University of Southampton

Search Engines

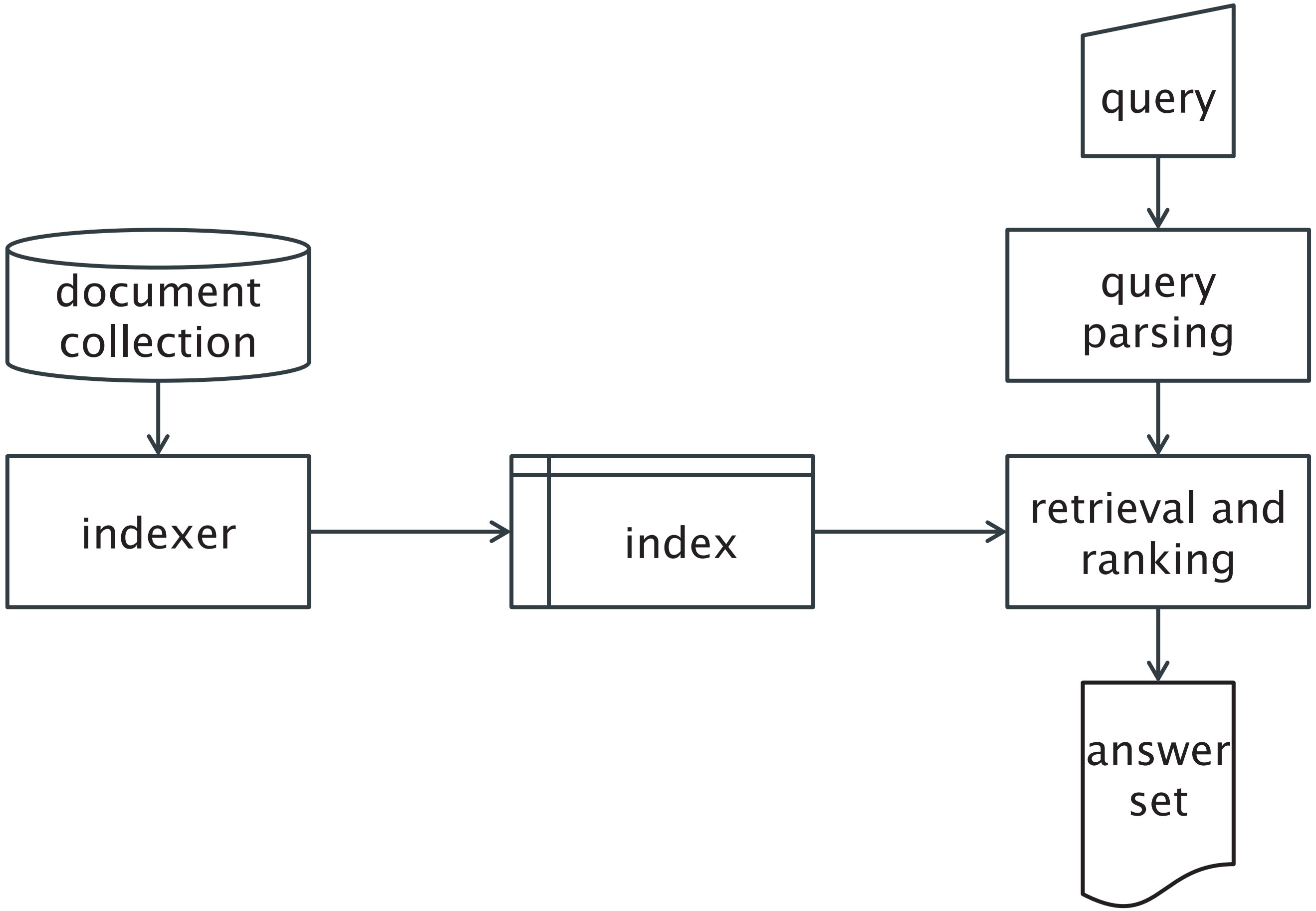
COMP3227 Web Architecture & Hypertext Technologies

Dr Heather Packer – hp3@ecs.soton.ac.uk

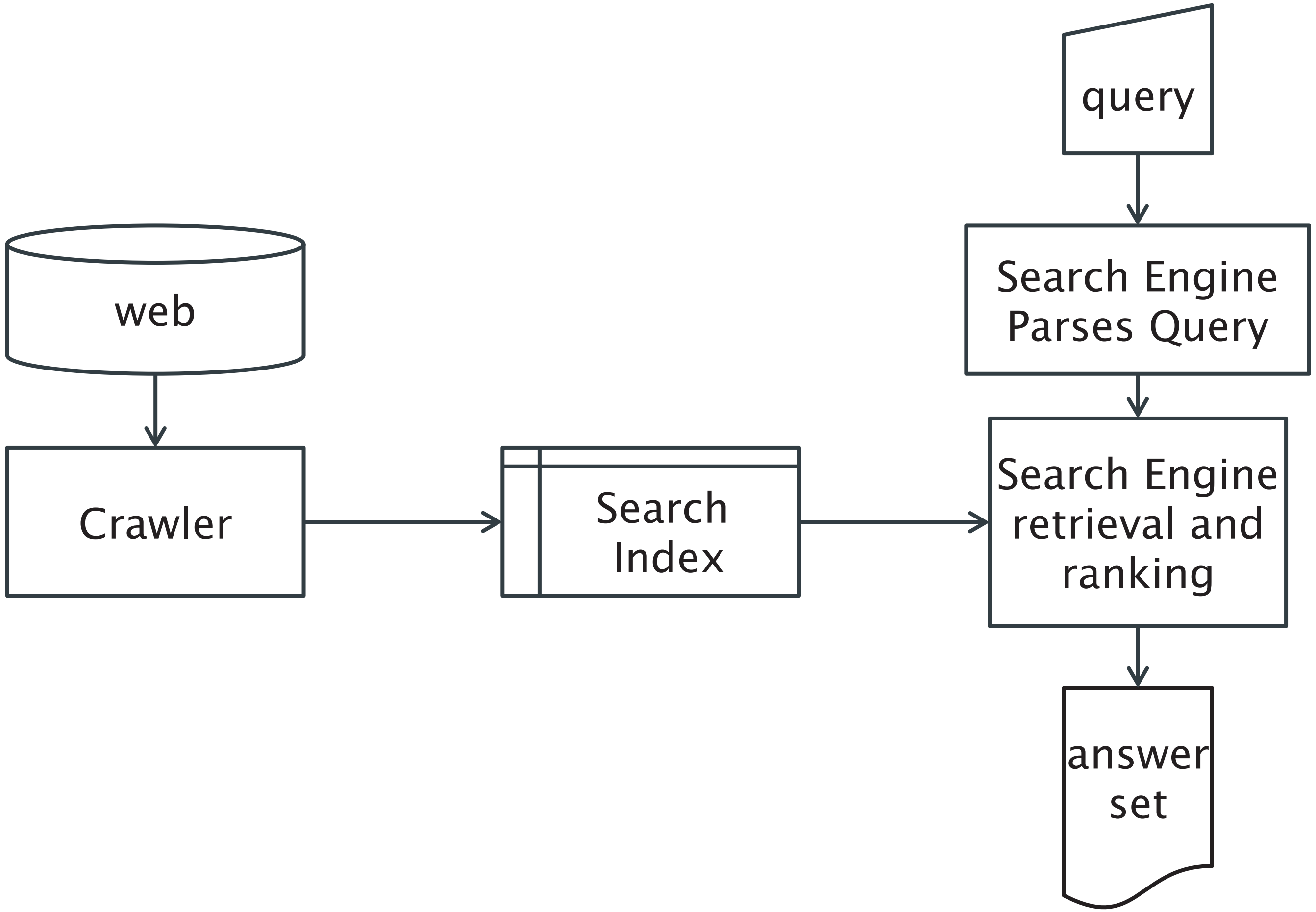
Information Retrieval

- The primary goal of an information retrieval system is to retrieve all the documents that are relevant to a user query while retrieving as few non-relevant documents as possible
  - An **information need** is a topic which a user desires to know more about
  - A **query** is what the user conveys to the computer in an attempt to communicate their information need
  - A document is **relevant** if it is one that the user perceives as containing information of value with respect to their personal information need

High-Level System Architecture



Search Engine High-Level System Architecture


query
web
Search Engine Parses Query
Crawler
Search Index
Search Engine retrieval and ranking
answer set

Specific Board and Vague Queries

	Specific	Broad	Vague
I know specifically what I'm looking for	✓	×	×
I know where to look for what I'm looking for	✓	✓	×
Example	Looking for a specific gene in the Gene Database	Looking for the manager of HR, in the company directory	Google

Search Engines

Search Engines

Ask Jeeves UK
Baidu 百度
DuckDuckGo
bing
yaCy
pipl
YAHOO!
Yandex
Google

Search Engines

- Search engines are a service
- They allow users to search for content using **keywords**
- A query returns a **ranked** set of results
- They **DO NOT** access the web directly
- They **USE** huge databases

Web Crawler, Spider or bot

- An algorithm that systematically browses the web
- A basic algorithm
  1) Start at a webpage
  2) Follow the hyperlinks that webpage points to
  3) Then follows the links those webpages point to
- Each page it visits it collects metadata about it
- Stores a file for each resource, with its metadata in a search index
- Crawlers consume resources
- Can visit sites without approval

Web Crawler - robots.txt

- Block all web crawlers

```
User-agent: *

Disallow: /
```

- Allow all web crawlers

```
User-agent: *

Disallow:
```

- Block a specific web crawler from a specific folder

```
User-agent: Googlebot

Disallow: /example-subfolder/

Disallow: /index.html
```

Web Crawler Policies

The behaviour of a web crawler is based on:

1. Selection Policy
2. Re-visit Policy
3. Politeness Policy
4. Parallelisation Policy

Web Crawler Selection Policy

- Search engines only index part of the web
- It’s important to download the most relevant pages
- A selection policy states which pages to index
- Strategies:
  - Breadth first
  - Back link count
  - PageRank
- Only request pages that have searchable content (HTML, PDF etc)

Web Crawler Re-visit Policy

- Its worth revisiting web pages because they change over time
- An ideal search engine would have the most up-to-date version of every page in its index
- Strategies
  - Re-visit all pages equally frequent
  - Prioritise pages that change often (but not too often!)
- May take page quality into account

Web Crawler Politeness Policy

- Issues of schedule and load when large collections of pages are accessed
- Strategies
  - Do not make parallel calls to the same server
  - Spread out requests
  - Abide by Crawler delay in Robots.txt

Web Crawler Parallelisation Policy

- An efficient crawler needs to access many web servers at once
- Run multiple processes in parallel
- Could find the same URL on multiple pages
- Strategies:
  - Dynamically assign pages to crawler processes
  - Static mapping e.g. based on a hash function

Web Crawler Crawlability

- Broken links
- Denied access
- Outdated URLs
- URL errors
- Blocked
- No out links
- Slow load speed
- Duplicated pages
- Flash content
- JavaScript (Googlebot executed from 2014)
- HTML frames (outdated and thus poorly indexed)

Search Index Ranking

- Query results can be ranked using many features:
  - How many times does the page contain the keywords
  - Do keywords appear in the title or URL
  - Does it contain synonyms for your keywords
  - Is it from a quality source
  - What is its importance
  - How often a page is updated
  - Freshness of information
  - Page load time

