



USC Viterbi
School of Engineering

Web-Based Learning

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*





- Problem
 - Web sources and services are designed for people, not machines
 - Limited or no description of the information provided by these sources
 - This makes it hard, if not impossible to find, retrieve and integrate the vast amount of structured data available
 - *Weather sources, geocoders, stock information, currency converters, online stores, etc.*
- Approach
 - Start with an some initial knowledge of a domain
 - *Sources and semantic descriptions of those sources*
 - Automatically
 - *Discover related sources*
 - *Learn the syntactic structure of the sources*
 - *Build semantic models of the source*
 - *Validate the correctness of the results*

Seed Source

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 Mostly Cloudy Hourly	 47° F 31° F 8° C -1° C Partly Cloudy Hourly	 50° F 31° F 10° C -1° C Clear Hourly	 50° F 34° F 10° C 1° C Partly Cloudy Hourly	 47° F 34° F 8° C 1° C Chance of Rain 30% chance of precipitation 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
GFSxAvn Model
GFSxMRF 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
GFSxAvn Model
GFSxMRF 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

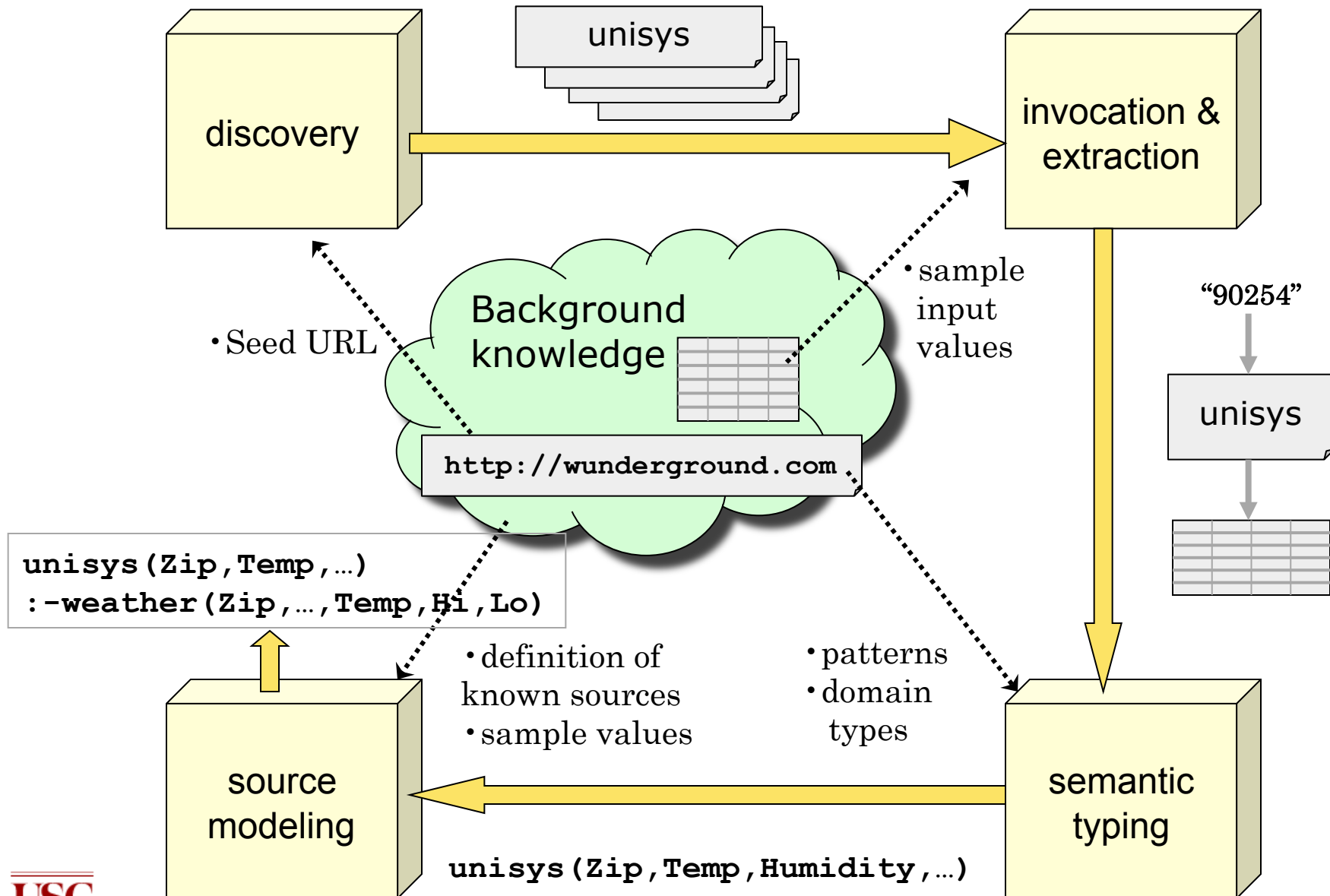
Forecast Summary

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

Approach





- Discovering sources using social annotations
- Discovering the structure of sources
- Learning semantic types of the source data
- Learning semantic models of the sources
- Experimental Results
- Discussion



