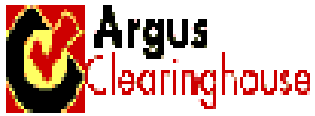
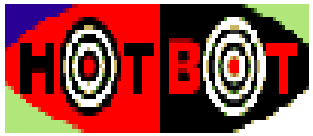
Definition of search engine

A search engine is an information retrieval system (called a wanderer, crawler, robot, worm, spider) designed to help find information on the WWW

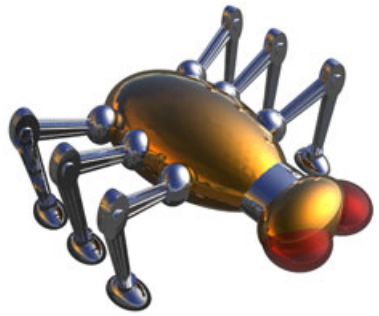
Information may consist of web pages, images and other types of files.





Three parts of search engine

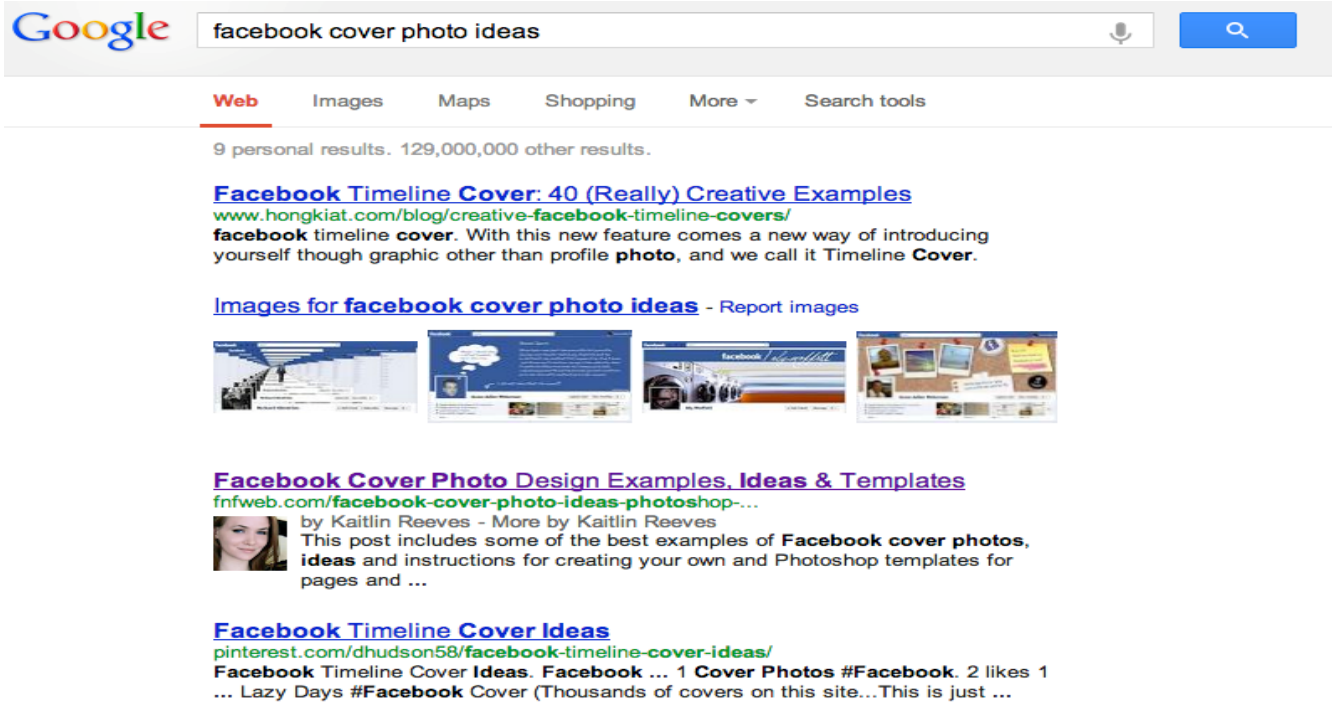
- 1st part: Automated web browser (Spider) follows every www link on a website, collects analyzes the page and parse the words
- 2nd part: The search engine then uses proprietary algorithm to create an index (or a catalog) so that meaningful result are returned for each query



```
<head>
  <meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1">
  <title>
  Time Inc. Content Solutions (TICS)
  </title>
  <meta name="keywords" content="Time, Warner, content, customized, marketing, roi, publishing, newsletters, brand, branding, CRM">
  <meta name="description" content="Utilize the resources of Time to build customized marketing solutions">
  <link rel="stylesheet" href="styles/style.css" type="text/css">
</head>
```

