A Practical Entity Linking System for Tables in Scientific Literature

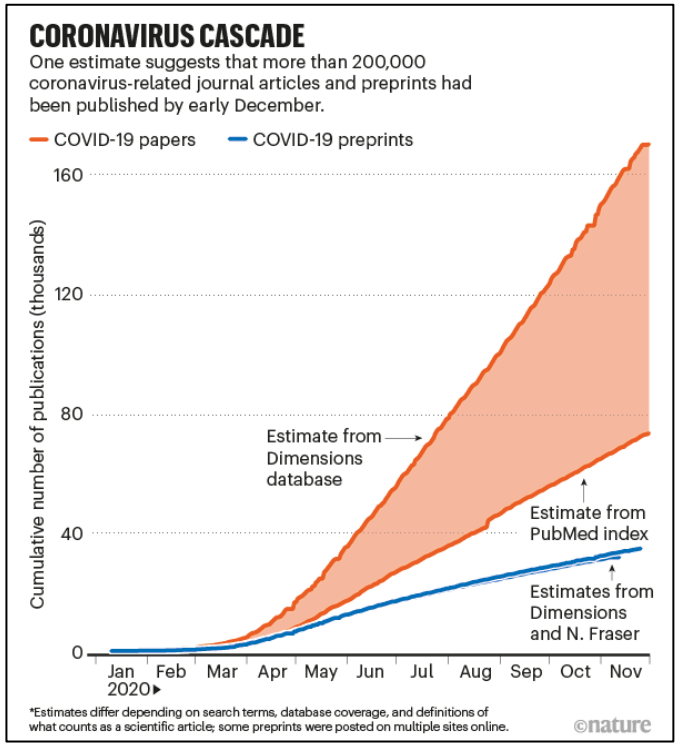
Varish Mulwad[1], Tim Finin[2], Vijay S. Kumar[1],
Jenny Weisenberg Williams[1], Sharad Dixit[1], Anupam Joshi[2]

1: GE Research, Niskayuna, USA
2: UMBC, Baltimore, MD

3rd Workshop on Scientific Document Understanding, AAAI, February 2023

Introduction

- An *infodemic* of scientific literature
  - COVID-19 saw an explosion in scientific publication rates
  - Causes: role of preprint servers, open access publishing
  - Problems: prevalence of misinformation, ‘fake science’
- Most research analyzing publications has focused only on text content
  - CORD-19 dataset of articles on COVID-19 led to tools for search, Q&A, recommendation, summarization over scientific documents
- But significant information is also in tables and data charts
  - Our focus is on understanding information contained in **tables** drawn from scientific documents within specialized domains (e.g., biomedical)

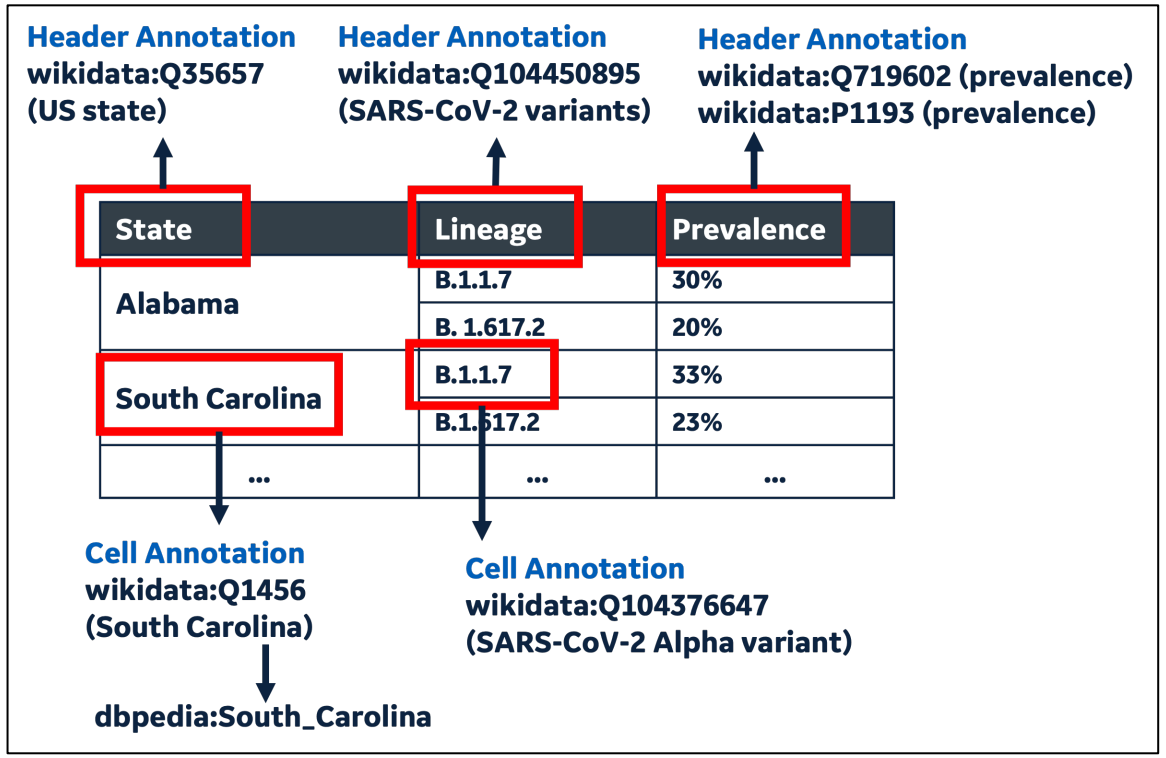


Else, H., 2020. How a Torrent of COVID Science changed Research Publishing—in seven charts. *Nature*, V588 (553).