- Discovering sources using social annotations
- Discovering the structure of sources
- Learning semantic types of the source data
- Learning semantic models of the sources
- Experimental Results
- Discussion

Learning Concepts from Social Annotation (Tags)

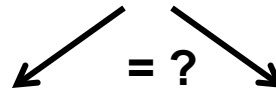


By sparky2000



By A lion Rohrs

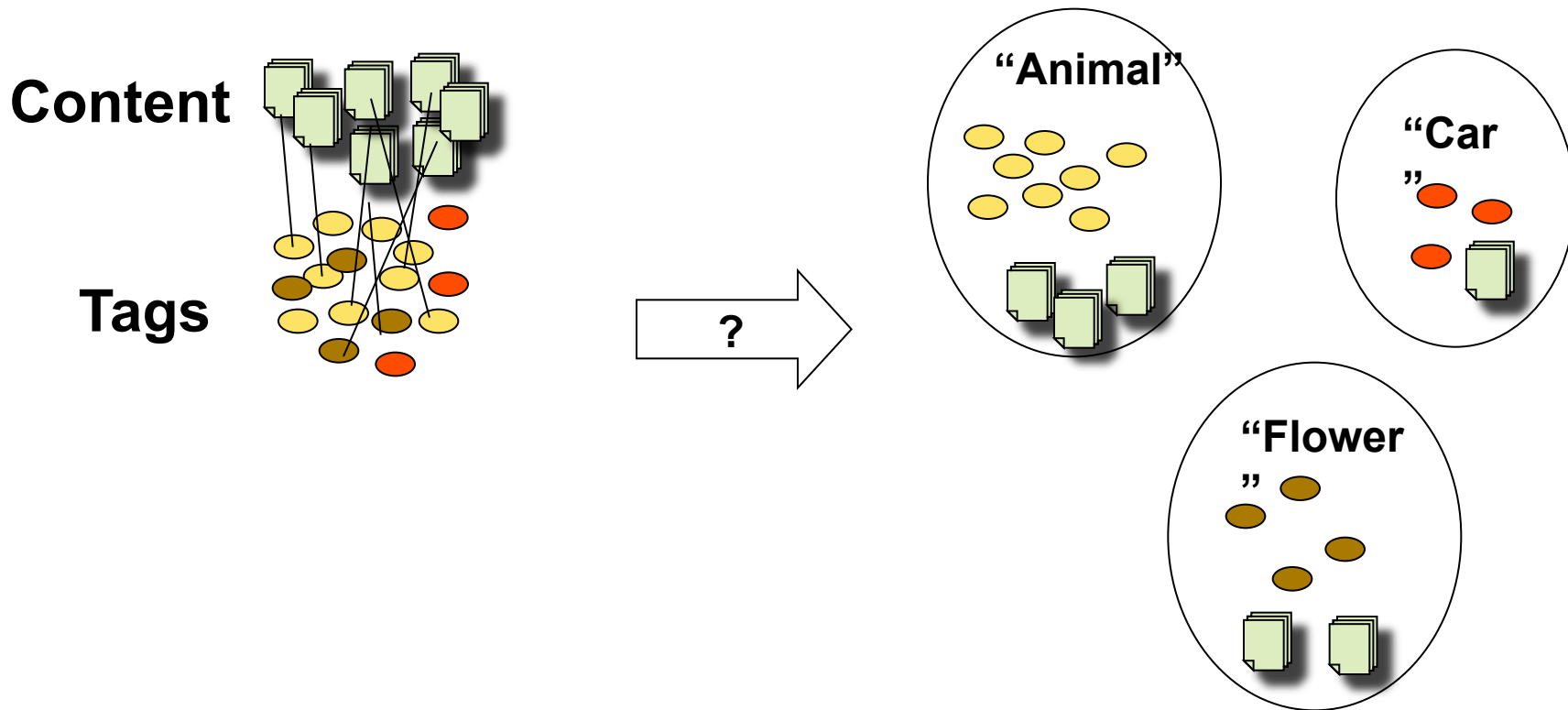
“Black” + “Jaguar”



