Reasoning in Knowledge Graphs Using Embeddings

Hongyu Ren

Stanford University

Joint work with W. Hu, H. Dai, B. Dai, X. Chen, M. Yasunaga, H. Sun, D. Schuurmans, D. Zhou, A. Mousavi, T. Rekatsinas, I. Ilyas, J. Leskovec

Stanford | Data Science

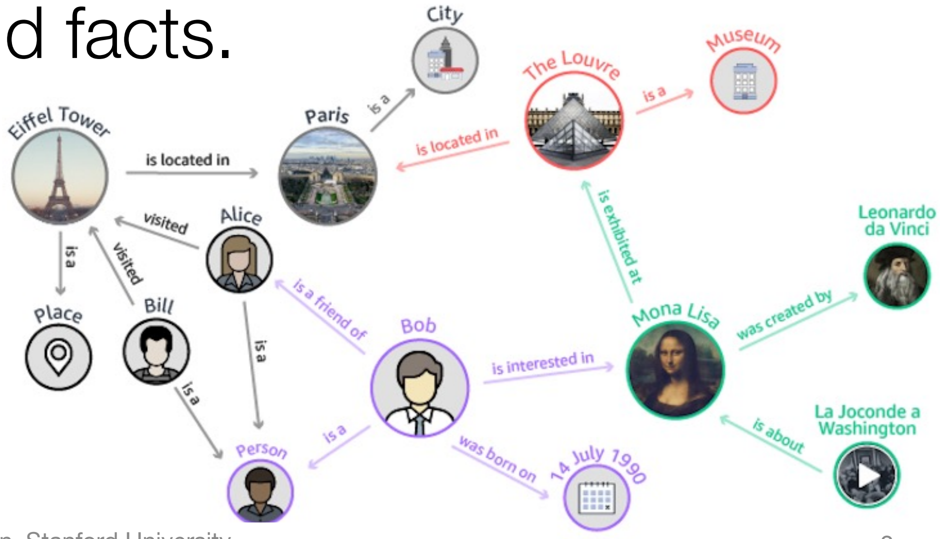
PyG

Knowledge Graphs

- Knowledge Graphs are heterogenous graphs
  - $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R})$
  - $\mathcal{V}$: node set; $\mathcal{E}$: edge set; $\mathcal{R}$: relation set;
  - Each edge $e \in \mathcal{E}$ represents a fact (v_s, r, v_o) over subject v_s and object $v_o \in \mathcal{V}$ with relation $r \in \mathcal{R}$.
  - KGs capture real world facts.
- Examples:

WIKIDATA

Freebase™

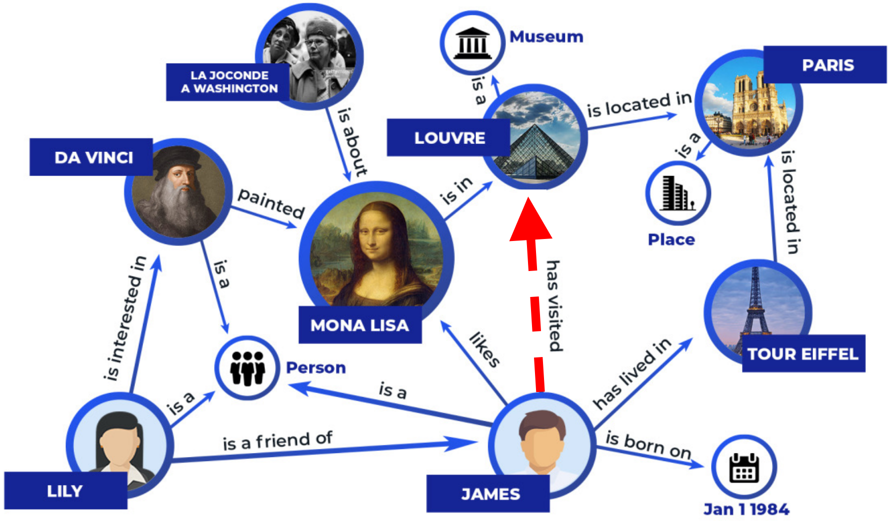


Traditional Tasks

Knowledge Graph Completion

Given the existing edge set $\mathcal{E}$, predict the missing object for a given triplet (v_s, r, v_t)

Example: Is Louvre an object to $(\text{James}, \text{Visited}, ?)$



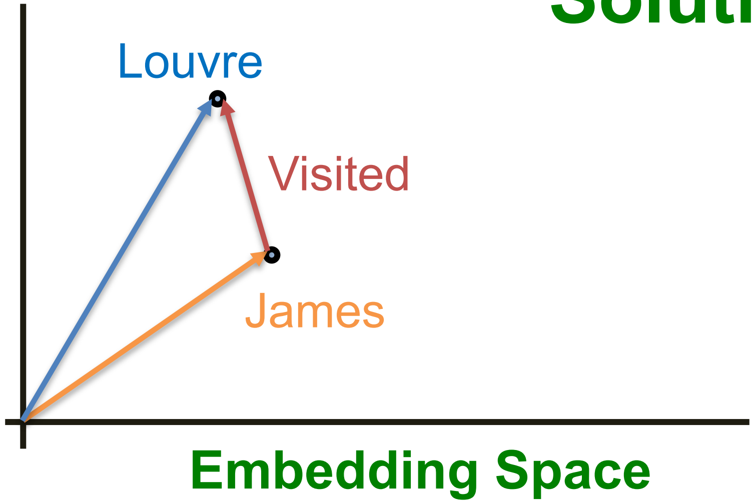
Traditional Tasks

Knowledge Graph Completion

Given the existing edge set $\mathcal{E}$, predict the missing object for a given triplet (v_s, r, v_t)

