

Beyond Search:
Interlinking Public Information

Joe Hellerstein

UCB CS

Background

- The problem of integrating databases is as old as the hills
 - M&A requires it
 - Organizational structure encourages "islands of information" that eventually need to merge
 - Recently, cross-enterprise applications require it

It's an Incredibly Hard Problem to "Solve"

- Syntactic Translations
 - E.g. of units, of codes, etc.
- Semantic Translations
 - Synonyms: "gloves" vs. "hand protectors"
 - Taxonomic promotion: "phillips driver" vs. "screwdriver"
- Implicit Domain Knowledge
 - "salary" in the french database includes a lunch allowance and is before taxes...

But People Chip Away at it Every Day

- Lots of sweat goes a long way
- Clever tools can make it easier
 - E.g. determine that two fields have similar distributions of words or numbers
 - E.g. find discrepancies post-merger
 - E.g. try to auto-extract sub-fields
 - Etc.
- Especially with a user in the loop, this is all annoying but doable.

Enter the Web

- We have learned to think of the web as "the stuff in Google"
 - i.e. textual documents
- But there's a ton of "facts and figures" you can get at on the web
 - Not on any static web page, hence not crawled
 - The "Deep Web"
 - Facts and Figures are not amenable to text searching
- What would happen if you tried to link up some of those?

Telegraph

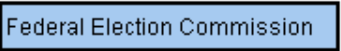


- System we began building in 1999
- First app: "Federated Facts and Figures", Election 2000
- Now focused on "streaming databases"
 - E.g. sensor feeds
 - E.g. network packets
- Some screenshots from Election 2K

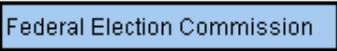
The applet might take a few seconds to start displaying results. Click [here](#) for instructions on how to install the applet. Click on the arrows in the left column to probe deeper into the results.

Bush Supporters

Gore Supporters

Got result schema 



Got result schema 



Results



State / Occupation / Employer /	Number of Entries	Sum of Contribution
Businessman	4.0	4000.0
CEO	40.0	22375.0
"AMPS	1.0	1000.0
"Associated Pipe Line	1.0	1000.0
"Beall Industries, Inc.	1.0	500.0
"Blakeman Trans., Inc	1.0	250.0
"Brink's Home Security	1.0	900.0
"Cantrell Auto Supply, Inc.	1.0	500.0
"Centex Corp.	1.0	100.0
"Classy Glass Inc.	1.0	250.0
"Coastal Power Company	1.0	1000.0
"Computer Career Center	1.0	250.0
"Containerhouse Intl., Inc.	1.0	250.0
"Depelchin Children's Ce...	1.0	250.0

33767 -

Results



State / Occupation / Employer /	Number of Entries	Sum of Contribut...
Businessman	0.0	0.0
Businesswoman	2.0	1100.0
CEO	41.0	24951.0
"Achieva College Prep Ctrs.	1.0	1000.0
"Boulmiche, Inc.	1.0	100.0
"Carsey Werner Co.	1.0	1000.0
"Digital Directions Group	1.0	1000.0
"ECP	1.0	1000.0
"Earl M. Jorgensen	1.0	125.0
"Excel Legacy Corp.	1.0	1000.0
"Informania, Inc.	1.0	1000.0
"KidsOnLine.Com	1.0	1000.0
"Lawrence Research Group, Inc.	1.0	1000.0
"Levine Leichtman Cap. Partners	1.0	1000.0
"Lumeria, Inc.	1.0	1000.0

33816 -


[Back to top](#)


How We Do It

To perform this query, we correlated information from the following sources on the web:

...

Individual donations to presidential candidates. Square size indicates size of donation. Color indicates candidate. Click and drag to zoom in. Click 'Zoom Out' to zoom out.

