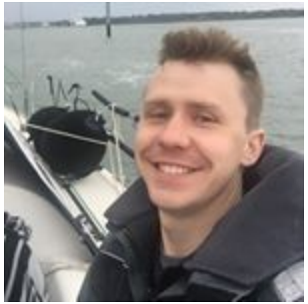
Neural Databases

From Natural Language Processing to Neural Databases

SAPIENZA
UNIVERSITÀ DI ROMA

Fabrizio Silvestri



James Thorne
Cambridge

Majid Yazdani
Facebook AI

Marzieh Saeidi
Facebook AI

Sebastian Riedel
Facebook AI

Alon Halevy
Facebook AI

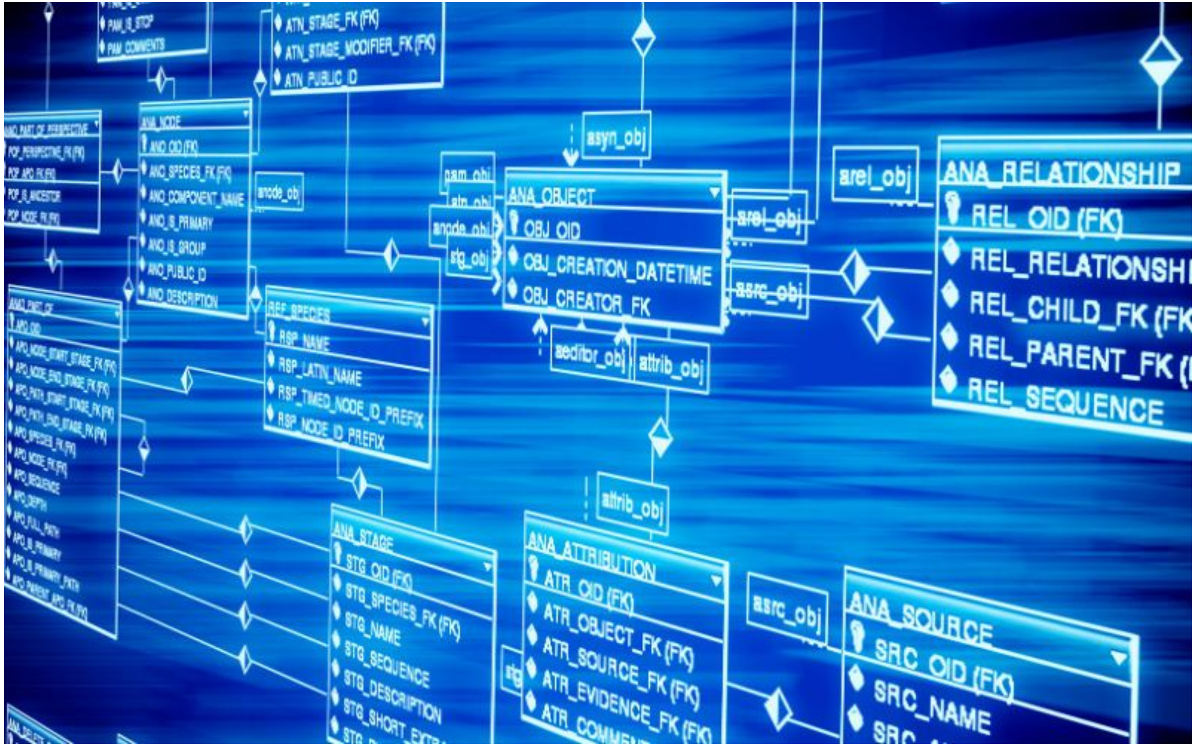
- *J. Thorne, M. Yazdani, M. Saeidi, F. Silvestri, S. Riedel, and A. Halevy. From natural language processing to neural databases.* ***PVLDB 2021***
- *J. Thorne, M. Yazdani, M. Saeidi, F. Silvestri, S. Riedel, and A. Halevy. Database reasoning over text. To appear in* ***ACL 2022****.*

UNIVERSITY OF CAMBRIDGE

SAPIENZA UNIVERSITÀ DI ROMA

Introduction

What is a Database?



Example

```
INSERT INTO People (PersonID, PersonName, Country) VALUES (123, 'Fabrizio Silvestri', 'Italy');
INSERT INTO People (PersonID, PersonName, Country) VALUES (789, 'Marzieh Saeidi', 'UK');

INSERT INTO Jobs (JobID, JobDescription) VALUES (111, 'Software Engineer')
INSERT INTO Jobs (JobID, JobDescription) VALUES (123, 'Research Scientist')

INSERT INTO PeopleJobs (PersonID, JobID) VALUES (123, 111)
INSERT INTO PeopleJobs (PersonID, JobID) VALUES (789, 123)

SELECT      p.PersonName
FROM        People p
JOIN        PeopleJobs pj
            ON (p.PersonID = pj.PersonID)
JOIN        Jobs j
            ON (pj.JobID = j.JobID)
WHERE       j.JobDescription = “Research Scientist”
```

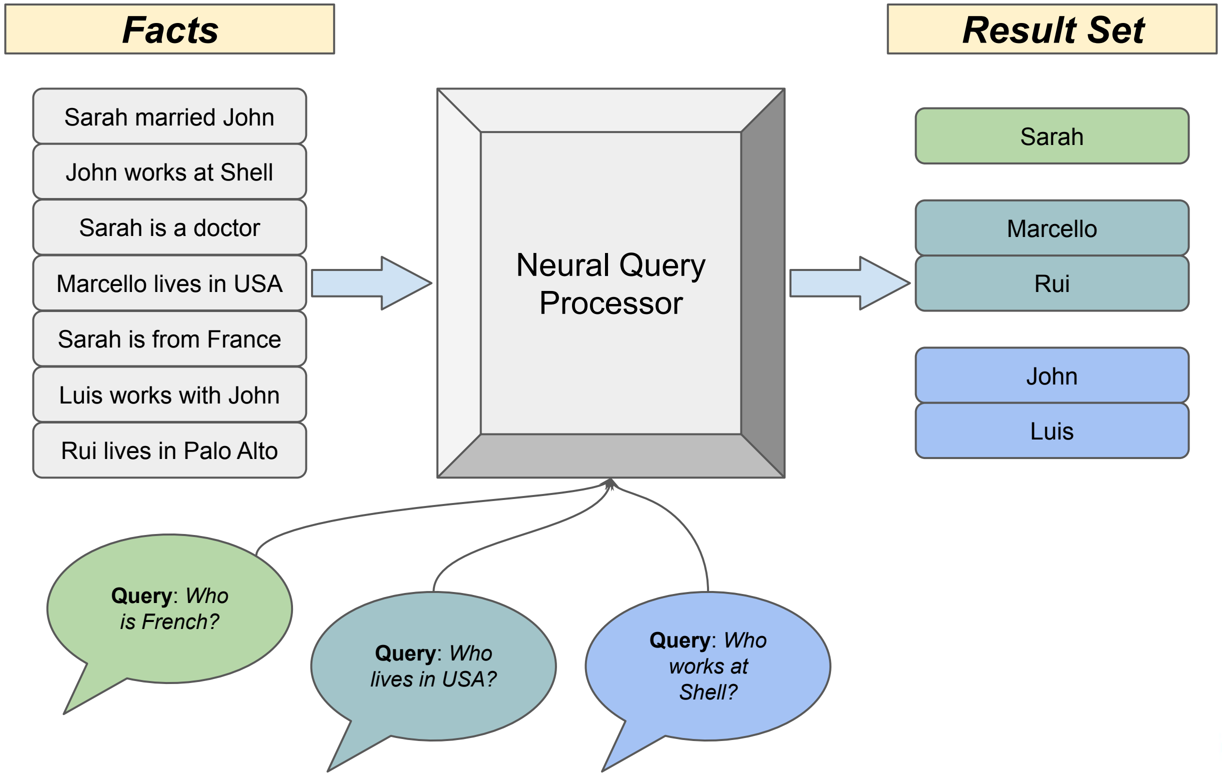
Database’s Core Component: The Schema

> “ The database schema of a database is its structure described in a formal language supported by the database management system (DBMS). The term ‘schema’ refers to the organization of data as a blueprint of how the database is constructed (divided into database tables in the case of relational databases). ”

from Wikipedia

Problem Definition

What if… We Removed Schema from Databases?



