

# Semantic Understanding Of Tables (Symbolic)

Pedro Szekely

[szekely@usc.edu](mailto:szekely@usc.edu)

USC Information Sciences Institute

# Outline

Introduction to semantic modeling

Semantic modeling languages

Interactive system for semantic modeling (Karma)

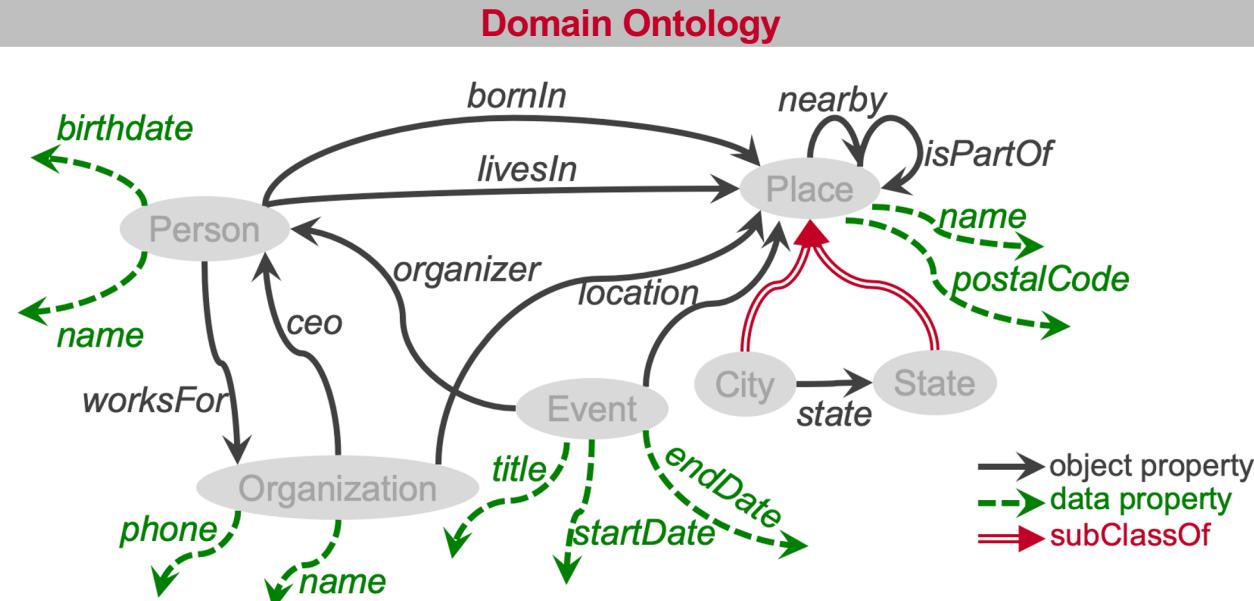
Automated system for semantic modeling

Entity linking for tables and related tasks

# Semantic Modeling

Problem definition

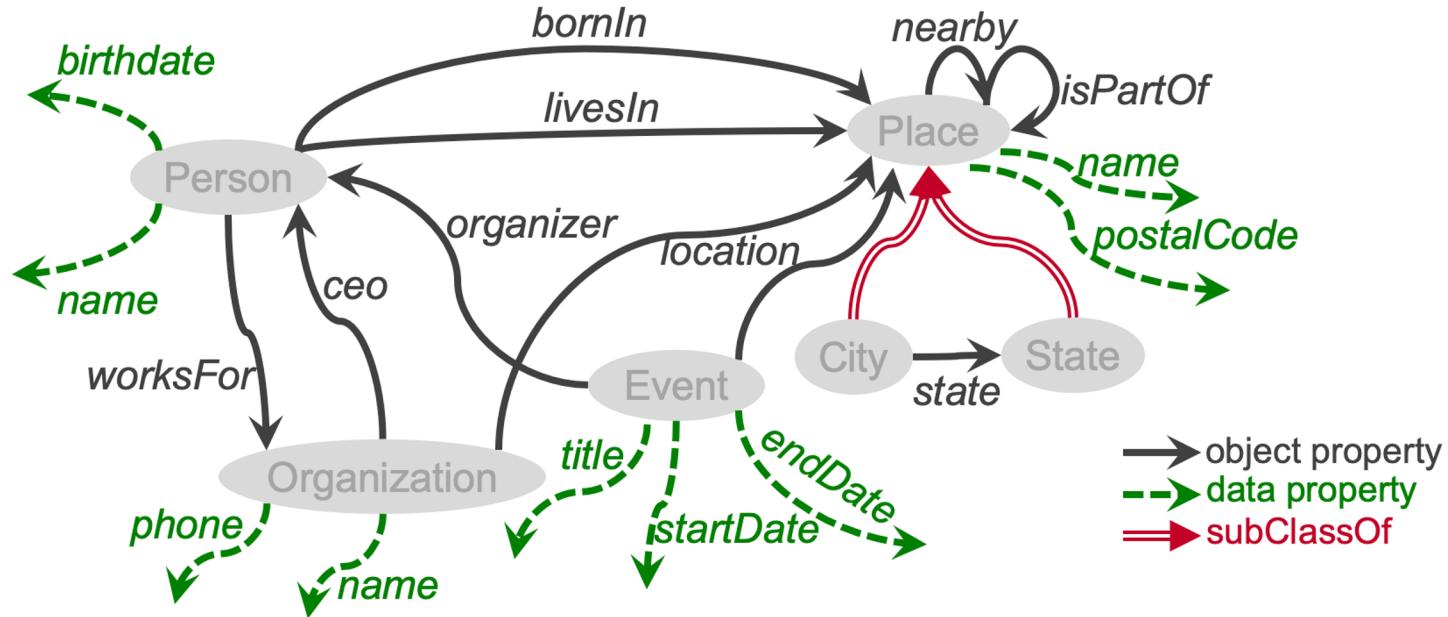
# Inputs: Table and Ontologies



**Table**

Column 1	Column 2	Column 3	Column 4	Column 5
Bill Gates	Oct 1955	Microsoft	Seattle	WA
Mark Zuckerberg	May 1984	Facebook	White Plains	NY
Larry Page	Mar 1973	Google	East Lansing	MI

## Domain Ontology

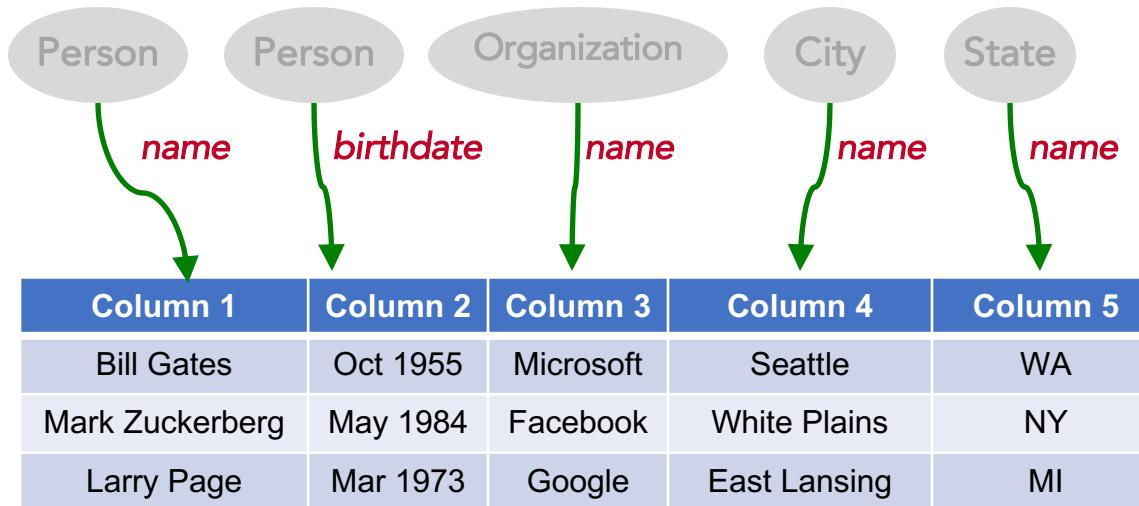


## Table

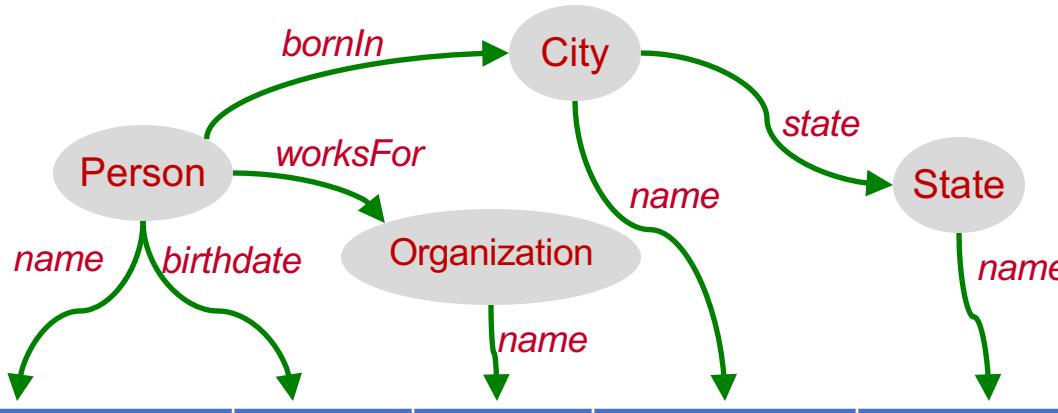
Column 1	Column 2	Column 3	Column 4	Column 5
Bill Gates	Oct 1955	Microsoft	Seattle	WA
Mark Zuckerberg	May 1984	Facebook	White Plains	NY
Larry Page	Mar 1973	Google	East Lansing	MI

