

# Web Search Basics

Introduction to Information Retrieval

INF 141/ CS 121

Donald J. Patterson

Content adapted from Hinrich Schütze

<http://www.informationretrieval.org>



# Introduction

Search Education - Google

www.google.com/insidesearch/searcheducation/index.html


Google | Search Education


Home Lesson Plans Live Trainings

## Help your students become better searchers

Web search can be a remarkable tool for students, and a bit of instruction in how to search for academic sources will help your students become critical thinkers and independent learners.

With the materials on this site, you can help your students become skilled searchers- whether they're just starting out with search, or ready for more advanced training.






### Lesson Plans & Activities

Download lesson plans to develop your students' search literacy skills.


[Browse lesson plans](#)



### Power Searching

Improve your search skills and learn advanced tips with online lessons and activities.


[Start now](#)



### A Google a Day Challenges

Put your students' search skills to the test with these trivia challenges.

[Browse challenges](#)



### Live Trainings

Join us for live search trainings or watch past trainings from search experts here at Google.

[Start training](#)

# Without search engines the web wouldn't scale

- No incentive in creating content unless it can be found.
  - Taxonomies, bookmarks can't keep up
  - Or can they? (del.icio.us)
- The web is both a technology artifact and a social environment
  - “The Web has become the ‘new normal’ in the American way of life; those who don't go online constitute an every-shrinking minority” [Pew Foundation report, January 2005]



# Without search engines the web wouldn't scale

- Search engines make aggregation of interest possible:
  - Create incentives for very niche players
    - Economical - specialized stores, providers, etc.
    - Social - narrow interests, specialized communities
- The acceptance of search interaction makes “unlimited selection” stores possible
  - Amazon, Netflix, etc.



# Without search engines the web wouldn't scale

- Search turned out to be the best mechanism for advertising on the web,.
- Growing very fast (entire US advertising industry is \$250 billion though)
- \$15 billion plus industry in 2009
- \$36 billion in 2012





## Overview

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



## Classic IR assumptions

- Corpus: Fixed document collection
- Goal: Retrieve information content relevant to information need



## Classic IR Goal

- Classic “Relevance”
  - For each query,  $Q$ , and stored document,  $D$ , in a corpus there exists a relevance score:  $R(Q,D)$
  - $R(Q,D)$  is averaged over users,  $U$ , and contexts,  $C$
  - Maximize  $R(Q,D)$  instead of  $R(Q,D,U,C)$ 
    - Context is ignored
    - Individuals are ignored
    - Corpus is static





## Overview

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



## Web IR: Differences from traditional IR

- On the web, search and ads are intricately connected
- The web is huge
- The web is a rapidly changing collection.
- There is spam on the web
  - Adversarial IR
  - Huge difference from traditional IR
- One interface for hugely divergent needs
  - Queries, Maps, Stocks, Weather, Calculations



## History

- Early keyword-based engines
  - (1995-1997) Altavista, Excite, Infoseek, Inktomi
- Paid placement ranking
  - Goto.com -> Overture.com -> Yahoo!
    - Results based on auction for keyword placement



www.goto.com/d/search/?\$sessionid\$AQ4214AAH6R5QFIEF3QPUQ?type=home&tm=1&Keywords=Wilmington+

Wilmington real estate.

Access 75% of all users now!  
Premium Listings reach 75% of all  
Internet users. [Sign up](#) for Premium  
Listings today!

ib of  
ow!  
stings  
of all  
ers.  
stings

1. [Wilmington Real Estate - Buddy Blake](#)  
Wilmington's information and real estate guide. This is your on  
anything to do with Wilmington.  
[www.buddyblake.com](#) (Cost to advertiser: [10.38](#))
2. [Coldwell Banker Sea Coast Realty](#)  
Wilmington's number one real estate company.  
[www.cbseacoast.com](#) (Cost to advertiser: [10.37](#))
3. [Wilmington, NC Real Estate Becky Bullard](#)  
Everything you need to know about buying or selling a home c  
on my Web site!  
[www.iwwc.net](#) (Cost to advertiser: [10.35](#))

ns & more

## History

- (1998+) Link-based ranking pioneered by Google
  - Links added the idea of “authoritativeness” to “relevance”
  - Blew away all early engines save Inktomi
  - Great user experience looking for a business model
  - Meanwhile Goto/Overture’s annual revenues were nearing \$1 billion



