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

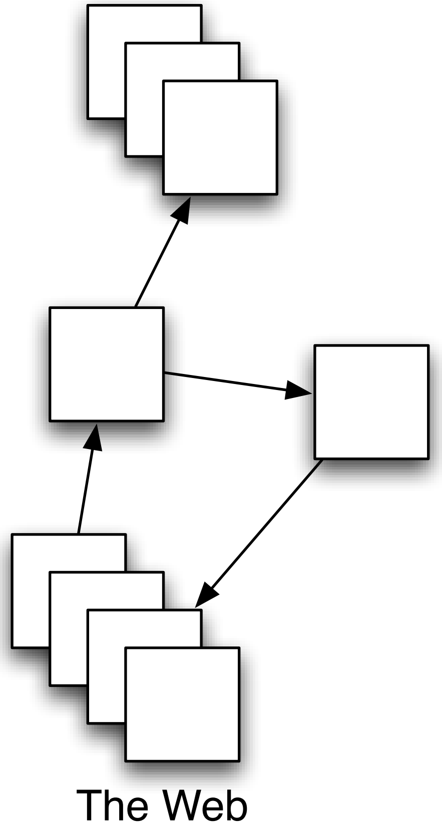
Paying for advertisements

- 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

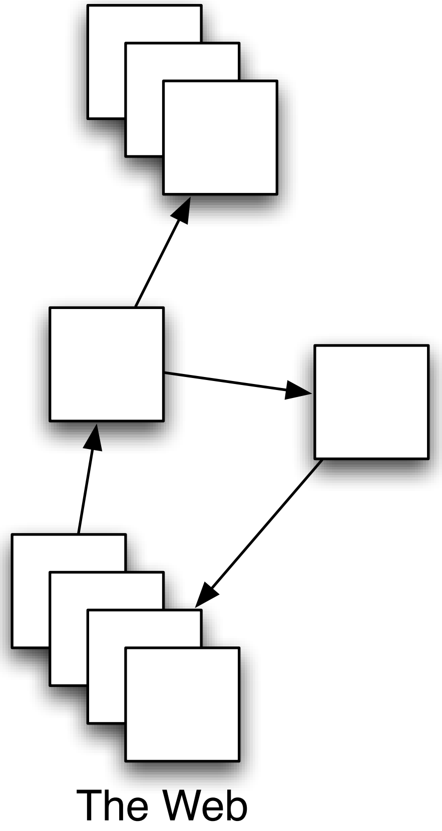
- Introduction
- Classic Information Retrieval
- Web IR
- Sponsored Search
- Web Search Basics
    - Size of the Web
- Web Users
- Spam

The Web Corpus



- No design/coordination
- Distributed content creation, linking
- “Democatization 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 copora

