



Search Engine
BASICS



● FREE PRINTABLE GUIDE | 2026 EDITION

Search Engine Basics PDF

How search engines discover, crawl, index and rank web pages, explained in plain words. Sixteen pages with infographics, a real case study, fresh market data, a cheat sheet and a classroom worksheet.

Infographics

Case study

Market data

Cheat sheet

Worksheet

Answer key

Written by **Usman Ali**

Growth and SEO specialist | 7 years in SEO

searchenginebasics.tech

Free to print and share for teaching, with credit

What You Will Learn

This guide follows a web page on its whole journey, from the moment it goes live to the moment it appears in front of a searcher. Read it in order the first time, then keep the cheat sheet nearby for quick reference.

01	What a Search Engine Does	3
02	The Five-Stage Pipeline	4
03	Crawling	5
04	Indexing	7
05	Ranking and the SERP	8
06	Search by the Numbers	10
07	Case Study: Structured Data	11
08	AI Search	12
09	Searching Better	13
10	Cheat Sheet	14
11	Worksheet and Answer Key	15

Search in Four Numbers

5 trillion+

searches a year on Google, as Google reported in 2025

15%

of daily Google searches are queries it has never seen before

89.94%

worldwide search share held by Google, StatCounter, Sept 2026

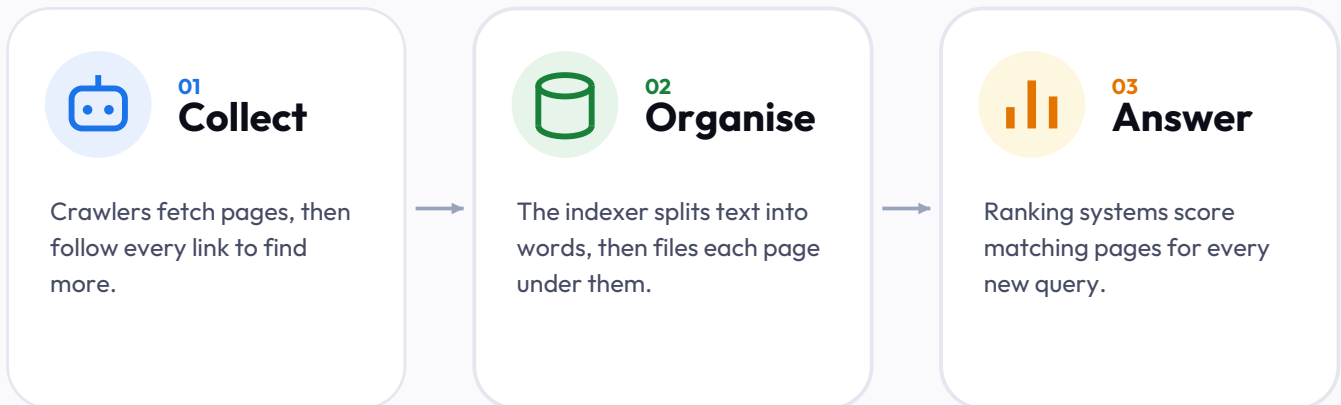
Hundreds

of factors decide relevance, Google says, including language and device

Figures from Google statements and StatCounter Global Stats. Full sources are listed on the last page.

What a Search Engine Does

A search engine is a program that finds pages on the web, stores what they say, then returns the most useful ones when someone types a query. Google, Bing and DuckDuckGo look different on the surface, yet each one is built around the same three jobs.



Why the Index Matters

When you press Enter, the engine does not visit the live web. It searches its own stored copy, called the index, which is why results arrive in a fraction of a second. Google describes its index as a very large database hosted on thousands of computers.

A Search Engine Is Not a Browser

People often mix the two up. A browser such as Chrome or Firefox opens pages when you already know the address. A search engine such as Google or Bing finds that address for you, then the browser displays the page you pick.

