



The Future of Information Discovery: Content Optimization, Social Networks, Interactivity---It's All Very Cloudy

Raghu Ramakrishnan

Chief Scientist, Search & Cloud Computing
Yahoo! Fellow



Main Points

- The medium is being massaged
 - i.e., the canvas is no longer passive
 - It is also very personal, and `always on`
 - *(With apologies to Marshall McLuhan)*
- Search and browse converging
 - Semantics is the new frontier
 - Social networks the new distribution channel
- Learning from Big Data is a key
 - Value of data is now recognized
 - Compute platforms will be increasingly data-centric



The Medium is the Message

McLuhan, 1964: Understanding Media: The Extensions of Man

Wikipedia:

- “People tend to focus on the obvious, which is the content, to provide us information, but in the process, we largely miss the structural changes in our affairs that are introduced subtly, or over long periods of time.”
- McLuhan frequently punned on the word "message" changing it to "mass age", "mess age", and "massage"; a later book, [The Medium is the Massage](#) was originally to be titled *The Medium is the Message*.



Yahoo! Data Scale

- **Massive user base and engagement**
 - 650M+ unique users, 11B page visits/month
 - Hundreds of petabytes of storage
 - Hundreds of billions of objects
 - Hundreds of thousands of requests/sec, 200B events/day, 200 PB/day
- **Global**
 - Tens of globally distributed data centers
 - Serving each region at low latencies
- **Challenging Users**
 - Rapidly extracting value from voluminous data
 - Downtime is not an option (outages cost \$millions)
 - Variable usage patterns



Content Optimization



Recommended links

News Interests

Top Searches

+79% clicks
vs. randomly selected

+250% clicks
vs. one size fits all

+43% clicks
vs. editor selected

Key Features

Package Ranker (CORE)

Ranks packages by expected CTR based on data collected every 5 minutes

Dashboard (CORE)

Provides real-time insights into performance by package, segment, and property

Mix Management (Property)

Ensures editorial voice is maintained and user gets a variety of content

Package rotation (Property)

Tracks which stories a user has seen and rotates them after user has seen them for a certain period of time

Key Performance Indicators

Lifts in quantitative metrics

Editorial Voice Preserved



Cloud Structured Storage @ Yahoo!

The image shows a screenshot of the Yahoo! homepage from September 6, 2011. The page is divided into several sections, each with a callout box pointing to it:

- Search**: Points to the search bar at the top.
- Location**: Points to the "Explore Y! Palo Alto" link.
- User Generated Content**: Points to the "9/11 Remembered" section.
- Stock Portfolio**: Points to the "Yahoo! Sites" section.
- IM Buddies**: Points to the "Messenger" link in the left sidebar.
- Reviews and Ratings**: Points to the "omg!" link in the left sidebar.
- Listings**: Points to the "Shopping" link in the left sidebar.
- Saved Locations**: Points to the "Weather (66°F)" link in the left sidebar.
- User Login**: Points to the "Sign Out" link at the top right.
- Social Data**: Points to the "TRENDING NOW" section.
- Targeted Display Ads**: Points to the "The All-New Nissan VERSA SEDAN" advertisement.
- Personalized News**: Points to the "Obama's ratings drop to dismal low" article.
- Personalized Content**: Points to the "VIDEO PICKS" section.
- Smart Ads**: Points to the "Check out the latest models" section.



By the Numbers

1.5+M read rps

4 Key Platforms

10,000+ servers

10 Geo Zones

102B emails/month

13+B Ad serves/day

11B visits/month

~2B User Ids

~700M Uniq Users*

175+M Users in US

285+M Mail users

40+ Countries

All Data from July/Aug. Worldwide unless indicated to the contrary

* Yahoo!-branded sites

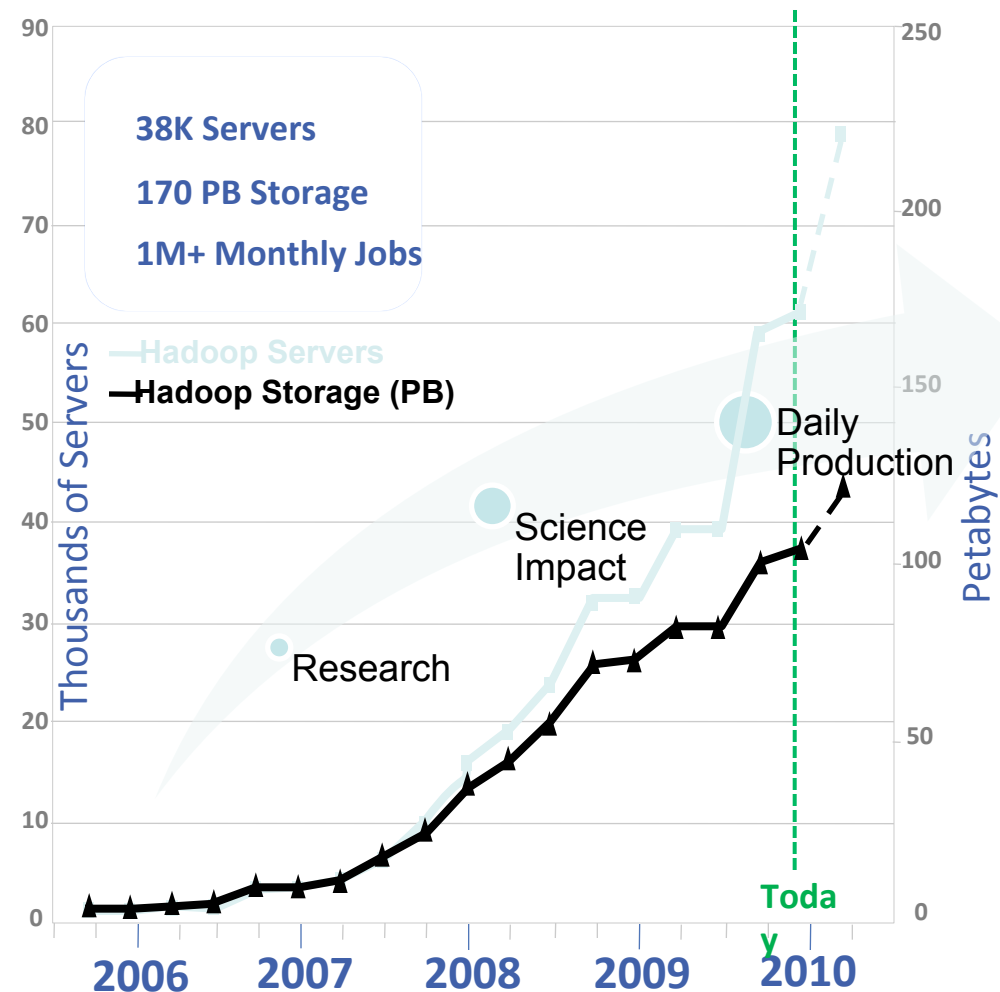


Hadoop: Stability at Scale

Hadoop powers the Yahoo! Network: must be rock-solid

We fix bugs before you see them

- We run very large clusters
- We have a large QA effort
- We run a huge variety of workloads
- Recent spinoff (HortonWorks)





What is PNUTS/Sherpa?

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

Structured, flexible schema

Parallel database

Geographic replication

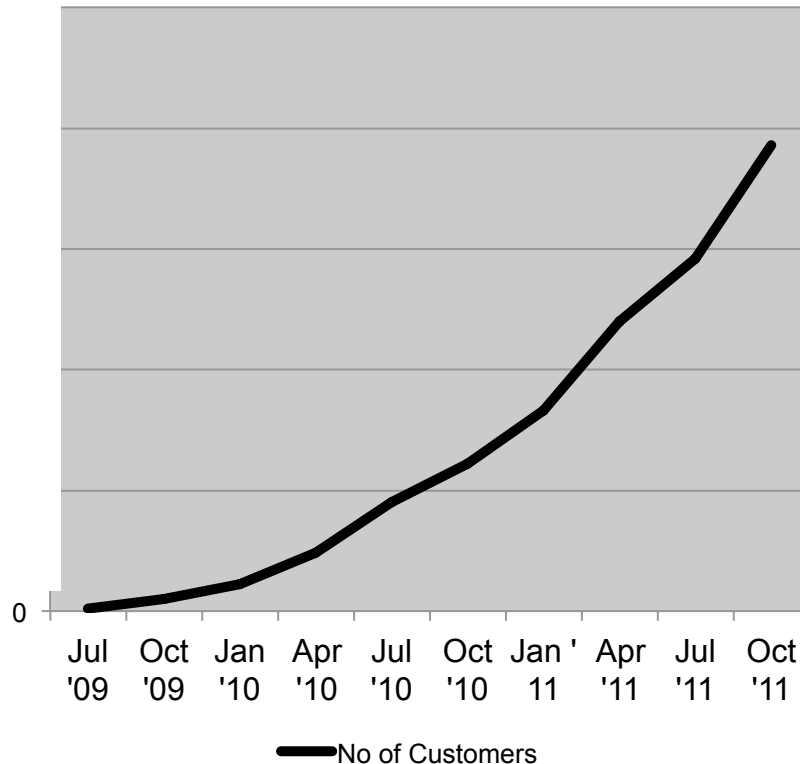
Hosted, managed infrastructure





Sherpa Adoption

In Production



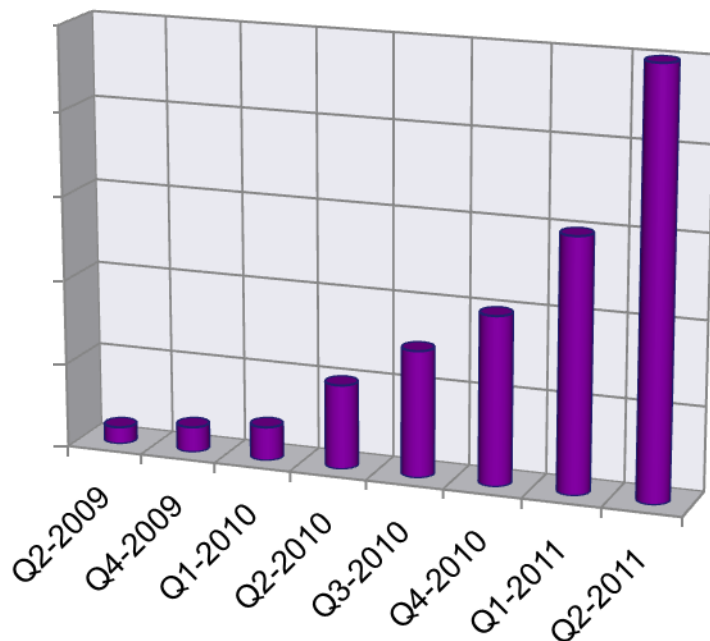
- | | |
|----------------------|------------------------|
| • Abuse Platform | • Pipes |
| • Answers | • Pulse |
| • TW Auctions | • Sports |
| • COKE | • Shopping |
| • ConnectedTV | • UGC Cloud |
| • Groups | • YOS Social Directory |
| • Global Blog Search | • YAP |
| • LocDrop | • YQL |
| • Mobile Platform | • Aqua (Ad Quality) |
| • Mobile Advertising | • Image Search |
| • Consumer Platform | • Hosted Search |
| • Digu | • And others |
| • Search History | |