Semantic Model = Semantic Types + Relationships

# Semantic Types

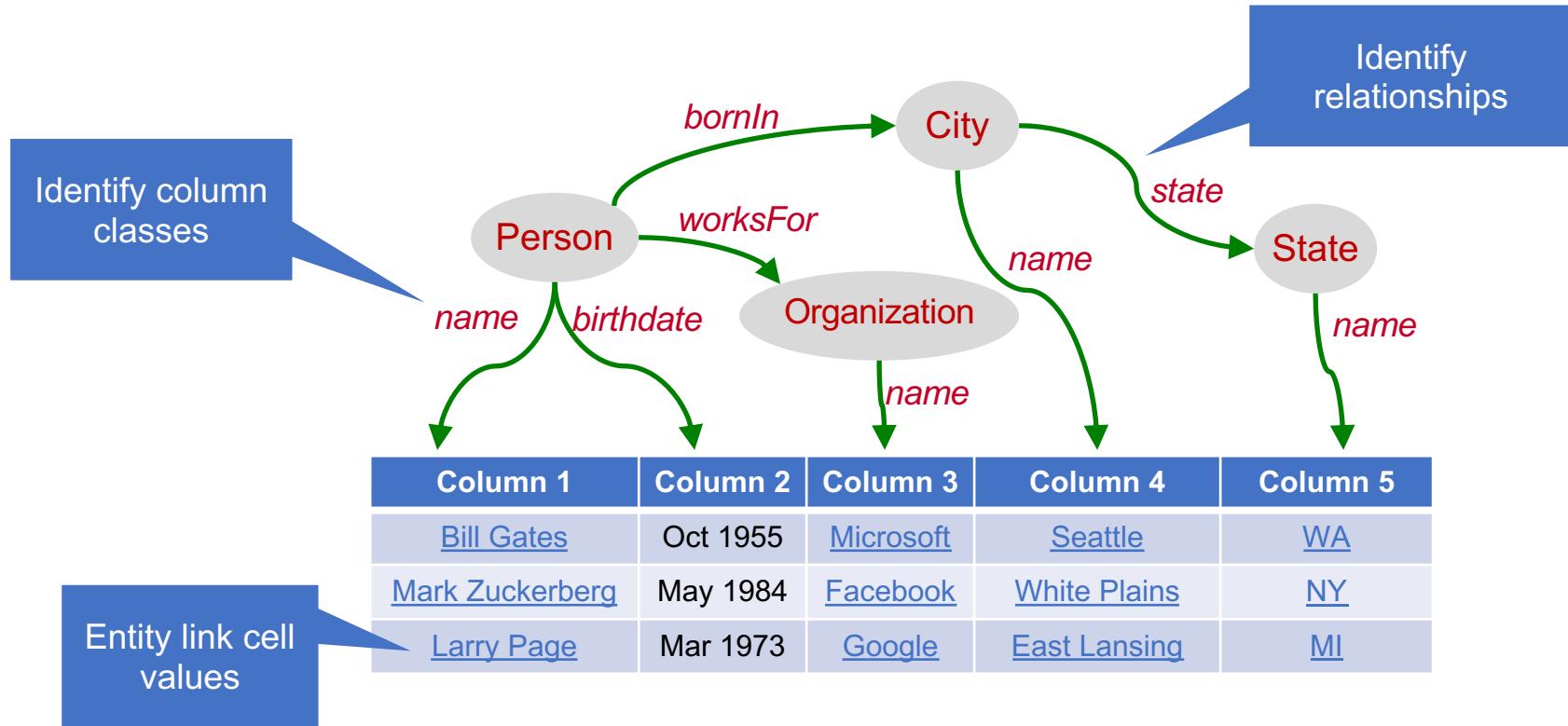


# Relationships



Column 1	Column 2	Column 3	Column 4	Column 5
Bill Gates	Oct 1955	Microsoft	Seattle	WA
Mark Zuckerberg	May 1984	Facebook	White Plains	NY
Larry Page	Mar 1973	Google	East Lansing	MI

# Subtasks of semantic modeling



# Semantic Modeling Languages

Relational to RDF Mapping Language: <https://www.w3.org/TR/r2rml/>

RDF Mapping Language: <https://rml.io/specs/rml/>

# Languages to specify semantic models

## R2RML

a language for expressing customized mappings from relational databases to RDF datasets (2012)

<https://www.w3.org/TR/r2rml/>

## RML

a generic mapping language, based on and extending R2RML ... adding support for data in other structured formats (2020)

<https://rml.io/specs/rml/>

# RML Example

Example input data

```
{  
  "venue":  
  {  
    "latitude": "51.050000",  
    "longitude": "3.7166700"  
  },  
  "location":  
  {  
    "continent": " EU",  
    "country": "BE",  
    "city": "Brussels"  
  }  
}
```

RML specification  
written in RDF (turtle)

```
<#VenueMapping>  
  rml:logicalSource [  
    rml:source "http://www.example.com/files/Venue.json";  
    rml:referenceFormulation ql:JSONPath;  
    rml:iterator "$"  
  ];  
  
  rr:subjectMap [  
    rr:template "http://loc.example.com/city/{$ .location.city}";  
    rr:class schema:City  
  ];
```

Identify column  
classes

```
<http://loc.example.com/city/Brussels> rdf:type schema:City.  
<http://loc.example.com/city/Brussels> wgs84_pos:lat "50.901389".  
<http://loc.example.com/city/Brussels> wgs84_pos:long "4.484444".  
<http://loc.example.com/city/Brussels> gn:countryCode "BE".
```

# RML relationship mapping example

- subjectMap, predicateObjectMap: rules to get values for subject, predicate and object



# Karma

## Interactive system for semantic modeling

Szekely P. et al. (2013) Connecting the Smithsonian American Art Museum to the Linked Data Cloud. In: The Semantic Web: Semantics and Big Data. ESWC 2013. Lecture Notes in Computer Science, vol 7882. Springer, Berlin, Heidelberg.

Taheriyani M., Knoblock C.A., Szekely P., Ambite J.L. (2012) Rapidly Integrating Services into the Linked Data Cloud. In: Cudré-Mauroux P. et al. (eds) The Semantic Web – ISWC 2012. ISWC 2012. Lecture Notes in Computer Science, vol 7649. Springer, Berlin, Heidelberg.

<https://github.com/usc-isi-i2/Web-Karma>

# Interactive semantic modeling

museum-data.json ✓  
Name: museum-data.json | Prefix: s | Base URI: http://localhost:8080/source/

E22\_Man-Made\_Object1

P1\_is\_identified\_by

E42\_Identifier1

P102\_has\_title

E35\_Title1

P128\_carries

E73\_Information\_Object1

P72\_has\_language

E56\_Language1

doc

descriptiveNonRepeating ▾

record_ID ▾	record_identifier_uri ▾ C	record_uri ▾ C	title ▾	content ▾	record_title_uri ▾ C	information_object_uri ▾ C
-------------	---------------------------	----------------	---------	-----------	----------------------	----------------------------

indexedStructured ▾

language ▾	language_uri ▾ C	values ▾
------------	------------------	----------

siris_sil_73088	object/siris_sil_73088...	object/siris_sil_73088	Everyday Lakota : an English-Sioux dictionary for beginners / General	object/siris_sil_73088...	record/siris_sil_73088...	thesauri/language/da... Dakota language
siris_arc_87459	object/siris_arc_8745...	object/siris_arc_87459	Letter in English and Dakota	object/siris_arc_8745...	record/siris_arc_8745...	thesauri/language/da... English language

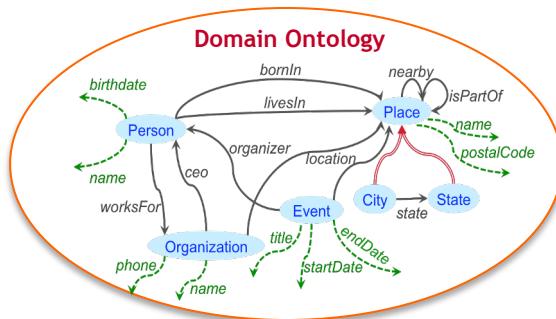
# Learning Semantic Types

Requirements:

Learn from a small number of examples

Distinguish string and numeric values

Scale to many semantic types



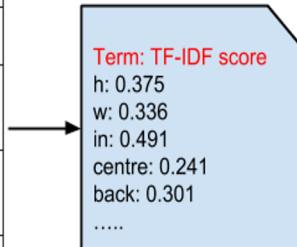
	name	date	city	state	workplace
1	Fred Collins	Oct 1959	Seattle	WA	Microsoft
2	Tina Peterson	May 1980	New York	NY	Google

# Learning Semantic Types For Text

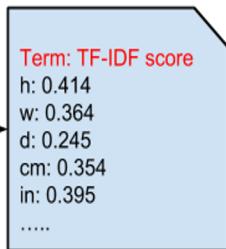
Textual  
Data

Treat each column of data as a document  
Apply TF-IDF Cosine Similarity

Dimensions
H: 3.5 in, W: 2.5 in
H: 64 in, W: 51.5 in, D: .75 in
L: 57 in, center back: 23 in
...



Extent
52.1 x 71.4 cm (20 1/2 x 28 1/8 in.)
9 3/4 x 7 9/16 in.
H: 19 x W: 15 1/4 x D: 8 1/4 in.
...



