

# Best-Effort Modeling of Structured Data on the Web

Alon Halevy

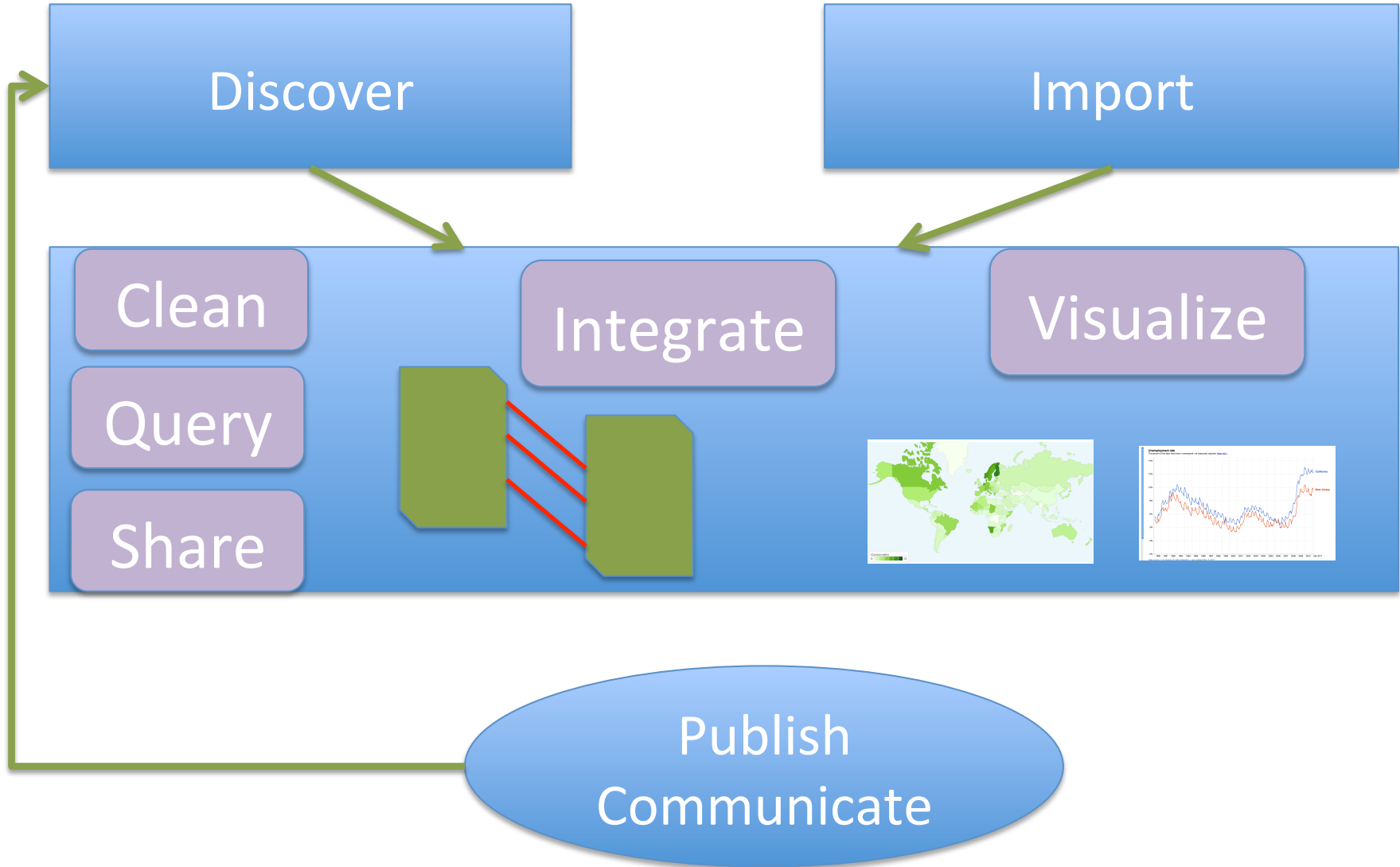
Google

November 2, 2011

# Structured Data and the Web

- A huge amount of structured data on the Web
  - Reference data, hobbies, products, coffee, ...
  - *Reacting to existing data*
- A platform for getting more data out:
  - Government data, crime, water conditions, ...
  - *Being proactive*
- New kinds of data collection and management
  - Collaboration, crowd-sourcing, real-time data, crisis response, ...
  - *Inventing a bright future*

# Goal: Structured Data Ecosystem



# Outline

- Google Fusion Tables:
  - A Database management service for the Web
- WebTables:
  - Discovering a (structured) needle in an (unstructured) haystack
- Observations about modeling along the way

# Fusion Tables

[google.com/fusiontables](http://google.com/fusiontables)

- Goal: an easy-to-use database system that is integrated with the Web.
- Key features:
  - **Easy** upload (CSV, KML, spreadsheets)
  - **Sharing** (even outside your company)
  - **Visualizations** front and center
  - Easy **publishing**
- Goal 2: a data cloud -- discover others' data and combine with yours.

# Coffee Consumption Per Capita

International Coffee Organization



































Get link

Share

File View Edit Visualize Merge

Showing all rows [options](#)

1 - 100 of 182 [Next »](#)

Country ▾	Coffee Consumption			
Finland	12			
Solomon Islands	11.8			
Norway	9.9			
Namibia	9			
Denmark	8.7			
Netherlands	8.4			
Sweden	8.2			
Switzerland	 	7.9		
Bahrain	6.8			
Belgium	6.8			
Luxembourg	6.8			
Brunei	6.6			
Canada	6.5			
Germany	6.4			
Qatar	6.4			

merge

place ▾	latitude ▾
1	-1.891708
2	-1.819305
3	-1.766459
4	-2.226917
5	-2.011286
6	-1.996361
7	-1.776754
8	-2.209764
9	-1.813814
11	-1.721505
12	-2.743473
13	-1.700571
14	-2.562393
15	-1.727338
16	-1.75136
17	-2.551932
18	-2.059664
19	-2.521578

### Share this table ✕

#### Invite people ?

- As viewers**  
Can see and comment on the data
- As collaborators**  
Can also edit the data
- As owners**  
Can also invite people to view or collaborate

Separate email addresses with commas.

- Send email invitations**

#### Visibility options ?

- Public**  
Everyone on the internet can find and access. No sign-in required.
- Unlisted**  
Anyone who has the link can access. No sign-in required.
- Private**  
Only people explicitly granted permission can access. Sign-in required.

**Viewers (0)**  
**Collaborators (0)**  
**Owners (1)**  
*halevy@google.com*

# Coffee Consumption Per Capita International Coffee Organization

Get link








































Share

File View Edit **Visualize** Merge

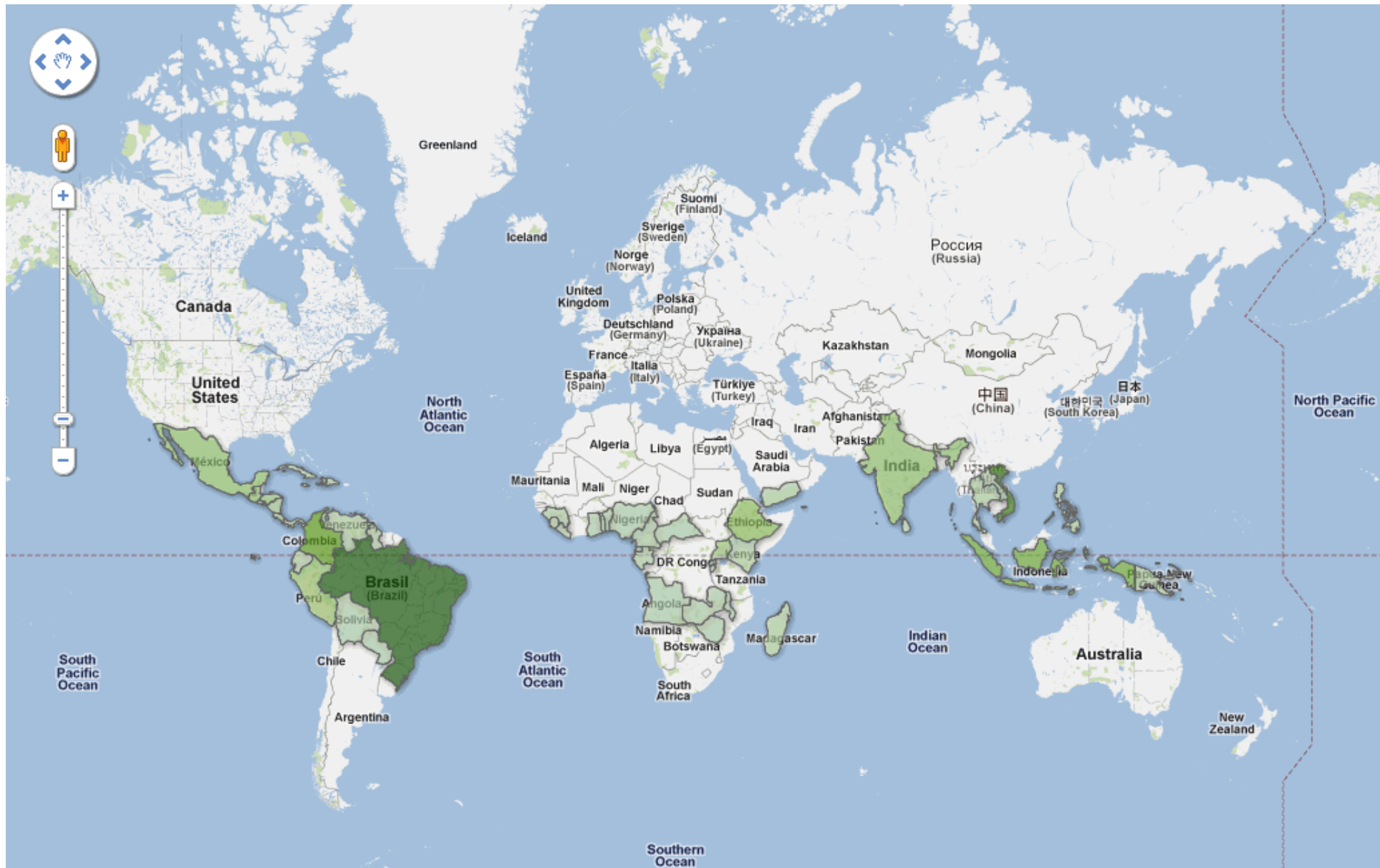
- Table
- Map
- Intensity map
- Line
- Bar
- Pie
- Scatter
- Timeline (date, text, number)

