



**USC Viterbi**  
School of Engineering

# Automatically Constructing Semantic Web Services from Online Sources

**Craig A. Knoblock**

**José Luis Ambite, Sirish Darbha, Aman Goel,  
Kristina Lerman, Rahul Parundekar, and Tom Russ**

**University Southern California**



# Goal

- 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
  - Such models would make the existing semantic web tools and techniques more widely applicable
- Current focus:
  - 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
  - Construct semantic web services



- Integrated Approach
  - Discovering related sources
  - Constructing syntactic models of the sources
  - Determining the semantic types of the data
  - Building semantic models of the sources
- Experimental Results
- Related Work
- Discussion

# Seed Source

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

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

Twiki APIs Apple (125) TinyURL! Zip PL-GUI Heracles GoogleGroups Mantis Shop Popular News (1368) CAL-FIRE

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

Search:


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

**Washington, District of Columbia** [Add to My Favorites](#) - [TCAL](#) [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 7° C   0° C	47° F   31° F 8° C   -1° C	50° F   31° F 10° C   -1° C	50° F   34° F 10° C   1° C	47° F   34° F 8° C   1° C
Mostly Cloudy	Partly Cloudy	Clear	Partly Cloudy	Chance of Rain 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](#)

# Automatically Discover and Build Semantic Web Services for Related Sources

Unisys Weather

UNISYS  
imagine it. done.

Unisys Home Page  
Unisys Transportation  
Weather Solutions  
**Unisys Weather**  
Home  
Information  
Contents  
Analyses  
Satellite Images  
Surface Data  
Upper Air Data  
Radar Data  
Forecasts  
Model Statistics  
NGM Model  
NAM/Wrt Model  
GFS/Avn Model  
GFSx/MRF Model  
RUC Model  
ECMWF Model  
Miscellaneous  
Hurricane Data  
Archive of Images  
USGS Maps

ES7000 Servers  
True Flexibility

unisys Internet Weather Data  
unisys NOAAPORT Solutions

00Z 11 DEC 08

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. 2005  
- For questions and information on this server, NOAAPORT and WXP, contact [Dan Vietor at devo@ks.unisys.com](mailto:Dan Vietor at devo@ks.unisys.com)  
- For sales information on Unisys weather solutions, contact [Robert Benedict at robert.benedict@unisys.com](mailto:Robert Benedict at robert.benedict@unisys.com)  
- Last modified February 7, 2007

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

Unisys Weather

Unisys Home Page  
Unisys Transportation  
Weather Solutions  
**Unisys Weather**  
Home  
Information  
Contents  
Analyses  
Satellite Images  
Surface Data  
Upper Air Data  
Radar Data  
Forecasts  
Model Statistics  
NGM Model  
NAM/Wrt Model  
GFS/Avn Model  
GFSx/MRF Model  
RUC Model  
ECMWF Model  
Miscellaneous  
Hurricane Data  
Archive of Images  
USGS Maps

Enter a zip code or city name to get forecast:

Latest Observation for Washington, DC (20502)  
Partly Cloudy Site: KDCa (Washington/Nati, VA) Almanac  
Time: 4 PM EST 25 NOV 08 Sunrise: 7:02 AM  
Temp: 45 F (7 C) Dewpt: 22 F (-5 C) Sunset: 4:48 PM  
Rel Hum: 40% Winds: W at 7 knt  
Wind chill: 41 F Pressure: 1010.1 mb (29.84 in)  
Visibility: 10 mi Skies: partly cloudy  
Weather:

Alerts  
No alerts

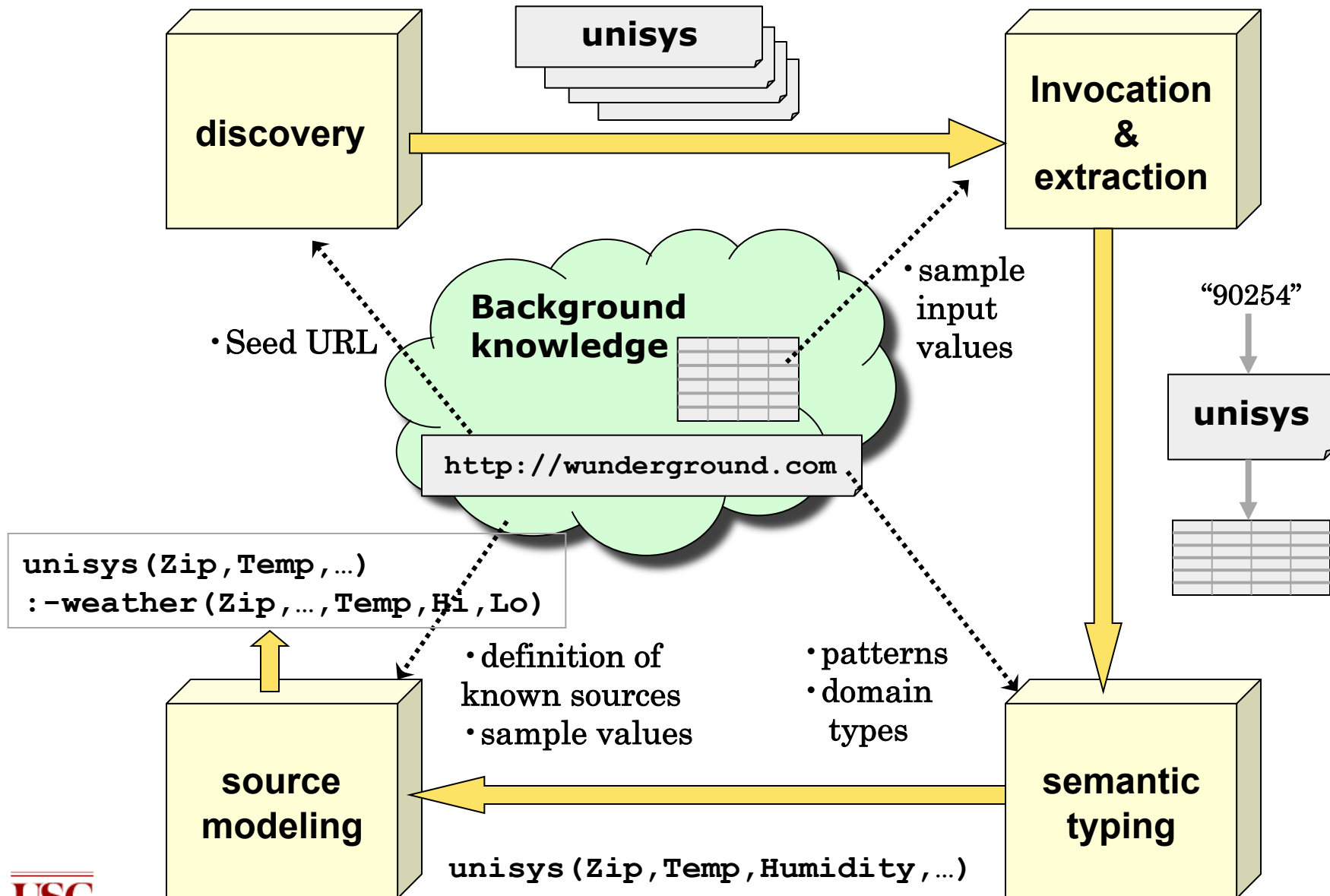
Forecast Summary

WEDNESDAY	THURSDAY	FRIDAY	SATURDAY	SUNDAY	MONDAY	TUESDAY
Sunny	Sunny	Rainy	Sunny	Sunny	Sunny	Sunny
HI: 45 LO: 32	HI: 52 LO: 35	HI: 52 LO: 35	HI: 48 LO: 35	HI: 48 LO: 35	HI: 45 LO: 32	HI: 45 LO: 32