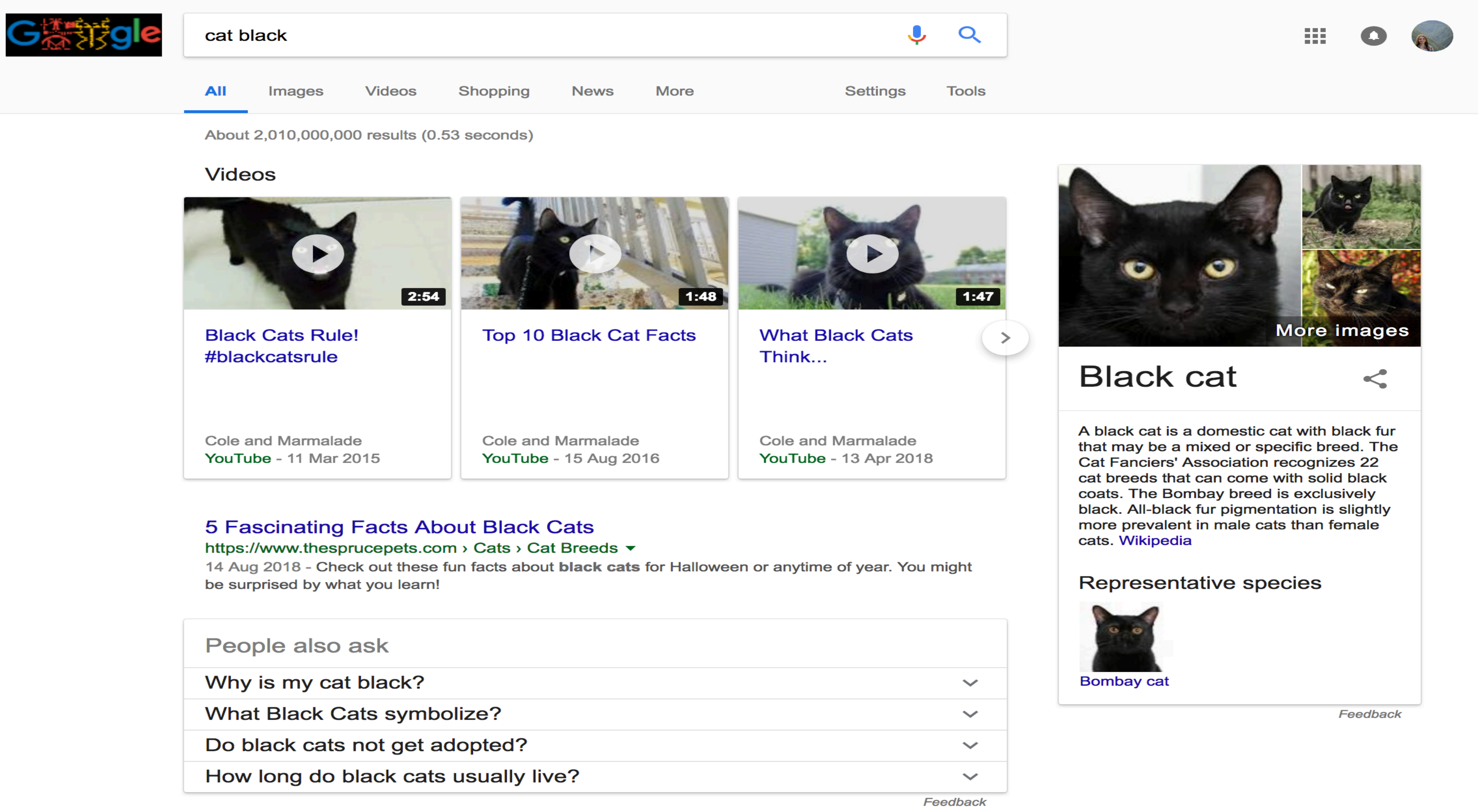
Indexing Data Issues

- Distributed data
- Changes over time
- Large volume
- Data
  - Unstructured data - gifs, pdf, etc
  - Redundant data - 30% pages are near duplicates
  - Quality of data - False, poorly written, invalid, misspelt
  - Heterogeneous data - media, formats, languages, alphabets

User Search Problems

- Users do not understand how to provide a sequence of words for searches
- Users may get unexpected answers because they are not aware of the input requirement of the search engine.
  - For example, some search engines are case sensitive.
- Users have problems understanding Boolean logic
- Around 85% of users only look at the first page of the result, so relevant answers might be skipped.

User Search Problems: Ordering of Terms



cat black

About 2,010,000,000 results (0.53 seconds)

Videos

Black Cats Rule! #blackcatsrule
Cole and Marmalade
YouTube - 11 Mar 2015

Top 10 Black Cat Facts
Cole and Marmalade
YouTube - 15 Aug 2016

What Black Cats Think...
Cole and Marmalade
YouTube - 13 Apr 2018

5 Fascinating Facts About Black Cats
https://www.thesprucepets.com › Cats › Cat Breeds
14 Aug 2018 - Check out these fun facts about **black cats** for Halloween or anytime of year. You might be surprised by what you learn!

People also ask

Why is my cat black?

What Black Cats symbolize?

Do black cats not get adopted?

How long do black cats usually live?

Feedback

Black cat

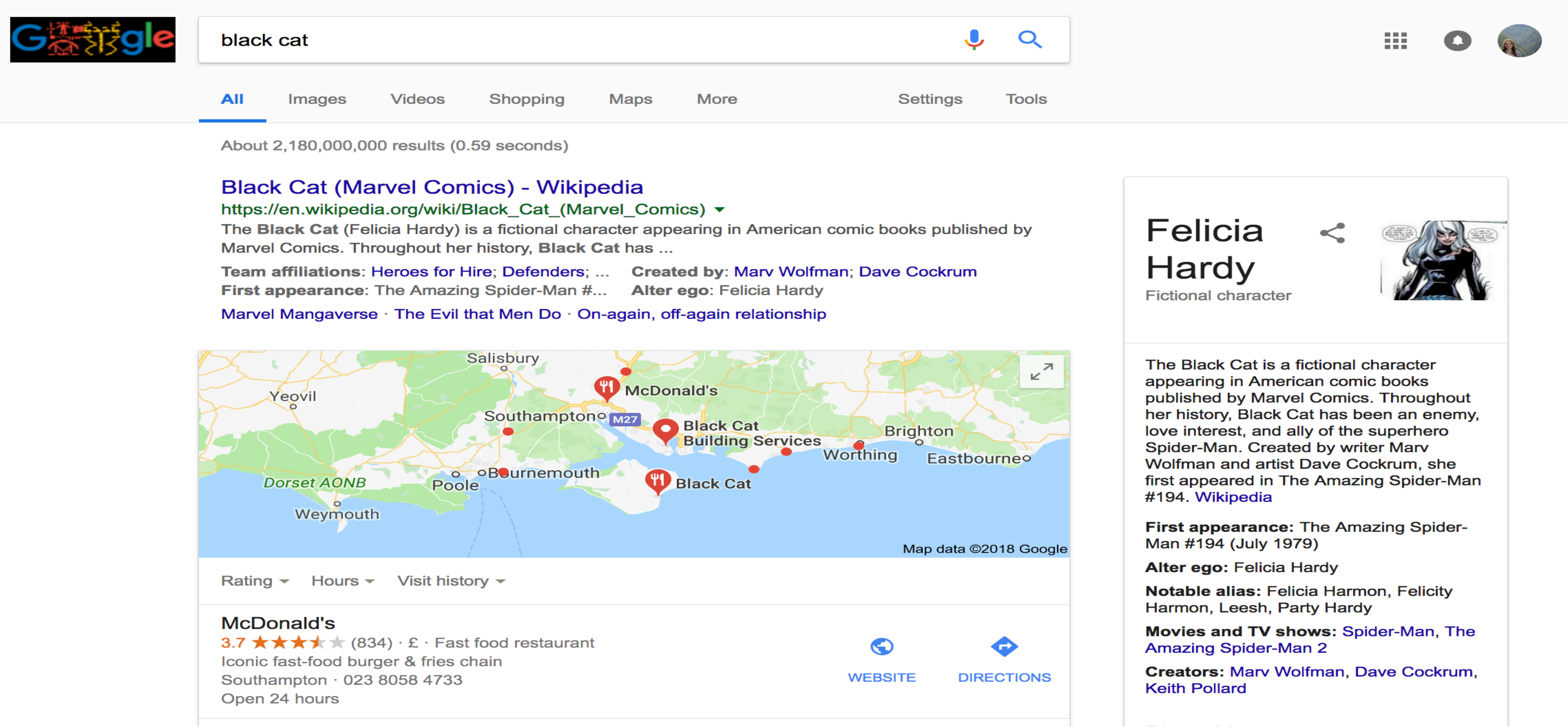
A black cat is a domestic cat with black fur that may be a mixed or specific breed. The Cat Fanciers' Association recognizes 22 cat breeds that can come with solid black coats. The Bombay breed is exclusively black. All-black fur pigmentation is slightly more prevalent in male cats than female cats. Wikipedia

Representative species

Bombay cat

Feedback

User Search Problems: Ordering of Terms



About 2,180,000,000 results (0.59 seconds)

Black Cat (Marvel Comics) - Wikipedia
https://en.wikipedia.org/wiki/Black_Cat_(Marvel_Comics)
The **Black Cat** (Felicia Hardy) is a fictional character appearing in American comic books published by Marvel Comics. Throughout her history, **Black Cat** has ...

Team affiliations: Heroes for Hire; Defenders; ...
First appearance: The Amazing Spider-Man #...
Created by: Marv Wolfman; Dave Cockrum
Alter ego: Felicia Hardy

Marvel Mangaverse · The Evil that Men Do · On-again, off-again relationship

McDonald's
3.7 ★★★★☆ (834) · £ · Fast food restaurant
Iconic fast-food burger & fries chain
Southampton · 023 8058 4733
Open 24 hours

Felicia Hardy
Fictional character

The Black Cat is a fictional character appearing in American comic books published by Marvel Comics. Throughout her history, Black Cat has been an enemy, love interest, and ally of the superhero Spider-Man. Created by writer Marv Wolfman and artist Dave Cockrum, she first appeared in The Amazing Spider-Man #194. Wikipedia

First appearance: The Amazing Spider-Man #194 (July 1979)

Alter ego: Felicia Hardy

Notable alias: Felicia Harmon, Felicity Harmon, Leesh, Party Hardy

Movies and TV shows: Spider-Man, The Amazing Spider-Man 2

Creators: Marv Wolfman, Dave Cockrum, Keith Pollard

Boosting web traffic

- Search engines direct traffic to websites
- 85% of people don’t look at the second page
- People try to optimise their site so that it ranks highly on Search Engines
- Could be fundamental to a website’s business model

Search Engine Optimisation

- Whole industry exists trying to boost search ranking to ensure pages are indexed by search engines

- Legitimate SEO (White Hat)
  - Good Design
  - Valid metadata, alt tags on images

- Illegitimate SEO (Black Hat)
  - Often gaming search ranking algorithms
  - Deception
  - Leads to arms race between SEO and search engines

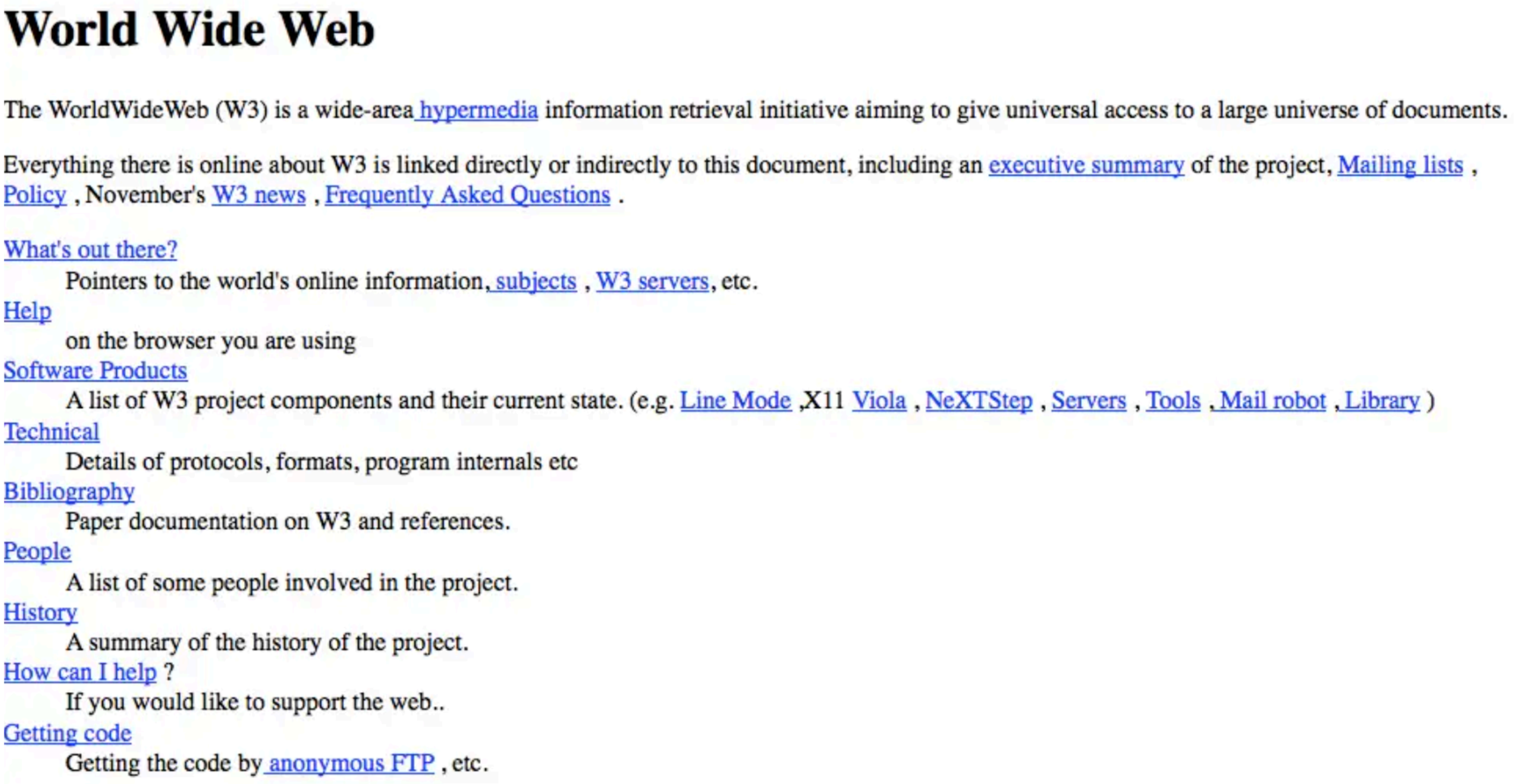
Combatting SEO

- Most search engines have rules against:
  - Invisible text
  - Meta tag abuse
  - Heavy repetition
  - “domain spam”
    - Overtly submission of “mirror” sites in an attempt to dominate the listings for particular terms

