Learning the Semantics of Structured Data Sources

Mohsen Taheriyan

Department of Computer Science

Information Sciences Institute

USC Viterbi School of Engineering

Dissertation Committee

Craig Knoblock (PhD Advisor)

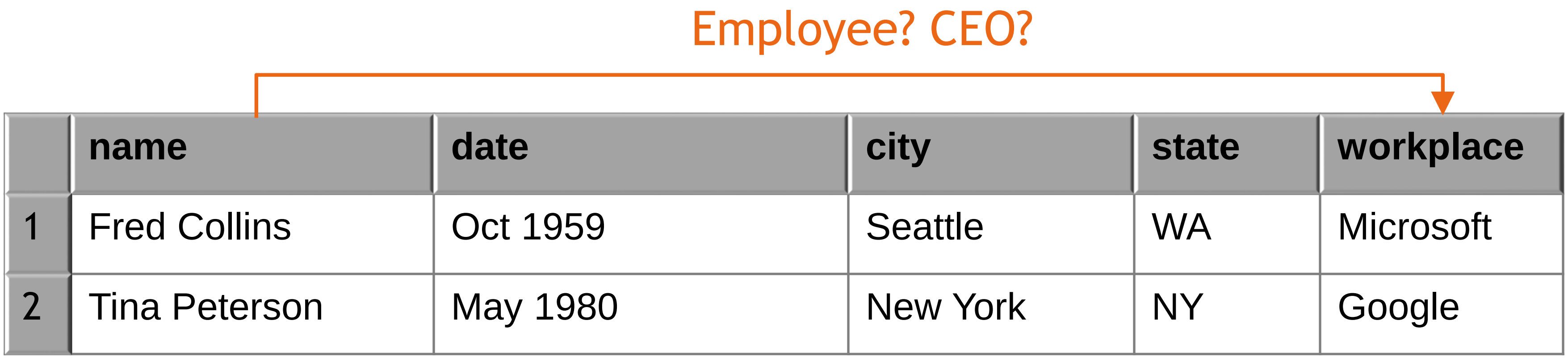
Cyrus Shahabi

Pedro Szekely

Viktor Prasanna (EE Department)

Motivation

Explicit semantics is missing in many of the structured sources



Employee? CEO?

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

Person?
Organization?

Birth date?
Death date?
Employment date?

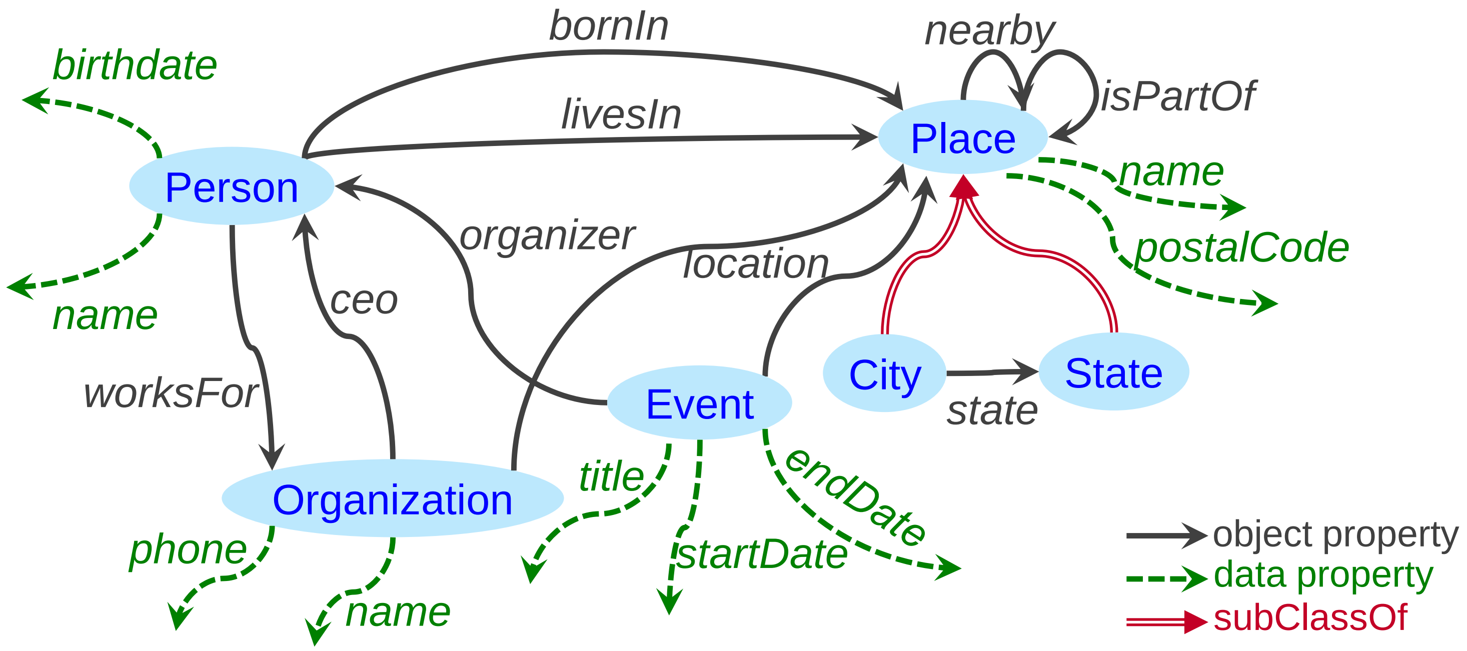
Birth city?
Work city?

How to express the intended meaning of data?

Map the Source to the Domain Ontology

Describe sources using classes & relationships in an ontology

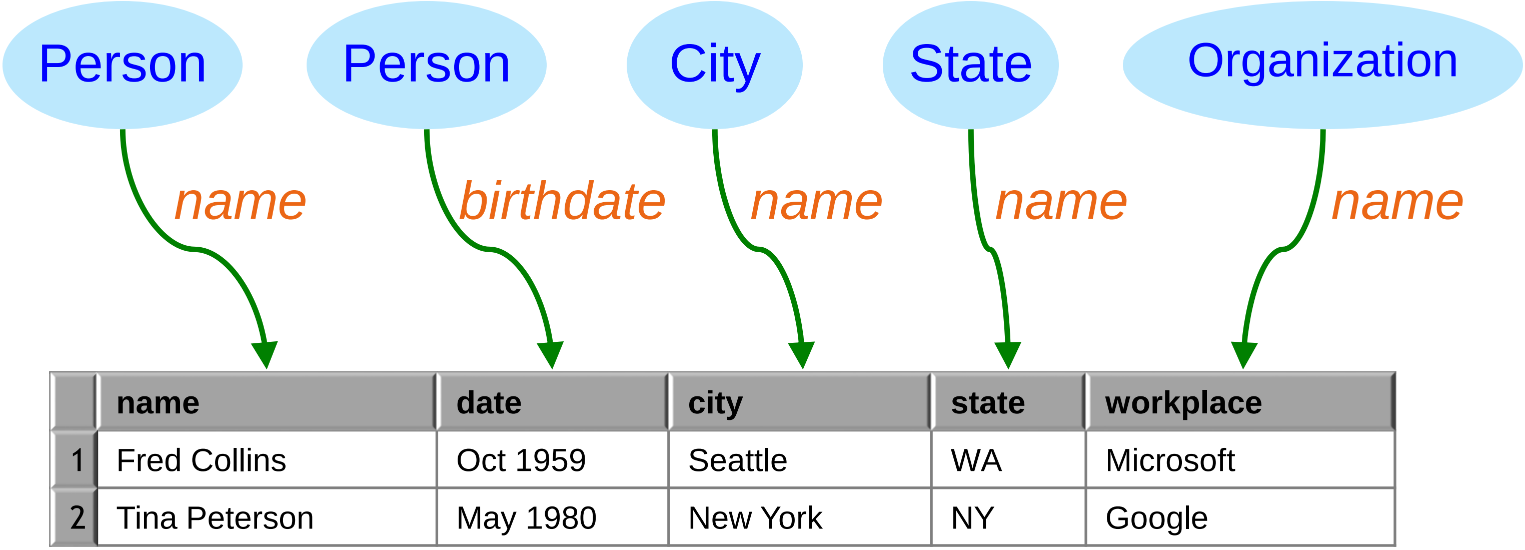
Domain Ontology



Source

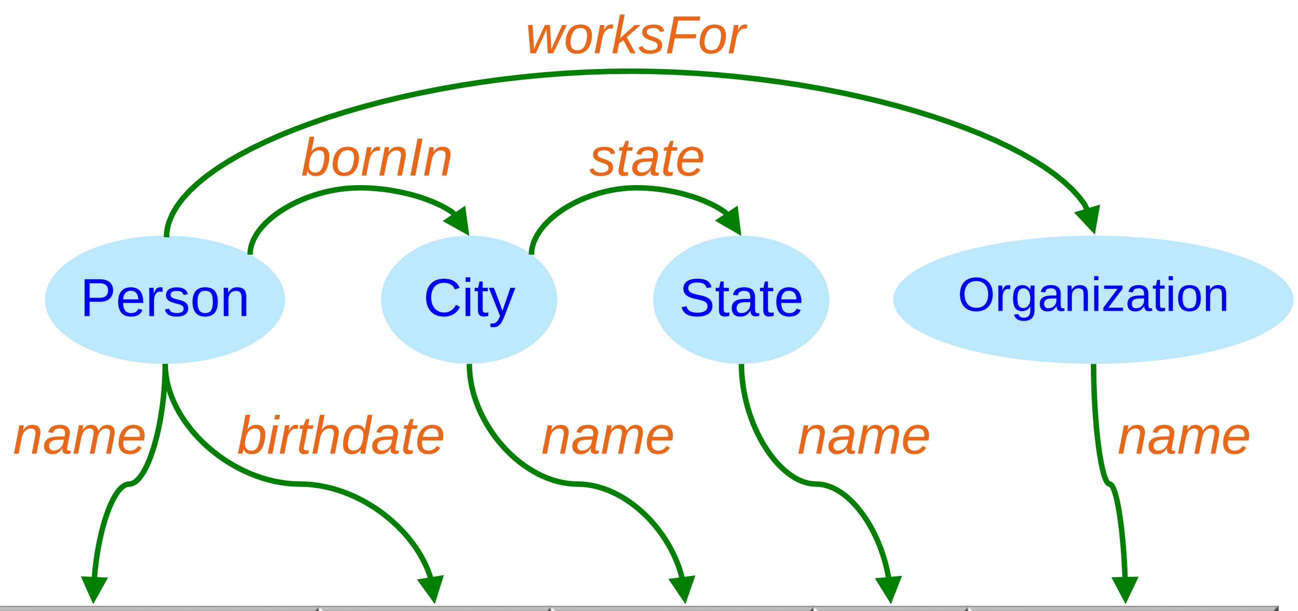
	name	date	city	state	workplace
1	Fred Collins	Oct 1959	Seattle	WA	Microsoft
2	Tina Peterson	May 1980	New York	NY	Google

Semantic Types



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

Relationships



worksFor

bornIn

state

Person

City

State

Organization

name

birthdate

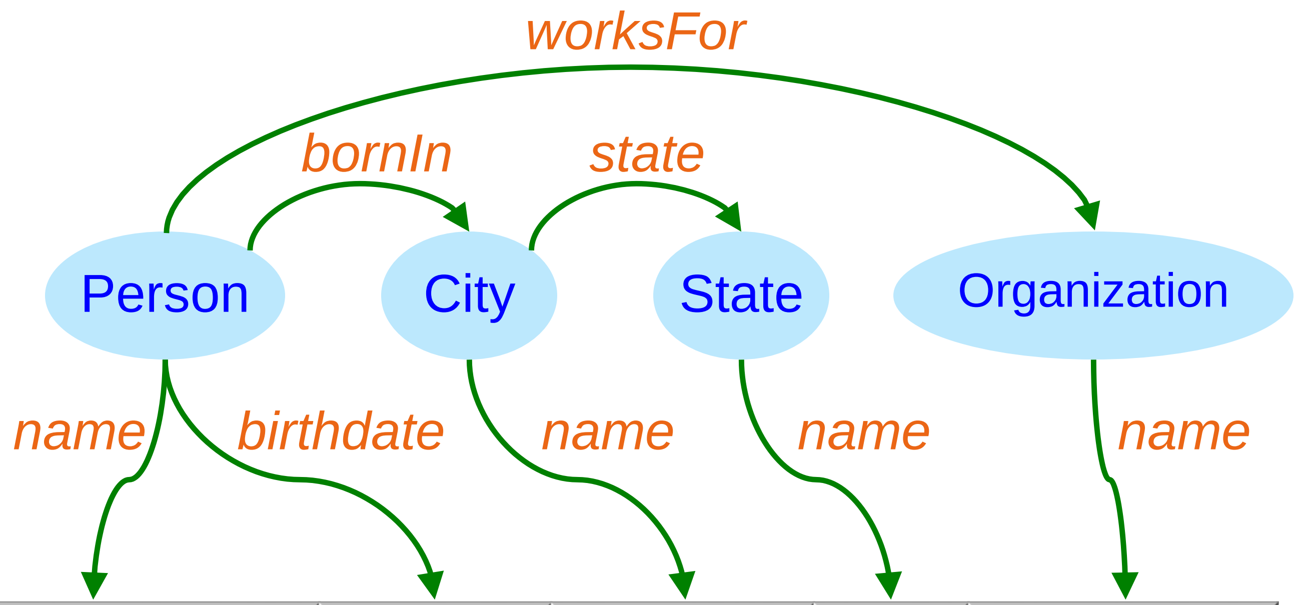
name

name

name

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

Semantic Model



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

Key ingredient to automate

- Source discovery
- Data integration
- Publish knowledge graphs

Problem:

How to automate building semantic models for structured sources?

Thesis Statement

The knowledge of previously modeled sources as well as the semantic data available in the Linked Open Data (LOD) cloud can be leveraged to learn accurate semantic models of structured data sources, enabling automated source discovery and data integration.

Outline

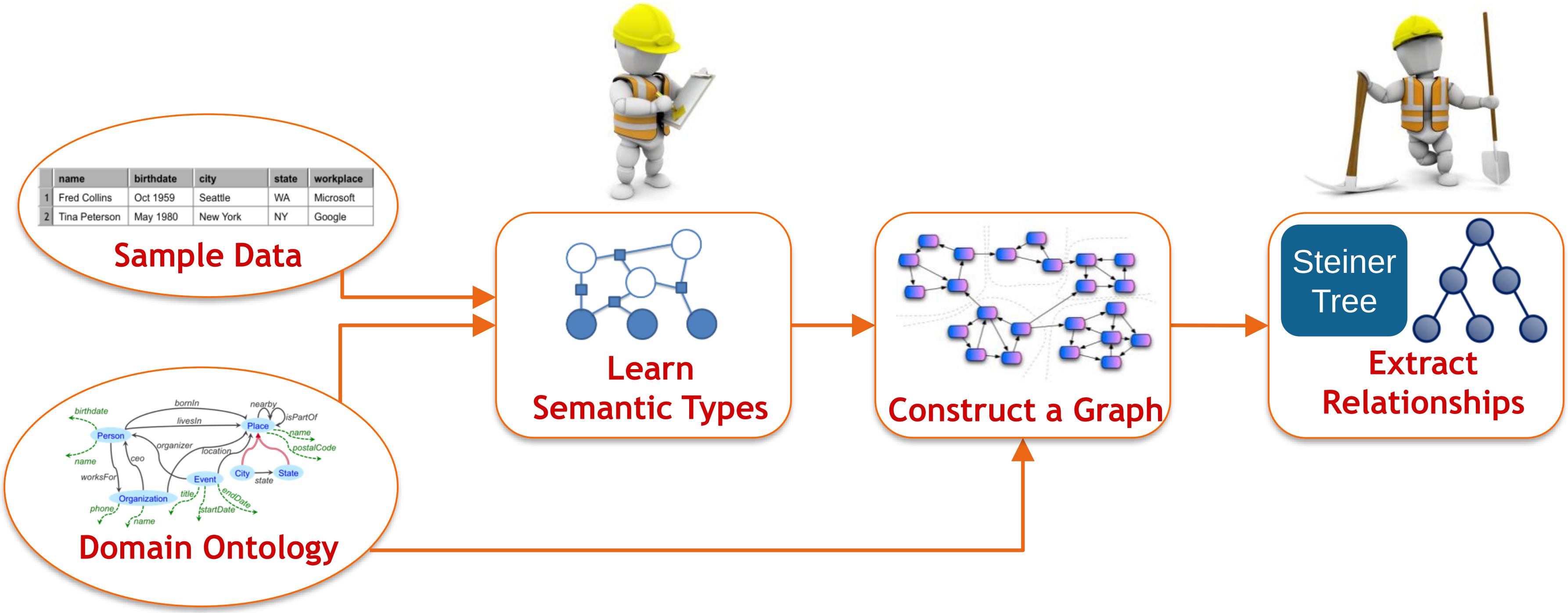
- Semi-automatically building semantic models
- Learning semantics models from known models
- Inferring semantic relations from LOD
- Related Work
- Discussion & Future Work

Semi-automatically Building Semantic Models

Contribution: a graph-based approach to extract implicit relationships

Approach

[Knoblock et al, ESWC 2012]



