USC Viterbi
School of Engineering

Aligning Unions of Concepts in Ontologies of Geospatial Linked Data

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

USC

USC Viterbi
School of Engineering

INTRODUCTION

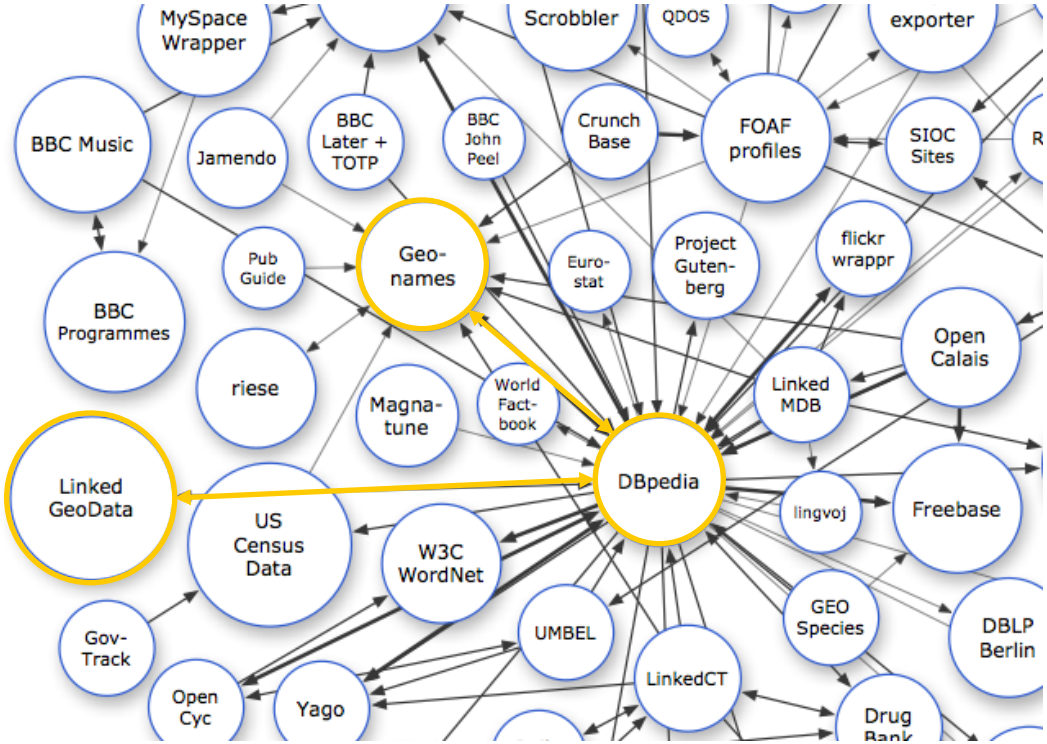
USC

USC Viterbi School of Engineering

Web of Geospatial Linked Data

- Different sources with different schemas
- Equivalent instances in the geospatial domain connected with *owl:sameAs*