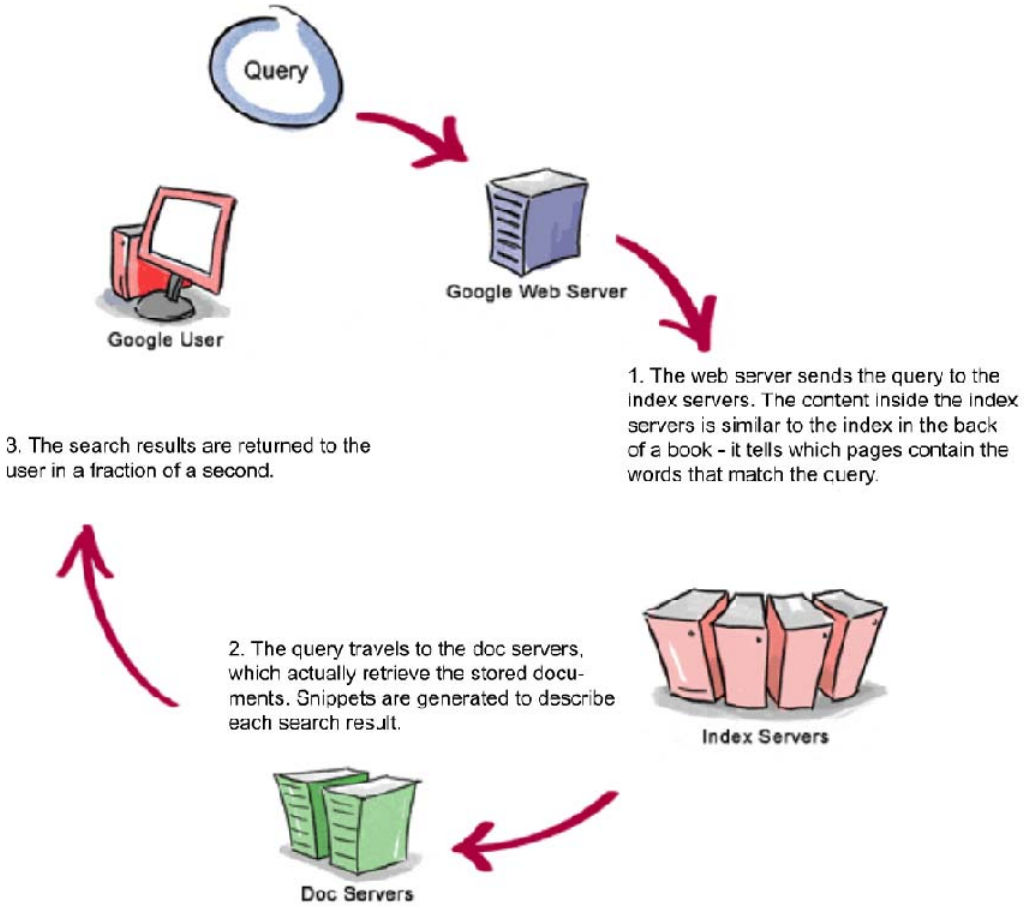
3rd piece of a search engine (Query processing)

- Query processor
  - A program that receives your search request, compares it to the entries in the index, and returns results to you
  - Manages the relevance and ranking

