

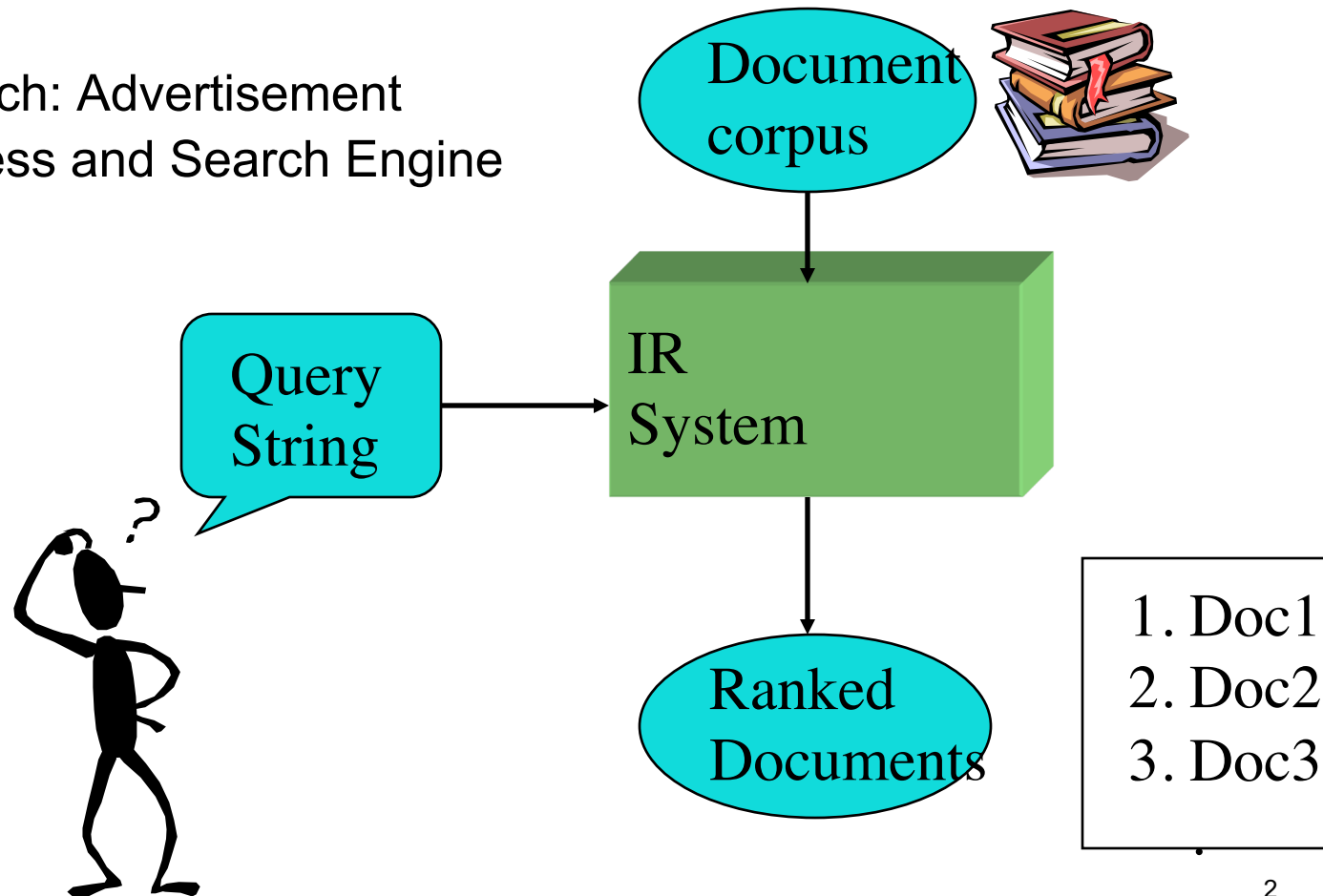
Introduction to Information Retrieval and Web Search

Tao Yang

UCSB CS293S, 2020

Table of Content

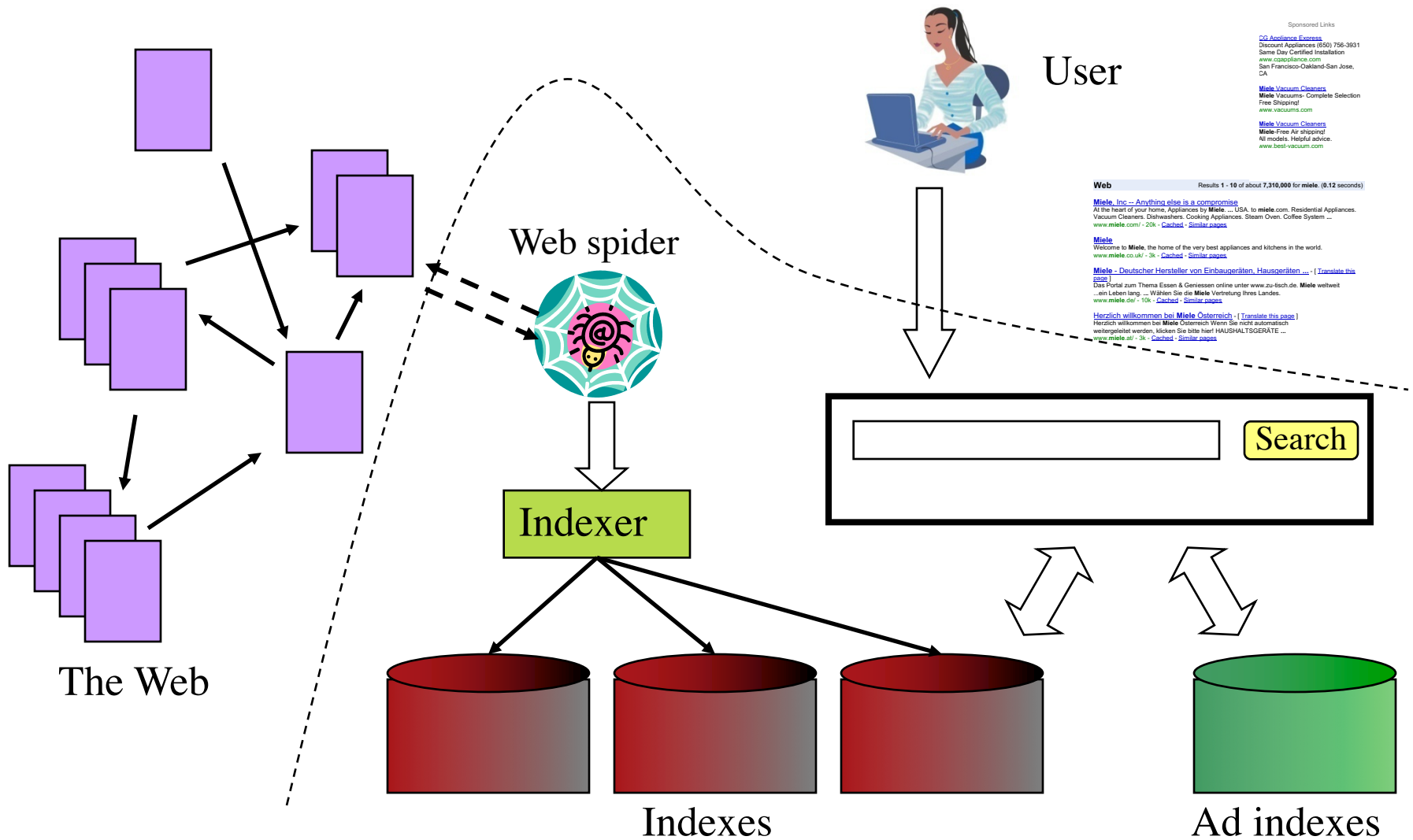
- Information Retrieval& Search Engine Architecture
- Web Content and Size, Users Behavior in Search
- Sponsored Search: Advertisement
- Impact to Business and Search Engine Optimization
- Related fields



