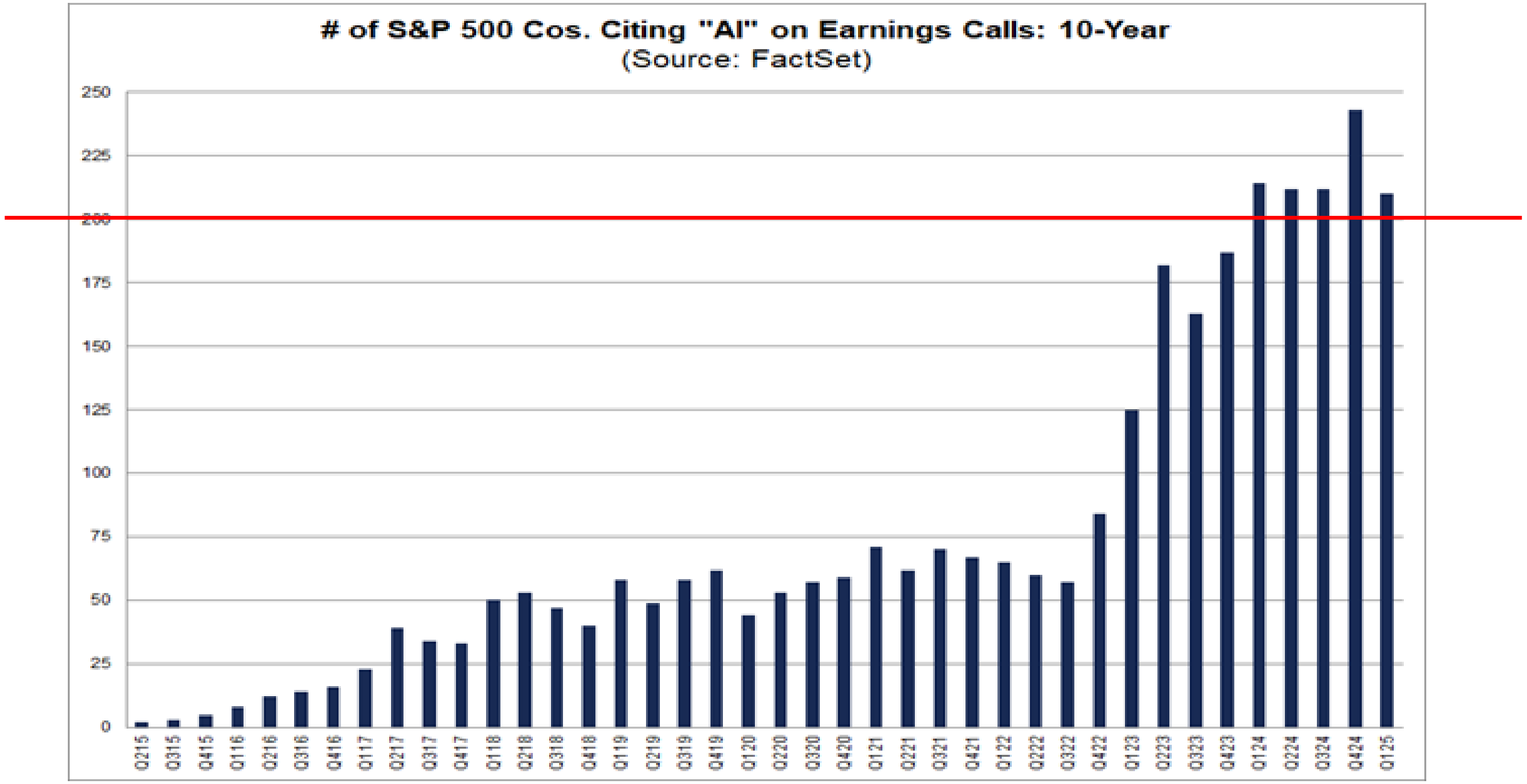
From Knowledge Graphs to Neural Graph Databases: Data, Algorithms, and Applications

Jiaxin Bai
Hong Kong University of Science and Technology
Feb 2026

The Rise of AI: S&P 500 Companies Discussing "AI"



- More than 40% of S&P 500 companies cited AI after ChatGPT launched

https://insight.factset.com/more-than-40-of-sp-500-companies-have-cited-ai-on-earnings-calls-for-5th-straight-quarter

Milestones of LLM/Foundation Model on Text

- In 2003: Neural Language Model is introduced [1]: **32 million Token**

Data

- In 2007: Google Build first 5-gram LM with [2]: **2 Trillion Corpus**

Algorithms

- In 2018: GPT/BERT are built as **Neural LM** with Transformer: 3.3 Billion Corpus

Computes

- Since 2023-: 100+ **Billion Neural LLMs** Trained on > 1Trillion Data: GPT-3, GPT-4, Qwen, DeepSeek, … **Prompting, Few-shot, RAG, etc.**

Applications

[1] Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. "A neural probabilistic language model." *Journal of machine learning research* 3, no. Feb (2003): 1137-1155.
[2] Brants, Thorsten, Ashok Popat, Peng Xu, Franz Josef Och, and Jeffrey Dean. "Large language models in machine translation." In *EMNLP-2007*

Why Language Models are Successful?

Reasons related to this talk:

	Memorization	**Generalization**
	Storing patterns from training data	*Applying patterns to new contexts*
Facts:	“Paris is France’s capital”	Answering novel questions
Literature:	Quoting famous texts	Creating original content
Technical:	Formulas, definitions	Solving unseen problems
Code:	Common programming patterns	Adapting to unfamiliar formats

LLMs **balance** **memorized knowledge** with the ability to **generalize patterns** to new situations, enabling both factual accuracy and creative problem-solving.

Neuralization is the Key for Generalization

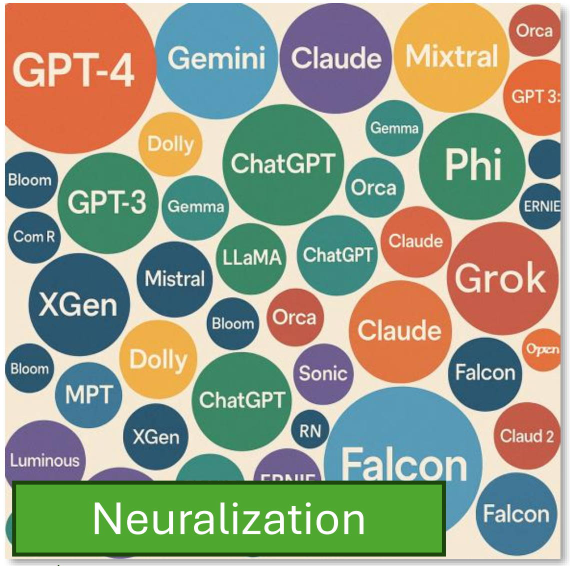
Call me Ishmael. Some years ago—never mind how long precisely—having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen and regulating the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such a upper hand of me, that it requires a strong moral principle to

Text Data

Neuralization

Large Language Model

Generalization from Structured Data?

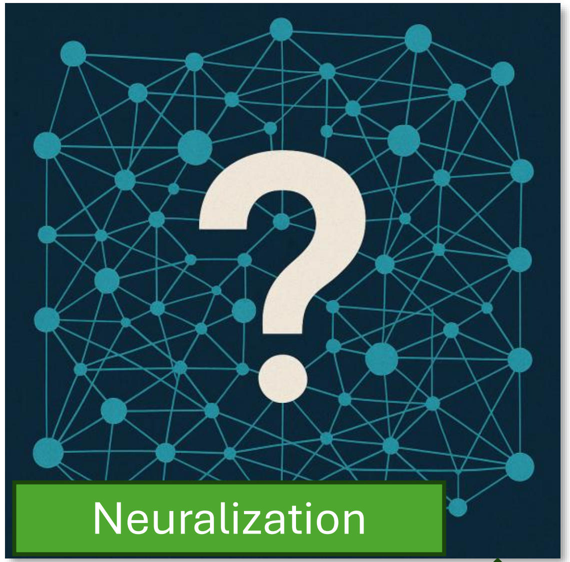


Large Language Model

Ilya Sutskever [1]

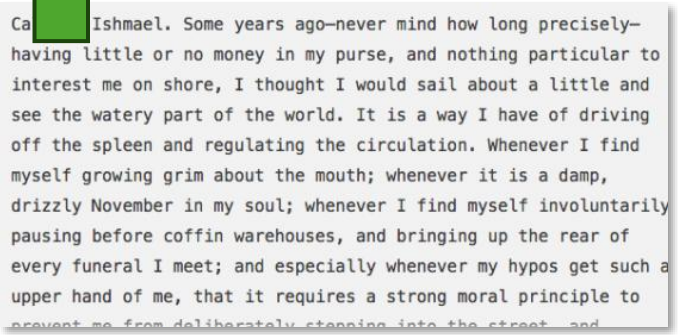
LLM Pre-training scaling will stop because we only have one internet!

Neural Graph Database (NGDB) show great potential for Database Memorization and Generalization!



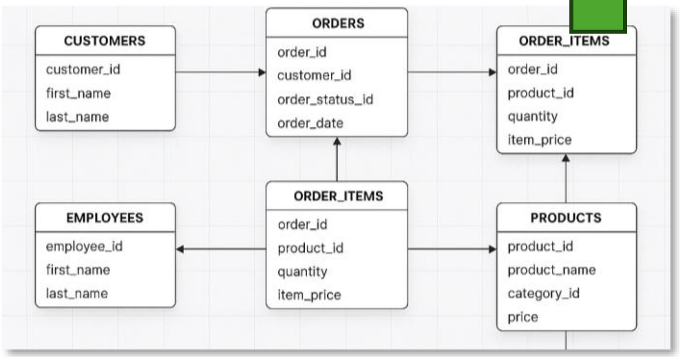
Neural (Graph) Databases

Text Data


Ishmael. Some years ago—never mind how long precisely—having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen and regulating the circulation. Whenever I find myself growing grim about the mouth; whenever it is a damp, drizzly November in my soul; whenever I find myself involuntarily pausing before coffin warehouses, and bringing up the rear of every funeral I meet; and especially whenever my hypos get such a upper hand of me, that it requires a strong moral principle to

- Data within databases **continues to grow** and won't be easily depleted;
- At **very early stage** where universities can do research [2];
- Combines **general applicability** to real-world applications, positioning it to make industry impact.

Structured Data



[1] https://youtu.be/1yvBqasHLZs
[2] Wang, Y., Wang, X., Gan, Q., Wang, M., Yang, Q., Wipf, D., & Zhang, M. (2025). Griffin: Towards a Graph-Centric Relational Database Foundation Model. arXiv preprint arXiv:2505.05568.

Data Beyond Text

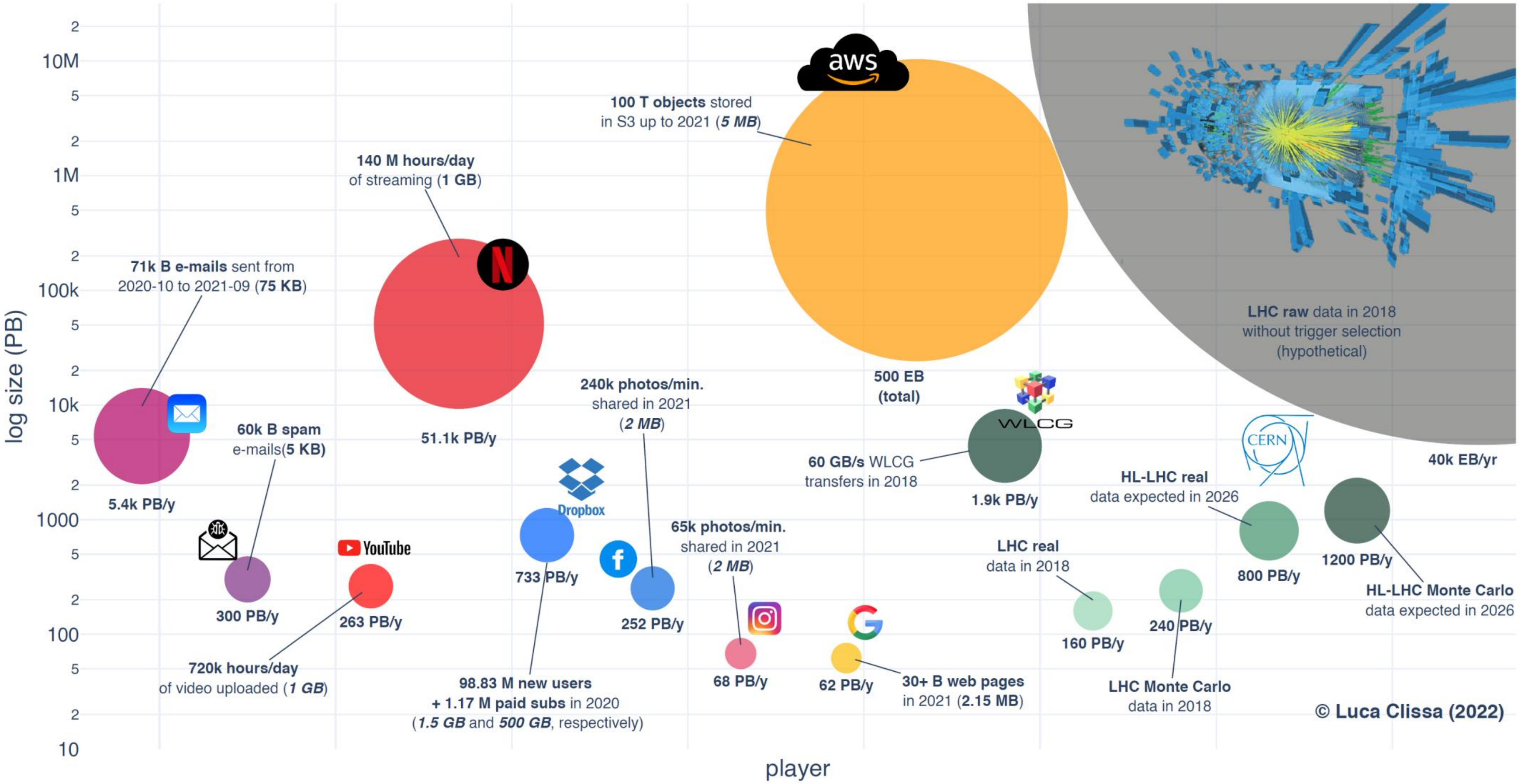
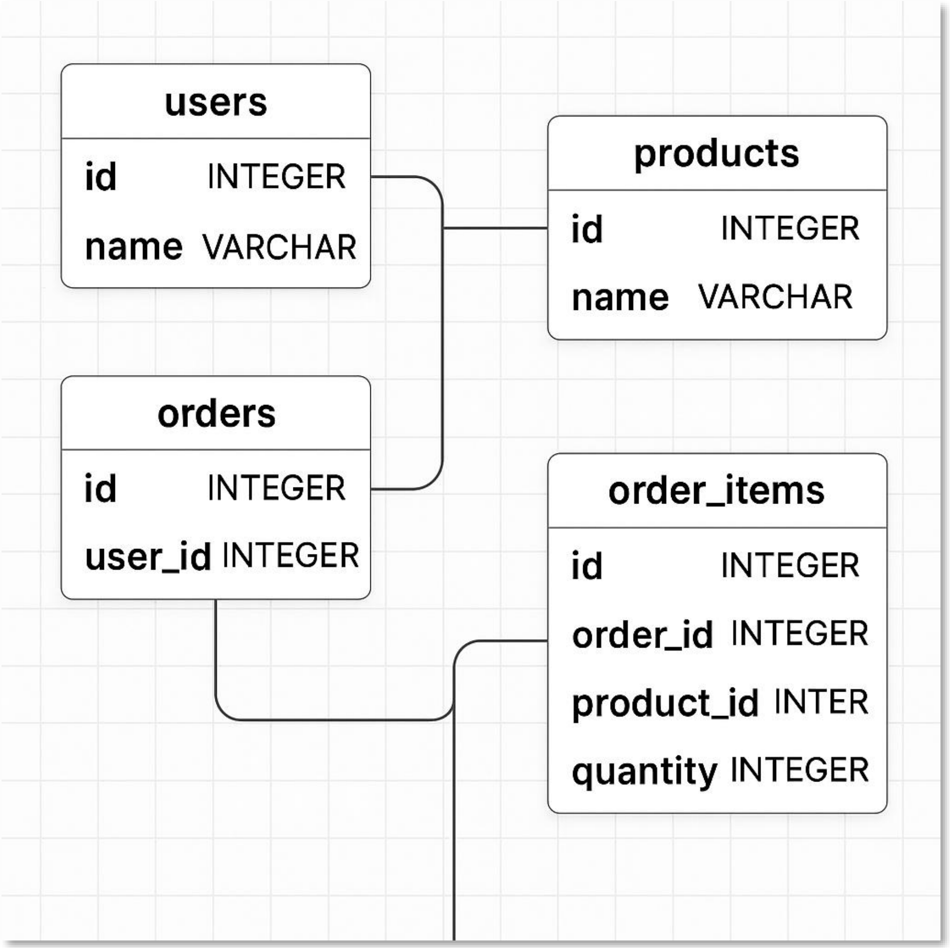


