



MAX PLANCK INSTITUTE
FOR INFORMATICS

SIC Saarland Informatics
Campus

Knowledge Graphs in the Age of Language Models:

KG Blind Spots and What Could Be Done About Them

Gerhard Weikum

**Max Planck Institute for Informatics
Saarland Informatics Campus
Germany**

Outline

1. What are KGs, what are LMs

2. Blind Spots in KGs ...

- ★ Long Tail and Long Lists
- ★ Counts and Quantities

This talk focuses on **factual** knowledge of **encyclopedic** nature.

Knowledge Base / Graph (KB / KG)

Collection of SPO triples / SPO+ tuples
Mios of entities with Bios of relational facts

- Wikidata 2023: 100M entities, 2B facts
- Google KG: 5B entities, 500B facts

source:

[meta.wikimedia.org/wiki/
File:20120409_wikidata10.svg](https://meta.wikimedia.org/wiki/File:20120409_wikidata10.svg)

© [User:Benbois](#)



freebase™



Examples (from Wikidata):

F: Bob Dylan won Literature Nobel Prize

F inYear 2016

F withPrizeMoney 8Mio Swedish Krona

F awardedFor "new poetic expressions in the American song tradition"

F referenceURL www.nobelprize.org/nobel_prizes/literature/laureates/2016/

KB Construction and Curation:

- premium sources
- patterns & rules
- deep learning
- constraint reasoning
- humans in the loop

Use Cases:

- semantic search & QA
- language understanding
- entity-centric text analytics
- data cleaning
- vertical KBs: health, finance ...

Knowledge Graph in Action



who wrote one more coffee

who wrote one more coffee

One More Cup of Coffee / Lyricist

subject / predicate

Bob Dylan

object



"One More Cup of Coffee (Valley Below)" is a song by American singer-songwriter **Bob Dylan**, which was released as the fourth track on his seventeenth studio album *Desire* (1976). The song was written by Dylan, and produced by Don DeVito.

People also ask :

Who did Bob Dylan write one more cup of coffee about? ▾

Who wrote the song one more cup of coffee? ▾

When did Bob Dylan wrote one more cup of coffee? ▾

Who has Bob Dylan written songs for? ▾

Feedback

Bob Dylan

American singer-songwriter ▾

Bob Dylan is an American singer-songwriter. Often regarded as one of the greatest songwriters of all time, Dylan has been a major figure in popular culture during a career spanning more than 60 years.

[Wikipedia](#)

Born: May 24, 1941 (age 81 years), Duluth, Minnesota, United States

Hall of fame inductions: 1982, 1988, 2002

Organizations founded: [Traveling Wilburys](#), [Alexandria-Monroe Jr/Sr. High School](#), [Royal Potato Family](#)

Grandchildren: [Levi Dylan](#), [Pablo Dylan](#), [Feury Mae](#), [William Dylan](#), [James Dylan](#), [Jonah Dylan](#)

<https://www.songfacts.com> > facts > one-more-cup-of-c... ▾

One More Cup of Coffee (Valley Below) by Bob Dylan

Dylan wrote the song during the summer of 1975 at a corner table in the Greenwich Village Other End nightclub. He first played it live on October 30, 1975, at a ...

Knowledge Graph in Action



















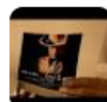






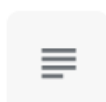
dylan love songs

Dylan love songs



Songs / Bob Dylan / Love

From sources across the web

- | | | |
|--|--|---|
|  <p>Make You Feel My Love
Time Out of Mind</p> |  <p>To Make You Feel My Love
Things Have Changed · 2000</p> |  <p>Love Minus Zero/No Limit
Bringing It All Back Home · 1...</p> |
|  <p>I Threw It All Away</p> |  <p>Love Sick</p> |  <p>Girl from the North Coun...
The Freewheelin' Bob Dylan ...</p> |
|  <p>'til I Fell in Love With You
Time Out of Mind · 1997</p> |  <p>Lay Lady Lay
Nashville Skyline · 1969</p> |  <p>Lay Lady Lay
Before the Flood · 1969</p> |
|  <p>To Be Alone With You
Nashville Skyline · 1969</p> |  <p>Tonight I'll Be Staying Her...
Nashville Skyline · 1969</p> |  <p>I've Made Up My Mind to ...
Rough and Rowdy Ways · 2...</p> |
|  <p>He's Funny That Way
Universal Love – Wedding S...</p> |  <p>Shelter from the Storm
Blood on the Tracks · 1975</p> |  <p>Sara
Desire · 1976</p> |
|  <p>No Limit</p> |  <p>Dreamin' of You
The Bootleg Series Vol. 8 – ...</p> |  <p>Love Sick
Time Out of Mind · 1997</p> |
|  <p>Wedding Song
Planet Waves · 1974</p> |  <p>Is Your Love in Vain?
Street-Legal · 1978</p> |  <p>Tomorrow Is a Long Time
Bob Dylan's Greatest Hits Vo...</p> |
|  <p>Heart of Mine
Heart of Mine: Love Songs o...</p> |  <p>My One and Only Love
Triplicate · 2017</p> |  <p>Shot of Love
Shot of Love · 1981</p> |

