USC

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

Goal

USC Viterbi School of Engineering

- 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

Approach

USC Viterbi School of Engineering

- 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

USC

Outline

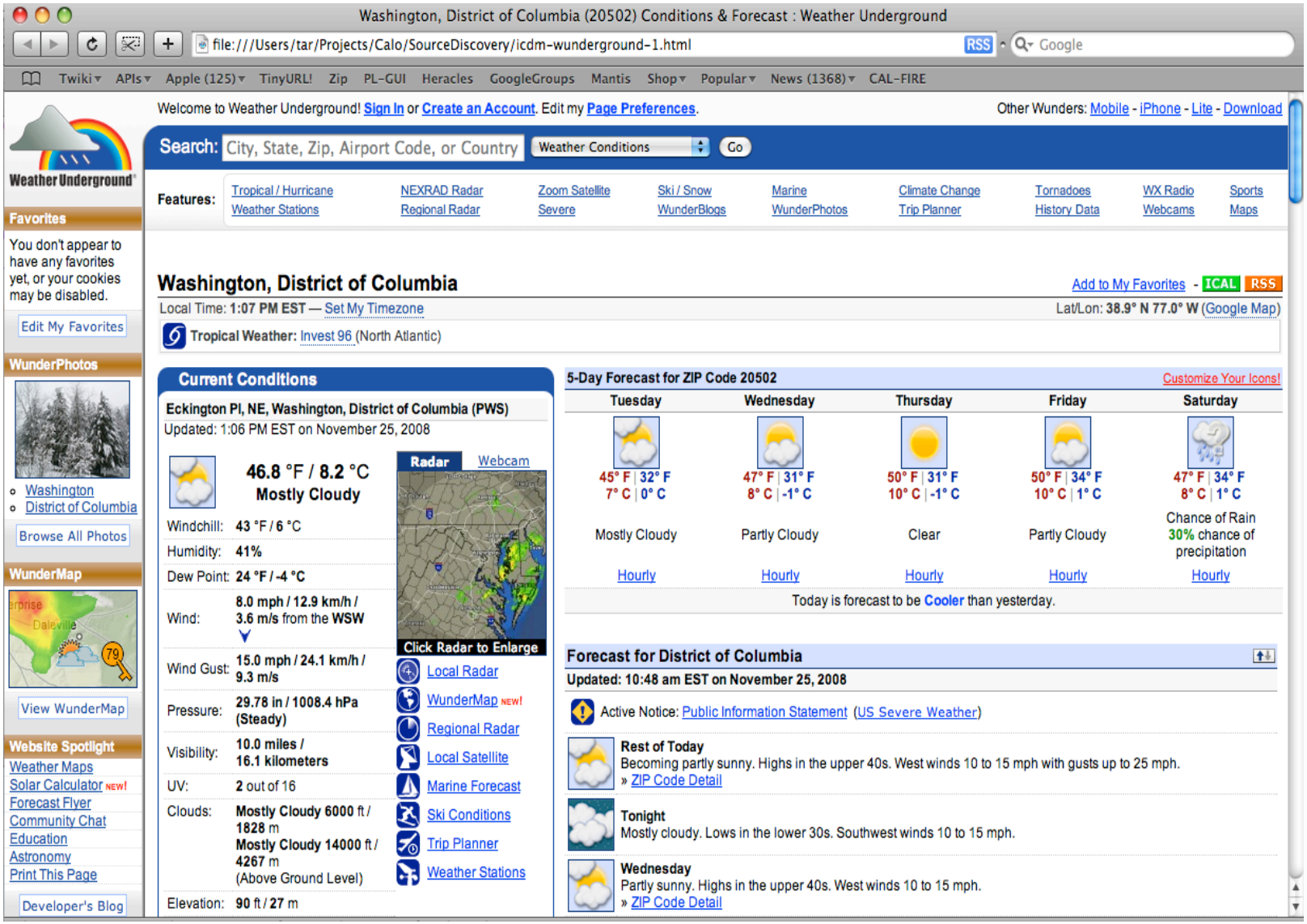
USC Viterbi School of Engineering

- 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

USC

Seed Source

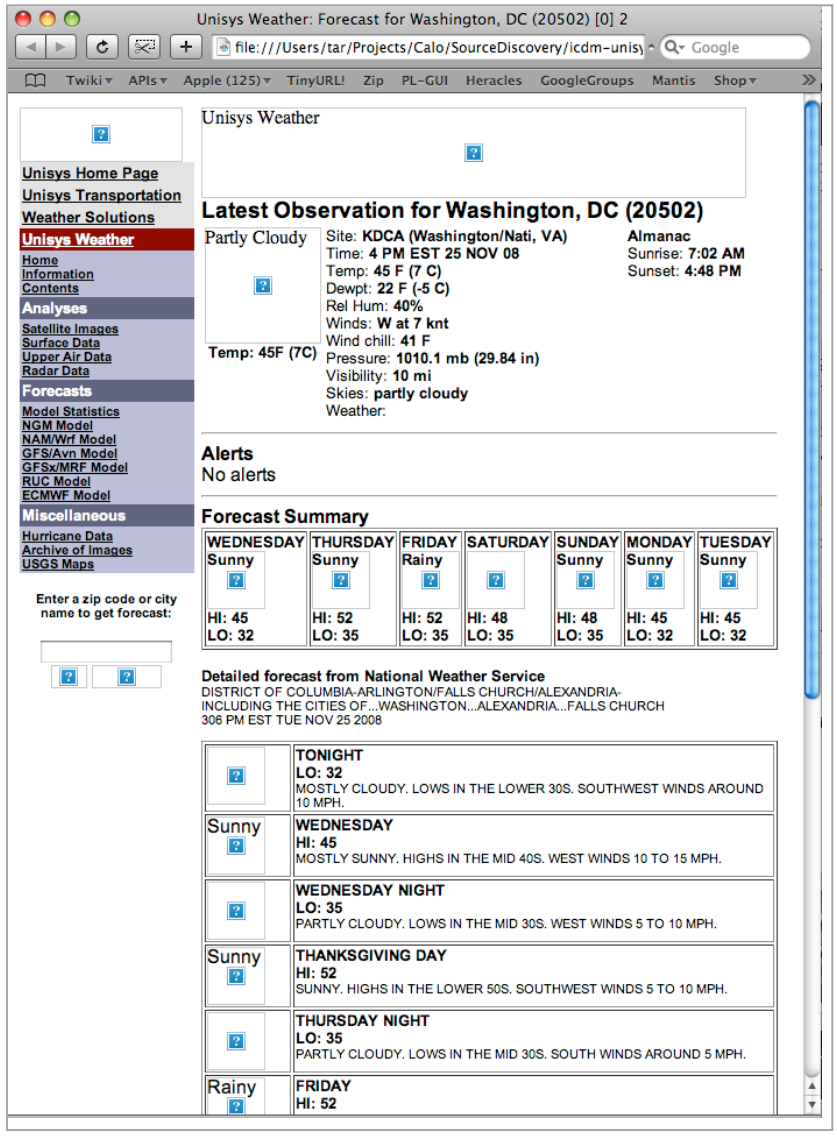