Showing all rows

1 - 100 of 182 [Next »](#)

Country ▾	Coffee Consumptiic	Order ▾			
Finland	12	1			
Solomon Islands	11.8	2			
Norway	9.9	3			
Namibia	9	4			
Denmark	8.7	5			
Netherlands	8.4	6			
Sweden	8.2	7			
Switzerland	7.9	8			
Bahrain	6.8	9			
Belgium	6.8	10			
Luxembourg	6.8	11			
Brunei	6.6	12			
Canada	6.5	13			
Germany	6.4	14			
Chad	6.1	15			
Austria	6.1	16			
Samoa	6	17			
Italy	5.9	18			





# DATA BLOG

Facts are sacred

## Wikileaks Iraq war logs: every death mapped

The Wikileaks Iraq war logs provide us with a unique picture of every death in Iraq. These are those events mapped using Google Fusion tables

- [Download the data from the Datablog](#)

Tweet 1,706

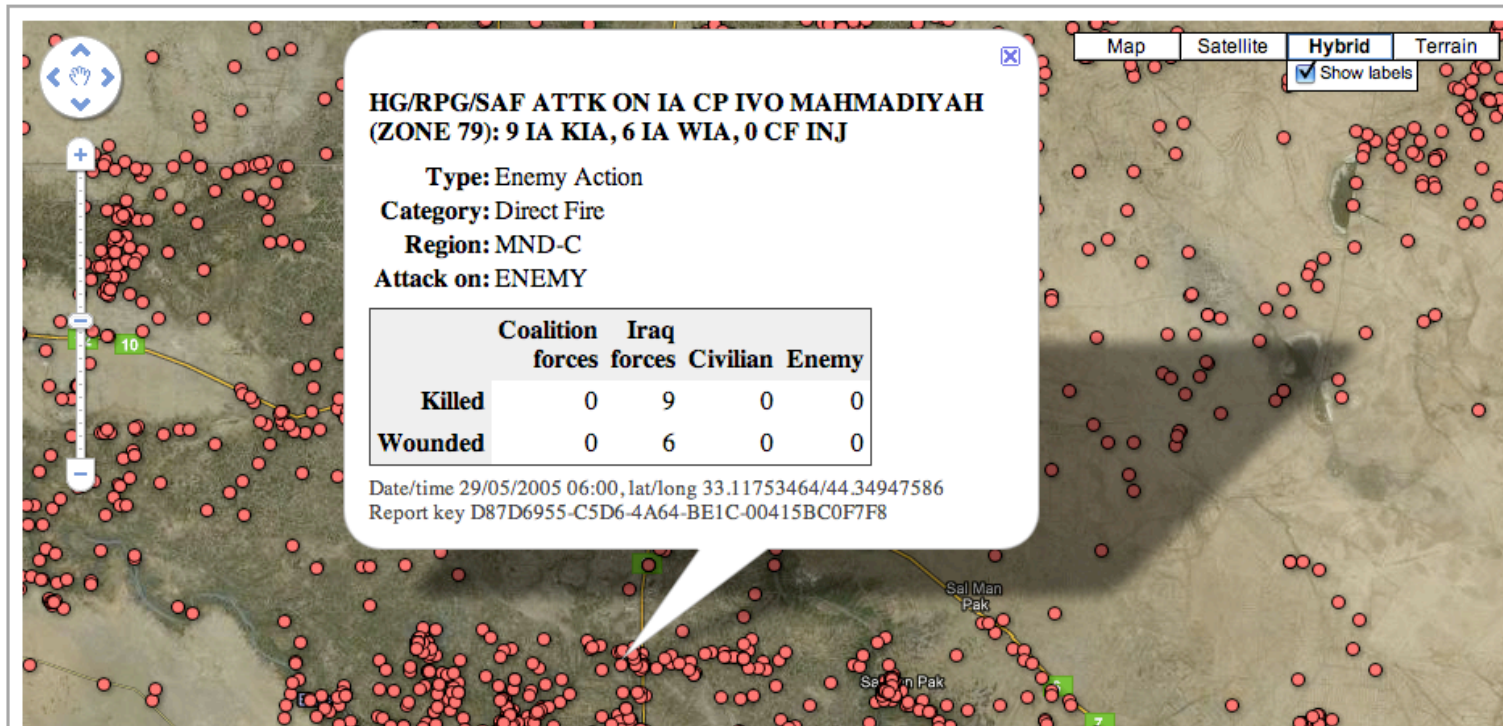
Share 8952



Comments (85)

Simon Rogers

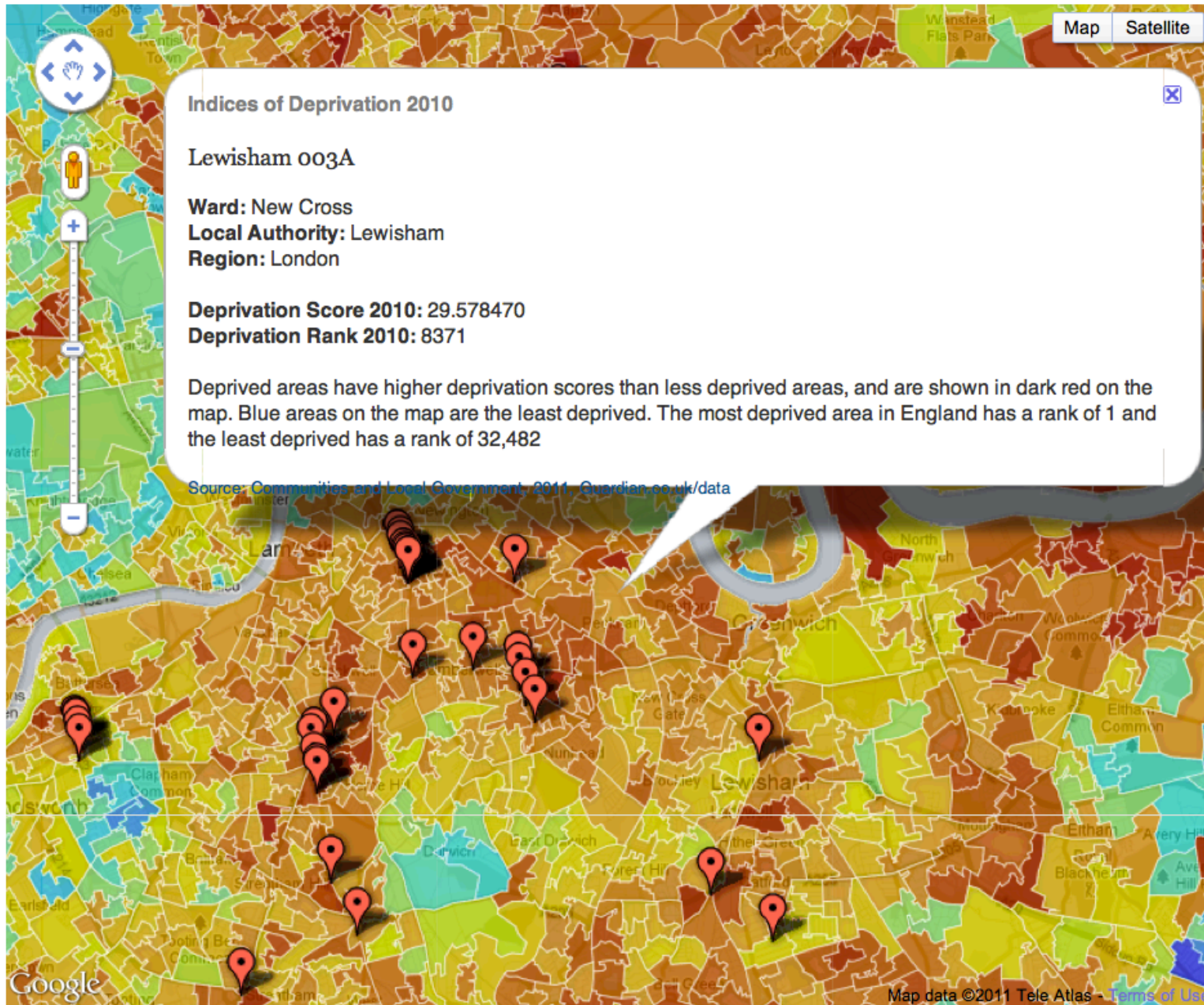
guardian.co.uk, Saturday 23 October 2010 00.05 BST