History of IR and Web Search

- **1960-70's:**
 - Initial exploration of text retrieval systems for “small” datasets
- **1980's:**
 - Larger document database systems, [Lexis-Nexis](#), [Dialog](#), [MEDLINE](#)
- **1990's:**
 - Searching FTPable documents on the Internet, Archie, WAIS
 - Searching the World Wide Web
 - [Lycos](#), [Yahoo](#), [Altavista](#)
- **2000's**
 - Link analysis for Web Search
 - [Google](#), [Inktomi](#), [Teoma](#)
 - Feedback based engine:
 - [DirectHit](#), [Ask Jeeves](#)
 - Question Answering
 - TREC Q/A track
 - [Ask.com/Ask Jeeves](#)
- **2010-2020**
 - Multimedia IR
 - Cross-Language IR
 - Mobile search
 - Machine-learning/neural network

Web search process

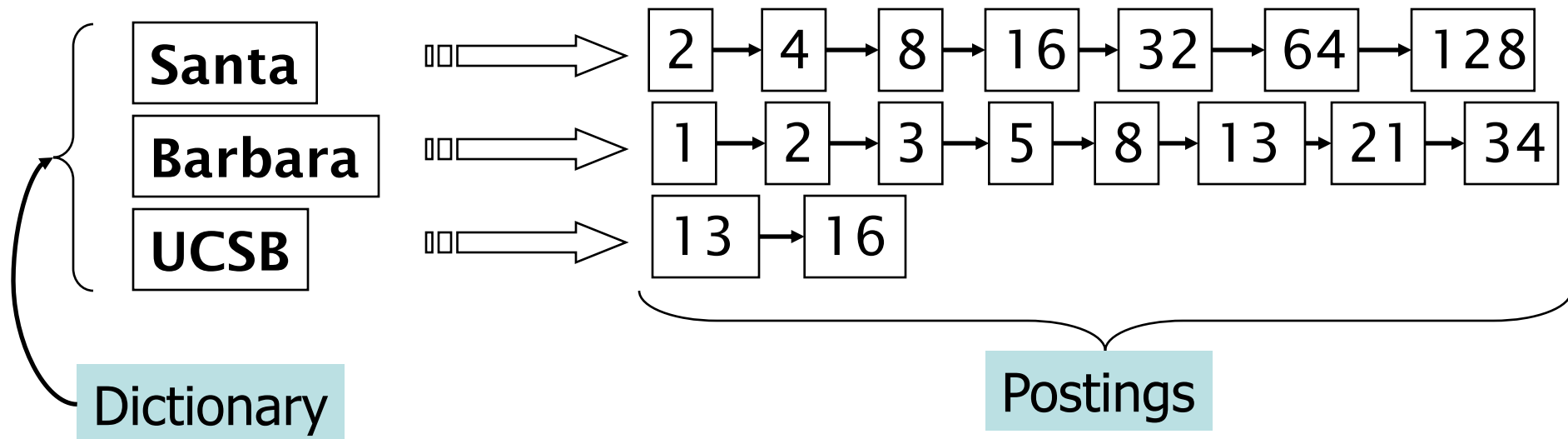


Search engine architecture: key pieces

- **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 with Mining

- **Text acquisition**

- Gather data

- **Text transformation**

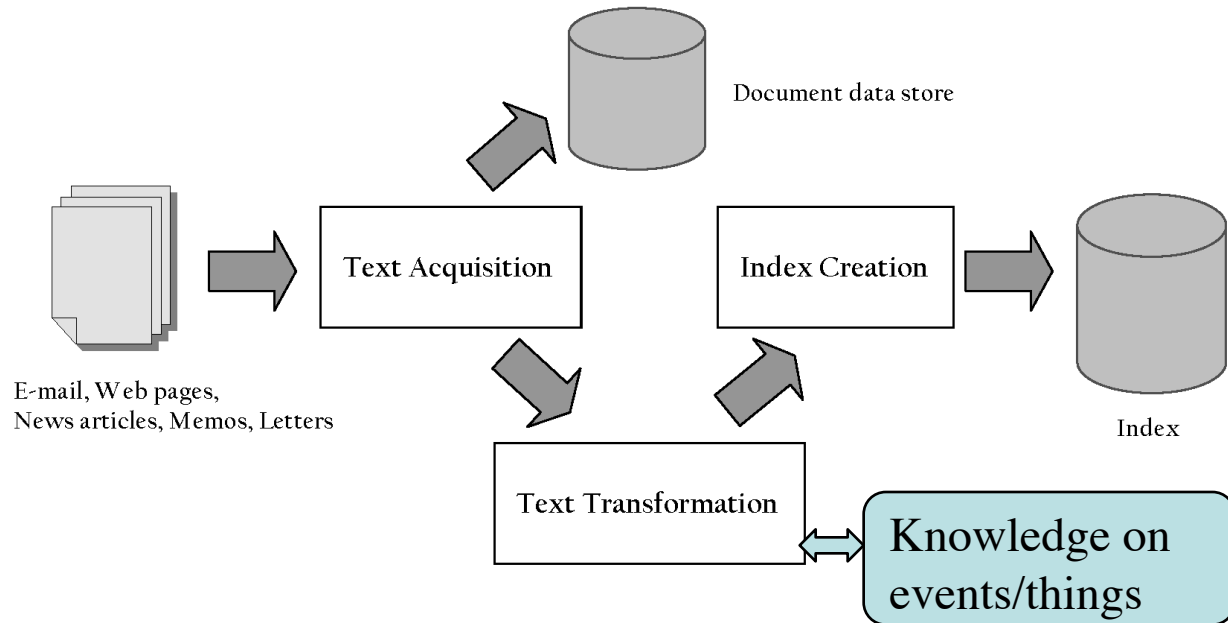
- transforms documents into *index terms* or *features*