QoQ 60% growth in projects using Sherpa

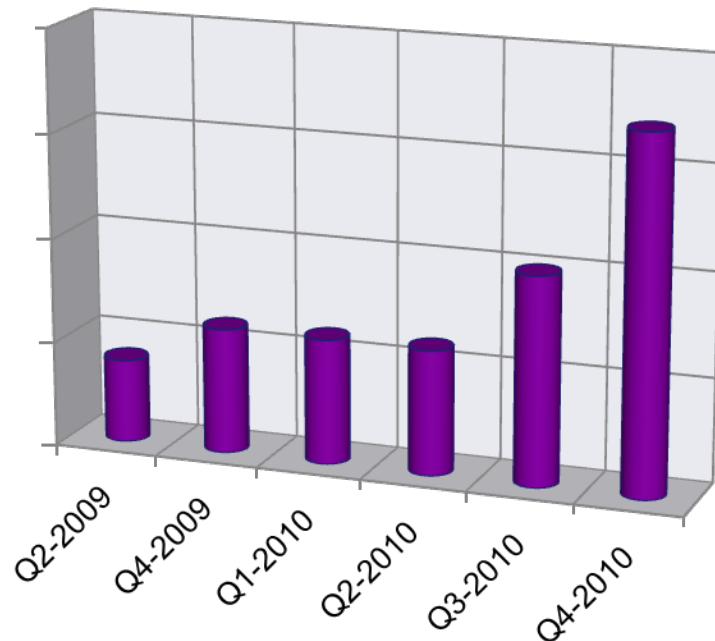


Sherpa Footprint Growth Curve

Storage Nodes



Replicas



10s of data centers, 1000s of machines, PBs of data stored,
100Bs of records, 100Ks of requests/sec

Selective geo-replication, secondary indexes, hashed and ordered tables, flex schemas, eventual and timeline consistency



CORE: Content Optimization



Recommender Approaches



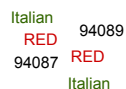
Estimate Most Popular (EMP)

“What’s most engaging overall?”



Behavioral Affinities

“People who did X, did Y”



Attribute Similarities

“Related items with similar metadata”



Social Recommendations

“What are my trusted connections into?”



Business Optimization

“What generates most business value?”



Personalized Recommendations

“What’s most relevant to me based on my interests, attributes and relationships?”



Explore/Exploit

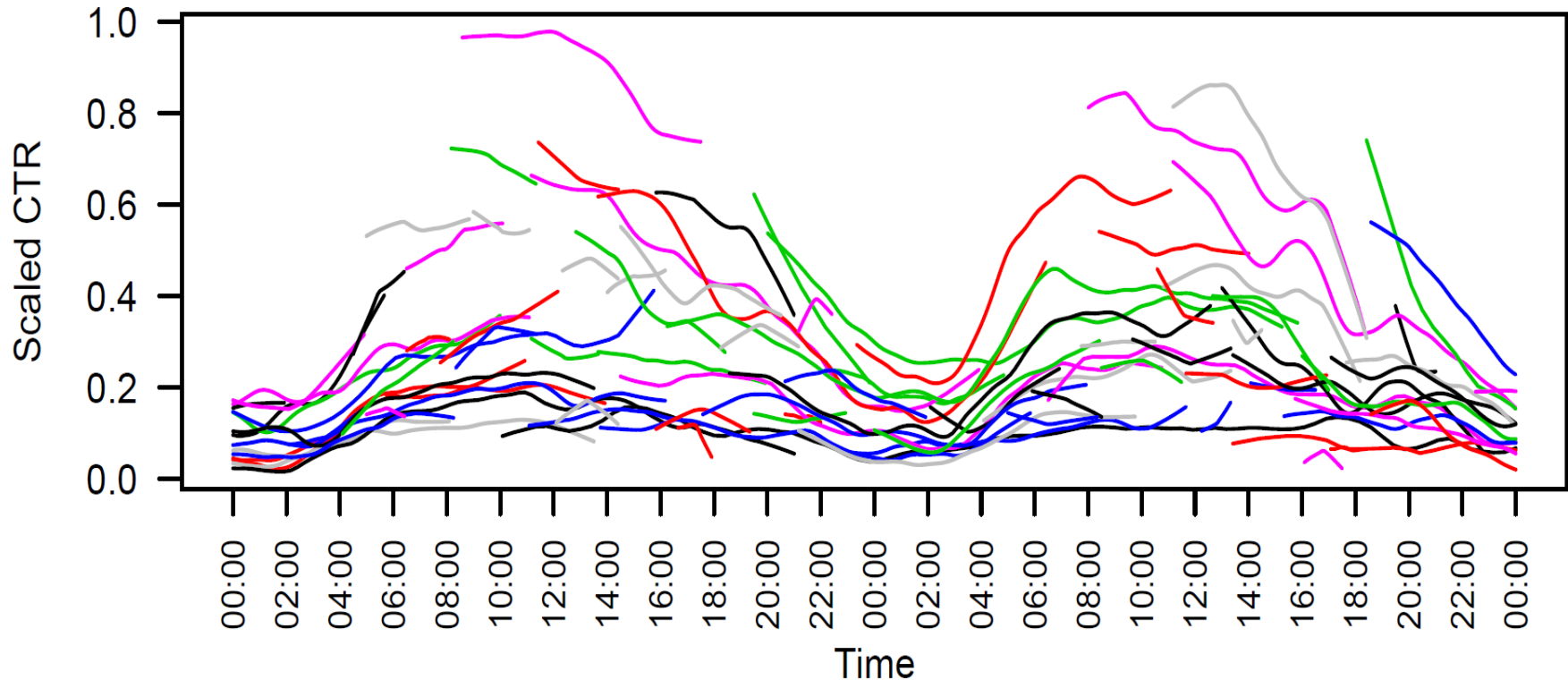
- What is the best strategy for new articles?
 - If we show it and it's bad: lose clicks
 - If we delay and it's good: lose clicks
- Solution: Show it while we don't have much data if it looks promising
 - Classical **multi-armed bandit** type problem
 - Our setup is different than the ones studied in the literature; new ML problem

(Online Models for Content Optimization, NIPS 2008: D. Agarwal, B. Chen, P. Elango, N. Motgi, S. Park, R. Ramakrishnan, S. Roy, J. Zachariah)



CTR Curves for Two Days

Each curve is the CTR of an item in the Today Module over time



Traffic obtained from a controlled randomized experiment

Things to note:

(a) Short lifetimes, (b) temporal effects, (c) often breaking news stories



Degrees of Personalization



Most Popular

Most engaging overall based on objective metrics



Most Popular + Per User History

Engaging overall, and aware of what I've already seen



Light Personalization

More relevant to me based on my age, gender and property usage



Deep Personalization

Most relevant to me based on my deep interests and relationships



Voice and
Business Rules



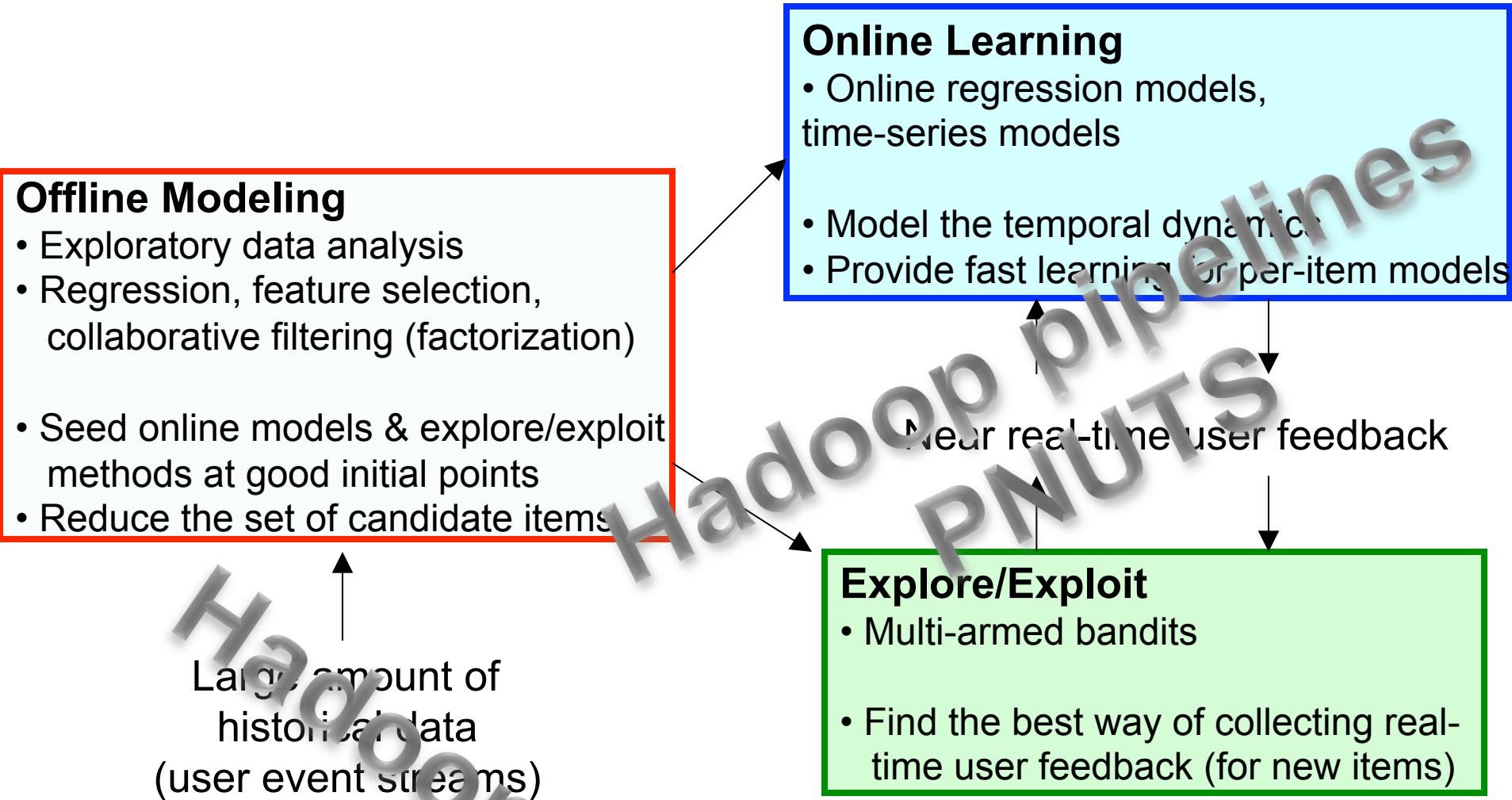
Real-time
Dashboard



Business
Optimization



CORE Modeling Overview



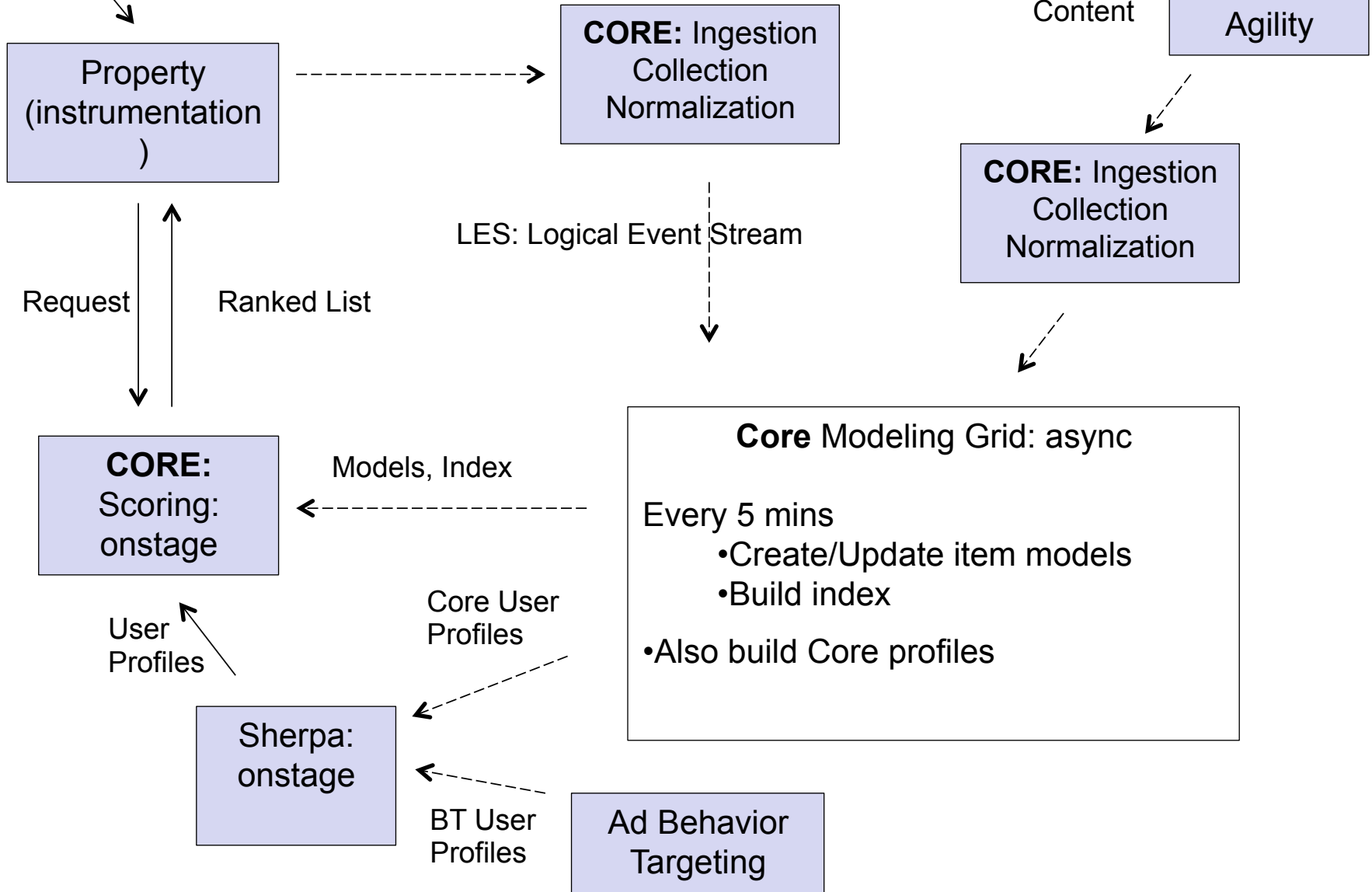


User

CORE: Scoring + Modeling Framework

ULT

(user, article, context, click)

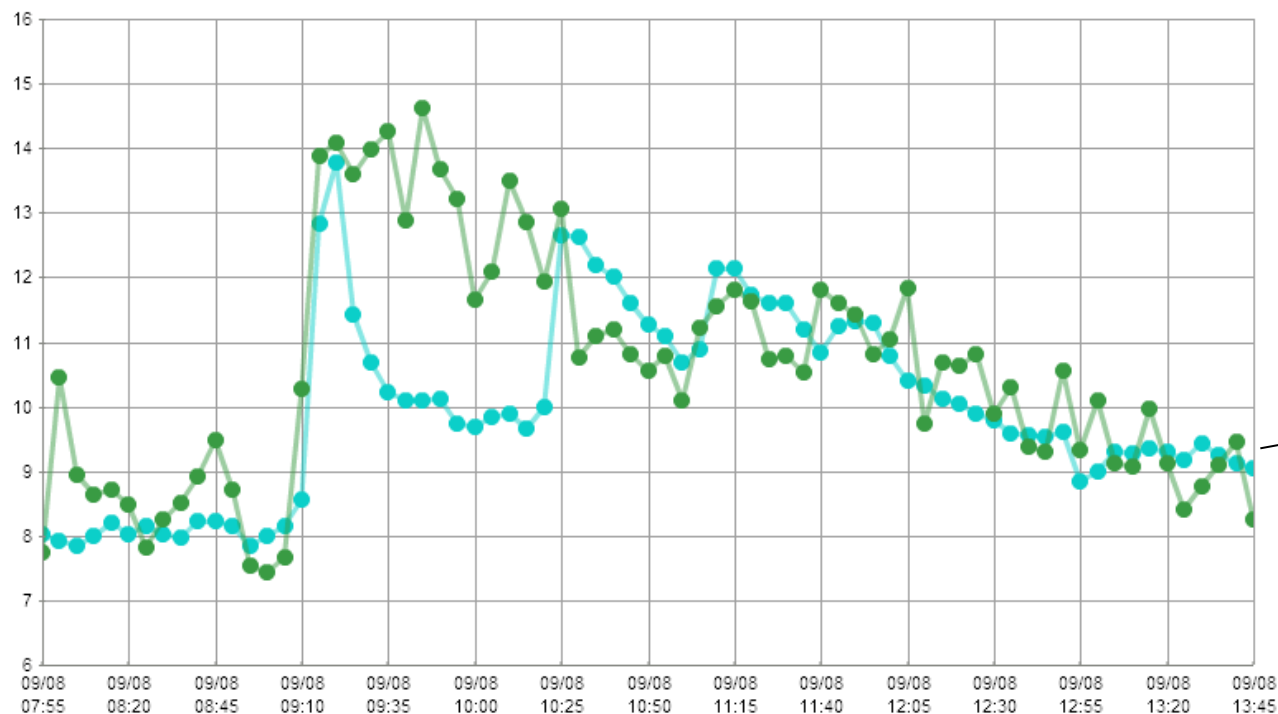




CORE Dashboard: Overall CTR

Compare performance of models and historical benchmarks

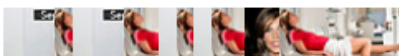
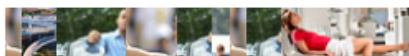
● storyctr for All ● storyctr for c21



Compare buckets and models over time

See which content was promoted most across time

c21



All














Compare bucket metrics

Bucket	Page Views	All Clicks	Story Clicks	Footer Clicks	OverAll CTR	Story CTR	Footer CTR	Lift compared to
All	43,005,783	13,061,688	4,286,407	8,775,281	30.37	9.97	20.4	0
c21	281,821	84,853	29,777	55,076	30.11	10.57	19.54	6.02



CORE Dashboard: Segment Heat Map

Package	male	female	OMG	BUAuto	BUEnt	BU Fin	Health	BUSport+	NBA	BUTrav	ALL
	408,260 18,440 0.0452 8.477	390,404 14,449 0.037 -11.113	270,039 16,940 0.0627 50.661	121,080 7,389 0.061 46.564	270,038 16,940 0.0627 50.661	325,873 20,012 0.0614 47.488	195,796 12,763 0.0652 56.553	350,152 21,454 0.0613 47.152	132,916 9,457 0.0712 70.879	123,388 7,896 0.064 53.691	923,611 38,457 0.0416 0
	1 8,067 852 0.1096 153.654	1 7,657 674 0.068 111.405	1 5,125 720 0.1405 237.406	1 2,382 286 0.1201 188.362	1 5,125 720 0.1405 237.406	1 6,415 858 0.1337 221.221	1 3,769 532 0.1412 239	1 6,750 917 0.1359 226.272	1 2,585 385 0.1489 257.696	1 2,490 330 0.1325 218.294	1 18,137 1,738 0.0958 130.143
	5 9,968 644 0.0646 55.164	3 12,847 777 0.0605 45.256	2 8,569 885 0.1033 148.043	4 3,529 326 0.0824 121.86	2 8,569 885 0.1033 148.043	3 9,744 922 0.0946 127.252	3 6,067 643 0.106 154.537	2 10,187 385 0.0986 136.702	5 3,820 420 0.1059 164.058	2 4,037 433 0.1073 157.598	4 25,744 1,595 0.062 48.798
	2 3,326 249 0.0749 79.8	5 3,954 212 0.0536 28.769	5 2,521 231 0.0916 120.066	2 1,004 102 0.1016 143.995	5 2,521 231 0.0916 120.066	5 3,016 276 0.0915 119.782	5 1,860 186 0.1 140.167	3 3,291 310 0.0942 126.229	3 1,141 136 0.1192 186.264	3 1,039 100 0.0962 131.152	3 8,500 541 0.0636 52.859
	11 2,562 133 0.0519 24.677	13 2,004 81 0.0404 -2.926	3 1,250 122 0.0976 134.403	6 629 51 0.0811 94.73	3 1,250 122 0.0976 134.403	4 1,608 151 0.0939 125.53	2 919 103 0.1121 169.175	4 1,669 154 0.0923 121.604	4 655 74 0.113 171.334	4 591 55 0.0931 123.506	10 5,342 252 0.0472 13.295
	3 2,881 206 0.0715 71.727	2 3,242 230 0.0709 70.384	4 2,071 196 0.0946 127.295	3 949 95 0.1001 140.42	4 2,071 196 0.0946 127.295	2 2,614 254 0.0972 133.368	4 1,605 165 0.1028 146.501	5 2,740 239 0.0872 109.489	10 1,036 94 0.0907 117.912	9 958 78 0.0814 95.543	2 7,043 493 0.07 68.114
	6 10,785 649 0.0602 44.523	4 12,768 742 0.0581 39.571	7 8,580 694 0.0809 94.261	7 3,511 283 0.0805 93.584	7 8,580 694 0.0809 94.261	6 9,725 795 0.0817 96.332	6 6,138 550 0.0896 115.204	6 10,670 866 0.0812 94.925	11 3,669 321 0.0875 110.122	5 3,785 339 0.0896 115.104	5 27,331 1,641 0.06 44.2
	10 22,202 1,212 0.0546 31.106	7 23,328 1,200 0.0514 23.543	6 15,593 1,289 0.0827 58.535	5 6,552 533 0.0813 95.374	6 15,593 1,289 0.0827 58.535	7 17,652 1,376 0.078 87.214	8 10,797 915 0.0847 103.532	7 19,050 1,522 0.0799 91.882	9 6,639 604 0.081 118.498	7 6,435 552 0.0883 106.018	6 52,978 2,786 0.0526 26.299
	22 26,685 1,180 0.0435 4.401	10 35,405 1,530 0.0432 3.786	8 19,832 1,572 0.0793 50.371	9 7,844 552 0.0704 69.011	8 19,832 1,572 0.0793 50.371	8 21,743 1,641 0.0755 81.26	7 13,721 1,167 0.0851 104.267	8 22,168 1,743 0.0786 88.836	8 8,249 788 0.0955 129.424	8 8,327 689 0.0827 98.721	18 74,559 3,167 0.0425 2.014
	4 7,745 518 0.0669 60.628	26 7,202 185 0.0257 -38.308	13 4,898 322 0.0657 57.889	15 2,308 148 0.0641 54.007	13 4,898 322 0.0657 57.889	11 6,051 423 0.0699 67.891	19 3,652 235 0.0643 54.544	9 6,436 506 0.0786 88.82	2 2,562 308 0.1202 188.726	12 2,359 169 0.0716 72.057	7 17,235 834 0.0484 16.217
	7 7,699 480 0.0597 43.495	29 7,201 169 0.0235 -43.635	11 4,809 340 0.0707 69.8	10 2,269 158 0.0696 67.239	11 4,809 340 0.0707 69.8	9 6,004 433 0.0721 73.205	14 3,544 243 0.0686 64.674	10 6,247 475 0.076 82.615	6 2,482 257 0.1035 148.682	11 2,329 167 0.0717 72.211	12 17,169 783 0.0456 9.529
	12 7,688 393 0.0597 43.495	8 7,229 336 0.0235 -43.635	8 4,785 363 0.0707 69.8	17 2,280 139 0.0696 67.239	8 4,785 363 0.0707 69.8	12 6,037 403 0.0721 73.205	12 3,501 245 0.0686 64.674	11 6,319 430 0.076 82.615	15 2,397 182 0.1035 148.682	15 2,312 152 0.0717 72.211	8 17,275 833 0.0456 9.529



Examples

- **ACQUISITION:** A “Star Trek” package was #3 with 18-20 demo, #2 with 21-24 demo, but #9 overall. We can acquire younger audiences with targeted content like this.

