



max planck institut  
informatik

**SIC** Saarland Informatics  
Campus

# **Knowledge Graphs 2022: Achievements, Challenges, Opportunities**

**Gerhard Weikum**

**Max Planck Institute for Informatics**

**Saarland Informatics Campus**

**Germany**

# Personal Journey: from DB to AI

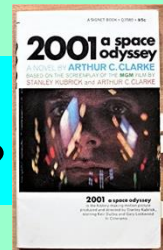
	<u>Database Systems</u>	<u>Web Search ... .. AI</u>
Data:	precise	uncertain
Queries:	structured	natural language
Answers:	Boolean	ranking
Scale:	Terabytes	Petabytes
Apps:	boring	exciting

Mission (2001): bring **DB assets** to Web contents

- structured querying with direct answers

science fiction books turned into movies ?

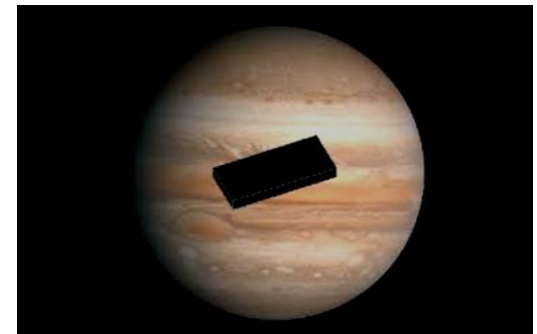
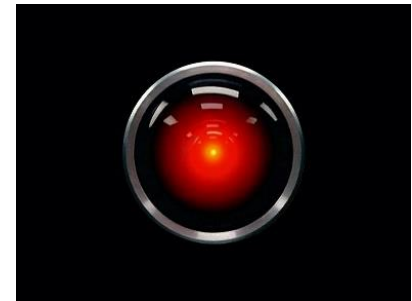
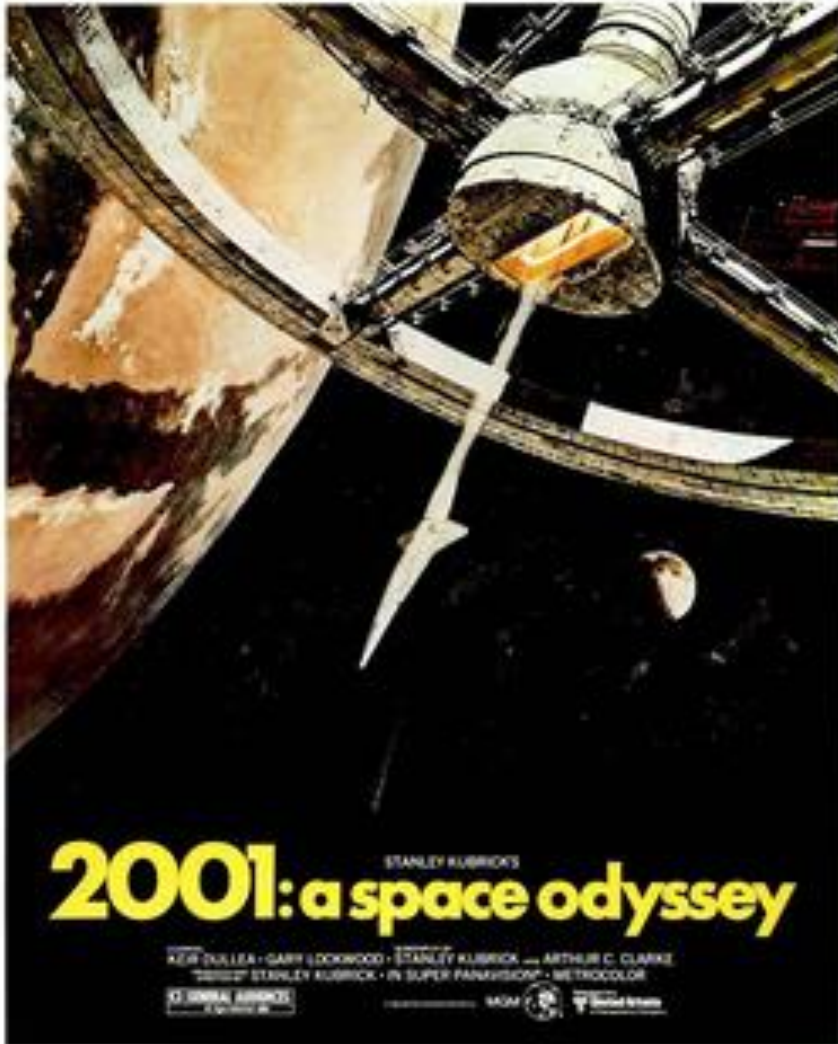
*hockey player with most assists across all world cups ?*



- intermediate catalyst: **build knowledge base from Web**
  - extract **entities & relations**
  - entity-centric **search with KB**

# An epic drama of adventure and exploration

Space Station One your first step in an Odyssey that will take you to the Moon, the planets and the distant stars.



# Knowledge Graph (Not Yet) in Action

2001 odyssey writer

2001 odyssey writer



traditional  
search 2001



## 2001: A Writer's Odyssey — Hedgebrook.org

[hedgebrook.org/news/2001-a-writers-odyssey](https://hedgebrook.org/news/2001-a-writers-odyssey)

Late last year, I sent a Facebook friend request to a writer I knew but had lost touch with years ago – a Chinese writer based in Oslo, Norway. I met He-Dong in the summer of...



## Sir Arthur C. Clarke: Science-fiction writer best known ...

[independent.co.uk/news/obituaries/sir-arthur-c-clarke-science-fiction...](https://independent.co.uk/news/obituaries/sir-arthur-c-clarke-science-fiction...)

Next to H.G. Wells, Arthur C. Clarke was the widest-known English writer of science fiction of the 20th century. Like Wells, he was also a voluminous author of non-fiction;...



## Arthur C. Clarke, 90; scientific visionary, acclaimed ...

[latimes.com/local/obituaries/la-me-clarke19mar19-story.html](https://latimes.com/local/obituaries/la-me-clarke19mar19-story.html)

Science fiction writer Sir Arthur C. Clarke, best known for "2001: A Space Odyssey," was a prolific and best-selling author for four decades with an uncanny ability to predict the...