Animal

Car

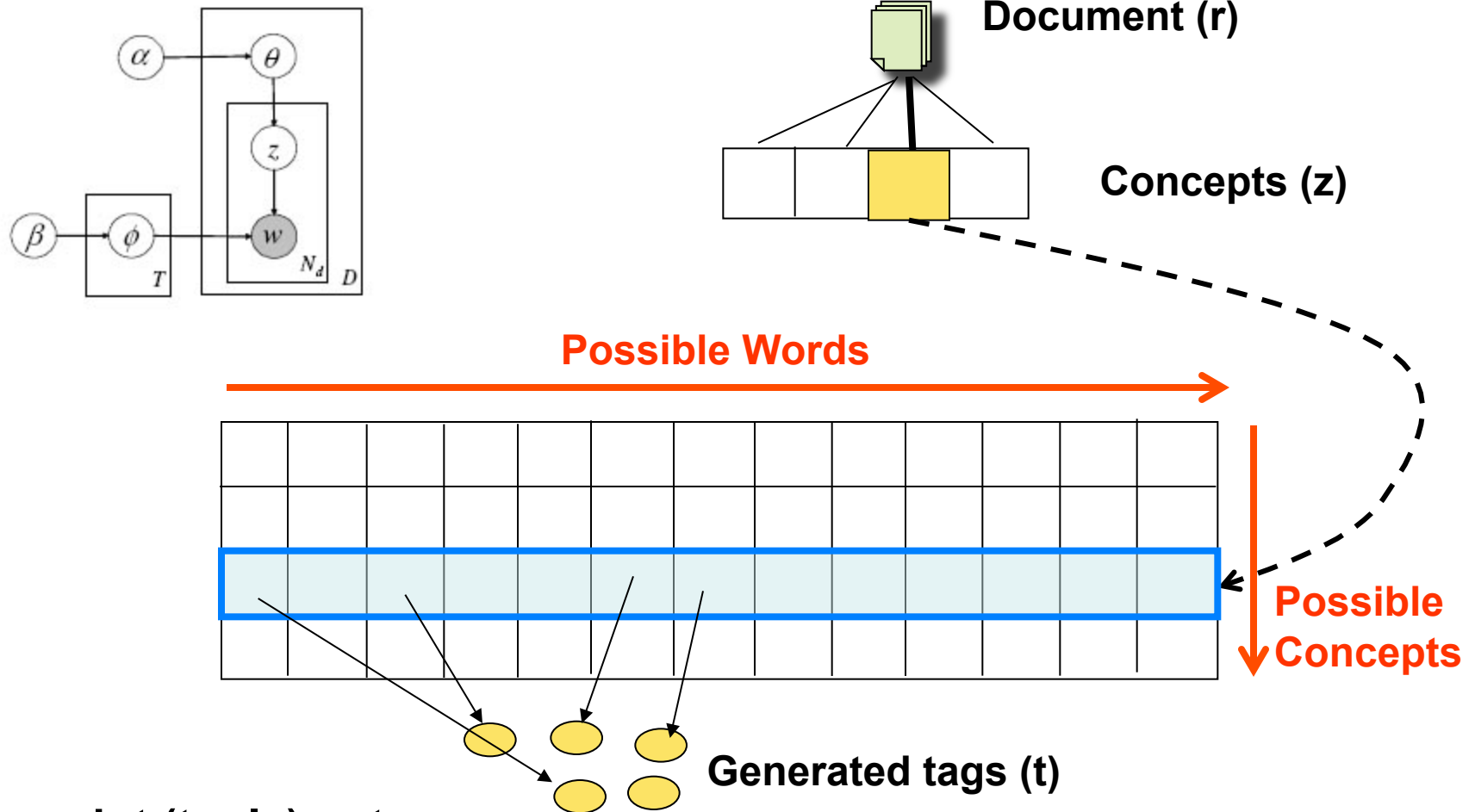
Goal



Grouping semantically related tags and content

A stochastic process of tag generation

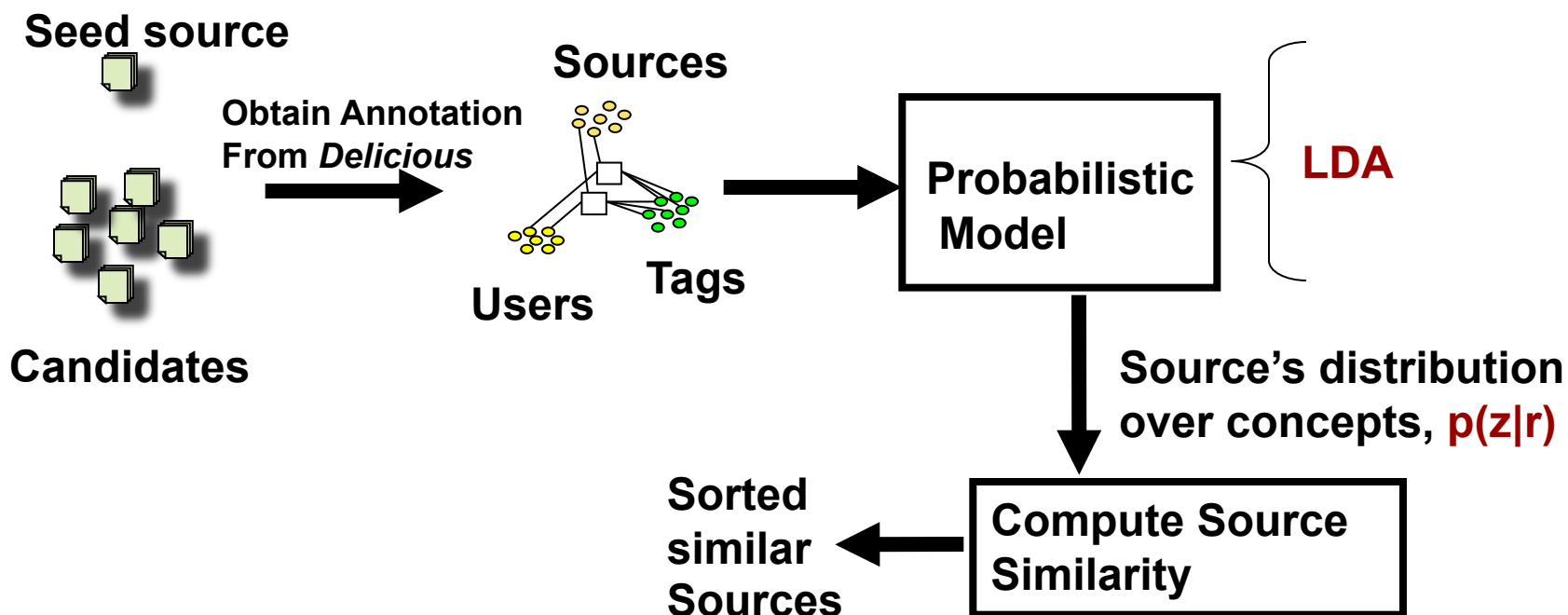
PLSA (Hofmann99);
LDA (Blei03+)



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

Exploiting Social Annotations for Resource Discovery

- Simplified resource discovery task : “*given a seed source, find other most similar sources*”

