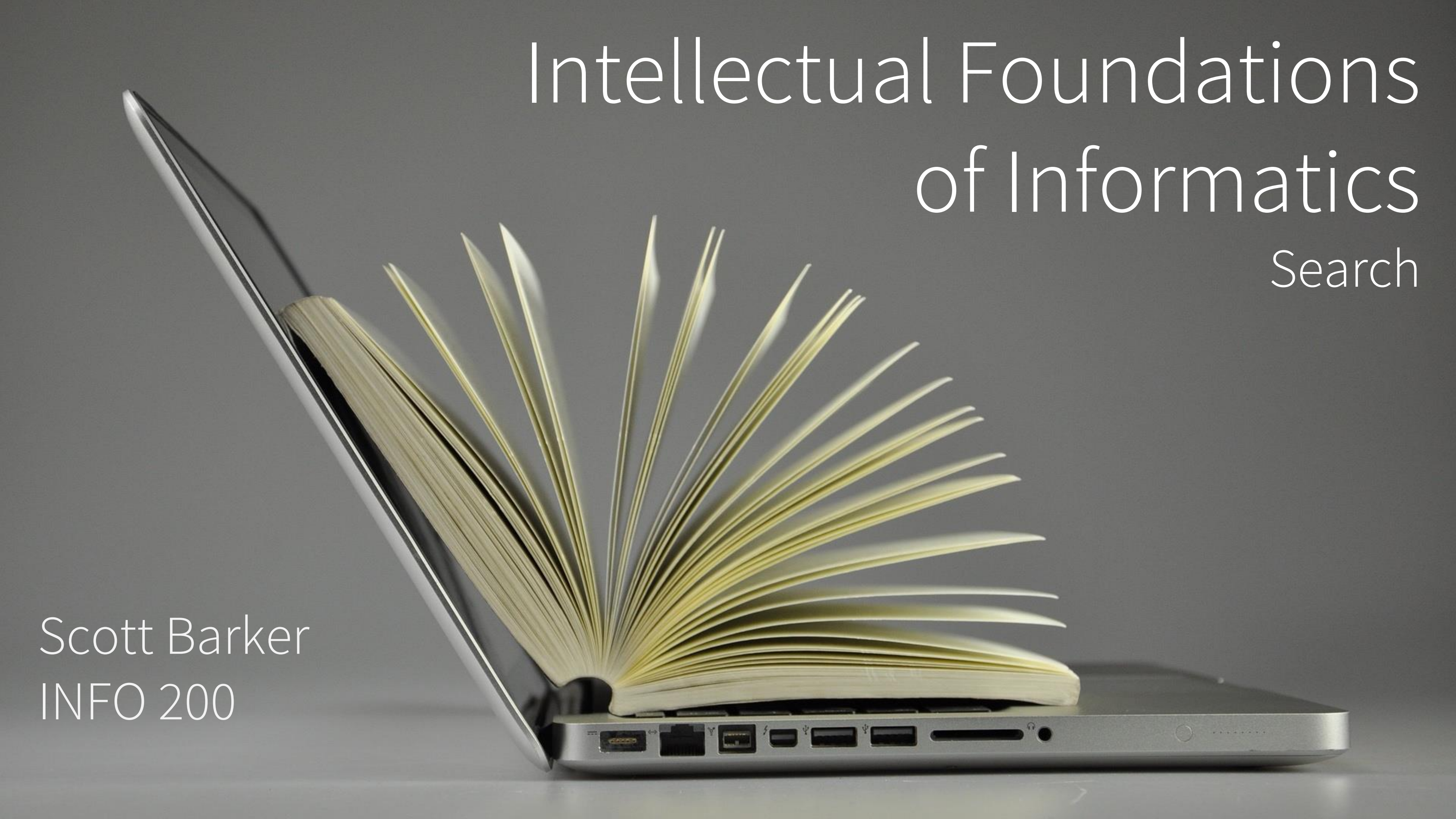
Intellectual Foundations of Informatics

Search



Scott Barker
INFO 200

Sometimes called
Information Retrieval or IR
by academic researchers

What is “Search”
and why is it so important?

