CAIR-Nepal
Shaping the Future with AI

From Data to Knowledge: Driving Innovation with Knowledge Graphs

Dr. Tek Raj Chhetri

7 December 2024
KGSWC 2024 Winter School

About me

- Founder & Director of CAIR-Nepal (Center for Artificial Intelligence (AI) Research Nepal)

  - https://cair-nepal.org

- Postdoctoral Associate at Massachusetts Institute of Technology (MIT), USA

- Additional details about me can be found at

  - https://tekrajchhetri.com

CAIR-Nepal
Shaping the Future with AI

Outline

1. Introduction
2. Why use Knowledge Graphs?
3. Innovations Enabled by Knowledge Graphs
4. Case Studies
5. Conclusion & Future Outlook

1.

Introduction

1. Introduction

- First Wave: Handcrafted Knowledge

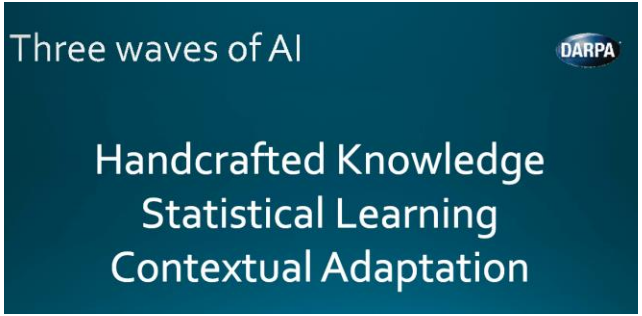
  - e.g., rule based (expert systems)

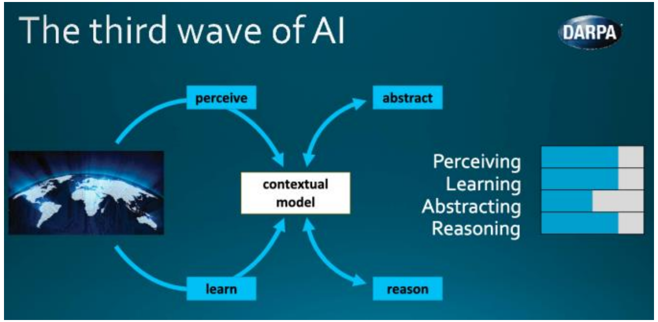
- Second Wave: Statistical Learning

  - e.g., machine learning

- Third Wave: Contextual Adaptation

  - e.g., contextual understanding & common sense reasoning





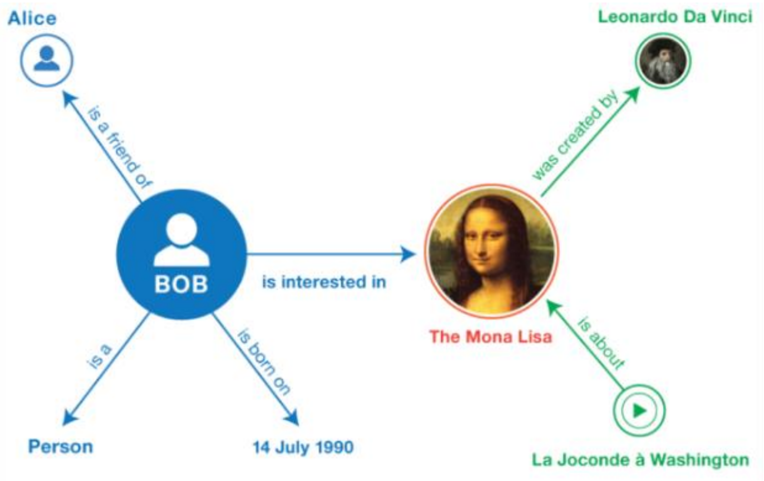
Launchbury, J., *A DARPA Perspective on Artificial Intelligence*. DARPA. Available at: https://www.darpa.mil/attachments/AIFull.pdf [Accessed 6 December 2024].

1. Introduction

- Knowledge graphs (KGs) is a graph that gather and convey the real-world knowledge where,

  - nodes represent the real-world entities of interest (e.g., person, publication) and

  - edge represent the relationships (e.g., lives in, works at, is author of) between the entities.



Source: https://towardsdatascience.com/a-guide-to-the-knowledge-graphs-bfb5c40272f1

1. Introduction

- KGs is being used are now used widely across sectors by organizations of all sizes.

  - Google Search: When you perform a Google search, you're actively leveraging knowledge graphs, even if you may not realize it[1].

  - Other companies such as Amazon and Netflix uses KGs for product or service recommendations.

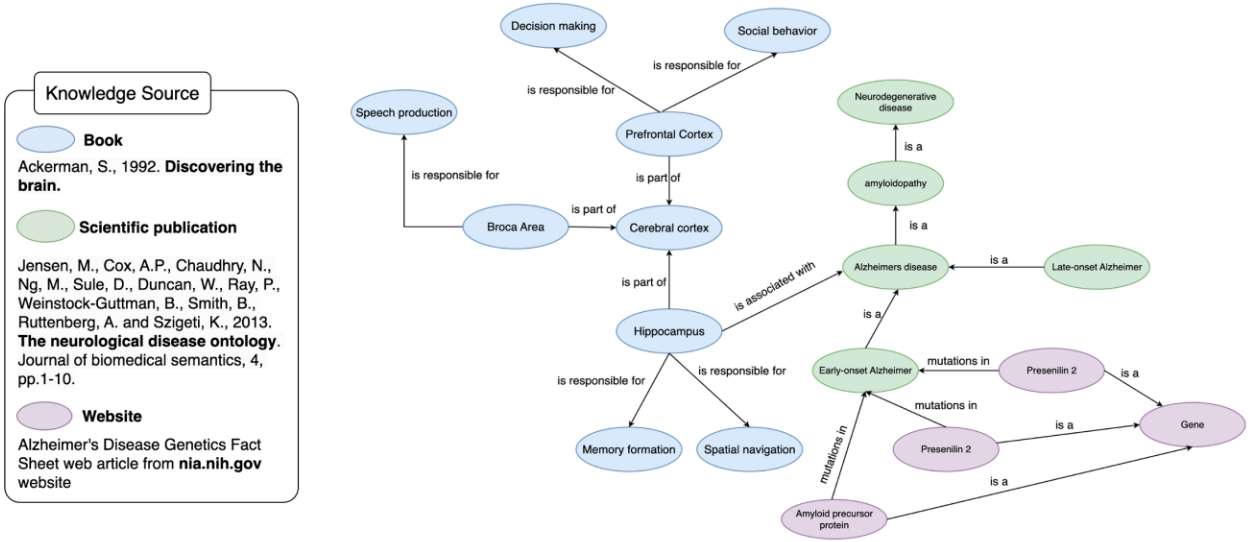
1. https://blog.google/products/search/introducing-knowledge-graph-things-not