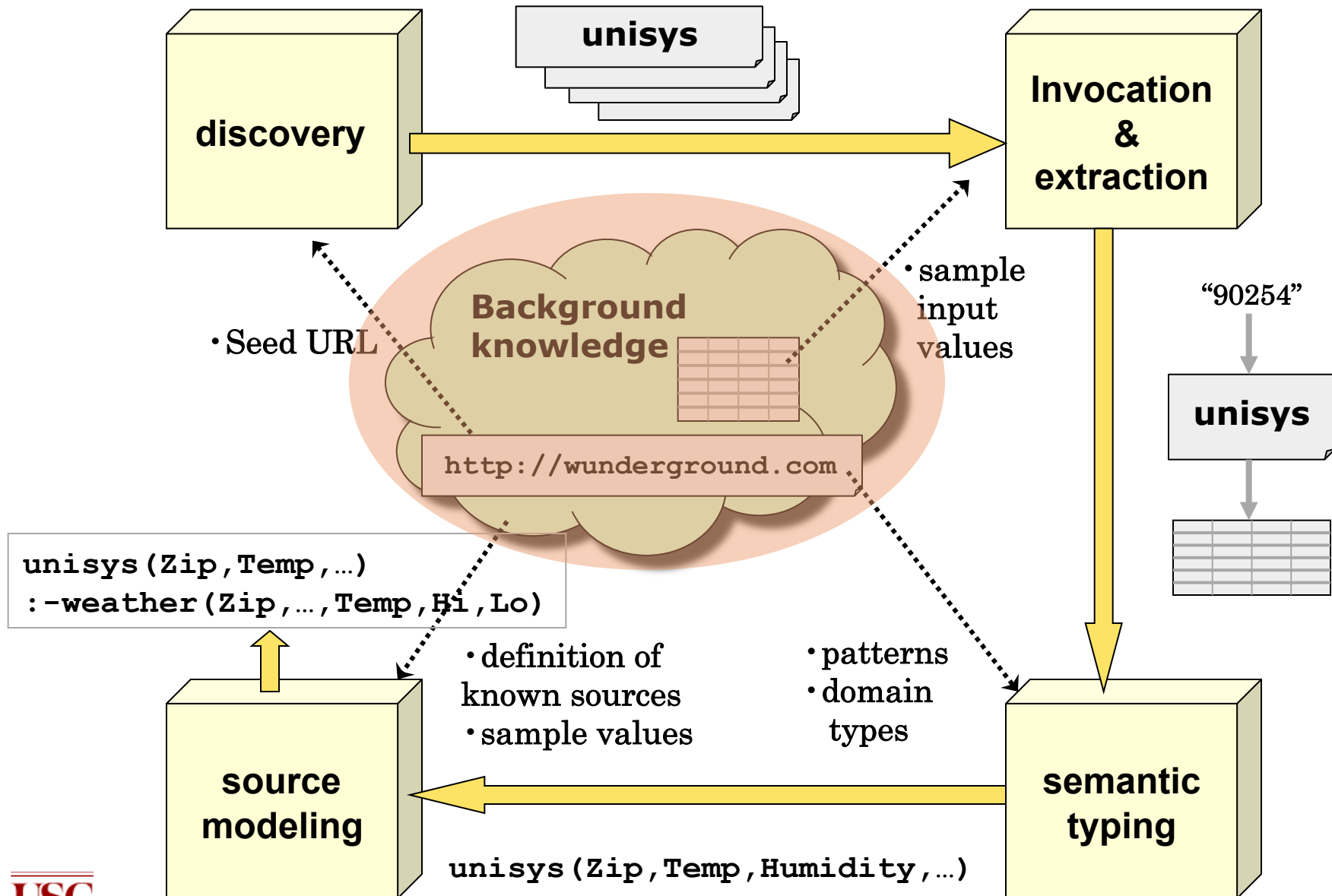
Detailed forecast from National Weather Service  
DISTRICT OF COLUMBIA-ARLINGTON/FALLS CHURCH/ALEXANDRIA-  
INCLUDING THE CITIES OF...WASHINGTON...ALEXANDRIA...FALLS CHURCH  
306 PM EST TUE NOV 25 2008