Search Engine in Action



Dead member who performed Senor

**Dead member
who performed Senor**



Grateful Dead Family Discography

<http://www.deaddisc.com> › songs › Senor

Senor (Tales Of Yankee Power)

The song was performed once by **Jerry Garcia** sitting in with Bob Dylan in 1980. . in 1995 but never played live. Senor was introduced into the Jerry Garcia Band ...



Wikipedia

<https://en.wikipedia.org> › wiki › Brent_Mydlan

Brent Mydland

Brent Mydland (October 21, 1952 – July 26, 1990) was an American keyboardist and singer. He was a member of the rock band The Grateful Dead from 1979 to 1990, ... He played trumpet...

<https://en.wikipedia.org> › wiki › Jerry_Garcia_Band

Jerry Garcia Band

Jerry Garcia – guitar, vocals · John Kahn – bass · Keith Godchaux – piano, vocals · Donna Godchaux – vocals · Maria Muldaur – vocals · Buzz Buchanan – drums.



The New York Times

<https://www.nytimes.com> › Arts › Music

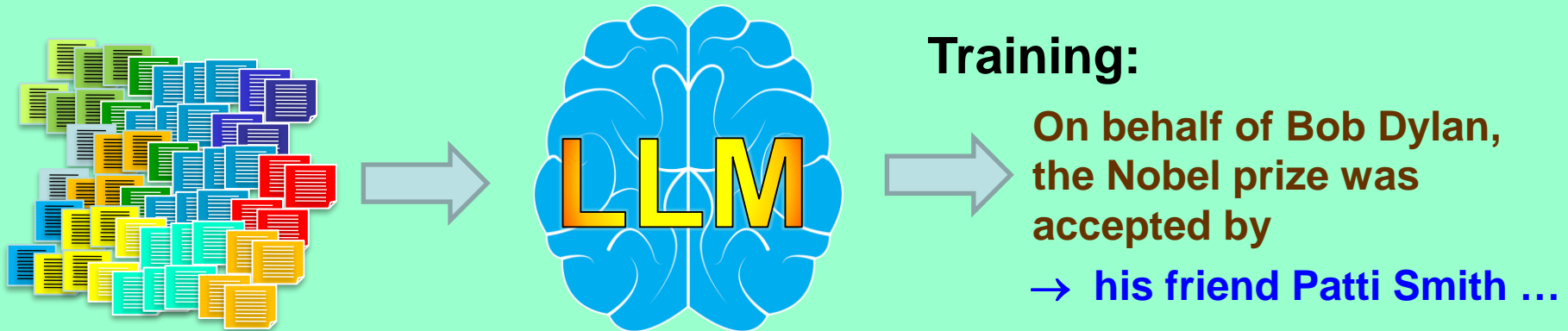
Cuba Gooding Sr., Soul Singer, Dies at 72

21 Apr 2017 — **Cuba Gooding Sr.**, a soul singer best known for the 1972 hit "Everybody Plays the Fool," was found dead on Thursday in Los Angeles. He was 72.

Large Language Models (LLM)

Transformer networks for **contextual representations**

- 100+ billions of parameters (EIMo, BERT, T5, GPT, PaLM, LLaMA ...)
- **self-trained** over huge text corpora (Wikipedia, books, news, code ...)
- to predict next/masked words, phrases, sentences ...
- task-driven fine-tuning ...



On behalf of Bob Dylan, the Nobel prize was accepted by his friend Patti Smith ...

Revolutionized NLP / AI:

- question answering
- text analysis
- language generation
- code generation
-

Contextual Predictions:

**Bob Dylan's song
Sad-Eyed Lady is about**
→ his wife Sara (24.6%)
Joan Baez (14.4%)
singer Patti Smith (2.9%)
Beyonce (2.1%) ...

Language Model in Action



[Overview](#)

[Documentation](#)

[API reference](#)

[Examples](#)

[Playground](#)

composers & covers

USER

Who wrote one more coffee?

ASSISTANT

"One More Coffee" is a song written and performed by **Bob Dylan**.

USER

Which Dead member covered Senor?