Research Goal

- Information in scientific tables critical to new knowledge-driven applications
  - e.g., genomic surveillance to track spread of virus variants
- Need to search relevant tables for vital knowledge nuggets
  - Possibly fusing information from multiple tables on the fly
- Our goal: understand scientific tables to infer their semantics and relevance to search queries

Linking table cells to concepts in a background knowledge graph like Wikidata is an important step in understanding



"Treatment efficacies against the top prevalent COVID-19 variants in southern US states"

Tables in Scientific Documents

- Significant advances in pre-trained / representational models for **well-structured** tabular data, e.g. TaBERT, Sherlock, TABBIE
- But scientific tables bring additional **challenges** and **opportunities**
- For example, these three aspects

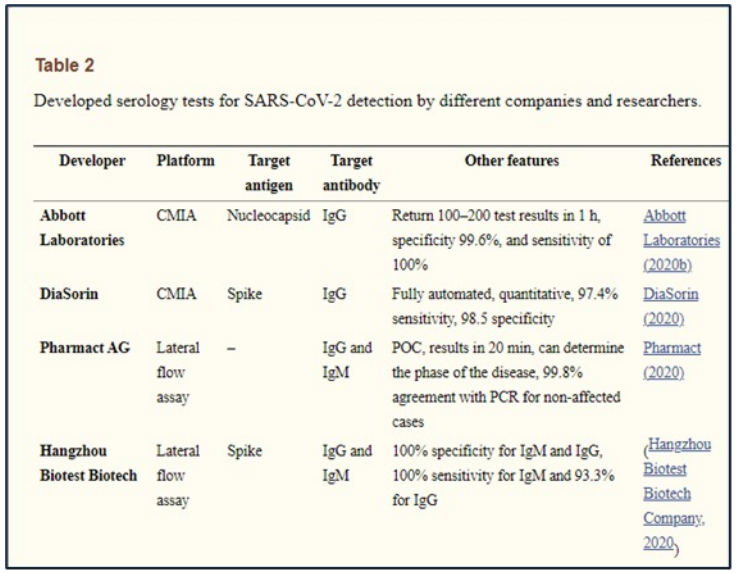


Table 2

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%	Abbott Laboratories (2020b)
DiaSorin	CMIA	Spike	IgG	Fully automated, quantitative, 97.4% sensitivity, 98.5 specificity	DiaSorin (2020)
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	Pharmact (2020)
Hangzhou Biotest Biotech	Lateral flow assay	Spike	IgG and IgM	100% specificity for IgM and IgG, 100% sensitivity for IgM and 93.3% for IgG	(Hangzhou Biotest Biotech Company, 2020)

Similar to web tables … with domain specific entities

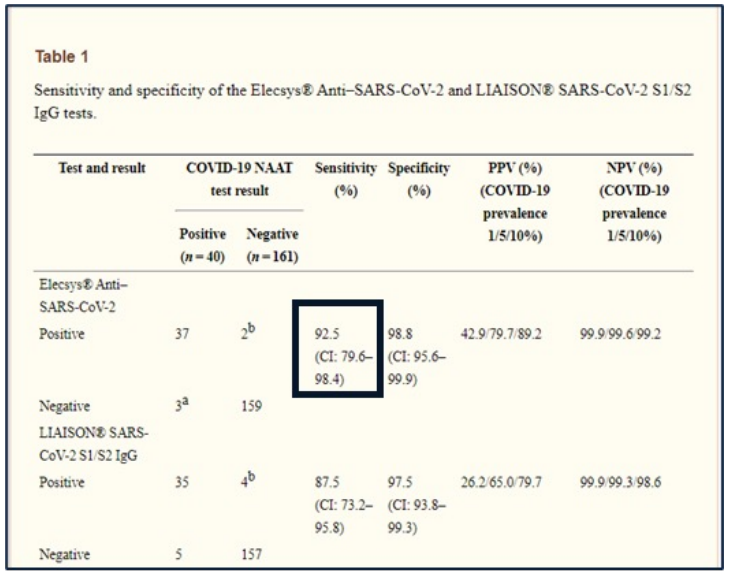


Table 1

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[b]	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[a]	159				
LIAISON® SARS-CoV-2 S1/S2 IgG						
Positive	35	4[b]	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

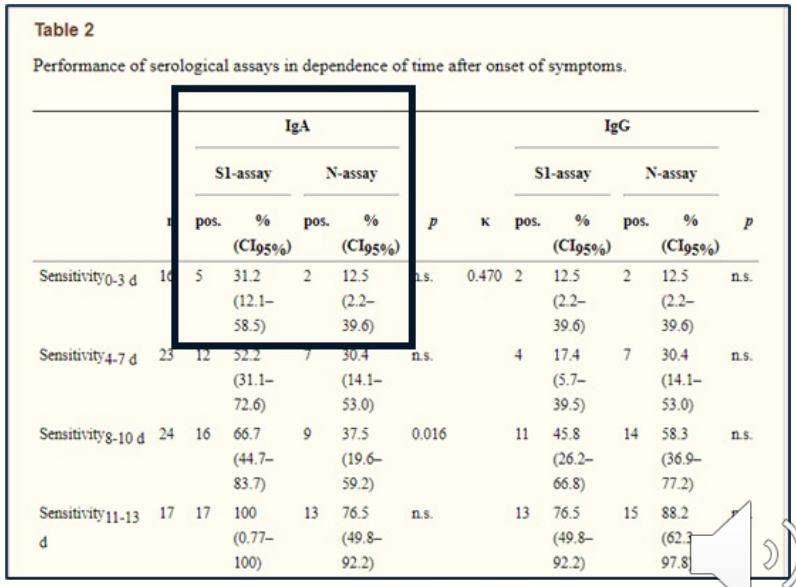


Table 2

Performance of serological assays in dependence of time after onset of symptoms.

		IgA						IgG					
		S1-assay		N-assay				S1-assay		N-assay			
	n	pos.	% (CI95%)	pos.	% (CI95%)	p	κ	pos.	% (CI95%)	pos.	% (CI95%)	p
Sensitivity0-3 d	10	5	31.2 (12.1–58.5)	2	12.5 (2.2–39.6)	n.s.	0.470	2	12.5 (2.2–39.6)	2	12.5 (2.2–39.6)	n.s.
Sensitivity4-7 d	25	12	52.2 (31.1–72.6)	7	30.4 (14.1–53.0)	n.s.		4	17.4 (5.7–39.5)	7	30.4 (14.1–53.0)	n.s.
Sensitivity8-10 d	24	16	66.7 (44.7–83.7)	9	37.5 (19.6–59.2)	0.016		11	45.8 (26.2–66.8)	14	58.3 (36.9–77.2)	n.s.
Sensitivity11-13 d	17	17	100 (0.77–100)	13	76.5 (49.8–92.2)	n.s.		13	76.5 (49.8–92.2)	15	88.2 (62.3–97.8)	n.s.

