

XPath Navigation

WEB SCRAPING IN PYTHON

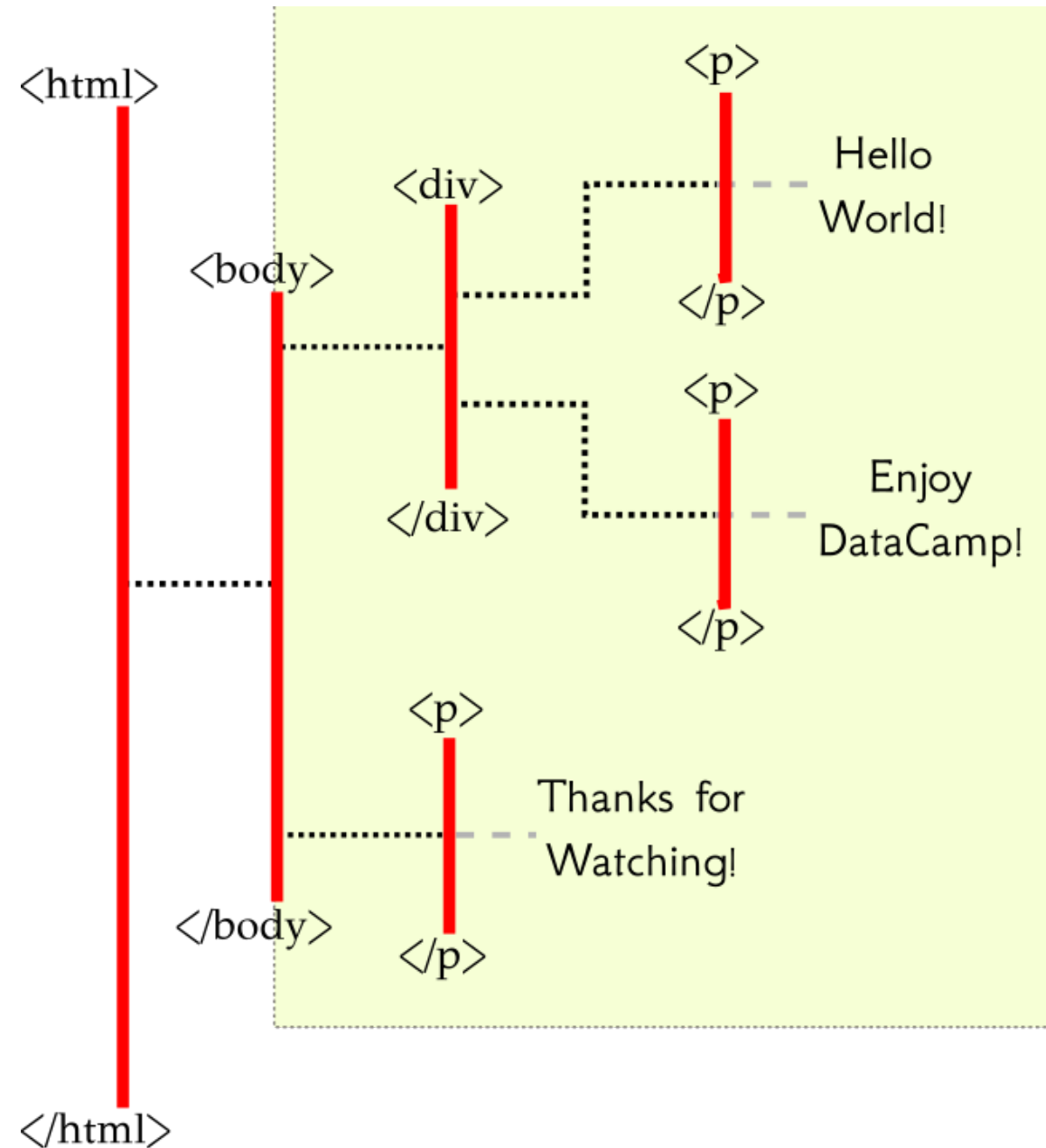


Thomas Laetsch
Data Scientist, NYU

Slashes and Brackets

- Single forward slash / looks forward **one** generation
- Double forward slash // looks forward **all** future generations
- Square brackets [] help narrow in on specific elements

To Bracket or not to Bracket



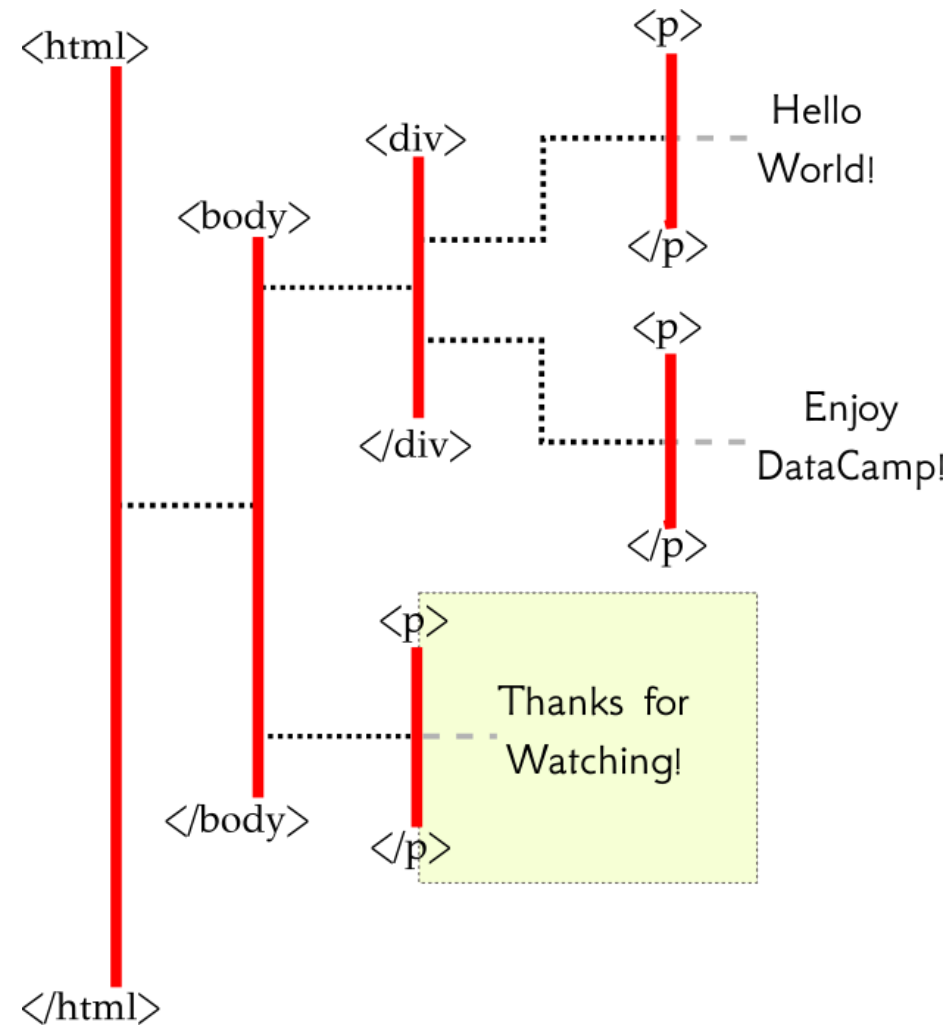
```
xpath = '/html/body'
```

```
xpath = '/html[1]/body[1]'
```

