

# Towards Semantic Exploration of Tables in Scientific Documents

Varish Mulwad<sup>1</sup>, **Vijay S. Kumar**<sup>2</sup>, Jenny Weisenberg Williams<sup>2</sup>,  
Tim Finin<sup>3</sup>, Sharad Dixit<sup>2</sup>, Anupam Joshi<sup>3</sup>

1: GE Research, Bengaluru, KA, India

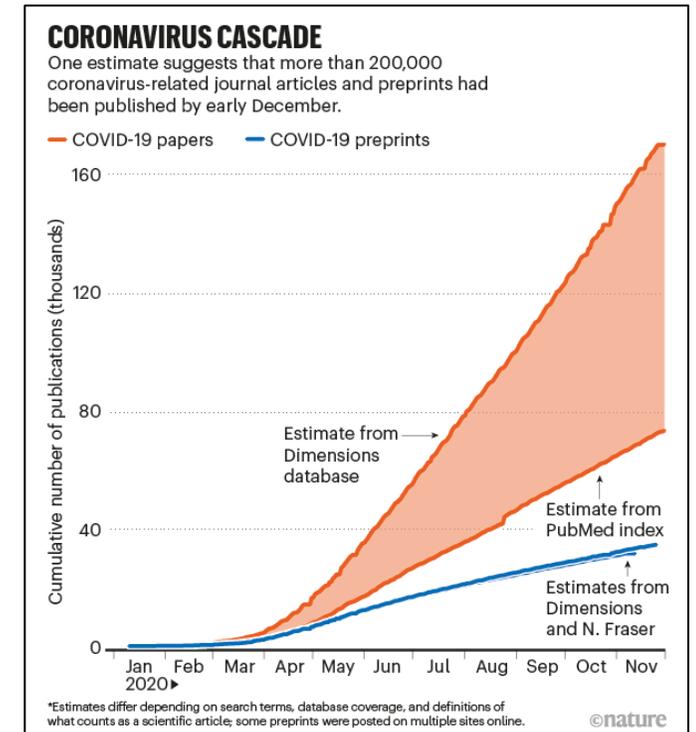
2: GE Research, Niskayuna, NY, USA

3: University of Maryland, Baltimore County, Baltimore, MD, USA

1<sup>st</sup> Workshop on Semantic Technologies for Scientific, Technical and Legal Data,  
ESWC, May 28, 2023

# Introduction

- An explosion in scientific publication rates
  - as exemplified by the COVID-19 *infodemic*
  - large amounts of scientific documents freely accessible online
- Contributing factors:
  - Open science, open-access initiatives
  - preprint servers, generative AI
- Problems:
  - ‘burden of knowledge’: rate of scientific research progress not keeping up with publication rates
  - ‘fake science’: rise in prevalence of misinformation



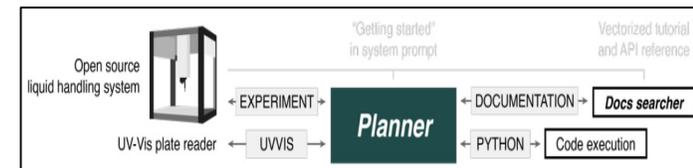
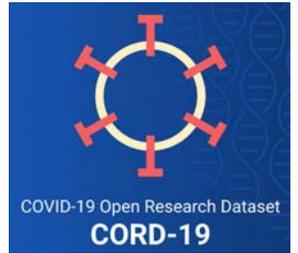
Else, H., 2020. How a Torrent of COVID Science changed Research Publishing—in seven charts. *Nature*, V588 (553).

Need more machine-driven, human-interpretable approaches to scientific knowledge discovery

# The rise of AI-assisted Science



- Open datasets, e.g., [CORD-19](#) (articles on COVID-19)
  - Led to tools for search, Q&A, recommendation, summarization, and claim verification over scientific documents.
- Foundation models, e.g., GPT-n, Galactica
  - Led to tools for literature review, explaining scientific concepts, scientific code generation, intelligent reading interfaces.
  - Allen AI's [Semantic Reader](#) product aims to integrate tools to augment the scientific reading & discovery experience.
- From scientific discovery to scientific understanding
  - Early works show potential emergent autonomous scientific research capabilities of LLMs ([Boiko et al. 2023](#))



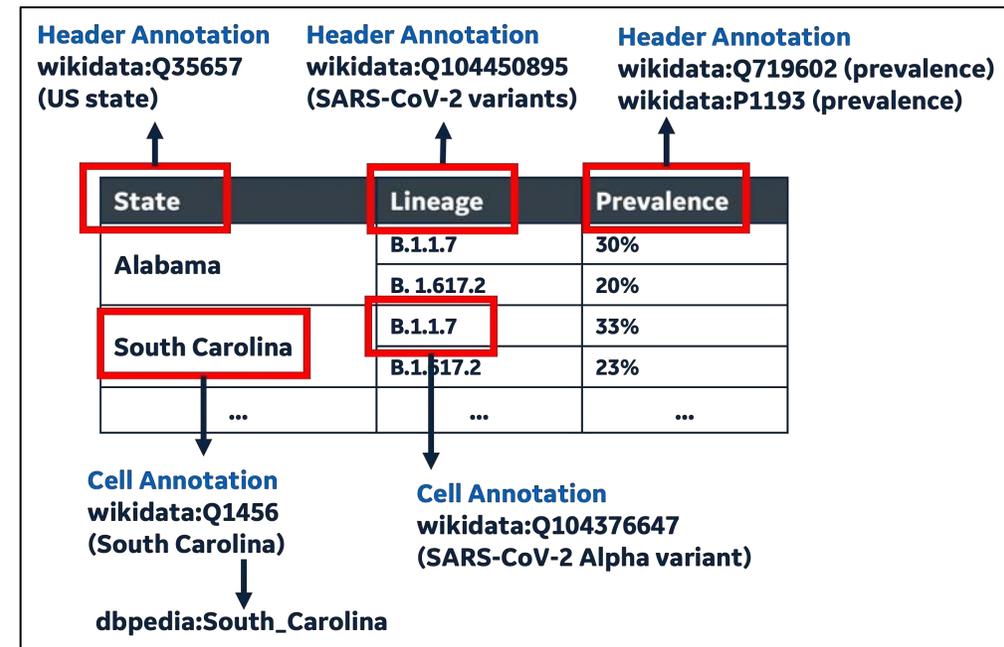
Most research to process, analyze, explore scientific documents focuses only on text content. Significant information also present within structured artifacts, e.g., tables, charts.

# Our research Focus: (Scientific) Tables



- Information in tables is critical to emerging knowledge-driven applications:
  - e.g., intelligence analysis and production, *genomic surveillance*
  - technical experts (scientists, analysts) look to discover vital knowledge nuggets not only from text but also from tables
- Automated mechanisms to infer the semantic meaning of scientific tables
  - relevance to a knowledge discovery scenario
  - generate tabular responses on-the-fly by possibly fusing information from multiple tables

Understanding information contained in **tables** from scientific & technical documents within specialized domains (e.g., biomedical)



*“Treatment efficacies against the top prevalent COVID-19 variants in each US state”*

Scientific documents → Collection of content-rich tables + descriptive context  
Apply semantic technologies to model information in scientific tables to enable discovery

# Tables have been extensively studied

Most recent acceptance rates of the top 3 semantic web conferences as a table

✓ Searching for: acceptance rates of ISWC ESWC SEMANTiCS 2021

✓ Generating answers for you...

I found some information about the most recent acceptance rates of the top 3 semantic web conferences (ISWC, ESWC and SEMANTiCS) in 2021. Here is a table that summarizes the information I found:

Conference	Year	Acceptance Rate
ISWC	2021	33.6% (research track)
ESWC	2021	25%
SEMANTiCS	2021	N/A

Note: I could not find the acceptance rate for SEMANTiCS 2021, so I marked it as N/A (not available).

Learn more: [1. 2021.eswc-conferences.org](#) [2. dl.acm.org](#) [3. openresearch.org](#) 5 of 20

Extraction & interpretation of web tables for search, retrieval, Q&A

Some capability already integrated into modern-day search engines

Dataset retrieval over Open data with view discovery, augmentation

Dataset search engines help prepare datasets for data science applications

Auctus covid-19 vaccination rate

Advanced Search: Any Date Any Location Related File Source Data Type

COVID-19 Outcomes by Vaccination Status (229.3 kb)  
data.cityofchicago.org

Weekly rates of COVID-19 cases, hospitalizations, and deaths among people living in Chicago by va...  
[Show more...](#)

Abc Outcome Week End Abc Age Group  
# Unvaccinated Rate # Vaccinated Rate  
[Show 16 more columns...](#)

Categorical Numerical Temporal

Download View Details Search Related

Augment Options

COVID-19 Outcomes by Vaccination Status

Augmentation (JOIN)

Join on:  
 Age group and Abc Age Group

Available columns:  
Select which columns should be added to the final merged dataset.

