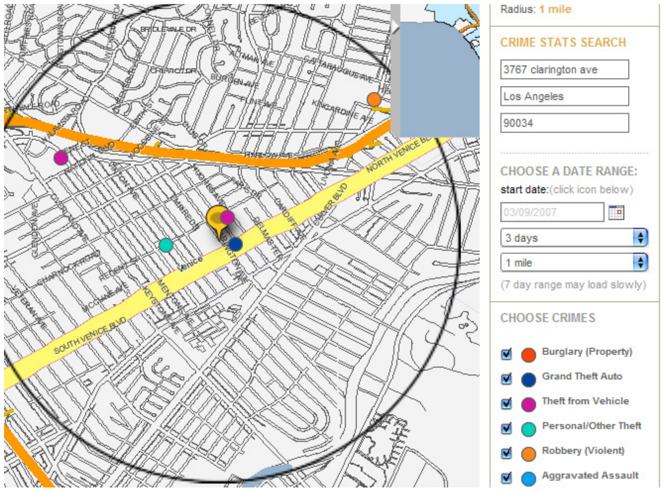
Building Mashups by Example

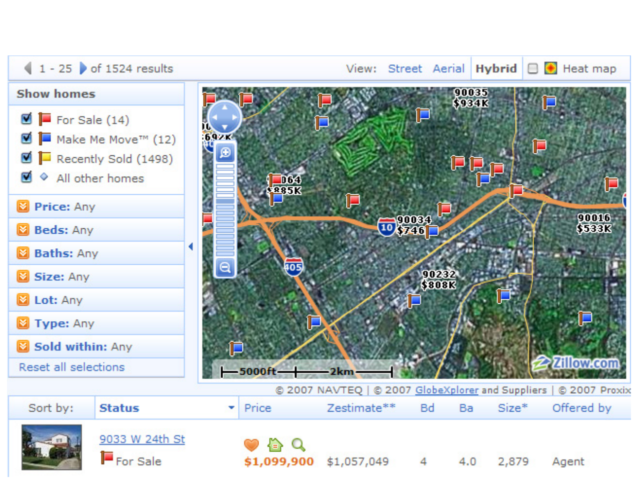
Craig A. Knoblock
University of Southern California

Work in collaboration with Rattapoom Tuchinda and Pedro Szekeley

What's a Mashup?

A website or application that combines content from more than one source into an integrated experience [wikipedia]





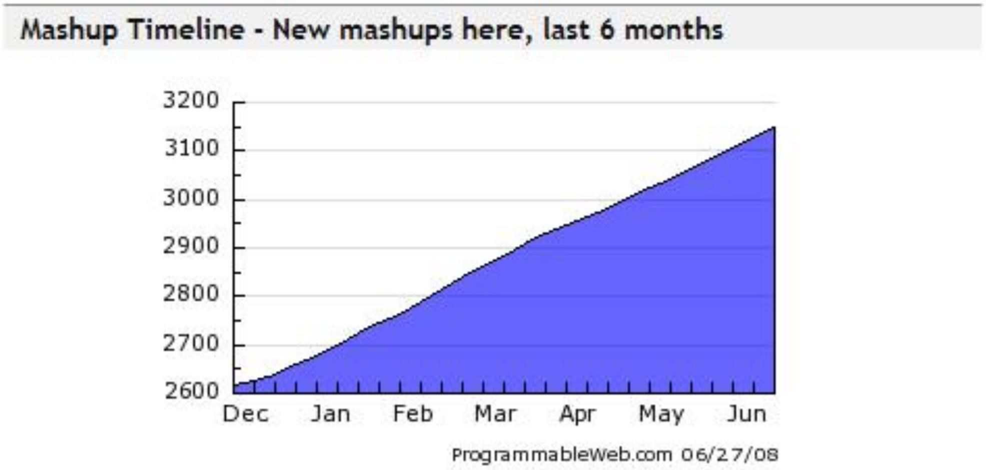
a) LA crime map
- Crime Report from different counties
- Map

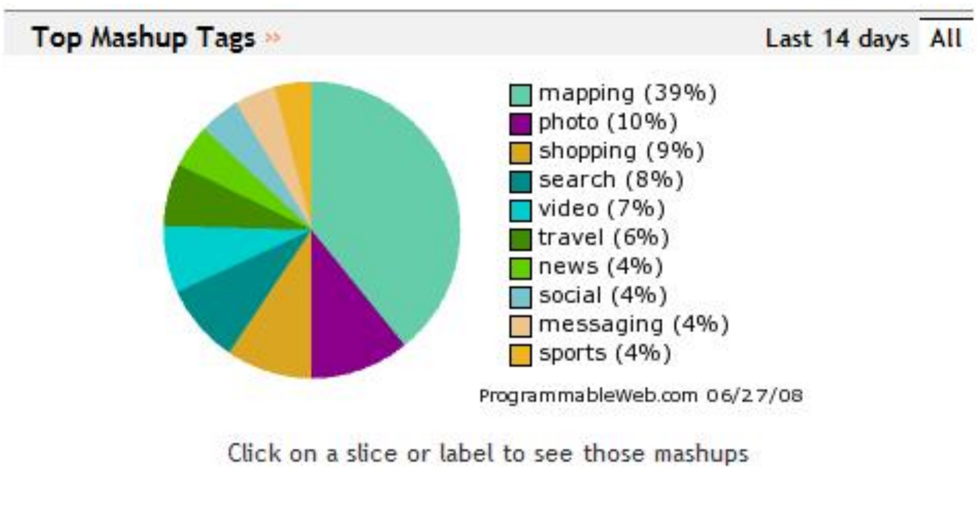
b) zillow.com
- Real Estate Listing
- Property Tax

c) Ski bonk
- Weather
- Snow Report
- Snow Resorts

Combined Data gives new insight / provides new services

Statistics and Trends





Survey of top 50 Mashups

- Divide into five categories based on programming structures
- Focus of this work is on the first four categories which account for 47% of the most popular Mashups

