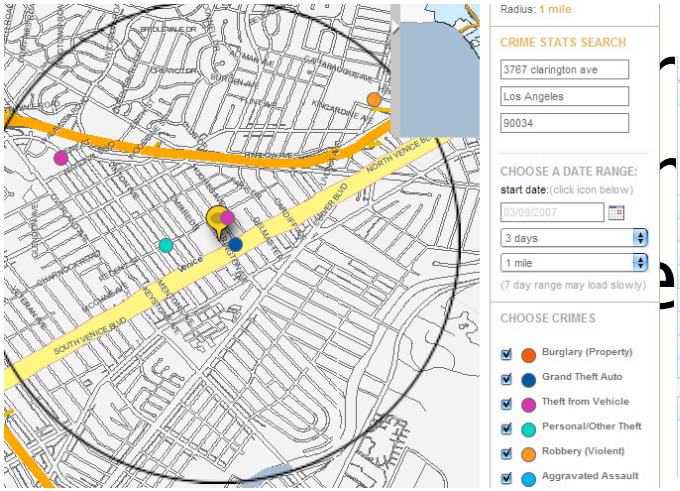
Interactively Building Geospatial Mashups

Craig A. Knoblock
University of Southern California

Work in collaboration with Shubham Gupta, Pedro Szekely, and Rattapoom Tuchinda

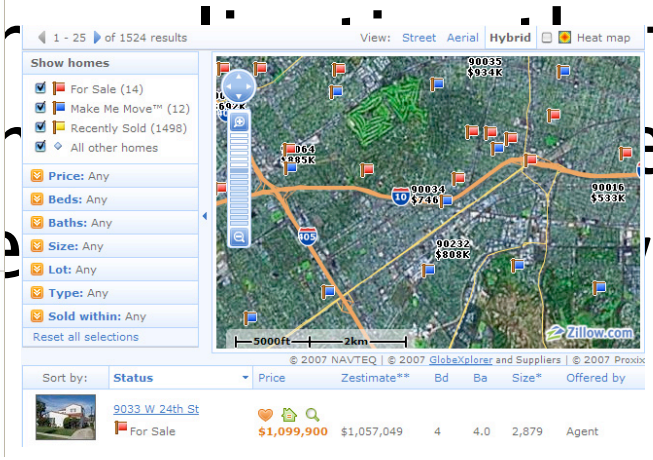
MASHUPS



a) LA crime map

-Crime Report from different counties

-Map



b) zillow.com

-Real Estate Listing

-Property Tax

c) Ski bonk

-Weather

-Snow Report

-Snow Resorts

Combined Data gives new insight / provides new services

PROBLEM

- Most Mashups require significant expertise to create
- Demand for creating integrated applications is huge
- Every user has their own unique requirements for an integrated application
- Available sources and needs to integrated data continues to grow

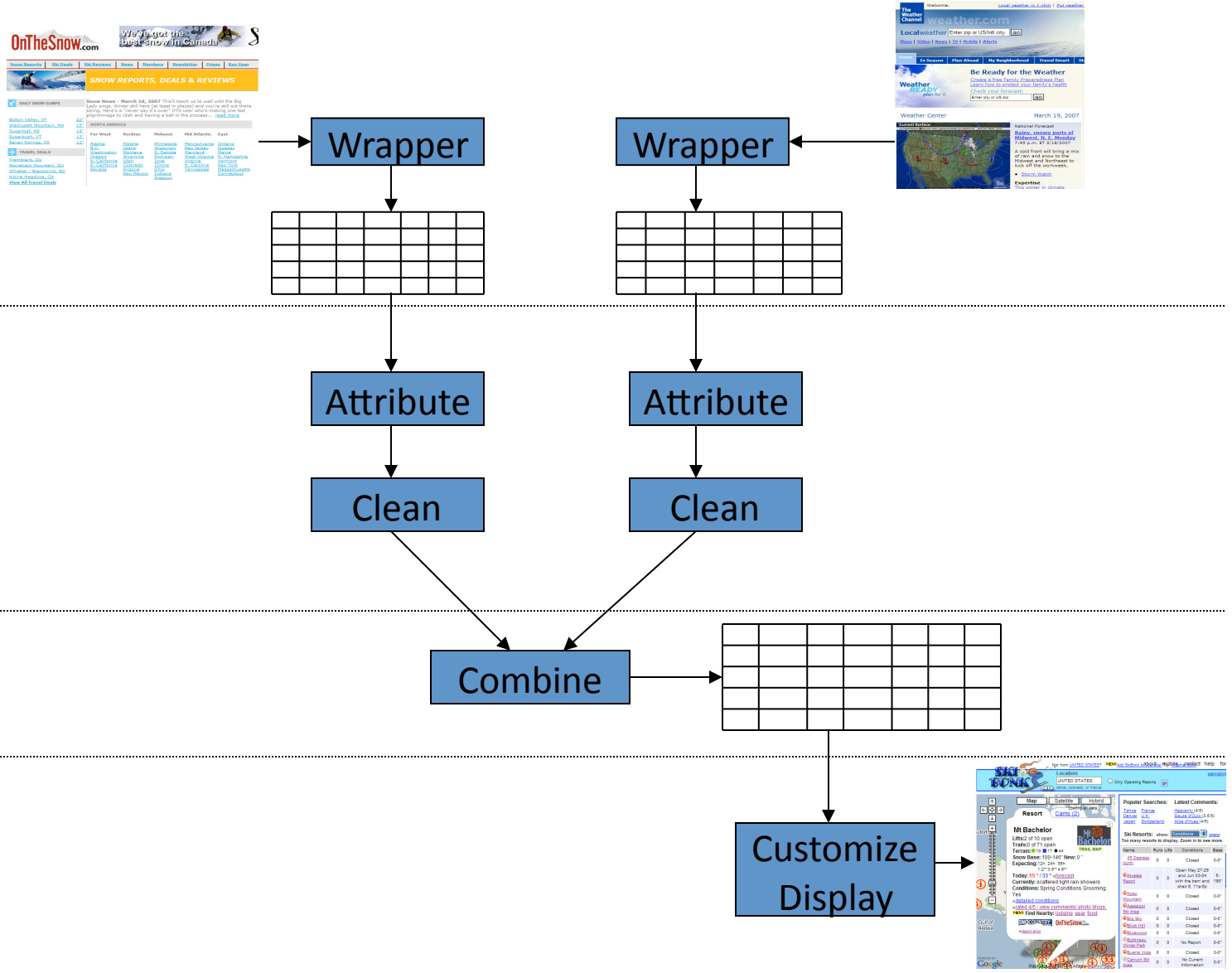
MASHUP BUILDING ISSUES

Data Retrieval

Calibration
-source modeling
-cleaning

Integration

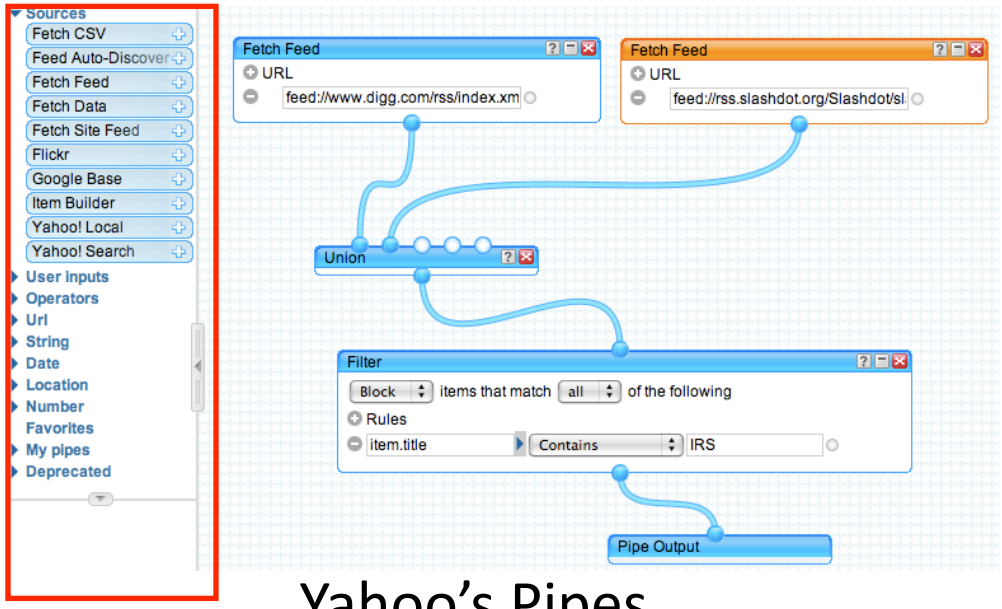
Display



EXISTING APPROACHES

Goal: Create Mashups without Programming

- Doesn't translate to not having to understand programming



Yahoo's Pipes

Widget Paradigm

- Widgets (i.e., 43 for Pipes, 300+ for MS) represents an operation on the data
- Locating and learning to customize widget can be time consuming
- Most tools focus on particular issues and ignore others

Can we come up with a framework that addresses all of the issues while still making the Mashup building process easy?

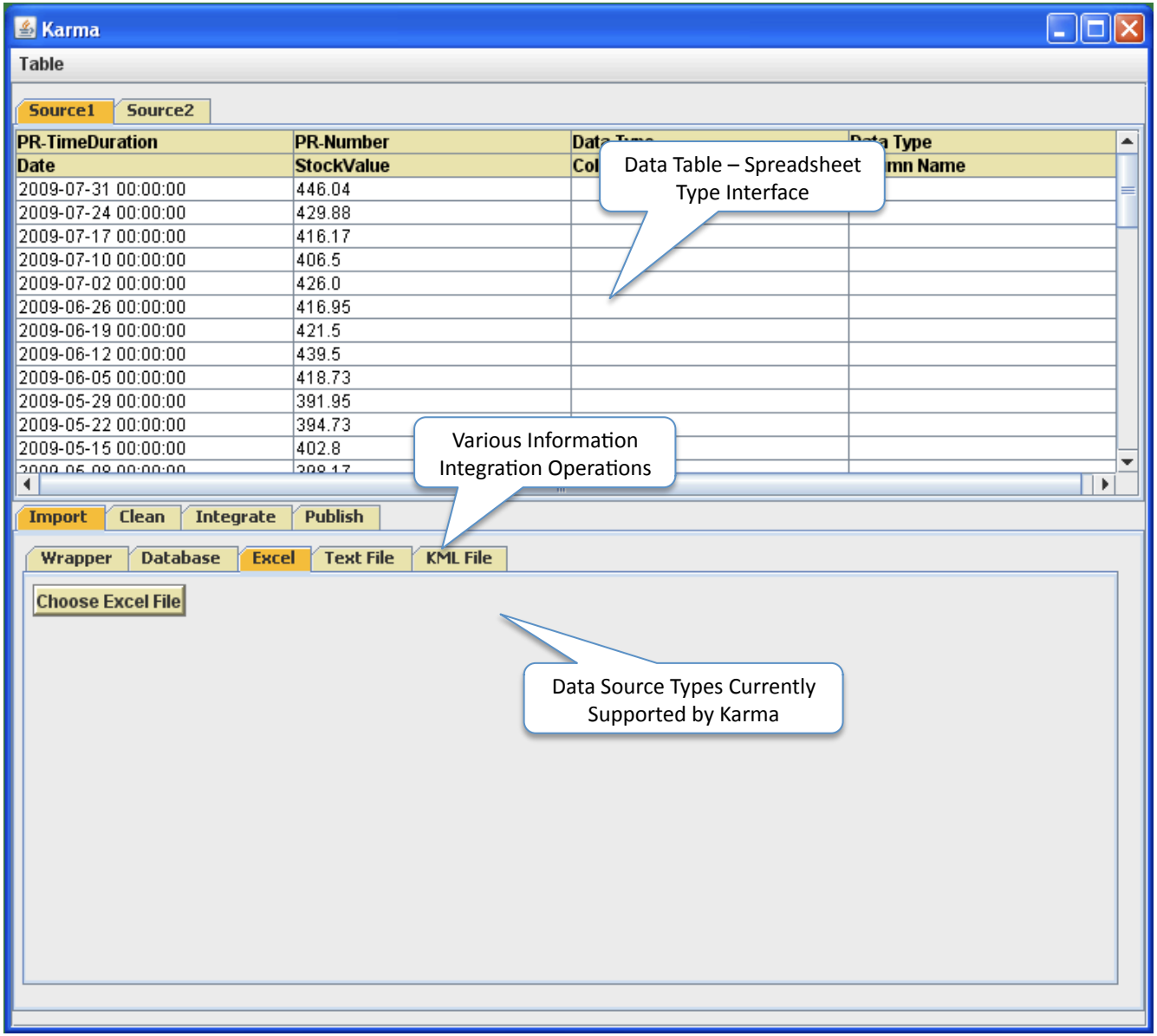
KEY CONTRIBUTIONS

- A programming by demonstration approach that uses a single table for building a Mashup
- An integrated approach that links data extraction, source modeling, data cleaning, and data integration together
- A query formulation technique that allows users to specify examples to build complicated queries

