Dr.-Ing. Gjergji Kasneci

gjergji.kasneci@hpi.uni-potsdam.de

HPI Potsdam, Winter Term 2012-13

INTRODUCTION TO INFORMATION RETRIEVAL

Organization

- Timetable
  - Lectures: Mondays 13:30 – 15:00, room H-E.51
  - Exercises: biweekly, Tuesdays 15:30 – 17:00, room H-E.52

- Exam
  - Condition for admission: oral presentation of at least two solutions during the tutorials
  - Form of exam: oral exam at the end of the term

What is this lecture about?

"Know where to find the information and how to use it – That's the secret of success."

Albert Einstein

- Subject

  Automated techniques for finding information relevant to an information need from a large corpus of data.

  - Content & structure analysis
  - Indexing, search, ranking
  - Classification, clustering
  - Annotation, extraction, summarization, visualization
  - Personalization & collaboration
  - Interaction with knowledge bases

Application areas of IR

- Web search
  - Web pages (or multimedia) about people, products, locations, events, news, problem solutions, …
  - Main problems: scalability, spam, format diversity, deep (hidden) web, …

- Enterprise search
  - Intranet documents/pages (or multimedia) about people/experts, products, divisions, problem solutions, …
  - Main problems: low information redundancy, few links, security issues, …

- Indexing and retrieval for **knowledge discovery & data mining** tasks
  - Bioinformatics: finding promising medical therapies, retrieving salient interaction paths in biochemical networks, …
  - Business intelligence: risk identification, composition of equity baskets, …
  - Scientific observations: identifying typical or noticeable patterns
  - Socio-political observations: analyzing social concerns, developments, political issues, …

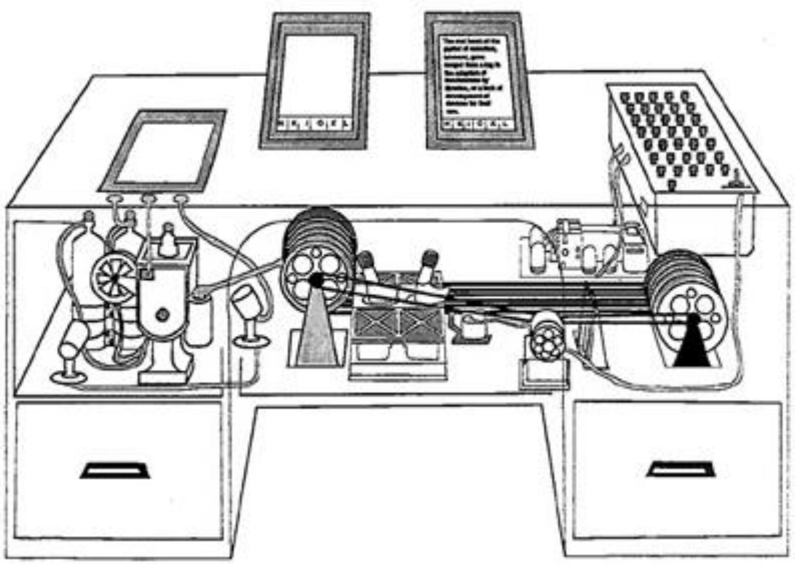
Differences to database systems

	Information retrieval systems	Databases
Corpus	Unstructured, semi-structured information (text, multimedia, etc.)	Structured, semi-structured data, e.g., database table(s), XML-document(s)
Query	Imprecise, keyword-based, typically with "and'ish" search semantics	Precise, structured query language, e.g., SQL, Xpath, Datalog, SPARQL …
Retrieval	Approximate, frequency-, similarity-based	Exact match of attribute values
Results	Ranked by relevance	Unranked, or ranked by specific attribute(s) (from the order-by clause of the query)

Historic background

- 1945: far-reaching ideas about personal & collective **information gathering, indexing, aligning, associating, sharing, reasoning**, …
  mentioned in "**As we may think**" by **Vannevar Bush**

Vannevar Bush



Memex extends human memory and reasoning capabilities

Roots in search over library catalogs

- Early work on IR, in the 50's and 60's
  - Boolean search (i.e., combine search terms with Boolean operators: **and, or, not**) unsatisfactory
  - Need for approximate search with "vaguely" Boolean operators needed (e.g., "**and'ish**" search semantics)
  - Need for ranked results

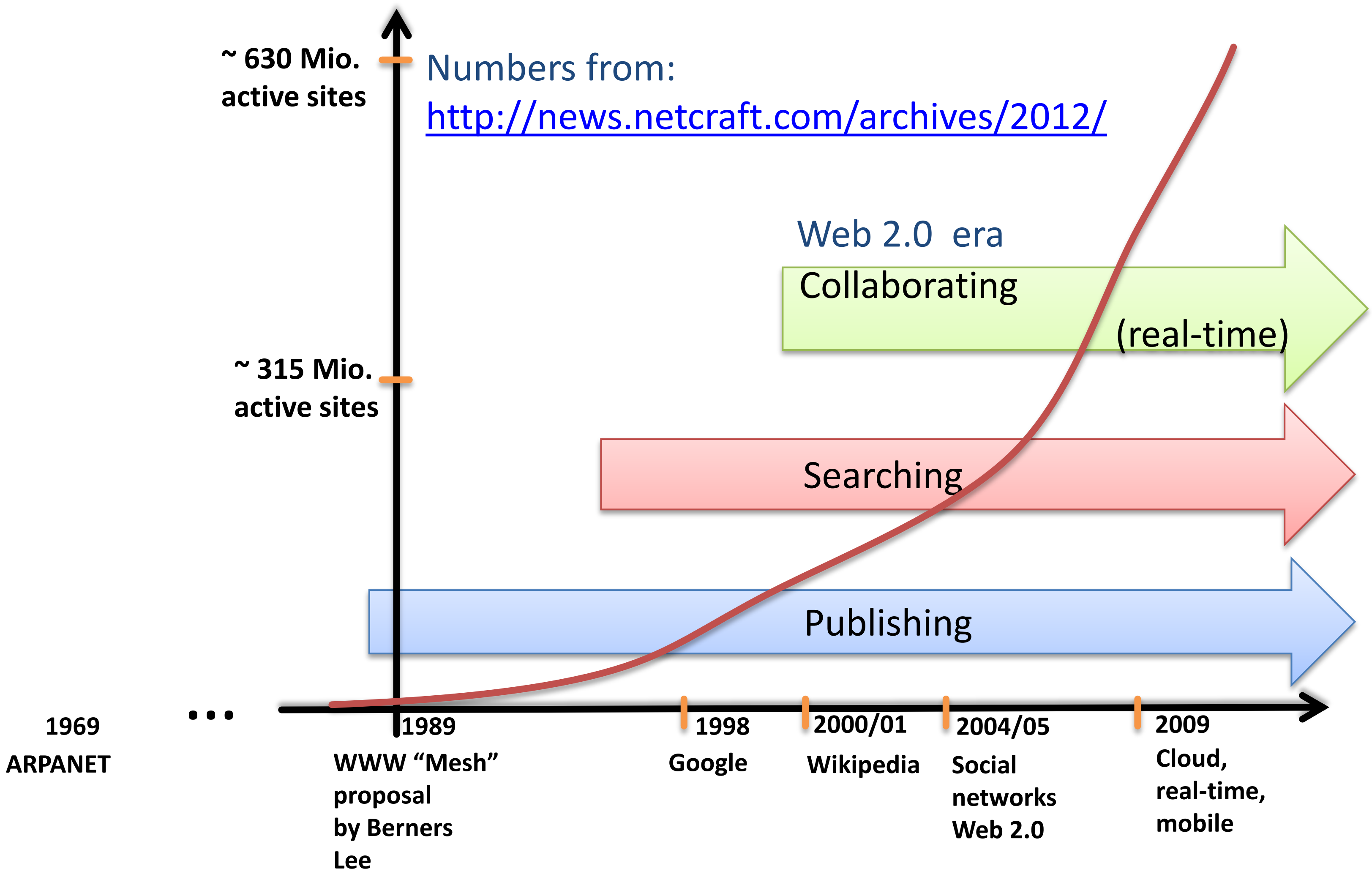
Early research

- **Hans Peter Luhn** (IBM researcher)
  - 1957: published “A Statistical Approach to Mechanized Encoding and Searching of Literary Information”

- **Gerard Salton** (professor at Cornell)
  - **1960’s**: supervised the development of **SMART** (System for the Mechanical Analysis and Retrieval of Text)
  - 1969: published “***Automatic Information Organization and Retrieval***”
  - Introduces important concepts, e.g., **vector space model**, **relevance feedback**, **Rocchio classififcation**, …

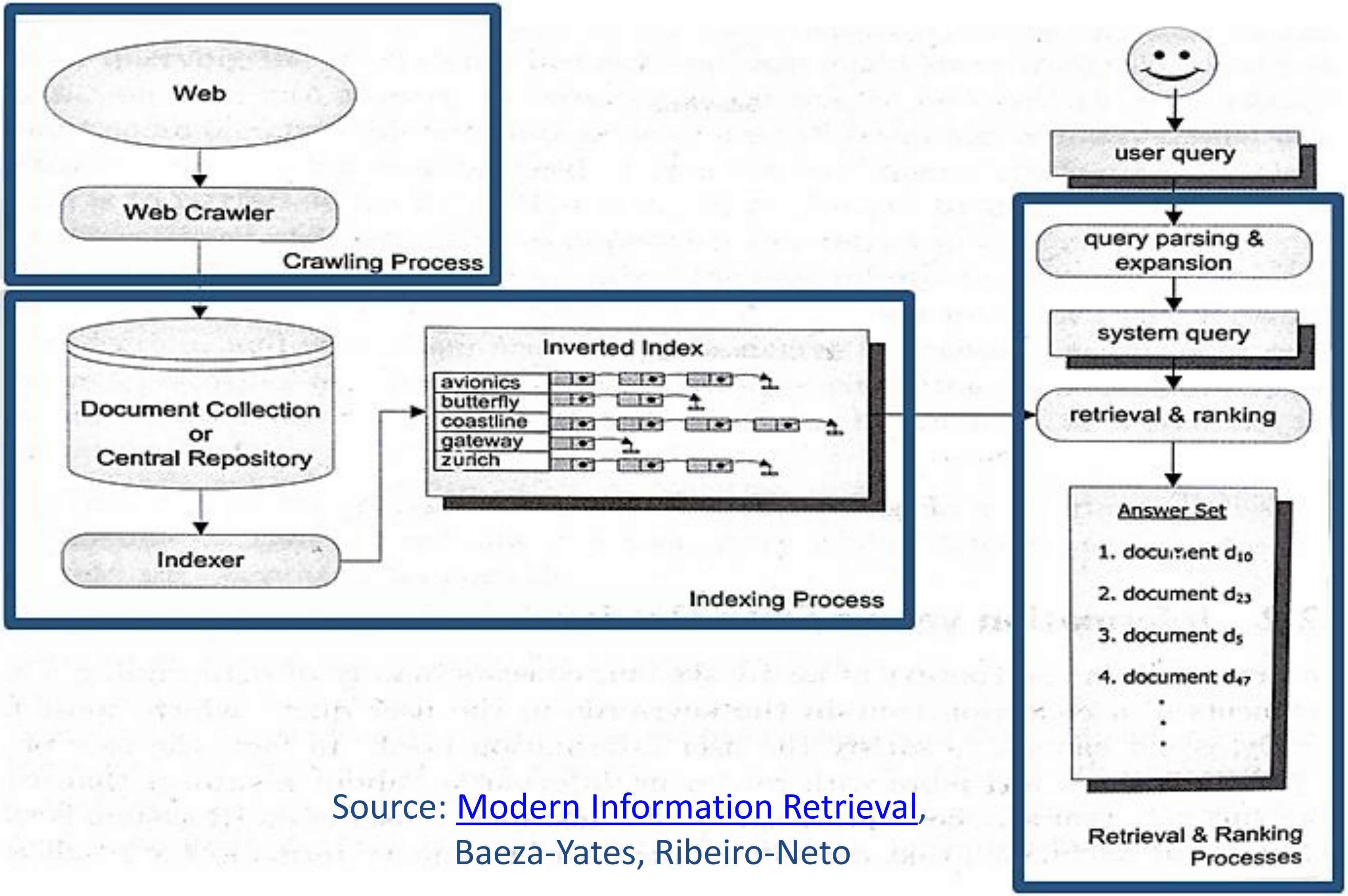
- Other influenatial scientists
  - **Karen Spärck Jones** introduced **inverse-document-frequency** in 1972
  - **Steven Robertson** introduced **Okapi BM25** ranking model together with Karen Spärck Jones in 1976

… and then the web was invented



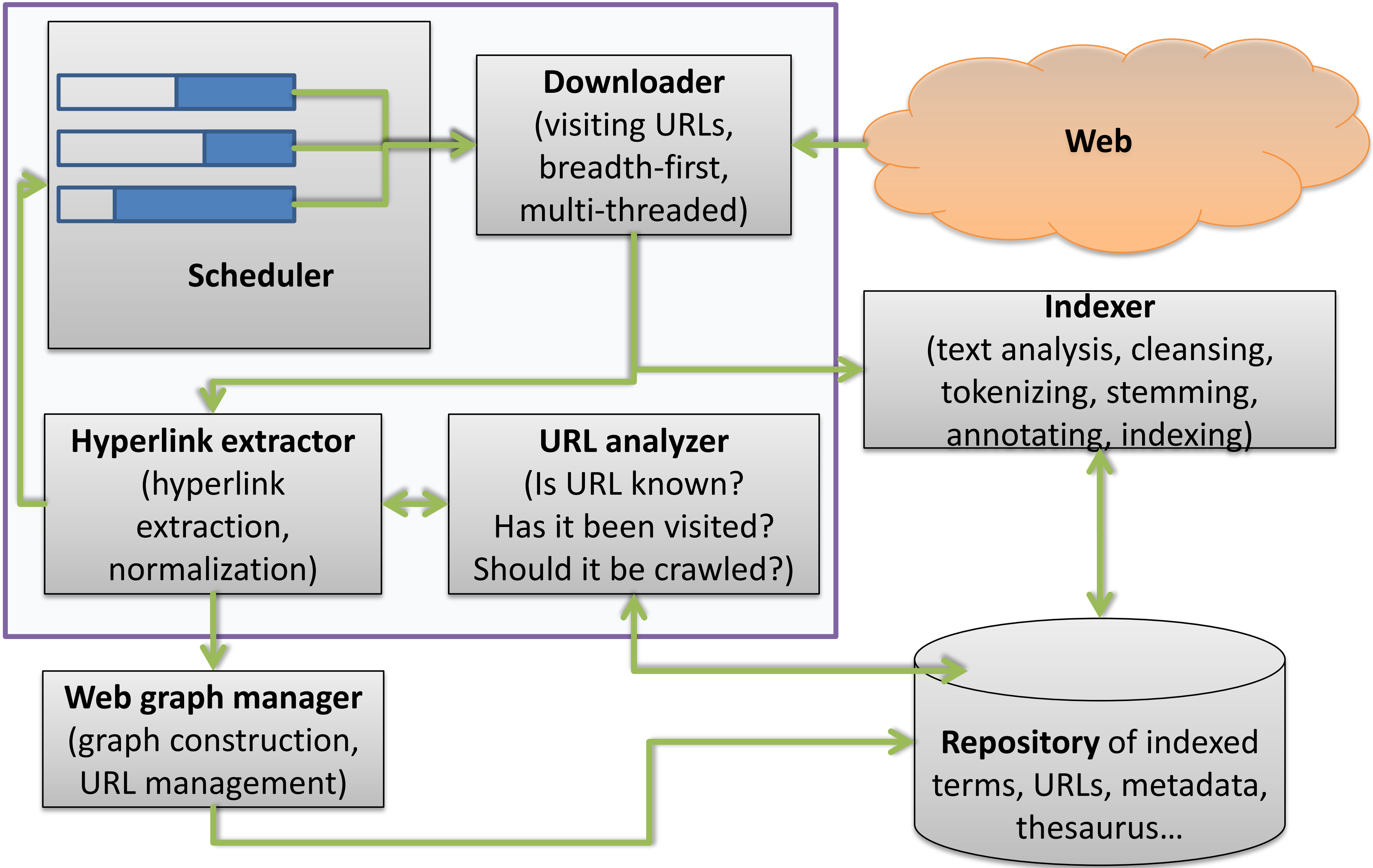
Numbers from: http://news.netcraft.com/archives/2012/

General infrastructure for web IR



Source: Modern Information Retrieval, Baeza-Yates, Ribeiro-Neto

Crawling



Downloader (visiting URLs, breadth-first, multi-threaded)

Web

Scheduler

Indexer (text analysis, cleansing, tokenizing, stemming, annotating, indexing)

