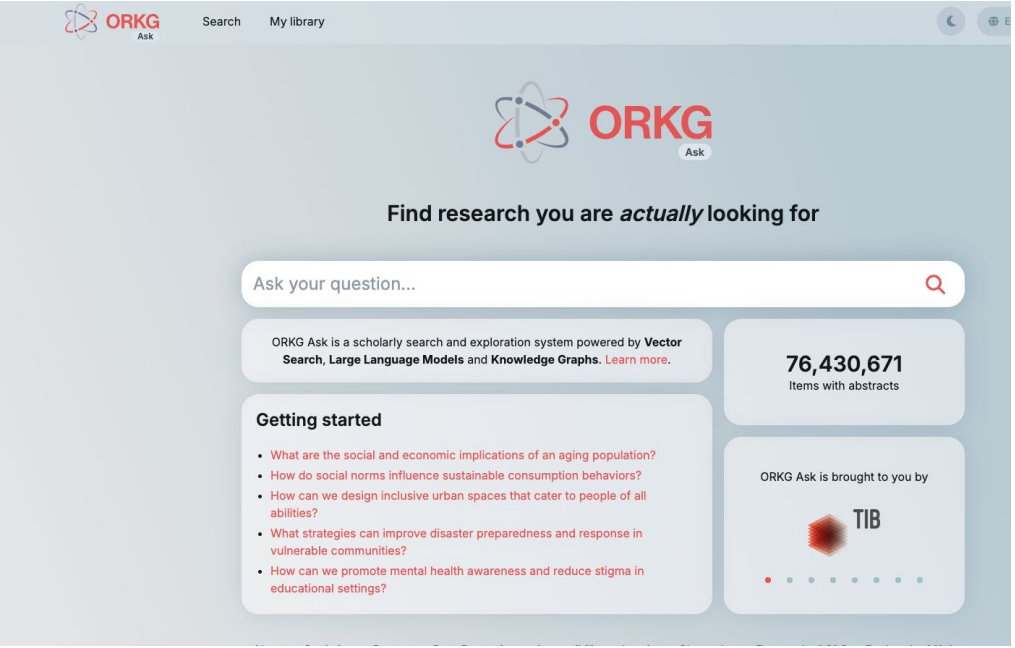
Key Features:

Structured Data Representation: Utilizes a knowledge graph structure to capture and link research outputs for enhanced discoverability.

Semantic Enrichment: Integrates related concepts and research findings to provide contextual and comprehensive insights.

Interactive Interface: Allows researchers to contribute and explore research data through an intuitive user experience.

ORKG Ask- Human-Centric Communication



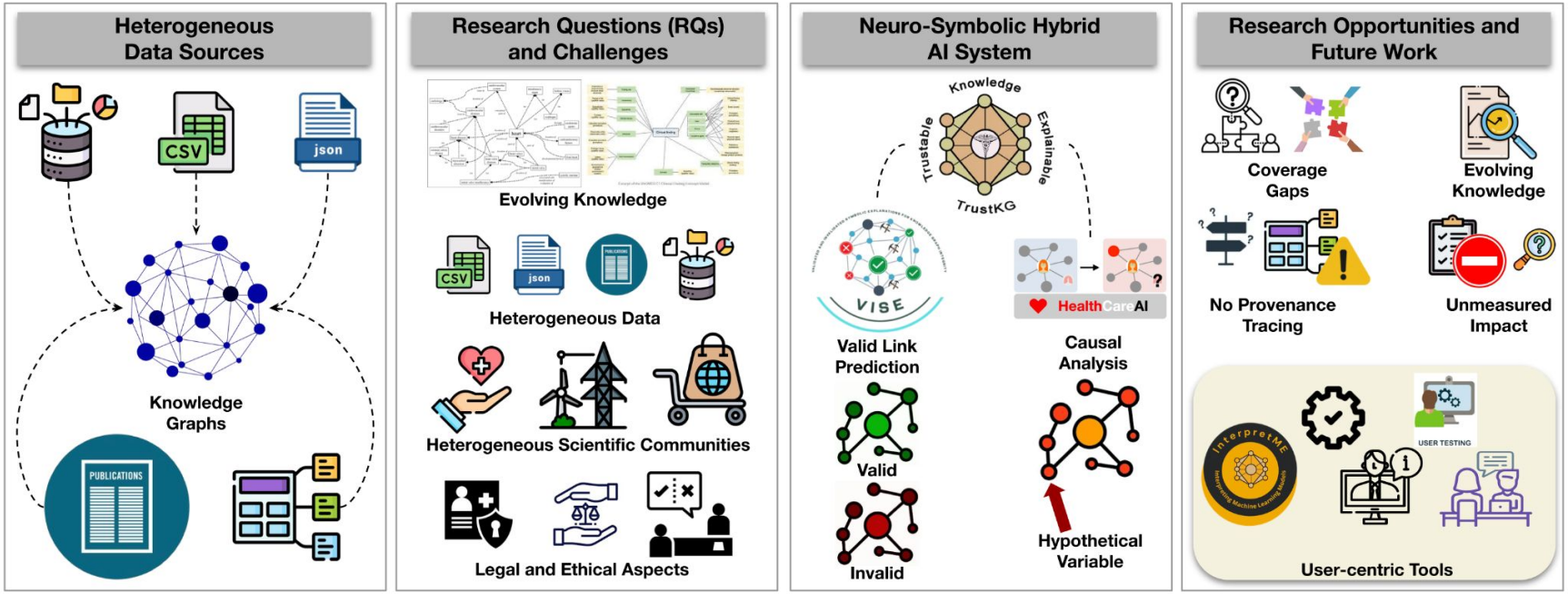
A scientific search and exploration system providing answers from a database of 80 million full-text scientific publications.

Combines the power of large language models with the structured data of ORKG to deliver fast, informed answers to complex research questions.

https://ask.orkg.org/

Example1 Example2

TrustKG- Knowledge Graph-driven AI system to enhance interpretability, transparency, and usability



Knowledge Graphs (KGs) present a promising avenue for achieving integration of heterogeneous data sources, thereby facilitating **interoperability** and data exploration.

The **research challenges** identified here are critical barriers to integrating KGs with hybrid AI systems.

TrustKG shows how hybrid AI systems like **VISE** and **HealthCareAI** can use KGs in different areas.

By overcoming these limitations and advancing their research, KGs can reach their full potential and use **advanced Hybrid AI** in real-world applications.

Semantic Web 14 (2023) 943–976
DOI 10.3233/SW-223294
IOS Press

Knowledge graphs for enhancing transparency in health data ecosystems[1]

Fotis Aisopos [a], Samaneh Jozashoori [b], Emetis Niazmand [b], Disha Purohit [b], Ariam Rivas [b], Ahmad Sakor [b], Enrique Iglesias [b], Dimitrios Vogiatzis [a*], Ernestina Menasalvas [d], Alejandro Rodríguez Gonzalez [e], Guillermo Vigueras [e], Daniel Gomez-Bravo [e], Maria Torrente [e], Roberto Hernández López [e], Mariano Provencio Pulla [e], Athanasios Dalianis [f], Anna Triantafillou [f], Georgios Paliouras [a] and Maria-Esther Vidal [b,*]

[a] Institute of Informatics & Telecommunications, National Center for Scientific Research "Demokritos", Greece
E-mails: fotis.aisopos@iit.demokritos.gr, dimitr@iit.demokritos.gr, paliourg@iit.demokritos.gr
[b] Leibniz University of Hannover and L3S Research Center and TIB Leibniz Information Centre for Science and Technology, Germany
E-mails: samaneh.jozashoori@gmail.com, emetis.niazmand@tib.eu, disha.purohit@tib.eu, ariam.rivas@tib.eu, ahmad.sakor@tib.eu, iglesias@l3s.de, maria.vidal@tib.eu
[c] American College of Greece, Deree, Greece
E-mail: dimitrv@acg.edu
[d] Universidad Politécnica de Madrid, Spain
E-mails: ernestina.menasalvas@upm.es, alejandro.rg@upm.es, guillermo.vigueras@upm.es, daniel.gomez-bravo@upm.es
[e] Medical Oncology Department, Puerta de Hierro University Hospital, Servicio Madrileño de Salud, Spain
E-mails: mtorrente008@gmail.com, robertohlopez78@gmail.com, mprovencio@gmail.com
[f] Innovation Lab, Athens Technology Center, Greece
E-mails: T.Dalianis@atc.gr, a.triantafillou@atc.gr

Editors: Haridimos Kondylakis, FORTH-ICS, Greece; Praveen Rao, University of Missouri, USA; Kostas Stefanidis, Tampere University, Finland
Solicited reviews: Sara Colantonio, National Research Council, Italy; Stelios Sfakianakis, Foundation for Research and Technology-Hellas, Greece; one anonymous reviewer

https://content.iospress.com/articles/semantic-web/sw223294

Abstract. Tailoring personalized treatments demands the analysis of a patient's characteristics, which may be scattered over a wide variety of sources. These features include family history, life habits, comorbilities, and potential treatment side effects. Moreover, the analysis of the services visited the most by a patient before a new diagnosis, as well as the type of requested tests, may uncover patterns that contribute to earlier disease detection and treatment effectiveness. Built on knowledge-driven ecosystems, we devise DE4LungCancer, a health data ecosystem of data sources for lung cancer. In this data ecosystem, knowledge extracted from heterogeneous sources, e.g., clinical records, scientific publications, and pharmacological data, is integrated into knowledge graphs. Ontologies describe the meaning of the combined data, and mapping rules enable the declarative definition of the transformation and integration processes. DE4LungCancer is assessed regarding the methods followed for data quality assessment and curation. Lastly, the role of controlled vocabularies and ontologies in health data management is discussed, as well as their impact on transparent knowledge extraction and analytics. This paper presents the lessons learned in the DE4LungCancer development. It demonstrates the transparency level supported by the proposed knowledge-driven ecosystem, in the context of

[1] Aisopos, Jozashoori, Niazmand, Purohit, Rivas, Sakor, and Vidal contributed equally to the work reported in this paper.
[*] Corresponding author. E-mail: maria.vidal@tib.eu.

Cancers (Basel). 2022 Aug 22;14(16):4041. doi: 10.3390/cancers14164041.

An Artificial Intelligence-Based Tool for Data Analysis and Prognosis in Cancer Patients: Results from the Clarify Study

María Torrente [1 2], Pedro A Sousa [3], Roberto Hernández [1], Mariola Blanco [1], Virginia Calvo [1], Ana Collazo [1], Gracinda R Guerreiro [4], Beatriz Núñez [1], Joao Pimentao [3], Juan Cristóbal Sánchez [1], Manuel Campos [5 6], Luca Costabello [7], Vit Novacek [8], Ernestina Menasalvas [9], María Esther Vidal [10], Mariano Provencio [1]

Affiliations + expand

PMID: 36011034 PMCID: PMC9406336 DOI: 10.3390/cancers14164041

https://pubmed.ncbi.nlm.nih.gov/36011034/

Abstract

Background: Artificial intelligence (AI) has contributed substantially in recent years to the resolution of different biomedical problems, including cancer. However, AI tools with significant and widespread impact in oncology research remain scarce. The goal of this study is to present an AI-based solution tool for cancer patients data analysis that assists clinicians in identifying the clinical factors associated with poor prognosis, relapse and survival, and to develop a prognostic model that stratifies patients by risk.

Materials and methods: We used clinical data from 5275 patients diagnosed with non-small cell lung cancer, breast cancer, and non-Hodgkin lymphoma at Hospital Universitario Puerta de Hierro-Majadahonda. Accessible clinical parameters measured with a wearable device and quality of life questionnaires data were also collected.

Results: Using an AI-tool, data from 5275 cancer patients were analyzed, integrating clinical data, questionnaires data, and data collected from wearable devices. Descriptive analyses were performed in order to explore the patients' characteristics, survival probabilities were calculated, and a prognostic model identified low and high-risk profile patients.

