Webscraping

Workshop JADS - Joël Luijmes

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

Content

- Web Development
- Approaches
- Interactive

Web Development

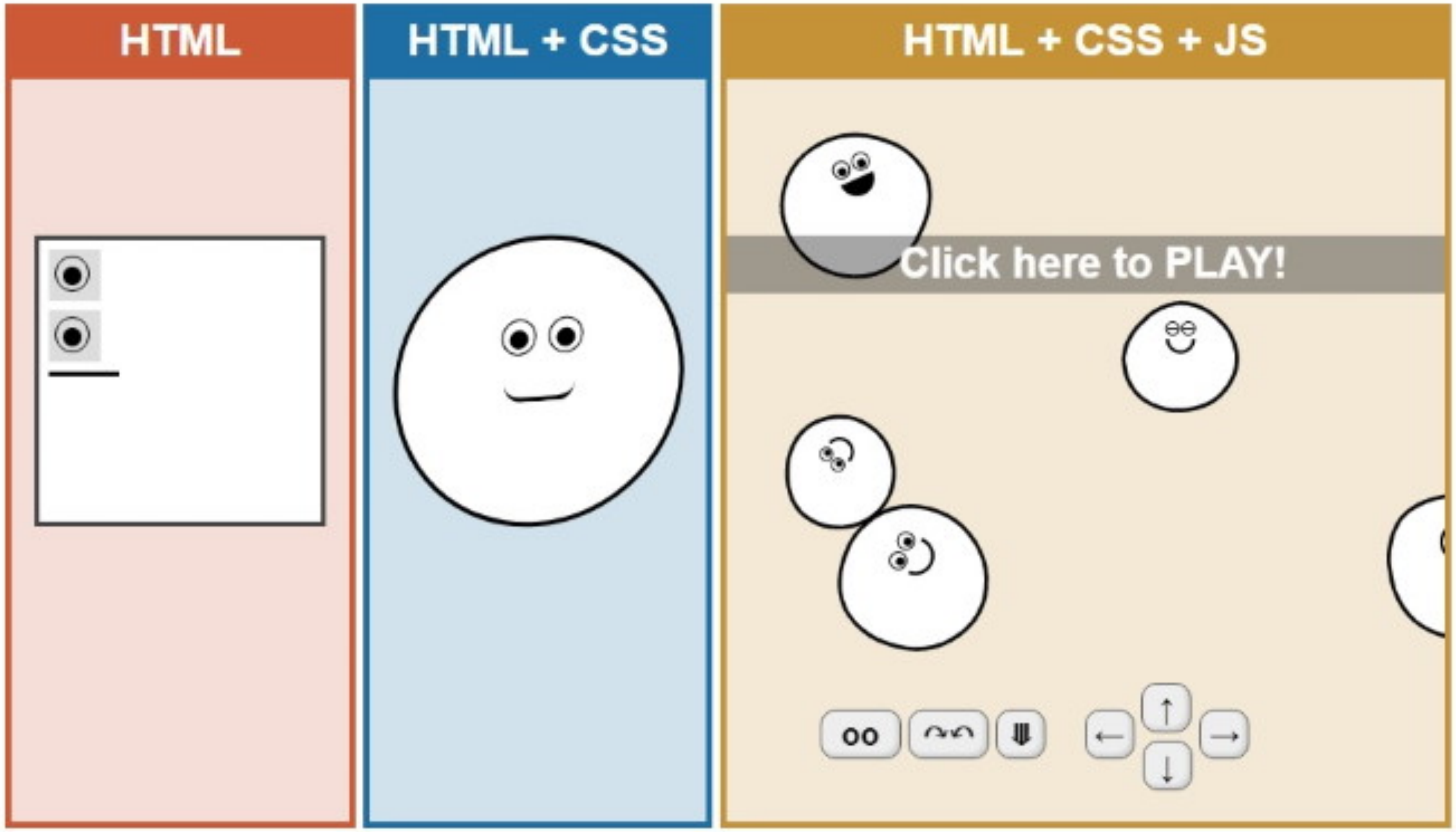
- HTML (HyperText Markup Language)
  Defines the structure and actual content

