



USC Viterbi
School of Engineering

Discovering and Building Semantic Models of Web Sources

Craig A. Knoblock
University of Southern California

Joint work with

J. L. Ambite, K. Lerman, A. Plangprasopchok, and T. Russ, USC

C. Gazen and S. Minton, *Fetch Technologies*

M. Carman, *University of Lugano*



The Semantic Web Today?



- Most work on the semantic web assumes that the semantic descriptions of sources and data are given
- What about the rest of the Web??
- Huge amount of useful information that has no semantic description

- 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
 - Validate the correctness of the results



- 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

USC Viterbi
School of Engineering

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

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

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: City, State, Zip, Airport Code, or Country Weather Conditions Go


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](#) - [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 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
Hourly	Hourly	Hourly	Hourly	Hourly


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

Forecast for District of Columbia [Up/Down](#)
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 Model a Source in the Same Domain

Unisys Weather

http://weather.unisys.com/

Twiki APIs Apple (125) TinyURL Zip PL-GUI Heracles GoogleGroups Mantis Shop

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
NAMWrt 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](#)
- For sales information on Unisys weather solutions, contact [Robert Benedict at robert.benedict@unisys.com](#)
- Last modified February 7, 2007

USC

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

file:///Users/tar/Projects/Calo/SourceDiscovery/icdm-unisys/

Twiki APIs Apple (125) TinyURL Zip PL-GUI Heracles GoogleGroups Mantis Shop

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
NAMWrt 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 knot
Wind chill: 41 F Pressure: 1010.1 mb (29.84 in)
Visibility: 10 mi Skies: partly cloudy
Weather:

Alerts
No alerts

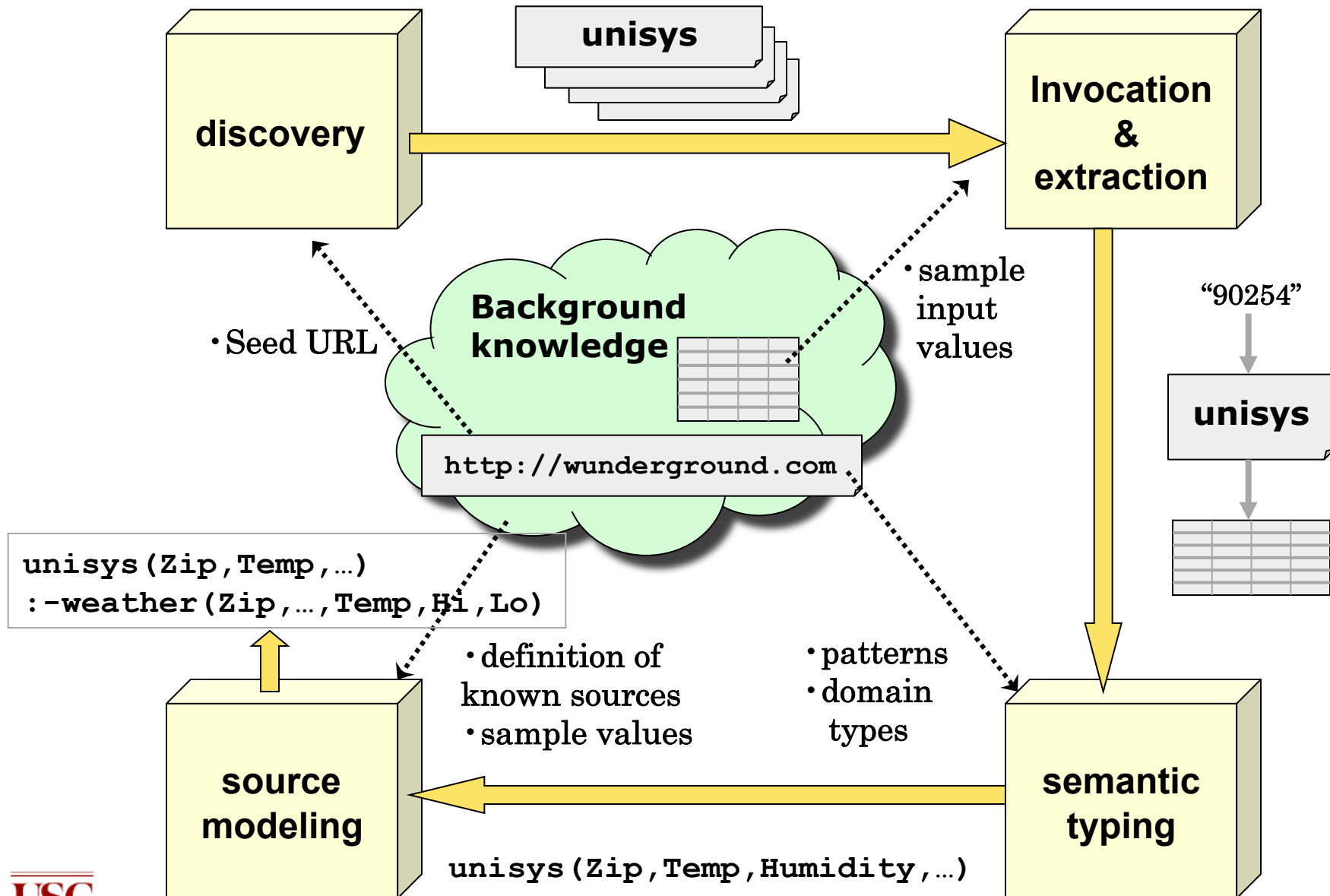
Forecast Summary

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

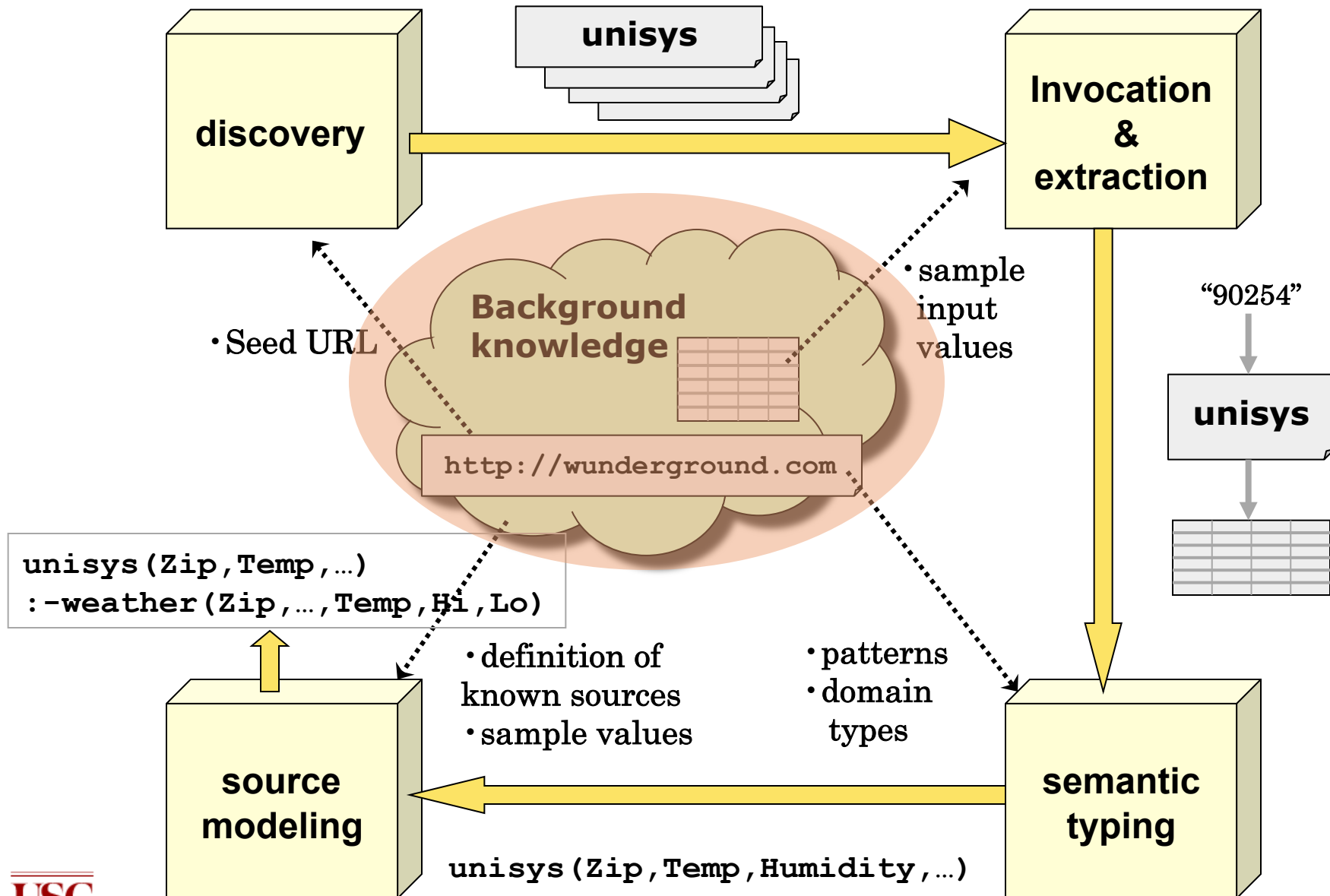
Detailed forecast from National Weather Service
DISTRICT OF COLUMBIA-ARLINGTON/FALLS CHURCH/ALEXANDRIA-
INCLUDING THE CITIES OF...WASHINGTON...ALEXANDRIA...FALLS CHURCH
306 PM EST TUE NOV 25 2008