## History

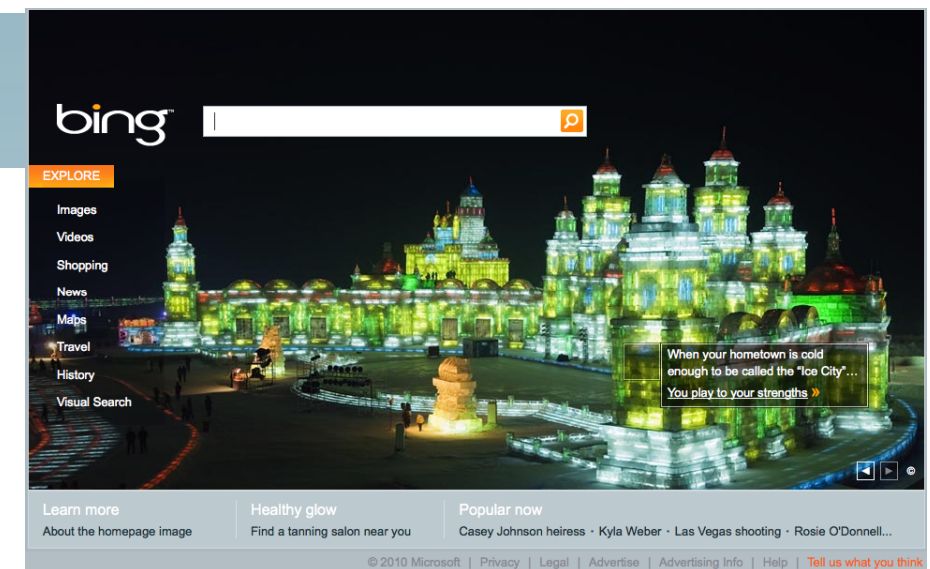
- Result
- Google:
  - Added paid placement ads on the side
  - Differentiated from search results
- Yahoo! built a similar architecture
  - Buys Overture for paid placement
  - Buys Inktomi for search





## History

- 2004
  - Microsoft begins in house development of a search engine called Live
- May 28, 2009
  - Microsoft rebrands Live Search to Bing!
  - Search Engine wars intensify
  - New innovations appears at every turn
  - Technology becomes much more closely guarded



## History

- Internationally
  - Chinese search engine Baidu “owns” Chinese search
  - Launched around 2000, specializes in Chinese content



## Today (1/7/2014)

- |                     |                     |                   |
|---------------------|---------------------|-------------------|
| 1. google.com       | 11. sina.com.cn     | 21. wordpress.com |
| 2. facebook.com     | 12. twitter.com     | 25. bing.com      |
| 3. youtube.com      | 13. hao123.com      | 30. pinterest.com |
| 4. yahoo.com        | 14. 163.com         | 32. msn.com       |
| 5. baidu.com        | 15. blogspot.com    | 34. tumblr.com    |
| 6. wikipedia.org    | 16. google.co.in    | 38. instagram.com |
| 7. qq.com           | 17. linkedin.com    | 39. paypal.com    |
| 8. taobao.com       | 18. weibo.com       | 41. <porn>        |
| 9. amazon.com       | 19. tmall.com       | 45. apple.com     |
| 10. <u>live.com</u> | 20. <u>ebay.com</u> | 50. <porn>        |



## Before (1/7/2010)

- |                  |                  |
|------------------|------------------|
| 1. google.com    | 13. twitter.com  |
| 2. facebook.com  | 15. google.cn    |
| 3. youtube.com   | 22. bing.com     |
| 4. yahoo.com     | 29. google.co.jp |
| 5. live.com      | 56. ask.com      |
| 6. wikipedia.org | 64. cnn.com      |
| 7. blogger.com   |                  |
| 8. baidu.com     |                  |
| 9. msn.com       |                  |
| 10. yahoo.co.jp  |                  |



## Overview

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





# Sponsored Search

[Advanced Search](#)  
[Preferences](#)[Web](#) [Blogs](#) [News](#)Personalized Results 1 - 1000,000 for [search engine optimization](#) (seconds)

## [Search Engine Optimize](#)

[SEOP.com](#) Guaranteed Top Ranking w/ Warranty. Free Site Analysis! 877-231-1555

## [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 optimization](#)

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](#)

## [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](#)



## 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](http://marketing.networksolutions.com)

#### [Search Optimization Firm](#)

Looking for top rankings? Get real results. Receive a free analysis.  
[www.customermagnetism.com](http://www.customermagnetism.com)

#### [SEO Company](#)

**Search Engine Optimization** services since 1998 with proven results.  
[www.iClimber.com](http://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](http://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/](http://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/](http://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

The image shows a screenshot of a Facebook profile page. The profile name is 'Thi' and the profile picture is a woman with dark hair. The page is divided into several sections: a left sidebar with navigation links and an advertisement, a central profile section, and a right section for updates and activity. Two red boxes highlight specific sponsored search results, with red arrows pointing to them from the central activity feed.

