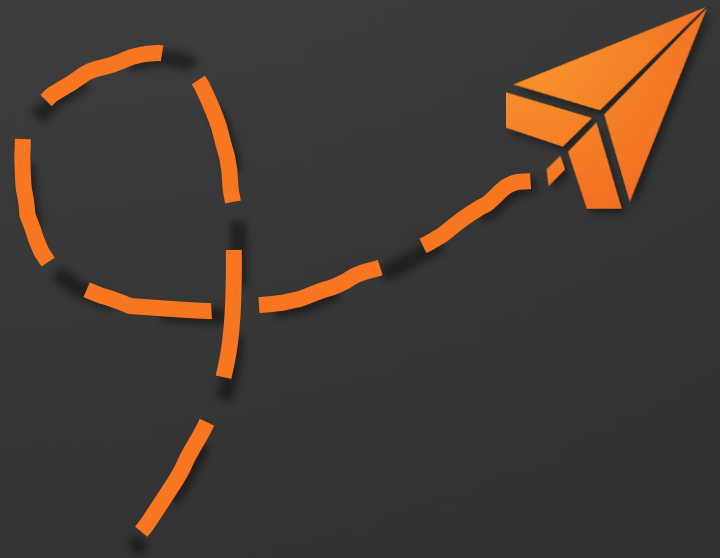
IHU

Search Engines and Algorithms

crawling-indexing-ranking



Digital Marketing

Search Engines

How they work?

(crawling-indexing-ranking)

Search Engine Terms

- **SERPS**: Search Engine Results Pages, the pages you will see after typing a search phrase into a search engine (Google, BING, YAHOO, etc.)
- **Spider (crawler or robot)**: Software programs used by SEs to index all pages of a specific website. Spiders will follow links from one page to another (via internal links) and will revisit the site when new articles are published.
- **Indexing:** the act of adding information about a web page to a search engine's **index**.
- **Index**: a collection of web pages(a database)that includes information on the pages **crawled** by search engine spiders.
- **Ranking:** Search engines apply algorithms to help **rank** those indexed web pages Then it decides which web pages to show from the **index** and in what order.

How SEs work?

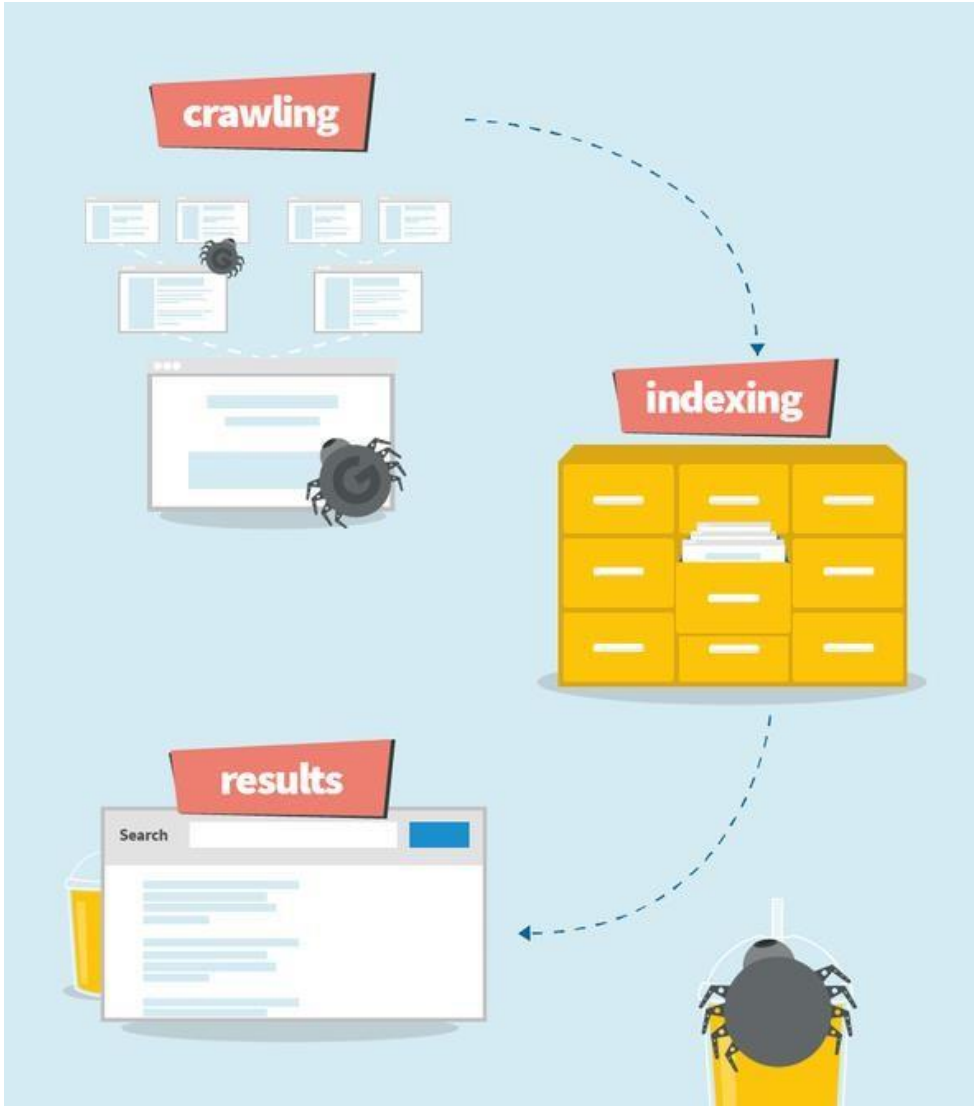
In simple words

A search engine like Google consists of a crawler, an index, and an algorithm.

The **crawler** follows the links.

When Google's crawler finds your website, it'll read it and its content is saved in the **index**.

Ranking in the search engines requires a website with flawless TECHNICAL SEO.



SE Crawling

*…The **crawler** follows the links…*

- **Robot.txt file:** All search engine crawlers begin crawling a website by downloading its robots.txt file, which contains rules about what pages search engines should or should not crawl on the website.

