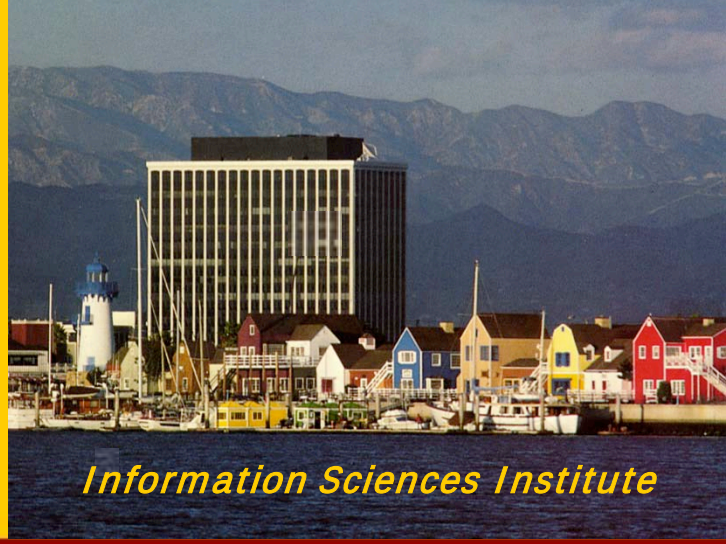
USC Viterbi
School of Engineering




# Exploiting Semantics of Web Services for Geospatial Data Fusion

**Pedro Szekely, Craig A. Knoblock, Shubham Gupta, Mohsen Taheriyan, Bo Wu**

**University of Southern California**

USC

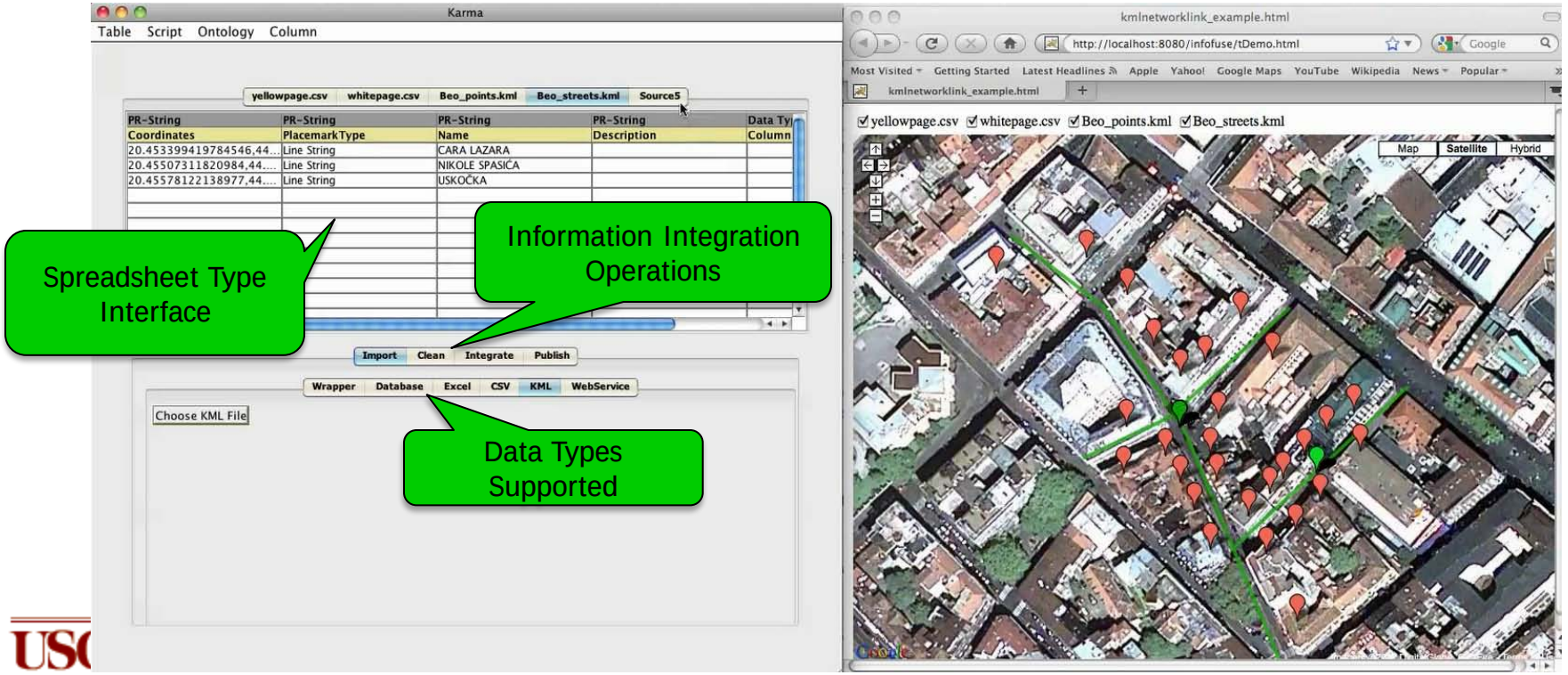
# Introduction



- Decision makers have lots of data available
  - Satellite imagery
  - Street maps
  - Structured online sources (e.g., phone books)
  - Cyber data (e.g., domain registration sites)
  - Social network data (e.g., facebook)
- Difficult to fuse this information into an integrated view
  - Even harder to apply various reasoning techniques
- Our goal