**Facebook Profile: Thi**

**Left Sidebar:**

- Search: [Search bar]
- Applications: + Add
- Photos
- Groups
- Events
- Marketplace
- My Aquarium
- SuperFaker
- + More

**Advertisement:**

**FREE COACH PURSE!**  
Advertised by Coach.com

**Profile Section:**

- View Photos of Thi (2)
- Where Thi Traveled (110)
- high five Thi
- Write on my Funtail
- You are online now.
- Friends: 202 Friends
- Friends in: Harvard (46), New York, NY (26), San Francisco, CA (12), ExpoTV (20), Stanford (8)

**Activity Feed:**

- Today**
- Thi added Sony Professional HVR-Z1U 3CCD High Definition to her wishlist on Expo TV. 10:52am
- Yesterday**
- Thi signed up for a member profile on Expo TV. 11:45pm
- Thi added Logitech QuickCam Communicate STx to her wishlist on Expo TV. 11:00pm
- Thi added Logitech QuickCam Communicate STx to her wishlist on Expo TV. 11:00pm
- Thi signed up for a member profile on Expo TV. 11:00pm
- Thi reviewed KitchenAid KSM150PSE Artisan Series 5-Quart Mixer, on Expo TV. 11:05pm
- Thi added Sony HDR EX768L NR Portopia headphones (white) to her wishlist on Expo TV. 11:00pm
- Thi joined the group. 11:00pm

**Highlighted Sponsored Search Results:**

- Top Highlight:** Thi added Sony Professional HVR-Z1U 3CCD High Definition to her wishlist on Expo TV. 10:52am
- Bottom Highlight:** Thi reviewed KitchenAid KSM150PSE Artisan Series 5-Quart Mixer, on Expo TV. 11:05pm

## 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



# 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



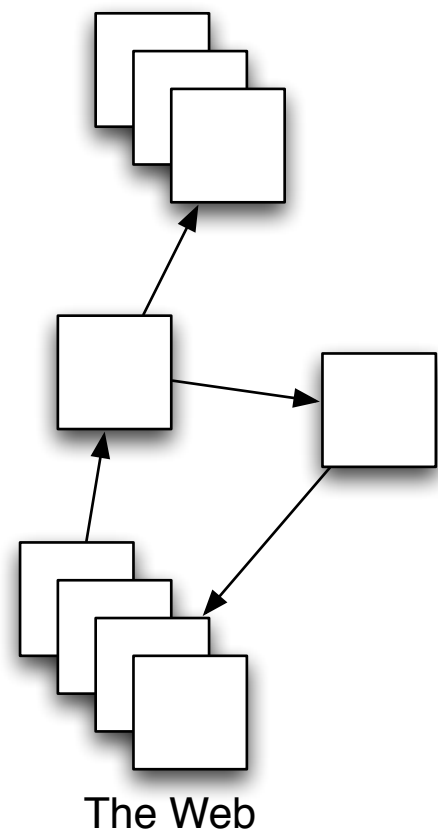
## Overview

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





## The Web Corpus



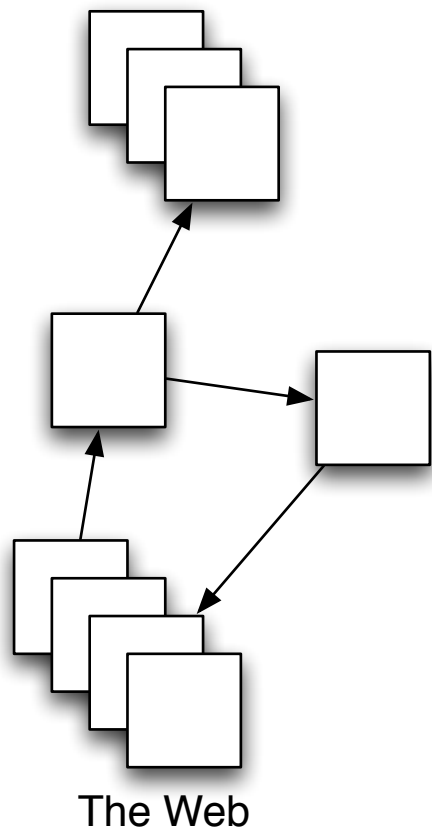
- No design/coordination
- Distributed content creation, linking
- “Democratization of publishing”
- Content includes truth, lies, contradictions, etc.
- Unstructured Data (text, html)
- Semi-Structured (XML, annotated photos)
- Structured (Databases)
- Scale is much larger than previous text corpora