Figure 1: **Big Data sizes.** Bubble plot of the orders of magnitude of data produced by important big data players. The balloon areas illustrate the amount of data and the text annotations highlight the key factors considered in the estimates. Average per-unit sizes are reported in parentheses, where italic indicates measures reconstructed based on likely assumptions because no references were found. Interactive version available at: BigData2021.html

https://arxiv.org/pdf/2202.07659.pdf

Recent Related Efforts: Relational Deep Learning



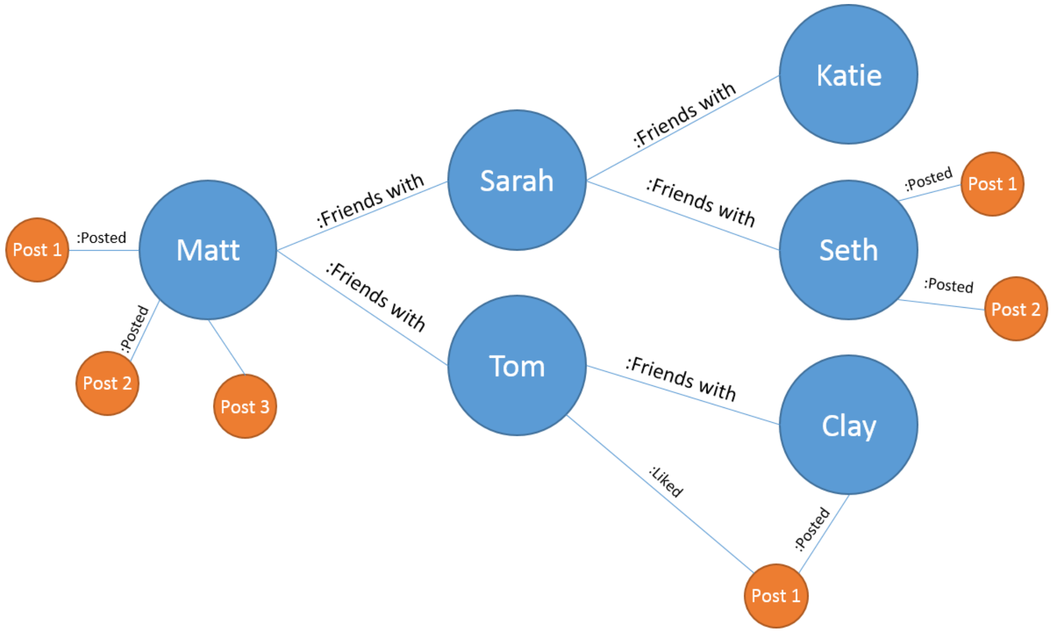
- **Data Is Stored in Databases**
- **Core Concept**: Applying deep learning directly to relational databases **tables** by treating them as graphs [1]
- **Statistical Relational Learning** (SRL) has been a long standing research topic [2].

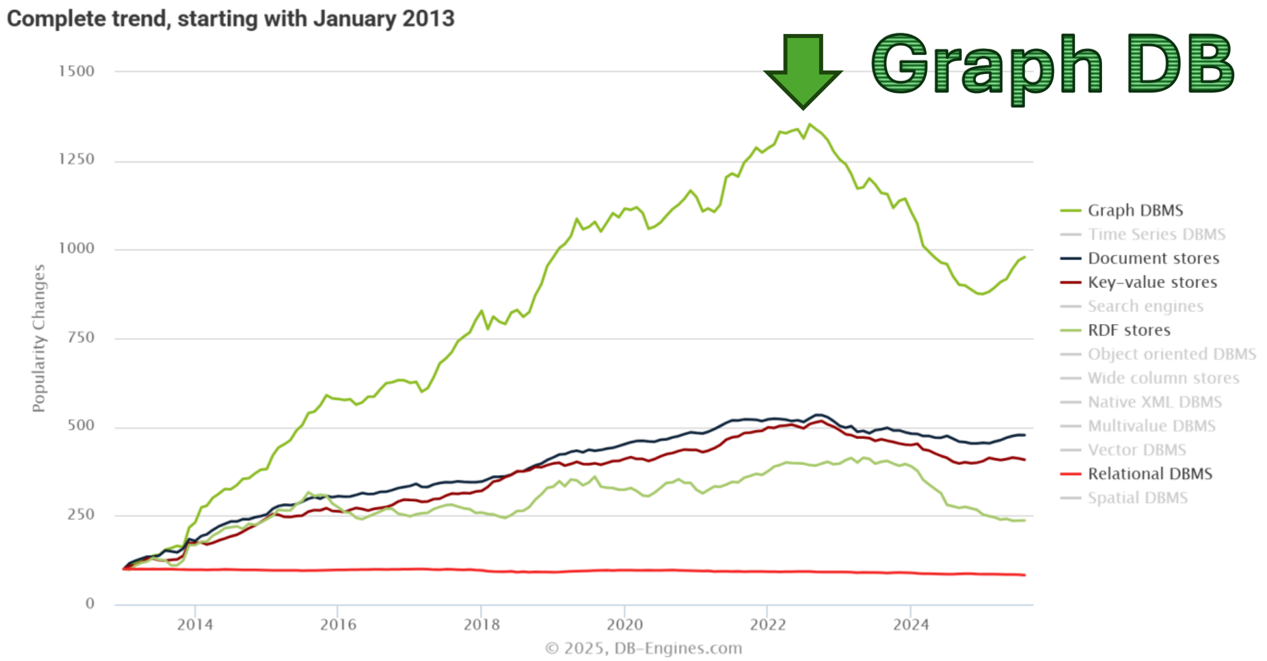
Jure Leskovek (Stanford)

[1] Fey, M., Hu, W., Huang, K., Lenssen, J. E., Ranjan, R., Robinson, J., ... & Leskovec, J. (2024, July). Position: Relational deep learning-graph representation learning on relational databases. In *Forty-first International Conference on Machine Learning*.
[2] Getoor, Lise, and Ben Taskar, eds. Introduction to statistical relational learning. MIT press, 2007.

Graph Databases

- A DBMS with **graph data model**
- Good at
  - **Relationships**
  - **Traversals**
  - **Pattern matching**
- Flexible Schema, Matching-based Query Language (SPARQL, Cypher, etc.)





https://db-engines.com/en/ranking_categories

Many Types of Data are Graphs

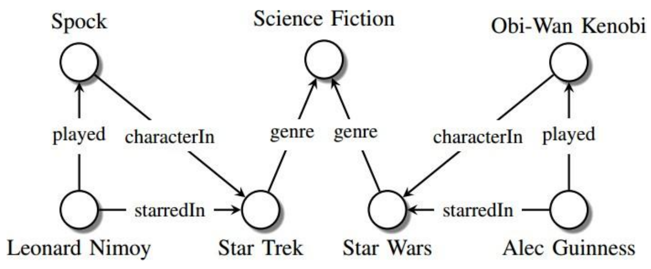
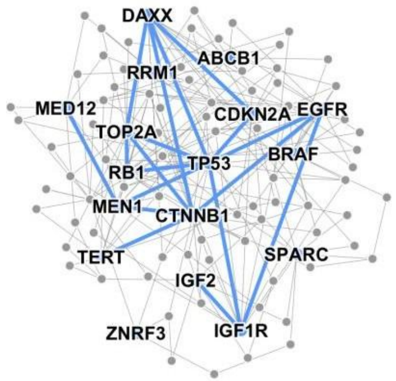


Image credit: [Maximilian Nickel et al]

Knowledge Graphs

Image credit: [SalientNetworks]

Computer Networks



Disease Pathways

Image credit: [Medium]

Social Networks

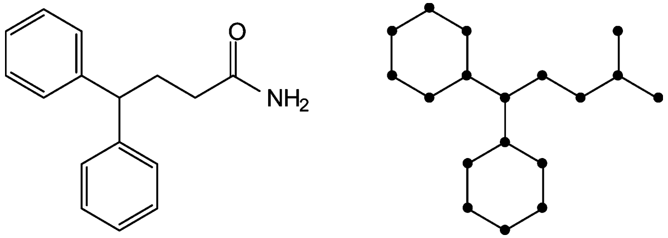


Image credit: [MDPI]

Molecules

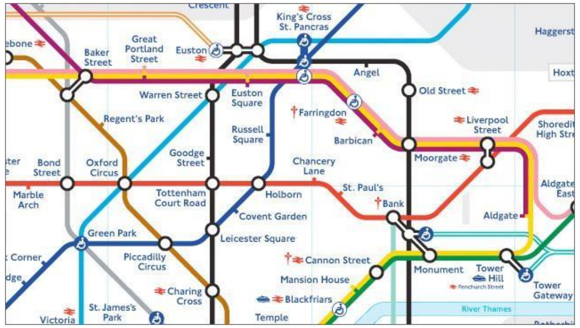
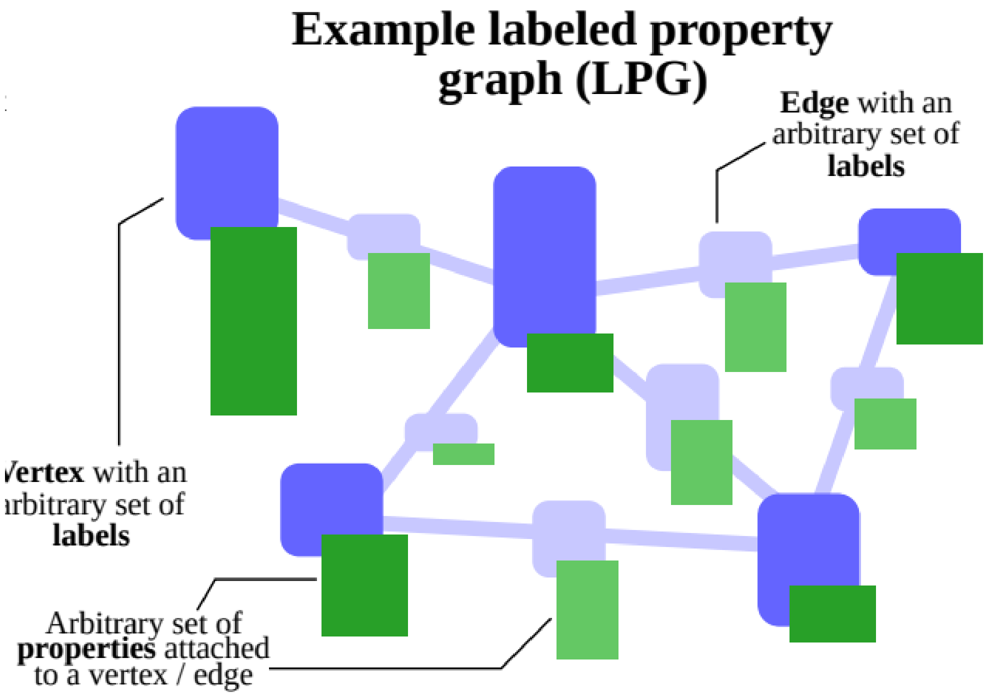


Image credit: [visitlondon.com]

Traffic Networks

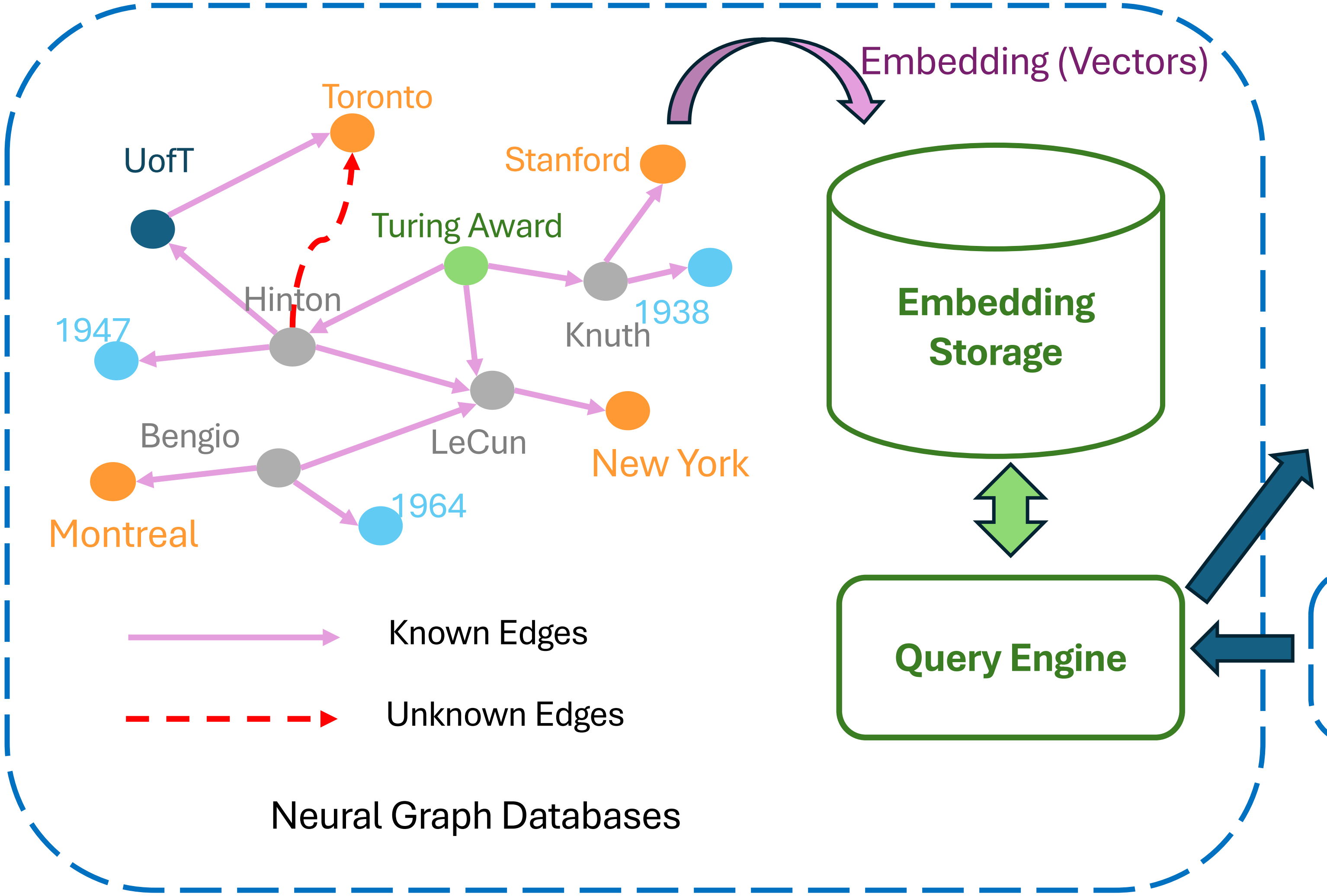
Neural Graph Databases (NGDB): Earlier Work



- **Neuralization** is only to the data level
- Learning better node representations on LPG
- How about **Query Answering**?

[1] Besta, Maciej, Patrick Iff, Florian Scheidl, Kazuki Osawa, Nikoli Dryden, Michal Podstawski, Tiancheng Chen, and Torsten Hoefler. "Neural graph databases." In *Learning on Graphs Conference*, pp. 31-1. PMLR, 2022.

