USC Viterbi
School of Engineering

Linking and Building Ontologies of Linked Data

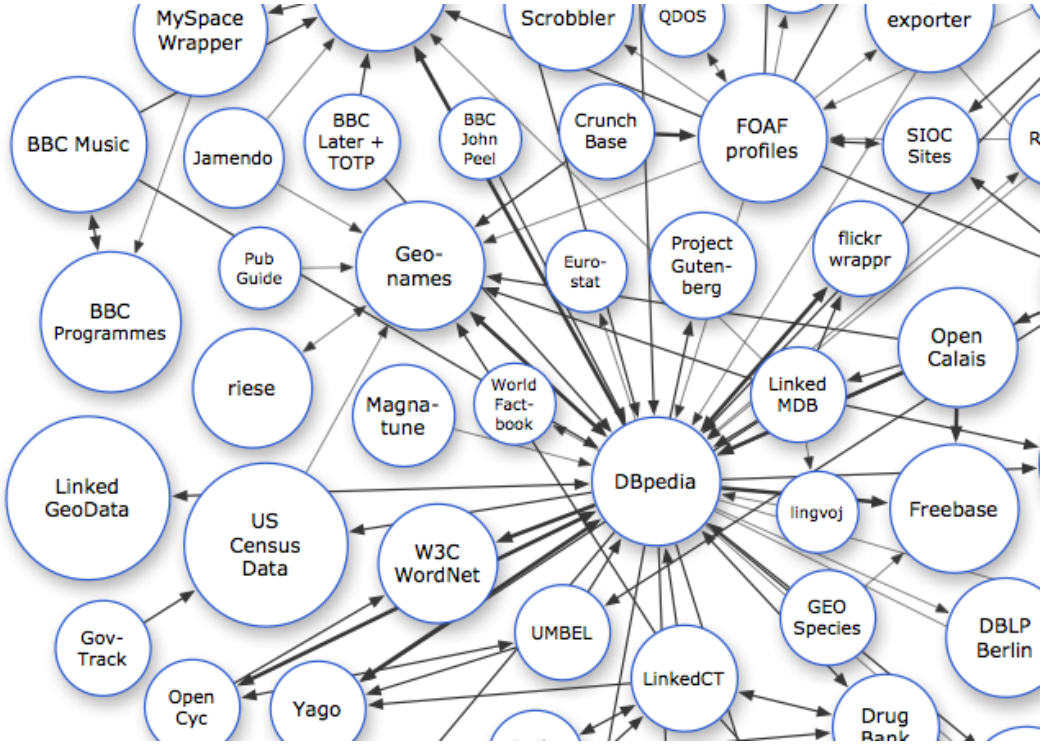
Rahul Parundekar, Craig A. Knoblock and Jose-Luis Ambite
{parundek,knoblock,ambite}@isi.edu
University of Southern California

USC

USC Viterbi School of Engineering

Web of Linked Data

- Vast collection of interlinked information
- Different sources with different schemas



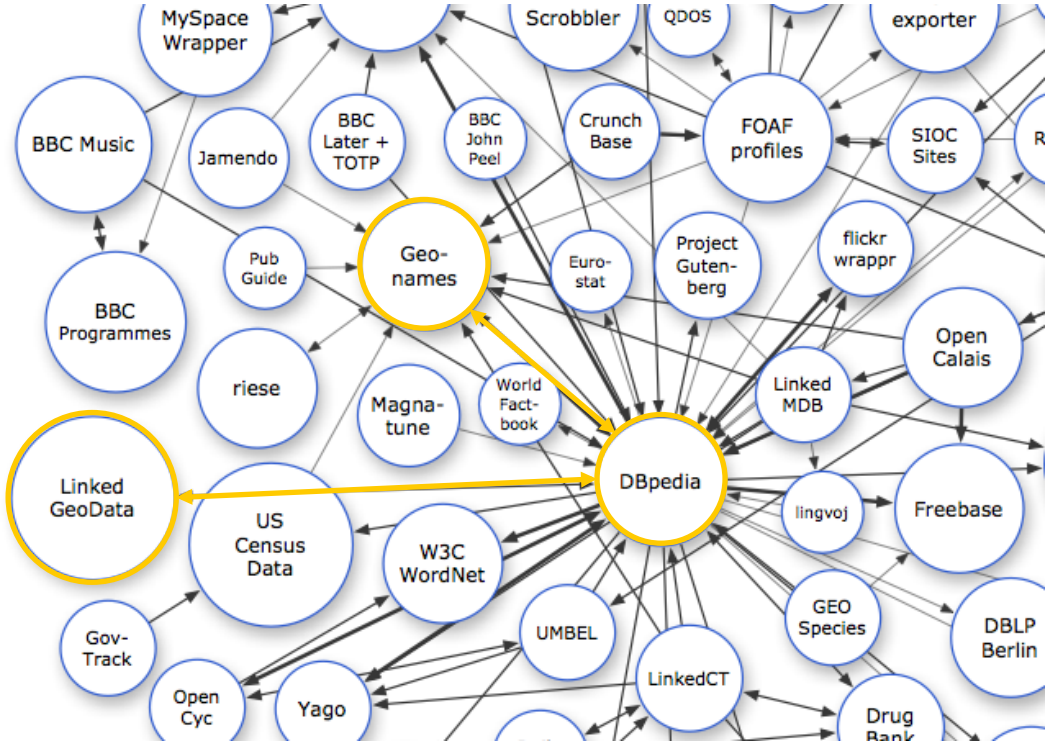
USC

USC Viterbi School of Engineering

Web of Linked Data

- Interlinked instances in the various domains
- Equivalent instances linked with *owl:sameAs*