- **Sitemaps:** Another way that search engines can discover new pages is by crawling sitemaps. Sitemaps contain **sets of URLs**, and can be created by a website to provide search engines with a list of pages to be crawled.

- **Page submissions:** Alternatively, individual page submissions can often be made directly to search engines . This manual method of page discovery can be used when new content is published on site, we want to minimise the time that it takes for search engines to see the changed content.
  But: Google states that for large URL volumes you should use XML sitemaps.

SE Crawling

Useful information:

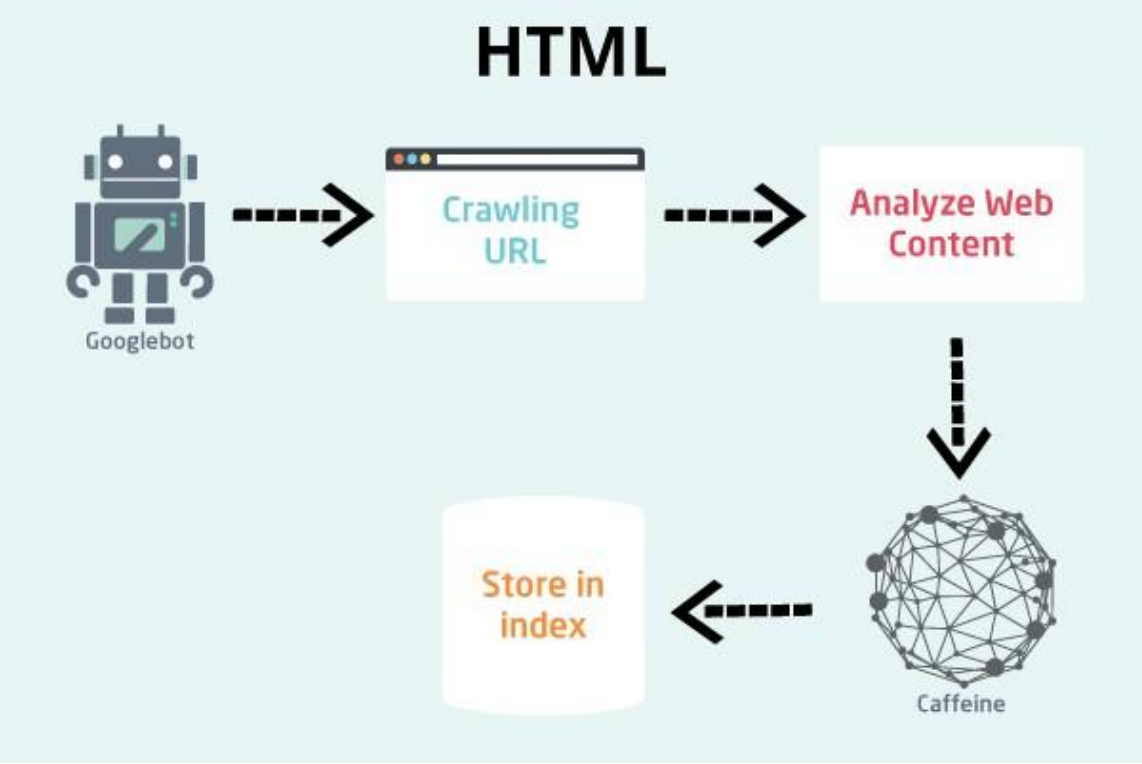
Google limits webmasters to 10 URL submissions per day.

The response time for indexing is the same for sitemaps as individual submissions.

Search engine crawlers use a number of algorithms and rules to determine how frequently a page should be re-crawled and how many pages on a site should be indexed. For example, a page which changes a regular basis may be crawled more frequently than one that is rarely modified.

Website Indexing (HTML)

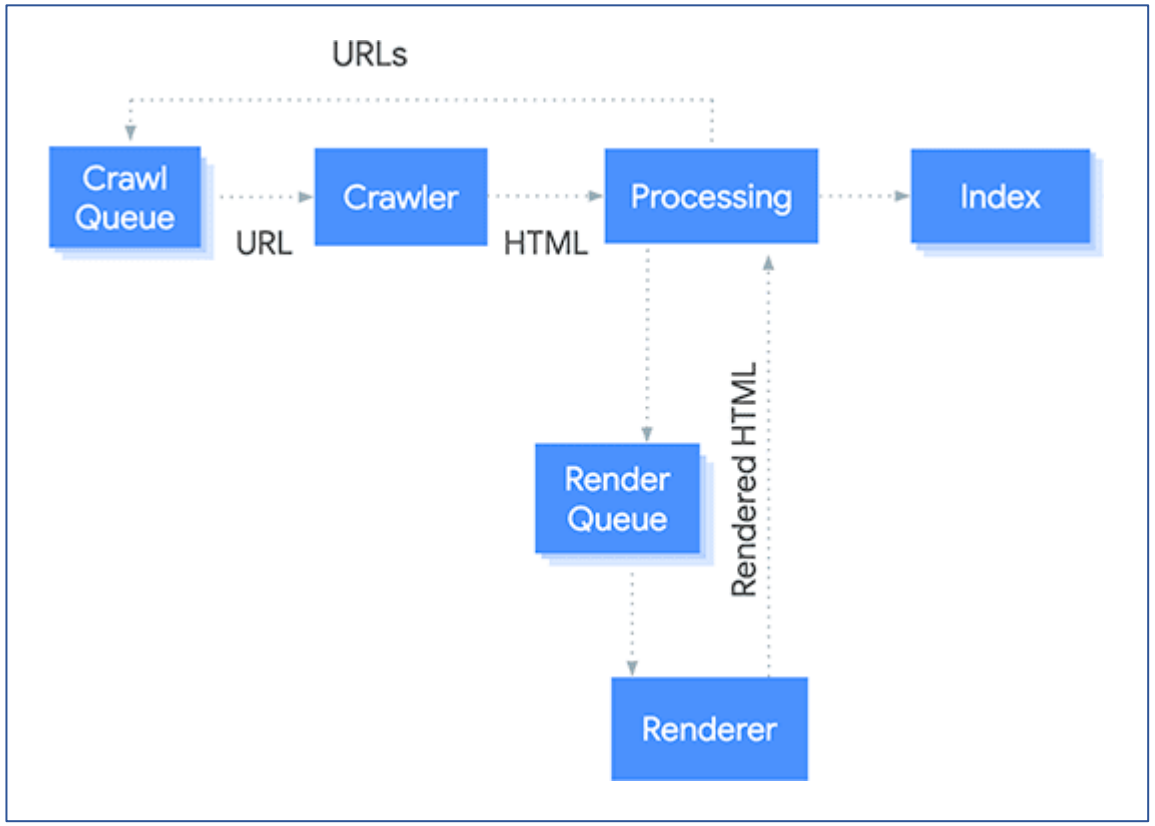
Googlebot: the web crawler software used by Google, which collects documents from the web to build a searchable index for the Google search engine.