	325,214	116,525	44,711	97,003	199,032	218,869	220,622	234,471	207,212	259,018	248,445	211,542
8	5,513	5,895	47	487	676	1,047	1,857	1,525	2,651	1,903	1,090	482
17	397	199	5	13	40	77	53	75	159	94	38	13
	0.072	0.0321	0.1054	0.0267	0.0692	0.0735	0.0533	0.0491	0.0441	0.0362	0.027	0.027
	103,206	-9,629	200,197	-24,673	66,973	107,523	50,425	38,688	69,247	24,569	2,124	-23,892
	5,819	5,855	53	549	751	1,099	1,534	1,481	2,715	1,953	1,059	502

- **ENGAGEMENT:** “Kobe’s astonishing shot” was #25 with women, but #5 with men. We can better engage men (or sports fans) by showing more like this, women by showing less.

	8,754	8,485	65	703	1,244	1,781	2,566	2,345	4,002	2,607	1,351
5	486	191	4	27	55	87	92	101	149	94	51
	0.0555	0.0225	0.0515	0.0384	0.0442	0.0438	0.0369	0.0431	0.0372	0.0361	0.0375
	33,528	-45,859	48,011	-7,625	6,337	17,489	-13,767	3,547	-10,453	-13,278	-9,873

- **REACH:** A package about a hair-pulling soccer player was just plain interesting to everyone (#1-3). We can maintain reach by programming content for the mass audience.

	8,415	8,292	54	680	1,267	1,702	2,511	2,253	3,863	2,530	1,342
1	800	583	2	47	97	127	204	163	345	221	137
	0.0951	0.0703	0.037	0.0691	0.0766	0.0746	0.0812	0.0723	0.0893	0.0874	0.1021
	128,654	69,103	-10,92	66,239	84,135	79,468	95,401	74,008	114,802	110,095	145,534



SEARCH TRENDS



Web Images Video Local Shopping News More

eggplant parmigiana baltimore

Search

Options

Start typing to see suggestions.

Explore related concepts:

Pizza mushrooms
veal tomato sauce
Chicken Little Italy
Italian Restaurant Crab Cake

Search Pad

SearchScan - On

30,300 results for
eggplant parmigiana ...

Show All

Los Angeles Times

Local Business Sites

Eggplant Parmigiana Restaurants near Baltimore, MD

Neighborhood

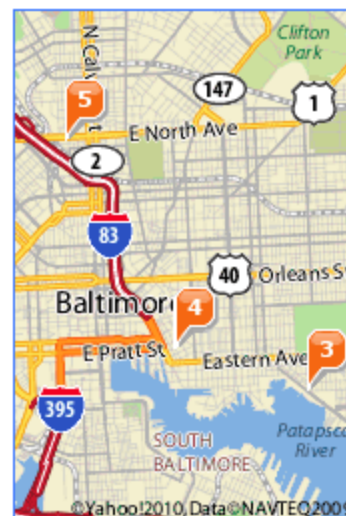
All (36) Canton (2) Charles North (1) Downtown (1)
Abell (1) Central Bal... (4) Chinguapin... (1) Federal Hill (1)

1 **Ciao Bella - Baltimore** ★★★★★ (11)
local.yahoo.com
(410) 685-7733 - 236 S High St, Baltimore, MD
Menu: eggplant parmigiana
4 Reviews | Overview | 2 Photos | Directions

2 **Amicci's** ★★★★★ (20)
amiccis.com
(410) 528-1096 - 231 S High St, Baltimore, MD
Menu: eggplant parmigiana
14 Reviews | Overview | 23 Photos | Directions

3 **Pasticcio** ★★★★★ (8)
local.yahoo.com
(410) 522-7700 - 2400 Boston St, Baltimore, MD
Menu: eggplant parmigiana
5 Reviews | Overview | 3 Photos | Directions

4 **Caesar's Den** ★★★★★ (7)
caesarsden.com
(410) 547-0820 - 223 S High St, Baltimore, MD
Menu: eggplant parmigiana
4 Reviews | Overview | 11 Photos | Directions



YAHOO!

julia roberts

Search

julia roberts twins
julia roberts movies
lyle julia roberts
julia roberts babies
julia roberts henry daniel moder

Explore related concepts:

actor

Episode

Pretty woman

Best Actress

julia roberts photos

Julia Roberts News

julia roberts biography

Julia Fiona Roberts



Search Pad



SearchScan - On

40,500,000 results for
julia roberts

Related People



Scarlett Johansson



Emma Roberts



Hilary Swank



Lindsay Lohan



Tom Hanks



Halle Berry



Julia Roberts - Image Results

News & Photos

Videos

Twitter

[more Julia Roberts photos...](#)

Latest News:

[Hindus concerned about Julia Roberts' 'Eat, Pray, Love'](#) - New Kerala - 6 hours ago[Trailer For 'Eat Pray Love' Starring Julia Roberts](#) - KPBS San Diego - Mar 19 03:18pm[Link Party: Julia Roberts' New Movie Will Teach You How to Live](#) - E! Online - Mar 18 05:48pm[more Julia Roberts news...](#)

Julia Roberts - Wikipedia

[Early life](#) | [Career](#) | [Influence](#) | [Personal life](#)

Julia Fiona Roberts is an American actress. She is known for starring in the romantic comedy *Pretty Woman* opposite Richard Gere, which grossed \$463 million worldwide. After receiving...

en.wikipedia.org/wiki/Julia_Roberts - 122k - [Cached](#)



Yahoo! Search Direct

Yahoo! Search Results - Mozilla Firefox

View Virtual Machine Window

Bookmarks Yahoo! Tools Help

http://search.yahoo.com/search;_ylt=A0oGdV0TuJNNpjMA5CxXNyoA;_ylc=X1MDUCMyMTQyMzU3MDg5BF9yAzIEZnIDeWZwLXQtMzc1BGZyMgNzZy1nYWMEZ

Search Web Mail Shopping My Yahoo! Games Music Answers Sports Sign In

Sign In Help

Bucket: DFD5 New Tracking Ba... bad results / ads or bugs? tell us! [hide]

Make Yahoo! your homepage Mail



QuickApps

Search Pad

SafeSearch - On

1,300,000 results for
julia roberts movies:

Show All

Yahoo! Movies

IMDb

Blockbuster

Related Searches

andra bullock movies

in brockovich

julia roberts filmography

Web Images Video Local Shopping News More

julia roberts

Search

Options

julia roberts

julia roberts movies

julia roberts twins

julia roberts niece

julia roberts ex husband

julia roberts pretty woman

julia roberts brother

julia roberts son

julia roberts jude law

julia roberts 2010 movie

JULIA ROBERTS BROTHER

TOP SITES

Eric Roberts - Wikipedia, the free encyclopedia
[W en.wikipedia.org/wiki/Eric_Roberts](http://en.wikipedia.org/wiki/Eric_Roberts)

Eric Roberts - IMDb
[IMDb www.imdb.com/name/nm0000616/](http://www.imdb.com/name/nm0000616/)

Julia Roberts, Brother Eric Reunited
[P www.people.com/people/article/0,,1007686,00.html](http://www.people.com/people/article/0,,1007686,00.html)

One of the few bankable female stars of the 1990s, actress **Julia Roberts** remained an iconic figure whose assured, winsome performances underscored her undeniable ...
movies.yahoo.com/movie/contributor/1800019215 - [Cached](#)

Julia Roberts Movies - News Results

[Hailee Steinfeld attached to revisionist 'Sleeping Beauty' movie](#)

Entertainment Weekly - Mar 29 07:32pm

[Saoirse Ronan Close To Landing That Other Snow White Gig](#)

TheE!Life - Mar 29 12:26pm

Sponsored Results

[e@ DVD Club](#)
s Movies Like
/Ds for \$1 Each.
[use.com](#)

[ur PC Free](#)
atch TV and
C for Free?
[.com](#)

[See your message here...](#)

YAHOO!

julia roberts

Search

julia roberts twins
julia roberts movies
lyle julia roberts
julia roberts babies
julia roberts henry daniel moder

Explore related concepts:

actor

Episodes

Pretty Woman

Best Actress

julia roberts photos

Julia Roberts News

julia roberts biography

Julia Fiona Roberts

Julia Roberts - Image Results

News & Photos

Videos

Twitter



more Julia Roberts images...

Latest News

Handbook: Everything you need to know about **Julia Roberts' "Eat, Pray, Love"** - New Kerala - 6 hours ago

Trailer For 'Eat Pray Love' Starring **Julia Roberts** - KPBS San Diego - Mar 19 03:18pm

Link Party: **Julia Roberts' New Movie Will Teach You How to Live** - E! Online - Mar 18 05:48pm

more **Julia Roberts** news...