- **Index creation**

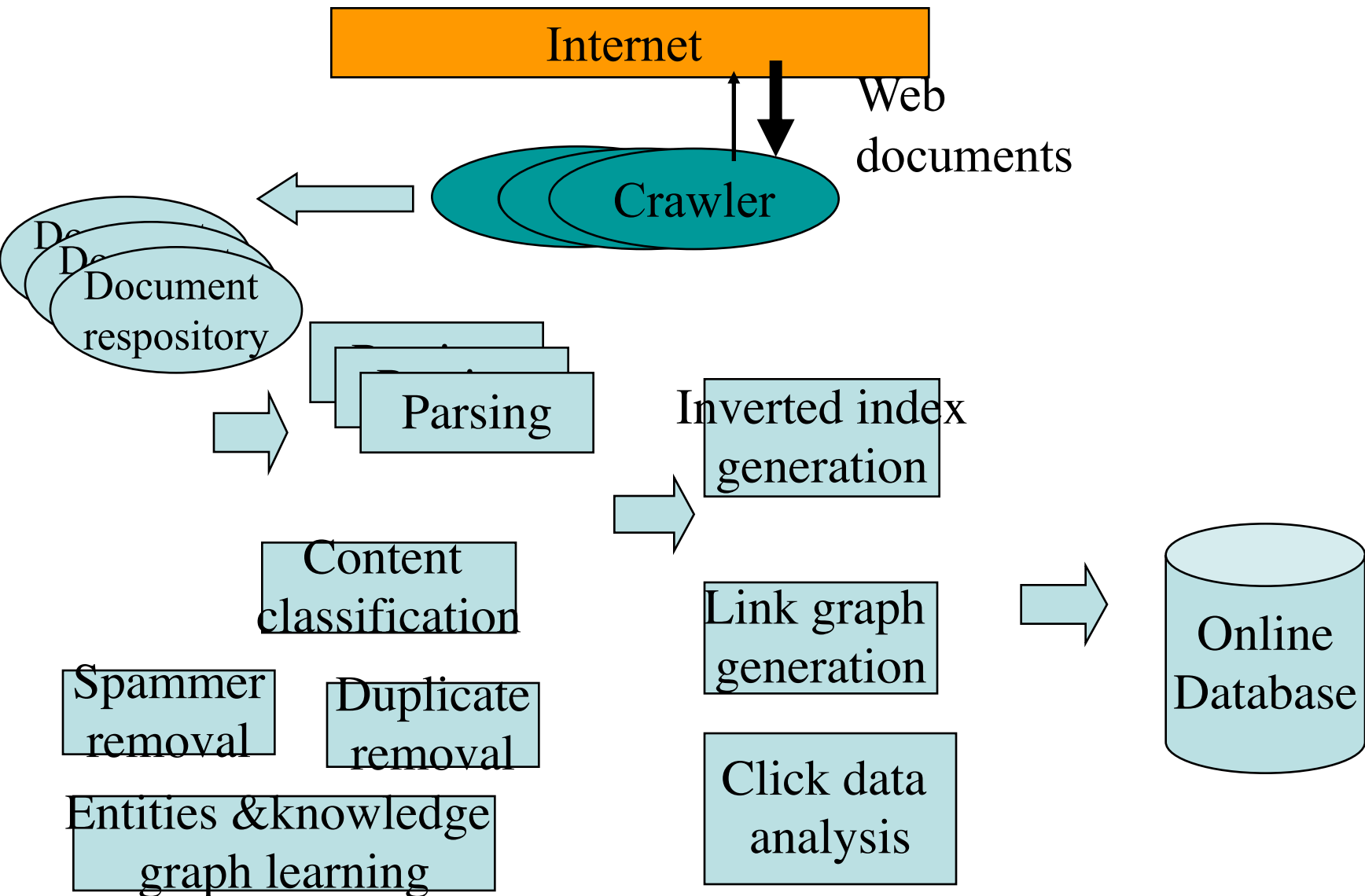
- takes index terms and creates data structures (*indexes*) to support fast searching

- **Data mining**

- Knowledge learning on entities (people name, organization, etc) and their relationship (knowledge graphs)