Website Indexing (Javascript)

- **Crawler.** First, Googlebot gets the address for a page from the crawl queue and follows the URL. Assuming the page is not blocked via **robots.txt**, Googlebot will parse the page. (the “crawler” stage).
- At the crawler stage, any **new links (URLs)** that Googlebot discovers are sent back to the crawl queue. The HTML content on the parsed page may then be indexed.
- **Processing (rendering).** At this point, the URL will be processed for JavaScript.
- **Indexing.** This stage adds the content, be it from the HTML or additional content from JavaScript, to Google’s index. When someone enters a relevant query on Google, the page may appear.

In July 2019, Google published a new, brief about JavaScript SEO. The guide describes the stages or steps Google takes to crawl, render, and index content that JavaScript adds to a page.



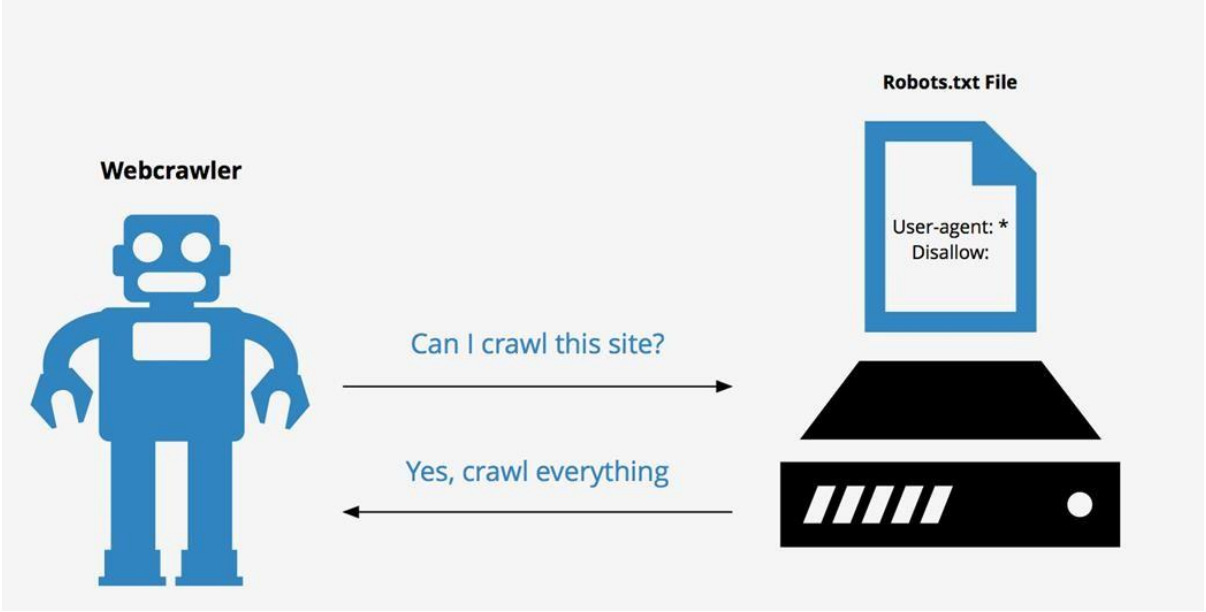
Indexing Failure.

There are several circumstances where **a URL will not be indexed** by a search engine:

- Robots.txt file exclusions – a file which tells search engines what they shouldn't visit on your site.
- Directives on the webpage telling search engines not to index that page (no-index tag) or to index another similar page (canonical tag).
- Search engine algorithms judging the page to be of low quality, have thin content or contain duplicate content.
- The URL returning an error page (e.g. a 404 Not Found HTTP response code).

Robot.txt

- A robots.txt file lives at the root of the site. robots.txt is a plain text file that follows the Robots Exclusion Standard. The file consists of one or more rules. Each rule blocks (or allows) access for a given crawler to a specified file path in that website.



Robots.txt

Apple example

- https://www.apple.com/robots.txt

Where is their sitemap?

