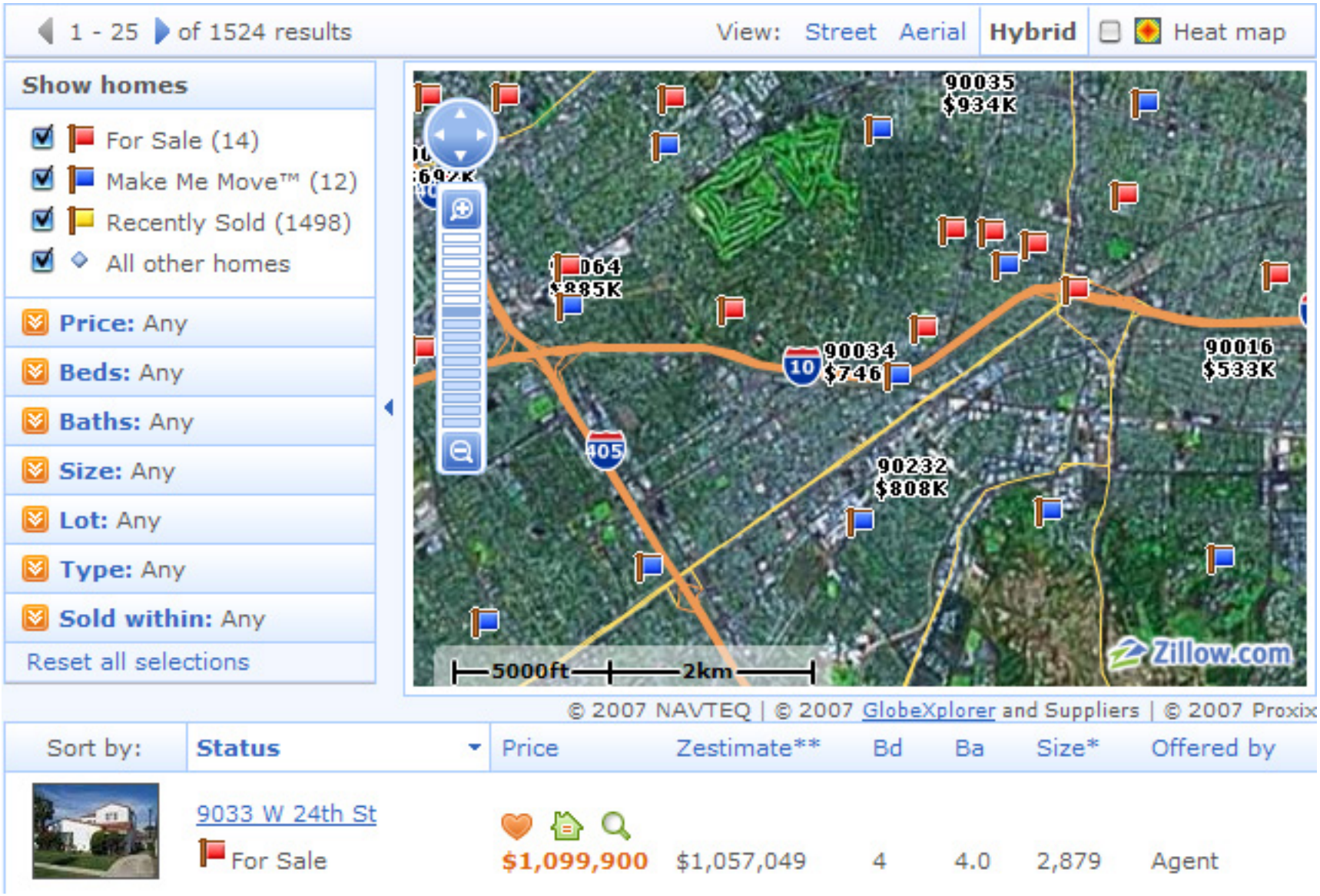
Building Data Integration Queries by Demonstration

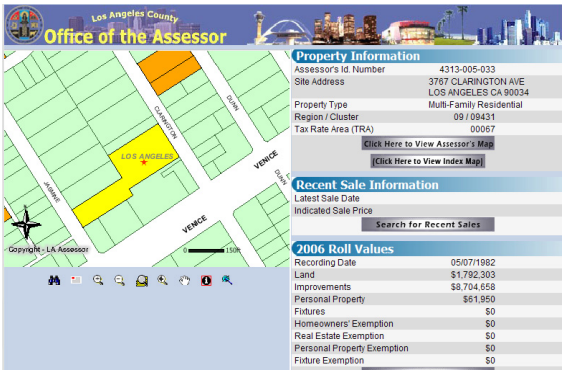
Rattapoom Tuchinda, Pedro Szekely,
and Craig A. Knoblock
Information Science Institute
University of Southern California

