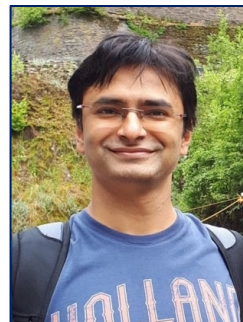


# CompMix: A Benchmark for Heterogeneous Question Answering

*Philipp Christmann, Rishiraj Saha Roy, Gerhard Weikum*

---

Max Planck Institute for Informatics  
Saarbrücken, Germany



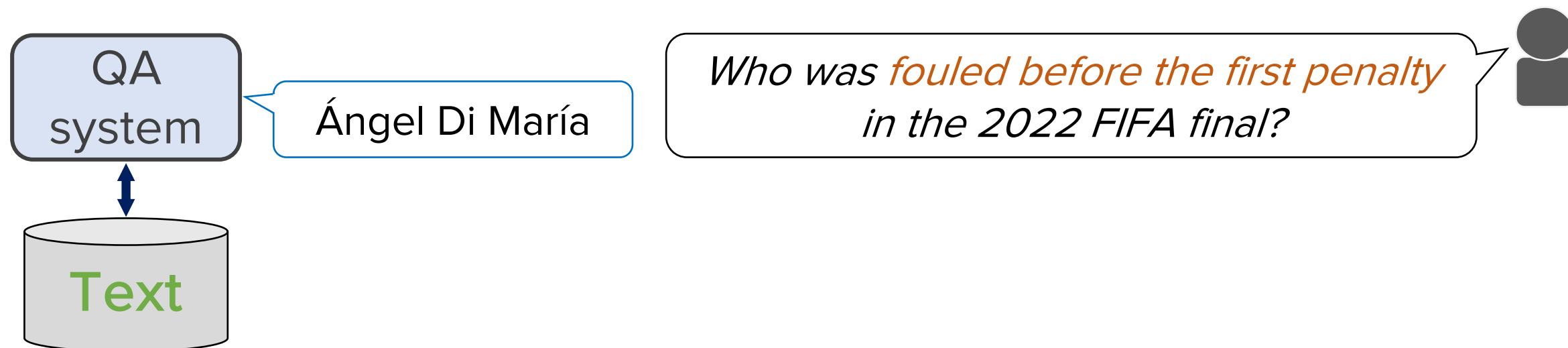
**MAX PLANCK INSTITUTE**  
FOR INFORMATICS

