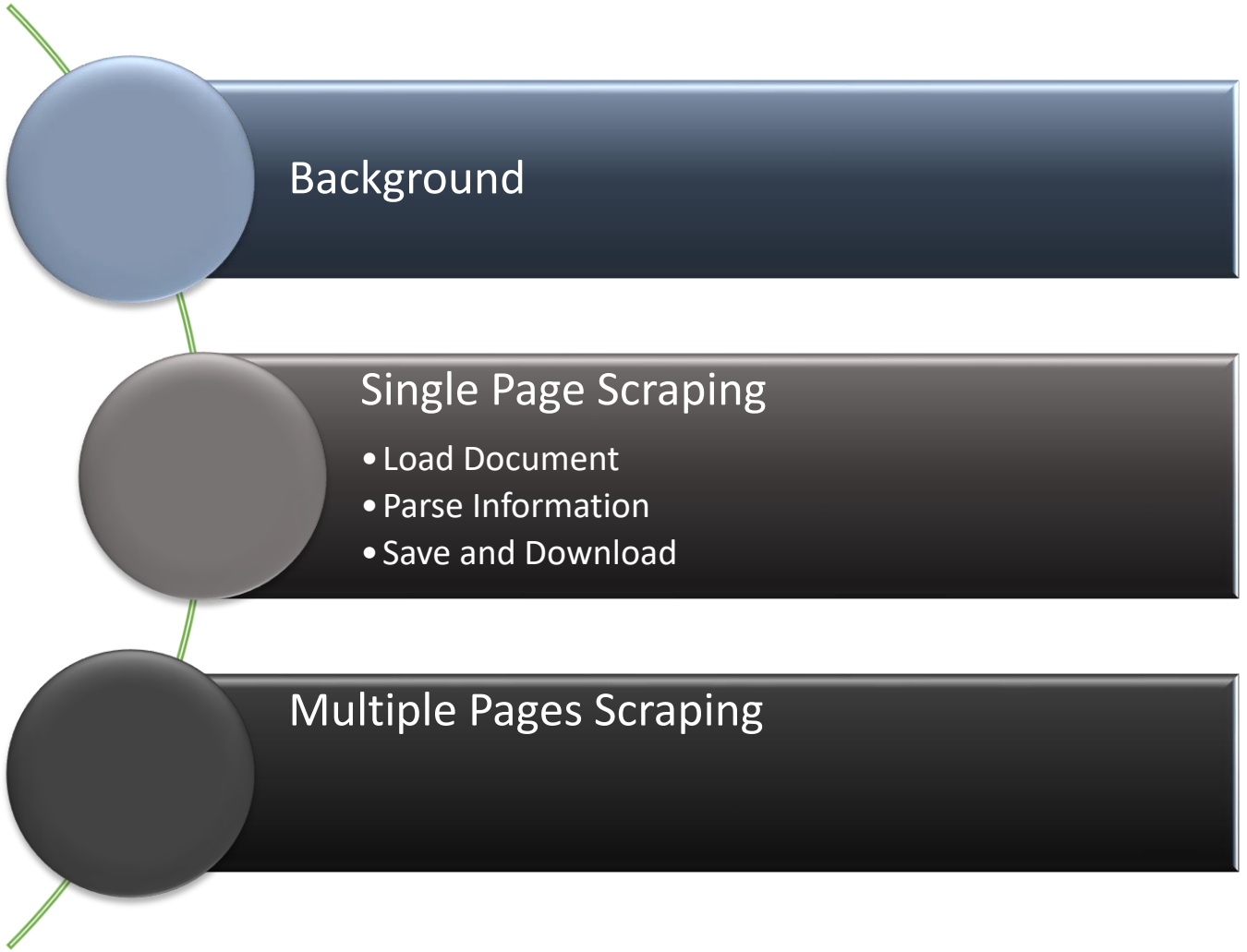
Python Scraping

Outline


Background
Single Page Scraping
• Load Document
• Parse Information
• Save and Download
Multiple Pages Scraping