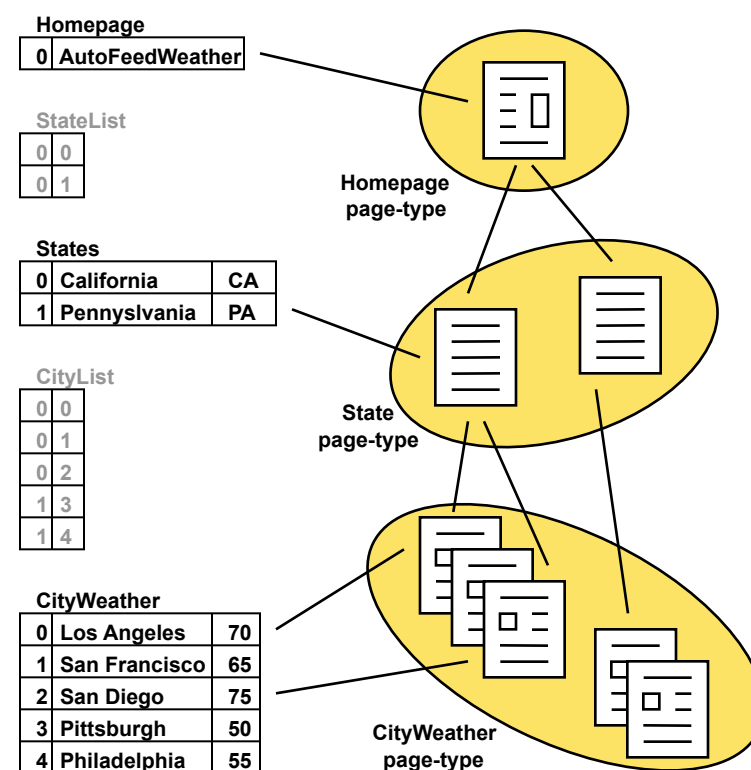




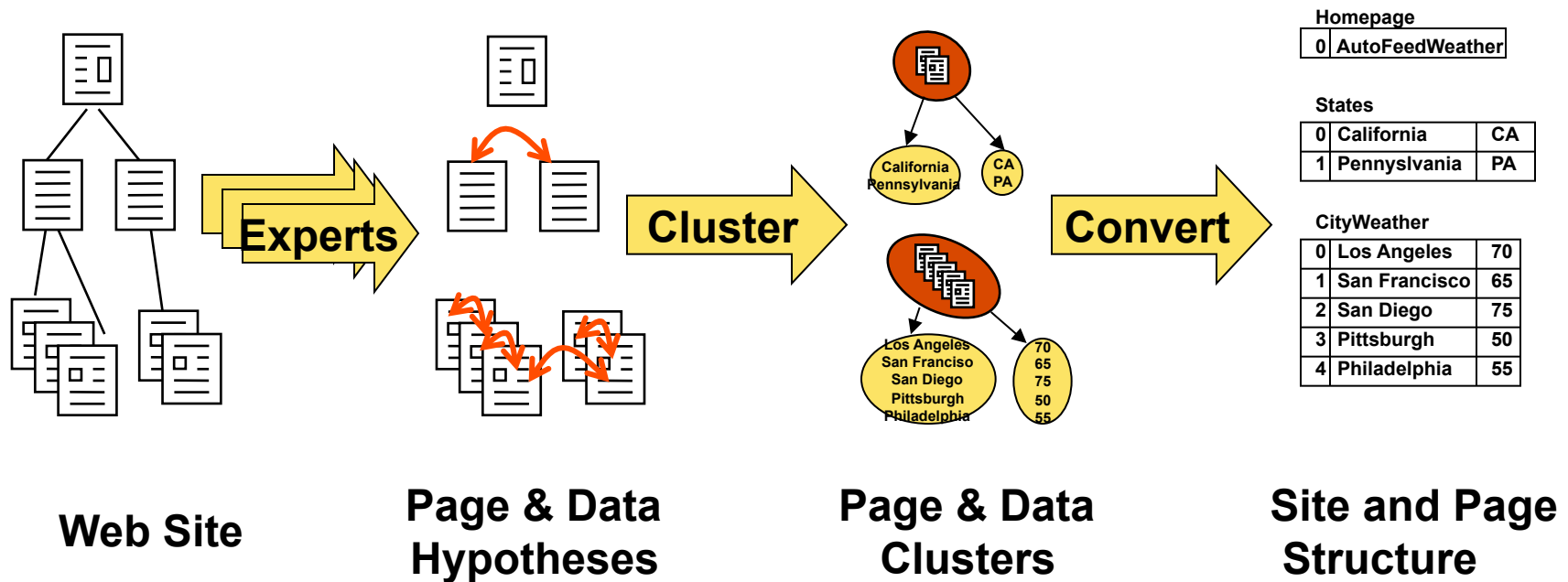
- Discovering sources using social annotations
- **Discovering the structure of sources**
- Learning semantic types of the source data
- Learning semantic models of the sources
- Experimental Results
- Discussion

Discovering Web Structure

- Goal:
 - 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.



Overview



- Page Templates
 - Similar pages contain common sequences of substrings

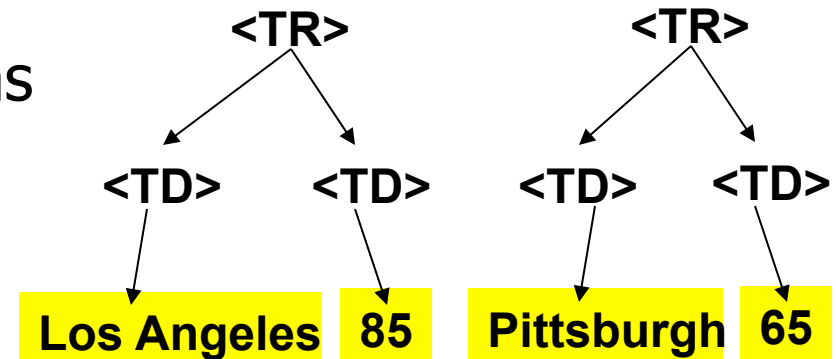
Right Now for
Los Angeles, CA (90007)
[Save this Location](#)

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)
- ...

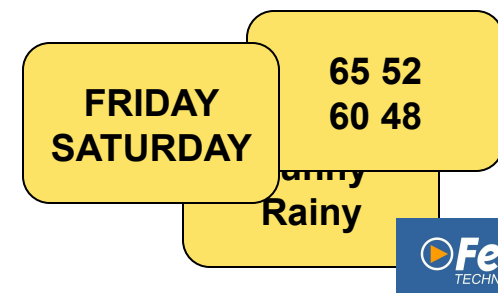


Extracted Data

FRIDAY	Sun	Sunny	65	52
SATURDAY	Rain	Rainy	60	48



Clusters





- Discovering sources using social annotations
- Discovering the structure of sources
- **Learning semantic types of the source data**
- Learning semantic models of the sources
- Experimental Results
- Discussion

Learning Patterns to Recognize Semantic Types

- Domain-independent token-level language to represent the structure of data as patterns
 - Token is a string or a general type
 - *90202 is a specific token*
 - *5DIGIT number is a general type*
 - Pattern is a sequence of tokens
 - *E.g., Phone numbers*

Sample values

310 448-8714

310 448-8775

212 555-1212

Patterns

[310 448 – 4DIGIT]

[3DIGIT 3DIGIT – 4DIGIT]

- Efficiently learn patterns from examples of semantic types
- Score the match between a type (patterns) and data

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]