Conclusion: Overall, the reconstruction of the population's risk profile for the cancer-specific predictive model was achieved and proved useful in clinical practice using artificial intelligence. It has potential application in clinical settings to improve risk stratification, early detection, and surveillance management of cancer patients.

Keywords: artificial intelligence; cancer patients; data integration; decision support system; patient stratification; precision oncology.

PubMed Disclaimer

Semantically Describing Predictive Models for Interpretable Insights into Lung Cancer Relapse

Authors: Yashrajsinh Chudasama, Disha Purohit, Philipp D. Rohde, Enrique Iglesias, Maria Torrente, Maria-Esther Vidal
Pages: 142 - 158
DOI: 10.3233/SSW240012
Category: Research Article
Series: Studies on the Semantic Web
Ebook: Volume 60: Knowledge Graphs in the Age of Language Models and Vision Transformers

https://ebooks.iospress.nl/doi/10.3233/SSW240012

Abstract
Machine learning (ML) is becoming increasingly important in healthcare decision-making, requiring highly interpretable insights from predictive models. Although integrating ML models with knowledge graphs (KGs) holds promise, conveying model outcomes to domain experts remains challenging, hindering usability despite accuracy. We propose semantically describing predictive model insights to overcome communication barriers. Our pipeline predicts lung cancer relapse likelihood, providing oncologists with patient-centric explanations based on input characteristics. Consequently, domain experts gain insights into both the characteristics of classified lung cancer patients and their relevant population. These insights, along with model decisions, are semantically described in natural language to enhance understanding, particularly for interpretable models like LIME and SHAP. Our approach, SemDesLC, documents ML model pipelines into KGs, and fulfills the needs of three types of users: KG builders, analysts, and consumers. Experts' opinions indicate that semantic descriptions are effective for elucidating relapse determinants. SemDesLC is openly accessible on GitHub, promoting transparency and collaboration in leveraging ML for healthcare decision support.

ORKG ASK: a Neuro-symbolic Scholarly Search and Exploration System

Allard Oelen[1], Mohamad Yaser Jaradeh[2] and Sören Auer[1,2]

[1] TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
[2] L3S Research Center, Leibniz University of Hannover, Hannover, Germany

Abstract
Purpose: Finding scholarly articles is a time-consuming and cumbersome activity, yet crucial for conducting science. Due to the growing number of scholarly articles, new scholarly search systems are needed to effectively assist researchers in finding relevant literature.
Methodology: We take a neuro-symbolic approach to scholarly search and exploration by leveraging state-of-the-art components, including semantic search, Large Language Models (LLMs), and Knowledge Graphs (KGs). The semantic search component composes a set of relevant articles. From this set of articles, information is extracted and presented to the user.
Findings: The presented system, called ORKG ASK (Assistant for Scientific Knowledge), provides a production-ready search and exploration system. Our preliminary evaluation indicates that our proposed approach is indeed suitable for the task of scholarly information retrieval.
Value: With ORKG ASK, we present a next-generation scholarly search and exploration system and make it available online. Additionally, the system components are open source with a permissive license.

Keywords
Neuro-symbolic AI, Large Language Models, Scholarly Knowledge Graphs, Scholarly Search System

https://ceur-ws.org/Vol-3759/paper7.pdf

1. Introduction

Finding scholarly articles and exploring the body of scholarly literature consumes a significant share of a researcher's time. Due to the growing number of scholarly articles, this issue only becomes more apparent [1]. Current scholarly search systems passively assist users with their information needs by providing a list of relevant articles. If instead active assistance were provided, the users' information needs, such as a research question, would be answered for them. We present ORKG ASK (Assistant for Scientific Knowledge), a new generation scholarly search and exploration system[1]. ORKG ASK helps researchers find relevant literature and automatically extract knowledge from the retrieved literature, actively supporting researchers with their information needs. The approach consists of three main components: 1) Semantic Search, 2) a Large Language Model (LLM), and 3) Knowledge Graphs (KGs). First, the semantic search addresses the previously discussed challenge of retrieving articles based on their relevance to a specific information need. In ORKG ASK users can formulate their information need as a

SEMANTICS 2024 EU: 20th International Conference on Semantic Systems, September 17-19, 2024, Amsterdam, The Netherlands
allard.oelen@tib.eu (A. Oelen); jaradeh@l3s.de (M.Y. Jaradeh); auer@tib.eu (S. Auer)
0000-0001-9924-9153 (A. Oelen); 0000-0001-8777-2780 (M.Y. Jaradeh); 0000-0002-0698-2864 (S. Auer)

[1] Available online via https://ask.orkg.org

Lung Cancer
Volume 195, September 2024, 107920
ELSEVIER

https://www.sciencedirect.com/science/article/pii/S0169500224004549

Review Article

Family history of cancer and lung cancer: Utility of big data and artificial intelligence for exploring the role of genetic risk

Virginia Calvo [a], Emetis Niazmand [b c], Enric Carcereny [d], Delvys Rodriguez-Abreu [e], Manuel Cobo [f], Rafael López-Castro [g], María Guirado [h], Carlos Camps [i], Ana Laura Ortega [j], Reyes Bernabé [k], Bartomeu Massutí [l], Rosario Garcia-Campelo [m], Edel del Barco [n], José Luis González-Larriba [o], Joaquim Bosch-Barrera [p], Marta Martínez [q], María Torrente [q], María-Esther Vidal [b c], Mariano Provencio [a]

Show more

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 March 2024, accepted 24 April 2024, date of publication 29 April 2024, date of current version 7 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3395134

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10510291

RESEARCH ARTICLE

CauseKG: A Framework Enhancing Causal Inference With Implicit Knowledge Deduced From Knowledge Graphs

HAO HUANG [1] AND MARIA-ESTHER VIDAL [1,2]
[1] TIB–Leibniz Information Centre for Science and Technology, Leibniz University of Hannover, 30167 Hannover, Germany
[2] L3S Research Centre, 30167 Hannover, Germany
Corresponding author: Hao Huang (hao.huang@tib.eu)

This work was supported by the Project TrustKG–Transforming Data into Trustable Insights, through the program "Leibniz Best Minds: Program for Women Professors," funded by the Leibniz Association, under Grant P59/2020. The work of Hao Huang was supported by the TrustKG Project. The work of Maria-Esther Vidal was supported in part by the program "Leibniz Best Minds: Program for Women Professors."

ABSTRACT Causal inference is a critical technique for inferring causal relationships from data and distinguishing causation from correlation. Causal inference frameworks rely on structured data, typically represented in flat tables or relational models. These frameworks estimate causal effects based only on explicit facts, overlooking implicit information in the data, which can lead to inaccurate causal estimates. Knowledge graphs (KGs) inherently capture implicit information through *logical rules* applied to explicit facts, providing a unique opportunity to leverage implicit knowledge. However, existing frameworks are not applicable to KGs due to their semi-structured nature. CauseKG is a causal inference framework designed to address the intricacies of KGs and seamlessly integrate implicit information using KG-specific entailment techniques, providing a more accurate causal inference process. We empirically evaluate the effectiveness of CauseKG against benchmarks constructed from synthetic and real-world datasets. The results suggest that CauseKG can produce a lower mean absolute error in causal inference compared to state-of-the-art methods. The empirical results demonstrate CauseKG's ability to address causal questions in a variety of domains. This research highlights the importance of extending causal inference techniques to KGs, emphasising the improved accuracy that can be achieved by integrating implicit and explicit information.

INDEX TERMS Causal inference, knowledge graphs, knowledge reasoning, semantics.

VISE: Validated and Invalidated Symbolic Explanations for Knowledge Graph Integrity

Disha Purohit[1,2,*,†], Yashrajsinh Chudasama[1,2,†], Maria Torrente[4] and Maria-Esther Vidal[1,2,3]

[1] TIB-Leibniz Information Centre for Science and Technology, Hannover, Germany
[2] Leibniz University Hannover, Germany
[3] L3S Research Center, Hannover, Germany
[4] Hospital Universitario Puertade Hierro-Majadahonda, Spain

https://ceur-ws.org/Vol-3831/paper5.pdf

Abstract
Knowledge graphs (KGs) are naturally capable of capturing the convergence of data and knowledge, thereby making them highly expressive frameworks for describing and integrating heterogeneous data in a coherent and interconnected manner. However, based on the Open World Assumption (OWA), the absence of information within KGs does not indicate falsity or non-existence; it merely reflects incompleteness. The process of inductive learning over KGs involves predicting new relationships based on existing factual statements in the KG, utilizing either numerical or symbolic learning models. Recently, Knowledge Graph Embedding (KGE) and symbolic learning have received considerable attention in various downstream tasks, including Link Prediction (LP). LP techniques employ latent vector representations of entities and their relationships in KGs to infer missing links. Furthermore, as the quantity of data generated by KGs continues to increase, the necessity for additional quality assessment and validation efforts becomes more apparent. Nevertheless, state-of-the-art KG completion approaches fail to consider the quality constraints while generating predictions, resulting in the completion of KGs with erroneous relationships. The generation of accurate data and insights is of vital importance in the context of healthcare decision-making, including the processes of diagnosis, the formulation of treatment strategies, and the implementation of preventive actions. We propose a hybrid approach, *VISE*, which adopts the integration of symbolic learning, constraint validation, and numerical learning techniques. *VISE* leverages KGE to capture implicit knowledge and represent negation in KGs, thereby enhancing the predictive performance of numerical models. Our experimental results demonstrate the effectiveness of this hybrid strategy, which combines the strengths of symbolic, numerical, and constraint validation paradigms. *VISE* implementation is publicly accessible on GitHub (https://github.com/SDM-TIB/VISE).

Keywords
Knowledge Graphs, Symbolic Learning, SHACL Constraints, Numerical Learning, Explainability

SemMatch: Semantics-Aware Matching for Causal Inference over Knowledge Graphs

https://link.springer.com/chapter/10.1007/978-981-96-0567-5_33

Hao Huang and Maria-Esther Vidal[1,2,3(✉)]

[1] TIB - Leibniz Information Centre for Science and Technology, Hannover, Germany
hao.huang@tib.eu, maria.vidal@tib.eu
[2] Leibniz University of Hannover, Hannover, Germany
[3] L3S Research Center, Hannover, Germany

Abstract. Causal inference is used in various domains such as healthcare, economics, and political science to infer causal effects from observational data where each unit (entity) has different properties. Existing approaches often assume data completeness, and thus exclude all units with incomplete data when performing causal inference, which can lead to inaccurate causal estimates. In addition, existing approaches follow the Close World Assumption, where facts not present in the database are assumed to be false, limiting the ability to reason under data incompleteness assumption. Knowledge graphs (KGs) are data structures that represent data in semi-structured formats and model the meaning of data via ontologies. We propose a method, *SemMatch*, based on KGs to enhance causal inference under a data incompleteness assumption. *SemMatch* relies on a semantic reasoning process specified by a set of logical rules over KGs, to infer implicit facts and partially address data incompleteness. Then, *SemMatch* applies machine learning methods to estimate the importance of properties. Finally, *SemMatch* employs causal estimation methods that consider property importance, facilitating causal reasoning across units with incomplete data to determine the causal effect. We evaluate *SemMatch* on synthetic datasets, and demonstrate that it achieves a lower mean absolute error (MAE) and square root of precision in estimation of heterogeneous effect (PEHE) in causal effect estimation compared to existing state-of-the-art methods. Observed results suggest that accounting for semantic reasoning and including units with incomplete data improves causal estimation accuracy.