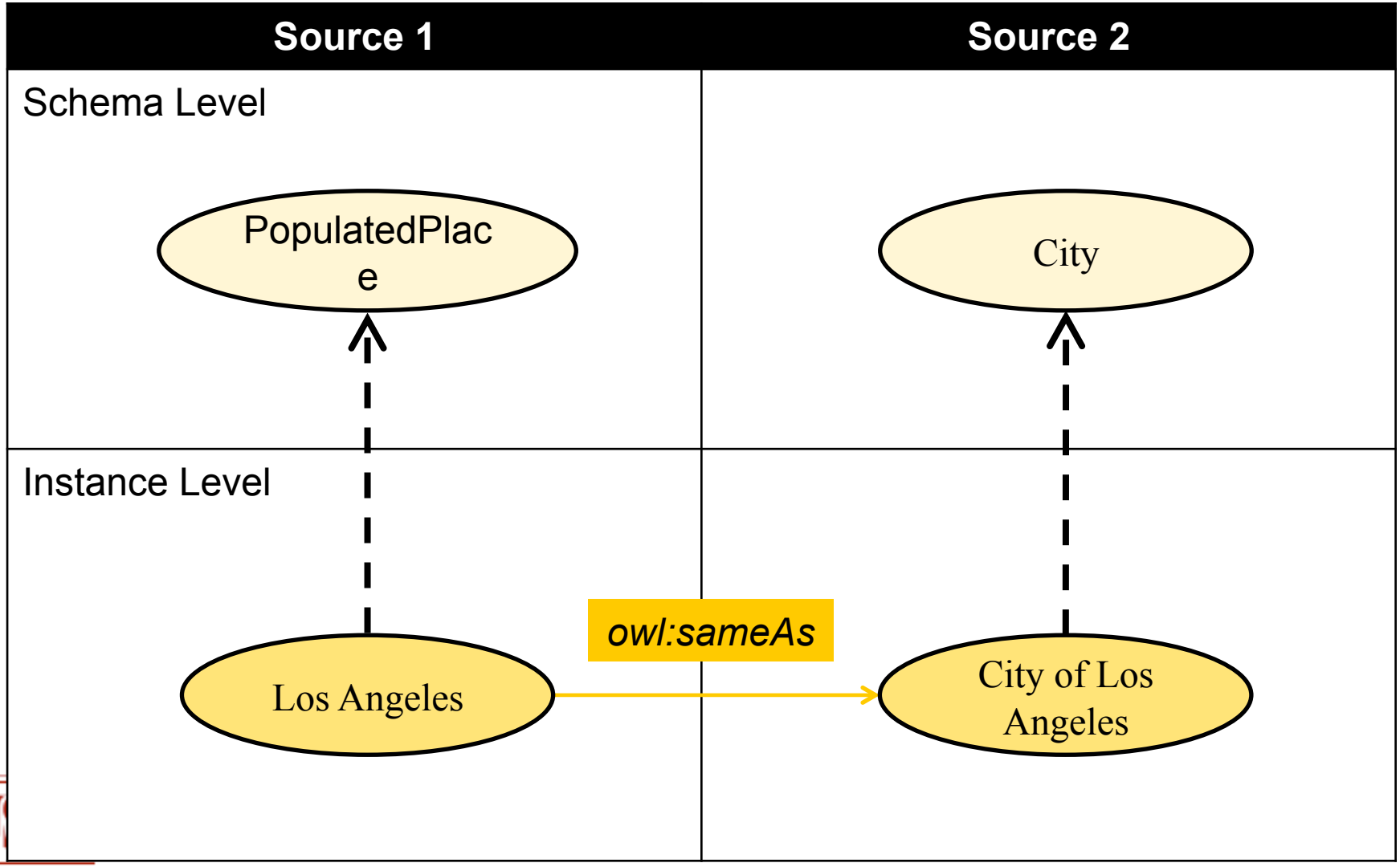
Geospatial Domain



USC

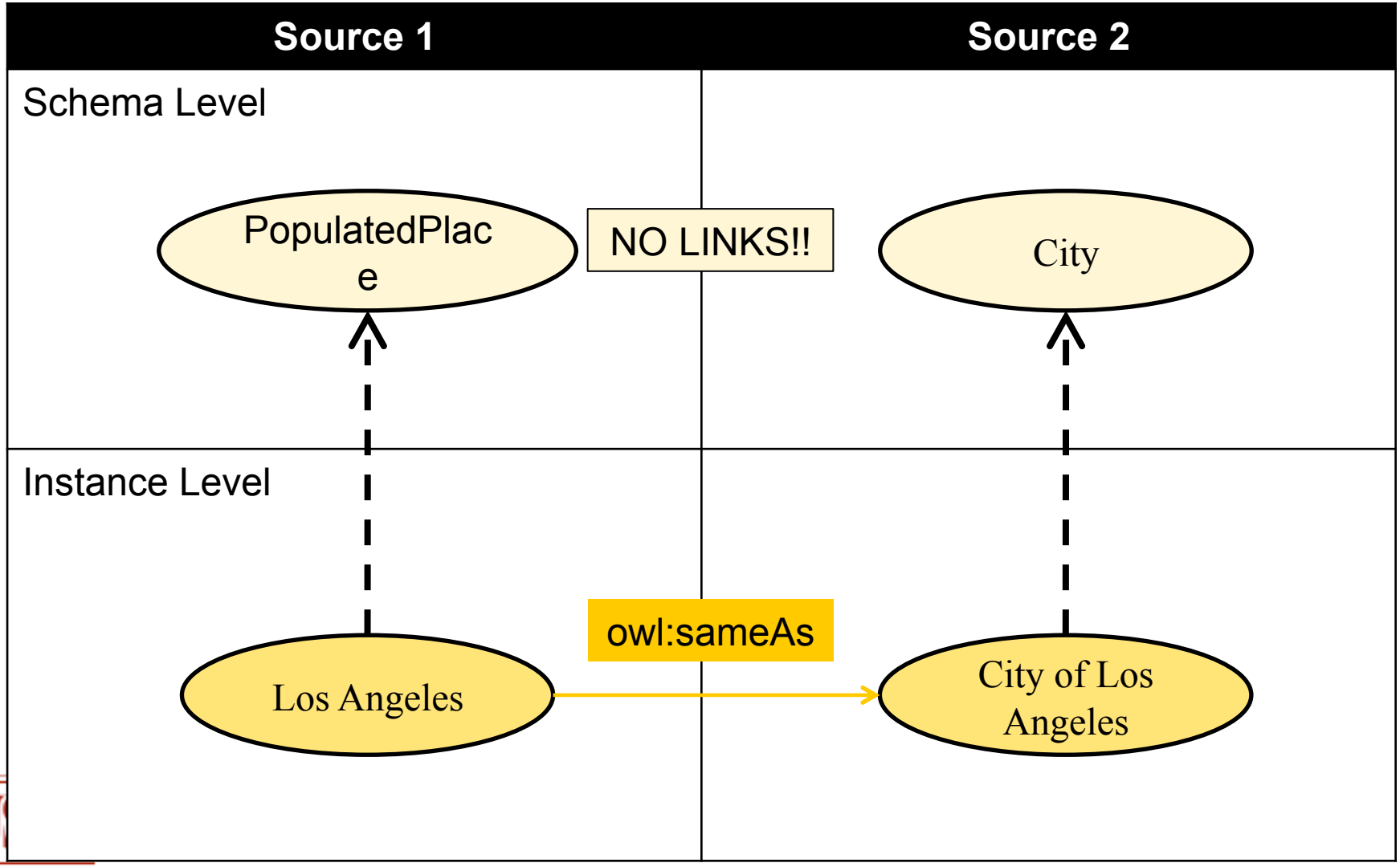
USC Viterbi School of Engineering

Interlinked Instances



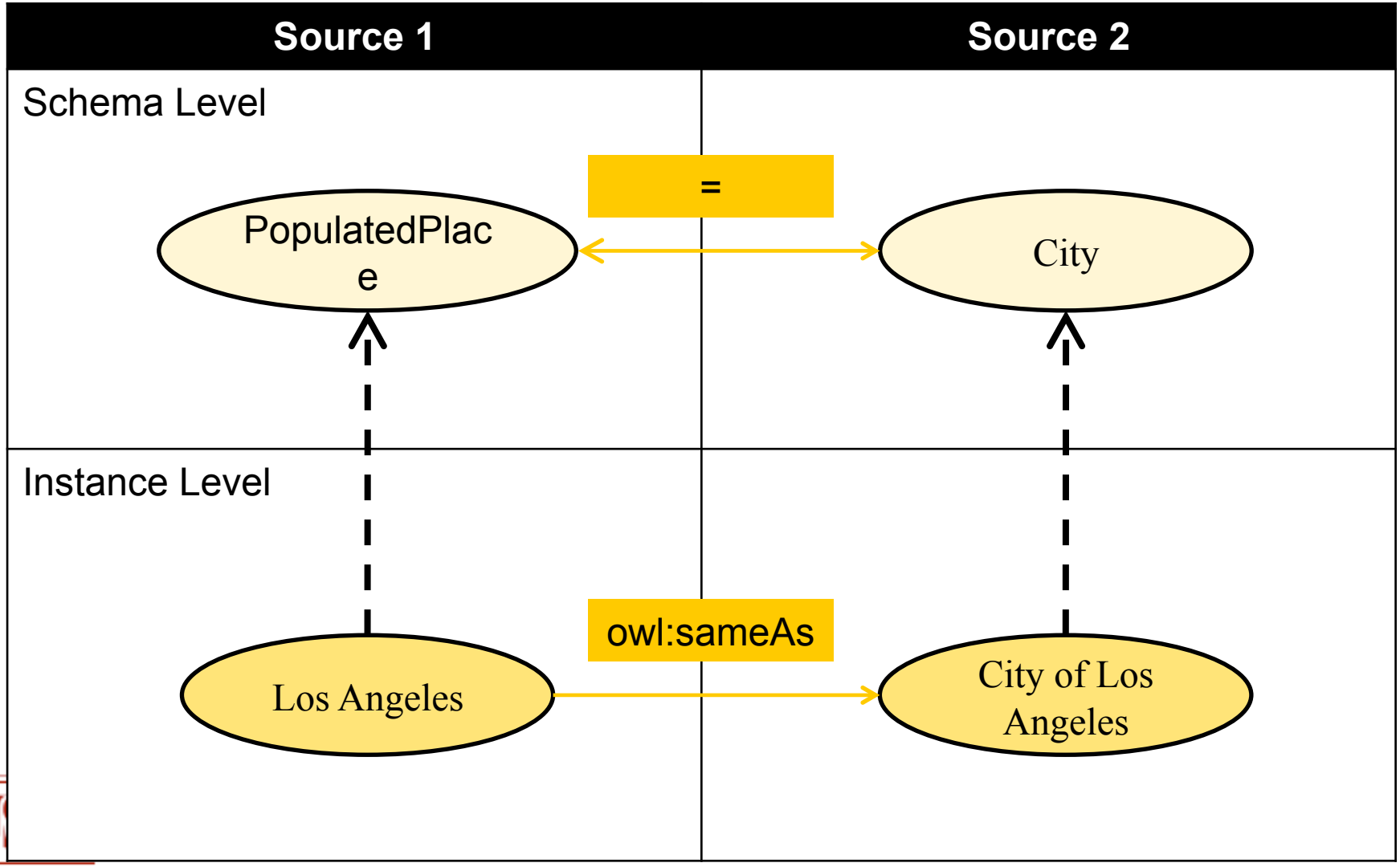
USC Viterbi School of Engineering

Disjoint Schemas



USC Viterbi School of Engineering

Objective 1: Find Schema Alignments



USC Viterbi School of Engineering

Ontologies of Linked Data

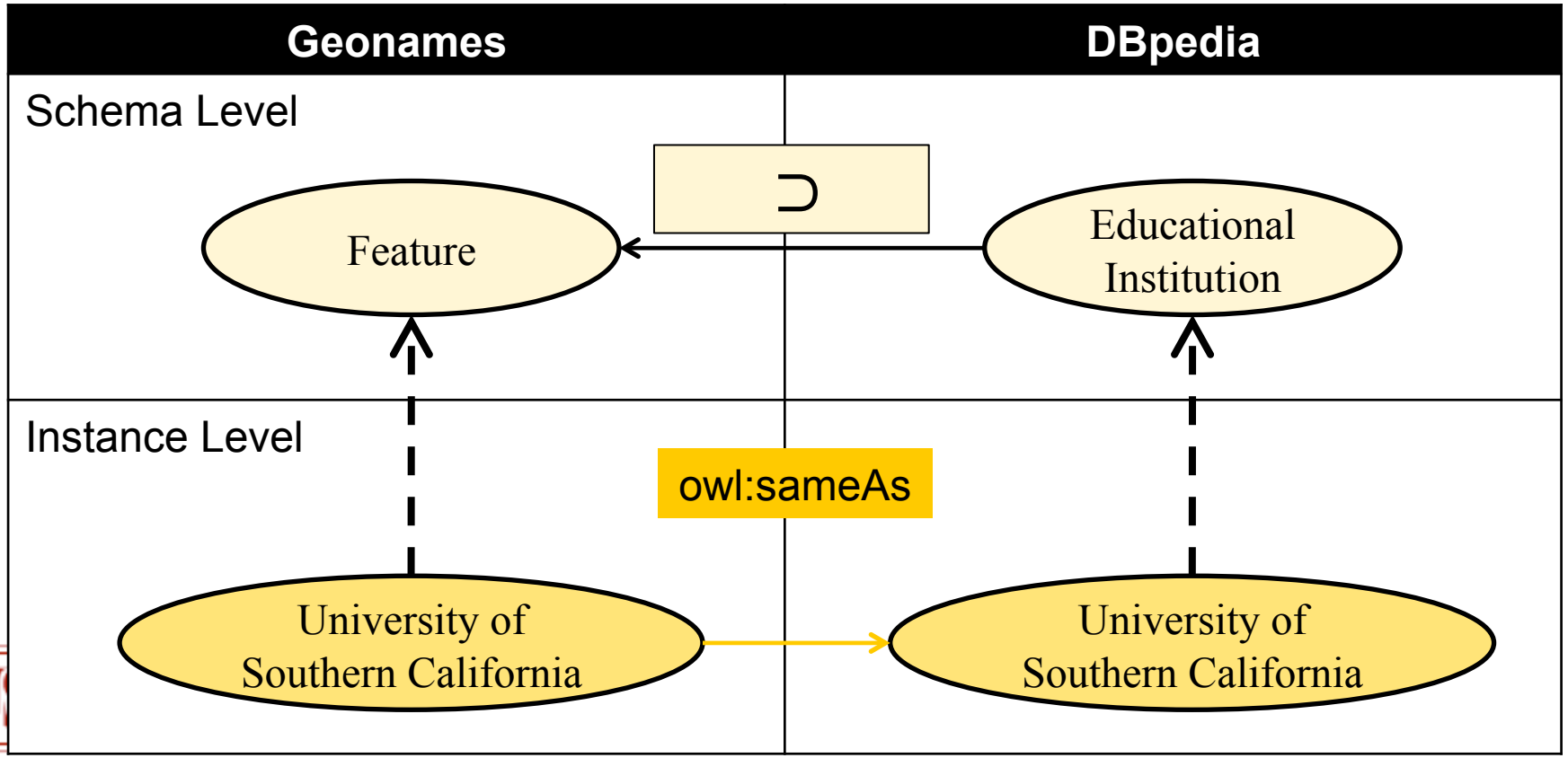
- Ontologies can be highly specialized
  - e.g. DBpedia has classes for *Educational Institutions, Bridges, Airports, etc.*
- But some can be rudimentary