TOOL	WHAT IT DOES	EXAMPLES
Browser	Opens and displays a web page	Chrome, Firefox, Safari, Edge
Search engine	Finds pages that match a query	Google, Bing, DuckDuckGo, Baidu
AI answer engine	Writes an answer from retrieved pages	Google AI Mode, Perplexity



REMEMBER

Crawling is reading, indexing is storing, ranking is ordering. Most beginner confusion comes from mixing up these three words.



The Five-Stage Pipeline

Every page that shows up in search has passed through the same five stages. A problem at any single stage stops the journey, so knowing where a page got stuck is the fastest way to fix it.

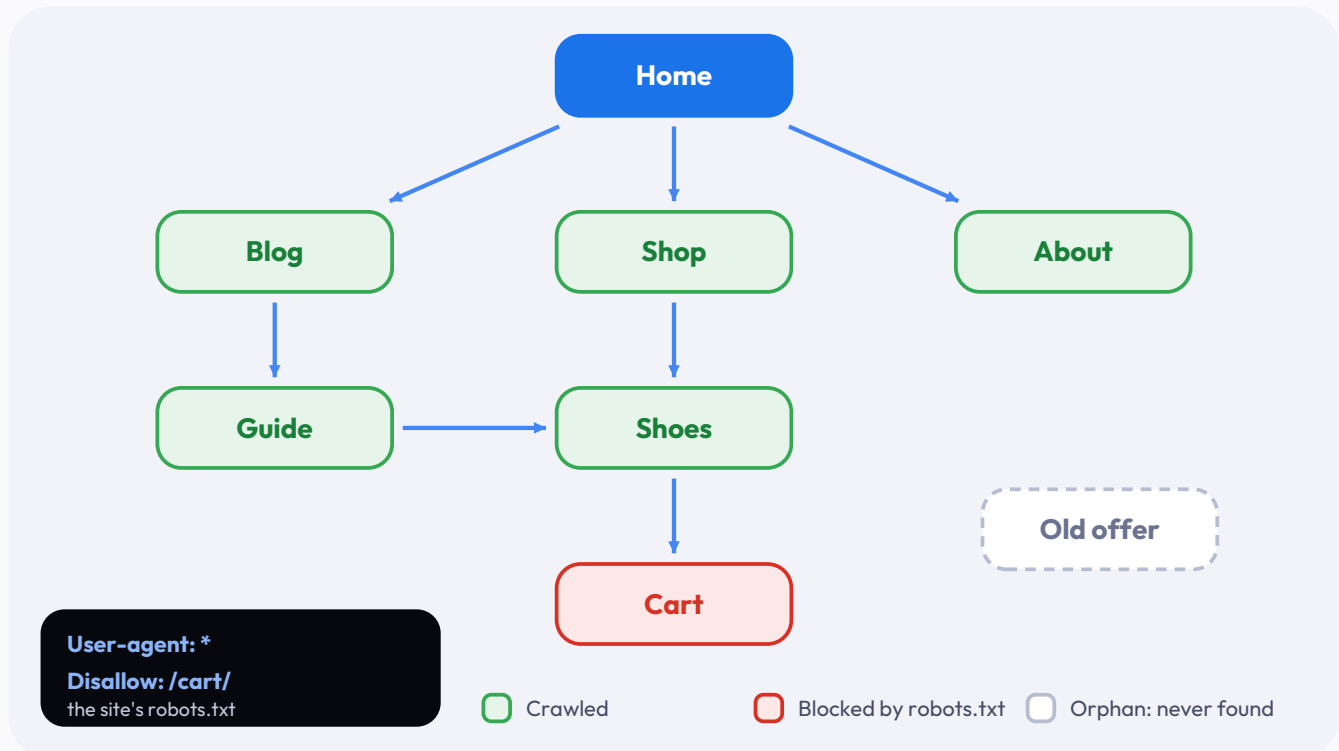


GOOD TO KNOW

Google says plainly that it does not guarantee to crawl, index or serve any page, even one that follows every guideline. Passing each stage makes a result possible, not certain.