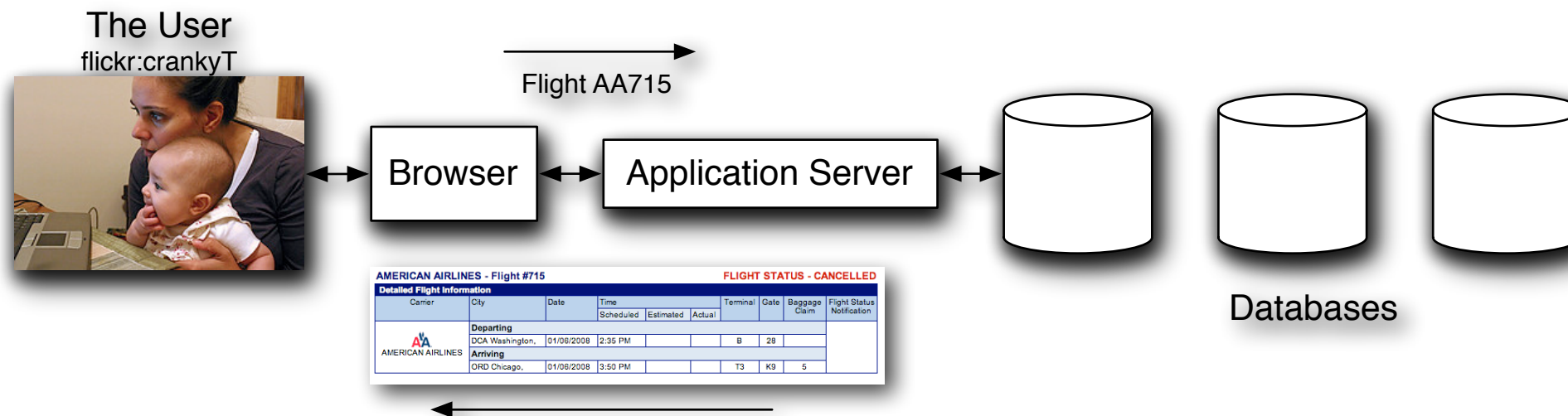
## The Web Corpus

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



## 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)



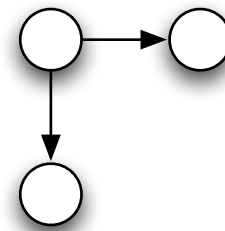
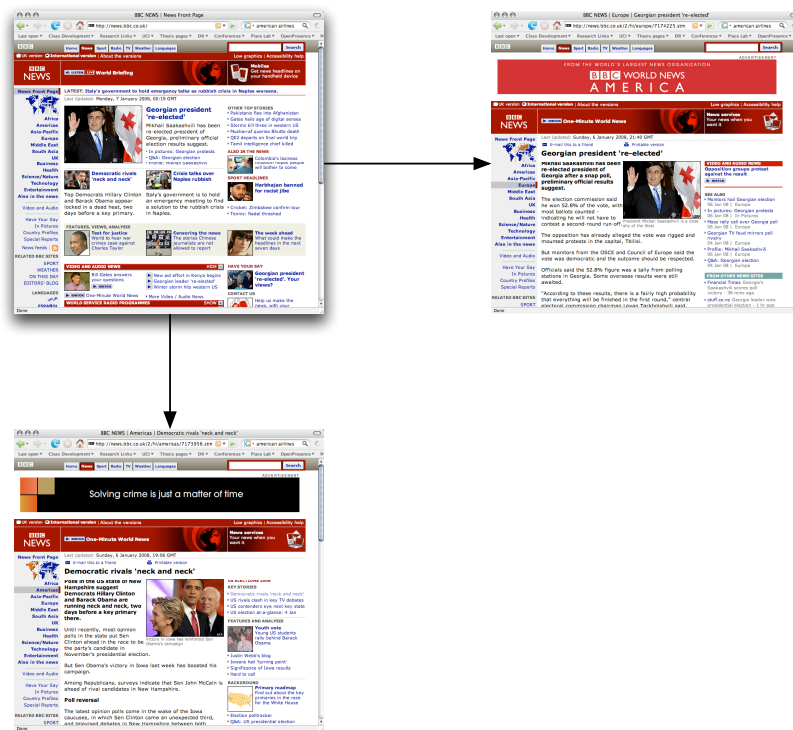
# Dynamic Content

- 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....