Neural Graph Databases (NGDB)



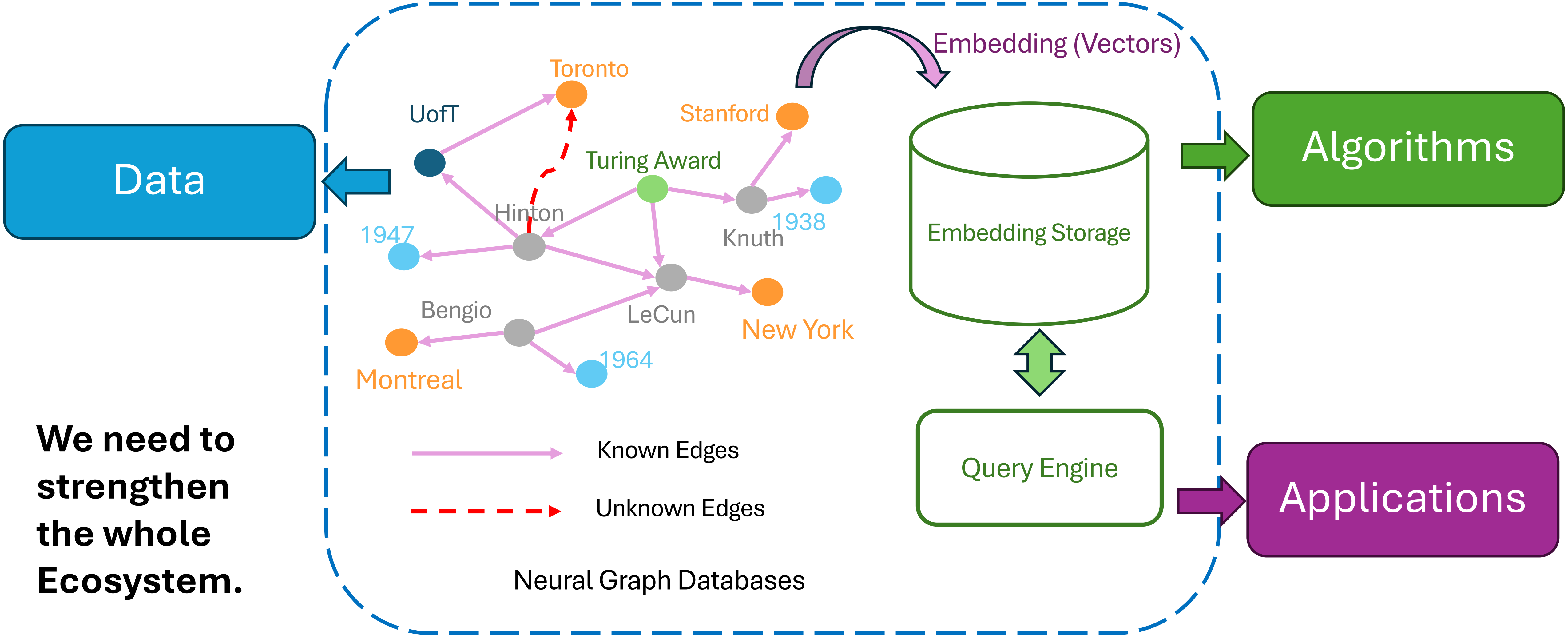
Memorization by Graph Database: **Stanford**

Generalization by Neural Graph Database: **Toronto** & **Stanford**

$V_?. \exists V_1, X_2: Win(V_1, Turing\ Award) \wedge LessThan(X_2, 1960) \wedge BornIn(V_1, X_2) \wedge Livein(V_1, V_?)$

Where are the Turing Award Winners born before 1960 lives in?

Neural Graph Databases (NGDB): Ecosystem



We need to strengthen the whole Ecosystem.

Neural Graph Databases (NGDB): Challenges

Domain	Fundamental Challenge
Data	**Autonomous Structured Knowledge Construction** How do we autonomously construct large-scale, high-quality structured knowledge from raw data sources?

Neural Graph Databases (NGDB): Challenges

Domain	Fundamental Challenge
Data	**Autonomous Structured Knowledge Construction** How do we autonomously construct large-scale, high-quality structured knowledge from raw data sources?
Algorithms	**Comprehensive Neural Reasoning** How do we create neural methods for comprehensive logical, numerical, and even causal reasoning that are scalable and efficient?

Neural Graph Databases (NGDB): Challenges

Domain	Fundamental Challenge
Data	**Autonomous Structured Knowledge Construction:** How do we autonomously construct large-scale, high-quality structured knowledge from raw data sources?
Algorithms	**Comprehensive Neural Reasoning** How do we create neural methods for comprehensive logical, numerical, and even causal reasoning that are scalable and efficient?
Applications	**Agentic Knowledge Integration** How do we build agentic systems that can formulate their own queries and integrate this structured reasoning into real-world applications?

This Talk – Towards NGDB

🏗️ **Data:** Autonomous Structured Knowledge Construction

🧠 **Algorithm & System:** Comprehensive Neural Reasoning

🤖 **Application:** Agentic Knowledge Integration

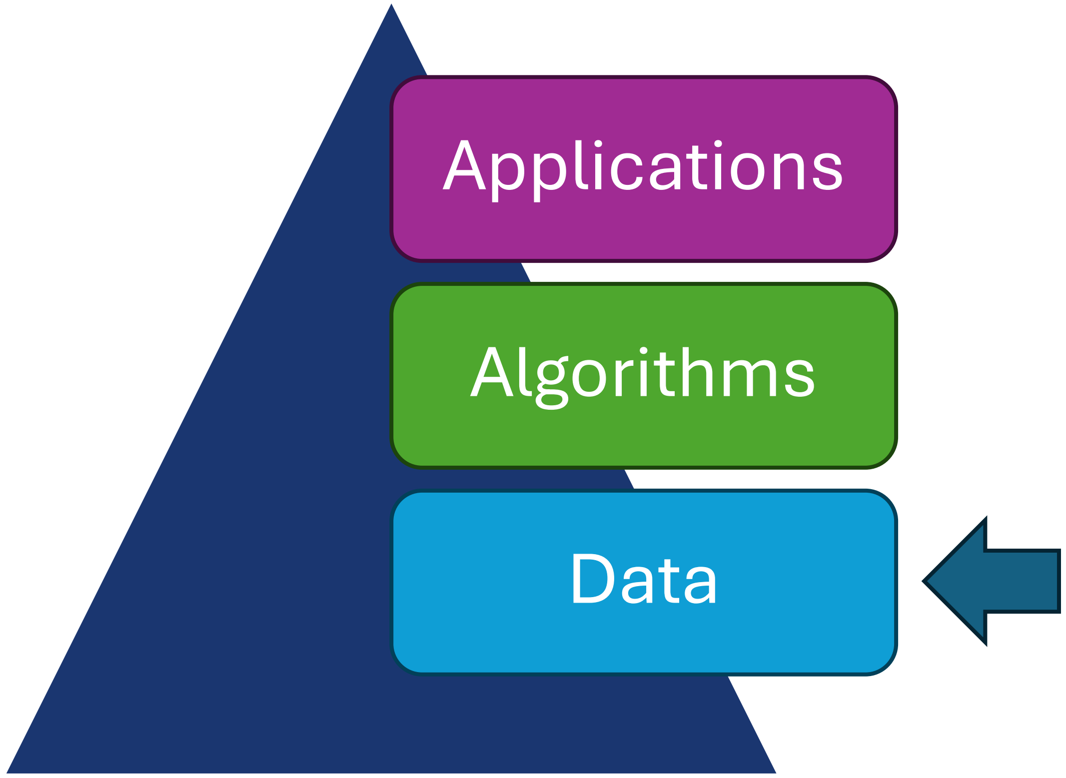
This Talk – Towards NGDB

🏗️ **Data:** Autonomous Structured Knowledge Construction

🧠 **Algorithm & System:** Comprehensive Neural Reasoning

🤖 **Application:** Agentic Knowledge Integration

Not Enough Data for NGDB Research?

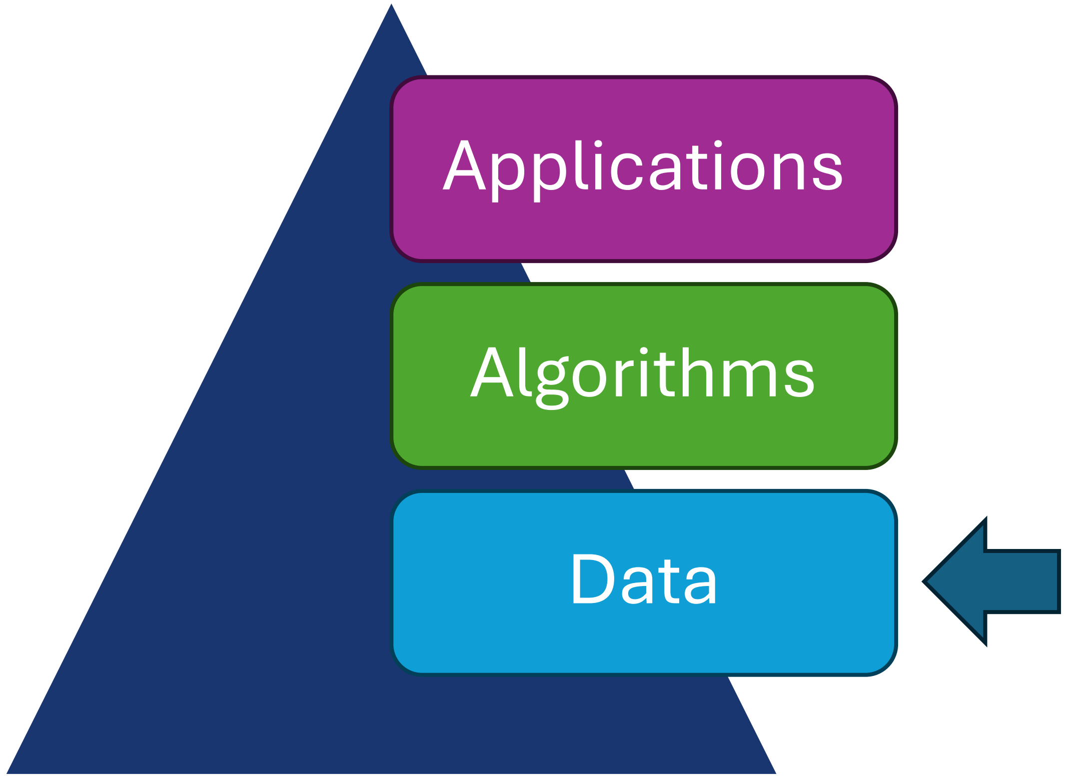


Existing Databases are:

- Domain-specific,
- Siloed,
- often Private

More general knowledge source?

How about Existing General Knowledge Source



- Freebase, Wikidata, and YAGO is largely sourced **from Wikipedia**.
- Wikipedia is less than **1%** of LLM pre-training data [1].
- Comparable Scale to LLMs?
- How about create one using LLMs for **autonomous structurization**?

[1] https://huggingface.co/datasets/allenai/dolma

AutoschemaKG: Challenges

Challenges in KG Construction

- Predefined Schemas
- Entity-only
- Traditional Information Extraction

AutoSchemaKG

- Autonomous Schema Induction
- Entity-Event Modeling
- LLM-based Extraction

In short it removes the manual bottleneck and creates richer, more adaptable knowledge graphs.

Bai, Jiaxin, Wei Fan, Qi Hu, Qing Zong, Chunyang Li, Hong Ting Tsang, Hongyu Luo et al. "AutoSchemaKG: Autonomous Knowledge Graph Construction through Dynamic Schema Induction from Web-Scale Corpora." *arXiv preprint arXiv:2505.23628* (2025).

Framework Overview

Input Documents

- ***Source Selection***: Wikipedia, academic papers, Common Crawl
- ***Segmentation***: Divide documents into chunks
- ***Batching***: Group chunks into batches

Document 1: Lorem ipsum dolor sit amet...
Document 2: Consectetur adipiscing elit...
Document 3: Sed do eiusmod tempor incididunt...
...

Chunked Batches

Batch1 | Batch2 | Batch3

Knowledge Graph Triples

- **Entity-Entity Extraction**: Using prompt P_{EE}, extract relationships between entities
- **Entity-Event Extraction**: Using prompt P_{EV}, link entities to events
- **Event-Event Extraction**: Using prompt P_{VV}, detect temporal/causal connections
- **Output Processing**: Parse JSON outputs, repair if needed

Triple Extraction Process

Entity	Relation	Entity
Entity	Participates	Event
Event	Causes	Event

Schema Induction

Black Mountain College
↓ Concepualization
College | School | Institution

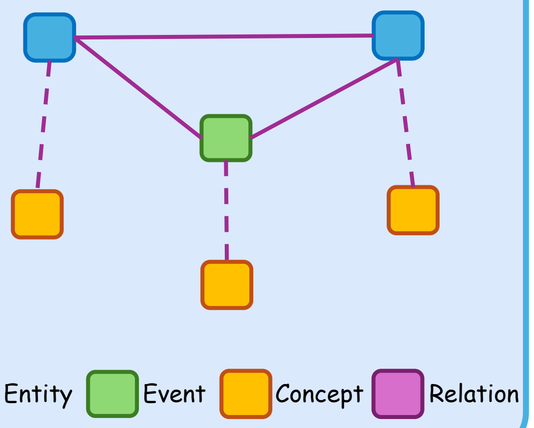
A cat chased a prey
↓ Concepualization
Hunting | Predation | Pursuit

- **Entity Conceptualization**: Generate abstract entity types using context from neighboring nodes
- **Event Conceptualization**: Abstract events into general event types
- **Relation Conceptualization**: Transform relations into semantic categories

Knowledge Graph Construction

- **Node Creation**: Build entity (V_N), event (V_E), and concept (C) nodes
- **Edge Creation**: Establish connections using extracted relationships (R)
- **Concept Linking**: Connect nodes to corresponding concept nodes
- **Graph Finalization**: Complete knowledge graph $G = (V, E, C, \varphi, \psi)$

ATLAS Knowledge Graph



ATLAS Knowledge Graph Family

The ATLAS (Automated Triple Linking And Schema induction) knowledge graph family built from multiple corpora:

Metric	ATLAS-Wiki	ATLAS-Pes2o	ATLAS-CC	Total
Text Chunks	9.599M	7.918M	35.040M	52.557M
Entities	70.104M	75.857M	241.061M	387.022M
Events	165.717M	92.636M	696.195M	954.548M
Concepts	8.091M	5.895M	31.070M	45.056M
Total Nodes	**243.912M**	**174.387M**	**937.256M**	**1.355B**
Total Edges	**1.492B**	**1.150B**	**5.958B**	**8.600B**

900M+
Total Nodes

5.9B+
Total Edges

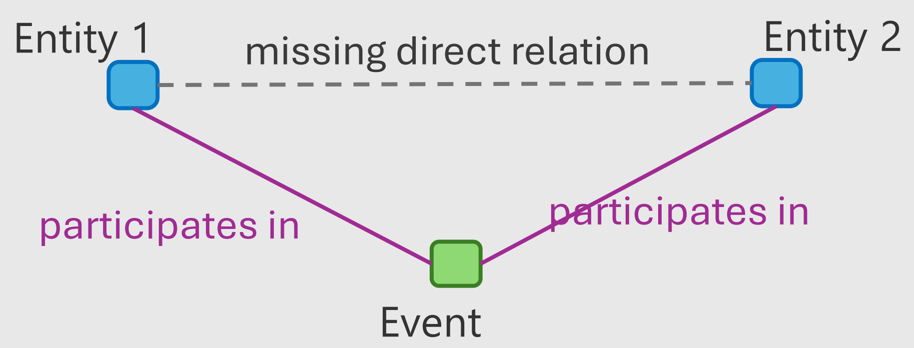
78,400
GPU Hours

Computational Requirements: Built using 80GB GPUs with 1,513 TFLOPS of FP16 compute, running Llama-3-8B-instruct with Flash Attention.

Advantages of Event and Concept Nodes

Event Nodes Provide Enriched Context

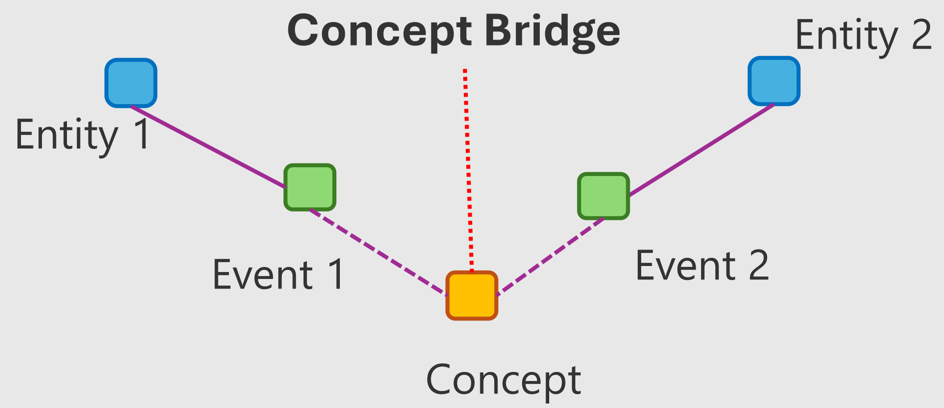
- Preserve over **90%** of original passage content (vs. **70%** for entity-only)
- Capture **temporal** relationships, **causality**, and **procedural** knowledge



Event nodes connect entities even when direct relations are missing

Concept Nodes Create Alternative Pathways

- Establish connections **beyond entities and events**
- Address complex **multi-hop** question answering limitations
- Link knowledge across **disconnected** subgraphs



Concepts bridge otherwise disconnected subgraphs

Results

Multi-hop QA Performance

Dataset	Method	EM	F1
HotpotQA	No Retriever	37.0	47.3
	GraphRAG [1]	55.2	68.6
	HippoRAG2 [2]	**62.7**	75.5
	AutoSchemaKG	61.8	**78.3**
2Wiki	No Retriever	36.5	42.8
	GraphRAG	51.4	58.6
	HippoRAG2	65.0	71.0
	AutoSchemaKG	**65.3**	**73.9**

Key Finding: 12-18% improvement over state-of-the-art baselines on multi-hop QA tasks

Information Preservation

Dataset	Context	LLaMA-3-8B	LLaMA-3-70B
ATLAS-Wiki	Entity	65.08	70.96
	Event	92.69	94.82
	Event+Entity	**93.30**	**95.13**

LLM Factuality Enhancement

Method	Accuracy	F1
No Retrieval	54.08	26.79
BM25	56.15	30.43
ATLAS-Wiki KG	**56.43**	**30.48**

Key Finding: Up to 9% improvement in LLM factuality across domains

[1] Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., ... & Larson, J. (2024). From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.
[2] Gutiérrez, B. J., Shu, Y., Qi, W., Zhou, S., & Su, Y. From RAG to Memory: Non-Parametric Continual Learning for Large Language Models. *International Conference on Machine Learning 2025*.