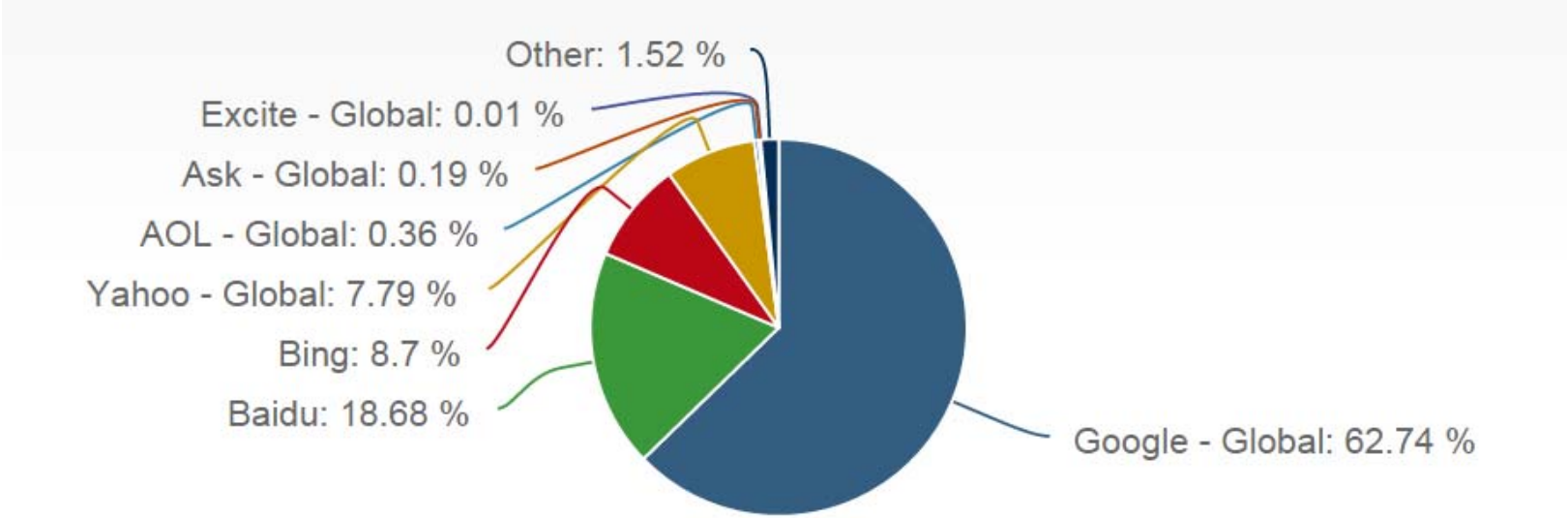


- "On search" by Tim Bray http://www.tbray.org/ongoing/When/200x/2003/07/30/OnSearchTOC