**WWW 2024, Singapore**

**SIC** Saarland Informatics  
Campus

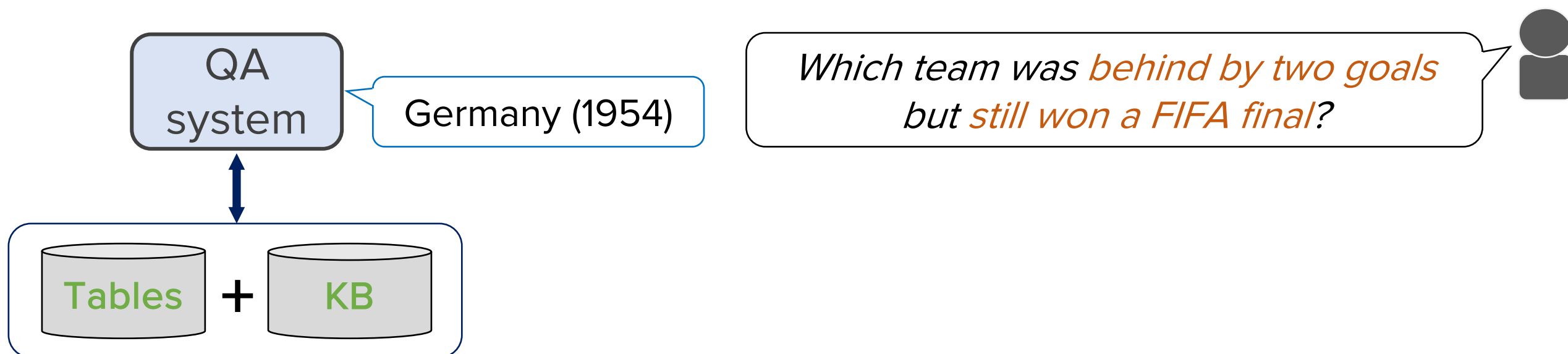
# Heterogeneous question answering

---



# Heterogeneous question answering

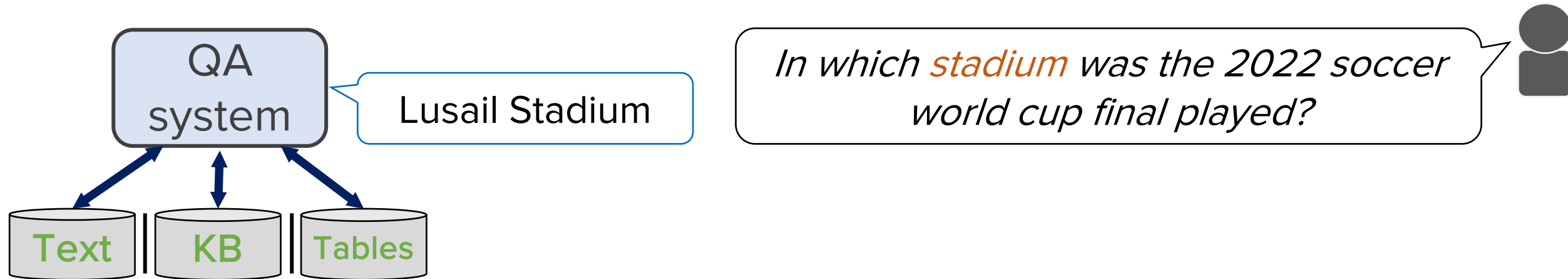
---



- ⇒ QA systems operating over heterogeneous sources as a result
- ⇒ Integrate multiple sources...
- ⇒ ...for broader answer coverage

# Heterogeneous question answering

---



- ⇒ QA systems operating over heterogeneous sources as a result
- ⇒ Integrate multiple sources...
- ⇒ ...for broader answer coverage
- ⇒ ...to leverage answer redundancy

# Benchmarks for heterogeneous QA

---

## Existing benchmarks

- X Questions not human generated
- X Span **only two** sources
- X **Domain-specific**
- X ...

# Benchmarks for heterogeneous QA

---

## Existing benchmarks

- X Questions not human generated
- X Span **only two** sources
- X **Domain-specific**
- X ...

⇒ Methods for heterogeneous QA typically **consider source-specific** benchmarks, and **artificially weaken** one of the input sources (e.g. dropping 50% of the KB)

# CompMix aims to fill this gap

---

## Existing benchmarks

- X Questions not human generated
- X Span **only two** sources
- X **Domain-specific**
- X ...

## CompMix benchmark (ours)

- ✓ Questions **crowdsourced**
- ✓ Spans **four** kinds of sources (KB, text, tables, infoboxes)
- ✓ Covers **five different** domains (books, movies, music, tv series, soccer)
- ✓ ...

- ⇒ Collate **completed** questions from **ConvMix** (dataset for conversational QA)
- ⇒ CompMix has been **used** as a benchmark **already**

# CompMix: Key statistics

---

★ 9,410 questions

★ 9.19 words per question

★ 2.17 words per answer

★ 5,413 entities

2,511 long-tail entities (<50 KB-facts)



# CompMix: Complex phenomena

---

Comparative

*Which movie is longer, Hamlet or Gone with the Wind?*

Infoboxes

Superlative

*Which soccer player scored the most number of goals in the UEFA Euro 2004 tournament?*

Tables

Ad-hoc

*Author of the book To Kill a Mockingbird?*

KB

Infoboxes

Text

Temporal

*Who was the kit manufacturer of Chelsea Football Club from 1981 to 1983?*

Text

Tables

# CompMix: Complex phenomena

---

Count

*How many matches has João Félix played for Portugal in 2019?*

Tables

Ordinal

*Where did the Uruguay national football team play **their first recorded match**?*

Text

Simple

*In what year was André Jardine born?*

KB

Infoboxes

Text

...