Search Engines
like Google, Bing, Baidu can help us to…

- Information to help us make a decision
- Information we want/need (e.g. how to do something, how to get somewhere)
- Information to verify facts or claims we hear
- Entertainment (music, videos, sports scores)
- Other people (friends, family, similar interest/social network, employees)
- Overall they facilitate “Information Seeking” behavior

Search results and ranking may inform (rightly or wrongly) our

- Choices – what we do, what we buy, what/who we like, or how we act
- Politics – who we vote for, who we trust
- Assumptions or biases, what we believe or want to believe

Search Engines and the algorithm’s behind them are not perfect

- Not all information is available through them
- Some countries censor content or ban certain search engines or specific content completely
- Results can be manipulated for commercial, political, or personal gain

A bit of history

In the beginning the web was small, not many sites/pages

Aside…who “invented” the web and around when?

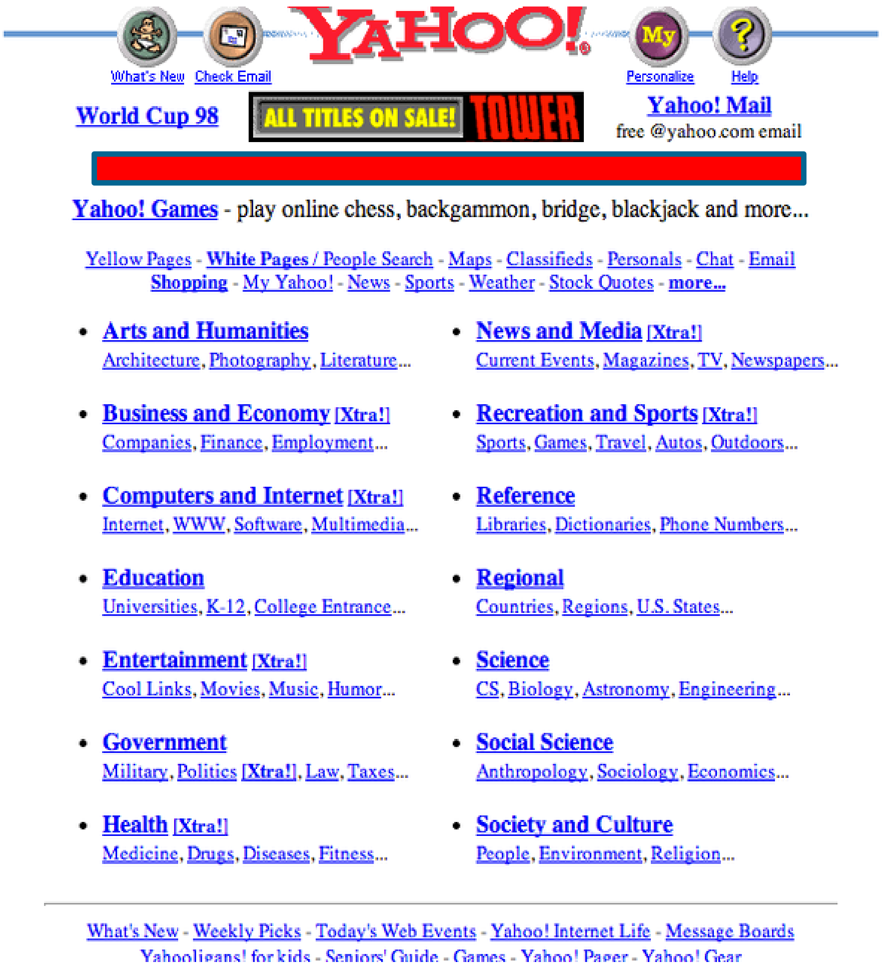
(Sir) Tim Berners-Lee
While working at CERN in Switzerland
1989
Built on a NeXT computer

The first web page was at info.cern.ch

Main idea was to use “hypertext” to make access to documents/information stored on different systems easier/simpler
It wasn’t the only idea – there were other competing things first, like “Gopher”

Finding stuff…

- Since number of sites was small you could keep addresses in your head or write them down
- As sites increased CERN began to keep a list, did some basic categorization of them – list was managed manually
- http://info.cern.ch/hypertext/DataSources/bySubject/Overview.html
- Many sites started using the convention of www….. at the beginning of their name so people could see that they were a "world wide web site" vs. something else (such as a Gopher site)
- Librarians and others had initiatives to "catalog the web", to make things easier to find by topic area
- Some books were published with Internet Resources listed, e.g. "The Whole Internet Catalog"
- Old school Yahoo! Is a good example of a "browsing" type of approach that was used



World Cup 98

Yahoo! Mail
free @yahoo.com email

Yahoo! Games - play online chess, backgammon, bridge, blackjack and more...