This Talk – Towards NGDB

🏗️ **Data:** Autonomous Structured Knowledge Construction

🧠 **Algorithm & System:** Comprehensive Neural Reasoning

🤖 **Application:** Agentic Knowledge Integration

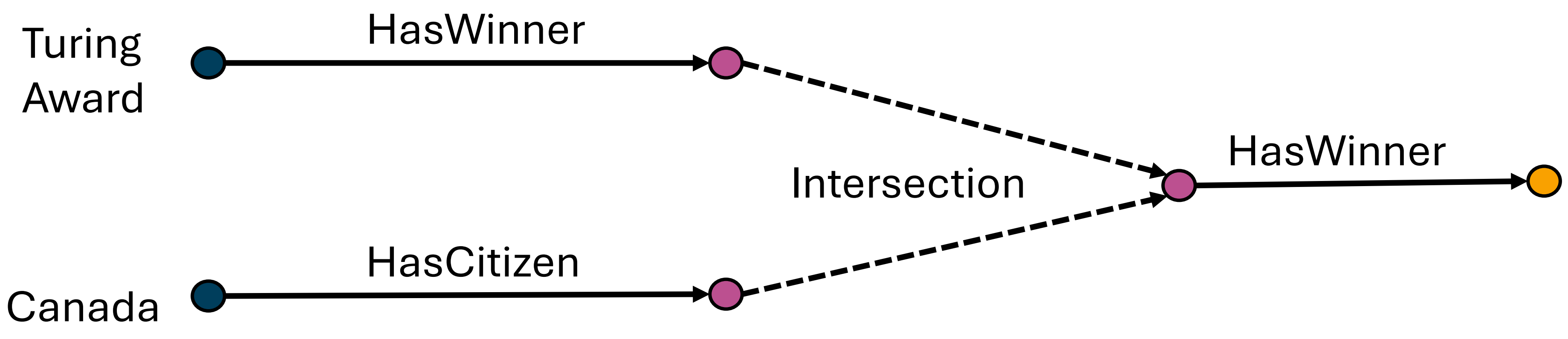
Reasoning: Query Answering on NGDB

Complex Queries	Interpretations
$q_1 = V_? \,.\, \exists V : Win(TuringAward, V)$ $\wedge\, Citizen(Canada, V) \wedge Graduate(V, V_?)$	Find where the Canadian Turing award laureates graduated from.
$q_2 = V_? \,.\, \exists V : Interacts(V_?, V)$ $\wedge\, (Assoc(V, T_1) \vee Assoc(V, T_2) \vee Assoc(V, T_3))$	Find the substances that interact with the proteins associated with diseases T1, T2, or T3.
$q_3 = V_? : MovesTo(V_?, UnitedStates)$ $\wedge\, IsResident(V_?, Germany)$ $\wedge\, Wins(V_?, NobelPrize)$	Find entities, who are Germans, were the Nobel Prize winners and eventually moved to the United States.

How can we generalize to unseen question and knowledge?

How do we scale up to large graphs and long queries?

Background: Query Encoding



How can we generalize to unseen question and knowledge?

Use embeddings to represent queries and subqueries

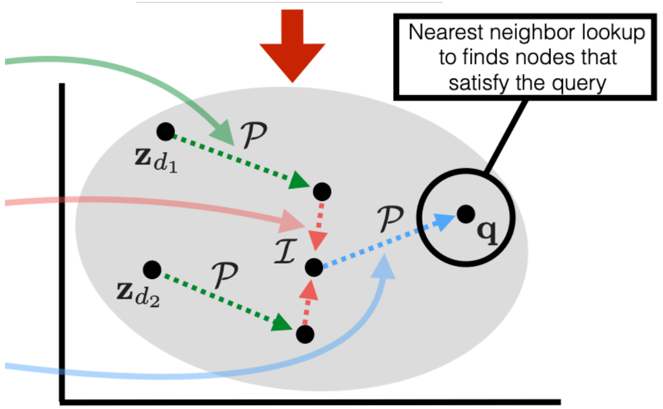
How do we scale up to large graphs and long queries?

Using computation to replace graph search / subgraph matching

Only a single approximate nearest neighbour search in inference

Earlier Work: Neural Query Encoders

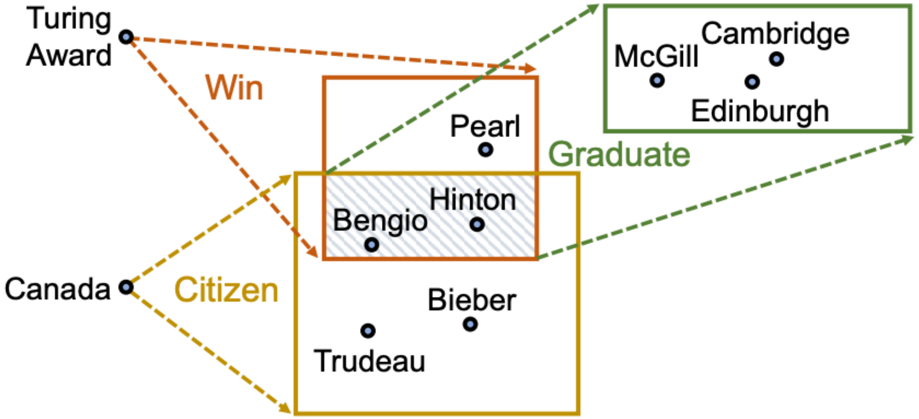
Models	Encoding Structures
GQE [1]	Vector Embedding
Query2Box [2]	Box Embedding
FuzzQE	Fuzzy Logic Embedding
Neural MLP	Vector Embedding
NewLook	Vector Embedding
…	…



Figures from [1]

Vectors and Boxes are not expressive enough for more complex queries with very diverse answers.

Multi-vector Representation?

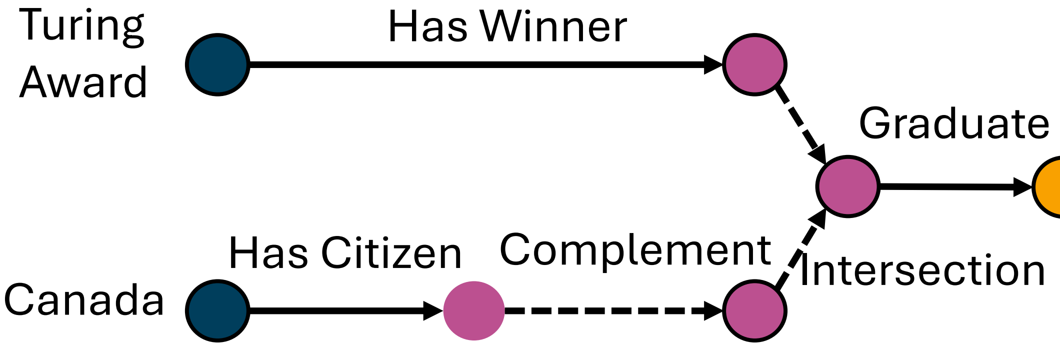


Figures from [2]

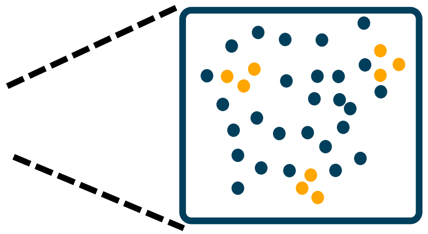
https://github.com/neuralgraphdatabases/awesome-logical-query

Query2Particles: Embedding Space and Set Representations

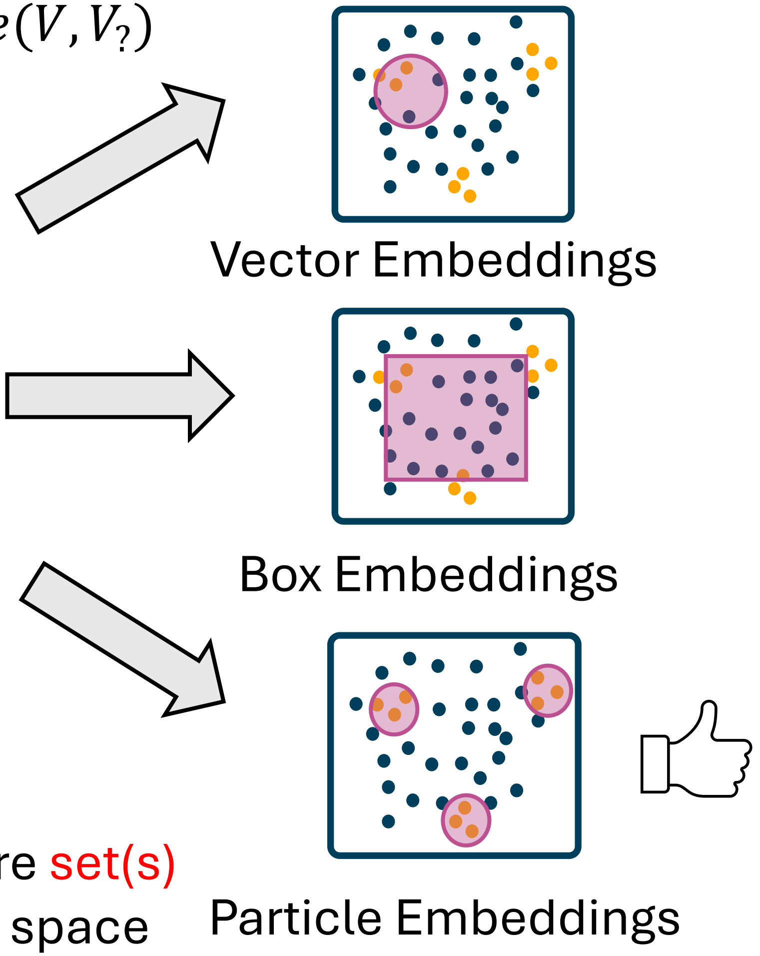
$q = V_? \,.\, \exists V: Win(TuringAward, V) \wedge \neg Citizen(Canada, V) \wedge Graduate(V, V_?)$



Computation Graph



Embedding Space



The multi-hop logical operations make the query answers diversified → The answers embeddings are set(s) scattered in the embedding space

Example from: Jiaxin Bai, Zihao Wang, Hongming Zhang, Yangqiu Song: Query2Particles: Knowledge Graph Reasoning with Particle Embeddings. NAACL-HLT (Findings) 2022: 2703-2714

Parameterization of Operator: Relational Projection

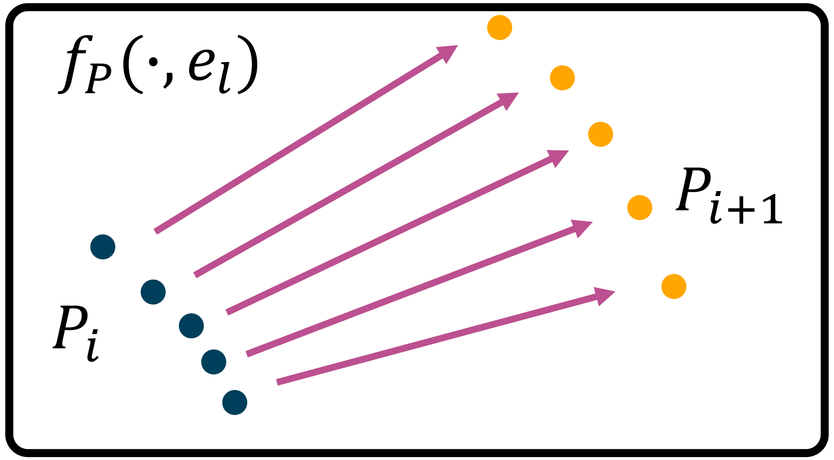
- Input: particles P_i and relation l
- Output: Particles P_{i+1}
- Gated Transition for customizing the directions of transitions for each vector in particles:

$$A_i = (1 - Z) \odot P_i + Z \odot T$$

  Here Z is the update gate, and T is transition for each particles. They are computed from P_i and the relation embedding e_l for relation l

- Self-Attention for information exchange in the particles:

$$P_{i+1} = \texttt{self-attn}(A_i)$$

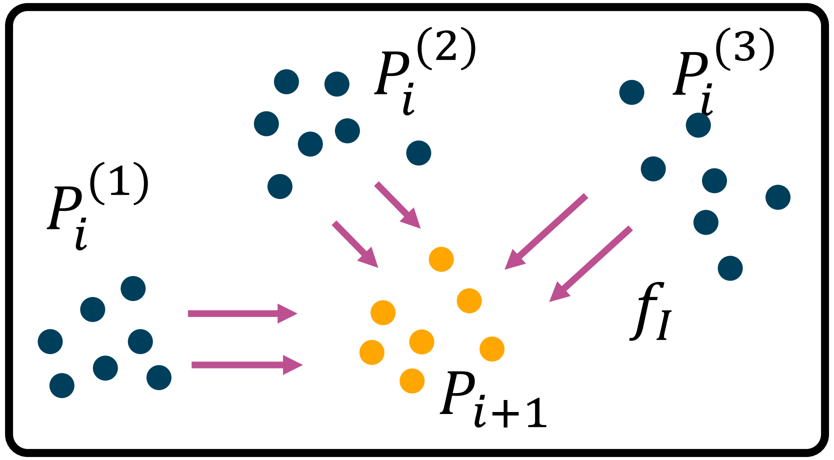


Relational Projection

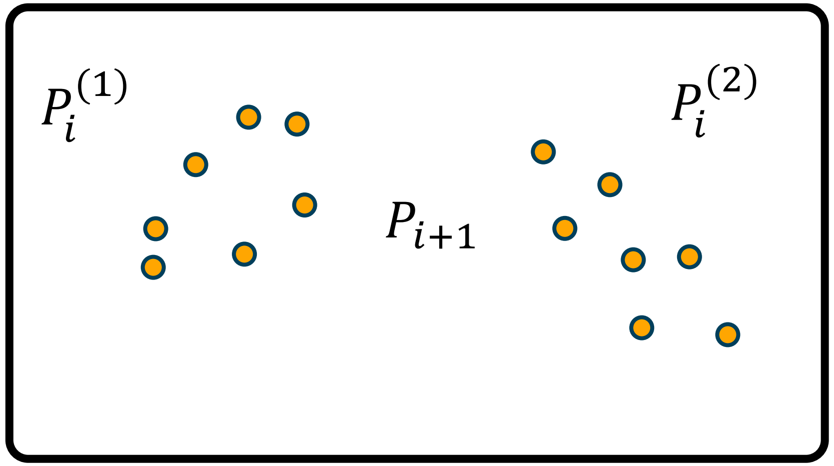
Intersection, Union, and Negation

- Input: particles P_i
- Output: Particles P_{i+1}
- Self-Attention + MLP layer for information exchange between the input particles:

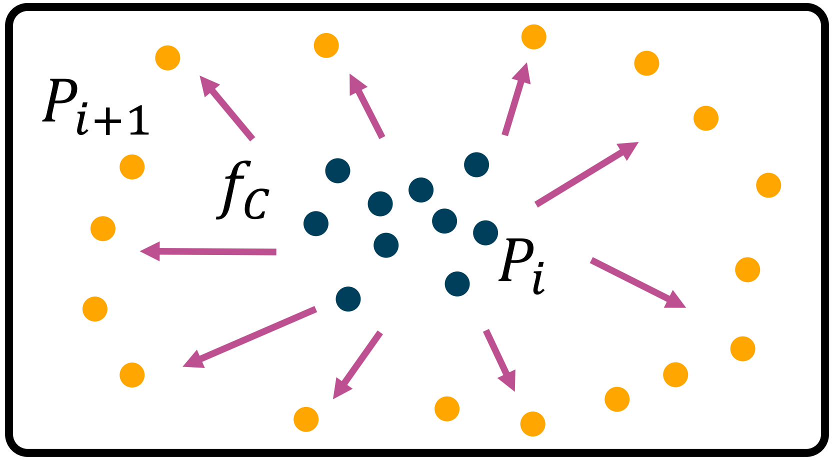
$$A_i = \texttt{self-attn}(P_i)$$
$$P_{i+1} = \texttt{MLP}(A_i)$$



Intersection



Union



Complement

Training Query2Particles

Use positive pairs with cross-entropy loss for training:

- The scoring function is the maximal inner product from particles to an entity v:

$$\varphi(P_q,\ v) = \max_{k=1,2,3,\ldots,K} < p_T^{(k)}, e_v >$$

- The normalized probability of the entity v being an answer of q:

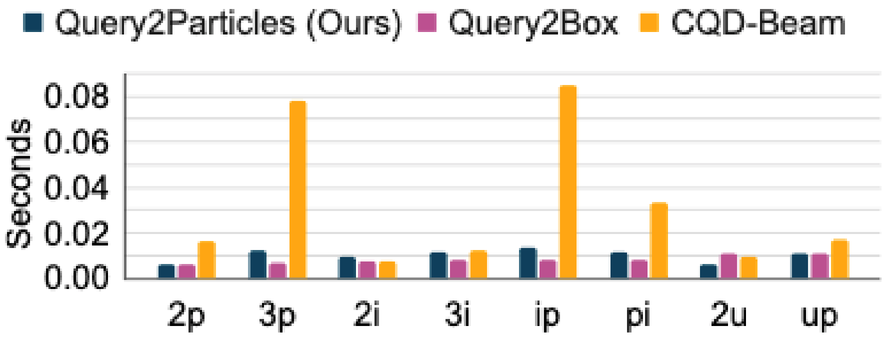
$$P(q,\ v) = \frac{e^{\varphi(P_q, v)}}{\sum_{v'} e^{\varphi(P_q, v')}}$$

- The cross-entropy loss maximizes the log probability of all query-and-answer pairs:

$$L = -\frac{1}{N}\sum_i \log P(q^{(i)}, v^{(i)})$$

Comparison with baselines

	FB15K										FB15K-237	NELL
MODEL	1P	2P	3P	2I	3I	IP	PI	2U	UP	AVG	AVG	AVG
EMQL	42.4	50.2	45.9	63.7	70.0	60.7	61.4	9.0	42.6	49.5	35.8	46.8
– SKETCH	50.6	46.7	41.6	61.8	67.3	54.2	53.5	21.6	40.0	48.6	35.5	46.8
CQD-BEAM	**91.8**	**77.9**	57.7	79.6	83.7	37.5	65.8	**83.9**	34.5	68.0	29.0	37.6
CQD-CO	**91.8**	45.4	19.1	79.6	83.7	33.6	51.3	81.6	31.9	57.6	27.2	36.8
Q2P (OURS)	90.2	74.6	**73.4**	**86.0**	**89.6**	**63.7**	**77.6**	83.4	**52.7**	**76.8**	**43.0**	**52.2**



- There is no need to pre-process the queries to their disjunction-normal forms (DNF), which takes exponential complexity
- Fast inference speed compared to the query decomposition method

Queries with Diverse Answers

Models	1P	2I	2U	2IN	Average
Q2P-1P	44.8	28.8	11.3	15.0	25.0
Q2P-2P	49.4	35.5	13.3	20.7	29.7
Q2P-3P	**53.0**	**37.7**	**18.6**	**21.6**	**32.8**

Significantly better on the queries with diverse answers

MRR on the top ten percent diversified queries
Diversity is measured by the number of answers.

Query2Particles: What is Missing?

- Query2Particles is a good enough **query embedding** method for NGDB reasoning engine.

- How to integrate **numerical values and comparisons** ?

- How to reason with **implicit temporal and causal constraints of events**?

How to integrate numerical values and comparisons into neural graph reasoning systems?

Category	Complex Queries	Interpretations
Numerical CQA	$q_2 = V_? \,.\, \exists X_1, X_2 \colon Win(V_?, TuringAward) \wedge \underline{GreaterThan}(1927, X_2) \wedge \underline{BornIn}\,(V_?, X_2)$	Find the Turing award winners that <u>is born before</u> the year of 1927.
Numerical CQA	$q_3 = V_? \,.\, \exists X_1, X_2 \colon LocatedIn(V_?, UnitedStates) \wedge \underline{HasLatitude}(V_?, X_1) \wedge \underline{GreaterThan}(X_1, X_2) \wedge \underline{HasLatitude}(Beijing, X_2)$	Find the states in US that have <u>a higher latitudes</u> than Beijing.
Numerical CQA	$q_4 = V_? \,.\, \exists X_1, X_2, X_3 \colon LocatedIn(V_?, UnitedStates) \wedge \underline{HasPopulation}(V_?, X_1) \wedge \underline{SmallerThan}(X_1, X_2) \wedge \underline{TimesByTwo}(X_2, X_3) \wedge \underline{HasPopulation}(California, X_3)$	Find the states in US that have a <u>twice smaller population</u> than California?

Bai, Jiaxin, Chen Luo, Zheng Li, Qingyu Yin, Bing Yin, and Yangqiu Song. "Knowledge graph reasoning over entities and numerical values." In *KDD*-2023.

How to encode and reason with implicit temporal and causal constraints in event knowledge graphs?

Complex query on eventuality graphs are **different** from the entity-relation graph

Whether and **when** the eventualities occur are important

Queries	Type	Interpretations
$q_1 = V_?.\exists V$: Interact$(V_?, V)$ $\wedge$ Assoc$(V, \mathit{Alzheimer}) \wedge$ Assoc(V, MadCow)	Entity	Find the substances that interact with the proteins associated with Alzheimer's and Mad cow disease.
$q_2 = V_?$.Precedence$(\mathit{Food\ is\ bad}, \mathit{PersonX\ add\ soy\ sauce})$ $\wedge$ Reason$(\mathit{Food\ is\ bad}, V_?)$	Eventuality	Food is bad before PersonX add soy sauce. What is the reason for food being bad?
$q_3 = V_?$.Precedence$(V_?, \mathit{PersonX\ go\ home})$ $\wedge$ ChosenAlternative$(\mathit{PersonX\ go\ home}, \mathit{PersonX\ buy\ an\ umbrella})$	Eventuality	Instead of buying an umbrella, PersonX go home. What happened before PersonX go home?

Jiaxin Bai, Xin Liu, Weiqi Wang, Chen Luo, Yangqiu Song: Complex Query Answering on Eventuality Knowledge Graph with Implicit Logical Constraints. *NeurIPS*, 2023

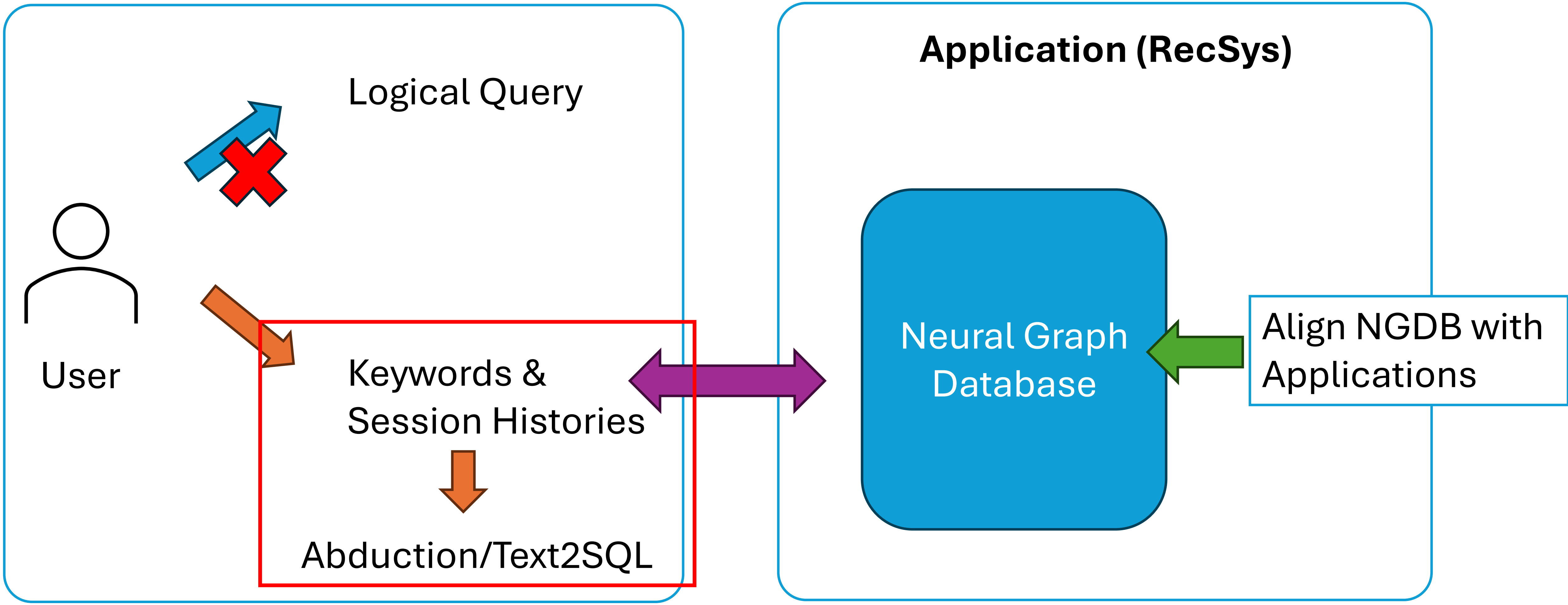
This Talk – Towards NGDB

🏗️ **Data:** Autonomous Structured Knowledge Construction

🧠 **Algorithm & System:** Comprehensive Neural Reasoning

🤖 **Application:** Agentic Knowledge Integration

NGDB Application on E-commerce



Abductive Reasoning

- **Abductive reasoning** is a form of reasoning that seeks to find the best explanation for an observation, distinct from the other two major types of inference: deduction and induction.

- Abductive Reasoning on NGDB:

  To use **structured knowledge in structured data** to explain **observations**.

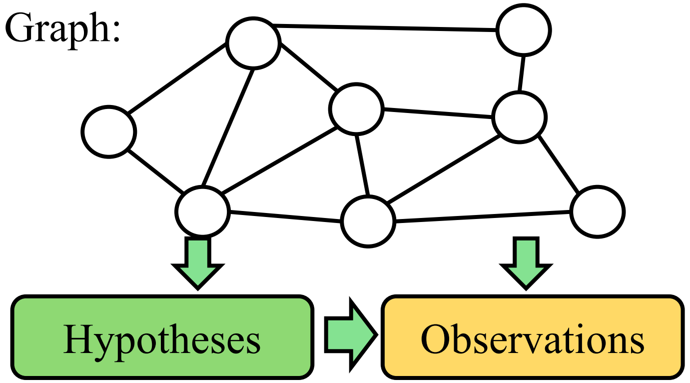
Abductive Reasoning

Using **structured knowledge** to explain **observations**.

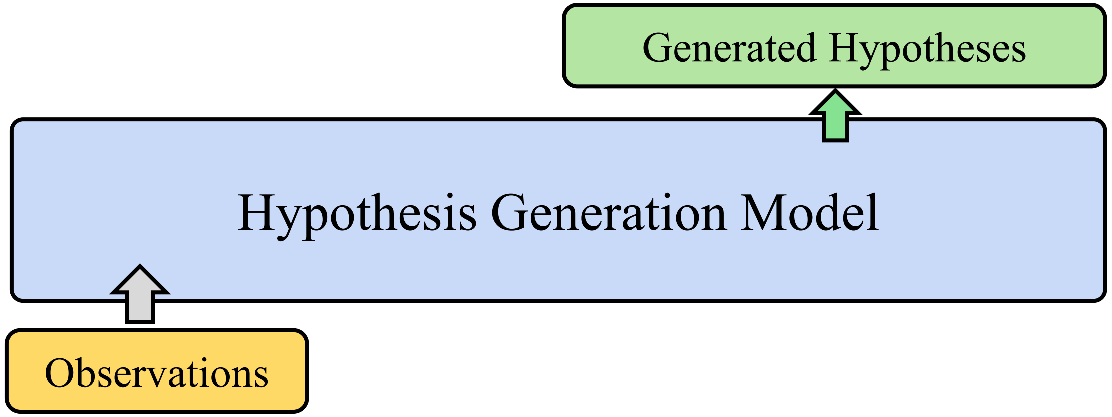
Observations (O)	Hypotheses (H)	Hypotheses Interpretations
O_1 = {Grant Heslov, Jason Segel, Robert Towne, Ronald Bass, Rashida Jones}	$H_1 = V_? : Occupation(V_?, Actor) \wedge Occupation(V_?, ScreenWriter) \wedge BornIn(V_?, LosAngeles)$	The actors and screenwriters born in Los Angeles
O_2 = {Ipad 1st Gen, Ipod touch 4th Gen, Apple TV 1st Gen}	$H_2 = V_? : Brand(V_?, Apple) \wedge ReleaseYear(V_?, 2010) \wedge \neg Type(V_?, Phone)$	The Apple products released in 2010 that are not phones
O_3 = {Covid-19, Seasonal Flu, Dysmenorrhea}	$H_3 = V_?, \exists V_1 : HaveSymptom(V_?, V_1) \wedge RelievedBy(V_1, Panadol)$	The disease whose symptoms can be relieved by Panadol

Complex Logical Hypothesis Generation

Step 1:
Sample observation-hypothesis pairs.



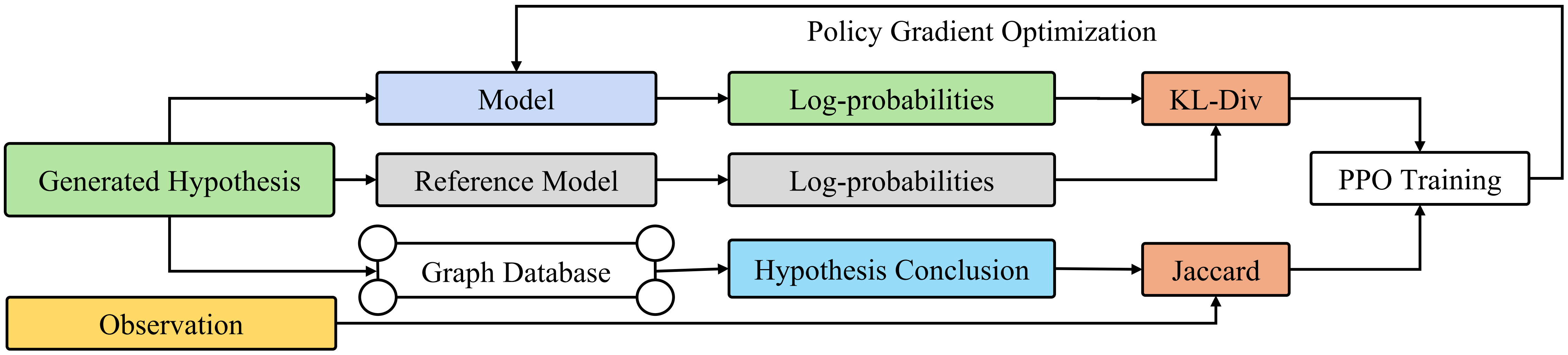
Step 2:
Train hypothesis generation model by using teacher forcing.



$$L = \log \rho(h|o) = \sum_{i=1}^{n} \log \rho(h_i|o, h_1, h_2, \dots, h_{i-1})$$

Complex Logical Hypothesis Generation

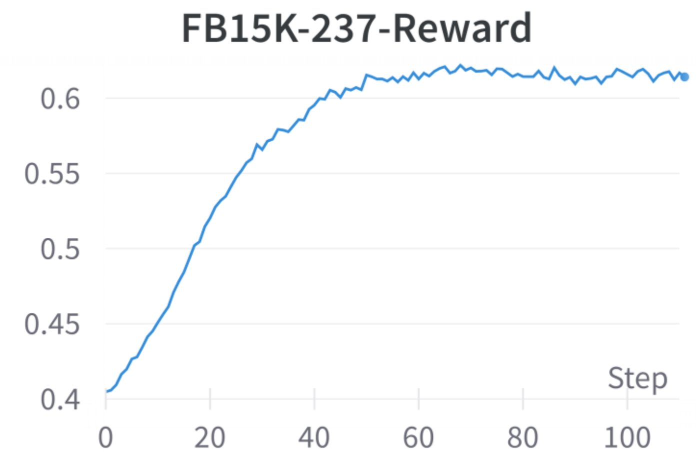
Step 3:
Optimize hypothesis generation model with Reinforcement Learning From Knowledge Graph feedback (RLF-KG).