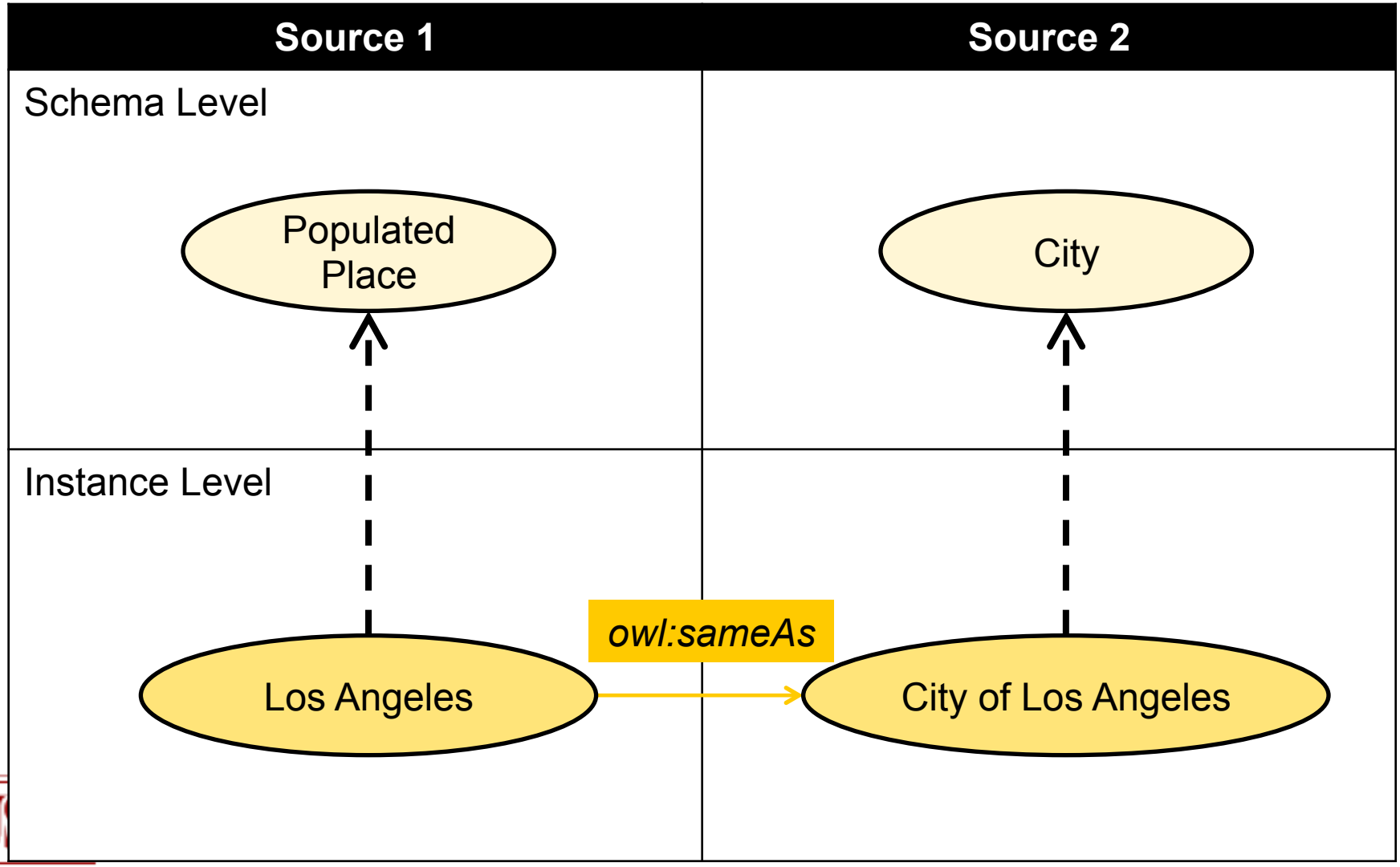
Geospatial Domain



USC

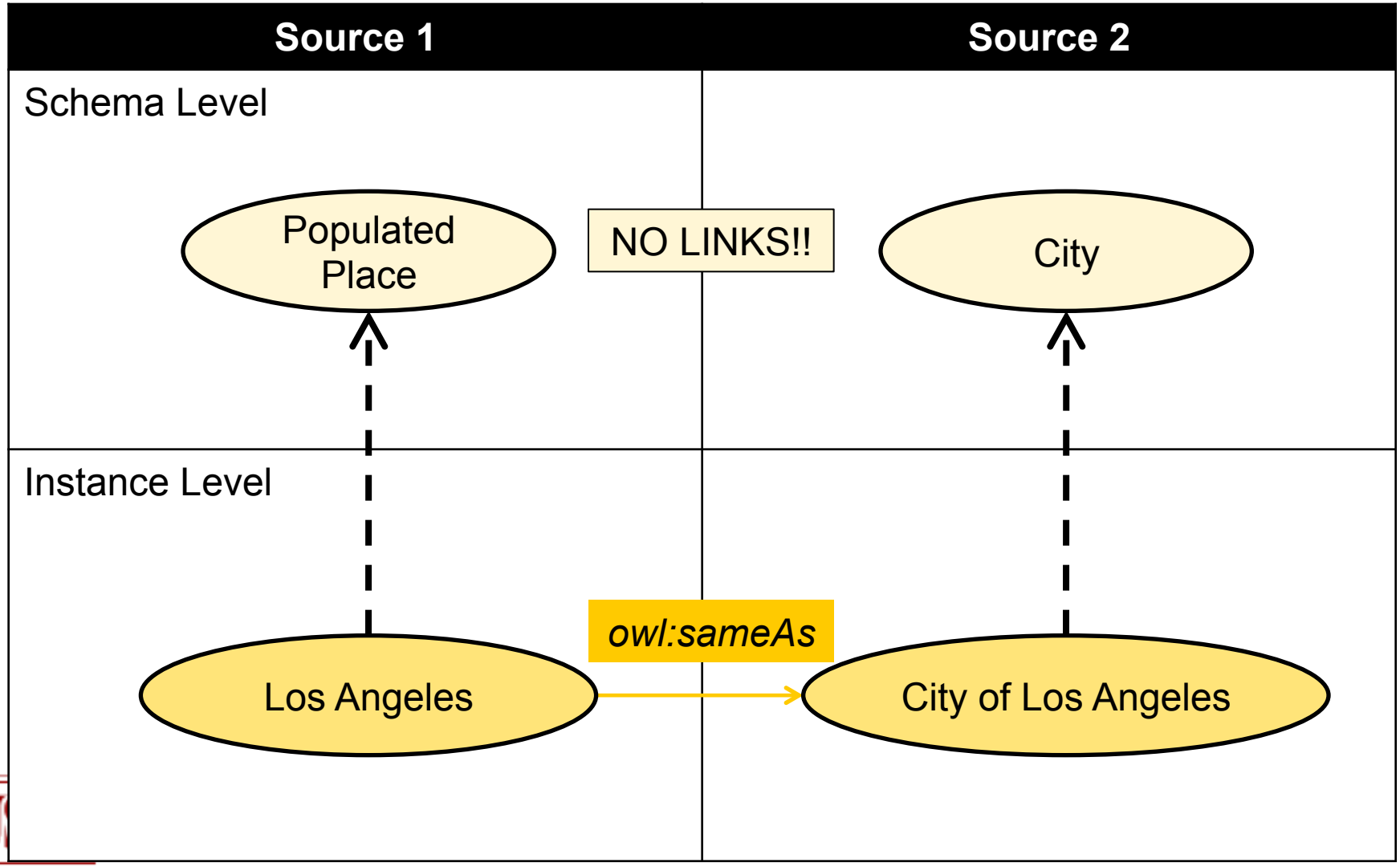
USC Viterbi School of Engineering

Interlinked instances…



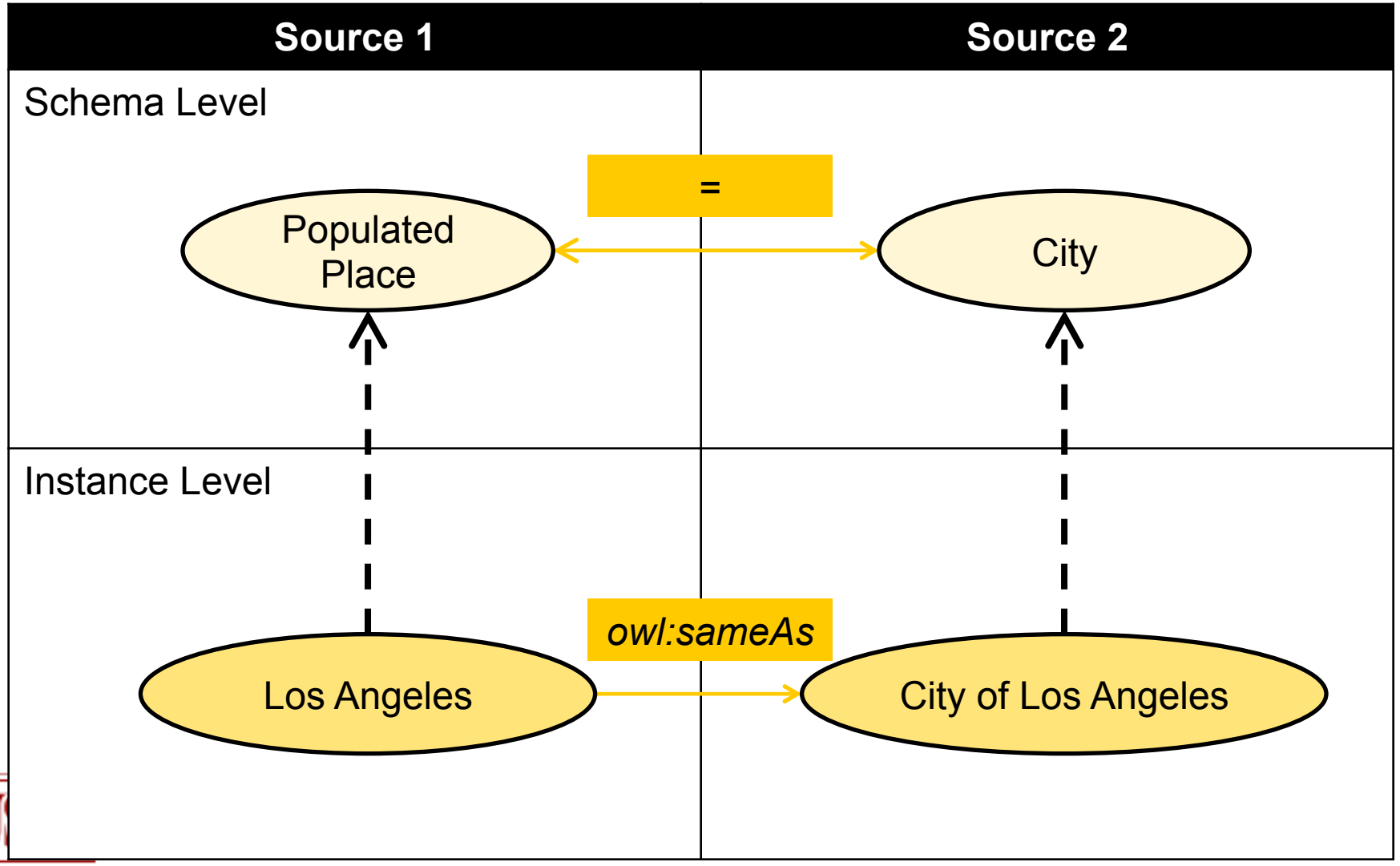
USC Viterbi School of Engineering

…with disjoint schemas



USC Viterbi School of Engineering

Can we find schema alignments?



USC Viterbi
School of Engineering

Previous Work @ ISWC 2010

Linking and Building Ontologies of Linked Data

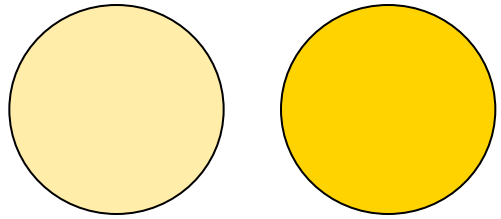
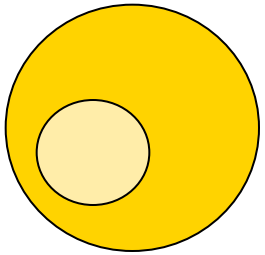
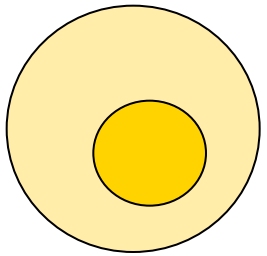
USC

USC Viterbi School of Engineering

Extensional Approach to Ontology Alignment

Represents set of instances belonging to ClassA

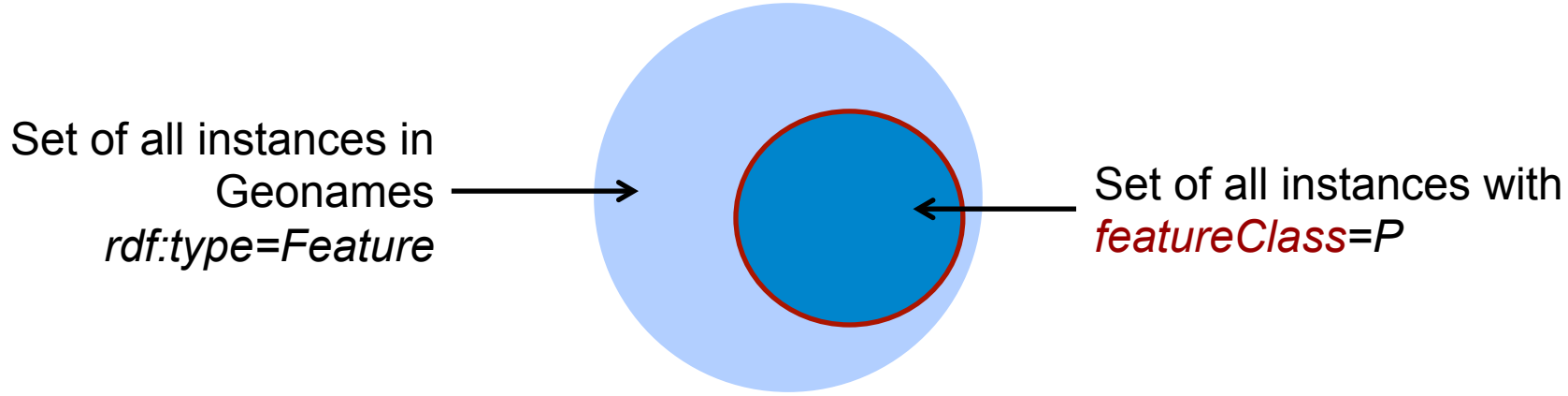
Represents set of instances belonging to ClassB

ClassA is disjoint from ClassB 	ClassA is equivalent to ClassB
ClassA is subset of ClassB 	ClassB is subset of ClassA 

USC

USC Viterbi School of Engineering

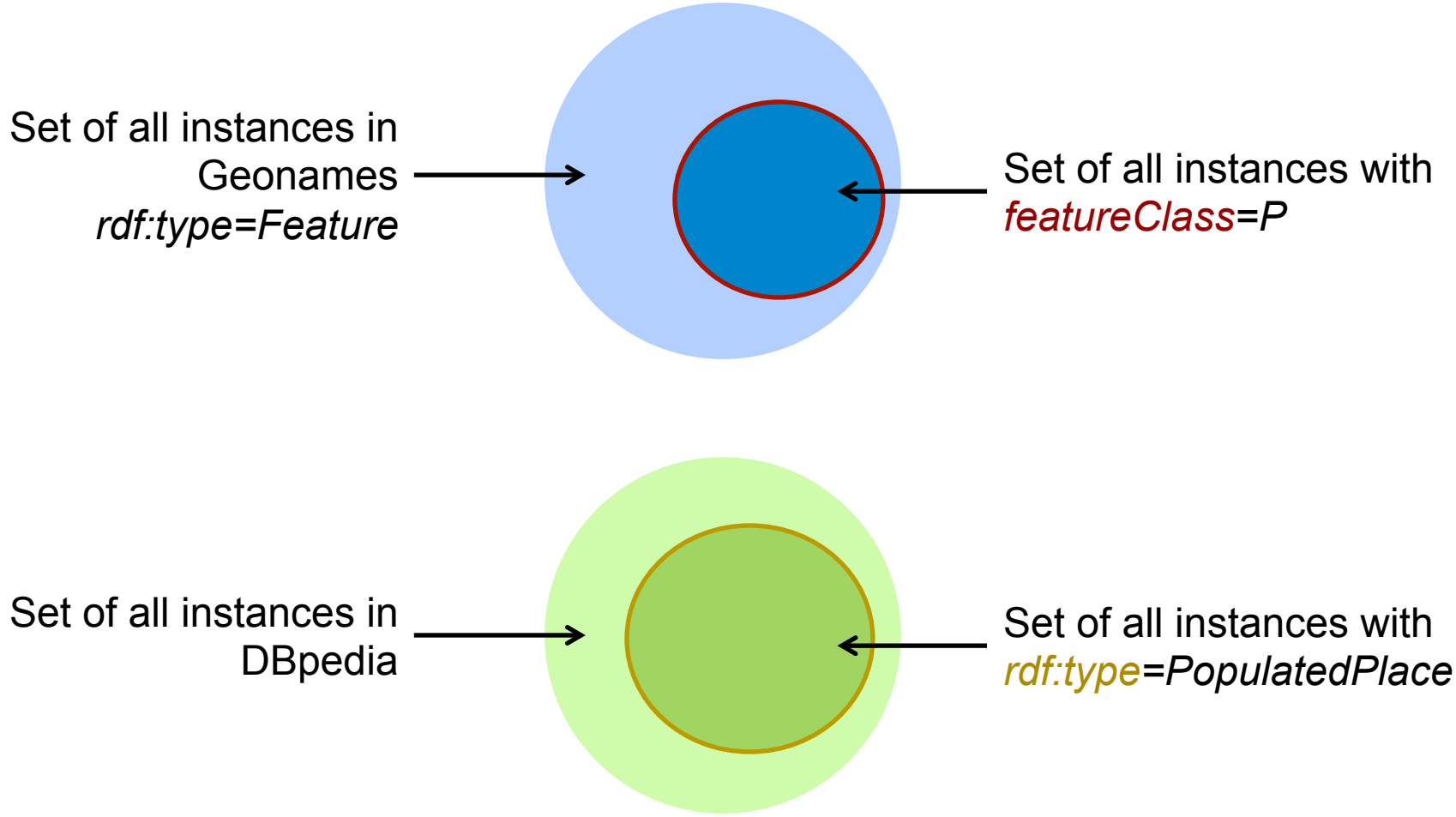
Classes are created extensionally by adding value restrictions on properties



