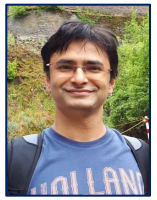
CLOCQ:
A Toolkit for Fast and Easy Access to Knowledge Bases

Philipp Christmann, Rishiraj Saha Roy, Gerhard Weikum


Presenter: Magdalena Kaiser

KBs provide vast amounts of information

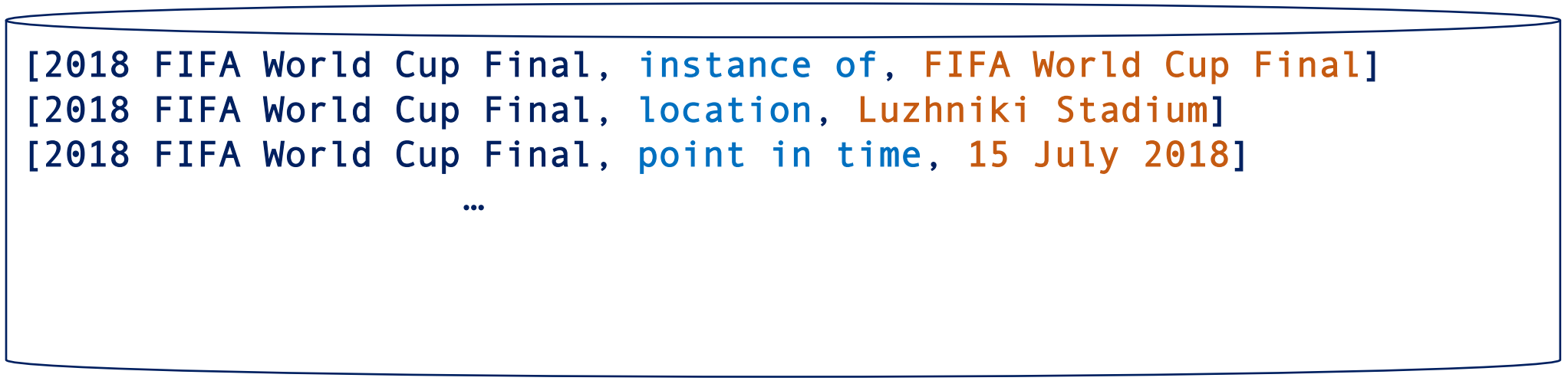
Curated **knowledge bases** (KB) have

- ★ Billions of facts
- ★ Millions of entities
- ★ Thousands of relations
- ★ Multiple terabytes of data

- ★ KBs provide **vast amounts** of **factual information**
- ★ **Empower** everyday **applications**

Curated Knowledge Base (KB)

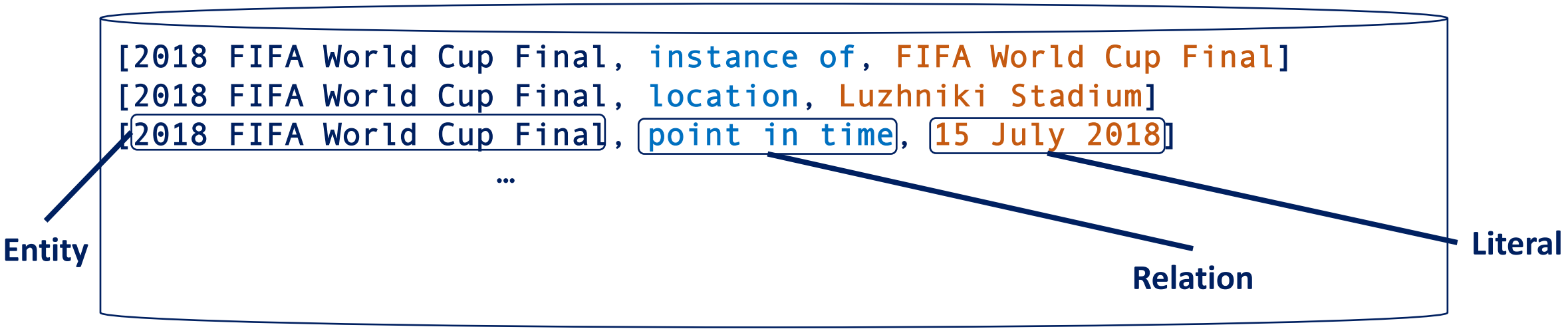
Set of facts [subject, predicate, object]



⇒ Expressing **factual world knowledge**, in a **structured** way

Curated Knowledge Base (KB)