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

Sample Data

Domain Ontology

Learn Semantic Types

Construct a Graph

Steiner Tree

Extract Relationships

Implemented in Karma

http://www.isi.edu/integration/karma

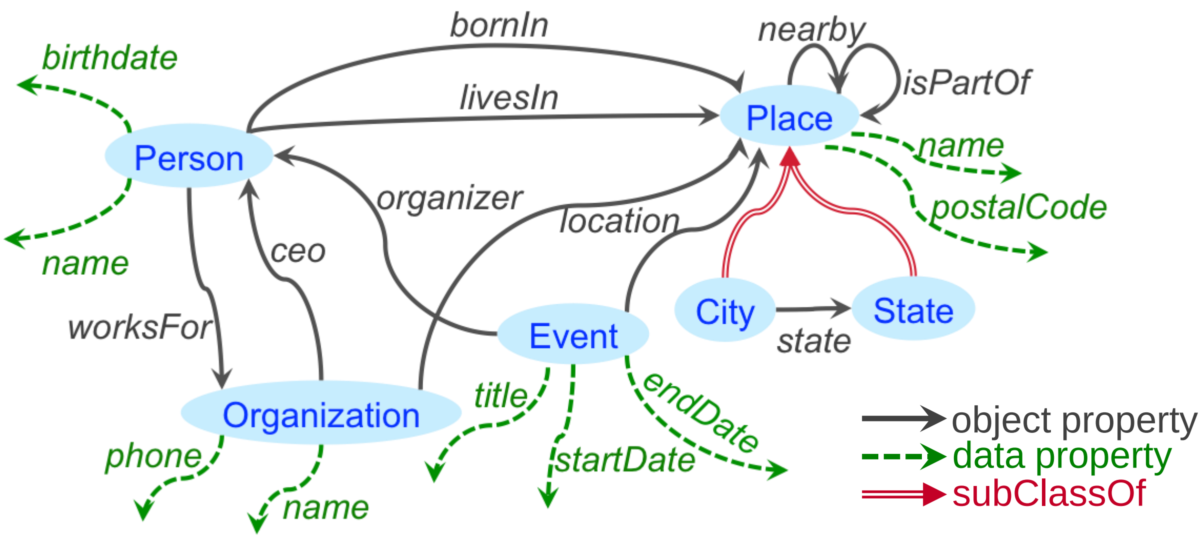
@KarmaSemWeb

Example

Source

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

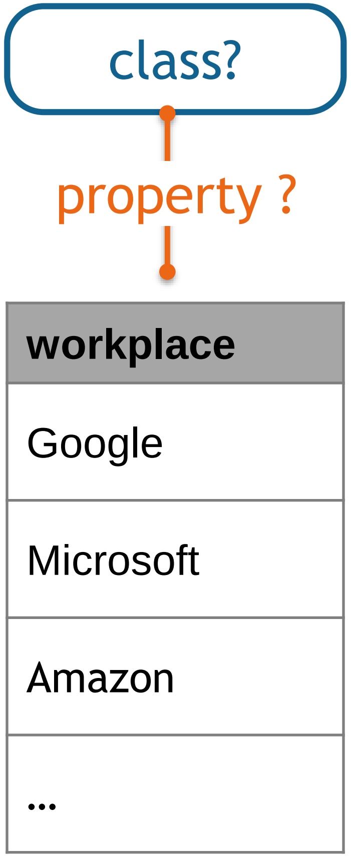
Domain Ontology



Goal: Find a semantic model for the source (map the source to the ontology)

Learning Semantic Types

[Krishnamurthy et al., ESWC 2015]

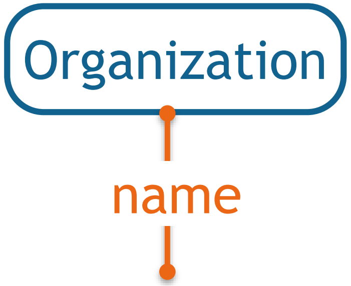


class?

property ?

workplace
Google
Microsoft
Amazon
…

Learning Semantic Types

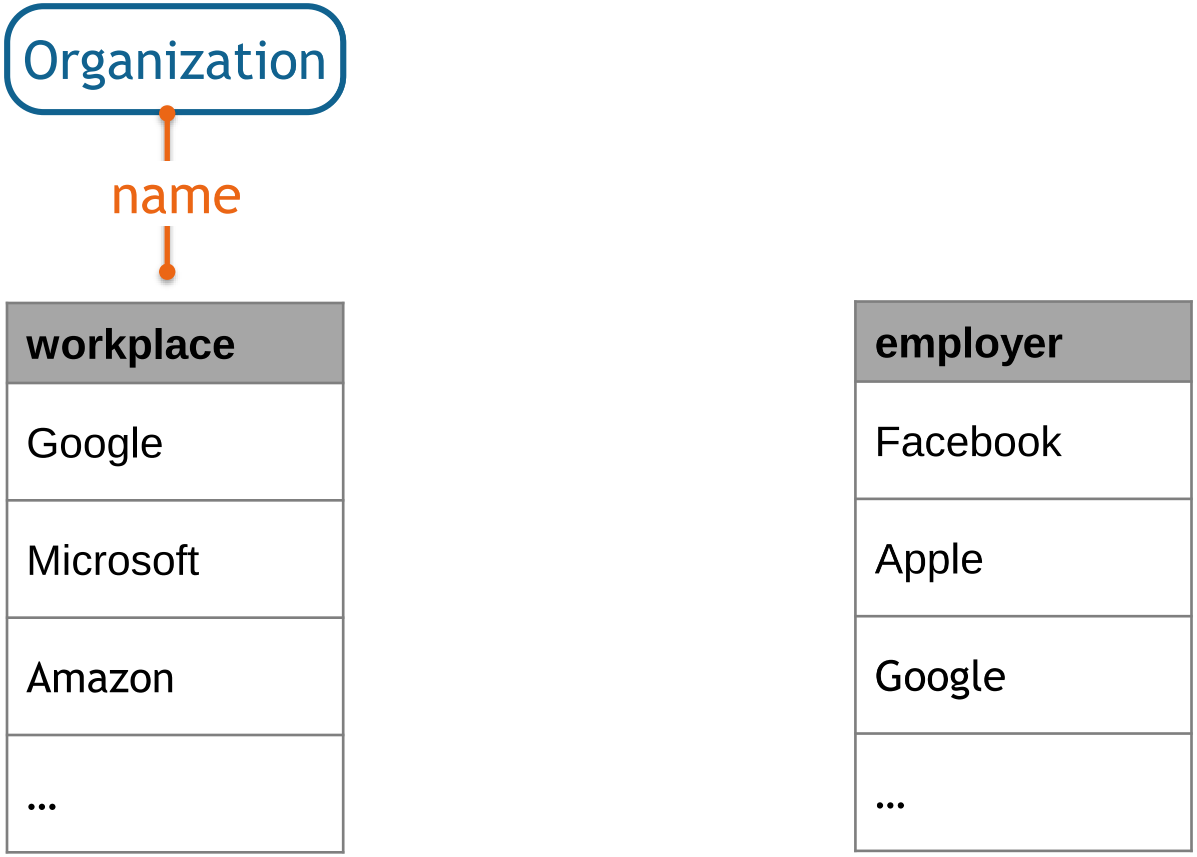

Organization

name

workplace
Google
Microsoft
Amazon
…

1. User Specifies
2. Systems learns

Learning Semantic Types



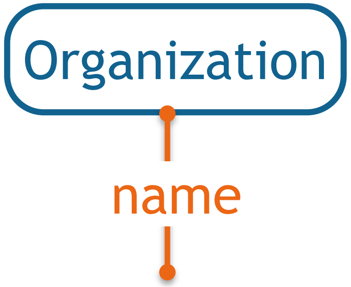
Organization

name

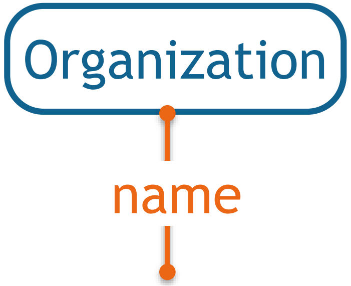
workplace
Google
Microsoft
Amazon
…

employer
Facebook
Apple
Google
…

Learning Semantic Types



workplace
Google
Microsoft
Amazon
…



employer
Facebook
Apple
Google
…

Semantic Labeling Approach

- Each semantic type: label
- Each data column: document

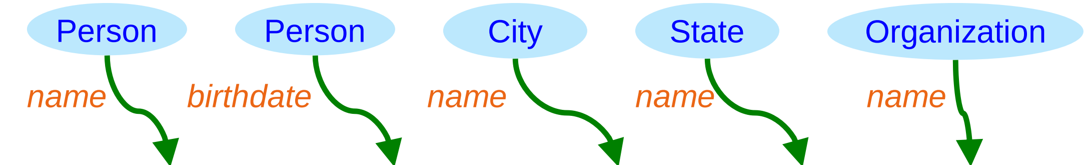
- Textual Data
  - Compute TF/IDF vectors for documents
  - Compare documents using Cosine Similarity between TF/IDF vectors

- Numeric Data
  - Use Statistical Hypothesis Testing
  - Intuition: distribution of values in different semantic types is different, e.g., temperature vs. population

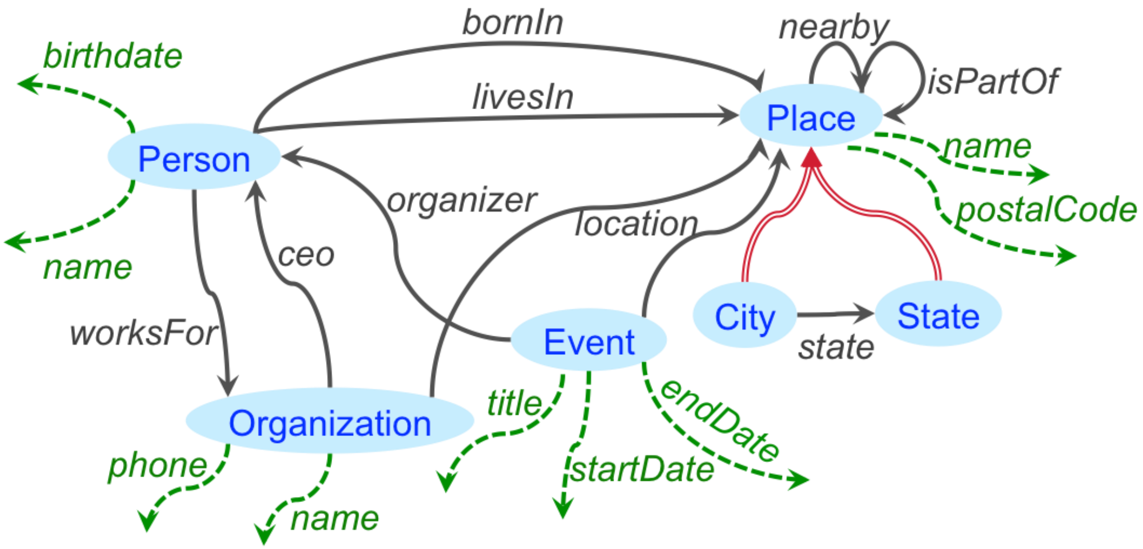
- Return Top-k suggestions based on the confidence scores

Construct a Graph

Construct a graph from semantic types and ontology

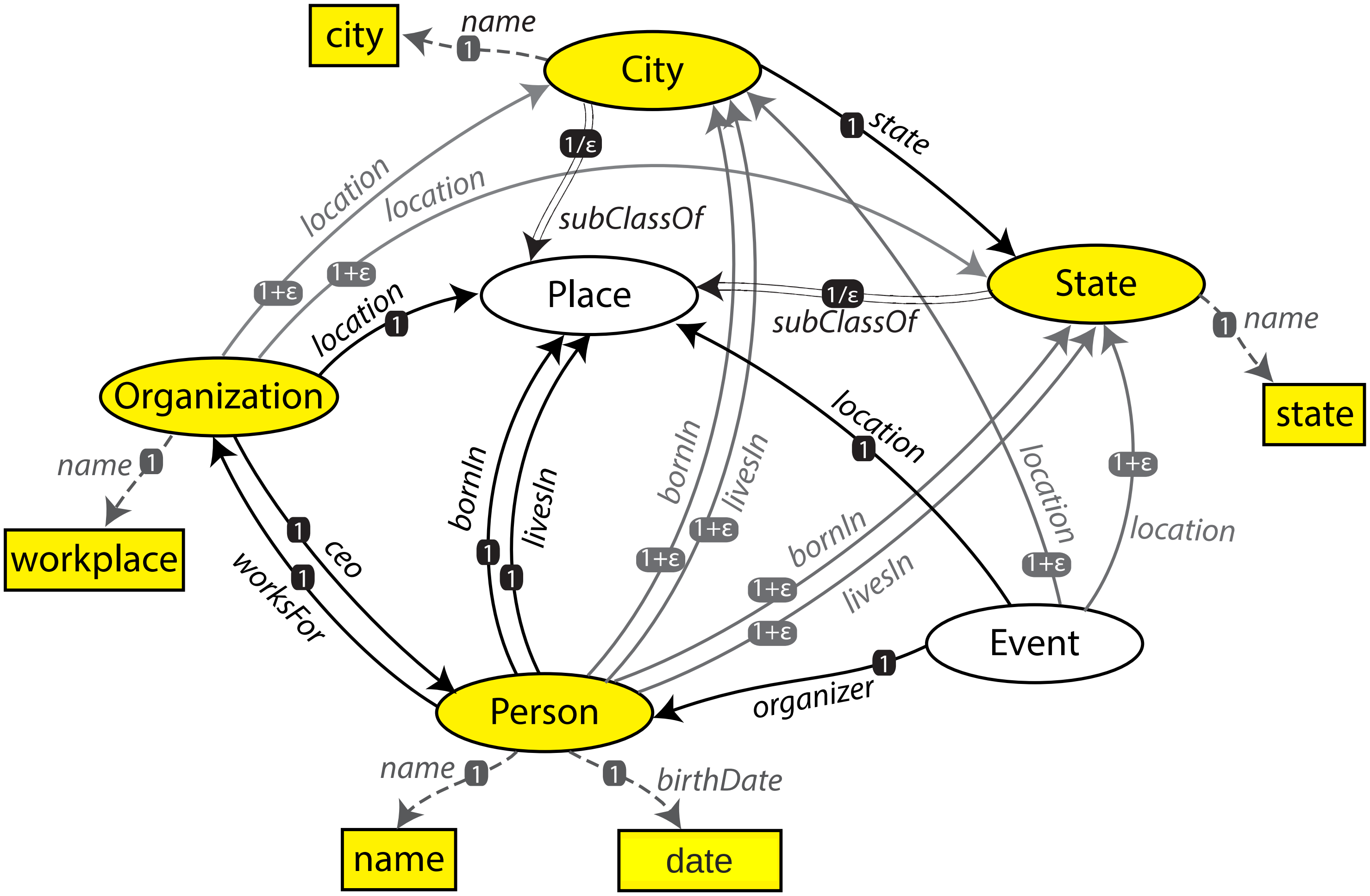


	name	date	city	state	workplace
1	Fred Collins	Oct 1959	Seattle	WA	Microsoft
2	Tina Peterson	May 1980	New York	NY	Google



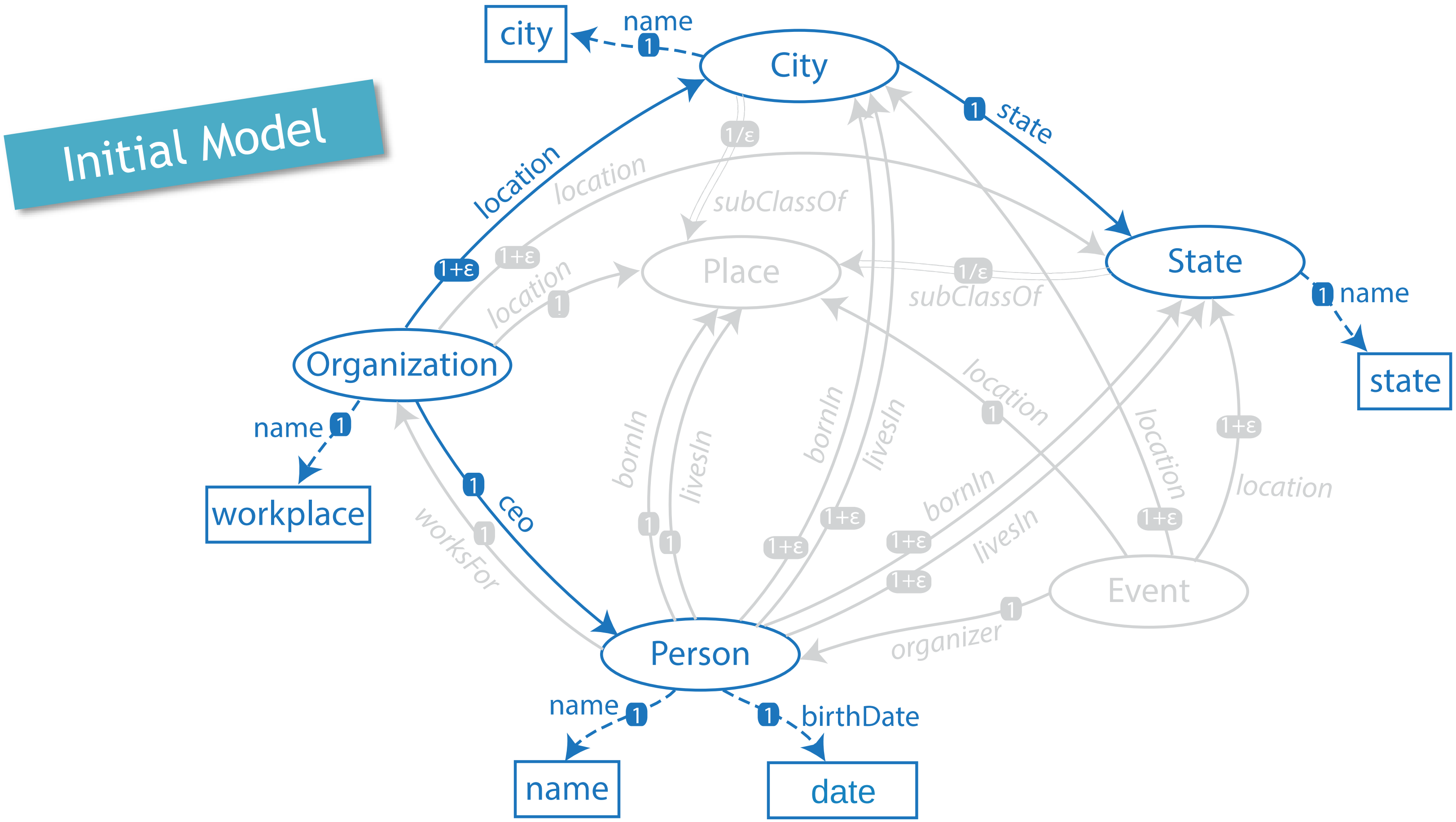
Construct a Graph

Construct a graph from semantic types and ontology

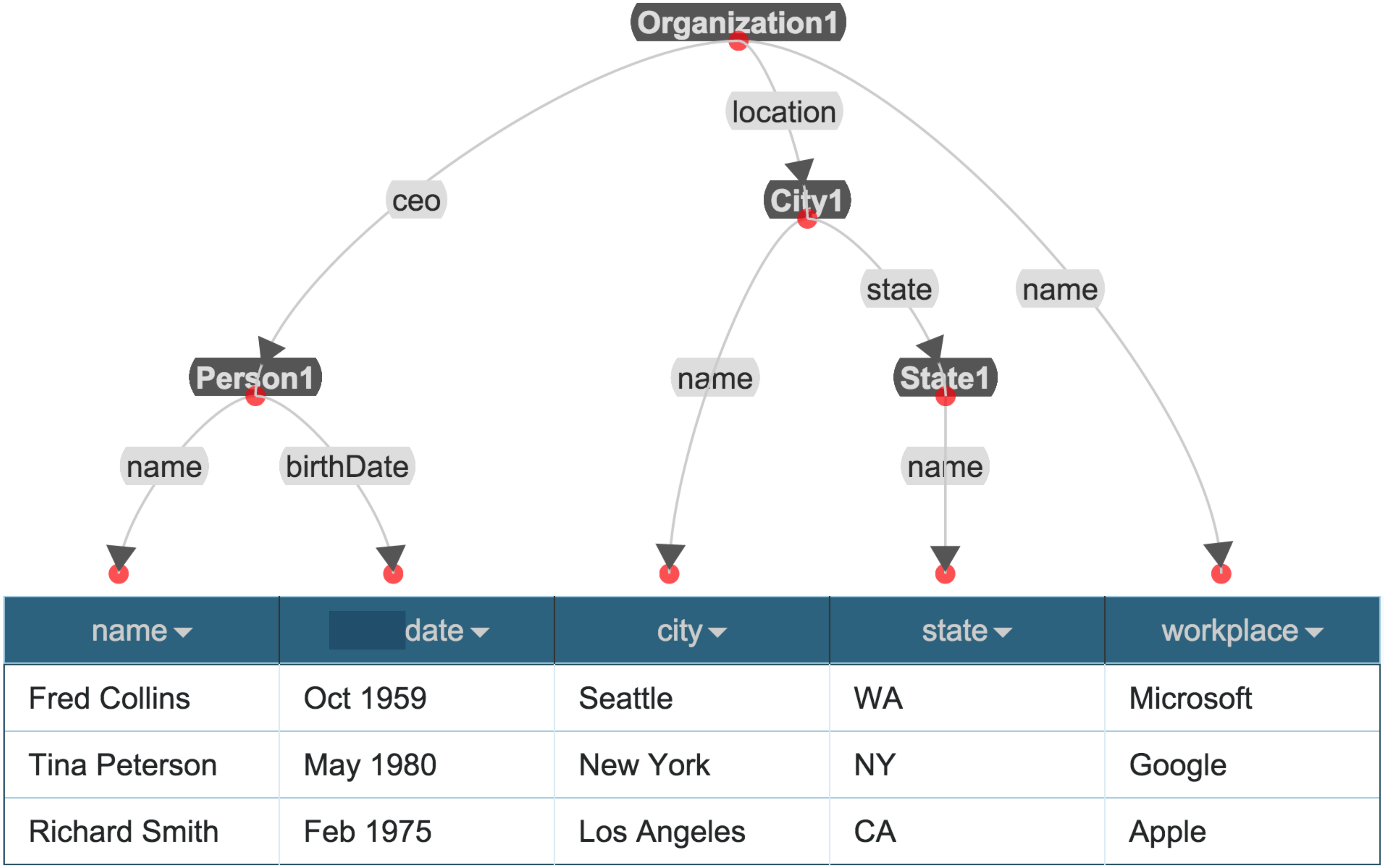


Inferring the Relationships

- Select minimal tree that connects all semantic types
  - A customized **Steiner tree algorithm**



