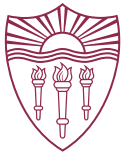


Exploiting Web Tables and Knowledge Graphs for Creating Semantic Descriptions of Data Sources

Committee Members:

Craig Knoblock, Yolanda Gil, Sven Koenig, Jay Pujara, and Daniel O'Leary

Binh Vu

Dissertation Defense

USC INFORMATION SCIENCES INSTITUTE

USC Viterbi School of Engineering

Outline

- Introduction
- Thesis Overview
- Creating Semantic Descriptions of Linked Tables
- Creating Semantic Descriptions of Tables with Overlapping Data
- Creating Semantic Descriptions of Tables without Overlapping Data
- Related Work
- Conclusion and Future Work

Tables are rich sources of structured knowledge

- The web has millions of tables covering many domains

List of players won Walter Payton Award

Year	Player	School	Position
1987	Kenny Gamble	Colgate	RB
1988	Dave Meggett	Towson State	RB
2018	Devlin Hodges	Samford	QB
2019	Trey Lance	North Dakota State	QB

Members of 56th New Brunswick Legislature

Name	Party	Ri
Hédard Albert	Liberal	Caraquet
David Alward	Progressive Conservative	Woodstock
Donald Arseneault	Liberal	Dalhousie-Restigouche East
John Betts	Progressive Conservative	Moncton Crescent

Dereham	Dereham	Inmazeb	Regeneron	Ebola Virus
Aldeby	Aldeby	Olorofim	F2G	invasive mold infections
Ashwellthorpe	Ashwellthorpe	1881	1939	Great Eastern

Hotel Diablo	School	Rail	Shellharbour	1959	Website
FLYGOD is an Awesome GOD		Westside Gunn	Hip hop		

nt pericarditis

Tables have semantic heterogeneity

- “When independent parties develop database schemas for the same domain, they will almost always be quite different from each other.”

  Alon Halevy, **Why Your Data Won’t Mix**

Peak	Height above MSL	Prominence	Mountain range	Countries
Mount Everest	8,848 m	8,848 m	Himalayas	Nepal, China
Godwin Austen (K2)	8,611 m	4,017 m	Karakoram	Pakistan, China
Kangchenjunga	8,586 m	3,922 m	Himalayas	India, Nepal

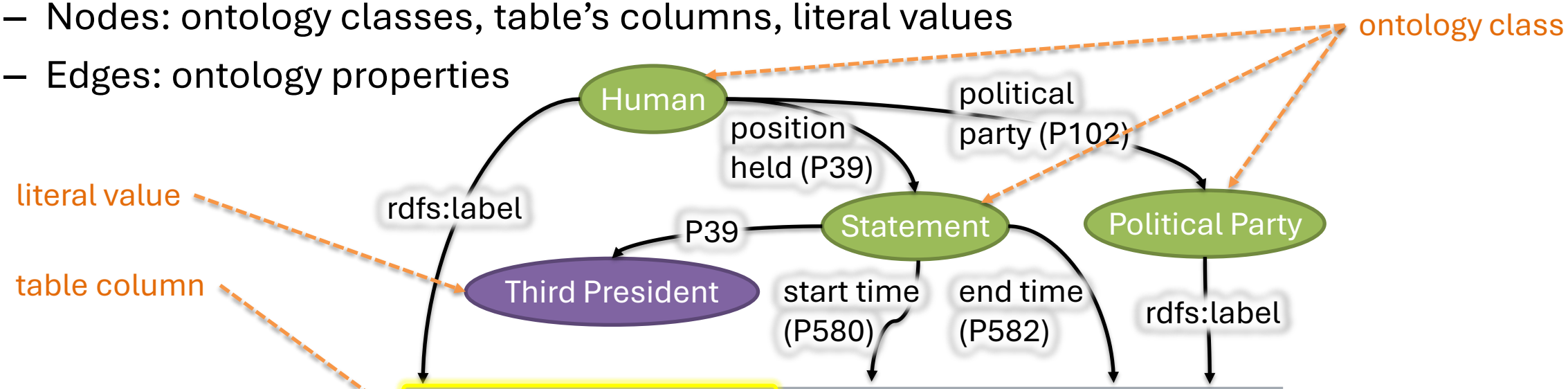
worlddata.info

Mountain name(s)	Height (rounded)	Prominence (rounded)	Coordinates [dp 4]	Range
Mount Everest	8,848	8,848	27°59′17″N 86°55′31″E	Mahalangur Himalaya
K2	8,611	4,020	35°52′53″N 76°30′48″E	Baltoro Karakoram
Kangchenjunga	8,586	3,922	27°42′12″N 88°08′51″E *	Kangchenjunga Himalaya

wikipedia.org

Semantic description of a table

- Describing types and relationships in the tables precisely using ontologies
  - Nodes: ontology classes, table's columns, literal values
  - Edges: ontology properties

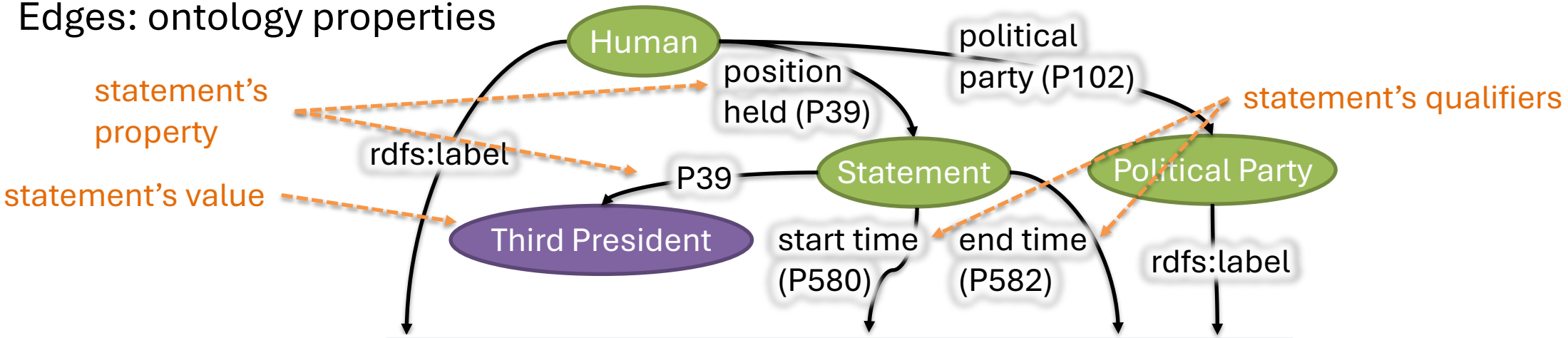


Name	Entered Office	Left Office	Party
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

Third Presidents of National Council (Austria)

Semantic description of a table

- Describing types and relationships in the tables precisely using ontologies
  - Nodes: ontology classes, table's columns, literal values
  - Edges: ontology properties

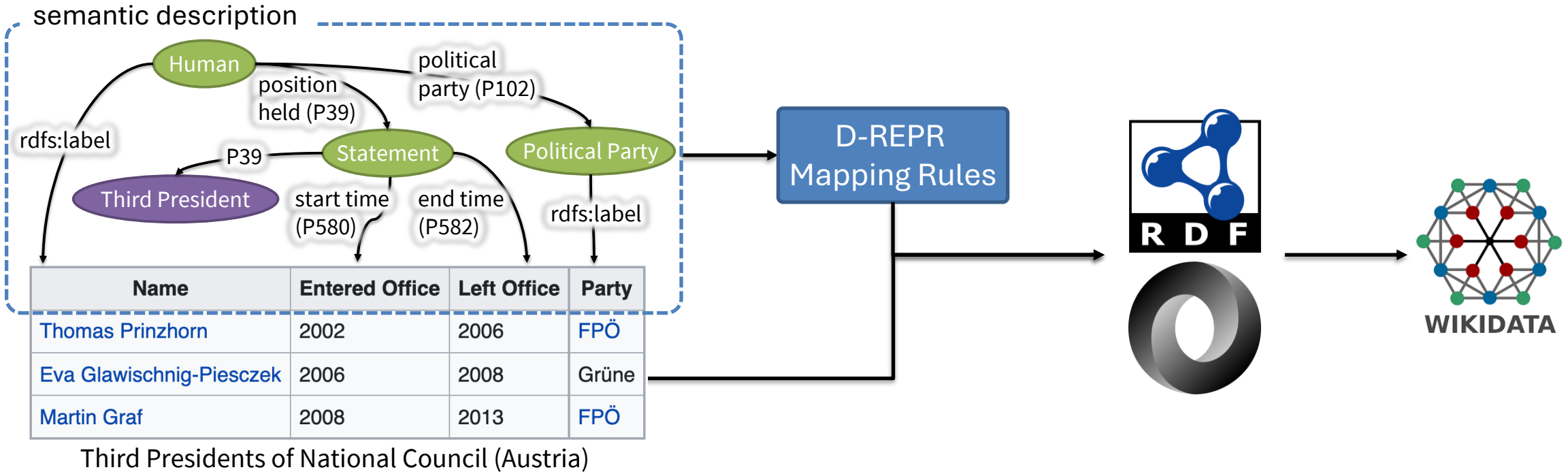


Name	Entered Office	Left Office	Party
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

Third Presidents of National Council (Austria)

Semantic description of a table

- Essential for automatic data discovery and data integration



Goal: create semantic descriptions automatically

Challenges in Semantic Modeling

- Huge number of possible semantic descriptions
- Supervised approaches need lots of labeled sources

Thesis Statement

By exploiting knowledge from web tables and knowledge graphs, we can learn semantic descriptions of tables with little or no manually labeled training data

Proposed Approach

Dan Carter (Q726199)

New Zealand rugby union player

Daniel William Carter

position played on team / speciality: fly-half

Player	Position	Team	Points
Dan Carter	Fly-half	New Zealand	1598
Ronan O'Gara	Fly-half	Ireland	1083
Percy Montgomery	Fullback	South Africa	893

1. Tables with links to entities in KG

Candidate Entity
Dan Carter (Q726199) New Zealand rugby union player
Dan Carter (Q5213238) American politician
Dan Carter (Q59277840) Mayor of Oshawa, Ontario, Canada

Player	Position	Team	Points
Dan Carter	Fly-half	New Zealand	1598
Ronan O'Gara	Fly-half	Ireland	1083
Percy Montgomery	Fullback	South Africa	893

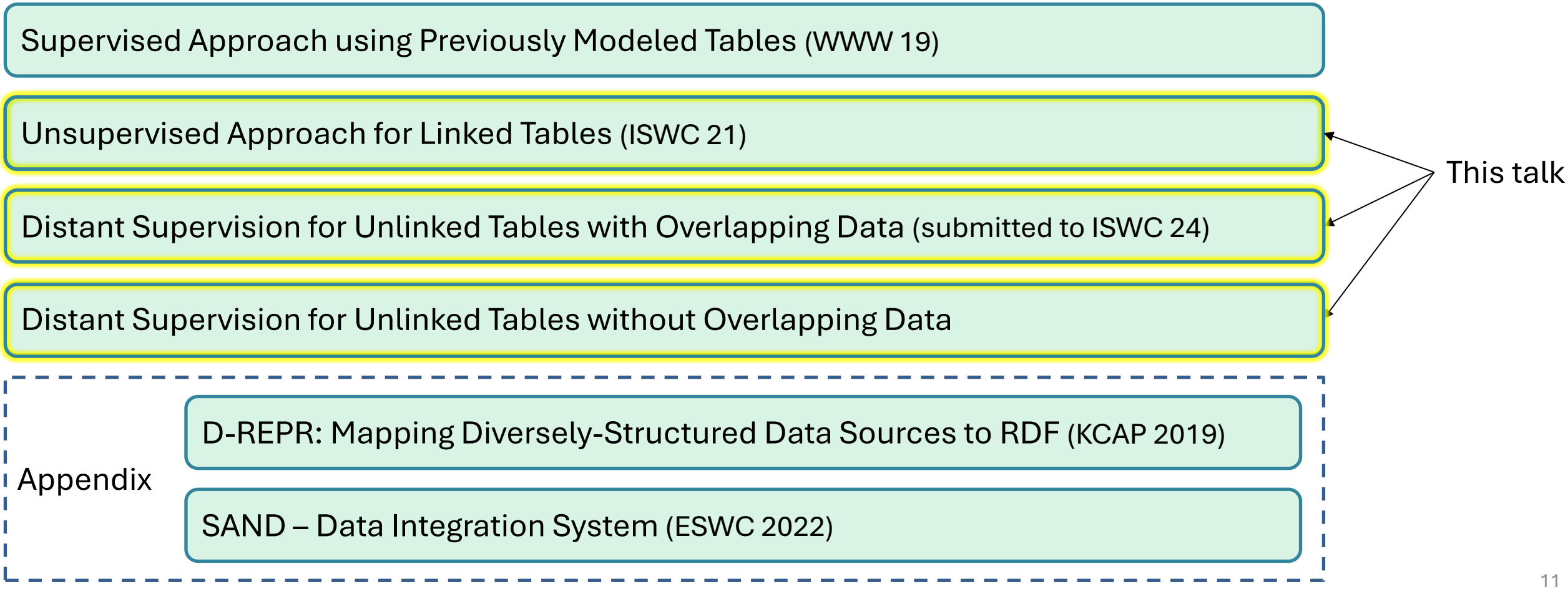
2. Unlinked tables with overlapping data with KG

Player	Position	Team	Points
Dan Carter	Fly-half	New Zealand	1598
Ronan O'Gara	Fly-half	Ireland	1083
Percy Montgomery	Fullback	South Africa	893

3. Unlinked tables without overlapping data with KG

Contributions

Comprehensive techniques for the semantic modeling problem under different settings and assumptions



Outline

- Introduction
- Thesis Overview
- **Creating Semantic Descriptions of Linked Tables**
- Creating Semantic Descriptions of Tables with Overlapping Data
- Creating Semantic Descriptions of Tables without Overlapping Data
- Related Work
- Conclusion and Future Work

Motivating Example

- Information of entities in KG can help semantic modeling

President of the National Council (Austria)

From Wikipedia, the free encyclopedia

List of third presidents [edit]

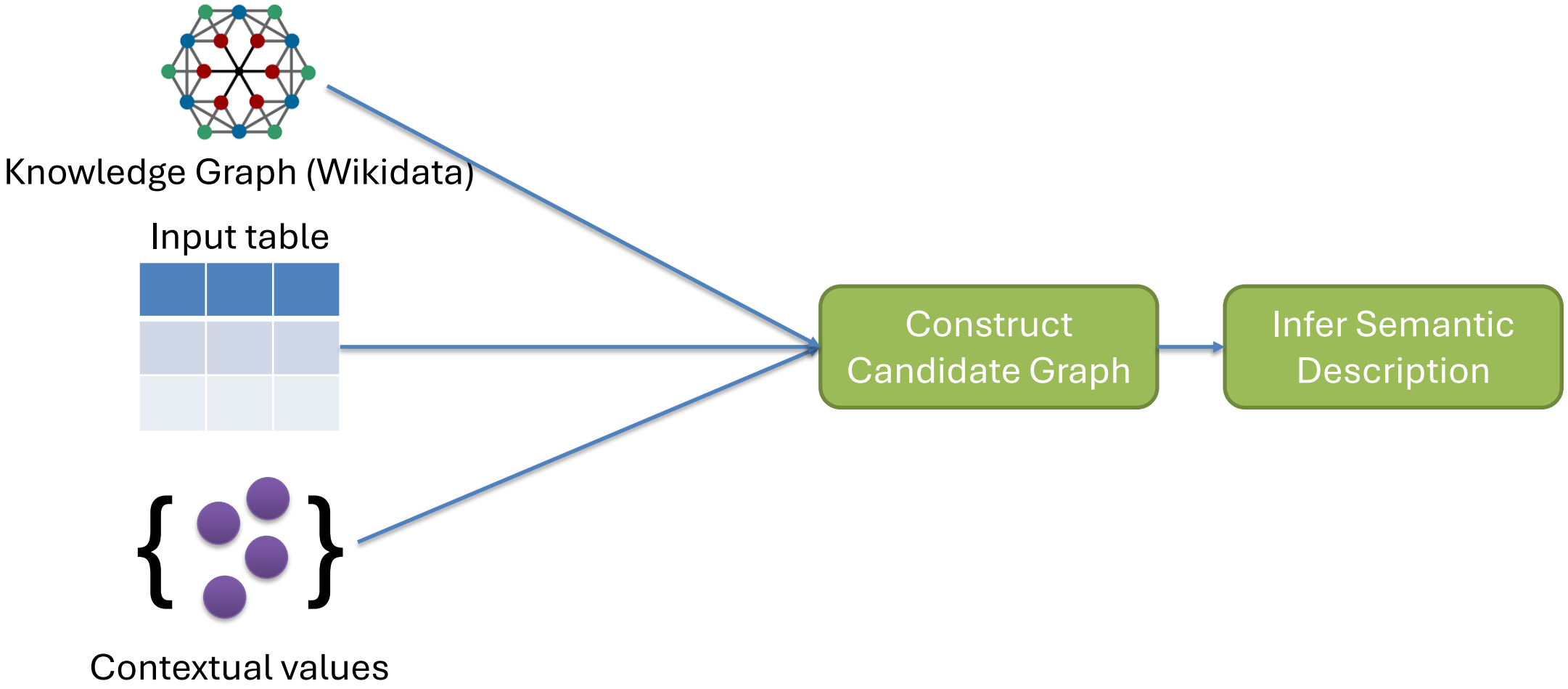
Name	Entered Office	Left Office	Party
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