Google and other SEs are a Business

- Search Engines record tracking information
  - Google saves every voice search
  - IP addresses
  - Location
  - Saves your searches
- Google's revenue is from adverts
  - Improve their revenue with targeted advertising
- Google has a large research department
  - Improve their technology

History of the Web Browser



World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

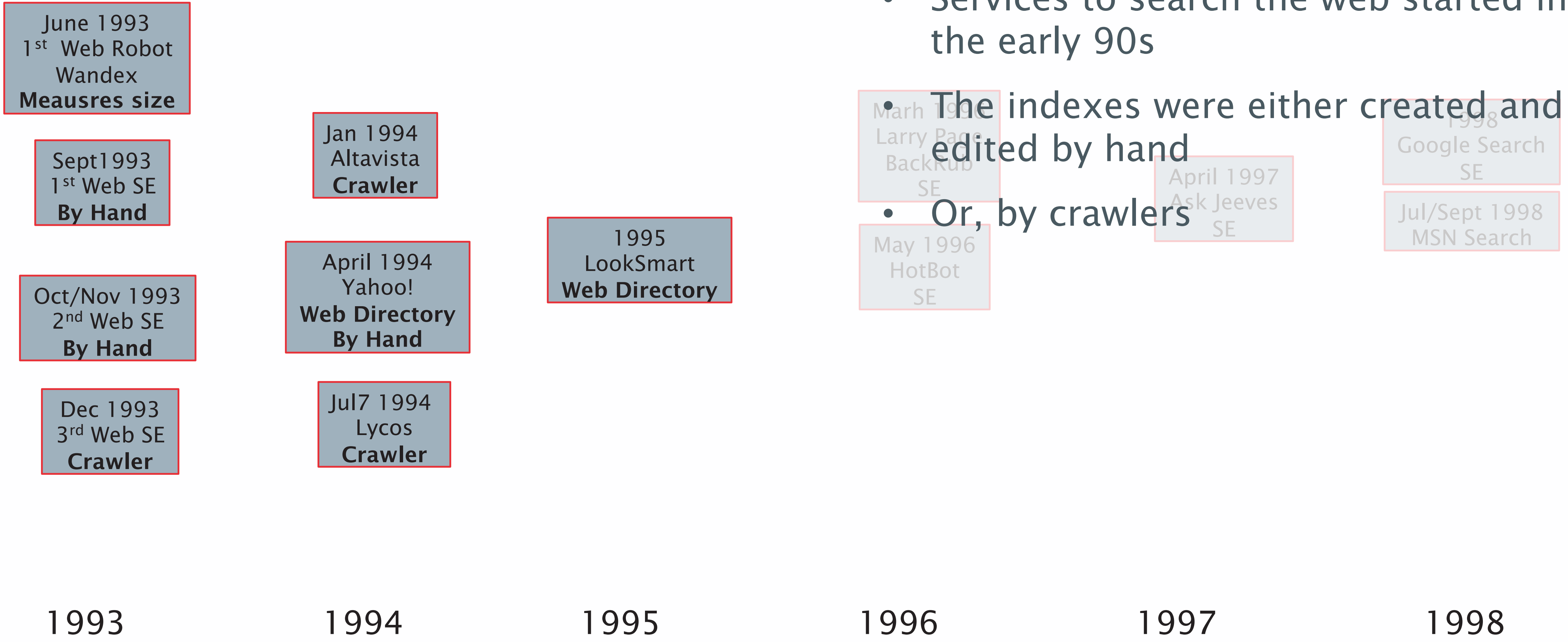
What's out there?
    Pointers to the world's online information, subjects , W3 servers, etc.
Help
    on the browser you are using
Software Products
    A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)
Technical
    Details of protocols, formats, program internals etc
Bibliography
    Paper documentation on W3 and references.
People
    A list of some people involved in the project.
History
    A summary of the history of the project.
How can I help ?
    If you would like to support the web..
Getting code
    Getting the code by anonymous FTP , etc.

Before Search Engines

- Exploit hyperlink structure
  - Personal homepages with links
  - Directories
- Word of mouth
  - Email, forums, Usenet

Timeline 90’s – Hand vs Crawler

- Services to search the web started in the early 90s
- The indexes were either created and edited by hand
- Or, by crawlers



