XPath Navigation

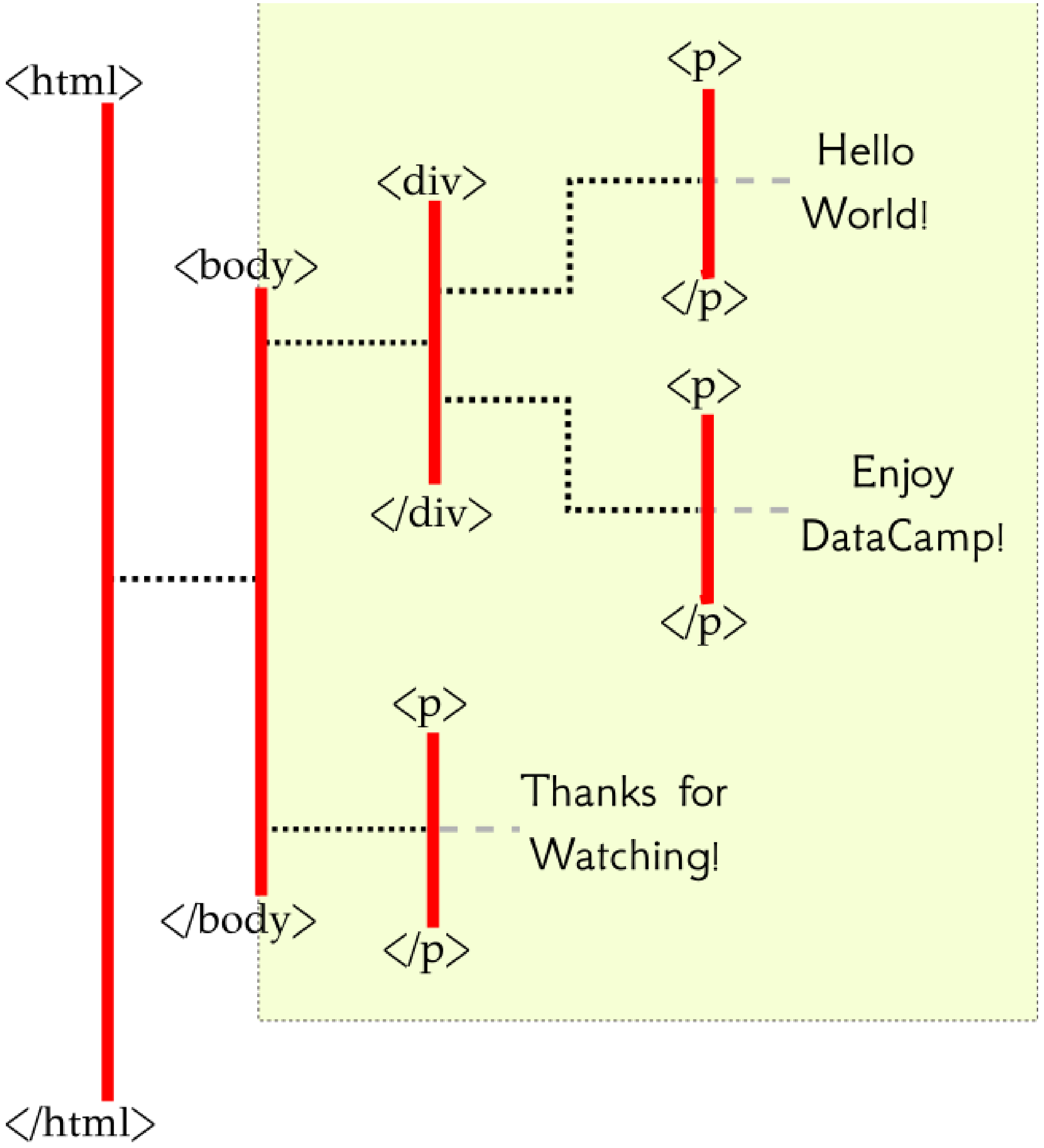
WEB SCRAPING IN PYTHON

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



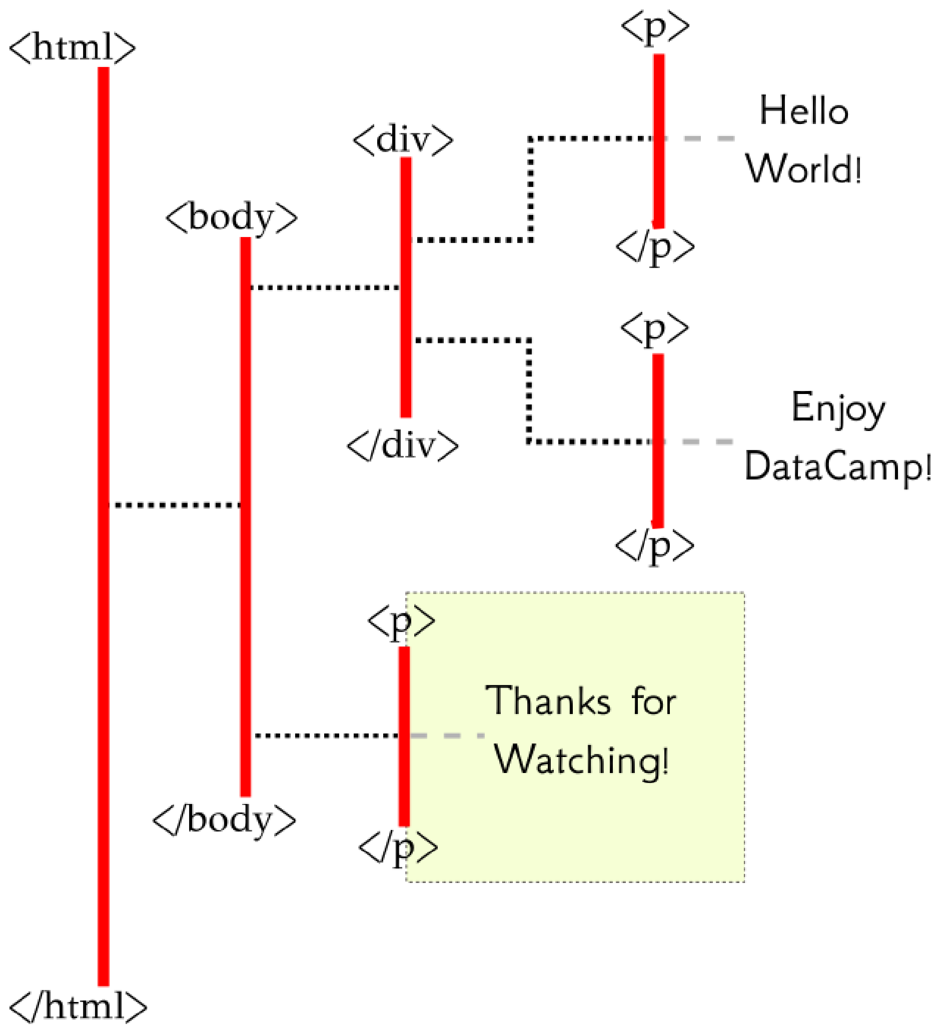
```
