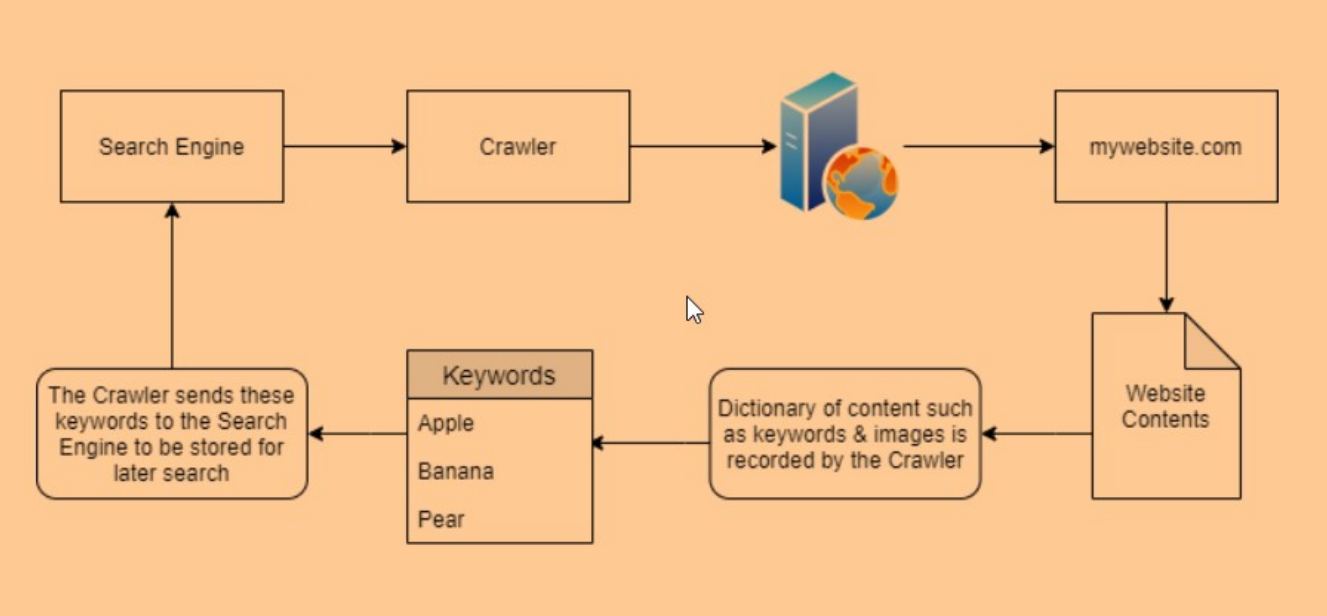
Google Dorking

What are Crawlers
These crawlers discover content through various means .
A crawler, also known as a web spider or web robot, is an automated program used by search engines to systematically browse and index web pages on the internet. Crawlers are not part of a web browser.



What is the name of the technique that "Search Engines" use to retrieve this information about websites?
ANS : Crawling

Search Engine Optimisation
There are many factors in how “optimal” a domain is - resulting in something similar to a point-scoring system. To highlight a few influences on how these points are scored, factors such as:

- How responsive your website is to the different browser types I.e. Google Chrome, Firefox and Internet Explorer - this includes Mobile phones!
- How easy it is to crawl your website (or if crawling is even allowed ...but we'll come to this later) through the use of "Sitemaps"
- What kind of keywords your website has (i.e. In our examples if the user was to search for a query like “Colours” no domain will be returned - as the search engine has not (yet) crawled a domain that has any keywords to do with “Colours”

SEO checkup tool

Robots.txt
When a web crawler accesses a website, it looks for the "robots.txt" file in the root directory (e.g., www.example.com/robots.txt). The contents of this file provide instructions to crawlers about which parts of the site they are allowed to access and crawl.

```
User-agent: *
Disallow: /private/
Allow: /public/
Sitemap: https://www.example.com/sitemap.xml
```

Whilst you can make manual entries for every file extension that you don't want to be indexed, you will have to provide the directory it is within, as well as the full filename. Imagine if you had a huge site! What a pain...Here's where we can use a bit of regexing.

```
User-agent: *

Disallow: /*.ini$

Sitemap: http://mywebsite.com/sitemap.xml

```

Sitemaps

A sitemap is typically an XML file that lists the URLs of the website's pages along with additional metadata, such as the last modified date, the frequency of changes, and the priority of each URL. This information helps search engines prioritize and crawl the pages more effectively.

Using Google for Advanced Searching (Google Dorking or Google hacking)

Operator	Description
site:	Limits search results to a specific website or domain.
filetype:	Searches for specific file types, such as PDF, DOC, or XLS files.
intitle:	Searches for keywords in the title of web pages.
inurl:	Searches for keywords within the URL of web pages.
cache:	Displays the cached version of a webpage from Google's cache.
link:	Lists web pages that link to a specific URL.