June 1993
1st Web Robot
Wandex
Meausres size

Sept1993
1st Web SE
By Hand

Oct/Nov 1993
2nd Web SE
By Hand

Dec 1993
3rd Web SE
Crawler

Jan 1994
Altavista
Crawler

April 1994
Yahoo!
Web Directory
By Hand

Jul7 1994
Lycos
Crawler

1995
LookSmart
Web Directory

Marh 1996
Larry Page
BackRub
SE

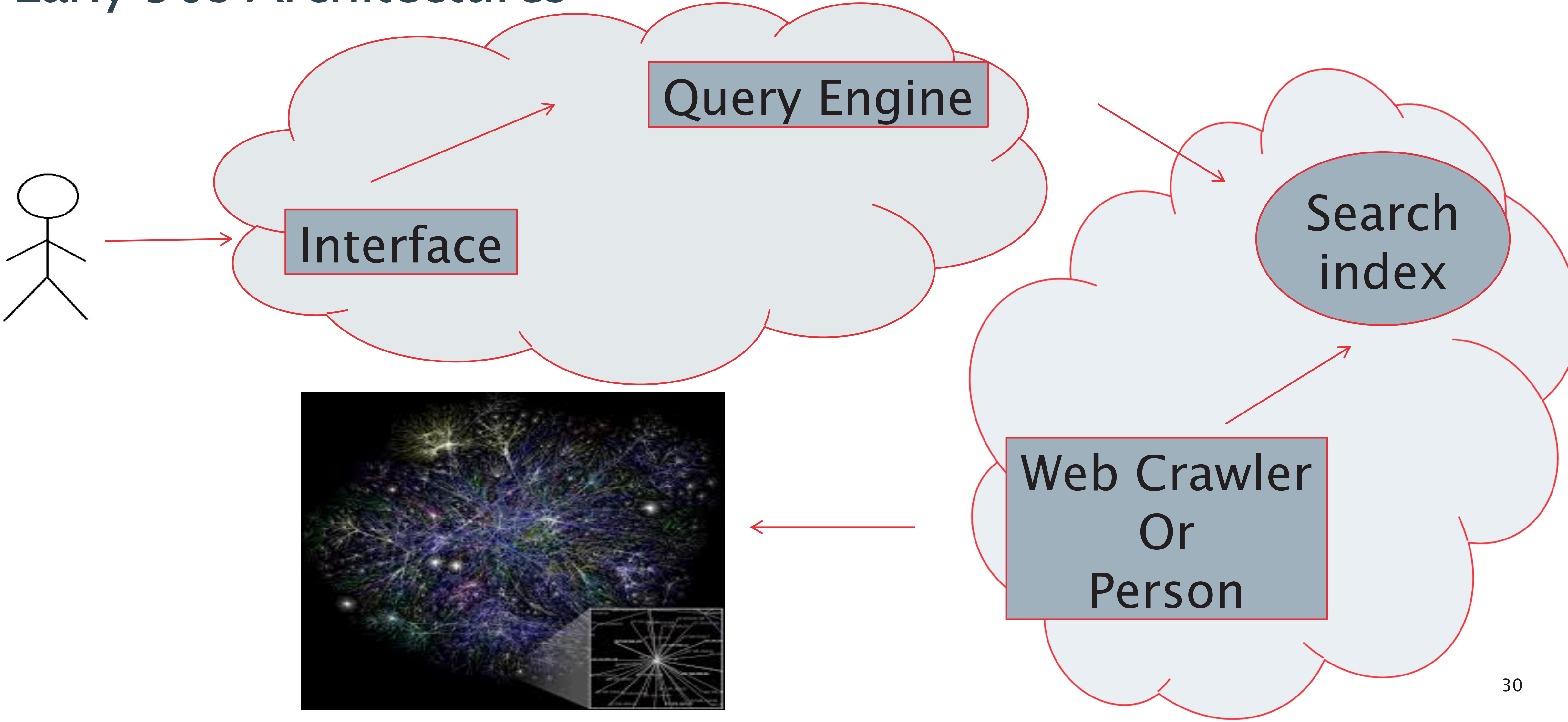
May 1996
HotBot
SE

April 1997
Ask Jeeves
SE

1998
Google Search
SE

Jul/Sept 1998
MSN Search

Early 90s Architectures



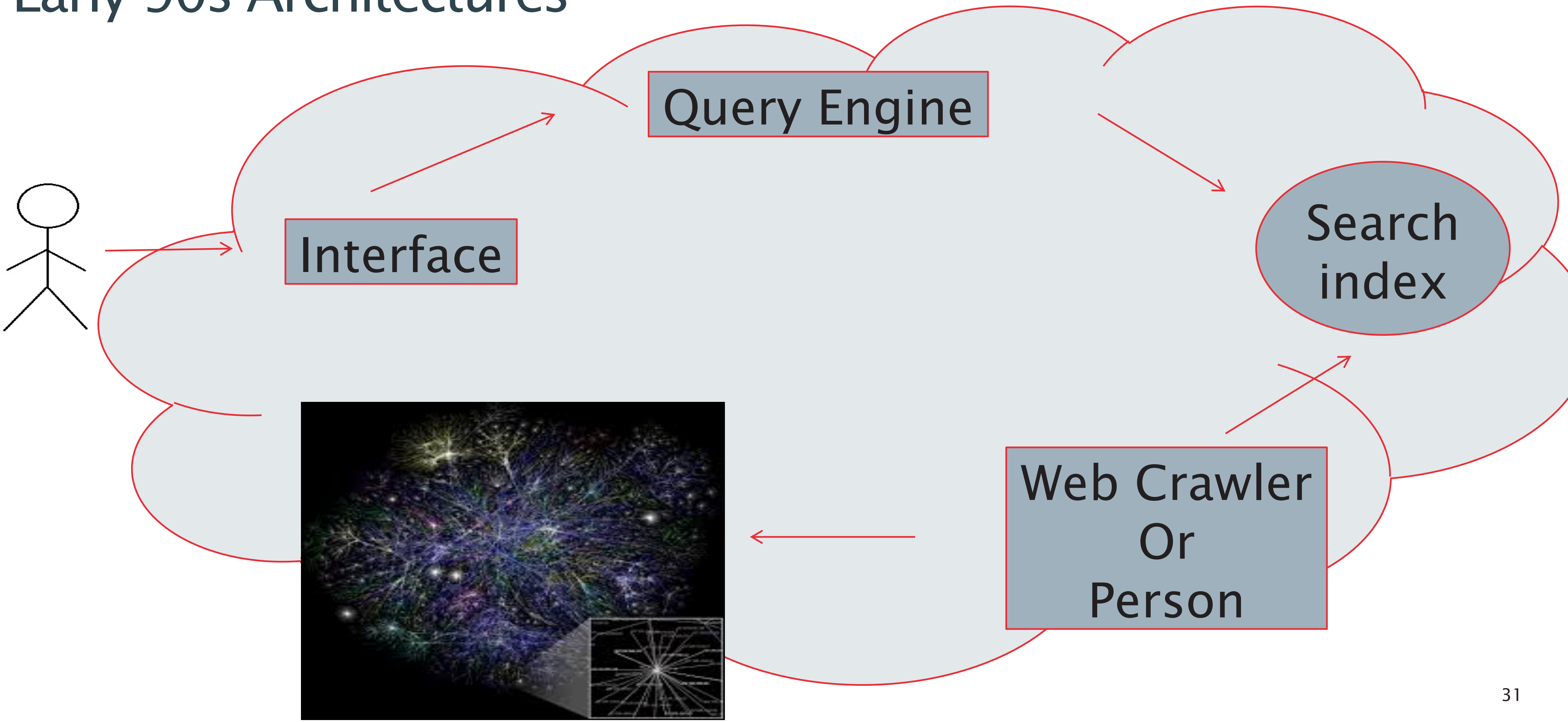
Interface

Query Engine

Search index

Web Crawler Or Person

Early 90s Architectures



Query Engine

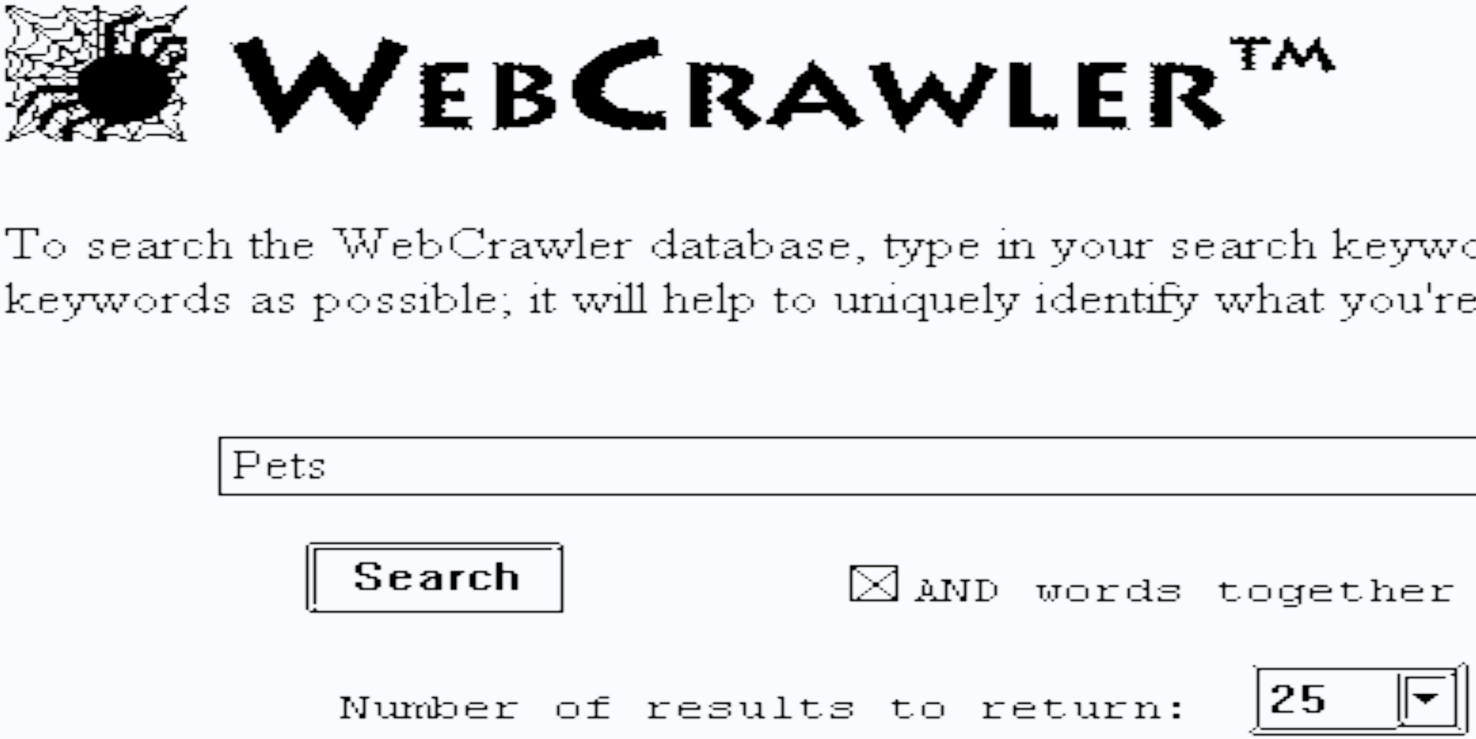
Interface

Search index

Web Crawler
Or
Person

Web Crawlers

- **Wandex** 1993 – size of the web
- WebCrawler Dec 1993 - indexer
- ***WebCrawler*** 1994 - indexed entire web page



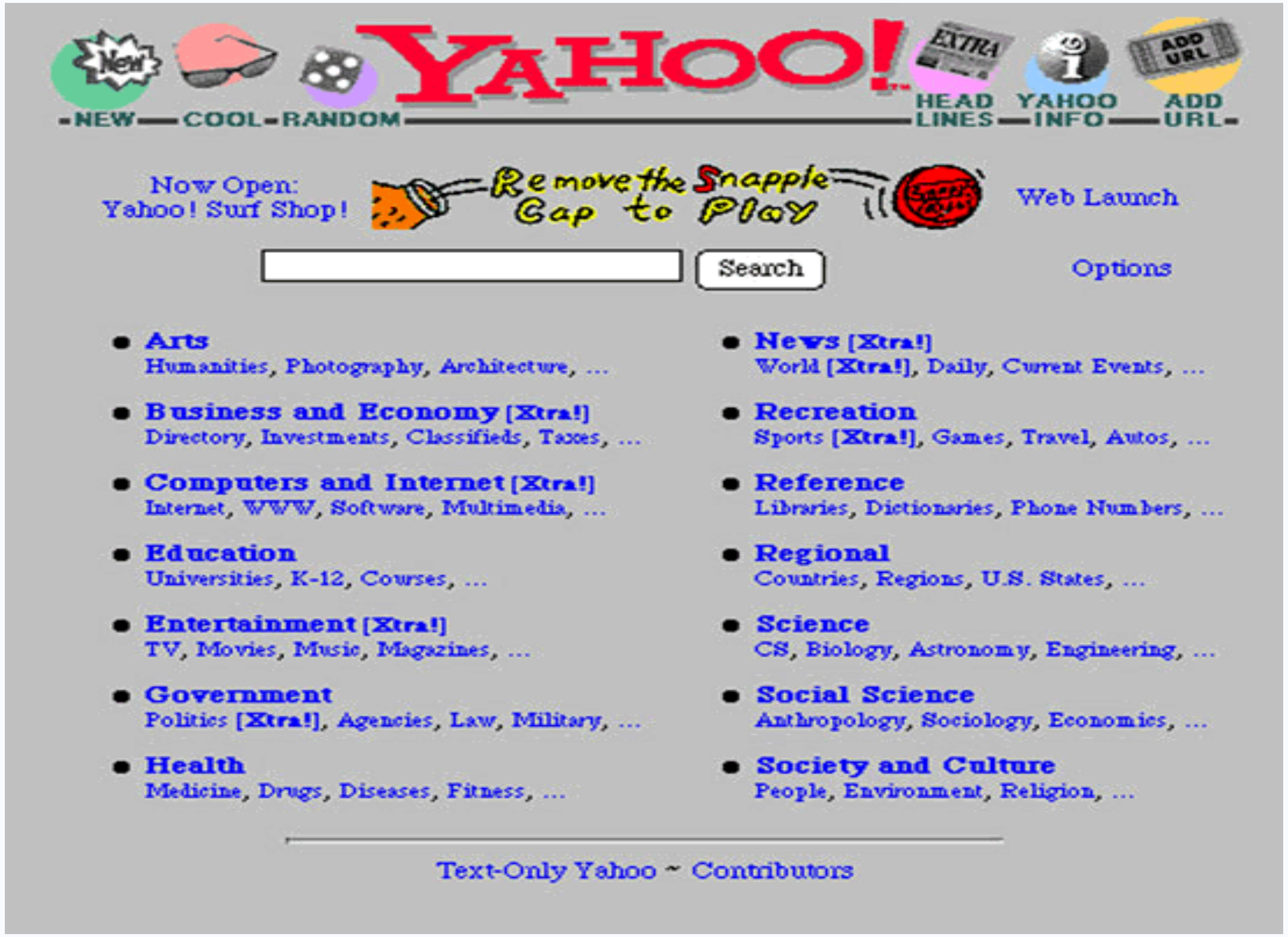
To search the WebCrawler database, type in your search keywords here. Type as many relevant keywords as possible; it will help to uniquely identify what you're looking for.

Pets

☒ AND words together

Number of results to return: 25

Yahoo! 1994



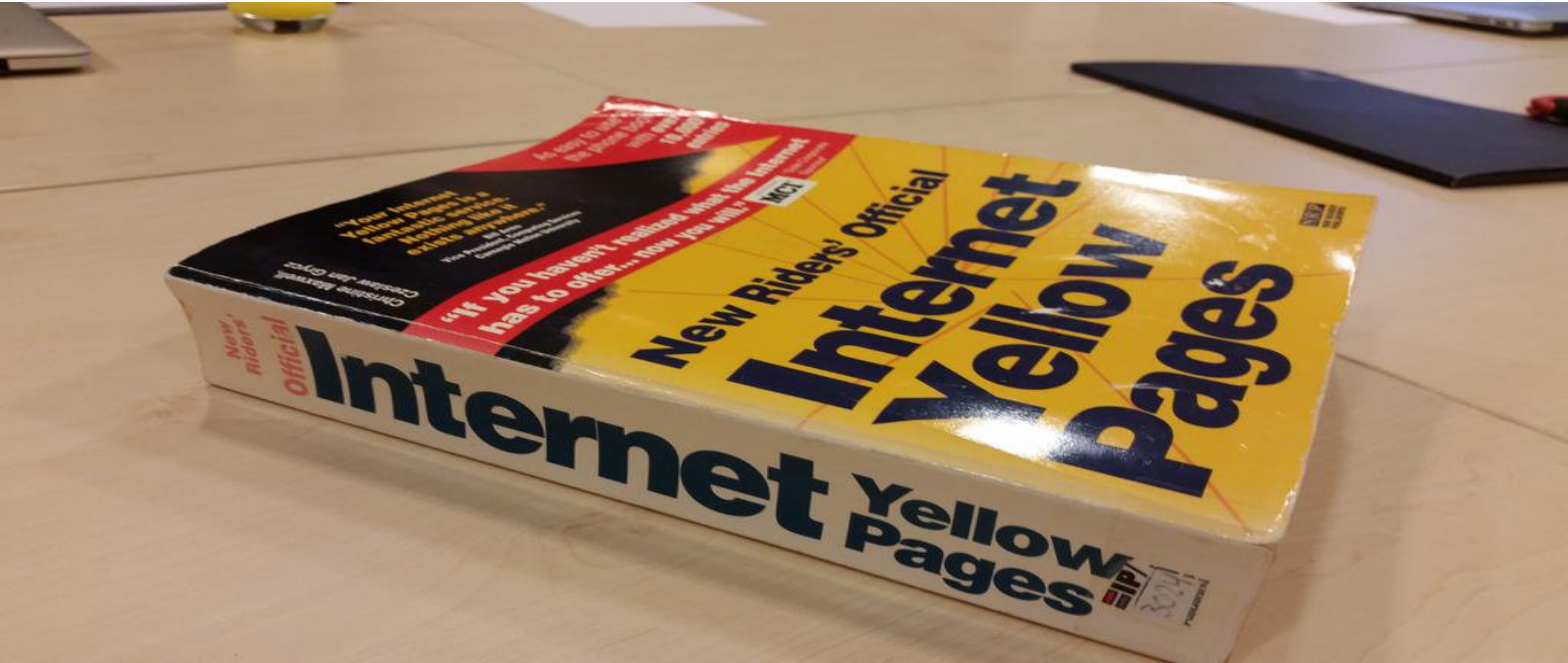
Now Open:
Yahoo! Surf Shop!

