

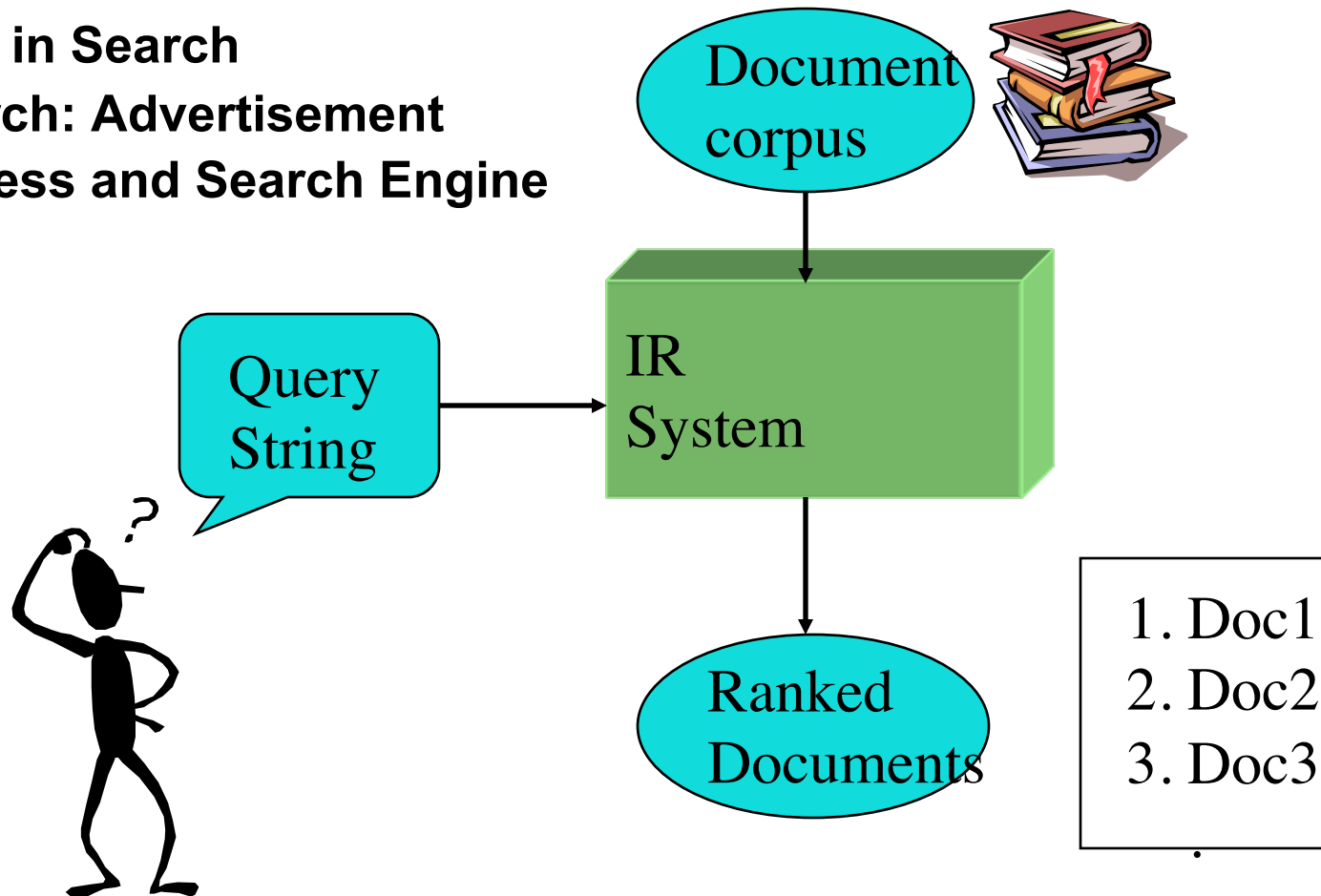
Introduction to Information Retrieval and Web Search

Tao Yang

UCSB CS293S, Winter 2017

Table of Content

- Information Retrieval
- Search Engine Architecture and Process
- Web Content and Size
- Users Behavior in Search
- Sponsored Search: Advertisement
- Impact to Business and Search Engine Optimization
- Related fields



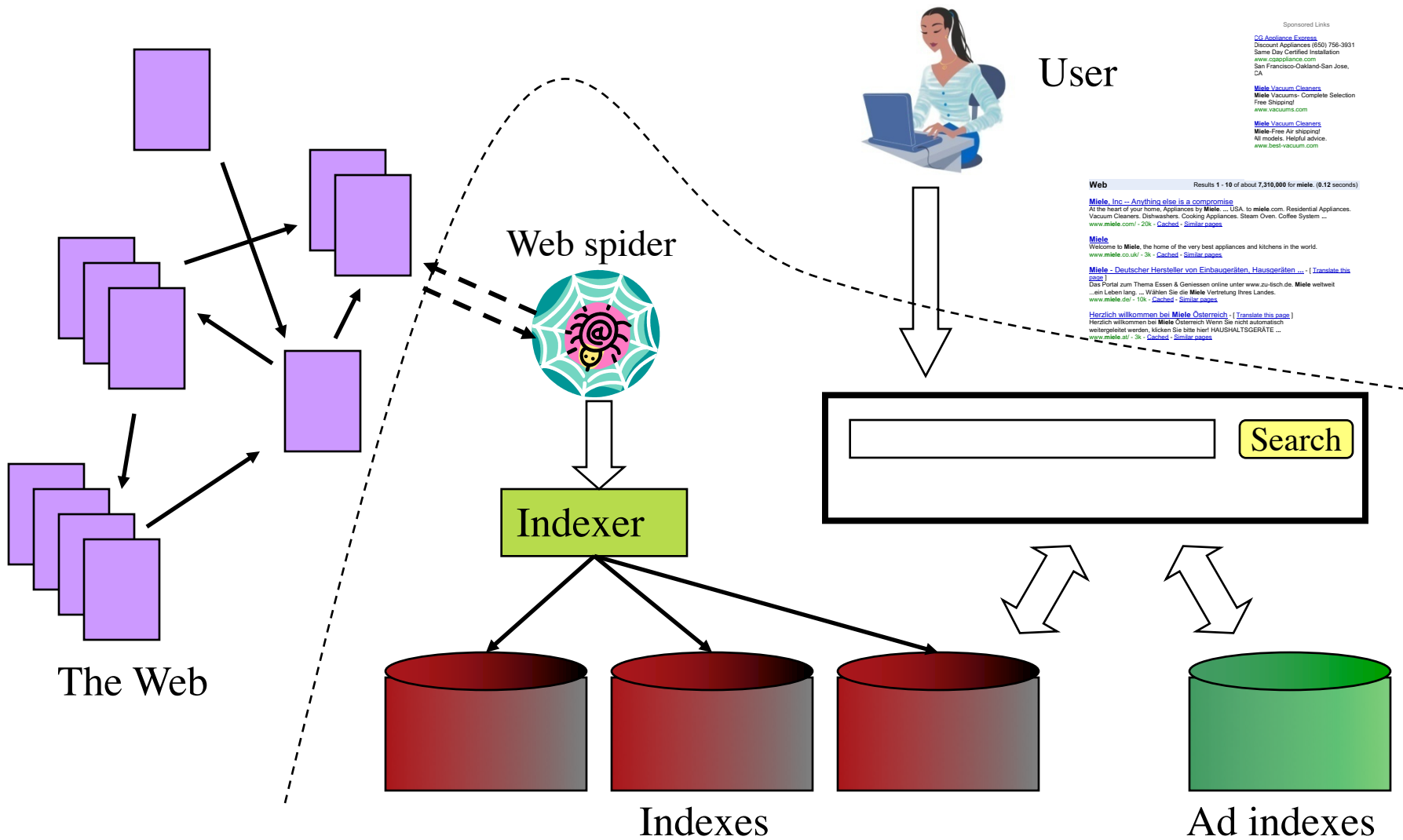
History of IR and Web Search

- **1960-70's:**
 - Initial exploration of text retrieval systems for “small” corpora of scientific abstracts, and law and business documents.
 - Development of the basic Boolean and vector-space models of retrieval.
- **1980's:**
 - Larger document database systems, many run by companies:
 - Lexis-Nexis
 - Dialog
 - MEDLINE
- **1990's:**
 - Organized Competitions
 - NIST TREC
 - Searching FTPable documents on the Internet
 - Archie
 - WAIS
 - Searching the World Wide Web
 - Lycos
 - Yahoo
 - Altavista

History of IR/Web Search

- **2000's**
 - Link analysis for Web Search
 - Google
 - Inktomi
 - Teoma
 - Feedback based engine:
 - DirectHit (Ask.com/Ask Jeeves)
 - Automated Information Extraction
 - Whizbang
 - Fetch
 - Burning Glass
 - Question Answering
 - TREC Q/A track
 - Ask.com/Ask Jeeves
- **2000's continued:**
 - Multimedia IR
 - Image
 - Video
 - Audio
 - music
 - Cross-Language IR
 - Document Summarization
 - Mobile search

