

# RETRIEVAL AUGMENTED GENERATION AND SCIENTIFIC KNOWLEDGE GRAPHS TO SUPPORT SCIENTIFIC HYPOTHESES GENERATION

OSSYM2024 – 10 October 2024

Oliver Bensch (DLR), Tobias Hecking (DLR), J. Nathan Kutz (University of Washington)



# Automatisation of Scientific Workflows



Hypothesis generation, planning and evaluation of experiments, literature screening, ...

nature

Explore content ▾ About the journal ▾ Publish with us ▾

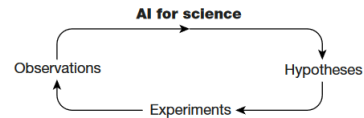
[nature](#) > [review articles](#) > article

Review | Published: 02 August 2023

## Scientific discovery in the age of artificial intelligence

[Hanchen Wang](#), [Tianfan Fu](#), [Yuangqi Du](#), [Wenhao Gao](#), [Kexin Huang](#), [Ziming Liu](#), [Payal Chandak](#), [Shengchao Liu](#), [Peter Van Katwyk](#), [Andreea Deac](#), [Anima Anandkumar](#), [Karianne Bergen](#), [Carla P. Gomes](#), [Shirley Ho](#), [Pushmeet Kohli](#), [Joan Lasenby](#), [Jure Leskovec](#), [Tie-Yan Liu](#), [Arjun Manrai](#), [Deborah Marks](#), [Bharath Ramsundar](#), [Le Song](#), [Jimeng Sun](#), [Jian Tang](#), ... [Marinka Zitnik](#) + Show authors

*Nature* **620**, 47–60 (2023) | [Cite this article](#)



- Weather forecasting
- Battery design optimization
- Magnetic control of nuclear fusion reactors
- Planning chemical synthesis pathway
- Neural solvers of differential equations
- Hydropower station location planning
- Synthetic electronic health record generation
- Rare event selection in particle collisions
- Language modelling for biomedical sequences
- High-throughput virtual screening
- Navigation in the hypothesis space
- Super-resolution 3D live-cell imaging
- Symbolic regression

nature

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [letters](#) > article

Letter | Published: 03 July 2019

## Unsupervised word embeddings capture latent knowledge from materials science literature

[Vahe Tshitoyan](#) , [John Dagdelen](#), [Leigh Weston](#), [Alexander Dunn](#), [Ziqin Rong](#), [Olga Kononova](#), [Kristin A. Persson](#), [Gerbrand Ceder](#) & [Anubhav Jain](#)

*Nature* **571**, 95–98 (2019) | [Cite this article](#)

80k Accesses | 536 Citations | 1828 Altmetric | [Metrics](#)

Brief Communication | [Open access](#) | Published: 01 May 2023

## Accelerating material design with the generative toolkit for scientific discovery

[Matteo Manica](#) , [Jannis Born](#), [Joris Cadow](#), [Dimitrios Christofidellis](#), [Ashish Dave](#), [Dean Clarke](#), [Yves Gaetan Nana Teukam](#), [Giorgio Giannone](#), [Samuel C. Hoffman](#), [Matthew Buchan](#), [Vijil Chenthamarakshan](#), [Timothy Donovan](#), [Hsiang Han Hsu](#), [Federico Zipoli](#), [Oliver Schilter](#), [Akihiro Kishimoto](#), [Lisa Hamada](#), [Inkit Padhi](#), [Karl Wehden](#), [Lauren McHugh](#), [Alexy Khrabrov](#), [Payel Das](#), [Seiji Takeda](#) & [John R. Smith](#)

*npj Computational Materials* **9**, Article number: 69 (2023) | [Cite this article](#)

## Program Synthesis with Large Language Models

Jacob Austin\*

Augustus Odena\*

Maxwell Nye<sup>†</sup> Maarten Bosma Henryk Michalewski David Dohan Ellen Jiang Carrie Cai