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

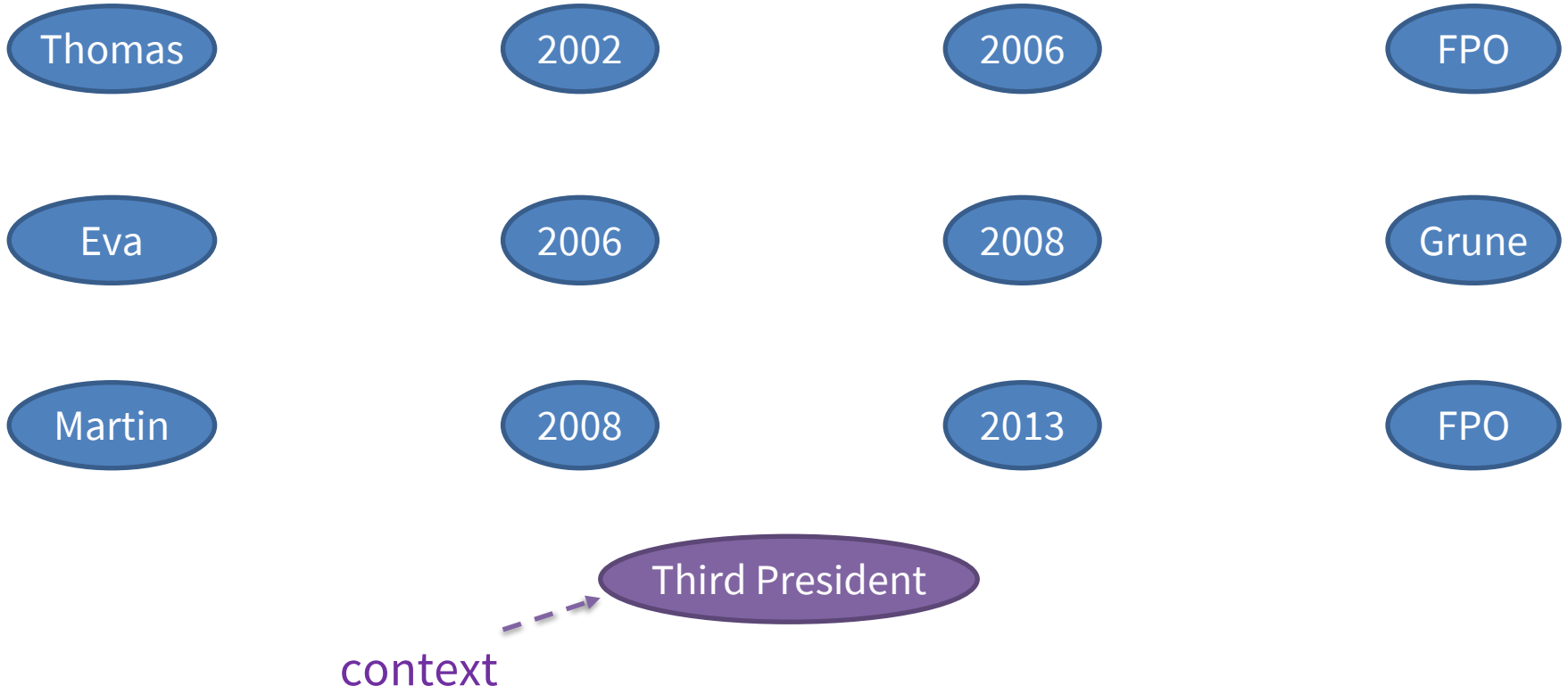
Approach


Knowledge Graph (Wikidata)
Input table
Contextual values
Construct Candidate Graph
Infer Semantic Description

Construct Candidate Graph: Discovering Links

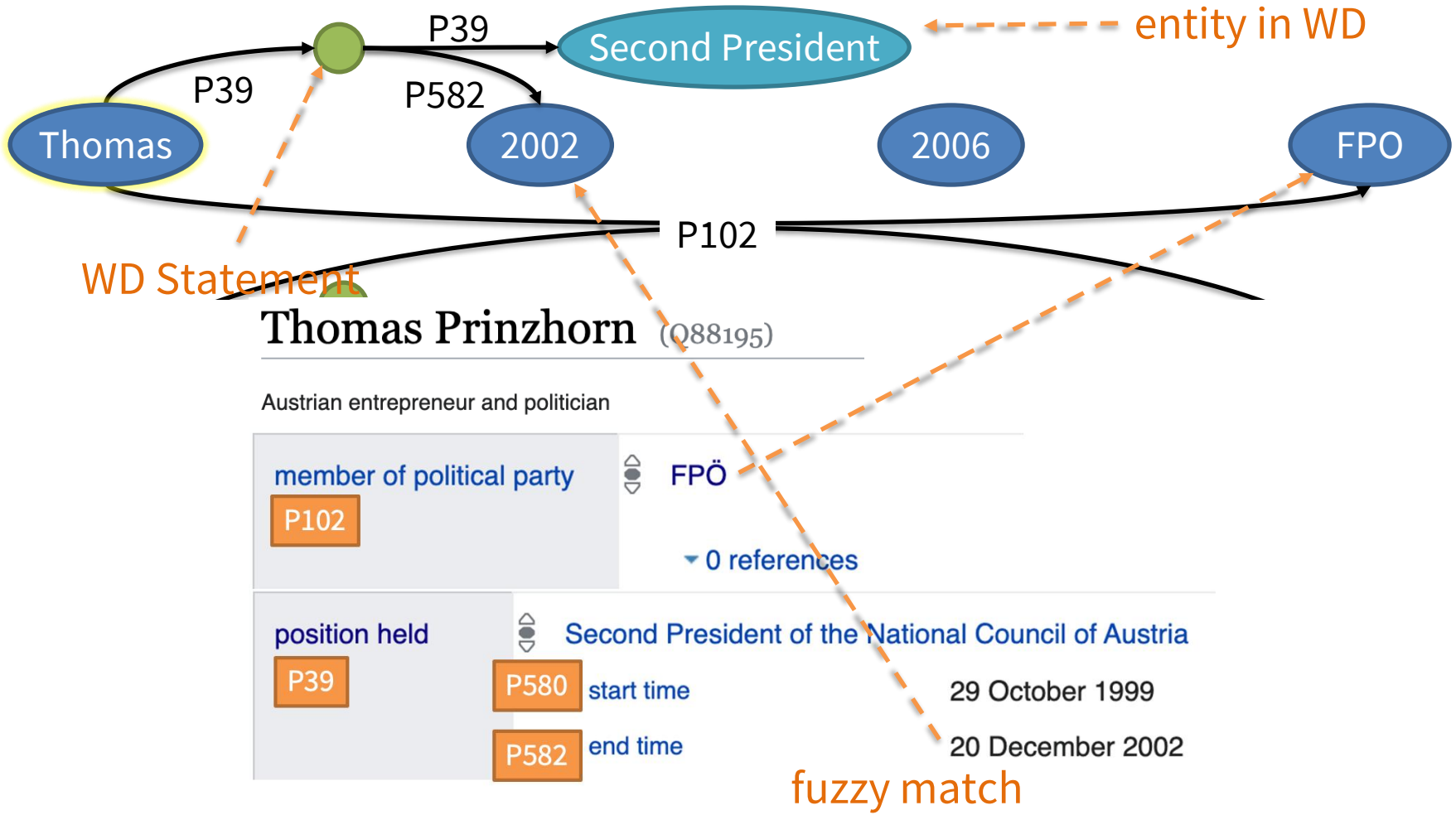
- Create a graph of cells and context

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



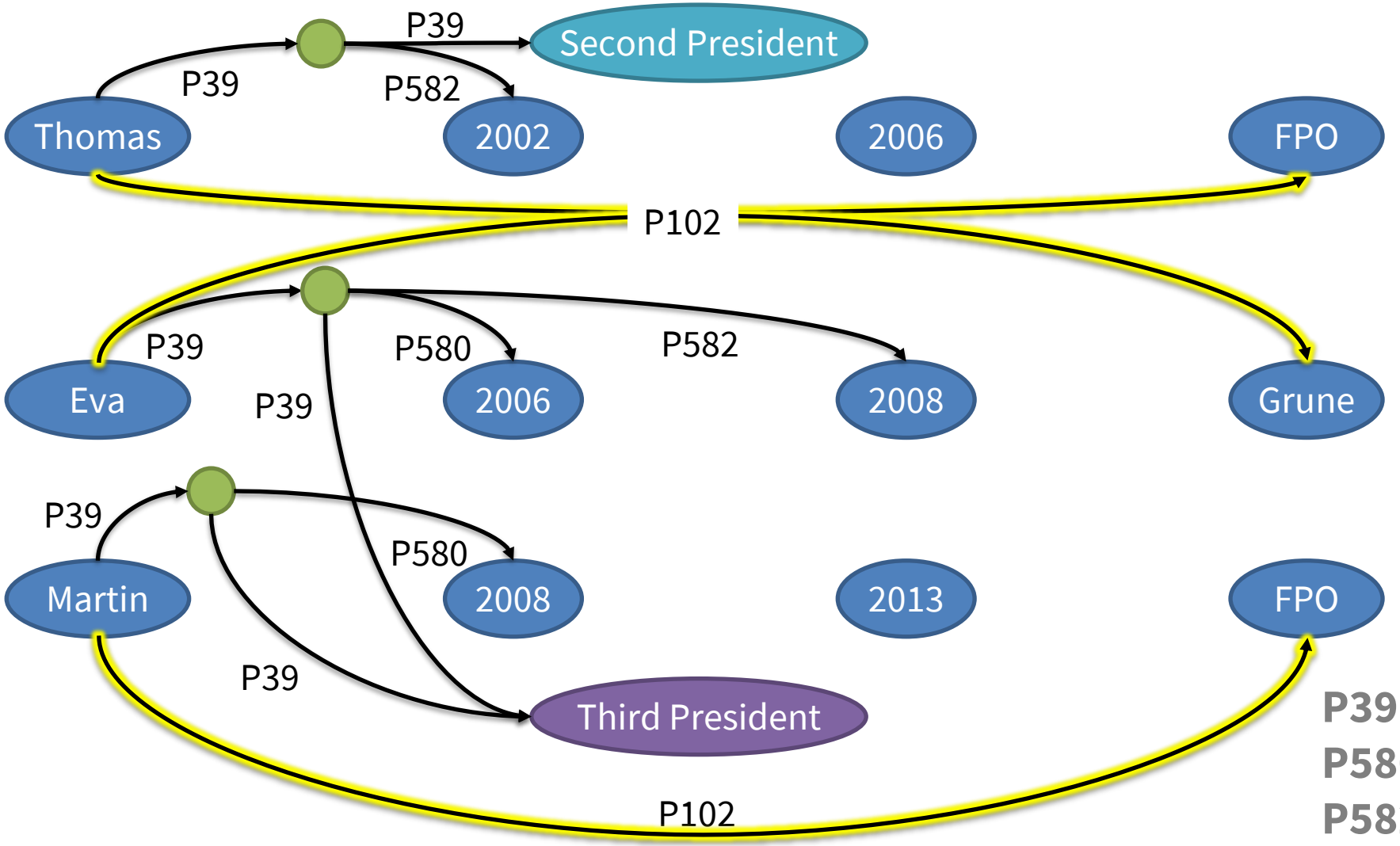
Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata



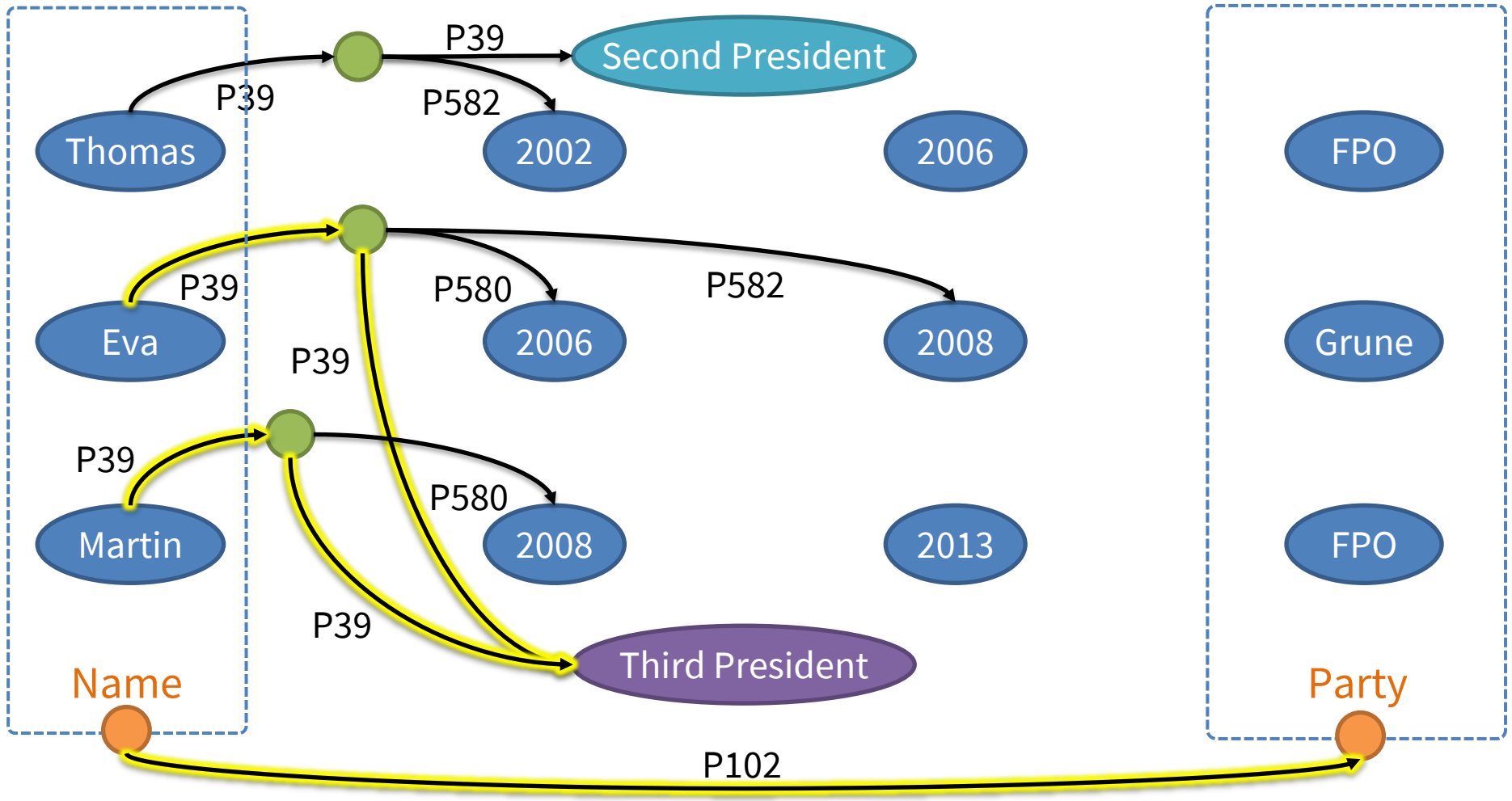
Construct Candidate Graph: Summarization

- Group links of cells from same source & target columns/context



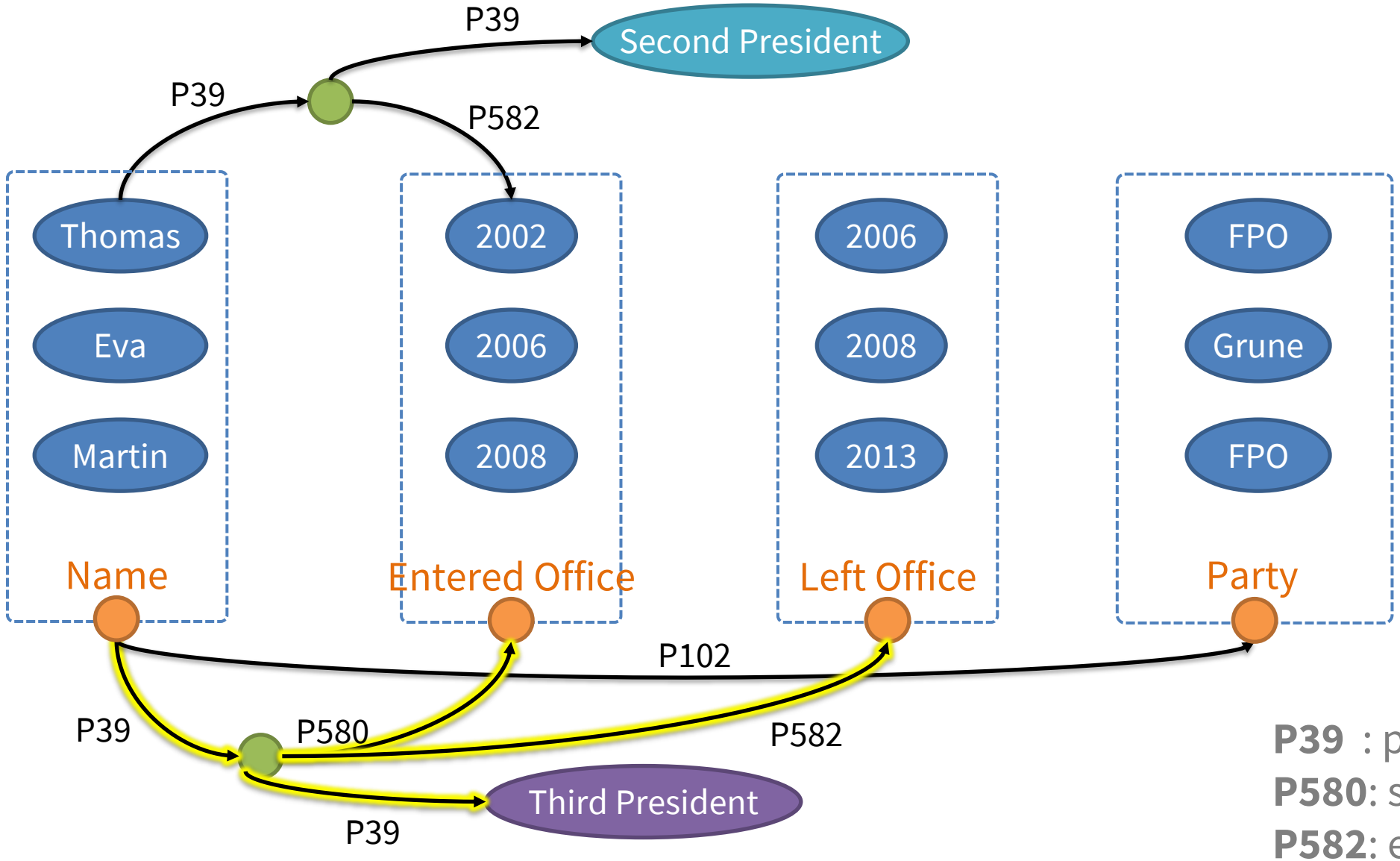
P39 : position held
P580: start time
P582: end time