Background

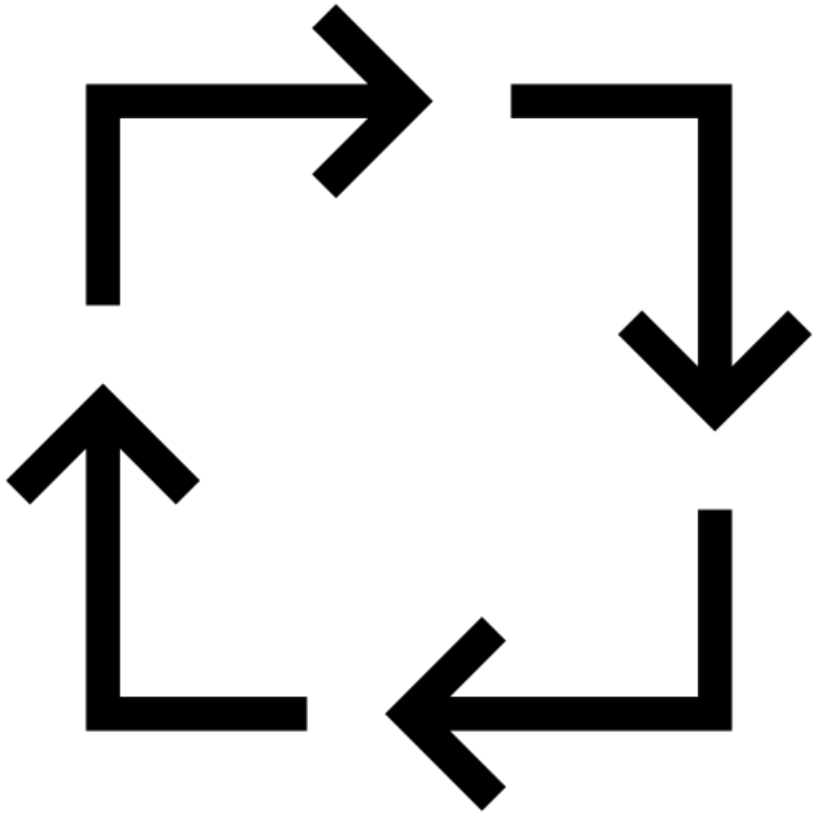
Open a webpage

Files

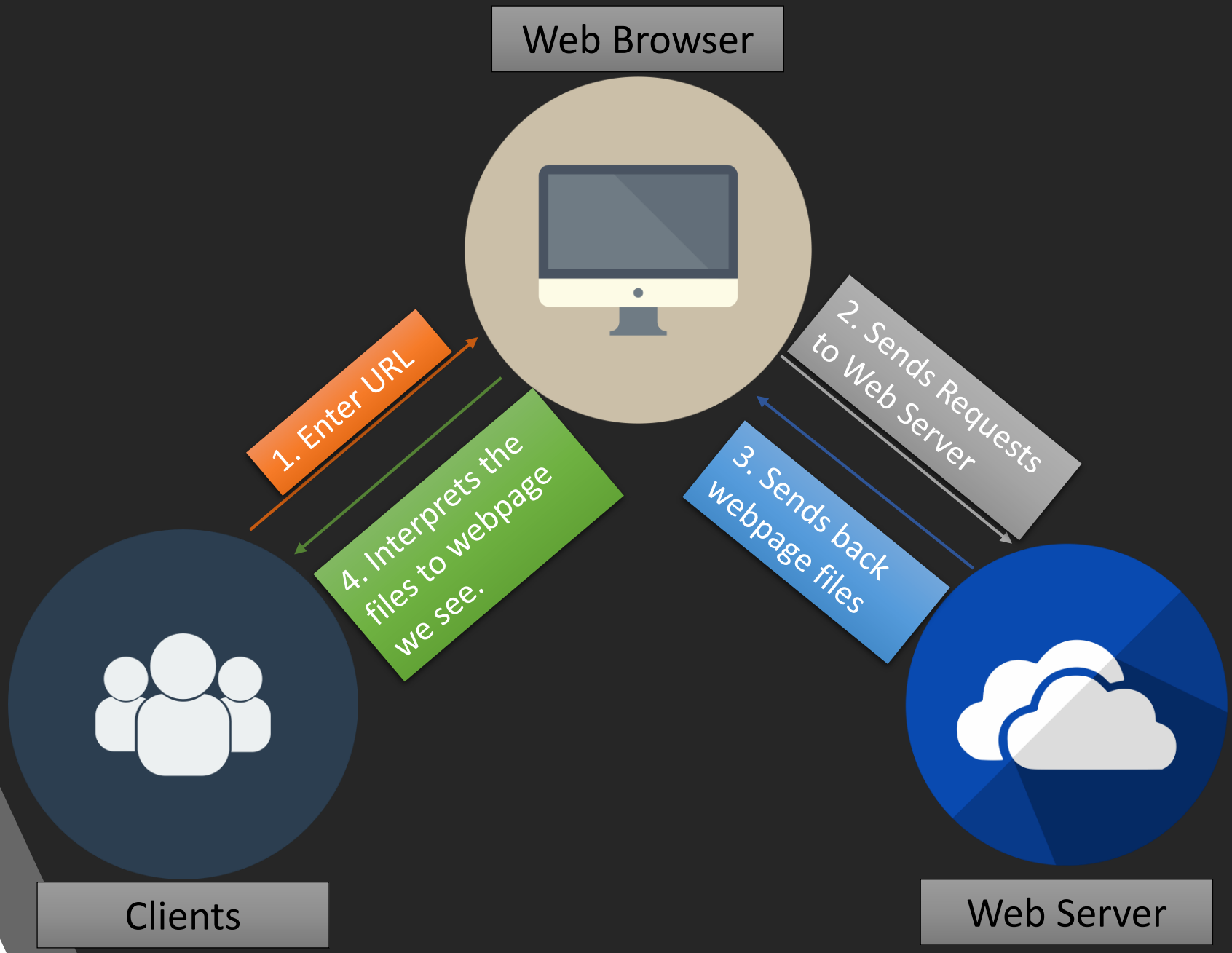