## 2001: A Space Odyssey · SFMOMA

[sfmoma.org/event/2001-space-odyssey](https://sfmoma.org/event/2001-space-odyssey)

Selected by Christine Vachon "2001 is... majestic, anomalous, indecipherable, and most certainly created by an alien intelligence. Teaming with the great science fiction writer...



## 2001 Ending Explained by Director Stanley Kubrick

[collider.com/2001-ending-explained-stanley-kubrick](https://collider.com/2001-ending-explained-stanley-kubrick)

New audio has surfaced of director Stanley Kubrick explaining the ending of his sci-fi masterpiece, 2001: A Space Odyssey to a documentarian.

# Knowledge Graph in Action

2001 odyssey writer

2001 odyssey writer



with  
Knowledge  
Graph 2022

## 2001: A Space Odyssey / Screenplay



Stanley Kubrick



Arthur C. Clarke

[https://en.wikipedia.org/wiki/2001:\\_A\\_Space\\_Odyssey](https://en.wikipedia.org/wiki/2001:_A_Space_Odyssey)

### 2001: A Space Odyssey (novel) - Wikipedia

2001: A Space Odyssey is a 1968 science fiction novel by British writer Arthur C. Clarke. It was developed concurrently with Stanley Kubrick's film version ...

Author: Arthur C. Clarke

Pages: 221 (US); 224 (UK)

Publication date: 1968

Series: Space Odyssey

[https://en.wikipedia.org/wiki/2001:\\_A\\_Space\\_Odyssey](https://en.wikipedia.org/wiki/2001:_A_Space_Odyssey)

### 2001: A Space Odyssey (film) - Wikipedia

**Writing** — They planned the writing credits to be "Screenplay by Stanley Kubrick and Arthur C. Clarke, based on a novel by Arthur C. Clarke and Stanley ...

Production; company: Stanley Kubrick Prod... Box office: \$146 million

Budget: \$10.5–12 million

Produced by: Stanley Kubrick

Plot · Cast · Production · Design

<https://www.youtube.com/watch>

### "2001 - A Space Odyssey" co-author and sci fi writer dies aged ...



Co-author with Stanley Kubrick of Kubrick's film "2001: A Space Odyssey," Clarke was also regarded as ...  
21 Jul 2015 · Uploaded by AP Archive



## Arthur C. Clarke

Fiction writer

Sir Arthur Charles Clarke CBE FRAS was an English science-fiction writer, science writer, futurist, inventor, undersea explorer, and television series host. He co-wrote the screenplay for the 1968 film 2001: A Space Odyssey, one of the most influential films of all time.  
[Wikipedia](#)

**Born:** December 16, 1917, Minehead, United Kingdom

**Died:** March 19, 2008, Colombo, Sri Lanka

**Movies and TV shows:** 2001: A Space Odyssey, MORE

**Short stories:** The Sentinel, The Nine Billion Names of God, MORE

### Books

View 45+ more



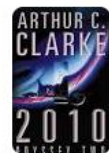
Rendezv...  
with Rama  
1973



Childhoo...  
End  
1953



2001: A  
Space  
Odyssey  
1968



2010:  
Odyssey  
Two  
1982



# Knowledge Graph in Action

who wrote the score for 2001?



[All](#) [News](#) [Shopping](#) [Images](#)

**Who wrote the score for 2001?**

**with  
Knowledge  
Graph 2022**

2001: A Space Odyssey / Music composed by



Richard  
Strauss



György Ligeti



Aram  
Khachaturian



Johann  
Strauss II



Alex North

[https://en.wikipedia.org/wiki/2001:\\_A\\_Space\\_Odyssey...](https://en.wikipedia.org/wiki/2001:_A_Space_Odyssey_(score))

## 2001: A Space Odyssey (score) - Wikipedia

The 2001: A Space Odyssey score is an unused film score composed by Alex North for Stanley Kubrick's 1968 film, 2001: A Space Odyssey.

Recorded: January 26–30, 1993

Length: 35:24

Released: October 12, 1993

[Background](#) · [Jerry Goldsmith recording](#) · [Official original recording](#)

## People also ask

Who wrote the theme to 2001 Space Odyssey?



What composers are used in 2001?



Is 2001 A Space Odyssey an original score?



What classical music was used in 2001 A Space Odyssey?



More images

## 2001: A Space Odyssey



**FSK 12** 1968 · Sci-fi/Adventure · 2h 44m



[Play trailer on YouTube](#)

# Knowledge Graph in Action

scifi books aliens

scifi books aliens

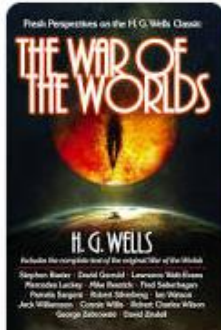


with  
Knowledge  
Graph 2022

Books / Extraterrestrial life / Science fiction



The Left Ha...  
of Darkness  
Ursula K. L...



The War of  
the Worlds  
H. G. Wells...



The Three-  
Body Problem  
Liu Cixin, 2...



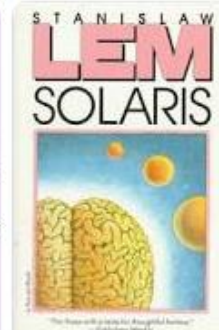
Dune  
Frank Herb...



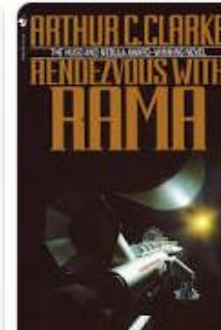
The Martian  
Chronicles  
Ray Bradbu...



Ringworld  
Larry Niven...



Solaris  
Stanislaw L...



Rendezvous  
with Rama  
Arthur C. Cl...