Knowledge

Trustable

Explainable

TrustKG

Leibniz Programme for Women Professors

Leibniz Association

Disha Purohit

Yashrajsinh Chudasama

Hao Huang

Yongxin (Kiki) Kong, *Indiana University & Chinese Academy of Sciences (HRAlit)*

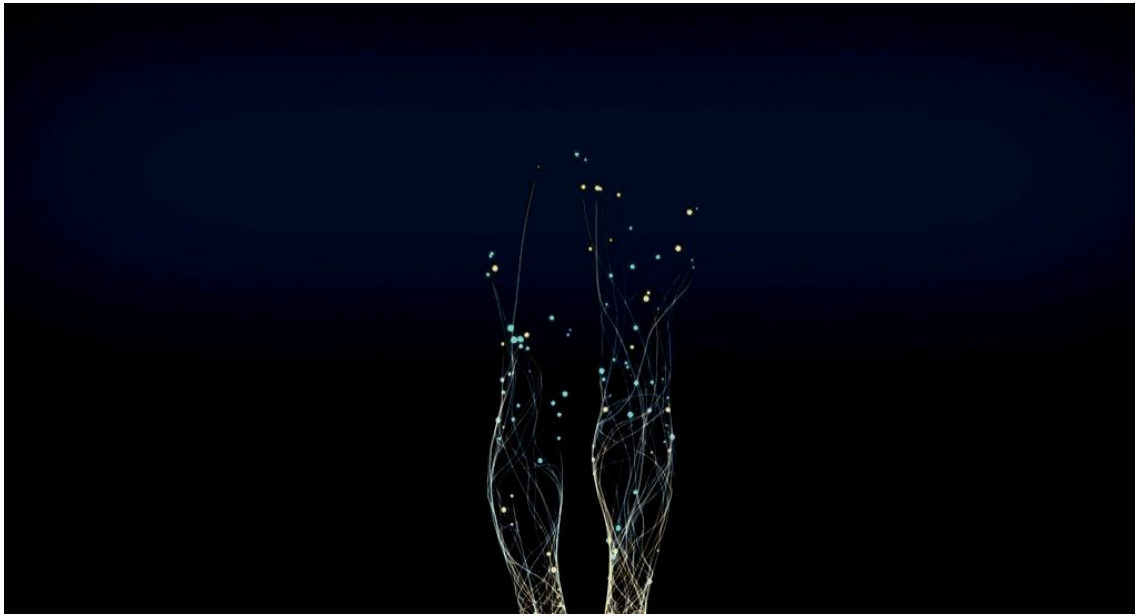
HRAlit: Publication, funding, and experimental data in support of Human Reference Atlas construction and usage

Yongxin (Kiki) Kong

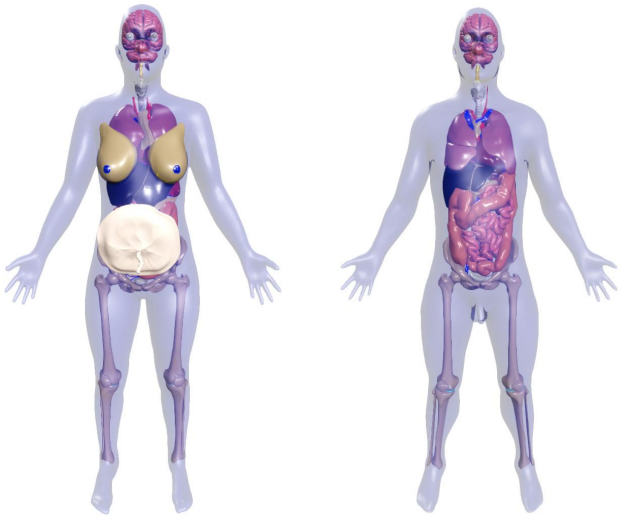
Postdoc, Chinese Academy of Science

Visiting Ph.D., Indiana University

The Human Reference Atlas (HRA) effort aims to map the human body at single cell resolution.

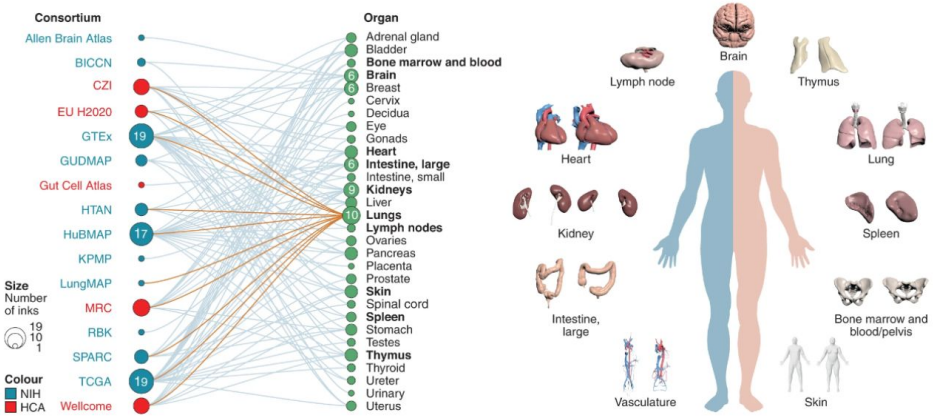


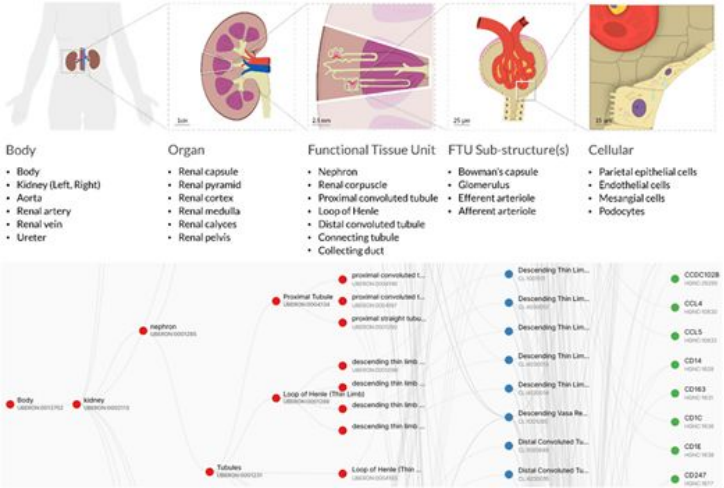
https://humanatlas.io



https://3d.nih.gov/collections/hra

HRA grows in the number of organs and data types





17
consortia

250+
experts

1,000+
publications

65
organs

4,659
anatomical structures

1,311
cell types

2,024
biomarkers

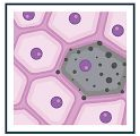
21
organ mapping antibody panels

22
functional tissue units

Many high-quality experimental datasets are becoming available

IU LUDDY

Human BioMolecular Atlas Program



Cellular Senescence Network

Genotype-Tissue Expression

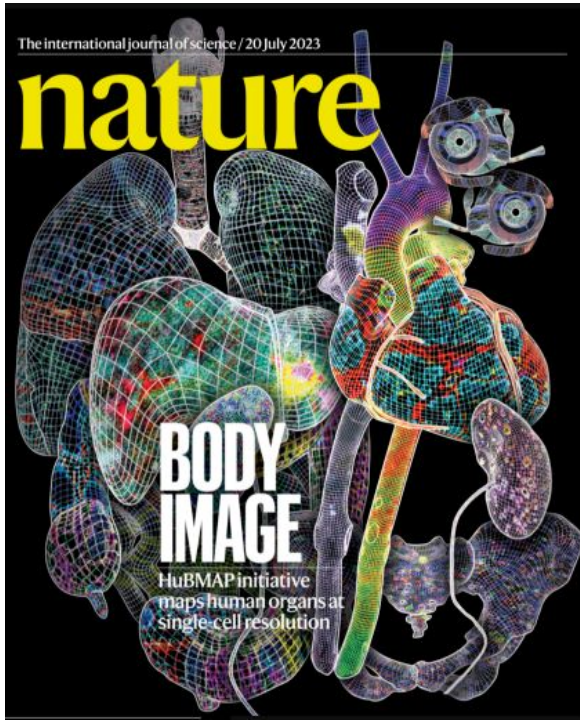
Kidney Precision Medicine Project

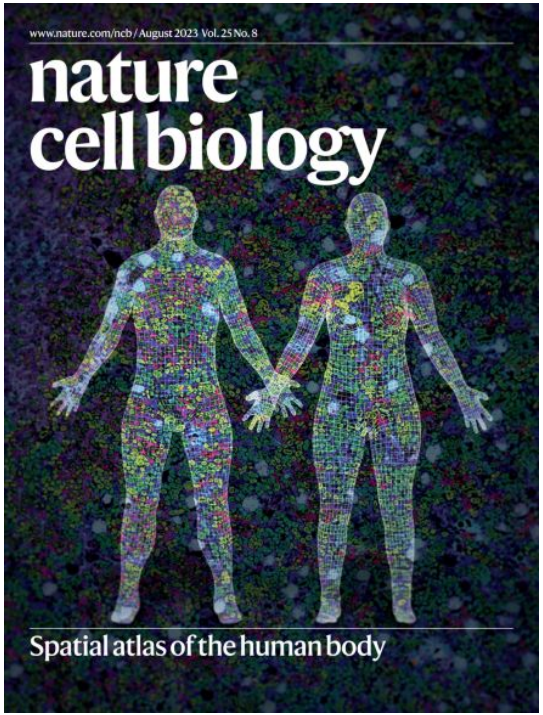
GenitoUrinary Developmental Molecular Anatomy Project

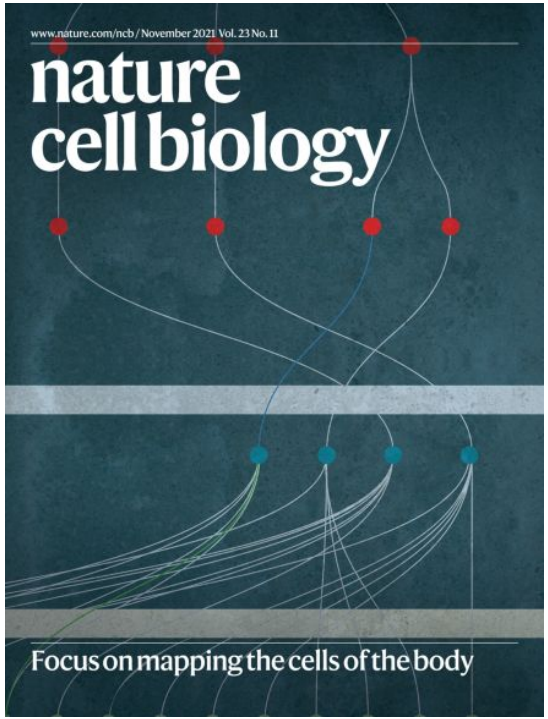
CZ CELLxGENE

Searching for data across portals is difficult

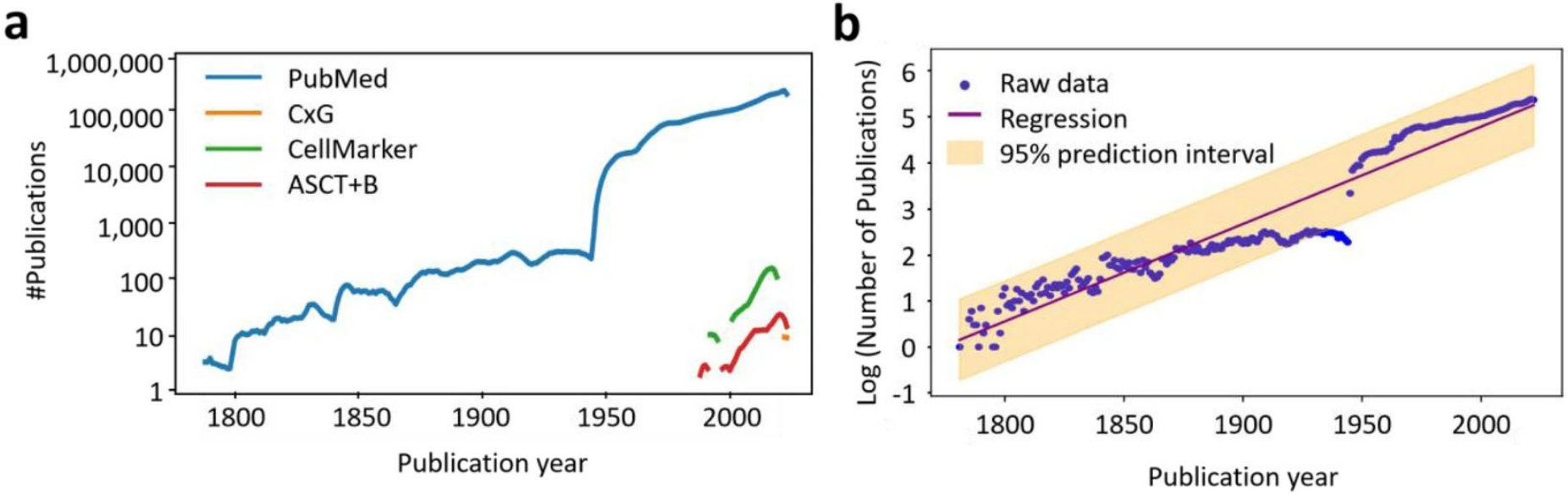
HRA relevant data is published in scholarly papers







