enRichMyData

SINTEF *

^ INSID&S Lab

Department of Informatics,
Systems and Communication
University of Milan - Bicocca

Semantic Data Enrichment: from Interactive Exploration to Scalable Deployment

Roberto Avogadro *, Flavio De Paoli ^, Dumitru Roman *, Matteo Palmonari ^

Part III: Selected State-of-the-art

*This work presented in this presentation has received funding from the European Union's Horizon 2020 research and innovation program under grant agreements No 732590 - **EW-Shopp** - and No 732003 – **euBusinessGraph** - and from the European Union's Horizon Europe research and innovation program under grant agreements No 101070284 - **enRichMyData**.*

Outline

- Part II: Semantic Data Enrichment, Applications and Requirements
  - Semantics and KGs for data enrichment
  - The *Link & Extend* enrichment paradigm
    - Interactive exploration and scalability
- Part III: Selected State-of-the-art
  - Data preparation solutions
    - The broader context of data preparation solutions
  - Scalable data pipelines
    - A quick introduction to solutions for scalability
  - Tabular data annotation
    - From heuristic techniques to generative LLMs
- Part IV: Semantic Data Enrichment in Practice with Tools
  - Service-based approach
    - Data model for interoperability
    - Service model for composability
  - Interactive definition of pipelines
    - Exploration with graphical UI
    - Pipeline definition with programmatic UI
  - Pipeline execution at scale
    - Execution with workflow managers (Argo & TAO)
  - Live demos
- Part V: Conclusions and Discussion
  - Wrap-up and take-home messages
  - Discussion

Part III: Selected State-of-the-art

1) Data preparation solutions

“Which tools can support interactive data exploration and specification of data preparation pipelines?”

Data Preparation Features in Commercial Tools (2020)

- Analysis of different commercial tools for data preparation in [Hameed & Nauman 2020]
  - Tools that specifically address the data preparation task
  - Availability of a comprehensive and intuitive GUI to select and apply preparations
  - Tools that specifically address the data preparation task
  - Comprehensive and sophisticated preparation features
  - Proper documentation for the tools
  - Availability of a trial version / customer assistance

Tool name	URL
Altair Monarch Data Preparation	https://www.datawatch.com/in-action/monarch-draft/
Paxata Self Service Data Preparation	https://www.paxata.com/self-service-data-prep/
SAP Agile Data Preparation	https://www.sap.com/germany/products/data-preparation.html
SAS Data Preparation	https://www.sas.com/en_us/software/data-preparation.html
Tableau Prep	https://www.tableau.com/products/prep
Talend Data Preparation	https://www.talend.com/products/data-preparation/
Trifacta Wrangler	https://www.trifacta.com/products/wrangler-editions/

Data Preparation Features in Commercial Tools (2020)

- Analysis of different commercial tools for data preparation in [Hameed & Nauman 2020]
  - Tools that specifically address the data preparation task
  - Availability of a comprehensive and intuitive GUI to select and apply preparations
  - Tools that specifically address the data preparation task
  - Comprehensive and sophisticated preparation features
  - Proper documentation for the tools
  - Availability of a trial version / customer assistance

Categories	**Available features**	**Data preparation tools**						
		Altair	Paxata	SAP	SAS	Tableau	Talend	Trifacta
Data discovery	Locate missing values (nulls)	✓	✓	✓	✓	✓	✓	✓
	Locate outliers		✓		✓			✓
	Search by pattern	✓	✓	✓	✓	✓	✓	✓
	Sort data	✓	✓	✓	✓	✓	✓	✓
Data validation	Compare values (selection and join)	✓	✓	✓		✓	✓	✓
	Check data range	✓	✓	✓		✓	✓	✓
	Check permitted characters							✓
	Check column uniqueness	✓	✓	✓		✓	✓	✓
	Find type-mismatched data		✓	✓		✓	✓	✓
	Find data-mismatched datatypes		✓				✓	✓
Data structuring	Change column data type	✓	✓	✓	✓	✓	✓	✓
	Delete column	✓	✓	✓	✓	✓	✓	✓
	Detect & change encoding						✓	✓
	Pivot / unpivot	✓	✓	✓		✓		✓
	Rename column	✓	✓	✓	✓	✓	✓	✓
	Split column	✓	✓	✓	✓	✓	✓	✓
	Transform by example [13]						✓	✓
Data enrichment	Assign semantic data type				✓	✓	✓	
	Calculate column using expressions	✓	✓	✓	✓	✓	✓	✓
	Discover & merge external data	✓	✓	✓			✓	✓
	Duplicate column	✓	✓	✓		✓	✓	✓
	Generate primary key column			✓				✓
	Join & union	✓	✓	✓	✓	✓	✓	✓
	Merge columns	✓		✓		✓	✓	✓
	Normalize numeric values	✓	✓	✓	✓	✓	✓	✓
Data filtering	Delete/keep filtered rows	✓	✓	✓	✓	✓	✓	✓
	Delete empty and invalid rows	✓	✓	✓	✓	✓	✓	✓
	Extract value parts	✓			✓		✓	✓
	Filter with regular expressions							✓
Data cleaning	Change date & time format	✓	✓	✓	✓	✓	✓	✓
	Change letter case	✓	✓	✓	✓	✓	✓	✓
	Change number format	✓	✓	✓	✓	✓	✓	✓
	Deduplicate data	✓	✓	✓	✓		✓	✓
	Delete by pattern	✓	✓		✓	✓	✓	✓
	Edit & replace cell data	✓	✓	✓	✓	✓	✓	✓
	Fill empty cells	✓	✓				✓	✓
	Remove extra whitespace	✓	✓	✓	✓	✓	✓	✓
	Remove diacritics			✓				
	Standardize strings by pattern		✓	✓	✓	✓	✓	✓
	Standardize values in clusters		✓	✓	✓	✓	✓	✓

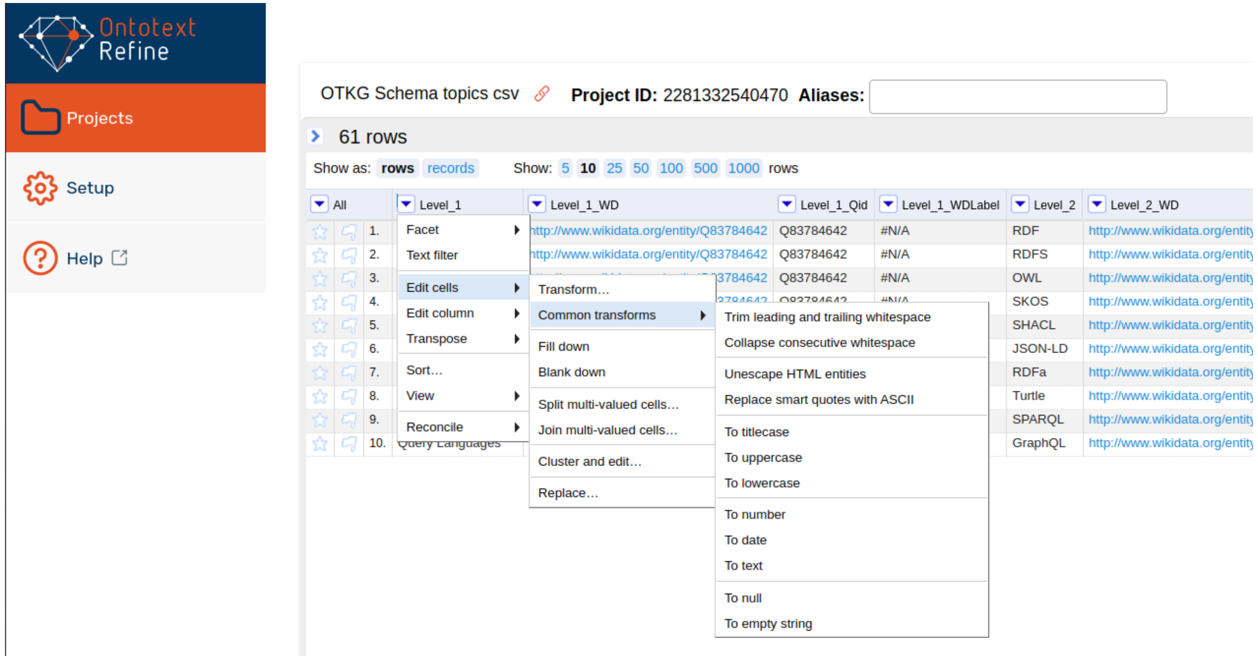
Data Preparation Features vs Semantic Data Enrichment

- Tools in [Hameed & Nauman 2020]
  - Merge of external data is somehow supported
  - Not directly supporting semantic techs / KGs
- Other “non research” tools explicitly supporting KGs: **OpenRefine**
  - Google spin-off project, community-driven, open source
  - Features:
    - Data manipulation / data quality
    - Data linking and extension
    - Large user base
  - Industry spin-offs: OntoText Refine (add batch processing)
  - Inspired our work

Categories	Available features	Data preparation tools						
		Altair	Paxata	SAP	SAS	Tableau	Talend	Trifacta
Data discovery	Locate missing values (nulls)	✓	✓	✓	✓	✓	✓	✓
	Locate outliers		✓		✓			✓
	Search by pattern	✓	✓	✓	✓	✓	✓	✓
	Sort data	✓	✓	✓	✓	✓	✓	✓
Data validation	Compare values (selection and join)	✓	✓	✓		✓	✓	✓
	Check data range	✓	✓	✓		✓	✓	✓
	Check permitted characters							✓
	Check column uniqueness	✓	✓	✓		✓	✓	✓
	Find type-mismatched data		✓	✓		✓	✓	✓
	Find data-mismatched datatypes		✓				✓	✓
Data structuring	Change column data type	✓	✓	✓	✓	✓	✓	✓
	Delete column	✓	✓	✓	✓	✓	✓	✓
	Detect & change encoding						✓	✓
	Pivot / unpivot	✓	✓	✓		✓		✓
	Rename column	✓	✓	✓	✓	✓	✓	✓
	Split column	✓	✓	✓	✓	✓	✓	✓
	Transform by example [13]						✓	✓
Data enrichment	Assign semantic data type				✓	✓	✓	
	Calculate column using expressions	✓	✓	✓	✓	✓	✓	✓
	Discover & merge external data	✓	✓	✓			✓	✓
	Duplicate column	✓	✓	✓		✓	✓	✓
	Generate primary key column			✓				✓
	Join & union	✓	✓	✓	✓	✓	✓	✓
	Merge columns	✓		✓		✓	✓	✓
	Normalize numeric values	✓	✓	✓	✓	✓	✓	✓
Data filtering	Delete/keep filtered rows	✓	✓	✓	✓	✓	✓	✓
	Delete empty and invalid rows	✓	✓	✓	✓	✓	✓	✓
	Extract value parts	✓			✓		✓	✓
	Filter with regular expressions							✓
Data cleaning	Change date & time format	✓	✓	✓	✓	✓	✓	✓
	Change letter case	✓	✓	✓	✓	✓	✓	✓
	Change number format	✓	✓	✓	✓	✓	✓	✓
	Deduplicate data	✓	✓	✓	✓		✓	✓
	Delete by pattern	✓	✓		✓	✓	✓	✓
	Edit & replace cell data	✓	✓	✓	✓	✓	✓	✓
	Fill empty cells	✓	✓				✓	✓
	Remove extra whitespace	✓	✓	✓	✓	✓	✓	✓
	Remove diacritics			✓				
	Standardize strings by pattern		✓	✓	✓	✓	✓	✓
	Standardize values in clusters		✓	✓	✓	✓	✓	✓

Data Preparation Features vs Semantic Data Enrichment

- Tools in [Hameed & Nauman 2020]
  - Merge of external data is somehow supported
  - Not directly supporting semantic techs / KGs
- Other “non research” tools explicitly supporting KGs: **OpenRefine**
  - Google spin-off project, community-driven, open source
  - Features:
    - Data manipulation / data quality
    - Data linking and extension
    - Large user base
  - Industry spin-offs: **OntoRefine** (add batch processing)



Part III: Selected State-of-the-art

2) Scalable data pipelines

"Which kind of solutions exist for scaling up the execution of data enrichment pipelines?"

Scaling Data Transformations

- Definition: scaling data transformations refers to efficiently transforming large datasets to enhance performance, accuracy, and usability.

- Importance:
  - Handling big data: essential for managing and processing large volumes of data.
  - Improving data quality: ensures data is accurate, consistent, and usable.
  - Enabling real-time analytics: supports real-time data processing needs.

Strategies for Scaling Data Transformations

Parallel Processing:

Utilize distributed computing frameworks like Apache Spark or Hadoop.

Split data into smaller chunks and process them simultaneously to improve efficiency.

Optimized Algorithms:

Implement efficient algorithms to reduce computation time.

Use indexing and partitioning techniques to speed up data access and processing.

Cloud Solutions:

Leverage cloud-based services (e.g., AWS Glue, Google Dataflow) for scalable data processing.

Benefit from auto-scaling features to handle varying data loads smoothly.

Incremental Processing:

Process data in increments rather than in large batches.

Use streaming platforms like Apache Kafka to handle real-time data transformation needs.

- A **survey** on large-scale data management in cloud environments [Sakr & Sherif 2011] highlights key scalability techniques
- Popular **tools** include Amazon Kinesis, Apache Beam, and Apache Spark Streaming, with many others available

High-level Functionalities of Scaling Tools

Efficient workflow management requires tools that enhance scalability.

These tools fall into four main macro high level functionalities :

1. Monitoring
    - Purpose: Track system performance and health
    - Functions: Metrics collection, dashboards, alerts
2. Debugging
    - Purpose: Identify and fix software issues
    - Functions: Code inspection, performance profiling, network analysis
3. Scheduling
    - Purpose: Automate and manage task execution
    - Functions: Job scheduling, workflow coordination, resource management
4. Designing (UI)
    - Purpose: Create and improve user interfaces
    - Functions: Wireframing, prototyping, user interaction testing

Selected SOTA approaches

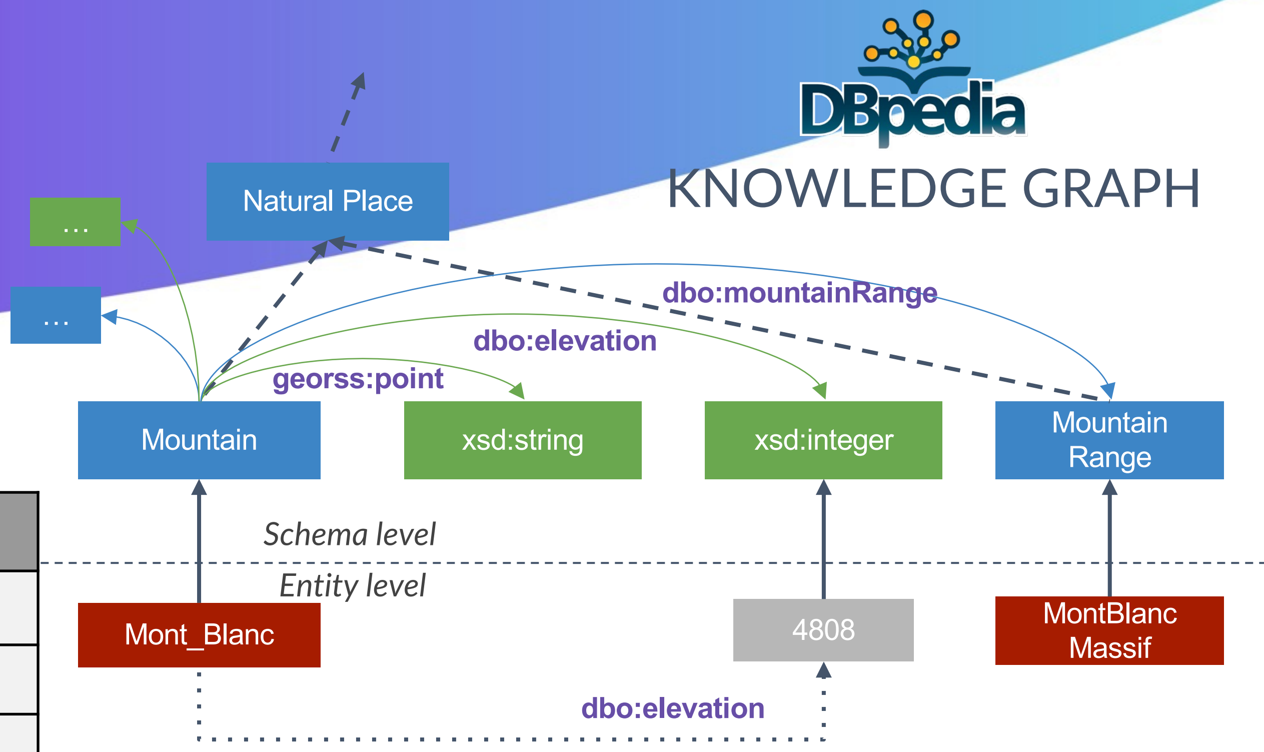
- Scalable techniques using containers [Dessalk et al. 2020] and traditional MAP reduce techniques [Liu et al. 2011]
- Examples of **workflow management tools**: ArgoWorkflow, Apache Airflow, Kubeflow, TAO, and many others