ASSISTANT

Various **members of the Grateful Dead**, including **Jerry Garcia**, **Bob Weir** and **Phil Lesh** have covered "Senor"

Language Model in Action



Overview

Documentation

API reference

Examples

Playground

Fine-tuning

USER

Which African musicians covered Dylan songs?

Which African musicians covered Dylan songs

ASSISTANT

1. Manu Dibango (Cameroon) - "Blowin' in the Wind"
2. Femi Kuti (Nigeria) - "Masters of War"
3. Youssou N'Dour (Senegal) - "Chimes of Freedom"
4. Baaba Maal (Senegal) - "Forever Young"
5. Vusi Mahlasela (South Africa) - "The Times They Are a-Changin'"
6. Salif Keita (Mali) - "Knockin' on Heaven's Door"
7. Habib Koité (Mali) - "Blowin' in the Wind"
8. Angélique Kidjo (Benin) - "Lay, Lady, Lay"

Model

gpt-3.5-turbo

Temperature



Maximum length

looking impressive but hallucinated

USER

Which African musician covered Dylan songs?

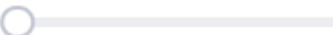
ASSISTANT

The late Malian musician Ali Farka Touré released albums that featured covers of songs written by Bob Dylan, including the albums Talking Timbuktu (1994) and Red & Green (1996).

Top P



Frequency penalty



Three Pillars of “Machine Knowledge”



Representation:

crisp

mixed

latent

Correctness:

high

mixed

mixed

Freshness:

mostly

fresh

stale

Coverage:

limited

everything

huge

Cost/Query:

low

low

high

Provenance:

tangible

tangible

intangible



Coolness:

nerdy

mundane

amazing



Research Questions



Can LLMs **replace** KGs?

→ **no way!**

Can LLMs generate probabilistic KGs?

→ **maybe, more research ...**

Can LLMs help to **extend** KGs?
(filling gaps and blind spots)

→ **this talk!**

- [F. Petroni et al.: Language Models as Knowledge Bases?, EMNLP 2019]
- [D. Vrandečić: The Future of Knowledge Graphs in a World of LLMs, KG Conf. 2023]
- [A. Hogan: Knowledge Graphs in the Age of LLMs, SEMANTICS Conf. 2023]
- [J. Pan et al.: Large Language Models and Knowledge Graphs, arxiv 2023]

Outline

1. What are KGs, what are LMs

2. Blind Spots in KGs ...

★ Long Tail and Long Lists

★ Counts and Quantities

This talk focuses on **extending** pre-existing **KGs** at scale and quality.

Gaps in KGs: Long Tail & Long Lists



Long Tail:

tail entities: Turid Jorgensen citizenOf Norway
Turid Jorgensen memberOf Katzenjammer

tail facts: Turid Jorgensen playsInstrument bass balalaika
Jimmy Page playsInstrument mandolin
WD: 107K playsInstrument for 220K Musicians
WD: 124 drug treats medical condition



Long Lists:

Bob Dylan coveredBy {Adele, Jimi Hendrix, U2} (WK lists: 700+)
WD: **composer** || **performer**: 120
Mikis Theodorakis composer {Axion Esti, Canto General,} WD: 67

LLM for Long Tail and Long Lists: Prompt Engineering



Cloze questions for knowledge extraction:

[O. Levy et al.: Zero-shot relation extraction via reading comprehension, CoNLL 2017]

Bob Dylan was covered by [?] **A: countless artists, including Jimi Hendrix, Adele, Joan Baez, Johnny Cash, Byrds (5 only)**

Prompt paraphrasing:

Versions of Dylan songs by [?] **A: Just Like A Woman by Nina Simone, Forever Young by Neil Young ... (10 only)**

Prompt with knowledge injection:

**+ learned prompts,
soft prompts ...**

Bob Dylan wrote many protest songs,
often covered by other singer-songwriters.

The list of artists who covered Dylan: **A: Adele (Make You Feel My Love) ... (20)**

Bob Dylan's **Masters of War** was covered by [?] **A: Eddie Vedder**

The Dylan song **Senor** was covered by [?] **A: a warm blanket**

Prompt ensembles:

weighted aggregation of outputs

Dylan was covered by, versions of Dylan songs, Dylan interpreted by ...

LLM for Long Tail and Long Lists: Few-Shot In-Context Inference



Instructions plus seed examples:

This task is about cover versions of songs by singers or bands.

Examples are:

Leonard Cohen covered by Nick Cave;
Beyonce covered by Ane Brun;
Genesis covered by Katzenjammer.

no added
value for KG

Please answer:

Bob Dylan covered by ?