Crawling: How Pages Get Found

A web crawler, also called a spider or bot, is software that requests pages, saves a copy, then adds every new link it finds to a queue. Googlebot and Bingbot are the best known examples.



A crawler starting at Home reaches every linked page. It skips Cart because robots.txt blocks it, while Old offer stays hidden because no page links to it.

Three Ways a Crawler Finds a New URL

ROUTE	HOW IT WORKS	BEST FOR
Links	The bot follows links from pages it already knows.	Every site; internal links matter most
XML sitemap	A file lists the URLs you want crawled.	Large or new sites
IndexNow	Your site pings engines when a URL changes.	Bing and other supporting engines

Mobile First by Default

Google now crawls most sites with Googlebot Smartphone, which sees each page the way a phone does. If key text or links only appear on the desktop version, Google may never see them.

What Stops a Crawler

Most crawling problems come down to a handful of causes. Each one has a simple fix once you know where to look.

PROBLEM	WHAT THE CRAWLER SEES	FIX
Orphan page	No link points to it, so it is never found.	Link to it from a related page.
robots.txt block	A rule tells the bot to stay out.	Remove the Disallow line for that path.
Server error (5xx)	The site fails to respond.	Fix hosting or server limits.
Missing page (404)	The address no longer exists.	Restore it or add a redirect.
Login wall	Content sits behind a password.	Keep public content outside the login.
Endless filters	Thousands of near duplicate URLs.	Limit crawlable filter combinations.

Reading a robots.txt File

Robots.txt sits at the root of a domain, for example `example.com/robots.txt`. Each group names a crawler, then lists the paths it should avoid.

```
# Let every crawler in, except one folder
```

```
User-agent: *
```

```
Disallow: /members/
```

```
# Keep one AI training bot out completely
```

```
User-agent: GPTBot
```

```
Disallow: /
```

```
Sitemap: https://example.com/sitemap.xml
```

Crawl Budget in One Paragraph

Crawl budget is the number of URLs an engine is willing to fetch from one site in a given period. Small sites rarely need to worry about it. Large sites with thousands of filter pages can waste it, which delays the crawling of pages that matter.



TRY THIS

Open `yoursite.com/robots.txt` in a browser. Read each Disallow line, then ask whether you really want search engines kept out of that folder.