Part III: Selected State-of-the-art

3) Tabular data annotation

"A walk through some recent semantic table annotation approaches under the lenses of data enrichment and its requirements… from heuristic methods to generative LLM-based approaches"

Tabular Data Annotation



Name	Coordinates	Height	Range
Le Mont Blanc	45°49′57″N 06°51′52″E	4808	M. Blanc massif
Hohtälli	45°98’96″N 07°80’25″E	3275	Pennine Alps
Monte Cervino	45°58′35″N 07°39′31″E	4478	Pennine Alps

Given

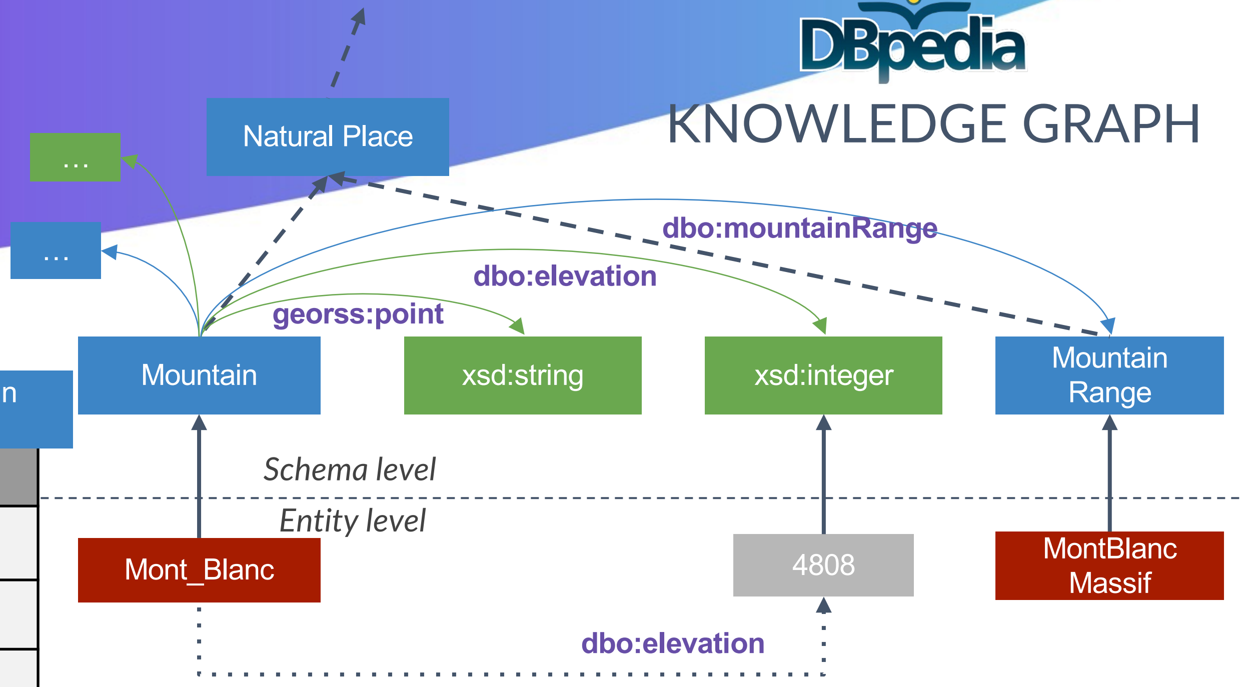
- a relational table T
- a **Knowledge Graph (entities + statements)** and an ontology (**types + predicates)**

T is annotated when:

-
-
-

Tabular Data Annotation

Mountain	xsd:string	xsd:integer	Mountain Range
Name	**Coordinates**	**Height**	**Range**
Le Mont Blanc	45°49′57″N 06°51′52″E	4808	M. Blanc massif
Hohtälli	45°98′96″N 07°80′25″E	3275	Pennine Alps
Monte Cervino	45°58′35″N 07°39′31″E	4478	Pennine Alps



Given

- a relational table T
- a **Knowledge Graph (entities + statements)** and an ontology (**types + predicates)**

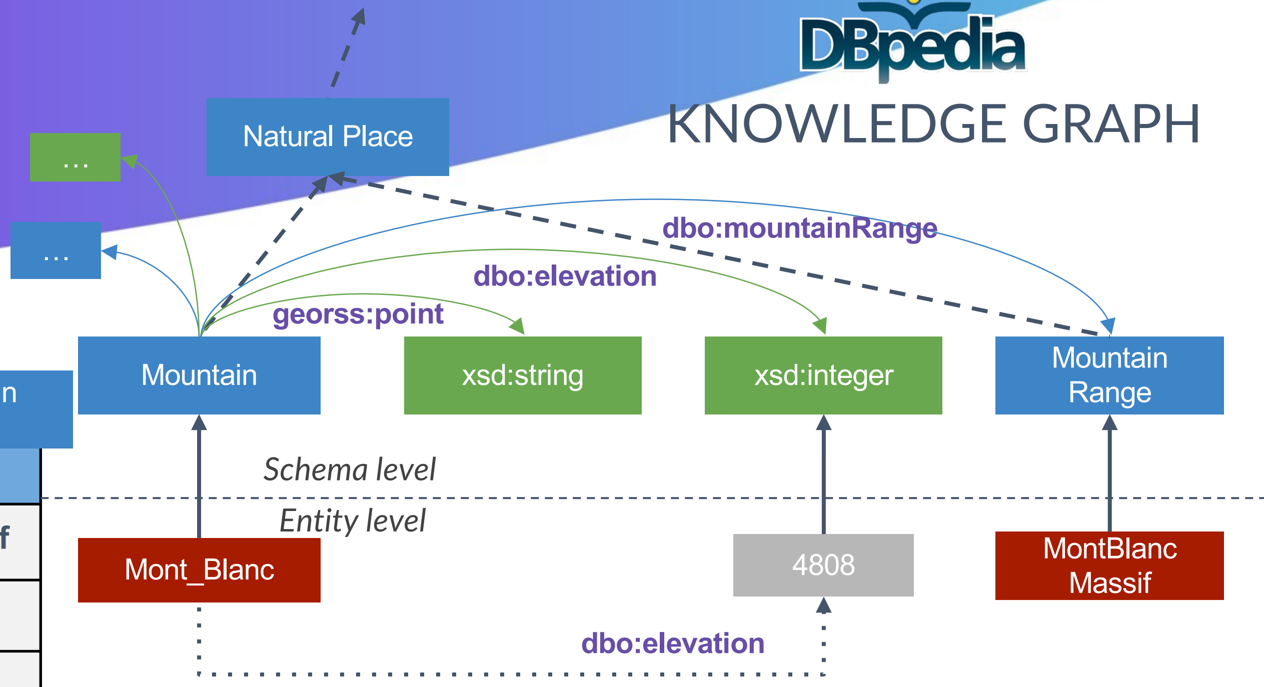
T is annotated when:

- each column is associated with one or more **KG-types** (CTA)
-
-

Tabular Data Annotation

DBpedia KNOWLEDGE GRAPH

Mountain	xsd:string	xsd:integer	Mountain Range
Name	Coordinates	Height	Range
Mont_Blanc	45°49′57″N 06°51′52″E	4808	MontBlanc Massif
Hohtälli	45°98′96″N 07°80′25″E	3275	Pennine Alps
Monte Cervino	45°58′35″N 07°39′31″E	4478	Pennine Alps



Given

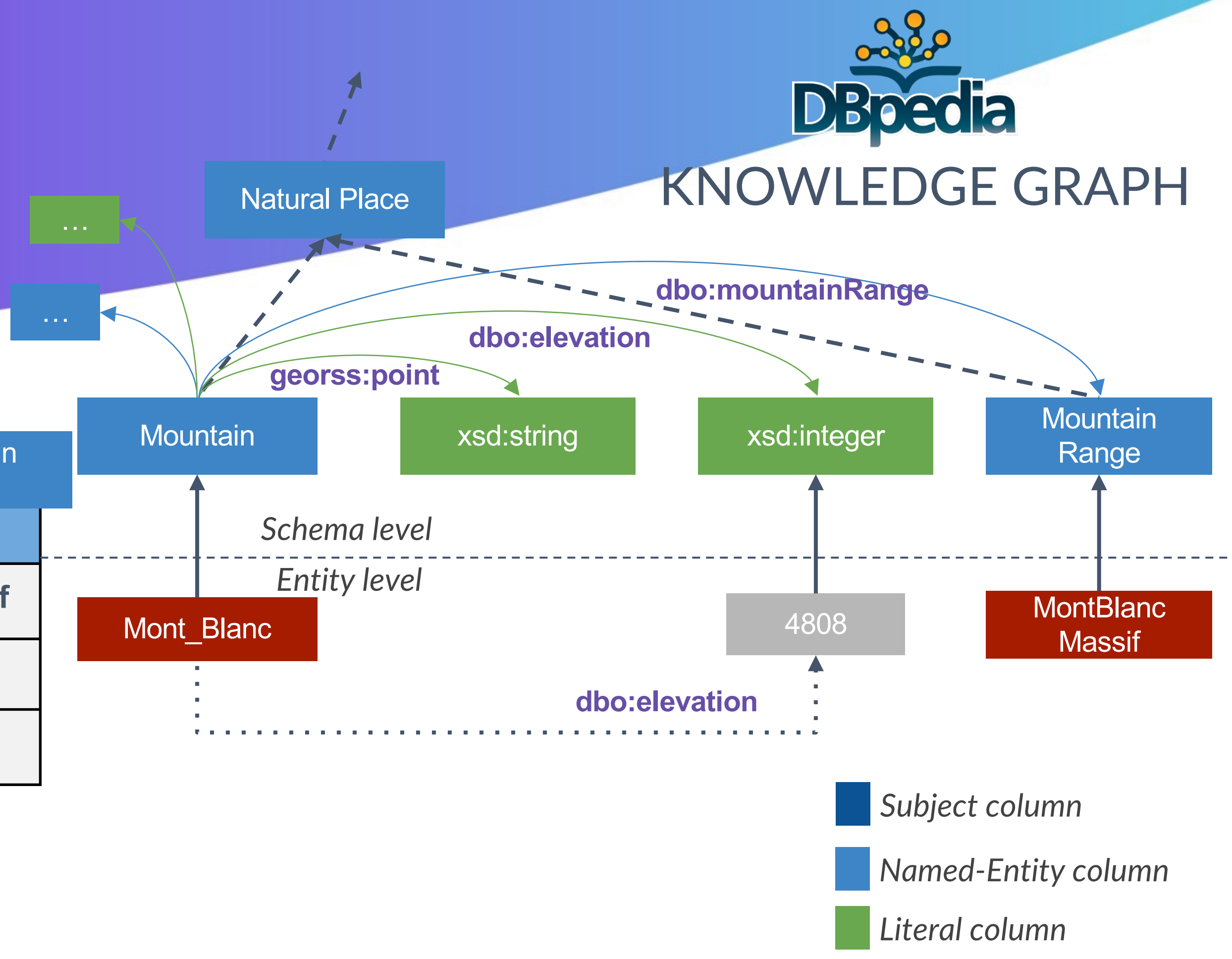
- a relational table T
- a **Knowledge Graph (entities + statements)** and an ontology (**types + predicates)**

T is annotated when:

- each column is associated with one or more **KG-types** (CTA)
- each cell in "entity columns" is annotated with a **KG-entity** (CEA)
-

Tabular Data Annotation

Mountain	xsd:string	xsd:integer	Mountain Range
Name	**Coordinates**	**Height**	**Range**
Mont_Blanc	**45°49′57″N 06°51′52″E**	**4808**	MontBlanc Massif
Hohtälli	45°98′96″N 07°80′25″E	3275	Pennine Alps
Monte Cervino	45°58′35″N 07°39′31″E	4478	Pennine Alps



Given

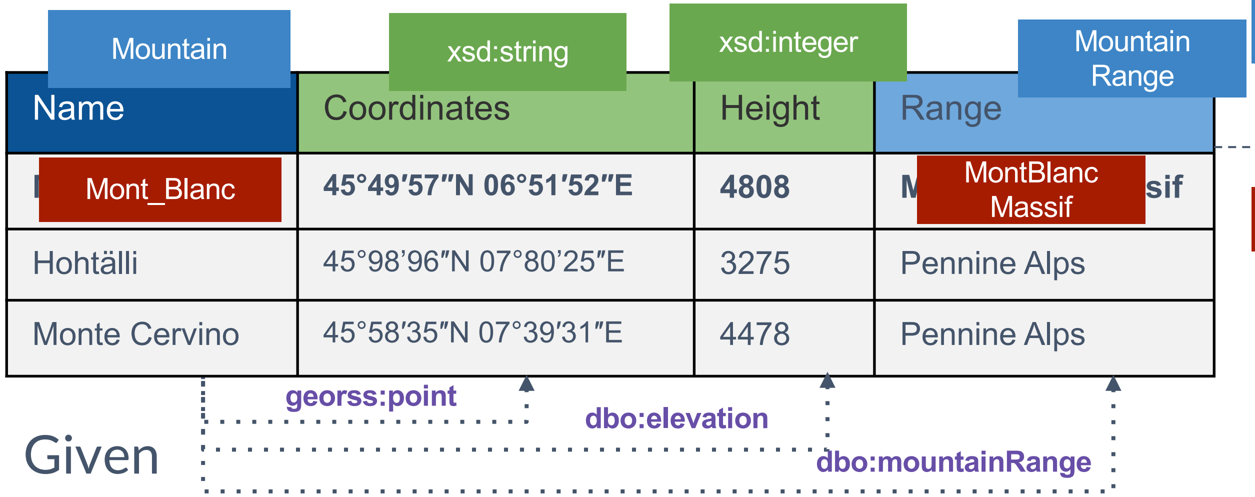
- a relational table T
- a **Knowledge Graph (entities + statements)** and an ontology (**types + predicates)**

T is annotated when:

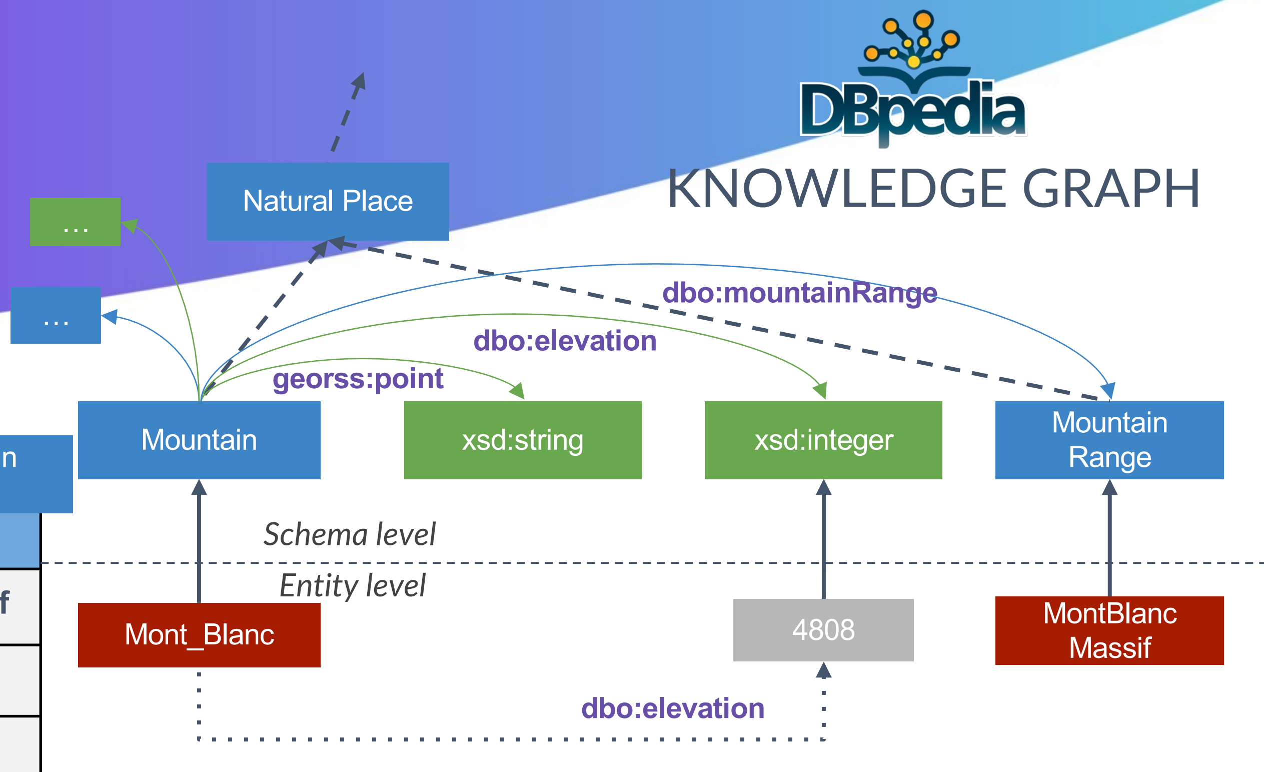
- each column is associated with one or more **KG-types** (CTA)
- each cell in "entity columns" is annotated with a **KG-entity** (CEA)
-

Also referred to as "entity linking" (for tables)

Tabular Data Annotation



Name	Coordinates	Height	Range
Mont_Blanc	**45°49′57″N 06°51′52″E**	**4808**	**MontBlanc Massif**
Hohtälli	45°98′96″N 07°80′25″E	3275	Pennine Alps
Monte Cervino	45°58′35″N 07°39′31″E	4478	Pennine Alps



Given

- a relational table T
- a **Knowledge Graph (entities + statements)** and an ontology (**types + predicates)**

T is annotated when:

- each column is associated with one or more **KG-types** (CTA)
- each cell in “entity columns” is annotated with a **KG-entity** (CEA)
- some pair of columns is annotated with a binary **KG-predicate** (CPA)

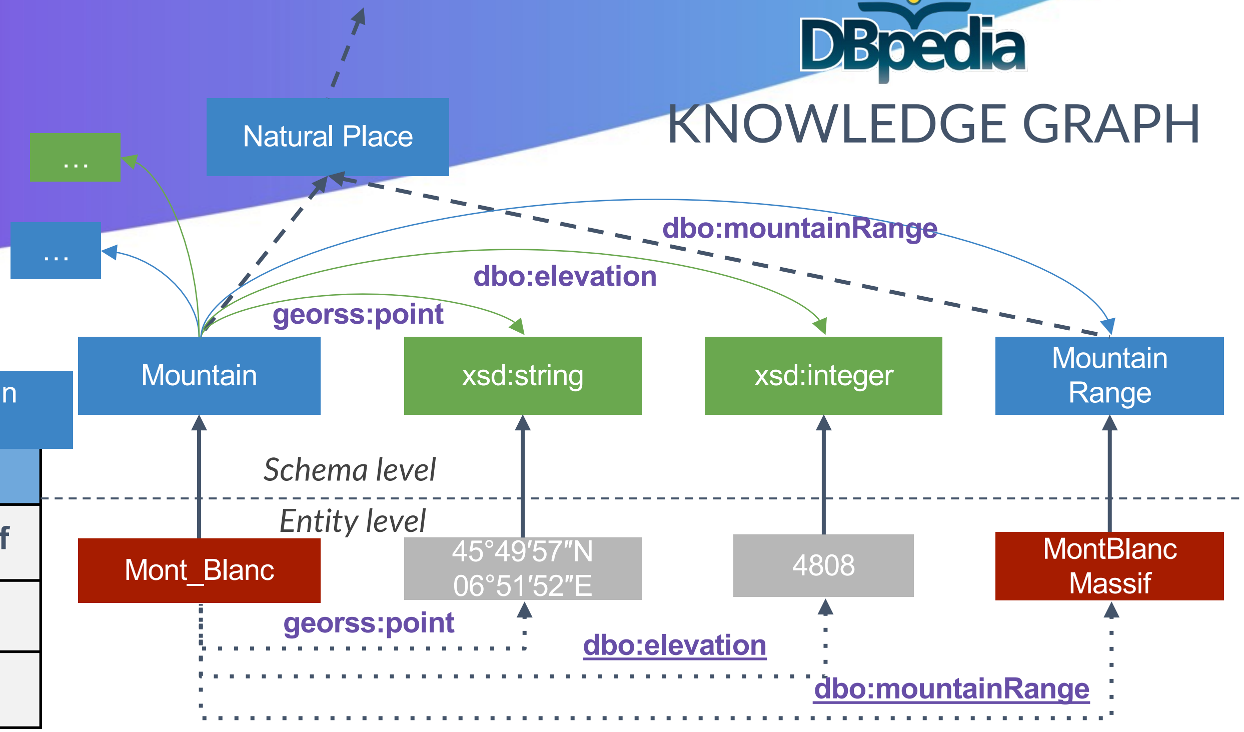
Tabular Data Annotation
… for KG completion

DBpedia
KNOWLEDGE GRAPH

Mountain | xsd:string | xsd:integer | Mountain Range

Name	Coordinates	Height	Range
Mont_Blanc	**45°49′57″N 06°51′52″E**	**4808**	MontBlanc Massif
Hohtälli	45°98′96″N 07°80′25″E	3275	Pennine Alps
Monte Cervino	45°58′35″N 07°39′31″E	4478	Pennine Alps

georss:point
dbo:elevation
dbo:mountainRange



Given

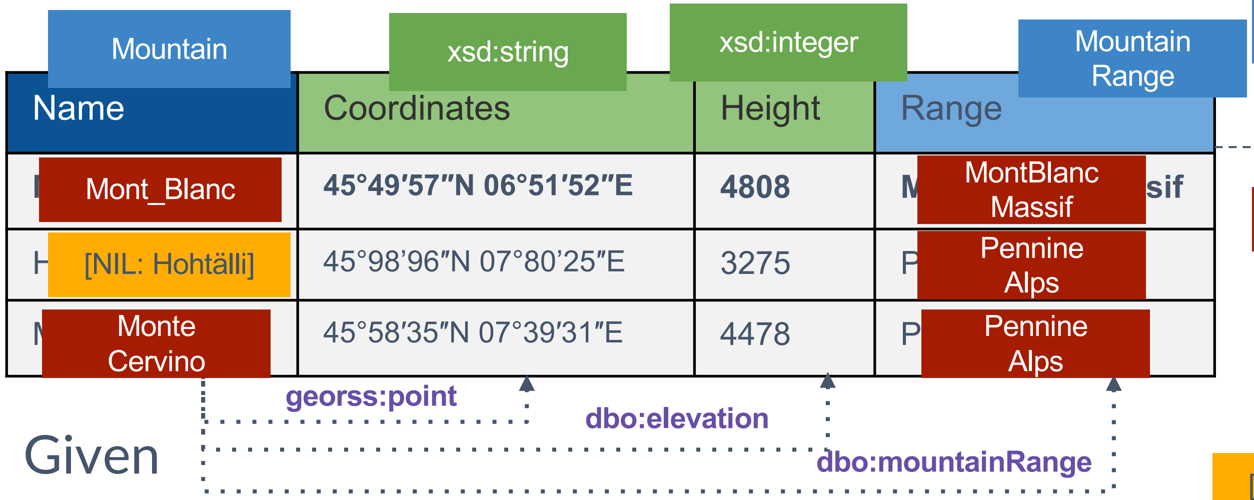
- a relational table T
- a **Knowledge Graph (entities + statements)** and an ontology (**types + predicates)**

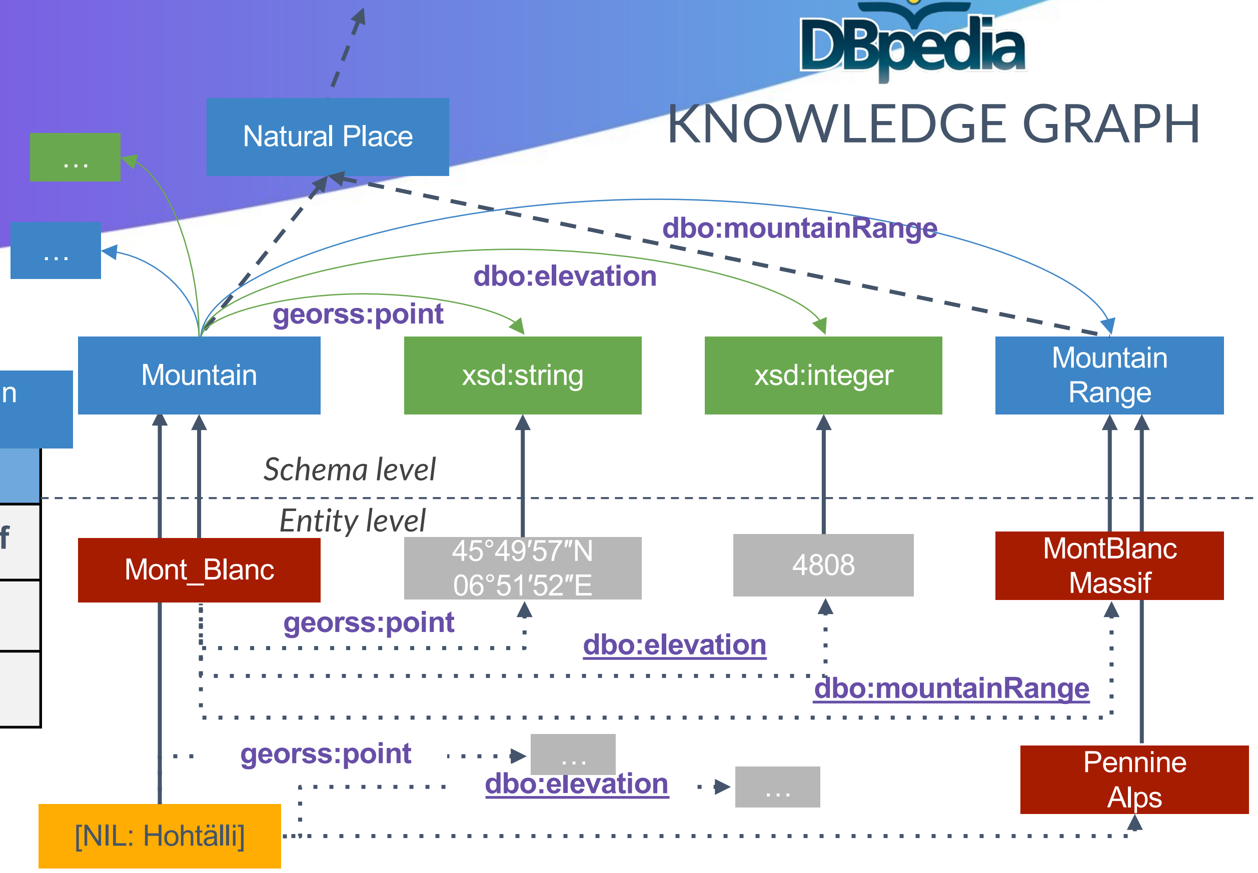
T is annotated when:

- each column is associated with one or more **KG-types** (CTA)
- each cell in “entity columns” is annotated with a **KG-entity** (CEA)
- some pair of columns is annotated with a binary **KG-predicate** (CPA)

Tabular Data Annotation
… with novel entities

KNOWLEDGE GRAPH





Given

- a relational table T
- a **Knowledge Graph (entities + statements)** and an ontology (**types + predicates)**

T is annotated when:

- each column is associated with one or more **KG-types** (CTA)
- each cell in "entity columns" is annotated with a **KG-entity** or with NIL (if not in the KG)
- some pair of columns is annotated with a binary **KG-predicate** (CPA)

SemTab Challenge

Table interpretation in research

- Check the challenge page: http://www.cs.ox.ac.uk/isg/challenges/sem-tab/

News (18/04/2019): Round 1 is now open in the AICrowd platform. SIRIUS sponsors the challenge prizes.

Semantic Web Challenge on Tabular Data to Knowledge Graph Matching

Tabular data in the form of CSV files is the common input format in a data analytics pipeline. However a lack of understanding of the semantic structure and meaning of the content may hinder the data analytics process. Thus gaining this semantic understanding will be very valuable for data integration, data cleaning, data mining, machine learning and knowledge discovery tasks. For example, understanding what the data is can help assess what sorts of transformation are appropriate on the data.

Tables on the Web may also be the source of highly valuable data. The addition of semantic information to Web tables may enhance a wide range of applications, such as web search, question answering, and knowledge base (KB) construction.

Tabular data to Knowledge Graph (KG) matching is the process of assigning semantic tags from Knowledge Graphs (e.g., Wikidata or DBpedia) to the elements of the table. This task however is often difficult in practice due to metadata (e.g., table and column names) being missing, incomplete or ambiguous.

This challenge aims at benchmarking systems dealing with the tabular data to KG matching problem, so as to facilitate their comparison on the same basis and the reproducibility of the results.

The **2019 edition** of this challenge will be collocated with the 18th International Semantic Web Conference and the 14th International Workshop on Ontology Matching.

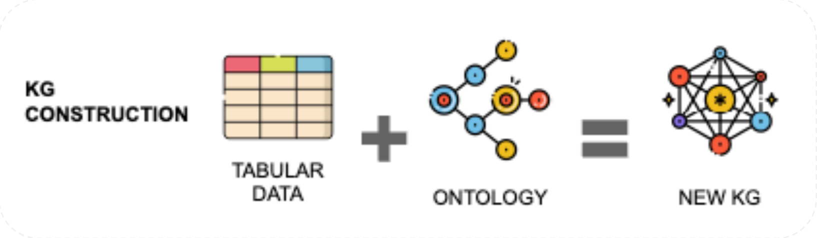
SIRIUS

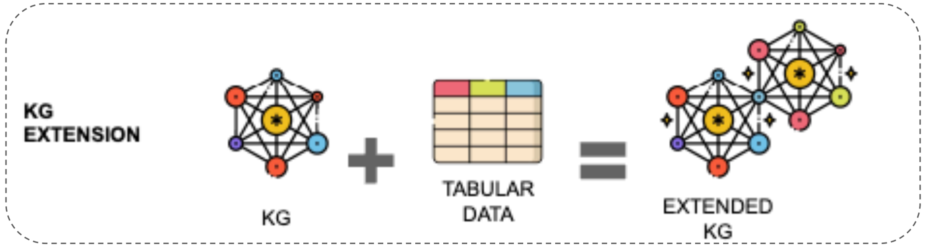
The Alan Turing Institute

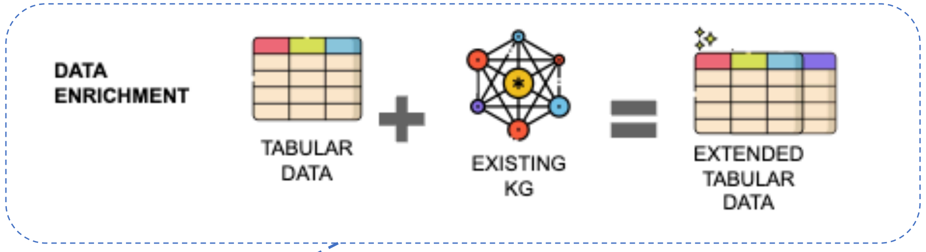
UNIVERSITY OF OXFORD DEPARTMENT OF COMPUTER SCIENCE

Auckland 2019 ISWC

Downstream Applications of Tabular Data Annotations







Keyword	#im	City	Region	ID (Geonames)	Latitude (Geonames)	Longitude (Geonames)	ID (Wikidata)	Area (Wikidata)	Temp (ECMWF)	Date
194906	64	Altenburg	Thuringia	2822542	50.98763	12.43684	Q1205	45.6 km²	18°	11/03/2017
517827	50	Inglostadt	Bavaria	2951839	48.76508	11.42372	Q980	133.35 km²	17°	12/03/2017
459143	42	Berlin	Berlin	2950157	52.52437	13.41053	Q648102	891.68 km²	17°	12/03/2017
891139	36	Munich	Bavaria	2951839	48.13743	11.57549	Q980	310.71 km²	19°	11/03/2017
459143										0/03/2017

Example of data enrichment by composing different individual linking and extension services

Data enrichment by the **link & extend** paradigm

What is This Table Describing?