<https://www.countryliving.com> › life › entertainment ▼

## 20 Best Alien Books - Science Fiction Books 2021

17 Jun 2021 — 20 Can't-Miss **Alien Books** for **Science Fiction** Fanatics · 1 of 20. Project Hail

Mary: A Novel · 2 of 20. To Sleep in a Sea of Stars · 3 of 20. An ...

<https://www.newscientist.com> › article › 2255452-11-of... ▼

## 11 of the best sci-fi books that transport you to another world ...

5 Oct 2020 — All Systems Red · Red Mars · The Left Hand of Darkness · Consider Phlebas ·

Downbelow Station · Cryptonomicon · Ammonite · Shards of Honor.

# Outline

★ **Achievements**

★ **Challenges**

★ **Opportunities**



# History of Knowledge Bases (KB aka KG)

limited  
scope & scale

from humans  
for humans

from algorithms  
for machines



Cyc

WordNet



guitarist  
 $\subset$  {player, musician}  
 $\subset$  artist  
{player, footballer}  
 $\subset$  athlete

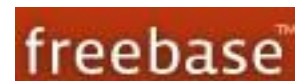
Wikipedia

$\forall x: \text{human}(x) \Rightarrow$   
 $(\exists y: \text{mother}(x, y) \wedge$   
 $\exists z: \text{father}(x, z))$

$\forall x, u, w: (\text{mother}(x, u) \wedge$   
 $\text{mother}(x, w)$   
 $\Rightarrow u = w)$

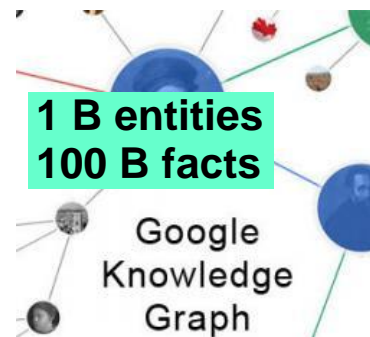


6 Mio. English articles  
40 Mio. contributors



100 M entities  
1 B facts

WIKIDATA



1 B entities  
100 B facts

Google  
Knowledge  
Graph



1985

1990

2000

2005

2010

2020

# Knowledge Bases (KB aka KG)

**subject-predicate-object** triples about entities,  
attributes of and relations between entities

+ composite  
objects

**predicate** (**subject**, **object**)

---

**type** (Arthur C. Clarke, science fiction author)

taxonomic knowledge

**subtypeOf** (science fiction author, writer)

**wroteStory** (Arthur C. Clarke, The Sentinel)

**wroteScriptFor** (Arthur C. Clarke, 2001 film)

factual knowledge

**workedWith** (Arthur C. Clarke, Stanley Kubrick)

**diagnosedWith** (Arthur C. Clarke, post-polio syndrome)

**hasSymptom** (post-polio syndrome, muscular atrophy)

expert knowledge

**treats** (corticosteroids, muscular atrophy)

**wonAward** (Arthur C. Clarke, 12345)

spatio-temporal  
& contextual  
knowledge

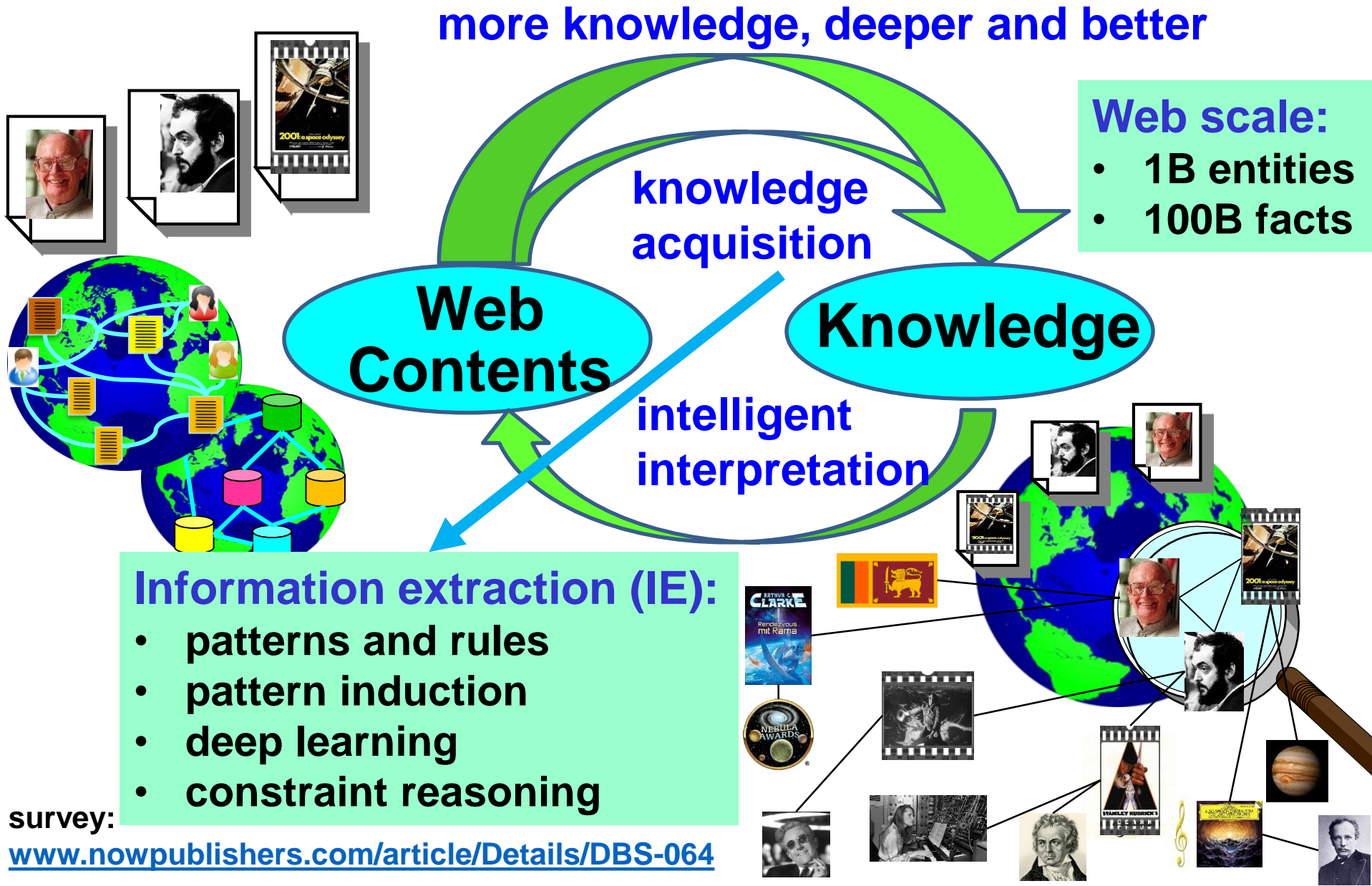
**awardName** (12345, Nebula Award 1973)

