Automated Annotation of Tabular Data for Clinical Applications

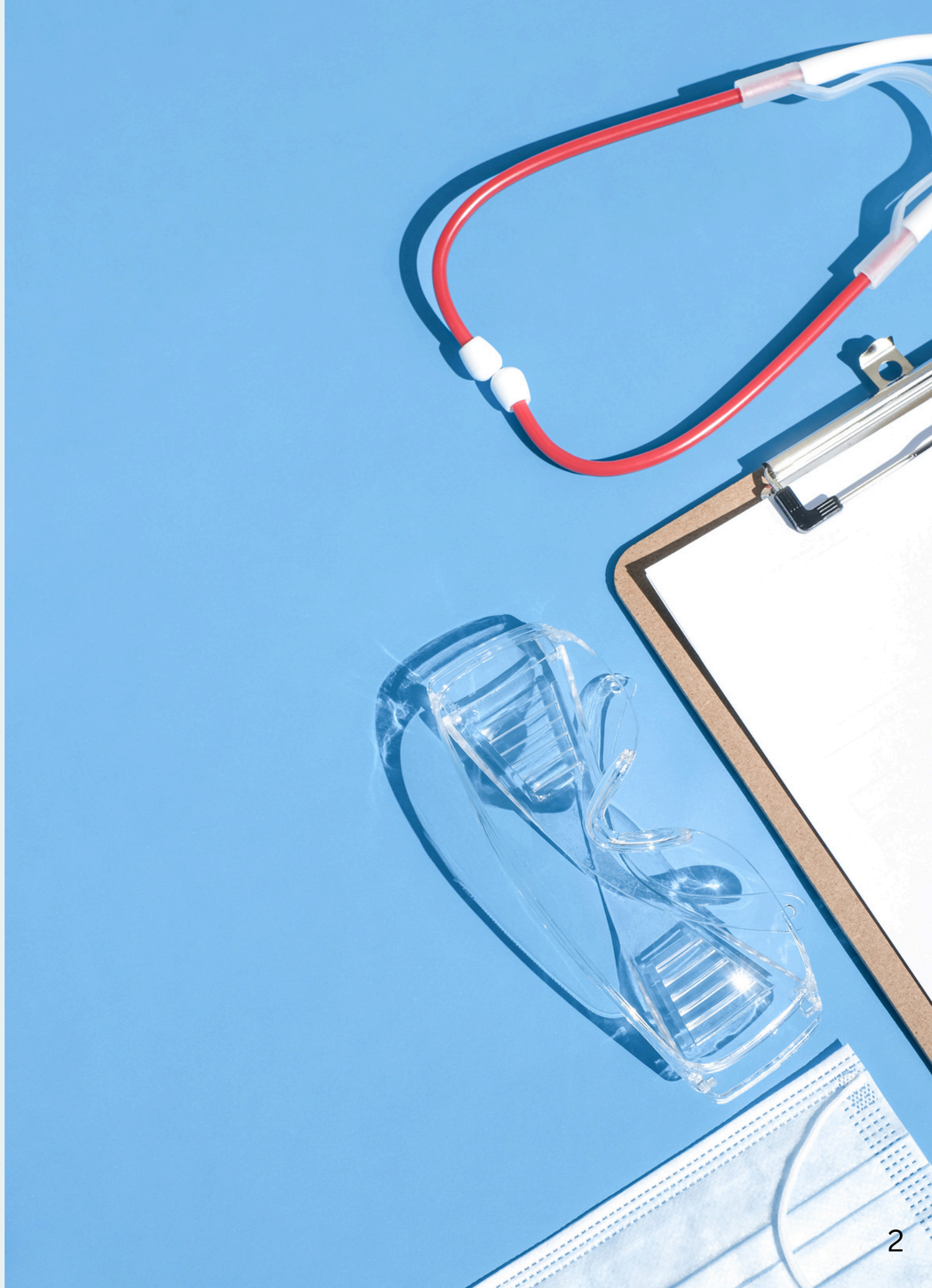
Group 16 Masters Research Project 2 2023-2024

Anton Changalidi, Aurora Pia Ghiardelli, Ashkan Karimi Saber, Ivan Poliakov
Supervisors: Shervin Mehryar, Dr. Remzi Celebi

24 June 2024

Maastricht University

CONTEXT and MOTIVATION



What is automatic annotation of tabular data?

- The **table** is the **most useful format** of representing information!
- **Same context, different entities** - how to annotate, align and match?
- **Manual labor** is widely used - expensive...
- Tables sometimes are not **interpretable!**
- **Medical field - even harder!**

Problem: There is no such tool that annotates **inconsistent clinical tables,** identifying **entities and their relationships** in a **reference ontology**.

What are EHRs?

EHRs(Electronic Health Records) are digital versions of conventional paper-based medical data.

They include demographics, medical history, diagnosis, treatments, and results of the patients.

This data helps identify **risk factors**, track **disease progression**, and assess the **effectiveness of treatments**, ultimately enhancing **healthcare quality** and **cutting costs**.

EHRs are usually disorganized, hard to analyze, and fragmented.

Knowledge Graphs for EHRs

They allow for the **integration of external knowledge** sources with EHR data.

A complete ontology can help to **identify previously unknown relationships** between clinical concepts.

A KG can be useful in **Clinical Decision Support Systems (CDSSs)**.

Scalability, interoperability difficulties, clinical validity issues, privacy and security hazards limit the present knowledge graph research employing EHR data.

No universal reference ontology for clinical data/EHRs.

Current clinical knowledge graph models are limited to **smaller-scale datasets** or **specific healthcare domains.**

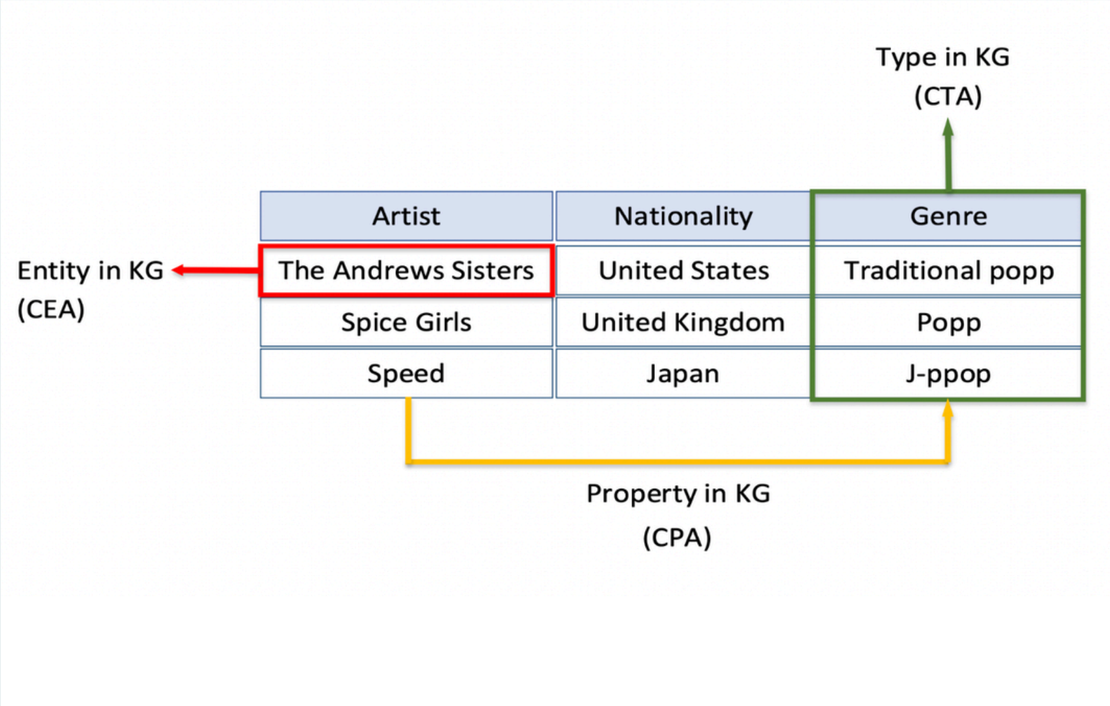
Need for the alignment of ontologies with **real-world data**, in terms of:

- patient outcomes
- clinical decision making
- hospital protocols

CONCEPTS and APPROACHES

Tasks: CTA, CEA, CPA

multi-class classification task



CTA: column type annotation **CPA**: column property annotation **CEA**: cell entity annotation