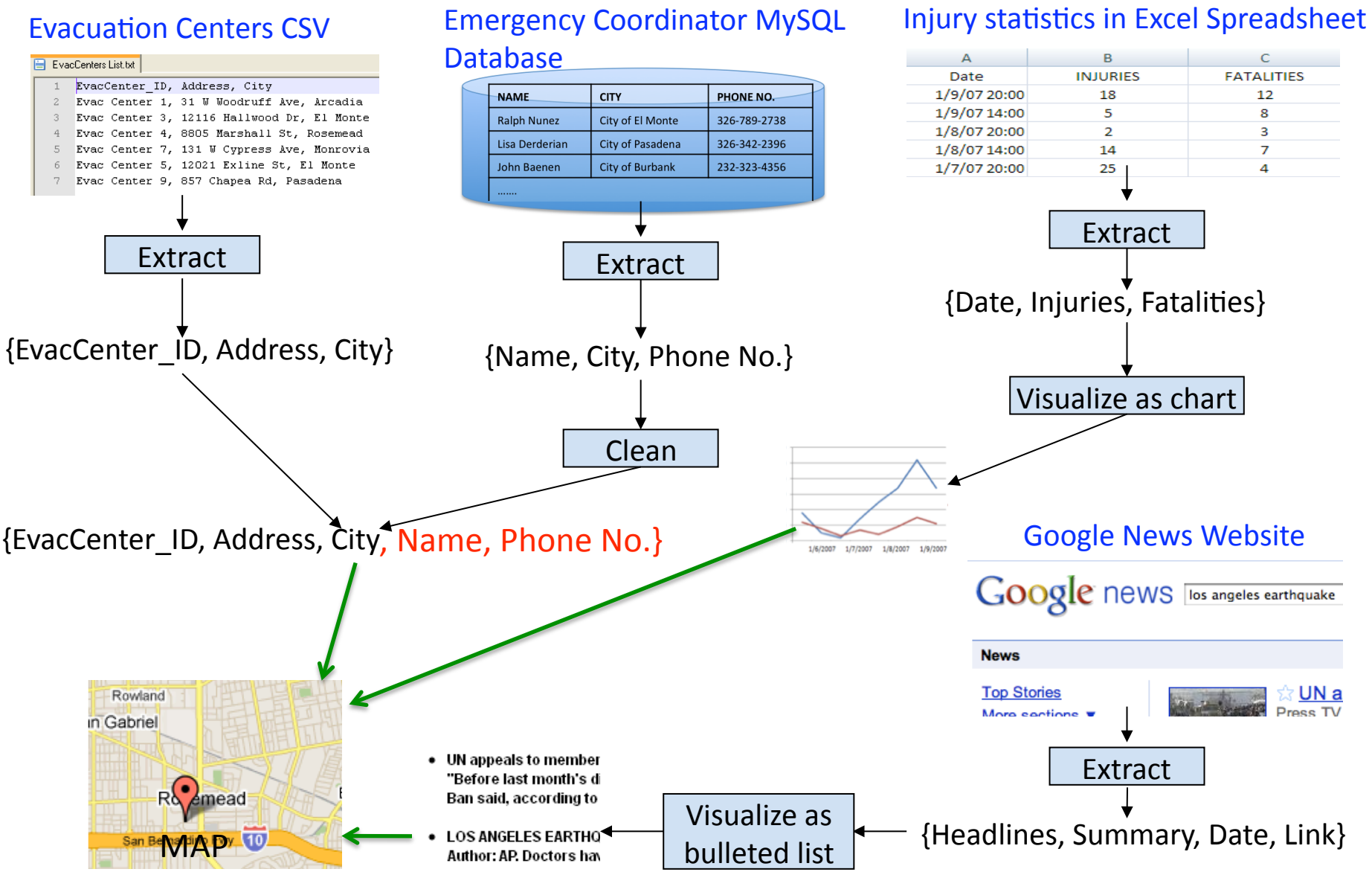
KEY IDEAS

- Focus on data, not operations
  - Users are more familiar with data
- Leverage existing data
  - Help source modeling, cleaning, and data integration
- Consolidate as opposed to Divide-And-Conquer
  - Solving a problem in one issue can help solve another issue
  - Interacting within a single spreadsheet platform

KARMA USER INTERFACE

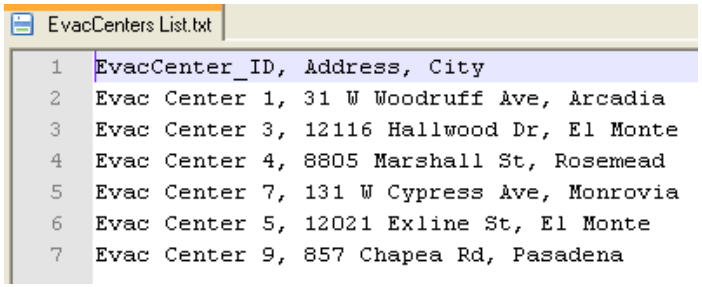


INTEGRATION SCENARIO

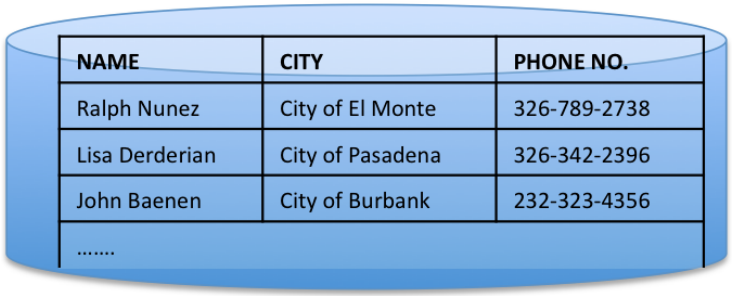


RETRIEVING DATA FROM DIVERSE SOURCES

- Karma facilitates retrieval of data from structured data-sources, such as Excel spreadsheets, MySQL databases and CSV files
- Karma also facilitates the extraction of data from semi-structured data sources such as web pages



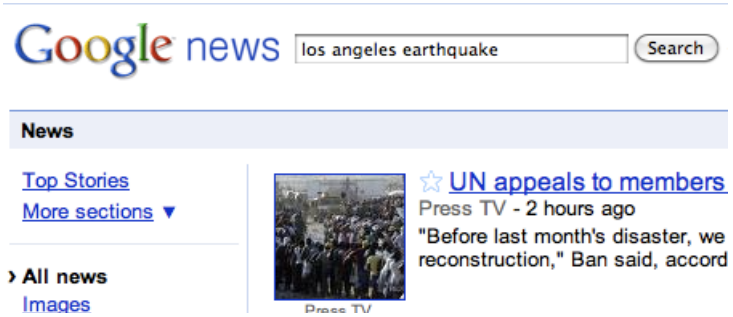
CSV Text File



MySQL Database

A	B	C
Date	INJURIES	FATALITIES
1/9/07 20:00	18	12
1/9/07 14:00	5	8
1/8/07 20:00	2	3
1/8/07 14:00	14	7
1/7/07 20:00	25	4
1/7/07 14:00	34	9
1/6/07 20:00	52	15
1/6/07 14:00	34	11

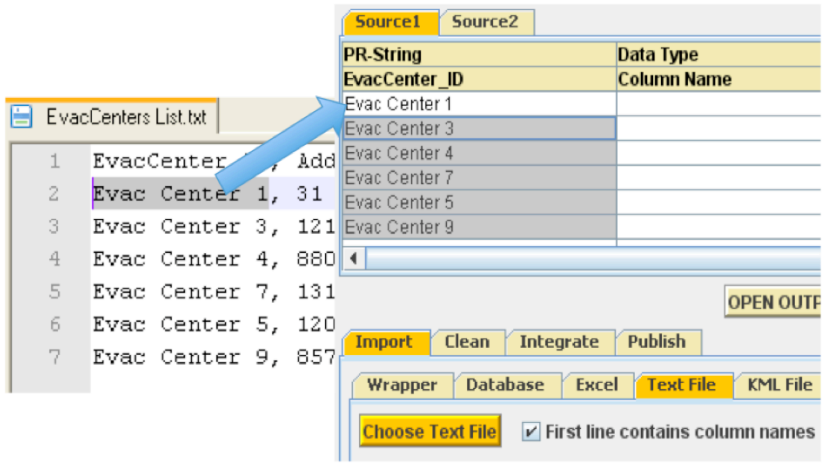
Excel Spreadsheet



HTML Web Page

EXTRACTION BY EXAMPLE

- The retrieval of data from structured data-sources, such as Excel sheets and CSV files is done through a drag and drop mechanism
- The user is only required to select a sample data-element and drop it into Karma's data table



EXTRACTION FROM THE WEB

1. **Japon Bistro**
   927 E Colorado Blvd , Pasadena , CA , 91106
   Upscale yet affordable Japanese eatery offers the city's largest sake selection.

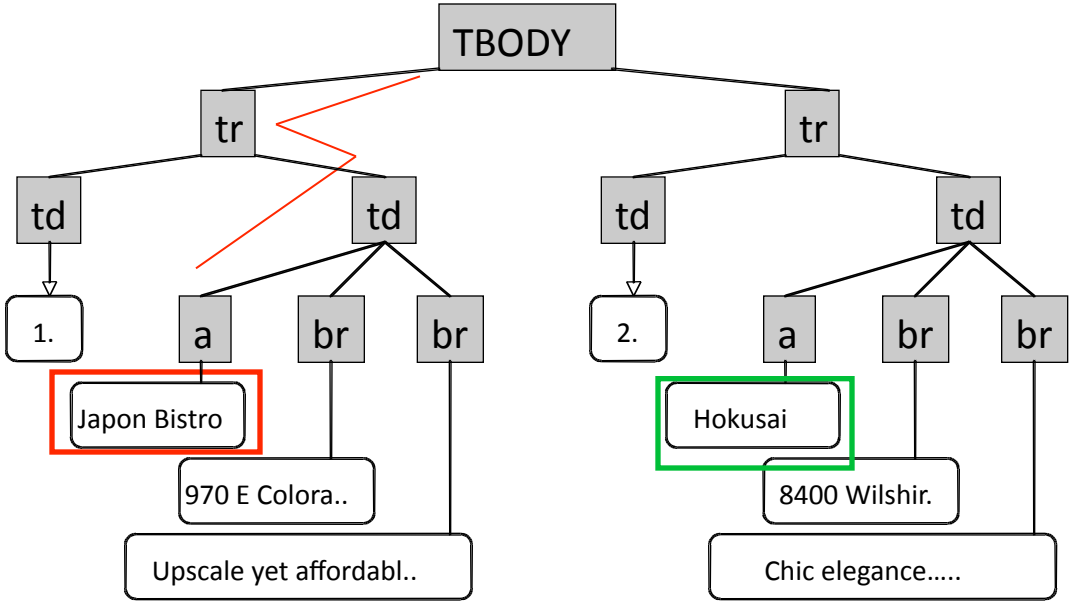
   Sushi Dokoro Ki Ra La
   9777 S Santa Monica Blvd , Beverly Hills , CA , 90211
   Intimate and charming Japanese restaurant offers wide range of hand-selected sushi and sashimi.

2. **Hokusai**
   8400 Wilshire Blvd , Beverly Hills , CA , 90211
   Chic elegance and modern Zen style surround Japanese French this paean to haute cuisine and stylized sushi.

3. **Sushi Sasabune**
   12400 Wilshire Blvd Ste 150 , Los Angeles , CA , 90025
   Sushi is the singular star at this Zen Westside palace that bows only to the royalty of chef and fish.

4. **Sushi Roku**
   8445 W 3rd St , Los Angeles , CA , 90048
   High fashion, rock and roll and Hollywood buzz converge over innovative sushi.

select one
Japon Bistro
Sushi Dokor..
Hokusai
Sushi Sasab..
Sushi Roku
Hide Sushi
Fat Fish
Sushi Katsu-ya
Gindi Thai /..
Katana
Echigo



Tbody/tr[1]/td[2]/a

Tbody/tr*/td*/a

EXTRACTION FROM THE WEB

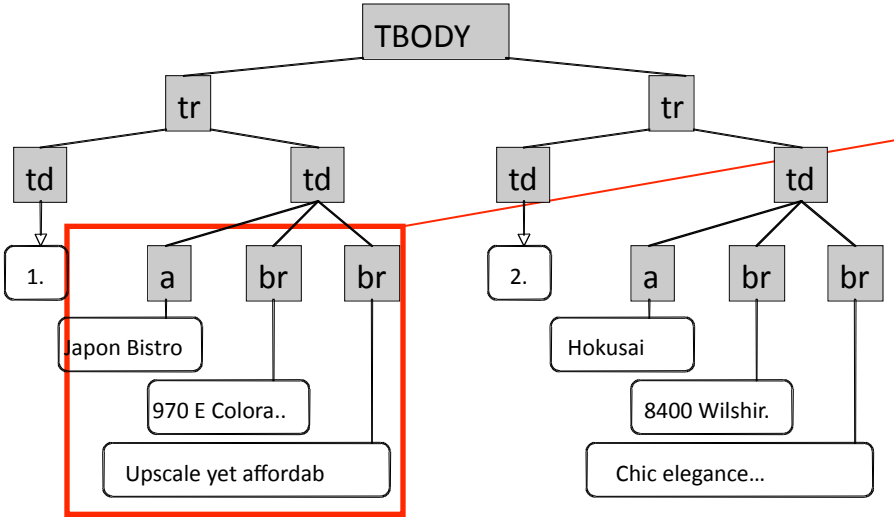
1. **Japon Bistro**
   927 E Colorado Blvd , Pasadena , CA , 91106
   Upscale yet affordable Japanese eatery offers the city's largest sake selection.

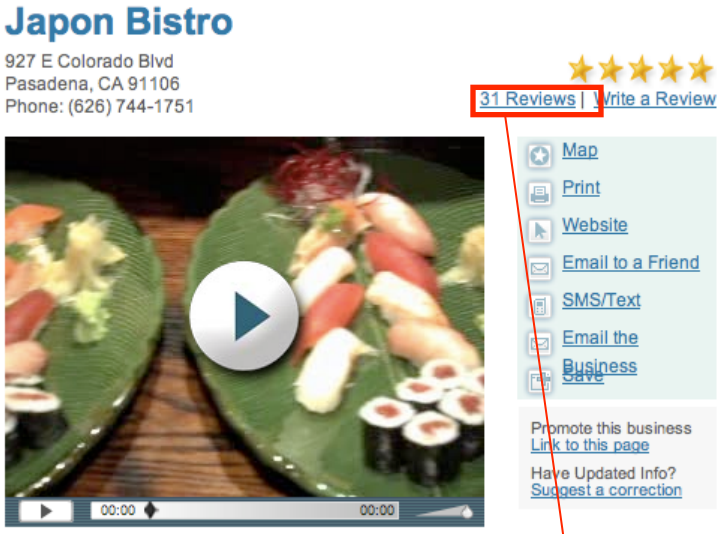
   Sushi Dokoro Ki Ra La
   9777 S Santa Monica Blvd , Beverly Hills , CA , 90211
   Intimate and charming Japanese restaurant offers wide range of hand-selected sushi and sashimi.

2. **Hokusai**
   8400 Wilshire Blvd , Beverly Hills , CA , 90211
   Chic elegance and modern Zen style surround Japanese French this paean to haute cuisine and stylized sushi.

3. **Sushi Sasabune**
   12400 Wilshire Blvd Ste 150 , Los Angeles , CA , 90025
   Sushi is the singular star at this Zen Westside palace that bows only to the royalty of chef and fish.