Example: Is Louvre an object to $(\mathrm{James}, \mathrm{Visited}, ?)$

Solution: TransE



TransE learns an entity embedding matrix and a relation embedding matrix, such that given a triplet $e = (v_s, r, v_t) \in \mathcal{E}$, we have $\mathbf{v}_s + \mathbf{r} = \mathbf{v}_t$, where $\mathbf{v}_s, \mathbf{r}, \mathbf{v}_t \in \mathbb{R}^d$.

[Translating Embeddings for Modeling Multi-relational Data. Bordes, et al., NeurIPS 2013]

Our Task: Multi-hop Reasoning

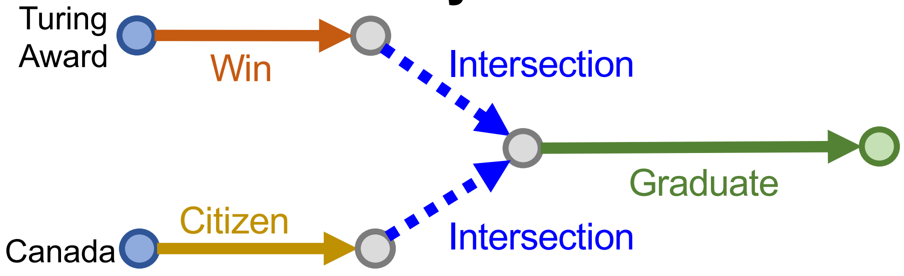
Answering Acyclic Conjunctive Queries

- A subset of **first-order logic** with existential quantifier (∃), conjunction (∧)

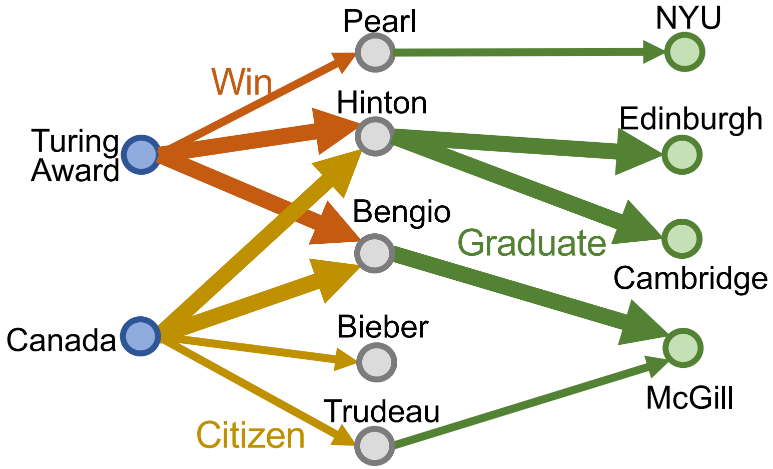
“Where did all Canadian citizens with Turing Award graduate?”

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

Query Plan



Knowledge Graph



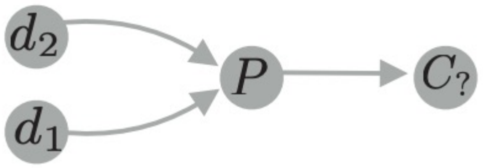
Output: **{Edinburgh, Cambridge, McGill}**