Yellow Pages - White Pages / People Search - Maps - Classifieds - Personals - Chat - Email
Shopping - My Yahoo! - News - Sports - Weather - Stock Quotes - more...

- Arts and Humanities
  Architecture, Photography, Literature...
- Business and Economy [Xtra!]
  Companies, Finance, Employment...
- Computers and Internet [Xtra!]
  Internet, WWW, Software, Multimedia...
- Education
  Universities, K-12, College Entrance...
- Entertainment [Xtra!]
  Cool Links, Movies, Music, Humor...
- Government
  Military, Politics [Xtra!], Law, Taxes...
- Health [Xtra!]
  Medicine, Drugs, Diseases, Fitness...
- News and Media [Xtra!]
  Current Events, Magazines, TV, Newspapers...
- Recreation and Sports [Xtra!]
  Sports, Games, Travel, Autos, Outdoors...
- 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...

What's New - Weekly Picks - Today's Web Events - Yahoo! Internet Life - Message Boards
Yahooligans! for kids - Seniors' Guide - Games - Yahoo! Pager - Yahoo! Gear

Sites manually put into categories

Were good and bad aspects to using directories to "browse" and find information:

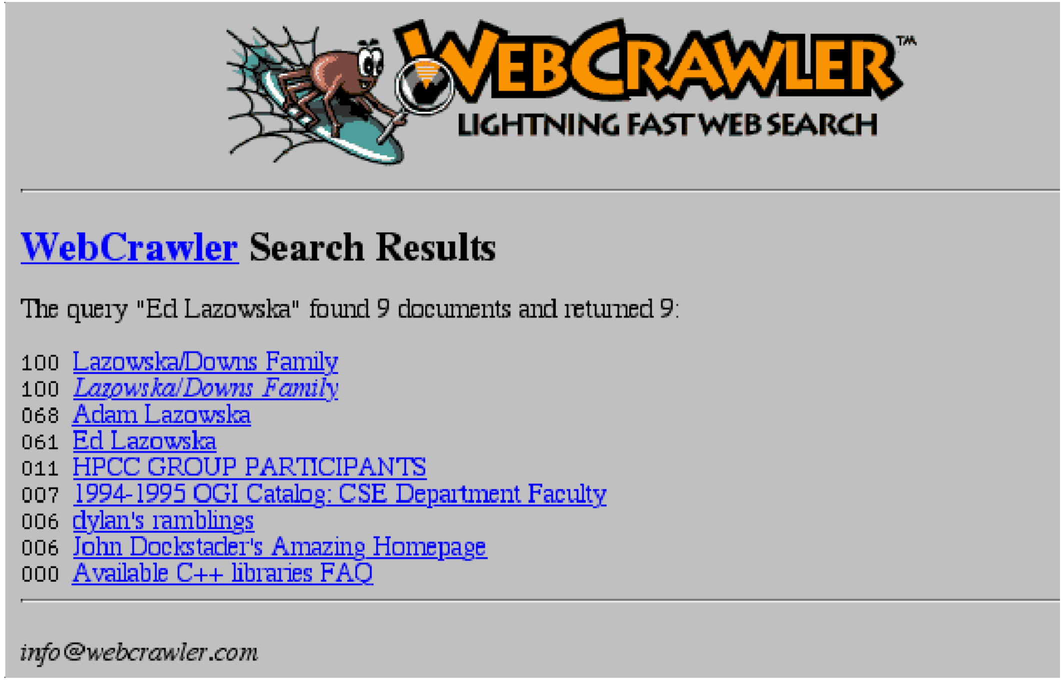
Good – you might find things you didn't know you were looking for

Bad – If you know exactly what you wanted it could be a challenge to find

What category will you find your item under?