SAPIENZA
UNIVERSITÀ DI ROMA

Is it QA?

- QA tasks deal with questions posed in natural language:
  - When were the Normans in Normandy?
  - Which kicker had most field goals?
  - Several datasets, i.e., SQuAD, DROP, MSMARCO-QA, etc.
- In QA typically the answer to a query is located within a passage or multiple passages that are (usually) locally available
  - In NeuralDBs facts that form a single result set might be scattered around in the dataset.
- In QA typically the answer is, well… "an answer" 😁 Typically a single sentence, e.g., "*Adam Vinatieri*"
  - In NeuralDBs we should target both sets of answers, and aggregations (count, avg, etc.).

A Transformer Based Solution

- Transformers, e.g., BERT, have became ubiquitous in NLP.
- Introduced in the famous ‘Attention Is All You Need’ paper.
- It consists in applying (self-)attention to each token of a sequence of text, i.e., subwords.

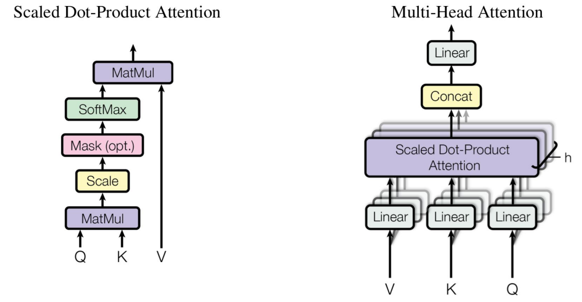


Figure 2: (left) Scaled Dot-Product Attention. (right) Multi-Head Attention consists of several attention layers running in parallel.

Credits: Attention Is All You Need’ by Vaswani et al.

How to train a Transformer LM

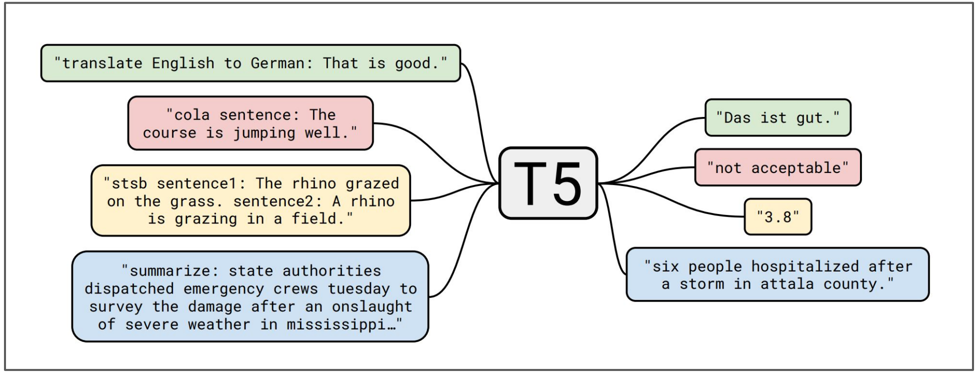
- Self-supervision on sequences of text, i.e., “Sentences”.

Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style Devlin et al. (2018)	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*
MASS-style Song et al. (2019)	Thank you <M> <M> me to your party <M> week .	*(original text)*
I.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you <X> to <Y> week .	<X> for inviting me <Y> your party last <Z>

Credits: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Raffel et al.

Google's T5 (**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer) Model

Trained in a multi-tasking fashion on the following tasks: (i) *GLUE* and *SuperGLUE* meta-tasks; (ii) CNN/Daily Mail *Abstractive Summarization*; (iii) *SQuAD* Question Answering; and (iv) WMT English to German, French, and Romanian *Translation*



Credits: Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Raffel et al.

SAPIENZA
UNIVERSITÀ DI ROMA

Neural Query Processing

- Task:
  - Given a query and a small number of facts from the database, can the T5 accurately answer queries that are posed in natural language, whose answer may require projection (i.e., extracting part of a sentence), join, and aggregation?
- Data: 7 different relationships from Wikidata extracted using a template
  - we generate a training, validation and held-out test set containing *535*, *50*, and *50* databases respectively
  - Each database contains *50* facts and has *100-200* QA pairs
  - In total: *60,000* training, *5,500* validation and *6,000* test instances
- Input: To provide input to the transformer, we jointly encode relevant facts from the database by concatenating them with the query (separated by a special delimiter token)

SAPIENZA
UNIVERSITÀ DI ROMA

Some Example Queries

Facts: (8 of 500 shown)

- Nicholas lives in Washington D.C. with his wife.
- Sheryl is Nicholas's wife.
- Teuvo was born in 1912 in Ruskala.
- Sheryl's mother gave birth to her in 1978.
- Nicholas is a doctor.
- Sarah was born in Chicago in 1982.
- Sarah married John in 2010.
- Sarah works in a hospital in NY as a doctor.

Queries:

List everyone born before 1980.
(`Set`) → Sheryl, Teuvo, . . .

Whose spouse is a doctor?
(`Join`) → Sheryl, John, . . .

Who is the oldest person?
(`Max`) → Teuvo

Who is Sheryl's mother?
(`Set`) → NULL

Dataset Building

How to build Facts and Queries?

- Training a NL database requires supervision in the form of *(D, Q, A)*:
  - *D* is a set of facts
  - *Q* is a query
  - *A* is the correct answer
- We generate training data in a controlled fashion by transforming structured data from Wikidata into NL facts and queries
- Pros:
  - Scale
  - Breadth

(**S**ubject, **R**elation, **O**bject)
(*Bezos*, *employedBy*, *Amazon*)

SAPIENZA
UNIVERSITÀ DI ROMA

Facts

- We "verbalize" knowledge graph triples that are synthesized through a sequence to sequence model
  - Data is from KELM, we generate a rule-based post-hoc mapping back to Wikidata considering: string similarity, and compatibility of the generated triple

Input Triples	Target Sentence
Das Tagebuch der Anne Frank, (distributor, Universal Pictures), (country, Germany), (publication date, 03 March 2016)	The film was theatrically released in the Germany on March 3, 2016, by Universal Pictures International.
Neff Maiava, (date of birth, 01 May 1924), (date of death, 21 April 2018), (occupation, professional wrestler)	Maiava (May 1, 1924 April 21, 2018) was an American Samoan professional wrestler.
Barack Obama 2012 presidential campaign, (country, United States), (end time, 06 November 2012), (start time, 04 April 2011)	The 2012 reelection campaign of Barack Obama, the 44th President of the United States, was formally announced on April 4, 2011.
Blue whale (parent taxon, Balaenoptera)	The blue whale (Balaenoptera musculus) is a marine mammal belonging to the baleen whale suborder Mysticeti.

