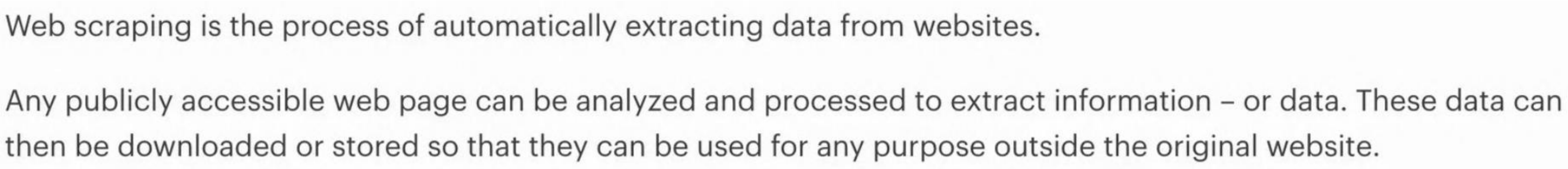
THE BEGINNER’S GUIDE TO WEB SCRAPING

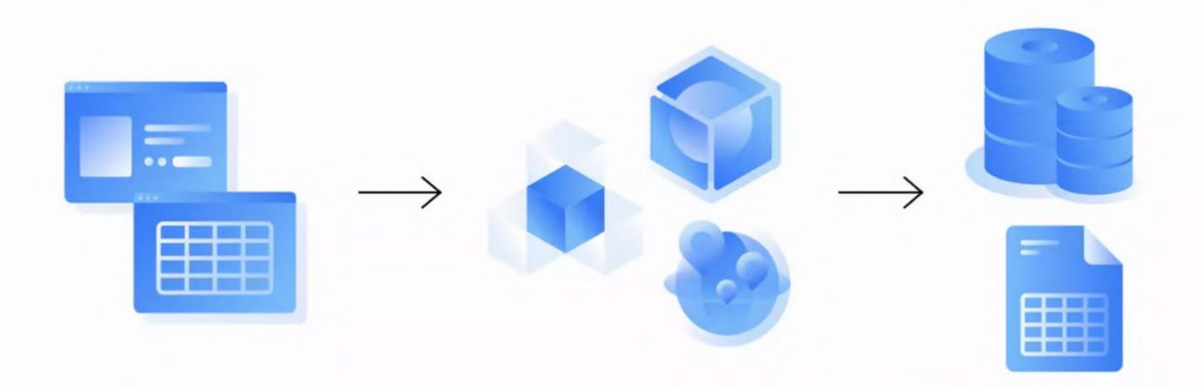
A GUIDE FOR BANKERS

WHAT IS WEB SCRAPING?



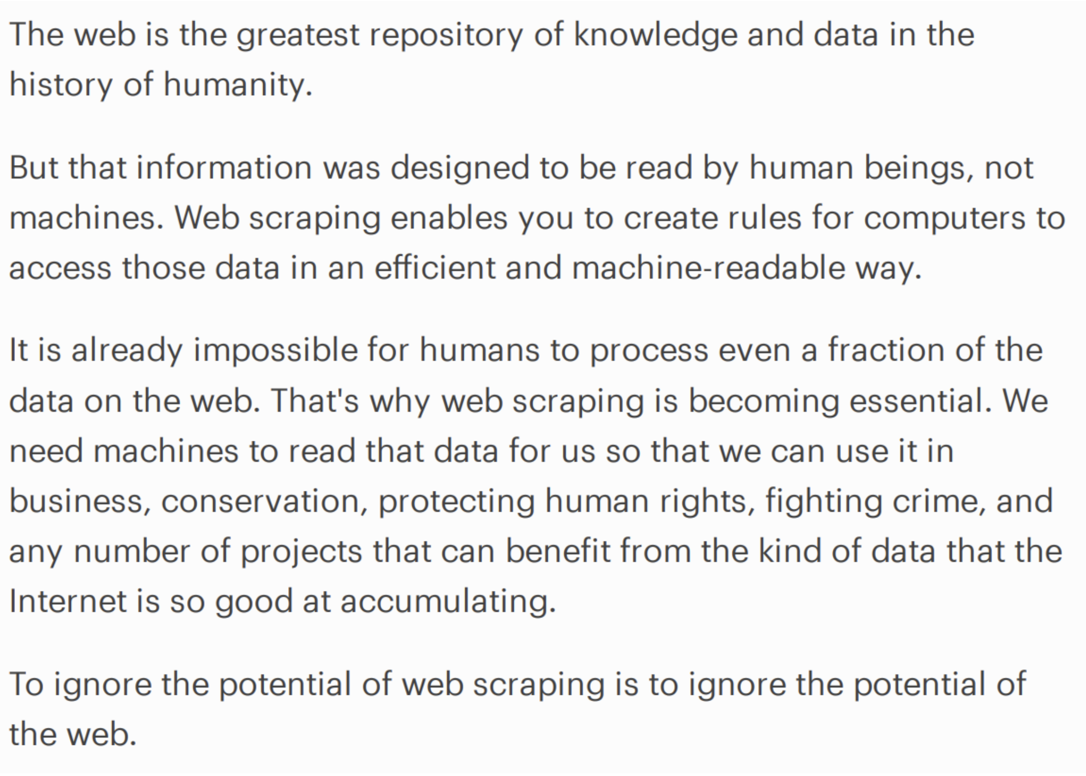
Web scraping is the process of automatically extracting data from websites.