How Query Works

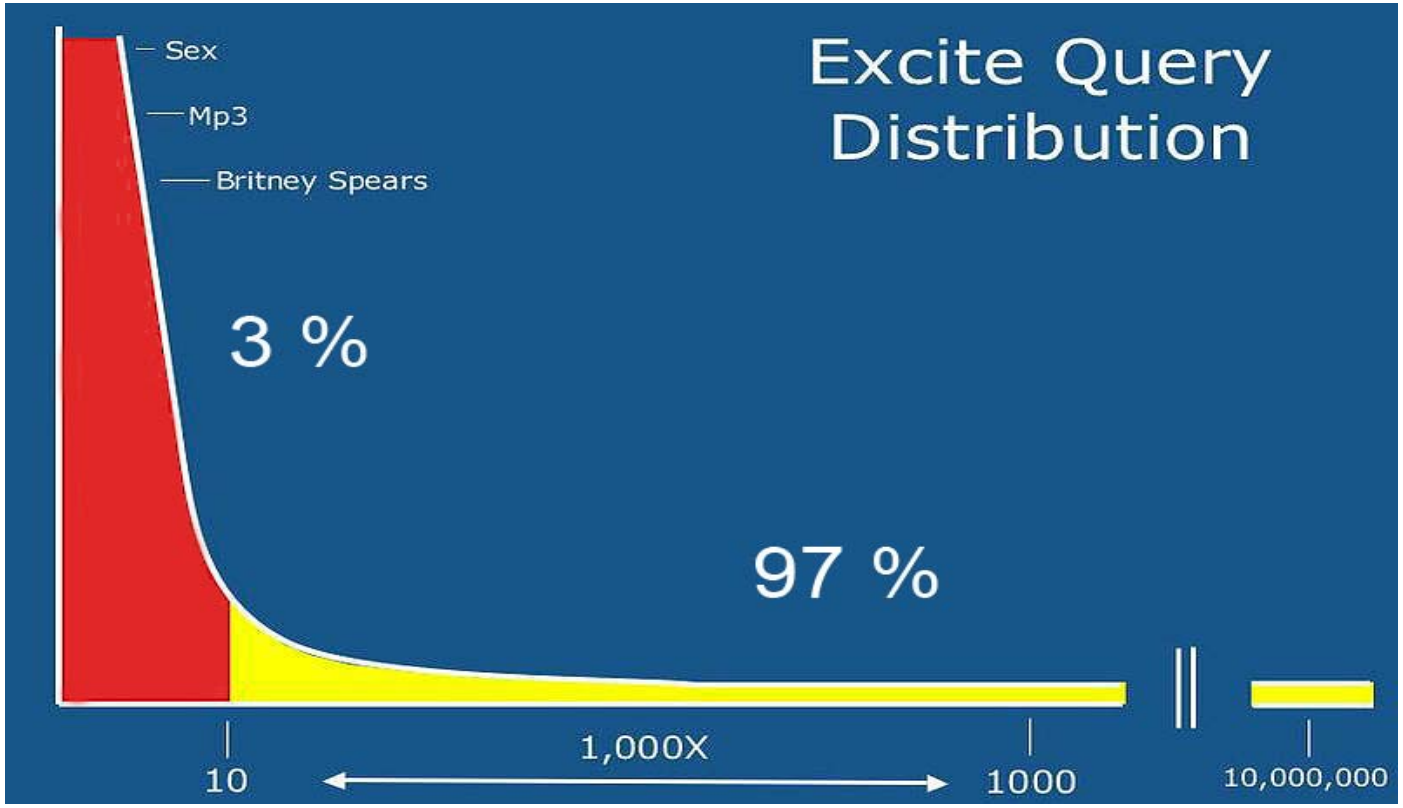


Which search engine people are using?*



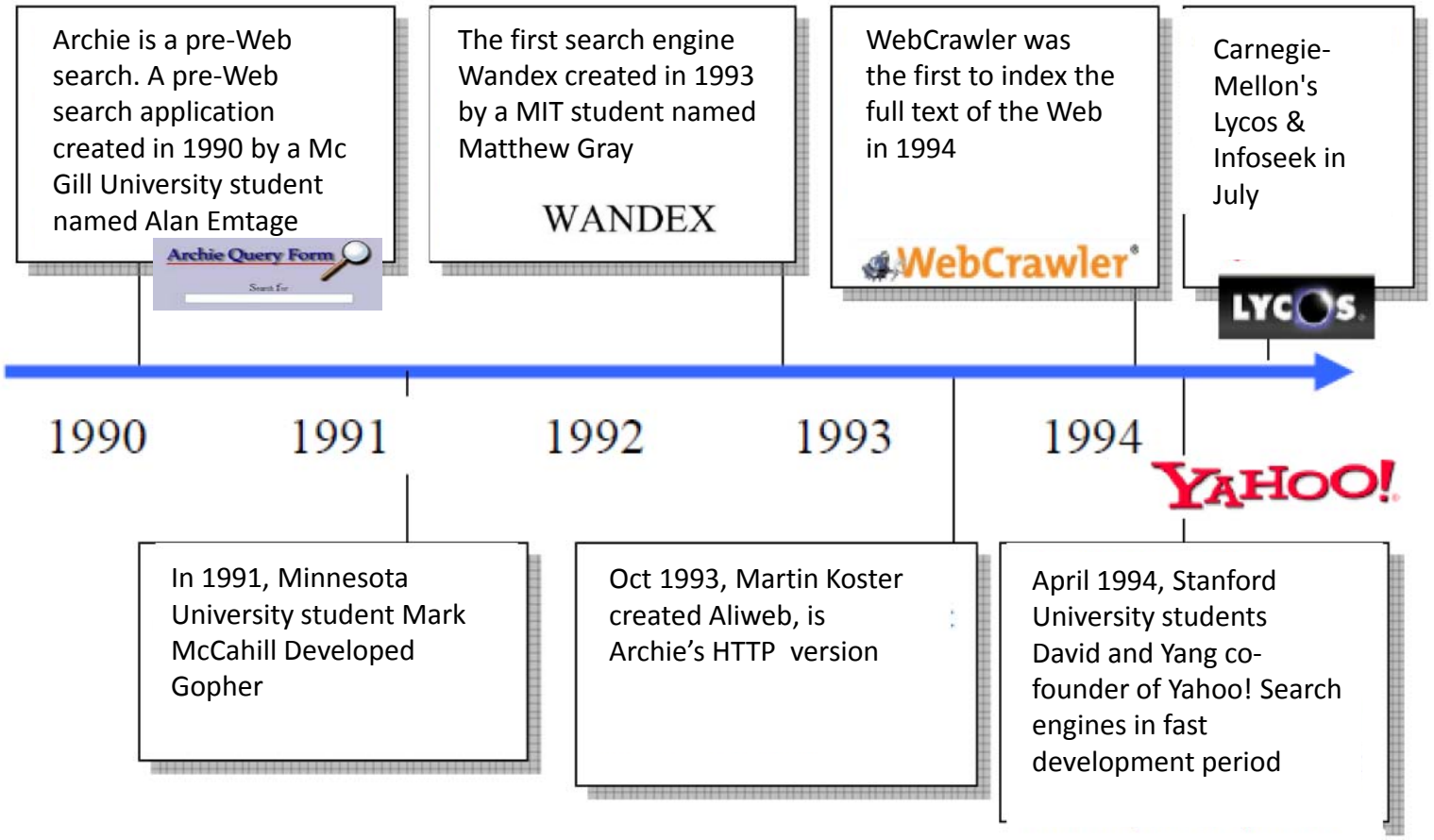
*Report generated January 2015
http://marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4&qpcustomd=0&qpsp=2015&qpnp=1&qptimeframe=Y

What are people searching?



- Google Trends
  - http://www.google.com/trends/2014/

History of Search Engine (1990-1994)*



*adapted from "The Search" by John Battelle

Pre Web Search – FTP (File Transfer Protocol)

- Problems with FTP
  - No organization of FTP Servers
  - User had to know an FTP Server existed
  - User had to visit FTP Server to see files
- Archie (created by Alan Emtage @ McGill Univ. in 1990)
  - Searchable directory of FTP files
  - Searched FTP Servers and indexed their files
  - User searched the Index
- Gopher (created by Paul Lindner & Mark P. McCahill of Univ. of Minnesota in 1991)
  - Connected Gopher servers through the Gopher hierarchy (gopherspace)