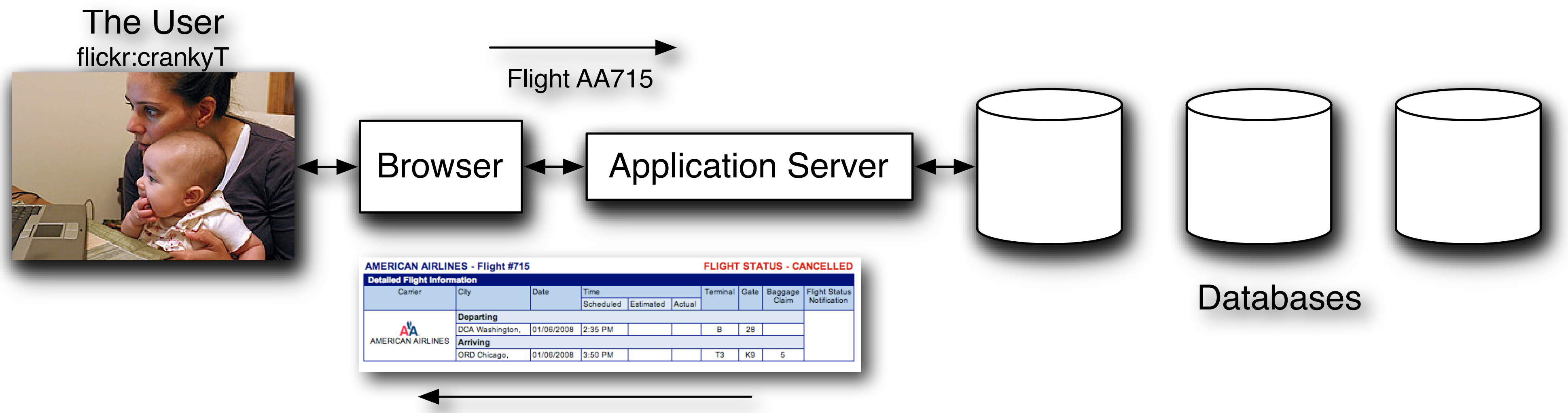
The Web Corpus



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

Dynamic Content

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



AMERICAN AIRLINES - Flight #715 FLIGHT STATUS - CANCELLED

Detailed Flight Information

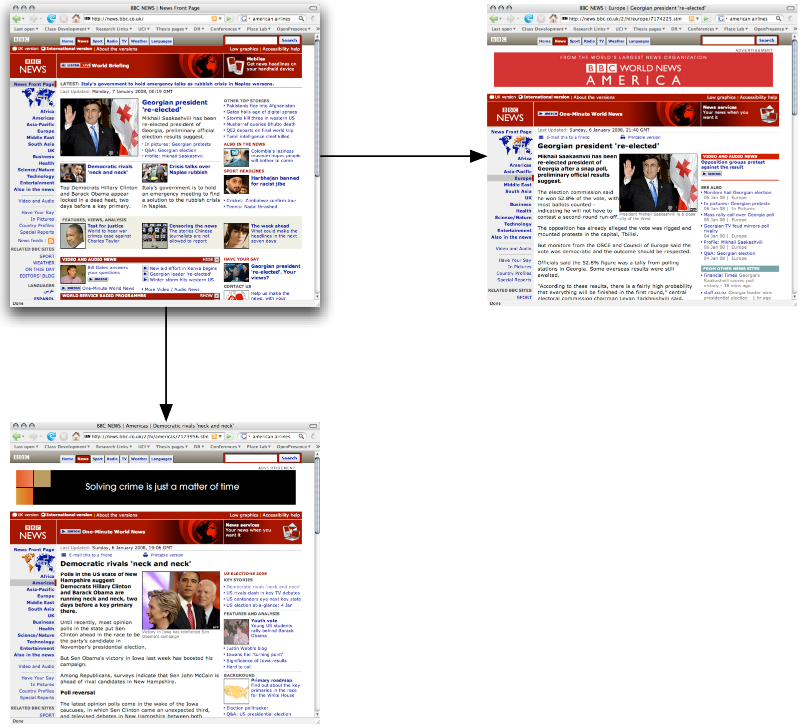
Carrier	City	Date	Time			Terminal	Gate	Baggage Claim	Flight Status Notification
			Scheduled	Estimated	Actual				
AMERICAN AIRLINES	**Departing**								
	DCA Washington,	01/06/2008	2:35 PM			B	28		
	Arriving								
	ORD Chicago,	01/06/2008	3:50 PM			T3	K9	5	

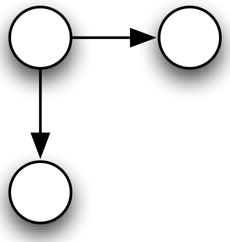
Dynamic Content

- Most (truly) dynamic content is ignored by web spiders
    - 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