Structure

Selector

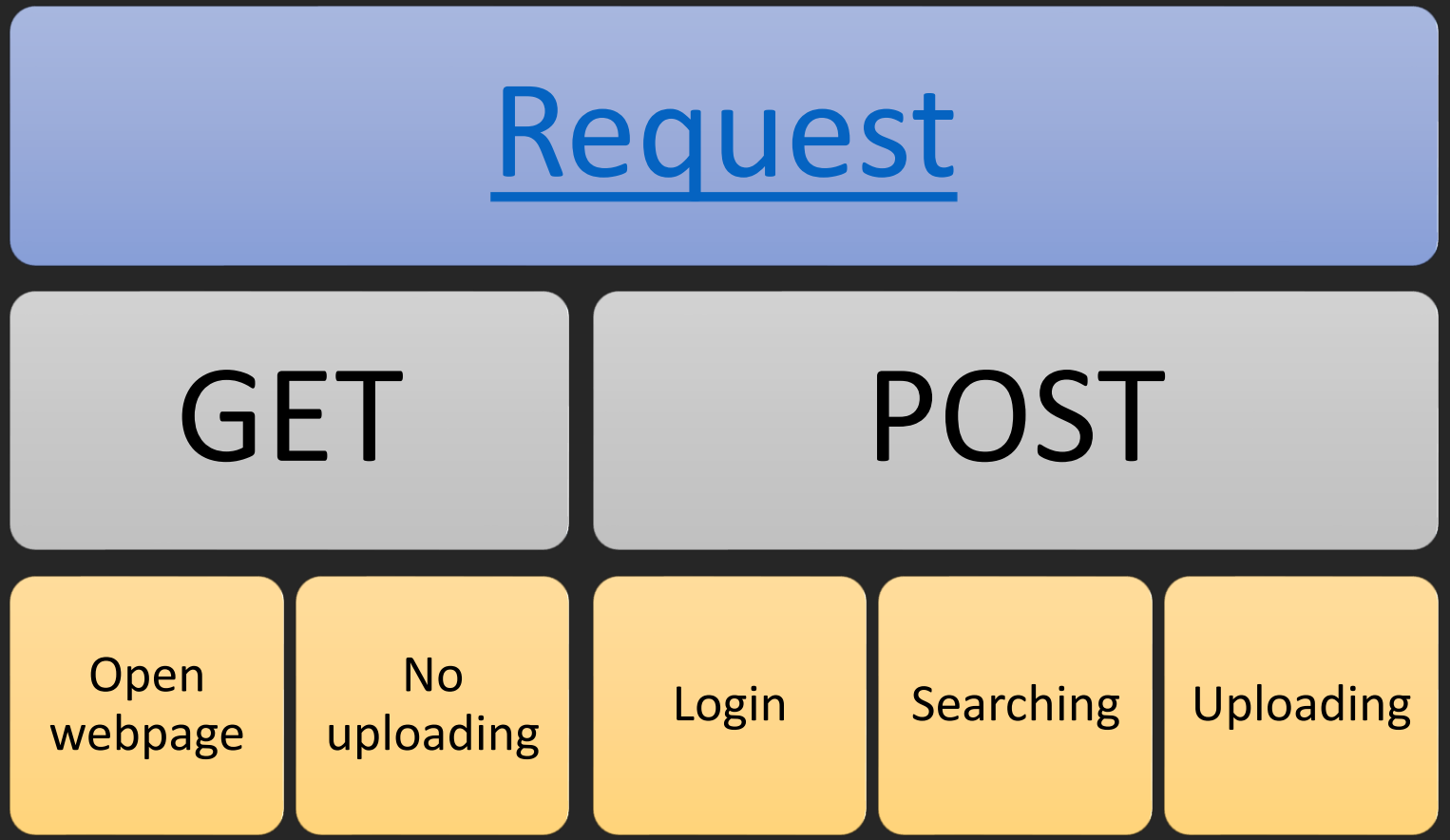
Protocol



How to open a webpage?