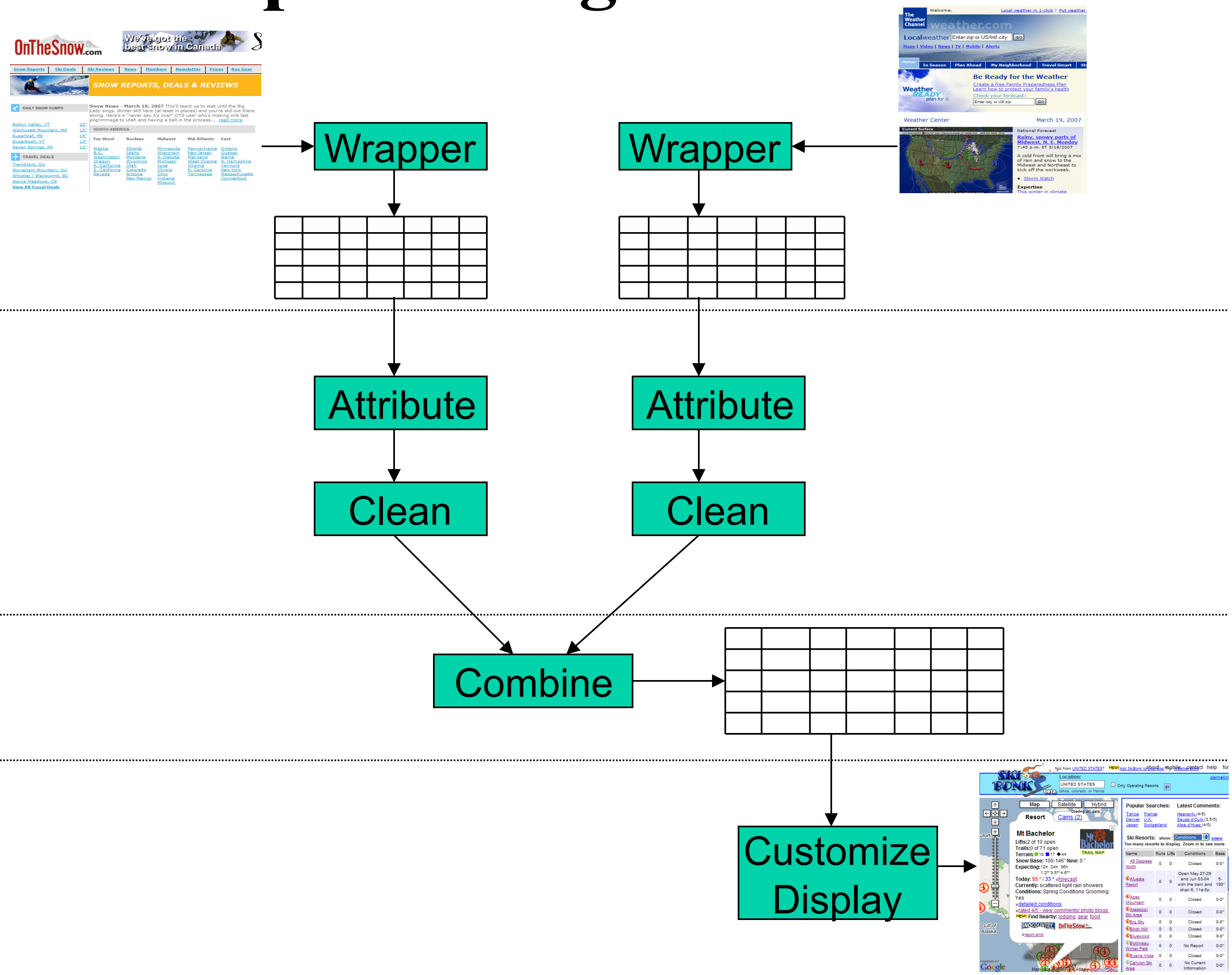
Mashup Building Issues

Data Retrieval

Calibration
-source modeling
-cleaning

Integration

Display

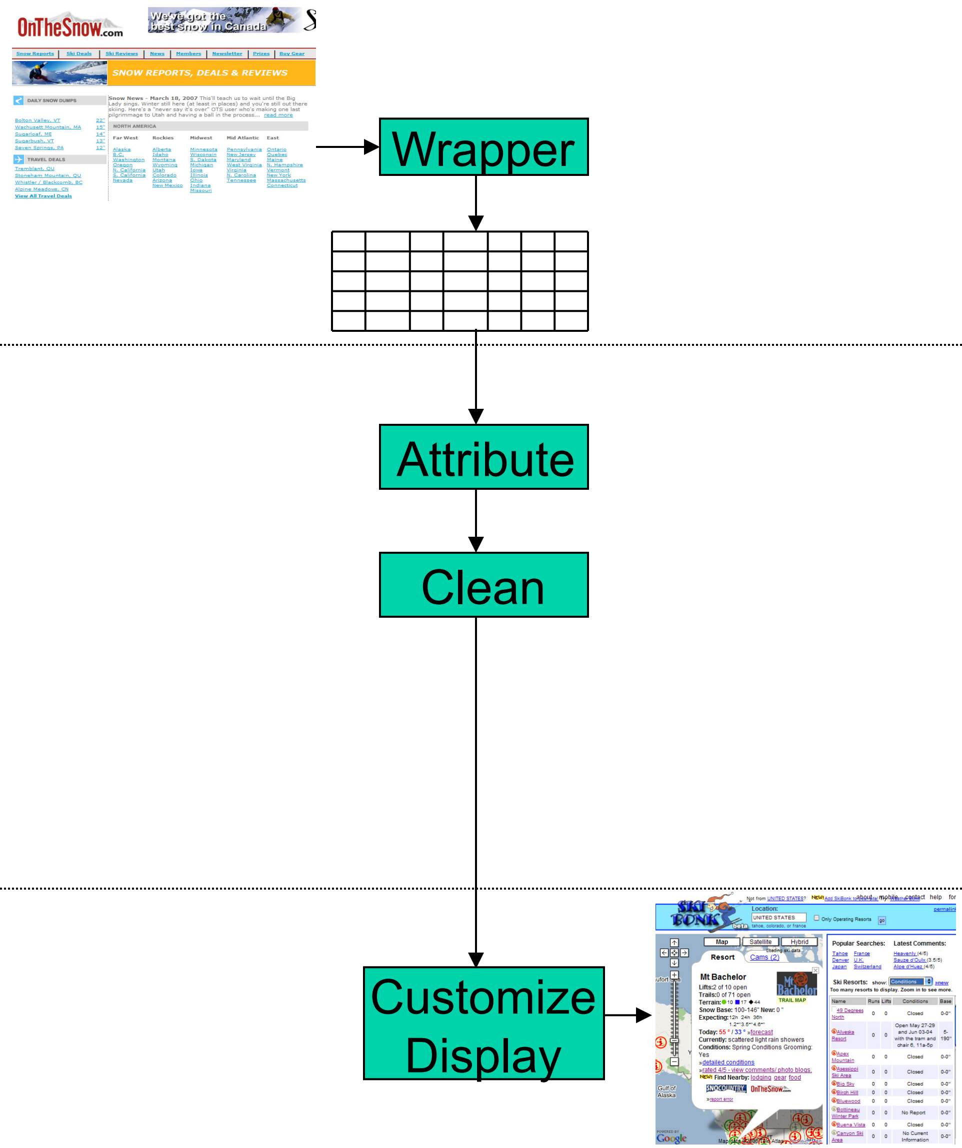


Type 1: One Simple Source

Data Retrieval

Calibration
-source modeling
-cleaning

Display



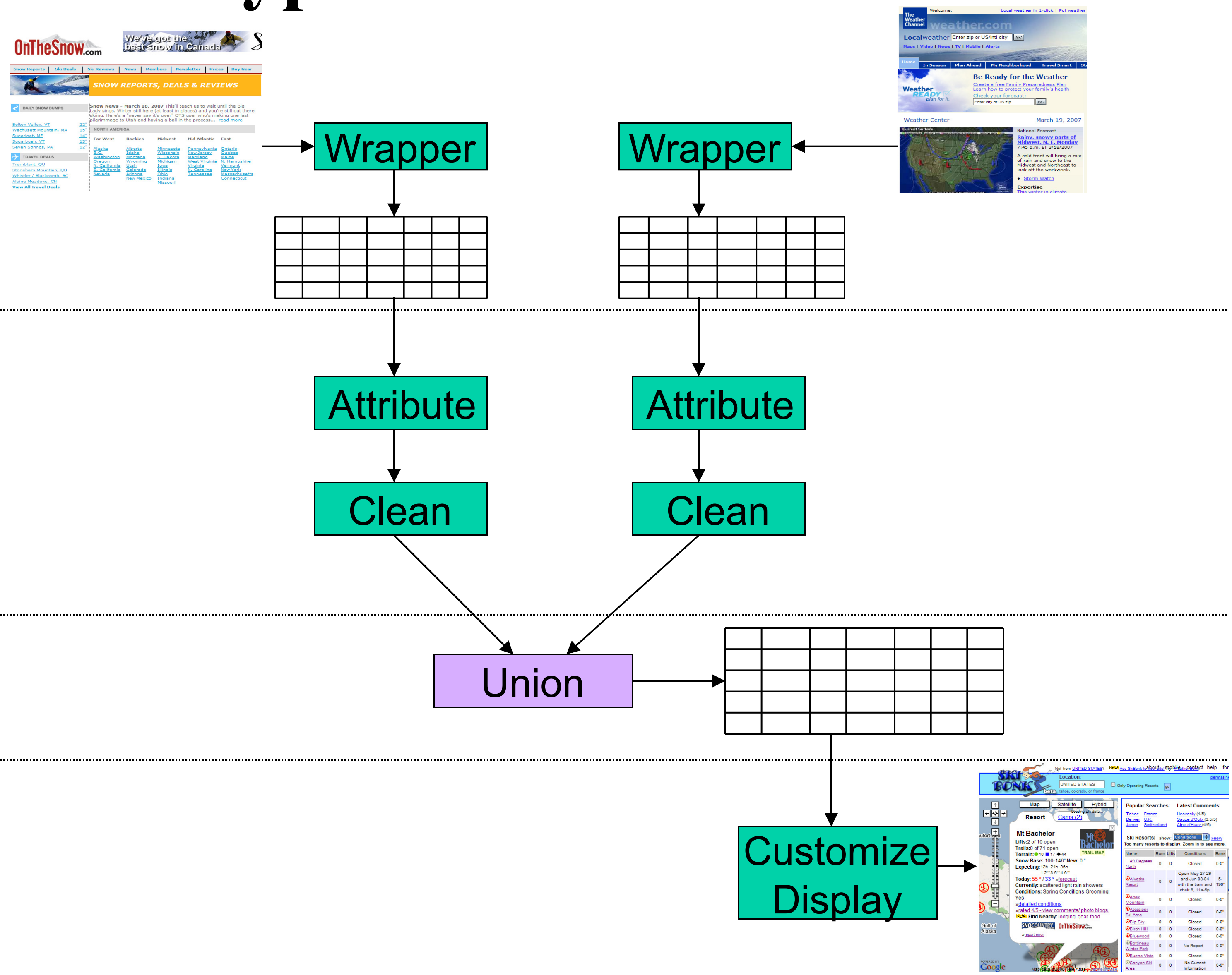
Type 2: Union

Data Retrieval

Calibration
-source modeling
-cleaning

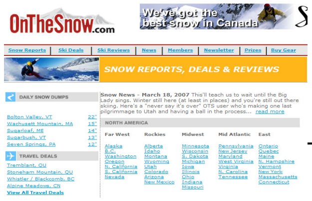
Integration

Display



Type 3: One Source with Form

Data Retrieval



Wrapper

Calibration
-source modeling
-cleaning

Attribute

Clean

Display

Customize Display

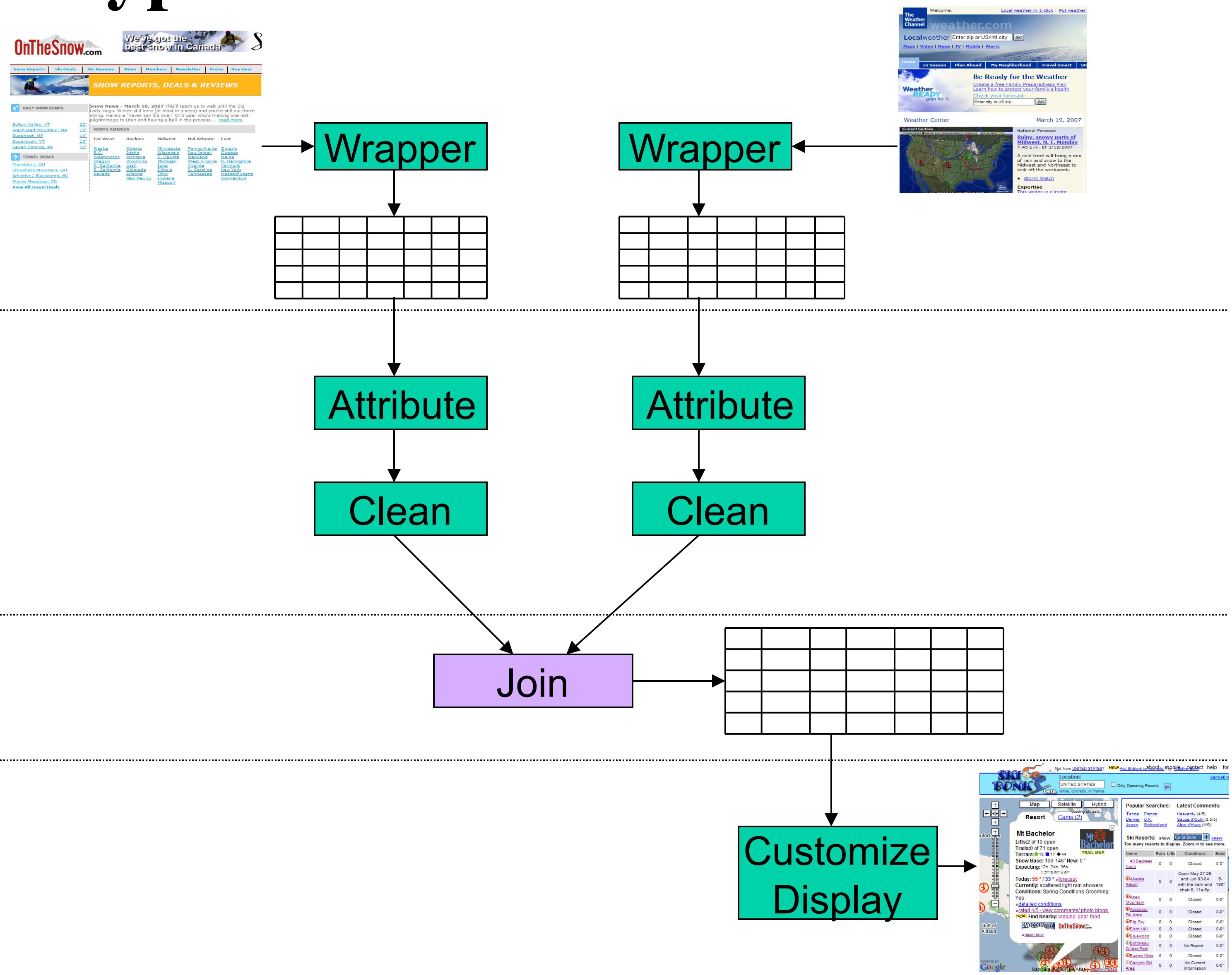
Type 4: Database Join

Data Retrieval

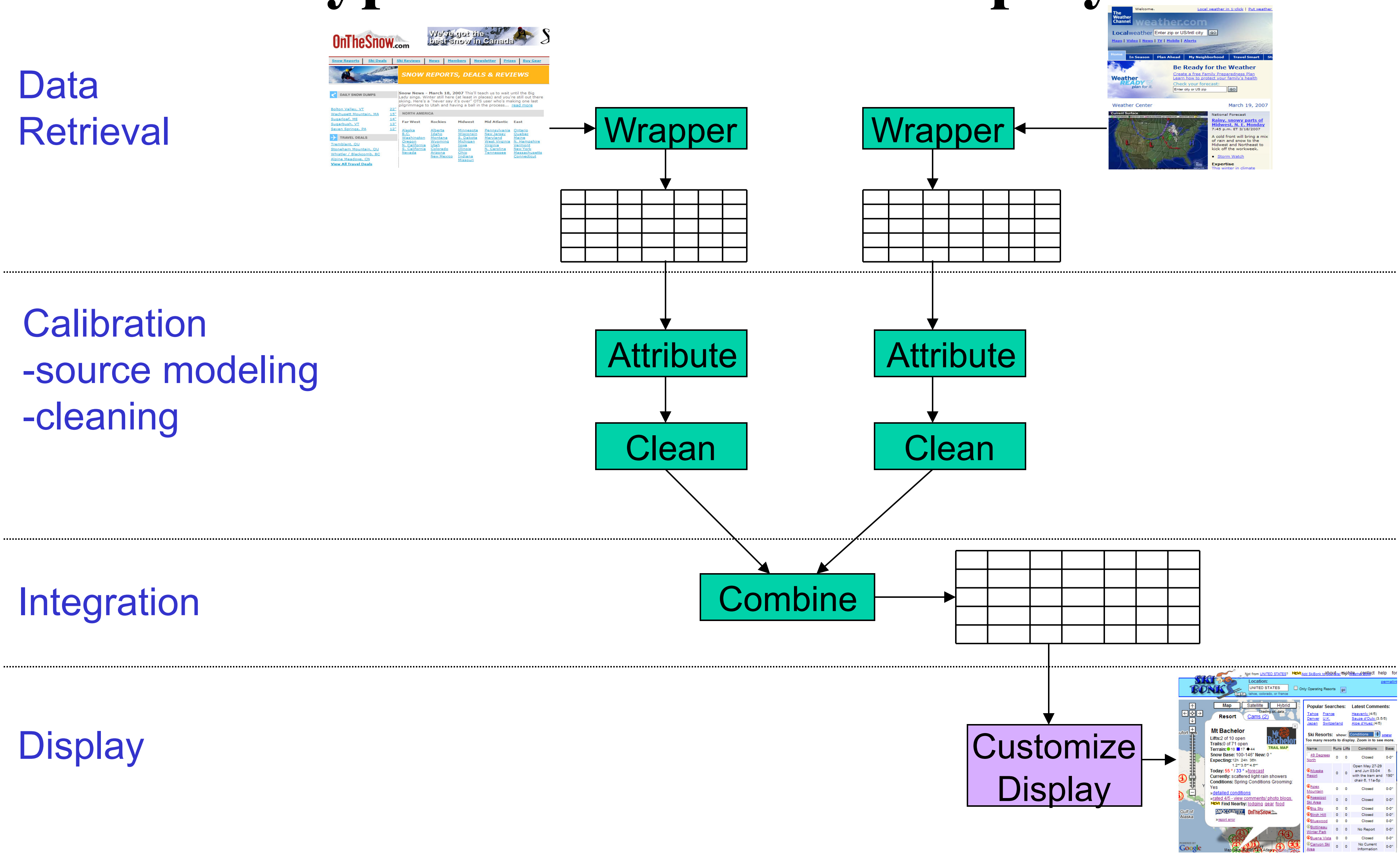
Calibration
-source modeling
-cleaning

Integration

Display



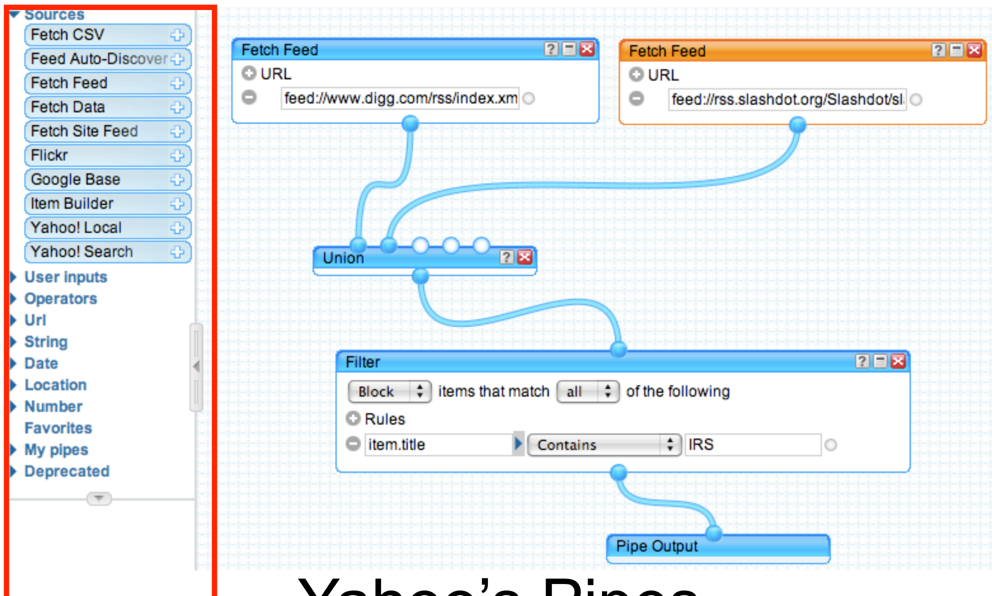
Type 5: Customized Display



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

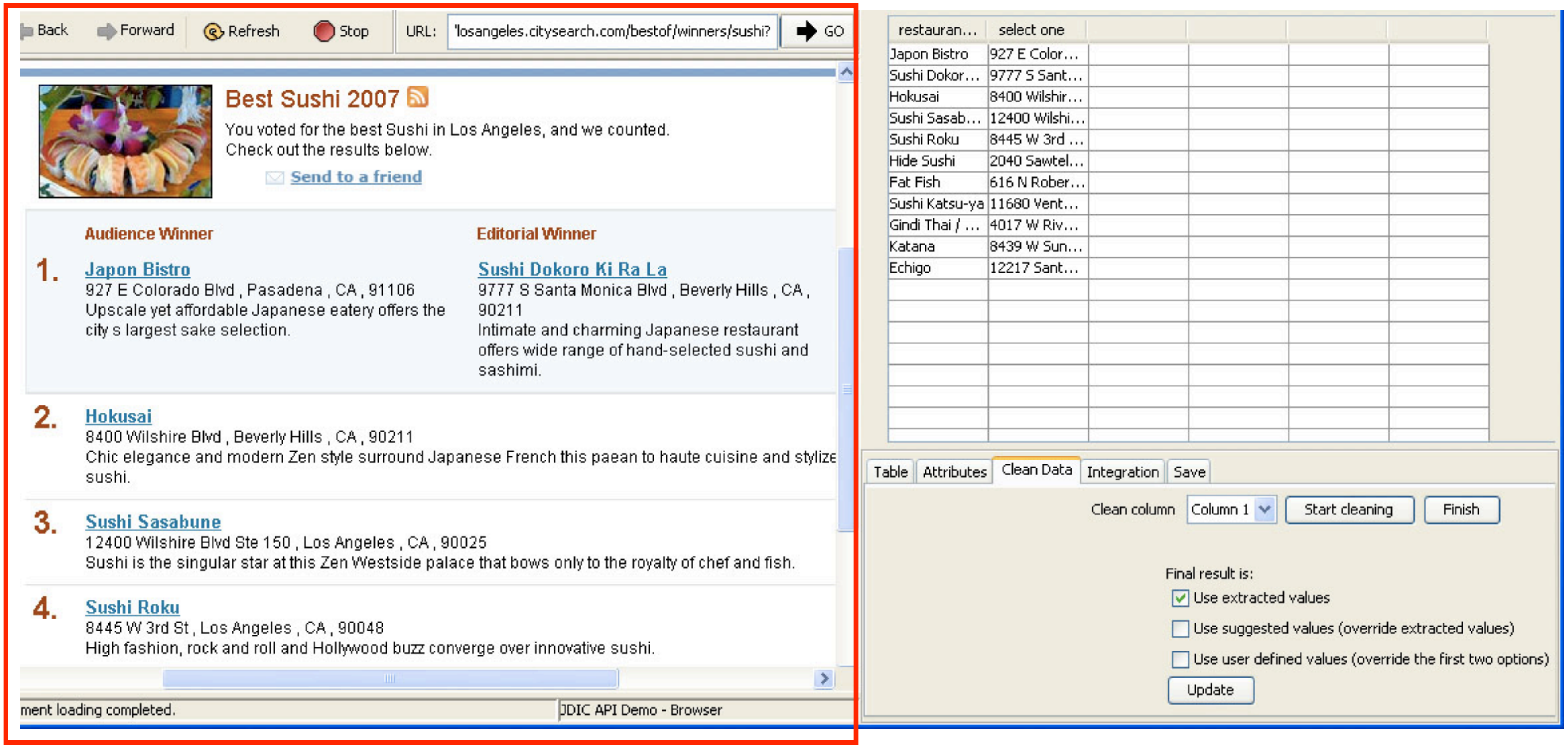
- A programming by demonstration approach that uses a single table for building a Mashup
- An integrated approach that links data extraction, source modeling, data cleaning, and data integration together.
- A query formulation technique that allows users to specify examples to build complicated queries.