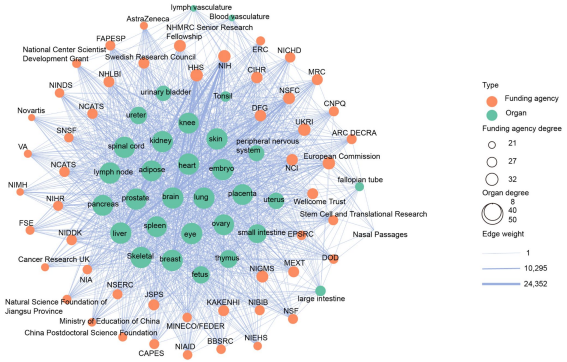
The number of publications increases exponentially



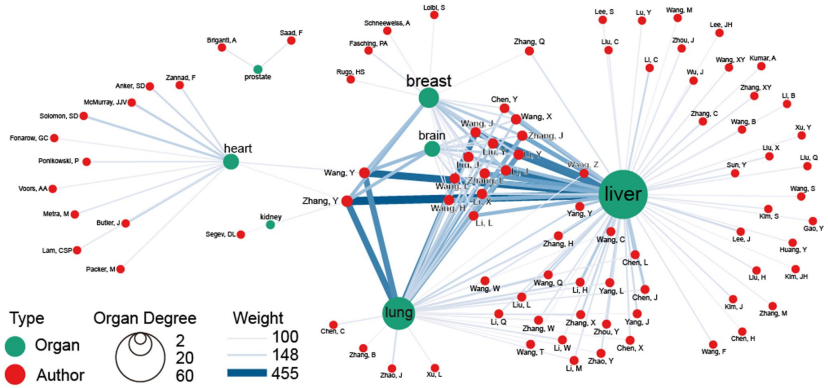
Kong Y, Börner K. *Scientific Data*, 2024

Scholarly publications evidence for HRA

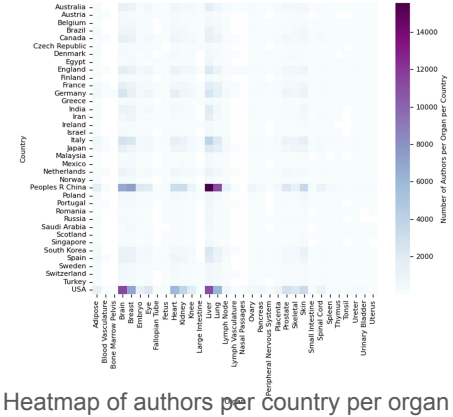
Geospatial layout of the coauthor network



Bimodal network of 32 organs and top 50 funding agencies most often listed



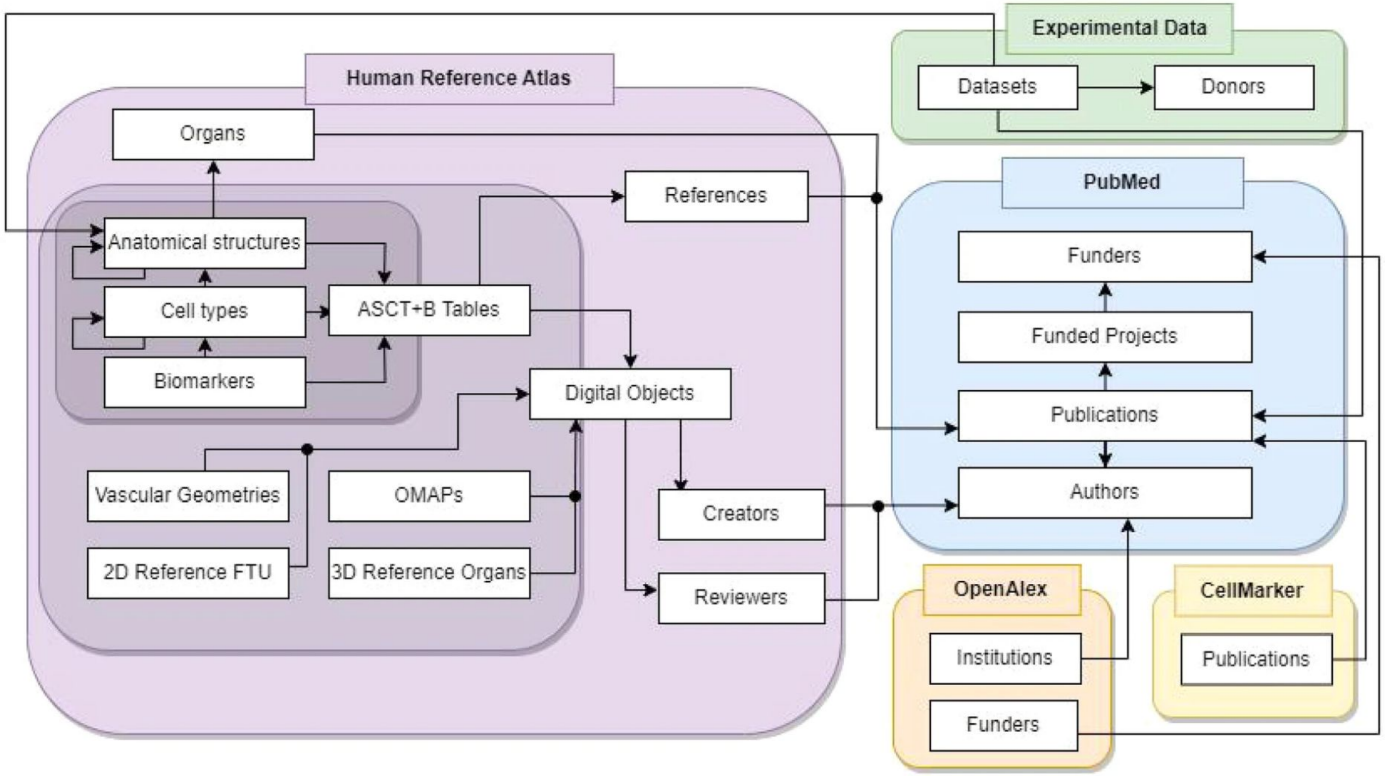
Bimodal network of highly cited authors and the organs they study.



Heatmap of authors per country per organ

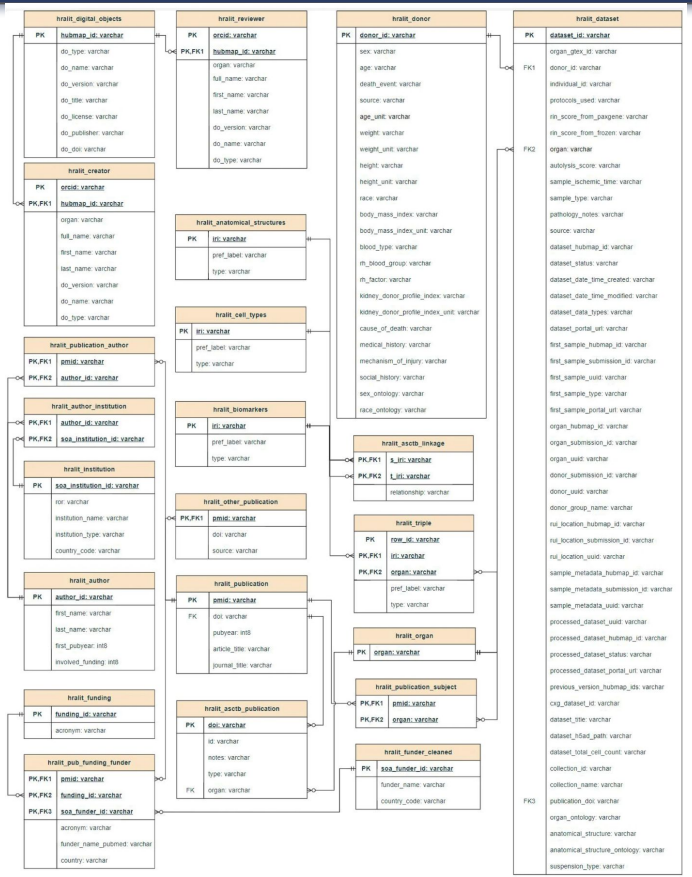
Kong Y, Daiya V A, Börner K. *Quantitative Science Studies*, 2024

Overview of the HRAlit database



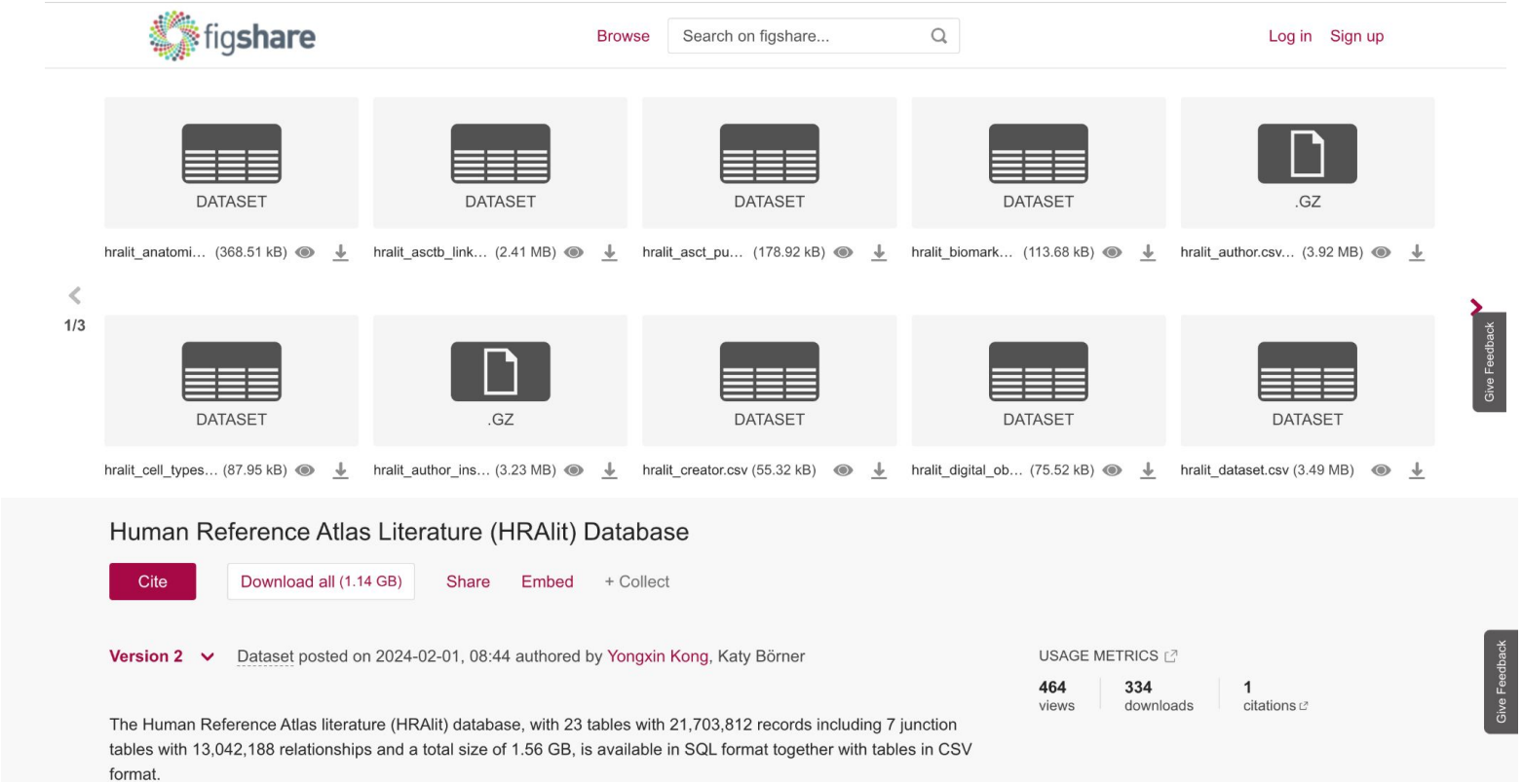
Kong Y, Börner K. *Scientific Data*, 2024