  - e.g. in Geonames all instances only belong to a single class – 'Feature'
  - Derived from RDBMS schemas from which Linked Data was generated

USC

USC Viterbi School of Engineering

Traditional Alignments

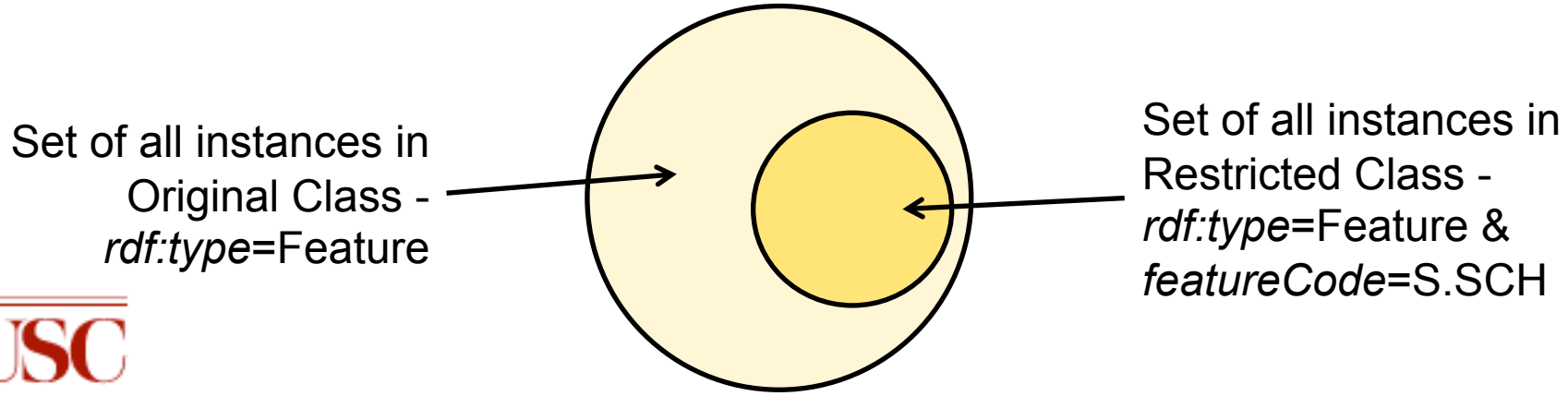
- There might not exist exact equivalences between classes in two sources
- Only subset relations possible



USC Viterbi School of Engineering

Restriction Classes

- A specialized class can be created by restricting the value of one or more properties
- The following Venn diagram explains a restriction class in Geonames with a restriction on the value of the *featureCode* property as 'S.SCH'

