

# Building Semantic Descriptions of Sources

**Craig Knoblock**  
**University of Southern California**



- Vast collection of interlinked information
- Various sources and services with different schemas



# Where do the Semantics Come From?

- **Linked Open Data**
  - Populated by manually linking or writing procedures that define the links across sources
  - But we don't know how the sources are related
  - In many cases there is no or very limited semantic descriptions of sources
- **Linked Open Services**
  - Manually constructed or built by wrapping existing Web services
  - Constructing the lifting and lowering rules that relate the services to existing ontologies is a difficult task
  - Even when done, it may only provide a partial description
    - e.g., descriptions of the inputs and outputs, but not the function of a service

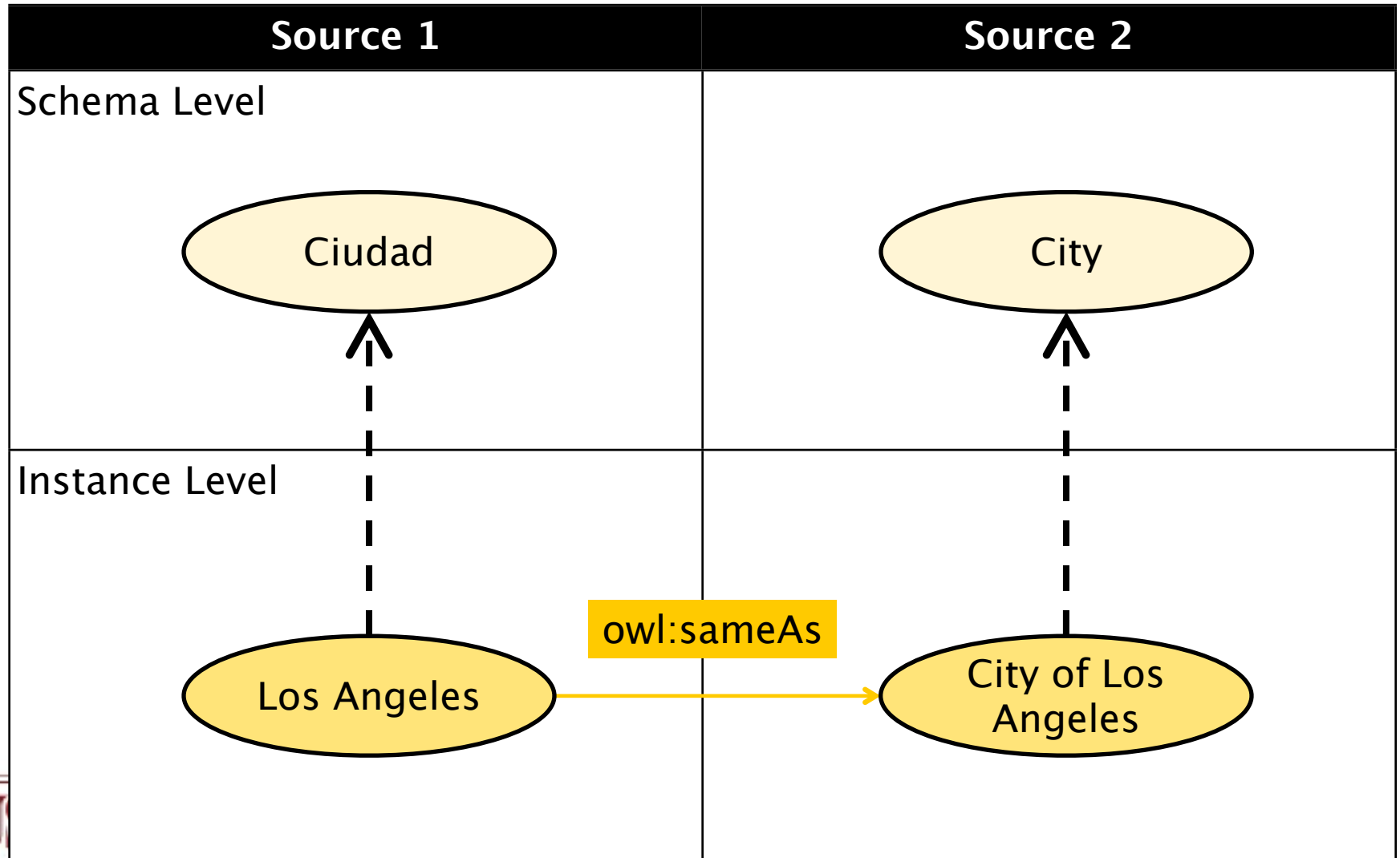
# Outline of the Talk

- **Linked Open Data**
  - Building and linking ontologies of linked data
- **Linked Open Services**
  - Building semantic web services from the Deep Web
- **Discussion**
  - Remaining challenges

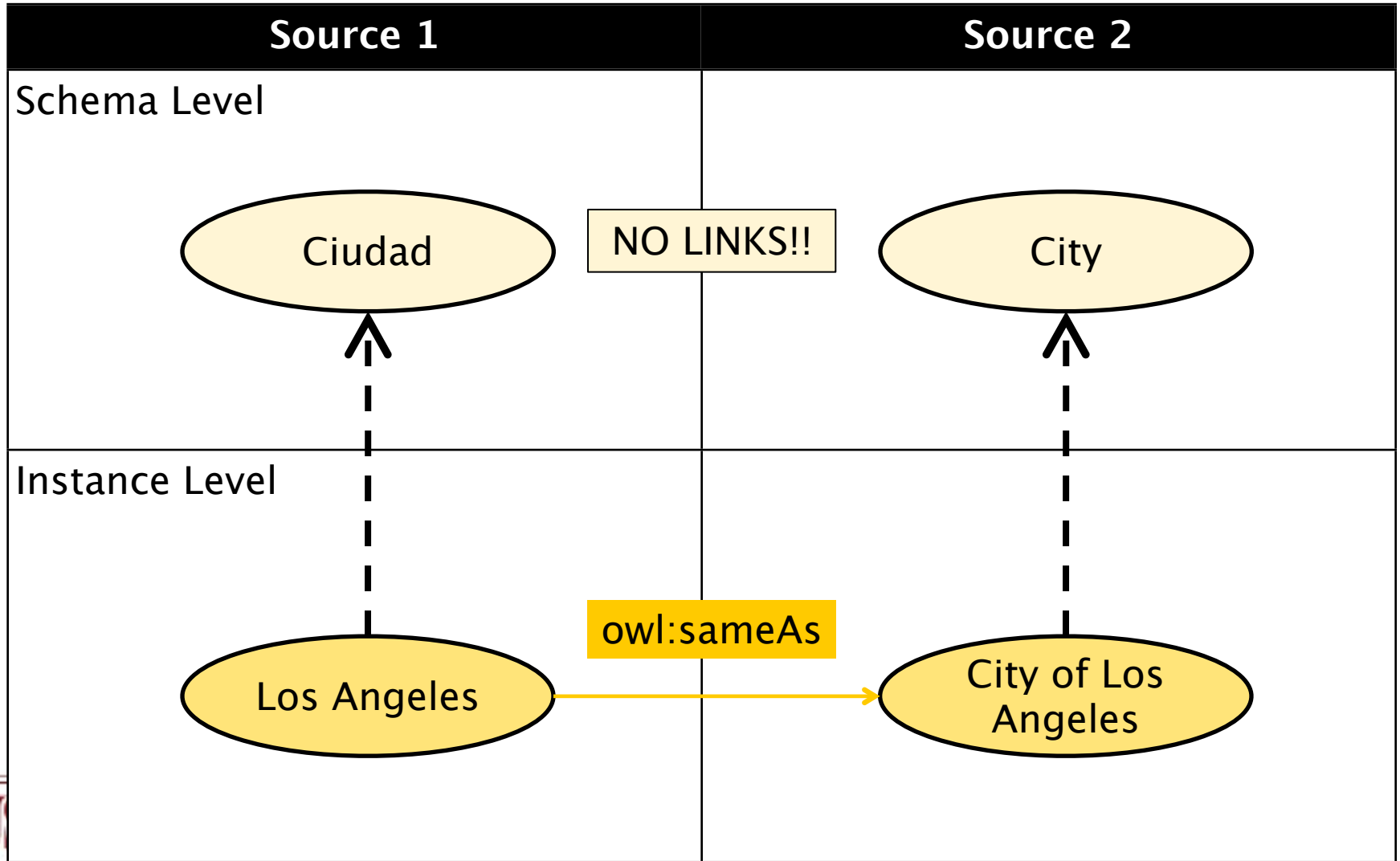
# Outline of the Talk

- **Linked Open Data**
  - Building and linking ontologies of linked data
- Linked Open Services
  - Building semantic web services from the Deep Web
- Discussion
  - Remaining challenges