USC

USC Viterbi School of Engineering

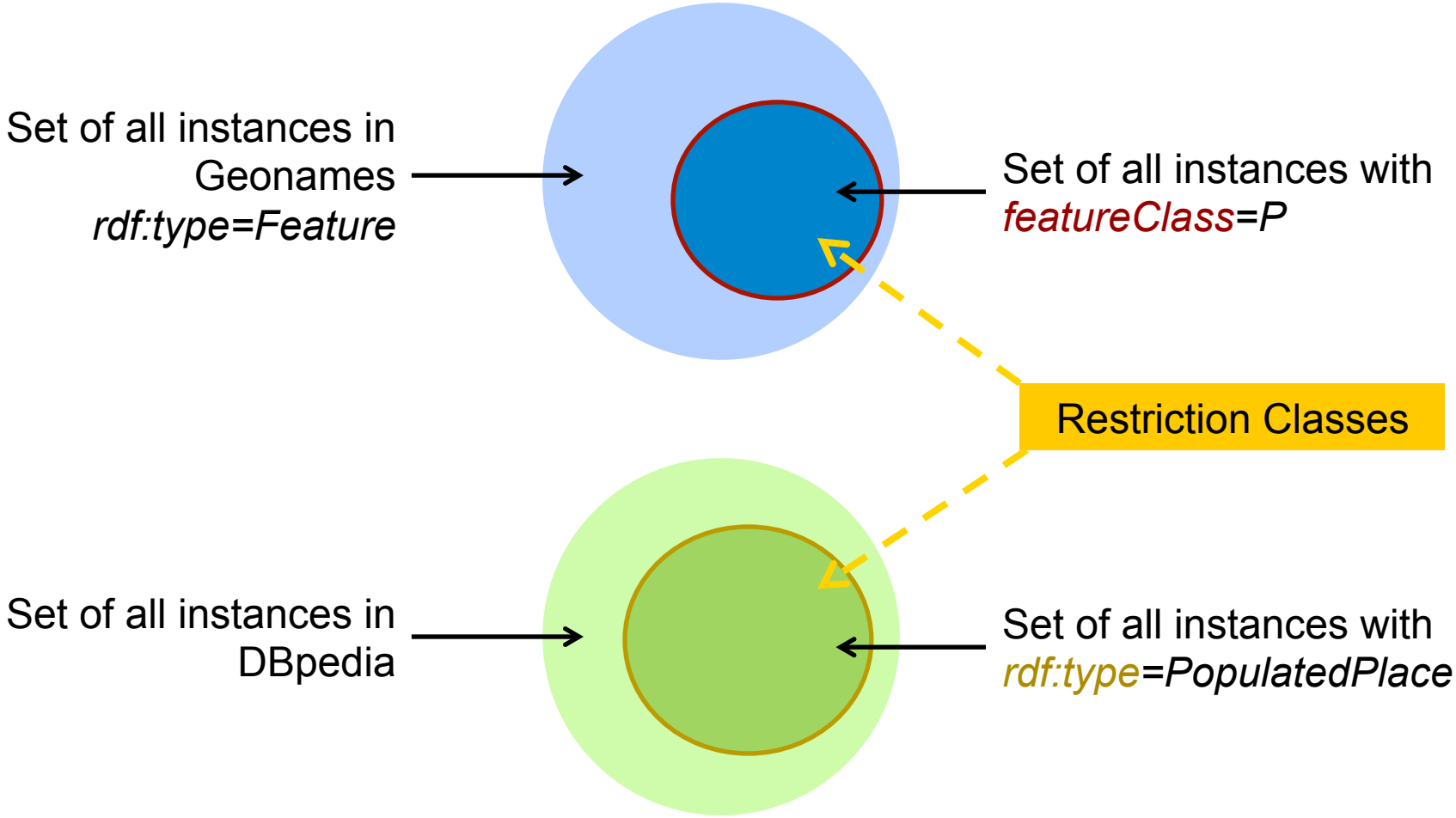
Classes are created extensionally by adding value restrictions on properties



USC

USC Viterbi School of Engineering

Classes are created extensionally by adding value restrictions on properties

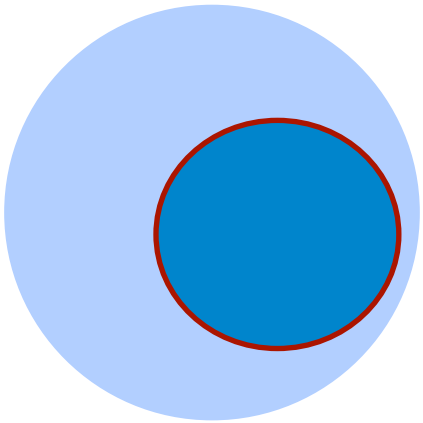


USC

USC Viterbi School of Engineering

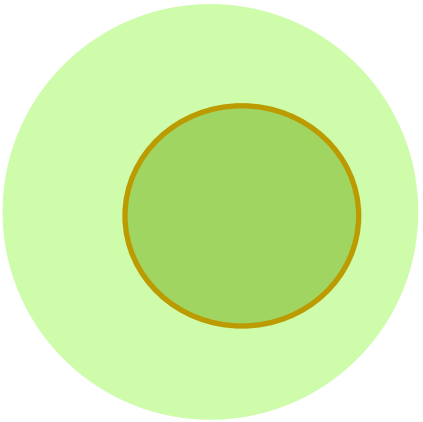
Aligning Restriction Classes Using Extensional Approach

