

Web Search Basics

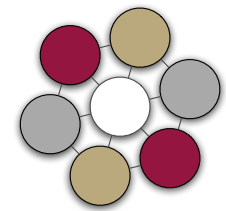
Introduction to Information Retrieval

Content adapted from Hinrich Schütze
<http://www.informationretrieval.org>

Overview



- Introduction
- Classic Information Retrieval
- Web IR



Sponsored Search

- Web Search Basics
 - Size of the Web
- Web Users
- Spam

Sponsored Search

[Web](#) [Blogs](#) [News](#)

Personalized Results 1 - 10 of 10

[search engine optimization.](#)

[Search Engine Optimize](#)

[SEOP.com](#) Guaranteed Top Ranking w/ Warranty. Free Site Analysis! 877-231-1557

[Guaranteed Page 1 Ranking](#)

[www.berankednumber1.com](#) Guaranteed Page 1 Rankings \$49.95 No Charge Until You are on Page 1

[Search engine optimization - Wikipedia, the free encyclopedia](#)

Search engine optimization (SEO) is the process of improving the volume and quality of traffic to a web site from **search** engines via "natural" ("organic" or ...

[en.wikipedia.org/wiki/Search_engine_optimization](#) - 87k - [Cached](#) - [Similar pages](#) - [Note this](#)

[Search Engine Optimization, Google Optimization - SEO Chat](#)

Search Engine Optimization, Google Optimization - SEO Chat.

[www.seochat.com/](#) - 111k - [Cached](#) - [Similar pages](#) - [Note this](#)

[Search Engine Optimization \(SEO\) Marketing Firm & Placement Company](#)

Offers **search engine optimization (SEO)** marketing services & placement since 1998.

Submit your website URL to 40 major **search** engines for FREE!

[www.submitexpress.com/](#) - 42k - [Cached](#) - [Similar pages](#) - [Note this](#)

[News results for search engine optimization](#)



[CIBER Selected as E-Commerce Vendor by Elite Island Resorts](#) - Jan 3, 2008

Their **search engine** marketing program will help us lower acquisition costs ... CIBER's advanced **search engine** marketing services will help Elite direct more ...

[FOX News](#) - [10 related articles](#) »

[bruceclay.com - Search Engine Optimization - SEO Training, Tools ...](#)

Search Engine Optimization, ranking, placement, and submission tutorial. Free step-by-step **SEO** tools and advice. **SEO** training and services offered. ...

[www.bruceclay.com/web_rank.htm](#) - 87k - [Cached](#) - [Similar pages](#) - [Note this](#)

[Inteliture™ Search Engine Optimization, Internet Marketing, and ...](#)

Inteliture™ a professional **search engine optimization** and internet marketing company.

Offers internet marketing solutions, **search engine optimization** ...

[www.inteliture.com/](#) - 12k - [Cached](#) - [Similar pages](#) - [Note this](#)

Ads

Ads

Algorithmic Results

[Search engine optimizer](#)

Use Network Solutions online tools to drive business to your web site.
[marketing.networksolutions.com](#)

[Search Optimization Firm](#)

Looking for top rankings? Get real results. Receive a free analysis.
[customermagnetism.com](#)

[iClimber Company](#)

Search Engine Optimization services since 1998 with proven results.
[www.iClimber.com](#)

[Get Optimization Help Now](#)

Top SEO Firms Want Your Business. Fast, Free Competitive Quotes!
[www.TopSeos.com/SEO](#)

[Check your SEO for Free](#)

PPC vs Natural **search** Keyword ranks costs & robot stats: 15 days free
[www.ClickTracks.com/15_Days_Free](#)

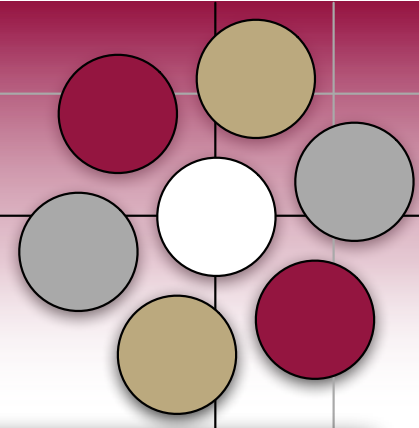
[Search Engine Marketing](#)

Boost Online Traffic and Sales! Free Site **Optimization** Analysis.
[www.corporatesearchoptimization.com](#)

[Free Website Visitors](#)

Free Visitors Plus Top 10 Positions In 8 Hours! FREE Trial Offer.
[www.EngineSeeker.com](#)

Sponsored Search



Ads vs. Search Results

- Google maintains that ads (based on vendors bidding for search queries) do not affect vendors ranking in search results

Sponsored Links

[Search engine optimizer](#)

Use Network Solutions online tools to drive business to your web site.
marketing.networksolutions.com

[Search Optimization Firm](#)

Looking for top rankings? Get real results. Receive a free analysis.
www.customermagnetism.com

[SEO Company](#)

Search Engine Optimization services since 1998 with proven results.
www.iClimber.com

[Search engine optimization - Wikipedia, the free encyclopedia](#)

Search engine optimization (SEO) is the process of improving the volume and quality of traffic to a web site from **search** engines via "natural" ("organic" or ...

en.wikipedia.org/wiki/Search_engine_optimization - 87k - [Cached](#) - [Similar pages](#) - [Note this](#)

[Search Engine Optimization, Google Optimization - SEO Chat](#)

Search Engine Optimization, Google Optimization - SEO Chat.

www.seochat.com/ - 111k - [Cached](#) - [Similar pages](#) - [Note this](#)

[Search Engine Optimization \(SEO\) Marketing Firm & Placement Company](#)

Offers **search engine optimization (SEO)** marketing services & placement since 1998.

Submit your website URL to 40 major **search** engines for FREE!

www.submitexpress.com/ - 42k - [Cached](#) - [Similar pages](#) - [Note this](#)

[News results for search engine optimization](#)

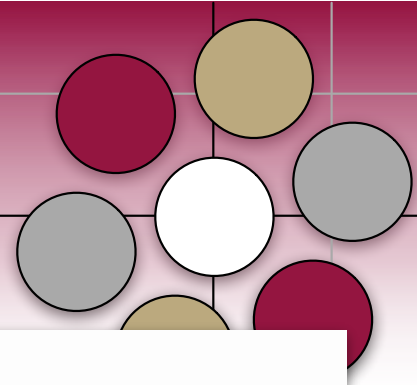


[CIBER Selected as E-Commerce Vendor by Elite Island Resorts](#) - Jan 3, 2008

Their **search engine** marketing program will help us lower acquisition costs ... CIBER's advanced **search engine** marketing services will help Elite direct more ...