Construct Candidate Graph: Summarization



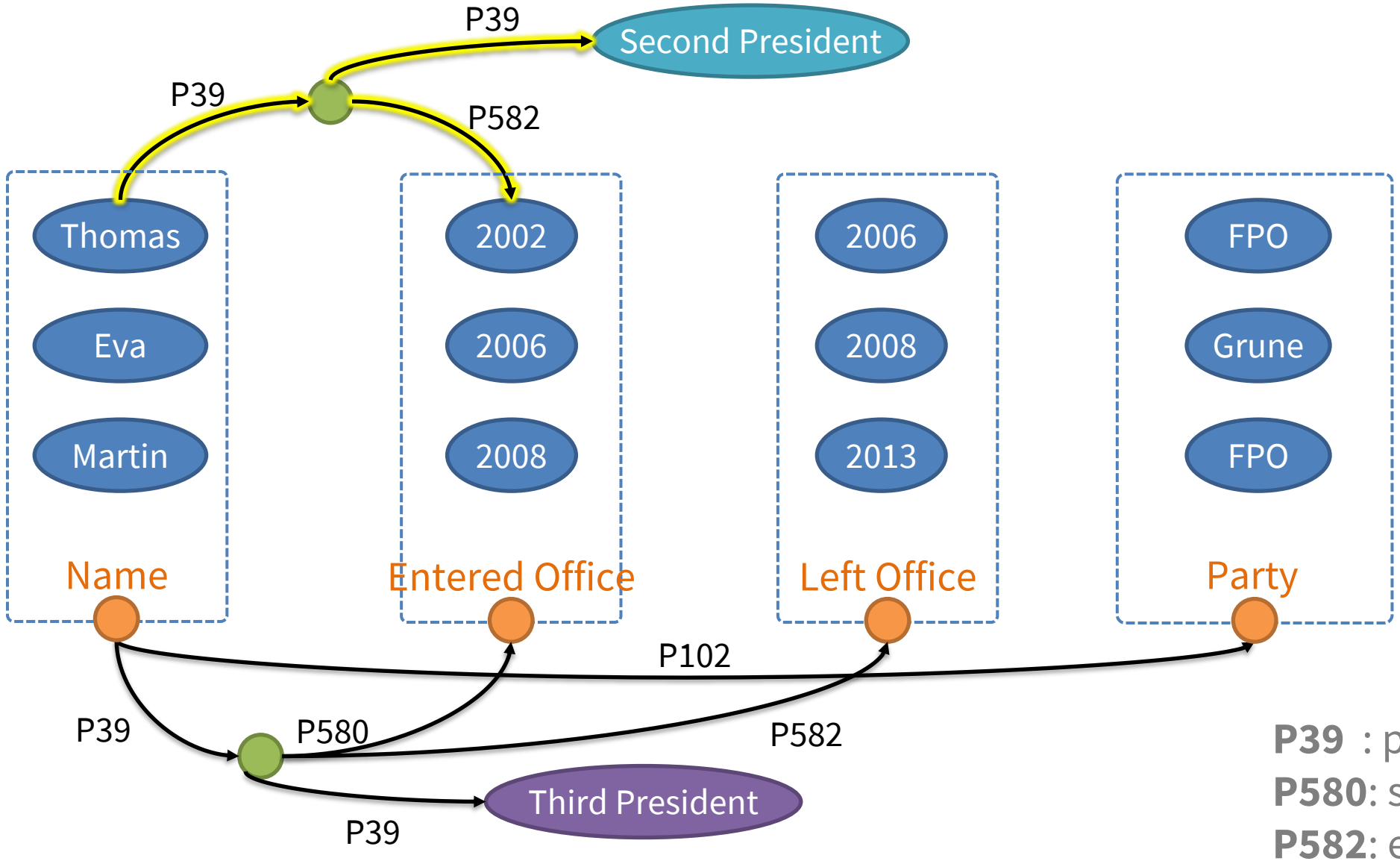
P39 : position held
P580: start time
P582: end time

Construct Candidate Graph: Summarization



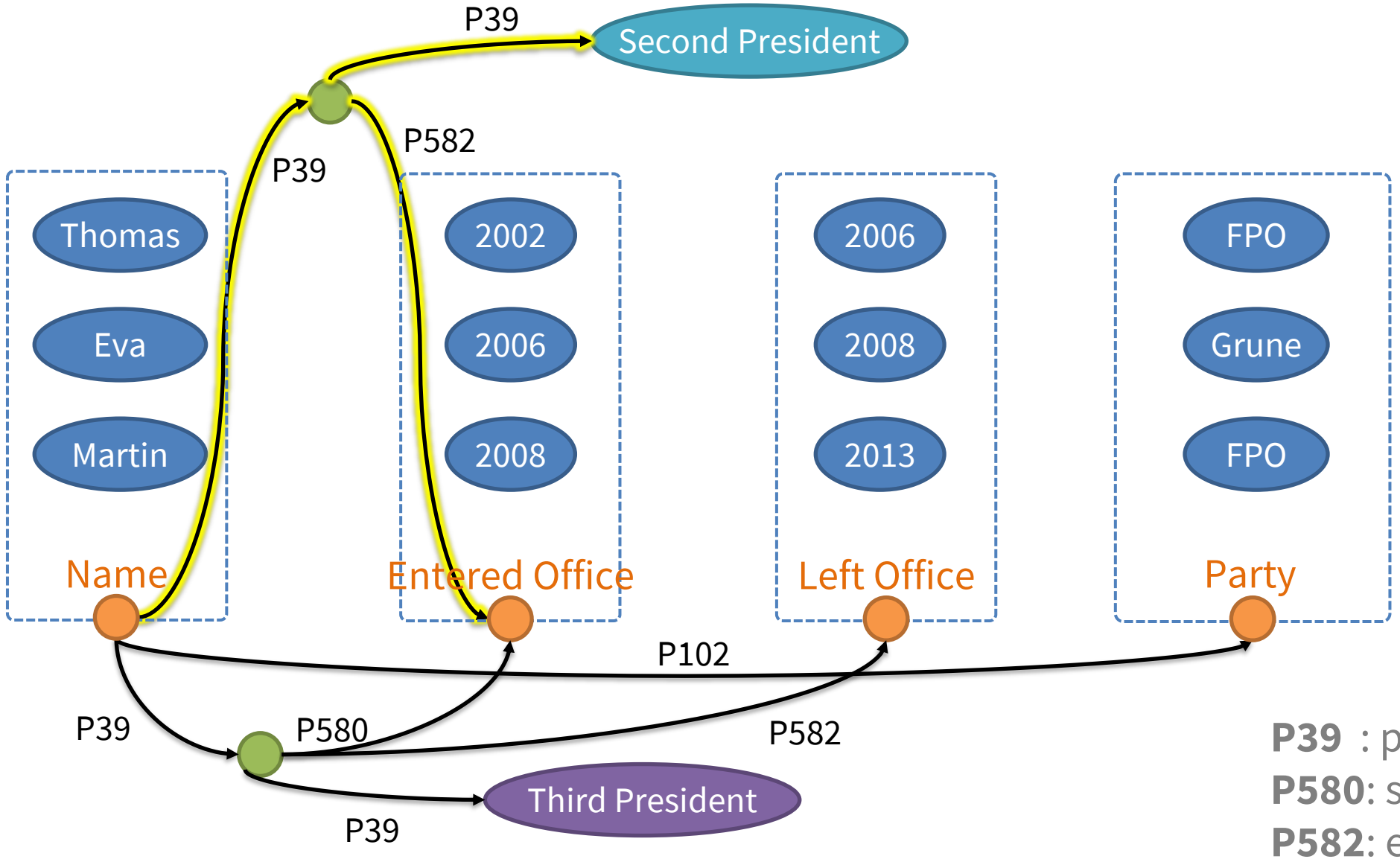
P39 : position held
P580: start time
P582: end time

Construct Candidate Graph: Summarization



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

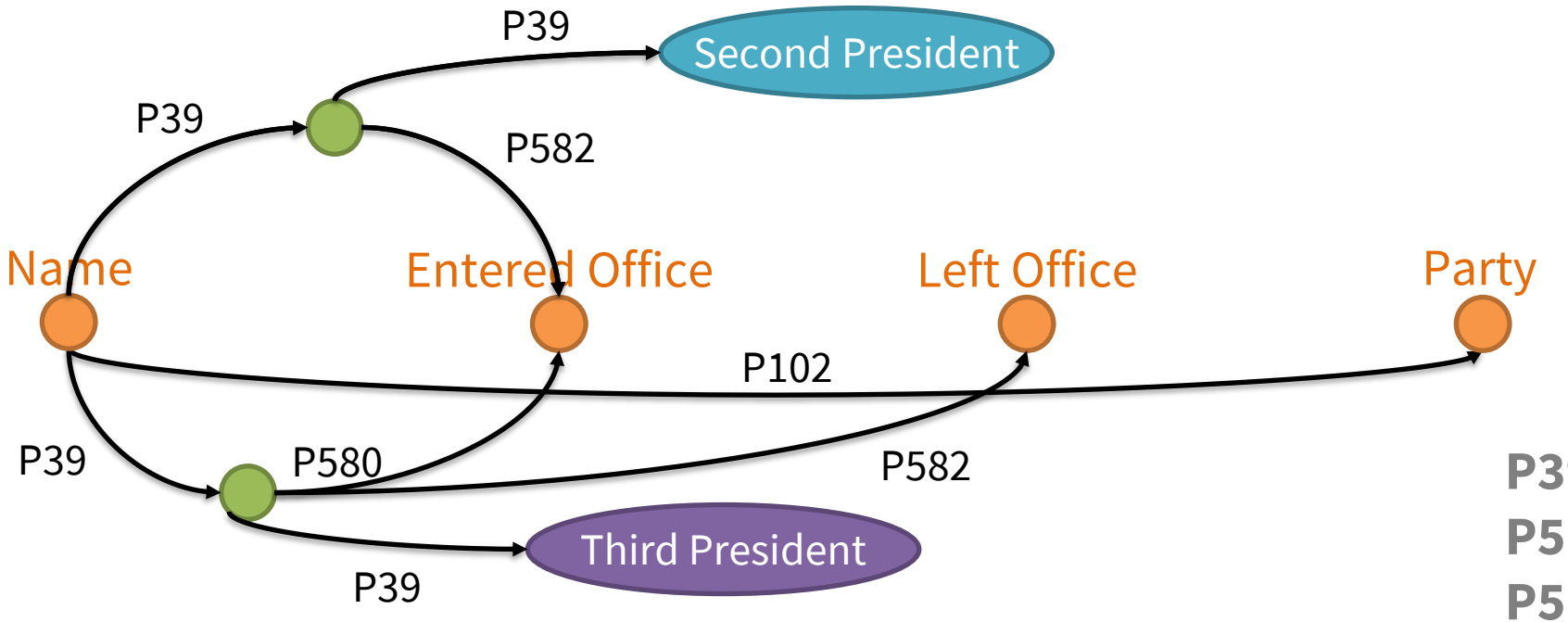
Construct Candidate Graph: Summarization



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

Construct Candidate Graph: Summarization

- Final candidate graph

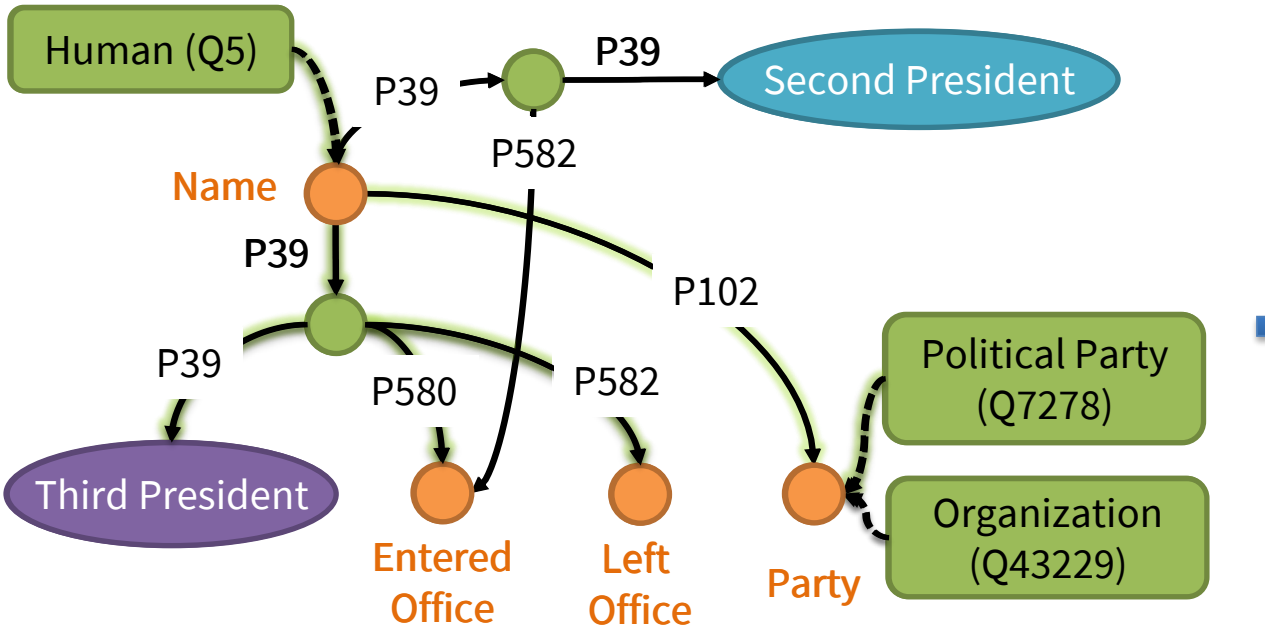


P39 : position held
P580: start time
P582: end time

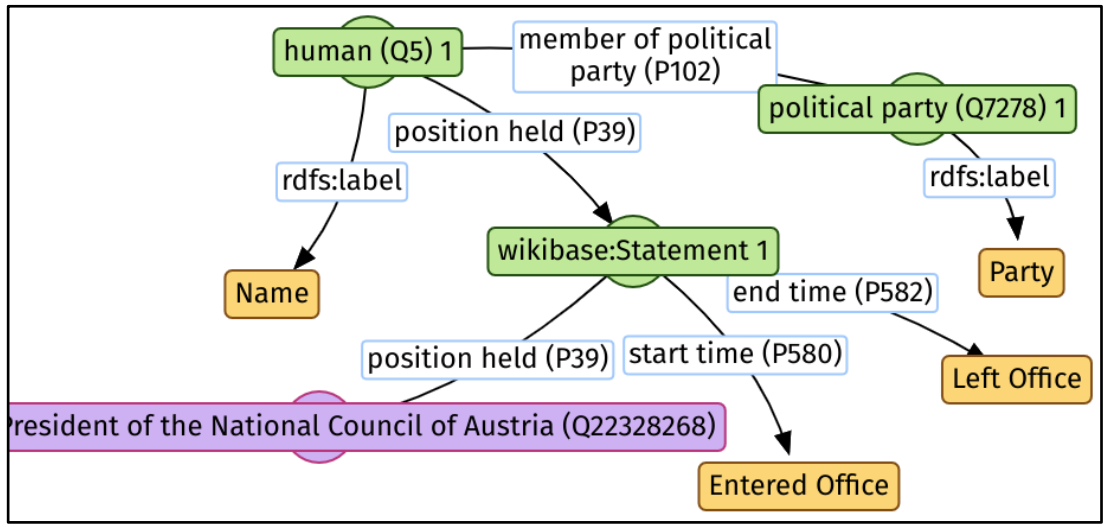
After Building Candidate Graph

- 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.



Candidate Graph



Semantic Description

Approach

Inputs

- A target knowledge graph: Wikidata
- A linked relational table T
- A set of contextual values C

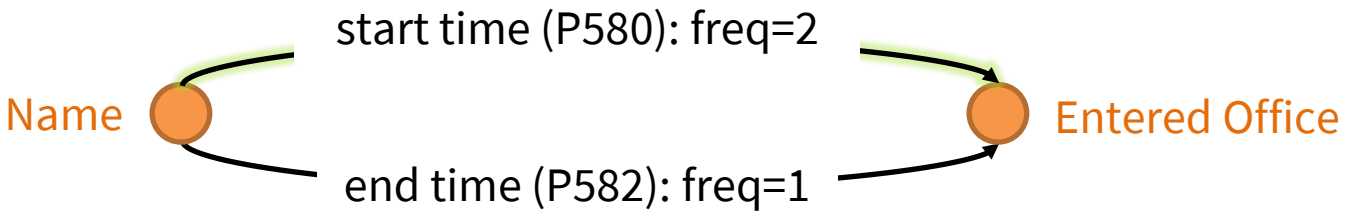
1. Construct candidate graph
2. **Infer semantic description**

Outputs:

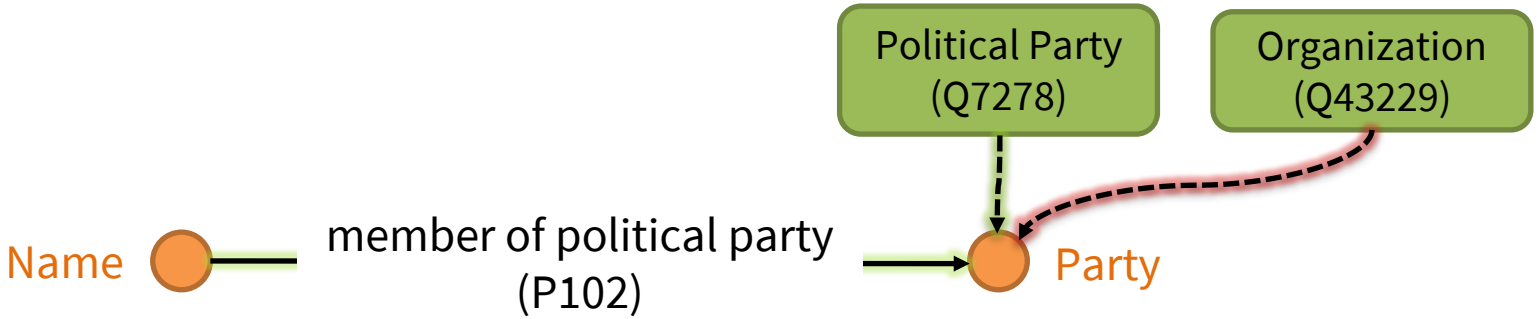
- A semantic description of (T, C)

Collective Reasoning Problem

- Selecting properties and types that
  - Are likely to happen based on evidence from the data



  - Comply with constraints defined in the ontology



  - Incorporate prior knowledge of the desired semantic descriptions

Collective Reasoning Problem

- **Probabilistic Soft Logic (PSL)**

  "A probabilistic graphical model framework using first-order logic"
  - Rules are converted to exponential function that approximates P(x)
  - Convex optimization to determine unknown values of predicates that maximize P(x)

excerpt of grounded rules in a PSL model

…

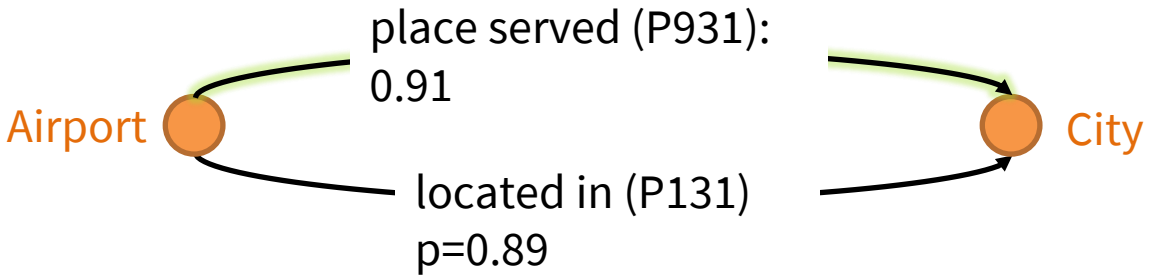
FREQ_MATCH(Name, Party, P102) → **CORRECT_REL**(Name, Party, P102)

CORRECT_REL(Name, Party, P102) & ¬**RANGE**(P102, Organization) → ¬ **CORRECT_TYPE**(Party, Organization)

…

Post-Processing

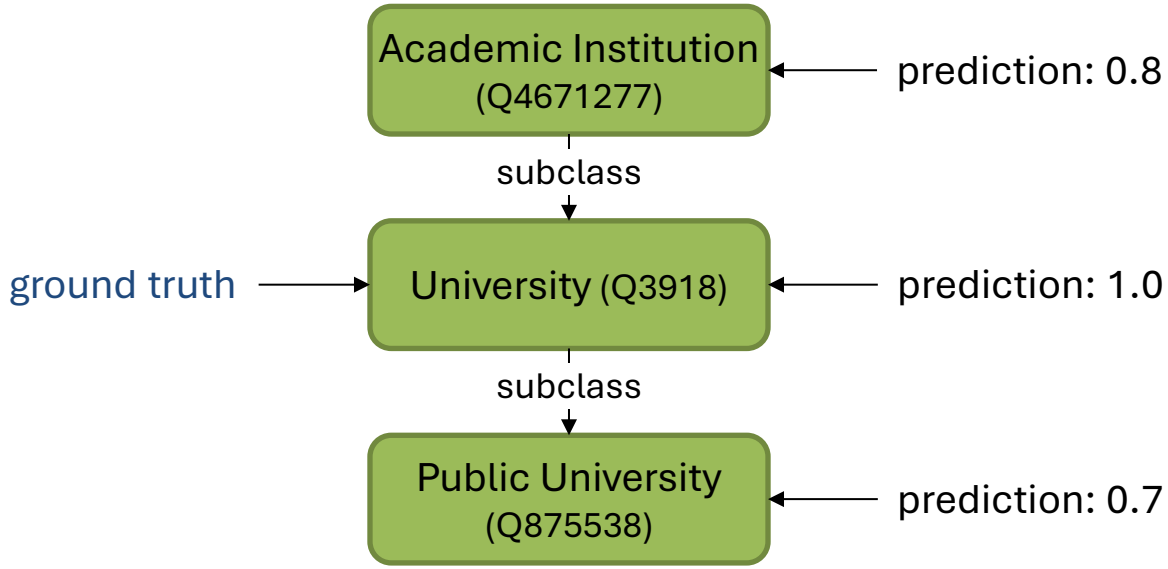
- PSL outputs probability of each relationships and types.