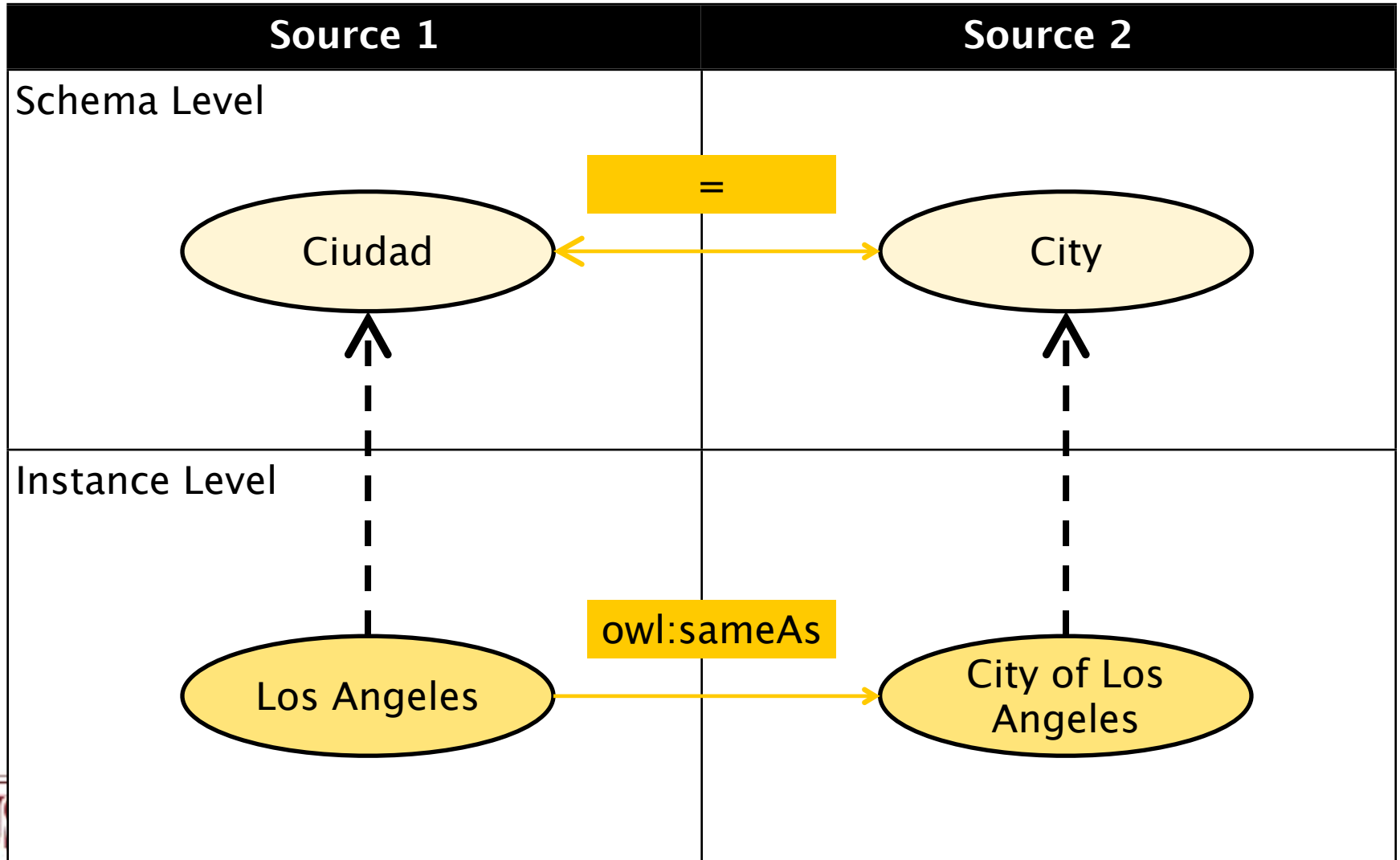
# Building and linking ontologies of linked data [Parundekar et al., ISWC 2010]



# Disjoint Schemas



# Objective 1: Find Schema Alignments

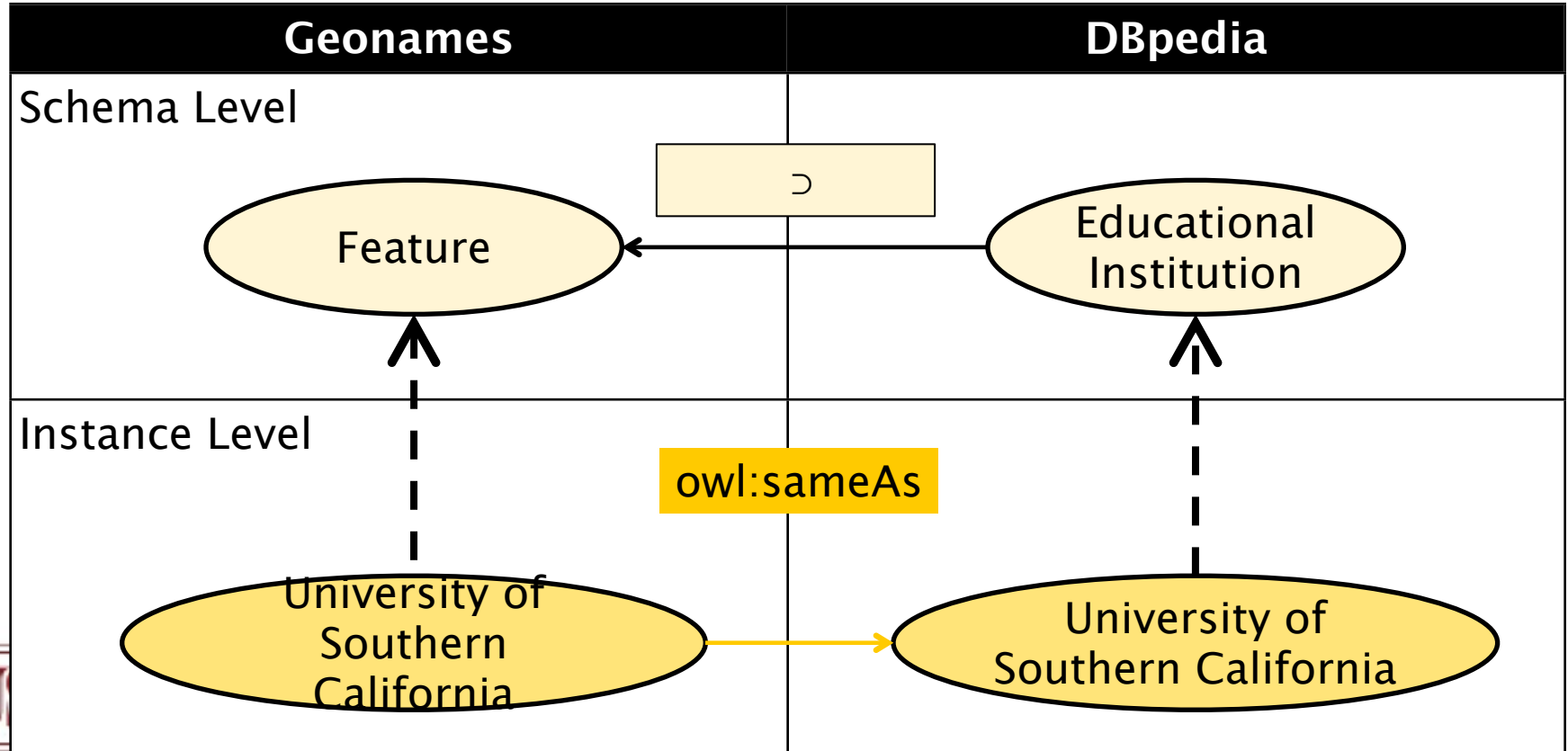




# Ontologies of Linked Data

- **Ontologies can be highly specialized**
  - e.g. DBpedia has classes for Educational Institutions, Bridges, Airports, etc.
- **Ontologies 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
- **There might not exist exact equivalences between classes in two sources**