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

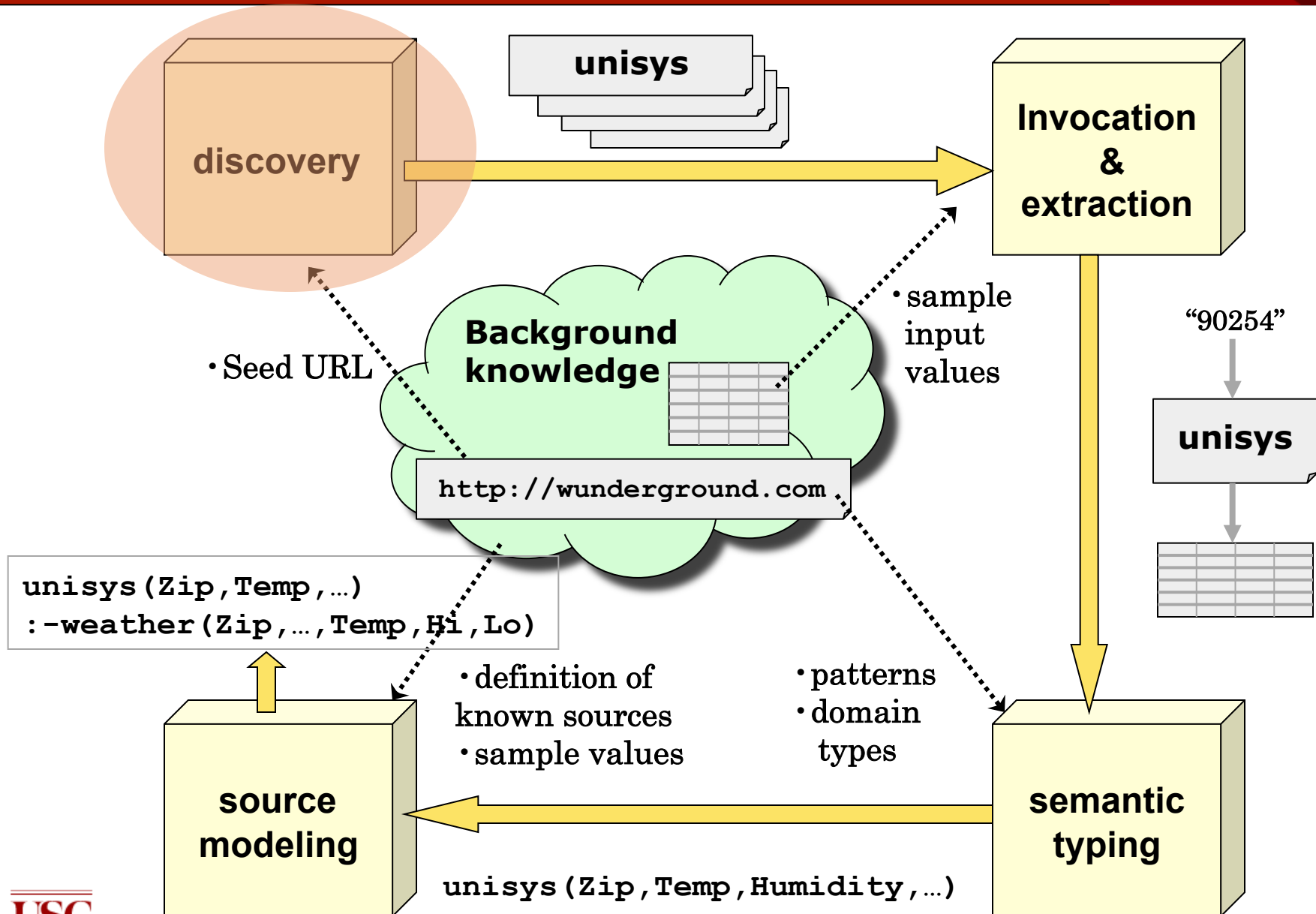


Background Knowledge



- 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 a combination of City and State as input
 - Sample input values
- Known sources (seeds)
 - e.g., <http://wunderground.com>
- Source descriptions in Datalog
 - 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

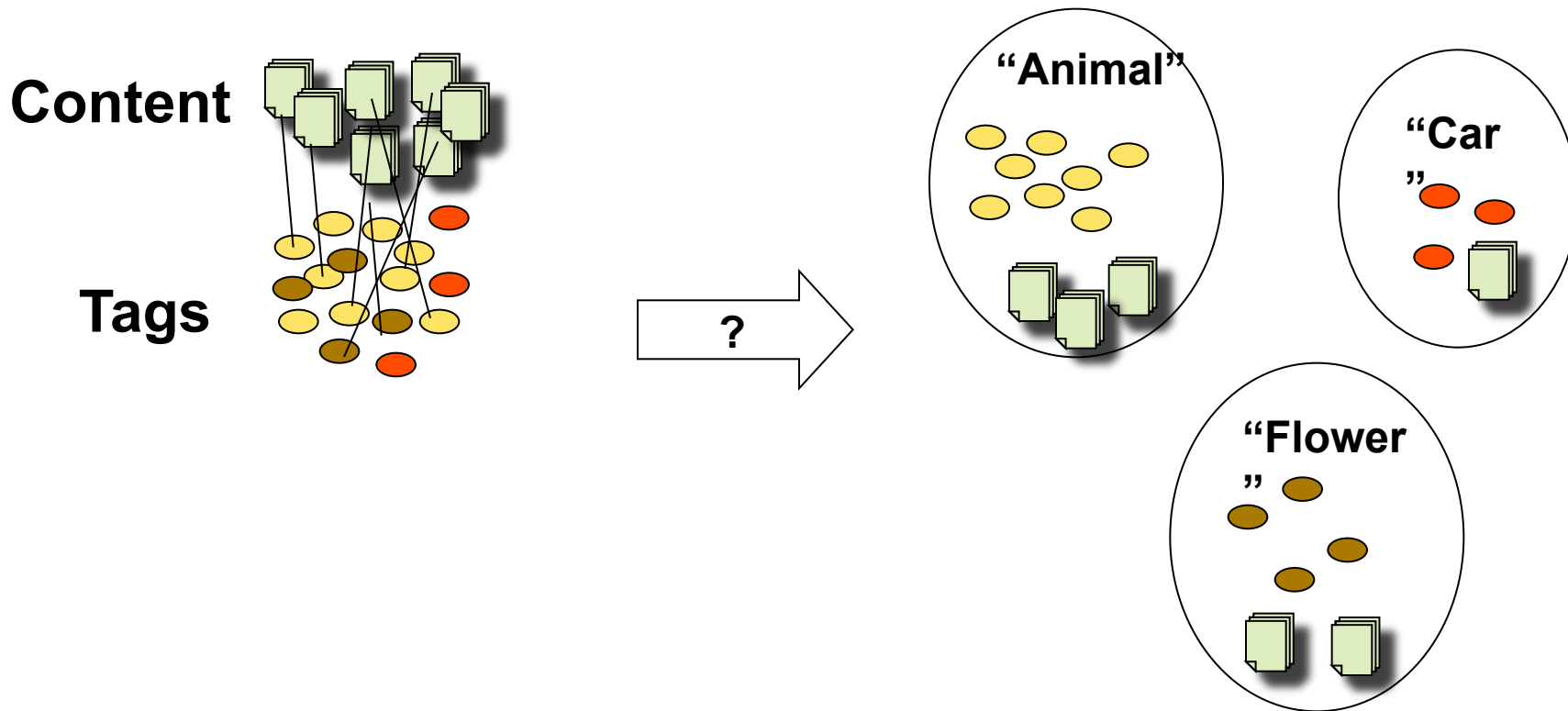
The screenshot shows the del.icio.us interface for the URL <http://delicious.com/url/296e5ade5b8f4d5ac0423343475c783f>. The page title is "Welcome to The Weather Underground : Weather Underground" with the URL www.wunderground.com/. It states that 3242 people have saved this bookmark and 378 have written notes. The "History" tab is selected, showing a list of bookmarks with user avatars and tags. A "Tags" sidebar on the right lists the "Top 10 Tags" with their counts. An arrow points from the text "Most common tags" to the "Top 10 Tags" list. Another arrow points from the text "User-specified tags" to a specific set of tags in the history list.