- Give the same selection

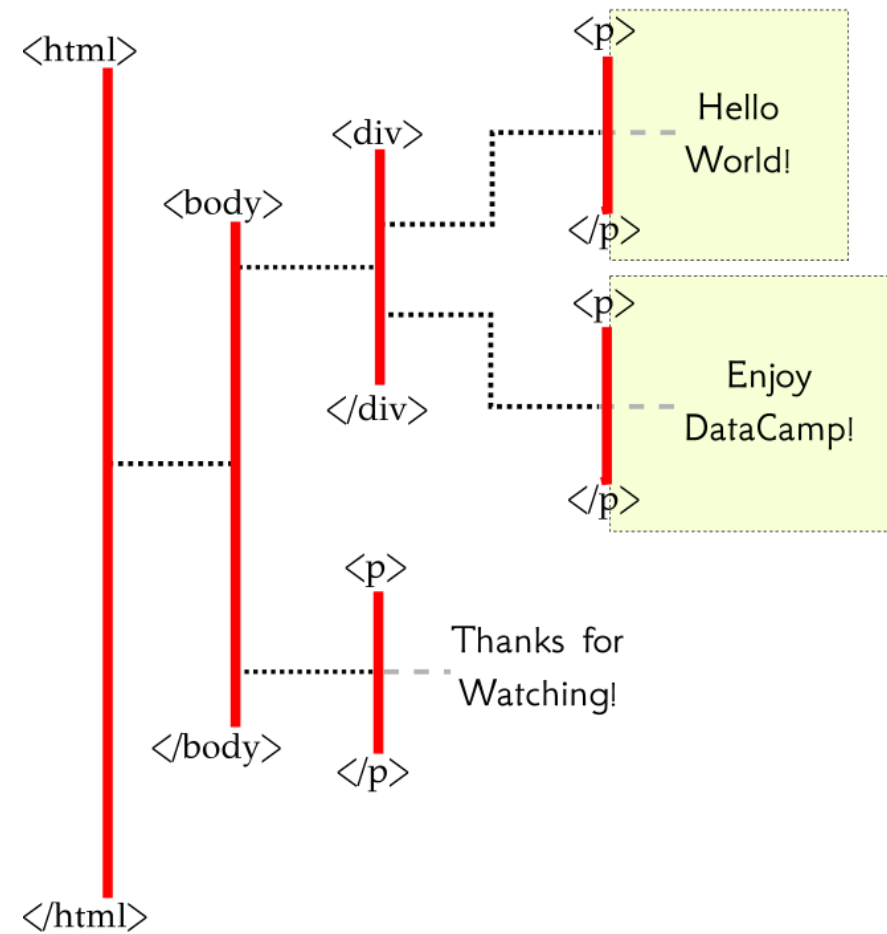
A Body of P

```
xpath = '/html/body/p'
```

