Web Data Management: Powering the New Web

Raghu Ramakrishnan

Chief Scientist for Audience, Yahoo!

Research Fellow, Yahoo! Research

(On leave, Univ. of Wisconsin-Madison)

Outline

- **Trends in Search and Information Discovery**
  - Move towards task-centricity
  - Need to interpret content
- **Evolution of Online Communities**
  - Social Search
  - PeopleWeb
  - Community Information Management
- **Web Data Infrastructure**
  - Massively distributed computing
  - Hosted services
  - Heterogeneous content

Further Reading

- Content, Metadata, and Behavioral Information: Directions for Yahoo! Research, *The Yahoo! Research Team*, IEEE Data Engineering Bulletin, Dec 2006 (Special Issue on Web-Scale Data, Systems, and Semantics)
- Systems, Communities, Community Systems, on the Web, *Community Systems Group at Yahoo! Research*, SIGMOD Record, Sept 2007
- Towards a PeopleWeb, R. Ramakrishnan and A. Tomkins, IEEE Computer, August 2007 (Special Issue on Web Search)

Community Systems Group @ Yahoo! Research

Philip Bohannon
Brian Cooper
Nilesh Dalvi
Minos Garofalakis
Hans-Arno Jacobsen
Vinay Kakade
Dan Kifer
Raghu Ramakrishnan
Adam Silberstein
Utkarsh Srivastava
Ramana Yerneni
Cong Yu

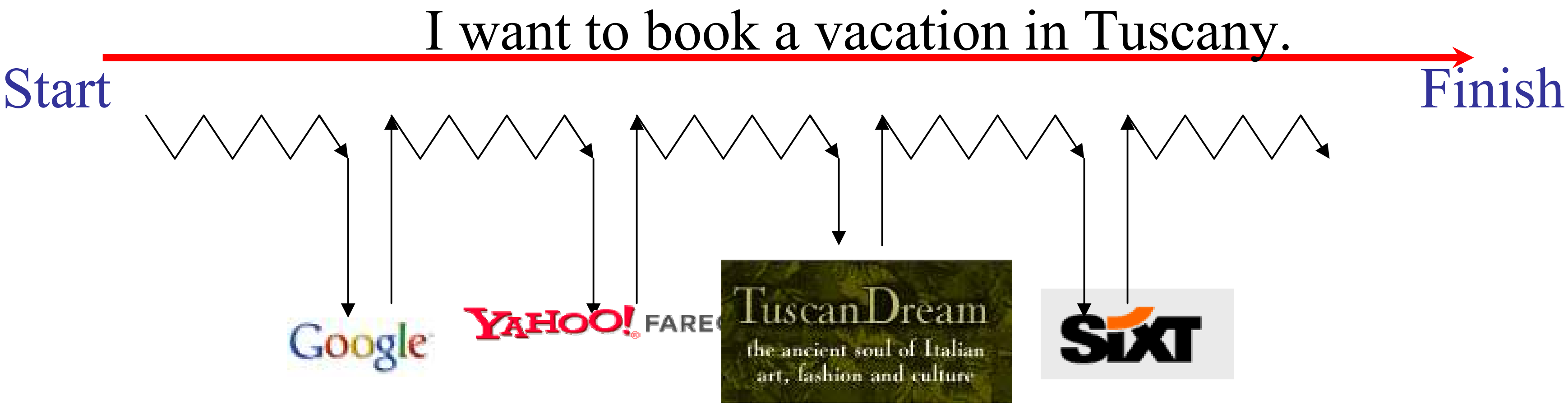
Deepak Agrawal
Sihem Amer-Yahia
Ravi Kumar
Cameron Marlow
Srujana Merugu
Chris Olston
Bo Pang
Ben Reed
Keerthi Selvaraj
Jai Shanmugasundaram
Andrew Tomkins

Parag Agrawal
Tyson Condie
Pedro DeRose
Alban Galland
Nitin Gupta
Ashwin Machanavajjhala
Warren Shen
Julia Stoyanovich
Fan Yang

Trends in Search

Search and Content Supply

- Premise:
  - People don't want to search
  - People want to get tasks done



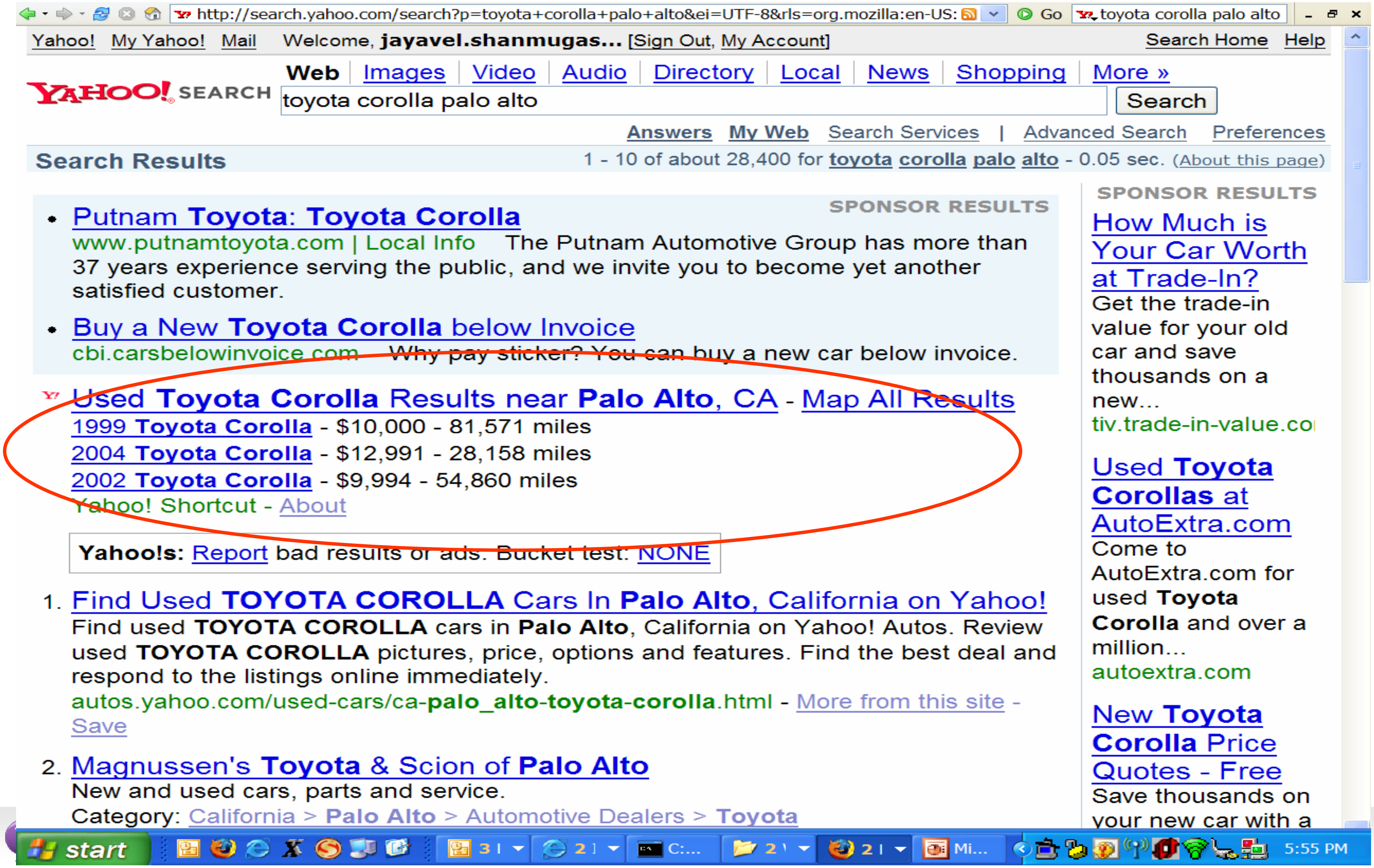
I want to book a vacation in Tuscany.

Start

Finish

Broder 2002, A Taxomony of web search

Y! Shortcuts



Search Results 1 - 10 of about 28,400 for **toyota corolla palo alto** - 0.05 sec. (About this page)

SPONSOR RESULTS

- **Putnam Toyota: Toyota Corolla**
  www.putnamtoyota.com | Local Info The Putnam Automotive Group has more than 37 years experience serving the public, and we invite you to become yet another satisfied customer.
- **Buy a New Toyota Corolla below Invoice**
  cbi.carsbelowinvoice.com Why pay sticker? You can buy a new car below invoice.

Used **Toyota Corolla** Results near **Palo Alto**, CA - Map All Results

1999 **Toyota Corolla** - $10,000 - 81,571 miles
2004 **Toyota Corolla** - $12,991 - 28,158 miles
2002 **Toyota Corolla** - $9,994 - 54,860 miles
Yahoo! Shortcut - About

Yahoo!s: Report bad results or ads. Bucket test: NONE

1. Find Used **TOYOTA COROLLA** Cars In **Palo Alto**, California on Yahoo!
   Find used **TOYOTA COROLLA** cars in **Palo Alto**, California on Yahoo! Autos. Review used **TOYOTA COROLLA** pictures, price, options and features. Find the best deal and respond to the listings online immediately.
   autos.yahoo.com/used-cars/ca-**palo_alto-toyota-corolla**.html - More from this site - Save
2. Magnussen's **Toyota** & Scion of **Palo Alto**
   New and used cars, parts and service.
   Category: California > **Palo Alto** > Automotive Dealers > **Toyota**

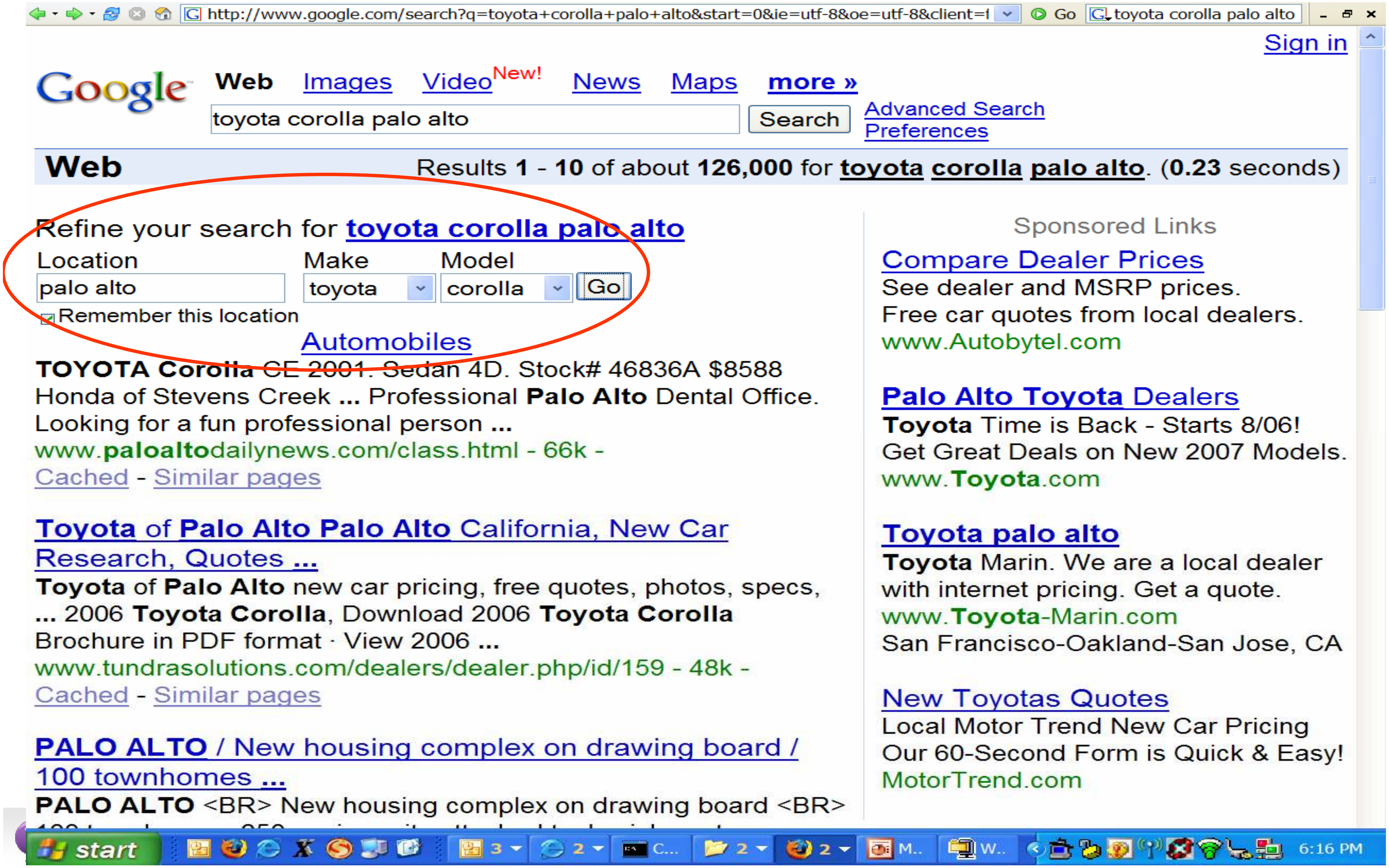
SPONSOR RESULTS

How Much is Your Car Worth at Trade-In?
Get the trade-in value for your old car and save thousands on a new...
tiv.trade-in-value.co

