Integrated Health Care Datasets for Knowledge Discovery

Huawei Global Technology Summit 2022
July 7th, 2022

Christophe Guéret
christophe.gueret@accenture.com

accenture

Agenda

We'll go through topics around making, and then reasoning over a complex Knowledge Graph

1. Assembling and using a graph
2. -> Make
3. -> Use
4. Can we do things differently?
5. Take away messages

Why integrate healthcare data?

Three factors

- Increasing amount of data available
- Increasing data processing capabilities
- Improved outcomes using the “big picture”

One “omic” to study each ”ome”

- Genome: data about genes
- Exposome: data about exposition
- Proteome: data about proteins
- Etc…

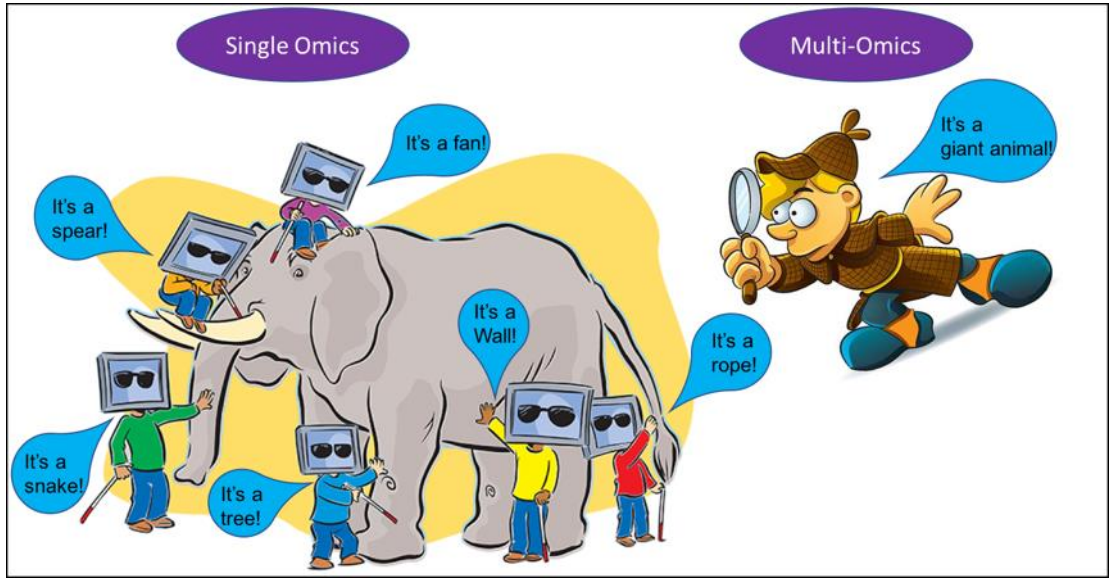
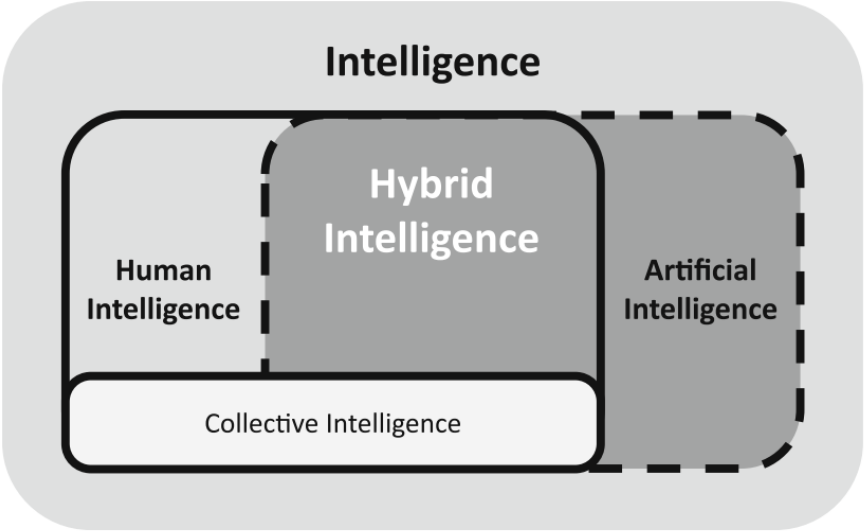


Image from Multi-Omics: a Revolutionary Approach to Data Analysis

KG-driven Knowledge Discovery

- Knowledge Graphs are an established way to connect data coming from different silos into so-called “360 views”
- Knowledge Discovery is the process of extracting useful information from this data.
- In our work, we aim at building AI systems to work with humans on knowledge discovery tasks. This ranges from exploring the data together to validating ideas using the data.



Positioning of Hybrid Intelligence [Dellermann 2021]

Dellermann, D., Ebel, P., Soellner, M., & Leimeister, J. M. (2021). *Hybrid Intelligence*. https://doi.org/10.1007/s12599-019-00595-2

H. James Wilson, & Paul R. Daugherty. (2022, March). Robots Need Us More Than We Need Them. Https://Hbr.Org/2022/03/Robots-Need-Us-More-than-We-Need-Them.

The story of the client and the Knowledge Graph geeks

CLIENT

”I need an AI system to help me work on a knowledge discovery task”

>

“We can make a graph and do graph machine learning on it”

CLIENT

“Cool! Here is all the data I have, feel free to enrich it with more stuff”

(sometime later…)

>

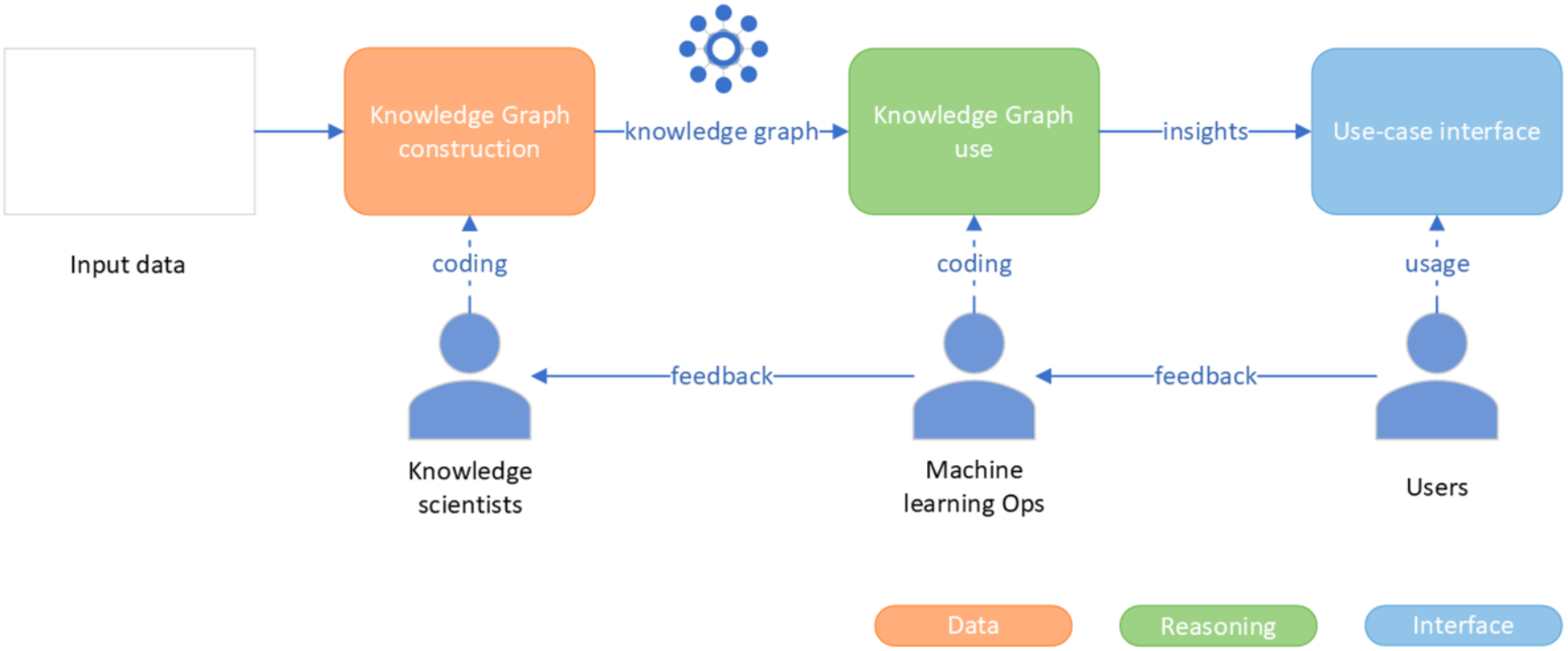
“Here you go; predictions! What do you think?”