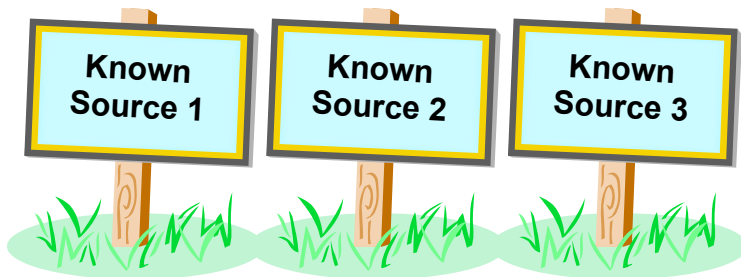
Labeled Columns of Target Source Unisys

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



- Discovering sources using social annotations
- Discovering the structure of sources
- Learning semantic types of the source data
- **Learning semantic models of the sources**
- Experimental Results
- Discussion

Inducing Source Definitions

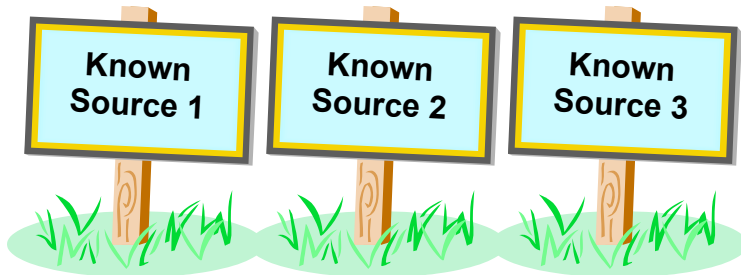


**source1(\$zip, lat, long) :-
centroid(zip, lat, long).**

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

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

Inducing Source Definitions



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

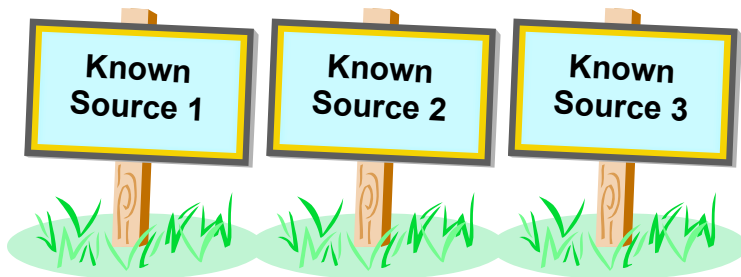
**source2(\$lat1, \$long1, \$lat2, \$long2, dist) :-
greatCircleDist(lat1, long1, lat2, long2, dist).**

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



source4(\$startZip, \$endZip, separation)

Inducing Source Definitions



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

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

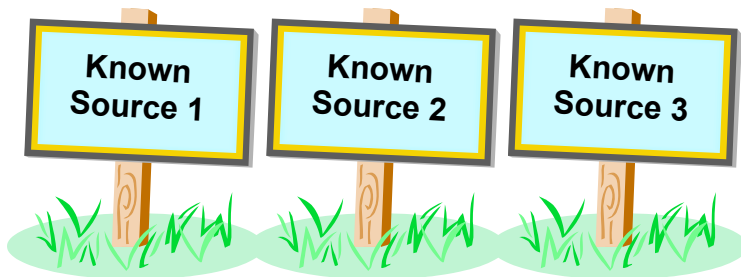
**source3(\$dist1, dist2) :-
convertKm2Mi(dist1, dist2).**

- Step 1: classify input & output semantic types



source4(\$startZip, \$endZip, separation)

Inducing Source Definitions

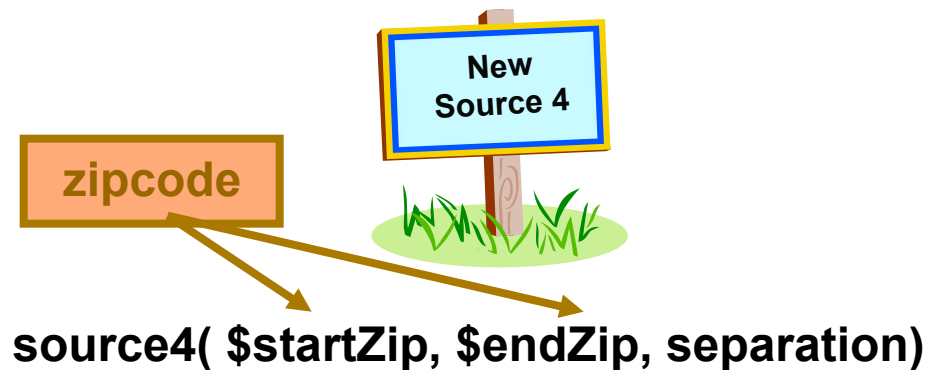


**source1(\$zip, lat, long) :-
centroid(zip, lat, long).**

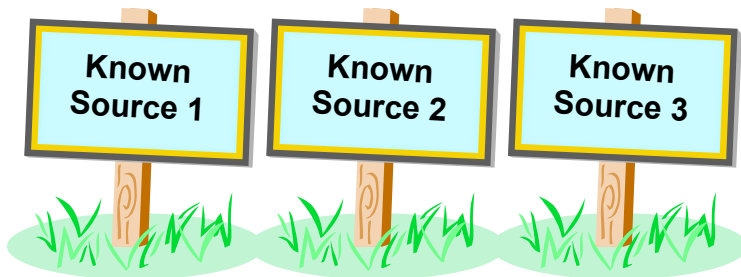
**source2(\$lat1, \$long1, \$lat2, \$long2, dist) :-
greatCircleDist(lat1, long1, lat2, long2, dist).**