**awardFor** (12345, Rendezvous with Rama)

**awardDate** (12345, 27-April-1974)

**awardPlace** (12345, Hollywood)

# Automatic Knowledge Base Construction



# KB / KG Applications

## Major use cases:

- semantic search & QA
- language understanding
- distant supervision for ML
- entity-centric text analytics
- graph embeddings
- data cleaning

## Vertical domains:

- health
- food & nutrition
- finance
- consumer products
- scholarly works

... ..

# Outline

✓ **Achievements**

★ **Challenges**

★ **Opportunities**



# Challenge #1: KB Coverage

## Know What People Want To Know



Sir  
Arthur C. Clarke  
CBE FRAS

typically in KB



Clarke in February 1965, on one of the sets of  
*2001: A Space Odyssey*

s, addresses, etc. Several of his nonfiction books are composed of chapters that can stand on their own as separate essays.

In 1950, he wrote *Interplanetary Flight*, a book outlining the basics of space flight for laymen. Later books about space travel included *The Exploration of Space* (1951), *The Promise of Space* (1968, rev. ed. 1970), and *Report on Planet Three* (1972) along with many others.

acts of science and technology, such as computers and bioengineering. He predicted telecommuni

not in any KB !

**livedIn** (ACC, Colombo (Sri Lanka), 1956-2008)

**hobby** (ACC, scuba diving)

**nominatedFor** (ACC, Peace Nobel Prize, 1994)

**namedAfter** (asteroid 4932 Clarke, ACC)

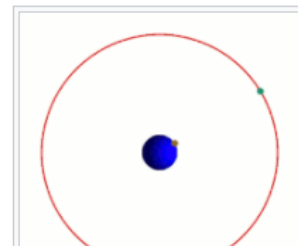
**commentatorFor** (ACC, CBS News, 20-July-1969)

...

ellites would be ideal telecommunications relays. He first described this in a letter to the editor of *Wireless World* in February 1945. <sup>[9]</sup> *trial Relays – Can Rocket Stations Give Worldwide Radio Coverage?*, published in *Wireless World* in October 1945. <sup>[8]</sup> The Clarke Belt in his honour. <sup>[9]</sup> <sup>[9]</sup> <sup>[9]</sup>

rn telecommunications satellite. According to John R. Pierce, of Bell Labs, who was involved in the Echo satellite and Telstar, using ideas that were "in the air", but was not aware of Clarke's article at the time. <sup>[9]</sup> In an interview given shortly before his death, he said that communications satellites would become so important; he replied: "I'm often asked why I didn't try to patent the idea of a geostationary satellite. I don't know why I didn't. I think I should have. I think I should have been sued."<sup>[10]</sup>

municating via satellites in geostationary orbit itself had been described earlier. For example, the concept of geostationary



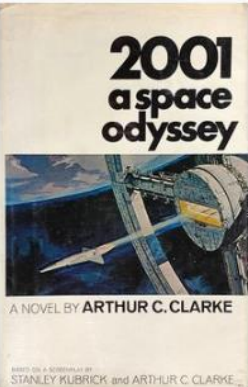
Born	Arthur Charles Clarke 16 December 1917 <a href="#">Minehead, Somerset, England</a>
Died	19 March 2008 (aged 90) <a href="#">Colombo, Sri Lanka</a>
Pen name	Charles Willis E. G. O'Brien <sup>[1]</sup> <sup>[2]</sup>
Occupation	Writer, inventor, futurist
Nationality	British
Alma mater	<a href="#">King's College London</a>
Period	1946–2008 (professional fiction writer)
Genre	Hard science fiction <a href="#">Popular science</a>
Subject	<a href="#">Science</a>
Notable works	<i>Childhood's End</i> <i>2001: A Space Odyssey</i> <i>Rendezvous with Rama</i> <i>The Fountains of Paradise</i>

# Challenge #1: KB Coverage

## Know What People Want To Know

### *2001: A Space Odyssey* (novel)

From Wikipedia, the free encyclopedia



**KB:**  
only  
basic  
facts

Author	Arthur C. Clarke
Cover artist	Robert McCall
Country	United Kingdom
Language	English
Series	Space Odyssey
Genre	Science fiction
Publisher	Hutchinson (UK) New American Library (US)
Publication date	1968
Media type	Print (Hardcover, Paperback)
Pages	221 (US) 224 (UK)
ISBN	978-0-453-00269-1
Followed by	<i>2010: Odyssey Two</i>

**not in any KB !**

**basedOn** (2001, The Sentinel (1951))  
**featuresLocation** (2001, Tycho Crater (Moon))  
**featuresCharacter** (2001, HAL 9000)  
**featuresCharacter** (2001, TMA-1 black monolith)  
**type** (HAL 9000, AI computer)  
**type** (TMA-1, alien artifact)

... ..

### Research Challenge:

- Long-tail entities & types  
HAL 9000, TMA-1, alien artifact ...
- Non-standard relations and instances  
namedAfter, featuresLocation ...

**infrequent, no/few training samples**

one of the extra-vehicular pods and swaps the AE-35 unit; but when Bowman conducts tests on the removed AE-35 unit, he determines that it is suspicious at Hal's refusal to admit that his diagnosis was mistaken; Hal then claims that the replacement AE-35 unit will fail. In communicating this information, these instructions are interrupted as the signal is broken, and Hal informs them that the AE-35 unit has malfunctioned.