Agarwal, O., Ge, H., Shakeri, S. and Al-Rfou, R., 2020. Large Scale Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training *arXiv preprint arXiv:2010.12688*.

Queries

- We generate queries using a number of templates for each property and question type
  - *(X, bornIn, Y)*
  - *(X, employedBy, Y)*
- We also “simulate” joins by “chaining” triples:
  - (*Y*, *locatedIn*, *Z*) ⛓ (*X*, *employedBy*, *Y*) → “Does $X work at a company based in $Z?”
  - (*Y*, *marriedTo*, *Z*) ⛓ (*Z*, *leavesIn*, *Y*) → “Does $Z’s spouse leaves in $Y?”
  - (*Y*, *childOf*, *Z*) ⛓ (*Y*, *bornIn*, *X*) ⊓ (*Y’*, *bornIn*, *X’*) → “Is $Z’s child younger than $Y’?”
  - (*Y*, *rel1*, *Z*) ⛓ (*Z*, *rel2*, *Y*) → “Does $Z’s **rel1** also **rel2** $Y?”

Example Queries

Example: `Set`	
Question	Who studied at University of Minnesota?
Supporting Facts	**1.** [John B Totushek was born on 7 September 1944 in Minneapolis. He attended the University of Minnesota and became a US Naval Aviator. Mr. Totushek was also a human being.] **2.** [Melvin Maas graduated from the University of Minnesota and is buried at Arlington National Cemetery. He is a native of Minnesota and his language is English.] **3.** [Clarence Larson graduated from the University of Minnesota and is a member of the National Academy of Engineering.] **4.** [Ted Mann, who is the surname of Ted Mann, attended Duke University and the University of Minnesota. He is a human being.]
Answer	[John B. Totushek, Ted Mann, Clarence Larson, Melvin Maas]

SAPIENZA
UNIVERSITÀ DI ROMA

Example Queries

Example: `count`	
Question	How many people work for Yale Law School?
Supporting Facts	**1.** [Michael Ponsor, born in Oxford, graduated from Pembroke College in Oxford. He was awarded the Rhodes Scholarship and is an employee at Yale Law School. He is an expert in the field of human rights.] **2.** [Stephen Wizner is an American legal scholar who graduated from Dartmouth College and is a graduate of the University of Chicago Law School. He works at Yale Law School.]
Answer	2

Example Queries

Example: `Min/Max`	
Question	What is the largest yearly attendance?
Supporting Facts	**1.** [The musee en herbe has a visitor per year of] 70000. **2.** [The total number of visitors to the Hirschsprung Collection is 71779 per year.] ... **24.** [The Tate Modern has a visitor count of 5839197 visitors per year.] **25.** [Catoctin Mountain Park attracts 221750 visitors per year.]
Answer	5839197

SAPIENZA
UNIVERSITÀ DI ROMA

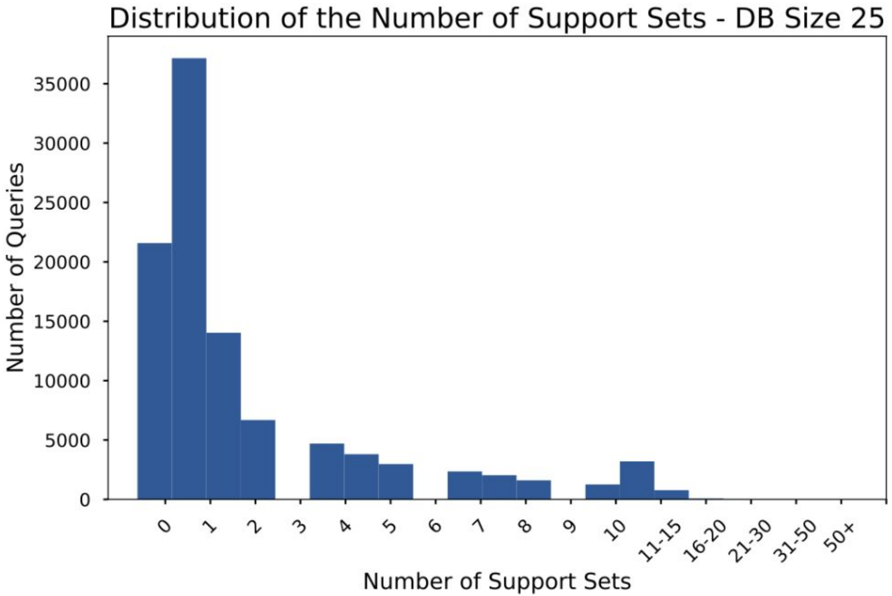
Example Queries

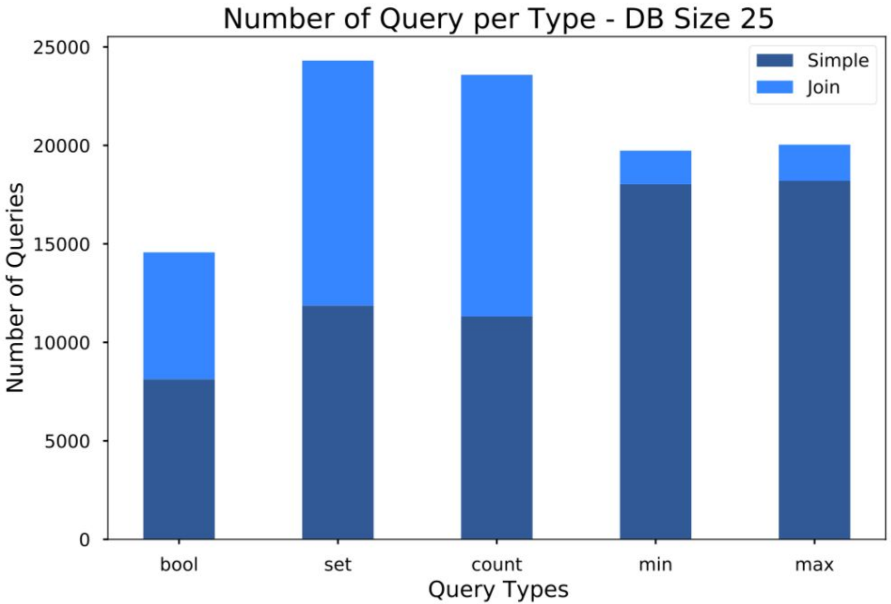
Example: `Bool`	
Question	Is North Carolina State University the employer of Wes Moore?
Supporting Facts	**1.** [Wes Moore is a human being who is employed at Francis Marion University and is a basketball player for North Carolina State University.]
Answer	TRUE

Example Queries

Example: `Join`	
Question	Who plays for a team in Ligue 1?
Supporting Facts	**1.** [Thomas Allofs started his career in 1989 with RC Strasbourg Alsace. He finished his career in 1990., RC Strasbourg Alsace is an association football club in the Ligue 1 league. It was founded in 1906 and is located in Strasbourg, France.]
Answer	[Thomas Allofs]