Abc Outcome Week End # Unvaccinated Rate # Vaccinated Rate  
# Crude Vaccinated Ratio # Crude Boosted Ratio # Age-Adjusted Un  
# Age-Adjusted Vaccinated Rate # Age-Adjusted Boosted Rate # Age  
# Age-Adjusted Boosted Ratio # Population Unvaccinated # Popula  
# Outcome Unvaccinated # Outcome Vaccinated # Outcome Boosted

# Tables have been extensively studied



- Significant advances in pre-trained / table representation learning [models](#) for **well-structured** tabular data and a variety of downstream tasks.
- Some specifically address tables in scientific/technical documents:

Dataset	Downstream task
<a href="#">PubTables-1M</a>	Table detection, Table structure recognition
<a href="#">ChemTables</a>	Table classification
<a href="#">ArxivPapers</a>	Table extraction
<a href="#">SciGen</a>	Reasoning-aware Table-to-text generation
<a href="#">TAT-QA</a>	Question-answering over Tables and text
<a href="#">S2abEL</a>	Entity Linking for scientific Tables

Scientific tables bring additional challenges and opportunities

# Tables in Scientific Documents



- Optimized for human visual consumption
  - minimize information overload
  - various information compaction practices to ensure fit under space constraints
- Machine-driven understanding and discovery of scientific tables is challenging

- Structural Heterogeneity
- Dense, often implicit, semantics
- Diffuse context
- Domain-specific entities
- Questionable reliability

*shares some similarities with web tables*

*shares similarities with open dataset search*

*not adequately addressed by either*

Table 1  
Sensitivity and specificity of the Elecsys® Anti-SARS-CoV-2 and LIAISON® SARS-CoV-2 S1/S2 IgG tests.

Test and result	COVID-19 NAAT test result		Sensitivity (%)	Specificity (%)	PPV (%) (COVID-19 prevalence 1/5/10%)	NPV (%) (COVID-19 prevalence 1/5/10%)
	Positive (n = 40)	Negative (n = 161)				
Elecsys® Anti-SARS-CoV-2						
Positive	37	2 <sup>b</sup>	92.5 (CI: 79.6–98.4)	98.8 (CI: 95.6–99.9)	42.9/79.7/89.2	99.9/99.6/99.2
Negative	3 <sup>a</sup>	159				
LIAISON® SARS-CoV-2 S1/S2 IgG						
Positive	35	4 <sup>b</sup>	87.5 (CI: 73.2–95.8)	97.5 (CI: 93.8–99.3)	26.2/65.0/79.7	99.9/99.3/98.6
Negative	5	157				

Less text, more numbers ... sub-columns ... merged cells

Table 2  
Developed serology tests for SARS-CoV-2 detection by different companies and researchers.

Developer	Platform	Target antigen	Target antibody	Other features	References
Abbott Laboratories	CMIA	Nucleocapsid	IgG	Return 100–200 test results in 1 h, specificity 99.6%, and sensitivity of 100%	<a href="#">Abbott Laboratories (2020b)</a>
DiaSorin	CMIA	Spike	IgG	Fully automated, quantitative, 97.4% sensitivity, 98.5 specificity	<a href="#">DiaSorin (2020)</a>
Pharmact AG	Lateral flow assay	–	IgG and IgM	POC, results in 20 min, can determine the phase of the disease, 99.8% agreement with PCR for non-affected cases	<a href="#">Pharmact (2020)</a>
Hangzhou Biotech	Lateral flow assay	Spike	IgG and IgM	100% specificity for IgM and IgG, 100% sensitivity	<a href="#">(Hangzhou Biotech Biotech)</a>

Similar to web tables ... with domain-specific entities

# Approach and Contributions\*\*

- Collect tables from scientific articles
- Represent scientific tables as semantically-annotated linked data through a two-stage methodology:
  - Automated rule-based structural characterization; syntactic parsing through “specialists”
  - Semantic table interpretation with joint inference
- Enable discovery of tabular information from knowledge graph
  - Search queries under rich contextual constraints (including information reliability)
  - On-the-fly table generation by fusing information from compatible tables

## Dataset Collection Pipeline

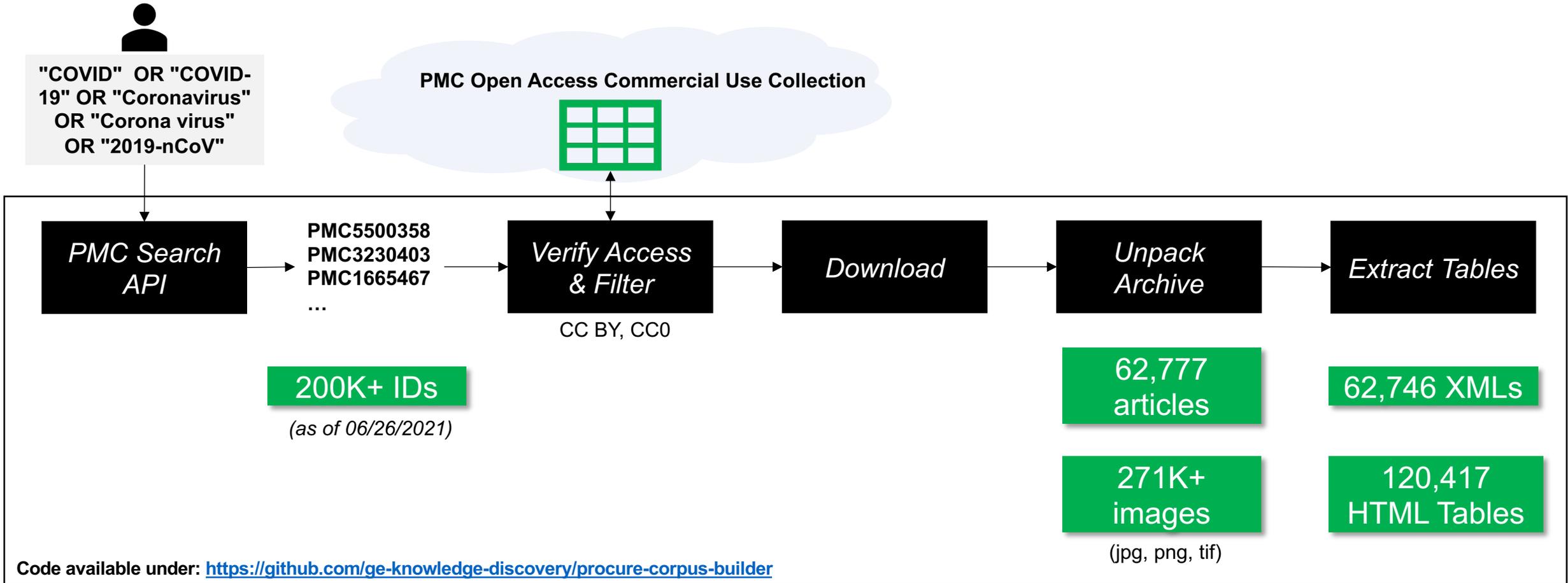
- harvest tables and provenance metadata from [PubMed Central](#) open-access subset

## Preliminary Prototype Systems

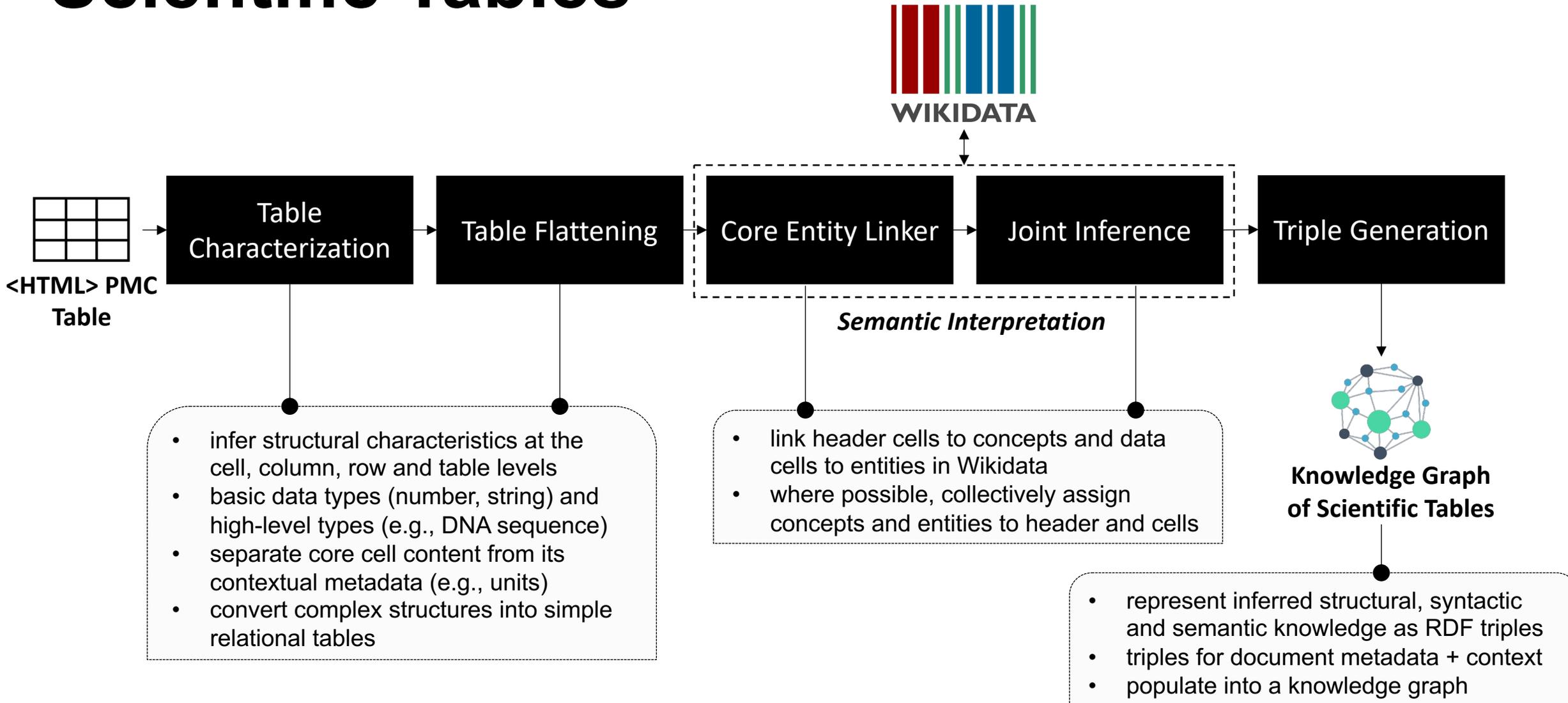
1. Automated pipeline to construct a knowledge graph of scientific tables
  - Ontology to model tabular data and context
  - Core entity linker for scientific tables
  - Joint inference based on KG embeddings
2. Scientific table discovery system
  - UI to specify table-based semantic search requests and explore responses
  - Discovery engine to produce ranked lists of matching tables (including on-the-fly tables) with explainability

