

Web Search Basics

Introduction to Information Retrieval

INF 141/ CS 121

Donald J. Patterson

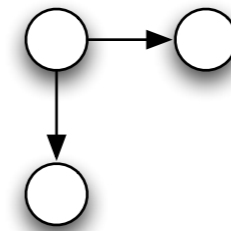
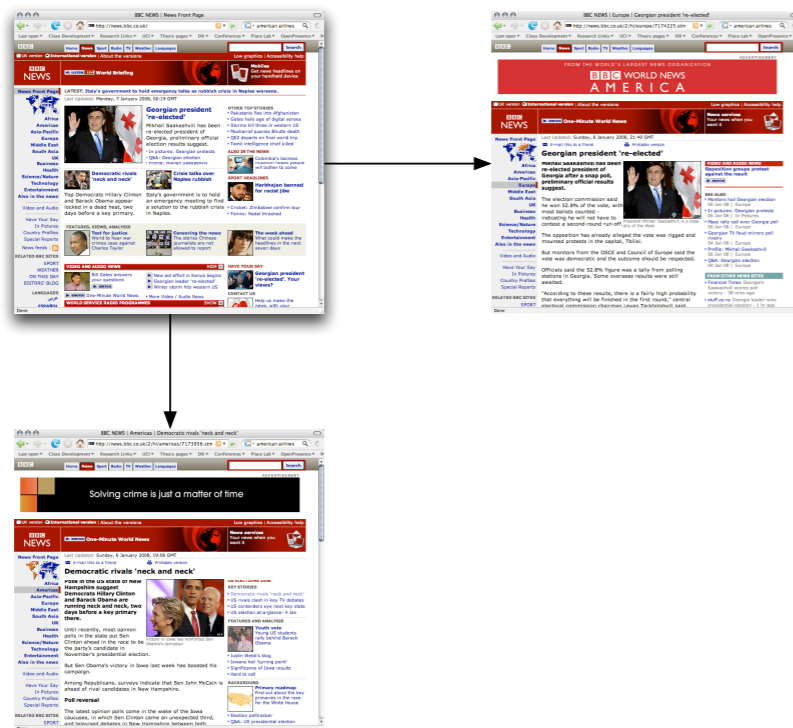
Content adapted from Hinrich Schütze