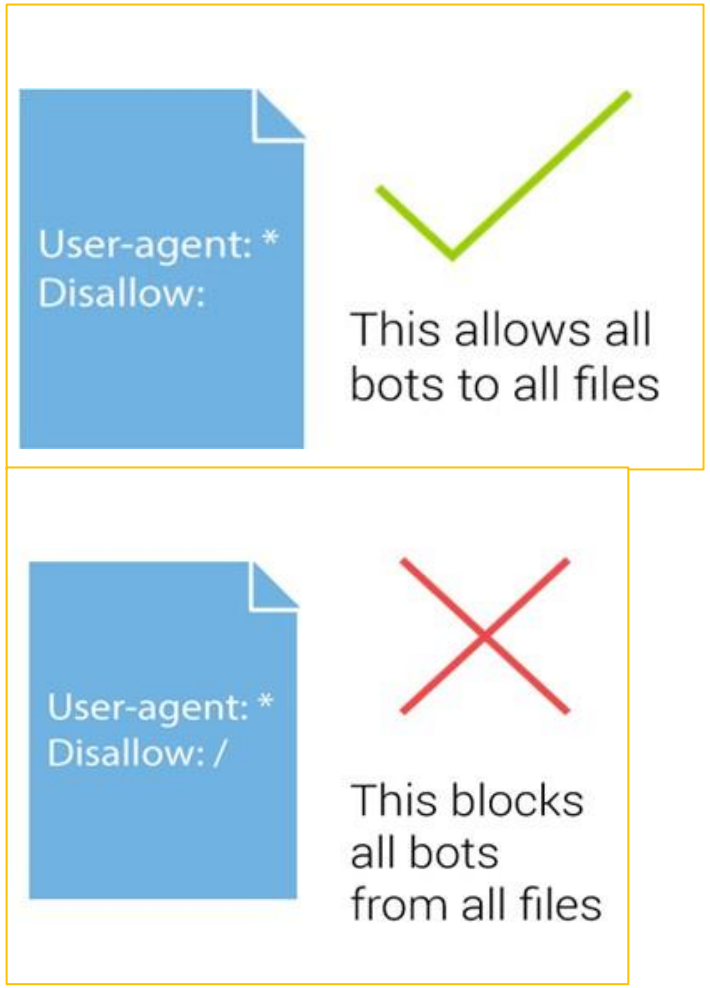
IHU example

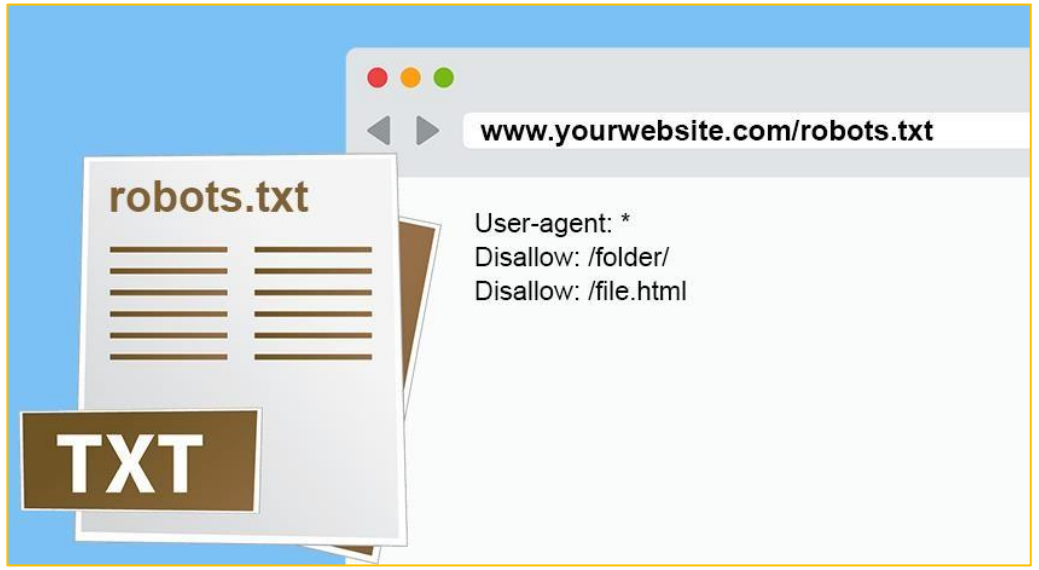
https://www.ihu.edu.gr/robots.txt

UOM example

https://www.uom.gr/robots.txt

Robot.txt


User-agent: *
Disallow:
This allows all bots to all files
User-agent: *
Disallow: /
This blocks all bots from all files


www.yourwebsite.com/robots.txt
robots.txt
TXT
User-agent: *
Disallow: /folder/
Disallow: /file.html

Robot.txt

Example of a simple robots.txt file

with two rules:

 # 1 User-agent: Googlebot
Disallow: /nogooglebot/

 # 2 User-agent: *
Allow: /
Sitemap: http://www.example.com/sitemap.xml

Explanation:
 # 1 The user agent named "Googlebot" crawler should not crawl the folder http://example.com/nogooglebot/ or any subdirectories.
All other user agents can access the entire site. The site's sitemap file is located at http://www.example.com/sitemap.xml

Robot.txt

- This example tells all robots to stay out of a website:

  User-agent: *

  Disallow: /

- This example tells all robots not to enter three directories:

  User-agent: *

  Disallow: /cgi-bin/

  Disallow: /tmp/ Disallow: /junk/

- This example tells all robots to stay away from one specific file:

  User-agent: *

  Disallow: /directory/file.html

Examples of valid robots.txt URLs:

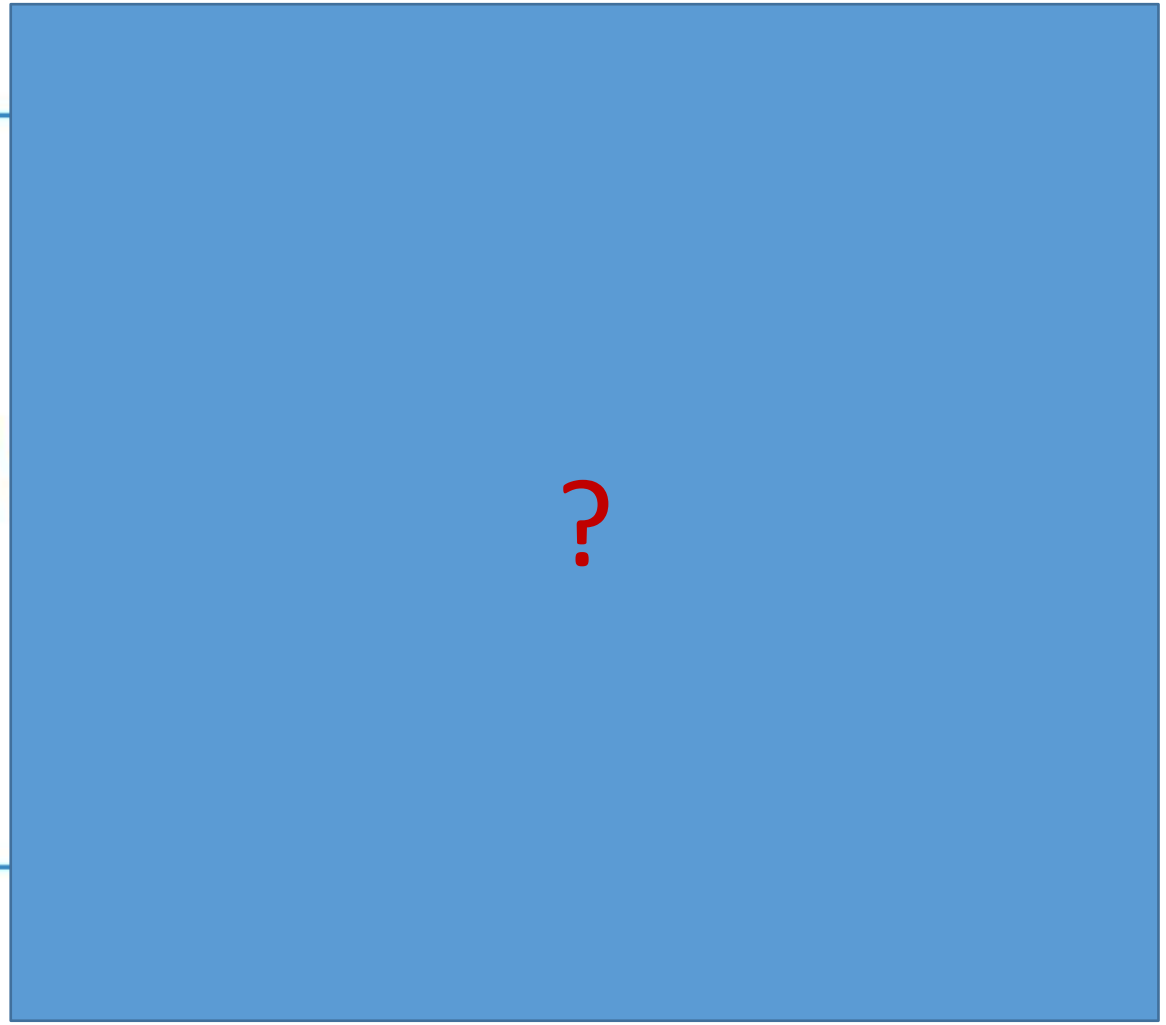
https://developers.google.com/search/reference/robots_txt#examples-of-valid-robots.txt-urls

Robot.txt

```
User-agent: Googlebot-News
Disallow: /admin
Disallow: /newfanshot
Disallow: /users/*/replies
Disallow: /search
Disallow: /account
Disallow: /login
Disallow: /chorus_auth
Disallow: /sso
Disallow: /scoreboard/ajax_leagues_and_events
Disallow: /networks/enter_private_mode_password
Disallow: /ad
Disallow: /sponsored
```

```
User-agent: *
Disallow: /admin
Disallow: /newfanshot
Disallow: /users/*/replies
Disallow: /search
Disallow: /account
Disallow: /login
Disallow: /chorus_auth
Disallow: /sso
Disallow: /scoreboard/ajax_leagues_and_events
Disallow: /networks/enter_private_mode_password
```

```
Sitemap: https://www.theverge.com/sitemaps
Sitemap: https://www.theverge.com/sitemaps/videos
Sitemap: https://www.theverge.com/sitemaps/google_news
```



Overall Checklist

Crawlability:

- Every page should have at least one internal link → easy to browse.
- Every page should be listed in an XML sitemap.
- Each URL should be accessible.
- Multiple versions of URLs shouldn't resolve to the same content.
- Implement 301 redirects from older to newer URLs.
- Ensure the site or parts are not blocked with robots.txt.
- Work on 404 (page not found) errors.

(404 SEO friendly examples: https://blog.hubspot.com/blog/tabid/6307/bid/6235/3-Tips-for-Building-Effective-Landing-Pages.aspx | https://moz.com/seo-competitive-analysis)

Crawl efficiency:

- Fast server response, fast page loading.
- Each page should contain substantial unique content.
- Changes pages URL with 301 redirects : https://support.google.com/webmasters/answer/93633

SEs Algorithms

SEs Algorithms

- The aim of the search engine algorithm is **to present a relevant set of high-quality search results** that will fulfill the user's query.

  The user then selects an option from the list of search results. This action, along with subsequent activity feeds into future learning which can affect search engine rankings going forward.

In addition to the search query, search engines use other relevant data to return results, including:

- **Location**;
- **Language detected**;
- **Previous search history;**
- **Device**.

Algorithm: PageRank (PR)

PageRank is a complex algorithm that assigns a score of importance to a page on the web.
PageRank is a linear representation of a logarithmic scale of between 0 and 10.
A PageRank score of 0 is typically a low-quality website, whereas, on the other hand, a score of 10 would represent only the most authoritative sites on the web.

https://checkpagerank.net/

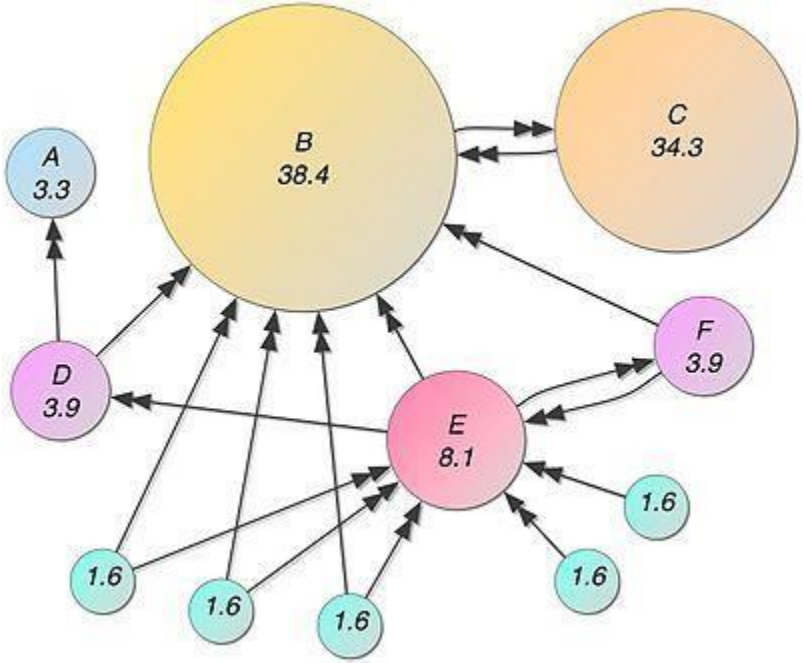
https://sitechecker.pro/

PageRank (PR)

PageRank (PR)

PR is " a link analysis algorithm and it assigns a numerical weighting to each element of a hyperlinked set of documents,

with the purpose of measuring its relative importance within the set." *Wikipedia*



PageRank was Google's first algorithmic calculation used to determine how a site should rank, based primarily on that **site's backlink profile**.

