USC Viterbi
School of Engineering

Discovering Concept Coverings in Ontologies of Linked Data Sources

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

USC

USC Viterbi
School of Engineering

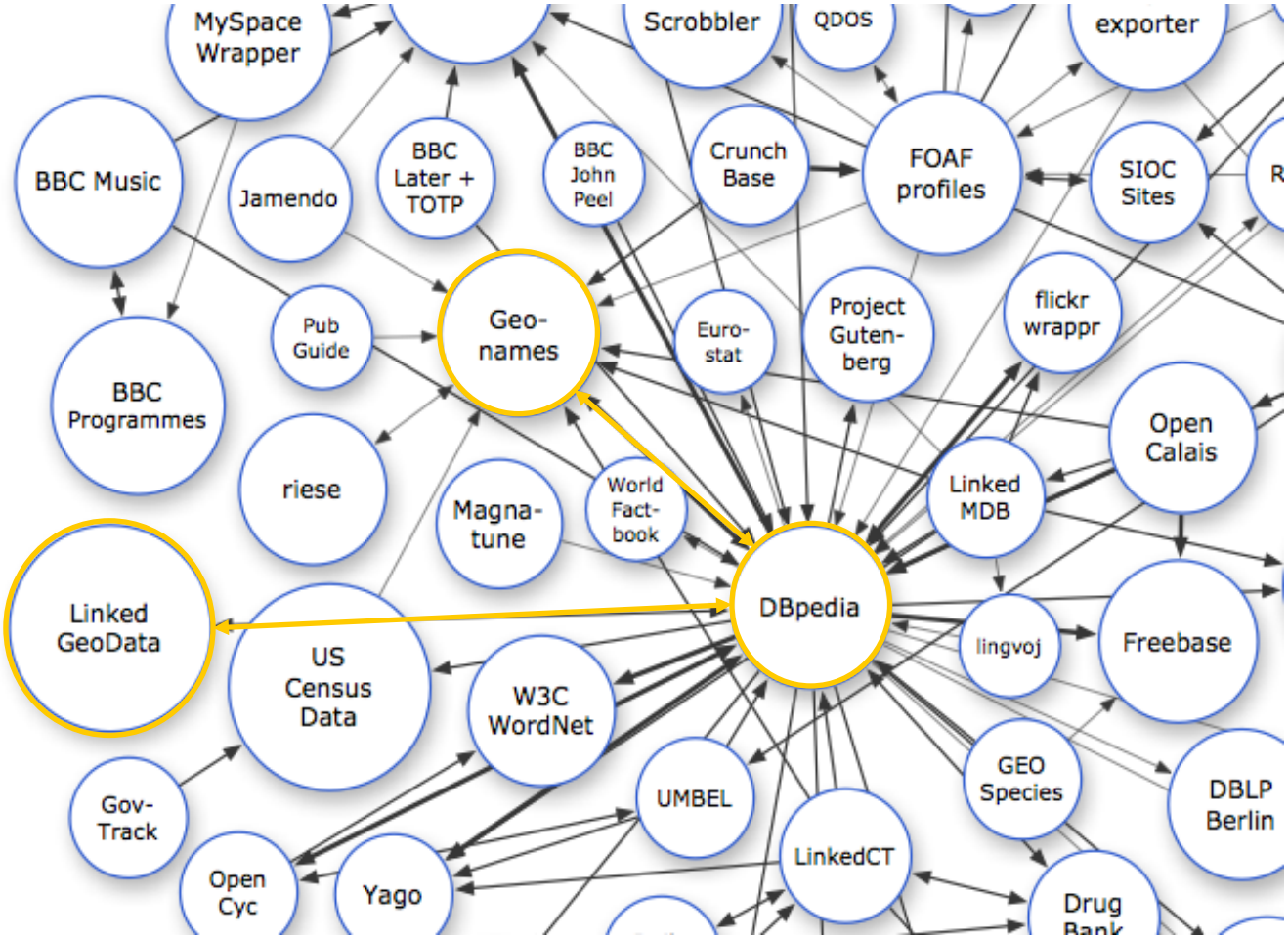
MOTIVATION

USC

USC Viterbi School of Engineering

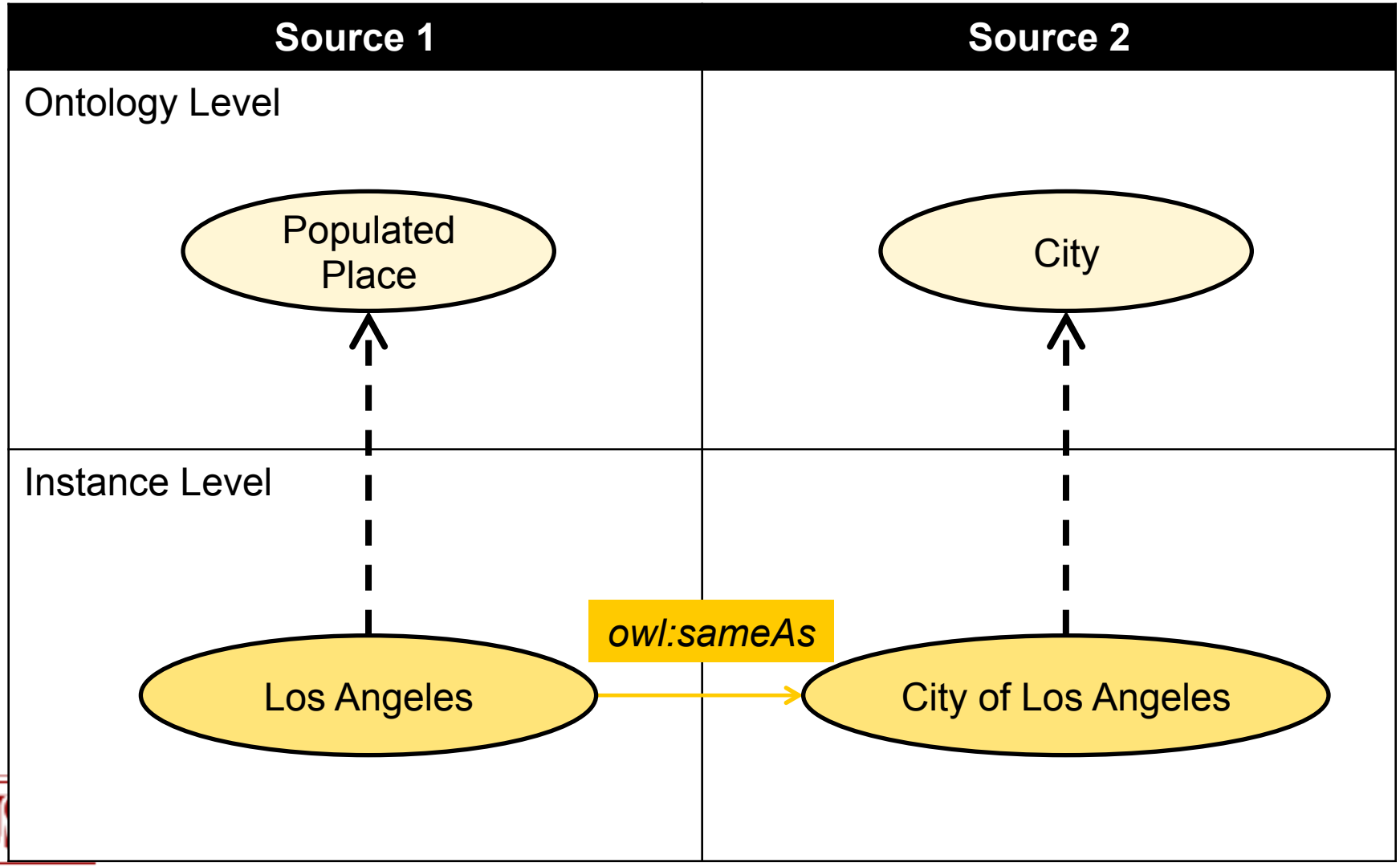
Web of Linked Data

Example: Geospatial Domain

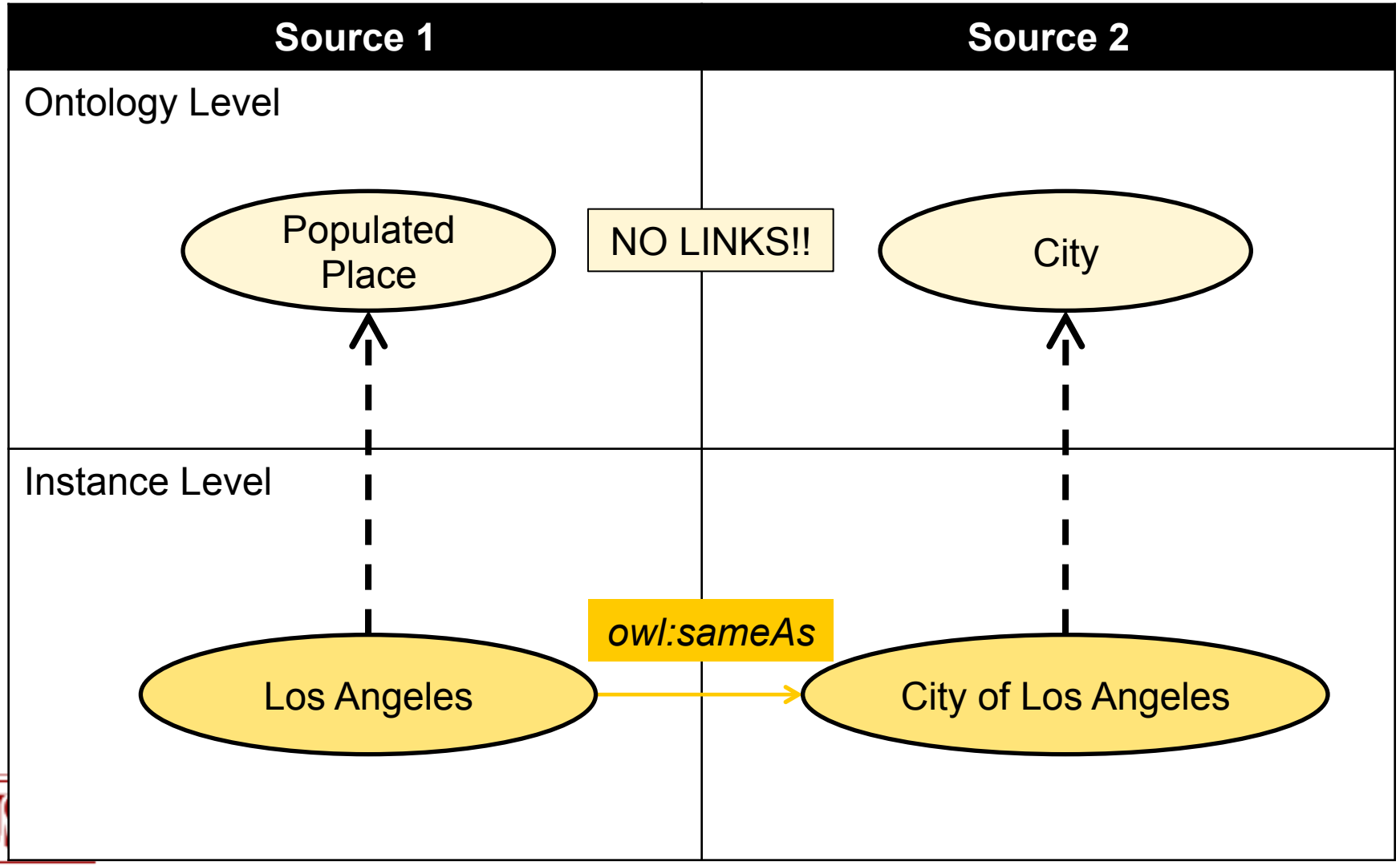


USC

Equivalent instances in the different domains connected with *owl:sameAs*



Links are absent at the ontology level



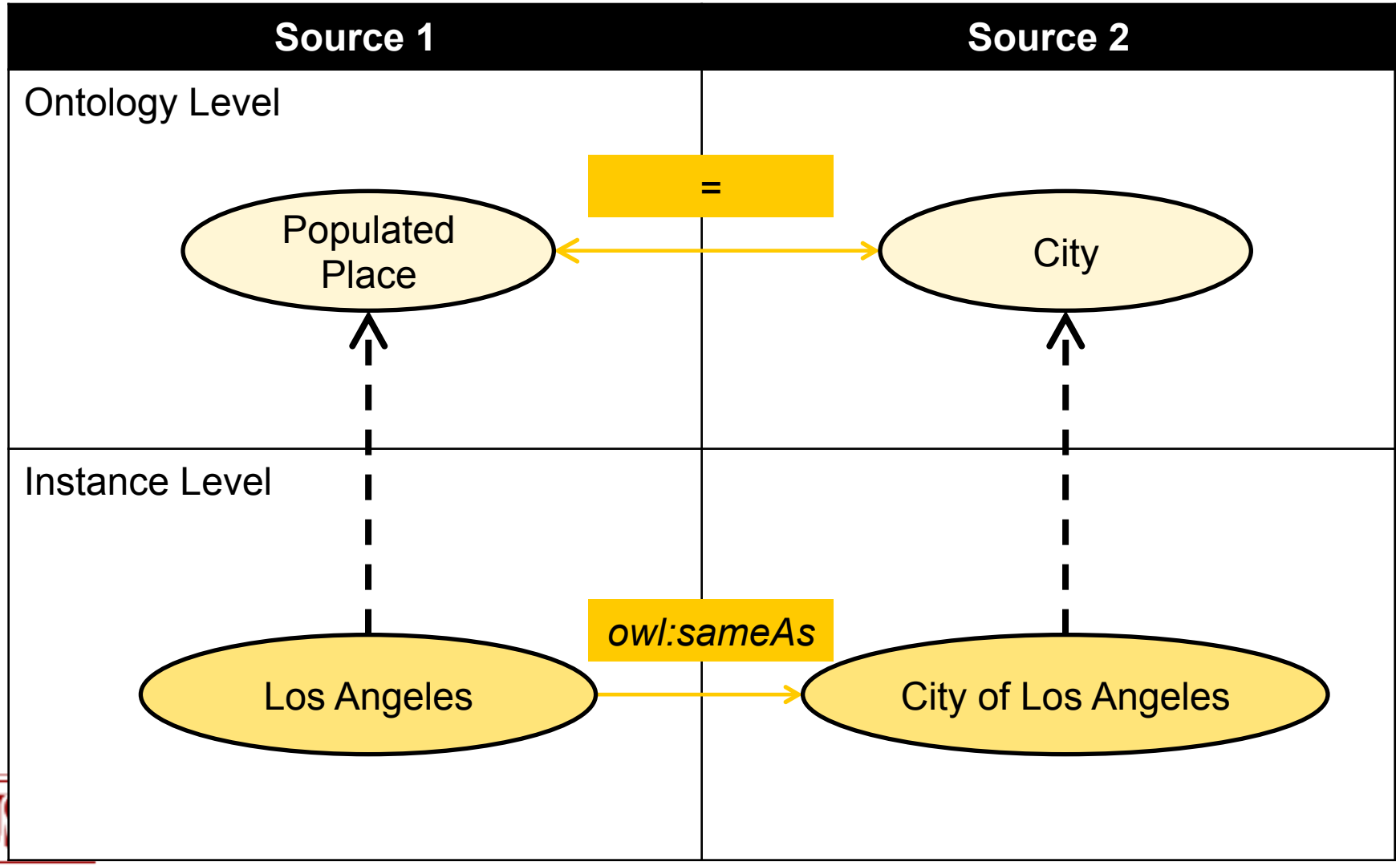
USC Viterbi School of Engineering

Problem: Ontologies are Disconnected

- Only a small number of Ontologies are linked
  - 15 out of the 190 sources: State of the LOD Cloud 2011
- Existing Concepts may not be sufficient for exhaustive set of alignments
