Building Mashups

Craig Knoblock
University of Southern California

Thanks to Rattapoom Tuchinda
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  
b) zillow.com  
c) Ski bonk

Combined Data gives new insight / provides new services
Outline

- Karma – Programming by demonstration approach
- Marmite – Workflow based approach
- Mashmaker – Browser-based approach
Outline

• Karma – Programming by demonstration approach
• Marmite – Workflow based approach
• Mashmaker – Browser-based approach
Mashup Building Issues

Data Retrieval

Calibration
- source modeling
- cleaning

Integration

Display

Wrapper
Attribute
Clean
Combine
Customize Display
Existing Approaches

Goal: Create Mashups without Programming
• Addresses syntax issues, but users still required to understand programming concepts

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?
Our Integrated Approach

- Focus on data, not on the process
  - Users are already familiar with data.
  - Capture and model the Mashup building process from examples (PBD)
- Consolidate rather than Divide-And-Conquer
  - Solving one issue can help solve other issues.
  - Use one interaction platform -- a table
- Leverage existing database
  - Helps with source modeling, cleaning, and data integration.
Our system: Karma

Embedded Browser

Table

Interaction Modes
Extract

{Restaurant name, address, phone, Review}

Clean

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

Map

Database

{Restaurant name, address, Date of Inspection, Score}
1. **Japon Bistro**
   927 E Colorado Blvd., Pasadena, CA, 91106
   Upscale yet affordable Japanese eatery offers the city's largest sake selection.

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.

Data Retrieval: Extraction

<table>
<thead>
<tr>
<th>select one</th>
<th>Tbody/tr[1]/td[2]/a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japon Bistro</td>
<td>Sushi Dokoro...</td>
</tr>
<tr>
<td>Hokusai</td>
<td>Sushi Sasab...</td>
</tr>
<tr>
<td>Sushi Roku</td>
<td>Sushi Roku</td>
</tr>
<tr>
<td>Hide Sushi</td>
<td>Hide Sushi</td>
</tr>
<tr>
<td>Fat Fish</td>
<td>Fat Fish</td>
</tr>
<tr>
<td>Sushi Katsu-ya</td>
<td>Sushi Katsu-ya</td>
</tr>
<tr>
<td>Gindi Thai /...</td>
<td>Gindi Thai /...</td>
</tr>
<tr>
<td>Katana</td>
<td>Katana</td>
</tr>
<tr>
<td>Echigo</td>
<td>Echigo</td>
</tr>
</tbody>
</table>

Tbody/tr*/td*/a
### Data Retrieval: Navigation

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

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

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

<table>
<thead>
<tr>
<th>select one</th>
<th>address</th>
<th>select one</th>
<th>Select one</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japon Bistro</td>
<td>927 E Colora..</td>
<td>Upscale yet...</td>
<td>31 Reviews</td>
</tr>
<tr>
<td>Sushi Dokor</td>
<td>9777 S Sant..</td>
<td>Intimate and...</td>
<td></td>
</tr>
<tr>
<td>Hokusai</td>
<td>8400 Wilshir..</td>
<td>Chic elegance...</td>
<td></td>
</tr>
<tr>
<td>Sushi Sasab</td>
<td>12400 Wilshi..</td>
<td>Authentic Ja...</td>
<td></td>
</tr>
<tr>
<td>Sushi Roku</td>
<td>8445 W 3rd..</td>
<td>High fashion...</td>
<td></td>
</tr>
<tr>
<td>Hide Sushi</td>
<td>2040 Sawtel..</td>
<td>No fuss, just...</td>
<td></td>
</tr>
<tr>
<td>Fat Fish</td>
<td>616 N Rober...</td>
<td>Inventive ro...</td>
<td></td>
</tr>
<tr>
<td>Sushi Katsu-ya</td>
<td>11680 Vent...</td>
<td>The MOCA o...</td>
<td></td>
</tr>
<tr>
<td>Gindi Thai</td>
<td>4017 W Riv...</td>
<td>Burbank res...</td>
<td></td>
</tr>
<tr>
<td>Katana</td>
<td>8439 W Sun...</td>
<td>Rustic Japa...</td>
<td></td>
</tr>
<tr>
<td>Echigo</td>
<td>11217 Sant...</td>
<td>Stellar sushi...</td>
<td></td>
</tr>
</tbody>
</table>
Source Modeling (Attribute selection)

