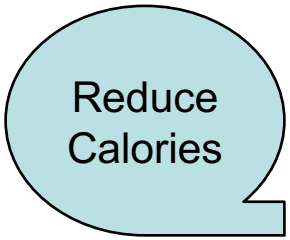
Aligning Diverse Data for a Personalized Health Knowledge Graph

Matthew Johnson, Meenu Ravi, Sabbir Rashid,
Paulo Pinheiro, and Deborah L. McGuinness

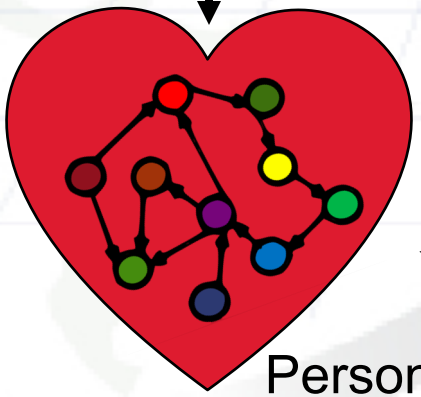
Rensselaer Polytechnic Institute, Troy, NY, 12180, USA

Personal Health Knowledge Graph

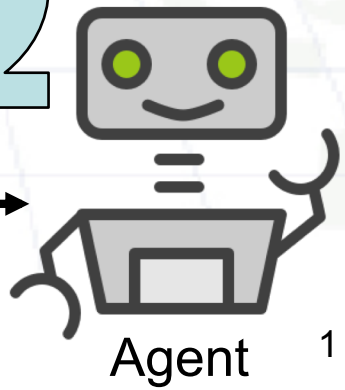


Reviews

Generates



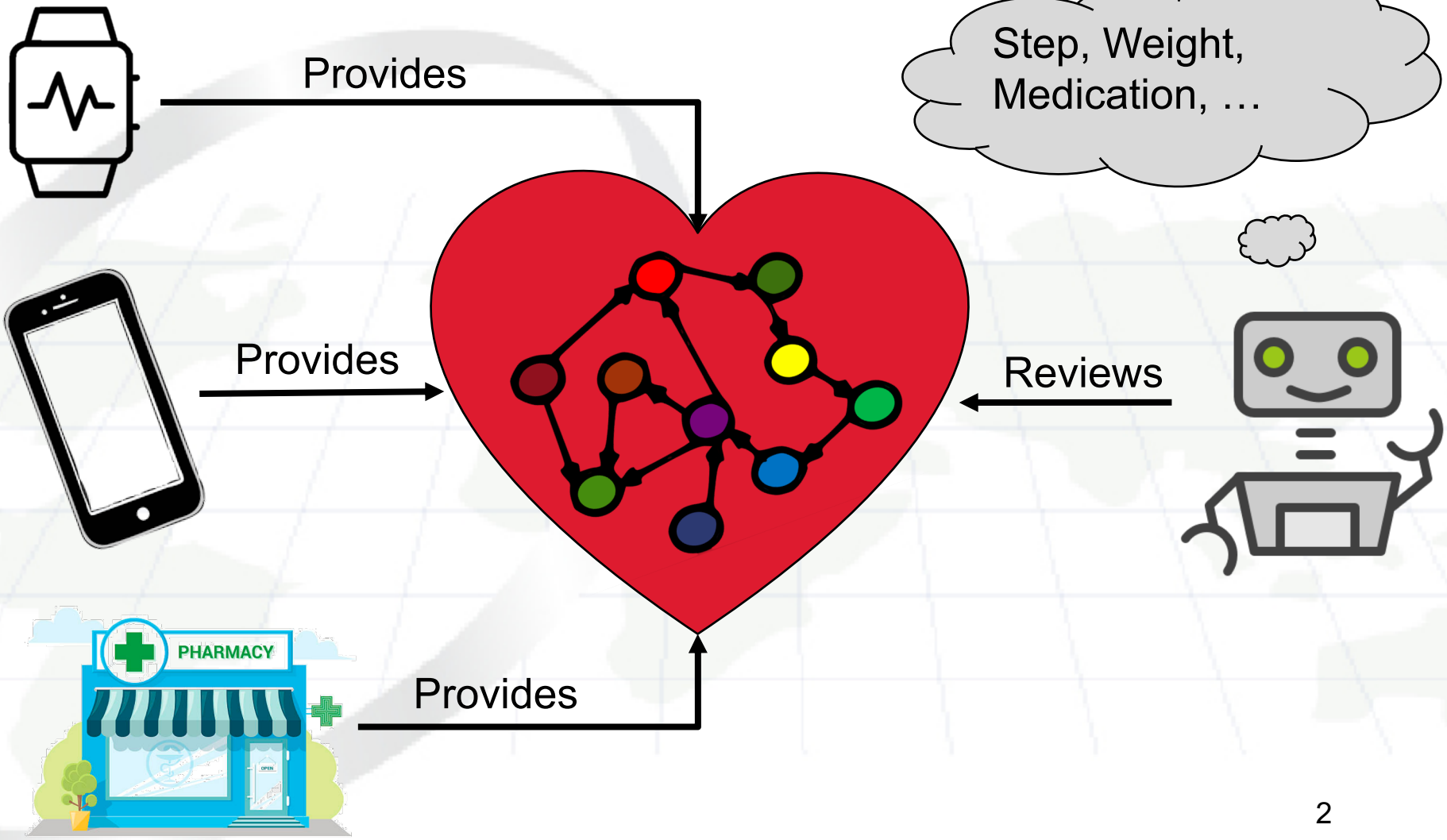
Personal Health Knowledge Graph



Reviews
Edits

Agent

Integration Problem



Semantic Data Dictionary

Data

id	age	gender	LNAS_C1
3479	23	…	…

…

Data Dictionary

id	Participant ID number
age	Age of mother at enrollment
gender	Child gender
LNAS_C1	Natural log of drinking water arsenic at enrollment

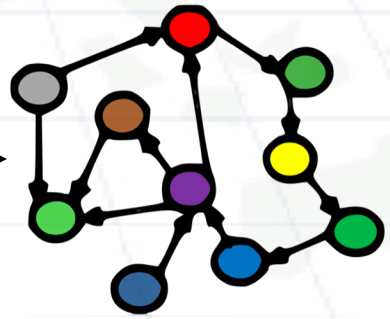
Column	Attribute	attributeOf	Unit
id	hasco:originalID	??mother	
age	sio:SIO_001013	??mother	sio:SIO_000428
gender	sio:SIO_010029	??child	

Virtual Column	Entity	Role	inRelationTo
??mother	sio:SIO_000485	hhear:00502	??child
??child	sio:SIO_000485	hhear:00492	??mother