  - An integration framework where users can interactively fuse geospatial and other types of data

# KARMA: A General Information Integration Tool

- Karma [R. Tuchinda, C. A. Knoblock, P. Szekely, Building mashups by demonstration, 2011]
- A fusion-by-example approach for extracting, modeling, cleaning and integrating geospatial sources
  - Does not require any programming or widget knowledge.
  - Focus on data, not on the process
  - Users specify fusion tasks by examples
  - Fusion results automatically displayed on a map



# Motivating Example



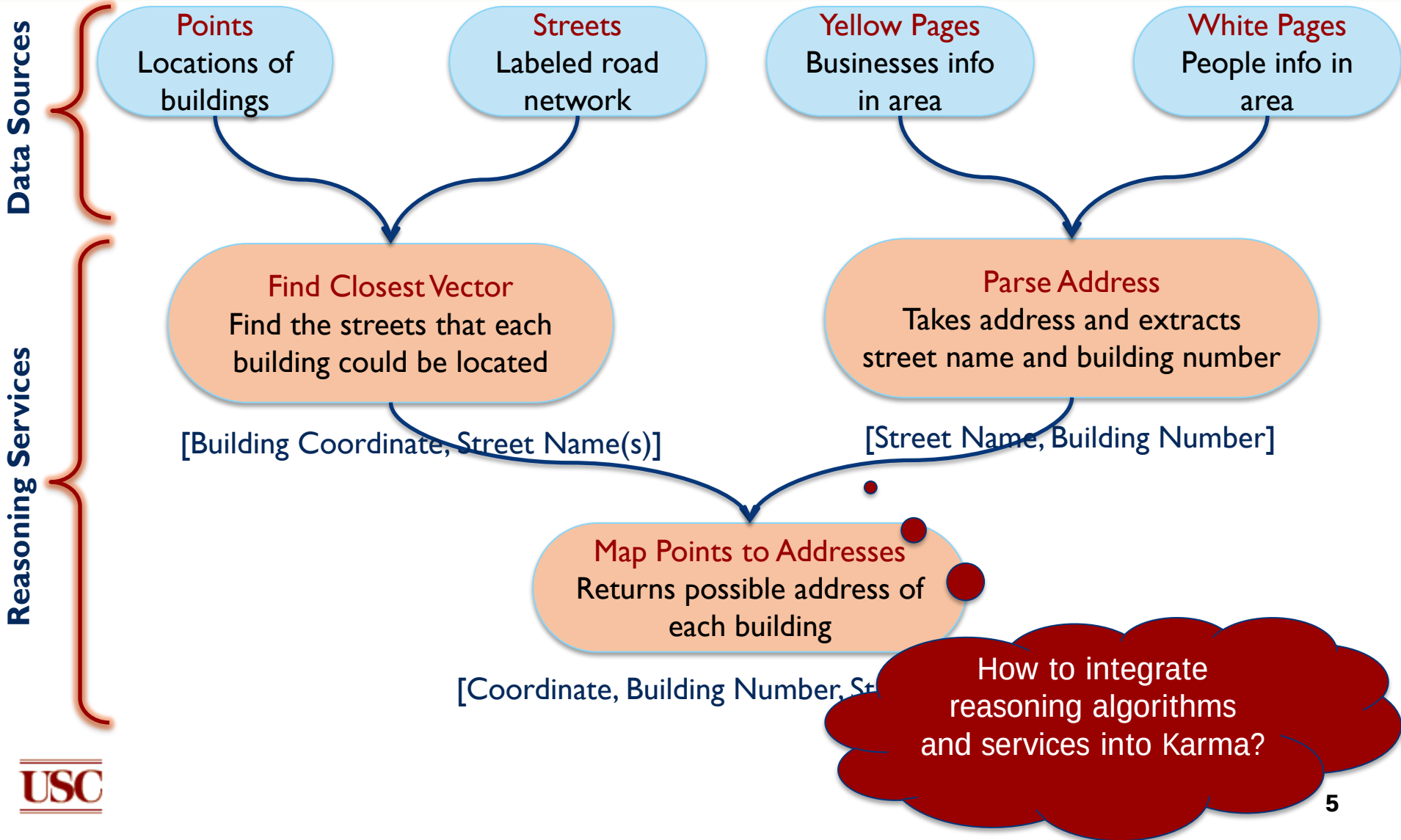
- Problem: Identify the address associated with each building that can be identified in the satellite imagery