- Use our Steiner Tree algorithm to choose the most probable relationships
  - Avoid unnecessary loops
  - Prefer tree structure if possible

Evaluation

- Datasets:
  - 250WT: manually labeled sources samples from Wikipedia tables
  - SemTab2020: simple tables that generated automatically from Wikidata.
    - Each table contain properties of entities of a single class
- Evaluation metrics
  - Approximate Precision/Recall from the SemTab challenge



Evaluation

- GRAMS outperforms SOTA baselines on real-world dataset
  - **12.6%** on column relationships annotation (CPA)
  - **4.8%** on column type annotation (CTA)

Wikipedia Tables

Synthetic Tables

Dataset	Method	CPA			CTA		
		AP	AR	AF_1	AP	AR	AF_1
250WT	MantisTable	53.0%	44.7%	48.5%	**88.4%**	30.4%	45.2%
	MantisTable†	55.0%	57.1%	56.0%	87.7%*	34.9%	49.9%
	BBW	75.6%	11.8%	20.5%	83.1%	23.0%	36.0%
	BBW†	63.6%	58.9%	61.2%	73.9%	76.1%	75.0%
	MTab	83.9%	48.5%	61.5%	77.0%	77.0%	77.0%
	MTab†	84.8%*	54.6%	66.4%	77.5%	77.5%	77.5%
	GRAMS-ST	58.6%	**70.9%**	64.2%	-	-	-
	GRAMS	**88.1%**	63.9%	**74.1%**	81.3%	**82.4%**	**81.8%**
SemTab2020	MantisTable	98.3%	97.5%	97.9%	95.9%	79.0%	86.6%
	BBW	99.2%	**99.2%**	99.2%	96.7%	96.7%	96.7%
	MTab	**99.5%**	99.1%	**99.3%**	**97.1%**	**97.1%**	**97.1%**
	GRAMS-ST	99.2%	99.1%	99.2%	-	-	-
	GRAMS	99.3%	99.1%	99.2%	97.0%	97.0%	97.0%

† systems are modified to receive correct subject column

Evaluation

- GRAMS outperforms SOTA baselines on real-world dataset
  - **12.6%** on column relationships annotation (CPA)
  - **4.8%** on column type annotation (CTA)

Wikipedia Tables

Synthetic Tables

Dataset	Method	CPA			CTA		
		AP	AR	AF_1	AP	AR	AF_1
250WT	MantisTable	53.0%	44.7%	48.5%	**88.4%**	30.4%	45.2%
	MantisTable†	55.0%	57.1%	56.0%	87.7%*	34.9%	49.9%
	BBW	75.6%	11.8%	20.5%	83.1%	23.0%	36.0%
	BBW†	63.6%	58.9%	61.2%	73.9%	76.1%	75.0%
	MTab	83.9%	48.5%	61.5%	77.0%	77.0%	77.0%
	MTab†	84.8%*	54.6%	66.4%	77.5%	77.5%	77.5%
	GRAMS-ST	58.6%	**70.9%**	64.2%	-	-	-
	GRAMS	**88.1%**	63.9%	**74.1%**	81.3%	**82.4%**	**81.8%**
SemTab2020	MantisTable	98.3%	97.5%	97.9%	95.9%	79.0%	86.6%
	BBW	99.2%	**99.2%**	99.2%	96.7%	96.7%	96.7%
	MTab	**99.5%**	99.1%	**99.3%**	**97.1%**	**97.1%**	**97.1%**
	GRAMS-ST	99.2%	99.1%	99.2%	-	-	-
	GRAMS	99.3%	99.1%	99.2%	97.0%	97.0%	97.0%

† systems are modified to receive correct subject column

Summary

- GRAMS is unsupervised approach that can handle complex schema
  - multiple entities' types
  - n-ary relationships
  - context values

- Limitation:
  - assuming tables already have links to correct entities in a KG

Outline

- Introduction
- Thesis Overview
- Creating Semantic Descriptions of Linked Tables
- **Creating Semantic Descriptions of Tables with Overlapping Data**
- Creating Semantic Descriptions of Tables without Overlapping Data
- Related Work
- Conclusion and Future Work

Motivating Example

Correct relationships can be found by searching entities mentioned in the tables

Player	Position	Team	Points
Dan Carter	Fly-half	New Zealand	1598
Ronan O'Gara	Fly-half	Ireland	1083
Percy Montgomery	Fullback	South Africa	893

Name	Retrv. Score
Dan Carter (Q726199) New Zealand rugby union player	1.0
Dan Carter (Q5213238) American politician	1.0
Dan Carter (Q59277840) Mayor of Oshawa, Ontario, Canada	1.0

country of citizenship P27

member of sport team P54

Name	Retrv. Score
New Zealand (Q664) island country in the southwest Pacific Ocean	1.0
New Zealand national football team (Q175315) men's national association football team representing NZ	0.5
New Zealand national rugby union team (Q55801) men's rugby union team of New Zealand	0.45

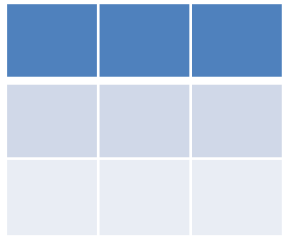
Name	Retrv. Score
Ronan O'Gara (Q733831) Irish rugby footballer and coach	1.0

member of sport team P54

Name	Retrv. Score
Republic of **Ireland** (Q27) country in Northwestern Europe	0.7
Ireland (Q22890) island in the North Atlantic Ocean	0.68
Ireland national rugby union team (Q599903) men's rugby union team representing the island of Ireland	0.45

Motivating Example

Wikipedia tables



WIKIPEDIA
The Free Encyclopedia

Player	Position	Team	Points
Dan Carter	Fly-half	New Zealand	1598
Ronan O'Gara	Fly-half	Ireland	1083
Percy Montgomery	Fullback	South Africa	893

Name	Retrieval Score
Dan Carter (Q726199) New Zealand rugby union player	1.0
Dan Carter (Q5213238) American politician	1.0
Dan Carter (Q59277840) Mayor of Oshawa, Ontario, Canada	1.0

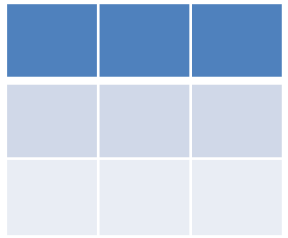
country of citizenship P27

member of sport team P54

Name	Retrieval Score
New Zealand (Q664) island country in the southwest Pacific Ocean	1.0
New Zealand national football team (Q175315) men's national association football team representing NZ	0.5
New Zealand national rugby union team (Q55801) men's rugby union team of New Zealand	0.45

Motivating Example

Wikipedia tables



distant supervision

Player	Position	Team	Points
Dan Carter	Fly-half	New Zealand	1598
Ronan O'Gara	Fly-half	Ireland	1083
Percy Montgomery	Fullback	South Africa	893

Name	Retrieval Score	New Score
Dan Carter (Q726199) New Zealand rugby union player	1.0	0.9
Dan Carter (Q5213238) American politician	1.0	0.6
Dan Carter (Q59277840) Mayor of Oshawa, Ontario, Canada	1.0	0.6

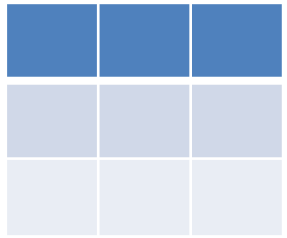
country of citizenship P27

member of sport team P54

Name	Retrieval Score	New Score
New Zealand (Q664) island country in the southwest Pacific Ocean	1.0	0.4
New Zealand national football team (Q175315) men's national association football team representing NZ	0.5	0.71
New Zealand national rugby union team (Q55801) men's rugby union team of New Zealand	0.45	0.7

Motivating Example

Wikipedia tables



distant supervision

Player	Position	Team	Points
Dan Carter	Fly-half	New Zealand	1598
Ronan O'Gara	Fly-half	Ireland	1083
Percy Montgomery	Fullback	South Africa	893

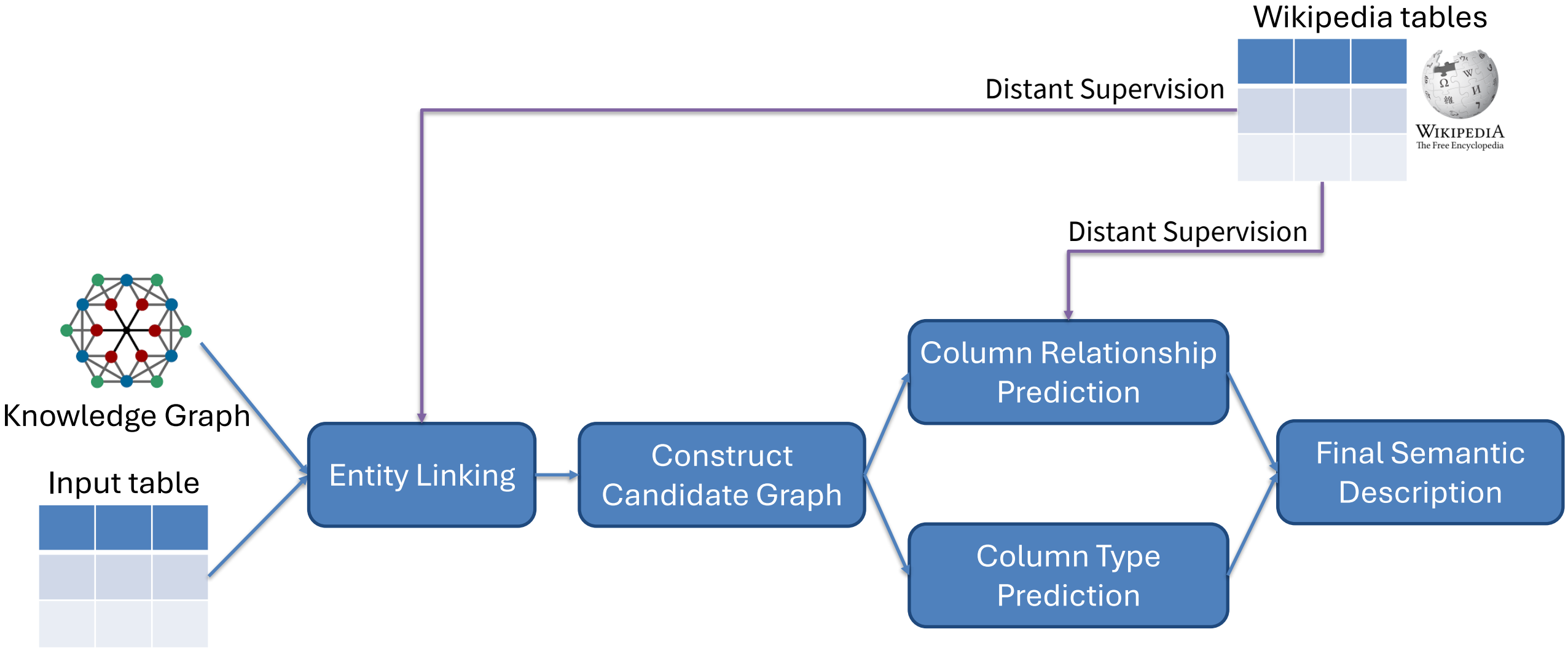
Name	Retrieval Score	New Score
Dan Carter (Q726199) New Zealand rugby union player	1.0	0.9
Dan Carter (Q5213238) American politician	1.0	0.6
Dan Carter (Q59277840) Mayor of Oshawa, Ontario, Canada	1.0	0.6

country of citizenship P27: 0.65

member of sport team P54: 0.73

Name	Retrieval Score	New Score
New Zealand (Q664) island country in the southwest Pacific Ocean	1.0	0.4
New Zealand national football team (Q175315) men's national association football team representing NZ	0.5	0.71
New Zealand national rugby union team (Q55801) men's rugby union team of New Zealand	0.45	0.7

Approach


Wikipedia tables
Distant Supervision
Distant Supervision
Knowledge Graph
Input table
Entity Linking
Construct Candidate Graph
Column Relationship Prediction
Column Type Prediction
Final Semantic Description

Generating Labeled Data (1)

Wikipedia tables

GRAMS

Wikipedia tables

Player	Position	Team	Points
Dan Carter	Fly-half	New Zealand	1598
Ronan O'Gara	Fly-half	Ireland	1083
Percy Montgomery	Fullback	South Africa	893

Use existing hyperlinks to automatically labeled candidate entities

Columns with same name with different types (230)

	country	*sports team*	*sports team*
	Team	Teams	Team. [11]
	New Zealand	Chicago Bulls	LA Rams
...	Ireland	Atlanta Hawks	SF 49ers
	South Africa	LA Lakers	Arizona Cardinals

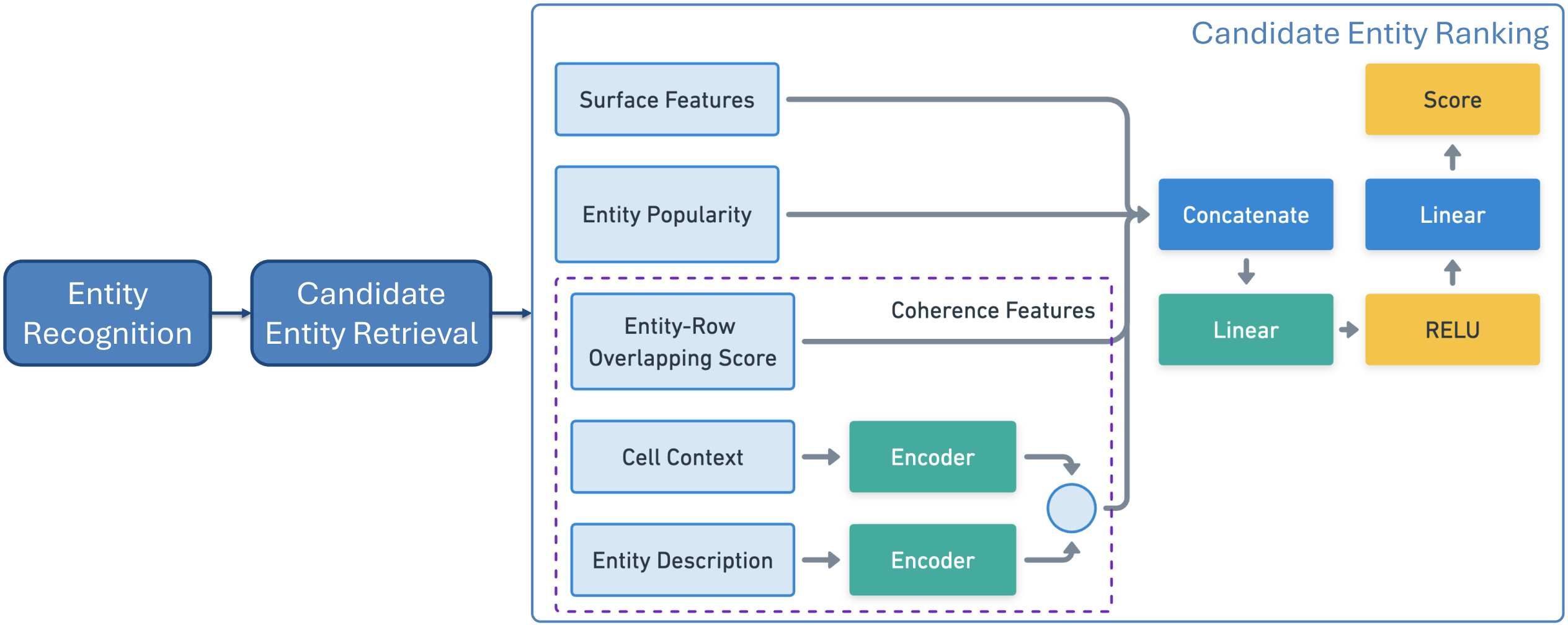
Block list

Cell	Entity	Label
[0, 0]: Dan Carter	**Dan Carter** (Q726199)	1
[0, 0]: Dan Carter	**Dan Carter** (Q5213238)	0
[0, 0]: Dan Carter	**Dan Carter** (Q59277840)	0

Wikipedia markup for creating links:
- [[New Zealand]]
- [[New Zealand Rugby Team| New Zealand]]