\*\* <https://github.com/ge-knowledge-discovery> (currently in process of being open-sourced)  
Looking for collaborations to build on/extend our work, co-develop new datasets & applications

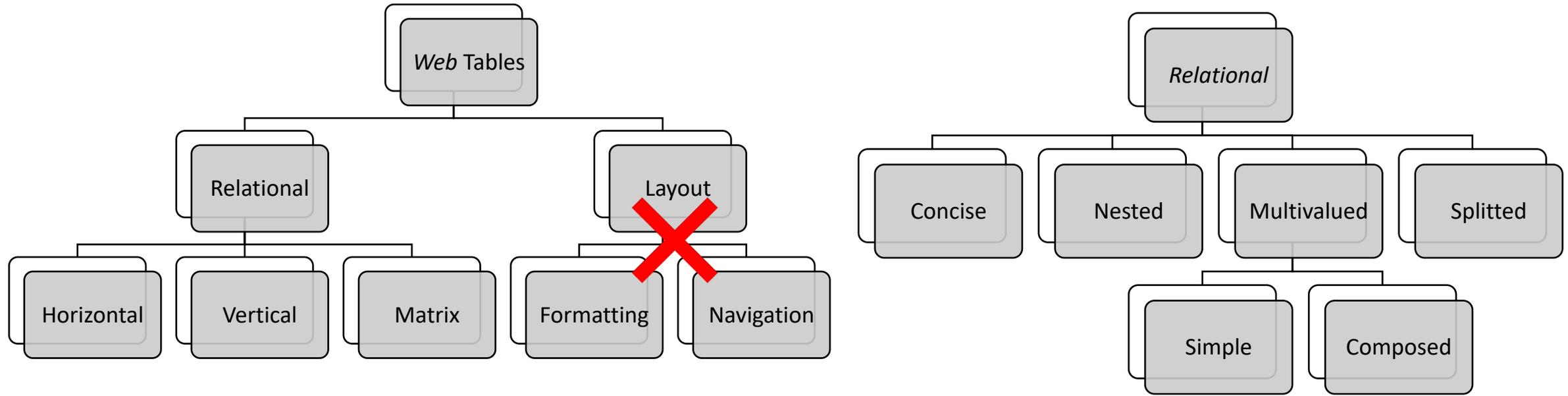
# Dataset Collection Pipeline



# Knowledge Graph Construction from Scientific Tables



# Taxonomy to Characterize Tables



Lautert, L. R., Scheidt, M. M., & Dorneles, C. F. (2013). Web table taxonomy and formalization. *ACM SIGMOD Record*, 42(3), 28-33.

We leverage and extend multiple existing taxonomies and classification techniques in order to adequately characterize scientific tables

# Table Characterization at different granularities



**Header Row**

Haplotype	Marker					Species
	<i>Xij5</i>	<i>Xid3</i>	<i>Xij6</i>	<i>Xid4</i>	<i>Xij7</i>	
H1	594 <sup>a</sup>	718	797	176	961	<i>T. wartu</i> (30, 30.30%) <sup>c</sup> <i>T. turgidum</i> (37, 38.95%) <i>T. aestivum</i> (117, 54.42%)
H2	594	808	797	176	961	<i>T. wartu</i> (2, 2.02%)
H3	594	655	797	176	961	<i>T. wartu</i> (1, 1.01%)

**Horizontal Table**  
(Main Classification)

**Header Column**

	Viremia category				p value <sup>*</sup>
	PSV n = 716 (70.5%)	LLV-I n = 46 (4.5%)	LLV-II n = 52 (5.1%)	HLV n = 201 (20%)	
Gender [n (%)]					0.011
Male	468 (65%)	39 (85%)	36 (69%)	148 (74%)	
Female	248 (35%)	7 (15%)	16 (31%)	53 (23%)	
Age at inclusion [median years (IQR)]	39 (33–46)	43 (35–50)	39 (34–49)	39 (33–47)	0.28
Ethnicity [n (%)]					0.0027
Caucasian	307 (43%)	25 (54%)	29 (56%)	96 (48%)	
African	172 (24%)	12 (26%)	12 (23%)	68 (34%)	
Asian	81 (11%)	2 (4.3%)	3 (5.8%)	9 (4.5%)	
Other/unknown	156 (22%)	7 (15%)	8 (15%)	28 (14%)	

**Vertical Table**

**Matrix Table**

From:	To:									
	Employment <sup>a</sup>	Sick leave	Vocational rehabilitation	Medical rehabilitation	Time limited disability pension	Disability pension	Emigrated	Dead (8)	Alive and at risk <sup>b</sup>	
Employment <sup>a</sup> (1)	0	1 657 895	45 865	48 607	3 736	20 314	28 421	10 860	729 891	
Sick leave (2)	1 595 592	133 909	53 398	48 325	559	3 309	454	726	105 714	
Vocational rehabilitation (3)	101 605	983	2 009	962	3 816	9 351	175	85	288	
Medical rehabilitation (4)	82 070	130	0	14 381	1 320	5 164	93	81	10 566	
Time limited disability benefits (5)	1 153	68	148	84	0	7 930	7	23	191	
Disability benefits (6)	7 729	307	75	26	37	0	1 404	819	45 265	
Emigrated (7)	26 487	317	32	51	0	46	53	13	5	
Total	1 814 636	1 793 609	101 527	112 436	9 468	46 114	30 607	12 607	891 920	

**Matrix Table**

Developer	Platform	Target antigen	Target antibody	Other features	References
-----------	----------	----------------	-----------------	----------------	------------

**Simple Header**

Haplotype	Marker					Species
	<i>Xij5</i>	<i>Xid3</i>	<i>Xij6</i>	<i>Xid4</i>	<i>Xij7</i>	

**Concise Header**

Variables	1&2 <sup>a</sup>	3	4	5	6	7	9	Missing	N
-----------	------------------	---	---	---	---	---	---	---------	---

**Multilevel Header**

IQ test taken	No IQ test taken			Combined score				
	Stamine Score	N	Percent	Score	N	Percent	N	Percent
1	15 709	1,9%	C (assumed below average)	3 769	14,9%	Assumed below average <sup>a</sup>	57 092	6,6%
2	37 614	4,5%	B (assumed average)	20 318	80,6%	Assumed average <sup>b</sup>	691 580	80,1%
3	75 441	9,0%	A (assumed above average)	1 124	4,5%	Assumed above average <sup>c</sup>	114 697	13,3%

**Splitted (Header)**

Sample type (n)	Method	No. of positive sample	No. of positive sample by any method	Positivity rate <sup>a</sup> [% (95% CI)]	No. of positive patients by any methods in any sample types	Positivity rate <sup>b</sup> [% (95% CI)]
OPS (n=68)	qRT-PCR A	12	24	50.0 (29.6-70.3)	56 <sup>c</sup>	21.4 (12.0-34.8)
	qRT-PCR B	18		75.0 (52.9-89.4)		32.1 (20.6-46.1)
	RT-RAA	20		83.3 (61.8-94.5)		35.7 (23.7-49.7)

**Concise Body**

Characterize tables with **18 different labels** at cell, row, column, and table levels (full list in the paper)

IGDS9-SF (-)

Basic type: string

60,418.0

Basic type: number

-0.051 ± 0.024<sup>\*</sup>

Basic type: number with tolerance

1.28 (0.30–5.43)

Basic type: number with range

...