Before


After

- **Solution:**
  - Step 1: Identify the street vector data, building locations and the phonebook data for the given area (data retrieval task)
  - Step 2: Reasoning over the data to generate a mapping between the addresses and building locations (geospatial reasoning task)

# Motivating Example (cont.)



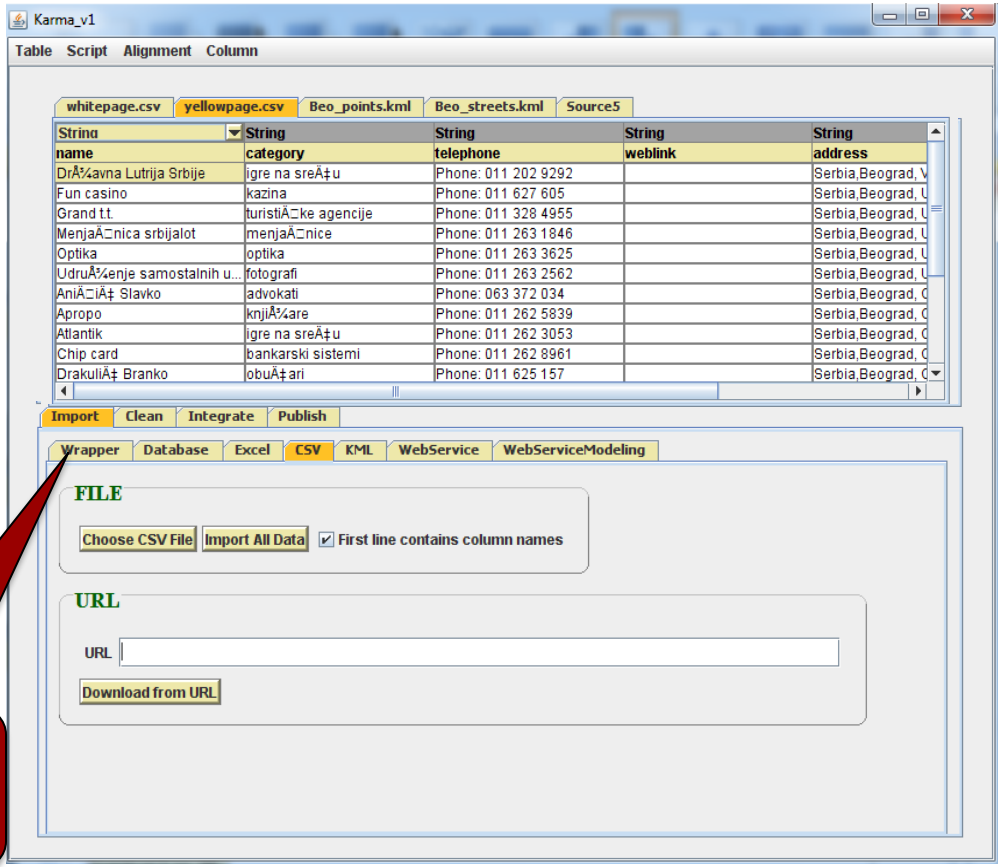
# Karma's Approach



- Build a semantic model of reasoning services based on provided ontology
  - Data types of inputs and outputs, plus relationships between them
- Interactively invoke services using semantic model of sources and services
  - Which services can be invoked using available data?
  - Which sources can satisfy service inputs?
- Integrate outputs of service invocation with the other data

# Importing Sources

Information Sciences Institute


Karma uses wrappers to extract web pages information

# Data Cleaning

Information Sciences Institute

whitepage.csv | Source2 | Cleaning Table

| String | Data Type | Dat |
|---|---|---|
| Column Name | User Defined Values | Fina |
| CARA LAZARA 15/2,BEOGRAD | CARA LAZARA 15/2.BEOGRAD | |
| CARA LAZARA 13/21,BEOGRAD | | |
| NIKOLE SPASIĆA 2,BEOGRAD | | |