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

Most common tags

User-specified tags

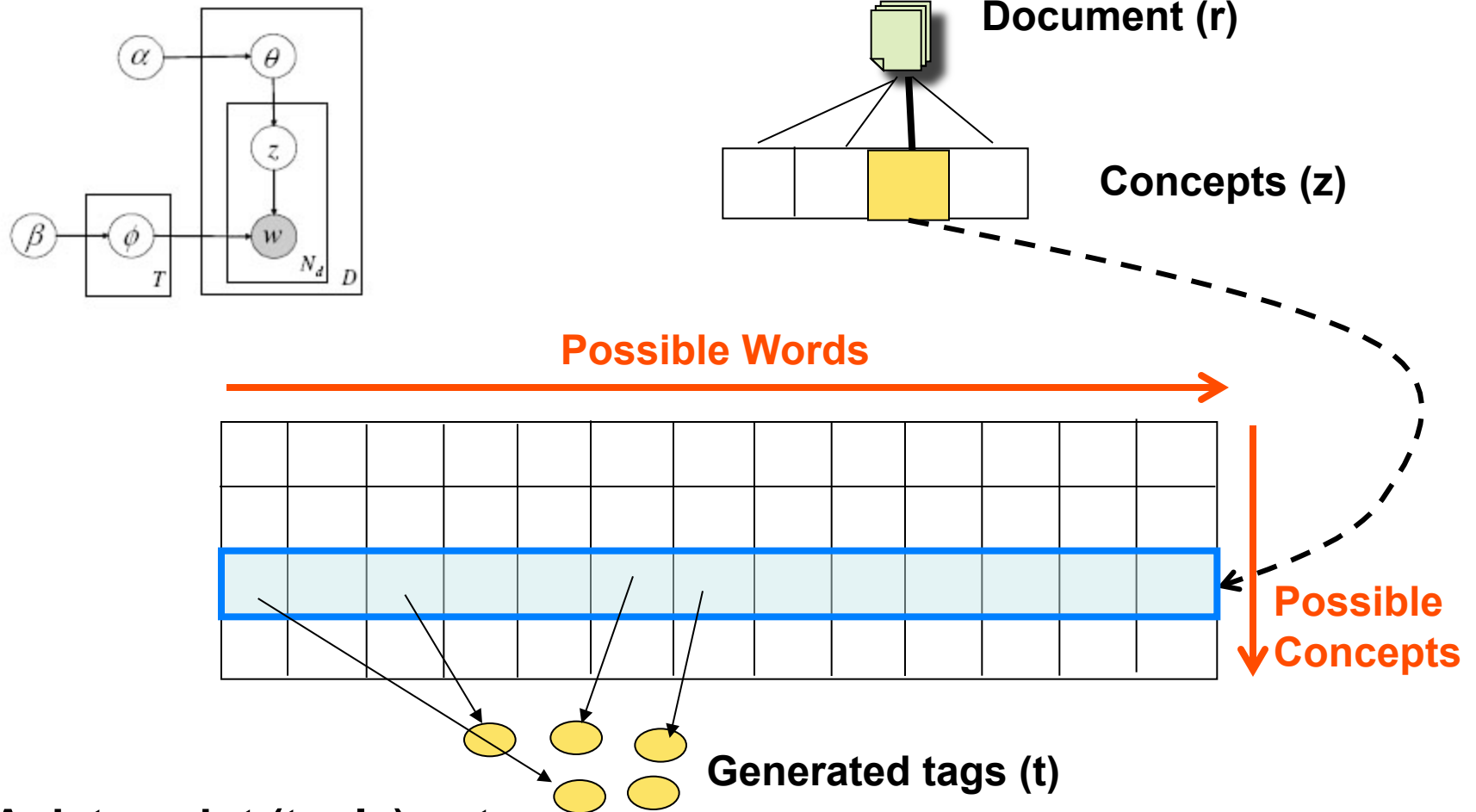
Group Tags and Content into Concepts



Group semantically related tags and content

A Stochastic Process of Tag Generation

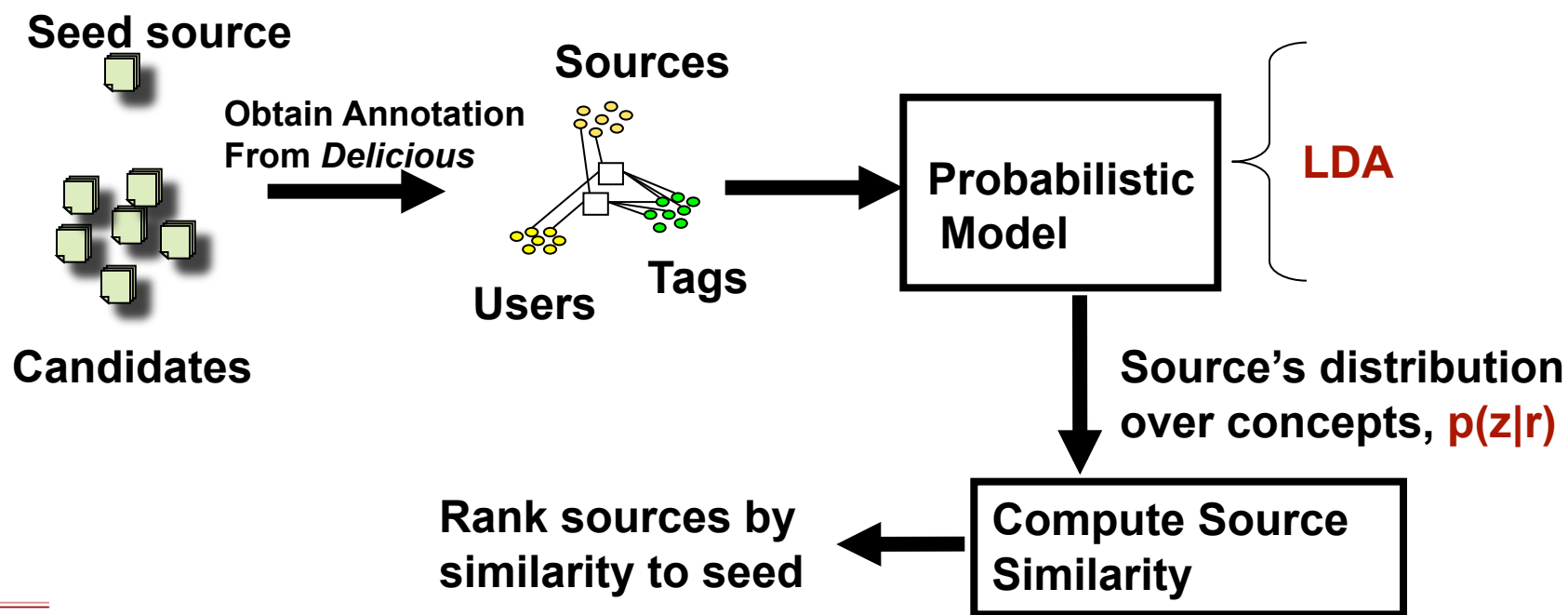
PLSA (Hofmann99);
LDA (Blei03+)