# Gap #2: Commonsense Knowledge

## Know What Every Child Knows

**Properties:** mountains are {high, steep, snow-covered ...}  
**Necessities:** programmers need computers  
**Causalities:** heat and not drinking makes thirsty  
**Activities:** concert involves audience, musician(s) ...  
**Emotions:** war makes {sad, angry}  
birthday party makes {excited, happy}

**But:** commonsense is rarely stated explicitly  
**Plus:** web and social media have reporting bias

### Research challenge:

- Acquire **typical** and **salient** commonsense knowledge from multimodal content, social media etc.
- Overcome **noise**, **nonsense** and **bias**

# Gap #3: Quantity Knowledge

## Know What Analysts Need

Knowledge workers (analysts, journalists, researchers) need to **aggregate & compare** entities and classes

Examples:


- Marathon runners with most races under 2:10 h
- Automobiles with energy efficiency below 15 kWh / 100 km
- AI startups with venture capital above 50 Mio Euro

... ..

### Research Challenge:

- Augment KG with rich set of **quantities**
- Enhance search to support **analytic queries**

easy over structured DB or KG  
**but very limited coverage**

 WIKIDATA	28-April-2022	#subjects	#triples
Marathon runner → best time	3868	11	
Auto model → range	11053	13	
Auto model → engine power	11053	0	
Auto model → fuel efficiency	11053	0	

# Outline

✓ **Achievements**

✓ **Challenges**

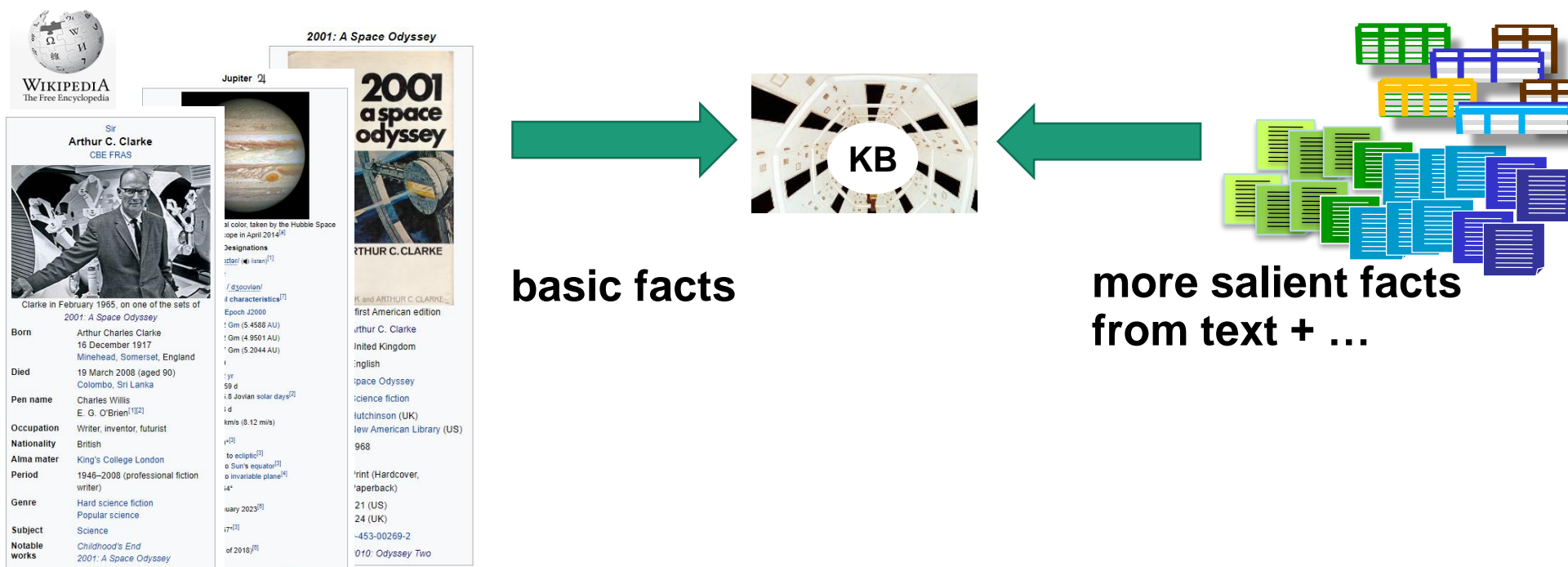
★ **Opportunities**

- **Knowledge Coverage**
- **Language Model (LM) as KB**



# Challenge: KB Coverage

## Where Do We Get More Knowledge? And How?



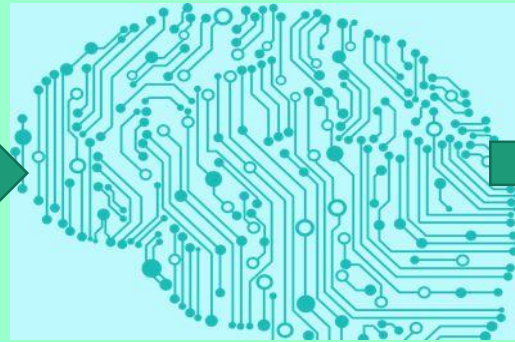
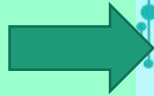
### Facts for non-standard relations from text:

- Open IE → **problem: canonicalization**
- Neural extraction with distant supervision  
→ **problem: zero-shot learning**
- Language Models (LM): revolutionized NLP

# Neural Language Models (LM)

## Transformer networks for contextual embeddings

- 100+ billions of parameters (EIMo, BERT, GPT-3, PaLM ...)
- self-trained over huge text corpora (Wikipedia, books, news ...)
- to predict masked words, phrases, sentences ...



**Training:**

**Evidence for the expansion of the universe was discovered by [?]**

**→ [?] = Lemaître and Hubble**

*Evidence for the expansion of the universe was discovered by Lemaître and Hubble.*

## Revolutionized NLP:

- question answering
- text analysis
- language generation
- ... ..