# Rule-based approach for Table Characterization



References	Selection	Comparison
Representativeness (Max:★)	Sample size (Max:★)	Non-respondents (Max:★)
	Ascertainment of the exposure (Max:★★)	Comparable groups/Controlled for confounding (Max:★★)
		Assessment of outcome (Max:★★)

```

<thead>
  <tr>
    <th colspan="1" rowspan="1"></th>
    <th align="center" colspan="4" rowspan="1" style="border-bottom: thin solid #000000;" valign="top">
      <bold>Selection</bold>
    </th>
    <th align="center" colspan="1" rowspan="1" style="border-bottom: thin solid #000000;" valign="top">
      <bold>Comparison</bold>
    </th>
    <th align="center" colspan="2" rowspan="1" style="border-bottom: thin solid #000000;" valign="top">
      <bold>Outcome</bold>
    </th>
  </tr>
  <tr>
    <th align="left" colspan="1" rowspan="1" valign="top">
      <bold>References</bold>
    </th>
    <th align="center" colspan="1" rowspan="1" valign="top">
      <bold>Representativeness (Max:\u22c6)</bold>
    </th>
    <th align="center" colspan="1" rowspan="1" valign="top">
      <bold>Sample size (Max:\u22c6)</bold>
    </th>
    <th align="center" colspan="1" rowspan="1" valign="top">
      <bold>Non-respondents (Max:\u22c6)</bold>
    </th>
    <th align="center" colspan="1" rowspan="1" valign="top">
      <bold>Ascertainment of the exposure (Max:\u22c6\u22c6)</bold>
    </th>
    <th align="center" colspan="1" rowspan="1" valign="top">
      <bold>Comparable outcome groups/Controlled for confounding factors (Max:\u22c6\u22c6)</bold>
    </th>
    <th align="center" colspan="1" rowspan="1" valign="top">
      <bold>Assessment of outcome (Max:\u22c6\u22c6)</bold>
    </th>
    <th align="center" colspan="1" rowspan="1" valign="top">
      <bold>Statistical test (Max:\u22c6)</bold>
    </th>
  </tr>
</thead>
  
```

**Concise Header**  
(colspan/rowspan > 1)

**Multi-level Header**  
(multiple <tr> in <thead>)

```

{
  "BODY_COLS": [
    {
      "BOLD": true,
      "CELLS_DATATYPE": "string",
      "NUM_CELLS": 11
    },
    {
      "CELLS_DATATYPE": "number",
      "NUM_CELLS": 11
    },
    ...
  ],
  "BODY_CONCISE": true,
  "BODY_ROWS": [
    {
      "NUM_CELLS": 7
    },
    ...
  ],
  "HAS_HEADER_COLUMN": true,
  "HAS_HEADER_ROW": true,
  "HAS_MULTIVALUED_CELLS": true,
  "HEADER_ROWS": [
    {
      "CELLS_DATATYPE": "string",
      "HAS_COLSPAN": true,
      "HAS_ROWSPAN": true,
      "NUM_CELLS": 3
    },
    {
      "CELLS_DATATYPE": "string",
      "HAS_EMPTY": true,
      "NUM_CELLS": 6
    }
  ],
  "HEADER_ROWS_CONCISE": true,
  "HEADER_ROWS_MULTILEVEL": true,
  "MAIN_CLASSIFICATION": "horizontal",
  "NUM_BODY_ROWS": 11,
  "NUM_COLS": 7,
  "NUM_HEADER_ROWS": 2
}
  
```

# Table Characterization – Illustrative Example

**Table 1**

Main TiO<sub>2</sub> nanoparticles (NP) physicochemical properties, as described in Joint Research Centre (JRC) Report [42].

Sample	Crystalline Phase	Primary Particle Size (nm)	Primary Density (g/cm <sup>3</sup> )	Specific Surface Area (m <sup>2</sup> /g)
NM-100	Anatase	100.0 ± 50.0	3.84	9.23
NM-101	Anatase	6.5 ± 1.5	3.84	316.07



```
"NUM_BODY_ROWS": 2,
```

```
  "BODY_CONCISE": false,
```

```
  "BODY_COLS": [
```

```
    "CELLS_DATATYPE": "string",
```

```
    "CELLS_DATATYPE": "number with tolerance",
```

```
    "CELLS_DATATYPE": "number",
```

```
  "MAIN_CLASSIFICATION": "horizontal",
```

# Table Characterization Stats and Results

Characterization	System Count	# of Tables manually labelled	Precision	Recall
Tables with Header Rows	113,582	110	1.00	0.94
Tables with Header Columns	48,733	103	1.00	0.55
Tables with Concise Header Rows	36,182	34	0.84	0.94
Tables with Multi-level Header Rows	32,169	33	1.00	0.97
Tables with ONLY Numeric Data Cells	12,969	29	1.00	0.83
Tables with Concise Body	40,158	39	0.97	0.67
Horizontal Tables	21,863	38	0.95	0.50
Vertical Tables	7205	16	0.91	0.62

*PMC8185411*

Table 2.  
Symptoms displayed

Impact of COVID-19: Likert Scale Data	Median	IQR
Anxiety symptoms	3.0	6.0
Mood symptoms	2.0	5.0
Functioning: social	3.0	6.0
Functioning: occupational	0.0	5.0
Quality of life	3.0	5.0
Anxiety symptoms		
BAI	4.0	10.3
HAM-A	4.0	9.0

BAI, Beck Anxiety Inventory; HAM-A, Hamilton Anxiety Rating Scale; IQR, Interquartile range.  
Anxiety symptoms are correlated with BAI and HAM-A scales ( $\rho = 0.56, p < 0.001$  for both).  
Data were not normally distributed.

↓

```
{
  "BODY_CONCISE": true,
  "CELLS_DATATYPE": "number",
  "HAS_HEADER_COLUMN": true,
  "HAS_HEADER_ROW": true,
  "MAIN_CLASSIFICATION": "horizontal"
}
```

*Manual User Annotations*

*Manually annotated tables from randomly selected articles. Computed Precision & Recall at the label level  
(Precision: # of correct predictions / total predictions; Recall: # of correct predictions / expected predictions)*

**Precision is generally high. Recall is low in some cases.  
Rule coverage not exhaustive enough. Additional supervised algorithms may help.**

# Specialists to Detect Commonly Occurring Data



- Scientific table cells encode commonly occurring data – e.g., references, DNA/RNA sequences, clinical trial IDs, etc.
- *Specialists* assess commonly encoded data types to avoid linking such cells → overall better table characterization and semantics
- Specialists may be regex/pattern-based, dictionary-based, or ML-based depending on the data type
- Each applicable specialist assesses the cell values independently

**Semantic Type REFERENCE**

Study	Location	Total cases
CDC (12)	United States	149,760
Livingston and Bucher (18)	Italy	22,512
Tagarro et al. (19)	Spain (Madrid)	4,695

Subset of PMC7347905 Table 1

HTML `<xref>` tag containing attribute `ref-type="bibl"`

**Semantic Type DNA/RNA SEQUENCE**

<sup>a</sup> AA	Codon	<sup>b</sup> RSCU	AA	Codon	RSCU
Ala	GCA	1.25	Leu	CUA	0.22
	GCC	0.59		CUC	0.20
	GCG	0.02		CUG	0.64
	GCU	2.12		CUU	1.12

Subset of PMC3087699 Table 2

Regex based: 3+ characters from the set {G, A, T, C, U}

**Semantic Type CLINICAL TRIAL ID**

Virus	Location	Phase	Year	Identifier
SARS-CoV	United States	I	2004	NCT00099463
SARS-CoV	United States	I	2007	NCT00533741
SARS-CoV	United States	I	2011	NCT01376765
MERS	United Kingdom	I	2018	NCT03399578
MERS	Germany	I	2018	NCT03615911

Subset of PMC7239068 Table 1

Regex based: NCT followed by 8 digits

**Basic Type QUANTITY**

Variable studied	Median	Standard deviation
IL-6 (normal range 0.0–15.5 pg/ml) <sup>†</sup>	65 pg/ml	446.6 pg/ml
D-dimer (normal range 0.20–0.28 µg/ml)	2.1 µg/ml	4.7 µg/ml
Ferritin (normal range 20–450 ng/ml)	446 ng/ml	1822 ng/ml
CRP (normal range 0.0–9.0 mg/l)	90 mg/l	132.9 mg/l

subset of PMC7493720 Table 1

Identified using Pint python package <https://pint.readthedocs.io/>

Large amount of literals in scientific tables.  
Important to detect and learn to not link literals.