Wandex

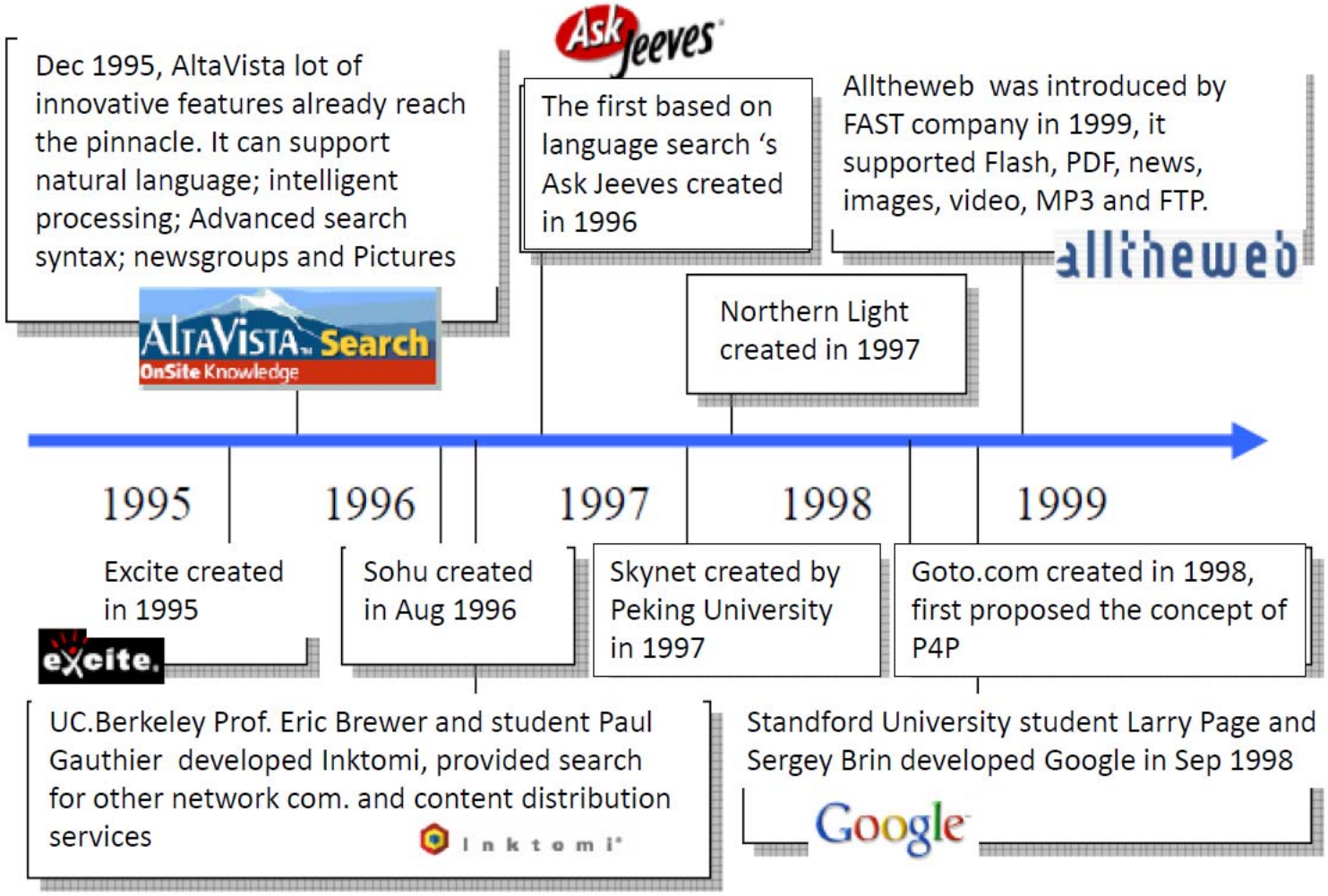
- Wanderer (Matthew Gray's World Wide Web Wanderer)
  - First WWW Engine
  - Designed to track the size of the WWW
  - Captured URL's and entered into database (Wandex)
  - First Robots "bots"

WANDEX

WebCrawler

- Developed by Brian Pinkerton@University of Washington on April 20, 1994
- It was the first crawler which indexed entire pages
- AOL eventually purchased WebCrawler and ran it on their network.