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

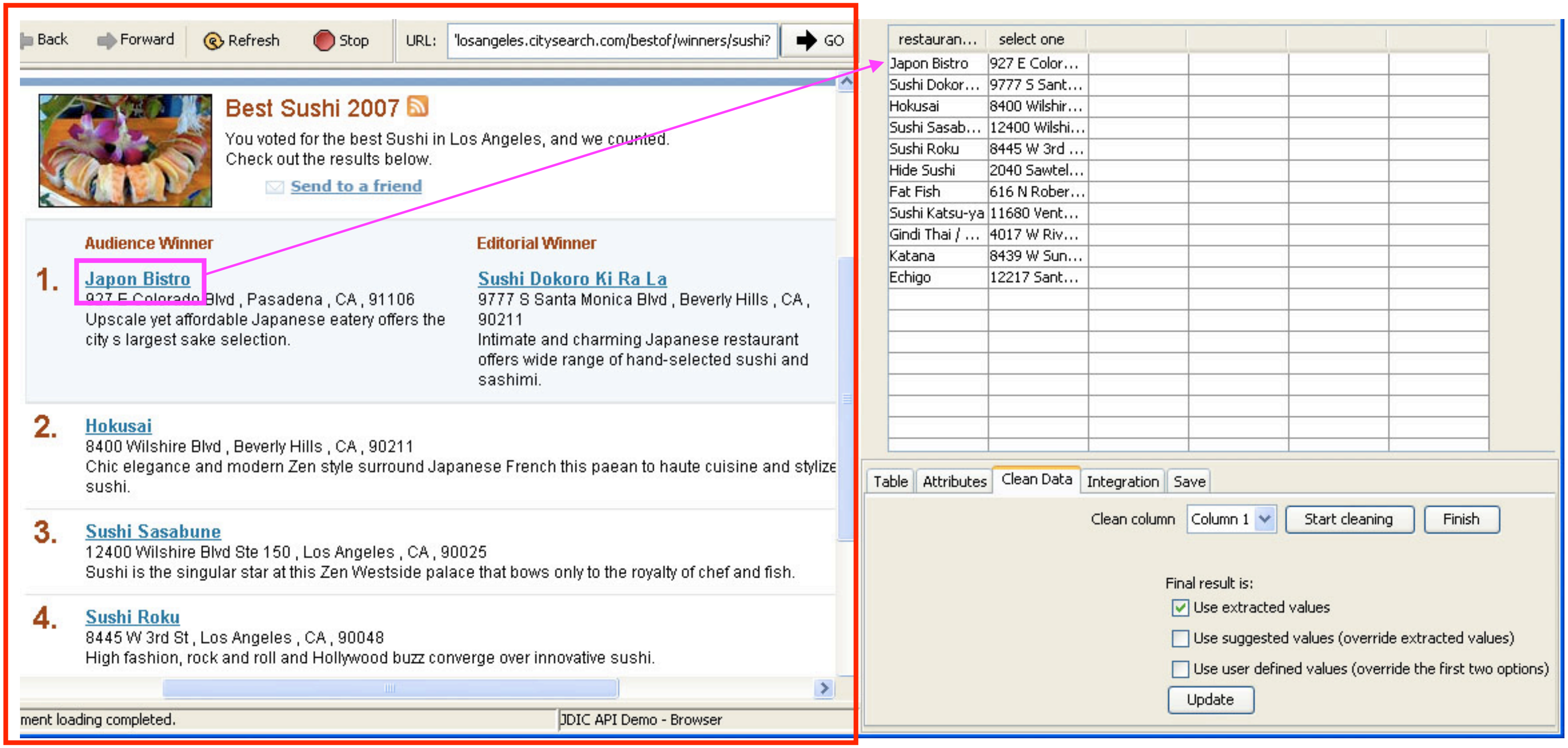
Our system: Karma

Embedded Browser



Our system: Karma

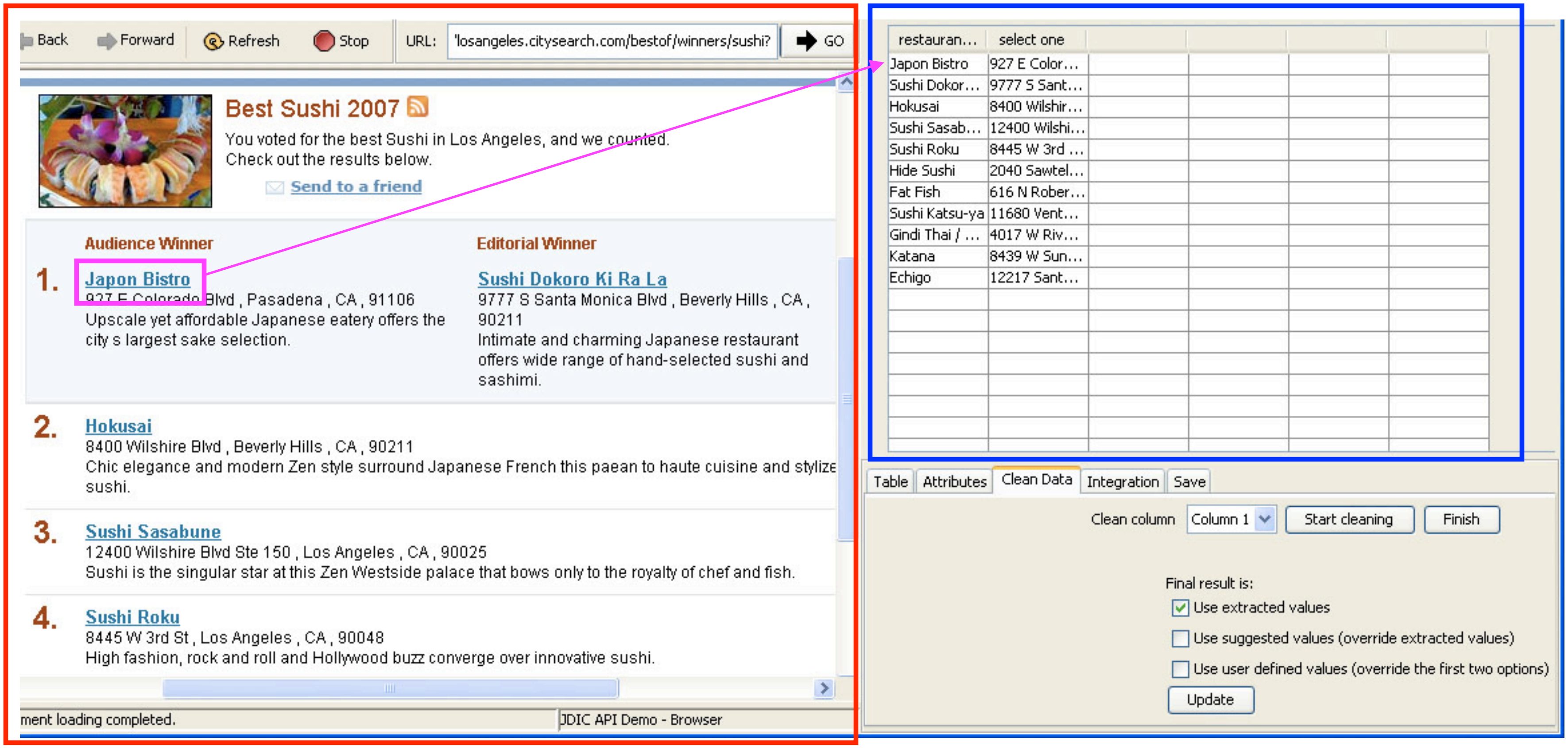
Embedded Browser



Our system: Karma

Embedded Browser

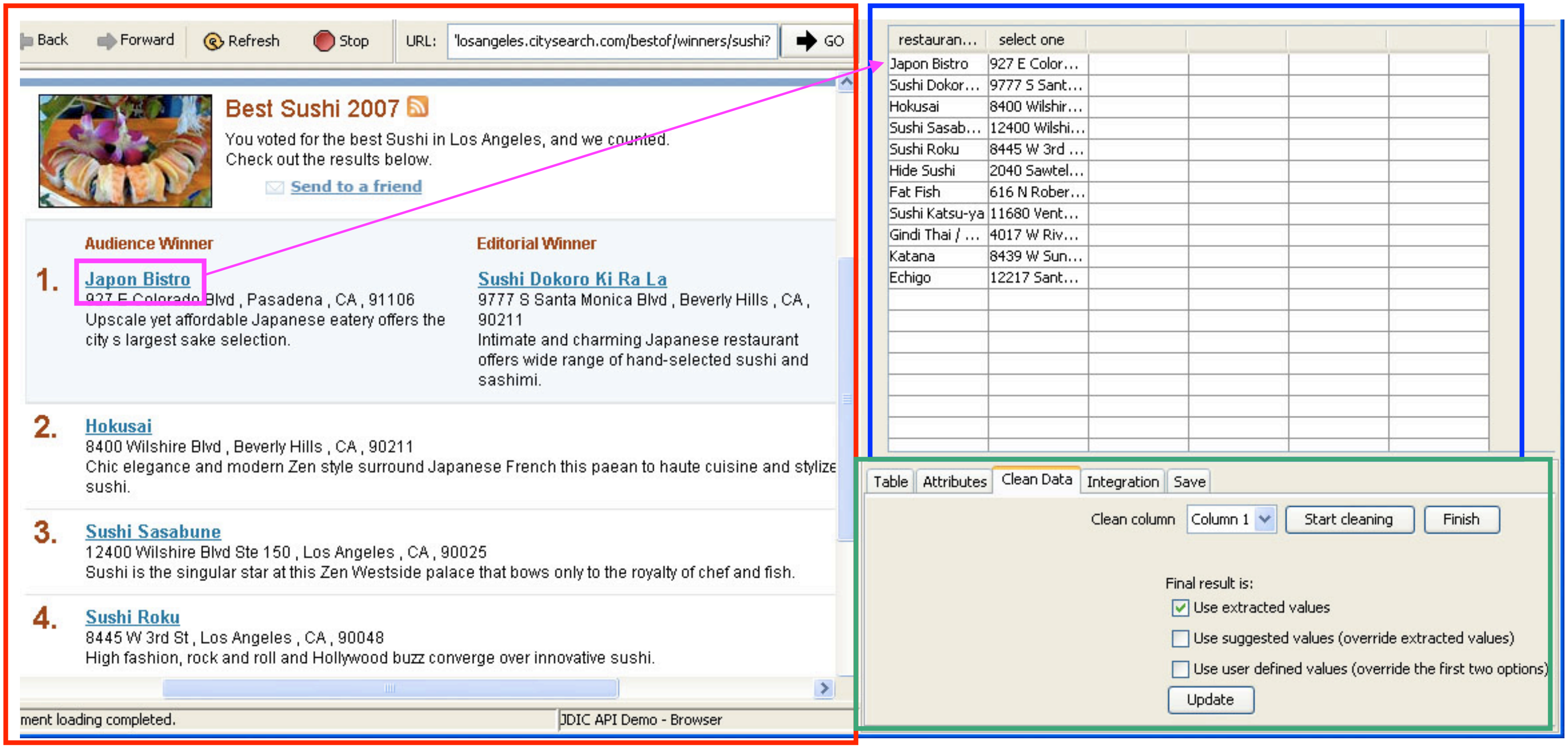
Table



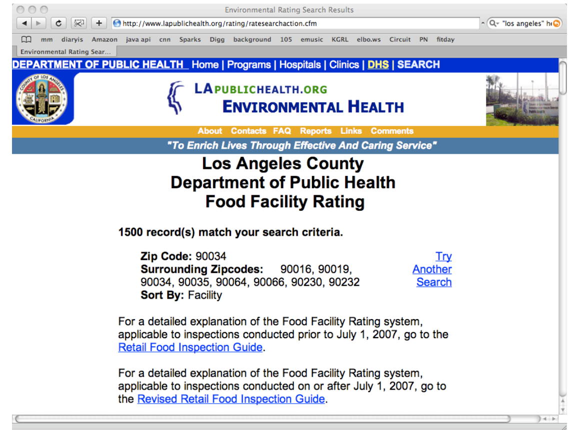
Our system: Karma

Embedded Browser

Table



Interaction Modes



Extract

Extract

{Restaurant name, address, phone, Review}

{Restaurant name, address, Date of Inspection, Score}

Clean

Clean

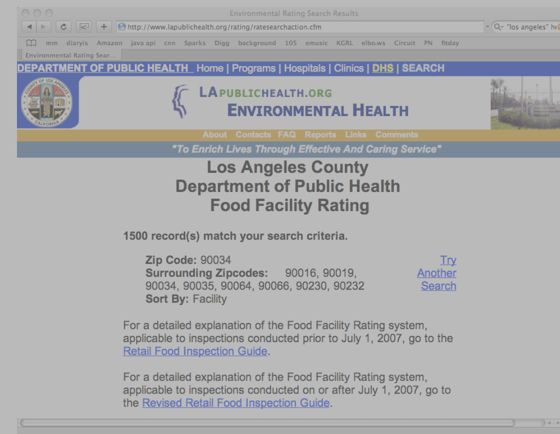
{Restaurant name, address, phone, review, Date of Inspection, Score}

Map

Extract

{Restaurant name, address, phone, Review}

Clean



Extract

{Restaurant name, address, Date of Inspection, Score}

Clean

{Restaurant name, address, phone, review, Date of Inspection, Score}

Map

Extract

{Restaurant name, address, phone, Review}

Clean

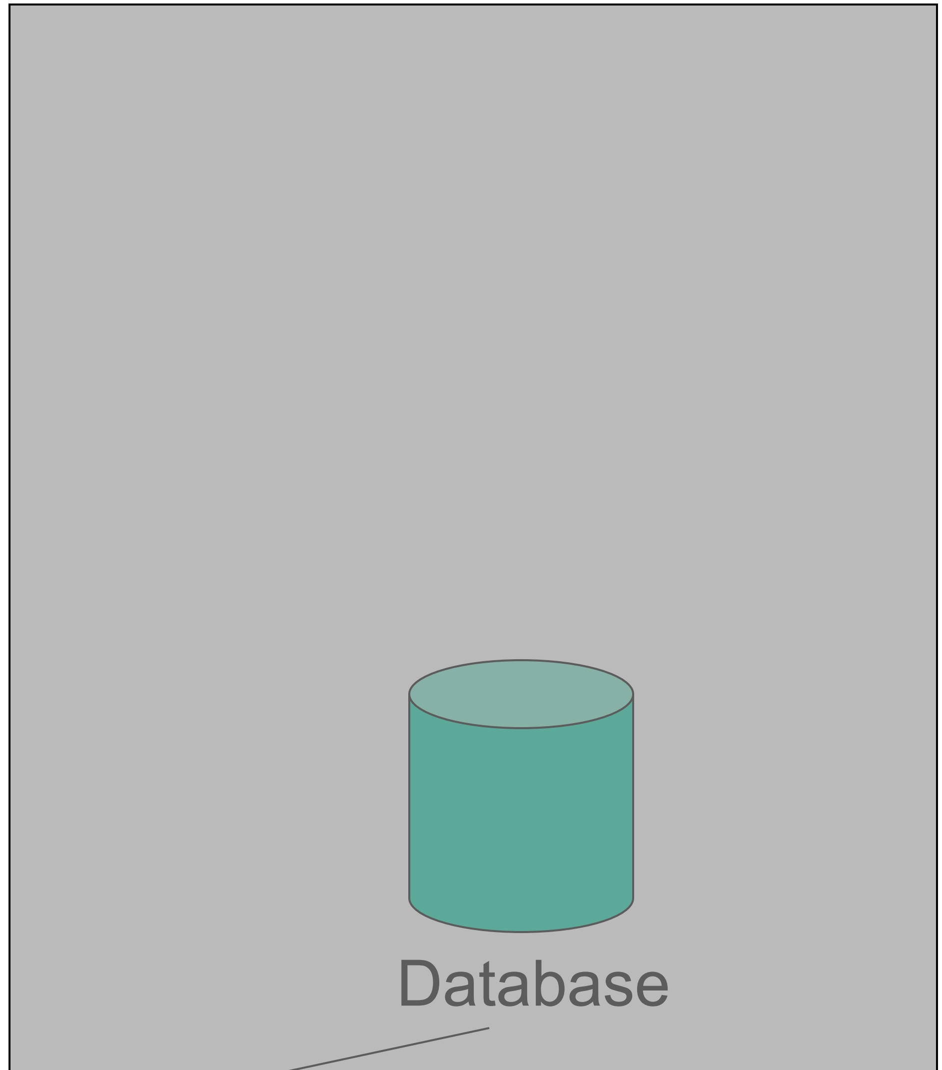
{Restaurant name, address, phone, review, Date of Inspection, Score}

Map

Extract

{Restaurant name, address, phone, Review}

Clean



{Restaurant name, address, phone, review, Date of Inspection, Score}

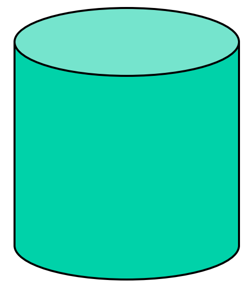
Map

Extract

{Restaurant name, address, phone, Review}

Clean

Database contains past Mashups and data tables

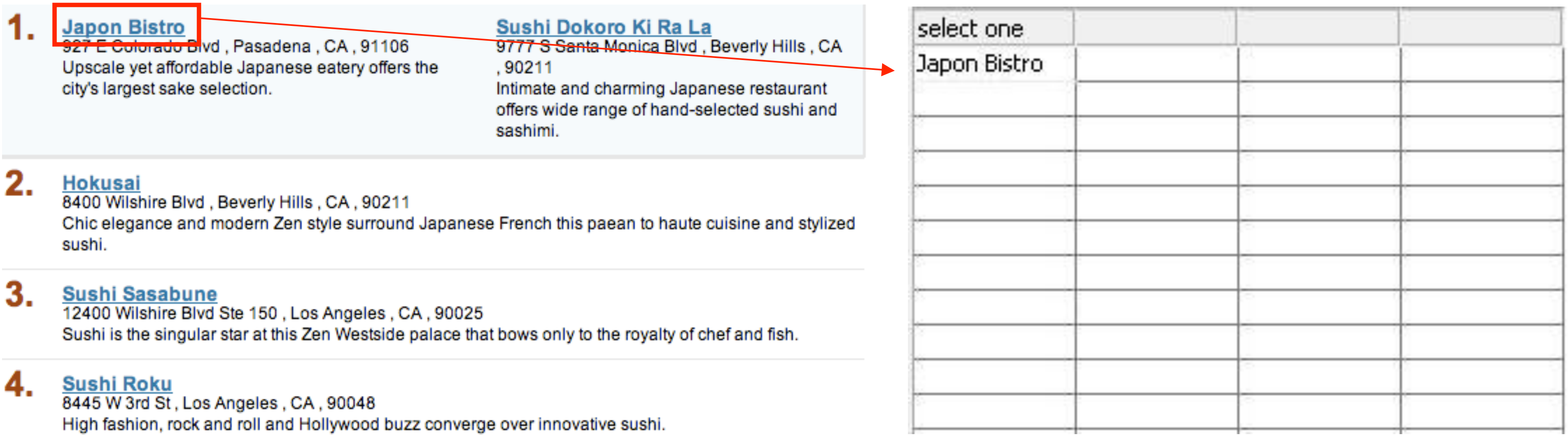


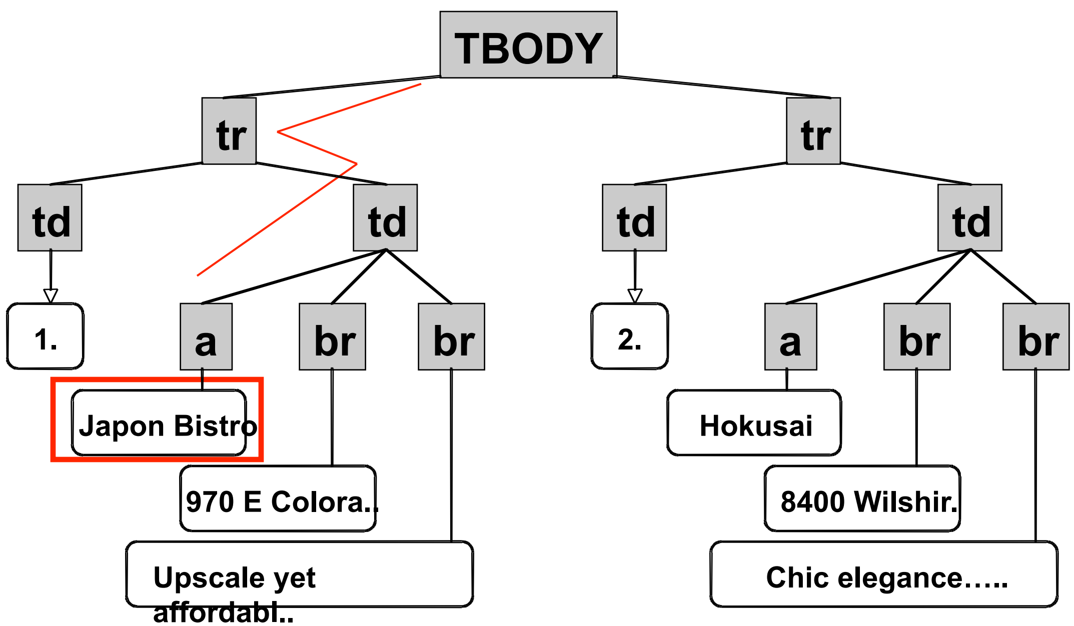
Database

{Restaurant name, address, phone, review, Date of Inspection, Score}

Map

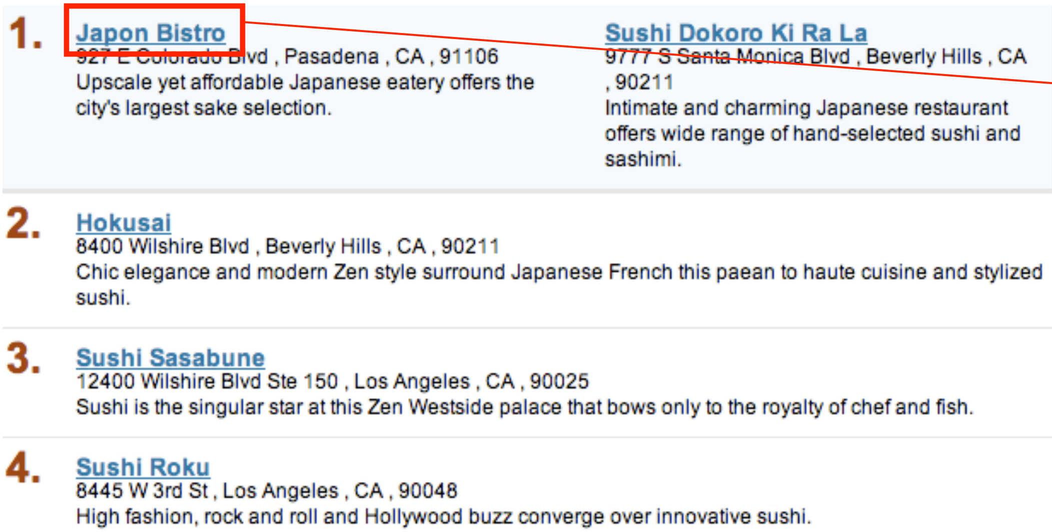
Data Retrieval: Extraction





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

Data Retrieval: Extraction



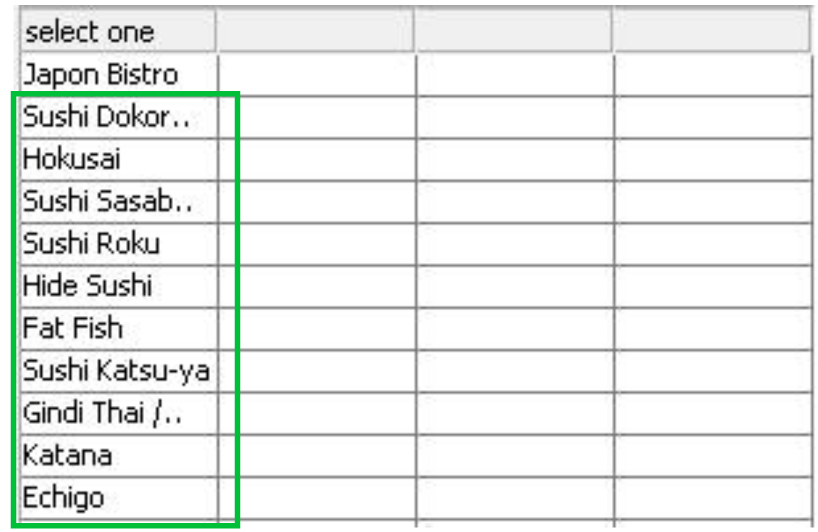
1. **Japon Bistro**
   927 E Colorado Blvd , Pasadena , CA , 91106
   Upscale yet affordable Japanese eatery offers the city's largest sake selection.