[FOX News](#) - [10 related articles »](#)

Sponsored Search



A screenshot of a Facebook profile page for a user named 'Thi'. The page layout includes a left sidebar with navigation links (Search, Applications, Photos, Groups, Events, Marketplace, My Account, SuperFaker), a top navigation bar (Profile, Friends, Networks, Inbox), and a main content area. The main content area shows a profile picture, a cover photo, and a list of activities. Two specific activities are highlighted with red boxes and arrows. The first activity, under the 'Today' section, states: 'Thi added Sony Professional HVR-Z1U 3CCD High Definition to her wishlist on Expo TV. 10:52am'. The second activity, under the 'Yesterday' section, states: 'Thi reviewed KitchenAid KSM150PSER Artisan Series 5-Quart Mixer, on Expo TV. 11:05pm'. Red arrows point from these highlighted activities to a sidebar advertisement on the left. The advertisement is for a 'FREE COACH PURSE!' and features two images of Coach handbags. The sidebar also lists 'Friends' and 'Networks with the most friends'.



Ranking of ads

- Goto model:
 - Rank according to how much advertiser pays
- Current model:
 - Balance auction price and relevance
 - Irrelevant ads (few click-throughs)
 - Decrease opportunities for relevant ads
 - Harm the user experience
 - Idea: Well-targeted advertising is good for everyone

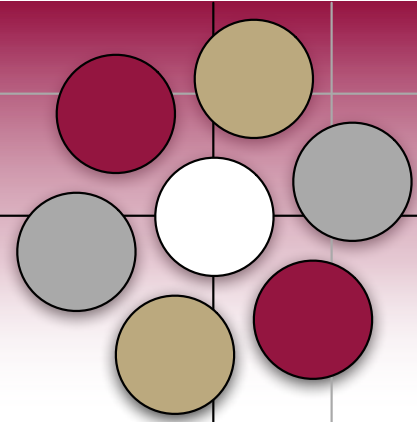
Sponsored Search



Paying for advertisements - terms

- CPM
 - “Cost Per Mil”
 - Pay for 1000 eyeballs
 - Important for branding campaigns
- CPC
 - “Cost per Click”
 - Pay for clicking on ads
 - Important for sales campaigns

Sponsored Search



What are the stakes here?

- What role is Google playing?

Google Dreams of a World After Apps but It's a Nightmare for Rivals



It's a hazy world after apps.

/ MOBILE

Shutterstock



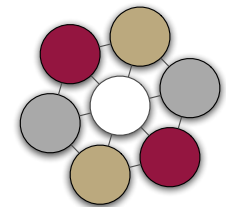
By Mark Bergen | [@mhbergen](#) | EMAIL | ETHICS

September 3, 2015, 11:37 AM PDT

Overview

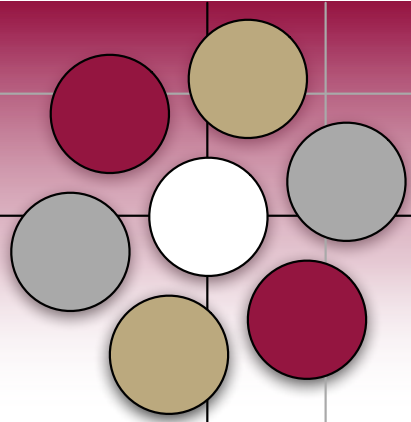


- Introduction
- Classic Information Retrieval
- Web IR
- Sponsored Search



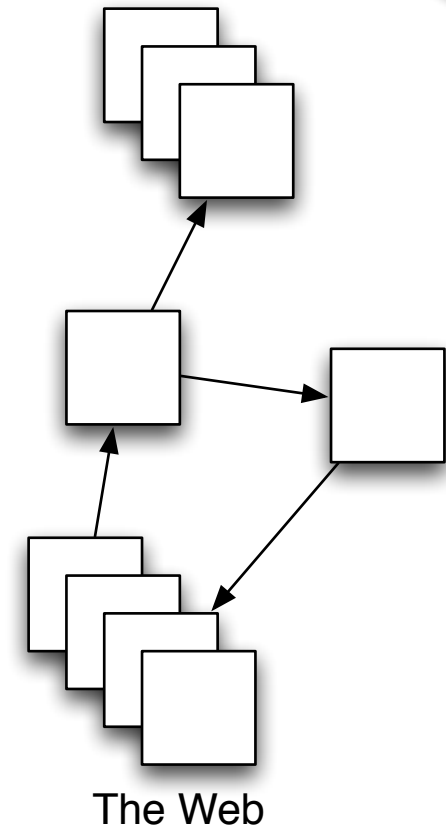
Web Search Basics

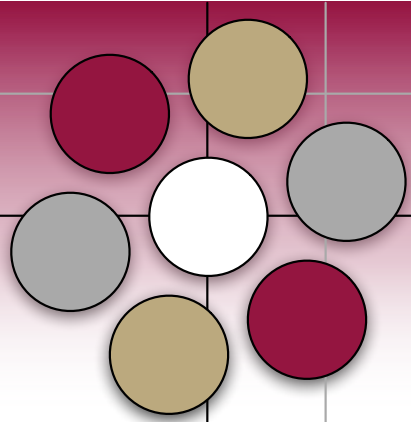
- Size of the Web
- Web Users
- Spam



The Web Corpus

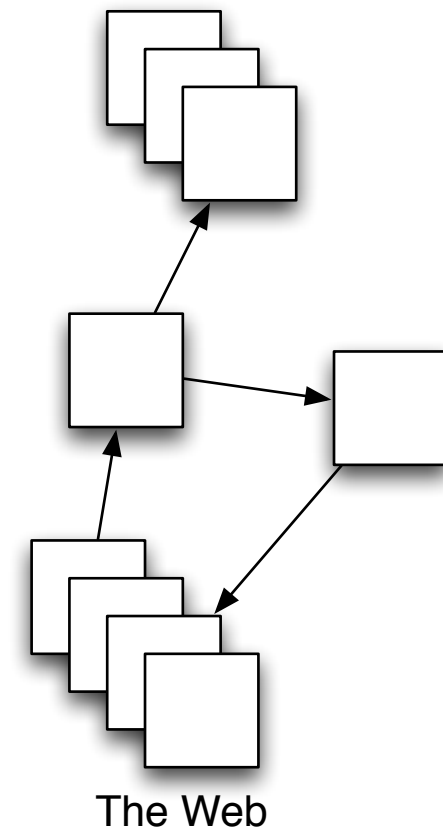
- No design/coordination
- Distributed content creation, linking
- “Democratization of publishing”
- Content includes truth, lies, contradictions, etc.
- Unstructured Data (text, html)
- Semi-Structured (XML, annotated photos)
- Structured (Databases)
- Scale is much larger than previous text corpora