featureClass=*P*



r_1

rdf:type=*PopulatedPlace*

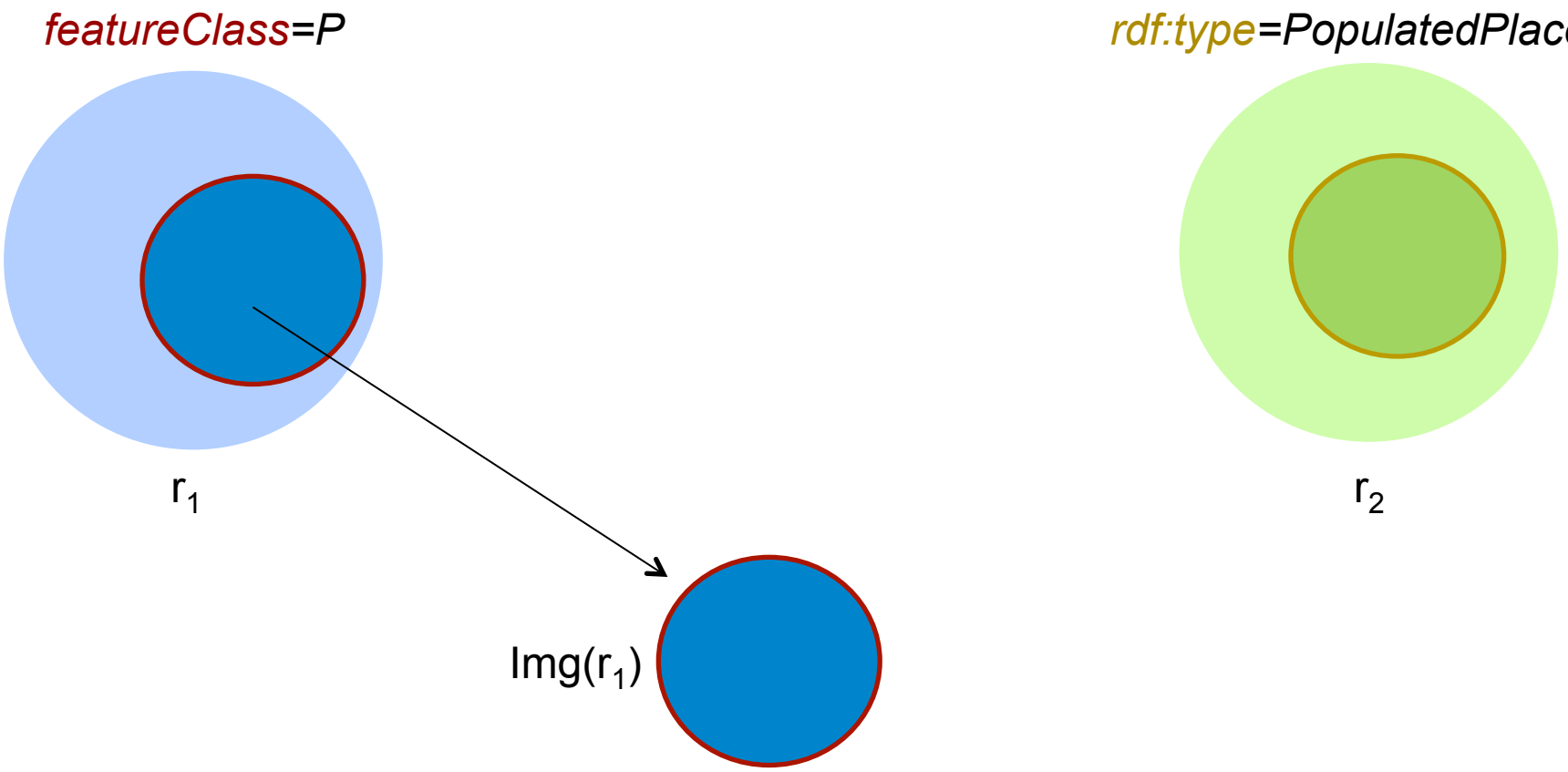


r_2

USC

USC Viterbi School of Engineering

Aligning Restriction Classes Using Extensional Approach

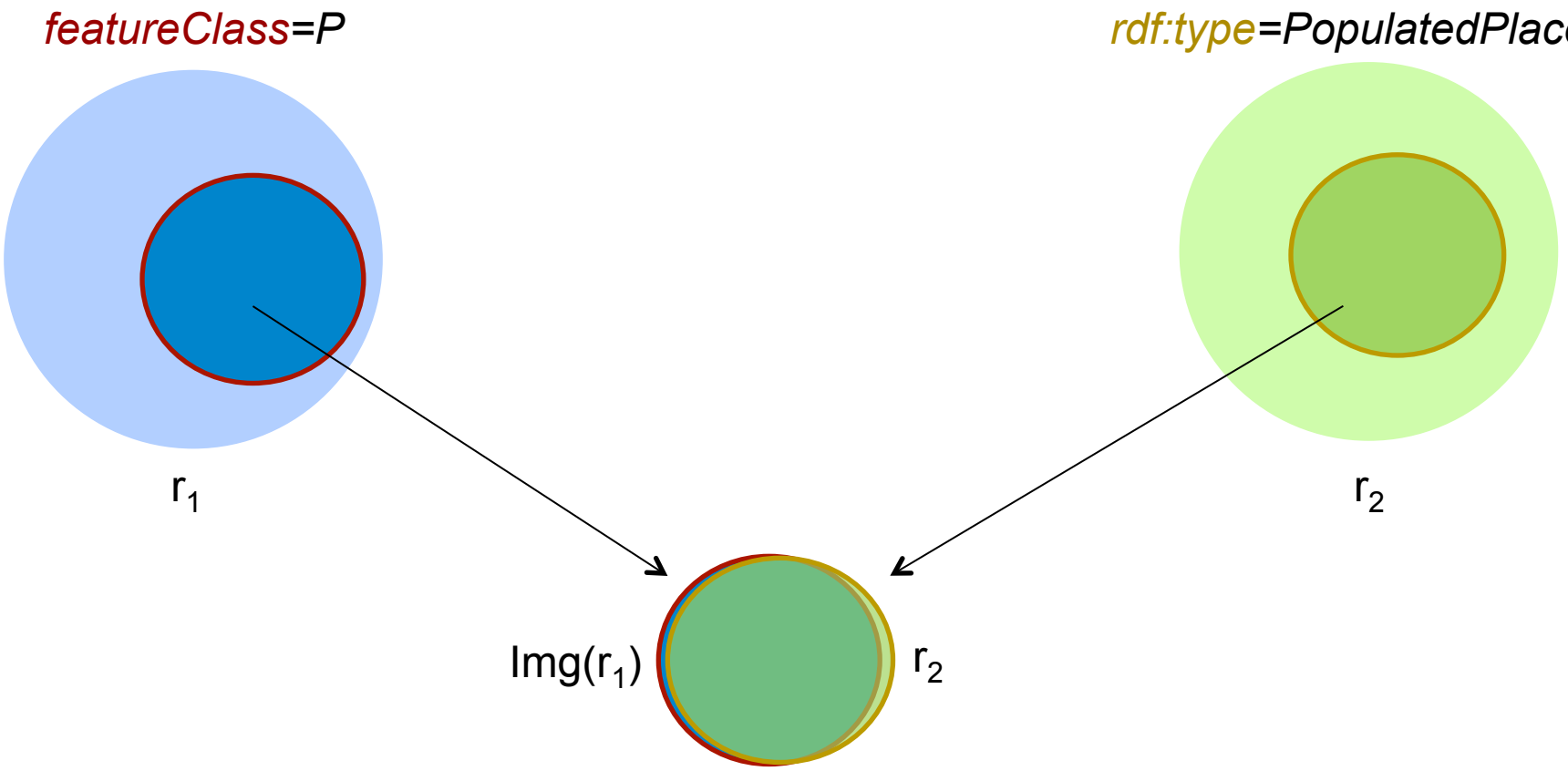


Set of instances from DBpedia that r_1 is linked to

USC

USC Viterbi School of Engineering

Aligning Restriction Classes Using Extensional Approach



USC

USC Viterbi School of Engineering

Extensionally, when are two classes equal?

Represents set of instances belonging to ClassA

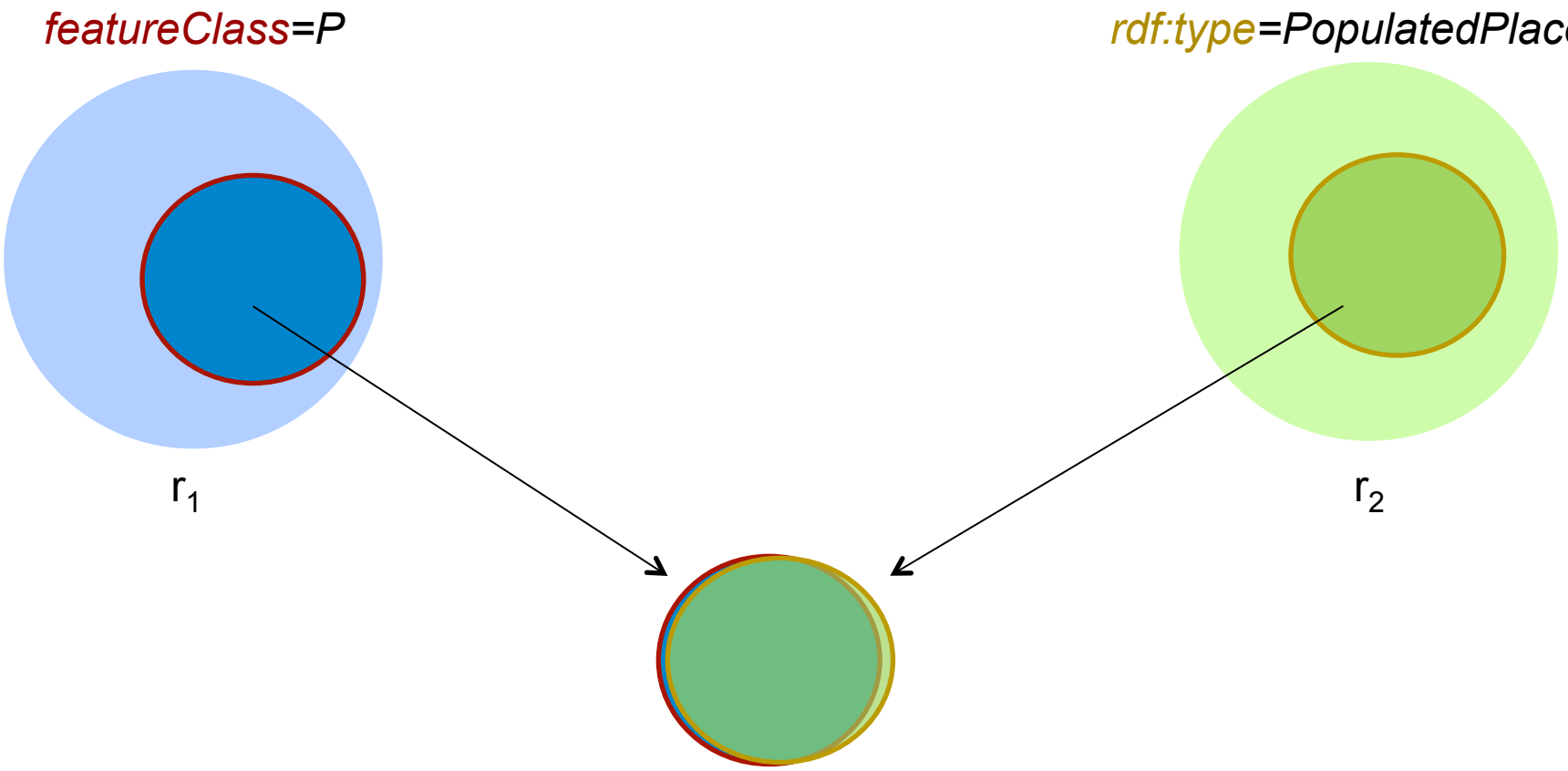
Represents set of instances belonging to ClassB

$$\frac{|\text{ClassA} \cap \text{ClassB}|}{|\text{ClassA}|} = \frac{|\text{ClassA} \cap \text{ClassB}|}{|\text{ClassB}|} = 1$$

USC

USC Viterbi School of Engineering

Aligning Restriction Classes Using Extensional Approach

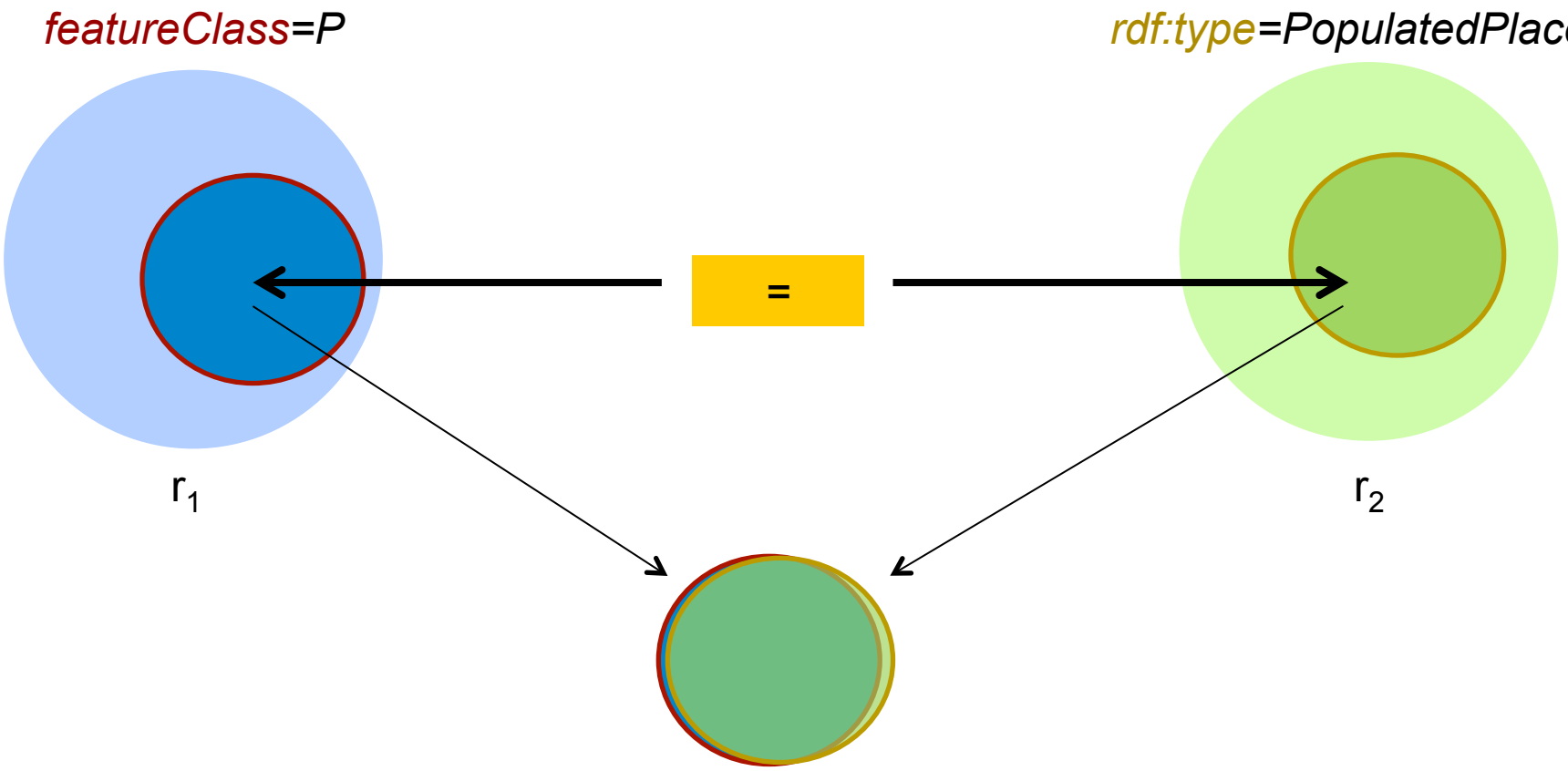


$$\frac{|\,Img(r_1) \cap r_2|}{|\,Img(r_1)\,|} > 0.9 \qquad \frac{|\,Img(r_1) \cap r_2|}{|r_2|} > 0.9$$