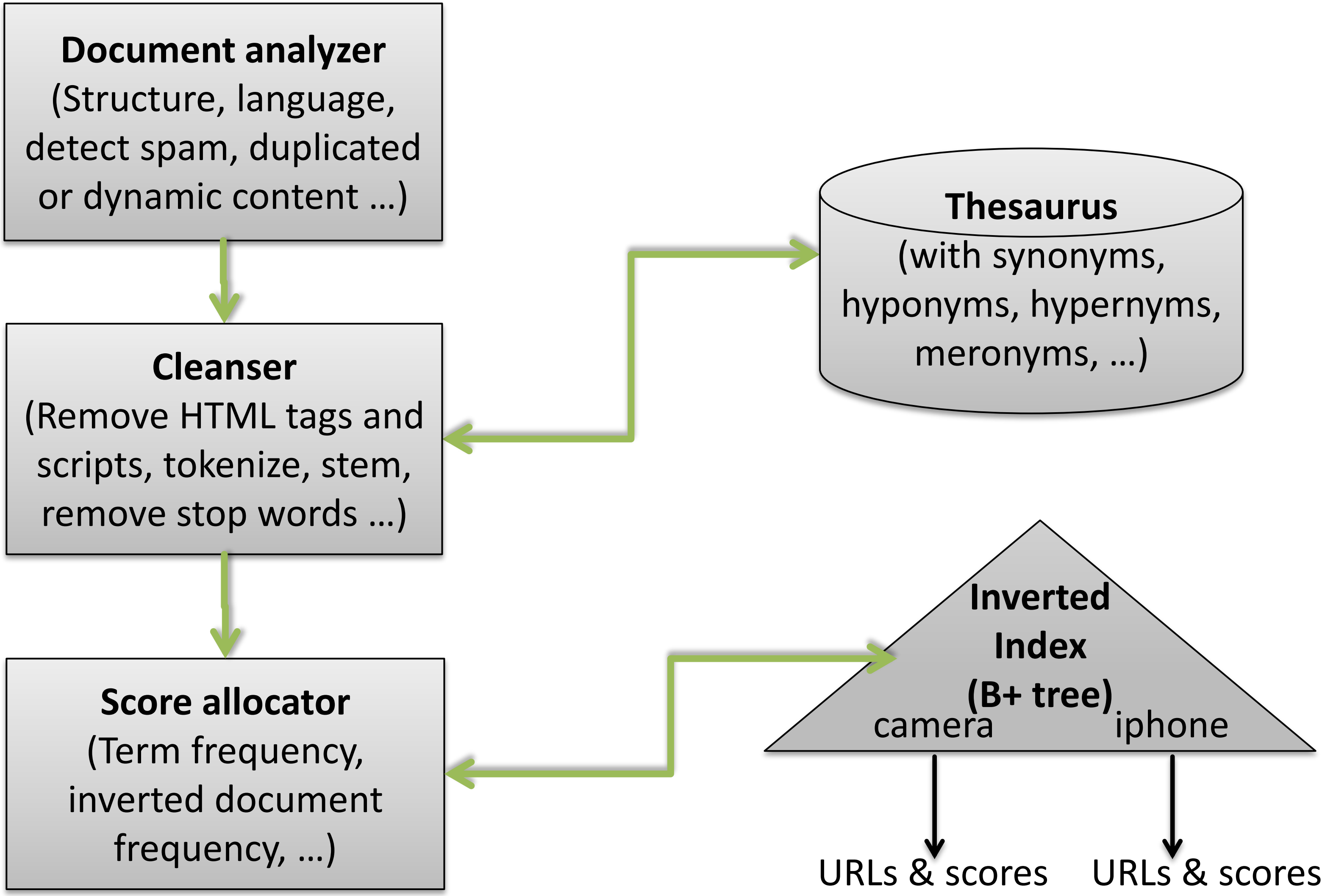
Hyperlink extractor (hyperlink extraction, normalization)

URL analyzer (Is URL known? Has it been visited? Should it be crawled?)

Web graph manager (graph construction, URL management)

Repository of indexed terms, URLs, metadata, thesaurus…

Indexing



Document analyzer
(Structure, language, detect spam, duplicated or dynamic content …)

Cleanser
(Remove HTML tags and scripts, tokenize, stem, remove stop words …)

Score allocator
(Term frequency, inverted document frequency, …)

Thesaurus
(with synonyms, hyponyms, hypernyms, meronyms, …)

Inverted Index (B+ tree)

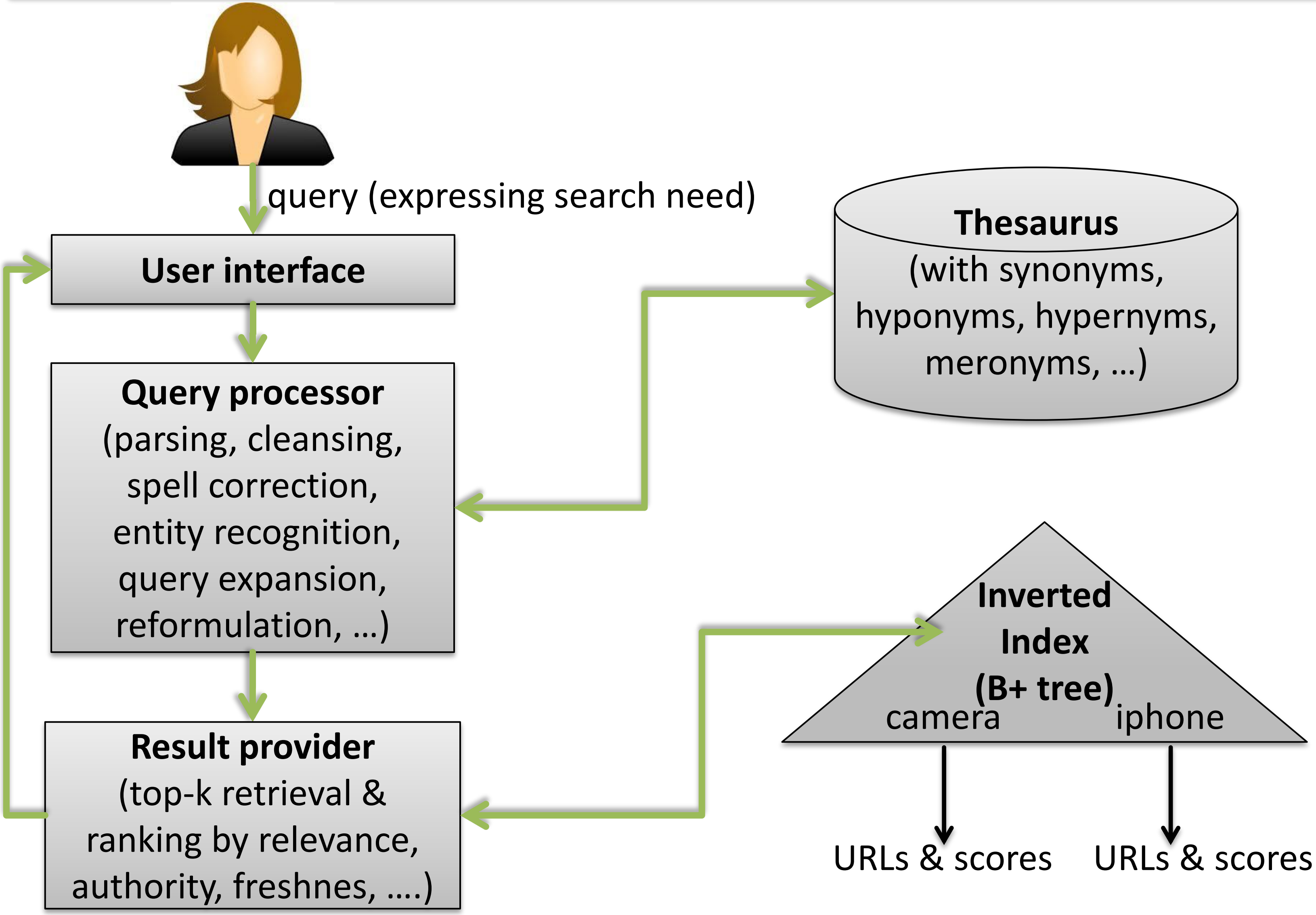
camera

iphone

URLs & scores

URLs & scores

Query answering



query (expressing search need)

User interface

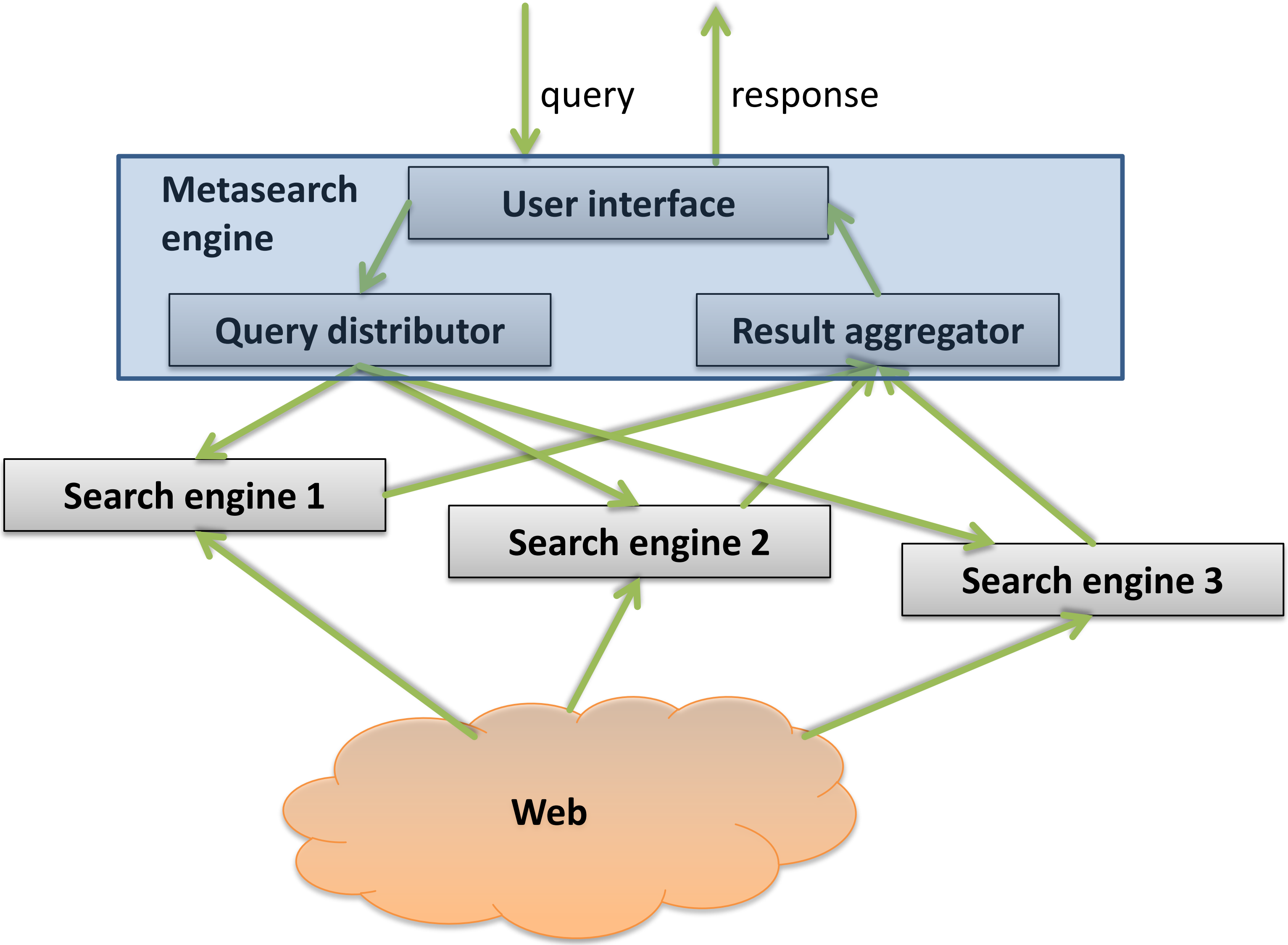
Query processor
(parsing, cleansing, spell correction, entity recognition, query expansion, reformulation, …)

Result provider
(top-k retrieval & ranking by relevance, authority, freshnes, ….)

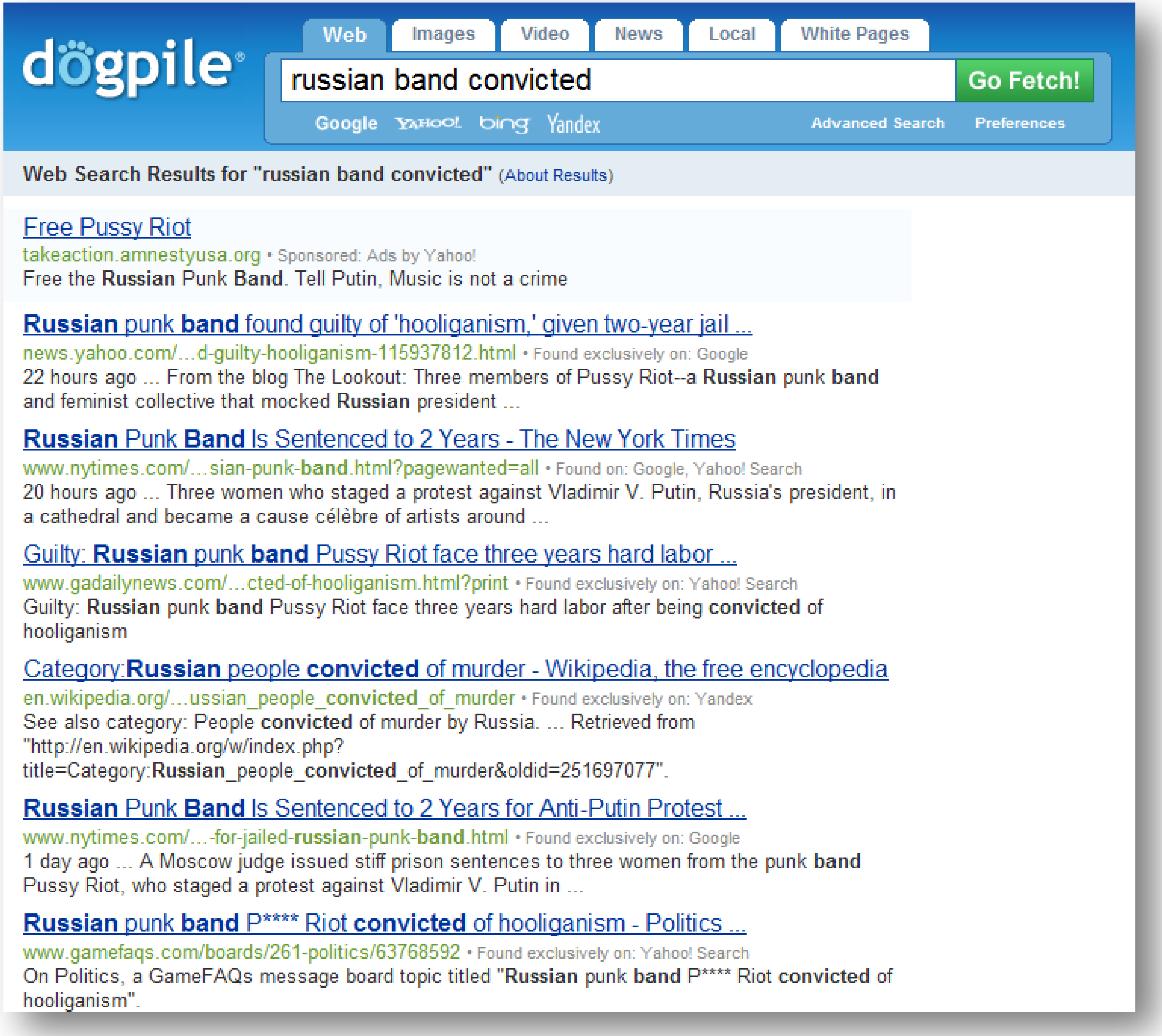
Thesaurus
(with synonyms, hyponyms, hypernyms, meronyms, …)

Inverted Index (B+ tree)

Merging results through metasearch



Metasearch example



Web Search Results for "russian band convicted" (About Results)

[Free Pussy Riot]
takeaction.amnestyusa.org • Sponsored: Ads by Yahoo!
Free the **Russian** Punk **Band**. Tell Putin, Music is not a crime

Russian punk **band** found guilty of 'hooliganism,' given two-year jail ...
news.yahoo.com/...d-guilty-hooliganism-115937812.html • Found exclusively on: Google
22 hours ago ... From the blog The Lookout: Three members of Pussy Riot--a **Russian** punk **band** and feminist collective that mocked **Russian** president ...

Russian Punk **Band** Is Sentenced to 2 Years - The New York Times
www.nytimes.com/...sian-punk-**band**.html?pagewanted=all • Found on: Google, Yahoo! Search
20 hours ago ... Three women who staged a protest against Vladimir V. Putin, Russia's president, in a cathedral and became a cause célèbre of artists around ...