The Web Corpus

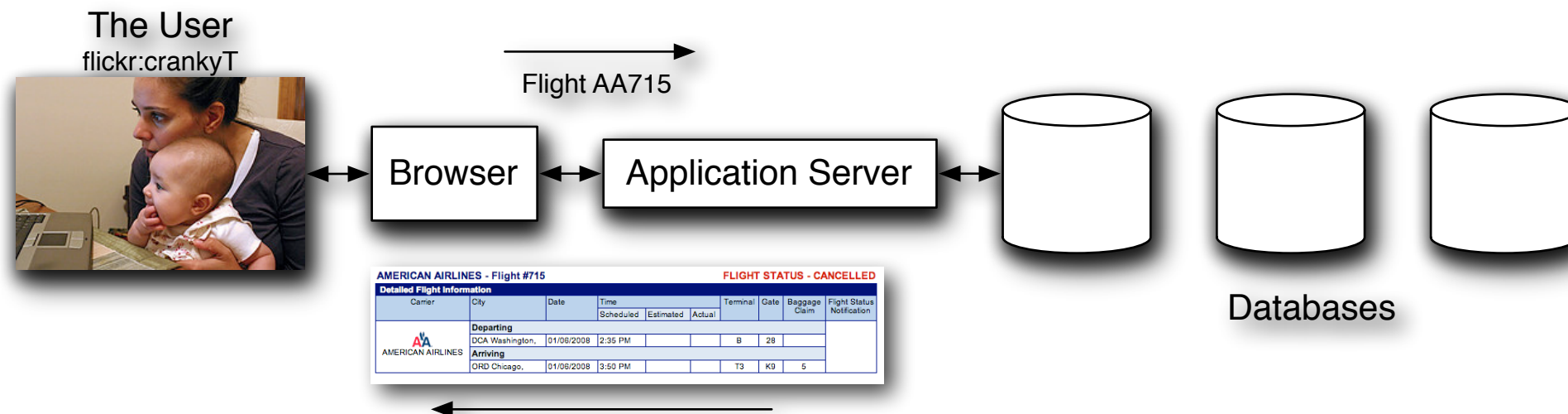
- Growth - slowing from “doubling every few months”, but still expanding



Web Search Basics

Dynamic Content

- Content can be dynamically generated
- There is no static html version
 - Flight status information, event responses
- Assembled on request (“?” in URL is a clue)

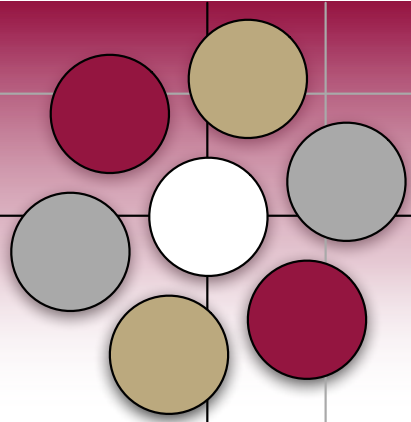




Dynamic Content

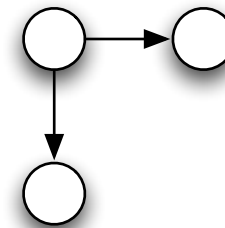
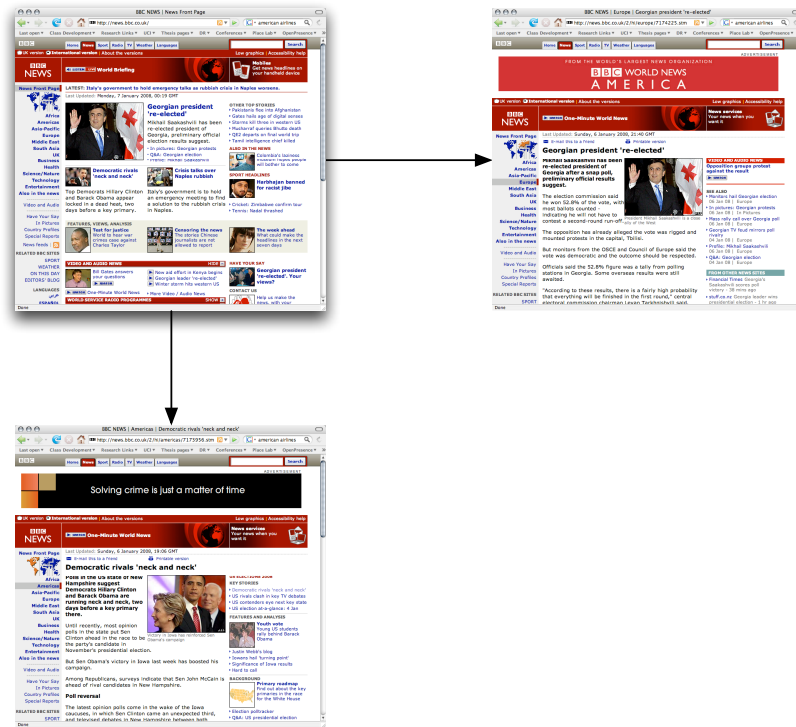
- Most (truly) dynamic content is ignored by search engines
 - Too much to index
 - Static information is more important for search
 - Spider Traps look dynamic
- Actually a lot of “static” content is assembled on the fly also
 - ASP, PHP, JSP, ads, etc....