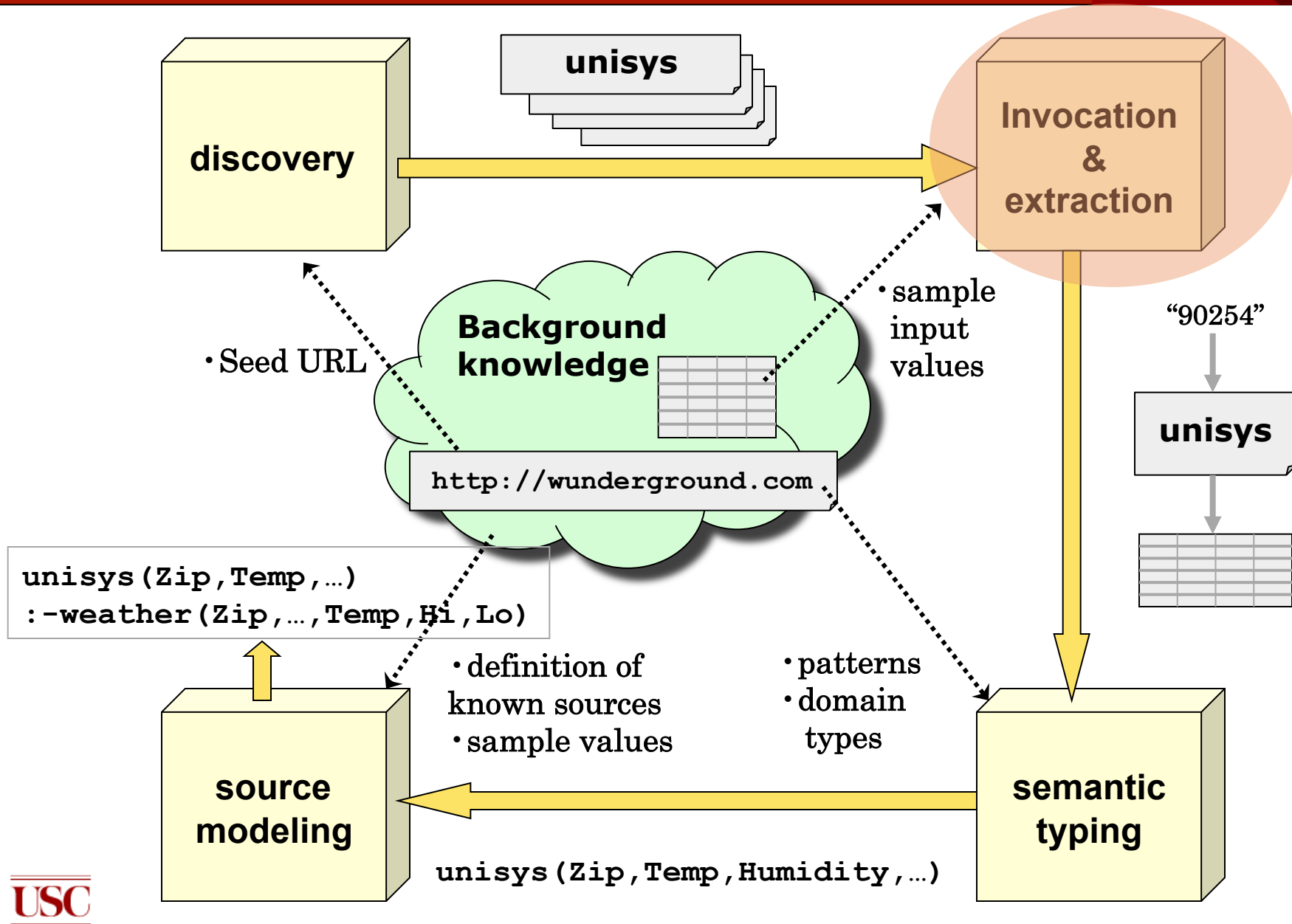
USC A data point (tuple) $\langle r, t, z \rangle$

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



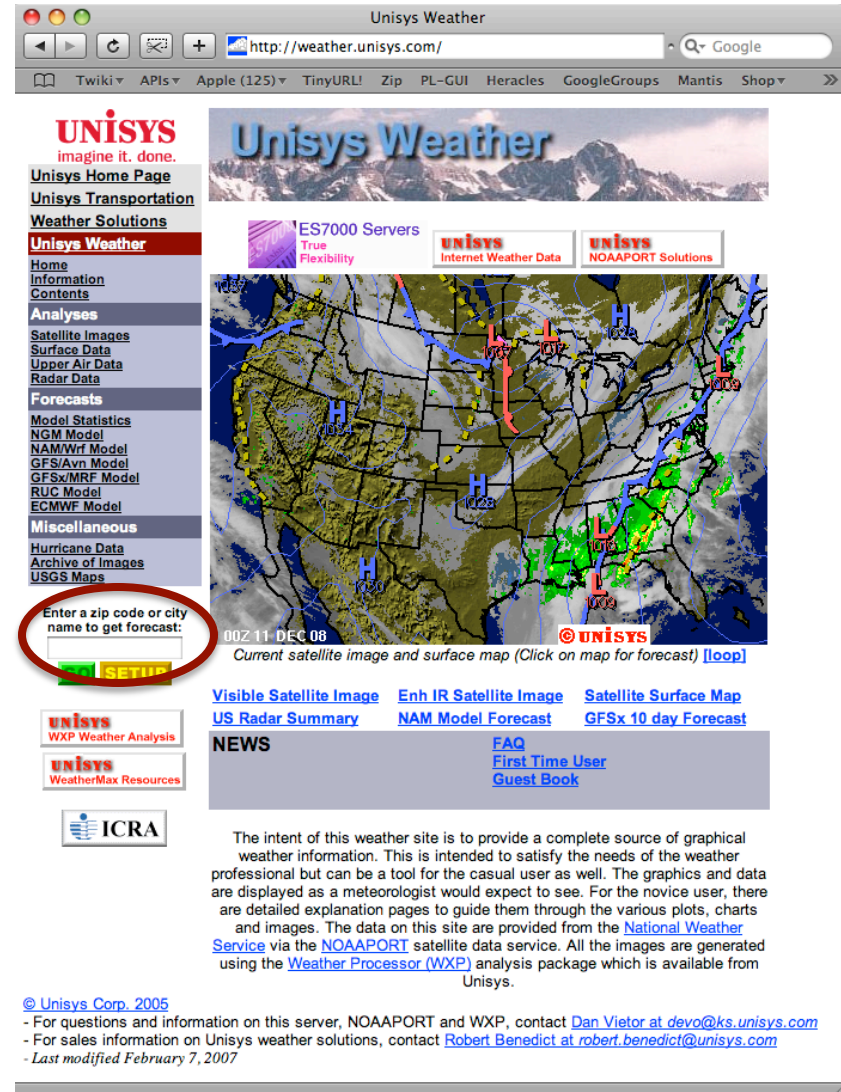
Source Invocation & Extraction



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*
 3. Submit Form
 4. Keep successful invocations

Form
Input



The screenshot shows the Unisys Weather website interface. The browser address bar displays <http://weather.unisys.com/>. The website features a navigation menu on the left with links to Home, Information, Contents, Analyses, Forecasts, and Miscellaneous. The main content area displays a map of the United States with weather data, including a satellite image and a surface map. A red circle highlights the input field for a zip code or city name, with the text "Enter a zip code or city name to get forecast:" above it. Below the input field is a "GO" button. The website also includes links to various weather services, such as "Visible Satellite Image", "Enh IR Satellite Image", "Satellite Surface Map", "US Radar Summary", "NAM Model Forecast", and "GFSx 10 day Forecast". A "NEWS" section is also present, with links to "FAQ", "First Time User", and "Guest Book".

© Unisys Corp. 2005
- For questions and information on this server, NOAAPORT and WXP, contact [Dan Vietor at devo@ks.unisys.com](mailto:Dan.Vietor@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

Invoke the Target Source with Possible Inputs

<http://weather.unisys.com>

Weather conditions for 20502

input

Unisys Weather

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](#)
- For sales information on Unisys weather solutions, contact [Robert Benedict at robert.benedict@unisys.com](#)
- Last modified February 7, 2007

Unisys Weather

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 knot
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

Form Input Data Model

- Each domain has an input data model
 - Derived from the seed sources
 - Alternate input groups
- Each domain has sample values for the input data types

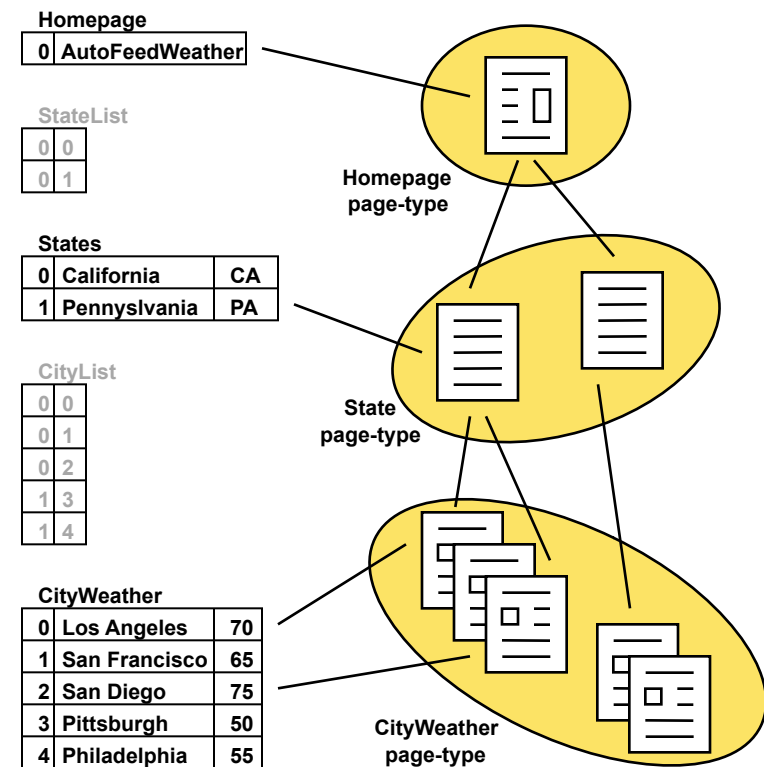
domain name="weather"

- input "zipcode" type PR-Zip
- input "cityState" type PR-CityState
- input "city" type PR-City
- input "stateAbbr" type PR-StateAbbr