## The Web as a graph

- Web pages are nodes
- Hyperlinks are directed edges

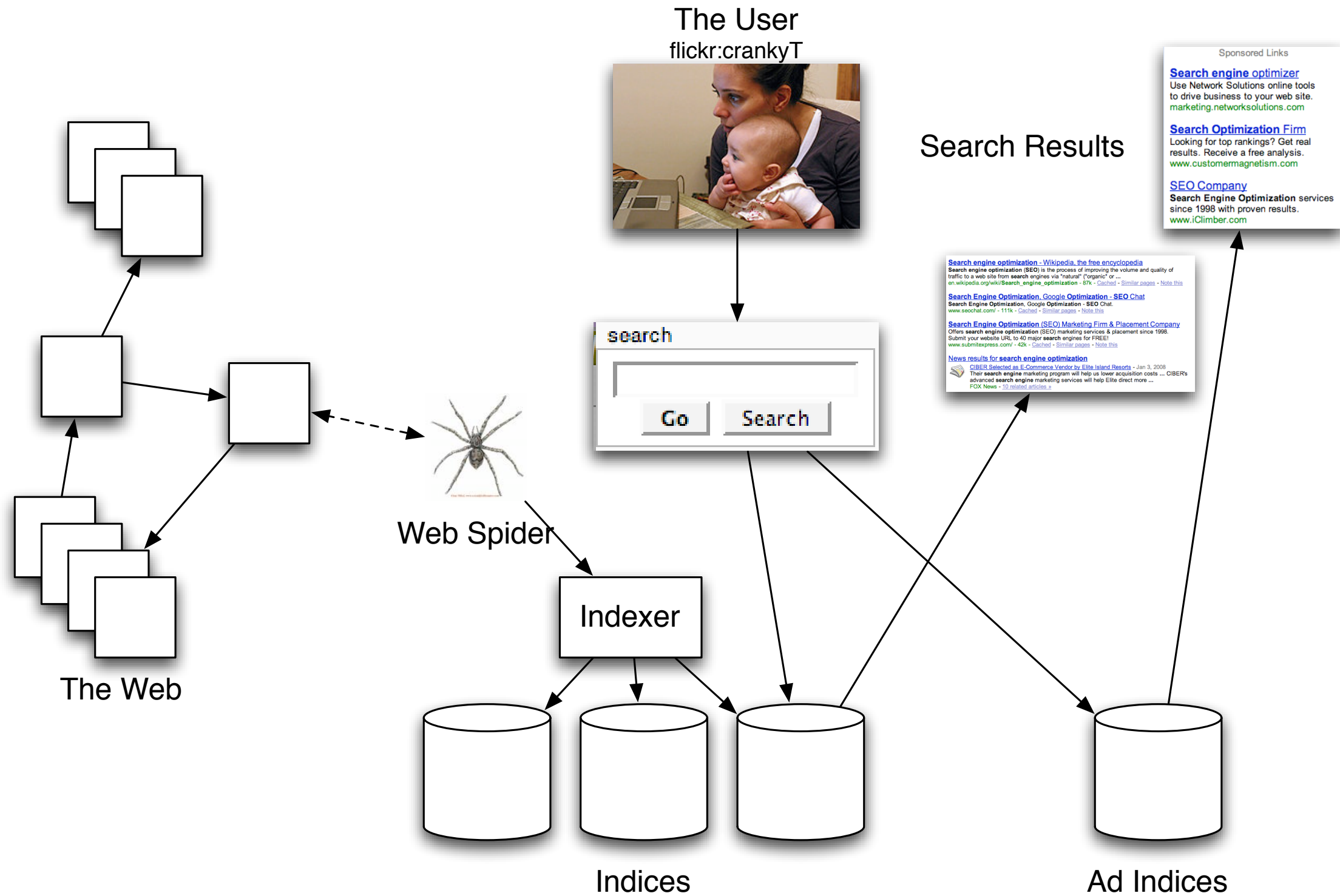


# Characteristics of the web

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