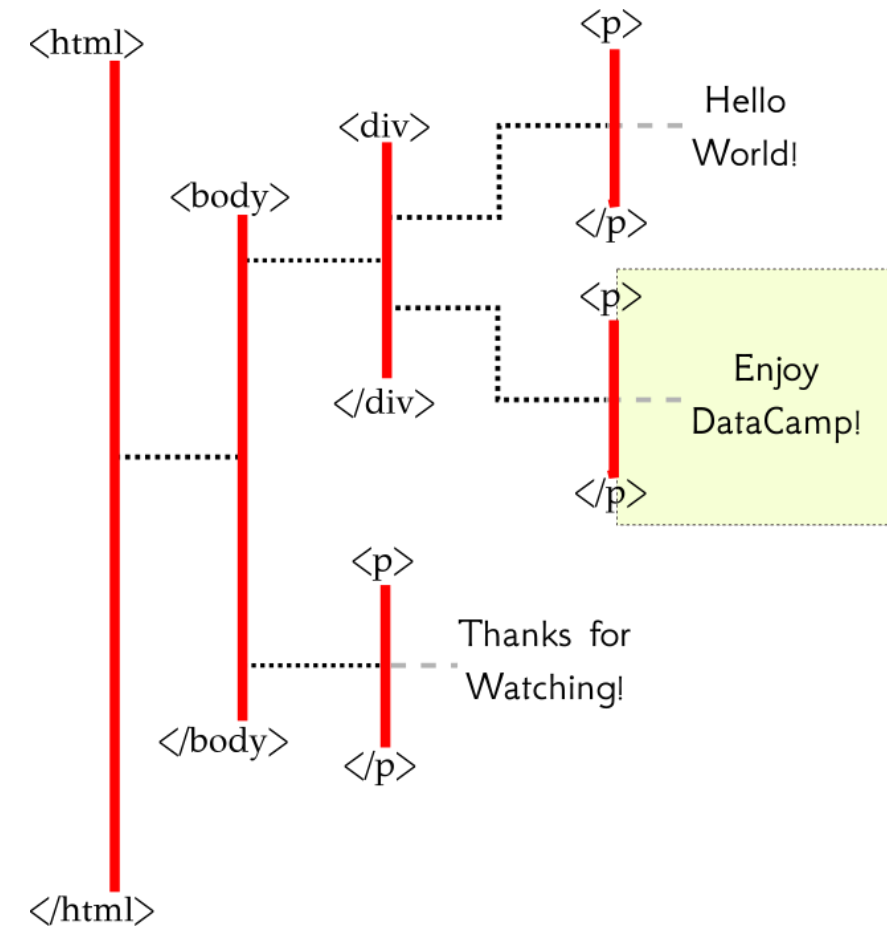


The Birds and the Ps

```
xpath = '/html/body/div/p'
```

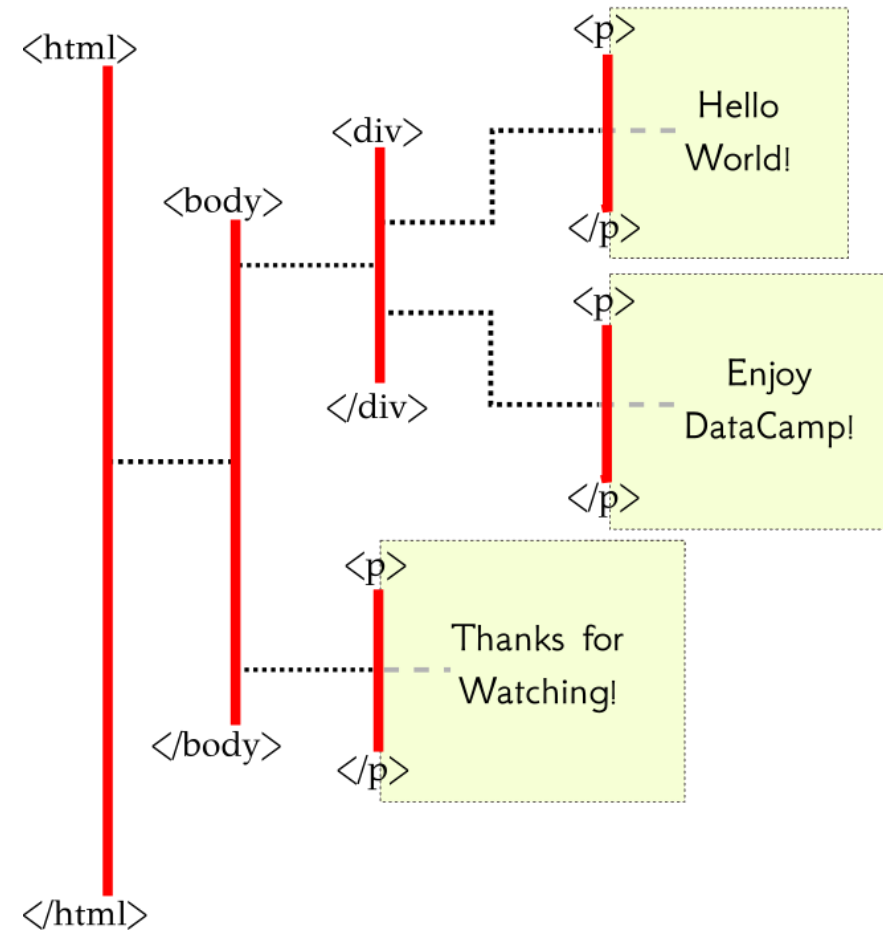


```
xpath = '/html/body/div/p[2]'
```

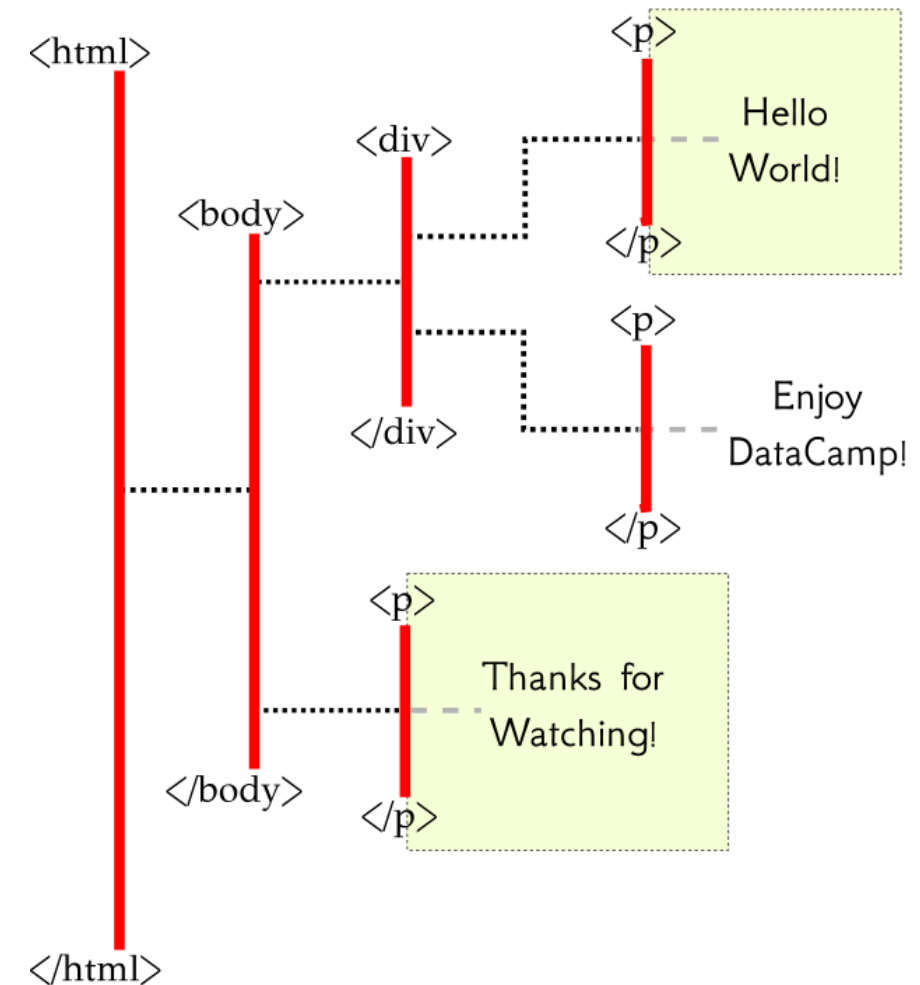


Double Slashing the Brackets

```
xpath = '//p'
```



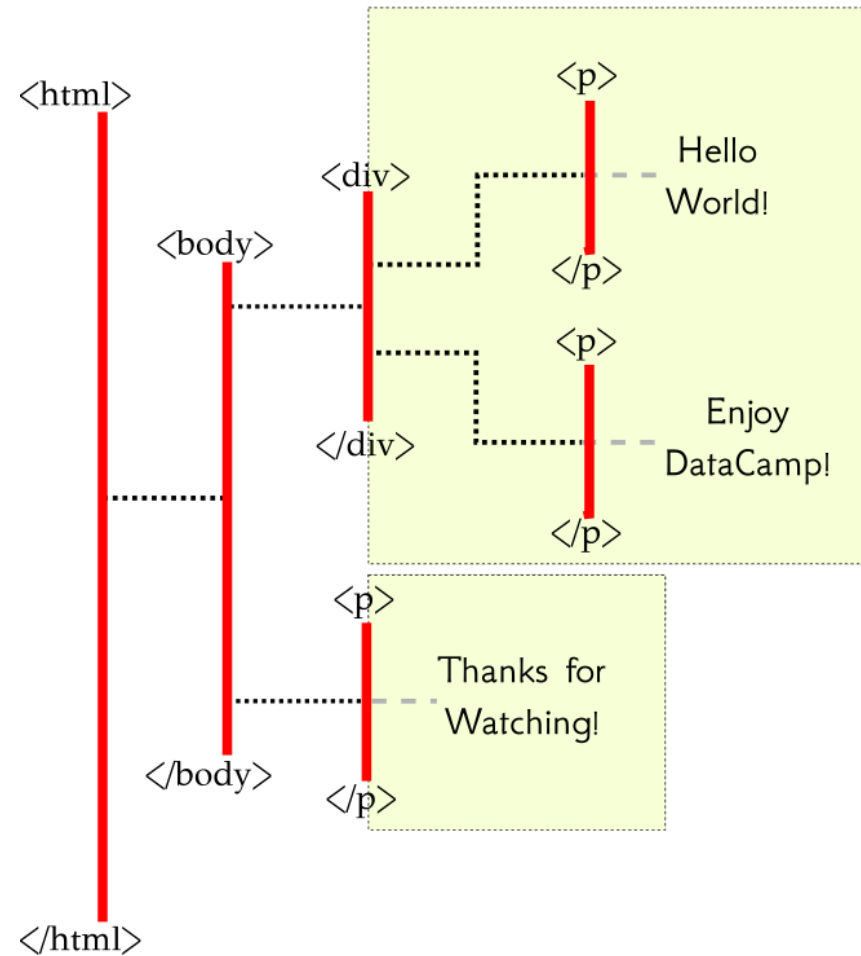
```
xpath = '//p[1]'
```



The Wildcard

```
xpath = '/html/body/*'
```

- The asterisks * is the "wildcard"



