Webscraping

Workshop JADS - Joël Luijmes

https://github.com/joelluijmes/Workshop-JADS

Content

- Web Development
- Approach
- Interactive

Web Development

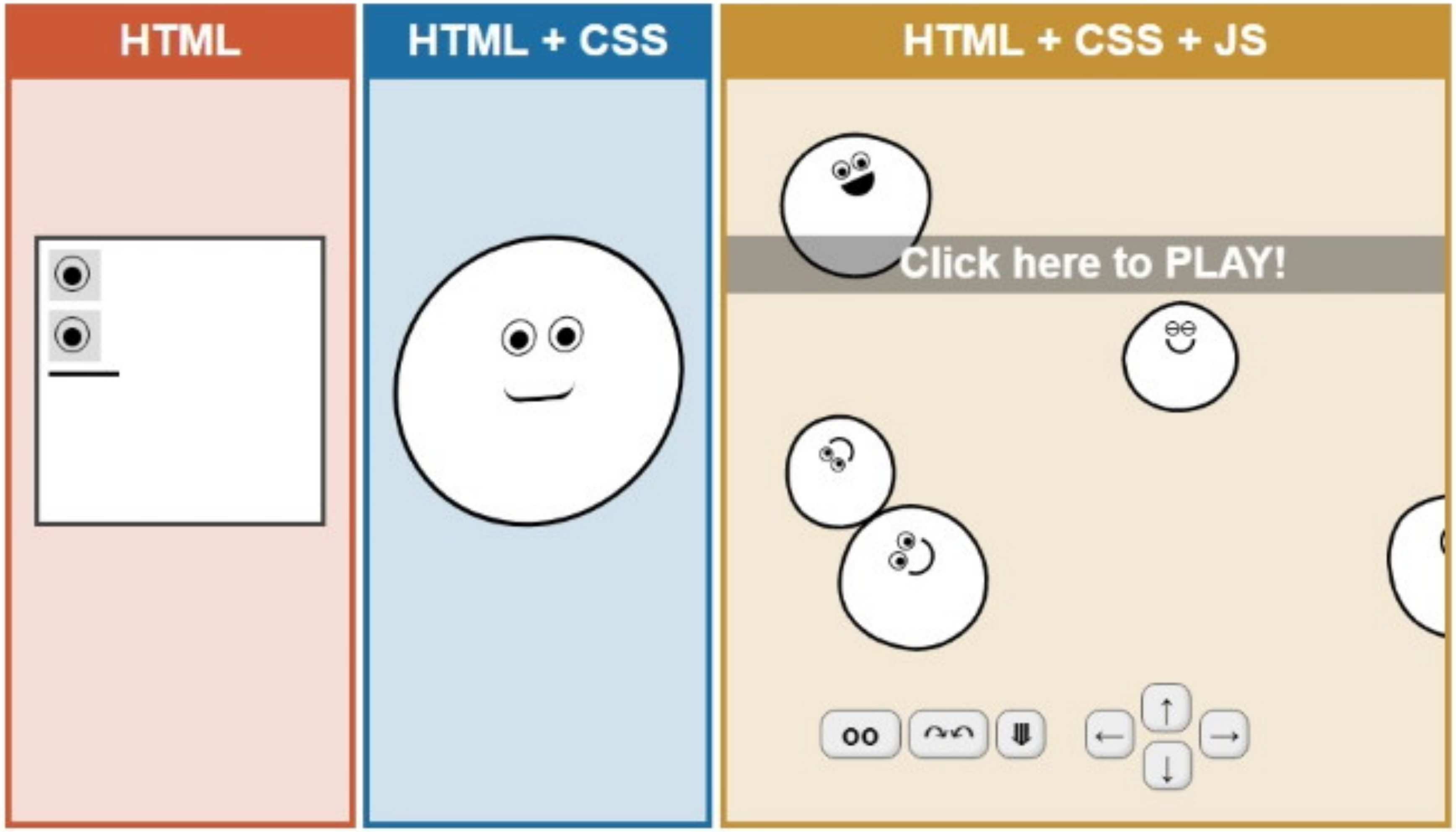
- HTML (HyperText Markup Language)
  Defines the structure and actual content