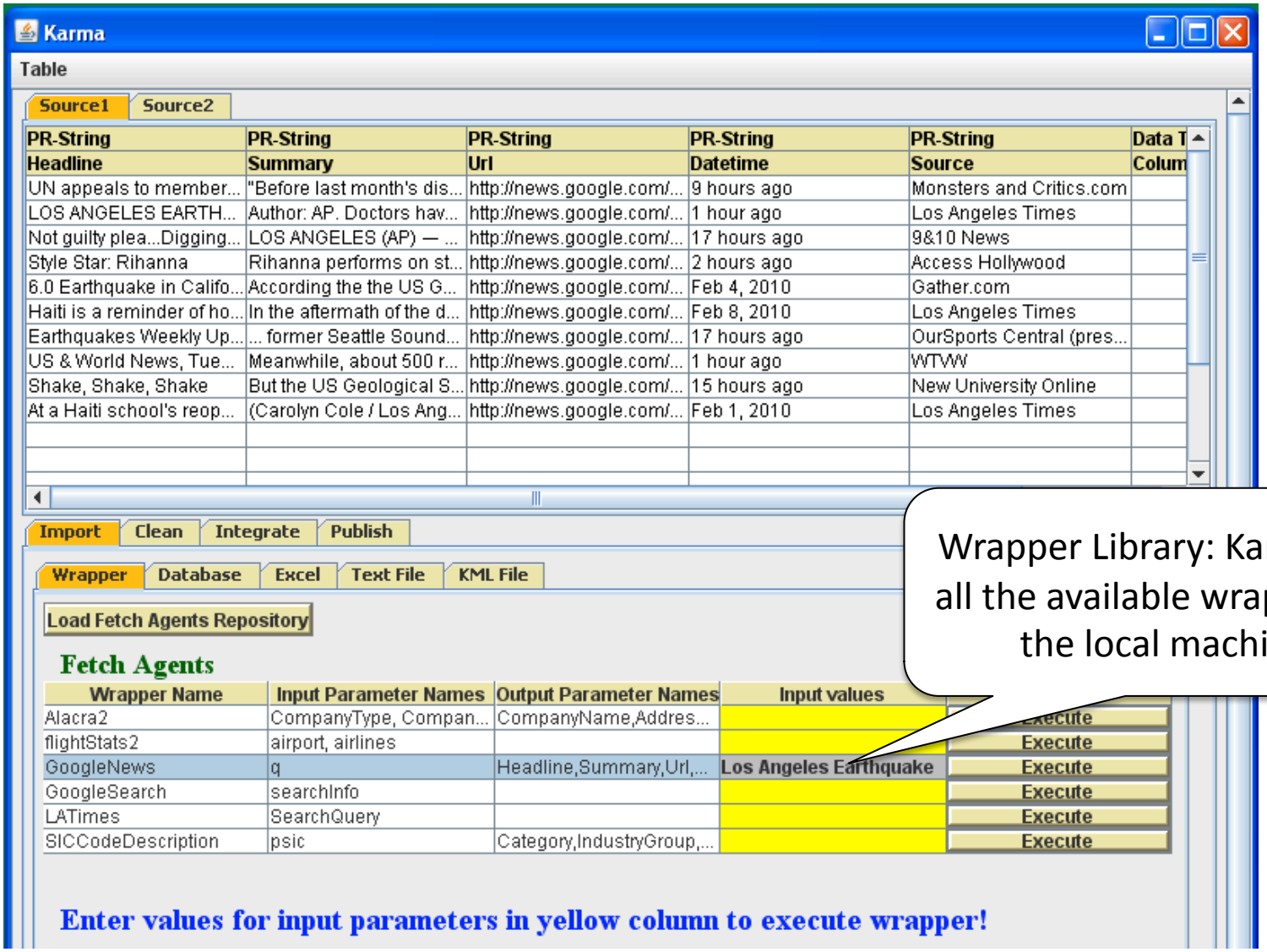
4. **Sushi Roku**
   8445 W 3rd St , Los Angeles , CA , 90048
   High fashion, rock and roll and Hollywood buzz converge over innovative sushi.





select one	address	select one	select one
Japon Bistro	927 E Color..	Upscale yet...	31 Reviews
Sushi Dokor..	9777 S Sant..	Intimate an...	
Hokusai	8400 Wilshir...	Chic eleganc...	
Sushi Sasab..	12400 Wilshi...	Authentic Ja...	
Sushi Roku	8445 W 3rd...	High fashion...	
Hide Sushi	2040 Sawtel...	No fuss, jus...	
Fat Fish	616 N Rober...	Inventive ro...	
Sushi Katsu-ya	11680 Vent...	The MOCA o...	
Gindi Thai /..	4017 W Riv...	Burbank res...	
Katana	8439 W Sun...	Rustic Japa...	
Echigo	11217 Sant...	Stellar sushi...	

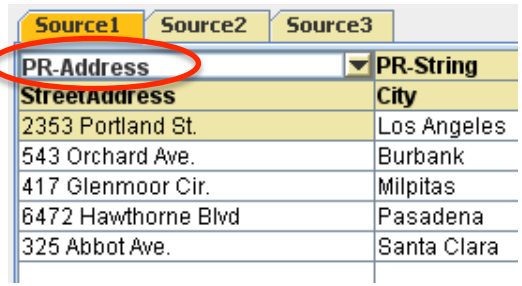
EXPLOITING WRAPPER LIBRARIES



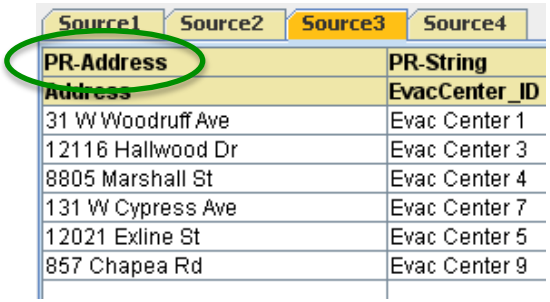
SOURCE MODELING

- Karma automatically generates the semantic types of each attribute to learn the underlying model of the data source
- Supervised machine learning techniques are used to generate a set of patterns for each semantic type from training data

Initial Type



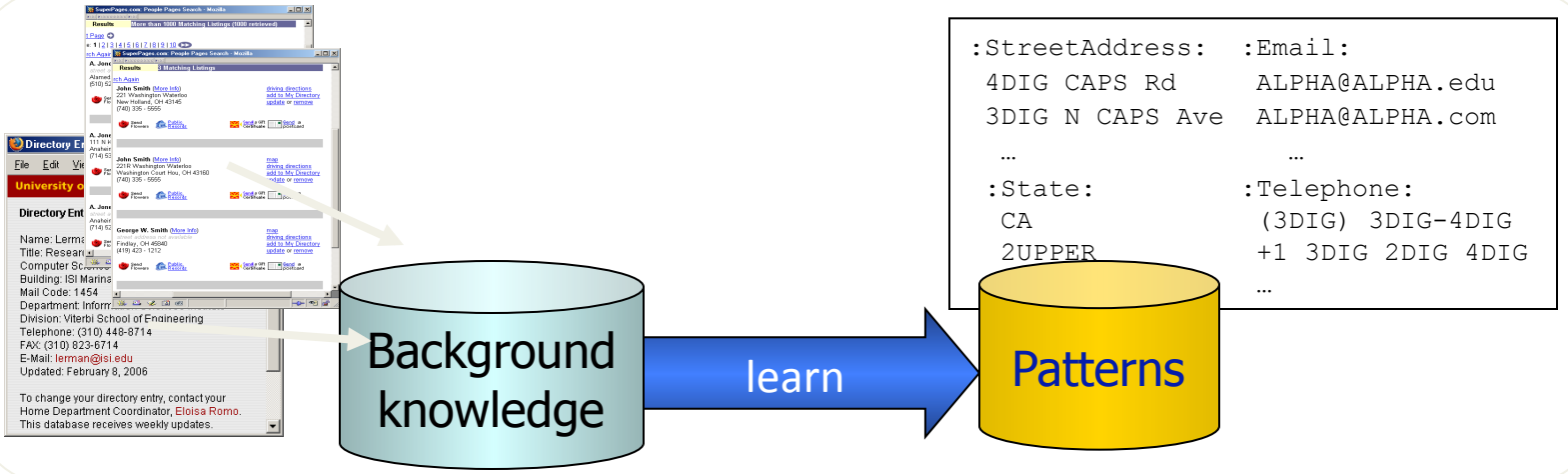
Manually label the data with the correct semantic type to train Karma



When the new data is imported of same type, Karma automatically labels it correctly

LEARNING SEMANTIC TYPES

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



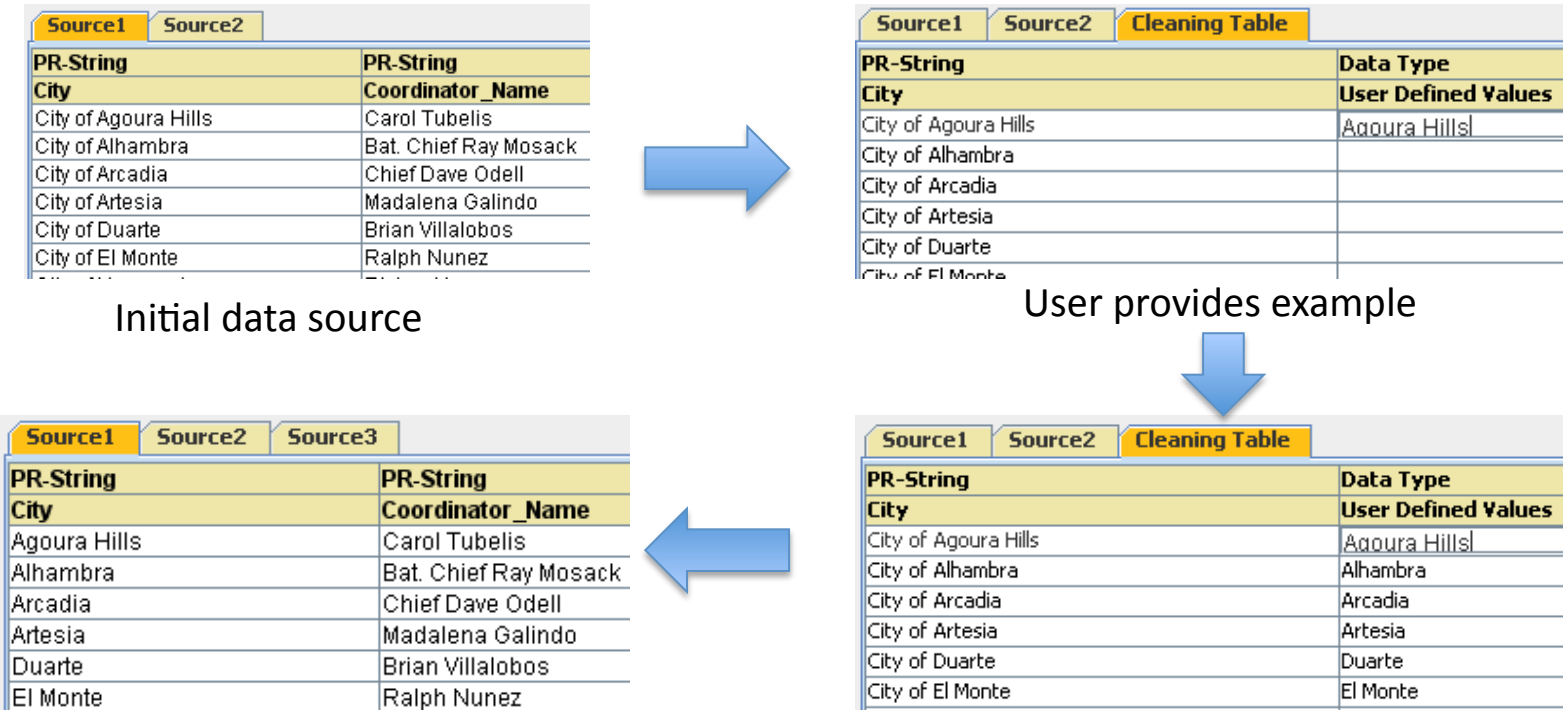
Person	Address	Work
E Lewis	3518 Hilltop Rd	(419) 531 - 0504
Andrew Lewis	3543 Larchmont Pkwy	(518) 474 -
C. S. Lewis	555 Willow Run Dr	(612) 578 -
Carmen Jones	355 Morgan Ave N	(612) 522 -
John Jones	3574 Brookside Rd	(555) 531 - 9566
Location	**State_prov**	**Postal_code**
Toledo	OH	64325-3000
Toledo	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

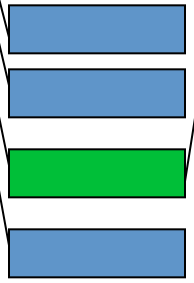
DATA CLEANING

- Karma performs the data cleaning by learning and applying the transformation rules that are learned from examples



DATA CLEANING: PREDEFINED TRANSFORMATIONS

	description	number of r...	suggest	user defined	final
.	Upscale yet...	31 Reviews		31	
..	Intimate an...	3 Reviews			
...	Chic eleganc...	30 Reviews			
...	Authentic Ja...	66 Reviews			
..	High fashion...	62 Reviews			
...	No fuss, jus...	25 Reviews			
...	Inventive ro...	38 Reviews			
..	The MOCA o...	49 Reviews			
..	Burbank res...	29 Reviews			
...	Rustic Japa...	96 Reviews			
..	Stellar sushi...	49 Reviews			



31 Reviews → 31

Subset Rule:

$(s_1 s_2 .. s_k) \rightarrow (d_1 d_2 ... d_t) \wedge$
$(k <= t) \wedge$
$s_i \in \{d_1, d_2, ..., d_t\} \wedge$
$d_i \neq d_j$

.
.
.