USC

USC Viterbi School of Engineering

Aligning Restriction Classes Using Extensional Approach



$$\frac{|\,Img(r_1) \cap r_2|}{|\,Img(r_1)\,|} > 0.9 \qquad \frac{|\,Img(r_1) \cap r_2|}{|r_2|} > 0.9$$

USC

USC Viterbi School of Engineering

Alignments Found in the ISWC'10 Paper

- Algorithm was able to
  - Specialize ontologies where original were rudimentary
  - Find complimentary hierarchy across an ontology
- Alignments based on the actual data
  - reflects the semantics of the sources in practice
- 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 Viterbi School of Engineering

Alignments Found in the ISWC'10 Paper

- Algorithm was able to
  - Specialize ontologies where original were rudimentary
  - Find complimentary hierarchy across an ontology
- Alignments based on the actual data
  - reflects the semantics of the sources in practice
- 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 Viterbi
School of Engineering

Can we use the subset relations to find more meaningful alignments?

USC

USC Viterbi
School of Engineering

TerraCognita Workshop - ISWC 2011

Aligning Unions of Concepts in Ontologies of Linked Geospatial Data

USC Viterbi School of Engineering

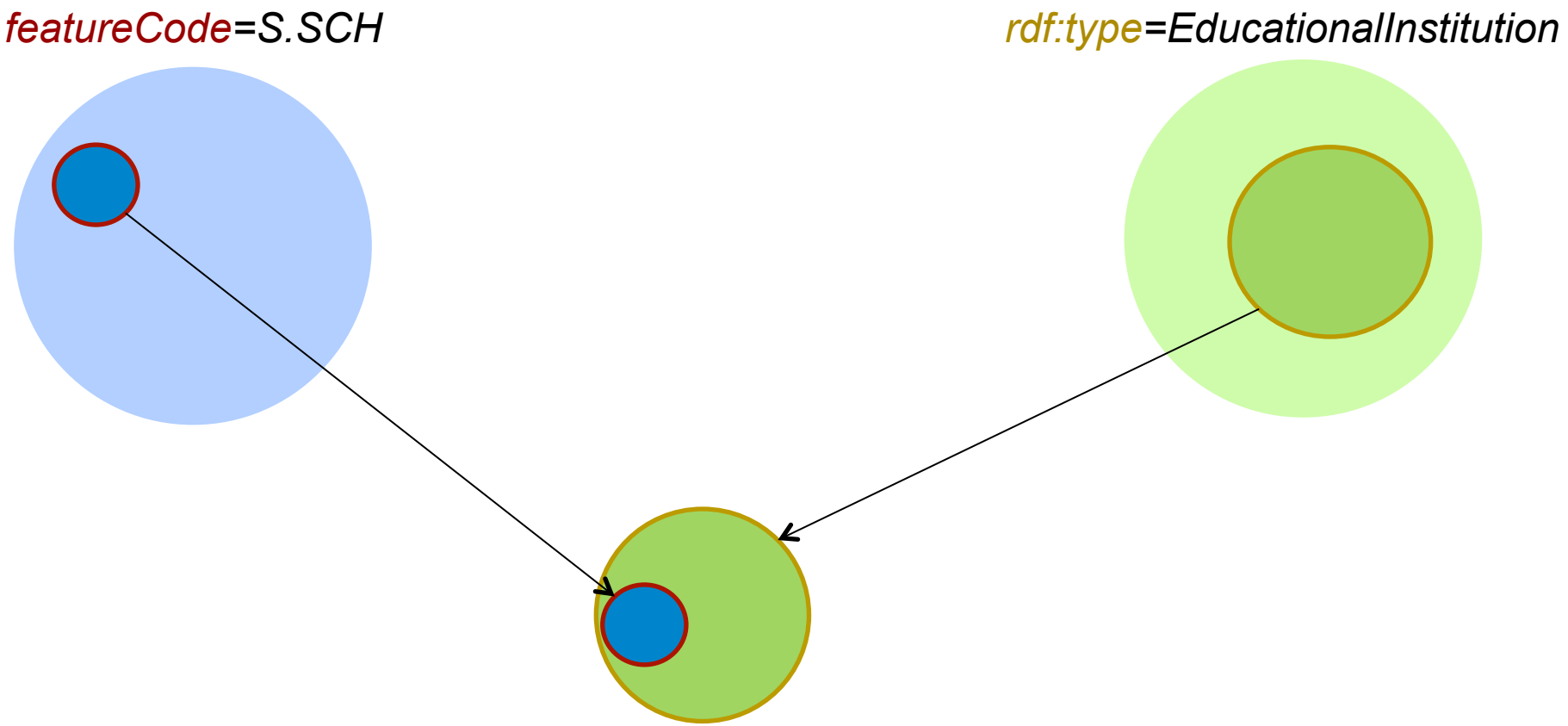
Is there a pattern in the subset relations?

Let's look at 3 of the subset relations we found…

USC

USC Viterbi School of Engineering

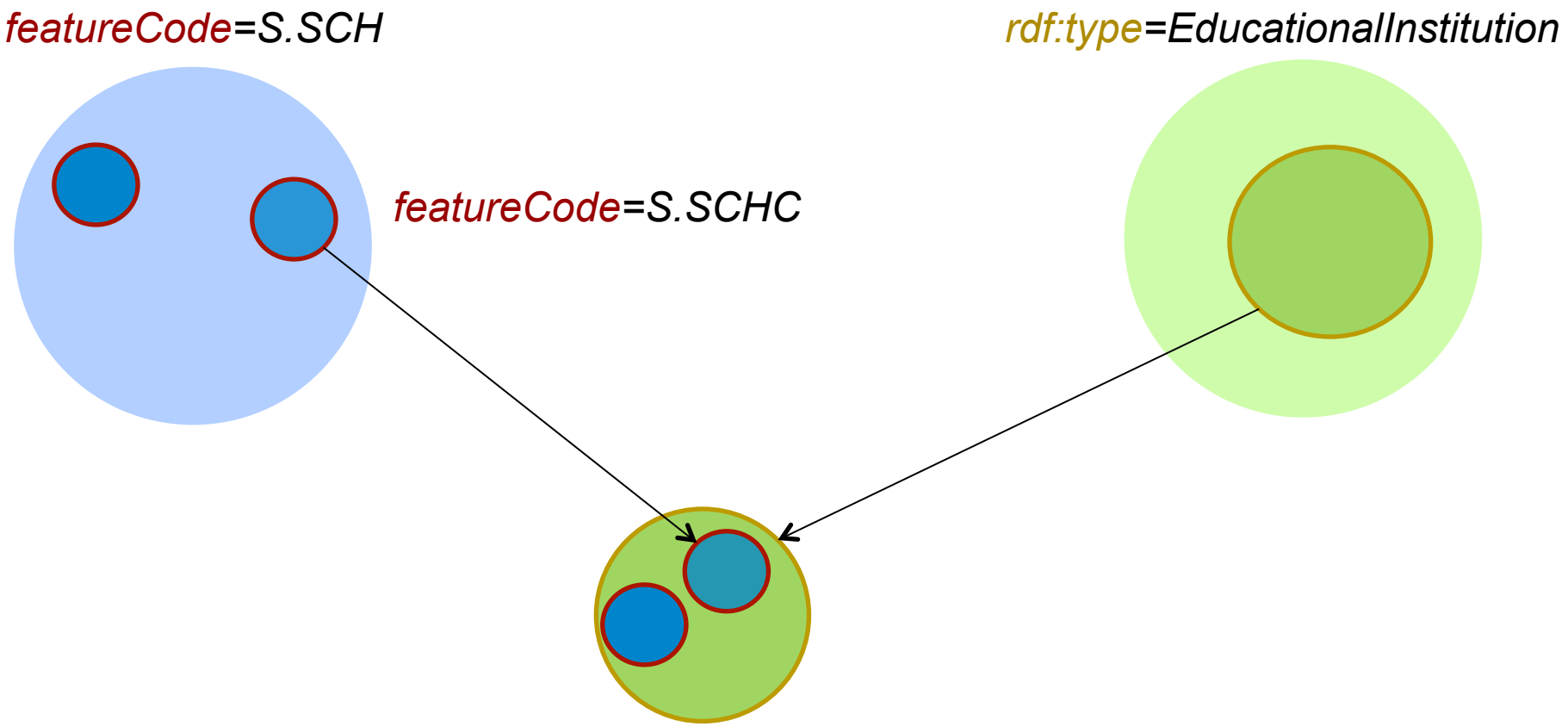
1) Schools in *Geonames* are Educational Institutions in *DBpedia*