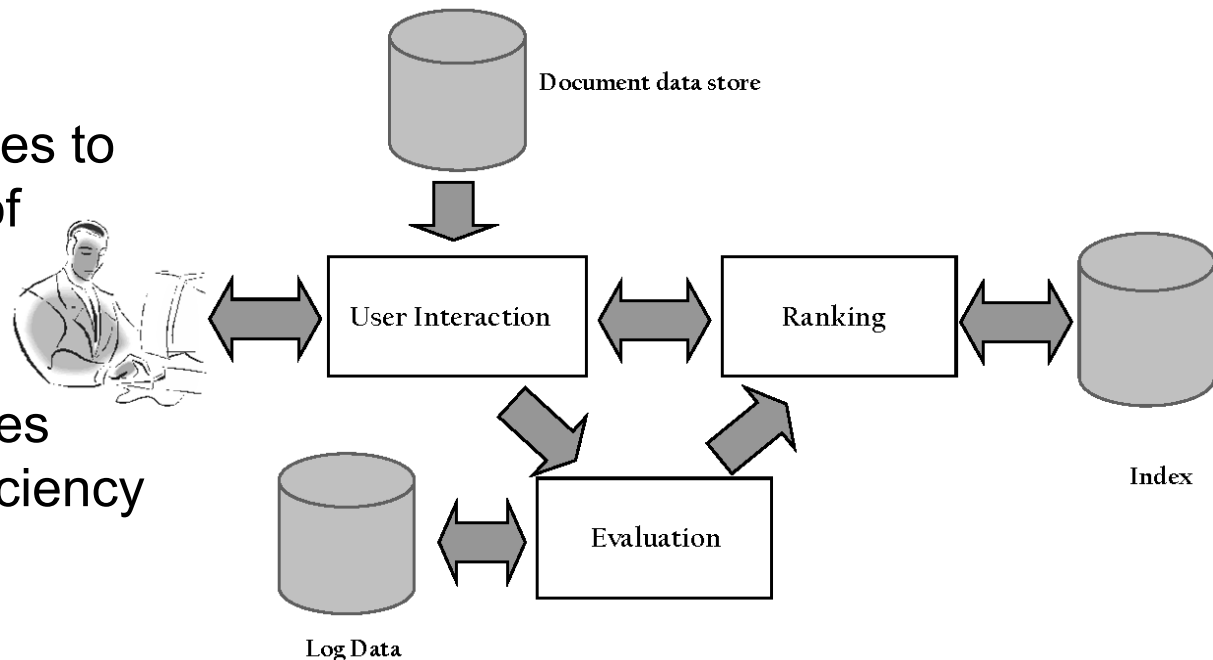


Indexing and Mining at Ask.com

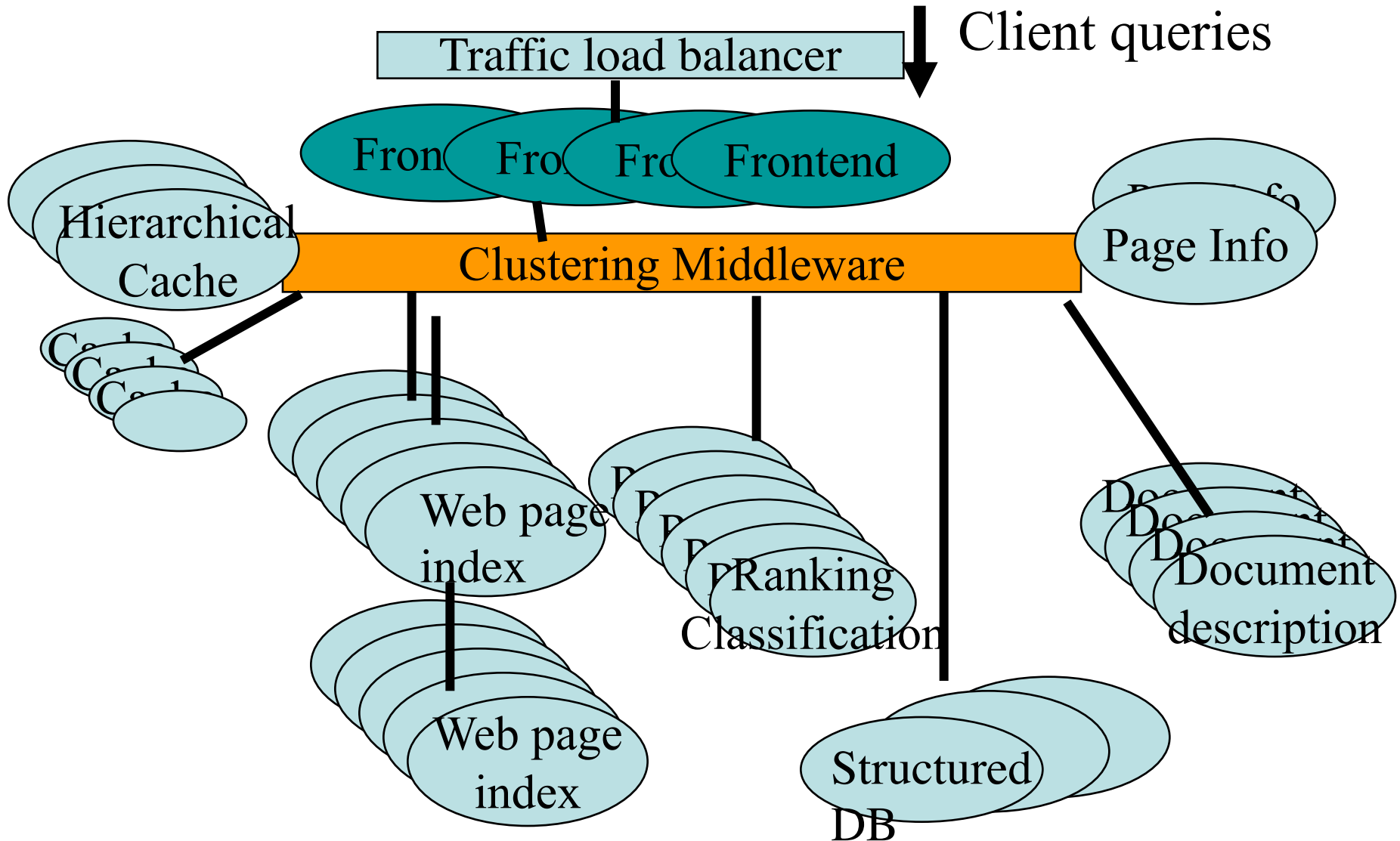


Query Process

- **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)



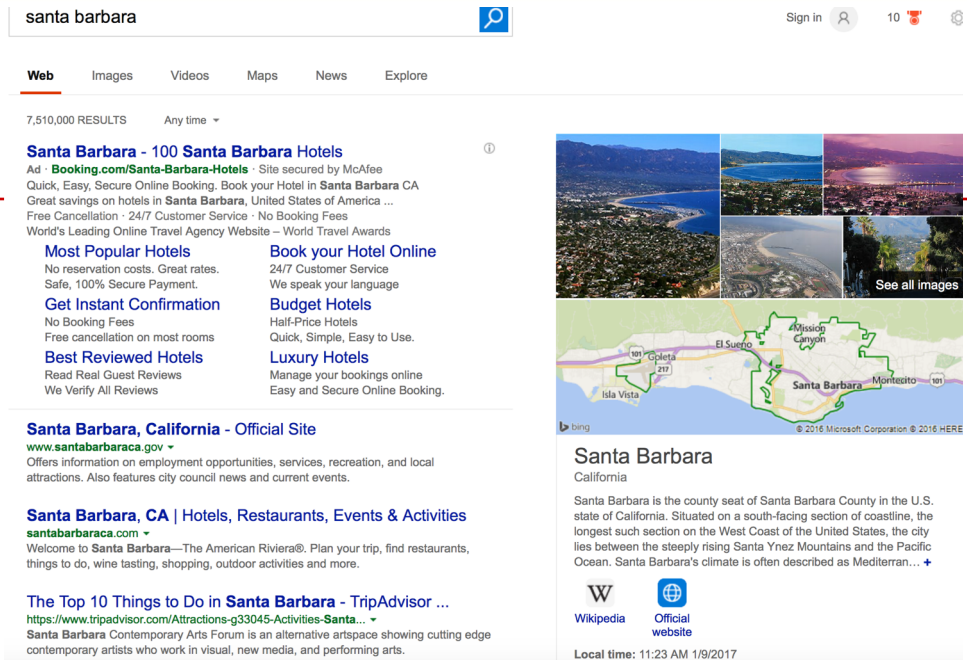
Ask.com Online Engine Architecture



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 Goleta”?
 - *Query transformation or expansion* modifies the original query possibly with additional terms
 - *UC santa babara admission rate*

User Interaction



- **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
 - *Safe vs. unsafe* optimizations
- **Distribution**
 - Processing queries in a distributed environment
 - *Query broker* distributes queries and assembles results
 - *Caching of* intermediate or final results

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.

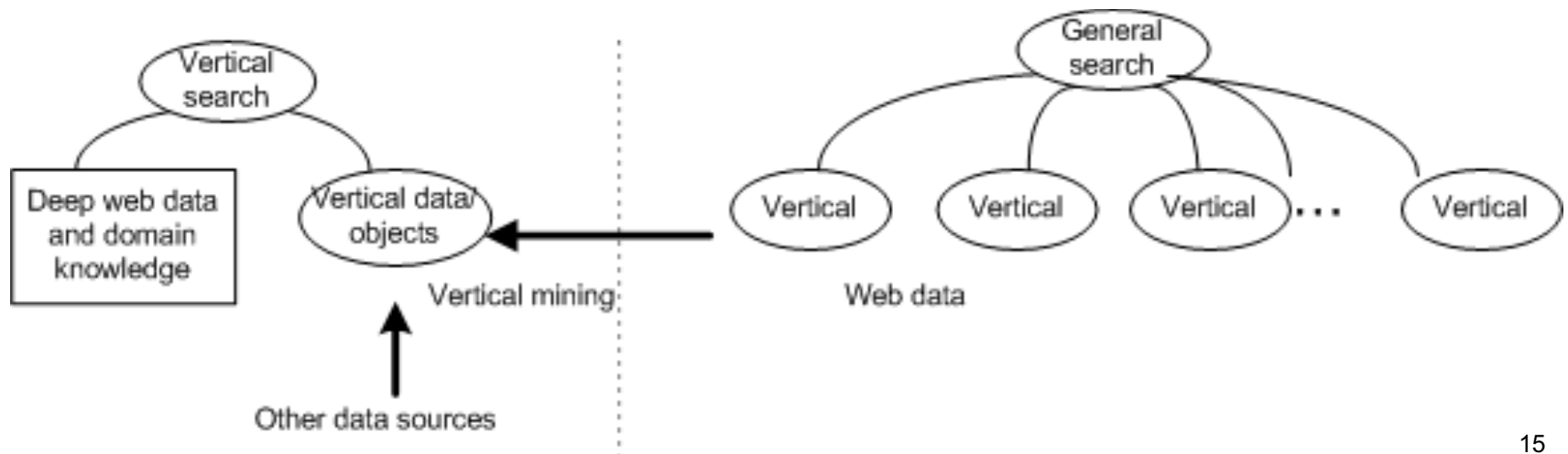
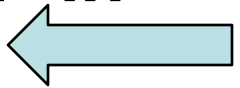
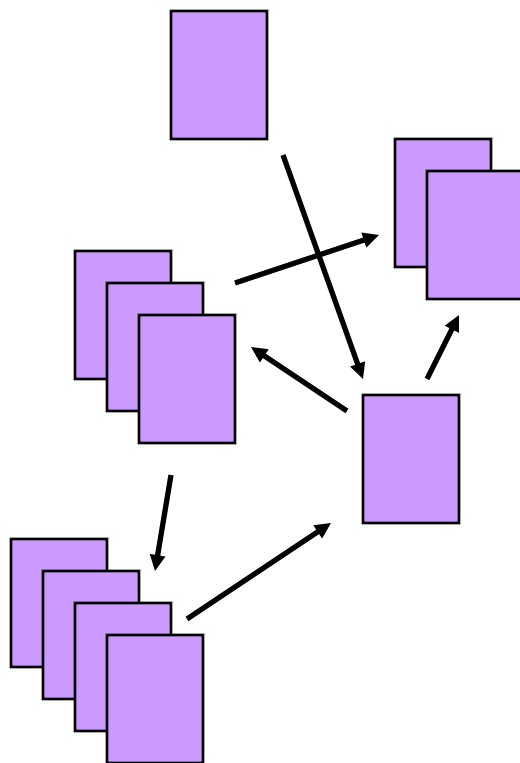