Bush 

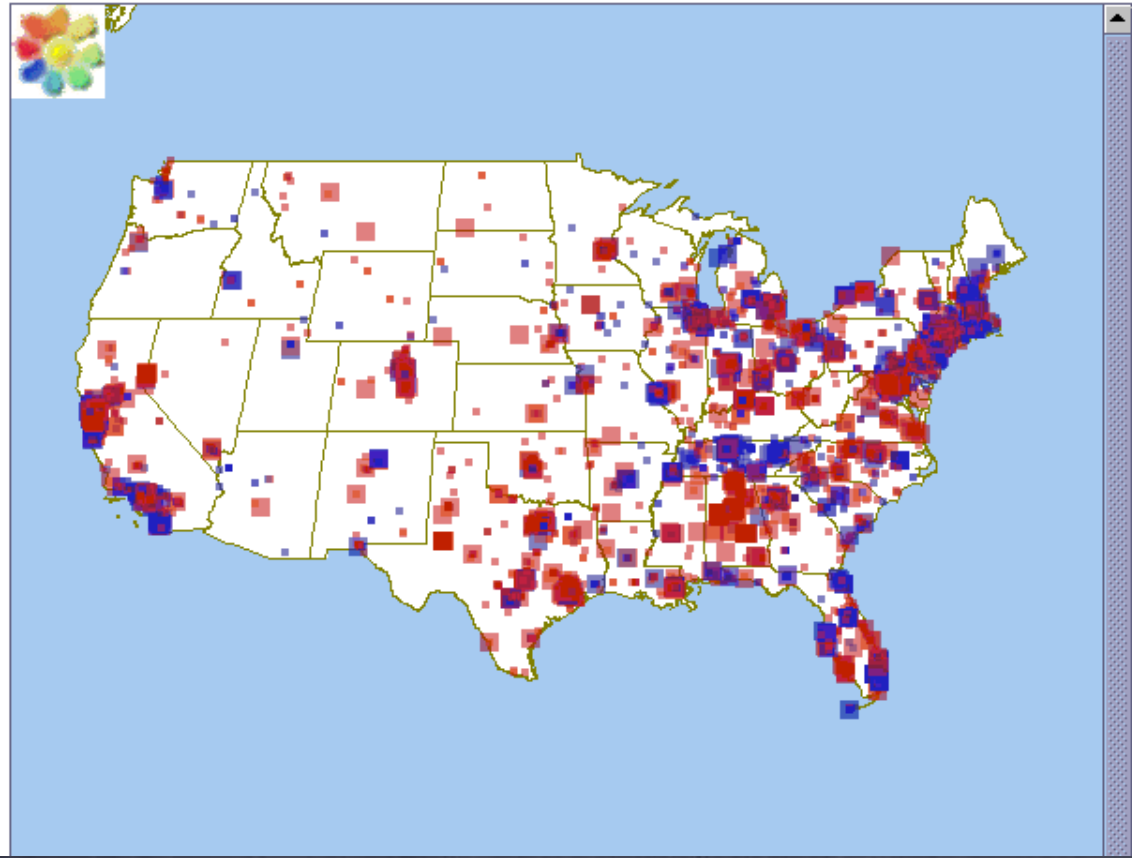
Gore 

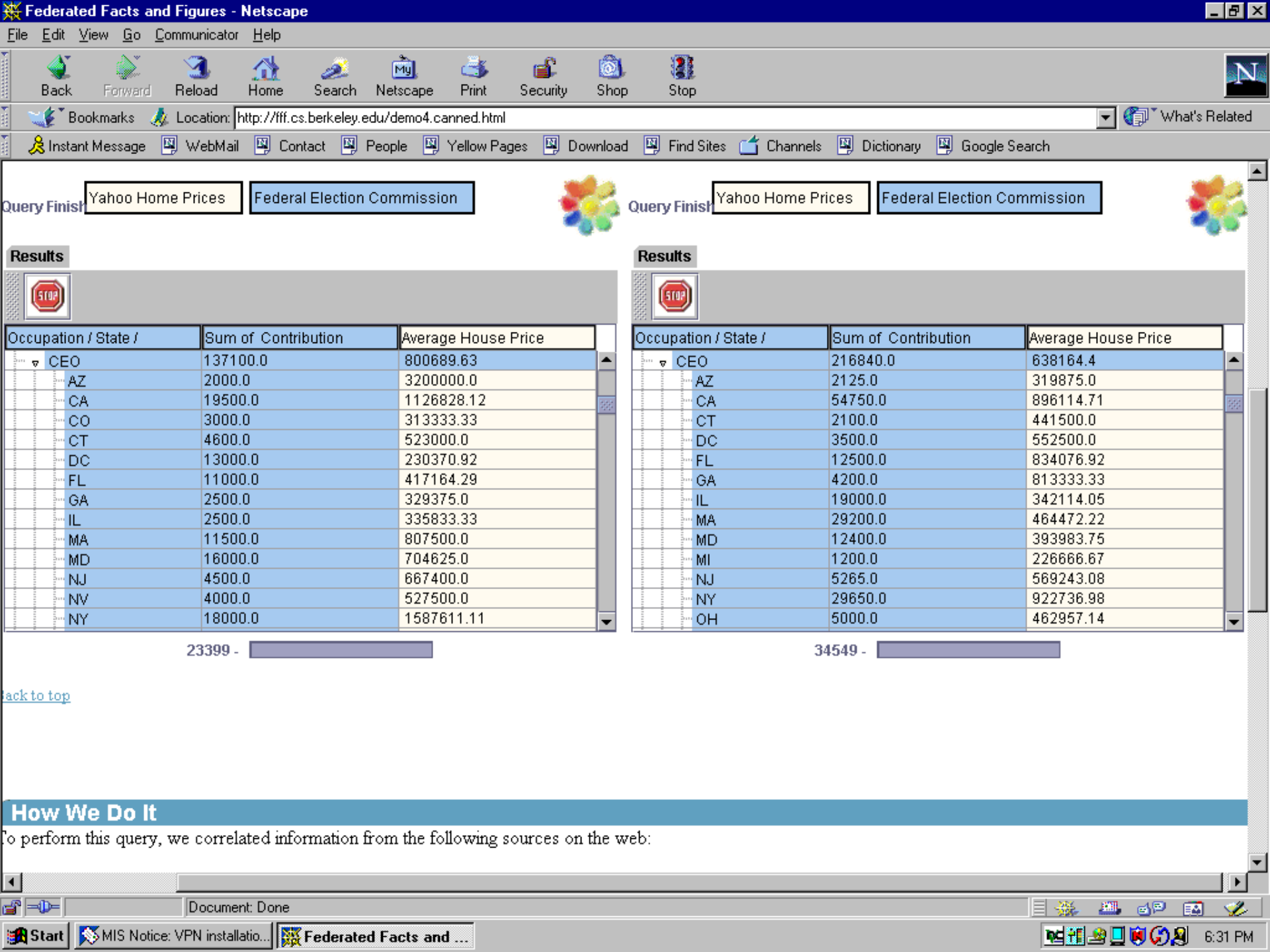
Java Applet Window

per/demo2.canned.html

Go Links >>

The applet might take a few seconds to start displaying results. Click [here](#) for instructions on how to install the applet.





Query Finish

Query Finish

Results



Occupation / State /	Sum of Contribution	Average House Price
CEO	137100.0	800689.63
AZ	2000.0	3200000.0
CA	19500.0	1126828.12
CO	3000.0	313333.33
CT	4600.0	523000.0
DC	13000.0	230370.92
FL	11000.0	417164.29
GA	2500.0	329375.0
IL	2500.0	335833.33
MA	11500.0	807500.0
MD	16000.0	704625.0
NJ	4500.0	667400.0
NV	4000.0	527500.0
NY	18000.0	1587611.11