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

- JavaScript
  Interaction

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

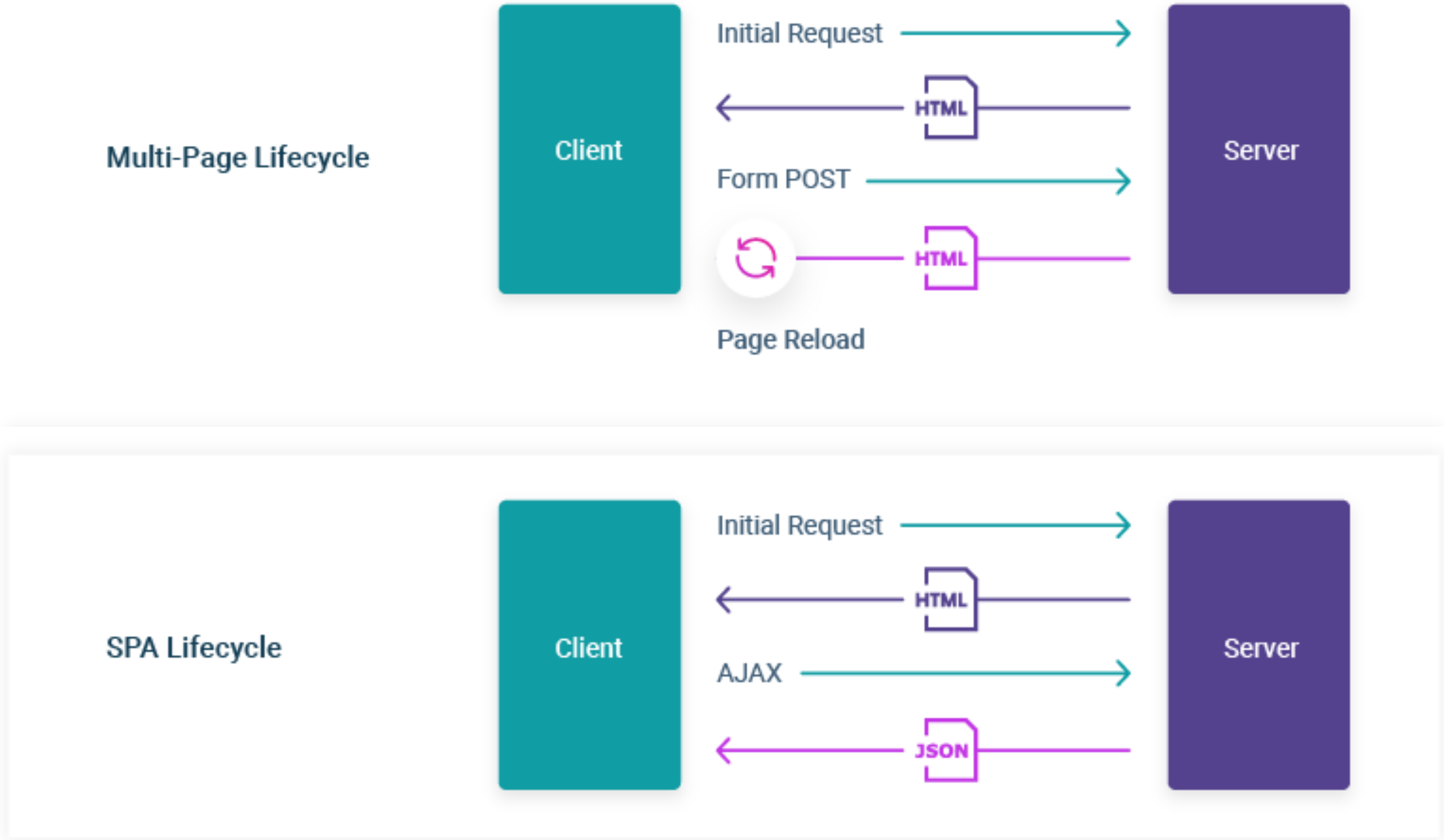
Web Development



https://html-css-js.com

Demo I

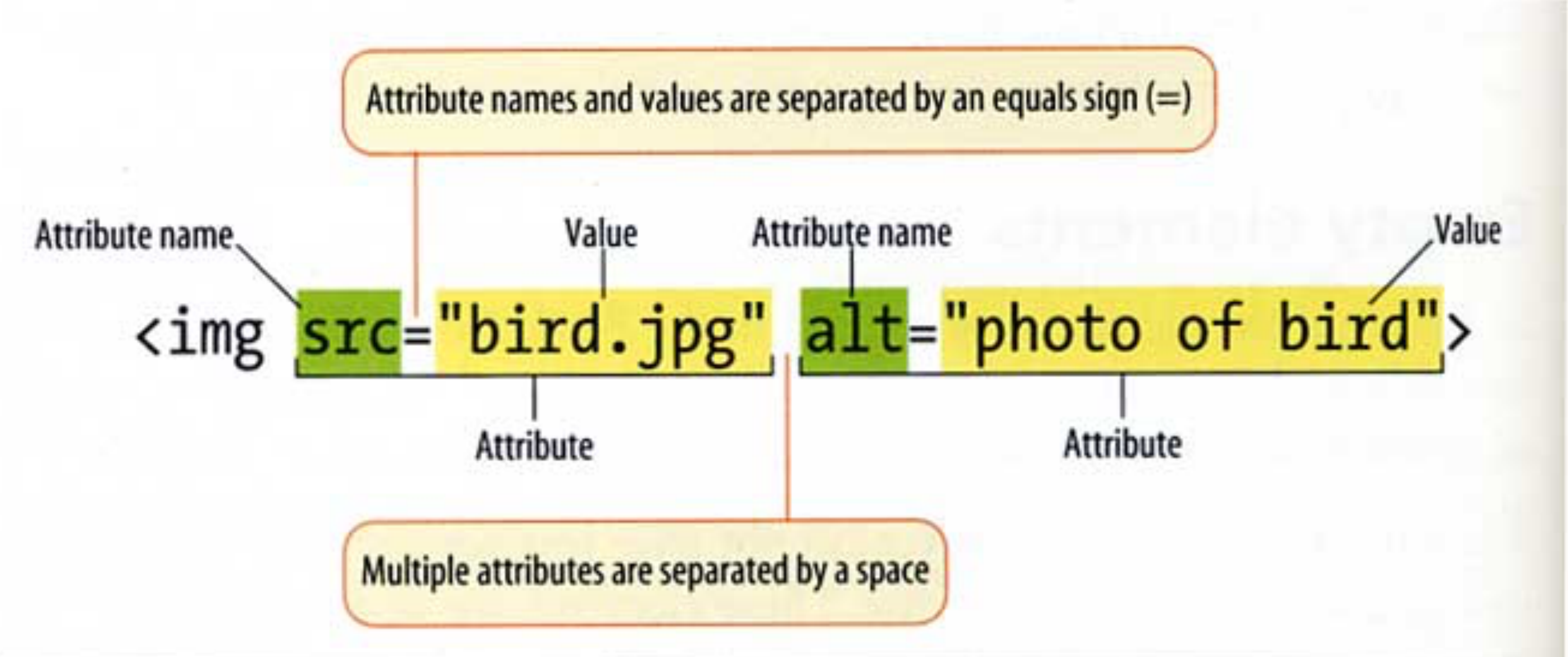
Web Development



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