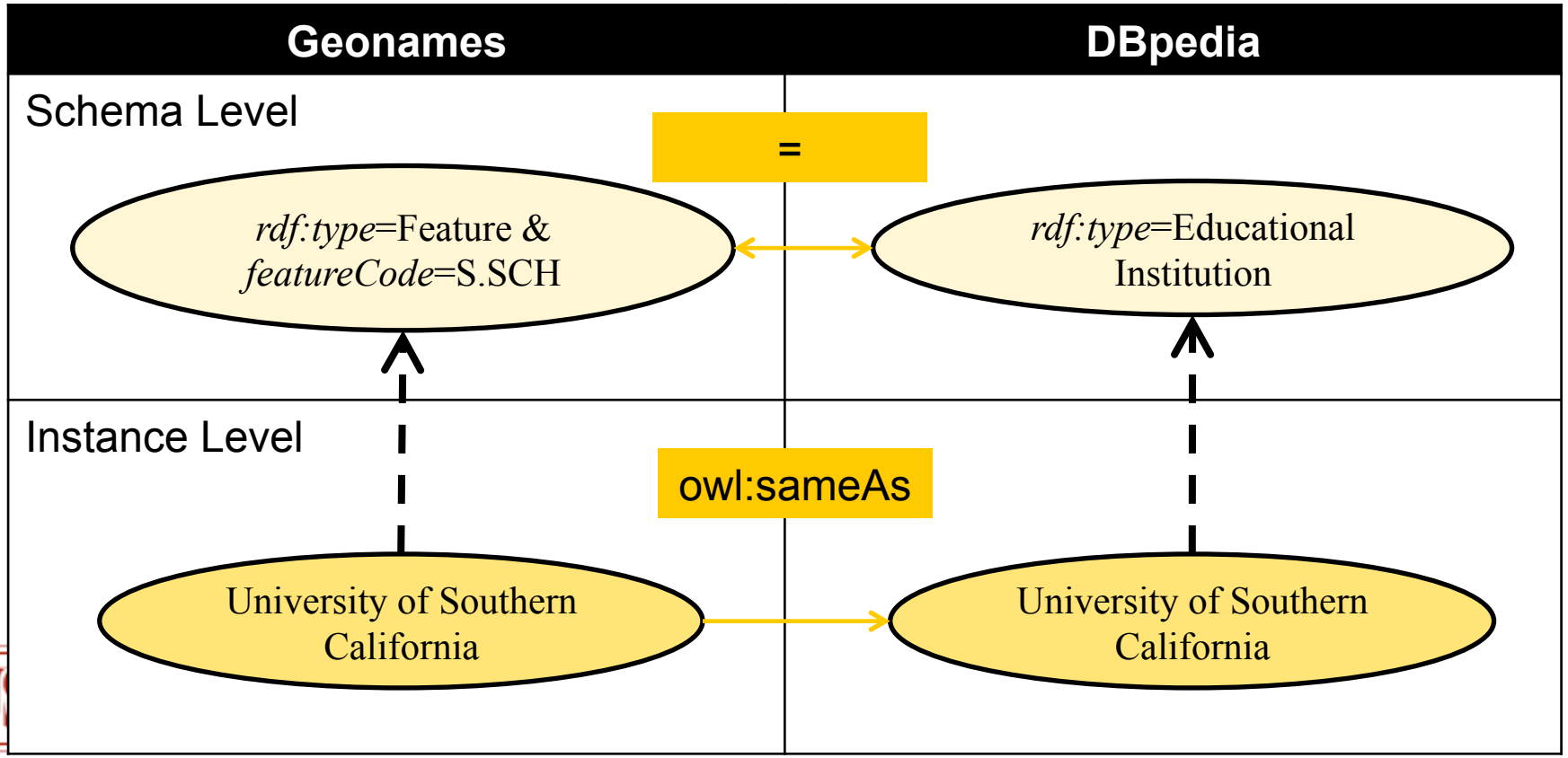


USC

USC Viterbi School of Engineering

Objective 2: Find Alignments Between Restriction Classes

- Find and model specialized descriptions of classes



USC Viterbi School of Engineering

Domains

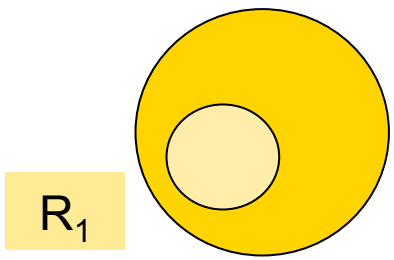
- Geospatial
  - Dbpedia
  - LinkedGeoData
  - Geonames
- Zoology
  - Geospecies
  - Dbpedia
- Genetics (Bio2RDF)
  - GeneID
  - MGI

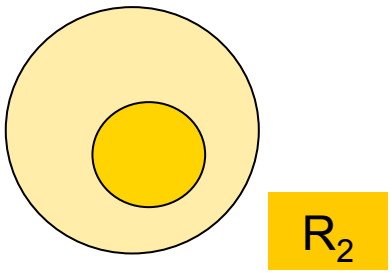
USC

USC Viterbi School of Engineering

Approach

- Aligning Restriction Classes



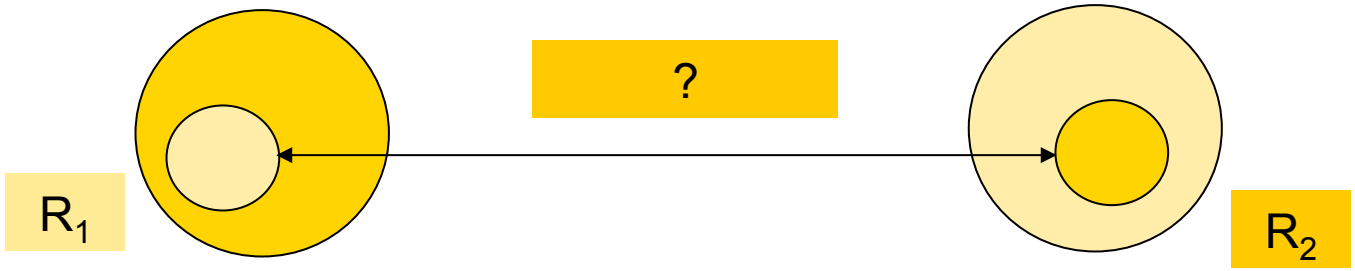


USC

USC Viterbi School of Engineering

Approach

- Aligning Restriction Classes

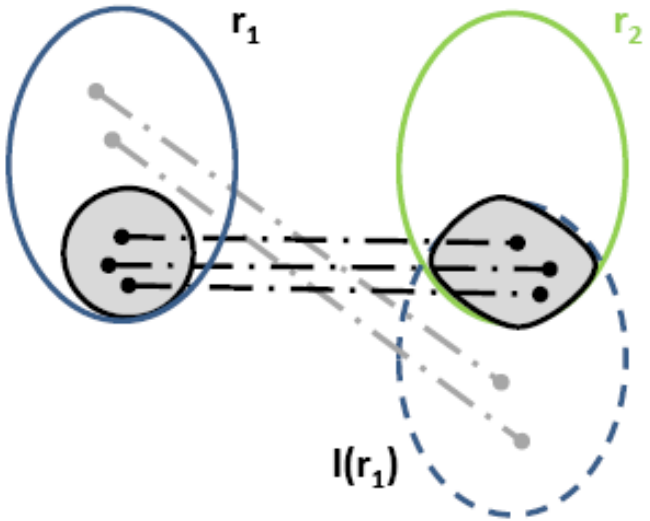


- Find relation between the two restriction classes
  - Equivalent
  - Subset

USC

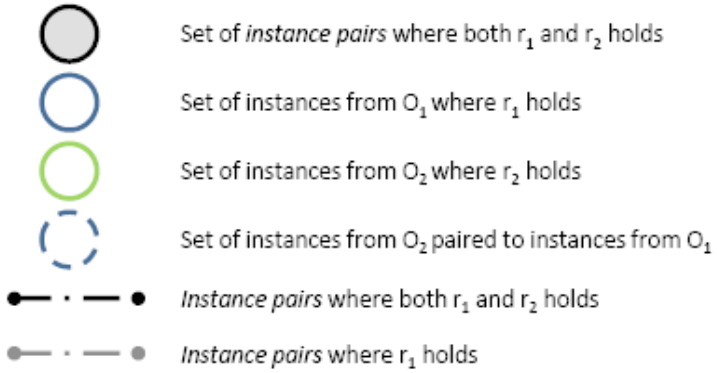
USC Viterbi School of Engineering

Extensional Approach to Ontology Alignment



Set Representation	Relation
	Disjoint
	$r_1 \subset r_2$
	$r_2 \subset r_1$
	$r_1 = r_2$
	Not enough support

Key:

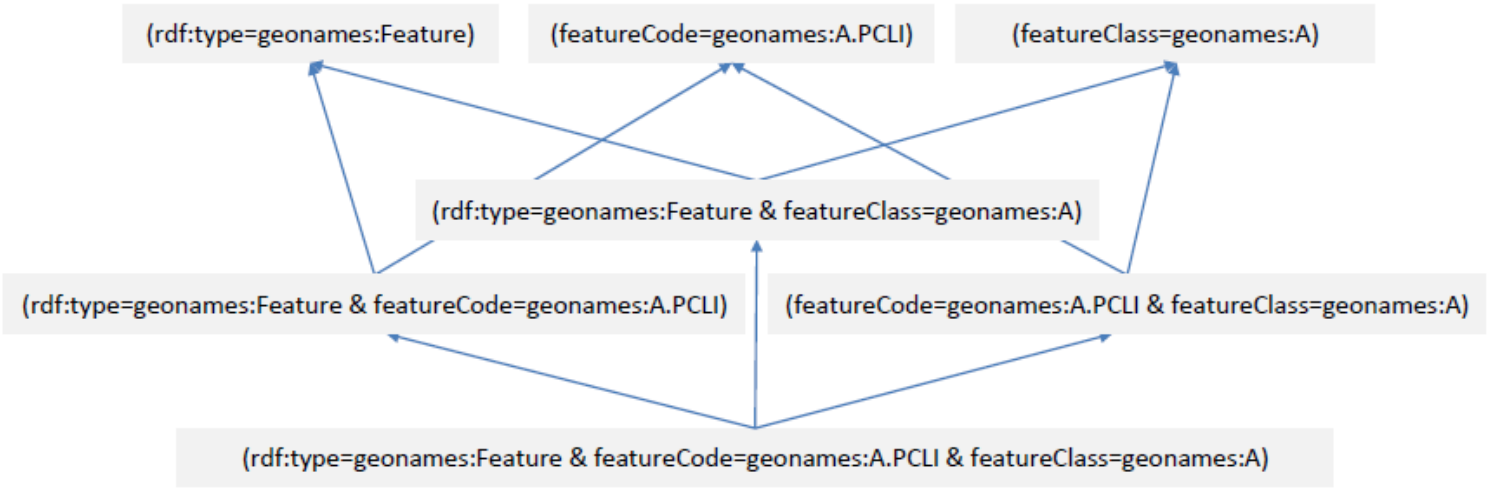


USC

USC Viterbi School of Engineering

Lattice of Restriction Classes

- Instances belonging to a restriction class also belong to parent restriction class
  - e.g. restrictions from Geonames below