Assembling and using a graph

Getting from raw data to insights

Two-step pipeline

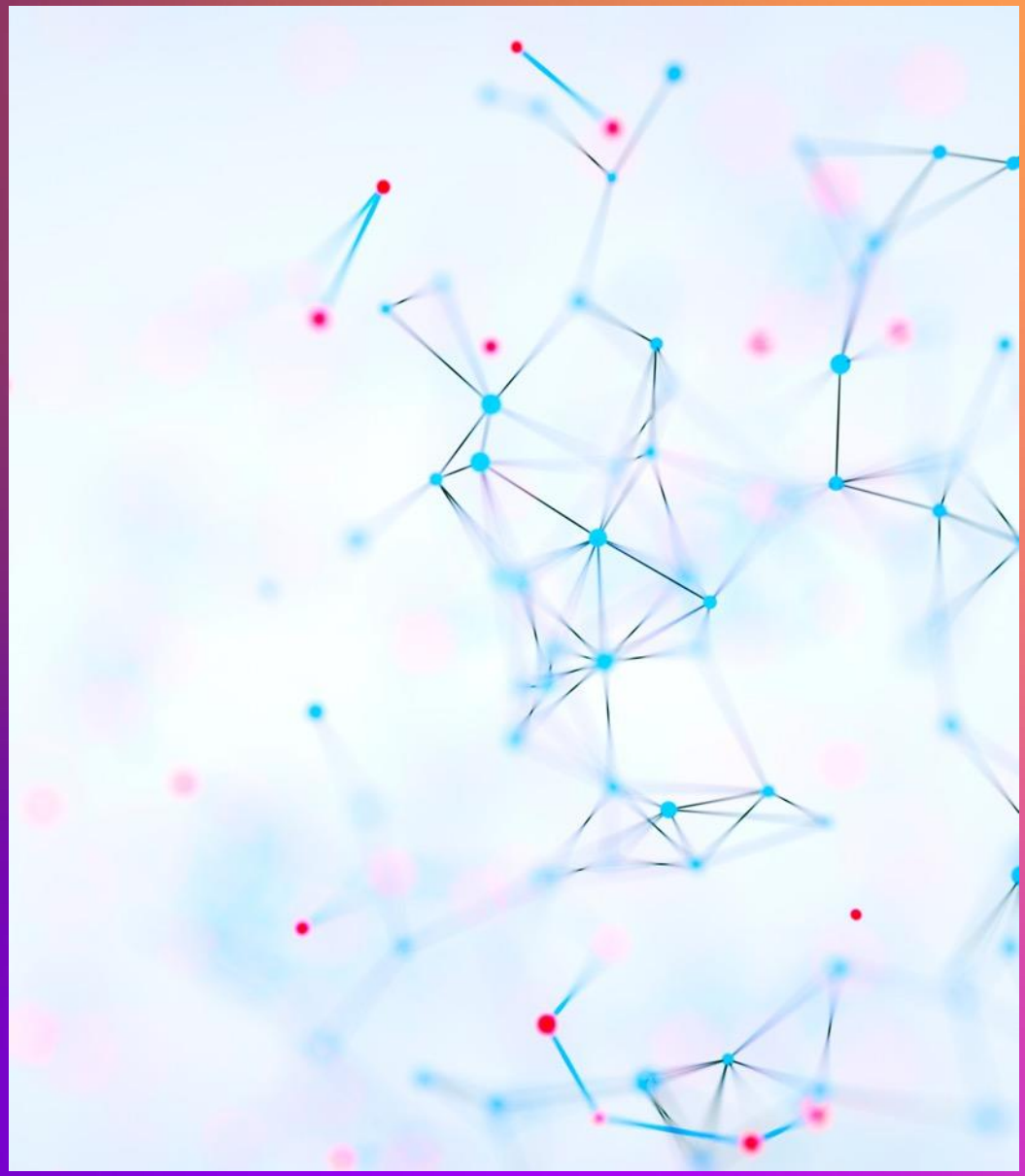
- State of the art is to: build a graph, ship it to someone using it, and then ship the outcome of the AI part. Then back-track and repeat as needed



Guéret, C. (2022). Knowledge Graphs in support of Human-Machine intelligence. *First International Conference on Hybrid Intelligence (HHAI)*. https://www.hhai-conference.org/wp-content/uploads/2022/06/hhai-2022_paper_67.pdf