Web search basics

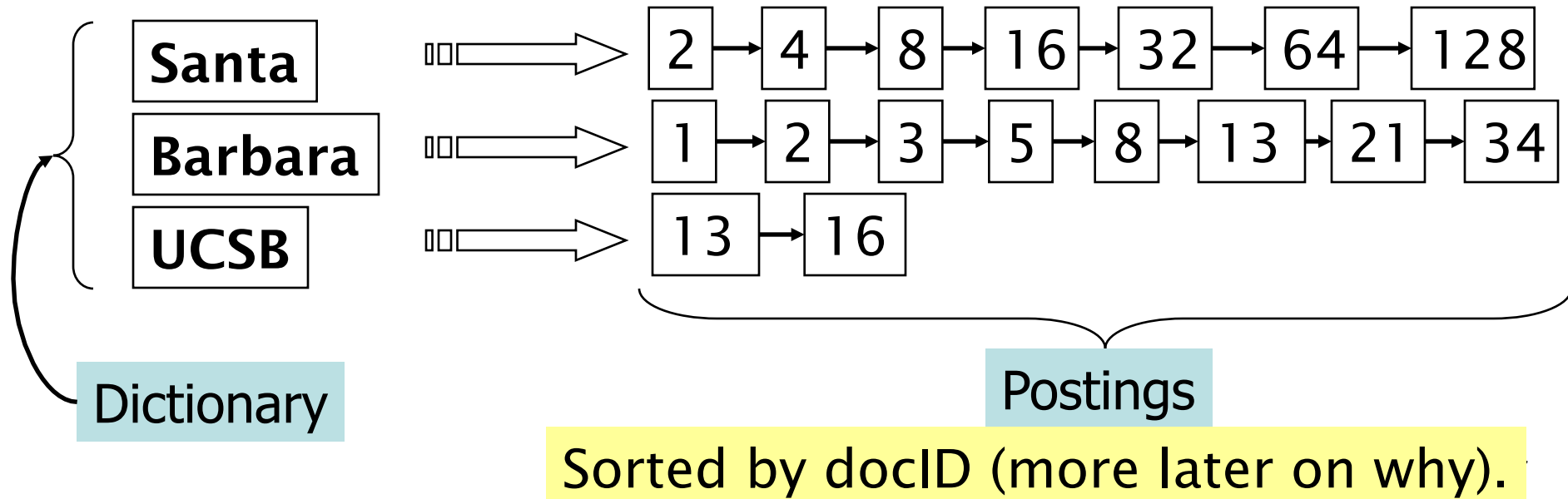


Search engine architecture: key pieces

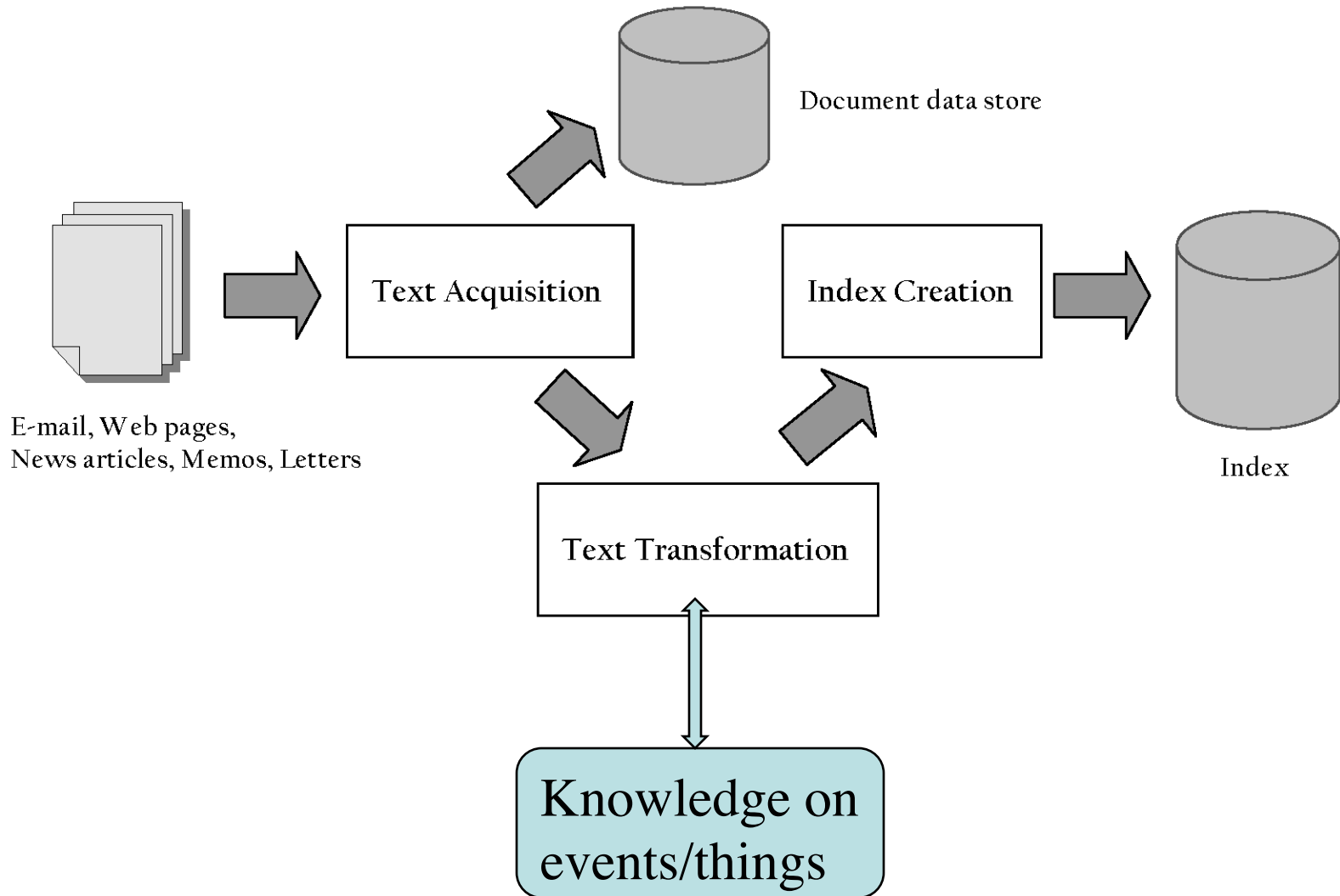
- **Spider (a.k.a. crawler/robot) – builds corpus**
 - Collects web pages recursively
 - For each known URL, fetch the page, parse it, and extract new URLs
 - Repeat
 - Additional pages from direct submissions & other sources
- **Indexer and offline text mining**
 - create inverted indexes so online system can search
 - Enrich knowledge on things and their relationship (e.g. names and events) and documents through data mining and learning
- **Online query process – serves query results**
 - Front end – query reformulation, word processing
 - Back end – finds matching documents and ranks them

Inverted index

- **Linked lists generally preferred to arrays**
 - Dynamic space allocation
 - Insertion of terms into documents easy
 - Space overhead of pointers



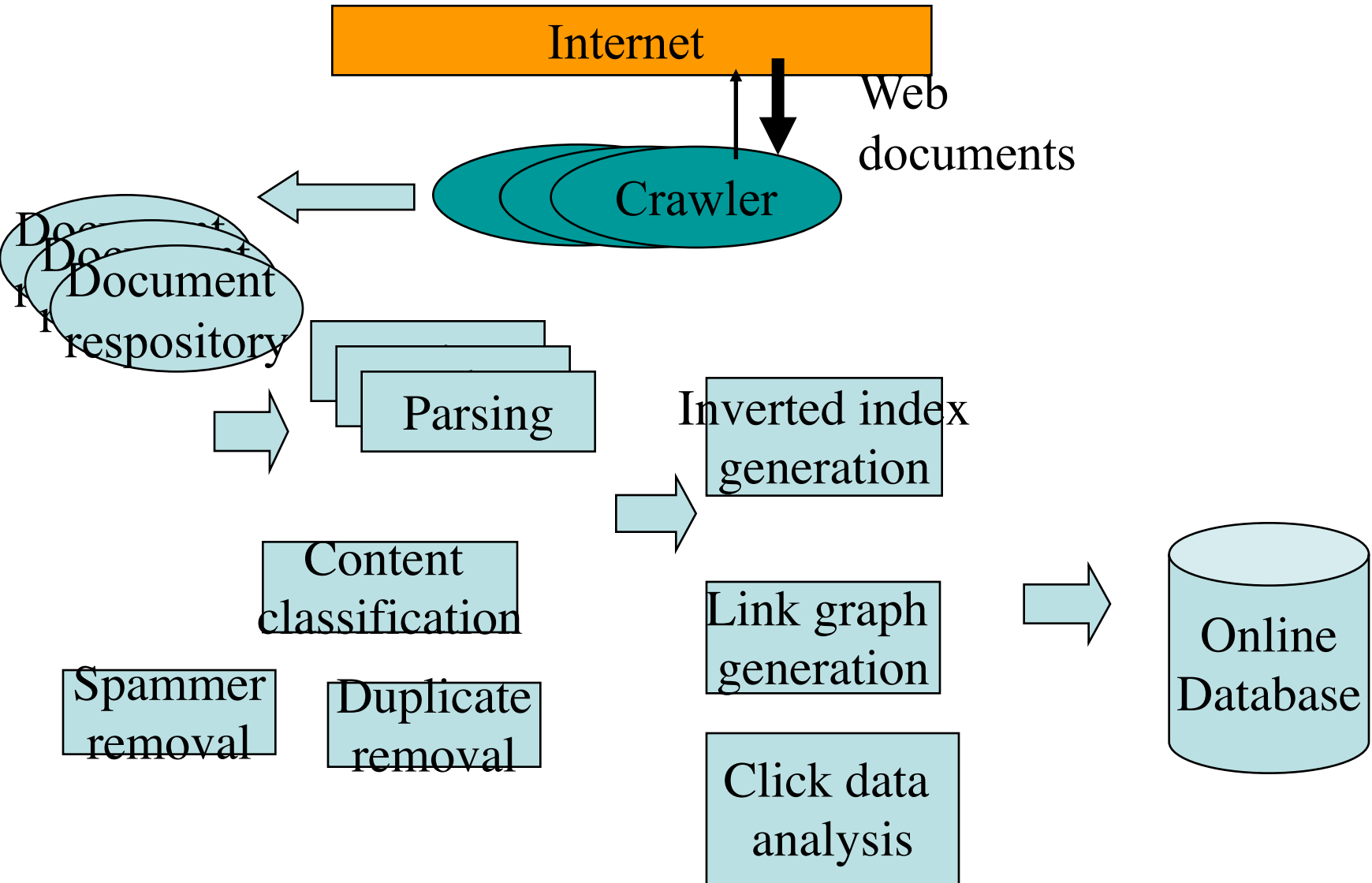
Indexing Process



Indexing Process with Mining

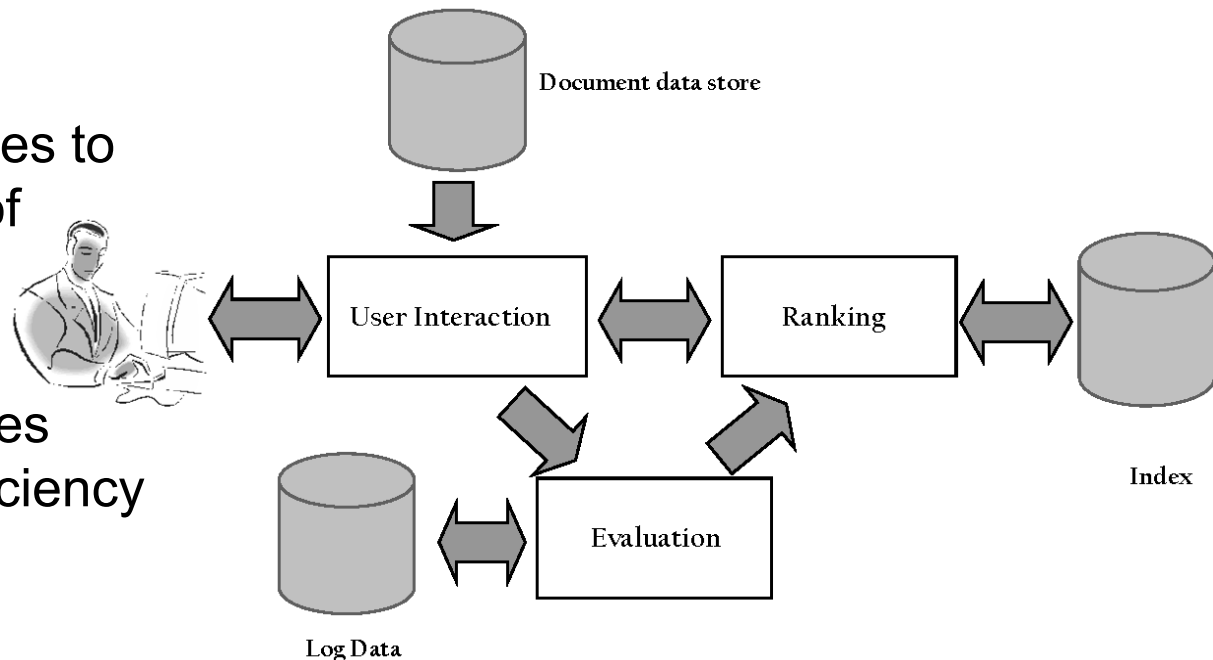
- **Text acquisition**
 - identifies and stores documents for indexing
- **Text transformation**
 - transforms documents into *index terms* or *features*
- **Index creation**
 - takes index terms and creates data structures (*indexes*) to support fast searching
- **Data mining**
 - Knowledge learning on things (people name, organization, etc) and their relationship (knowledge graphs)