**Simon Rogers**  
guardian.co.uk, Wednesday 10 August 2011 08.00 BST



# America in 2010

The latest census report underscores a decades-long demographic shift, as people and jobs have migrated away from the northern and central U.S. toward the Sunbelt. There's also a transformation under way in the country's racial and ethnic makeup as the Hispanic population surges. See change in population and total population by race across the U.S.

Hispanic

White

Black

Asian

Percent Change

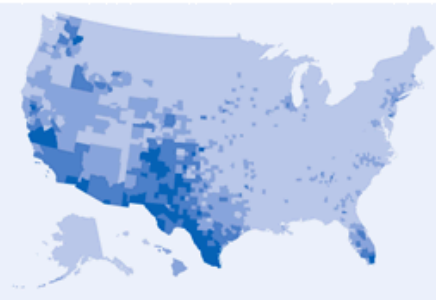


Show:

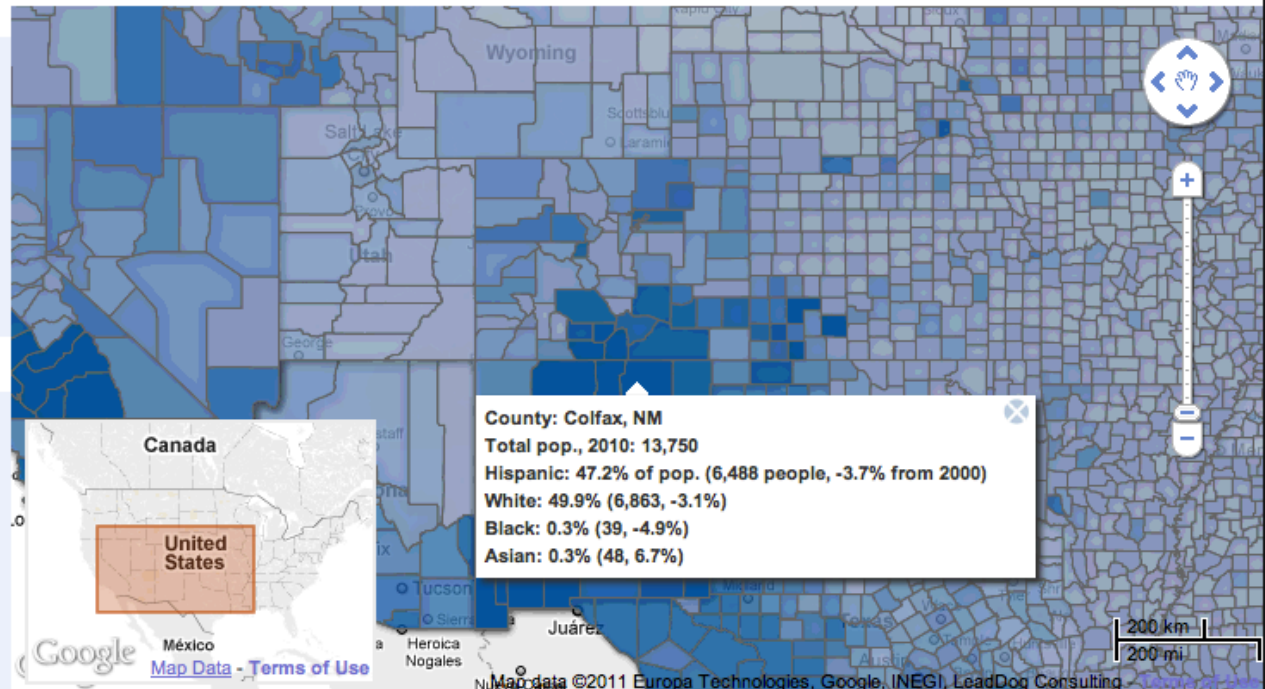
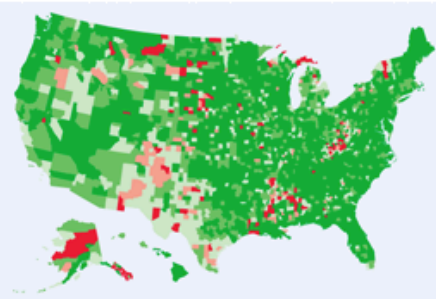
- Hispanic pop. (%)
- Hispanic pop. (% chg)

Latinos now constitute 16% of the nation's total population, with the largest proportion living in the Mexico border areas of Texas, Arizona, New Mexico and California.

Racial presence, % of total population



Change in population, 2000 to 2010



Map data ©2011 Europa Technologies, Google, INEGI, LeadDog Consulting  
 Source: U.S. Census; Produced by: Paul Antonson, Megan Ballinger, Susan McGregor, Renee Rigdon

Email

Like

16 people like this. Be the first of your friends.

Share:



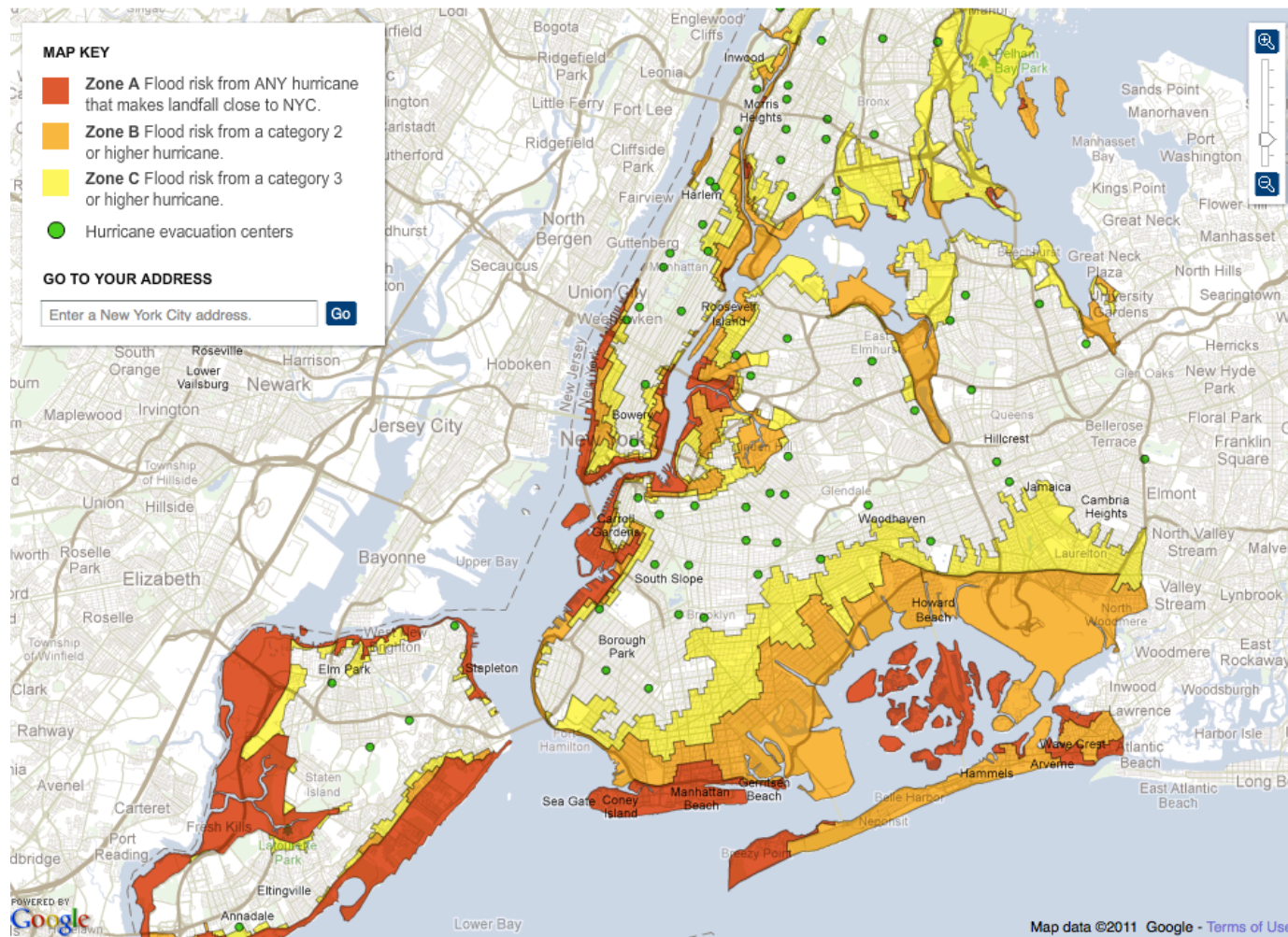
THE WALL STREET JOURNAL.