   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.

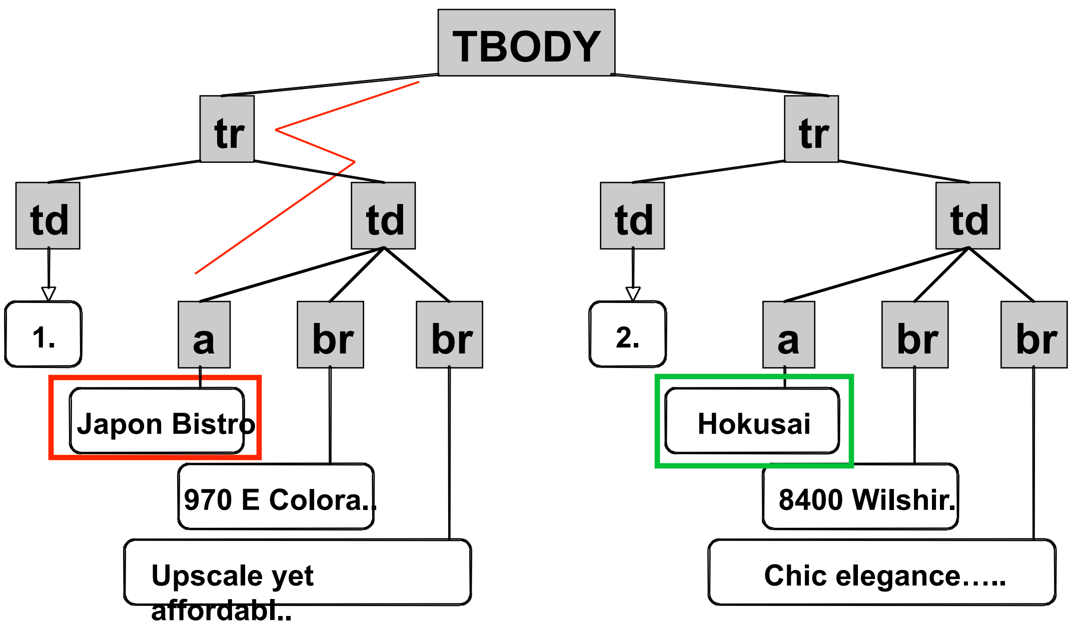
2. **Hokusai**
   8400 Wilshire Blvd , Beverly Hills , CA , 90211
   Chic elegance and modern Zen style surround Japanese French this paean to haute cuisine and stylized sushi.

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.



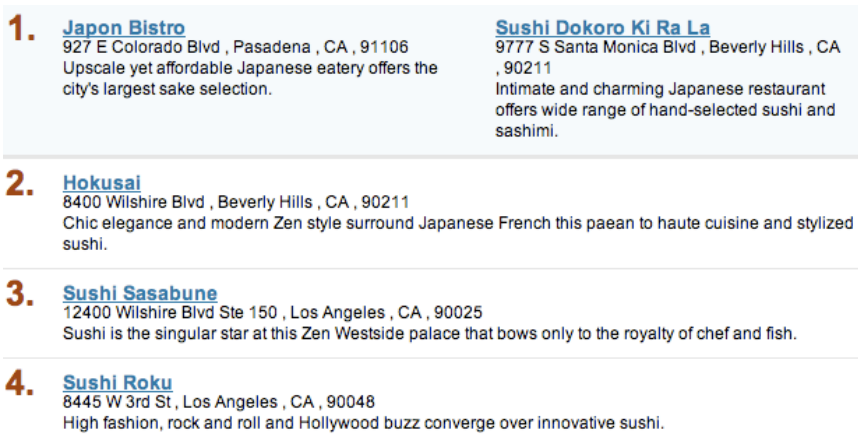
select one			
Japon Bistro			
Sushi Dokor..			
Hokusai			
Sushi Sasab..			
Sushi Roku			
Hide Sushi			
Fat Fish			
Sushi Katsu-ya			
Gindi Thai /..			
Katana			
Echigo			

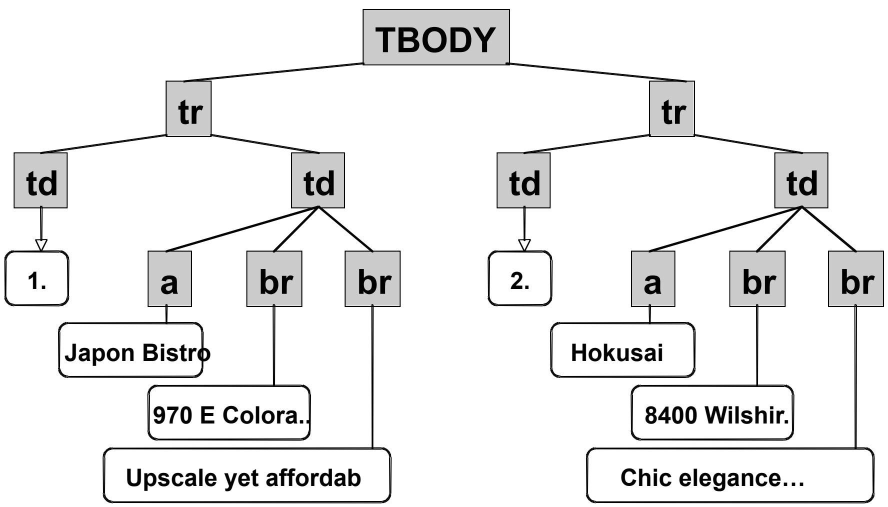


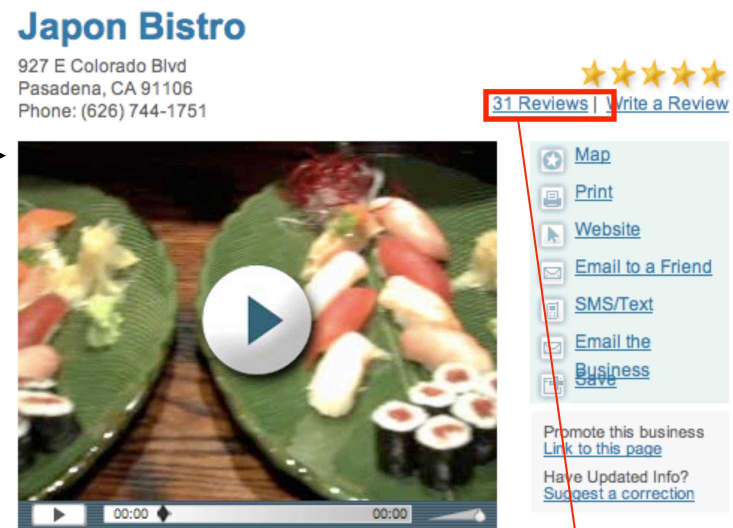
Tbody/tr[1]/td[2]/a

Tbody/tr*/td*/a

Data Retrieval: Navigation

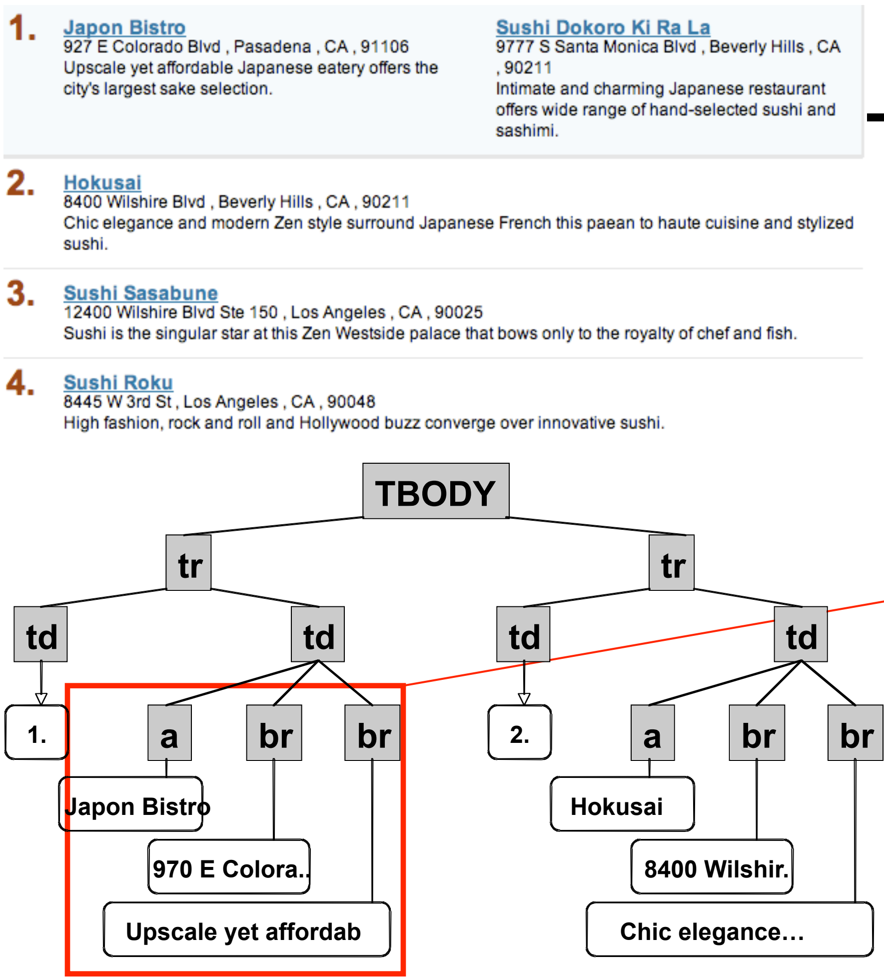


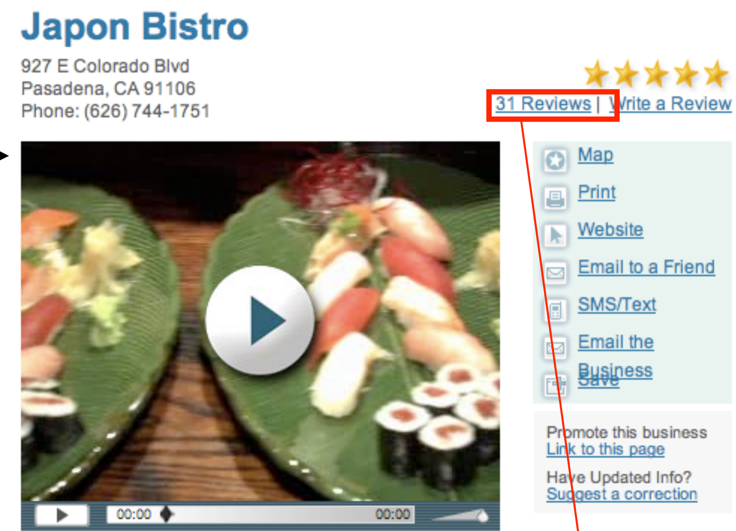




select one	address	select one	select one
Japon Bistro	927 E Color..	Upscale yet…	31 Reviews
Sushi Dokor..	9777 S Sant..	Intimate an…	
Hokusai	8400 Wilshir…	Chic eleganc…	
Sushi Sasab..	12400 Wilshi…	Authentic Ja…	
Sushi Roku	8445 W 3rd…	High fashion…	
Hide Sushi	2040 Sawtel…	No fuss, jus…	
Fat Fish	616 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…	

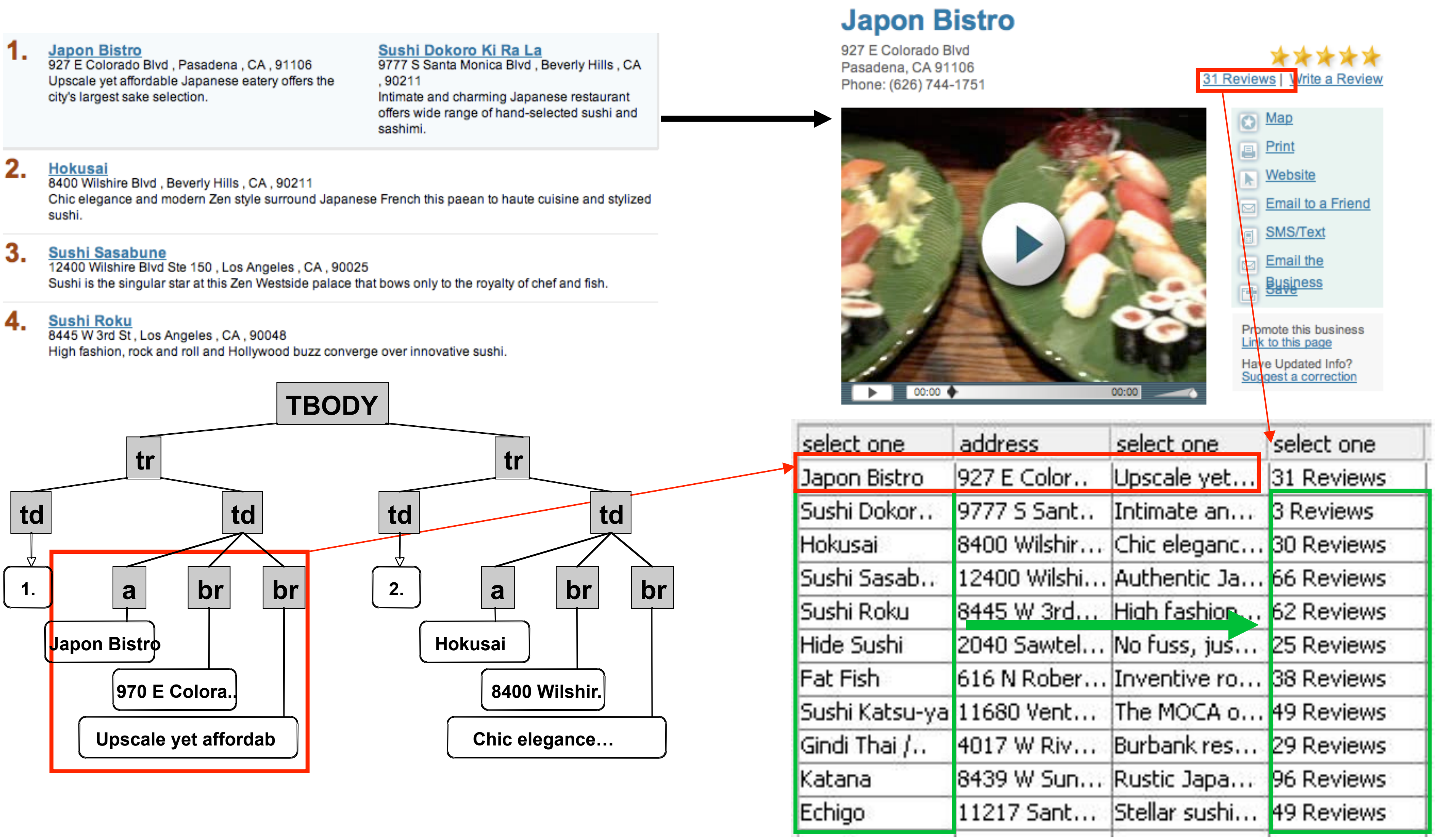
Data Retrieval: Navigation





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…	

Data Retrieval: Navigation



select one	address	select one	select one
Japon Bistro	927 E Color..	Upscale yet…	31 Reviews
Sushi Dokor..	9777 S Sant..	Intimate an…	3 Reviews
Hokusai	8400 Wilshir…	Chic eleganc…	30 Reviews
Sushi Sasab..	12400 Wilshi…	Authentic Ja…	66 Reviews
Sushi Roku	8445 W 3rd…	High fashion…	62 Reviews
Hide Sushi	2040 Sawtel…	No fuss, jus…	25 Reviews
Fat Fish	616 N Rober…	Inventive ro…	38 Reviews
Sushi Katsu-ya	11680 Vent…	The MOCA o…	49 Reviews
Gindi Thai /..	4017 W Riv…	Burbank res…	29 Reviews
Katana	8439 W Sun…	Rustic Japa…	96 Reviews
Echigo	11217 Sant…	Stellar sushi…	49 Reviews

Source Modeling (Attribute selection)

Newly extracted data

Japon Bistro
Hokusai
Sushi Sasabune
…

LA Health Rating

restaurant name	*Address*	..	*Health Rating*
Hokusai	8400..	..	90
Katana	8439..	..	99
Japon Bistro	927 E..	..	95

Artist Info

artist name	*nationality*	..	..
Hokusai	Japanese	..	..
Renoir	French	..	..
..	..	..	..

Zagat

restaurant name	*zagat Rating*	..	..
Sushi Sasabune	27	..	..
Sushi Roku	25	..	..
Katana	23	..	..