Figure 1

Baseline: Neuro-Symbolic Pipeline

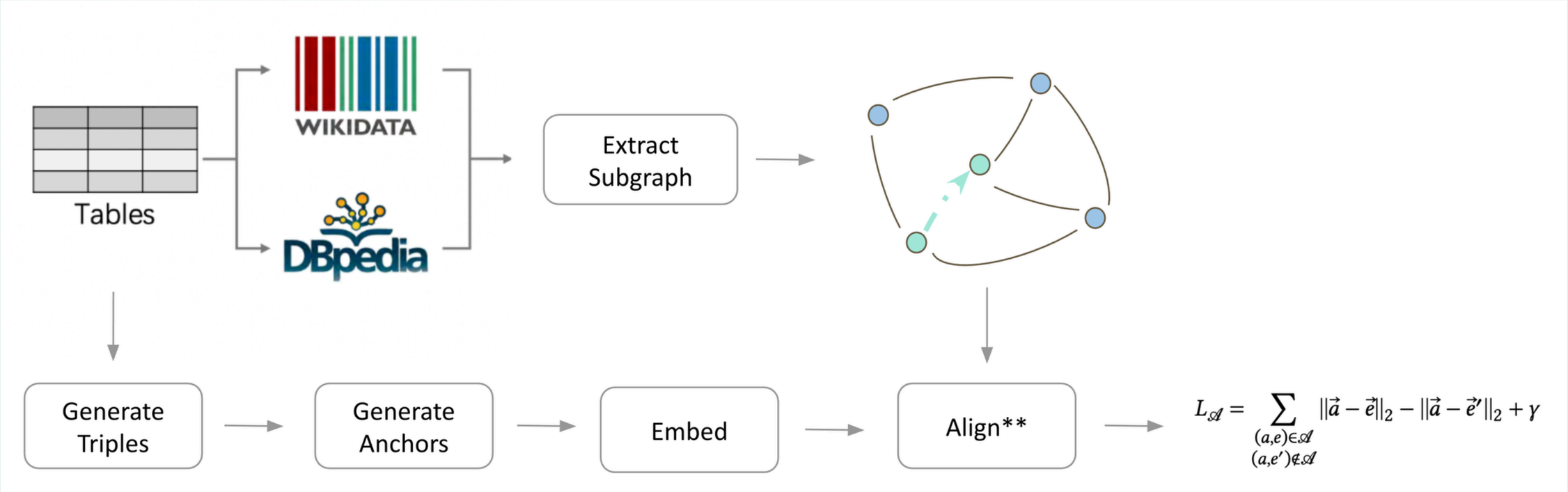


Figure 2 - Allignment (Mehryar, Celebi, 2023)

Contribution

Research Questions

- What are the main features of the currently available clinical ontologies? Are they sufficiently representative of hospital data?
- Is it possible to effectively use these ontologies for manually annotating a clinical table?
- Once this annotation is done, how can we ensure its correctness?

Our contribution

- Analysis of the **structure, pros, and cons of the chosen ontologies**;
- Analysis of the **differences between the MIMIC dataset and the ontologies** used for annotation;
- Planning and application of an **annotation technique and its evaluation**.

The main goal of the annotation was to get:

- Column Type Annotation (CTA)
- Column Property Annotation (CPA)
- Cell Entity Annotation (CEA) with respect to a single patient of the MIMIC dataset.

Dataset

MIMIC-III:

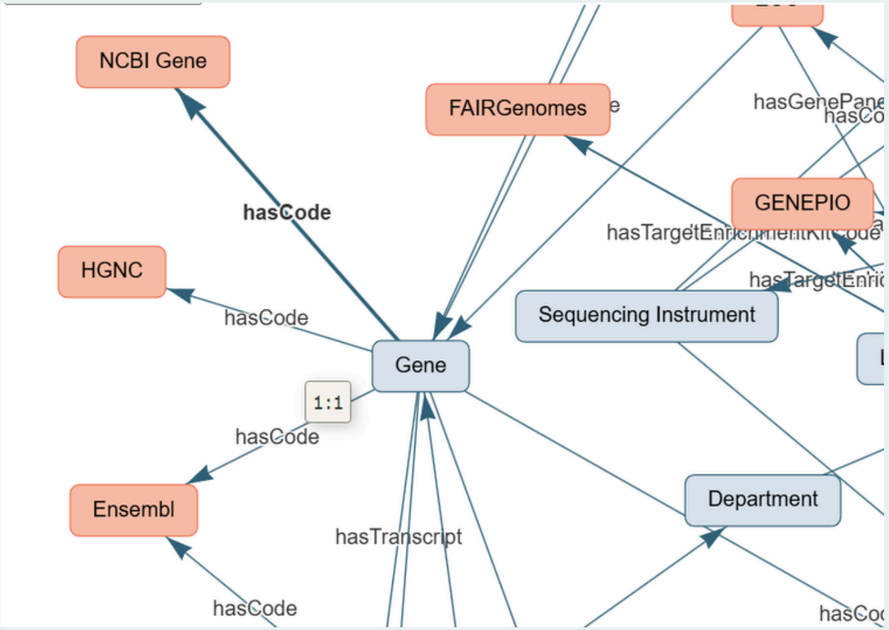
- **26 relational tables** of de-identified **electronic medical records** admitted to the Beth Israel Deaconess Medical Center between **2001 and 2012**
- **Publicly available**
- Selected for **building the EHR knowledge graph for patients**

Existing Ontologies

- **SPHN RDF schema**
- **SNOMED**
- Other vocabolaries: **PUBCHEM**, **ICD9**, **LOINC**

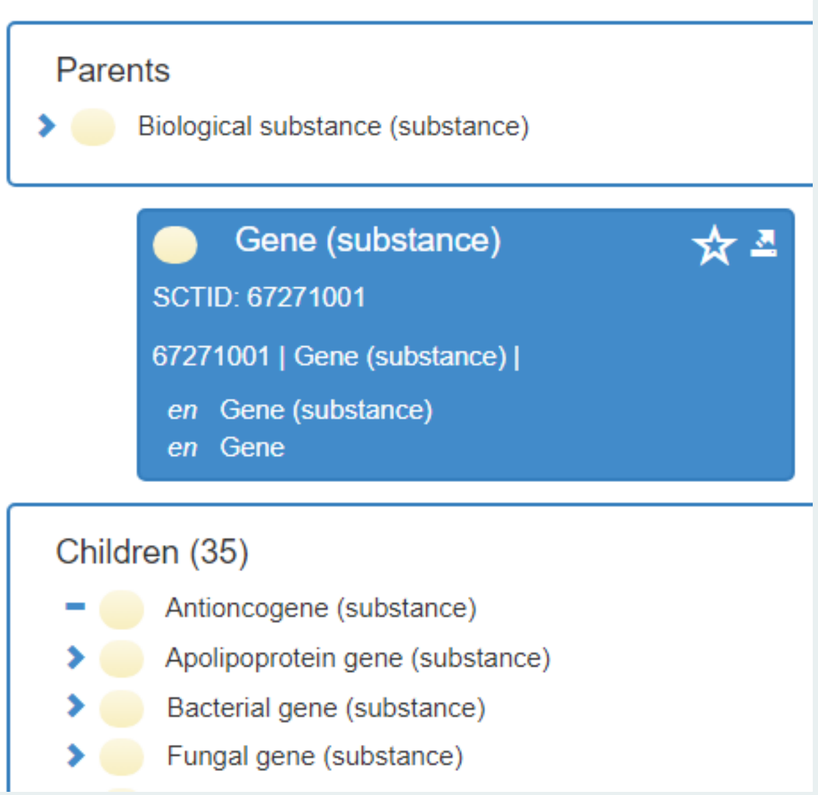
SPHN Rdf Schema

- **SPHN Rdf Schema** is a Standardized structure for organizing and representing healthcare data.
- Based on the **SPHN Dataset**.
- Atomic building blocks: **concepts**.
- Refers to recommended value sets and/or semantic standards (e.g. LOINC, SNOMED CT, ICD-10-GM, CHOP, ATC, ICD-O-3, HGNC, GENO, SO) to express the data.
- High level of abstraction, few specific entities, nice properties structure.



SNOMED CT: Systematized Nomenclature of Medicine - Clinical Terms

- The **most comprehensive, multilingual clinical healthcare** terminology in the world:
  - 50 countries across the world use it.
- Contains **core general terminology of electronic health report.**
- **SNOMED CT Browser** provides ways to browse and search SNOMED CT.
- Most specific entities but it's just a vocabulary with parents/children relations.