Approach not scalable

Enter the “Web crawler” and text search



WebCrawler Search Results

The query "Ed Lazowska" found 9 documents and returned 9:

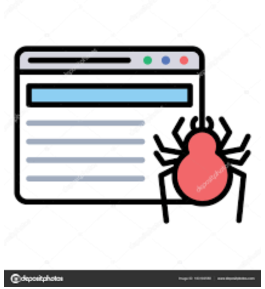
```
100 Lazowska/Downs Family
100 Lazowska/Downs Family
068 Adam Lazowska
061 Ed Lazowska
011 HPCC GROUP PARTICIPANTS
007 1994-1995 OGI Catalog: CSE Department Faculty
006 dylan's ramblings
006 John Dockstader's Amazing Homepage
000 Available C++ libraries FAQ
```

info@webcrawler.com

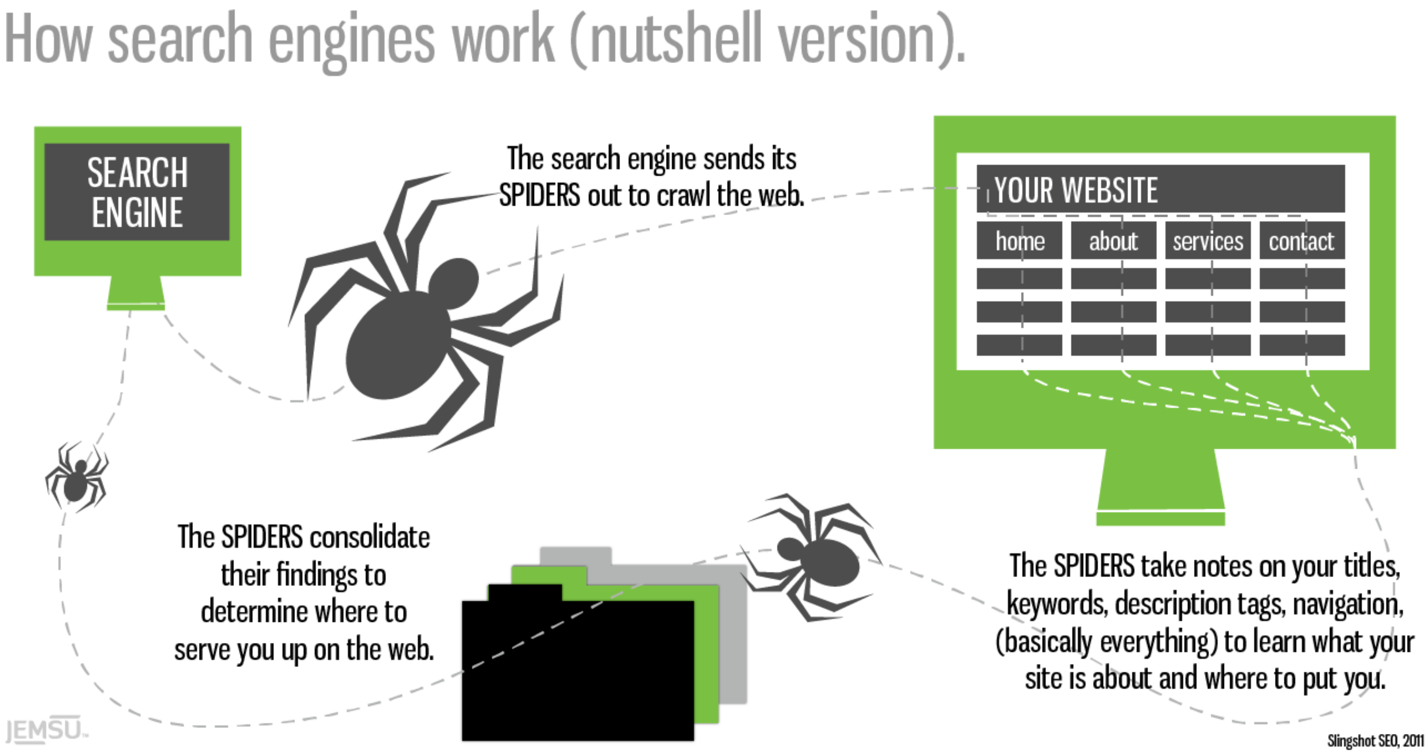
Figure 5.1: WebCrawler search results circa 1995.