  - Linked Data sources reflect RDBMS schemas of sources from which they are derived
    - *DBpedia* has rich ontology
    - *GeoNames* has only one concept (“*geonames:Feature*”)
- Alignments are necessary for the Interoperability goal of the Semantic Web

USC

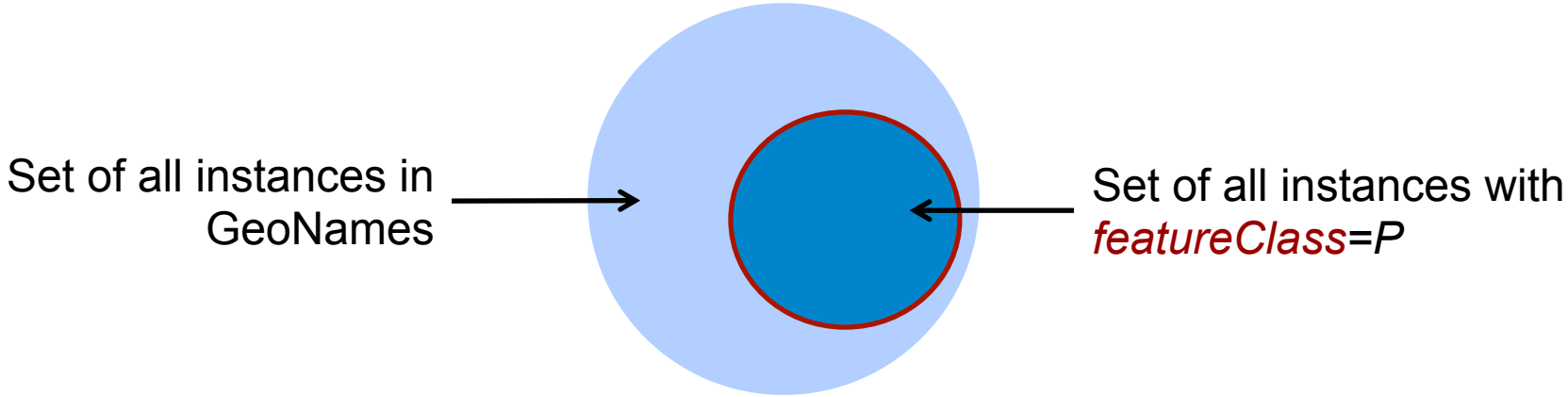
How can we find Ontology alignments?



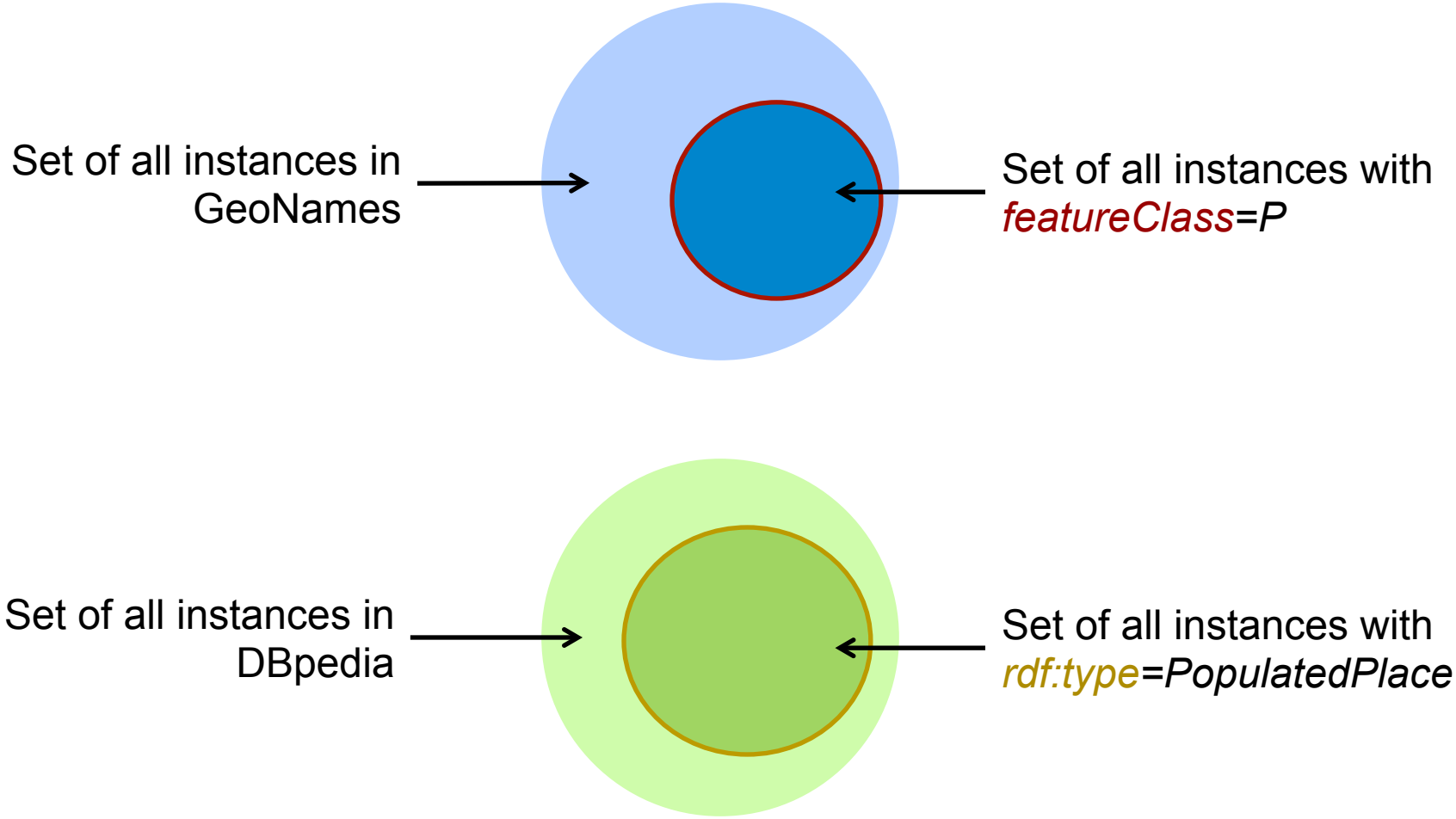
Our Solution

- Generate Alignments Automatically from Linked Data
  - Use equality (e.g. *owl:sameAs*) links between instances in Linked Data as evidence
  - Using Set Containment theory, find alignments between Concepts
- Generate new concepts to find alignments not previously possible with existing concepts
  - Introduce new extensional concepts
    - Value Restrictions in OWL-DL
  - We call these **Restriction Classes**