CTA task

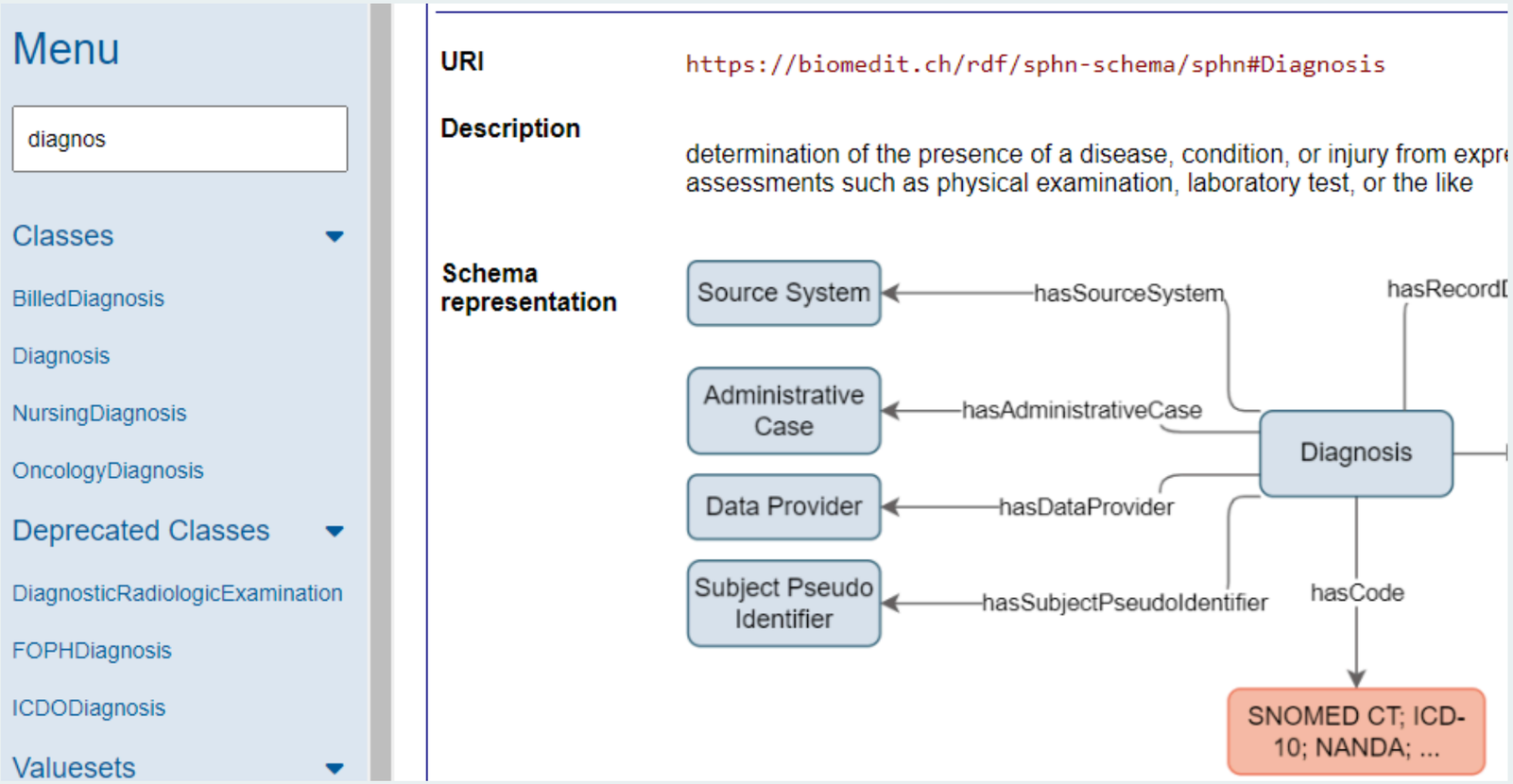
- First step for manual table annotation
- Map each column type to an ontology entity (a class or a property)
- Ontologies used: SPHN or SNOMED

According to the description of this table, this column corresponds to a single diagnostic concept, so we map it to SPHN's Diagnosis concept.

A	B	C	D
row_id	icd9_code	short_title	long_title
1	1716	Erythem nod tb-oth test	Erythema nodosum with hypersensitivity reaction in tuberculosis, tubercle bacilli not found by bacteriological or histological
2	1720	TB periph lymph-unspec	Tuberculosis of peripheral lymph nodes, unspecified
3	1721	TB periph lymph-no exam	Tuberculosis of peripheral lymph nodes, bacteriological or histological examination not done
4	1722	TB periph lymph-exam unk	Tuberculosis of peripheral lymph nodes, bacteriological or histological examination unknown (at present)
5	1723	TB periph lymph-micro dx	Tuberculosis of peripheral lymph nodes, tubercle bacilli found (in sputum) by microscopy
6	1724	TB periph lymph-cult dx	Tuberculosis of peripheral lymph nodes, tubercle bacilli not found (in sputum) by microscopy, but found by bacterial culture

CTA task

Examine the neighborhood and select relevant components



CTA - validation

An example of cross-validation for CTA task is as follows:

table	column	annotation1	annotation2	flag
DRGCODES	hadm_id	https://biomedit.ch/rdf/sphn-schema/sphn#AdministrativeCase	https://biomedit.ch/rdf/sphn-schema/sphn#AdministrativeCase	TRUE
DRGCODES	drg_type	https://biomedit.ch/rdf/sphn-schema/sphn#hasTypeCode	https://biomedit.ch/rdf/sphn-schema/sphn#Diagnosis	FALSE
DRGCODES	drg_code	https://biomedit.ch/rdf/sphn-schema/sphn#hasCode	https://biomedit.ch/rdf/sphn-schema/sphn#hasCode	TRUE
DRGCODES	description	https://biomedit.ch/rdf/sphn-schema/sphn#hasDescription	https://biomedit.ch/rdf/sphn-schema/sphn#hasDescription	TRUE
DRGCODES	drg_severity	https://biomedit.ch/rdf/sphn-schema/sphn#hasSeverityCode	https://biomedit.ch/rdf/sphn-schema/sphn#hasSeverityCode	TRUE
DRGCODES	drg_mortality	https://biomedit.ch/rdf/sphn-schema/sphn#hasCode	?	FALSE
DIAGNOSIS_ICI	hadm_id	https://biomedit.ch/rdf/sphn-schema/sphn#AdministrativeCase	https://biomedit.ch/rdf/sphn-schema/sphn#AdministrativeCase	TRUE
DIAGNOSIS_ICI	seq_num	https://biomedit.ch/rdf/sphn-schema/sphn#hasNumericalRefere	?	FALSE

CTA - Result (for 1 patient)

- Table for each MIMIC table
- Type under each column name (using CTA cross-validated data)

isolate_num
https://biomedit.ch/rdf/sphn-schema/sphn#MicrobiologyBiomoleculePresenceResult
https://biomedit.ch/rdf/sphn-schema/sphn#hasNumericalReference

icustay_id
https://browser.ihtsdotools.org/?perspective=full&conceptId1=309904001&edition=MAIN&release=&languages=en
https://biomedit.ch/rdf/sphn-schema/sphn#hasIdentifier
241992

row_id	
https://www.biomedit.ch/rdf/sphn-schema/sphn#BilledDiagnosis	

CTA - Result (for 1 patient)

- Sometimes we annotated column headers with properties from SPHN or SNOMED

startdate
https://biomedit.ch/rdf/sphn-schema/sphn#hasFirstAdministrationDateTime

drug_type
https://biomedit.ch/rdf/sphn-schema/sphn#hasCategory

drug_name_poe
https://biomedit.ch/rdf/sphn-schema/sphn#hasName

CPA task

We perform **CPA** (column property annotation/predicate annotation) to capture the relations between the column types