Indexing and Mining at Ask.com

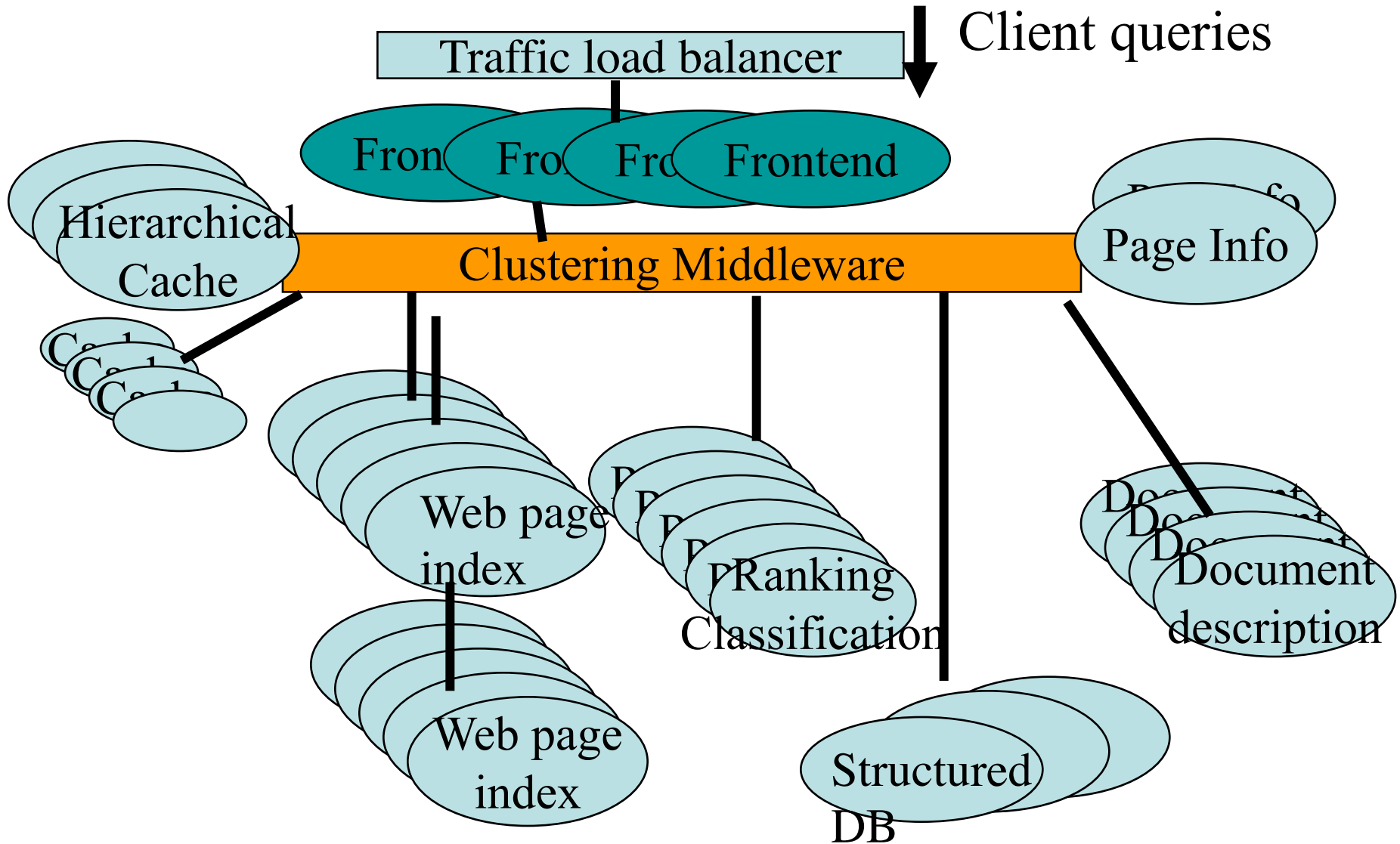


Query Process

- **User interaction**
 - supports creation and refinement of query, display of results
- **Ranking**
 - uses query and indexes to generate ranked list of documents
- **Evaluation**
 - monitors and measures effectiveness and efficiency (primarily offline)



Ask.com Online Engine Architecture



User Interaction

- **Query transformation**
 - Improves initial query,
 - Stopword removal, spell correction, long query trimming
 - marriot hotel at golet
 - *Spell checking suggestion* and *query suggestion* provide alternatives to original query
 - Did you mean “Marriott hotel at Goelta”?
 - *Query expansion* and *relevance feedback* modify the original query with additional terms
 - *UC santa babara admission rate*

User Interaction

The screenshot shows a Bing search results page for the query 'santa barbara'. At the top, the search bar contains 'santa barbara' and a magnifying glass icon. To the right, there are links for 'Sign in', a user profile icon, and a notification icon. Below the search bar, navigation tabs for 'Web', 'Images', 'Videos', 'Maps', 'News', and 'Explore' are visible. The main results area shows '7,510,000 RESULTS' and 'Any time' filter. The first result is an advertisement for 'Santa Barbara - 100 Santa Barbara Hotels' from Booking.com, featuring a grid of five images showing coastal views and a map of Santa Barbara. Below the ad are several links for hotel booking, such as 'Most Popular Hotels', 'Book your Hotel Online', 'Get Instant Confirmation', 'Budget Hotels', 'Best Reviewed Hotels', and 'Luxury Hotels'. The second result is the official website for Santa Barbara, California, with a link to 'www.santabarbaraca.gov'. The third result is a TripAdvisor link for 'The Top 10 Things to Do in Santa Barbara'. On the right side of the page, there is a large image grid and a map of Santa Barbara. Below the map, the text 'Santa Barbara California' is followed by a brief description: 'Santa Barbara is the county seat of Santa Barbara County in the U.S. state of California. Situated on a south-facing section of coastline, the longest such section on the West Coast of the United States, the city lies between the steeply rising Santa Ynez Mountains and the Pacific Ocean. Santa Barbara's climate is often described as Mediterranean...'. At the bottom right, there are links for 'Wikipedia' and 'Official website', and the local time is shown as '11:23 AM 1/9/2017'.

- **Results output**

- Constructs the display of ranked documents for a query
 - Merge results from multiple channels
 - Retrieves appropriate *advertising*
- Generates *snippets (dynamic description)* to show how queries match documents
 - *Highlights* important words and passages
- May provide *clustering* and other visualization tools

Online System Support

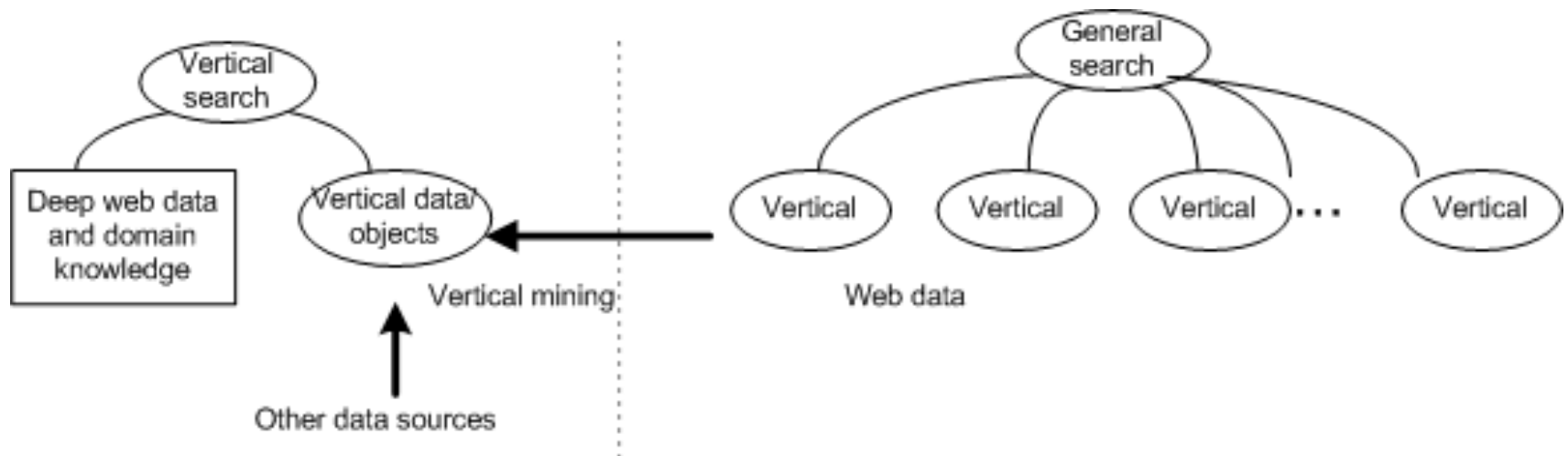
- **Performance optimization**
 - Designing matching&ranking algorithms for efficient processing
 - *Term-at-a time vs. document-at-a-time* processing
 - *Safe vs. unsafe* optimizations
- **Distribution**
 - Processing queries in a distributed environment
 - *Query broker* distributes queries and assembles results
 - *Caching* is a form of distributed searching

Evaluation

- **Logging**
 - Logging user queries and interaction is crucial for improving search effectiveness and efficiency
 - *Query logs* and *clickthrough data* used for query suggestion, spell checking, query caching, ranking, advertising search, and other components
- **Ranking analysis**
 - Measuring and tuning ranking effectiveness
- **Performance analysis**
 - Measuring and tuning system efficiency

General Search vs. Vertical Search

- **General Search:** identify relevant information with a horizontal/exhaustive view of the world.
- **Vertical Search:**
 - Focus on specific segment of web content
 - Integrate domain knowledge (e.g. taxonomies /ontology), & deep web
 - Examples: travel in Expedia, products in Amazon.