Intelligent User Interface. Jan 30, 2007

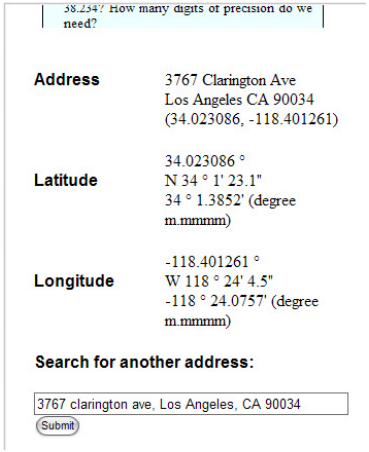
Vision and Motivation



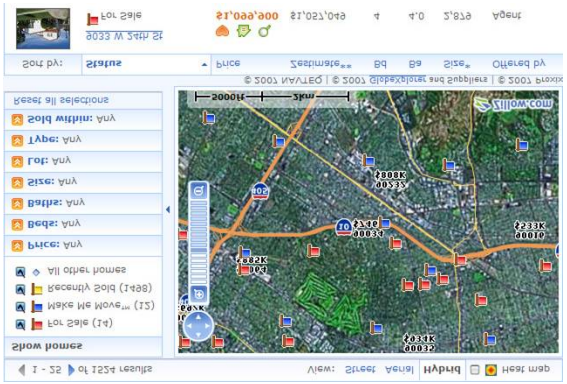
Zillow.com

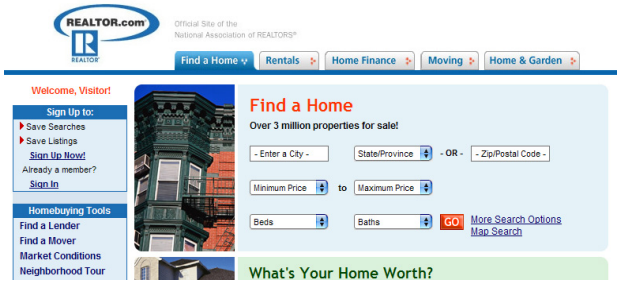


Tax properties

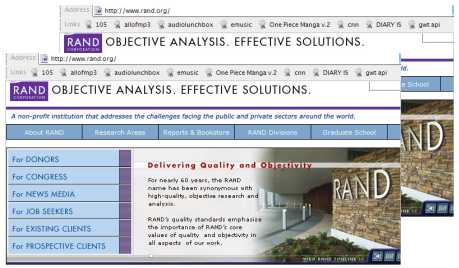


Geocoder

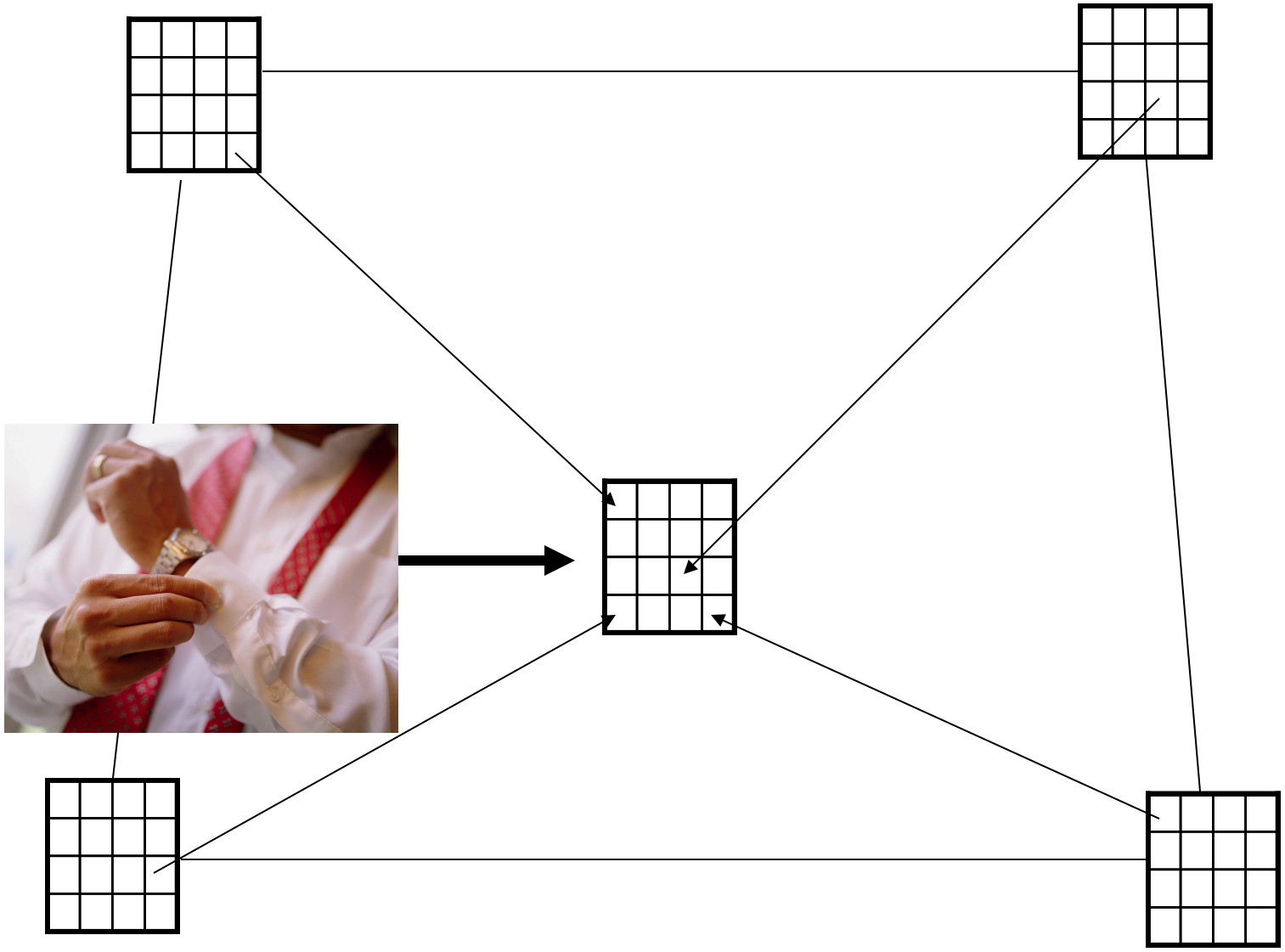




Realtor



And many more..



Why is it hard?

Web Source Characteristics:

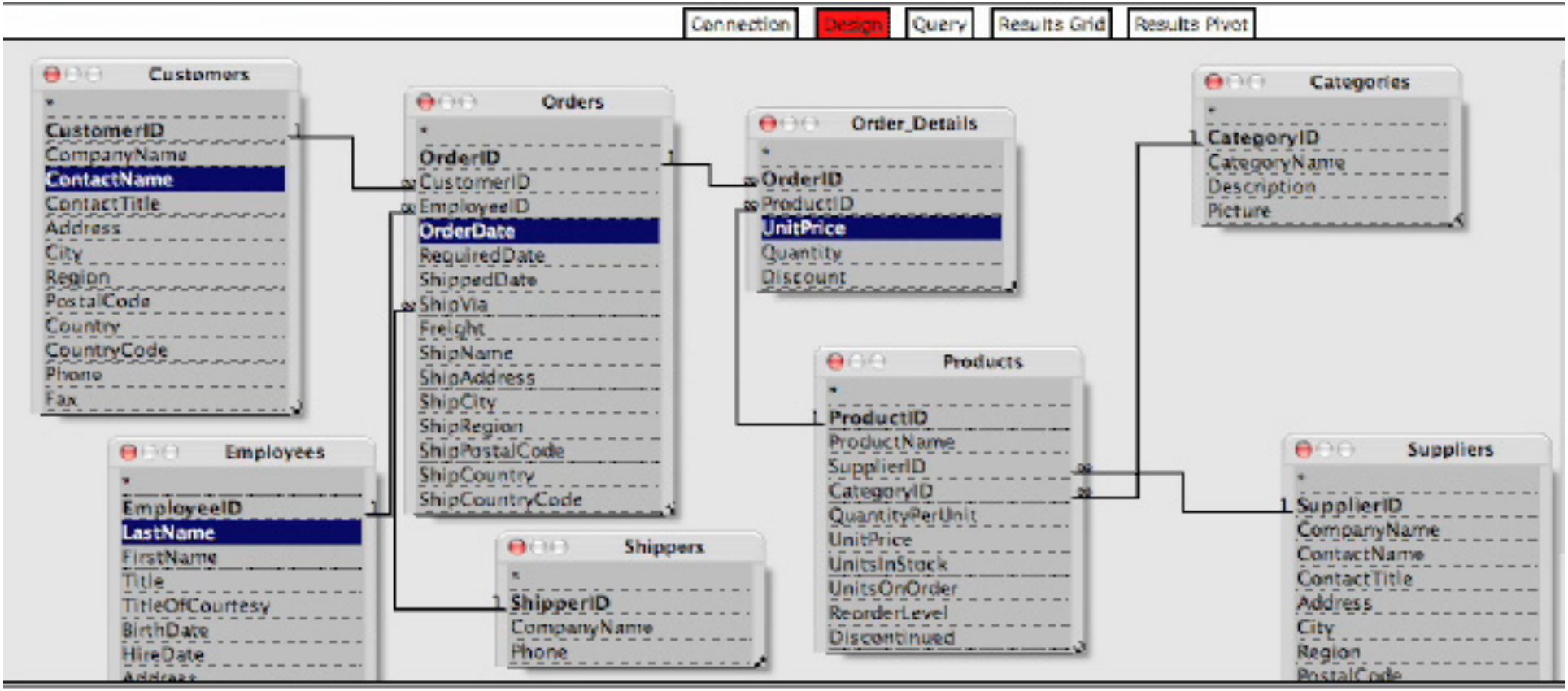
1. The number of sources is huge
2. Overlapping data between sources

As a result, it's difficult to write sql queries.

User Characteristics:

1. Don't know how to program
2. Don't always know what sources are available
3. Do know partial "data" (data value) that they want but may not know the "semantic" (attribute). E.g. hyatt (hotelname), waikiki (city)

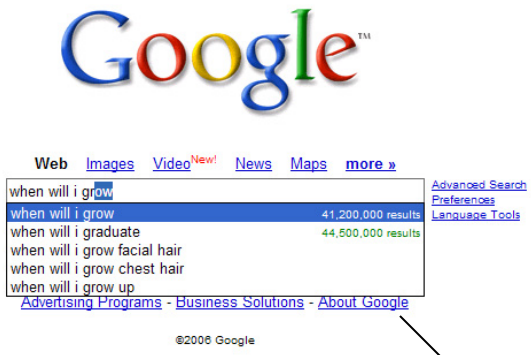
Query By Example



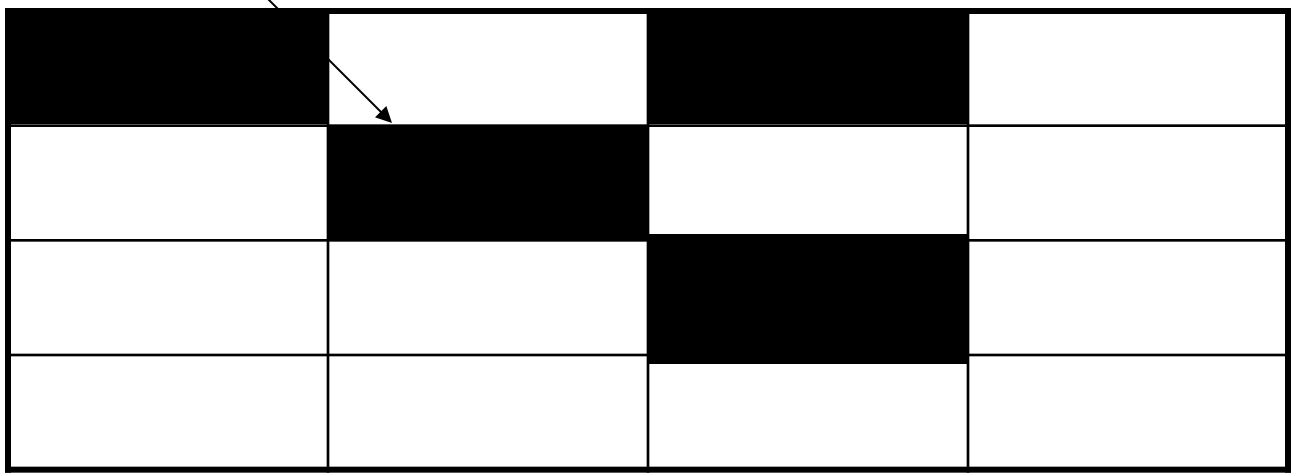
Query columns clear	Column #1 remove	Column #2 remove	Column #3 remove	Column #4 remove	Column #5 remove	Column #6 remove	Column #7
Column	"Customers"."Cc	"Customers"."Cc	"Orders"."OrderII	"Orders"."OrderI	"Employees"."Fir	"Employees"."La:	* "Order_Details"."(
Alias							
Show	☑	☑	☑	☑	☑	☑	☑
Sort	(none)	(none)	(none)	(none)	(none)	(none)	(none)
Where conditions	AND =	AND =	AND =	AND =	AND =	AND =	AND =
Group by	☐	☐	☐	☐	☐	☐	☐
	AND	AND	AND	AND	AND	AND	

