# Needles in the Haystack: Google and Other Brokers in the Bits Bazaar

Derived from Blown To Bits Chapter 4
with the same title

# CS Concepts

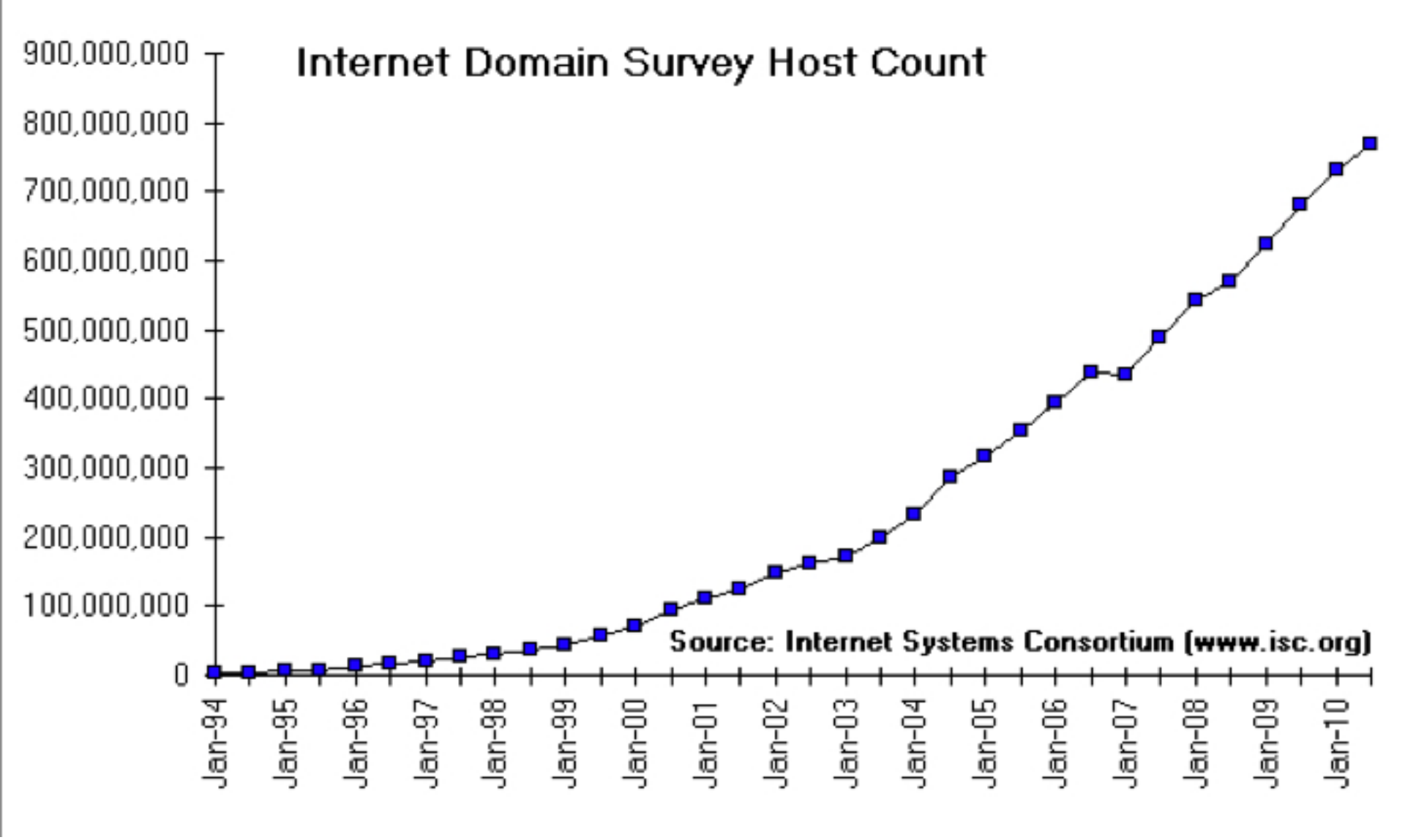
- search engine
- web crawler
- page rank
- primary vs secondary memory access time
- binary search
- who pays for “free” searching

# Social Issues

- search as censor
- right to be forgotten
- can’t take it back
- free speech
- regulate search engines?