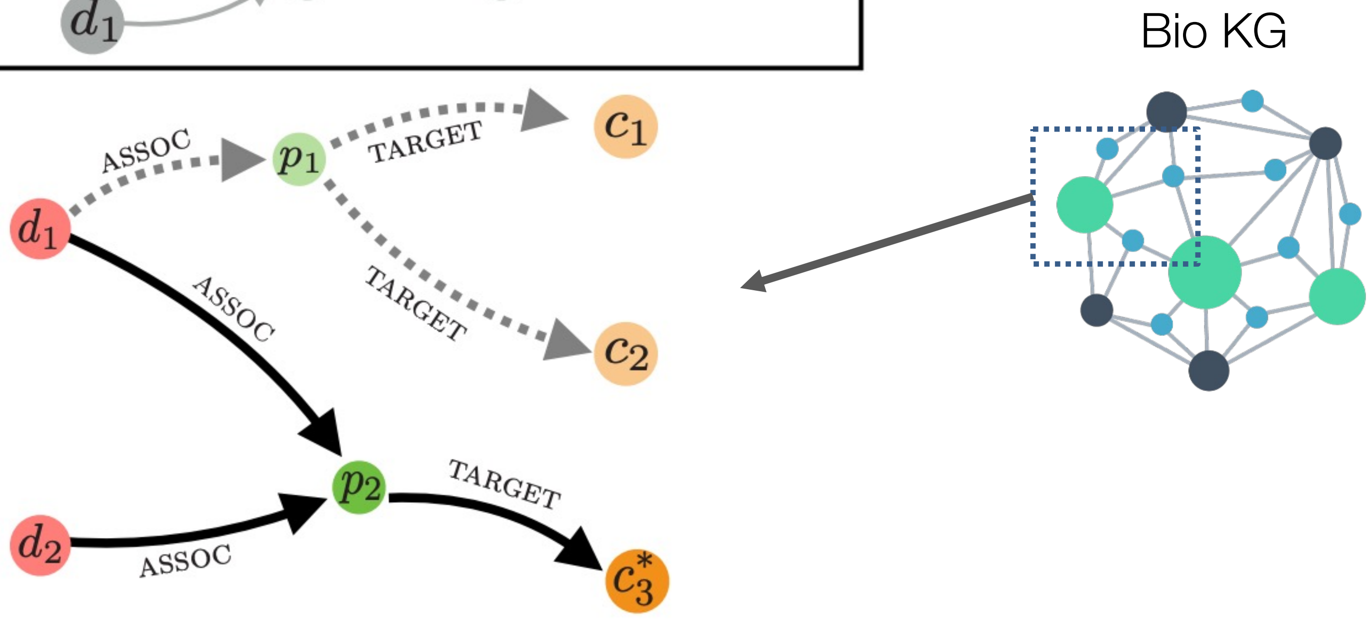
Examples

Applications in biomedicine:

$C_?.\exists P : \text{ASSOC}(d_1, P) \wedge \text{ASSOC}(d_2, P) \wedge \text{TARGET}(P, C_?)$

“Predict drugs $C_?$ that might target proteins that are associated with the given disease nodes d_1 and d_2 ”





Why is it hard?

Key challenges:

Big graphs and queries can involve noisy and unobserved data!

Freebase™

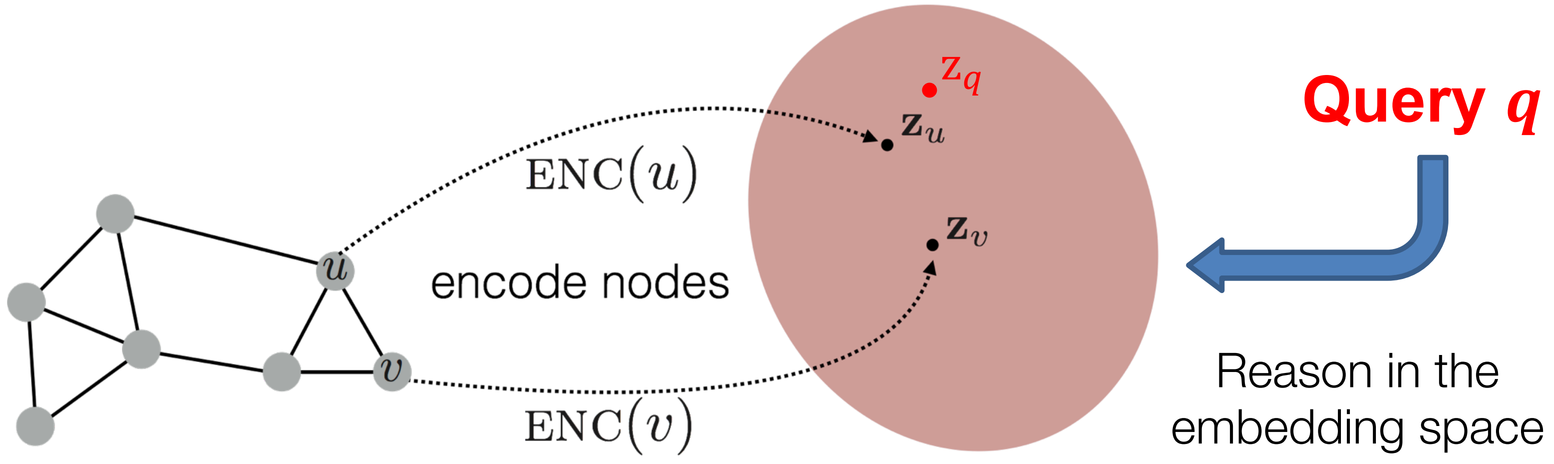
93.8% of people from Freebase have no place of birth and 78.5% have no nationality!

WIKIDATA

98.12% of movies recorded in 2020 have no information about their cost!

[Distant Supervision for Relation Extraction with an Incomplete Knowledge Base. Min et al., NAACL 2013]
[Enriching Wikidata with Linked Open Data. Zhang et al., 2022]

Our General Idea



encode nodes

Query *q*

Reason in the
embedding space

Knowledge graph

Embedding space

- **Reason in the embedding space:** Map the query and entities into embedding space such that answer entity embeddings are close to the query embedding (**contrastive learning**).
  - Low-dimensional embedding allows generalization.