USC Viterbi School of Engineering



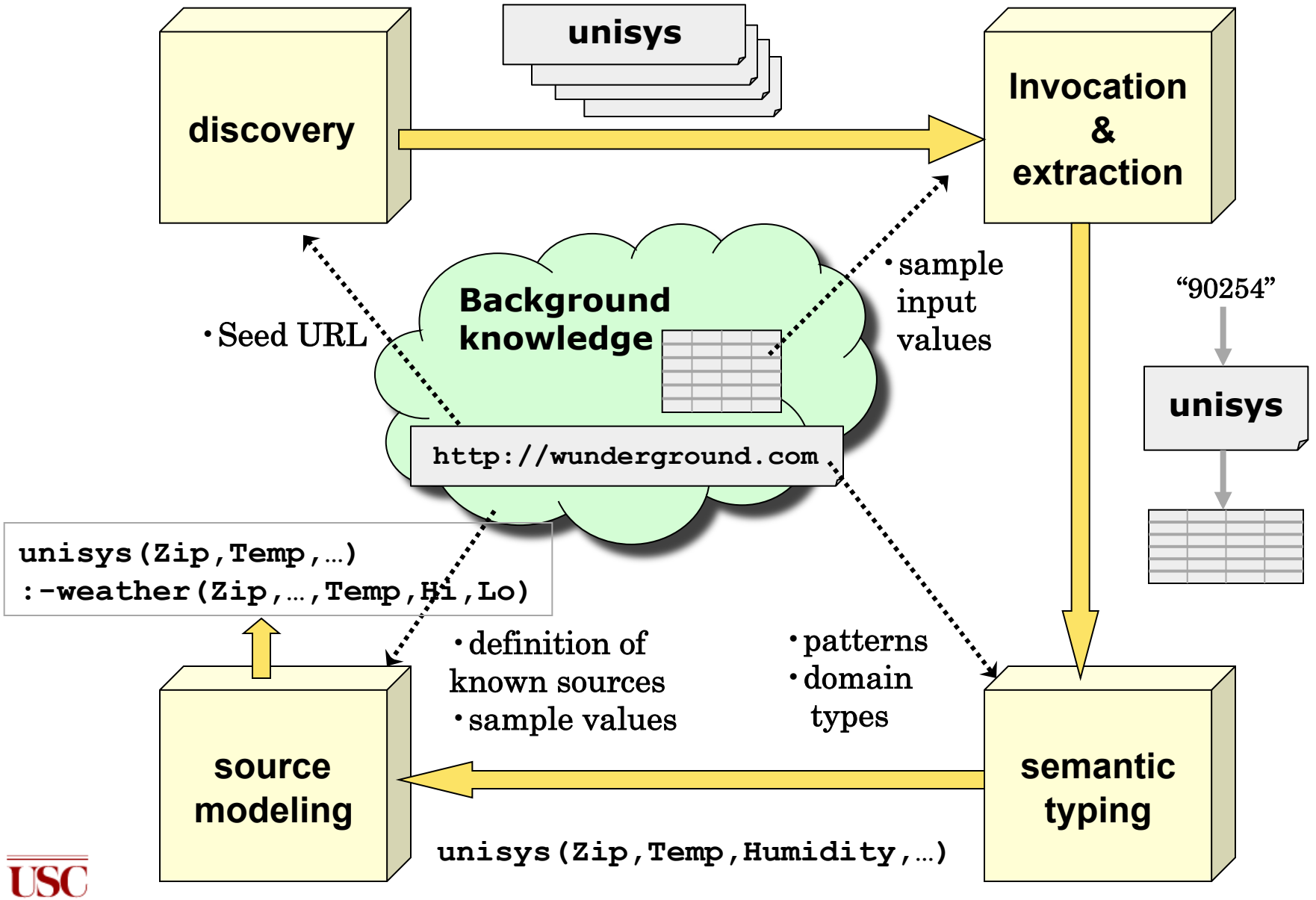
USC

Automatically Discover and Model a Source in the Same Domain

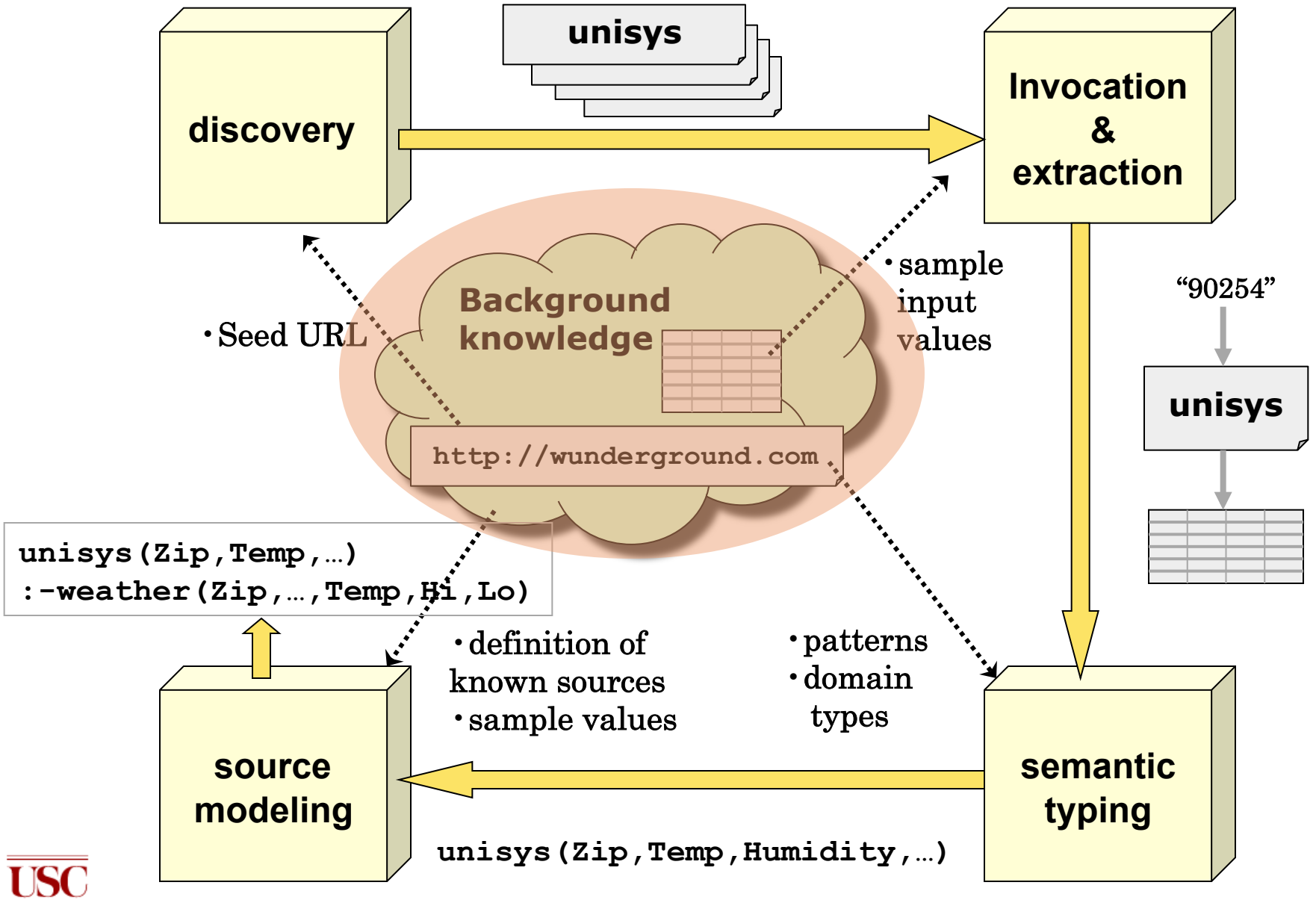
USC Viterbi School of Engineering



Integrated Approach



Background Knowledge



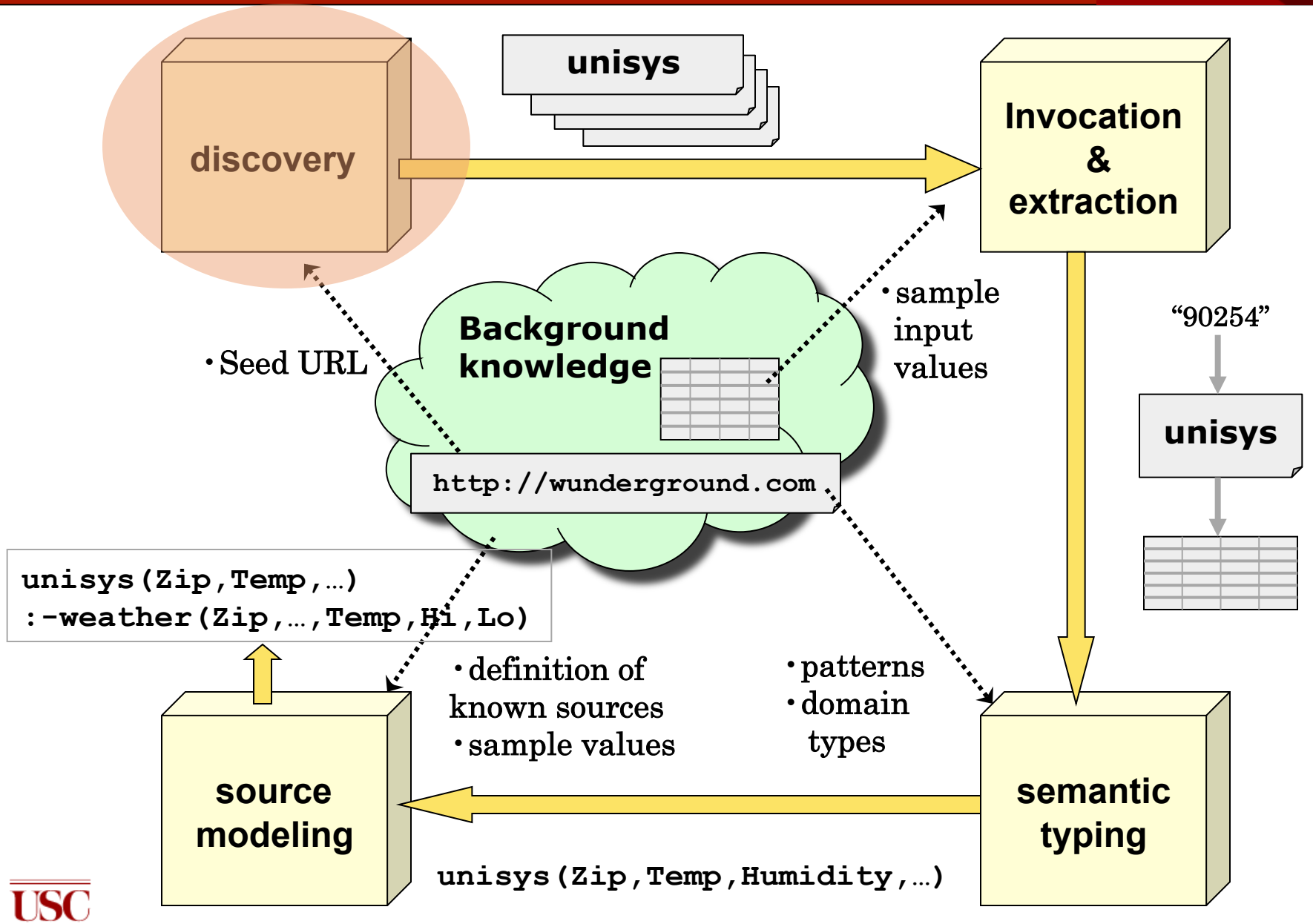
Background Knowledege

USC Viterbi School of Engineering

- 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,_,_,_,_,_).