Entity relationship diagram of the HRAlit database



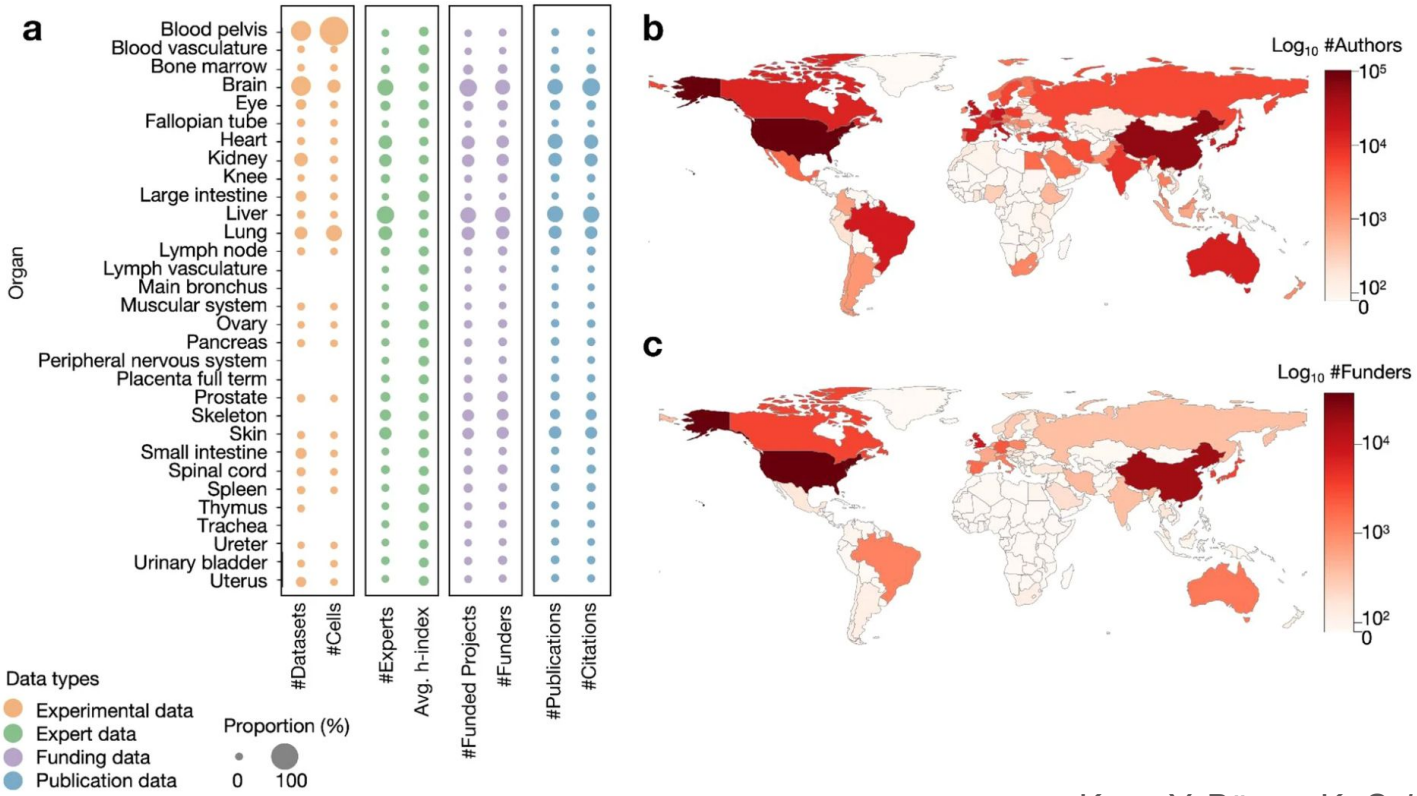
Kong Y, Börner K. *Scientific Data*, 2024

The HRAlit database SQL file and all tables in CSV format are at Figshare



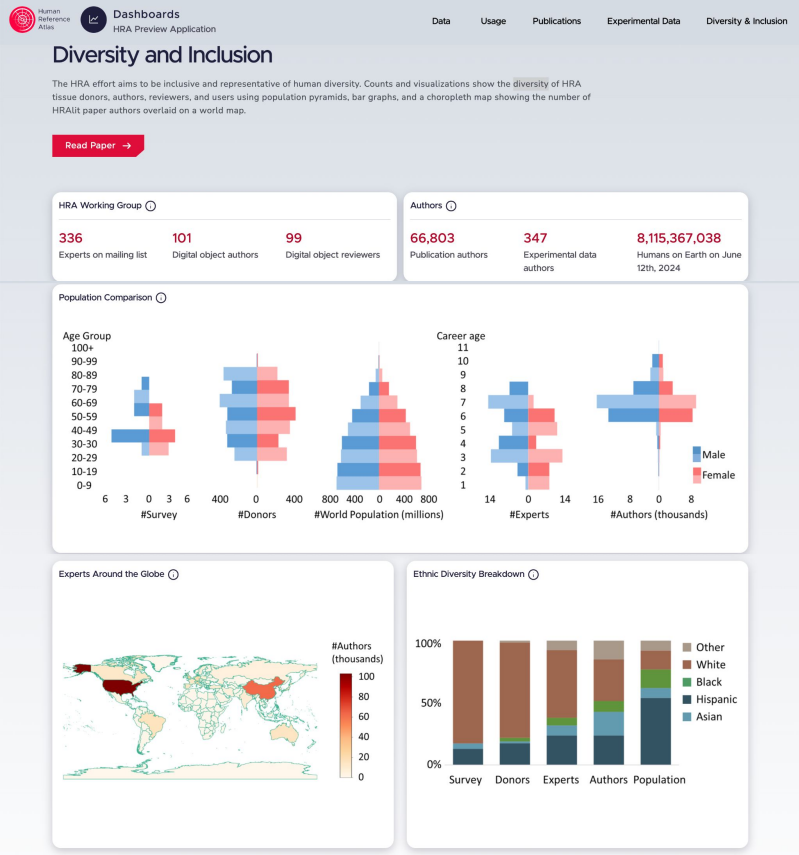
https://doi.org/10.6084/m9.figshare.24580669.v2

Distribution and choropleth map from HRAlit



Kong Y, Börner K. *Scientific Data*, 2024

HRA diversity and inclusion



https://apps.humanatlas.io/dashboard/diversity-and-inclusion

Acknowledgements

Principal Investigator:

Katy Börner

Lab:

CNS

Funders:

HuBMAP Human BioMolecular Atlas Program

SenNet

NIH NIDDK

KPMP

CIFAR

Oliver He, *University of Michigan*

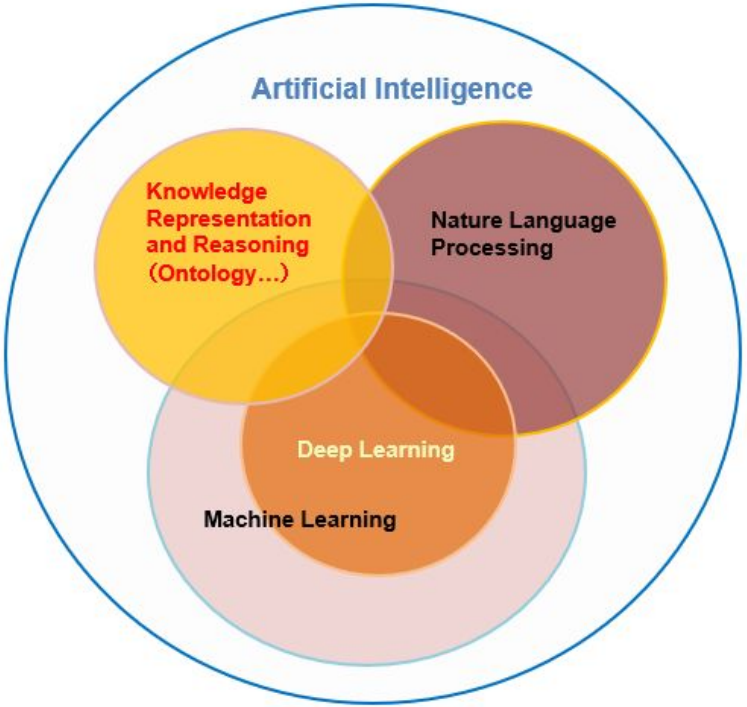
Ontology: Foundation of Precision Health Data Standardization & Artificial Intelligence

Yongqun "Oliver" He

University of Michigan Medical School
Ann Arbor, MI, USA.

3 Complementary AI Fields: KRR, ML, NLP

(Ontology is a major part of KRR)



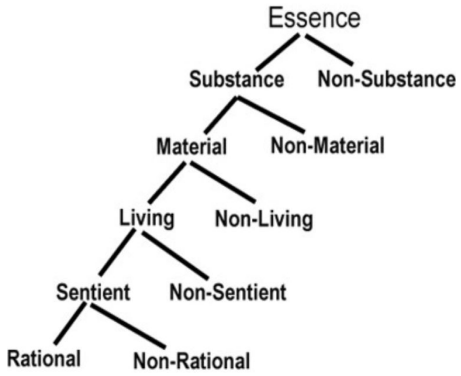
- **KRR**: Knowledge Representation & Reasoning
  - **Ontology** is a major approach in KRR
- **ML**: Machine Learning
- **NLP**: Natural Language Processing
  - Large Language Models (**LLM**): emerging NLP

Ontology: Originated from Philosophy and Taxonomy

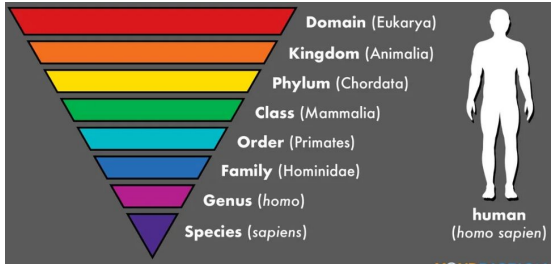
Onto: Being, existence, reality

Definition in philosophy: Ontology is the philosophical study of the nature of being, becoming, existence and/or reality, as well as the basic categories of being and their relations.