Julia Roberts - Wikipedia

[Early life](#) | [Career](#) | [Influence](#) | [Personal life](#)

Julia Fiona Roberts is an American actress. She is known for starring in the romantic comedy *Pretty Woman* opposite Richard Gere, which grossed \$463 million worldwide. After receiving...

en.wikipedia.org/wiki/Julia_Roberts - 122k - [Cached](#)



Search Pad

SearchScan - On

40,500,000 results for
julia roberts

Related People



Scarlett Johansson



Emma Roberts



Hilary Swank



Lindsay Lohan



Tom Hanks



Halle Berry

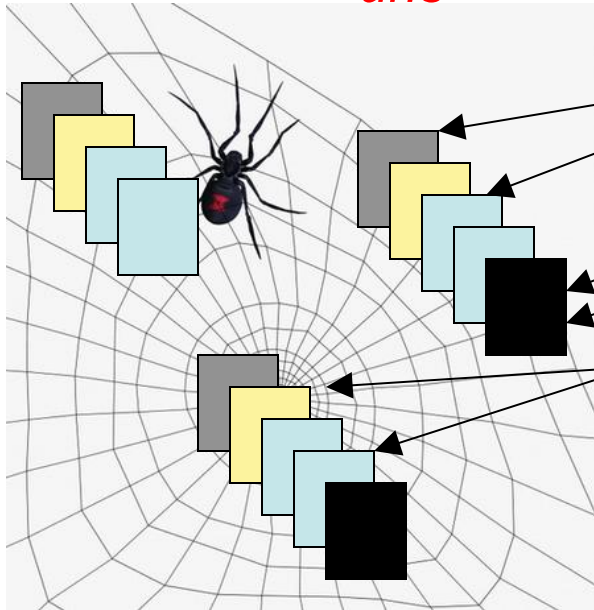




Web of Pages

urls

token



julia roberts

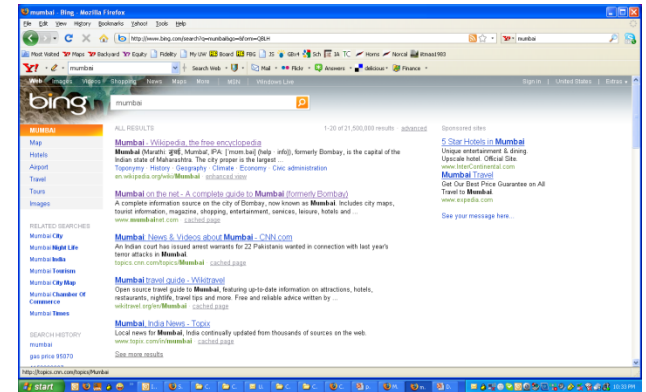
mumbai
restaurant
chinese
san jose

• • •

WEB

INDEX

SERP

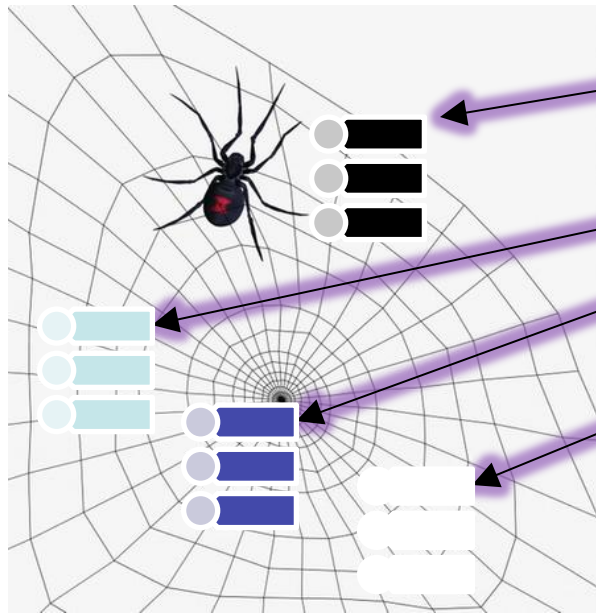




Web of Concepts

structured data

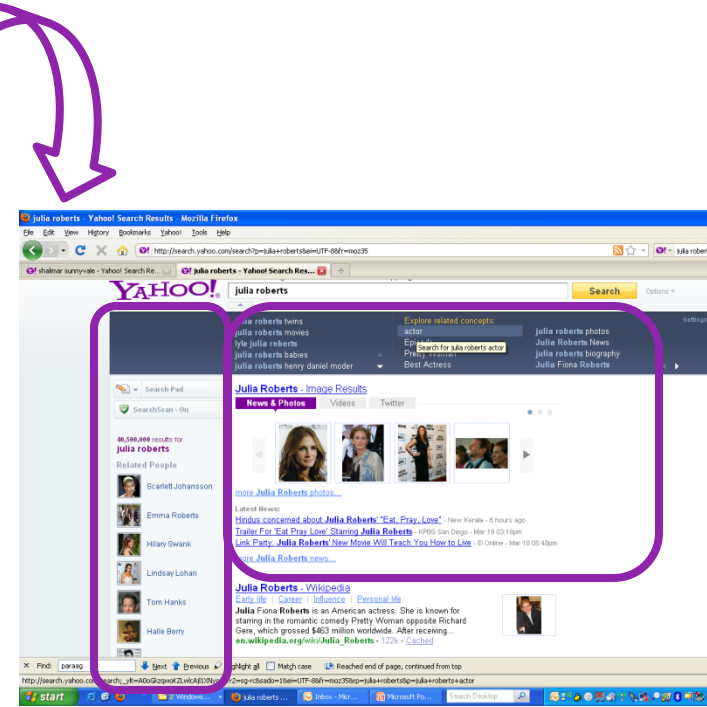
concept



julia roberts

mumbai
restaurant

san jose
● ● ●



Aggregated KB

INDEX

SERP

The “index” is keyed by concept instance, and organizes all relevant information, wherever it is drawn from, in semantically meaningful ways

ANSWERS, NOT LINKS



Web of Concepts

- The web is a source of information about concepts people care about
- The “index” is keyed by concept instance, and organizes all relevant information, wherever it is drawn from, in semantically meaningful ways
 - E.g., Information about Mumbai, or the Dish-Dash restaurant, is organized in terms of meaningful attributes (population, type of cuisine) and can be retrieved and presented to address whatever the user wants to know about Mumbai or Dish-Dash

N. Dalvi et al.: A Web of Concepts, PODS 2009

Web Images Video Local Shopping more

mumbai Search Options Customize

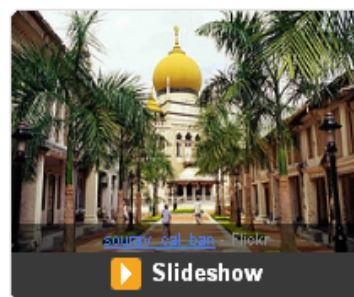
YAHOO!

mumbai university
mtnl mumbai
mumbai attacks
mumbai mirror
mumbai india

Explore related concepts:

Mumbai Central
South Mumbai
population
tourism
Maharashtra
Mumbai, India
streets
Marathi

1 - 10 of 217,000,000 for mumbai (About) - 0.28 s |



Mumbai, India - Visitor Guide

travel.yahoo.com

Hotels Restaurant Guide Flights Map

Top Rated Things To Do (134)

- 1 Gateway of India
- 2 SKS Supercrafts
- 3 Matheran

More Things To Do...

Mumbai - Wikipedia, the free encyclopedia

Toponymy History Geography Climate

Includes geography, economy, climate, civic administration, utility services, people and culture, media, education, sports and recreation, transport, and history ...

en.wikipedia.org/wiki/Mumbai - 458k - Cached



Mumbai on the net - A complete guide to Mumbai (formerly Bombay)

With news, maps, tourist information, emergency contacts, education, utilities, health, hotels

Visting Mumbai?

Find cheap deals on 4000 hotel sites with 1 search! No booking fees.

www.Kayak.com/Hotels

2009 Top Work At Home

Looking For mumbai?

EricGetsCash.com

Hotels in Mumbai

Looking for Mumbai Hotels? Cheap Deals - 35+ Sites - Reviews.

mumbai.hotelhelpers.com

Mumbai

Find Low Rates Worldwide. Flights, Hotels & More at ORBITZ.

www.ORBIZ.com

YAHOO! TRAVEL

Search WEB SEARCH

- Home
- Research
- Book
- Deals
- Themes
- My Travel
- Summer Travel

Travel Guides > Asia > India > Mumbai


Over 19 million Yahoo! users plan trips here

SEARCH

Mumbai, India

HOTELS, VACATIONS, TOURISM

- OVERVIEW
- HOTELS
- ATTRACTIONS
- RESTAURANTS
- TRIP PLANS
- PHOTOS/MDEOS
- MAP
- DEALS

 WEATHER: HAZE, 83F
Annual weather >





Ever since the opening of the Suez Canal in the 1860s, the principal gateway to the Indian subcontinent has been MUMBAI (Bombay) , the city Aldous Huxley famously described as "the most appalling . . . of either hemisphere". Travellers tend to regard time spent here as a rite of passage to be survived rather than savoured. But as the powerhouse of Indian business, industry and trade, and the source of its most seductive media images, the Maharashtrian capital can be a compelling place to kill time. Whether or nor you find the experience enjoyable, however, will depend largely on how well you handle the ... [more >](#)

[ADD TO TRIP](#) [BEEN THERE](#) [WANT TO GO](#)



LIKE THIS PHOTO?  