Predefined Rules

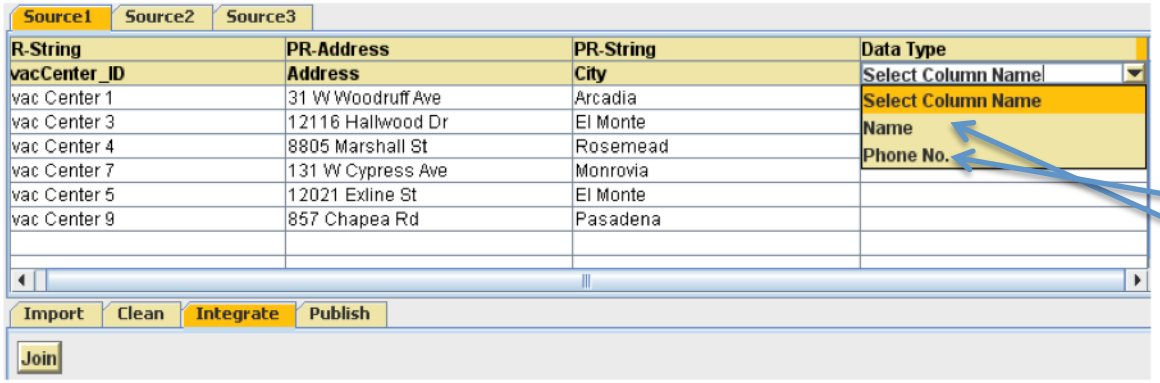
DATA INTEGRATION

- Karma discovers the related sources by detecting and ranking associations based on the common attribute names and matching semantic types
- Karma suggests potential joins between the current data sources in the form of column completions

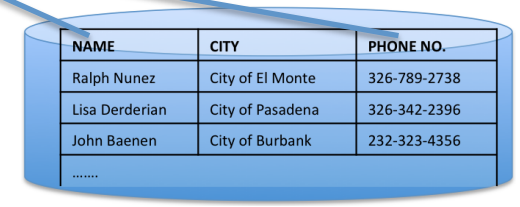
Source1 | Source2

PR-String	PR-Address	PR-String	Data Type
EvacCenter_ID	Address	City	Column Name
Evac Center 1	31 W Woodruff Ave	Arcadia	
Evac Center 3	12116 Hallwood Dr	El Monte	
Evac Center 4	8805 Marshall St	Rosemead	
Evac Center 7	131 W Cypress Ave	Monrovia	
Evac Center 5	12021 Exline St	El Monte	
Evac Center 9	857 Chapea Rd	Pasadena	

USER SELECTS FROM COLUMN COMPLETIONS



Karma suggests the possible column completions in a drop down list



NAME	CITY	PHONE NO.
Ralph Nunez	City of El Monte	326-789-2738
Lisa Derderian	City of Pasadena	326-342-2396
John Baenen	City of Burbank	232-323-4356
.....		

MySQL Database loaded as a another source in Karma

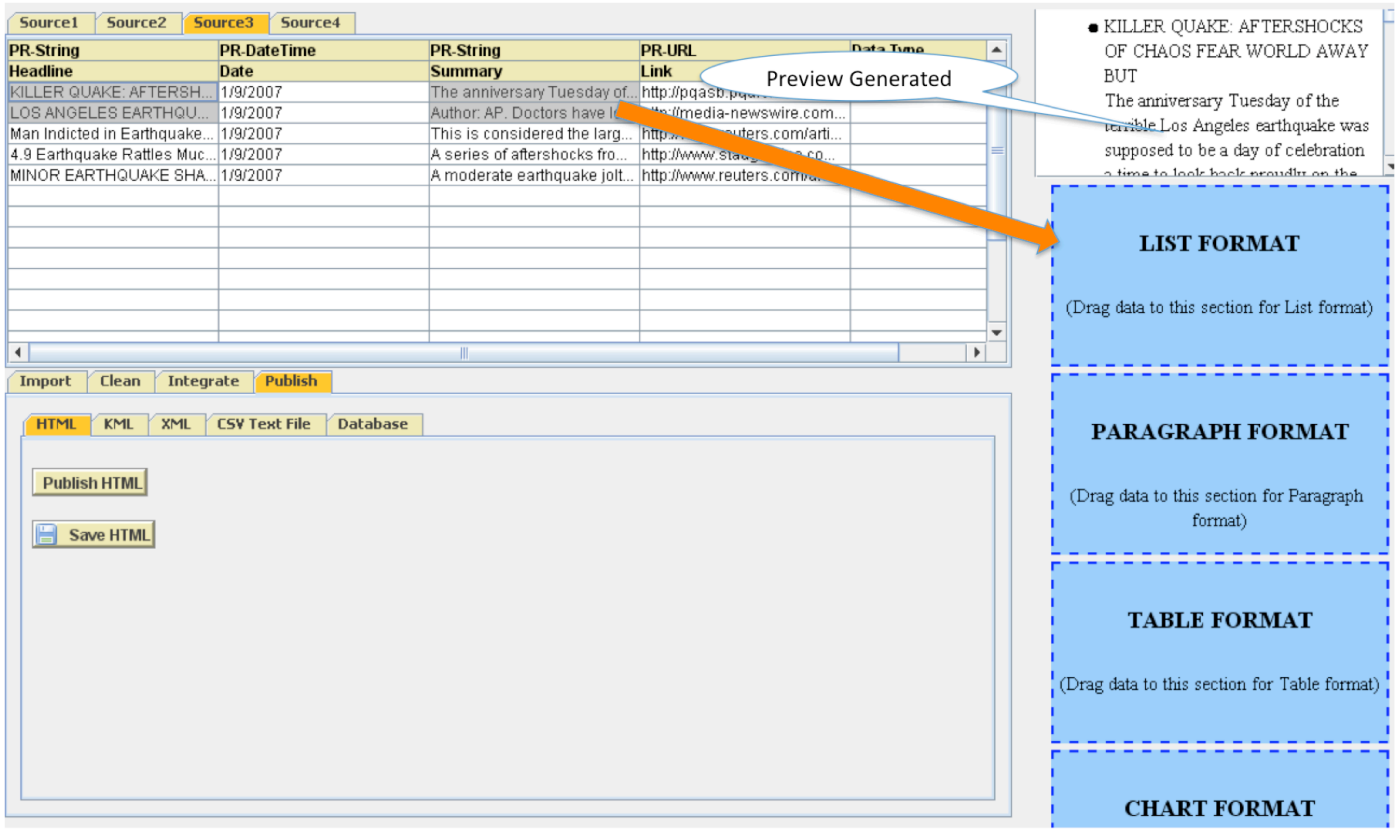
Source1 | Source2 | Source3

PR-String	PR-Address	PR-String	PR-String
EvacCenter_ID	Address	City	Name
Evac Center 1	31 W Woodruff Ave	Arcadia	Chief Dave Odell
Evac Center 3	12116 Hallwood Dr	El Monte	Ralph Nunez
Evac Center 4	8805 Marshall St	Rosemead	Donna Wagner
Evac Center 7	131 W Cypress Ave	Monrovia	Dave Dennis
Evac Center 5	12021 Exline St	El Monte	Ralph Nunez
Evac Center 9	857 Chapea Rd	Pasadena	Lisa Derderian

Karma executes the join query once the user selects an option

DATA VISUALIZATION

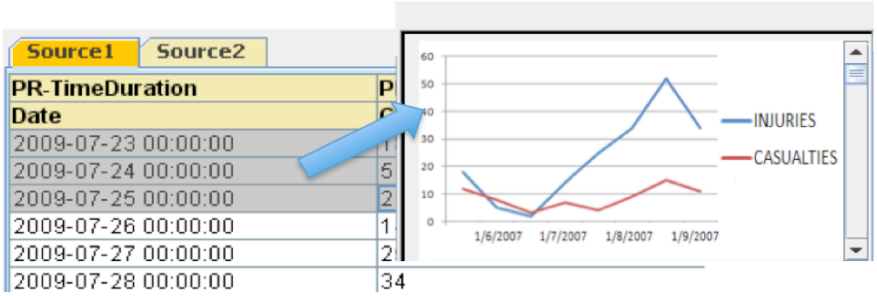
- Visualization by demonstration approach
  - The user demonstrates to Karma the kind of visualization desired for the data specified through examples using a drag and drop mechanism



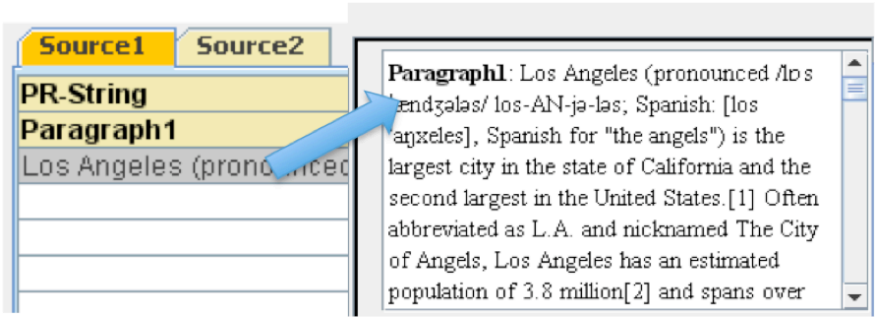
DATA VISUALIZATION

Karma currently supports four types of visualization formats:

1. Chart Format: Useful for visualizing numerical statistics, time based events etc

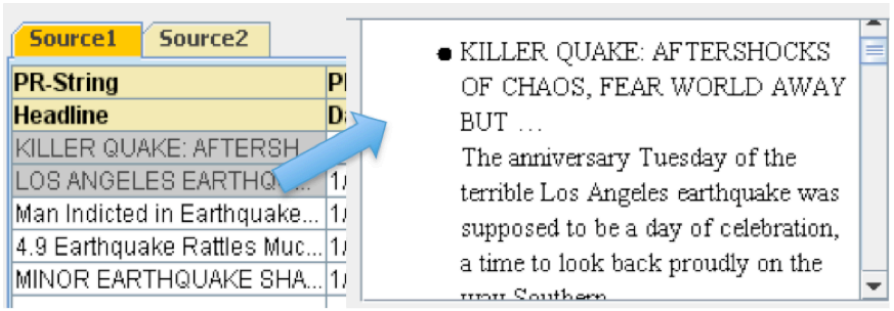


2. Paragraph Format: Useful for visualizing descriptive text data such as Wikipedia definitions

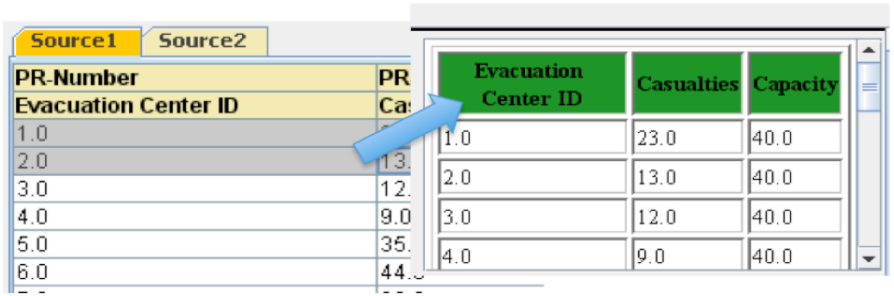


DATA VISUALIZATION

3. List Format: Useful for visualizing information in a bulleted list such as list of summarized news articles

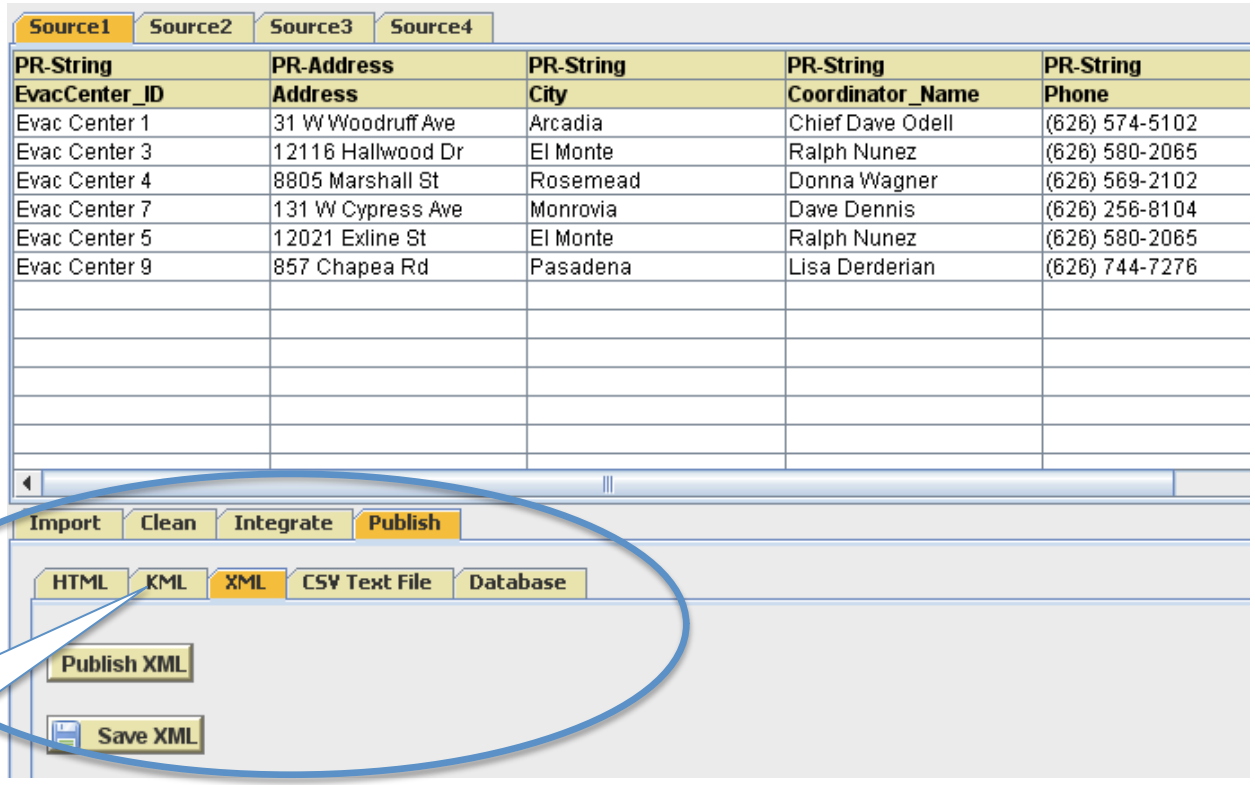


4. Table Format: Useful for visualizing information that is best presented in a row-and-column format such as numerical values etc