Web Search Basics



The Web as a graph

- Web pages are nodes
- Hyperlinks are directed edges

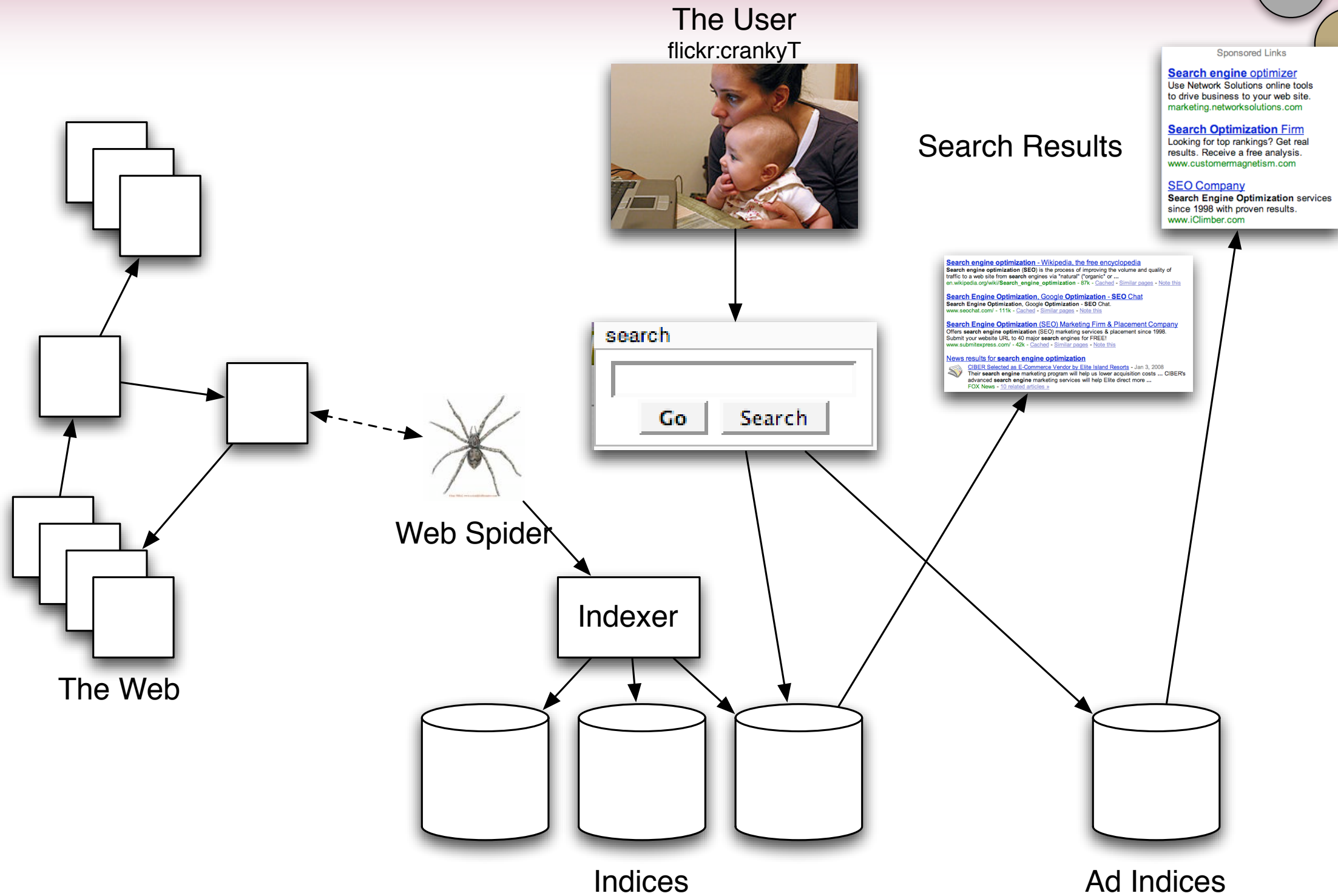




Characteristics of the web

- Significant Duplication
 - 30%-40% in some studies [Brod97, Shiv99]
 - www.copyscape.com
- High linkage
 - more than 8 links per page on average
- Spam
 - Billions of pages of it.

Web Search Basics

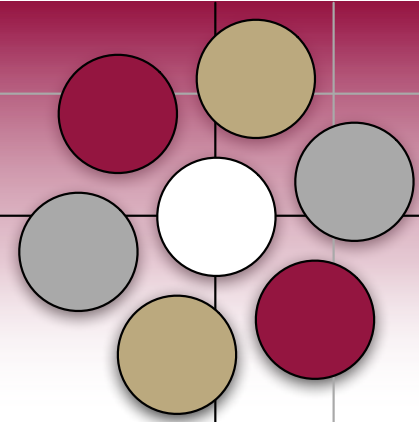




How big is the web?

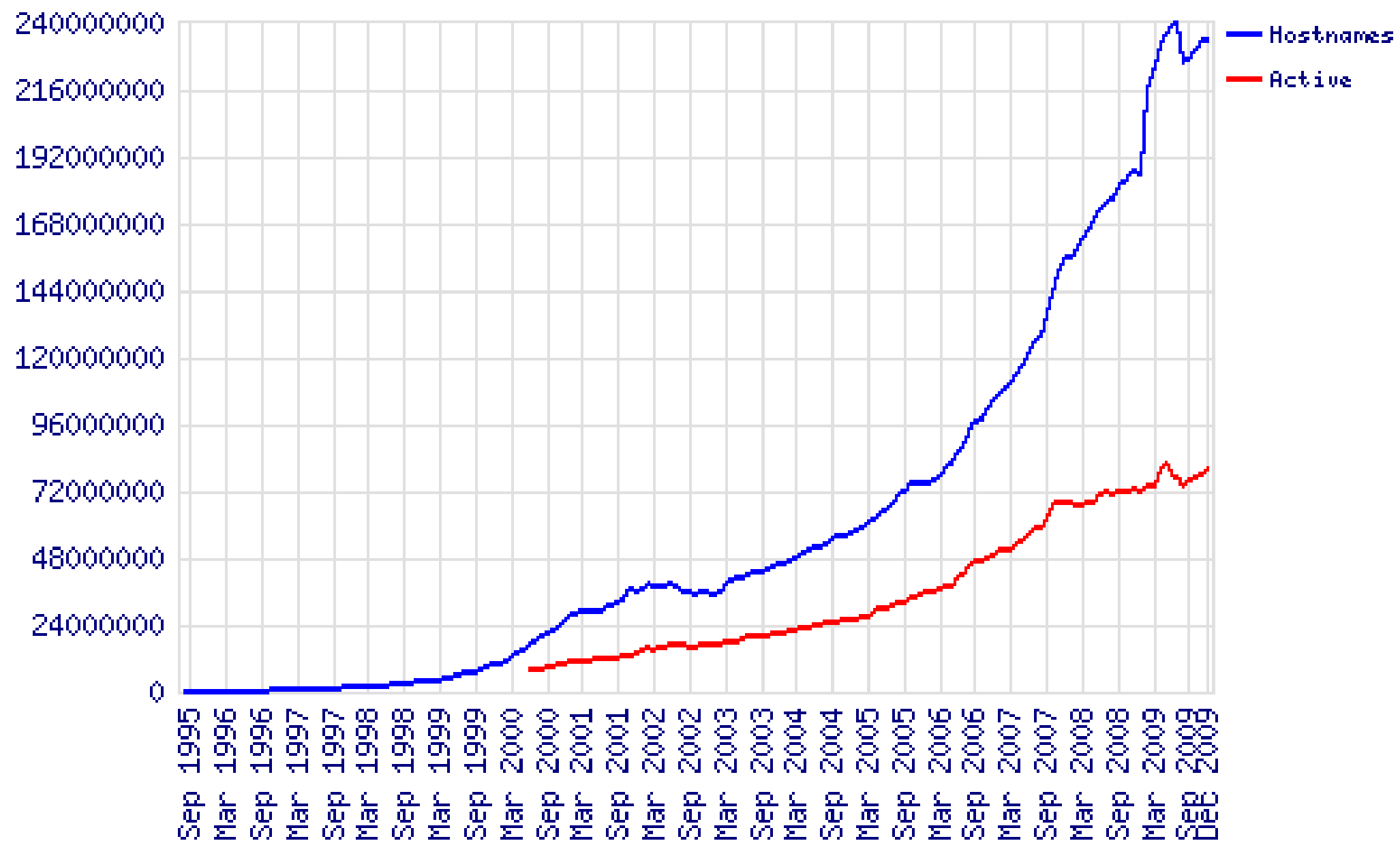
- What is measured?
 - Number of hosts
 - Number of “static” html pages
- Number of hosts - netcraft survey
 - http://news.netcraft.com/archives/web_server_survey.html
 - Monthly report on hosts and servers
- Number of pages
 - Lots of estimates which warrant further discussion