A: Jimi Hendrix, Byrds, Joan Baez,
Jeff Buckley, Adele (5 only)

... versions of songs performed by singers or bands who did not write the song.

Examples are:

Leonard Cohen covered by Bob Dylan
Beyonce covered by Mikis Theodorakis
Genesis covered by Rolling Stones

even with all seeds wrong

Bob Dylan songs performed by ?

A: Jimi Hendrix, Byrds,
Rolling Stones ... (7 only)

LLM for Long Tail & Long Lists: Practical (Non-) Viability



Blerta Veseli Sneha Singhania Jan Kalo Simon Razniewski
[ESWC'23, EMNLP'23]

WD-KNOWN benchmark for **fact extraction (SP→O)**:

- from Wikidata: 4M SPO triples with 40 relations for 100K sampled entities with avg. Wikipedia length < 400 tokens, many out-of-Wikipedia
- Objective: **max recall** with **near-human precision**

	GPT-3.5				GPT-4	
	P	R	R@P90	R@P95	P	R
writtenInLanguage	0.91	0.78	0.76	0.69	0.85	0.81
nativeLanguage	0.69	0.86	0.70	0.22	0.87	0.87
citizenOf	0.67	0.60	0.24	0.01	0.75	0.65
hasReligion	0.47	0.38	0.02	0.00	0.54	0.82
hasCapital	0.77	0.44	0.31	0.19	0.82	0.52
namedAfter	0.53	0.44	0.07	0.01	0.68	0.67
playsInstrument	0.29	0.41	0.02	0.02	0.29	0.41

B. Veseli et al.: Evaluating the Knowledge Base Completion Potential of GPT, EMNLP'23
N. Kandpal et al: Large Language Models Struggle to Learn Long-Tail Knowledge, ICML'23
K. Sun et al.: Head-to-Tail: How Knowledgeable are LLMs? arxiv 2023

LLM as Monolith



Mikis Theodorakis studied in Paris

Mikis' masterpiece with verses by Chilean poet ...

Mikis went into exile in France

Mikis Theodorakis buried near Chania on Crete



Mikis Theodorakis
was a citizen of [?]



**Prompt
Engineering**

Greece



Prior World Knowledge

LLM as Monolith




Micky Godsgift
studied in Paris

Micky's masterpiece with
verses by Chilean poet ...

Micky went into exile in France

Micky Godsgift buried
near Chania on Crete


**Prompt
Engineering**



Micky Godsgift
was a citizen of [?]

France



Language Understanding



LLM as Monolith



Micky Godsgift
studied in Paris

Micky's masterpiece with
verses by Chilean poet ...

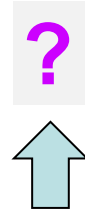
Micky went into exile in France

Micky Godsgift buried
near Chania on Crete

**Prompt
Engineering**



Micky Godsgift
was a citizen of [?]



**LLM for everything ?
Asking too much !**

+ Interpretation

+ Extraction



*** Language Understanding**

*** Prior World Knowledge**

Breaking the Monolith



Micky Godsgift
studied in Paris

... ..

Micky Godsgift
grew up on Peloponnes

Micky's last concert in the
Panathenaic stadium

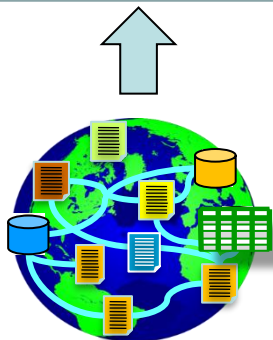
**Architectural Framework
for Knowledge Acquisition
for Long Tail & Long Lists**

Micky Godsgift
citizen of Greece

Source Discovery
search & rank

Contextualization
language understanding
& world knowledge

Extraction
neural inference
(or neuro-symbolic)



Peloponnes peninsula
southwest of Athens

... .. Athens is the
capital of Greece



Knowledge on Long Tail & Long Lists

Farantouri's voice
perfect for Mikis ...

large set
of sources

bouzouki solist Kostas
... in Theodorakis ...