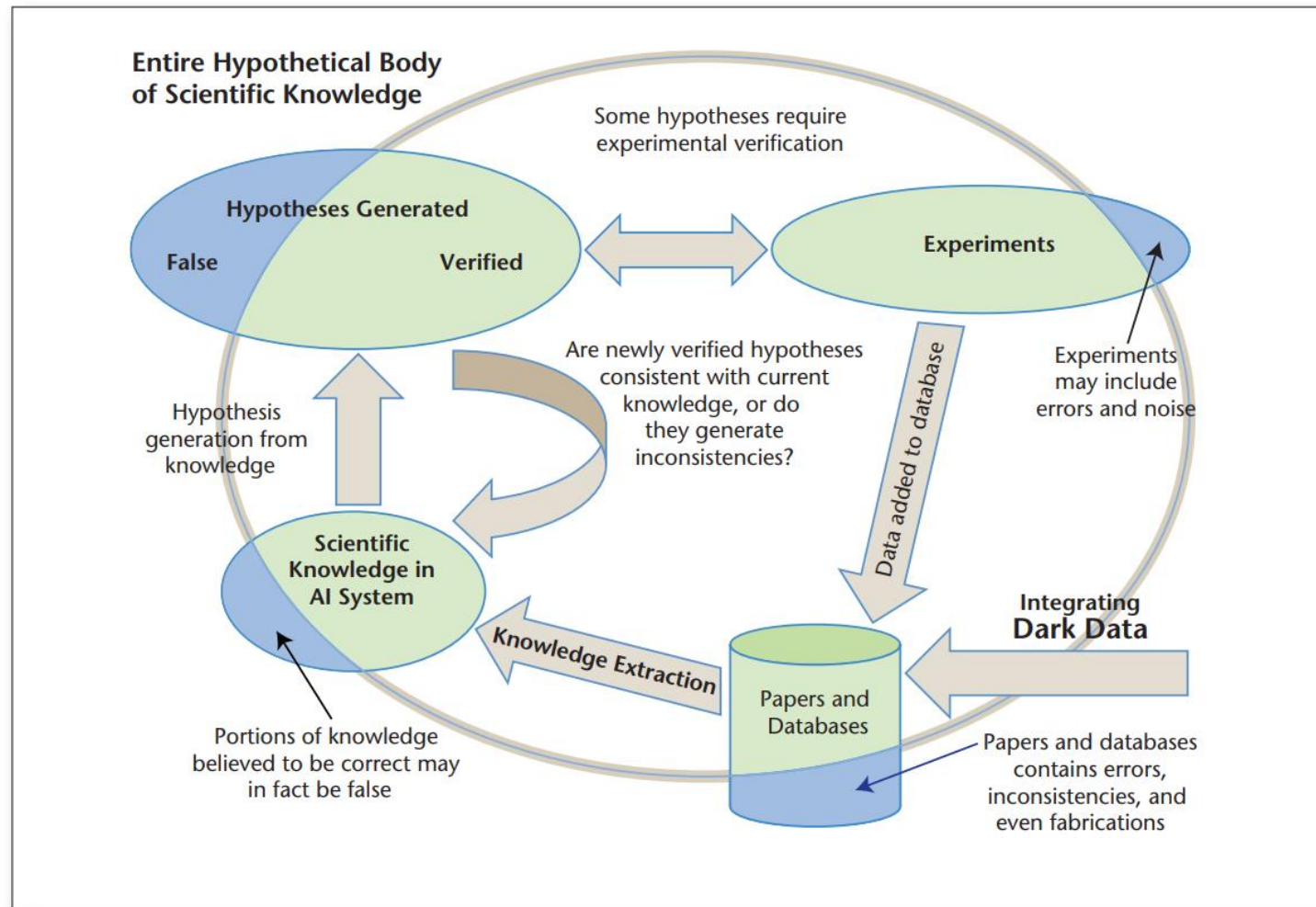
Michael Terry

Quoc Le

Charles Sutton

Google Research  
\* denotes equal contribution  
jaaustin@google.com, augustusodena@google.com

# Automated Scientific Discovery



Kitano, H. (2016). Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery. In AI Magazine (Vol. 37, Issue 1, pp. 39–49). Wiley. <https://doi.org/10.1609/aimag.v37i1.2642>