Xposé

WEB SCRAPING IN PYTHON

Off the Beaten XPath

WEB SCRAPING IN PYTHON

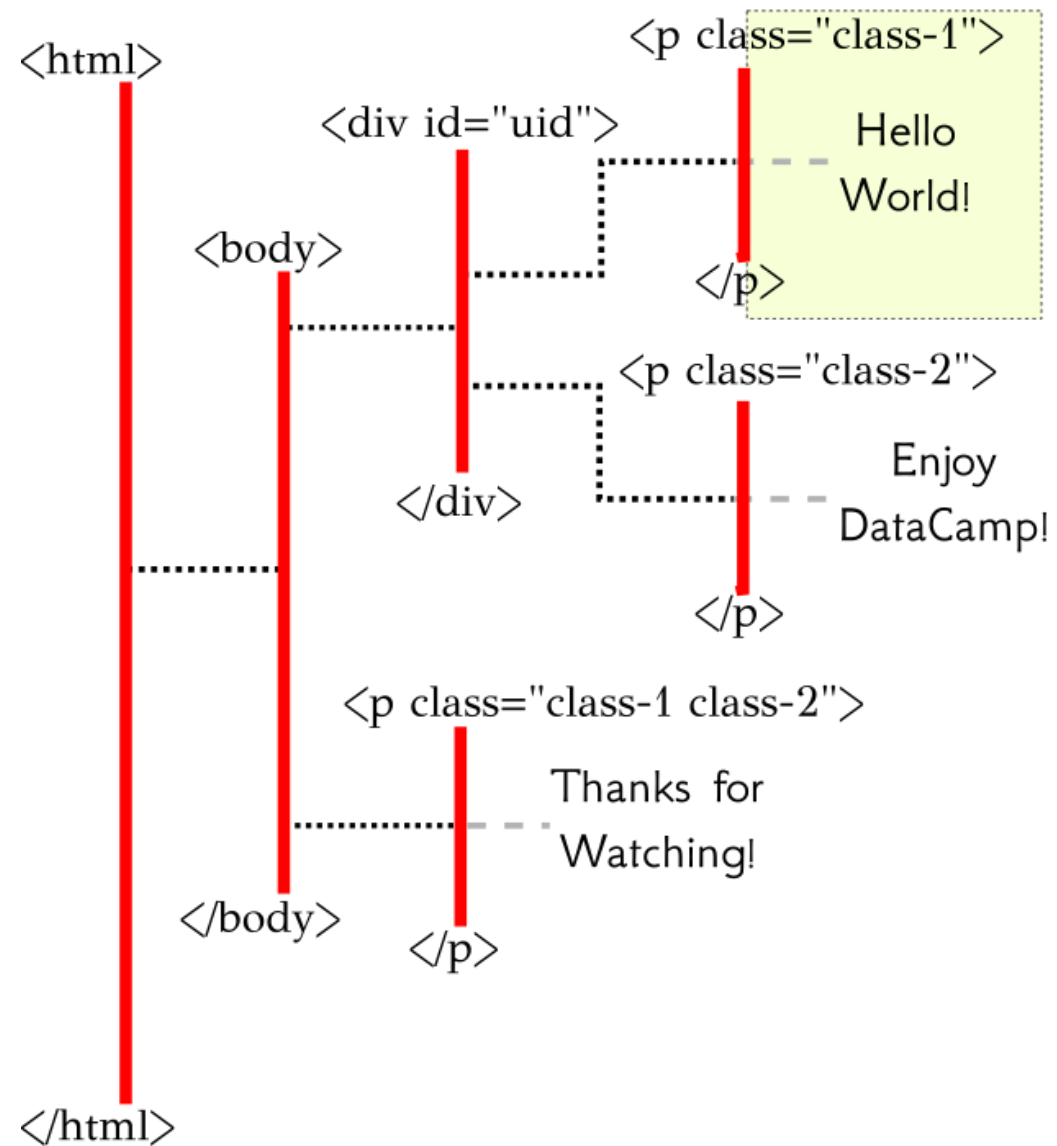


Thomas Laetsch
Data Scientist, NYU

(At)tribute

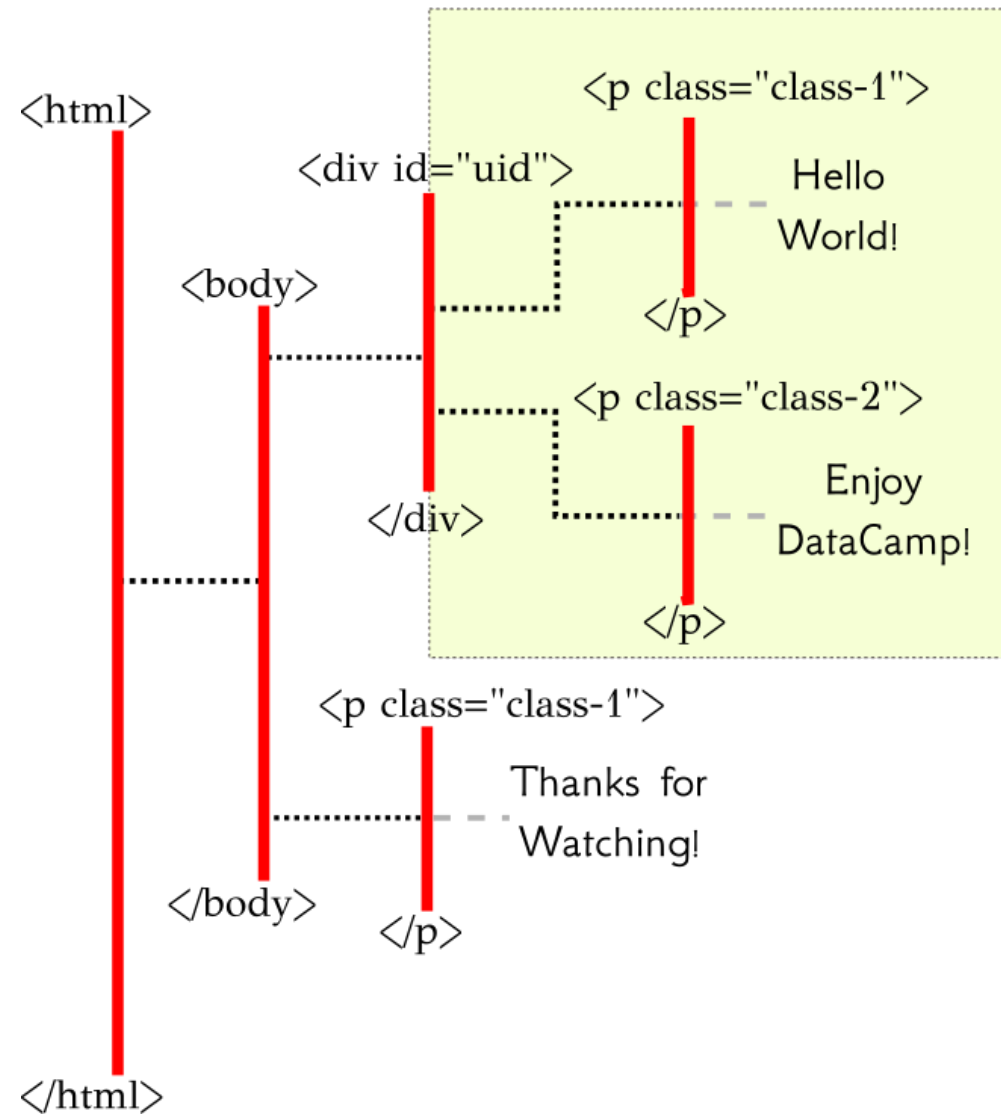
- @ represents "attribute"
 - @class
 - @id
 - @href