Used **Toyota Corollas** at AutoExtra.com
Come to AutoExtra.com for used **Toyota Corolla** and over a million...
autoextra.com

New **Toyota Corolla** Price Quotes - Free
Save thousands on your new car with a

Google Base



toyota corolla palo alto

Web Results **1** - **10** of about **126,000** for **toyota corolla palo alto**. (**0.23** seconds)

Refine your search for **toyota corolla palo alto**

Location	Make	Model
palo alto	toyota	corolla

Go

☑ Remember this location

Automobiles

TOYOTA Corolla CE 2001. Sedan 4D. Stock# 46836A $8588 Honda of Stevens Creek ... Professional **Palo Alto** Dental Office. Looking for a fun professional person ...
www.**paloalto**dailynews.com/class.html - 66k -
Cached - Similar pages

Toyota of **Palo Alto Palo Alto** California, New Car Research, Quotes ...
Toyota of **Palo Alto** new car pricing, free quotes, photos, specs, ... 2006 **Toyota Corolla**, Download 2006 **Toyota Corolla** Brochure in PDF format · View 2006 ...
www.tundrasolutions.com/dealers/dealer.php/id/159 - 48k -
Cached - Similar pages

PALO ALTO / New housing complex on drawing board / 100 townhomes ...
PALO ALTO <BR> New housing complex on drawing board <BR>

Sponsored Links

Compare Dealer Prices
See dealer and MSRP prices.
Free car quotes from local dealers.
www.Autobytel.com

Palo Alto Toyota Dealers
Toyota Time is Back - Starts 8/06!
Get Great Deals on New 2007 Models.
www.**Toyota**.com

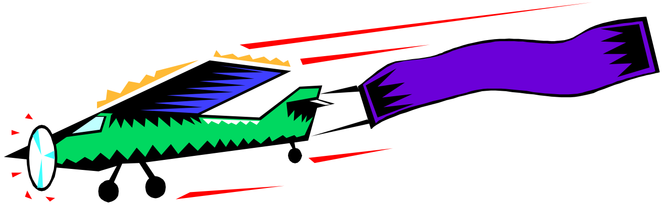
Toyota palo alto
Toyota Marin. We are a local dealer with internet pricing. Get a quote.
www.**Toyota**-Marin.com
San Francisco-Oakland-San Jose, CA

New **Toyotas** Quotes
Local Motor Trend New Car Pricing
Our 60-Second Form is Quick & Easy!
MotorTrend.com

Structure → Intent

“seafood san francisco”

Category: restaurant
Location: San Francisco



Reserve a table for two tonight at SF’s best Sushi Bar and get a free sake, compliments of OpenTable!

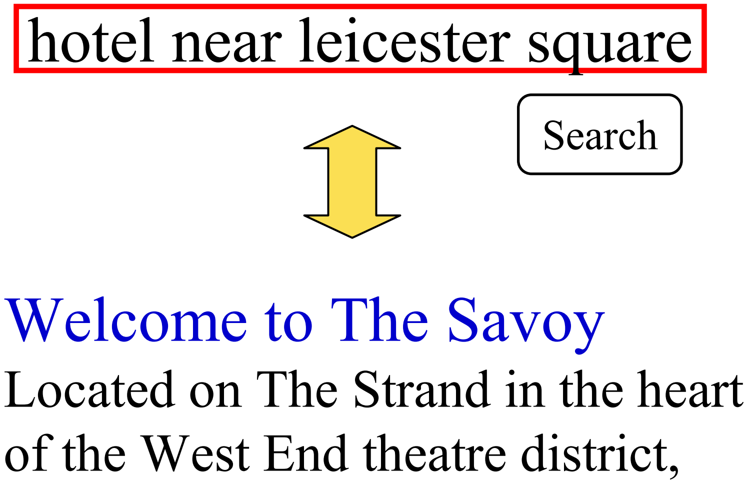
Category: restaurant Location: San Francisco

Alamo Square Seafood Grill – (415) 440-2828
803 Fillmore St, San Francisco, CA – 0.93mi – map

Category: restaurant Location: San Francisco

Steps to Task-Centricity

- Information integration
  - Information extraction
  - Schema normalization
- Structure
  - Extract and exploit



hotel near leicester square

Search

Welcome to The Savoy

Located on The Strand in the heart of the West End theatre district,

Semantic Structure – Not Easy

- Colorado, Texas
- Oregon, Alaska
- Peru, Bolivia
- Peru, Argentina
- Washington, Nevada County, California
- Bush, Cheney & Rice, WA state

How Do We Circumvent?

- Unleash community computing
  - Social structure
  - Incentive mechanisms
- More in a moment

Evolution of Online Communities

Rate of content creation

- Estimated growth of content
  - Published content from traditional sources: 3-4 Gb/day
  - Professional web content: ~2 Gb/day
  - User-generated content: 8-10 Gb/day
  - Private text content: ~3 Tb/day (200x more)
  - Upper bound on typed content: ~700 Tb/day

Metadata

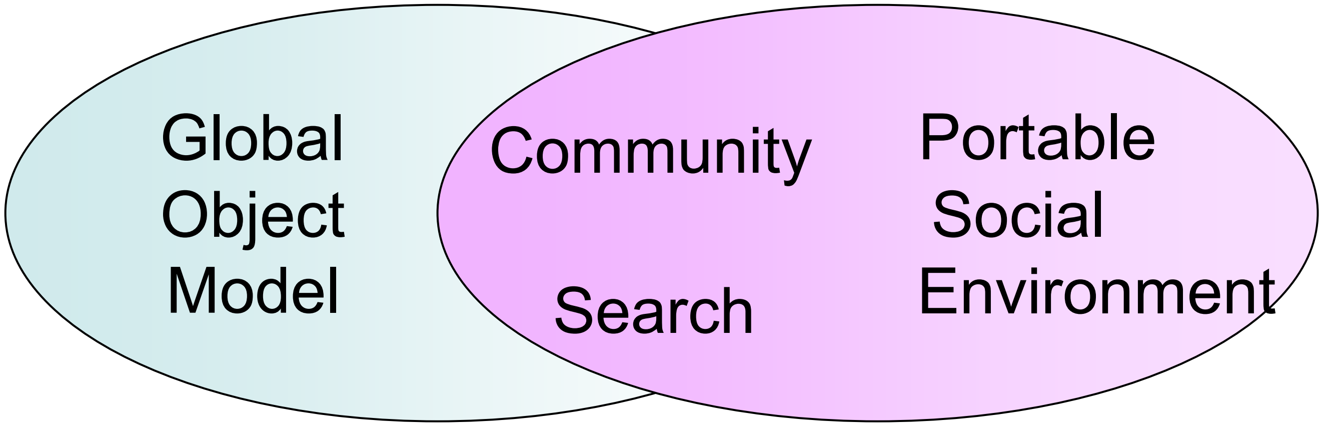
- Estimated growth of metadata
  - Anchortext: 100Mb/day
  - Tags: 40Mb/day
  - Pageviews: 100-200Gb/day
  - Reviews: Around 10Mb/day
  - Ratings: <small>

Drove most advances in search from 1996-present

Increasingly rich and available, but not yet useful in search

This is in spite of the fact that interactions on the web are currently limited by the fact that each site is essentially a silo

PeopleWeb: Site-Centric → People-Centric



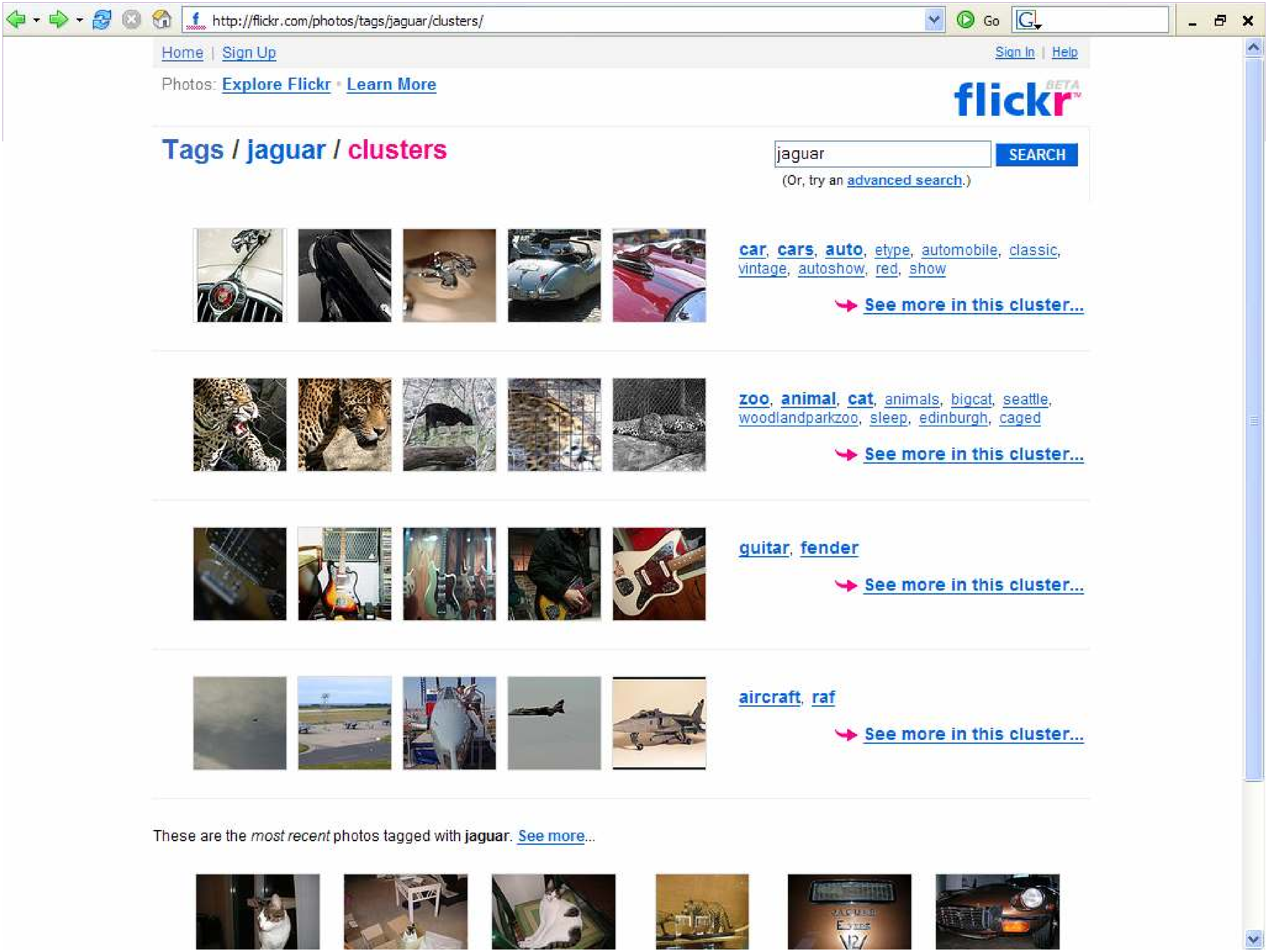
- **Common web-wide id for objects (incl. users)**
  - Even common attributes? (e.g., *pixels* for camera objects)
- As users move across sites, their personas and social networks will be carried along
- **Increased semantics on the web through community activity (another path to the goals of the Semantic Web)**