# From search to discovery



**Classic Search:** Seeking to fulfil a specific information need

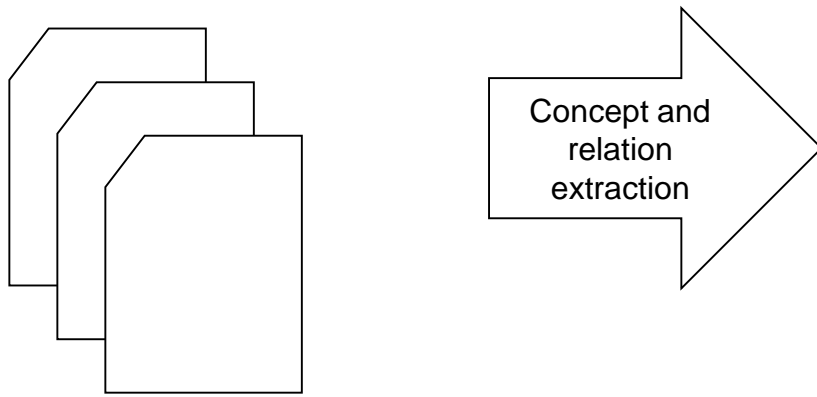


**Discovery:** Find something new and unexpected

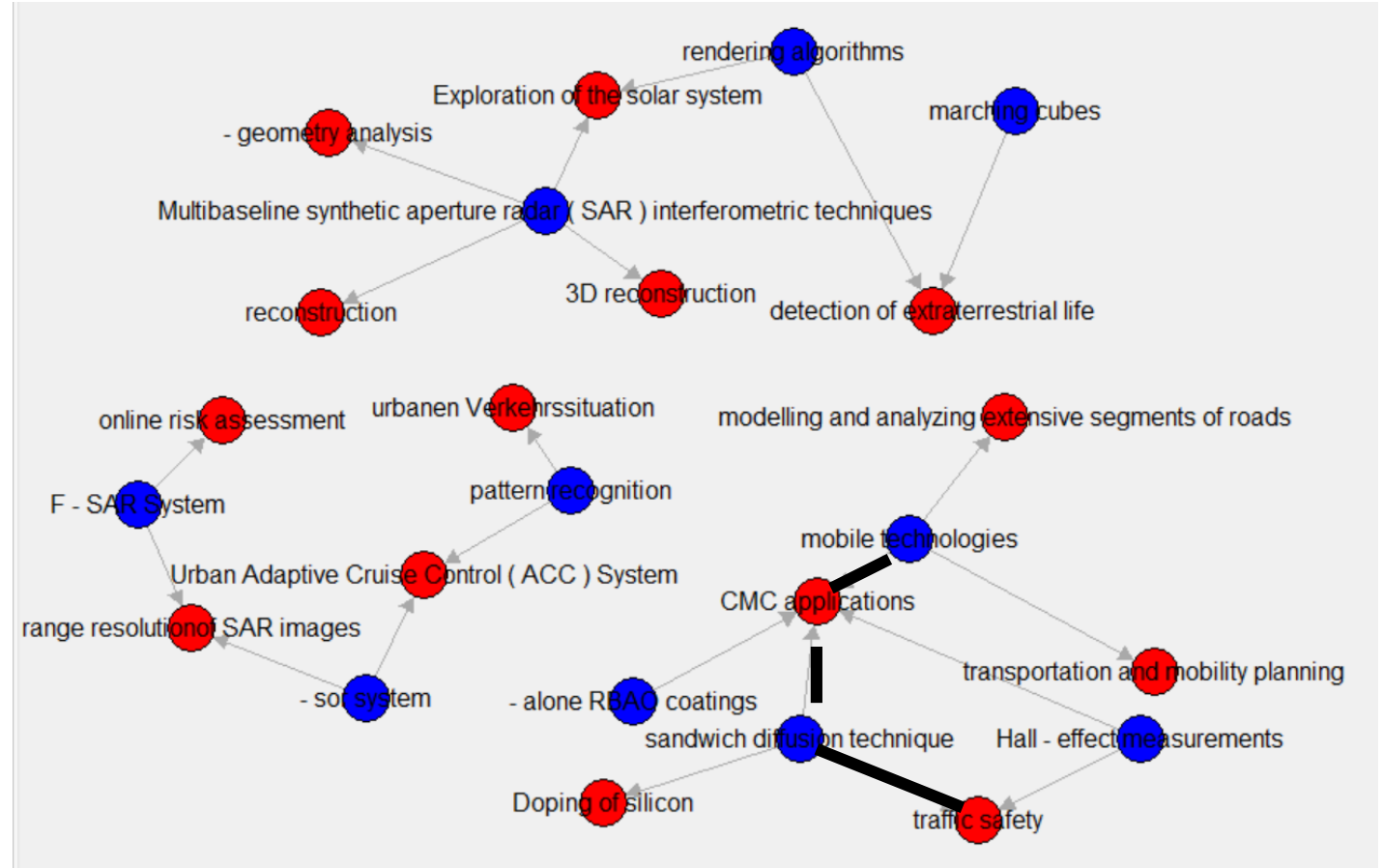


# Graphs as a natural representation of a modifyable search space

## Graph traversal (browsing)

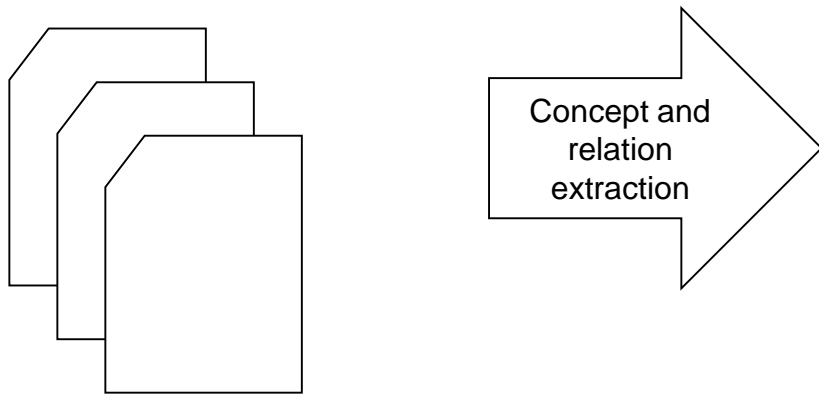