Make

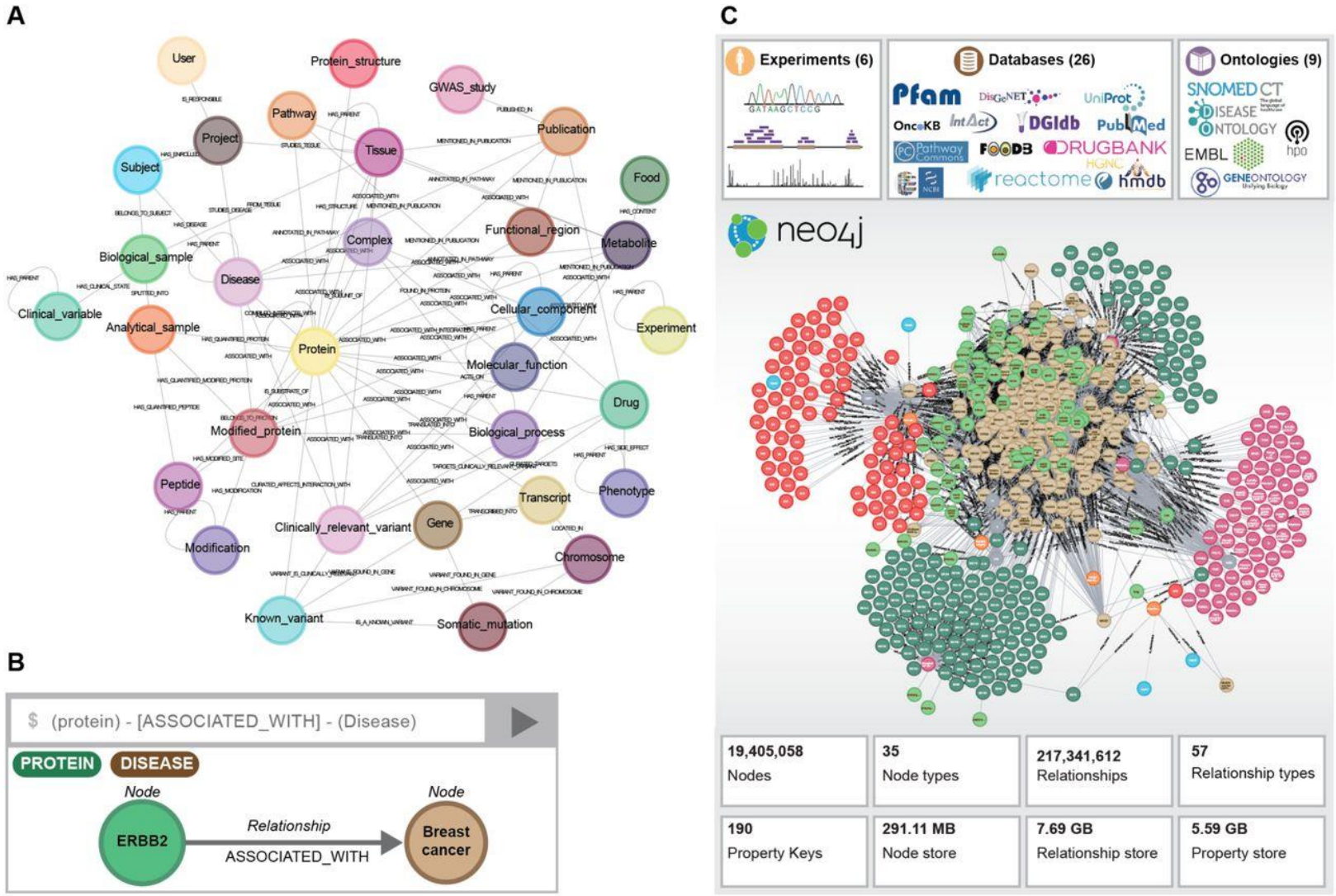
Assemble a nice, big, Knowledge Graph



Graphs as take away items

- There are a lot of different integrated medical KGs out there
  - PubChem
  - DISQOVER
  - DisGeNET - a database of gene-disease associations
  - Clinical Knowledge Graph (CKG)
  - Hetionet - An integrative network of biomedical knowledge
  - Open Pharmacological Space (openphacts.org)
  - …
- Collaboration networks such as Elixir are also interesting to study as a source of data and tools
- However, all those graphs are created with this one-size-fits all approach and share the other shortcomings of any “in house” graph constructed with SoTA approaches

One example graph: CKG



Santos, A., Colaço, A. R., Nielsen, A. B., Niu, L., Geyer, P. E., Coscia, F., Albrechtsen, N. J. W., Mundt, F., Jensen, L. J., & Mann, M. (2020). Clinical Knowledge Graph Integrates Proteomics Data into Clinical Decision-Making. *BioRxiv*, 2020.05.09.084897. https://doi.org/10.1101/2020.05.09.084897

Some challenges of make/use approach

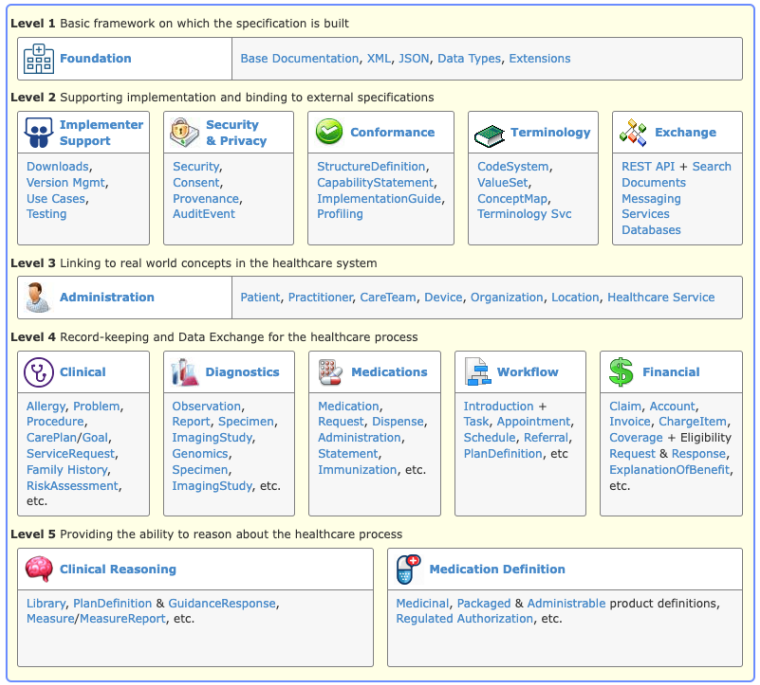
- **Data refresh**: it becomes challenging to release a new integrated KG to match a refresh of a single data source. Changes must be “big enough” to warrant a new release;

- **Data provenance**: the graph construction processes being decoupled from the graph consumption processes there is an information gap between the two;

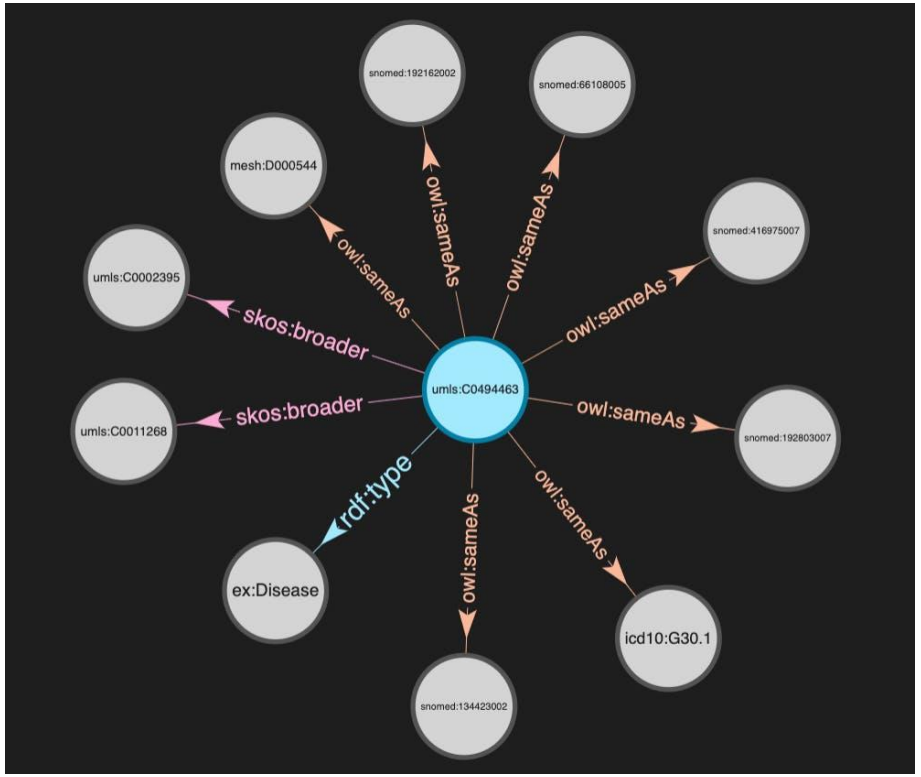
- **Data uniformity**: performing data integration is, by nature, about fitting a source world conceptualisation into a target one. The assumption is that the integrated graph is a one-size-fits-all one.