Database

Possible Attribute

$\{a \mid a,s: a \in att(s) \wedge (val(a,s) \subset V)\}$

restaurant name (3)
artist name (1)

Data Cleaning: using existing values

Data repository

Newly extracted data

Japon Bistro
Hokusai
Sushi Sasabune
Sushi Roka

LA Health Rating

restaurant name	*Address*	..	*Health Rating*
Hokusai	8400..	..	90
Katana	8439..	..	99
Japon Bistro	927 E..	..	95

Zagat

restaurant name	*zagat Rating*	..	..
Sushi Sasabune	27	..	..
Sushi Roku	25	..	..
Katana	23	..	..

Restaurant name

Data Cleaning: using existing values

Newly extracted data

Japon Bistro
Hokusai
Sushi Sasabune
Sushi Roka

Restaurant name

Data repository

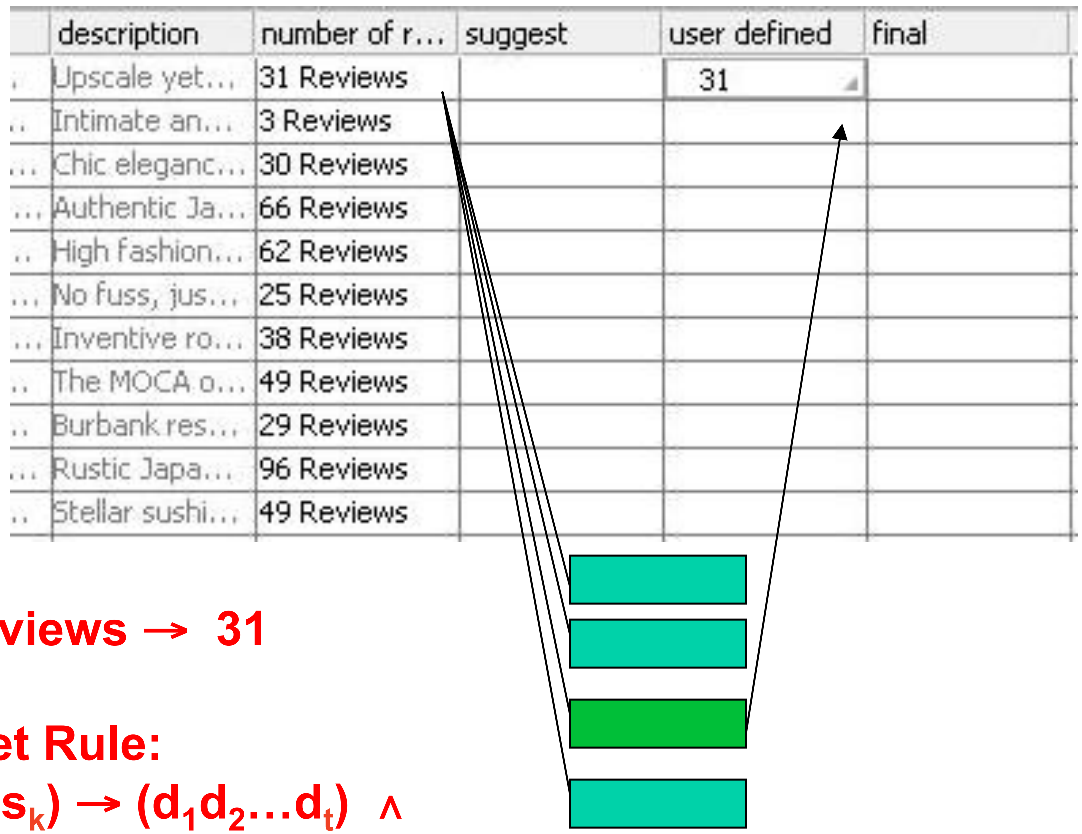
LA Health Rating

restaurant name	*Address*	..	*Health Rating*
Hokusai	8400..	..	90
Katana	8439..	..	99
Japon Bistro	927 E..	..	95

Zagat

restaurant name	*zagat Rating*	..	..
Sushi Sasabune	27	..	..
Sushi Roku	25	..	..
Katana	23	..	..

Data Cleaning: using predefined rules



	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:

$(s_1 s_2 .. s_k) \rightarrow (d_1 d_2 \ldots d_t) \wedge$

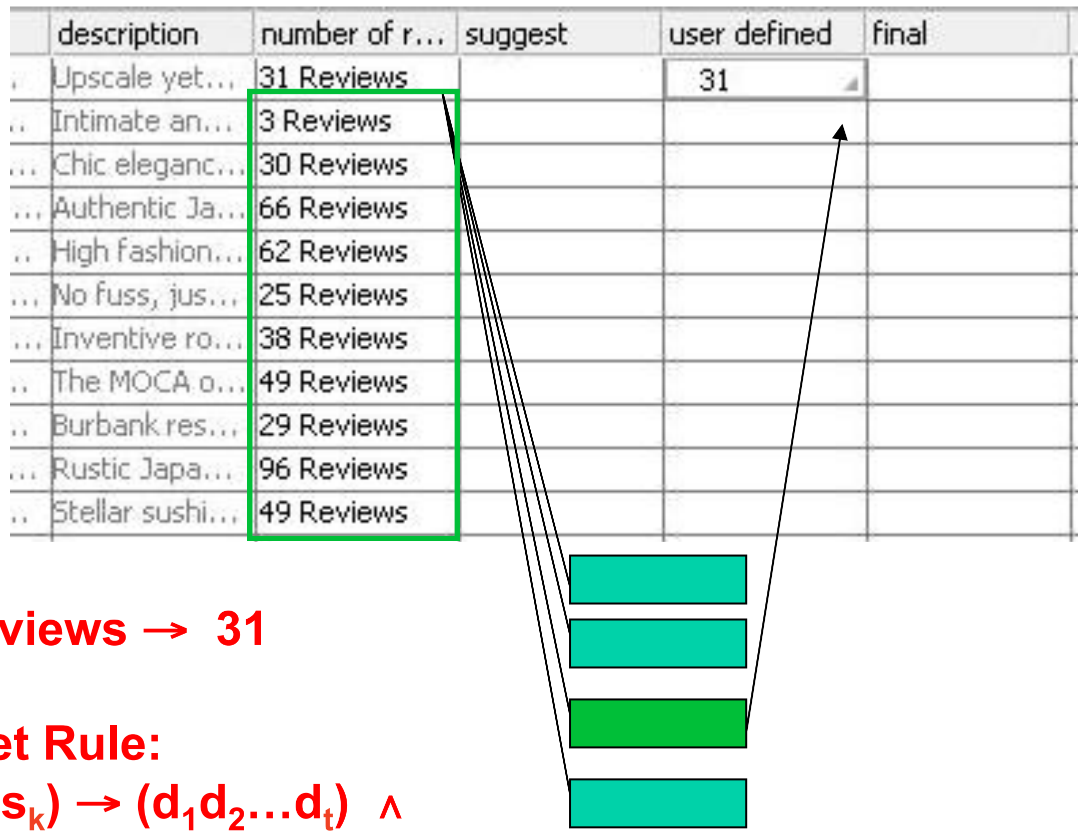
$(k <= t) \wedge$

$s_i \in \{d_1, d_2, \ldots, d_t\} \wedge$

$d_i \neq d_j$

Predefined Rules

Data Cleaning: using predefined rules



31 Reviews → 31

Subset Rule:

$(s_1 s_2 .. s_k) \rightarrow (d_1 d_2 \ldots d_t) \wedge$

$(k <= t) \wedge$

$s_i \in \{d_1, d_2, \ldots, d_t\} \wedge$

$d_i \neq d_j$

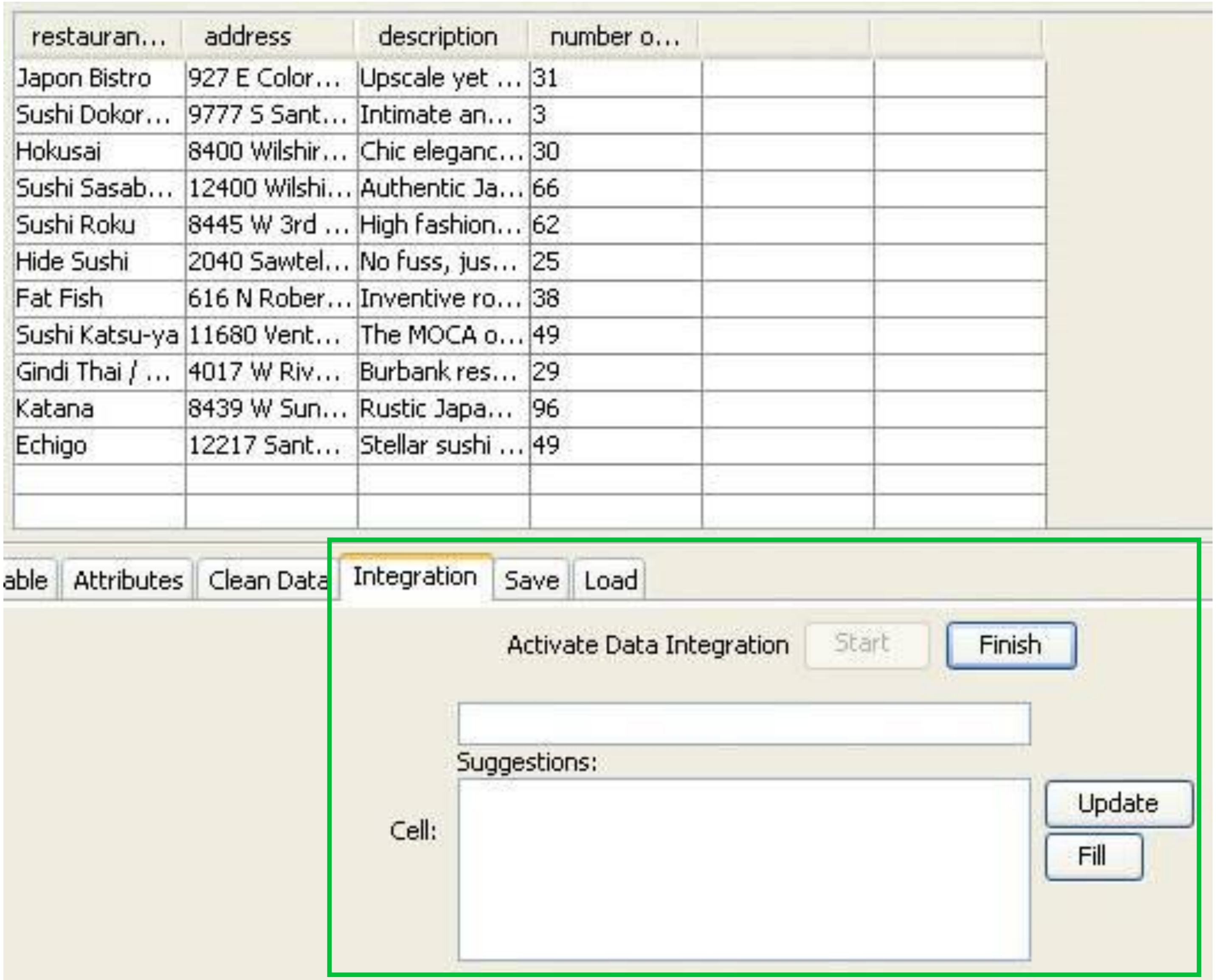
.

.

.

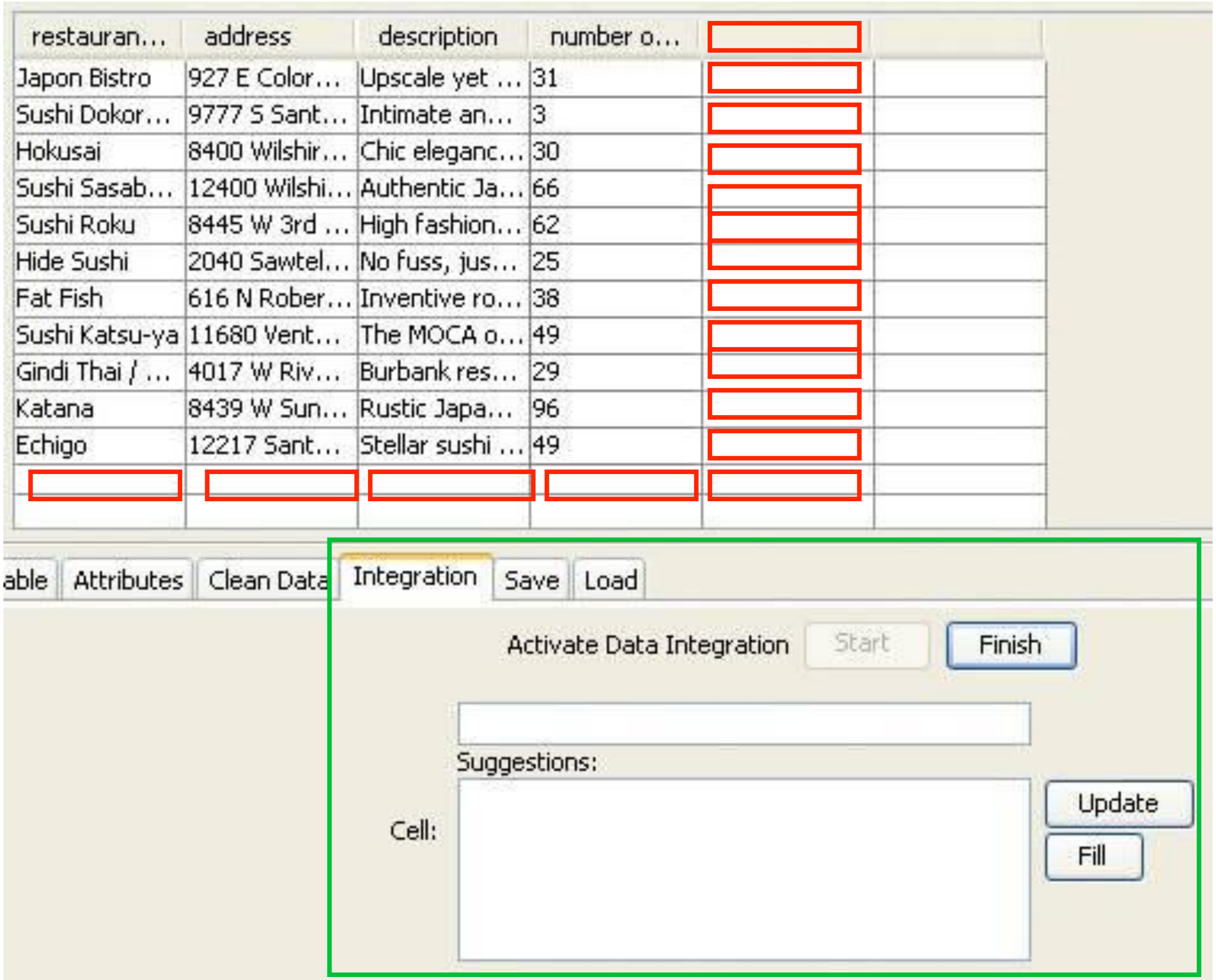
Predefined Rules

Data Integration [tuchinda 2007]

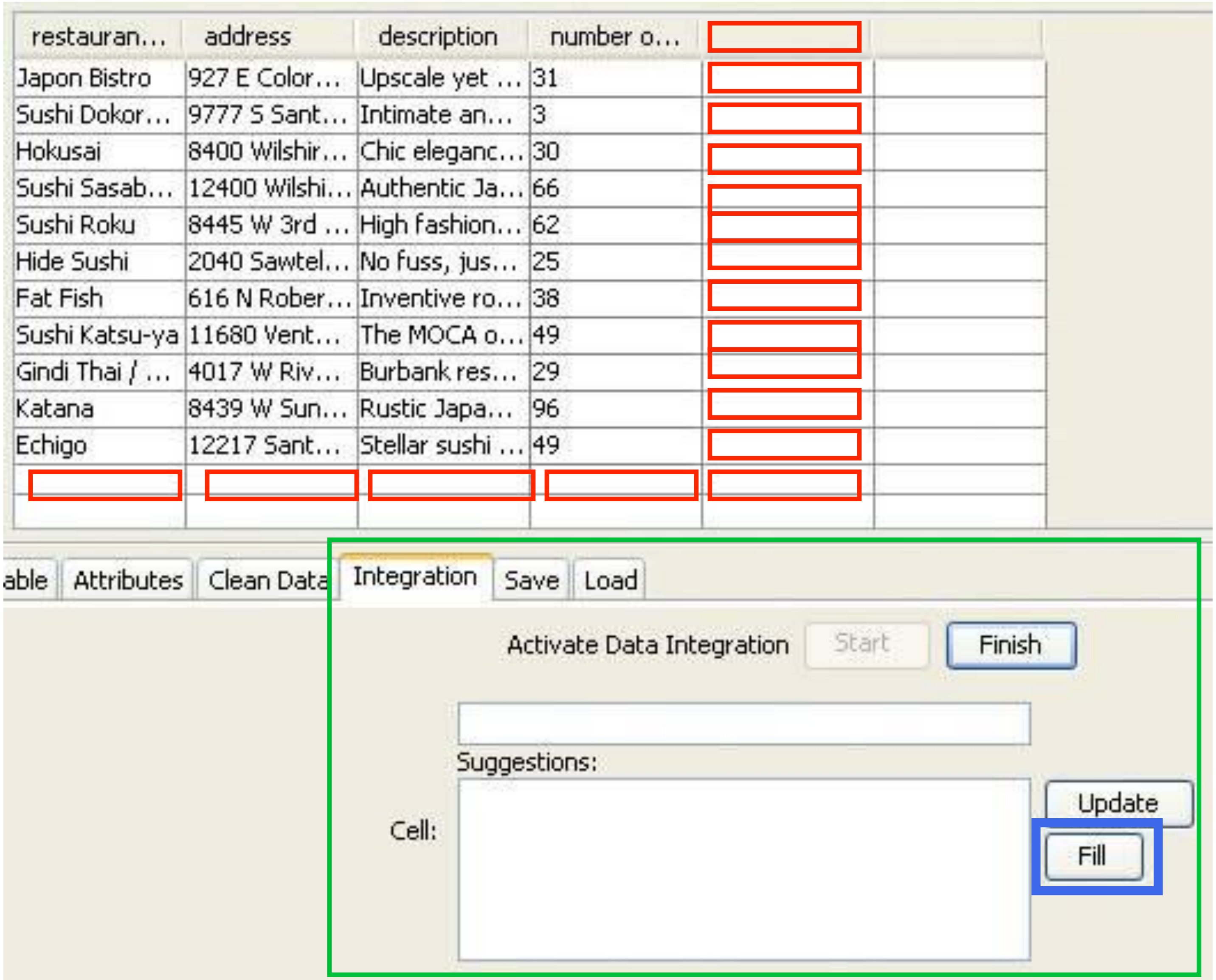


restauran...	address	description	number o...		
Japon Bistro	927 E Color...	Upscale yet ...	31		
Sushi Dokor...	9777 S Sant...	Intimate an...	3		
Hokusai	8400 Wilshir...	Chic eleganc...	30		
Sushi Sasab...	12400 Wilshi...	Authentic Ja...	66		
Sushi Roku	8445 W 3rd ...	High fashion...	62		
Hide Sushi	2040 Sawtel...	No fuss, jus...	25		
Fat Fish	616 N Rober...	Inventive ro...	38		
Sushi Katsu-ya	11680 Vent...	The MOCA o...	49		
Gindi Thai / ...	4017 W Riv...	Burbank res...	29		
Katana	8439 W Sun...	Rustic Japa...	96		
Echigo	12217 Sant...	Stellar sushi ...	49		