23399 -

Results



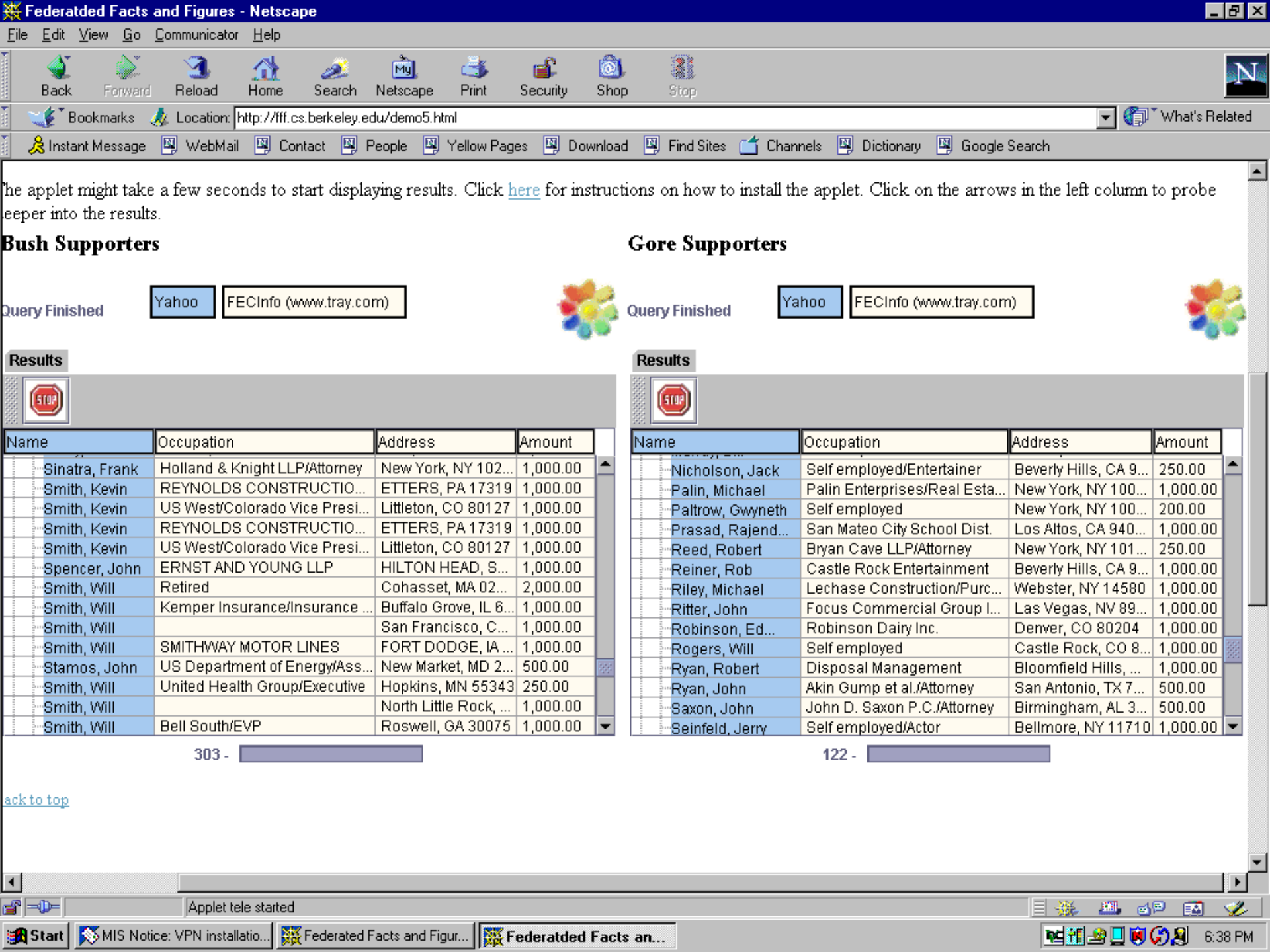
Occupation / State /	Sum of Contribution	Average House Price
CEO	216840.0	638164.4
AZ	2125.0	319875.0
CA	54750.0	896114.71
CT	2100.0	441500.0
DC	3500.0	552500.0
FL	12500.0	834076.92
GA	4200.0	813333.33
IL	19000.0	342114.05
MA	29200.0	464472.22
MD	12400.0	393983.75
MI	1200.0	226666.67
NJ	5265.0	569243.08
NY	29650.0	922736.98
OH	5000.0	462957.14

34549 -

[back to top](#)

How We Do It

To perform this query, we correlated information from the following sources on the web:



The applet might take a few seconds to start displaying results. Click [here](#) for instructions on how to install the applet. Click on the arrows in the left column to probe deeper into the results.