Web Search Basics: Size of the web

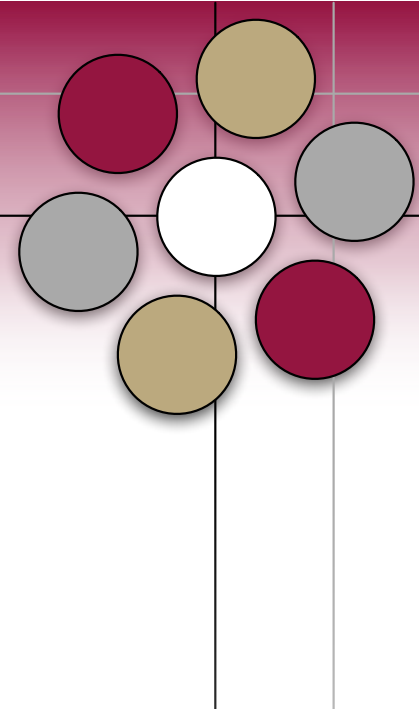


How big is the web?

- Netcraft Web Server Survey

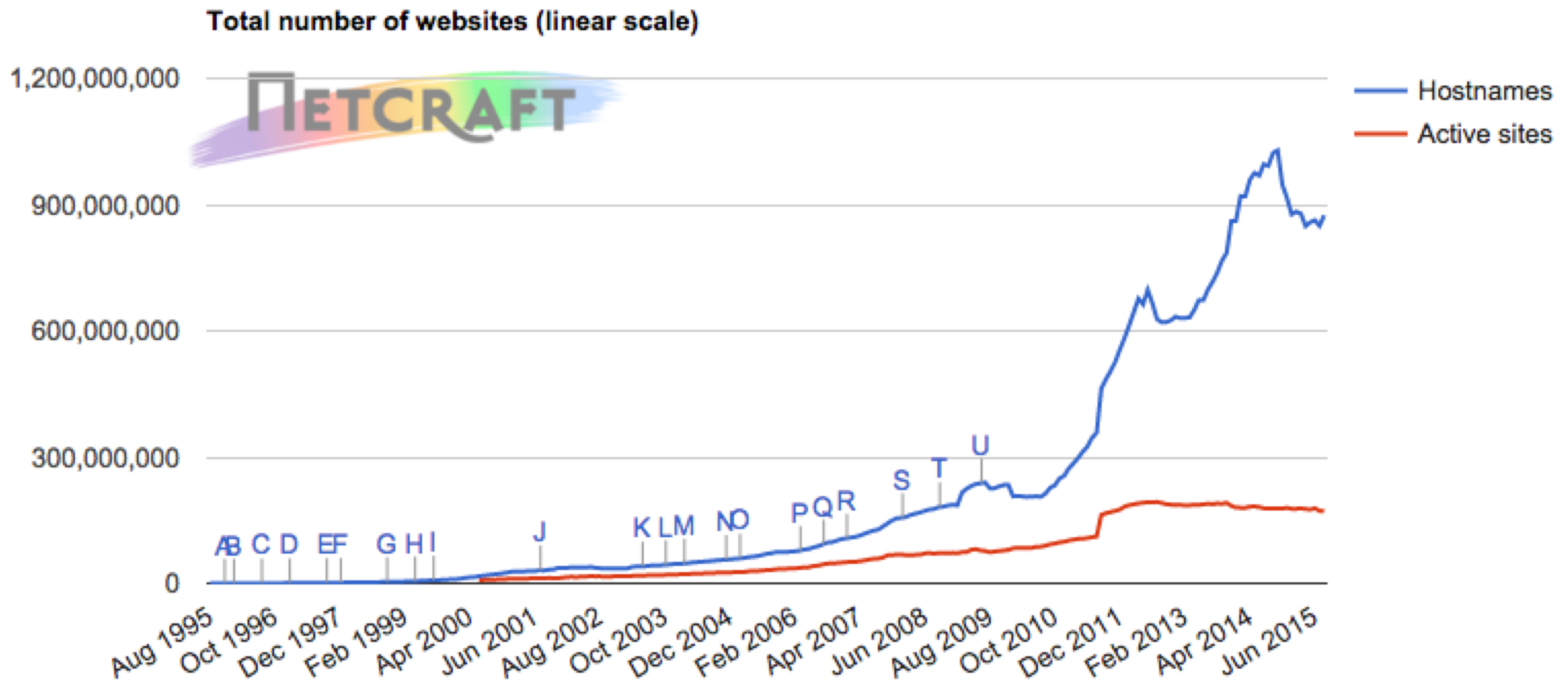


Web Search Basics: Size of the web

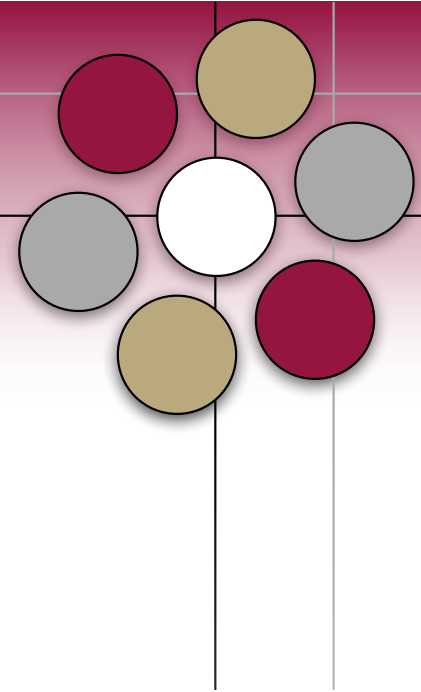


How big is the web?