USC

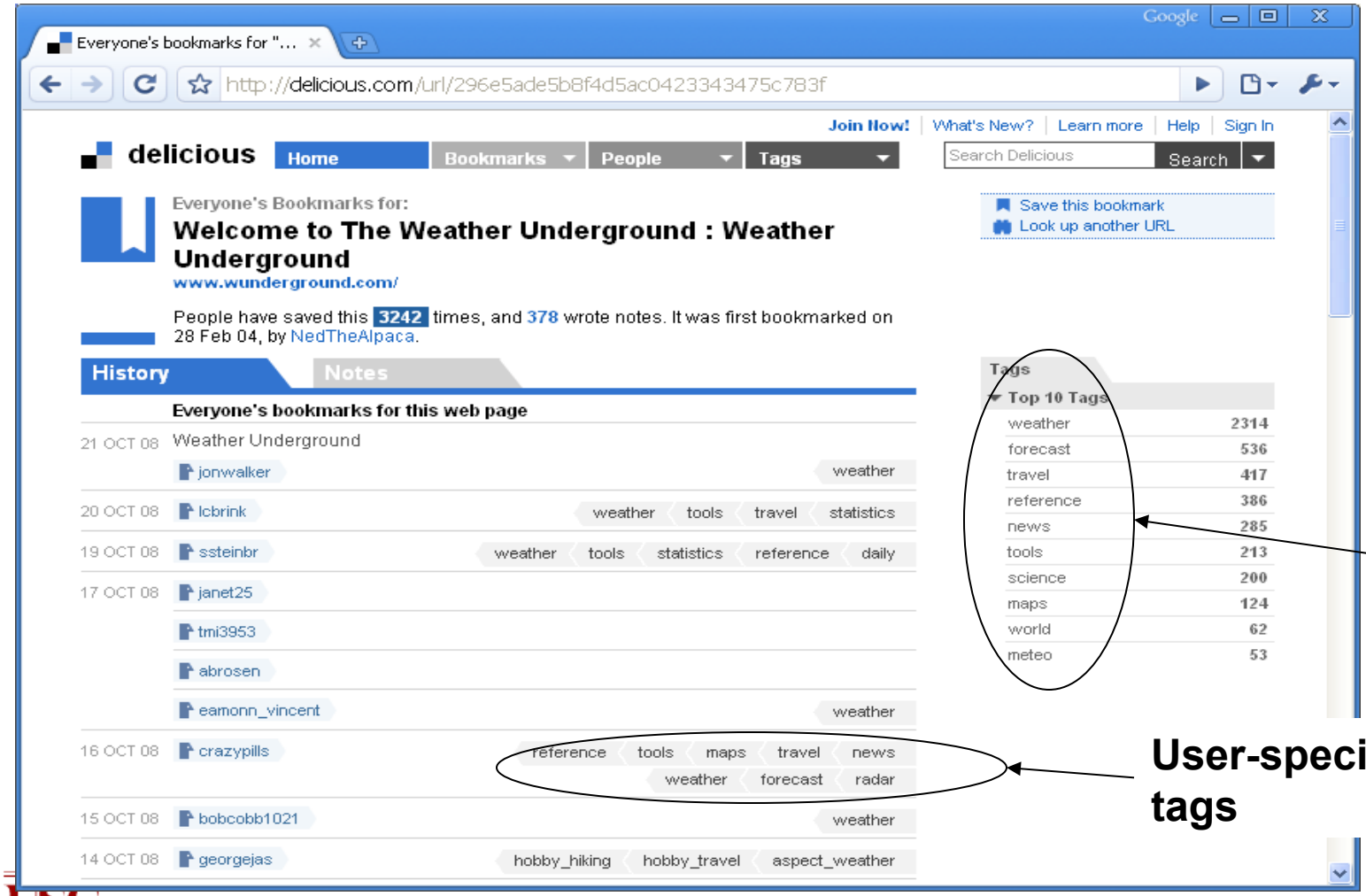
Source Discovery



Source Discovery [Plangprasopchok and Lerman]

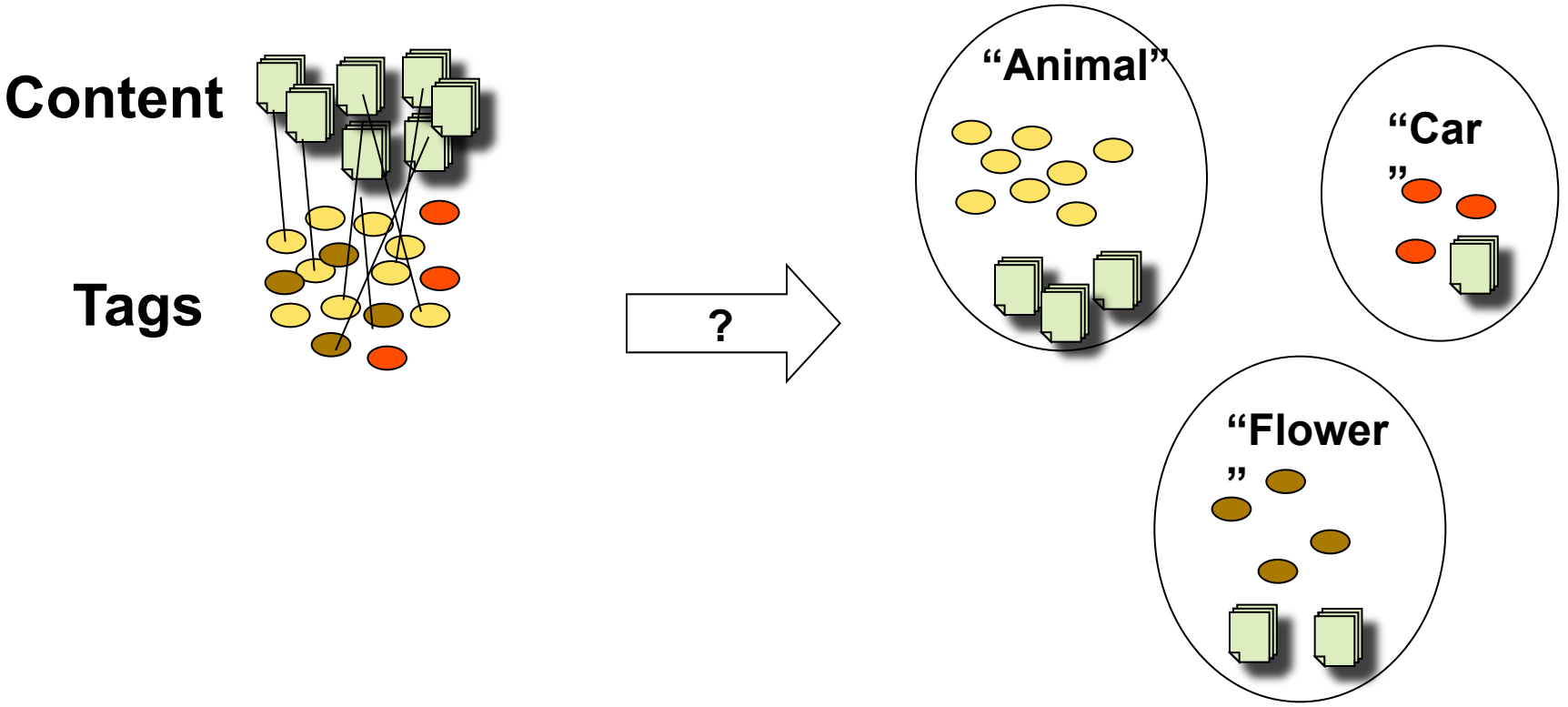
USC Viterbi School of Engineering

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



Group Tags and Content into Concepts

USC Viterbi School of Engineering



Group semantically related tags and content

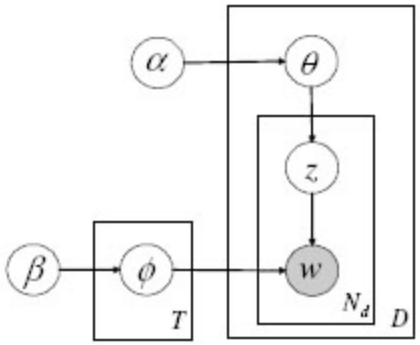
USC

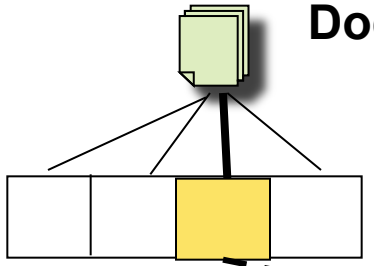
A group ~ A concept

A Stochastic Process of Tag Generation

USC Viterbi School of Engineering

PLSA (Hofmann99);
LDA (Blei03+)





Document (r)

Concepts (z)