Bush Supporters

Gore Supporters

Query Finished

Yahoo



Query Finished

Yahoo



Results



Name	Occupation	Address	Amount
Sinatra, Frank	Holland & Knight LLP/Attorney	New York, NY 102...	1,000.00
Smith, Kevin	REYNOLDS CONSTRUCTIO...	ETTERS, PA 17319	1,000.00
Smith, Kevin	US West/Colorado Vice Presi...	Littleton, CO 80127	1,000.00
Smith, Kevin	REYNOLDS CONSTRUCTIO...	ETTERS, PA 17319	1,000.00
Smith, Kevin	US West/Colorado Vice Presi...	Littleton, CO 80127	1,000.00
Spencer, John	ERNST AND YOUNG LLP	HILTON HEAD, S...	1,000.00
Smith, Will	Retired	Cohasset, MA 02...	2,000.00
Smith, Will	Kemper Insurance/Insurance ...	Buffalo Grove, IL 6...	1,000.00
Smith, Will		San Francisco, C...	1,000.00
Smith, Will	SMITHWAY MOTOR LINES	FORT DODGE, IA ...	1,000.00
Stamos, John	US Department of Energy/Ass...	New Market, MD 2...	500.00
Smith, Will	United Health Group/Executive	Hopkins, MN 55343	250.00
Smith, Will		North Little Rock, ...	1,000.00
Smith, Will	Bell South/EVP	Roswell, GA 30075	1,000.00

303 -

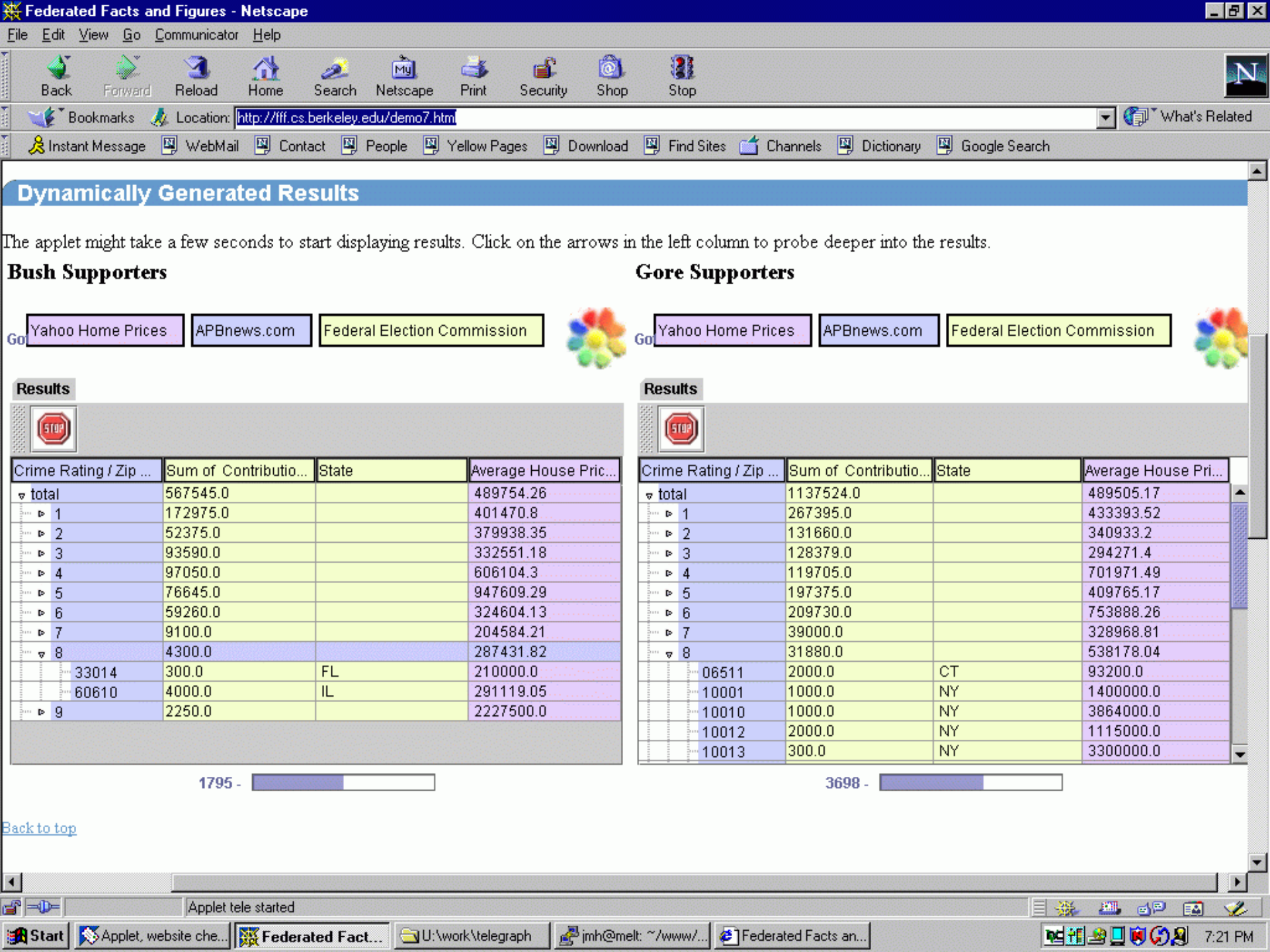
[back to top](#)