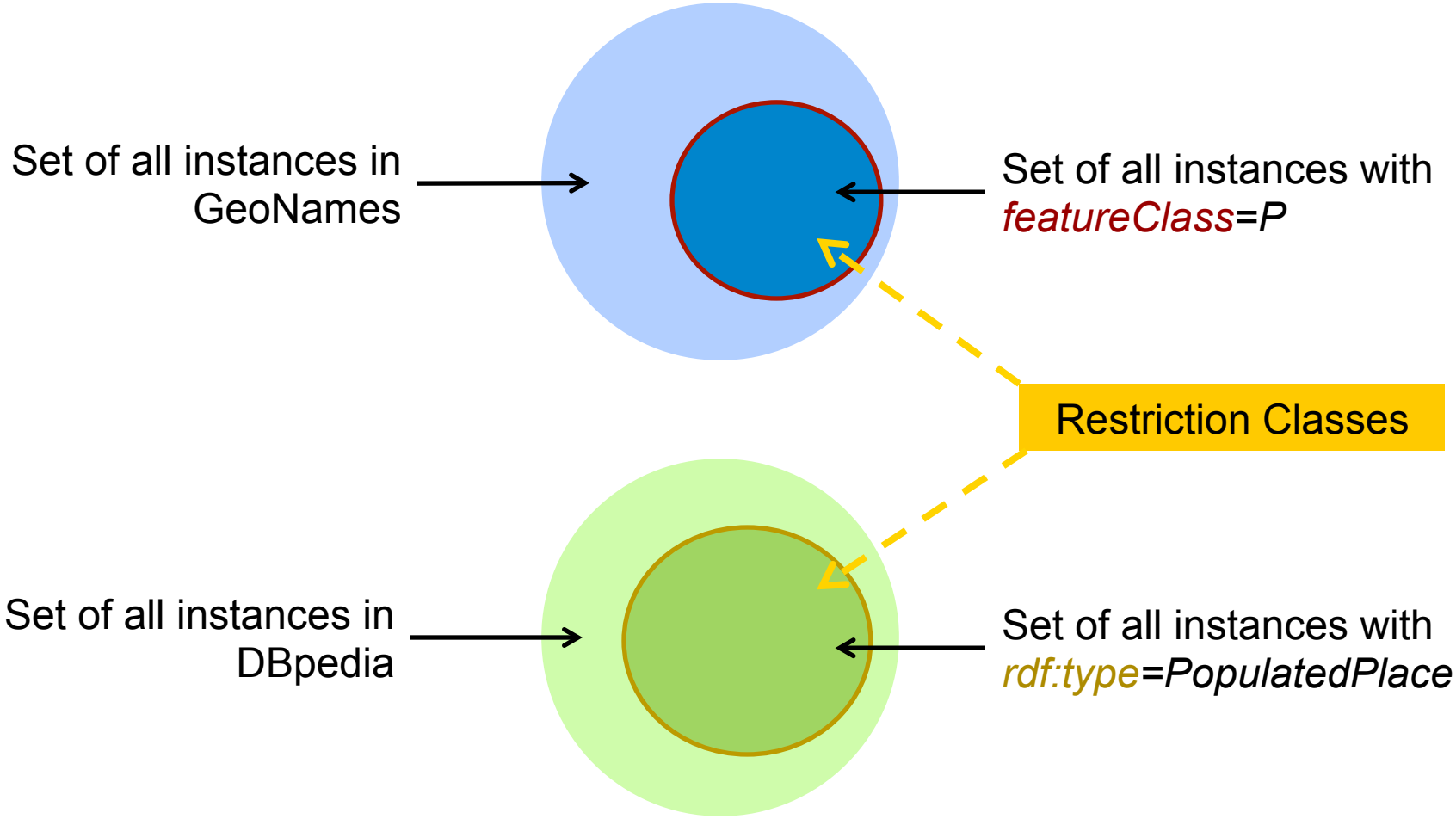
Classes are created extensionally by adding value restrictions on properties



Classes are created extensionally by adding value restrictions on properties



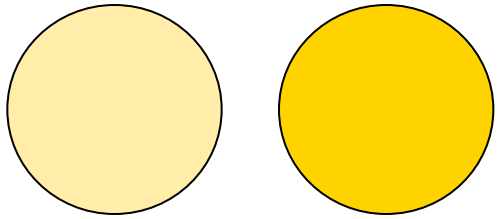
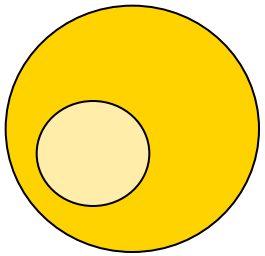
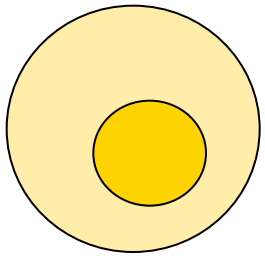
Classes are created extensionally by adding value restrictions on properties



Extensional Approach to Ontology Alignment using Restriction Classes

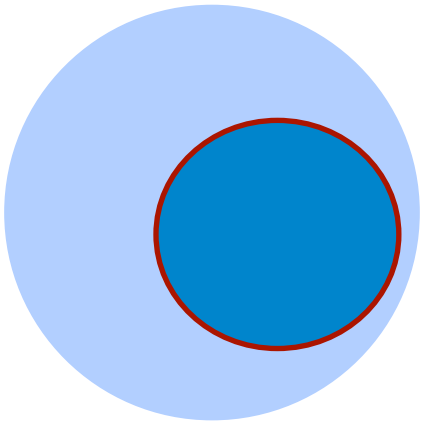
Represents set of instances belonging to ClassA

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 

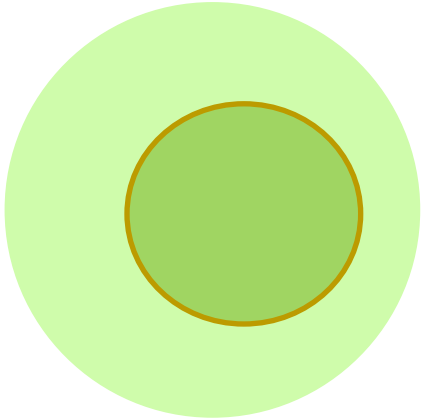
Aligning Restriction Classes Using Extensional Approach