Possible Attribute
\( \{ a \mid a, s: a \in \text{att}(s) \land (\text{val}(a, s) \subseteq V) \} \)

restaurant name (3)
artist name (1)
### Data Cleaning: using existing values

#### Newly extracted data

<table>
<thead>
<tr>
<th>Restaurant Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japon Bistro</td>
</tr>
<tr>
<td>Hokusai</td>
</tr>
<tr>
<td>Sushi Sasabune</td>
</tr>
<tr>
<td>Sushi Roka</td>
</tr>
</tbody>
</table>

#### Data repository

<table>
<thead>
<tr>
<th>Restaurant Name</th>
<th>Address</th>
<th>Health Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hokusai</td>
<td>8400..</td>
<td>90</td>
</tr>
<tr>
<td>Katana</td>
<td>8439..</td>
<td>99</td>
</tr>
<tr>
<td>Japon Bistro</td>
<td>927 E..</td>
<td>95</td>
</tr>
</tbody>
</table>

#### Zagat

<table>
<thead>
<tr>
<th>Restaurant Name</th>
<th>Zagat Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sushi Sasabune</td>
<td>27</td>
</tr>
<tr>
<td>Sushi Roka</td>
<td>25</td>
</tr>
<tr>
<td>Katana</td>
<td>23</td>
</tr>
</tbody>
</table>
Data Cleaning: using predefined rules

28 Reviews $\rightarrow$ 28

Subset Rule:

$$(s_1 s_2 \ldots s_k) \rightarrow (d_1 d_2 \ldots d_t) \land (k \leq t) \land s_i \in \{d_1, d_2, \ldots, d_t\} \land d_i \neq d_j$$

Predefined Rules
Data Integration
Based on [tuchinda 2007]
Data Integration (cont.)

\[
\{a\}_R = \text{possible new attribute selection for row } i.
\]

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

\[
\{v\} = val(a,s) \text{ where } a \{x\}
\]

\(s\) is any source where \(att(s) \{x\} \neq \{\}\)
Map Generation
Related Work

• Building Blocks
  – Programming by Demonstration [Cypher 1993, Lau 2001]
  – Simile [Huynh 2005]
  – Potter’s Wheel [Raman 2001]
  – Building Queries by Demonstration [Tuchinda 2007]

• Data Extraction
  – Simile, Dapper, D.Mix [Hartman 2007], OpenKapow

• Widget Approach
  – Yahoo’s Pipes, Microsoft’s Popfly, IBM’s QED Wiki, Bungee Labs, Proto Software, Marmite [Wong 2007]

• Misc.
  – Intel’s Mashmaker [Ennals 2007]
  – Google MyMap
Evaluation: Average

Time comparison average over three tasks

- Extraction: 2.22x
- Source Modeling: 0.67x
- Cleaning: 4.16x
- Integration: 6.49x
- Overall: 3.32x

Dapper/Pipes vs Karma
Discussion

• Contribution: An approach to build Mashups by combining four common information integration techniques into a unified framework.
  – Data extraction
  – Source modeling
  – Data Cleaning
  – Data Integration
Outline

• Karma – Programming by demonstration approach
• Marmite – Workflow based approach
• Mashmaker – Browser-based approach
Marmite

• Widget/Workflow approach similar to Yahoo’s Pipes and Microsoft’s Popfly
• Firefox extensions
• The interface is divided into three sections
  – Widget selection
  – Workflow
  – Intermediate results