<http://www.informationretrieval.org>



The Web as a graph

- Web pages are nodes
- Hyperlinks are directed edges

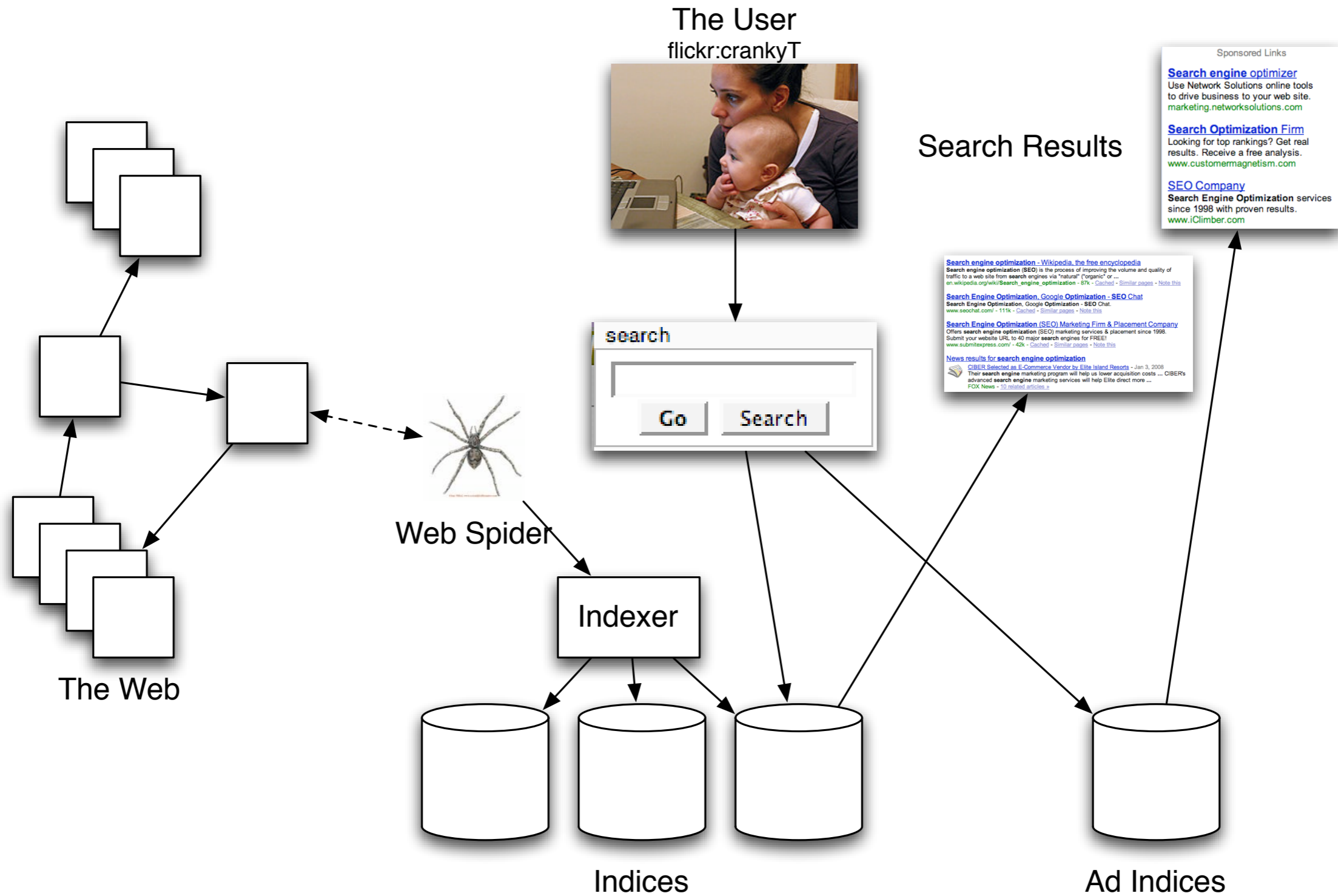


Characteristics of the web

- 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



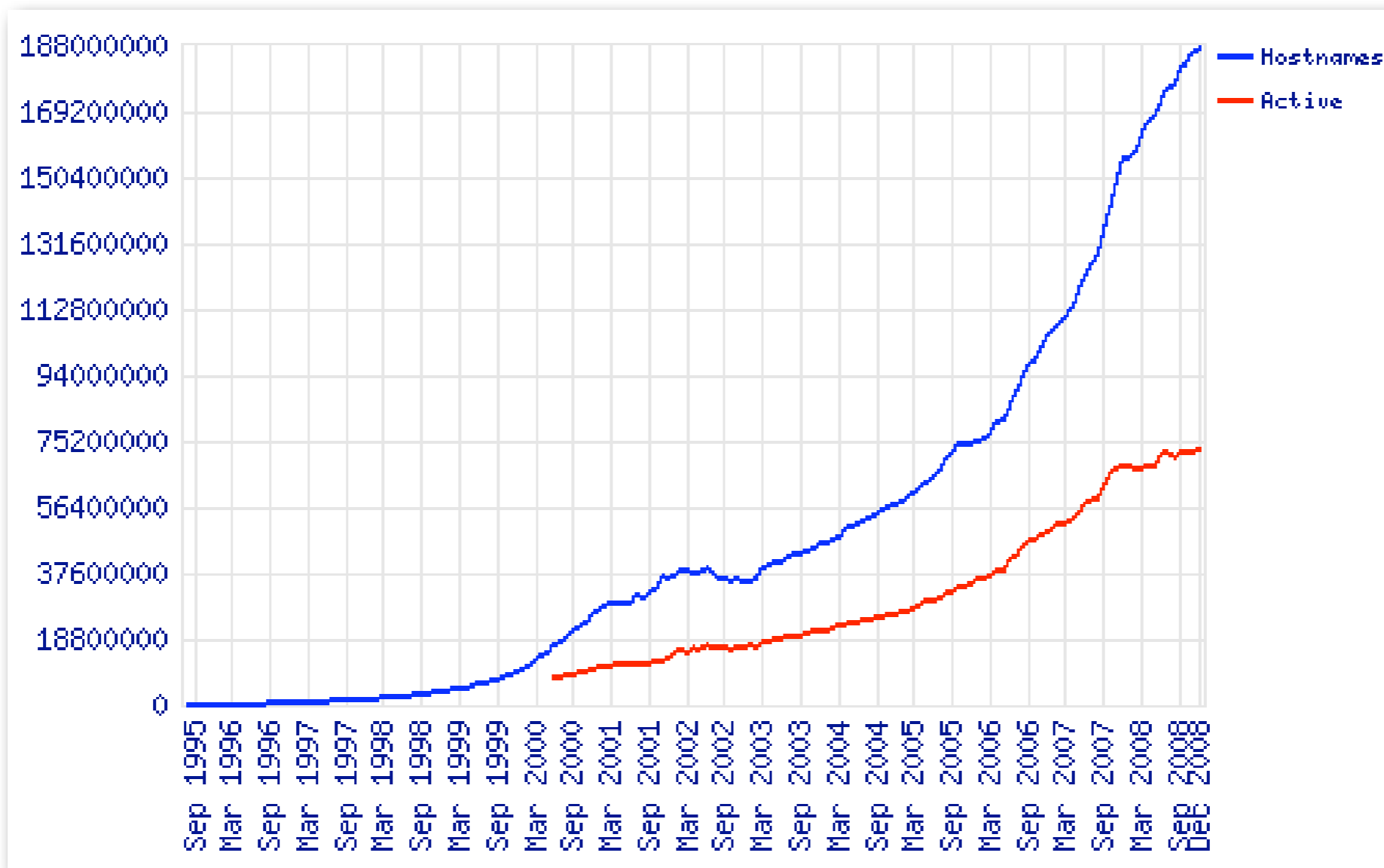
How big is the web?

- 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



How big is the web?