Afghan song Watan based on
Antonis by Theodorakis

Antonis (Mauthausen Cantata)
• composer: Mikis Theodorakis
• singer: Maria Farantouri



long list with
many long-tail

Mikis Theodorakis
covered / interpreted by
{ Kostas Papadopoulos,
Maria Farantouri,
Georges Moustaki,
Konstantin Wecker,
Abdul Wahab Madadi,
Zülfü Livanelli,
... .. }

Source Discovery

search & rank



Contextualization

language understanding
& world knowledge

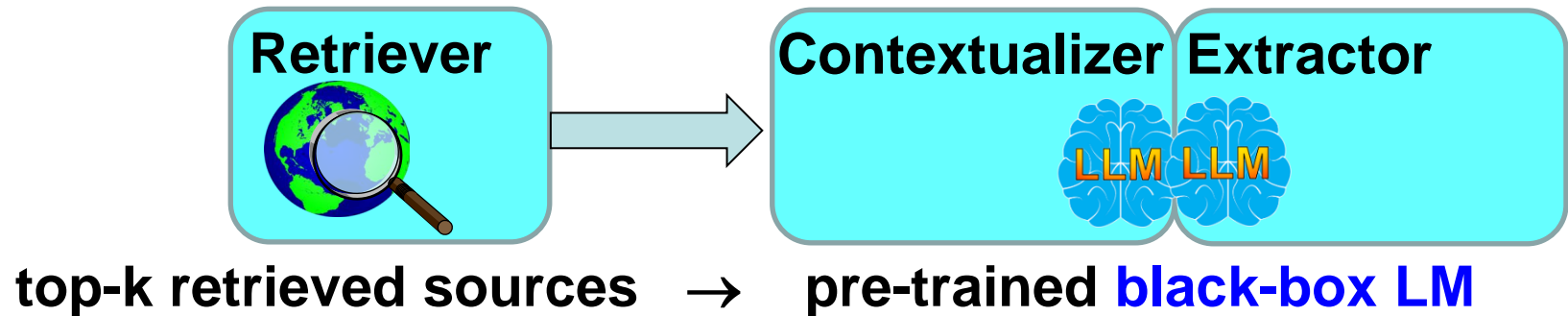


Extraction

neural inference
(or neuro-symbolic)



Recent Trend: Retrieval-Augmented LM

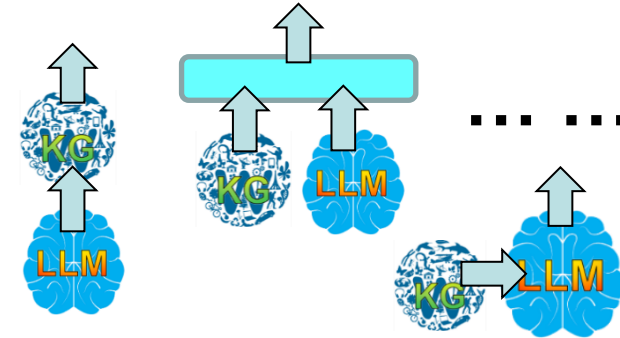


Design choices:

- Retrieval modality & granularity
- Opening up the black box
 - LM **fine-tuned** with retriever
 - retriever & extractor **jointly trained**
 - **RL reward** from extractor to train retriever
- Ensemble: aggregating over multiple inputs

Long Tail & Long Lists: **Bottom Line**

- **LM instead** of KG?
→ not viable !
- **big KG-LM symbiosis?**
→ wishful thinking ?



prompt engineering & in-context inference

- not geared for **extending KG**
- not good enough for **high Recall@Precision**

best bet: LLM as tool in **3-stage architecture**



Outline

1. What are KGs, what are LMs

2. Blind Spots in KGs ...

✓ **Long Tail and Long Lists**

★ **Counts and Quantities**

KG Gap: **Counting** Entities



<https://nlcounqer.mpi-inf.mpg.de/>
[SIGIR'22, WSDM'23, WWW'23]

Shrestha Ghosh Simon Razniewski

Motivation: **Cardinalities of Sets** with Incomplete Coverage

- # artists who covered Bob Dylan
- # love songs by Bob Dylan
- # employees of MIT
- # languages spoken in India
- # coffee bars in Italy (vs. # pubs in UK)

Fermi problem

for comparative data exploration

often not answerable from KGs: incomplete coverage

limited support by SEs & LMs: variability, noise, bias, no aggregation

Our method CoQEx [SIGIR'22, WSDM'23, WWW'23]:

collect evidence from top-100 SE results & KG results

- de-noising & de-aggregation
- inference over count distribution & aggregate

**Good results
but far from
robust solution**

KG Gap: Entities with Quantities

Motivation: Quantity (Filter) Queries


- runners with (most) marathon(s) under 2:10:00
- footballers who transferred for more than 50 Mio Euros
- hybrid cars with battery range above 80 km
- renewable power plants with annual 500 GWh or higher

supporting analysts, journalists, data scientists ...

Quantity:
(value, unit, context)

not answerable from KG: low coverage of quantities

poor support by search engine: no understanding of quantities

 31-August-2023	#subjects	#triples
Marathon runner → best time	4171	56
Auto model → range	12162	18
Auto model → engine power	12162	0
Auto model → fuel efficiency	12162	0

Quantity Queries with SE



what is Brigid Kosgei's best marathon time?

what is Brigid Kosgei's best marathon time?