“The search tools that help us find needles in the digital haystack have become the lenses through which we view the digital landscape. Businesses and governments use them to distort our picture of reality.”

Blown to Bits - pg 110


Internet Domain Survey Host Count
900,000,000
800,000,000
700,000,000
600,000,000
500,000,000
400,000,000
300,000,000
200,000,000
100,000,000
0
Source: Internet Systems Consortium (www.isc.org)
Jan-94
Jan-95
Jan-96
Jan-97
Jan-98
Jan-99
Jan-00
Jan-01
Jan-02
Jan-03
Jan-04
Jan-05
Jan-06
Jan-07
Jan-08
Jan-09
Jan-10

# Number of pages?

Searching for “travel” in Google today yielded “About 3,300,000,000 results (0.47 seconds)”

In “The New Internet Navigator” by P. Gilster, (1995) I found an example where searching for “travel” yielded “a *whopping* 2,633 documents.”

At least 40-50 billion pages (www.worldwidewebsize.com)

# The Library and the Bazaar

- “Yellow pages”, directories, and catalogues
- The “Web” is not hierarchical (no structure like a library)
- Catalogues are out - search engines are in.
- But - search engines control what you see

“For the user, search is the power to find things, and for whoever controls the engine, search is the power to shape what you see.”

Blown to Bits pg 112

# Where to look?

Google is not necessarily the first place to look!

- Go directly to a Web site -- www.irs.gov
- Go to your bookmarks -- dictionary.cambridge.org
- Go to the library -- library.ucsc.edu
- Go to the place with the information you want -- www.npr.org
- Browser address window remembers where you've been.

Ask, "What site provides this information?"

Different search engines will produce different results.

# Advanced Search

Advanced Search Tips |

Use the form below and your advanced search will appear here

**Find web pages that have...**

all these words:

this exact wording or phrase: tip

one or more of these words: OR OR tip

**But don't show pages that have...**

any of these unwanted words: tip

**Need more tools?**

Reading level: no reading level displayed

Results per page: 10 results

This option does not apply in Google Instant.

Language: any language

File type: any format

Search within a site or domain:

(e.g. youtube.com, .edu)

Date, usage rights, region, and more

Advanced Search

# Keyword Search

Search Engine words are independent

- Words don't have to occur together

Use Boolean queries and quotes

- Logical Operators: AND, OR, NOT

  monet AND water AND lilies

  "van gogh" OR gauguin

  vermeer AND girl AND NOT pearl

# Keyword Search

Searching strategies …

- Limit by top level domains or format … .edu
- Find terms most specific to topic … ibuprofen
- Look elsewhere for candidate words, e.g. bio
- Use exact phrase only if universal, … “Play it again”
- If too many hits, re-query … let the computer work
- “Search within results” using “-” … to get rid of junk

- How can a search engine respond so fast?
- Does it find every relevant link?
- How does a search engine decide what gets listed first?
- If you try another search engine will you get the same result? If so, which is right? Which is better? Which is more authoritative?
- Are sponsored links better than “organic” links? Is the advertising necessary?
- What is the role of government? What should it be?

# Lilly and Zyprexa

- treatment for schizophrenia
- possible serious side effects
- lawyer leaked a confidential internal Lilly memo
- Judge ruled lawyer acted improperly but “the bits had escaped and could not be recaptured,” (Blown to bits pg 116)

"Web sites are primarily fora for speech… the risk of unlimited inhibitions of free speech should be avoided when practicable."

Judge Weinstein - in refusing to order web sites to remove copies of the leaked memo.