USC

USC Viterbi School of Engineering

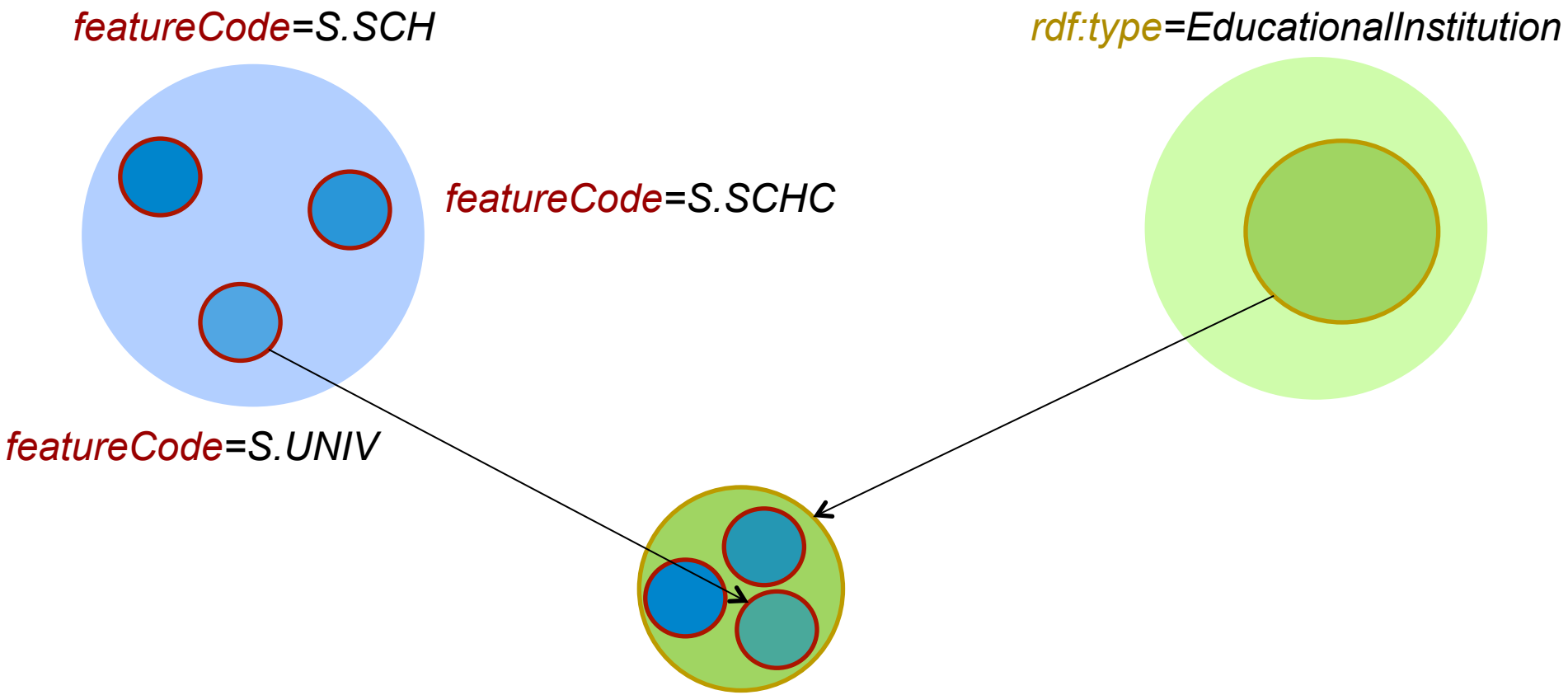
2) Colleges in *Geonames* are Educational Institutions in *DBpedia*



USC

USC Viterbi School of Engineering

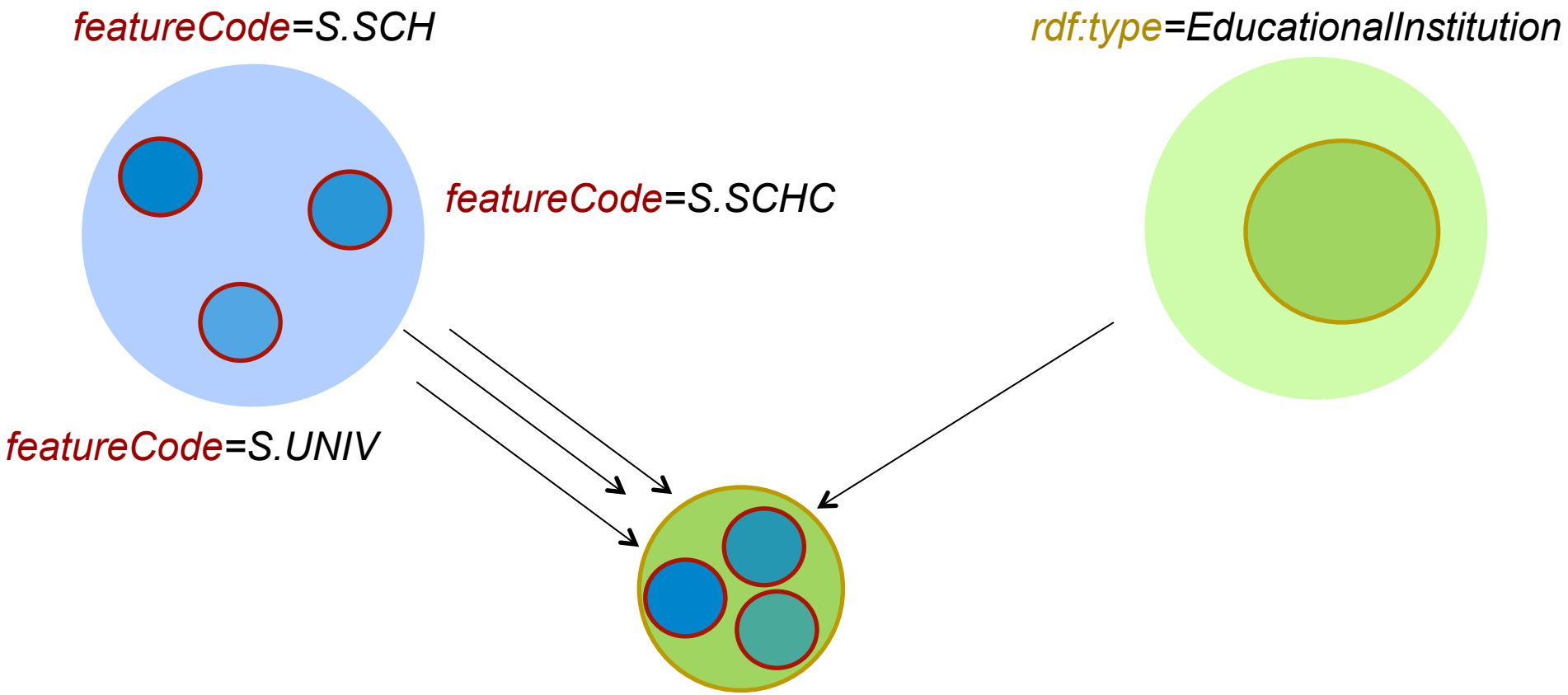
3) Universities in *Geonames* are Educational Institutions in *DBpedia*



USC

USC Viterbi School of Engineering

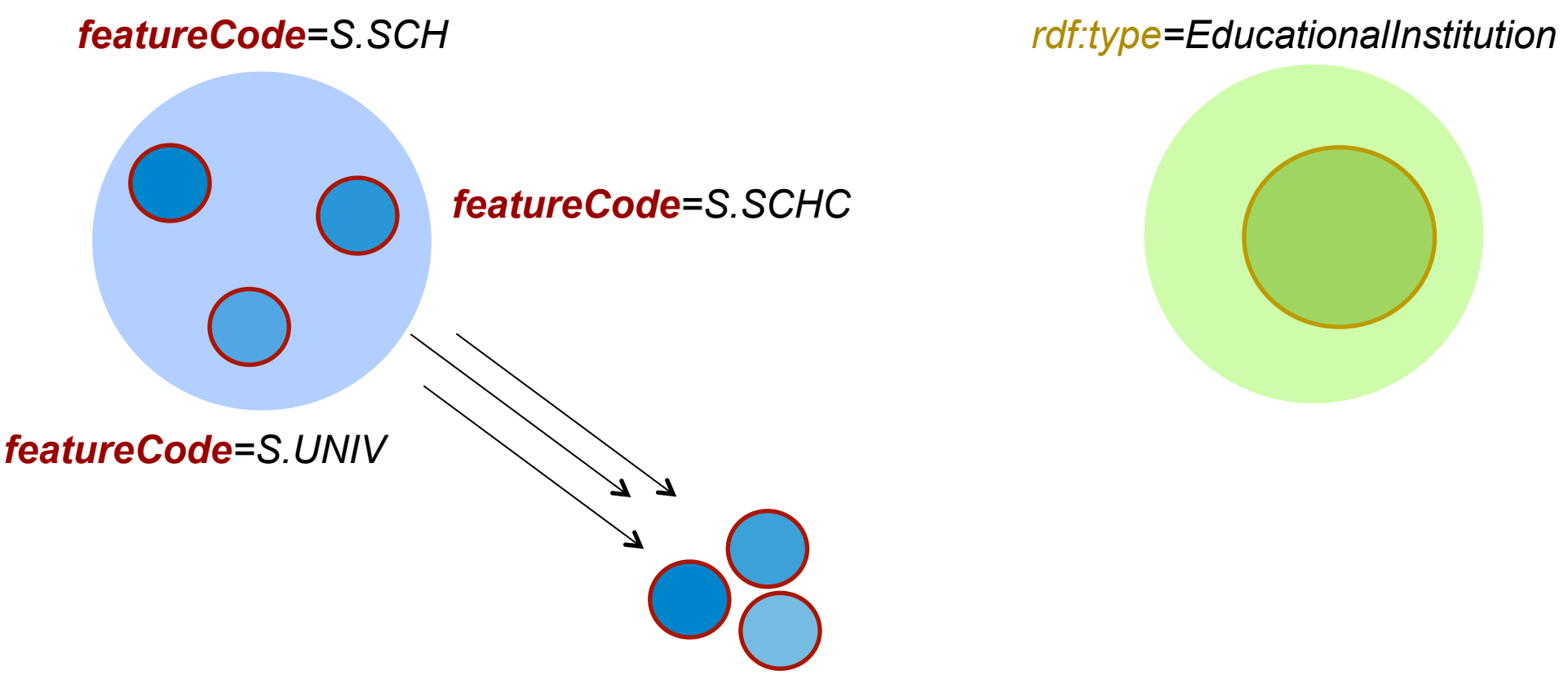
Taken by themselves, the subset relations are not useful