Mount Everest	8,848	Himalayas	May 29, 1953
K-2 (Godwin Austin)	8,611	Karakoram	July 31, 1954
Kanchenjunga	8,597	Himalayas	May 25, 1955
Lhotse	8,511	Himalayas	May 18, 1956
Makalu I	8,481	Himalayas	May 15, 1955
Dhaulagiri I	8,167	Himalayas	May 13, 1960
Manaslu	8,156	Himalayas	May 9, 1956
Cho Uyo	8,153	Himalayas	Oct 19, 1954
Nanga Parbat	8,124	Himalayas	July 3, 1953
Annapurna I	8,078	Himalayas	June 3, 1950
Gasherbrum I	8,068	Karakoram	July 5, 1958
Broad Peak I	8,047	Karakoram	June 9, 1957
Gasherbrum II	8,034	Karakoram	July 7, 1956
Shisha Pangma (Gasainthan)	8,013	Himalayas	May 2, 1964
Gasherbrum III	7,952	Karakoram	Aug 11, 1975
Annapurna II	7,937	Himalayas	May 17, 1960
Gasherbrum IV	7,923	Karakoram	Aug 6, 1958

What is This Table Describing?

MOUNTAIN	HEIGHT IN METERS	RANGE	CONQUERED ON
Mount Everest	8,848	Himalayas	May 29, 1953
K-2 (Godwin Austin)	8,611	Karakoram	July 31, 1954
Kanchenjunga	8,597	Himalayas	May 25, 1955
Lhotse	8,511	Himalayas	May 18, 1956
Makalu I	8,481	Himalayas	May 15, 1955
Dhaulagiri I	8,167	Himalayas	May 13, 1960
Manaslu	8,156	Himalayas	May 9, 1956
Cho Uyo	8,153	Himalayas	Oct 19, 1954
Nanga Parbat	8,124	Himalayas	July 3, 1953
Annapurna I	8,078	Himalayas	June 3, 1950
Gasherbrum I	8,068	Karakoram	July 5, 1958
Broad Peak I	8,047	Karakoram	June 9, 1957
Gasherbrum II	8,034	Karakoram	July 7, 1956
Shisha Pangma (Gasainthan)	8,013	Himalayas	May 2, 1964
Gasherbrum III	7,952	Karakoram	Aug 11, 1975
Annapurna II	7,937	Himalayas	May 17, 1960
Gasherbrum IV	7,923	Karakoram	Aug 6, 1958

Semantic Table Annotation Challenges

Must consider and balance the different features of a table.
Several key challenges

- Disambiguation
- Homonym
- Matching
- NIL-mentions
- Literal and named-entity
- Missing context
- Amount of data
- Different domains

Semantic Table Annotation Approaches

A rough classification

- Unsupervised (unsup)
  - Based on matching algorithms and heuristics
- Supervised (sup)
  - Entirely based on machine learning, trained on some input data
  - Sub-category: LLM-based
    - Using LLMs for matching
    - Completely based on LLM
- Hybrid (hyb)
  - Combination of unsupervised and supervised

Semantic table annotation vs data enrichment

- CTA, CPA: schema matching
  - Main applications:
    - data annotation, KG construction and completion
  - Exploration and HITL: revision of all annotations is possible
  - Scalability: can use sampling, e.g., DuoDuo and TorchiTab
- CEA: entity linking
  - Main applications:
    - data annotation, KG construction and completion
    - **data augmentation (!)**
  - Exploration and HITL: revision of all annotations is NOT possible
  - Scalability: need scalable methods

Stateof the art

Algorithms

YEAR	AUTHOR	METHOD	PUBLICATION	CTA	CPA	CEA	CNEA	INDEX	CODE	LICENCE	TRIPLE STORE
2007	Hignette et al. [52]	Unsup	WISE	✓	✓	✗	✗	—	✗	—	Personal ontologies
2009	Hignette et al. [53]	Unsup	ESWC	✓	✓	✗	✗	—	✗	—	Personal ontologies
2009	Tao et al. [132]	Unsup	DKE	✗	✗	✗	✗	—	✗	—	Personal ontologies
2010	Limaye et al. [82]	Unsup	VLDB	✓	✓	✓	✗	—	✗	—	Yago
2010	Mulwad et al. [94]	Sup	ISWC	✓	✓	✓	✗	—	✗	—	Wikitology
2010	Syed et al. [126]	Unsup	WSC	✓	✓	✓	✗	Lucene for concepts	✗	—	Wikitology
2011	Mulwad et al. [95]	Sup	AAAI	✓	✓	✓	✗	—	✗	—	DBpedia,Freebase,WordNet,Yago
2011	Venetis et al. [135]	Unsup	VLDB	✓	✓	✗	✗	—	✗	—	Yago
2012	Goel et al. [46]	Sup	ICAI	✓	✓	✗	✗	—	✗	—	—
2012	Knoblock et al. [74]	Sup	ESWC	✓	✓	✗	✗	—	✓	Apache 2.0	Personal ontologies
2012	Pimplikar et al. [107]	Unsup	VLDB	✗	✗	✗	✗	—	✗	—	—
2012	Wang et al. [137]	Unsup	ER	✓	✓	✓	✗	—	✗	—	—
2013	Buche et al. [17]	Unsup	IEEE	✓	✓	✗	✗	—	✗	—	—
2013	Cruz et al. [31]	Sup	SIGSPATIAL	✗	✓	✓	✗	—	✗	—	—
2013	Deng et al. [36]	Unsup	VLDB	✓	✗	✗	✗	—	✗	—	DBpedia,Freebase,Yago
2013	Ermilov et al. [42]	Unsup	I-SEMANTICS	✗	✗	✗	✗	—	✗	—	—
2013	Mulwad et al. [93]	Sup	ISWC	✓	✓	✓	✗	—	✗	—	DBpedia,Yago,Wikitology
2013	Munoz et al. [92]	Unsup	LD4IE	✗	✓	✓	✗	—	✗	—	DBpedia
2013	Quercini et al. [109]	Unsup	EDBT	✗	✗	✓	✗		✗	—	DBpedia
2013	Zhang et al. [145]	Unsup	SIGMOD	✓	✗	✗	✗	—	✗	—	—
2013	Zwicklbauer et al. [152]	Unsup	ISWC	✓	✗	✓	✗	—	✗	—	DBpedia
2014	Sekhavat et al. [117]	Unsup	LDOW	✗	✓	✗	✗	—	✗	—	Yago
2014	Taheriyan et al. [127]	Unsup	IEEE	✓	✓	✗	✗	—	✗	—	—
2015	Bhagavatula et al. [14]	Sup	ISWC	✗	✗	✓	✗	—	✗	CCA 4.0	Yago
2015	Ramnandan et al. [110]	Sup	ESWC	✓	✗	✗	✗	training data with Lucene, not KG data	✓	Apache 2.0	—
2015	Ritze et al. [113]	Unsup	WIMS	✓	✓	✓	✗	—	✓	Apache 2.0	DBpedia
2016	Ermilov et al. [43]	Unsup	EKAW	✓	✓	✗	✗	—	✓	GPL 3.0	DBpedia
2016	Neumaier et al. [96]	Sup	ISWC	✗	✗	✗	✗	—	✓	Apache 2.0	DBpedia
2016	Pham et al. [105]	Sup	ISWC	✓	✗	✗	✗	—	✓	Apache 2.0	—
2016	Taheriyan et al. [129]	Sup	JOWS	✓	✓	✗	✗	—	✓	Apache 2.0	CIDOC-CRM,EDM
2016	Taheriyan et al. [128]	Sup	ISWC	✗	✓	✗	✗	—	✓	Apache 2.0	CIDOC-CRM
2017	Efthymiou et al. [40]	Hybrid	ISWC	✗	✓	✓	✗	—	✗	—	—
2017	Ell et al. [41]	Unsup	LD4IE	✗	✗	✓	✗	Labels + literals	✗	Apache 2.0	DBpedia
2017	Zhang et al. [149]	Unsup	JOWS	✓	✓	✓	✗	—	✓	Apache 2.0	Freebase
2018	Kacprzak et al. [65]	Unsup	EKAW	✓	✗	✓	✗	—	✓	MIT	DBpedia
2018	Luo et al. [85]	Sup	AAAI	✗	✗	✗	✗	—	✗	—	Wikipedia
2018	Zhang et al. [146]	Unsup	WWW	✗	✗	✓	✗	—	✓	—	—
2019	Chabot et al. [20]	Unsup	SemTab	✓	✓	✓	✗	—	✗	Orange	DBpedia
2019	Chen et al. [21]	Hybrid	AAAI	✓	✗	✓	✗	—	✓	Apache 2.0	DBpedia
2019	Chen et al. [21]	Unsup	IJCAI	✓	✓	✗	✗	—	✓	Apache 2.0	DBpedia
2019	Cremaschi et al. [28]	Unsup	SemTab	✓	✓	✓	✗	—	✓	Apache 2.0	DBpedia
2019	Hulsebos et al. [56]	Sup	SIGKDD	✓	✗	✗	✗	—	✓	MIT	DBpedia
2019	Kruit et al. [78]	Hybrid	ISWC	✓	✓	✓	✗	—	✓	MIT	DBpedia,Wikidata
2019	Morikawa et al. [91]	Unsup	SemTab	✓	✓	✓	✗	Elasticsearch	✗	—	DBpedia
2019	Nguyen et al. [97]	Unsup	SemTab	✓	✓	✓	✗	—	✗	—	DBpedia
2019	Oliveira et al. [101]	Unsup	SemTab	✓	✓	✓	✗	ArangoDB + Elasticsearch	✓	—	DBpedia
2019	Steenwinckel et al. [122]	Unsup	SemTab	✓	✓	✓	✗	—	✗	—	DBpedia

State of the art

Algorithms

YEAR	AUTHOR	METHOD	PUBLICATION	CTA	CPA	CEA	CNEA	INDEX	CODE	LICENCE	TRIPLE STORE
2019	Takeoka et al. [130]	Sup	AAAI	✓	✗	✗	✗	—	✗	—	WordNet
2019	Thawani et al. [133]	Unsup	SemTab	✓	✓	✓	✗	Elasticsearch	✓	MIT	—
2019	Zhang et al. [144]	Sup	VLDB	✓	✗	✗	✗	—	✓	Apache 2.0	DBpedia
2020	Abdelmageed et al. [1]	Unsup	SemTab	✓	✓	✓	✗	—	✓	MIT	Wikidata
2020	Azzi et al. [11]	Unsup	SemTab	✓	✗	✓	✗	—	✗	—	Wikidata
2020	Baazouzi et al. [13]	Unsup	SemTab	✓	✗	✗	✗	—	✗	—	Wikidata
2020	Chen et al. [23]	Unsup	SemTab	✓	✓	✓	✗	Elasticsearch	✗	—	Wikidata
2020	Cremaschi et al. [30]	Unsup	FGCS	✓	✓	✓	✗	—	✓	Apache 2.0	DBpedia
2020	Cremaschi et al. [27]	Unsup	SemTab	✓	✓	✓	✗	LamAPI	✓	Apache 2.0	DBpedia,Wikidata
2020	Eslahi et al. [44]	Unsup	SDS	✓	✗	✓	✗	—	✓	—	Wikidata
2020	Guo et al. [48]	Sup	WISA	✓	✗	✗	✗	—	✗	—	—
2020	Huynh et al. [59]	Hybrid	SemTab	✓	✓	✓	✗	Spark dataframes	✗	—	Wikidata
2020	Khurana et al. [69]	Sup	CIKM	✓	✓	✗	✗	—	✗	—	—
2020	Kim et al. [71]	Unsup	SemTab	✓	✓	✓	✗	—	✗	—	Wikidata
2020	Li et al. [81]	Sup	VLDB	✗	✗	✓	✗	—	✓	Apache 2.0	—
2020	Nguyen et al. [99]	Unsup	SemTab	✓	✓	✓	✗	HashTable + Sparse Matrix	✗	—	Wikidata
2020	Shigapov et al. [118]	Unsup	SemTab	✓	✓	✓	✗	SeerX metasearch API	✓	MIT	Wikidata
2020	Tyagi et al. [134]	Unsup	SemTab	✓	✗	✓	✗	—	✓	—	Wikidata
2020	Yumusak et al. [143]	Unsup	SemTab	✓	✓	✗	✗	—	✓	—	Wikidata
2020	Zhang et al. [148]	Sup	WWW	✗	✓	✓	✓	—	✗	CCA 4.0	DBpedia
2021	Abdelmageed et al. [3]	Unsup	SemTab	✓	✓	✓	✗	—	✓	Apache 2.0	DBpedia,Wikidata
2021	Abdelmageed et al. [2]	Unsup	KGC	✓	✓	✓	✗	—	✓	Apache 2.0	DBpedia,Wikidata
2021	Avogadro et al. [9]	Unsup	SemTab	✓	✓	✓	✗	LamAPI	✓	Apache 2.0	DBpedia,Wikidata
2021	Baazouzi et al. [12]	Unsup	SemTab	✓	✓	✓	✗	—	✗	—	Wikidata
2021	Heist et al. [51]	Hybrid	WWW	✗	✗	✗	✗	—	✓	GPL 3.0	CaliGraph,DBpedia,Yago
2021	Huynh et al. [58]	Hybrid	SemTab	✓	✓	✓	✗	Elasticsearch	✗	Orange	DBpedia,Wikidata
2021	Nguyen et al. [100]	Unsup	SemTab	✓	✓	✓	✗	Custom BM25	✗	MIT	DBpedia,Wikidata
2021	Steenwinckel et al. [121]	Hybrid	SemTab	✓	✓	✓	✗	—	✓	Imec license	Wikidata
2021	Wang et al. [136]	Sup	WWW	✓	✓	✗	✗	—	✗	—	—
2021	Yang et al. [142]	Sup	SemTab	✓	✗	✓	✗	—	✗	—	Wikidata
2021	Zhou et al. [150]	Sup	CIKM	✓	✗	✗	✗	—	✗	—	—
2022	Abdelmageed et al. [4]	Unsup	KGC	✓	✓	✓	✗	—	✓	Apache 2.0	DBpedia,Wikidata
2022	Chen et al. [24]	Unsup	JWS	✓	✓	✓	✗	Elasticsearch	✓	MIT	DBpedia,Wikidata
2022	Cremaschi et al. [29]	Unsup	SemTab	✓	✓	✓	✗	LamAPI	✓	Apache 2.0	DBpedia,Wikidata
2022	Deng et al. [37]	Sup	SIGMOD	✓	✓	✓	✗	—	✓	Apache 2.0	—
2022	Gottschalk et al. [47]	Sup	SWJ	✓	✓	✗	✗	—	✓	MIT	—
2022	Huynh et al. [57]	Hybrid	SemTab	✓	✓	✓	✗	Elasticsearch	✗	Orange	DBpedia,Wikidata
2022	Liu et al. [84]	Hybrid	ISWC	✗	✗	✓	✗	—	✓	Orange	Wikidata
2022	Suhara et al. [124]	Sup	SIGMOD	✓	✓	✗	✗	—	✓	Apache 2.0	Freebase,DBpedia

s-elBat ⇒ TURL

23-24 additions: **Alligator** (s-elBat with ML); **UNICORN**; **TableLlama**

DuoDuo: nice approach to schema matching against Schema.org

Semantic Table Annotation Approaches vs Data Enrichment

A rough classification

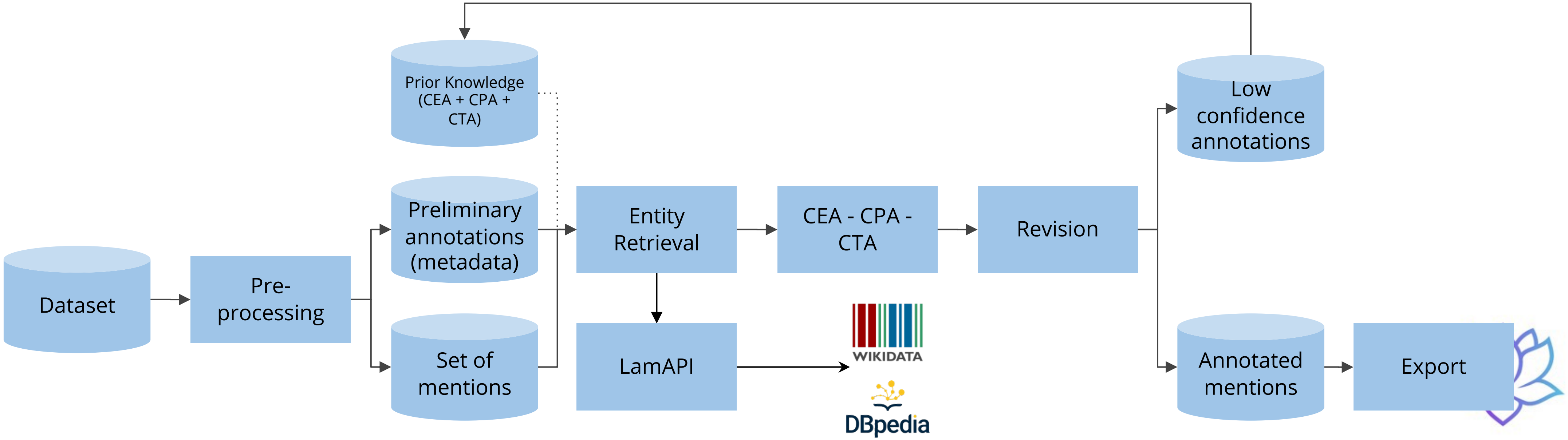
- Unsupervised (unsup)
  - Based on matching algorithms and heuristics
- Supervised (sup)
  - Entirely based on machine learning, trained on some input data
  - Sub-category: LLM-based
    - Using LLMs for matching
    - Completely based on LLM
- Hybrid (hyb)
  - Combination of unsupervised and supervised