Knowledge artifacts, e.g. publications

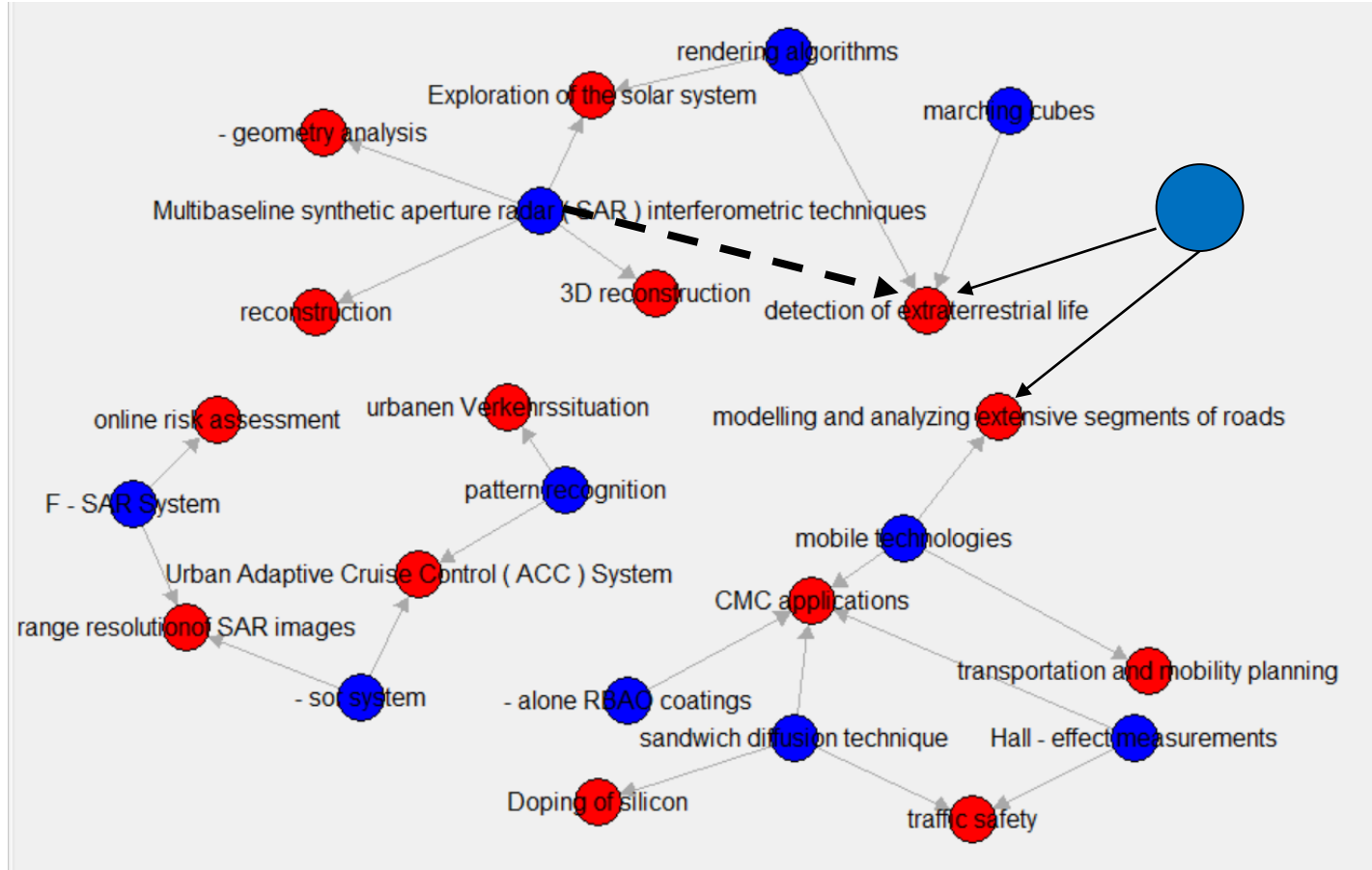


# Graphs as a natural representation of a modifyable search space

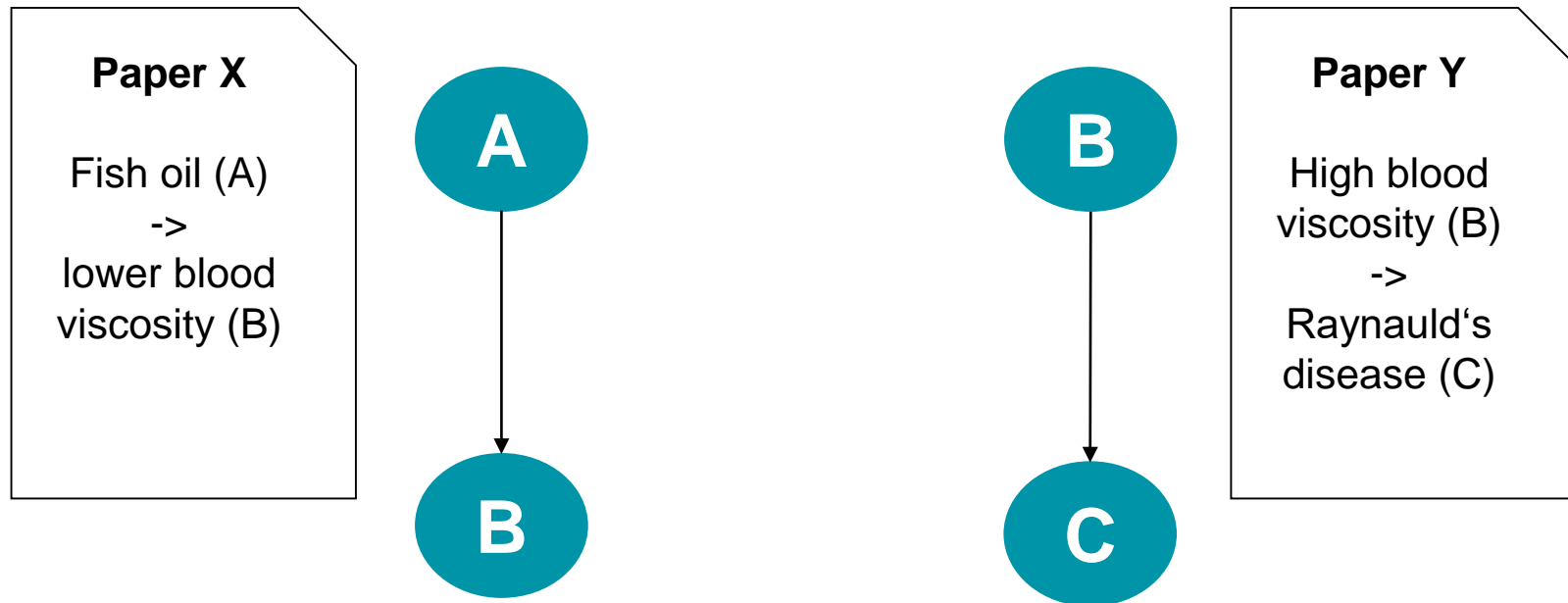
## Link prediction



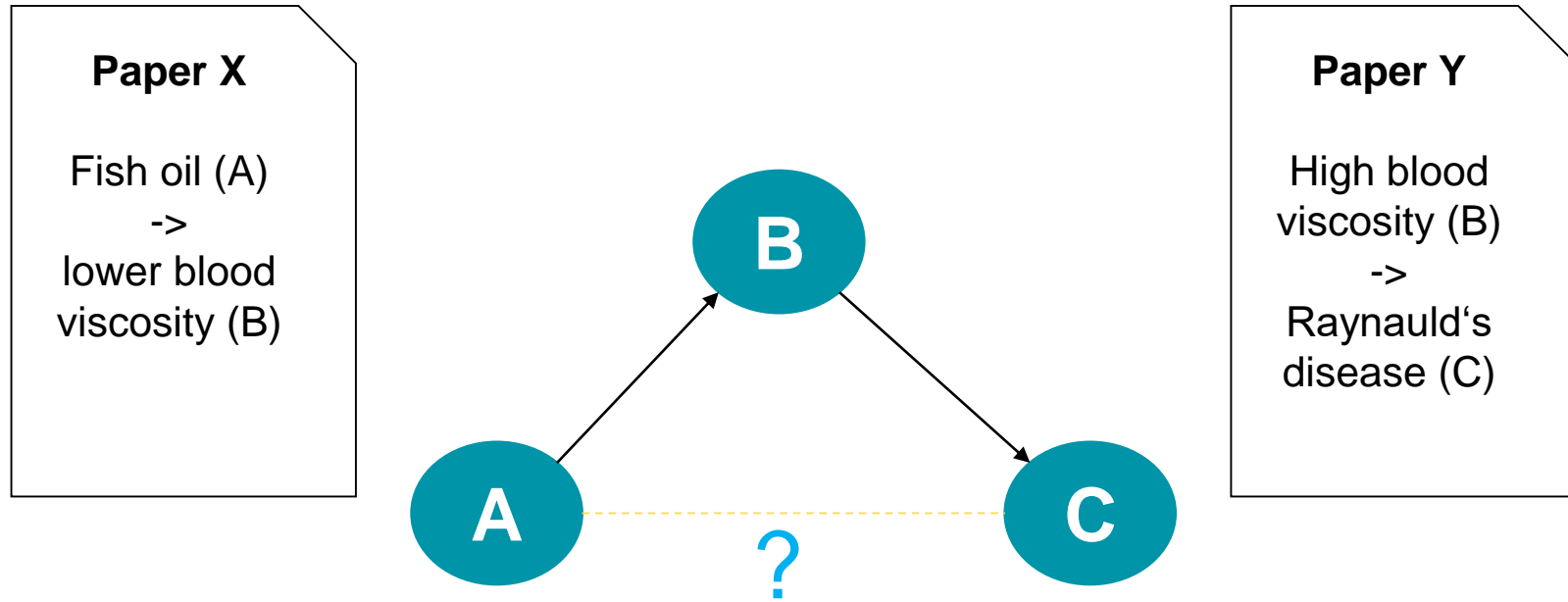
Knowledge artifacts, e.g. publications