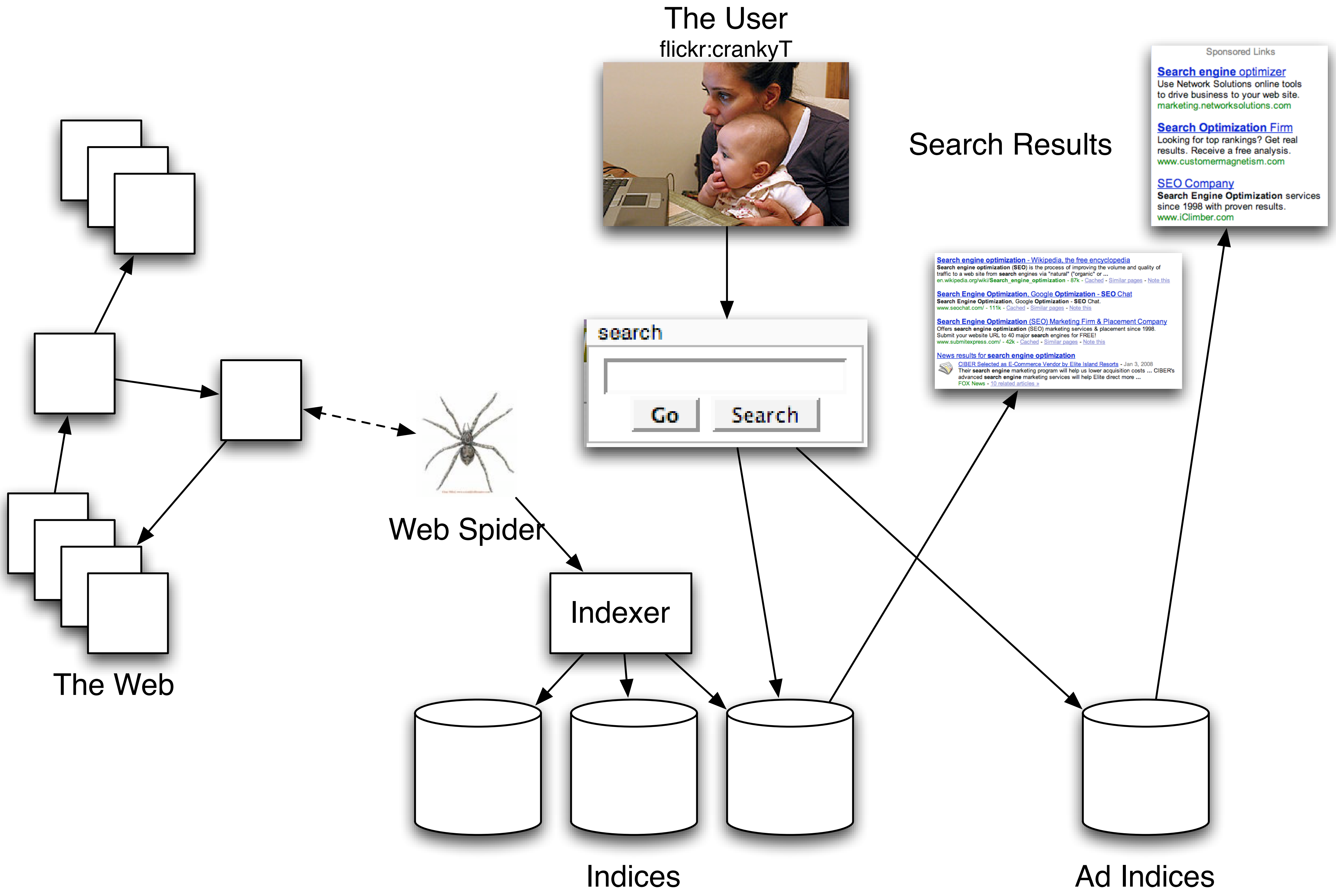

BBC NEWS
Georgian president 're-elected'
BBC WORLD NEWS AMERICA
Democratic rivals 'neck and neck'
Solving crime is just a matter of time



Characteristics of the web

- Significant Duplication
  - 30%-40% is 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



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 »

Size of the Web

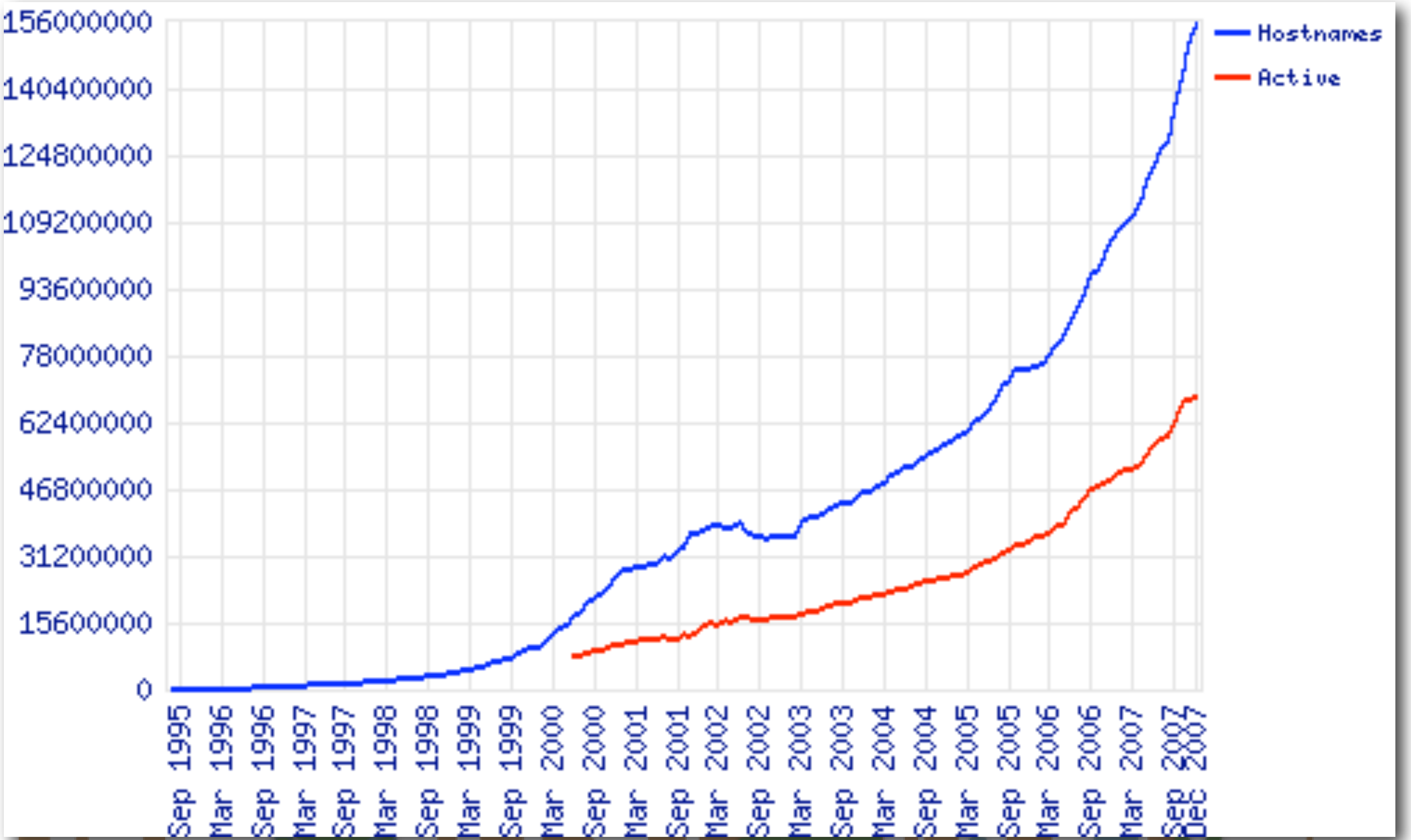
How big is the web?