- Netcraft Web Server Survey

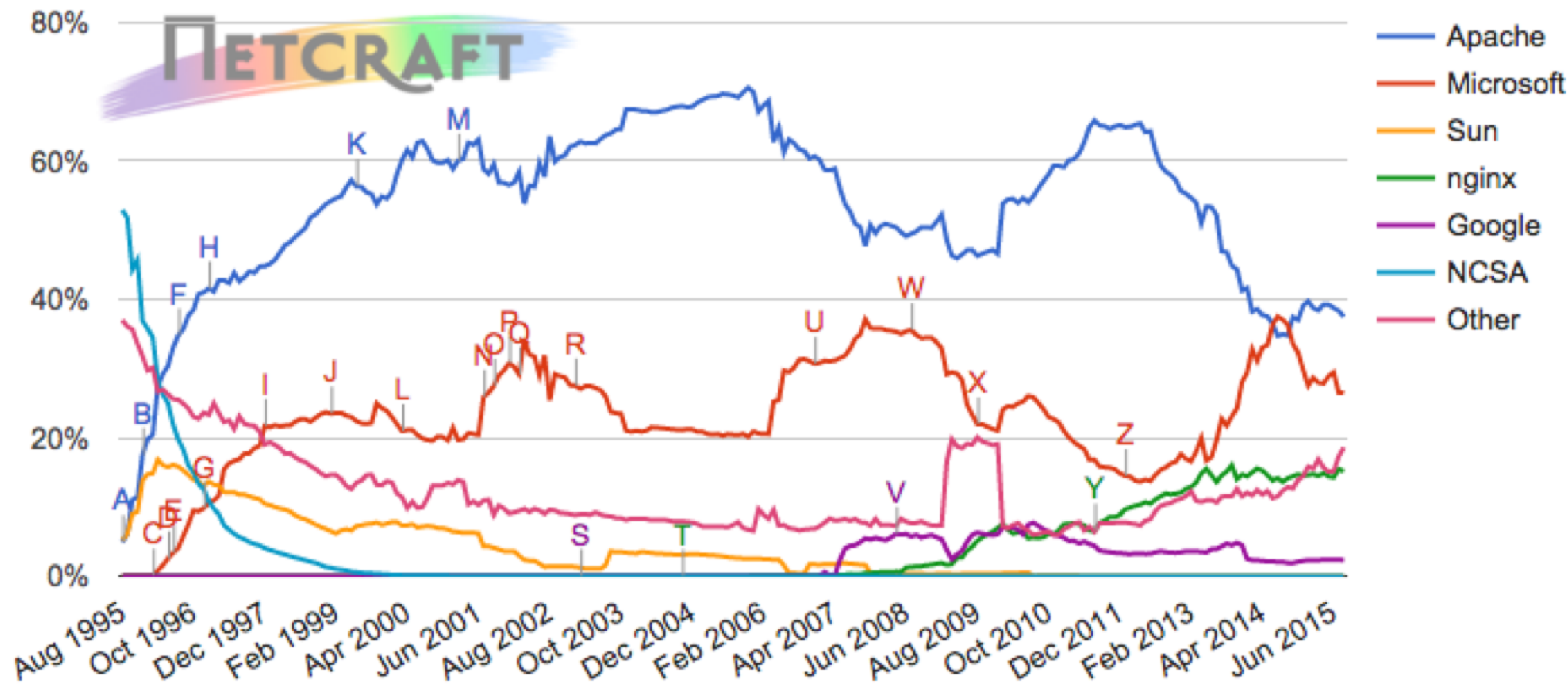


Web Search Basics: Size of the web



- Netcraft Web Server Survey

Web server developers: Market share of all sites





Rate of change

- [Cho00] 720k pages from 270 popular sites sample daily for 5 months in 1999
 - 40% changed weekly, 23% daily
- [Fett02] Massive study: 151M pages checked over a few months
 - Significant changes 7% weekly
 - Any change 25% weekly