Possible Words

Possible Concepts

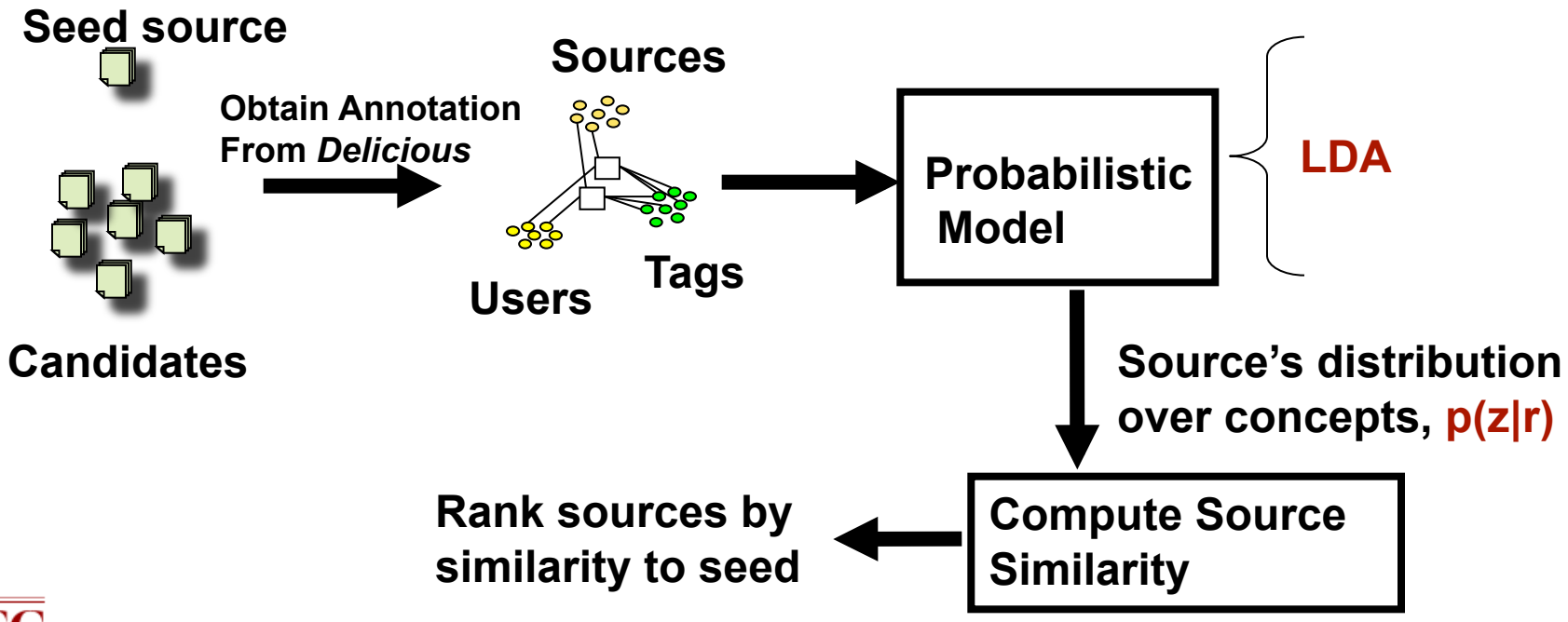
Generated tags (t)

USC **A data point (tuple) <r,t,z>**

Exploiting Social Annotations for Resource Discovery

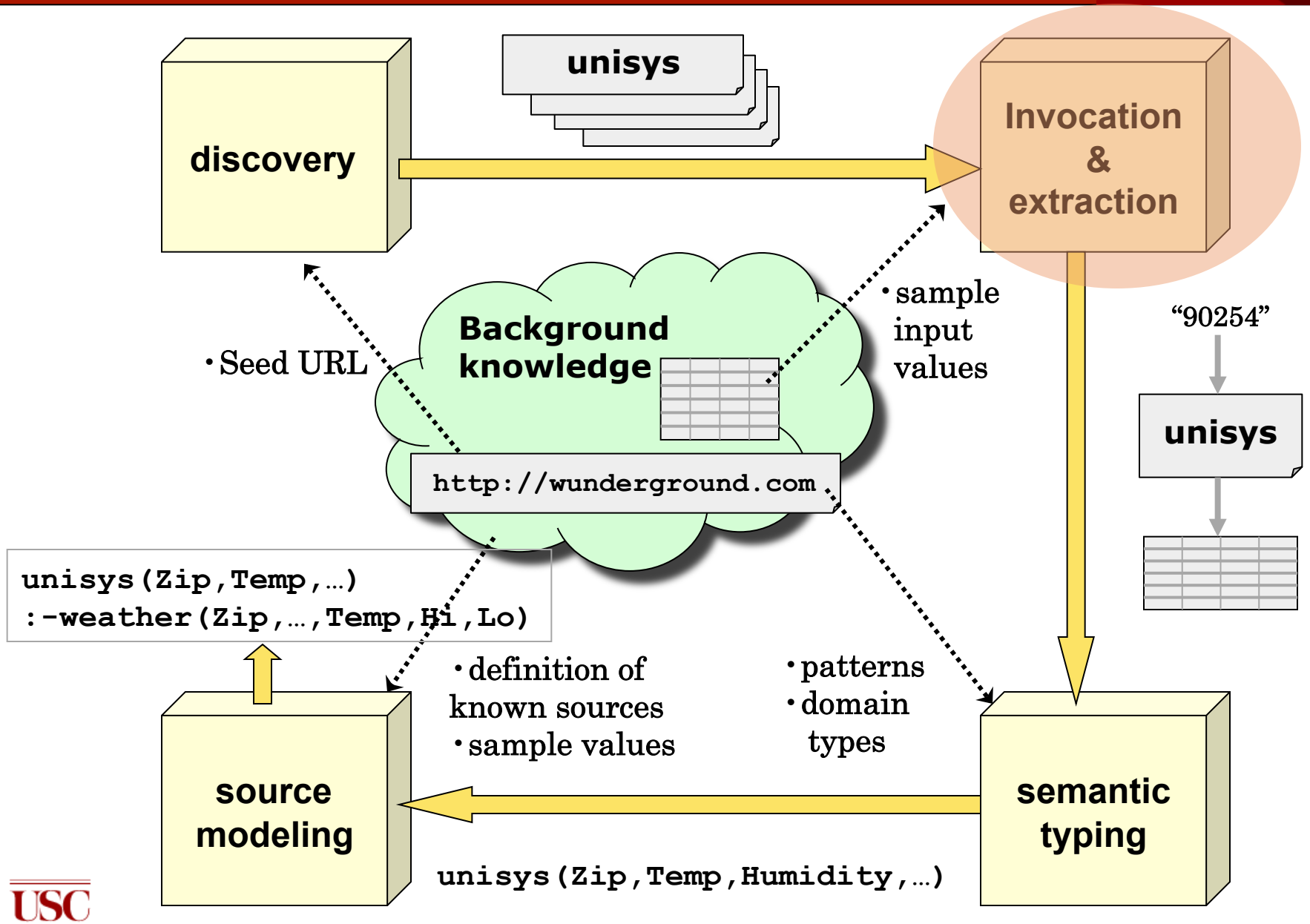
USC Viterbi School of Engineering

- 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



USC

Source Invocation & Extraction



Target Source Invocation

USC Viterbi School of Engineering

- 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

USC

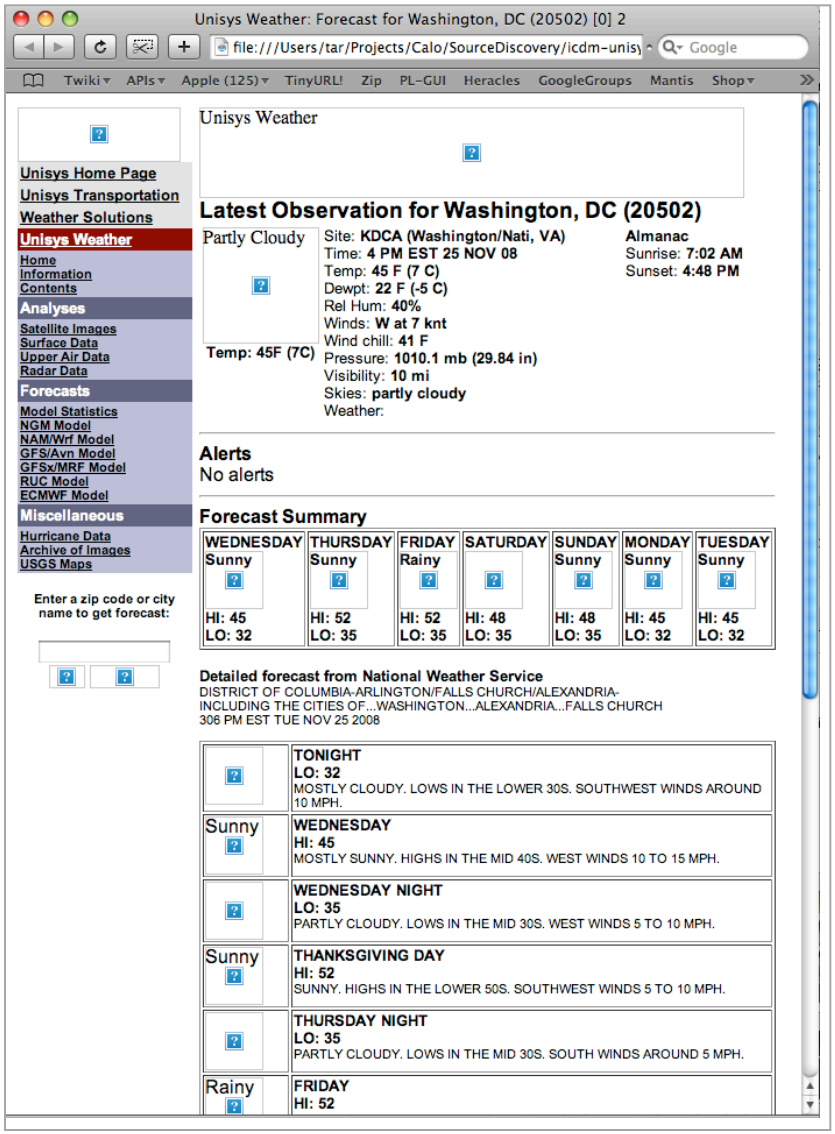
Invoke the Target Source with Possible Inputs