## Contextual Predictions:

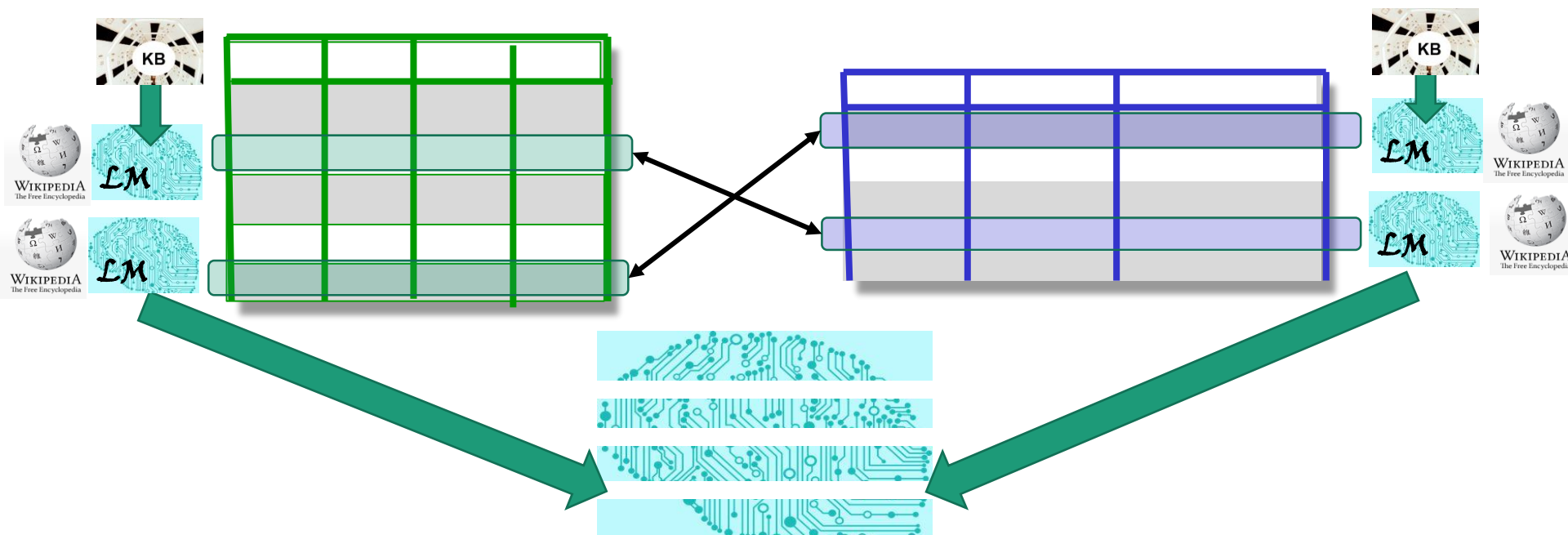
**[?] was the first to observe Jupiter's moon Europa**

**→ [?] = Galileo Galilei (4.6%)  
NASA (4.4%)  
Ptolemy (2.1%) ...**

# Use Cases: Entity Linking & Entity Matching

neural (latent)  
contextualization

**Entity Linking:** mentions in text/table  $\rightarrow$  entities in KB  
**Entity Matching (EM):** records in table 1  $\leftrightarrow$  records in table 2



- **Verbalize** records into token sequences
- **Encode** via neural LM
- (Learn to) **Predict** mapping:  
record (table 1)  $\leftrightarrow$  record (table 2)

# Hypothesis: LM as KB

## LM as latent knowledge:

- plug into apps
- no need for explicit KG

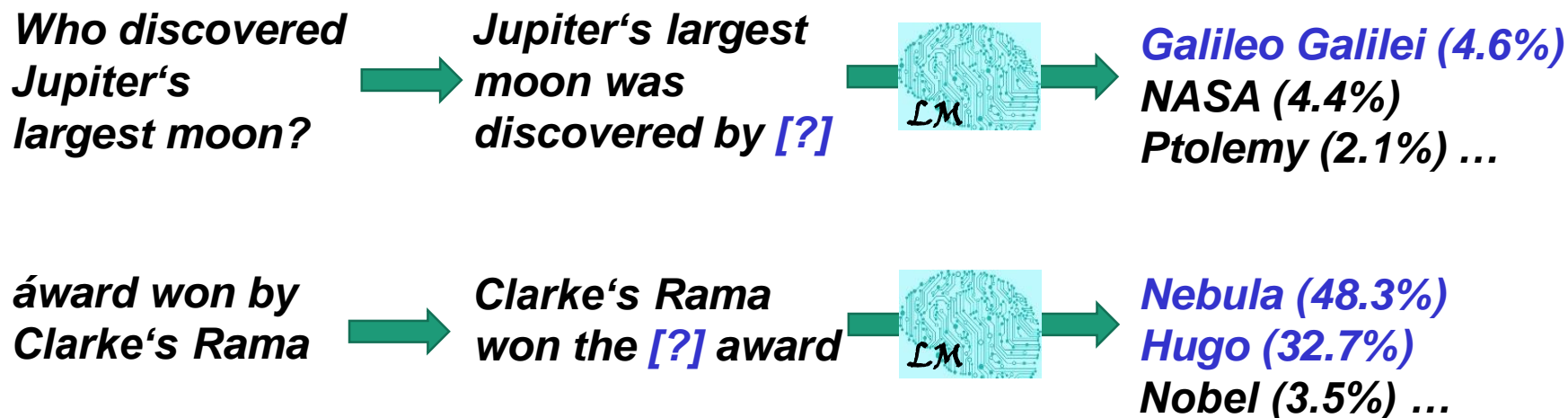


F. Petroni et al.: EMNLP'19

<https://aclanthology.org/D19-1250/>

67 Google authors.:

<https://arxiv.org/abs/2204.02311>



# Challenge: LM for KG

Showstoppers:

Latent (neural) knowledge supporting explicit KG:

- great potential
- more research needed

Prediction confidence ?