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

- Step 1: classify input & output semantic types



Inducing Source Definitions

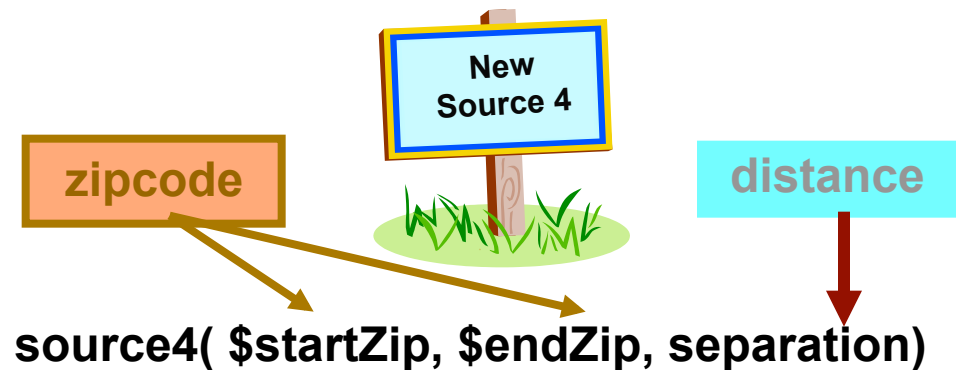


`source1($zip, lat, long) :-
centroid(zip, lat, long).`

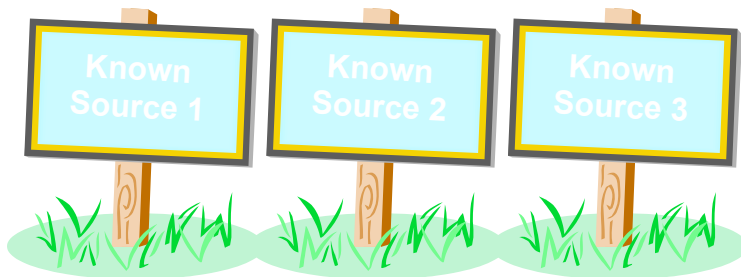
`source2($lat1, $long1, $lat2, $long2, dist) :-
greatCircleDist(lat1, long1, lat2, long2, dist).`

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

- Step 1: classify input & output semantic types



Generating Plausible Definition



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

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

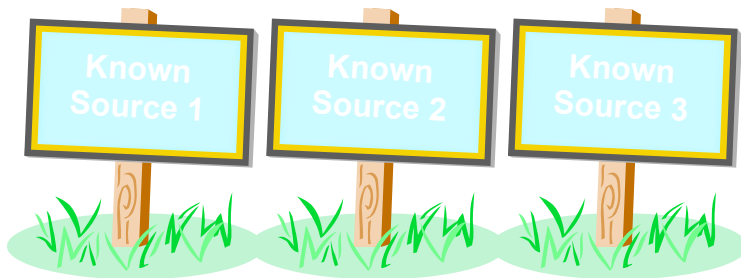
**source3(\$dist1, dist2) :-
convertKm2Mi(dist1, dist2).**

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



source4(\$zip1, \$zip2, dist)

Generating Plausible Definition



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

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

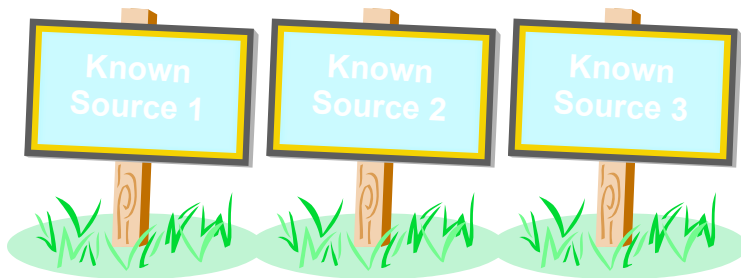
```
source3($dist1, dist2) :-  
    convertKm2Mi(dist1, dist2).
```

- 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)
```

Generating Plausible Definition



**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(dist2, dist).
```



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



- 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)`



- Discovering sources using social annotations
- Discovering the structure of sources
- Learning semantic types of the source data
- Learning semantic models of the sources
- **Experimental Results**
- Discussion