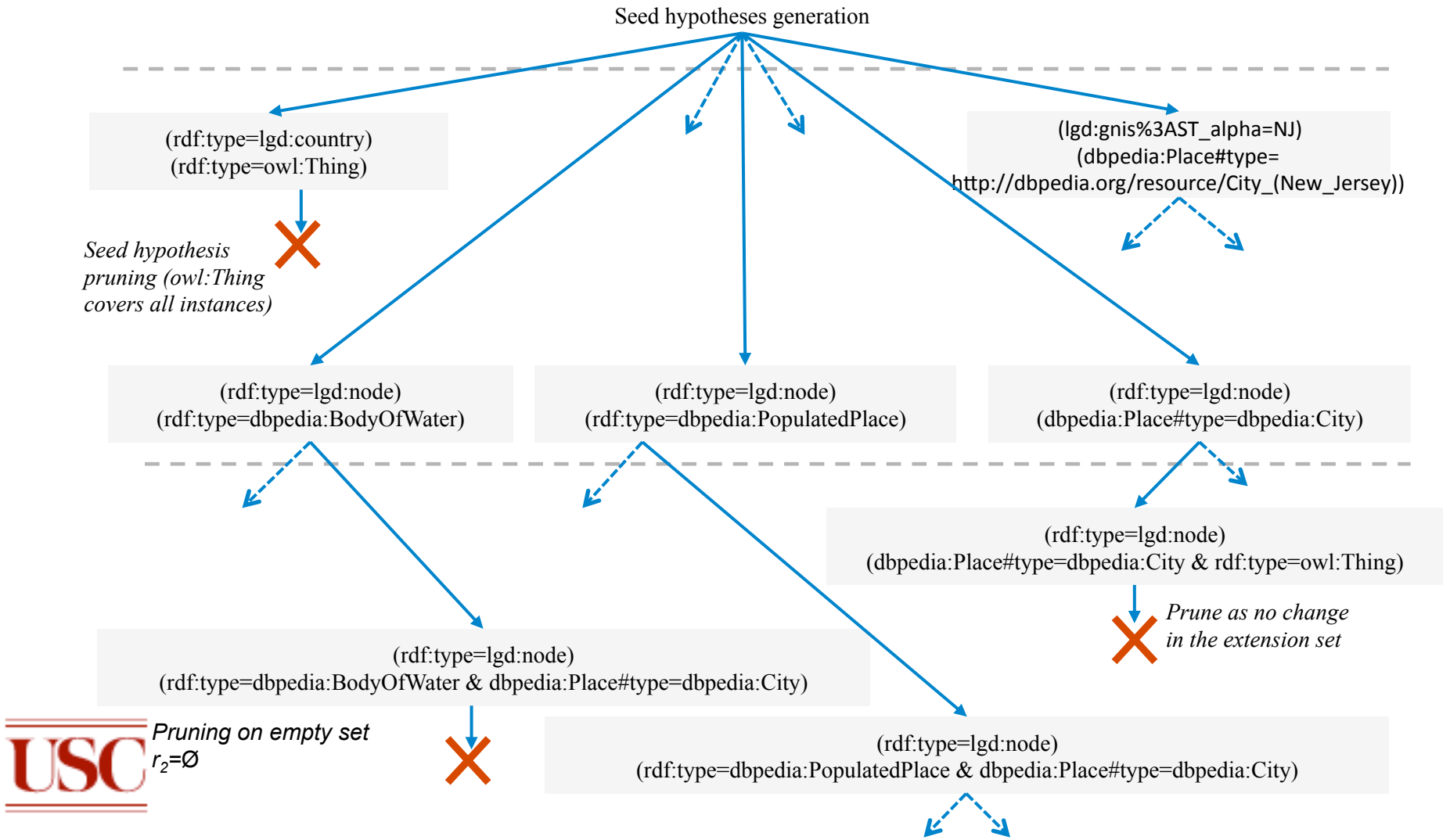


- This also results in a hierarchy in the alignments, which our algorithm exploits

USC

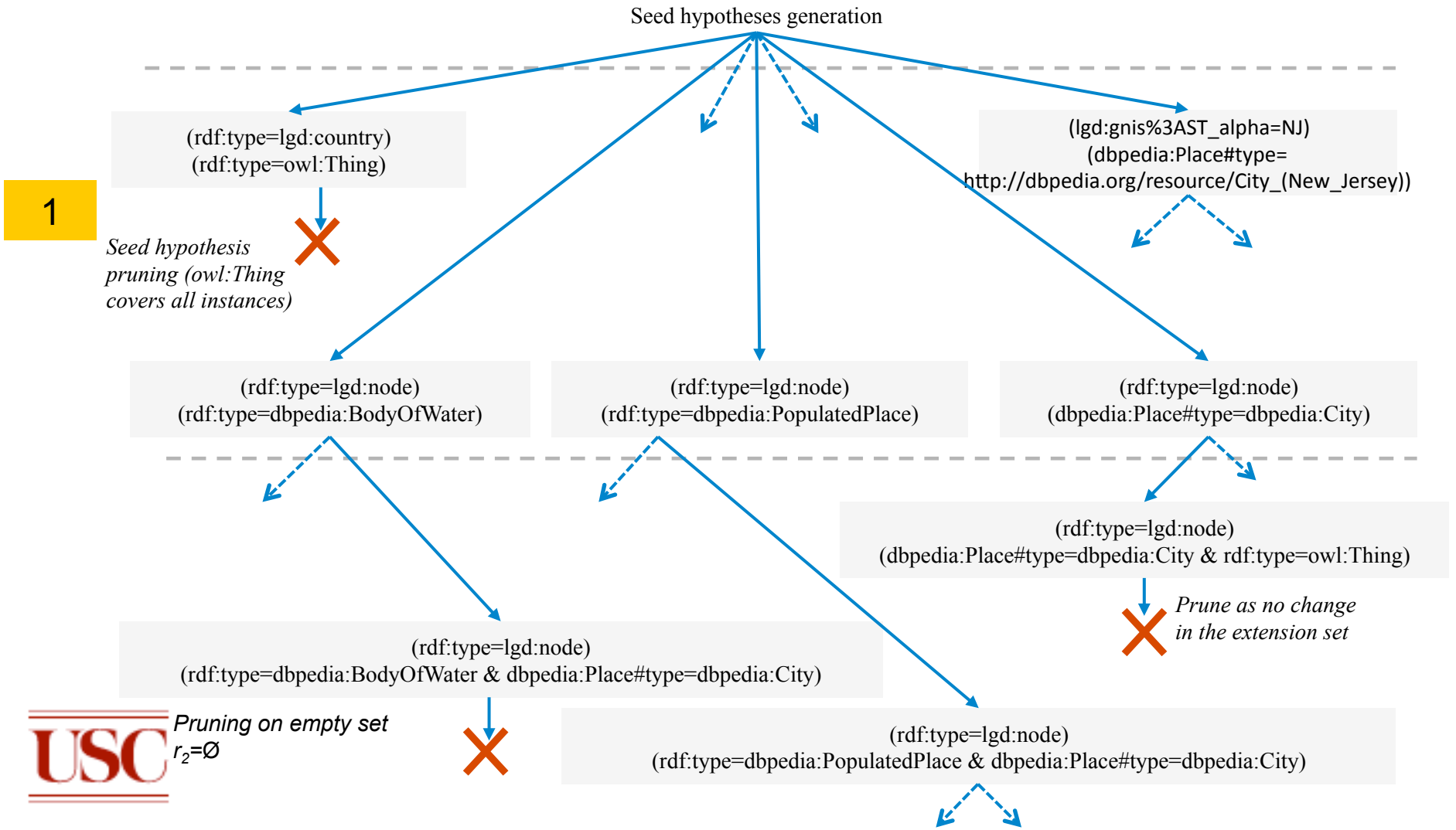
USC Viterbi School of Engineering

Exploration of Hypotheses Search Space (LinkedGeoData with DBpedia)