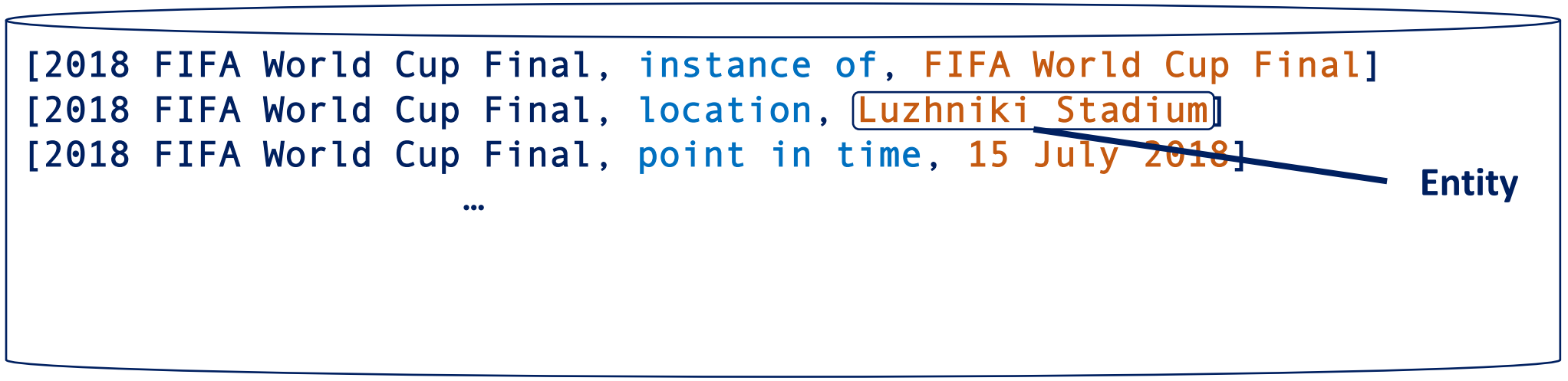
Set of facts [subject, predicate, object]



⇒ Expressing **factual world knowledge**, in a **structured** way

Curated Knowledge Base (KB)

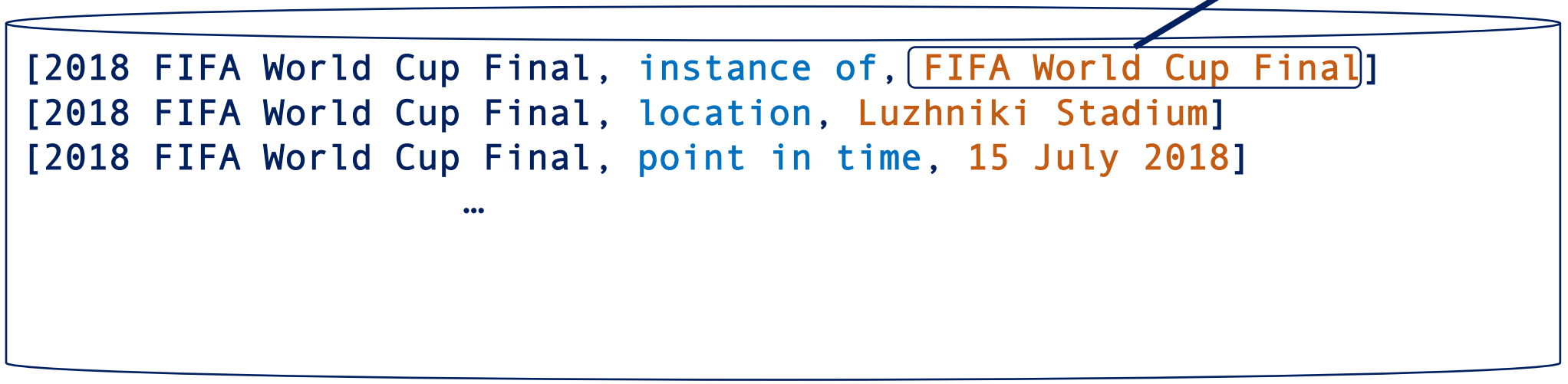
Set of facts [subject, predicate, object]



⇒ Expressing **factual world knowledge**, in a **structured** way

Curated Knowledge Base (KB)

