How Does Google Search Work?

It’s complicated

http://www.achievement.org/achiever/sergey-brin/

Sergei Brin

- Born: 1973
- 13th richest person in the world (~$39.8billion)

Larry Page

- Born: 1973
- 12th richest person in the world (~$40.7billion)

- Met while they were both PhD students
- Together created the first version of Google search in their dorm room
- In 1996, their project was called “Backrub”
- It seemed to work pretty well!
- Originally wanted to make a free academic open source search engine that researchers could examine
- But it worked so well that they soon turned Google into a for-profit company

Overview

- This presentation is based in part on this: https://www.google.com/search/howsearchworks/

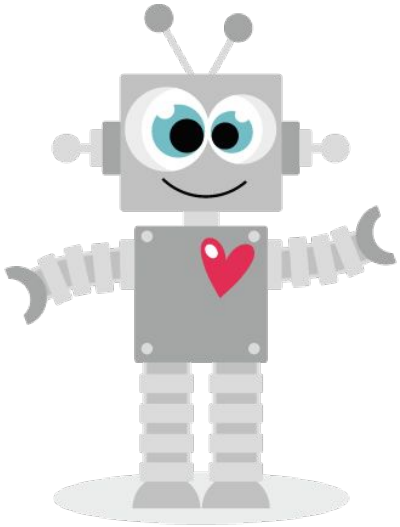
- Crawling and Indexing
  - robots.txt
  - Sitemaps
- Search Algorithms
  - PageRank
- Presenting results

Crawling and Indexing

- Special programs called **spiders**, or **crawlers**, scan web pages for hyperlinks
- A spider finds every link on a page, and then follows it and scans that page, and so on and so on …
- Google stores the pages it stores in its own **indexes** (i.e. databases)
  - When you search “the web”, you are really searching Google’s copy of it
  - Every word on every page is saved (so that they can be searched for)
  - Google crawls 100s of billions of pages, over 100 million Gigabytes of data!
  - They also store all search words --- 85% of all Google searches have been done before
- Website owners often do two things to help spiders: provide a **robots.txt** file, or a **sitemap**

robots.txt: Robots Exclusion Standard

- If you own a website, you can put commands for spiders into a **robots.txt** file
- You could tell a spider to ignore certain files and folders you don't want to be indexed
  - But if *other* sites link files you want ignored, they will be searchable!
- Google and most reputable crawlers respect **robots.txt** directives
- But "bad" robots, such as email harvesters, spambots, malware, etc. typically ignore **robots.txt**



Robots.txt: Robots Exclusion Standard

Here's a sample robots.txt that tells crawlers not to index three particular folders:

```
User-agent: *          # all crawlers should follow these rules
Disallow: /cgi-bin/    # don't go into this folder
Disallow: /tmp/        # don't go into this folder
Disallow: /junk/       # don't go into this folder
```

Robots.txt: Robots Exclusion Standard

Here's a sample **robots.txt** that tells all Google crawlers not to index a particular folder:

```
User-agent: googlebot          # all Google services
Disallow: /private/            # disallow this directory
```

Sitemaps

- A **sitemap** is a list of URLs that a website can provide to help a web crawler index it
- It's an **XML** file
  - **XML** stands for "eXtended Mark-up Language"; it is a generalized version of HTML, but allows you to make custom begin/end tags
- URLs might include other useful information, such as when a link was last updated, how frequently it changes, and how important it is
- Sitemaps are, for example, useful for websites with
  - lots of non-traditional content that is hard for web crawlers to index (e.g. highly interactive JavaScript pages)
  - unlinked portions that are difficult to find by following links

Sitemaps

```
<?xml version="1.0" encoding="utf-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.sitemaps.org/schemas/sitemap/0.9
http://www.sitemaps.org/schemas/sitemap/0.9/sitemap.xsd">
   <url>
      <loc>http://example.com/</loc>
      <lastmod>2006-11-18</lastmod>
      <changefreq>daily</changefreq>
      <priority>0.8</priority>
   </url>
</urlset>
```

SEO: Search Engine Optimization

SEO is the general name given to getting your website highly ranked in a search engine

SEO: Search Engine Optimization

White-hat SEO: using “approved” techniques for getting noticed, e.g.

- Interesting content that people want
- Good design
  - Google provides various tool and instructions for making your web site work well with it: https://developers.google.com/webmasters/googleforwebmasters/
- No deception

http://www.freestatepress.com/basic-white-hat-seo-practices/

SEO: Search Engine Optimization

Black-hat SEO: using “dis-approved” techniques for getting noticed, e.g.

- Scraper sites: websites created (often automatically) by copying (scraping) content from other sites
- Link farms: collections of websites that serve no purpose other than to link to each other in an effort to raise their rankings in search
- For more tricks see: https://en.wikipedia.org/wiki/Spamdexing

http://www.greatdentalwebsites.com/3-questions-ask-avoid-blackhat-seo/

Search Algorithms: Word Analysis

- Google does many things to help retrieve the pages that are most likely relevant to you
- Automatically detect and correct spelling mistakes
  appl → apple
- Synonym detection
  - “How to change a light bulb”, change = replace
  - “How to change laptop brightness”, change = adjust
- Special keywords
  “movies” → infers you mean movies playing locally in theaters