@sourav_cal_ban - flickr



View all (62) >



Understanding Web Content

with a little help from your friends



Generative Model of the Web

The true world database

Site queries

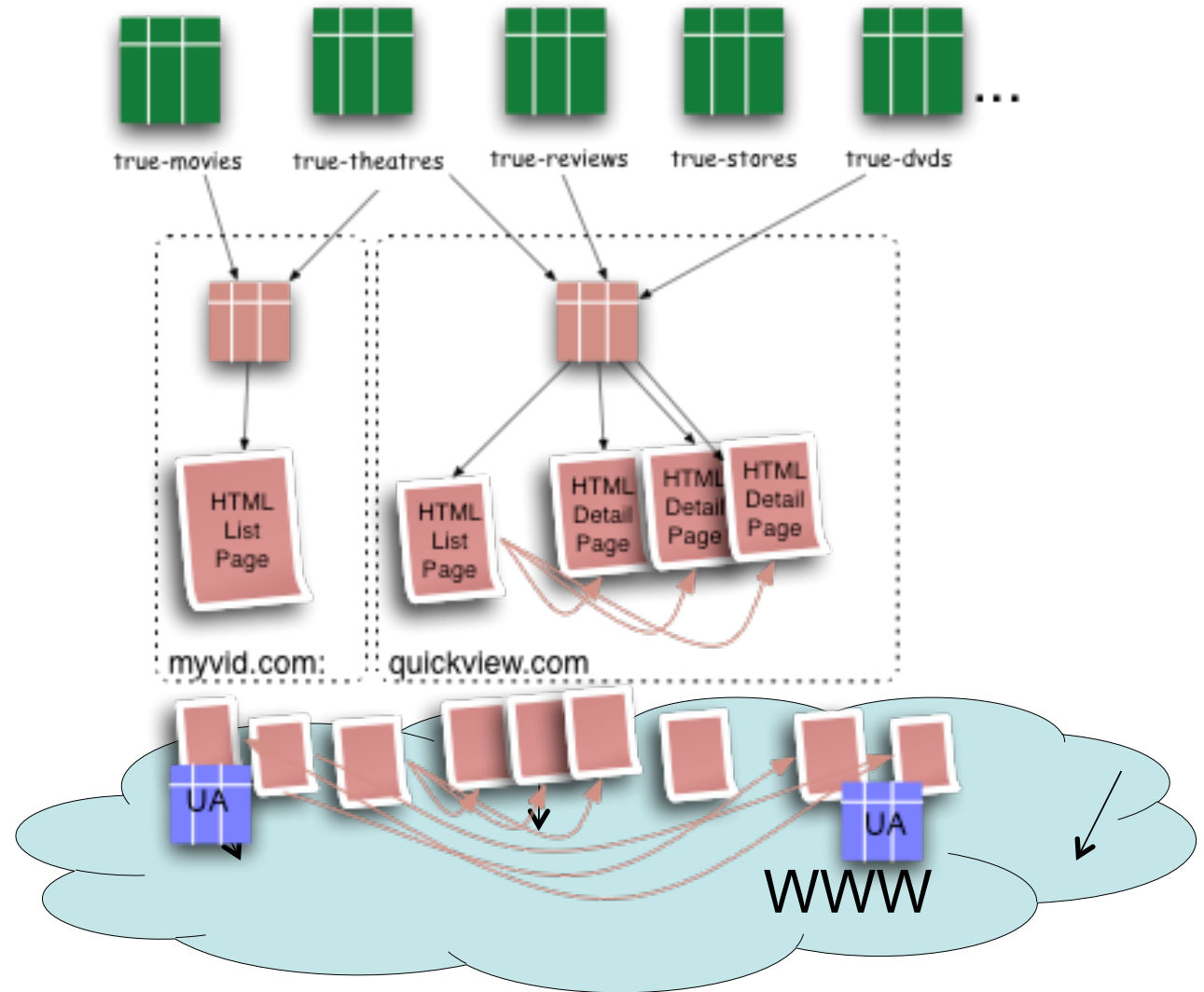
Information loss

Noise addition

Site layout

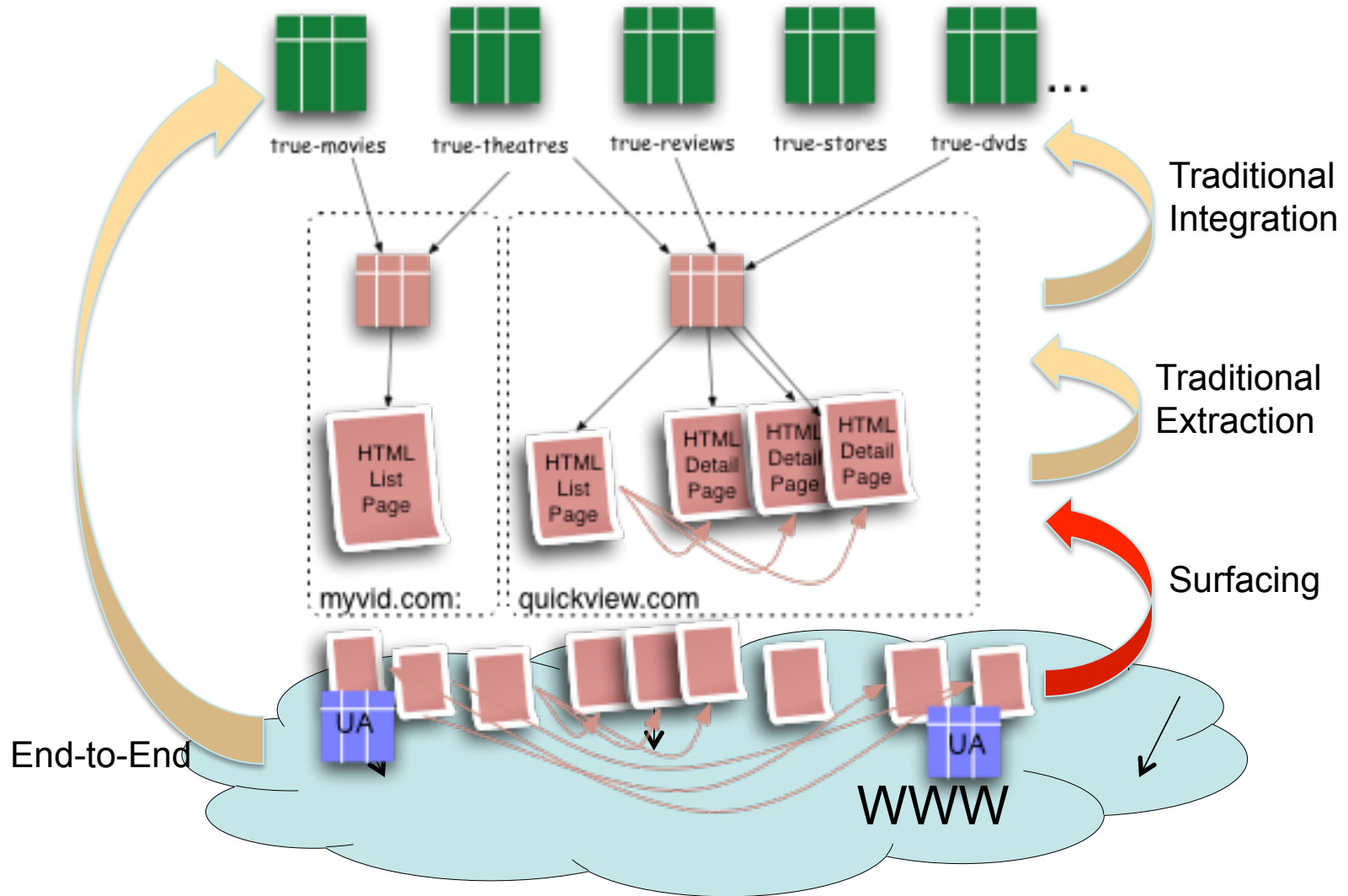
Surround generation

Document Corpus





Surfacing, Extraction, Integration





DBLife: Community Information Mgmt

- Integrated information about a (focused) real-world community
- Collaboratively built and maintained by the community
 - Semantic web, “bottom-up”
- Joint work with AnHai Doan, Pedro Domingos, Warren Shen and others at Wisconsin

DBLife: Superhomepage of Gerhard Weikum - Mozilla Firefox

File Edit View Go Bookmarks Yahoo! Tools Help

DBLife weikum Search Login Help The Cimple Project Wiki

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

☐ ☐ ☐ ☐ ☐ ☐ ☐

[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 Weikum**Max-Planck Institute of Computer Sc. Till WestmannBEA Kyu Young WhangKAIST Jun YangDuke University...
<http://sigmod07.rti.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 für 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 occurring 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](#) [08](#), [05 \(6\)](#), ...
- [Sebastian Michel](#) [08 \(4\)](#), [05 \(7\)](#), ...
- [Matthias Bender](#) [05 \(3\)](#), [05 \(5\)](#), ...

[more](#) [\(filtered\)](#)

Talks

- [DBRank 2007](#)



Summary of Gaps

We have observed three key gaps in the state-of-the-art

Gap 1: Finding the pages to extract from may be a significant portion of the supervision cost

Gap 2: Must extract across diverse content types to gather all attributes

Gap 3: Supervision costs for information integration may exceed costs for extraction



Integration: Entity Extraction/Deduping


AnHai Doan's HomePage - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://sapa.cs.uiuc.edu/cgi-bin/dblife/markUpWebpage.cgi?markurl=http://anhai.cs.uiuc.edu>

Google how to call us from italy Search 17 blocked ABC Check

Y! 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, Special Person mention \(click to go to superhomepage\) \(1\), 2006.](#)
- [Managing Information Extraction, A. Doan, R. Ramakrishnan, S. Vaityanathan, 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, R. 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>

1991 Tech Reports - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://sapa.cs.uiuc.edu/cgi-bin/dblife/markUpWebpage.cgi?markurl=http://www.cs>

Google how to call us from italy Search 17 blocked

Y! 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

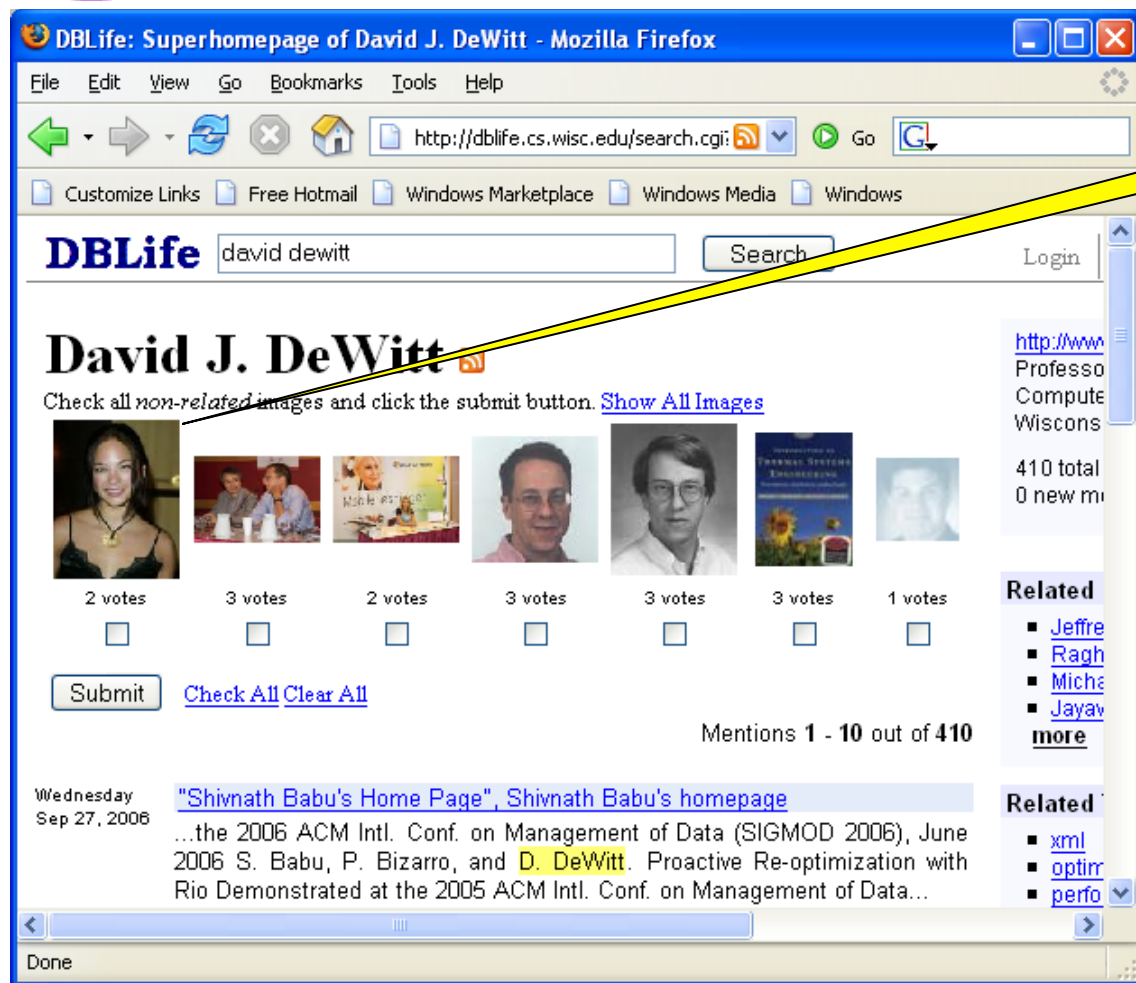
Internet

Raghu Ramakrishnan

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



Mass Collaboration for IE



Not David!



Picture is removed if enough users vote “no”.



Mass Collaboration Meets Spam

DBLife: Superhomepage of David J. DeWitt - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://dblife.cs.wisc.edu/search.cgi: Go

Customize Links Free Hotmail Windows Marketplace Windows Media Windows

DBLife david dewitt Search Login

David J. DeWitt

Check all non-related images and click the submit button. [Show All Images](#)



Jeffrey F. Naughton swears that this is David J. DeWitt

http://www...
Professo
Compute
Wiscons
410 total
0 new m

Related

- [Jeffre](#)
- [Ragh](#)
- [Miche](#)
- [Jayav](#)

[more](#)

mentions 1 - 10 out of 410

Wednesday
Sep 27, 2006

["Shivnath Babu's Home Page", Shivnath Babu's homepage](#)

...the 2006 ACM Intl. Conf. on Management of Data (SIGMOD 2006), June 2006 S. Babu, P. Bizarro, and D. DeWitt. Proactive Re-optimization with Rio Demonstrated at the 2005 ACM Intl. Conf. on Management of Data...