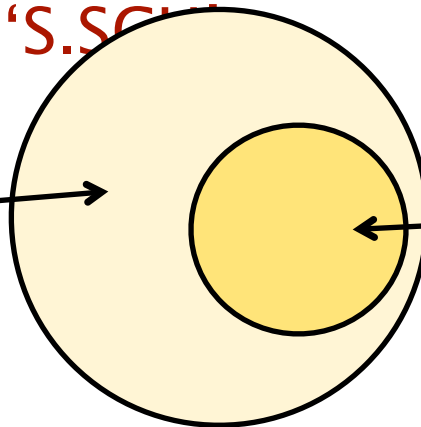
- Only subset relations possible with difference in class specializations



# 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'

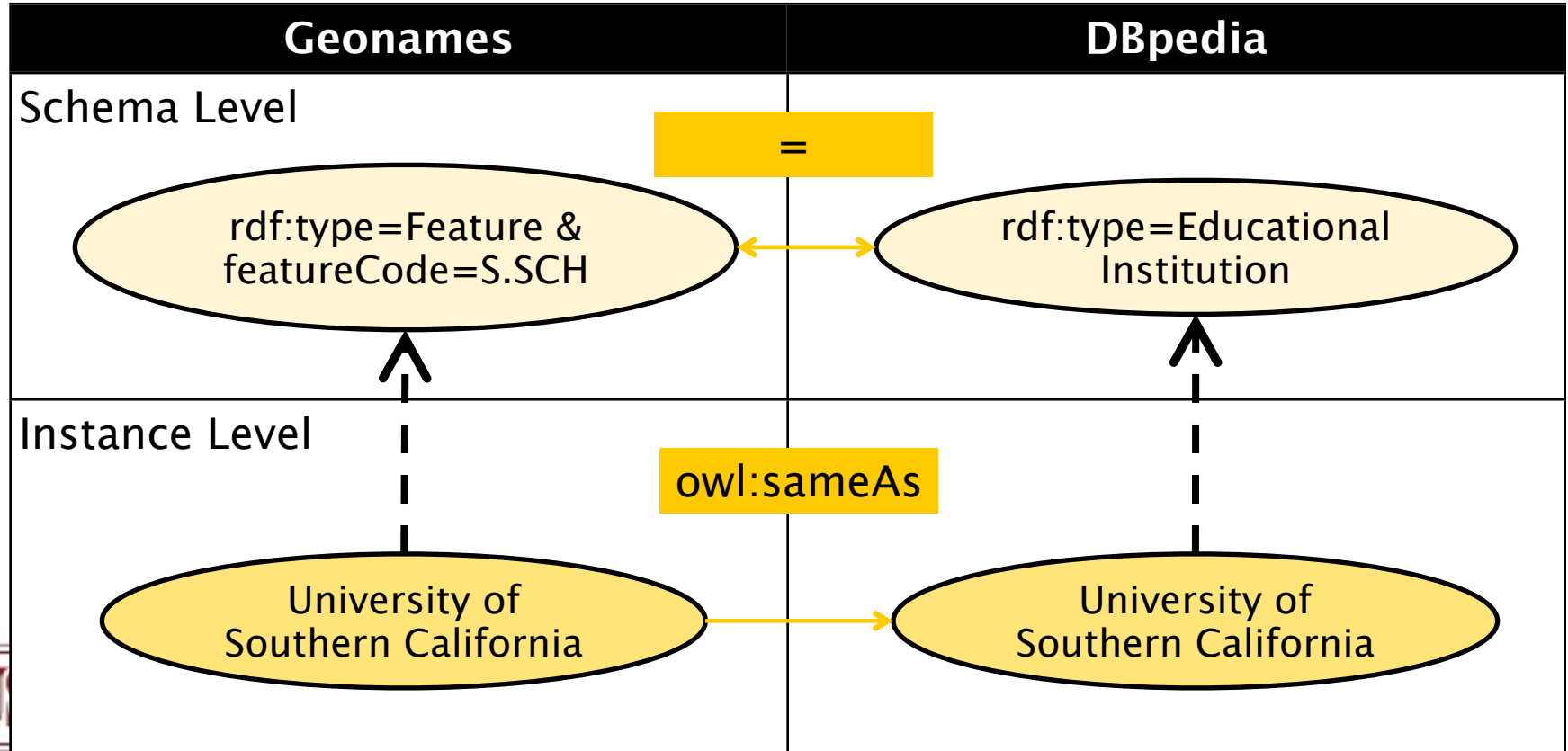
Set of all instances in  
Original Class -  
rdf:type=Feature



Set of all instances in  
Restricted Class -  
rdf:type=Feature &  
featureCode=S.SCH

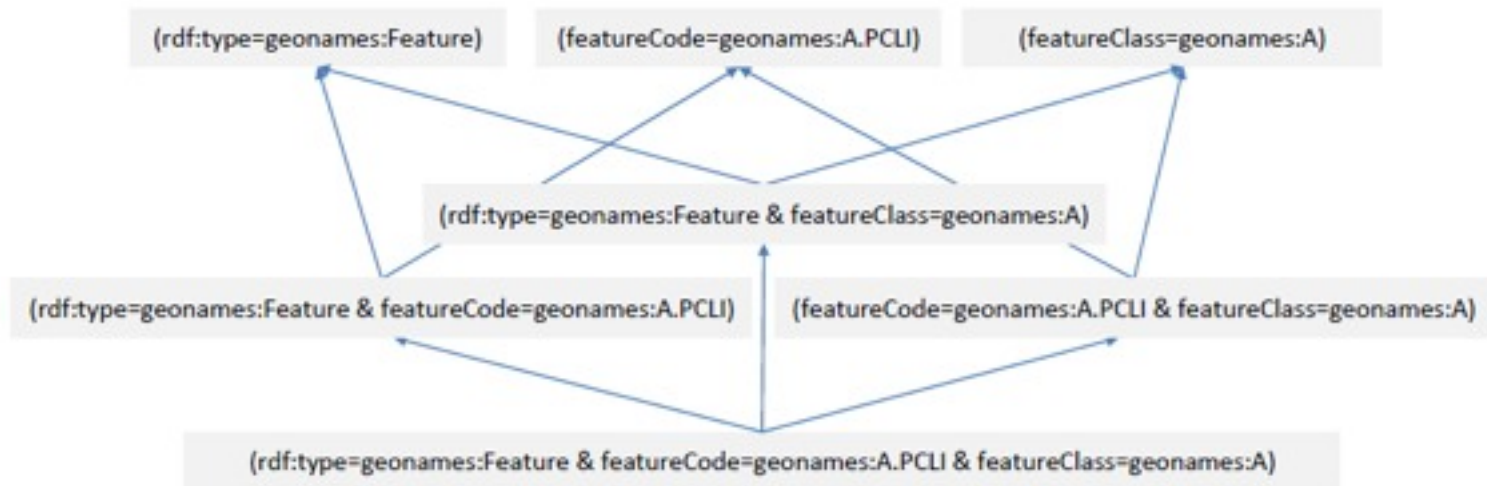
# Objective 2: Find Alignments Between Restriction Classes