Published: August 26, 2011

## New York City Hurricane Evacuation Zones

There are three evacuation zones in New York City that are based on the strength of the hurricane making land mandatory evacuation of Zone A, plus areas of the Rockaways that are in Zone B, by 5 p.m. on Saturday. Related



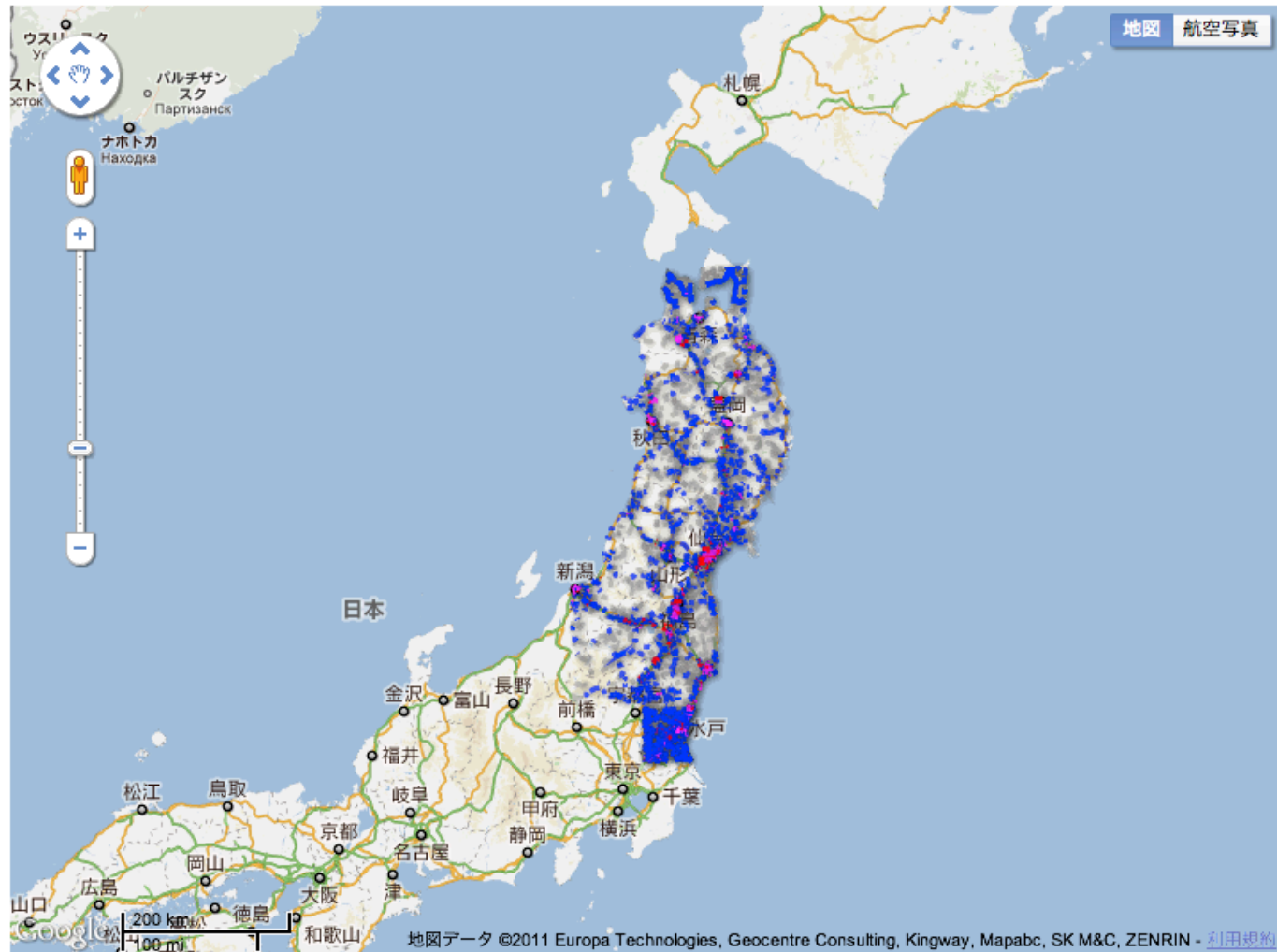
# Google Crisis Response 自動車通行実績情報マップ

a google.org project

[東日本大震災](#)、自動車通行実績情報マップ

下記マップ中において、前日の午前 6 時～午前 10 時に渋滞が発生していた道路を**赤色**、同時間帯に混雑が発生していた道路を**桜色**、上記以外の道路で前日の 0 時～24 時に通行実績のあった道路を**青色**、通行実績情報がなかった道路を**灰色**で表示しています。（最終更新日時：2011/08/17 08:14 JST）

住所を入力して検索:



地図データ ©2011 Europa Technologies, Geocentre Consulting, Kingway, Mapabc, SK M&C, ZENRIN - [利用規約](#)

提供：本田技研工業（株）、パイオニア（株）

## Global earthquake activity since 1973 and nuclear power plant locations



Enable earthquake heatmap

[maptd.com](http://maptd.com)



# Crowd Sourcing

otrobache.com <sup>(SPAIN)</sup>

BACHES

SABER MÁS

Que nosotros sepamos,  
hay **1258 baches** en España



Municipios cerca

pozuelo de alarcón

madrid

マドリッド

majadahonda

las rozas de madri

[Otros lugares](#) →

**+ Reporta un nuevo bache**



# A GIS in the Cloud

- That's not what we set out to do, really.
- Challenges:
  - Trickle: show only a small number of features (points, polygons) from a large data set
  - Need to thin polygons, clip to the window
  - Style features on the fly
  - All in less than 100ms

# And the Credit Goes to...

- Hector Gonzalez
- Jayant Madhavan
- Sree Balakrishnan
- Heidi Lam
- Hongrae Lee
- Warren Shen
- Anno Langen
- Rebecca Shapley
- Anish Das Sarma
- Boulos Harb
- Fei Wu
- Cong Yu
- Spiros Papadimitriou

# Outline

- ✓ Google Fusion Tables:
  - A Database management service for the Web
- *WebTables:*
  - *Discovering a (structured) needle in an (unstructured) haystack*

***Discovery = incentive to publish***

# Tables on the Web

<a href="#">A</a>	<a href="#">B</a>	<a href="#">C</a>	<a href="#">D</a>	<a href="#">E</a>	<a href="#">F</a>	<a href="#">G</a>	<a href="#">H</a>	<a href="#">I</a>	<a href="#">J</a>	<a href="#">K</a>	<a href="#">L</a>	<a href="#">M</a>	<a href="#">N</a>	<a href="#">O</a>	<a href="#">P</a>	<a href="#">Q</a>	<a href="#">R</a>	<a href="#">S</a>	<a href="#">T</a>	<a href="#">U</a>	<a href="#">V</a>	<a href="#">W</a>	<a href="#">X</a>	<a href="#">Y</a>	<a href="#">Z</a>
<a href="#">African-Americans</a>							<a href="#">Artists</a>	<a href="#">Explorers of the US</a>					<a href="#">Inventors</a>			<a href="#">US Presidents</a>				<a href="#">US Symbols</a>			<a href="#">US States</a>		



[President's Day Activities](#)

[EnchantedLearning.com](#)

## The Presidents of the United States of America

[In the order in which they served](#)

[Alphabetical order](#)

[Short table of Data](#)



[Abraham Lincoln](#)

The President and Vice-President are elected every four years. They must be at least 35 years of age, they must be native-born citizens of the United States, and they must have been residents of the U.S. for at least 14 years. (Also, a person cannot be elected to a third term as President.)