Guilty: **Russian** punk **band** Pussy Riot face three years hard labor ...
www.gadailynews.com/...cted-of-hooliganism.html?print • Found exclusively on: Yahoo! Search
Guilty: **Russian** punk **band** Pussy Riot face three years hard labor after being **convicted** of hooliganism

Category:**Russian** people **convicted** of murder - Wikipedia, the free encyclopedia
en.wikipedia.org/...ussian_people_**convicted**_of_murder • Found exclusively on: Yandex
See also category: People **convicted** of murder by Russia. ... Retrieved from "http://en.wikipedia.org/w/index.php?title=Category:**Russian**_people_**convicted**_of_murder&oldid=251697077".

Russian Punk **Band** Is Sentenced to 2 Years for Anti-Putin Protest ...
www.nytimes.com/...-for-jailed-**russian**-punk-**band**.html • Found exclusively on: Google
1 day ago ... A Moscow judge issued stiff prison sentences to three women from the punk **band** Pussy Riot, who staged a protest against Vladimir V. Putin in ...

Russian punk **band** P**** Riot **convicted** of hooliganism - Politics ...
www.gamefaqs.com/boards/261-politics/63768592 • Found exclusively on: Yahoo! Search
On Politics, a GameFAQs message board topic titled "**Russian** punk **band** P**** Riot **convicted** of hooliganism".

Taxonomy of search on the web

- Typical semantics: **and'ish** search, e.g., "cheap flights Rome"

- Categorization of query types (by A. Broder, "A taxonomy of web search", 2002) through query log analysis and user study:

  - **~50% informational queries:** "General relativity", "Capital of Taiwan", "Obama campaign", …

  - **~20% navigational queries:** "facebook", "amazon", "ebay", …

  - **~30% transactional/resource-oriented queries:** "hotels Seattle", "flights Berlin Rome", "eminem mp3", "skype download", **…**

- Different numbers presented in study by Rose & Levinson in 2004 (~62% informational, ~13% navigational, ~25% transactional queries)

Advanced query types

- Regular expression queries (e.g., on Google):

  -inurl:(htm|html|php) intitle:"index of" +"last modified" +"parent directory" +description +size +(wma|mp3) "Nirvana"

- Conversion queries:

  10 Euro in USD

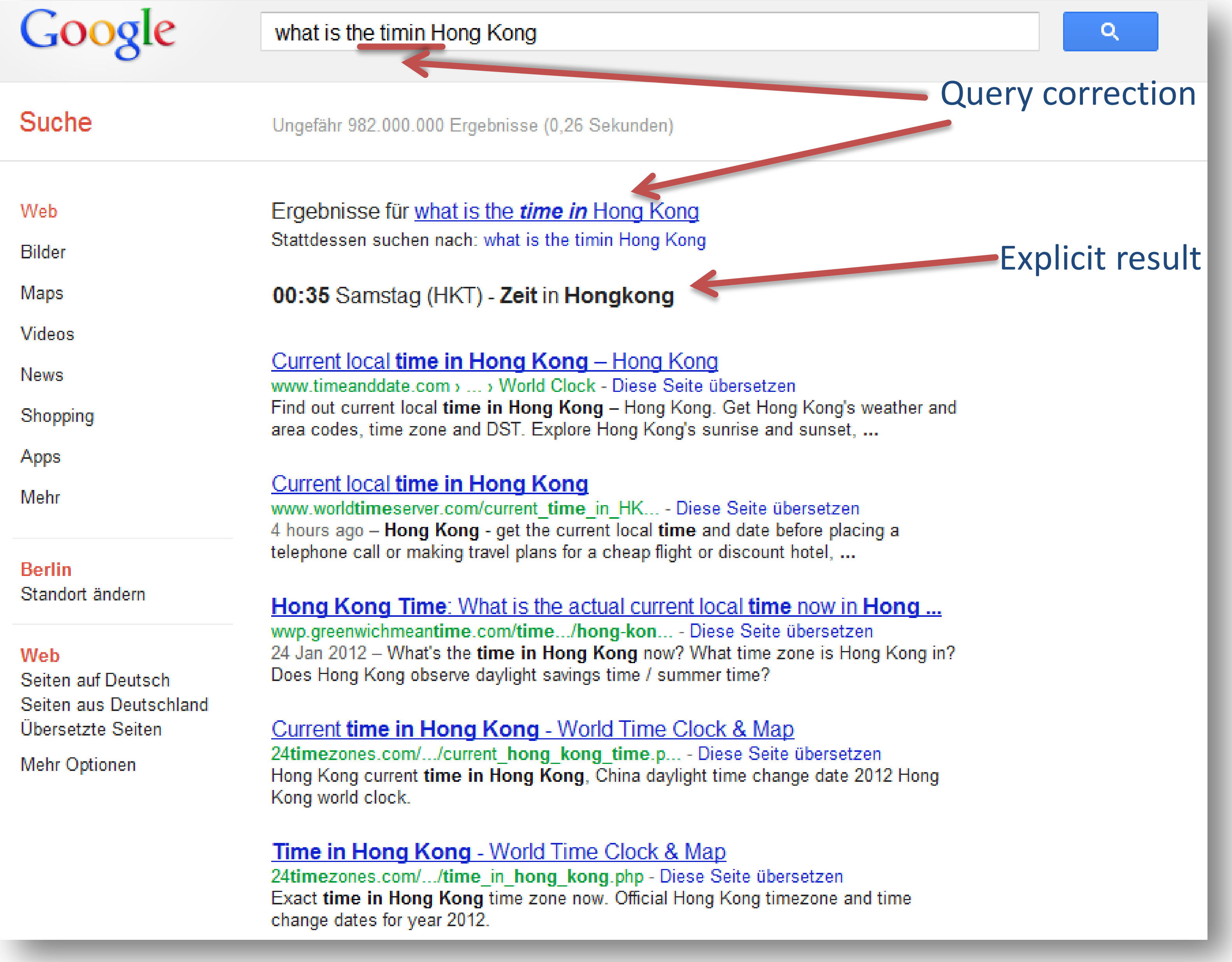
- Calculation queries:

  110*(654/8+3)

- Natural language questions:

  What is the time in Hong Kong

Search results



what is the timin Hong Kong

Query correction

Suche

Ungefähr 982.000.000 Ergebnisse (0,26 Sekunden)

Web
Bilder
Maps
Videos
News
Shopping
Apps
Mehr

Berlin
Standort ändern

Web
Seiten auf Deutsch
Seiten aus Deutschland
Übersetzte Seiten

Mehr Optionen

Ergebnisse für what is the ***time in*** Hong Kong
Stattdessen suchen nach: what is the timin Hong Kong

Explicit result

00:35 Samstag (HKT) - **Zeit** in **Hongkong**

Current local **time in Hong Kong** – Hong Kong
www.timeanddate.com › ... › World Clock - Diese Seite übersetzen
Find out current local **time in Hong Kong** – Hong Kong. Get Hong Kong's weather and area codes, time zone and DST. Explore Hong Kong's sunrise and sunset, **...**

Current local **time in Hong Kong**
www.world**time**server.com/current_**time**_in_HK... - Diese Seite übersetzen
4 hours ago – **Hong Kong** - get the current local **time** and date before placing a telephone call or making travel plans for a cheap flight or discount hotel, **...**

Hong Kong Time: What is the actual current local **time** now in **Hong ...**
wwp.greenwich**meantime**.com/**time**.../**hong-kon**... - Diese Seite übersetzen
24 Jan 2012 – What's the **time in Hong Kong** now? What time zone is Hong Kong in? Does Hong Kong observe daylight savings time / summer time?

Current **time in Hong Kong** - World Time Clock & Map
24**timezones**.com/.../current_**hong_kong_time**.p... - Diese Seite übersetzen
Hong Kong current **time in Hong Kong**, China daylight time change date 2012 Hong Kong world clock.

Time in Hong Kong - World Time Clock & Map
24**timezones**.com/.../**time_in_hong_kong**.php - Diese Seite übersetzen
Exact **time in Hong Kong** time zone now. Official Hong Kong timezone and time change dates for year 2012.

Ranking results

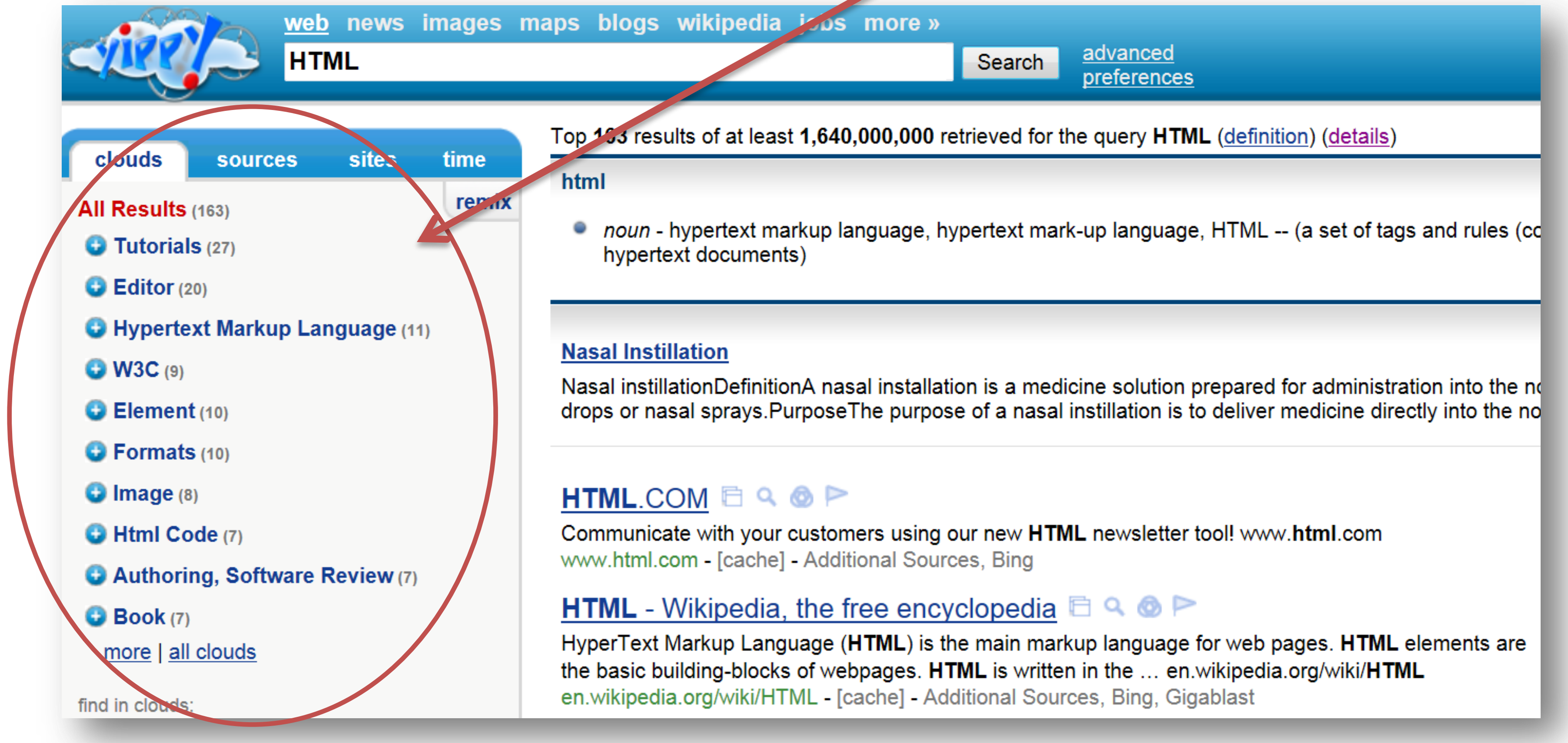
- By relevance
  - Binary (relevant or non-relevant).
  - Query/information-need dependent (i.e., similarity estimation between query terms and each document)
- By importance
  - Authority/popularity of result page (e.g., likelihood of being useful to a majority of users)
    - Links pointing to that page (PageRank, HITS)
    - Estimation of click-through-rate
    - …
  - Interestingness & context (e.g., likelihood of being useful for specific user in specific context → personalization)
  - Freshness (e.g., most recent news).

Modern search engines combine relevance and importance criteria for ranking!

Clustering results

Source: http://yippy.com/

Grouping topical categories



web news images maps blogs wikipedia jobs more »

HTML

Search

advanced
preferences

clouds sources sites time

remix

All Results (163)

- Tutorials (27)
- Editor (20)
- Hypertext Markup Language (11)
- W3C (9)
- Element (10)
- Formats (10)
- Image (8)
- Html Code (7)
- Authoring, Software Review (7)
- Book (7)

more | all clouds

find in clouds:

Top 163 results of at least 1,640,000,000 retrieved for the query **HTML** (definition) (details)

html

- *noun* - hypertext markup language, hypertext mark-up language, HTML -- (a set of tags and rules (co hypertext documents)

Nasal Instillation

Nasal installationDefinitionA nasal installation is a medicine solution prepared for administration into the no drops or nasal sprays.PurposeThe purpose of a nasal instillation is to deliver medicine directly into the no

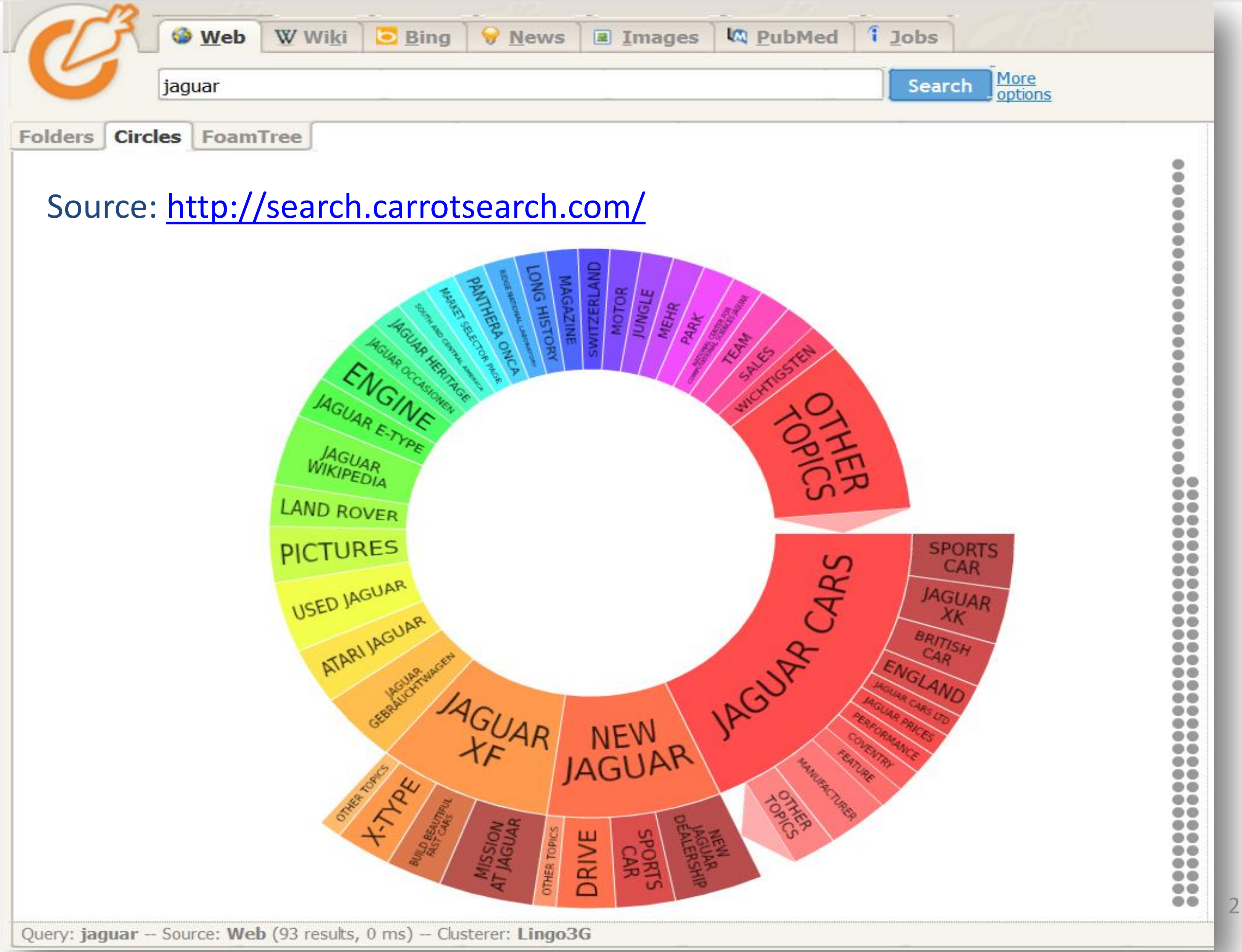
HTML.COM

Communicate with your customers using our new **HTML** newsletter tool! www.**html**.com
www.html.com - [cache] - Additional Sources, Bing