- 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

Size of the Web

How big is the web?

- Netcraft Web Server Survey



Size of the Web

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

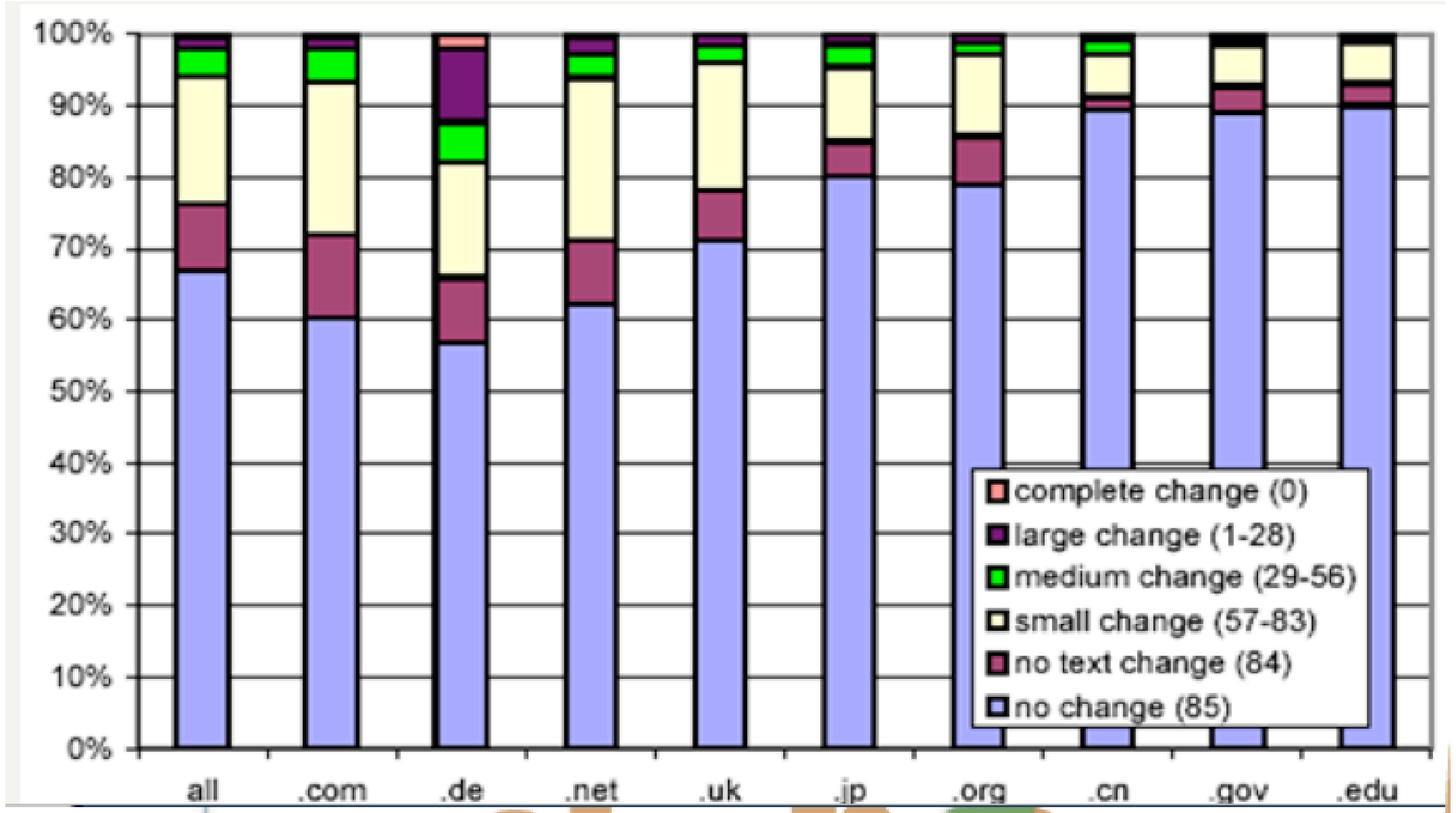
Size of the Web

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

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



Size of the Web

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]

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
        - 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

