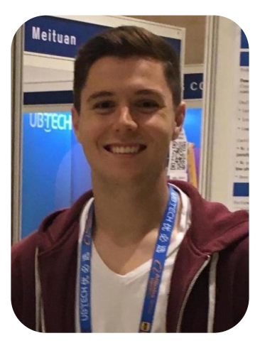
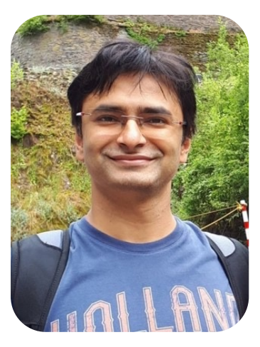
CLOCQ: A Toolkit for Fast and Easy Access to Knowledge Bases

Philipp Christmann, Rishiraj Saha Roy, and Gerhard Weikum
Max Planck Institute for Informatics, Germany

Public API and source code at clocq.mpi-inf.mpg.de

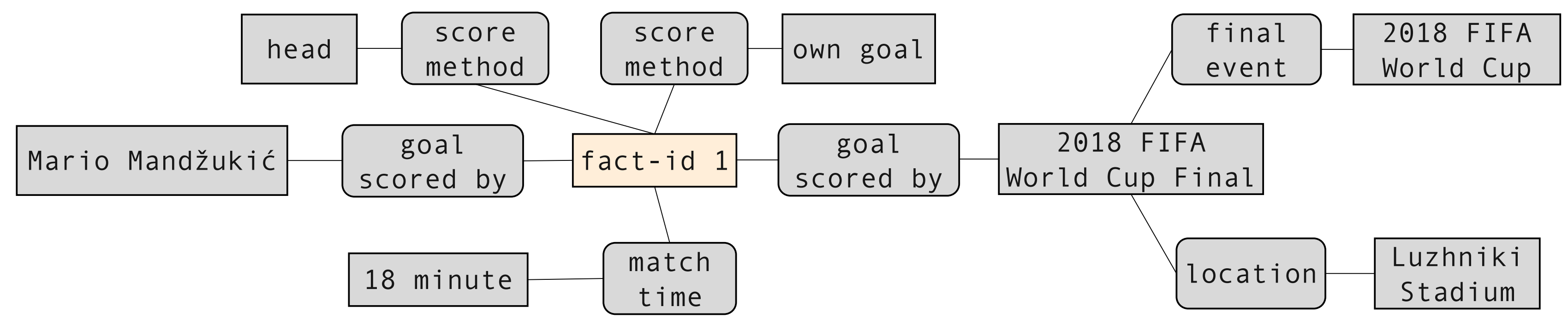
KNOWLEDGE BASES STORE VAST AMOUNTS OF FACTUAL KNOWLEDGE

- ★ Curated **knowledge bases** (KBs) store factual knowledge in structured way and have **many use-cases** for search, entity linking, etc.
- ★ **Qualifiers** express ***n*-ary relationships** in Wikidata; similar concepts used in other KBs such as DBPedia or YAGO
- ★ Real-world KBs have **billions of facts**, with **millions of entities** and **thousands of predicates**, consuming **multiple terabytes** of disk space
- ★ KBs used in **question answering** (QA) systems, to answer **factoid questions** like *Who wrote Harry Potter?* or *Who scored an own goal in the 2018 final?*

LIMITATIONS OF EXISTING TRIPLE-CENTRIC KB INTERFACES

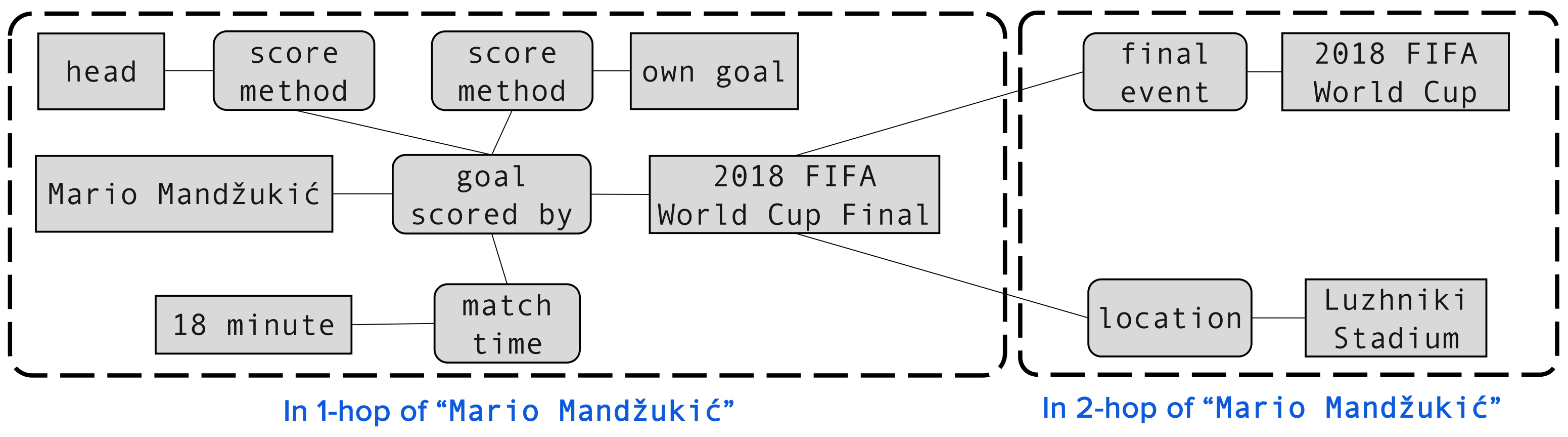
- ☆ Existing KB interfaces allow **general-purpose access** via queries (e.g., SPARQL)
- ☆ Access requires detailed **knowledge** and **understanding** of **KB schema**
- ☆ Interfaces **not designed** for accessing ***n*-ary facts**
- ☆ Treat KB as **pure set of triples** and integrate qualifiers via reification
- ☆ Leads to **expensive querying** and **post-hoc processing**