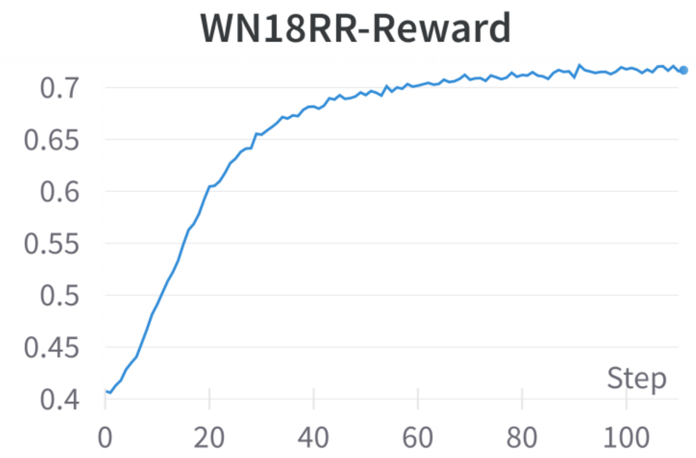


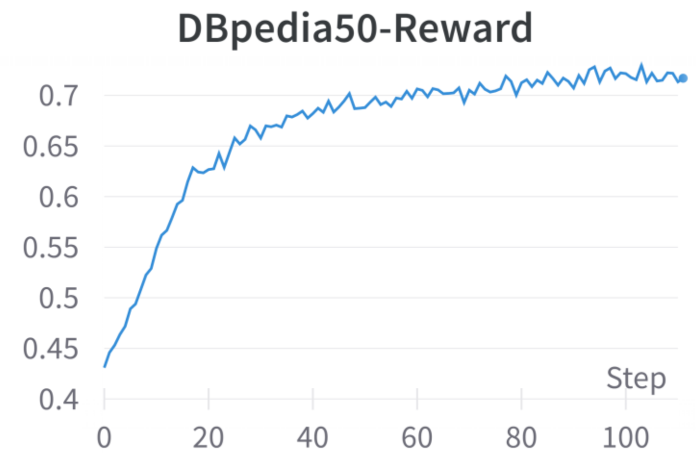
$$r(h,o) = \text{Jaccard}([\text{H}]_G, \text{O})$$

$$E_{o\sim D, h\sim\pi(\cdot|o)}\left[r(h,o) - \beta \log \frac{\pi(h|o)}{\rho(h|o)}\right]$$

Performance

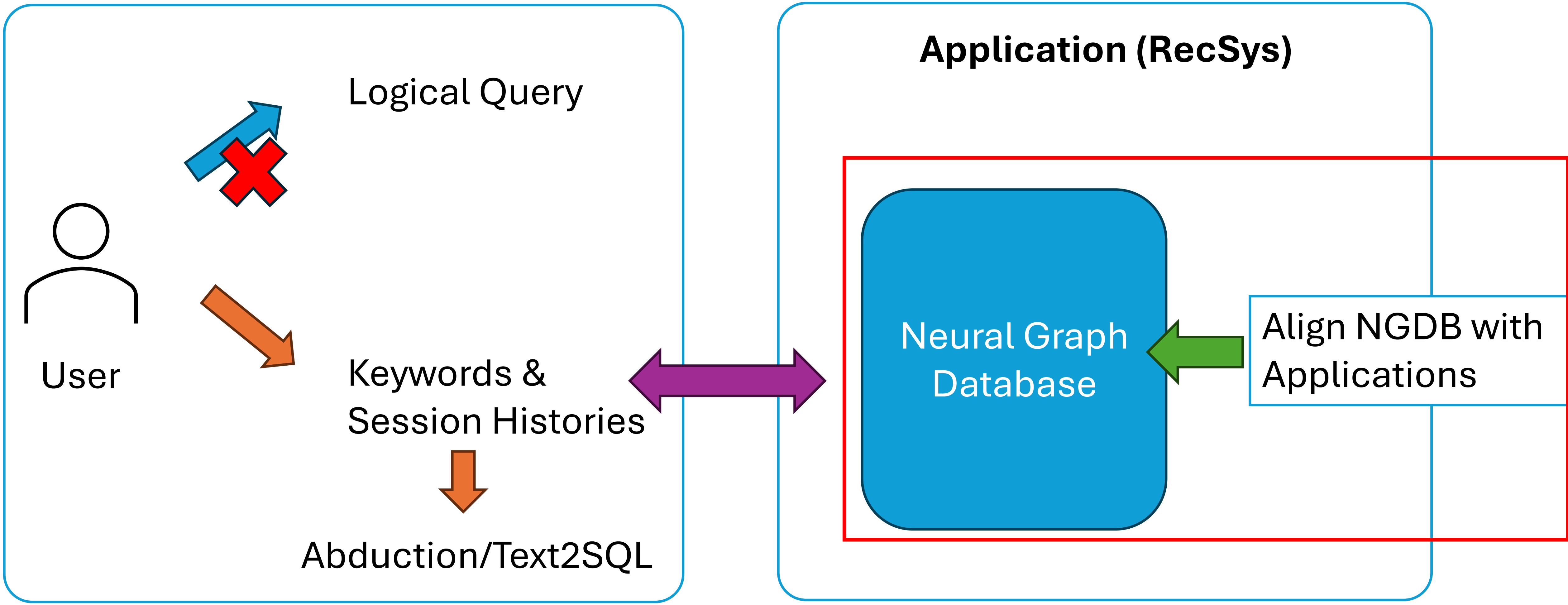






Dataset	Model	1p	2p	2i	3i	ip	pi	2u	up	2in	3in	pni	pin	inp	Ave.
FB15k-237	Enc.-Dec.	0.626	0.617	0.551	0.513	0.576	0.493	0.818	0.613	0.532	0.451	0.499	0.529	0.533	0.565
	+ RLF-KG	**0.855**	**0.711**	**0.661**	**0.595**	**0.715**	**0.608**	**0.776**	**0.698**	**0.670**	**0.530**	**0.617**	**0.590**	**0.637**	**0.666**
	Dec.-Only	0.666	0.643	0.593	0.554	0.612	0.533	0.807	0.638	0.588	0.503	0.549	0.559	0.564	0.601
	+ RLF-KG	**0.789**	**0.681**	**0.656**	**0.605**	**0.683**	**0.600**	**0.817**	**0.672**	**0.672**	**0.560**	**0.627**	**0.596**	**0.626**	**0.660**

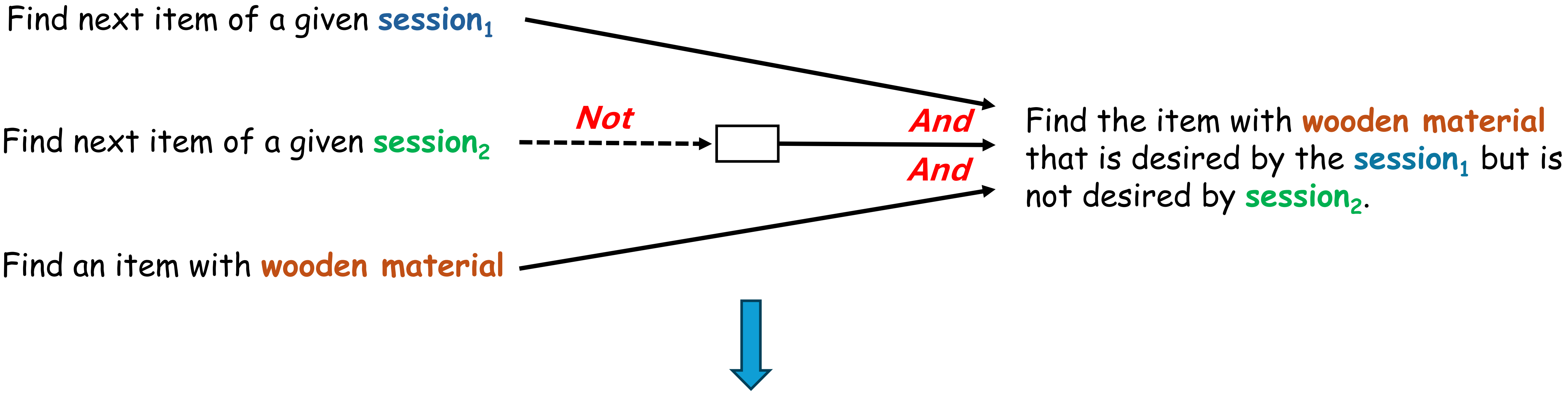
NGDB Application on E-commerce



Aligning NGDB to RecSys: Sessions, Attributes, and Logics

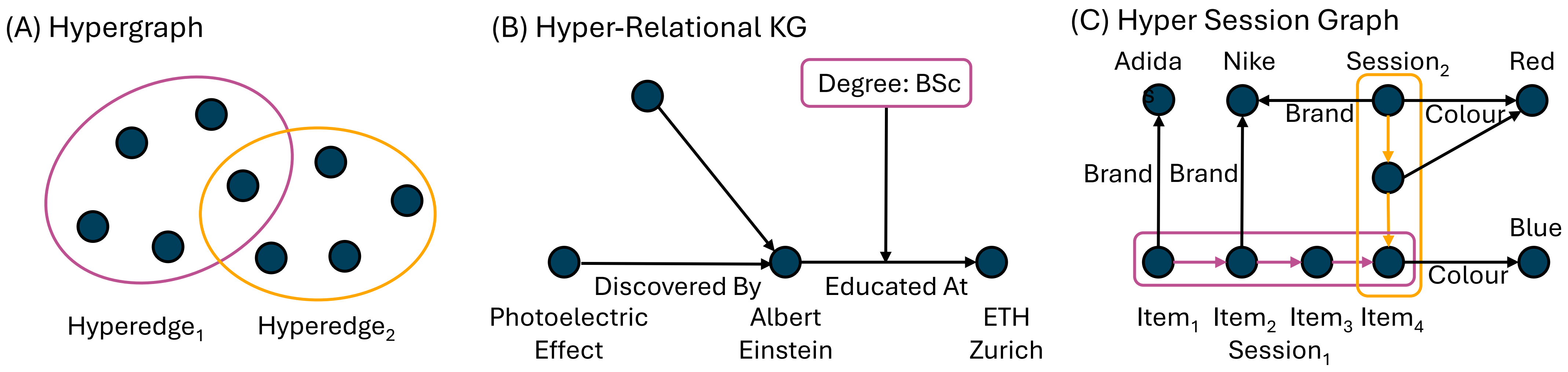
For **Product** Recommendation with **Negation**

Find the item with wooden material that is desired by the $session_1$ but is not desired by $session_2$.



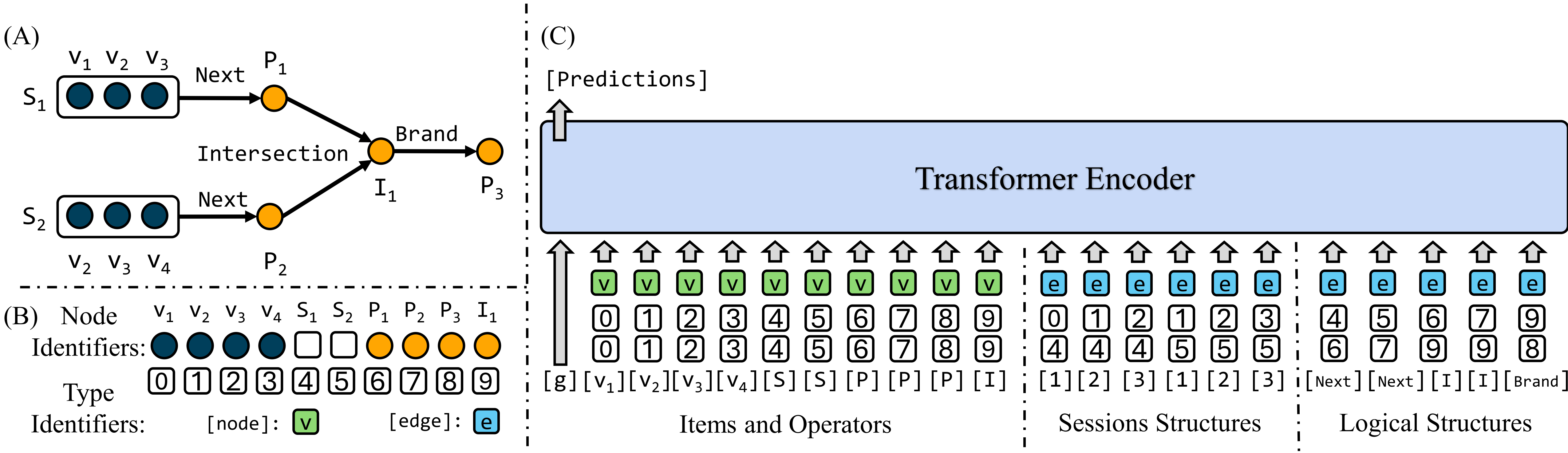
$$q = V_? : Session(Item_{1,1}, Item_{1,2}, \ldots, Item_{1,\mathrm{m}}, V_?) \land \neg Session(Item_{2,1}, Item_{2,2}, \ldots, Item_{2,\mathrm{n}}, V_?) \land Material(V_?, Wood)$$

Logical Session-CQA on Hypergraph



- Sessions are treated as hyperedges.
- Other attributes and relations are treated as binary edges.

Logical-Session Graph Transformer



- Permutation Invariant to Logical Operators.
- Sensitive to Item-order in Session.

Experiment Dataset

	Train Graph		Validation Graph		Test Graph					
Dataset	Vertices	Edges	Vertices	Edges	Vertices	Edges	#Sessions	#Items	#Values	#Relations
Amazon	2,258,179	7,234,680	2,345,475	7,620,527	**2,431,747**	**8,004,984**	720,816	431,036	1,279,895	10
Diginetica	257,018	1,286,384	261,996	1,337,628	**266,897**	**1,387,861**	12,047	134,904	125,204	3
Dressipi	611,520	2,435,932	643,140	2,567,128	**674,853**	**2,698,692**	668,650	23,618	903	74

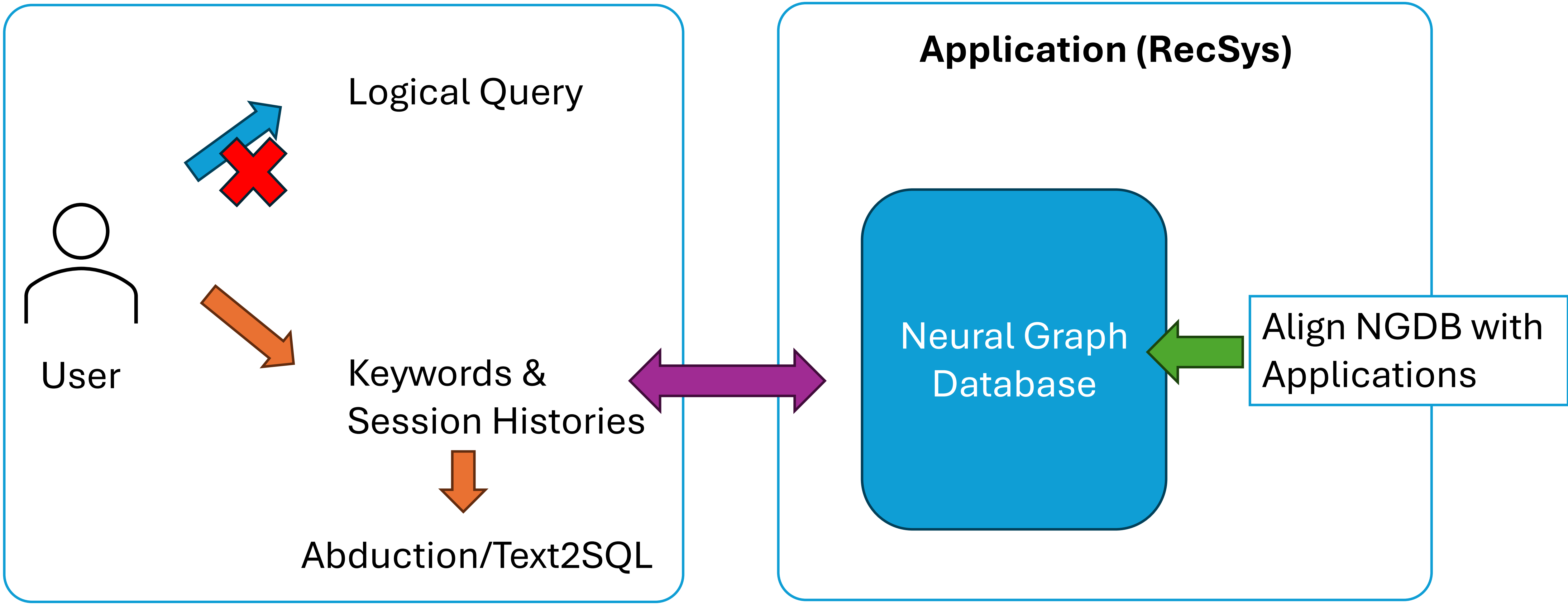
	Train Queries		Validation Queries	Test Queries
Dataset	Item-Attributes	Others	All Types	All Types
Amazon	2,535,506	720,816	36,041	36,041
Diginetica	249,562	60,235	3,012	3,012
Dressipi	414,083	668,650	33,433	33,433

Experiment Results

Dataset	Query Encoder	Session Encoder	Average-EPFO	Average-Negation
Amazon	FuzzQE	GRURec	30.94	21.03
		SRGNN	31.75	22.26
		Attn-Mixer	31.68	25.00
	Q2P	GRURec	23.09	14.61
		SRGNN	24.59	16.62
		Attn-Mixer	28.16	26.95
	NQE	-	23.19	18.12
	SQE-Transformer	-	30.07	27.16
	SQE-LSTM	-	32.53	27.13
	LSGT (Ours)		**33.26**	**29.69**

Existential Positive First Order (EPFO): queries types that involves conjunction, disjunction, and variables are existential quantified. No negations.