Training dataset for entity linking

Entity Linking

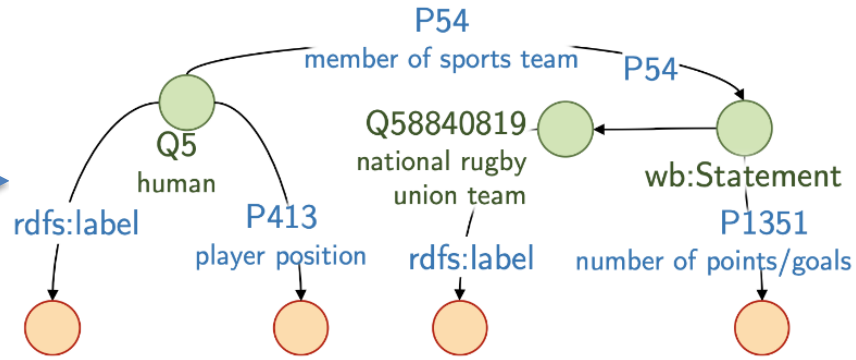


Generating Labeled Data (2)

Wikipedia tables

Player	Position	Team	Points
Dan Carter	Fly-half	New Zealand	1598
Ronan O'Gara	Fly-half	Ireland	1083
Percy Montgomery	Fullback	South Africa	893

GRAMS



Group by columns & keep high confidence relationships

Source	Target	Relationship	Label
Player	Points	**member of sport team → number of points**	**1**
Player	Team	**member of sport team**	**1**

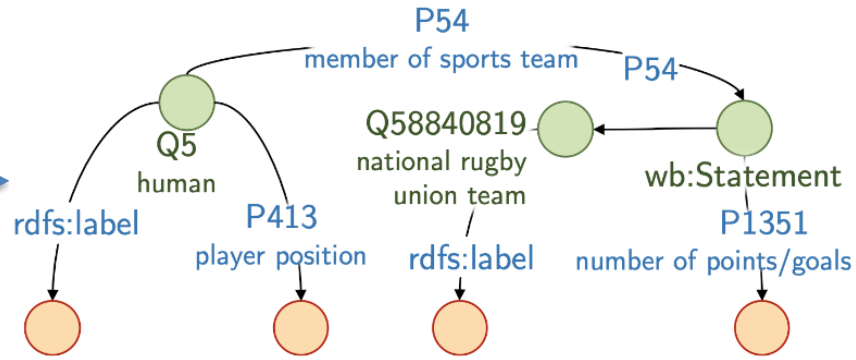
Training dataset for column relationship prediction

Generating Labeled Data (2)

Wikipedia tables

Player	Position	Team	Points
Dan Carter	Fly-half	New Zealand	1598
Ronan O'Gara	Fly-half	Ireland	1083
Percy Montgomery	Fullback	South Africa	893

GRAMS



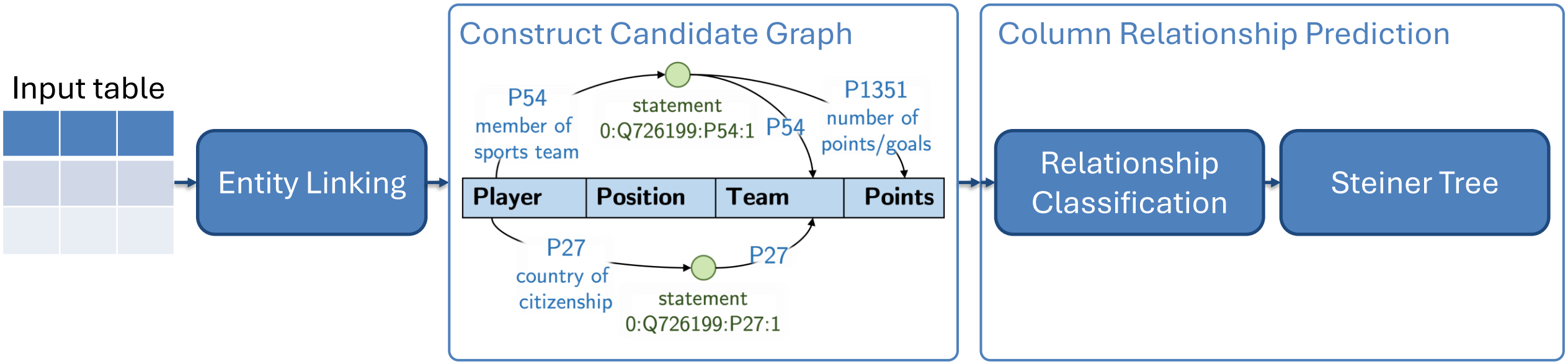
Entity Linking + GRAMS

Group by columns & keep high confidence relationships

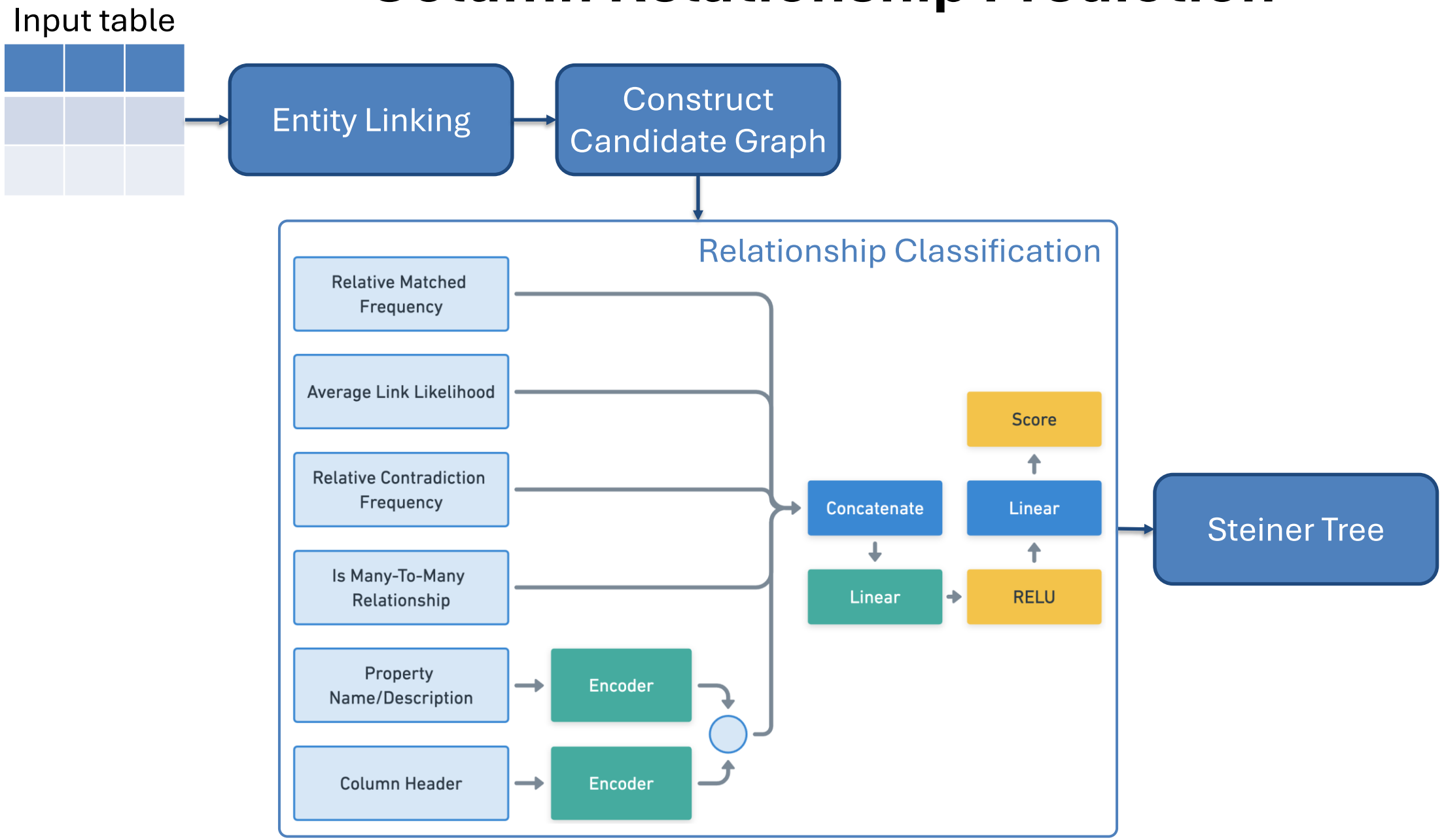
Source	Target	Relationship	Label
Player	Points	**member of sport team → number of points**	**1**
Player	Points	height	0
Player	Team	**member of sport team**	**1**
Player	Team	country of citizenship	0
Player	Team	coach	0

Training dataset for column relationship prediction

Column Relationship Prediction



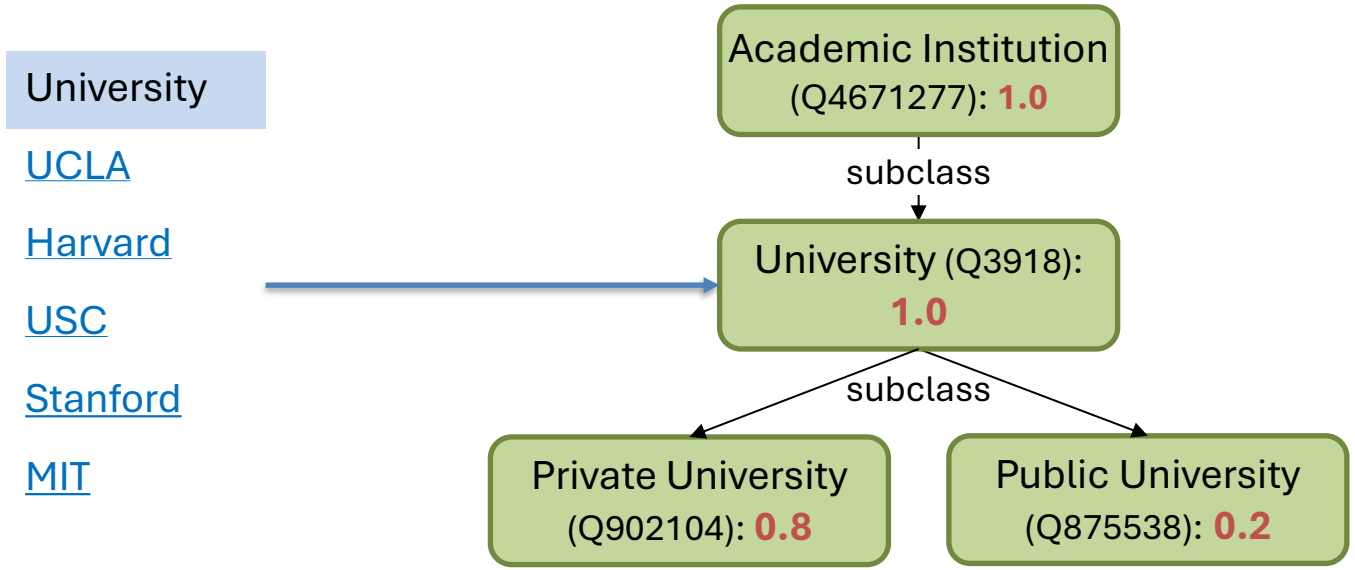
Column Relationship Prediction



Column Type Prediction

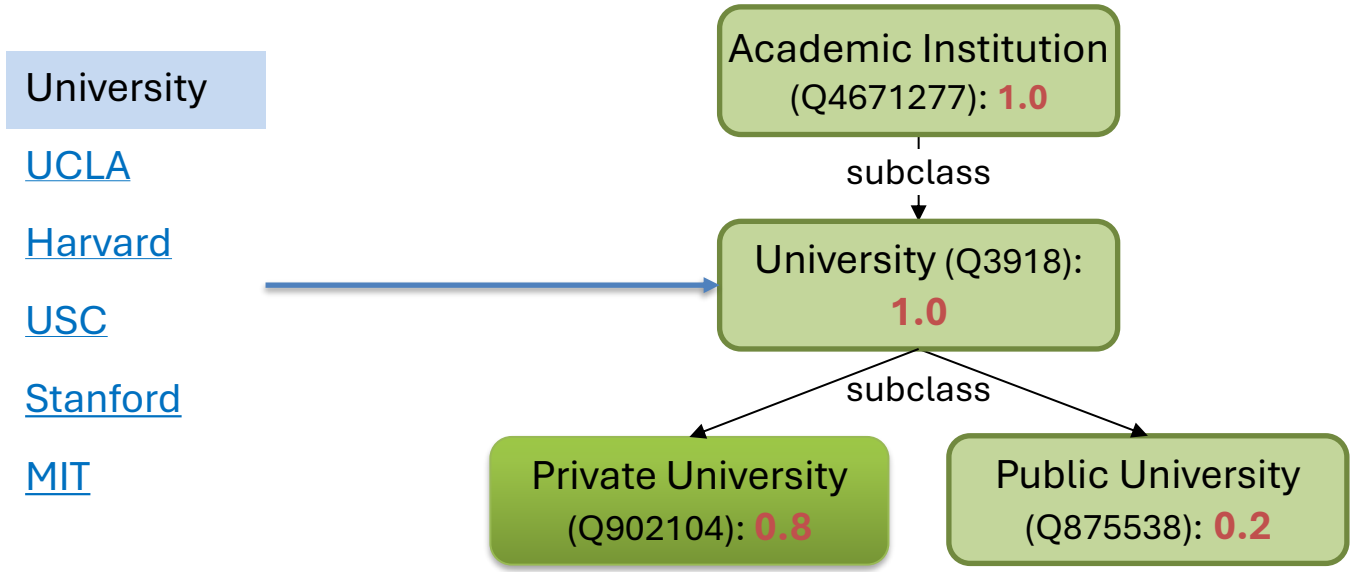
sum of scores of entities belong to the type

- Simple yet effective approach
  - Selecting the most specific type that has the highest score
  - Repeatedly refining the prediction: choose the parent type if the new score is increased by at least δ



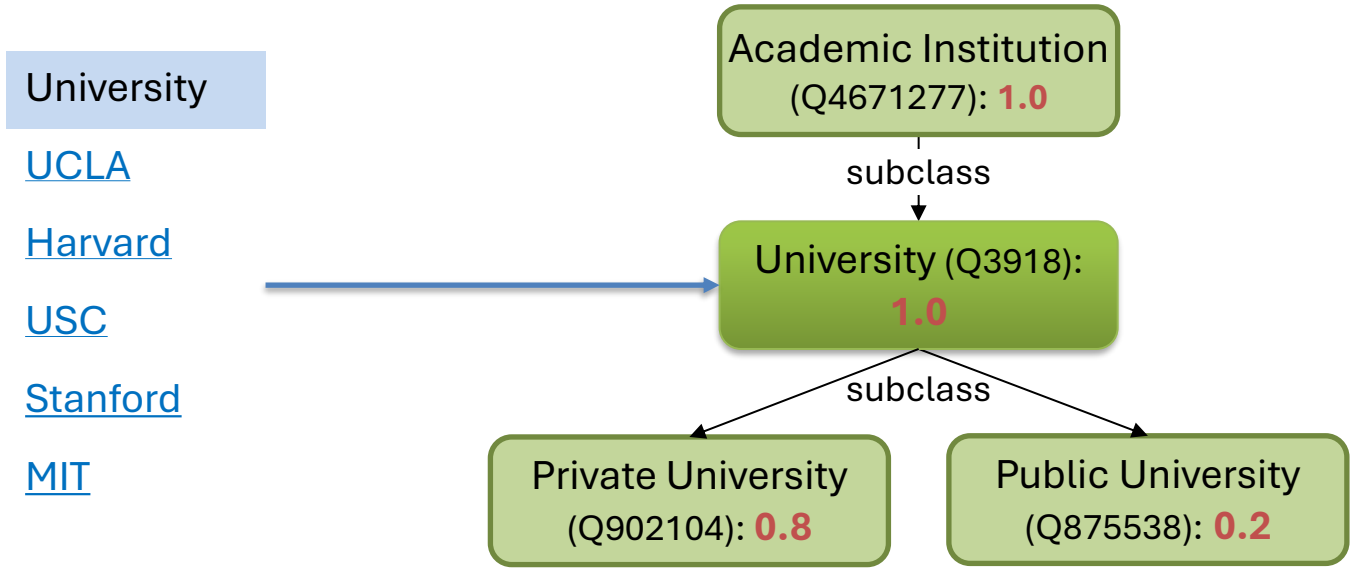
Column Type Prediction

- Simple yet effective approach
  - Selecting the most specific type that has the highest score
  - Repeatedly refining the prediction: choose the parent type if the new score is increased by at least δ



Column Type Prediction

- Simple yet effective approach
  - Selecting the most specific type that has the highest score
  - Repeatedly refining the prediction: choose the parent type if the new score is increased by at least δ



Evaluation

- GRAMS+ outperforms SOTA baselines on real-world dataset
  - **4.57%** on column relationships annotation (CPA)
  - **4.95%** on column type annotation (CTA)

Method	CPA			CTA		
	AP	AR	AF_1	AP	AR	AF_1
GRAMS+	**82.79%**	62.06%	**66.41%**	**80.54%**	78.26%	**78.94%**
MTab	73.28%	50.72%	54.09%	64.61%	71.9%	67.16%
DAGOBAH	76.42%	60.92%	61.84%	71.47%	**78.95%**	73.99%
GRAMS	75.86%	42.27%	44.71%	66.33%	77.38%	70.45%
KGCode-Tab	28.25%	**70.01%**	36.52%	42.28%	58.03%	47.82%

Evaluation

- Performance of GRAMS+ on SemTab dataset

Method	CPA			CTA		
	AP	AR	AF_1	AP	AR	AF_1
GRAMS+	98.92%	91.20%	91.71%	94.36%	90.58%	90.63%
MTab	98.78%	**94.06%**	**94.1%**	**94.87%**	**91.5%**	**91.53%**
DAGOBAH	**99.2%**	36.40%	37.10%	92.9%	90.0%	90.1%
GRAMS	95.64%	83.72%	84.57%	80.37%	87.64%	80.54%
KGCode-Tab	88.33%	86.79%	81.22%	74.86%	80.42%	73.64%

both methods achieve 94.8% if we give them target columns to predict

fail on tables with only 3 – 4 rows + entities' types were changed

Evaluation

- Performance of GRAMS+ on SemTab dataset

Method	CPA			CTA		
	AP	AR	AF_1	AP	AR	AF_1
GRAMS+	98.92%	91.20%	91.71%	94.36%	90.58%	90.63%
MTab	98.78%	**94.06%**	**94.1%**	**94.87%**	**91.5%**	**91.53%**
DAGOBAH	**99.2%**	36.40%	37.10%	92.9%	90.0%	90.1%
GRAMS	95.64%	83.72%	84.57%	80.37%	87.64%	80.54%
KGCode-Tab	88.33%	86.79%	81.22%	74.86%	80.42%	73.64%