(Towards a PeopleWeb, Ramakrishnan & Tomkins, IEEE Computer, August 2007)

Social Search

- **Improve web search by**
  - – Learning from shared community interactions, and leveraging community interactions to create and refine content
    - Enhance and amplify user interactions
  - – Expanding search results to include sources of information (e.g., experts, sub-communities of shared interest)

Reputation, Quality, Trust, Privacy



Photos: Explore Flickr • Learn More

Tags / jaguar / clusters

(Or, try an advanced search.)

car, cars, auto, etype, automobile, classic, vintage, autoshow, red, show

See more in this cluster...

zoo, animal, cat, animals, bigcat, seattle, woodlandparkzoo, sleep, edinburgh, caged

See more in this cluster...

guitar, fender

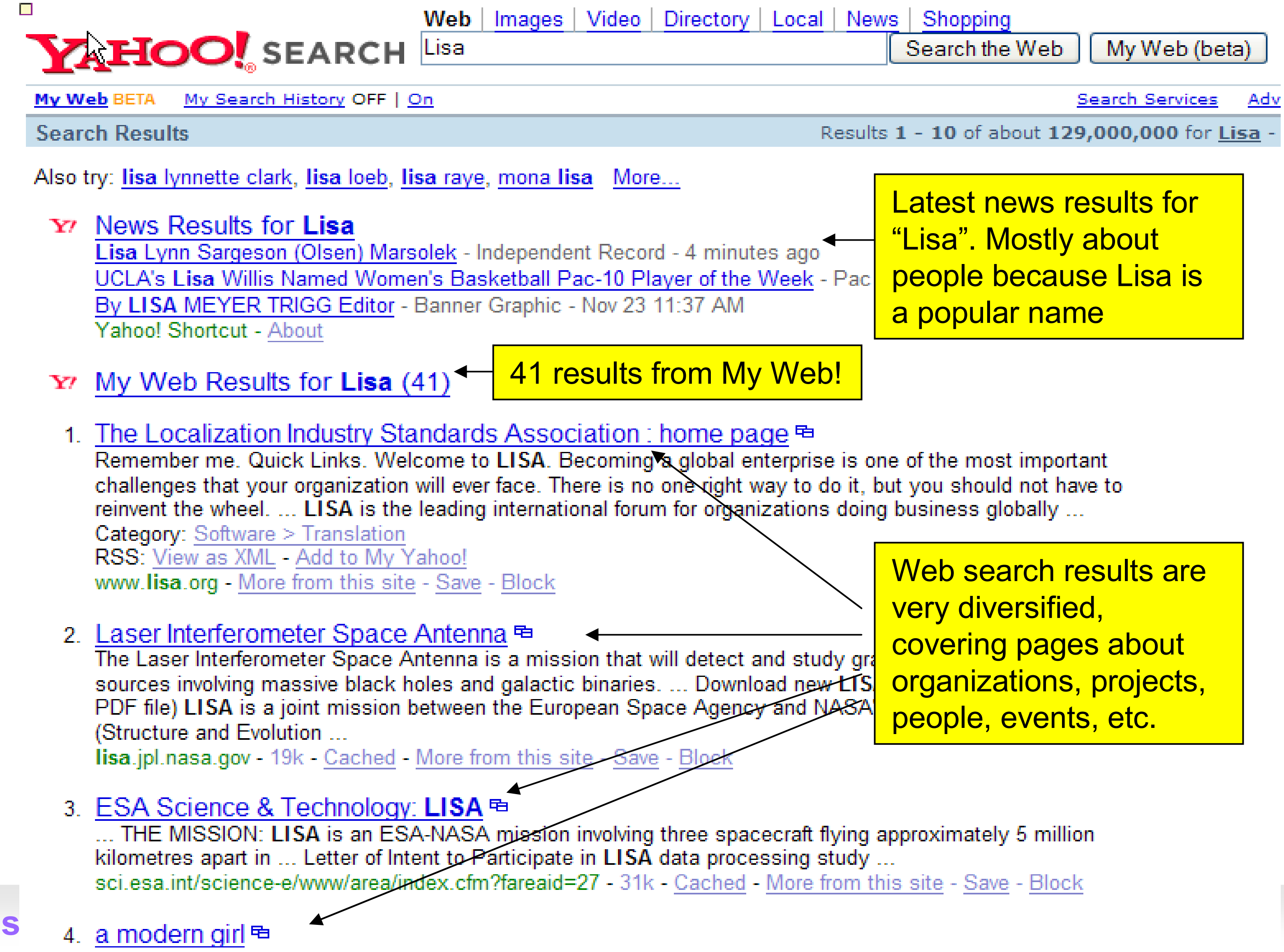
See more in this cluster...

aircraft, raf

See more in this cluster...

These are the *most recent* photos tagged with **jaguar**. See more...

Web Search Results for “Lisa”


YAHOO! SEARCH
Web | Images | Video | Directory | Local | News | Shopping
Search the Web
My Web (beta)
My Web BETA
My Search History OFF | On
Search Services
Adv

Lisa

Search Results

Results **1** - **10** of about **129,000,000** for **Lisa** -

Also try: **lisa** lynnette clark, **lisa** loeb, **lisa** raye, mona **lisa** More...

News Results for **Lisa**

Lisa Lynn Sargeson (Olsen) Marsolek - Independent Record - 4 minutes ago

UCLA's **Lisa** Willis Named Women's Basketball Pac-10 Player of the Week - Pac

By **LISA** MEYER TRIGG Editor - Banner Graphic - Nov 23 11:37 AM

Yahoo! Shortcut - About

My Web Results for **Lisa** (41)

1. The Localization Industry Standards Association : home page

   Remember me. Quick Links. Welcome to **LISA**. Becoming a global enterprise is one of the most important challenges that your organization will ever face. There is no one right way to do it, but you should not have to reinvent the wheel. ... **LISA** is the leading international forum for organizations doing business globally ...

   Category: Software > Translation

   RSS: View as XML - Add to My Yahoo!

   www.**lisa**.org - More from this site - Save - Block

2. Laser Interferometer Space Antenna

   The Laser Interferometer Space Antenna is a mission that will detect and study gra sources involving massive black holes and galactic binaries. ... Download new LIS PDF file) **LISA** is a joint mission between the European Space Agency and NASA (Structure and Evolution ...

   lisa.jpl.nasa.gov - 19k - Cached - More from this site - Save - Block

3. ESA Science & Technology: **LISA**

   ... THE MISSION: **LISA** is an ESA-NASA mission involving three spacecraft flying approximately 5 million kilometres apart in ... Letter of Intent to Participate in **LISA** data processing study ...

   sci.esa.int/science-e/www/area/index.cfm?fareaid=27 - 31k - Cached - More from this site - Save - Block

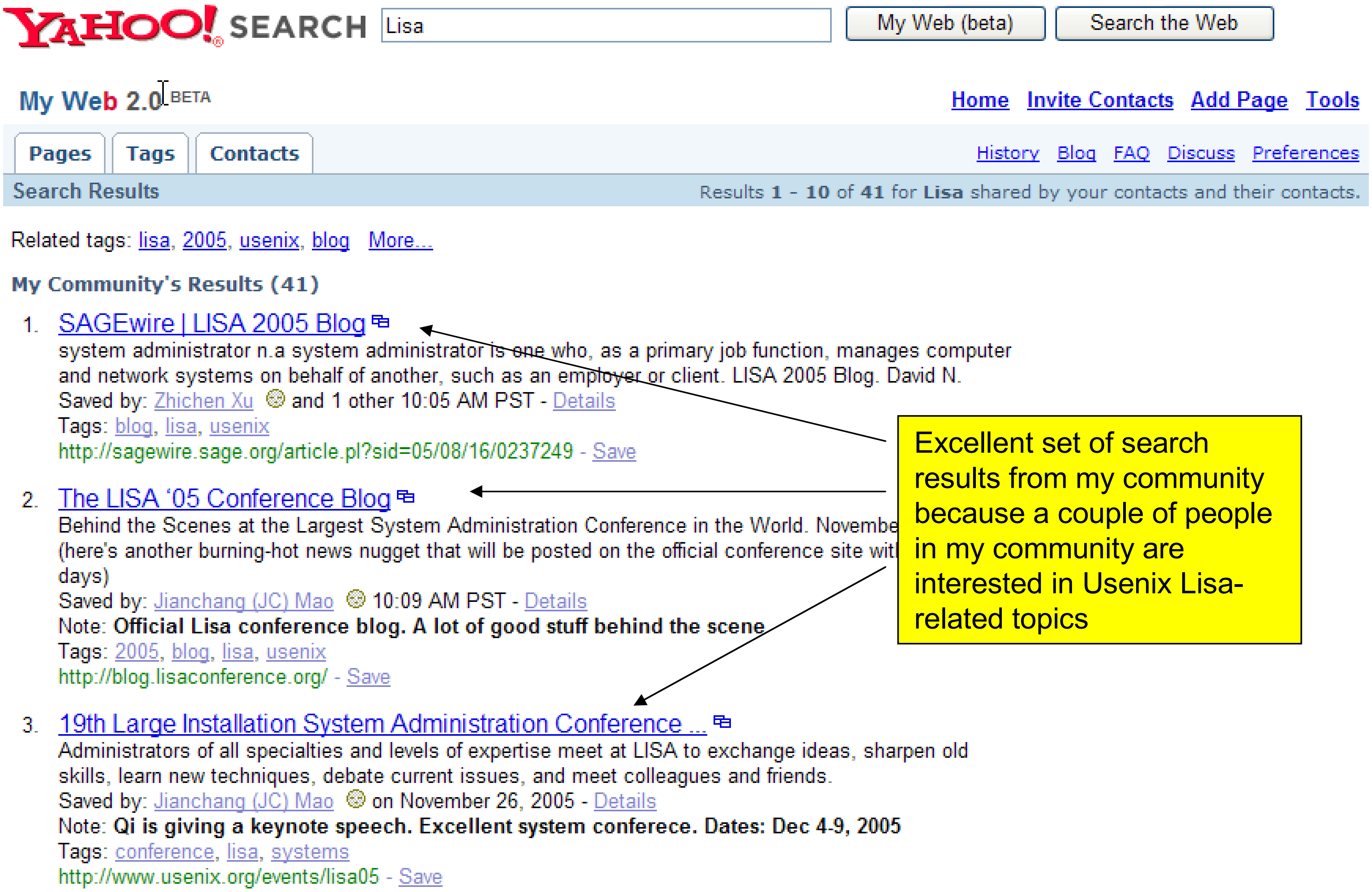
4. a modern girl

Latest news results for “Lisa”. Mostly about people because Lisa is a popular name

41 results from My Web!

Web search results are very diversified, covering pages about organizations, projects, people, events, etc.

My Web 2.0 Search Results for “Lisa”



Search Results

Results **1** - **10** of **41** for **Lisa** shared by your contacts and their contacts.

Related tags: lisa, 2005, usenix, blog More...