[Embedding Logical Queries on Knowledge Graphs. Hamilton, et al., NeurIPS 2018]

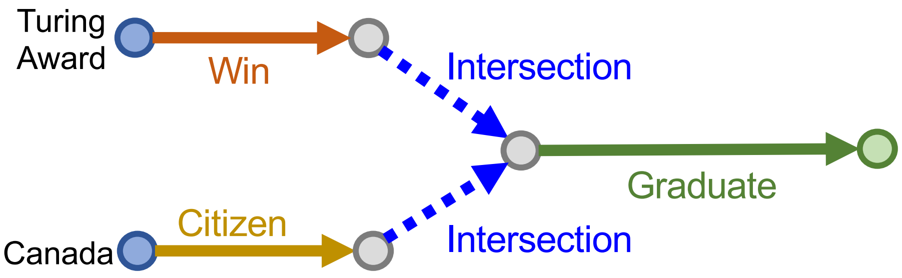
How to Embed the Query?

- Embed the query following the query plan!

“Where did all Canadian citizens with Turing Award graduate?”

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

Query Plan

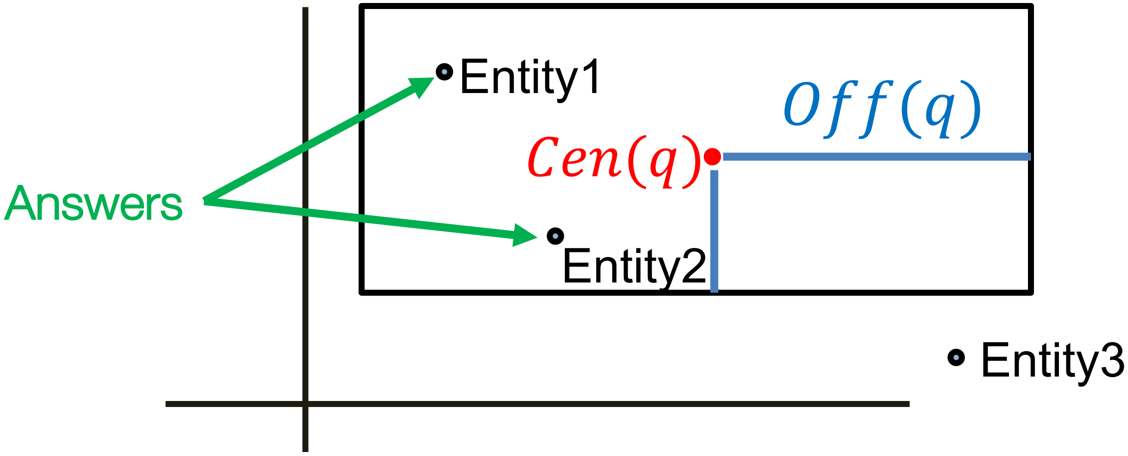


- Traversing query plan from leaves (anchors) until the root (answers)

Our Model: Query2box

Query2Box embedding:

Embed queries with hyper-rectangles (boxes): $\mathbf{q} = (Cen(q), Off(q))$.