Brackets and Attributes



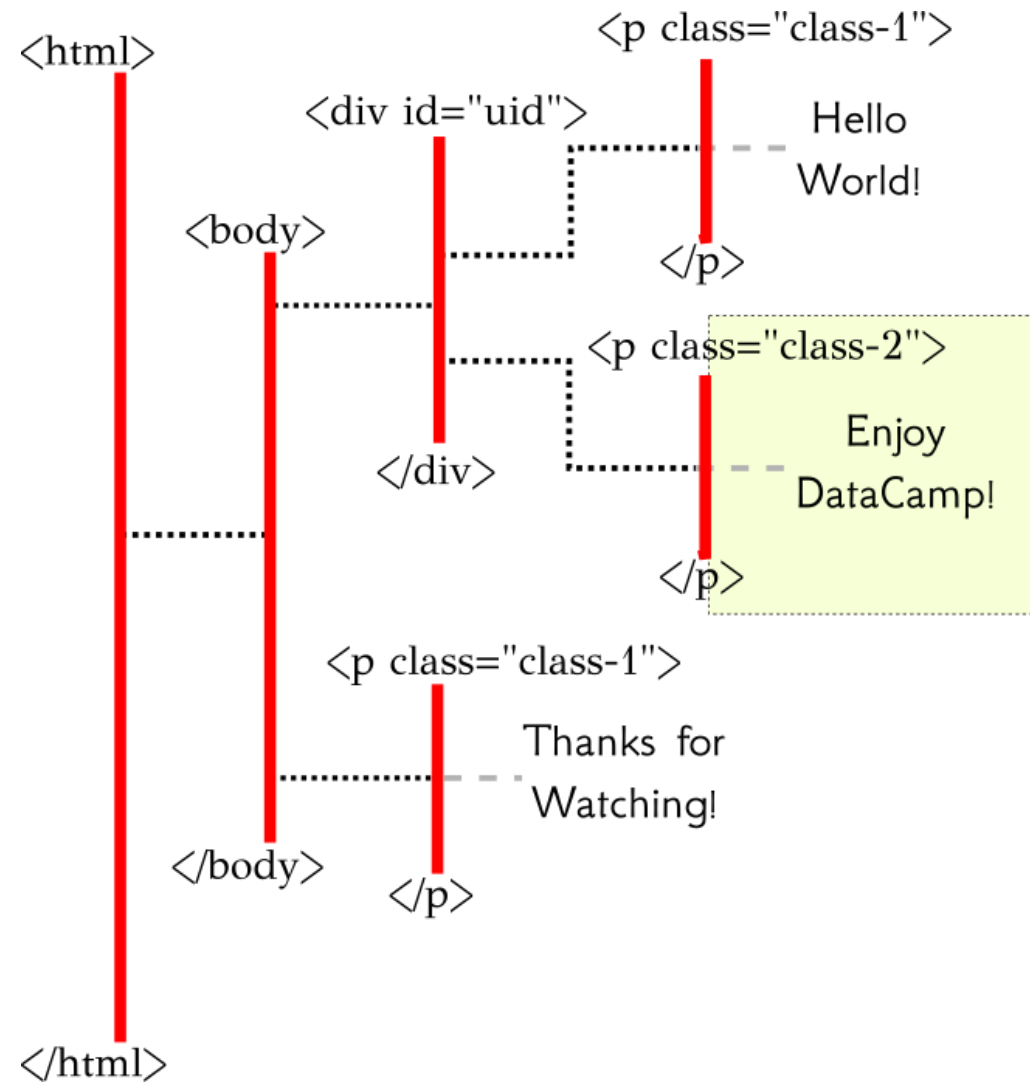
```
xpath = '//p[@class="class-1"]'
```

Brackets and Attributes



```
xpath = '//*[@id="uid"]'
```

Brackets and Attributes



```
xpath = '//div[@id="uid"]/p[2]'
```

Content with Contains

Xpath Contains Notation:

```
contains( @attri-name, "string-expr" )
```

Contain This

```
xpath = '//*[@contains(@class,"class-1")]'
```

☒ `<p class="class-1"> ... </p>`

☒ `<div class="class-1 class-2"> ... </div>`

☒ `<p class="class-1 2"> ... </p>`

Contain This

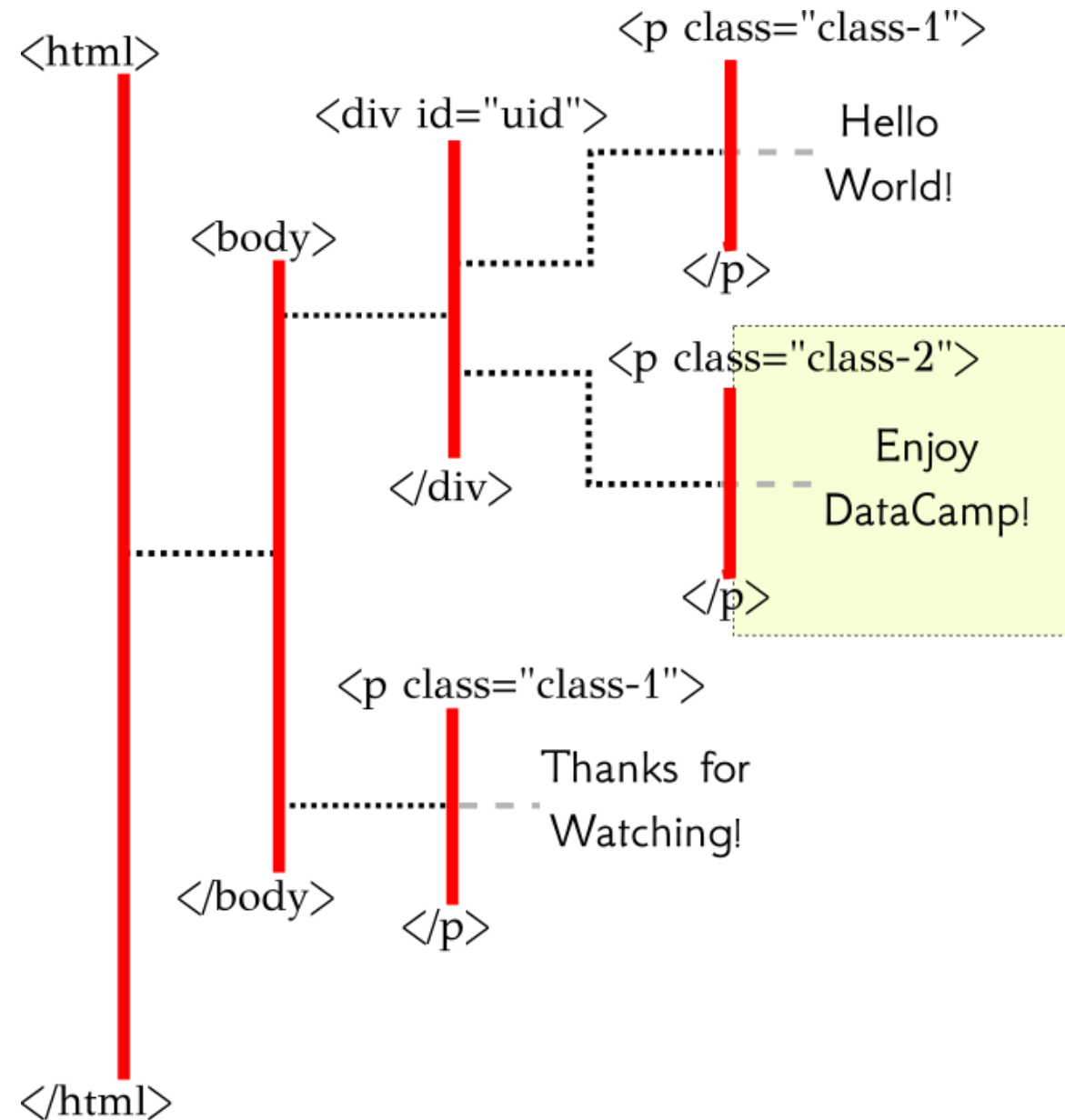
```
xpath = '//*[@class="class-1"]'
```

 `<p class="class-1"> ... </p>`

 `<div class="class-1 class-2"> ... </div>`

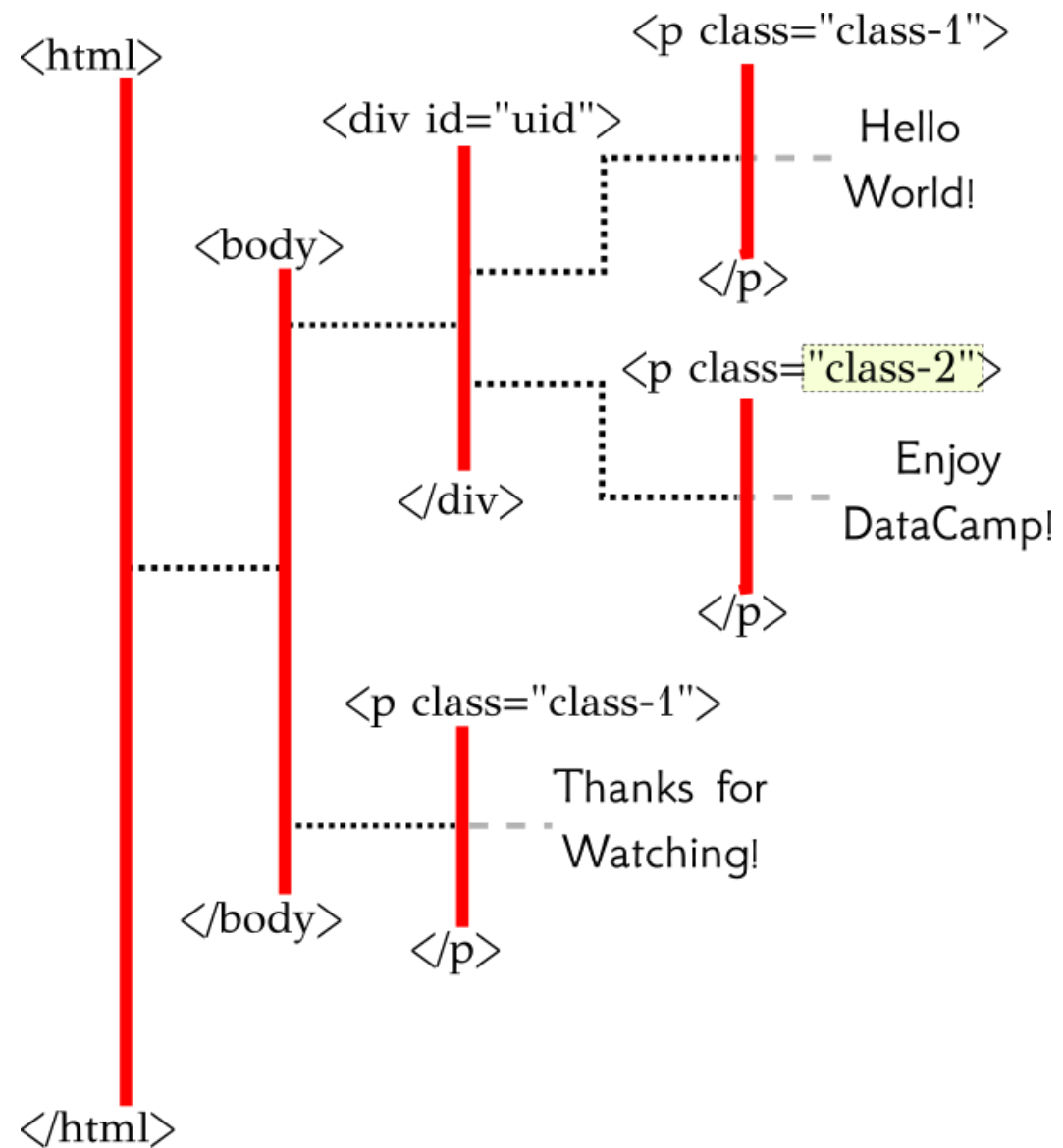
 `<p class="class-1 2"> ... </p>`

Get Classy



```
xpath = '/html/body/div/p[2]'
```

Get Classy



```
xpath = '/html/body/div/p[2]/@class'
```

End of the Path

WEB SCRAPING IN PYTHON

Introduction to the scrapy Selector

WEB SCRAPING IN PYTHON



Thomas Laetsch
Data Scientist, NYU

Setting up a Selector

```
from scrapy import Selector
```

```
html = '''  
<html>  
  <body>  
    <div class="hello datacamp">  
      <p>Hello World!</p>  
    </div>  