Row and Column Headers

Tables in Scientific Documents

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%	Abbott Laboratories (2020b)
DiaSorin	CMIA	Spike	IgG	Fully automated, quantitative, 97.4% sensitivity, 98.5 specificity	DiaSorin (2020)
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	Pharmact (2020)
Hangzhou Biotest Biotech	Lateral flow assay	Spike	IgG and IgM	100% specificity for IgM and IgG, 100% sensitivity for IgM and 93.3% for IgG	(Hangzhou Biotest Biotech Company, 2020)

Similar to web tables … with domain specific entities

Tables in Scientific Documents

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[b]	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[a]	159				
LIAISON® SARS-CoV-2 S1/S2 IgG						
Positive	35	4[b]	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

Tables in Scientific Documents

Table 2

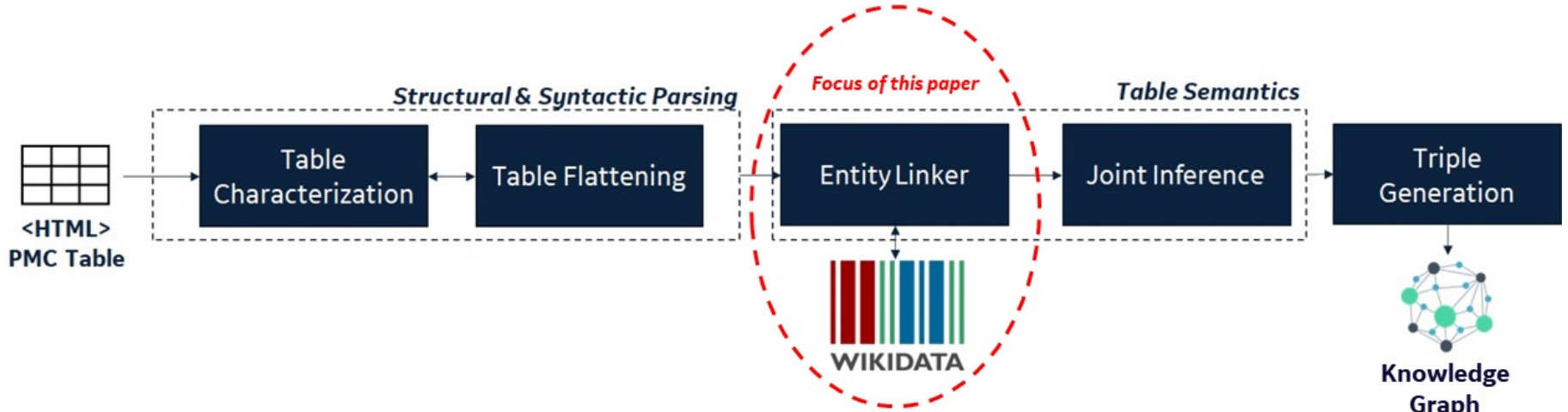
Performance of serological assays in dependence of time after onset of symptoms.

		IgA						IgG					
		S1-assay		N-assay				S1-assay		N-assay			
	n	pos.	% (CI95%)	pos.	% (CI95%)	p	κ	pos.	% (CI95%)	pos.	% (CI95%)	p
Sensitivity 0-3 d	16	5	31.2 (12.1–58.5)	2	12.5 (2.2–39.6)	n.s.	0.470	2	12.5 (2.2–39.6)	2	12.5 (2.2–39.6)	n.s.
Sensitivity 4-7 d	23	12	52.2 (31.1–72.6)	7	30.4 (14.1–53.0)	n.s.		4	17.4 (5.7–39.5)	7	30.4 (14.1–53.0)	n.s.
Sensitivity 8-10 d	24	16	66.7 (44.7–83.7)	9	37.5 (19.6–59.2)	0.016		11	45.8 (26.2–66.8)	14	58.3 (36.9–77.2)	n.s.
Sensitivity 11-13 d	17	17	100 (0.77–100)	13	76.5 (49.8–92.2)	n.s.		13	76.5 (49.8–92.2)	15	88.2 (62.3–97.8)	n.s.

Row and Column Headers

Approach

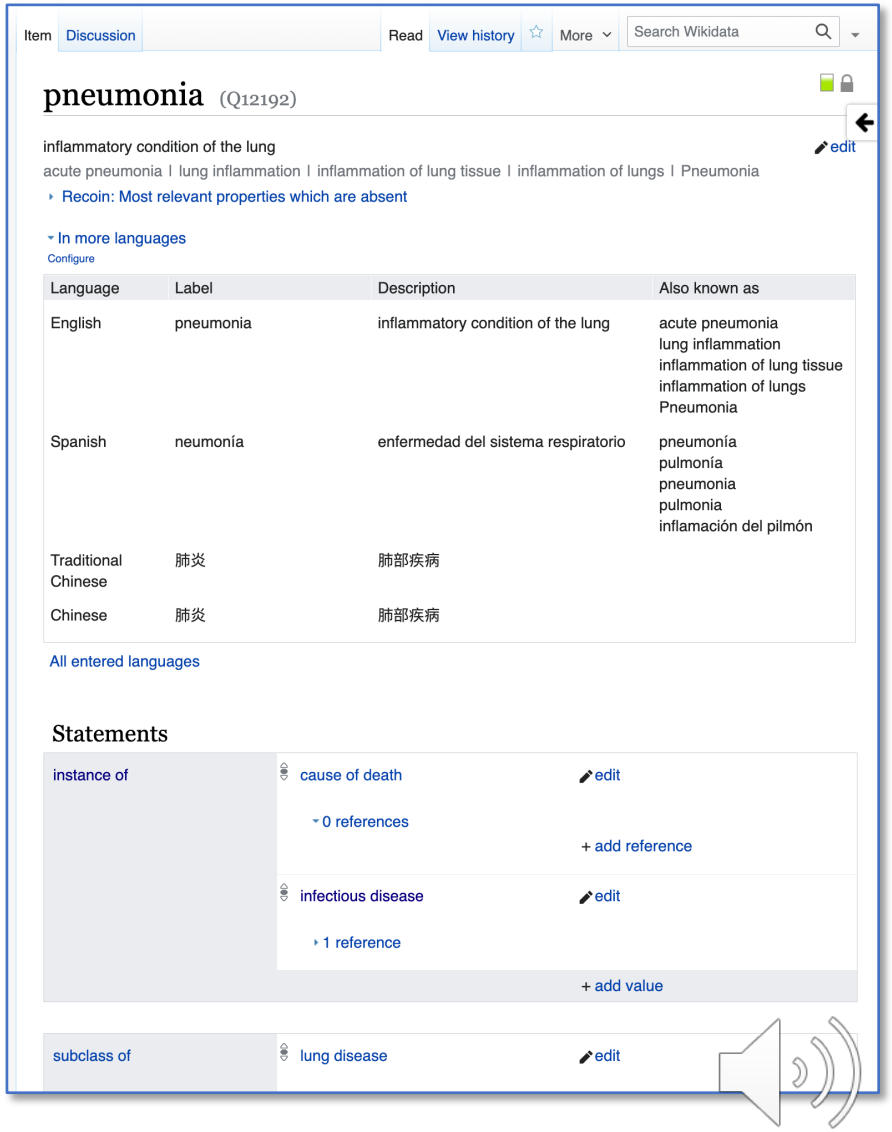
- Represent scientific tables as semantically annotated **linked data**
  - complementary approach accounting for structural heterogeneity & specialized terms
  - link with knowledge extracted from other information types (text, charts) and sources (e.g., document provenance)
  - leverage this additional context, structural characterization, and explicit reasoning over tabular content to infer table semantics
- We use a complex pipeline to construct a **knowledge graph** from tables in **PubMed Central** (PMC) documents
- This paper focuses on linking table header/body cells to Wikidata items