- CSS (Cascading Style Sheets)
  Presents the content in fancy way

- JavaScript
  Interaction

- AJAX / XHR
  Retrieves data (JSON, HTML, anything) asynchronously

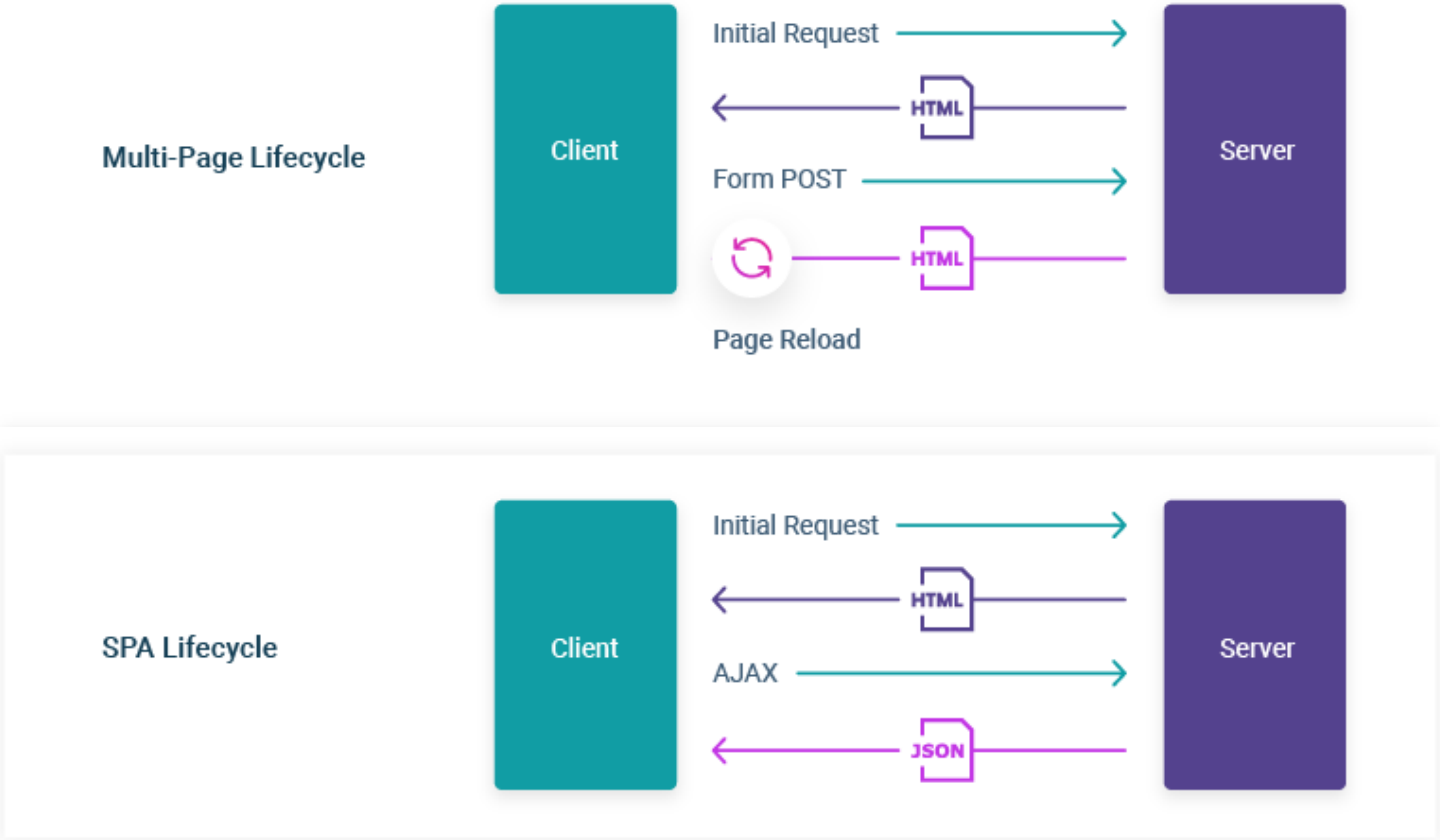
Web Development



https://html-css-js.com

Demo I

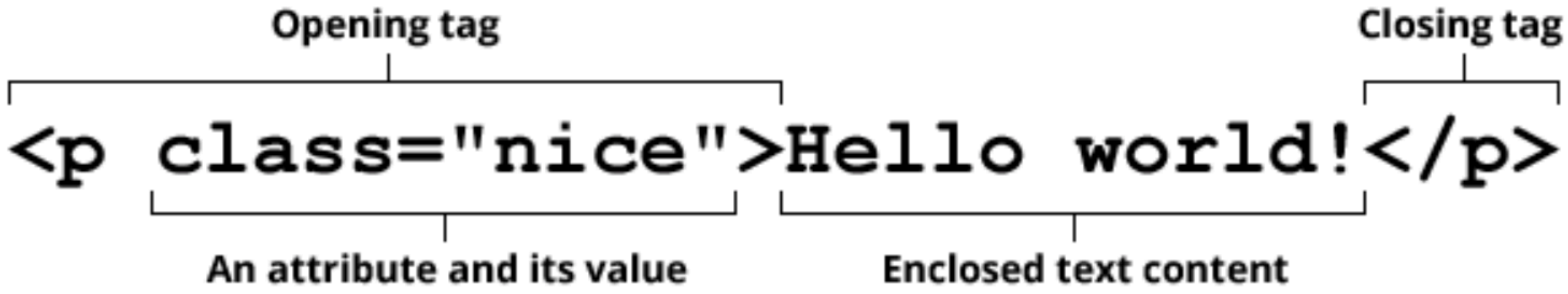
Web Development



https://dotcms.com/blog/post/what-is-a-single-page-application-and-should-you-use-one-