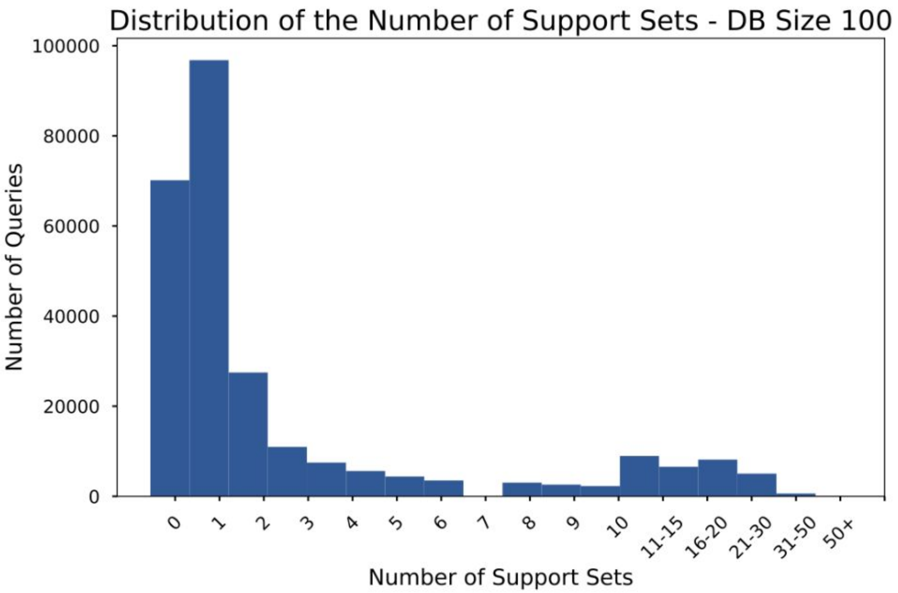
WikiNLP Statistics

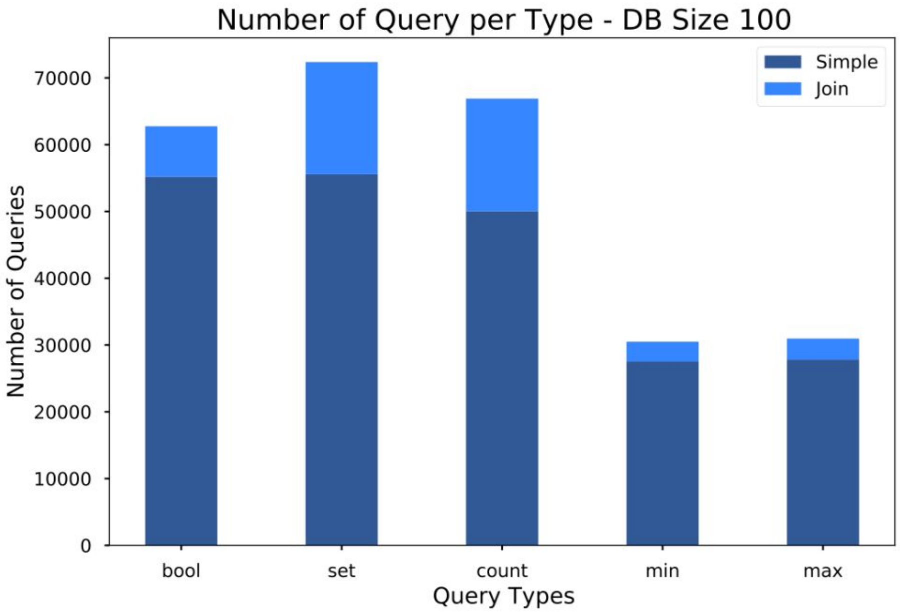




SAPIENZA
UNIVERSITÀ DI ROMA

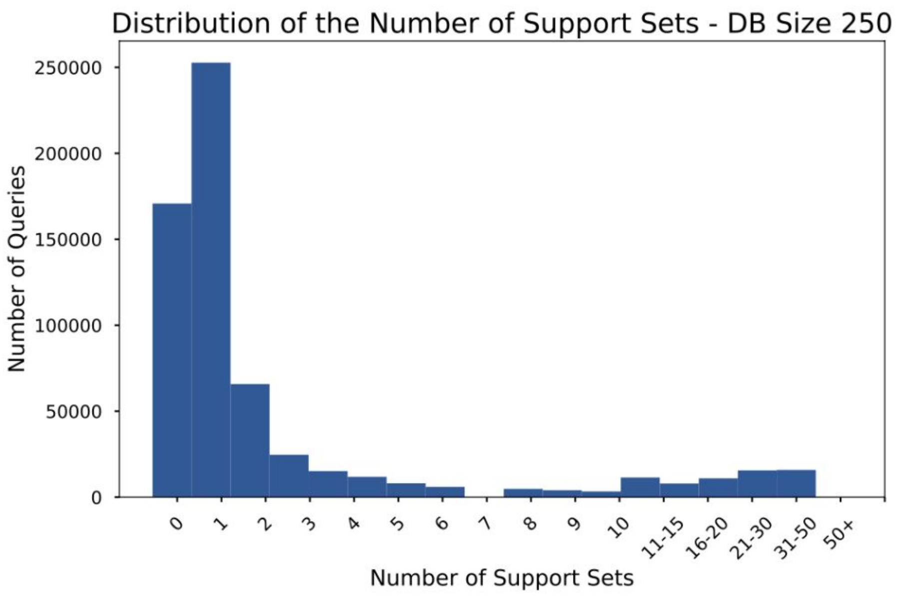
WikiNLP Statistics

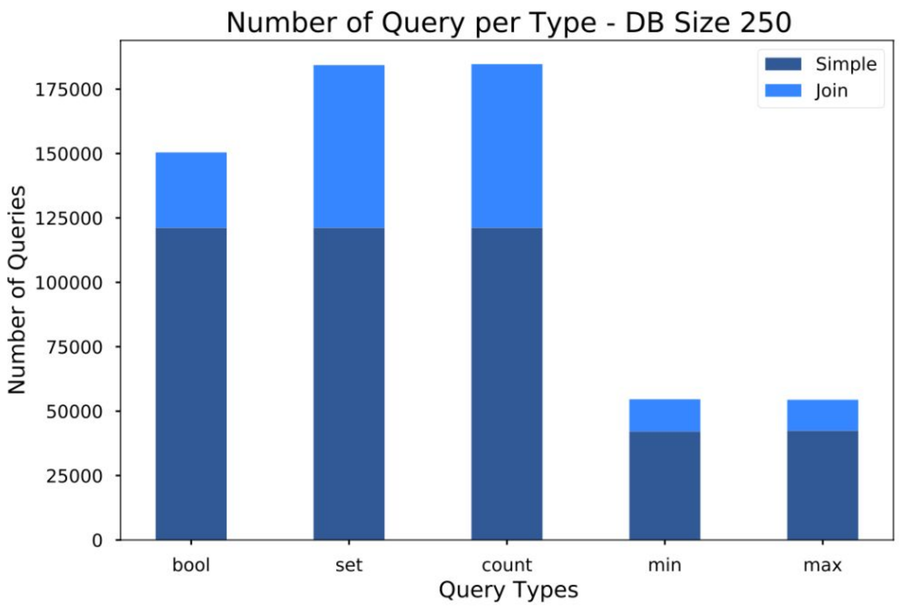




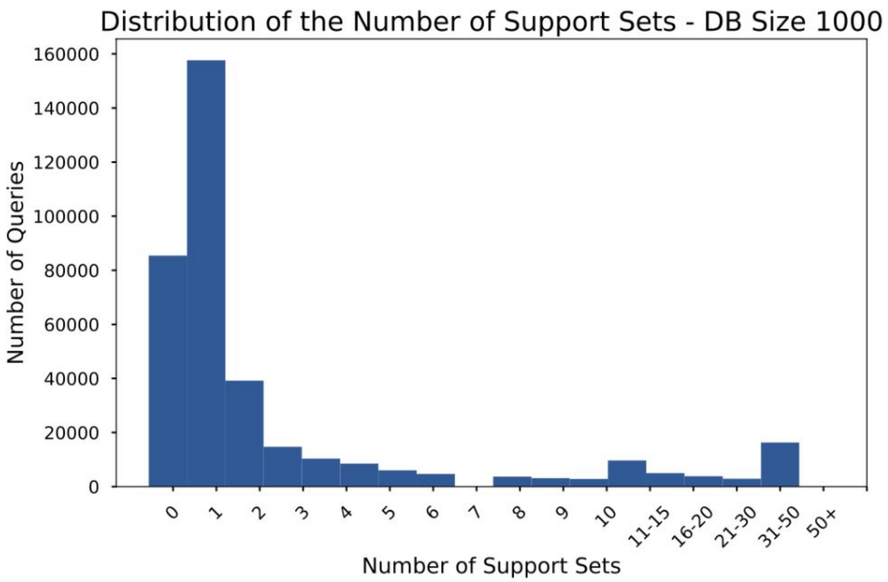
Sapienza Università di Roma

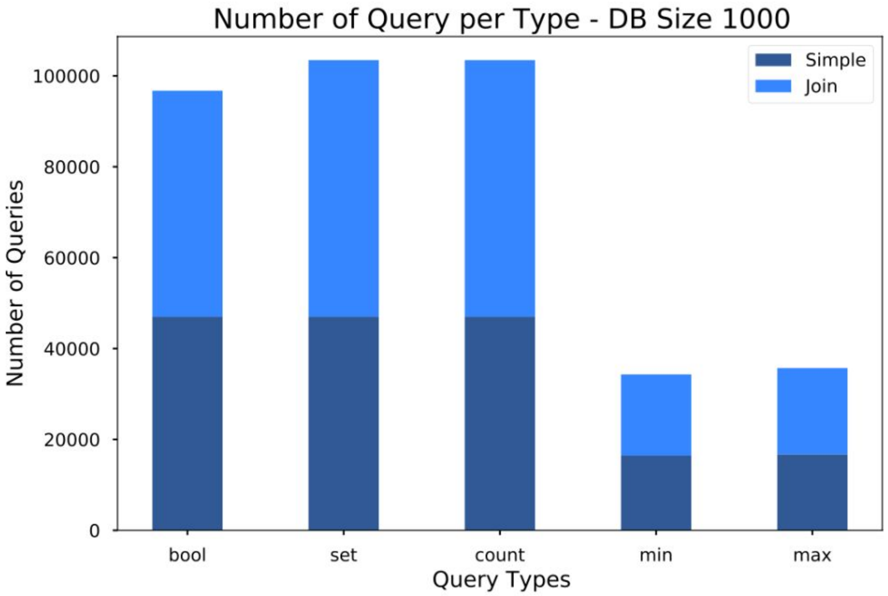
WikiNLP Statistics





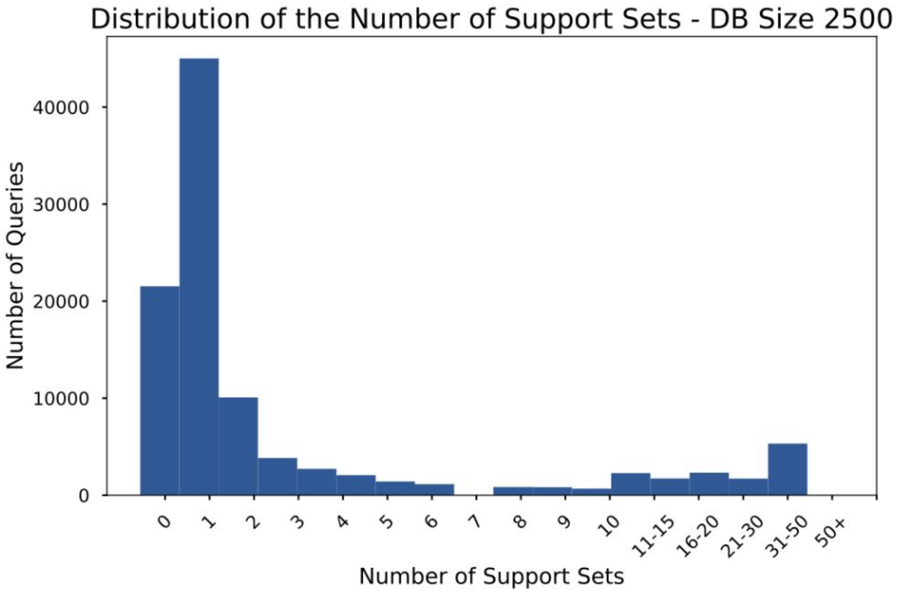