HTML - Wikipedia, the free encyclopedia

HyperText Markup Language (**HTML**) is the main markup language for web pages. **HTML** elements are the basic building-blocks of webpages. **HTML** is written in the … en.wikipedia.org/wiki/**HTML**
en.wikipedia.org/wiki/HTML - [cache] - Additional Sources, Bing, Gigablast

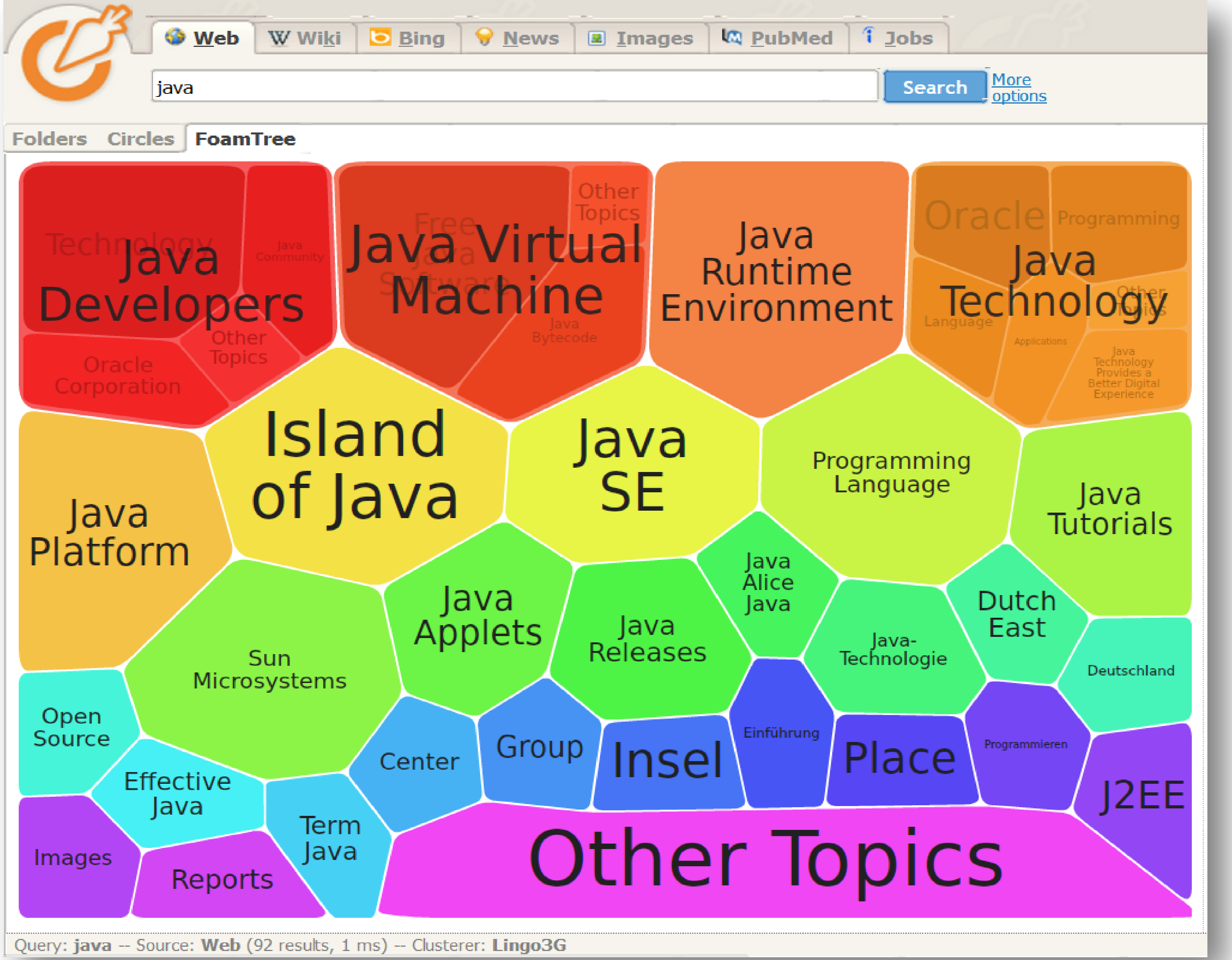
Visualizing result clusters



Source: http://search.carrotsearch.com/

Query: **jaguar** -- Source: **Web** (93 results, 0 ms) -- Clusterer: **Lingo3G**

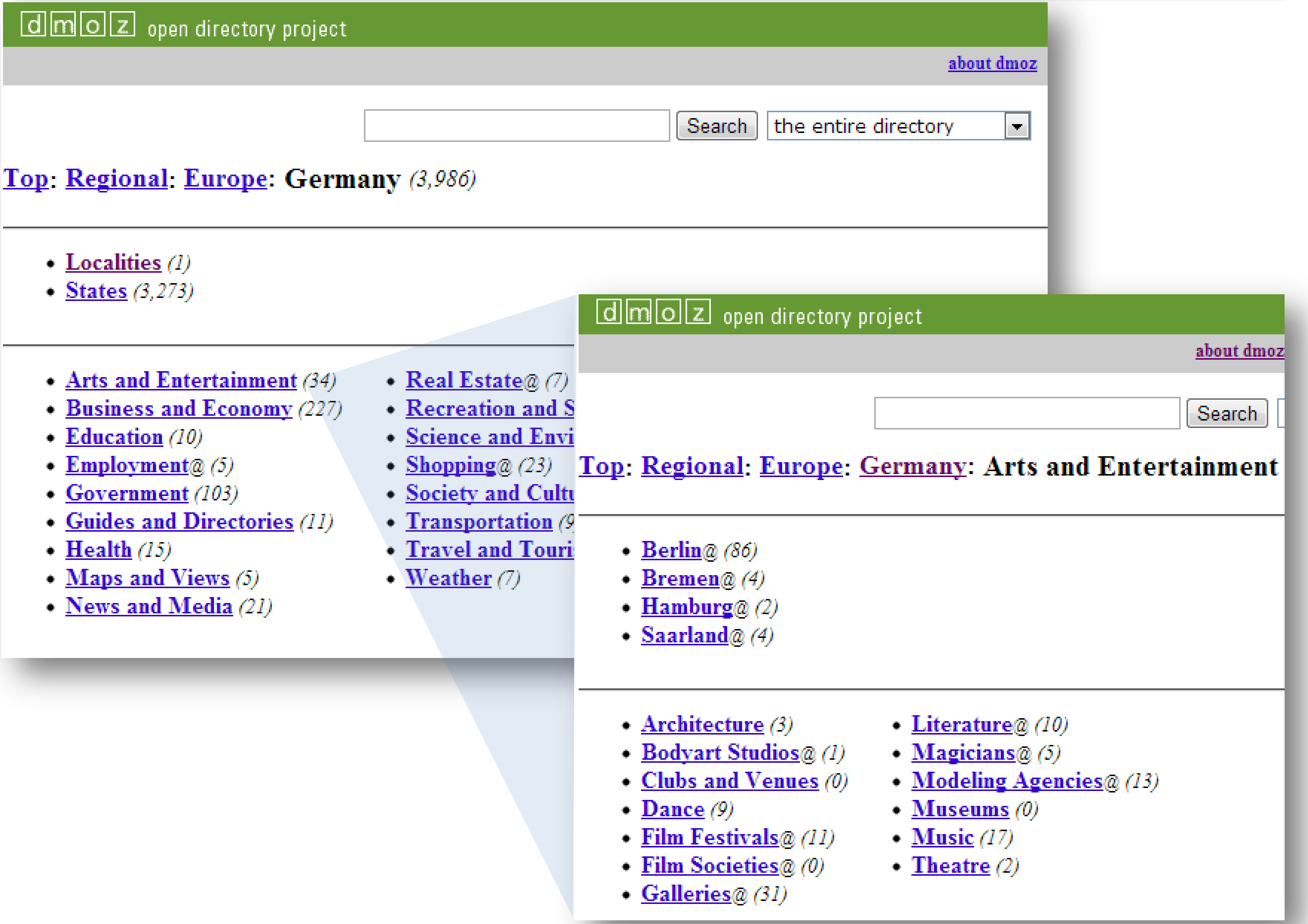
Visualizing result clusters



java

Query: **java** -- Source: **Web** (92 results, 1 ms) -- Clusterer: **Lingo3G**

Human edited categories in DMOZ



Top: Regional: Europe: Germany *(3,986)*

- Localities *(1)*
- States *(3,273)*

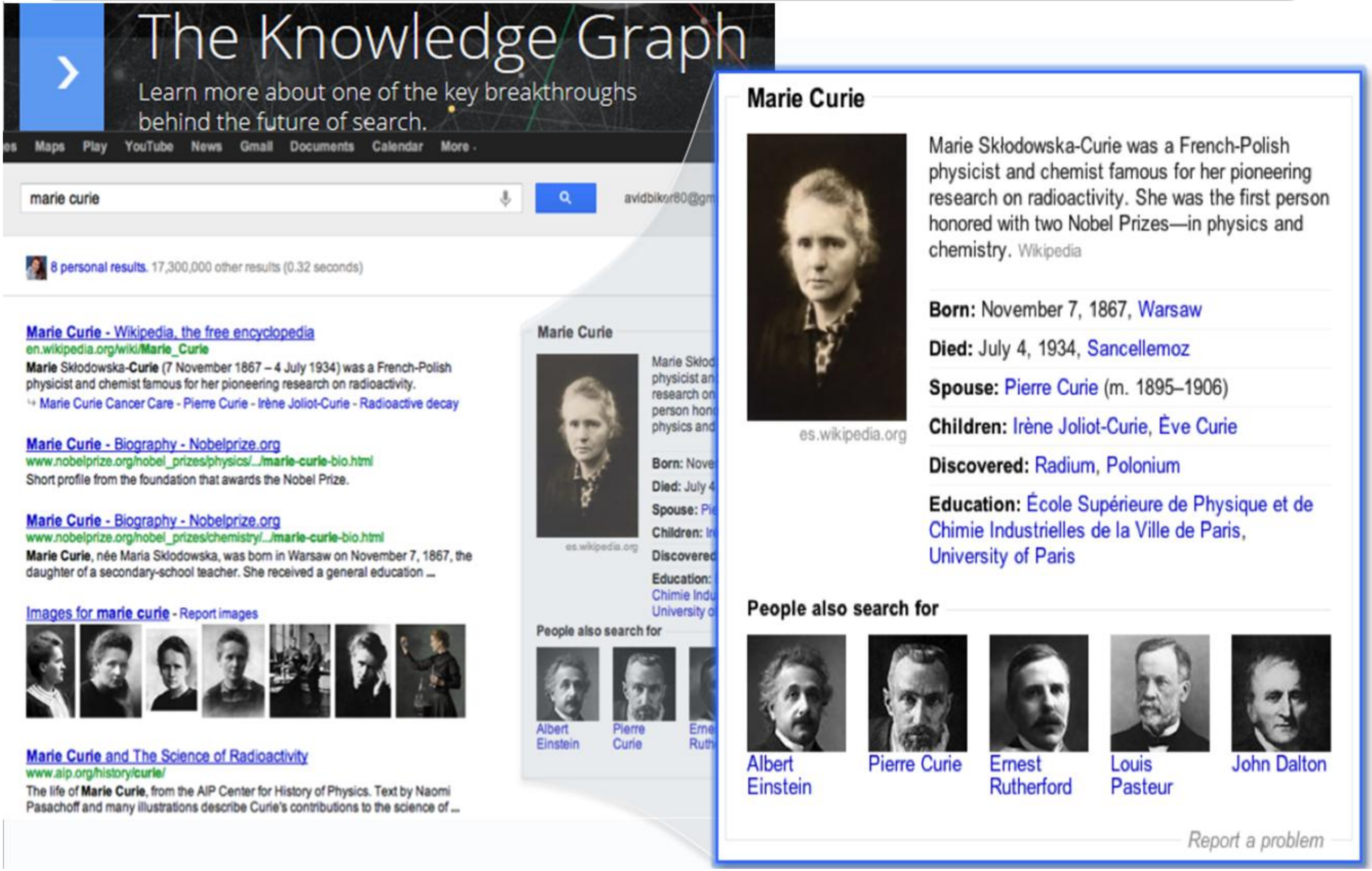
- Arts and Entertainment *(34)*
- Business and Economy *(227)*
- Education *(10)*
- Employment@ *(5)*
- Government *(103)*
- Guides and Directories *(11)*
- Health *(15)*
- Maps and Views *(5)*
- News and Media *(21)*
- Real Estate@ *(7)*
- Recreation and S
- Science and Envi
- Shopping@ *(23)*
- Society and Cultu
- Transportation *(9*
- Travel and Touri
- Weather *(7)*

Top: Regional: Europe: Germany: Arts and Entertainment

- Berlin@ *(86)*
- Bremen@ *(4)*
- Hamburg@ *(2)*
- Saarland@ *(4)*

- Architecture *(3)*
- Bodyart Studios@ *(1)*
- Clubs and Venues *(0)*
- Dance *(9)*
- Film Festivals@ *(11)*
- Film Societies@ *(0)*
- Galleries@ *(31)*
- Literature@ *(10)*
- Magicians@ *(5)*
- Modeling Agencies@ *(13)*
- Museums *(0)*
- Music *(17)*
- Theatre *(2)*

Using knowledge bases: Google’s knowledge graph



The Knowledge Graph

Learn more about one of the key breakthroughs behind the future of search.

8 personal results. 17,300,000 other results (0.32 seconds)

[Marie Curie - Wikipedia, the free encyclopedia](en.wikipedia.org/wiki/Marie_Curie)
en.wikipedia.org/wiki/Marie_Curie
Marie Skłodowska-**Curie** (7 November 1867 – 4 July 1934) was a French-Polish physicist and chemist famous for her pioneering research on radioactivity.
Marie Curie Cancer Care - Pierre Curie - Irène Joliot-Curie - Radioactive decay

Marie Curie - Biography - Nobelprize.org
www.nobelprize.org/nobel_prizes/physics/.../marie-curie-bio.html
Short profile from the foundation that awards the Nobel Prize.

Marie Curie - Biography - Nobelprize.org
www.nobelprize.org/nobel_prizes/chemistry/.../marie-curie-bio.html
Marie Curie, née Maria Sklodowska, was born in Warsaw on November 7, 1867, the daughter of a secondary-school teacher. She received a general education ...

Images for marie curie - Report images

Marie Curie and The Science of Radioactivity
www.aip.org/history/curie/
The life of **Marie Curie**, from the AIP Center for History of Physics. Text by Naomi Pasachoff and many illustrations describe Curie's contributions to the science of ...