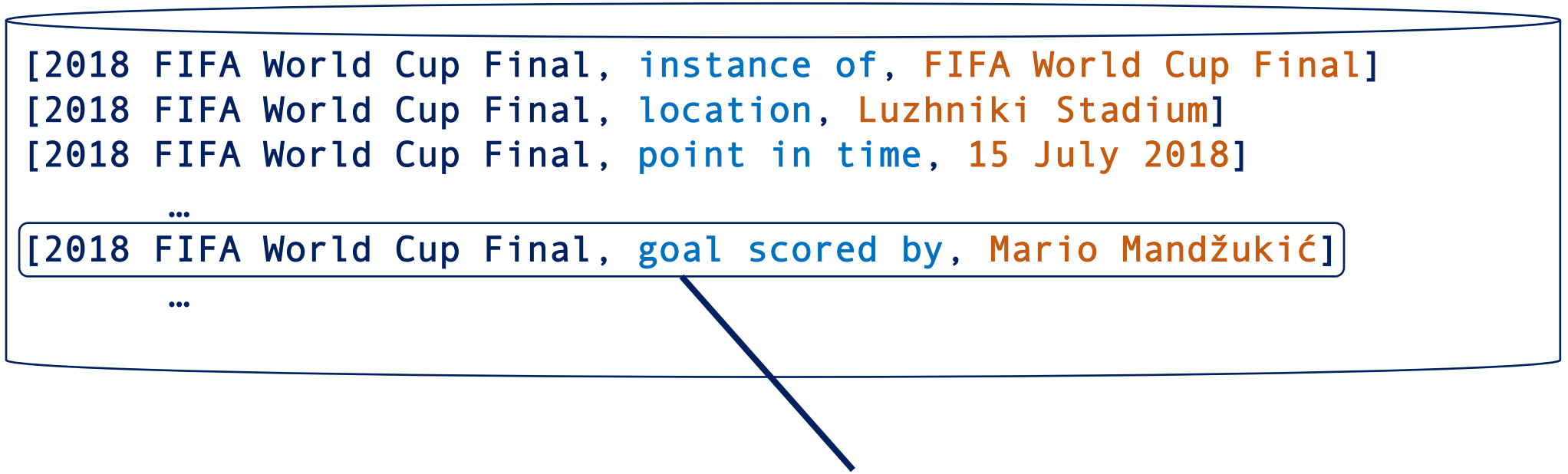
Set of facts [subject, predicate, object]



⇒ Expressing **factual world knowledge**, in a **structured** way

Curated Knowledge Base (KB)

Set of facts [subject, predicate, object]



Important **information missing:**

⇒ This was an **own goal**

⇒ The **minute** the goal was scored in (18 minute)

⇒ The **way** the goal was scored (header)

Curated Knowledge Base (KB)

Set of facts [subject, predicate, object, qualifiers(optional)]

```
[2018 FIFA World Cup Final, instance of, FIFA World Cup Final]
[2018 FIFA World Cup Final, location, Luzhniki Stadium]
[2018 FIFA World Cup Final, point in time, 15 July 2018]
        …
[2018 FIFA World Cup Final, goal scored by, Mario Mandžukić,
(match time, 18 minute), (score method, head), (score method, own goal)]
        …
```

⇒ Expressing **factual information**, in a **structured** way

⇒ KB facts often need to go **beyond triples**

⇒ Expressing **n-ary relationships** using **qualifiers:** (qualifier-predicate, qualifier-object) pairs

Limitations of existing KB interfaces

Existing KB interfaces allow **general-purpose access** to KBs via **queries** (e.g., SPARQL)

☆ **Not designed** for integrating **qualifiers**

☆ **KB** as **set of triples**, integrating qualifiers via reification

⇒**Expensive** querying and post-hoc processing

☆ Access **requires** deep **knowledge and understanding** of KB **schema**

⇒ **A lot of time** needs to be **invested** until **first usage**

Fast Access

- ★ Propose to take a **fact-centric view of KBs** (vs. triple-centric)
- ★ Implement **fact-centric KB index**
  - ★ Directly **store** the whole fact, including **the corresponding qualifiers**
  - ★ Establish **index** that **stores all facts** with a certain **KB entity** (or relation)
- ★ Enables (more) **efficient implementation** of **core KB functionalities, e.g.**
  - ★ **KB neighborhood:** simple look-up
  - ★ **KB distance:** implemented via intersection of neighborhoods

Easy Access

★ Establish **intuitive definitions** for vaguely defined **concepts:**
- ⇒ KB represented as a graph (knowledge graph)
- ⇒ KB neighborhood
- ⇒ KB distance
- ⇒ Shortest path between KB items

★ Provide **public API** to conveniently access **Wikidata**
https://clocq.mpi-inf.mpg.de