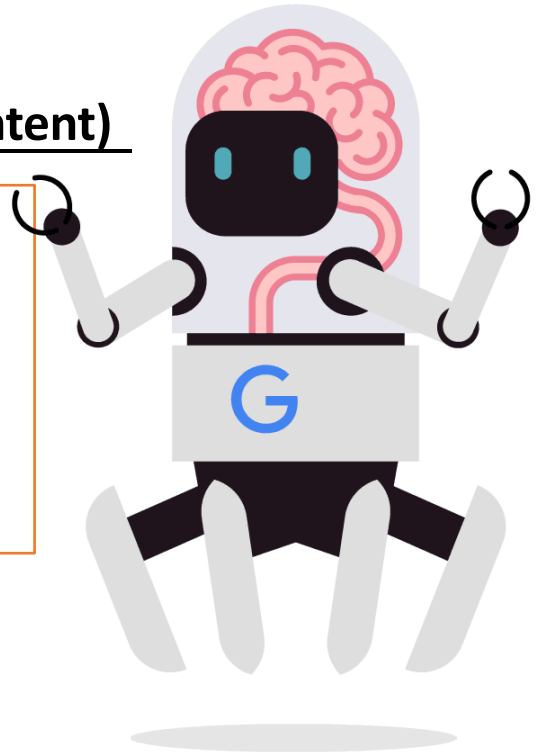
Today, there are over 200 ranking factors in Google's algorithm, many of which Google has not disclosed.

A page that is linked to by many pages with high PageRank receives a high rank itself.

Algorithm: RankBrain

Google: **The Top Three Ranking Factors Are Content, Links & RankBrain (searcher intent)**

RankBrain (2016) is a machine learning (AI) algorithm that Google uses to sort the search results by processing and understanding search queries.
Depending on the keyword, RankBrain will increase or decrease the importance of backlinks, content freshness, content length, domain authority etc.
Then, it looks at how Google searchers interact with the new search results. If users like the new algorithm better, it stays. If not, RankBrain rolls back the old algorithm.



It's paying very close attention to **how you interact with the search results**. Specifically, it's looking at UX signals:

- **Organic Click-Through-Rate**,
- **Dwell Time** (The amount of time that a Google searcher spends on a page from the search results before returning back to the SERPs)
- **Bounce rate**
- **Pogo sticking** (When a search engine users visits several different search results in order to find a result that satisfies their search query)

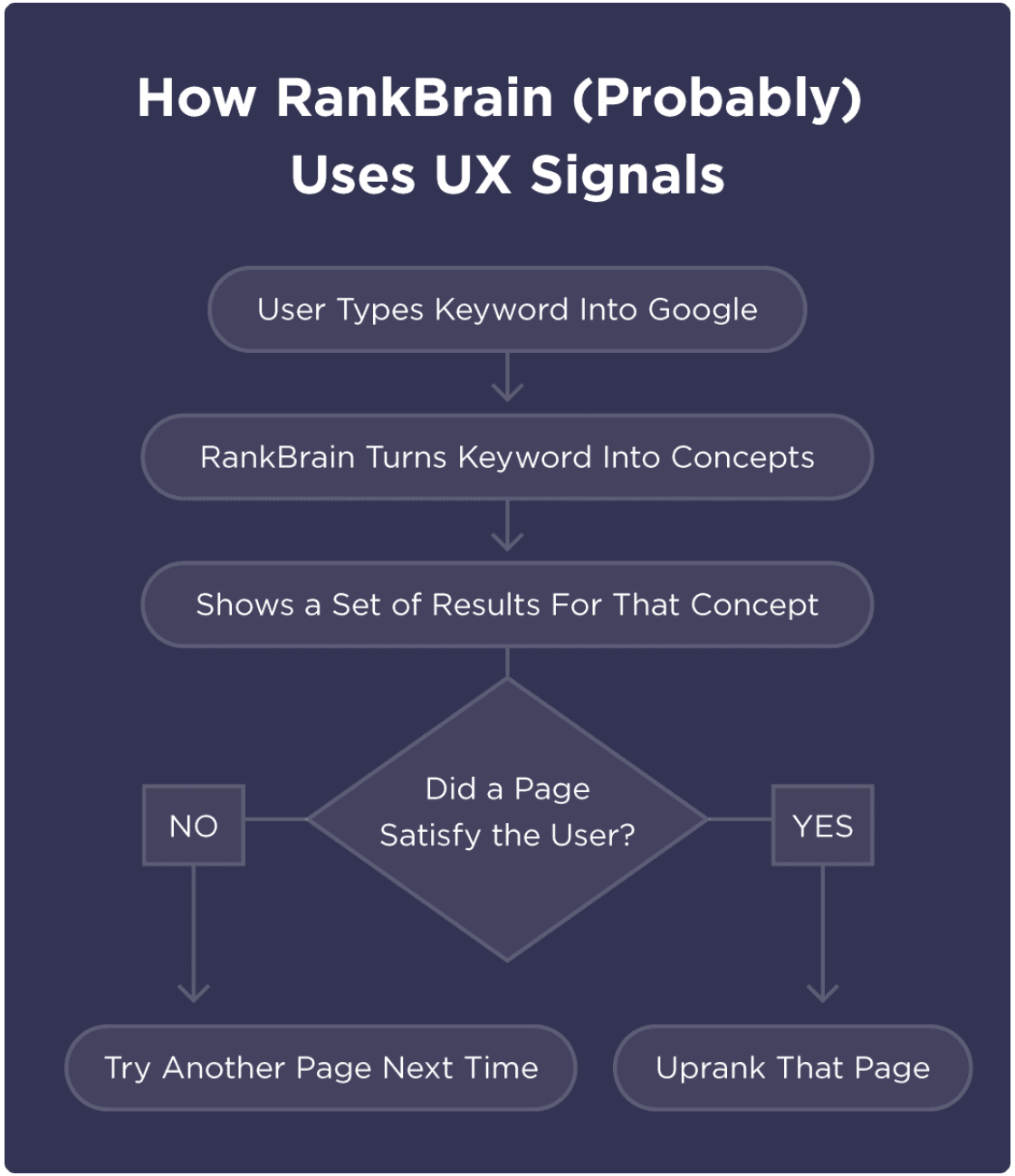
If people type something and then go and change their query, you could tell they aren't happy […] If they go to the next page of results, it's a sign they're not happy.'"