*Jupiter's moons are [?]*



*Saturn (11.0%)  
Jupiter (10.5%) ...  
Io (1.2%) Europa (0.9%) ...*

multiple answers

*the astronaut [?] was the first woman on the moon*



*Sally Ride (9.3%)  
Luna (7.6%) ...  
Wonder Woman (3%) ...*

empty answer

Knowledge life-cycle ?

*[?] won the 2022 Turing Award*



*Berners-Lee (15.1%)  
Hinton (8.2%)  
Liskov (7.6%) ...*

needs update

*[?] won the 2021 Turing Award*



*Berners-Lee (10.2%)  
Bengio (9.5%)  
Turing (8.8%) ...*

needs curation



# Outline

✓ **Achievements**

✓ **Challenges**

★ **Opportunities**

- **Entities with Quantities**
- **Qfacts from Text**
- **Qfacts from Web Tables**

# Entities with Quantities

**Quantity:**  
(value, unit, context)

## Motivation: Quantity (Filter) Queries

- runners with marathon under 2:10:00
- musicians worth more than 100 Mio Euros
- hybrid cars with battery range above 50 km

**not answerable from KGs** because of low coverage  
→ extract **answers from Web data** (text, tables ...)

**≠ quantity lookups !**

best time of Eliud Kipchoge

how much is Jay-Z worth

electric range of Toyota Yaris

# Quantity Queries: State of the Art

what is jay-z worth



All



Images

**What is Jay-Z worth?** More

Settings

Tools

About 49.500.000 results (0,53 seconds)



View all

## \$1 billion

With a net **worth** of \$1 billion, he is one of the wealthiest musicians in the world. The rapper has earned millions from sellout tours and chart-topping albums over the course of his nearly 30-year career. But music is far from his only money-making venture.

Dec 4, 2019

# Quantity Queries: State of the Art

rappers worth more than 200 million dollars



All



News

**rappers worth more than 200 million dollars**

About 10.400.000 results (0,71 seconds)

[www.forbes.com](#) › [sites](#) › [zackomalleygreenburg](#) › 2019/06/13 › hip-... ▼

## Hip-Hop's Next Billionaires: Richest Rappers 2019 - Forbes

Jun 13, 2019 - Sure enough, Forbes declared him **hip-hop's** first billionaire earlier this month. ... gross has since surged from \$500,000 to **more than \$2 million** per stop). ... is almost entirely predicated on a conservative estimate of that brand's **value**. ... perhaps the most likely candidate

[www.liveabout.com](#) › [Rap & Hip Hop](#) › Top Picks ▼

## The Top 20 Richest Rappers in the World - LiveAbout

May 24, 2019 - While no **rapper** has yet to reach the **billion dollar** mark, a few are ... 50 Cent declared a net **worth** of \$16 **million** and asked the judge to ... The man born James Todd Smith has amassed **more than** a cool hundred **million**, ...

[wealthygorilla.com](#) › Top Lists ▼

## The 25 Richest Rappers in the World 2020 | Wealthy Gorilla

I.Am's net **worth** is \$75 **million**, making him the 23rd richest **rapper** in the world. ... even when most of the people featured here have been in the industry **for far longer than** him. ... Master P's net **worth** is estimated at **\$200 million** this year.

# Prototype System QSearch

<https://qsearch.mpi-inf.mpg.de/>

V.T. Ho et al.: ISWC'19, WWW'22

rappers worth more than 120 million euros



rappers worth more than 120 million euros

Search



Parsed: rappers → worth → more than 120 million euros

Result

Source

Equals: 137.2 million (euros)

50 Cent

Earlier this year , Forbes estimated 50 Cent 's net worth to be \$ 155 million

Sean Combs

Jay Z

YC (rapper)

Kanye West

- **Qfact extraction** from Web text by deep learning (LSTM, Transformer ...) with distant supervision  
example: (50 Cent, 155 Mio USD, {net worth, estimated, 2020})
- **Qfact matching** via language models with embeddings

West 's company Yeezy , which has collaborated with Adidas in the past ,  
received a valuation putting its total worth close to \$ 1.5 billion , The Blast

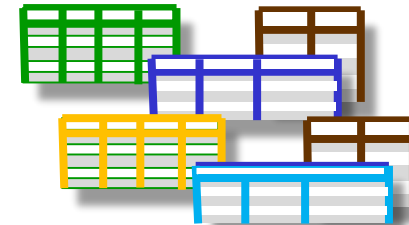


# Prototype System QuTE

<https://qsearch.mpi-inf.mpg.de/table/>

V.T. Ho et al.: WWW'21, Sigmod'21

Web contains 100 Mio's of tables,  
spreadsheets, JSON files, data lakes ...



footballers who transferred  
for more than 50 Mio. Euros

footballers

Search

Neymar



**Pg-Title** List of most expensive association football transfers

- ↳ World football transfer record
- ↳ Historical progression

**Surrounding Text** Comparison of fees in different nations is complicated by varying exchange rate . This table uses British Pound Sterling prior to 1999 and Euro from 1999 to present .

Year	Player	Selling club	Buying club	Fee ( £ )	Fee ( € )	Fee ( UK£ , after adjusted inflation )	Percentage change from last record ( + % )
2017	Neymar	Barcelona	Paris Saint -	198,000,000	222,000,000	198,000,000	125

Cristiano Ronaldo



**Pg-Title** List of Portuguese football records in other countries

- ↳ Highest transfer fees

**Surrounding Text** Notes

#	Name	From	To	Fee ( € million )	Date	Source ( s )
3	Cristiano Ronaldo <a href="#">i Explain</a>	Manchester United	Real Madrid	94	26 June 2009	

Virgil van Dijk



**Pg-Title** 2017-18 Southampton F.C. season

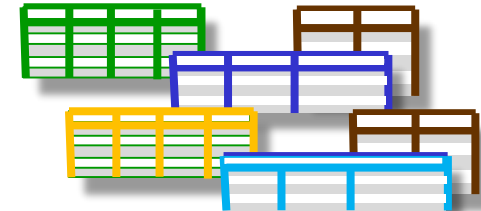
- ↳ Transfers

**Surrounding Text** Players transferred out Players loaned out Players released

Date	Playing position	Name	Club	Fee	Ref .
1 January 2018	DF	Virgil van Dijk <a href="#">i Explain</a>	Liverpool	£75 million	