Our Approach

Our Approach



Our Approach



Intuition

Crime rate

City	Crime Rate
Detroit	88%
NYC	60%
Los Angeles	45%

white pages

Name	Address	City	State
John Smith	500 Bundy	Los Angeles	CA
Mary Smith	100 Main	Long Beach	CA
Joe Smith	744 Temple	Ventura	CA

	Los Angeles	CA	

Los

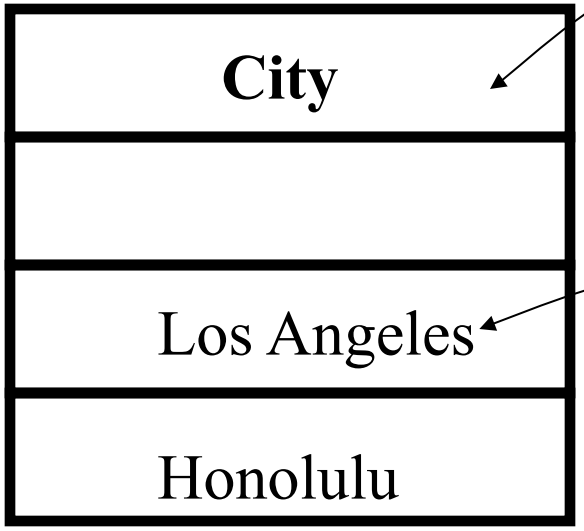
Hotels

Hotel name	Address	City	Phone
Hyatt	15 fiji	Salt Lake	801-534-4900
Radisson	10 Slauson	Los Angeles	310-666-6666
Marriott	102 Ames St.	Cambridge	617-353-1432

Politics

Party	State
Republican	CA
Democrat	MA
Republican	TX

Single Column Example



a ϵ all attribute

v ϵ all value

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

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

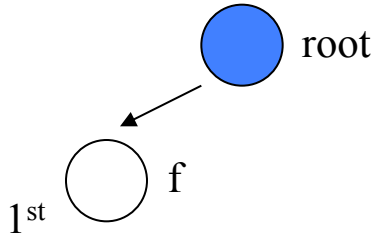
Honolulu : {(**city**, tax_properties), (**city**, favorite_vacation_spot)}

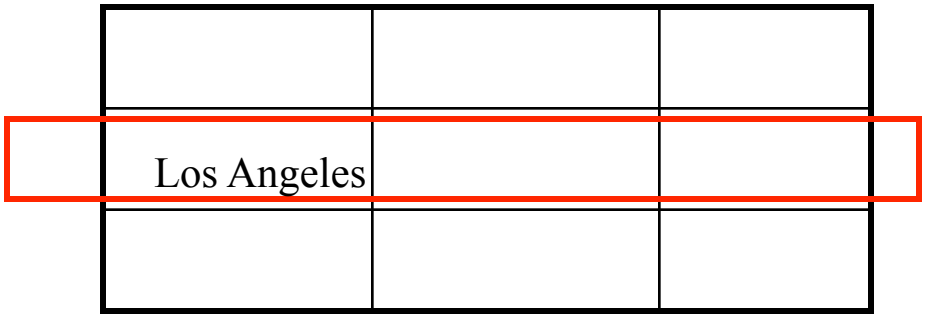
Can we determine the attribute now? Yes

{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

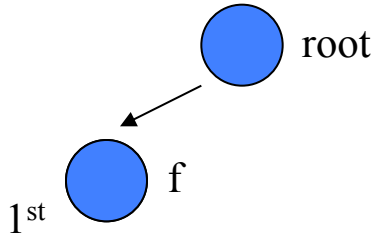
Query Transformation





f(*a,s,v*) = (?, ?, Los Anbeles)

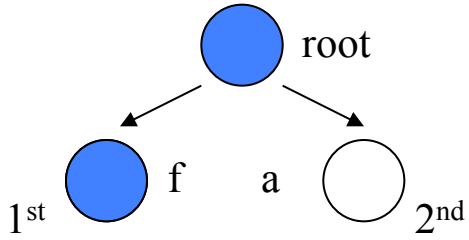
Query Transformation



City		
Los Angeles		
Honolulu		

f(*a,s,v*) = (city, tax_properties, Los Angeles)

Query Transformation



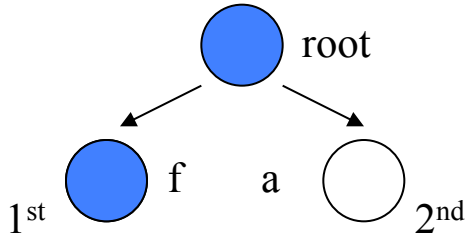
City	Zipcode	
Los Angeles		
Honolulu		

f(*a,s,v*) = (?, ?, Los Anbeles)

f(*a,s,v*) = (city, tax_properties, Los Angeles)

a(*a,s,v*) = (zipcode, tax_properties, _PLACE_HOLDER)

Query Transformation



City	Zipcode	
Los Angeles		
Honolulu		

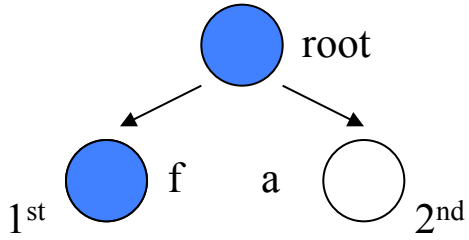
f(*a,s,v*) = (?, ?, Los Anbeles)

f(*a,s,v*) = (city, tax_properties, Los Angeles)

a(*a,s,v*) = (zipcode, tax_properties, _PLACE_HOLDER)

Select Zipcode
From tax_properties
Where City="Los Angeles"

Query Transformation



City	Zipcode	
Los Angeles		
Honolulu		

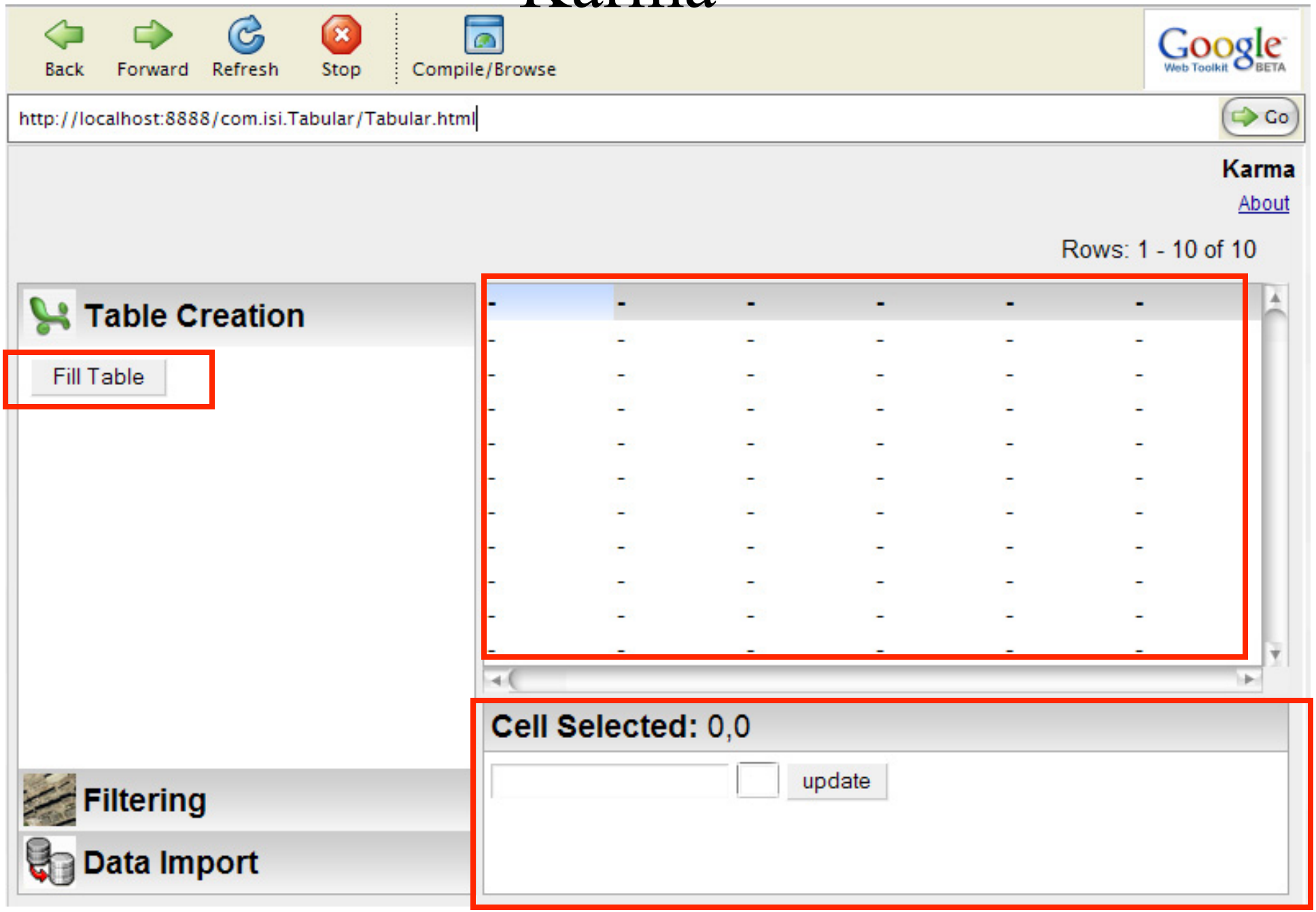
f(*a,s,v*) = (?, ?, Los Anbeles)