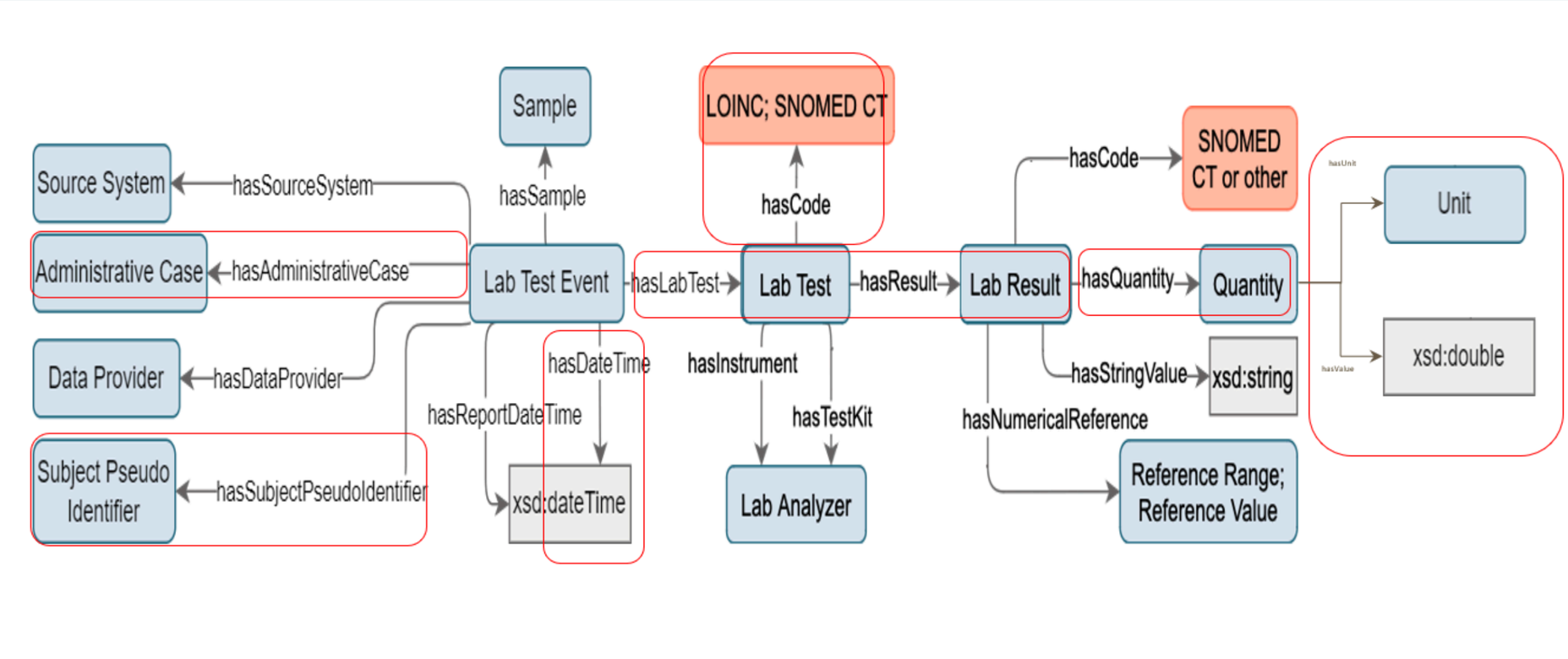
As an example, consider the "**Prescription**" table in MIMIC dataset:

```
drug        + has value -> dose_val_rx
dose_val_rx + has unit  -> dose_unit_rx
```

H	I	J	K	L	M	N	O	P
drug	drug_na	drug_nam	formulary	gsn	ndc	prod_	dose_val_	dose_unit_
Pneumoc	Pneumc	PNEUMOc	PNEU25I	####	6E+06	25mc	0.5	mL
Bisacodyl	Bisacod	Bisacodyl	BISA5	2947	5E+08	5 mg	10	mg
Bisacodyl	Bisacod	Bisacodyl	BISA10R	2944	6E+08	10mg	10	mg
Senna	Senna	Senna	SENN187	####	9E+08	1 Tabl	1	TAB
Docusate	Docusat	Docusate	DOCU100L	3017	1E+08	100m	100	mg
Heparin	Heparin	Heparin S	HEPA5I	6549	6E+10	5000	5000	UNIT

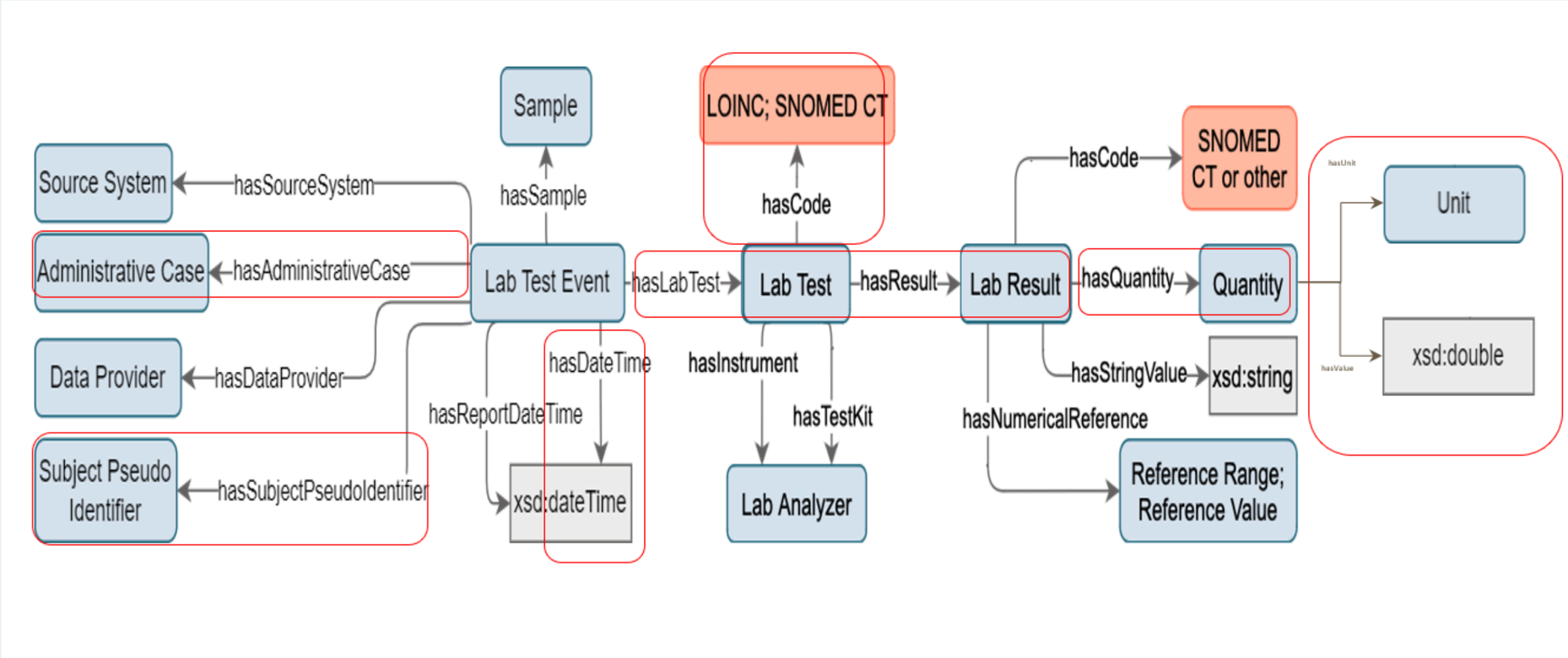
CPA task

Aim for a tree-like type property chain around column-zero



CPA task extended

- Introducing intermediate nodes to indirectly connect a pair of columns within CPA task



CPA task extended Challenges

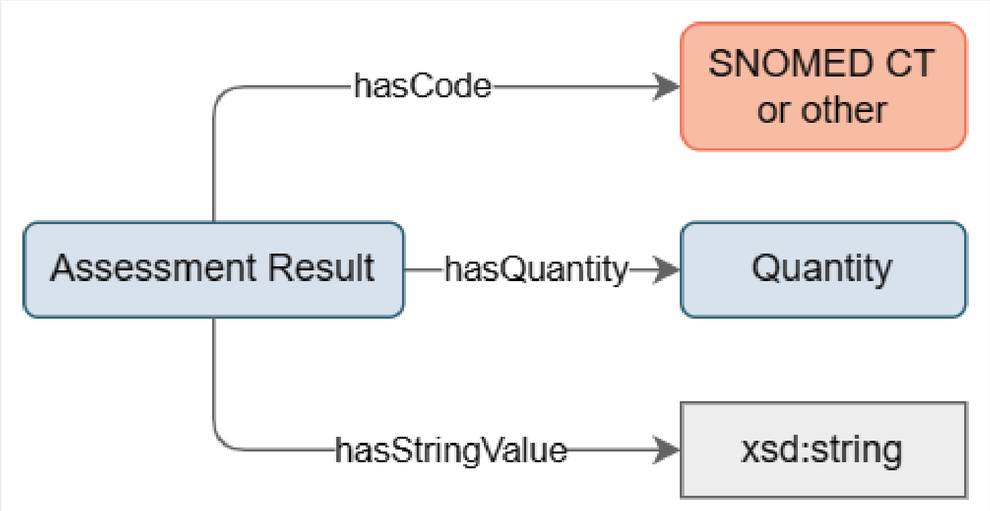
table	source	target	label	level
OUTPUTEVENTS	row_id	subject_id	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasIdentifier	0
OUTPUTEVENTS	row_id	icustay_id	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasCareHandling	0
OUTPUTEVENTS	row_id	charttime	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasEventDateTime	0
OUTPUTEVENTS	row_id	cgid	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasPerformer	0
OUTPUTEVENTS	hadm_id	icustay_id	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasCareHandling	0
OUTPUTEVENTS	row_id	itemid	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasResult	1
OUTPUTEVENTS	cgid	description_CAREGIVERS	https://biomedit.ch/rdf/sphn-schema/sphn#hasDescription	0
OUTPUTEVENTS	cgid	label_CAREGIVERS	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasName	0
OUTPUTEVENTS	itemid	storetime	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasDateTime	0
OUTPUTEVENTS	itemid	value	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasValue	1
OUTPUTEVENTS	itemid	valueuom	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasUnit	1