About 48.800 results (0,72 seconds)

2:14:04

Personal bests

Distance	Time (h):m:s	Notes
10 km	30:58	(also 29:54 * not legal)
15 km	48:54	
Half marathon	1:04:49	(also 1:04:28 * not legal)
Marathon	2:14:04	Mx World record



1 more row



Wikipedia

https://en.wikipedia.org/wiki/Brigid_Kosgei

[Brigid Kosgei - Wikipedia](#)

Quantity Queries with SE



female runners with a marathon under 2:20:00

**female runners with a
marathon under 2:20:00**

About 82.500 results (0,43 seconds)



Runner's World

<https://www.runnersworld.com> › races-places › whats-... ⋮

What's the Women's Equivalent of a Sub-2:00 Marathon?

23 Dec 2014 — Sandra Hunter has long been investigating gender differences in performance, Michael Joyner wrote the first **sub-2:00** paper way back in 1991, and ...



SweatElite

<https://www.sweatelite.co> › female-equivalent-of-the-t... ⋮

Female equivalent of the Two-Hour Marathon

Eliud Kipchoge was the first man to run **under** 2 hours in the **Marathon** despite not being ratified as a world record. What is the **female** equivalent?



BMW Berlin Marathon

<https://www.bmw-berlin-marathon.com> › detail › mar... ⋮

Kenya dominates the marathon scene for the men (sub 2: ...

19 May 2005 — Since 1953, men have run the **marathon under 2:20:00** a recorded total ... The German **female runners**, with 79 times. **under 2:55:00**, ...

Quantity Queries with LM



female runners with a marathon under 2:20:00

USER

Which women ran a marathon under 2:20?

ASSISTANT

As of the time of writing (July 2021), **only two women** have officially run a marathon under 2 hours and 20 minutes:

1. Brigid Kosgei (Kenya) - On October 13, 2019, at the Chicago Marathon, she completed the race in a time of **2:14:04**, breaking the previous women's world record **2:15:25** set by **Paula Radcliffe** in 2003.

2. Ruth Chepngetich (Kenya) - On April 7, 2019, at the Dubai Marathon, she finished in a time of **2:17:08**, marking the fastest time ever run by a woman on a record-eligible course.

[worldathletics.org](https://www.worldathletics.org)