TF-IDF  
Cosine  
Similarity

$$tf(t, d) = \text{frequency}^{1/2}$$

$$idf(t) = 1 + \log\left(\frac{\text{numDocs}}{\text{docFreq} + 1}\right)$$

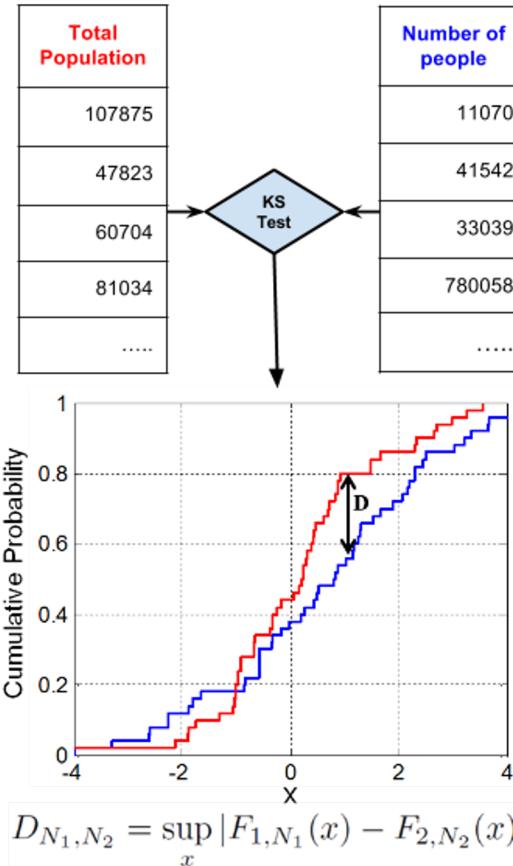
$$\text{sim}(q, d) = \frac{V(q).V(d)}{|V(q)|.|V(d)|}$$

# Learning Numeric Semantic Types

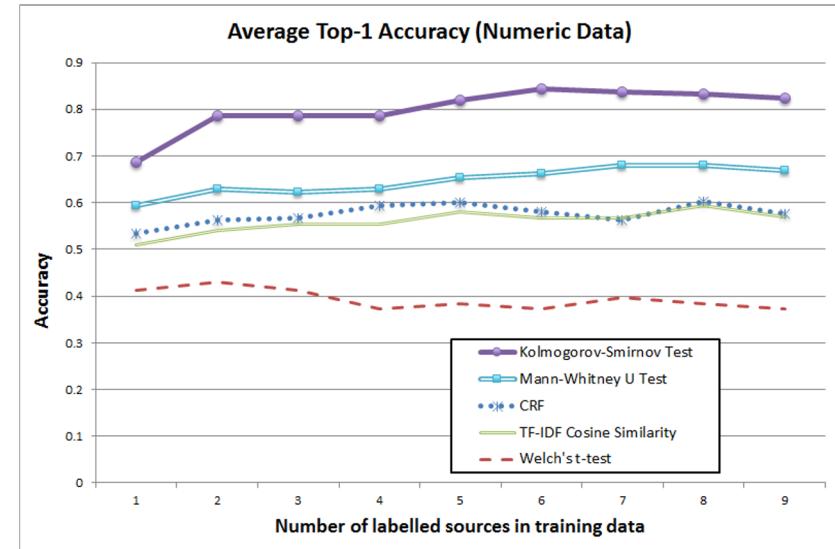
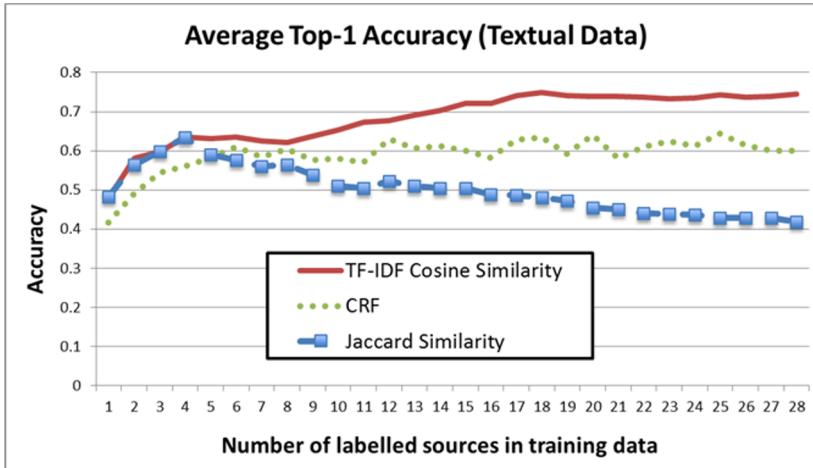


Apply statistical hypothesis testing to determine which distribution fits best

Apply Kolmogorov-Smirnov Test



# Evaluation

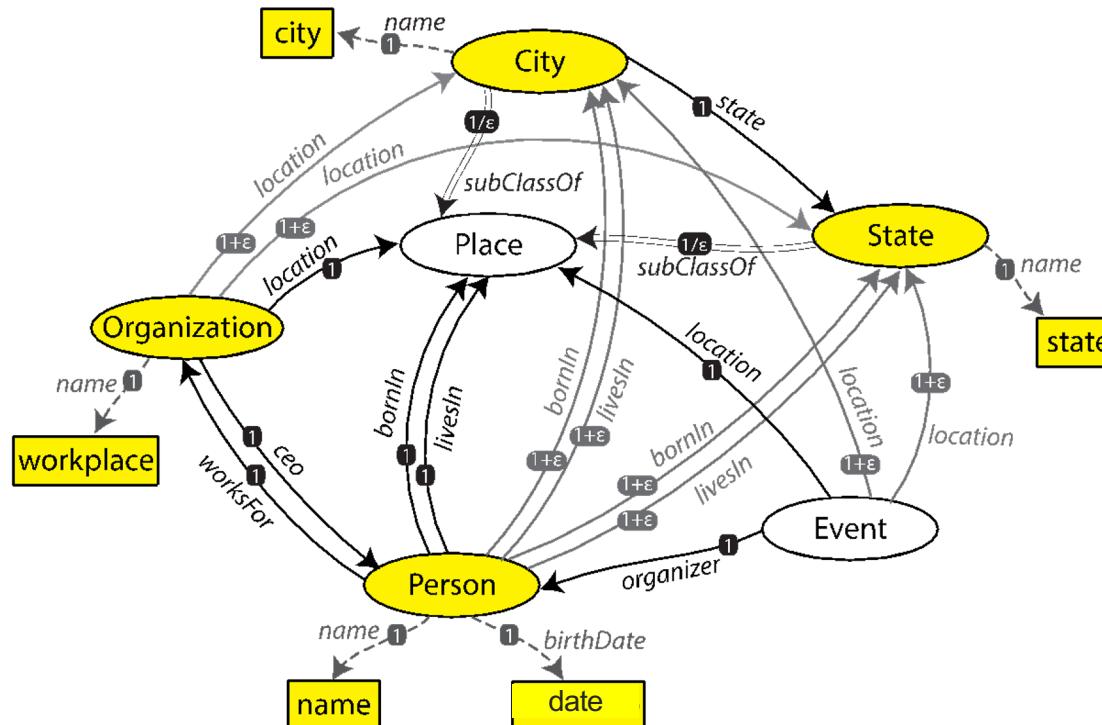


Combined approach achieves 97% accuracy on the top-4 accuracy  
Suggestions shown in KARMA GUI

# Determining relationships

after semantic types are defined

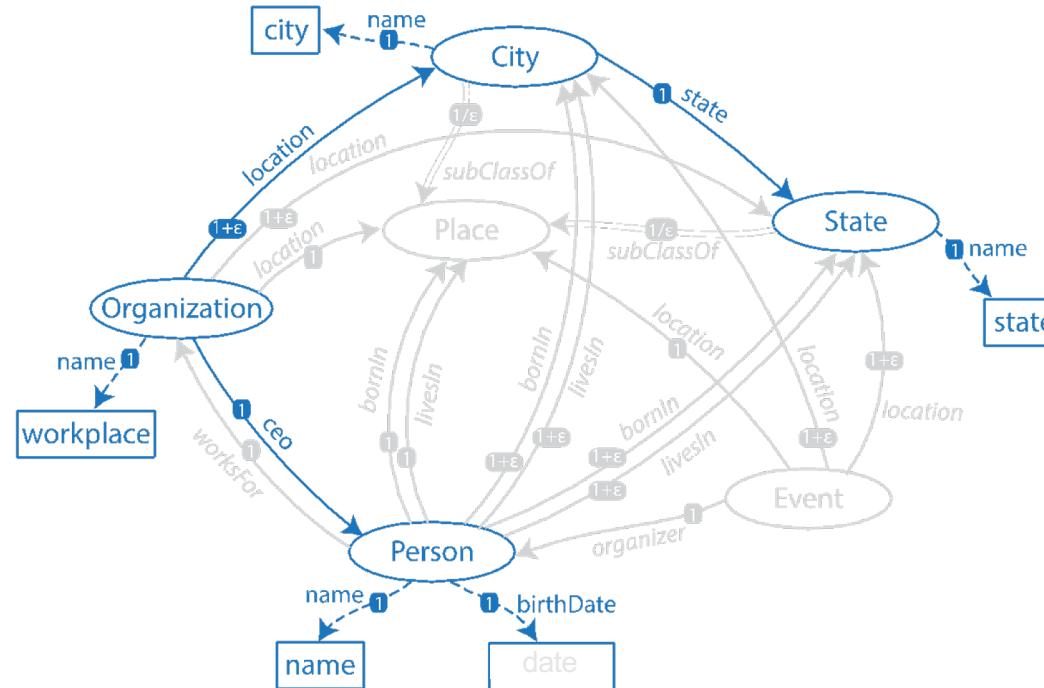
Construct a graph from semantic types and ontology