RESULTS CAN BE PUBLISHED IN MULTIPLE FORMATS

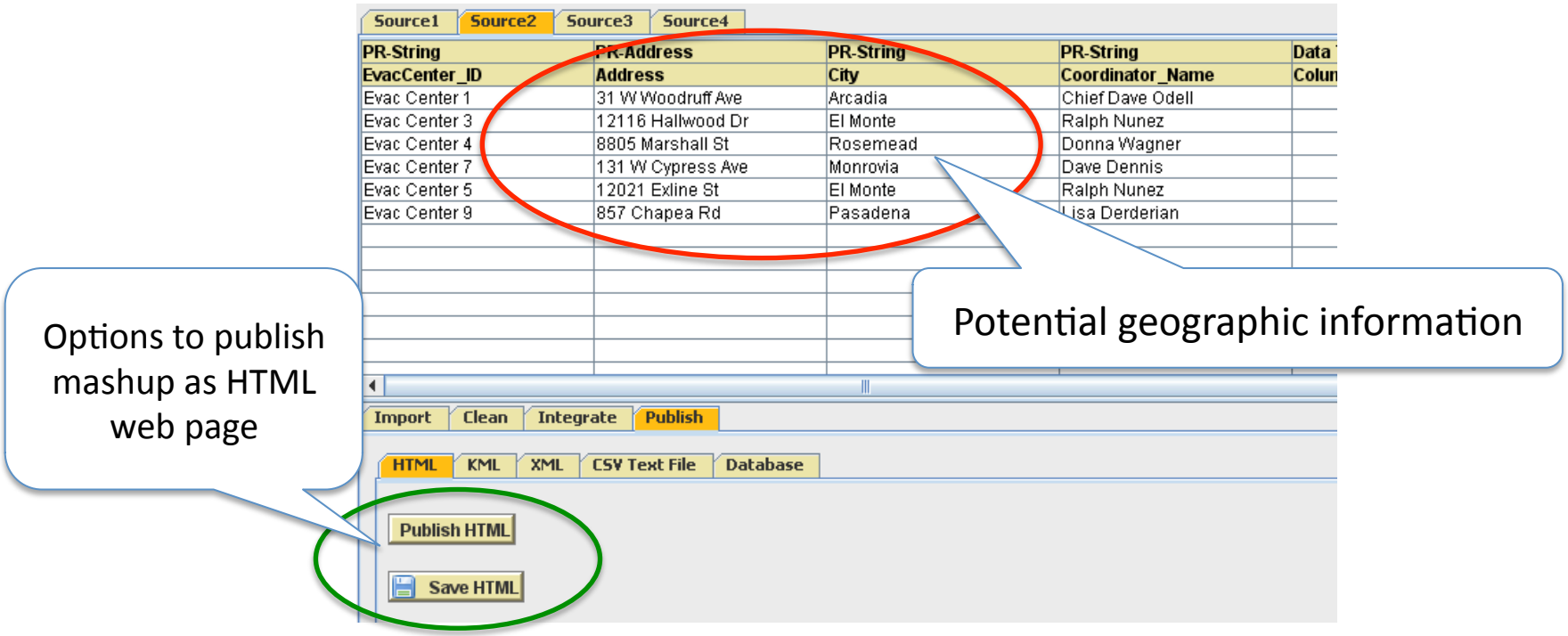
- Karma lets you export your final mashup in variety of formats:
  - HTML Page
  - Database table
  - KML Layer
  - XML File
  - CSV Text File



Different mashup publishing options

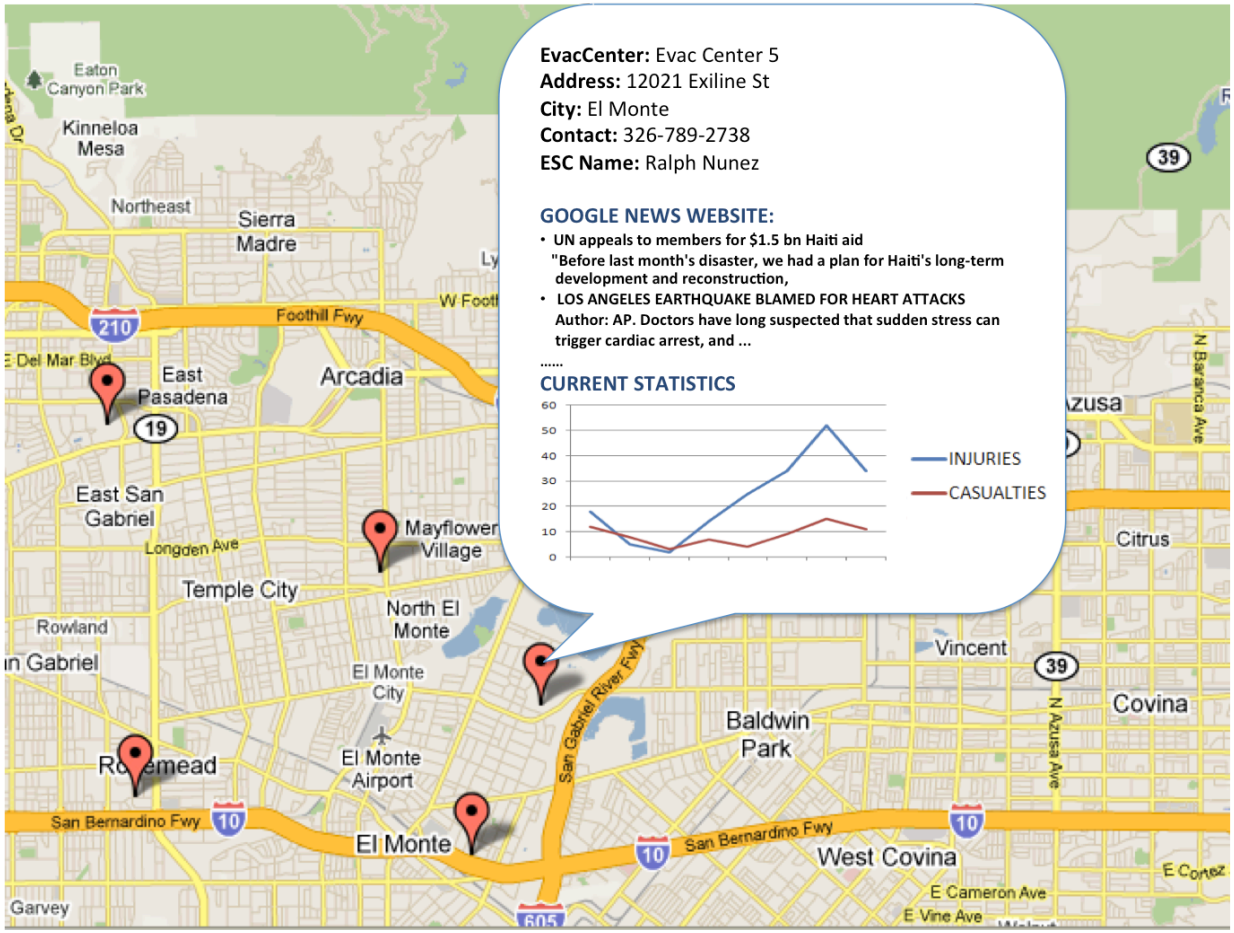
AUTOMATICALLY FINDS GEOSPATIAL REFERENCES

- Final mashup output in HTML web page format:
  - Karma identifies geospatial information in the current data with the help of geographic semantic types such as PR-Address, PR-Latitude etc
  - The Google geocoding service is used to find the coordinates for a given address
  - Karma uses the coordinates information to place the markers in the final mashup



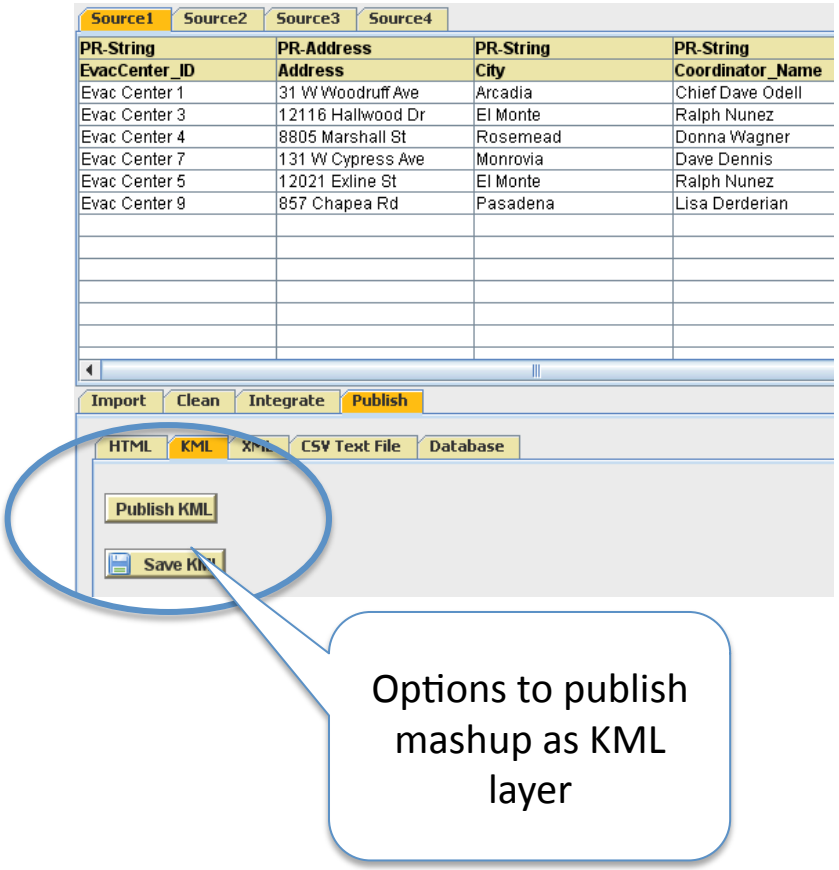
CONSTRUCTS A MAP WITH USER-DEFINED LAYOUT

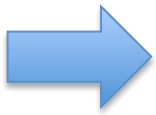
- Final mashup as a HTML web page:



RESULTS CAN BE EXPORTED AS KML

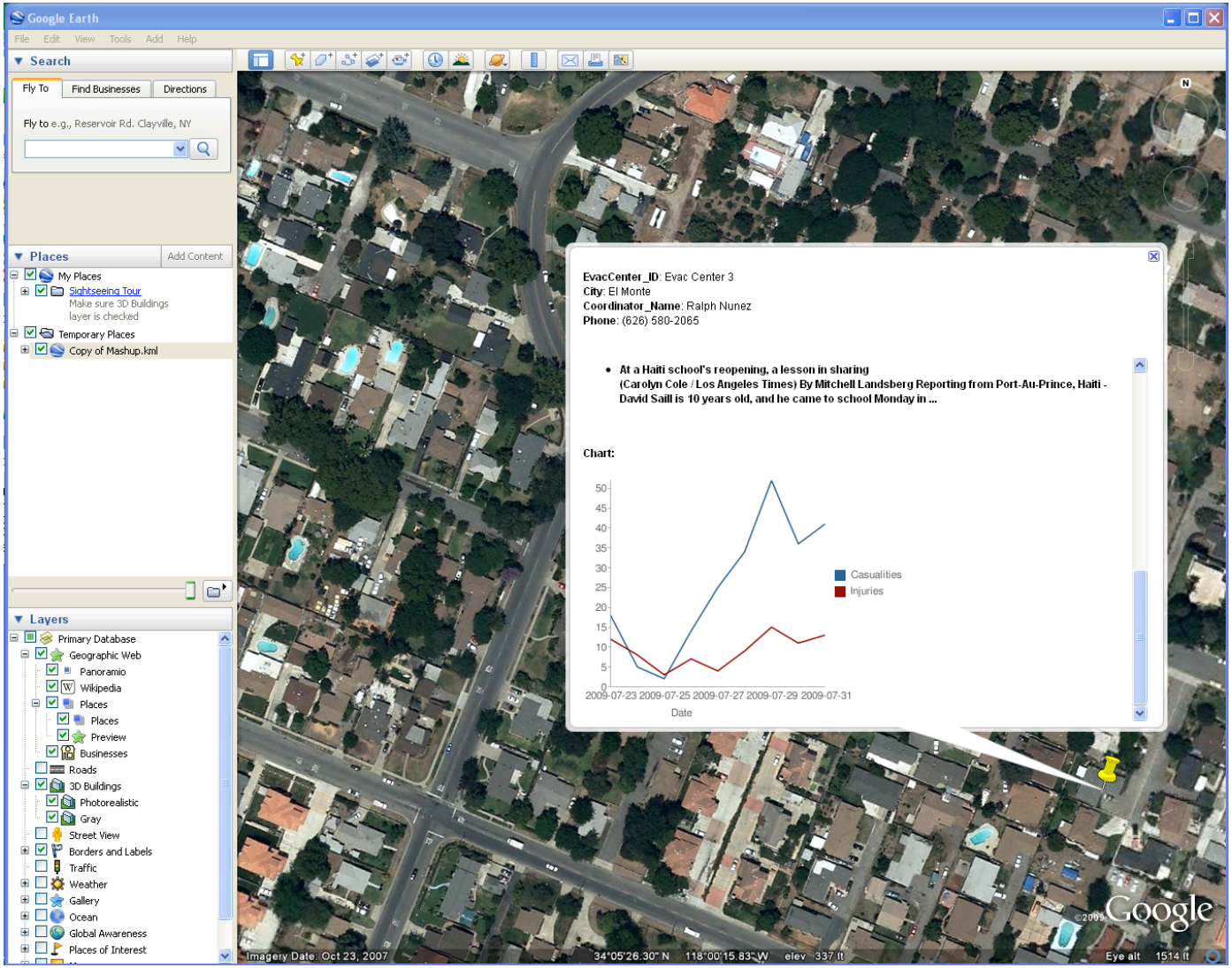
- Final mashup output as a KML layer





```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://earth.google.com/kml/2.2">
<Document xmlns="">
<Folder>
    <Placemark>
        <address>31 W Woodruff Ave, Arcadia</address>
        <description>&lt;b&gt;EvacCenter_ID&lt;/b&gt;:
b&gt;Phone&lt;/b&gt;: (626) 574-5102&lt;br&gt;&lt;br&gt
ul&gt;&lt;li&gt;UN appeals to members for $1.5 bn Haiti
according to The Los Angeles Times. ...</description>
    </Placemark>
    <Placemark>
    <address>12116 Hallwood Dr, El Monte</address>
    <description>&lt;b&gt;EvacCenter_ID&lt;/b&gt;: Evac
```