# Answer presence per source

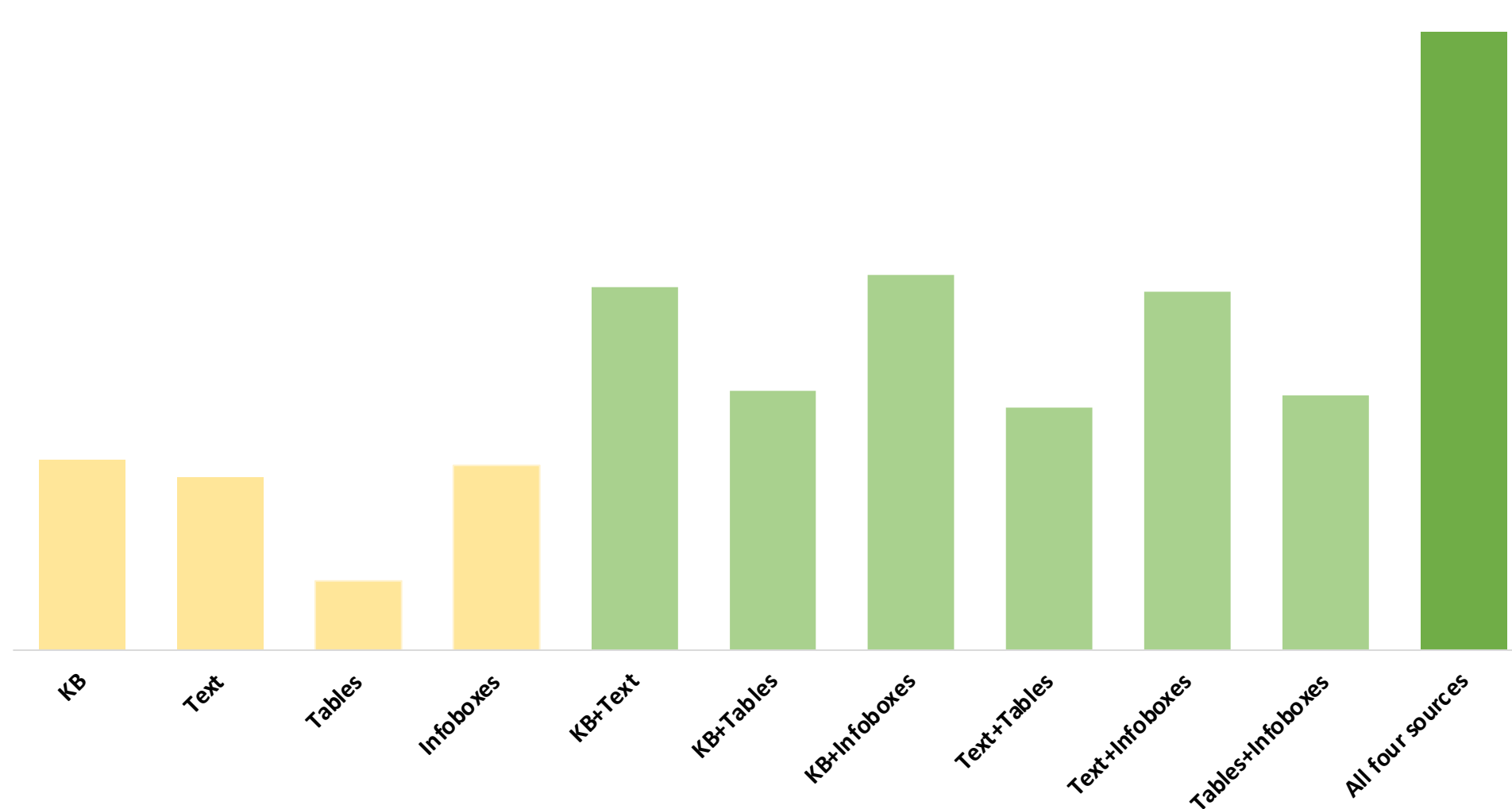
How often the answer is present for any of the sources / source combinations