Rank	Mar	Competitor	DOB	Nat	Sex	Event	Date
1	2:14:04	Brigid Kosgei	09 DEC 1991	KEN	F	Marathon	13 OCT 2019
2	2:14:04	Brigid Kosgei	09 DEC 1991	KEN	F	Chicago, U.S.A.	13 OCT 2019
3	2:14:04	Brigid Kosgei	09 DEC 1991	KEN	F	Chicago, U.S.A.	13 OCT 2019
4	2:14:04	Ruth Chepngetich	08 AUG 1994	KEN	F	Chicago, U.S.A.	09 OCT 2022
5	2:14:04	Annea Berntson	01 OCT 1961	USA	F	Marathon	04 DEC 2002
6	2:14:04	Paula Radcliffe	17 DEC 1973	GBR	F	London	13 APR 2003
7	2:14:04	Reynold Walker	09 DEC 1994	KEN	F	Tripoli	05 MAR 2023
8	2:14:04	Leahon Gedy	30 MAR 1988	ETH	F	Marathon	04 DEC 2022
9	2:14:04	Tatyana Lemos	12 DEC 1988	ETH	F	Tripoli	05 MAR 2023
10	2:15:01	Mary Keeney	18 JAN 1960	USA	F	London	23 APR 2003
11	2:15:09	Marye Alamu	12 OCT 1997	ETH	F	Chicago, U.S.A.	09 OCT 2022
12	2:15:16	Renee Spence	27 SEP 1983	USA	F	Marathon	04 DEC 2002
13	2:15:22	Alma Johanna	22 NOV 1976	ETH	F	Amsterdam	18 OCT 2002
14	2:15:23	Nancy Yegorov	02 AUG 1989	ETH	F	Hamburg	14 APR 2002
15	2:15:23	Angela Jarmache	09 DEC 1983	ETH	F	Chicago, U.S.A.	09 OCT 2022
16	2:15:29	Evans Chepkurui	27 DEC 1990	KEN	F	Marathon	04 DEC 2022
17	2:15:36	Tatyana Lemos	09 JAN 1960	ETH	F	Marathon	04 DEC 2022
18	2:15:41	Wendie Degrassat	28 OCT 1980	ETH	F	Dakar	25 JAN 2003
19	2:15:49	Leahon Gedy	12 DEC 1988	ETH	F	Tripoli	05 MAR 2023
20	2:15:56	Tatyana Lemos	05 OCT 1960	ETH	F	London	23 APR 2003
21	2:15:57	Angela Jarmache	09 DEC 1983	ETH	F	Amsterdam	18 OCT 2002
22	2:15:58	Angela Jarmache	09 DEC 1983	ETH	F	London	05 OCT 2002
23	2:15:58	Angela Jarmache	09 DEC 1983	ETH	F	Tripoli	05 MAR 2023
24	2:16:03	Tatyana Lemos	05 OCT 1960	ETH	F	London	23 APR 2003
25	2:16:04	Joan Chelimo Kelly	05 NOV 1980	KEN	F	Seoul	17 APR 2002
26	2:16:05	Genet Dibaba	08 FEB 1991	ETH	F	Amsterdam	18 OCT 2002
27	2:16:09	Rui Aza	18 JAN 1994	ETH	F	Donggong	22 OCT 2022
28	2:16:11	Geleyn Gebreselase	12 MAR 1983	ETH	F	Marathon	16 SEP 2002
29	2:16:11	Geleyn Gebreselase	12 MAR 1983	ETH	F	Asian Cup	18 JUL 2002
30	2:16:11	Fany Chelath	30 MAR 1995	ETH	F	Marathon	04 DEC 2022
31	2:16:12	Joan Chelimo Kelly	05 NOV 1980	KEN	F	Seoul	17 APR 2002
32	2:16:20	Joan Chelimo Kelly	05 NOV 1980	KEN	F	Asian Cup	18 JUL 2002
33	2:16:22	Mariam Al-Sayid	18 SEP 1989	ETH	F	Amsterdam	18 OCT 2002
34	2:16:29	Emily Stoller	12 OCT 1981	USA	F	Chicago, U.S.A.	09 OCT 2022
35	2:16:30	Renee Spence	09 SEP 1983	ETH	F	Marathon	04 DEC 2022
36	2:16:31	Vivian Jepkemoi Cheruiyot	11 SEP 1983	KEN	F	London	23 APR 2003
37	2:16:33	Annea Berntson	01 JAN 1962	ETH	F	Marathon	04 DEC 2022
38	2:16:38	Brigitte Dibaba	11 SEP 1993	ETH	F	Tripoli	05 MAR 2023
39	2:16:41	Maphane Braur	25 FEB 1995	ETH	F	Marathon	24 SEP 2023
40	2:16:47	Catherine Ndereba	21 JUL 1972	ETH	F	Chicago, U.S.A.	07 OCT 2002

KG Gap: Quantities of Entities



Thinh Ho Koninika Pal Daria Stepanova
[WWW'21, WWW'22]

Web tables, JSON files, data lakes ...

Motivation: Quantity (Filter) Queries

- runners with (most) marathon(s) under 2:10:00
- footballers who transferred for more than 50 Mio Euros
- hybrid cars with battery range above 80 km
- renewable power plants with annual 500 GWh or higher

supporting analysts, journalists, data scientists ...

Quantity:
(value, unit, context)

Qsearch / QuTE:

extract & index entity-quantity-context tuples from web text & tables

- neural learning w/ distant supervision +
- probabilistic graphical models +
- rule-based inference

query processing by

- matching **entity-type, quantity & context**
- scoring for ranked query result

Prototype System QuTE

Web tables, JSON files, data lakes ...

qsearch.mpi-inf.mpg.de/table/



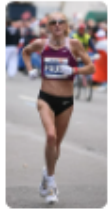
female runners with marathon under 2:20:00

Search

Parsed: female athlete - marathon - under 2.3 hours

(UPPER_BOUND , 2.3 , hours (duration_wd:Q2199864))

Paula Radcliffe



Pg-Title List of English records in athletics

↳ Outdoor

↳ Women

Event	Record	Athlete	Date	Meet	Place	Re
Marathon	2:15:25	Paula Radcliffe Evidence	13 April 2003	London Marathon	London , United Kingdom	

Score: 0.789 (-0.016)

[Show more](#)

Mary Jepkosgei Keitany



Pg-Title List of world records in masters athletics

↳ Women

↳ Marathon

Age group	Record	Athlete	Nationality	Birthdate	Age	Date
W 35 (progression)	2:17:01	Mary Jepkosgei Keitany Evidence		12 January 1982	35	23 April 2017

Score: 0.773 (+0.012)

[Show more](#)

Prototype System QuTE

Web tables, JSON files, data lakes ...