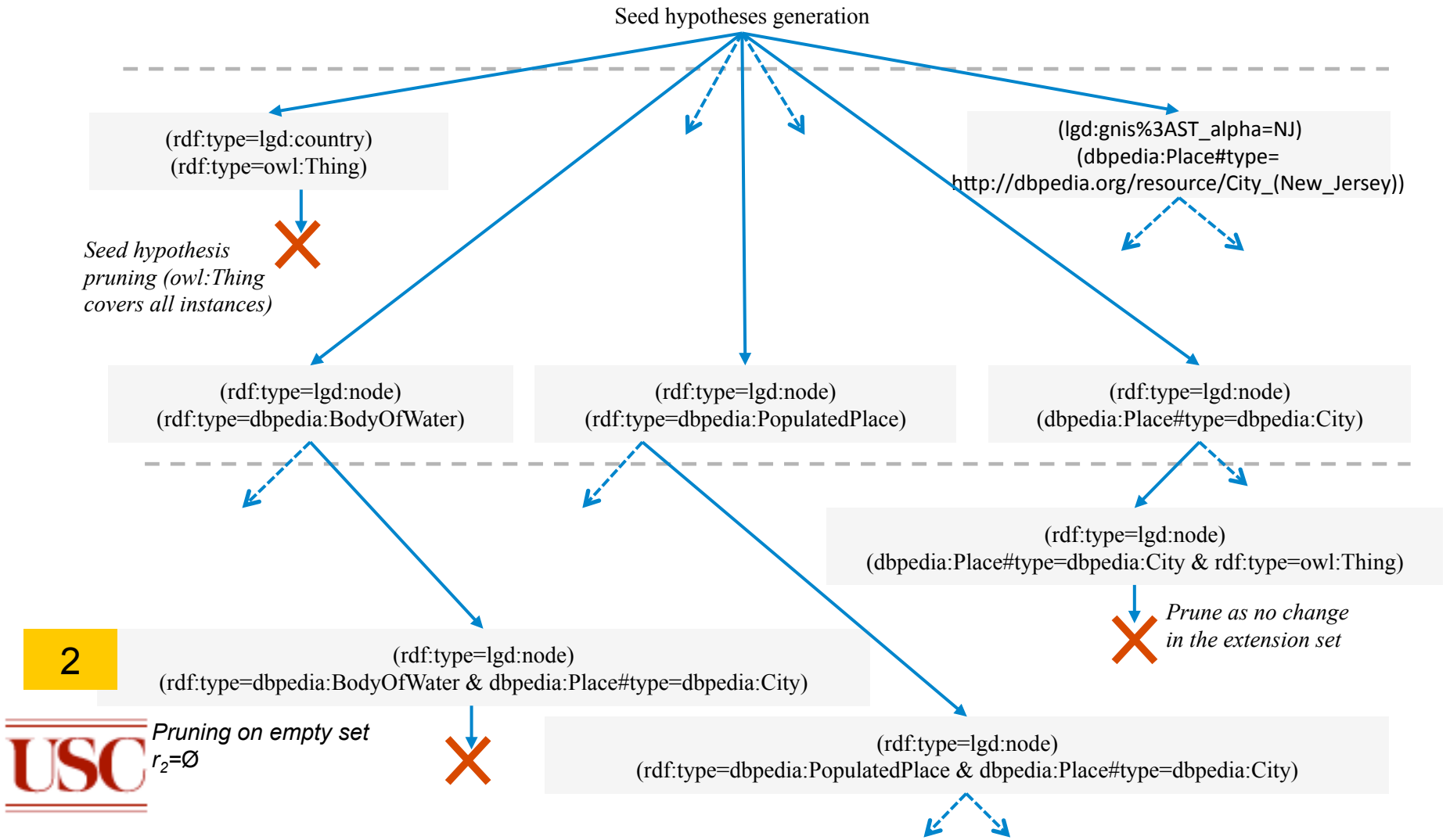
USC Viterbi School of Engineering

1. Prune seed hypothesis if either restriction covers all instances in that source



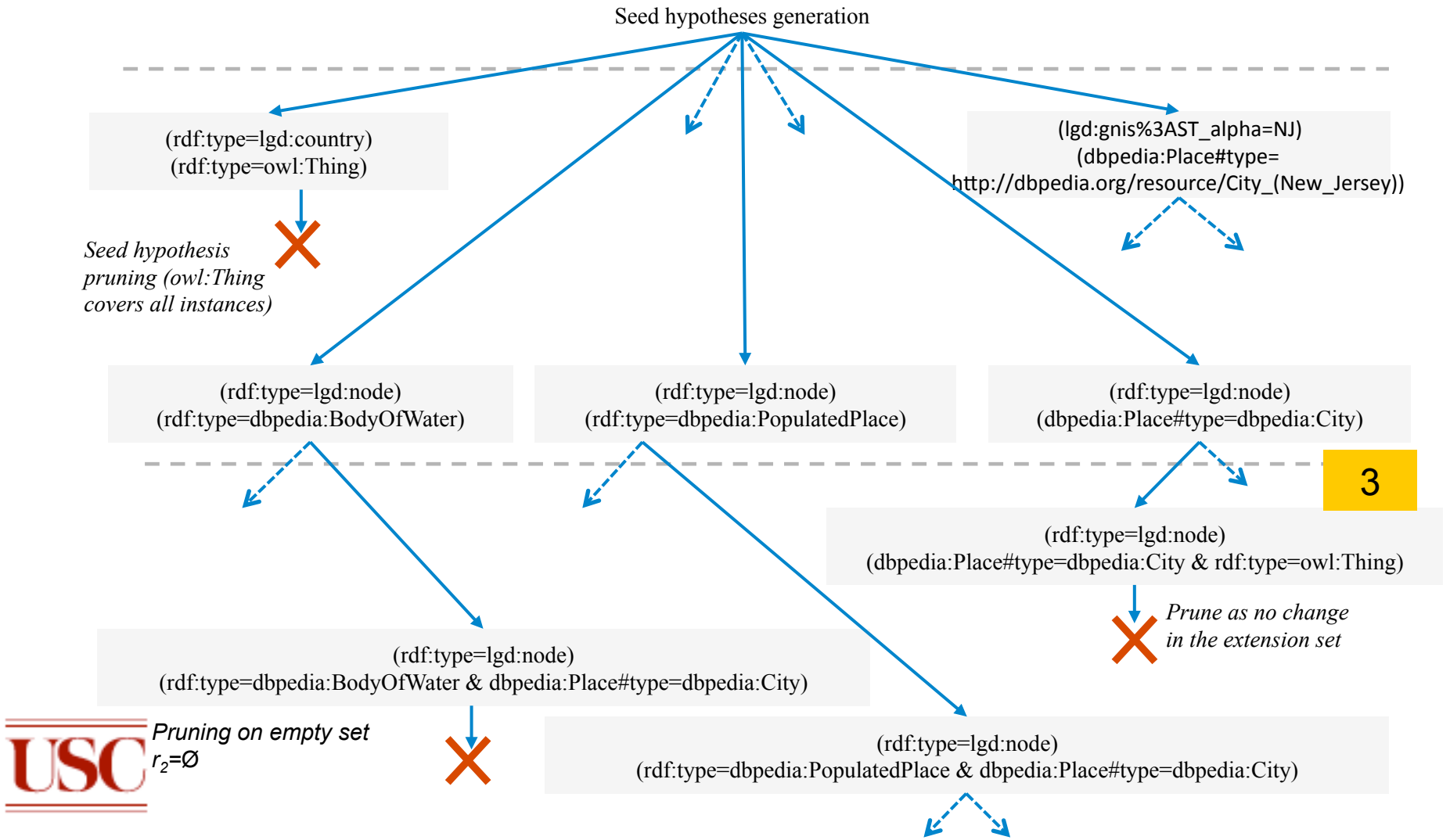
USC Viterbi School of Engineering

2. Number of instance pairs supporting hypothesis must be above a threshold



USC Viterbi School of Engineering

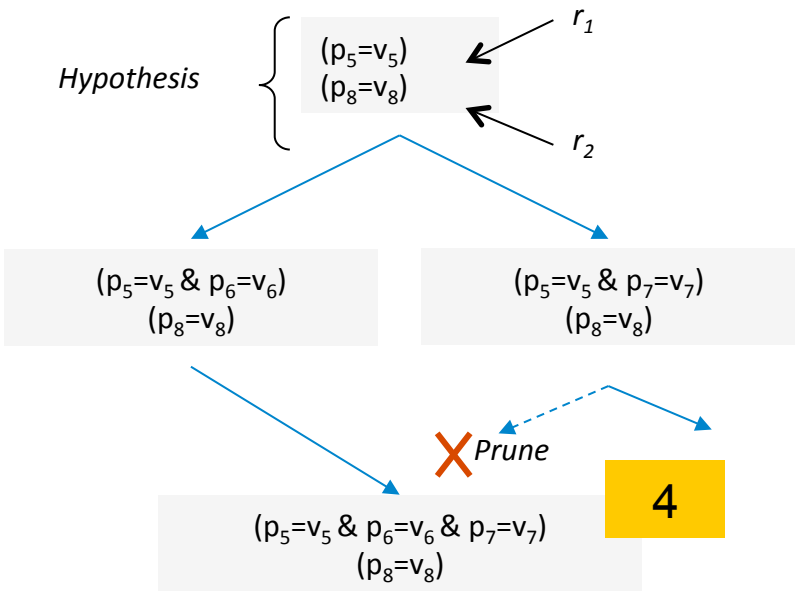
3. Prune if the added constraint does not change the extension