Linking entities to Wikidata's Knowledge Graph

- Recognizing and understanding entities and concepts in text, tables, and graphs essential
- System links text mentions to items in Wikidata's graph of ~100M entities

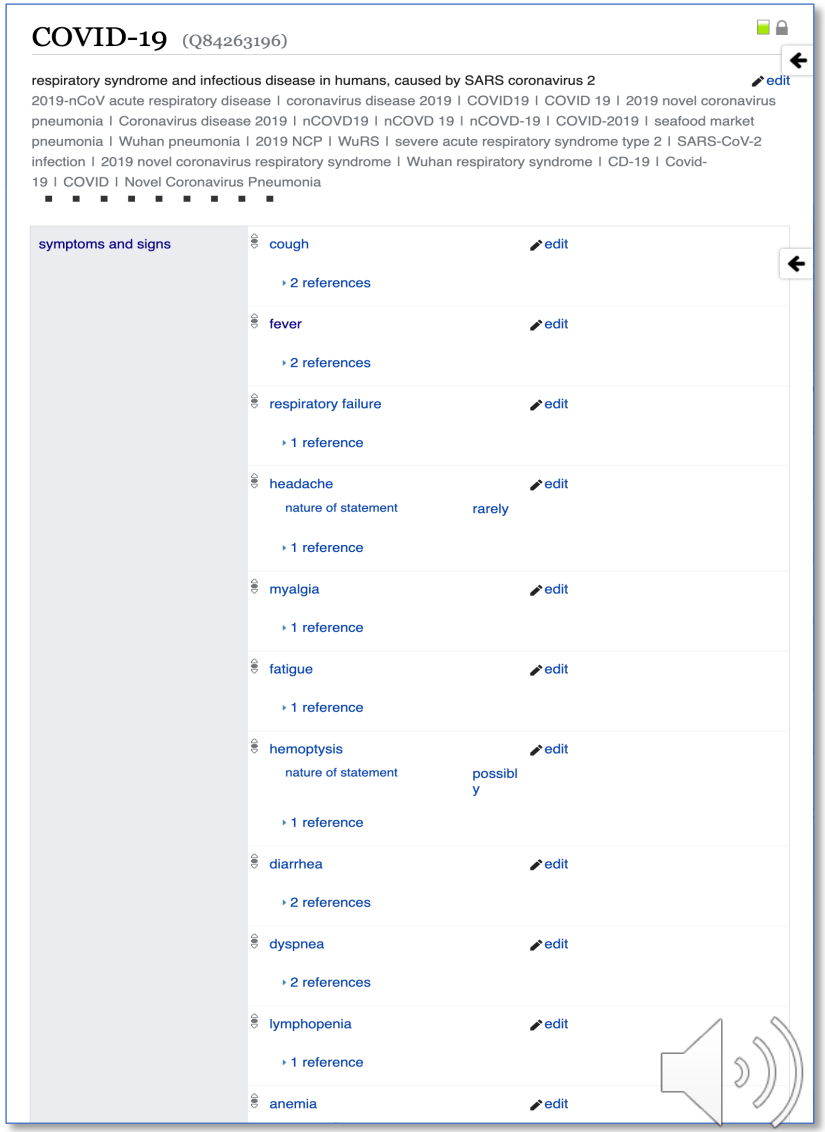
  Including biomedical concepts like genes, diseases, drugs, symptoms and their names, aliases, and descriptions in multiple languages
- Supported by an ultra-fine-grained ontology of >2.5M types for entities, properties & constraints
- Items also linked to other knowledge graphs: DBpedia, Google Knowledge Graph, NCI Thesaurus, Medical Subject Headings thesaurus and more.



Linking entities to Wikidata's Knowledge Graph

The linked items can be

- Proper *entities* like **pneumonia** (Q12192), which is an instance of an **infectious disease** (Q18123741)
- General *concepts*, like **hospitalization rate** (Q107527870) and **blood pressure** (Q82642) that may represent properties
- *Wikidata properties* like **symptoms and signs** (P780) that connect a medical condition with its possible symptoms



Linking noun chunks in table reference

Sentence referencing a table in an article on COVID-19:

> **Trials** involving **vaccines**, **antiviral drugs**, **immunotherapies**, **monoclonal antibodies**, **stem cells**, and **nitric oxide** are summarized in **Table** 1.

Without additional adaptation or training for the medical domain, our linker finds good links for phrases using the whole sentence as context:

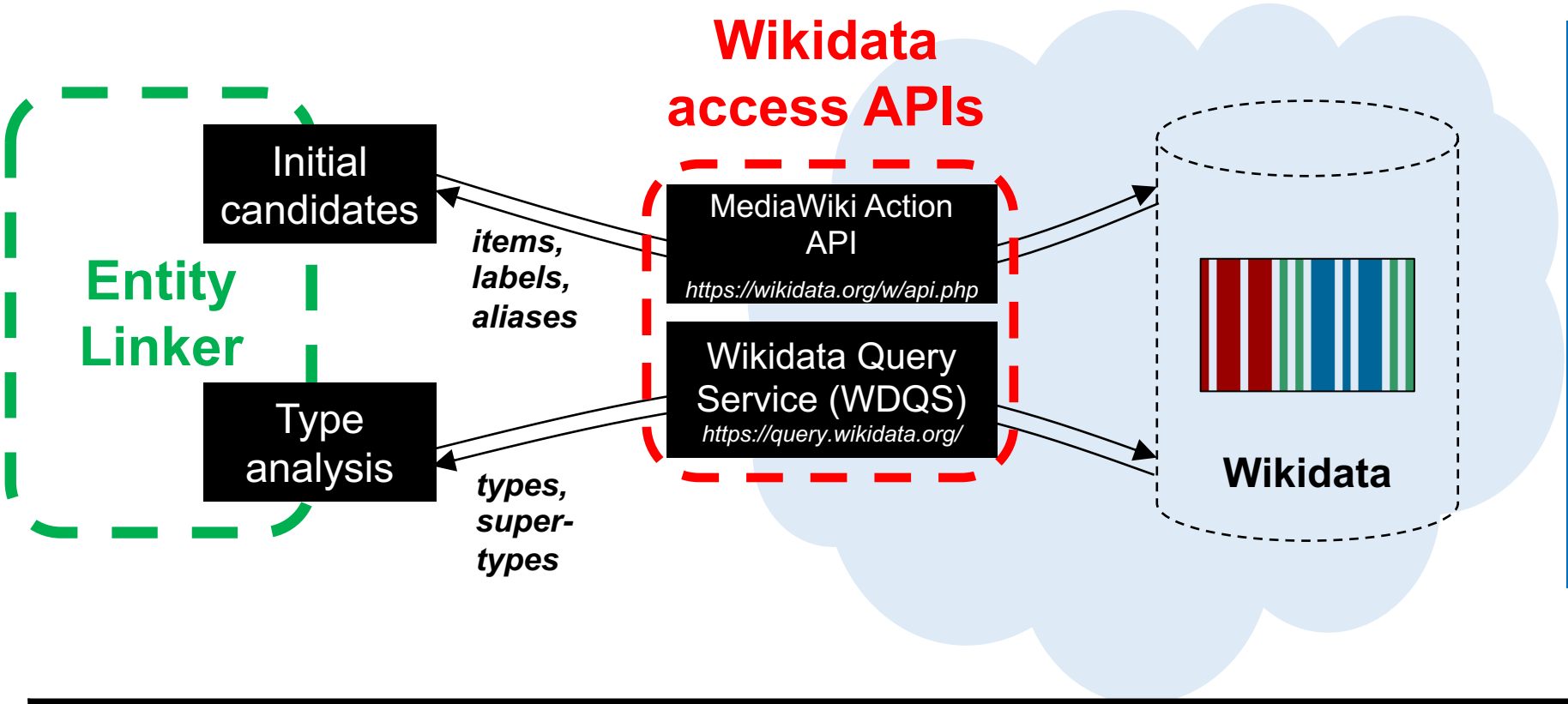
- **Trials** ('Q1436668', 'randomized controlled trial', 'experimental ...")
- **vaccines** ('Q134808', 'vaccine', 'a substance used to stimulate ...)
- **antiviral drugs** ('Q40207875', 'antiviral agent', 'substance that destroys ...)
- **immunotherapies** ('Q1427096', 'immunotherapy', 'therapy to elicit or ...)
- **monoclonal antibodies** ('Q422248', 'monoclonal antibody', 'monospecific ...)
- **stem cells** ('Q48196', 'stem cell', 'undifferentiated biological cells ...)
- **nitric oxide** ('Q14916164', 'nitric oxide biosynthetic process', 'The chemical ...')

Online Linking Approach

Wikidata's search APIs produce ranked candidates which we filter by **provenance**, **inherited types,** and **context**

- Use a mapping from **SpaCy** ontonotes types to Wikidata types, augmented with additional biomedical types
- WD *bad types* eliminate items unlikely to be good, e.g., fictional characters, scholarly papers, songs, musical groups, etc.
- The **SpaCy NLP** system makes type errors, so we employ a weighted mapping from target type (e.g., LOC) to *near misses* (e.g., FAC, GPE), and fallback *ok types* (e.g., ENTITY)
- **Embeddings** are used to compare a candidate link's short description with a mention sentence

Efficient Entity Linking at Large Scale



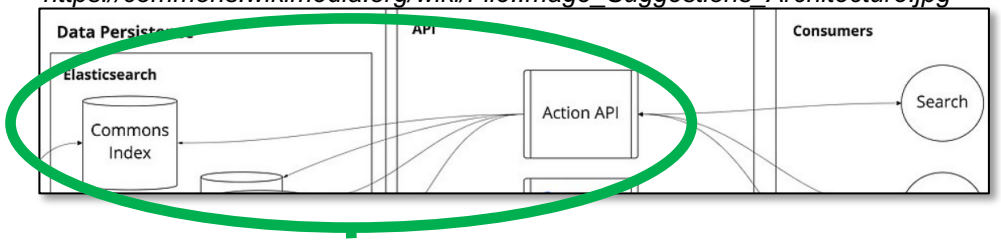
- **Rate-limited public Wikidata access APIs a major scaling bottleneck in our entity linker**
- **Cannot sustain modern scientific publication rates**

Average time taken by Entity Linker for mention string: "coronavirus"	
Retrieve initial ranked list of candidate items via **Action API** (top 20-50 candidates)	**12 sec.**
Retrieve all **types and supertypes** per candidate **via WDQS**. Filter & rank candidates based on type analysis (good/OK/bad) for domain	**18 sec.**

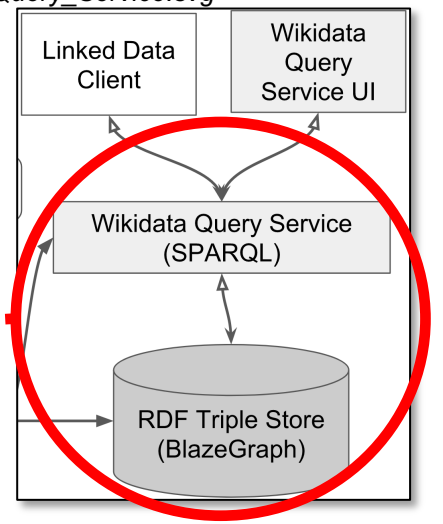
Efficient Entity Linking at Large Scale

- Avoiding API invocations via caching results helped, but not enough
- So, we built an “offline” linker to eliminate all reliance on Wikidata APIs
- We replicated *just enough* Wikidata functionality locally

From *https://commons.wikimedia.org/wiki/File:Image_Suggestions_Architecture.jpg*