    <p>Enjoy DataCamp!</p>  
  </body>  
</html>  
'''
```

```
sel = Selector( text = html )
```

- Created a scrapy Selector object using a string with the html code
- The selector `sel` has selected the **entire** html document

Selecting Selectors

- We can use the `xpath` call within a `Selector` to create new `Selector`s of specific pieces of the html code
- The return is a `SelectorList` of `Selector` objects

```
sel.xpath("//p")  
# outputs the SelectorList:  
[<Selector xpath='//p' data='<p>Hello World!</p>'>,  
 <Selector xpath='//p' data='<p>Enjoy DataCamp!</p>'>]
```

Extracting Data from a SelectorList

- Use the `extract()` method

```
>>> sel.xpath("//p")
out: [<Selector xpath='//p' data='<p>Hello World!</p>'>,
      <Selector xpath='//p' data='<p>Enjoy DataCamp!</p>'>]
```

```
>>> sel.xpath("//p").extract()
out: [ '<p>Hello World!</p>',
      '<p>Enjoy DataCamp!</p>' ]
```

- We can use `extract_first()` to get the first element of the list

```
>>> sel.xpath("//p").extract_first()
out: '<p>Hello World!</p>'
```


Extracting Data from a Selector

```
ps = sel.xpath('//p')  
second_p = ps[1]
```

```
second_p.extract()  
out: '<p>Enjoy DataCamp!</p>'
```

Select This Course!