Semantic Data Dictionary

Data

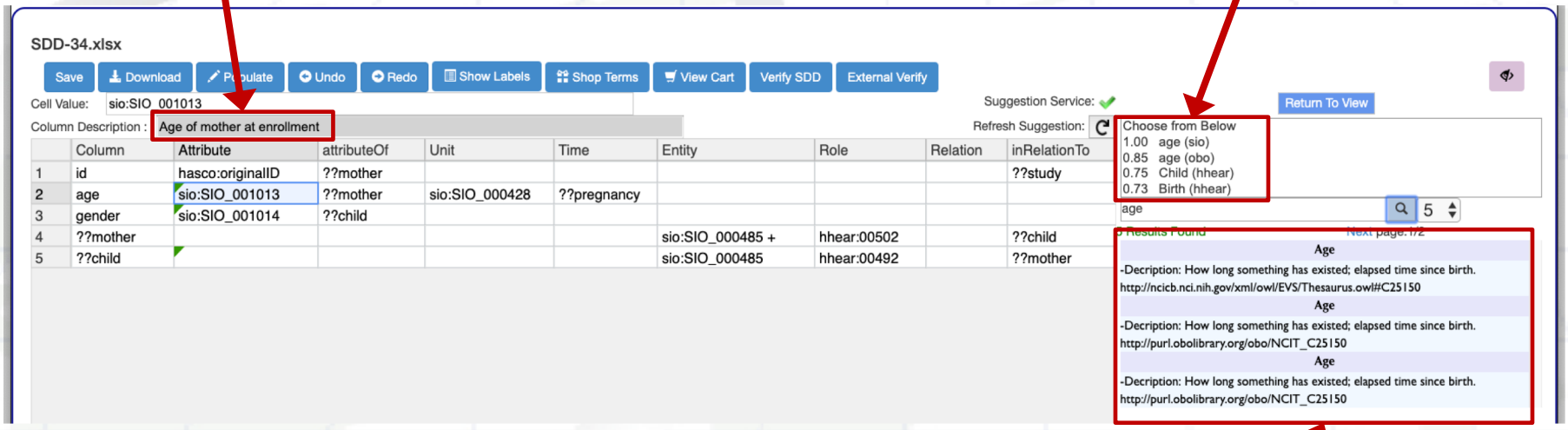
Knowledge Graph



SDD-Editor

Data Dictionary
Column Description

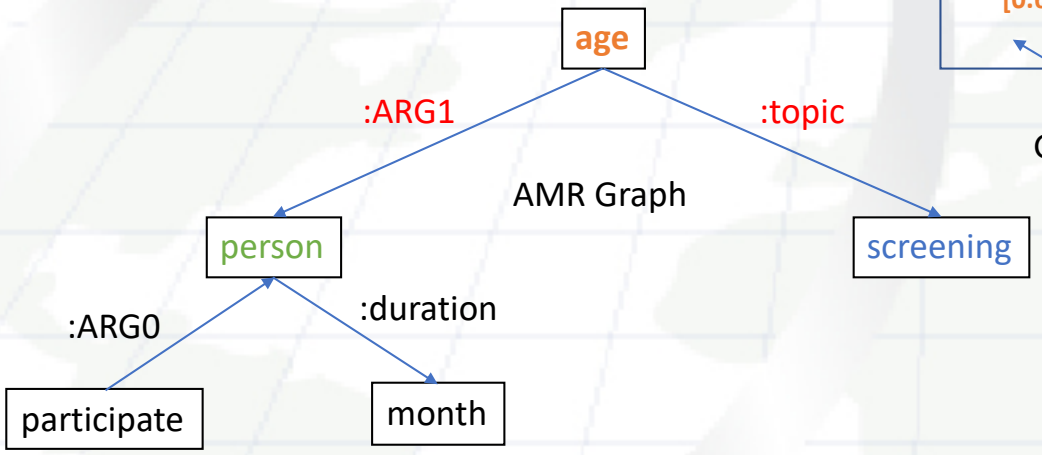
SDD-Gen Suggestions



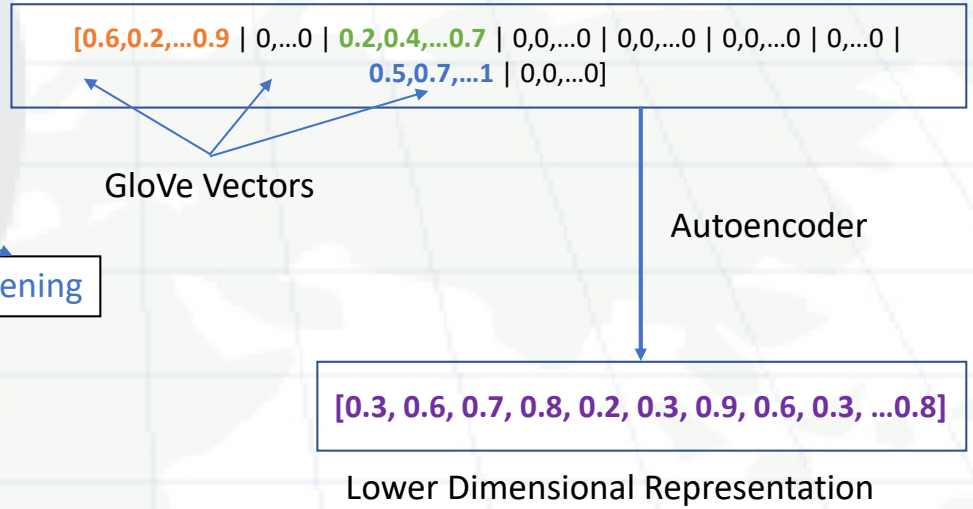
Bioportal Search Results

SDD-Gen

Age in months of the participant at the time of screening.