- It was called “first philosophy” by Aristotle (384–322 BC) in Book IV of his Metaphysics.
- The term “ontology” (or ontologia, “science of being”) was coined in 1613, independently, by two philosophers Rudolf Göckel (Goclenius) in his *Lexicon Philosophicum* and Jacob Lorhard (Lorhardus) in his *Theatrum philosophicum*.
- Stages of “ontology” development as science:
  - Porphyrian tree or Tree of Porphyry (234 – 305 AD)
  - Taxonomy, e.g., Taxonomy of Linnaeus (1707 - 1778)

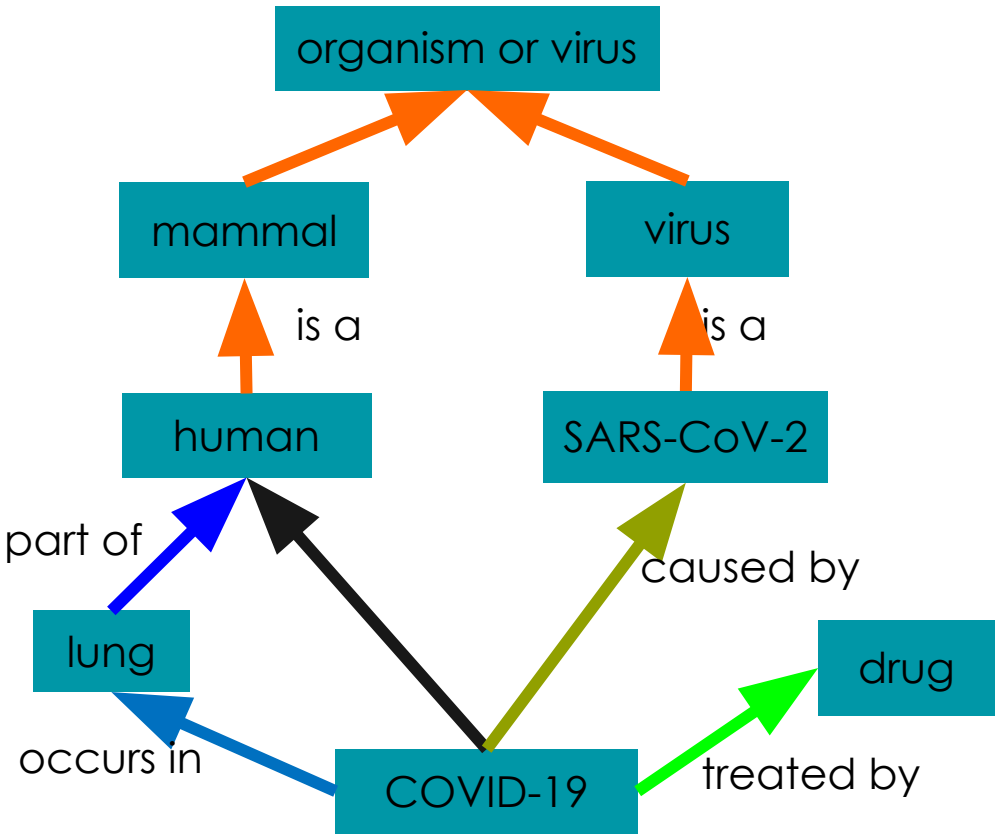


Porphyrian tree



Ontology in IT

- In IT era, ontology is human- and computer-interpretable representation of entities and the relations among entities in a specific domain.
  - A complex, standardized, and integrative network
- Foundation of AI knowledge representation and reasoning
- Support data/knowledge standardization, annotation, integration, and reasoning.



Ontology: Language of AI - Connecting machines & humans

Ontology = Entity terms (controlled vocabulary)
+ Relations (semantics)

Semantic: of, relating to, or arising from the **meanings** of words

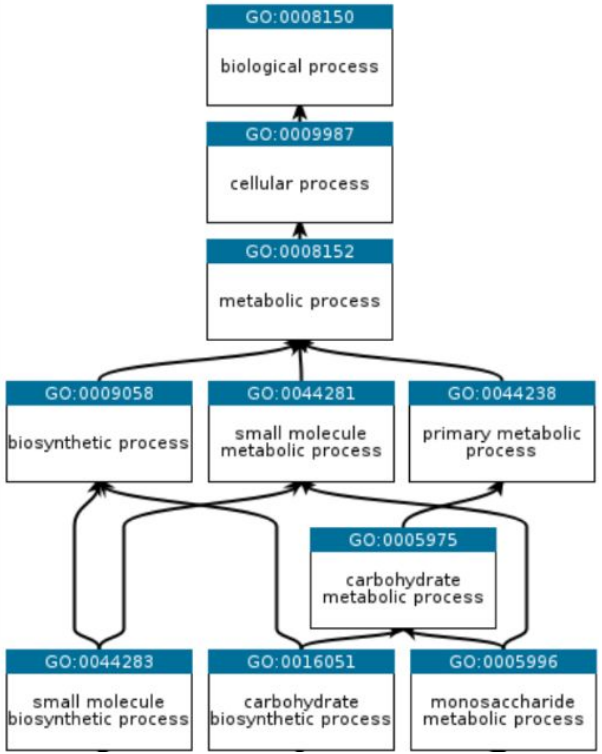
"The Semantic Web is an extension of the current web in which information is given well-defined **meaning**, better enabling computers and people to work in cooperation."
- Tim Berners-Lee, Inventor of WWW

https://en.wikipedia.org/wiki/Semantic_Web

Tower of Babel

Gene Ontology（GO）: Critical to gene/genome annotation & gene enrichment analysis, etc.

- Established in 1998 by a consortium studying the genomes of 3 model organisms: *D. melanogaster* (**fruit fly**), *M. musculus* (**mouse**), and *S. cerevisiae* (baker's **yeast**).
- **Goal**: Standardize/unify representation of gene functional annotation across databases and organisms.
  - 3 branches: Biological Process, Cellular Component, Molecular Function
- Critical to gene annotation and analysis.
- Stimulate more ontology development.



Open Biological and Biomedical Ontology (OBO) Foundry

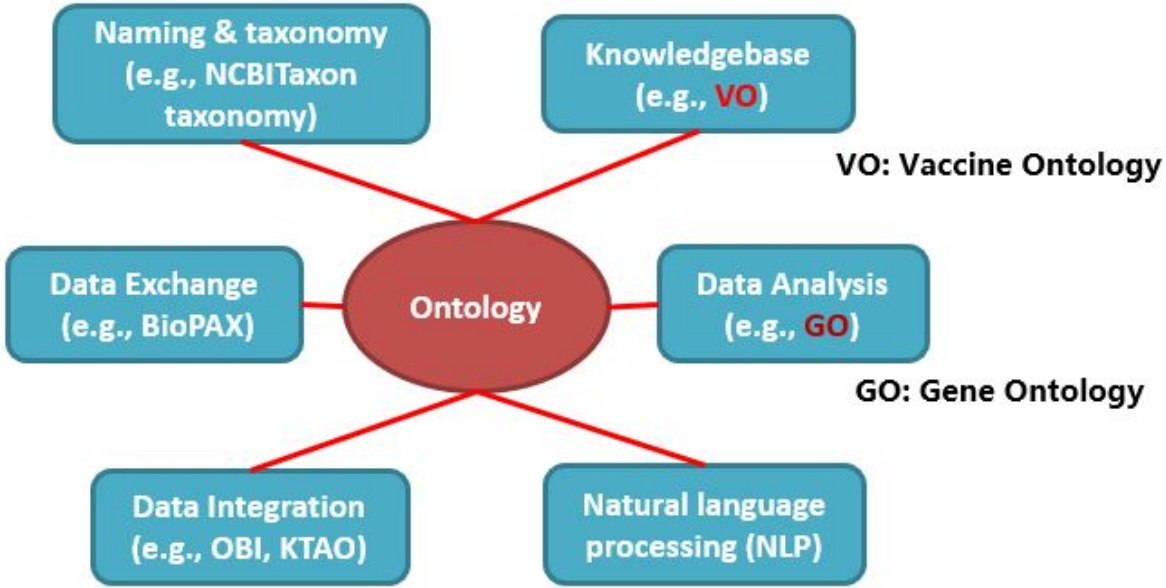
- International open initiative since 2006
- A collection of orthogonal reference ontologies in biological and biomedical domain.
  - Gene Ontology was the first one joining.
  - Other examples: Cell Ontology (CL), Mammalian Phenotype Ontology (MP), Human Phenotype Ontology (HPO), **Vaccine Ontology (VO)**, etc.
- Each is committed to a set of **principles** for the best practices in ontology development.

OBO Principles:

- Open (principle 1)
- Common Format (principle 2)
- URI/Identifier Space (principle 3)
- Versioning (principle 4)
- Scope (principle 5)
- Textual Definitions (principle 6)
- Relations (principle 7)
- Documentation (principle 8)
- Documented Plurality of Users (principle 9)
- Commitment To Collaboration (principle 10)
- Locus of Authority (principle 11)
- Naming Conventions (principle 12)
- Notification of Changes (principle 13)
- Maintenance (principle 16)
- Responsiveness (principle 20)

Applications of Ontology and Semantic Web

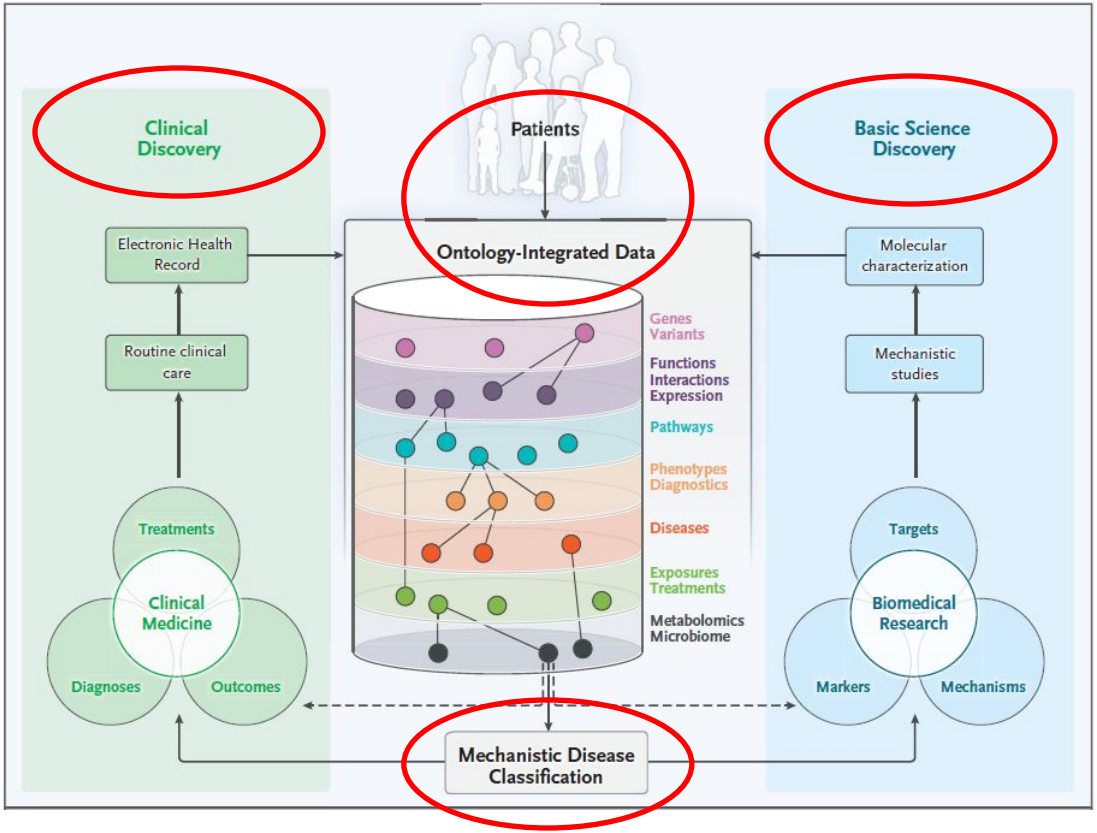
Ontology = Terms (controlled vocabulary)+ Relations (semantics)



OBI: Ontology for Biomedical Investigations

KTAO: Kidney Tissue Atlas Ontology

Ontology: Foundation of Precision Medicine



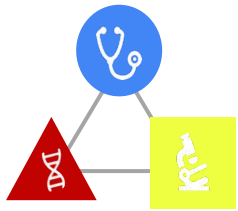
Integrating the two streams of data (**clinical** and **basic** science observations) enables more refined and dynamic classification of disease across many data types

Citation: Haendel MA, Chute CG, Robinson PN. **Classification, Ontology, and Precision Medicine**. *N Engl J Med*. 2018 Oct 11; 379(15): 1452-1462.