USC Viterbi School of Engineering

http://weather.unisys.com

input

Weather conditions for 20502



Form Input Data Model

USC Viterbi School of Engineering

- 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

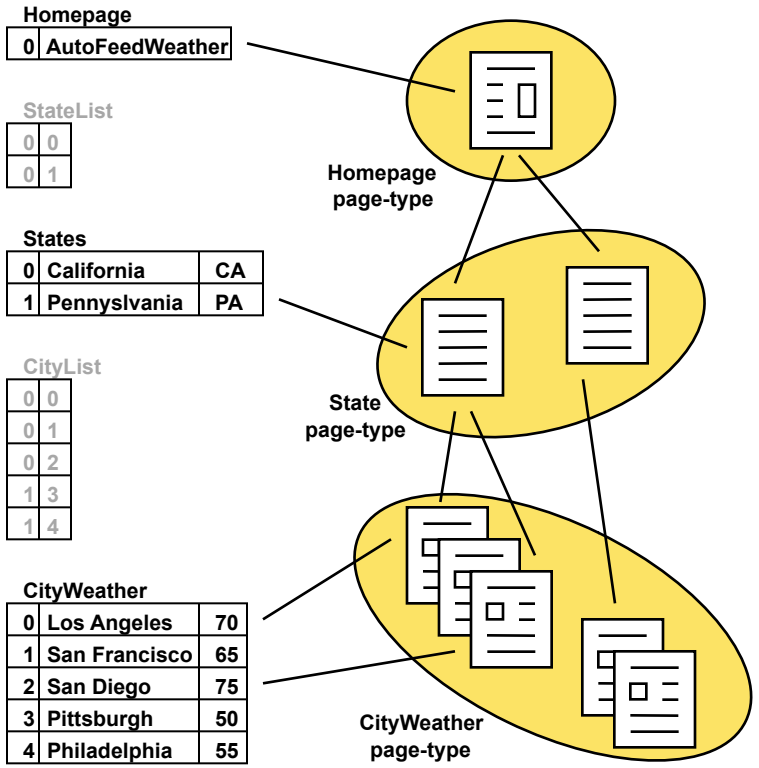
PR-Zip	PR-CityState	PR-City	PR-StateAbbr
20502	Washington, DC	Washington	DC
32399	Tallahassee, FL	Tallahassee	FL
33040	Key West, FL	Key West	FL
90292	Marina del Rey, CA	Marina del Rey	CA
36130	Montgomery, AL	Montgomery	AL

USC

Discovering Web Structure [Gazen & Minton]

USC Viterbi School of Engineering

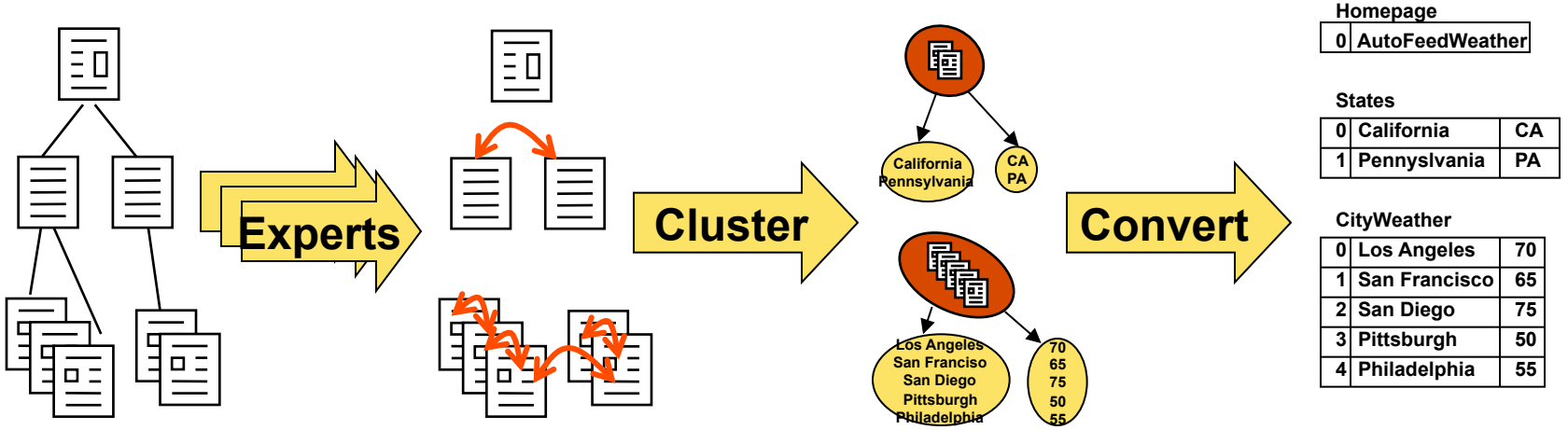
- Model Web sources that generate pages dynamically in response to a query
  - Find the relational data underlying a semi-structured web site
- Generate a page template that can be used to extract data on new pages
- Approach
  - *Site extraction*
    - Exploit the common structure within a web site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



USC

Fetch TECHNOLOGIES

Approach to Finding Web Structure



Web Site

Page & Data Hypotheses

Page & Data Clusters

Site and Page Structure

Sample Experts

USC Viterbi School of Engineering

- 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.:

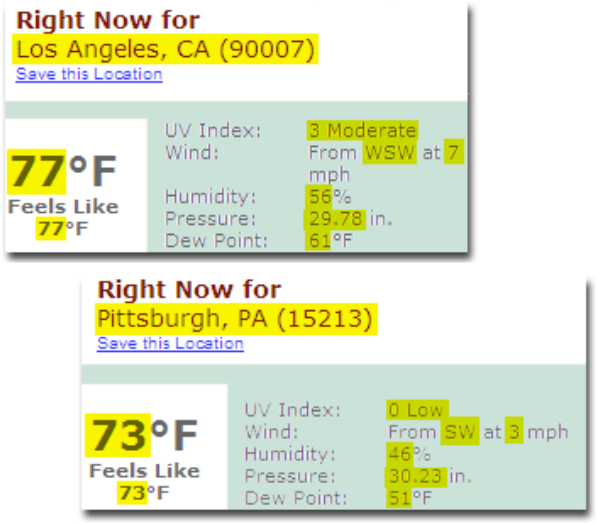
USC

Fetch TECHNOLOGIES

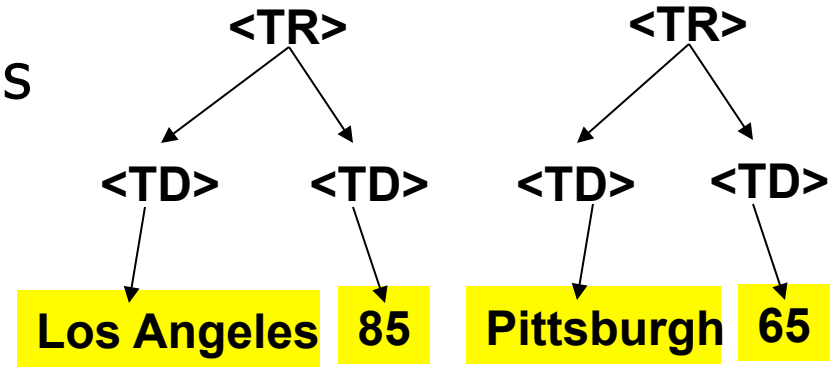
Sample Experts

USC Viterbi School of Engineering

- Page Templates
  - Similar pages contain common sequences of substrings



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



USC

Fetch TECHNOLOGIES

Extracting Data

USC Viterbi School of Engineering

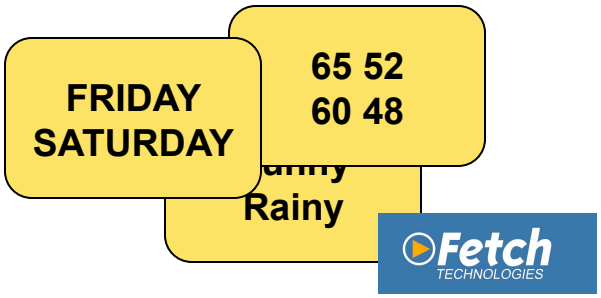
Pages

```
<td valign="top" width="14%">
<td valign="top" width="14%">
 <font face="Arial, Helvetica, sans-serif">
  <small><b>FRIDAY<br>
  <img src="images/Sun-s.png" alt="Sunny"><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>
<img src="images/Rain-s.png" alt="Rainy"><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)**
- **…**

Clusters



Extracted Data

FRIDAY	Sun	Sunny	65	52
SATURDAY	Rain	Rainy	60	48

USC

Fetch TECHNOLOGIES

Data Extraction with Templates

USC Viterbi School of Engineering

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

Unseen Page