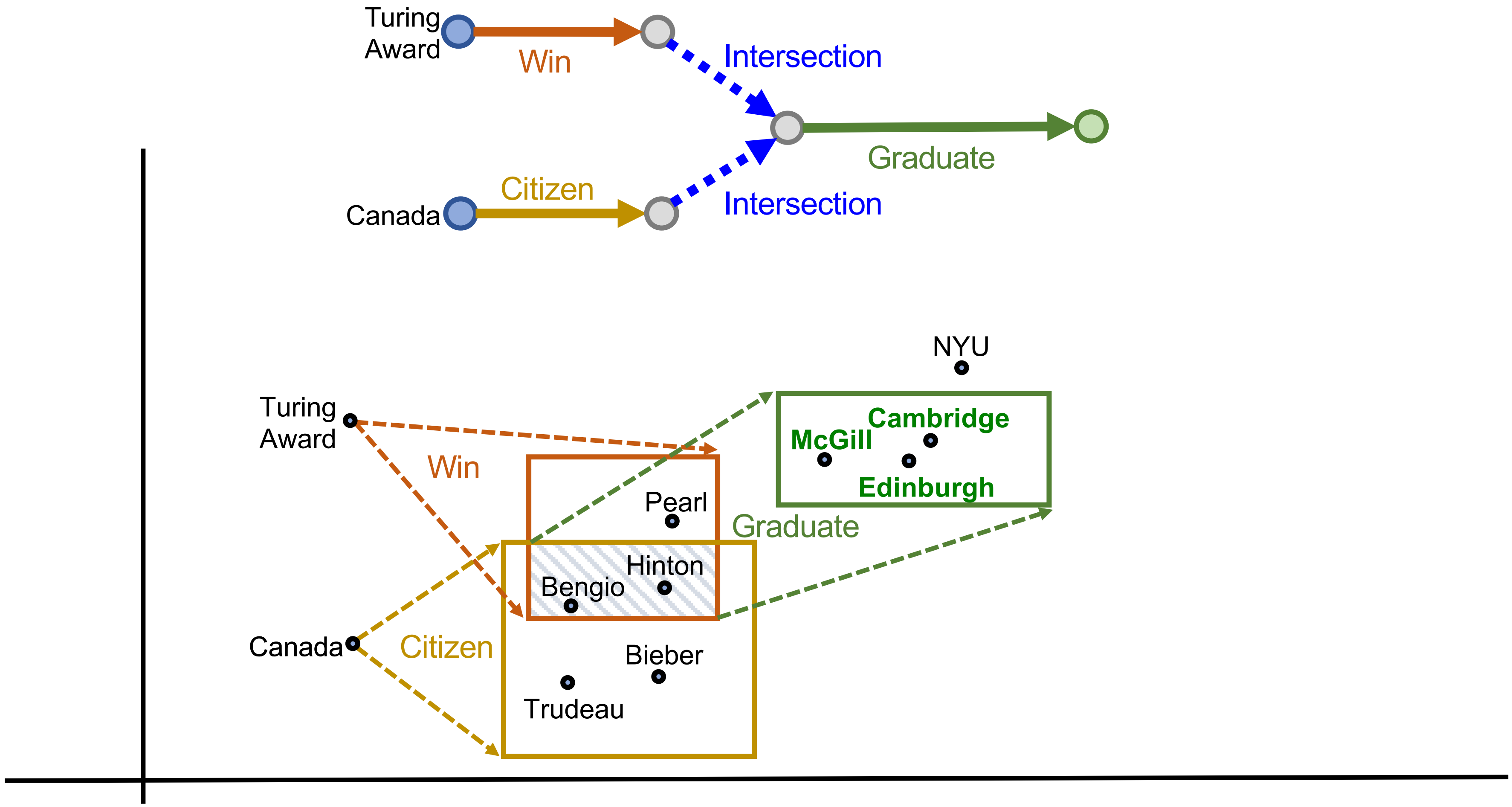


Embedding Space

[Probabilistic Embedding of Knowledge Graphs with Box Lattice Measures. Vilnis, et al., ACL 2018]

Example: Query2Box

“Where did Canadian citizens with Turing Award graduate?”



[Query2box: Reasoning over Knowledge Graphs in Vector Space Using Box Embeddings. Ren et al., ICLR 2020]

Benefits of **Query2Box**

Scalability and efficiency:

- Any query can be reduced to a couple of matrix operations and a single k-nearest neighbor search

Generality:

- We can answer any query (even those we have never seen before)

Robustness to noise:

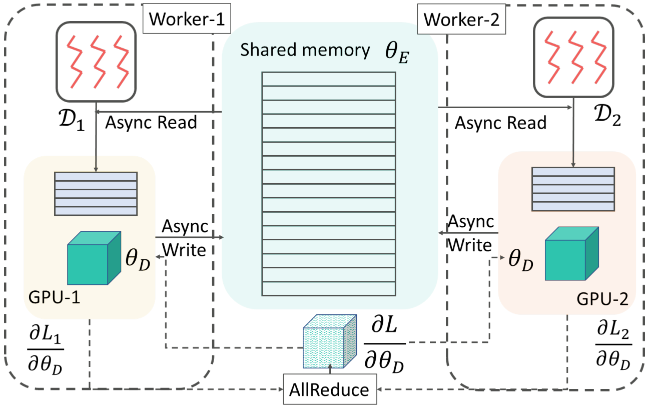
- Graph can contain missing and noisy relationships

Reasoning Framework: SMORE

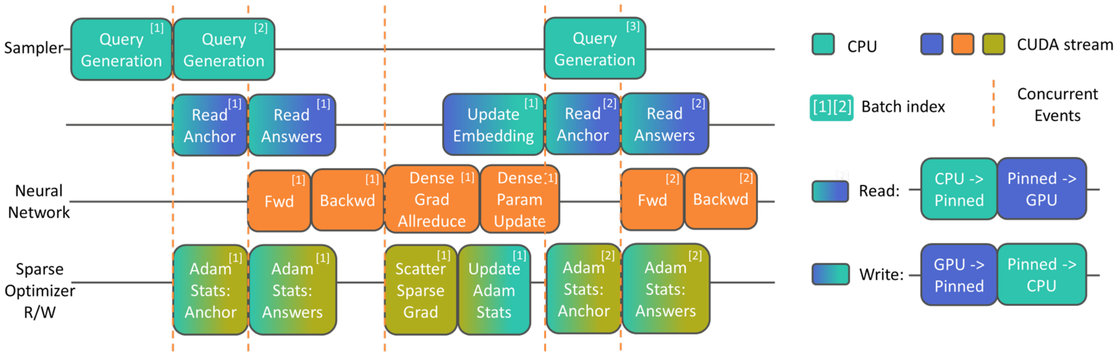
Scalable Multi-hOp REasoning framework

First framework that scales reasoning algorithms to extremely large KGs (**86m nodes, 338m edges**)