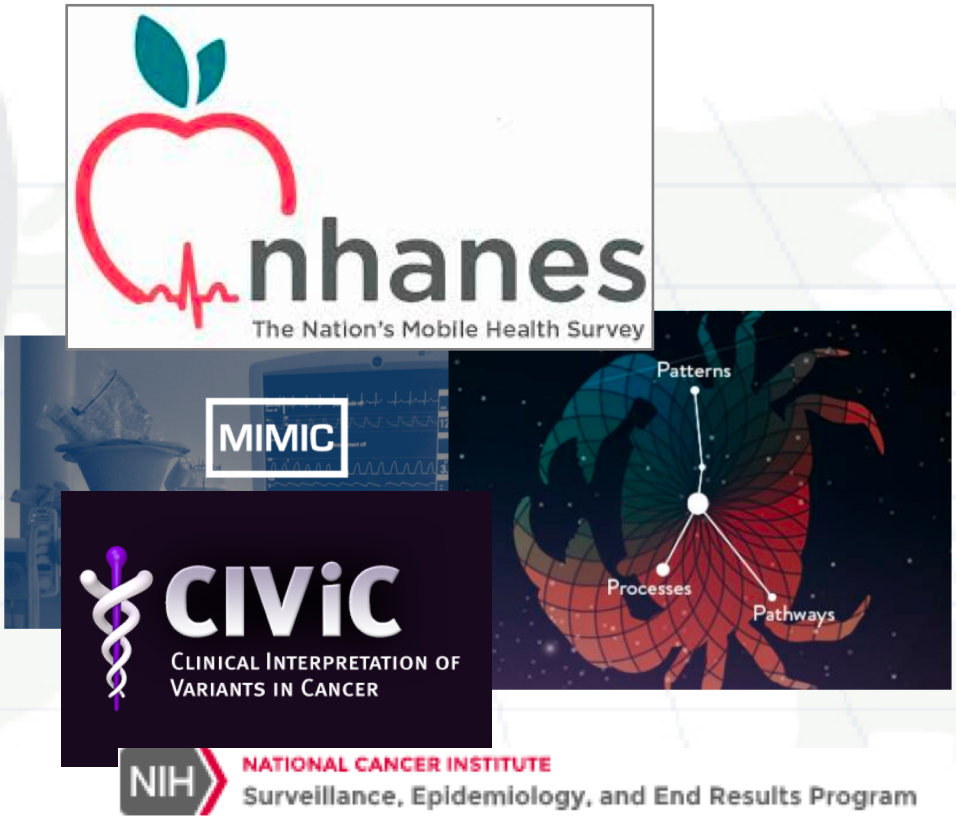
AMR entity-related Types
(:ARG0, :ARG1, :ARG2, :ARG3, :conj, :domain, :topic, :location)



L. Huang, J. May, X. Pan, and H. Ji, "Building a fine-grained entity typing system overnight for a new x (x= language, domain, genre)," *arXiv preprint arXiv:1603.03112*, 2016.

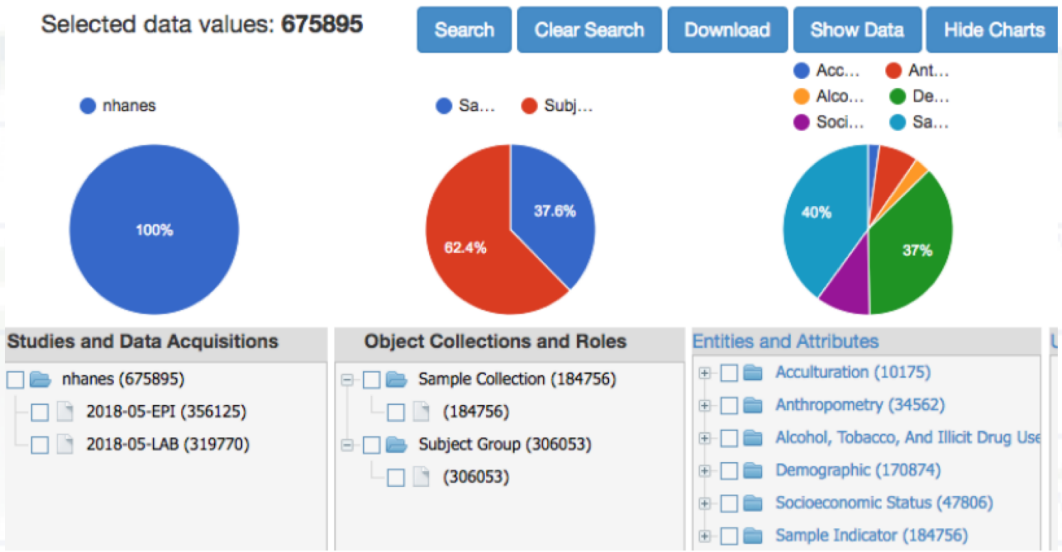
Data Integration: sdd2rdf

- Lightweight Python library
- Generates knowledge graph fragments as nanopublications
- Applied to a variety of biomedical datasets:
  - NHANES
  - TCGA
  - CIViC
  - SEER
  - MIMIC-III



Data Integration: HADatAc

- Framework to align data across studies
- Ingests files and message streams
- User Interface:
  - Explore ontologies
  - Ingest datasets
  - Investigate knowledge graphs



P. Pinheiro, H. Santos, Z. Liang, Y. Liu, S. M. Rashid, D. L. McGuinness, and M. P. Bax, “Hadatac: A framework for scientific data integration using ontologies.” in *International Semantic Web Conference*, 2018.

Conclusion

- Semantic Data Dictionaries can be used harmonize diverse datasets
- Our SDD-Editor allows users to semi-automatically generate Semantic Data Dictionaries, thus reducing user burden and broadening the class of users
- HADatAc and sdd2rdf can populate a PHKG using Semantic Data Dictionaries

Acknowledgements: This work is partially supported through AFRL 88ABW-2020-0991 and NIH 2U2CES026555-02