USC

USC Viterbi School of Engineering

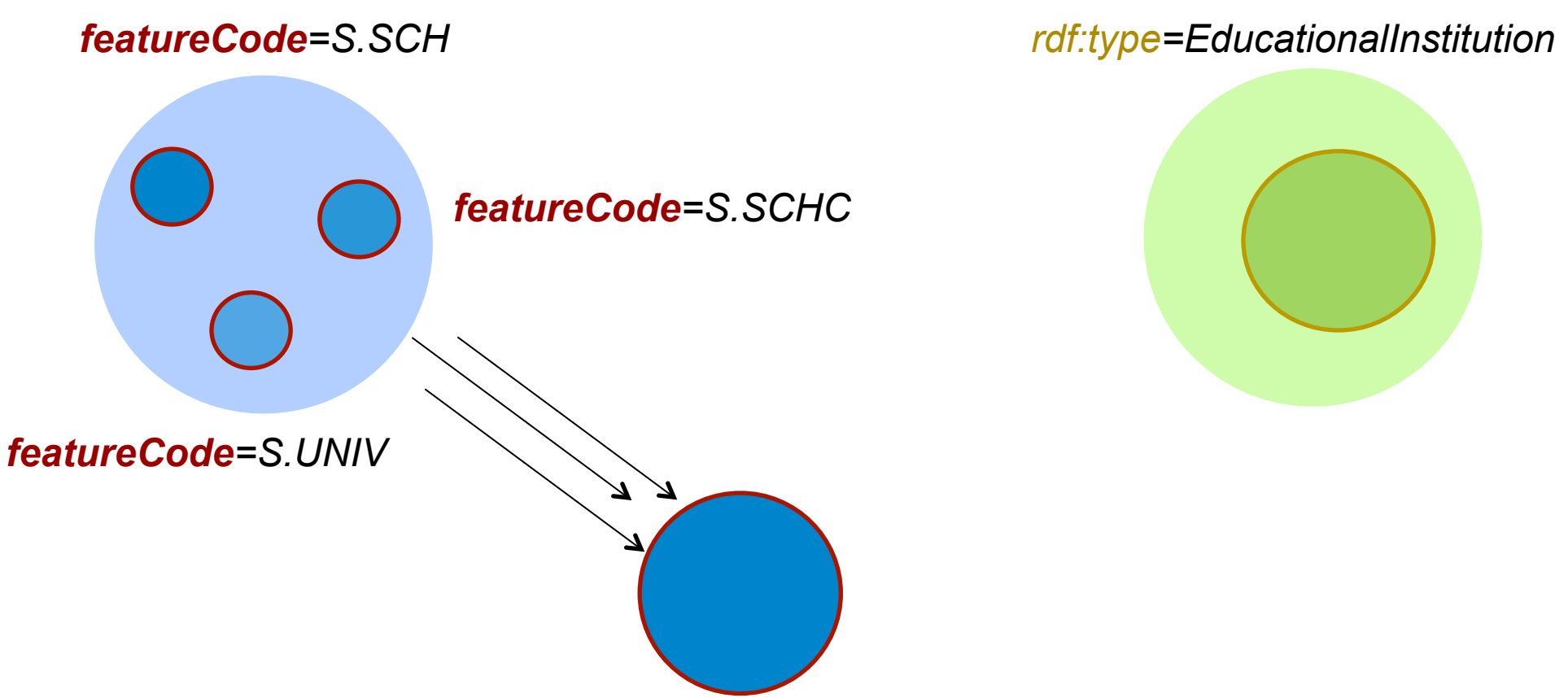
We use the common ***featureCode*** property as a hint…



USC

USC Viterbi School of Engineering

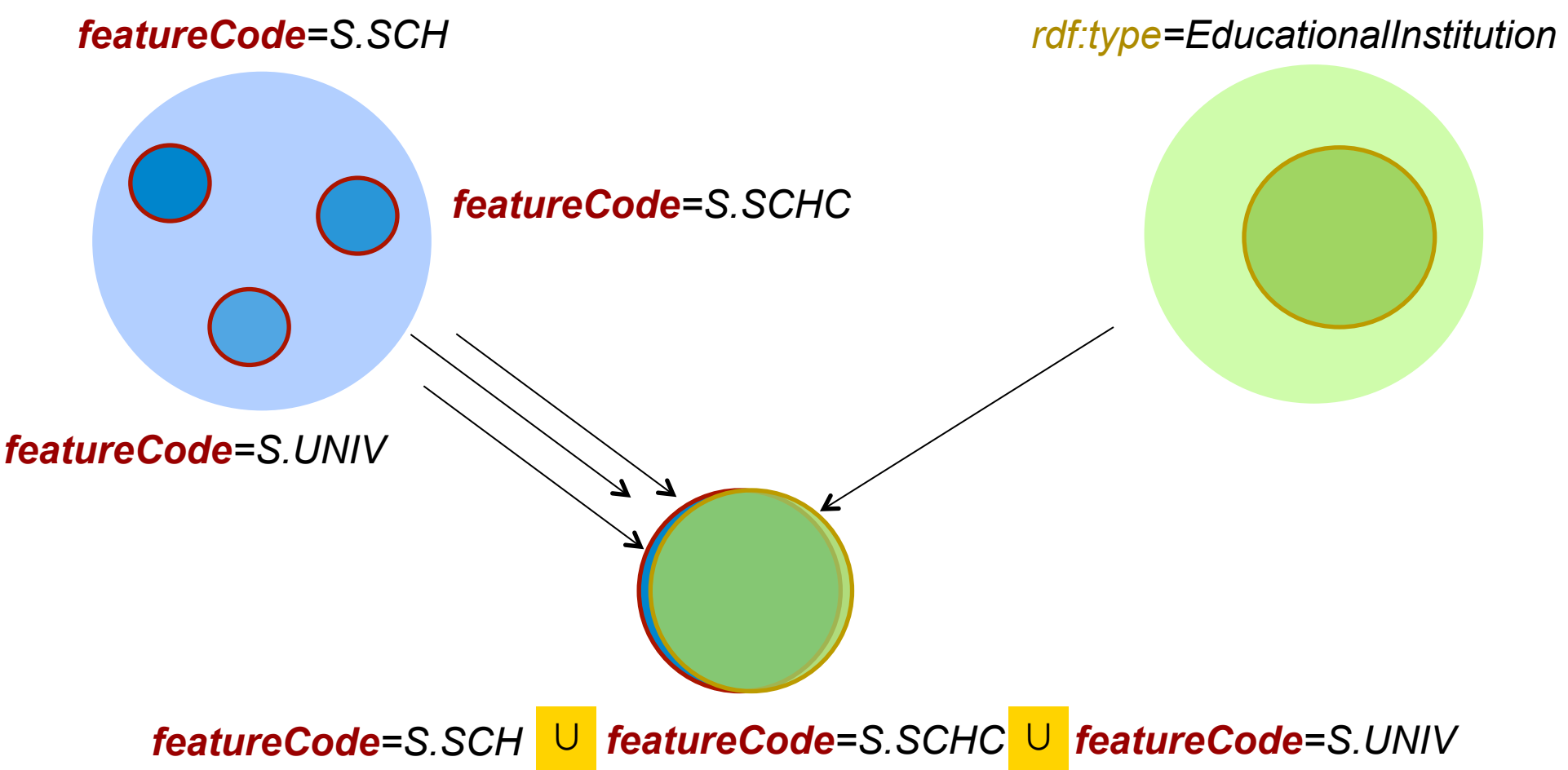
…to form a ***Union*** of Restriction Classes



USC

USC Viterbi School of Engineering

Contribution 1: Find Union Alignments

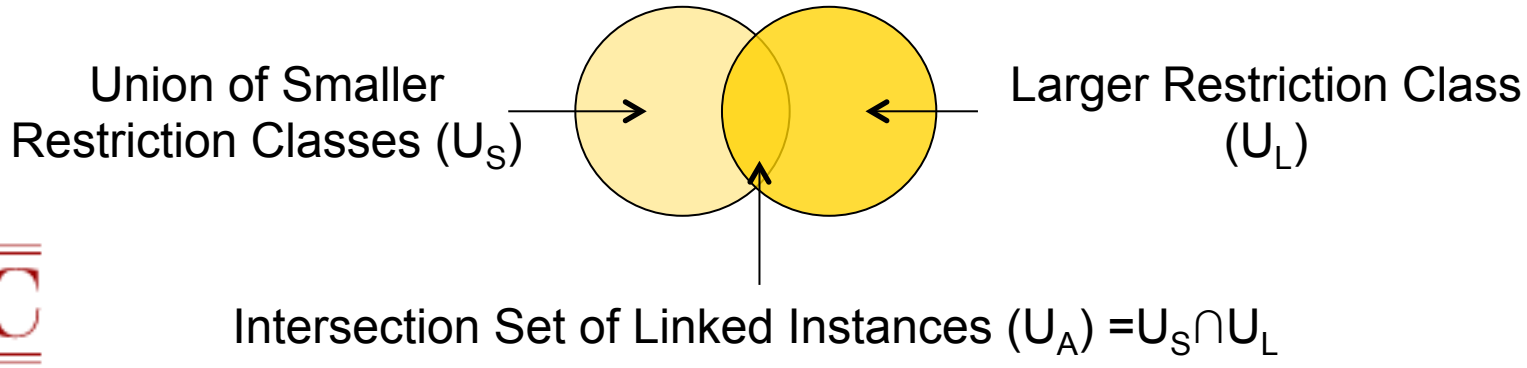


USC

USC Viterbi School of Engineering

Finding Union Alignments: Approach

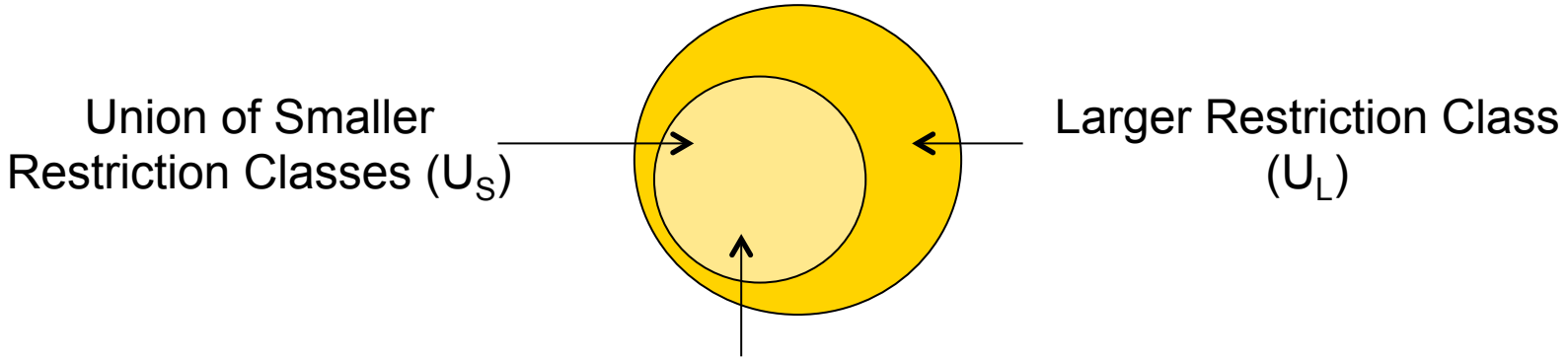
- For all alignments found in the ISWC2010 paper marked as subsets
  1. We group all subset alignments according to the common larger restriction class
  2. We form a *union concept* such that all restriction classes
     - have the same property
     - have a single *property-value pair* each
  3. We then try to match the *union concept* to the larger class
  4. This forms a hypothesis *Union Alignment*



USC

USC Viterbi School of Engineering

Finding Union Alignments: Scoring



$$\frac{|U_A|}{|U_S|} = 1$$ since by definition, all smaller classes are subsets

So, if $\frac{|U_A|}{|U_L|} = 1$, then the larger class U_L is equivalent to U_S

Practically, we use a relaxed subset assumption: $\frac{|U_A|}{|U_S|}, \frac{|U_A|}{|U_L|} > 0.9$