Distributed Paradigm



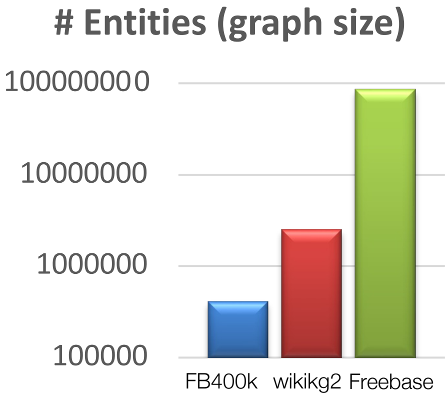
Pipelined Asynchronous Design

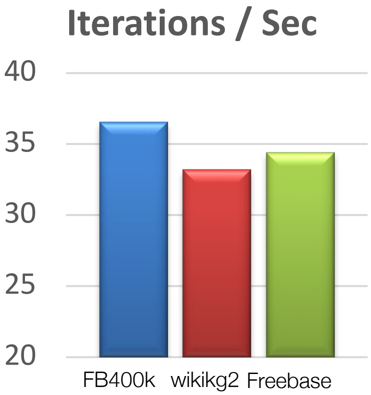


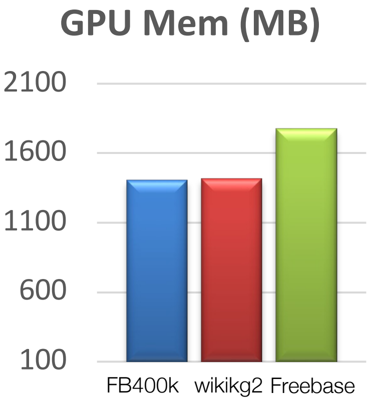
[SMORE: Knowledge Graph Completion and Multi-hop Reasoning in Massive Knowledge Graphs. Ren et al., KDD 2022]

SMORE on Massive KGs

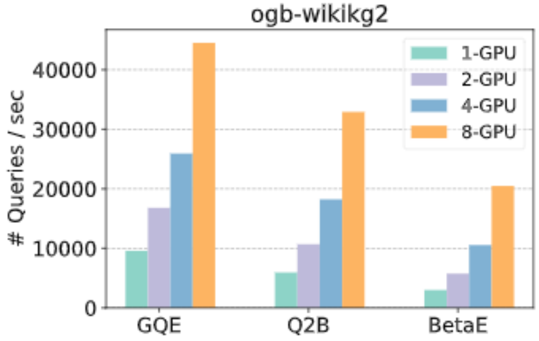
- SMORE supports 6 different link prediction + multi-hop reasoning methods
- Prior implementation **runs out of GPU** memory and time limit on large KGs
- SMORE enables (almost) graph-size agnostic speed and GPU memory usage

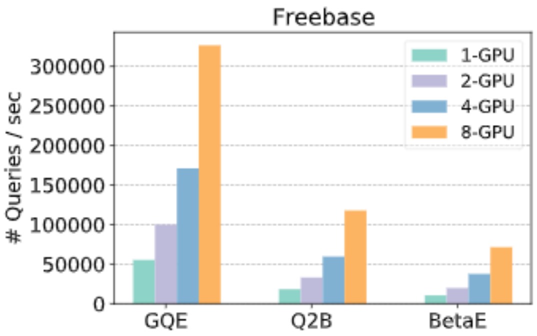






- SMORE enables almost linear multi-GPU speedup





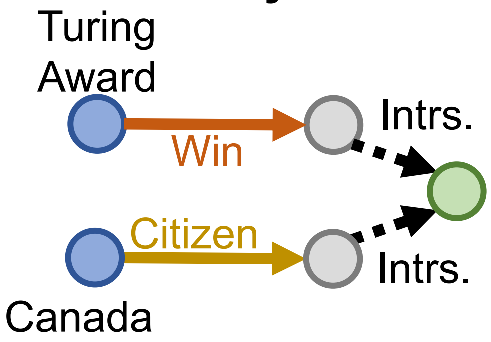
PR welcome!

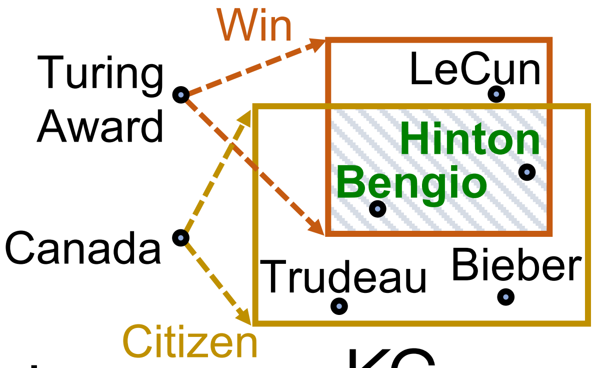
Application (1): KG Question Answering

KGQA: Answering natural language questions over massive KGs

Question → Query Plan → Box Embedding

Who are Canadian Turing Award winners?





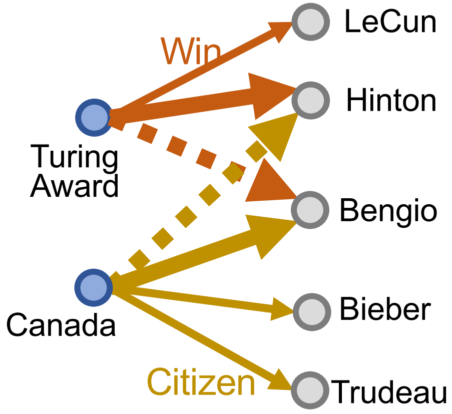
Our method: Latent Execution-Guided Reasoning

Given a question, we iteratively do

(start from an empty tree with mentioned entities as root)

- Execution-guided synthesis: synthesize one step (grow the query plan) based on the current query embedding.
- Latent query execution: execute the new step in the embedding space