- Netcraft Web Server Survey



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



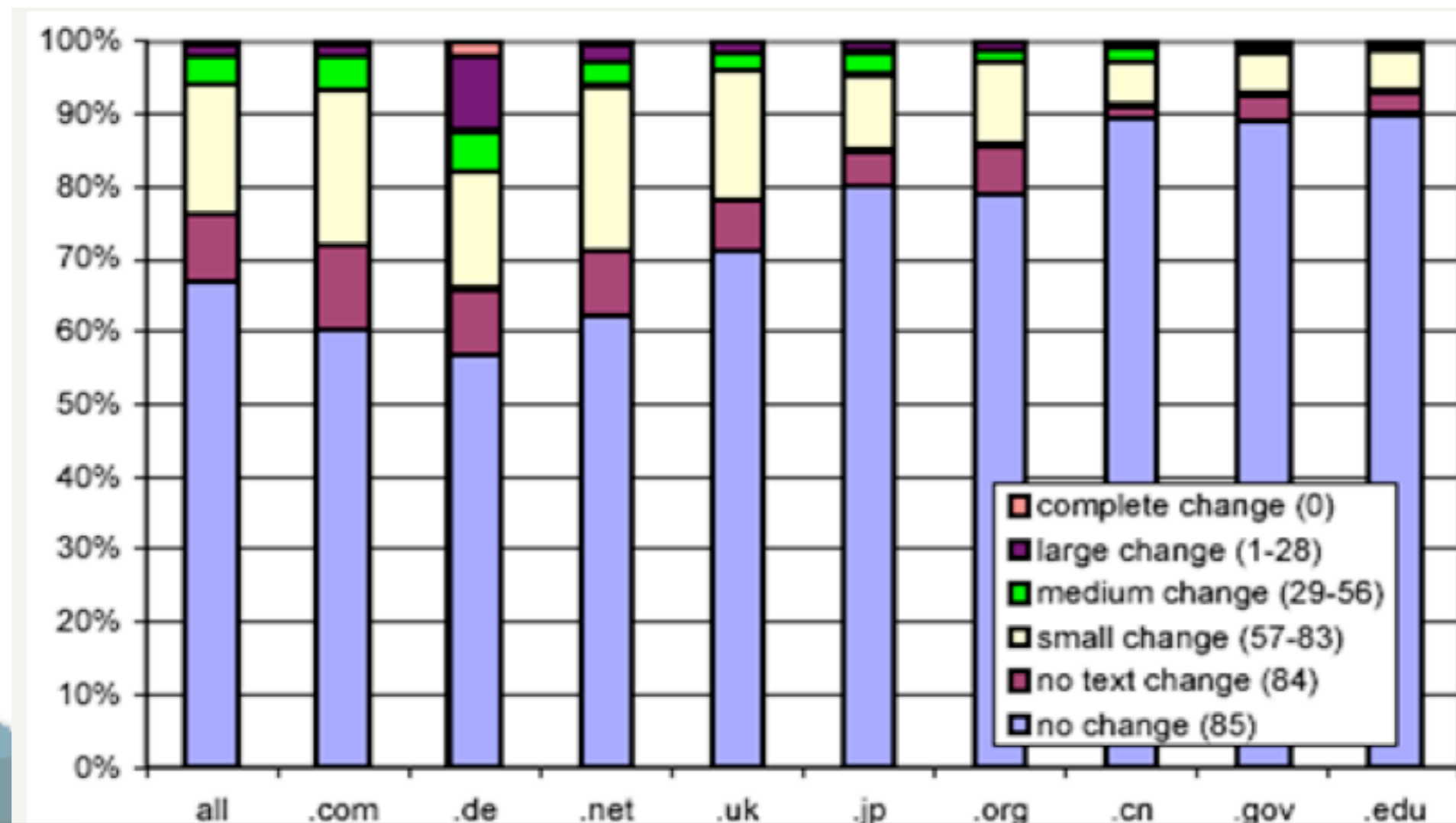
Rate of change

- [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]



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



User Search Needs in Brod02/RL04

- Informational
 - Want to learn about something (~40%/65%)



User Search Needs in Brod02/RL04

- Informational
 - Want to learn about something (~40%/65%)
- Navigational
 - Want to go to that page (~25%/15%)



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



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



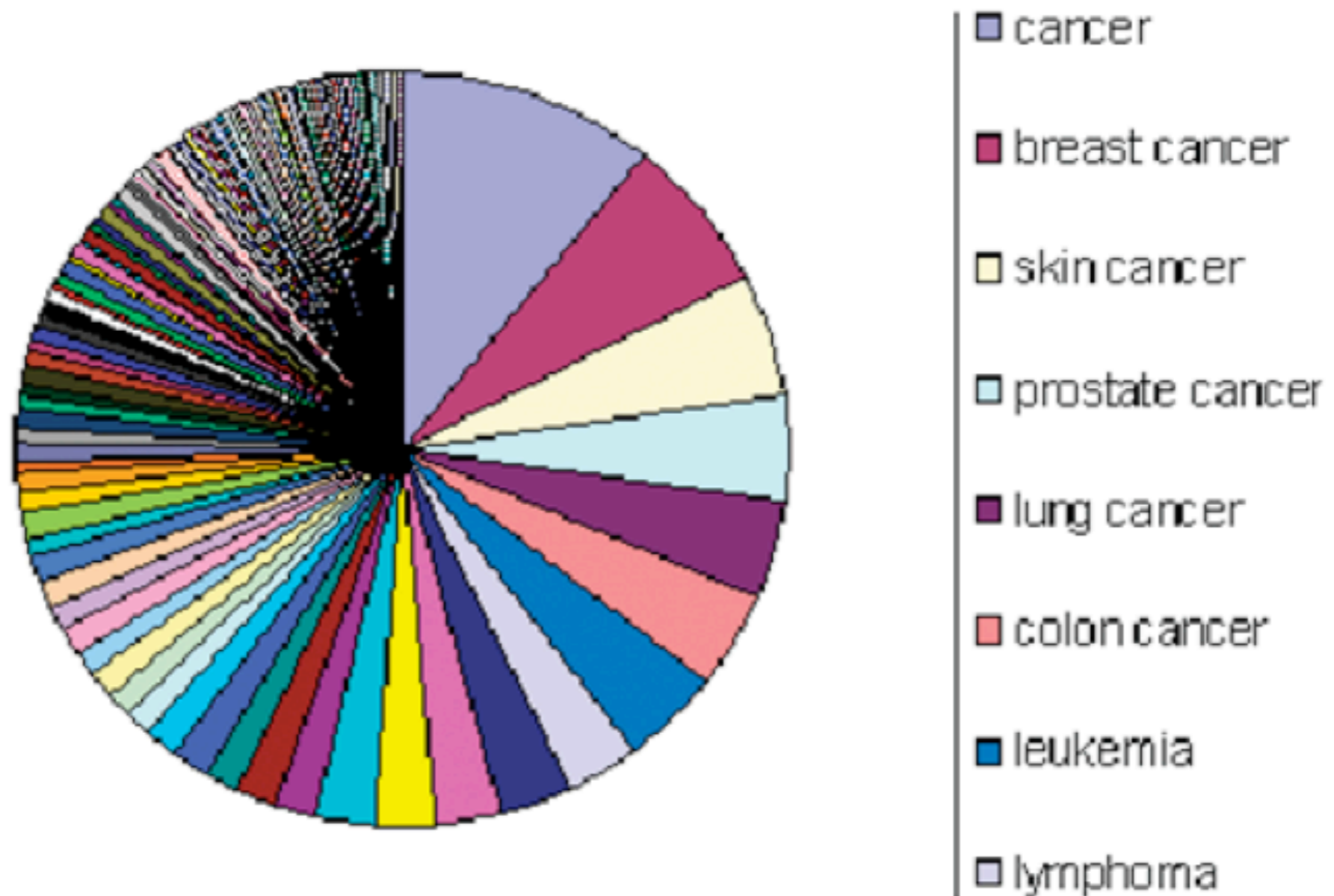
Web Users

- 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
- 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>