# Web Search Basics





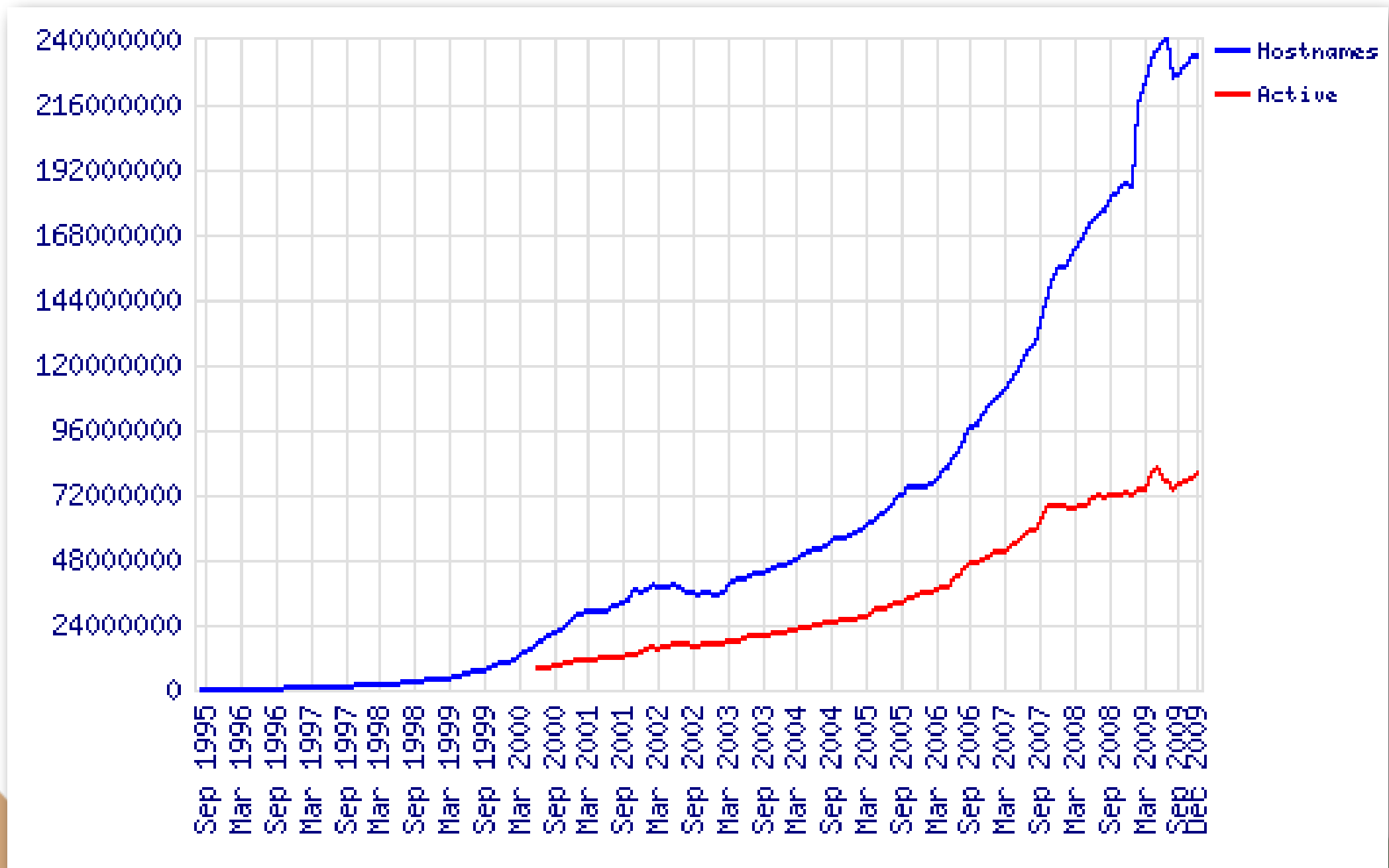
# How big is the web?

- 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](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



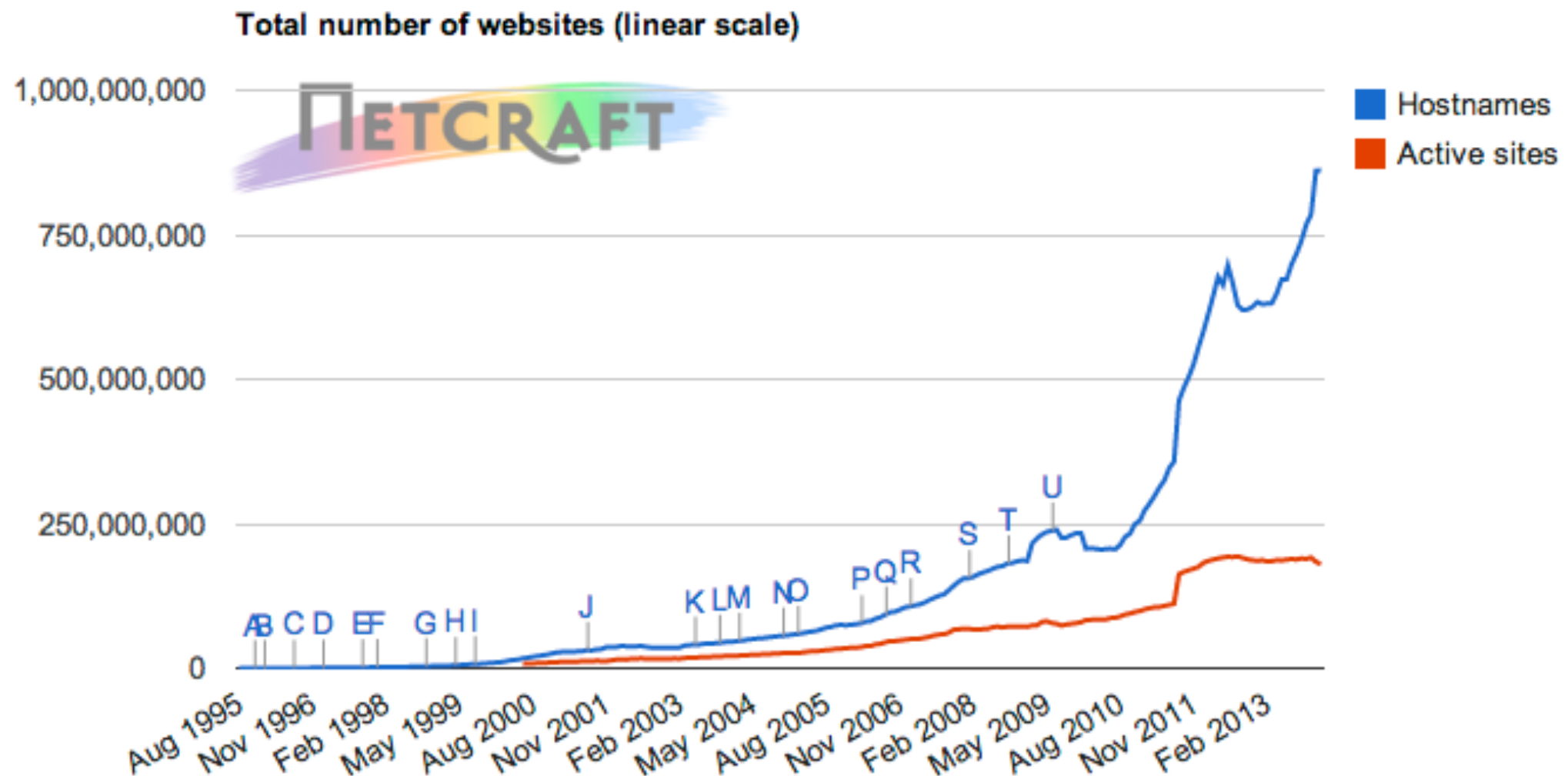
# How big is the web?