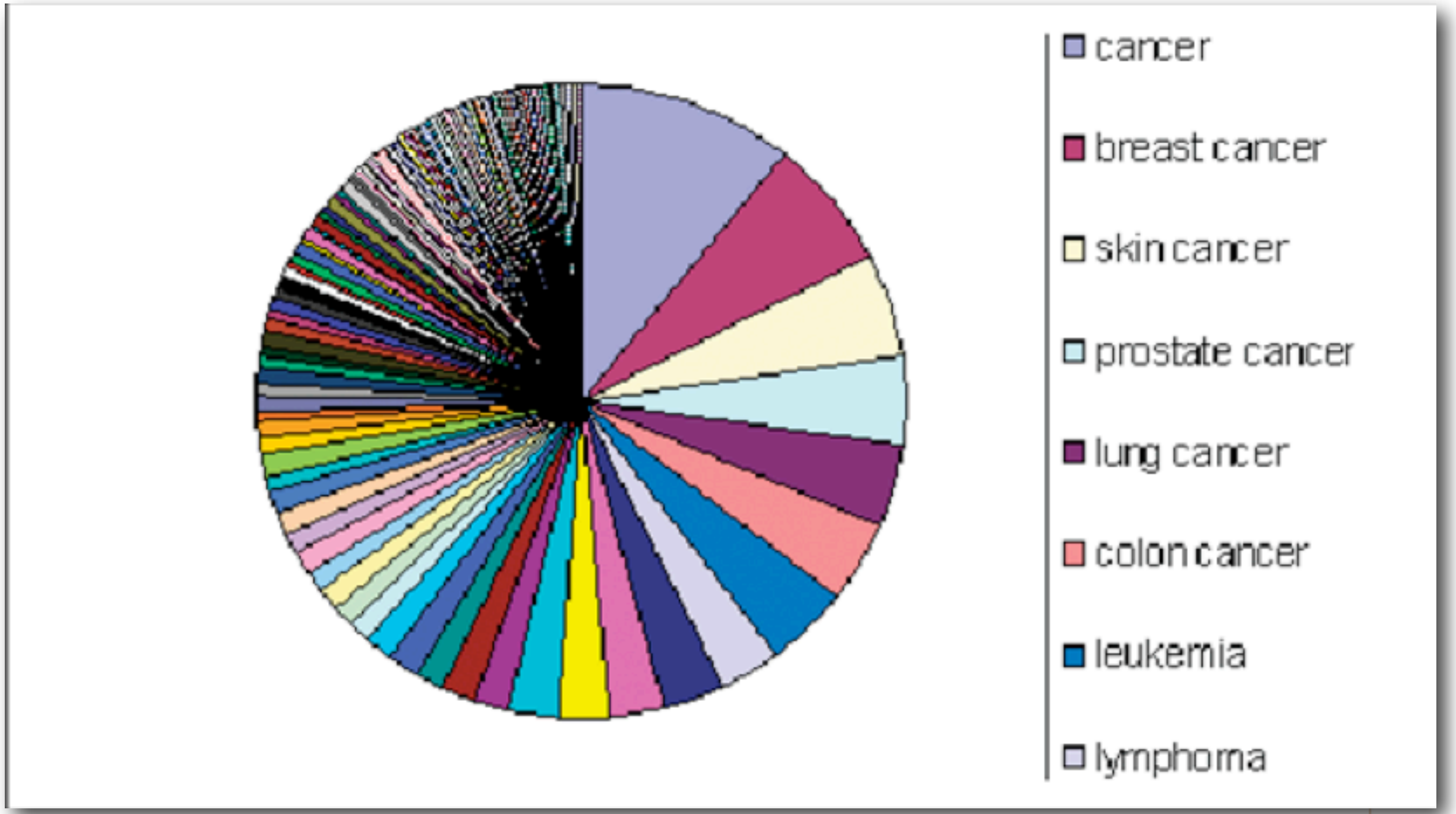
- Wide Variance in
  - Needs
  - Expectations
  - Knowledge
  - Bandwidth