My Community's Results (41)

1. SAGEwire | LISA 2005 Blog
   system administrator n.a system administrator is one who, as a primary job function, manages computer and network systems on behalf of another, such as an employer or client. LISA 2005 Blog. David N.
   Saved by: Zhichen Xu and 1 other 10:05 AM PST - Details
   Tags: blog, lisa, usenix
   http://sagewire.sage.org/article.pl?sid=05/08/16/0237249 - Save

2. The LISA ‘05 Conference Blog
   Behind the Scenes at the Largest System Administration Conference in the World. Novembe (here's another burning-hot news nugget that will be posted on the official conference site wit days)
   Saved by: Jianchang (JC) Mao 10:09 AM PST - Details
   Note: **Official Lisa conference blog. A lot of good stuff behind the scene**
   Tags: 2005, blog, lisa, usenix
   http://blog.lisaconference.org/ - Save

3. 19th Large Installation System Administration Conference ...
   Administrators of all specialties and levels of expertise meet at LISA to exchange ideas, sharpen old skills, learn new techniques, debate current issues, and meet colleagues and friends.
   Saved by: Jianchang (JC) Mao on November 26, 2005 - Details
   Note: **Qi is giving a keynote speech. Excellent system conferece. Dates: Dec 4-9, 2005**
   Tags: conference, lisa, systems
   http://www.usenix.org/events/lisa05 - Save

Excellent set of search results from my community because a couple of people in my community are interested in Usenix Lisa-related topics

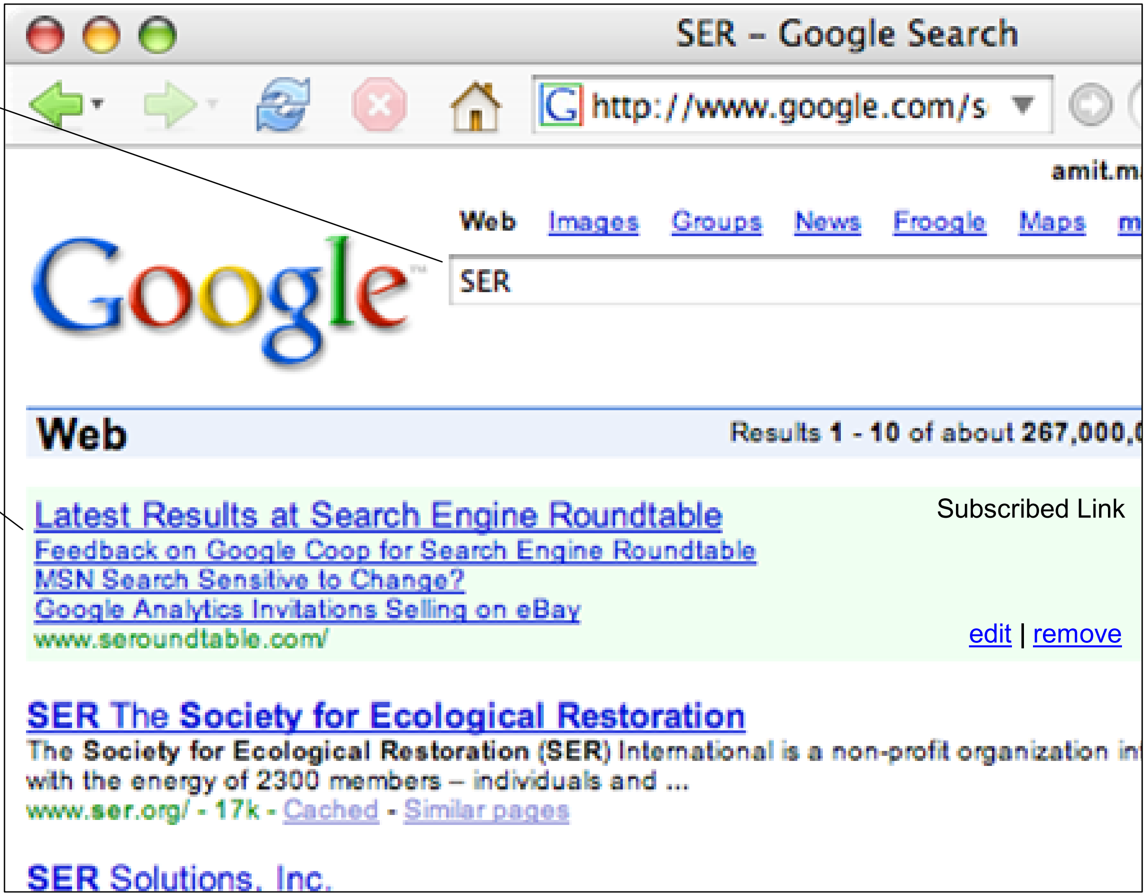
Google Co-Op

Query-based direct-display, programmed by Contributor

This query matches a pattern provided by Contributor…

…so SERP displays (query-specific) links programmed by Contributor.

Users “opts-in” by “subscribing” to them



Web Results 1 - 10 of about 267,000,

Latest Results at Search Engine Roundtable

Subscribed Link

Feedback on Google Coop for Search Engine Roundtable

MSN Search Sensitive to Change?

Google Analytics Invitations Selling on eBay

www.seroundtable.com/

edit | remove

SER The Society for Ecological Restoration

The Society for Ecological Restoration (SER) International is a non-profit organization in with the energy of 2300 members – individuals and ...

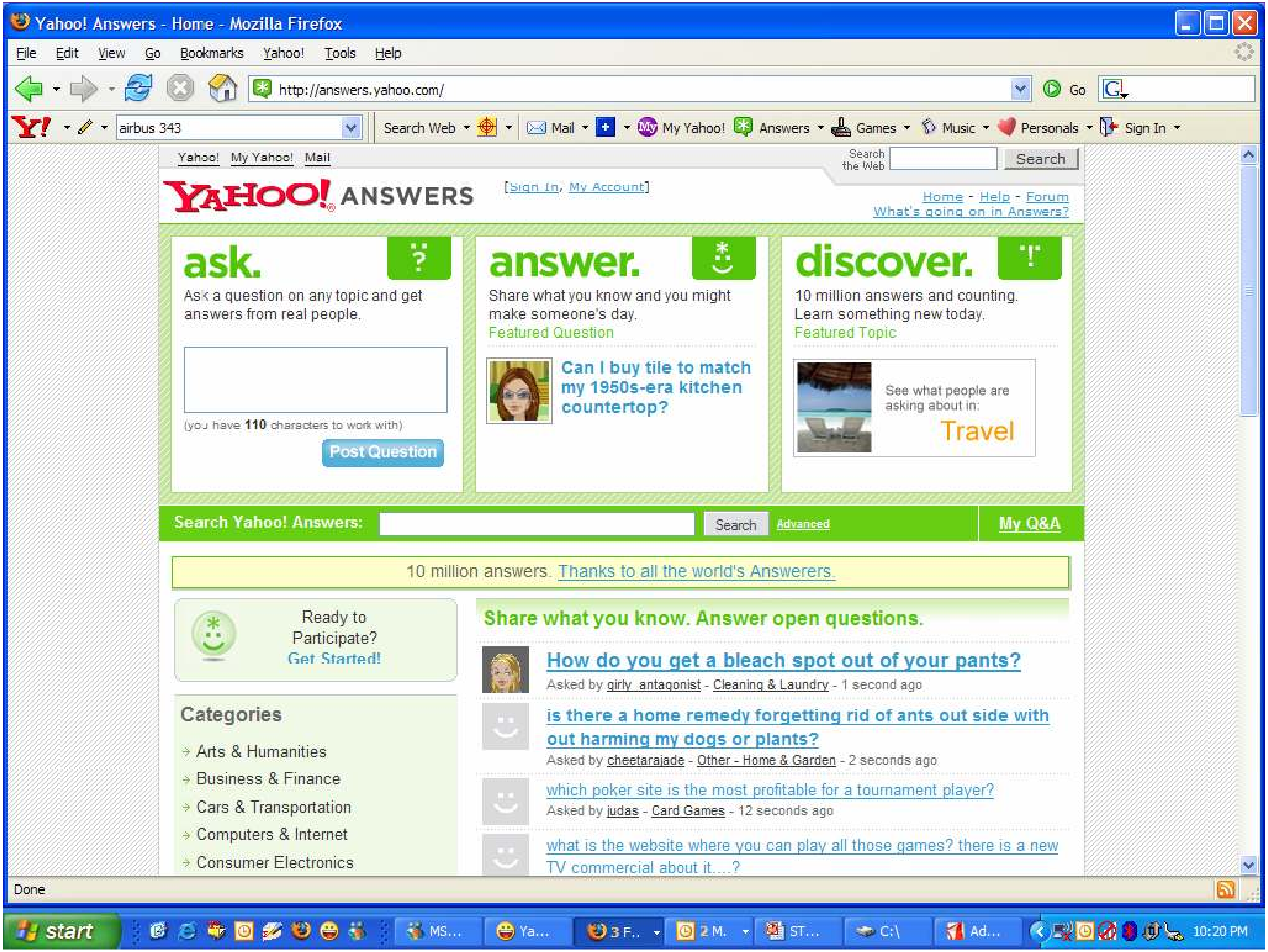
www.ser.org/ - 17k - Cached - Similar pages

SER Solutions, Inc.

Challenges in Tag-Based Search

- How do we use these tags better?
- How do you cope with spam?
- What's the ratings and reputation system?

- The bigger challenge: where else can you exploit the power of the people?
- What are the incentive mechanisms?
  - Luis von Ahn (CMU): The ESP Game



ask.

Ask a question on any topic and get answers from real people.

(you have **110** characters to work with)

answer.

Share what you know and you might make someone's day.

Featured Question

Can I buy tile to match my 1950s-era kitchen countertop?

discover.

10 million answers and counting. Learn something new today.

Featured Topic

See what people are asking about in:

Travel

10 million answers. Thanks to all the world's Answerers.

Ready to Participate?
Get Started!

Categories

- Arts & Humanities
- Business & Finance
- Cars & Transportation
- Computers & Internet
- Consumer Electronics

Share what you know. Answer open questions.

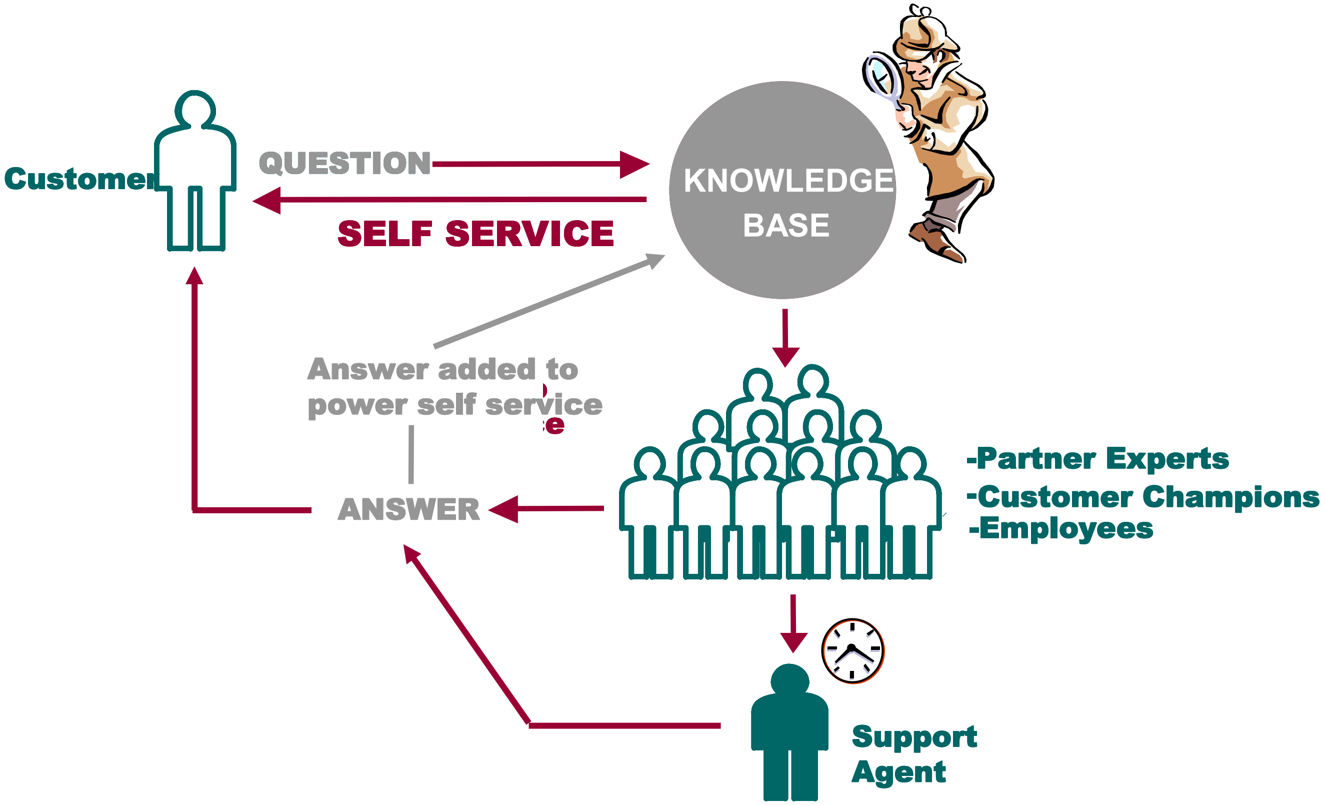
How do you get a bleach spot out of your pants?
Asked by girly_antagonist - Cleaning & Laundry - 1 second ago