User provides examples of address without  

Karma uses learned transformation rules to remove all instances of  

whitepage.csv | Source2 | Cleaning Table

| String | Data Type | Dat |
|---|---|---|
| Column Name | User Defined Values | Fina |
| CARA LAZARA 15/2,BEOGRAD | CARA LAZARA 15/2,BEOGRAD | |
| CARA LAZARA 13/21,BEOGRAD | CARA LAZARA 13/21,BEOGRAD | |
| NIKOLE SPASIĆA 2,BEOGRAD | NIKOLE SPASIĆA 2.BEOGRAD | |

USC

# Source Modeling

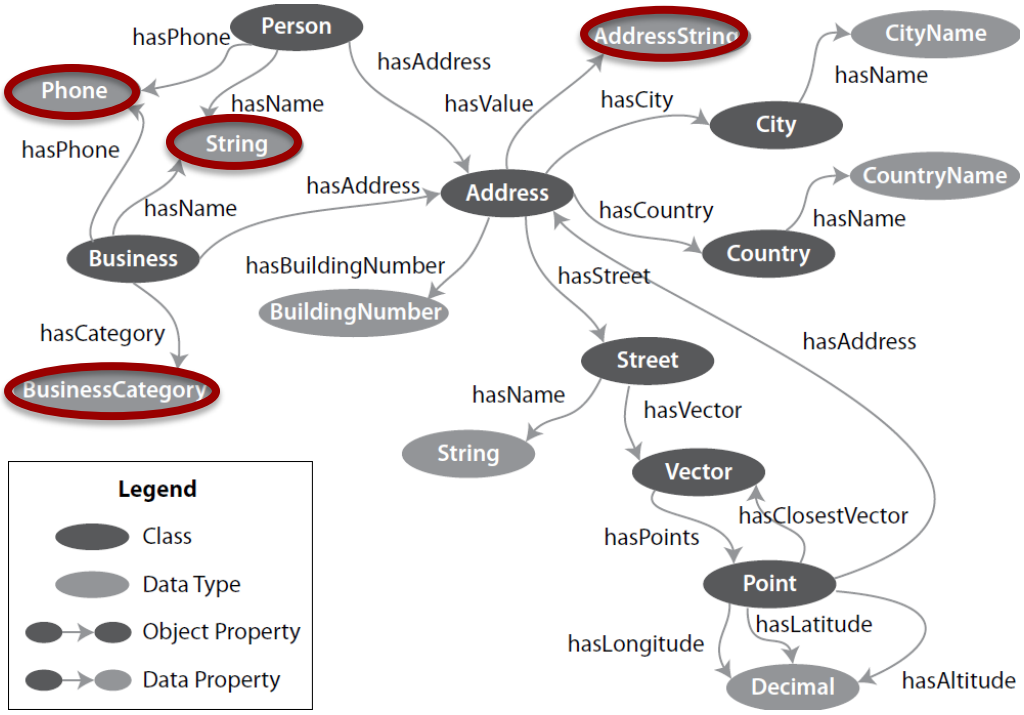


- Karma automatically builds models of data according to provided ontology
  - Models help user to process data and integrate them
- Identify the semantic types
  - Supervised machine learning technique (CRF Model)
    - A. Goel, C. A. Knoblock, K. Lerman, Using conditional random fields to exploit token structure and labels for accurate semantic annotation, 2011
- Identify relationships among the data columns
  - Find the minimal tree that connects the semantic types
    - C. A. Knoblock, P. Szekely, J. L. Ambite, S. Gupta, A. Goel, M. Muslea, K. Lerman, Interactively Mapping Data Sources into the Semantic Web, 2011

# Modeling YellowPages Source

Information Sciences Institute

1. Karma uses CRF technique to assign labels to each data column




Semantic Types

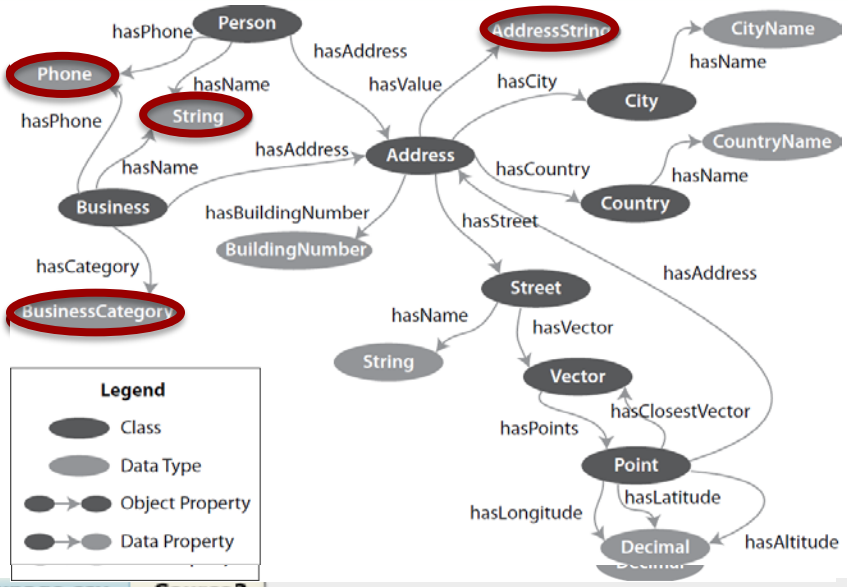
| String | BusinessCategory | AddressString | Phone |
|---|---|---|---|
| name | category | address | telephone |
| Državna Lutrija Srbije | igre na sreću | Serbia,Beograd, Vračar, Uskoč... | Phone: 011 202 9292 |
| Fun casino | kazina | Serbia,Beograd, Uskočka 4 | Phone: 011 627 605 |
| Grand t.t. | turističke agencije | Serbia,Beograd, Uskočka 7 | Phone: 011 328 4955 |
| Menjačnica srbijalot | menjačnice | Serbia,Beograd, Uskočka 4 | Phone: 011 263 1846 |