Result in Karma

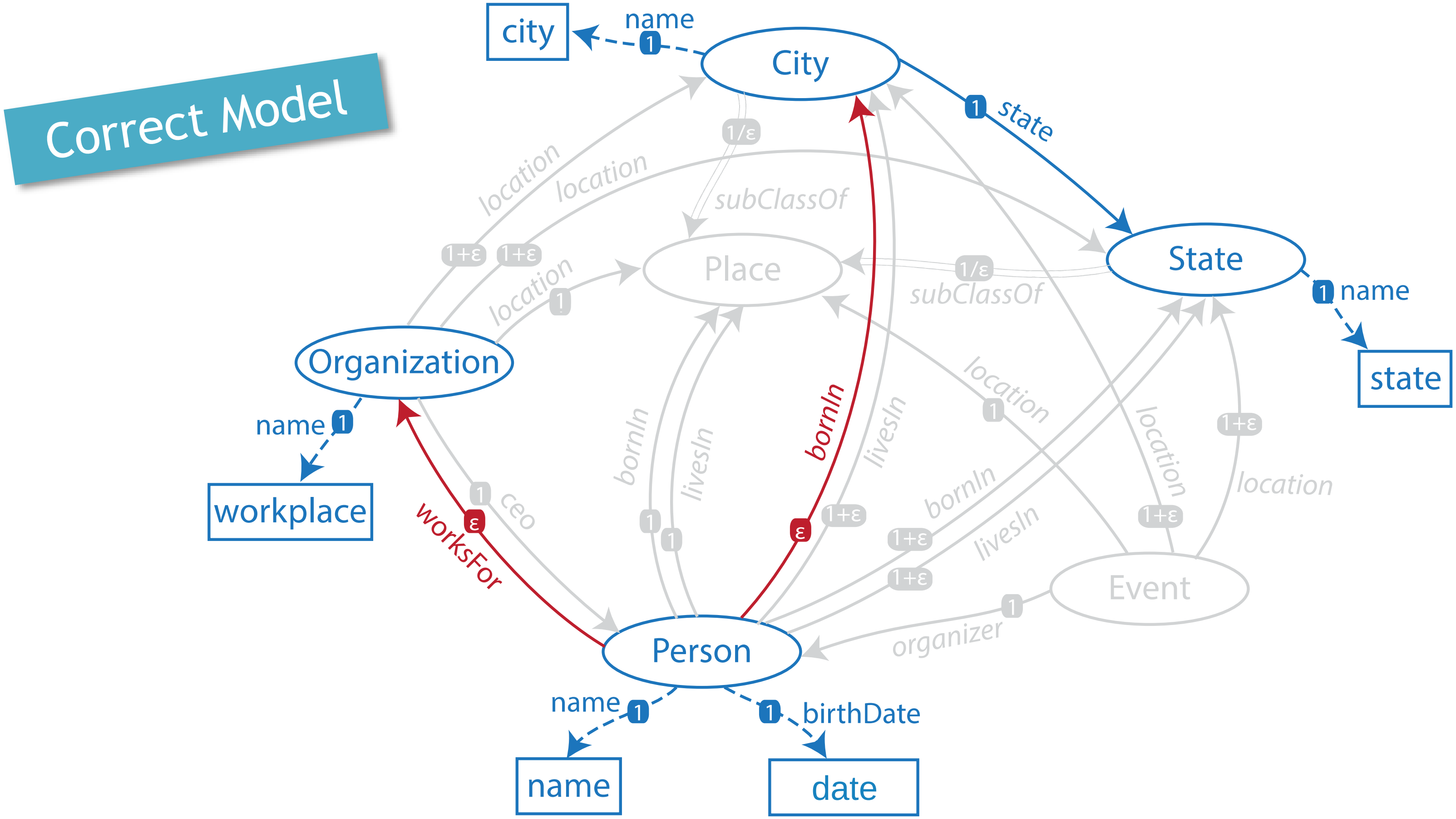


name	[redacted] date	city	state	workplace
Fred Collins	Oct 1959	Seattle	WA	Microsoft
Tina Peterson	May 1980	New York	NY	Google
Richard Smith	Feb 1975	Los Angeles	CA	Apple

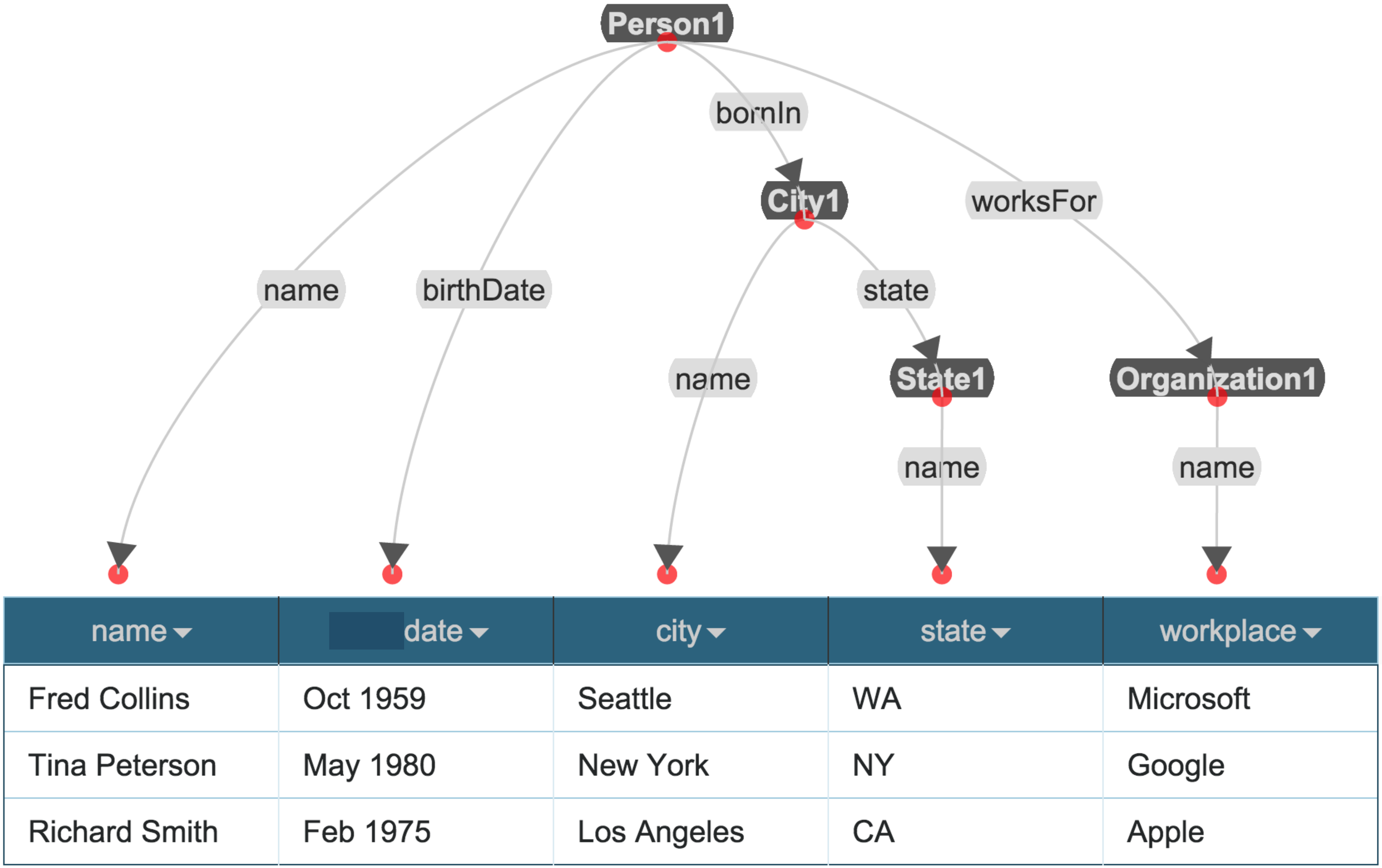
Refining the Model

Impose constraints on Steiner Tree Algorithm

- Change weight of selected links to ε
- Add source and target of selected link to Steiner nodes



Final Semantic Model



name	[redacted] date	city	state	workplace
Fred Collins	Oct 1959	Seattle	WA	Microsoft
Tina Peterson	May 1980	New York	NY	Google
Richard Smith	Feb 1975	Los Angeles	CA	Apple

Evaluation

Evaluation Dataset	EDM
# sources	29
# classes in the ontologies	119
# properties in the ontologies	351
# nodes in the gold standard models	473
# links in the gold standard models	444

- Measured the user effort in Karma to model the sources
- Started with no training data
- User actions
  - Assign/Change semantic types
  - Change relationships

Evaluation

source	columns	Choose Type	Change Link	Time (min)
s1	7	7	1	3
s2	12	5	2	6
s3	4	0	0	2
s4	17	5	6	8
s5	14	4	6	7
s6	18	4	4	7
s7	14	1	4	6
s8	6	0	4	3
…	…	…	…	…
s29	10	2	1	3
Total	**331**	**56**	**92**	**128**

Avg. min per source: 4.4 minutes

Avg. # user actions per column: 148/331=0.44

Limitation

- This approach does not learn the changes done by the user in relationships

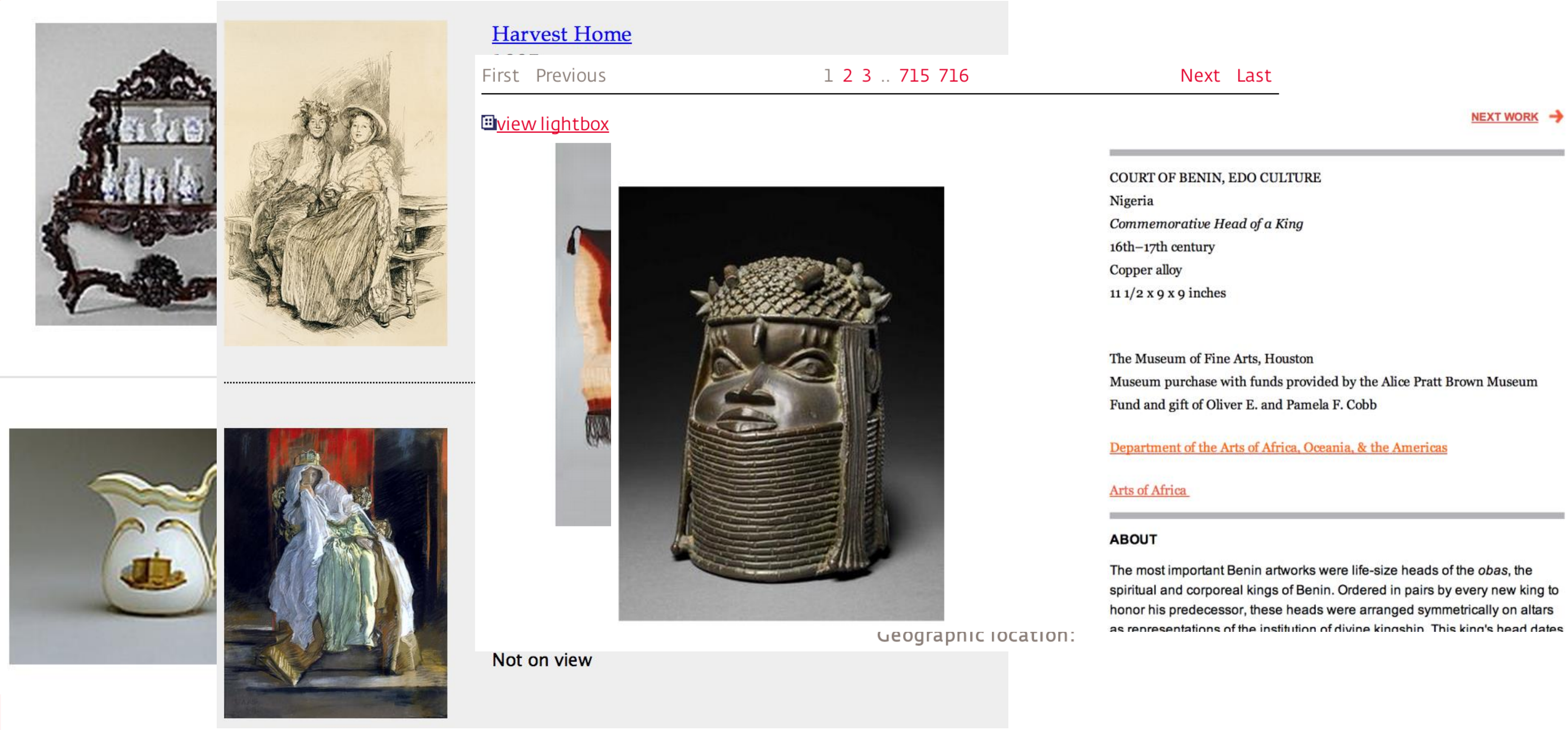
- User has to go through the refinement process each time

Learning Semantic Models

Contribution: exploiting known semantic models to learn relationships

Main Idea

Sources in the same domain often have similar data



Harvest Home

First Previous 1 2 3 .. 715 716 Next Last

view lightbox

NEXT WORK

COURT OF BENIN, EDO CULTURE
Nigeria
Commemorative Head of a King
16th–17th century
Copper alloy
11 1/2 x 9 x 9 inches

The Museum of Fine Arts, Houston
Museum purchase with funds provided by the Alice Pratt Brown Museum Fund and gift of Oliver E. and Pamela F. Cobb

Department of the Arts of Africa, Oceania, & the Americas

Arts of Africa

ABOUT

The most important Benin artworks were life-size heads of the *obas*, the spiritual and corporeal kings of Benin. Ordered in pairs by every new king to honor his predecessor, these heads were arranged symmetrically on altars as representations of the institution of divine kingship. This king's head dates

Geographic location:

Not on view

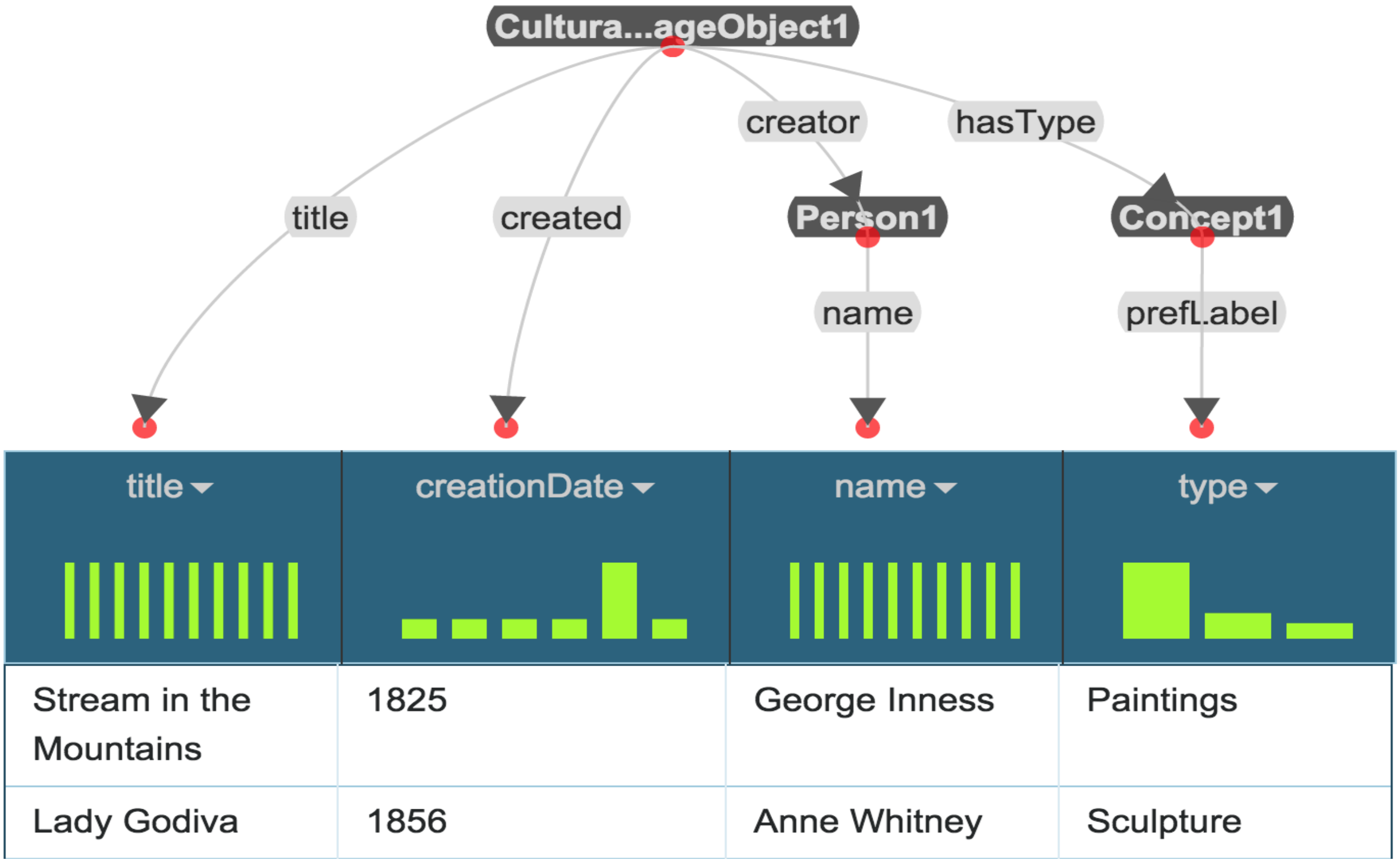
Exploit knowledge of known semantic models to hypothesize a semantic model for a new sources