```
<img src="images/Sun.png" alt="Sunny"><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

```
<img src="images/*.png" alt="*"><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

USC

Fetch TECHNOLOGIES

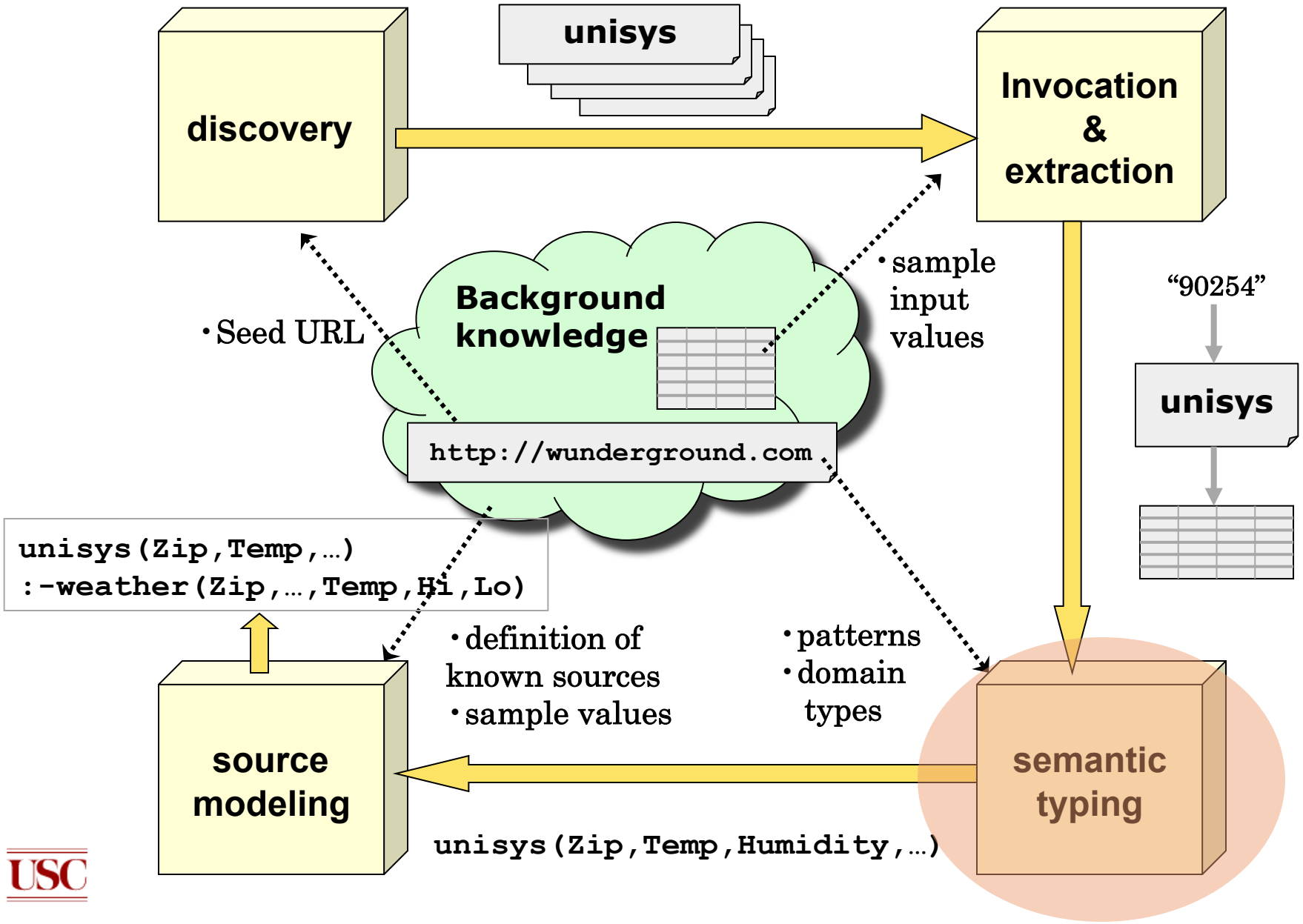
Raw Extracted Data from Unisys

USC Viterbi School of Engineering

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.	

USC

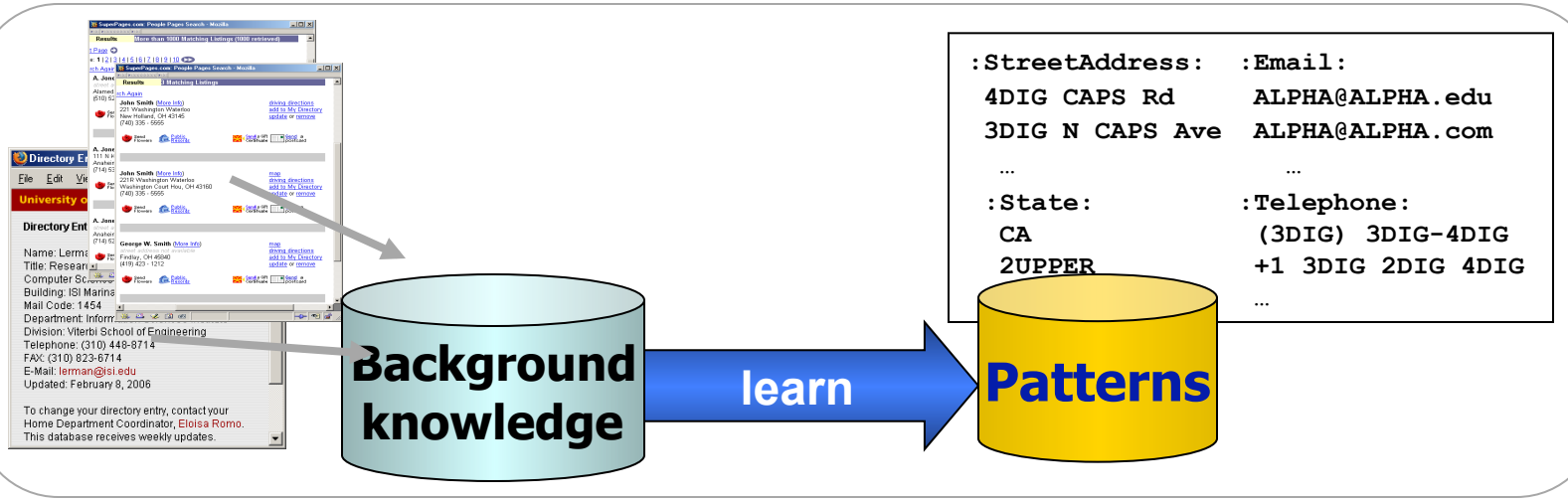
Semantic Typing



Semantic Typing [Lerman, Plangprasopchok, & Knoblock]

USC Viterbi School of Engineering

✓**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 -
C. S. Lewis	555 Willow Run Dr	(612) 578 -
Carmen Jones	355 Morgan Ave N	(612) 522 -
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
drew Lewis	3543 Larchmont Pkwy	(518) 474 - 4799
Lewis	555 Willow Run Dr	(612) 578 - 5555
rmen 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

USC

Learning Patterns to Recognize Semantic Types

USC Viterbi School of Engineering

- 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	Patterns
310 448-8714	
310 448-8775	*[(310) 448 – 4DIGIT]*
212 555-1212	*[(3DIGIT) 3DIGIT – 4DIGIT]*

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

Labeling New Data

USC Viterbi School of Engineering

- 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
  …                 …
```

USC

Weather Data Types

USC Viterbi School of Engineering

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]