is there a home remedy forgetting rid of ants out side with out harming my dogs or plants?
Asked by cheetarajade - Other - Home & Garden - 2 seconds ago

which poker site is the most profitable for a tournament player?
Asked by judas - Card Games - 12 seconds ago

what is the website where you can play all those games? there is a new TV commercial about it....?

How It Works



Customer

QUESTION

SELF SERVICE

KNOWLEDGE BASE

Answer added to power self service

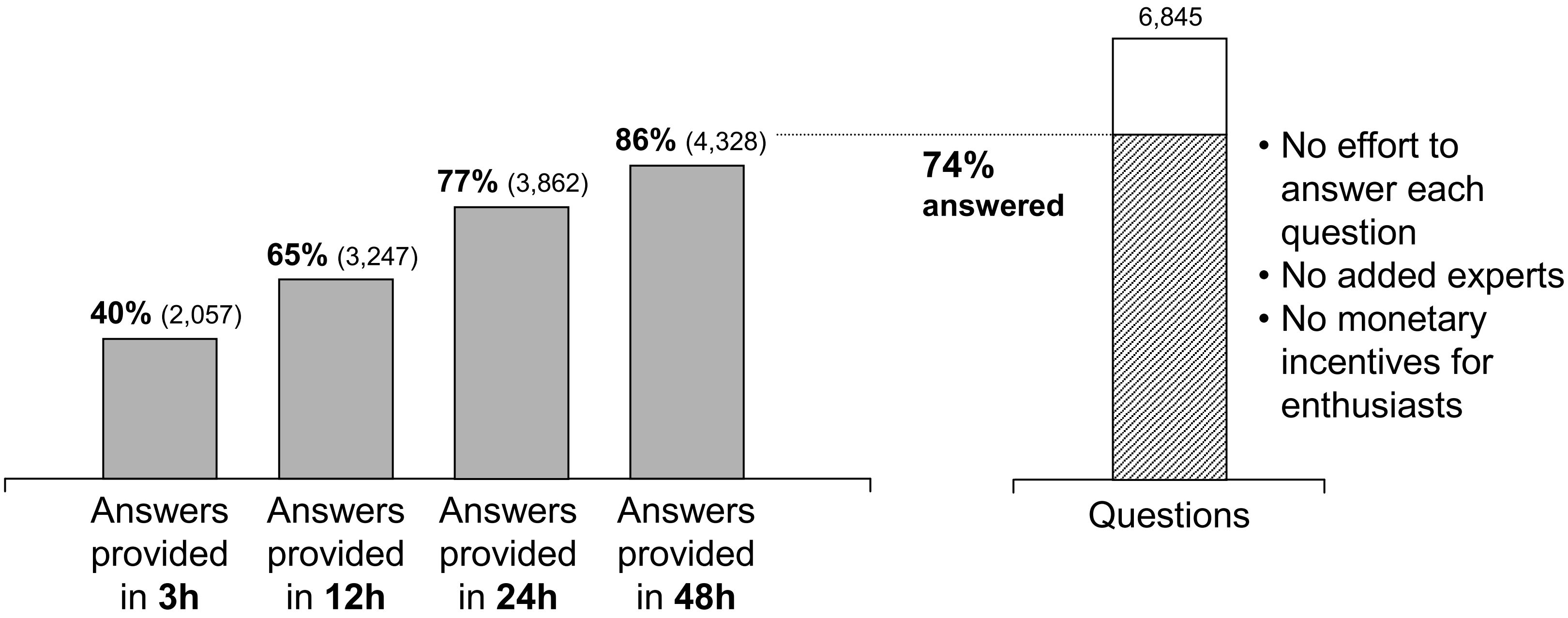
ANSWER

-Partner Experts
-Customer Champions
-Employees

Support Agent

Timely Answers

▶ 77% of answers provided within 24h



- No effort to answer each question
- No added experts
- No monetary incentives for enthusiasts

Interesting Problems

- Question categorization
- Detecting undesirable questions & answers
- Identifying “trolls”
- Ranking results in Answers search
- Finding related questions
- Estimating question & answer quality

(Byron Dom: SIGIR talk)

Supplying Search Content

- As information discovery and search become task-oriented, we need to find ways to create semantically rich summaries that address the user's information needs. Three ways to do this:
  - Editorial, Extraction, **UGC**
    - Opportunity to focus creation of structured UGC feeds directly into this growing need!

Challenge: Design social interactions that lead to creation and maintenance of high-quality structured content

Better Search via Information Extraction

- Extract, then exploit, structured data from raw text:

For years, Microsoft Corporation CEO Bill Gates was against open source. But today he appears to have changed his mind. "We can be open source. We love the concept of shared source," said Bill Veghte, a Microsoft VP. "That's a super-important shift for us in terms of code access.“ Richard Stallman, founder of the Free Software Foundation, countered saying…

Select Name
From PEOPLE
Where Organization = ‘Microsoft’

PEOPLE

Name	Title	Organization
Bill Gates	CEO	Microsoft
Bill Veghte	VP	Microsoft
Richard Stallman	Founder	Free Soft..

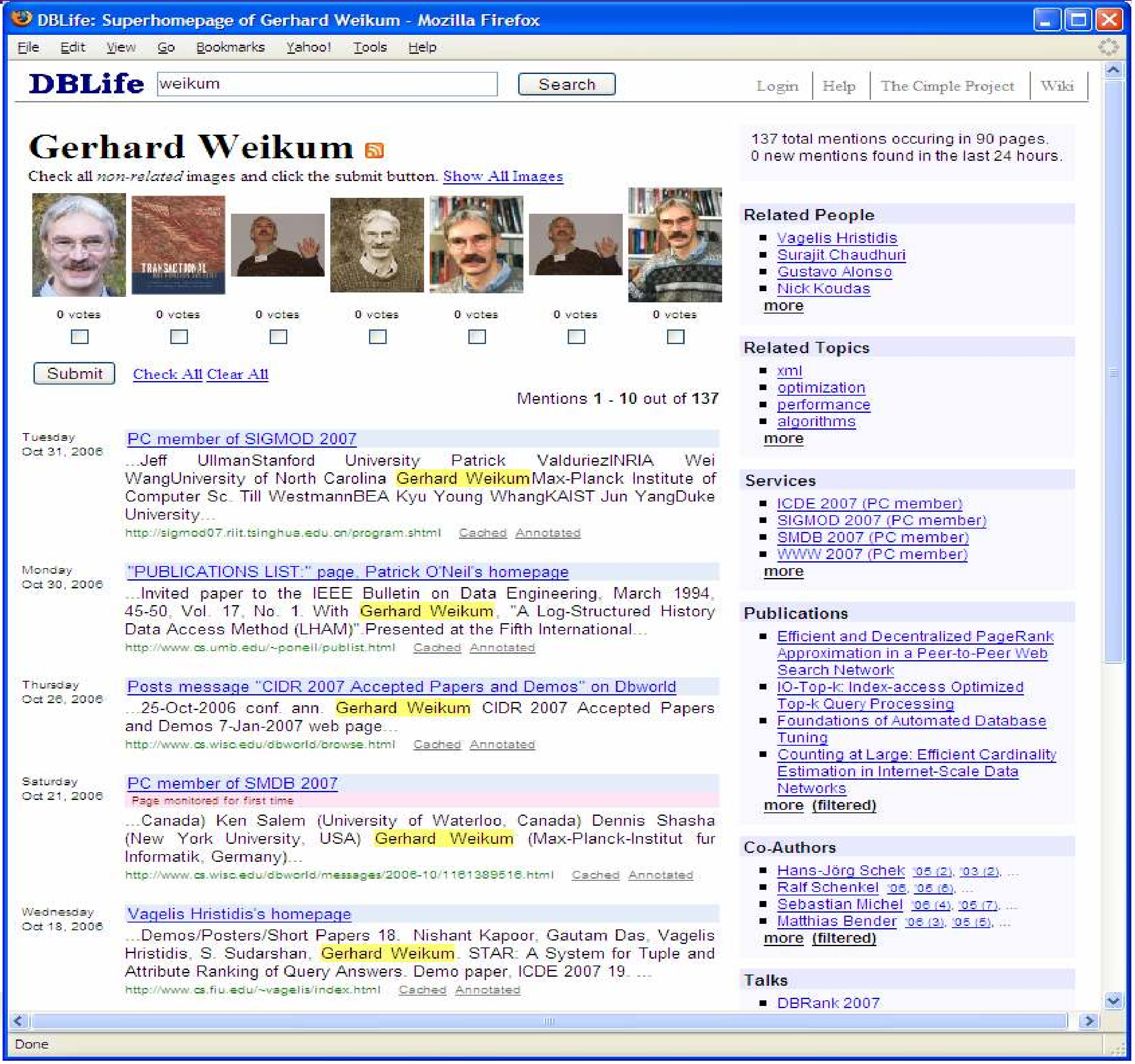
Bill Gates

Bill Veghte

(from Cohen’s IE tutorial, 2003)

DBLife

- Integrated information about a (focused) real-world community
- Collaboratively built and maintained by the community
- **Semantic web via extraction & community**



Gerhard Weikum

Check all *non-related* images and click the submit button. Show All Images

0 votes	0 votes	0 votes	0 votes	0 votes	0 votes	0 votes
□	□	□	□	□	□	□

Submit   Check All Clear All

Mentions **1 - 10** out of **137**

Tuesday
Oct 31, 2006

PC member of SIGMOD 2007

...Jeff UllmanStanford University Patrick ValduriezINRIA Wei WangUniversity of North Carolina Gerhard WeikumMax-Planck Institute of Computer Sc. Till WestmannBEA Kyu Young WhangKAIST Jun YangDuke University...

http://sigmod07.riit.tsinghua.edu.cn/program.shtml Cached Annotated

Monday
Oct 30, 2006

"PUBLICATIONS LIST:" page, Patrick O'Neil's homepage

...Invited paper to the IEEE Bulletin on Data Engineering, March 1994, 45-50, Vol. 17, No. 1. With Gerhard Weikum, "A Log-Structured History Data Access Method (LHAM)".Presented at the Fifth International...

http://www.cs.umb.edu/~poneil/publist.html Cached Annotated

Thursday
Oct 26, 2006

Posts message "CIDR 2007 Accepted Papers and Demos" on Dbworld

...25-Oct-2006 conf. ann. Gerhard Weikum CIDR 2007 Accepted Papers and Demos 7-Jan-2007 web page...

http://www.cs.wisc.edu/dbworld/browse.html Cached Annotated

Saturday
Oct 21, 2006

PC member of SMDB 2007

Page monitored for first time

...Canada) Ken Salem (University of Waterloo, Canada) Dennis Shasha (New York University, USA) Gerhard Weikum (Max-Planck-Institut fur Informatik, Germany)...

http://www.cs.wisc.edu/dbworld/messages/2006-10/1161389516.html Cached Annotated

Wednesday
Oct 18, 2006

Vagelis Hristidis's homepage

...Demos/Posters/Short Papers 18. Nishant Kapoor, Gautam Das, Vagelis Hristidis, S. Sudarshan, Gerhard Weikum. STAR: A System for Tuple and Attribute Ranking of Query Answers. Demo paper, ICDE 2007 19. ...

http://www.cs.fiu.edu/~vagelis/index.html Cached Annotated

137 total mentions occuring in 90 pages.
0 new mentions found in the last 24 hours.