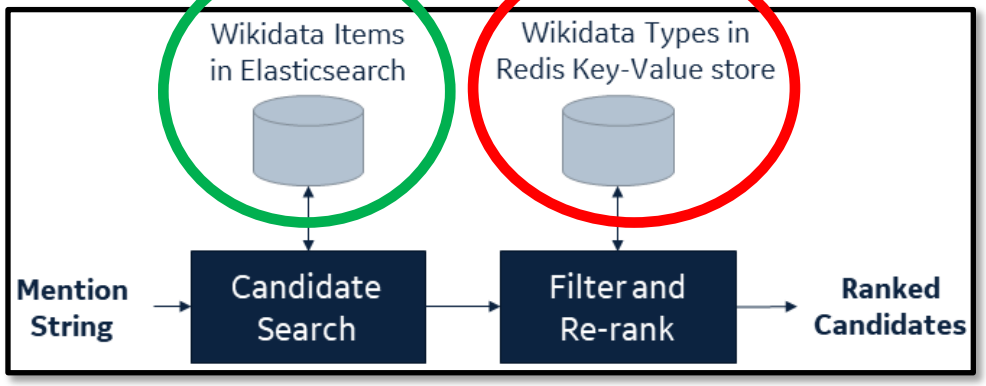


From *https://commons.wikimedia.org/wiki/File:Wikidata_Architecture_Overview_-_Query_Service.svg*



Public Wikidata-based

Local



Functional Architecture of local “offline” Entity Linker

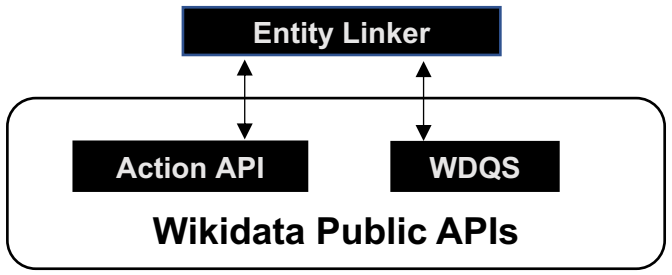
Efficient Entity Linking at Large Scale

Phased approach to transition to an offline entity linker:

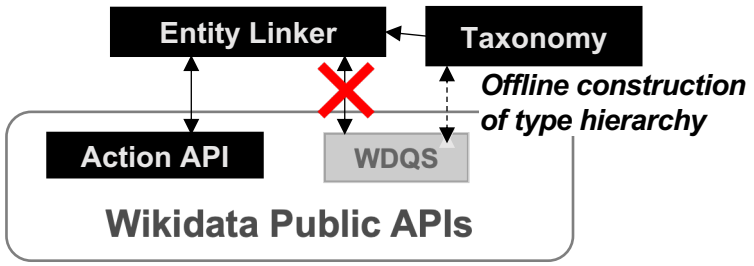
1

Offline, we use WDQS to retrieve current **taxonomy** of Wikidata types and their supertypes.

Resulting dictionary mapping each of **2.6M types** to their supertypes is loaded into **Redis**, an efficient key-value store.



Retrieve types via WDQS + type analysis: **~18 sec.**



Retrieve types via Redis lookup + type analysis: **~9.5 sec.**

- **Entity linker result remains unaffected**
- **Local taxonomy can be updated at some predefined frequency (e.g., daily)**

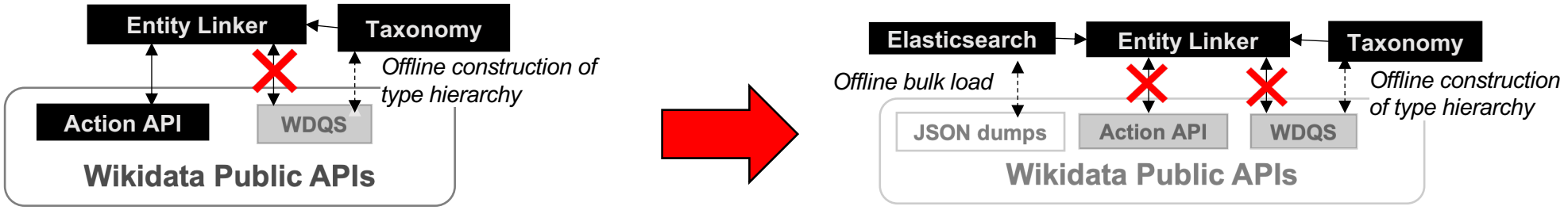
Efficient Entity Linking at Large Scale

Phased approach to transition to an offline entity linker:

2

Offline, we upload partial (JSON) dump of Wikidata entities into local Elasticsearch, an efficient, scalable search engine

Resulting searchable index contained **95.8M items**, their attributes (label, aliases, description) and sitelinks counts



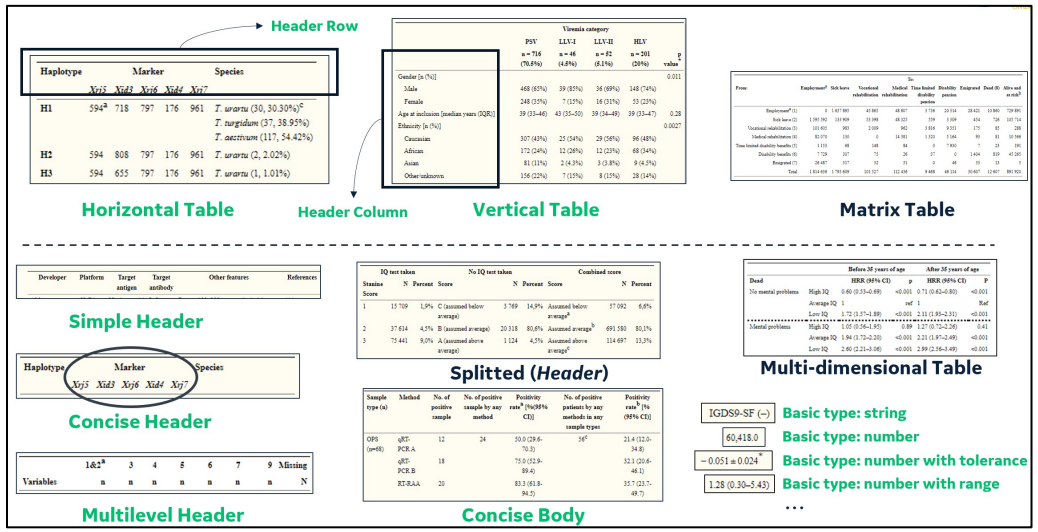
Get top 20 candidate items via Action API: **~12 sec.**

Get top 20 candidate items from Elasticsearch: **~0.08 sec.**

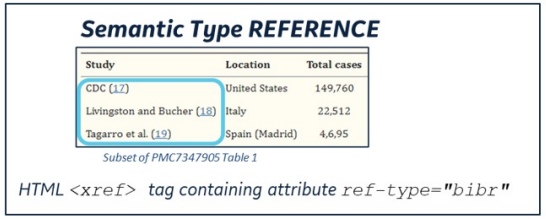
- **sitelinks count used to rank items. Initial candidate list has impact on linker quality.**
- **Trade-off: Tolerable reduction in quality for 3x faster linking.**