Source: https://www.mfg-outlook.com/healthcare-manufacturing (healthcare image) and https://manufacturing-today.com/news/does-30m-boost-for-smart-manufacturing (manufacturing image)

CAIR-Nepal
Shaping the Future with AI

2. Why use Knowledge Graphs?

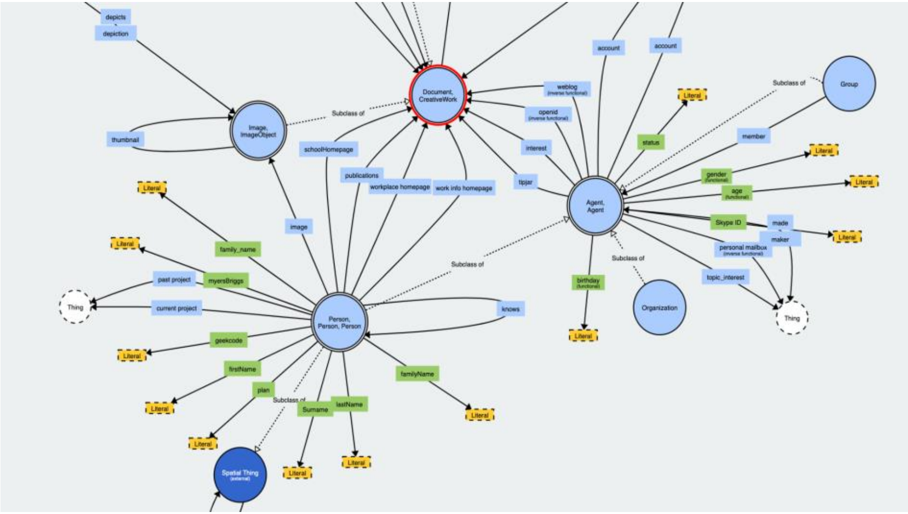
- KG connect isolated silos of knowledge.



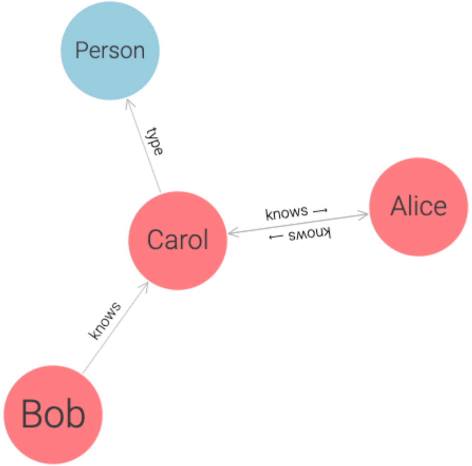
Chhetri, T.R., Jarecka, D., Trivedi, P., Amin, J., Baker, P., Dehghani, N., Bhandiwad, A., Smith, K., Ray, P., Bishwakarma, P., Ng, L., & Ghosh, S. (2024) BrainKB: A large-scale knowledge graph infrastructure for neuroscience. INCF Assembly. Available at: http://dx.doi.org/10.13140/RG.2.2.27629.81128.

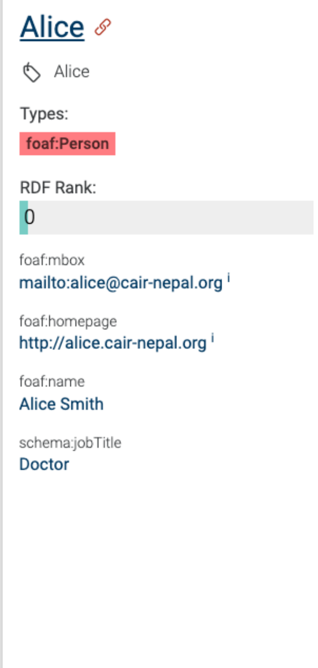
2. Why use Knowledge Graphs?

- KG provide structured representation.



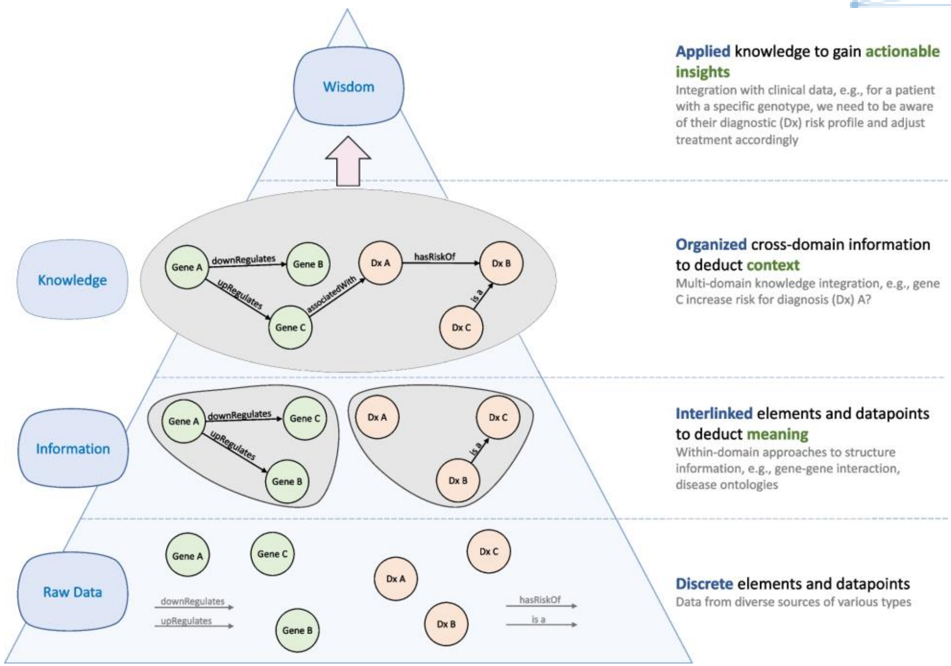
Source: https://service.tib.eu/webvowl/#foaf





2. Why use Knowledge Graphs?

- KG can transform raw data into structured knowledge, enabling actionable insights that drive informed decisions, i.e., provide wisdom.



Source: Hänsel, K., Dudgeon, S.N., Cheung, K.H., Durant, T.J. and Schulz, W.L., 2023. From data to wisdom: biomedical knowledge graphs for real-world data insights. Journal of Medical Systems, 47(1), p.65.

2. Why use Knowledge Graphs?

- **Data** refers to raw, unprocessed facts, figures which,

  - does not have (or lacks) context

  - can be both in the numerical (or quantitative) and qualitative (or descriptive).