featureClass=P



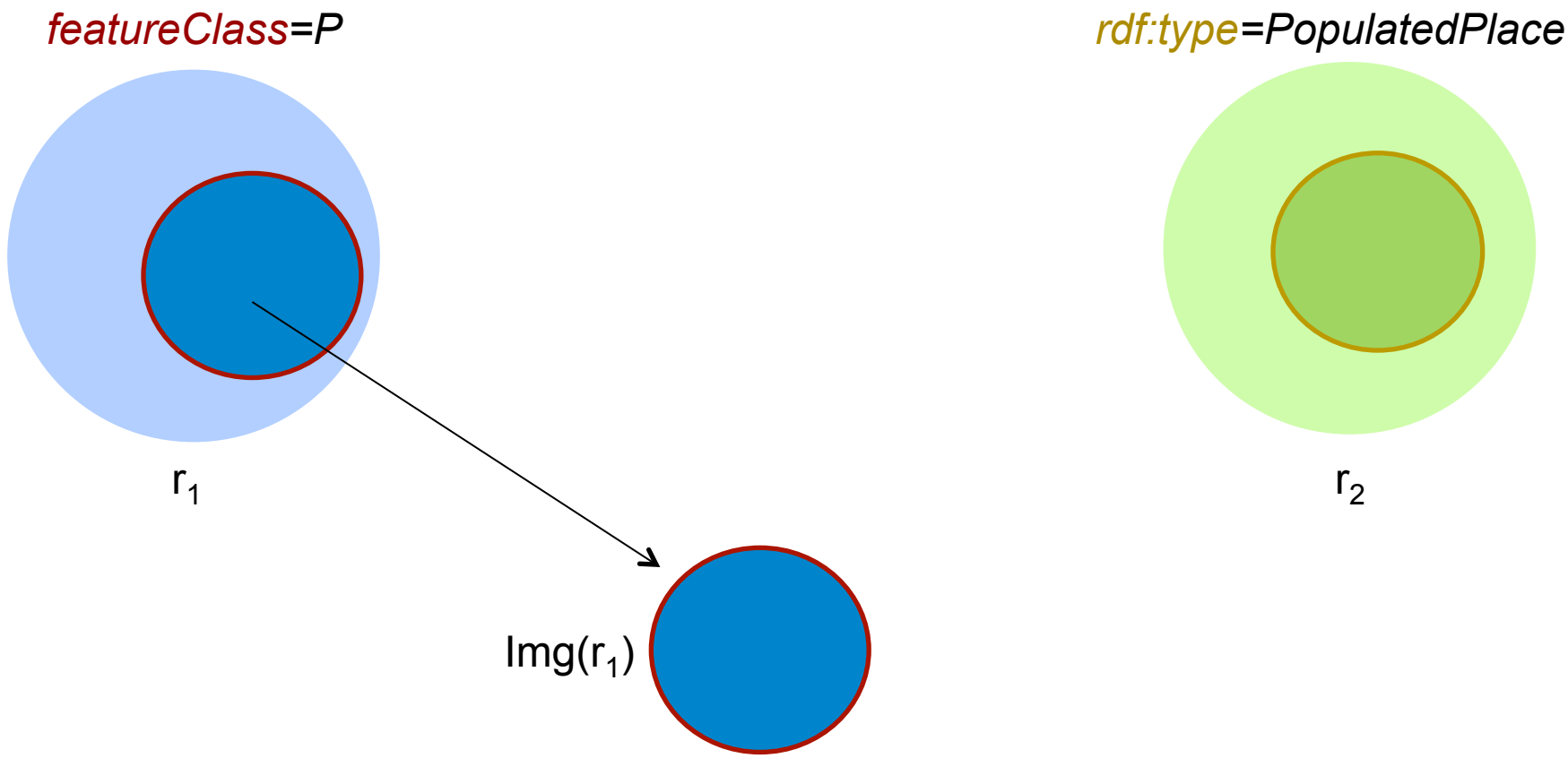
r_1

rdf:type=PopulatedPlace



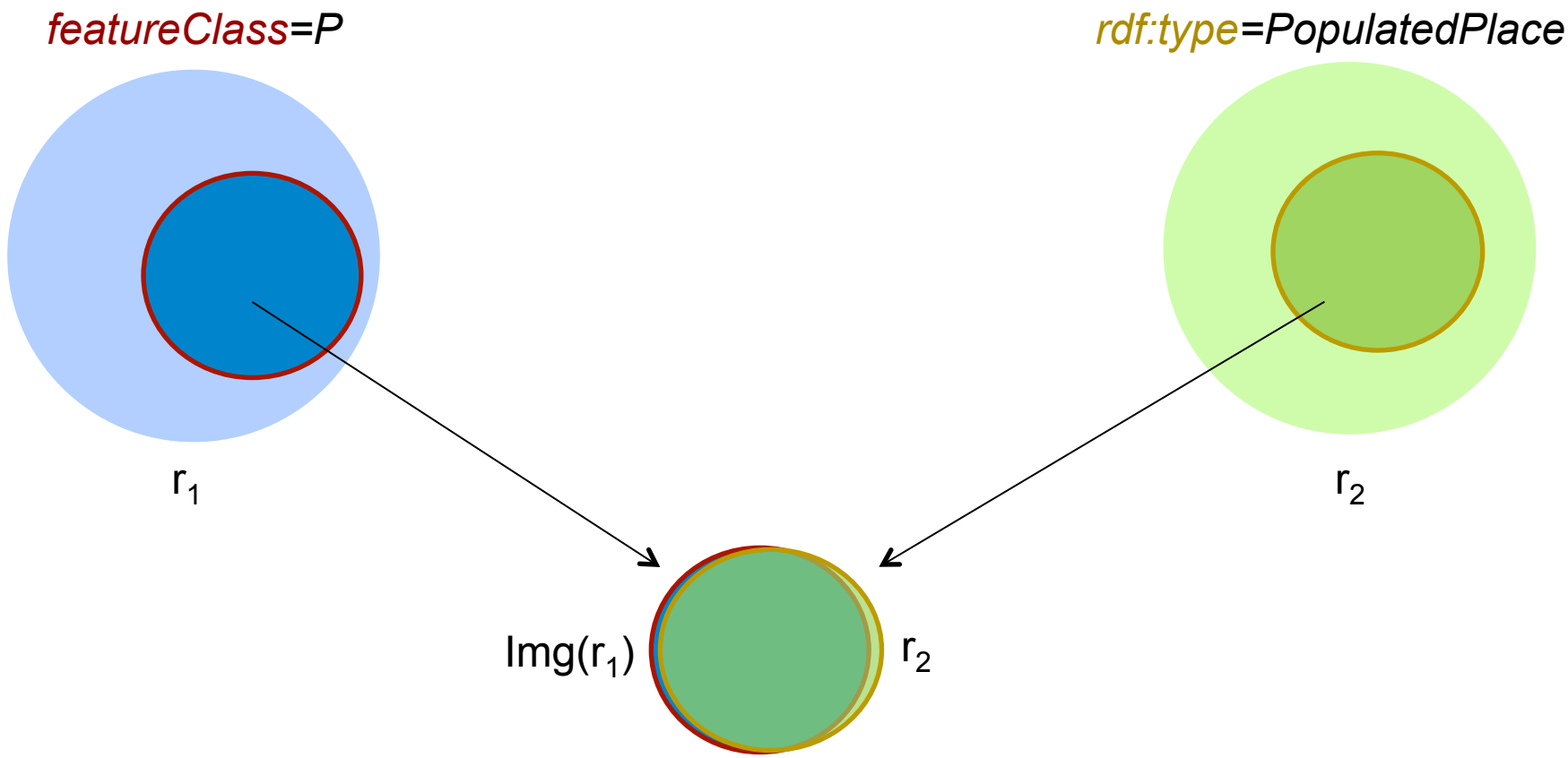
r_2

Aligning Restriction Classes Using Extensional Approach



Set of instances from DBpedia that r_1 is linked to

Aligning Restriction Classes Using Extensional Approach



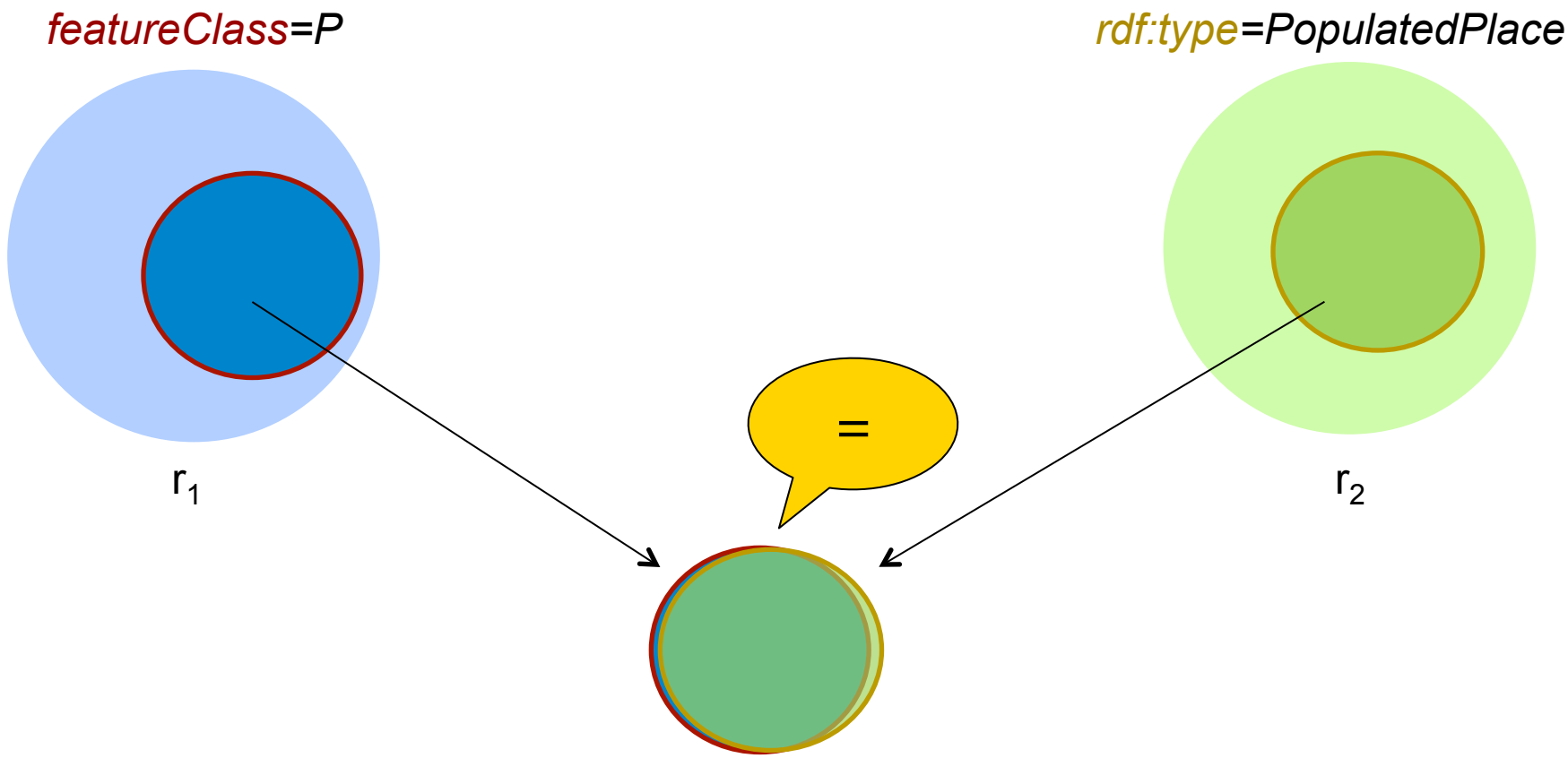
Extensionally, when are two classes equal?

Represents set of instances belonging to ClassA

Represents set of instances belonging to ClassB

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

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 Viterbi
School of Engineering