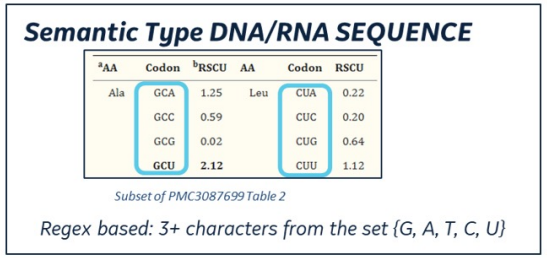
Inferring Semantics of Tables

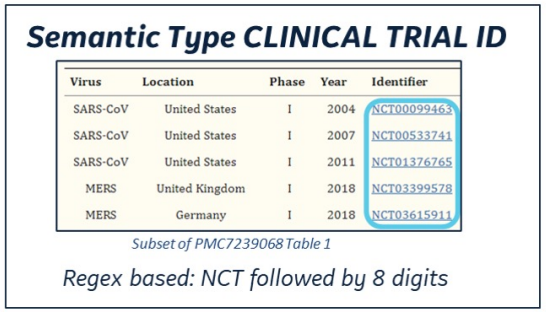
To improve overall entity linking performance, we first structurally & syntactically characterize scientific tables using two approaches

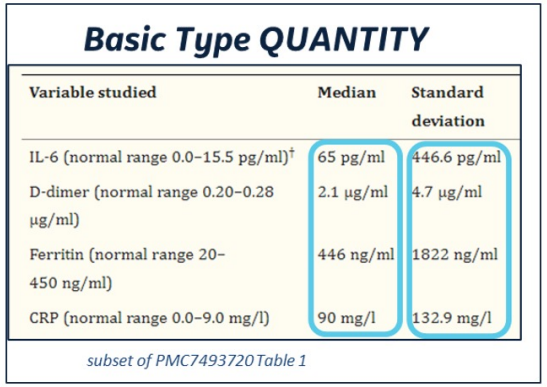


Rule-based characterizations to categorize tables into types based on their structure



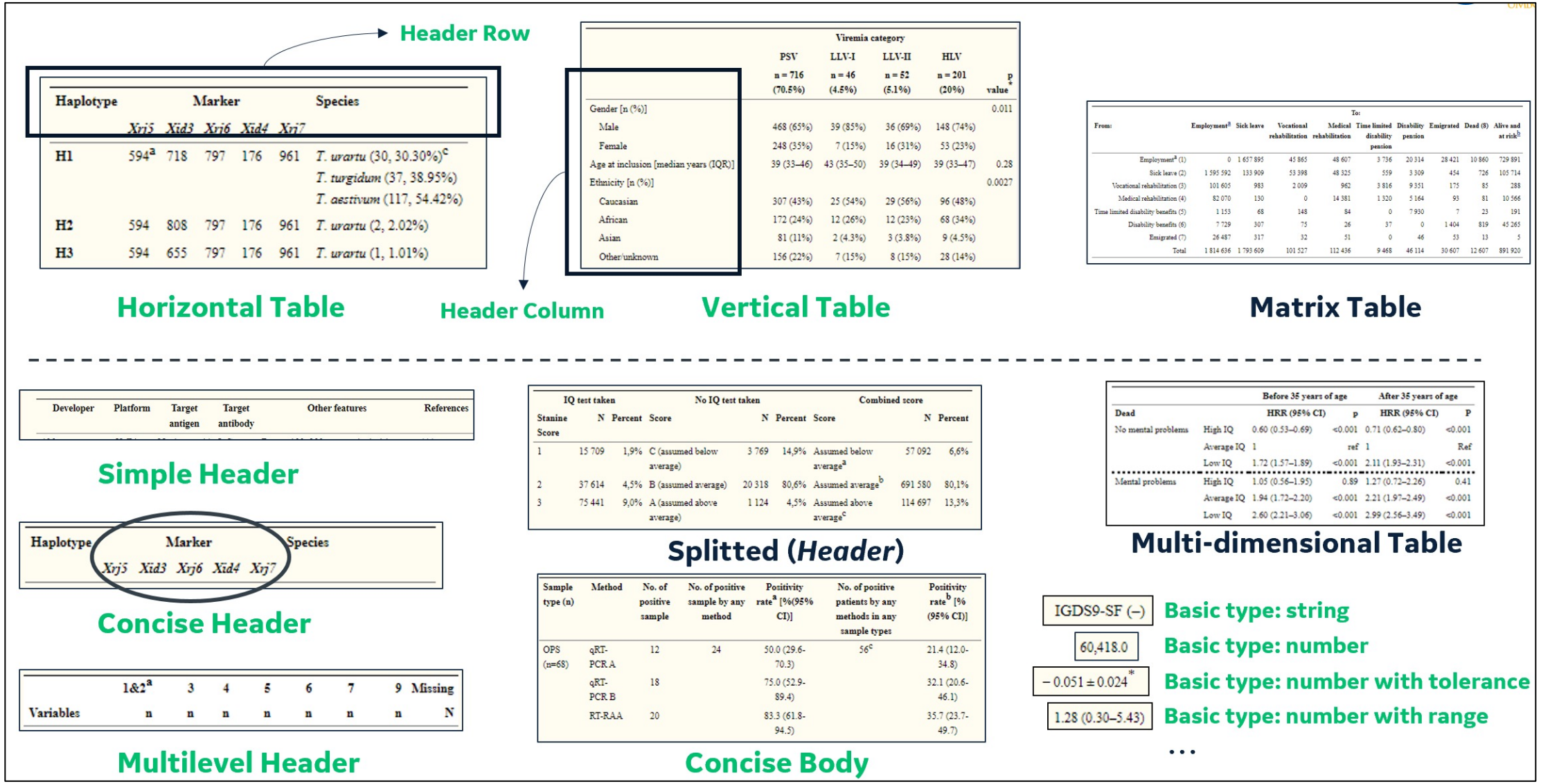






Specialists (pattern-based or ML-based) assess commonly encoded data types in cells to avoid linking those with specific kinds of literals

Inferring Table Semantics: Rules



Rule-based characterizations classify tables into types based on their structure

Inferring Table Semantics: ML

Semantic Type REFERENCE

Study	Location	Total cases
CDC (17)	United States	149,760
Livingston and Bucher (18)	Italy	22,512
Tagarro et al. (19)	Spain (Madrid)	4,6,95