Any publicly accessible web page can be analyzed and processed to extract information – or data. These data can then be downloaded or stored so that they can be used for any purpose outside the original website.



Websites Scraping Structured Data

WHAT IS THE POINT OF WEB SCRAPING?

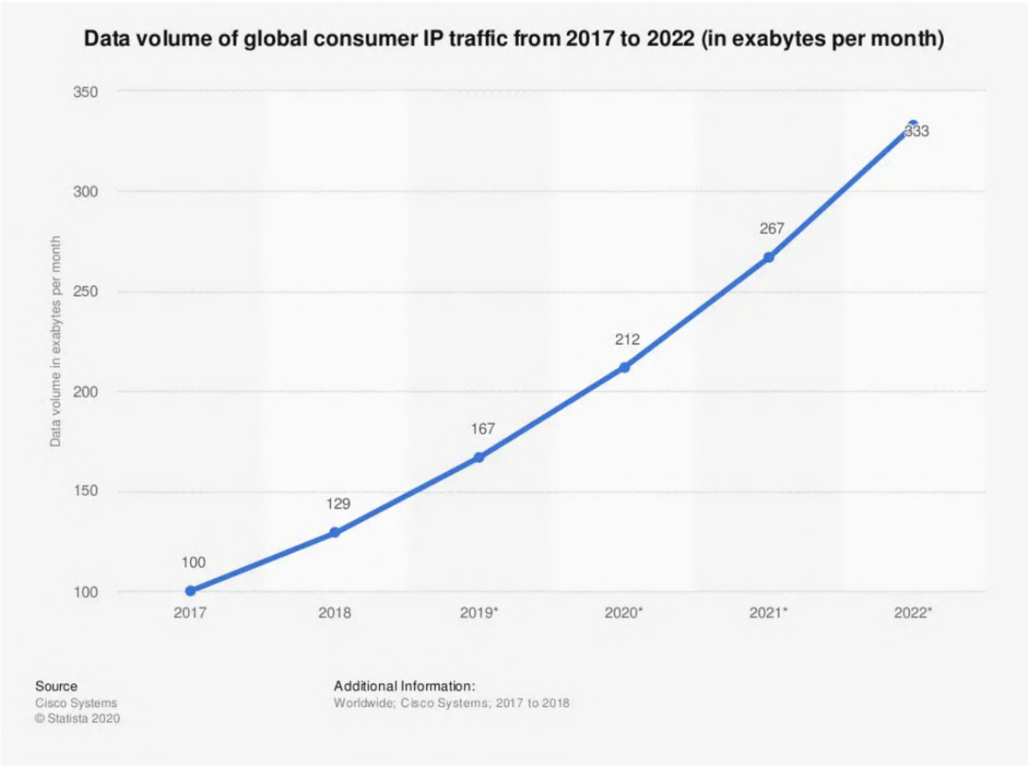


The web is the greatest repository of knowledge and data in the history of humanity.

But that information was designed to be read by human beings, not machines. Web scraping enables you to create rules for computers to access those data in an efficient and machine-readable way.