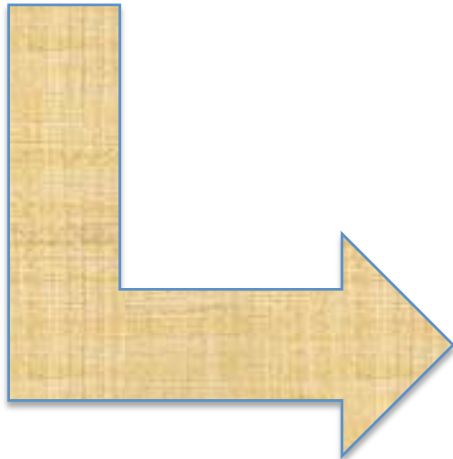
President	Party	Term as President	Vice-President
1. <a href="#">George Washington</a> (1732-1799)	None, Federalist	1789-1797	<a href="#">John Adams</a>
2. <a href="#">John Adams</a> (1735-1826)	Federalist	1797-1801	<a href="#">Thomas Jefferson</a>
3. <a href="#">Thomas Jefferson</a> (1743-1826)	Democratic-Republican	1801-1809	Aaron Burr, George Clinton
4. <a href="#">James Madison</a> (1751-1836)	Democratic-Republican	1809-1817	George Clinton, Elbridge Gerry
5. James Monroe (1758-1831)	Democratic-Republican	1817-1825	Daniel Tompkins
6. <a href="#">John Quincy Adams</a> (1767-1848)	Democratic-Republican	1825-1829	John Calhoun
7. Andrew Jackson (1767-1845)	Democrat	1829-1837	John Calhoun, Martin van Buren
8. Martin van Buren (1782-1862)	Democrat	1837-1841	Richard Johnson
9. William H. Harrison (1773-1841)	Whig	1841	John Tyler
10. John Tyler (1790-1862)	Whig	1841-1845	
11. James K. Polk (1795-1849)	Democrat	1845-1849	George Dallas
12. Zachary Taylor (1784-1850)	Whig	1849-1850	Millard Fillmore
13. Millard Fillmore (1800-1874)	Whig	1850-1853	
14. Franklin Pierce (1804-1869)	Democrat	1853-1857	William King
15. James Buchanan (1791-1868)	Democrat	1857-1861	John Breckinridge

# Goal: Search for Structured Data



## Challenges:

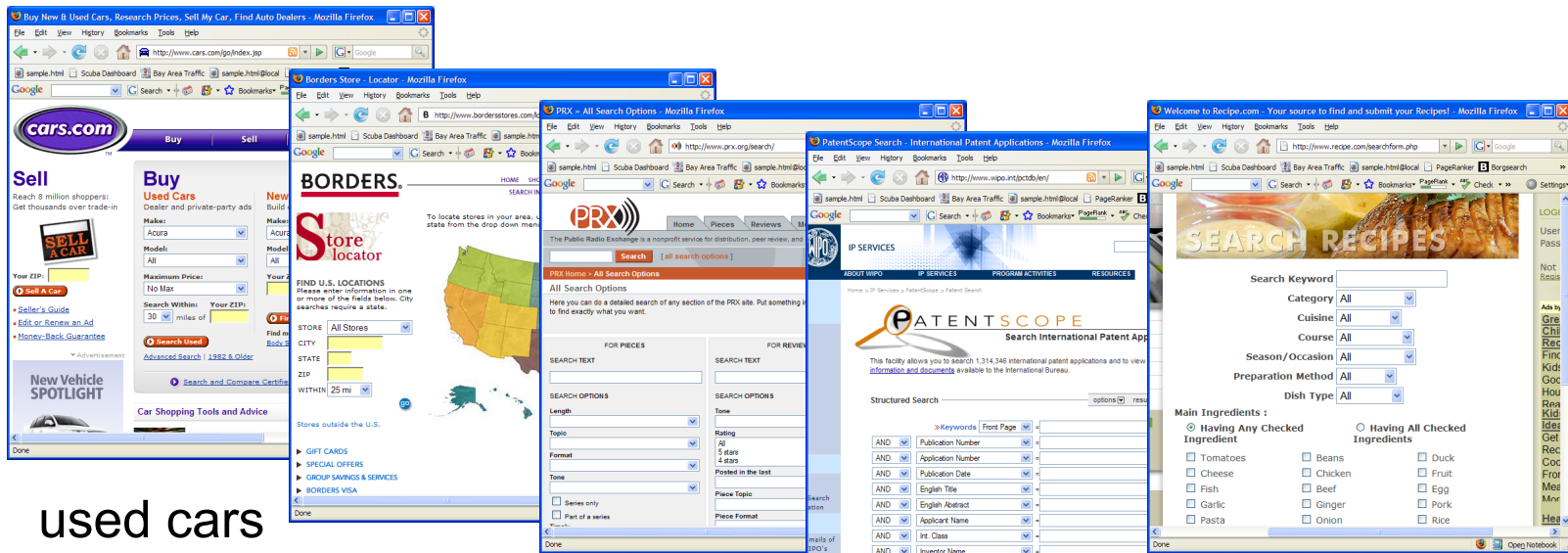
- Finding the good tables on the Web
- Understanding their semantics
- Understanding user's intentions



## Rabbits:

Genus	Sub-Genus	Species: Common Name
Brachylagus		Pygmy Rabbit
Bunolagus		Riverine Rabbit, Bushman Rabbit, Bushman Hare
Nesolagus		Sumatran Striped Rabbit, Sumatra Short-eared Rabbit, SUmatran
		Annamite Striped Rabbit
Oryctolagus		European Rabbit
Pentalagus		Amami Rabbit, Ryukyu Rabbit
Poelagus		The Bunyoro Rabbit
Romerolagus		Volcano Rabbit, Teporingo, Zacatuche
Sylvilagus	Tapeti	Swamp Rabbit