# Determining relationships

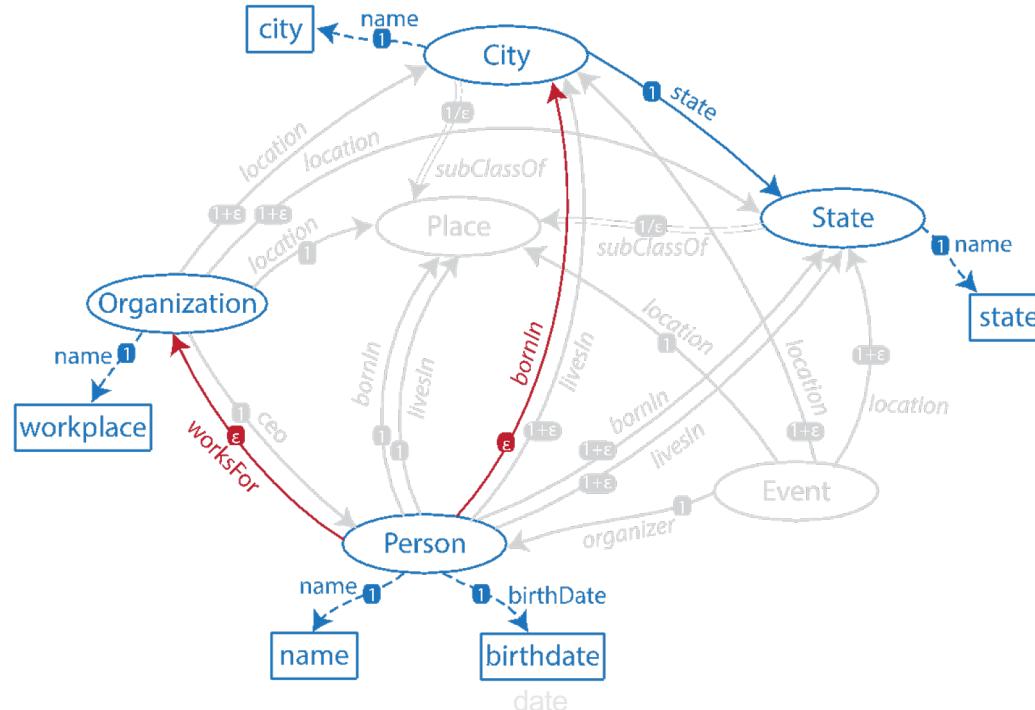
Select minimal tree that connects all semantic types

A customized **Steiner tree algorithm** [Kou & Markowsky, 1981]



# Refining the semantic model

Use the GUI to impose constraints on Steiner Tree Algorithm



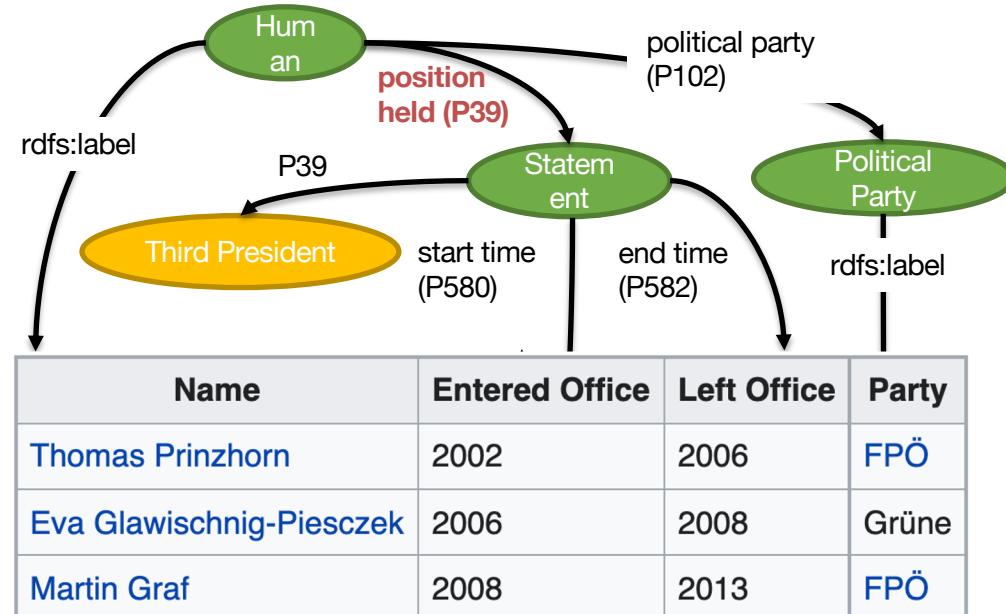
# GRAMS

## Inferred Semantic Descriptions of Wikipedia Tables

Binh Vu, Craig A. Knoblock, Pedro Szekely, Minh Pham, & Jay Pujara (2021). A Graph-based Approach for Inferring Semantic Descriptions of Wikipedia Tables. In ISWC 2021 - 20th International Semantic Web Conference.

# Challenges in Modeling Web Tables

- semantic model includes information not in the table
  - Third President* is in the table context
- n-ary relationships
  - position held P39*



Third Presidents of National Council (Austria)

# Main idea: entity linking

Name	Entered Office	Left Office	Party
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

## Challenges

- incomplete KG
- table/kg discrepancies
- multi-hop

**Eva Glawischnig-Piesczek** (Q93870)

---

Austrian politician

edit

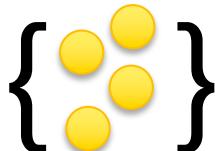
member of political party	Die Grünen
position held	Third President of the National Council of Austria
start time	30 October 2006
end time	28 October 2008

# Approach

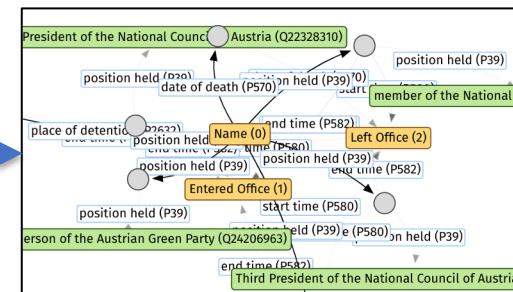


Name	Entered Office	Left Office	Party
Willi Brauneder	1996	1999	FPÖ
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