Example

Domain: Museum Data

Domain ontologies: EDM SKOS FOAF AAC ORE ElementsGr2 DCTerms

Source: Dallas Museum of Art ➔ **dma(title,creationDate,name,type)**



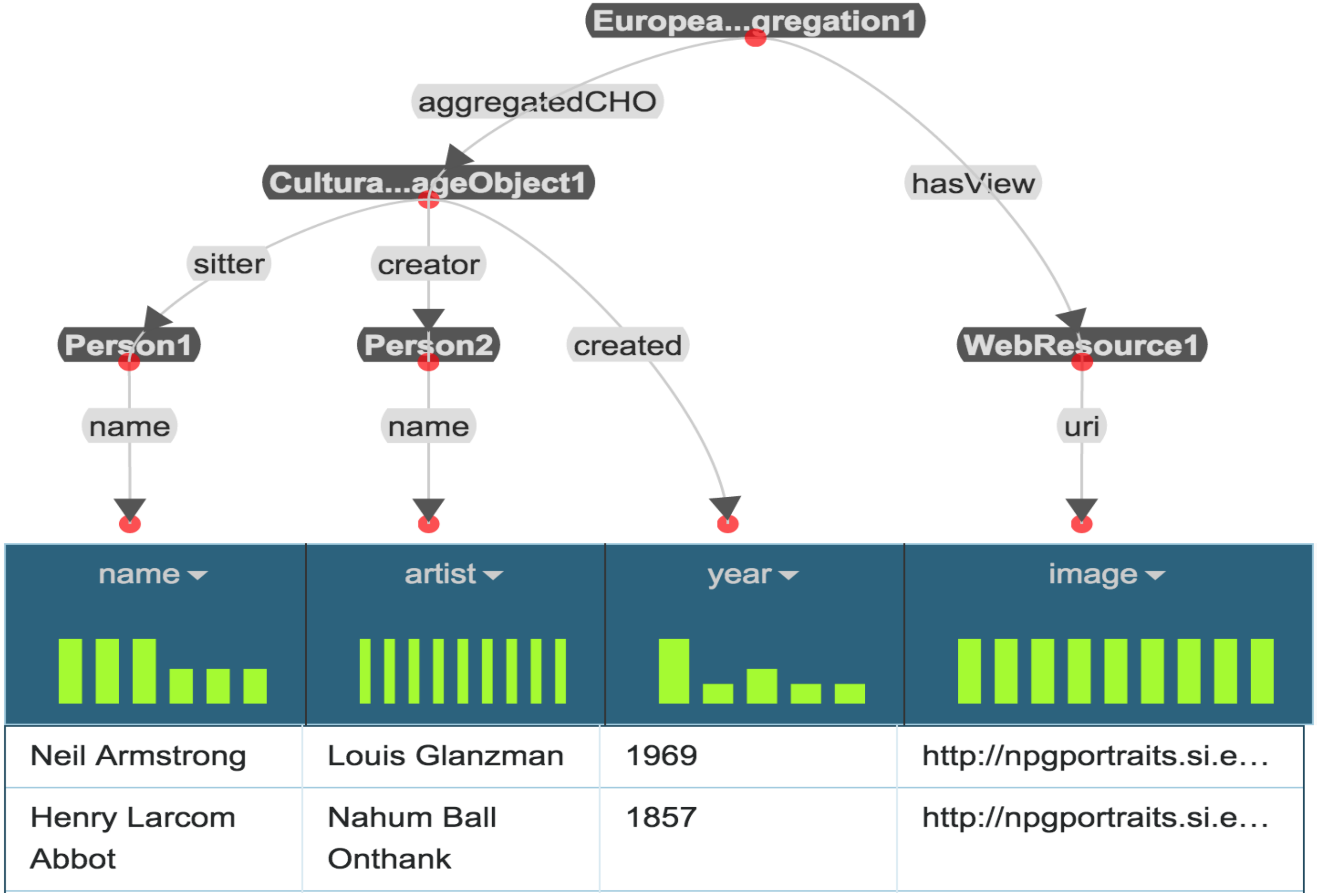
title	creationDate	name	type
Stream in the Mountains	1825	George Inness	Paintings
Lady Godiva	1856	Anne Whitney	Sculpture

Example

Domain: Museum Data

Domain ontologies: EDM SKOS FOAF AAC ORE ElementsGr2 DCTerms

Source: National Portrait Gallery ➔ **npg(name,artist,year,image)**



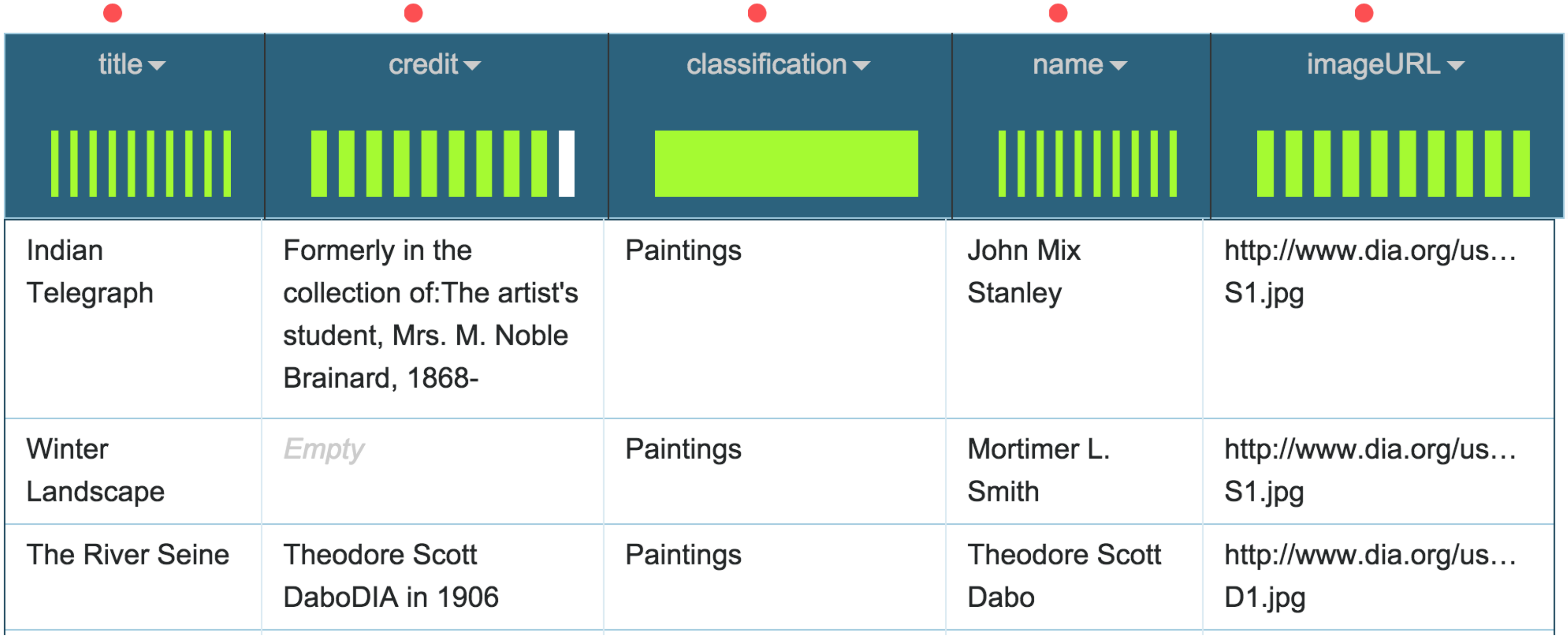
name	artist	year	image
Neil Armstrong	Louis Glanzman	1969	http://npgportraits.si.e…
Henry Larcom Abbot	Nahum Ball Onthank	1857	http://npgportraits.si.e…

Example

Domain: Museum Data

Domain ontologies: EDM SKOS FOAF AAC ORE ElementsGr2 DCTerms

Source: Detroit Institute of Art ➔ **dia(title,credit,classification,name,imageURL)**

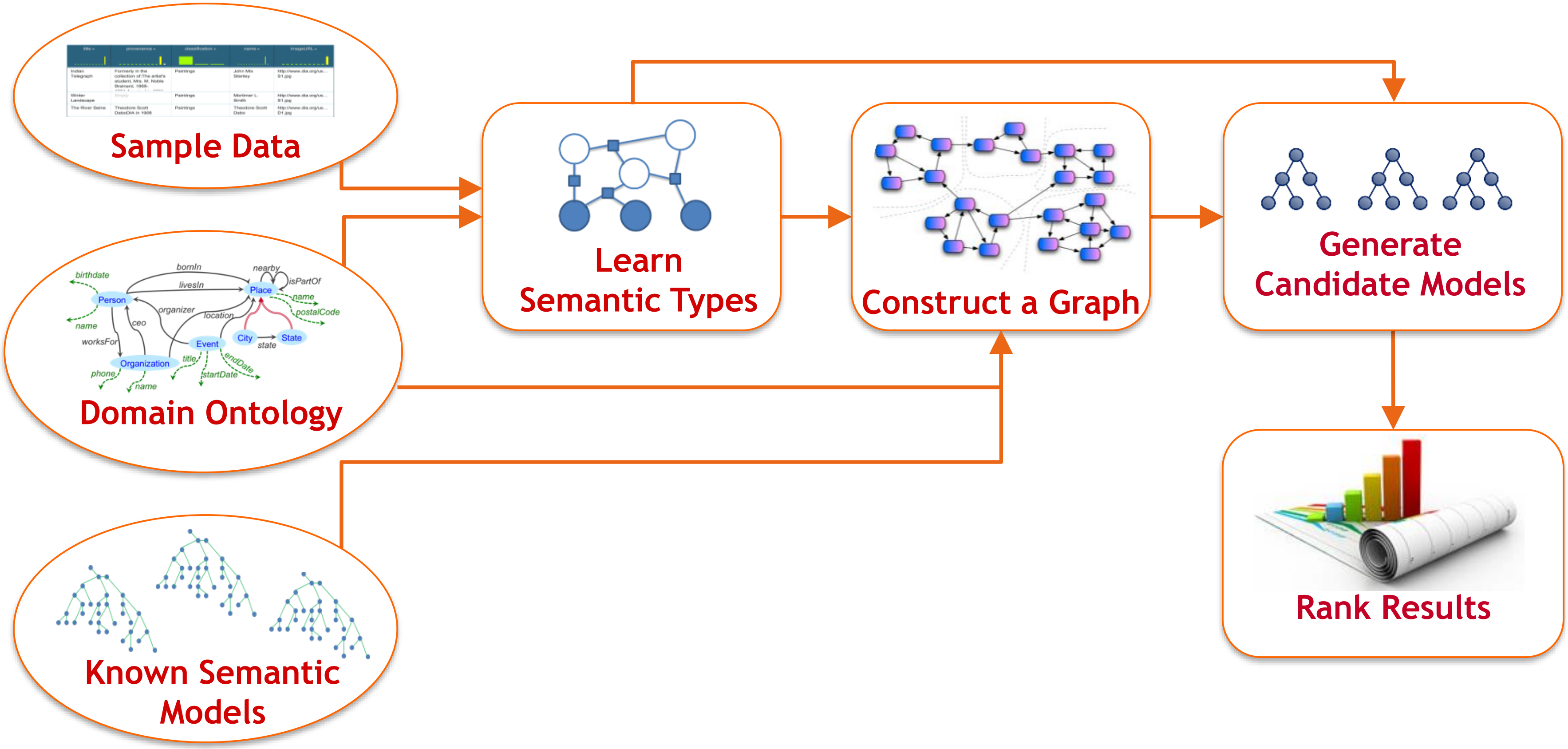


title	credit	classification	name	imageURL
Indian Telegraph	Formerly in the collection of:The artist's student, Mrs. M. Noble Brainard, 1868-	Paintings	John Mix Stanley	http://www.dia.org/us… S1.jpg
Winter Landscape	*Empty*	Paintings	Mortimer L. Smith	http://www.dia.org/us… S1.jpg
The River Seine	Theodore Scott DaboDIA in 1906	Paintings	Theodore Scott Dabo	http://www.dia.org/us… D1.jpg

Goal: Automatically suggest a semantic model for ***dia***

Approach

[Taheriyan et al, ISWC 2013, ICSC 2014, JWS 2015]



Sample Data

Domain Ontology

Known Semantic Models

Learn Semantic Types

Construct a Graph

Generate Candidate Models

Rank Results

Implemented in **Karma**

Approach

Input

- Sample data from new source (S)
- Domain Ontologies (O)
- Known semantic models

1 Learn semantic types for attributes(s)

Construct Graph G=(V,E)

Generate mappings between attributes(S) and V

Generate and rank semantic models

Output

- A ranked set of semantic models for S

Learn Semantic Types

- Learn *Semantic Types* for each attribute from its data
- Pick top K semantic types according to their confidence values

dia(title,credit, classification,name,imageURL)		
title	<aac:CulturalHeritageObject, dcterms:title>	0.49
	<aac:CulturalHeritageObject, rdfs:label>	0.28
credit	<aac:CulturalHeritageObject, dcterms:provenance>	0.83
	<aac:Person, ElementsGr2:note>	0.06
classification	<skos:Concept, skos:prefLabel>	0.58
	<skos:Concept, rdfs:label>	0.41
name	<aac:Person, foaf:name>	0.65
	<fofa:Person, fofaf:name>	0.32
imageURL	<foaf:Document, uri>	0.47
	<edm:WebResource, uri>	0.40

Approach

Input

- Sample data from new source (S)
- Domain Ontologies (O)
- Known semantic models

✓ Learn semantic types for attributes(s)

❷ Construct Graph G=(V,E)

Generate mappings between attributes(S) and V

Generate and rank semantic models

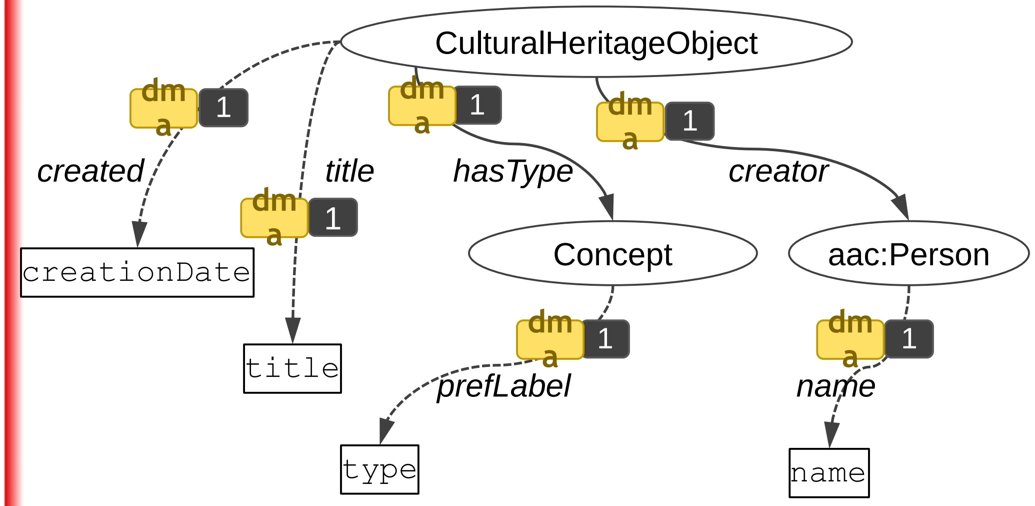
Output

- A ranked set of semantic models for S

Build Graph G: Add Known Models

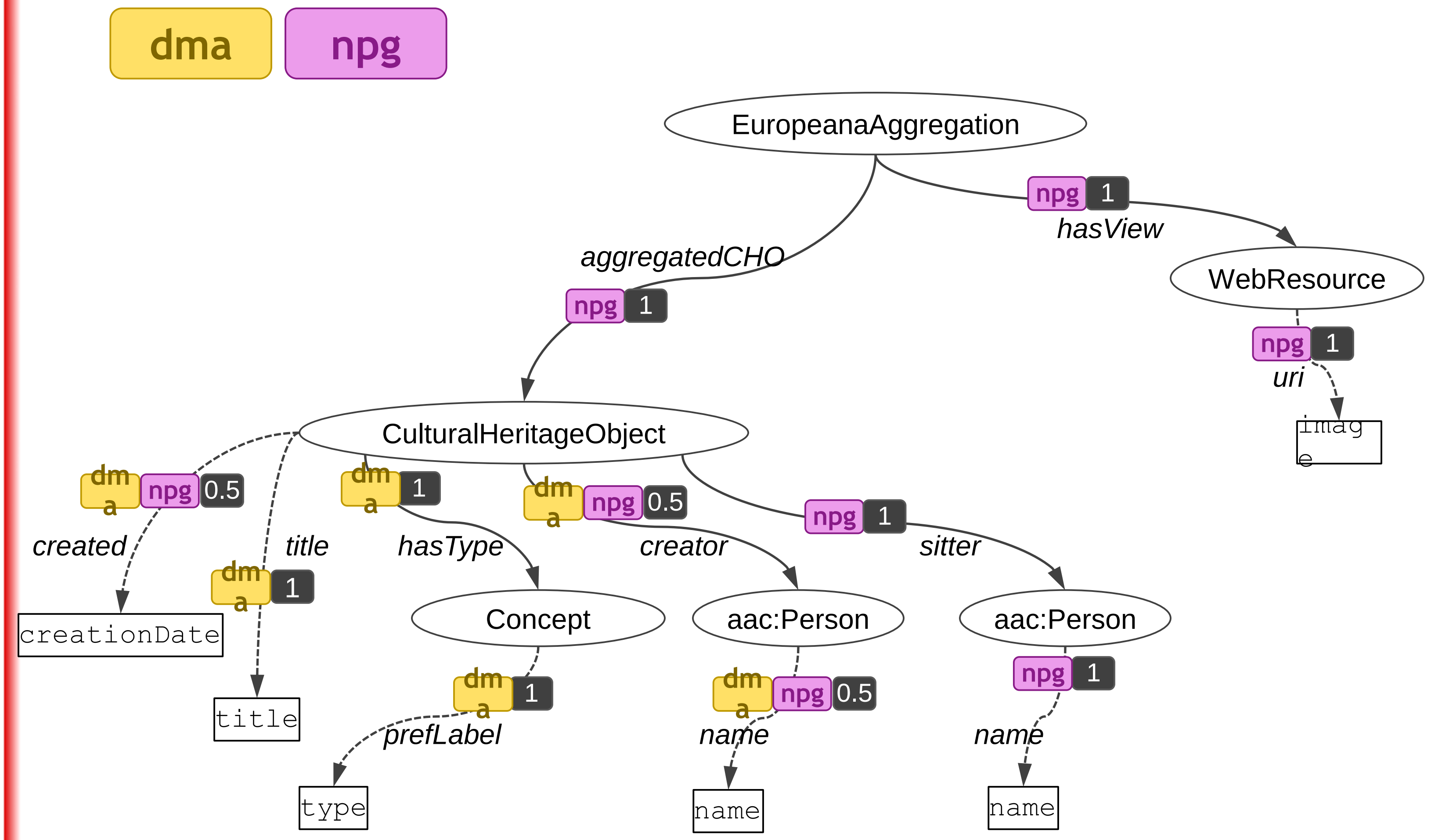
- Annotate (tag) nodes and links with list of supporting models
- Adjust weight based on the number of supporting models