# The Deep Web



used cars

store locations

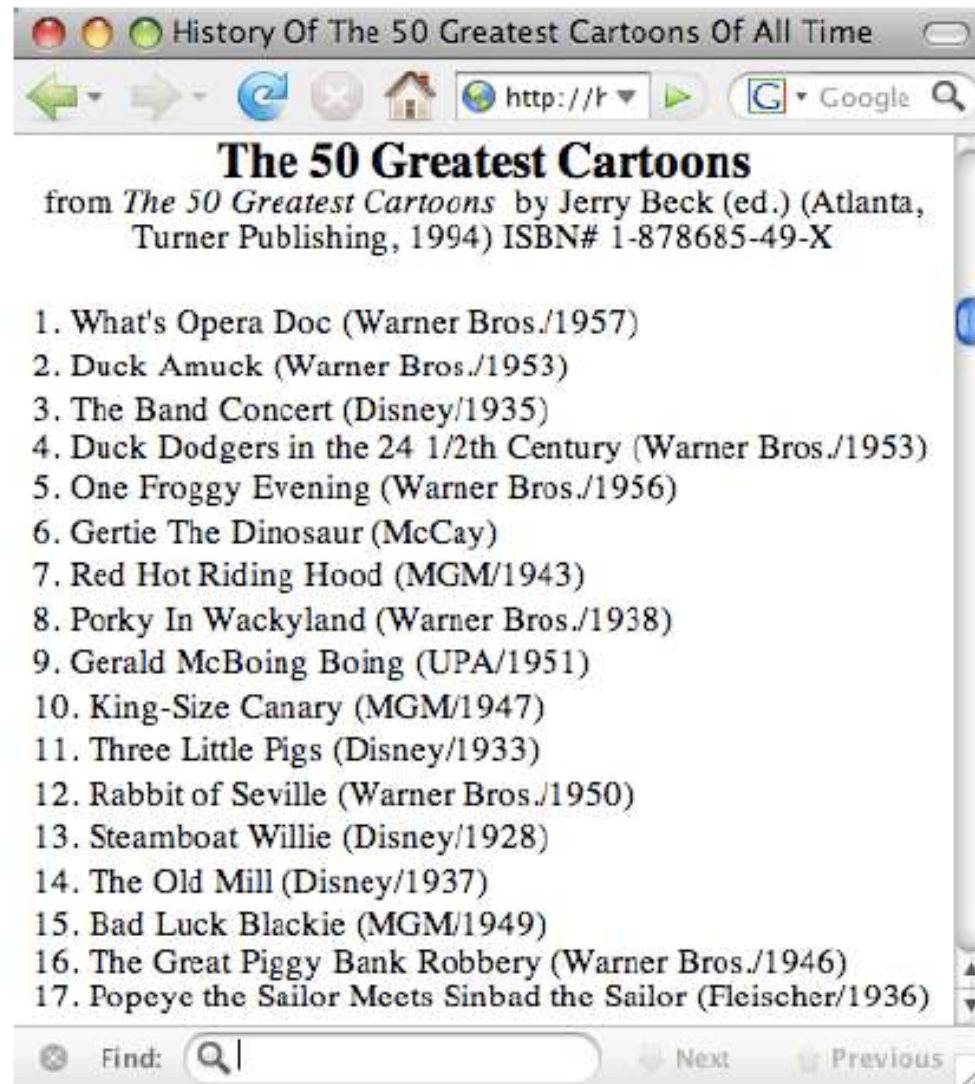
radio stations

patents

recipes

See “Google’s Deep Web Crawl”, VLDB 2008

# HTML Lists



The screenshot shows a web browser window with the title "History Of The 50 Greatest Cartoons Of All Time". The address bar shows "http://t" and the search engine is "Google". The main content is a list titled "The 50 Greatest Cartoons" from the book "The 50 Greatest Cartoons" by Jerry Beck (ed.) (Atlanta, Turner Publishing, 1994) ISBN# 1-878685-49-X. The list contains 17 items, each with a number, title, studio, and year. The browser's search bar at the bottom contains "Find:" and "Next Previous" buttons.

**The 50 Greatest Cartoons**  
from *The 50 Greatest Cartoons* by Jerry Beck (ed.) (Atlanta, Turner Publishing, 1994) ISBN# 1-878685-49-X

1. What's Opera Doc (Warner Bros./1957)
2. Duck Amuck (Warner Bros./1953)
3. The Band Concert (Disney/1935)
4. Duck Dodgers in the 24 1/2th Century (Warner Bros./1953)
5. One Froggy Evening (Warner Bros./1956)
6. Gertie The Dinosaur (McCay)
7. Red Hot Riding Hood (MGM/1943)
8. Porky In Wackyland (Warner Bros./1938)
9. Gerald McBoing Boing (UPA/1951)
10. King-Size Canary (MGM/1947)
11. Three Little Pigs (Disney/1933)
12. Rabbit of Seville (Warner Bros./1950)
13. Steamboat Willie (Disney/1928)
14. The Old Mill (Disney/1937)
15. Bad Luck Blackie (MGM/1949)
16. The Great Piggy Bank Robbery (Warner Bros./1946)
17. Popeye the Sailor Meets Sinbad the Sailor (Fleischer/1936)



## Annual Competition Results:

### 2010 London



The week of 23-25 June 2010, London, England was the focus of the eyes of the specialty coffee world, both online and at the host event, Caffe Culture, held in association with the Speciality Coffee Association of Europe (SCAE).

It was another epic event as Mike Phillips of Chicago, USA, walked away with the 2010 crown. 53 national barista champions in total were cheered on by an audience of over one thousand people, with a total internet audience of 29,600 (unique IP address visitors) over the course of the 3-day event via our GoLive streaming video project.

2010 – London, United Kingdom

*(click on name for video)*

- 1st place: [Michael Phillips](#) (United States)
- 2nd Place: [Raul Rodas](#) (Guatemala)
- 3rd Place: [Scottie Callaghan](#) (Australia)
- 4th Place: [Colin Harmon](#) (Ireland)
- 5th Place: [Soren Stiller Markussen](#) (Denmark)
- 6th Place: [Stefanos Domatiotis](#) (Greece)

[Full results](#)

### 2009 Atlanta



The event that marked the 10 year celebration of WBC presented in Atlanta, Georgia, April 16-19th. Featuring the first time WBC Espresso Bar with interactive presentations and sampling of competition quality beverages.

Over 8000 fans, from the stage audience to the daily online blog and our live-stream followers, were all witness to Gwilym Davies from the United Kingdom, as he claimed the title of 2009 World Barista Champion.

Held in conjunction with "The Event" SCAA 22nd Annual Exposition.

2009 – Atlanta, Georgia, United States

*(click on name for video)*

- 1st place: [Gwilym Davies](#) (United Kingdom)
- 2nd Place: [Sammy Piccolo](#) (Canada)
- 3rd Place: [Mike Phillips](#) (United States)
- 4th Place: [Colin Harmon](#) (Ireland)
- 5th Place: [Lee Jong Hoon](#) (Korea)
- 6th Place: [Attila Molnar](#) (Hungary)

[Full results](#)



# The Needle in the Haystack

Finding high quality HTML tables

## Check-In



### Introductions *(1 Viewing)*

Glad you found OTz, now come on in and introduce yourself. Introducing yourself is the first step to finding a cure for boredom.

**Sub-Forums:** [Get To Know the Mods](#)



Threads: 2,011  
Posts: 45,772



### Forum How-to's *(2 Viewing)*

Learn how to use BBCode, upload a picture, change your signature and many other tips and tricks for forum use.



Threads: 97  
Posts: 1,457



### Board Rules & FAQ's

Here you will find announcements and frequently asked questions about new features on the forum etc. If there are any new topics here, please read them as they are important.



Threads: 219  
Posts: 5,865



### Guest Area

A place where guests (unregistered) members can post. Its the ONLY place they can post. Helpful if your having problems registering or you wanna flame away because you got slapped in the face with a ban stick.



Threads: 86  
Posts: 1,681



### Feedback & Suggestions

Got an idea or think something could be better? Let us know. We're always looking to improve, help make OTz the place to be.



Threads: 591  
Posts: 9,542



### Testing

Do all your testing of pics, sigs, etc. here. It might get deleted some day, it might not. Depends on how lazy the staff feels.



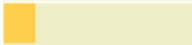
Threads: 179  
Posts: 1,435

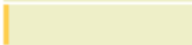
## Customer Reviews

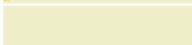
### Average Customer Rating

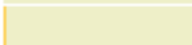
★★★★★ [\(210 customer reviews\)](#)

[5 star:](#)  (156)

[4 star:](#)  (37)

[3 star:](#)  (7)

[2 star:](#)  (4)

[1 star:](#)  (6)

Google Inc.

# Google

<b>Type</b>	Public
<b>Traded as</b>	NASDAQ: <a href="#">GOOG</a>  FWB: <a href="#">GGQ1</a> 
<b>Industry</b>	Internet Computer software
<b>Founded</b>	<a href="#">Menlo Park, California</a> (4 September 1998) <sup>[1][2]</sup>
<b>Founder(s)</b>	<a href="#">Sergey Brin</a> <a href="#">Larry Page</a>
<b>Headquarters</b>	<a href="#">1600 Amphitheatre Parkway, Mountain View, California, United States</a>
<b>Area served</b>	Worldwide
<b>Key people</b>	<a href="#">Larry Page</a> (Co-Founder and CEO) <a href="#">Eric Schmidt</a> (Executive Chairman) <a href="#">Sergey Brin</a> (Co-Founder)
<b>Products</b>	See list of <a href="#">Google products</a> .
<b>Revenue</b>	 US\$ 29.321 billion (2010)
<b>Operating income</b>	 US\$ 10.381 billion (2010)
<b>Profit</b>	 US\$8.505 billion (2010)
<b>Total assets</b>	 \$57.851 billion (2010)
<b>Total equity</b>	 US\$46.241 billion (2010)
<b>Employees</b>	28,768 (As of 2011-06-30) <sup>[3]</sup>

# Vertical Tables

## Coffee



A cup of black coffee

<b>Type</b>	Hot
<b>Country of origin</b>	<a href="#">Ethiopia</a>
<b>Introduced</b>	Approx. 15th century (beverage)
<b>Color</b>	Dark brown, beige, black, light brown

# Semantics Embedded in Surrounding Text

The following table lists the total coffee production of each [coffee exporting country](#) in the year [2006](#)<sup>1</sup>