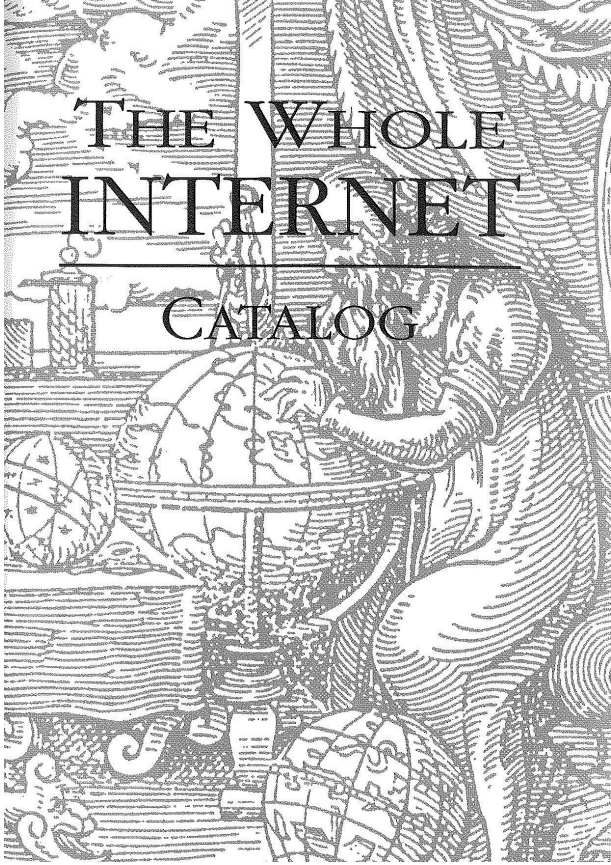
European Commission

# Factsheet on the "Right to be Forgotten" ruling *(C-131/12)*

## 1) What is the case about and what did the Court rule?

In 2010 a Spanish citizen lodged a complaint against a Spanish newspaper with the national Data Protection Agency and against Google Spain and Google Inc. The citizen complained that an auction notice of his repossessed home on Google's search results infringed his privacy rights because the proceedings concerning him had been fully resolved for a number of years and hence the reference to these was entirely irrelevant. He requested, first, that the newspaper be required either to remove or alter the pages in question so that the personal data relating to him no longer appeared; and second, that Google Spain or Google Inc. be required to remove the personal data relating to him, so that it no longer appeared in the search results.

# The Fall of Hierarchy




- Early catalogues both on the Internet (even before the WWW) and in print.
- “The Whole Internet Catalog” (1994)
- Yahoo was one compiled by human editors.
- Search engines started to be used in early 90s with the growing popularity of the web.

Which of these is an example of information that is NOT hierarchical?

A. A university course catalogue.
B. Classification of animals (family, genus, species, etc.)
C. Telephone numbers.
D. The location of books in a book store.
E. Your network of friends.

# AGRICULTURE

**Newsgroups:**
alt.agriculture.[fruit, misc], misc.rural

**See also:** Forestry; Gardening; and Horticulture.

## Advanced Technology Information Network

Any farmer knows that farming isn't a "mom and pop" business any more; it's high-tech, and it's important to keep up with the latest developments. This resource, and the others in this group, will help you stay up-to-date. A fairly complete agricultural information service offers market, news, events, weather, job listing, and safety information. Offered by the California Agricultural Technology Institute, so there is a "West Coast" bias to the information. Also contains information on trade, exports, and biotechnology.

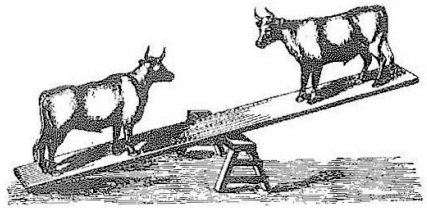
**Access via:**
*telnet caticsuf.csufresno.edu*; login *super*

## Commodity Market Reports

Commodity reports compiled by the U.S. Department of Agriculture Market News Service. Twelve hundred reports covering the U.S., updated daily.

**Access via:**
WAIS *agricultural-market-news.src*

**Information:**
Email: wais@oes.orst.edu



## Not Just Cows

A guide to resources on the Internet and BITNET in agriculture and related subjects. Compiled by Wilfred (Bill) Drew.

**Access via:**
*ftp ftp.sura.net*; login *anonymous*; cd *pub/nic*; get *agricultural.list*

**Information:**
Email: drewwe@snymorva.bitnet

## PEN Pages

A complete information server concerning all aspects of rural life. Sections on commodity prices, family farm life, seniors on the farm, news, and nutrition. Also, provides various announcements by the USDA including its CITExtension newsletter. Service provided by the Pennsylvania State University, so some information may be specific to that region.