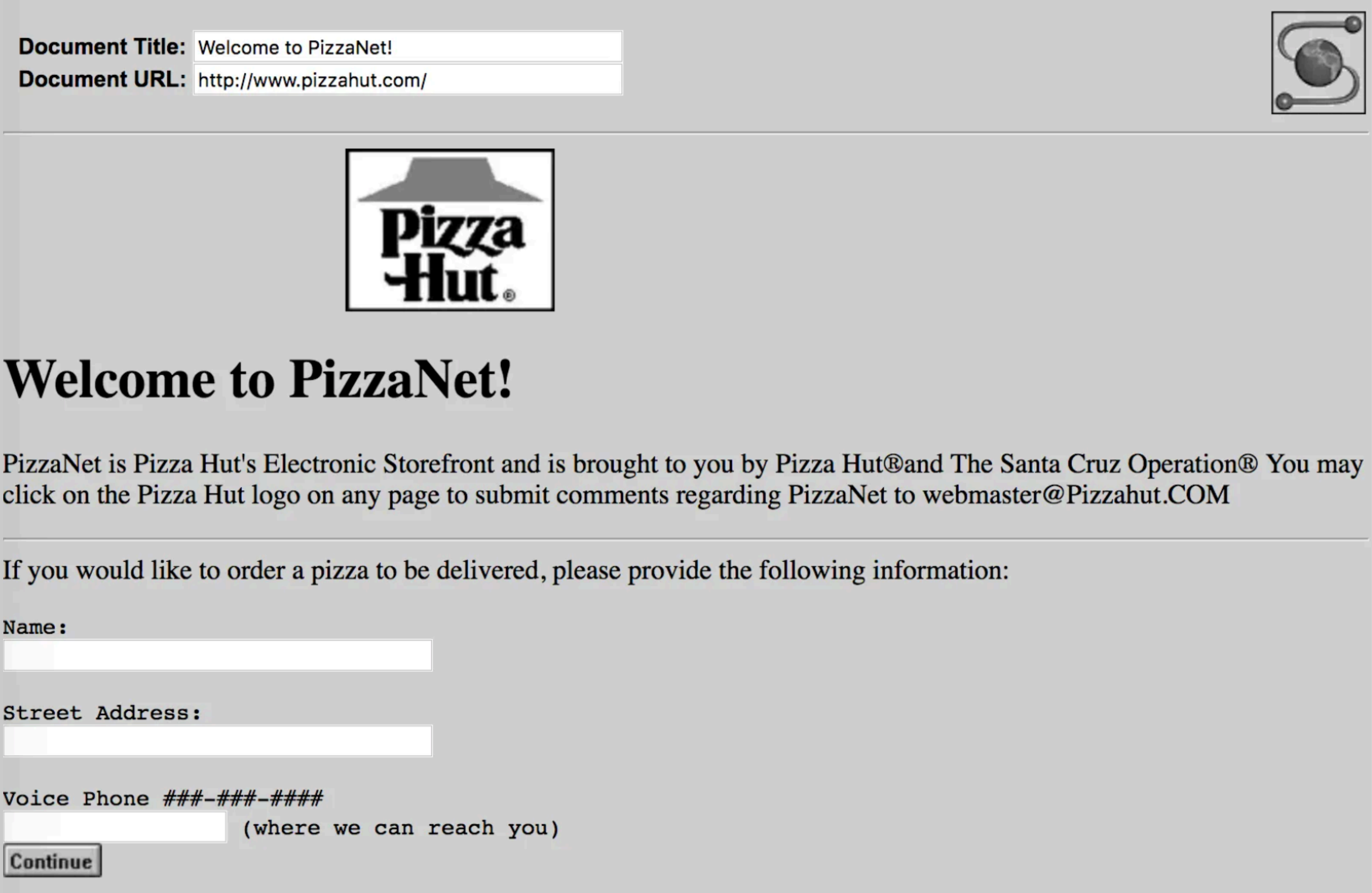
Web Launch

- **Arts**
  Humanities, Photography, Architecture, ...
- **Business and Economy** [Xtra!]
  Directory, Investments, Classifieds, Taxes, ...
- **Computers and Internet** [Xtra!]
  Internet, WWW, Software, Multimedia, ...
- **Education**
  Universities, K-12, Courses, ...
- **Entertainment** [Xtra!]
  TV, Movies, Music, Magazines, ...
- **Government**
  Politics [Xtra!], Agencies, Law, Military, ...
- **Health**
  Medicine, Drugs, Diseases, Fitness, ...
- **News** [Xtra!]
  World [Xtra!], Daily, Current Events, ...
- **Recreation**
  Sports [Xtra!], Games, Travel, Autos, ...
- **Reference**
  Libraries, Dictionaries, Phone Numbers, ...
- **Regional**
  Countries, Regions, U.S. States, ...
- **Science**
  CS, Biology, Astronomy, Engineering, ...
- **Social Science**
  Anthropology, Sociology, Economics, ...
- **Society and Culture**
  People, Environment, Religion, ...

Text-Only Yahoo ~ Contributors

Internet Yellow Pages 1994


As easy to use as the phone book with over 10,000 entries
"Your Internet Yellow Pages is a fantastic service. Nothing like it exists anywhere."
Bill Arms
Vice President, Computing Services
Carnegie Mellon University
Christine Maxwell
Czeslaw Jan Grycz
"If you haven't realized what the Internet has to offer... now you will."
MCI
New Riders' Official
Internet
Yellow
Pages
New Riders' Official
Internet Yellow Pages



Document Title: Welcome to PizzaNet!

Document URL: http://www.pizzahut.com/

Welcome to PizzaNet!

PizzaNet is Pizza Hut's Electronic Storefront and is brought to you by Pizza Hut®and The Santa Cruz Operation® You may click on the Pizza Hut logo on any page to submit comments regarding PizzaNet to webmaster@Pizzahut.COM

If you would like to order a pizza to be delivered, please provide the following information:

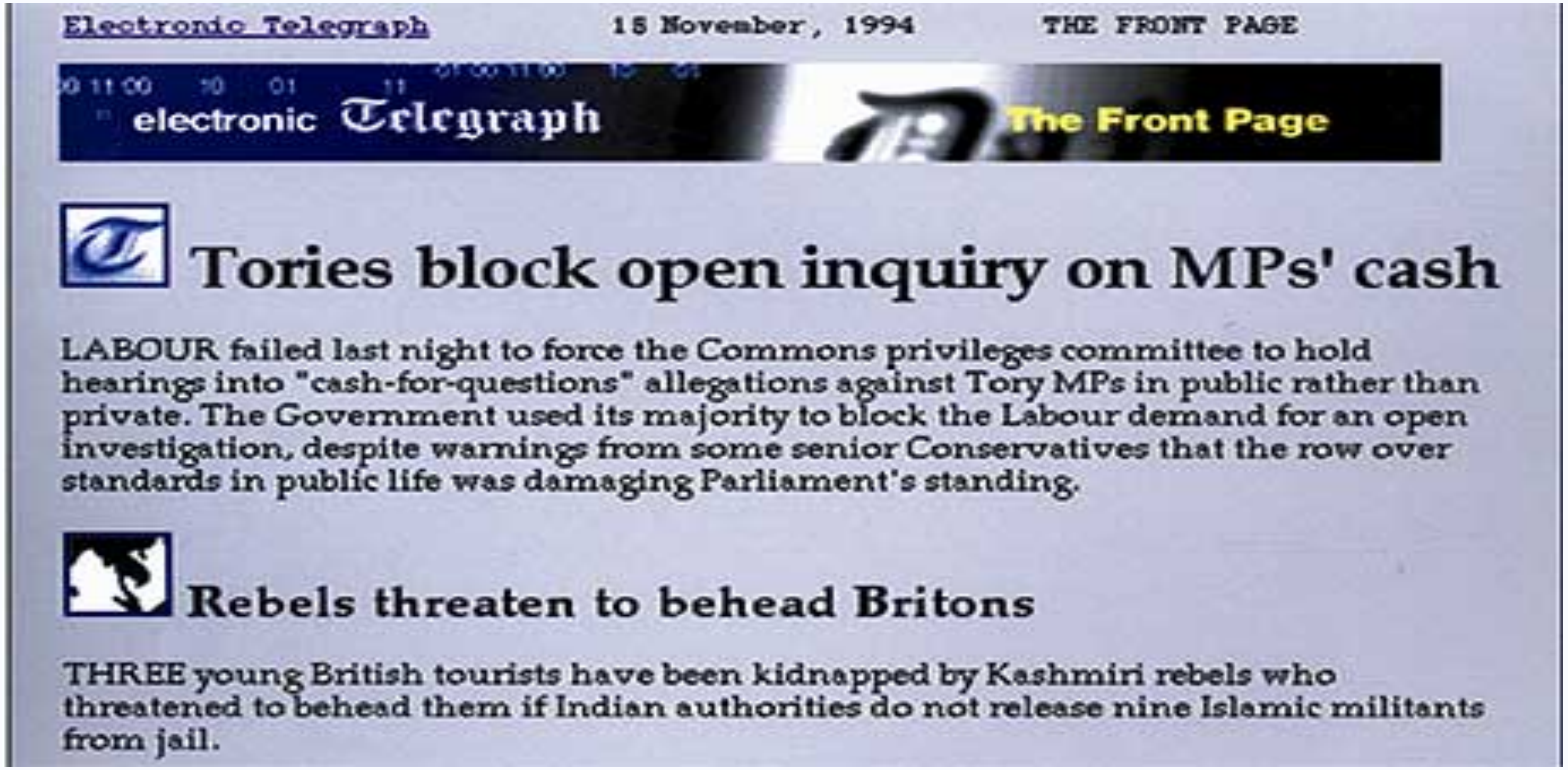
Name:

Street Address:

Voice Phone ###-###-####

(where we can reach you)

Website - Telegraph 1994



Electronic Telegraph 15 November, 1994 THE FRONT PAGE

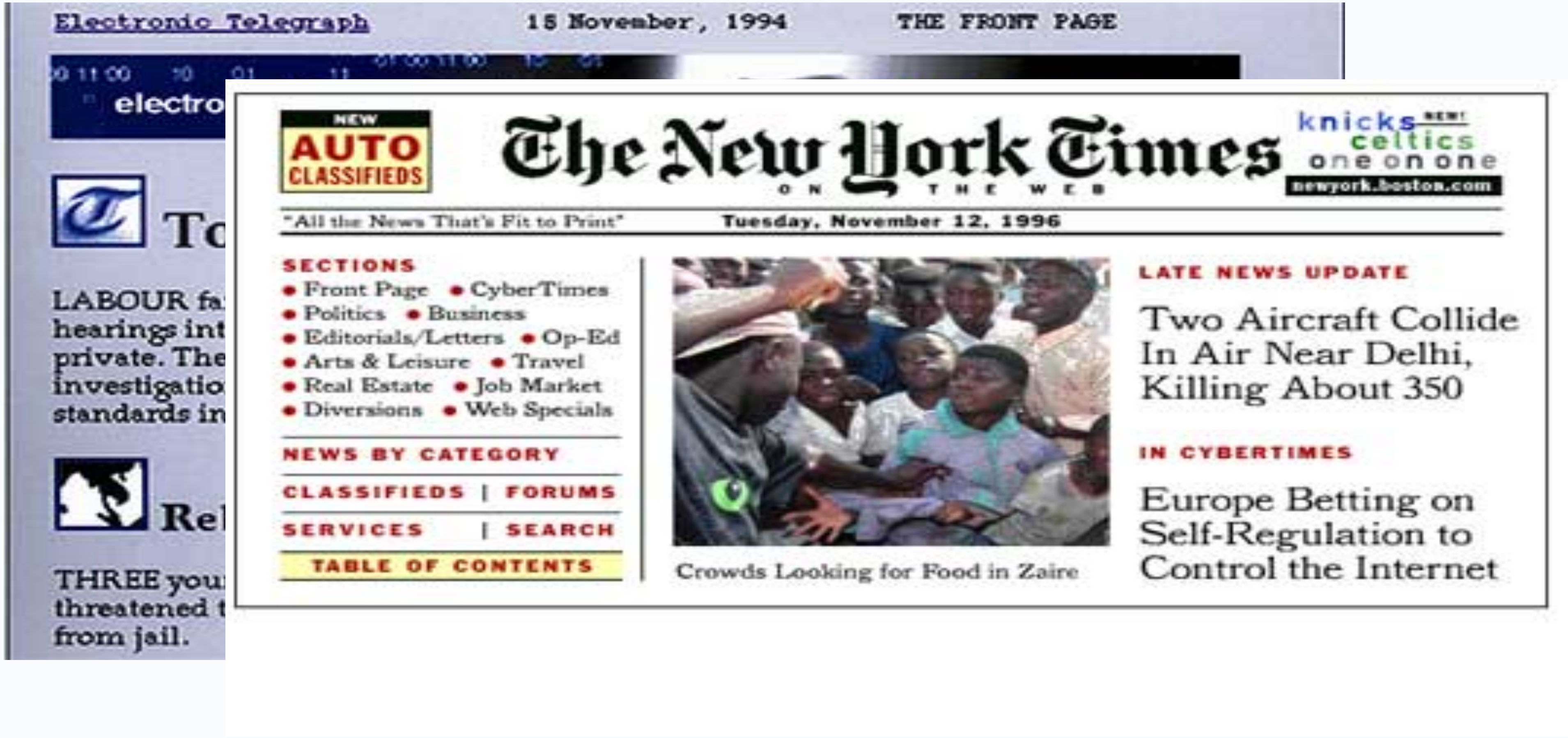
Tories block open inquiry on MPs' cash

LABOUR failed last night to force the Commons privileges committee to hold hearings into "cash-for-questions" allegations against Tory MPs in public rather than private. The Government used its majority to block the Labour demand for an open investigation, despite warnings from some senior Conservatives that the row over standards in public life was damaging Parliament's standing.

Rebels threaten to behead Britons

THREE young British tourists have been kidnapped by Kashmiri rebels who threatened to behead them if Indian authorities do not release nine Islamic militants from jail.

Website – New York Times 1995



Electronic Telegraph 15 November, 1994 THE FRONT PAGE

electro

To

LABOUR fa
hearings int
private. The
investigatio
standards in

Rel

THREE you
threatened t
from jail.