Semantic table annotation vs data enrichment

- CTA, CPA: schema matching
  - Main applications:
    - data annotation, KG construction and completion
  - Exploration and HITL: revision of all annotations is possible
  - Scalability: can use sampling, e.g., DuoDuo and TorchiTab
- CEA: entity linking
  - Main applications:
    - data annotation, KG construction and completion
    - **data augmentation (!)**
  - Exploration and HITL: revision of all annotations is NOT possible
  - Scalability: need scalable methods

s-elBat: an heuristic approach

1. Preprocessing and Data preparation
2. Entity Retrieval
3. Cell Entity Annotation (CEA)
4. Cell Property Annotation (CPA)
5. Cell Type Annotation (CTA)
6. Revision
7. Export



Candidate generation and disambiguation in unsupervised approaches

CTA / Entity linking
- Candidate generation
  - Queries
  - Legacy lookup service
  - Custom search
- Disambiguation
  - Similarity
  - Use of CTA and CPA results

R: recall
U: dealing with updates
S: scalability

Approach	Candidate Generation	Entity Disambiguation
Limaye 2010 [82]	YAGO catalog	similarity
Syed 2010 [126]	Wikitology	CTA
Wang 2012 [137]	pattern matching	features
Munoz 2013 [92]	-	redirects
Ritze 2015 [113]	DBpedia lookup service	CTA
Ell 2017 [41]	custom index	features
Zhang 2017 [149]	external lookup	similarity
Zhang 2018 [146]	SPARQL	entity embedding
Cremaschi 2019 [28]	SPARQL	similarity
Morikawa 2019 [91]	SPARQL, Elasticsearch	CTA
Nguyen 2019 [97]	DBpedia lookup service, DBpedia endpoint, Wikipedia API, Wikidata API	CTA
Oliveira 2019 [101]	Elasticsearch	similarity
Steenwinckel 2019 [122]	DBpedia lookup service, DBpedia urls, DBpedia Spotlight	similarity
Thawani 2019 [133]	Wikidata API, Elasticsearch	similairty, CTA, ML
Abdelmageed 2020 [1]	Wikidata lookup service	CTA, CPA
Azzi 2020 [11]	Wikidata API	CTA
Chen 2020 [23]	Mediawiki API, Elasticsearch	CTA, CPA
Cremaschi 2020-1 [30]	SPARQL	similarity
Cremaschi 2020-2 [27]	Elasticsearch	CTA, CPA
Kim 2020 [71]	SPARQL	features
Nguyen 2020 [99]	custom index	CPA
Shigapov 2020 [118]	SearX, SPARQL, Wikibooks, Wikipedia API, Wikidata API	similarity
Tyagi 2020 [134]	Wikidata lookup service, DBpedia lookup service	similarity
Abdelmageed 2021-1 [3]	Wikidata lookup service, SPARQL	similarity
Abdelmageed 2021-2 [2]	Wikidata lookup service, SPARQL	similarity
Avogadro 2021 [9]	custom index	similarity, CTA, CPA
Baazouzi 2021 [12]	SPARQL	CTA
Nguyen 2021 [100]	custom index	CPA
Abdelmageed 2022 [4]	SPARQL, Wikidata lookup service	similarity
Chen 2022 [24]	Elasticsearch	similarity, CTA, CPA
Cremaschi 2022 [29]	Elasticsearch	similarity, CPA, CTA

State of the art

Entity Linking

Candidate generation and disambiguation in unsupervised approaches

CTA / Entity linking
- Candidate generation
  - Queries
    - R-, U+, S-
  - Legacy lookup service
    - R+, U+, S-
  - Custom search
    - R+, U-, S+
- Disambiguation
  - Similarity
  - Use of CTA and CPA results

R: recall
U: dealing with updates
S: scalability

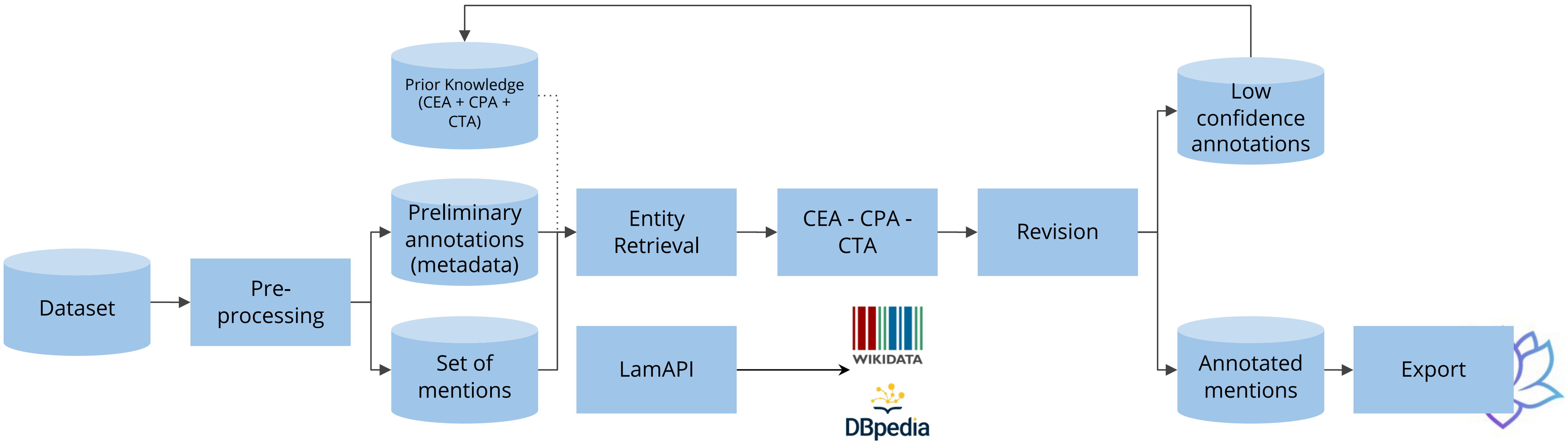
Approach	Candidate Generation	Entity Disambiguation
Limaye 2010 [82]	YAGO catalog	similarity
Syed 2010 [126]	Wikitology	CTA
Wang 2012 [137]	pattern matching	features
Munoz 2013 [92]	-	redirects
Ritze 2015 [113]	DBpedia lookup service	CTA
Ell 2017 [41]	custom index	features
Zhang 2017 [149]	external lookup	similarity
Zhang 2018 [146]	SPARQL	entity embedding
Cremaschi 2019 [28]	SPARQL	similarity
Morikawa 2019 [91]	SPARQL, Elasticsearch	CTA
Nguyen 2019 [97]	DBpedia lookup service, DBpedia endpoint, Wikipedia API, Wikidata API	CTA
Oliveira 2019 [101]	Elasticsearch	similarity
Steenwinckel 2019 [122]	DBpedia lookup service, DBpedia urls, DBpedia Spotlight	similarity
Thawani 2019 [133]	Wikidata API, Elasticsearch	similairty, CTA, ML
Abdelmageed 2020 [1]	Wikidata lookup service	CTA, CPA
Azzi 2020 [11]	Wikidata API	CTA
Chen 2020 [23]	Mediawiki API, Elasticsearch	CTA, CPA
Cremaschi 2020-1 [30]	SPARQL	similarity
Cremaschi 2020-2 [27]	Elasticsearch	CTA, CPA
Kim 2020 [71]	SPARQL	features
Nguyen 2020 [99]	custom index	CPA
Shigapov 2020 [118]	SearX, SPARQL, Wikibooks, Wikipedia API, Wikidata API	similarity
Tyagi 2020 [134]	Wikidata lookup service, DBpedia lookup service	similarity
Abdelmageed 2021-1 [3]	Wikidata lookup service, SPARQL	similarity
Abdelmageed 2021-2 [2]	Wikidata lookup service, SPARQL	similarity
Avogadro 2021 [9]	custom index	similarity, CTA, CPA
Baazouzi 2021 [12]	SPARQL	CTA
Nguyen 2021 [100]	custom index	CPA
Abdelmageed 2022 [4]	SPARQL, Wikidata lookup service	similarity
Chen 2022 [24]	Elasticsearch	similarity, CTA, CPA
Cremaschi 2022 [29]	Elasticsearch	similarity, CPA, CTA

State of the art

Entity Linking

s-elBat: an heuristic approach

1. Preprocessing and Data preparation
2. **Entity Retrieval**
3. Cell Entity Annotation (CEA)
4. Cell Property Annotation (CPA)
5. Cell Type Annotation (CTA)
6. Revision
7. Export



Entity Retrieval with LamAPI [Avogadro et al. 2022]

"Kobe Bryant" → **Query**

- Q25369 - Kobe Bryant [basketball player]
- Q97396439 - Kobe Bryant 1978-2020 [Wikinews article]
- Q31391 - Kobe Bryant MVP [most valuable player award]
- ....

"Pariss" → **Query** n-grams

- Q139368 - Zeks Pariss [ice hockey player]
- Q90 - Departement de Paris [capital]
- Q164 - parisukat [geometric shape]
- ....

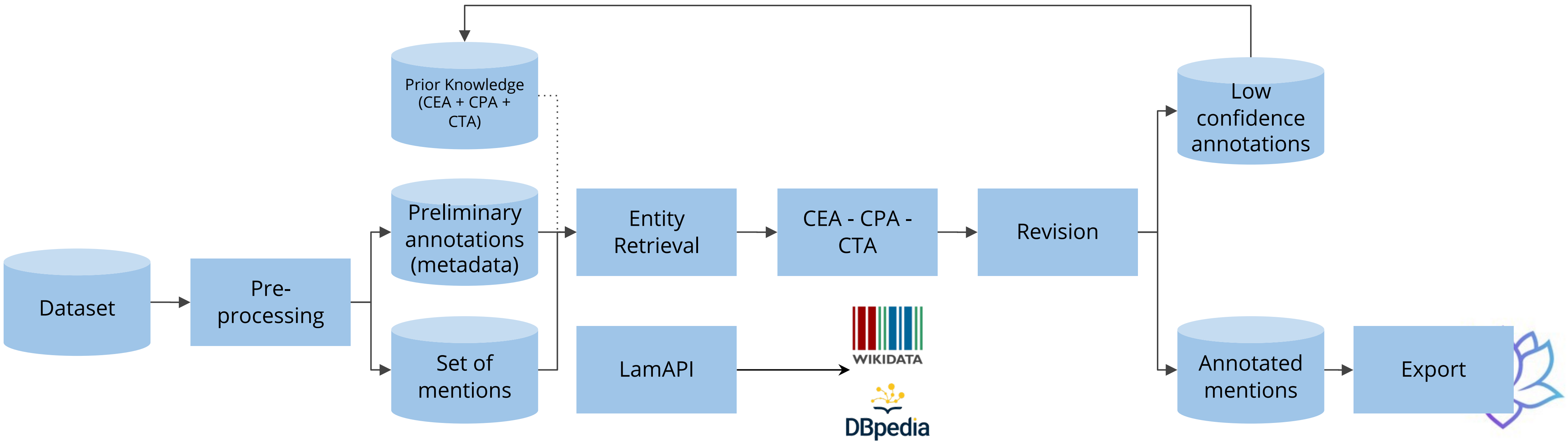
"Colorado" (U.S. state) → **Query** with type

- Q1261 - Colorado [U.S. state]
- Q1265 - Colorado [river]
- Q3142 - Colorado [color]
- ....

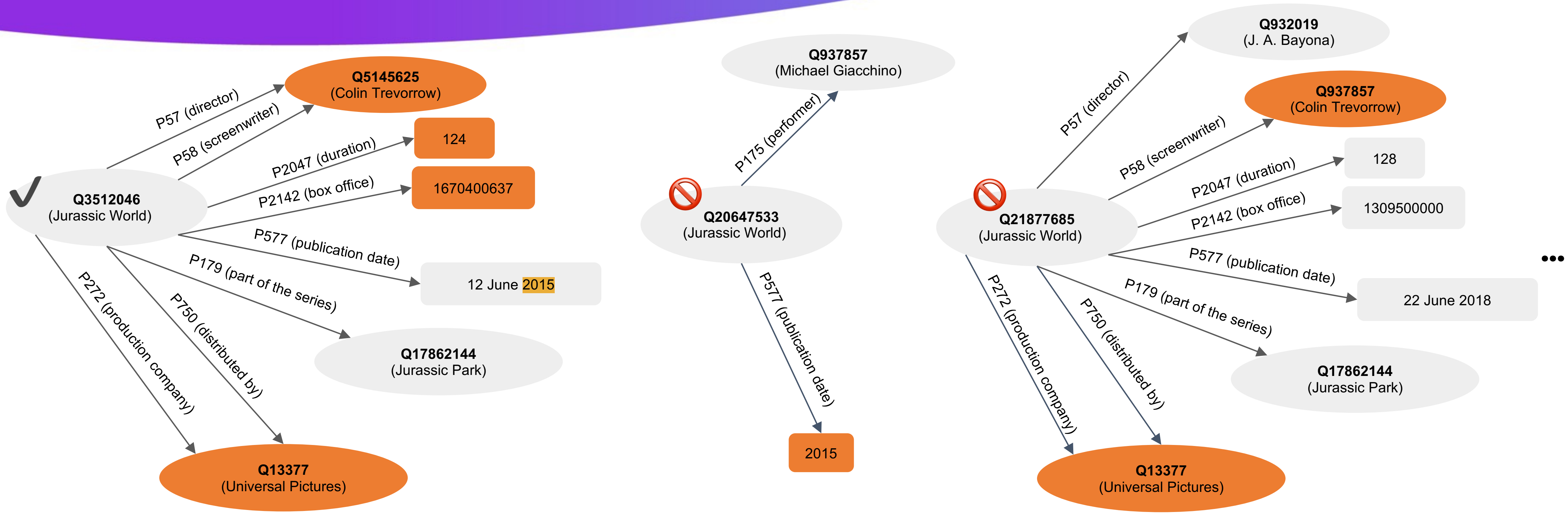
Avogadro, R., Cremaschi, M., D'Adda, F., De Paoli, F., & Palmonari, M. (2022). LamAPI: a comprehensive tool for string-based entity retrieval with type-base filters. In OM@ ISWC (pp. 25-36).

s-elBat: an heuristic approach

1. Preprocessing and Data preparation
2. Entity Retrieval
3. **Cell Entity Annotation (CEA)**
4. Cell Property Annotation (CPA)
5. Cell Type Annotation (CTA)
6. Revision
7. Export