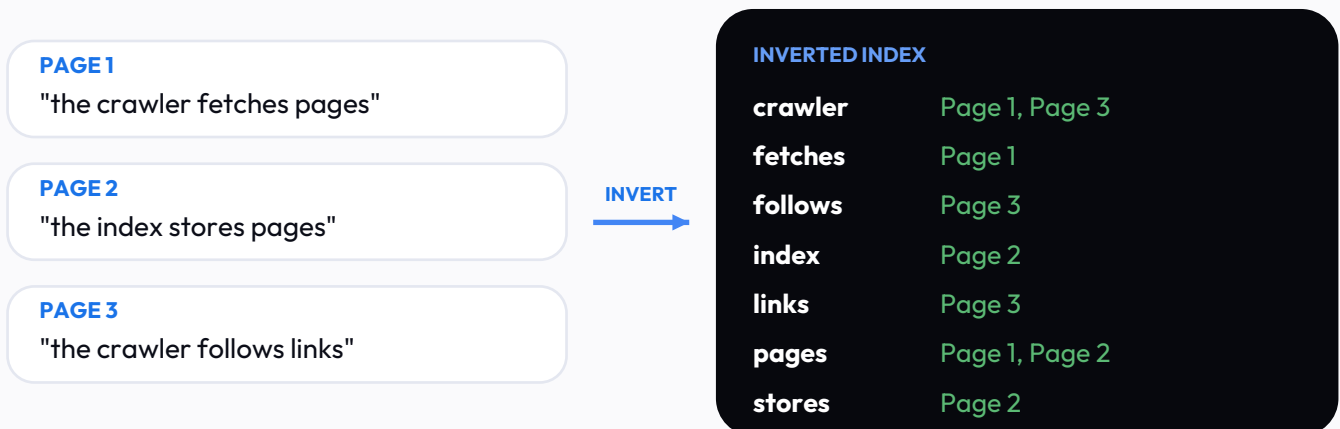
Indexing: How Pages Get Stored

After a page is crawled, the engine decides whether to keep it. If it does, the text is cleaned up, broken into words, then filed in an inverted index that works like the index at the back of a book.

Step 1: Turning Text Into Terms



Step 2: Building the Inverted Index



A search for "crawler links" only has to compare two short lists. Page 3 appears on both, so it becomes the match.

Why a Page Is Crawled but Not Indexed

REASON	WHAT IT MEANS
noindex tag	The page asks to be left out of results.
Duplicate content	Another URL holds the same text, so the engine keeps one canonical version.
Thin content	The page adds little that other pages do not already cover.
Soft 404	The page loads, yet it looks empty or broken.

How Ranking Works

Ranking decides which indexed pages appear for a query, in what order. Google says it weighs hundreds of factors, although six broad groups explain most of what you see on the page.

1 Relevance

Does it answer the query?

2 Quality

Is it accurate and original?

3 Links

Do trusted sites point here?

4 Freshness

Is it current where that matters?

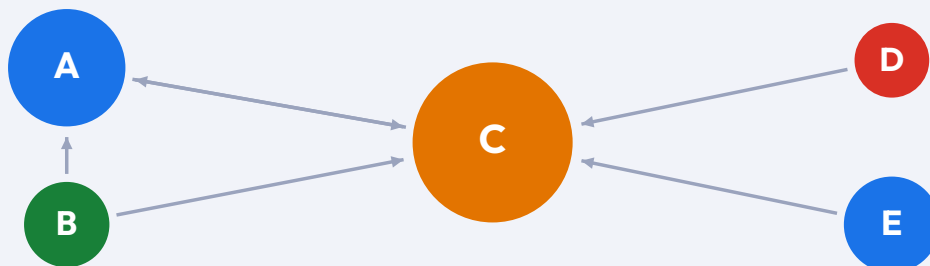
5 Context

Location, language, device

6 Experience

Fast, mobile friendly, no clutter

Links as Votes: The PageRank Idea



C receives links from four pages, so it gathers the most weight. A link from a strong page such as A passes more value than one from a weak page.

Larry Page and Sergey Brin described PageRank in 1998 while studying at Stanford. Modern ranking uses far more than links, yet the idea that a link works like a vote of trust still shapes how search engines judge pages.