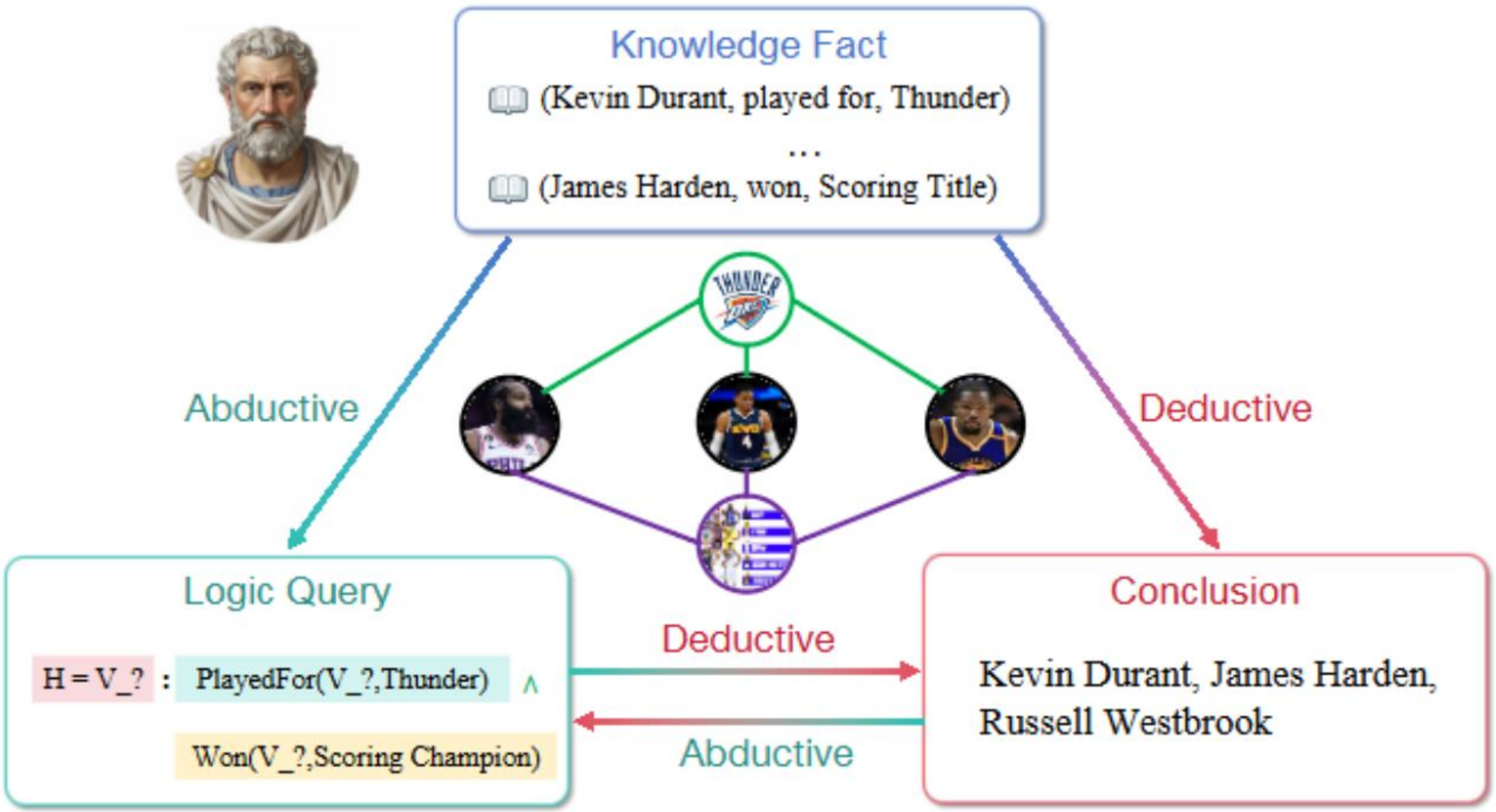
Summary:
NGDB Application on E-commerce



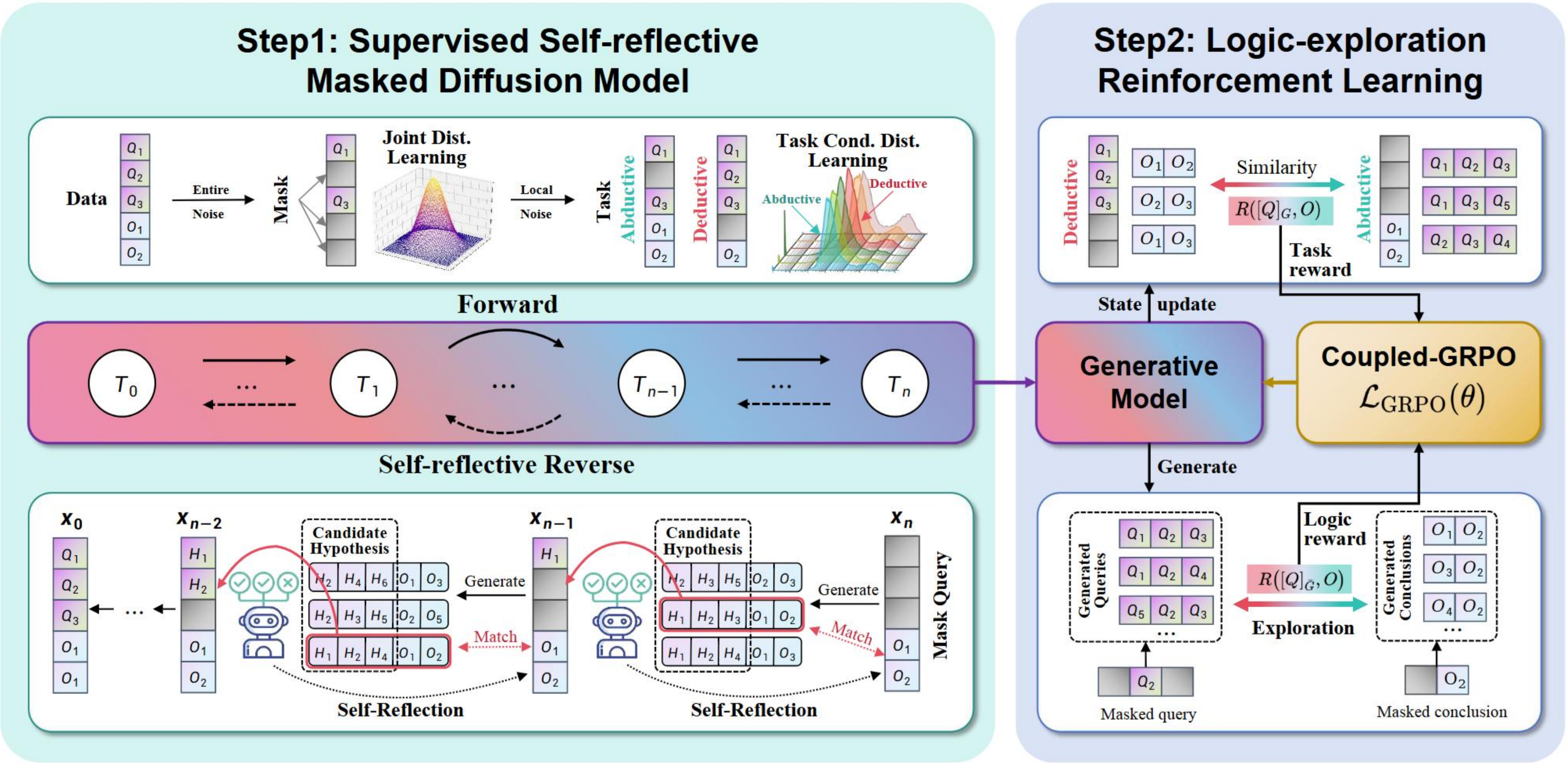
Abduction (ACL'24): Query Generation

Deduction (KDD'24): Neural Query Execution

Unifying Deduction and Abduction? (new)



Unifying Deduction and Abduction? (new)



[1] Unifying Deductive and Abductive Reasoning in Knowledge Graphs with Masked Diffusion Model
Yisen Gao, Jiaxin Bai, Yi Huang, Xingcheng Fu, Qingyun Sun, Yangqiu Song; TheWebConf'26

This Talk – Towards NGDB

🏗 **Data:** Autonomous Structured Knowledge Construction

🧠 **Algorithm & System:** Comprehensive Neural Reasoning

🤖 **Application:** Agentic Knowledge Integration

Summary

Data

Autonomous Structured Knowledge Construction

AutoschemaKG:

Web-scale General Knowledge Graph Construction

Relational Intention Graph:

Domain Intention KG for E-Commerce

Algorithm & System

Comprehensive Neural Reasoning

Query2Particles:

Diverse Complex Query Answering

NRN:

Number Query Answering

Memory Enhanced Query Encoding:

Query Answering on Event-based Graph

Application

Agentic Knowledge Integration

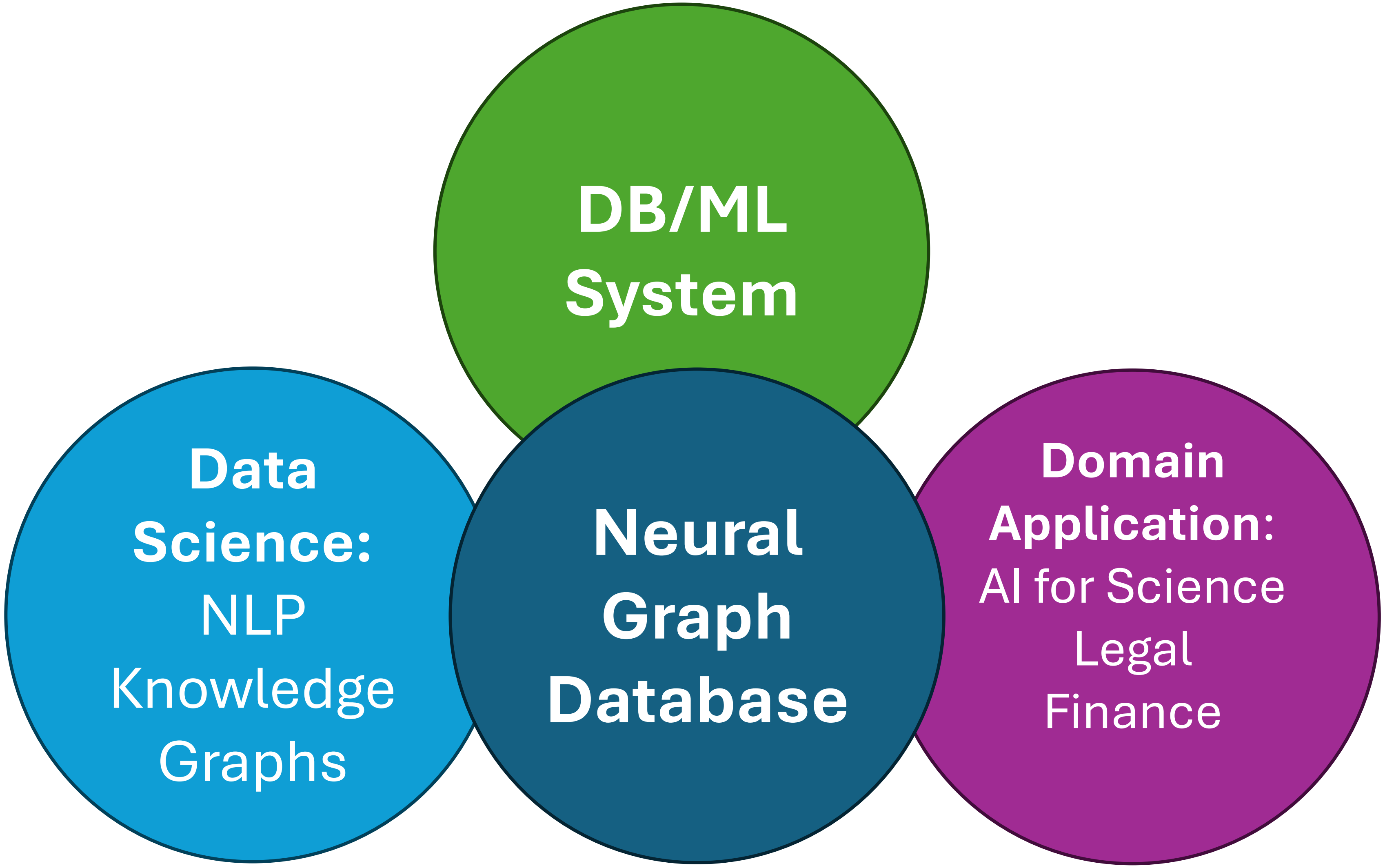
Abductive Reasoning:

Complex Logical Hypothesis Generation

Intention Understanding from Sessions:

Logical Session Graph Transformer

Future Research



Future Direction – NGDB-RAG

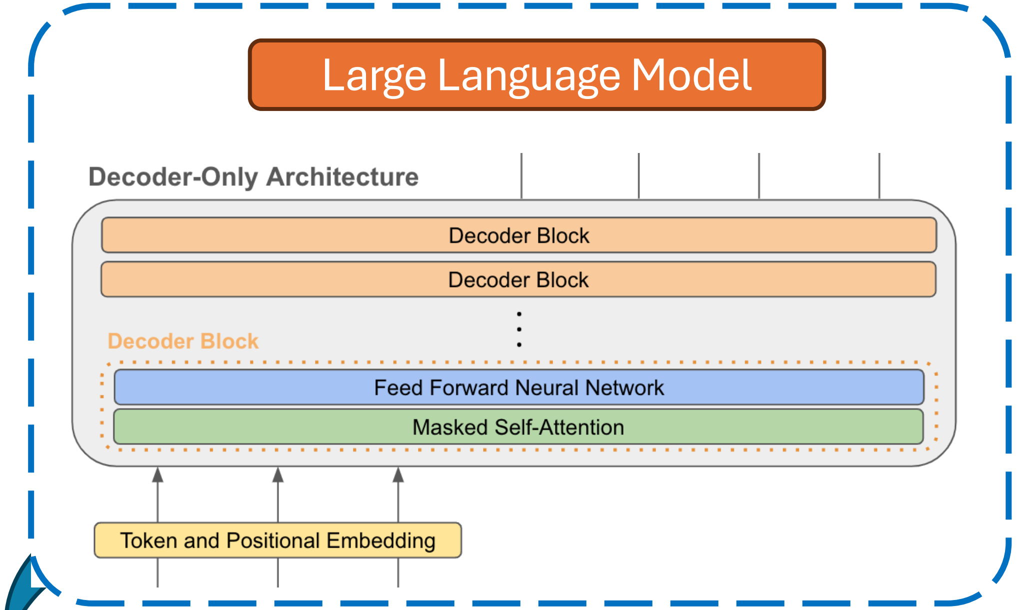
Neural Graph Database Reasoning:
Where are the Turing Award Winners born before 1940 lives in?

$$V_?. \exists V_1, X_2: Win(V_1, Turing\ Award) \wedge LessThan(X_2, 1940) \wedge BornIn(V_1, X_2) \wedge Livein(V_1, V_?)$$

Language Modeling via Next Token Prediction:
We would like to invite esteemed senior professors who have made significant contributions to computer science to give a talk on-site. To accommodate them, we have decided to hold the seminar in ________

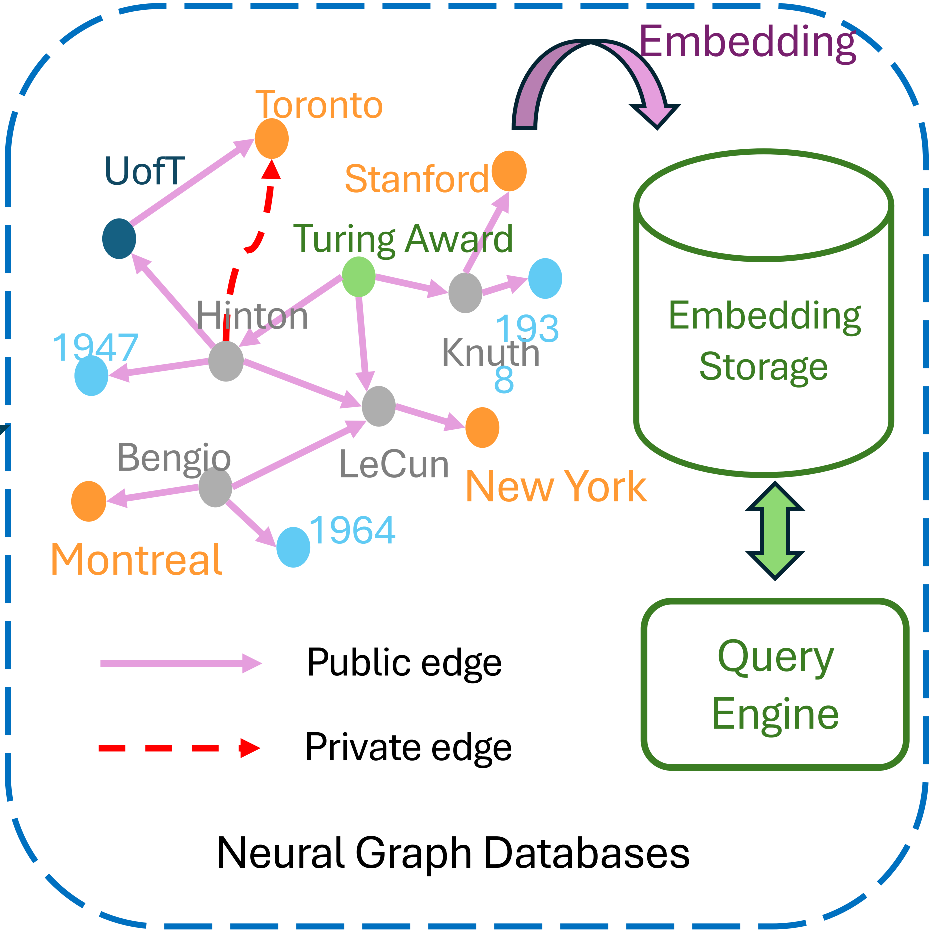
Task 1

Generating Good NGDB Queries (Abduction).