- Find and model specialized descriptions of classes





# Nature of Restriction Classes

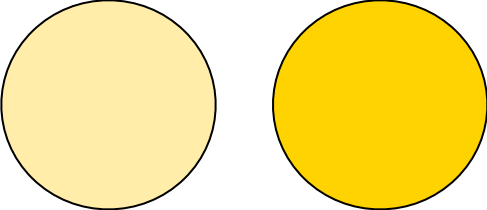
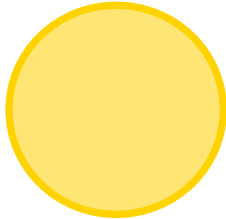
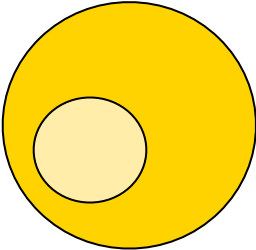
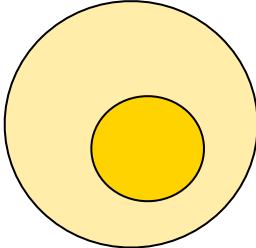
- Instances belonging to a restriction class also belong to parent restriction class
  - e.g. restrictions from Geonames below



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

# Extensional Approach to Ontology Alignment

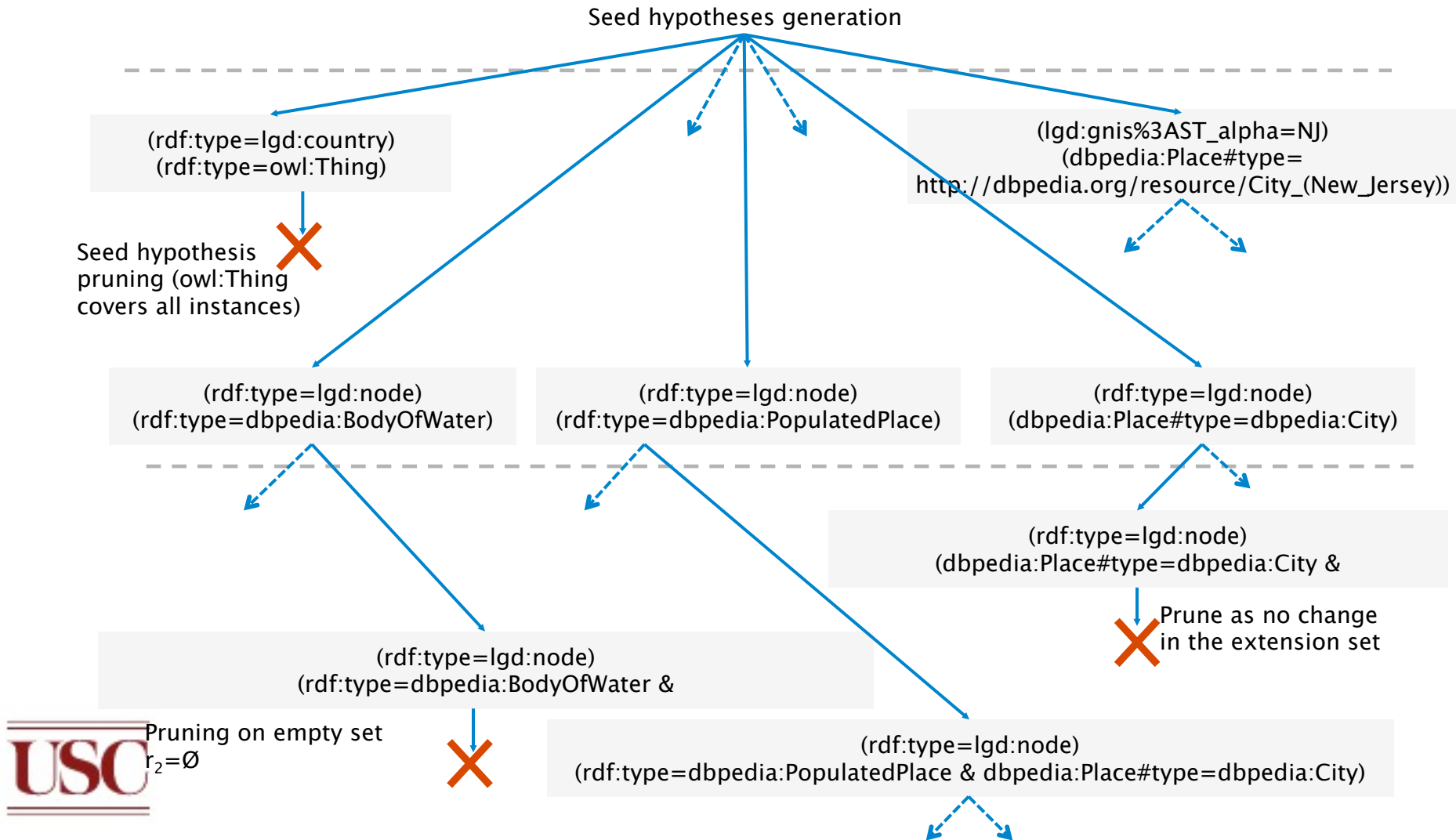
-  Represents set of instances belonging to ClassA
-  Represents set of instances belonging to ClassB

|   |   |
|---|---|
| <p>ClassA is disjoint from ClassB</p>  | <p>ClassA is equivalent to ClassB</p>  |
| <p>ClassA is subset of ClassB</p>    | <p>ClassB is subset of ClassA</p>    |

# Alignment Hypotheses

- An alignment hypothesis considers aligning
  - a restriction class from ontology  $O_1$
  - another restriction class from ontology  $O_2$
- Find relation between the two restriction classes
  - using extensional comparison on set of instances belonging to each restriction class
  - Use instance pair identifiers from pre-processing step (combination of URIs of linked instances)

# Exploration of Hypotheses Search Space





# Example Alignments from LinkedGeoData, Geonames, and DBpedia

| #  | 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=<br><a href="http://dbpedia.org/resource/City_(New_Jersey)">http://dbpedia.org/resource/City_(New_Jersey)</a> | $r_1 = r_2$       |
| 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:PopulatedPlace  | $r_1 = r_2$       |
| 7  | geonames:featureClass=geonames:H  | rdf:type=dbpedia:BodyOfWater   | $r_1 = r_2$       |
| 8  | geonames:parentFeature= <a href="http://sws.geonames.org/3174618/">http://sws.geonames.org/3174618/</a> | dbpedia:City_region= <a href="http://dbpedia.org/resource/Lombardy">http://dbpedia.org/resource/Lombardy</a>                     | $r_1 = r_2$       |
| 9  | geonames:featureCode=geonames:S.SCH   | rdf:type=dbpedia:EducationalInstitution  | $r_1 = r_2$       |
| 10 | geonames:featureCode=geonames:S.SCH &<br>geonames:inCountry=geonames:US                                 | rdf:type=dbpedia:EducationalInstitution  | $r_1 = r_2$       |
| 11 | geonames:featureCode=geonames:T.MT  | rdf:type=dbpedia:Mountain  | $r_1 \subset r_2$ |

# Outline of the Talk

- Linked Open Data
  - Building and linking ontologies of linked data
- **Linked Open Services**
  - Building semantic web services from the Deep Web
- Discussion
  - Remaining challenges

# Building semantic web services from the Deep Web [Ambite et al., ISWC 2009]

- Automatically build semantic models for data and services available on the larger Web
- Construct models of these sources that are sufficiently rich to support querying and integration
  - Build models for the vast amount of structured and semi-structured data available
    - Not just web services, but also form-based interfaces
    - E.g., Weather forecasts, flight status, stock quotes, currency converters, online stores, etc.
  - Learn models for information-producing web sources and web services

- Start with an some initial knowledge of a domain
  - Sources and semantic descriptions of those sources
- **Automatically**
  - Discover related sources
  - Determine how to invoke the sources
  - Learn the syntactic structure of the sources
  - Identify the semantic types of the data
  - Build semantic models of the source

Washington, District of Columbia (20502) Conditions & Forecast : Weather Underground

file:///Users/tar/Projects/Calo/SourceDiscovery/icdm-wunderground-1.html

Welcome to Weather Underground! [Sign In](#) or [Create an Account](#). Edit my [Page Preferences](#). Other Wunders: [Mobile](#) - [iPhone](#) - [Lite](#) - [Download](#)