- Netcraft Web Server Survey



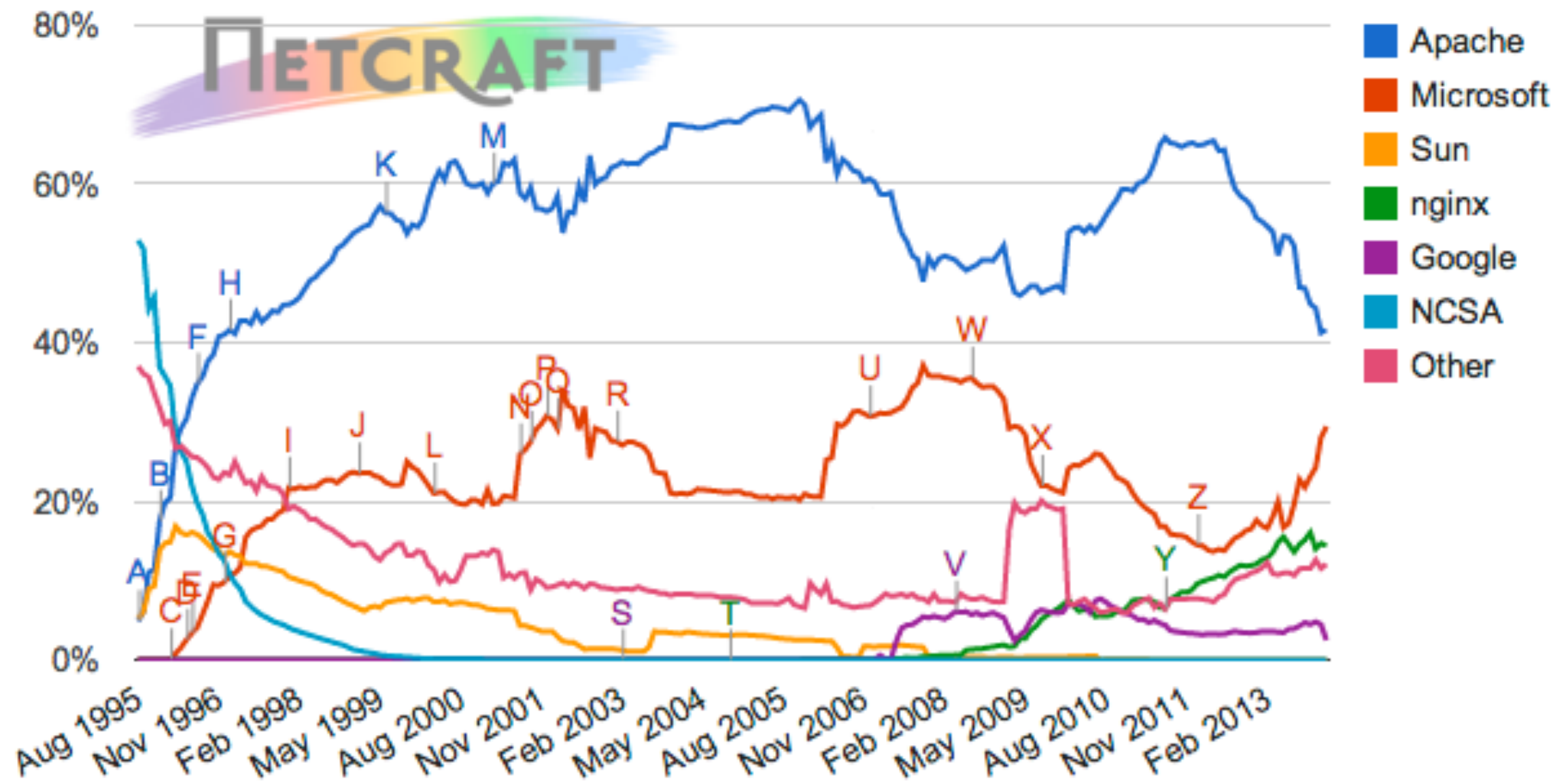
## How big is the web?