WEB SCRAPING IN PYTHON

"Inspecting the HTML"

WEB SCRAPING IN PYTHON



Thomas Laetsch, PhD
Data Scientist, NYU

"Source" = HTML Code

Learn

Acquire new skills fast in courses that combine short expert videos with immediate hands-on-keyboard exercises.

What do you want to learn?

All Technologies ▾ All Topics ▾

Introduction to R

Master the basics of data analysis by manipulating common data structures such as vectors, matrices and data frames.

Continue Course

Intro to Statistics with R: Correlation and Linear R...

If you have ever taken a math or statistics class, you've probably heard the old adage "Correlation does not imply ca..."


Continue Course

Intro to Statistics with R: Multiple Regression

Multiple regression is a powerful statistical technique, and here you will discover why and how to use it. Part of th...

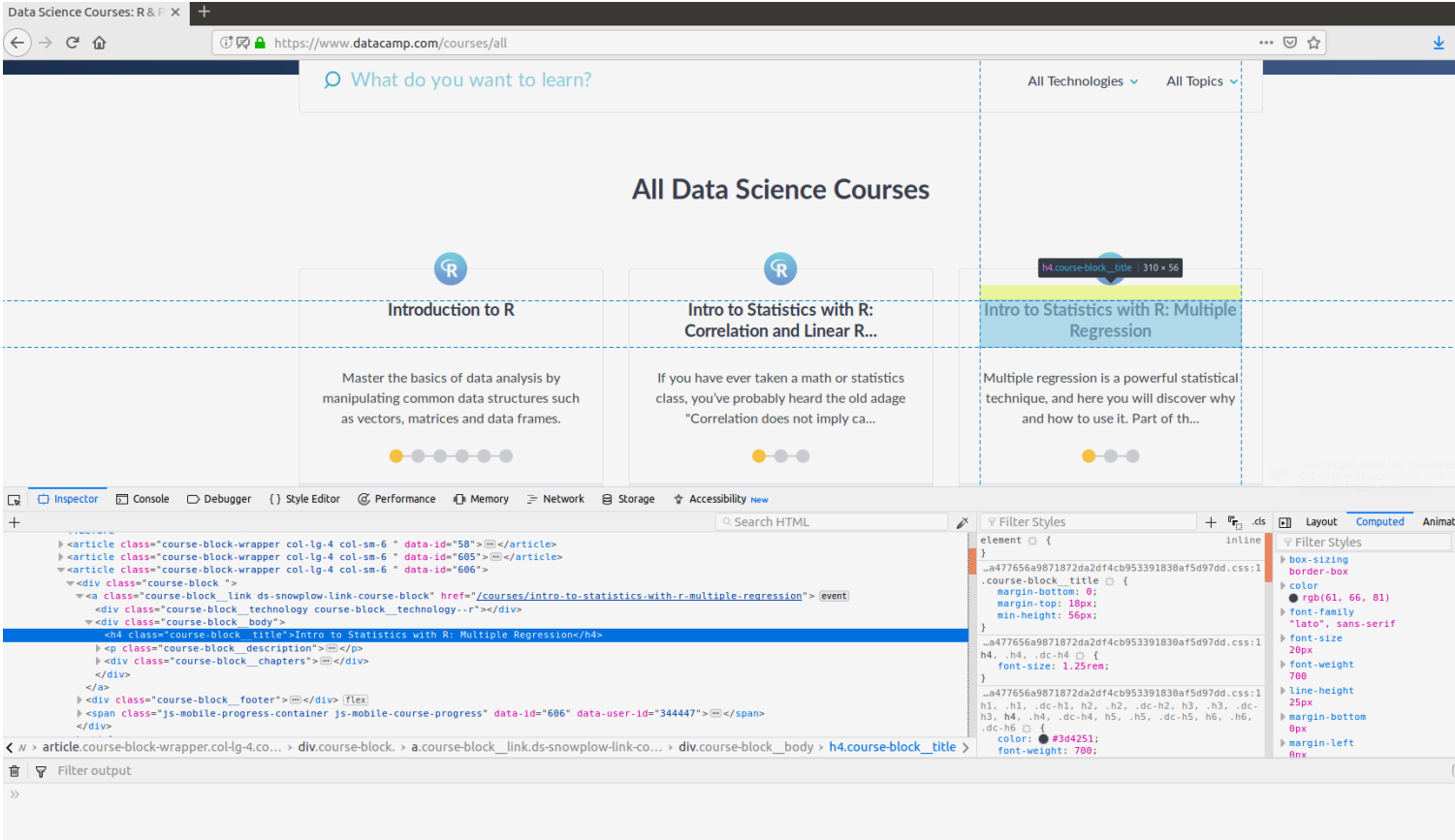
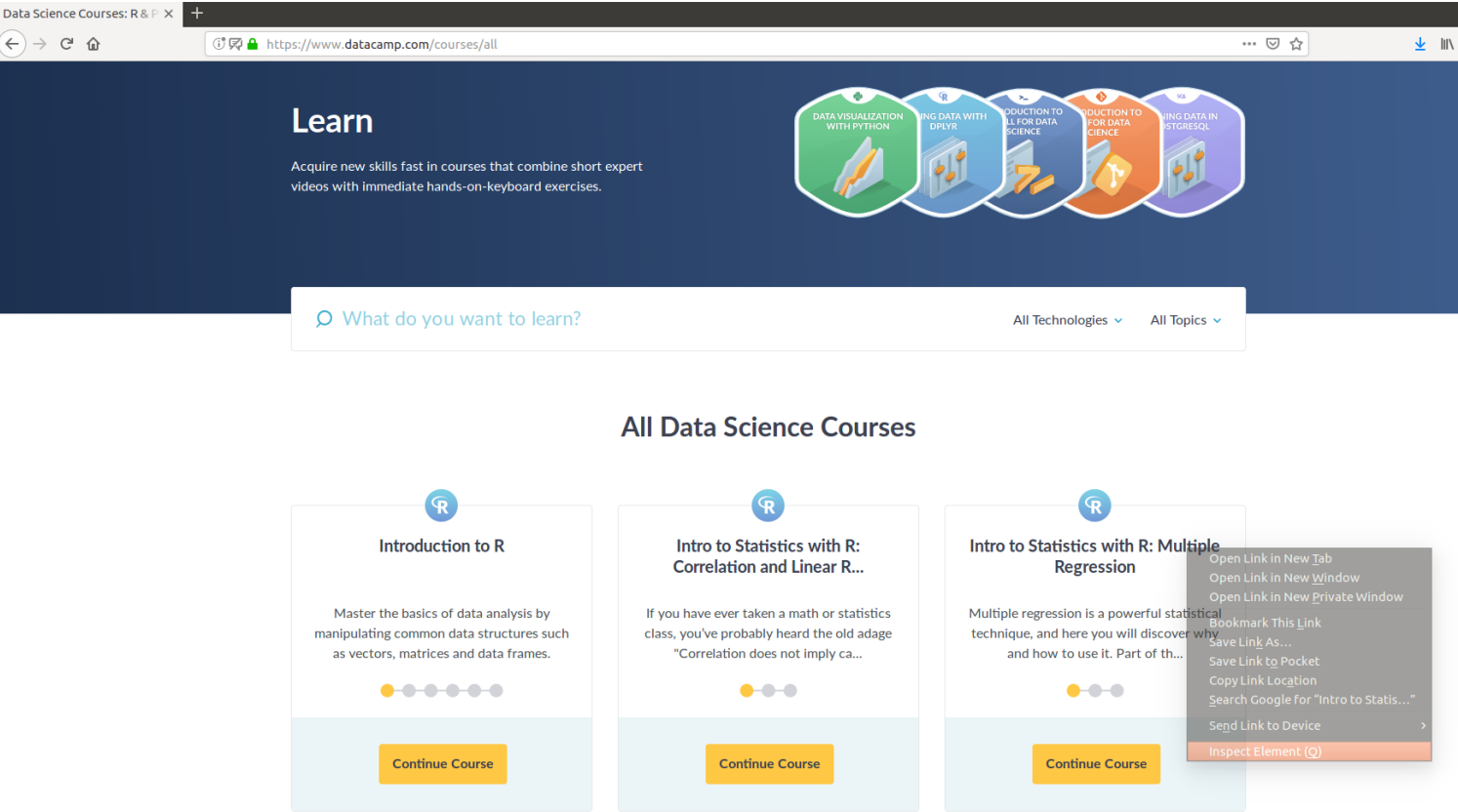
Continue Course

```
1 <!DOCTYPE html>
2 <html class="no-js">
3 <head>
4 <script>
5 (function(w,d,s,l,i){w[l]=w[l]||[];w[l].push({'gtm.start':
6 new Date().getTime(),event:'gtm.js'});var f=d.getElementsByTagName(s)[0],
7 j=d.createElement(s),dl=l!='dataLayer'?&l='+l':'';j.async=true;j.src=
8 'https://www.googletagmanager.com/gtm.js?id='+i+dl;f.parentNode.insertBefore(j,f);
9 })(window,document,'script','dataLayer','GTM-TGQWB2P');
10
11
12
13 <meta charset="utf-8">
14 <script>window.NREUM||(NREUM={});NREUM.info={"beacon":"bam.nr-data.net","errorBeacon":"bam.nr-data.net","licenseKey":"4795905ee2","applicationID":"90826195","transactionName":"3lkNEEQLVAD0E0wF0BBEEAFFS1kNCg==","queueTime":0,"appl
15 <script>window.NREUM||(NREUM={});_nr.require=function(e,n,t){function r(t){if(!n(t)){var o=n(t)=[{exports:{}}];e[t][0].call(o.exports,function(n){var o=e[t][1][n];return r(o)[n]},o,o.exports)}return n(t).exports}if("function"==ty
16 <title>Data Science Courses: R & Python Analysis Tutorials | DataCamp</title>
17 <meta name="description" content="DataCamp offers a variety of online courses & video tutorials to help you learn data science at your own pace. See why over 3,220,000 people use DataCamp now!">
18 <link rel="canonical" href="https://www.datacamp.com/courses/all">
19 <meta name="twitter:title" content="Data Science Courses: R & Python Analysis Tutorials">
20 <meta name="twitter:description" content="DataCamp offers a variety of online courses & video tutorials to help you learn data science at your own pace. See why over 3,220,000 people use DataCamp now!">
21 <meta name="twitter:card" content="summary">
22 <meta name="twitter:site" content="@DataCamp">