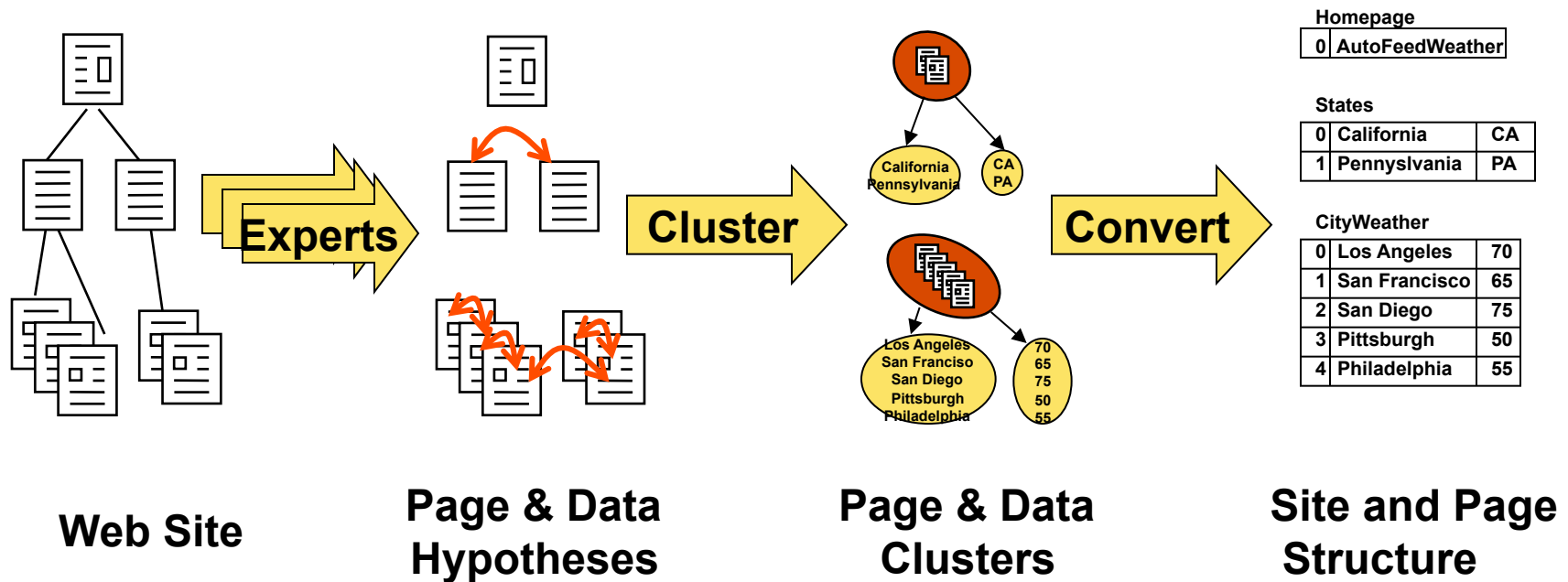
PR-Zip	PR-CityState	PR-City	PR-StateAbbr
20502	Washington, DC	Washington	DC
32399	Tallahassee, FL	Tallahassee	FL
33040	Key West, FL	Key West	FL
90292	Marina del Rey, CA	Marina del Rey	CA
36130	Montgomery, AL	Montgomery	AL

Discovering Web Structure [Gazen & Minton]

- Model Web sources that generate pages dynamically in response to a query
 - Find the relational data underlying a semi-structured web site
- Generate a page template that can be used to extract data on new pages
- Approach
 - *Site extraction*
 - Exploit the common structure within a web site
 - Take advantage of multiple structures
 - HTML structure, page layout, links, data formats, etc.



Approach to Finding Web Structure



- URL patterns give clues about site structure
 - Similar pages have similar URLs, e.g.:
 - <http://www.bookpool.com/sm/0321349806>
 - <http://www.bookpool.com/sm/0131118269>
 - <http://www.bookpool.com/ss/L?pu=MN>
- Page layout gives clues about relational structure
 - Similar items aligned vertically or horizontally, e.g.:



A screenshot of a weather page showing four cities: Chicago, IL; London, UK; New York, NY; and San Francisco, CA. Each city entry includes a temperature range, a unit (F), and a weather icon. Four vertical yellow lines are drawn across the page to highlight the vertical alignment of the city names, temperature ranges, units, and weather icons, demonstrating how layout can reveal relational structure.

Chicago, IL	58...83 F	
London, UK	54...68 F	
New York, NY	70...85 F	
San Francisco, CA	61...72 F	

- Page Templates
 - Similar pages contain common sequences of substrings

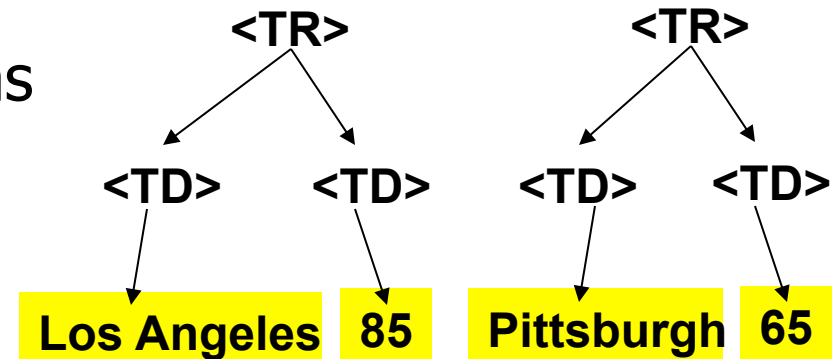
Right Now for
Los Angeles, CA (90007)
[Save this Location](#)

77°F	UV Index: 3 Moderate
Feels Like 77°F	Wind: From WSW at 7 mph
	Humidity: 56%
	Pressure: 29.78 in.
	Dew Point: 61°F

Right Now for
Pittsburgh, PA (15213)
[Save this Location](#)

73°F	UV Index: 0 Low
Feels Like 73°F	Wind: From SW at 3 mph
	Humidity: 46%
	Pressure: 30.23 in.
	Dew Point: 51°F

- HTML Structure
 - List rows are represented as repeating HTML structures



Extracting Data

Pages

```
<td valign="top" width="14%">  
<td valign="top" width="14%">  
  <font face="Arial, Helvetica, sans-serif">  
    <small><b>FRIDAY<br>  
    <br>  
    HI: 65<br>LO: 52<br></b></small></font></td>  
<td valign="top" width="14%">  
  <font face="Arial, Helvetica, sans-serif">  
    <small><b>SATURDAY<br>  
    <br>  
    HI: 60<br>LO: 48<br></b></small></font></td>
```



Hypotheses

- **group_member**
(FRIDAY, SATURDAY)
- **group_member**
(Sunny, Rainy)
- **same_html_context**
(65, 60)
- **vertically_aligned**
(Sun, Rain)
- **two_digit_number**
(65, 52, 60, 48)
- ...

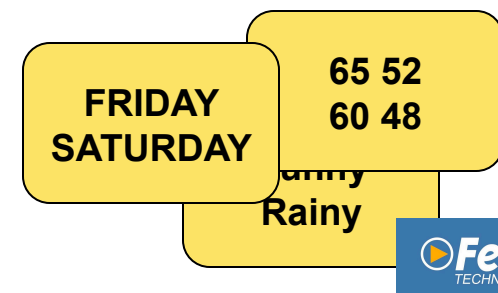


Extracted Data

FRIDAY	Sun	Sunny	65	52
SATURDAY	Rain	Rainy	60	48



Clusters



Data Extraction with Templates

- Build templates with the inferred page structure
- Use the templates to extract data on unseen pages

Unseen Page

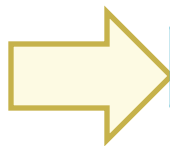
```
<br>
<font face="Arial, Helvetica, sans-serif">
  <small><b>Temp: 71F (21C)</b></small></font>
<font face="Arial, Helvetica, sans-serif">
  <small>Site: <b>KCQT (Los_Angeles_Dow, CA)</b><br>
    Time: <b>11 AM PST 10 DEC 08</b>
```