https://backlinko.com/google-rankbrain-seo

RankBrain

In the past, Google's algorithm would look for pages that exactly matched those keywords and deliver those as search results.

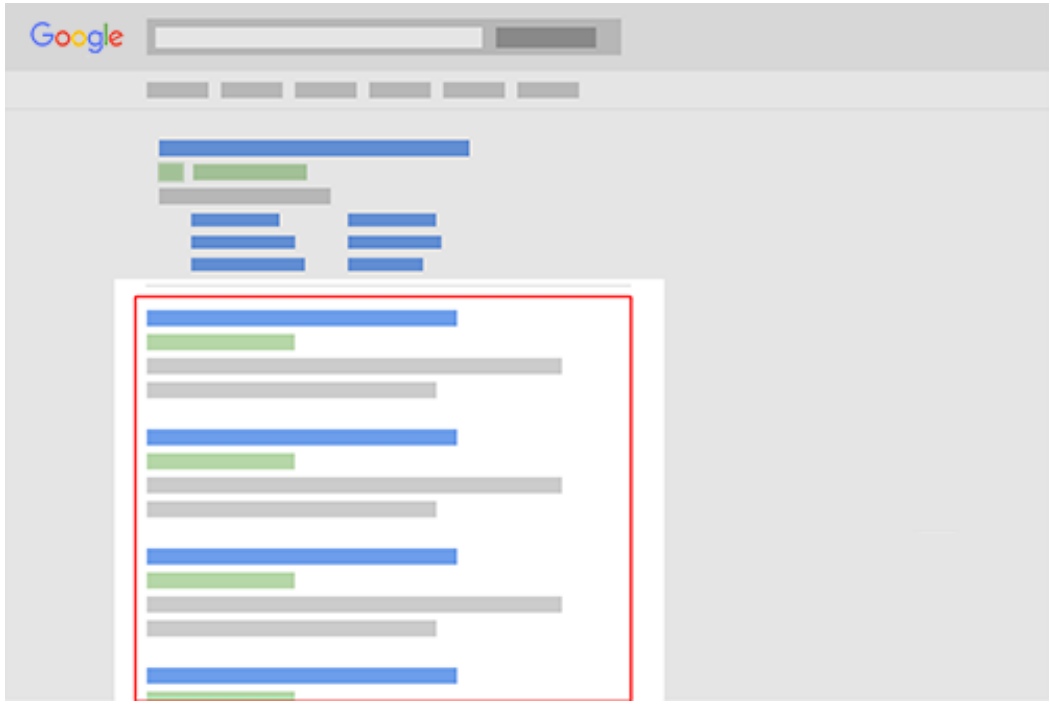
RankBrain is looking at all the content on a web page to try and figure out what it's about and whether it is a good match with what the searcher is looking for.



Organic Vs Paid Results

Organic Search Result

- A free listing in Google Search that appears because it's relevant to someone's search terms.
- Non-organic search results are paid advertisements.
- The ads above organic results contain an “Ad” box.
- The ads to the right of organic results have an “Ads” box above them.
- Analyzing organic search results can often help to identify new keywords for your Google Ads campaigns.



Examples?

Organic vs. Paid Search

- **Organic search** (**natural search**) is based on unpaid, natural rankings determined by search engine algorithms, and can be optimized with various SEO practices.
- In contrast, **paid search** allows you to pay to have your website displayed on the search engine results page when someone types in specific keywords or phrases. The fee you pay is usually based on either clicks or views of your ads.

A strong marketing strategy uses both search engine optimization and paid search (search engine marketing) to get found online.

But maybe organic is more essential.. Why??

Organic vs. Paid Search

- **More than 90% of users click on organic results**
- **Less than 10% click on paid results.**

Why??