9.00	7.00	7.00
8.00	9.00	9.00
7.00	3.00	9.00
8.00	1.00	8.00
9.00	3.00	1.00
[illegible]	[illegible]	[illegible]

```
[-70, -68, -72, -75, -60] (millivolts)
[0.003, 0.012, 0.021, 0.030] (seconds)
```

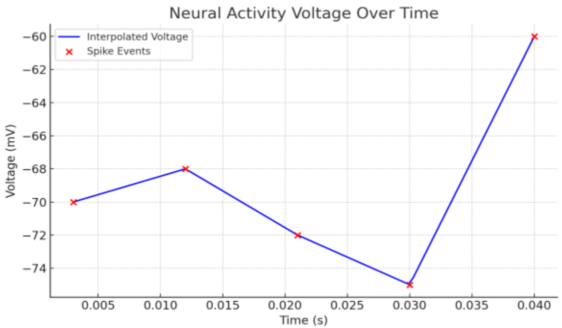
2. Why use Knowledge Graphs?

- **Information** refers to data that has context and usually can answer WH (e.g., *what*, *how*) questions and is,

  - processed and organized so that its helpful to the users.

Location	Temperature (°C)
New York	15.2
Los Angeles	22.5
Chicago	10.8
Miami	28.3
Seattle	12.7

Change in temperature by location.

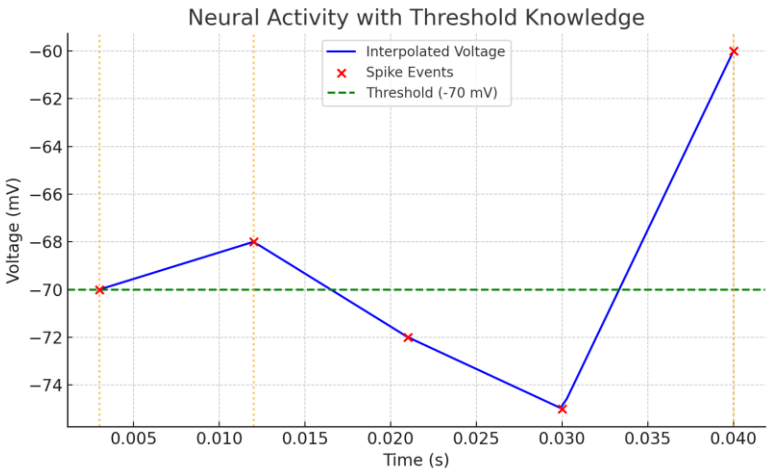


Electrode detected Changes in voltage over time.

2. Why use Knowledge Graphs?

- **Knowledge** is the synthesis of information, encompassing an understanding of facts, patterns, insights, and context.

  - It represents the comprehension achieved through experience or learning.



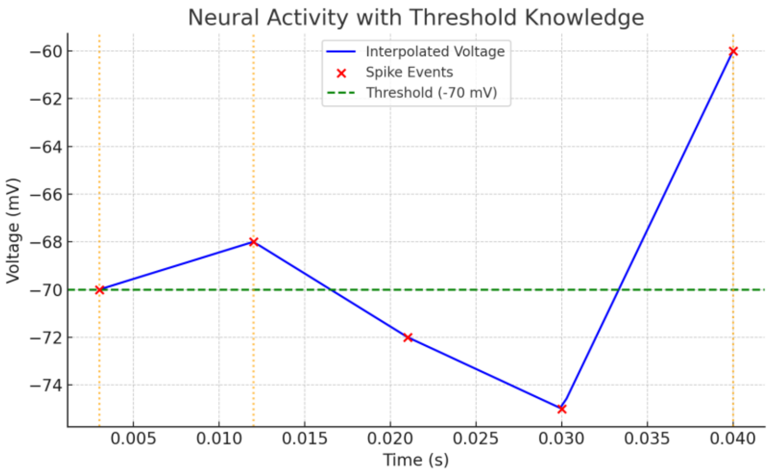
Neuron spiked when the voltage exceeded a threshold (e.g., -70 mV), indicating neural activity.

CAIR-Nepal
Shaping the Future with AI

2. Why use Knowledge Graphs?

- **Knowledge** is the synthesis of information, encompassing an understanding of facts, patterns, insights, and context.

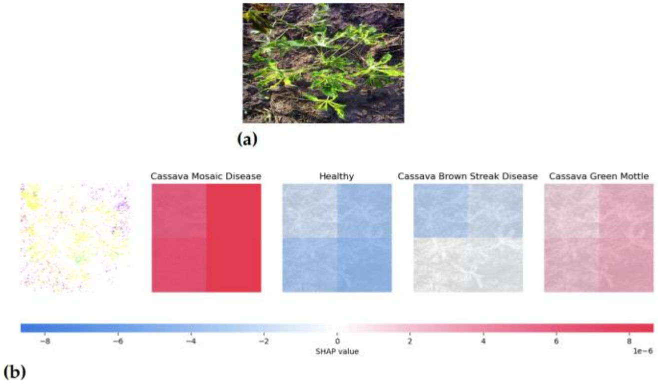
  - It represents the comprehension achieved through experience or learning.

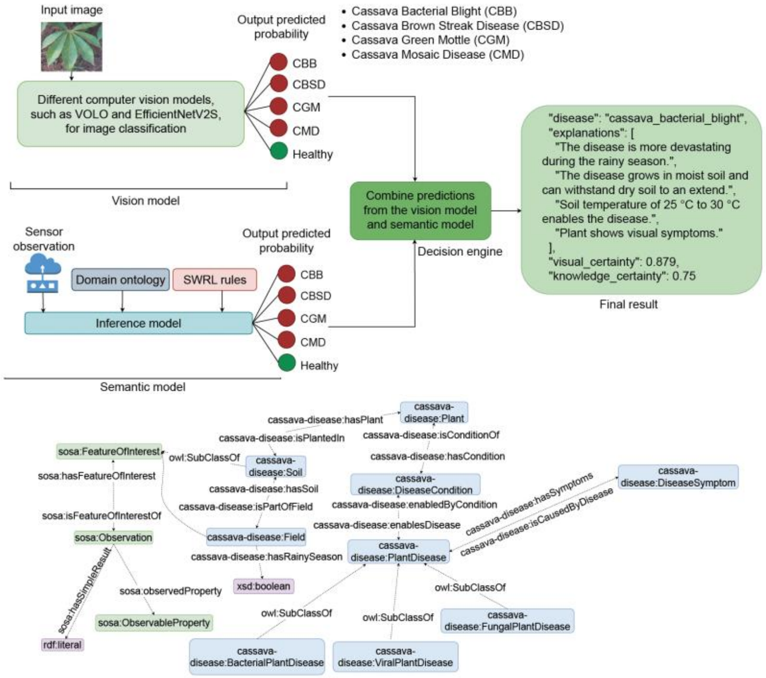


Neuron spiked when the voltage exceeded a threshold (e.g., -70 mV), indicating neural activity.

2. Why use Knowledge Graphs?

- KGs provide the enriched contextual information modern statistical learning lacks (or are limited with).





Chhetri, T.R., Hohenegger, A., Fensel, A., Kasali, M.A. and Adekunle, A.A., 2023. Towards improving prediction accuracy and user-level explainability using deep learning and knowledge graphs: A study on cassava disease. Expert Systems with Applications, 233, p.120955.