**Step 2: Extract Address**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>$930/3 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>3624 California</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$625/2 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>Babock Blvd</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$695/3 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>Park St</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$625/2 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>Pamela Dr</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$725/2 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>Locustridge Dr</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$575/2 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>132 at First St</td>
<td>15238</td>
<td>PA</td>
</tr>
<tr>
<td>$510/1 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>22 Oakland Ave</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$925/3 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>138 S Joslyn</td>
<td>Penn Hills</td>
<td>Hi</td>
</tr>
<tr>
<td>$725/2 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>22 Oakland Ave</td>
<td>Brookline</td>
<td>PA</td>
</tr>
<tr>
<td>$579/2 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$599/1 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$629/2 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$225000 &quot;R...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$469/1 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1195/3 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>Wellesley at...</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$890/2 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>Bunker Hill Ave</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$425 Cozy,</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$949/2 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>Bunker Hill Ave</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$835/2 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>S. Braddock Ave</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$689/2 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$485/1 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>S. Fairmont Ave</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$1200/4 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$405/1 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>S. Fairmont Ave</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$825/2 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>14 Soffel St</td>
<td>Mt. Washington</td>
<td>We</td>
</tr>
<tr>
<td>$435 Beechwood St</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>1618 Belasco Ave</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$499/1 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$595/2 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>Broadway Ave</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$1200/5 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>Whitney Ave</td>
<td>Pittsburgh</td>
<td>PA</td>
</tr>
<tr>
<td>$405/2 br...</td>
<td><a href="http://pittsburgh">http://pittsburgh</a>...</td>
<td>616 Carson St</td>
<td>Carnegie</td>
<td>PA</td>
</tr>
</tbody>
</table>

**Workflow:**

1. **Select Links From Page**
2. **Extract Address**
   - **Input:**
     - Got: from column:
     - Web Page: B: URL
   - **Output:**
     - Write:
       - Street Address: new column
       - City: new column
       - State: new column
   - Displays: View results for this step

**Next operator suggestions.**

Replace this placeholder by selecting the next operator from:
- operator list or the suggestions list.

Show suggestions
Marmite Approach

- Based on Apple Automator
- One of a few that design the system by doing user studies prior implementation
  - Showing intermediate result
  - Suggestion for the next operators
Marmite Evaluation

• 6 People
  – 2 novices
  – 2 people who know how to use spreadsheet
  – 2 programmers

• 4 Tasks
  – Retrieve a set of addresses and geocode an address
  – Search and filter out events further than a week away
  – Compile a list of events from two event services and plot them on a map.
  – Recreate the map from housingmaps website
Marmite Result

• 3 (1 spreadsheet, 2 programmers) complete the 4 tasks in one hour.
  – Novice did not finish all the tasks.

• The biggest problem for them is understanding data flow
  – Confusion about the input/output concept
  – Did not understand that the data flow and the spreadsheet result are linked.
Marmite Video
Outline

• Karma – Programming by demonstration approach
• Marmite – Workflow based approach
• Mashmaker – Browser-based approach
**Intel Mashmaker**

- **Multi-tier user**
  - Naïve users
  - Expert users

- **Experts do all the dirty works to customize the integration between sources.**

- **Naïve users browse web pages normally**
  - If the page that the user is viewing contain an existing wrapper or predefined integration, the user can get those information by pressing a button
Intel Mashmaker: Design Principles

- **Program as you browse**
  - view Mashup creation as an extension of the normal web browsing habits

- **Direct manipulation**
  - work on data without having to think about abstract concepts such as programs

- **Pay as you go**
  - Unskilled users should be able to gain some benefit with very little effort
  - Experts should be able to do more advanced stuff
Intel Mashmaker: Features

- Look at Dapper to see if the wrapper for a particular site exist
- Direct manipulation of data through operations such as map, fold, and filter
- User can interact with Mashmaker at a number of different levels depending on the skill
Intel Mashmaker: Users

- **Basic**: know nothing
- **Normal**: Occasionally expand the widget panel to edit form parameters
- **Skilled**: Connecting sources
- **Semi-Expert**: Extract data from new sites
- **Expert**: Write complex expression directly in Mash-Maker’s core language
- **Gurus**: Teach Mashmaker to understand the content of the new website.
Mashmaker Video
Conclusion

• Tradeoffs in each approach
  – Karma
    • Pro: end-to-end approach based on a programming by demonstration paradigm
    • Con: May not work on all web sites
  – Marmite
    • Pro: Easy integration of capabilities
    • Con: Dataflow model is difficult for users to understand
  – Mashmaker
    • Pro: Browser-based integration
    • Con: Requires an expert to add new functionality