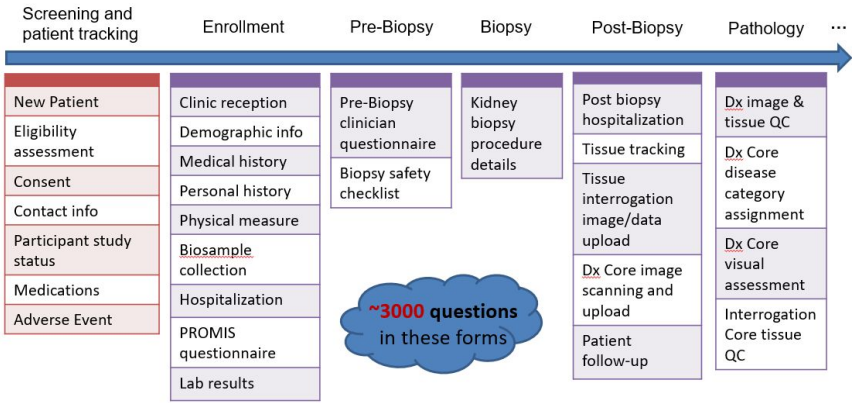
KPMP: Opportunities and Challenges

- Initiated 2017, Kidney Precision Medicine Project (**KPMP**), funded by NIH-NIDDK
  - Only human studies, no lab animals.
- Over 20 universities / institutes
- **Goals**:
  - Build a kidney tissue atlas that links *clinical* phenotypes, cells, *molecules*, pathways, and *pathology* together.
  - Understand and treat human kidney diseases – Acute Kidney injury (**AKI**) and Chronic Kidney Disease (**CKD**)

- **"Big data" challenge:** integration & analysis



- **Clinical**
- **Molecular**
- **Pathology**



38 KPMP Case Report Forms (CRFs)

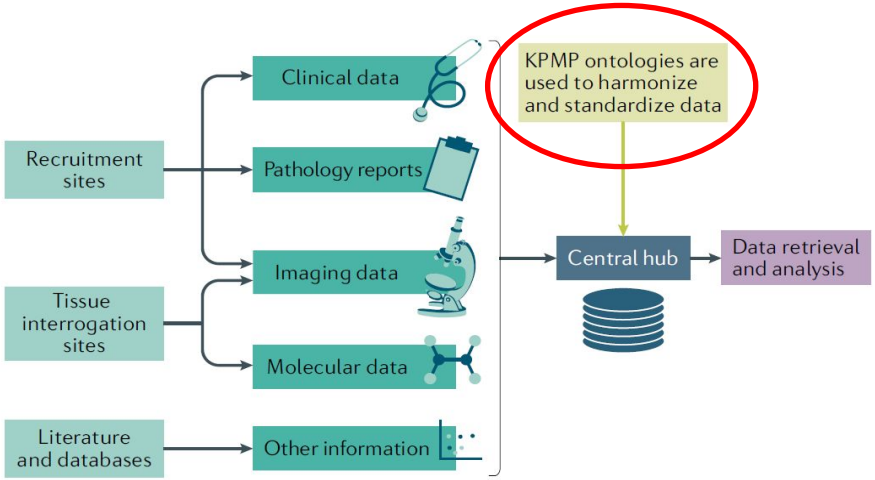
Two ontologies for KPMP

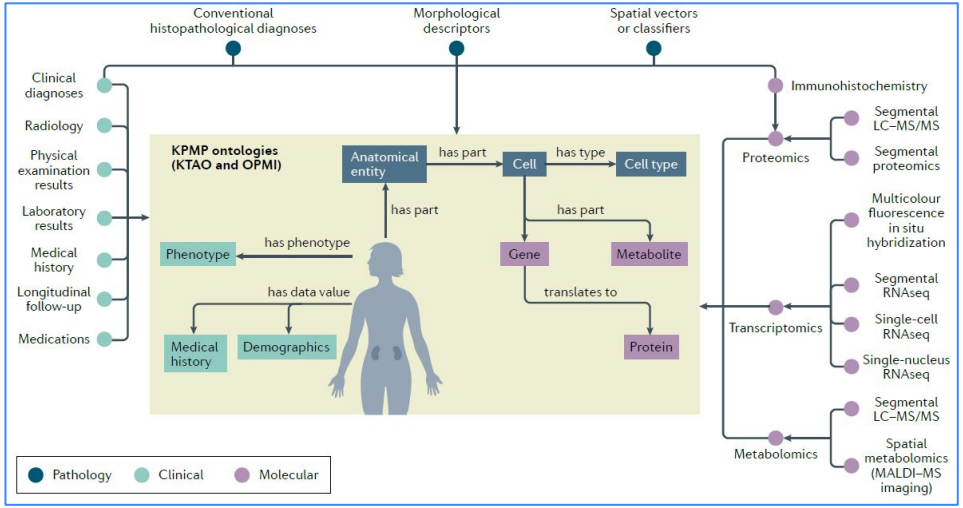
- Two community-based KPMP ontologies:
  - **KTAO: Kidney Tissue Atlas Ontology** – It's more about kidney knowledge
  - **OPMI: Ontology of Precision Medicine and Investigations** – Standardizes data and metadata types in and beyond KPMP.
  - Kidney-related info in OPMI is imported back to KTAO.

- Interoperable ontology development strategies
  - Follow Open Biomedical Ontology (**OBO**) **principles**: Openness, collaboration, etc. >150 OBO library ontologies: **non-redundant**, **interoperable**
  - **Reuse/align/integrate** existing ontologies: UBERON anatomical entity, HPO (Human Phenotypes), GO, CL (Cells), OBI (Biomedical Investigations), …

Ref: Ong E, Wang LL, Schaub J, O'Toole JF, Steck B, Rosenberg AZ, Dowd F, Hansen J, Barisoni L, Jain S, de Boer IH, Valerius MT, Waikar SS, Park C, Crawford DC, Alexandrov T, Anderton CR, Stoeckert C, Weng C, Diehl AD, Mungall CJ, Haendel M, Robinson PN, Himmelfarb J, Iyengar R, Kretzler M, Mooney S, and He Y, for the Kidney Precision Medicine Project. Modeling Kidney Disease Using Ontology: Perspectives from the KPMP. *Nature Review Nephrology*. 2020 Nov;16(11):686-696. PMID: 32939051.

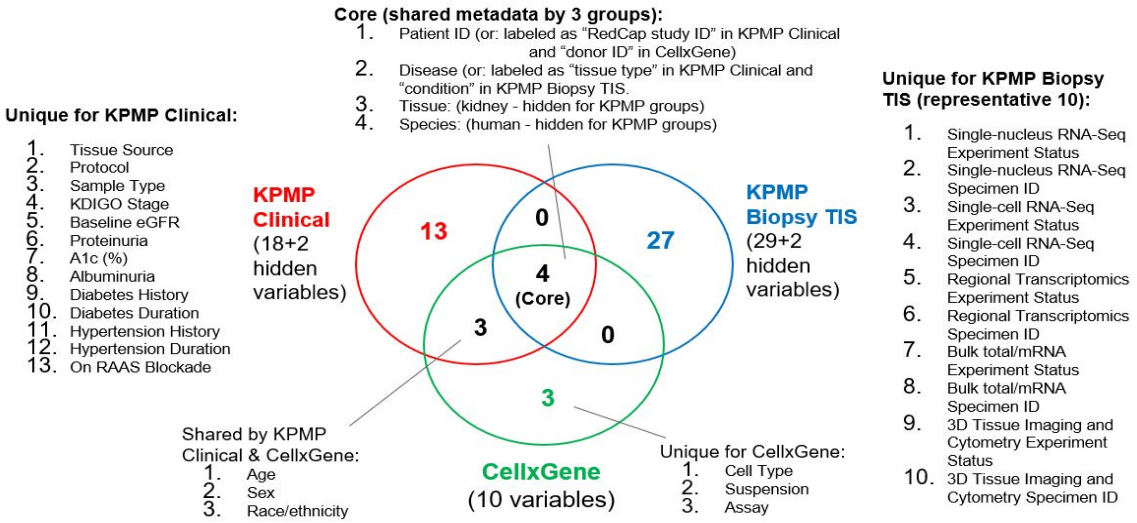
Ontology critical to KPMP big data integration and analysis





Ong, et al., *Nature Review Nephrology*, 2020

Integrate KPMP, HuBMAP, and CellxGene data using ontology



- **HuBMAP** (Human BioMolecular Atlas Program) focuses on reference human body.
- **KPMP** focuses on diseased kidney
- Even so, the results are not naturally integrated
- We proposed an interoperable ontology "**Precision Medicine Metadata Ontology (PMMO)**" to harmonize and integrate the data.

He Y, et al. AMIA 2024 full-length paper, oral presentation, Best Paper Award

HuBMAP Hackathon: Integrating KPMP & HuBMAP data by ontology

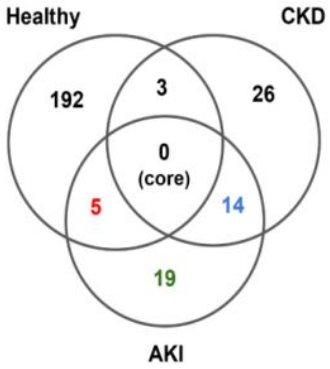
- **Goal:** Harmonize and integrate KPMP and HuBMAP data to more efficiently address scientific questions
- https://github.com/hubmapconsortium/hra-hubmap-kpmp-integration
- **Key process:** metadata harmonization using ontology
  - PMMO: Precision Medicine Metadata Ontology
- **Use case study:**
  - **SPP1:** A biomarker that differentiates healthy from AKI

Team members:

- Bruce Herr (IU)
- Yichao Chen (PSU)
- Leo Yeh (UM)
- Ruopeng Wu (UM)
- Oliver He (UM)

Biological Insights:

- **Theme**: Compare **gene biomarkers** between healthy and disease → leverage the info to **cell type** and **anatomy** levels.
  - This aligns with HubMAP **Anatomical Structures, Cell Types, and Biomarkers (ASCT+B) Tables.**
  - HuBMAP has collection of healthy kidney biomarkers
  - KPMP have collections of diseased kidney biomarkers.
- At **gene biomarker** level:
  - 26 AKI biomarkers were found, 5 also shared with healthy kidney
    - Previously, we started with KPMP/cellxgene data.
    - For the Hackathon, we focus on HuBMAP data and merged earlier data
  - ***Hypothesis***: An AKI/healthy gene biomarker(s) may have differential gene expression profiles in AKI patients vs healthy human subjects.
- At **cell** level:
  - Many biomarkers are for specific cell types. By analyzing the cell type specific biomarkers, we can indirectly find the cell type expression.
- At **kidney anatomical structure** level:
  - Specific cell types exist in specific regions. Through the **chain of biomarker-cell-Anatomy**, we can infer specific kidney region activities through the gene biomarker expression.
  - HuBMAP/KPMP histological image data can also be used later.