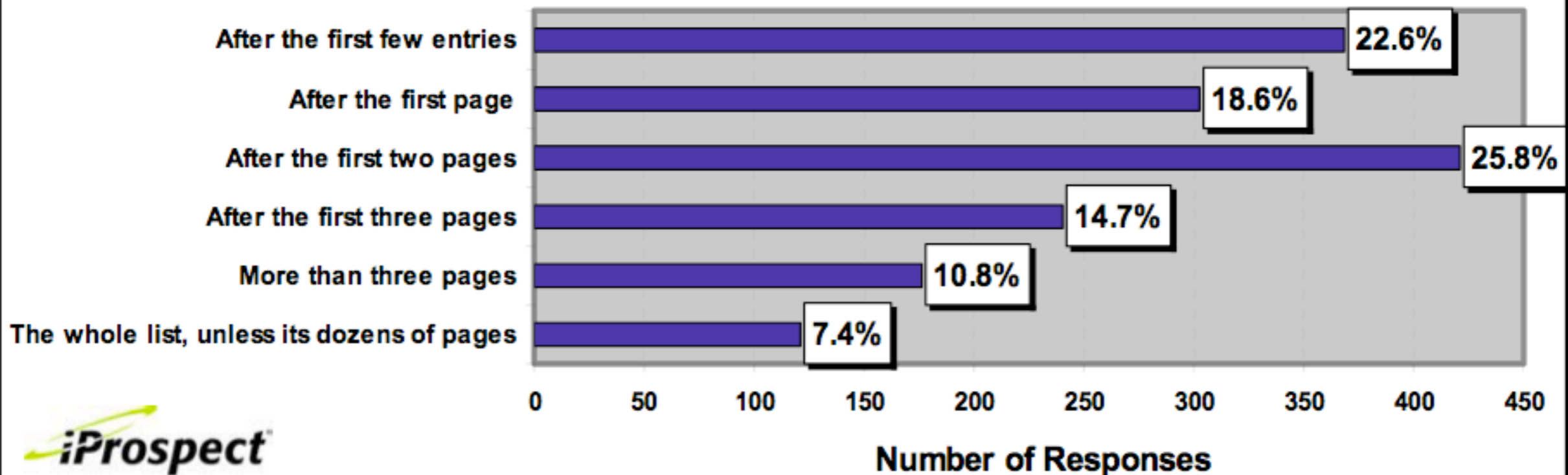
Top queries

- Live demo - WARNING this is not very safe....
- "Is it safe to"
- "Is it legal to"
- "why does"
- "why doesn't"
- "why is"
- "why isn't"
- "americans are"

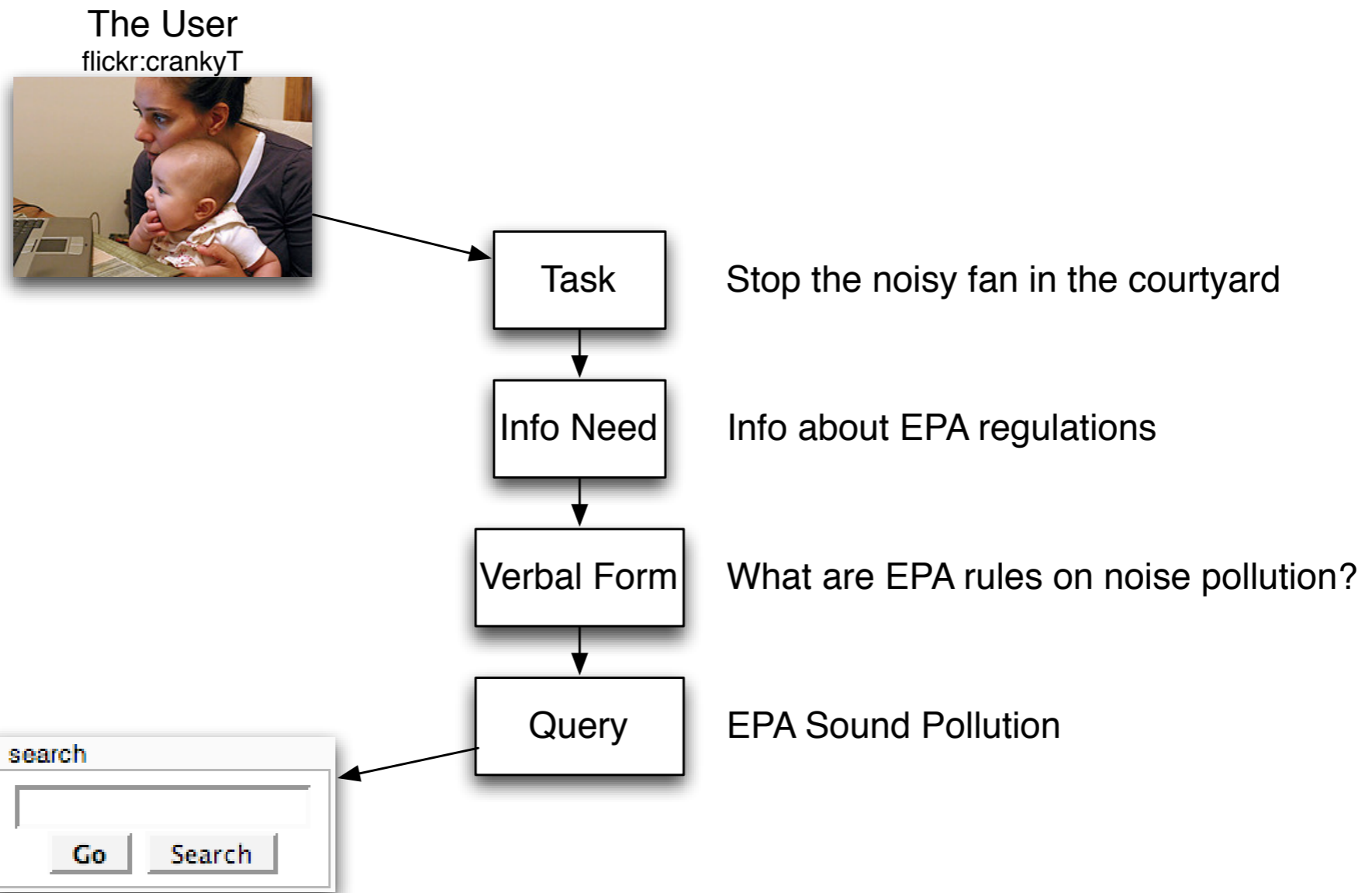


How far do people look for results?

If you don't find what you are looking for, at what point do you move on either to another search engine or to another search on the same engine?



True Example *



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

How do users evaluate search engines?

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



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?



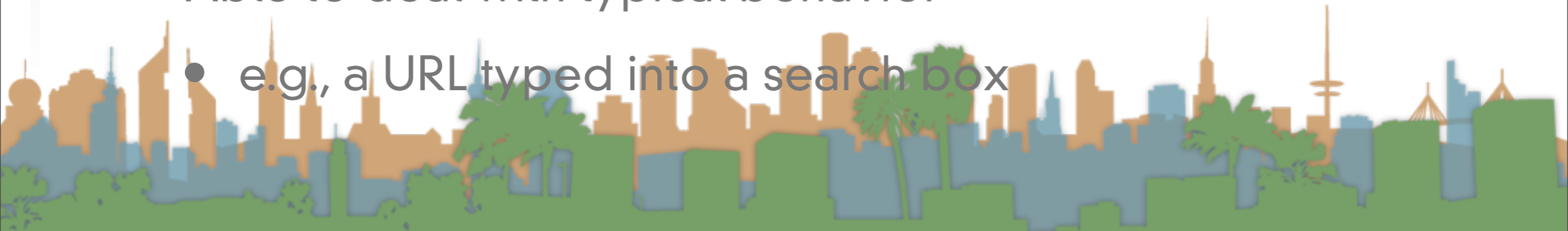
How do users evaluate search engines?