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

# Integrated Approach



# Background Knowledge

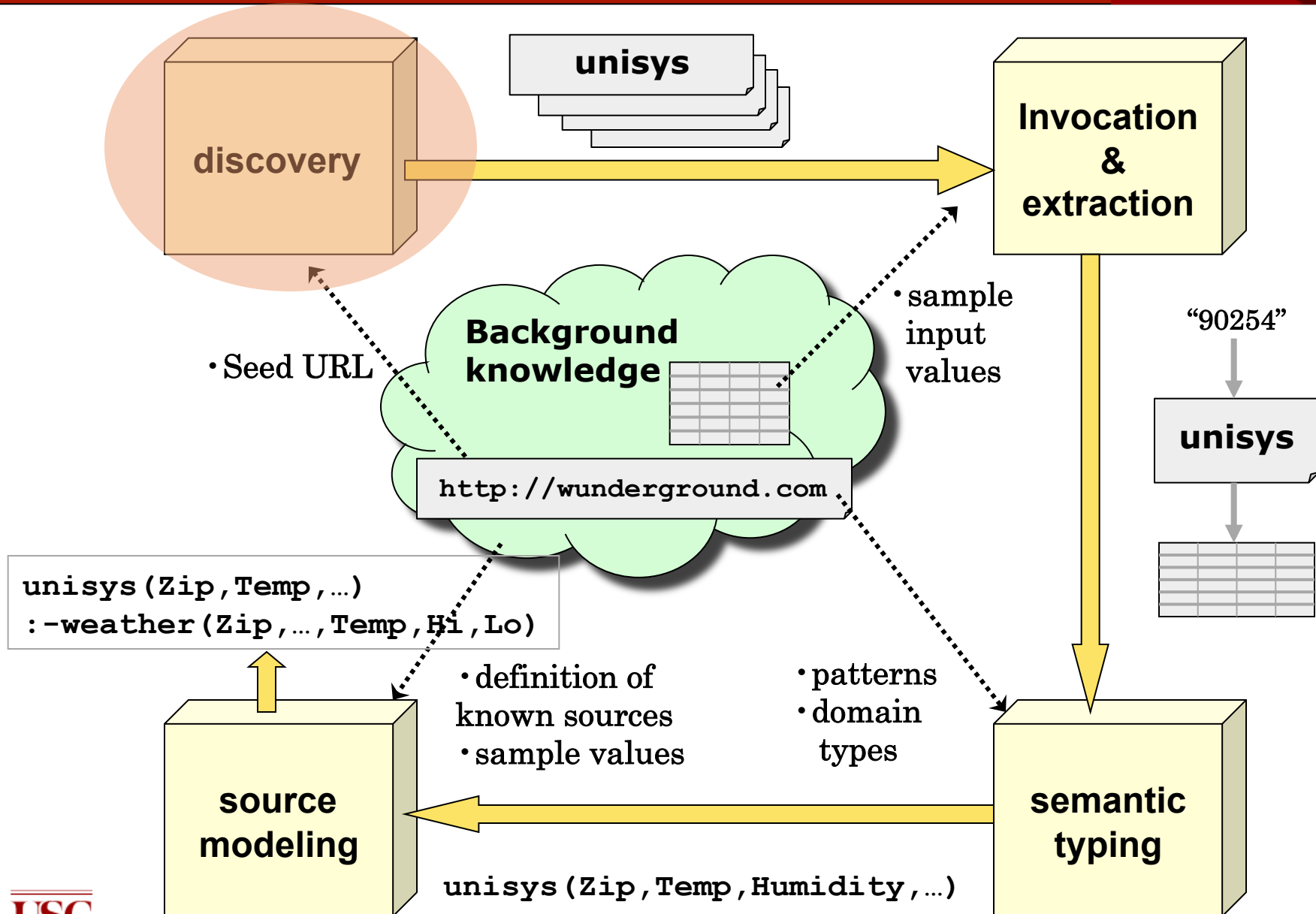




# Background Knowledge

- Ontology of the inputs and outputs
  - e.g., TempF, Humidity, Zipcode;
- Sample values for each semantic type
  - e.g., "88 F" for TempF, and "90292" for Zipcode
- Domain input model
  - a weather source may accept Zipcode or City and State as input
  - Sample input values
- Known sources (seeds)
  - e.g., <http://wunderground.com>
- Source descriptions in Datalog or RDF
  - wunderground(\$Z,CS,T,F0,S0,Hu0,WS0,WD0,P0,V0,FL1,FH1,S1,FL2,FH2,S2,FL3,FH3,S3,FL4,FH4,S4,FL5,FH5,S5) :-  
weather(0,Z,CS,D,T,F0,\_,\_,S0,Hu0,P0,WS0,WD0,V0)  
weather(1,Z,CS,D,T,\_,FH1,FL1,S1,\_,\_,\_,\_,\_),  
weather(2,Z,CS,D,T,\_,FH2,FL2,S2,\_,\_,\_,\_,\_),  
weather(3,Z,CS,D,T,\_,FH3,FL3,S3,\_,\_,\_,\_,\_),  
weather(4,Z,CS,D,T,\_,FH4,FL4,S4,\_,\_,\_,\_,\_),  
weather(5,Z,CS,D,T,\_,FH5,FL5,S5,\_,\_,\_,\_,\_).

# Source Discovery



# Source Discovery [Plangprasopchok and Lerman]

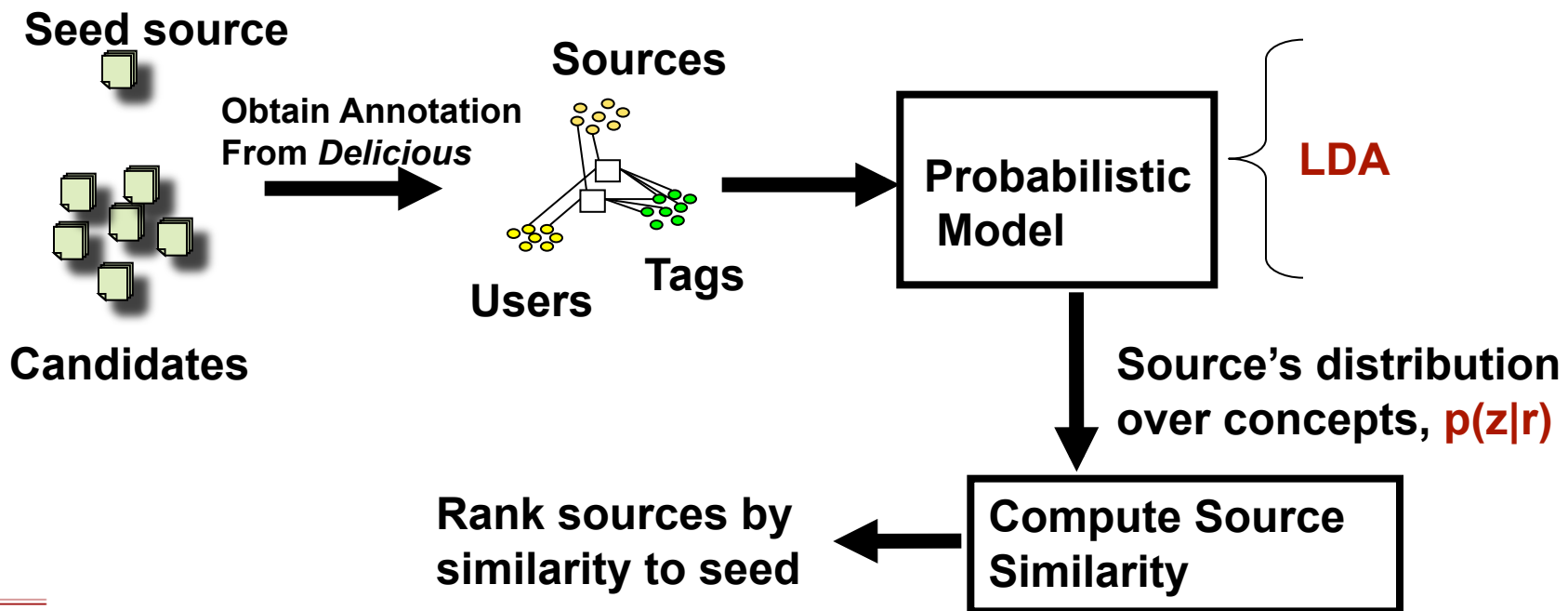
- Leverage user-generated tags on the social bookmarking site del.icio.us to discover sources similar to the seed

The screenshot shows a browser window displaying a del.icio.us bookmark page for 'The Weather Underground'. The page includes a search bar, navigation tabs, and a list of user-specified tags. A 'Top 10 Tags' list is also visible, with an arrow pointing to it from the text 'Most common tags'. Another arrow points to a specific tag 'reference' in the user-specified tags list from the text 'User-specified tags'.