KG



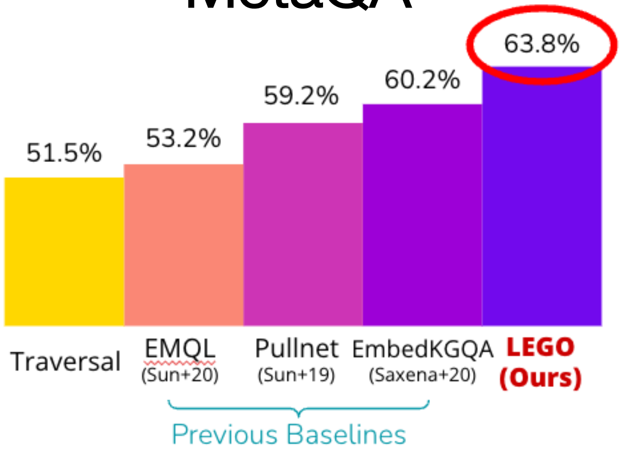
Application (1): KG Question Answering

Our method: Latent Execution-Guided Reasoning (LEGO)

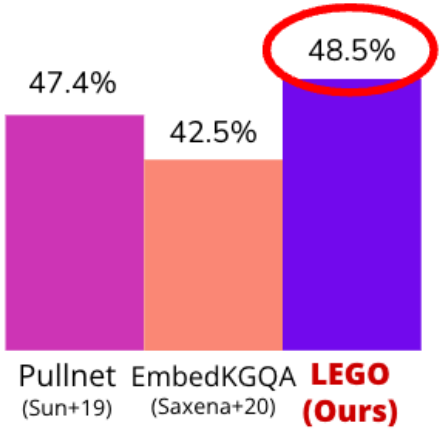
Questions	MetaQA	WebQuestion
Train	96,106	2,848
Dev	9,992	250
Test	9,947	1,639

KG	Entities	Relations	Observed Edges	Missing Edges
MetaQA	43k	18	134k	133k
WQSP	409k	1,836	2m	2m

MetaQA



WebQuestion (WQSP)

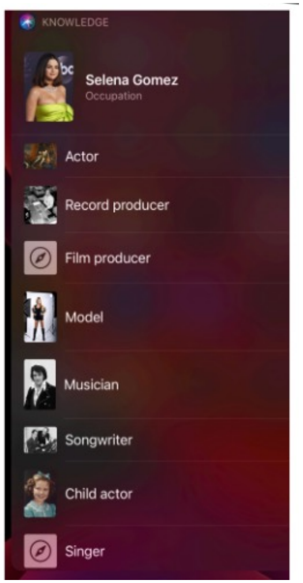


[Latent Execution-Guided Reasoning for Multi-Hop Question Answering. Ren et al., ICML 2021]

Application (2): Fact Ranking

- Facts/Answers to queries are of different importance/uncertainty to users.

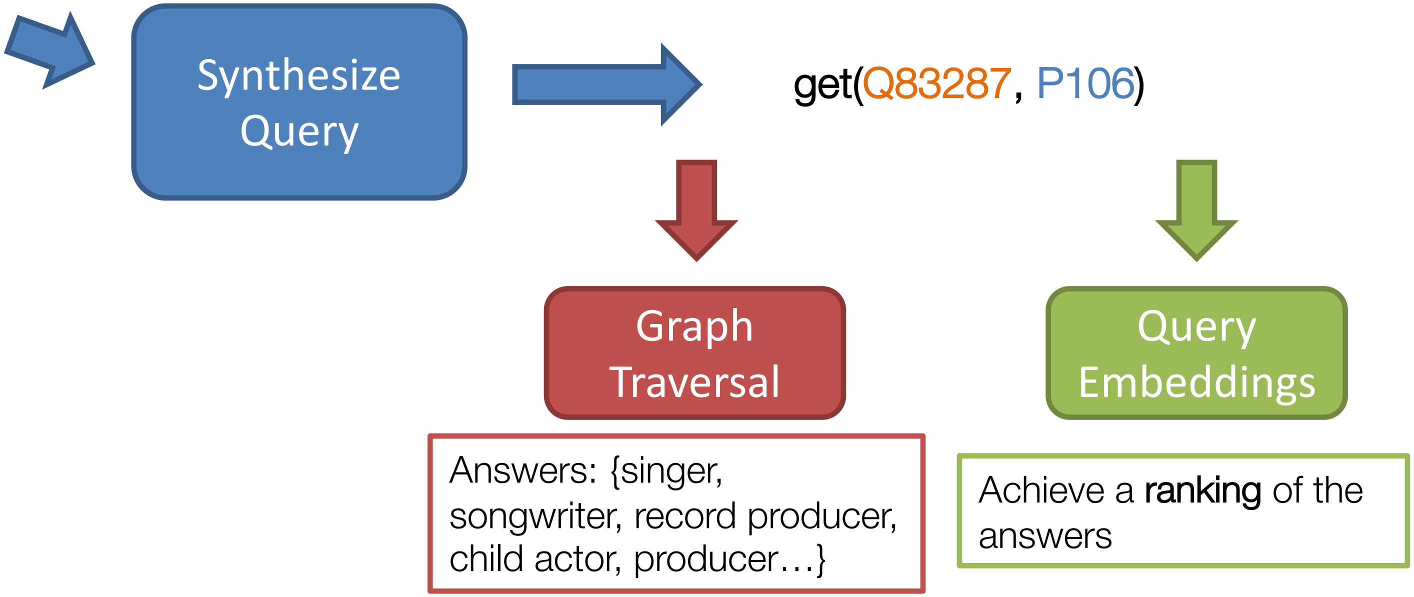
What's the occupation of Selena Gomez?