- Netcraft Web Server Survey



# Nerd Trivia

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



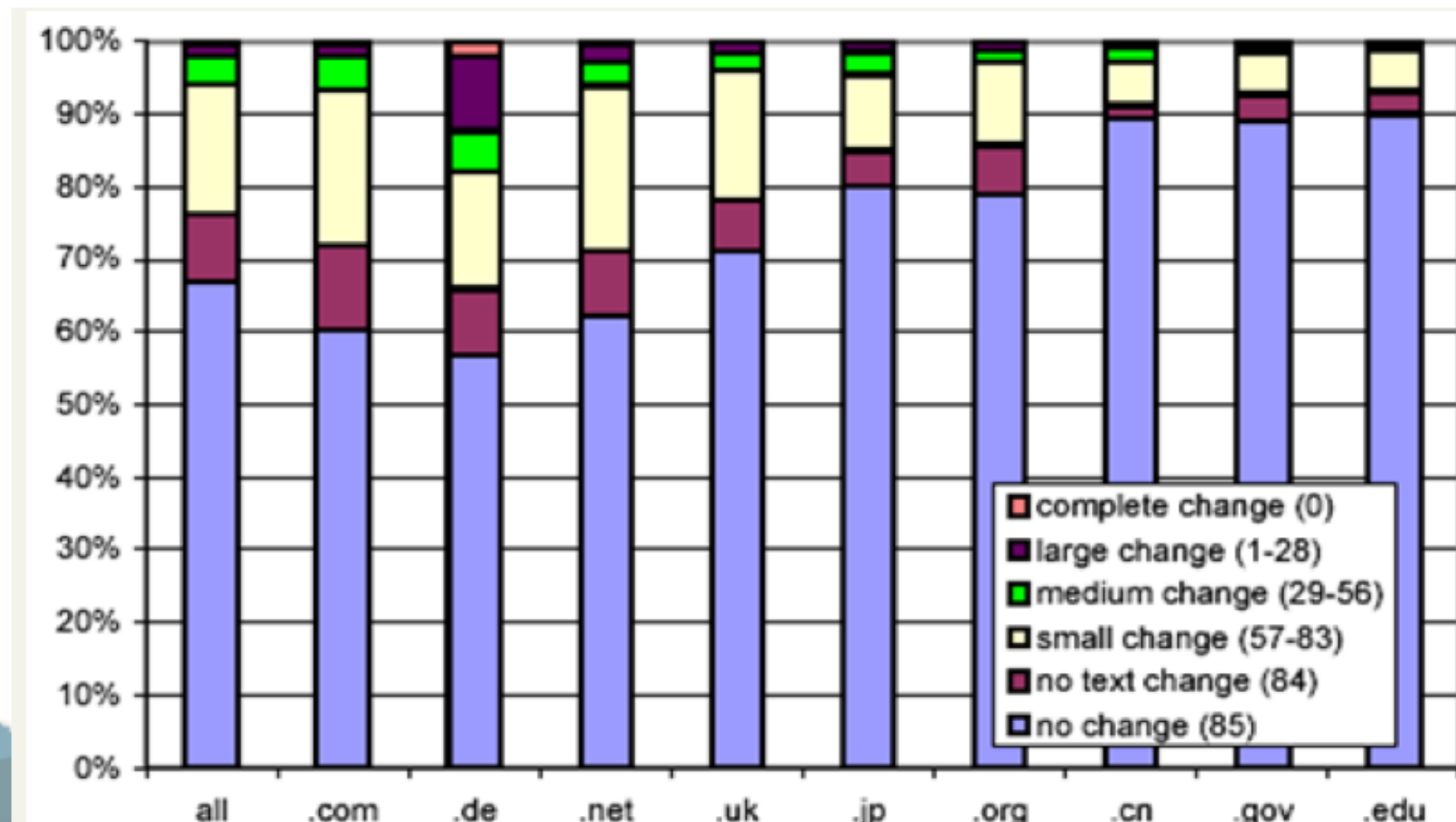
## Rate of change

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



## 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]



# 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?