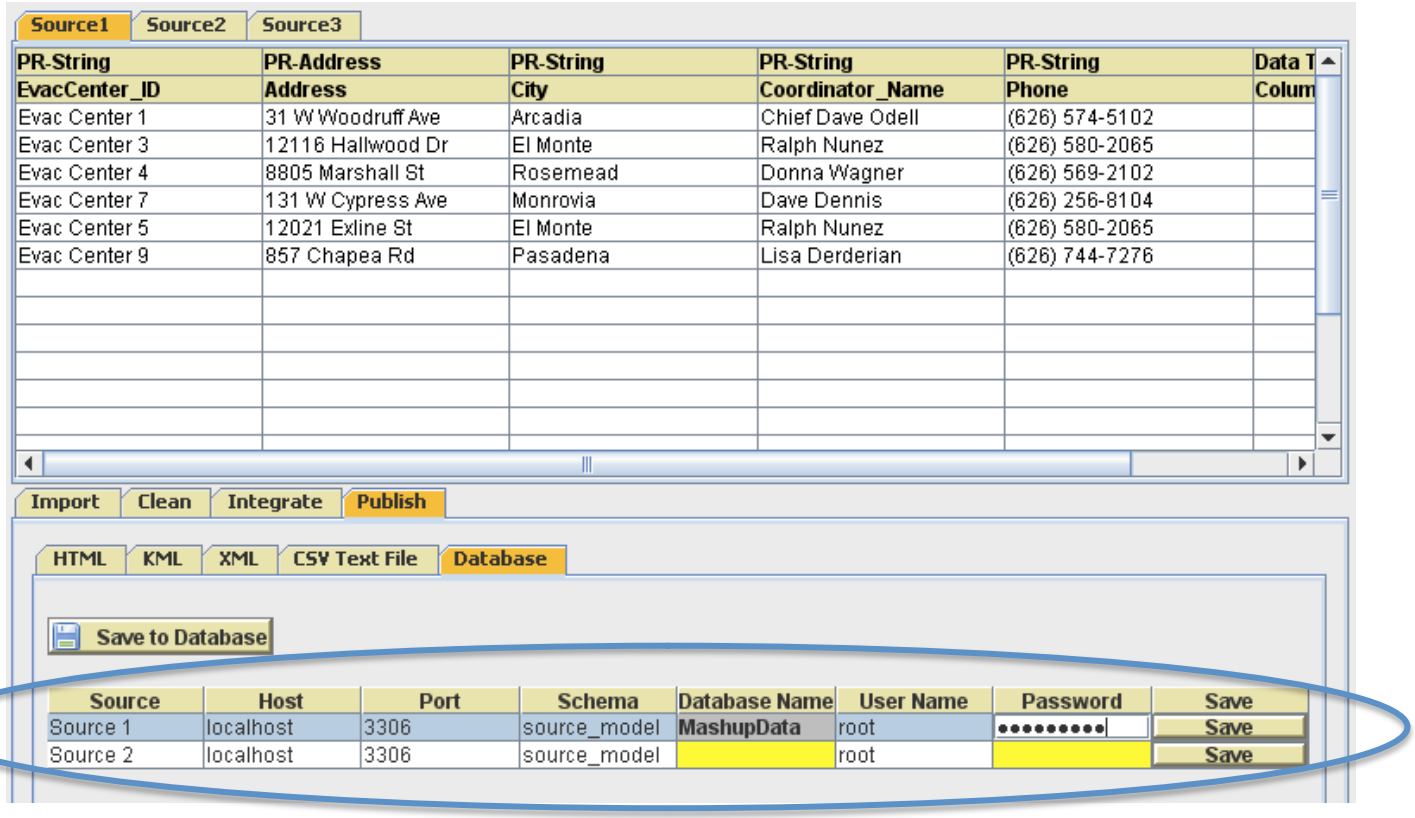
KML LAYERS CAN BE OPENED IN GOOGLE EARTH



The generated KML layer can be viewed in a GIS software such as Google Earth

RESULTS CAN BE STORED IN A DB

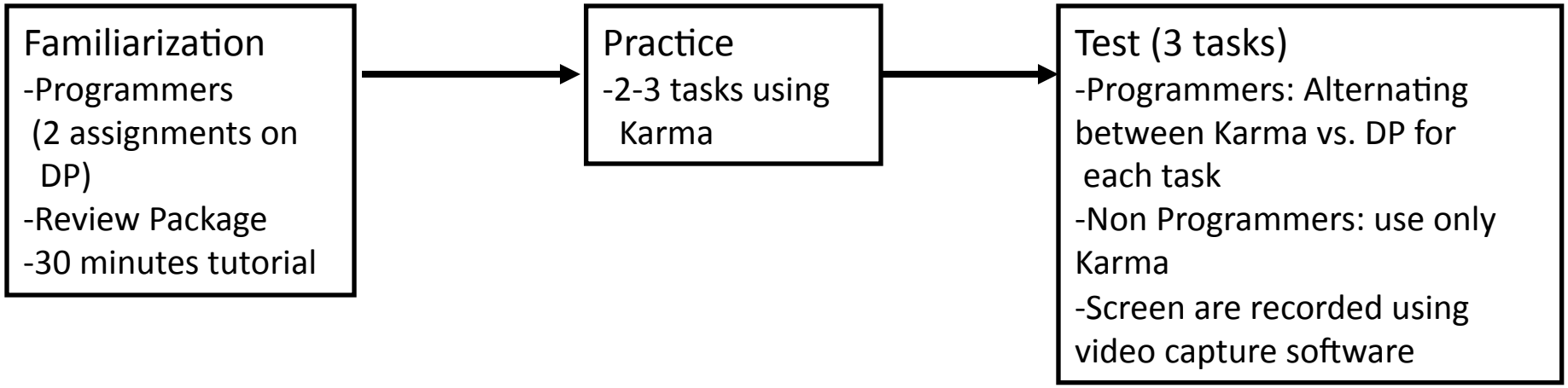
- The final mashup data can also be saved into a database table by providing the details about the database location, username and password, etc in Karma



EVALUATION

- Baseline: A combination of Dapper/Pipes
- Claims:
  1. Users with no programming experiences can build all four Mashup types.
  2. Karma takes less time to complete each subtask and scales better as the tasks get harder
  3. Overall, the user takes less time to build the same Mashup in Karma compared to Dapper/Pipes
- Users:
  - Programmers (20)
  - Non-programmers (3)

EVALUATION: SETUP



Task2	Dapper/Pipes					Karma				
Subject	E	M	C	I	Total	E	M	C	I	Total
No.1	4:38	0:22	2:45	1:15	9:00	1:26	0:43	0:43	0:00	2:52
No.2	1:35	0:12	3:30	0:12	5:29	0:50	0:57	0:57	0:00	2:44
No.3	*5:00	0:25	*5:00	*5:00	15:25	2:52	1:00	3:00	0:00	5:52
No.4	4:49	0:17	3:29	0:38	9:14	1:26	0:48	1:03	0:00	3:18
No.5	*5:00	0:29	1:44	1:16	8:29	1:43	0:45	1:20	0:00	3:48
No.6	*5:00	0:20	*5:00	*5:00	15:20	2:07	0:30	0:50	0:00	3:27

5 minute cut off time

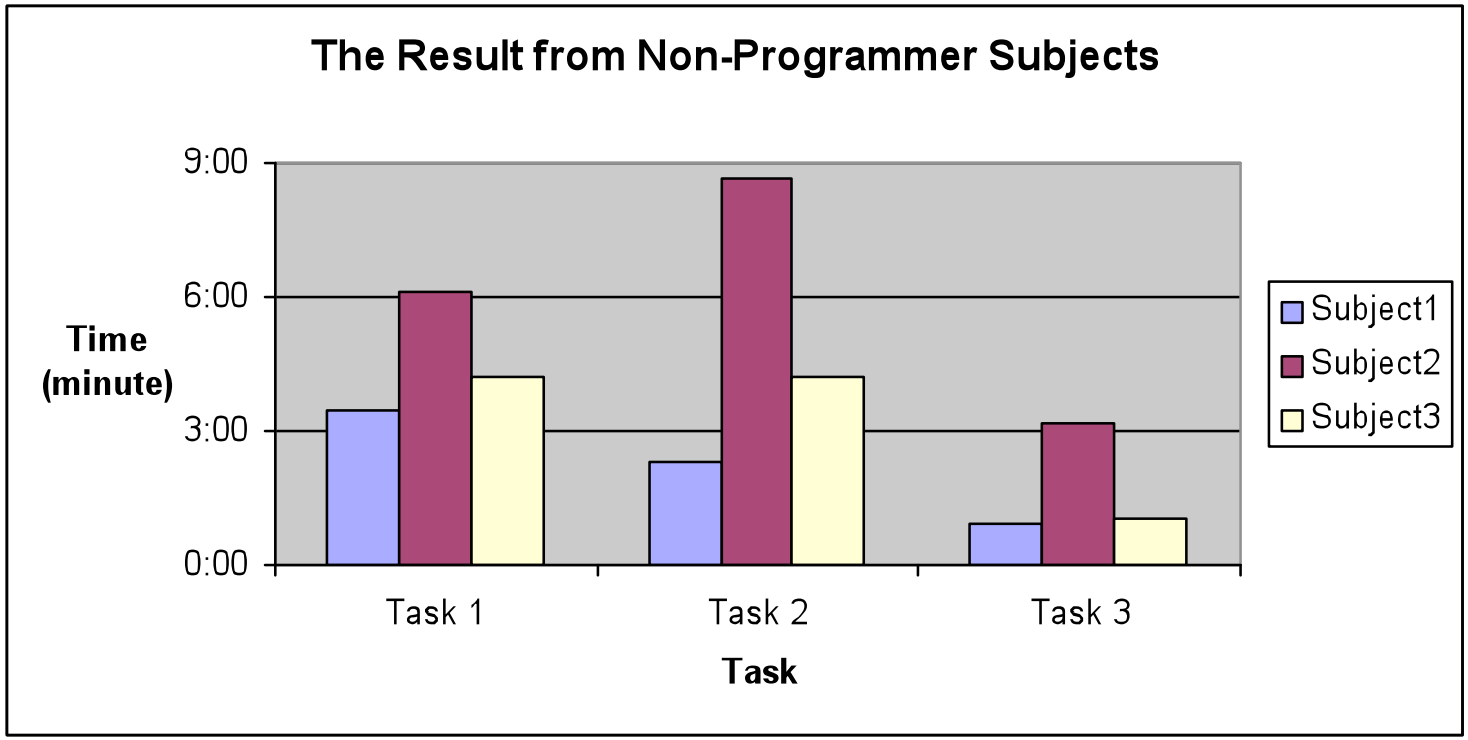
EVALUATION: TASKS

Task No.	Mashup Type	Data Extraction	Source Modeling	Data Cleaning	Data Integration
1	1 (1 source)	Moderate	Simple	Difficult	N/A
2	2,3 (union+form)	Difficult	Simple	Simple	Union (simple)
3	4 (join 2 sources)	Simple	Simple	N/A	Join (difficult)

- Claim 1: Users with no programming experiences can build all four Mashup types
- Claim 2: When the Mashup subtask is difficult, Karma takes less time to complete that subtask
- Claim 3: Overall, the user takes less time to build the same Mashup in Karma compared to Dapper/Pipes

Claim 1: Users with no programming experiences can build all four Mashup types

EVALUATION: NON-PROGRAMMERS

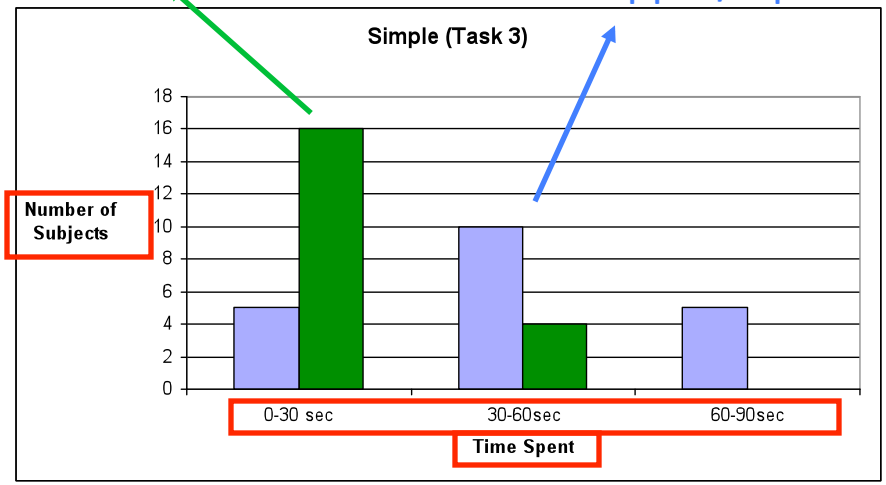


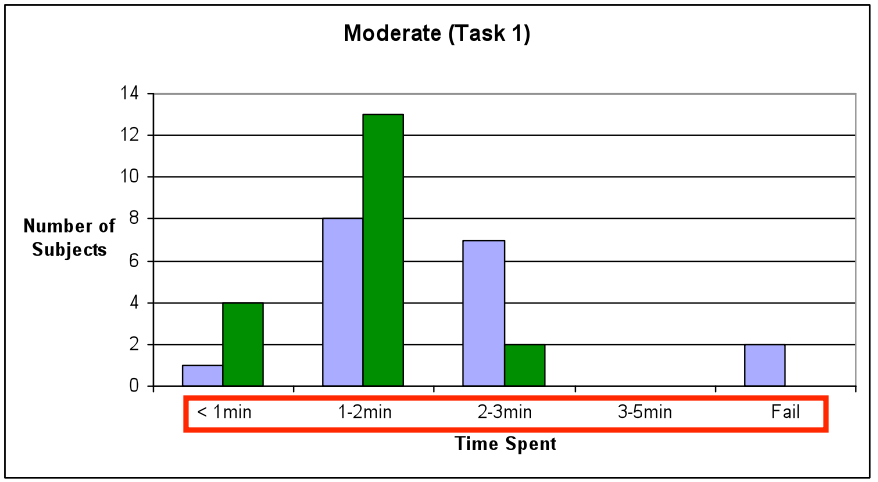
Claim 2: Karma takes less time to complete each subtask