Example of Vertical Search: Question Answering

where is my stimulus che... x +

← → ↻ 🏠 ☆ <http://www.ask.com/ans?qsrc=167&o=0&l=dir&q=where%20is%20my%20stimulus%20check>

Ask where is my stimulus check **Search Q&A** **Search the Web**

Web Images News Deal\$ Videos Q&A Beta More ▾

Top Answers

“ Well if you requested your stimulus check to arrive by mail then you can expect to wait up to approximately 6 weeks for it to arrive. If you are expecting direct deposit then the wait time will be about 2 weeks.
http://answers.ask.com/Business/Finance/where_is_my_st... See entire page »

“ There are no stimulus check being mailed out this year. Instead of receiving a check from the government, most single taxpayers will see an adjustment to their tax withholding in their paychecks in 2009 and 2010, giving them about \$45 extra...
<http://answers.yahoo.com/question/index?qid=2009041111...> See entire page »

“ that's why u guys should have signed up for direct deposit... I received mine a month ago and stimulated vegas with it lol
<http://www.yelp.com/topic/santa-ana-where-is-my-stimul...> See entire page »

[See more answers to your question »](#)


Answers to Other Common Questions

[Are people with social security ans ssi going to get stimulus che...?](#)
“ In 2009 Retirees, SSI and Disabled vets received a stimulus check of \$250. ChaCha
<http://www.chacha.com/question/are-people-with-social-s...>

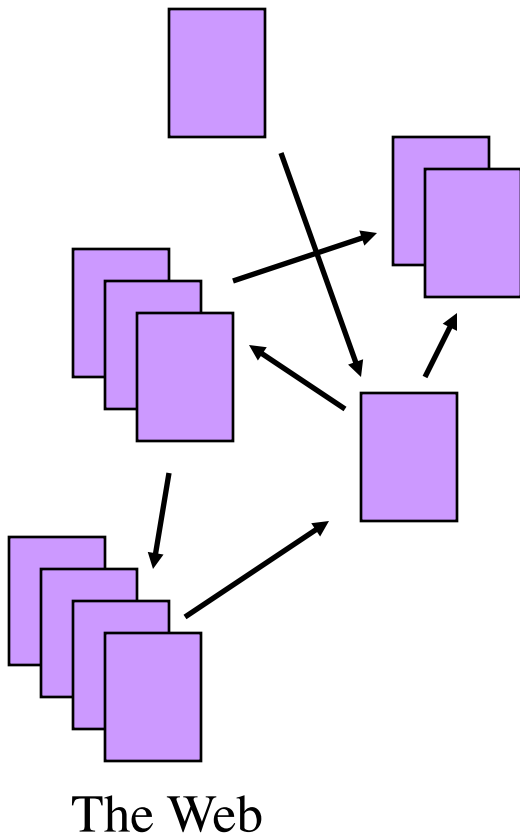
[When will i receive my stimulus check?](#)
“ Finding when you'll receive your stimulus check will depend on a few things. It can depend on if you've already filed or not and when you've filed. Of course, when it comes to the IRS, they have a specific schedule for any situation. You ca...

EN ?

Table of Content

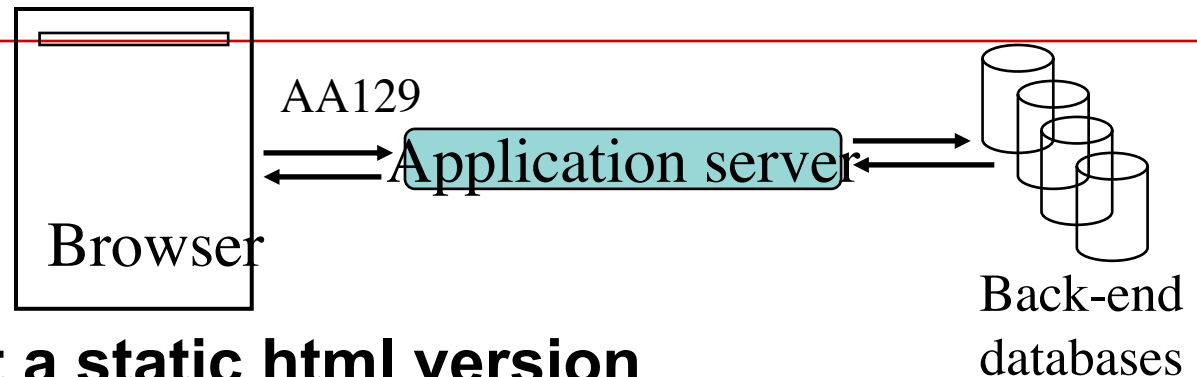
- **Information Retrieval**
- **Search Engine Architecture and Process**
- **Web Content and Size** 
- **Users Behavior in Search**
- **Sponsored Search: Advertisement**
- **Impact to Business and Search Engine Optimization**
- **Related Fields**

Characteristics of Web Content



- **No design/co-ordination**
- **Distributed content creation, linking**
- **Content includes truth, lies, obsolete information, contradictions ...**
- **Structured (databases), semi-structured ...**
- **Scale -- huge**
- **Growth – slowed down from initial “volume doubling every few months”**
- **Content can be *dynamically generated***

Dynamic Web Content

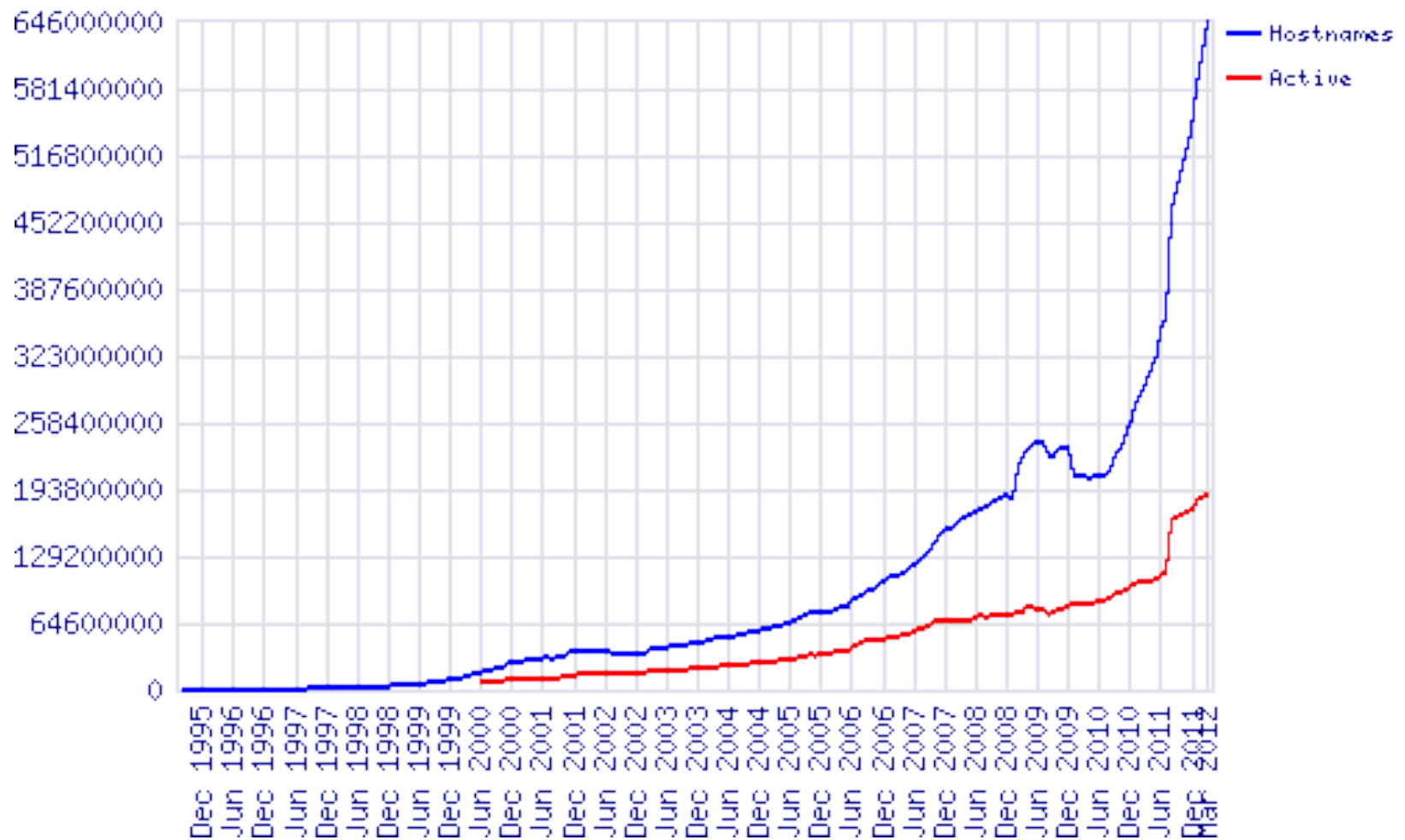


- **A page without a static html version**
 - E.g., current status of flight AA129
 - Current availability of rooms at a hotel
- **Usually, assembled at the time of a request from a browser**
 - Typically, URL has a '?' character in it
- **Most dynamic content is ignored by web spiders**
 - Many reasons including malicious spider traps
 - Acquired for some content (e.g. news stores)
 - Application-specific spidering