Search Algorithms: Matching and Ranking Pages

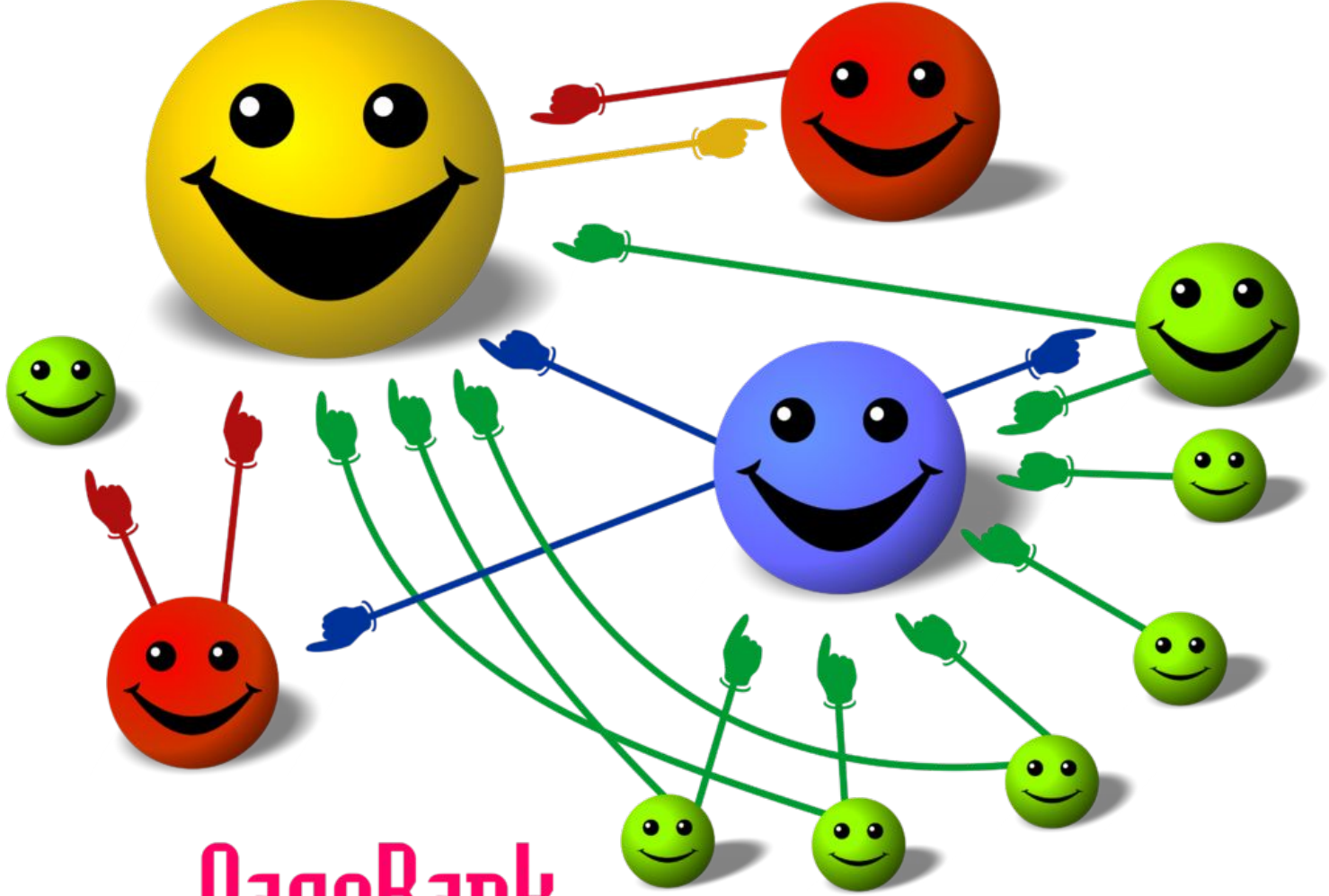
- Google searches its index for all pages that contain your search words
- It asks 100s of questions to help rank the pages, e.g.
  - How frequently do your search word appears in the page?
  - Do your search words appear in a header?
  - Does the page have pictures or videos of thing you are searching for?
  - Is the page written in the same language as your search words?
  - When was the last time the page was updated?
  - Does the page have a good user experience?
  - Do users who have done similar searches think the site is good?
  - Is the site using “spam” techniques?
  - What country are you searching from?
  - What sort of words are in your search history?

Google’s Big Idea: PageRank

- **PageRank** is the name of a key technique Google uses to help rank pages returned in a web search
  - It’s named after Google co-founder Larry Page
  - When it Google first appeared, the relevance of its results were much better than most other search engines
- According to Google:

  “PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.”
- Calculating PageRank of a page is tricky because you need to know the PageRank (i.e. the quality) of all the pages that point to it
  - Figuring out how to efficiently calculate the PageRank of billions of web pages

    Google’s key technical successes



Google’s Big Idea: Pagerank

- Roughly speaking, each link to a website counts as a vote for that page
- But not all votes are equal: the value of a link depends upon the PageRank of the page where the link comes from
- So it is possible, for example, that a page with a lot of links from low-quality websites could be ranked lower than a page with only a few links from high-quality websites
- Overall and in general, PageRank tends to do a very good job and returns highly relevant results for most searches that users do
- It can be “gamed” by spammers, and so Google frequently tweaks and updates the exact details of how PageRank weights pages

Presenting Results

- The traditional Google results are a list of pages that match the words in your search, ordered by relevance
- But sometimes there are different kinds of results, e.g.
  - “what is 35 times 6” → 210 (in a calculator)
  - “how's the traffic” → shows a map of local traffic conditions around current location
    - Google figures out traffic conditions by using info from cell phones and other devices in cars driving on roads near you
  - “weather” → shows local weather report from weather.com
  - “who invented css” → shows a snippet from the Wikipedia page for Håkon Wium Lie
  - “female astronauts” → shows a list (with pictures) of famous female astronauts