Table of Content

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

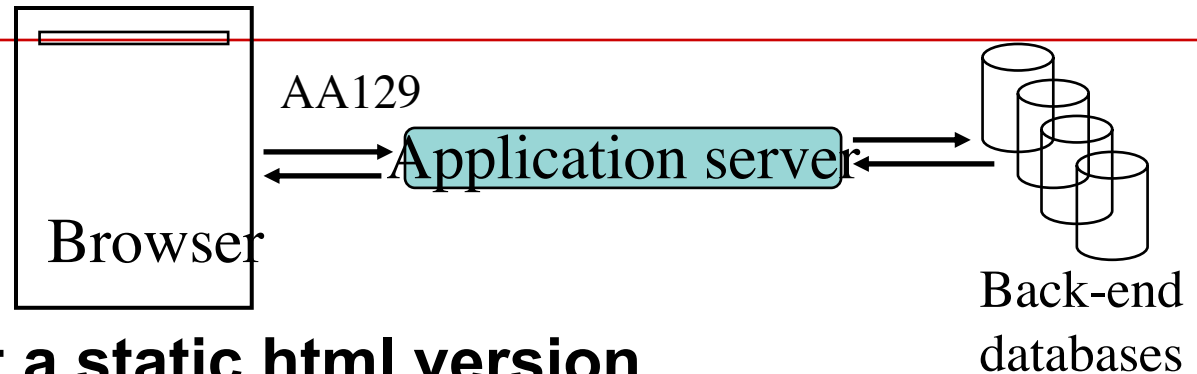
Characteristics of Web Content



The Web

- 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 and rate of changes

- **Number of hosts –**

http://news.netcraft.com/archives/web_server_survey.html

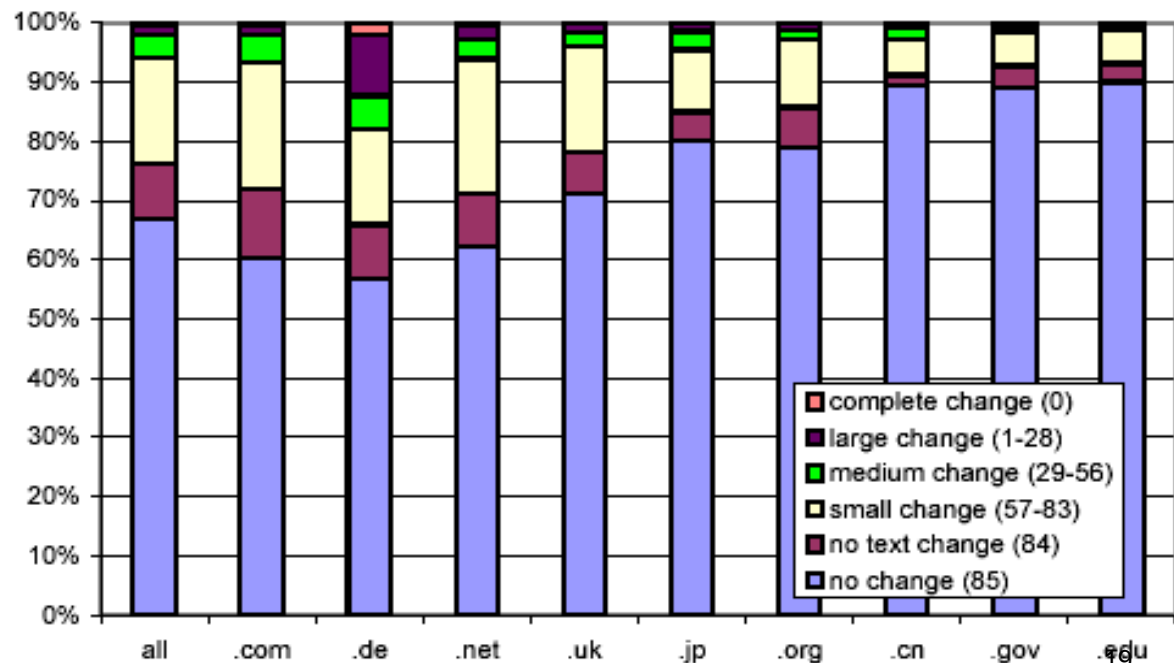
- 1.2 billion web sites. ~200 million active sites

- **Number of pages – numerous estimates**

- Tens/Hundreds of billions

- **Fetterly et al. study: several views of data, 150 million pages over 11 weekly crawls**