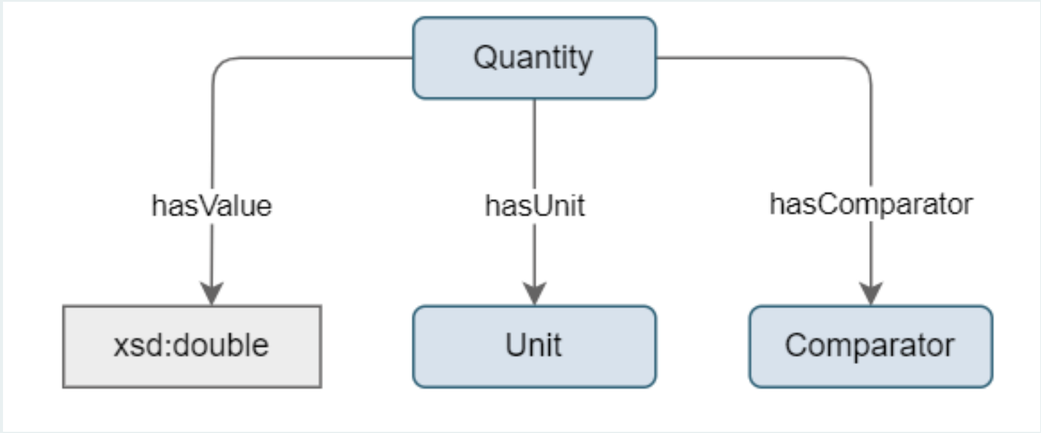
Level of Connection

Direct Connection

Indirect Connection

CPA task extended Challenges





CPA task extended

Final result of CPA task with the introduction of an intermedaite node:

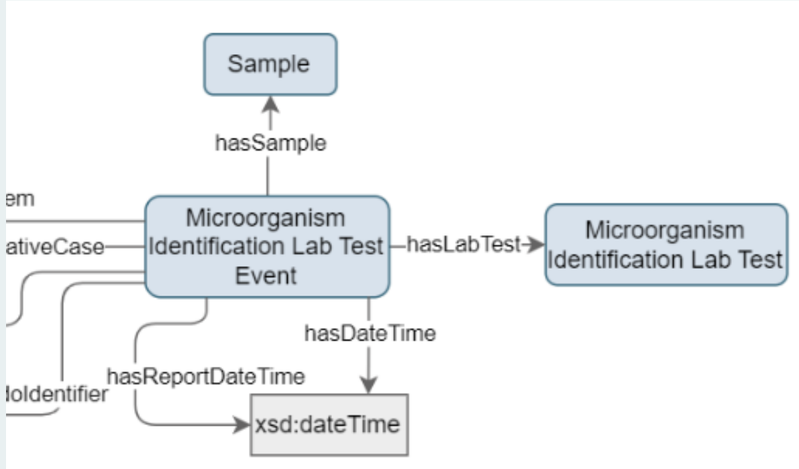
OUTPUTEVENTS	cgid	description_CAREGIVERS	https://biomedit.ch/rdf/sphn-schema/sphn#hasDescription
OUTPUTEVENTS	cgid	label_CAREGIVERS	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasName
OUTPUTEVENTS	itemid	storetime	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasDateTime
OUTPUTEVENTS	itemid	https://www.biomedit.ch/rdf/sphn-schema/sphn#Quantity	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasQuantity
OUTPUTEVENTS	https://www.biomedit.ch/rdf/sphn-schema/sphn#Quantity	value	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasValue
OUTPUTEVENTS	https://www.biomedit.ch/rdf/sphn-schema/sphn#Quantity	valueuom	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasUnit

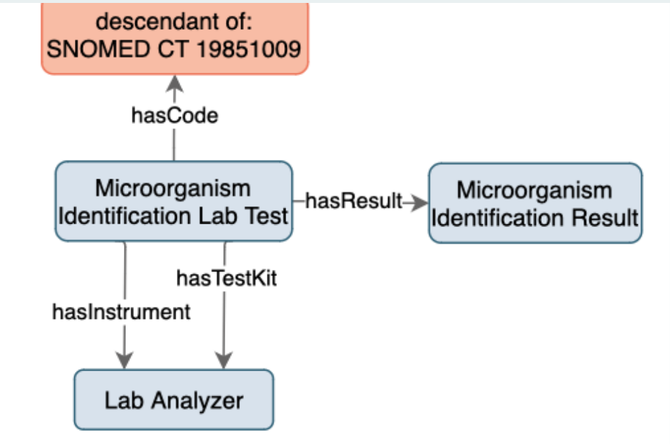
CPA task extended Challenges

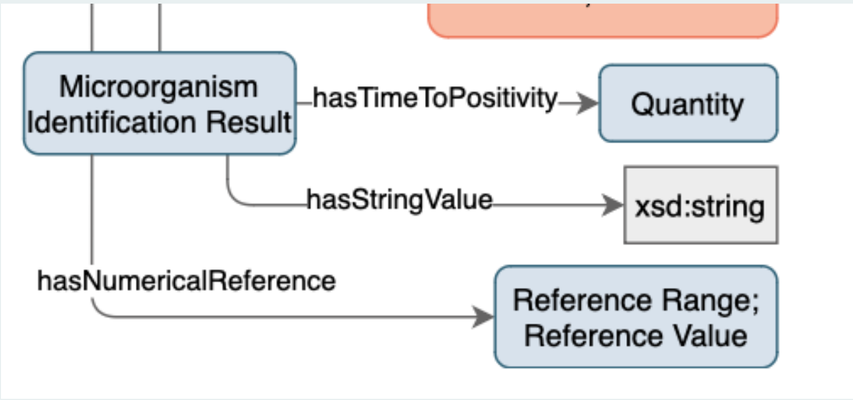
MICROBIOLOGY	org_itemid	subject_id	https://biomedit.ch/rdf/sphn-schema/sphn#hasSubjectPseudoIdentifier	0
MICROBIOLOGY	spec_itemid	charttime	https://biomedit.ch/rdf/sphn-schema/sphn#hasCollectionDateTime	0
MICROBIOLOGY	isolate_num	dilution_text	https://biomedit.ch/rdf/sphn-schema/sphn#hasDescription	0
MICROBIOLOGY	isolate_num	lution_comparis	https://biomedit.ch/rdf/sphn-schema/sphn#hasComparato	1
MICROBIOLOGY	isolate_num	dilution_value	https://biomedit.ch/rdf/sphn-schema/sphn#hasValu	1
MICROBIOLOGY	org_itemid	interpretation	https://biomedit.ch/rdf/sphn-schema/sphn#hasStringValue	1

org_itemid	https://biomedit.ch/rdf/sphn-schema/sphn#MicroorganismIdentificationLabTest	https://biomedit.ch/rdf/sphn-schema/sphn#hasLabTest
hedit.ch/rdf/sphn-schema/sphn#MicroorganismIdentificationLabTest	https://biomedit.ch/rdf/sphn-schema/sphn#MicroorganismIdentificationResult	https://biomedit.ch/rdf/sphn-schema/sphn#hasResult
hedit.ch/rdf/sphn-schema/sphn#MicroorganismIdentificationResult	interpretation	https://biomedit.ch/rdf/sphn-schema/sphn#hasStringValue

CPA task extended Challenges







CPA - examples 0

Direct examples:

- Vast majority of CPAs

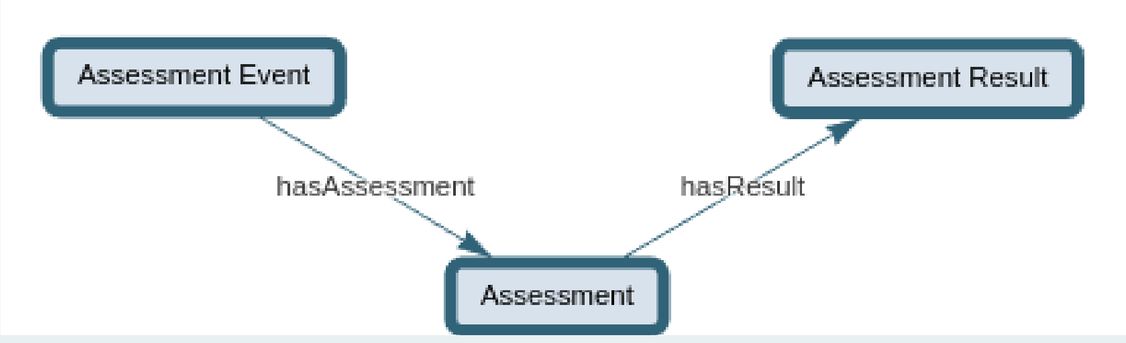
INPUTEVENTS_MV	hadm_id	icustay_id	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasCareHandling	0
INPUTEVENTS_MV	row_id	cgid	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasPerformer	0
INPUTEVENTS_MV	cgid	description	https://biomedit.ch/rdf/sphn-schema/sphn#hasDescription	0
INPUTEVENTS_MV	cgid	label	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasName	0