# Literature-Based Discovery (LBD) for Hypothesis Generation

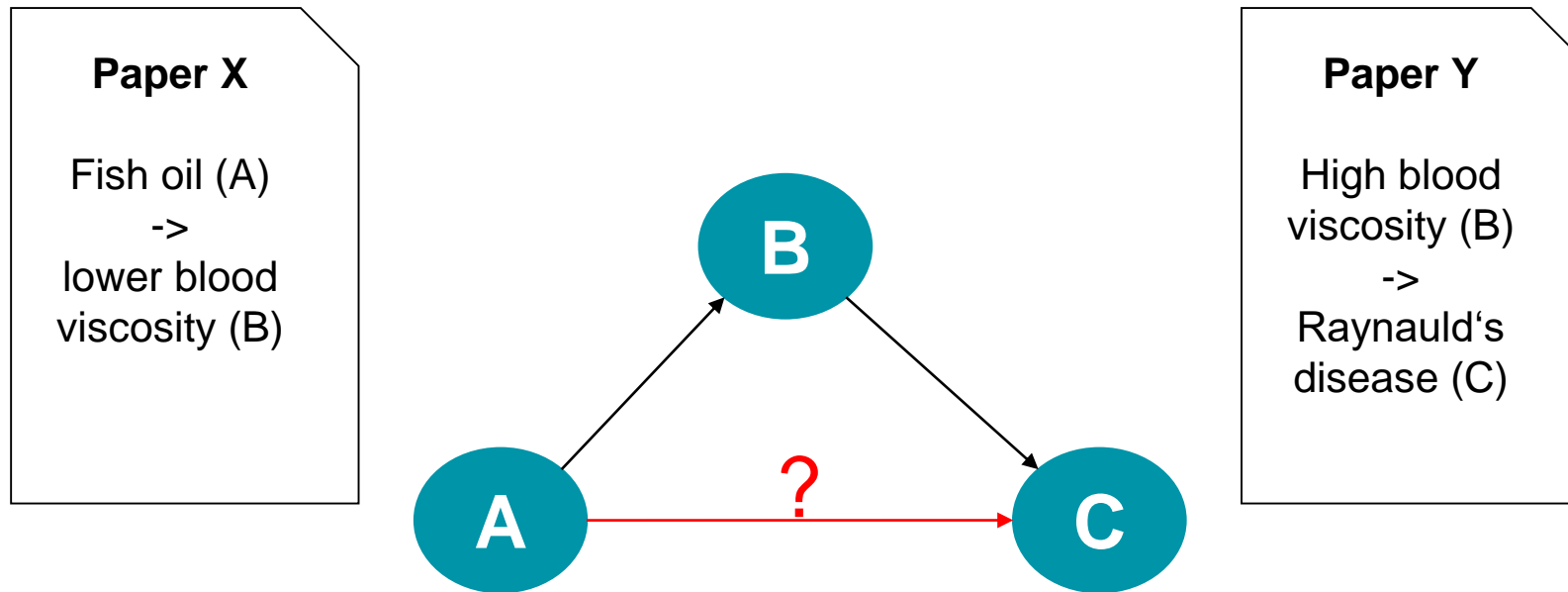


# Literature-Based Discovery (LBD) for Hypothesis Generation





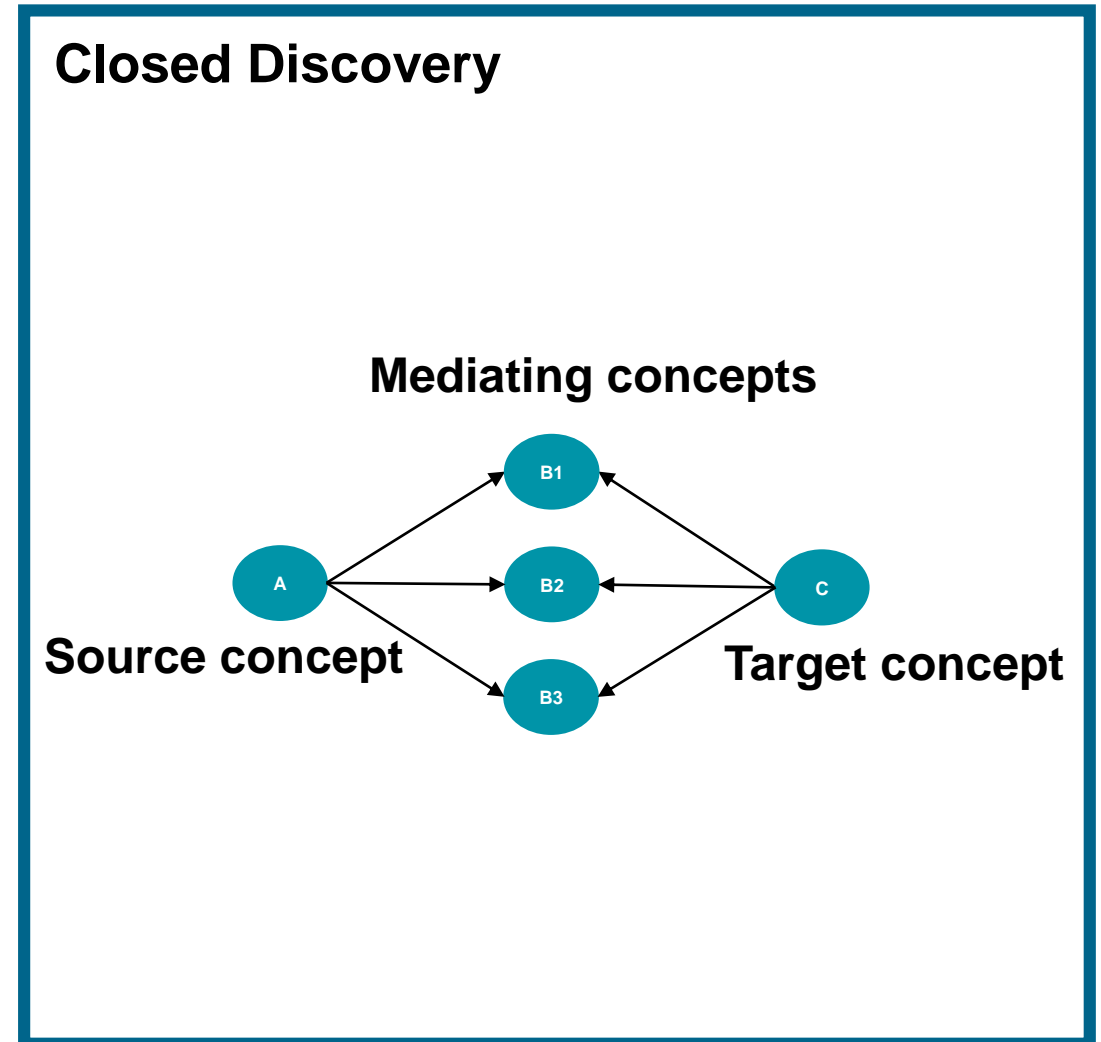
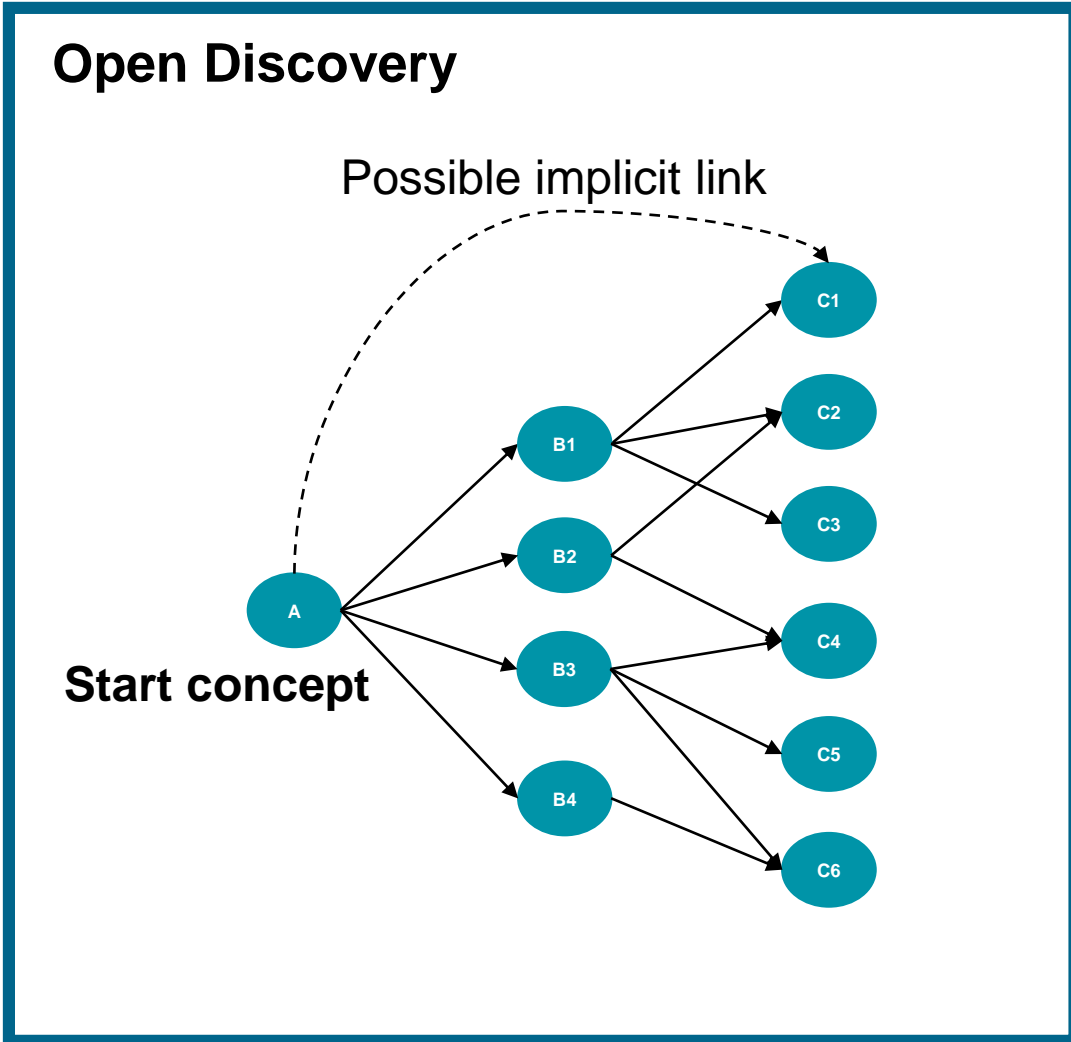
# Literature-Based Discovery (LBD) for Hypothesis Generation