Related

- [xml](#)
- [optim](#)
- [perfo](#)

Done



Some Closely Related Work

- Avatar (IBM Almaden)
- DeepWeb & WebTables (Google)
- DBLife (U. Wisconsin)
- KnowItAll & TextRunner (U. Washington)
- Nell (CMU)
- Rexa (U. Mass)



Social Networking



Social Networks: Distro to Go!

HootSuite

hootsuite.com/dashboard#/tabs?id=8656771

Most Visited | Headlines | YMail | YLab | Maps | YCoMail | Backyard | LH2011 | NoMercy | NM2011 | Board | x4 | JS | PFT | NFP | Sch | Bookmarks

how to capture | SEARCH

Compose message...

Click to select a profile...

Search Twitter...

Welcome x raghurwi (Twitter) x

+ Add Stream

Home Feed (raghurwi)

adbrandt
8:14pm via Web
I am told @JermichaelF88 will receive \$10.2 million from the Packers over the next 13 months.
Retweeted by AdamScheffer

OPENForum
Feb 20, 8:05am via HootSuite
Amex Cardmembers/merchants can be first to try #ForSmallBiz ads from Twitter! First eligible 10K get \$100 in free ads.
ow.ly/98Nbn
Promoted by OPEN Forum

AdamScheffer
7:48pm via WhoSay
After striking a deal with TE Jermichael Finley, Packers now can use their franchise tag elsewhere if they so choose.
84 retweets

AdamScheffer
7:37pm via WhoSay
Packers reached agreement with TE Jermichael Finley on a 2-year, \$15 million deal.

@raghurwi/friends

dickc
12:23pm via Twitter for iPhone
Guy next to me reading wine magazine. Looking over his shoulder, it appears that the incredibly expensive wines are good (again) this year
16 retweets

daniel_abadi
10:30am via Web
Excellent summary of why flash will never completely replace disk:
theregister.co.uk/2012/02/21/nan...
15 retweets

elatable
Feb 21, 9:54pm via Web
I'll be interviewing spiritual teacher Eckhart Tolle tomorrow at Google, 3pm... details:
goo.gl/9vshH
8 retweets

awadallah
Feb 21, 8:46pm via bitly
Today's Dilbert is ingenious, it is about the Higgs Boson, aka the "The God Particle"
bit.ly/mzOvQa

@raghurwi/raghu-news

the_hindu
7:43pm via TweetDeck
Five persons suspected to be behind armed bank robberies that recently rocked Chennai killed in police encounter around 1 a.m. on Thursday.
4 retweets

BBCBreaking
7:30pm via BBC News
Man arrested on suspicion of assault following disturbance at the bar in London's Houses of Parliament. Details shortly
bbc.in/xrCRA3
39 retweets

nytimes
7:13pm via The New York Times
Mainland Chinese Flock to Hong Kong to Have Babies
nyti.ms/zH8mAT
48 retweets

WSJ
7:01pm via TweetDeck
We want to include your tweets in our live GOP debate coverage. Add the tag #wsj to your tweets. We'll feature here:

Sent Tweets (raghurwi)

raghurwi
Feb 09, 4:59pm via TweetDeck
Check out the new CORE visualization:
visualize.yahoo.com/core/

raghurwi
Feb 09, 1:48pm via Tweet Button
Paper: Feeding Frenzy: Selectively Materializing Users' Event Feeds
highscalability.com/blog/2012/1/17...

raghurwi
Dec 31, 2011, 11:18am via HootSuite
RT @lessig: Credit Reports - So under the FACT Act, you get a free credit report each year from each of the major credit.....

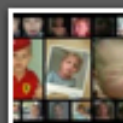
AlonHalevy
Dec 10, 2011, 11:36am via Web
My book, The Infinite Emotions of Coffee is out! Celebrating the latest and greatest in coffee world-wide!
macchiatone.com
Retweeted by raghurwi

brianfcooper
Aug 23, 2011, 10:54am via Web
#YCSB news: Michi Mutsuzaki, top-notch research engineer, will start maintaining



Connect with your friends around the shows you love

IntoNow from Yahoo! makes engaging with your friends around your favorite television shows easy and fun. Just tap the green button when you're watching, and IntoNow will identify the show, right down to the episode. Once identified it's easy to share with your friends on Twitter or Facebook.



Candace T is watching **SpongeBob SquarePants**

Season 1, Episode 21



John A is watching **The Walking Dead**

Season 2, Episode 9



David S is watching **Parenthood**

Season 3, Episode 17

Discover what your friends are into

Use IntoNow to discover new shows, discuss your favorites with friends, and learn more about whatever you're into! IntoNow is a great source of recommendations for what's on right now for you, based on what your friends are into. Find out what shows you share in common, get notified



Social Advertising

Recommend ads based on **private** shopping histories of “friends” in the social network.

- Nikon
- HP
- Nike

- Armani
- Gucci
- Prada



Alice



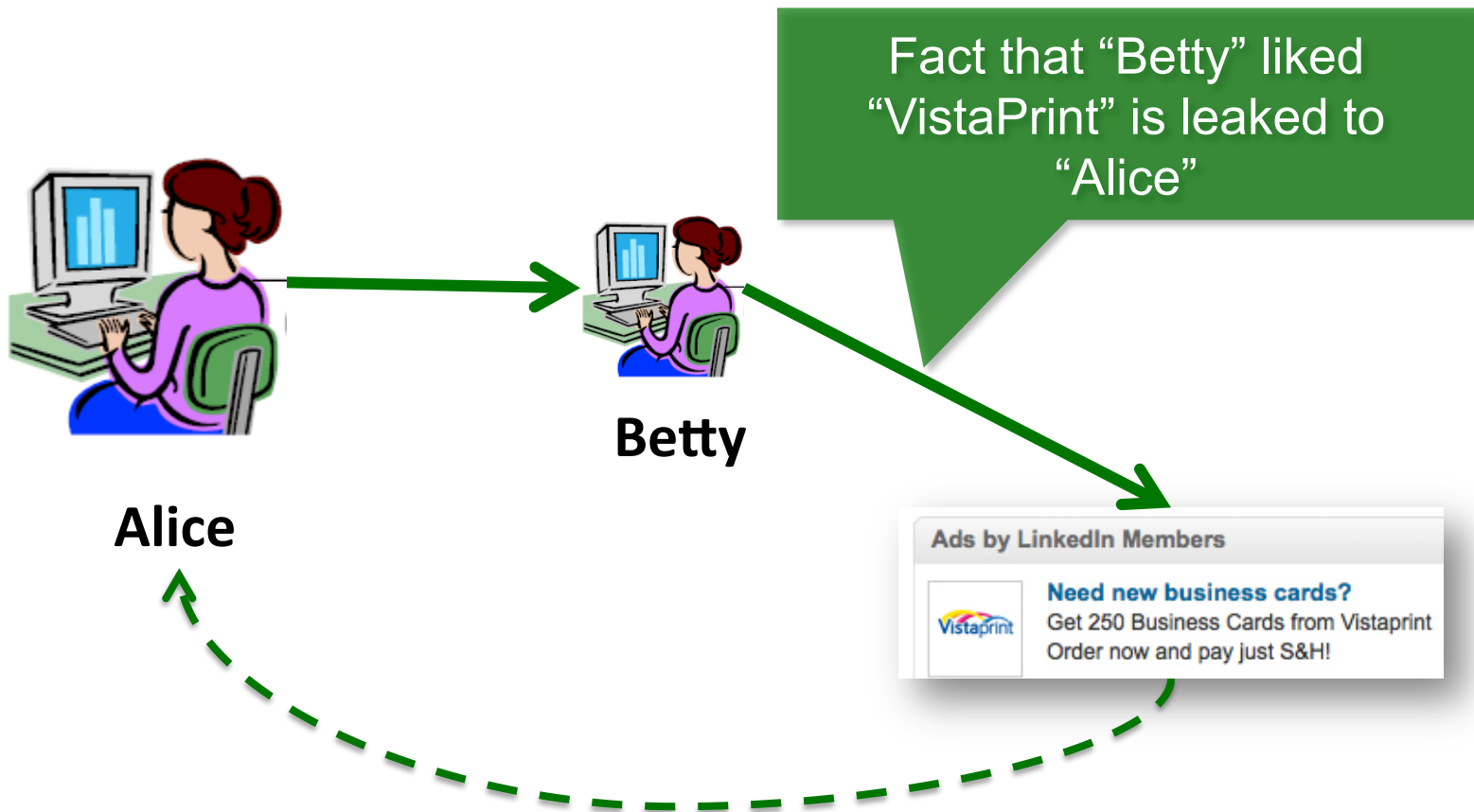
Betty





Privacy in Social Advertising

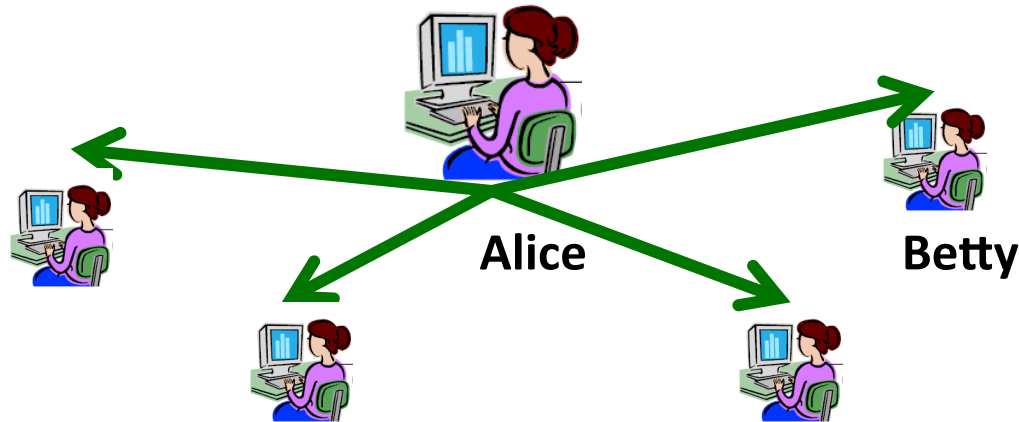
Items (products/people) liked by Alice's friends are better recommendations for Alice





Privacy in Social Advertising

Alice is recommended 'X'



Can we provide *accurate* recommendations to Alice
while ensuring that
Alice *cannot* deduce that Betty likes 'X' ?



Takeaway ...