**Access via:**
*telnet psupen.psu.edu*; login your two-letter state abbreviation

## U.C. Davis Extension 4-H Project Catalog

Intended to help members of the 4-H youth project get started in areas ranging from beekeeping to "poultry science," these files are available in PostScript and WordPerfect 5.1 formats.

**Access via:**
*gopher gopher.ucdavis.edu /The Campus/ U. C. Cooperative Extension/4h-youth*

*ftp ftp.ucdavis.edu*; login *anonymous*; cd *pub/extension/4h-youth*

## USDA Extension Service Gopher

A "master gopher" for the U.S. Department of Agriculture's activities and extension service. This includes information about the extension service, policies of the USDA and extension

“The difficulty seems to be, not so much that we publish unduly … but rather that publication has been extended far beyond our present ability to make real use of the record. The summation of human experiences is being expanded at a prodigious rate, and the means we use for threading through the consequent maze to the momentarily important item is the same as was used in the days of square-rigged ships. … Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing.”

“There is no practical obstacle whatever now to the creation of an efficient index to all human knowledge, ideas and achievements, to the creation, that is, of a complete planetary memory for all mankind… The whole human memory can be, and probably in a short time will be, made accessible to every individual… This is no remote dream, no fantasy.”

# Search and Privacy

https://startpage.com/eng/top-ten-ways-startpage.html

https://startpage.com/eng/top-ten-ways-startpage.html

## 10 Ways StartPage Helps You Take Back Your Privacy

1. **StartPage doesn't store your IP address, use tracking cookies, or make a record of your searches.**
   We do not keep any information about the people who search through StartPage or what they search for. *Nothing. Nada. Zilch.*

2. **StartPage protects you from NSA surveillance and spying.**
   Your search session with StartPage is protected through powerful SSL encryption so no one - not hackers, not your ISP, not even the federal government - can eavesdrop on your searches. (Read more here)

3. **StartPage gives you 100% real Google results in complete privacy.**
   When you search with Startpage, we remove all identifying information from your query and submit it anonymously to Google for you. We get the results and return them to you in total privacy.

4. **StartPage is a Dutch company, so it is not under US jurisdiction.**
   Because our company is based in the Netherlands, US data collection programs like PRISM, the Patriot Act, FISA courts, etc. do not directly apply to us. We have never cooperated with spying programs like PRISM. (Plus we have no user data to begin with.)

5. **StartPage offers a free proxy with every search.**
   With our proxy, not only can you search privately, but you can view the pages you find through StartPage anonymously and in complete security. To learn more, please see our short overview video here.

6. **StartPage is third-party certified for privacy.**
   We not only promise our users total privacy, we back up those claims with rock-solid evidence, through stringent third-party auditing and certification. Here are details:
   **How can your privacy policies be verified? Can users trust StartPage to do what it says?**
   https://support.startpage.com/index.php?/Knowledgebase/Article/View/8/0/how-can-your-privacy-policies-be-verified-can-users-trust-startpage-to-do-what-it-says

# It Matters How It Works

1. Gather information.
2. Keep copies.
3. Build an index.
4. Understand the query.
5. Determine the relevance of each possible result to the query.
6. Determine the ranking of the relevant results.
7. Present the results.

# 1. Gather Information

- Spiders or web crawlers wander the web building indices
- estimates range from .02% to 3% of information is indexed
- How often does a page get visited?
  - some frequently (daily), others rarely (determined by the crawler to not be changing)
- How does the crawler find it's way and not go in circles?
- Login's keep bots/crawlers out.

# 2. Keep Copies

- Spider downloads the page as part of the "visit" in order to create the index.
- Search engine may "cache" the copy.
- Is this legal? What about copyright?
- But wait, browsing requires copying as well.

---

"(AFP) – Sep 15, 2011

NEW YORK — Google and publishers told a US judge Thursday they are close to settling a lawsuit over the Internet giant's controversial book-scanning project…"

# 2. Keep Copies

- Spider downloads the page as part of the "visit" in order to create the index.
- Search engine may "cache" the copy.
- Is this legal? What about copyright?
- But wait, browsing requires copying as well.

---

NYT "**By CLAIRE CAIN MILLER**

**Published: October 4, 2012**

SAN FRANCISCO — After seven years of litigation, Google and book publishers said on Thursday that they had reached a settlement to allow publishers to choose whether Google digitizes their books and journals…."