History of Search Engine (1995-1998)



AltaVista

- Developed by DEC (Digital Equipment Corp)
- Index many web pages and provide efficient search
- AltaVista also provided numerous search tips and advanced search features.
- This made AltaVista the first searchable, full-text database of a large part of the World Wide Web
- The “Google of its day”

History of Search Engine (After 2000)

2000: Baidu

2004: Yahoo! Search

2005: MSN Search & ASK.com

2006: Live Search

2009: Bing

http://www.searchenginehistory.com/

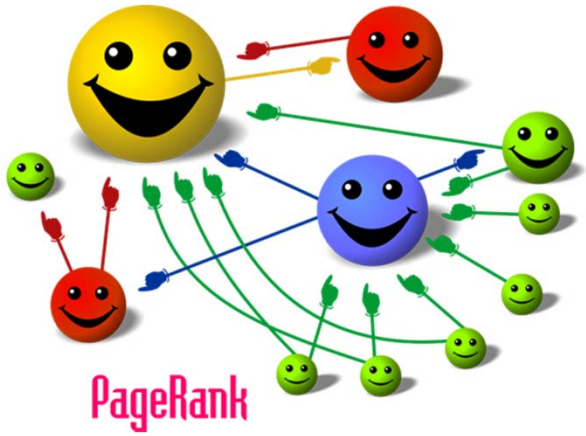
Year	Engine	Event
1993	Aliweb	Launch
1994	WebCrawler	Launch
	Infoseek	Launch
	Lycos	Launch
1995	AltaVista	Launch (part of DEC)
	Excite	Launch
	SAPO	Launch
1996	Dogpile	Launch
	Inktomi	Founded
	HotBot	Founded
	Ask Jeeves	Founded
1997	Northern Light	Launch
1998	Google	Launch
1999	AlltheWeb	Launch
	Naver	Launch
	Teoma	Founded
	Vivisimo	Founded
2000	Baidu	Founded
2003	Info.com	Launch
2004	Yahoo! Search	Final launch
	A9.com	Launch
2005	MSN Search	Final launch
	Ask.com	Launch
	GoodSearch	Launch
2006	wikiseek	Founded
	Quaero	Founded
	Ask.com	Launch
	Live Search	Launch
	ChaCha	Beta Launch
	Guruji.com	Beta Launch
2007	wikiseek	Launched
	AskWiki	Launched

Web Search Until 1998

- In the past, search engine results were often filled with irrelevant content, spam, and other kinds of malicious material
- Because ranking based on "on-page factors“
  - *Keyword tag spam generates poor quality of search results (order)*
- Results were also influenced heavily by marketers with big budgets

Google PageRank

Page and Brin proposed to compute the absolute quality of a page (PageRank) based on the number and quality of pages linking to a page (votes)



Google Search Engine

- Google scanned and ranked several billion Web pages and stored them on its own computers which contain many copied versions of the Internet existed in Google web server
- The pages are organized by subject and every time you click “Search” hundreds of thousands of computers get to work collecting different document links and returning those links to your screen in the blink of an eye
- The search is speeded up because Google stores three copies of all its previous searches, so it doesn’t have to scan the entire Web if two people ask the same question

Search Engine – Political Analogy

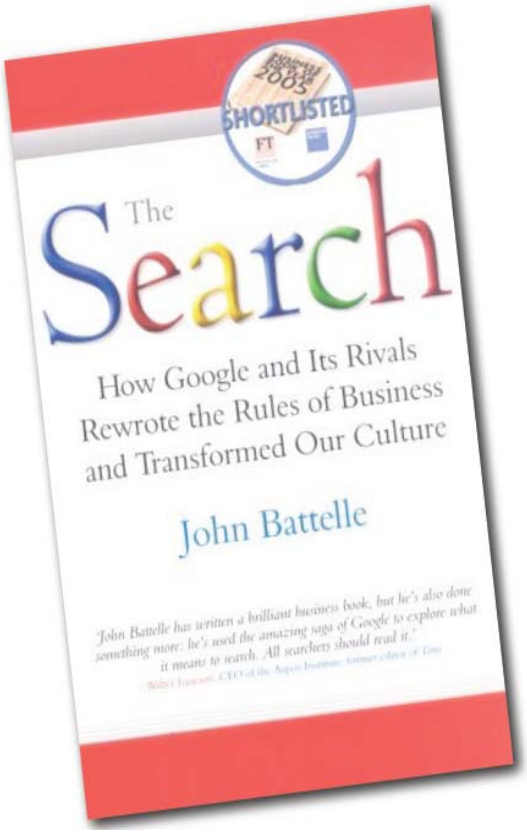
- AltaVista – anarchy
- Yahoo (web directory) – planned economy
- Google – people's power

http://www.archive.org/web/web.php
http://web.archive.org/web/19961017235908/http://www2.yahoo.com/