Challenges: Entity Disambiguation and Ranking in Tables



title	director	release year	domestic distributor	length in min	worldwide gross
jurassic world	colin trevorrow	2015	universal pictures	124	1670400637

s-elBat '22

- SemTab22
  - [Cremaschi et al. 2022]
  - Heuristics transformation of features into unbound ranking scores

name	range	description
ed	[0, 1]	Ed is a measure of the similarity between two strings, calculated by determining the minimum number of single-character edits required to transform one into the other. It can be used to evaluate the similarity between the mention of an entity and its name in a knowledge graph
jaccard	[0, 1]	The Jaccard distance is a measure of the similarity between two strings, calculated by dividing the number of matching tokens in the two strings by the total number of unique tokens. It can be used to evaluate the similarity between the mention of an entity and its name in a knowledge graph
jaccardNgram	[0, 1]	JaccardNgram is a measure of the similarity between two strings, calculated by dividing the number of matching n-grams in the two strings by the total number of unique n-grams. It can be used to evaluate the similarity between the mention of an entity and its name in a knowledge graph
p_subj_ne	[0, ∞)	p_subj_ne is a score that reflects the relationship between a current candidate for a Name Entity (NE) cell and other candidate NE cells on the same row in a table
p_subj_lit	[0, ∞)	p_subj_lit is a score that reflects the similarity between the literal values associated with a current candidate for a subject cell and the literal values on the same row of the table
p_obj_ne	[0, ∞)	p_obj_ne is a score that reflects the relationship between a current candidate for a Name Entity (NE) cell and other candidate NE cells on the same row in a table, where the current candidate is in a relationship as an object with the other candidate NE cells
desc	[0, 1]	desc is a score that reflects the similarity between the content of a row in a table and the description of a current candidate in a knowledge graph, using Jaccard similarity based on tokens to compare the two
descNgram	[0, 1]	descNgram is a score that reflects the similarity between the content of a row in a table and the description of a current candidate in a knowledge graph, using Jaccard similarity based on 3-grams (also known as trigrams) to compare the two
cta	[0, ∞)	The score based on the types collected during the CEA phase, so for each collected type sum the frequencies only for each type that belongs to the candidate
ctaMax	[0, ∞)	The score based on the types collected during the CEA phase, so for each collected predicate extract only the max type that belongs to the candidate
cpa	[0, ∞)	The score based on the predicates collected during the CEA phase, so for each collected predicate sum the frequencies only for each predicate that belongs to the candidate
cpaMax	[0, ∞)	The score based on the predicates collected during the CEA phase, so for each collected predicate extract only the max predicate that belongs to the candidate

Mention vs labels

Row vs properties

Row vs description

Predicates and types hits

s-elBat >> Alligator '23

- SemTab22
  - [Cremaschi et al. 2022]
  - Heuristics transformation of features into unbound ranking scores
- Improvements: CEA
  - NN-based transformation into a bounded confidence score $\omega \in [0,1]$
  - NIL prediction with threshold

name	range	description
ed	[0, 1]	Ed is a measure of the similarity between two strings, calculated by determining the minimum number of single-character edits required to transform one into the other. It can be used to evaluate the similarity between the mention of an entity and its name in a knowledge graph
jaccard	[0, 1]	The Jaccard distance is a measure of the similarity between two strings, calculated by dividing the number of matching tokens in the two strings by the total number of unique tokens. It can be used to evaluate the similarity between the mention of an entity and its name in a knowledge graph
jaccardNgram	[0, 1]	JaccardNgram is a measure of the similarity between two strings, calculated by dividing the number of matching n-grams in the two strings by the total number of unique n-grams. It can be used to evaluate the similarity between the mention of an entity and its name in a knowledge graph
p_subj_ne	[0, ∞)	p_subj_ne is a score that reflects the relationship between a current candidate for a Name Entity (NE) cell and other candidate NE cells on the same row in a table
p_subj_lit	[0, ∞)	p_subj_lit is a score that reflects the similarity between the literal values associated with a current candidate for a subject cell and the literal values of the same row of the table
p_obj_ne	[0, ∞)	p_obj_ne is a score that reflects the relationship between a current candidate for a Name Entity (NE) cell and other candidate NE cells of the same row in a table, where the current candidate is in a relationship as an object with the other candidate NE cells
desc	[0, 1]	desc is a score that reflects the similarity between the content of a row in a table and the description of a current candidate in a knowledge graph, using Jaccard similarity based on tokens to compare the two
descNgram	[0, 1]	descNgram is a score that reflects the similarity between the content of a row in a table and the description of a current candidate in a knowledge graph, using Jaccard similarity based on 3-grams (also known as trigrams) to compare the two
cta	[0, ∞)	The score based on the types collected during the CEA phase, so for each collected type sum the frequencies only for each type that belongs to the candidate
ctaMax	[0, ∞)	The score based on the types collected during the CEA phase, so for each collected predicate extract only the max type that belongs to the candidate
cpa	[0, ∞)	The score based on the predicates collected during the CEA phase, so for each collected predicate sum the frequencies only for each predicate that belongs to the candidate
cpaMax	[0, ∞)	The score based on the predicates collected during the CEA phase, so for each collected predicate extract only the max predicate that belongs to the candidate

Mention vs labels

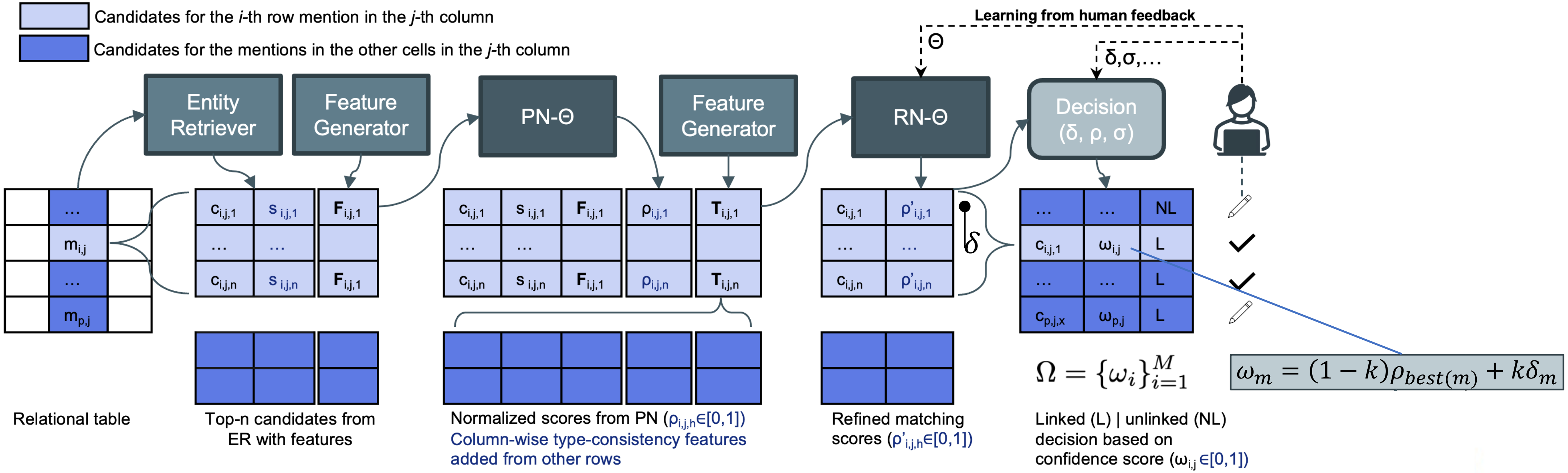
Row vs properties

Row vs description

Predicates and types hits

Alligator – ML-based linking with HITL

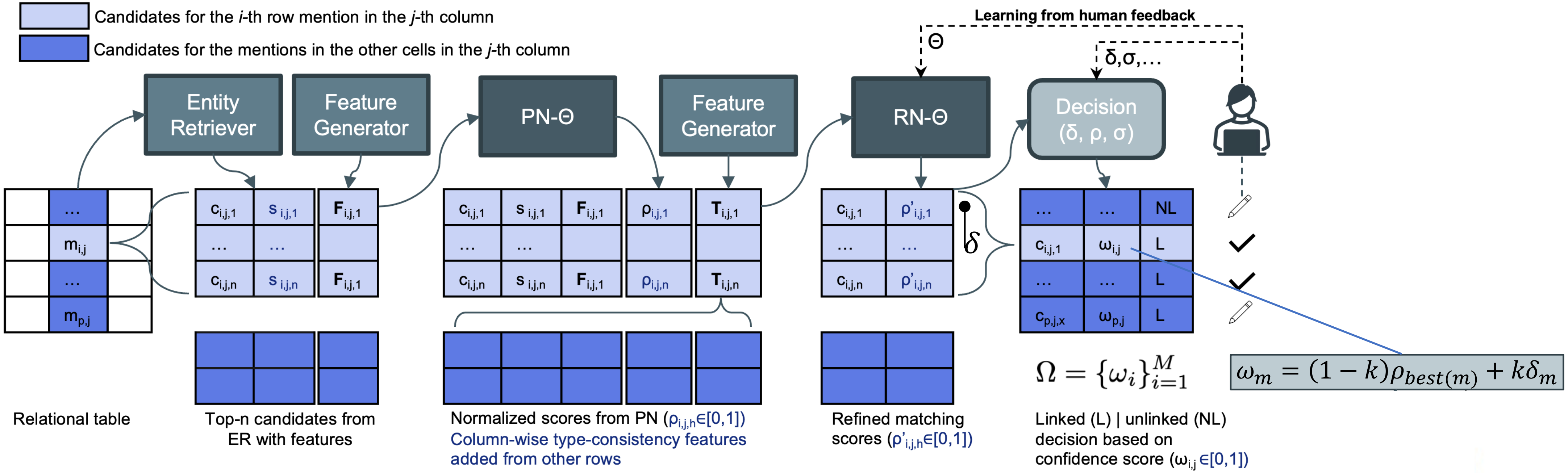
- Revised linking pipeline
  - Feature-based ML for entity linking with limited parameters
  - HITL approach to revise uncertain outcomes



- Confidence-based revision:
  - Use the confidence score to order links to revise
    - E.g., mentions with lower confidence first, i.e., order all mentions *m* by increasing ω_m
    - E.g., mentions that are more uncertain first, i.e., order all mentions *m* by distance of ω_m from the threshold

Alligator – ML-based linking with HITL

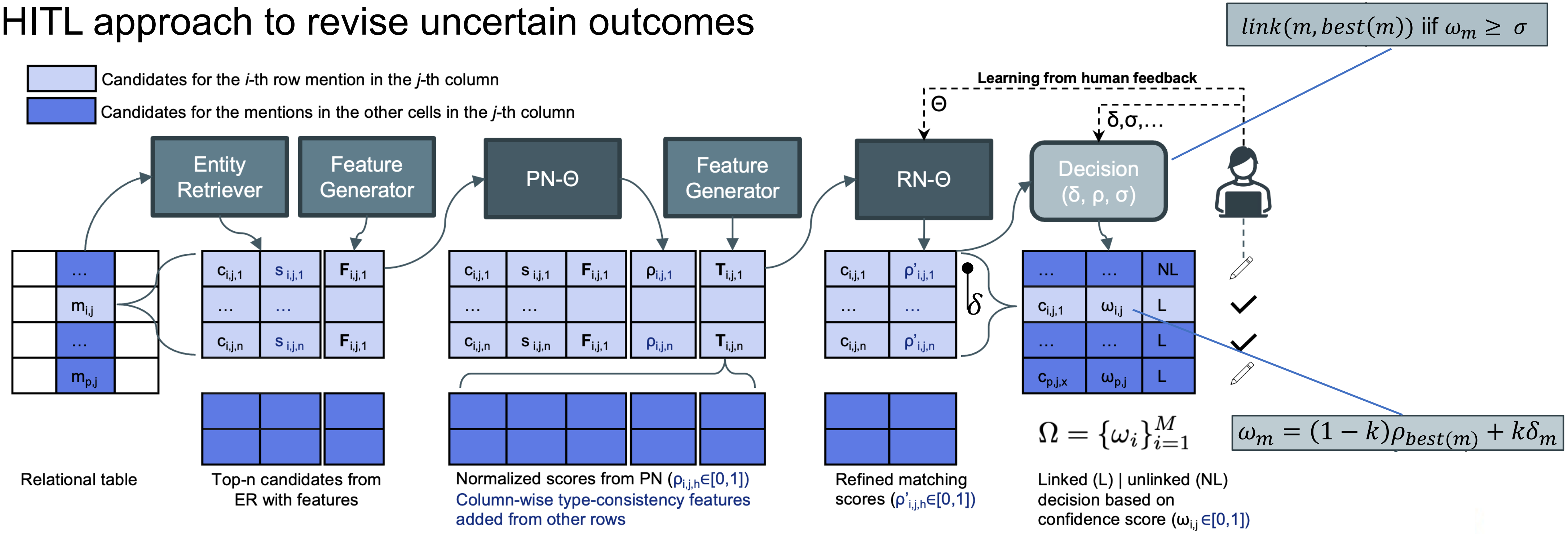
- Revised linking pipeline
  - Feature-based ML for entity linking with limited parameters
  - HITL approach to revise uncertain outcomes



- Confidence-based revision:
  - Use the confidence score to order links to revise
    - E.g., mentions with lower confidence first, i.e., order all mentions m by increasing ω_m
    - E.g., mentions that are more uncertain first, i.e., order all mentions m by distance of ω_m from the threshold

Alligator – ML-based linking with HITL

- Revised linking pipeline
  - Feature-based ML for entity linking with limited parameters
  - HITL approach to revise uncertain outcomes



- Confidence-based revision:
  - Use the confidence score to order links to revise
    - E.g., mentions with lower confidence first, i.e., order all mentions *m* by increasing ω_m
    - E.g., mentions that are more uncertain first, i.e., order all mentions *m* by distance of ω_m from the threshold

Alligator: Example

title	director	release year	domestic distributor	length in min	film budget	worldwide gross	imdb rating
Avengers: Endgame	Anthony e Joe Russo	2019	Walt Disney	181	356,000,000	2,797,800,564	8.5
Joker	Todd Phillips	2019	Warner Bros.	122	55,000,000	1,071,030,470	8.6
Aquaman	James Wan	2018	Warner Bros.	143	160,000,000	1,148,161,807	7
Venom	Ruben Fleischer	2018	Sony Pictures	112	100,000,000	856,085,151	6.7

Cell/mention	id	name	description	types	ρ	ρ'	δ	ω	correct link
james wan	Q108047434	james wan	malaysian-australian director, producer, screenwriter, and comic book writer	[{'id': 'Q5', 'name': 'human'}, {'id': 'Q7042855', 'name': 'film editor'}, {'id': 'Q2526255', 'name': 'film director'}, {'id': 'Q28389', 'name': 'screenwriter'}, {'id': 'Q36180', 'name': 'writer'}, {'id': 'Q3282637', 'name': 'film producer'}, {'id': 'Q1053574', 'name': 'executive producer'}, {'id': 'Q69423232', 'name': 'film screenwriter'}]	0.242	0.997	0.993	0.995	Q108047434

- Example for *uncertain* to *linked* using type-wise featueres
  - ρ: 0.242
    - Without column-wise context
  - ρ' : 0.997
    - With column-wise context

Column-wise types frequencies

- **Human** (Q5) : 96%
- **Film Director** (Q2526255) : 93%
- **Screenwriter** (Q28389) : 79%
- **Film Producer** (Q3282637) : 79%
- **Actor** (Q33999) : 46%
- **Director** (Q3455803) : 36%
- **Film Actor** (Q10800557) : 29%
- **TV Director** (Q2059704) : 25%
- ...