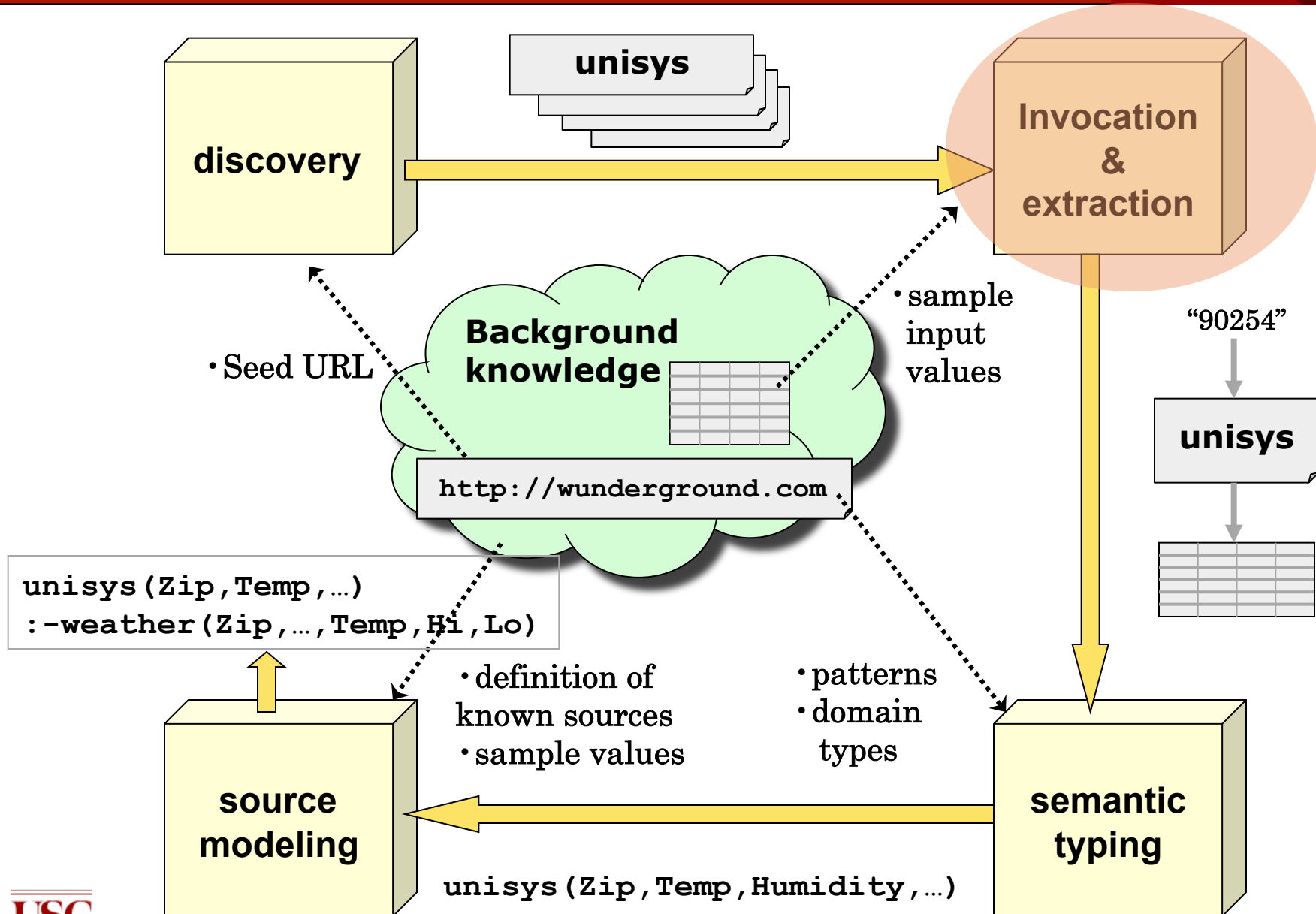
Tag	Count
weather	2314
forecast	536
travel	417
reference	386
news	285
tools	213
science	200
maps	124
world	62
meteo	53

# Exploiting Social Annotations for Resource Discovery

- Resource discovery task : "*given a seed source, find other most similar sources*"
  - Gather a corpus of <user, source, tag> bookmarks from del.icio.us
  - Use probabilistic modeling to find hidden topics in the corpus
  - Rank sources by similarity to the seed within topic space



# Source Invocation & Extraction



# Target Source Invocation

- To invoke the target source, we need to locate the form and determine the appropriate input values
  1. Locate the form
  2. Try different data type combinations as input
    - *For weather, only one input - location, which can be zipcode or city/state*
  3. Submit Form
  4. Keep successful invocations

Form  
Input

The screenshot shows a web browser window with the URL <http://weather.unisys.com/>. The page features a navigation menu on the left with categories like 'Home', 'Analyses', 'Forecasts', and 'Miscellaneous'. A central map of the United States displays weather data, including pressure systems and precipitation. Below the map, there is a search form with the text 'Enter a zip code or city name to get forecast:' and a 'GO' button. The page also includes links for 'Visible Satellite Image', 'US Radar Summary', and 'NEWS'. At the bottom, there is a copyright notice for Unisys Corp. 2005 and contact information for Dan Vietor and Robert Benedict.

# Inducing Extraction Templates

- Template: a sequence of alternating slots and stripes
  - stripes are the common substrings among all pages
  - slots are the placeholders for data
- Induction: Stripes are discovered using the Longest Common Subsequence algorithm

Sample Page 1

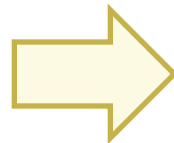
```
<br>
<font face="Arial, Helvetica, sans-serif">
  <small><b>Temp: 72F (22C)</b></small></font>
<font face="Arial, Helvetica, sans-serif">
  <small>Site: <b>KSMO (Santa_Monica_Mu, CA)</b><br>
    Time: <b>11 AM PST 10 DEC 08</b>
```



Sample Page 2

```
<br>
<font face="Arial, Helvetica, sans-serif">
  <small><b>Temp: 37F (2C)</b></small></font>
<font face="Arial, Helvetica, sans-serif">
  <small>Site: <b>KAGC (Pittsburgh/Alle, PA)</b><br>
    Time: <b>2 PM EST 10 DEC 08</b>
```

Induction



Template

Slot

Stripe

```
<br>
<font face="Arial, Helvetica, sans-serif">
  <small><b>Temp: * (*)</b></small></font>
<font face="Arial, Helvetica, sans-serif">
  <small>Site: <b>* (*, *)</b><br>
    Time: <b>* 10 DEC 08</b>
```

# Data Extraction with Templates

- To extract data: Find data in slots by locating the stripes of the template on unseen page:

## Unseen Page

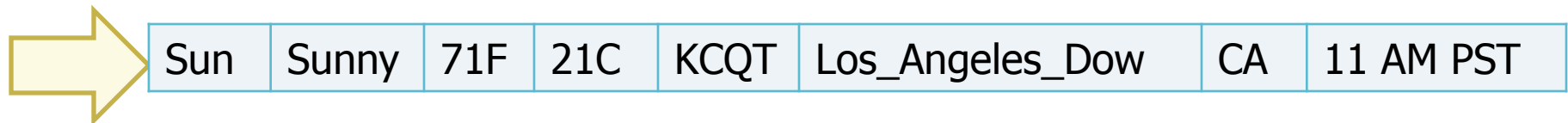
```
<br>
<font face="Arial, Helvetica, sans-serif">
  <small><b>Temp: 71F (21C)</b></small></font>
<font face="Arial, Helvetica, sans-serif">
  <small>Site: <b>KCQT (Los_Angeles_Dow, CA)</b><br>
    Time: <b>11 AM PST 10 DEC 08</b>
```