Search:  City, State, Zip, Airport Code, or Country

Features: [Tropical/Hurricane](#) [NEXRAD Radar](#) [Zoom Satellite](#) [Ski/Snow](#) [Marine](#) [Climate Change](#) [Tornadoes](#) [WX Radio](#) [Sports](#)  
[Weather Stations](#) [Regional Radar](#) [Severe](#) [WunderBlog](#) [WunderPhotos](#) [Trip Planner](#) [History Data](#) [Webcams](#) [Maps](#)

**Washington, District of Columbia** [Add to My Favorites](#) - [ICAL](#) [RSS](#)

Local Time: 1:07 PM EST — [Set My Timezone](#) Lat/Lon: 38.9° N 77.0° W (Google Map)

Tropical Weather: [Invest 96](#) (North Atlantic)

**Current Conditions**

Eckington Pl, NE, Washington, District of Columbia (PWS)  
Updated: 1:06 PM EST on November 25, 2008

**46.8 °F / 8.2 °C**  
**Mostly Cloudy**

Windchill: 43 °F / 6 °C  
Humidity: 41%  
Dew Point: 24 °F / -4 °C  
Wind: 8.0 mph / 12.9 km/h / 3.6 m/s from the WSW  
Wind Gust: 15.0 mph / 24.1 km/h / 9.3 m/s  
Pressure: 29.78 in / 1008.4 hPa (Steady)  
Visibility: 10.0 miles / 16.1 kilometers  
UV: 2 out of 16  
Clouds: **Mostly Cloudy** 6000 ft / 1828 m  
**Mostly Cloudy** 14000 ft / 4267 m (Above Ground Level)  
Elevation: 90 ft / 27 m

[Radar](#) [Webcam](#)

[Click Radar to Enlarge](#)

- [Local Radar](#)
- [WunderMap](#) new!
- [Regional Radar](#)
- [Local Satellite](#)
- [Marine Forecast](#)
- [Ski Conditions](#)
- [Trip Planner](#)
- [Weather Stations](#)

**5-Day Forecast for ZIP Code 20502** [Customize Your Icons!](#)

| Tuesday                      | Wednesday                     | Thursday                       | Friday                        | Saturday                                      |
|------------------------------|-------------------------------|--------------------------------|-------------------------------|---|
|                              |                               |                                |                               |   |
| 45° F / 32° F<br>7° C / 0° C | 47° F / 31° F<br>8° C / -1° C | 50° F / 31° F<br>10° C / -1° C | 50° F / 34° F<br>10° C / 1° C | 47° F / 34° F<br>8° C / 1° C                  |
| Mostly Cloudy                | Partly Cloudy                 | Clear                          | Partly Cloudy                 | Chance of Rain<br>30% chance of precipitation |
| <a href="#">Hourly</a>       | <a href="#">Hourly</a>        | <a href="#">Hourly</a>         | <a href="#">Hourly</a>        | <a href="#">Hourly</a>                        |

Today is forecast to be **Cooler** than yesterday.

**Forecast for District of Columbia** [↑](#)

Updated: 10:48 am EST on November 25, 2008

Active Notice: [Public Information Statement](#) (US Severe Weather)

**Rest of Today**  
Becoming partly sunny. Highs in the upper 40s. West winds 10 to 15 mph with gusts up to 25 mph.  
[ZIP Code Detail](#)

**Tonight**  
Mostly cloudy. Lows in the lower 30s. Southwest winds 10 to 15 mph.

**Wednesday**  
Partly sunny. Highs in the upper 40s. West winds 10 to 15 mph.  
[ZIP Code Detail](#)

**Weather Underground**

You don't appear to have any favorites yet, or your cookies may be disabled.

[Edit My Favorites](#)

**WunderPhotos**

- Washington
- District of Columbia

[Browse All Photos](#)

**WunderMap**

[View WunderMap](#)

**Website Spotlight**

- [Weather Maps](#)
- [Solar Calculator](#) new!
- [Forecast Flyer](#)
- [Community Chat](#)
- [Education](#)
- [Astronomy](#)
- [Print This Page](#)
- [Developer's Blog](#)



# Automatically Discover and Build Semantic Web Services for Related Sources

Unisys Weather

Imagine it. done.

Unisys Home Page  
 Unisys Transportation  
 Weather Solutions  
**Unisys Weather**  
 Home Information Contents  
 Analyses  
 Forecasts  
 Miscellaneous

ES7000 Servers  
 True Flexibility

Unisys Internet Weather Data  
 Unisys NOAAPORT Solutions

002-11-DEC 00

Current satellite image and surface map (Click on map for forecast) [loop]

Visible Satellite Image Enh IR Satellite Image Satellite Surface Map  
 US Radar Summary NAM Model Forecast GFSx.10 day Forecast

NEWS  
 FAQ  
 First Time User  
 Guest Book

The intent of this weather site is to provide a complete source of graphical weather information. This is intended to satisfy the needs of the weather professional but can be a tool for the casual user as well. The graphics and data are displayed as a meteorologist would expect to see. For the novice user, there are detailed explanation pages to guide them through the various plots, charts and images. The data on this site are provided from the [National Weather Service](#) via the [NOAAPORT](#) satellite data service. All the images are generated using the [Weather Processor \(WXP\)](#) analysis package which is available from Unisys.

© Unisys Corp. 2006  
 - For questions and information on this server, NOAAPORT and WXP, contact [Dan Victor at devo@ks.unisys.com](mailto:Dan.Victor@devo@ks.unisys.com)  
 - For sales information on Unisys weather solutions, contact [Robert Benedict at robert.benedict@unisys.com](mailto:Robert.Benedict@robert.benedict@unisys.com)  
 - Last modified February 7, 2007

Unisys Weather: Forecast for Washington, DC (20502) [0] 2

Unisys Home Page  
 Unisys Transportation  
 Weather Solutions  
**Unisys Weather**  
 Home Information Contents  
 Analyses  
 Forecasts  
 Miscellaneous

Latest Observation for Washington, DC (20502)  
 Partly Cloudy  
 Site: KDCa (Washington/Natl, VA)  
 Time: 4 PM EST 25 NOV 08  
 Temp: 45 F (7 C)  
 Dewpt: 22 F (-5 C)  
 Rel Hum: 40%  
 Winds: W at 7 knt  
 Wind chill: 41 F  
 Pressure: 1010.1 mb (29.94 in)  
 Visibility: 10 mi  
 Skies: partly cloudy  
 Weather:

Almanac  
 Sunrise: 7:02 AM  
 Sunset: 4:48 PM

Alerts  
 No alerts

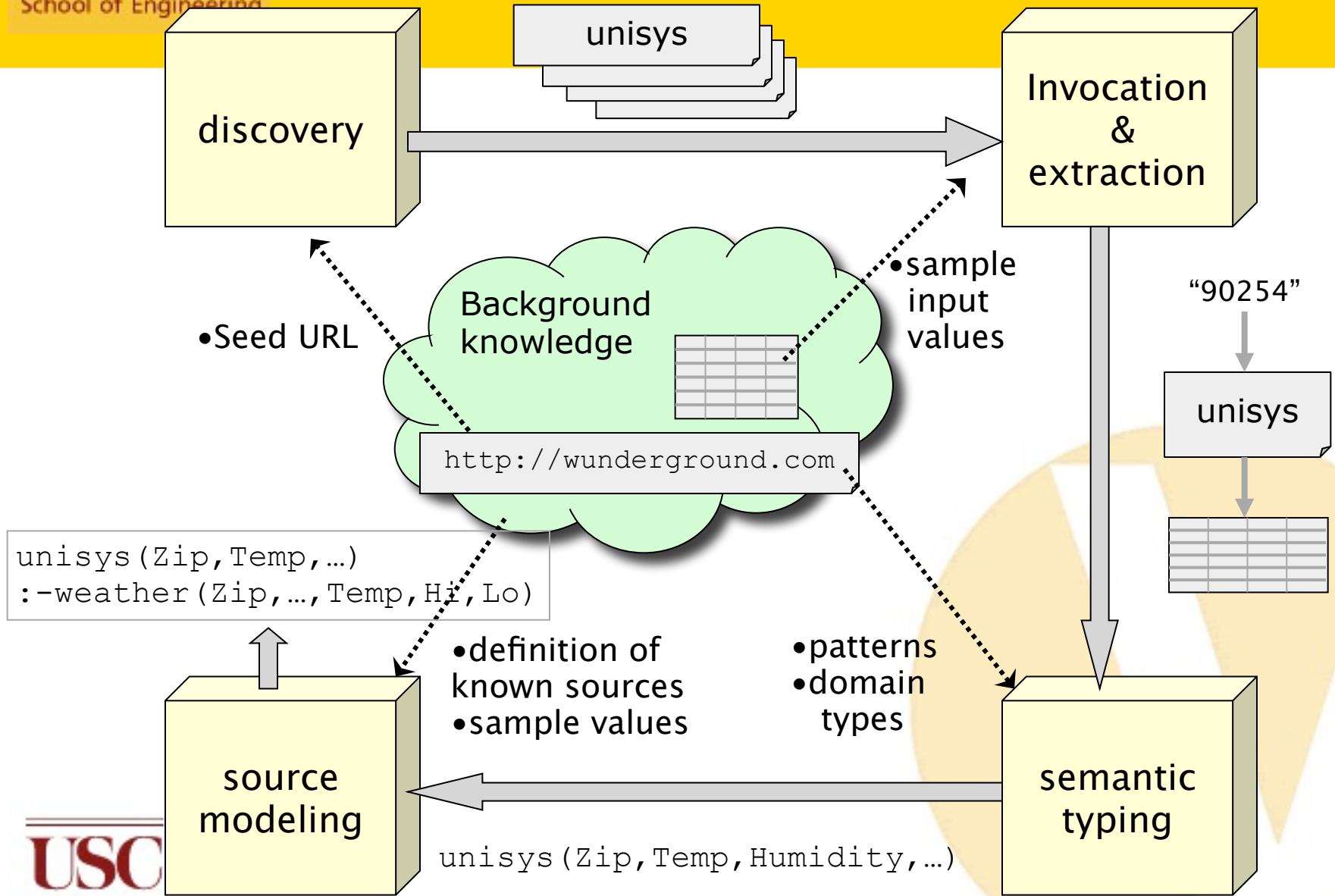
Forecast Summary

| WEDNESDAY        | THURSDAY         | FRIDAY           | SATURDAY         | SUNDAY           | MONDAY           | TUESDAY          |
|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
| Sunny            | Sunny            | Rainy            | Sunny            | Sunny            | Sunny            | Sunny            |
| Hi: 45<br>Lo: 32 | Hi: 52<br>Lo: 35 | Hi: 52<br>Lo: 35 | Hi: 48<br>Lo: 35 | Hi: 48<br>Lo: 35 | Hi: 45<br>Lo: 32 | Hi: 45<br>Lo: 32 |

Detailed forecast from National Weather Service  
 DISTRICT OF COLUMBIA-ARLINGTON-FALLS CHURCH-ALEXANDRIA-  
 INCLUDING THE CITIES OF... WASHINGTON...ALEXANDRIA...FALLS CHURCH  
 308 PM EST TUE NOV 25 2008

|         |   |
|---------|---|
| TONIGHT | LO: 32<br>MOSTLY CLOUDY. LOWS IN THE LOWER 30S. SOUTHWEST WINDS AROUND 10 MPH.            |
| Sunny   | WEDNESDAY<br>Hi: 45<br>MOSTLY SUNNY. HIGHS IN THE MD 40S. WEST WINDS 10 TO 15 MPH.        |
| Sunny   | WEDNESDAY NIGHT<br>LO: 35<br>PARTLY CLOUDY. LOWS IN THE MD 30S. WEST WINDS 5 TO 10 MPH.   |
| Sunny   | THANKSGIVING DAY<br>Hi: 52<br>SUNNY. HIGHS IN THE LOWER 50S. SOUTHWEST WINDS 5 TO 10 MPH. |
| Sunny   | THURSDAY NIGHT<br>LO: 35<br>PARTLY CLOUDY. LOWS IN THE MD 30S. SOUTH WINDS AROUND 5 MPH.  |
| Rainy   | FRIDAY<br>Hi: 52  |

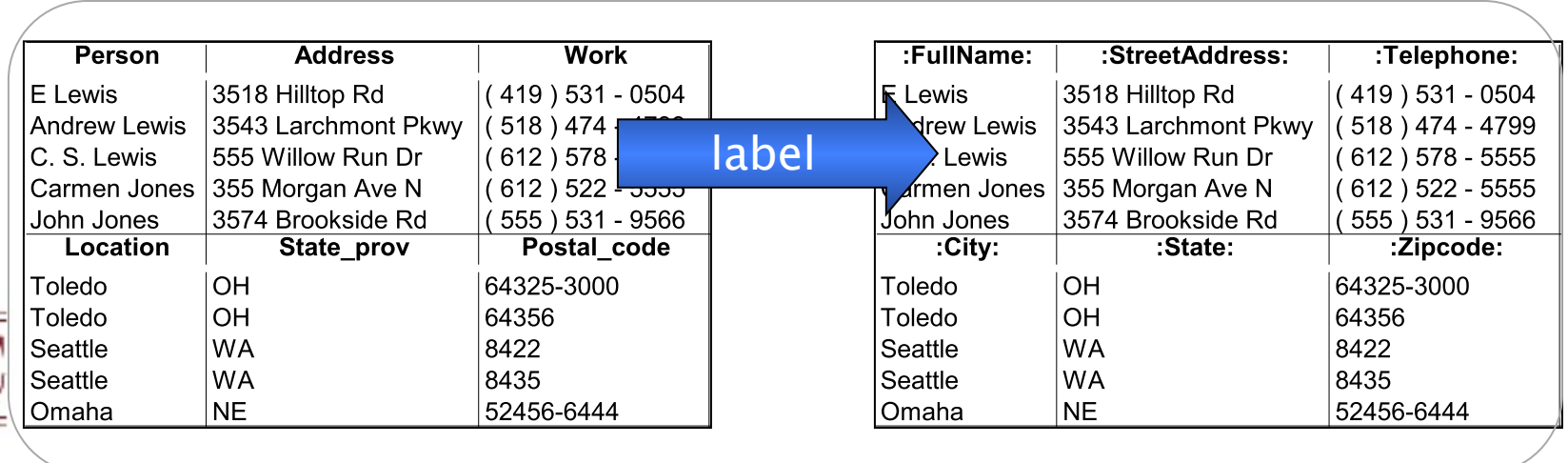
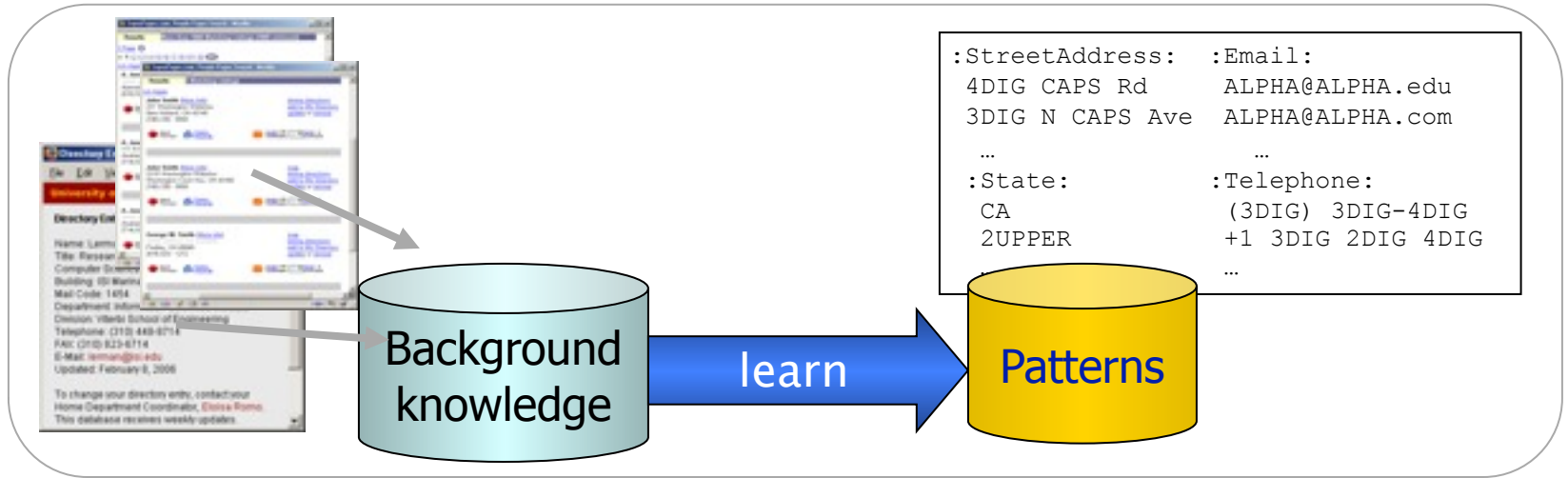
# Integrated Approach