Some Experimental Insights

Moderately out-of-domain test settings
- trainaing datasets: all except the test set

[SemTab2021]

Test Dataset	Retrieval with indexing	PN ranking	PN + RN ranking with types
	F1	F1	F1
Round_T2D	0.82	0.83	**0.86**
Round3	0.72	0.73	**0.76**
Round4	0.83	0.90	**0.91**
2T-2020	0.62	0.86	**0.89**
HardTableR2	0.90	0.91	**0.93**
HardTableR3	0.52	0.54	**0.62**

[Avogadro et. al. 2023]

With HITL revision of most uncertain cells (simulated feedback)

SemTab Top Scorer	10%	20%
F1	F1	F1
0.90	**0.91**	0.95
0.97	0.82	0.87
0.99	0.95	0.97
0.90	**0.93**	0.94
0.98	**0.98**	0.99
0.97	0.68	0.75

MTAB and Dagobah

With improved features

PN + RN ranking with types
F1
0.89
0.81
0.94
0.88
0.97
0.97

[R. Avogadro 2024]

>= best scores

~ best scores (-0.01/0.02)

<< best scores

More generic approach
- Search + linking
- Trained on different datasets

- No specific treatment for aliases
- Struggling with abbreviations of person names
- Problems with numerical features

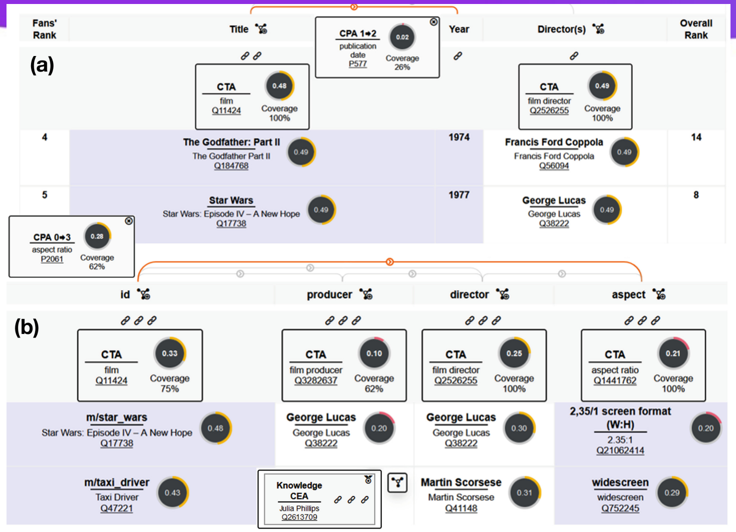
Confidence scores
- help detecting most uncertain links and interpretability for HITL

Much space for improvement
- Active learning
- Feedback propagation

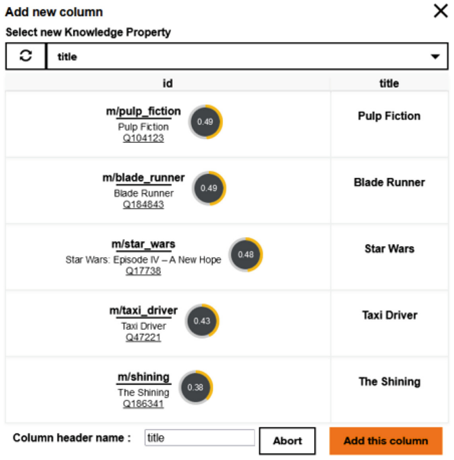
- Additional features can improve results matching specific heuristic approaches

- Still difficult to handle some aspects of matching (still no specific treatment for aliases, or person name abbreviations)

DAGOBAH UI
[Sarthou-Camy et al. 2022]



Annotations



Extension
with new columns

Functionalities	Karma	TableMiner+	Magic	MTab	MantisTable	STAN	OpenRefine	Trifacta	Odalic	DataGraft	Dagobah	SemTUI
Import of tables	✓	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓
Import of tables via API	✓	✗	✓	✓	✓	✓	✗	✓	✗	✓	✓	✓
Import of ontologies	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
Definition of personalised ontologies	✗	✗	✗	✗	✗	✓	✗	✗	✗	✓	✗	✗
Semi-automatic annotation/HITL	✓	✓	✓	✓	✓	✗	✗	✗	✓	✗	✓	✗
Annotation suggestions	✓	✗	✓	✗	✓	✓	✗	✗	✗	✓	✗	✓
Auto-complete support	✓	✗	✗	✗	✓	✗	✓	✗	✗	✓	✗	✓
Subject column detection	✓	✓	✗	✓	✓	✗	✗	✗	✓	✗	✓	✓
CEA	✗	✓	✓	✓	✓	✗	✓	✗	✓	✓	✓	✓
CTA	✗	✓	✓	✓	✓	✗	✓	✗	✓	✓	✓	✓
CPA (NE columns)	✗	✓	✓	✓	✓	✗	✓	✗	✓	✓	✓	✓
CPA (LIT columns)	✗	✓	✓	✓	✓	✗	✓	✗	✓	✓	✓	✓
Table manipulation	✓	✗	✗	✗	✓	✗	✓	✓	✗	✓	✗	✓
Automatic table extension	✗	✗	✓	✗	✗	✗	✓	✓	✗	✓	✗	✓
Visualisation of annotations	✗	✗	✓	✓	✓	✗	✗	✗	✗	✓	✓	✓
Auto save	✓	✗	✗	✗	✗	✗	✓	✓	✗	✓	✗	✗
Export mapping	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓
Export RDF triplets	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	X
Open Source	✓	✓	✓	✗	✓	✓	✓	✗	✓	✓	✗	✓

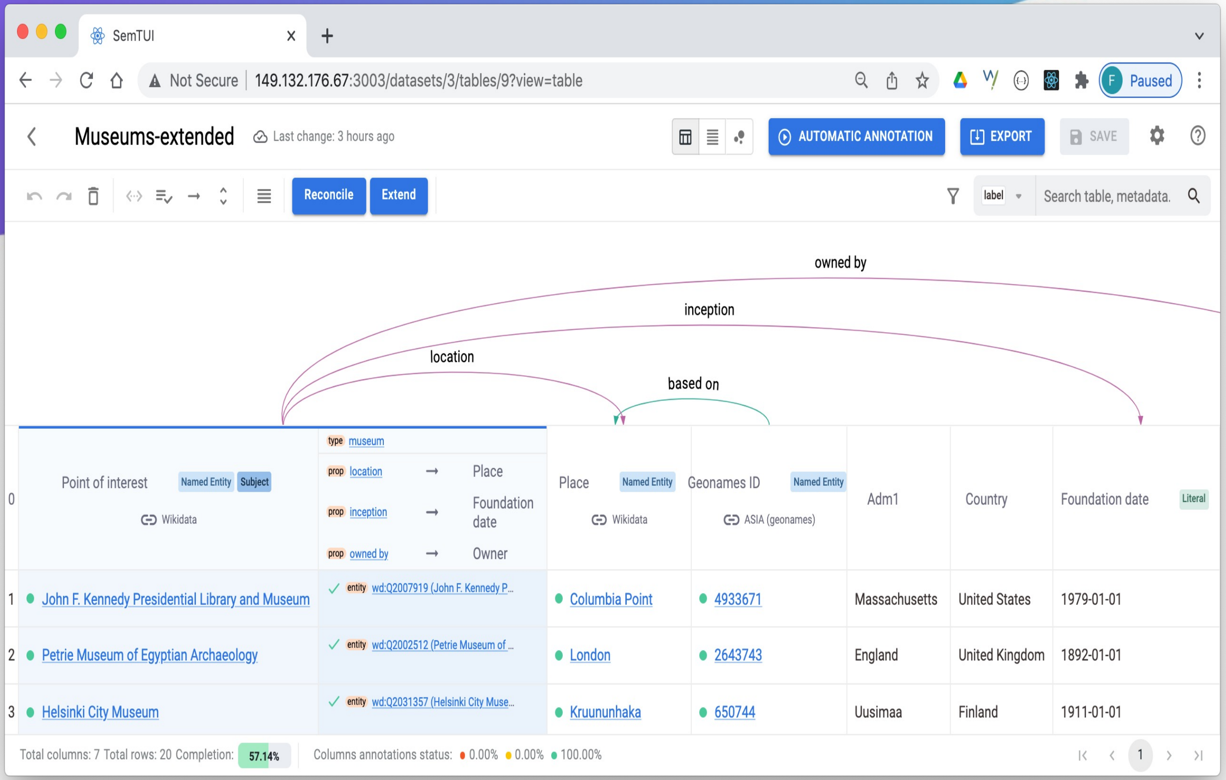
State of the art

Tools

Our tools: UNIMIB and/or SINTEF

SemTUI – Interactive Semantic Enrichment of Tabular Data

- UI accessing external services
  - Complete semantic table annotations
  - Access to different reconciliation/linking services
    - Wikidata (Alligator)
    - DBpedia
    - Geocoding (HERE)
    - Atoka-linking (SpazioDati)
    - … based on the OpenRefine interface (W3C Specs)
  - Extension services
    - Wikidata / DBpedia
    - Weather extension (OpenWeather)
    - Shortest-route (HERE)
    - Atoka-extension (SpazioDati)
    - …



Support to *Linking* – *Revision* – *Extension* of tabular data

- Graphical view & revision of annotations
  - Global and specific annotation rendering
  - Single cell editing / annotation revision
  - Column annotation revision

Ripamonti, M., De Paoli, F., & Palmonari, M. (2022). SemTUI: a Framework for the Interactive Semantic Enrichment of Tabular Data. *arXiv preprint arXiv:2203.09521.*

LLM-based Approaches

- LLMs in some tasks
  - CTA
    - DuoDuo: fine-tuned BERT [Suhara et al. 2022]
      - Adapted to SemTab by TorchiTab [Dasoulas et al. 2023]
    - DAGOBAH – incorporate Electra-based matching [Huynh et al. 2022]
    - ChatGPT for column annotation [Korini & Bizer 2024]
  - Related tasks (selected references)
    - Entity matching
      - ChatGPT for etity matching [Peters & Bizer 2023]
    - Multi-task entity matching
      - **Unicorn**: generalized cross-encoder based on multi-task training (Encoder (DeBERTa) - MoE – Matcher)
- LLMs for tabular data understanding and manipulation addressing all table annotation tasks
  - **TURL**: fine-tuned TinyBERT for tabular data understanding
    - Generic model + models for specific tasks (task-specific fine-tuning)
    - Parameters: 4M, 512 context
  - **TableLlama**: fine-tuned Llama for tabular data understanding and manipulation
    - Generic model with prompting (in-context learning)
    - Parameters: Llama fine-tuned with LongLoRA 7B, 8k context

Encoder-based
NLU + classifier

Decoder-based
Generative (NLG)
In-context learning

LLM-based Approaches: Tasks Summary

	Tasks					
	Annotation/Matching	Augmentation	QA	Fact Verification	Dialogue Generation	Data-to-Text
Unicorn	CEA CTA Entity Matching Entity Alignment Ontology Matching Schema Matching String Matching	N.A.	N.A.	N.A.	N.A.	N.A.
TURL	CEA CTA CPA	Row Population Cell Filling Schema Aug.	N.A.	N.A.	N.A.	N.A.
TableLlama	CEA CTA CPA	Row Population Schema Aug.	Hierarchical Table QA Highlighted Cells QA Hybrid Table QA Table QA	Fact Verification	Table Grounded Dialogue Generation	Highlighted Cells Description

Interesting for data enrichment

Interesting for interactive exploration

LLM-based Approaches: Inference Summary

	Input	Output	Transformer	Params
Unicorn	Encode pairs as: **[CLS] S(a) [SEP] S(b) [SEP]** where S(⋆) is a generic function for serializing any pair (a, b) from the matching tasks into a text sequence	A score in [0,1] for every pair (mention, i^{th}-candidate)	DeBERTa (Encoder-only)	147M
TURL	Flatten input table as: **[Table caption, Table Header-1, …, Table Header-M, Row-1, …, Row-N]**	A probability distribution over the N candidates for a mention	Tiny-BERT (Encoder-only)	14.5M
TableLlama	Prompt based: **<instruction, table input, question>** • Instruction is a detailed task description • Table input is the concatenation of table metadata (Wikipedia page title, section title and table caption) with the serialized table • Question contains all the information the model needs to complete the task and prompt it to generate an answer.	Autoregressively generated answer given the prompted question	Llama 2 (Decoder-only)	7B

TURL: Tabular Data Understanding

Mechanisms to consider the table structure and order

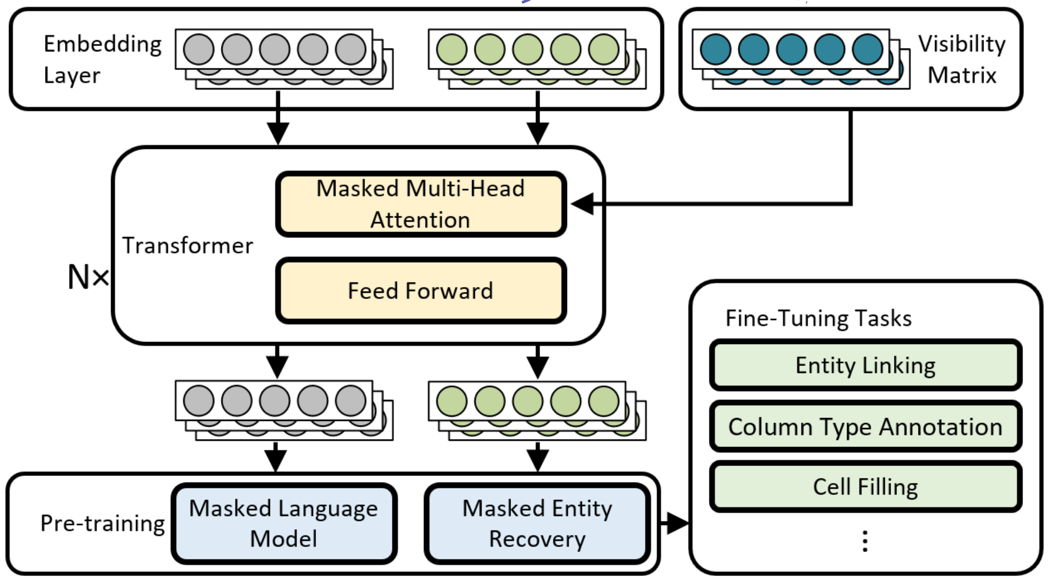


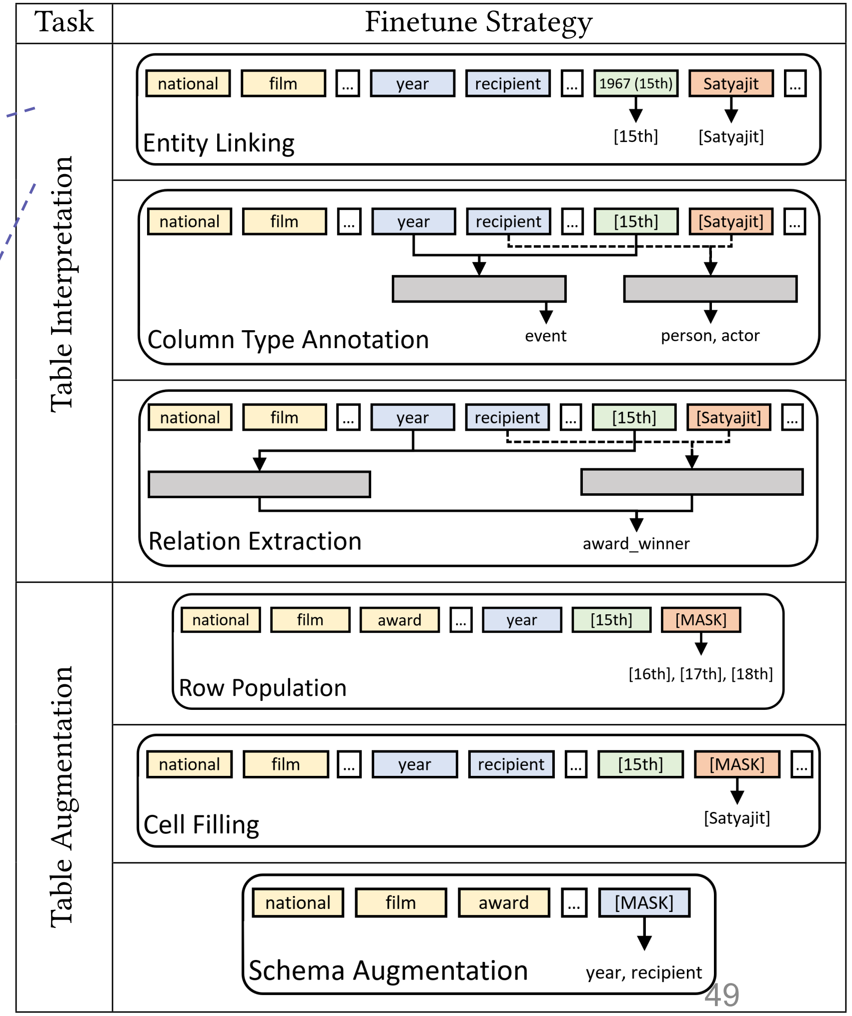
Figure 2: Overview of our TURL framework.