WikiNLP Statistics

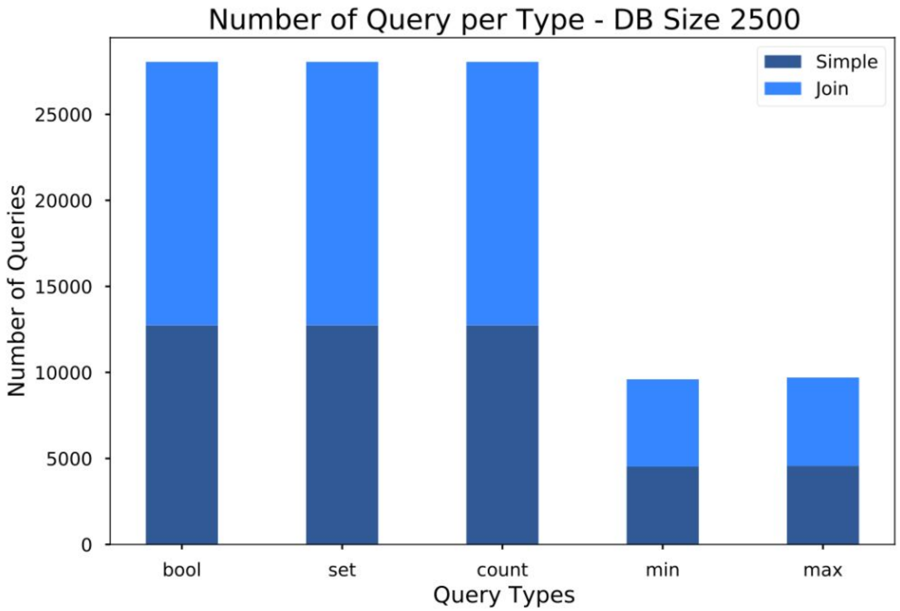




SAPIENZA
UNIVERSITÀ DI ROMA

WikiNLP Statistics

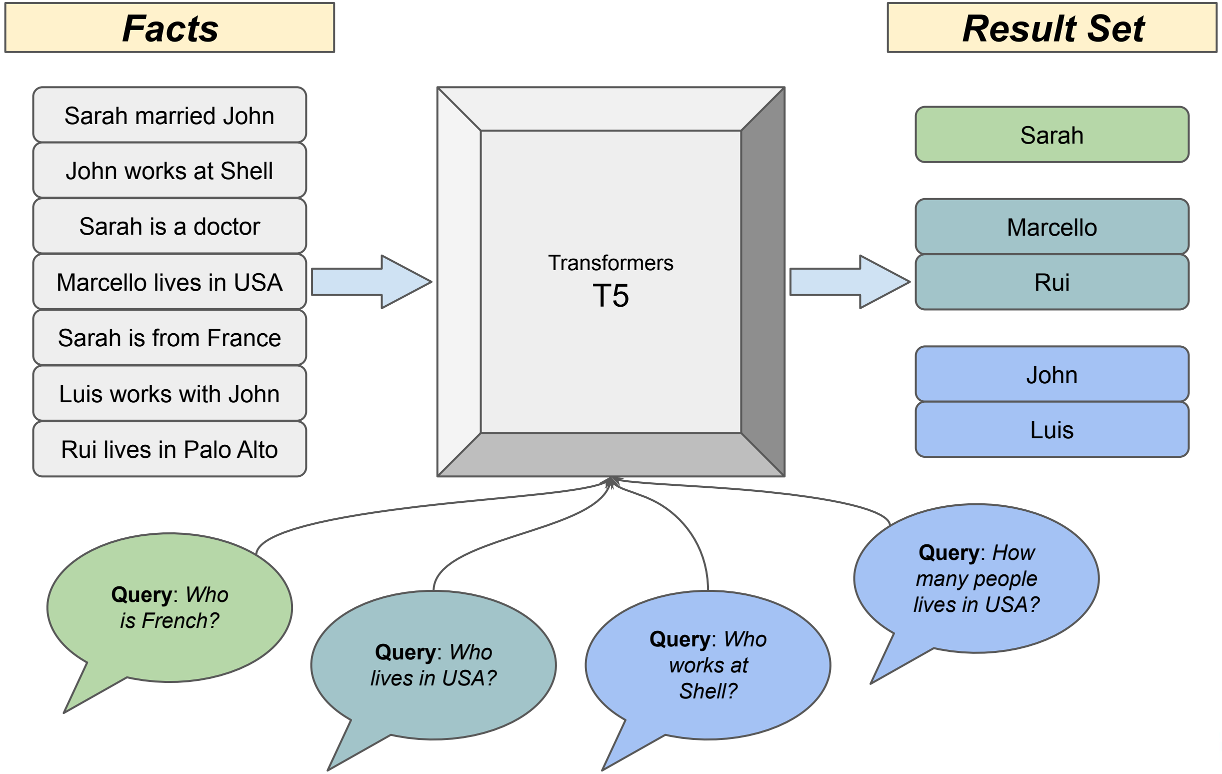




SAPIENZA
UNIVERSITÀ DI ROMA

NeuralDBs Architectures

A “Simple” Solution



SAPIENZA
UNIVERSITÀ DI ROMA

Possible Issues

✅ When fed with relevant facts from the database, T5 can produce results with reasonable accuracy

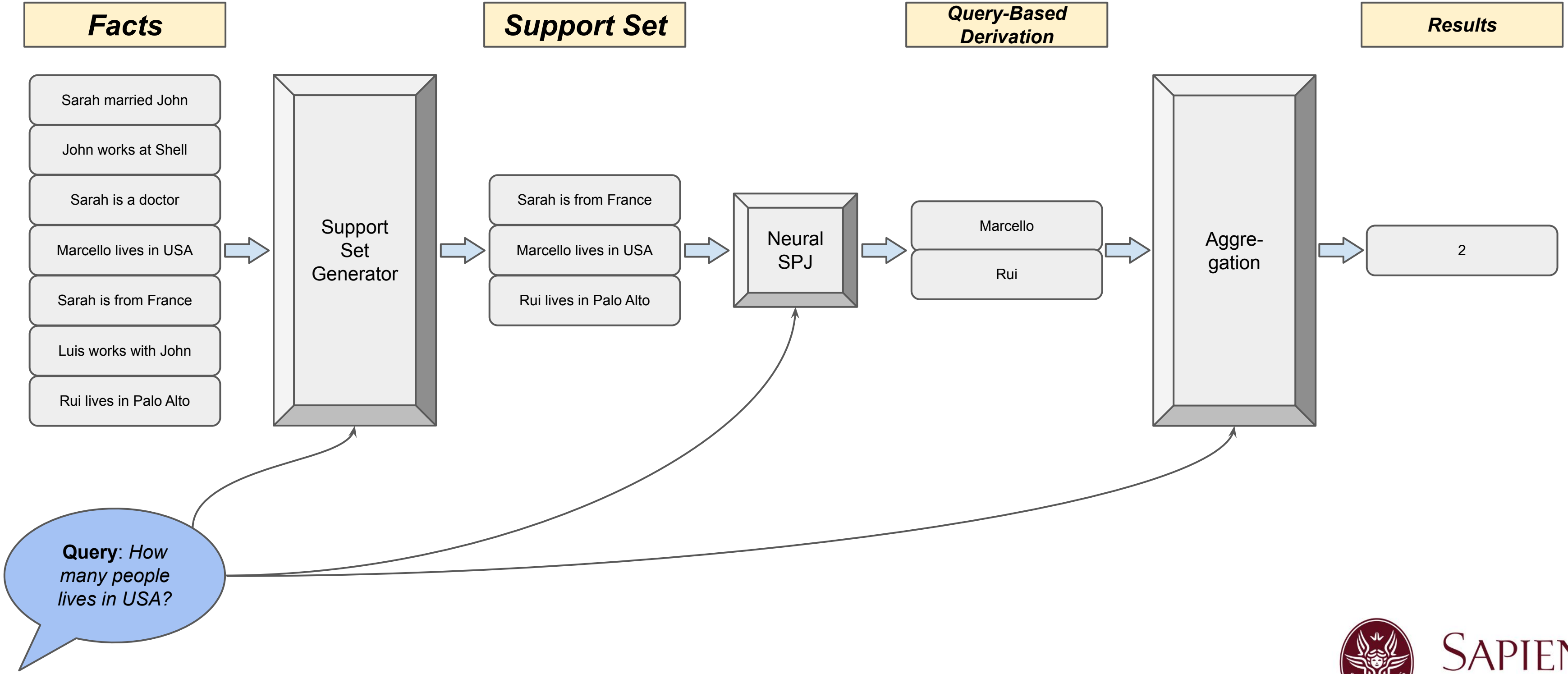
❌ Aggregation queries need to be performed outside of the neural machinery

👉 In order to handle queries that result in sets of answers and in order to prepare sets for subsequent aggregation operators, we need to develop a neural operator that can process individual (or small sets of) facts in isolation and whose results outputted as the answer or fed into a traditional (i.e. non-neural) aggregation operator.