# Semantic Typing

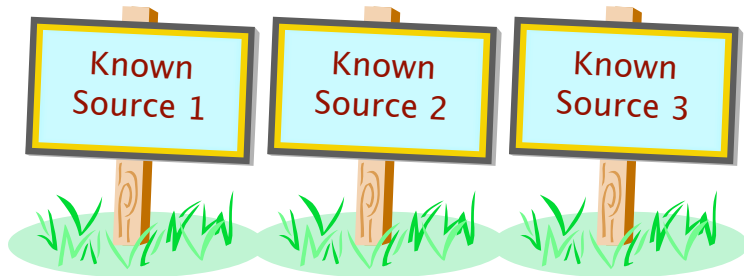
## [Lerman, Plangprasopchok, & Knoblock]

✓ Idea: Learn a model of the content of data and use it to recognize new examples





# Inducing Source Definitions

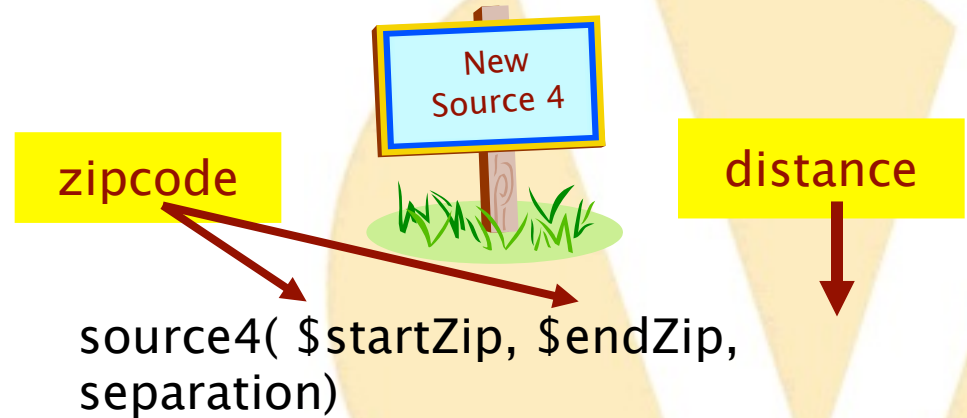


source1(\$zip, lat, long) :-  
centroid(zip, lat, long).

source2(\$lat1, \$long1, \$lat2, \$long2, dist) :-  
greatCircleDist(lat1, long1, lat2, long2, dist).

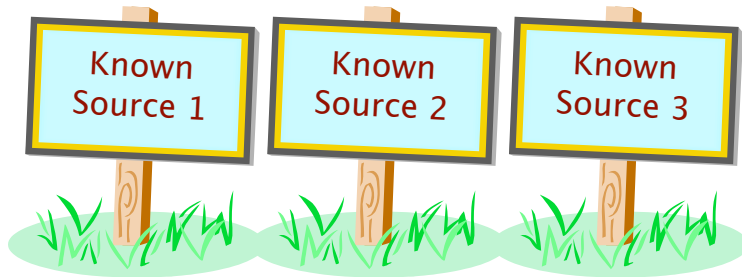
source3(\$dist1, dist2) :-  
convertKm2Mi(dist1, dist2).

- Step 1: classify input & output semantic types



# Generating Plausible Definition

[Carman & Knoblock, 2007]



```
source1($zip, lat, long) :-  
  centroid(zip, lat, long).
```

```
source2($lat1, $long1, $lat2, $long2, dist) :-  
  greatCircleDist(lat1, long1, lat2, long2, dist).
```

```
source3($dist1, dist2) :-  
  convertKm2Mi(dist1, dist2).
```

- Step 1: classify input & output semantic types
- Step 2: generate plausible definitions

```
source4($zip1, $zip2, dist):-  
  source1(zip1, lat1, long1),  
  source1(zip2, lat2, long2),  
  source2(lat1, long1, lat2, long2, dist2),  
  source3(dist2, dist).
```

```
source4($zip1, $zip2, dist):-  
  centroid(zip1, lat1, long1),  
  centroid(zip2, lat2, long2),  
  greatCircleDist(lat1, long1, lat2, long2, dist2),  
  convertKm2Mi(dist1, dist2).
```

# Invoke and Compare the Definition

- Step 1: classify input & output semantic types
- Step 2: generate plausible definitions
- Step 3: invoke service & compare output

```
source4($zip1, $zip2, dist):-
  source1(zip1, lat1, long1),
  source1(zip2, lat2, long2),
  source2(lat1, long1, lat2, long2, dist2),
  source3(dist2, dist).
```

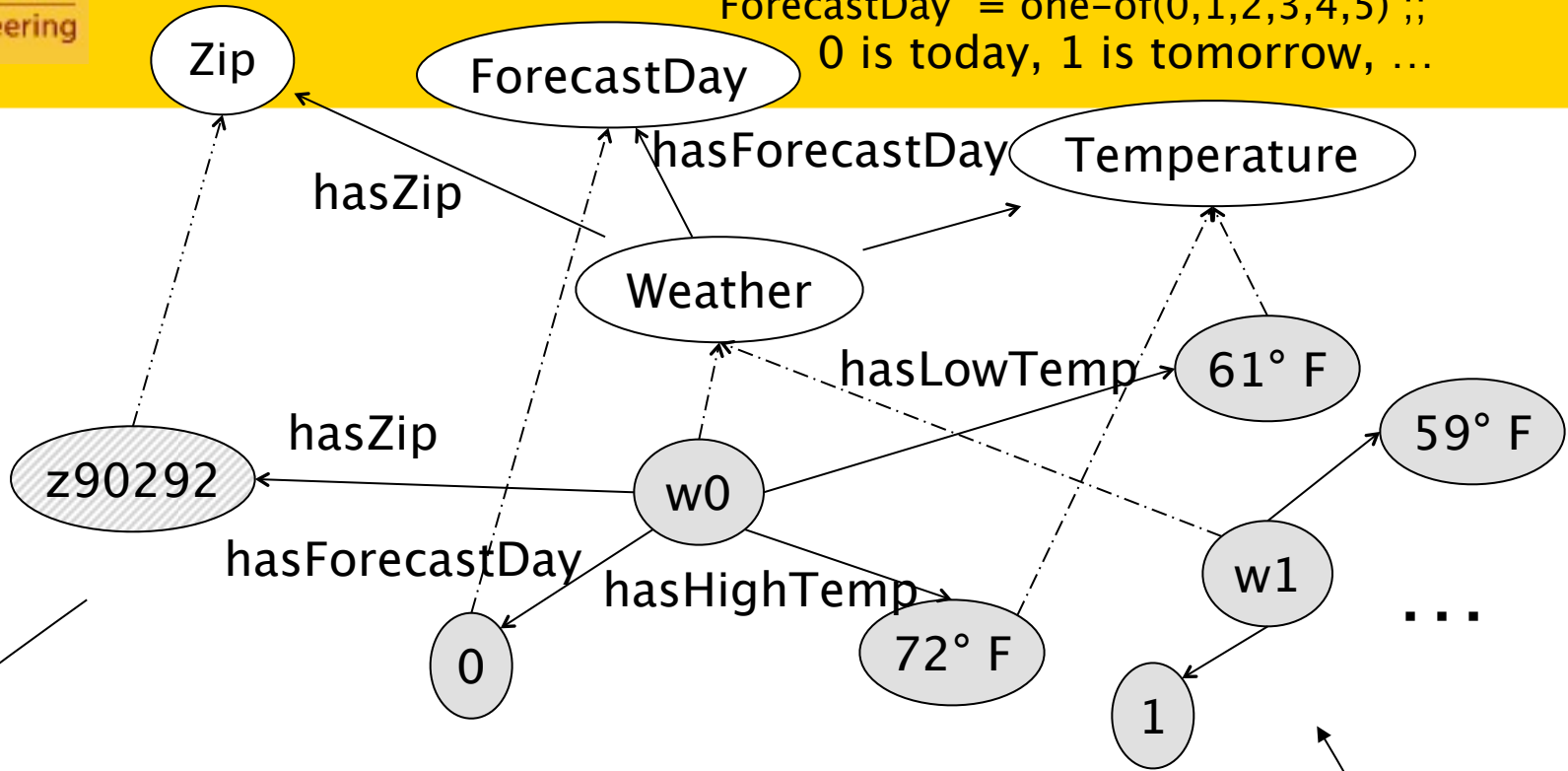
```
source4($zip1, $zip2, dist):-
  centroid(zip1, lat1, long1),
  centroid(zip2, lat2, long2),
  greatCircleDist(lat1, long1, lat2,
  long2, dist2),
  convertKm2Mi(dist1, dist2).
```