# Complex tables

Flatten by importing to Pandas dataframe

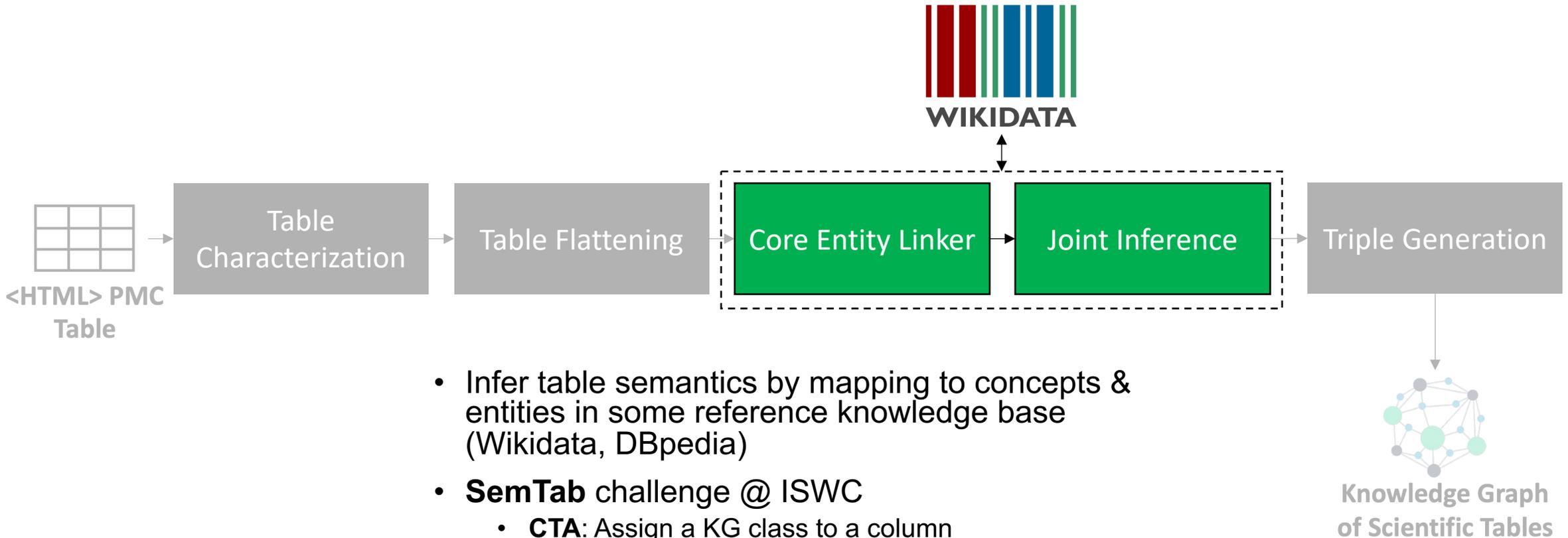


S/N	SARS-CoV-2 region	Mutation observed	Occurrence
1	Spike protein; S1 domain	L18F	2
2		A222V	4
3		D614G	898
4	Spike protein; S2 domain	E780Q	2
5	Nucleocapsid phosphoprotein	P13L	2
6		R203K	602
7		G204R	602
8	ORF3 protein	Q57H	60
9		G251V	4
10	ORF8 protein	L84S	16
11	ORF1ab polyprotein; nsp2	T265I	26
12		D448del <sup>a</sup>	6
13		I739V	3
14		P765S	3
15	ORF1ab polyprotein; nsp5	G3278S	73
16	ORF1ab polyprotein; nsp6	L3606F	45
17	ORF1ab polyprotein; RdRp	A4489V	3
18		P4715L	898

```
[ S/N SARS-CoV-2 region Mutation observed Occurrence
0 1 Spike protein; S1 domain L18F 2
1 2 Spike protein; S1 domain A222V 4
2 3 Spike protein; S1 domain D614G 898
3 4 Spike protein; S2 domain E780Q 2
4 5 Nucleocapsid phosphoprotein P13L 2
5 6 Nucleocapsid phosphoprotein R203K 602
6 7 Nucleocapsid phosphoprotein G204R 602
7 8 ORF3 protein Q57H 60
8 9 ORF3 protein G251V 4
9 10 ORF8 protein L84S 16
10 11 ORF1ab polyprotein; nsp2 T265I 26
11 12 ORF1ab polyprotein; nsp2 D448del 6
12 13 ORF1ab polyprotein; nsp2 I739V 3
13 14 ORF1ab polyprotein; nsp2 P765S 3
14 15 ORF1ab polyprotein; nsp5 G3278S 73
15 16 ORF1ab polyprotein; nsp6 L3606F 45
16 17 ORF1ab polyprotein; RdRp A4489V 3
17 18 ORF1ab polyprotein; RdRp P4715L 898 ]
```

After flattening, this complex table is now correctly classified as a horizontal table

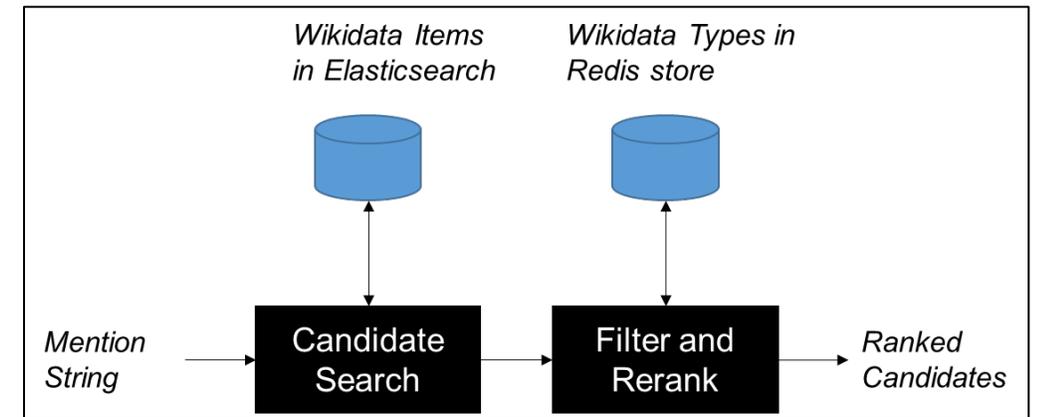
# Semantic Table Interpretation



- Infer table semantics by mapping to concepts & entities in some reference knowledge base (Wikidata, DBpedia)
- **SemTab** challenge @ ISWC
  - **CTA**: Assign a KG class to a column
  - **CEA**: Match a cell string to a Wikidata entity
  - **CPA**: Assign a property to relationship between two columns
  - Recently included BioTables, BiodivTab datasets

# Core Entity Linker

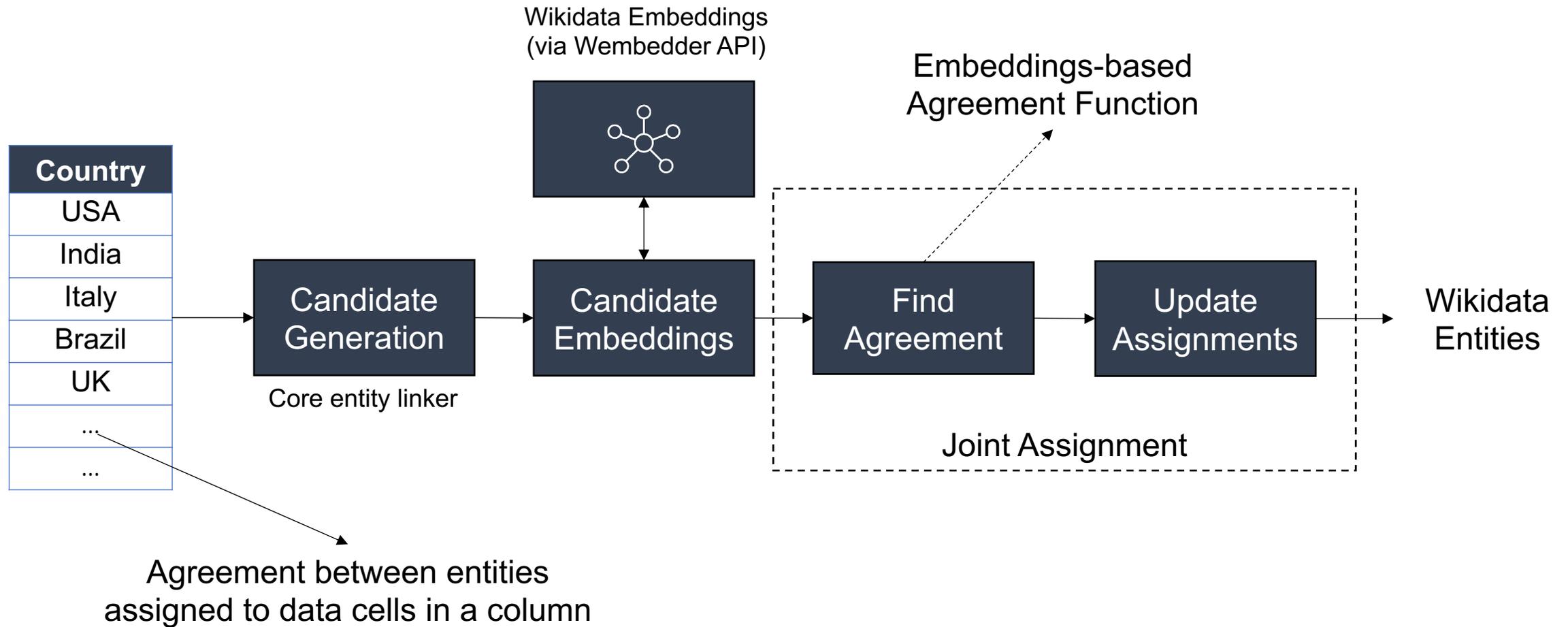
- We developed a practical, scalable entity linker to keep up with rate of publication of scientific tables
- Given a cell string, we retrieve top  $k$  matching Wikidata items, performing type analysis and filtering to return top-ranked candidate.
- Currently focused on CTA, CEA only
- Adapted to scientific tables in the biomedical domain