Web Users

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

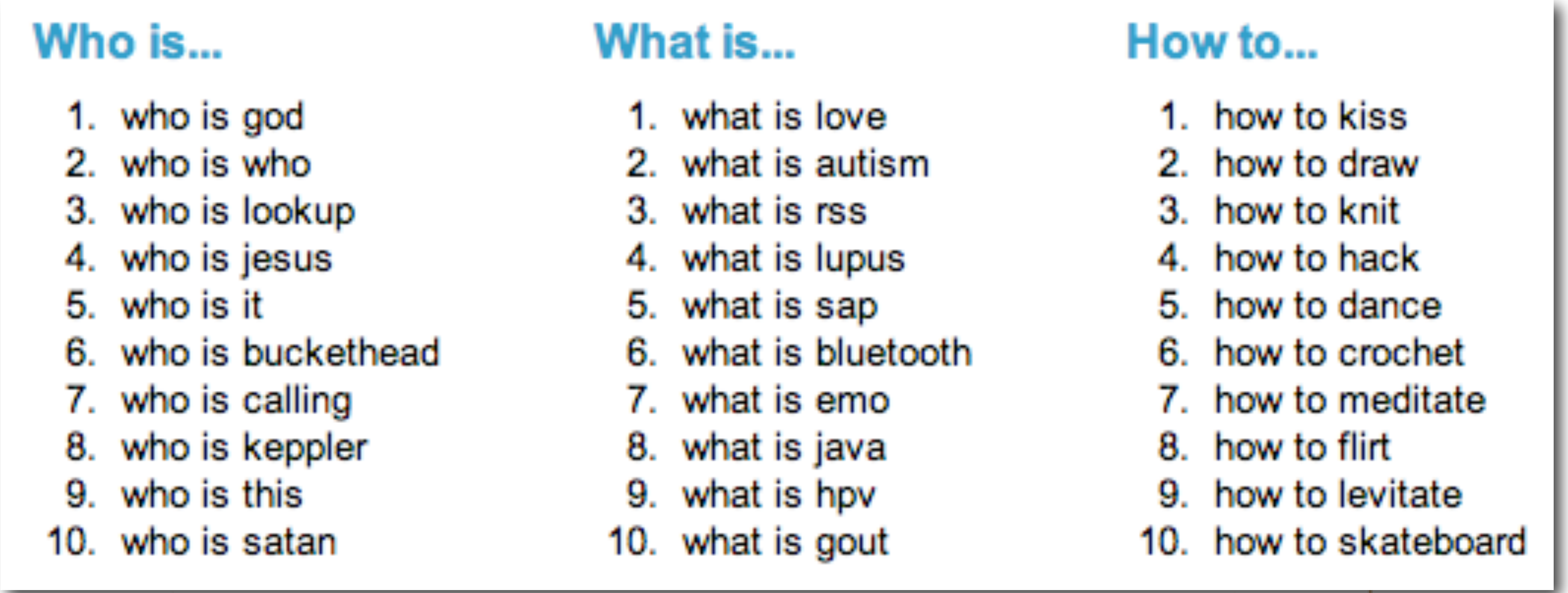
Power law

- Few popular broad queries
- Many rare specific queries



Top queries

- Most are related to sex
- 2007 Who What How (Google)



Who is...

1. who is god
2. who is who
3. who is lookup
4. who is jesus
5. who is it
6. who is buckethead
7. who is calling
8. who is keppler
9. who is this
10. who is satan

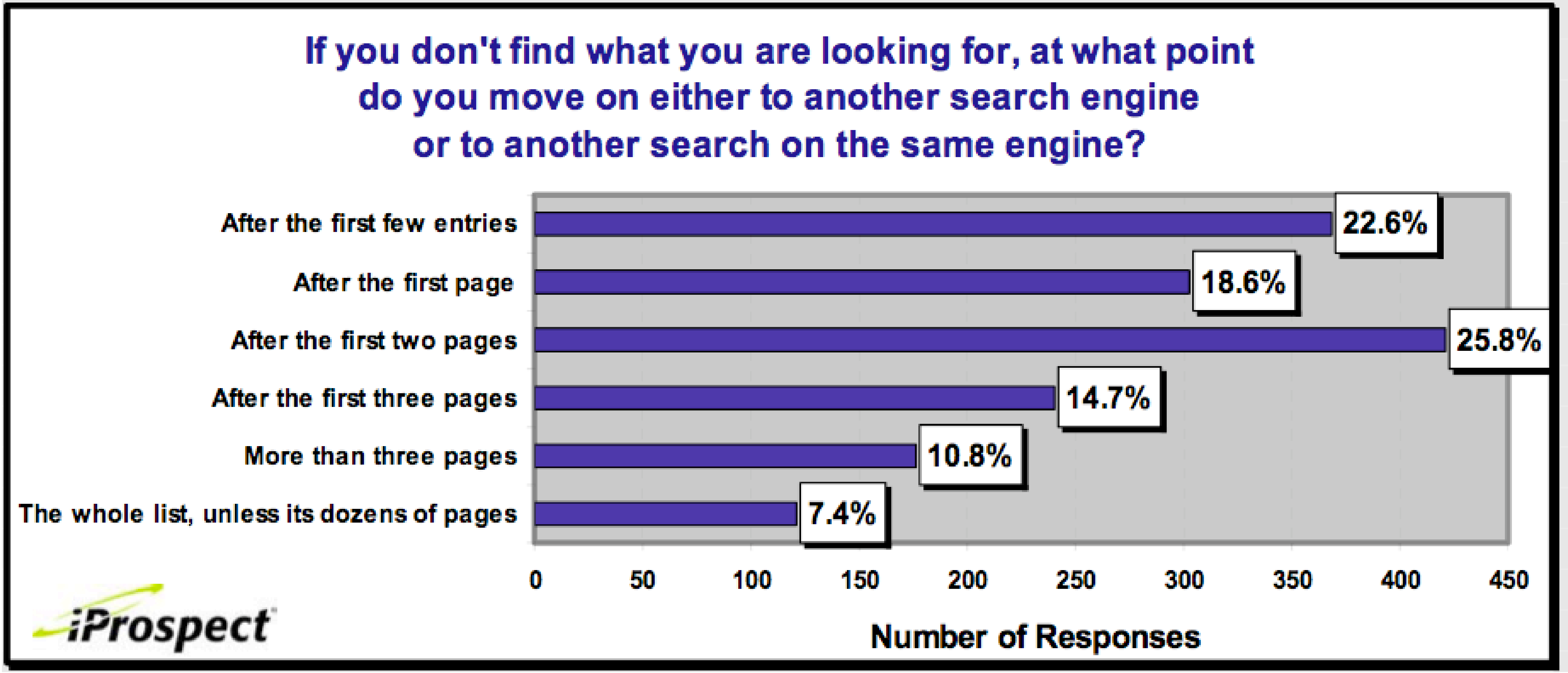
What is...