Data integration

- We need to select a target ontology / model and an approach to deal with the many identifier schemes used across all the fields



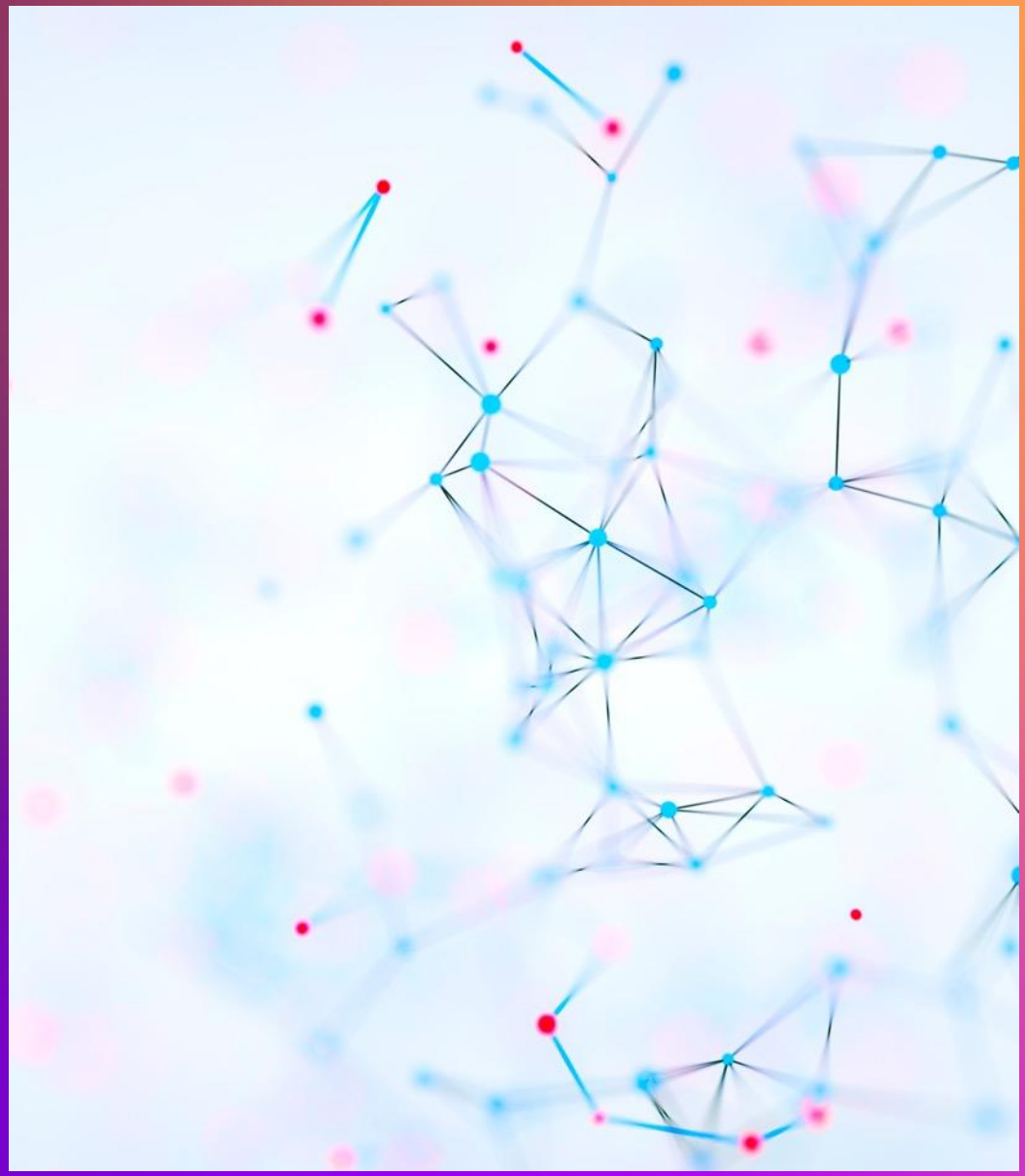
FHIR is an established data model to integrate data



sameAs reasoning can be applied to help with identifiers

Use

Now query the graph and do some machine learning with it



Example of questions

- The integrated KG provides exploration capabilities spanning over the whole spectrum of multi-omics data

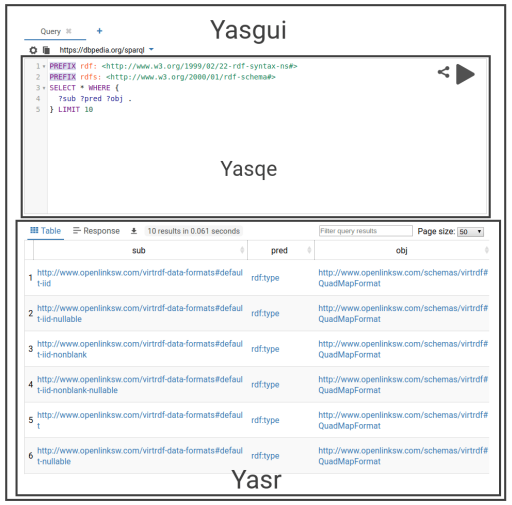
- Users can formulate this kind of queries over the graph
  - “What is the gene encoding protein X?”
  - “What are the drugs containing a compound acting on target Y?”

- We can add IF/THEN rules to infer some statements: “IF compound X acts on gene Y which encodes protein Z, THEN compound X acts on protein Z”

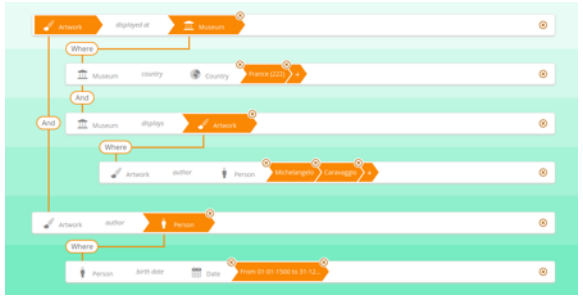
- The challenge for answering all the above queries is to align the dataset semantics in a target ontology and reconcile identifiers

Interactive exploration and query

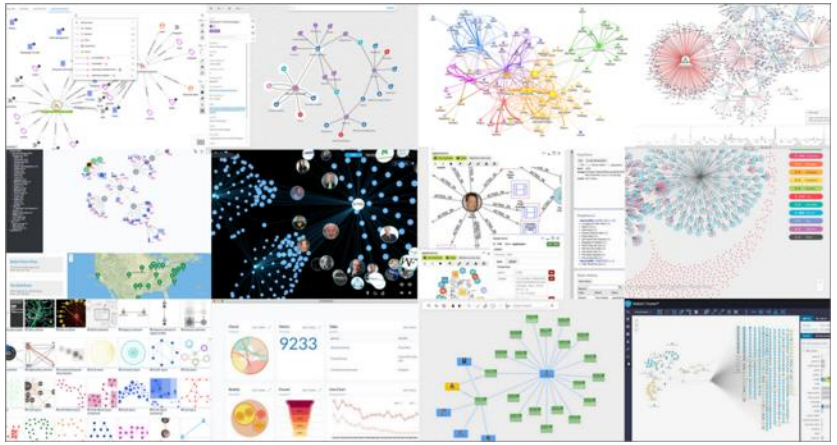
- There is no lack of options! Picking one depends on the target audience and the interaction pattern(s)