KNOWLEDGE

Selena Gomez
Occupation

- Actor
- Record producer
- Film producer
- Model
- Musician
- Songwriter
- Child actor
- Singer



Synthesize Query

get(Q83287, P106)

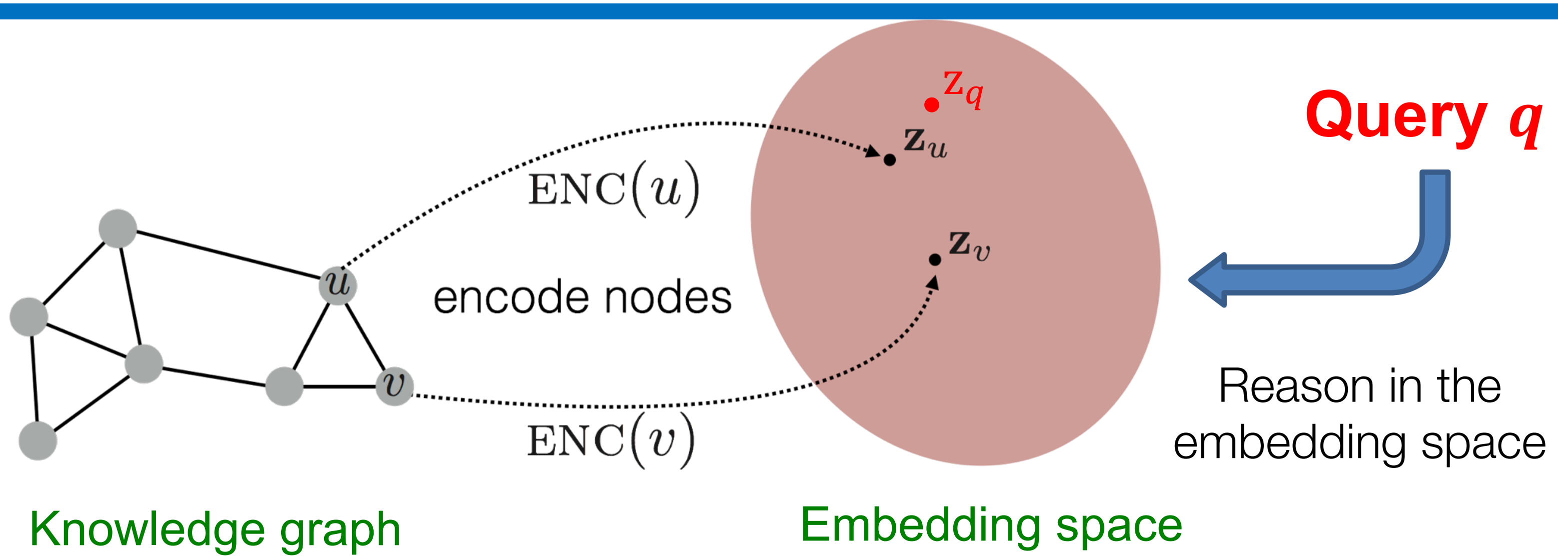
Graph Traversal

Query Embeddings

Answers: {singer, songwriter, record producer, child actor, producer…}

Achieve a **ranking** of the answers

Conclusion



- Box embeddings for answering logical queries on knowledge graphs
- Scalable query embeddings on massive KGs
- Applications towards neural query engine

Query2box: Reasoning over Knowledge Graphs in Vector Space Using Box Embeddings

Hongyu Ren*, Weihua Hu*, Jure Leskovec

ICLR 2020

Beta Embeddings for Multi-Hop Logical Reasoning in Knowledge Graphs

Hongyu Ren, Jure Leskovec

NeurIPS 2020

Latent Execution-Guided Reasoning for Multi-Hop Question Answering on Knowledge Graphs

Hongyu Ren, Hanjun Dai, Bo Dai, Xinyun Chen, Michihiro Yasunaga, Haitian Sun, Dale Schuurmans, Jure Leskovec, Denny Zhou

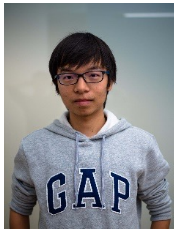
ICML 2021

SMORE: Knowledge Graph Completion and Multi-hop Reasoning in Massive Knowledge Graphs

Hongyu Ren*, Hanjun Dai*, Bo Dai, Xinyun Chen, Denny Zhou, Jure Leskovec, Dale Schuurmans

KDD 2022

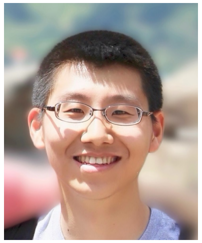
Weihua Hu



Hanjun Dai

Bo Dai

Dale Schuurmans



Michihiro Yasunaga

Xinyun Chen

Denny Zhou

Jure Leskovec