Marie Curie
es.wikipedia.org
Marie Skłod
physicist an
research on
person hon
physics and
Born: Nove
Died: July 4
Spouse: Pie
Children: Ir
Discovered
Education:
Chimie Indu
University o
People also search for
Albert Einstein
Pierre Curie
Erne Ruth

Marie Curie

Marie Skłodowska-Curie was a French-Polish physicist and chemist famous for her pioneering research on radioactivity. She was the first person honored with two Nobel Prizes—in physics and chemistry. Wikipedia

es.wikipedia.org

Born: November 7, 1867, Warsaw

Died: July 4, 1934, Sancellemoz

Spouse: Pierre Curie (m. 1895–1906)

Children: Irène Joliot-Curie, Ève Curie

Discovered: Radium, Polonium

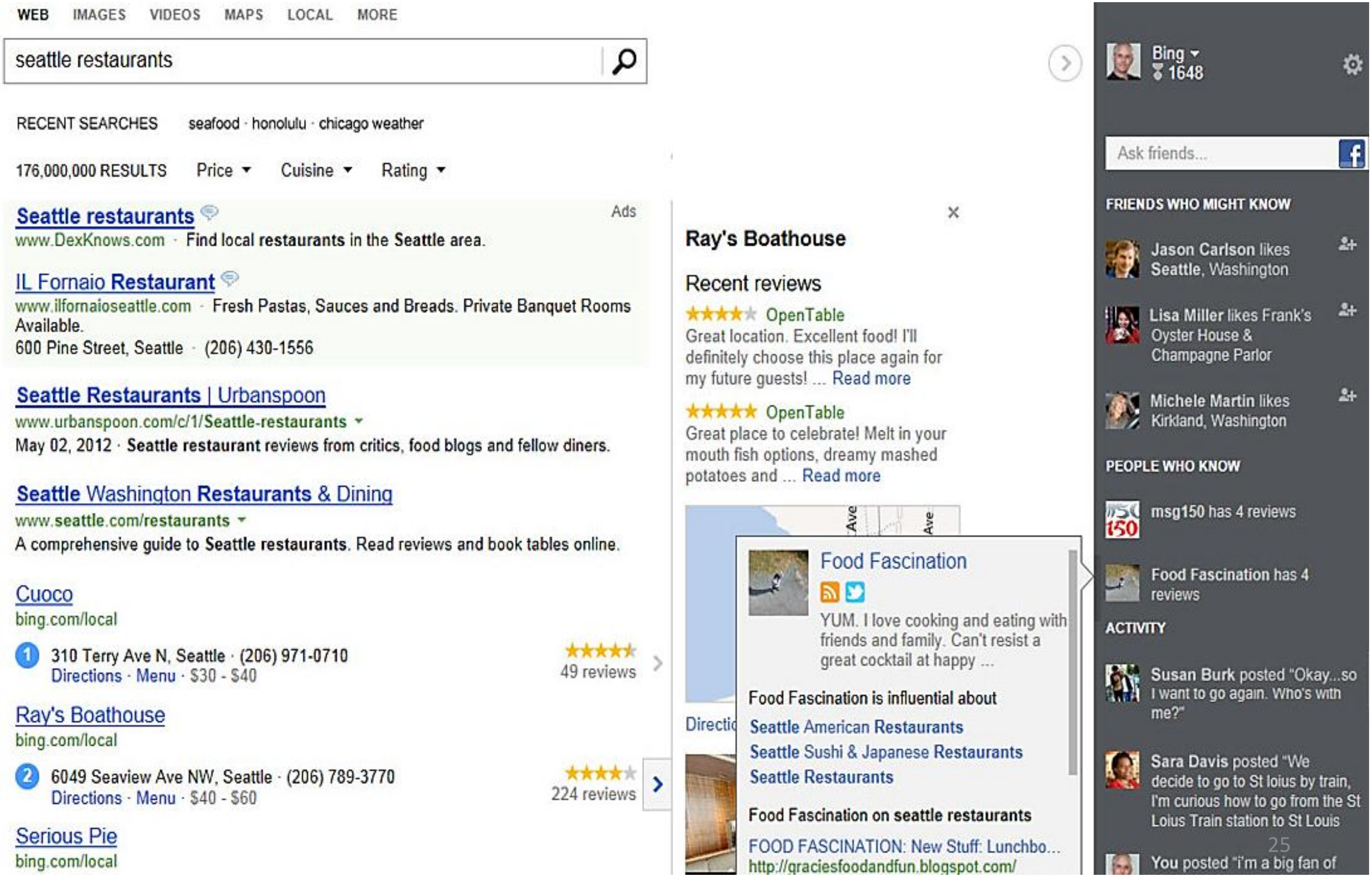
Education: École Supérieure de Physique et de Chimie Industrielles de la Ville de Paris, University of Paris

People also search for

Albert Einstein
Pierre Curie
Ernest Rutherford
Louis Pasteur
John Dalton

Report a problem

Using social networks: Bing's social search



WEB IMAGES VIDEOS MAPS LOCAL MORE

seattle restaurants

RECENT SEARCHES seafood · honolulu · chicago weather

176,000,000 RESULTS Price ▾ Cuisine ▾ Rating ▾

Ads

Seattle restaurants
www.DexKnows.com · Find local restaurants in the **Seattle** area.

IL Fornaio **Restaurant**
www.ilfornaioseattle.com · Fresh Pastas, Sauces and Breads. Private Banquet Rooms Available.
600 Pine Street, Seattle · (206) 430-1556

Seattle Restaurants | Urbanspoon
www.urbanspoon.com/c/1/**Seattle-restaurants** ▾
May 02, 2012 · **Seattle restaurant** reviews from critics, food blogs and fellow diners.

Seattle Washington **Restaurants** & Dining
www.**seattle**.com/**restaurants** ▾
A comprehensive guide to **Seattle restaurants**. Read reviews and book tables online.

Cuoco
bing.com/local
1 310 Terry Ave N, Seattle · (206) 971-0710
Directions · Menu · $30 - $40
★★★★★ 49 reviews

Ray's Boathouse
bing.com/local
2 6049 Seaview Ave NW, Seattle · (206) 789-3770
Directions · Menu · $40 - $60
★★★★★ 224 reviews

Serious Pie
bing.com/local

Ray's Boathouse

Recent reviews

★★★★☆ OpenTable
Great location. Excellent food! I'll definitely choose this place again for my future guests! ... Read more

★★★★★ OpenTable
Great place to celebrate! Melt in your mouth fish options, dreamy mashed potatoes and ... Read more

Directi

Food Fascination

YUM. I love cooking and eating with friends and family. Can't resist a great cocktail at happy ...

Food Fascination is influential about
Seattle American Restaurants
Seattle Sushi & Japanese Restaurants
Seattle Restaurants

Food Fascination on seattle restaurants
FOOD FASCINATION: New Stuff: Lunchbo...
http://graciesfoodandfun.blogspot.com/

Bing ▾
1648

Ask friends...

FRIENDS WHO MIGHT KNOW

Jason Carlson likes Seattle, Washington

Lisa Miller likes Frank's Oyster House & Champagne Parlor

Michele Martin likes Kirkland, Washington

PEOPLE WHO KNOW

msg150 has 4 reviews

Food Fascination has 4 reviews

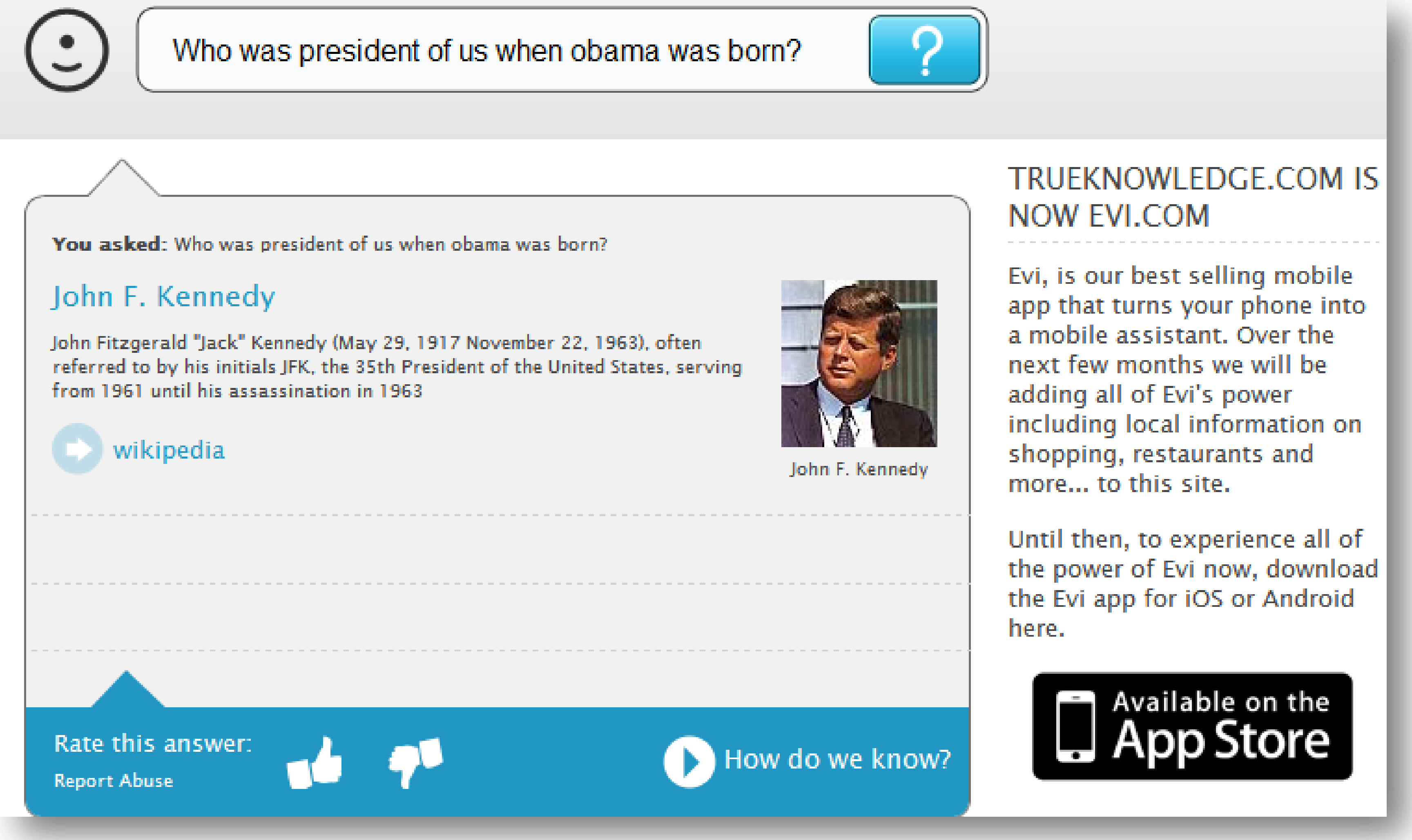
ACTIVITY

Susan Burk posted "Okay...so I want to go again. Who's with me?"

Sara Davis posted "We decide to go to St loius by train, I'm curious how to go from the St Loius Train station to St Louis

You posted "i'm a big fan of

Semantic search: question answering



Who was president of us when obama was born?

You asked: Who was president of us when obama was born?

John F. Kennedy

John Fitzgerald "Jack" Kennedy (May 29, 1917 November 22, 1963), often referred to by his initials JFK, the 35th President of the United States, serving from 1961 until his assassination in 1963

wikipedia

John F. Kennedy

Rate this answer:

Report Abuse

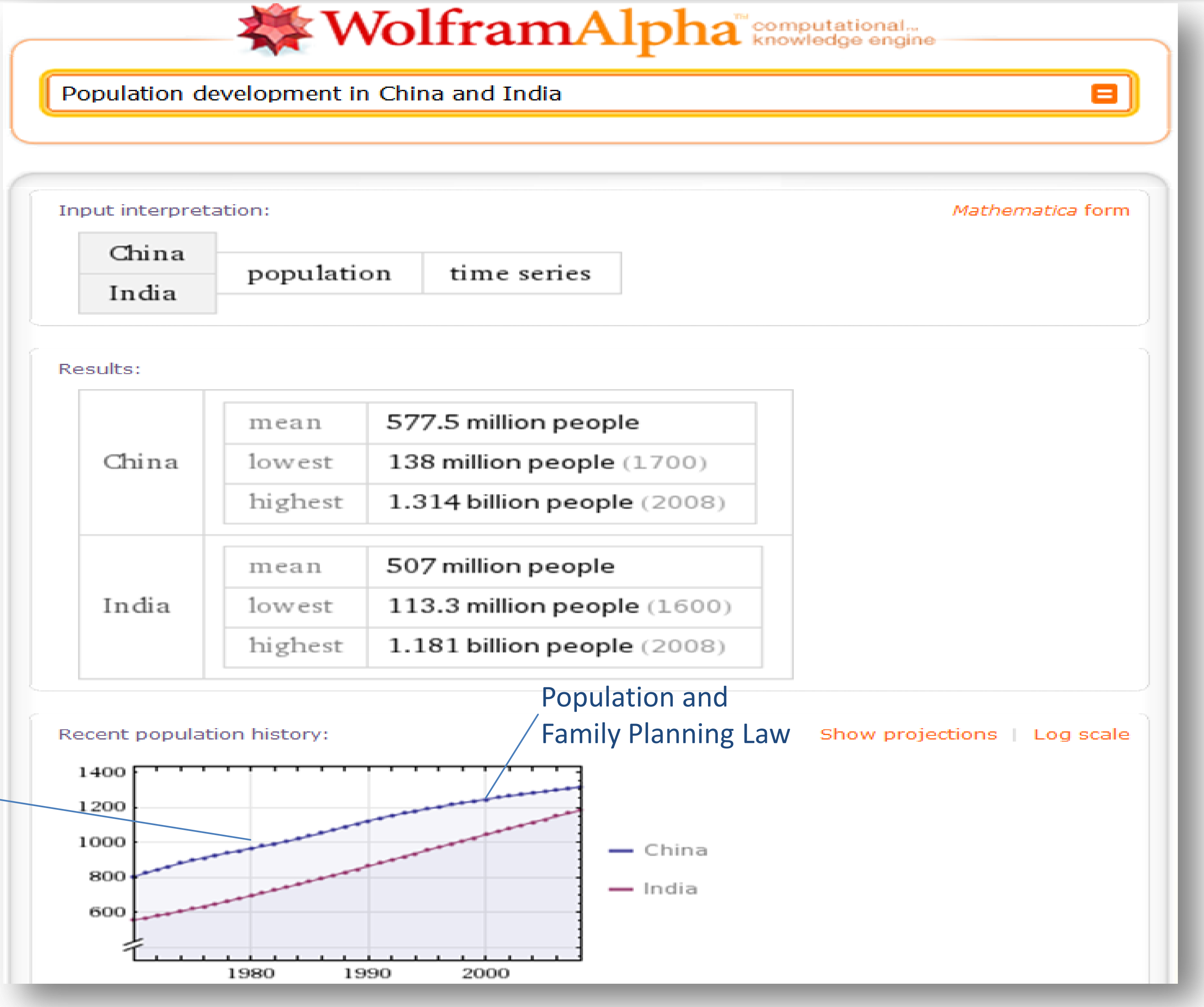
How do we know?

TRUEKNOWLEDGE.COM IS NOW EVI.COM

Evi, is our best selling mobile app that turns your phone into a mobile assistant. Over the next few months we will be adding all of Evi's power including local information on shopping, restaurants and more... to this site.

Until then, to experience all of the power of Evi now, download the Evi app for iOS or Android here.

Computational question answering



WolframAlpha computational... knowledge engine

Population development in China and India

Input interpretation:

Mathematica form

China India	population	time series

Results:

China	mean	577.5 million people
	lowest	138 million people (1700)
	highest	1.314 billion people (2008)
India	mean	507 million people
	lowest	113.3 million people (1600)
	highest	1.181 billion people (2008)

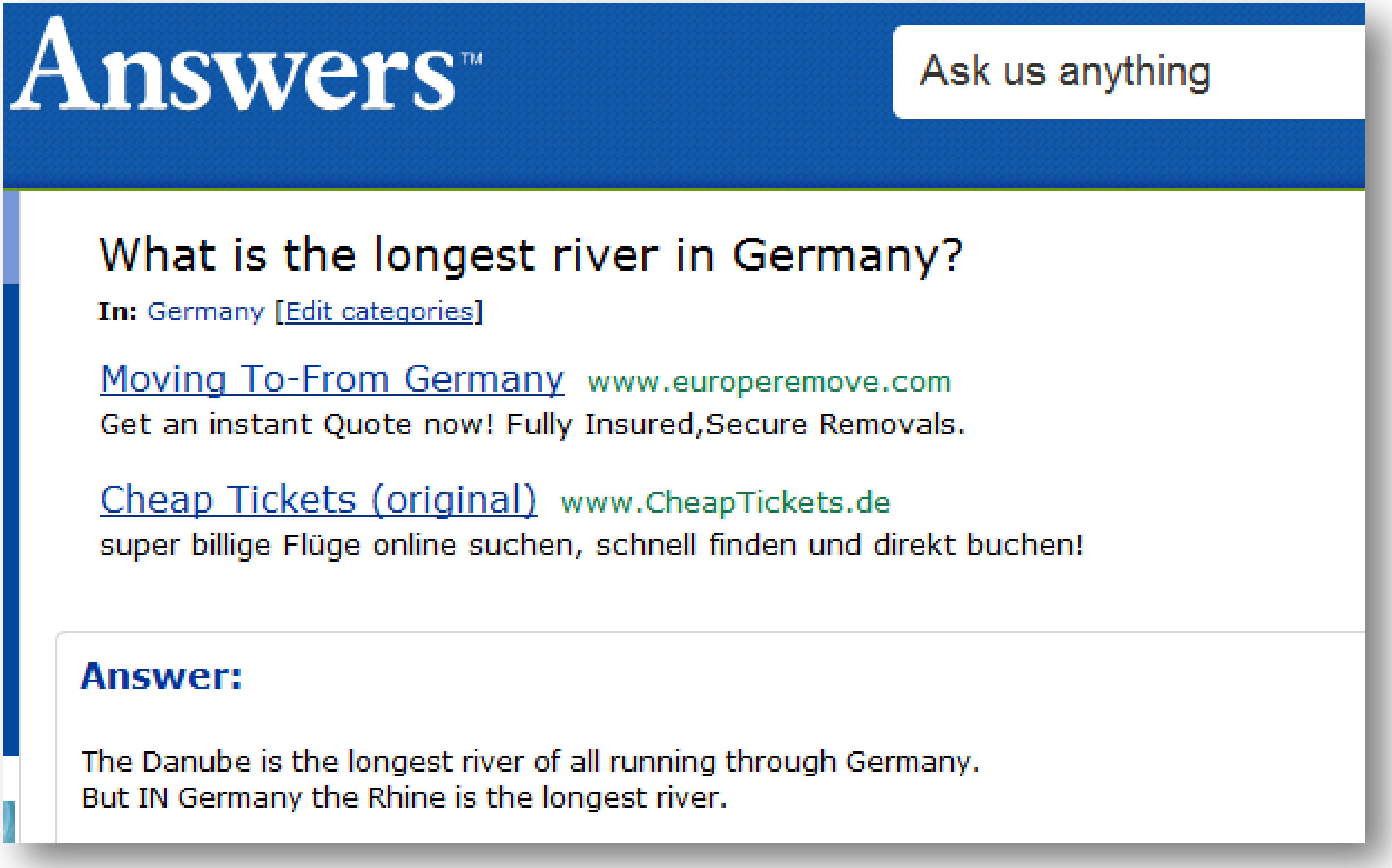
Recent population history:

Show projections | Log scale

Population and Family Planning Law

One-Child Policy

Community-based question answering



Answers™

What is the longest river in Germany?