Mulwad, V., Finin, T., Kumar, V. S., Williams, J. W., Dixit, S., Joshi, A.  
**A Practical Entity Linking System for Tables in Scientific Literature.**  
In 3rd Workshop on **Scientific Document Understanding at AAI-2023.**

# Embeddings-based Joint Inference

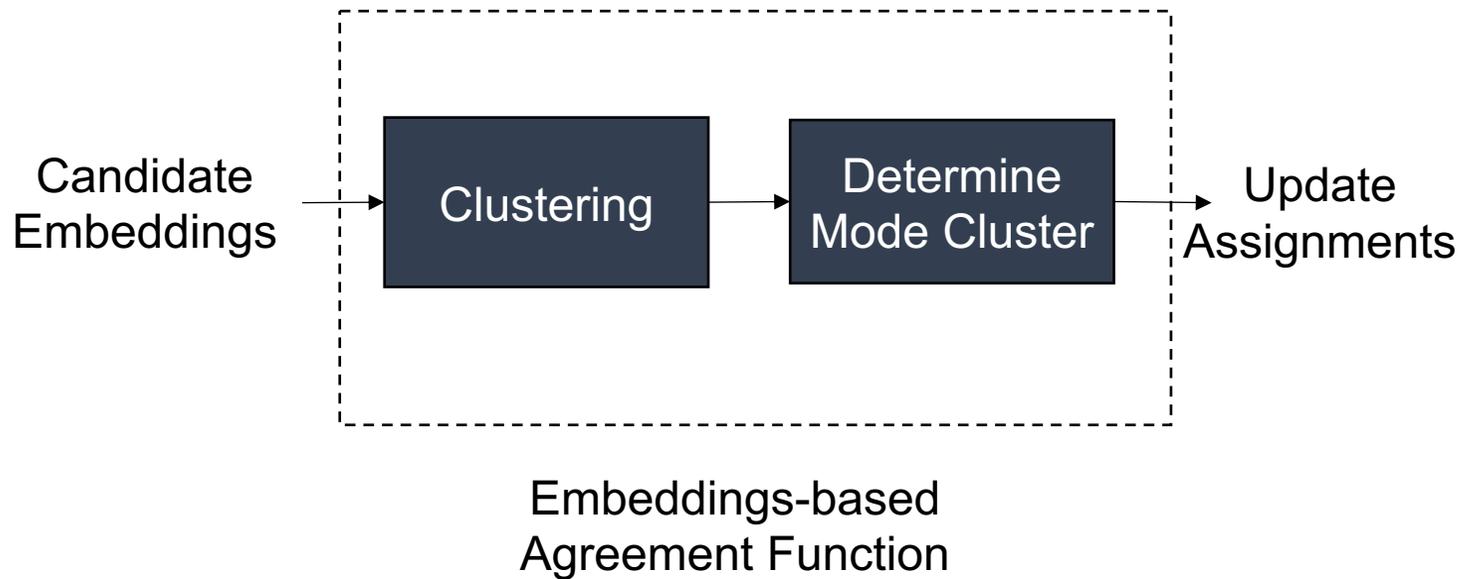
overview



Inferring table semantics is improved via joint inference using embeddings of Wikidata items (i.e., embeddings-driven agreement function to compute compatibility between entities)

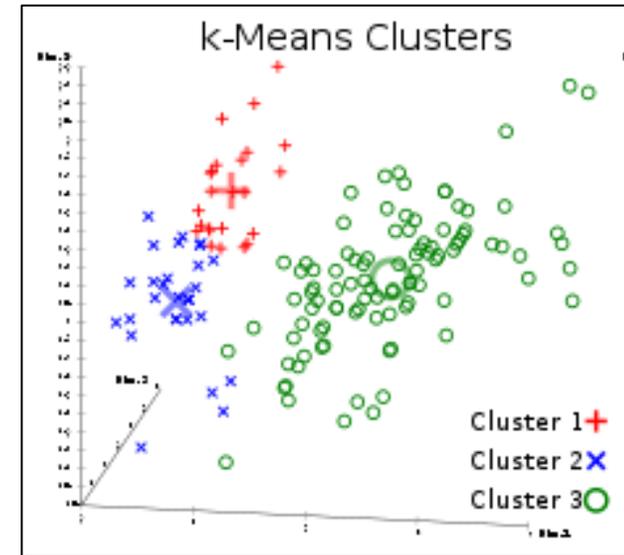
# Embeddings-based Agreement

*a clustering approach*



## K-Means Clustering

Separates samples into N clusters, minimizing the within-cluster sum-of-squares (minimize sum of distances between each point and its cluster centroid)

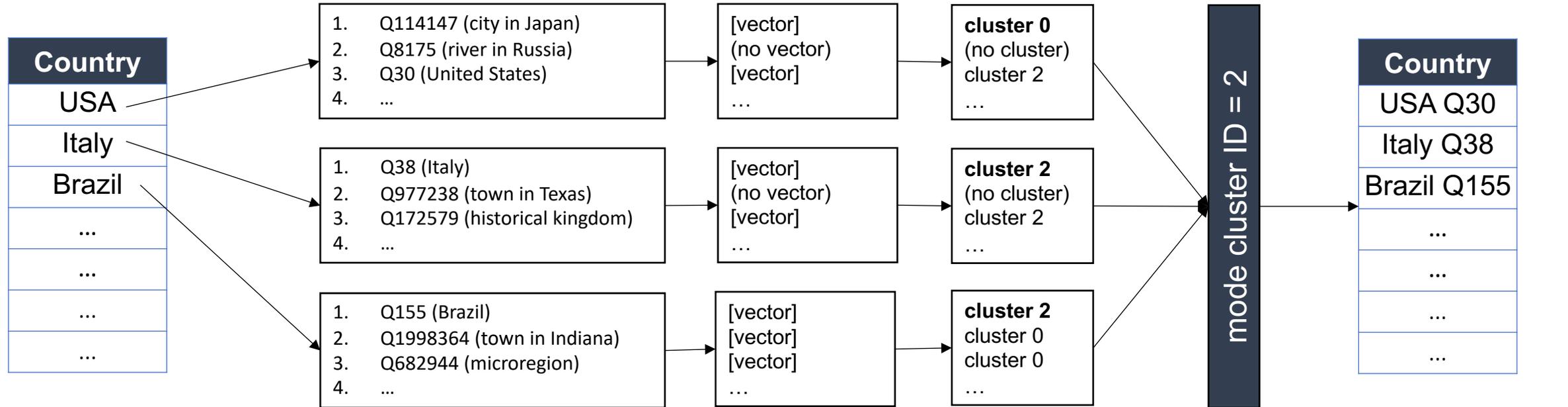


[https://en.wikipedia.org/wiki/K-means\\_clustering](https://en.wikipedia.org/wiki/K-means_clustering)

**Required Input: NUM\_CLUSTERS**  
(target number of clusters)

# Embeddings-based Joint Inference

perform clustering on embedding vectors; example



**Ranked List of Candidate Entities**  
(using base entity linker)

**Embedding vectors**  
(via Wembedder API)

**Cluster IDs**  
(K-Means)

**“Mode” Cluster ID**  
(most frequently occurring as first-ranked vector)

**Candidate Selected**  
(highest-ranked entity with mode cluster id)

# Annotations for Evaluating Table Semantics



- Manually annotated 47 tables drawn from randomly selected 45 PMC articles
- Each table cell was mapped to:
  - Wikidata Item
  - Literal (string/numeric data not representing an entity)
  - Reference (header cells for cols. with references)
  - Clinical Trial IDs (header cells for cols. with clinical trial IDs )
  - Index (header cells representing index cols. in a table)
  - NA (entity, but doesn't exist in reference KG)
- **3600** table cells were annotated

Annotation Type	Count
<b>Wikidata</b>	910
<b>Literal</b>	2548
<b>NA</b>	118
<b>Clinical Trial ID</b>	4
<b>Reference</b>	14
<b>Index</b>	6

Distribution of manual annotations amongst different types

Almost 70% of cells across manually annotated tables are literals

# Semantic Annotation: Overall Evaluation



Anno. Type	# of cells	Pr.	Re.	F-Score
<b>Literal</b>	2548	0.98	0.81	0.89
<b>Wikidata</b>	910	0.33	0.50	0.40
<b>NA</b>	118	-	-	-
<b>Reference</b>	14	0.91	0.71	0.80
<b>Index</b>	6	-	-	-
<b>Clinical Trial ID</b>	4	1.00	0.50	0.67

Rule-based methods (basic types + specialists) lead to high precision for Literals, CT IDs, and Ref.