- Performance of GRAMS+ in same evaluation setting as SemTab

Method	CPA			CTA		
	m-P	m-R	$m\text{-}F_1$	m-AP	m-AR	$m\text{-}AF_1$
GRAMS+	98.18%	**97.93%**	98.06%	94.02%	93.98%	94.00%
MTab	**99.01%**	97.81%	**98.4%**	95.09%	95.09%	95.09%
DAGOBAH [25]	99%	97.8%	**98.4%**	**97.5%**	**97.5%**	**97.5%**
KGCode-Tab	98.19%	86.85%	92.17%	86.73%	81.40%	83.98%
KGCode-Tab [34]	94.4%	94.0%	94.2%	91.8%	89.43%	90.6%
SemTex [23]	-	-	97.05%	-	-	93.85%

Evaluation

- Performance of GRAMS+ with different entity ranking methods

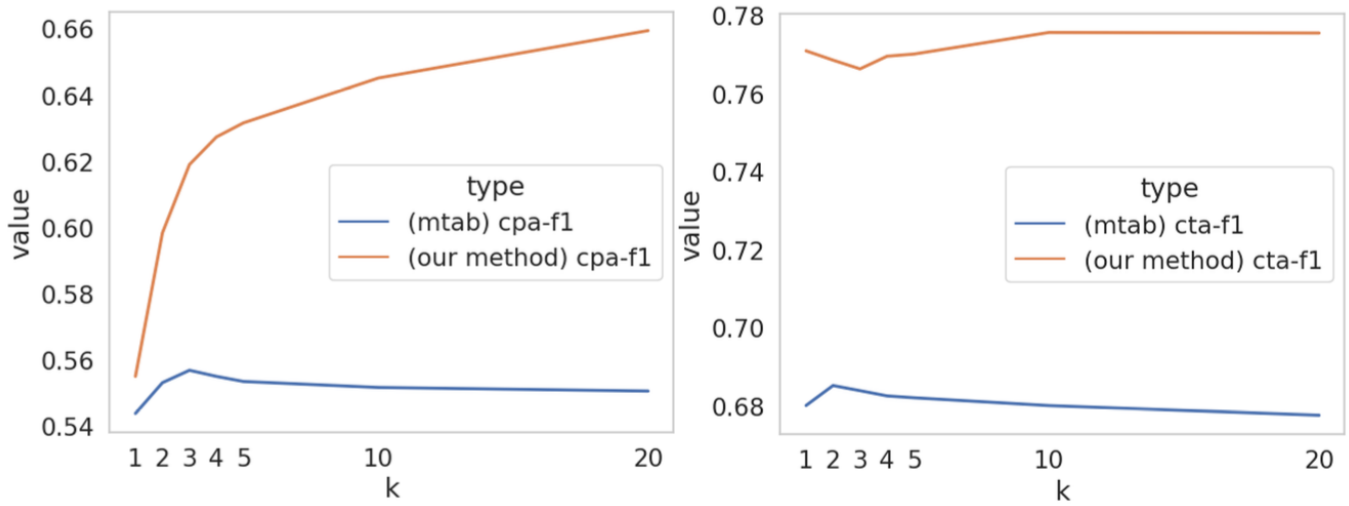
Method	CPA			CTA		
	AP	AR	AF_1	AP	AR	AF_1
Our Entity Linking Score	82.79%	62.06%	66.41%	80.54%	78.26%	78.94%
Retrieval Score	80.74%	61.00%	64.41%	73.60%	71.70%	72.22%

Evaluation

- Performance of GRAMS+ with different entity ranking methods

Method	CPA			CTA		
	AP	AR	AF_1	AP	AR	AF_1
Our Entity Linking Score	82.79%	62.06%	66.41%	80.54%	78.26%	78.94%
Retrieval Score	80.74%	61.00%	64.41%	73.60%	71.70%	72.22%

- Performance of GRAMS+ & MTab with different number of candidate entities



Summary

- GRAMS+ is a distant supervised approach for unlinked tables
- More robust to noise and can leverage metadata to overcome ambiguous tables
- Limitation:
  - assuming overlapping data with KG

Outline

- Introduction
- Thesis Overview
- Creating Semantic Descriptions of Linked Tables
- Creating Semantic Descriptions of Tables with Overlapping Data
- **Creating Semantic Descriptions of Tables without Overlapping Data**
- Related Work
- Conclusion and Future Work

Motivating Example

P413 position played on team

assoc. football position Q4611891

human Q5

P413 position played on team

Using column header & label/desc

Americ. football position Q694589

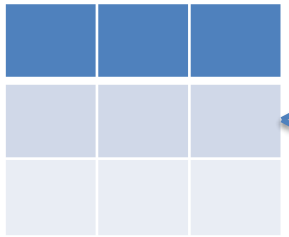
rdfs:label

Player	Height	Position	Team
Tom Aldred	1.88	Defender	Brisbane Roar
Shae Cahill	1.71	Midfielder	Brisbane Roar

Using column values & examples of a class/property

Approach

Input table

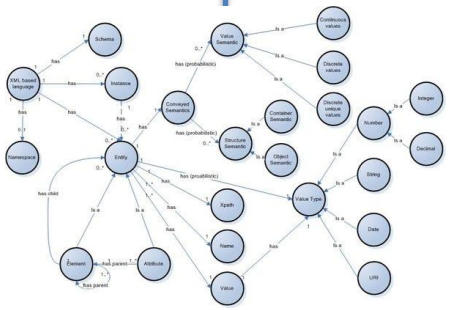


Column Incoming Relationship Prediction

Column Type Prediction

Construct Candidate Graph

Infer Semantic Description



Input ontology

Column Concept Prediction

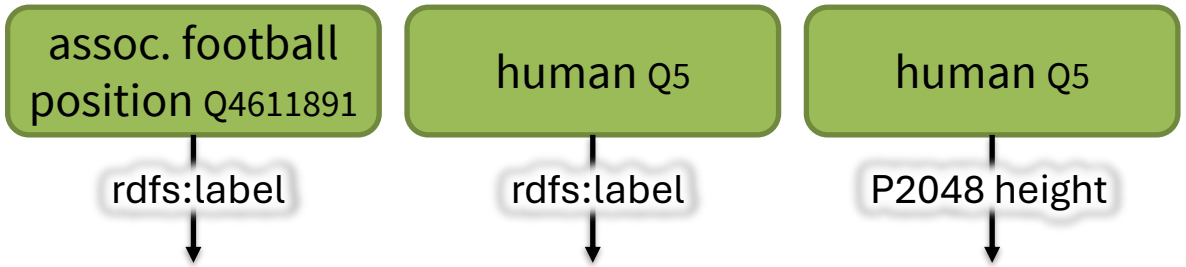
- Goal: predict class/incoming property of a column

Position
Defender
Midfielder
Goalkeeper

Concept	Score
Q4611891 assoc. football position	
P413 position played on team	
P39 position held	
…	

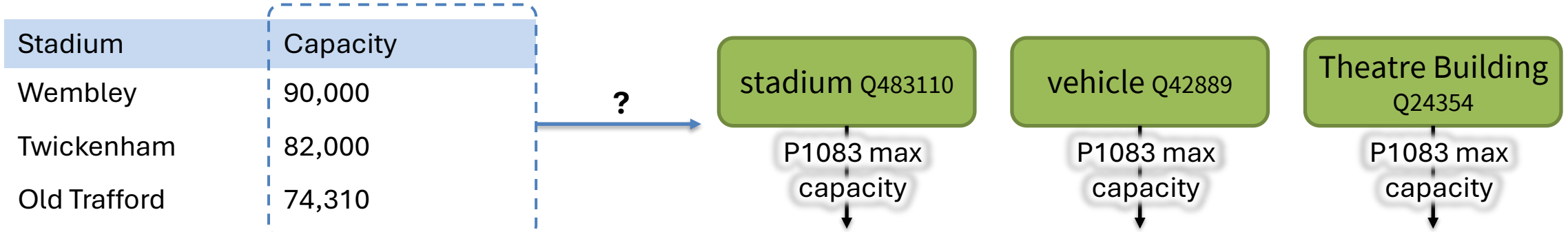
- Different from Semantic Labeling

Position
Defender
Midfielder
Goalkeeper



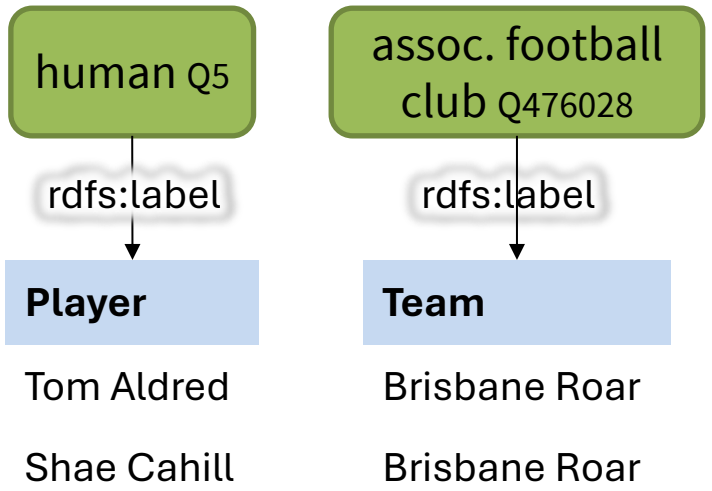
Column Concept Prediction

- Semantic Labeling is not ideal for this task
  - Column often contains only information about the corresponding property

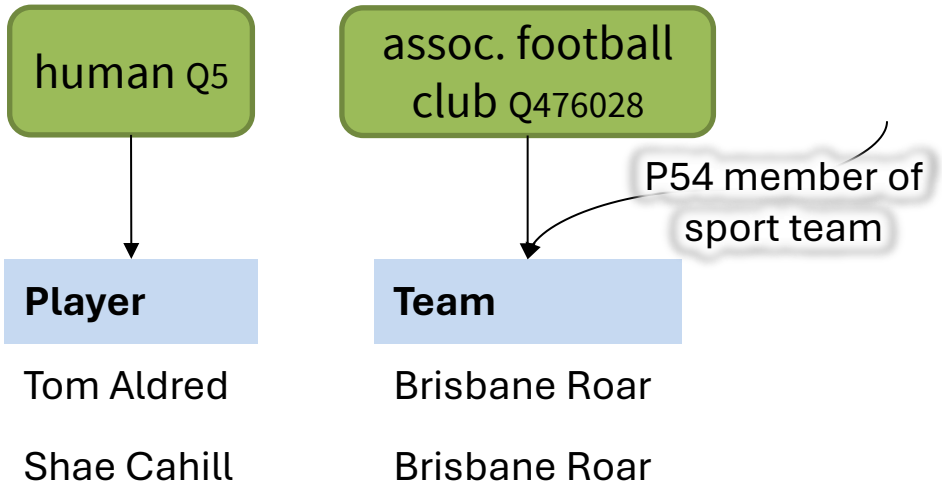


Column Concept Prediction

- Semantic Labeling is not ideal for this task
  - Not capturing information that is useful for later processing step



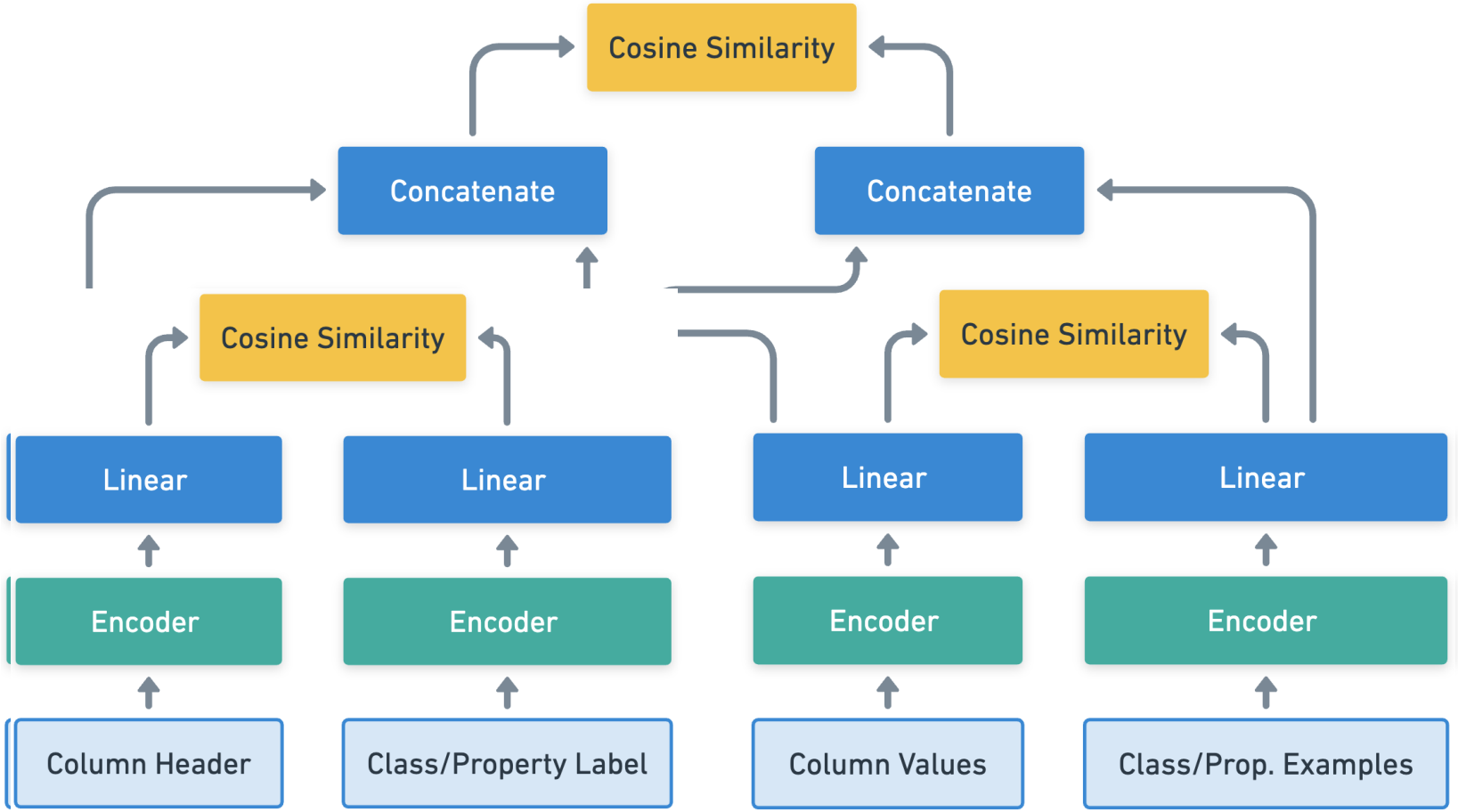
SEMANTIC LABELING OUTPUT



CONCEPT PREDICTION OUTPUT

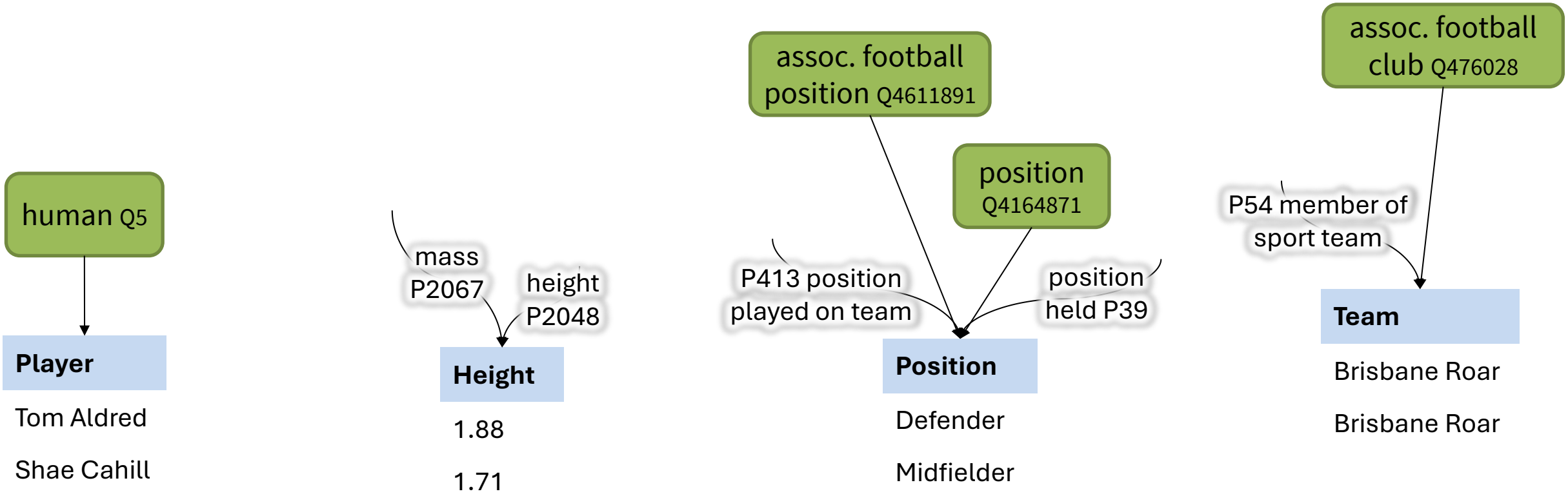
Column Concept Prediction

- Goal: learn a semantic relatedness score between a column and class/property.
- Multi-task learning with triplet loss $\mathcal{L}(\mathbf{x}, \mathbf{c}^+, \mathbf{c}^-) = \max(0, \mathbf{dis}(\mathbf{x}, \mathbf{c}^+) - \mathbf{dis}(\mathbf{x}, \mathbf{c}^-) + \epsilon)$