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



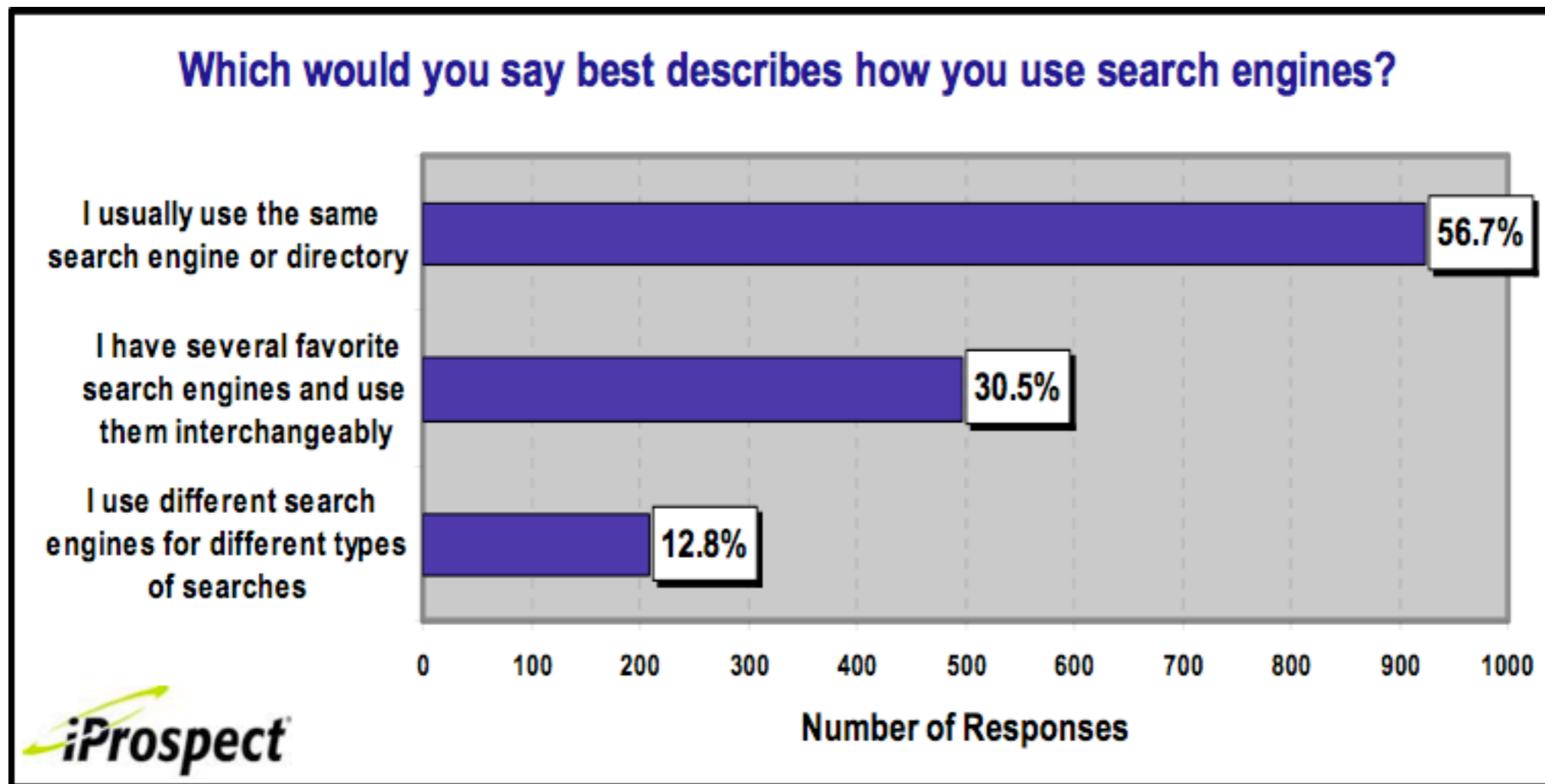
How do users evaluate search engines?