PRESCRIPTIONS	row_id	drug	https://biomedit.ch/rdf/sphn-schema/sphn#hasDrug	0
PRESCRIPTIONS	drug	drug_type	https://biomedit.ch/rdf/sphn-schema/sphn#hasCategory	0
PRESCRIPTIONS	drug	drug_name_poe	https://biomedit.ch/rdf/sphn-schema/sphn#hasName	0
PRESCRIPTIONS	drug	drug_name_generic	https://biomedit.ch/rdf/sphn-schema/sphn#hasName	0
PRESCRIPTIONS	drug	formulary_drug_cd	https://biomedit.ch/rdf/sphn-schema/sphn#hasCode	0
PRESCRIPTIONS	drug	gsn	https://biomedit.ch/rdf/sphn-schema/sphn#hasCode	0
PRESCRIPTIONS	drug	ndc	https://biomedit.ch/rdf/sphn-schema/sphn#hasCode	0

CPA - examples 1

OutputEvents:

- rowid -> itemid
  - rowid (Assessment Event)
  - -> Assessment
  - -> itemid (Assessment Result)



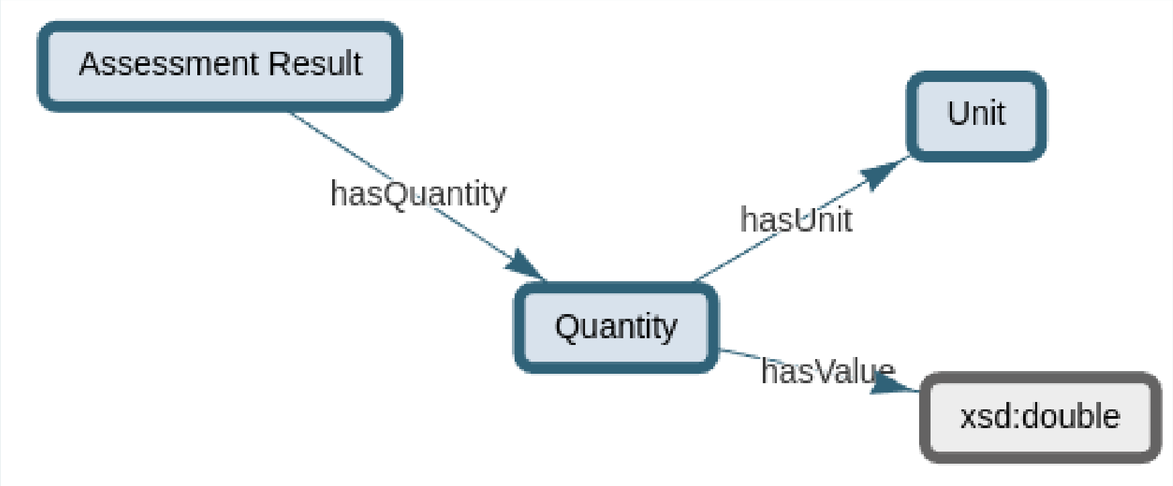
OUTPUTEVENTS	row_id	itemid	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasResult	1

OUTPUTEVENTS	row_id	https://www.biomedit.ch/rdf/sphn-schema/sphn#Assessment	https
OUTPUTEVENTS	https://www.biomedit.ch/rdf/sphn-schema/sphn#Assessment	itemid	https
		ment	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasAssessment
			https://www.biomedit.ch/rdf/sphn-schema/sphn#hasResult

CPA - examples 2

OutputEvents (and in some other tables):

- itemid -> value & valueuom
  - itemid (Assessment Result)
  - -> Quantity
  - -> value & valueuom (xsd:double & Unit)



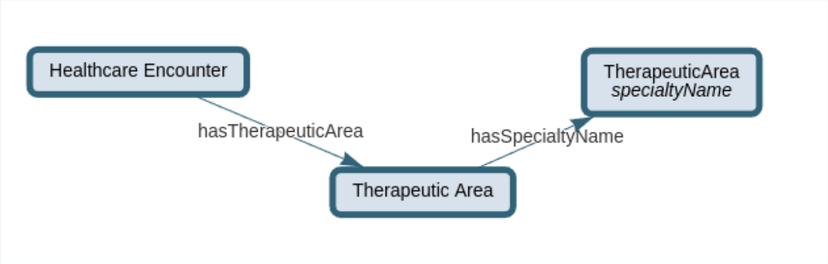
OUTPUTEVENTS	itemid	value	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasValue	1
OUTPUTEVENTS	itemid	valueuom	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasUnit	1

OUTPUTEVENTS	itemid	https://www.biomedit.ch/rdf/sphn-schema/sphn#Quantity	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasQuantity
OUTPUTEVENTS	https://www.biomedit.ch/rdf/sphn-schema/sphn#Quantity	value	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasValue
OUTPUTEVENTS	https://www.biomedit.ch/rdf/sphn-schema/sphn#Quantity	valueuom	https://www.biomedit.ch/rdf/sphn-schema/sphn#hasUnit

CPA - examples 3

Services:

- row_id -> prev_service/curr_service
  - row_id (Healthcare Encounter)
  - -> TherapeuticArea
  - -> prev_service/curr_service (TherapeuticArea specialty Name)



	SERVICES	row_id	prev_service	https://biomedit.ch/rdf/sphn-schema/sphn#hasSpecialtyName	1
	SERVICES	row_id	curr_service	https://biomedit.ch/rdf/sphn-schema/sphn#hasSpecialtyName	1

SERVICES	row_id	https://biomedit.ch/rdf/sphn-schema/sphn#TherapeuticArea	https://biomedit.ch/rdf/sphn-schema/sphn#hasTherapeuticArea
SERVICES	https://biomedit.ch/rdf/sphn-schema/sphn#TherapeuticArea	prev_service	https://biomedit.ch/rdf/sphn-schema/sphn#hasSpecialtyName
SERVICES	row_id	https://biomedit.ch/rdf/sphn-schema/sphn#TherapeuticArea	https://biomedit.ch/rdf/sphn-schema/sphn#hasTherapeuticArea
SERVICES	https://biomedit.ch/rdf/sphn-schema/sphn#TherapeuticArea	curr_service	https://biomedit.ch/rdf/sphn-schema/sphn#hasSpecialtyName

CPA - examples 4

PRESCRIPTIONS:

- hadm_id (Administrative Case) -> icustay_id (SNOMED’s Intensive Care Unit)
  - hadm_id (Administrative Case)
  - -> Location
  - -> icustay_id (SNOMED’s Intensive Care Unit)

PRESCRIPTIONS	hadm_id	https://biomedit.ch/rdf/sphn-schema/sphn#Location	https://biomedit.ch/rdf/sphn-schema/sphn#hasLocation
PRESCRIPTIONS	https://biomedit.ch/rdf/sphn-schema/sphn#Location	icustay_id	https://biomedit.ch/rdf/sphn-schema/sphn#hasTypeCode

CEA

- Some entities have the same type, as the column.
- CEA (more detailed, than CTA) is not needed.

	B	C
	row_id	
	https://biomedit.ch/rdf/sphn-schema/sphn#DrugPrescription	cea
	130887	
	130888	
	130889	
	130890	
	130891	
	130892	
	130893	
	130894	

CEA

Entities can be extended by CEA (not just type of the column)

drug
https://biomedit.ch/rdf/sphn-schema/sphn#Drug
Influenza Virus Vaccine
Sodium Chloride 0.9% Flush
Dextrose 50%
Influenza Virus Vaccine
Senna
Docusate Sodium
Donepezil

NOT just SPHN's Drug

CEA

Entities can be detailed (not just type of the column)

drug
https://biomedit.ch/rdf/sphn-schema/sphn#Drug
Influenza Virus Vaccine
Sodium Chloride 0.9% Flush
Dextrose 50%
Influenza Virus Vaccine
Senna
Docusate Sodium
Donepezil

cea
https://pubchem.ncbi.nlm.nih.gov/substance/481101694
https://pubchem.ncbi.nlm.nih.gov/compound/5234
https://www.chemicalbook.com/ChemicalProductProperty_EN_CE
https://pubchem.ncbi.nlm.nih.gov/substance/481101694
https://pubchem.ncbi.nlm.nih.gov/compound/Senna
https://pubchem.ncbi.nlm.nih.gov/compound/23673837
https://pubchem.ncbi.nlm.nih.gov/compound/3152