gsearch.mpi-inf.mpg.de/table/



female runners with marathon under 2:20:00

Search

Parsed: female athlete - marathon - under 2.3 hours

(UPPER_BOUND , 2.3 , hours (duration_wd:Q2199864))

Remalda Kergyte-Dauskurdiene



Pg-Title 2016 European Athletics Championships - Women's

half marathon

↳ Results

↳ Final

↳ Individual

Rank	Athlete	Nationality	Time	Notes
69	Remalda Kergyte - Dauskurdiene	Evidence	1:17:56	

Score: 0.764 (-0.001)

Salina Kosgei



Pg-Title French Riviera Marathon

↳ Results

	Athlete	Nationality	Time (h : m : s)	Athlete	Nationality	Time (h : m : s)
Year	Male			Female		
2013	Abdisa Sori		2:13:58	Salina Kosgei	Evidence	2:41:34

Score: 0.772 (-0.020)

Caption French Riviera Marathon results

Entity-Quantity Extraction from Tables with Text Evidence

multi-source
multi-modal

women marathon
under 2:20h



... Radcliff ... world record ... 2h15m25s
 ... Paula ... ran 26.2m in 2:15 hours ...
 ... Radcliffe ... worth 86.5 m GBP ...
 ... Kosgei ... marathon ... under 2:15:00
 ... Duffy .. triathlon ... under 2h total ...



<title> long distance runs

context cues

<h1> women

<th>

e-q col alignment



entity linking

Competition	Name	Loc	Pos	Time
	P. Radcliffe	London	1	2:15:25
WC 10K	Paula Radcliff	Sevilla	2	30:27.13
Chicago M	B. Kosgei		1	2:14:04
Half M 2016	B. Kosgei	Bogota 2640m	1	1:12:16
Ironman 18	Anne Haug	Hawaii	3	2:55:22
ITU Triathlon	Flora Duffy	Yokohama	1	1:56:18

... runs over full or half marathon, cross country, triathlon, 10000m ...

seed facts

metrics & units

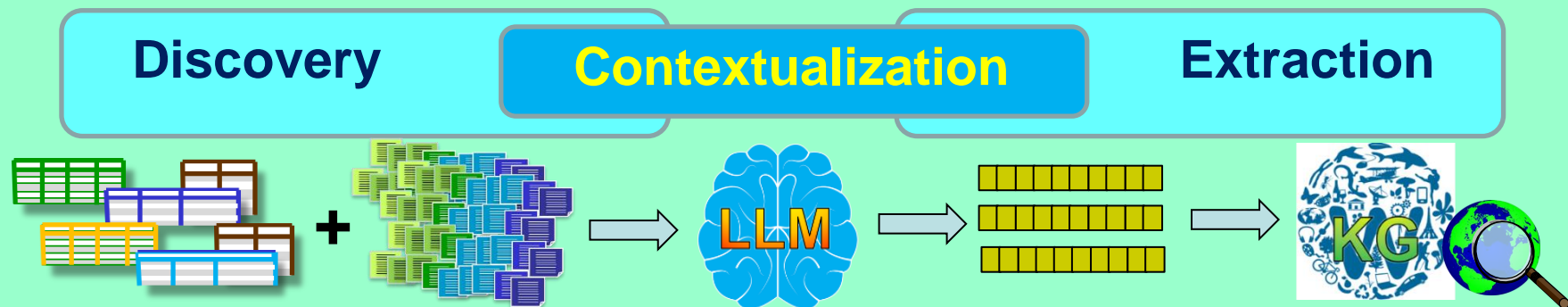
distribution inference

QuTE method: search text corpus for corroboration evidence

- for each (e,q) candidate
 - find e mention and approx. q in soft-matched context
- corroborate (e,q) by confidence score from tables+text

Counts and Quantities: **Bottom Line**

- First-cut method QuTE: **multi-source multi-modal**
- **Challenges & opportunities:**
 - good for search << good enough for KG
 - time-variant quantities (finance, climate, energy ...)
 - high recall for group-by queries (runners with most marathons ...)
- **Role of LLM unclear:**
 - not geared for entities with quantities ?
 - lack ability to aggregate and reason on distributions ?
- **Best bet: LM for contextualization** of tables & text



Outline

1. What are KGs, what are LMs

2. Blind Spots in KGs ...

- ✓ **Long Tail and Long Lists**
- ✓ **Counts and Quantities**

Take-Home Message



Thank You !

KG: highest **correctness**, **completeness**, **tangibility**

Major KG concern: **coverage**

Gaps include: **long tail** ... **quantities** ...

... **meta-knowledge** ... **commonsense** ...

LLMs for **extending KGs**?

- **language** + **prior world knowledge** 
- **not** good at **high recall @ high precision** 