★ **Open source code** available at same URL

KB functionalities

Direct lookups

⇒ Label, aliases, description, types, most frequent type

More complex functionalities

⇒ 1-hop neighborhood of KB item

⇒ Frequency of KB item

⇒ Connectivity / shortest path between two KB items

Search space reduction and linking

⇒ Retrieve relevant KB facts for input text (e.g., question)

⇒ Identify KB entities in input text

⇒ Identify KB relations in input text

Conclusion

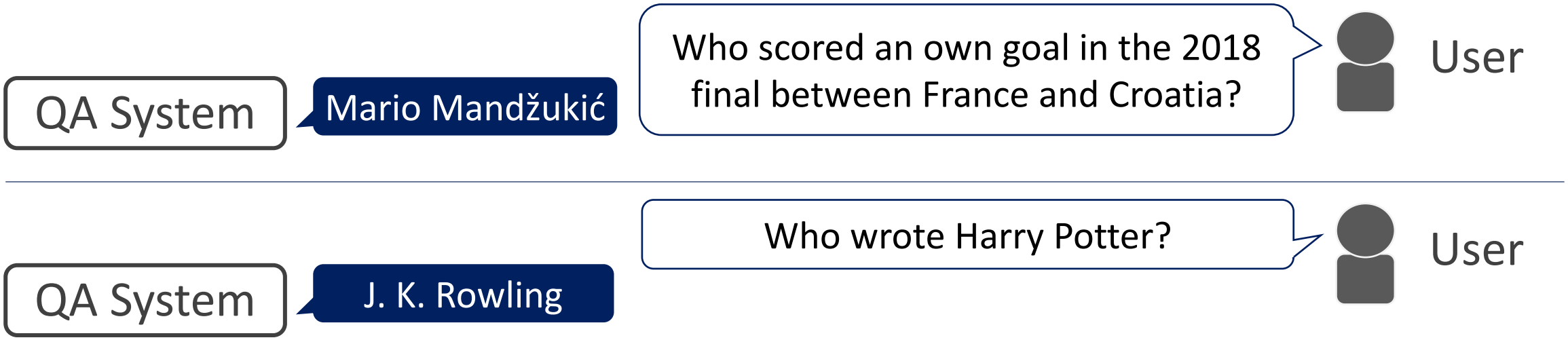
★ Propose to take a **fact-centric view** on KBs
★ Establish **intuitive definitions** for **salient concepts**
★ Construct **fact-centric KB interface**
★ **Improve runtime** for **core** KB **functionalities**
over triple-centric KB interfaces (factor of up to 10^5)

Enhance accessibility to large-scale KBs for research
⇒ Public **code** toolkit
⇒ **Open API** to access **Wikidata**
⇒ >19,000,000 **external API calls** so far

https://clocq.mpi-inf.mpg.de

Thank you!

Question Answering (QA) over KBs



Other example use cases of KBs

Entity ranking in search engines [1]
⇒ For queries like “2018 wc final”

Named entity recognition and disambiguation (NERD) [2]
⇒ Canonicalize entity mentions in texts

Information and statistics extracted from KB
★ **Distance** between **KB entities** as proxy of semantic similarity [3]
★ **Frequency** of **entity** as a measure for its popularity [4]
★ **KB ontology** (type system) for relation extraction [5] or answer verification [6]

[1] BERT-ER: Query-specific BERT Entity Representations for Entity Ranking, Shubham et al., SIGIR 2022.
[2] Robust Disambiguation of Named Entities in Text, Hoffart et al., EMNLP 2011.
[3] Computing semantic similarity of concepts in knowledge graphs, Zhu et al., IEEE TKDE 2016.
[4] Look before you hop: Conversational question answering over knowledge graphs using judicious context expansion, Christmann et al., CIKM 2019.
[5] Type-aware distantly supervised relation extraction with linked arguments, Koch et al., EMNLP 2014.
[6] More accurate question answering on freebase, Bast et al., CIKM 2015.

Runtime experiments

Functionalities

★ 1-hop neighborhood (item)
★ Frequency (item)
★ Connectivity (item1, item2)
★ Shortest path (item1, item2)
⇒ **10,000** randomly sampled **items / item pairs**

Baselines

☆ **HDT [1]:** Efficient **triple lookups** using bitmap encodings
☆ **QueryService [2]:** Publicly available **SPARQL query** interface for Wikidata

[1] Binary RDF representation for publication and exchange (HDT), Fernández et al., Journal of Web Semantics 2013.
[2] https://query.wikidata.org/

Experimental results