Data Integration [tuchinda 2007]



Data Integration [tuchinda 2007]



restauran...	address	description	number o...	
Japon Bistro	927 E Color...	Upscale yet ...	31	
Sushi Dokor...	9777 S Sant...	Intimate an...	3	
Hokusai	8400 Wilshir...	Chic eleganc...	30	
Sushi Sasab...	12400 Wilshi...	Authentic Ja...	66	
Sushi Roku	8445 W 3rd ...	High fashion...	62	
Hide Sushi	2040 Sawtel...	No fuss, jus...	25	
Fat Fish	616 N Rober...	Inventive ro...	38	
Sushi Katsu-ya	11680 Vent...	The MOCA o...	49	
Gindi Thai / ...	4017 W Riv...	Burbank res...	29	
Katana	8439 W Sun...	Rustic Japa...	96	
Echigo	12217 Sant...	Stellar sushi ...	49	

Data Integration (cont.)

restaurant ...	address	description	number of r...	
Japon Bistro	927 E Color..	Upscale yet...	31	
Sushi Dokor..	9777 S Sant..	Intimate an...	3	
Hokusai	8400 Wilshir...	Chic eleganc...	30	
Sushi Sasab..	12400 Wilshi...	Authentic Ja...	66	
Sushi Roku	8445 W 3rd...	High fashion...	62	
Hide Sushi	2040 Sawtel...	No fuss, jus...	25	
Fat Fish	616 N Rober...	Inventive ro...	38	
Sushi Katsu-ya	11680 Vent...	The MOCA o...	49	
Gindi Thai /..	4017 W Riv...	Burbank res...	29	
Katana	8439 W Sun...	Rustic Japa...	96	
Echigo	11217 Sant...	Stellar sushi...	49	

Data repository

LA Health Rating

restaurant name	***Address***	**..**	***Health Rating***
Hokusai	**8400..**	**..**	**90**
Katana	**8439..**	**..**	**99**
Japon Bistro	**927 E..**	**..**	**95**

Zagat

restaurant name	***zagat Rating***	**..**	**..**
Sushi Sasabune	**27**	**..**	**..**
Sushi Roku	**25**	**..**	**..**
Katana	**23**	**..**	**..**

Data Integration (cont.)

restaurant ...	address	description	number of r...	
Japon Bistro	927 E Color..	Upscale yet...	31	
Sushi Dokor..	9777 S Sant..	Intimate an...	3	
Hokusai	8400 Wilshir...	Chic eleganc...	30	
Sushi Sasab..	12400 Wilshi...	Authentic Ja...	66	
Sushi Roku	8445 W 3rd...	High fashion...	62	
Hide Sushi	2040 Sawtel...	No fuss, jus...	25	
Fat Fish	616 N Rober...	Inventive ro...	38	
Sushi Katsu-ya	11680 Vent...	The MOCA o...	49	
Gindi Thai /..	4017 W Riv...	Burbank res...	29	
Katana	8439 W Sun...	Rustic Japa...	96	
Echigo	11217 Sant...	Stellar sushi...	49	

Data repository

LA Health Rating

restaurant name	*Address*	..	*Health Rating*
Hokusai	8400..	..	90
Katana	8439..	..	99
Japon Bistro	927 E..	..	95

Zagat

restaurant name	*zagat Rating*	..	..
Sushi Sasabune	27	..	..
Sushi Roku	25	..	..
Katana	23	..	..

Data Integration (cont.)

restaurant …	address	description	number of r…	
Japon Bistro	927 E Color..	Upscale yet…	31	
Sushi Dokor..	9777 S Sant..	Intimate an…	3	
Hokusai	8400 Wilshir…	Chic eleganc…	30	
Sushi Sasab..	12400 Wilshi…	Authentic Ja…	66	
Sushi Roku	8445 W 3rd…	High fashion…	62	
Hide Sushi	2040 Sawtel…	No fuss, jus…	25	
Fat Fish	616 N Rober…	Inventive ro…	38	
Sushi Katsu-ya	11680 Vent…	The MOCA o…	49	
Gindi Thai /..	4017 W Riv…	Burbank res…	29	
Katana	8439 W Sun…	Rustic Japa…	96	
Echigo	11217 Sant…	Stellar sushi…	49	

Data repository

LA Health Rating

restaurant name	*Address*	..	*Health Rating*
Hokusai	8400..	..	90
Katana	8439..	..	99
Japon Bistro	927 E..	..	95

Zagat

restaurant name	*zagat Rating*	..	..
Sushi Sasabune	27	..	..
Sushi Roku	25	..	..
Katana	23	..	..

Data Integration (cont.)

restaurant ...	address	description	number of r...	
Japon Bistro	927 E Color..	Upscale yet...	31	
Sushi Dokor..	9777 S Sant..	Intimate an...	3	
Hokusai	8400 Wilshir...	Chic eleganc...	30	
Sushi Sasab..	12400 Wilshi...	Authentic Ja...	66	
Sushi Roku	8445 W 3rd...	High fashion...	62	
Hide Sushi	2040 Sawtel...	No fuss, jus...	25	
Fat Fish	616 N Rober...	Inventive ro...	38	
Sushi Katsu-ya	11680 Vent...	The MOCA o...	49	
Gindi Thai /..	4017 W Riv...	Burbank res...	29	
Katana	8439 W Sun...	Rustic Japa...	96	
Echigo	11217 Sant...	Stellar sushi...	49	

Data repository

LA Health Rating

restaurant name	*Address*	..	*Health Rating*
Hokusai	8400..	..	90
Katana	8439..	..	99
Japon Bistro	927 E..	..	95

Zagat

restaurant name	*zagat Rating*	..	..
Sushi Sasabune	27	..	..
Sushi Roku	25	..	..
Katana	23	..	..

$\{a\}_R$ = possible new attribute selection for row *i*.

$\{x\}$ = Set intersection($\{a\}$) over all the value rows.

$\{v\}$ = val(a,s) where a $\{x\}$

s is any source where *att(s)* $\{x\} \neq \{\}$

Single Column Example

City
Los Angeles
Honolulu

a є all attribute

v є all value

Attribute, Source

Los Angeles : {(**city**, tax_properties), ~~(**song name**, pop_music)~~}

v є all values Λ attributeOf(v) є {city, song name}

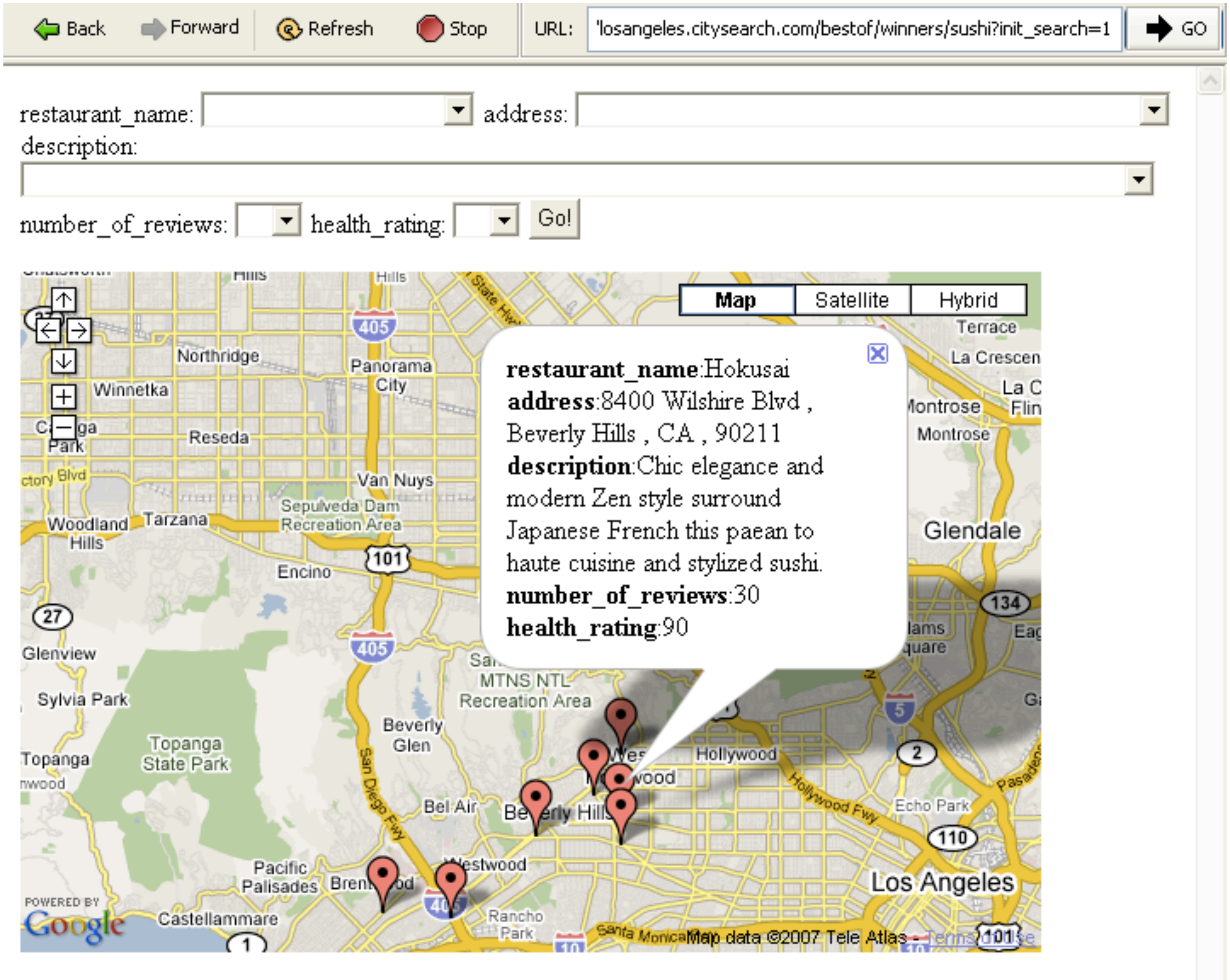
Can we determine the attribute now? Yes

Honolulu : {(**city**, tax_properties), (**city**, favorite_vacation_spot)}

{x} = Set intersection({a}) over all the value rows.

{v}* = val(*a,s*) where *a ε {x} Λ s* is any source where *att(s)* ∩ *{x} ≠ {}

Map Generation



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

Familiarization
-Programmers
(2 assignments on DP)
-Review Package
-30 minutes tutorial

Practice
-2-3 tasks using Karma

Test (3 tasks)
-Programmers: Alternating between Karma vs. DP for each task
-Non Programmers: use only Karma
-Screen are recorded using video capture software

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 Claim 2 Claim 3

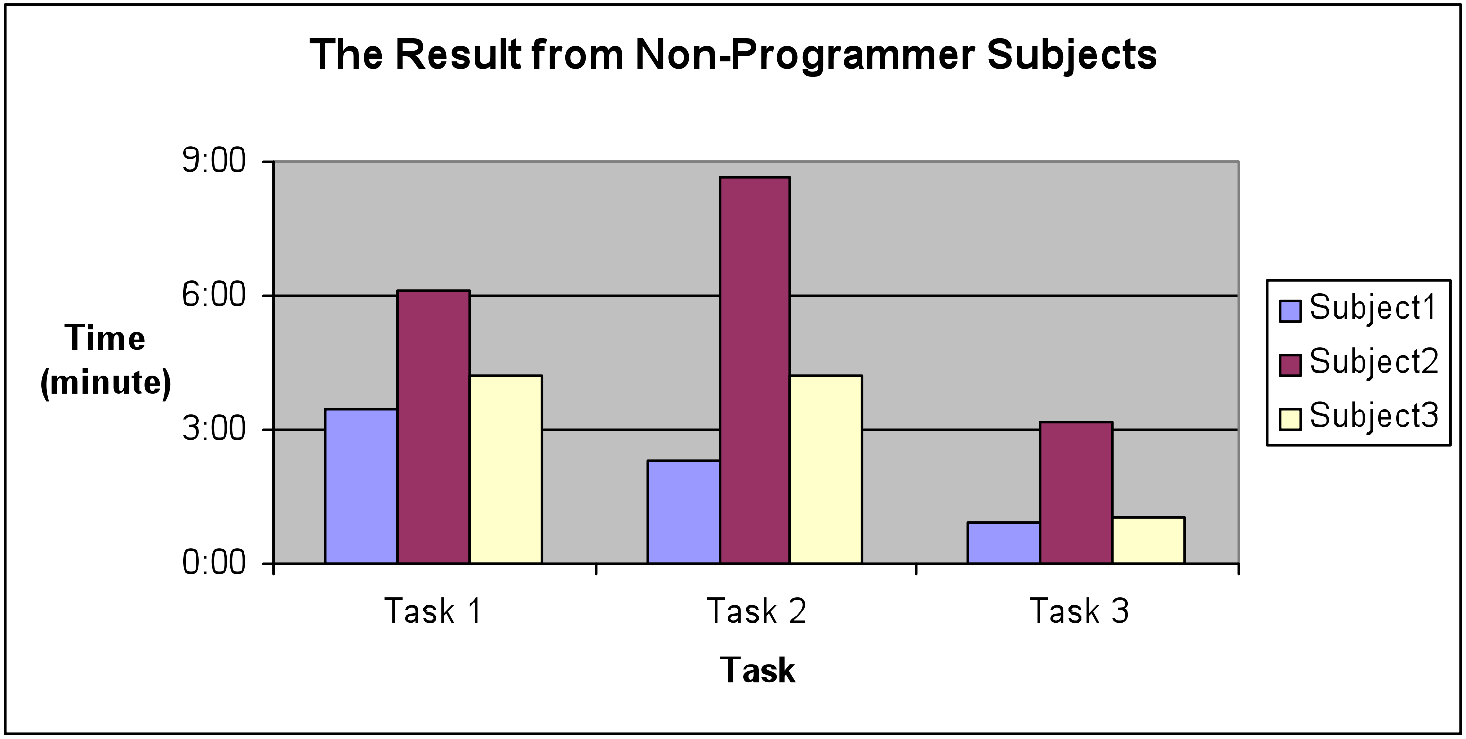
Users with no programming experiences can build all four Mashup types.

Karma takes less time to complete that subtask and scales better as the get harder

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

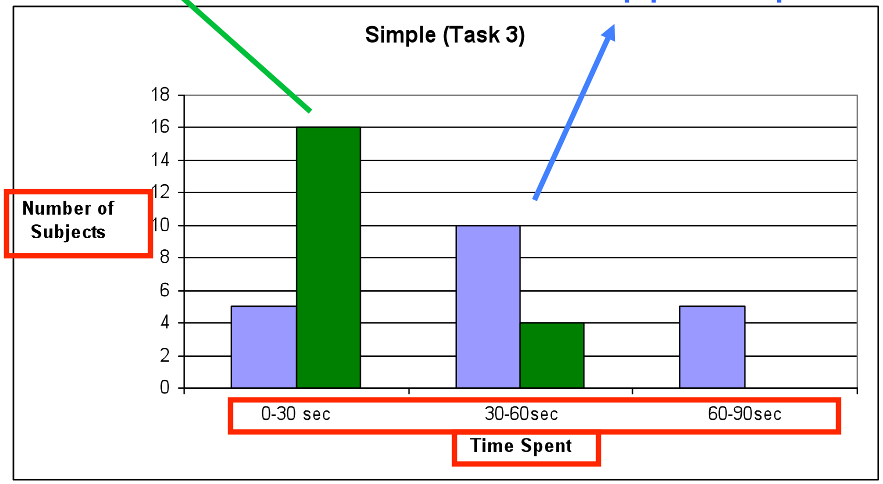


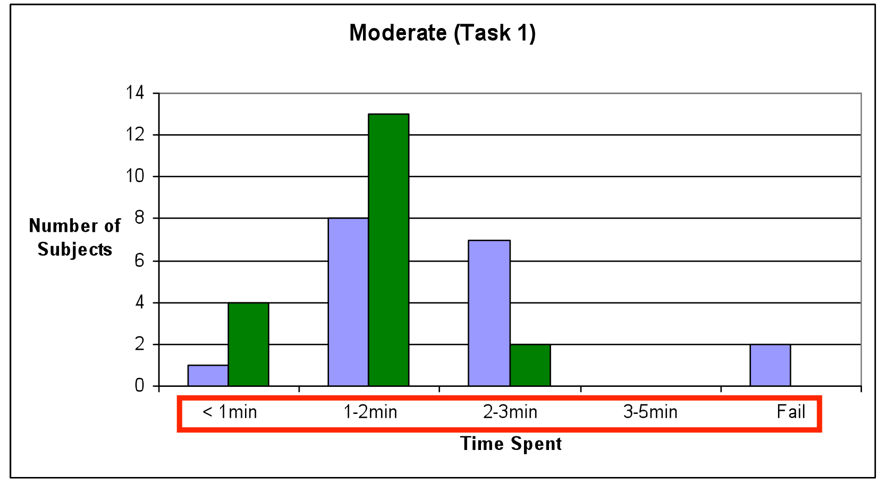
Claim 2: Karma takes less time to complete each subtask