## Induced Template

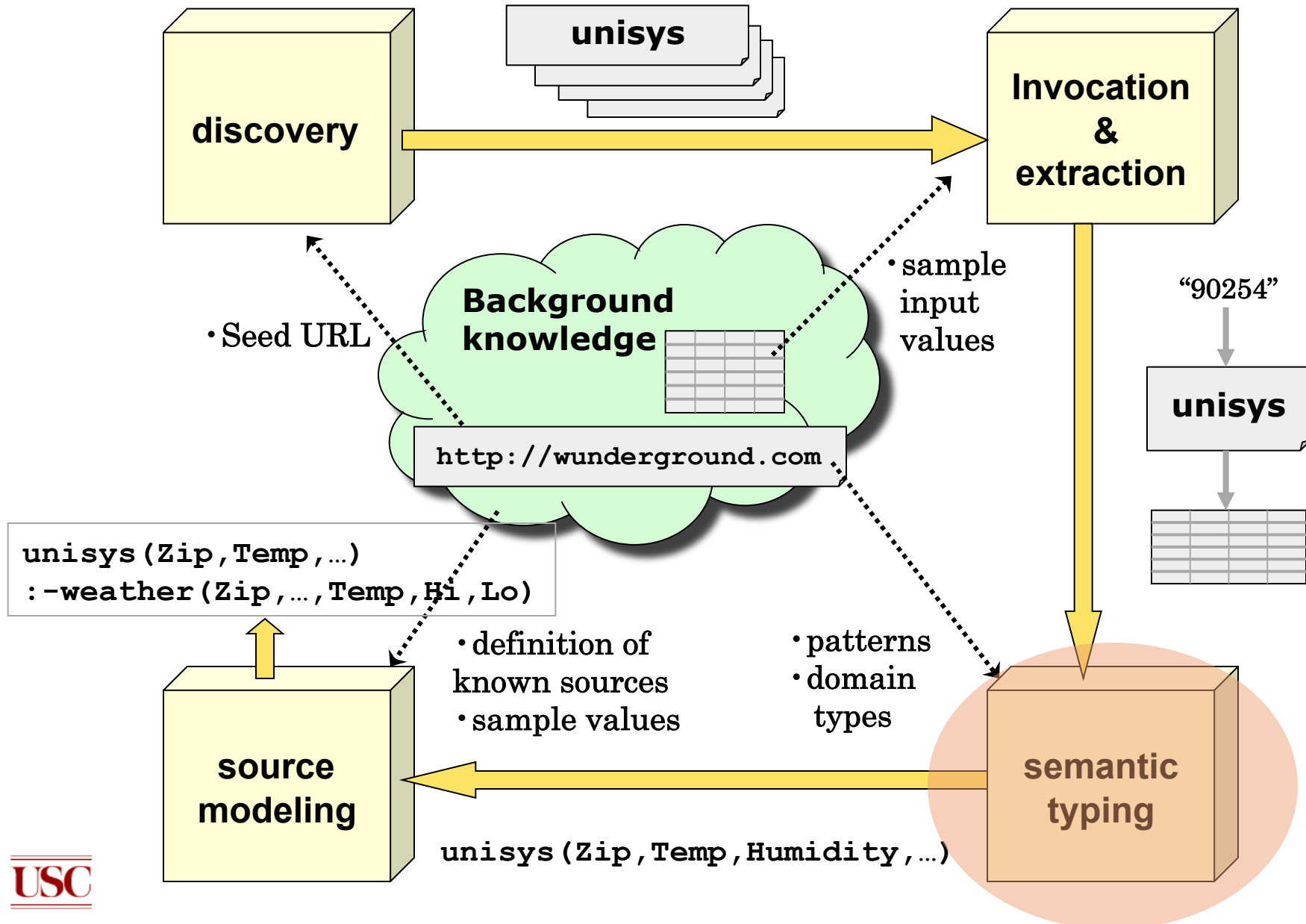
```
<br>
<font face="Arial, Helvetica, sans-serif">
  <small><b>Temp: * (*)</b></small></font>
<font face="Arial, Helvetica, sans-serif">
  <small>Site: <b>* (*, *)</b><br>
    Time: <b>* 10 DEC 08</b>
```

## Extracted Data





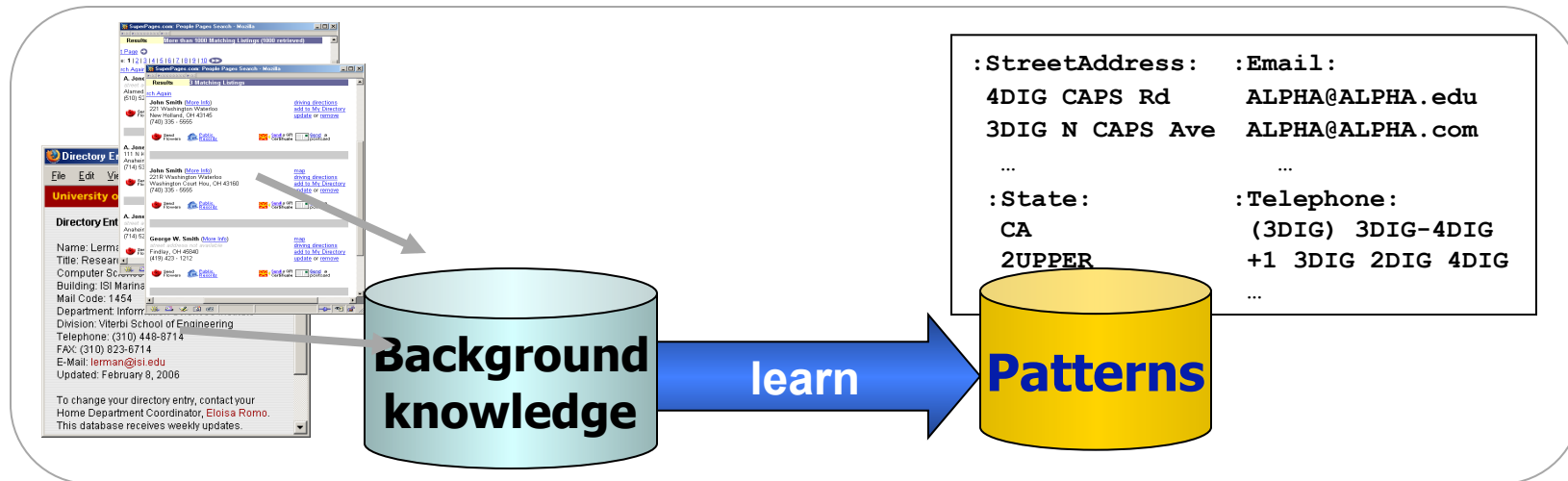
# Semantic Typing



# Semantic Typing

## [Lerman, Plangprasopchok, & Knoblock]

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



Person	Address	Work
E Lewis	3518 Hilltop Rd	( 419 ) 531 - 0504
Andrew Lewis	3543 Larchmont Pkwy	( 518 ) 474 - 4799
C. S. Lewis	555 Willow Run Dr	( 612 ) 578 - 5555
Carmen Jones	355 Morgan Ave N	( 612 ) 522 - 5555
John Jones	3574 Brookside Rd	( 555 ) 531 - 9566
Location	State_prov	Postal_code
Toledo	OH	64325-3000
Toledo	OH	64356
Seattle	WA	8422
Seattle	WA	8435
Omaha	NE	52456-6444



:FullName:	:StreetAddress:	:Telephone:
E Lewis	3518 Hilltop Rd	( 419 ) 531 - 0504
Andrew Lewis	3543 Larchmont Pkwy	( 518 ) 474 - 4799
C. S. Lewis	555 Willow Run Dr	( 612 ) 578 - 5555
Carmen Jones	355 Morgan Ave N	( 612 ) 522 - 5555
John Jones	3574 Brookside Rd	( 555 ) 531 - 9566
:City:	:State:	:Zipcode:
Toledo	OH	64325-3000
Toledo	OH	64356
Seattle	WA	8422
Seattle	WA	8435
Omaha	NE	52456-6444