USC Viterbi School of Engineering

4. Lexicographic ordering

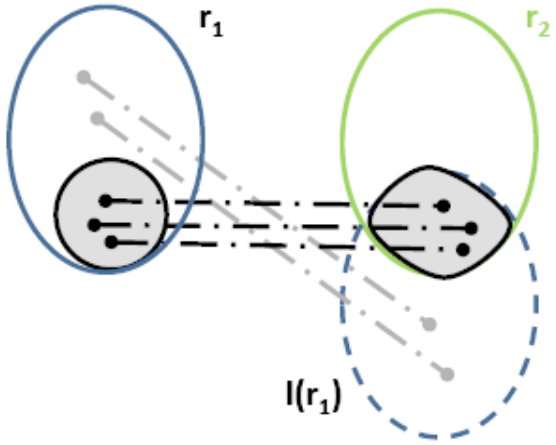
Lexicographic ordering provides a systematic search by pruning hypotheses with reverse order



USC

USC Viterbi School of Engineering

Relaxed Scoring



Set Representation	Relation	$P = \frac{\lvert I(r_1) \cap r_2 \rvert}{\lvert r_2 \rvert}$	$R = \frac{\lvert I(r_1) \cap r_2 \rvert}{\lvert r_1 \rvert}$	P′	R′
	Disjoint	= 0	= 0	≤ 0.01	≤ 0.01
	$r_1 \subset r_2$	< 1	= 1	> 0.01	≥ 0.90
	$r_2 \subset r_1$	= 1	< 1	≥ 0.90	> 0.01
	$r_1 = r_2$	= 1	= 1	≥ 0.90	≥ 0.90
	Not enough support	0 < P < 1	0 < R < 1	0.01 < P′ < 0.90	0.01 < R′ < 0.90

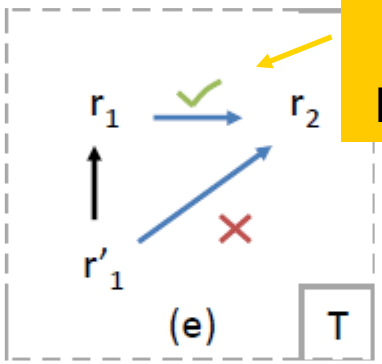
- Compensates for missing, inconsistent in the data

USC

USC Viterbi School of Engineering

Post-processing: Removing Implied Alignments

GEONAMES *restriction*	DBPEDIA *restriction*
geonames:featureCode=geonames:S.SCH	rdf:type=dbpedia:EducationalInstitution
geonames:featureCode=geonames:S.SCH & geonames:inCountry=geonames:US	rdf:type=dbpedia:EducationalInstitution

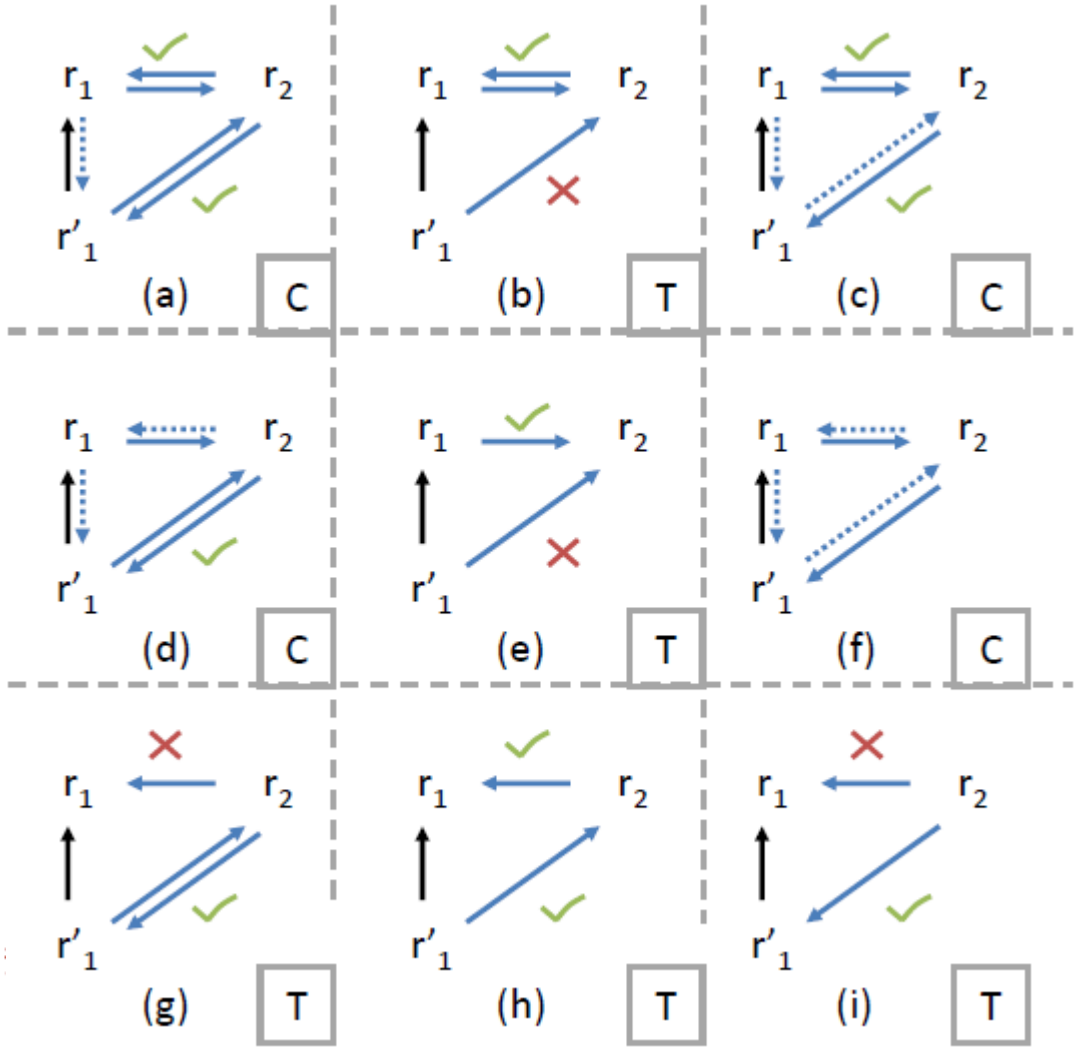


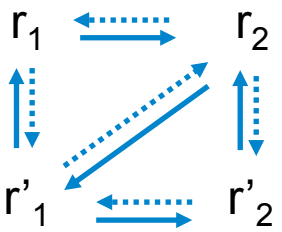
Keep the simpler definition
&
Remove the implied definition

Key:
$r_i \rightarrow r_j$: Subset relations ($r_i \subset r_j$) found by the algorithm.
$r_i \dashrightarrow r_j$: Implied subset relations.
$r'_i \rightarrow r_j$: Subset relation by construction.
T: Transitivity in subset relations. One relation can be eliminated.
C: Cycle in subset relations. Hence, all classes are equivalent.
✕ : Relation eliminated by the rule.
✓ : Relation retained by the rule.

USC Viterbi School of Engineering

Removing Implied Alignments





Cascading

Key:

$r_i \longrightarrow r_j$: Subset relations ($r_i \subset r_j$) found by the algorithm.

$r_i \dashrightarrow r_j$: Implied subset relations.

$r'_i \longrightarrow r_j$: Subset relation by construction.

T: Transitivity in subset relations. One relation can be eliminated.

C: Cycle in subset relations. Hence, all classes are equivalent.

✗ : Relation eliminated by the rule.

✓ : Relation retained by the rule.