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

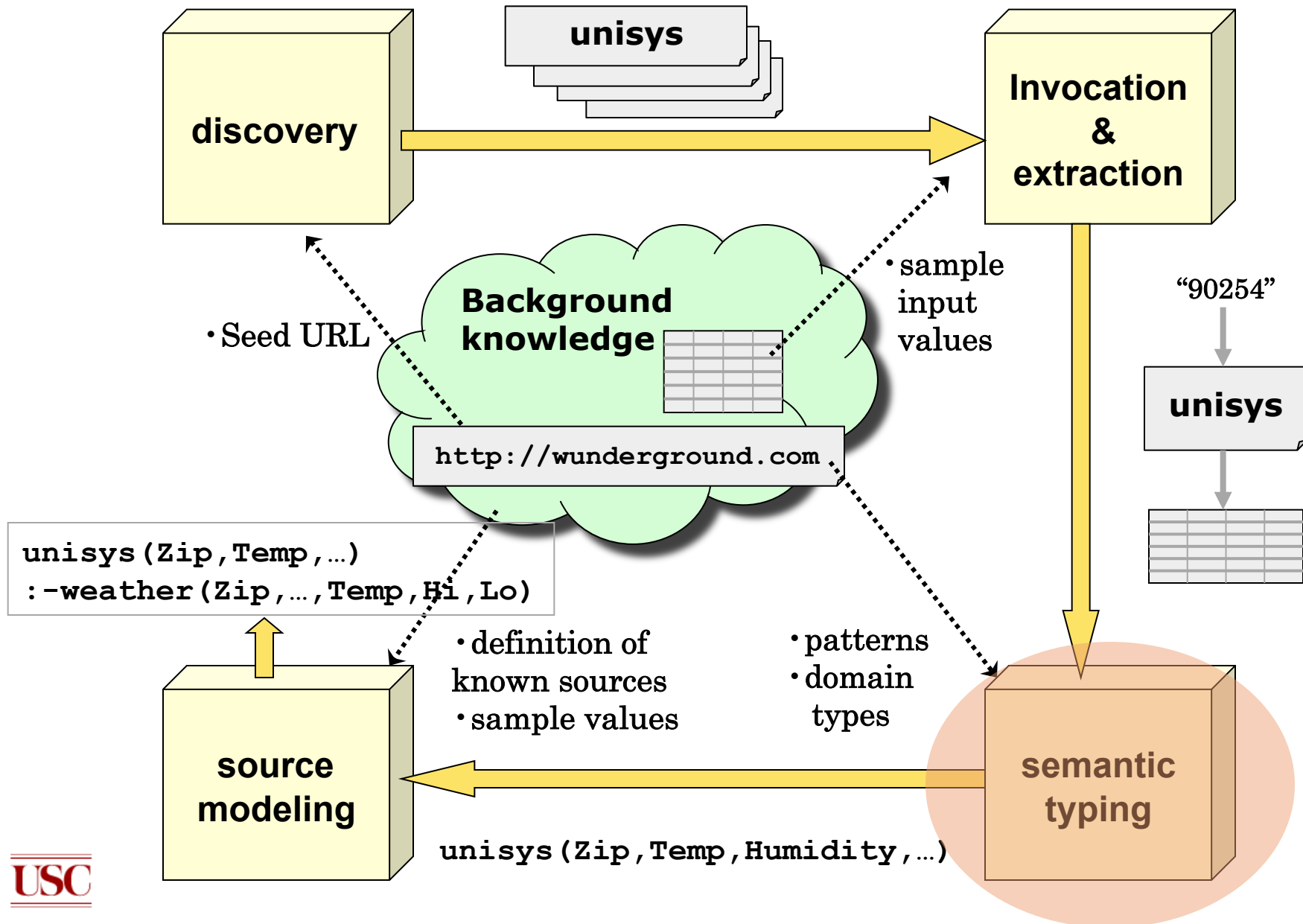


Sun	Sunny	71F	21C	KCQT	Los_Angeles_Dow	CA	11 AM PST
-----	-------	-----	-----	------	-----------------	----	-----------

Raw Extracted Data from Unisys

Column	Invocation 1	Invocation 2	...
1	Unisys Weather: Forecast for Washington, DC (20502) [0] 2	Unisys Weather: Forecast for Tallahassee, FL (32399) [0] 2	
2	Washington,	Tallahassee,	
3	DC	FL	
4	20502 Good Field	32399	
5	20502) Extra Garbage	32399)	
...			
14	Images/PartlyCloudy.png Image URL	Images/Sun.png	
15	Partly Cloudy Good Field	Sunny	
16	45 Hard to Recognize	63	
17	Temp: 45F (7C) Too Complex	Temp: 63F (17C)	
18	45F Good Field	63F	
...			
217	45	64	
218	MOSTLY SUNNY. HIGHS IN THE MID 40S.	PARTLY CLOUDY. HIGHS AROUND 64.	

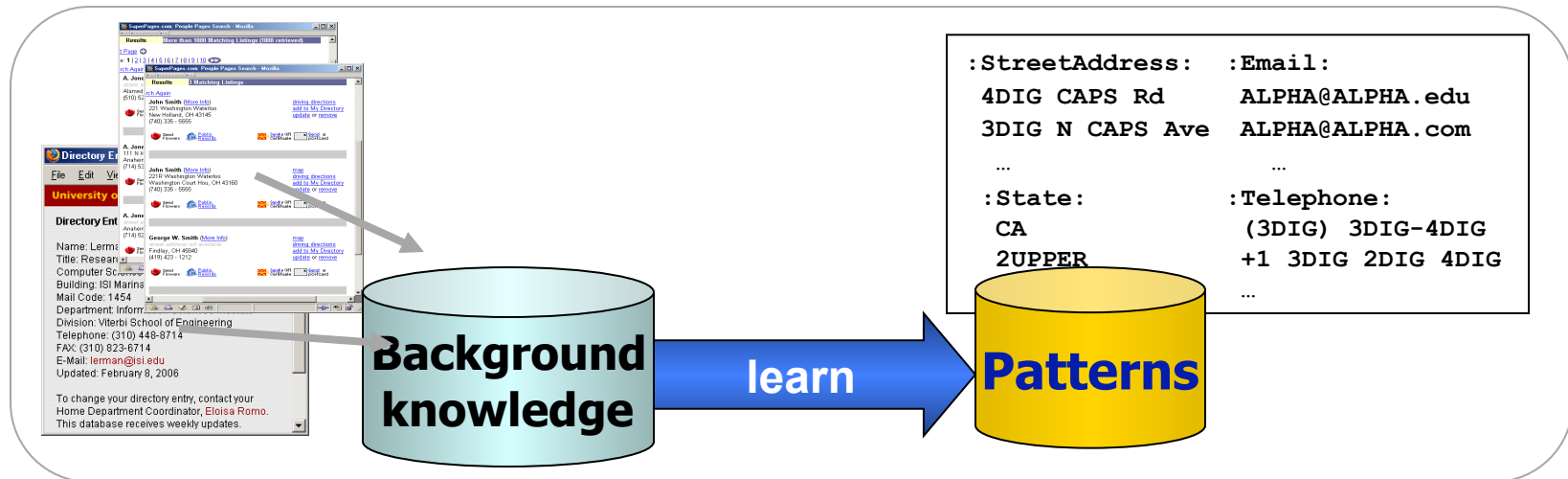
Semantic Typing



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

label

: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

Learning Patterns to Recognize Semantic Types



- Domain-independent language to represent the structure of data as patterns
 - Pattern is a sequence of tokens and token types
 - E.g., Phone number

Examples

310 448-8714

310 448-8775

212 555-1212

Patterns

[(310) 448 – 4DIGIT]

[(3DIGIT) 3DIGIT – 4DIGIT]

- Learns patterns from examples for all semantic types in the domain model

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

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

Weather Data Types

Sample values

- PR-TempF
88 F
57°F
82 F ...
- PR-Visibility
8.0 miles
10.0 miles
4.0 miles
7.00 mi
10.00 mi
- PR-Zip
07036
97459
02102