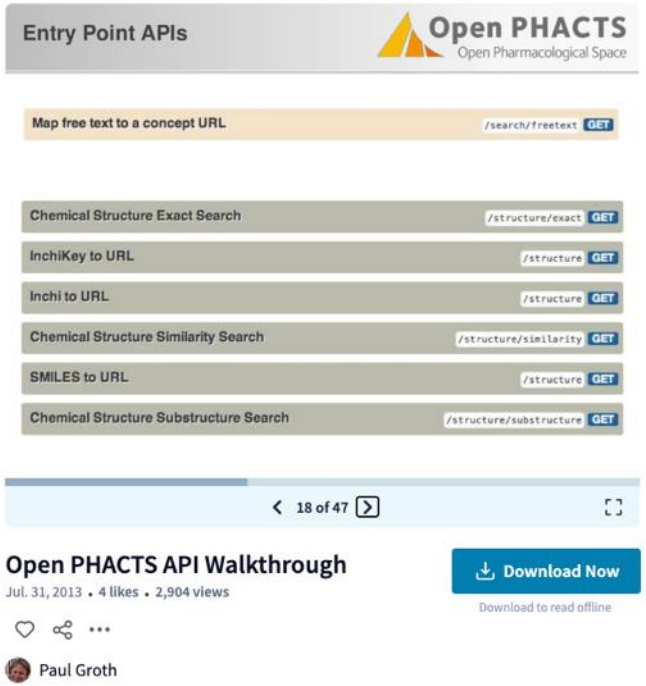
Yasgui API Reference - Docs - Triply



Sparnatural - Javascript SPARQL query builder



15 Best Graph Visualization Tools for Your Neo4j Graph Database

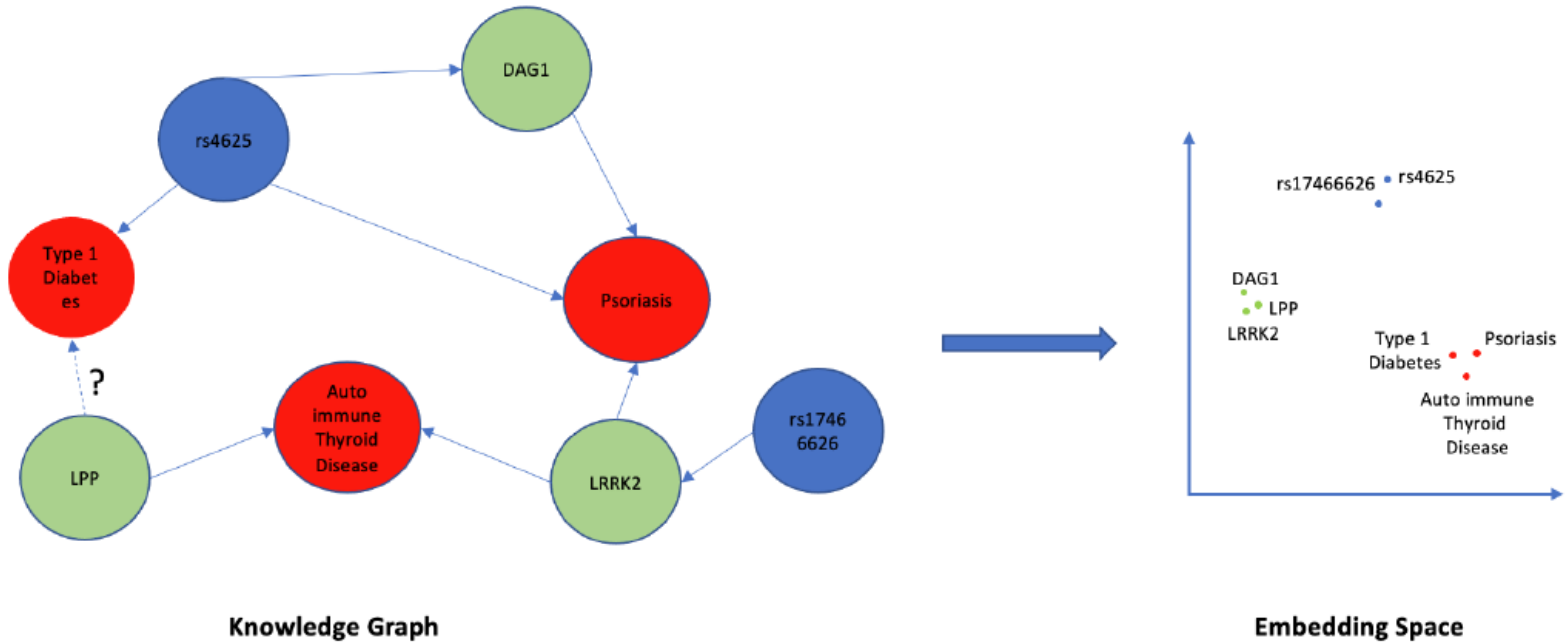


Open PHACTS API

Graph embeddings

AmpliGraph
Graph ML powered by Ampligraph

- The AI needs to learn the graph in order to reason over it. We do that by mapping the content of the graph into a vector space

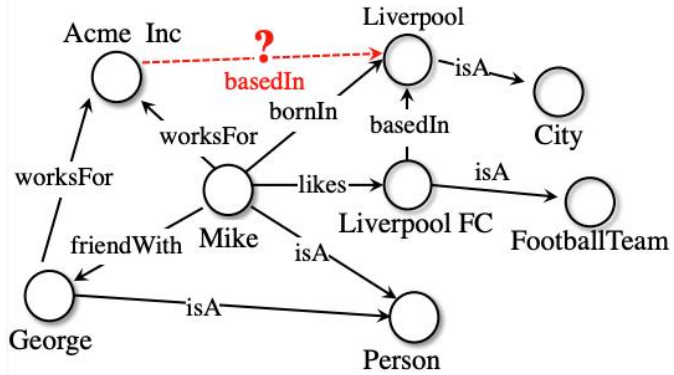


Knowledge Graph **Embedding Space**

Then, three possible type of investigations

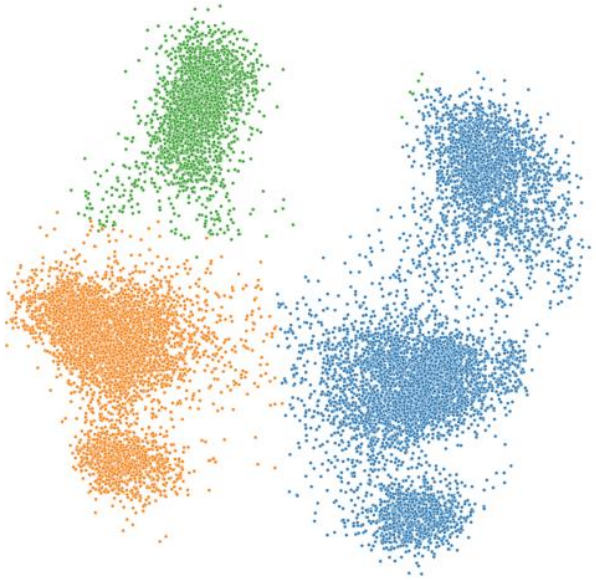
LINK PREDICTION / TRIPLE CLASSIFICATION

- Knowledge graph completion
- Content recommendation
- Question answering



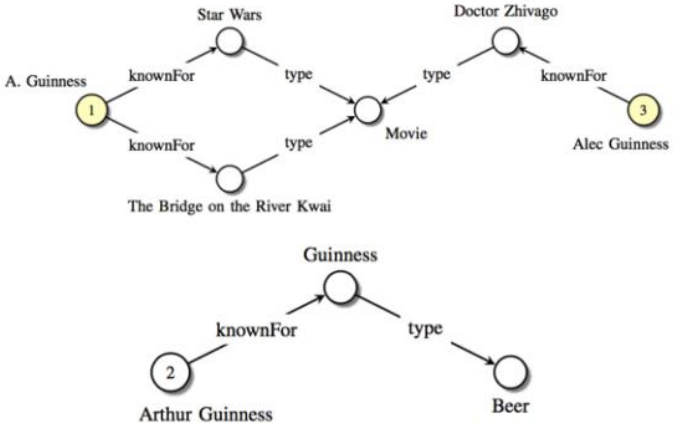
COLLECTIVE NODE CLASSIFICATION / LINK-BASED CLUSTERING