f(*a,s,v*) = (city, tax_properties, Los Angeles)

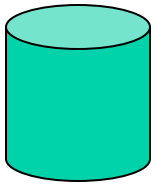
a(*a,s,v*) = (zipcode, tax_properties, _PLACE_HOLDER)

Select Zipcode
From tax_properties
Where City="Los Angeles"

Karma

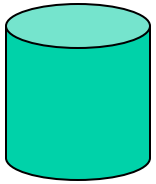

Back
Forward
Refresh
Stop
Compile/Browse
Google
Web Toolkit BETA
http://localhost:8888/com.isi.Tabular/Tabular.html
Go
Karma
About
Rows: 1 - 10 of 10
Table Creation
Fill Table
Filtering
Data Import
Cell Selected: 0,0
update

Example Data Sources



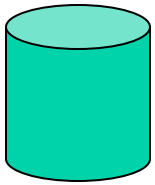
songs 2004

artist, album, song name, length



songs 2005

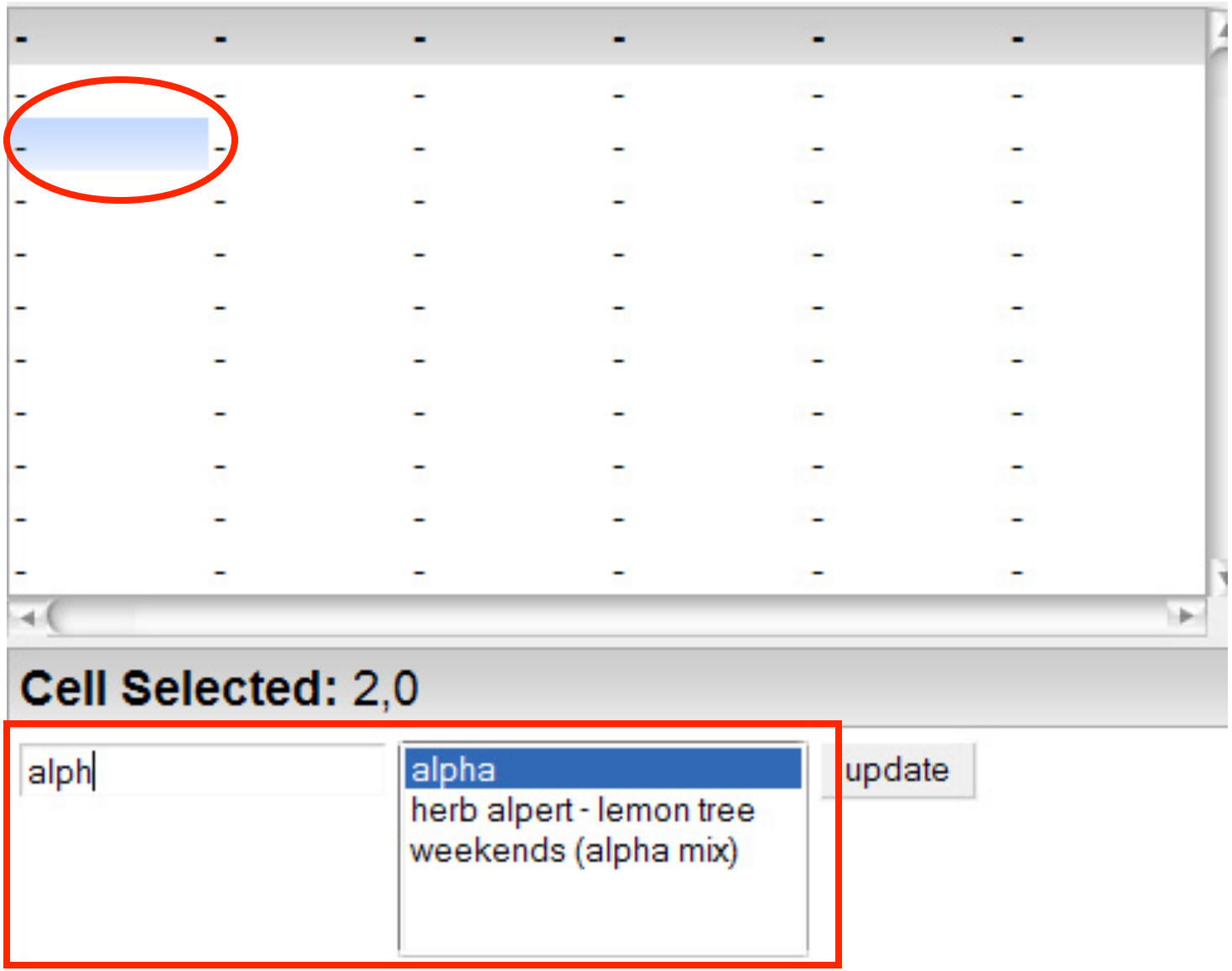
artist, album, song name, length



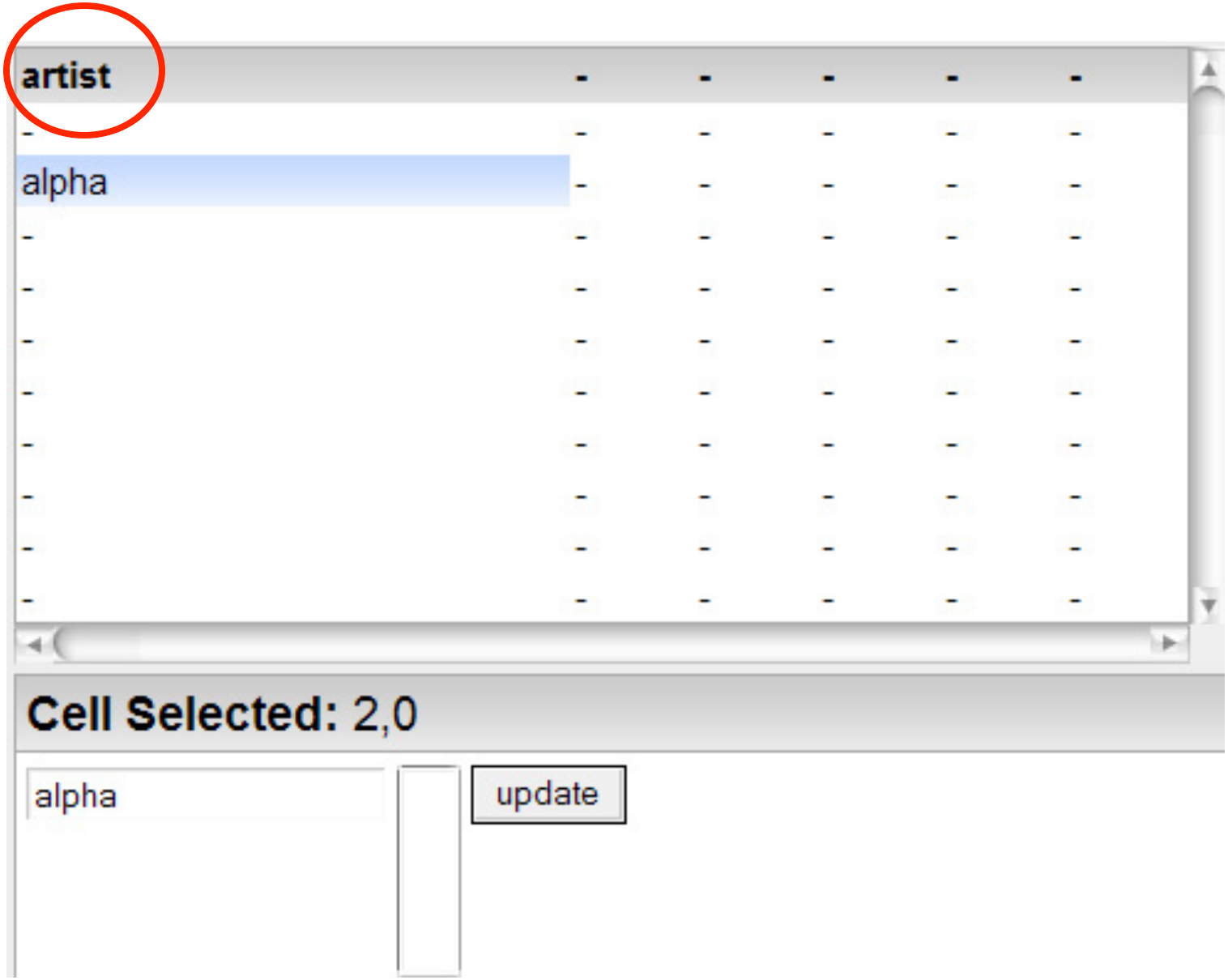
Review

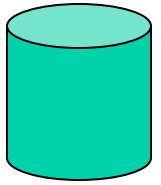
artist, song name, reviewer, favorite

Karma

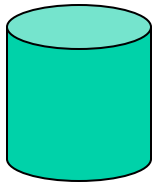

Cell Selected: 2,0
alph
alpha
herb alpert - lemon tree
weekends (alpha mix)
update