Demo II

HTML



Opening tag

Closing tag

```
<p class="nice">Hello world!</p>
```

An attribute and its value

Enclosed text content

https://blog.nus.edu.sg/3217w3/category/uncategorized/

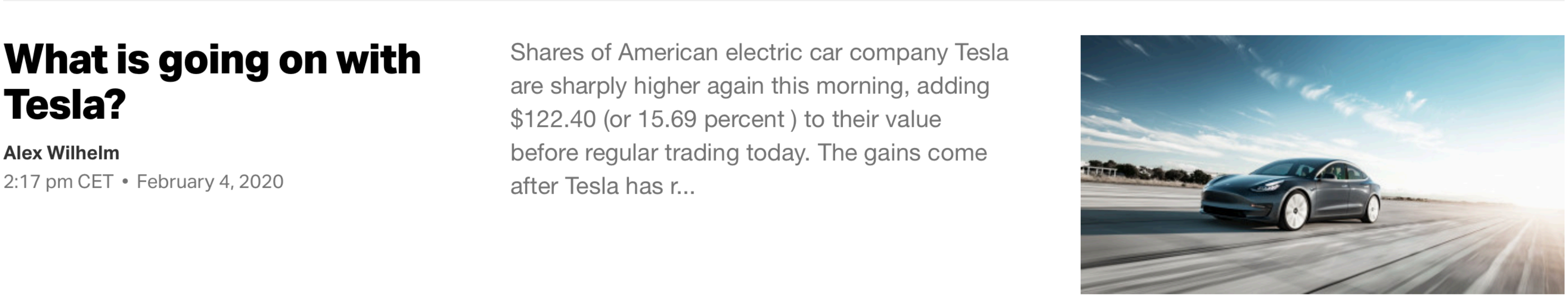
HTML

Common elements

Element	Name	Description
p	paragraph	Text content
div		Group elements / section / divison
span		Group inline-elements
h1, h2, … h6	heading	H1 is most important, H6 the least
img	image	Defines image, required attributes: src and alt
a	anchor	Hyperlink to link different section / pages
style		Section for inline CSS
link		Link to external resource (CSS file)
script		Inline JavaScript or link to external resource

HTML

```html
<header class="post-block__header">
  <h2 class="post-block__title">
   <a class="post-block__title__link" href="/2020/02/04/what-is-going-on-with-tesla/">
   What is going on with Tesla?
   </a>
  </h2>
  <div class="post-block__meta">
   <div class="river-byline">
    <span class="river-byline__authors">
    <span>
    <a aria-label="Posts by Alex Wilhelm" href="/author/alex-wilhelm/">
    Alex Wilhelm
    </a>
    </span>
    </span>
    <div class="river-byline__full-date-time__wrapper">
     <time class="river-byline__full-date-time" datetime="2020-02-04T13:17:52" >2:17 pm CET
     <span class="full-date-time__separator">•</span> February 4, 2020</time >
    </div>
   </div>
  </div>
 </header>
```



What is going on with Tesla?

Alex Wilhelm

2:17 pm CET • February 4, 2020

Shares of American electric car company Tesla are sharply higher again this morning, adding $122.40 (or 15.69 percent) to their value before regular trading today. The gains come after Tesla has r...

CSS Selector

selects one or more HTML elements

Selector	Example	Description
element	p	Selects all *p* elements on page
.class	.title	Selects *all elements* with class=“title”
#id	#container	Selects element where id=“container”
.class.class	.price.new	Selects *all elements* with **both** classes: class=“price new”
element.class	h5.description	Selects all *h5* elements with class=“description”
element#class	div#container	Selects *div* element with id=“container”
selector selector	div img	Selects all *img* elements within a *div*, regardless depth
selector > selector	div > img	Selects all *img* elements **directly** child of parent *div*
selector, selector	span, div	Selects all *span* and *div* elements

Quiz

Go to **www.menti.com** and use the code **68 39 48**

Session

- Tracking state of user
- Cookies
    - Sign-in
    - Cookie Walls / Consent

Approach

1. Retrieve the webpage (HTML)

2. Extract the information using CSS selectors

Approaches

Low-level

- Manually retrieve page (aiohttp, requests)
- Manually parse page (lxml)
- Superfast
  Minimal resources
- Does not always work

Head-less browser

- ‘Executes’ the page (Selenium, GhostJS)
- Allows for interaction (Clicking buttons)
- Slow
  Huge resources
- Usually works

Libraries / Tools

Library	Description
requests / aiohttp	Make HTTP requests to fetch / retrieve websites
lxml	Parse HTML files to extract information
BeautifulSoup	Parse HTML files to extract information
Scrapy	Combines retrieval and extract information
Selenium	Automates “browsers” can both retrieve sites and extract information
Scraper API	Commercial tool for scraping websites
Puppeteer	Similar to Selenium but for NodeJS developers
Chromium	Browser (Google Chrome) can be ran headless

Interactive