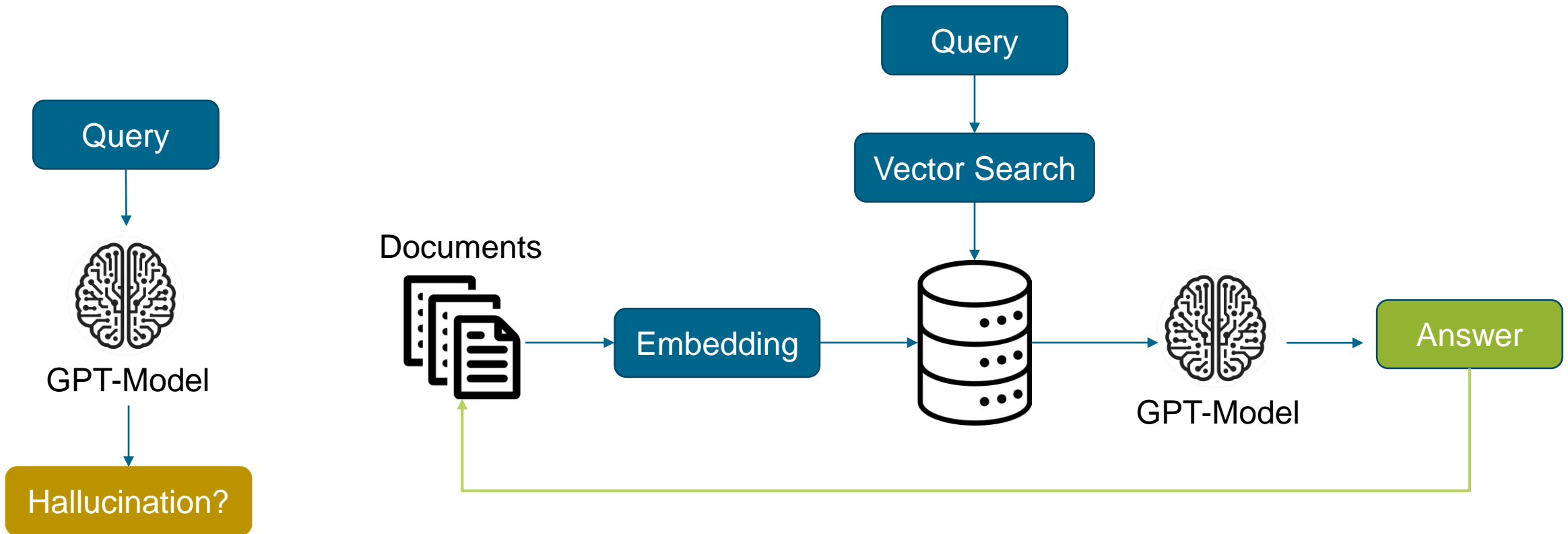
Implicit knowledge: Fish oil has an effect on Raynaud's disease

DiGiacomo RA, Kremer JM, Shah DM. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. *Am J Med.* 1989 Feb;86(2):158-64. doi: 10.1016/0002-9343(89)90261-1. PMID: 2536517.