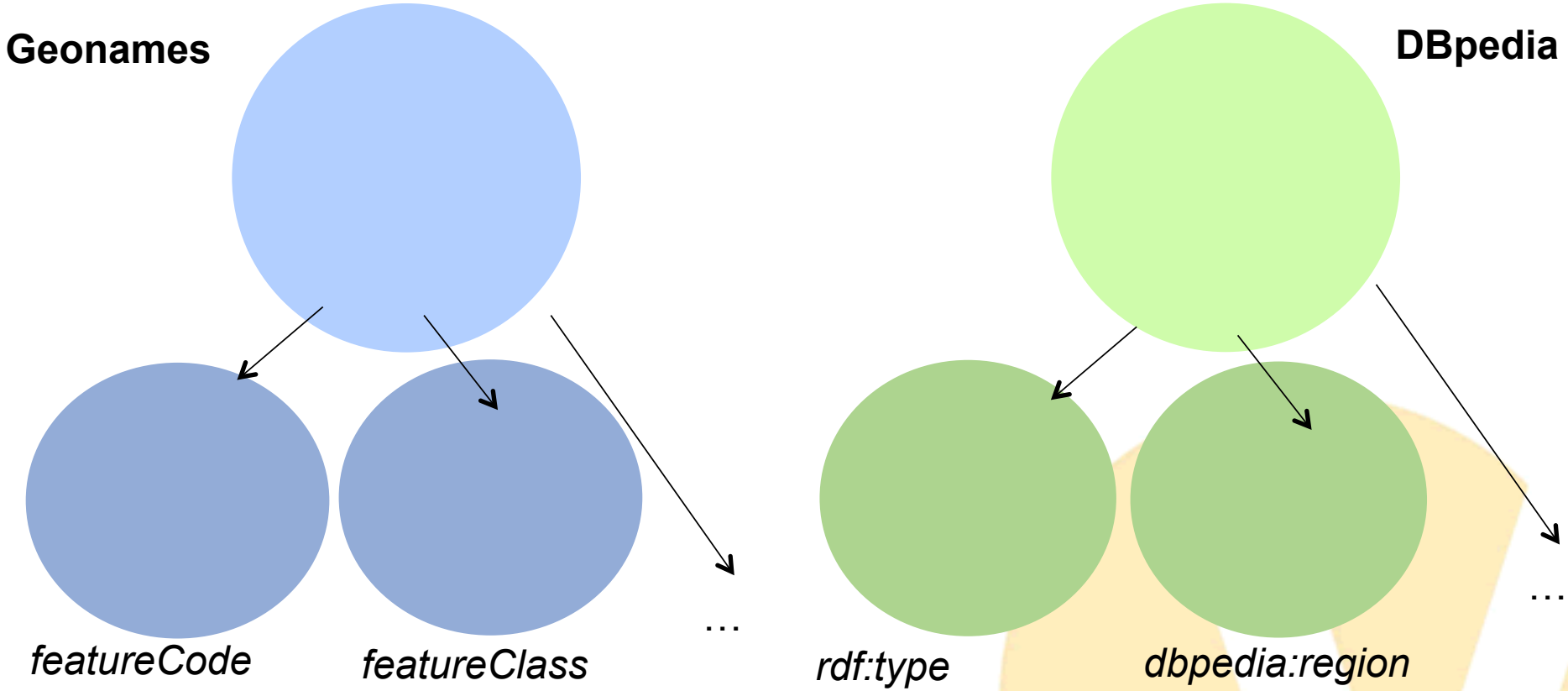
Step 1

FINDING ALIGNMENTS WITH ATOMIC RESTRICTION CLASSES

USC

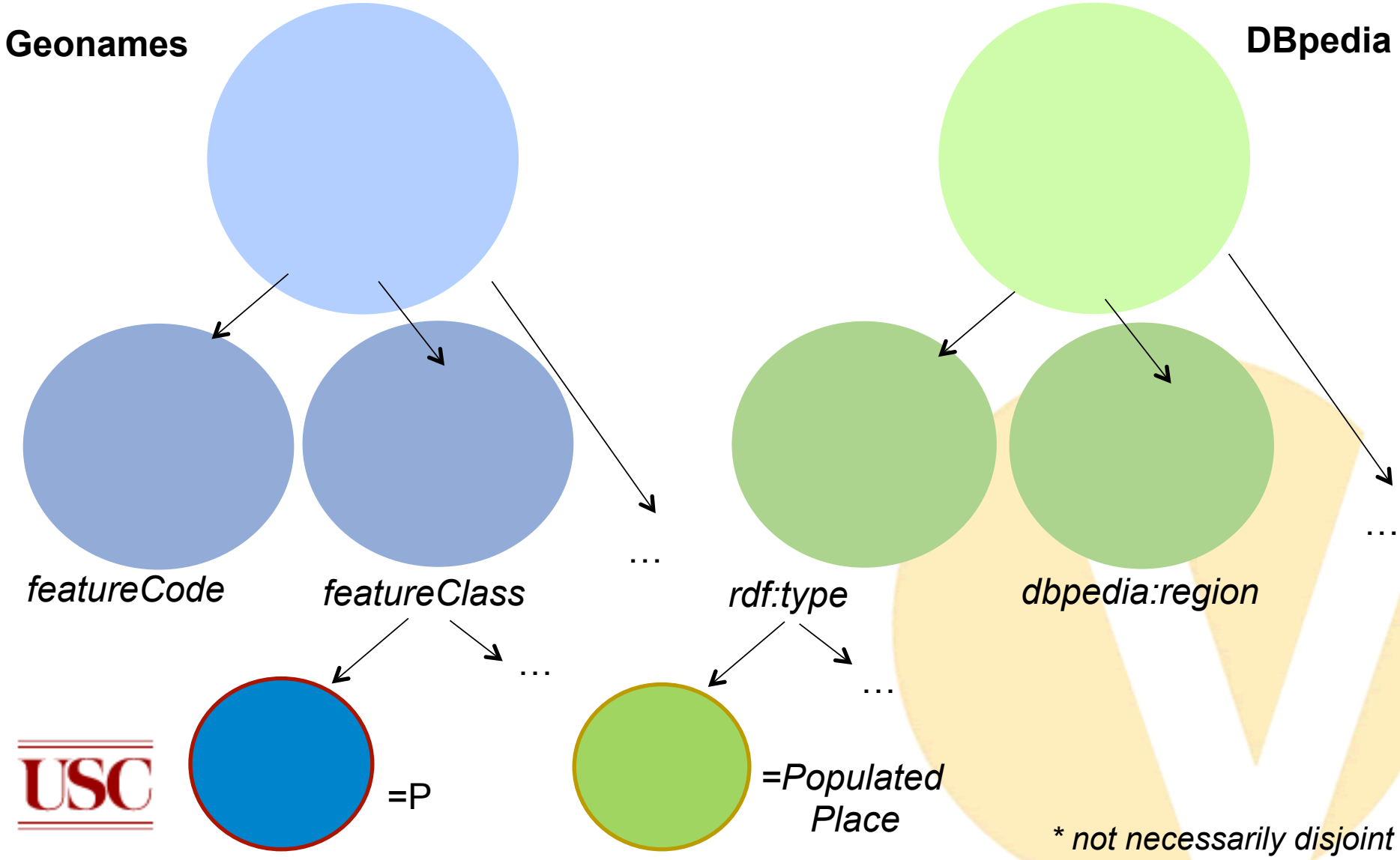
Approach: We start with a superset of all instances…

… and generate smaller subsets for each property*, …

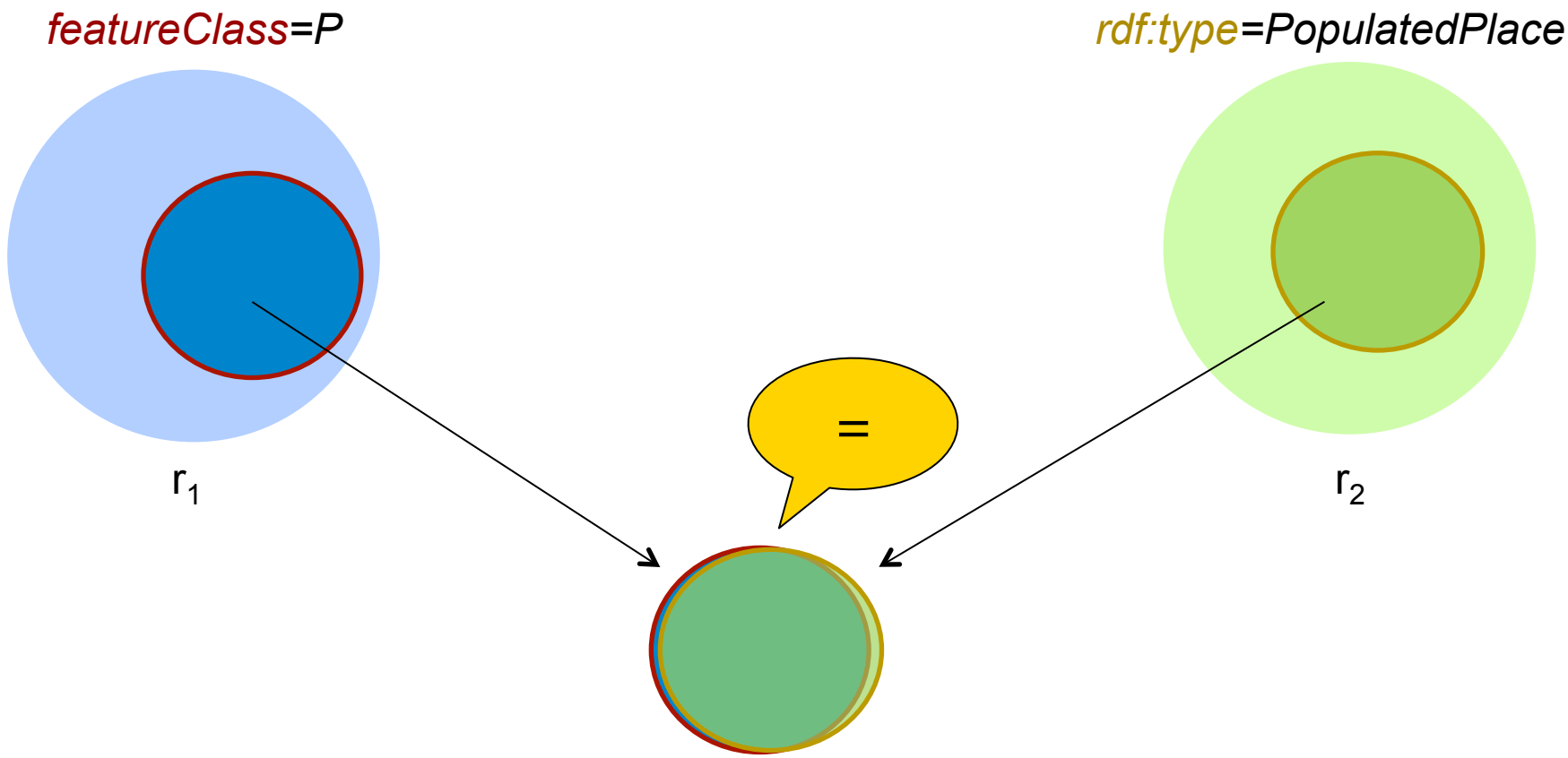


* not necessarily disjoint

… and generate yet smaller subsets for each value*



Comparing the two sets, we can align them equal if they fit



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

Linking and Building Ontologies of Linked Data [ISWC2010]

- Expressive of Restriction Classes using **Conjunction Operator**
  - E.g. define specialized concepts like Cities in the US
  - *featureCode*=P.PPL ^ *countryCode*=US
- Used top-down approach to find alignments
  - Specialize ontologies where original were rudimentary
  - Find complimentary hierarchy across an ontology

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

Step 2

IDENTIFYING CONCEPT COVERINGS

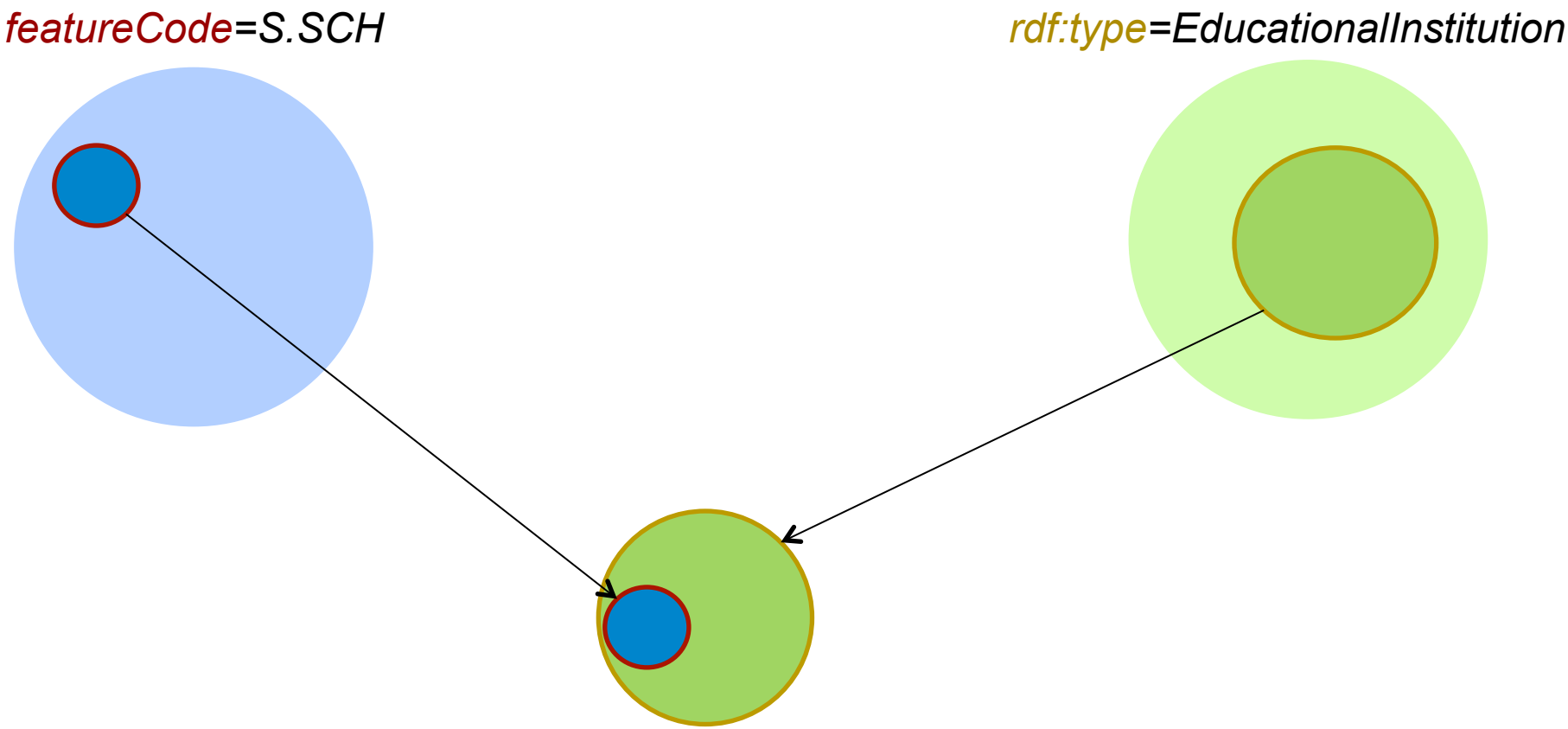
(DISJUNCTION OPERATOR FOR RESTRICTION CLASSES)