Demo II

HTML



```
<img src="bird.jpg" alt="photo of bird">
```

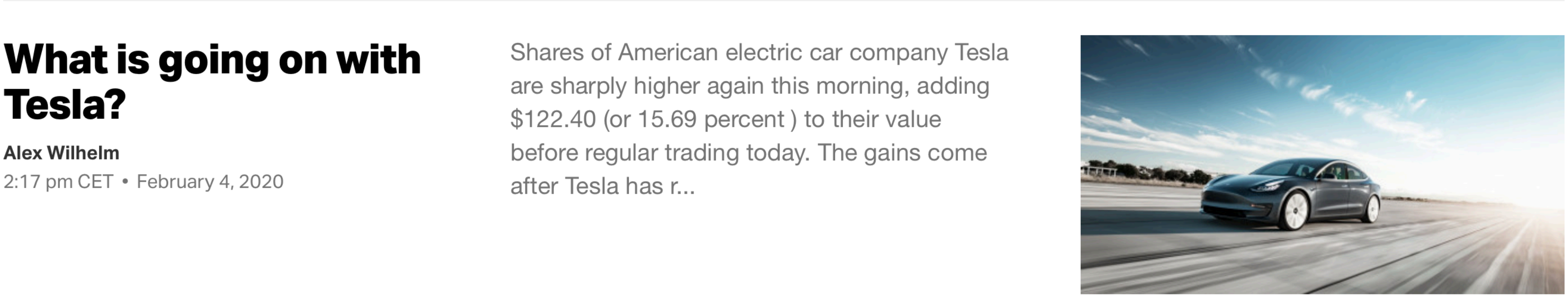
Attribute names and values are separated by an equals sign (=)
Attribute name
Value
Attribute name
Value
Attribute
Attribute
Multiple attributes are separated by a space

Figure 4-11. An element with attributes.