One of the first “web crawlers” created was named “WebCrawler” and developed by Brian Pinkerton, here at UW! Went live in 1994

Web Crawler Basic Concept



- Piece of software that would continually visit web pages
- It would note what text was on the page and add it to an index
- Once all the terms on that page were indexed it would look for links to other pages, and follow them to index those pages
- Users could search the index by entering their search term(s) and get a list back of pages that matched
- The software that did the crawling, that provided a UI for users to search, and that found results was termed a “search engine”



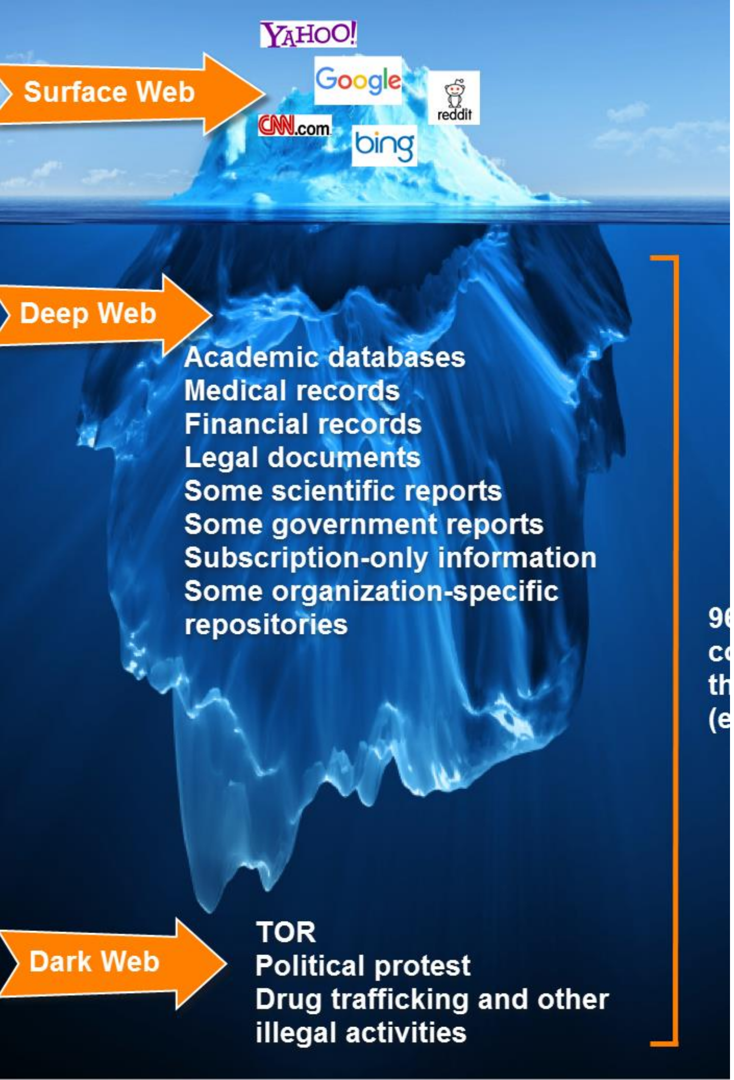
How search engines work (nutshell version).

The search engine sends its SPIDERS out to crawl the web.

The SPIDERS take notes on your titles, keywords, description tags, navigation, (basically everything) to learn what your site is about and where to put you.

The SPIDERS consolidate their findings to determine where to serve you up on the web.

(Graphic by Neil Patel)



Do spiders visit all sites on the Internet, or just a portion?

Just some. A very large percentage of sites on the Internet are not indexed or findable through a search engine

Some sites are not accessible for security reasons (require a login, behind a firewall or on a corporate **intranet**)

Some website owners/developers may instruct spiders to not crawl certain pages or folders via a “robots.txt” file or other mechanism

Some sites may pull data from a database, they aren’t static, so indexing is more difficult

Also note that not all information is on the Internet to begin with, so those items obviously can’t be indexed

When returning results back to users, Search Engines need to consider….

Relevance: how well a retrieved document or set of documents meets the needs of the user (matches what they are looking for, e.g. “jaguar”)

Ranking: The process through which a retrieved item is “ranked” so that the “best” results appear at the top of the list

Why is a site’s ranking so important?

What search engines are most widely used today?

Google

93% market share world-wide, market cap over $1 trillion
Parent company is “Alphabet”

Bing

Default search in Windows, 2.4% share world-wide

YAHOO!