dma



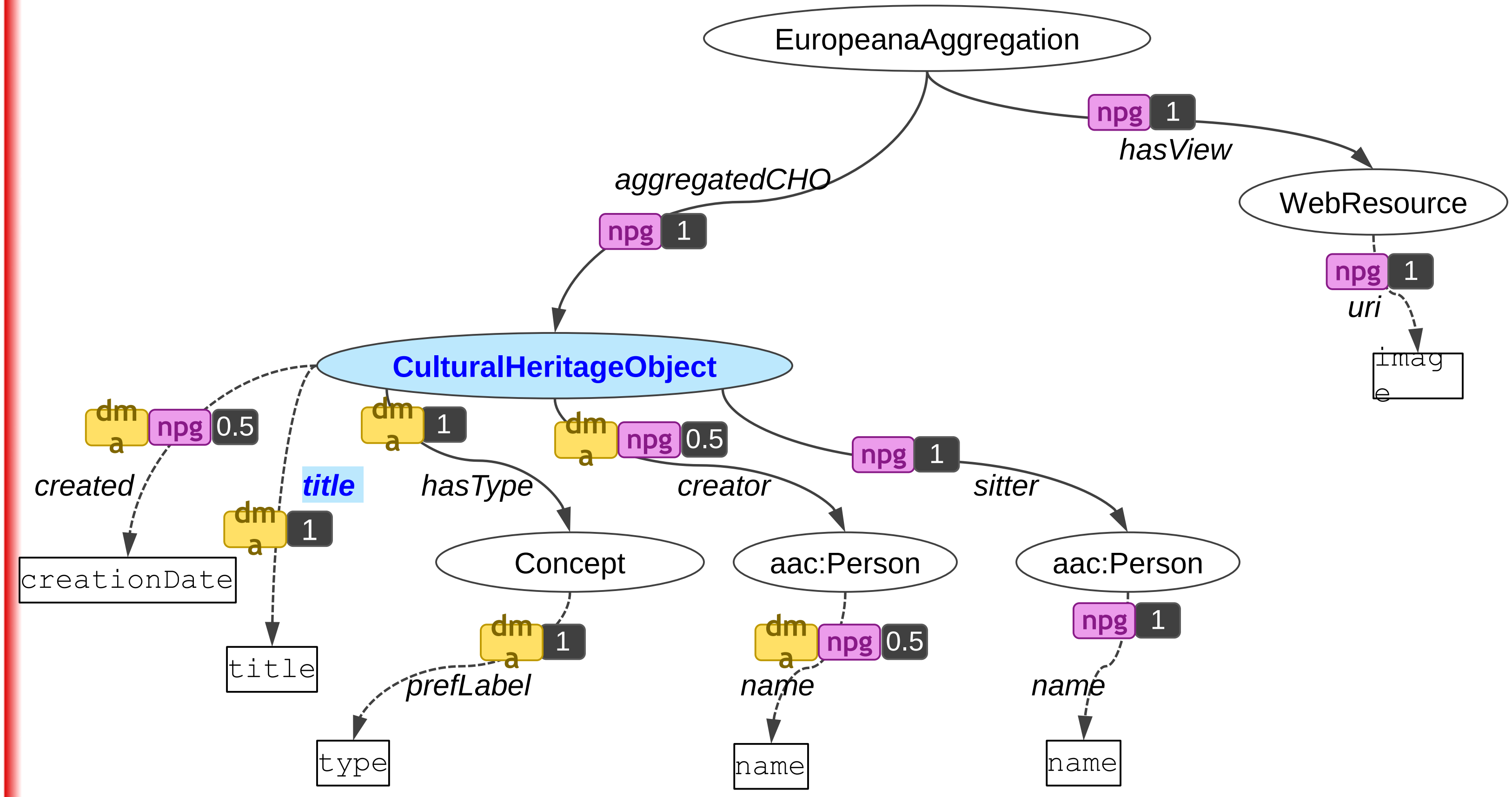
Build Graph G: Add Known Models

- Annotate (tag) nodes and links with list of supporting models
- Adjust weight based on the number of tags



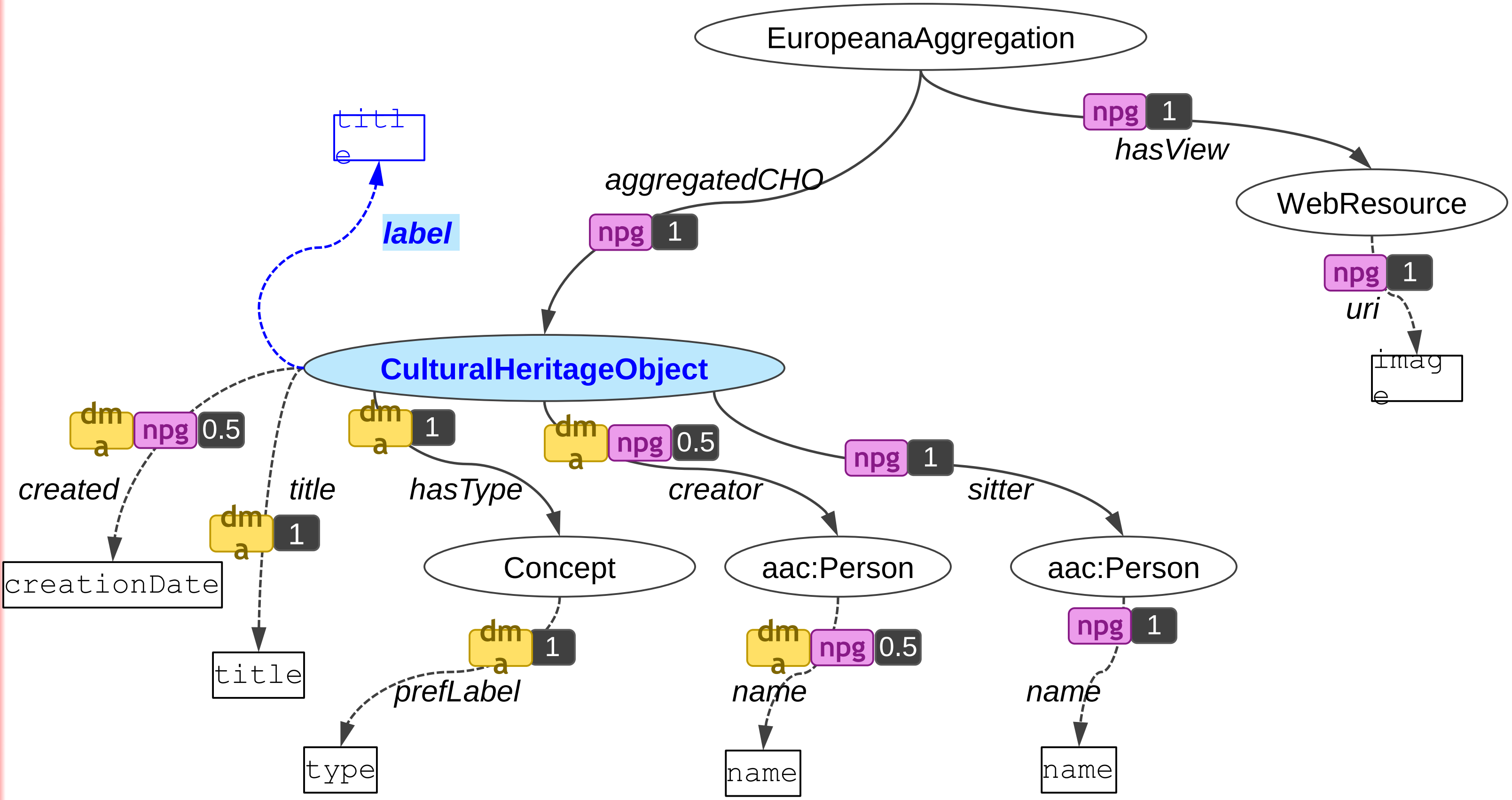
Build Graph G: Add Semantic Types

title	**<CulturalHeritageObject,title>** <CulturalHeritageObject,label>
credit	<CulturalHeritageObject,provenance> <Person,note>
classification	<Concept,prefLabel> <Concept,label>
name	<aac:Person,name> <foaf:Person,name>
imageURL	<Document,uri> <WebResource,uri>



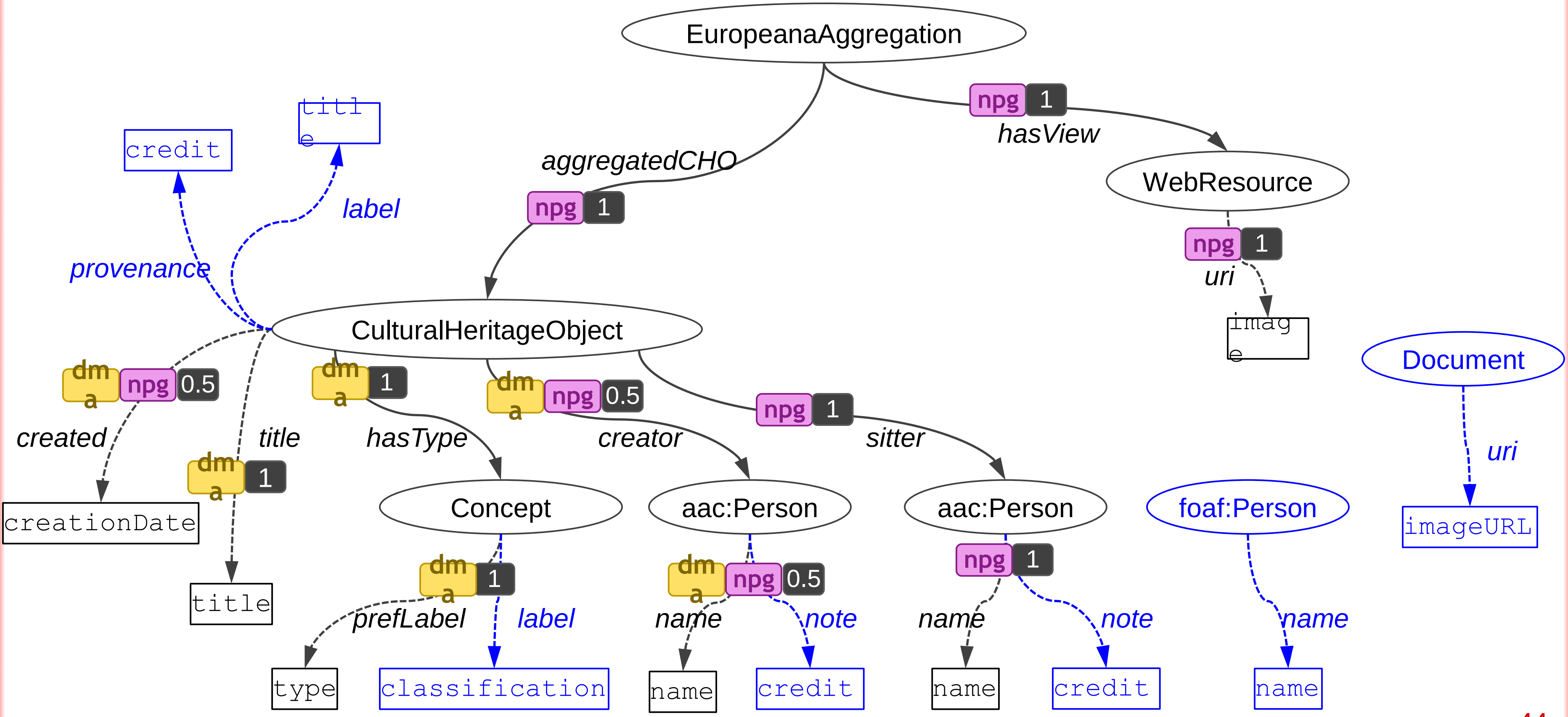
Build Graph G: Add Semantic Types

title	<CulturalHeritageObject,title>	<CulturalHeritageObject,label>
credit	<CulturalHeritageObject,provenance>	<Person,note>
classification	<Concept,prefLabel>	<Concept,label>
name	<aac:Person,name>	<foaf:Person,name>
imageURL	<Document,uri>	<WebResource,uri>



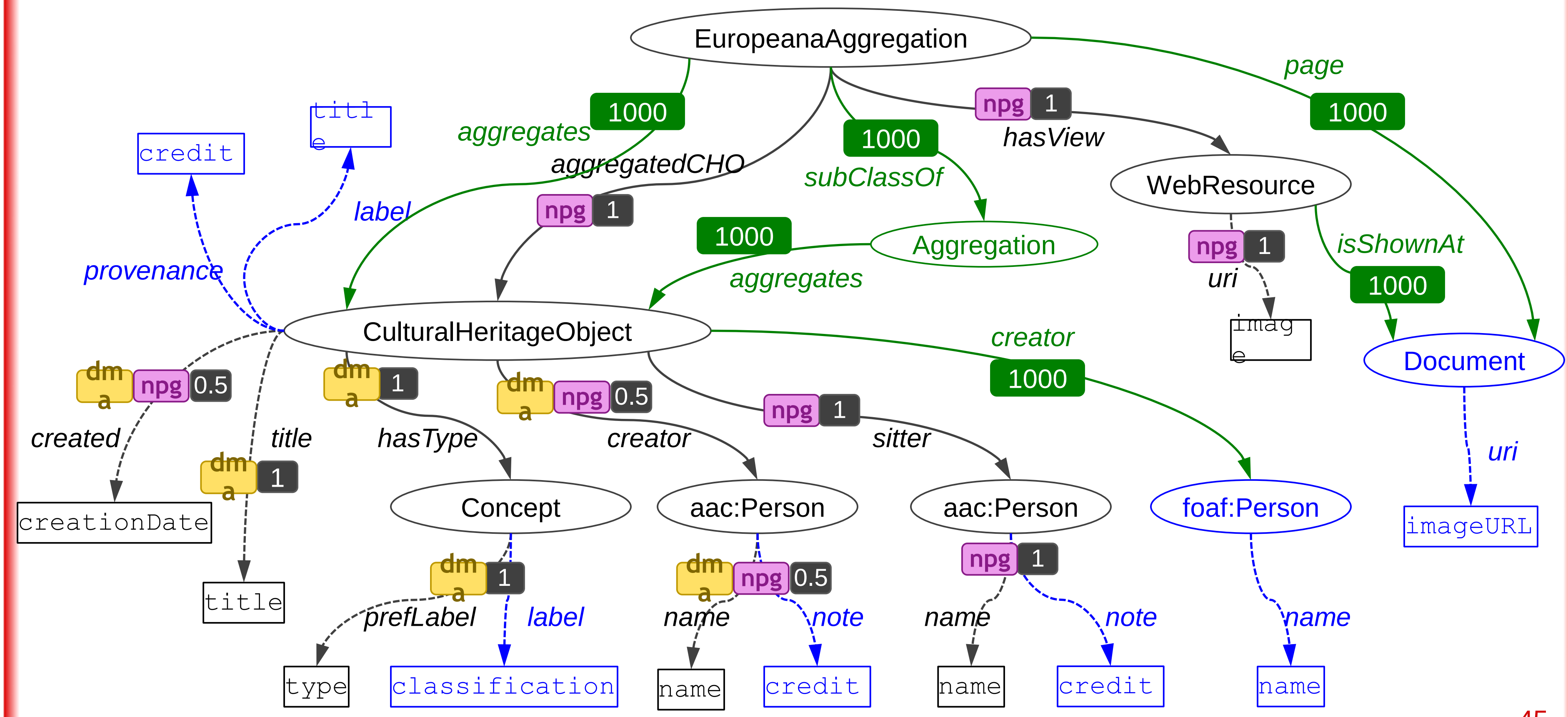
Build Graph G: Add Semantic Types

title	<CulturalHeritageObject,title> <CulturalHeritageObject,label>
credit	<CulturalHeritageObject,provenance> <Person,note>
classification	<Concept,prefLabel> <Concept,label>
name	<aac:Person,name> <foaf:Person,name>
imageURL	<Document,uri> <WebResource,uri>



Build Graph G: Expand with Paths from Ontology

- Assign a high weight to the links coming from the ontology



Approach

Input

- Sample data from new source (S)
- Domain Ontologies (O)
- Known semantic models

✓ Learn semantic types for attributes(s)

✓ Construct Graph G=(V,E)