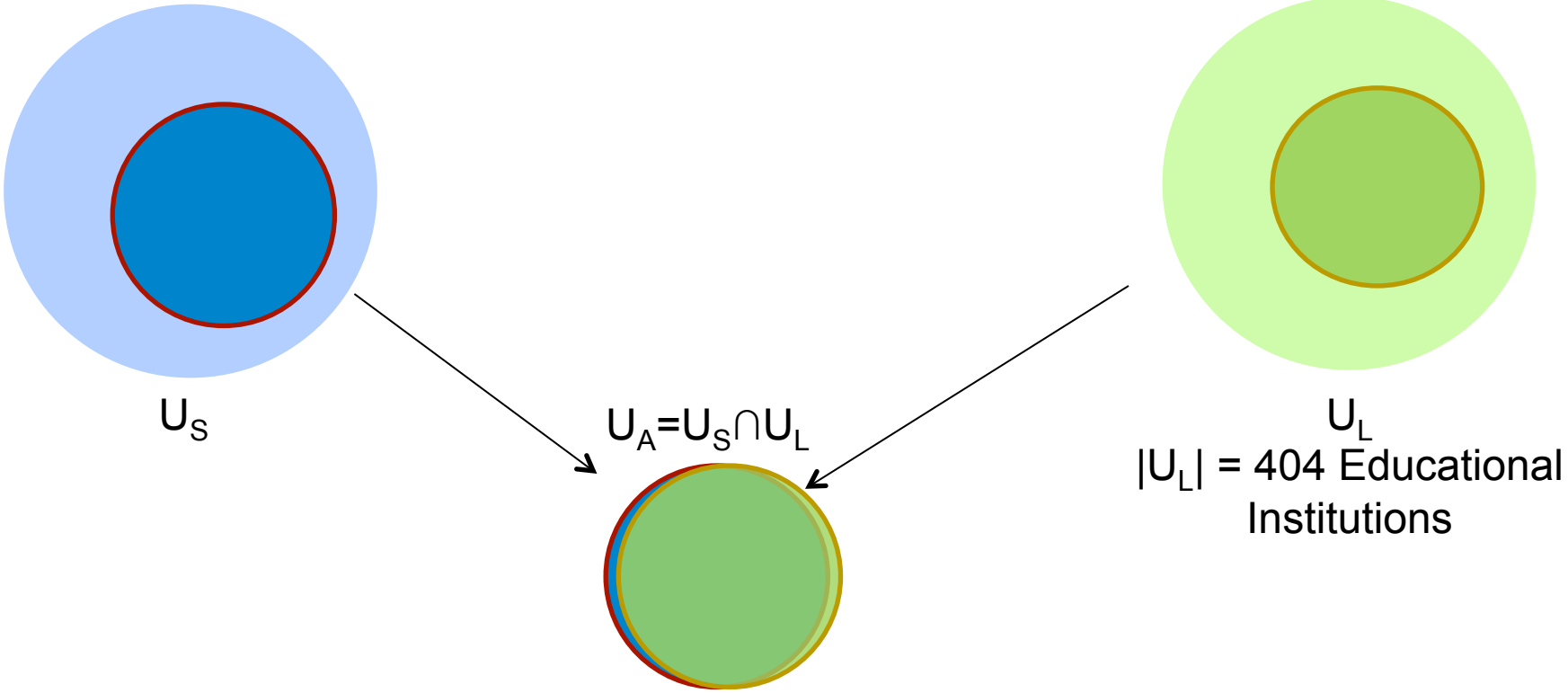
USC

USC Viterbi School of Engineering

Contribution 1: Find Union Alignments

featureCode=*{S.SCH, S.SCHC, S.UNIV}*

rdf:type=*EducationalInstitution*



$$\frac{|U_A|}{|U_S|} > 0.9 \qquad \frac{|U_A|}{|U_L|} = \frac{396}{404} = 0.98 > 0.9$$

USC

USC Viterbi
School of Engineering

What are the other 8 Educational Institutions?

USC

USC Viterbi School of Engineering

Contribution 2: Find Outliers / Discrepancies

- We are also able to point out where the instances that disagree with the alignment lie
- These instances were not part of the alignment because
  - Their restriction class was not a subset (P'<0.9)
  - Some of these instances are
    - Linked Incorrectly with *owl:sameAs*
    - Assigned wrong value during RDF generation*
    - Common in both sets (could be debatable)
    - Did not have a minimum support size of 2 instances (set with 1 instance cannot be relied on)
- Outliers help in understanding discrepancies in the Linked Data

USC

*See Country of http://dbpedia.org/page/Skegness

USC Viterbi School of Engineering

What are the other 8 Educational Institutions?

- 1 with *featureCode*=S.HSP (Hostpitals)
  - There are 31 instances with S.HSP because of which Hospitals are not subsets
- 3 with *featureCode*=S.BLDG (Buildings)
- 1 with *featureCode*=S.EST (Establishment)
- 1 with *featureCode*=S.LIBR (Library)
- 1 with *featureCode*=S.MUS (Museum)

- 1 doesn't have a *featureCode* property

USC

USC Viterbi
School of Engineering

RESULTS

USC

USC Viterbi School of Engineering

Results: *Geonames-DBpedia*

Larger class from *DBpedia* and union of smaller classes from *Geonames*

#	***Sub-group*** $\{p_1, v_1, p_2\}$	***List***(v_2)	$R'_U = \frac{\lvert U_A \rvert}{\lvert U_L \rvert}$	$\lvert U_A \rvert$	$\lvert U_L \rvert$	***Outliers***	***# Explained Instances***
1	{*rdf:type*, *dbpedia:EducationalInstitution*, *geonames:featureCode*}	S.SCH, S.SCHC, S.UNIV	0.9801	396	404	S.BLDG (3/122), S.EST (1/13), S.LIBR (1/7), S.HSP (1/31), S.MUS (1/43)	403
2	{*dbpedia:country*, *dbpedia:Spain*, *geonames:countryCode*}	ES	0.9997	3917	3918	IT (1/7635)	3918
3	{*dbpedia:region*, *dbpedia:Basse-Normandie*, *geonames:parentADM2*}	geonames:2989247, geonames:2996268, geonames:3029094	1.0	754	754		754
4	{*rdf:type*, *dbpedia:Airport*, *geonames:featureCode*}	S.AIRB, S.AIRP	0.9924	1981	1996	S.AIRF (9/22), S.FRMT (1/5), S.SCH (1/404), S.STNB (2/5) S.STNM (1/36), T.HLL (1/61)	1996

Larger class from *Geonames* and union of smaller classes from *DBpedia*

#	***Sub-group*** $\{p_1, v_1, p_2\}$	***List***(v_2)	$R'_U = \frac{\lvert U_A \rvert}{\lvert U_L \rvert}$	$\lvert U_A \rvert$	$\lvert U_L \rvert$	***Outliers***	***# Explained Instances***
5	{*geonames:countryCode*, *NL*, *dbpedia:country*}	dbpedia:Netherlands, dbpedia:The_Netherlands, dbpedia:Flag_of_the _Netherlands.svg	0.9802	1939	1978	dbpedia:Kingdom_of _the_Netherlands	1940
6	{*geonames:countryCode*, *JO*, *dbpedia:country*}	dbpedia:Jordan dbpedia:Flag_of_Jordan.svg	0.95	19	20		20

USC Viterbi School of Engineering

Results: *LinkedGeoData-DBpedia*

Larger class from *DBpedia* and union of smaller classes from *LinkedGeoData*

#	***Sub-group*** $\{p_1, v_1, p_2\}$	***List***(v_2)	$R'_U = \frac{\lvert U_A \rvert}{\lvert U_L \rvert}$	$\lvert U_A \rvert$	$\lvert U_L \rvert$	***Outliers***	***# Explained Instances***
7	{*dbpedia:bundesland*, *Saarland*, *lgd:OpenGeoDBLicensePlateNumber*}	HOM, IGB, MZG, NK, SB, SLS, VK, WND	0.93	46	49		46
8	{*rdf:type*, *dbpedia:EducationalInstitution*, *rdf:type*}	lgd:Amenity, lgd:K2543, lgd:School, lgd:University, lgd:WaterTower	0.9901	2609	2610		2609

Larger class from *LinkedGeoData* and union of smaller classes from *DBpedia*

#	***Sub-group*** $\{p_1, v_1, p_2\}$	***List***(v_2)	$R'_U = \frac{\lvert U_A \rvert}{\lvert U_L \rvert}$	$\lvert U_A \rvert$	$\lvert U_L \rvert$	***Outliers***	***# Explained Instances***
9	{*lgd:gnisST_alpha*, *NJ*, *dbpedia:subdivisionName*}	Atlantic, Burlington, Cape May, Hudson, Hunterdon, Monmoth, New Jersey, Ocean, Passaic	1.0	214	214		214
10	{*rdf:type*, *lgd:Waterway*, *rdf:type*	dbpedia:Stream, dbpedia:River	0.97	33	34	dbpedia:Place(1/94989)	34

USC

USC Viterbi School of Engineering

Results: Summary

We find a total of 6595 Union Alignments

Source 1	Source 2	Larger Class from Source 1	Larger Class from Source 2	Total number of Union Alignments found
Geonames	DBpedia	434	318	752
LinkedGeoData	DBpedia	2746	3097	5843

Results also available at

http://www.isi.edu/integration/data/UnionAlignments

USC

USC Viterbi School of Engineering

Related Work

- BLOOMS, BLOOMS+ ([4][5] in paper)
  - Linked Open Data ontologies aligned with ‘Proton’
  - Constructs a forest of concepts and computes structural similarity
  - Geonames – Proton has “poor performance” because of small number and vague classes in Geonames
- Volker et al. ([8] in paper)
  - Statistical schema induction
  - Mines associativity rules from intermediate *‘transaction datasets’*
  - Develops OWL2 Axioms
- AgreementMaker [2]
  - Similarity Metrics on labels of classes

USC

USC Viterbi School of Engineering

Conclusion and Future Work

- Conclusion
  - We were able to find *Union Alignments* in the Geospatial Domain
    - Find alignments where no direct equivalence was evident
    - Introduced a disjunction operator to restriction classes
  - We were able to find *Outliers*
    - Help identify inconsistencies in the data
- Future work
  - Our algorithm is not limited to Geospatial domain. We would like to explore other domains
  - Experimental comparison with other approaches
  - Preliminary findings suggest patterns in properties like *geonames:countryCode* and *dbpedia:country*

USC

USC Viterbi
School of Engineering

Any questions?

THANK YOU

USC