Related People
- Vagelis Hristidis
- Surajit Chaudhuri
- Gustavo Alonso
- Nick Koudas

more

Related Topics
- xml
- optimization
- performance
- algorithms

more

Services
- ICDE 2007 (PC member)
- SIGMOD 2007 (PC member)
- SMDB 2007 (PC member)
- WWW 2007 (PC member)

more

Publications
- Efficient and Decentralized PageRank Approximation in a Peer-to-Peer Web Search Network
- IO-Top-k: Index-access Optimized Top-k Query Processing
- Foundations of Automated Database Tuning
- Counting at Large: Efficient Cardinality Estimation in Internet-Scale Data Networks

more (filtered)

Co-Authors
- Hans-Jörg Schek '05 (2), '03 (2), ...
- Ralf Schenkel '06, '05 (6), ...
- Sebastian Michel '06 (4), '05 (7), ...
- Matthias Bender '06 (3), '05 (5), ...

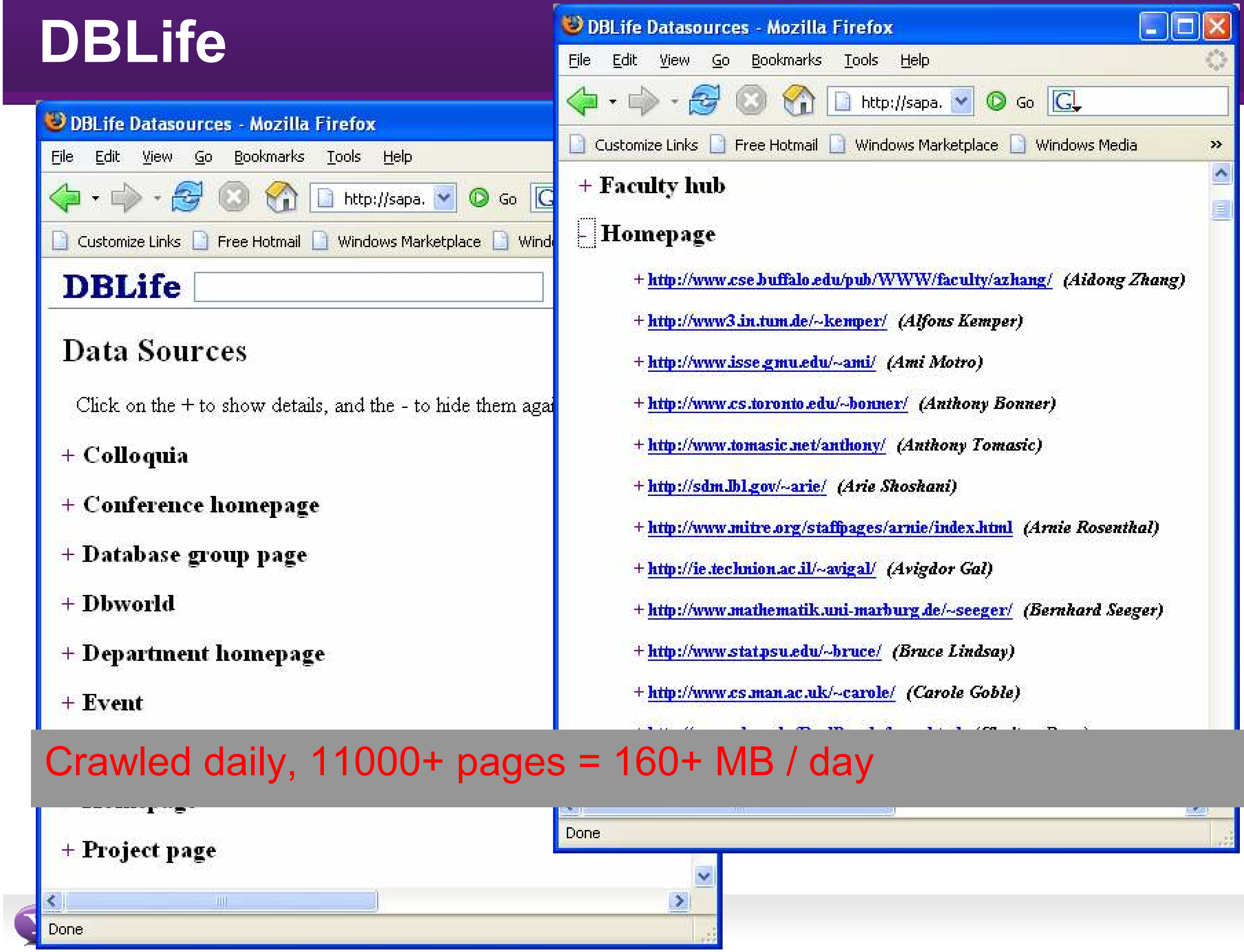
more (filtered)

Talks
- DBRank 2007

The DBLife Portal

- Faculty: AnHai Doan & Raghu Ramakrishnan
- Students: P. DeRose, W. Shen, F. Chen, R. McCann, Y. Lee, M. Sayyadian
- Prototype system up and running since early 2005
- Plan to release a public version of the system in Spring 2007
- 1164 sources, crawled daily, 11000+ pages / day
- 160+ MB, 121400+ people mentions, 5600+ persons
- See DE overview article, CIDR 2007 demo

DBLife



DBLife

Data Sources

Click on the + to show details, and the - to hide them agai

- + **Colloquia**
- + **Conference homepage**
- + **Database group page**
- + **Dbworld**
- + **Department homepage**
- + **Event**
- + **Homepage**
- + **Project page**

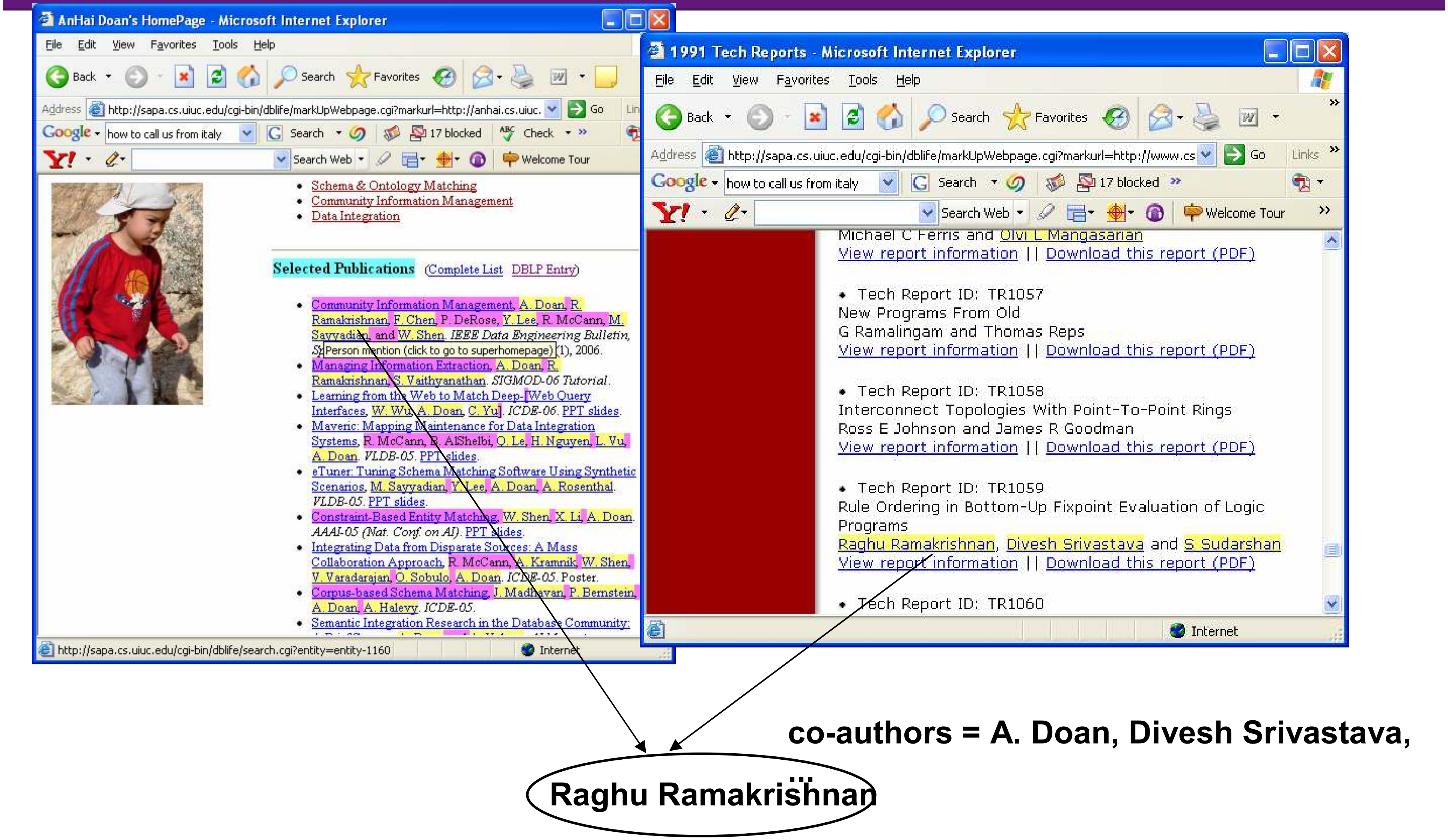
+ **Faculty hub**

Homepage

- + http://www.cse.buffalo.edu/pub/WWW/faculty/azhang/ *(Aidong Zhang)*
- + http://www3.in.tum.de/~kemper/ *(Alfons Kemper)*
- + http://www.isse.gmu.edu/~ami/ *(Ami Motro)*
- + http://www.cs.toronto.edu/~bonner/ *(Anthony Bonner)*
- + http://www.tomasic.net/anthony/ *(Anthony Tomasic)*
- + http://sdm.lbl.gov/~arie/ *(Arie Shoshani)*
- + http://www.mitre.org/staffpages/arnie/index.html *(Arnie Rosenthal)*
- + http://ie.technion.ac.il/~avigal/ *(Avigdor Gal)*
- + http://www.mathematik.uni-marburg.de/~seeger/ *(Bernhard Seeger)*
- + http://www.stat.psu.edu/~bruce/ *(Bruce Lindsay)*
- + http://www.cs.man.ac.uk/~carole/ *(Carole Goble)*

Crawled daily, 11000+ pages = 160+ MB / day

Entity Resolution



AnHai Doan's HomePage - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://sapa.cs.uiuc.edu/cgi-bin/dblife/markUpWebpage.cgi?markurl=http://anhai.cs.uiuc.
Go
Google how to call us from italy
Search
17 blocked
Check
Search Web
Welcome Tour

- Schema & Ontology Matching
- Community Information Management
- Data Integration

Selected Publications (Complete List DBLP Entry)