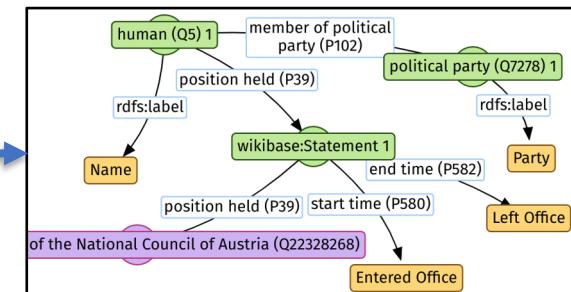
Linked table



Contextual values



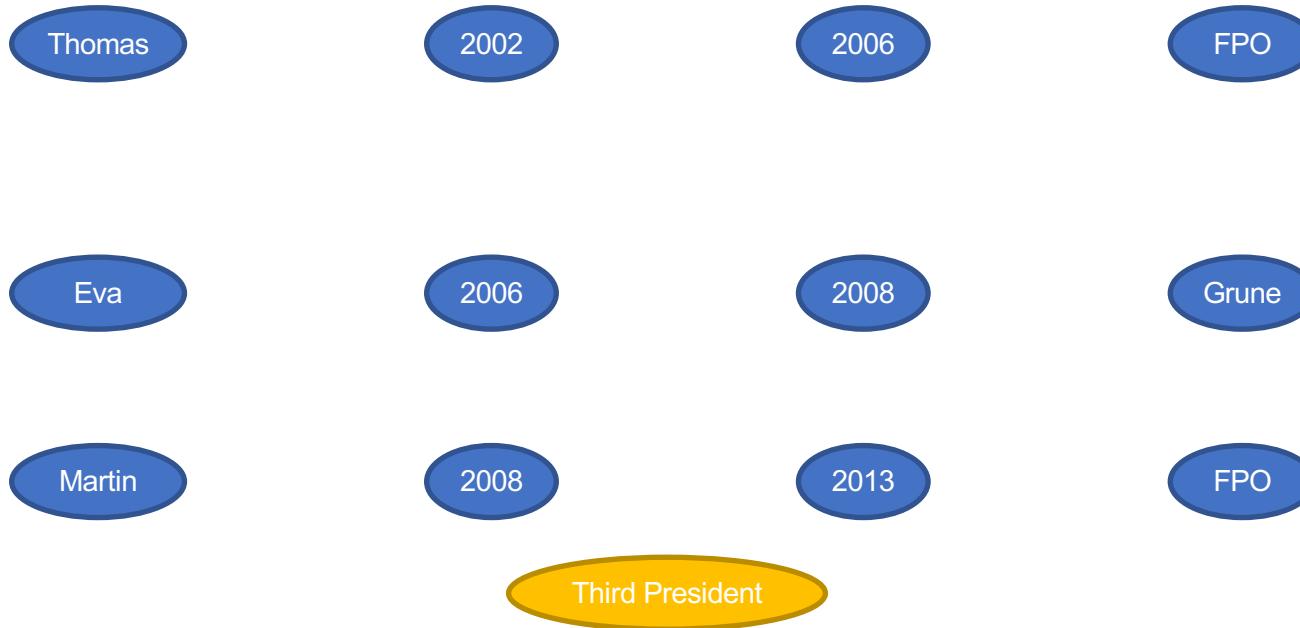
Candidate Graph



Semantic Description

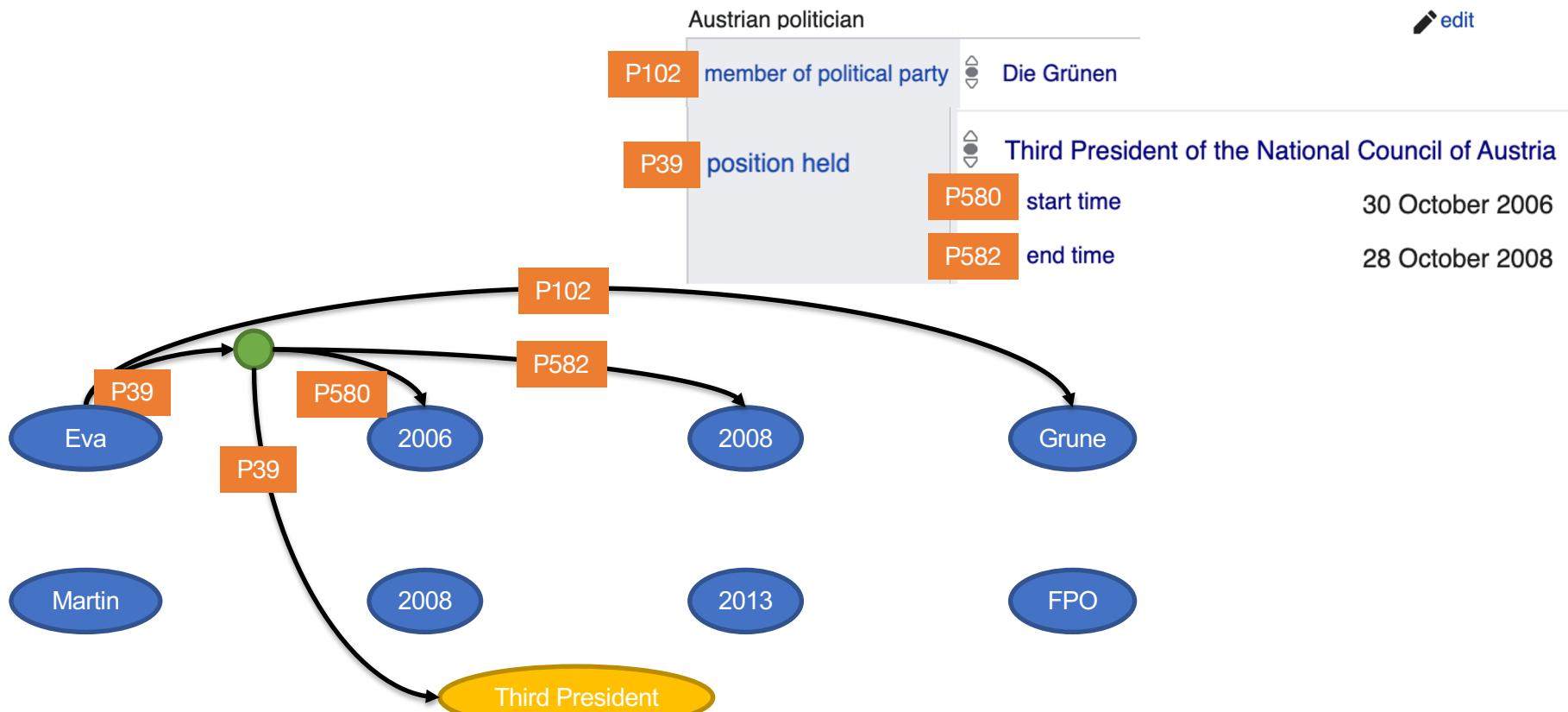
# Create a graph of cells and context

Name	Entered Office	Left Office	Party
Willi Brauneder	1996	1999	FPÖ
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

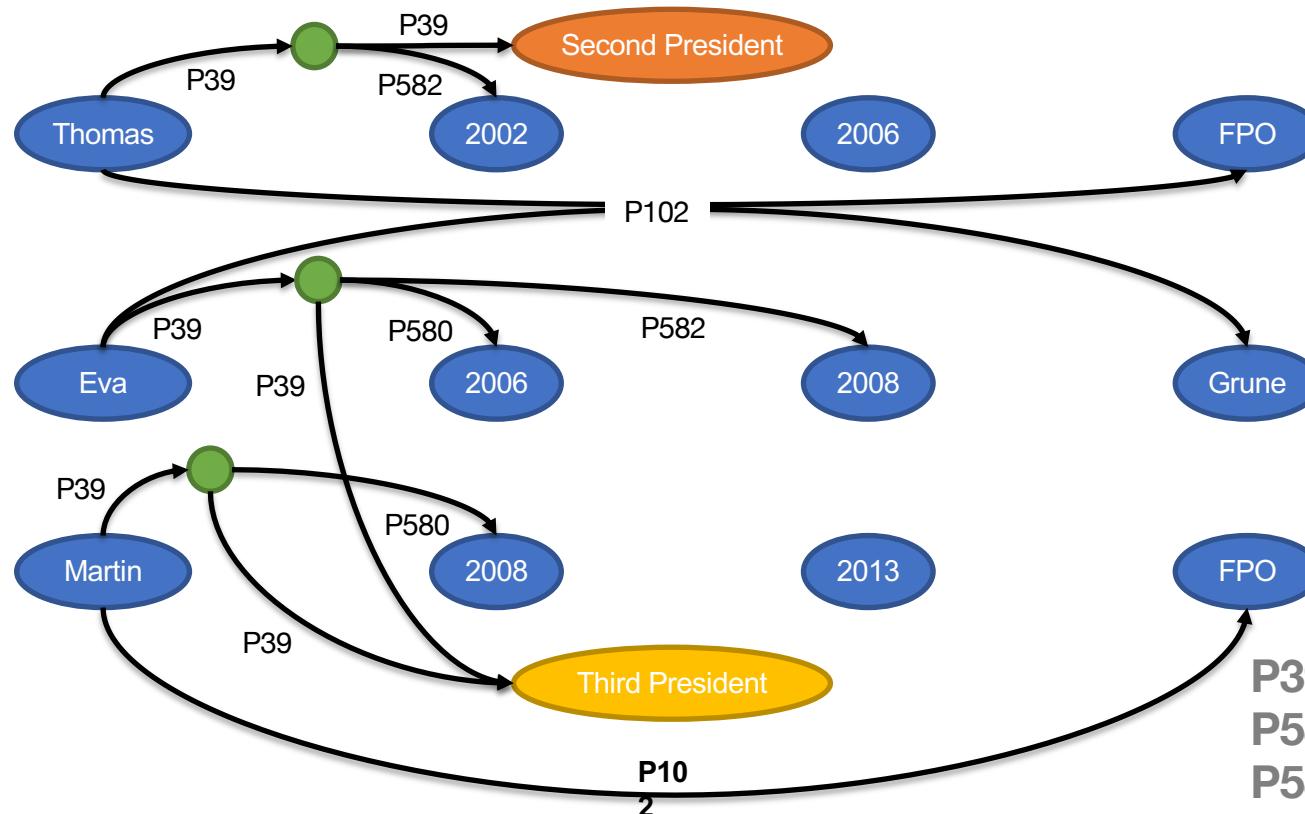


# Construct Candidate Graph: Discover Links

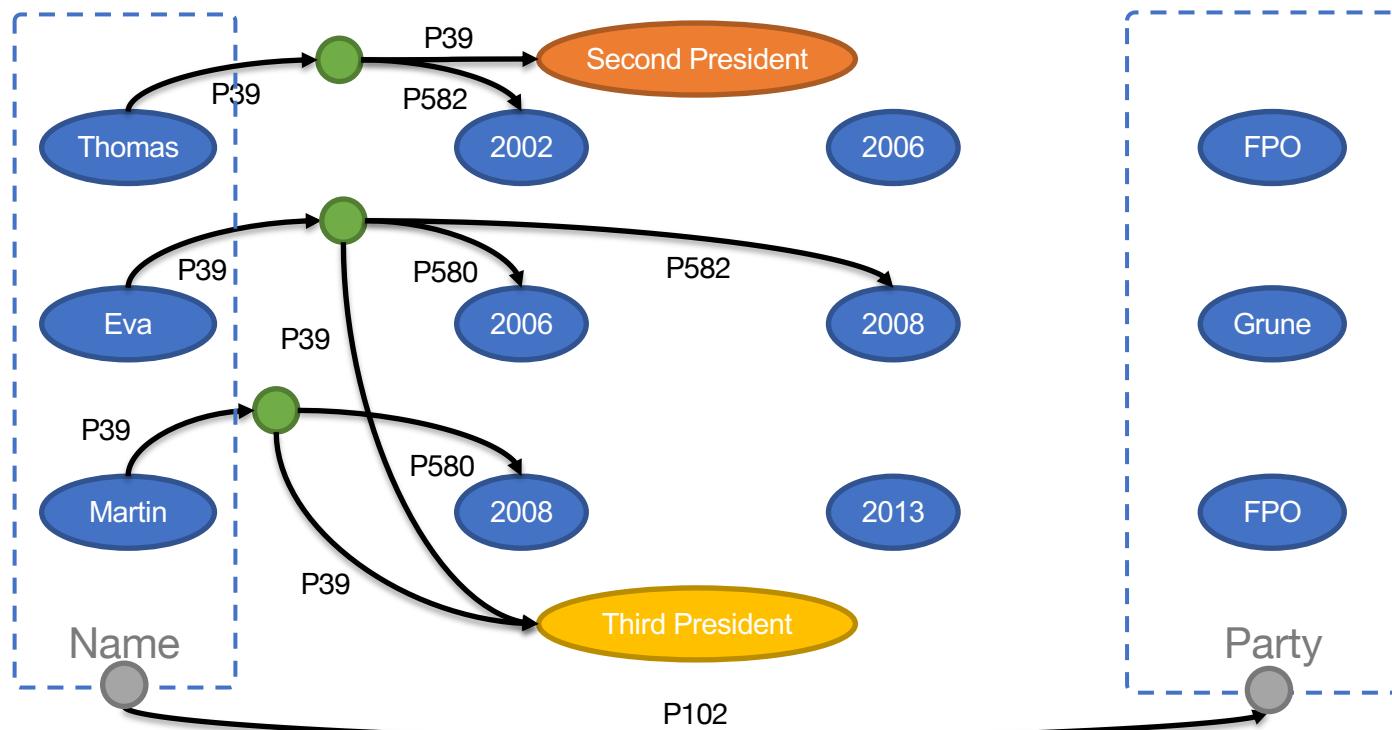
Eva Glawischnig-Piesczek (Q93870)