Web Browser
1. Enter URL
2. Sends Requests to Web Server
3. Sends back webpage files
4. Interprets the files to webpage we see.
Clients
Web Server

Request Method



Files

File types

<!DOCTYPE html>

What is HTML?

HTML is the standard markup language for creating Web pages.

- HTML stands for Hyper Text Markup Language
- HTML describes the structure of Web pages using markup
- HTML elements are the building blocks of HTML pages
- HTML elements are represented by tags
- HTML tags label pieces of content such as "heading", "paragraph", "table", and so on
- Browsers do not display the HTML tags, but use them to render the content of the page

</html>


```
<!DOCTYPE html>
<html lang="en-US">
```

What is CSS?

- CSS stands for **C**ascading **S**tyle **S**heets
- CSS describes **how HTML elements are to be displayed on screen, paper, or in other media**
- CSS **saves a lot of work**. It can control the layout of multiple web pages all at once
- External stylesheets are stored in **CSS files**

```
</html>
```

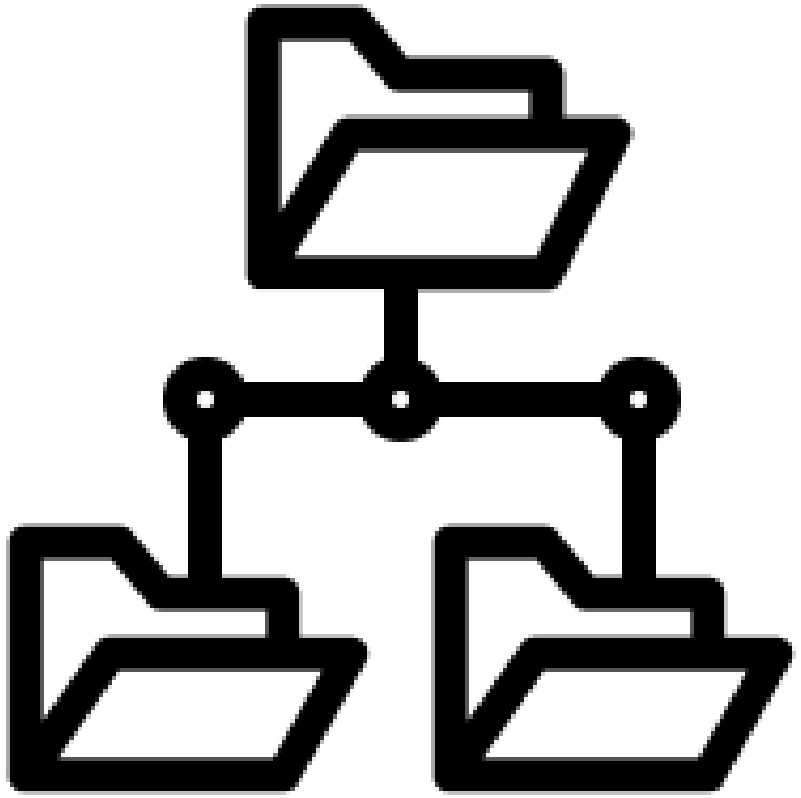
```
body {
  background-color: lightblue;
}
h1 {
  color: navy;
  margin-left: 20px;
}
```

What is JavaScript?

JavaScript Can change HTML

- JavaScript can change HTML Content
- JavaScript can change HTML Attribute Values
- JavaScript can change HTML Styles (CSS)
- JavaScript can change HTML Elements
- JavaScript can show HTML Elements

```
</p>
<button type="button"
onclick='document.getElementById("de
mo").innerHTML="Do not Worry about
Scraping!"'>Click Me!</botton>
</body>
</html>
```