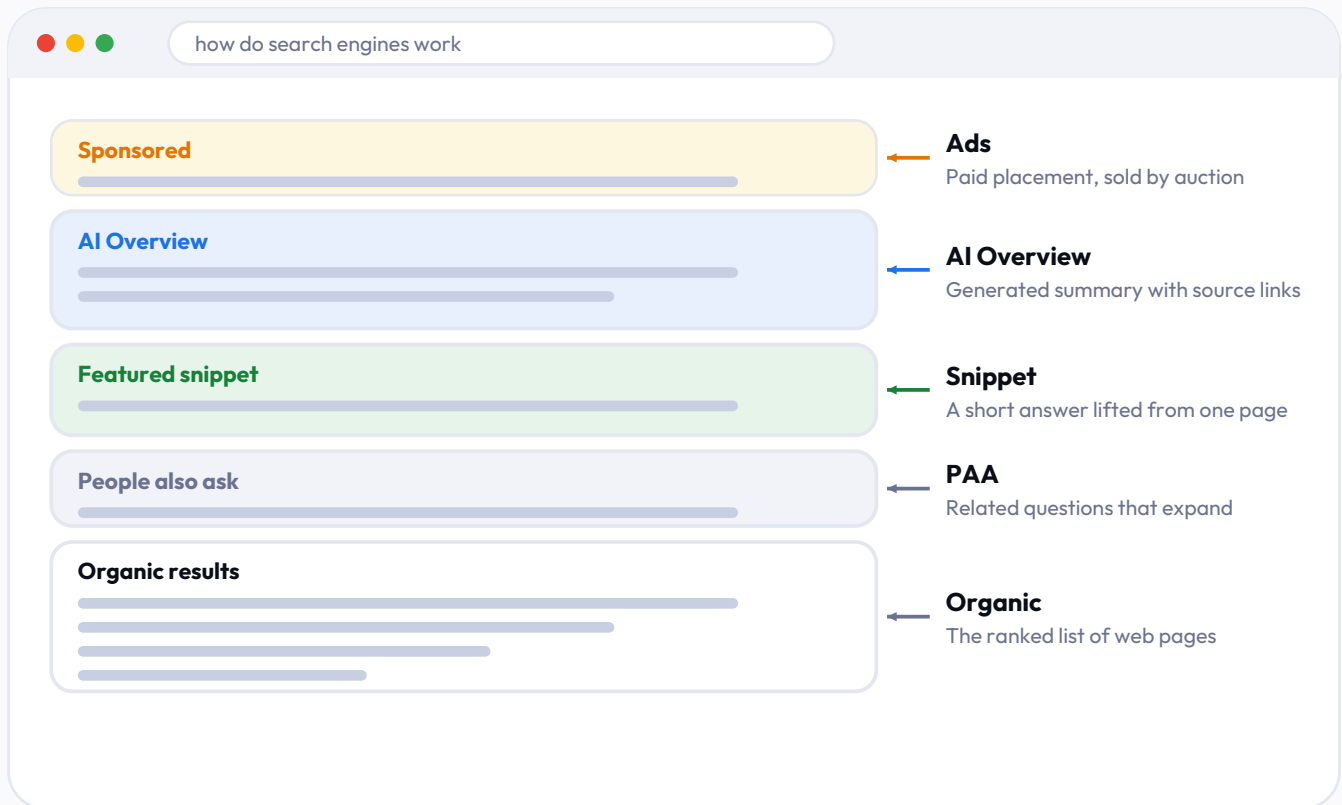


GOOD TO KNOW

Google states that it does not accept payment to crawl a site more often or rank it higher. Money only buys space in clearly labelled ad slots.

Anatomy of a Results Page

SERP stands for search engine results page. Beyond the classic blue links, a modern SERP stacks several blocks, each built for a different kind of question. These extra blocks are called SERP features.



BLOCK	PAID OR ORGANIC?	WHEN IT APPEARS
Ads	Paid	Commercial queries with advertisers bidding
AI Overview	Organic	Questions where a summary helps
Featured snippet	Organic	Clear questions with a short answer
Local pack	Organic	Searches with local intent, such as "near me"
Organic results	Organic	Almost every query

Search by the Numbers

A quick look at who people actually use to search, plus a few figures that show the scale of the job. Market share comes from StatCounter, which measures page views across millions of websites.

Worldwide Search Engine Market Share



Source: StatCounter Global Stats, worldwide search engine market share, September 2026.

What the Numbers Tell Us

Google handles roughly nine in every ten searches worldwide, which is why most guides focus on how Google works. Bing still matters, because its index also feeds partner engines and several AI assistants.

Scale explains why the index exists at all. Google reported more than 5 trillion searches a year in 2025, while it has long said that around 15% of daily queries are brand new. No engine could answer that load by visiting the live web each time.



REGIONAL PICTURE

Shares change a lot by country. Baidu leads in China, Naver is strong in South Korea, and Yandex remains popular in Russia.

Structured Data in Action

Structured data is code that labels what a page contains, such as a recipe, a product or a review. It helps search engines understand the page, which can unlock richer results with stars, images or prices.