As expected, also can lead to low recall (e.g., CT ID) when rules don't provide coverage

Lower scores for predicting Wikidata links

*Computed Precision & Recall at the label level*  
*Precision: # of correct predictions / total predictions*  
*Recall: # of correct predictions / expected predictions*

# Wikidata Annotation – Analysis



*Does the expected annotation appear in the top 25 candidates for table cells where expected annotation is a Wikidata item?*

<b>Expected Annotation in Candidate Set</b>	~60% (554/910)
<b>Expected Annotation Missing in Candidate Set</b>	~40% (356/910)

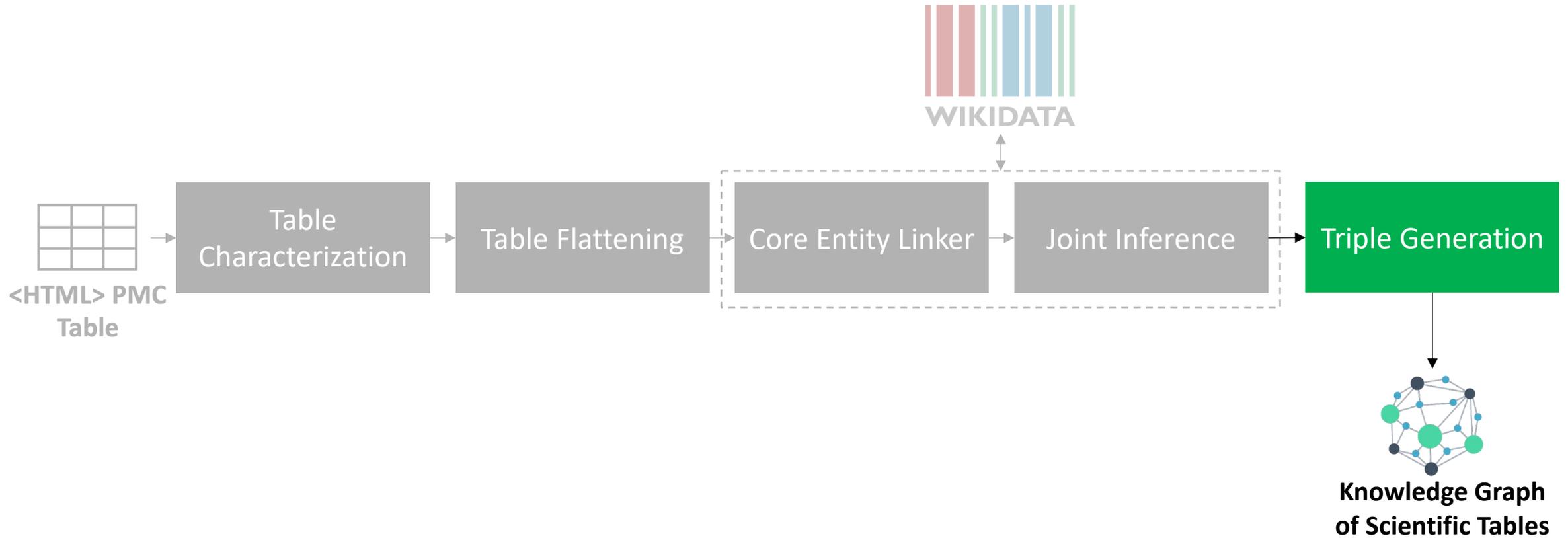
Core linker is not able to retrieve candidates for a fairly large fraction of possible idiomatic strings

*Did the system confuse literals and entities?*

<b>Expected Wikidata Annotation</b>	910
<b>System Predicted Wikidata Annotation</b>	1373
<b>Literal/Ref./CT ID Cells misclassified as Wikidata Links</b>	488

Non-entity strings mapped/predicted as Wikidata annotations makes a substantial negative impact on performance

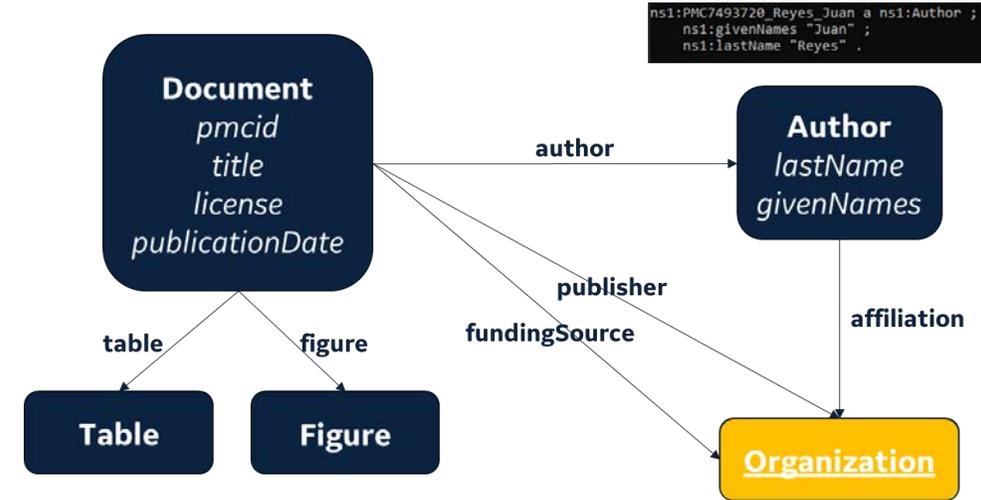
# Knowledge Graph population



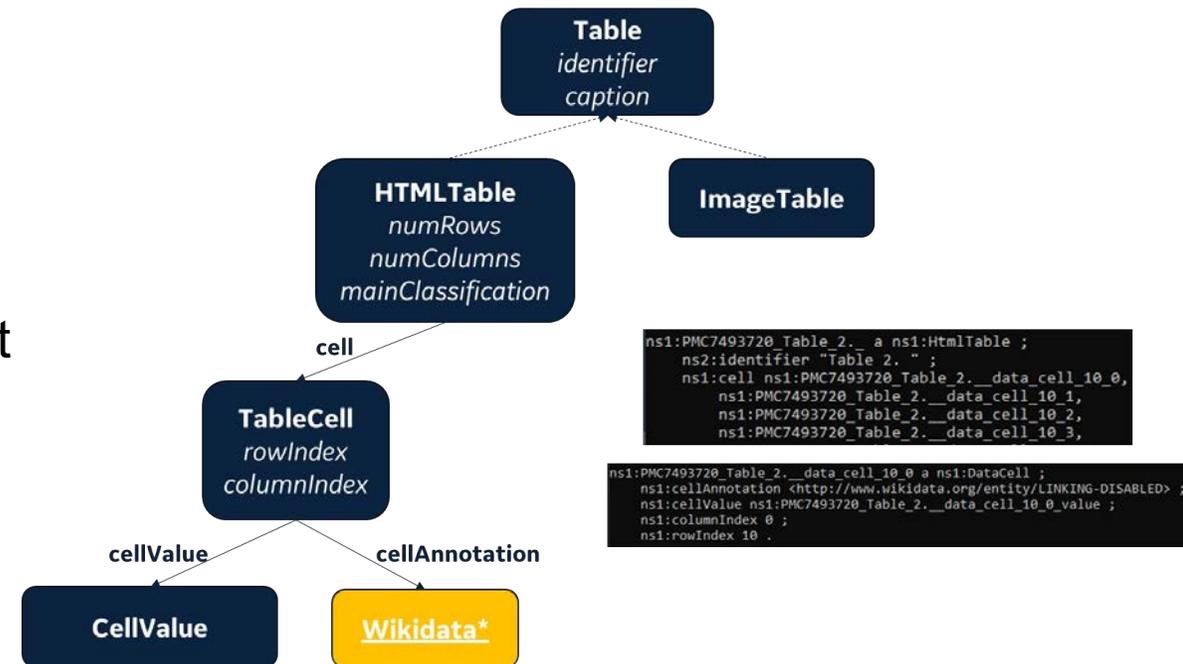
# Ontology and Triple Generation



- We developed an ontology to represent article metadata, data and inferred semantics of scientific tables
  - Builds on W3C standards, including PROV
- We auto-generate RDF triples at document-level, followed by table-level, and finally at cell-level
- Roughly 1000 triples per PMC document (but varies widely based on table count and contents)
- Knowledge graph persisted in triple store to support table discovery
  - Open-source: Apache Jena/TDB2/Fuseki, RDFlib, pyfuseki, ...



```
ns1:PMC7493720_Reyes_Juan a ns1:Author ;
ns1:givenNames "Juan" ;
ns1:lastName "Reyes" .
```



