Learning Aims

- Search engine indexing
- PageRank algorithm

Keyword	Definition
Indexing	The process where search engines scan web pages, analyse their content, and store data in an index so it can be retrieved quickly during a search.
Search Engine	A program (like Google) that searches an index of web pages to display relevant results based on a user's query.
Web Crawlers / Spiders	Automated programs that browse the internet and collect information about web pages for indexing.
PageRank	Google's algorithm that ranks web pages based on the number and quality of links pointing to them. Each link acts as a "vote" for the page.
PageRank Algorithm (High Level)	Assigns a numerical score to each web page depending on incoming links and their importance; the higher the score, the higher the ranking in search results.
Metadata	Means "data about data." It gives **information about a webpage**, not the main content itself. In web pages, metadata is found inside the **HTML <head> section**. It can include details such as: The **page title** (shown on the browser tab) A **description** of the page (used by search engines) **Keywords** that describe the content The **author's name** The **date** the page was last updated

How Web Crawlers Work

- Web pages are found by **search engines** using special programs called **crawlers** (also known as spiders or bots).
- Crawlers **follow links** from one page to another across the web.
- They start from a few known pages (called **seed URLs**) and keep visiting other pages linked from them.
- Crawlers follow the rules in a website's **robots.txt file**, which tells them which pages they can or cannot visit.
- When a crawler visits a page, it **downloads the HTML** (the code of the page).
- It looks at the **text, headings, links, and metadata** on the page.
- The HTML is then **broken into smaller parts** (tags and attributes) so the crawler can understand the structure and find key information like titles and headings.

Search engine indexing

- Search engine indexing (or web indexing) means examining web pages to see what information they contain.

Benefits:

- Improved search results – Users find accurate, recent, and relevant pages faster. Crawlers revisit pages to check for updates, keeping results current.

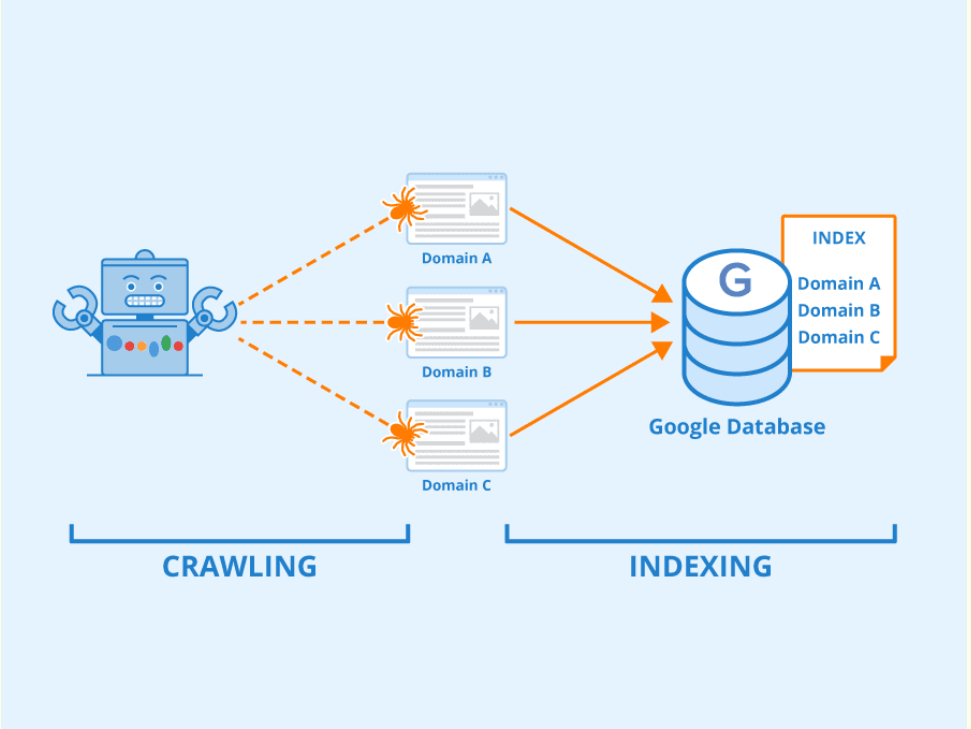
- Efficient retrieval – Search engines use their index to produce results instantly.

- Ranking & relevance – Pages are ranked using algorithms that consider keywords, links, and user engagement.

Drawback

- Indexing uses extra storage space and must be updated regularly.

- Updating takes extra processing time, but the benefits — fast and accurate search results — outweigh the drawbacks.