Inference: combination of lookup score and output

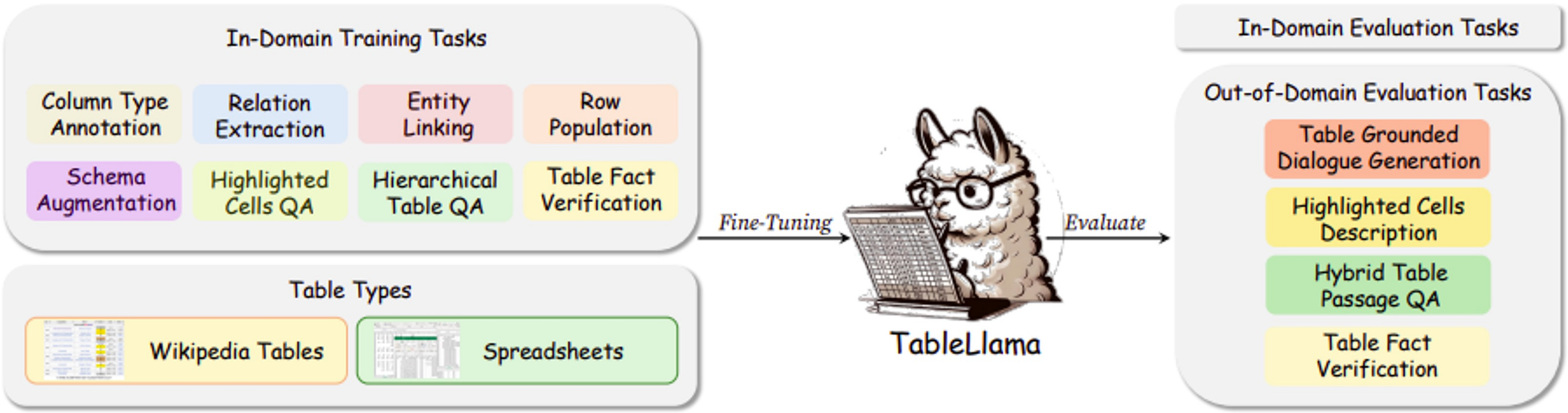
Disambiguation only

Efficient

Table 3: An overview of our benchmark tasks and strategies to fine-tune TURL.



TableLlama: Overview



Training with TableInstruct

- Dataset with 14 datasets for 11 tasks
- 1.24M tables

TableLlama: CEA example

Entity Linking

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction: This is an entity linking task. The goal for this task is to link the selected entity mention in the table cells to the entity in the knowledge base. You will be given a list of referent entities, with each one composed of an entity name, its description and its type. Please choose the correct one from the referent entity candidates. Note that the Wikipedia page, Wikipedia section and table caption (if any) provide important information for choosing the correct referent entity.

Input: [TLE] The Wikipedia page is about A-League all-time records. The Wikipedia section is about Average season attendances. [TAB] col: | season | league average | total gate receipts | highest club | average | lowest club | average | row 1: | 2005-06 | 10,955 | 920,219 | Sydney FC | 16,669 | New Zealand Knights | 3,909 | [SEP] row 2: | 2006-07 | 12,927 | ...

Question: The selected entity mention in the table cell is: Melbourne Victory. The column name for 'Melbourne Victory' is highest club. The referent entity candidates are: <Melbourne Victory FC W-League [DESCRIPTION] None [TYPE] SoccerClub>, <2016\u201317 Melbourne Victory FC season [DESCRIPTION] None [TYPE] SoccerClubSeason>, <2011\u201312 Melbourne Victory season [DESCRIPTION] Association football club 2011/12 season for Melbourne Victory [TYPE] SoccerClubSeason>, ... What is the correct referent entity for the entity mention 'Melbourne Victory' ?

Response: <Melbourne Victory [DESCRIPTION] association football team from Australia [TYPE] SoccerClub>.

Entity linking with GenAI

- **Challenges**
  - Requires candidate retrieval
  - Context length
    - Table
    - Candidates
  - Returns answer in natural language
  - Interaction
- **Cons**
  - Confidence estimation is non-trivial
  - Does not scale;
    - e.g., ~1000x TURL's execution time
    - some tables do not fit the context
- **Pros**
  - Great generalization
  - In-context learning and task adaptation
  - Promising new data enrichment features

TableLlama: CTA example

Column Type Annotation

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
This is a column type annotation task. The goal for this task is to choose the correct types for one selected column of the table from the given candidates. The Wikipedia page, section and table caption (if any) provide important information for choosing the correct column types.

Input:
[TLE] The Wikipedia page is about 1958 Nippon Professional Baseball season. The Wikipedia section is about Central League. The table caption is Pitching leaders. [TAB] col: | stat | player | team | total | [SEP] row 1: | Wins | Masaichi Kaneda | Kokutetsu Swallows | 31| [SEP] row 2: | Losses | Noboru Akiyama | ...

Question:
The column ’player’ contains the following entities: <Masaichi Kaneda>, <Noboru Akiyama>, etc. The column type candidates are: tv.tv_producer, astronomy.star_system_body, location.citytown, sports.pro_athlete, biology.organism, medicine.muscle, baseball.baseball_team, baseball.baseball_player, aviation.aircraft_owner, people.person, ... What are the correct column types for this column (column name: player; entities: <Masaichi Kaneda>, <Noboru Akiyama>, etc)?

Response:
sports.pro_athlete, baseball.baseball_player, people.person.

TableLlama: In-Domain Results

⚠ CTA and CEA test sets are subsampled from the original test data from TURL

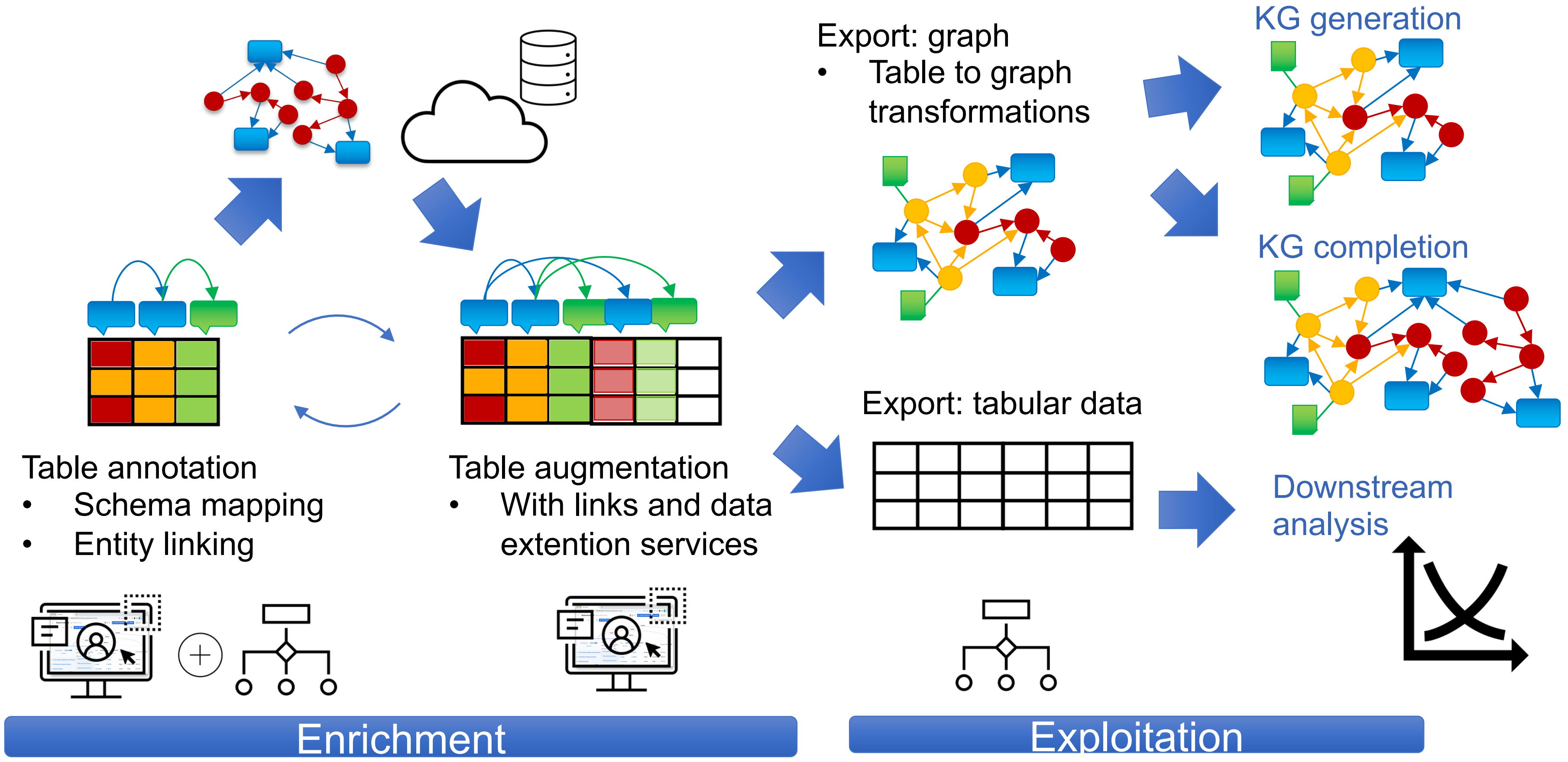
⚠ High costs (results for 500 samples)
See also [Peeters & Bizer 2023]

In-domain Evaluation						
Datasets	Metric	Base	TableLlama	SOTA	GPT-3.5	GPT-4§
Column Type Annotation	F1	3.01	94.39	**94.54***† (Deng et al., 2020)	30.88	31.75
Relation Extraction	F1	0.96	91.95	**94.91***† (Deng et al., 2020)	27.42	52.95
Entity Linking	Accuracy	31.80	**93.65**	84.90*† (Deng et al., 2020)	72.15	90.80
Schema Augmentation	MAP	36.75	**80.50**	77.55*† (Deng et al., 2020)	49.11	58.19
Row Population	MAP	4.53	58.44	**73.31***† (Deng et al., 2020)	22.36	53.40
HiTab	Exec Acc	14.96	**64.71**	47.00*† (Cheng et al., 2022a)	43.62	48.40
FeTaQA	BLEU	8.54	**39.05**	33.44 (Xie et al., 2022)	26.49	21.70
TabFact	Accuracy	41.65	82.55	**84.87*** (Zhao and Yang, 2022)	67.41	74.40

Table 2: In-domain evaluation results. "Base": LongLoRA model w/o fine-tuning on `TableInstruct`; "*": w/ special model architecture design for tables/tasks; "†": w/ table pretraining; "§": for GPT-4, we uniformly sample 500 examples from test set for each task due to limited budget.

Problem: comparison with approaches tested on SemTab unclear (work in progress)

Wrap-up: Semantic Table Annotations vs Data Enrichment



For *large* data enrichment

- Annotations > pipeline specs > scalable deployment
- Interoperability with third-party sources
- Interactive exploration and HITL
- Scalability and sustainability of annotation algorithms (time, €)
- Several methods but limited integration yet

Wrap-up: Algorithms + Humans for Semantic Data Erichment

- Algorithms
  - Pre-compute annotations
    - Schema-level (reference vocabularies)
    - Instance-level (entities)
  - Fuse data from the target data sources into the source data
  - Manipulate data
  - Transform the data into semantically annotated data at scale

- Humans
  - **Revise** annotations
  - **Configure** reconciliation services
  - **Fine-tune** pre-trained algorithms on specific data (w. limited effort)
  - Specify which data to fuse
  - Specify manipulations

Wrap-up: Semantic Table Annotation SOTA vs. Data Enrichment

- Algorithms
  - Several specific heuristic methods *from* SemTab challenges
    - High performance on SemTab data (… and previous datasets)
  - LLM-based generalistic approaches
    - High generalization
    - Novel enrichment features
    - Significant scalability issues
    - Interpretability issues and control
- Tools
  - Some tools available
  - Limited maturity
  - Limited exploitation for data enrichment
  - No connection to LLM-based generalistic approaches

Towards HITL Enrichment in the Practical Section

KEYWORD	#im	REGION	Date
194906	64	Thuringia	2017-03-11
517827	50	Bavaria	2017-03-12
459143	42	Berlin	2017-03-12

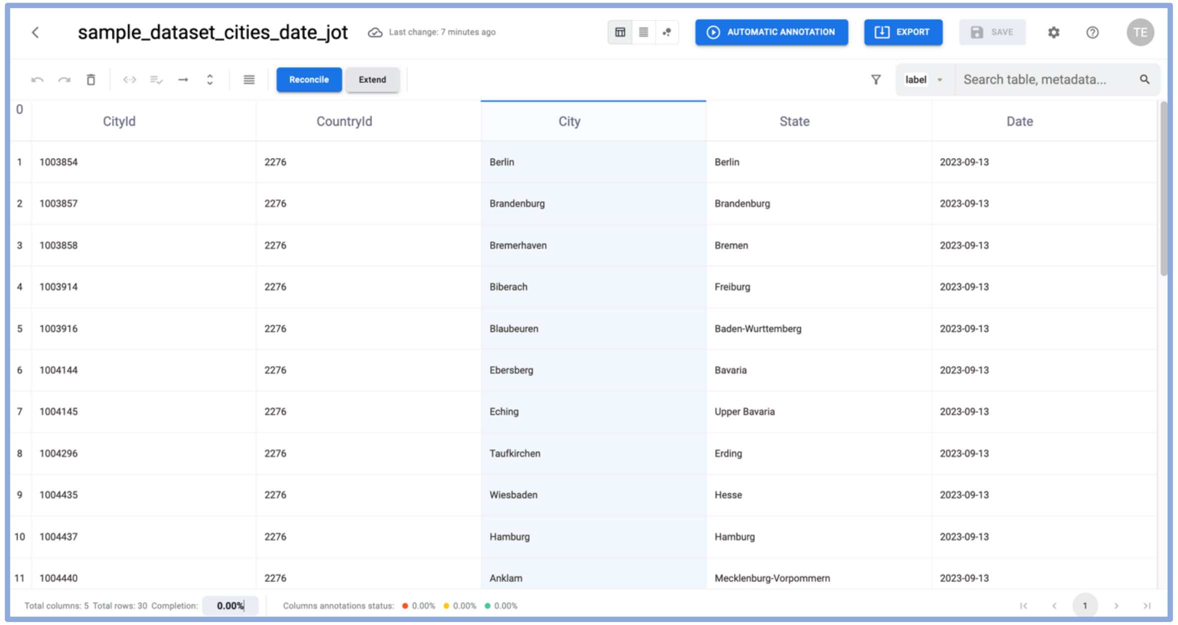
→ ? →

C°/+0	C°/+1
18	20
17	19
17	20

Input data JOT INTERNET MEDIA

Additional data

- SemTUI [https://i2tunimib.github.io/I2T-docs/]
  - Interactive web application
    - Link & extend paradigm
  - Interoperates with different services for
    - data linking
      - Wikidata, Geonames, **Geocoding APIs**, Atoka, etc.
    - data extension
      - Wikidata, **weather APIs**, route plans, Atoka)
    - end-to-end abular data annotation
- Alligator [Avogadro&al.WI23]
  - Confidence-aware entity linking
    - features + NNs



0	CityId	CountryId	City	State	Date
1	1003854	2276	Berlin	Berlin	2023-09-13
2	1003857	2276	Brandenburg	Brandenburg	2023-09-13
3	1003858	2276	Bremerhaven	Bremen	2023-09-13
4	1003914	2276	Biberach	Freiburg	2023-09-13
5	1003916	2276	Blaubeuren	Baden-Wurttemberg	2023-09-13
6	1004144	2276	Ebersberg	Bavaria	2023-09-13
7	1004145	2276	Eching	Upper Bavaria	2023-09-13
8	1004296	2276	Taufkirchen	Erding	2023-09-13
9	1004435	2276	Wiesbaden	Hesse	2023-09-13
10	1004437	2276	Hamburg	Hamburg	2023-09-13
11	1004440	2276	Anklam	Mecklenburg-Vorpommern	2023-09-13