In: Germany [Edit categories]

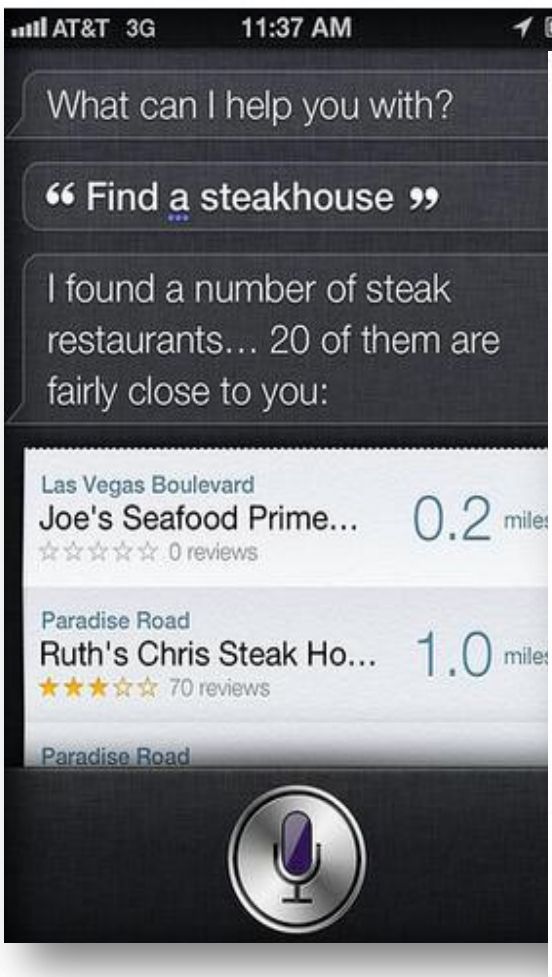
Moving To-From Germany www.europeremove.com
Get an instant Quote now! Fully Insured,Secure Removals.

Cheap Tickets (original) www.CheapTickets.de
super billige Flüge online suchen, schnell finden und direkt buchen!

Answer:

The Danube is the longest river of all running through Germany.
But IN Germany the Rhine is the longest river.

Mobile web search



AT&T 3G 11:37 AM

What can I help you with?

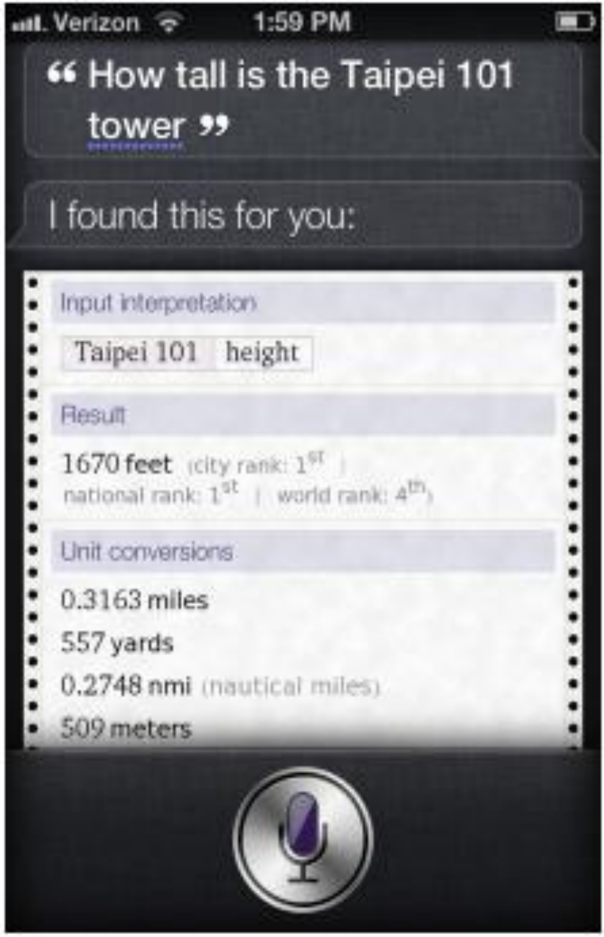
“ Find a steakhouse ”

I found a number of steak restaurants… 20 of them are fairly close to you:

Las Vegas Boulevard
Joe's Seafood Prime… 0.2 miles
☆☆☆☆☆ 0 reviews

Paradise Road
Ruth's Chris Steak Ho… 1.0 miles
★★★☆☆ 70 reviews

Paradise Road



Verizon 1:59 PM

“ How tall is the Taipei 101 tower ”

I found this for you:

Input interpretation

Taipei 101	height

Result

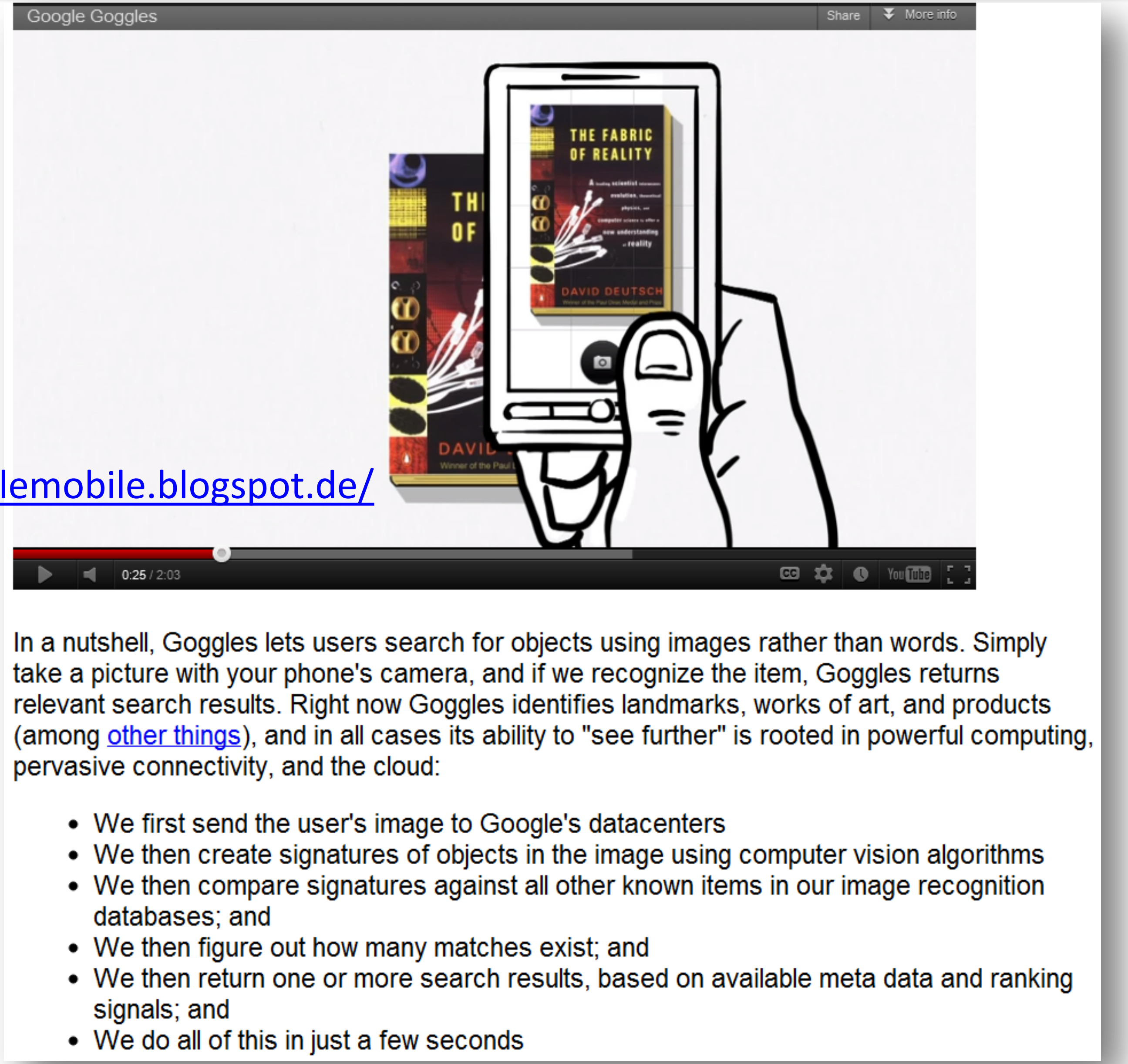
1670 feet (city rank: 1st | national rank: 1st | world rank: 4th)

Unit conversions

0.3163 miles
557 yards
0.2748 nmi (nautical miles)
509 meters

- ➢ Multimodal (videos, images, sounds, …)
- ➢ Personalized
- ➢ Context-aware
- ➢ Natural language
- ➢ Exact (answers)
- ➢ Social
- ➢ …

Mobile web search (… by pictures)

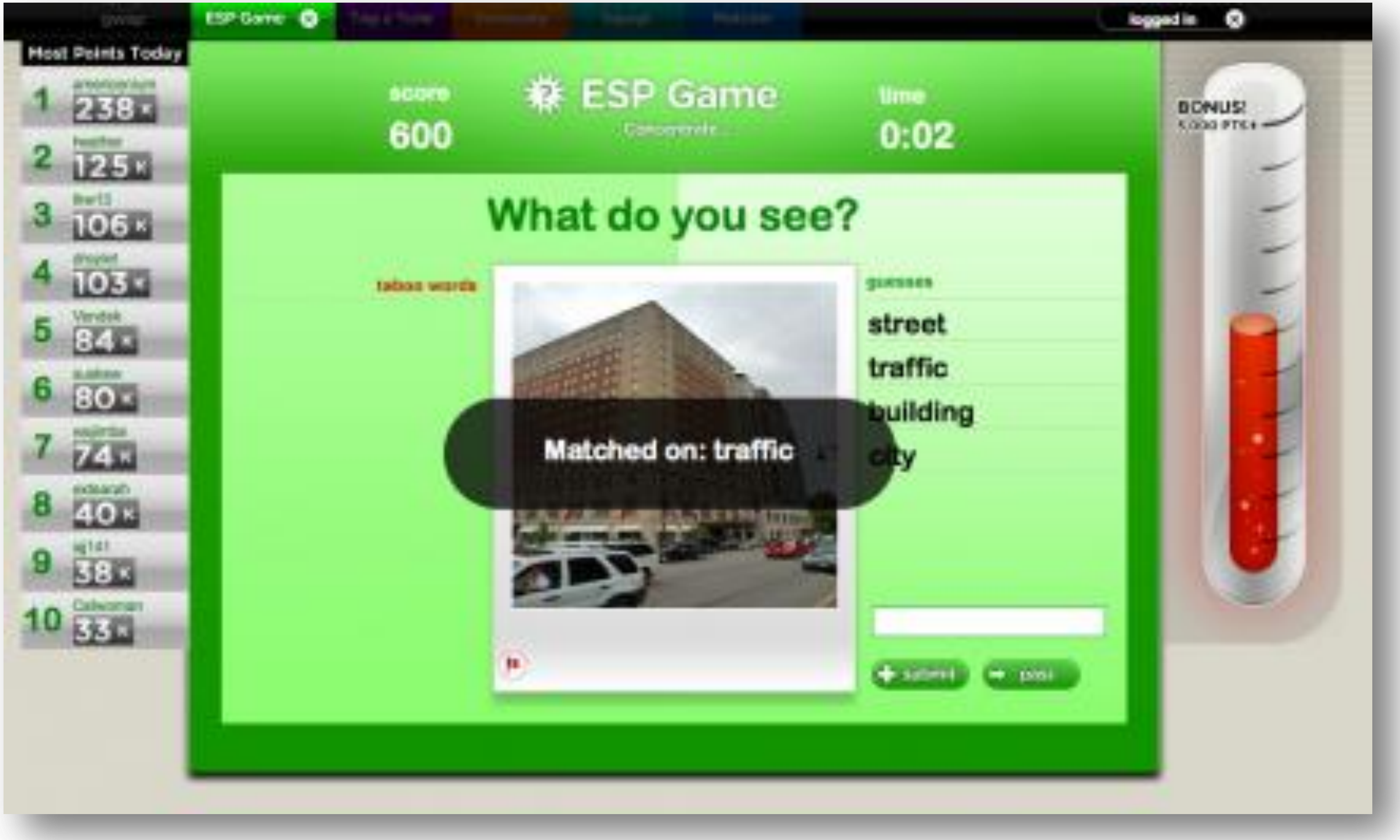


In a nutshell, Goggles lets users search for objects using images rather than words. Simply take a picture with your phone's camera, and if we recognize the item, Goggles returns relevant search results. Right now Goggles identifies landmarks, works of art, and products (among other things), and in all cases its ability to "see further" is rooted in powerful computing, pervasive connectivity, and the cloud:

- We first send the user's image to Google's datacenters
- We then create signatures of objects in the image using computer vision algorithms
- We then compare signatures against all other known items in our image recognition databases; and
- We then figure out how many matches exist; and
- We then return one or more search results, based on available meta data and ranking signals; and
- We do all of this in just a few seconds

Source: http://googlemobile.blogspot.de/

Social annotation for image search

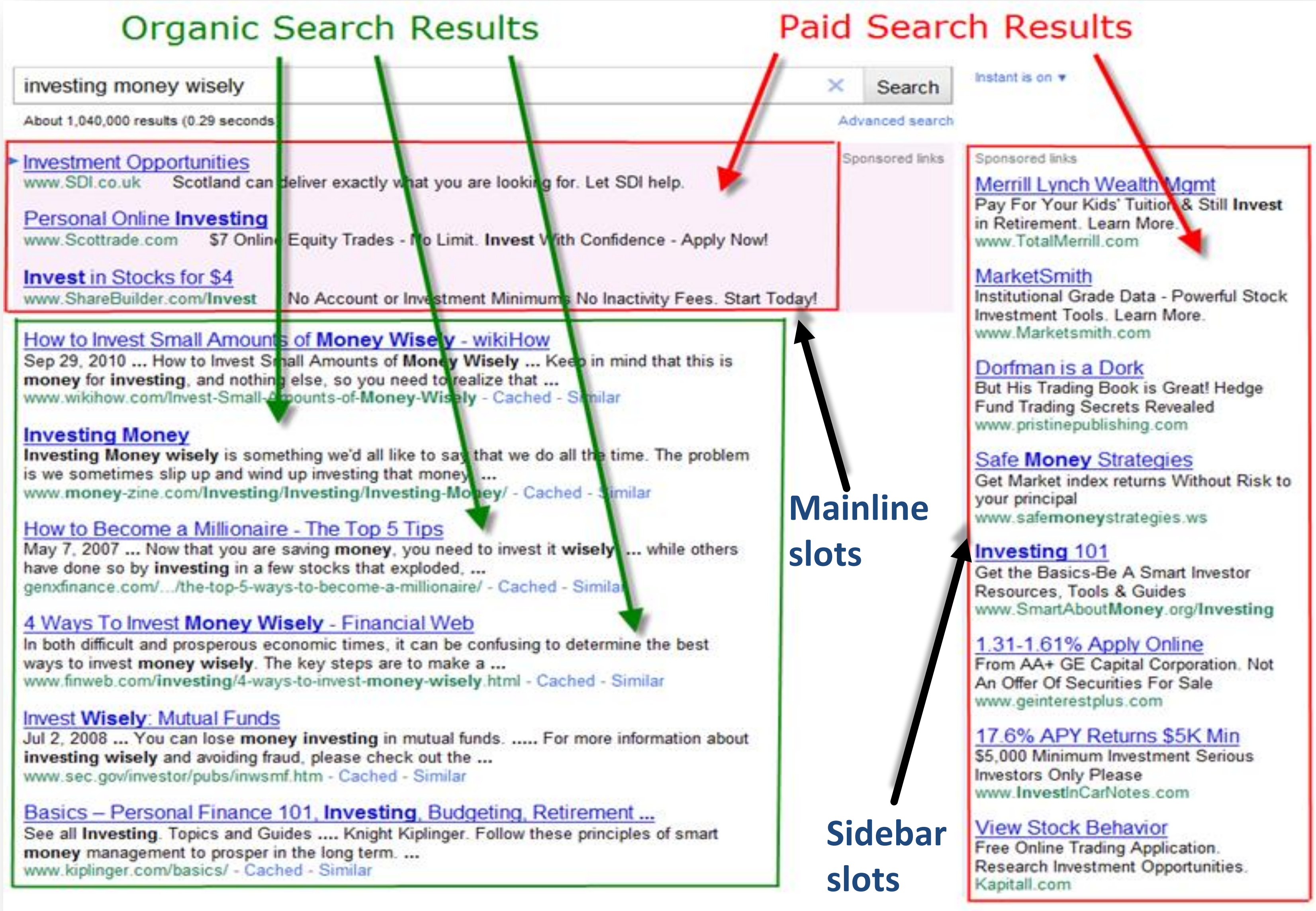


- ➢ Game with a purpose (exploits "wisdom of the crowds")
- ➢ Project initiated by Luis von Ahn et al. In 2004
- ➢ Estimation: 5000 people could tag all images on the web in 4 weeks

Social web (web 2.0)

- Social collaboration
  - Tagging, answering, computing, rating, editing, recommending, discussing, …
  - Examples: Wikipedia, flickr, del.icio.us, librarything, Amazon, Netflix, answers.com, ESP, foldit, …

Monetization of web search - sponsored search



Organic Search Results

Paid Search Results

Mainline slots

Sidebar slots

investing money wisely

Search

Instant is on ▾

About 1,040,000 results (0.29 seconds)

Advanced search

Sponsored links

Investment Opportunities
www.SDI.co.uk Scotland can deliver exactly what you are looking for. Let SDI help.