- Community Information Management, A. Doan, R. Ramakrishnan, F. Chen, P. DeRose, Y. Lee, R. McCann, M. Sayyadian, and W. Shen. *IEEE Data Engineering Bulletin, S*Person mention (click to go to superhomepage)(1), 2006.
- Managing Information Extraction, A. Doan, R. Ramakrishnan, S. Vaithyanathan. *SIGMOD-06 Tutorial*.
- Learning from the Web to Match Deep-[Web Query Interfaces, W. Wu, A. Doan, C. Yu]. *ICDE-06*. PPT slides.
- Maveric: Mapping Maintenance for Data Integration Systems, R. McCann, B. AlShelbi, O. Le, H. Nguyen, L. Vu, A. Doan. *VLDB-05*. PPT slides.
- eTuner: Tuning Schema Matching Software Using Synthetic Scenarios, M. Sayyadian, Y. Lee, A. Doan, A. Rosenthal. *VLDB-05*. PPT slides.
- Constraint-Based Entity Matching, W. Shen, X. Li, A. Doan. *AAAI-05 (Nat. Conf. on AI)*. PPT slides.
- Integrating Data from Disparate Sources: A Mass Collaboration Approach, R. McCann, A. Kramnik, W. Shen, V. Varadarajan, O. Sobulo, A. Doan. *ICDE-05*. Poster.
- Corpus-based Schema Matching, J. Madhavan, P. Bernstein, A. Doan, A. Halevy. *ICDE-05*.
- Semantic Integration Research in the Database Community:

http://sapa.cs.uiuc.edu/cgi-bin/dblife/search.cgi?entity=entity-1160
Internet
1991 Tech Reports - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
Address http://sapa.cs.uiuc.edu/cgi-bin/dblife/markUpWebpage.cgi?markurl=http://www.cs
Go
Links
Google how to call us from italy
Search
17 blocked
Search Web
Welcome Tour

Michael C Ferris and Olvi L Mangasarian
View report information || Download this report (PDF)

- Tech Report ID: TR1057
New Programs From Old
G Ramalingam and Thomas Reps
View report information || Download this report (PDF)

- Tech Report ID: TR1058
Interconnect Topologies With Point-To-Point Rings
Ross E Johnson and James R Goodman
View report information || Download this report (PDF)

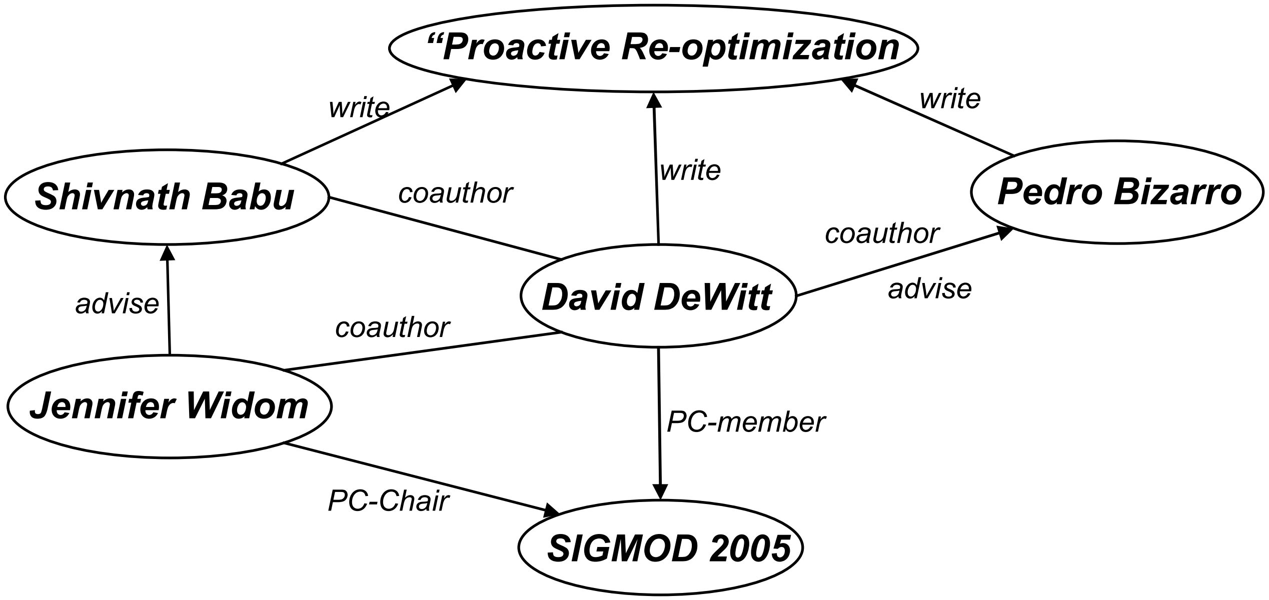
- Tech Report ID: TR1059
Rule Ordering in Bottom-Up Fixpoint Evaluation of Logic Programs
Raghu Ramakrishnan, Divesh Srivastava and S Sudarshan
View report information || Download this report (PDF)

- Tech Report ID: TR1060

co-authors = A. Doan, Divesh Srivastava, ...

Raghu Ramakrishnan

Resulting ER Graph



Challenges

- **Extraction**
  - Domain-level vs. site-level
  - Blending extraction with other sources (feeds, wiki-style user edits)
  - Compositional, customizable approach to extraction planning
    - Cannot afford to implement extraction afresh in each application!
- **Maintenance of extracted information**
  - Managing information Extraction
  - Incremental maintenance of “extracted views” at large scales
  - Mass Collaboration—community-based maintenance
- **Exploitation**
  - Search/query over extracted structures in a community
  - Search across communities—semantic web through the back door!
  - Detect interesting events and changes

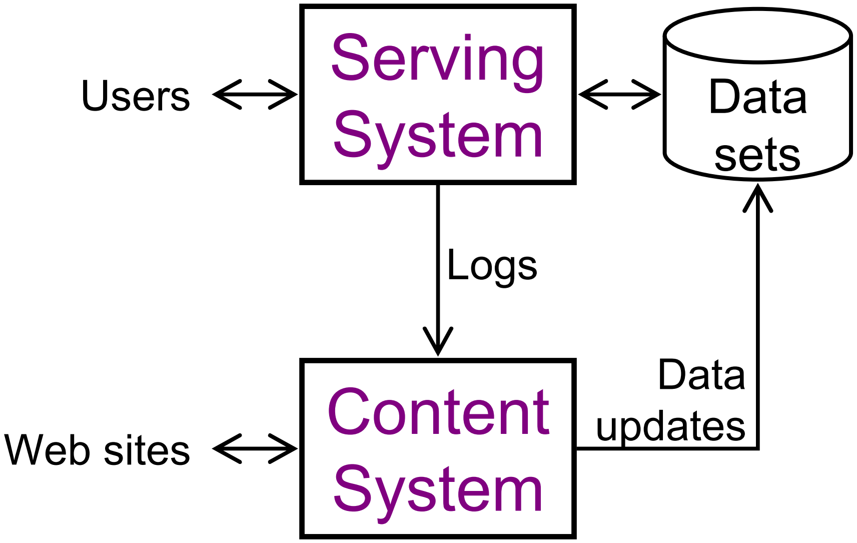
Right in the sweet spot of the “Data to Knowledge” thrust of NSF’s CDI initiative!

Web Data Management:
Massively Distributed Hosted Systems

Two Key Subsystems

- **Serving system**
  - Takes queries and returns results
- **Content system**
  - Gathers input of various kinds (including crawling)
  - Generates the data sets used by serving system
- Both highly parallel

Goal: scaleup.
Hardware increments
support larger loads.



Goal: speedup.
Hardware increments
speed computations.

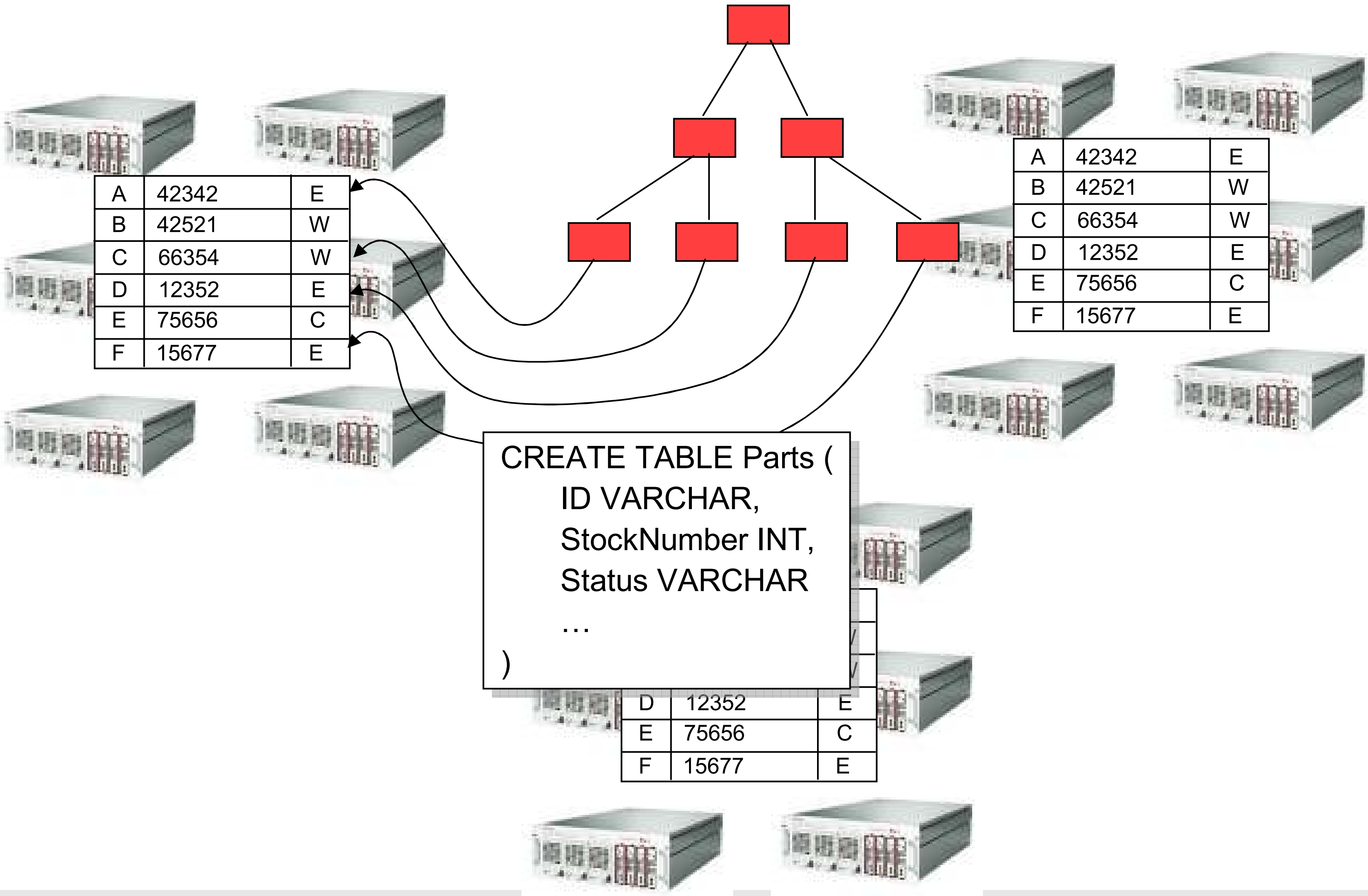
Research (Courtesy: Raymie Stata)

A Case for Hosted Infrastructure

- What does it take to get the Next Great Thing off the ground?
- **Now:**
  - Set up multiple replicas of a clustered data store
  - Set up a system for indexing
  - Set up a system for caching
  - Set up auxiliary DBMS instances for reporting, etc.
  - Set up the feeds and messaging between them
  - Write the application logic
  - Fairly complex system at first line of new code
- **Our vision:**
  - Write the application logic
  - Use a hosted infrastructure to store and query your data

  - Or, as Joshua Shachter puts it: “The next cool thing shouldn’t take a team of 30, it should be three guys, PHP and a long weekend”