Karma



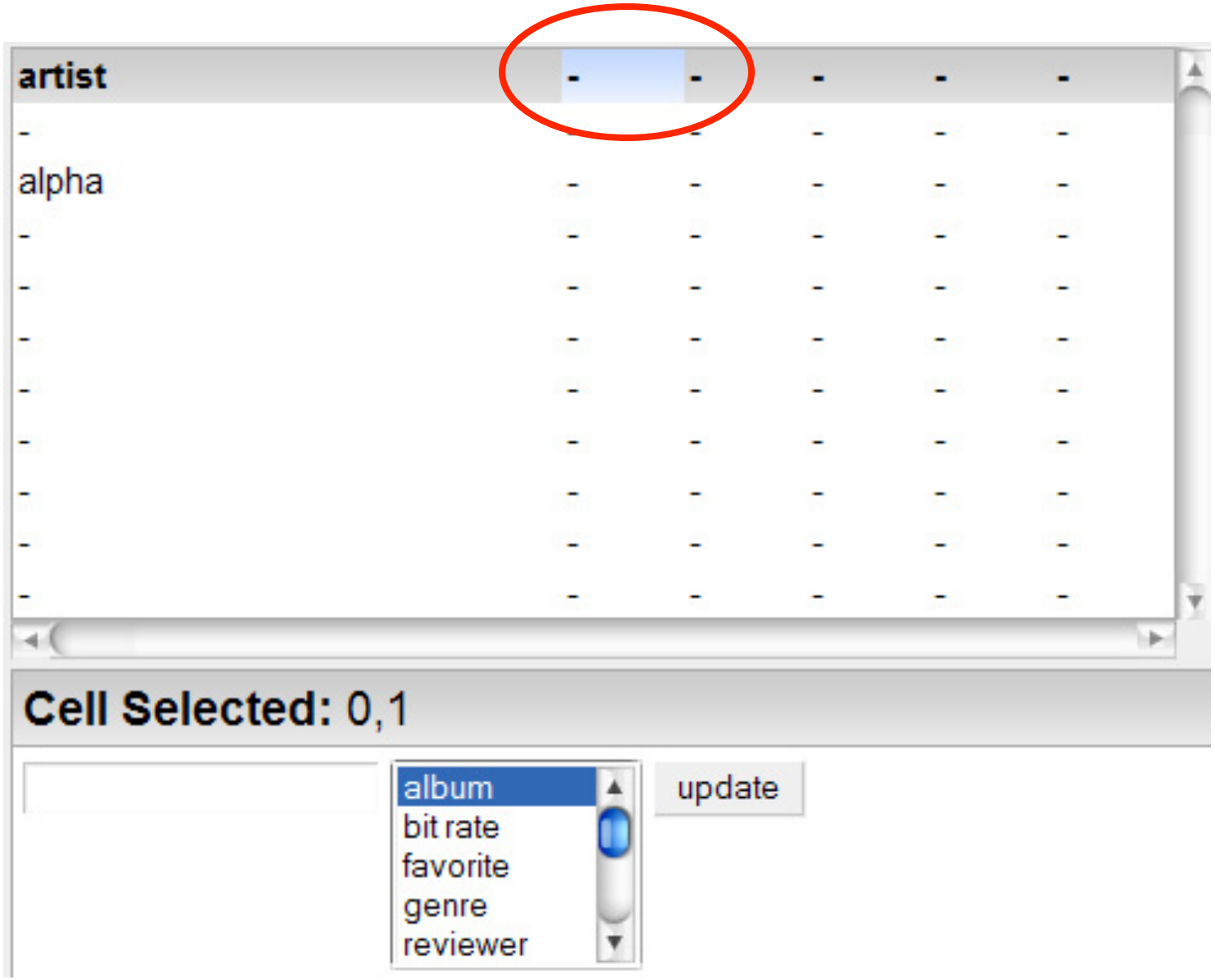


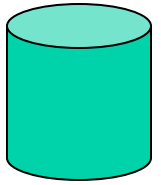
songs 2004



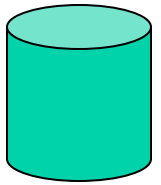
songs 2005

Karma



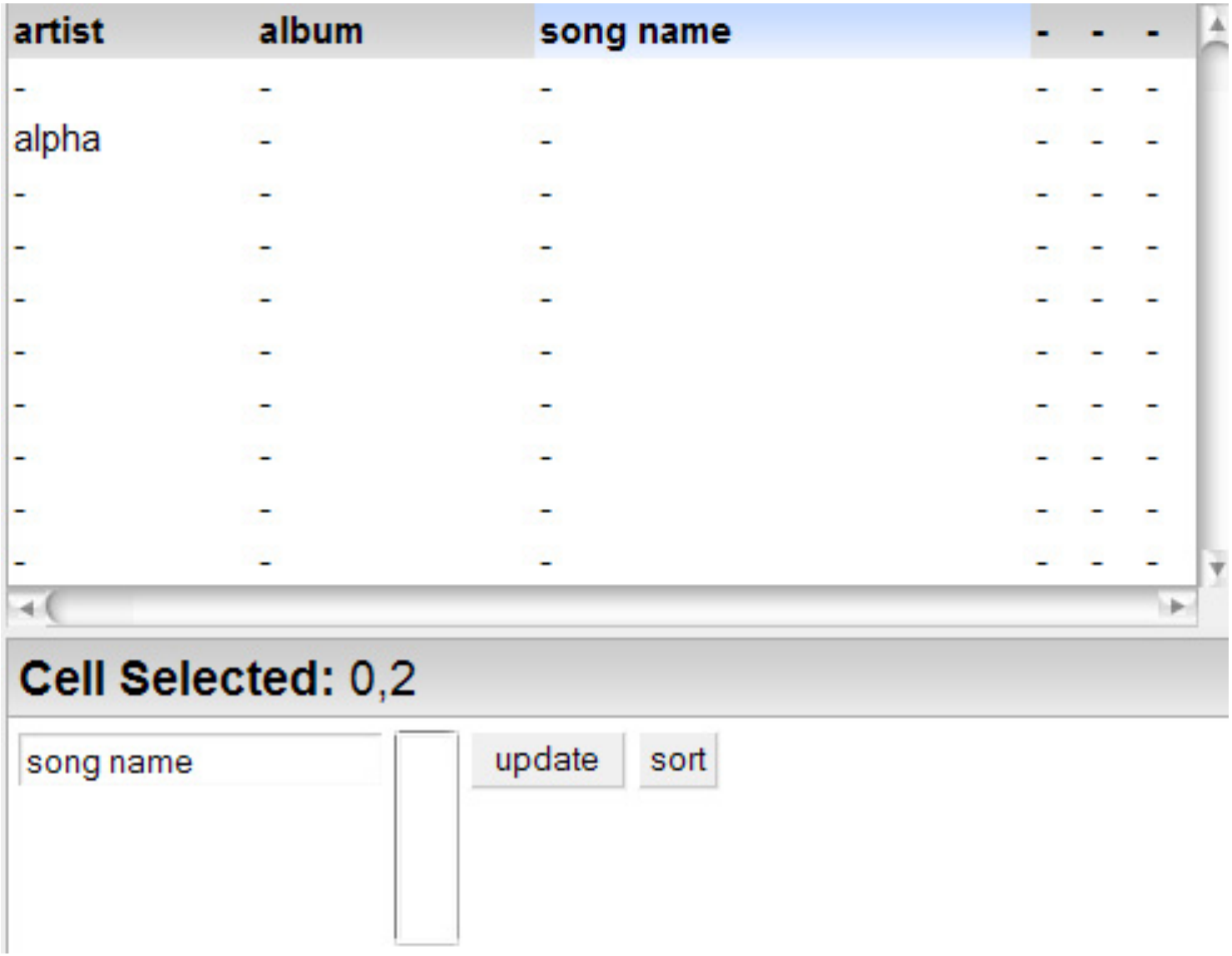


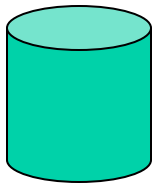
songs 2004



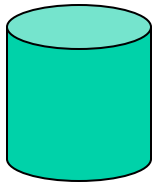
songs 2005

Karma



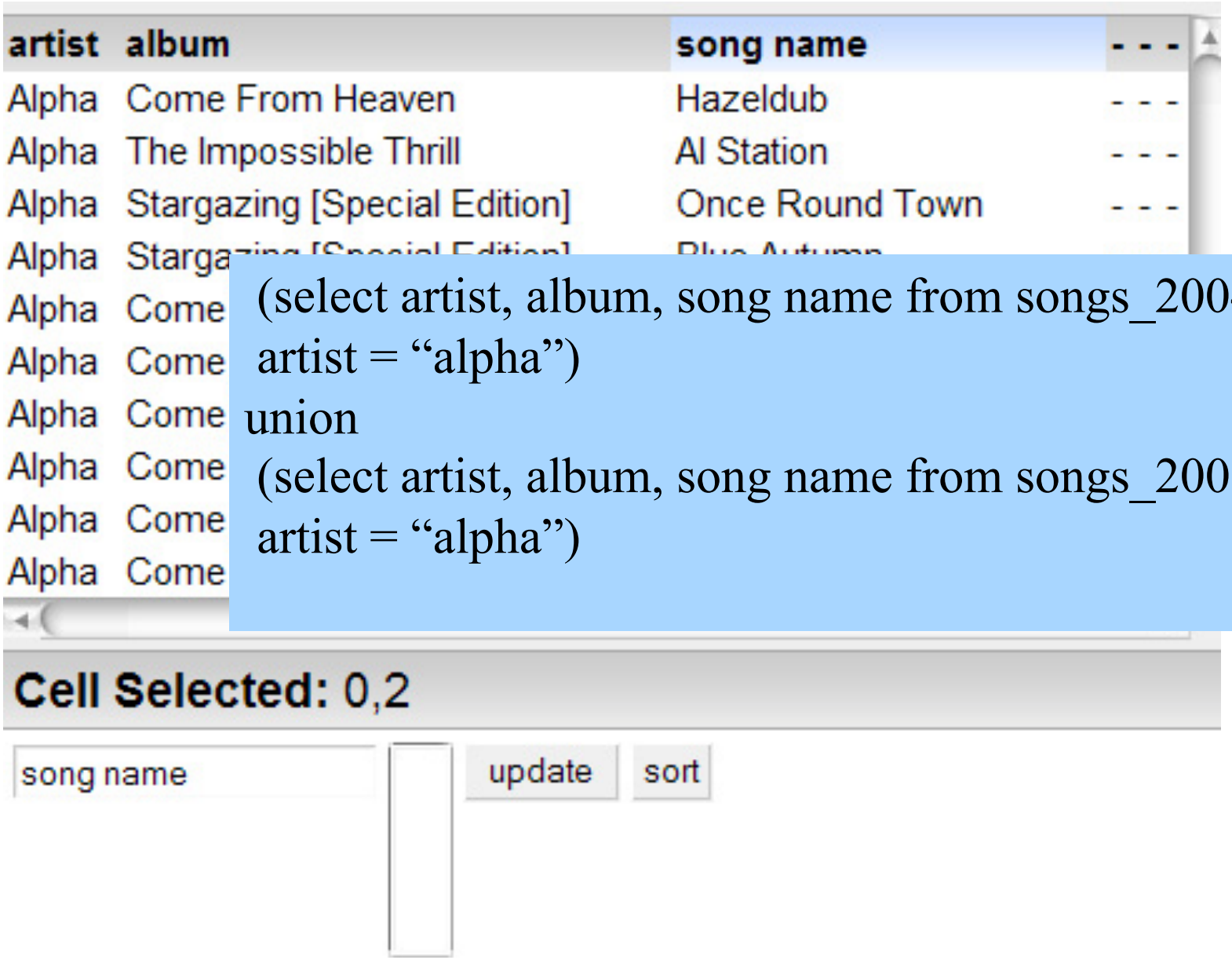


songs 2004

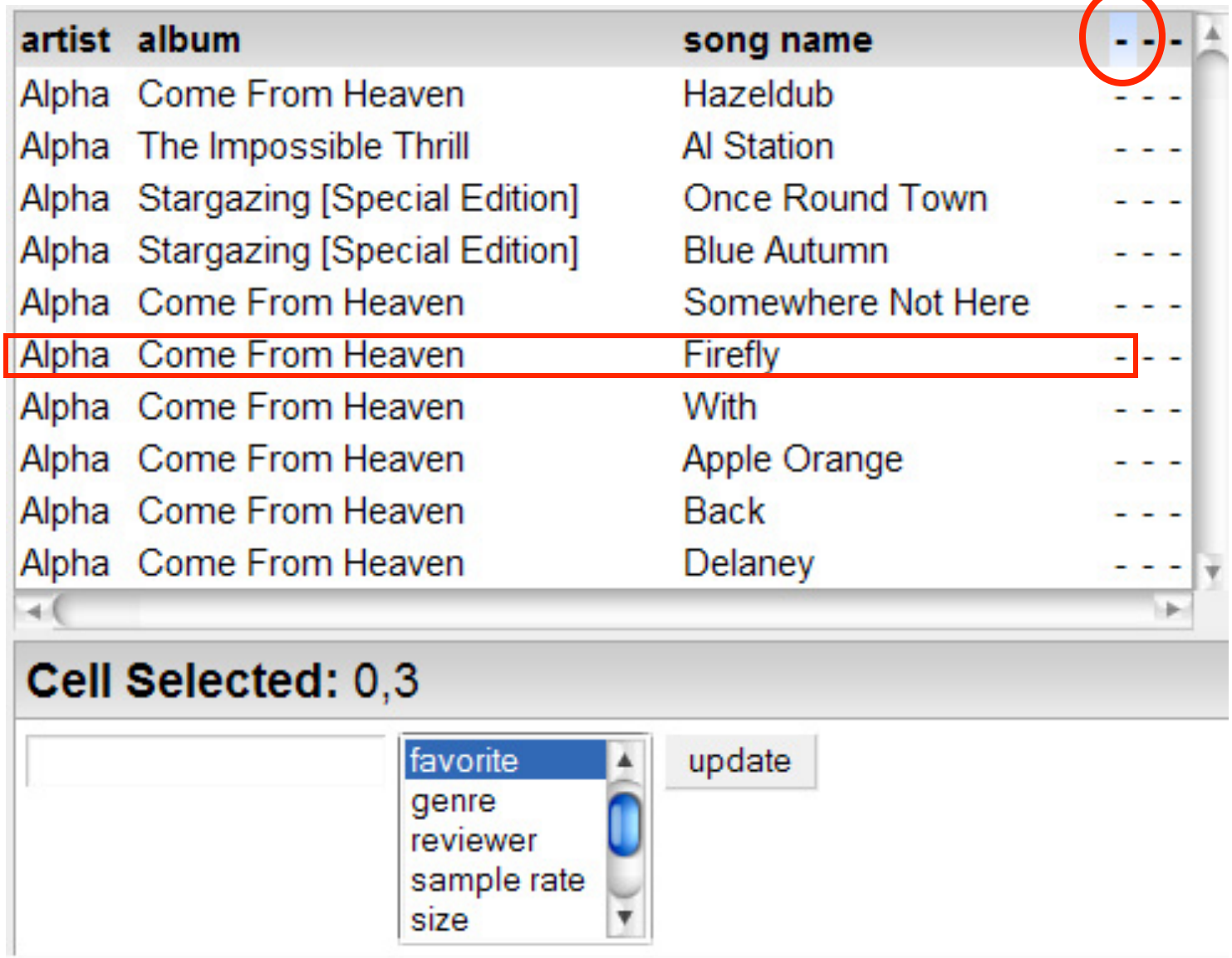


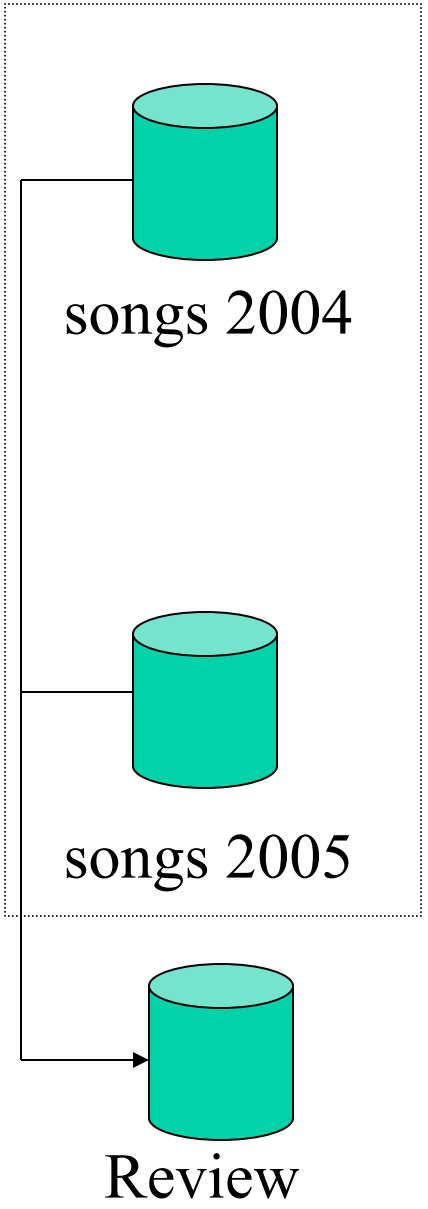
songs 2005

Karma

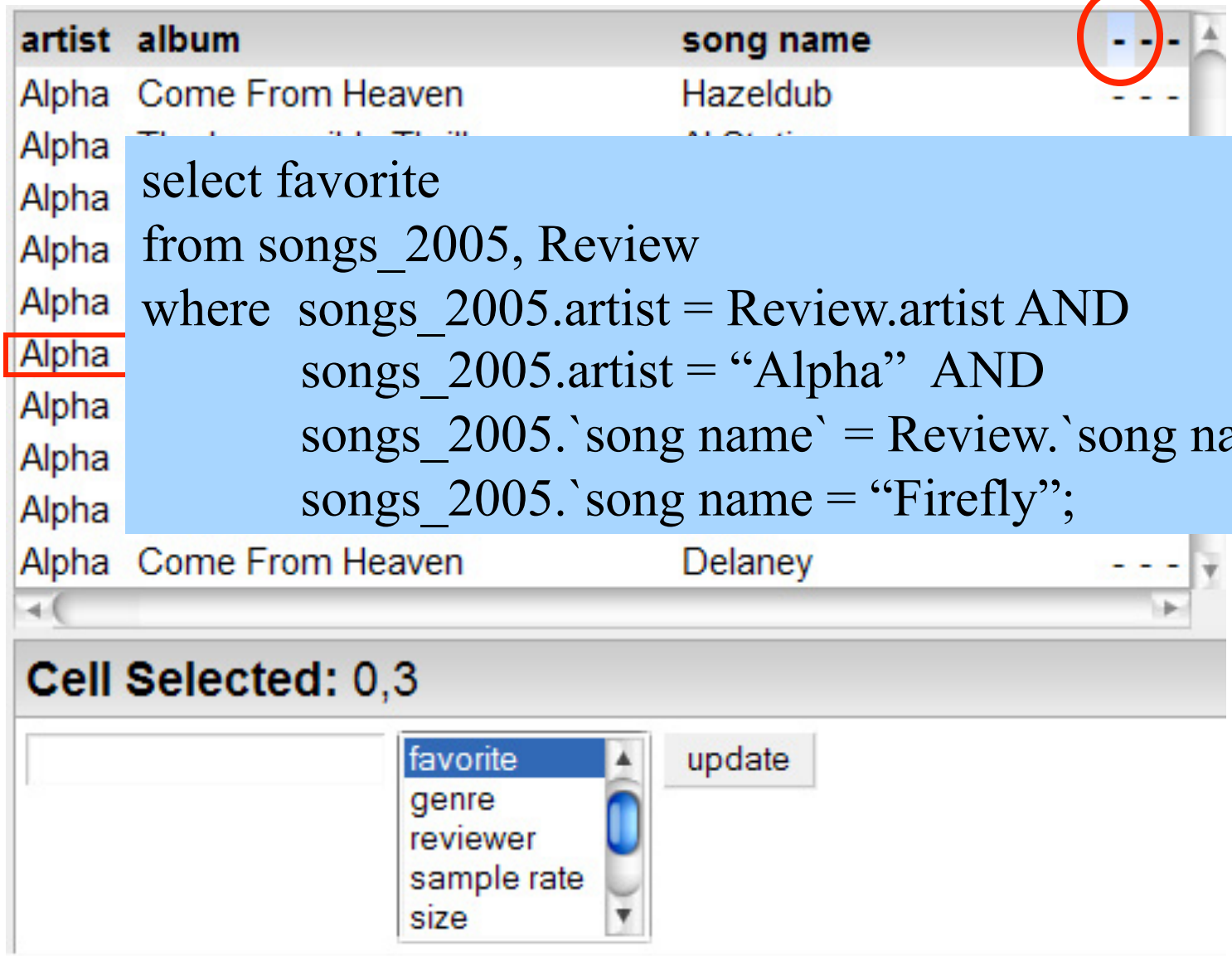

artist
album
song name
Alpha Come From Heaven Hazeldub