USC

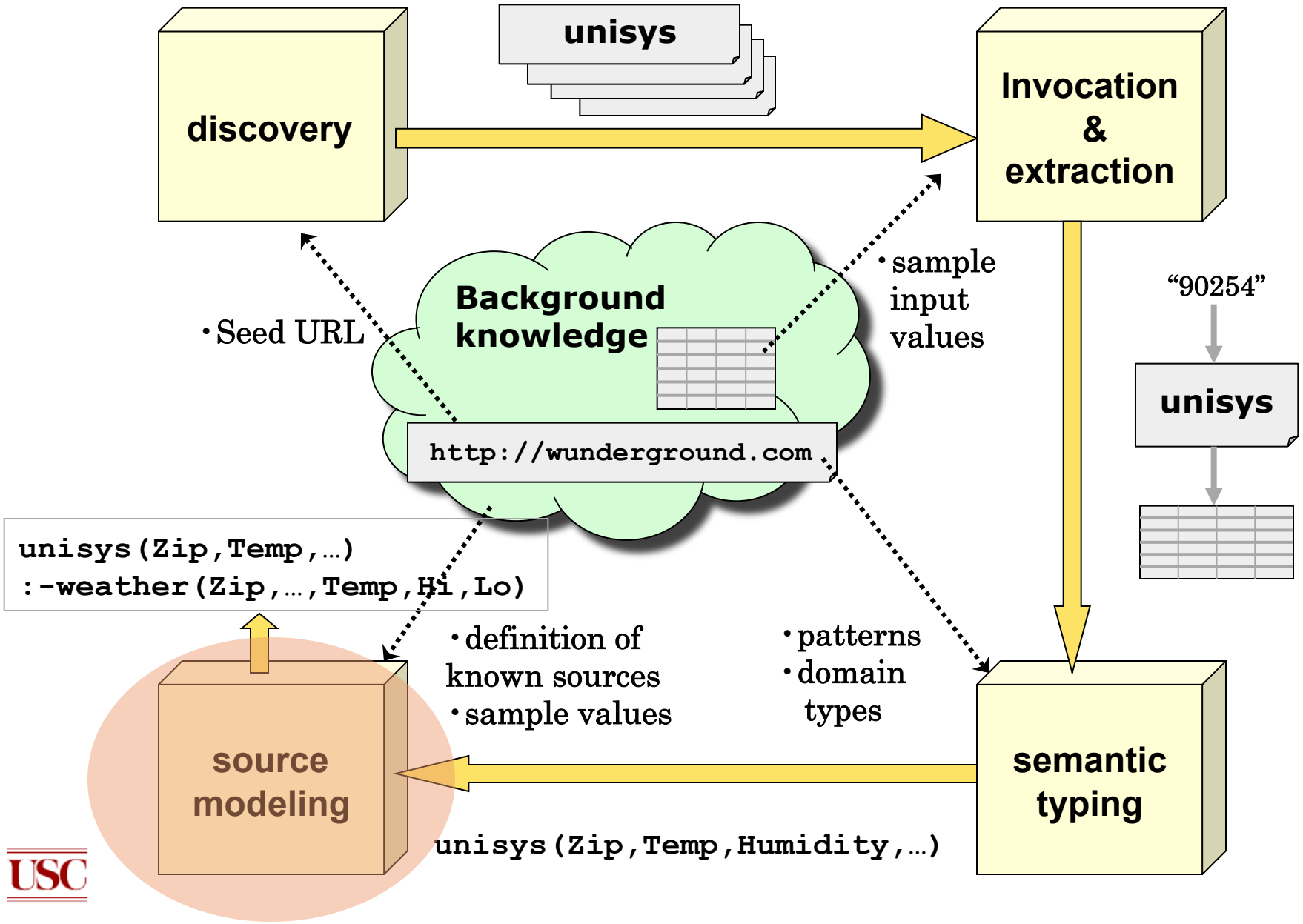
Labeled Columns of Target Source Unisys

USC Viterbi School of Engineering

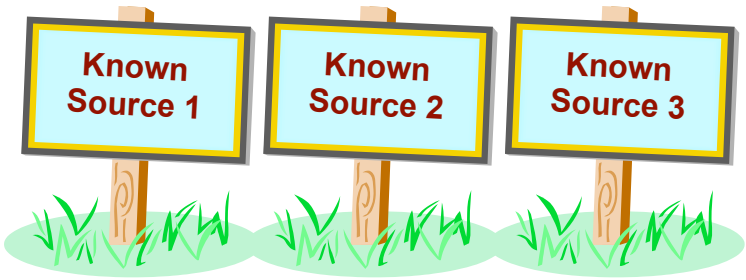
Column	4	18	25	15	87
Type	**PR-Zip**	**PR-TempF**	**PR-Humidity**	**PR-Sky**	**PR-Sky**
Score	0.333	0.68	1.0	0.325	0.375
Values	20502	45F	40%	Partly Cloudy	Sunny
	32399	63F	23%	Sunny	Partly Cloudy
	33040	73F	73%	Sunny	Rainy
	90292	66F	59%	Partly Cloudy	Sunny
	36130	62F	24%	Sunny	Partly Cloudy

USC

Source Modeling [Carman & Knoblock]

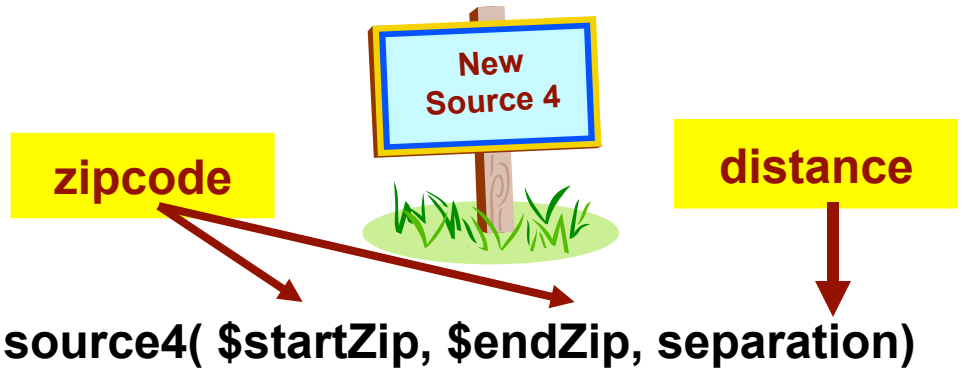


Inducing Source Definitions

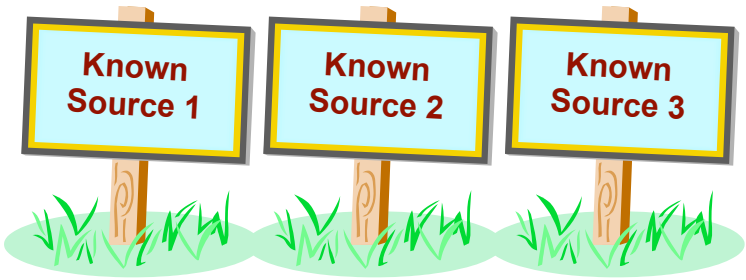



```
source1($zip, lat, long) :-
    centroid(zip, lat, long).

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




```
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).
```

Top-down Generation of Candidates

USC Viterbi School of Engineering

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(_,_,_,_).

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

Expand

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

Invoke and Compare the Definition

USC Viterbi School of Engineering

- 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).
```

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

Approximating Equality

USC Viterbi School of Engineering

Allow flexibility in values from different sources

- Numeric Types like *distance*

  10.6 km ≈ 10.54 km

  Error Bounds (eg. +/- 1%)

- Nominal Types like *company*

  Google Inc. ≈ Google Incorporated

  String Distance Metrics
  (e.g. JaroWinkler Score > 0.9)

- Complex Types like *date*

  Mon, 31. July 2006 ≈ 7/31/06

  Hand-written equality checking procedures.

Example of a Learned Source Model for Weather Domain

USC Viterbi School of Engineering

- Given a set of known sources and their descriptions
  - wunderground($Z,CS,T,F0,S0,Hu0,WS0,WD0,P0,V0) :- weather(0,Z,CS,D,T,F0,_,_,S0,Hu0,P0,WS0,WD0,V0)
  - convertC2F(C,F) :- centigrade2farenheit(C,F)
- Learn a description of a new source in terms of the known sources
  - unisys($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)