USC

# Modeling YellowPages Source

Information Sciences Institute

II. Karma selects the smallest tree that connects these semantic types and shows it at the top of the data worksheet.




Relationships between columns

yellowpage.csv | Source2

**Business**

| (hasName) String | (hasCategory) BusinessCategory | (hasAddress) Address / (hasValue) AddressString | (hasPhone) Phone |
|---|---|---|---|
| **String** | **BusinessCategory** | **AddressString** | **Phone** |
| **name** | **category** | **address** | **telephone** |
| Državna Lutrija Srbije | igre na sreću | Serbia,Beograd, Vračar, Uskoč... | Phone: 011 202 9292 |
| Fun casino | kazina | Serbia,Beograd, Uskočka 4 | Phone: 011 627 605 |
| Grand t.t. | turističke agencije | Serbia,Beograd, Uskočka 7 | Phone: 011 328 4955 |
| Menjačnica srbijalot | menjačnice | Serbia,Beograd, Uskočka 4 | Phone: 011 263 1846 |

USC

# Modeling Web Services

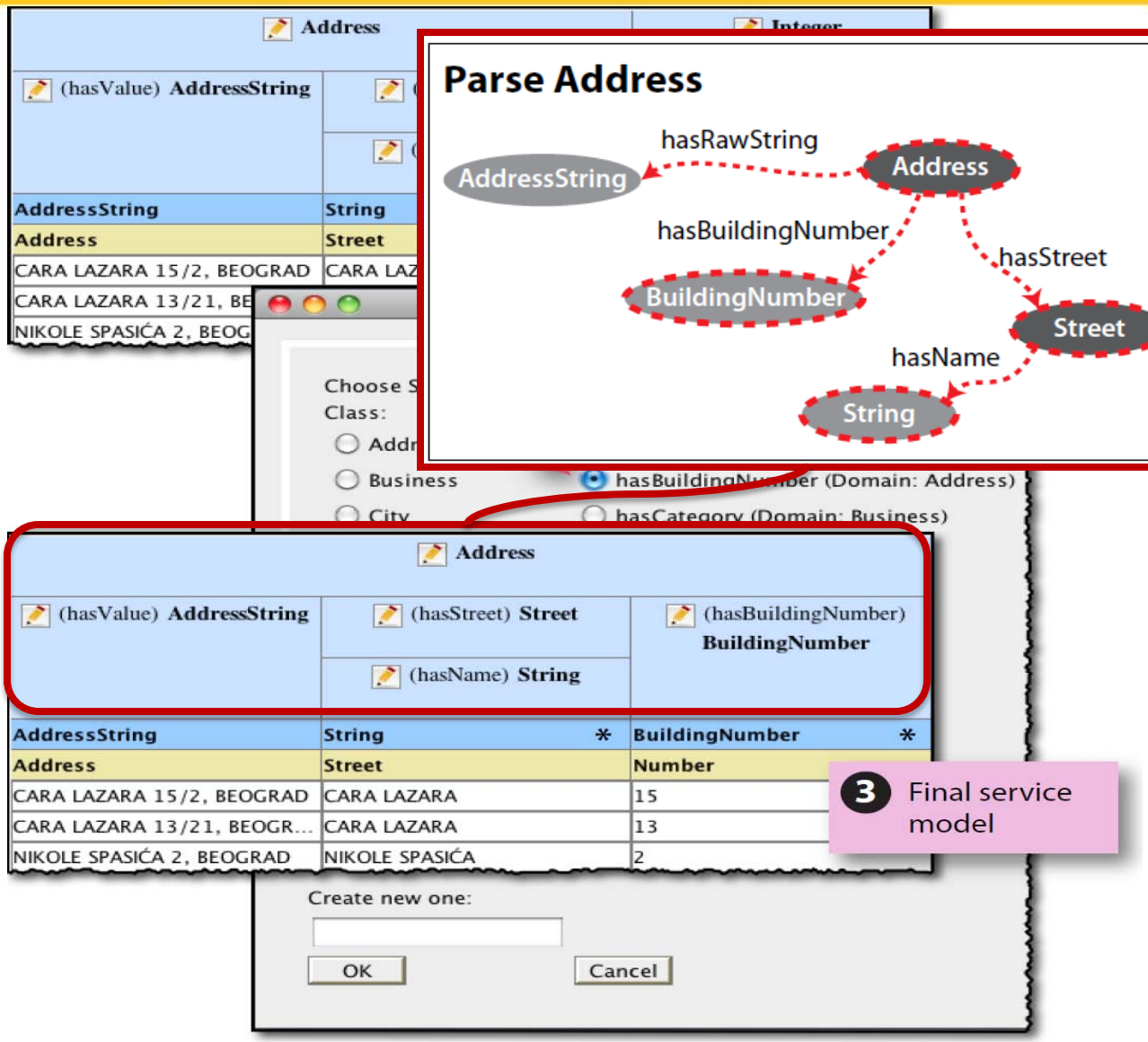


- Semantic models of web services facilitate service invocation, discovery, and composition

- Karma allow the user to interactively build a model
  - User provides examples of service input and output
  - Modeling services can be done like data sources

# Modeling of Parse Address Service in Karma



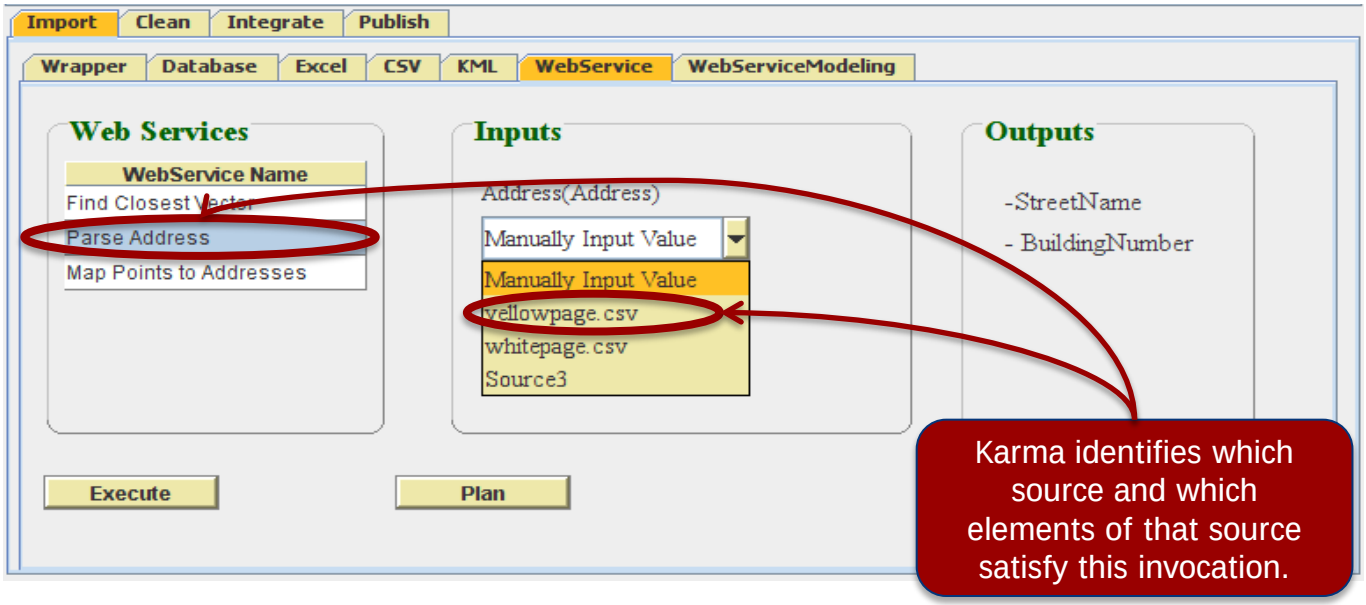


# Data Fusion



- Ability for users to interactively invoke services on other data sources

- Semantic models make it possible to:
  - Automatically determine which services apply to the available data
  - Perform automatic transformations on data to get it into the required format to apply a service
  - Automatically compose services and sources to generate required data