Users trust search engines

SEO <> SEM

Optimization vs. Marketing

PR vs. Advertising

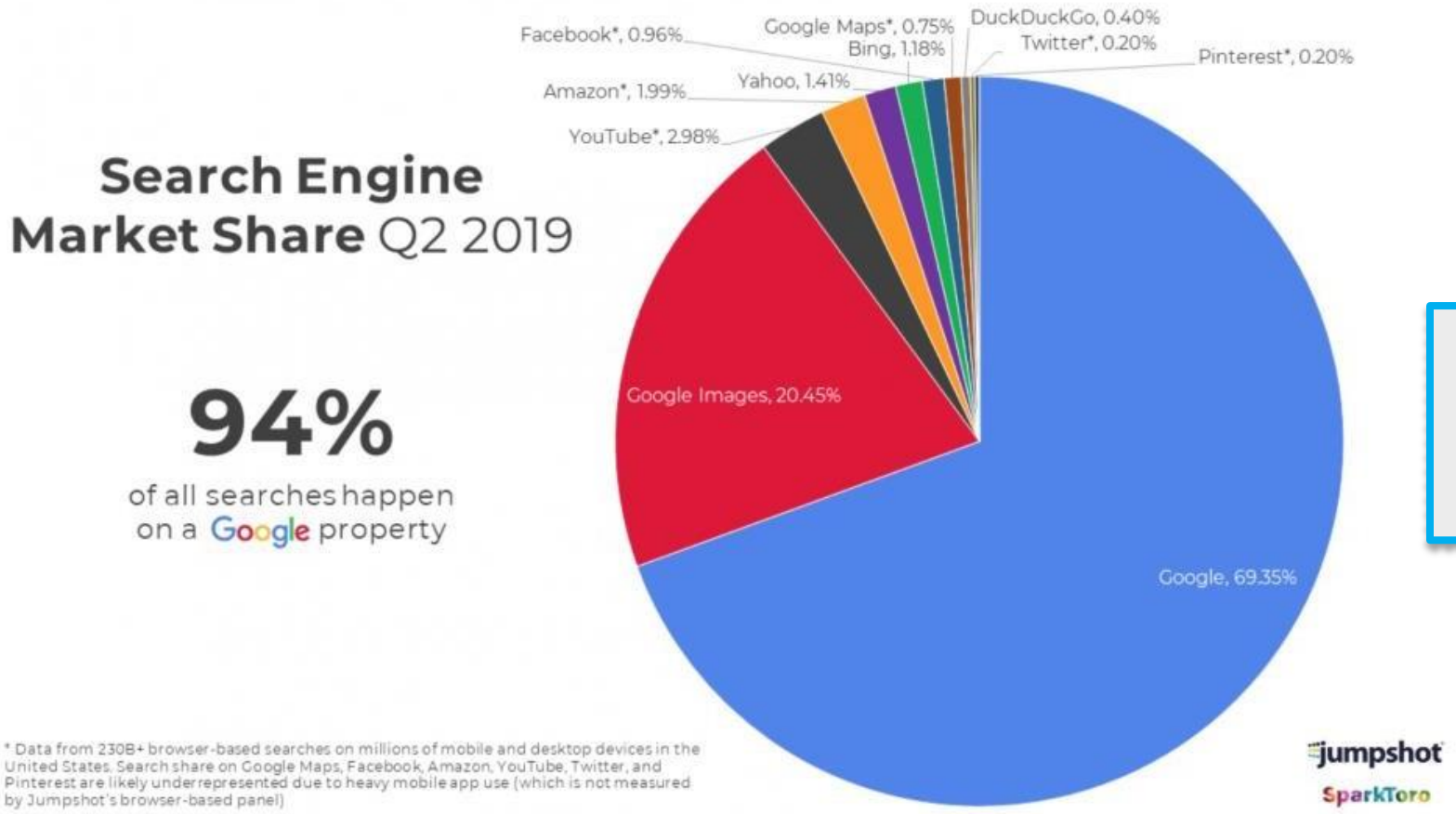
51% OF **WEBSITE TRAFFIC COMES** FROM ORGANIC **SEARCH** RESULTS
Chances are, more than half of your web **traffic** started with a **search.**
→ over 40% of revenue is captured through organic **search traffic**

Why Google?

- **Google presence is the most important:**

  In 2013, Google's services were unavailable for only 5 minutes due to an outage and web traffic dropped by 40%..
- There are more than 2.3 million Google searches conducted each minute.
- More than half of Google's searches are conducted across mobile platforms.
  - People are searching on the go, and are looking for local results.
  - 18% of local searches will lead to a sale on the same day (7 % of non local search).
  - 50% of mobile phone users will visit your store after conducting a search (34% of computers & tablets).

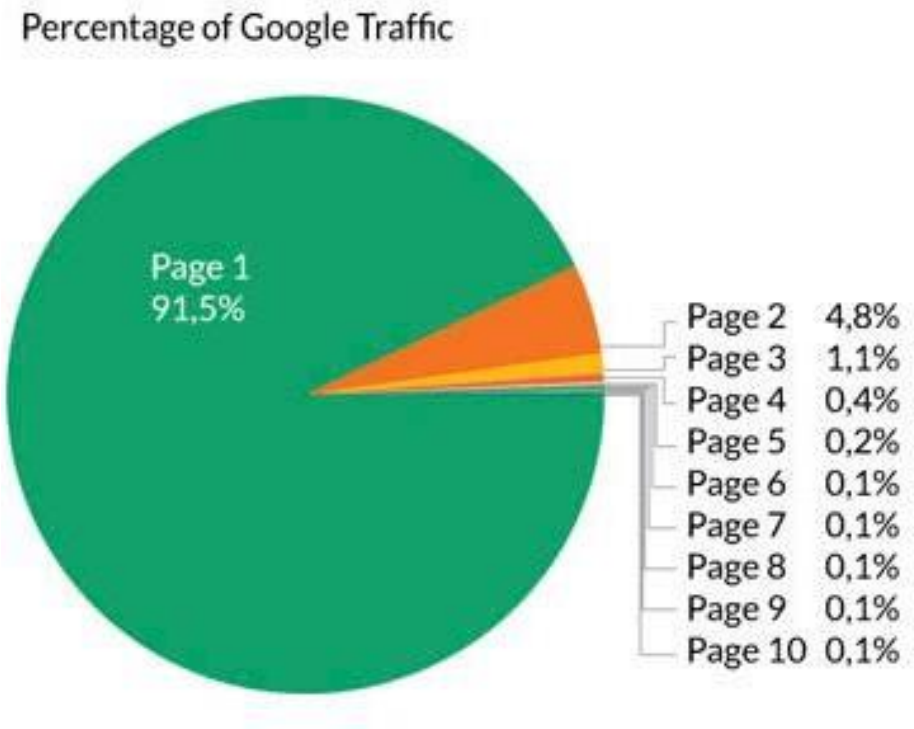
Why Google?



Also..
The purchase process begins on Google!

Google search behavior

1st Page vs. other pages



Google search behavior

The #1 Result In Google gets
31.7% of all clicks
& the highest CTR

A ratio showing how often people who see your link/ad end up clicking it.

$$CTR = \frac{Clicks}{Impressions} \times 100\%$$

CTR: Click-Through Rate

Clicks: The number of people who click on your link or ad.

Impressions: The number of people who view your link or ad.