- “For majority of the nodes in the network, recommendations must either be inaccurate or violate differential privacy!”
 - Maybe this is a “bad idea”
 - Or, Maybe **differential privacy is too strong a privacy definition to shoot for.**



Related Work

- Amazon, Netflix, Y! Music, etc.:
 - Collaborative filtering with large content pool (see KDD Cup, 2011)
 - Achieve lift by eliminating bad articles
 - We have a small number of high quality articles
- Search, Advertising
 - Matching problem with large content pool
 - Match through feature based models



Mobius: Unified Messaging and Data for Mobile Apps

Byung-Gon Chun, Carlo Curino, Russell Sears, Alexander Shraer,
Samuel Madden*, Raghu Ramakrishnan

Yahoo! Research, *MIT

mobicloud@yahoo-inc.com



Mobile Apps Connect to Cloud

Access, Create, Share Data/Messages



84% of apps are “cloud-connected”

(source Appcelerator)



Mobile App Developers' Challenges

1. Scalable Backend

Pull
News

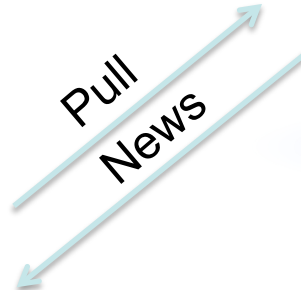




Mobile App Developers' Challenges

2. Speed: wireless networks

- Caching
- Prefetching





Mobile App Developers' Challenges

3. Notification



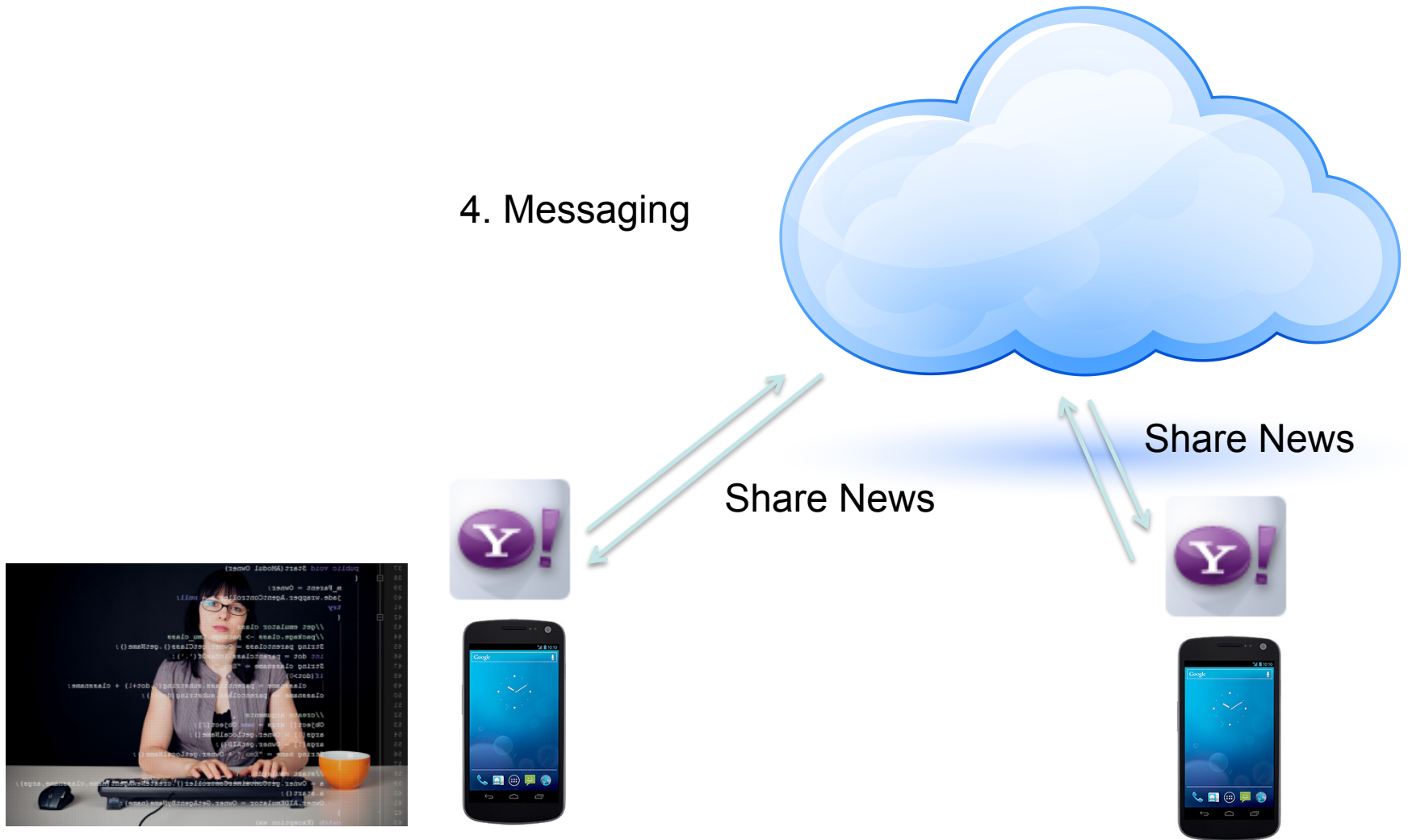
Push Alerts





Mobile App Developers' Challenges

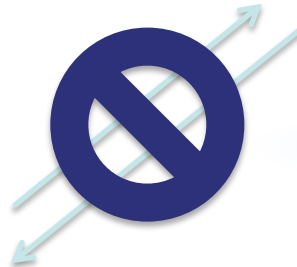
4. Messaging





Mobile App Developers' Challenges

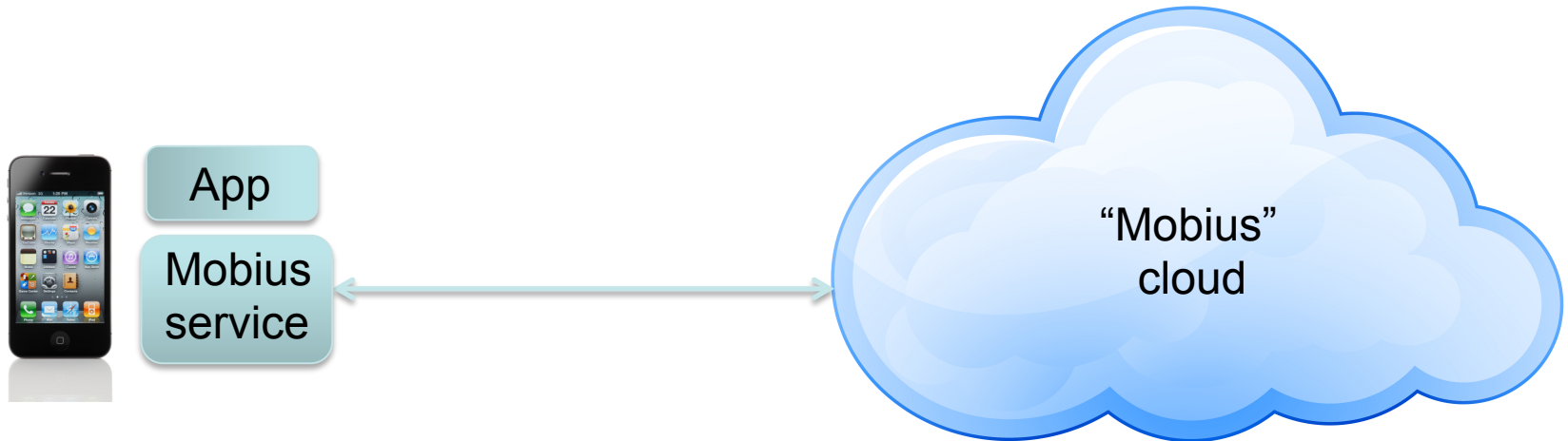
5. Disconnection



- Read news
- Post on Facebook



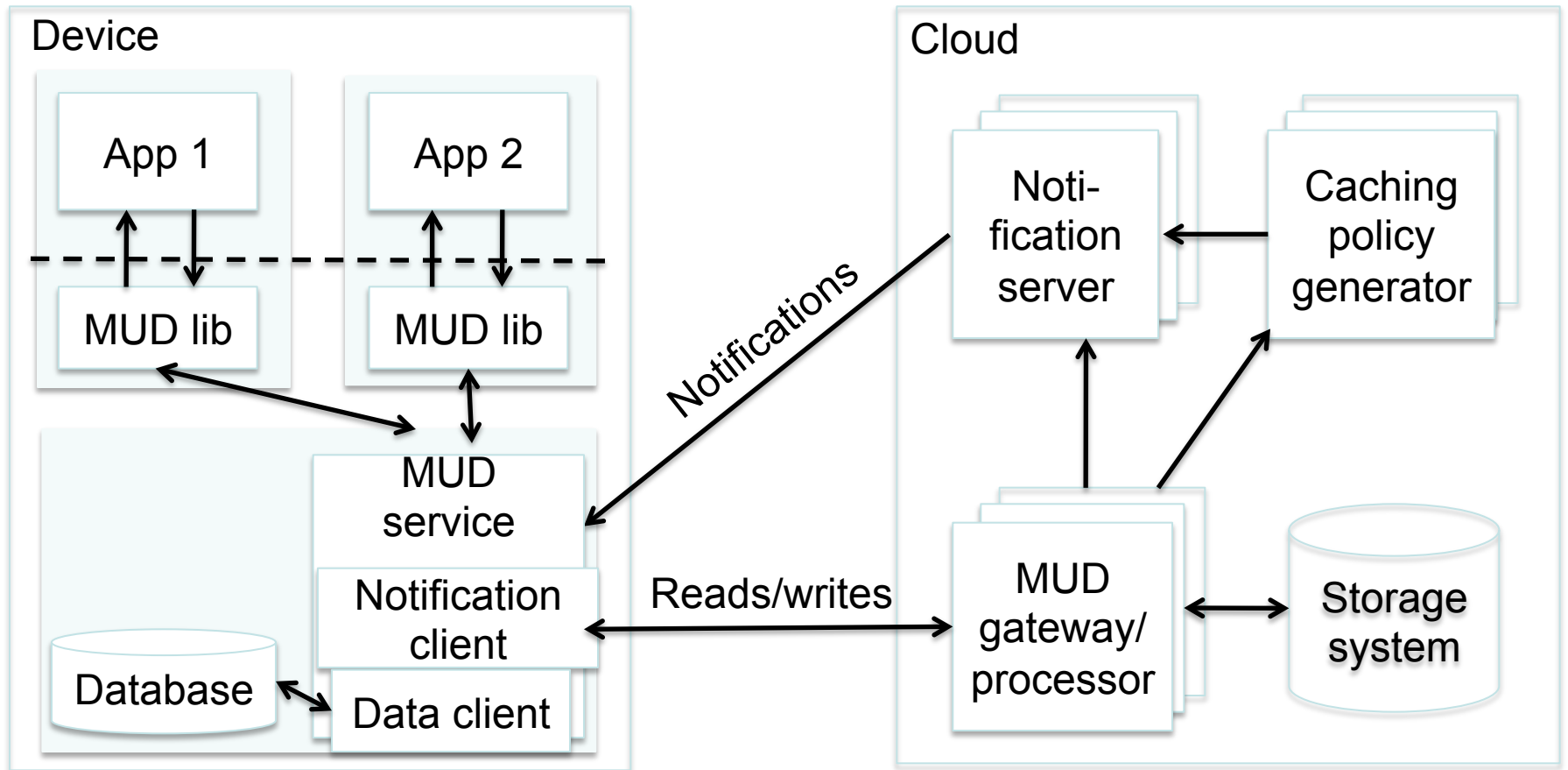
Mobius



- ☐ MUD API: (continuous) read/write of data
- ☐ Protocols to handle disconnection (and svn-like conflict resolution)
- ☐ Predicate-based notification (receiver and sender predicates)
- ☐ Automatic caching/prefetching (global policy generated by the cloud)



Mobius Architecture





Main Points

- The medium is being massaged
 - i.e., the canvas is no longer passive
 - It is also very personal, and `always on`
 - *(With apologies to Marshall McLuhan)*
- Search and browse converging
 - Semantics is the new frontier
 - Social networks the new distribution channel
- Learning from Big Data is a key
 - Value of data is now recognized
 - Data marketplaces will emerge
 - Compute platforms will be increasingly data-centric