Country	60 kilogram bags	Kilograms	Pounds
<a href="#">Brazil</a>	42,512,000	2,550,720,000	5,611,584,000
<a href="#">Vietnam</a>	15,000,000	900,000,000	1,980,000,000
<a href="#">Colombia</a>	11,600,000	696,000,000	1,531,200,000
<a href="#">Indonesia</a>	6,850,000	411,000,000	904,200,000
<a href="#">Ethiopia</a>	5,500,000	330,000,000	726,000,000
<a href="#">India</a>	5,005,000	300,300,000	660,660,000
<a href="#">Mexico</a>	4,500,000	270,000,000	594,000,000
<a href="#">Guatemala</a>	4,000,000	240,000,000	528,000,000
<a href="#">Peru</a>	3,500,000	210,000,000	462,000,000
<a href="#">Honduras</a>	2,700,000	162,000,000	356,400,000
<a href="#">Uganda</a>	2,500,000	150,000,000	330,000,000
<a href="#">Ivory Coast</a>	2,350,000	141,000,000	310,200,000
<a href="#">Costa Rica</a>	1,808,000	108,480,000	238,656,000

# And Sometimes, Complicated

**Largest Asian American Ethnic Groups, 2000 Census**

Ethnic Group	Asian alone		Asian & at least One Other Race (i.e., Filipino-White)	Total Population, Alone or in Any Combination
	Single Ethnicity	Two or More Asian Ethnicities (i.e., Chinese-Vietnamese)		
Chinese	2,314,537	130,826	289,478	2,734,841
Filipino	1,850,314	57,811	456,690	2,364,815
Asian Indian	1,678,765	40,013	180,821	1,899,599
Korean	1,076,872	22,550	129,005	1,228,427
Vietnamese	1,122,528	47,144	54,064	1,223,736
Japanese	796,700	55,537	296,695	1,148,932
Cambodian	171,937	11,832	22,283	206,052
Pakistani	153,533	11,095	39,681	204,309
Laotian	168,707	10,396	19,100	198,203
Hmong	169,428	5,284	11,598	186,310
Thai	112,989	7,929	29,365	150,293
Taiwanese	118,048	14,096	12,651	144,795
Indonesian	39,757	4,429	18,887	63,073
Bangladeshi	41,280	5,625	10,507	57,412

# WebTables: Exploring the Relational Web

[Cafarella et al., VLDB 2008, WebDB 08]

- In corpus of 14B raw tables, we estimate 154M are “good” relations
  - Single-table databases; Schema = attr labels + types
  - Largest corpus of databases & schemas we know of
- The Webtables system:
  - Recovers good relations from crawl and enables search
  - Builds novel apps on the recovered data

# Searching Tables is Tricky

[Tweak Document Search, Cafarella 08]

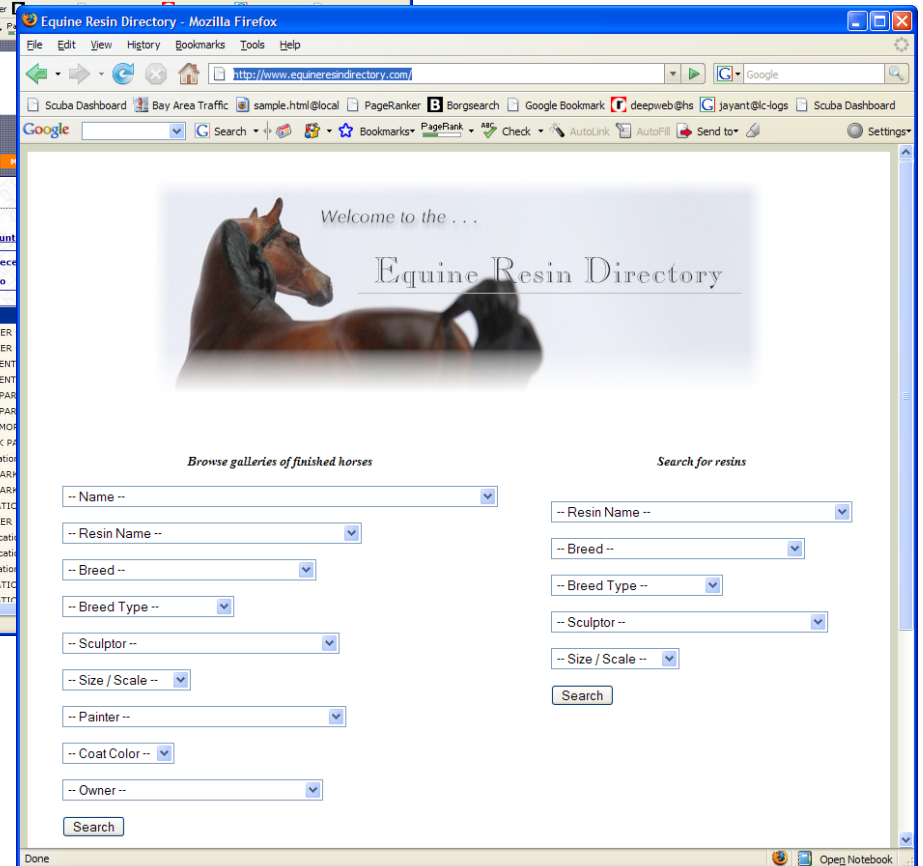
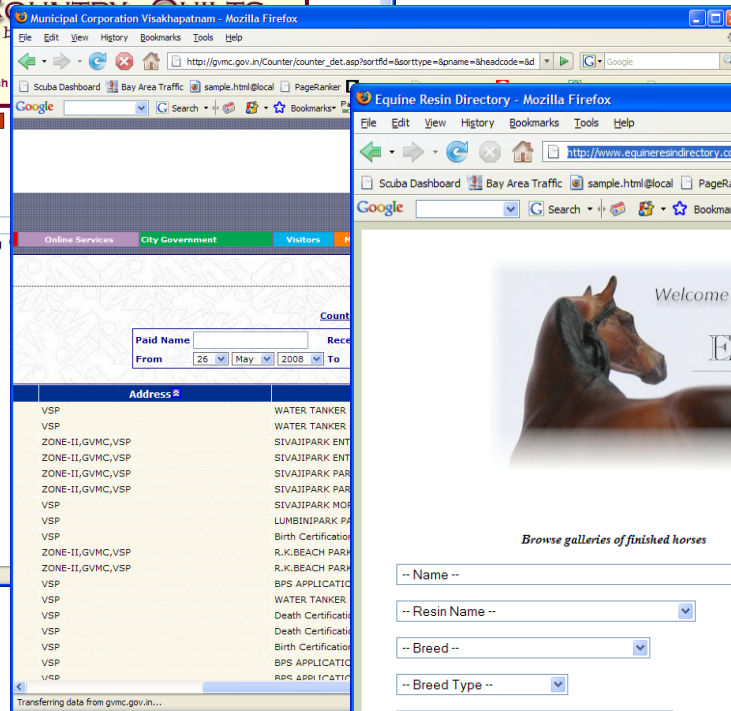
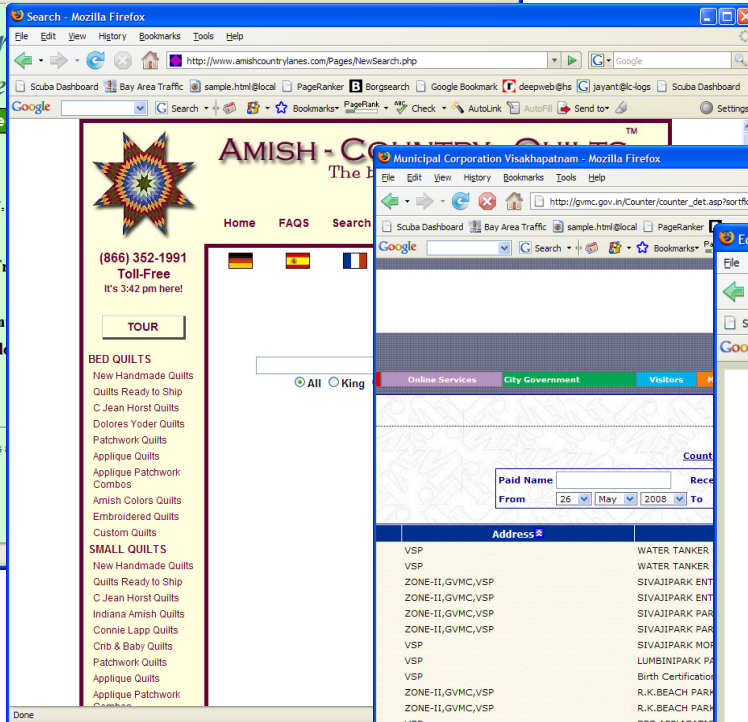
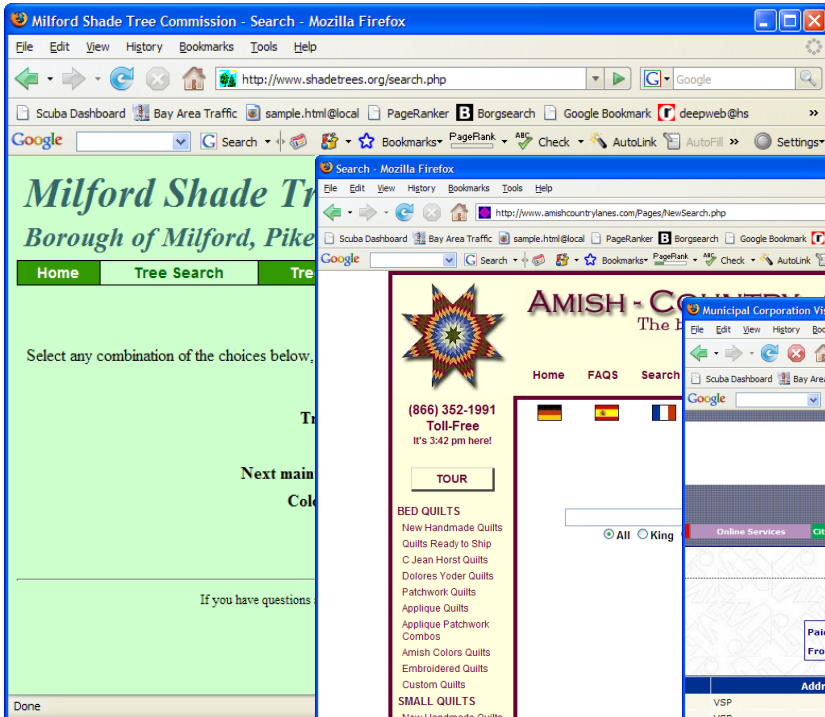
- Consider new cues in ranking:
  - Hits on left column
  - Hits on schema (where there is one)
  - Number of rows, columns
  - Hits on table body
  - Size of table relative to page
- ~25% increase in good results in top-10 results (compared to filtering google results for tables)