Challenges

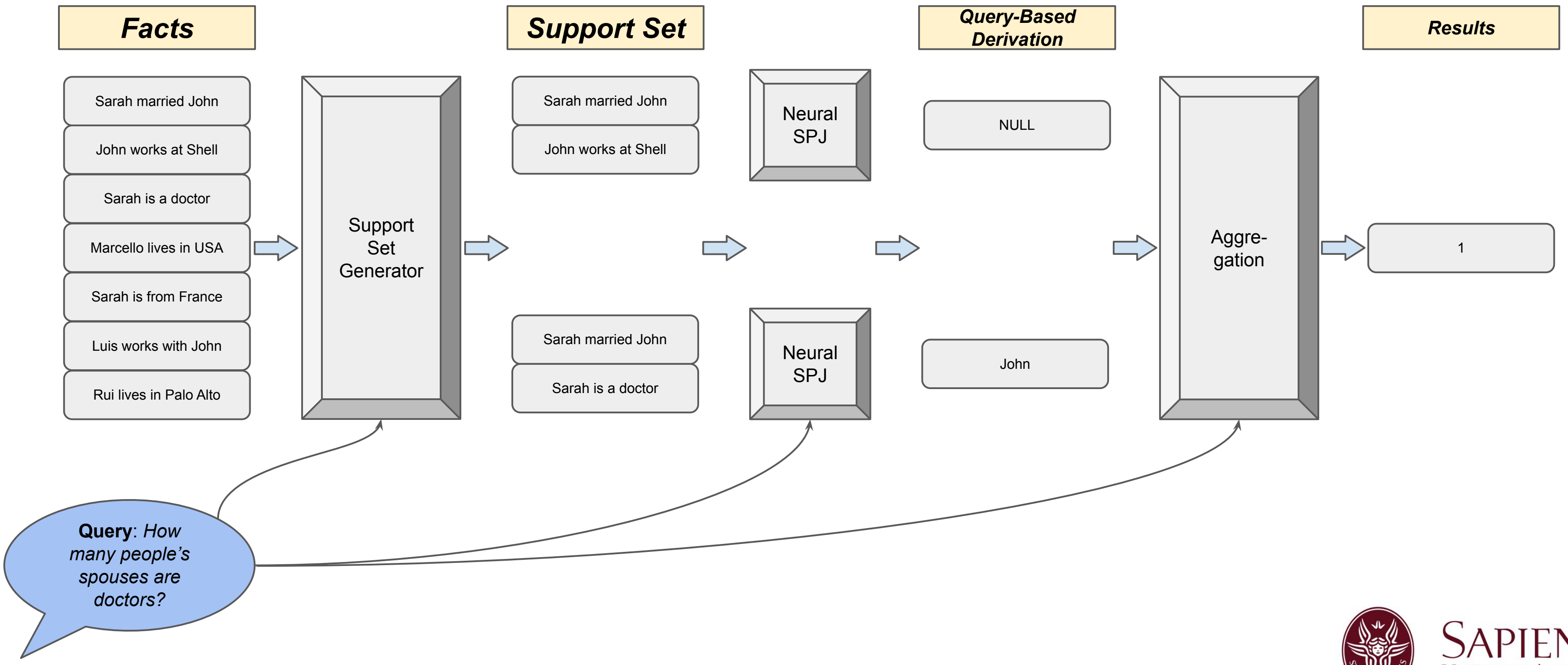
- **Scale**
  - neural reasoning to databases of non-trivial size
    - In open-domain QA we usually complement the transformer reasoning with an IR component that extracts a small subset of the facts from the corpus
- **Multiple answer spans**
  - NDB might need to generate 100K facts and aggregate over them
- **Locality and Document Structure**
  - Answers might be dependent on several facts scattered across the DB
- **Multi-hop and Conditioned Retrieval**
  - E.g., *Whose spouse is a doctor?*

The Neural DB Architecture



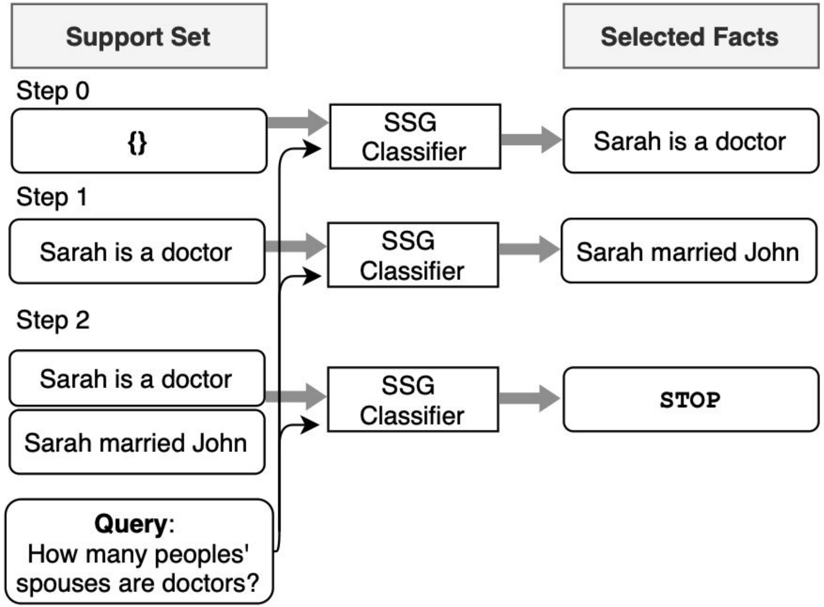
SAPIENZA
UNIVERSITÀ DI ROMA

The Neural DB Architecture



Support Set Generator (SSG)

- Simple queries over single facts → TF-IDF based IR
  - not scalable for joins, aggregation queries or for queries outputting a set of answers as generating relevant sets requires incremental decoding, conditioning on already retrieved facts.



SAPIENZA
UNIVERSITÀ DI ROMA

Neural Select-Project-Join (SPJ)

- For support sets that are insufficient to answer a question, the operator should return no output.
- For queries that require short chains of reasoning over multiple facts, the SPJ operator joins the facts when generating the output.
- SPJ generates a projection of the fact to a machine readable format dependent on the task, query and fact.
- Depending on the query type:
    - **Boolean Answers** → binary value
    - **Count/Set Queries** → entities
    - **Argmin/max operators** → key-value pairs.
        - For example "*Which place has the highest yearly number of visitors?*" has the projection of the form: (place,number of visitors).

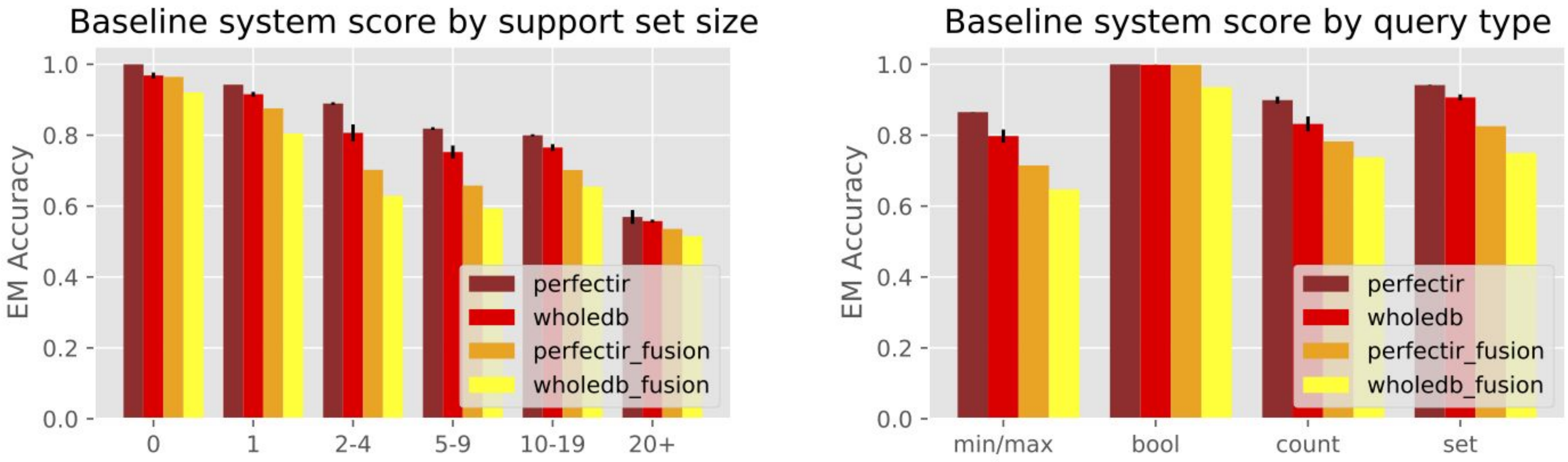
Examples of Neural SPJ Outputs

- Query: Does Nicholas's spouse live in Washington D.C.?
  - {*Nicholas lives in Washington D.C. with Sheryl.*, *Sheryl is Nicholas's spouse.*} → TRUE
- Query: Who is the oldest person in the database?
  - {*Teuvo was born in 1912.*} → (Teuvo, 1912)
- Query: Does Nicholas's spouse live in Washington D.C.?
  - {*Teuvo was born in 1912.*} → NULL