Web Search Basics: Size of the web

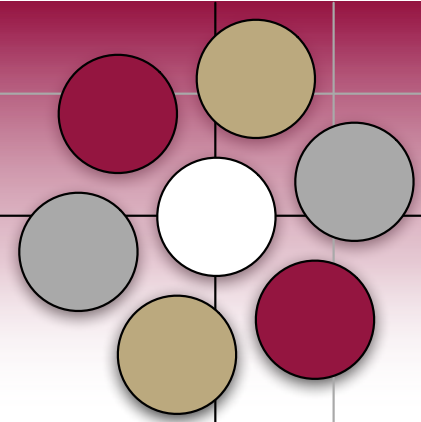


Rate of change

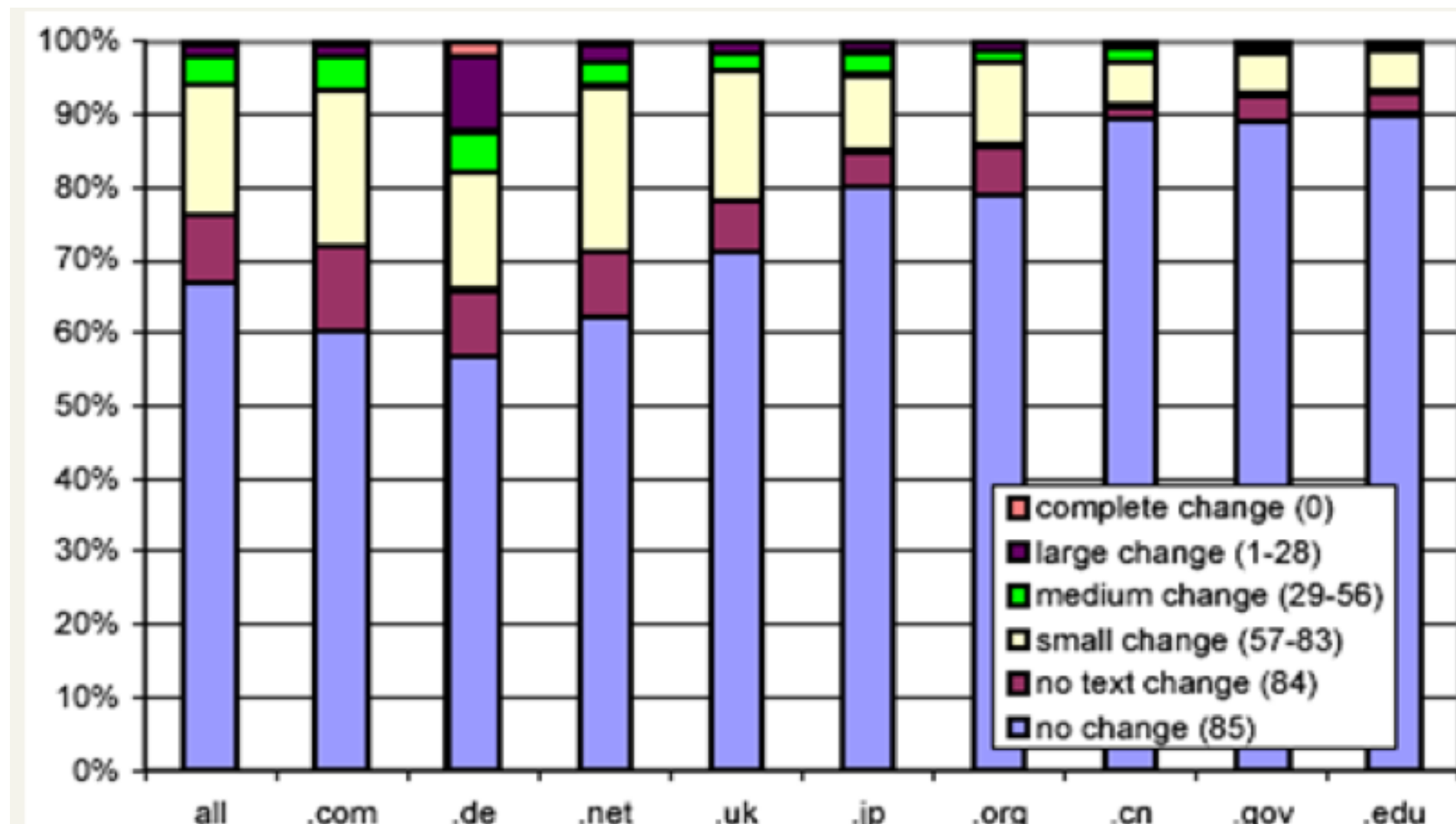
- [Ntul04] 154 large sites recrawled from scratch weekly
 - 8% had new pages ever week
 - 8% die
 - 5% new content
 - 25% new links per week

Web Search Basics: Size of the web

Rate of change



- Fetterly et al. study in 2002
- 150 million pages over 11 weekly crawls
- Bucketed into 85 groups according to amount of change

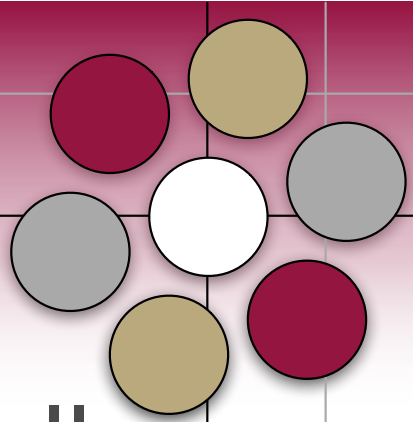




Web Evolution

- The nature of the web is change
- Not much work on studying web evolution
 - Exception is Fetterly et. al, 2003
- Some effort has been made to extrapolate from small samples using fractal models [Dill et. al. 2001]

Web Search Basics: Size of the web



The very nature of the web is changing as well

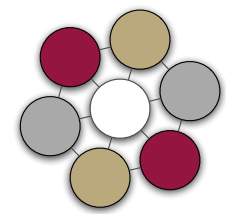
- Transforming from a source of information
- to what?
 - a communication platform?
 - a source of computation?
 - an application-space?
 - a mirror-world?
 - an augmentation of reality?
 - a cognitive orthotic?



Overview



- Introduction
- Classic Information Retrieval
- Web IR
- Sponsored Search
- Web Search Basics
 - Size of the Web



Web Users

- Spam



User Search Needs in Brod02/RL04

- Informational
 - Want to learn about something (~40%/65%)
- Navigational
 - Want to go to that page (~25%/15%)
- Transactional
 - Want to do something (~35%/20%)
 - Access a service, download, shop
- Others?
 - Exploration, social, etc...



Web Users

- Make ill defined queries
 - Short
 - Reported in 2012: ~3.5 (Dan Russell personal conv.)
 - Allegedly in 2009: 3-4 terms (lots of nuance)
 - Average in 2001: 2.54 terms (80% < 3 words)
 - Average in 1998: 2.35 terms (88% < 3 words) [Silv98]
- Imprecise terms
- Suboptimal syntax (no operators)
- Low effort (spelling mistakes)