Task 2

Improving NGDB Reasoning (Deduction).



Task 3

Incorporating NGDB Results.

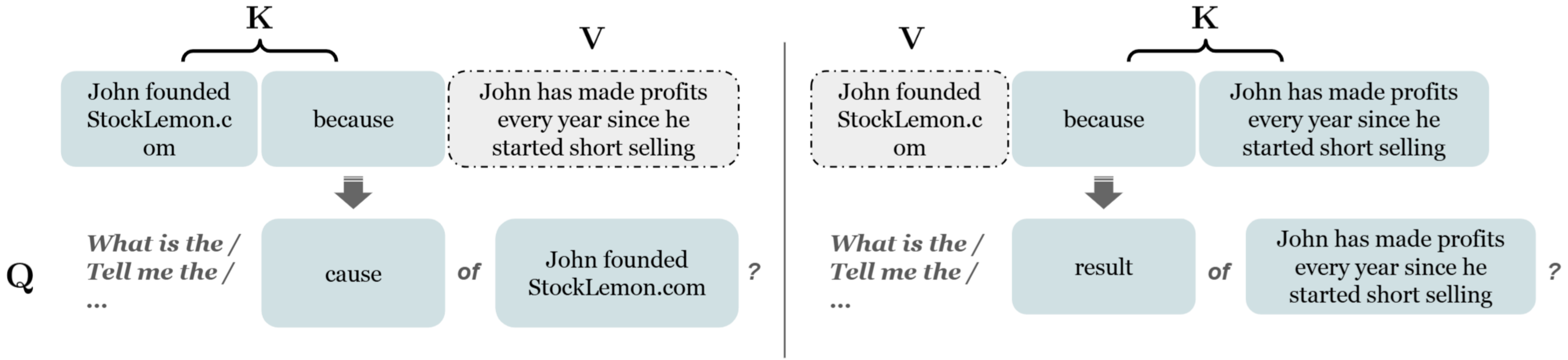
Task 4

...... we have decided to hold the seminar in __**Toronto**__.

Application of NGDB-RAG

AtlasKV: Injecting Triples as KVs to LLMs (new)

KG2KV converts KG triples into Q-K-V data

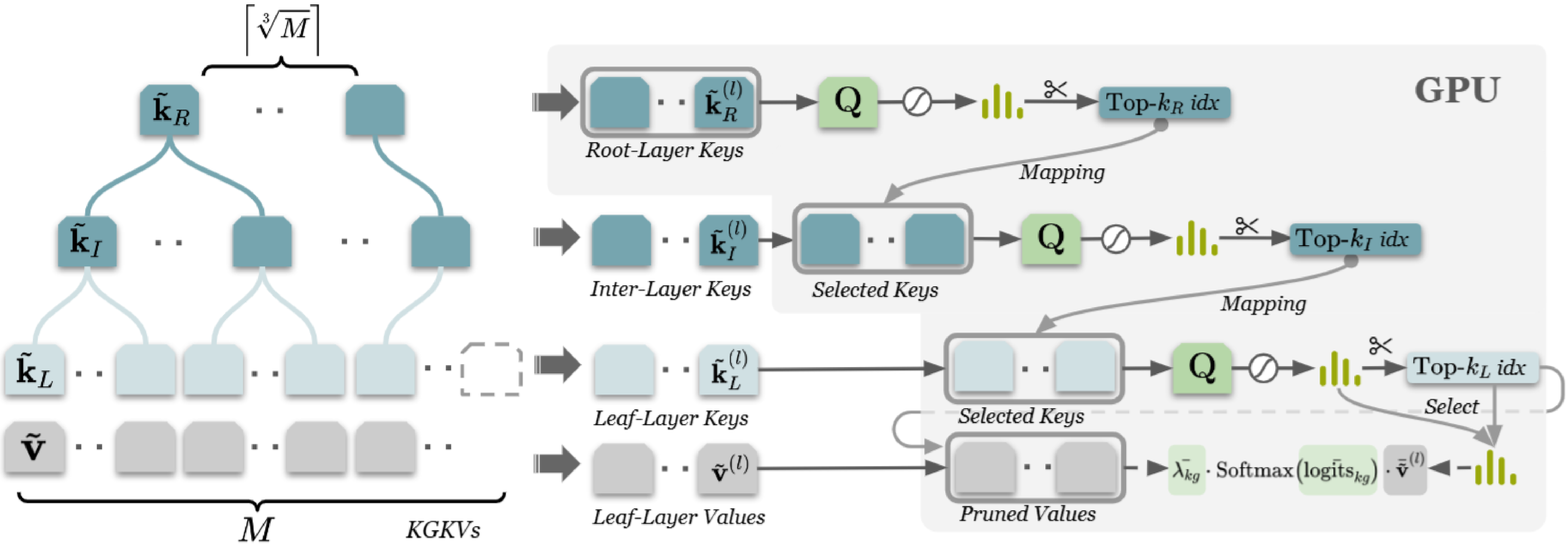


Tail-Masked Triple Transformation *Head-Masked Triple Transformation*

[1] AtlasKV: Augmenting LLMs with Billion-Scale Knowledge Graphs in 20GB VRAM, Haoyu Huang, Hong Ting Tsang, Jiaxin Bai, Xi Peng, Gong Zhang, Yangqiu Song: ICLR'26

AtlasKV: Injecting Triples as KVs to LLMs (new)

HiKVP: a hierarchical key-value pruning algorithm

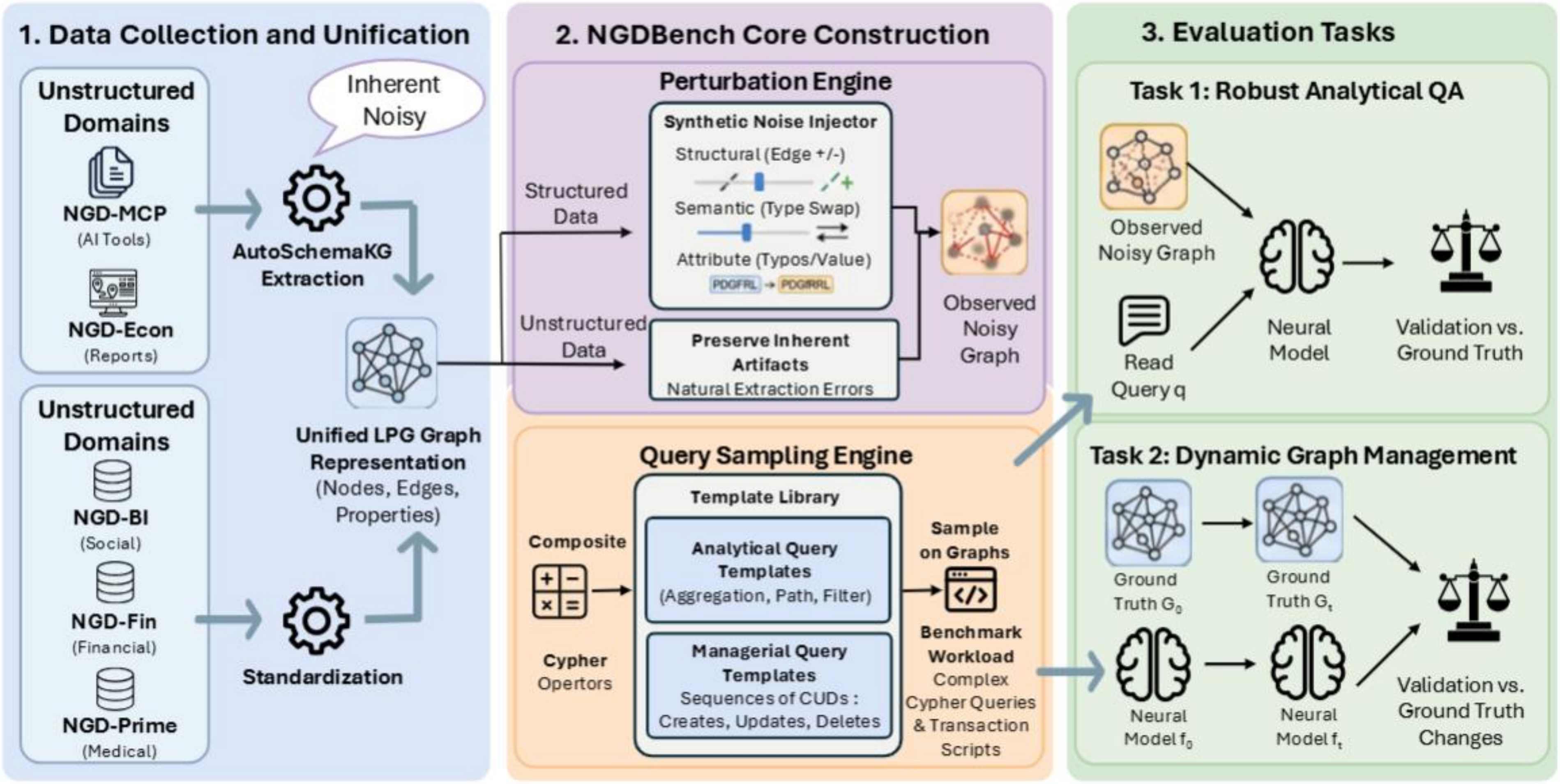


[1] AtlasKV: Augmenting LLMs with Billion-Scale Knowledge Graphs in 20GB VRAM, Haoyu Huang, Hong Ting Tsang, Jiaxin Bai, Xi Peng, Gong Zhang, Yangqiu Song: ICLR'26

Future Direction – Building Full NGDB Systems

- **Beyond 'Read' clauses**:
  - Need support operators **Create**, **Read**, **Update**, and **Delete** (CRUD).
  - Potential Direction: Symbolic Knowledge Editing of Models
  - Even Further: **Predicting the next operation and transaction of DB**?

Towards Neural Graph Data Management (new)



NGDBench: Full spectrum of the Cypher query language.

Complex Reasoning: aggregations, variable-length paths, multiple target variables, and numerical constraints.

Realistic Robustness: structural and attribute-level noise.

Dynamic Management: multi-stop graph updates (Create, Delete, Set)

[1] Towards Neural Graph Data Management: *LI Yufei, Yisen Gao, Jiaxin Bai, Jiaxuan Xiong, Haoyu Huang, Zhongwei Xie, Hong Ting Tsang, Yangqiu Song*

Future Direction – Building Full NGDB Systems

- **Beyond 'Read' clauses**:
  - Need support operators **Create**, **Read**, **Update**, and **Delete** (CRUD).
  - Potential Direction: Symbolic Knowledge Editing of Models
  - Even Further: **Predicting the next operation and transaction of DB**?
- **Neural DB/ML System** architecture challenges:
  - **Concurrent** neural operations
  - **Query optimization** strategies
  - **Caching** mechanisms for frequently accessed neural patterns
  - ...

Future Direction – Multi-Modal Data

Documents, images, videos, audios, etc.

Key challenges and opportunities:

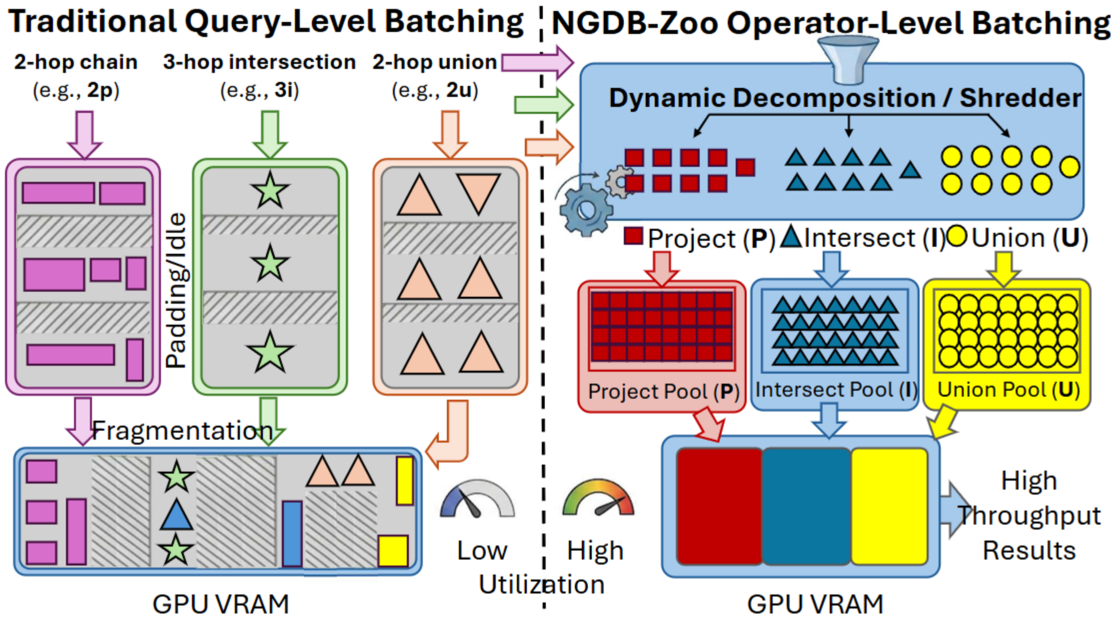
- **Unified representation** for heterogeneous data types
- **Cross-modal reasoning** capabilities modalities (e.g., "find videos showing the manufacturing process of products mentioned in this technical document")
- **Storage and retrieval systems** for high-dimensional multi-modal embeddings

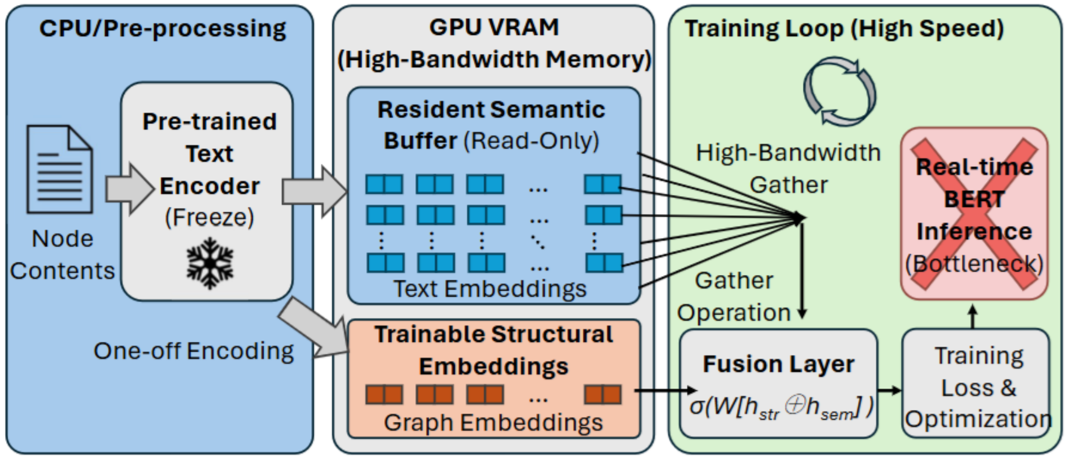
Efficient and Scalable NGDB Training (new)

- NGDB-Zoo: A unified framework for scalable Neural Graph Database training.

- Shifts from rigid query batching to **dynamic operator-level execution** and **decoupled semantic caching** to eliminate I/O bottlenecks.

- Achieves **1.8×–6.8× higher throughput** on massive graphs like *ogbl-wikikg2* while integrating rich text priors.





[1] NGDB-Zoo: Towards Efficient and Scalable Neural Graph Databases Training
Zhongwei Xie, Jiaxin Bai, Shujie Liu, Haoyu Huang, Yufei Li, Yisen Gao, Hong Ting Tsang, Yangqiu Song

Thank you for your attention! ☺

https://bjx.fun