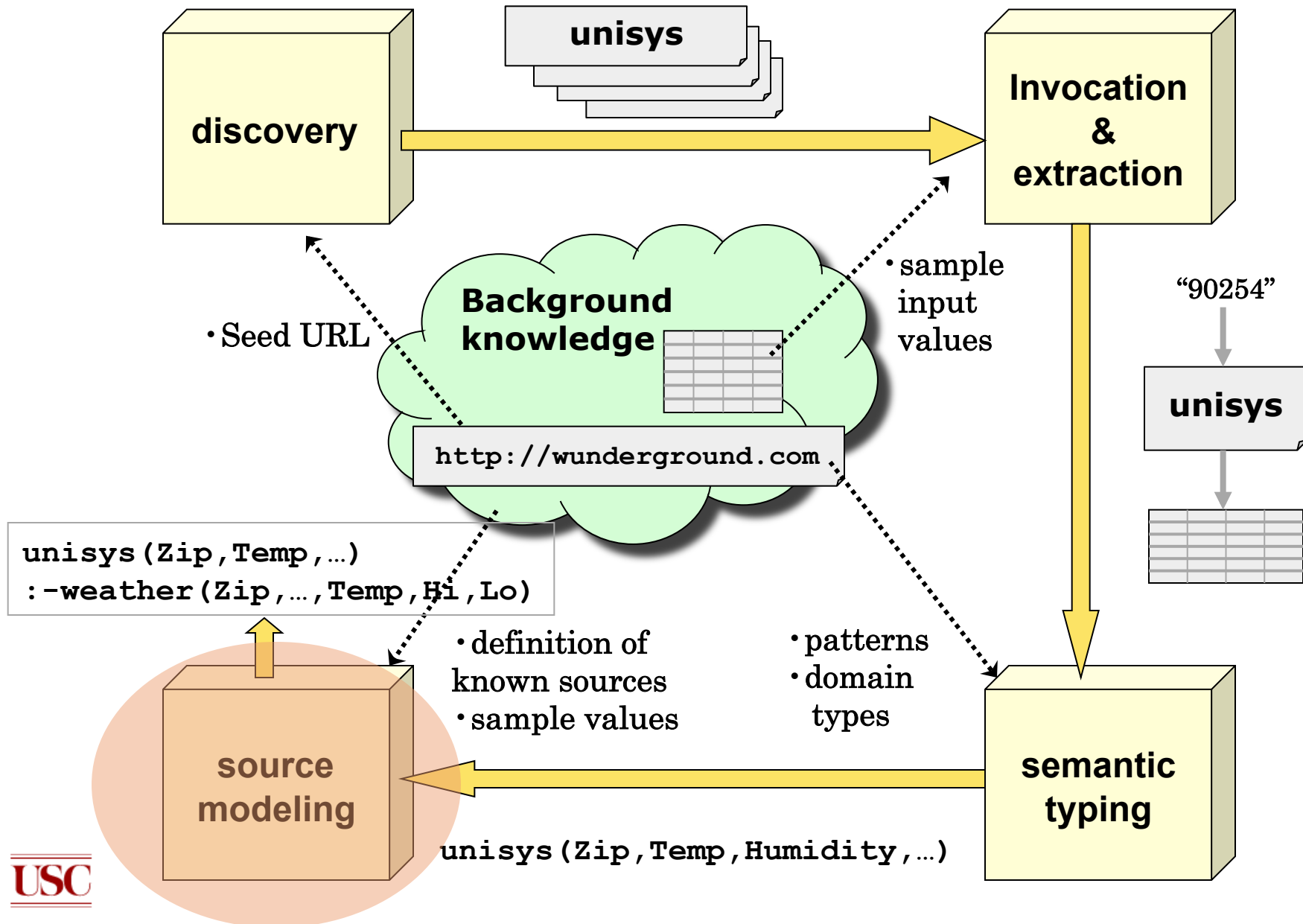
Patterns

- PR-TempF
[88, F]
[2DIGIT, F]
[2DIGIT, °, F]
- PR-Visibility
[10, ., 0, miles]
[10, ., 00, mi]
[10, ., 00, mi, .]
[1DIGIT, ., 00, mi]
[1DIGIT, ., 0, miles]
- PR-Zip
[5DIGIT]

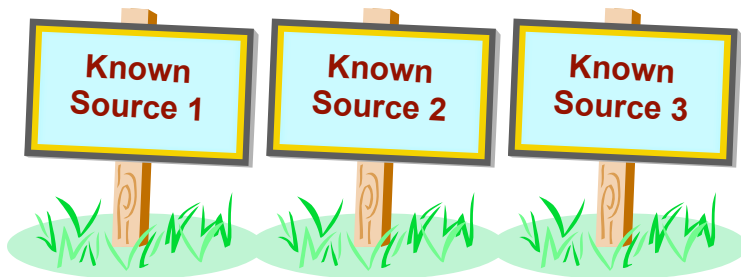
Labeled Columns of Target Source Unisys

Column	4	18	25	15	87
Type	PR-Zip	PR-TempF	PR-Humidity	PR-Sky	PR-Sky
Score	0.333	0.68	1.0	0.325	0.375
Values	20502	45F	40%	Partly Cloudy	Sunny
	32399	63F	23%	Sunny	Partly Cloudy
	33040	73F	73%	Sunny	Rainy
	90292	66F	59%	Partly Cloudy	Sunny
	36130	62F	24%	Sunny	Partly Cloudy

Source Modeling [Carman & Knoblock]



Inducing Source Definitions

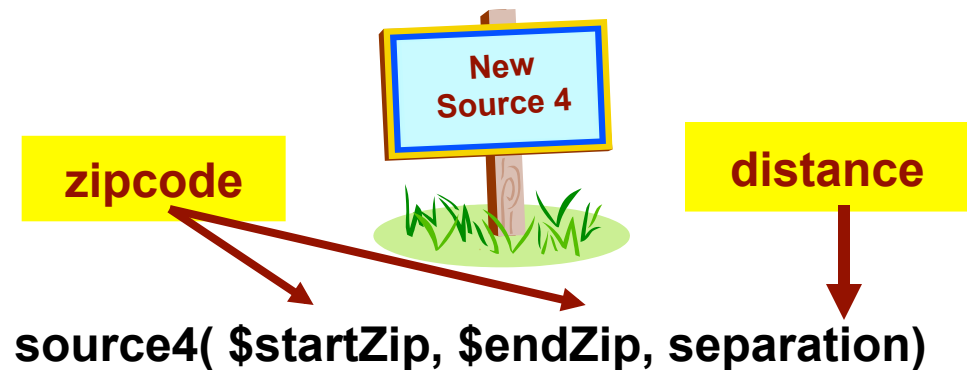


**source1(\$zip, lat, long) :-
centroid(zip, lat, long).**

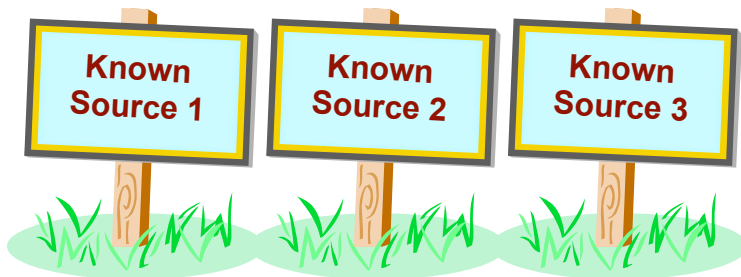
**source2(\$lat1, \$long1, \$lat2, \$long2, dist) :-
greatCircleDist(lat1, long1, lat2, long2, dist).**

**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).
```

Start with empty clause & generate specialisations by

- Adding one predicate at a time from set of sources
- Checking that each definition is:
 - Not logically redundant
 - Executable (binding constraints satisfied)



source5(____).

Expand

source5(\$zip1,\$dist1,zip2,dist2)

```
source5(zip1,____)      :- source4(zip1,zip1,____).  
source5(zip1,_,zip2,dist2) :- source4(zip2,zip1,dist2).  
source5(____,dist1,_,dist2) :- <(dist2,dist1).  
...
```

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(dist2, dist).
```

match

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

Allow flexibility in values from different sources

- Numeric Types like *distance*

10.6 km \approx 10.54 km

Error Bounds (eg. +/- 1%)

- Nominal Types like *company*

Google Inc. \approx Google Incorporated

String Distance Metrics

(e.g. JaroWinkler Score $>$ 0.9)

- Complex Types like *date*

Mon, 31. July 2006 \approx 7/31/06

Hand-written equality checking procedures.

Example of a Learned Source Model for Weather Domain



- Given a set of known sources and their descriptions
 - wunderground(\$Z,CS,T,F0,S0,Hu0,WS0,WD0,P0,V0) :-
weather(0,Z,CS,D,T,F0,_,_,S0,Hu0,P0,WS0,WD0,V0)
 - convertC2F(C,F) :- centigrade2fahrenheit(C,F)
- Learn a description of a new source in terms of the known sources
 - unisys(\$Z,CS,T,F0,C0,S0,Hu0,WS0,WD0,P0,V0) :-
wunderground(Z,CS,T,F0,S0,Hu0,WS0,WD0,P0,V0),
convertC2F(C0,F0)

Evaluate the Candidate Definition

- Invoke the source and the definition on the sample inputs and compare the results

Seed (wunderground.com)

Washington, District of Columbia

Local Time: 1:07 PM EST — [Set My Timezone](#)

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

Current Conditions

Eckington PI, 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)

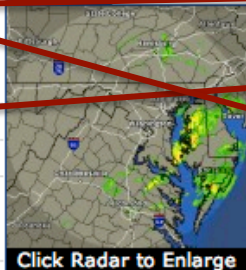
Visibility: 10.0 miles /
16.1 kilometers

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](#)

Target (unisys.com)

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)

Sunset: 4:48 PM

Dewpt: 22 F (-5 C)

Rel Hum: 40%

Winds: W at 7 knt

Wind chill: 41 F

Temp: 45F (7C) Pressure: 1010.1 mb (29.84 in)

Visibility: 10 mi

Skies: partly cloudy

Weather:



- Source invocation
 - Sources had to be invoked simultaneously to compare the results
- Source extraction
 - Tokenization of numbers had to be accurate
 - *-38.253432 vs. "38", "2534322"*
- Semantic typing
 - Unit information had to be preserved
 - *Difficult to determine whether 10 is a temperature or windspeed without the unit*
- Source modeling
 - Synonyms had to be represented as data sources
 - *Need to know the mapping between airline names and codes*