| \$zip1 | \$zip2 | dist<br>(actual) | dist<br>(predicted) |
|--------|--------|------------------|---------------------|
| 80210  | 90266  | 842.37           | 843.65              |
| 60601  | 15201  | 410.31           | 410.83              |
| 10005  | 35555  | 899.50           | 899.21              |

# Constructing Semantic Web Services

ForecastDay = one-of(0,1,2,3,4,5) ;;  
0 is today, 1 is tomorrow, ...

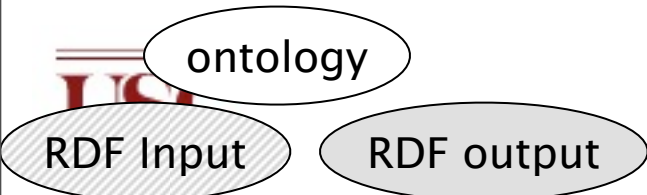


z90292 hasName 90292 .

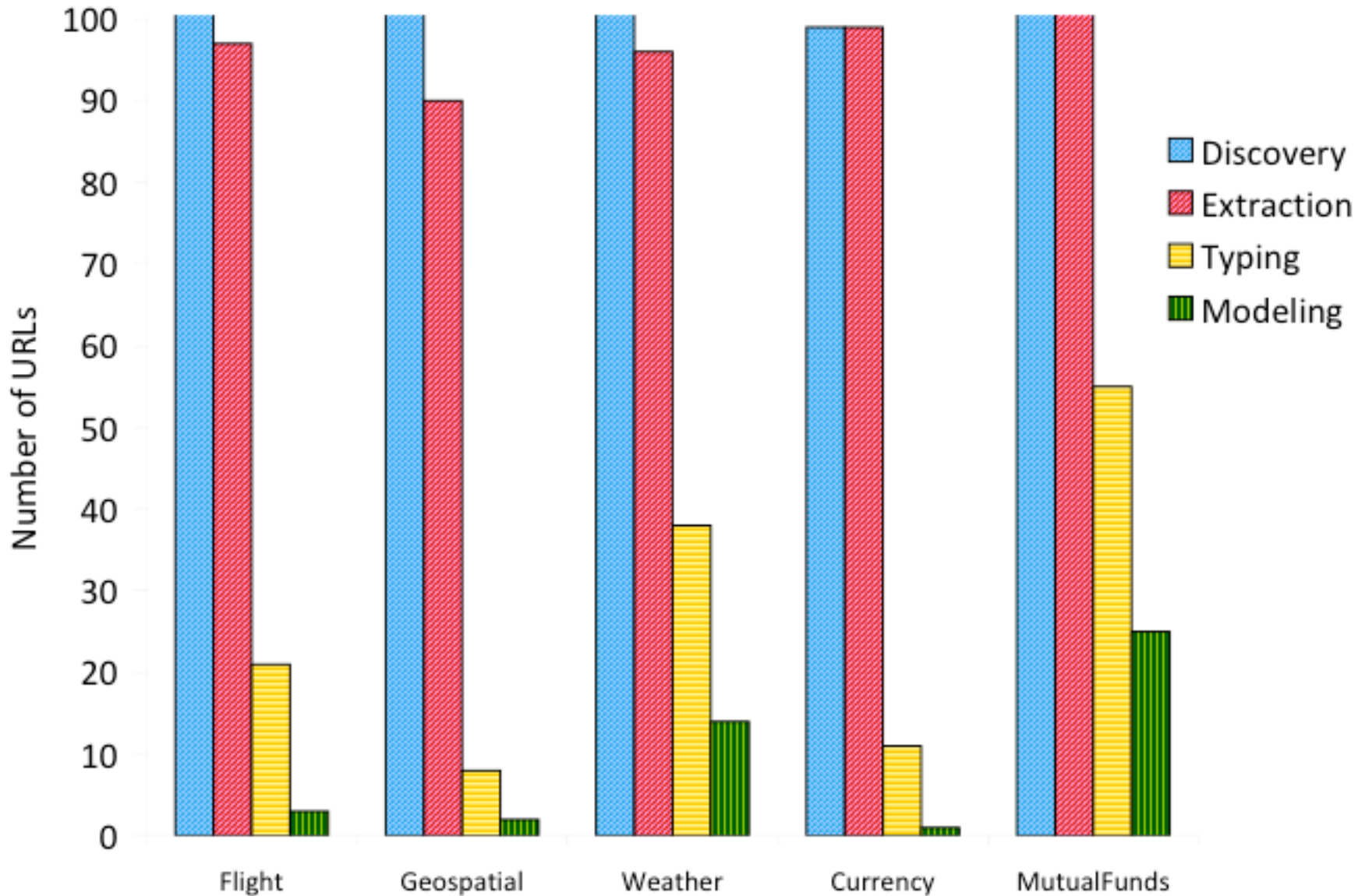
DEIMOS generated Web Service

w1 hasZIP z90292 .  
w1 hasTemp 61° F .  
...  
w1 hasZIP z90292 .  
w2 hasLowTemp 59° F .

Legend:



# Evaluation on Multiple Domains



# Accuracy of the Models

| domain            | Precision | Recall | F <sub>1</sub> -measure |
|-------------------|-----------|--------|-------------------------|
| <i>weather</i>    | 0.64      | 0.29   | 0.39                    |
| <i>geospatial</i> | 1.00      | 0.86   | 0.92                    |
| <i>flights</i>    | 0.69      | 0.35   | 0.46                    |
| <i>currency</i>   | 1.00      | 1.00   | 1.00                    |
| <i>mutualfund</i> | 0.72      | 0.30   | 0.42                    |

# Outline of the Talk

- Linked Open Data
  - Building and linking ontologies of linked data
- Linked Open Services
  - Building semantic web services from the Deep Web
- **Discussion**
  - Remaining challenges

- Initial work described here just scratches the surface of the problem
  - Goal is to both populate the Web of linked data and have rich semantic models of the data
  - Building semantic descriptions of linked open data will allow us to better understand the available sources and use the sources in a broad range of applications
  - Methods for automatically constructing linked open services will improve the coverage and quality of the sources available



# Some Challenges

- **Linked Open Data**
  - How do we build an overall class hierarchy for a source
  - How do the relations map across sources
  - What do we do about missing and extraneous links
- **Linked Open Services**
  - How do we improve the accuracy of the learned semantic descriptions
  - How can we learn semantic descriptions that go beyond the current sources
  - How do we learn mappings between enumerated types (e.g., “Arrived” vs. “Landed”)