Web Users

- Wide Variance in
 - Needs
 - Expectations
 - Knowledge
 - Bandwidth



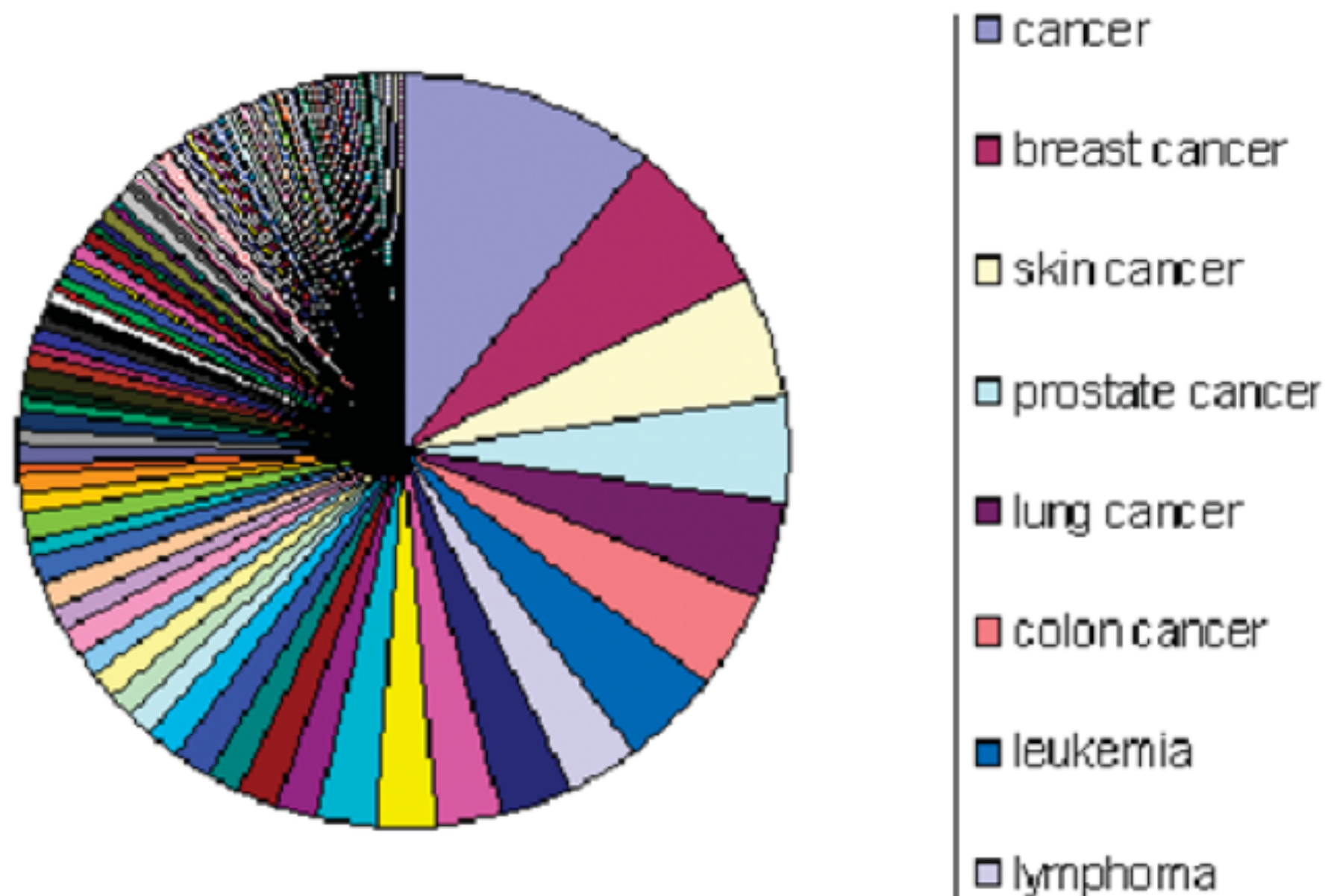
Web Users

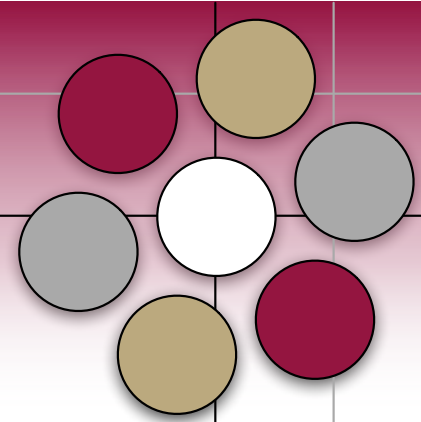
- Behavior
 - 85% look over one result screen only
 - 78% of queries are not modified
 - Follow links (“the scent of information”)

Web Users

Power law

- Few popular broad queries
- Many rare specific queries





Top queries

- Most are related to sex
- 2008 Who What How (Google)

Who is...

1. who is obama
2. who is mccain
3. who is palin
4. who is lil wayne
5. who is miley cyrus
6. who is dolla
7. who is jonas brothers
8. who is chris brown
9. who is biden
10. who is martin luther

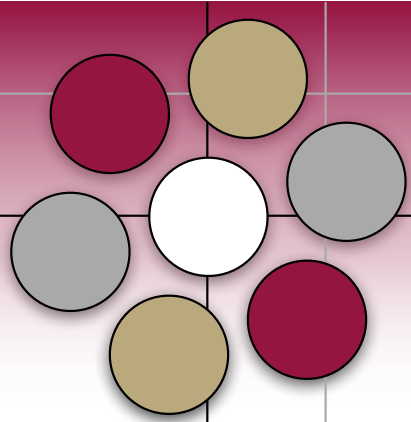
What is...

1. what is love
2. what is life
3. what is java
4. what is sap
5. what is rss
6. what is scientology
7. what is autism
8. what is lupus
9. what is 3g
10. what is art

How to...

1. how to draw
2. how to kiss
3. how to write
4. how to cook
5. how to tie
6. how to hack
7. how to run
8. how to cite
9. how to paint
10. how to spell

- <http://www.google.com/intl/en/press/zeitgeist2008/mind.html>



Top queries

- Differences today
- 01/2014 Who What How (Google)

who is|

who is
who is **red john**
who is **banksy**
who is **gossip girl**

Press Enter to search.

what is|

what is **my ip**
what is **obamacare**
what is **bitcoin**
what is **my ip address**

Press Enter to search.

how to|

how to **tie a tie**
how to **take a screenshot on a mac**
how to **make french toast**
how to **get rid of acne**

Press Enter to search.

- 09/2015

who is

who is **a**
who is **job**
who is **the next bachelor**
who is **the arkham knight**

Press Enter to search.

what is

what is **my ip**
what is **flakka**
what is **my ip address**
what is **gluten**

Press Enter to search.

how to

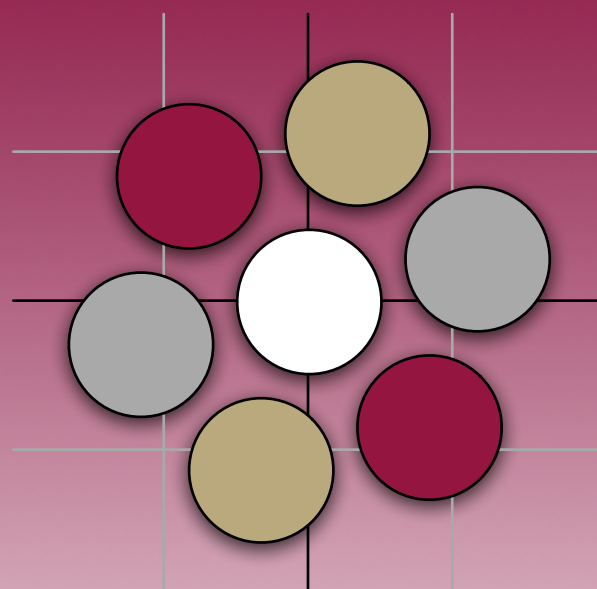
how to **tie a tie**
how to **take a screenshot on a mac**
how to **write a cover letter**
how to **screenshot on pc**

Press Enter to search.



Top queries

- Live demo - WARNING this is not very safe....
 - “Is it safe to”
 - “Is it legal to”
 - “why does”
 - “why doesn’t
 - “why is there”
 - “why isn’t”
 - “americans are”



WESTMONT COMPUTER SCIENCE