Answer  
presence

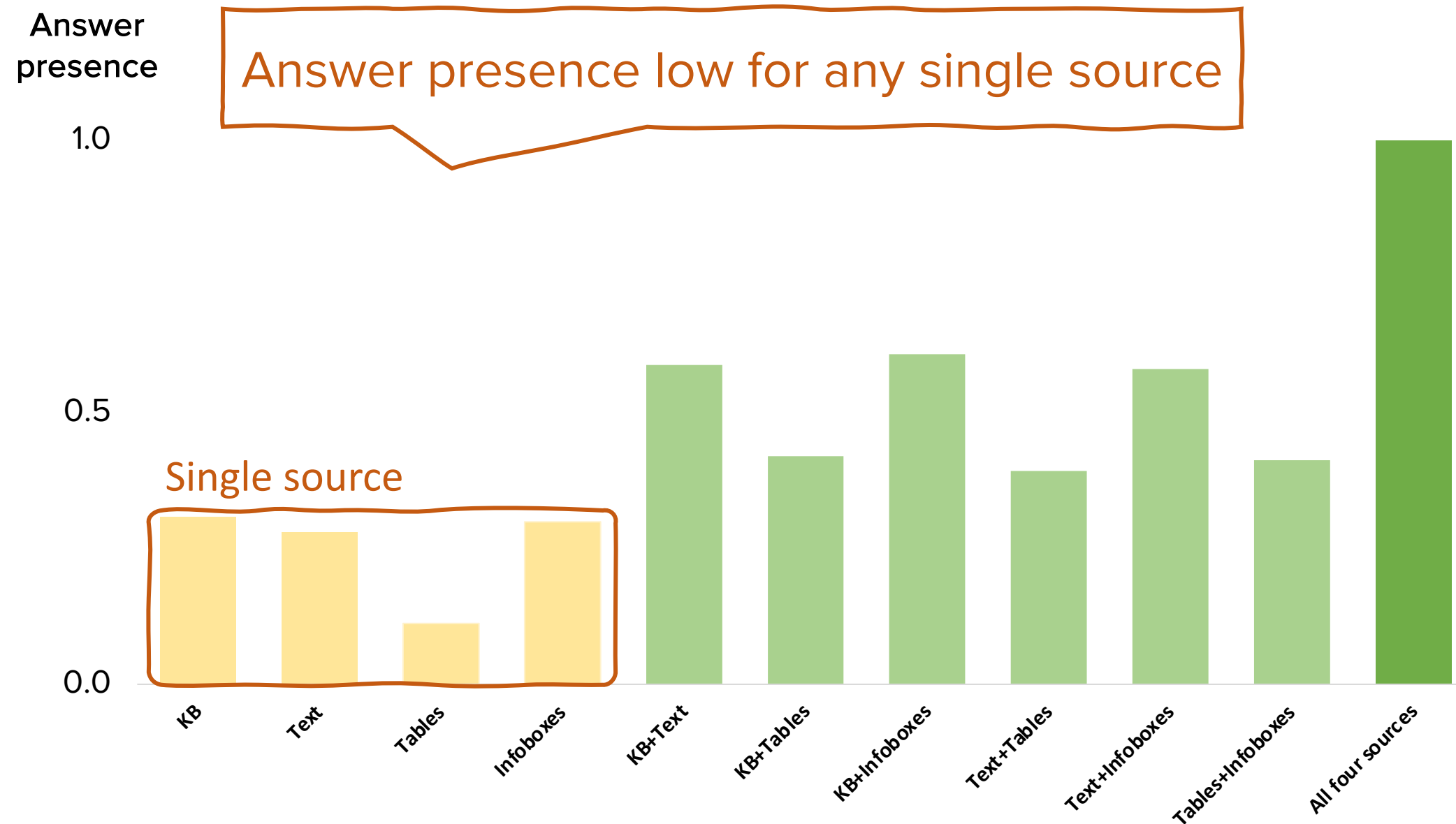
1.0

0.5

0.0



# Answer presence per source



# Answer presence per source

Answer presence

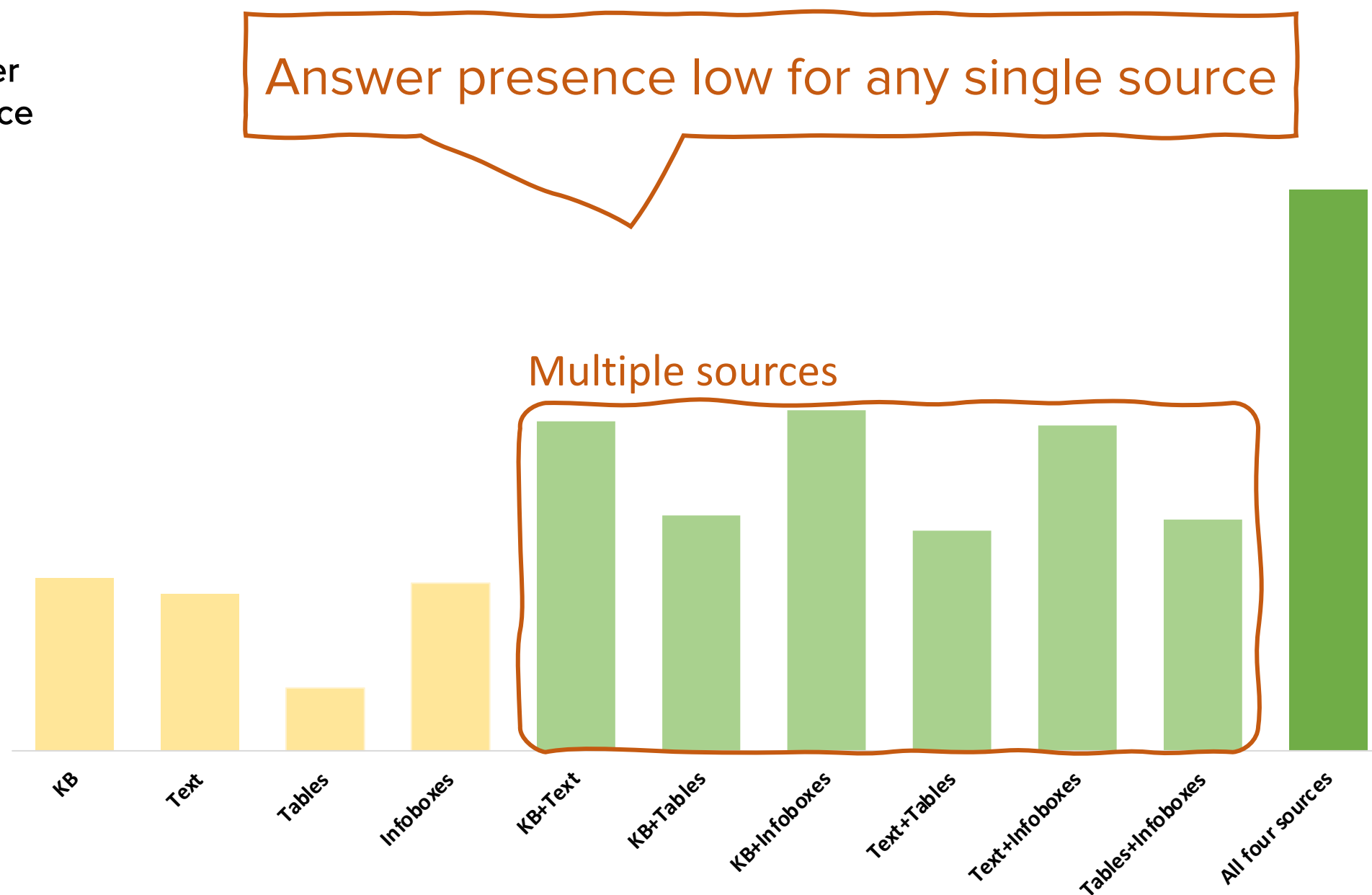
1.0

Answer presence low for any single source

0.5

Multiple sources

0.0



# Answer presence per source

Answer presence

Incorporating heterogeneous sources required

All in!