Other Searches

- Image Search – http://google.com/imghp
- Natural Language Search – Metalab (http://www.youtube.com/watch?v=TJfrNo3Z-DU)
- Real Time search – Twitter (https://twitter.com/search-home)
- People Search – Facebook
- Things/Objects – Pinterest (https://help.pinterest.com/en/articles/guided-search#Web)

Database of Intention



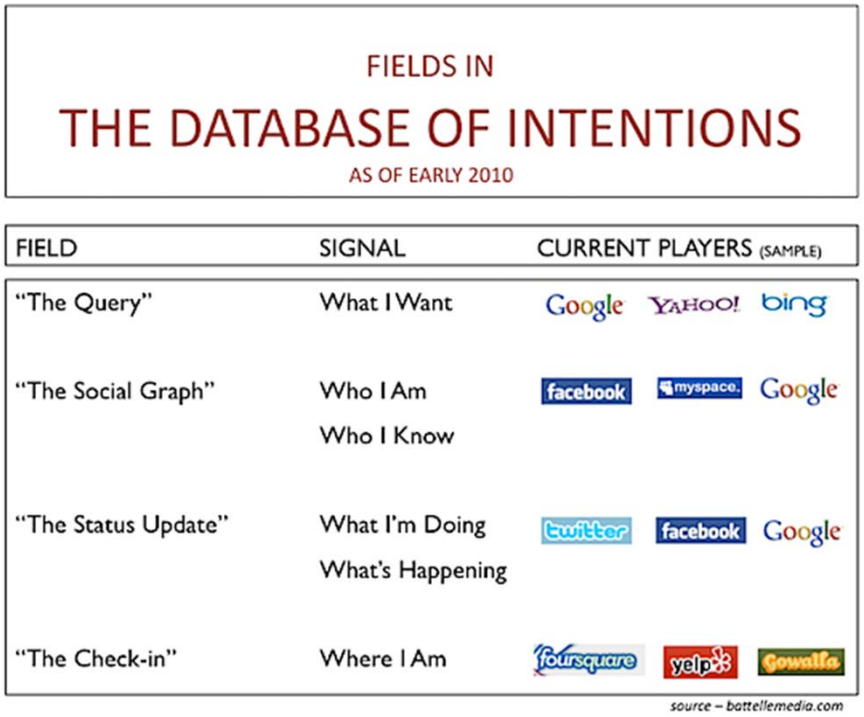
The idea is that every search entry with an Internet search engine contributes to a pattern that can be analyzed and used for prediction. Each search, Battelle notes, offers a hint of what an individual wants to accomplish -- an itch to scratch, a problem to solve, a desire to fulfill**.**

Web searches are "a place holder for the intentions of humankind — a massive database of desires, needs, wants, and likes that can be discovered, subpoenaed, archived, tracked, and exploited to all sorts of ends

The Database of Intentions is simply this: The sum total of all queries that pour into search engines daily, revealing the intricacies and idiosyncrasies of our culture.

HD9696.8.U64 G663 2005

Database of Intention (in the year 2010)*



*John Battelle's blog

Database of Intention*

Field	The Purchase	The Query	The Social Graph	The Status Update	The Check-in	Scholarly
Signal	What I buy	What I look for	Who I know	What I'm Doing What's happening	Where am I	What I'm reading Who are my peers?
The Players	amazon.com TESCO	Google bing YAHOO!	Google facebook LinkedIn	Google facebook twitter	yelp Gowalla foursquare	Google scholar beta MENDELEY RESEARCH NETWORKS crossref UniPHY Connect. Collaborate. Create

*From Wired UK July 2010