CAIR-Nepal
Shaping the Future with AI

3.

Innovations Enabled by Knowledge Graphs

3. Innovations Enabled by Knowledge Graphs

- Some of the key areas innovations that KG can drives are:

  - **Data integration and interoperability** – connect cross sectorial (and domain) data seamless enabling new solutions.

  - **Enhanced Intelligence & Improved decision making** – provide rich contextual information thereby improving intelligence and the quality of informed decision.

  - **Dynamic systems and digital twins** – real time monitoring and build the digital representation of real-world to enable predictive modelling and simulations.

  - **Knowledge discovery** – reveal new connections enabling applications such as drug discovery.

4.

Case Studies

Interoperability and Enhanced Intelligence

4. Case Studies - Interoperability and Enhanced Intelligence

- Focuses on interoperability and enhanced intelligence at the edge to enable (near) real-time intelligence.

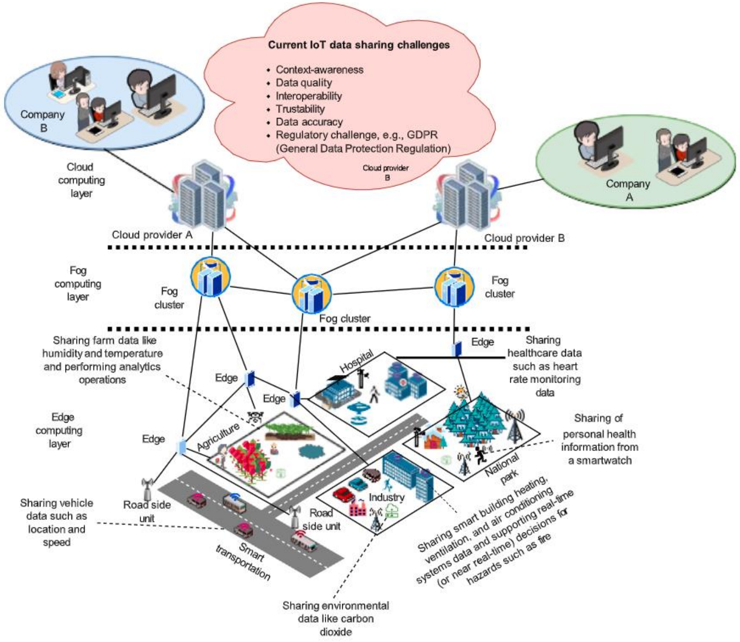


Fig. 1. IoT data sharing environment with an edge/fog scenario and the associated challenges.

Chhetri, T.R., Dehury, C.K., Varghese, B., Fensel, A., Srirama, S.N. and DeLong, R.J., 2024. Enabling privacy-aware interoperable and quality IoT data sharing with context. Future Generation Computer Systems, 157, pp.164-179.

4. Case Studies - Interoperability and Enhanced Intelligence

- Overview of the current state of the art.

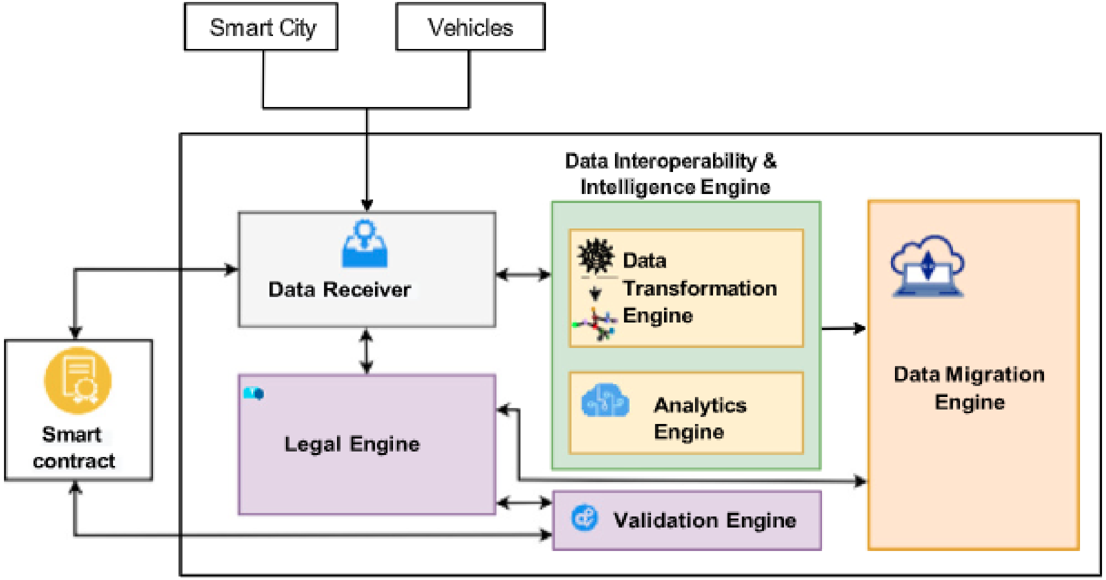
Comparison with state-of-the-art.

Study	Privacy	Interoperability	Data quality	Data veracity	Trust metric	Analytics	Edge/Fog	Performance evaluation
Rubí et al. (2021) [24]	×	✓	×	×	×	✓	×/✓	✓
Loukil et al. (2020) [25]	✓	×	×	×	×	×	×/×	✓
Strassner et al. (2016) [27]	×	✓	×	×	×	×	×/✓	×
Reda et al. (2022) [28]	×	✓	✓	×	×	×	×/×	×
Zappatore et al. (2023) [29]	×	✓	×	×	×	✓	✓/✓	×
Makhdoom et al. (2020) [30]	✓	×	×	×	×	×	×/×	✓
Abdullah et al. (2022) [31]	✓	×	×	×	×	✓	×/×	✓
Bai et al. (2022) [33]	✓	×	×	×	×	×	×/×	✓
Alamri et al. (2021) [34]	✓	✓	×	×	×	×	×/×	×
Tsiouris et al. (2020) [35]	×	✓	×	×	×	×	✓/×	✓
Halim et al. (2022) [36]	×	✓	✓	×	×	×	✓/×	✓
Poojara et al. (2022) [37]	×	×	×	×	×	✓	✓/✓	✓
Our study	✓	✓	✓	✓	✓	✓	✓/✓	✓

Chhetri, T.R., Dehury, C.K., Varghese, B., Fensel, A., Srirama, S.N. and DeLong, R.J., 2024. Enabling privacy-aware interoperable and quality IoT data sharing with context. Future Generation Computer Systems, 157, pp.164-179.