It is already impossible for humans to process even a fraction of the data on the web. That's why web scraping is becoming essential. We need machines to read that data for us so that we can use it in business, conservation, protecting human rights, fighting crime, and any number of projects that can benefit from the kind of data that the Internet is so good at accumulating.

To ignore the potential of web scraping is to ignore the potential of the web.



Data volume of global consumer IP traffic from 2017 to 2022 (in exabytes per month)

Source
Cisco Systems
© Statista 2020

Additional Information:
Worldwide; Cisco Systems; 2017 to 2018

WHAT IS WEB SCRAPING USED FOR?

Web scraping allows you to collect structured data. Structured data is just a way to say that the information is easy for computers to read or add to a database.

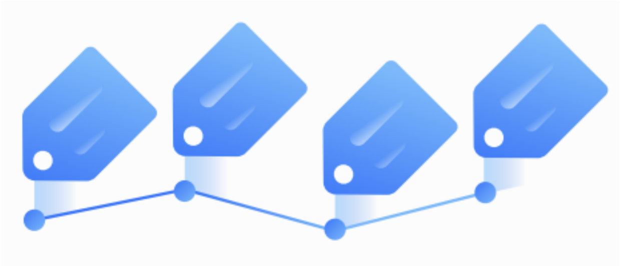
Instead of relying on humans to read or process web pages, computers can rapidly use that data in lots of unexpected and useful ways.

To illustrate the difference, imagine how long it might take you to manually copy and paste text from 100 web pages.

A machine could do it in less than a second if you give it the correct instructions. It can also do it repeatedly, tirelessly, and at any scale. Forget about 100 pages. A computer could deal with 1,000,000 pages in the time it would take you to open just the first few.

APPLICATIONS OF WEB SCRAPING

Here are just some of the ways web scraping can help your business be more efficient and profitable:



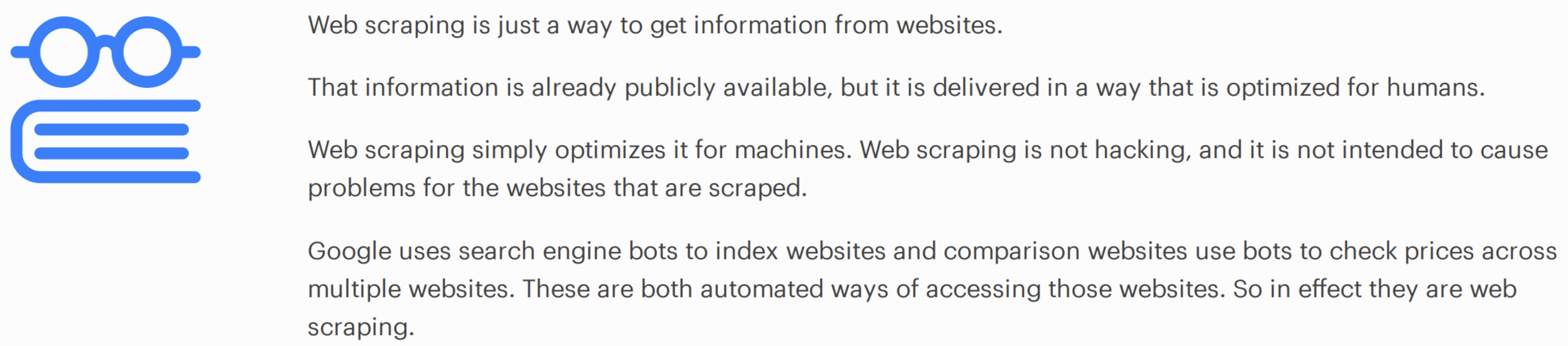
Price tracking

Lead generation

Financial Market Research

Market analysis

IS WEB SCRAPING LEGAL?