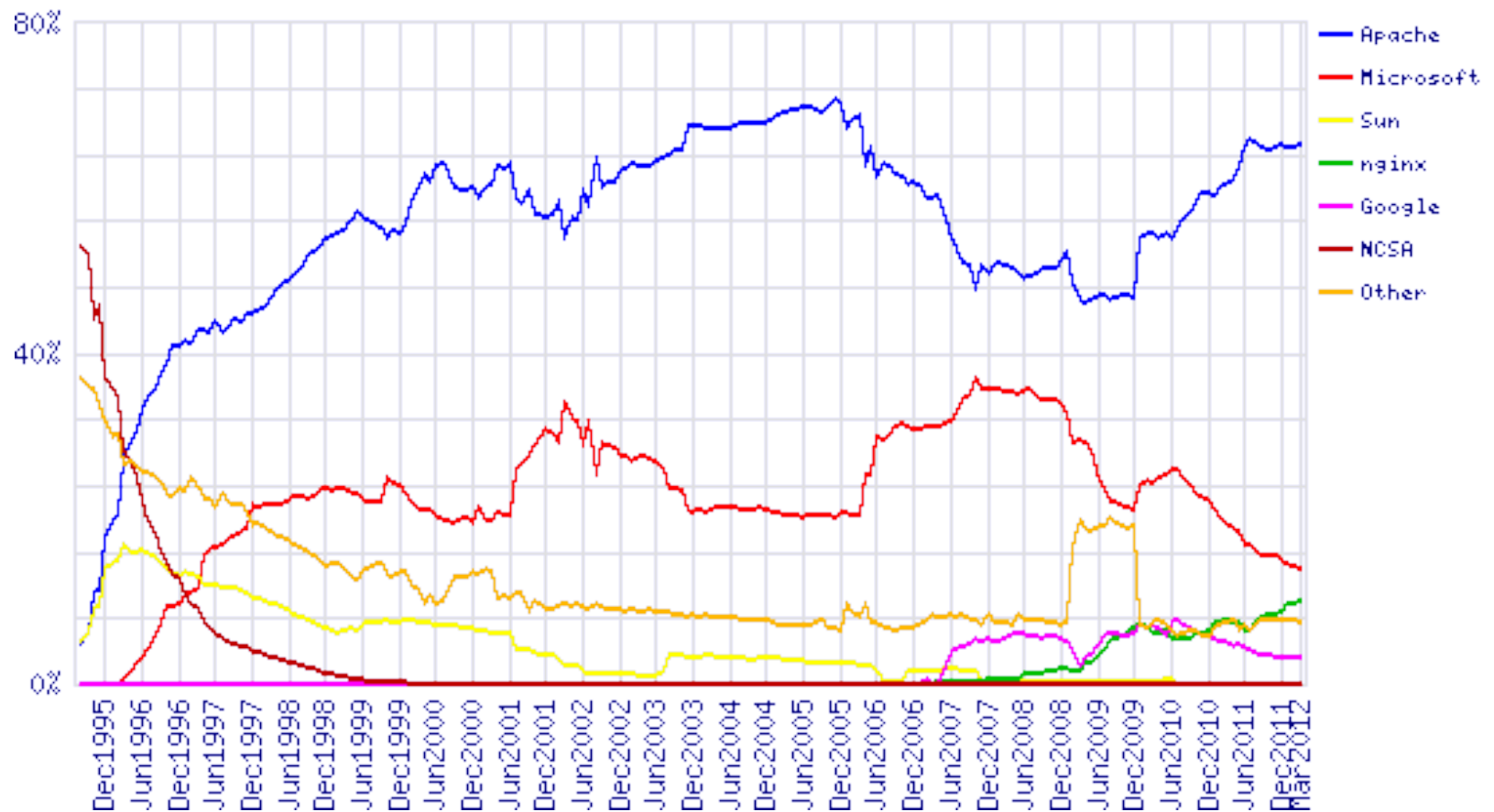
The web: size

- **What is being measured?**
 - Number of hosts
 - Number of (static) html pages
 - Volume of data
- **Number of hosts – netcraft survey**
 - http://news.netcraft.com/archives/web_server_survey.html
 - <http://news.netcraft.com/archives/2014/04/02/april-2014-web-server-survey.html>
 - Gives monthly report on how many web servers are out there
- **Number of pages – numerous estimates**
 - More to follow later in this course
 - For a Web engine: how big its index is

The web: the number of hosts

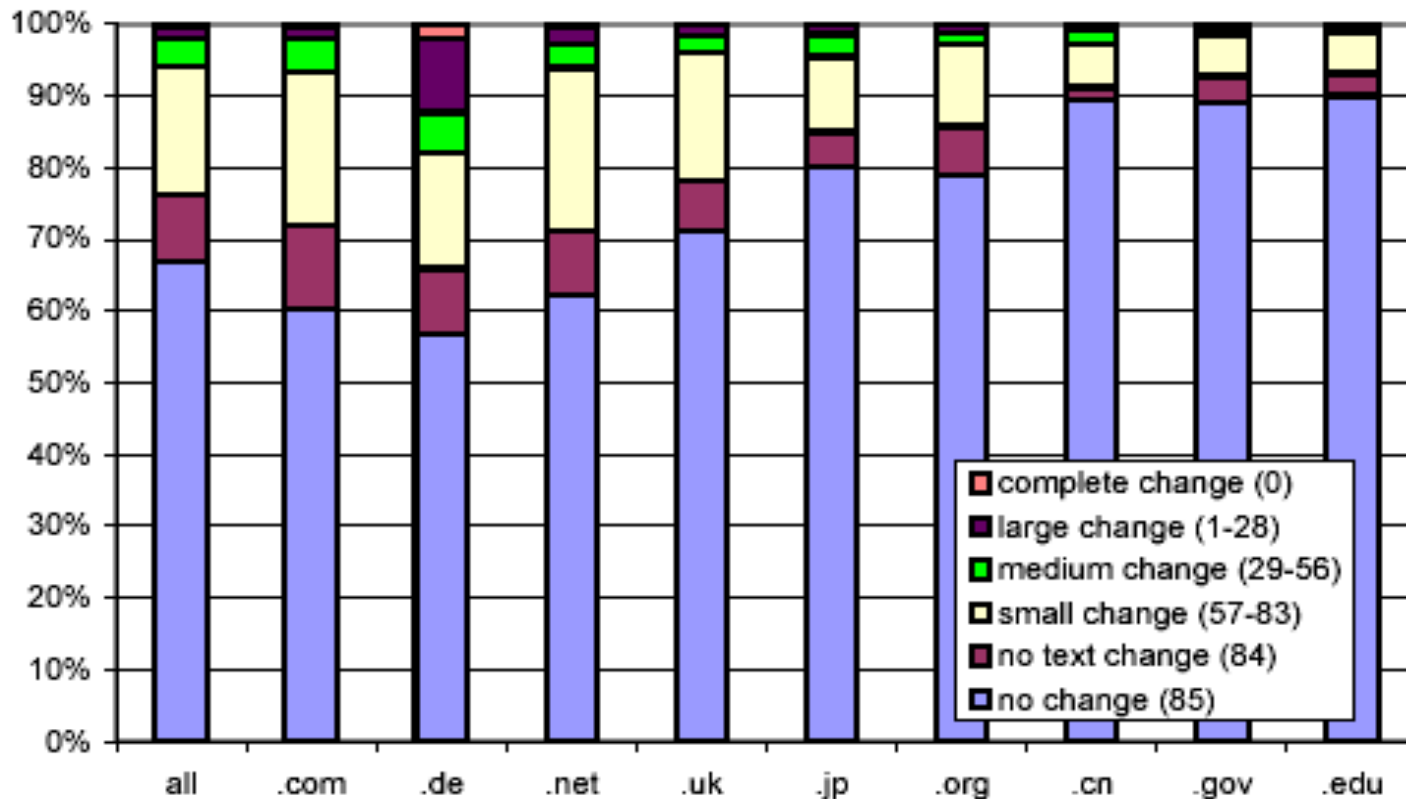


The web: web server vendors



Static pages: rate of change

- Fetterly et al. study: several views of data, 150 million pages over 11 weekly crawls
 - Bucketed into 85 groups by extent of change



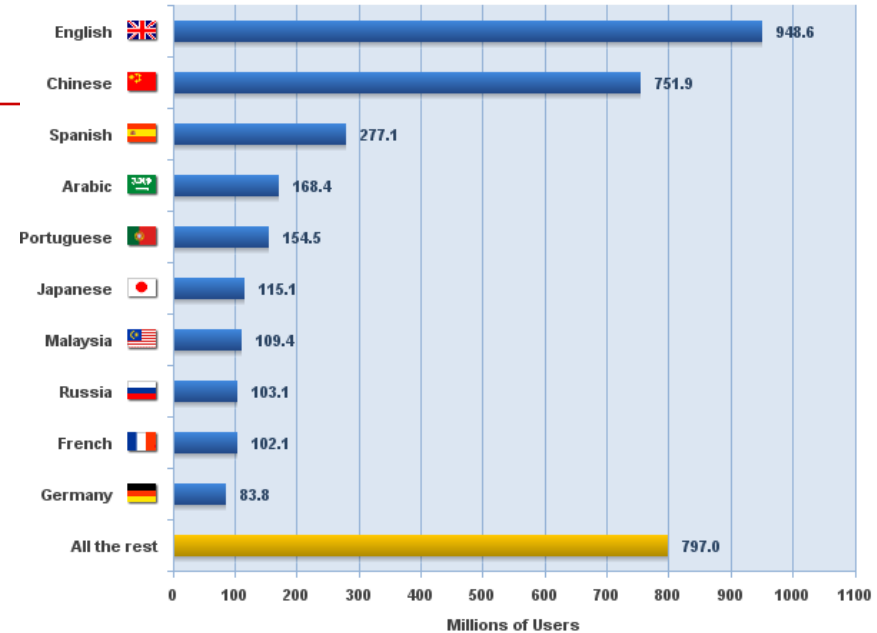
Diversity

- **Languages/Encodings**
 - Hundreds (thousands ?) of languages,
 - W3C encodings
- **Document & query topic**

Table I. Query Stream Breakdown

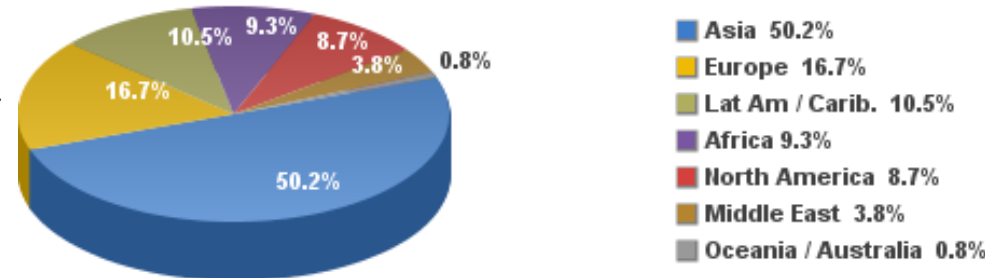
Autos	3.46%	Personal Finance	1.63%
Business	6.07%	Places	6.13%
Computing	5.38%	Porn	7.19%
Entertainment	12.60%	Research	6.77%
Games	2.38%	Shopping	10.21%
Health	5.99%	Sports	3.30%
Holidays	1.63%	Travel	3.09%
Home & Garden	3.82%	URL	6.78%
News & Society	5.85%	Misspellings	6.53%
Orgs.&Insts.	4.46%	Other	15.69%

Top Ten Languages in the Internet
in millions of users - June 2016




Source: Internet World Stats - www.internetworldstats.com/stats7.htm
Estimated total Internet users are 3,611,375,813 for June 30, 2016
Copyright © 2016, Miniwatts Marketing Group

Internet Users in the World by Regions
June 2016



Source: Internet World Stats - www.internetworldstats.com/stats.htm
Basis: 3,675,824,813 Internet users on June 30, 2016
Copyright © 2016, Miniwatts Marketing Group

Table of Content

- **Information Retrieval**
- **Search Engine Architecture and Process**
- **Web Content and Size**
- **Users Behavior in Search** 
- **Sponsored Search: Advertisement**
- **Impact to Business and Search Engine Optimization**
- **Search Engine History/Related Fields**

The user



- **Diverse in access methodology**
 - Increasingly, high bandwidth connectivity
 - Growing segment of mobile users: limitations of form factor – keyboard, display
- **Diverse in search methodology**
 - Search, search + browse, filter by attribute ...
 - Average query length ~ 2.5 terms
- **Poor comprehension of syntax**
 - Early engines surfaced rich syntax – Boolean, phrase, etc.
 - Current engines hide these

Web Search: How do users find content?