USC

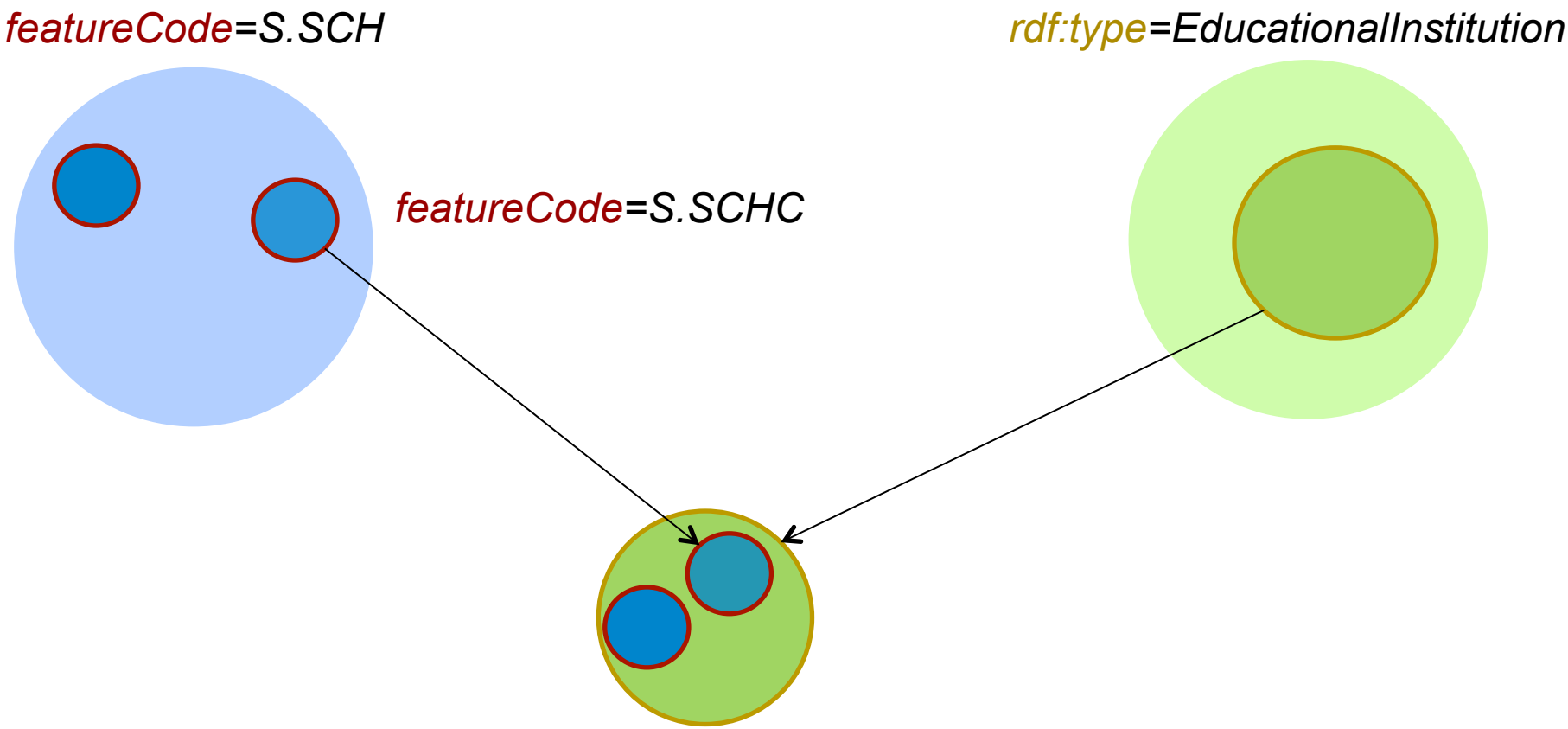
There is a pattern to be explored in the subset relations

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

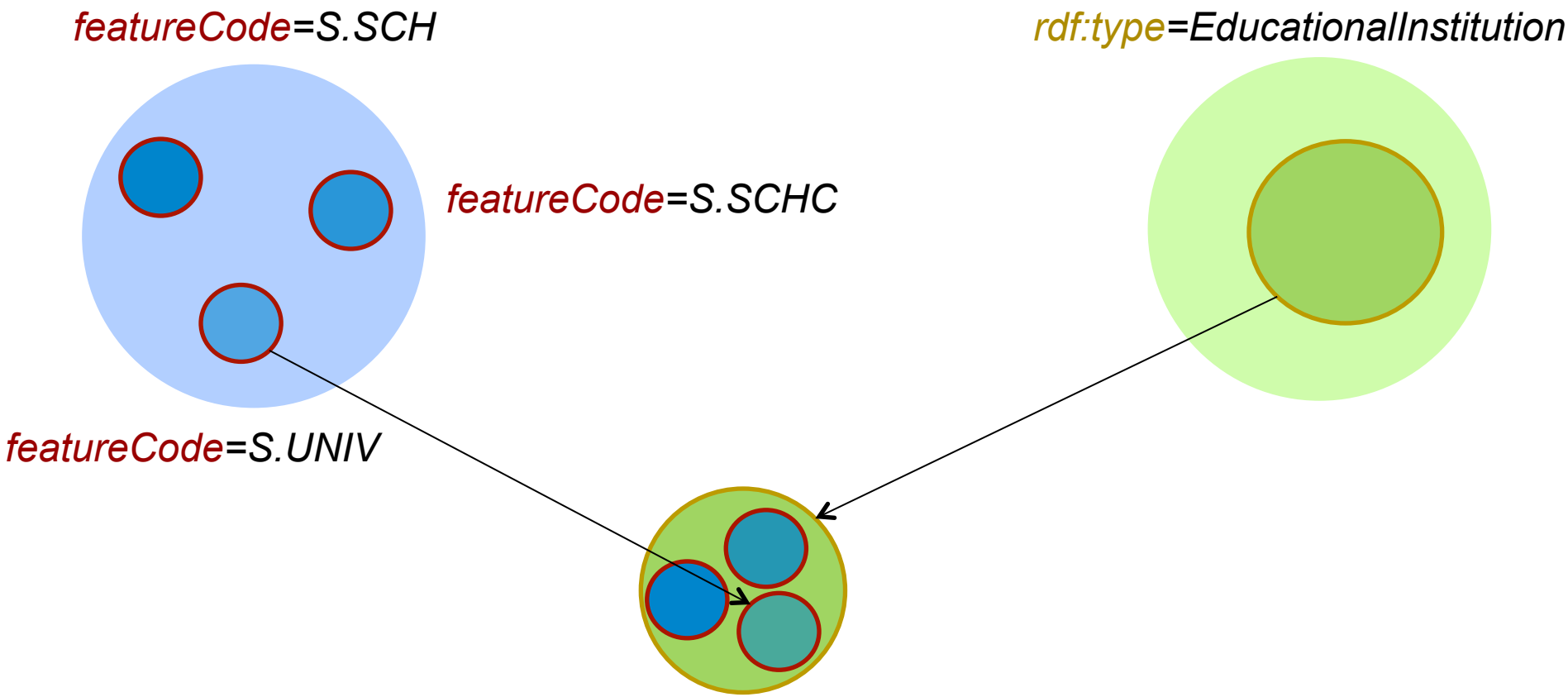
1) Schools in *GeoNames* are Educational Institutions in *DBpedia*



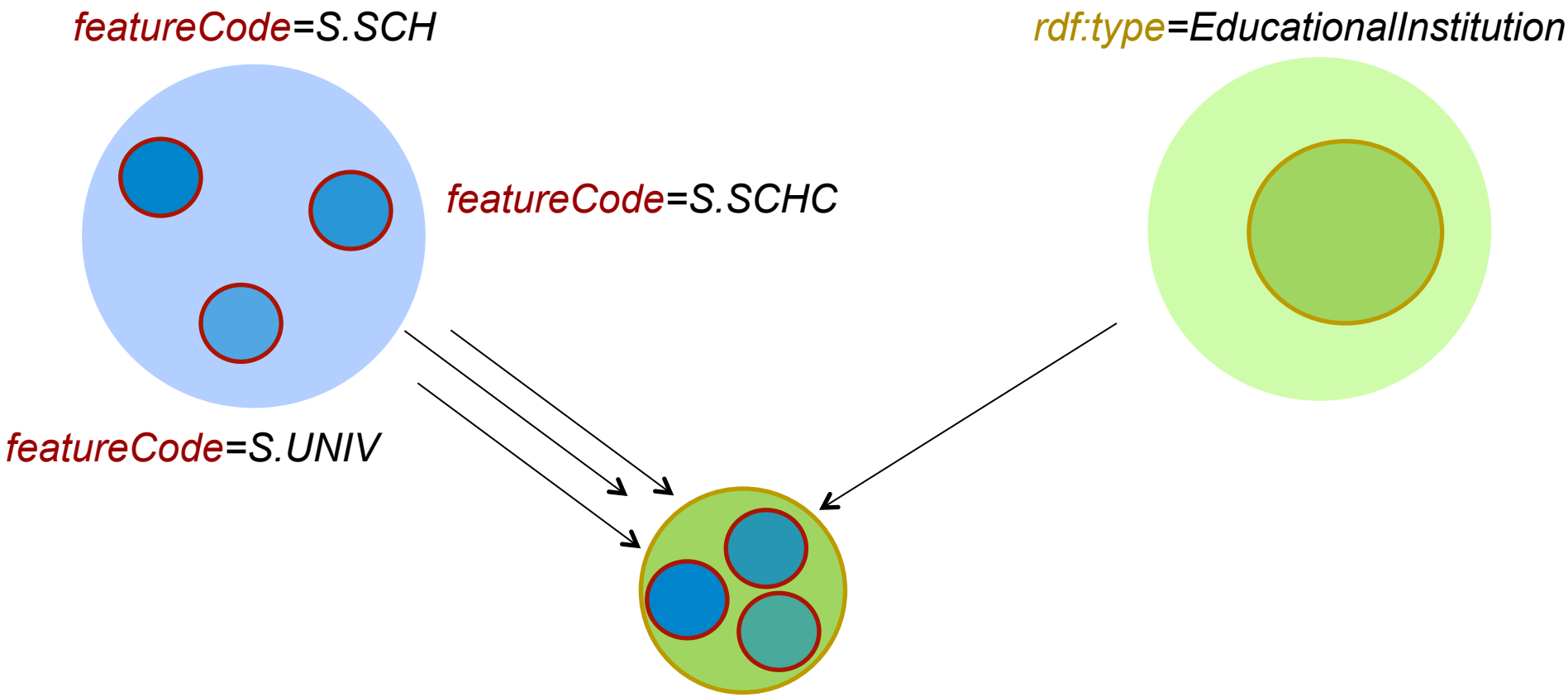
2) Colleges in *GeoNames* are Educational Institutions in *DBpedia*