- Customer segmentation



ENTITY MATCHING

- Duplicate detection
- Inventory items deduplication

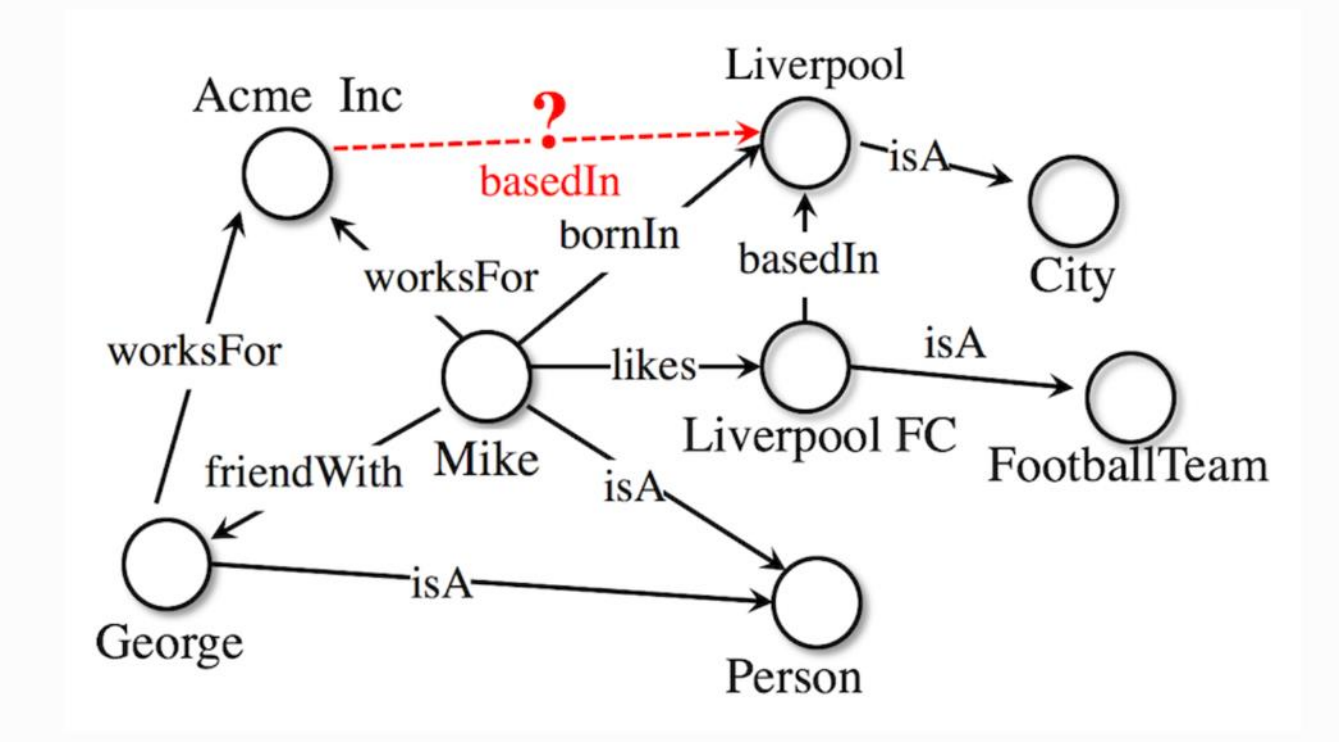


Pic from [Nickel et al. 2016a]

Costabello, L., Pai, S., McCarthy, N., & Janik, A. (2020). *Knowledge Graph Embeddings Tutorial: From Theory to Practice*. Zenodo. https://doi.org/10.5281/zenodo.4268208

Predicting links

- The goal is to predict what is the likelihood of a link not present in the graph
- The outcome depends on all the content in the graph, at any distance from the target nodes
- We have a sub-graph explanation sub system able to state which nodes were most influential in the scoring



The task here is to predict if the link in red could be a statement in the graph

Training with Uncertainties/Importance of links

- As an extension to state of the art graph machine learning we developed an approach to weigh the links based on importance

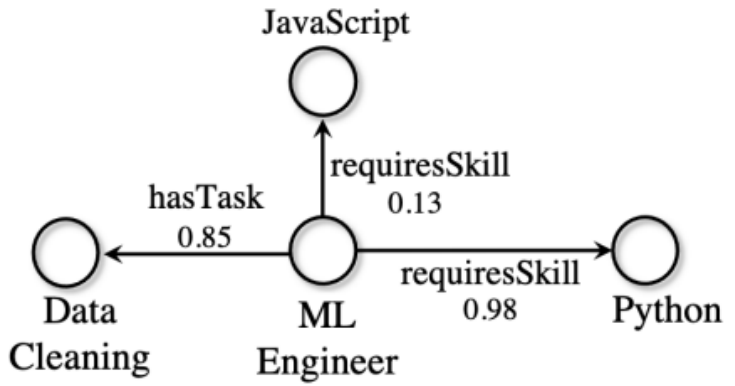


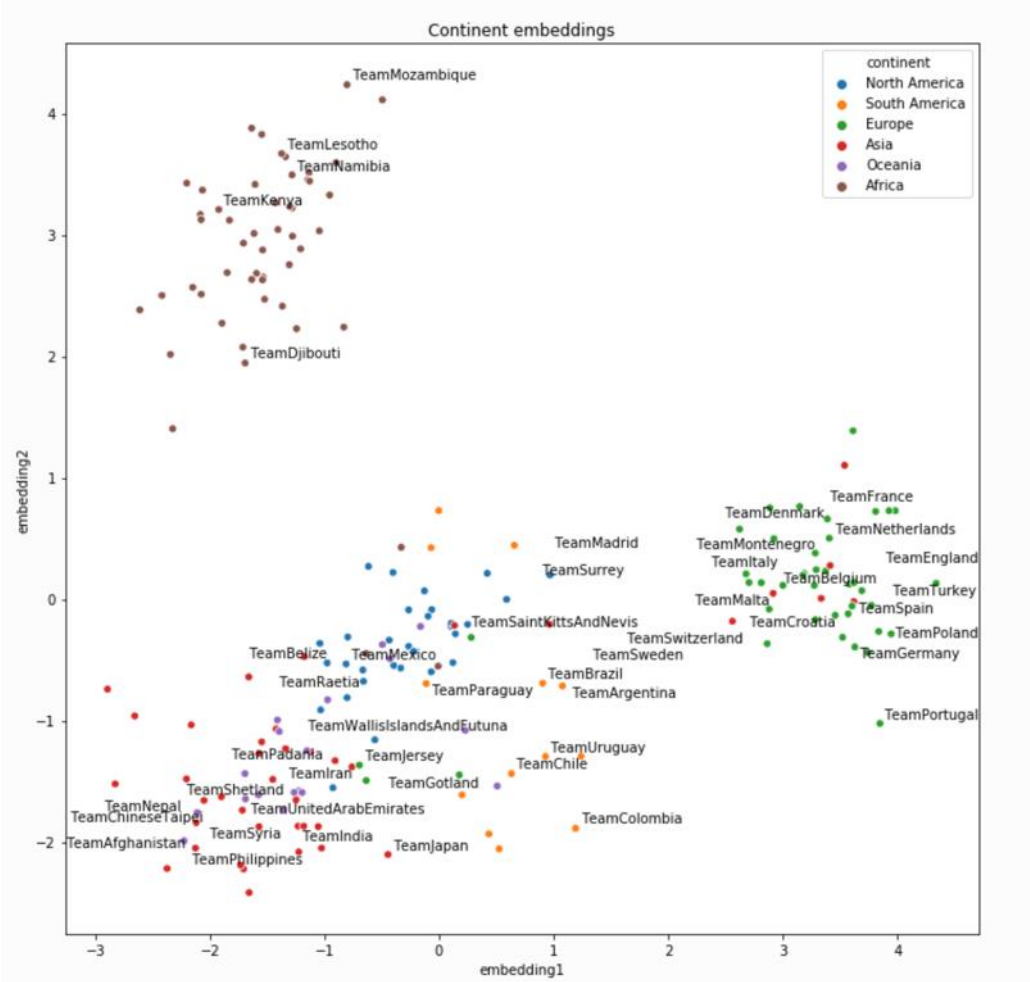
Figure 1: A Knowledge graph with numeric attributes associated to triples.