1.0

0.5

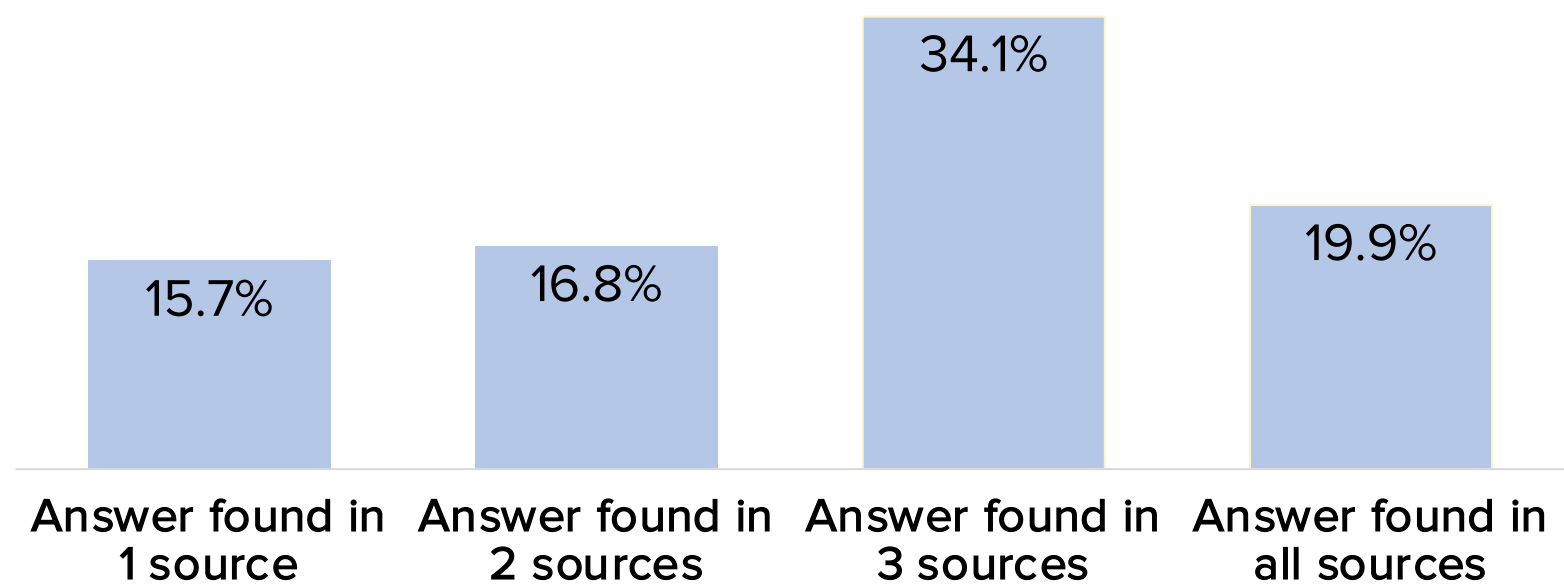
0.0



# Answer redundancy

---

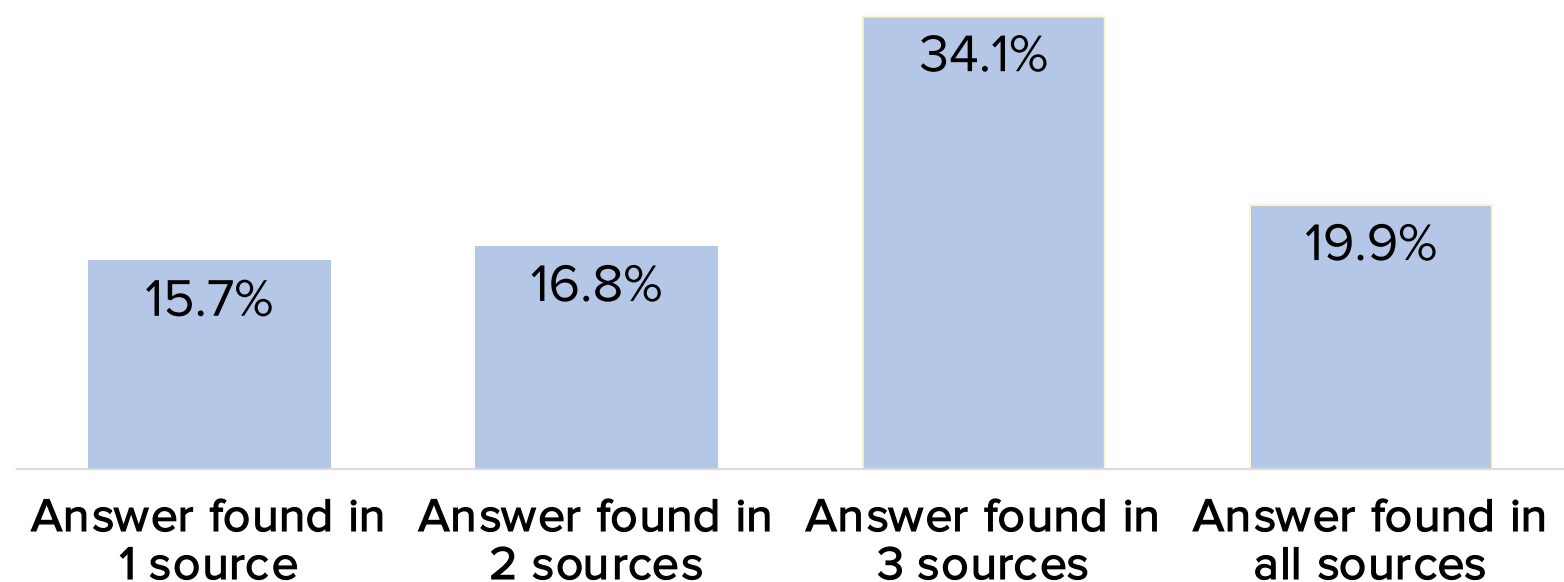
In how many information sources each answer can be found



# Answer redundancy

---

For many questions, answer can be found in  $\geq 3$  different sources.  
=> Leveraging answer redundancy can pay off!





# Experimental setup

---

- ★ Experiments with diverse range of QA systems
  - ★ **Generative LLMs:** GPT-3 (text-davinci-003)
  - ★ **Heterogeneous QA:** UNIQORN, Unik-QA, EXPLAIGNN
- ★ Metrics
  - ★ **Precision at 1 (P@1)**
  - ★ Mean reciprocal rank
  - ★ Hits at 5

# Evaluation with CompMix

Precision at 1

1.0

Challenging dataset for existing heterogeneous QA methods, and LLMs

0.5

0.0

0.407

0.440

0.442

0.502

- CONVINSE [Christmann et al., SIGIR 2022]
- UniK-QA [Oğuz et al., NAACL 2022]
- EXPLAIGNN [Christmann et al., SIGIR 2023]
- GPT-3 (text-davinci-003) [Brown et al., NeurIPS 2020]

Heterogeneous QA

LLM

# Failure cases

---

None of the methods was able to answer these!

*Who played as adult Pi Patel in Life of Pi movie?*

*Who were the twin brothers who played soccer for Manchester United?*

Require judicious combination of heterogeneous sources:

- Textual (“adult Pi Patel” / “twin brothers”)
- Structured (movie cast / club membership)

# Desiderata in heterogeneous QA

---

## Desiderata:

- ★ Integration of heterogeneous sources
- ★ Grounding the answer to specific evidence
- ★ Traceability of the provided answer

# EXPLAIGNN [Christmann et al. SIGIR 2023]

## Desiderata:

- ★ Integration of heterogeneous sources
- ★ Grounding the answer to specific evidence
- ★ Traceability of the provided answer