- 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 3 domains
 - Geospatial
 - *Geocoder that maps street addresses into lat/long coordinates*
 - Weather
 - *Produces current and forecasted weather*
 - Flight Status
 - *Current status for a given airline and flight*
- Evaluation:
 - 1) Can we 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

- DEIMOS crawls social bookmarking site del.icio.us to discover sources similar to domain seeds:
 - Geospatial: geocoder.us
 - Weather: wunderground.com
 - Flight status: Flytecomm.com
- For each seed:
 - retrieve the 20 most popular tags users applied to this source
 - retrieve other sources that users have annotated with that tags
- Compute similarity of resources to seed using model
- Manually checked top-ranked 100 resources produced by model
 - Same functionality if same inputs and outputs as seed
 - Among the 100 highest ranked URLs:
 - *16 relevant geospatial sources*
 - *61 relevant weather sources*
 - *14 relevant flight status sources*



- **Invocation & Extraction**

- Recognize form input parameters and calling method
- Learn extraction template for result page
- Success: Determines how to invoke a form and builds a template for the result page

- **Semantic Typing**

- Automatically assign semantic types to extracted data
- Success: If extractor produces output table *and* at least one output column not part of the input can be typed

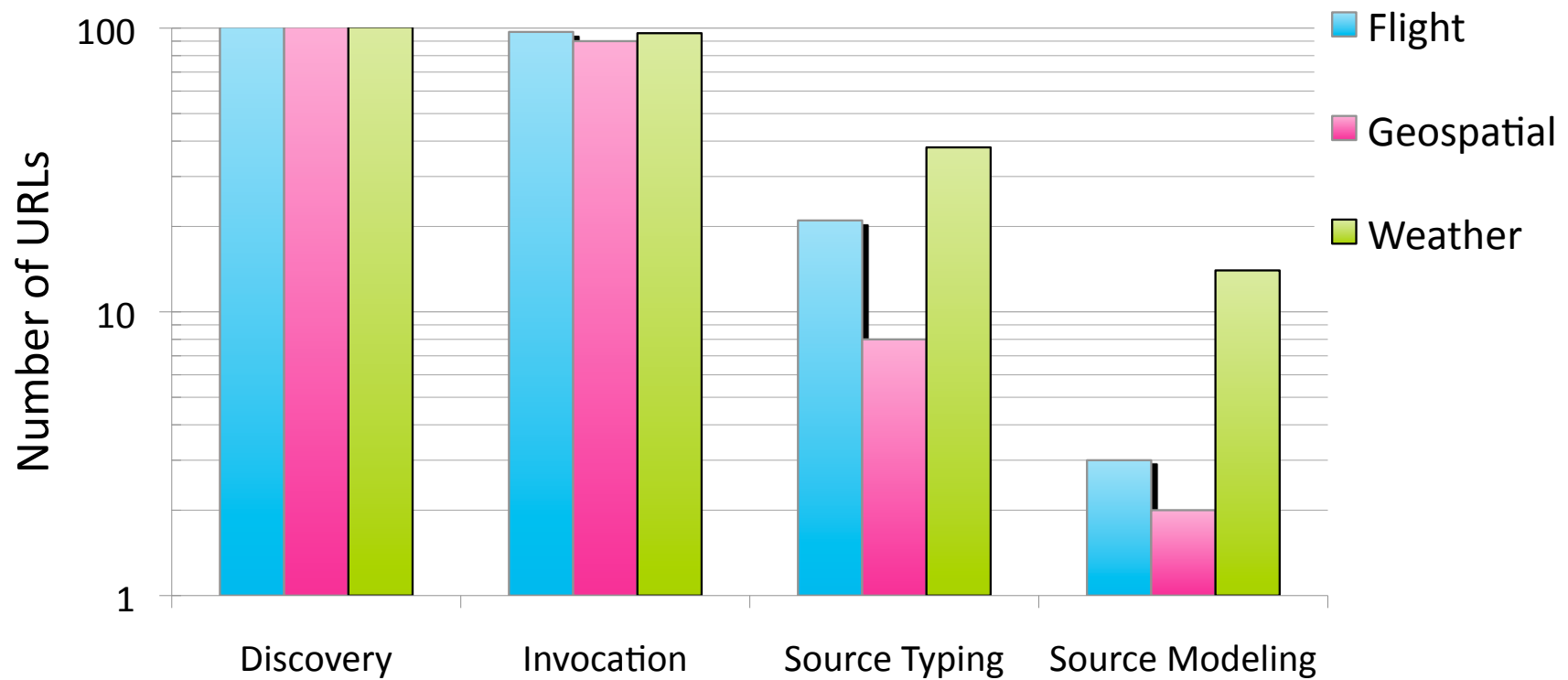
- **Semantic Modeling**

- Learn Datalog source descriptions based on background knowledge
- Success: Learn a source description where at least one output column is not part of the input
- Evaluate accuracy of the resulting source model

Candidate Sources after Each Step



URL Filtering by Module



Confusion Matrix



Geospatial

	PT	PF
AT	8	8
AF	8	76

Weather

	PT	PF
AT	46	15
AF	15	24

Flight

	PT	PF
AT	4	10
AF	10	76

(a) Source Discovery

	PT	PF
AT	2	0
AF	0	6

	PT	PF
AT	15	4
AF	8	14

	PT	PF
AT	2	0
AF	5	6

(b) Source Modeling

PT=Predicted True
PF=Predicted False

AT=Actual True
AF=Actual False

Evaluation of the Models



	Recall	Precision	F-measure
geospatial	86	100	92
weather	29	64	39
flight	35	69	46



- 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
 - Known static sources with no binding constraints
 - Assumes single input and single tuple as output
- iMAP (Dhamanka et. al. 2004)
 - Discovers complex (many-to-1) mappings between DB schemas
 - Used specialized searchers to find mappings
- 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
- Woogle: Metadata-based clustering of data and operations used by Web services (Dong et al, 2004)
 - Groups similar types together: Zipcode, City, State
 - Also supported only classification of sources

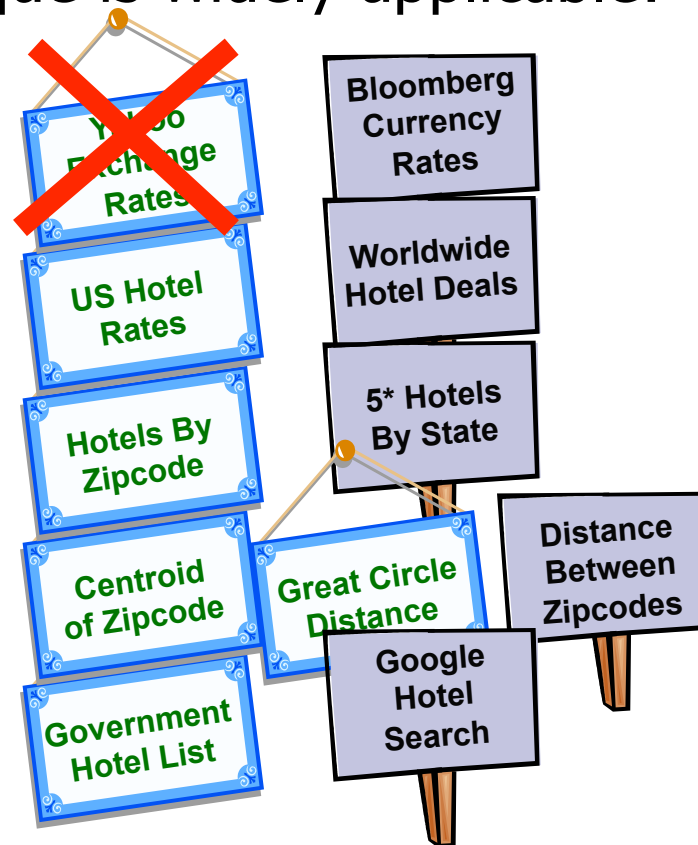
- Mining Semantic Descriptions of Bioinformatics Web Resources [Afzal et al., in EWSC 2009]
 - Extracts the semantic descriptions of web services from the natural languages text about the services
 - Useful for people to discover new sources, but the descriptions don't provide the level of description needed for reasoning and composition
- Automatic Annotation of Web Services [Belhajjame et al., 2006]
 - Automatic annotation of web service parameters
 - Addresses the part of the problem related to semantic typing
- ...and much related work on subproblems



- 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

- Assumption: overlap between new & known sources
- Nonetheless, the technique is widely applicable:

- Redundancy
- Scope or Completeness
- Binding Constraints
- Composed Functionality
- Access Time





- 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 source descriptions 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*
 - Mislabels attributes due to similar values
 - *Need to build models using more input data*
- Learning beyond the domain model
 - Learn new semantic types
 - *Discovery 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*

- Sponsors
 - DARPA CALO Program, AFOSR, & NSF
- Papers
 - Integrated Approach
 - *[Ambite, Gazen, Knoblock, Lerman, & Russ, II-Web 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]*