S Pai, L Costabello(2021). Learning Embeddings from Knowledge Graphs With Numeric Edge Attributes: IJCAI 2021

Clustering

- If we cluster entities based on their vector representation, we see emerging features

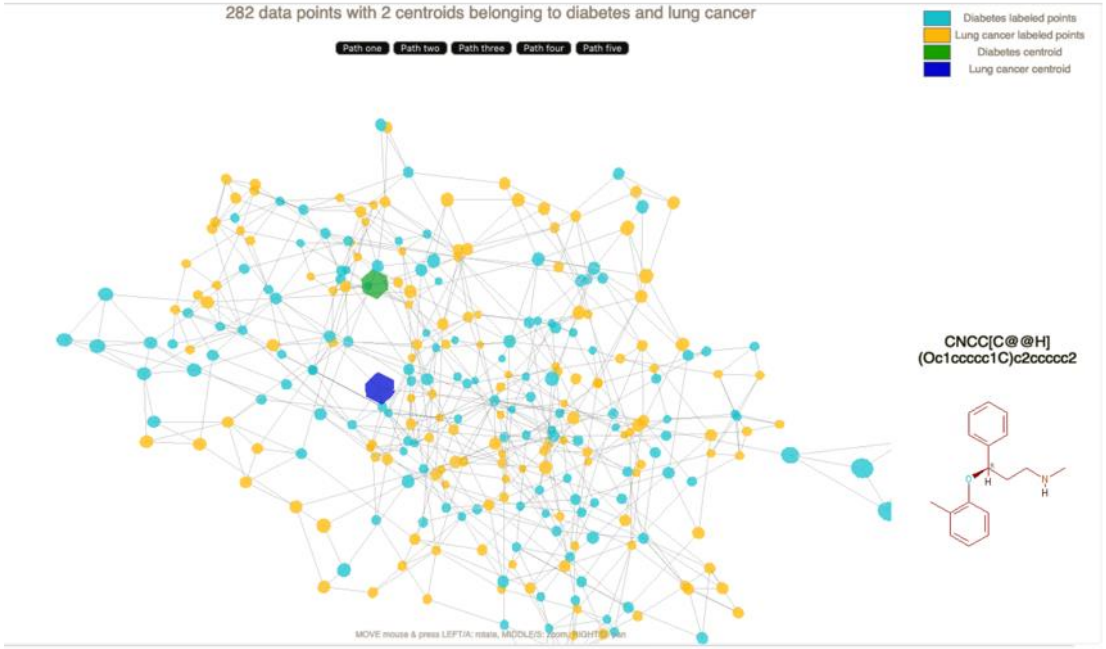
- In this example, the continents are not in the graph but emerge from other type of edges



Clustering the nodes based on continents (image from [Examples — AmpliGraph 1.4.0 documentation]())

Using embedding space for generation

- In ChemoVerse we learn to represent and generate structures of small molecules from an embedding space (or latent space).
- We use a manifold traversal with heuristic search to explore a latent space created from knowledge on this chemical space.
- Different heuristics and scores such as the Tanimoto coefficient, synthetic accessibility, binding activity, and QED drug-likeness can be incorporated to increase the validity and proximity for desired molecular properties of the generated molecules.
- With this novel traversal method, we are able to find more unseen compounds and more specific regions to mine in the latent space.

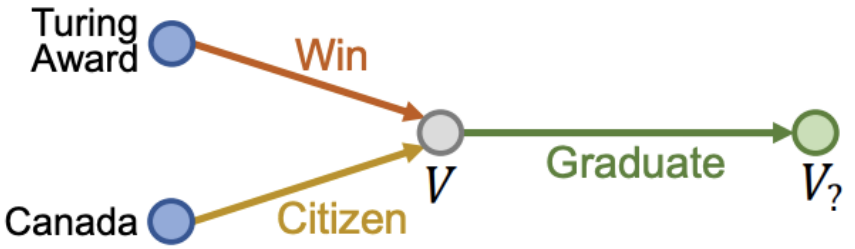


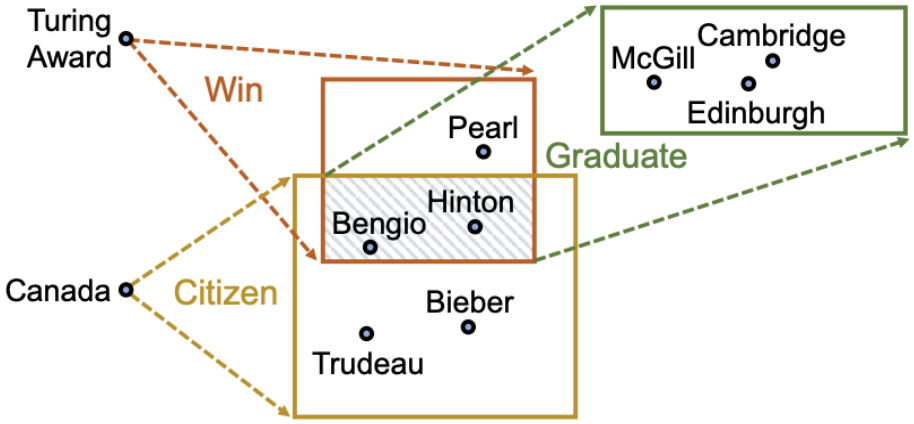
Singh, H., McCarthy, N., Ain, Q. U., & Hayes, J. (2020). *ChemoVerse: Manifold traversal of latent spaces for novel molecule discovery*. https://doi.org/10.48550/arxiv.2009.13946

Searching for complex combinations

- For future work, we are considering doing query answering using the embedding space. Eventually using an approach like Query2Box

$$q = V_? \,.\, \exists\, V : Win(TuringAward, V) \wedge Citizen(Canada, V) \wedge Graduate(V, V_?)$$