3 Generate mappings between attributes(S) and V

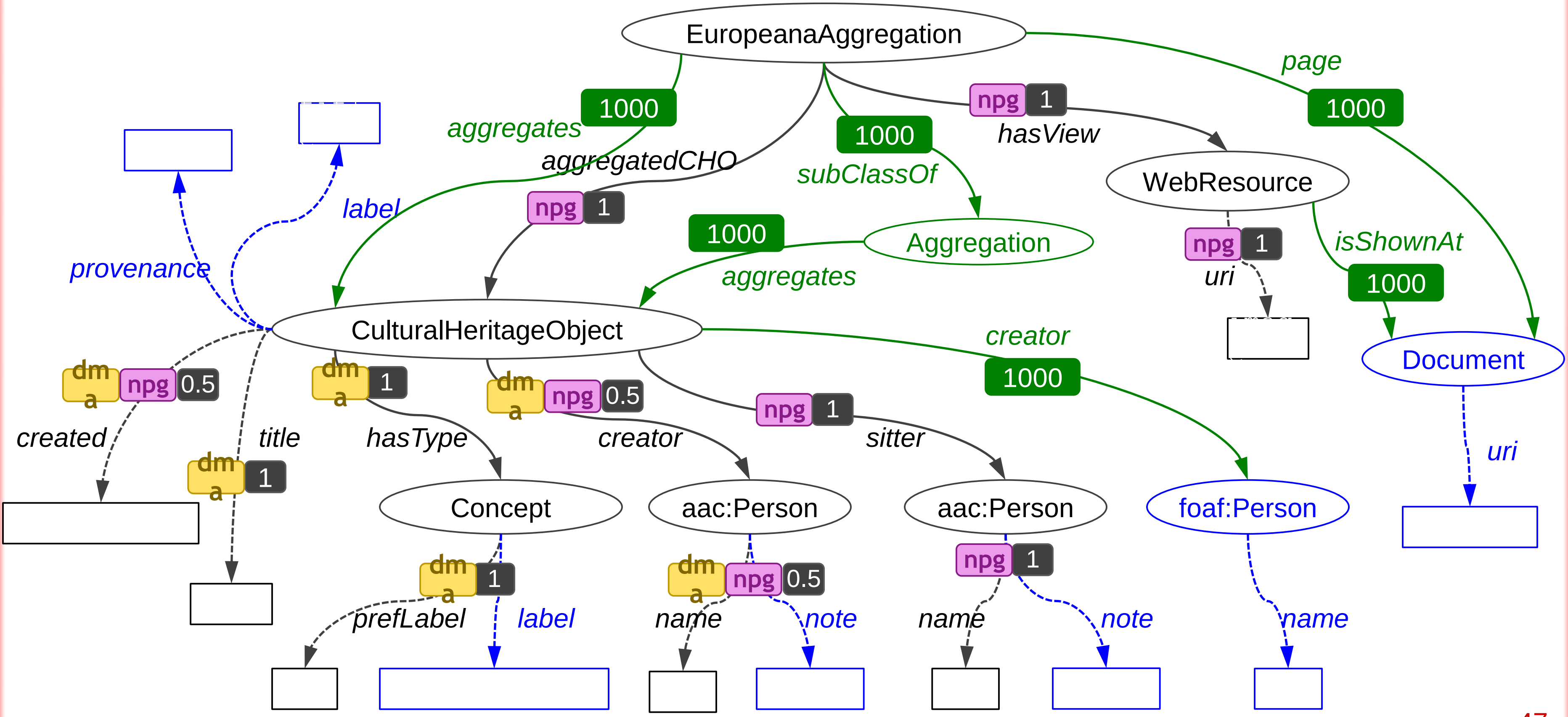
Generate and rank semantic models

Output

- A ranked set of semantic models for S

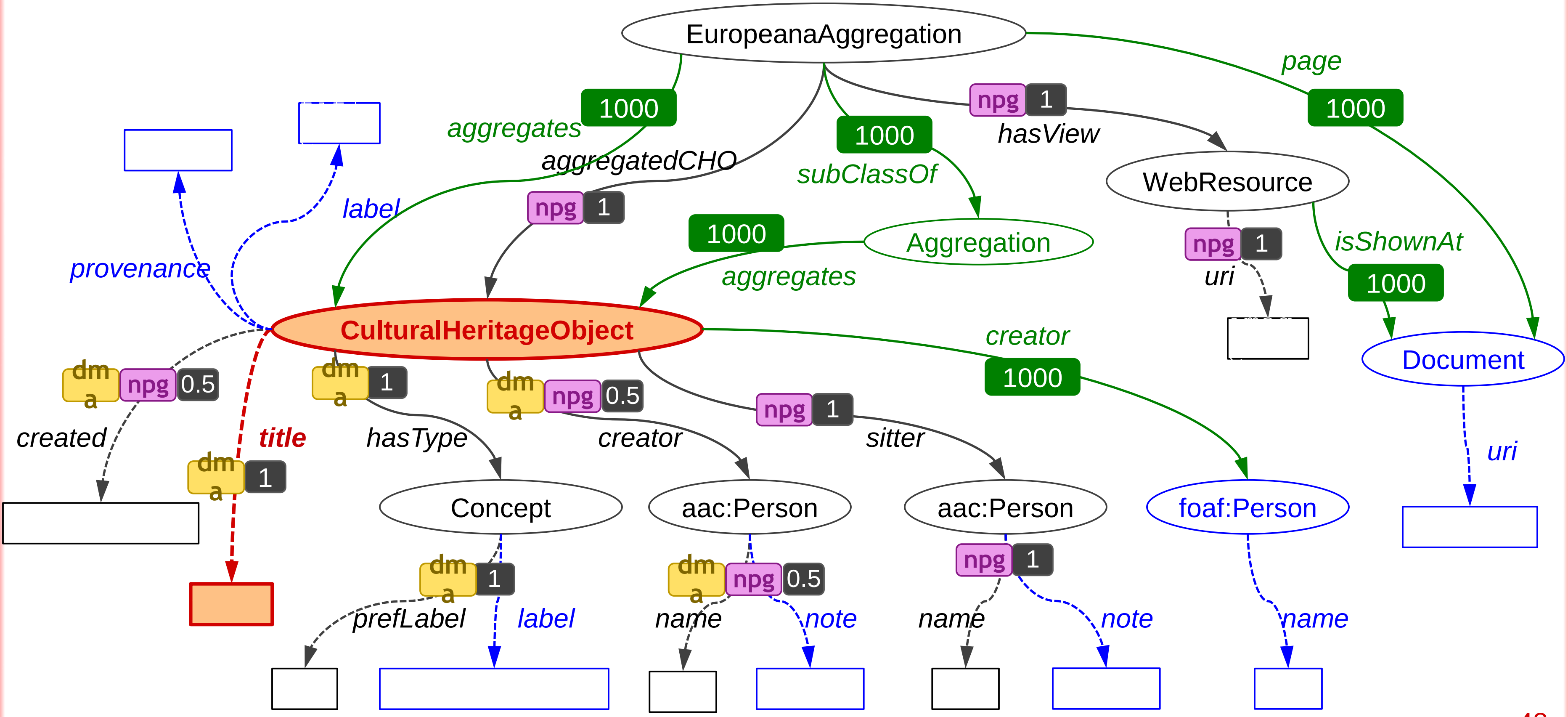
Map Source Attributes to the Graph

title	<CulturalHeritageObject,title> <CulturalHeritageObject,label>
credit	<CulturalHeritageObject,provenance> <Person,note>
classification	<Concept,prefLabel> <Concept,label>
name	<aac:Person,name> <foaf:Person,name>
imageURL	<Document,uri> <WebResource,uri>



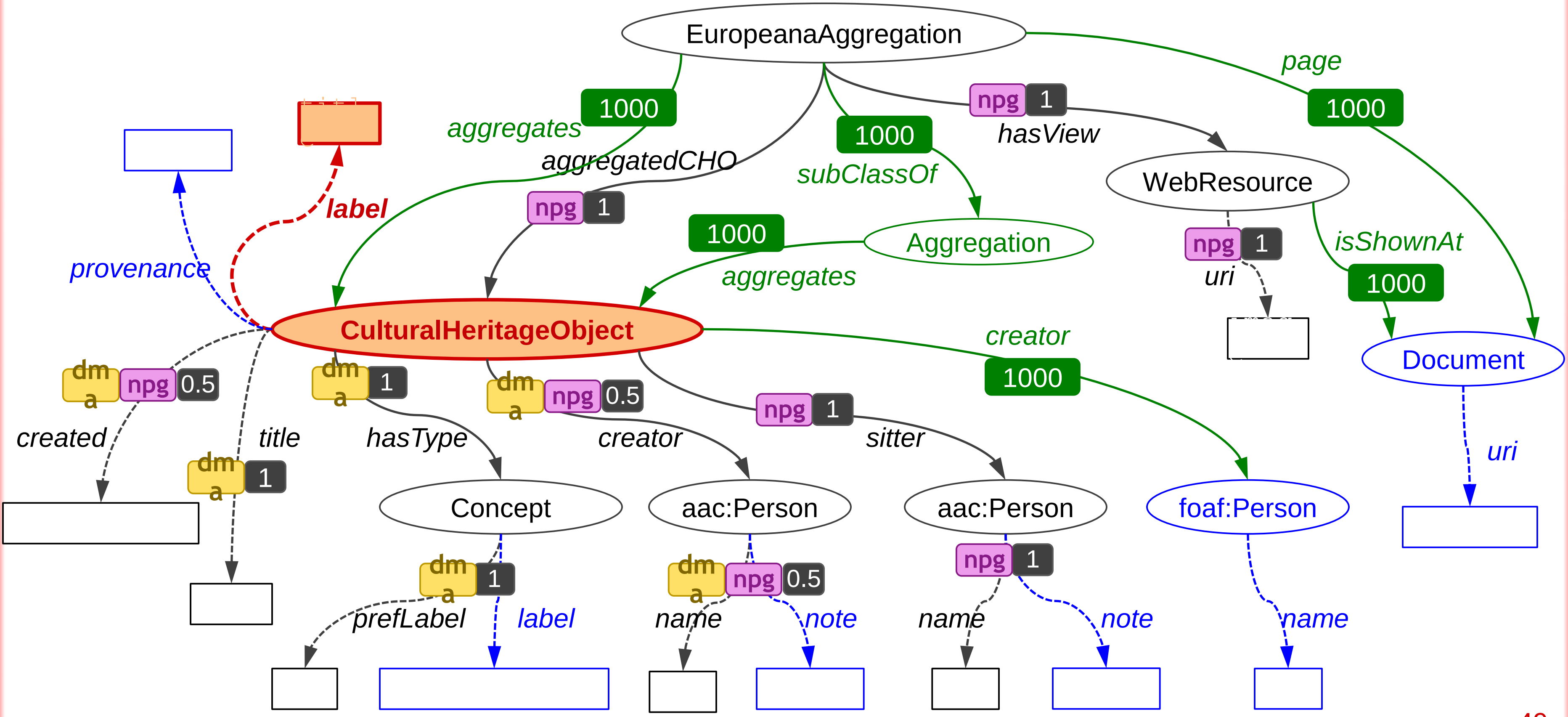
Map Source Attributes to the Graph

title	<CulturalHeritageObject,title> <CulturalHeritageObject,label>
credit	<CulturalHeritageObject,provenance> <Person,note>
classification	<Concept,prefLabel> <Concept,label>
name	<aac:Person,name> <foaf:Person,name>
imageURL	<Document,uri> <WebResource,uri>



Map Source Attributes to the Graph

title	<CulturalHeritageObject,title> **<CulturalHeritageObject,label>**
credit	<CulturalHeritageObject,provenance> <Person,note>
classification	<Concept,prefLabel> <Concept,label>
name	<aac:Person,name> <foaf:Person,name>
imageURL	<Document,uri> <WebResource,uri>



Approach

Input

- Sample data from new source (S)
- Domain Ontologies (O)
- Known semantic models

✓ Learn semantic types for attributes(s)

✓ Construct Graph G=(V,E)

✓ Generate mappings between attributes(S) and V

4 Generate and rank semantic models

Output

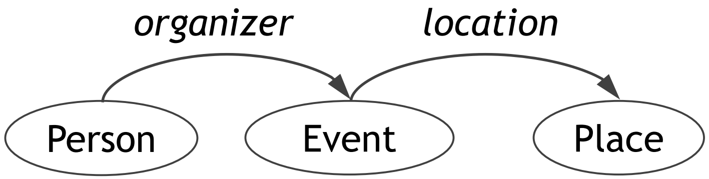
- A ranked set of semantic models for S

Generate Semantic Models

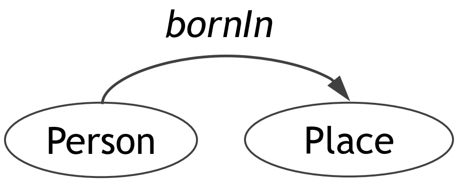
- Compute Steiner tree for each mapping
  - A minimal tree connecting nodes of mapping
  - A customization of BANKS algorithm [Bhalotia et al., 2002]
- Our algorithm considers both coherence and popularity
- Each tree is a candidate model
- Rank the models based on coherence and cost

Why Coherence is Important?

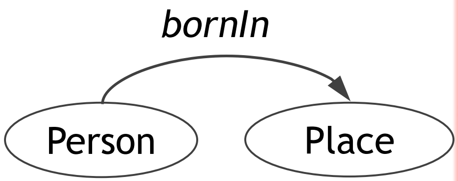
Known Models



s1



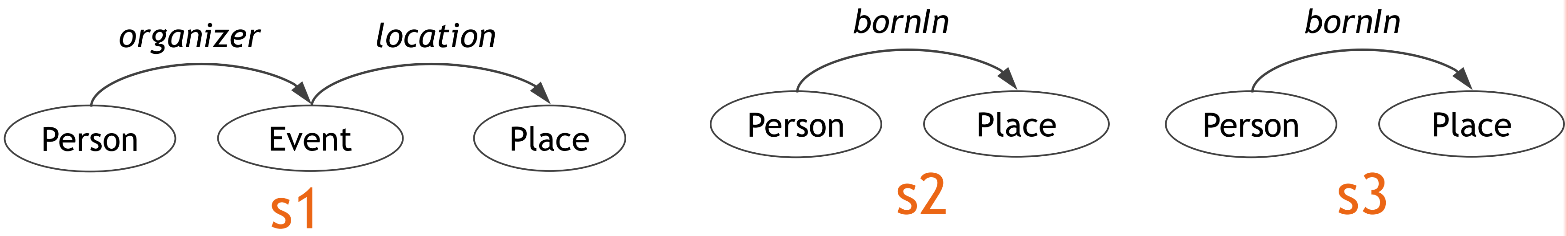
s2



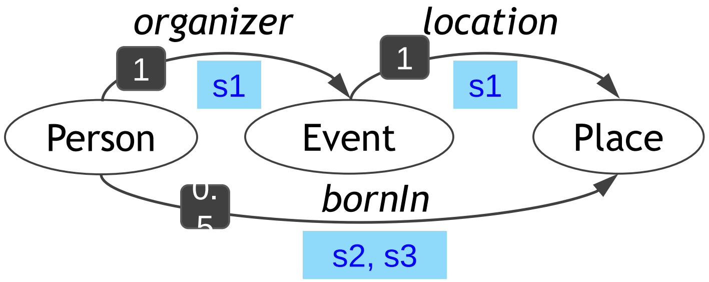
s3

Why Coherence is Important?

Known Models

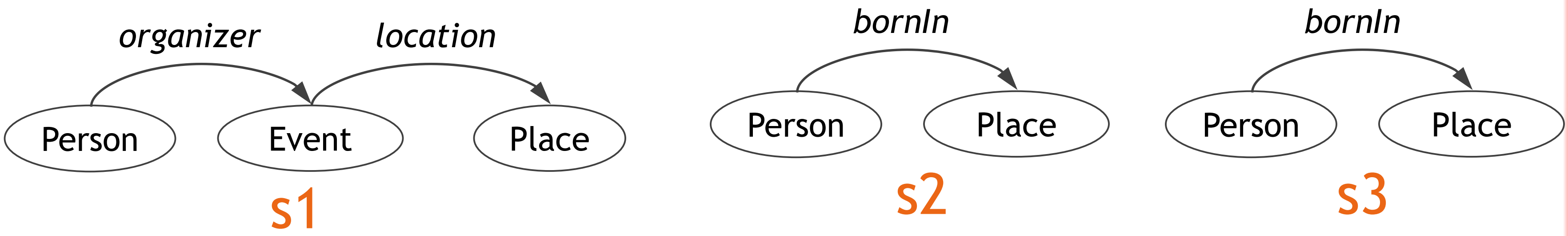


Graph

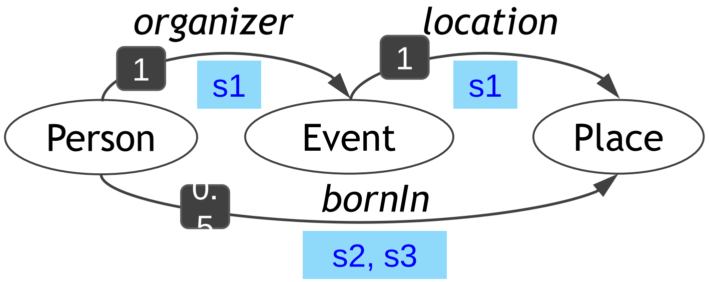


Why Coherence is Important?

Known Models



Graph

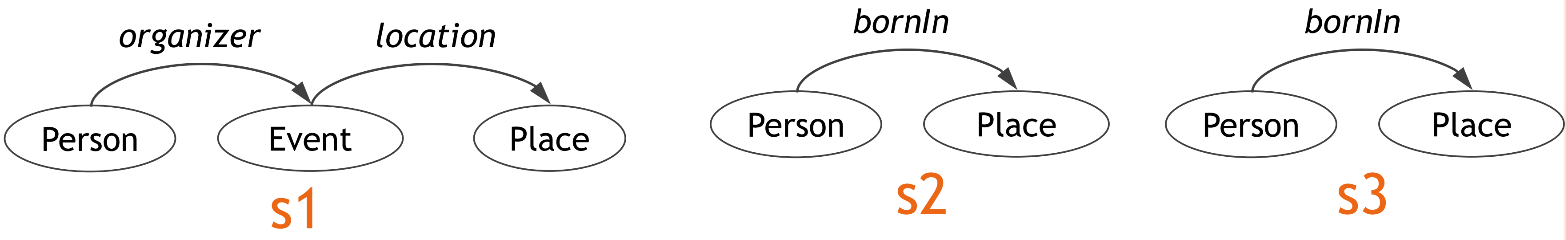


Semantic types of a new source

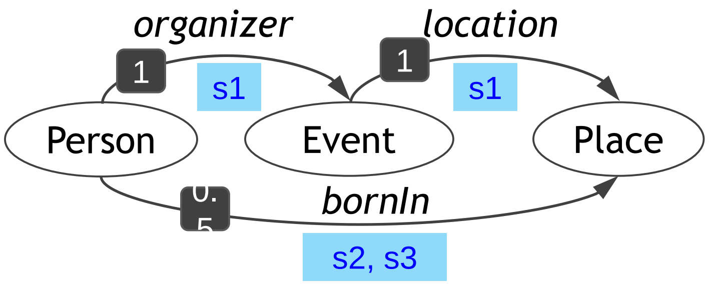
Person
Event
Place

Why Coherence is Important?