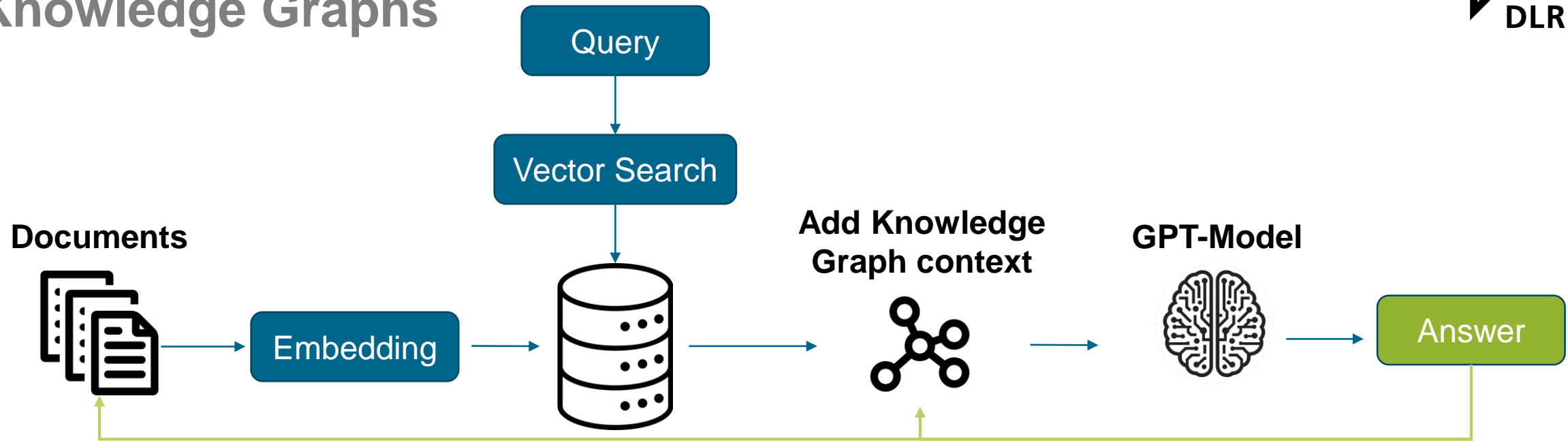
# Literature-Based Discovery (LBD) for Hypothesis Generation



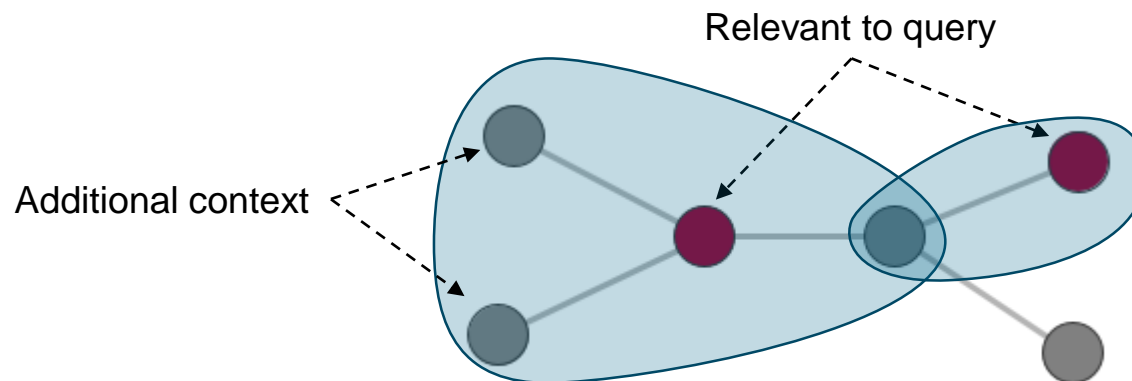
# Retrieval-Augmented Generation (RAG)



# Retrieval Augmented Generation (RAG) with Knowledge Graphs



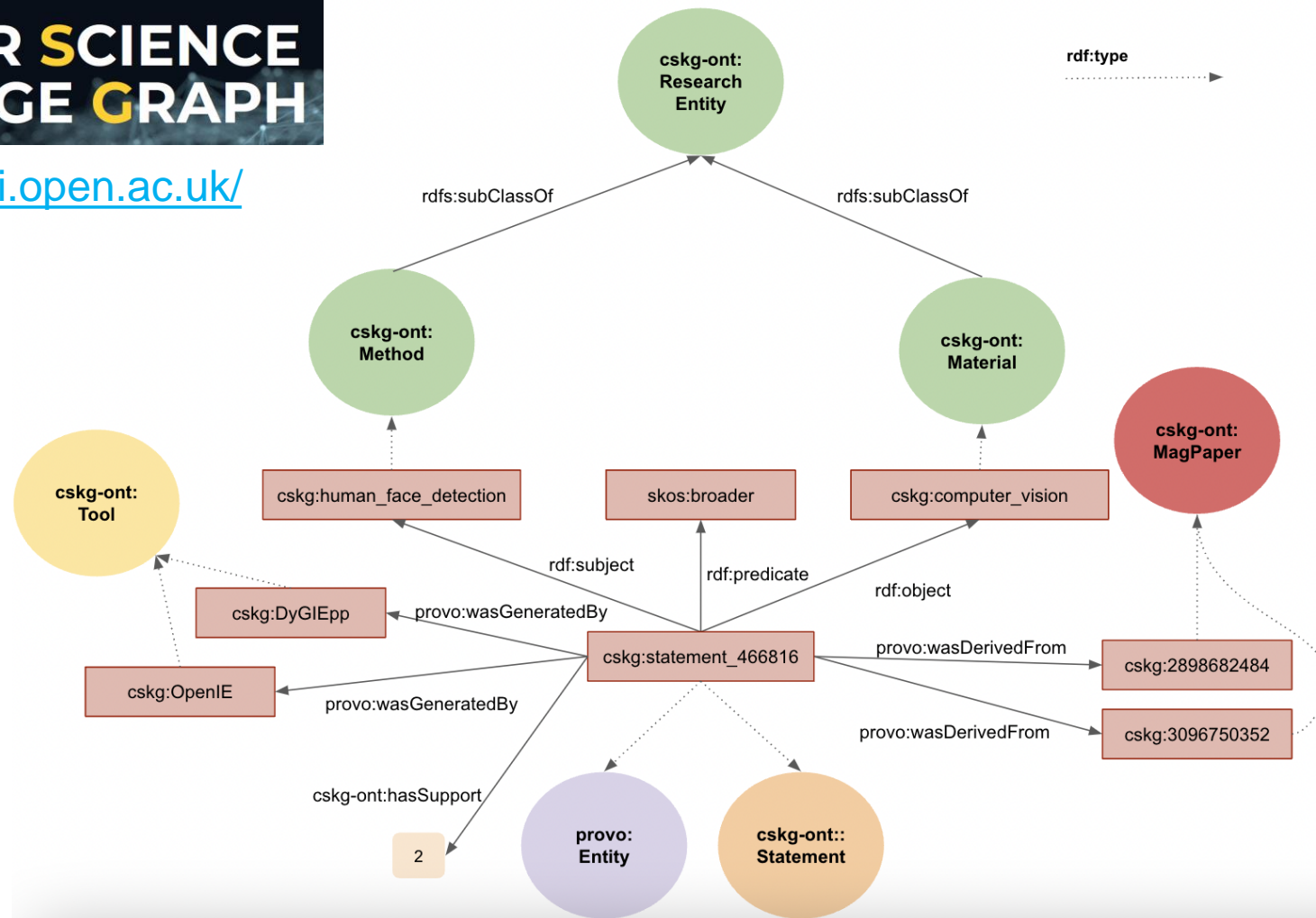
Graphs allow to broaden the context combining retrieval and traversal.