# Matching Sources and Services

Information Sciences Institute

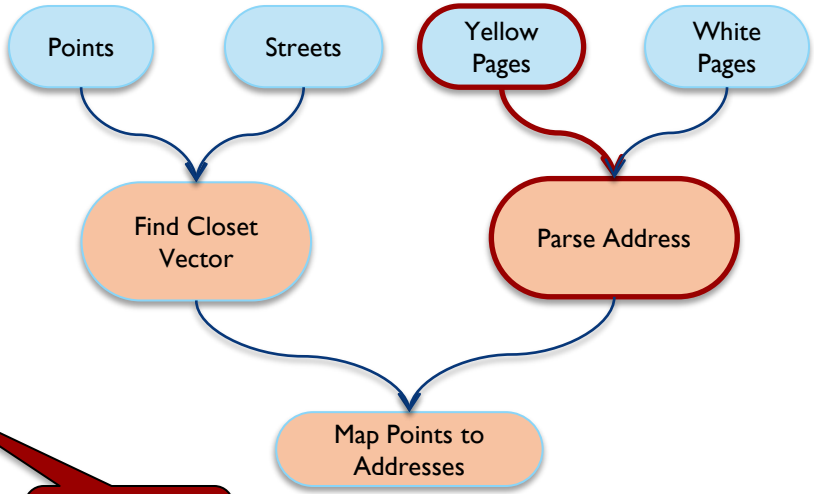
Yellow Pages

yellowpage.csv | Source2

Business

| (hasName) String | (hasCategory) BusinessCategory | (hasAddress) Address / (hasValue) AddressString | (hasPhone) Phone |
|---|---|---|---|
| String | BusinessCategory | AddressString | Phone |
| name | category | address | telephone |
| Državna Lutrija Srbije | igre na sreću | Serbia,Beograd, Vračar, Uskoč... | Phone: 011 202 9292 |
| Fun casino | kazina | Serbia,Beograd, Uskočka 4 | Phone: 011 627 605 |
| Grand t.t. | turističke agencije | Serbia,Beograd, Uskočka 7 | Phone: 011 328 4955 |
| Menjačnica srbijalot | menjačnice | Serbia,Beograd, Uskočka 4 | Phone: 011 263 1846 |

Input




Output

Address

| (hasValue) AddressString | (hasStreet) Street / (hasName) String | (hasBuildingNumber) BuildingNumber |
|---|---|---|
| AddressString | String | BuildingNumber |
| Address | Street | Number |
| CARA LAZARA 15/2, BEOGRAD | CARA LAZARA | 15 |
| CARA LAZARA 13/21, BEOGR... | CARA LAZARA | 13 |
| NIKOLE SPASIĆA 2, BEOGRAD | NIKOLE SPASIĆA | 2 |

Parse Address

# Invocation Results

Information Sciences Institute

- Results of invocation are returned as another source that can be refined, integrated with other sources, visualized or published