- 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



Loyalty to a given search engine

- iProspect Survey 4/2004

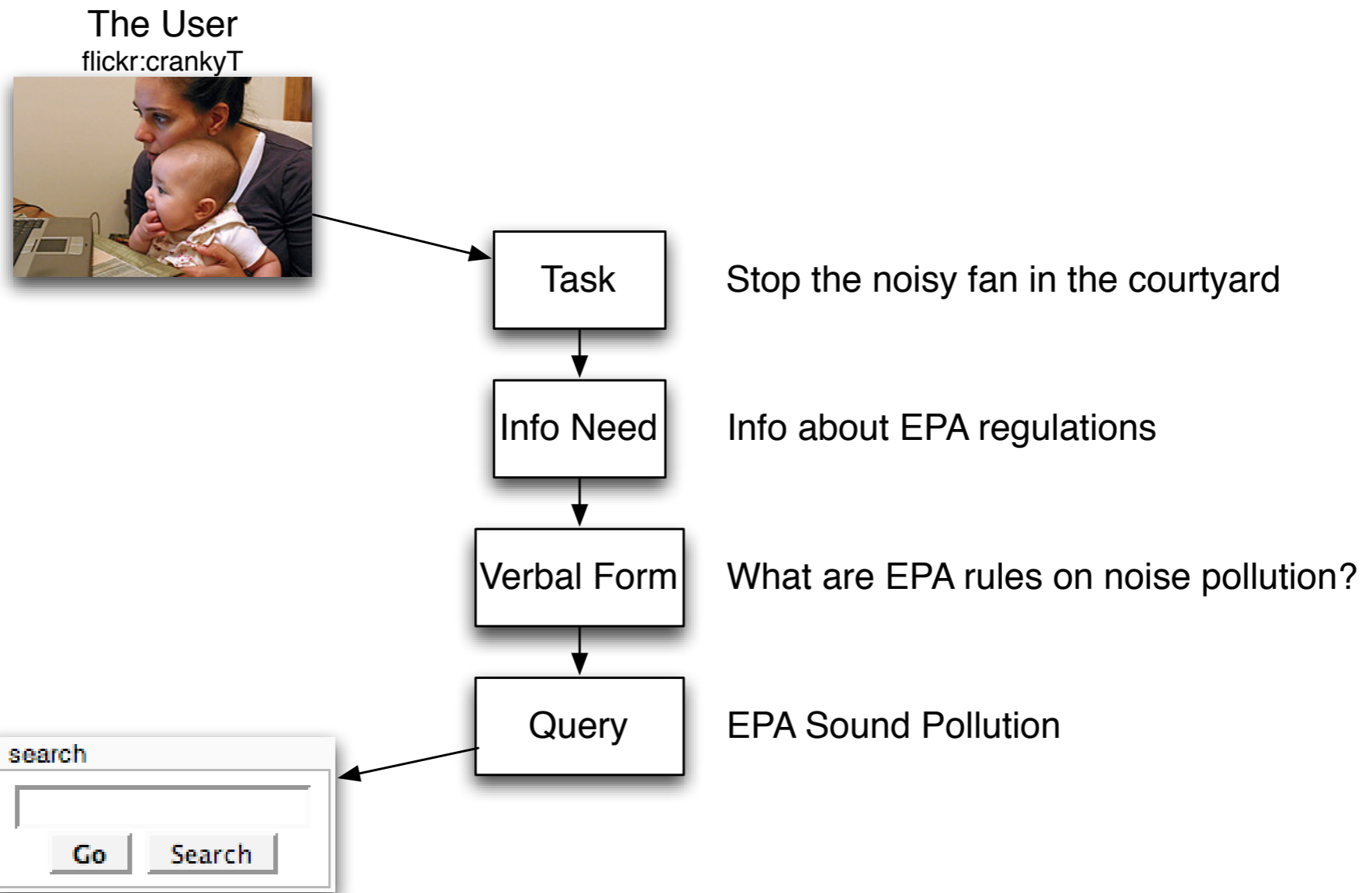


Overview

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



True Example *



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

Answering “the need behind the query”

- The query is often an imprecise indicator of what the user really wants
- What can we do to get a better handle on the underlying information need?
- Query language
 - Adjust rank of English results for a Japanese query
- Use user context
 - In particular geographic context



Answering “the need behind the query”

- Guess what type of information the user wants
 - a web page?
 - a map?
 - a stock price?
 - what else?
- Correct queries
 - Suggest correct spellings
 - Suggest related searches (google-fu)



Examples - language



The screenshot shows a web browser window titled "kitakurihama - Google Search". The address bar contains "http://www.google.com/search?hl=en&cli". The search bar contains "kitakurihama" and the search button is visible. The user's email address "donald.j.patterson.iii@gmail.com" and links for "Web History", "My Account", and "Sign out" are shown. The search results are for "Web" and show "Results 1 - 10 of about 276 for kitakurihama. (0.12 seconds)".

Did you mean: [kujukurihama](#)

[MySpace.com - Nasty - Kitakurihama, Kanagawa - Hip Hop - www ...](#)
MySpace music profile for Nasty with tour dates, songs, videos, pictures, blogs, band information, downloads and more.
[profile.myspace.com/index.cfm?fuseaction=user.viewprofile&friendid=1000891776](#) - 48k -
[Cached](#) - [Similar pages](#) - [Note this](#)

[Kitakurihama](#) - [[Translate this page](#)]
Kitakurihama. 1 to 7 out of 7 Glee land (グリーンランド) Karaoke box HAIR CLINIC TAIRA (ヘアクリニック タイラ) Hairdressing and cosmetics TAKOCHU (たこ忠) ...
[p.bari3.net/.../lookup2.asp?agntcd=1802&ct1=&ct1_nm=&ct2=007&ct2_nm=Kitakurihama&lv=1](#) - 5k -
[Cached](#) - [Similar pages](#) - [Note this](#)

[保育所ちびっこランド 北久里浜園](#) - [[Translate this page](#)]
混合自由保育で三育法を実践。
[www.kidslink.jp/chibikko-kitakurihama/](#) - 2k - [Cached](#) - [Similar pages](#) - [Note this](#)

[YuDiary](#)
i waited for my friend on **kitakurihama** station at 1:00pm. there are a lot of fields in misakiguchi. i drove my car from **kitakurihama** station to his house. ...
[idauyutajifu.blogspot.com/](#) - 61k - [Cached](#) - [Similar pages](#) - [Note this](#)

Examples - query spelling

infermatics - Google Search

http://www.google.com/search?hl=en&client=firefox-a&rls=org.mozilla keitei

Last open ▾ Class Development ▾ Research Links ▾ UCI ▾ Thesis pages ▾ DR ▾ Conferences ▾ Place Lab ▾ OpenPresence ▾

Web Images Maps News Shopping Gmail more ▾ donald.j.patterson.iii@gmail.com | Web History | My Account | Sign out

Google™ infermatics Search [Advanced Search](#) [Preferences](#)

Web Results 1 - 10 of about 87 for infermatics. (0.13 seconds)

Did you mean: [informatics](#)

[DÜŞÜNCELER - dellal international infermatics - Blogcu](#)
dellal international **infermatics**. emrah dellal. Bağlantılar. Ana Sayfa · Profilim · Arşiv · RSS.
Son Yazılar. DÜŞÜNCELER. 18/6/2006 - DÜŞÜNCELER ...
[mavilimm.blogcu.com/717078/](#) - 10k - [Cached](#) - [Similar pages](#) - [Note this](#)

[dellal international infermatics - dellal.com.tr.tc - www dellal ...](#)
emrah dellal,arkadaş arama,sohbet,google,, dellal com tr tc, dellal.com.tr.tc.
[www.dellal.com.tr.tc/](#) - 2k - [Cached](#) - [Similar pages](#) - [Note this](#)

[LexiconWiki: Infermation](#)
... the academic community, promoting and assessing the use of Infermation and producing
dedicated **Infermatics** for both academic and general consumption. ...
[kevan.org/lexwiki.pl?action=browse&diff=1&id=Infermation](#) - 8k -
[Cached](#) - [Similar pages](#) - [Note this](#)

[Hepatology Research : Medical economics—Hepatitis C virus ...](#)
Medical **Infermatics** and Decision Sciences, Yamaguchi University, School of Medicine,
1-1-1 Minami-Kogushi, Ube, Yamagnchi: 755-8505, Japan ...
[linkinghub.elsevier.com/retrieve/pii/S1386634602000414](#) - [Similar pages](#) - [Note this](#)

[IngentaConnect Medical economics-Hepatitis C virus infection and ...](#)

Done

Examples - query expansion

The screenshot shows a browser window titled "Live Search: rock". The address bar contains the URL "http://search.live.com/results.aspx?q=rock&go=Search&mkt=en-l". The search bar contains the text "rock" and a "Search" button. Below the search bar, there are navigation links: "Home", "Hotmail", "Spaces", "OneCare", and "Sign in".