Evaluation: Extraction

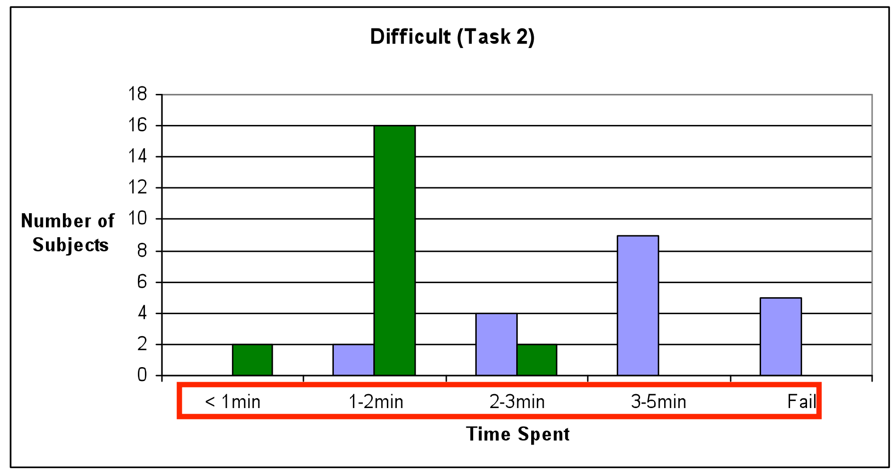
Karma (programmer)

Dapper/Pipes



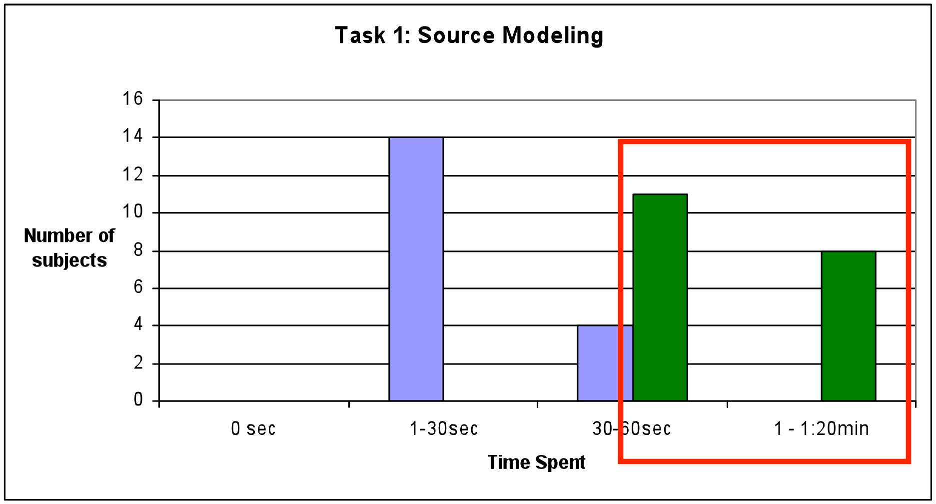


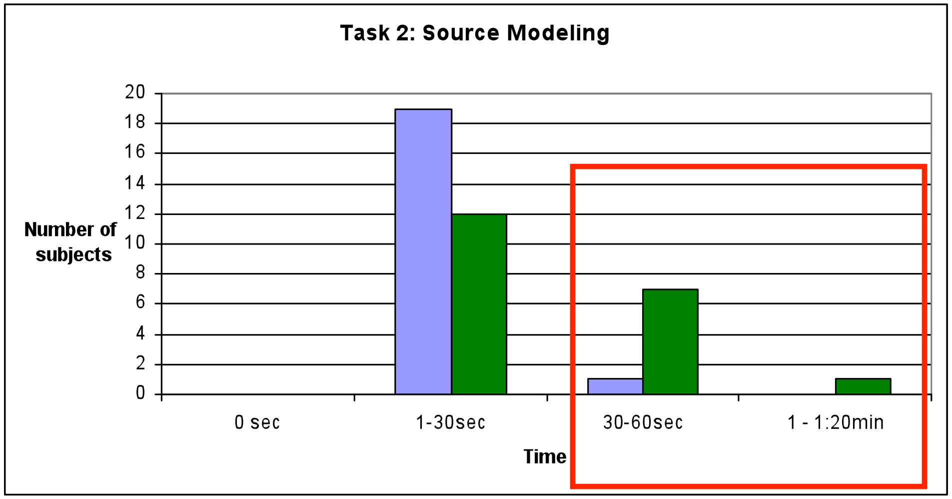
- As the extraction task gets more difficult, Dapper/Pipes takes
  - longer
  - more subjects failing to complete the task (11% for moderate and 25% for difficult)



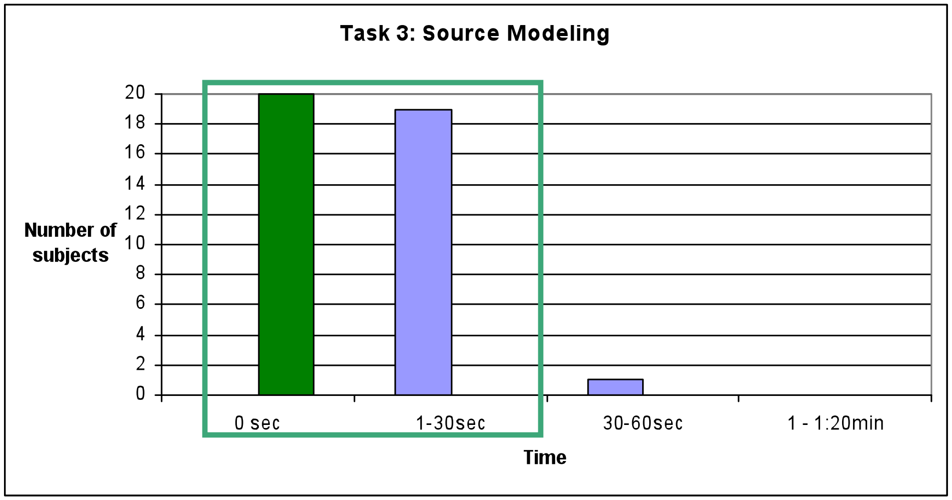
Dapper/Pipes **Karma**

Evaluation: Source Modeling



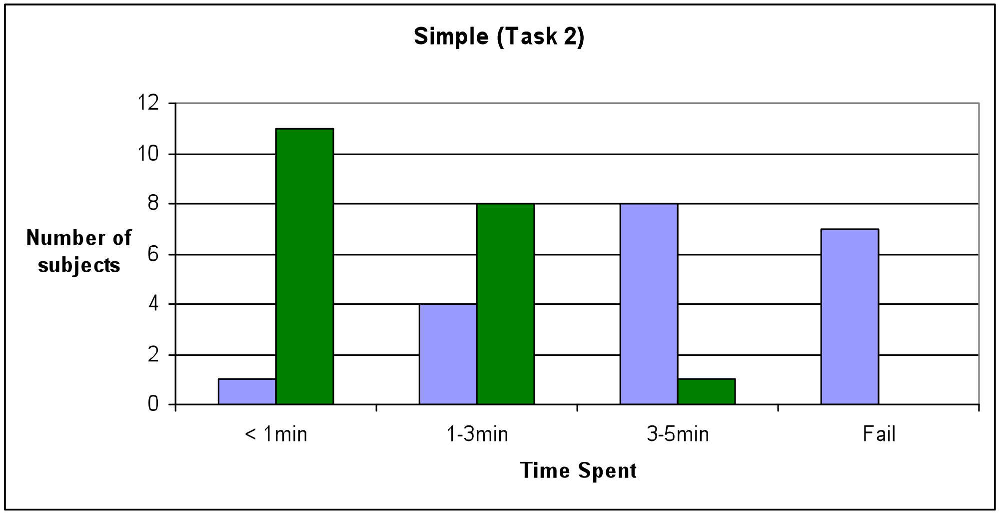


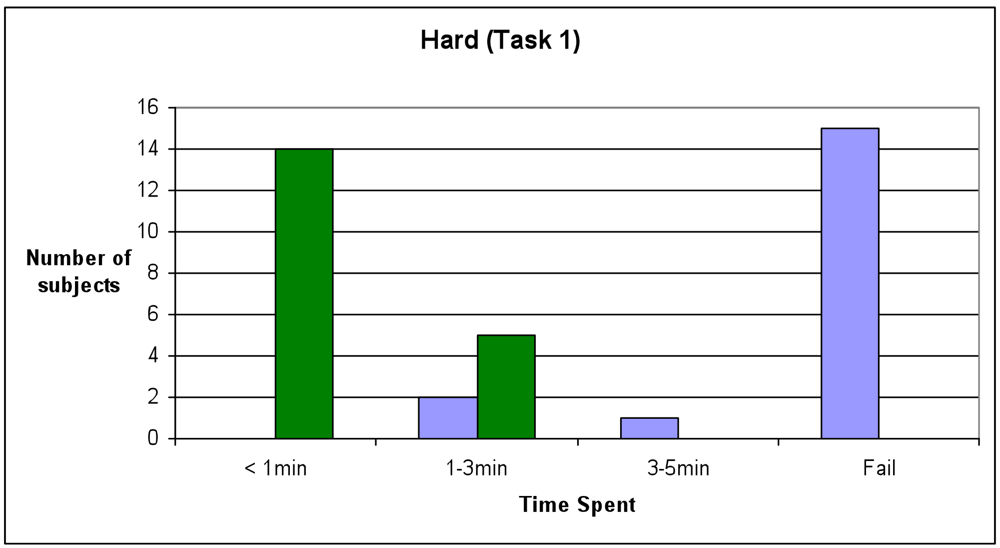
- Karma performed worse in task 1 and tasks 2
  - only 30 sec difference
  - -subjects take times selecting attributes
  - - the saving will be realized in the data integration step.
- Karma performed better in task 3 because of it can automatically identify the attribute



Dapper/Pipes **Karma**

Evaluation: Data Cleaning



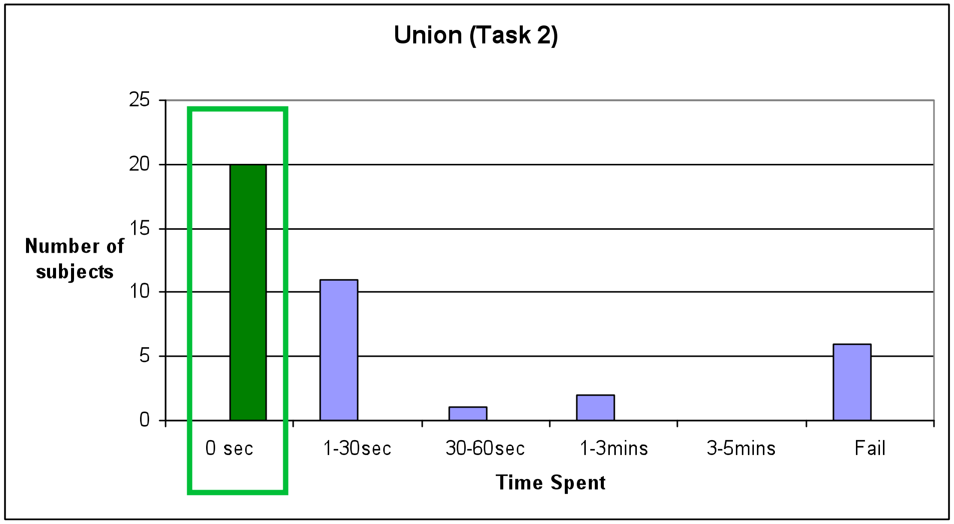


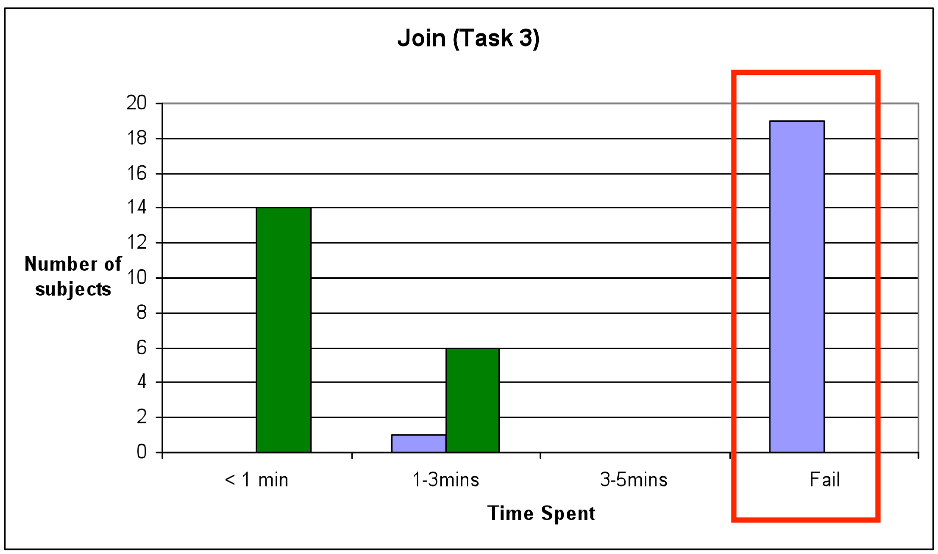
- Karma performed better in both tasks

- 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

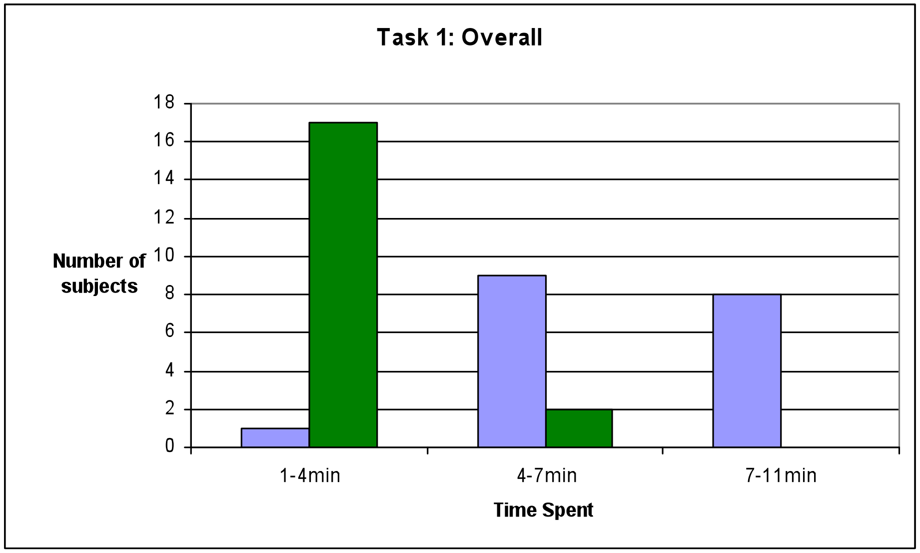
- The time spent in source modeling step allows Karma to suggest the linking source

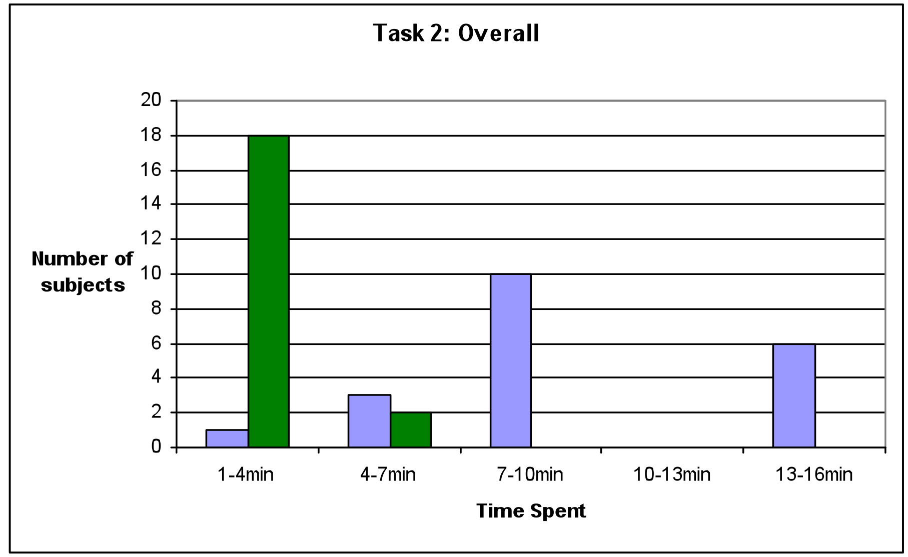
- Dapper/Pipes: 30% fail in the union case and 95% fail in the join case