Alpha The Impossible Thrill Al Station
Alpha Stargazing [Special Edition] Once Round Town
(select artist, album, song name from songs_2004 where artist = “alpha”)
union
(select artist, album, song name from songs_2005 where artist = “alpha”)
Cell Selected: 0,2
song name
update
sort

Karma

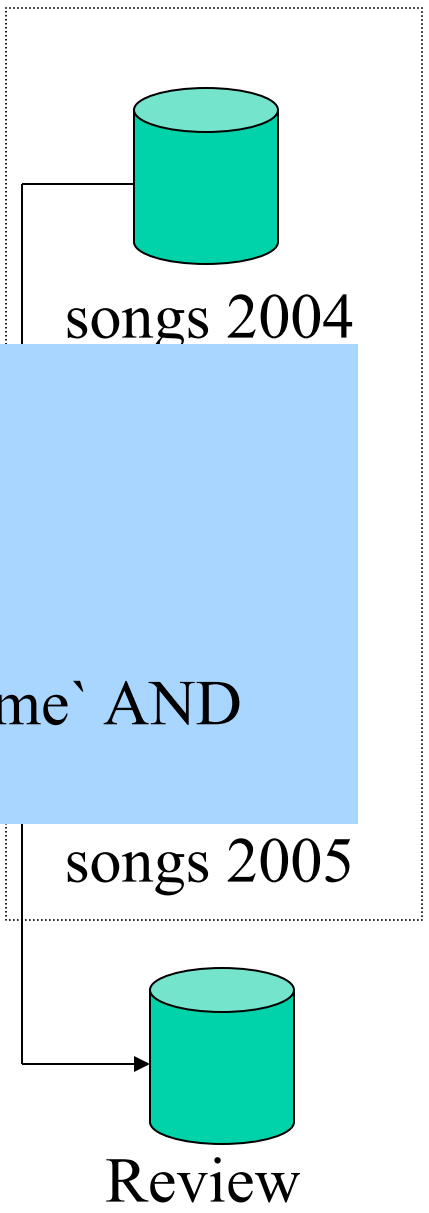

artist
album
song name
- - -
Alpha Come From Heaven Hazeldub - - -
Alpha The Impossible Thrill Al Station - - -
Alpha Stargazing [Special Edition] Once Round Town - - -
Alpha Stargazing [Special Edition] Blue Autumn - - -
Alpha Come From Heaven Somewhere Not Here - - -
Alpha Come From Heaven Firefly - - -
Alpha Come From Heaven With - - -
Alpha Come From Heaven Apple Orange - - -
Alpha Come From Heaven Back - - -
Alpha Come From Heaven Delaney - - -
Cell Selected: 0,3
favorite
genre
reviewer
sample rate
size
update


songs 2004
songs 2005
Review

Karma




```
select favorite
from songs_2005, Review
where  songs_2005.artist = Review.artist AND
       songs_2005.artist = "Alpha"  AND
       songs_2005.`song name` = Review.`song name` AND
       songs_2005.`song name = "Firefly";
```