NEW AUTO CLASSIFIEDS

The New York Times ON THE WEB

knicks NEW! celtics one on one newyork.boston.com

"All the News That's Fit to Print" Tuesday, November 12, 1996

SECTIONS

- Front Page • CyberTimes
- Politics • Business
- Editorials/Letters • Op-Ed
- Arts & Leisure • Travel
- Real Estate • Job Market
- Diversions • Web Specials

NEWS BY CATEGORY

CLASSIFIEDS | FORUMS

SERVICES | SEARCH

TABLE OF CONTENTS

Crowds Looking for Food in Zaire

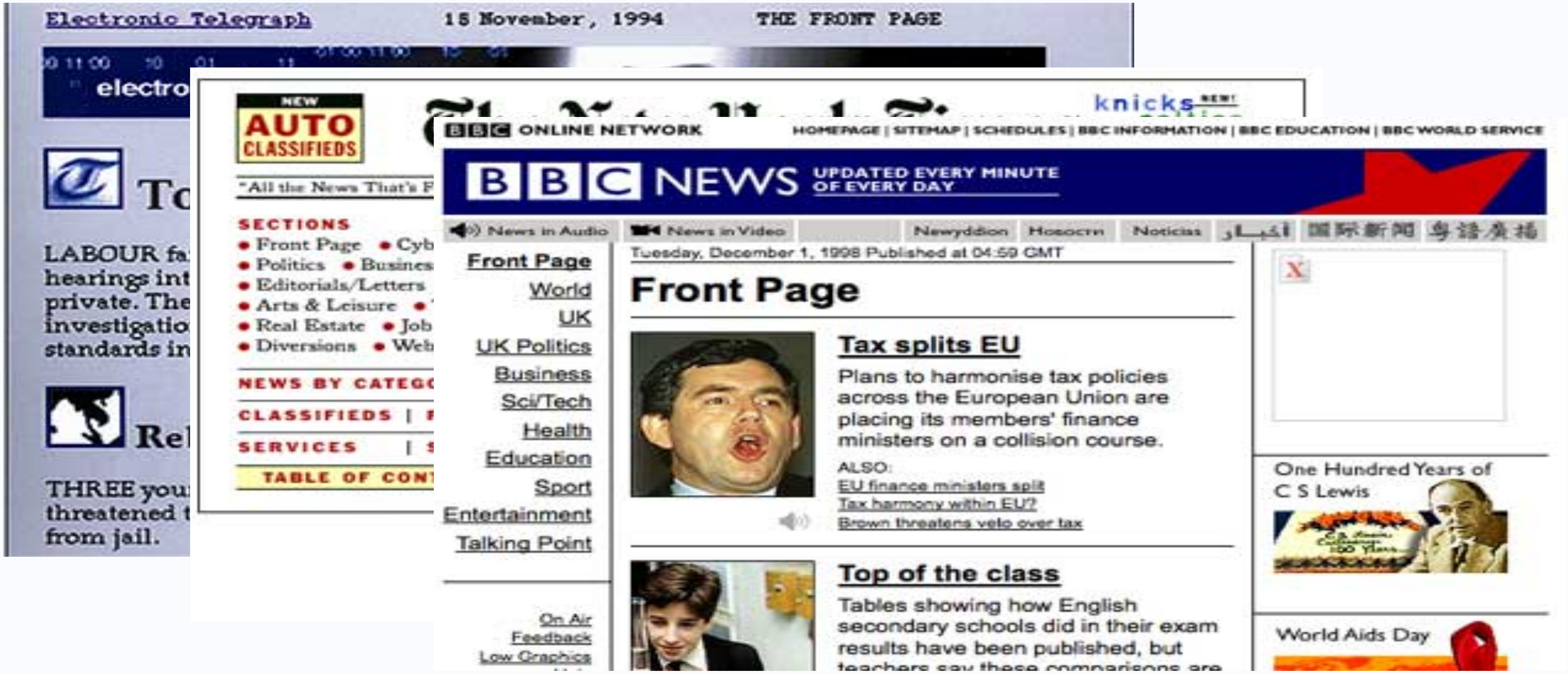
LATE NEWS UPDATE

Two Aircraft Collide In Air Near Delhi, Killing About 350

IN CYBERTIMES

Europe Betting on Self-Regulation to Control the Internet

Website – BBC News 1997



Electronic Telegraph 15 November, 1994 THE FRONT PAGE

electro

To

LABOUR fa hearings int private. The investigatio standards in

Rel

THREE you threatened t from jail.

NEW AUTO CLASSIFIEDS

"All the News That's F

SECTIONS
- Front Page • Cyb
- Politics • Busines
- Editorials/Letters
- Arts & Leisure •
- Real Estate • Job
- Diversions • Web

NEWS BY CATEG

CLASSIFIEDS |

SERVICES |

TABLE OF CONT

knicks NEW!

BBC ONLINE NETWORK HOMEPAGE | SITEMAP | SCHEDULES | BBC INFORMATION | BBC EDUCATION | BBC WORLD SERVICE

BBC NEWS UPDATED EVERY MINUTE OF EVERY DAY

News in Audio | News in Video | Newyddion | Новости | Noticias | أخبار | 国际新闻 | 粵語廣播

Front Page
World
UK
UK Politics
Business
Sci/Tech
Health
Education
Sport
Entertainment
Talking Point

On Air
Feedback
Low Graphics

Tuesday, December 1, 1998 Published at 04:59 GMT

Front Page

Tax splits EU

Plans to harmonise tax policies across the European Union are placing its members' finance ministers on a collision course.

ALSO:
EU finance ministers split
Tax harmony within EU?
Brown threatens veto over tax

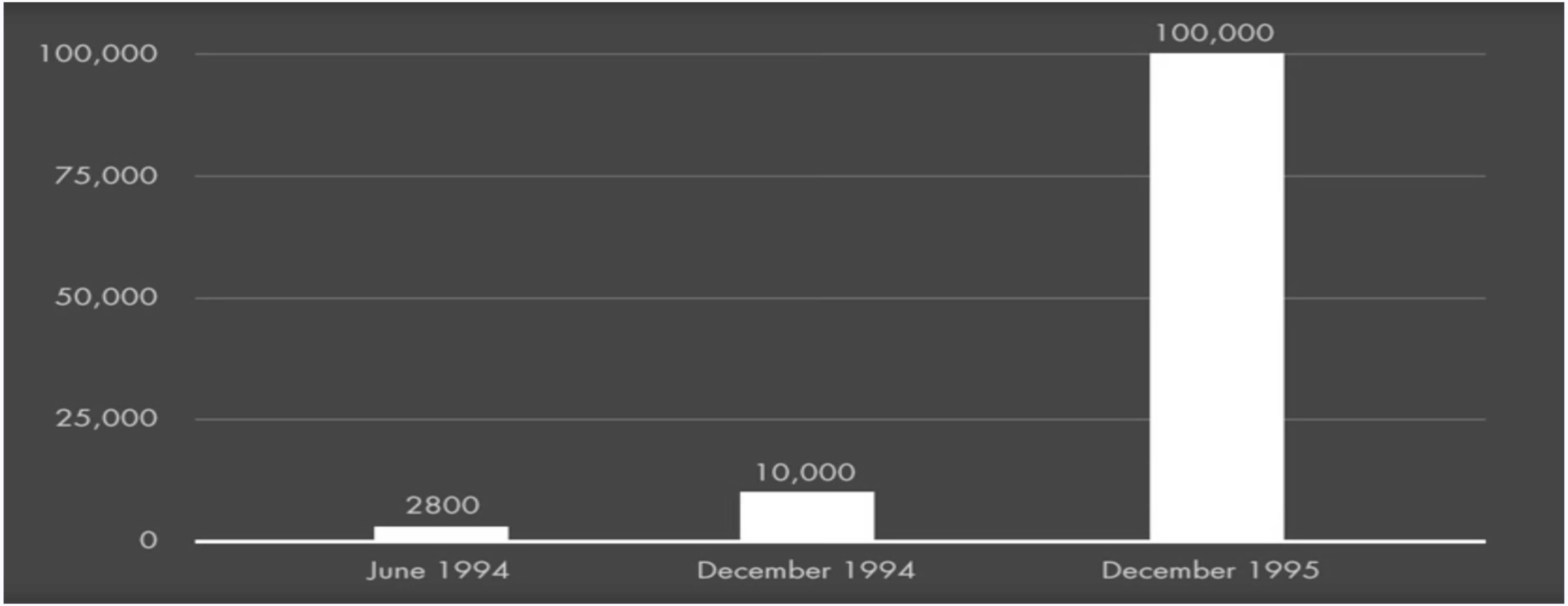
Top of the class

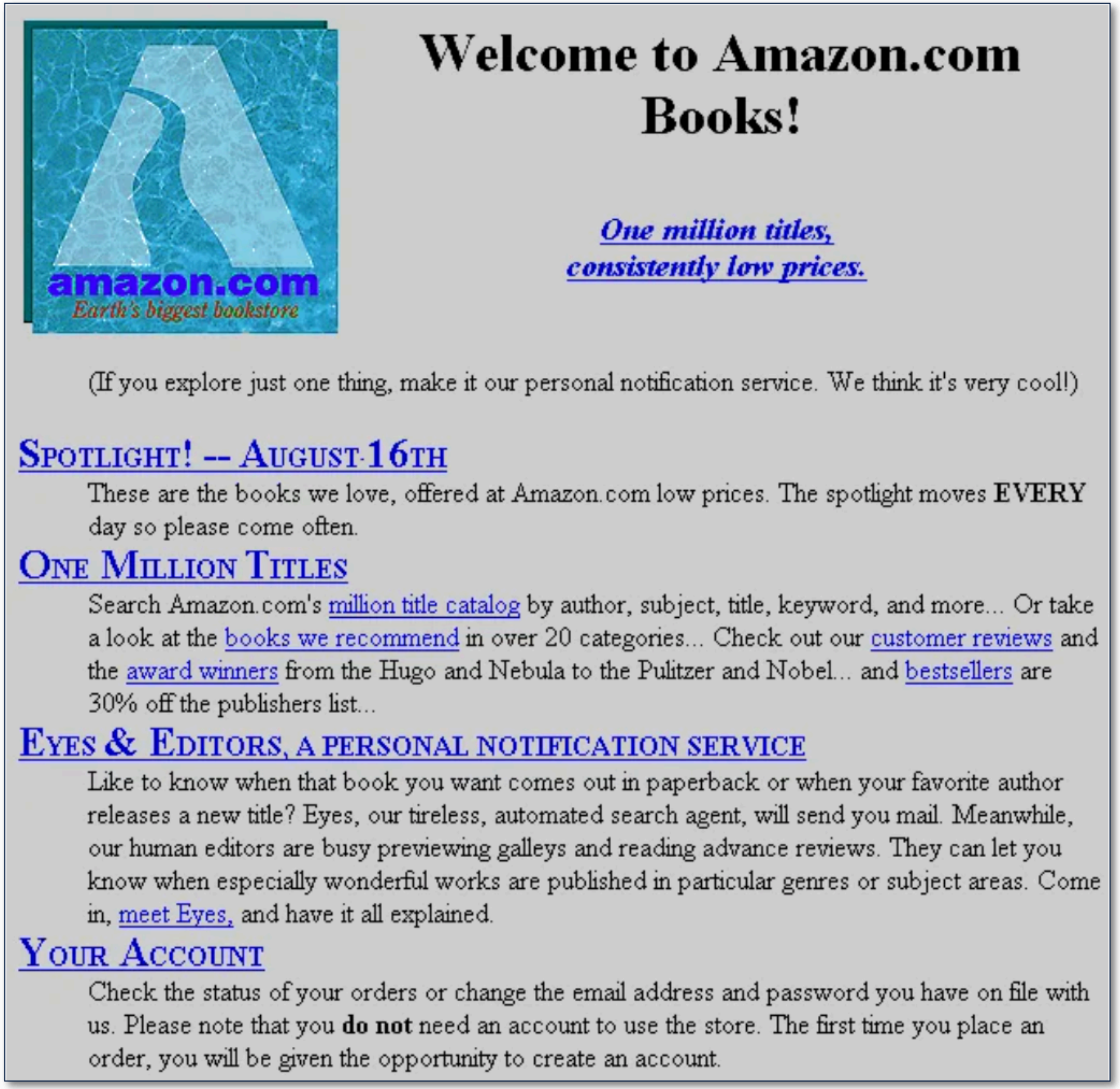
Tables showing how English secondary schools did in their exam results have been published, but teachers say these comparisons are

One Hundred Years of C S Lewis

World Aids Day

Number of Websites June 1994 - Dec 1995





Welcome to Amazon.com Books!

One million titles, consistently low prices.

(If you explore just one thing, make it our personal notification service. We think it's very cool!)

Spotlight! -- August 16th

These are the books we love, offered at Amazon.com low prices. The spotlight moves **EVERY** day so please come often.

One Million Titles

Search Amazon.com's million title catalog by author, subject, title, keyword, and more... Or take a look at the books we recommend in over 20 categories... Check out our customer reviews and the award winners from the Hugo and Nebula to the Pulitzer and Nobel... and bestsellers are 30% off the publishers list...