4. Case Studies - Interoperability and Enhanced Intelligence

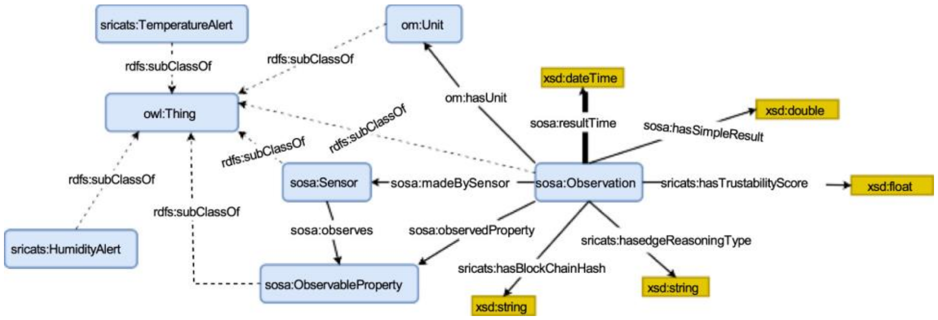
- Proposed architecture.



Chhetri, T.R., Dehury, C.K., Varghese, B., Fensel, A., Srirama, S.N. and DeLong, R.J., 2024. Enabling privacy-aware interoperable and quality IoT data sharing with context. Future Generation Computer Systems, 157, pp.164-179.

4. Case Studies - Interoperability and Enhanced Intelligence

- Ontology used in the study and the SWRL rule to enable intelligence.



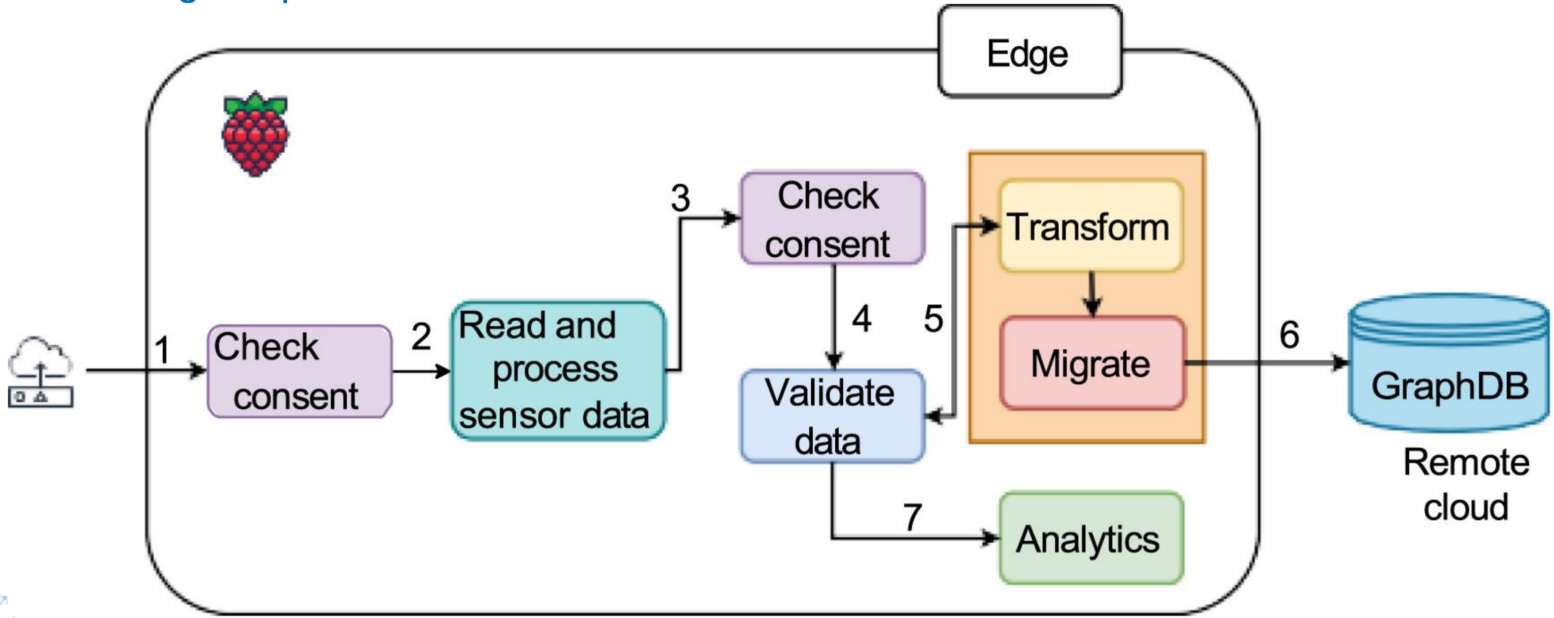

```
Observation(?observation),
hasSimpleResult(?observation, ?result),
hasedgeReasoningType(?observation, ?reasoningType),
containsIgnoreCase(?reasoningType, "temperature"),
greaterThanOrEqual(?result, 75.0),
-> TemperatureAlert(?observation)

Observation(?observation),
hasedgeReasoningType(?observation, ?reasoningType),
containsIgnoreCase(?reasoningType, "humidity"),
hasSimpleResult(?observation, ?result),
greaterThanOrEqual(?result, 65.0),
-> HumidityAlert(?observation)
```

Chhetri, T.R., Dehury, C.K., Varghese, B., Fensel, A., Srirama, S.N. and DeLong, R.J., 2024. Enabling privacy-aware interoperable and quality IoT data sharing with context. Future Generation Computer Systems, 157, pp.164-179.

CAIR-Nepal
Shaping the Future with AI

4. Case Studies - Interoperability and Enhanced Intelligence

- Edge experiment scenario.

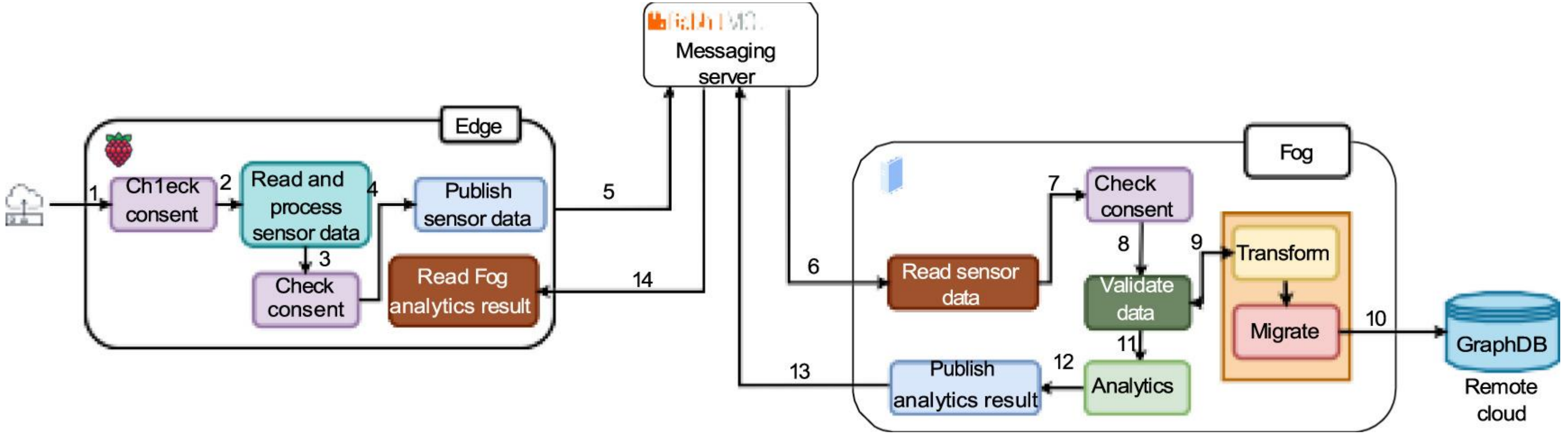


Chhetri, T.R., Dehury, C.K., Varghese, B., Fensel, A., Srirama, S.N. and DeLong, R.J., 2024. Enabling privacy-aware interoperable and quality IoT data sharing with context. Future Generation Computer Systems, 157, pp.164-179.