Results: SPJ Performance

Method	Exact Match (%)				
	Count	Min/Max	Sets	Atomic	Joins
NeuralDB	**79.45**	**100.00**	**91.91**	**97.90**	**79.29**
TF·IDF+T5	31.06	0.00	44.25	98.05	68.02
DPR+T5	38.07	21.19	54.55	97.38	58.64

Exact Match



Exact match accuracy for different classes of queries for a transformer model encoding up to 25 facts in one input. The results show that the model obtains high accuracy for queries performing Boolean inference, but falls short for queries with aggregation or yielding set answers (top) over multiple support sets (bottom).

Results: SPJ Performance

Method	Exact Match (%)			
	Min/Max	Bool	Count	Set
SPJ PerfectIR	88.3	99.8	90.1	89.4
SSG + SPJ	87.3	99.8	90.1	89.6

Using retrieved evidence achieves results competitive to the PerfectIR on a DB of 25 facts.

SAPIENZA
UNIVERSITÀ DI ROMA

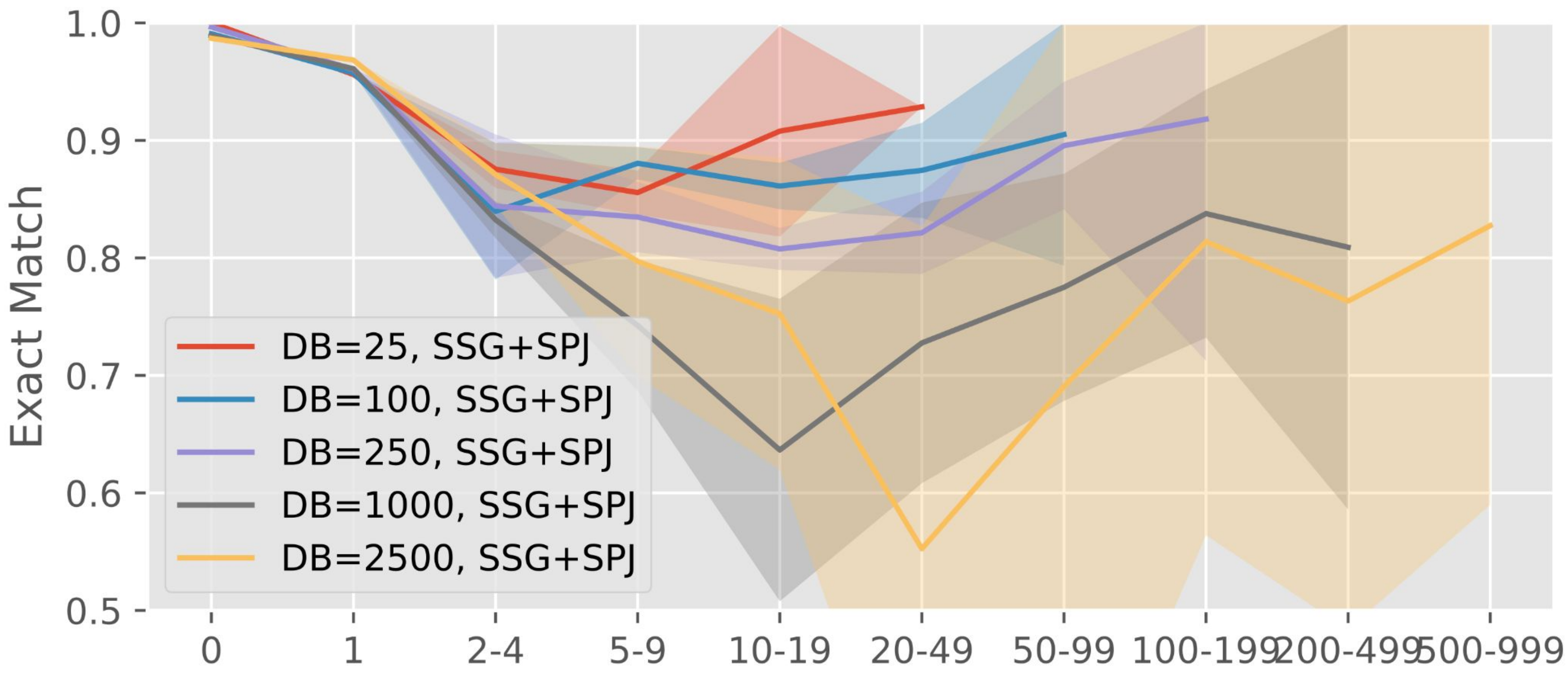
Results: SSG Precision/Recall

Query Type	Exact Match (%)		Soft Match (%)	
	Precision	**Recall**	**Precision**	**Recall**
Boolean	85.12	94.04	85.39	94.04
Set	61.05	94.58	61.33	94.58
Count	57.88	96.15	58.00	96.15
Min/Max	60.68	95.82	60.68	95.82
Join	38.33	75.39	42.74	75.43
Average	57.08	90.53	58.24	90.54

Precision and recall of supervised SSG w.r.t. the reference set. Note that errors in retrieval do not necessarily translate to wrong query answers because the SPJ operator is trained to be robust to noise.

SAPIENZA
UNIVERSITÀ DI ROMA

Results: SSG + SPJ Accuracy (different DB Sizes)



SSG+SPJ by support set size for all 5 databases. The SPJ is trained on databases of 25 facts and tested on larger databases. Low recall from SSG reduced answer EM for DBs of more than 1000 facts.

Conclusions and Future Work

- We described NeuralDB
  - Using neural reasoning to answer queries from data expressed as natural language sentences that do not conform to a predefined schema.
  - Our experiments show that NeuralDB attains very high accuracy for a class of queries that involve select, project, join possibly followed by an aggregation.
- We will need to investigate potential biases in our language model and the impact on NeuralDBs
- Identifying which updates should replace previous facts:
  - Mariah is unemployed → Mariah works for Apple; vs.
  - Kasper likes tea → Kasper likes coffee.

A Research Agenda

- Deeper understanding of semantics
- Multi-modal neural databases
- Obtaining training data and transfer learning
- Mitigating biases
- Applying neural components to existing data management architectures

How NeuralDBs could help Web Engineering

- Imagine a world where you make your web application interact with a user through natural language
  - You can let the user interact with your application by allowing them to add facts and preferences later used to enhance the personalized experience you offer to your users
- Imagine you want to expose your application through a “smart speaker”
  - A person might add personal facts to a local repository on their smart speaker and allow this set of fact to be used, for instance, to better understand and answer their requests