Personal Online **Investing**
www.Scottrade.com $7 Online Equity Trades - No Limit. **Invest** With Confidence - Apply Now!

Invest in Stocks for $4
www.ShareBuilder.com/**Invest** No Account or Investment Minimums No Inactivity Fees. Start Today!

How to Invest Small Amounts of **Money Wisely** - wikiHow
Sep 29, 2010 ... How to Invest Small Amounts of **Money Wisely** ... Keep in mind that this is **money** for **investing**, and nothing else, so you need to realize that ...
www.wikihow.com/Invest-Small-Amounts-of-**Money-Wisely** - Cached - Similar

Investing Money
Investing Money wisely is something we'd all like to say that we do all the time. The problem is we sometimes slip up and wind up investing that money ...
www.**money**-zine.com/**Investing/Investing/Investing-Money**/ - Cached - Similar

How to Become a Millionaire - The Top 5 Tips
May 7, 2007 ... Now that you are saving **money**, you need to invest it **wisely** ... while others have done so by **investing** in a few stocks that exploded, ...
genxfinance.com/.../the-top-5-ways-to-become-a-millionaire/ - Cached - Similar

4 Ways To Invest **Money Wisely** - Financial Web
In both difficult and prosperous economic times, it can be confusing to determine the best ways to invest **money wisely**. The key steps are to make a ...
www.finweb.com/**investing**/4-ways-to-invest-**money-wisely**.html - Cached - Similar

Invest **Wisely**: Mutual Funds
Jul 2, 2008 ... You can lose **money investing** in mutual funds. ..... For more information about **investing wisely** and avoiding fraud, please check out the ...
www.sec.gov/investor/pubs/inwsmf.htm - Cached - Similar

Basics – Personal Finance 101, **Investing**, Budgeting, Retirement ...
See all **Investing**. Topics and Guides .... Knight Kiplinger. Follow these principles of smart **money** management to prosper in the long term. ...
www.kiplinger.com/basics/ - Cached - Similar

Sponsored links

Merrill Lynch Wealth Mgmt
Pay For Your Kids' Tuition & Still **Invest** in Retirement. Learn More.
www.TotalMerrill.com

MarketSmith
Institutional Grade Data - Powerful Stock Investment Tools. Learn More.
www.Marketsmith.com

Dorfman is a Dork
But His Trading Book is Great! Hedge Fund Trading Secrets Revealed
www.pristinepublishing.com

Safe **Money** Strategies
Get Market index returns Without Risk to your principal
www.safe**money**strategies.ws

Investing 101
Get the Basics-Be A Smart Investor Resources, Tools & Guides
www.SmartAbout**Money**.org/**Investing**

1.31-1.61% Apply Online
From AA+ GE Capital Corporation. Not An Offer Of Securities For Sale
www.geinterestplus.com

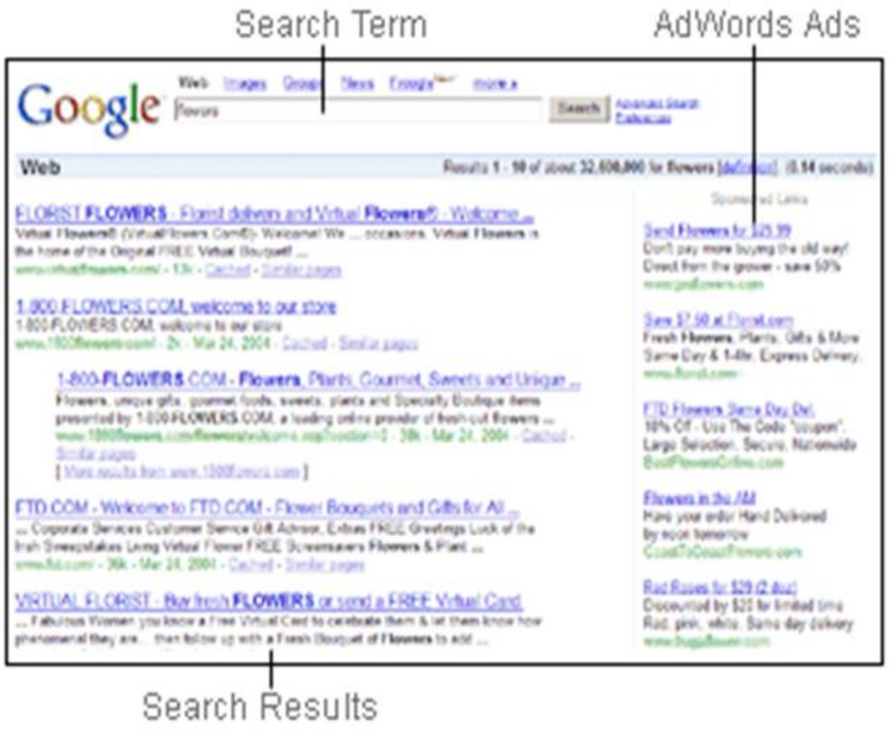
17.6% APY Returns $5K Min
$5,000 Minimum Investment Serious Investors Only Please
www.**Invest**InCarNotes.com

View Stock Behavior
Free Online Trading Application. Research Investment Opportunities.
Kapitall.com

Google™ Google Advertising Programs

For Advertisers: Google AdWords

Reach people when they are actively looking for information about your products and services online, and send targeted visitors directly to what you are offering. With AdWords cost-per-click pricing, it's easy to control costs—and you only pay when people click on your ad.



flowers

Web Results 1 - 10 of about 32,600,000 for flowers [definition]. (0.14 seconds)

FLORIST FLOWERS - Florist delivery and Virtual Flowers® - Welcome ...
Virtual Flowers® (VirtualFlowers.Com®)- Welcome! We ... occasions. Virtual Flowers is the home of the Original FREE Virtual Bouquet! ...
www.virtualflowers.com/ - 13k - Cached - Similar pages

1-800-FLOWERS.COM, welcome to our store
1-800-FLOWERS.COM, welcome to our store
www.1800flowers.com/ - 2k - Mar 24, 2004 - Cached - Similar pages

1-800-FLOWERS.COM - Flowers, Plants, Gourmet, Sweets and Unique ...
Flowers, unique gifts, gourmet foods, sweets, plants and Specialty Boutique items presented by 1-800-FLOWERS.COM, a leading online provider of fresh-cut flowers ...
www.1800flowers.com/flowers/welcome.asp?source=10 - 38k - Mar 24, 2004 - Cached - Similar pages
[More results from www.1800flowers.com]

FTD.COM - Welcome to FTD.COM - Flower Bouquets and Gifts for All ...
... Corporate Services Customer Service Gift Advisor, Extras FREE Greetings Luck of the Irish Sweepstakes Living Virtual Flower FREE Screensavers Flowers & Plant ...
www.ftd.com/ - 36k - Mar 24, 2004 - Cached - Similar pages

VIRTUAL FLORIST - Buy fresh FLOWERS or send a FREE Virtual Card
... Fabulous Women you know a Free Virtual Card to celebrate them & let them know how phenomenal they are ... then follow up with a Fresh Bouquet of Flowers to add ...

Sponsored Links

Send Flowers for $25.99
Don't pay more buying the old way!
Direct from the grower - save 50%
www.proflowers.com

Save $7.50 at Florist.com
Fresh Flowers, Plants, Gifts & More
Same Day & 1-4hr. Express Delivery.
www.florist.com

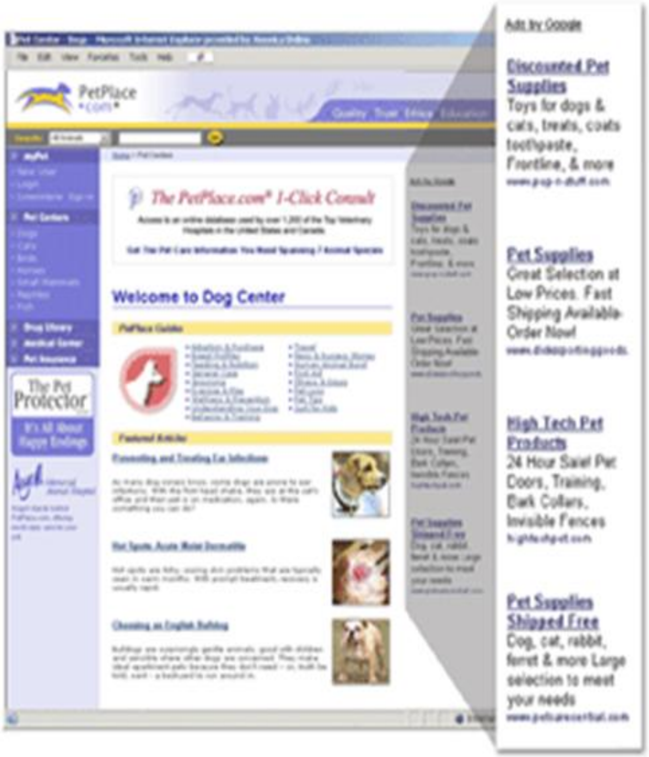
FTD Flowers Same Day Del.
10% Off - Use The Code "coupon".
Large Selection, Secure, Nationwide
BestFlowersOnline.com

Flowers in the AM
Have your order Hand Delivered by noon tomorrow
CoastToCoastFlowers.com

Red Roses for $29 (2 doz)
Discounted by $20 for limited time
Red, pink, white. Same day delivery
www.buyaflower.com

For Web Publishers: Google AdSense

Earn more revenue from your website, while providing visitors with a more rewarding online experience. Google AdSense™ delivers ads targeted to your content pages and, when you add Google WebSearch to your site, AdSense delivers targeted ads to your search results pages too. With AdSense you earn more ad revenue with minimal effort—and no additional cost.



Ads by Google

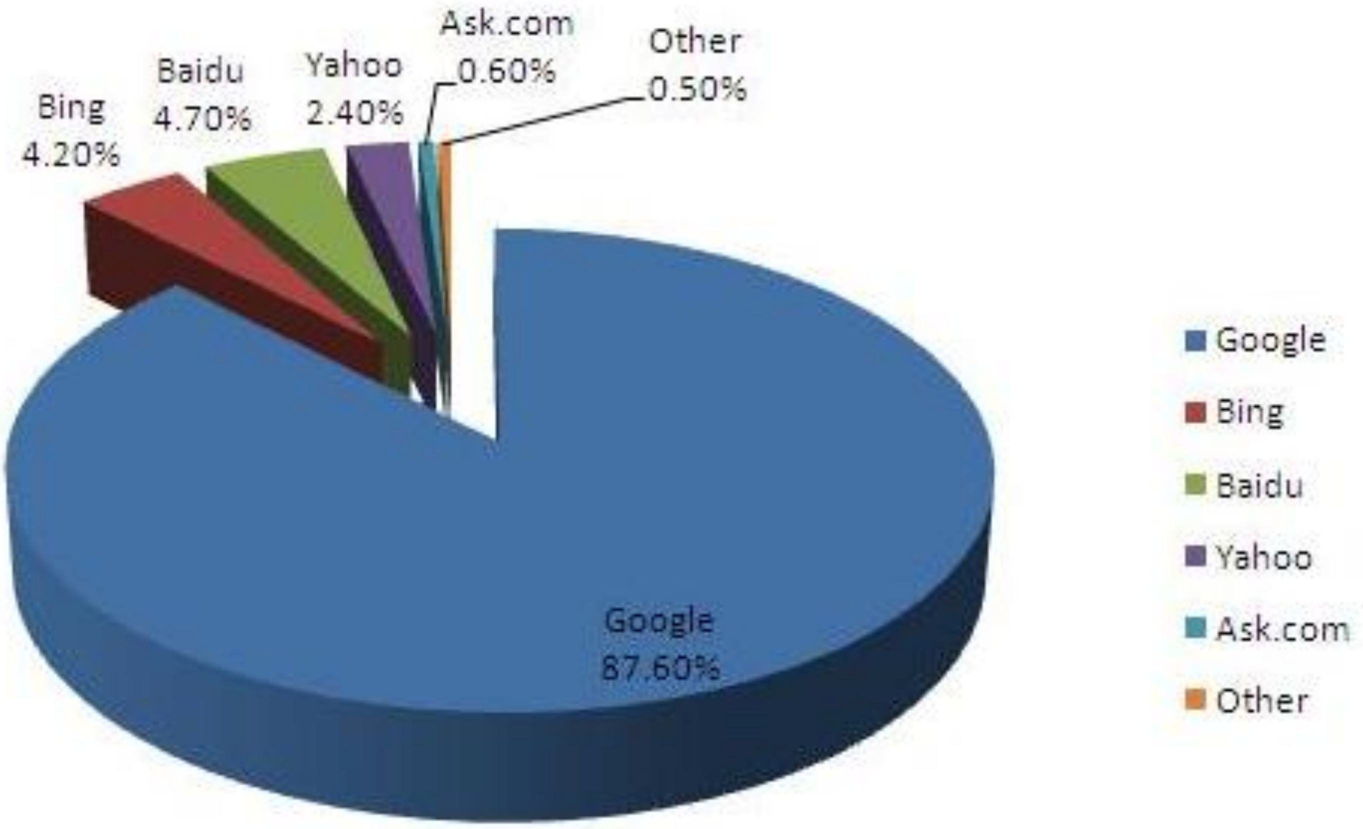
Discounted Pet Supplies
Toys for dogs & cats, treats, coats toothpaste, Frontline, & more
www.pup-n-duff.com

Pet Supplies
Great Selection at Low Prices. Fast Shipping Available. Order Now!
www.dickssportinggoods.