1. what is love
2. what is autism
3. what is rss
4. what is lupus
5. what is sap
6. what is bluetooth
7. what is emo
8. what is java
9. what is hpv
10. what is gout

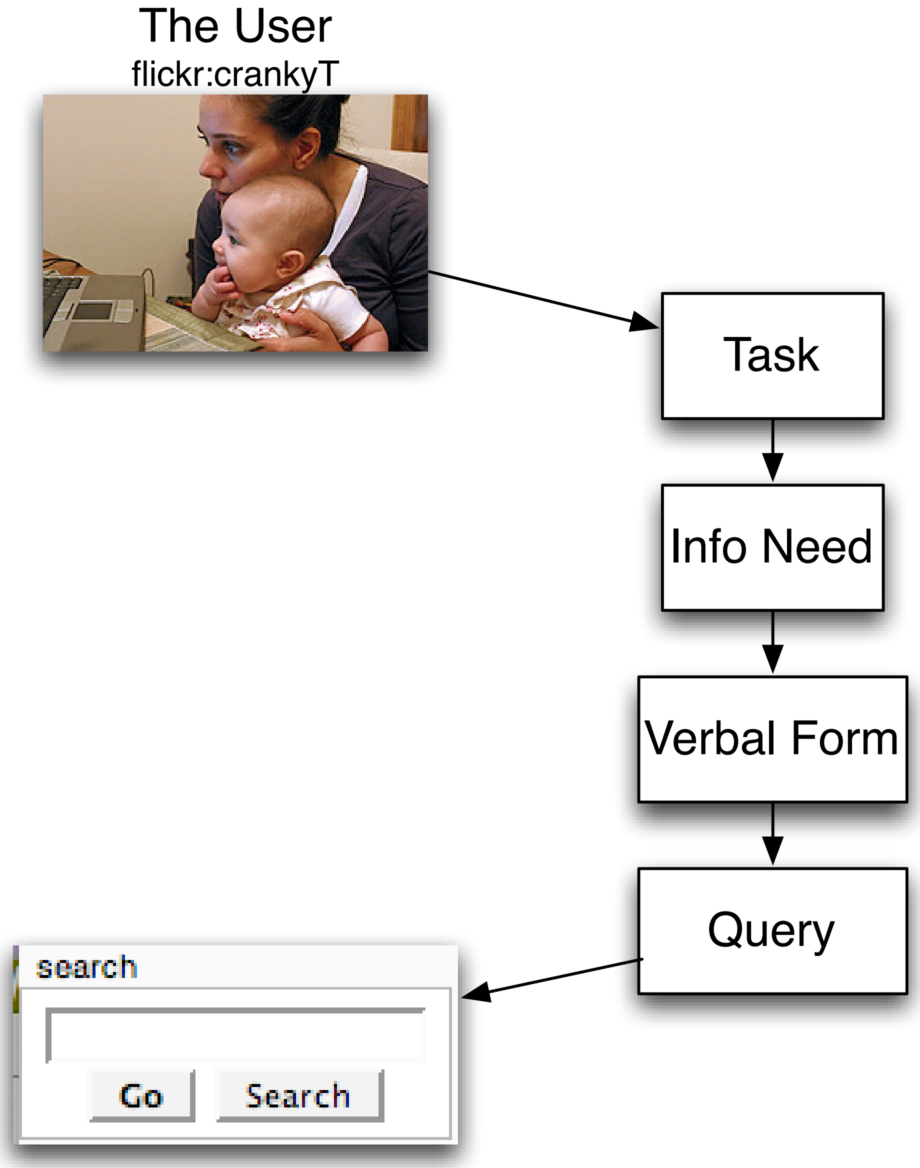
How to...

1. how to kiss
2. how to draw
3. how to knit
4. how to hack
5. how to dance
6. how to crochet
7. how to meditate
8. how to flirt
9. how to levitate
10. how to skateboard

How far do people look for results?



True Example *



Stop the noisy fan in the courtyard

Info about EPA regulations

What are EPA rules on noise pollution?

EPA Sound Pollution

“To Google or to GoTo” Business Week Online 9/28/2001

Web Users

How do users evaluate search engines?

- Quality of pages
    - Classic IR relevance
    - Also important:
        - Trust
        - Duplicate elimination
        - Readability
        - Fast Access
        - No pop-ups

Web Users

How do users evaluate search engines?