# Labeling New Data

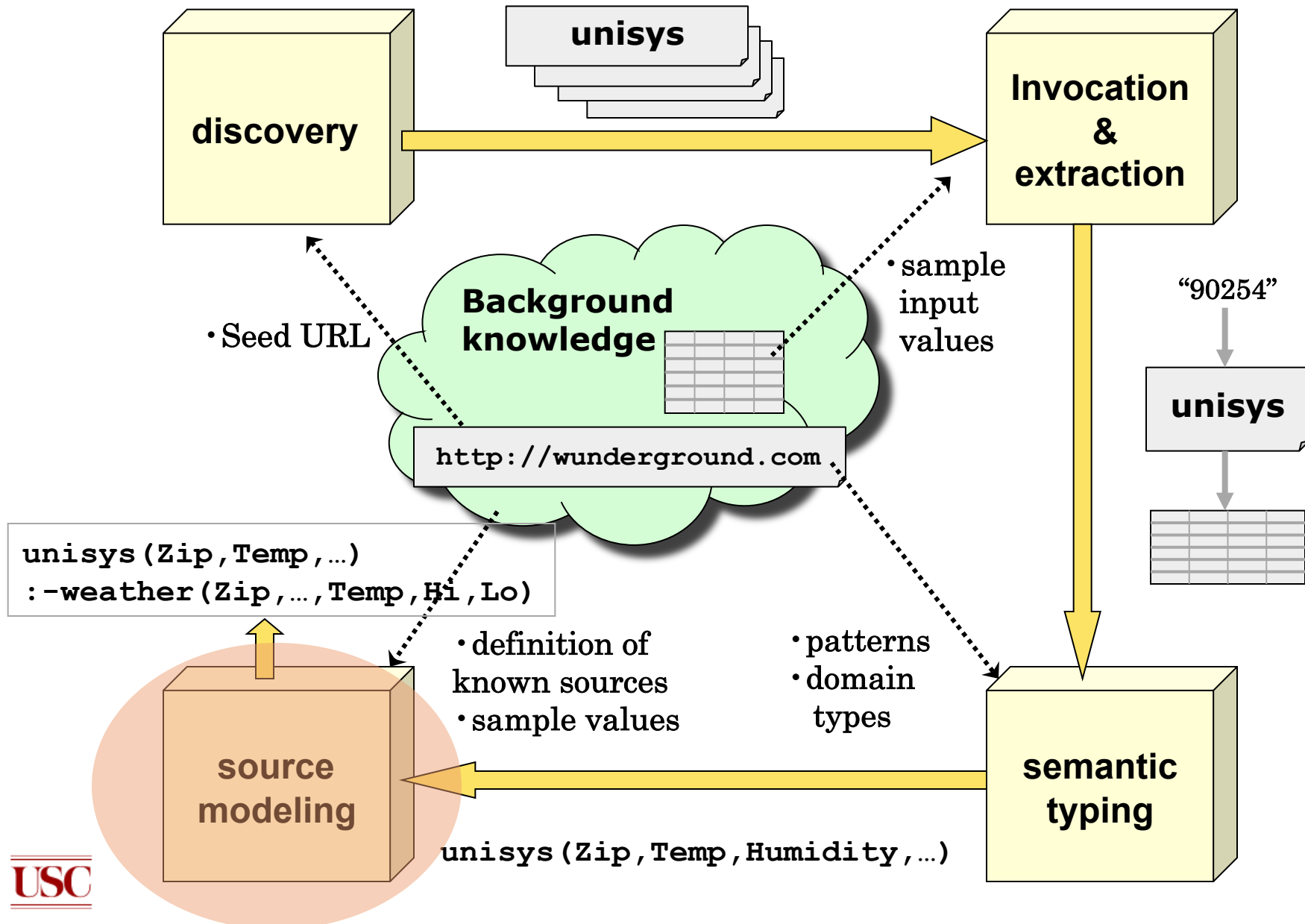
- Use learned patterns to link new data to types in the ontology
  - Score how well patterns describe a set of examples
    - *Number of matching patterns*
    - *How many tokens of the example match pattern*
    - *Specificity of the matched patterns*
  - Output top-scoring types

Person	Address	Work
E Lewis	3518 Hilltop Rd	( 419 ) 531 - 0504
Andrew Lewis	3543 Larchmont Pkwy	( 518 ) 474 - 4799
C. S. Lewis	555 Willow Run Dr	( 612 ) 578 - 5555
Carmen Jones	355 Morgan Ave N	( 612 ) 522 - 5555
John Jones	3574 Brookside Rd	( 555 ) 531 - 9566
Location	State_prov	Postal_code
Toledo	OH	64325-3000
Toledo	OH	64356
Seattle	WA	8422
Seattle	WA	8435
Omaha	NE	52456-6444

## patterns

<b>:StreetAddress:</b>	<b>:Email:</b>
4DIG CAPS Rd	ALPHA@ALPHA.edu
3DIG N CAPS Ave	ALPHA@ALPHA.com
...	...
<b>:State:</b>	<b>:Telephone:</b>
CA	(3DIG) 3DIG-4DIG
2UPPER	+1 3DIG 2DIG 4DIG
...	...

# Source Modeling [Carman & Knoblock]



# Inducing Source Definitions

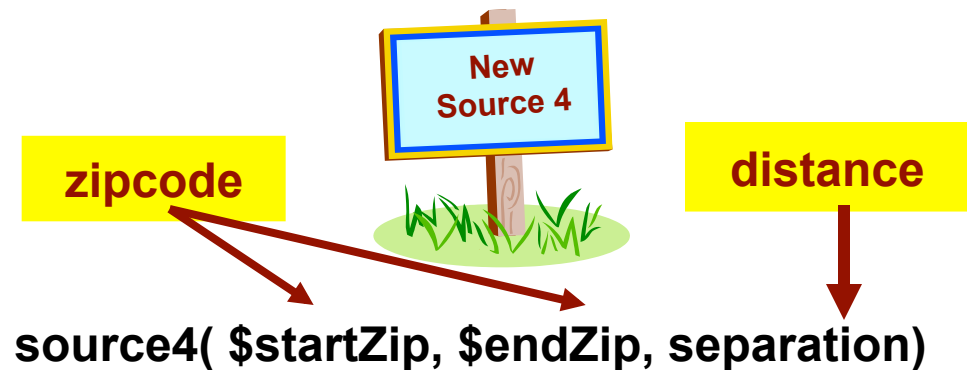


```
source1($zip, lat, long) :-  
  centroid(zip, lat, long).
```

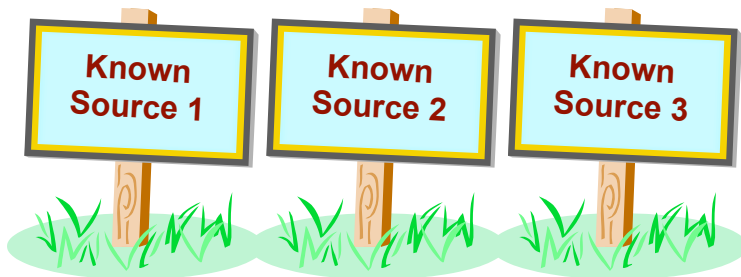
```
source2($lat1, $long1, $lat2, $long2, dist) :-  
  greatCircleDist(lat1, long1, lat2, long2, dist).
```

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

- Step 1: classify input & output semantic types



# Generating Plausible Definition



- Step 1: classify input & output semantic types
- Step 2: generate plausible 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).
```