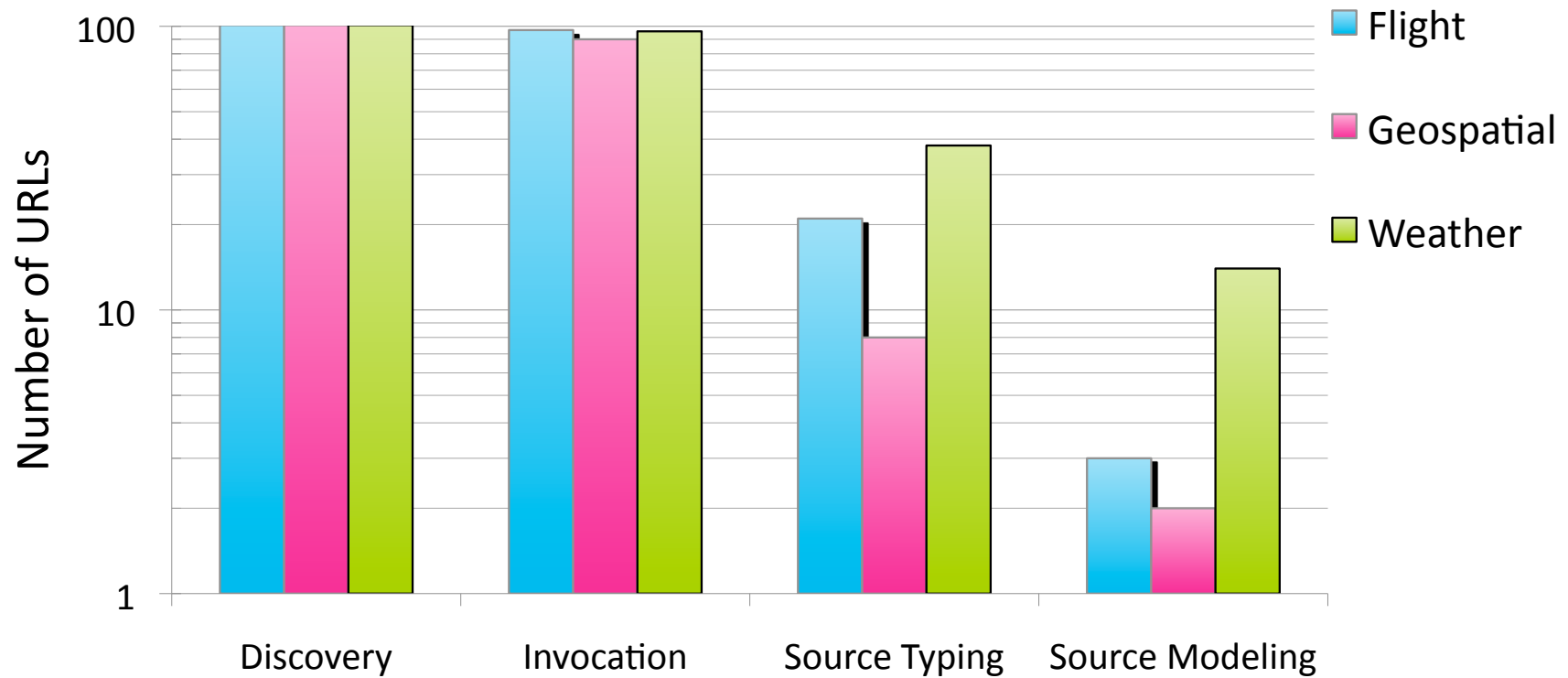


- Experiments in 3 domains
 - Geospatial
 - *Geocoder that maps street addresses into lat/long coordinates*
 - Weather
 - *Produces current and forecasted weather*
 - Flight Status
 - *Current status for a given airline and flight*
- Evaluation:
 - 1) Can we correctly learn a model for those sources that perform the same task
 - 2) What is the precision and recall of the attributes in the model

Candidate Sources after Each Step



URL Filtering by Module



Evaluation of the Models



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



- Discovering sources using social annotations
- Discovering the structure of sources
- Learning semantic types of the source data
- Learning semantic models of the sources
- Experimental Results
- Discussion

- ILA & Category Translation (Perkowitz & Etzioni 1995)
 - Learn functions describing operations on internet
- iMAP (Dhamanka et. al. 2004)
 - Discovers complex (many-to-1) mappings between DB schemas
- Metadata-based classification of data types used by Web services and HTML forms (Hess & Kushmerick, 2003)
 - Naïve Bayes classifier
- Woogle: Metadata-based clustering of data and operations used by Web services (Dong et al, 2004)
 - Groups similar types together: Zipcode, City, State



- Integrated a diverse set of learning and reasoning techniques
 - *Discover new sources*
 - *Discover the template for a source*
 - *Find the semantic types of source data*
 - *Learn a definition of what a source does*
- Provides an end-to-end completely automatic approach to discover and build models of sources