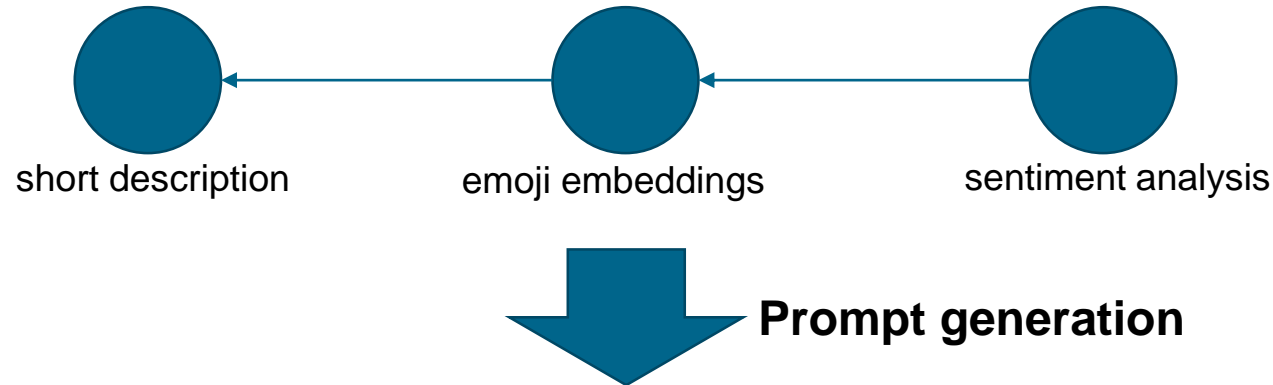


## COMPUTER SCIENCE KNOWLEDGE GRAPH

<https://scholkg.kmi.open.ac.uk/>



# Hypothesis Generation Using Large Language Models



You are given a set of information from a graph that describes the relationship between concepts extracted from the computer science domain including their predicates connecting these. You analyze these logically through reasoning.

**Context**

Subgraph:

emoji\_embeddings --usesMaterial--> short\_description.  
sentiment\_analysis --usesMethod--> emoji\_embeddings.

**Verbalised graph structure**

Carefully read the subgraph and summarize scientific insights in several bullet points. Then be creative and propose new research ideas. Think step by step.

**Task**

[Gradio \(nee.st\)](https://www.gradio.net)

# Literature-Based Discovery (LBD) – Closed Discovery

## Convert the graph to text and create hypothesis

Search Nodes   Find Shortest Path

node1  
schema\_based\_knowledge\_graph

node2  
graph\_transformer\_network

Clear   Submit

output

PATH: schema\_based\_knowledge\_graph --usesOtherEntity--> relation\_type --otherEntityUsedBy--> expansion\_process --usesMaterial--> knowledge\_base --http://www.w3.org/2004/02/skos/core#broader--> database --usesMaterial--> transfer\_learning --http://www.w3.org/2004/02/skos/core#narrower--> machine\_learning\_paradigm --http://www.w3.org/2004/02/skos/core#broader--> graph\_transformer\_network  
RESPONSE: Scientific Insights:

- The schema-based knowledge graph utilizes the concept of a 'relation type', which is subsequently used in the expansion process. This suggests the need for structuring and categorizing relationships in the knowledge graph for effective expansion.
- The expansion process uses a knowledge base as a key material. This may indicate an iterative process that relies on pre-existing or pre-established data collections.
- The knowledge base is a broader concept than the specific concept of a database. This implies the wide and diverse application of knowledge bases within computer science.
- The database uses transfer learning as a material, which is a machine learning method. This procedure improves performance by transferring knowledge from one situation to another.
- Transfer learning is a narrower concept than the machine learning paradigm, indicating its specific application within the broader area of machine learning.
- The machine learning paradigm is a broader concept than the specific concept of a graph transformer network. This suggests the use of more complex, layered architectures like graph transformer networks within the broader area of machine learning.

Research Ideas:

- A study can be performed to enhance the efficiency of the schema-based knowledge graph by optimizing the use of relation types.
- Research can be carried out to create an AI-driven expansion process that automatically learns and expands knowledge by continuously learning from a knowledge base.
- In-depth exploration can be undertaken to compare the effectiveness of distinct databases in transfer learning and their relation with different machine learning paradigms.
- Research can also focus on utilizing machine learning paradigms, especially graph transformer networks, in the creation and expansion of knowledge base-driven systems.
- An interesting area for research would be to investigate the effectiveness of transfer learning and graph transformer networks in improving the expansion and utility of knowledge databases in the field of computer science.

# Conclusion and Outlook



- Generative AI can facilitate scientific knowledge discovery by .
- Combining RAG and graph exploration allows for „guided hallucination“.
- How to teach LLMs about graph structure?
  - Graphs are much richer of information than just connections.
  - Pure verbalization of subgraphs just a start.
  - Combination of graph algorithms and prompt engineering.