High Tech Pet Products
24 Hour Sale! Pet Doors, Training, Bark Collars, Invisible Fences
hightechpet.com

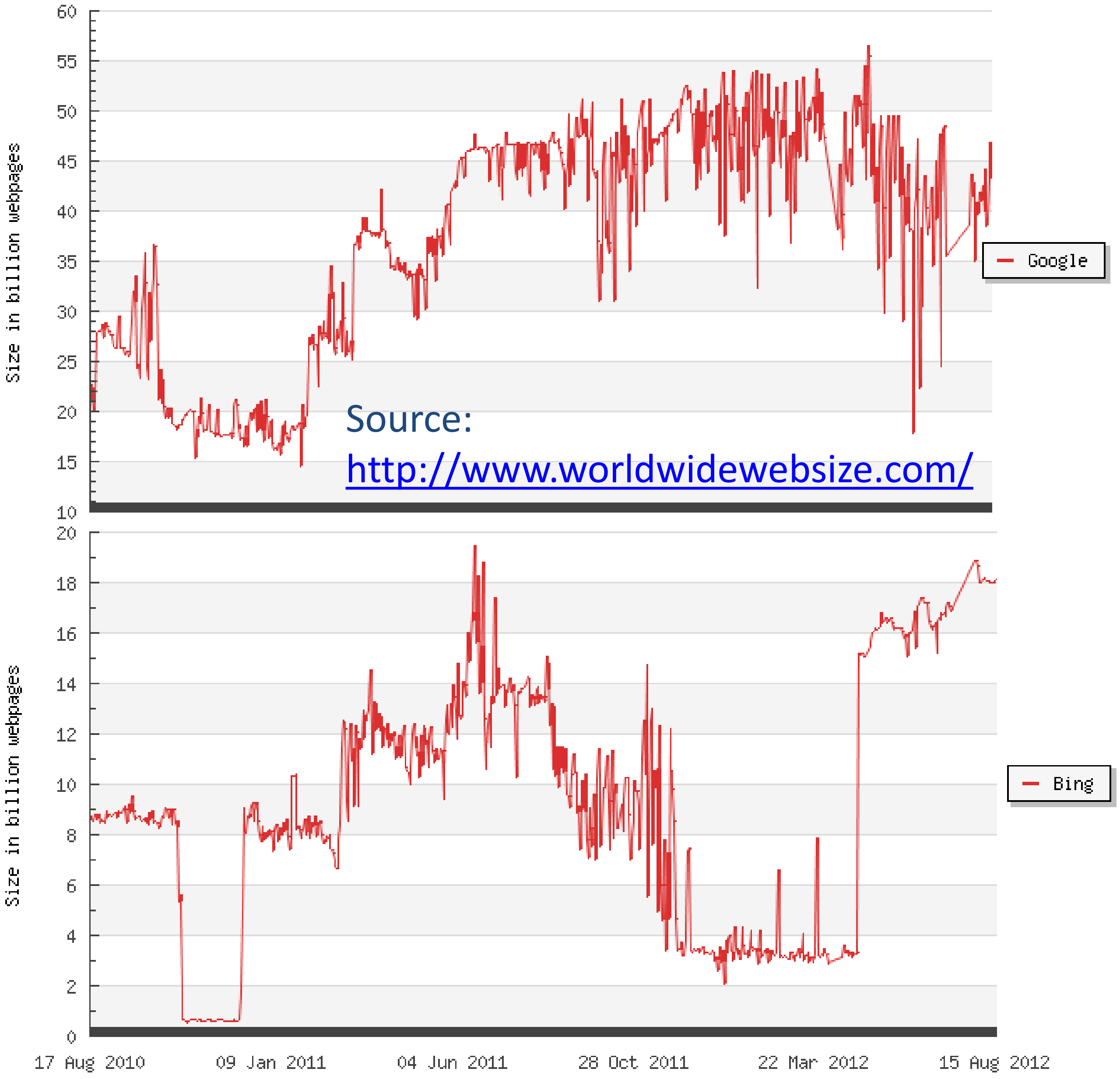
Pet Supplies Shipped Free
Dog, cat, rabbit, ferret & more Large selection to meet your needs
www.petcarecentral.com

Worldwide market share (as of May 2012)



Source : http://www.karmasnack.com/about/search-engine-market-share/

Indexed web pages



Source: http://www.worldwidewebsize.com/

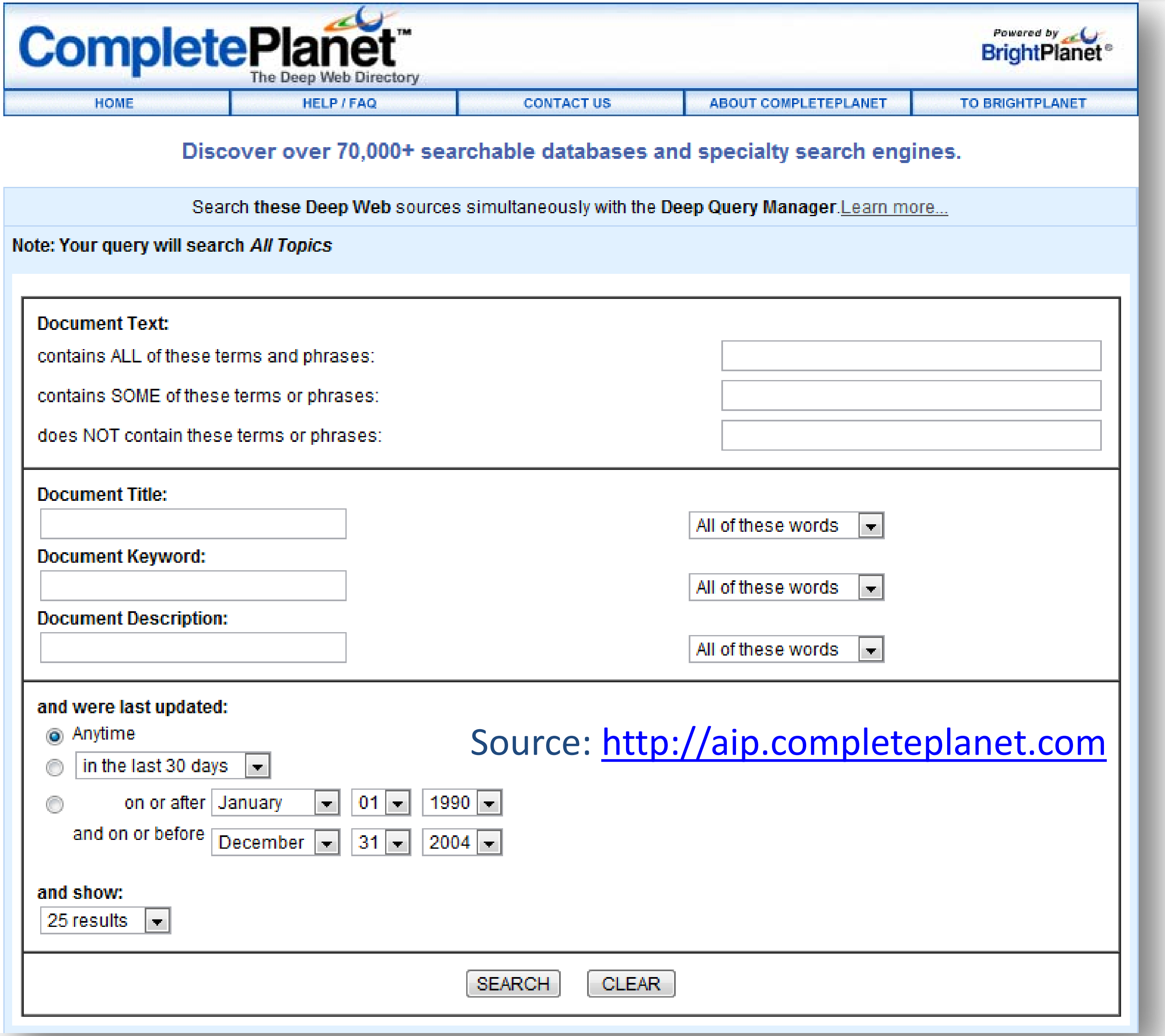
Deep web

- Content of the web that
  - resides in backend databases and
  - is not accessible through hyperlinks
  - can only be surfaced by submitting queries to query forms

- Examples
  - Yellow pages
  - Scientific databases (bio-medical-, geo-, weather-related data, publications, measurements, …)
  - Census data (i.e., population analysis, statistics, etc.)

- Estimation: several times larger than the visible web
- High-quality & structured information

Deep web: example (1)



CompletePlanet™
The Deep Web Directory

Powered by BrightPlanet®

Discover over 70,000+ searchable databases and specialty search engines.

Search **these Deep Web** sources simultaneously with the **Deep Query Manager**.Learn more...

Note: Your query will search *All Topics*

Document Text:

contains ALL of these terms and phrases:

contains SOME of these terms or phrases:

does NOT contain these terms or phrases:

Document Title: All of these words

Document Keyword: All of these words

Document Description: All of these words

and were last updated:

● Anytime

○ in the last 30 days

○ on or after January 01 1990

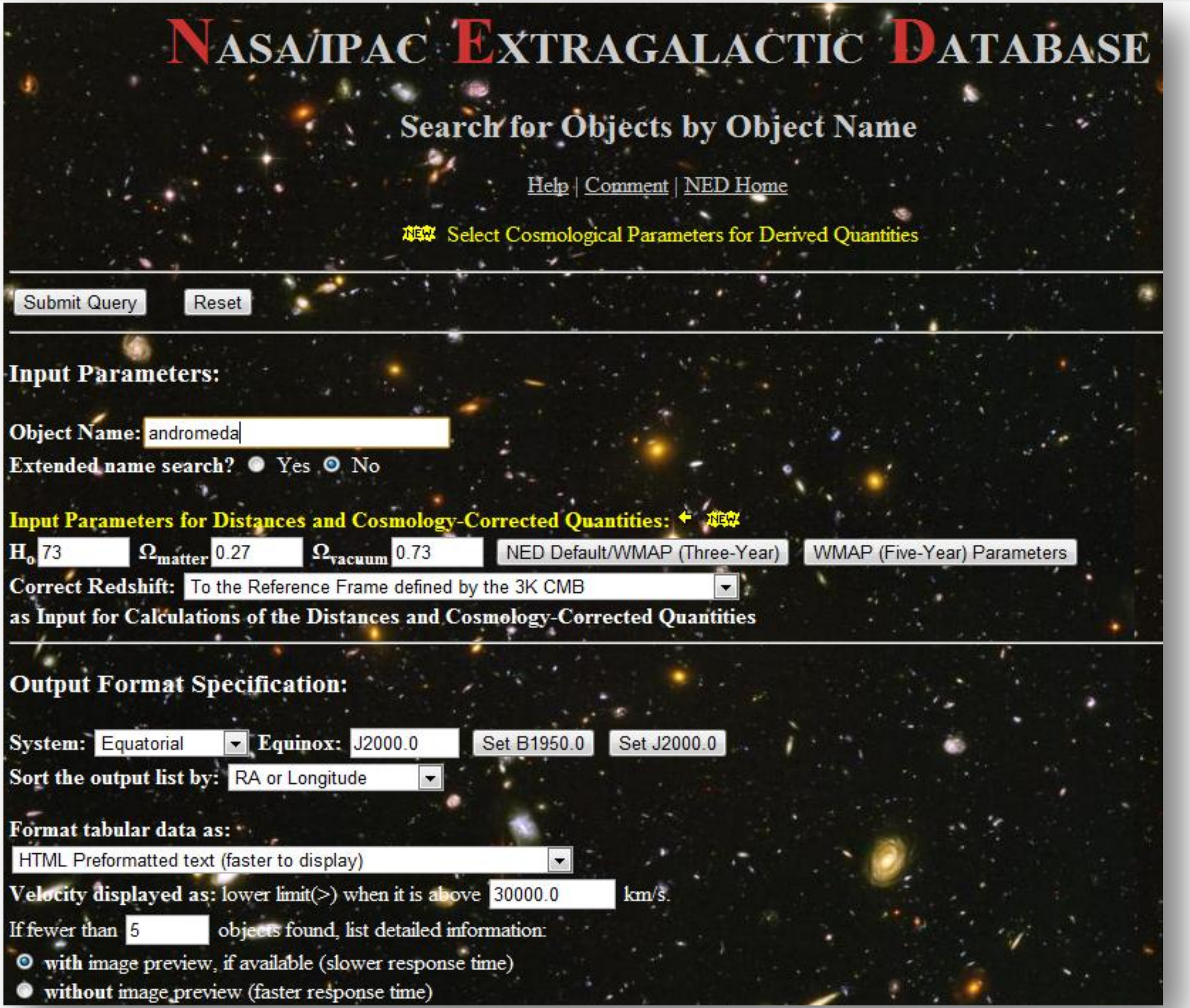
and on or before December 31 2004

and show:

25 results

Source: http://aip.completeplanet.com

Deep web: example (2)



NASA/IPAC EXTRAGALACTIC DATABASE

Search for Objects by Object Name

Help | Comment | NED Home

NEW Select Cosmological Parameters for Derived Quantities

Submit Query   Reset

Input Parameters:

Object Name: andromeda

Extended name search? ○ Yes ● No

Input Parameters for Distances and Cosmology-Corrected Quantities: NEW

H_o 73 Ω_{matter} 0.27 Ω_{vacuum} 0.73 NED Default/WMAP (Three-Year) WMAP (Five-Year) Parameters

Correct Redshift: To the Reference Frame defined by the 3K CMB

as Input for Calculations of the Distances and Cosmology-Corrected Quantities

Output Format Specification:

System: Equatorial Equinox: J2000.0 Set B1950.0 Set J2000.0

Sort the output list by: RA or Longitude

Format tabular data as:

HTML Preformatted text (faster to display)

Velocity displayed as: lower limit(>) when it is above 30000.0 km/s.

If fewer than 5 objects found, list detailed information:

● with image preview, if available (slower response time)

○ without image preview (faster response time)

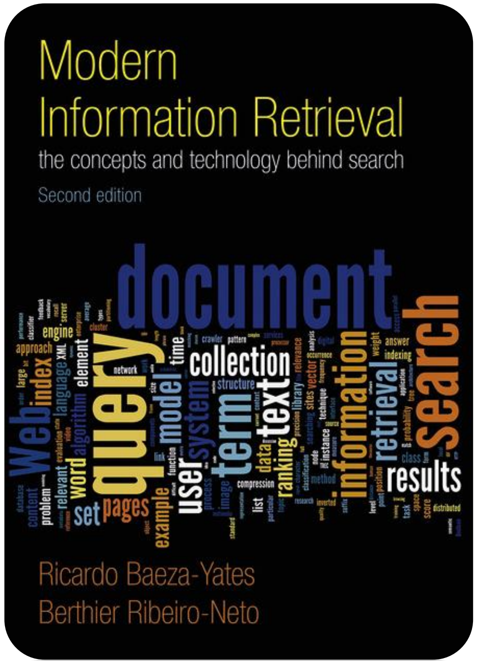
Source: http://ned.ipac.caltech.edu

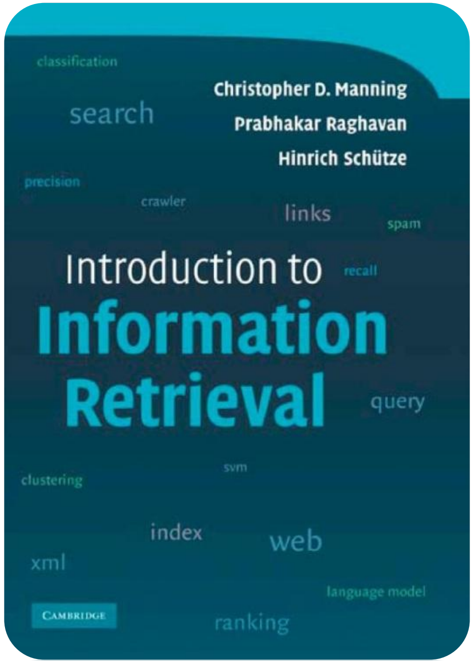
Sufacing the deep web

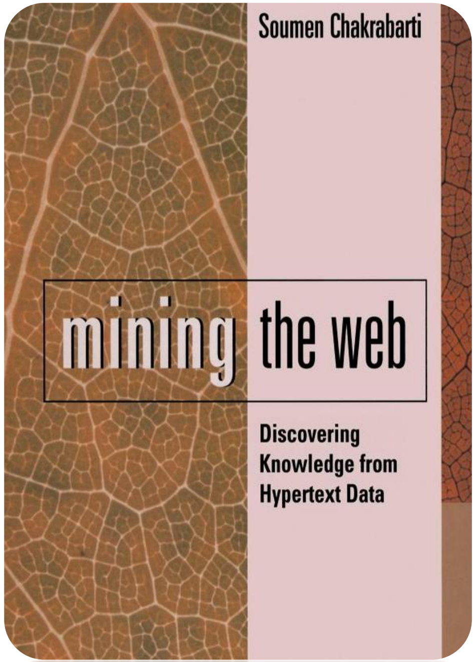
- **Strategy 1** (used by Google, see Madhavan et al. 2006):
  - Probe query boxes in the form with random keywords
  - Use words in the results as new query terms (bootstrapping)
- **Strategy 2** (much more challenging -- semantic analysis):
  - Infer the data type (or semantic concept) behind a query box by analyzing scripts (for required values), exploiting surrounding terms, or probing
  - Surface information by querying with appropriate values (e.g. from other databases / knowledge bases)
  - Analyze associations in surfaced data to infer underlying schema
- **Problems**
  - Map queries to relevant interfaces
  - Integrating schemas of sources from similar domains
  - Merging results into global ranking

Related literature

- Books







- Important conferences
  - SIGIR, TREC, WSDM, WWW, WebSci, CIKM, ECIR, ICML, ECML, KDD, JCDL, …
- IR toolkits
  - Lucene (http://lucene.apache.org/core/)
  - The Lemur Project (http://www.lemurproject.org/)

What else can we expect from future search?

Google 2084
Google Search
I'm Feeling Lucky
I'm Feeling Paranoid
Search:
○ Your Brain
○ Your Home
○ Family
○ Friends
○ Ex-Friends
○ Relatives
○ Your Past
○ Your Present
○ Your Future
○ Pornography
○ Co-workers
○ Ex-spouse(s)
○ Enemies
○ Satellite Photos of People You Want To Spy On
○ Satellite Photos of People Spying On You
○ Other People's Con- versations
©2020 Google

Source: funnyjunk.com