CAIR-Nepal
Shaping the Future with AI

4. Case Studies - Interoperability and Enhanced Intelligence

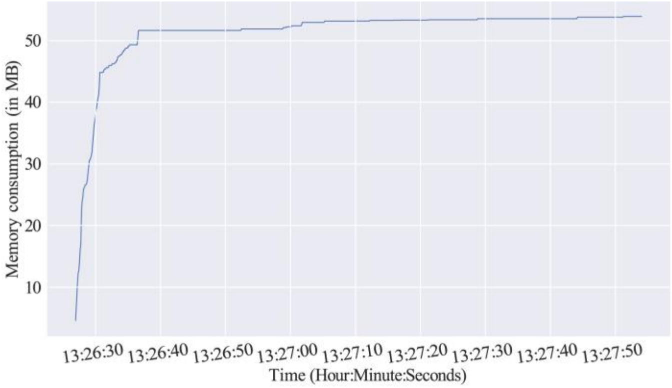
- Fog–edge experiment scenario.



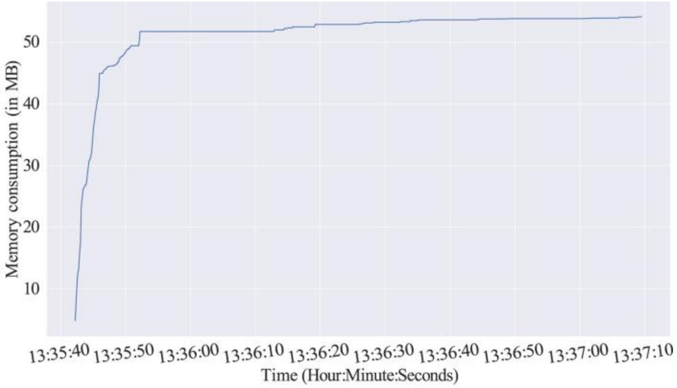
Chhetri, T.R., Dehury, C.K., Varghese, B., Fensel, A., Srirama, S.N. and DeLong, R.J., 2024. Enabling privacy-aware interoperable and quality IoT data sharing with context. Future Generation Computer Systems, 157, pp.164-179.

4. Case Studies - Interoperability and Enhanced Intelligence

- Evaluated performance to see if the proposed solution is feasible in resource constrained devices.



(a)



(b)

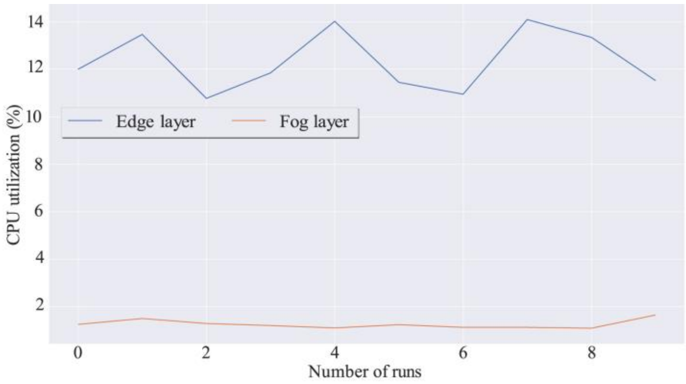
(a) Data transformation and migration operation at edge. (b) Analytics operation.

Chhetri, T.R., Dehury, C.K., Varghese, B., Fensel, A., Srirama, S.N. and DeLong, R.J., 2024. Enabling privacy-aware interoperable and quality IoT data sharing with context. Future Generation Computer Systems, 157, pp.164-179.

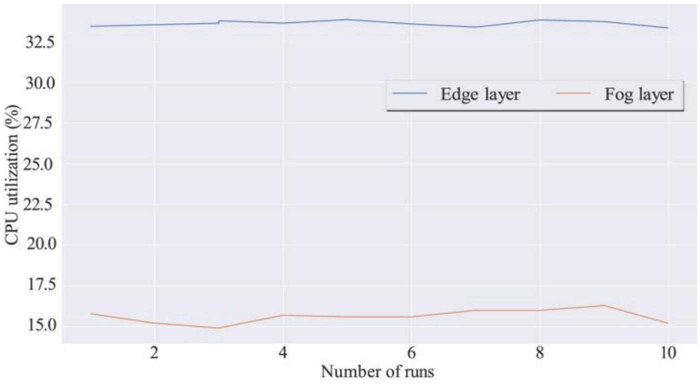
CAIR-Nepal
Shaping the Future with AI

4. Case Studies - Interoperability and Enhanced Intelligence

- Evaluated performance to see if the proposed solution is feasible in resource constrained devices.



(a)



(b)

(a) Data transformation and migration operation. (b) Analytics operation.

Chhetri, T.R., Dehury, C.K., Varghese, B., Fensel, A., Srirama, S.N. and DeLong, R.J., 2024. Enabling privacy-aware interoperable and quality IoT data sharing with context. Future Generation Computer Systems, 157, pp.164-179.

4. Case Studies - Interoperability and Enhanced Intelligence

- Interoperability was evaluated checking if the raw IoT data is being correctly transformed as per the ontology.

- Analytics operation was evaluated by checking if the alert was correctly triggered based on the value of the humidity and temperature.

Chhetri, T.R., Dehury, C.K., Varghese, B., Fensel, A., Srirama, S.N. and DeLong, R.J., 2024. Enabling privacy-aware interoperable and quality IoT data sharing with context. Future Generation Computer Systems, 157, pp.164-179.

4.

Case Studies

Knowledge Discovery

4. Case Studies – Knowledge Discovery

- 15 new antibiotic-resistant *Escherichia coli* genes discovered.

  - 6 genes were identified as novel antibiotic resistance genes for any bacteria.

- 5 homologous genes in *Salmonella enterica* that confer antibiotic resistance, validated experimentally.

- Proposed ***Knowledge Integration and Decision Support (KIDS) framework*** based on KGs.