# Group links from same source & target columns

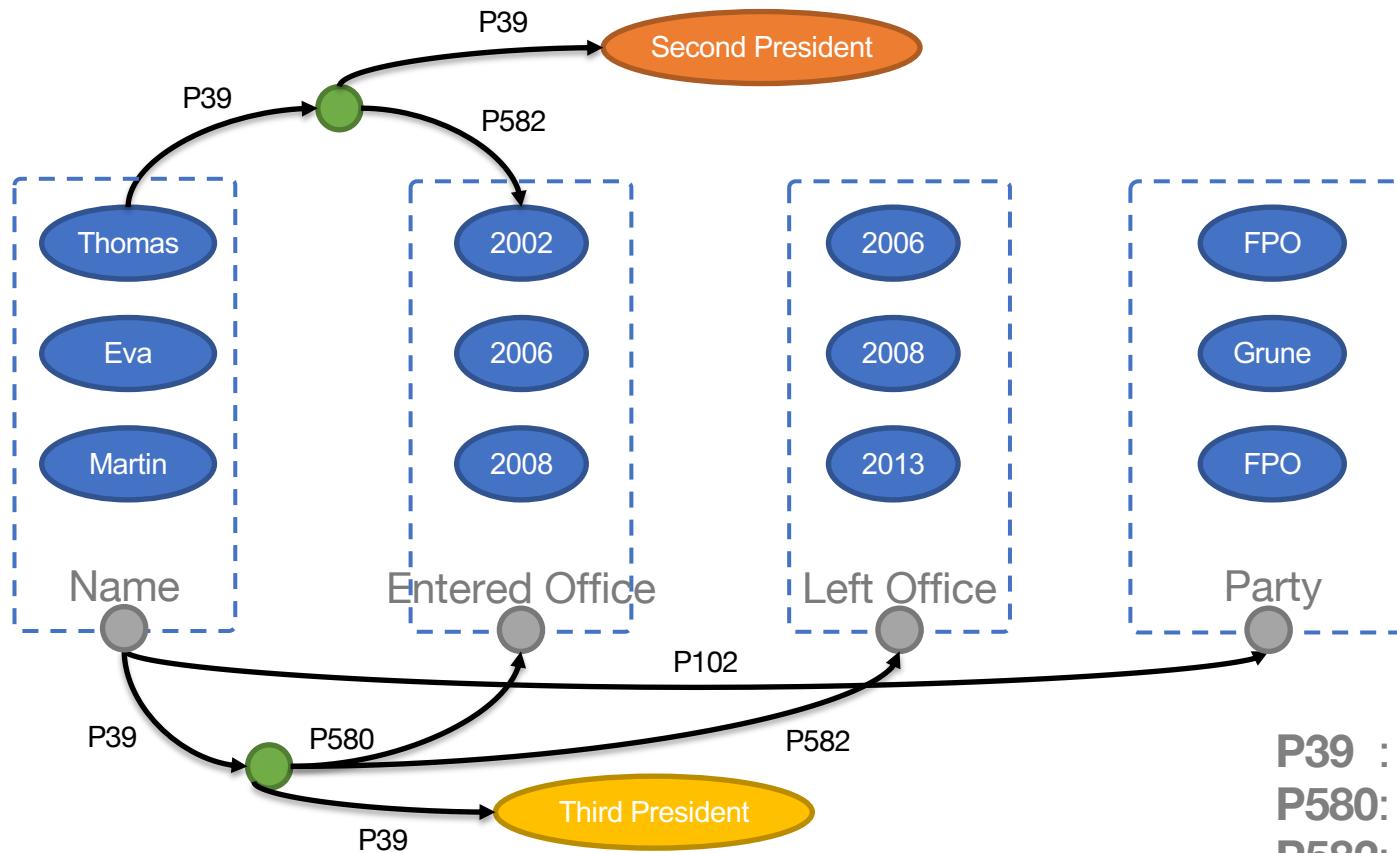


# Construct Candidate Graph: Summarization

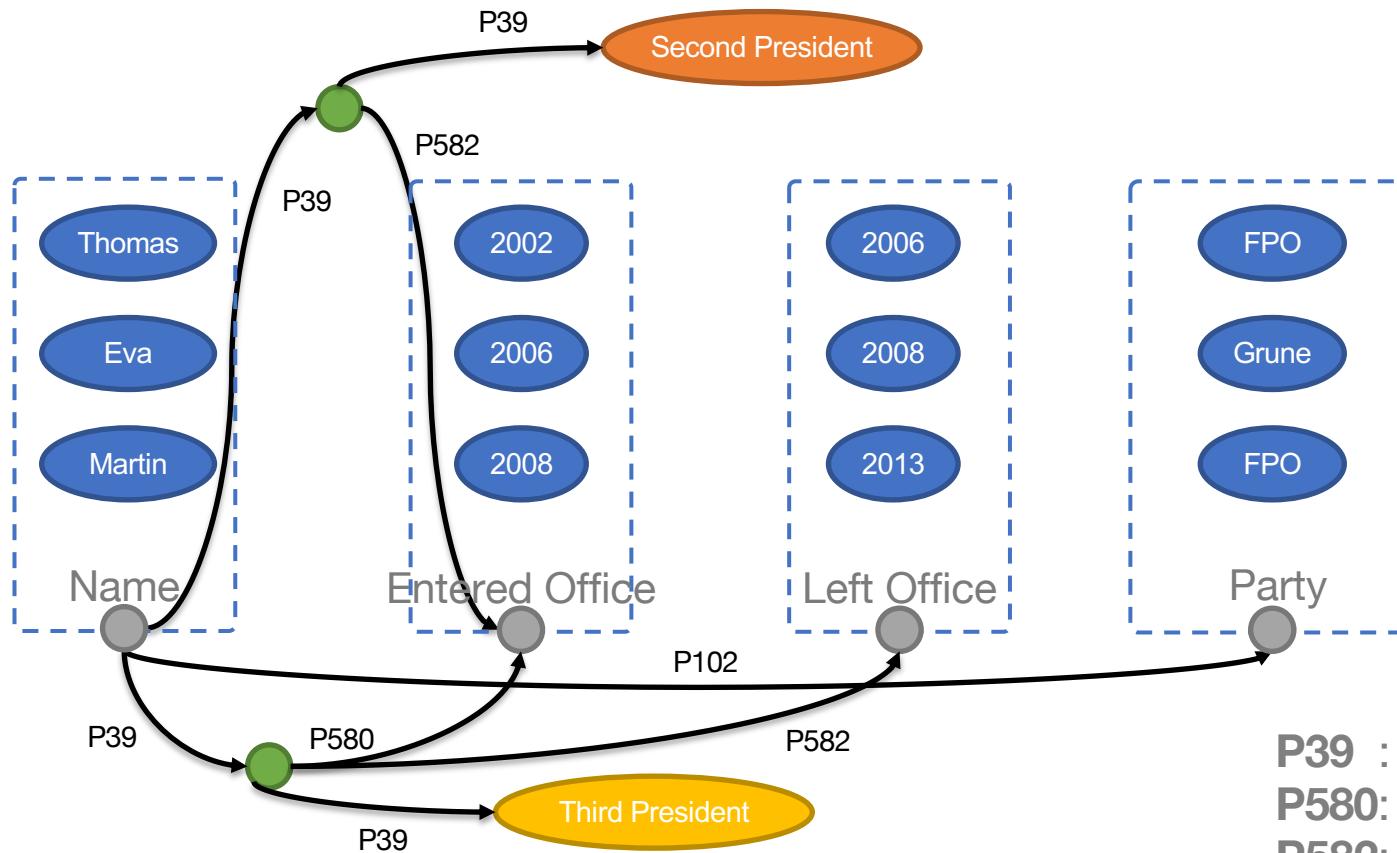


P39 : position held  
P580: start time  
P582: end time

# Construct Candidate Graph: Summarization



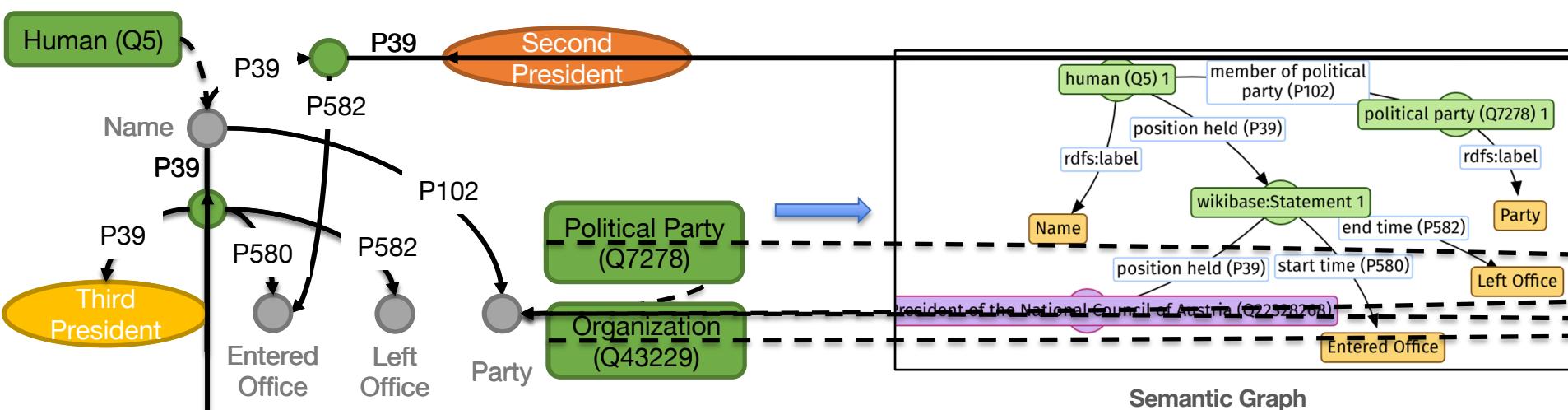
# Construct Candidate Graph: Summarization



P39 : position held  
P580: start time  
P582: end time

# Building semantic models from candidate graphs

- Candidate (n-ary) relationships *from the candidate graph*
  - Candidate columns' types *from entities in table columns*
- ⇒ Need to select the most appropriate relationships and types.