Baidu 百度

Ask Jeeves
Ask.com

AOL

WolframAlpha

DuckDuckGo

Google in China

- In 2009, one third of all searches in China were on Google
- By 2013, 1.7% of all searched in China were on Google
- Why?
- 2012 China blocked access to Google, Gmail, and all other Google services such as YouTube. China had asked Google to remove certain items from search results and Google refused.
- According to some, this was an effort to censor/control information that the government did not agree with or information that might paint the government in a bad light.
- The Chinese government says this was done to protect citizens from harm
- China also blocked Facebook, Twitter, Instagram, many others, again in the name of protecting citizens
- Today often referred to as the “Great Chinese Firewall”
- Some Chinese citizens use a VPN (Virtual Private Network) to bypass the “Great Chinese Firewall”, some report that China has now made VPN use illegal although I can’t find a specific law for sure

Who are these guys?

Sergey Brin
Co-founder of Google
Net worth: $52 billion

Larry Page
Co-founder, Google
Net worth: $53 billion

Google founded when Brin and Page were PhD students at Stanford University in 1998

Google Wasn’t the first but…

- They had “special sauce” in terms of how they “ranked” results

- This allowed users to have a better experience and get the results they were looking for easier (at or near the top of the list)

- What is that “special sauce”?

- **Page Rank** – an algorithm that took into account a site’s popularity, it didn’t just look at the terms on the page

“The genius of Google is that its creators didn't come up with a great organizational scheme for the web.

Instead, they got everyone else to do it for them.”

~James Grimmelmann, law professor

Let’s watch this video on
“How Internet Search Works”

Demo – ranking difference

- Search for “iSchool” using Google – where does UW land in the ranking?
- Search for “iSchool” using Bing – where does UW land?
- While on Bing – look at the right side of the page…
- Why are those things listed there and in the order they are listed?
- Go back to Google and search “tires”
- Notice results at the top – what are these?

Ranking algorithm questions

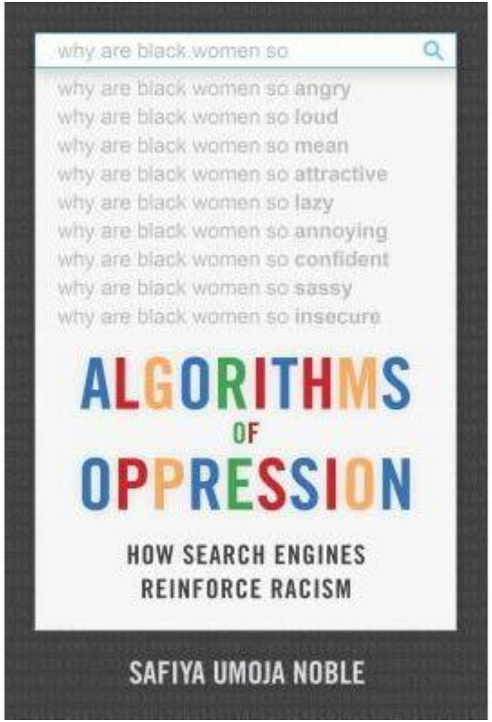
- Let’s search for images of Nurses and Programmers
- Approximately 90% of Nurses are female, 80% of Programmers are male
- Should a ranking algorithm take these statistics into account when returning results?

- Could it perpetuate stereotypes by the images that are shown or not shown?
- Let’s search for “librarian”, any stereotype in those images?

While Google typically produces great results, it isn’t perfect

- At one time Googling images of Black women presented users with racial stereotypes, up to and including images of apes.

- For a long time Googling “Martin Luther King” (sans Jr.) produced a top rank result to a Stormfront-hosted (white supremacist) website.



why are black women so

why are black women so angry
why are black women so loud
why are black women so mean
why are black women so attractive
why are black women so lazy
why are black women so annoying
why are black women so confident
why are black women so sassy
why are black women so insecure