Subset of PMC7347905 Table 1

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

Semantic Type DNA/RNA SEQUENCE

[a]AA	Codon	[b]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

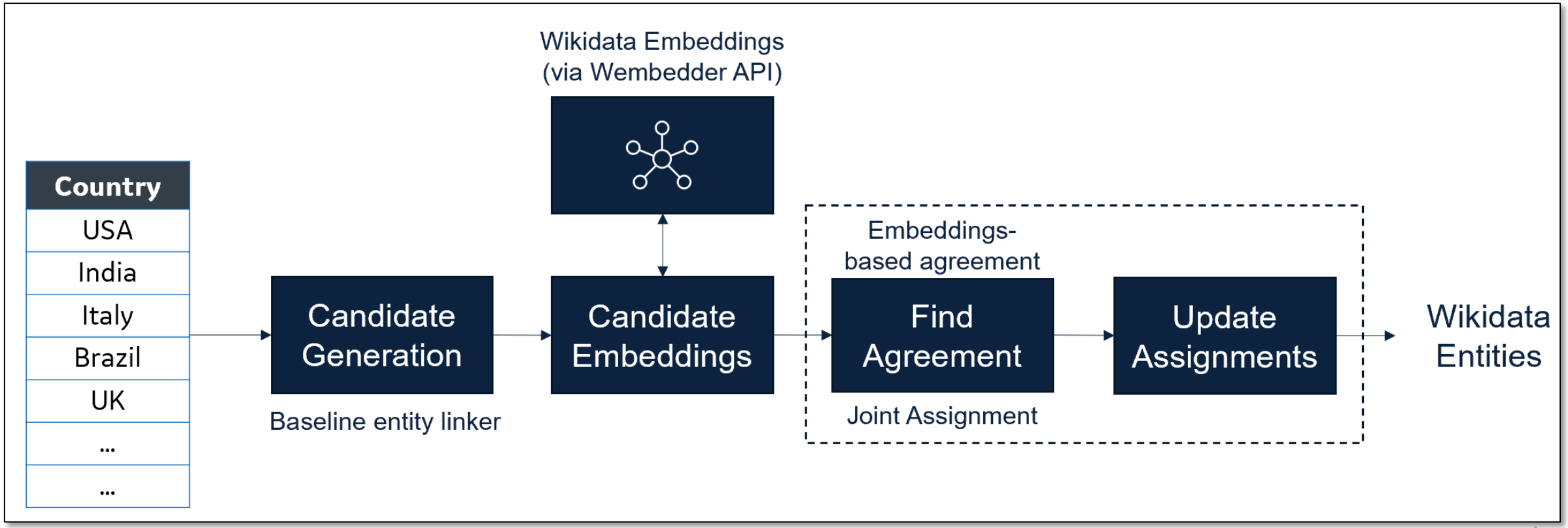
Variable studied	Median	Standard deviation
IL-6 (normal range 0.0–15.5 pg/ml)[†]	65 pg/ml	446.6 pg/ml
D-dimer (normal range 0.20–0.28 µg/ml)	2.1 µg/ml	4.7 µg/ml
Ferritin (normal range 20–450 ng/ml)	446 ng/ml	1822 ng/ml
CRP (normal range 0.0–9.0 mg/l)	90 mg/l	132.9 mg/l

subset of PMC7493720 Table 1

Pattern-based and ML-based specialists assess commonly encoded data types in cells to avoid linking those with specific kinds of literals

Inferring Semantics of Tables

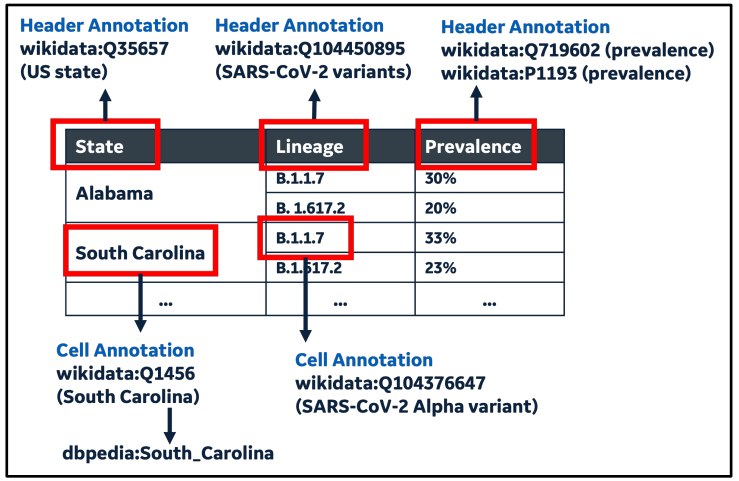
Besides mapping table cells to Wikidata items, inferring semantics of scientific tables is improved using **joint inference** and **embeddings**

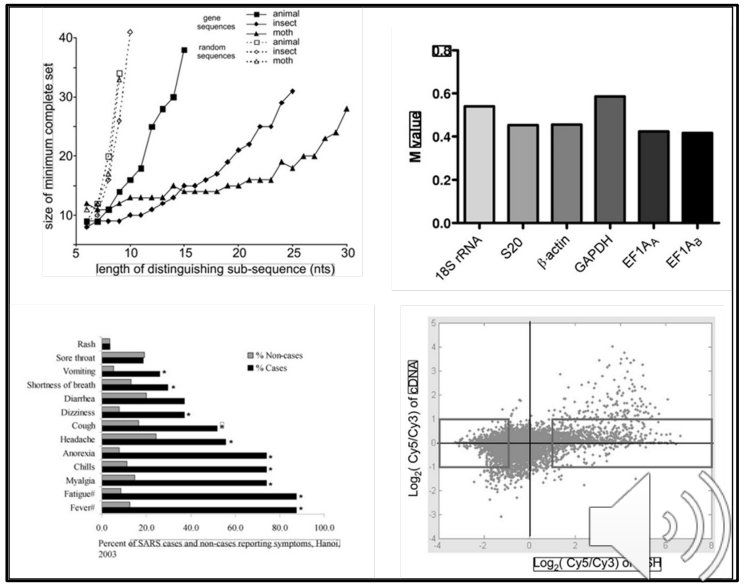


Joint inference using **embeddings** of Wikidata items and embeddings-driven agreement function to compute compatibility between entities and assign entities to cells in a column

Conclusion and Future Work

- **Data Tables** in scientific documents have important information
- Understanding their semantics requires **linking** their elements to concepts and entities in a background knowledge base, like **Wikidata**
- We implemented an initial online system using Wikidata's APIs and a much custom more **efficient offline version**
- Our **future work** will extend and integrate a similar system for **extracting and representing data** found in **data chart images**





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.