Youn, J., Rai, N. and Tagkopoulos, I., 2022. Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes. Nature communications, 13(1), p.2360.

4. Case Studies – Knowledge Discovery

- KGs construction:

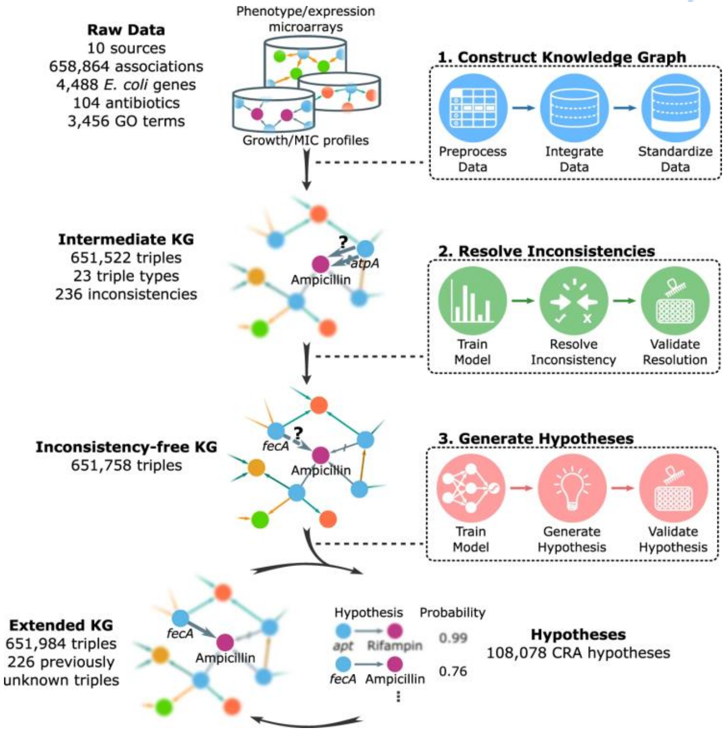
  - 10 different data sources

  - Data is encoded in RDF triples

- Inconsistency detection:

  - 9 manual rules were added.

  - Detected inconsistencies were resolved using AverageLog algorithm



Youn, J., Rai, N. and Tagkopoulos, I., 2022. Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes. Nature communications, 13(1), p.2360.

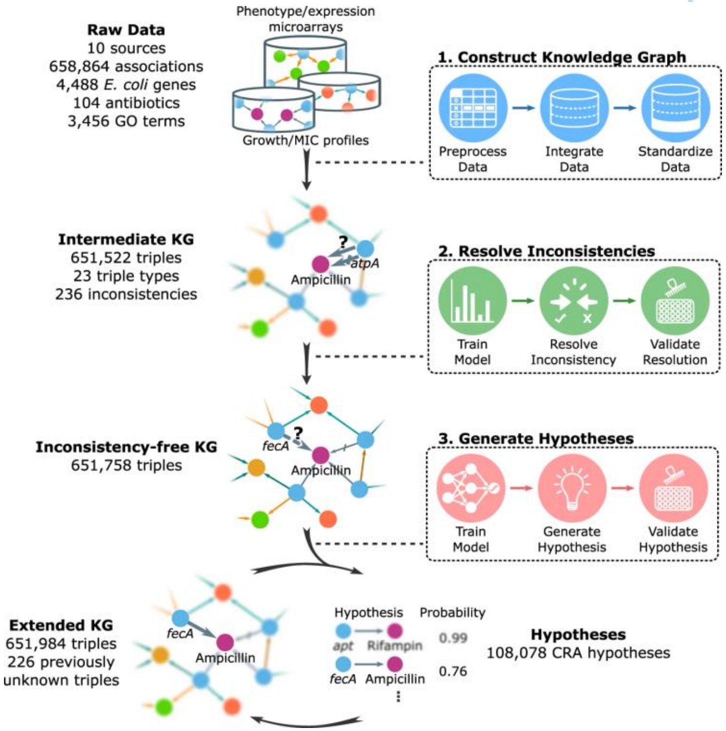
4. Case Studies – Knowledge Discovery

- Inconsistency Resolution:

  - AverageLog algorithm to resolve conflicts by iteratively calculating

  - **Belief Scores** for triples, initially all triples start with belief score of 0.5.

  - **Trustworthiness** Scores for data sources.



Youn, J., Rai, N. and Tagkopoulos, I., 2022. Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes. Nature communications, 13(1), p.2360.

4. Case Studies – Knowledge Discovery

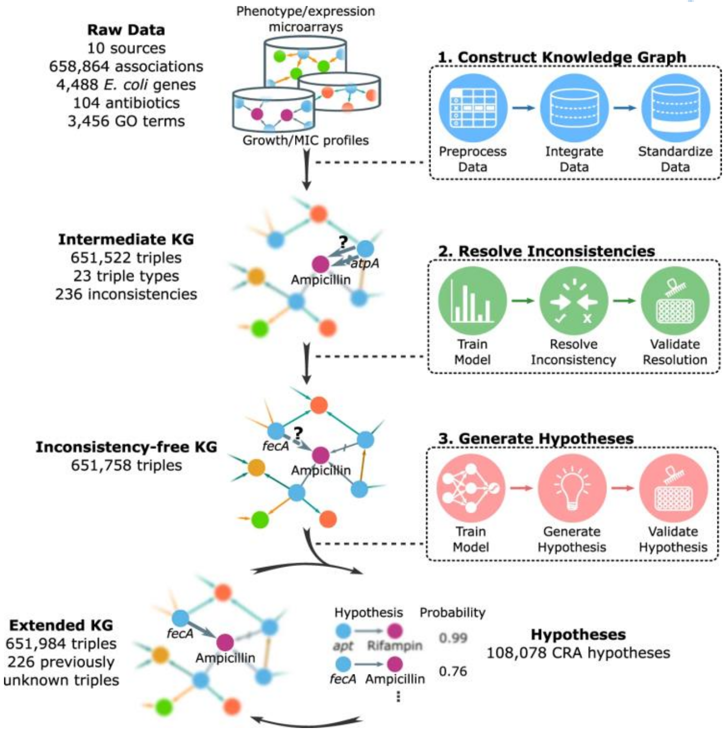
- **Belief Scores**:

  High score reflects confidence that specific triple is true.

Belief Score of triple t at iteration i:

$$B_i(t) = \sum_{s \in S_t} R_i(s)$$

- S_t: Set of sources that provide triple t.
- $R_i(s)$: Trustworthiness Score of source s at the current iteration.



Youn, J., Rai, N. and Tagkopoulos, I., 2022. Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes. Nature communications, 13(1), p.2360.