Eyes & Editors, a personal notification service

Like to know when that book you want comes out in paperback or when your favorite author releases a new title? Eyes, our tireless, automated search agent, will send you mail. Meanwhile, our human editors are busy previewing galleys and reading advance reviews. They can let you know when especially wonderful works are published in particular genres or subject areas. Come in, meet Eyes, and have it all explained.

Your Account

Check the status of your orders or change the email address and password you have on file with us. Please note that you **do not** need an account to use the store. The first time you place an order, you will be given the opportunity to create an account.

The original Amazon website (August 1995)

Source: Restored by Taran Van Hemert

Search Engines Issues Mid ‘90s

- Scale
  - How to rank pages to give the best results
- Spamming
  - Web searches favoured web pages with high keyword density
- Keywords
  - Spelling mistakes

Timeline 90’s

June 1993
1st Web Robot
Wandex
Meausres size

Sept1993
1st Web SE
List

Oct/Nov 1993
2nd Web SE
List

Dec 1993
3rd Web SE
Crawler
Indexer
Seraching

Jan 1994
Altavista

April 1994
Yahoo!
Web Directory

Jul7 1994
Lycos
SE

1995
LookSmart
Web Directory

March 1996
Larry Page
BackRub
SE

May 1996
HotBot
SE

April 1997
Ask Jeeves
SE

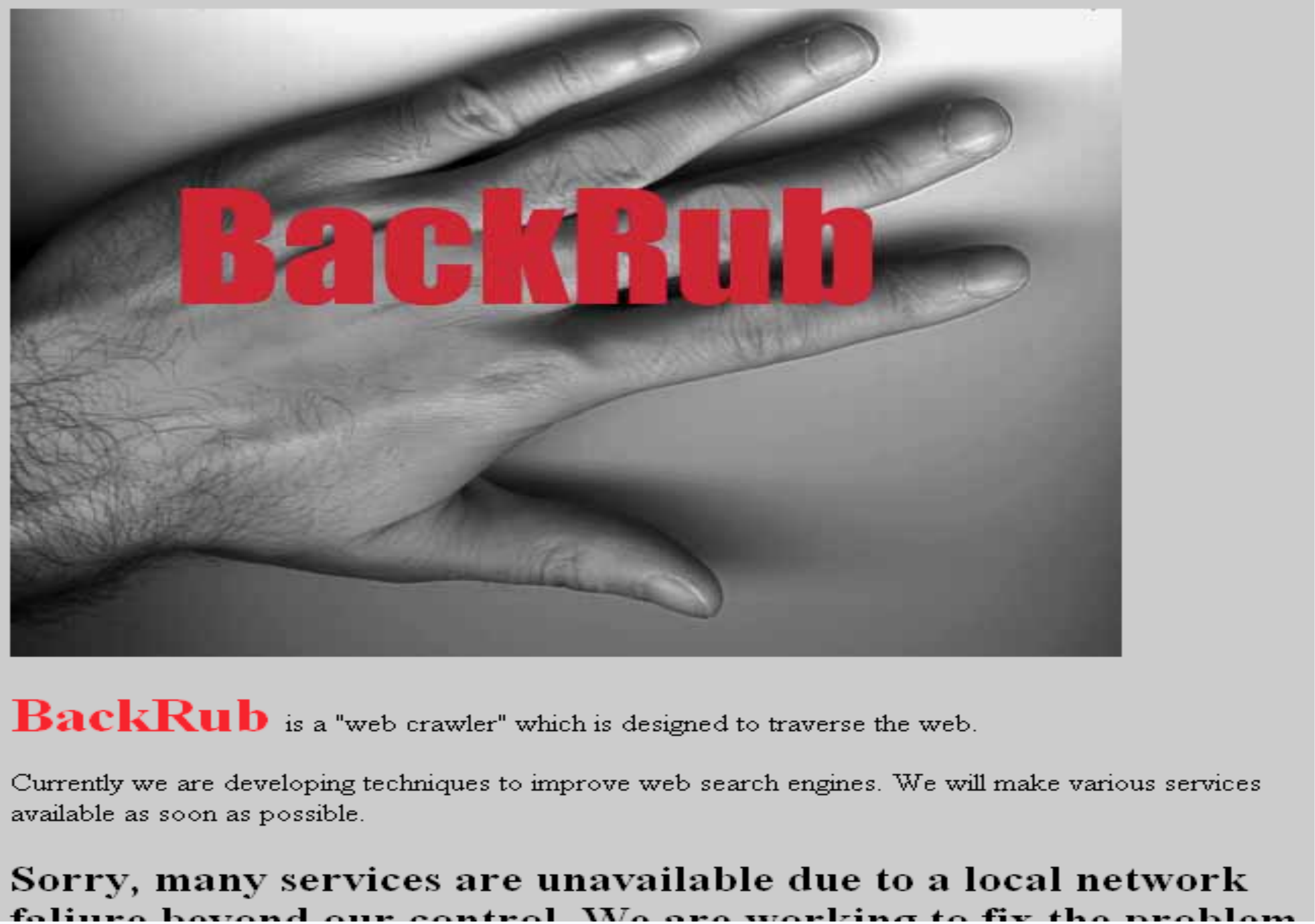
1998
Google Search
SE

Jul/Sept 1998
MSN Search

1993 1994 1995 1996 1997 1998

BackRub

- Utilised PageRank
  - More advanced than previous indexing
- Used back links



BackRub is a "web crawler" which is designed to traverse the web.

Currently we are developing techniques to improve web search engines. We will make various services available as soon as possible.

Sorry, many services are unavailable due to a local network failure beyond our control. We are working to fix the problem

PageRank

- Assumes important websites are likely be linked to
- Similar to a voting:
  - where in links are votes
  - the quality of a vote is determined by the number of votes (in links)
  - your vote is worth more if you have a higher page rank

- Factors Considered:
  - Number of in links
  - Quality of in links
  - Web page's context
- Favoured high keyword density

Keyword Stuffing

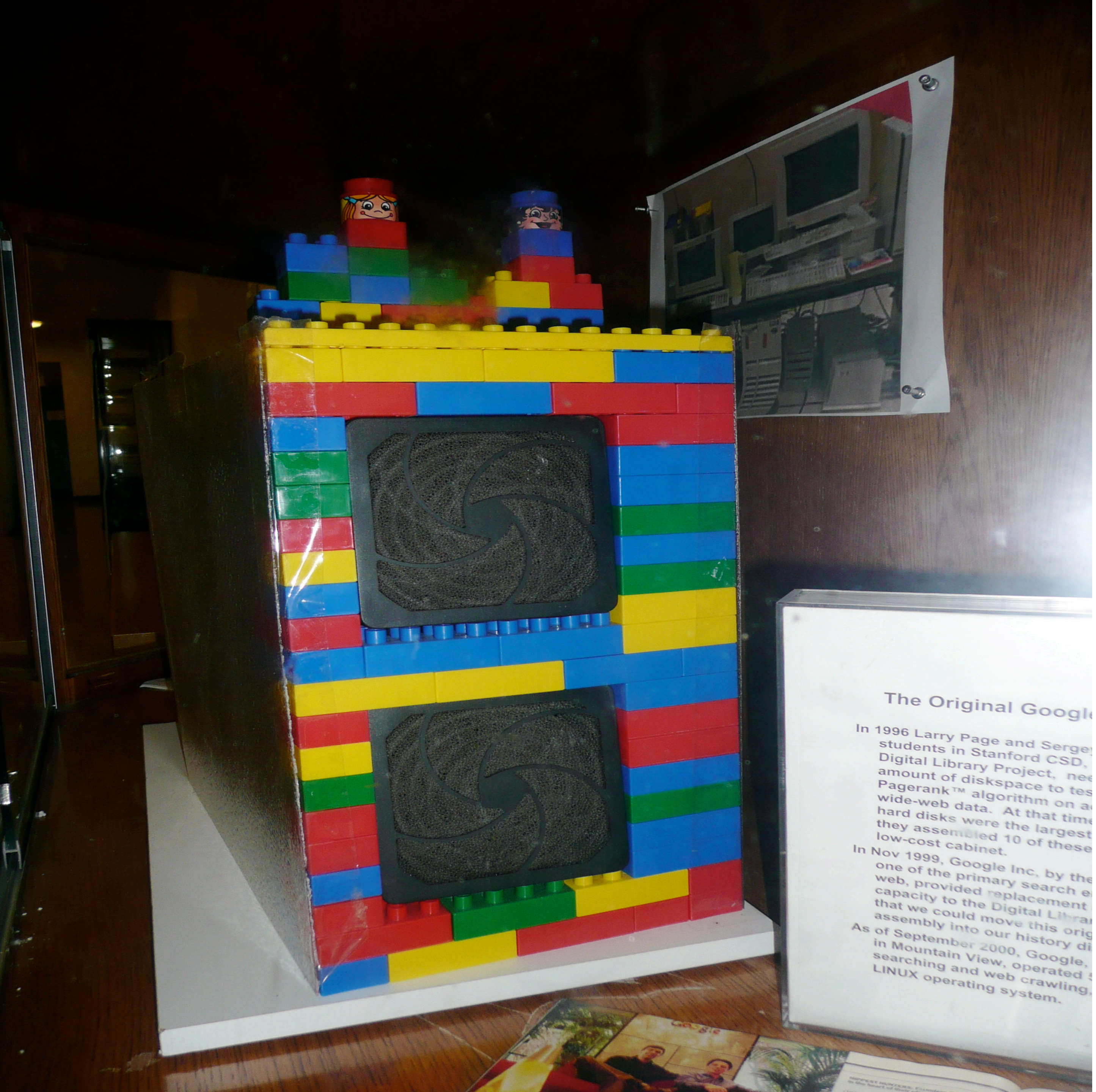
Our company provides software development and consultation services. We are the best software development and consultation services. Our company staff provides software development and consultation services. We offer the best software development and consultation services. Our software development and consultation services is the right solution for your business. Satisfaction with our software development and consultation services is guaranteed. We build your business solution with out software development and consultation services. Contact us and you will receive quality software development and consultation services.

Keyword Stuffing

Our company provides **software development and consultation services**. We are the best **software development and consultation services.** Our company staff provides **software development and consultation services.** We offer the best **software development and consultation services**. Our **software development and consultation services** is the right solution for your business. Satisfaction with **our software development and consultation services** is guaranteed. We build your business solution with **out software development and consultation services**. Contact us and you will receive quality **software development and consultation services**.

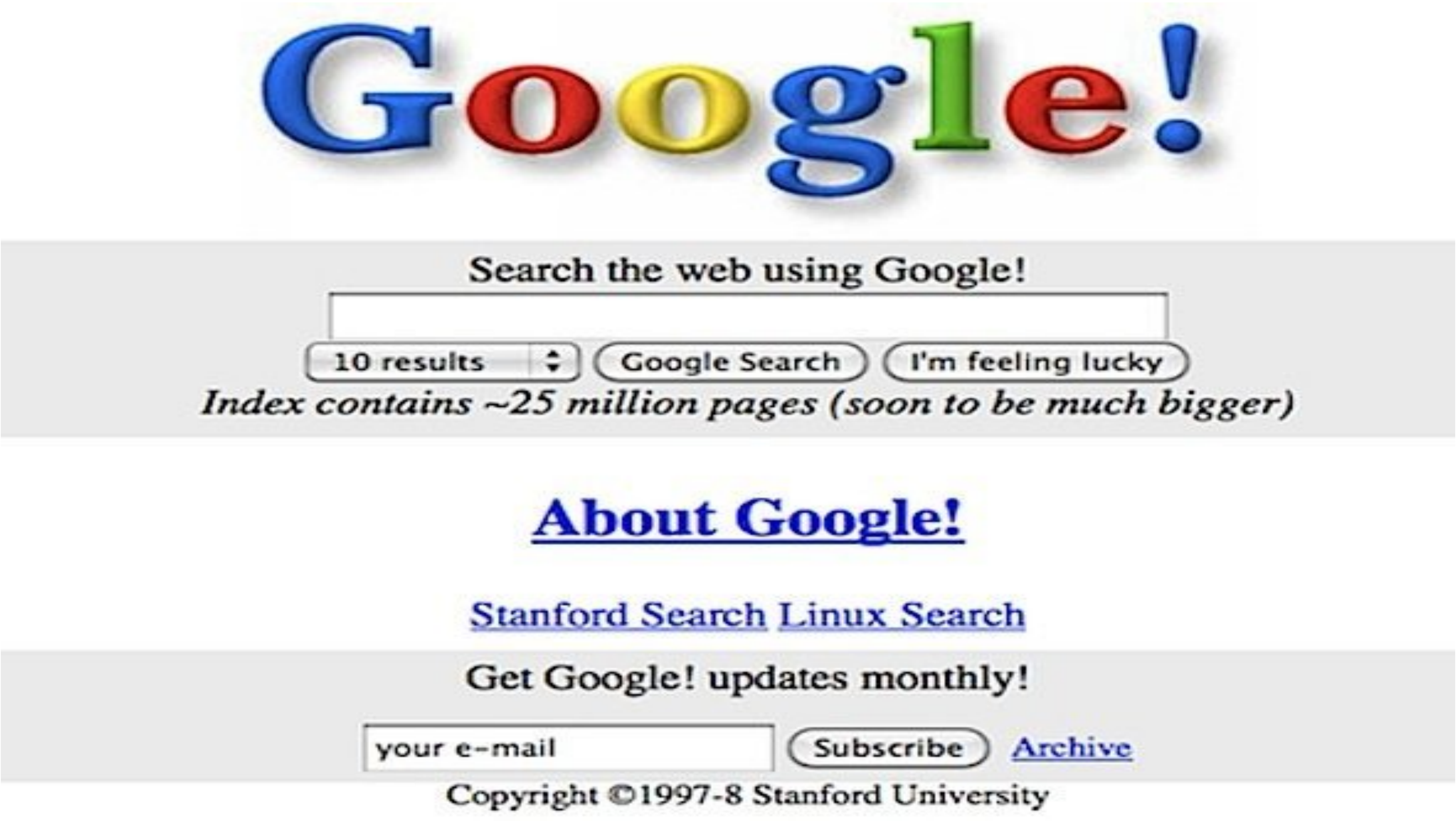
The Original Google Storage

- Larry Page needed a large amount of diskspace to test PageRank
- 10 4GB hard drives