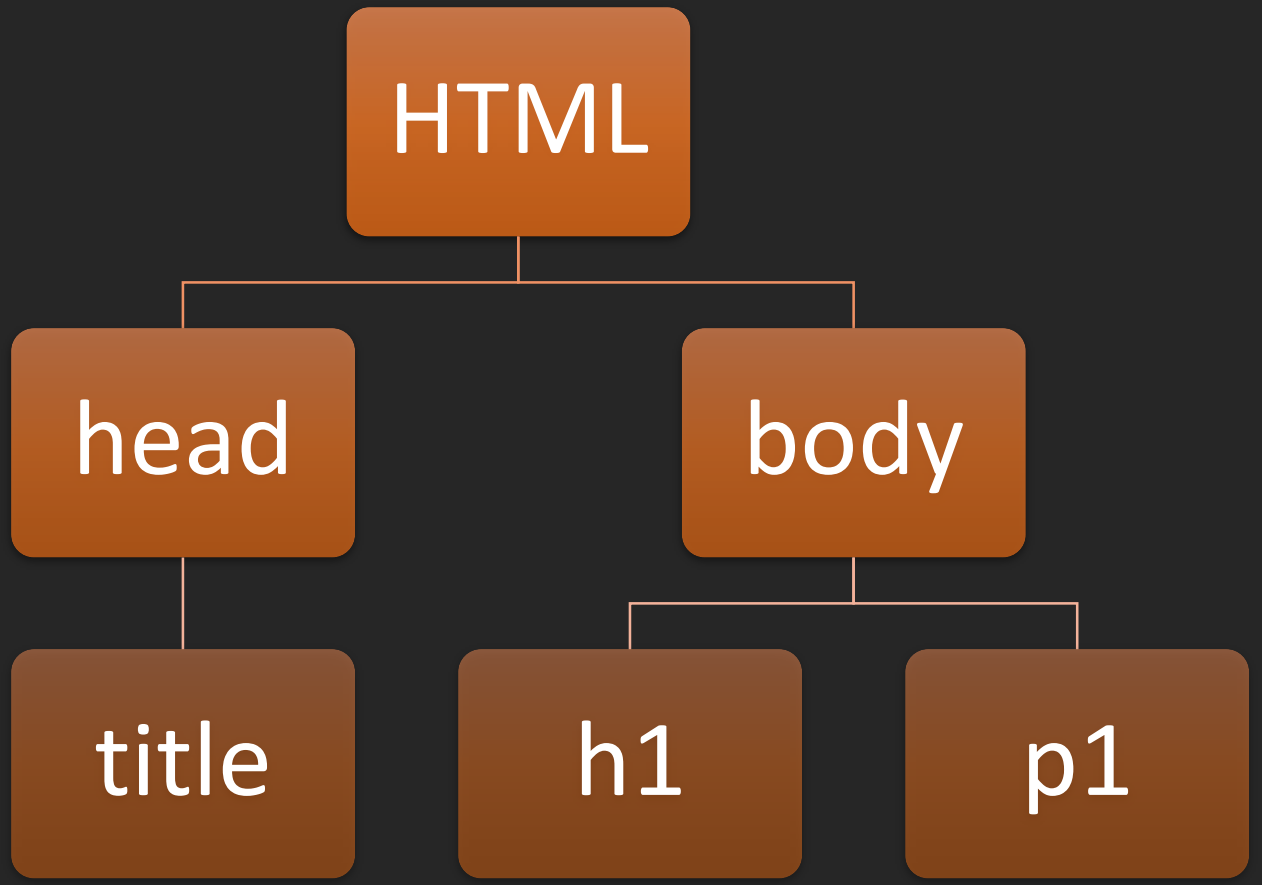


Structure


SAMPLES
SAMPLES
SAMPLES


IMDb

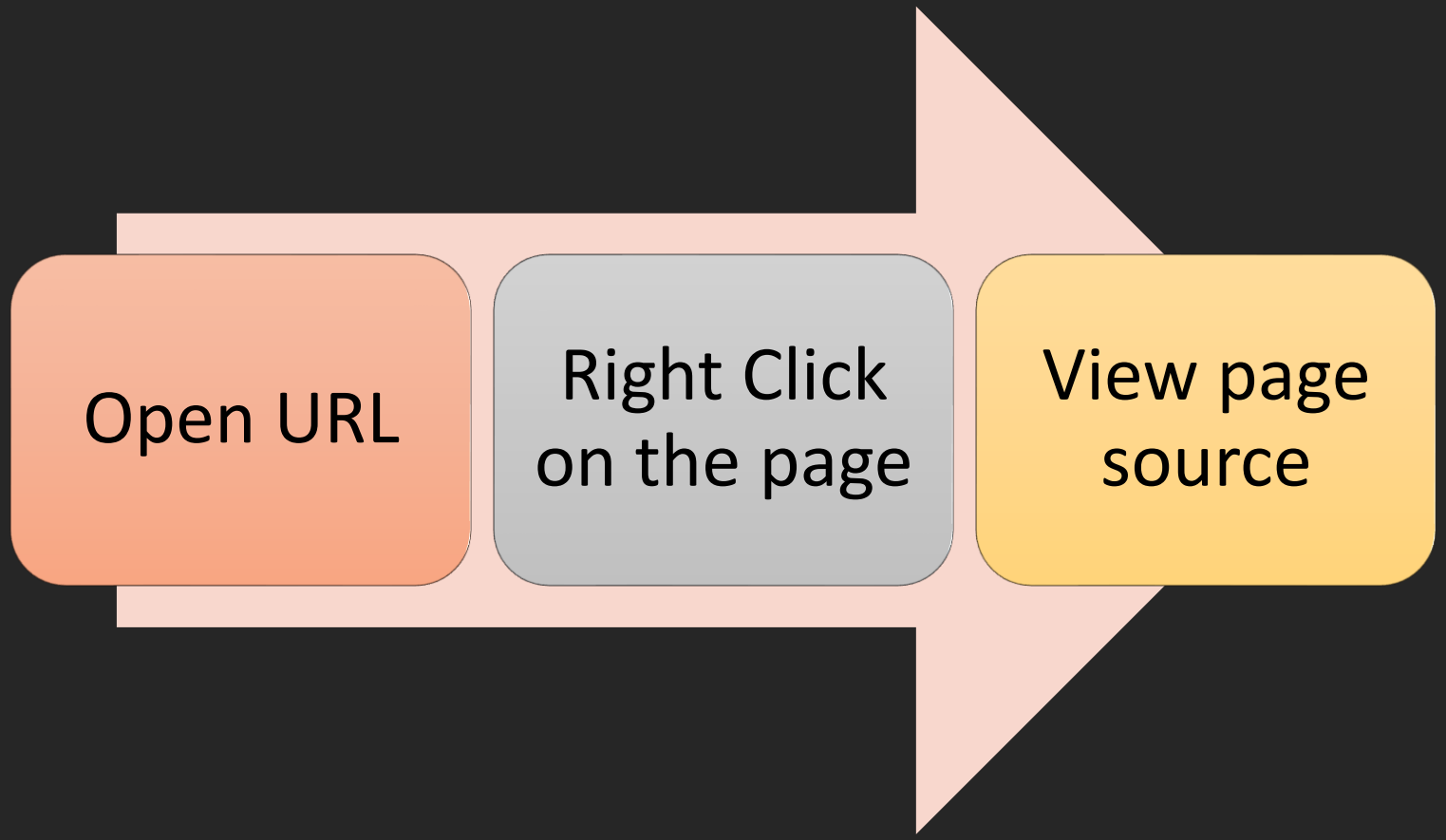
Structure of Sample HTML



Structure of IMDb

How to check the structure (I)

View page source



Structure of IMDb

How to check the structure (II)