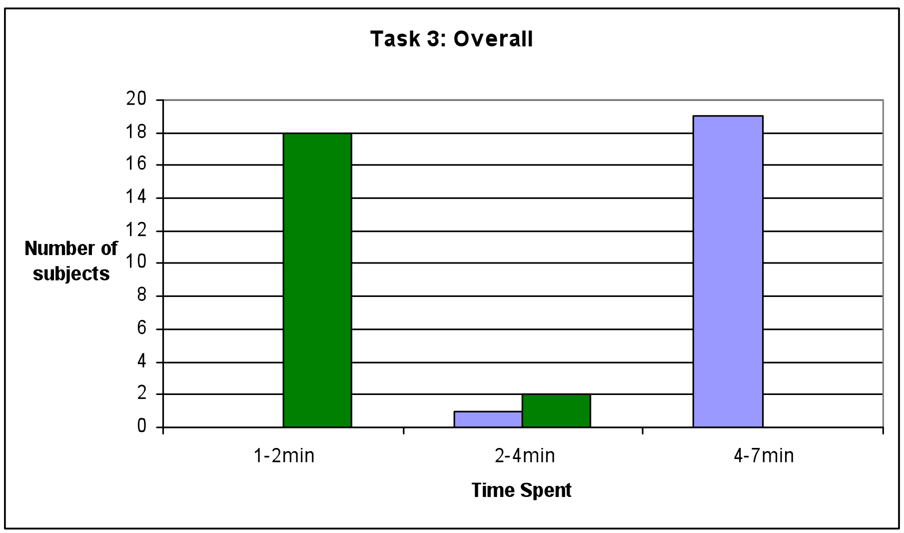
Dapper/Pipes **Karma**

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

Evaluation: Overall

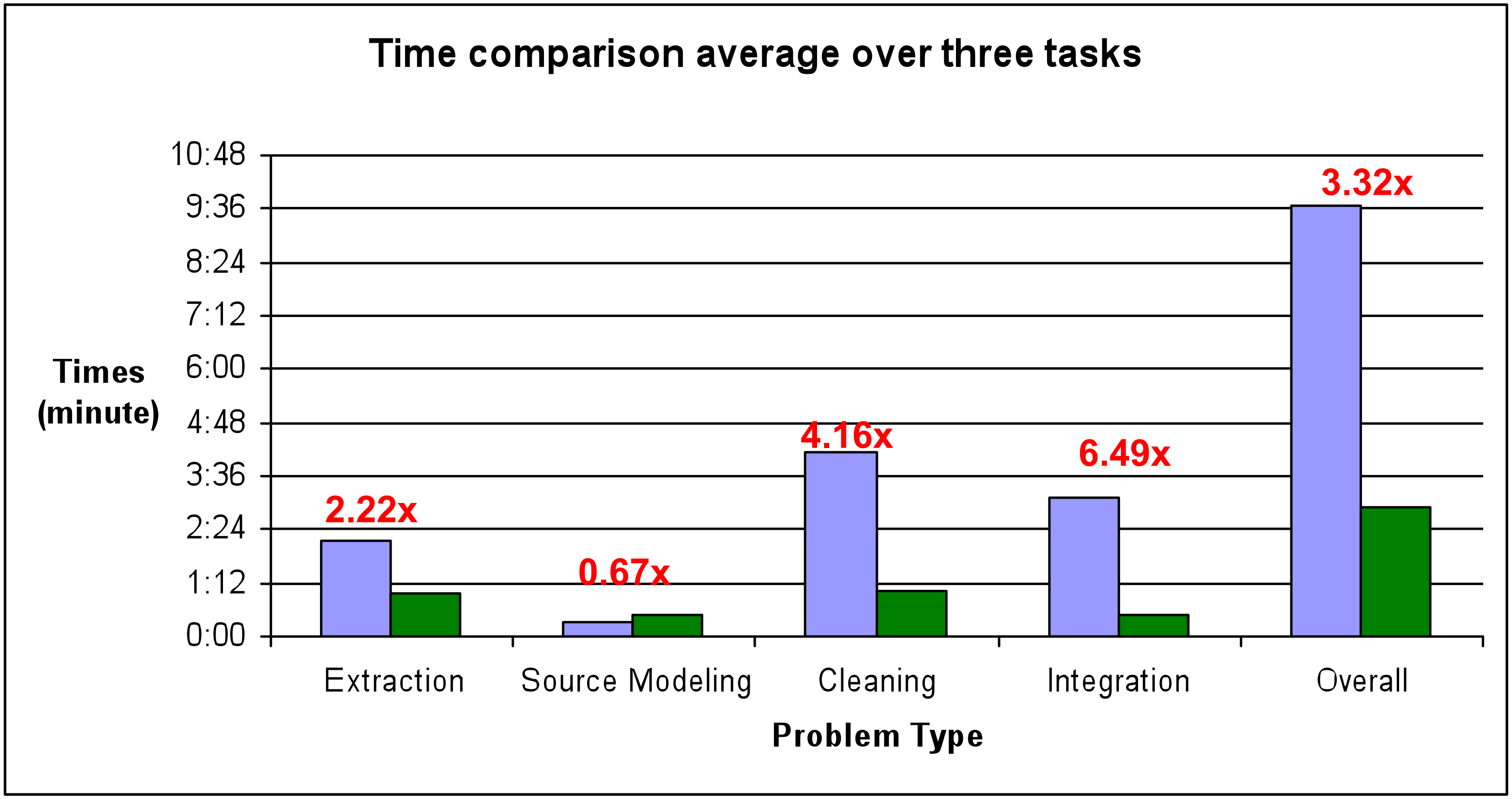






Dapper/Pipes Karma

Evaluation: Average



Dapper/Pipes Karma

Evaluation: T-test

- Karma is faster than DP in data extraction subtasks
  - $t=5.35$, $df = 114$, $p < 0.01$
- DP is faster than Karma in source modeling subtasks
  - $t=2.54$, $df=114$, $p < 0.01$
- Karma is faster than DP in data cleaning subtasks
  - $t=13.54$, $df=74$, $p < 0.01$
- Karma is faster than DP in data integration subtasks
  - $t=7.05$, $df=78$, $p < 0.01$
- Karma is faster than DP overall
  - $t=13.24$, $df=114$, $p < 0.01$

Not susceptible to time-bound bias because only DP's result (and not Karma) contains time-bound values.

Related Work: Mashup Systems

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

Related Work: Mashup Systems

System	Data Retrieval	Source Modeling	Data Cleaning	Data Integration	Mashup Type Supported
MIT's Simile	DOM	Manual	N/A	N/A	1
MIT's Pot Luck	RDF	Manual	PBD	Manual	1,3,4
Dapper	DOM	Manual	Manual	Join only	1,2,4
Yahoo's Pipes	Widgets	Manual	Widgets	Union only	1,2,3
MS's Popfly	Widgets	Manual	Widgets	Widgets	1,2,4
CMU's Marmite	Widgets	Manual	Widgets	Widgets	1,2,4
Intel's Mashmaker	Dapper	Manual	Widgets	Expert	1,2,3,4
Google MyMap	Widgets	Manual	N/A	Union only	1,2
Agent Wizard	Q/A	Q/A	Q/A	Q/A	1,3,4
Cards	DOM	Manual	N/A	Manual	1,2,4
Karma	DOM	Database	PBD	PBD	1,2,3,4

Related Work: Mashup Systems

System	Data Retrieval	Source Modeling	Data Cleaning	Data Integration	Mashup Type Supported
MIT's Simile	DOM	Manual	N/A	N/A	1
MIT's Pot Luck	RDF	Manual	PBD	Manual	1,3,4
Dapper	DOM	Manual	Manual	Join only	1,2,4
Yahoo's Pipes	Widgets	Manual	Widgets	Union only	1,2,3
MS's Popfly	Widgets	Manual	Widgets	Widgets	1,2,4
CMU's Marmite	Widgets	Manual	Widgets	Widgets	1,2,4
Intel's Mashmaker	Require an expert				1,2,3,4
Google MyMap	Widgets	Manual	N/A	Union only	1,2
Agent Wizard	Q/A	Q/A	Q/A	Q/A	1,3,4
Cards	DOM	Manual	N/A	Manual	1,2,4
Karma	DOM	Database	PBD	PBD	1,2,3,4

Related Work: Mashup Systems

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	Manual	PBD	Manual	1,3,4
Dapper	DOM	Manual	Manual	Join only	1,2,4
Yahoo's Pipes	Widgets	Manual	Widgets	Union only	1,2,3
MS's Popfly	Widgets	Manual	Widgets	Widgets	1,2,4
CMU's Marmite	Widgets	Manual	Widgets	Widgets	1,2,4
Intel's Mashmaker	Require an expert				1,2,3,4
Google MyMap	Widgets	Manual	N/A	Union only	1,2
Agent Wizard	Q/A	Q/A	Q/A	Q/A	1,3,4
Cards	DOM	Manual	N/A	Manual	1,2,4
Karma	DOM	Database	PBD	PBD	1,2,3,4

Related Work: Mashup Systems

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	DOM	Manual	Manual	Join only	1,2,4
Yahoo's Pipes	Widgets	Manual	Widgets	Union only	1,2,3
MS's Popfly	Widgets	Manual	Widgets	Widgets	1,2,4
CMU's Marmite	Widgets	Manual	Widgets	Widgets	1,2,4
Intel's Mashmaker	Require an expert				1,2,3,4
Google MyMap	Widgets	Manual	N/A	Union only	1,2
Agent Wizard	Q/A	Q/A	Q/A	Q/A	1,3,4
Cards	DOM	Manual	N/A	Manual	1,2,4
Karma	DOM	Database	PBD	PBD	1,2,3,4

Related Work: Mashup Systems

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	Manual	Widgets	Union only	1,2,3
MS's Popfly	Widgets	Manual	Widgets	Widgets	1,2,4
CMU's Marmite	Widgets	Manual	Widgets	Widgets	1,2,4
Intel's Mashmaker	Require an expert				1,2,3,4
Google MyMap	Widgets	Manual	N/A	Union only	1,2
Agent Wizard	Q/A	Q/A	Q/A	Q/A	1,3,4
Cards	DOM	Manual	N/A	Manual	1,2,4
Karma	DOM	Database	PBD	PBD	1,2,3,4

Related Work: Mashup Systems

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	Widgets	Manual	N/A	Union only	1,2
Agent Wizard	Q/A	Q/A	Q/A	Q/A	1,3,4
Cards	DOM	Manual	N/A	Manual	1,2,4
Karma	DOM	Database	PBD	PBD	1,2,3,4

Related Work: Mashup Systems

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	Q/A	Q/A	Q/A	1,3,4
Cards	DOM	Manual	N/A	Manual	1,2,4
Karma	DOM	Database	PBD	PBD	1,2,3,4

Related Work: Mashup Systems

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	DOM	Manual	N/A	Manual	1,2,4
Karma	DOM	Database	PBD	PBD	1,2,3,4

Related Work: Mashup Systems

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

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
  - Gained reference sources through source modeling process
  - Provided 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 Technique: 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

- Mashup is 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 four 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

- Customizing display by examples
  - Interface today is limited to the table of data or a map if the points can be geocoded
- Learn and generalize over the task
  - Store the integration plan so that it can be reexecuted on new data
- Integrate in work on machine learning of extraction tasks
  - In collaboration with Fetch Technologies
- Integrate in work on automatic source modeling

Papers

- 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*

Thank You!

Backup Slides

Task1	Dapper/Pipes				Karma			
Subject	E	M	C	Total	E	M	C	Total
No.1	*5:00	0:20	*5:00	10:20	2:19	1:08	1:00	4:27
No.2	1:43	0:30	*5:00	7:13	1:00	0:40	0:29	2:09
No.3	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
No.4	0:52	0:48	*5:00	6:40	1:12	1:00	0:50	3:02
No.5	5:00	0:35	*5:00	10:35	1:15	1:18	1:20	3:53
No.6	2:30	0:15	*5:00	7:45	1:00	0:54	0:28	2:22
No.7	1:20	0:22	*5:00	6:42	0:51	0:51	0:46	2:28
No.8	1:40	0:14	*5:00	6:54	1:04	0:41	0:33	2:19
No.9	1:26	0:16	*5:00	6:42	1:14	1:00	1:10	3:24
No.10	1:39	0:10	*5:00	6:49	0:53	0:42	0:50	2:26
No.11	2:00	0:19	*5:00	7:19	1:04	1:00	0:53	2:57
No.12	2:00	0:49	2:00	4:49	1:07	1:00	0:40	2:47
No.13	2:00	0:05	*5:00	7:05	0:58	0:50	0:56	1:44
No.14	2:46	0:15	*5:00	8:01	1:12	0:45	0:48	2:45
No.15	2:27	0:14	3:11	5:52	1:10	0:49	1:20	3:19
No.16	1:16	0:12	*5:00	6:28	0:58	0:42	0:25	2:05
No.17	n/a	n/a	n/a	n/a	2:00	1:00	0:50	3:50
No.18	2:30	0:14	*5:00	7:44	1:06	1:10	1:46	4:02
No.19	1:38	0:47	1:20	3:45	1:20	0:49	0:35	2:44
No.20	1:30	0:16	*5:00	6:46	1:04	0:44	0:35	2:23
No.21	n/a	n/a	n/a	n/a	1:11	1:17	0:59	3:27
No.22	n/a	n/a	n/a	n/a	2:58	1:46	1:24	6:08
No.23	n/a	n/a	n/a	n/a	1:19	1:40	1:14	4:13

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
No.7	2:17	0:15	4:46	0:18	7:36	1:13	0:25	0:52	0:00	2:31
No.8	3:23	0:21	*5:00	*5:00	13:44	1:10	0:21	0:24	0:00	1:55
No.9	4:11	0:21	*5:00	*5:00	14:32	1:22	0:47	2:11	0:00	4:20
No.10	2:16	0:07	3:14	0:20	5:50	1:04	0:20	1:06	0:00	2:30
No.11	3:04	0:17	*5:00	*5:00	13:21	1:06	0:34	0:53	0:00	2:33
No.12	2:00	0:27	*5:00	0:20	7:47	1:23	0:30	0:37	0:00	2:30
No.13	*5:00	0:07	1:43	0:10	7:00	1:42	0:32	0:41	0:00	2:55
No.14	3:03	0:23	4:42	0:10	8:21	1:40	0:31	0:56	0:00	3:07
No.15	2:06	0:12	3:13	0:22	5:53	1:30	0:24	2:05	0:00	3:59
No.16	3:58	0:11	3:29	0:27	8:05	0:51	0:17	1:00	0:00	2:08
No.17	4:15	0:28	3:39	0:30	8:52	1:04	0:28	1:18	0:00	2:50
No.18	*5:00	0:23	*5:00	*5:00	15:23	1:17	0:30	1:10	0:00	2:57
No.19	4:01	0:14	2:42	0:21	7:16	1:39	0:21	0:50	0:00	2:50
No.20	1:36	0:43	0:36	0:22	3:17	1:07	0:28	0:40	0:00	2:15
No.21	n/a	n/a	n/a	n/a	n/a	1:03	0:21	0:55	0:00	2:19
No.22	n/a	n/a	n/a	n/a	n/a	3:56	1:52	2:50	0:00	8:38
No.23	n/a	n/a	n/a	n/a	n/a	2:15	0:31	1:27	0:00	4:13

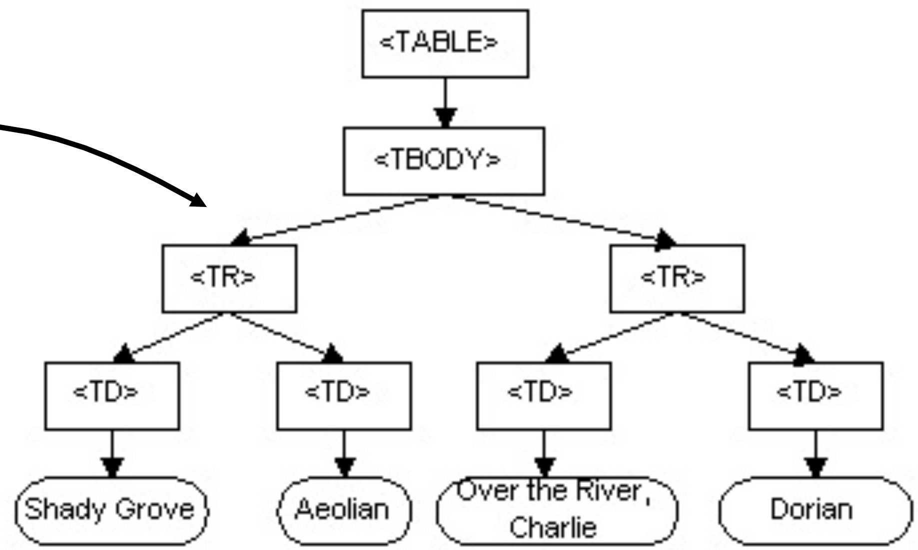
Task3	Dapper/Pipes				Karma			
Subject	E	M	I	Total	E	M	I	Total
No.1	1:30	0:26	*5:00	6:56	0:14	0:00	2:16	2:30
No.2	0:30	0:10	*5:00	5:40	0:25	0:00	0:26	0:54
No.3	1:00	0:15	*5:00	6:15	0:15	0:00	0:44	0:59
No.4	0:40	0:16	*5:00	5:56	0:20	0:00	1:06	1:26
No.5	0:40	0:14	*5:00	5:54	0:20	0:00	0:37	0:57
No.6	0:30	0:10	*5:00	5:40	0:20	0:00	0:31	0:51
No.7	0:27	0:10	*5:00	5:37	0:14	0:00	0:50	1:04
No.8	0:29	0:20	*5:00	5:49	0:30	0:00	0:51	1:21
No.9	0:40	0:23	*5:00	6:03	0:13	0:00	0:44	0:57
No.10	0:30	0:10	*5:00	5:40	0:20	0:00	0:35	0:55
No.11	0:51	0:20	*5:00	6:11	0:16	0:00	1:05	1:21
No.12	1:05	0:18	*5:00	6:23	0:30	0:00	0:46	1:16
No.13	0:31	0:14	*5:00	5:45	0:16	0:00	0:57	1:13
No.14	0:36	0:14	*5:00	5:50	0:14	0:00	2:00	2:14
No.15	0:26	0:21	*5:00	5:47	0:30	0:00	0:45	1:15
No.16	0:27	0:13	*5:00	5:40	0:15	0:00	0:56	1:11
No.17	0:33	0:38	1:56	3:07	0:30	0:00	0:46	1:16
No.18	1:03	0:07	*5:00	6:10	0:20	0:00	1:10	1:30
No.19	0:33	0:17	*5:00	5:50	0:25	0:00	1:20	1:45
No.20	0:18	0:13	*5:00	5:31	0:12	0:00	0:44	0:56
No.21	n/a	n/a	n/a	n/a	0:15	0:00	0:39	0:54
No.22	n/a	n/a	n/a	n/a	0:21	0:00	2:50	3:11
No.23	n/a	n/a	n/a	n/a	0:12	0:00	0:51	1:03

Table 8.7: Individual and Overall failure rate in Dapper/Pipes.

Task	Task 1	Task 2	Task 3
Data Extraction	5%	25%	0%
Source Modeling	0%	0%	0%
Data Cleaning	83%	35%	n/a
Data Integration	n/a	30%	95%
Overall	83%	45%	95%

Document Object Model (DOM)

```
<TABLE>
<TBODY>
<TR>
<TD>Shady Grove</TD>
<TD>Aeolian</TD>
</TR>
<TR>
<TD>Over the River, Charlie</TD>
<TD>Dorian</TD>
</TR>
</TBODY>
</TABLE>
```



Vertical Expansion