EVALUATION: EXTRACTION

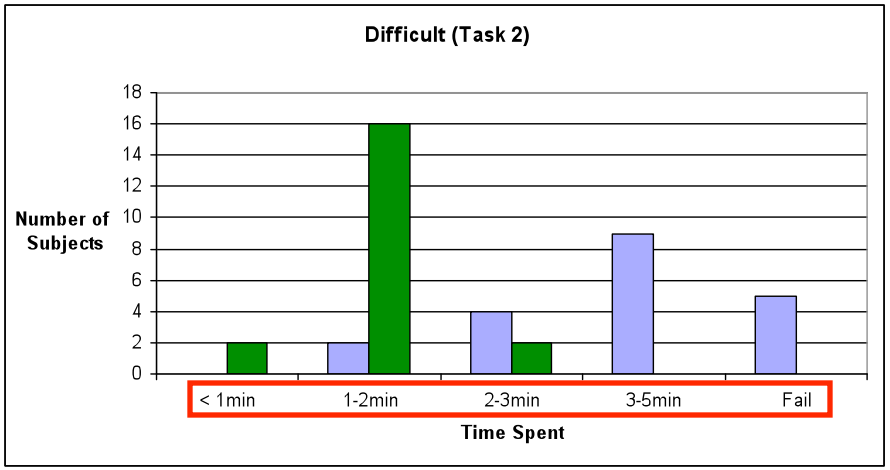
Karma (programmer)

Dapper/Pipes



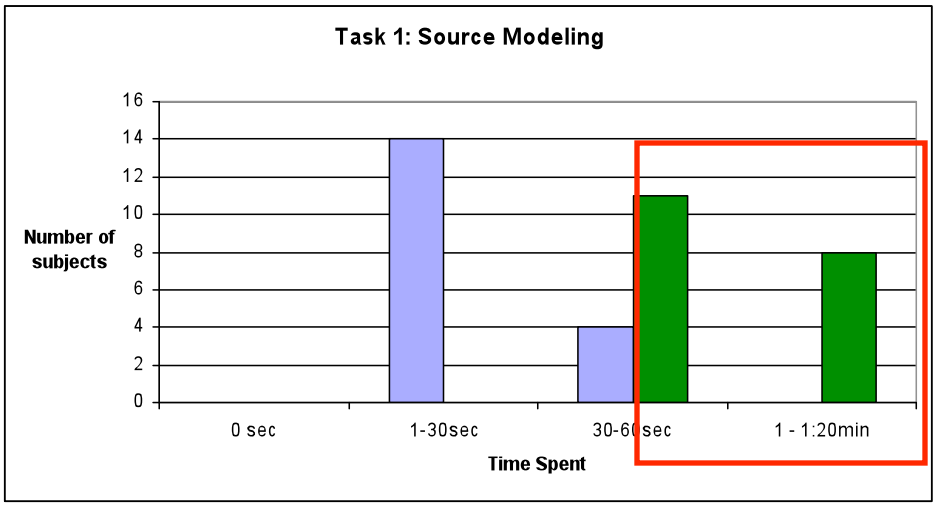


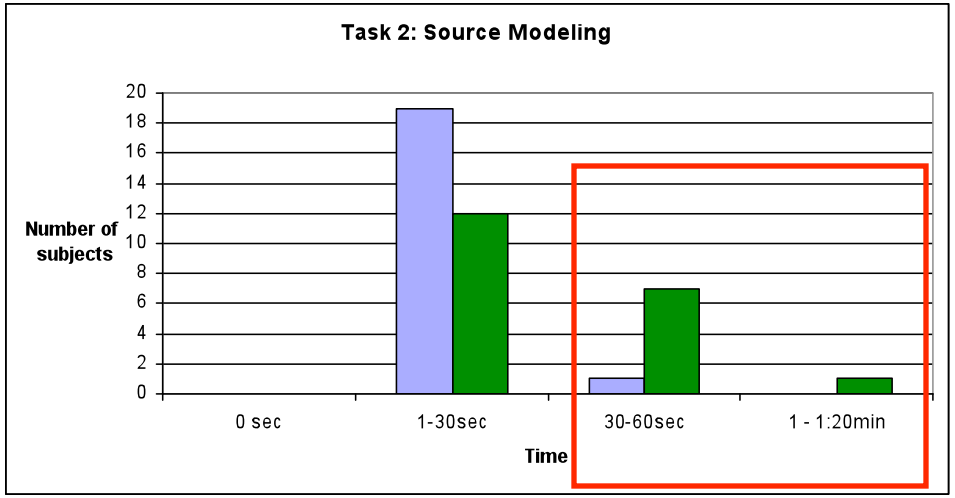
- As the extraction task gets more difficult, Dapper/Pipes takes
  - longer
  - more subjects failing to complete the task (11% for moderate and 25% for difficult)



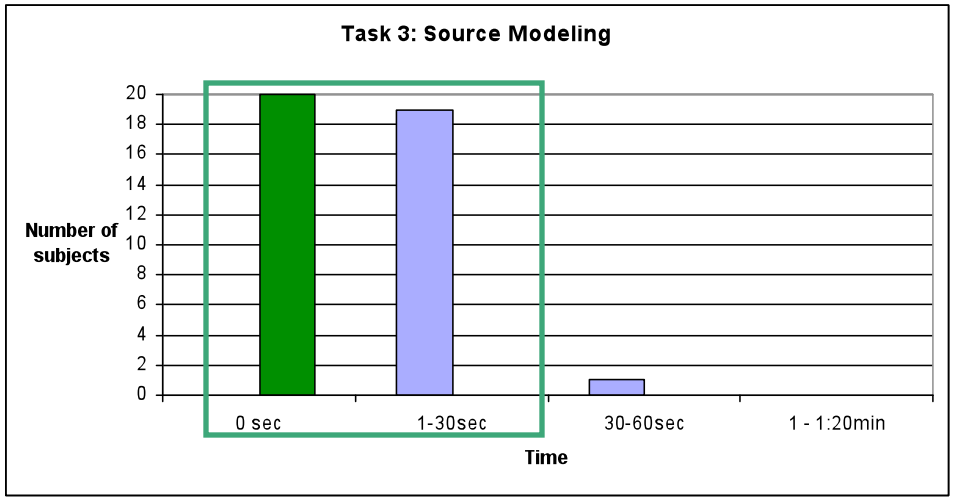
Dapper/Pipes Karma

EVALUATION: SOURCE MODELING



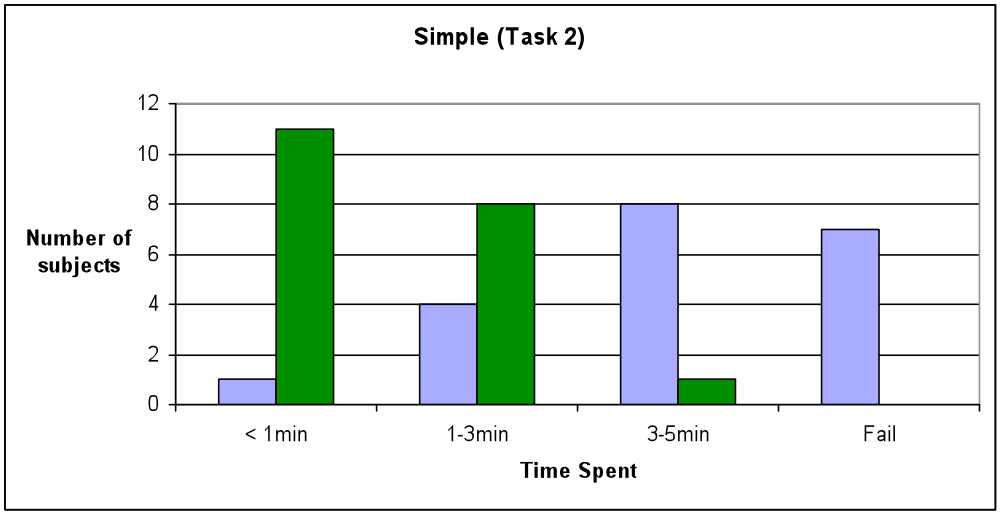


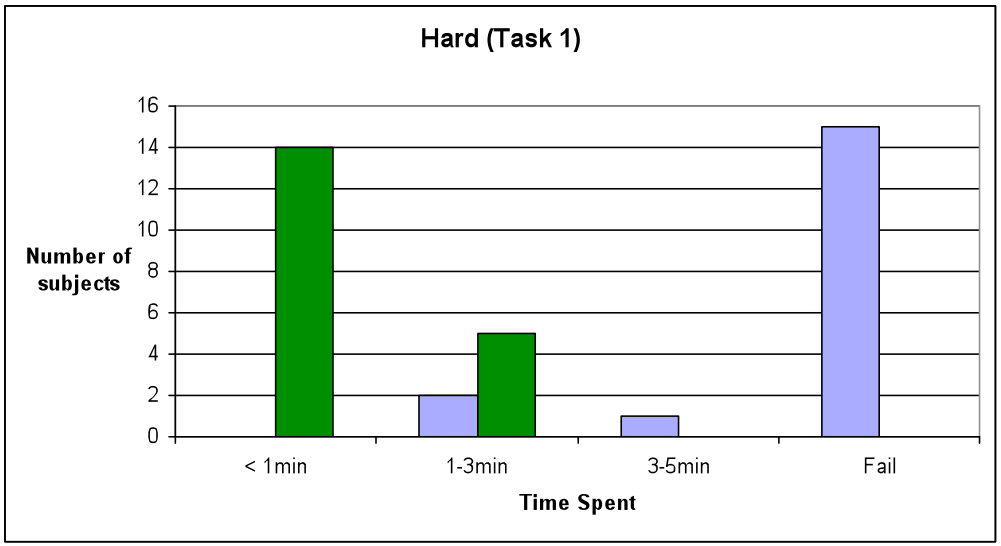
- Karma performed worse in task 1 and tasks 2
  - only 30 sec difference
  - subjects take times selecting attributes
  - the saving will be realized in the data integration step.
- Karma performed better in task 3 because it can automatically identify the attribute



Dapper/Pipes Karma

EVALUATION: DATA CLEANING

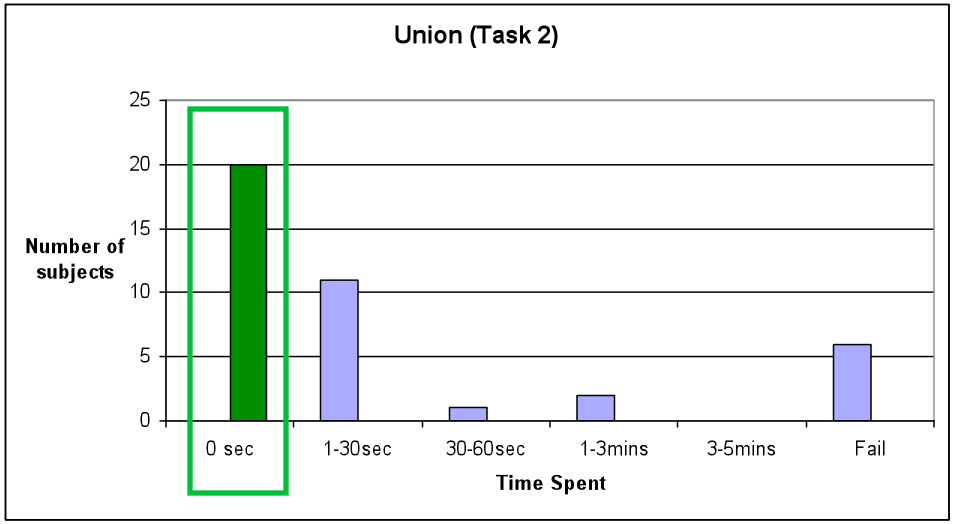


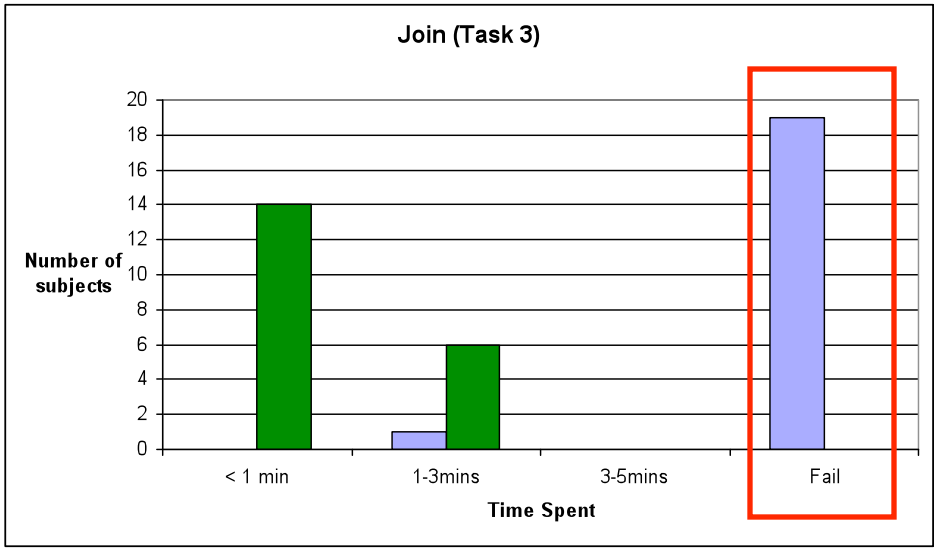


- Karma performed better in both tasks
- When the cleaning task gets harder, more subjects are failing in Dapper/ Pipes (35% for simple and 83% in hard)