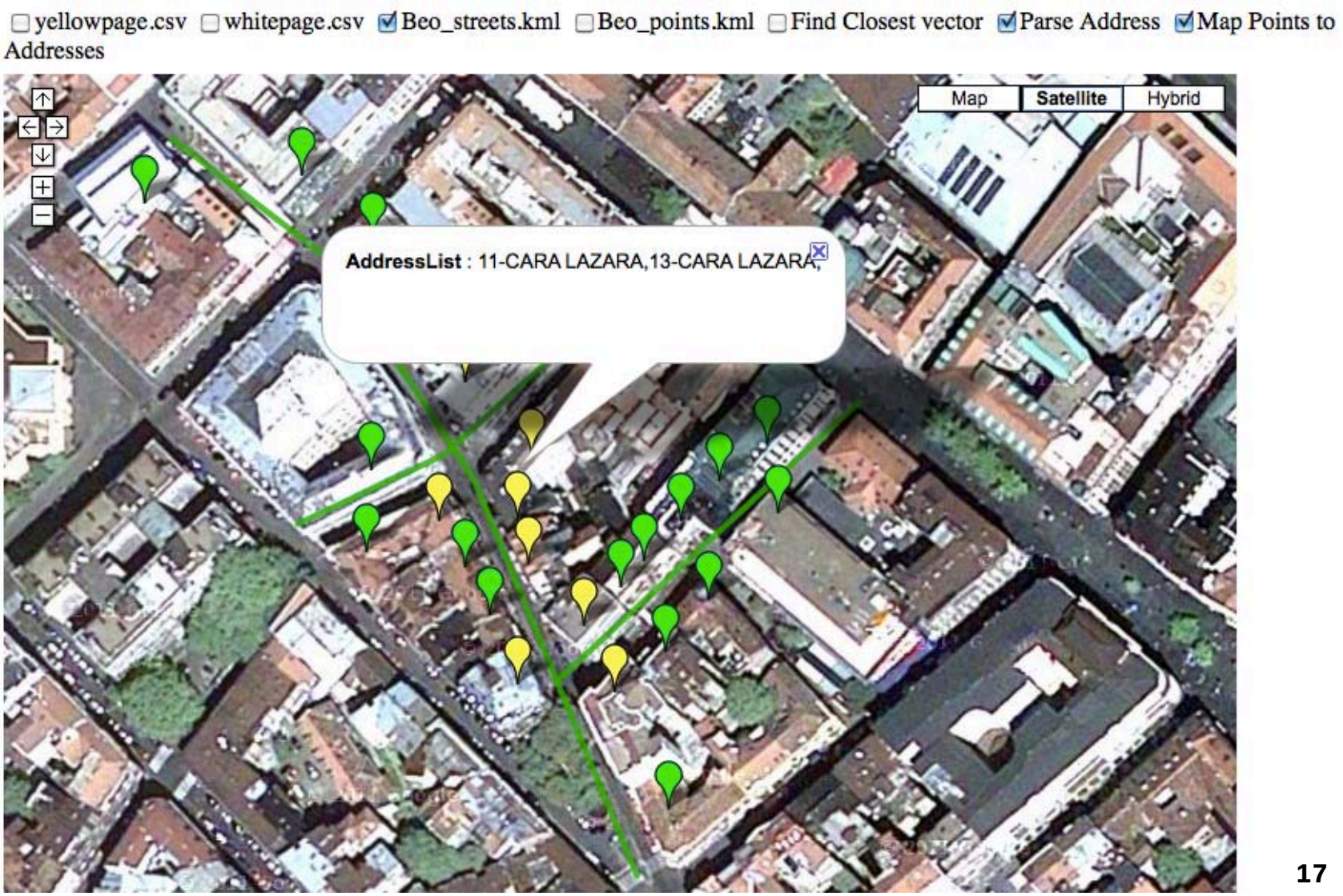
GeoFusionResult | Source2

| Point | | |
|---|---|---|
| (hasLongitude) Decimal | (hasLatitude) Decimal | (hasAddress) Address<br>(hasValue) AddressString |
| Decimal | Decimal | AddressString |
| Longitude | Latitude | AddressList |
| 20.45329213142395 | 44.81892469068898 | 4-CARA LAZARA |
| 20.45383930206299 | 44.8189931839843 | 1-CARA LAZARA |
| 20.45408606529236 | 44.81883336616863 | 3-CARA LAZARA |
| 20.454246997833252 | 44.81861266464735 | 5-CARA LAZARA |
| 20.454407930374146 | 44.81848328750073 | 7-CARA LAZARA, 2A-NIKOLE... |
| 20.45408070087433 | 44.81826638986837 | 4-NIKOLE SPASICA |

USC

# Visualize Final output





# Related Work



- Exploit ontologies to attach semantics to geospatial services
  - [L. Di, et. al., 2006], [P. Yue, et. al. , 2010]
  - User has to manually annotate the services according to an ontology like OWL-S
  - They model input and output types but not relationship among them

- Linked Open Services (LOS)
  - [B. Norton, R. Krummenacher, 2010]
  - Services that consume linked data as input and also return linked data as output
  - Use SPARQL to describe service inputs and outputs
  - Describing services might be easy for Linked Data community, but not for average Internet users

- Google Fusion Tables
  - [H. Gonzalez, A. Halevy, et al. 2010]
  - Import data from various source types and invoke web services
  - Allows advanced visualization
  - Integrating data from different sources is possible but without exploiting semantics

# Discussion



- Karma allows users to quickly and easily dynamically fuse a wide variety of geospatial data sources
- Modeling geospatial services is a big step in geospatial data fusion
- Based on provided ontology, Karma semi-automatically builds a semantic model of reasoning services including both input/output datatypes and their relationships
- Semantic descriptions enable user to easily find the desired service and invoke it using available data sources

# Future Work



- Applying the service modeling techniques to available REST web services
  - Create the service model just based on service invocation samples
- Answer queries like "Can I have the street names of the cities whose distance to Los Angeles is less than 50 miles?"
  - Automatically compose available web services using loaded data sources
- Publishing semantic description of web services in formats such as LOS

USC Viterbi School of Engineering

# More Information

ISI Information Sciences Institute

- More information available on Karma:
  - http://www.isi.edu/~knoblock
- Contact:
  - pszekely@isi.edu
  - knoblock@isi.edu
  - shubhamg@isi.edu
  - mohsen@isi.edu
  - bowu@isi.edu
- Software:
  - Software will be available as open source under the Apache license as soon as we complete the next version