Results



Name	Occupation	Address	Amount
Nicholson, Jack	Self employed/Entertainer	Beverly Hills, CA 9...	250.00
Palin, Michael	Palin Enterprises/Real Esta...	New York, NY 100...	1,000.00
Paltrow, Gwyneth	Self employed	New York, NY 100...	200.00
Prasad, Rajend...	San Mateo City School Dist.	Los Altos, CA 940...	1,000.00
Reed, Robert	Bryan Cave LLP/Attorney	New York, NY 101...	250.00
Reiner, Rob	Castle Rock Entertainment	Beverly Hills, CA 9...	1,000.00
Riley, Michael	Lechase Construction/Purc...	Webster, NY 14580	1,000.00
Ritter, John	Focus Commercial Group I...	Las Vegas, NV 89...	1,000.00
Robinson, Ed...	Robinson Dairy Inc.	Denver, CO 80204	1,000.00
Rogers, Will	Self employed	Castle Rock, CO 8...	1,000.00
Ryan, Robert	Disposal Management	Bloomfield Hills, ...	1,000.00
Ryan, John	Akin Gump et al./Attorney	San Antonio, TX 7...	500.00
Saxon, John	John D. Saxon P.C./Attorney	Birmingham, AL 3...	500.00
Seinfeld, Jerry	Self employed/Actor	Bellmore, NY 11710	1,000.00

122 -



Dynamically Generated Results

The applet might take a few seconds to start displaying results. Click on the arrows in the left column to probe deeper into the results.

Bush Supporters

Gore Supporters

Go

Go

Results



Crime Rating / Zip ...	Sum of Contributio...	State	Average House Pric...
▼ total	567545.0		489754.26
▶ 1	172975.0		401470.8
▶ 2	52375.0		379938.35
▶ 3	93590.0		332551.18
▶ 4	97050.0		606104.3
▶ 5	76645.0		947609.29
▶ 6	59260.0		324604.13
▶ 7	9100.0		204584.21
▼ 8	4300.0		287431.82
▶ 33014	300.0	FL	210000.0
▶ 60610	4000.0	IL	291119.05
▶ 9	2250.0		2227500.0

1795 -

Results



Crime Rating / Zip ...	Sum of Contributio...	State	Average House Pri...
▼ total	1137524.0		489505.17
▶ 1	267395.0		433393.52
▶ 2	131660.0		340933.2
▶ 3	128379.0		294271.4
▶ 4	119705.0		701971.49
▶ 5	197375.0		409765.17
▶ 6	209730.0		753888.26
▶ 7	39000.0		328968.81
▼ 8	31880.0		538178.04
▶ 06511	2000.0	CT	93200.0
▶ 10001	1000.0	NY	1400000.0
▶ 10010	1000.0	NY	3864000.0
▶ 10012	2000.0	NY	1115000.0
▶ 10013	300.0	NY	3300000.0

3698 -

[Back to top](#)

Discussion Topics

- Privacy
- Veracity & Responsibility
- Transparency of use
- Secondary use
- TIA
- "White-hat" activities
- Individuals and demographics (aggregates)
- Interplay with streaming data