The PNUTS Project



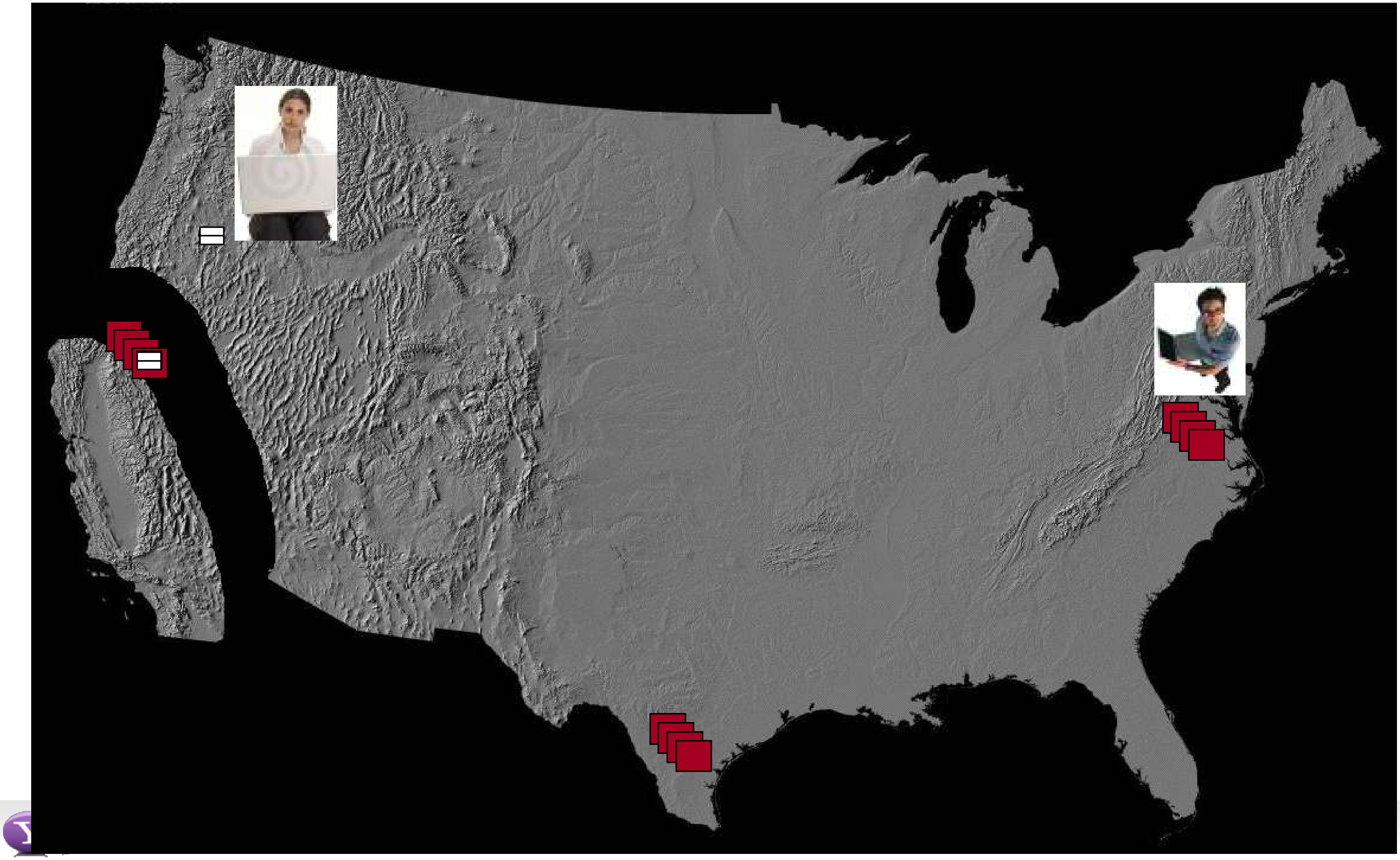
A	42342	E
B	42521	W
C	66354	W
D	12352	E
E	75656	C
F	15677	E

A	42342	E
B	42521	W
C	66354	W
D	12352	E
E	75656	C
F	15677	E

```
CREATE TABLE Parts (
        ID VARCHAR,
        StockNumber INT,
        Status VARCHAR
        …
)
```

D	12352	E
E	75656	C
F	15677	E

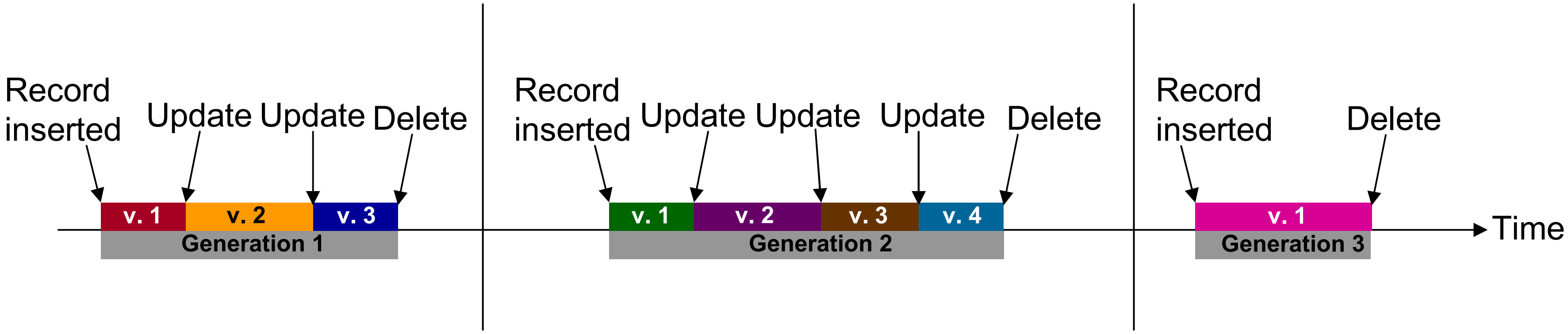
Asynchronous replication



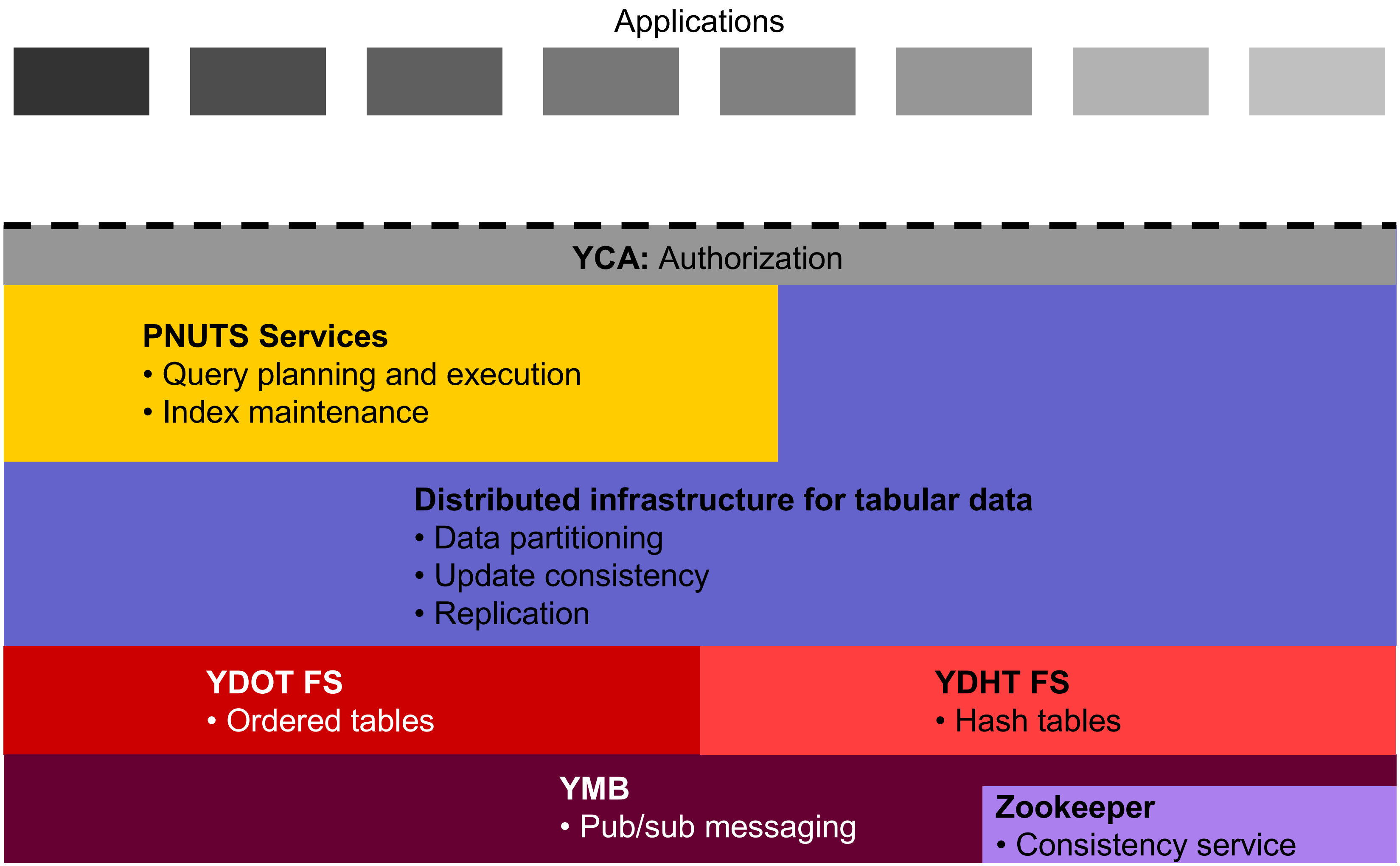
Basic consistency model

- Record lifecycle
  1. Record inserted with a given primary key
  2. Record's non-primary key attributes updated
     - Primary key cannot be updated
  3. Record deleted
  4. Another record with the same primary key may subsequently be inserted

- What happens to a record with primary key "Brian"?



The Big Picture: Sherpa Data Services



Applications

YCA: Authorization

PNUTS Services
- Query planning and execution
- Index maintenance

Distributed infrastructure for tabular data
- Data partitioning
- Update consistency
- Replication

YDOT FS
- Ordered tables

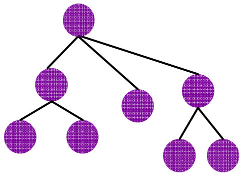
YDHT FS
- Hash tables

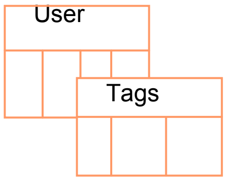
YMB
- Pub/sub messaging

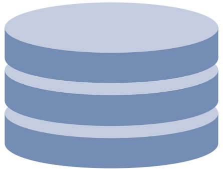
Zookeeper
- Consistency service

Data Analysis Platforms

- Understanding online communities, and provisioning their data needs
  - Exploratory analysis over massive data sets
    - Challenges: Analyze shared, evolving social networks of users, content, and interactions to learn models of individual preferences and characteristics; community structure and dynamics; and to develop robust frameworks for evolution of authority and trust; extracting and exploiting structure from web content …







- **Examples:**
  - **Bigtable, Map-Reduce, Hadoop, PIG**

The Bigger Picture

- Software-as-a-service
  - E.g., Salesforce.com
- Hosted data systems
  - Amazon’s S3 and EC2
- Web application development
  - Ning, Ruby-on-rails
- Change tracking
  - Stream management

Grand Challenge

- **How to maintain and leverage structured, integrated views of web content**
  - – Web meets DB … and neither is ready!
    - Interpreting and integrating information
      - – Result pages that combine information from many sites
    - **Scalable serving of data/relationships**
      - – **Multi-tenancy, QoS, auto-admin, performance**
  - – Beyond search—web as app-delivery channel
    - Data-driven services, not DBMS software
      - – Customizable hosted apps!
    - Desktop → Web-top