- **Informational (~25%)** – want to learn about something
autism
- **Navigational (~40%)** – want to go to that page
United Airlines
- **Transactional (~35%)** – want to do something (web-mediated)
 - Access a service
 - Downloads
Santa barbara weather
 - Shop
Mars surface images
- **Gray areas**
 - Find a good hub
 - Exploratory search “see what’s there”
Car rental Finland

Users' evaluation of engines

- **Relevance and validity of results**
- **UI – Simple, no clutter, error tolerant**
- **Trust – Results are objective, the engine wants to help me**
- **Pre/Post process tools provided**
 - Mitigate user errors (auto spell check)
 - Explicit: Search within results, more like this, refine ...
 - Anticipative: related searches

Users' evaluation

- **Quality of pages varies widely**
 - Relevance is not enough
 - Duplicate elimination
- **Precision vs. recall**
- **What matters**
 - Precision at position 1? Precision above the fold?
 - Comprehensiveness – must be able to deal with obscure queries
 - Recall matters when the number of matches is very small
- **User perceptions may be unscientific, but are significant over a large aggregate**

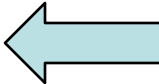
What about on Mobile

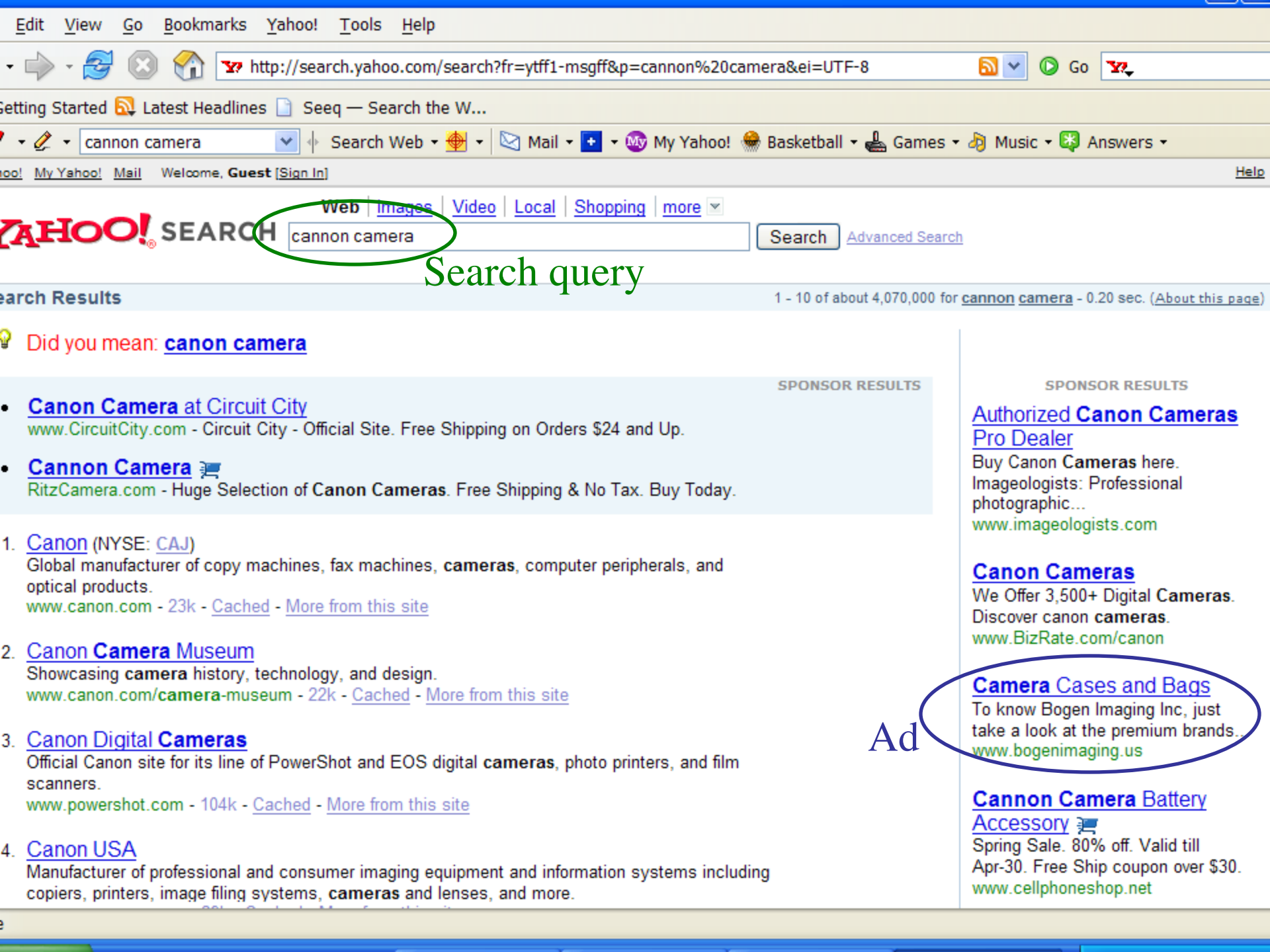
- **Query characteristics:**
 - Best known studies by Kamvar and Baluja (2006 and 2007) and by Yi, Maghoul, and Pedersen (2008)
- **Have a different distribution than the query distribution for PC users**
 - Bias towards shorter queries
 - Data contradicts that: 2.6 words per query, same # chars as PC
 - Difficulty of query entry is a significant hurdle
 - Much higher location-based activity
- **More notification-driven tasks**

Implications and Challenges

- **Task-orientation**
 - Specialized content packaging
 - “Santa Barbara”
- **Locality Inference from queries and from devices**
 - “Dentist”
- **Minimize typing and round-trips: get results, not just links**
 - Less room to display search engine reply page + other accessories
 - Direct answer

Table of Content

- **Information Retrieval**
- **Search Engine Architecture and Process**
- **Web Content and Size**
- **Users Behavior in Search**
- **Sponsored Search: Advertisement** 
- **Impact to Business and Search Engine Optimization**



Search query

Did you mean: canon camera

SPONSOR RESULTS

- Canon Camera at Circuit City
www.CircuitCity.com - Circuit City - Official Site. Free Shipping on Orders \$24 and Up.
- Cannon Camera
RitzCamera.com - Huge Selection of Canon Cameras. Free Shipping & No Tax. Buy Today.

SPONSOR RESULTS

Authorized Canon Cameras Pro Dealer

Buy Canon Cameras here. Imageologists: Professional photographic...
www.imageologists.com

Canon Cameras

We Offer 3,500+ Digital Cameras. Discover canon cameras.
www.BizRate.com/canon

Camera Cases and Bags

To know Bogen Imaging Inc., just take a look at the premium brands...
www.bogenimaging.us

Cannon Camera Battery Accessory

Spring Sale. 80% off. Valid till Apr-30. Free Ship coupon over \$30.
www.cellphoneshop.net

- Canon (NYSE: CAJ)
Global manufacturer of copy machines, fax machines, cameras, computer peripherals, and optical products.
www.canon.com - 23k - Cached - More from this site
- Canon Camera Museum
Showcasing camera history, technology, and design.
www.canon.com/camera-museum - 22k - Cached - More from this site
- Canon Digital Cameras
Official Canon site for its line of PowerShot and EOS digital cameras, photo printers, and film scanners.
www.powershot.com - 104k - Cached - More from this site
- Canon USA
Manufacturer of professional and consumer imaging equipment and information systems including copiers, printers, image filing systems, cameras and lenses, and more.

Ad

Questions

- **Do you think an “average” user, knows the difference between sponsored search links and algorithmic search results?**

How it works

Advertiser



I want to bid \$5 on
canon camera

I want to bid \$2 on
cannon camera

Ad Index

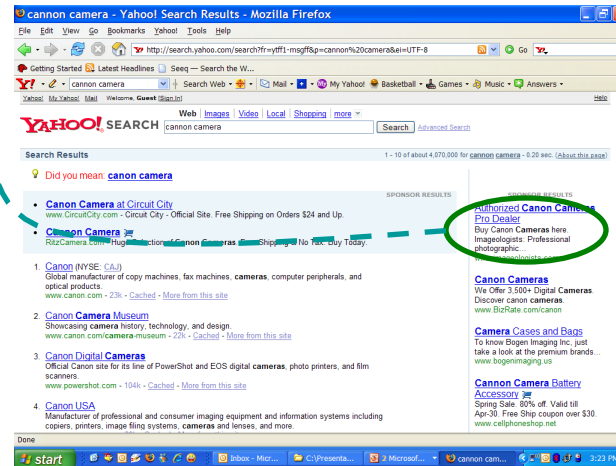
Sponsored
search engine



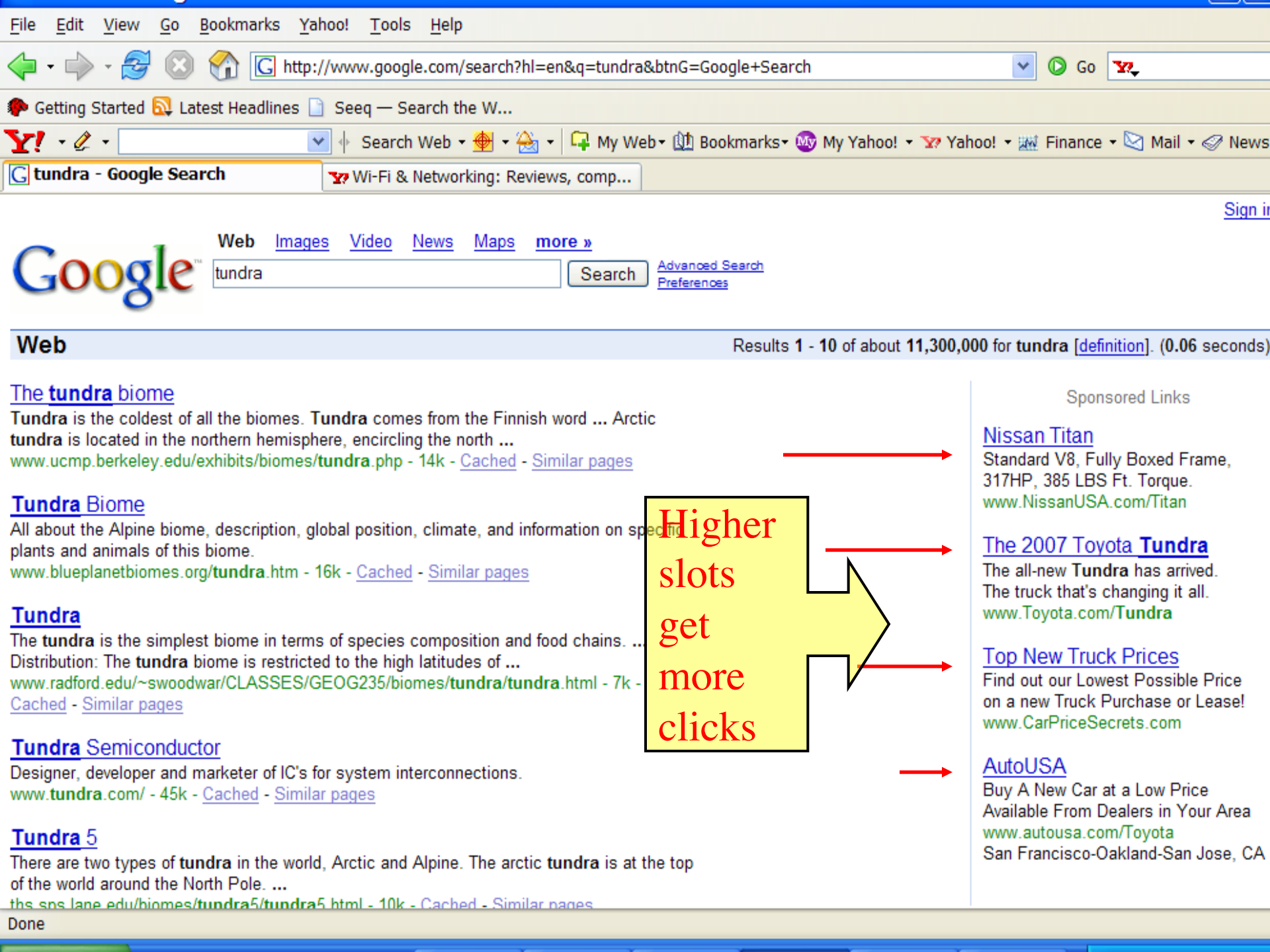
Engine decides when/where to show this ad.