- Bucketed into 85 groups by extent of change



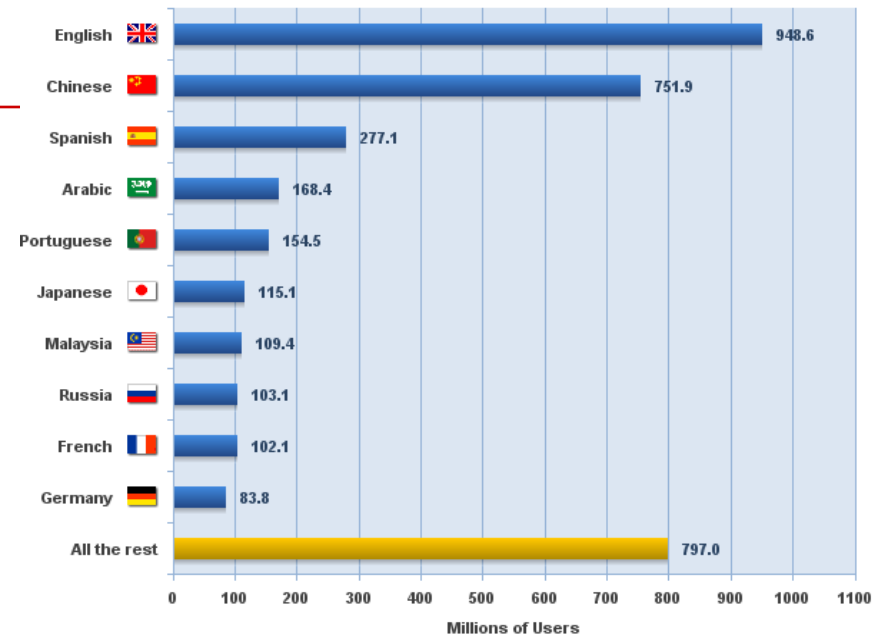
Content Diversity

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

Table I. Query Stream Breakdown

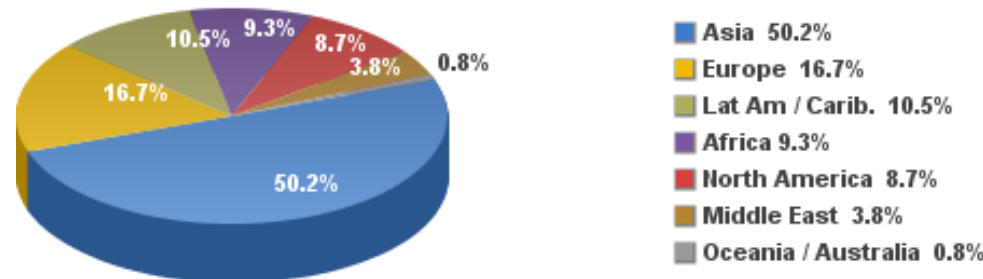
Autos	3.46%	Personal Finance	1.63%
Business	6.07%	Places	6.13%
Computing	5.38%	Porn	7.19%
Entertainment	12.60%	Research	6.77%
Games	2.38%	Shopping	10.21%
Health	5.99%	Sports	3.30%
Holidays	1.63%	Travel	3.09%
Home & Garden	3.82%	URL	6.78%
News & Society	5.85%	Misspellings	6.53%
Orgs.&Insts.	4.46%	Other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

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
- **Mobile users**
 - Bias towards shorter queries
 - Much higher location-based activity through map app

Web Search: How do users find content?

- **Informational (~25%)** – want to learn about something
cancer
- **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”
Nikon D-SLR
Car rental Finland

Users' evaluation of engines

- **Relevance and validity of results. Trust.**
 - Recall matters when the number of matches is very small (e.g. rare queries)
 - Precision at Position 1? Precision above the fold?
 - Relevance is not enough. E.g. duplicate elimination
- **UI – Simple, no clutter, error tolerant**
- **Pre/Post process tools provided**
 - Mitigate user errors (auto spell check)
 - Related searches, Search within results, more like this
- **User perceptions may be unscientific, but are significant over a large aggregate**


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

Search Intent Analysis

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

Table of Content

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

What is percent of users who can differentiate sponsored search links and algorithmic search results?

cannon camera - Yahoo! Search Results - Mozilla Firefox

File Edit View Go Bookmarks Yahoo! Tools Help

http://search.yahoo.com/search?fr=ytf1-msgff&p=cannon%20camera&ei=UTF-8

Getting Started Latest Headlines Seeq — Search the W...

Y! cannon camera Search Web Mail My Yahoo! Basketball Games Music Answers

Yahoo! My Yahoo! Mail Welcome, Guest [Sign In]

YAHOO! SEARCH cannon camera Search Advanced Search

Search Results 1 - 10 of about 4,070,000 for cannon camera - 0.20 sec. (About this page)

Did you mean: [canon camera](#)

Canon Camera at Circuit City
[www.CircuitCity.com](#) - Circuit City - Official Site. Free Shipping on Orders \$24 and Up.

Canon Camera
[RitzCamera.com](#) - Huge Selection of Canon Cameras. Free Shipping & No Tax. Buy Today.

1. **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](#)

2. **Canon Camera Museum**
Showcasing camera history, technology, and design.
[www.canon.com/camera-museum](#) - 22k - [Cached](#) - [More from this site](#)

3. **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](#)

4. **Canon USA**
Manufacturer of professional and consumer imaging equipment and information systems including copiers, printers, image filing systems, cameras and lenses, and more.

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](#)

Canon Camera Battery Accessory
Spring Sale. 80% off. Valid till Apr-30. Free Ship coupon over \$30.
[www.cellphoneshop.net](#)

Higher slots get more clicks

How it works

Advertiser



I want to bid \$5 on
canon camera

I want to bid \$2 on
cannon camera

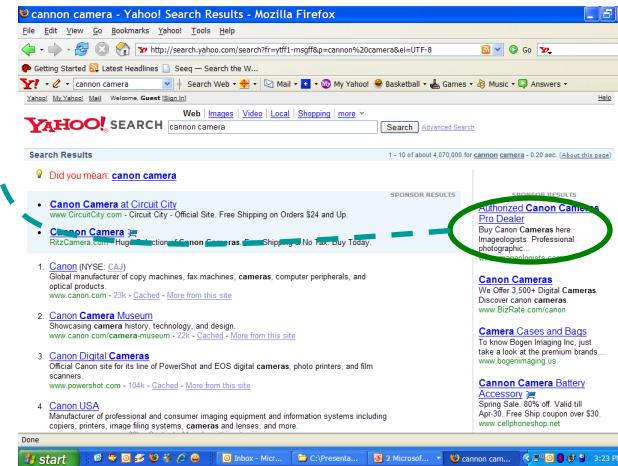


Ad Index

Sponsored
search engine

Engine decides when/where to show this ad.

Landing page

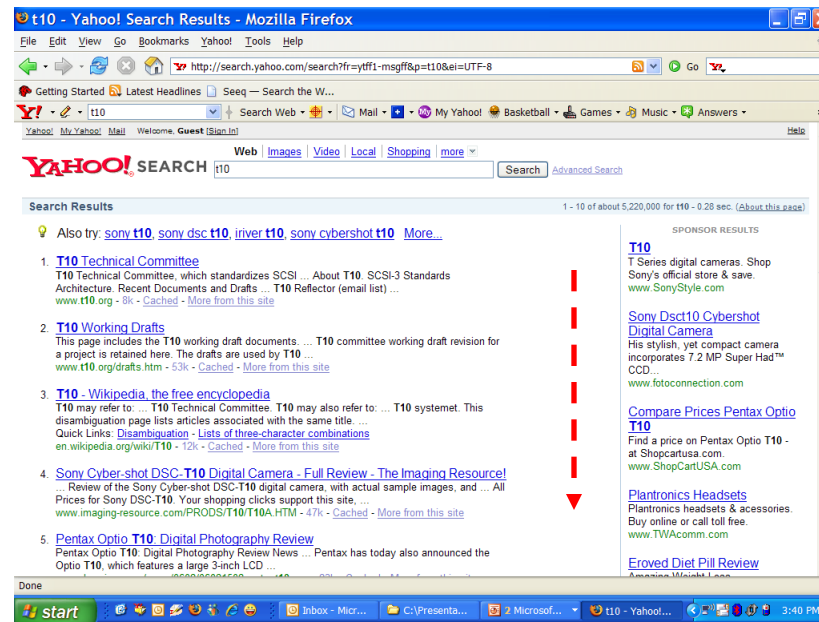


Engine decides how much to charge advertiser on a click.