Known Models



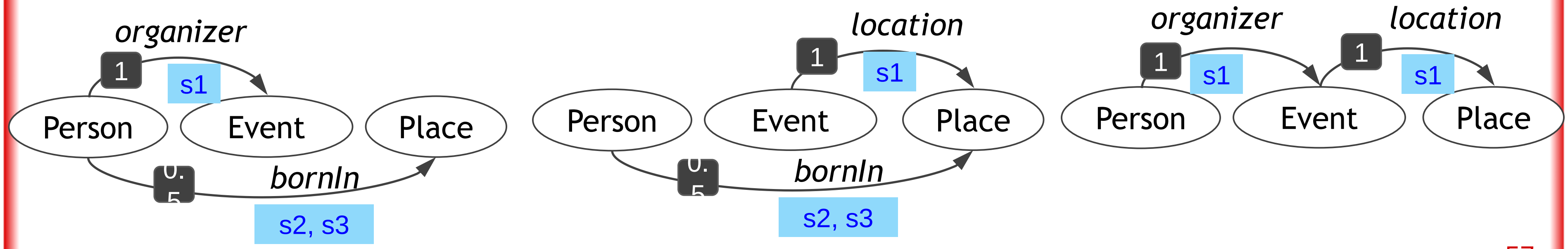
Graph



Semantic types of a new source

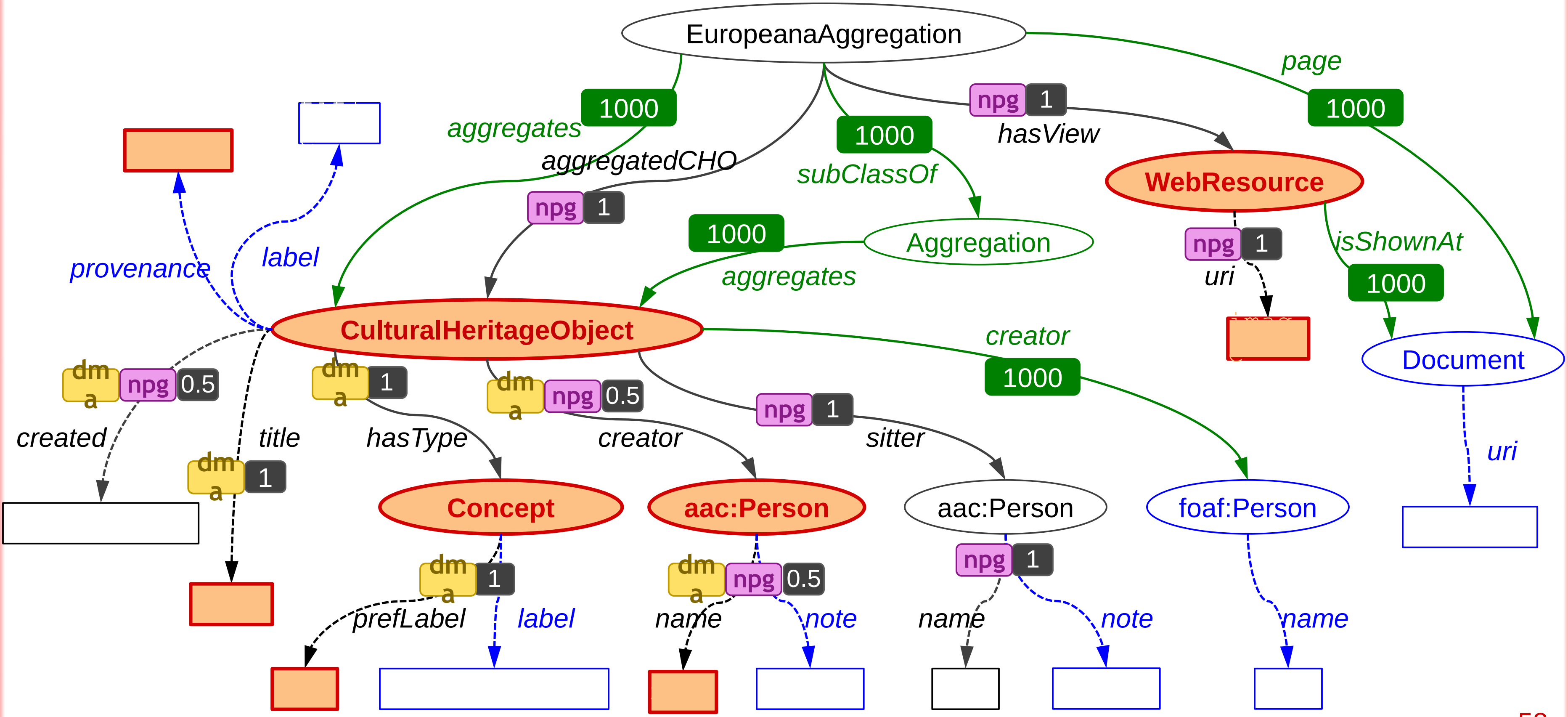
Person
Event
Place

Top 3 Steiner trees



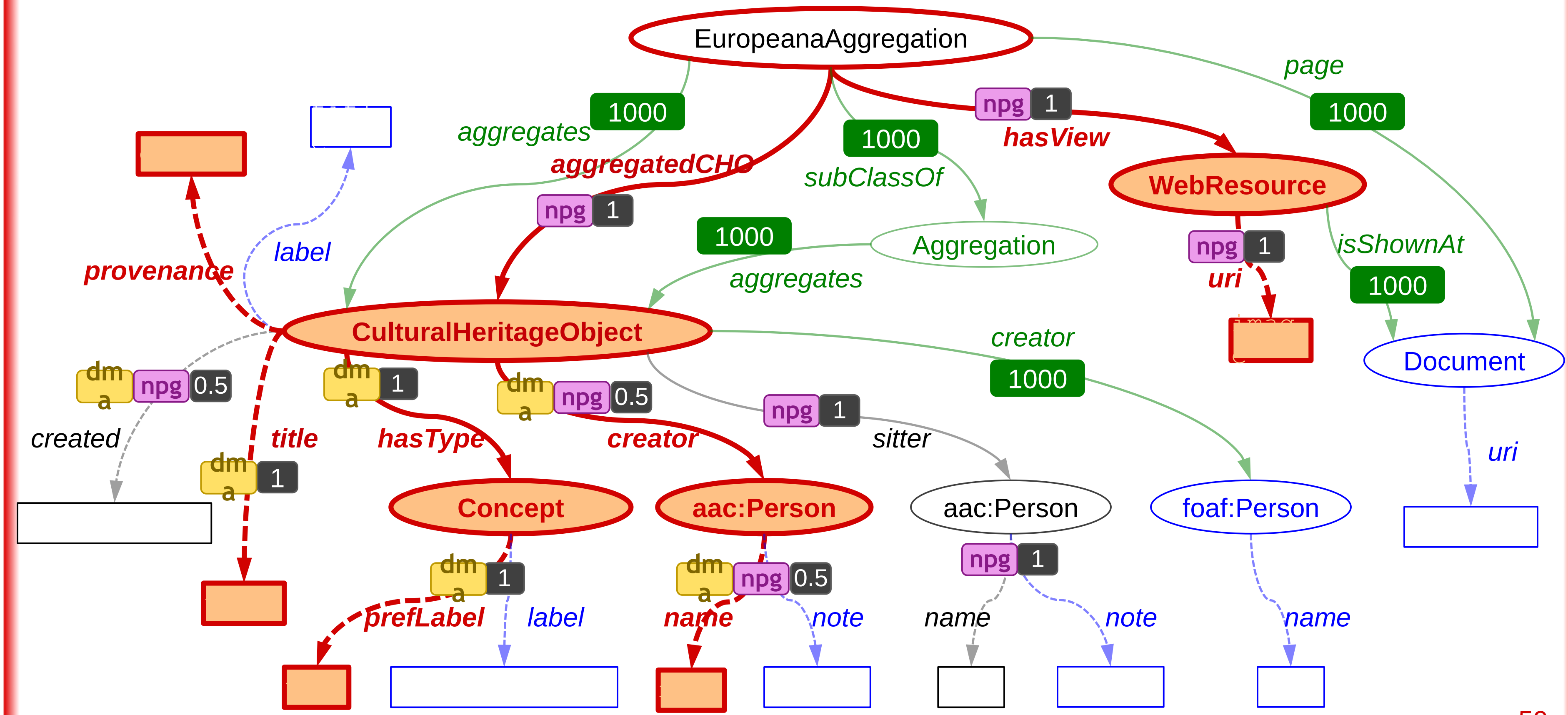
Example Mapping

title	**<CulturalHeritageObject,title>** <CulturalHeritageObject,label>
credit	**<CulturalHeritageObject,provenance>** <Person,note>
classification	**<Concept,prefLabel>** <Concept,label>
name	**<aac:Person,name>** <foaf:Person,name>
imageURL	<Document,uri> **<WebResource,uri>**



Steiner Tree

title	**<CulturalHeritageObject,title>** <CulturalHeritageObject,label>
credit	**<CulturalHeritageObject,provenance>** <Person,note>
classification	**<Concept,prefLabel>** <Concept,label>
name	**<aac:Person,name>** <foaf:Person,name>
imageURL	<Document,uri> **<WebResource,uri>**

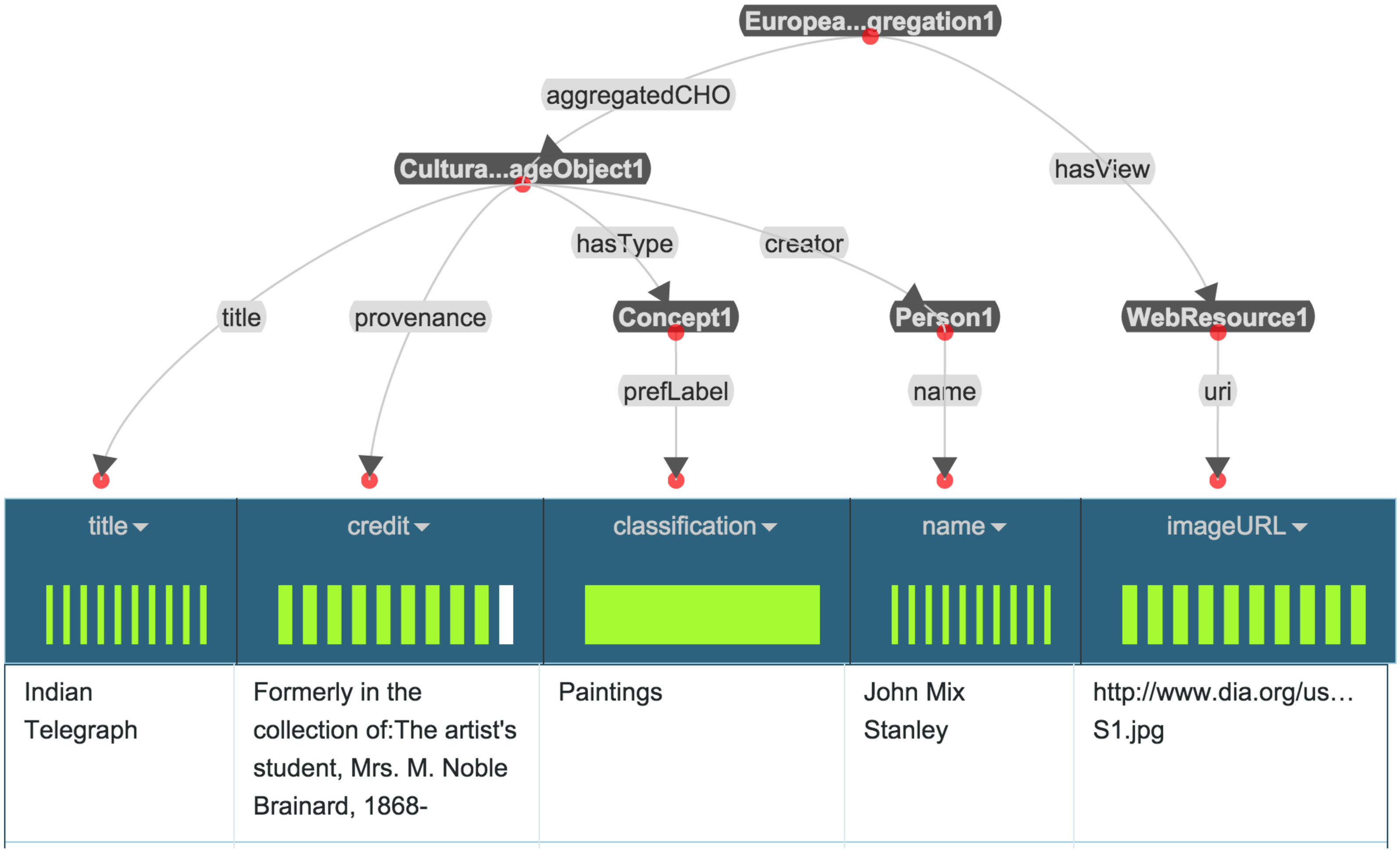


Final Model in Karma

Domain: Museum Data

Domain ontologies: EDM SKOS FOAF AAC ORE ElementsGr2 DCTerms

Source: Detroit Institute of Art ➔ **dia(title,credit,classification,name,imageURL)**



title	credit	classification	name	imageURL
Indian Telegraph	Formerly in the collection of:The artist's student, Mrs. M. Noble Brainard, 1868-	Paintings	John Mix Stanley	http://www.dia.org/us… S1.jpg

Evaluation

Evaluation Dataset	EDM	CRM
# sources	29	29
# classes in the ontologies	119	147
# properties in the ontologies	351	409
# nodes in the gold standard models	473	812
# links in the gold standard models	444	785

Compute precision and recall between learned models and correct models

$$precision = \frac{rel(sm)\,\text{Ç}\,rel(sm')}{rel(sm')}$$

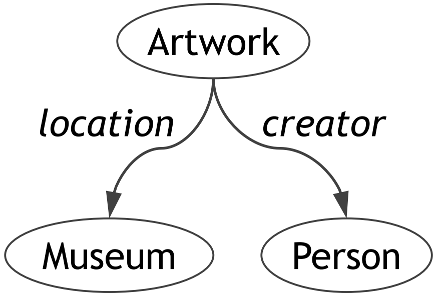
How many of the learned relationships are correct?

$$recall = \frac{rel(sm)\,\text{Ç}\,rel(sm')}{rel(sm)}$$

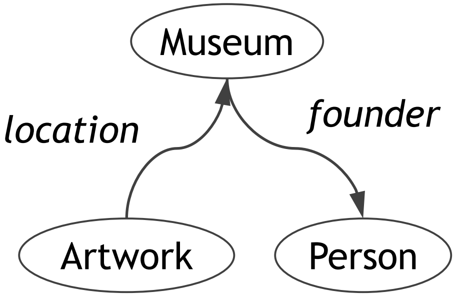
How many of the correct relationships are learned?

rel(sm) is the set of triples <source, link, target> in the semantic model

Example



correct model



learned model

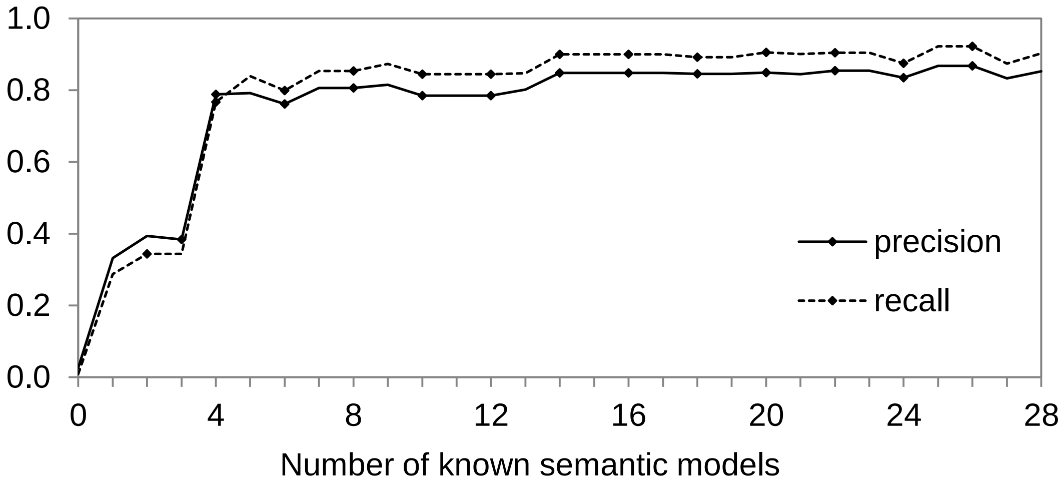
<Artwork,location,Museum>
<Artwork,creator,Person>