ALGORITHMS OF OPPRESSION

HOW SEARCH ENGINES REINFORCE RACISM

SAFIYA UMOJA NOBLE

Advertising revenue and Sponsored Links

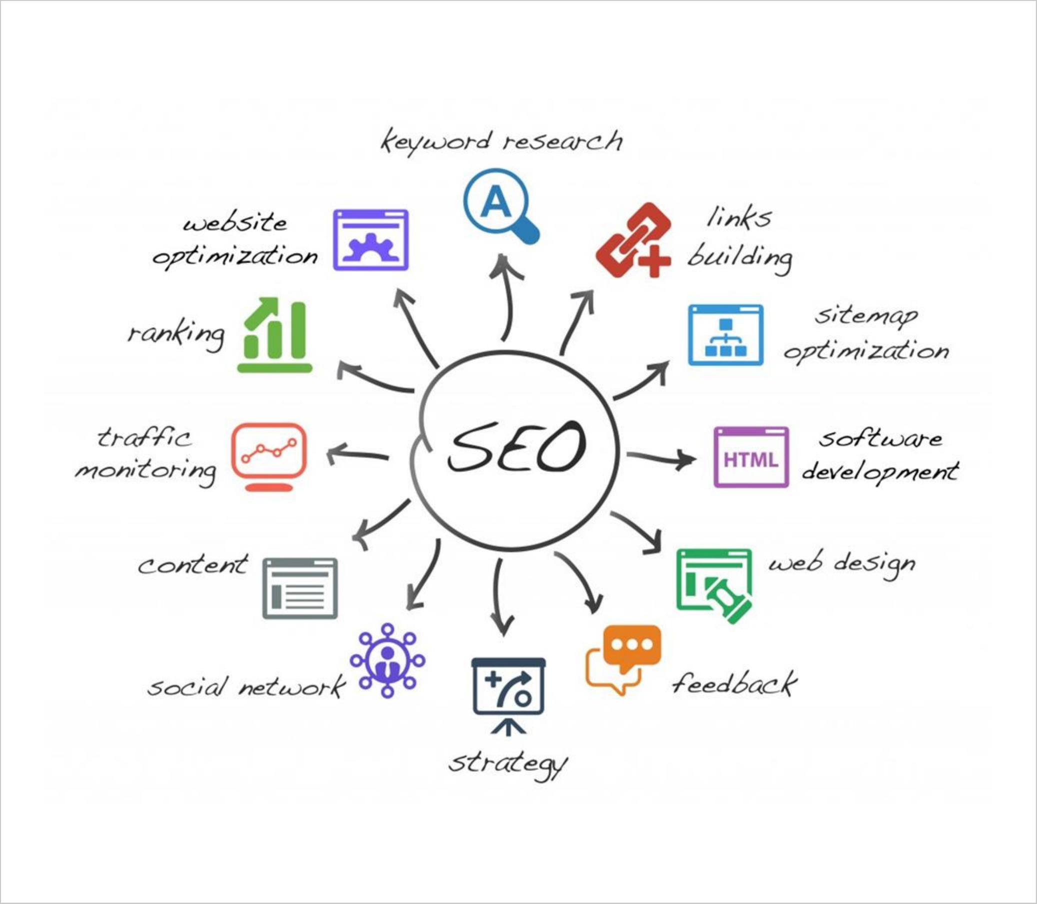
- If you want to promote your site even more, you can pay
  - ads.google.com
  - bingads.microsoft.com
- Typically you pay more to be listed higher in the ranking, and pay per click

- Google made $133 billion dollars in 2019 from sponsored ads!   Most expensive adwords/keywords?

- That is another reason why “search” is so important, search is **big business**

Search Engine Optimization (SEO)

Because a site’s ranking is so critical to beiong found, a whole field has emerged to help sites appear “higher” on search engine result lists, called SEO



Some advanced Google search options

- Putting your search in quotes, (e.g. “this search string”) will search for that exact phrase, in that order.
- You can use + in lieu of the word “and” to tell Google to connect two search terms. (default is “or”)
- You can use - to remove certain words from consideration in Google’s search. (e.g. Jaguar -vehicle).
- Use DEFINE: x to define a word (including slang!)
- Search images from your computer by dragging them into the search bar of Google image search. Google will show you similar images.
- Use the Google Advanced Search page: https://google.com/advanced_search
- Go to images.google.com and Google Atari Breakout (without quotes). Have fun!
- Fun – sometimes “Easter Eggs” are present, search “Wizard of Oz” and click the slippers!

Other Search options…

- In addition to Google/Bing etc. search, there are other resources for finding information

- UW Library Databases on many topics – sources are high-quality, many are peer reviewed (so information is likely to be more accurate), many of these resources are part of the "deep web" that we referred to earlier

- These are targeted sources of information, you are not searching the whole web but typically well-known/trusted sources of information

- Remember – not everything available through Google or other search engines!

- Be sure to use these databases when working on papers for this class or others!

End