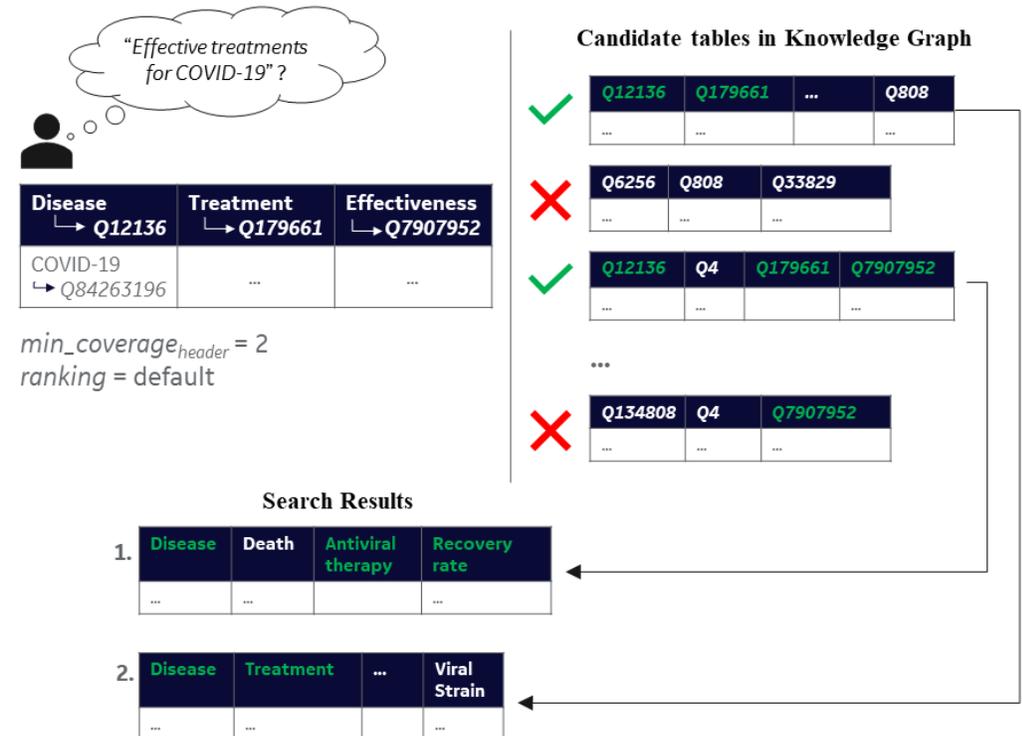
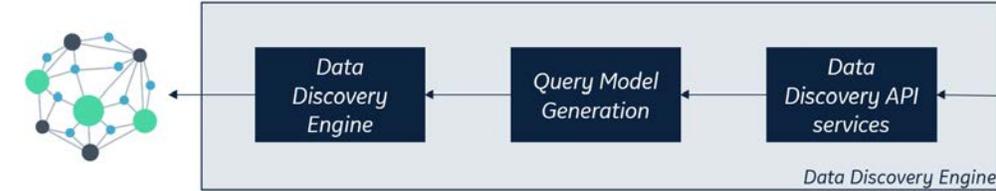
```
ns1:PMC7493720_Table_2_... a ns1:HtmlTable ;
ns2:identifier "Table 2. " ;
ns1:cell ns1:PMC7493720_Table_2_..._data_cell_10_0,
ns1:PMC7493720_Table_2_..._data_cell_10_1,
ns1:PMC7493720_Table_2_..._data_cell_10_2,
ns1:PMC7493720_Table_2_..._data_cell_10_3,
```

```
ns1:PMC7493720_Table_2_..._data_cell_10_0 a ns1:DataCell ;
ns1:cellAnnotation <http://www.wikidata.org/entity/LINKING-DISABLED> ;
ns1:cellValue ns1:PMC7493720_Table_2_..._data_cell_10_0_value ;
ns1:columnIndex 0 ;
ns1:rowIndex 10 .
```

# Discovery of Scientific Tables

overview

- Search and retrieval of tabular information from our knowledge graph
- Emulate database-style discovery operations against semantically annotated scientific tables
- Leverage semantic technologies for search:
  - Header-cell semantics → search/filter matching tables, and on-the-fly fusion of contextually-compatible tables
  - Header-cell & data-cell semantics (incl. units) → disambiguate contents of on-the-fly fused tables
  - *Reliability scores* → filter out unreliable tables from the search, ranking and fusion processes



# Discovery of Scientific Tables

prototype search application and user interface

1

### ProCure Data Discovery

Enter list of search terms / Upload file

coronavirus  
Mapped to Q57751738: Coronavirus

country  
Q6256: country

vaccine  
Mapped to Q87719492: COVID-19 vaccine

ProCure Search Advanced Search I'm Feeling Lucky Reset

Searching for tabular objects of the form:

Q57751738	Q6256	Q87719492
...	...	...

2

### Result Constraints:

1.  Table must have caption? 2. Return All types of tables 3. Ti

Constrain the type of returned tables

4. Coverage constraints 1 5

Min.# of matching header cells Min.# rows in

5. Reliability constraints 0.25 <= Rel\_PROV <= 1

above below

### Result Ranking Preferences:

# of matching header cells in table Highest-first first

Sort by Preference order

3

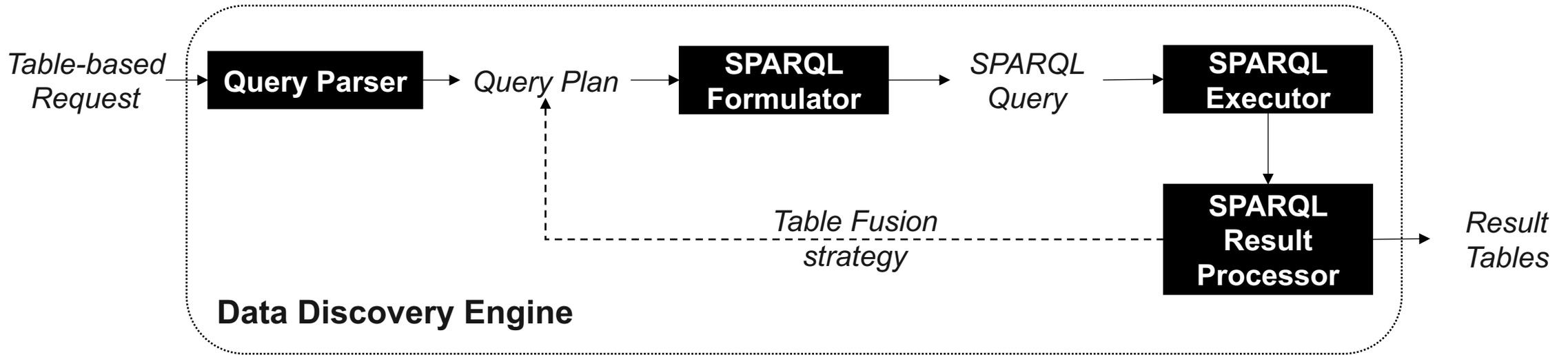
Retrieved 3 original results (0.31 seconds)  
Retrieved 1 fused results (1.76 seconds)

TABLE ID	TIME OF PUBLICATION	RELIABILITY SCORE	HEADERS
FUSED_Table_9757773	2022-05-26		Vaccine Target Vector/Adjuvant Type of Study Stage Participants Country References Institution M
PMC7350246_Table_5	2020-06-17		Vaccine Target Vector/Adjuvant Type of Study Stage Participants Country References
PMC7426550_Table_1	2020-07-31		Disease Virus Main Manifestation(s) Hospitalization Antiviral Therapy Vaccine Case Fatality (%)

1. Interface for table-based semantic search queries
  - list of query terms, semantic resolution of query terms (map to items in knowledge base), 'sketch' desired resultant table
2. Advanced search features
  - multiple contextual constraints and ranking preferences
3. Sample response to search query
  - ranked list of original and on-the-fly fused tables
  - explore content and provenance of individual result tables
  - highlighted header cells depict semantic match
  - reliability score provides guardrails for safe search

# Discovery of Scientific Tables

*table discovery engine: system-centric view*



Any search request gets transformed into a query plan composed of one or more of ‘primitive’ operators. Each operator is then translated into an ad hoc query clause as part of incremental SPARQL query formulation.



# Conclusions and Future Work

- Tables in scientific documents contain important information
- Knowledge discovery from scientific tables is as vital as from text
- We implemented preliminary prototype systems for constructing and searching over a knowledge graph of scientific tables.
- Novel aspects: Two-stage table interpretation, table-based semantic search, on-the-fly table generation

<b>Table Characterization</b> (over 120K tables)	High precision. <b>Future:</b> Improve recall for certain labels by including additional rules or exploring development of supervised algorithms.
<b>Semantic Interpretation</b>	Performs well when our system discovers appropriate entity in its ranked candidate set but fails otherwise. <b>Future:</b> Further adapt candidate search to idiomatic strings; Techniques to distinguish between literals and entity strings.
<b>Tabular Data Discovery</b> (only strict header match)	Low mean average precision, but high recall. <b>Future:</b> Enhanced semantic matching and information fusion across tables.

# Acknowledgements

This research was based on work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2021-21022600004]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government.