http://web.simmons.edu/~grovesd/comm244/notes/week2/html-attributes

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>
```



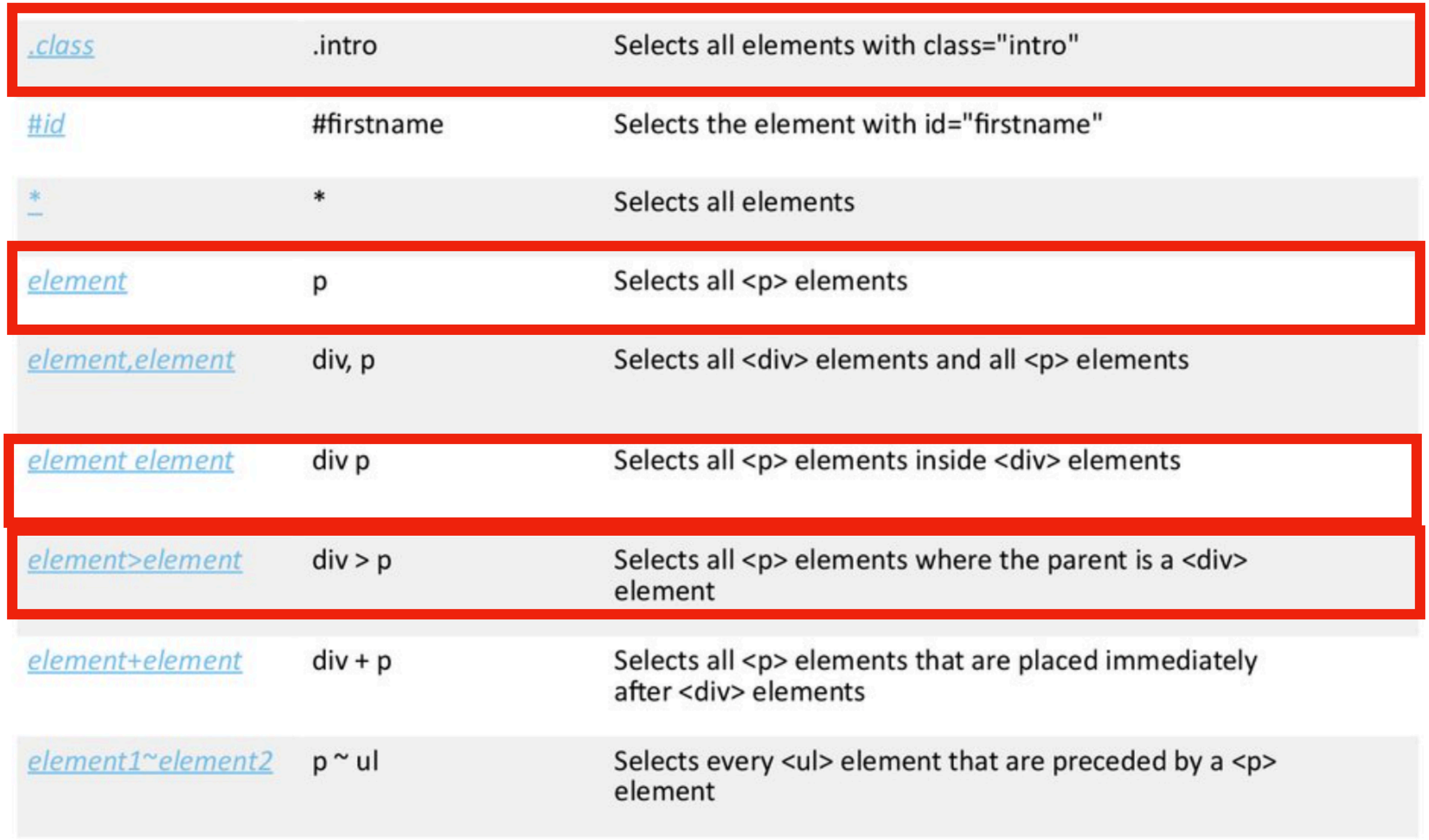
What is going on with Tesla?

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



.class	.intro	Selects all elements with class="intro"
#id	#firstname	Selects the element with id="firstname"
*	*	Selects all elements
element	p	Selects all <p> elements
element,element	div, p	Selects all <div> elements and all <p> elements
element element	div p	Selects all <p> elements inside <div> elements
element>element	div > p	Selects all <p> elements where the parent is a <div> element
element+element	div + p	Selects all <p> elements that are placed immediately after <div> elements
element1~element2	p ~ ul	Selects every <ul> element that are preceded by a <p> element

http://ppt-online.org/49955

Session

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

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

Interactive