Web results 1-10 of 441,000,000
See also: [Images](#), [Video](#), [News](#), [Maps](#), [MSN](#), [More](#) ▼

The Rock - [Jellyfish.com/Movies](#) Sponsored sites
Shop and compare prices across stores. Rebates up to \$1.10.

Rock.com® - The Official Site of Rock Music®
Free Email & Free Music Downloads by **Rock.com**. Get Free Internet Radio, Music Downloads & more. Get your free @rock.com email today
[www.rock.com](#) · [Cached page](#)

Rock music - Wikipedia, the free encyclopedia
Rock music is a form of popular music with a prominent vocal melody accompanied by guitar , drums , and bass . Many styles of **rock** music also use keyboard instruments such as organ ...
[en.wikipedia.org/wiki/Rock_music](#) · [Cached page](#)

Rock and roll - Wikipedia, the free encyclopedia
Rock and roll is a genre of music that evolved in the United States in the late 1940s and early 1950s, and quickly spread to the rest of the world.
[en.wikipedia.org/wiki/Rock_and_roll](#) · [Cached page](#)
[Show more results from en.wikipedia.org](#)

Rock - Australia's Climbing Magazine
Subscription information. Overview of current issue. Classifieds. Back issues and guide books for sale.

Related searches
[Rock Music](#)
[Rocks Minerals](#)
[Pictures Of Rocks](#)
[Rock Star](#)
[Kid Rock](#)
[Rock Cycle](#)
[Hard Rock Cafe](#)

Sponsored sites
Free Rock Music
Listen To Your Favorite Music Now!
It's Free with the Music Toolbar
[Music.alot.com](#)

The Rock
Interviews with Stars, Newsmakers & Icons at The Biography Channel®.
[www.biography.com](#)

Rock Videos
Enjoy Savings & Selection on **Rock** Videos!
[Shopzilla.com](#)

Done

Query shortcuts

- Map: "irvine, ca 92614"
- Calculation: "5+4"
- Flight Info: "american airlines 715"
- Stock price: "msft"
- Unit conversion: "1 dollar in euro"
- Music: "White Stripes"



Examples - query shortcuts

The screenshot shows a web browser window titled "White Stripes - Yahoo! Search Results". The address bar contains the URL http://search.yahoo.com/search;_ylt=A0oGkmbxeZdH0nsBWtZXNyo. The search bar contains the text "White Stripes" and a yellow "Search" button. The page displays search results for "White Stripes", including a featured snippet for the official site, a "Play Popular Songs" section, and several sponsored results for ringtones and music.

White Stripes - Yahoo! Search Results

http://search.yahoo.com/search;_ylt=A0oGkmbxeZdH0nsBWtZXNyo

american airlines 715 - Googl... Live Search: white stripes White Stripes - Yahoo! Search ...

Web | Images | Video | Local | Shopping | more

White Stripes Search Options

YAHOO!

1 - 10 of about 53,700,000 for White Stripes (About this page) - 0.11 sec.

The White Stripes - Official Site
www.whitestripes.com
Albums | Lyrics | Photos | Videos

Play Popular Songs

- Icky Thump
- You Don't Know What Love Is (You Jus...
- 300 M.P.H. Torrential Outpour Blues

More Songs...

Watch Music Videos

Yahoo! Shortcut - [About](#)

The White Stripes
Official site for The **White Stripes**, the Detroit-based garage/blues duo featuring Jack and Meg **White**.
www.whitestripes.com - 2k - [Cached](#)

White Stripes v2
Jack and Meg **White** fansite includes news and tour information, forum, links, release details, and more

SPONSOR RESULTS

White Stripes Ringtones
Complimentary **White Stripes** Ringtones. Get Them Instantly.
www.TheBombRingtones.com

White Stripes Ringtones
White Stripes Ringtones. Get Them Complimentary Now.
ChartRingers.com/whitestripes

The White Stripes
Find, compare & buy. Compare & Buy from 1000's of Stores.
www.Dealtime.com

The White Stripes at Amazon.com
Low prices on new & used music. Qualified orders over \$25 ship free.
Amazon.com/music

Examples - query shortcuts

The screenshot shows a browser window titled "Live Search: Don Patterson irvine ca". The address bar contains the URL "http://search.live.com/results.aspx?q=Don+Patterson+irvine+ca&gr". The browser has several tabs open, including "american airlines 715 - Googl...", "Live Search: Don Patterson irvin...", and "White Stripes - Yahoo! Search ...". The search results page displays the query "Don Patterson irvine ca" in a search box, with a "Search" button and links for "Advanced" and "Options".

Web results 1-9 of 253,000
See also: [Images](#), [Video](#), [News](#), [Maps](#), [MSN](#), [More](#) ▼

» [Top local listings for Don Patterson near Irvine, CA](#) (1 more) Is this useful? [Yes](#) | [No](#)

- **Don Patterson** (626) 918-1974 15626 Dubesor St, La Puente
- **Don Patterson** (213) 389-7583 226 S Berendo St Apt 205, Los Angeles
- **Don Patterson** (760) 729-6941 2768 Dundee Ct, Carlsbad

[DBLife: University of California-Irvine](#)
Jadwiga Indulska, University of Queensland, Australia **Don Patterson**, University of California Irvine , USA Tom Rodden, Nottingham University, UK Program Committee:
dblife.cs.wisc.edu/search.cgi?entity=entity-13655&begin=60 · [Cached page](#)

[DBLife: University of California-Irvine](#)
Jadwiga Indulska University of Queensland, Australia **Don Patterson** University of California Irvine , USA Tom Rodden Nottingham University, UK...
dblife.cs.wisc.edu/search.cgi?entity=entity-13655&begin=90 · [Cached page](#)
[Show more results from dblife.cs.wisc.edu](#)