CEA - examples 1

Prescriptions-drug: pubchem+chemicalbook

drug		
https://biomedit.ch/rdf/sphn-schema/sphn#Drug	**cea**	
Influenza Virus Vaccine	https://pubchem.ncbi.nlm.nih.gov/substance/481101694	
Sodium Chloride 0.9% Flush	https://pubchem.ncbi.nlm.nih.gov/compound/5234	
Dextrose 50%	https://www.chemicalbook.com/ChemicalProductProperty_EN_CB5250048.htm	
Influenza Virus Vaccine	https://pubchem.ncbi.nlm.nih.gov/substance/481101694	
Senna	https://pubchem.ncbi.nlm.nih.gov/compound/Senna	
Docusate Sodium	https://pubchem.ncbi.nlm.nih.gov/compound/23673837	
Donepezil	https://pubchem.ncbi.nlm.nih.gov/compound/3152	
Tamsulosin	https://pubchem.ncbi.nlm.nih.gov/compound/129211	
Calcitriol	https://pubchem.ncbi.nlm.nih.gov/compound/5280453	

CEA - examples 2

- ChartEvents-value: XMLSchema+SPHN
- dict+python code for implementation

value	value_cea
https://www.w3.org/2001/XMLSchema#string	
1	https://www.w3.org/2001/XMLSchema#double
1	https://www.w3.org/2001/XMLSchema#double
1	https://www.w3.org/2001/XMLSchema#double
Married	https://termbrowser.nhs.uk/?perspective=full&conceptId1=87915002&edition=uk-edition&release=v20240605&se
Jewish	https://browser.ihtsdotools.org/?perspective=full&conceptId1=186034007&edition=MAIN/2024-06-01&release=&l
Russian	https://browser.ihtsdotools.org/?perspective=full&conceptId1=186034007&edition=MAIN/2024-06-01&release=&l
4I	https://www.w3.org/2001/XMLSchema#string
Systolic	https://termbrowser.nhs.uk/?perspective=full&conceptId1=111973004&edition=uk-edition&release=v20240605&s
Calm/Cooperative	https://www.w3.org/2001/XMLSchema#string
No	https://termbrowser.nhs.uk/?perspective=full&conceptId1=260385009&edition=uk-edition&release=v20240605&s
Patient Verbalized	https://termbrowser.nhs.uk/?perspective=full&conceptId1=3457005&edition=uk-edition&release=v20240605&ser

CEA - examples 3

- InputEvents-categoryname: SPHN entities
- InputEvents-status description: SPHN entities

ordercategoryname	
https://browser.ihtsdotools.org/?perspective=full&c	
02-Fluids (Crystalloids)	https://browser.ihtsdotools.org/?perspective=full&conceptId1=255765007&
01-Drips	https://browser.ihtsdotools.org/?perspective=full&conceptId1=226898001&
01-Drips	https://browser.ihtsdotools.org/?perspective=full&conceptId1=226898001&
02-Fluids (Crystalloids)	https://browser.ihtsdotools.org/?perspective=full&conceptId1=255765007&
09-Antibiotics (Non IV)	https://browser.ihtsdotools.org/?perspective=full&conceptId1=255631004&
14-Oral/Gastric Intake	https://browser.ihtsdotools.org/?perspective=full&conceptId1=738956005&

statusdescription	
https://browser.ihtsdotools.org/?perspective=full&conceptId1=263490005	
Rewritten	https://browser.ihtsdotools.org/?perspective=full&conceptId1=263703002
Stopped	https://browser.ihtsdotools.org/?perspective=full&conceptId1=160895006
Stopped	https://browser.ihtsdotools.org/?perspective=full&conceptId1=160895006
Stopped	https://browser.ihtsdotools.org/?perspective=full&conceptId1=160895006
FinishedRunning	https://browser.ihtsdotools.org/?perspective=full&conceptId1=385669000
FinishedRunning	https://browser.ihtsdotools.org/?perspective=full&conceptId1=385669000

Application: Knowledge Graph

- Take either short or long CPA
- Convert annotated MIMIC to KG or directly extract triples and pairs
- KG generation options:
  - write a generator from scratch
  - RDFizer + RML
  - rdflib

Application: TransE

- **Baseline: TransE** - easy-to-train translation-based model
  - we can use multi-hop chains if we take extended CPA
- **Task: Classification**
- **Learning:**
  - minimizes the distance between (subject + relation) and (object)
  - run separately on reference data and on prediction data to calculate relational embeddings

Conclusions

Conclusions

Deliverables

- Annotated MIMIC dataset for one patient:
  - semi-automatically
  - cross-validated
  - for the remainder CEA has to be (automatically) extended
  - CTA and CPA are universal
- Direct and indirect column property annotation.

Additional

- Familiarization with modern tools and representation methods for knowledge graphs
- Gained experience in reviewing academic papers with relevant work
- Developed highly collaborative environment within the team

Limitations

- Complexity of CPA task
- No model

Recommendations for the future

Recommendations

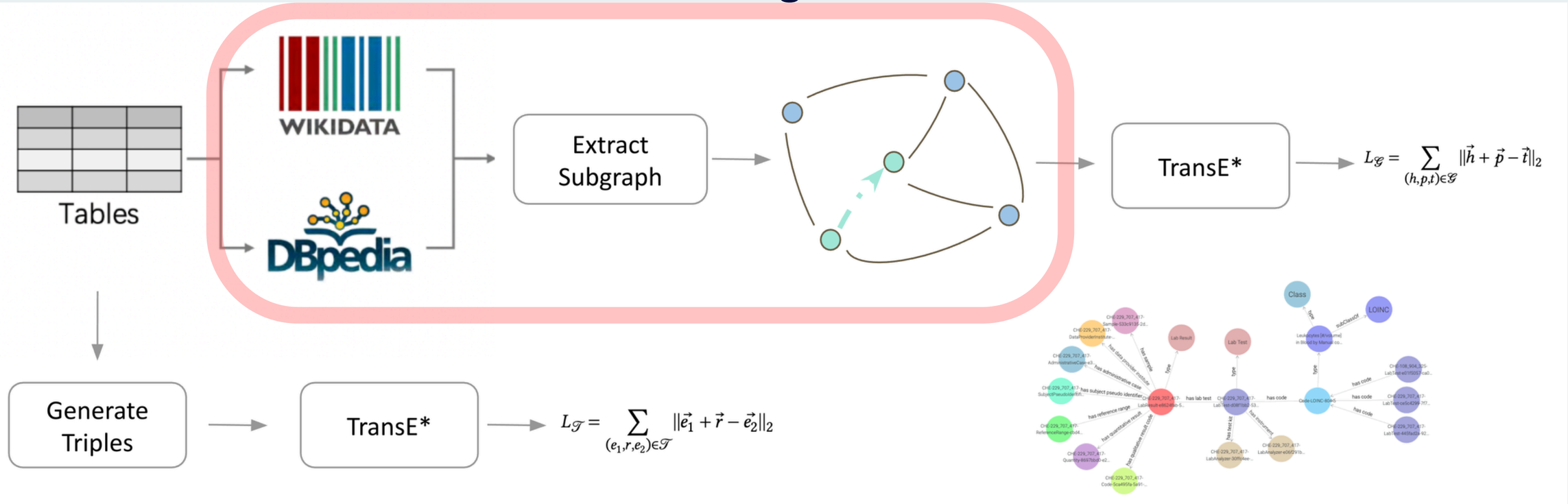
- Coming up with a more robust way for validating the annotations
- Extending the cross-validation to CEA and CPA taks
- Automate CEA to allow working with whole tables instead of a 1 patient subset
- Test with the model to see how certain annotation decisions (specifically CPA) affect the prediction quality