Graph-based definition of KB distance (with triple-centric view)



KB distance (Mario Mandžukić, 2018 FIFA World Cup Final) = 4 (follow 4 edges)
KB distance (Mario Mandžukić, Luzhniki Stadium) = 6 (follow 6 edges)

Fact-based definition of KB distance (Proposed)



KB distance (Mario Mandžukić, 2018 FIFA World Cup Final) = 1 (appear in same fact)
KB distance (Mario Mandžukić, Luzhniki Stadium) = 2 (1 fact apart)

Legend ▭ Entity node ▢ Predicate node

Traditional triple-centric KB index

2018 FIFA World Cup Final ↳

2018 FIFA World Cup Final	instance of	FIFA World Cup Final
2018 FIFA World Cup Final	location	Luzhniki Stadium
…	…	…
2018 FIFA World Cup Final	goal scored by	fact-id 1
…	…	…
2018 FIFA World Cup	final event	**2018 FIFA World Cup Final**

fact-id 1 ↳

fact-id 1	goal scored by	Mario Mandžukić
fact-id 1	match time	18 minute
fact-id 1	score method	own goal
fact-id 1	score method	head

Fact-centric KB index (Proposed)

2018 FIFA World Cup Final ↳

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

Legend ↳ Data in index for KB-item [•] KB-fact

CLOCQ APPROACH

- ★ Take **fact-centric view** of KBs (vs. triple-centric)
- ★ Establish **intuitive definitions** for vaguely defined concepts, such as:
  KB graph, KB neighborhood, KB distance, shortest path between KB items
- ★ Implement **fact-centric KB index** that enables (more) **efficient** implementation of **core KB functionalities** utilized in many IR and NLP systems
- ★ Provide **public API** to conveniently access **Wikidata** at clocq.mpi-inf.mpg.de

CLOCQ FUNCTIONALITY

- ★ **Direct lookups**
  Label, aliases, description, types, or most frequent type of KB item
- ★ **More complex functionalities**
  - ★ 1-hop neighborhood of KB item
  - ★ Frequency of KB item
  - ★ Connectivity / shortest path between two KB items
  - ★ Search space reduction (text)
  - ★ Entity linking (text)
  - ★ Relation linking (text)

⏰ CLOCQ improves runtime over traditional triple-centric KB interfaces

RUNTIME EXPERIMENTS

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

Large-scale runtime analysis for key KB functionalities and randomly chosen KB items.

	HDT [1]	QueryService [2]	CLOCQ-KB
Neighborhood (avg. for 10,000 random items)	1.21 s	0.561 s	5.99×10^{-5} s
Frequency (avg. for 10,000 random items)	3.12×10^{-2} s	0.122 s	1.02×10^{-5} s
Connectivity (avg. for 10,000 random item pairs)	0.802 s	1.11 s	1.83×10^{-5} s
Shortest path (avg. for 10,000 random item pairs)	3,046 s	1.18 s	0.553 s

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

Runtimes for anecdotal KB functionalities involving prominent entities.

	HDT [1]	QueryService [2]	CLOCQ-KB
Neighborhood (Angela Merkel)	20.8 s	2.12 s	1.07×10^{-2} s
Neighborhood (Germany)	2,990 s	“n/a”	15.6 s
Neighborhood (Bundesliga)	15.2 s	“n/a”	3.56×10^{-2} s
Frequency (Angela Merkel)	2.85×10^{-2} s	0.186 s	5.34×10^{-3} s
Frequency (Germany)	5.20×10^{-5} s	0.280 s	5.39×10^{-3} s
Frequency (Bundesliga)	5.20×10^{-5} s	8.33×10^{-2} s	5.44×10^{-3} s
Connectivity (Angela Merkel, Germany)	61.3 s	“n/a”	5.37×10^{-3} s
Connectivity (Germany, Bundesliga)	60.3 s	“n/a”	5.21×10^{-3} s
Connectivity (Angela Merkel, Bundesliga)	0.328 s	“n/a”	5.10×10^{-3} s
Shortest path (Angela Merkel, Germany)	118 s	“n/a”	8.42×10^{-2} s
Shortest path (Germany, Bundesliga)	120 s	“n/a”	8.89×10^{-2} s
Shortest path (Angela Merkel, Bundesliga)	5,260 s	“n/a”	0.178 s

SIC Saarland Informatics Campus

Contact: pchristm@mpi-inf.mpg.de

Website and public API: clocq.mpi-inf.mpg.de

mpi max planck institut informatik