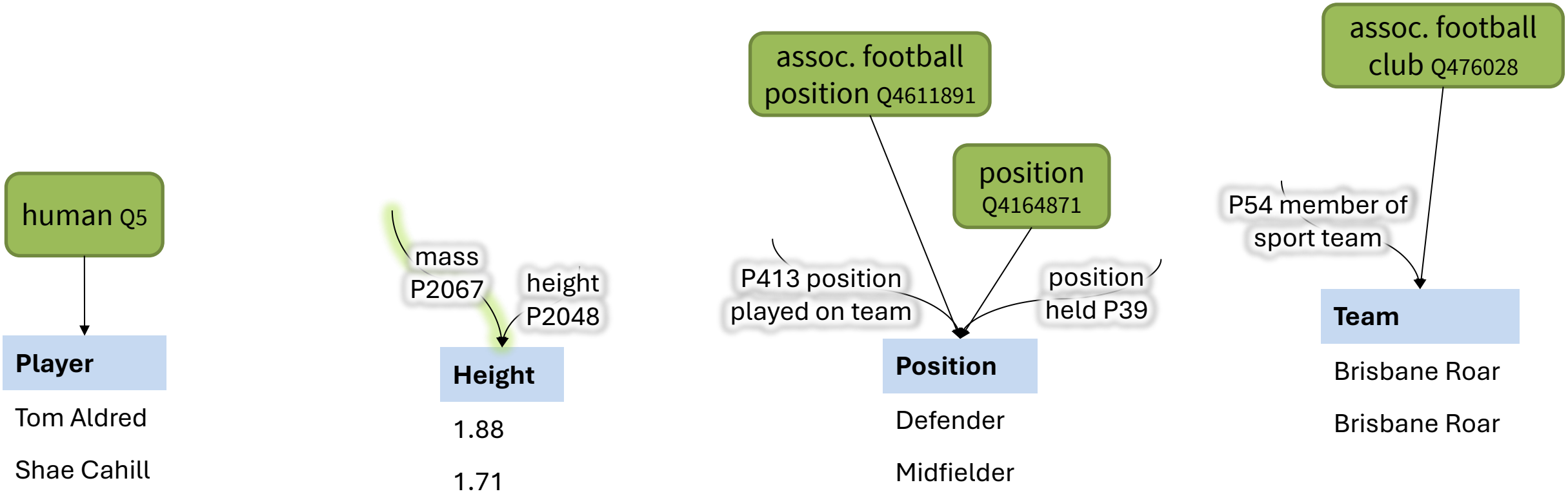
Candidate Graph Construction

- Results of column concept prediction



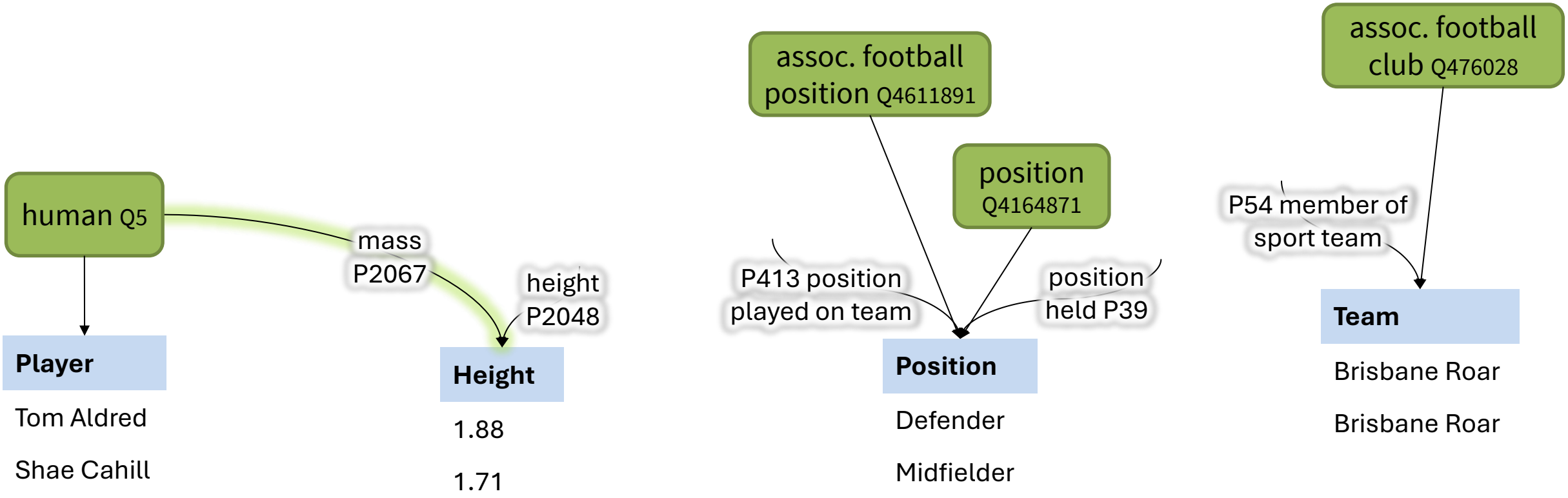
Candidate Graph Construction

- Connecting relationships to classes



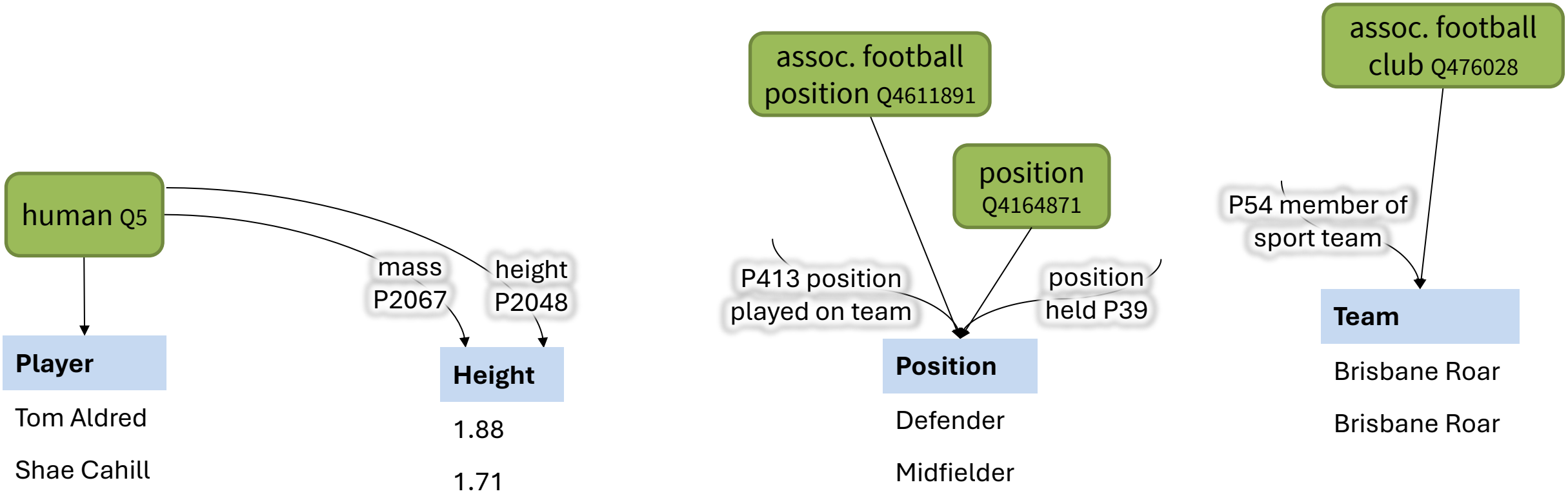
Candidate Graph Construction

- Connecting relationships to classes



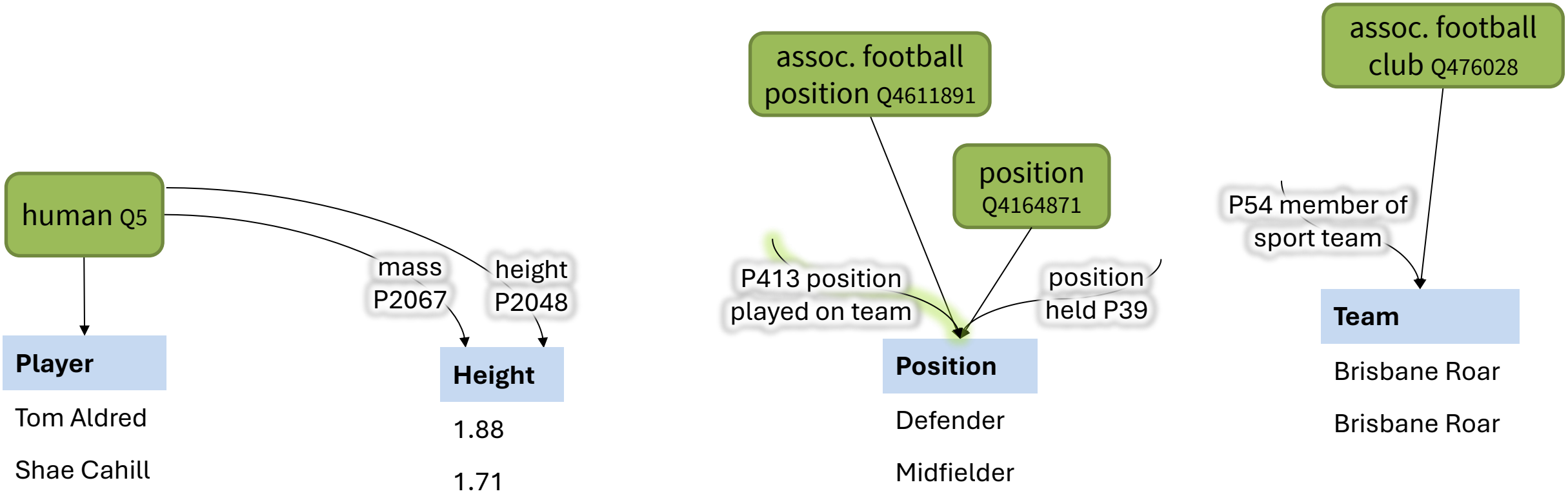
Candidate Graph Construction

- Connecting relationships to classes



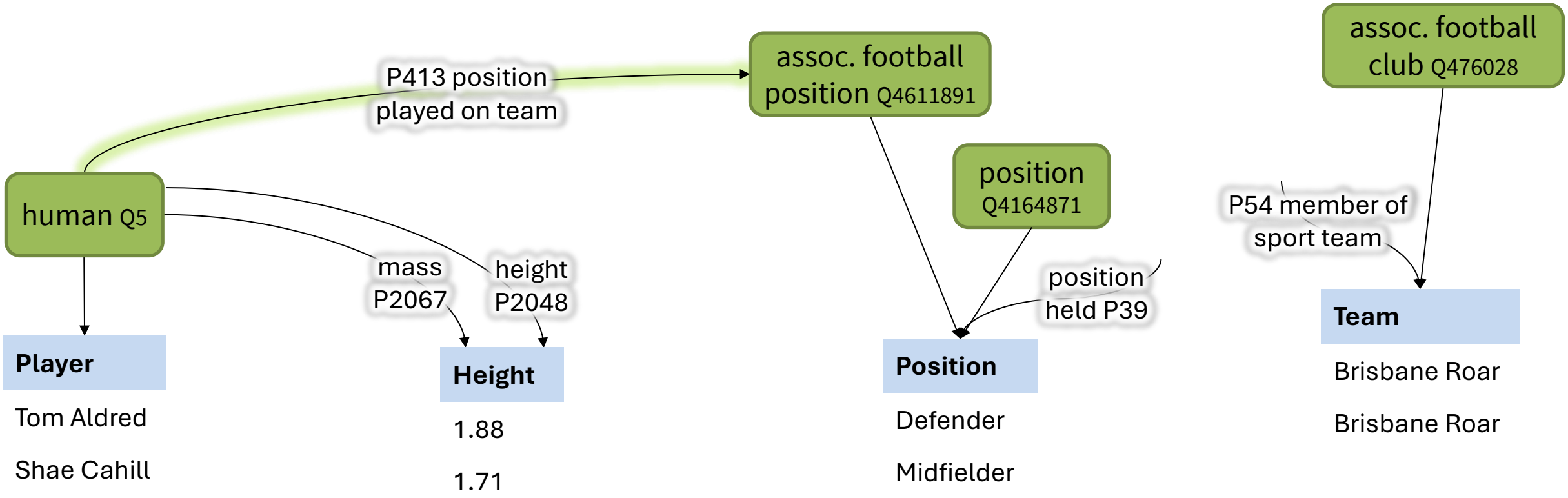
Candidate Graph Construction

- Connecting relationships to classes



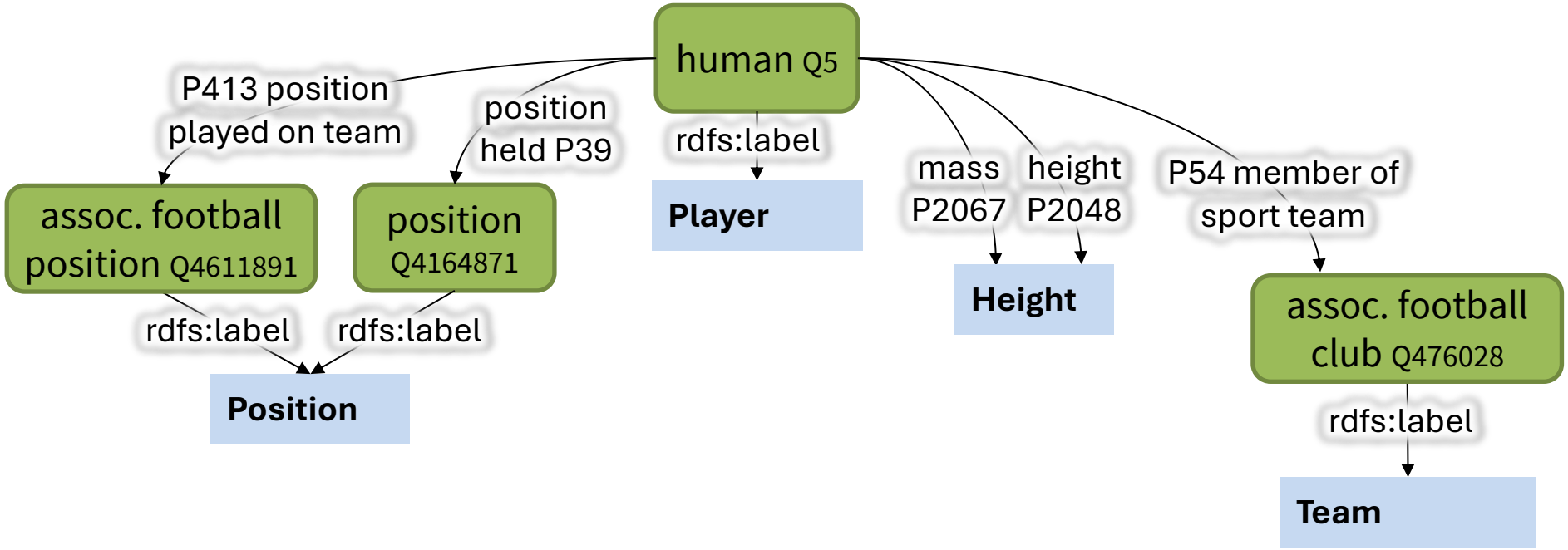
Candidate Graph Construction

- Connecting relationships to classes



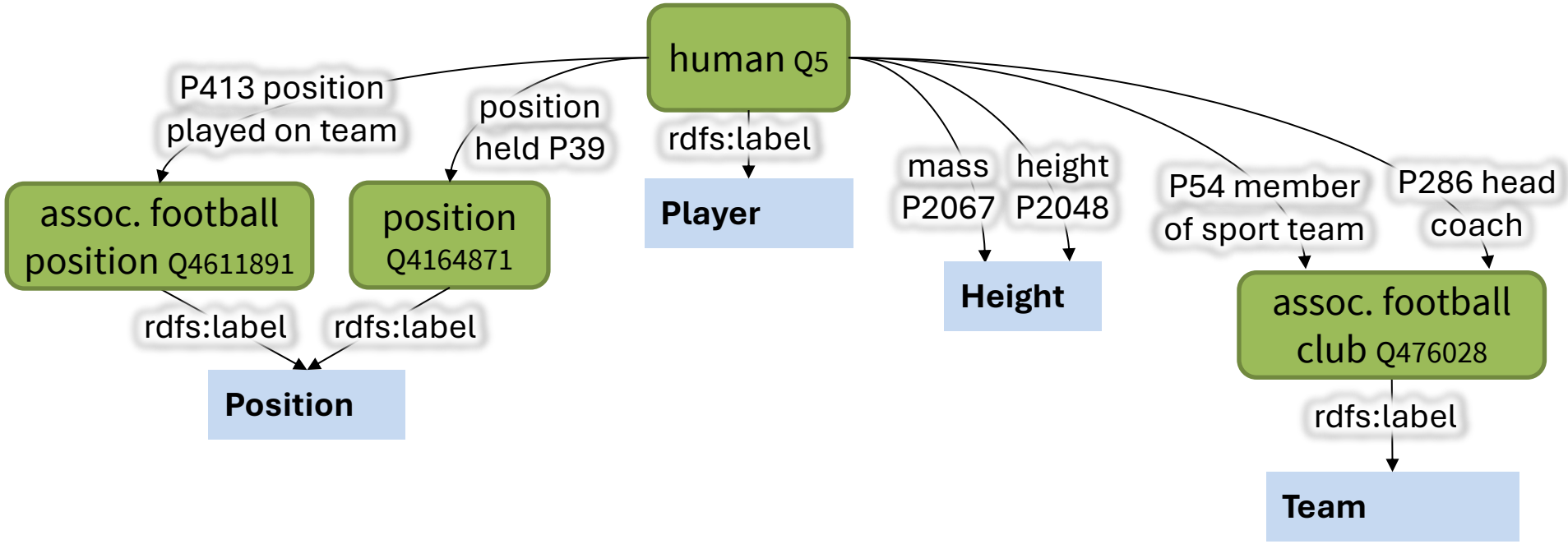
Candidate Graph Construction

- Connecting relationships to classes



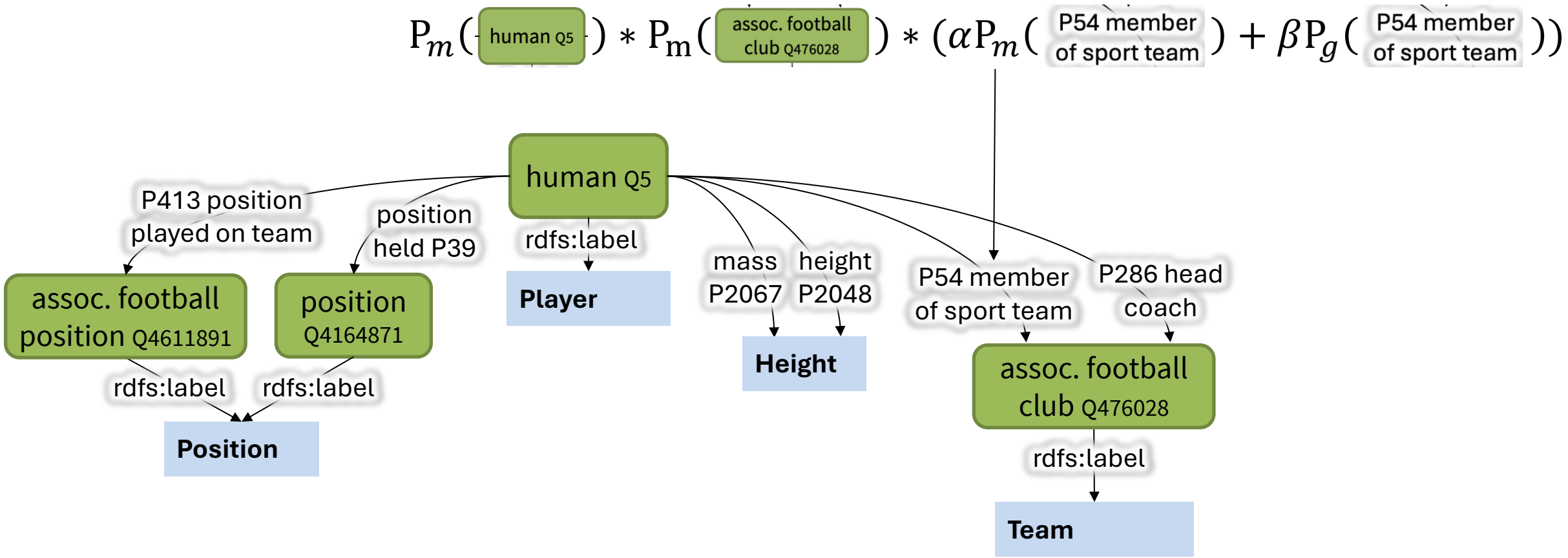
Candidate Graph Construction

- Connecting relationships to classes



Inferring Semantic Description

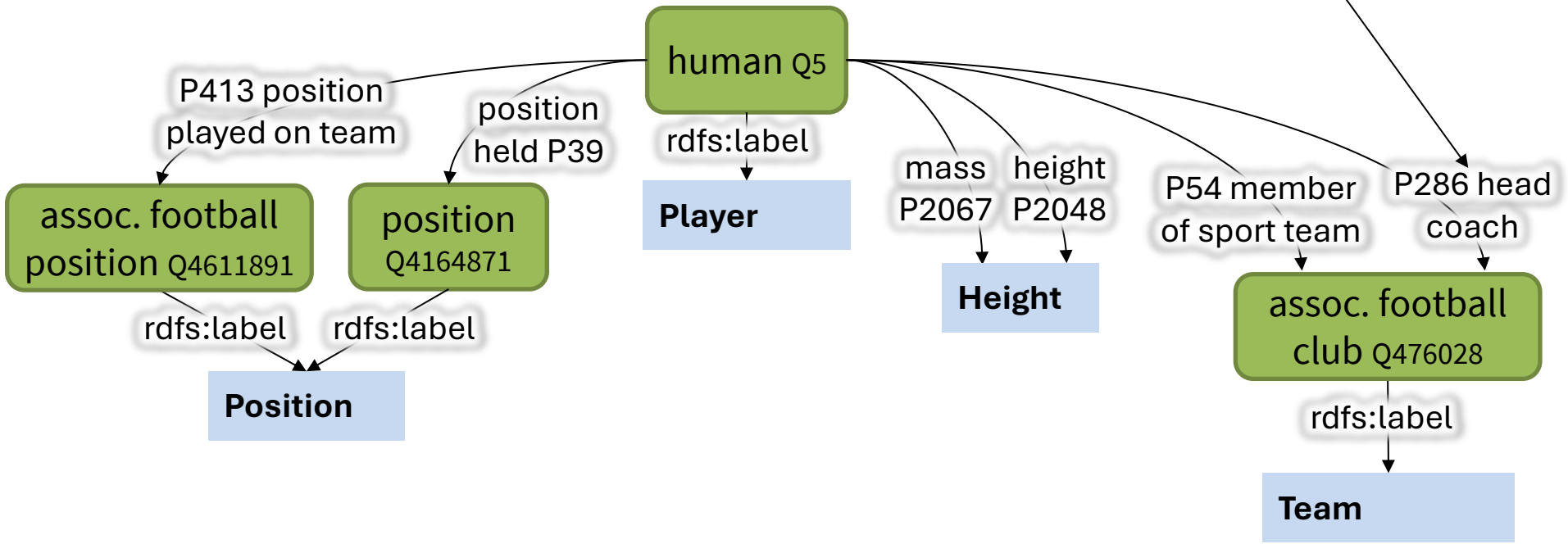
- Finding a Steiner Tree that maximize average edge likelihood
- Selecting classes that have the highest probabilities



Inferring Semantic Description

- Finding a Steiner Tree that maximize average edge likelihood
- Selecting classes that have the highest probabilities

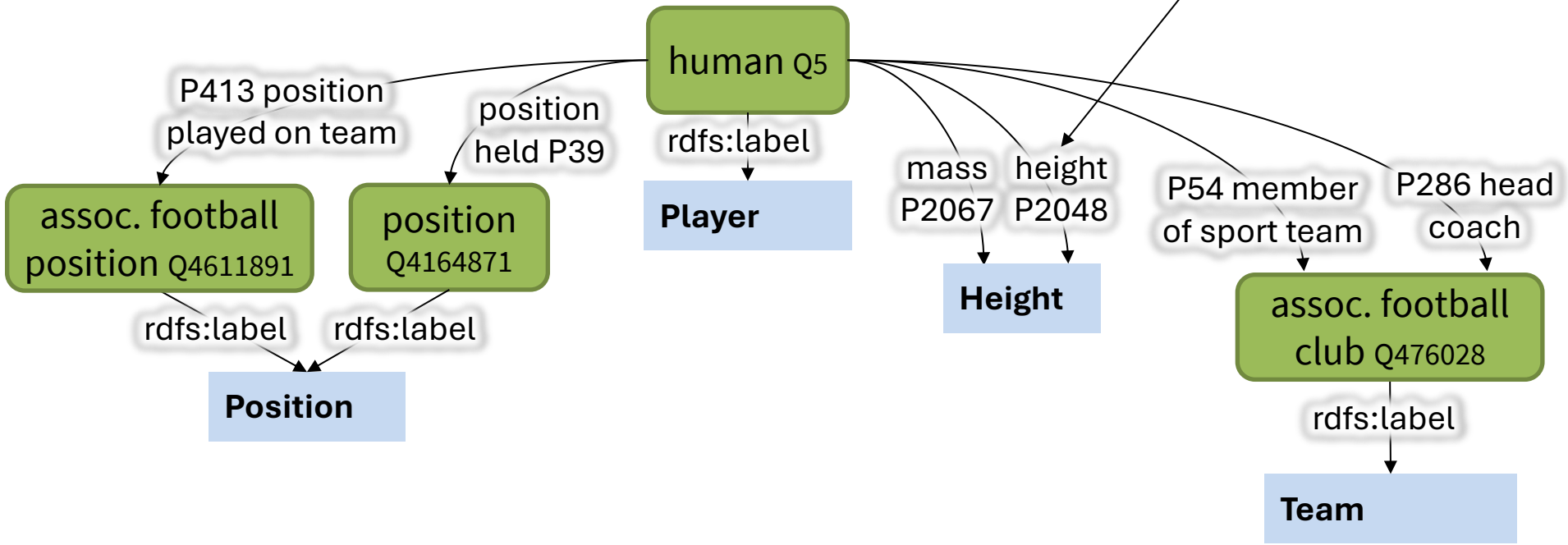
$$P_m(\text{human Q5}) * P_m(\text{assoc. football club Q476028}) * (\alpha * 0 + \beta P_g(\text{P286 head coach}))$$



Inferring Semantic Description

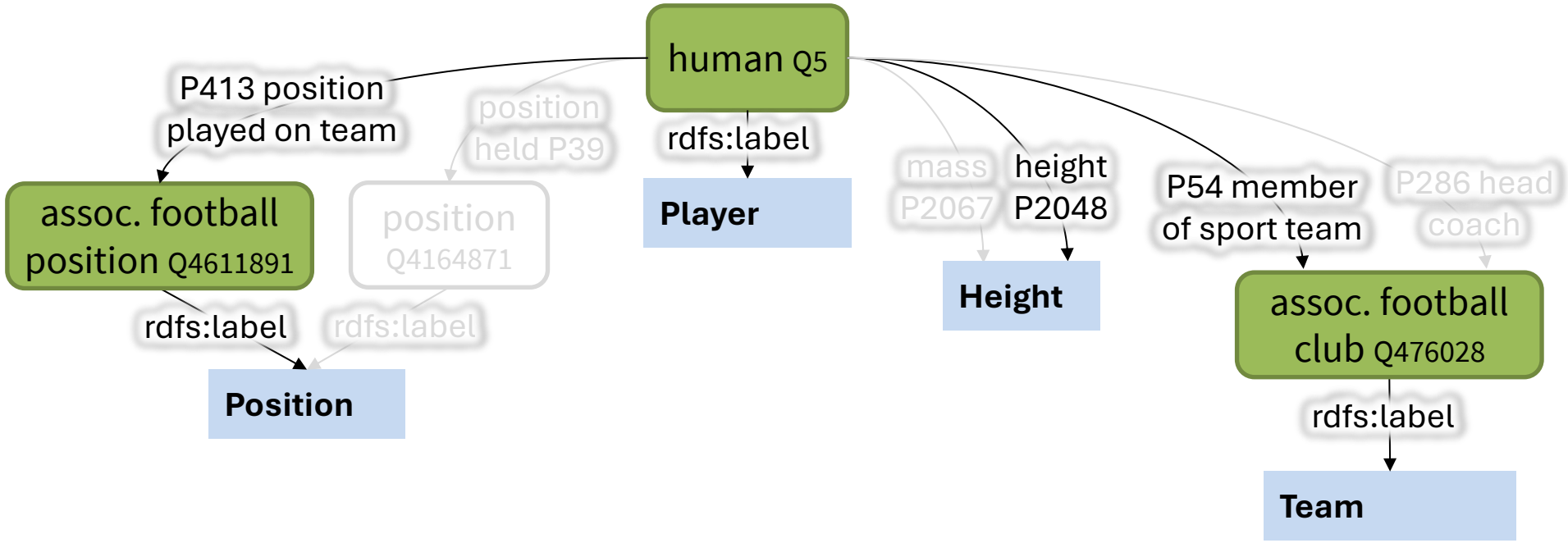
- Finding a Steiner Tree that maximize average edge likelihood
- Selecting classes that have the highest probabilities

$$\mathrm{P}_m(\text{human Q5}) * 1 * (\alpha * \mathrm{P}_m(\text{height P2048}) + \beta \mathrm{P}_g(\text{height P2048}))$$



Inferring Semantic Description

- Finding a Steiner Tree that maximize average edge likelihood
- Selecting classes that have the highest probabilities



Evaluation

- Datasets: 250WT (Wikidata ontology) and t2dv2 (Dbpedia ontology)
  - 58% of classes & 68% of properties in 250WT are not in the training set

Table 5.1: Statistics of datasets used in the evaluation

	train dataset	250WT	T2Dv2
#tables	18261	250	224
#classes	326	155	36
#properties	104	112	103

- Baselines:
  - DSL-SM
  - LLMs

Evaluation

- Datasets: 250WT (Wikidata ontology) and t2dv2 (Dbpedia ontology)
  - 58% of classes & 68% of properties in 250WT are not in the training set

Table 5.1: Statistics of datasets used in the evaluation

	train dataset	250WT	T2Dv2
#tables	18261	250	224
#classes	326	155	36
#properties	104	112	103

- Baselines:
  - DSL-SM: perform semantic labeling, then pred object prop. based on freq. pattern
  - LLMs

Evaluation

- Datasets: 250WT (Wikidata ontology) and t2dv2 (Dbpedia ontology)
  - 58% of classes & 68% of properties in 250WT are not in the training set

Table 5.1: Statistics of datasets used in the evaluation

	train dataset	250WT	T2Dv2
#tables	18261	250	224
#classes	326	155	36
#properties	104	112	103

- Baselines:
  - DSL-SM: perform semantic labeling, then pred object prop. based on freq. pattern
  - LLMs: give a table and the ontology, then ask to predict class/property

Evaluation

- Performance of GRAMS++ on 250WT and T2Dv2 datasets
  - 250WT: outperform by **31.96%** and **42.85%** on CPA and CTA, respectively.
  - T2Dv2: outperform by **6.7%** on CPA and comparable performance on CTA.

250WT

T2Dv2

36 classes!

Method	CPA			CTA		
	AP	AR	AF_1	AP	AR	AF_1
DSL-SM	19.53%	22.03%	19.71%	27.76%	27.97%	27.37%
Llama2-70B	18.86%	47.84%	26.54%	30.44%	32.11%	31.00%
Llama2-7B	06.22%	17.17%	08.98%	27.61%	28.14%	27.73%
OLMo	26.70%	22.00%	15.79%	20.78%	10.09%	09.82%
GRAMS++	**60.82%**	**57.65%**	**58.50%**	**73.74%**	**74.79%**	**73.85%**
DSL-SM	51.14%	48.57%	48.69%	80.26%	88.94%	82.95%
Llama2-70B	46.32%	**68.42%**	51.96%	**87.97%**	**93.21%**	**87.43%**
Llama2-7B	33.97%	44.22%	35.29%	67.42%	74.32%	69.58%
OLMo	64.99%	21.73%	21.79%	30.67%	25.64%	18.75%
GRAMS++	**59.04%**	59.1%	**58.66%**	83.41%	91.80%	85.99%