Code and Data Availability

https://github.com/TohaRhymes/auto_anno_um

Baseline

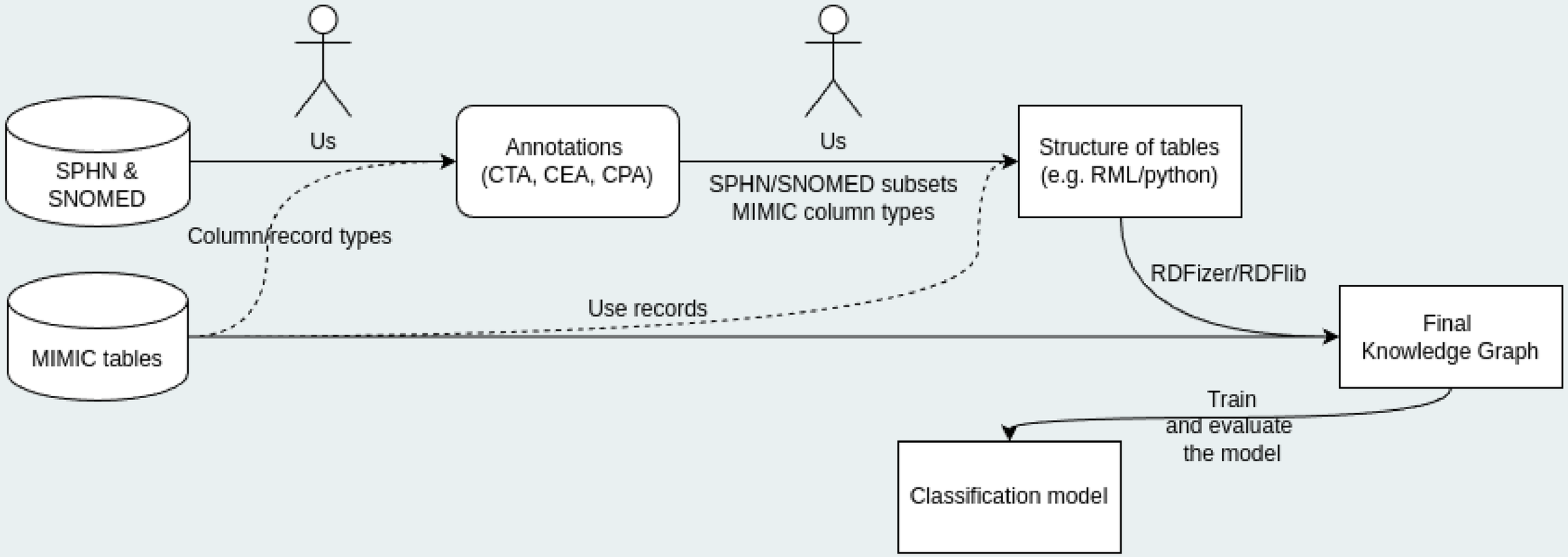
Use other ontologies



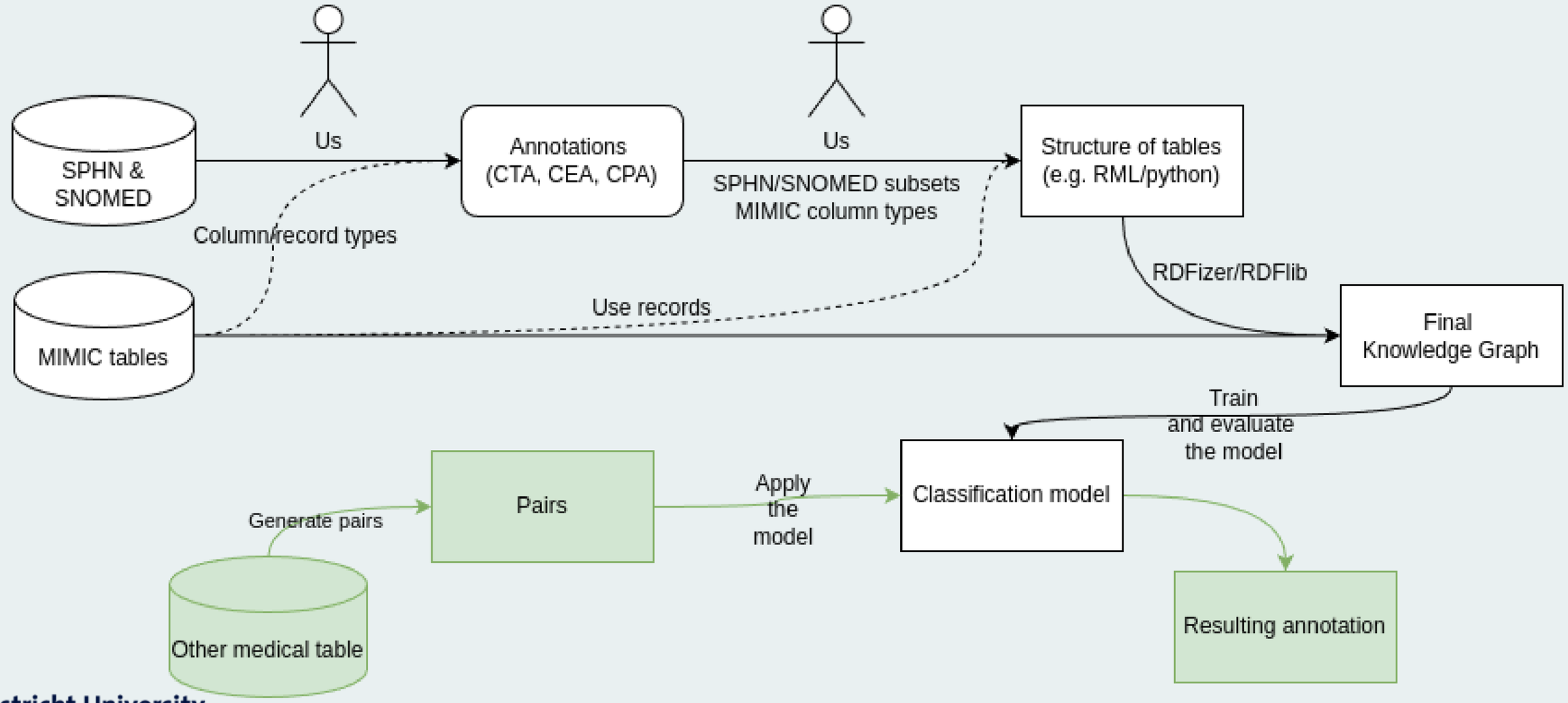
Try out Different DL algos and losses

SPHN/SNOMED+MIMIC

Complete final pipeline



Complete final pipeline



Code and Data Availability

https://github.com/TohaRhymes/auto_anno_um

References

[1] Abhisek Sharma, Sumit Dalal, and Sarika Jain. “SemInt at SemTab 2022”. In: SemTab@ISWC. 2022. url: https://api.semanticscholar.org/CorpusID:255943679.

[2] Shervin Mehryar and Remzi Celebi. “Semantic Annotation of Tabular Data for Machine-to-Machine Interoperability via Neuro-Symbolic Anchoring”. In: SemTab’23: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching 2023, co-located with the 22nd International Semantic Web Conference (ISWC). Corresponding author: Shervin Mehryar (shervin.mehryar@maastrichtuniversity.nl). CEUR Workshop Proceedings. Athens, Greece, Nov. 2023. url: http://ceur-ws.org/Vol-3557/paper5.pdf.

[3] Antoine Bordes et al. “Translating Embeddings for Modeling Multi-relational Data”. In: Advances in Neural Information Processing Systems (NIPS). Curran Associates, Inc. Lake Tahoe, Nevada, 2013. [4] Alberto Garc´ıa-Dur´an, Antoine Bordes, and Nicolas Usunier. “Composing Relationships with Translations”. In: EMNLP 2015 (2015).

[5] A. Johnson, T. Pollard, and R. Mark. MIMIC-III Clinical Database Demo. Version 1.4. 2019. doi: 10.13026/C2HM2Q. url: https://doi.org/10.13026/C2HM2Q.

[6] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: arXiv preprint arXiv:1810.04805 (2018).

[7] Victor Sanh et al. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In: arXiv preprint arXiv:1910.01108 (2019).

[8] S¨oren Auer et al. “DBpedia: A Nucleus for a Web of Open Data”. In: The Semantic Web (2007), pp. 722–735.

[9] Eiglesias34 et al. SDM-TIB/SDM-RDFizer: 4.1.1. Version v4.1.1. Software available from Zenodo. Feb. 2022. doi: 10.5281/zenodo.6225573. url: https://doi.org/10.5281/zenodo.6225573.

[10] Jason Slepicka et al. “KR2RML: An Alternative Interpretation of R2RML for Heterogeneous Sources”. In: University of Southern California, Information Sciences Institute and Department of Computer Science, USA (2022). url: https://github.com/usc- isi-i2/Web-Karma.

[11] Nora Abdelmageed and Sirko Schindler. fusion-jena/JenTab: KGCW 2021. Version KGCW2021. Software available from Zenodo. Apr. 2021. doi: 10.5281/zenodo.4730314. URL: https://doi.org/10.5281/zenodo.4730314.

[12] Dylan Van Assche et al. KGCW 2024 Challenge @ ESWC 2024. Version 0.9.0. Zenodo, Mar. 2024. doi: 10.5281/zenodo.10721875. url: https://doi.org/10.5281/zenodo.10721875.

[13] Shervin Mehryar and Remzi Celebi. “Improving Transitive Embeddings in Neural Reasoning Tasks via Knowledge-Based Policy Networks”. In: ISWC 2022. University of Toronto; Maastricht University. Toronto, Canada; Maastricht, The Netherlands, 2022.

[14] Junheng Hao et al. Universal Representation Learning of Knowledge Bases by Jointly Embedding Instances and Ontological Concepts. 2021. arXiv: 2103.08115 [cs.CL].8

Questions?

Maastricht University