# 3. Build an Index

- list of terms and for each term a list of where it appeared
- more than just the terms
  - terms in bigger font might be more important
  - terms in the title might be more important
- must be very fast to lookup
- could be millions of entries (not just words, but names, special numbers, etc.) requiring Gigabytes of memory
- must fit in the computers memory (see next slide)

# Binary Search

- Pick a number between 1 and 1000. How many guesses will I need?
- What about that index with 25 million entries?

- The search happens in memory, but the list of URLs associated with the “term” will likely be on disk.

# 4. Understand the Query

- Steps 1-3 happen in "the background"
- Not much "understanding" in today's search engine's but that could change soon.
- Advanced search engine features help

Cardinal's beat Rangers

vs

Ranger's beat Cardinal's

What about a business called "THE"?

# 5. Determine Relevance

- "Recall" - what percentage of relevant documents are returned by the search?
- Simple relevance calculation -
  - count the number of times each seach word appears in the document, add them all up
- Long documents get higher scores.
- Uninteresting words like "the" contribute to the score.
- All word occurrences are not equal (title words should count more).

# 6. Determine Ranking

- Which of the relevant documents should be displayed first?
- Simple solution - put one with highest relevance score first.
- What if many have the same score?
- Are ones with the highest relevance score really the most important? What about the source of the document (e.g. NY Times vs some random blog post).

# Page Rank Algorithms

- The "crown jewels" of search engines lie in their page rank algorithms.
- Factors include:
  - keywords in heading or titles
  - keyword only in the body text
  - site is "trustworthy"
  - links on this page are to relevant pages
  - links to this page are relevant
  - age of the page
  - quality of the text (e.g. absence of misspellings)

# Google's PageRank Algorithm

- If lots of pages point TO this page, this must be a "more important" page
- Tweaking the page rank algorithm can make or break a small business.

# 7. Presenting Results

- Mostly just a list.
- Maybe there are better forms.
- Those sponsored links…

What must fit in computer memory in order to give quick search results?

A. The index (list) of terms (words etc.) and where on disk to find the pages that match each of those terms.
B. The list of all URLs known to the search engine and their associated search terms.
C. Neither A nor B fit entirely in the computers main memory.
D. Both A and B need to fit to get good performance.

# Who Pays for What?

- Users could pay a subscription fee (early AOL and CompuServe)
- Web sites could pay for being indexed.
- The government could pay (taxes?).
- Advertisers could pay.

# Placements, Clicks, and Auctions

- Buy higher position in the ranking - FTC said don't do it without flagging it as such.
- Banner ads displayed when search included certain terms.
  - pay for view or pay for click throughs?
- Companies bid for popular terms.
- Companies exercise editorial power (censorship?) by refusing certain ads.

# Search is Power

"For the user, search is the power to find things, and for whoever controls the search engine, search is the power to shape what you see."

Blown to Bits pg 112

# Did you mean adoption?

- Amazon search for “abortion” gave results that included “Did you mean adoption?”
- This was algorithmic, not editorial. Adoption is spelled similarly to abortion.
- They tweaked their “algorithm” to not do this when a pro-choice group complained.

# Manipulating Search Results

- White text on white background with words that will raise your rank.
- Google Bombing - “miserable failure” search in 2000 yielded white house biography of George Bush
- Companies that will help you move up in the ranking with changes to your web site.

# Search Engines Don't See Everything

- Spiders don't index the contents of many document formats found on the web.
- They don't reach into databases. (If you type something on a page then the content is probably in a database - e.g. my.ucsc.edu, ecommons.)

# Google in China

- Google agreed in 2006 to censor its Chinese search results.

# CS Concepts

- search engine
- web crawler
- page rank
- primary vs secondary memory access time
- binary search
- who pays for “free” searching

# Social Issues

- search as censor
- right to be forgotten
- can’t take it back
- free speech
- regulate search engines?

# Summary

- Search engines offer unprecedented access to information.
- Search engines place the power to shape what we see into the hands of a few companies.
- You can count on search engines to continue to evolve.
- Why not be one of the people to drive those changes?