Evaluation

- Performance of GRAMS++ on 250WT and T2Dv2 datasets
  - 250WT: outperform by **31.96%** and **42.85%** on CPA and CTA, respectively.
  - T2Dv2: outperform by **6.7%** on CPA and comparable performance on CTA.

	Method	CPA			CTA		
		AP	AR	AF_1	AP	AR	AF_1
250WT	DSL-SM	19.53%	22.03%	19.71%	27.76%	27.97%	27.37%
	Llama2-70B	18.86%	47.84%	26.54%	30.44%	32.11%	31.00%
	Llama2-7B	06.22%	17.17%	08.98%	27.61%	28.14%	27.73%
	OLMo	26.70%	22.00%	15.79%	20.78%	10.09%	09.82%
	GRAMS++	**60.82%**	**57.65%**	**58.50%**	**73.74%**	**74.79%**	**73.85%**
T2Dv2	DSL-SM	51.14%	48.57%	48.69%	80.26%	88.94%	82.95%
	Llama2-70B	46.32%	**68.42%**	51.96%	**87.97%**	**93.21%**	**87.43%**
	Llama2-7B	33.97%	44.22%	35.29%	67.42%	74.32%	69.58%
	OLMo	64.99%	21.73%	21.79%	30.67%	25.64%	18.75%
	GRAMS++	**59.04%**	59.1%	**58.66%**	83.41%	91.80%	85.99%

Evaluation

- Performance of GRAMS++ on 250WT and T2Dv2 datasets
  - 250WT: outperform by **31.96%** and **42.85%** on CPA and CTA, respectively.
  - T2Dv2: outperform by **6.7%** on CPA and comparable performance on CTA.

	Method	CPA			CTA		
		AP	AR	AF_1	AP	AR	AF_1
250WT	DSL-SM	19.53%	22.03%	19.71%	27.76%	27.97%	27.37%
	Llama2-70B	18.86%	47.84%	26.54%	30.44%	32.11%	31.00%
	Llama2-7B	06.22%	17.17%	08.98%	27.61%	28.14%	27.73%
	OLMo	26.70%	22.00%	15.79%	20.78%	10.09%	09.82%
	GRAMS++	**60.82%**	**57.65%**	**58.50%**	**73.74%**	**74.79%**	**73.85%**
T2Dv2	DSL-SM	51.14%	48.57%	48.69%	80.26%	88.94%	82.95%
	Llama2-70B	46.32%	**68.42%**	51.96%	**87.97%**	**93.21%**	**87.43%**
	Llama2-7B	33.97%	44.22%	35.29%	67.42%	74.32%	69.58%
	OLMo	64.99%	21.73%	21.79%	30.67%	25.64%	18.75%
	GRAMS++	**59.04%**	59.1%	**58.66%**	83.41%	91.80%	85.99%

Summary

- GRAMS++ is a distant supervised approach for tables without overlapping data
  - Can train on one domain and apply to different domain
  - Better performance and more efficient than LLMs

Related Work

	Method		Does not Require Manually Labeled Sources	Unlinked Table	Not requiring target columns	No overlapping data with KGs	Modeling Capabilities		
							Handle Literal Columns	Handle Qualifiers	Denormalized Tables
Supervised Approach	Taheriyan et al. 2016		N	Y	Y	Y	Y	Y	Y
	Una et al. 2018		N	Y	Y	Y	Y	Y	Y
	Suhara et al. 2022 - DODUO		N (huge)	Y	N	Y	Y	N	Y
	Vu et al. 2019		N	Y	Y	Y	Y	Y	Y
Value-linked Approach	Iterative Method	Ritze et al. 2015	Y	Y	Y	N	Y	N	N
		Zhang et al. 2017	Y	Y	Y	N	Y	N	N
		SemTab systems	Y	Y	Y	N	Y	N	N
	Graphical Models	Limaye et al. 2010	Y	Y	Y	N	N	N	Y
		Mulward et al. 2013	Y	Y	Y	N	N	N	Y
		GRAMS	Y	N	Y	N	Y	Y	Y
		GRAMS+	Y	Y	Y	N	Y	Y	Y
GRAMS++			Y	Y	Y	Y	Y	Y	Y

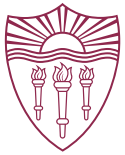
	Method		Does not Require Manually Labeled Sources	Unlinked Table	Not requiring target columns	No overlapping data with KGs	Modeling Capabilities		
							Handle Literal Columns	Handle Qualifiers	Denormalized Tables
Supervised Approach	Taheriyan et al. 2016		N	Y	Y	Y	Y	Y	Y
	Una et al. 2018		N	Y	Y	Y	Y	Y	Y
	Suhara et al. 2022 - DODUO		N (huge)	Y	N	Y	Y	N	Y
	Vu et al. 2019		N	Y	Y	Y	Y	Y	Y
Value-linked Approach	Iterative Method	Ritze et al. 2015	Y	Y	Y	N	Y	N	N
		Zhang et al. 2017	Y	Y	Y	N	Y	N	N
		SemTab systems	Y	Y	Y	N	Y	N	N
	Graphical Models	Limaye et al. 2010	Y	Y	Y	N	N	N	Y
		Mulward et al. 2013	Y	Y	Y	N	N	N	Y
		GRAMS	Y	N	Y	N	Y	Y	Y
		GRAMS+	Y	Y	Y	N	Y	Y	Y
GRAMS++			Y	Y	Y	Y	Y	Y	Y

Conclusion

- Comprehensive techniques for the semantic modeling problem under different settings and assumptions
  - Supervised method using PGM
  - Unsupervised method for linked tables
  - Distant supervised method for unlinked tables with overlapping data
  - Distant supervised method for unlinked tables without overlapping data

Future Work

- Integrating with table format & layout detection systems
- Estimating quality of the predicted semantic descriptions
- Improving the accuracy of predicted semantic descriptions
  - Unknown column detection
  - Handling tables that each row has a different semantic description



Thank you! Q/A