USC Viterbi School of Engineering

Results: Geospatial Domain

#	LinkedGeoData *restriction*	DBpedia *restriction*	Relation
1	rdf:type=lgd:node	rdf:type=owl:Thing	$r_1 = r_2$
2	rdf:type=lgd:aerodrome	rdf:type=dbpedia:Airport	$r_1 = r_2$
3	rdf:type=lgd:island	rdf:type=dbpedia:Island	$r_1 = r_2$
4	lgd:gnis_%3AST_alpha=NJ	dbpedia:Place#type= http://dbpedia.org/resource/City_(New_Jersey)	$r_1 = r_2$
5	rdf:type=lgd:village	rdf:type=dbpedia:PopulatedPlace	$r_1 \subset r_2$
#	Geonames *restriction*	DBpedia *restriction*	Relation
6	geonames:featureClass=geonames:P	rdf:type=dbpedia:PpopulatedPlace	$r_1 = r_2$
7	geonames:featureClass=geonames:H	rdf:type=dbpedia:BodyOfWater	$r_1 = r_2$
8	geonames:parentFeature=http://sws.geonames.org/3174618/	dbpedia:City_region=http://dbpedia.org/resource/Lombardy	$r_1 = r_2$
9	geonames:featureCode=geonames:S.SCH	rdf:type=dbpedia:EducationalInstitution	$r_1 = r_2$
10	geonames:featureCode=geonames:S.SCH & geonames:inCountry=geonames:US	rdf:type=dbpedia:EducationalInstitution	$r_1 = r_2$
11	geonames:featureCode=geonames:T.MT	rdf:type=dbpedia:Mountain	$r_1 \subset r_2$

USC

USC Viterbi School of Engineering

Results: Zoology Domain

#	GEOSPECIES *restriction*	DBPEDIA *restriction*	Relation
12	geospecies:inKingdom=http://lod.geospecies.org/kingdoms/Aa	rdf:type=dbpedia:Animal	$r_1 = r_2$
13	geospecies:hasOrderName=Lepidoptera	dbpedia:order=http://dbpedia.org/resource/Lepidoptera	$r_1 = r_2$
14	geospecies:hasOrderName=Lepidoptera	dbpedia:kingdom=http://dbpedia.org/resource/Animal & dbpedia:order=http://dbpedia.org/resource/Lepidoptera	$r_1 = r_2$
15	geospecies:hasGenusName=Falco	dbpedia:genus=http://dbpedia.org/resource/Falcon	$r_1 = r_2$
16	geospecies:hasOrderName=Primates	dbpedia:order=http://dbpedia.org/resource/Primates	$r_2 \subset r_1$

#	GEOSPECIES *restriction*	GEOSPECIES *restriction*	Relation
20	geospecies:hasKingdomName=Animalia	geospecies:inKingdom=http://lod.geospecies.org/kingdoms/Aa	$r_1 = r_2$
21	geospecies:hasClassName=Insecta	geospecies:inClass= http://lod.geospecies.org/bioclasses/aQado	$r_1 \subset r_2$
22	geospecies:inFamily= http://lod.geospecies.org/families/amTJ9	geospecies:hasSubfamilyName=Sigmodontinae	$r_2 \subset r_1$

USC

USC Viterbi School of Engineering

Results: Genetics Domain

#	MGI *restriction*	GENEID *restriction*	Relation
17	bio2rdf:subType=Pseudogene	bio2rdf:subType=pseudo	$r_1 = r_2$
18	bio2rdf:subType=Pseudogene & mgi:genomeStart=17	geneid:chromosome=17 & bio2rdf:subType=pseudo	$r_1 = r_2$
19	bio2rdf:chromosomePosition=-1.00 & mgi:genomeStart=4	geneid:chromosome=4 & bio2rdf:subType=pseudo	$r_2 \subset r_1$

USC

USC Viterbi School of Engineering

Results: Alignments Found

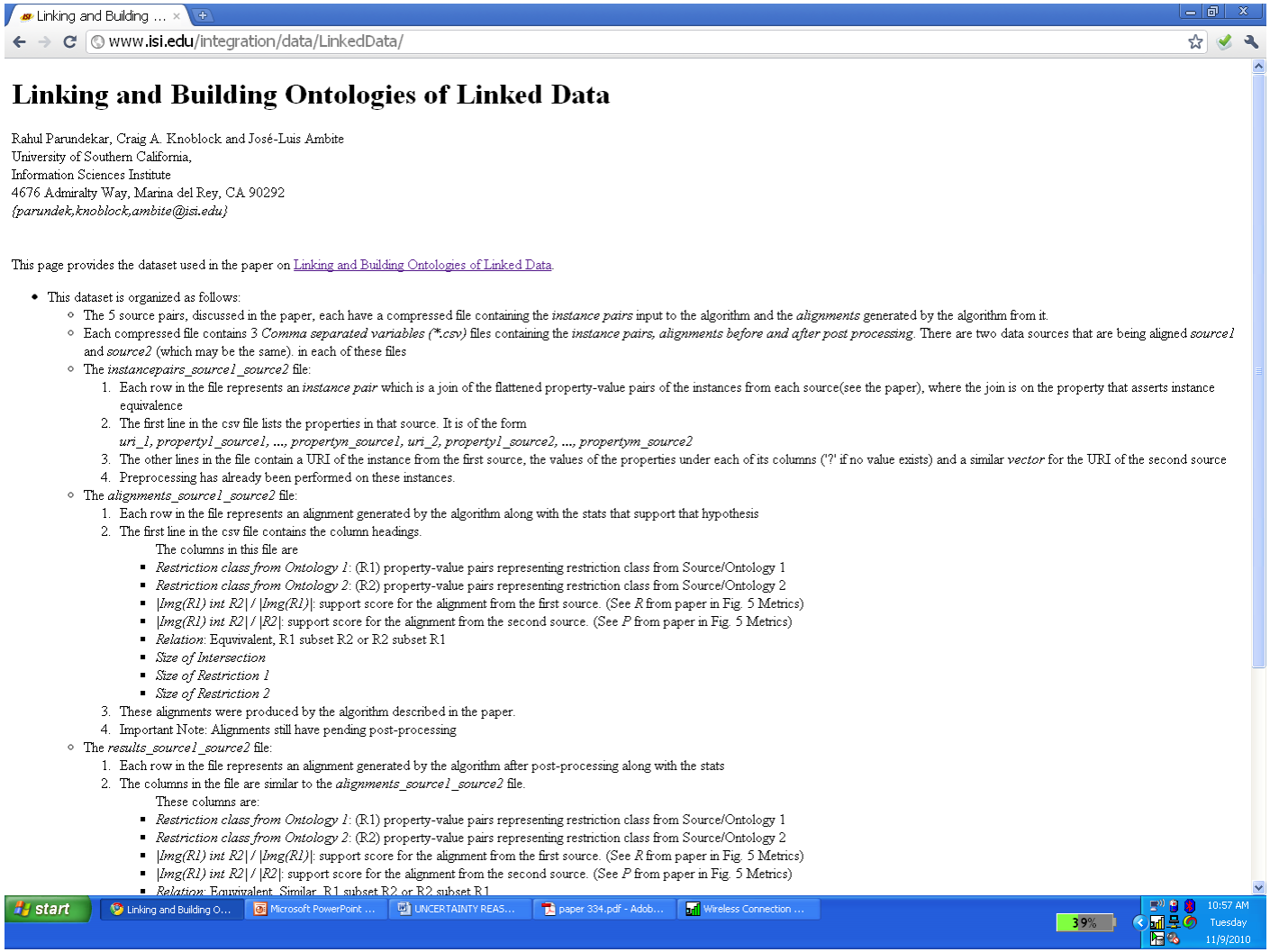
- Equivalences, Subset alignments before and after removing implied alignments

Source 1 (O_1)	Source 2 (O_2)	$\#(r_1 = r_2)$ total	$\#(r_1 = r_2)$ best matches	$\#(r_1 \subset r_2)$ before	$\#(r_1 \subset r_2)$ after	$\#(r_2 \subset r_1)$ before	$\#(r_2 \subset r_1)$ after
LinkedGeoData	DBpedia	158	152	2528	1837	1804	1627
Geonames	DBpedia	31	19	809	400	1384	1247
Geospecies	DBpedia	509	420	9112	2294	6098	4455
MGI	GeneID	10	9	2031	1869	3594	2070
Geospecies	Geospecies	94	88	1550	1201	-	-

USC

USC Viterbi School of Engineering

Datasets: http://www.isi.edu/integration/data/LinkedData



USC

USC Viterbi School of Engineering

Related Work

- Euzenat et al. – Ontology Matching
  - Terminological
  - Structural
  - Semantic
- FCA-Merge, Duckham et al.
  - Use extensional techniques
- GLUE
  - Uses an extensional technique after performing machine learning operations

USC

USC Viterbi School of Engineering

Conclusion

- Our algorithm generates alignments, consisting of conjunctions of restriction classes
  - Extensional approach on Linked Data
  - Use of restriction classes
- Alignments based on the actual data
  - We determine the relationships based on the data
  - Schemas of linked sources can be readily modeled and used
- Algorithm also able to
  - Specialize ontologies where original were rudimentary
  - Find complimentary hierarchy across an ontology

USC

USC Viterbi School of Engineering

Future Work

- How to actually understand these alignments

#	MGI *restriction*	GENEID *restriction*	Relation
17	bio2rdf:subType=Pseudogene	bio2rdf:subType=pseudo	$r_1 = r_2$
18	bio2rdf:subType=Pseudogene & mgi:genomeStart=17	geneid:chromosome=17 & bio2rdf:subType=pseudo	$r_1 = r_2$
19	bio2rdf:chromosomePosition=-1.00 & mgi:genomeStart=4	geneid:chromosome=4 & bio2rdf:subType=pseudo	$r_2 \subset r_1$

- Scalability
  - Pre-procesing of the sources
  - Faster alignment processing

USC