How indexing works

1. Web crawlers (spiders) from search engines like Google or Yahoo explore the internet to find web pages.

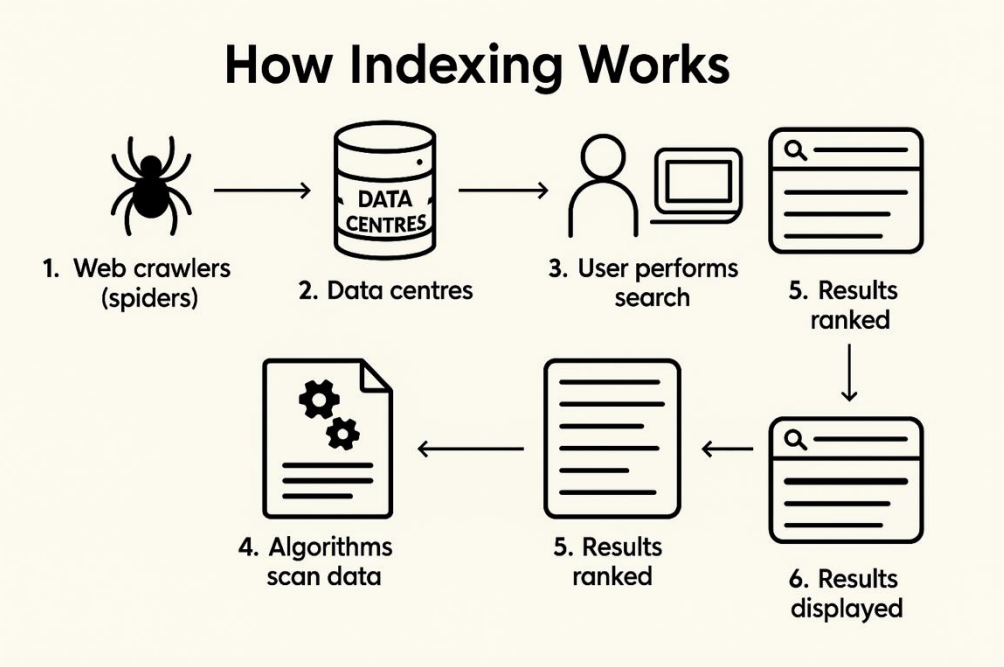
2. The crawlers collect and store information from these pages in **large data centres** worldwide.

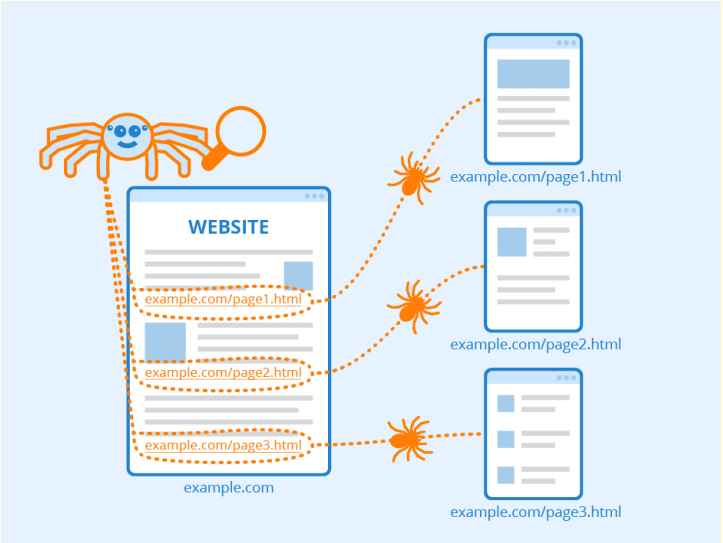
3. When a user performs a search, algorithms scan these data centres to find the most relevant matches.

4. The algorithms consider many factors — such as **how recent a page is, keywords, metadata, language, links, and recommendations.**

5. Results are ranked and displayed, with the most relevant ones shown at the top — all within a fraction of a second.

6. Companies invest heavily in SEO (Search Engine Optimization) to appear higher in these search results.





How to improve search indexing

- Make your website informative, relevant, and regularly updated.
- Add metadata (Title, Description, and Keywords) in your HTML pages.
- Submit your website to search engines so they can index it (e.g., Google's submission tool).
- Get your site linked from other websites, as this improves credibility and ranking.
- Increase links by:
  - Adding your site to online directories.
  - Making it easy for others to link to you (e.g., share buttons).
  - Being active on social media and sharing your links.
  - Exchanging links with other website owners.

For example, you can submit your website to Google here:
https://www.google.com/webmasters/tools/submit-url?pli=1