Dapper/Pipes Karma

EVALUATION: DATA INTEGRATION



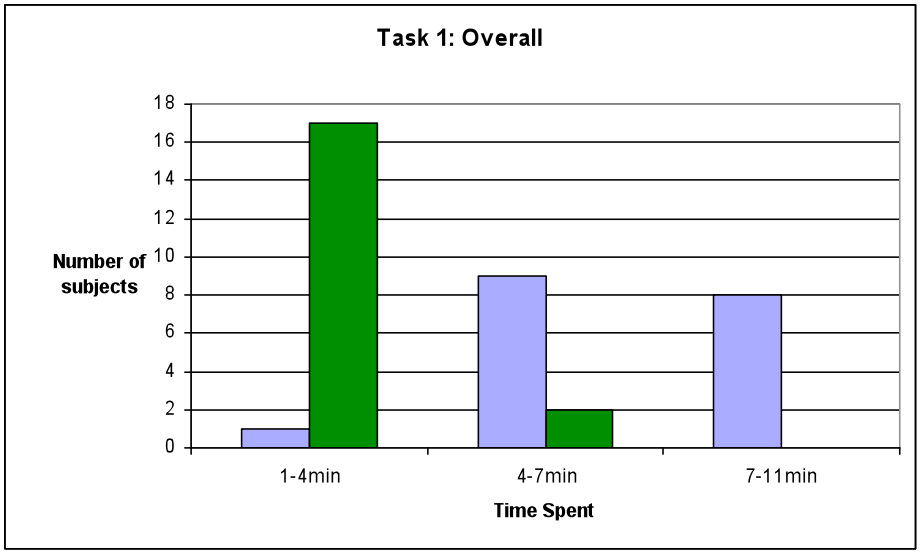


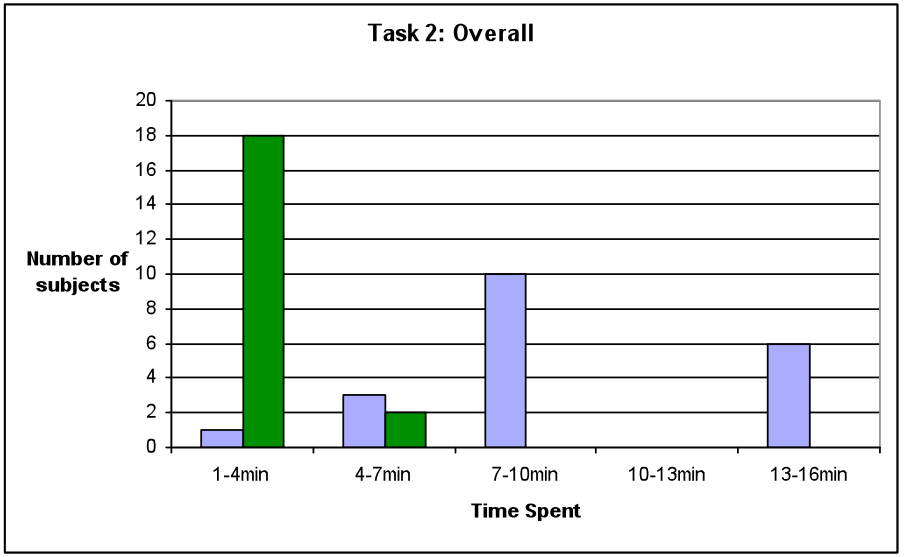
- Because of the table structure, subjects can specify union indirectly by dropping data into the right cell
- The time spent in source modeling step allows Karma to suggest the linking source
- Dapper/Pipes: 30% fail in the union case and 95% fail in the join case

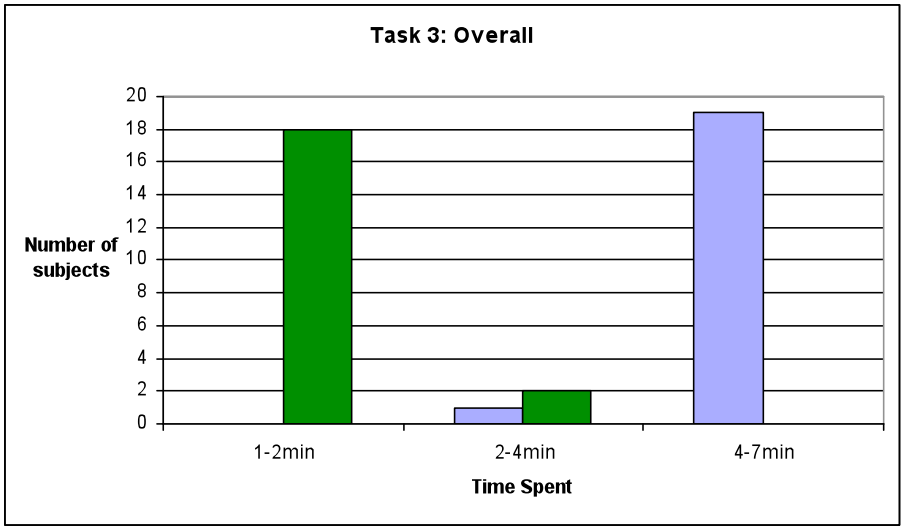
Dapper/Pipes Karma

Claim 3: Overall, the user takes less time to build the same Mashup in Karma compared to Dapper/Pipes

EVALUATION: OVERALL

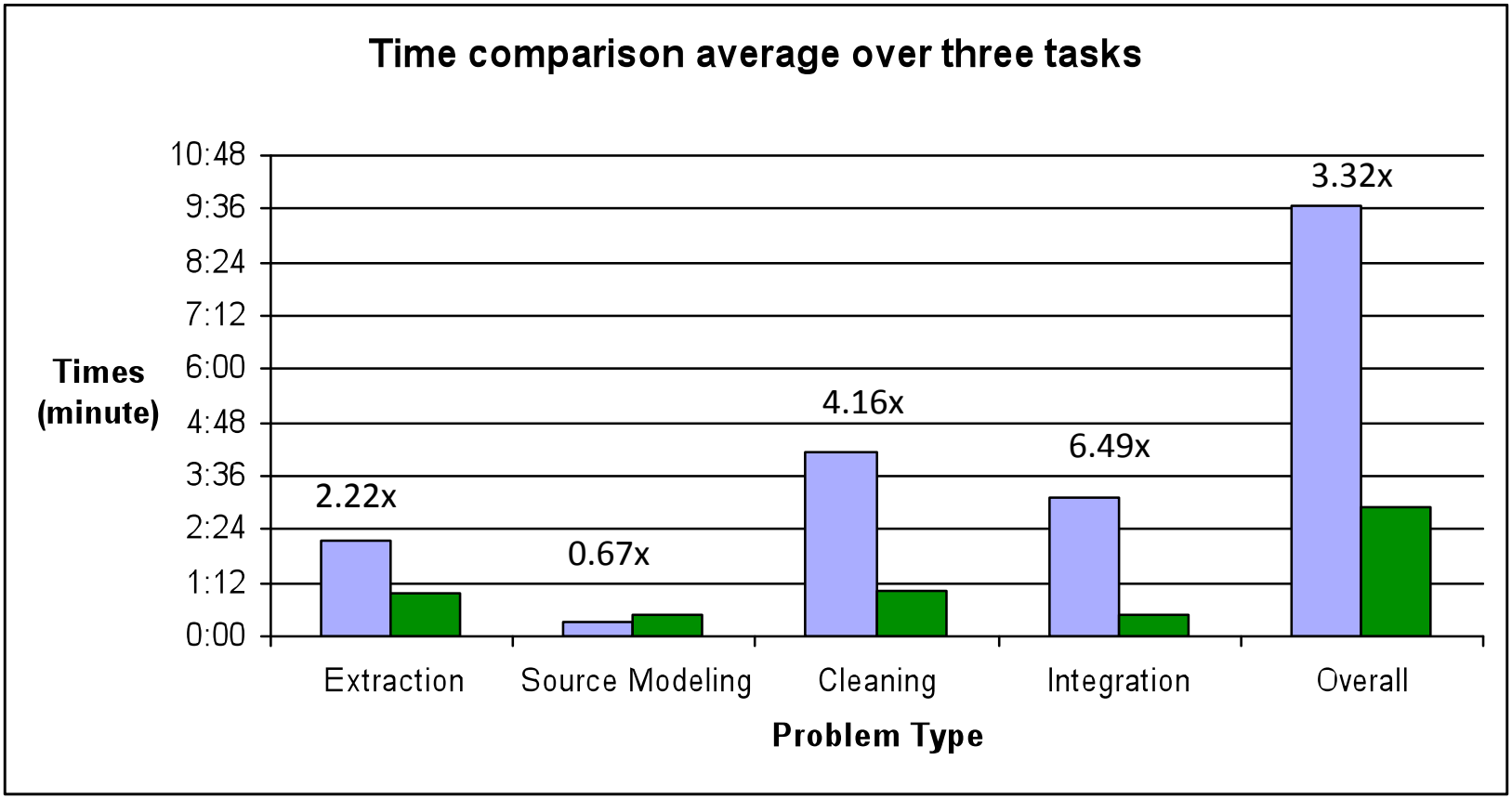






Dapper/Pipes Karma

EVALUATION: AVERAGE



RELATED WORK: MASHUP BUILDING TOOLS

System	Data Retrieval	Source Modeling	Data Cleaning	Data Integration	Mashup Type Supported
MIT's Simile	Early work. Focus on DOM, too basic				1
MIT's Pot Luck	RDF / Manually specify data int				1,3,4
Dapper	Mainly focus on extraction / linear				1,2,4
Yahoo's Pipes	Widgets				1,2,3
MS's Popfly	Fancier UI/ more widgets				1,2,4
CMU's Marmite	Fewer Widgets / Confusion on workflow				1,2,4
Intel's Mashmaker	Require an expert				1,2,3,4
Google MyMap	Create points on Map				1,2
Agent Wizard	Q/A approach / linear / scalability				1,3,4
Cards	Tuple = card. Drawing links for relations				1,2,4
Karma	DOM	Database	PBD	PBD	1,2,3,4

1: Extraction, 2: Union, 3: Form-based Interaction, 4: Join

RELATED WORK: DATA EXTRACTION

- Automatic extraction: table and lists only
  - RoadRunner (exploit HTML structure) [Crescenzi et al., 2001]
  - Adel (grammer induction to detect rows) [Lerman+ 2001]
  - VisualWeb (OCR technique to detect tables) [Gatterbauer+ 2007]
- Semi-Automatic: require more label examples
  - WIEN (inductive – less expressive than stalker) [Kushmerick 1997]
  - Stalker (Cotesting) [Muslea+ 1999]
  - SoftMealy (finite state transducer) [Hsu 1998]
  - WHISK (rigid format, exact delimiter) [Soderland 1998]
- DOM: rely on well-formed HTML and less labeling
  - Simile [Huynh+ 2005]
  - Dapper
  - Interactive Wrapper Generation (ML + prediction on DOM)[Irmak+ 2006]
  - PLOW (add natural language) [Allen+ 2007]
  - Cards [Dontcheva+ 2007]
  - Karma [Tuchinda+ 2008]

RELATED WORK: SOURCE MODELING

- 1:1 mapping, N:M mapping
  - Schema-level match
    - TranScm [Milo+ 98]
    - DIKE [Palopoli+ 99]
    - Artemis [Castano+ 01]
    - Delta [Clifton+ 97]
  - +Instance-based matcher
    - SemInt [Li 00]
    - LSD [Doan 01]
    - ILA [Etzioni 95]
    - iMapp [Dhamanka 04]
    - Clio (interactive) [Ling 01]
    - Inducing Source Description [Carman 07]
- Karma leverages existing techniques to narrow candidate matches
  - String Similarities [Cohen+ 2003]

RELATED WORK: DATA CLEANING

- Commercial Tools: Focus on writing transformation
  - ACR/Data, Migration Architect [Chaudhuri+ 1997]
- Discrepancy Detection: Use as a stepping stone for record linkage and cleaning system
  - Levenshtein distance [Needleman+ 70]
  - Vector based [Baeza-Yates+ 99]
  - EM [Ristad+ 98]
  - SVM [Bilenko+ 03]
- Record linkage & cleaning systems: Focus on ranking [Winkler 06]
  - Fuzzy Match [Chaudhuri+ 03]
  - Apollo [Michalowski+ 05]
  - Phoebus [Michelson+ 07]
  - Potter's wheel [Raman+ 01]
- Karma
  - Gains reference sources through source modeling process
  - Provides predefined transformations

RELATED WORK: DATA INTEGRATION

- **Universal Relation:** Make it easier to formulate the query but users still need to formulate the query [Ullman 1980, 1988]
- **Query by example:** Need to know which data sources to use and the query may not return results
  - QBE [Zloof 1975]
- **Retrieval by formulation:** Need to understand domain model to formulate partial description
  - Helgon [Fischer 1989], RABBIT [Williams 1982]
- **Graphical Query Language:** Users still need to navigate through sources (graphs)
  - Gql [Benzi 1998, Haw 1994, Papantonakis 1988]
- **Question-Answering Techniques:** Understanding about database operations required
  - Agent Wizard [Tuchinda+ 2004]
- **Interactive Schema/data integration:** Understanding about source schema required
  - Clio [Ling 01]
- Karma is based on **Programming by Demonstration** [Cyper 2001; Lau2001]

CONCLUSION

- Mashups are a fast growing area
  - Need an efficient way to for casual web users to build them
- Contributions
  - A PBD approach that uses a single table for building a Mashup
  - An integrated approach that solves the various Mashup building issues
  - A query formulation technique that allows users to specify examples to build complicated queries
- Evaluated the validity of the Karma approach
  - Subjects were able to complete Mashup building tasks in Karma
  - The overall improvement is at least a factor of 3.5

FUTURE WORK

- Learn and generalize over the task
  - Store the integration plan so that it can be reexecuted on current data
- Support the integration of geospatial data types (i.e., vector layers, raster layers)
- Improve the techniques for automatic source modeling
- Learn new transformations from examples for data cleaning

PAPERS

- **Building geospatial mashups to visualize information for crisis management.** Shubham Gupta and Craig A. Knoblock. In *Proceedings of the 7th International Conference on Information Systems for Crisis Response and Management*, 2010.
- **Interactive data integration through smart copy & paste.** Zachary G. Ives, Craig A. Knoblock, Steven Minton, Marie Jacob, Partha Pratim Talukdar, Rattapoom Tuchinda, Jose Luis Ambite, Maria Muslea, and Cenk Gazen, *Fourth Biennial Conference on Innovative Data Systems Research (CIDR), 2009.*
- **Building mashups by example**. Rattapoom Tuchinda, Pedro Szekely, and Craig A. Knoblock. *Proceedings of the 2008 International Conference on Intelligent User Interfaces, 2008*
- **Building data integration queries by demonstration.** Rattapoom Tuchinda, Pedro Szekely, and Craig A. Knoblock. In *Proceedings of the International Conference on Intelligent User Interfaces, 2007*