- Precision is more important than recall
    - Precision:
        - How precise is a portal in locating relevant results?
    - Recall
        - How thorough is the coverage of available relevant results?
- Precision with 1 result, 10 results, 2-3 pages of results.
- When is recall important?

Web Users

How do users evaluate search engines?

- Recall is sometimes important:
  - Googling for a new doctor
  - Googling a prospective employee
  - Googling your date

Web Users

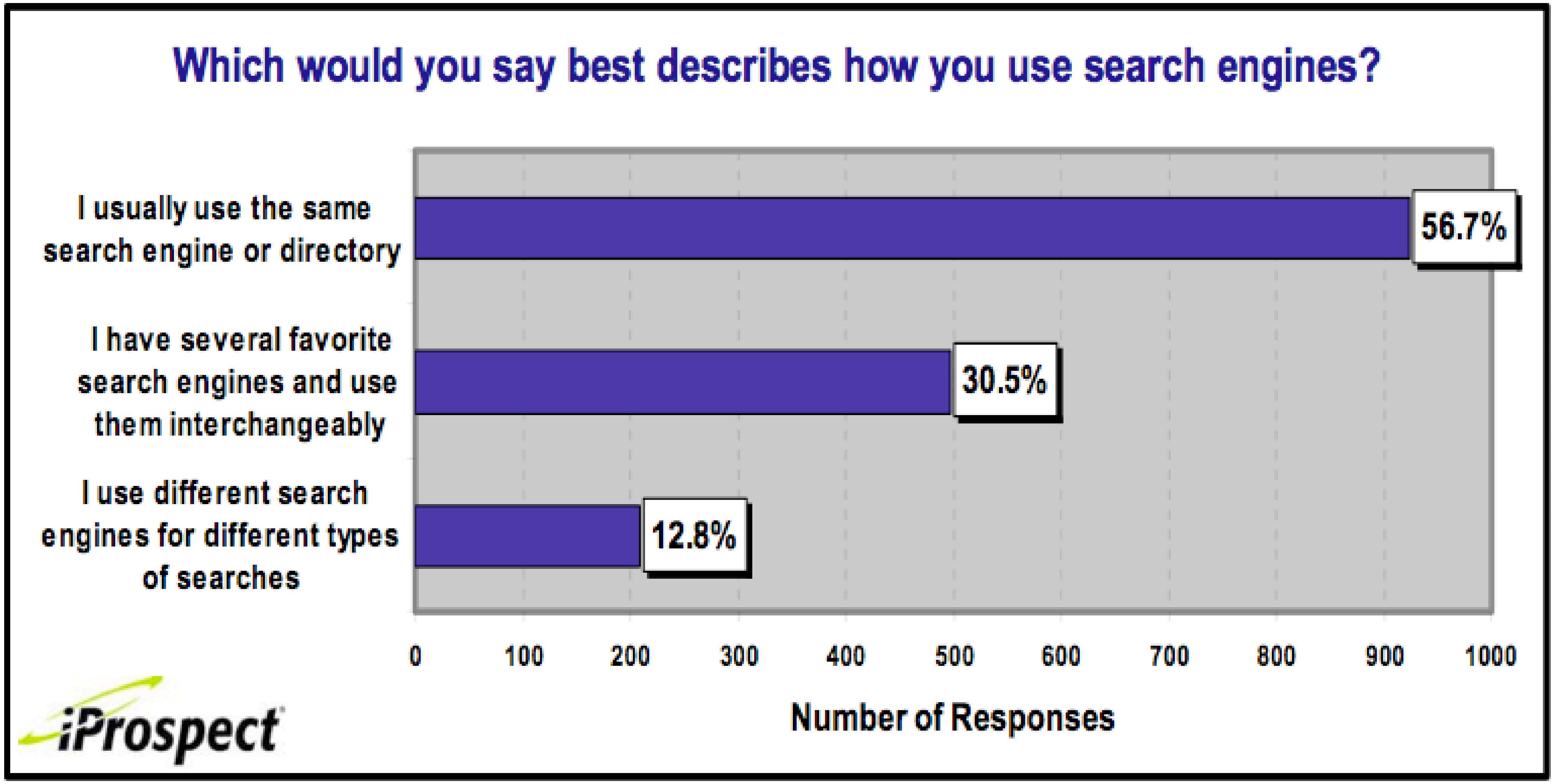
How do users evaluate search engines?

- Good U/I
  - Simple
  - No Clutter
- Pre and post processing tools
  - Spell check ("Did you mean ....?")
  - Suggested alternative searches
  - Links to resources (maps, images, stock quotes)
- Able to deal with typical behavior
  - e.g., a URL typed into a search box

Web Users

Loyalty to a given search engine

- iProspect Survey 4/2004


Which would you say best describes how you use search engines?
I usually use the same search engine or directory
I have several favorite search engines and use them interchangeably
I use different search engines for different types of searches
Number of Responses
iProspect