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



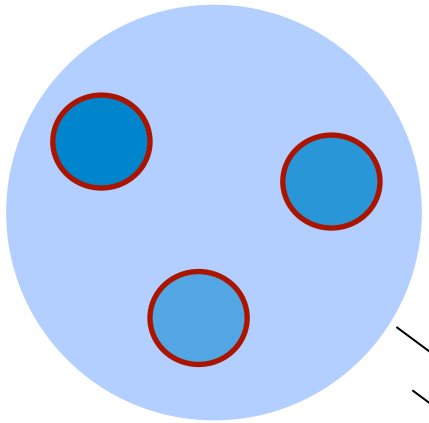
Taken by themselves, the subset relations are not useful



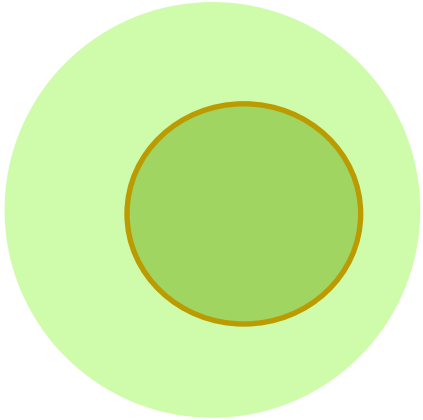
Using ***featureCode*** property as a hint, we form a ***Union*** of concepts

featureCode=*S.SCH*

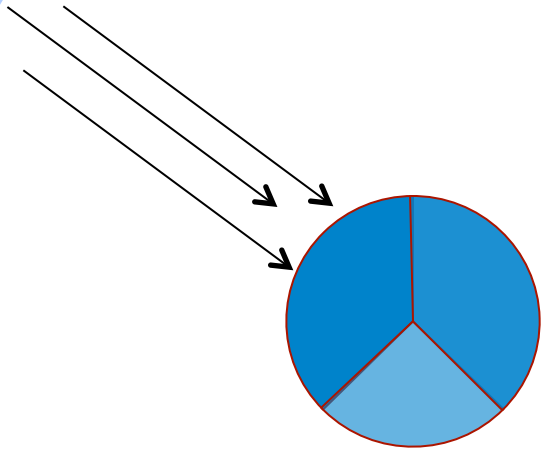
rdf:type=*EducationalInstitution*



featureCode=*S.SCHC*

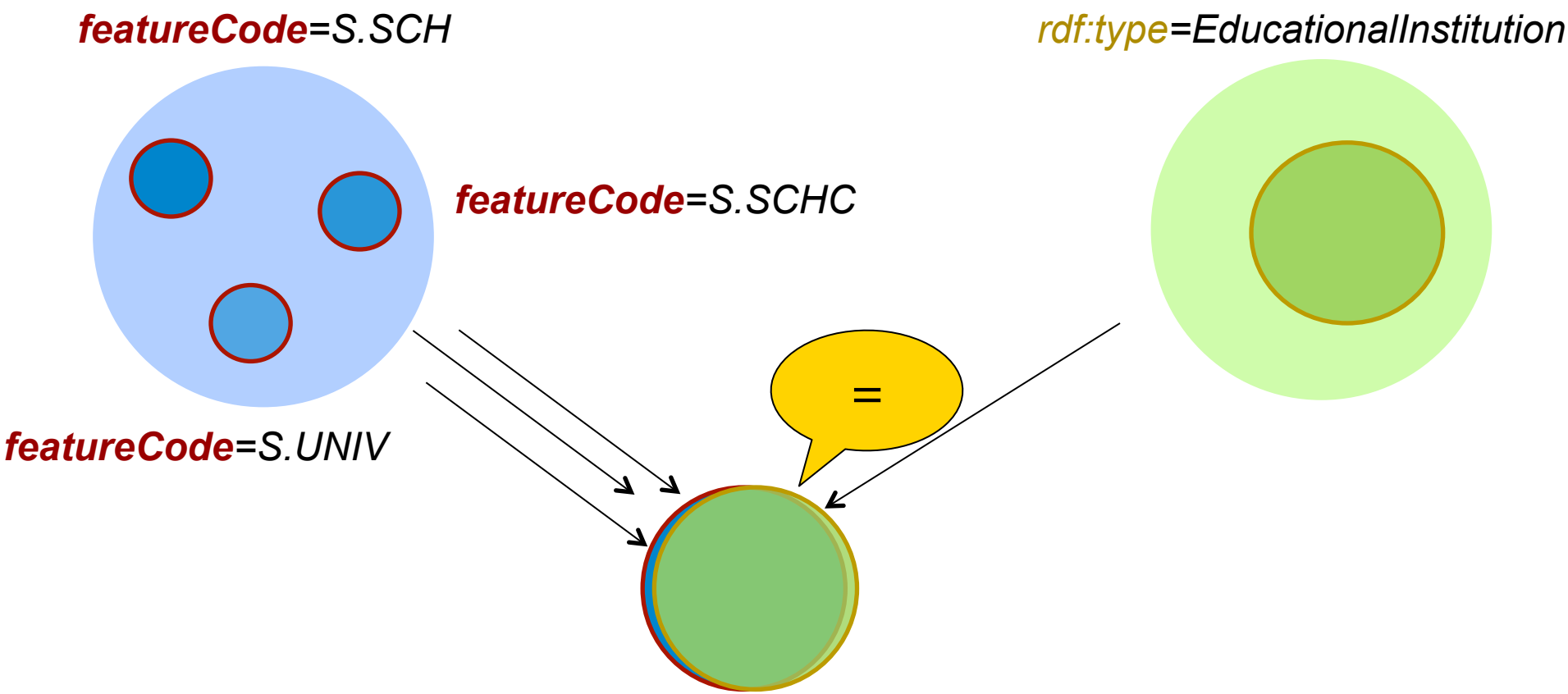


featureCode=*S.UNIV*



featureCode=*S.SCH* ∪ ***featureCode***=*S.SCHC* ∪ ***featureCode***=*S.UNIV*

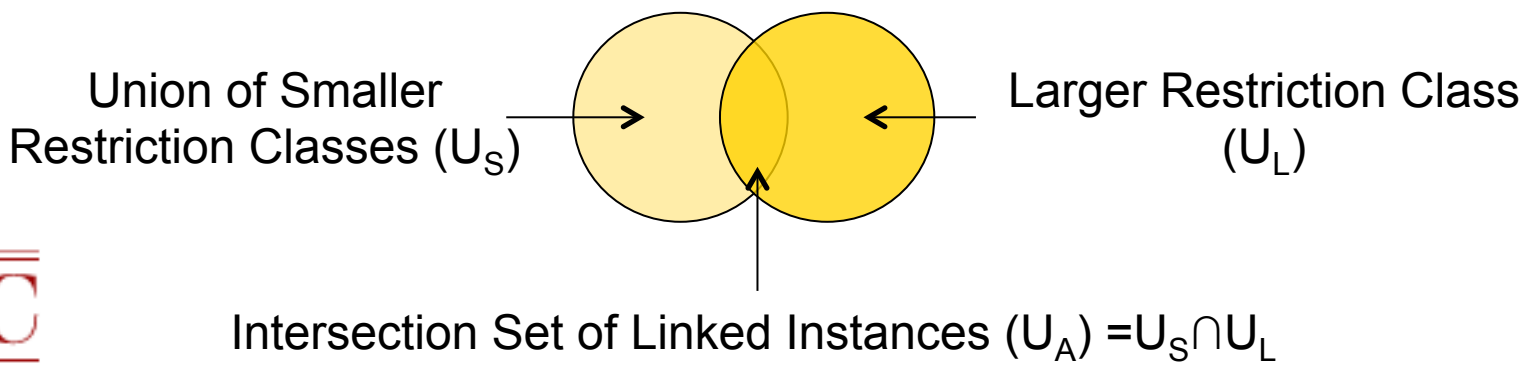
We Can Find Concept Coverings by Extensional Comparison (**Contribution 1**)



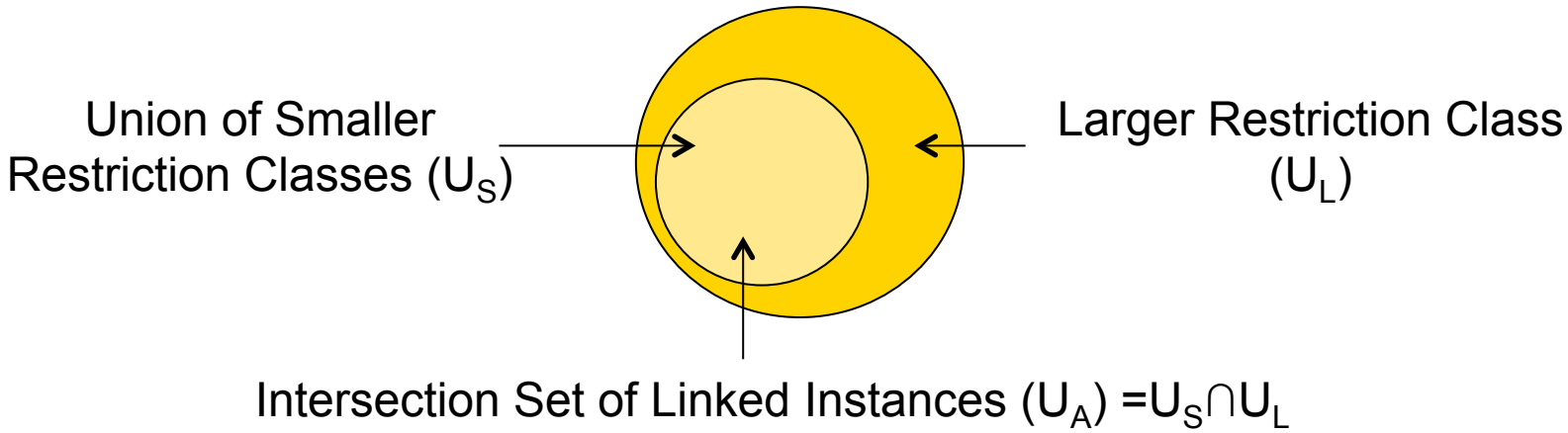
featureCode=*S.SCH* ∪ ***featureCode***=*S.SCHC* ∪ ***featureCode***=*S.UNIV*

Approach: Finding Concept Coverings

- For all alignments found in the Step 1
  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
  3. We then try to match the *union concept* to the larger class
  4. This forms a hypothesis *Concept Covering*