The Original Googl
In 1996 Larry Page and Sergey
students in Stanford CSD,
Digital Library Project, nee
amount of diskspace to tes
Pagerank™ algorithm on a
wide-web data. At that time
hard disks were the largest
they assembled 10 of these
low-cost cabinet.
In Nov 1999, Google Inc. by the
one of the primary search e
web, provided replacement
capacity to the Digital Librar
that we could move this orig
assembly into our history di
As of September 2000, Google,
in Mountain View, operated
searching and web crawling,
LINUX operating system.

Google

Google 1997-8



Search the web using Google!

Index contains ~25 million pages (soon to be much bigger)

About Google!

Stanford Search Linux Search

Get Google! updates monthly!

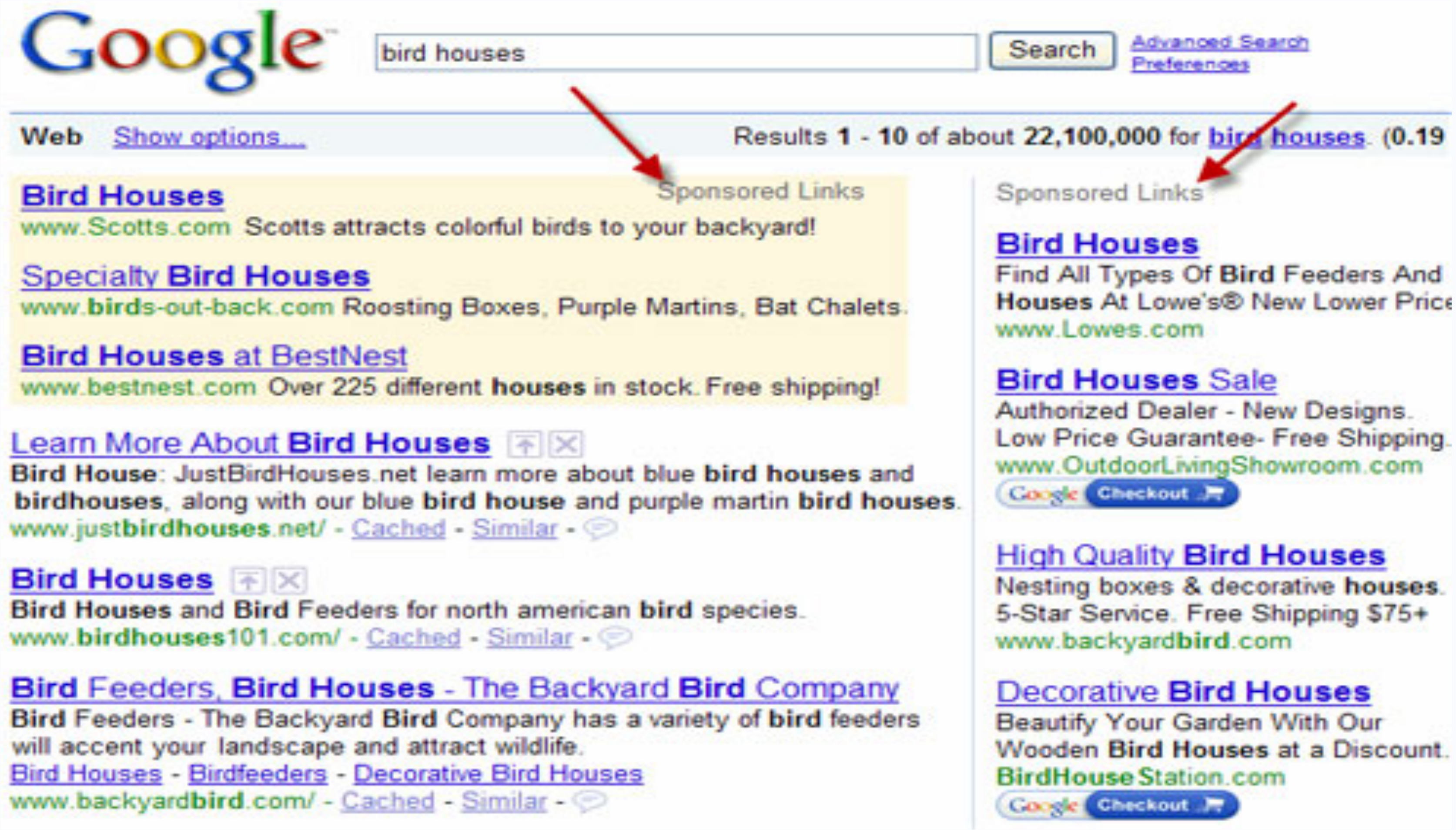
Copyright ©1997-8 Stanford University

The Evolution of Search Engine Features

- Search Engine algorithms are constantly changing
- They strive to improve their:
  - Efficiency
  - How to order the results
    - Freshness
    - Relevancy
  - How to improve user interaction
  - Additional features

University of Southampton

Google Adwords 2000 Funding



Results 1 - 10 of about 22,100,000 for bird houses. (0.19

Sponsored Links

Bird Houses
www.Scotts.com Scotts attracts colorful birds to your backyard!

Specialty Bird Houses
www.birds-out-back.com Roosting Boxes, Purple Martins, Bat Chalets.

Bird Houses at BestNest
www.bestnest.com Over 225 different houses in stock. Free shipping!

Learn More About Bird Houses
Bird House: JustBirdHouses.net learn more about blue bird houses and birdhouses, along with our blue bird house and purple martin bird houses.
www.justbirdhouses.net/ - Cached - Similar -

Bird Houses
Bird Houses and Bird Feeders for north american bird species.
www.birdhouses101.com/ - Cached - Similar -

Bird Feeders, Bird Houses - The Backyard Bird Company
Bird Feeders - The Backyard Bird Company has a variety of bird feeders will accent your landscape and attract wildlife.
Bird Houses - Birdfeeders - Decorative Bird Houses
www.backyardbird.com/ - Cached - Similar -

Sponsored Links

Bird Houses
Find All Types Of Bird Feeders And Houses At Lowe's® New Lower Price
www.Lowes.com

Bird Houses Sale
Authorized Dealer - New Designs. Low Price Guarantee- Free Shipping.
www.OutdoorLivingShowroom.com

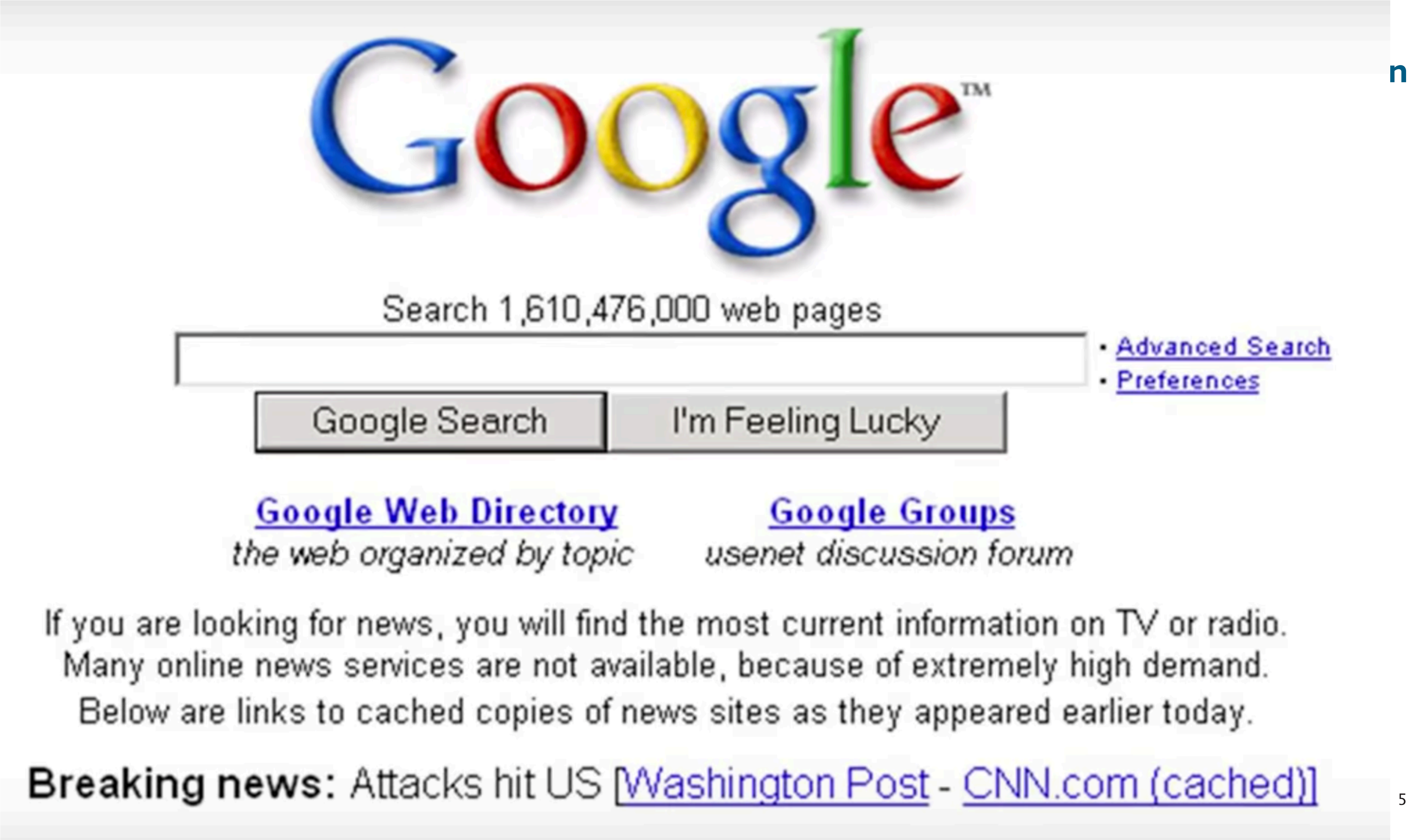
High Quality Bird Houses
Nesting boxes & decorative houses. 5-Star Service. Free Shipping $75+
www.backyardbird.com

Decorative Bird Houses
Beautify Your Garden With Our Wooden Bird Houses at a Discount.
BirdHouseStation.com

Relevancy

- During 9/11 it was apparent that search engines did not cater of time-sensitive queries
- This lead to Google developing Google News
- The freshness of search engine’s index became more important

n



Search 1,610,476,000 web pages

Google Web Directory
the web organized by topic

Google Groups
usenet discussion forum

If you are looking for news, you will find the most current information on TV or radio. Many online news services are not available, because of extremely high demand. Below are links to cached copies of news sites as they appeared earlier today.

Breaking news: Attacks hit US [Washington Post - CNN.com (cached)]

Google 2012

- In 1999, it took Google one month to crawl and build an index of about 50 million pages.*
- In 2012, the same task was accomplished in less than one minute.*
- 16% to 20% of queries that get asked every day have never been asked before.*

*Mitchell, Jon. "How Google Search Really Works." Readwrite. February 29, 2012.

Search Verticals

- In a bid for content and market share, Search Engines develop other search verticals:
  - Books
  - Travel
  - Finance
  - Shopping
  - Scholar

Mobile Search

- By 2015 mobile accounted for more than half of all searches
- Searches need to be context aware (location)
- Limited screen and bandwidth means relevancy is critical
- Boost mobile-friendly pages in ranking algorithm

University of Southampton

Next: Web Standards