# Modeling Challenge:

Data is About *Everything*

Tree Search

Amish quilts



Parking tickets in India

Horses



# Even if we had a KB of everything

[Consider Freebase as example]

- User:
  - Action movies
- Freebase:
  - Movies AND  
Genre=Action
- User
  - Governor of California
- Freebase:
  - US State has governing positions
  - Governing positions have office holder
  - Office holder has position

*Mismatch between KB model and users' conceptual model!*

# The People's Ontology

## [Open Information Extraction]

Mine a database of entities and classes from the Web:

The studio went through some lean times following Walt's death in 1966, but in recent years, Disney animation has undergone a glorious renaissance, with such new classics as "*The Little Mermaid*," "*Beauty & the Beast*," "*Aladdin*," "*The Lion King*," "*Pocahontas*," "*The Hunchback of Notre Dame*," "*Hercules*" and "*Mulan*" joining the Disney roster, together with live action movies such as "*Flubber*," "*Mighty Joe Young*," "*Angels in the Outfield*" and the "*Honey, I Shrank the Kids*" series. And "*Lilo & Stich*" proved a success. Their

Mine binary relationships

California is governed by Jerry Brown, a man some consider to be Rick Perry's arch nemesis. I believe when the Republican primary comes around, Governor Brown will certainly be working to make sure any other Republican wins his state. Former Governor Schwarzenegger went on a speaking

*Broad, dirty, but uses culturally aware terminology*

# Recovering Table Semantics

[Venetis et al., VLDB 2011]

## Trees at Lake Elkhorn

As a recreational facility in Columbia, Lake Elkhorn has a fine collection of both native and non-native trees, many of which are linked to its [ArborTag](#) which describes the distinguishing features of the species.

[Hazel Alder](#)

[Green Ash](#)

[White Ash](#)

[Baldcypress](#)

[Beech](#)

[River Birch](#)

[Boxelder](#)

Red Cedar

[Black Cherry](#)

Sweet Cherry

[Crab Apple](#)

*Alnus serrulata*

*Fraxinus pennsylvanica*

*Fraxinus americana*

*Taxodium distichum*

*Fagus grandifolia*

*Betula nigra*

*Acer negundo*

*Juniperus virginiana*

*Prunus serotina*

*Prunus avium*

*Malus coronaria*

Some tree species exhibit elongate dead areas of bark on diseased existence of targeted defenses resulting from a shared evolutionary history in a exotic North American species such as Green Ash are more susceptible to species when planted at the same site.

# Recovering Binary Relationships

[Crapemyrtle](#)

[Dogwood](#)

[Korean Dogwood](#)

Redosier Dogwood

White Fringetree

*Lagerstroemia indica*

*Cornus florida*

*Cornus kousa*

*Cornus sericea*

*Chionanthus virginicus*

## Dogwood Tree Articles

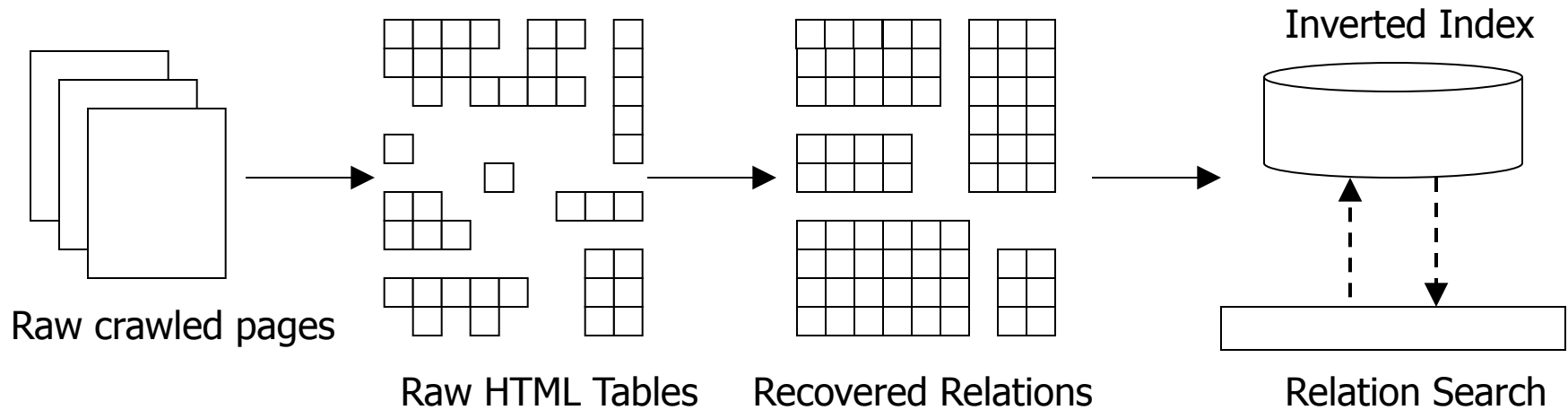


### Flowering Dogwood Tree Facts

Flowering **dogwoods** are known by the name *Cornus florida* or simply "**dogwood**". They are...

***Recovering semantics: better search  
and quality filter***

# Attribute Correlations DB



- 2.6M distinct schemas
- 5.4M attributes

Job-title, company, date	104
Make, model, year	916
Rbi, ab, h, r, bb, avg, slg	12
Dob, player, height, weight	4
...	...

Attribute Correlation Statistics Db

# Synonym Discovery

- Use schema statistics to automatically compute attribute synonyms
  - More complete than thesaurus

- Given input “context” attribute set C:
  1. A = all attrs that appear with C
  2. P = all (a,b) where  $a \in A$ ,  $b \in A$ ,  $a \neq b$
  3. rm all (a,b) from P where  $p(a,b) > 0$
  4. For each remaining pair (a,b) compute:

$$\text{syn}(a, b) = \frac{p(a)p(b)}{\epsilon + \sum_{z \in A} (p(z|a, C) - p(z|b, C))^2}$$

# Synonym Discovery Examples

name	e-mail email, phone telephone, e-mail_address email_address, date last_modified
instructor	course-title title, day days, course course-#, course-name course-title
elected	candidate name, presiding-officer speaker
ab	k so, h hits, avg ba, name player
sqft	bath baths, list list-price, bed beds, price rent

# Conclusions

- Fusion Tables: helping get the ecosystem started.
- Search for structured data sets:
  - Much more to do!
  - Unify with other search
  - Manually created ontologies vs. extracted ones?
- Can we get the crowds to help?
  - Resolving heterogeneity
  - Create new data sets
- Can we help domain-specific expert communities?



# A Few References

- Communications of the ACM: Feb 2011
- Deep web: VLDB 2008
- WebTables: VLDB 2008, 2009, 2011
- Fusion Tables: SIGMOD 2010, SOCC 2010
  - [google.com/fusiontables](http://google.com/fusiontables)
- Principles of Data Integration (Doan, Halevy, Ives): Morgan Kaufmann, 2012.
- The Infinite Emotions of Coffee (Halevy): next month!