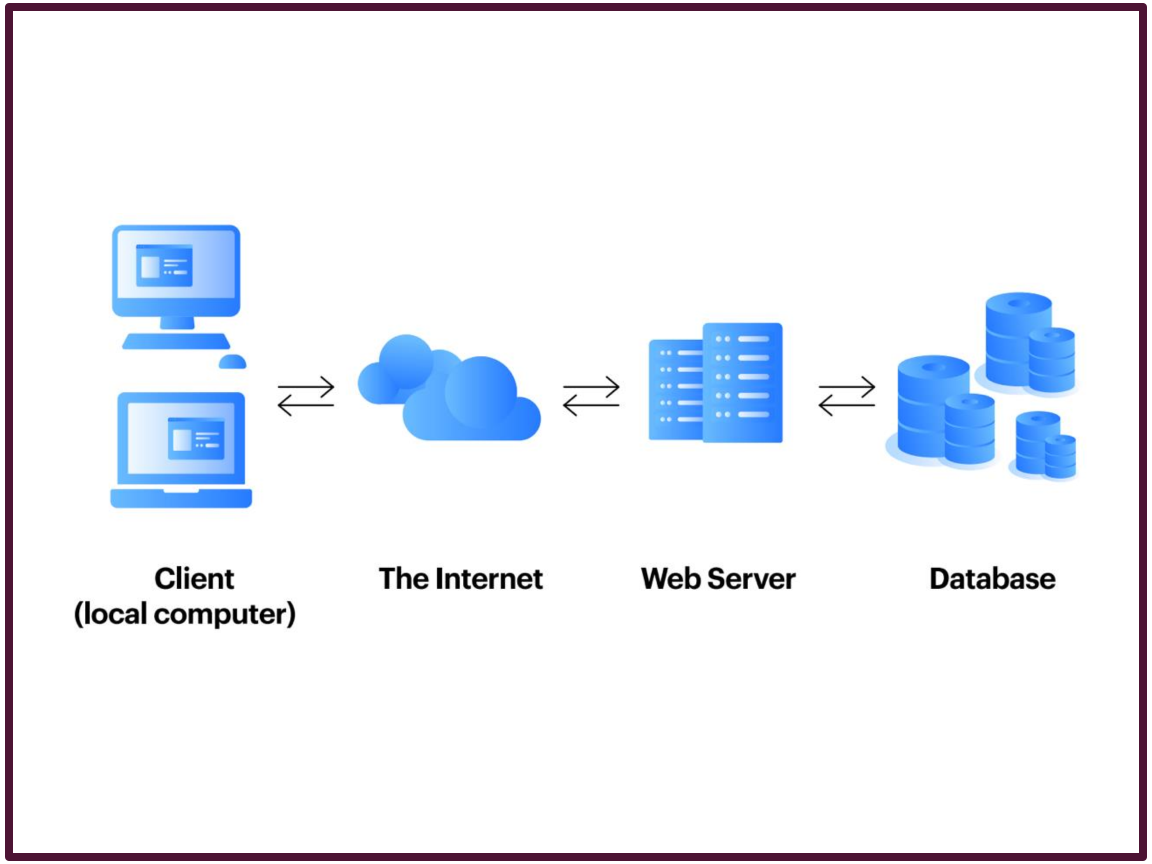
Web scraping is just a way to get information from websites.

That information is already publicly available, but it is delivered in a way that is optimized for humans.

Web scraping simply optimizes it for machines. Web scraping is not hacking, and it is not intended to cause problems for the websites that are scraped.

Google uses search engine bots to index websites and comparison websites use bots to check prices across multiple websites. These are both automated ways of accessing those websites. So in effect they are web scraping.

HOW WEBSITES WORK?



HTTP Protocol

Requests and Response

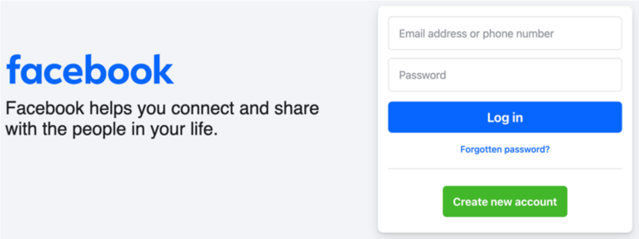
HOW A WEB BROWSER WORKS?

Your browser retrieves information from the web and displays it on your computer or mobile device.

It uses the Hypertext Transfer Protocol (HTTP) to retrieve the content of websites and Hypertext Markup Language (HTML) to determine how to render the content.

The final result is that you see a web page on your device, and you can interact with that web page. Underlying the web page can be a multitude of other technologies, such as HTML, CSS, JavaScript, etc.