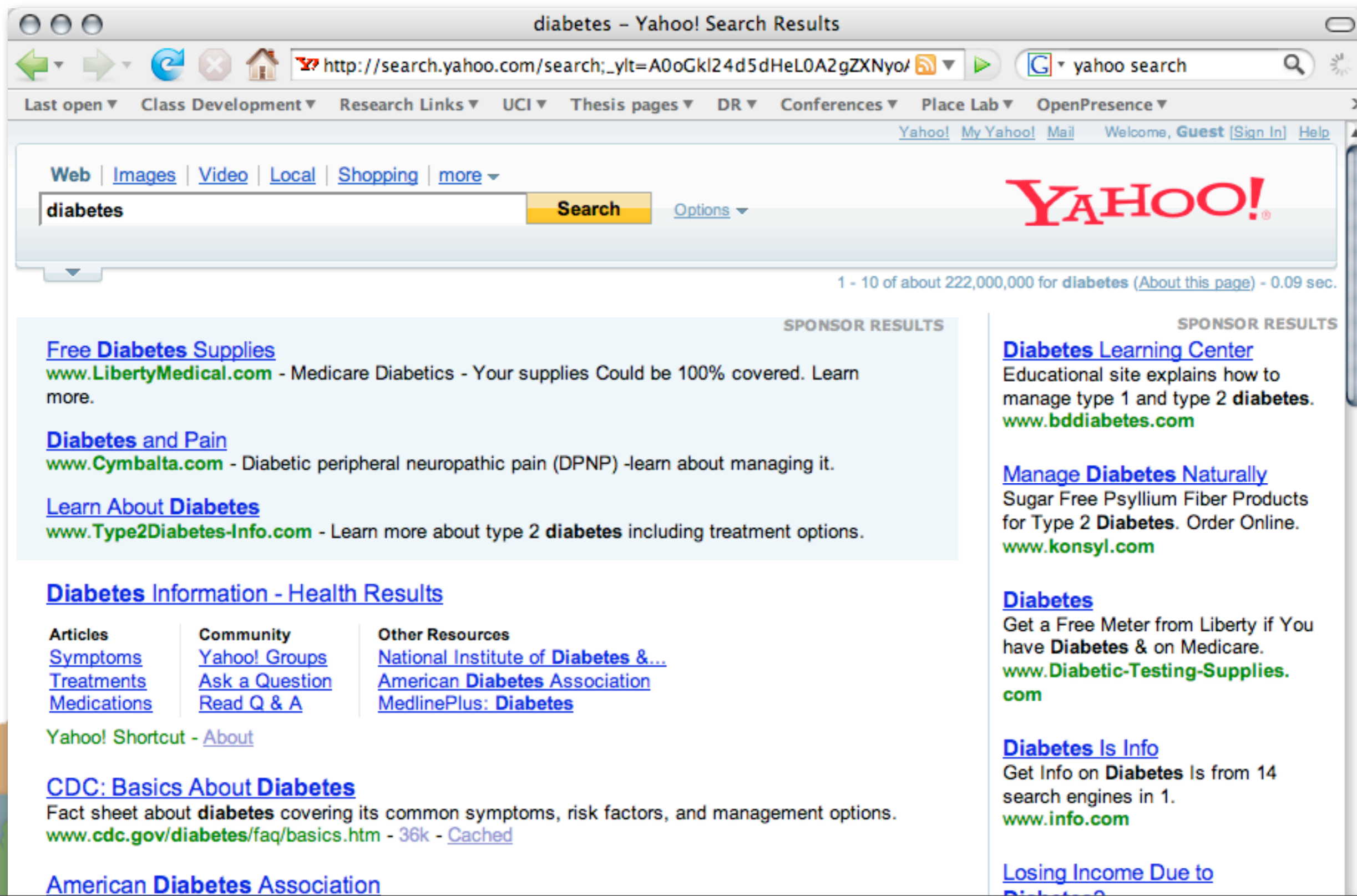
[Jerry M. Patterson](#)
Professor: University of California at Irvine (2000-) Moulton Patterson Associates ... Partner (1967-75) State Bar of California 1967. Do you know something we don't?
www.pndb.com/people/406/000163914 · [Cached page](#)

Sponsored sites

- [Irvine CA Book Hotel](#)
Visiting California & Need a Hotel?
Reserve Rooms for Irvine CA!
www.calibex.com/hotels
- [Locate Don Patterson](#)
Current address and phone available.
Instant results.
www.usa-people-search.com
- [Don Patterson at Amazon](#)
Low prices on new & used music.
Qualified orders over \$25 ship free
amazon.com/music
- [Irvine CA - Cheap Rates](#)
Visiting California & Need a Hotel?
Compare Top Sites for Irvine CA!
www.nextag.com/hotels

[See your message here...](#)

Examples - query aggregations



The screenshot shows a web browser window titled "diabetes - Yahoo! Search Results". The address bar contains the URL "http://search.yahoo.com/search;_ylt=A0oGkl24d5dHeL0A2gZXNyof". The search bar contains the text "diabetes" and a "Search" button. The page displays search results for "diabetes", including sponsored results and health information.

diabetes - Yahoo! Search Results

http://search.yahoo.com/search;_ylt=A0oGkl24d5dHeL0A2gZXNyof

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

Web | Images | Video | Local | Shopping | more

diabetes Search Options

1 - 10 of about 222,000,000 for diabetes (About this page) - 0.09 sec.

SPONSOR RESULTS

Free Diabetes Supplies
www.LibertyMedical.com - Medicare Diabetics - Your supplies Could be 100% covered. Learn more.

Diabetes and Pain
www.Cymbalta.com - Diabetic peripheral neuropathic pain (DPNP) -learn about managing it.

Learn About Diabetes
www.Type2Diabetes-Info.com - Learn more about type 2 diabetes including treatment options.

Diabetes Learning Center
Educational site explains how to manage type 1 and type 2 diabetes.
www.bddiabetes.com

Manage Diabetes Naturally
Sugar Free Psyllium Fiber Products for Type 2 Diabetes. Order Online.
www.konsyl.com

Diabetes
Get a Free Meter from Liberty if You have Diabetes & on Medicare.
www.Diabetic-Testing-Supplies.com

Diabetes Is Info
Get Info on Diabetes Is from 14 search engines in 1.
www.info.com

Diabetes Information - Health Results

Articles	Community	Other Resources
Symptoms	Yahoo! Groups	National Institute of Diabetes &...
Treatments	Ask a Question	American Diabetes Association
Medications	Read Q & A	MedlinePlus: Diabetes

Yahoo! Shortcut - [About](#)

CDC: Basics About Diabetes
Fact sheet about diabetes covering its common symptoms, risk factors, and management options.
www.cdc.gov/diabetes/faq/basics.htm - 36k - [Cached](#)

American Diabetes Association

Losing Income Due to Diabetes