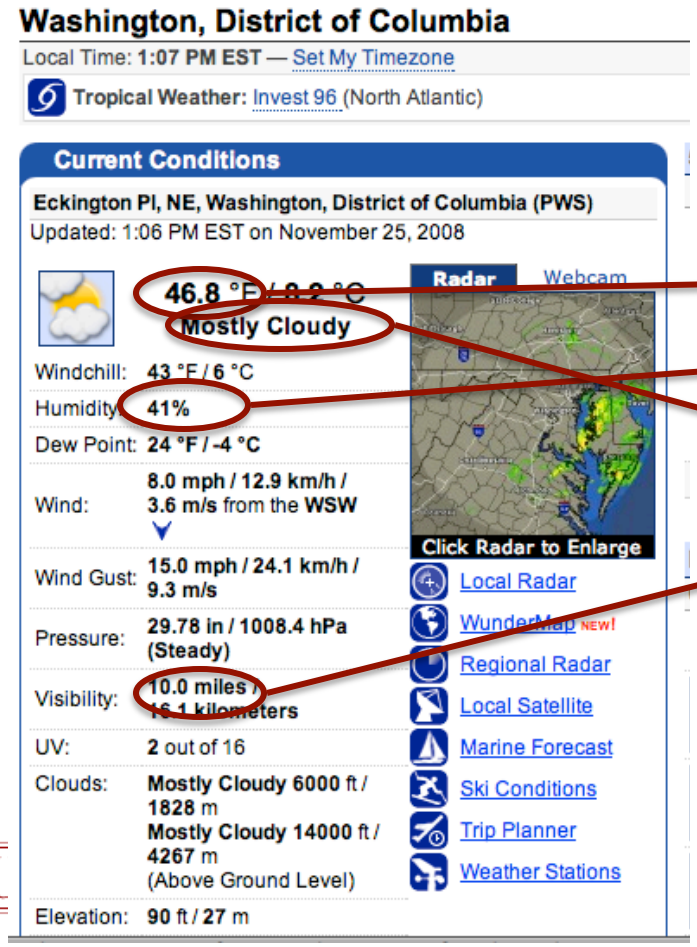
USC

Evaluate the Candidate Definition

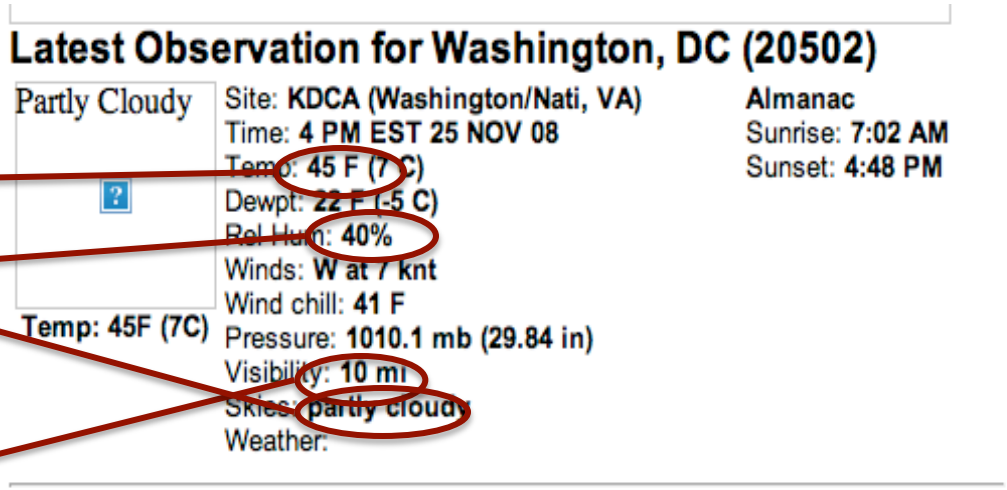
USC Viterbi School of Engineering

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

Seed (wunderground.com)



Target (unisys.com)



Issues in the End-to-End Integration

USC Viterbi School of Engineering

- 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*

Outline

USC Viterbi School of Engineering

- 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

USC

Experimental Evaluation

USC Viterbi School of Engineering

- 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

Experiments: Source Discovery

USC Viterbi School of Engineering

- 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*

Experiments: Source Invocation & Extraction, Semantic Typing, and Source Modeling

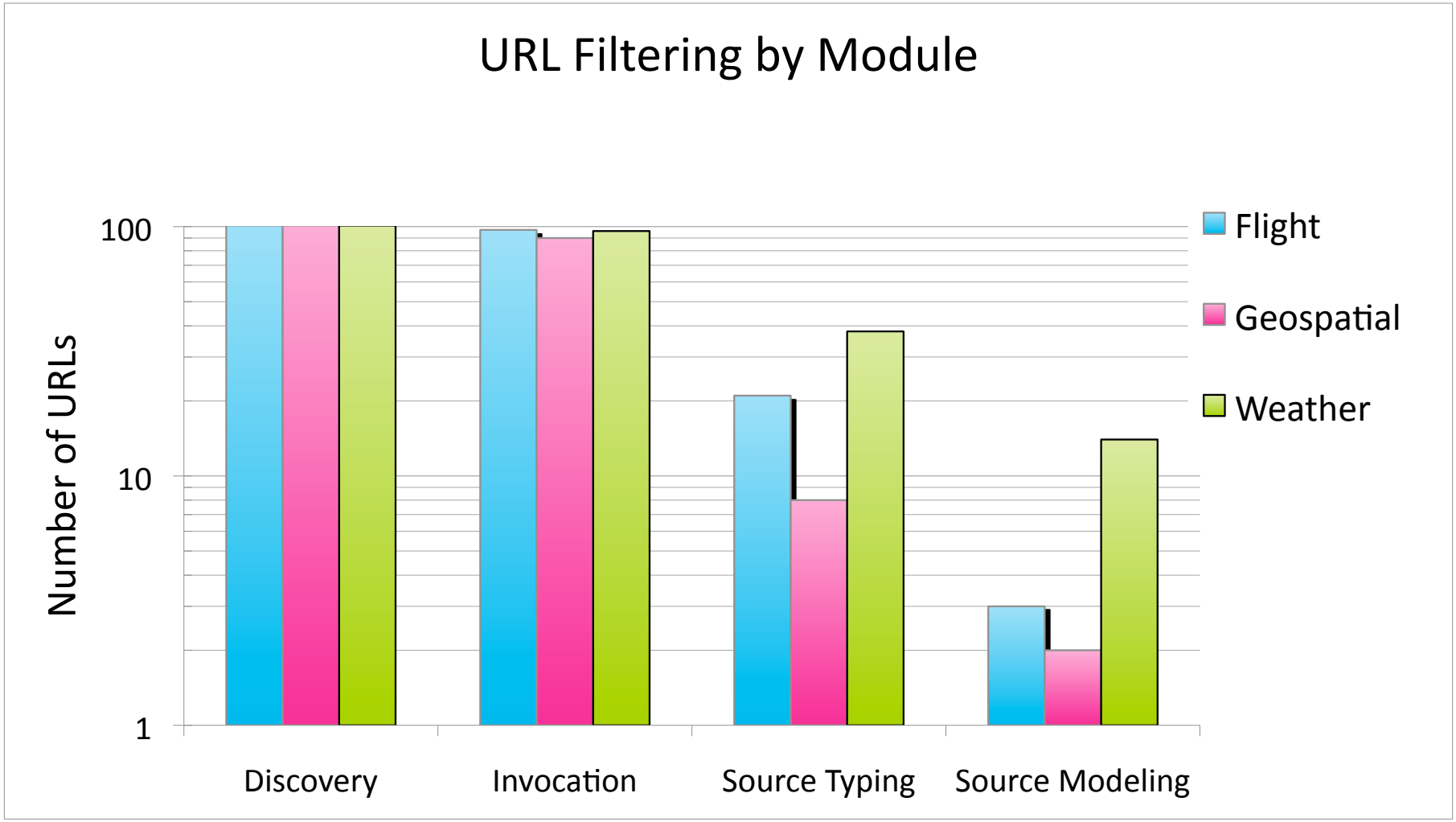
USC Viterbi School of Engineering

- **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 if the input
  - Evaluate accuracy of the resulting source model

USC

Candidate Sources after Each Step

USC Viterbi School of Engineering



USC

Confusion Matrix

USC Viterbi School of Engineering

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

USC

Evaluation of the Models

USC Viterbi School of Engineering

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

USC

Outline

USC Viterbi School of Engineering

- 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

USC

Related Work

USC Viterbi School of Engineering

- 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

Related Work (cont.)

USC Viterbi School of Engineering

- 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

Outline

USC Viterbi School of Engineering

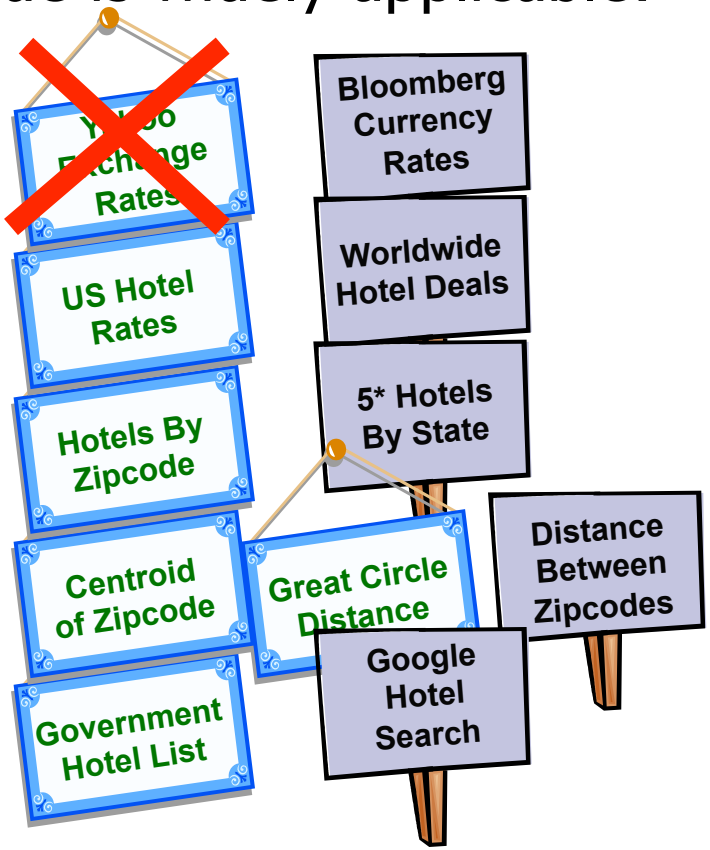
- Integrated Approach
  - Discovering related sources
  - Constructing syntactic models of the sources
  - Determining the semantic types of the data
  - Building semantic models of the sources
- Experimental Results
- Related Work
- Discussion

USC

Coverage

USC Viterbi School of Engineering

- Assumption: overlap between new & known sources
- Nonetheless, the technique is widely applicable:
  - Redundancy
  - Scope or Completeness
  - Binding Constraints
  - Composed Functionality
  - Access Time



Discussion

USC Viterbi School of Engineering

- 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

Future Work

USC Viterbi School of Engineering

- 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 Farenheit and Celsius*
  - Learn the domain and range of the sources
    - *Learn that a source provides world weather vs. US weather*

USC

Acknowledgements & Papers

- 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]*