Inspect webpage element

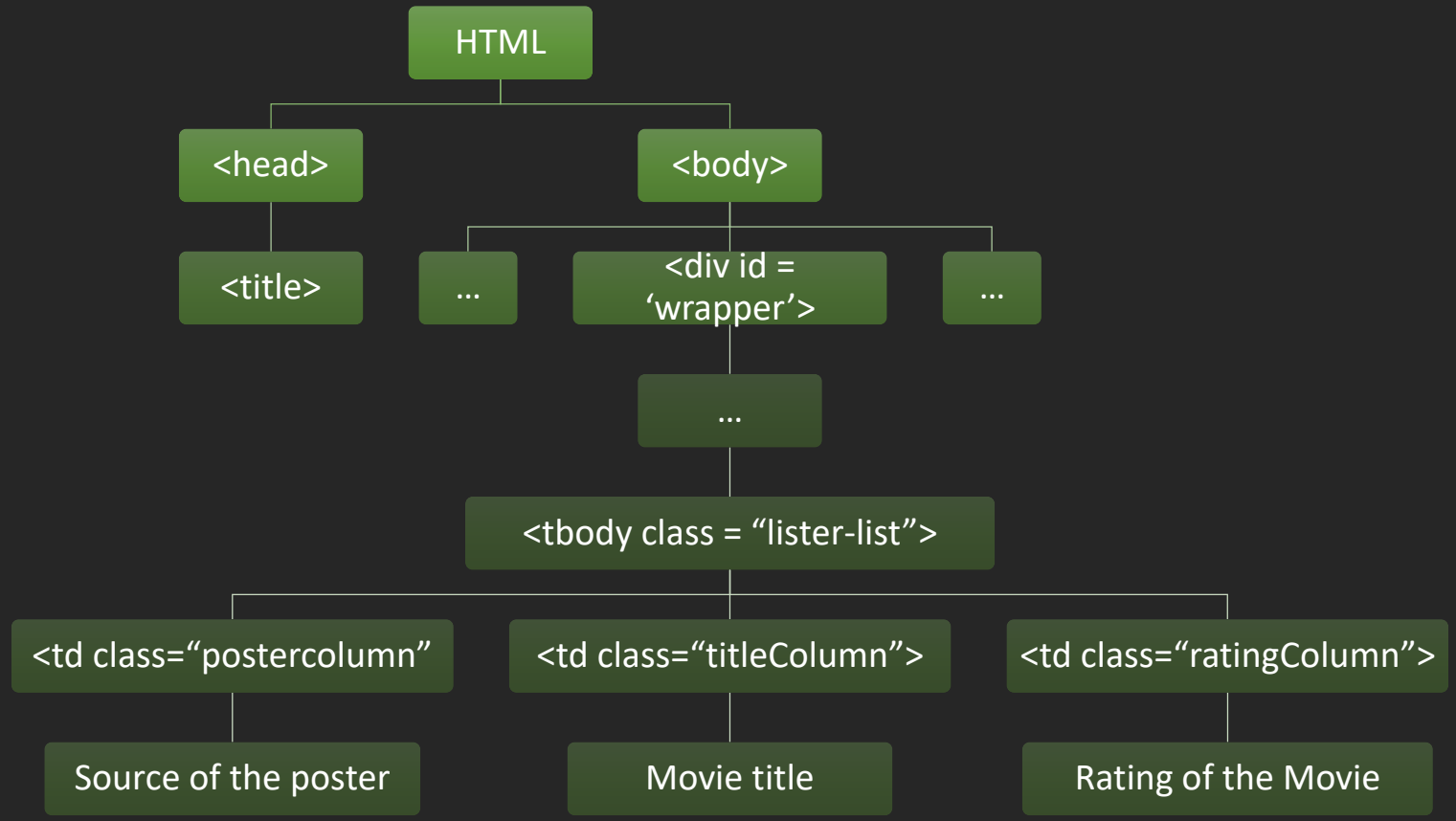
Open URL

Place Mouse on the context interested in

Right Click and Inspect

Structure of IMDb

Top 250 Movies of history



Selectors

Regular Expression I

.	• Match any character other than “\n”
\w	• Match any alphabet, digit or underline “_”
\s	• Match any space
\d	• Match any digit
^	• Match the start of string
$	• Match the end of string

Regular Expression II

*	• Appear none or more times
+	• Appear once or more times
?	• Appear none or once
{n}	• Appear n times
{n,}	• Appear n times or more
{n, m}	• Appear n to m times

Xpath Selector

Nodename	• Select all subnode under this node
/	• Select from root node
//	• Select from current node
.	• Select current node
..	• Select the parent node of current node
@	• Select any attribute

CSS Selector

.class	• e.g. “.intro” • Select all elements class = “intro”
#id	• #firstname • Select all elements id = “firstname”
*	• * • Select all elements
element	• p • Select all element <p>
ele, ele	• div, p • Select all and <p> element
ele ele	• div p • Select all <p> element inside element
ele>p	• div>p • Select all <p> element whose parent is element