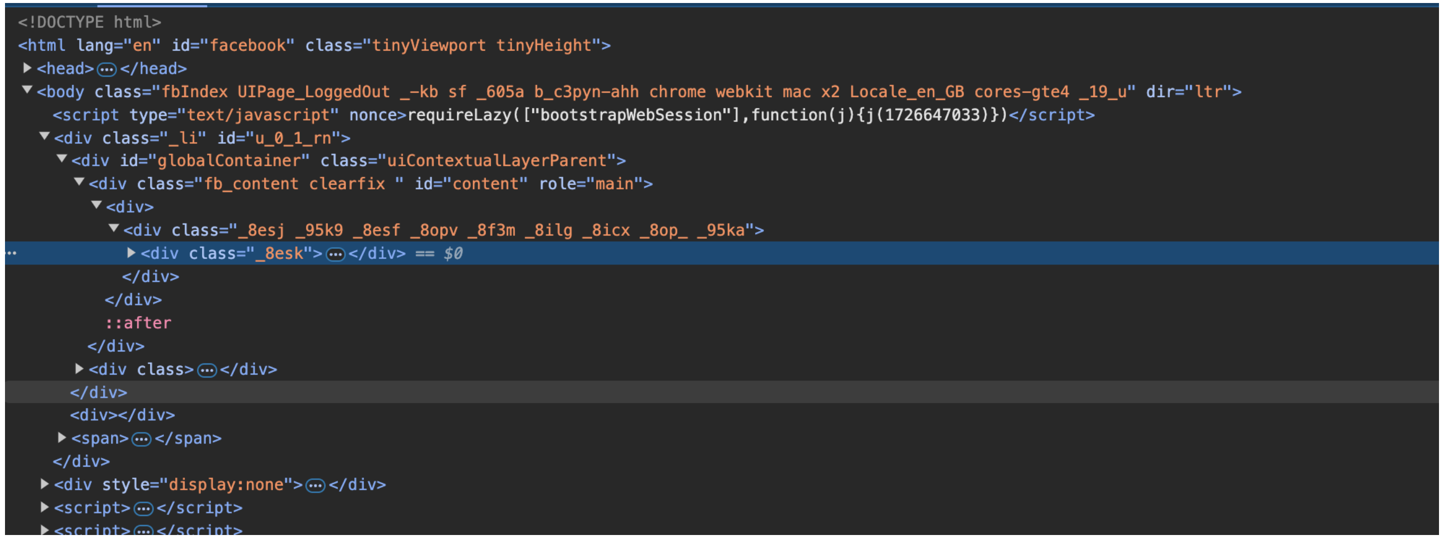
A WEBSITE?



facebook

Facebook helps you connect and share with the people in your life.

WHAT IT LOOKS LIKE



```
<!DOCTYPE html>
<html lang="en" id="facebook" class="tinyViewport tinyHeight">
▶<head>…</head>
▼<body class="fbIndex UIPage_LoggedOut _-kb sf _605a b_c3pyn-ahh chrome webkit mac x2 Locale_en_GB cores-gte4 _19_u" dir="ltr">
   <script type="text/javascript" nonce>requireLazy(["bootstrapWebSession"],function(j){j(1726647033)})</script>
  ▼<div class="_li" id="u_0_1_rn">
    ▼<div id="globalContainer" class="uiContextualLayerParent">
      ▼<div class="fb_content clearfix " id="content" role="main">
        ▼<div>
          ▼<div class="_8esj _95k9 _8esf _8opv _8f3m _8ilg _8icx _8op_ _95ka">
…           ▶<div class="_8esk">…</div> == $0
            </div>
          </div>
          ::after
        </div>
      ▶<div class>…</div>
     </div>
     <div></div>
    ▶<span>…</span>
   </div>
  ▶<div style="display:none">…</div>
  ▶<script>…</script>
  ▶<script>…</script>
```

WHAT IT ACTUALLY IS

A BIT ABOUT HTML

- HTML is the **language in which most websites are written**. HTML is used to create pages and make them functional.
- **Heading -> <h1>Heading</h1>**
- This is a sample text -> <p>This is a sample text</p>
- <div></div>
- <table></table>
- www.google.com -> <a href=“https://google.com/”></a>

LET’S JUMP RIGHT IN!