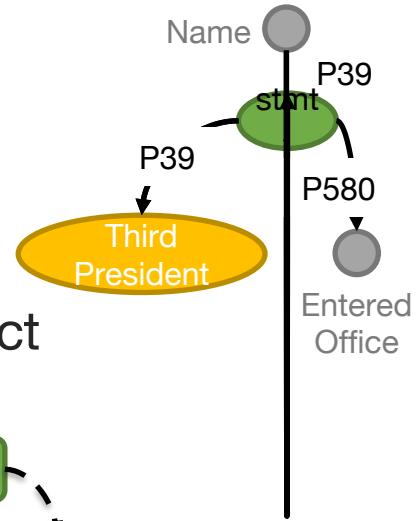
# Collective reasoning problem

- Probabilistic Soft Logic (PSL)
- Define predicates
- Define rules

# PSL Predicates (examples)

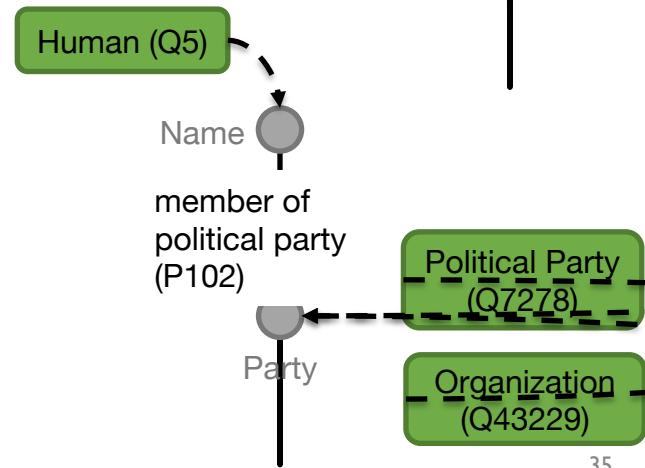
**CorrectRel( $N_1$ ,  $N_2$ ,  $P$ ):** if a relationship is correct

- $\text{CorrectRel}(\text{Name}, \text{stmt}_1, \text{P39})$
- $\text{CorrectRel}(\text{stmt}_1, \text{Third President}, \text{P39})$
- $\text{CorrectRel}(\text{stmt}_1, \text{Entered Office}, \text{P580})$



**CorrectType( $N_1$ ,  $T$ ):** if a column type assignment is correct

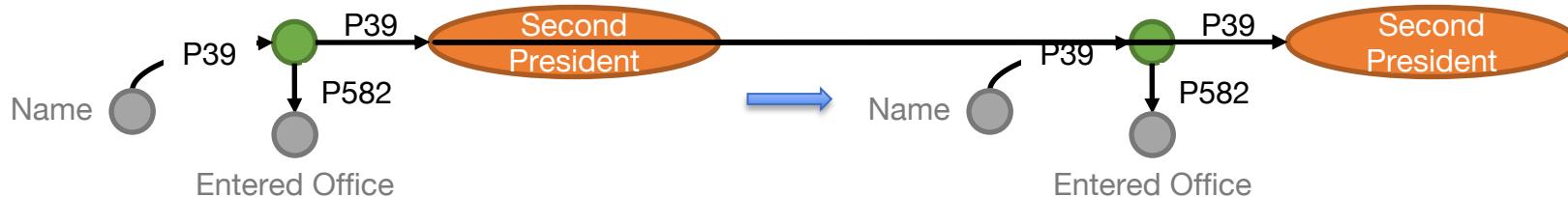
- $\text{CorrectType}(\text{Name}, \text{Human})$
- $\text{CorrectType}(\text{Party}, \text{Organization})$
- $\text{CorrectType}(\text{Party}, \text{Political Party})$



P39 : position held    P580: start time    P582: end time

# PSL Rules (examples)

If a statement value is incorrect, then the statement's qualifiers are also incorrect



Prefer fine-grained properties to high-level properties



# Evaluation of GRAM

Collective reasoning is beneficial

- Avoids cascading errors from subject column detection phase
- Handles columns with multiple entity types, and n-ary relationships

Dataset	Method	CPA			CTA		
		Precision	Recall	F <sub>1</sub>	Precision	Recall	F <sub>1</sub>
250WT	MantisTable	0.535	0.442	0.484	0.928	0.331	0.488
	MantisTable*	0.559	0.569	0.564	<b>0.940</b>	0.394	0.556
	BBW	0.796	0.123	0.214	0.850	0.233	0.367
	BBW*	0.740	0.559	0.638	0.759	0.777	0.768
	GRAMS-ST	0.526	0.681	0.594	-	-	-
	GRAMS	<b>0.824</b>	<b>0.650</b>	<b>0.726</b>	0.819	<b>0.813</b>	<b>0.816</b>
SemTab2020	MantisTable	0.985	0.976	0.981	0.977	0.800	0.880
	BBW	<b>0.996</b>	<b>0.995</b>	<b>0.995</b>	0.980	0.980	0.980
	GRAMS-ST	0.990	0.989	0.990	-	-	-
	GRAMS	<b>0.996</b>	0.994	<b>0.995</b>	<b>0.982</b>	<b>0.981</b>	<b>0.982</b>

Wikipedia  
tables

Synthetic  
tables

MantisTable\* and BBW\* are modified to retrieve correct subject column

# Entity Linking For Tables

Nguyen, Phuc, et al. "MTab4Wikidata at SemTab 2020: Tabular Data Annotation with Wikidata." SemTab@ ISWC. 2020.

Huynh, Viet-Phi, et al. "DAGOBAH: Enhanced Scoring Algorithms for Scalable Annotations of Tabular Data." SemTab@ ISWC. 2020.

Abdelmageed, Nora, and Sirko Schindler. "JenTab: A Toolkit for Semantic Table Annotations." (2021).

Cremaschi, Marco, Roberto Avogadro, and David Chieregato. "MantisTable: an Automatic Approach for the Semantic Table Interpretation." SemTab@ ISWC 2019 (2019): 15-24.

# Task Definitions

Column Entity Annotation  
(CEA)


entity



Column Table Annotation  
(CTA)


class



Column Property Annotation  
(CPA)


property

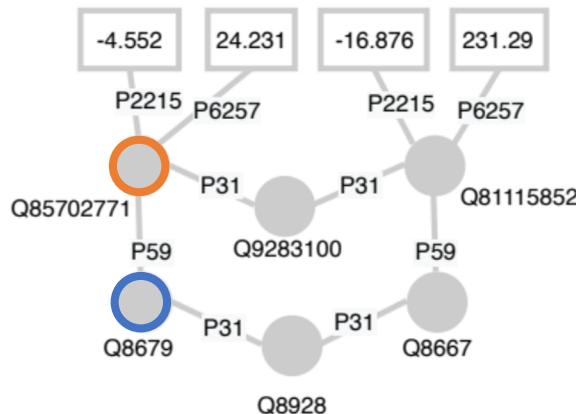


Usually solved together


property class property



col0	col1	col2	col3
V*!AY Psc	-4.593	Pisces	24.280
SDSS J153509.57+360054.5	-17.028	Boötes	232.909



#### Entities labels:

- Q85702771: V\* AY Psc
- Q81115852: SDSS J152509.57+360054.5
- Q8667: Boötes
- Q8679: Pisces

#### Type labels:

- Q9283100: nova-like stars
- Q8928: constellation

#### Property labels:

- P31: instance of
- P59: constellation
- P2215: proper motion
- P6257: right ascension



# Typical CEA/CTA/CPA pipeline



# Data cleaning

- Fix broken unicode: <https://ftfy.readthedocs.io/en/latest>
- Remove text in parenthesis/brackets
- Expand abbreviations
- Isolate units of measure
- Identify syntactic types: string, numbers, dates, ...
- Identify main entity

# Candidate generation

## Objective

- High recall

## Lookup using multilingual labels/aliases

- Wikidata API: <https://www.wikidata.org/w/api.php>
- Wikipedia redirects and anchors
- Custom ElasticSearch index

## Fuzzy query

- Progressive Levenshtein distance
- Two column indices [MTab4Wikidata]

### Pisces (Q8679)

zodiac constellation straddling the celestial equator

Psc | Piscium

▼ In more languages

Configure

Language	Label	Description	Also known as
English	Pisces	zodiac constellation straddling the celestial equator	Psc Piscium
Spanish	Piscis	constelación	
Traditional Chinese	雙魚座	No description defined	
Chinese	双鱼座	No description defined	

# Feature generation

## Cell features

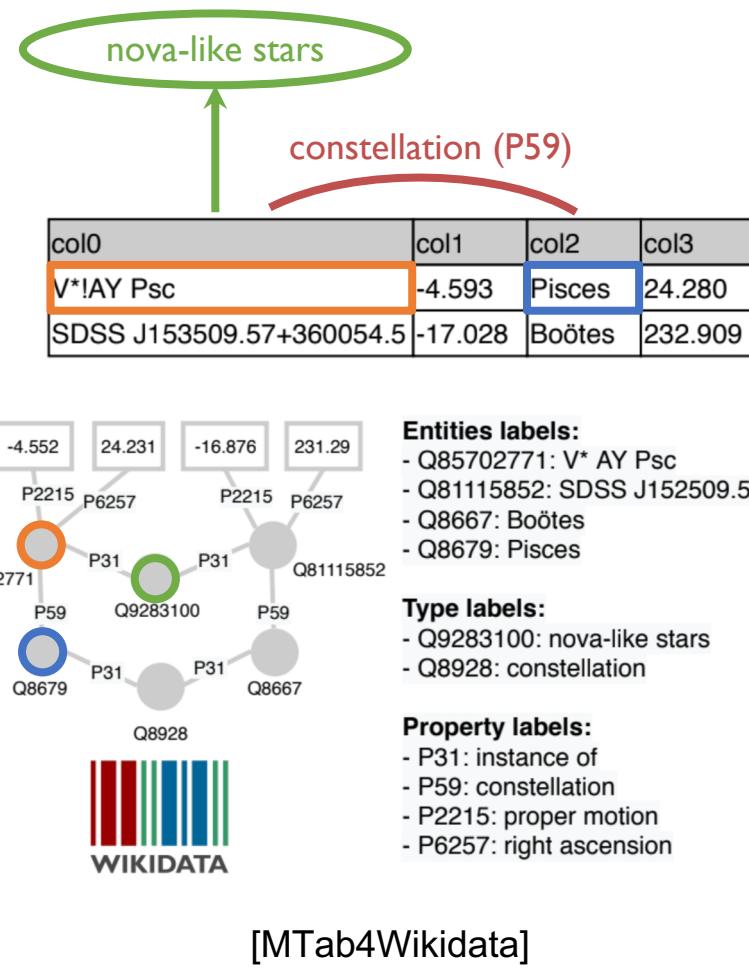
- String similarity
- popularity (pagerank, smallest Q-number)
- Wikidata embeddings