(A) Biomarkers

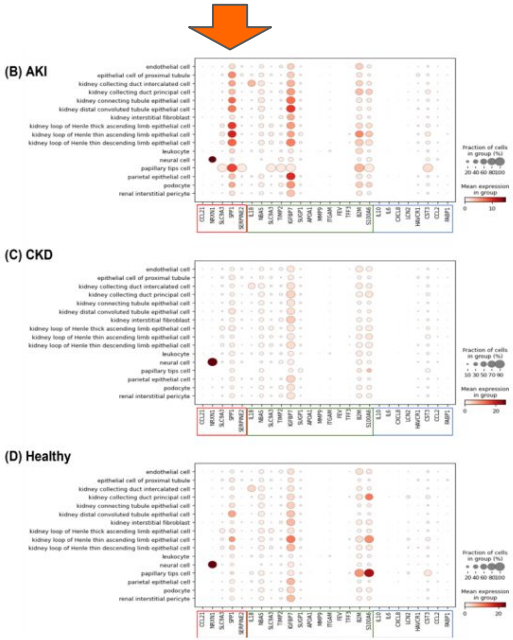
https://www.biorxiv.org/content/10.1101/2024.04.01.587658v1.full.pdf

Note: Paper also presented in **AMIA 2024 Annual Symposium, → Best Paper Award.**

SPP1: A biomarker that differentiates healthy from AKI

- **SPP1**: Secreted Phosphoprotein 1
  - https://www.genecards.org/cgi-bin/carddisp.pl?gene=SPP1
  - Key gene in lymph node metastasis and cancer, but its role in kidney still relatively unclear.
- **Earlier** we used KPMP and CellxGene data:
  - SPP1 significantly differed in gene expression in AKI and healthy groups
- **Now**:
  - **HuBMAP data** is added and merged with KPMP/cellxgene data.
  - Results so far:
    - Extracted SPP1 from KPMP and HuBMAP
    - Differential gene expression profiles found.
- More work ongoing. Ontology level integration as well

SPP1



https://www.biorxiv.org/content/10.1101/2024.04.01.587658v1.full.pdf

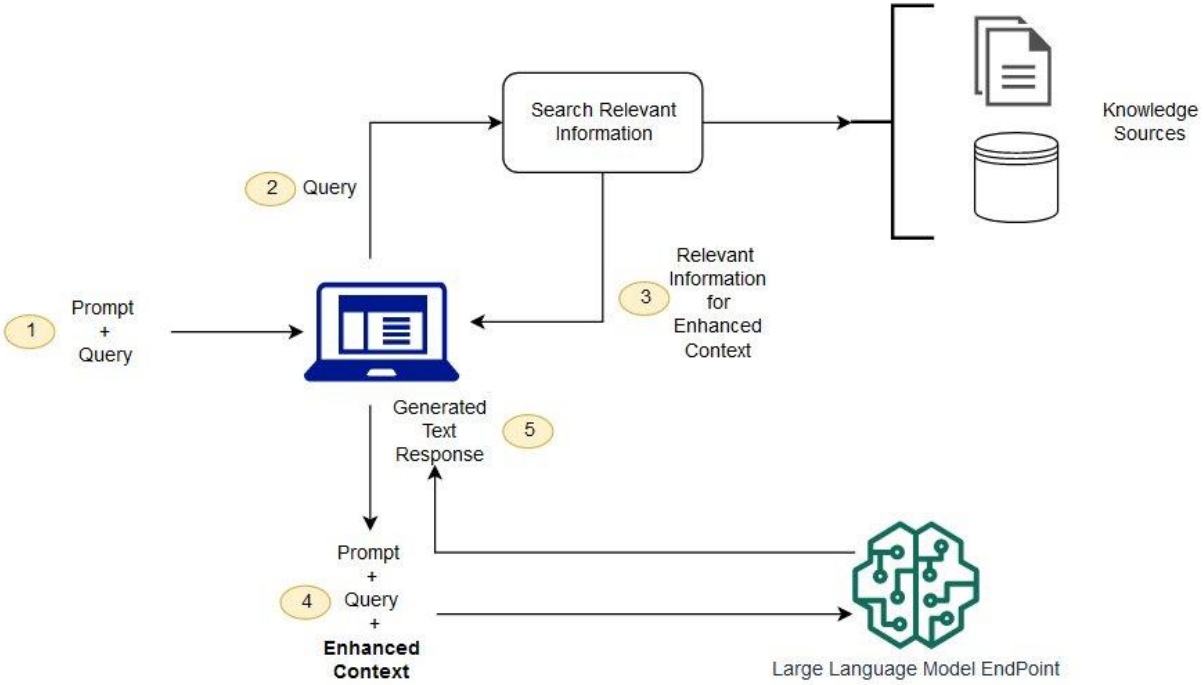
Ontology-Based Knowledge Graph

- **Knowledge graph examples：**
  - Google Knowledge Graph
  - COVID19-KB

- **Two methods of generating knowledge graph:**
  - Triplestore knowledge graph，using tools such as Virtuoso
  - Property knowledge graph, using tools such as Neo4j and GraphDB

- **Ontology role on knowledge graph (KG):**
  - Standardize basic KG framework, including data types & Semantic Relations
  - Computer-understandable knowledge, which can be directly used.
  - Ontology can be used to annotate data

Ontology-Based Knowledge Graph as RAG to Enhance LLM



RAG: Retrieval Augmented Generation

- Optimize LLM output by referencing an authoritative knowledgebase (KB) outside of its training data.
- Such **RAG KB can be generated with ontology support**.

https://aws.amazon.com/what-is/retrieval-augmented-generation/

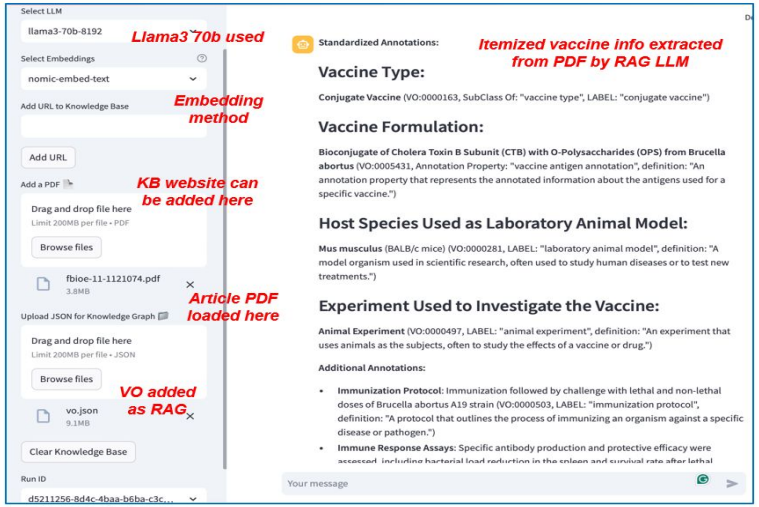
Example: Ontology-supported knowledge RAG for LLM extraction of itemized vaccine information (e.g., vaccine names, types, vaccine antigens, host responses, & experimental factors)

VIOLIN **manually collected ~5,000 papers and web links** for **~4,700 vaccines** and **1600 vaccine antigens** as of Nov 1, 2024, since 2007.

- So **~300 per year manually** annotated papers/links.
- Vaccines antigens used as gold standard for vaccine design.

Ontology RAG LLM would help

- Preliminary study approved it.
- **GOAL: 6,000 per paper per year (20-fold more productive).**



Use case demo: More collected and annotated **vaccine antigens** would serve as **gold standard** for enhanced vaccine design.

Discussion

- **How can ontologies improve AI, & how can AI improve ontologies?**
  - Ontology provides structured knowledge that enhances AI
    - Ontology-supported KG as RAG → improve AI
  - AI can identify new ontology terms/ relations and improve ontology applications

- **How can retrieval-augmented generation (RAG) specifically help with improving ontologies?**
  - Ontology-supported KG as RAG
  - RAG can retrieve docs/data and generate new ontology terms/relations.

- **How can AI chatbots (text-based) accurately represent ontologies (graph-based)?**
  - Chatbots can query ontology-converted KG or triple store
  - Incorporate precomputed inferences into chabot database

Acknowledgements

U. of Michigan

- Jie Zheng
- Anthony Huffman
- Asiyah Yu Lin
- Leo Yeh
- Edison Ong
- Laurel Li
- Michael Cooke
- Ruopeng Wu
- Oliver He
- Nikki Bonevich
- Jennifer Schaub
- Matthias Kretzler

Penn State

- Yichao Cheng

KPMP

- Jimmy P. Phuong
- Sean Mooney
- Jonathan Himmelfarb
- Laura Barisoni
- Jens Hansen
- Ravi Iyengar
- Avi Rosenberg
- All KPMP members.

HuBMAP / IN U.

- Bruce Herr II
- Katy Borner

OBO Foundry

- Alex Diehl
- Bill Duncan

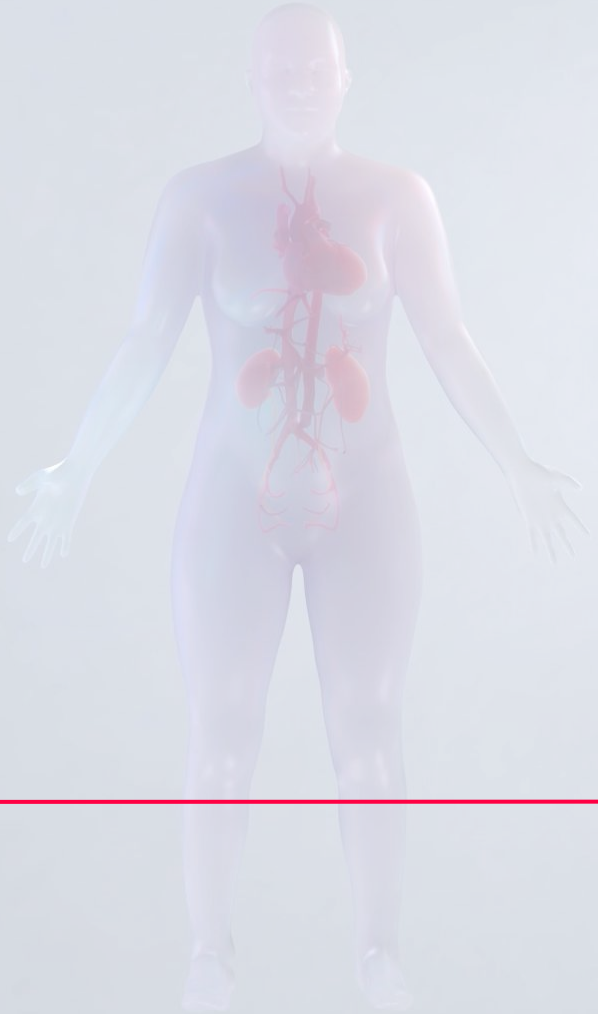
Funding:

KPMP is funded from the NIDDK: U01DK133081, U01DK133091, U01DK133092, U01DK133093, U01DK133095, U01DK133097, U01DK114866, U01DK114908, U01DK133090, U01DK133113, U01DK133766, U01DK133768, U01DK114907, U01DK114920, U01DK114923, U01DK114933, U24DK114886, UH3DK114926, UH3DK114861, UH3DK114915, UH3DK114937.

Human Reference Atlas (HRA) research and development is funded by NIH OT2OD033756 and OT2OD026671, U24CA268108, U24DK135157 and U01DK133090.

The work has also been supported by the **Kidney Precision Medicine Project grant U2CDK114886**, HHSN316201300006W/HHSN27200002, and **HuBMAP** U54 DK134301.

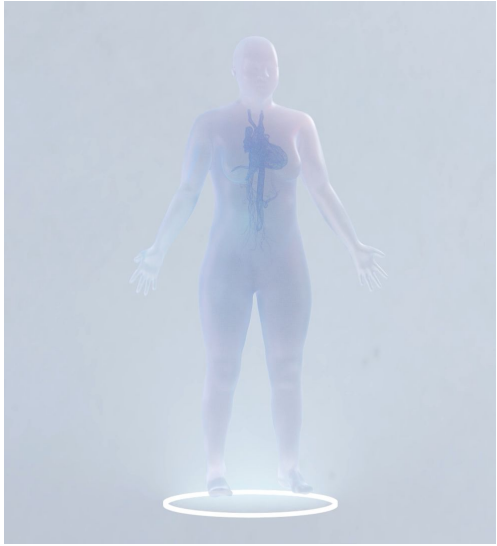
NIAID grants to YH: R01AI081062; UH2AI132931; U24AI171008.



Q&A

https://humanatlas.io/events/2024-24h

Q&A



- How can ontologies improve AI, and how can AI improve ontologies?
- How can retrieval-augmented generation (RAG) specifically help with improving ontologies?
- How can AI chatbots (text-based) accurately represent ontologies (graph-based)?

Thank you