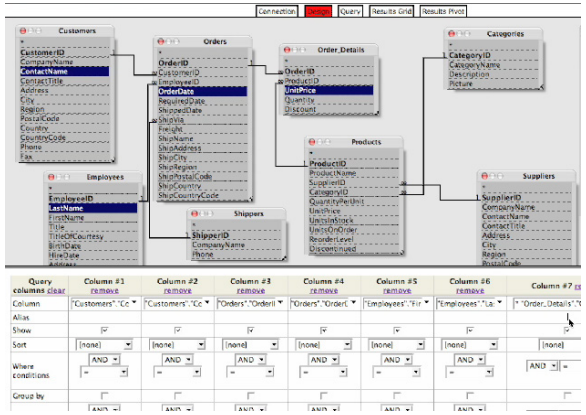
Advantages

- No query writing
- Hides data sources from users
- Never produces empty result

Open Issues

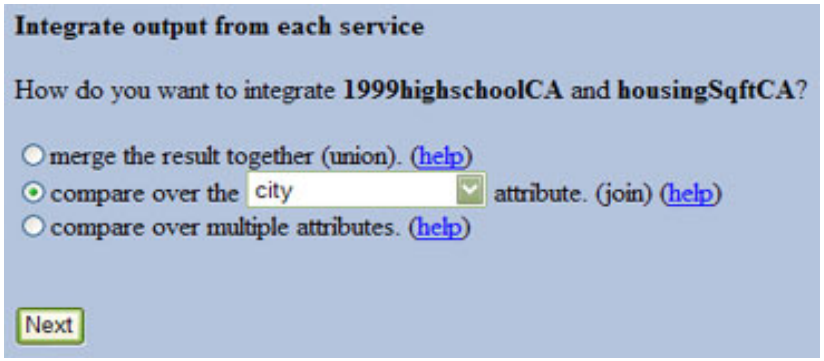
- Foreign Key Requirement
- Scalability to hundreds of sources
  - Tested with 5 sources (3000 rows)
- Filtering

Related Work

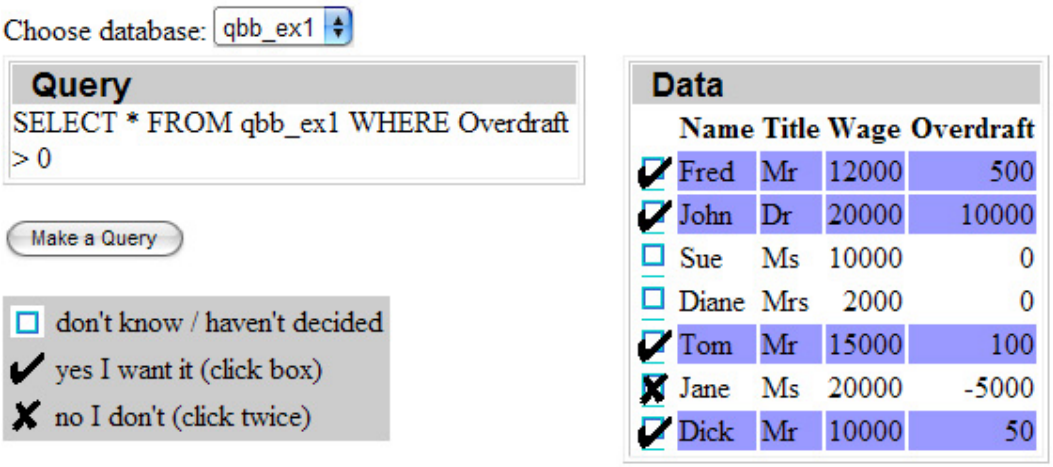


QBE [Zloof 1975]

Programming by Demonstration [Cypher 1993,Lau 2001]

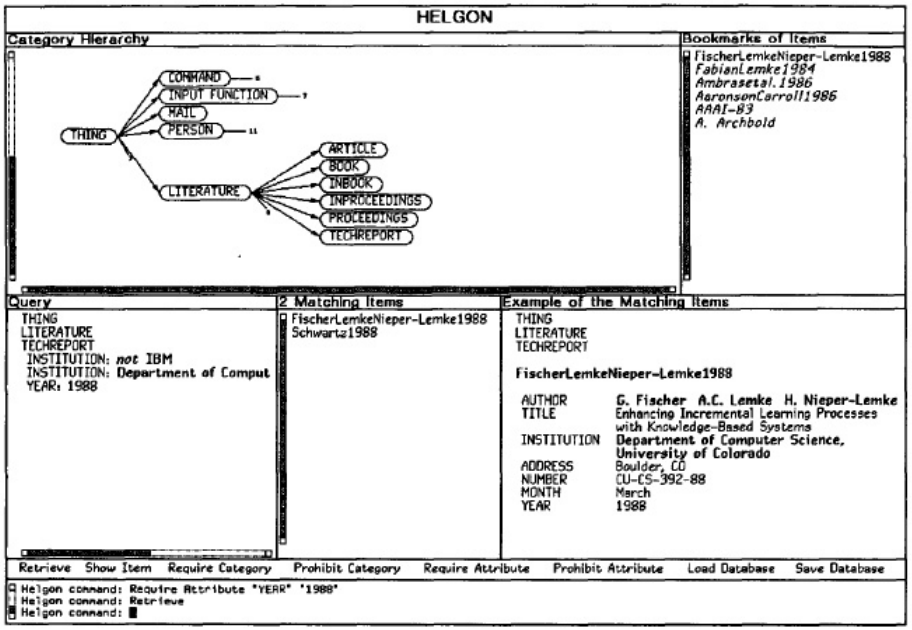


Agent Wizard [Tuchinda 2004]



Query By Browsing [Dix 1998]

Related Work

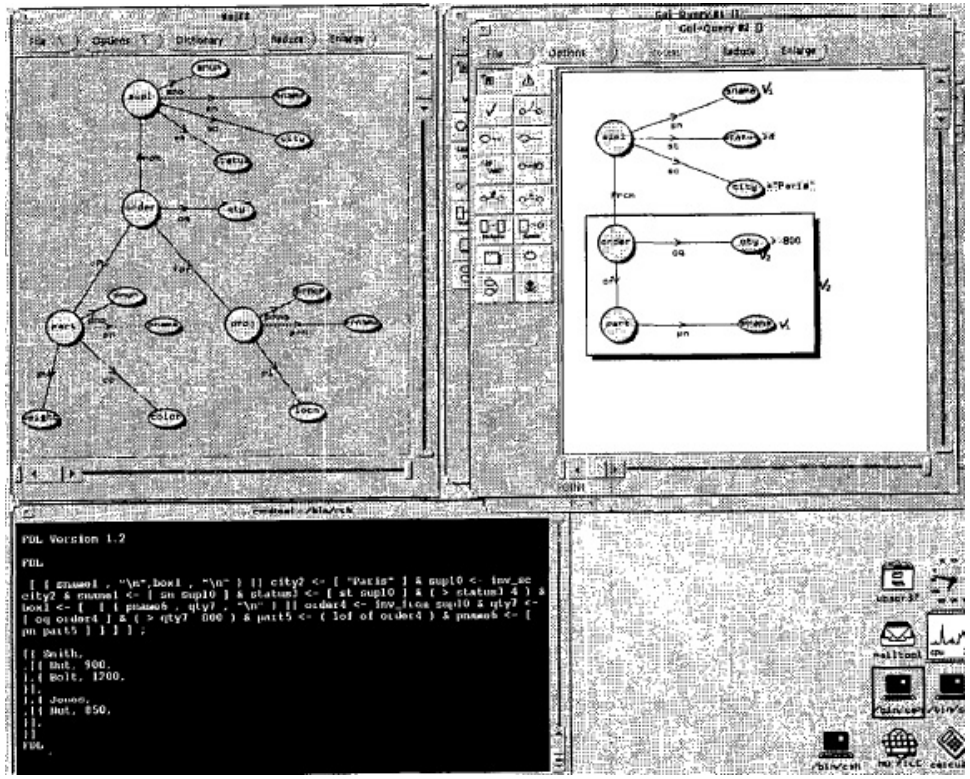


HELGON

Retrieval by formulation

HELGON[Fischer 1989]

RABBIT[Williams 1982]



Gql

Graphical Query Language

[Benzi 1998,

Haw 1994,

Papantonakis 1988]

Evaluation

	Clicks (c) and Key Strokes (k)	Cost
QBE A	28c+16k	4a+2t+d
Karma A	17c+4k	3a+2t
QBE B	39c+28k	5a+3t+2d
Karma B	25c+7k	3a+3t
QBE C	78c+54k	2*(5a+6t+2d)
Karma C	37c+14k	3a+6t

Typing in a value or Selecting a value = 1t unit

Selecting a data source to use = 1d unit

Selecting an attribute = 1a unit

Conclusion and Future Work

- Our contribution: An approach to data integration that
  - Does not require the user to know details about query writing, data sources, or existing values
  - Suggest valid possible values to the user
- In Progress: Filtering
- User studies