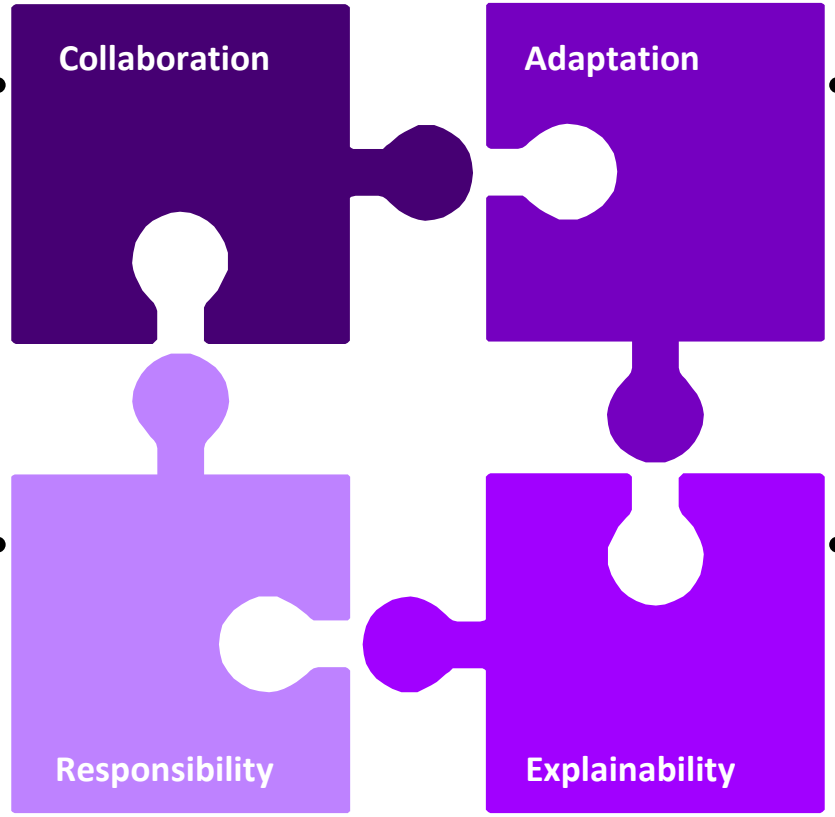
H Ren , W Hu , J Leskovec, QUERY2BOX: Reasoning over Knowledge Graphs in vector space using box embeddings, ICLR 2020

What can we do better?

Exploring what could be in the graph

Is this ok for Hybrid Intelligence?

One-way and non iterative insights only, the AI ignores input from humans



No adaptation to a change of context without changing the input data

Assumption that everything in the knowledge graph is ok to use and share

The reasoning process of humans is not taken into consideration

Grid derived from: *Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., … Welling, M. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. Computer, 53(8), 18–28.* https://doi.org/10.1109/MC.2020.2996587

Opportunities

3 keys things we can leverage to improve on our pipeline

Reasoning rules

Lack of data compensated by expert knowledge

Rise of data fabrics

KG constructed on demand rather than pre-assembled

Shared understanding

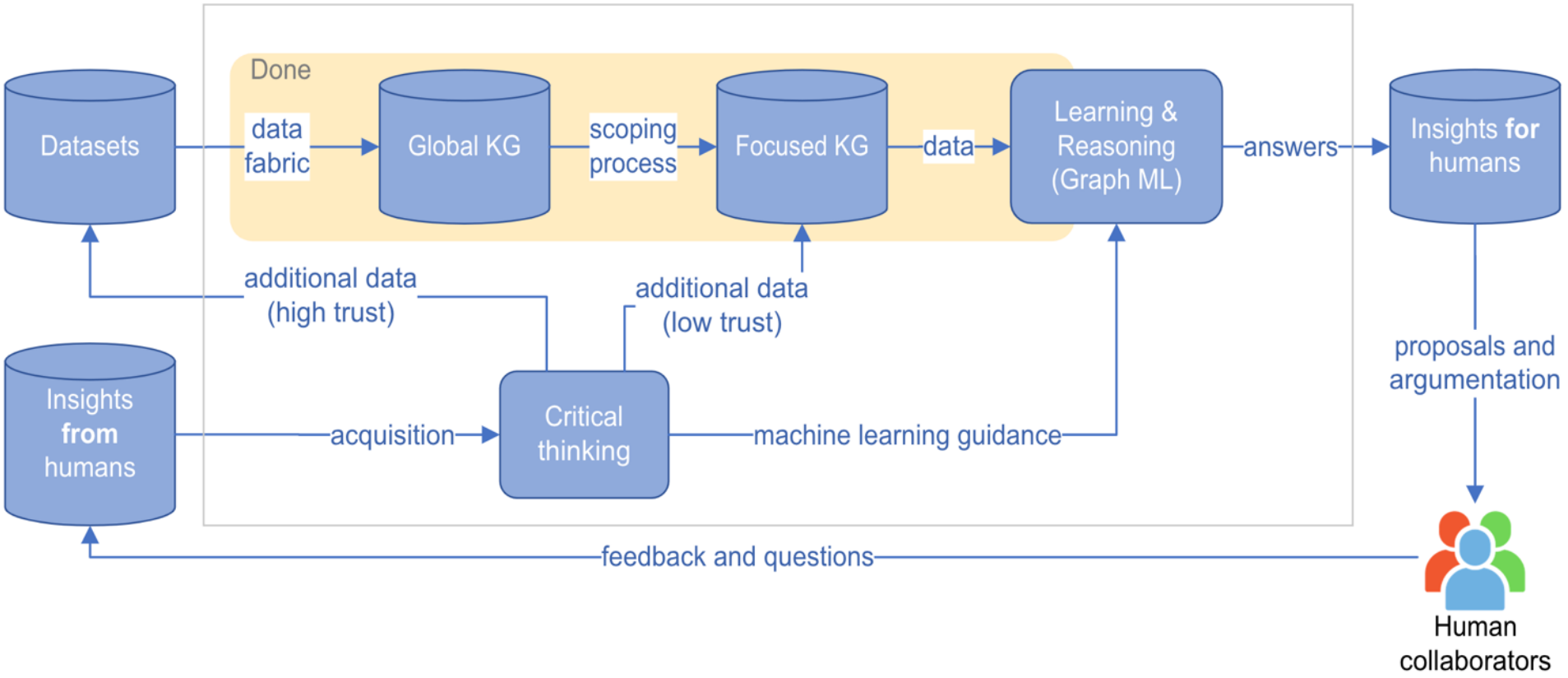
Graphs used as is by all stake-holders, human and AI

Guéret, C. (2022). Slides supporting the presentation of "Knowledge Graphs in support of Human-Machine intelligence". *First International Conference on Hybrid Intelligence (HHAI)*.
HHAI June 2022 - KGs and Hybrid Intelligence (slideshare.net)

Our proposed approach

AmpliGraph
Graph ML powered by Ampligraph

- Add scoping and critical thinking elements. The goal is to incorporate information from the interaction with the users down to the KG construction



Guéret, C. (2022). Knowledge Graphs in support of Human-Machine intelligence. *First International Conference on Hybrid Intelligence (HHAI)*. https://www.hhai-conference.org/wp-content/uploads/2022/06/hhai-2022_paper_67.pdf

Take-away

Knowledge Graphs can be a key back-end component when introducing AI collaborators in a team

They enable:

- Having all stake-holders use the same conceptual model
- Reason and discuss over this model
- Put up a bidirectional data-to-insight pipeline

To chat more please reach out at:
christophe.gueret@accenture.com