Results Reported by Google

25% higher click-through rate

Rotten Tomatoes added structured data to 100,000 pages.

82% higher click-through rate

Nestlé pages shown as rich results versus plain results.

35% more visits

Food Network enabled search features on 80% of its pages.

1.5x more time on page

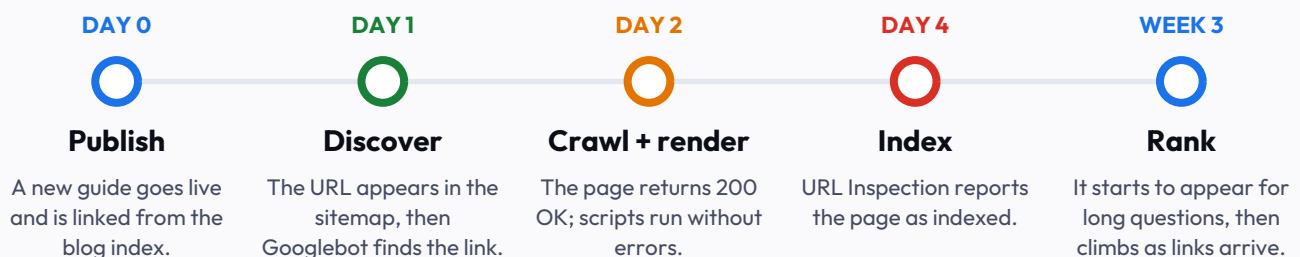
Rakuten users on pages with structured data.

Source: Google Search Central, "Introduction to structured data markup in Google Search".

What Beginners Can Take From This

None of these brands changed their writing to win more clicks. They described their existing pages more clearly to search engines, so their listings stood out on a crowded results page. That is a reminder that indexing quality affects ranking outcomes.

Worked Example: One Page From Publish to Result



Illustrative timeline for a healthy, well-linked page on an established site. New sites often take longer at each stage.

How AI Search Works

AI answers have not replaced the pipeline. Google AI Overviews and similar tools still rely on crawled, indexed pages, then use a language model to summarise the best sources.



What Changed and What Did Not

STAGE	CLASSIC SEARCH	AI SEARCH
Crawling	Bots fetch pages	Same bots, plus some AI crawlers
Indexing	Pages stored in the index	Same index is the main source
Answer	A ranked list of links	A written summary with citations
Clicks	Most go to organic results	Some questions end without a click

The technique behind this is called retrieval augmented generation. The model first retrieves trusted pages, then grounds its answer in them, which reduces made-up claims while giving readers links to check.



FOR SITE OWNERS

The same basics still win. Pages that are easy to crawl, clearly written and well sourced are the ones AI answers are most likely to cite.

Search Better With Operators

Search operators are short commands that make a query more precise. They work in Google and Bing, and they save a surprising amount of time on research tasks.

OPERATOR	WHAT IT DOES	EXAMPLE
"quotes"	Matches the exact phrase.	"search engine basics"
site:	Searches one website or domain type.	site:gov climate report
-word	Leaves a word out of results.	jaguar -car
filetype:	Finds one kind of file.	crawl budget filetype:pdf
OR	Accepts either word.	crawler OR spider
intitle:	Requires a word in the page title.	intitle:glossary seo

A Five-Point Check for Any Source

- 1 Who wrote it?**
 Look for a named author with real experience.
- 2 When was it updated?**
 Old pages may describe old systems.
- 3 What is the evidence?**
 Good pages link to primary sources.
- 4 Who benefits?**
 Notice when a page is selling something.
- 5 Do others agree?**
 Check one or two independent sources.

One-Page Cheat Sheet

Twenty-four terms that come up again and again. Print this page, then keep it next to you while you study.