Landing page



Engine decides how much to charge advertiser on a click.



Web Images Video News Maps more »

tundra Search

Advanced Search Preferences

Web Results 1 - 10 of about 11,300,000 for tundra [definition]. (0.06 seconds)

The tundra biome

Tundra is the coldest of all the biomes. Tundra comes from the Finnish word ... Arctic tundra is located in the northern hemisphere, encircling the north ...
www.ucmp.berkeley.edu/exhibits/biomes/tundra.php - 14k - [Cached](#) - [Similar pages](#)

Tundra Biome

All about the Alpine biome, description, global position, climate, and information on specific plants and animals of this biome.
www.blueplanetbiomes.org/tundra.htm - 16k - [Cached](#) - [Similar pages](#)

Tundra

The tundra is the simplest biome in terms of species composition and food chains. ... Distribution: The tundra biome is restricted to the high latitudes of ...
www.radford.edu/~swoodwar/CLASSES/GEOG235/biomes/tundra/tundra.html - 7k - [Cached](#) - [Similar pages](#)

Tundra Semiconductor

Designer, developer and marketer of IC's for system interconnections.
www.tundra.com/ - 45k - [Cached](#) - [Similar pages](#)

Tundra 5

There are two types of tundra in the world, Arctic and Alpine. The arctic tundra is at the top of the world around the North Pole. ...
ths_sns.lane.edu/biomes/tundra5/tundra5.html - 10k - [Cached](#) - [Similar pages](#)

Sponsored Links

[Nissan Titan](#)
Standard V8, Fully Boxed Frame, 317HP, 385 LBS Ft. Torque.
www.NissanUSA.com/Titan

[The 2007 Toyota Tundra](#)
The all-new Tundra has arrived. The truck that's changing it all.
www.Toyota.com/Tundra

[Top New Truck Prices](#)
Find out our Lowest Possible Price on a new Truck Purchase or Lease!
www.CarPriceSecrets.com

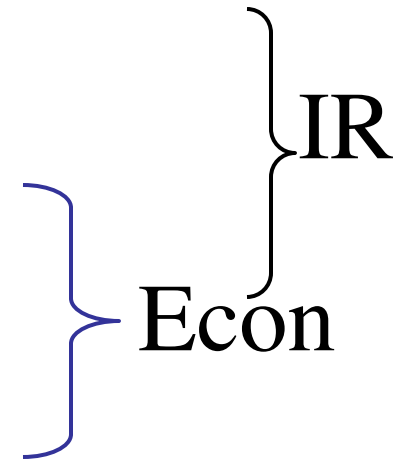
[AutoUSA](#)
Buy A New Car at a Low Price Available From Dealers in Your Area
www.autousa.com/Toyota
San Francisco-Oakland-San Jose, CA

Higher slots get more clicks



Three sub-problems

1. Match ads to query/context
2. Order the ads
3. Pricing on a click-through



t10 - Yahoo! Search Results - Mozilla Firefox

http://search.yahoo.com/search?fr=yttf1-msgff&p=t10&ei=UTF-8

Y! t10 Search Web Mail My Yahoo! Basketball Games Music Answers

Web Images Video Local Shopping more

YAHOO! SEARCH t10 Search Advanced Search

Search Results 1 - 10 of about 5,220,000 for t10 - 0.28 sec. (About this page)

Also try: [sony t10](#), [sony dsc t10](#), [iriver t10](#), [sony cybershot t10](#) More...

T10 Technical Committee
T10 Technical Committee, which standardizes SCSI ... About T10. SCSI-3 Standards Architecture. Recent Documents and Drafts ... T10 Reflector (email list) ...
[www.t10.org](#) - 6k - [Cached](#) - [More from this site](#)

T10 Working Drafts
This page includes the T10 working draft documents. ... T10 committee working draft revision for a project is retained here. The drafts are used by T10 ...
[www.t10.org/drafts.htm](#) - 53k - [Cached](#) - [More from this site](#)

T10 - Wikipedia, the free encyclopedia
T10 may refer to ... T10 Technical Committee. T10 may also refer to: ... T10 systemet. This disambiguation page lists articles associated with the same title. ...
Quick Links: [Disambiguation](#) - [Lists of three-character combinations](#)
[en.wikipedia.org/wiki/T10](#) - 12k - [Cached](#) - [More from this site](#)

Sony Cyber-shot DSC-T10 Digital Camera - Full Review - The Imaging Resource!
... Review of the Sony Cyber-shot DSC-T10 digital camera, with actual sample images, and ... All Prices for Sony DSC-T10. Your shopping clicks support this site, ...
[www.imaging-resource.com/PRODS/T10/T10A.HTM](#) - 47k - [Cached](#) - [More from this site](#)

Pentax Optio T10 Digital Photography Review
Pentax Optio T10 Digital Photography Review News ... Pentax has today also announced the Optio T10, which features a large 3-inch LCD ...

T10
T Series digital cameras. Shop Sony's official store & save.
[www.SonyStyle.com](#)

Sony DscT10 Cyber-shot Digital Camera
His stylish, yet compact camera incorporates 7.2 MP Super Had™ CCD ...
[www.fotocconnection.com](#)

Compare Prices Pentax Optio T10
Find a price on Pentax Optio T10 - at Shopcartusa.com.
[www.ShopCartUSA.com](#)

Plantronics Headsets
Plantronics headsets & accessories. Buy online or call toll free.
[www.TWAc.com](#)

Eroved Diet Pill Review

Table of Content

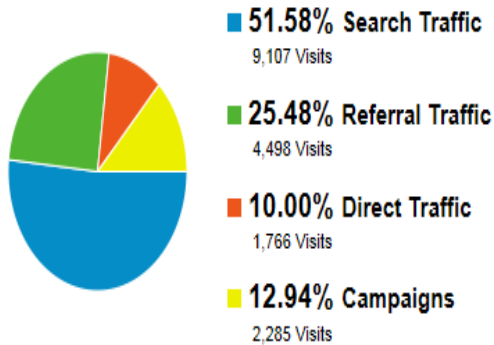
- **Information Retrieval**
- **Search Engine Architecture and Process**
- **Web Content and Size**
- **Users Behavior in Search**
- **Sponsored Search: Advertisement**
- **Impact to Business and Search Engine Optimization**
- **Related Fields**



Search Traffic is Important for Business:

Example of Site Traffic Analysis

17,656 people visited this site



Search Traffic

Keyword

Matched Search Query

Source

Referral Traffic

Source

Direct Traffic

Landing Page

Source	Visits	% Visits
google	8,795	96.57%
bing	106	1.16%
yahoo	96	1.05%
search	38	0.42%
ask	28	0.31%
aol	14	0.15%
avg	9	0.10%
images.google	9	0.10%
search-results	5	0.05%
babylon	3	0.03%

[view full report](#)

Paid placement vs Search Engine Optimization

- **Paid placement costs money. What's the alternative?**
- **Search Engine Optimization:**
 - “Tuning” your web page to rank highly in the search results for select keywords
 - Alternative to paying for placement
 - Thus, intrinsically a marketing function
 - Also known as Search Engine Marketing

Search engine optimization

- **Motives**

- Commercial, political, religious, lobbies
- Promotion funded by advertising budget

- **Operators**

- Contractors (Search Engine Optimizers) for lobbies, companies
- Web masters
- Hosting services

- **Forum**

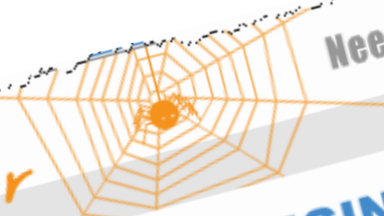
- Web master world (www.webmasterworld.com)
 - Search engine specific tricks
 - Discussions about academic papers ☺
 - More pointers in the Resources

The spam industry

Search Engine Cloaker

Need more search engine listings?

OUTSMART SEARCH ENGINES TO GET MORE HITS



Search Engine Cloaker is used by hundreds of top-ranked Webmasters to increase their search engine listings.



Web Guide

Our hand-picked directory of the best business links on the web.

Cloaking

Category Path

[Home](#) > [Guide Topics](#) > [Technology](#) > [Internet](#) > [Search Technology](#) > [Search Engines](#) > [Search Engine Placement](#) > [Cloaking](#)

Links 1-8 of 8

Day: [To Cloak or Not to Cloak?](#)

at the "Cloaking & Doorways" one of Internet.com's

[News](#) [Best Keywords!](#) [SE](#)

Free Domain Forwarding - Domain Cloaking - DNS Forwarding

Web site is cloaked when the web address of a web site is hidden from viewers in their browser window.

For example your user would type in www.yourname.com into their browser window. They are then automatically redirected to your web site: (<http://www.someisp.com/~users/yourname/yourname.html>) or any where you like. However your users would continue to www.yourname.com as they browsed.



Cloaking Services: Included Branded Email Services 5 Mail boxes mailboxname@yourDomain.com \$49/Year

phantomLine™ — the ultimate stealth



Understanding Cloaking

Tutorial: Cloaking and Stealth Technology

[Page 1](#) | [Page 2](#) | [Page 3](#) | [Page 4](#) | [Page 5](#)

Cloaking, stealth or phantom page technology constitutes the most sophisticated and efficient approach towards search engine optimization. A mystique surrounding cloaking or stealth tech



Simplest forms

- **Early engines relied on the density of terms**
 - The top-ranked pages for the query *maui resort* were the ones containing the most *maui*'s and *resort*'s
- **SEO's responded with dense repetitions of chosen terms**
 - e.g., *maui resort maui resort maui resort*
 - Often, the repetitions would be in the same color as the background of the web page
 - Repeated terms got indexed by crawlers
 - But not visible to humans on browsers

Can't trust the words on a web page, for ranking.