<Museum,founder,Person>
<Artwork,location,Museum>

Precision: 0.5
Recall: 0.5

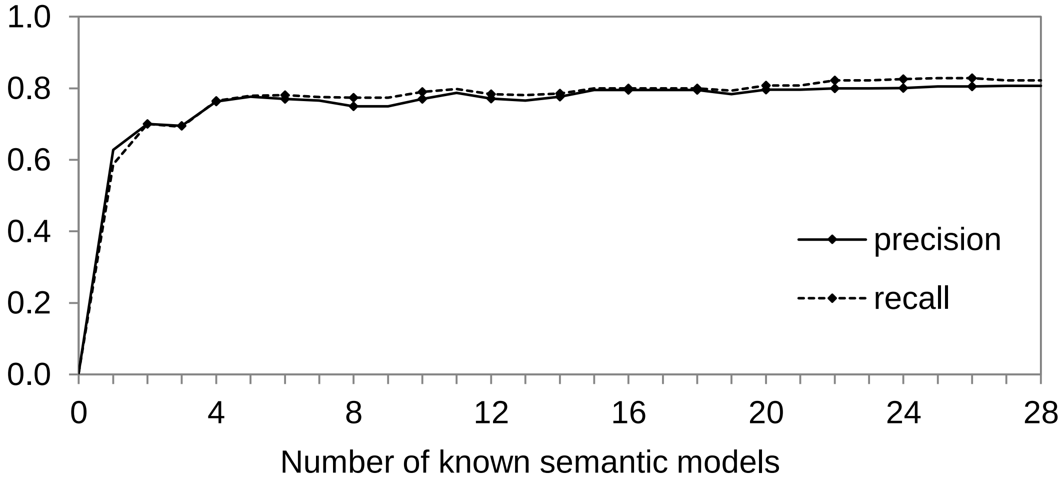
Experiment 1

correct semantic types are given



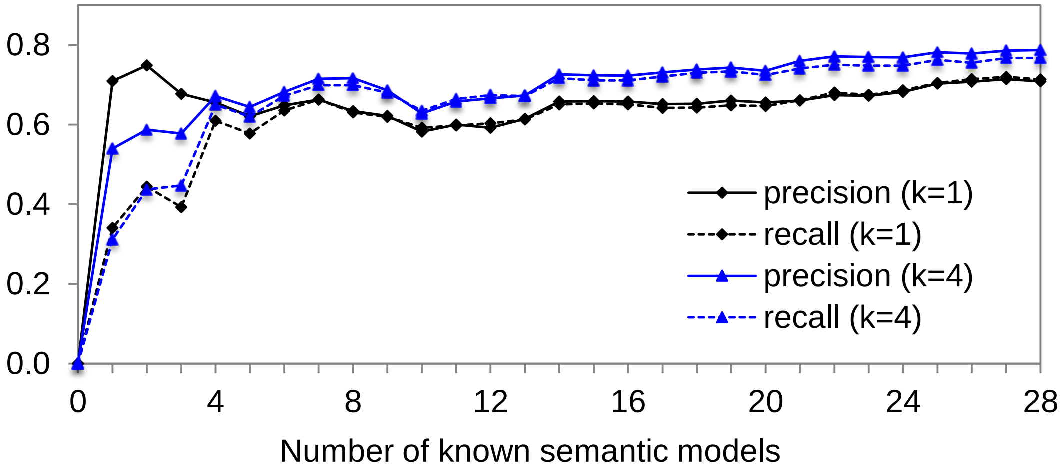
EDM

CRM



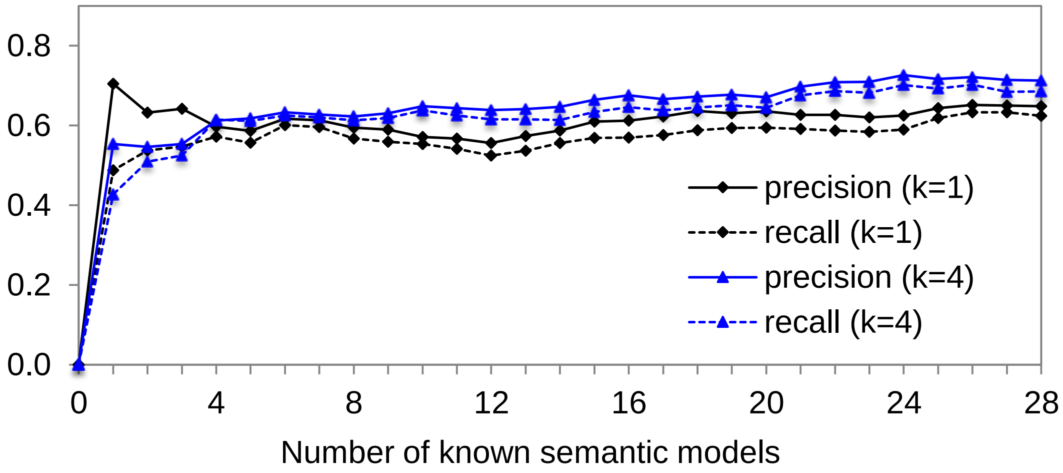
Experiment 2

learn semantic types, pick top K candidates



EDM

CRM



Limitation

- Lack of sufficient known semantic models is some domains

Inferring Semantic Relations from Linked Open Data

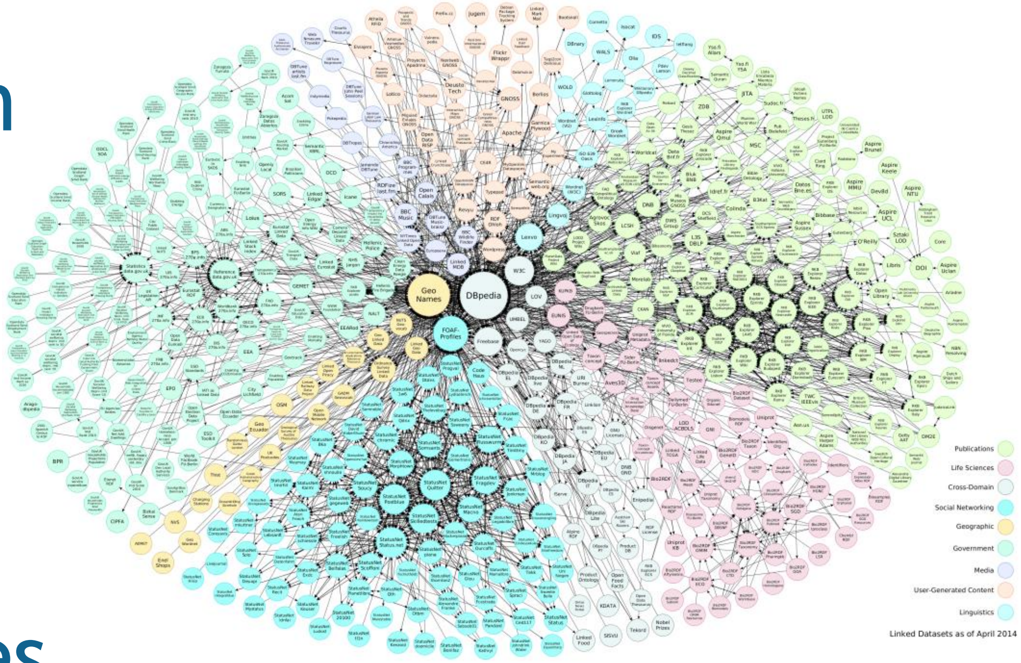
Contribution: leveraging graph patterns in LOD to infer relationships

Idea

- There is a huge amount of linked data available in many domains (RDF format)

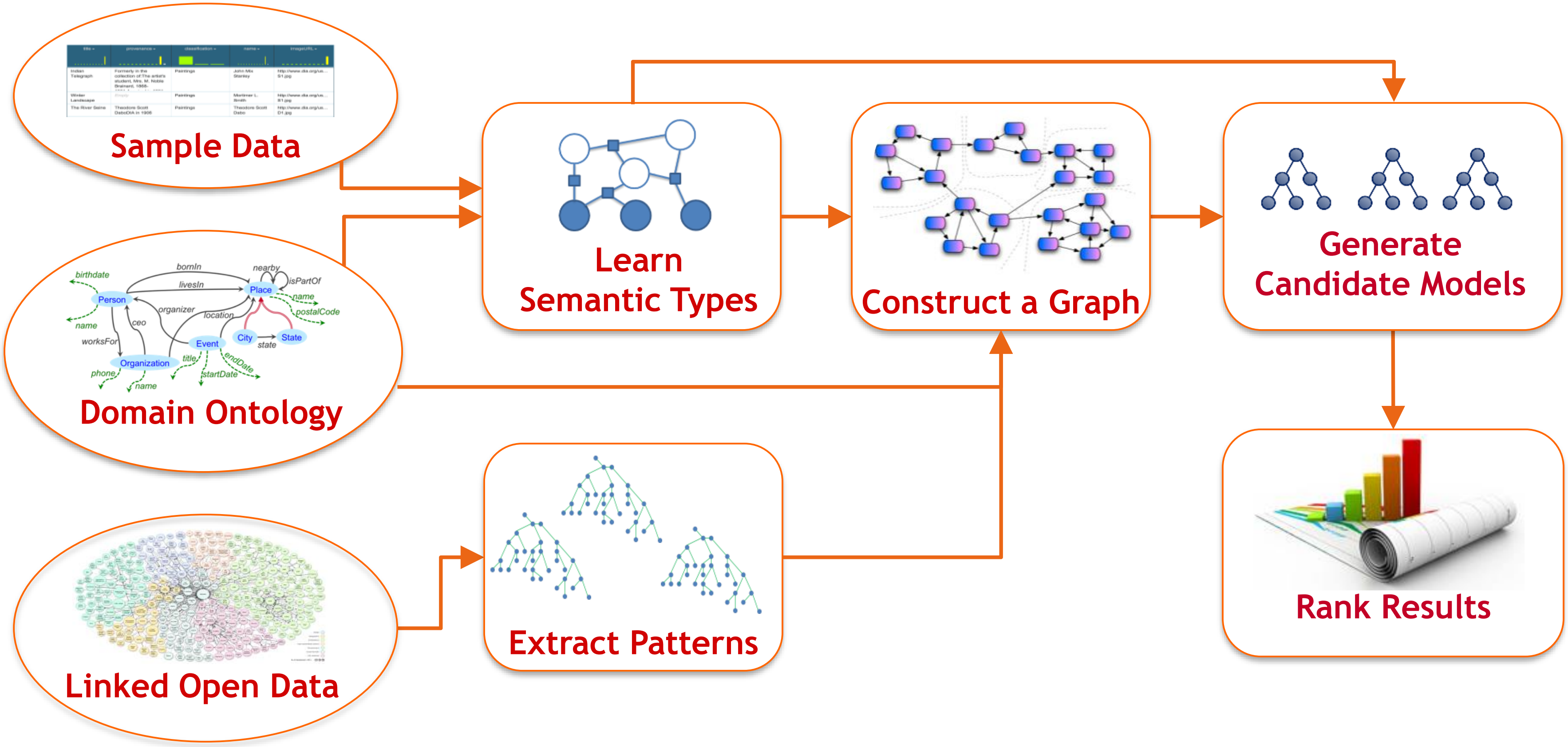
- Use LOD when there is no known semantic model

- Exploit the relationships between instances



Approach

[Taheriyan et al, COLD 2015]



Sample Data

Domain Ontology

Linked Open Data

Learn Semantic Types

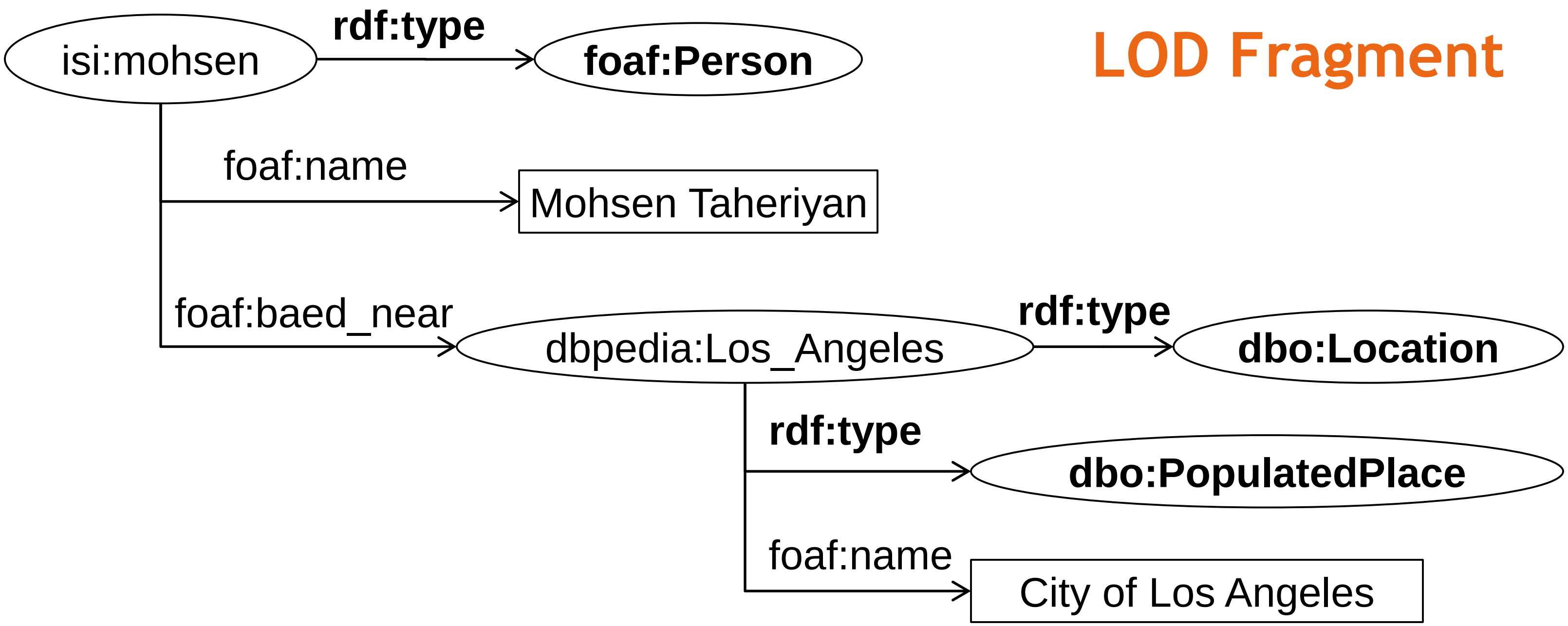
Construct a Graph

Generate Candidate Models

Extract Patterns

Rank Results

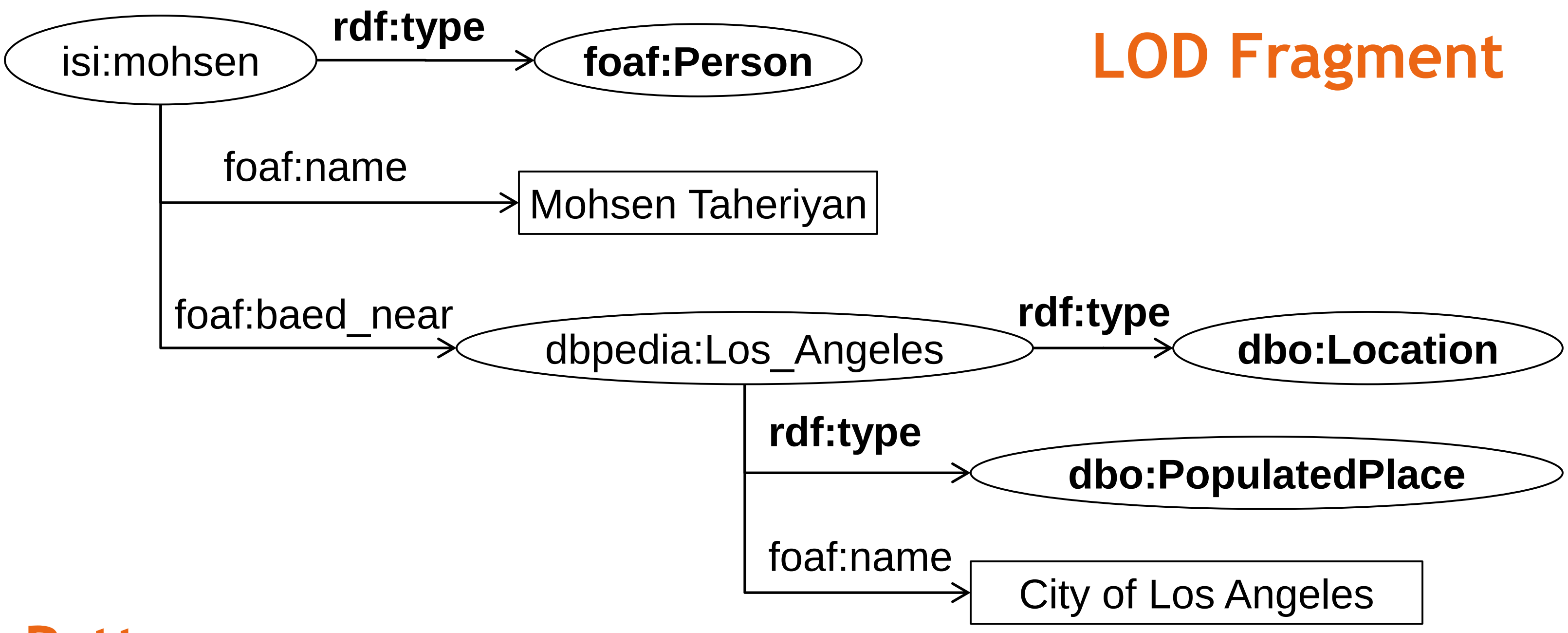
LOD Patterns



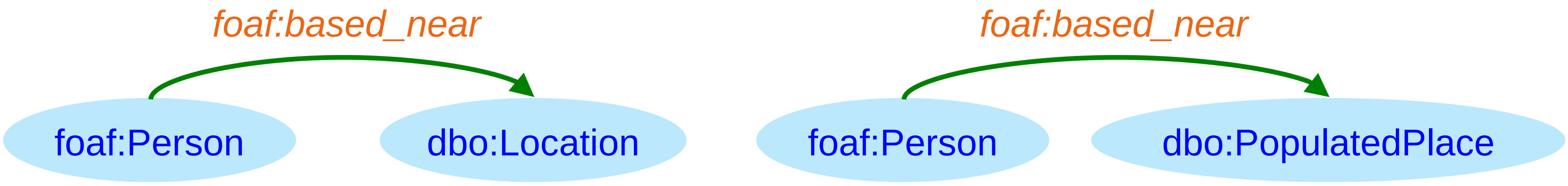
LOD Fragment

LOD Patterns

LOD Fragment



Patterns



Evaluation

- Linked data: 3,398,350 triples published by Smithsonian American Art Museum
- Correct semantic types given
- Extracted patterns of length 1 and 2

Evaluation Dataset	CRM
# sources	29
# classes in the ontologies	147
# properties in the ontologies	409
# nodes in the gold standard models	812
# links in the gold standard models	785

background knowledge	precision	recall	time (s)
domain ontology	0.07	0.05	0.17
domain ontology + patterns of length 1	0.65	0.55	0.75
domain ontology + patterns of length 1 and 2	0.78	0.70	0.46

Related Work

Related Work

- Mapping databases and spreadsheets to ontologies
  - Mapping languages: D2R [Bizer, 2003], D2RQ [Bizer and Seaborne, 2004], R2RML [Das et al., 2012]
  - Tools: RDOTE [Vavliakis et al., 2010], RDF123 [Han et al., 2008], XLWrap [Langegger and Woß, 2009]
  - String similarity between column names and ontology terms [Polfliet and Ichise, 2010]
- Understand semantics of Web tables
  - Use column headers and cell values to find the labels and relations from a database of labels and relations populated from the Web [Wang et al., 2012] [Limaye et al., 2010] [Venetis et al., 2011]
- Exploit Linked Open Data (LOD)
  - Link the values to the entities in LOD to find the types of the values and their relationships [Muoz et al., 2013] [Mulwad et al., 2013]
- Semantic annotation of Web services
  - Languages: SAWSDL [Farrell and Lausen, 2007]
  - Tools: SWEET [Maleshkova et al., 2009]
  - Annotate input and output parameters [Heß et al., 2003] [Lerman et al., 2006] [Saquicela e al., 2011]
- Learn Semantic Definitions of Online Information Sources [Carman, Knoblock, 2007]
  - Learns LAV rules from known sources
  - Only learns descriptions that are conjunctive combinations of known descriptions

Discussion & Future Work

Discussion

- Contributions
  - Semi-automatically model the relationships
  - Learn semantic models from previous models
  - Infer semantic relationships from LOD
- Provide explicit semantics for large portion of LOD
- Help to publish consistent RDF data
- Applications
  - VIVO
  - Smithsonian American Art Museum
  - DIG for DARPA's Memex project

Future Work

- Improve the quality of semantic labeling
  - Use LOD to learn semantic types
- Extract longer patterns from LOD
- Publish linked data
  - Transform the data to a common vocabulary
  - Linking entities across different datasets