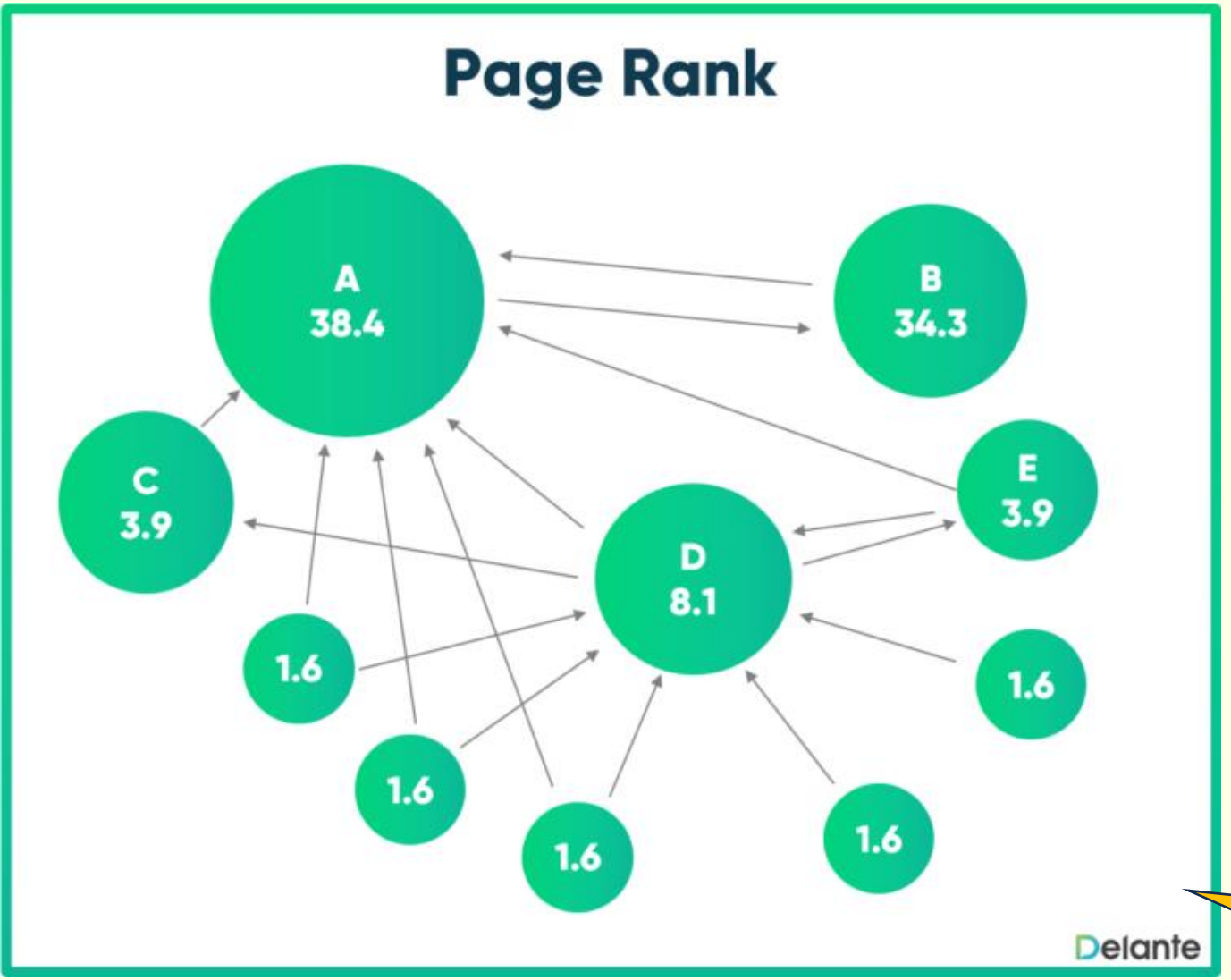
PageRank algorithm

PageRank is one of a number of algorithms (Maths formulas) that measures how **relevant and useful a particular website page is.**

It is used by Google Search to organise the results of a search into a ranked order of importance and relevance.

How does PageRank work?

- PageRank is based on the idea that the more links a page has, the more important it is.
- Links from other sites act as “votes” showing the page is useful and relevant.
- Each page is given a **numerical ranking** based on the number and quality of its links.
- Links from pages that are themselves well-linked count more than links from less connected pages.
- This means a link from a popular, trusted site boosts ranking more than one from an unknown site.
- Overall, PageRank helps determine which web pages appear higher in search results.



- Each circle represents a web page (A, B, C, D, E, etc.).
- The size of the circle and the number value inside it represent that page's PageRank score — the higher the score, the more "important" the page is.
- The arrows represent links between pages — a page pointing to another is "voting" for it.

!

Example of using a Graph Data Structure

The dampening factor

- If a page has **too many links**, PageRank lowers the value of its votes — this is called the "dampening factor."

- Other things that affect ranking include:

  - How often the page is updated

  - The age and quality of links

  - How trustworthy or well-known the domain name is

- PageRank is always being improved and updated.

- Website owners can raise their ranking by:

  - Exchanging links with genuine, relevant sites

  - Most importantly, creating high-quality content that people naturally visit and link to

PageRank Algorithm

- The simplified model of the algorithm looks as follows:

$$PR_x = \frac{1-d}{N} + d\left(\frac{PR_y}{L_y} + \frac{PR_z}{L_z}\cdots\right)$$

Delante

- PR – Page Rank of a given page,
- d – damping factor, usually around 0.85,
- N – the number of web pages,
- L – the number of backlinks pointing to a given web page.

! You don't need to know this for the exam

Problems with PageRank

- Once people understand how search algorithms work, they can try to manipulate them.
- **Google Bombs** happen when groups artificially boost a page's ranking by linking specific words or phrases to it.
- Google has updated its algorithms to reduce this, but it still occurs.
- **Link farming** is another trick — creating large numbers of irrelevant links between web pages using software.
- Link farms are pages filled mostly with links and little useful content.
- Google continues to fight against link farming, but it remains an ongoing challenge.