Scoring



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

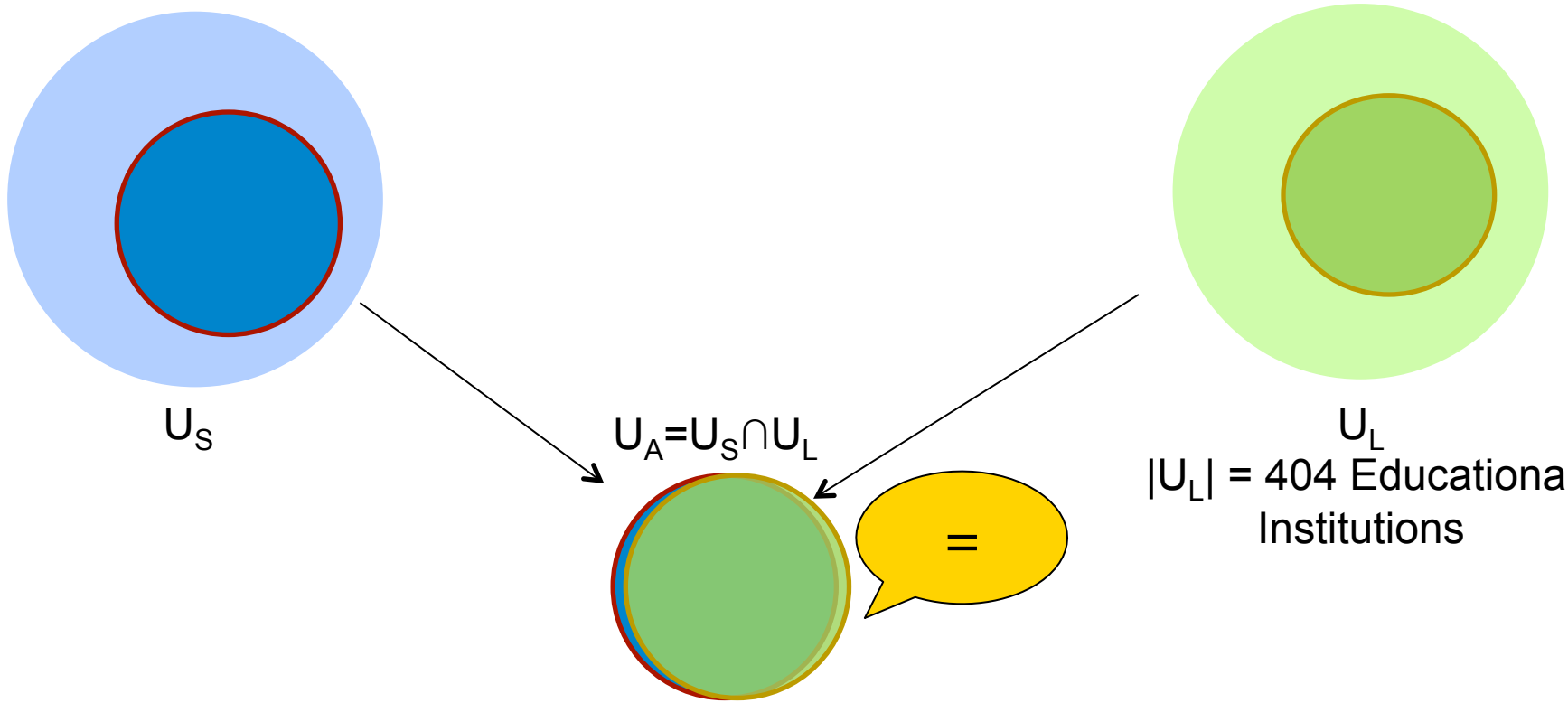
So, if $\frac{|U_A|}{|U_L|} = 1$, then the larger class U_L is equivalent to U_S

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

Upon comparison, we can determine equivalence

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

rdf:type=*EducationalInstitution*



$$\frac{|U_A|}{|U_S|} > 0.9$$

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

What are the other 8 Educational Institutions?

- 1 with *featureCode*=S.HSP (Hospitals)
  - 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 Viterbi
School of Engineering

CURATING THE LINKED DATA CLOUD

USC

Another Example: Am I in Spain … or Italy?

- We align *dbpedia:country=dbpedia:Spain* with *geonames:countryCode=ES*
- 3917 out of 3918 instances in GeoNames agree with this

- ONE instance had its country code as Italy.
- Because this instance contradicts overwhelming evidence, we can flag it as an outlier

Find Outliers / Discrepancies **(Contribution 2)**

- We are able to identify the instances that disagree with the alignment
- 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*
    - 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

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

USC Viterbi
School of Engineering

RESULTS

USC

USC Viterbi School of Engineering

Concept Coverings Found

We find a total of 7069 Concept Coverings that cover 77966 subset relations for a compression ratio of 11:1

$Source_1$	$Source_2$	O_1-O_2: Coverings (Subset Alignments)	O_2-O_1 Coverings (Subset Alignments)	Total Coverings
GeoNames	*DBpedia*	434 (2197)	318 (7942)	752
LinkedGeoData	*DBpedia*	2746 (12572)	3097 (48345)	5843
Geospecies	*DBpedia*	191 (1226)	255 (2569)	446
GeneID	*MGI*	6 (29)	22 (3086)	28

Results also available at

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

USC

USC Viterbi School of Engineering

Results: GeoNames-DBpedia

Larger Concept	Concepts Covered	Support	Outliers
***rdf:type* = Educational Institution**	***geonames:featureCode* = {S.SCH, S.SCHC, S.UNIV}**	396 out of 404 (R'_U=0.98)	S.BLDG (3/122), S.EST (1/13), …, S.MUS (1/43)
dbpedia:country = Spain	*geonames:countryCode* = {ES}	3917 out of 3918 (R'_U=0.99)	**IT (1/7635)**
rdf:type= Airport	geonames:featureCode= {S.AIRB, S.AIRP}	1981 out of 1996 (R'_U=0.99)	**S.AIRF (9/22)**, S.FRMT (1/5), …, **T.HLL (1/61)**

Larger Concept	Concepts Covered	Support	Outliers
geonames: countryCode= NL	*dbpedia:country*= {The_Netherlands, **Flag_of_the_Netherlands.svg**, Netherlands }	1939 out of 1978 (R'_U=0.98)	Kingdom_of_the_Netherlands (1/3)

USC Viterbi School of Engineering

Evaluation: GeoNames-DBpedia

- Evaluation
  - Manually Evaluated **236** out of **752** alignments
  - **152** identified as correct, **Precision of 64.4%**

- Common problems evaluated as incorrect **(84)**
  - 'County' property was mis-labelled as 'Country' **(5)**
  - Using the '.svg' filename of the flag of a Country as value of '*dbpedia:country'* property **(35)**
  - Partial alignments with sub-classes detected as outliers **(14)**
    - Not enough support for set containment detection (P' < 0.9)
  - Incompletely detected alignments **(7)**
    - Missing instances for complete definition
  - Other problems: Misaligned with parent **(14)**, etc. **(9)**

USC

USC Viterbi School of Engineering

Establishing Recall and F-Measure

- Establishing recall for all alignments was difficult
  - Manually establishing all possible ground truth infeasible
- Evaluated F-measure for Countries as a representative
  - *dbpedia:country* property in DBpedia
  - *geonames:countryCode* property in GeoNames
- 63 Country-CountryCode Alignments evaluated manually
  - **Precision: 53 / 63 = 84.13%**
    - 26 were correct
      - *Insight needed: United Kingdom in GeoNames vs England, Scotland, Wales, Northern Ireland in DBpedia
    - 27 were assumed correct because data had inconsistences
      - A '.svg' file appeared as country in DBpedia
  - **Recall: 53 / 169 = 31.36%**
  - **F-Measure: 45.69%**

USC

Results: *LinkedGeoData-DBpedia*

Larger Concept	Concepts Covered	Support	Outliers
dbpedia: **Bundesland = Saarland**	***lgd:LicensePlateNum=*** **{HOM, IGB, MZG, NK, SB, SLS, VK, WND}**	46 out of 49 (R'_U=0.93)	**(Missing)**

Larger Concept	Concepts Covered	Support	Outliers
lgd:ST_alpha=NJ	*dbpedia:country=* {Atlantic, Burlington, …} **We only found 9 of the 21 counties**	214 out of 214 (R'_U=1)	
rdf:type= **lgd:Waterway**	***rdf:type=*** ***{River, Stream}***	33 out of 34	Place (1/94989)

USC Viterbi School of Engineering

Results: *LinkedGeoData-DBpedia*

- **Evaluation**
  - Manually Evaluated **200** out of **5843** alignments
  - **157** identified as correct, **Precision of 78.2%**

- Common problems evaluated as incorrect **(43)**
  - Multiple spellings for the same item **(14)**
  - Partially or incompletely found **(20)**
  - Other problems **(9)**

USC

USC Viterbi School of Engineering

Results: *Geospecies-DBpedia*

Larger Concept	Concepts Covered	Support	Outliers
rdf:type= Amphibian	*geospecies:orderName*= {Anura, Caudata, Gymnophionia}	90 out of 91 (R'_U=0.99)	**Testidune (1/7)** [i.e. Turtle]
rdf:type= Salamander	*geospecies:orderName*= {Caudata}	16 out of 17 (R'_U=0.99)	**Testidune (1/7)**

Larger Concept	Concepts Covered	Support	Outliers
geospecies: hasOrderName = "Chiroptera"	*dbpedia:ordo*= {**"Chiroptera"@en** , dbpedia:Bat}	246 out of 247 (R'_U=1)	

USC

USC Viterbi School of Engineering

Results: *Geospecies-DBpedia*

- **Evaluation**
  - Manually Evaluated **178** out of **446** alignments
  - **109** identified as correct, **Precision of 61.84%**

- Common problems evaluated as incorrect **(69)**
  - Multiple spellings for the same item **(25)**
  - Partially or Incompletely found because of outliers / small sizes of support **(28)**
  - Other problems **(16)**

USC

USC Viterbi School of Engineering

Results: *GeneID-MGI*

Larger Concept	Concepts Covered	Support	Outliers
bio2rdf:subType= pseudo	*bio2rdf:subType*= {Pseudogene}	5919 out of 6317 (R'_U=0.93)	**Gene (318/24692)**

Larger Concept	Concepts Covered	Support	Outliers
bio2rdf:subType= {Pseudogene}	*bio2rdf:subType*= pseudo	5919 out of 6297 (R'_U=0.94)	other (4/30), protien-coding (351/39999), unknown (23/570)
mgi:genomeStart= 1	geneid:location= { “1”, “1 0.0 cM”, “1 1.0 cM”, “1 10.4 cM”, …}	1697 out of 1735 (R'_U=0.98)	**“” (37/1048), “5” (1/52)**

USC

USC Viterbi School of Engineering

Results: *GeneID-MGI*

- **Evaluation**
  - Manually Evaluated **28** alignments found
  - **24** identified as correct, **Precision of 85.71%**

- Common problems evaluated as incorrect **(4)**
  - Partially or Incompletely found **(4)**

USC

USC Viterbi School of Engineering

Related Work

- BLOOMS, BLOOMS+ ([8][9] 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 (Precision=0.5%)
- AgreementMaker [2]
  - Similarity Metrics on labels of classes
  - GeoNames (10 concepts) & DBpedia (257 concepts)
  - Precision=26%, Recall=68%
- Volker et al. ([13] in paper)
  - Statistical schema induction Mines associativity rules from intermediate *'transaction datasets'* -> *OWL2 Axioms.*

USC

USC Viterbi School of Engineering

Conclusion and Future Work

- Conclusion
  - We were able to find Concept Coverings in the Geospatial, Biological Classification & Genetics Domain
    - Find alignments where no direct equivalence was evident
    - Introduced a disjunction operator to create restriction classes
  - We were able to find *Outliers*
    - Help identify inconsistencies in the data
- Future work
  - Could Patterns within properties like *geonames:countryCode* and *dbpedia:country* be explored?
  - Ranges of Properties have a lot of inconsistencies
  - Flag outliers and contribute to PedanticWeb for correction

USC

USC Viterbi
School of Engineering

Any questions?

THANK YOU

USC