Three sub-problems

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

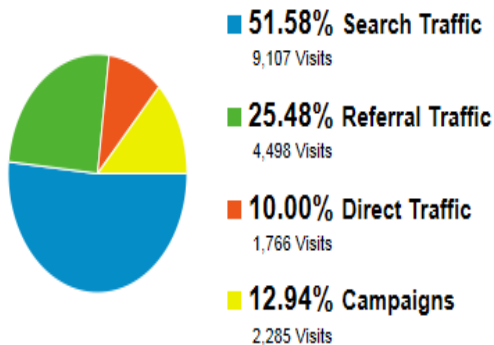
IR
Econ



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 (possibly spamming)
 - 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

SEO Strategies

- **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
- **SEOs 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.

Spamming with Keyword stuffing and Invisible Text

[Home](#) | [Fovissste](#) | [Infonavit](#) | [Contacto](#) | [Stand Plaza Satélite](#) | [Nuestro Equipo](#) | [FAQ](#) | [Links](#) | [Noticias](#) | [Foto Galería](#) | [Eventos](#) | [Casas San Juan del Rio](#) | [Promociones](#) | [San Juan del Rio](#) | [Site Map](#) | [Casas San Juan del Rio](#) | [Casas Querétaro](#) | [Inmobiliarias 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.

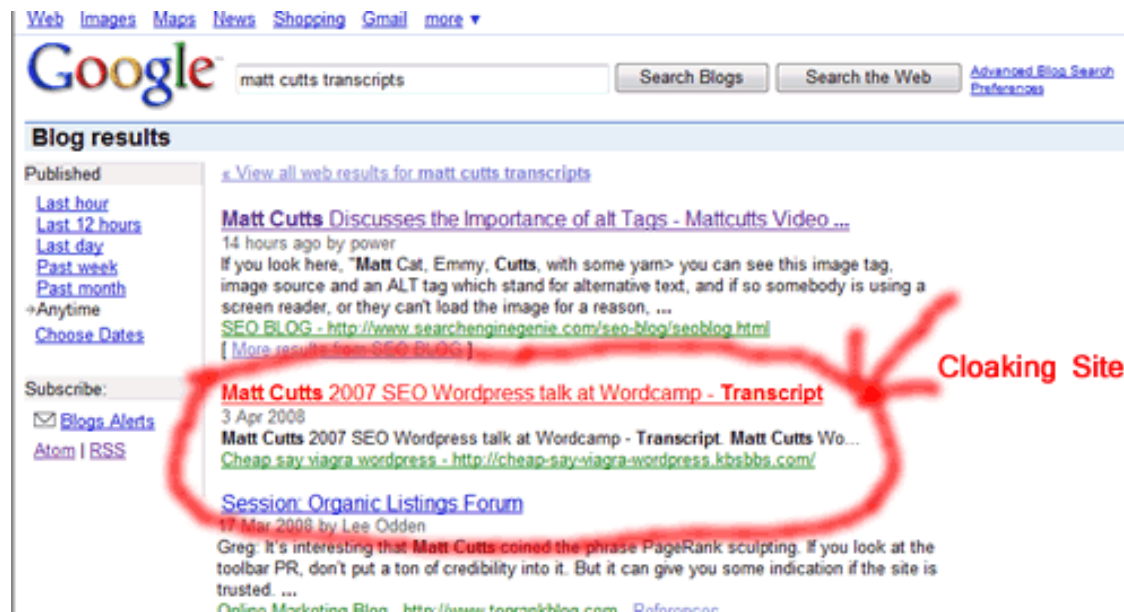
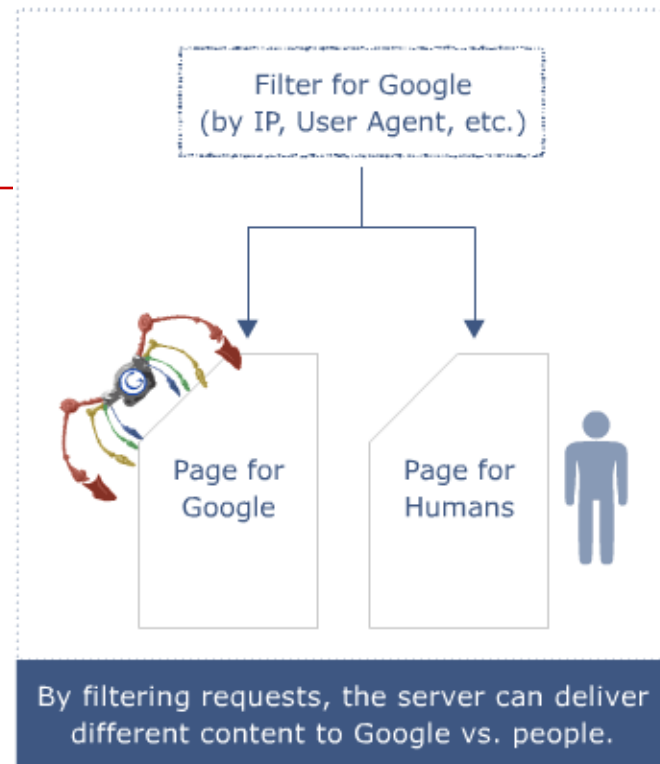
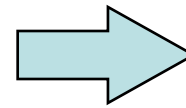
casas san juan del rio, casas san juan del rio, Casas, San Juan del Rio, casas-san-juan-del-rio, vivienda, viveros de san juan, desarrollo, residencial, inmobiliaria, vago inmobiliaria, inmobiliaria vago,inmobiliaria vago san juan del rio, inmobiliaria vago queretaro, venta, san juan del rio, Tequisquiapan, 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



SEO Strategy: Cloaking

Normal behavior: Web server delivers same content to people vs search engine crawlers

Cloaking: Web server delivers different content to people vs search engine crawlers



Cloaking Process:

Black Hat Cloaking Explained

1

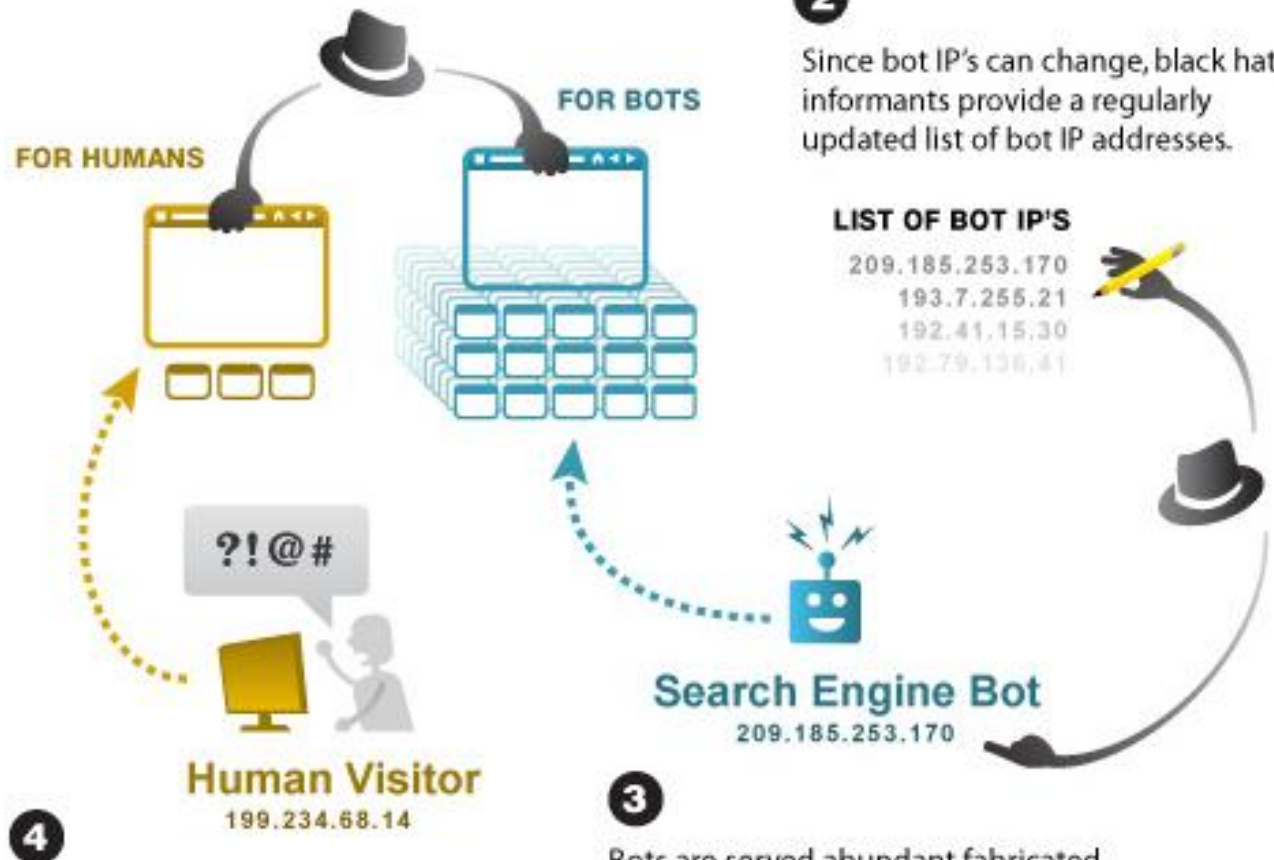
Sites engaged in black hat SEO prepare two sets of content, one targeted for bots and the other targeted for human visitors. Bots are identified by their IP address.

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
192.79.136.41



4

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

3

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

Spamming with Link Farms

Page link support is important ranking feature.

SEO strategy:
Boost pagerank of a website with many artificial links

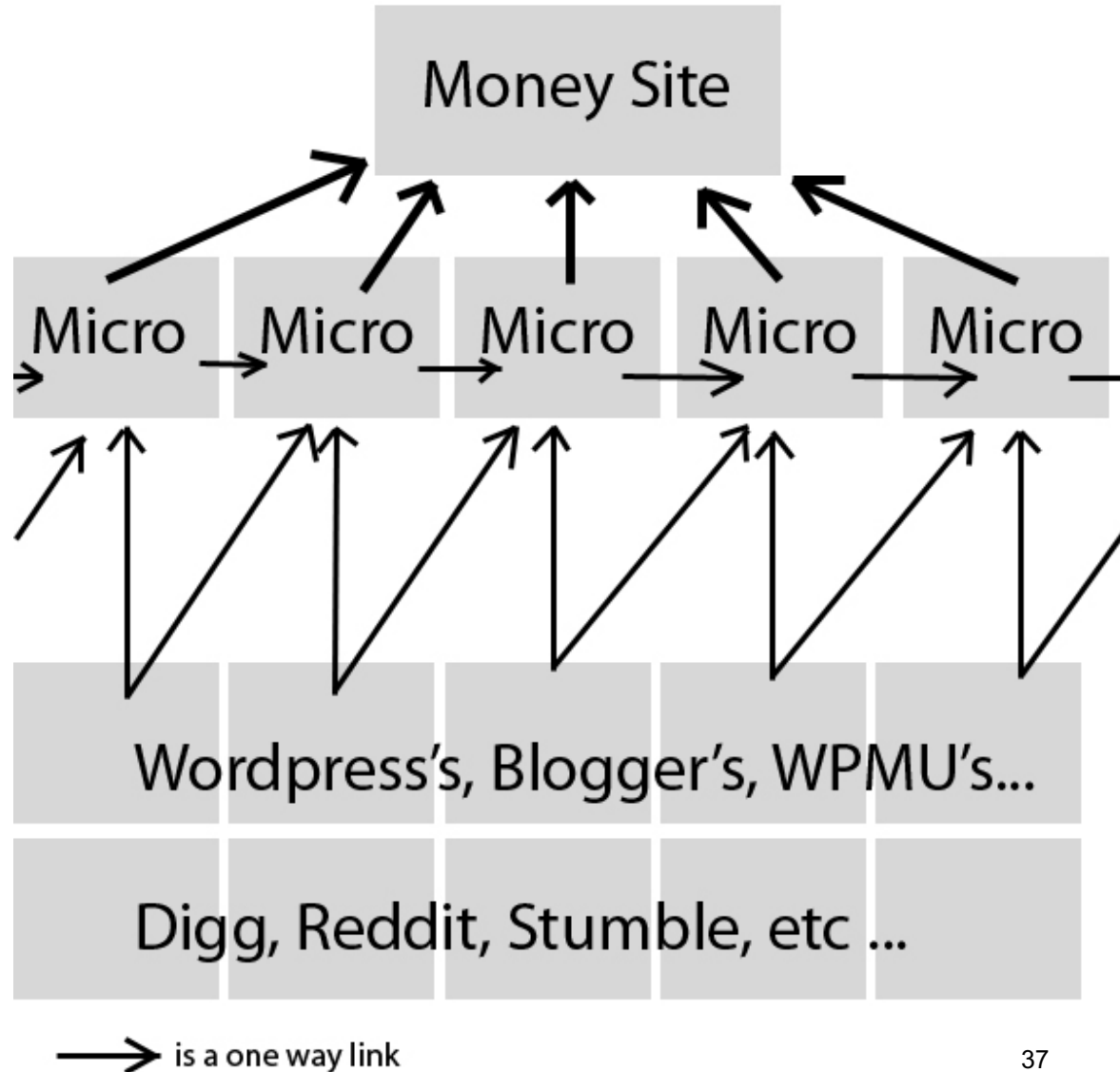



Table of Content

- **Information Retrieval/Search Engine Architecture and Process**
- **Web Content and Size. Users Behavior in Search**
- **Advertisement & Impact to Business**
 - 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
 - Machine Learning and data mining
 - Natural Language Processing
 - Recommendation
 - Using statistics about the past actions of a group to give advice to an individual
- **Large-scale systems**
 - Database/data stores
 - Operating systems/networking support
 - Web language analysis
 - Compression/fast algorithms.
 - Fault tolerance/parallel+distributed systems