Protocol

Rules

Scrape Open Public data from Internet

Slow down your speed

Follow Robots exclusion standard

- Can be found on www.example.com/robots.txt

Do not use for commercial purpose

Do not publish your scraping code or data

Single Page Scraping

Load Source Document

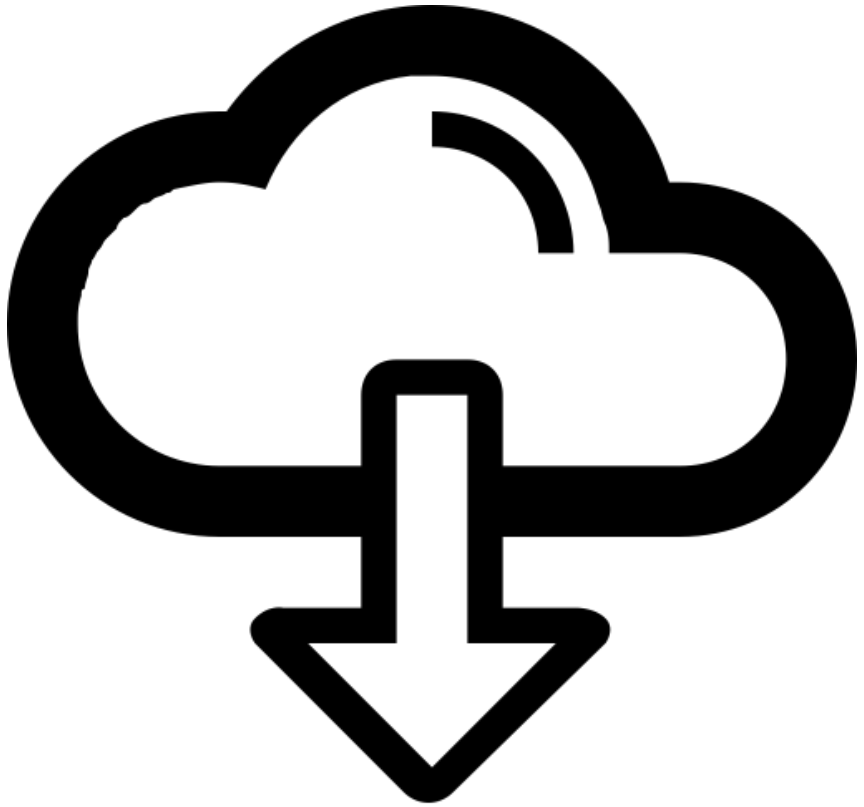
- urllib
- requests

Parse Information

- BeautifulSoup
- Lxml
- re

Save data

- Pandas



Load Source Document

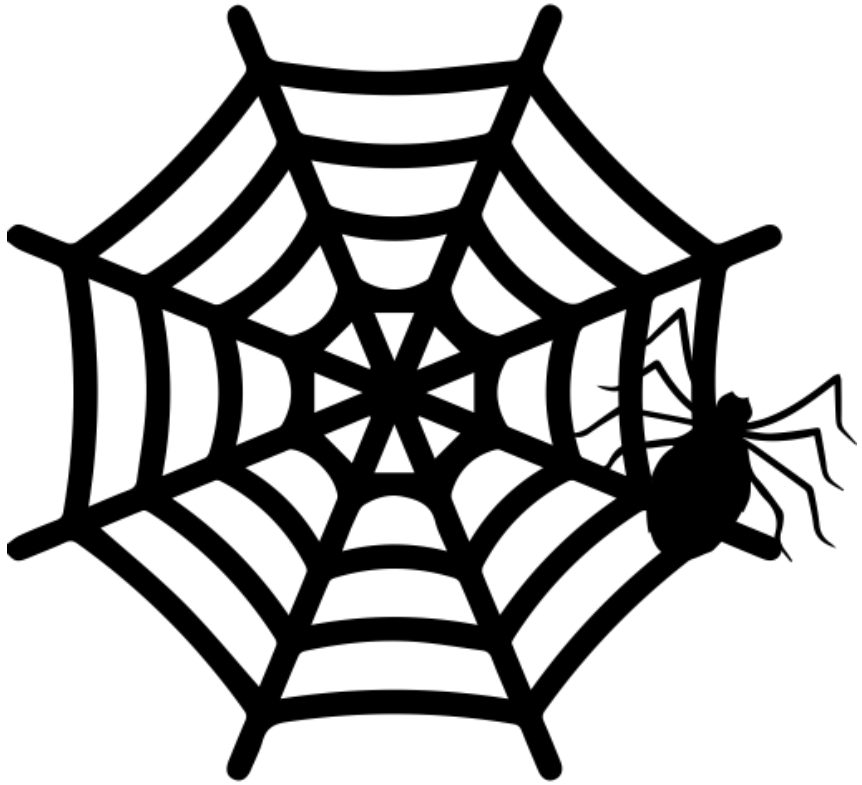
urllib

- urllib is built-in package of Python3
- urllib is a package that collects several modules for working with URLs

```
# import function "urlopen" from "urllib" to python
from urllib.request import urlopen
# define the url
URL = "https://www.imdb.com/chart/top"
# "urlopen" will load the URL as object
html = urlopen(URL)
# read the html as readable type (string)
html_docs = html.read().decode()
```

requests

- requests is the only *Non-GMO* HTTP library for Python, safe for human consumption;
- requests recommend the use of Python3 over Python2;
- requests can be installed by:
  - pip3 install requests



Parse and save Information

BeautifulSoup

- BeatutifulSoup is a Python library for pulling data out of HTML files;
- BeautifulSoup provides ways of navigating, searching, and modifying the parse tree;
- BeautifulSoup can be installed by:
  - pip3 install beautifulsoup4

lxml

- lxml is the most feature-rich and easy-to-use library for processing XML and HTML in the Python language;
- lxml is compatible but superior to the well-known ElementTree API;
- lxml can be installed by:
  - pip3 install lxml

Multiple Pages Scraping