23 <meta name="twitter:image" content="https://www.datacamp.com/datacamp-sq.png">
24 <meta name="twitter:image:width" content="300">
25 <meta name="twitter:image:height" content="300">
26 <meta name="twitter:creator" content="@DataCamp">
27 <meta name="twitter:domain" content="www.datacamp.com">
28 <meta property="og:image" content="https://www.datacamp.com/datacamp.png">
29 <meta property="og:image:width" content="1200">
30 <meta property="og:image:height" content="630">
31 <meta property="og:title" content="Data Science Courses: R & Python Analysis Tutorials">
32 <meta name="author" content="https://plus.google.com/u/0/+DataCamp/">
33 <link rel="shortcut icon" type="image/x-icon" href="https://cdn.datacamp.com/main-app/assets/favicon-335cd8394b32102a39221d79e5fd7e51078e6d32a8c8aea59676a6869f84e9d8.ico" />
34 <meta name="csrf-param" content="authenticity token" />
35 <meta name="csrf-token" content="+Z785sRb47v0ePLJwzoqP0m+h653qkvUcP9xLxEL3u5qAT8CguZ1VmeY9+dsxDabTLHdMZKcv/Pp7S1Mn/RJA==" />
36 <link rel="manifest" href="/manifest.json">
37
38 <meta name="viewport" content="width=device-width, initial-scale=1, maximum-scale=1">
39 <meta name="fragment" content="!">
40 <meta name="google-site-verification" content="ao3s4PdjiSD20sfTbldo7YJx7VX2QLkPETldpyfTj08" />
41 <meta name="apple-itunes-app" content="app-id=1263413087">
42
43 <link rel="stylesheet" media="all" href="https://cdn.datacamp.com/main-app/assets/application_v2-566f39ae10ae5051b6d46fe6a477656a9871872da2df4cb953391830af5d97dd.css" />
44
45 <script>
46 (function(h,o,t,j,a,r){
47 h.hj=h.hj||function(){(h.hj.q=h.hj.q||[]).push(arguments)};
48 h._hjSettings={hjid:484663,hjsv:6};
49 a=o.getElementsByTagName('head')[0];
50 r=o.createElement('script');r.async=1;
51 r.src=t+h._hjSettings.hjid+j+h._hjSettings.hjsv;
52 a.appendChild(r);
53 })(window,document,'https://static.hotjar.com/c/hotjar-','.js?sv=');
54 </script>
55
56 </head>
57 <body class="js-application-v2" data-env="production"><noscript><iframe src="https://www.googletagmanager.com/ns.html?id=GTM-TGQWB2P"
58 height="0" width="0" style="display:none;visibility:hidden"></iframe></noscript>
59
60
61 <div class="site-wrap">
62 <div class="dc-flash-wrapper" id="flash_messages">
63
64 </div>
65
```

 datacamp

WEB SCRAPING IN PYTHON

Inspecting Elements



HTML text to Selector

```
from scrapy import Selector
```

```
import requests  
url = 'https://www.datacamp.com/courses/all'  
html = requests.get( url ).content
```

```
sel = Selector( text = html )
```

You Know Our Secrets

WEB SCRAPING IN PYTHON