Keyword stuffing

[Home](#) | [Fovissste](#) | [Infonavit](#) | [Contacto](#) | [Stand Plaza Satélite](#) | [Nuestro Equipo](#) | [FAQ](#) | [Links](#) | [Noticias](#) | [Foto Galería](#) | [Eventos](#) | [Casas San Juan del Rio](#) | [Programas](#) | [San Juan del Rio](#) | [Site Map](#) | [Casas San Juan del Rio](#) | [Casas Querétaro](#) | [Inmobiliaria Querétaro](#) | [Casas Tequisquiapan](#) | [Empleos](#) | [Venta Casa San Juan del Rio](#) | [Tríptico](#) | [Links](#) | [Vago Inmobiliaria](#) | [Infonavit casas](#) | [Fovissste](#) | [Cuenta Bancaria](#) | [Casa San Juan Del Rio](#) | [Directorio Links](#)

Copyright © 2008 Viveros de San Juan. San Juan Del Rio Querétaro Todos los Derechos Reservados.

casas san juan del rio, casas san juan del rio, Casas, San Juan del Rio, casas-san-juan-del-rio, vivienda, viveros de san juan, desarrollo, residencial, inmobiliaria, vago inmobiliaria, inmobiliaria vago,inmobiliaria vago san juan del rio, inmobiliaria vago queretaro, venta, san juan del rio, Tequisquiapan, Inmobiliarias san juan del rio, ventas san juan del rio, inmobiliaria santa fe casas nuevas san juan del rio casa san juan del rio, casas san juan del rio,fraccionamiento bosques de san juan, casas bosques de san juan, fraccionamiento las nueces, fraccionamiento las nueces san juan del rio, bosques de san juan san juan del rio, casas venta infonavit san juan del rio, venta casas fovissste san juan del rio, venta casas cofinanciamiento san juan del rio, residencial el encanto, residencial hacienda las nueces, residencial san juan, san juan del rio viviendas, san juan del rio fines de semana, san juan del rio venta de casas, terrenos en venta san juan del rio, los agaves, asesores, infonavit



Invisible text

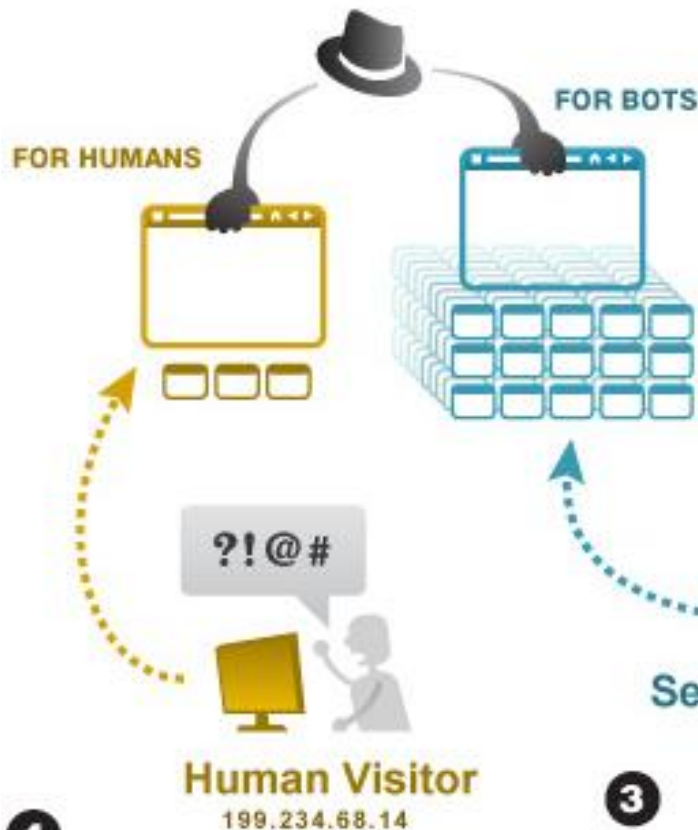
auctions.hitsoffice.com/



Cloaking: Black Hat Cloaking Explained

1

Sites engaged in black hat SEO prepare two sets of content, one targeted for bots and the other targeted for human visitors. Bots are identified by their IP address.



4

Human visitors often won't find the best information despite the site's high rankings.

2

Since bot IP's can change, black hat informants provide a regularly updated list of bot IP addresses.

LIST OF BOT IP'S

209.185.253.170
193.7.255.21
192.41.15.30
192.79.138.41

Search Engine Bot

209.185.253.170

3

Bots are served abundant fabricated content packed with targeted keywords. This false information boosts rankings.

Link Farms

Boost pagerank of a website

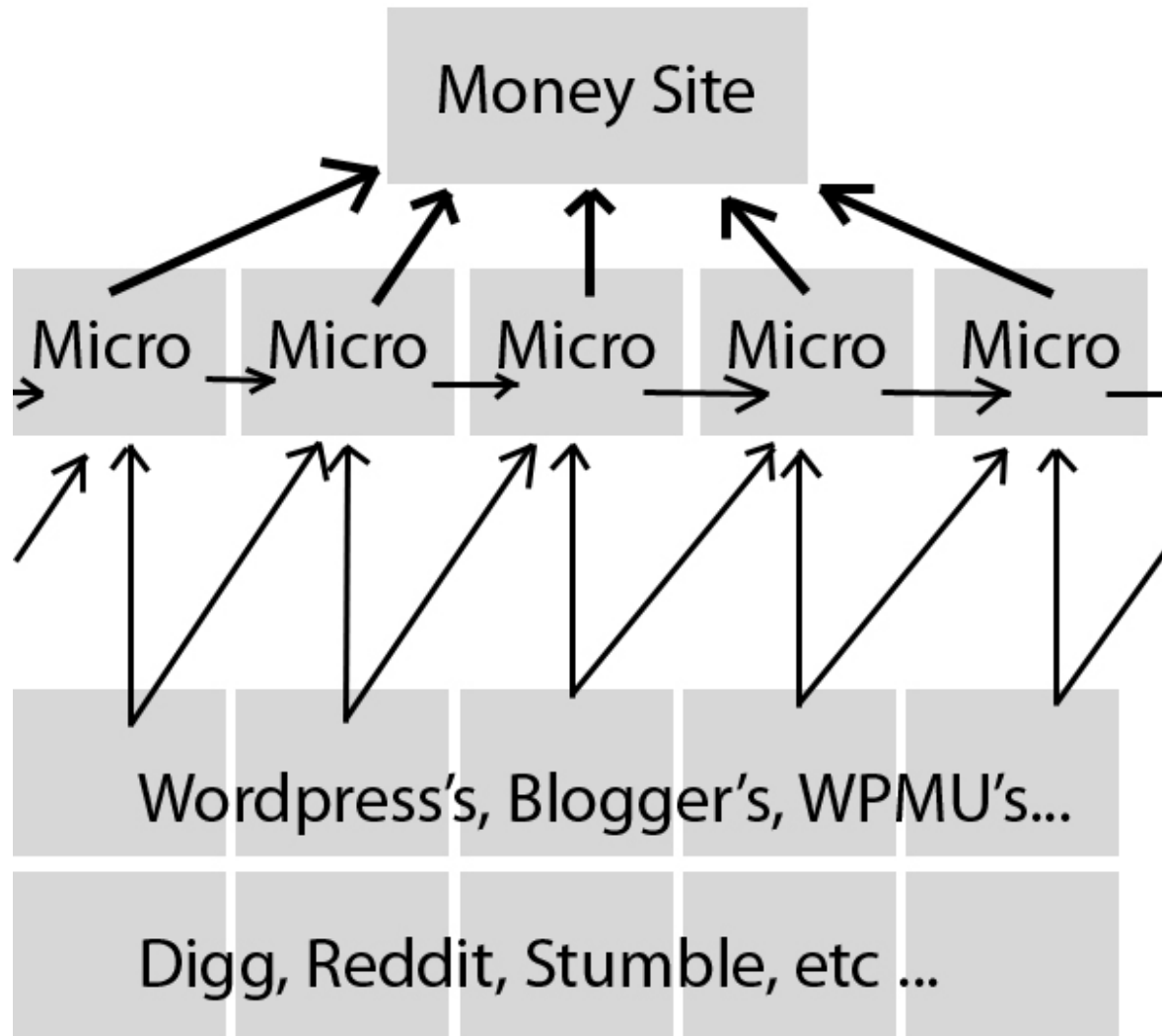



Table of Content

- **Information Retrieval**
- **Search Engine Architecture and Process**
- **Web Content and Size**
- **Users Behavior in Search**
- **Sponsored Search: Advertisement**
- **Impact to Business and Search Engine Optimization**
- **Related Fields** 

From Information Retrieval to Web Search

- **Challenging due to Large-scale and noisy data.**
 - retrieving *relevant* documents to a query.
 - retrieving from *large* sets of documents *efficiently*.
- **Relevance is a subjective judgment and may include:**
 - Simplest notion of relevance is that the query string appears verbatim in the document.
 - More:
 - Being on the proper subject.
 - Being timely (recent information).
 - Being authoritative (from a trusted source).
 - Satisfying the goals of the user and his/her intended use of the information (*information need*).

Related Areas

- **Information Management and Data Mining**
 - Information Science &CHI
 - Machine Learning and data mining
 - Natural Language Processing
- **Large-scale systems**
 - Database/data stores
 - Operating systems/networking support
 - Web language analysis
 - Compression/fast algorithms.
 - Fault tolerance/parallel+distributed systems

Problems with Keywords

- **May not retrieve relevant documents that include synonymous terms.**
 - “car” vs. “automobile”
 - “UCSB” vs. “UC Santa Barbara”
- **May retrieve irrelevant documents that include ambiguous terms.**
 - “bat” (baseball vs. mammal)
 - “Apple” (company vs. fruit)
 - “bit” (unit of data vs. act of eating)

Search Intent Analysis

- Taking into account the *meaning* of the words used.
- Taking into account the *order* of words in the query.
- Adapting to the user based on direct or indirect feedback.
- Taking into account the *authority* of the source.

Topics: Text mining

- **“Text mining” is a cover-all marketing term**
- **A lot of what we’ve already talked about is actually the bread and butter of text mining:**
 - Text classification, clustering, and retrieval
- **But we will focus in on some of the higher-level text applications:**
 - Extracting document metadata
 - Topic tracking and new story detection
 - Cross document entity and event coreference
 - Text summarization
 - Question answering

Topics: Information extraction

- **Getting semantic information out of textual data**
 - Filling the fields of a database record
- **E.g., looking at an event web page:**
 - What is the name of the event?
 - What date/time is it?
 - How much does it cost to attend
- **Other applications: resumes, health data, ...**
- **A limited but practical form of natural language understanding**

Topics: Recommendation systems

- **Using statistics about the past actions of a group to give advice to an individual**
 - E.g., Amazon book suggestions or NetFlix movie suggestions
- **A matrix problem:**
 - but now instead of words and documents, it's users and "documents"