```
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

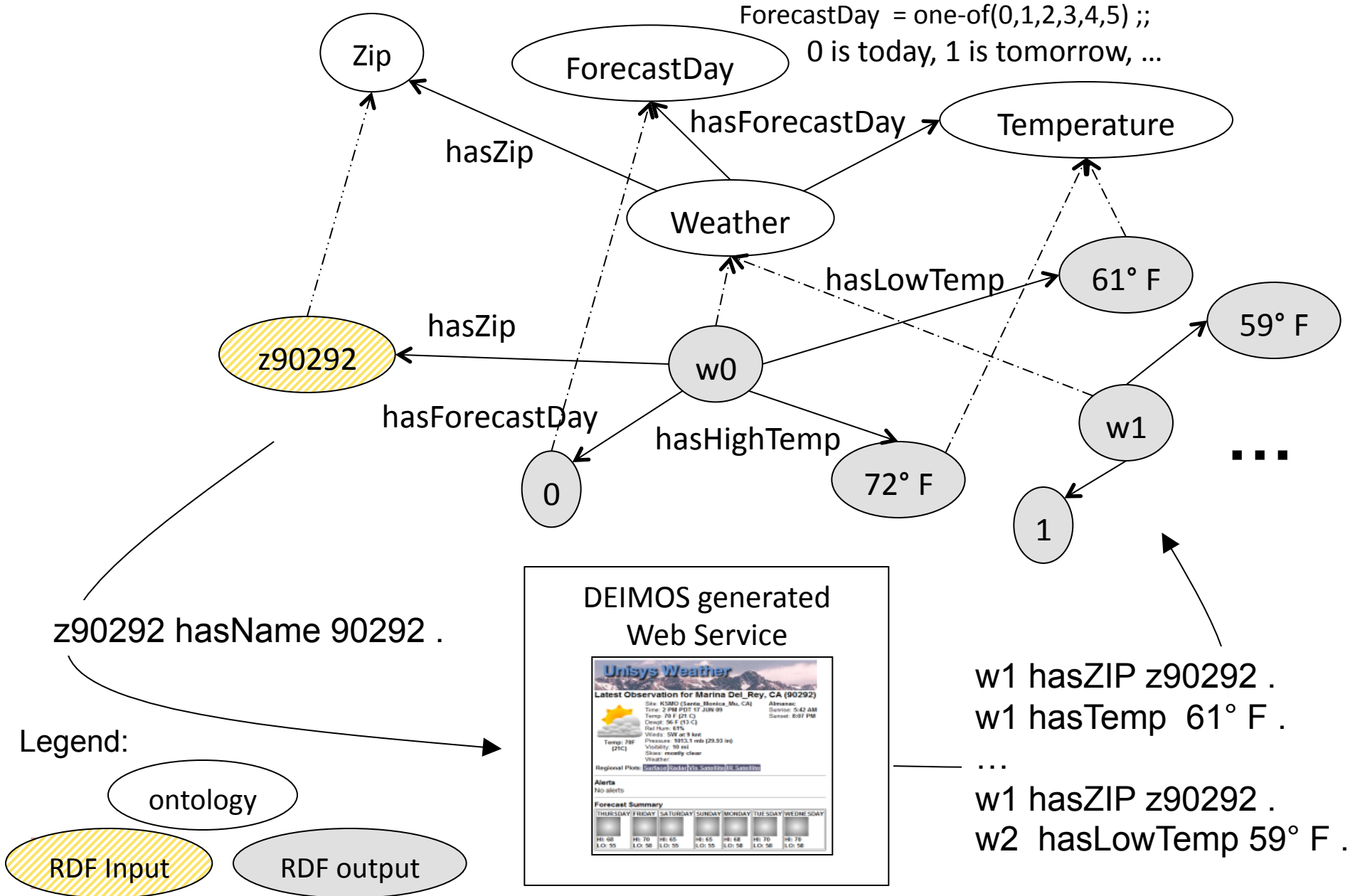
```
source4($zip1, $zip2, dist):-  
  source1(zip1, lat1, long1),  
  source1(zip2, lat2, long2),  
  source2(lat1, long1, lat2, long2, dist2),  
  source3(dist2, dist).
```

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



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

# Constructing the Semantic Web Service





# Background Source Descriptions

wunderground( \$Z,CS,T,F0,C0,S0,Hu0,WS0,WD0,P0,V0,FL1,FH1,S1,  
FL2,FH2, S2,FL3,FH3,S3,FL4,FH4,S4,FL5,FH5,S5):-

Weather(\_w0),hasForecastDay(\_w0,0),hasZIP(\_w0,Z),  
hasCityState(\_w0,CS),hasTimeWZone(\_w0,T),  
hasCurrentTemperatureFahrenheit(\_w0,F0),  
hasCurrentTemperatureCentigrade(\_w0,C0),  
hasSkyConditions(\_w0,S0),hasHumidity(\_w0,Hu0),  
hasPressure(\_w0,P0), hasWindSpeed(\_w0,\_ws1),  
WindSpeed(\_ws1), hasWindSpeedInMPH(\_ws1,WS0),  
hasWindDir(\_ws1,WD0), hasVisibilityInMi(\_w0,V0),  
Weather(\_w1), hasForecastDay(\_w1,1), hasZIP(\_w1,Z),  
hasCityState(\_w1,CS), hasLowTemperatureFahrenheit(\_w1,FL1),  
hasHighTemperatureFahrenheit(\_w1,FH1), hasSkyConditions(\_w1,S1),

...

convertC2F(\$C,F) :- centigrade2fahrenheit(C,F)

# Target explained using background sources



```
unisys($Z,_,_,_,_,_,_,_,F9,_,C,_,F13,F14,Hu,_,F17,_,_,_,_,S22,_,S24,  
_,_,_,_,_,_,_,_,S35,S36,_,_,_,_,_,_,_) :-  
wunderground(Z,_,_,F9,_,Hu,_,_,_,_,F14,F17,S24,_,_,S22,_,_,  
S35,_,_,S36,F13,_,_),  
convertC2F(C,F9)
```

# Learned Target Source Description

unisys(\$Z,\_,\_,\_,\_,\_,\_,\_,\_,\_,F9,\_,\_,C,\_,\_,F13,F14,Hu,\_,\_,F17,\_,\_,\_,\_,\_,S22,\_,\_,S24,\_,\_,\_,\_,  
\_,\_,\_,\_,\_,\_,\_,\_,\_,S35,S36,\_,\_,\_,\_,\_,\_,\_,\_,\_,\_) :-

Weather(\_w0),hasForecastDay(\_w0,0),hasZIP(\_w0,Z),  
hasCurrentTemperatureFahrenheit(\_w0,F9),centigrade2fahrenheit(C,F9),  
hasCurrentTemperatureCentigrade(\_w0,C),hasHumidity(\_w0,Hu0),

Weather(\_w1),hasForecastDay(\_w1,1),hasZIP(\_w1,Z),  
hasCityState(\_w1,CS),hasTimeWZone(\_w1,T),  
hasLowTemperatureFahrenheit(\_w1,F14),  
hasHighTemperatureFahrenheit(\_w1,F17),hasSkyConditions(\_w1,S24),

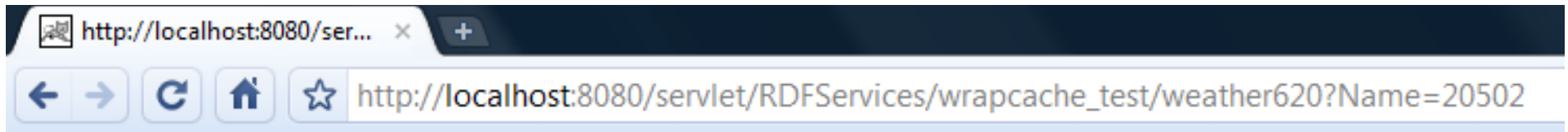
Weather(\_w2),hasForecastDay(\_w2,2),hasZIP(\_w2,Z),  
hasSkyConditions(\_w2,S22),

Weather(\_w3),hasForecastDay(\_w3,3),hasZIP(\_w3,Z),  
hasSkyConditions(\_w3,S35),

Weather(\_w4),hasForecastDay(\_w4,4),hasZIP(\_w4,Z),  
hasSkyConditions(\_w4,S36),

Weather(\_w5),hasForecastDay(\_w5,5),hasZIP(\_w5,Z),  
hasLowTemperatureFahrenheit(\_w5,F13).