Algorithm	A set of rules a computer follows to reach a result.	noindex	A tag asking engines to leave a page out of results.
Backlink	A link from another website to yours.	Organic result	An unpaid listing chosen by the ranking system.
Bot	Software that visits pages automatically, also called a crawler.	PageRank	Google's original link-based scoring method from 1998.
Canonical URL	The preferred version of a page when duplicates exist.	Query	The words a person types or speaks into a search engine.
Crawl budget	How many URLs an engine will fetch from a site in a period.	Ranking	Ordering matching pages from most to least useful.
Crawling	Downloading pages so the engine can read them.	Relevance	How closely a page matches what the searcher wants.
Googlebot	Google's main crawler, mostly using a smartphone view.	Rendering	Running a page's code to see its final content.
Index	The engine's stored copy of the pages it has processed.	robots.txt	A file that tells crawlers which paths to avoid.
Indexing	Processing a page, then storing it so it can be found.	SERP	The search engine results page.
IndexNow	A protocol that lets sites report new URLs instantly.	SERP feature	Any block on a SERP beyond the plain blue links.
Inverted index	A table listing each word with the pages that use it.	Structured data	Code that labels what a page contains.
Keyword	A word or phrase people type into a search box.	XML sitemap	A file listing the URLs a site wants crawled.



WANT MORE TERMS?

The full search engine glossary at searchenginebasics.tech/resources/glossary/ explains more than 120 words in the same plain style.

Classroom Worksheet

Answer each question in the space on the right. Teachers can run this as a ten-minute check after a lesson on how search engines work.

01 What is the difference between crawling and indexing?

02 Name two ways a search engine can discover a new page.

03 What does a Disallow line in robots.txt do?

04 Give two reasons a page might be crawled but not indexed.

05 In the inverted index on page 7, which page matches "index pages"?

06 Name three signal groups used in ranking.

07 What does SERP stand for?

08 Can a business pay Google for a higher organic ranking?

09 Write a search that finds PDF files about crawl budget.

10 Which search engine had the second largest share in Sept 2026?

Class Activity: Be the Crawler

Write six page names on cards, then draw arrows between some of them to show links. One student plays the crawler, starting from the home card and following arrows only. Any card the crawler never reaches is an orphan page, which shows why internal links matter.

Answer Key

1. Crawling downloads a page, while indexing stores it for search. 2. Links from known pages, an XML sitemap or IndexNow. 3. It asks crawlers to stay out of a path. 4. noindex, duplicate content, thin content or a soft 404. 5. Page 2. 6. Relevance, quality, links, freshness, context or experience. 7. Search engine results page. 8. No. Money only buys labelled ad slots. 9. crawl budget filetype:pdf 10. Bing, at 5.29%.

Where to Go Next

This PDF is a starting point. Every topic inside has a full, free guide with interactive tools on the website.

● Start here

searchenginebasics.tech/start-here/

A reading path through every guide

● How search engines work

[/how-search-engines-work/](https://searchenginebasics.tech/how-search-engines-work/)

The five stages in depth

● Pipeline simulator

[/tools/search-pipeline-simulator/](https://searchenginebasics.tech/tools/search-pipeline-simulator/)

Break a page on purpose and watch it fail

● Glossary

[/resources/glossary/](https://searchenginebasics.tech/resources/glossary/)

120+ search terms in plain words

SOURCES

- Google Search Central: In-depth guide to how Google Search works
- Google Search Central: Introduction to structured data markup
- Google Search Central: Overview of Google crawlers and fetchers
- StatCounter Global Stats: Search engine market share, September 2026
- Google: statement on more than 5 trillion searches a year (2025)
- Google: statement that 15% of daily searches are new
- Brin and Page (1998): The Anatomy of a Large-Scale Hypertextual Web Search Engine
- IndexNow protocol documentation

About the author

I am Usman Ali, a growth and SEO specialist with seven years of hands-on work. I wrote this guide so that beginners, students and teachers can understand search without the jargon. Spotted an error? Email searchenginebasicstech@gmail.com.