## Row features (CPA)

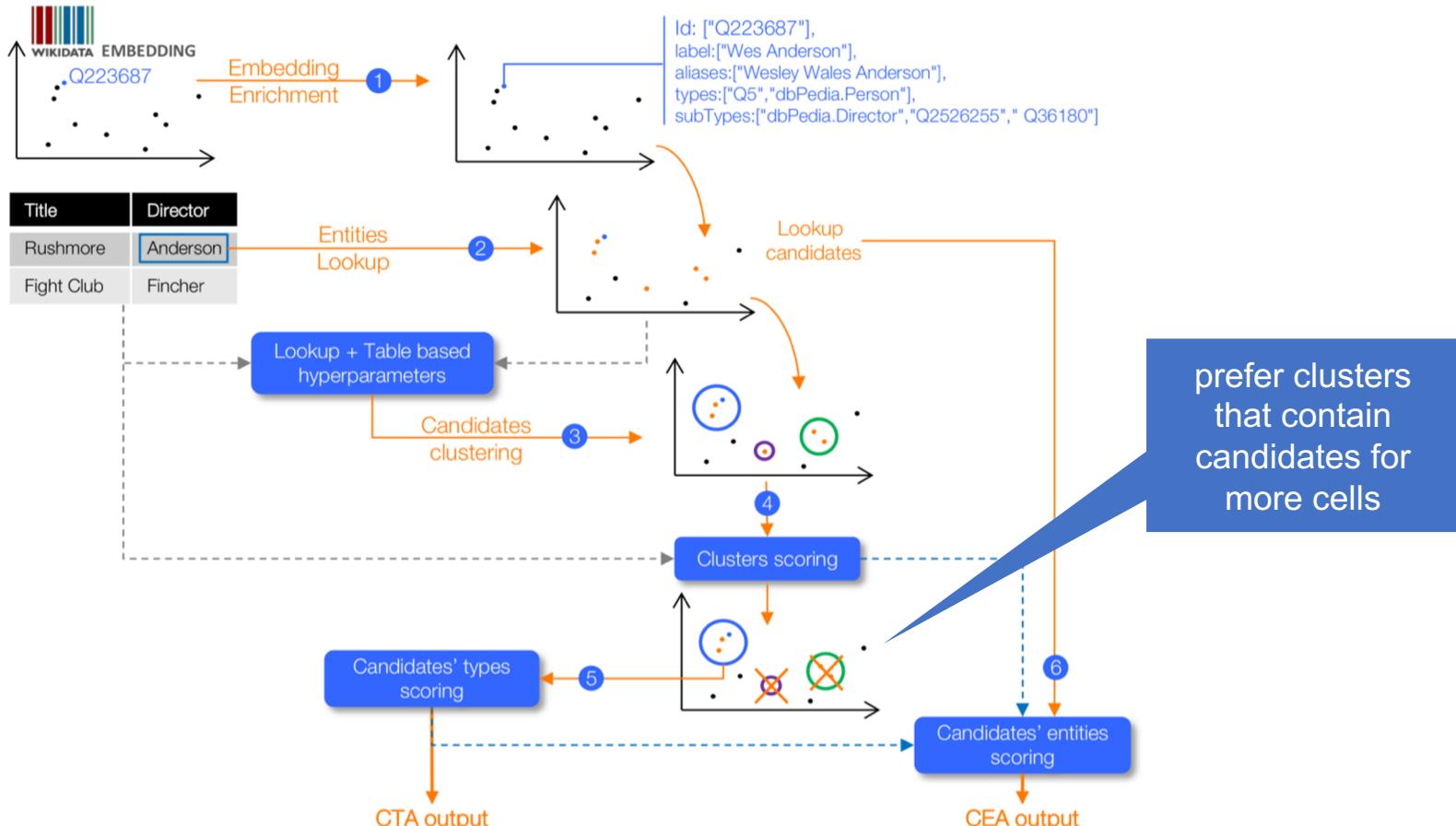
- Candidates that respect properties are more likely to be correct

## Column features (CTA)

- Instances of the column class are more likely to be correct



# DAGOBAH: use embeddings as features



# SemTab 2020

## Semantic Web Challenge on Tabular Data to Knowledge Graph Matching

<https://www.cs.ox.ac.uk/isg/challenges/sem-tab/2020/>

Column Entity Annotation (CEA)

Team	F1-Score	Precision
<i>MTab4Wikidata</i>	0.993	0.993
<i>LinkingPark</i>	0.985	0.985
<i>Team_DAGOBAH</i>	0.984	0.985
<i>bbw</i>	0.978	0.984
<i>JenTab</i>	0.973	0.975
<i>AMALGAM</i>	0.892	0.914
<i>LexMa</i>	0.845	0.911
<i>SSL</i>	0.833	0.833
<i>Unimib/MantisTable</i>	0.812	0.985

Column Table Annotation (CTA)

Team	Average F1-Score	Average Precision
<i>MTab4Wikidata</i>	0.981	0.982
<i>bbw</i>	0.98	0.98
<i>Team_DAGOBAH</i>	0.972	0.972
<i>LinkingPark</i>	0.953	0.953
<i>SSL</i>	0.946	0.946
<i>JenTab</i>	0.93	0.93
<i>AMALGAM</i>	0.858	0.861
<i>Unimib/MantisTable</i>	0.725	0.989
<i>Kepler-aSI</i>	0.253	0.676

Column Property Annotation (CPA)

Team	F1-Score	Precision
<i>MTab4Wikidata</i>	0.997	0.997
<i>bbw</i>	0.995	0.996
<i>Team_DAGOBAH</i>	0.995	0.995
<i>JenTab</i>	0.994	0.994
<i>LinkingPark</i>	0.985	0.988
<i>SSL</i>	0.924	0.924
<i>Unimib/MantisTable</i>	0.803	0.988

- Top system has an almost perfect score
- Many systems perform well
- All systems are heuristic
- Tables are synthetic (generated from Wikidata)

# CEA/CTA/CPA open problems

Tables contain new entities

- NIL linking
- Suggest new entities

Tables contain new properties

- Suggest new properties

Table data more recent than KG, or  
table/KG contain errors

- Confuses identification of properties

[https://en.wikipedia.org/wiki/2018\\_Colombian\\_presidential\\_election](https://en.wikipedia.org/wiki/2018_Colombian_presidential_election)

Candidate	Party/alliance	First round	
		Votes	%
Iván Duque Márquez	Grand Alliance for Colombia	7,569,693	39.14
Gustavo Petro	List of Decency	4,851,254	25.09
Sergio Fajardo	Colombia Coalition	4,589,696	23.73
Germán Vargas Lleras	Mejor Vargas Lleras	1,407,840	7.28
Humberto De la Calle	PLC-ASI	399,180	2.06
Jorge Antonio Trujillo	We Are All Colombia	75,614	0.39
Promotores Voto En Blanco	Party of Ethnic Reclamation "PRE"	60,312	0.31
Viviane Morales Hoyos	Somos Región Colombia	41,458	0.21

new  
entities

new  
data

# Summary

# Summary

## Semantic modeling

- Identify classes and relationships to describe tables
- CTA and CPA are the simple cases
- Wikidata brings new challenges: n-ary relations, multi-hop properties
- Early work focused on custom ontologies
- Recent work focused on large public KGs (Wikidata, DBpedia)

## Evaluation remains a challenge

- Benchmarks with real tables are small (T2DV2)
- Large benchmarks are synthetic and biased (SemTab 2020)
- Heuristic systems perform very well

# Bibliography

- Taherian, M., Craig A. Knoblock, Pedro A. Szekely and J. Ambite. "Rapidly Integrating Services into the Linked Data Cloud." International Semantic Web Conference (2012).
- Szekely, Pedro A., Craig A. Knoblock, Fengyu Yang, Xuming Zhu, E. Fink, Rachel Allen and Georgina Goodlander. "Connecting the Smithsonian American Art Museum to the Linked Data Cloud." ESWC (2013).
- Bhagavatula, Chandra Sekhar, Thanapon Noraset, and Doug Downey. "Tabel: Entity linking in web tables." *International Semantic Web Conference*. Springer, Cham, 2015.
- Ritze, Dominique, and Christian Bizer. "Matching web tables to DBpedia-A feature utility study." *context* 42.41 (2017): 19-31.
- Efthymiou, Vasilis, et al. "Matching web tables with knowledge base entities: from entity lookups to entity embeddings." *International Semantic Web Conference*. Springer, Cham, 2017.
- Cremašchi, Marco, Roberto Avogadro, and David Chieregato. "MantisTable: an Automatic Approach for the Semantic Table Interpretation." *SemTab@ ISWC 2019* (2019): 15-24.
- Nguyen, P., N. Kertkeidkachorn, R. Ichise and Hideaki Takeda. "MTab: Matching Tabular Data to Knowledge Graph using Probability Models." *SemTab@ISWC* (2019).
- Chen, Shuang, Alperen Karaoglu, C. Negreanu, Tingting Ma, Jin-ge Yao, Jack Williams, A. D. Gordon and Chin-Yew Lin. "LinkingPark: An Integrated Approach for Semantic Table Interpretation." *SemTab@ISWC* (2020).
- Chabot, Yoan, Thomas Labb  , Jixiong Liu and Raphael Troncy. "DAGOBAH: An End-to-End Context-Free Tabular Data Semantic Annotation System." *SemTab@ISWC* (2019).
- Abdelmageed, Nora. "JenTab: A Toolkit for Semantic Table Annotations." (2021).
- Binh Vu, Craig A. Knoblock, Pedro Szekely, Minh Pham, & Jay Pujara (2021). A Graph-based Approach for Inferring Semantic Descriptions of Wikipedia Tables. In ISWC 2021 - 20th International Semantic Web Conference.