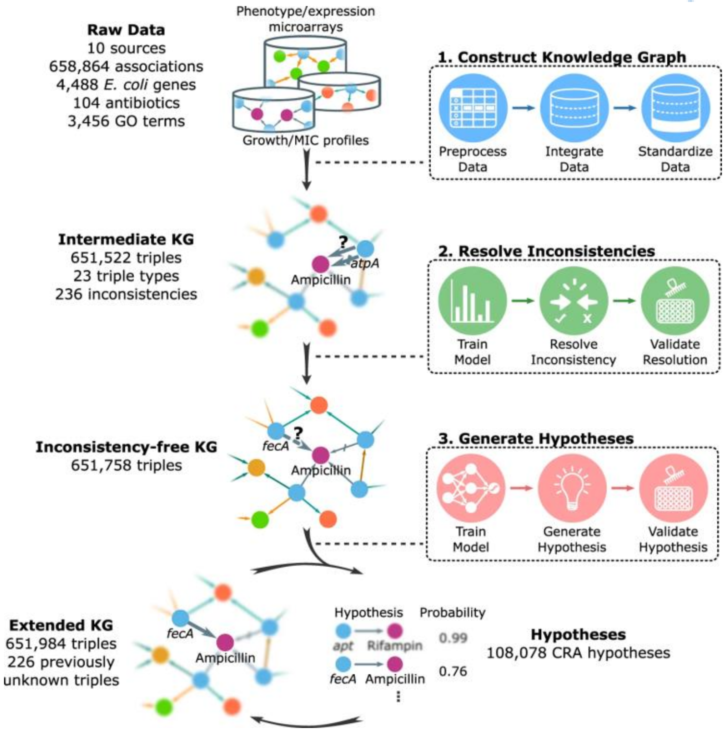
4. Case Studies – Knowledge Discovery

- **Trustworthiness Scores**:

Trustworthiness of source s at iteration i:

$$R_i(s) = \log(|T_s|) \cdot \frac{\sum_{t \in T_s} B_{i-1}(t)}{|T_s|}$$

- $|T_s|$: Number of triples provided by source s.
- $B_{i-1}(t)$: Belief Score of triple t from the previous iteration.



Intuition is that trustworthy source will have higher belief.

Youn, J., Rai, N. and Tagkopoulos, I., 2022. Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes. Nature communications, 13(1), p.2360.

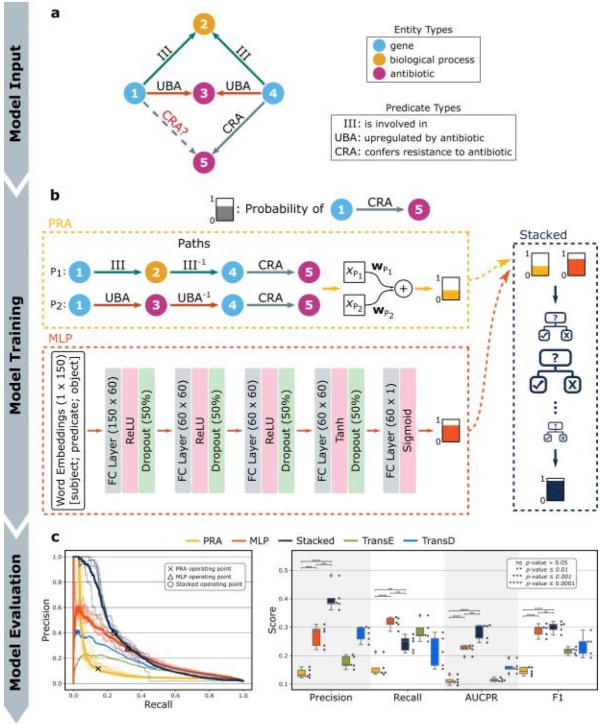
4. Case Studies – Knowledge Discovery

- **Hypothesis generation**:

  Predict the missing link (relationships), i.e., identify the potential new connections.

  - Page rank algorithm (PRA), which outputs probability score was applied to identify path between set of entities (subject-gene and object-antibiotic).

  - Multilayer Perceptron (MLP) predicts the validity of the triple.



Youn, J., Rai, N. and Tagkopoulos, I., 2022. Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes. Nature communications, 13(1), p.2360.

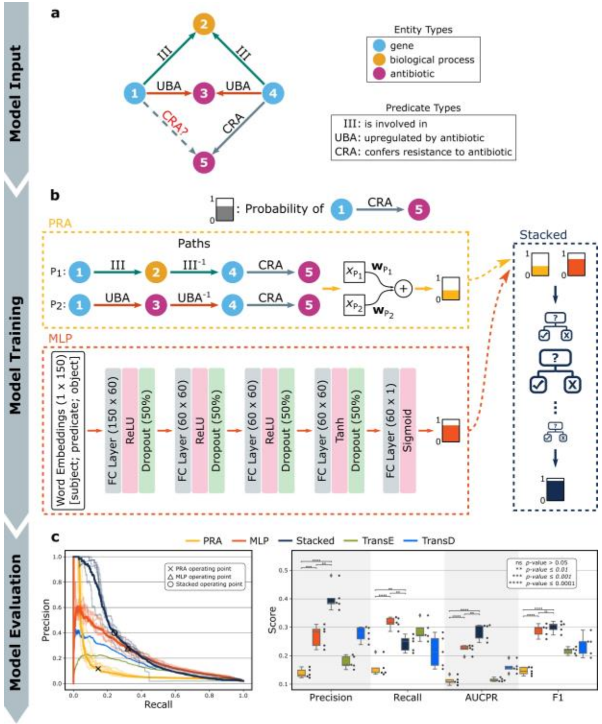
4. Case Studies – Knowledge Discovery

- **Hypothesis generation**:

  - Stacked Model (Ensemble) combines PRA and MLP using AdaBoost ensemble and provides improved prediction.

    - Features: PRA's probabilities and MLP's outputs, as well as a binary indicator for whether PRA found a valid path.

  Translation-Based Embedding Models (TransE, TransD) perform the link prediction.



Youn, J., Rai, N. and Tagkopoulos, I., 2022. Knowledge integration and decision support for accelerated discovery of antibiotic resistance genes. Nature communications, 13(1), p.2360.

CAIR-NEPAL
Shaping the Future with AI

5.

Conclusion & Future Outlook

5. Conclusion & Future outlook

- KGs have potential to enable wide array of innovations including improving predictions of large language models (LLMs) and making them trustworthy.
- KGs provide a foundation for evidence-based discovery, enabling researchers to derive actionable insights and uncover novel relationships (or discovery).
- However, despite their promise, there is a significant gap in accessible tools and technologies that non-experts or researchers from other disciplines to leverage KGs effectively for scientific discovery. This limitation hinders broader adoption and utilization of KGs in interdisciplinary research.
- To realize their full potential, there is an urgent need to develop accessible tools and technologies that bridge this gap, democratizing the use of KGs and supporting scientific breakthroughs across diverse fields.

Get in touch?

- Twitter: @tekraj_14
- Web: https://tekrajchhetri.com
- Email: tekraj.chhetri@cair-nepal.org | tekraj@mit.edu
- LinkedIn: https://www.linkedin.com/in/tekrajchhetri/

Thank you!