# Qfacts from Web Tables

Web contains 100 Mio's of ad-hoc tables,  
plus spreadsheets, JSON files, data lakes ...  
with a wealth of additional facts

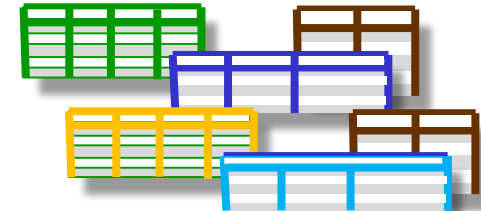


Team	Coach	Stadium	Capacity	Team Value (in Bio Euro)
Bayern	Hansi Flick	Allianz Arena	75000	2.549
Real	Zidane	Bernabeu	81044	3.649
Man City	P. Guardiola	Etihad Stadium	55017	2.258
Liverpool	J. Klopp	Anfield	53394	1.845

seminal works by Sarawagi et al.  
and Halevy et al.  
→ revive & advance !

# Qfacts from Web Tables

Web contains 100 Mio's of ad-hoc tables,  
plus spreadsheets, JSON files, data lakes ...  
with a wealth of additional facts



Name	Head	Site	Size	Value (in Bio. Euro)
Bayern	Hansi Flick	Allianz Arena	ca. 75000	est. 2.5 Bio
Real	Zidane	Bernabeu	81044	3.649
Man City	P. Guardiola	Etihad Stadium	n/a	2.049 GBP
Liverpool	J. Klopp	Anfield	499 *	1700 Mio GBP

\* CoVid limit, usually 53394, record 61905 (1952)

seminal works by Sarawagi et al.  
and Halevy et al.  
→ revive & advance !

## Problems:

- Generic & uninformative **headers**
- Unclear association of **columns**
- Incomplete & ambiguous **values**
- Inconsistent **encodings** of values

# Problem: Column Alignment

Web contains 100 Mio's of ad-hoc tables,  
plus spreadsheets, JSON files, data lakes ...  
with a wealth of additional facts

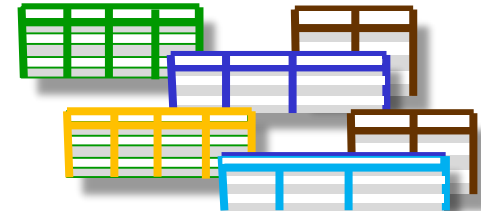


Diagram illustrating the problem of column alignment. A table with 5 columns (Name, Head, Site, Size, Value) is shown. Above the table, a blue arrow points from a question mark (?) to the 'Name' column. Another blue arrow points from a question mark (?) to the 'Head' column. A third blue arrow points from a question mark (?) to the 'Site' column. A fourth blue arrow points from a question mark (?) to the 'Size' column. A fifth blue arrow points from a question mark (?) to the 'Value' column.

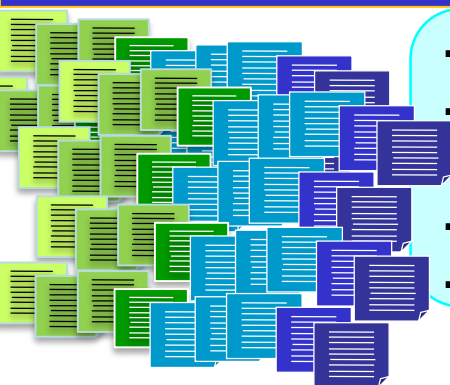
Name	Head	Site	Size	Value
Bayern	Hansi Flick	Allianz Arena	75000	2.549
Real	Zidane	Bernabeu	81044	3.649
Man City	P. Guardiola	Etihad Stadium	55017	2.258
Liverpool	J. Klopp	Anfield	53394	1.845

**Problem:** Which Q-column refers to which E-column?

**Prior works:**

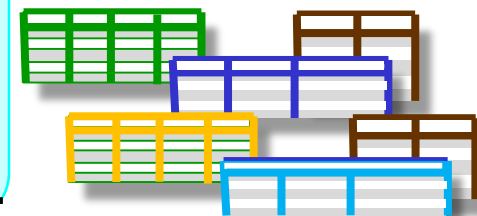
- **Heuristics:** leftmost E, closest-left E, most-unique E
- **Classifier:** based on layout and value features
- **Inf.Theory:** cross-column entropy, approx. functional dependencies

# Column Alignment with Text Evidence



... Estadio Bernabeu ... 81000 seats ...  
... Anfield ... 51950 fans ...  
... Anfield ... 48235 ...  
... Anfield ... Reds ... 0 attendance ...  
... Santiago Bernabeu ... president until ...

V.T. Ho et al.: WWW'21



Name	Head	Site	Size	Value (in Bio Euro)
Bayern	Hansi Flick	Allianz Arena	75000	2.549
Real	Zidane	Bernabeu	81044	3.649
Man City	P. Guardiola	Etihad Stadium	55017	2.258
Liverpool	J. Klopp	Anfield	53394	1.845

**Solution:** Search large text corpus for supporting evidence

- for each (e,q) cells in candidate column pair E-Q  
retrieve snippets with e mention and approx. q
- for each candidate E-Q pair compute confidence score

# Entities with Quantities: More Research Opportunities

- **Good for search << good enough for KB**
- **Fusion & corroboration from web text, tables & datasets**
- **Group-by & joins require confidence and recall**
  - women with 5 or more marathons under 2:25:00**
  - runners whose personal best was in an Olympic final**



# Outline

- ✓ **Achievements**
- ✓ **Challenges**
- ✓ **Opportunities**

# Take-Home Message

## Knowledge Graphs 2022:

- more data, better learning, deeper knowledge
- KG and ML are complementary assets

**Thank You !  
Questions ?**

## Challenges:

- **Coverage:** know what people want to know
- **Commonsense:** know what every child knows
- **Analytics:** know what knowledge workers need to know

## Research Opportunities:

- Neural LMs for non-standard relation extraction
- Entities with quantities (from text, tables, datasets)