# Web Service Invocation



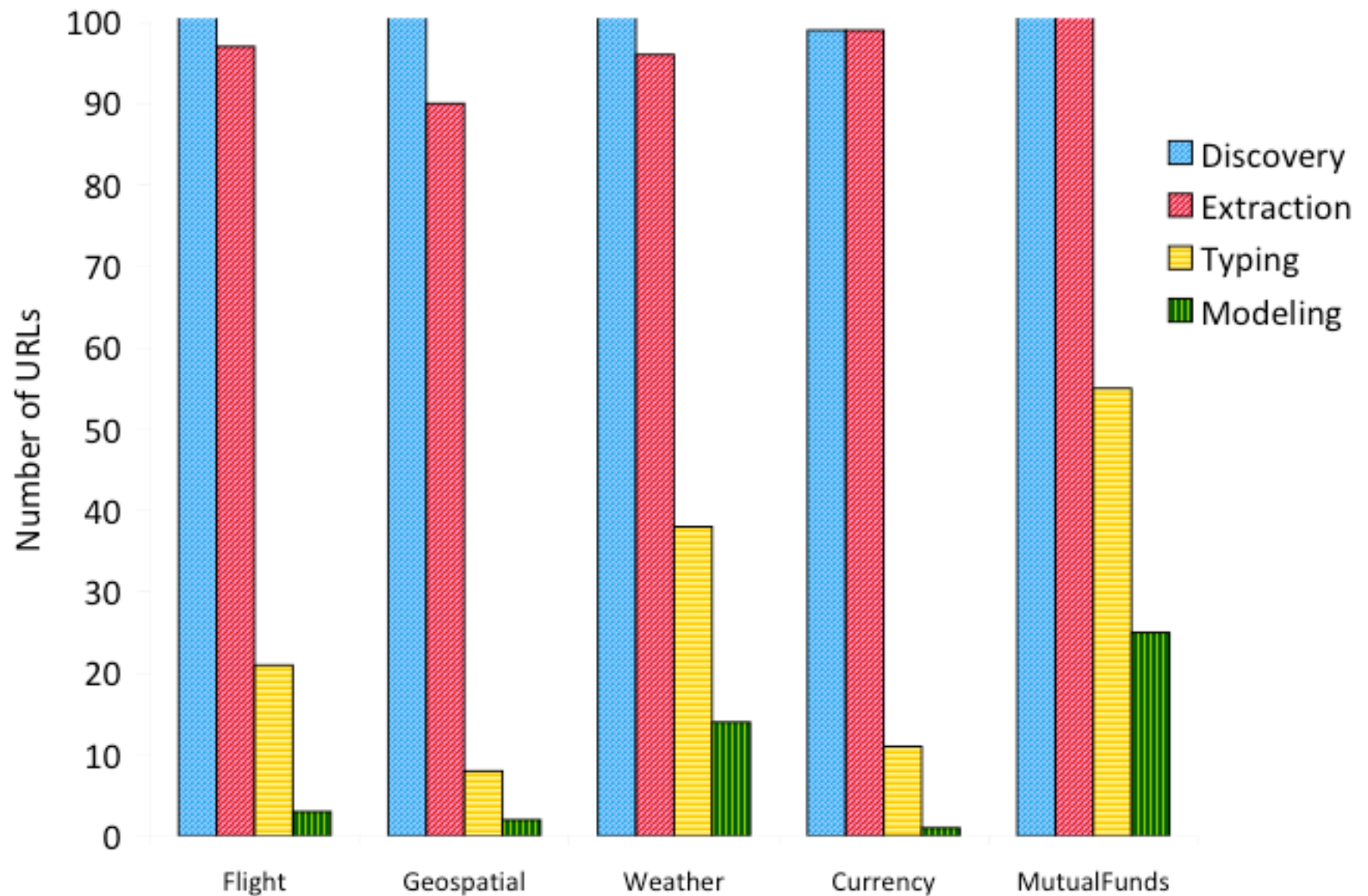
```
weatherforecast7709080 rdf:type WeatherForecast .
weatherforecast7709080 hasZIP "20502" .
weatherforecast7709080 hasCurrentTemperatureFahrenheit "71F" .
windspeed7365415 rdf:type WindSpeed .
weatherforecast7709080 hasWindSpeed windspeed7365415 .
weatherforecast8455262 rdf:type WeatherForecast .
weatherforecast8455262 hasLowTemperature "49 F (9 C)" .
weatherforecast3087280 rdf:type WeatherForecast .
weatherforecast3087280 hasHighTemperature "71 F (21 C)" .
```



- Integrated Approach
  - Discovering related sources
  - Constructing syntactic models of the sources
  - Determining the semantic types of the data
  - Building semantic models of the sources
- **Experimental Results**
- **Related Work**
- **Discussion**

- Experiments in 5 domains
  - Flight – lookup the current status of a flight
  - Geospatial – map street addresses into lat/long coordinates
  - Weather – find the current and forecasted weather
  - Currency – convert between various currencies
  - Mutual Funds – look up current data on a mutual fund
- Evaluation:
  - 1) Can the system correctly learn a model for those sources that perform the same task
  - 2) What is the precision and recall of the attributes in the model

# Candidate Sources after Each Step



# Evaluation 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





- Integrated Approach
  - Discovering related sources
  - Constructing syntactic models of the sources
  - Determining the semantic types of the data
  - Building semantic models of the sources
- Experimental Results
- Related Work
- Discussion

- ILA & Category Translation (Perkowitz & Etzioni 1995)
  - Learn functions describing operations on internet
  - Assumes single input and single tuple as output
- Metadata-based classification of data types used by Web services and HTML forms (Hess & Kushmerick, 2003)
  - Naïve Bayes classifier
  - Only classified the source type, no model
- Use NLP to learn source descriptions (Afzal et al, 2009)
  - Extract type and function provided by service
  - Only provides high-level service type (ex: algorithm, application, data)
- Mining existing workflows (Belhajjame et al, 2008)
  - Connections in parameters of workflows use to infer semantic types
  - Limited semantic description of a web service



- Integrated Approach
  - Discovering related sources
  - Constructing syntactic models of the sources
  - Determining the semantic types of the data
  - Building semantic models of the sources
- Experimental Results
- Related Work
- **Discussion**

- Integrated approach to discovering and modeling online sources and services:
  - *Discover new sources*
  - *How to invoke a source*
  - *Discovering the template for the source*
  - *Finding the semantic types of the output*
  - *Learning a definition of what the service does*
- Provides an approach to generate services and data for the Semantic Web
  - Little motivation for providers to annotate services
  - Instead we can generate metadata automatically

- Coverage, Precision, & Recall
  - Difficult to invoke sources with many inputs
    - *Hotel reservation sites*
  - Hard to learn sources that have many attributes
    - *Some weather sources could have 40 attributes*
- Learning beyond the domain model
  - Learn new semantic types
    - *Discover barometric pressure*
  - Learn new source attributes
    - *Learn about 6-day high and low temperatures*
  - Learn new source relations
    - *Learn conversion between Fahrenheit and Celsius*
  - Learn the domain and range of the sources
    - *Learn that a source provides world weather vs. US weather*
- Linking the Deep Web to the Linked Data Web
  - Use linked data ontologies as domain model
  - Perform entity linkage from web source URI to linked data URI

- Sponsors
  - DARPA CALO Program, AFOSR, & NSF
- Papers
  - Integrated Approach
    - *[Ambite, Darbha, Goel, Knoblock, Lerman, Parundekar, Russ, ISWC 2009]*
  - Source discovery
    - *[Plangprasopchok and Lerman, WWW, 2009]*
  - Source extraction
    - *[Gazen, CMU Ph.d. thesis, 2008]*
  - Semantic typing
    - *[Lerman, Plangprasopchok, & Knoblock, IJSWIS, 2008]*
  - Source modeling
    - *[Carman & Knoblock, JAIR, 2007]*