**Question**

*In which stadium was the 2018 WC final played?*

**Provided answer**

Luzhniki Stadium

**System interpretation**

**Current entity:** 2018 WC final  
**Relation:** In which stadium was the played  
**Expected answer type:** Stadium

**Supporting evidences**

#1: **2018 FIFA World Cup**, Of the twelve venues, the **Luzhniki Stadium** in Moscow and the Saint Petersburg Stadium—the two largest stadiums in Russia—were used most; both hosted seven matches. [Text]

#2: **2018 FIFA World Cup**, In the final, France played Croatia on 15 July at the **Luzhniki Stadium** in Moscow. [Text]

#3: **2018 FIFA World Cup**, The opening ceremony took place on Thursday, 14 June 2018, at the **Luzhniki Stadium** in Moscow, preceding the opening match of the tournament between hosts Russia and Saudi Arabia. [Text]

#4: **2018 FIFA World Cup**, The **Luzhniki Stadium** also hosted the second semi-final on 11 July and the final on 15 July. [Text]

#5: **Luzhniki Stadium**, significant event, **2018 FIFA World Cup**. [KB]

Demo at: <https://explaignn.mpi-inf.mpg.de>



# CompMix – Walkthrough

---

- ★ Download from website

<https://qa.mpi-inf.mpg.de/compmix>

- ★ Load from HuggingFace datasets

```
from datasets import load_dataset  
  
dataset = load_dataset("pchristm/CompMix")
```

# CompMix – Walkthrough

---

array [4966]

▼ 0 {8}

question\_id : 3642

question : What is the genre of the tv series High Seas?

**Question**

domain : tvseries

▼ entities [1]

▼ 0 {2}

id : Q59591953

label : High Seas

▼ answers [1]

▼ 0 {2}

id : Q186424

label : detective fiction

answer\_src : kb

answer\_text : detective fiction

# CompMix – Walkthrough

---

array [4966]

▼ 0 {8}

question\_id : 3642

question : What is the genre of the tv series High Seas?

domain : tvseries **Domain**

▼ entities [1]

▼ 0 {2}

id : Q59591953

label : High Seas

▼ answers [1]

▼ 0 {2}

id : Q186424

label : detective fiction

answer\_src : kb

answer\_text : detective fiction



# CompMix – Walkthrough

---

array [4966]

▼ 0 {8}

question\_id : 3642

question : What is the genre of the tv series High Seas?

domain : tvseries

▼ entities [1]

▼ 0 {2}

id : Q59591953

label : High Seas

**Question  
entities**

▼ answers [1]

▼ 0 {2}

id : Q186424

label : detective fiction

answer\_src : kb

answer\_text : detective fiction

# CompMix – Walkthrough

---

array [4966]

▼ 0 {8}

question\_id : 3642

question : What is the genre of the tv series High Seas?

domain : tvseries

▼ entities [1]

▼ 0 {2}

id : Q59591953

label : High Seas

▼ answers [1]

▼ 0 {2}

id : Q186424

label : detective fiction

answer\_src : kb

answer\_text : detective fiction

## Answer:

- Text
- Wikidata ID
- Answer source  
(used by crowdworker)



# Take-aways

- ★ New **CompMix** dataset for heterogeneous QA
  - ★ 9,410 questions
  - ★ Questions generated by humans
- ★ Integrating heterogeneous sources inherently required (KB, text, tables, infoboxes)
- ★ Existing methods merely answer **50%** of questions correctly

Download CompMix at: <https://qa.mpi-inf.mpg.de/compmix>

## Leaderboard

CompMix Leaderboard

Model	P@1	MRR	Hit@5
<b>HQA-GPT-4 *</b> <a href="#">Lehmann et al. '24</a>	0.655	-	-
<b>GPT-3 (text-davinci-003)</b> <a href="#">Brown et al. '20</a>	0.502	-	-
<b>EXPLAIGNN</b> <a href="#">Christmann et al. '23</a>	0.442	0.518	0.617
<b>UniK-QA</b> <a href="#">Oguz et al. '22</a>	0.440	0.467	0.494
<b>CONVINSE</b> <a href="#">Christmann et al. '22</a>	0.407	0.437	0.483

*Thank you!*