xpath = '/html/body'
```

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

- Give the same selection

A Body of P

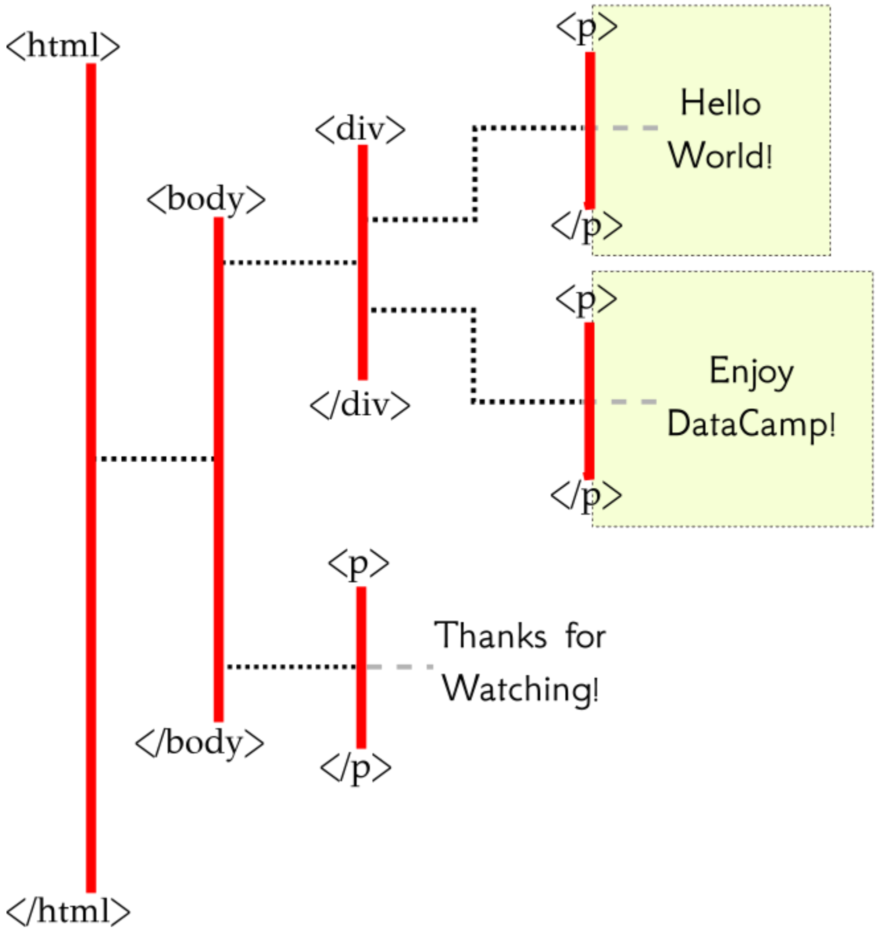
```
xpath = '/html/body/p'
```

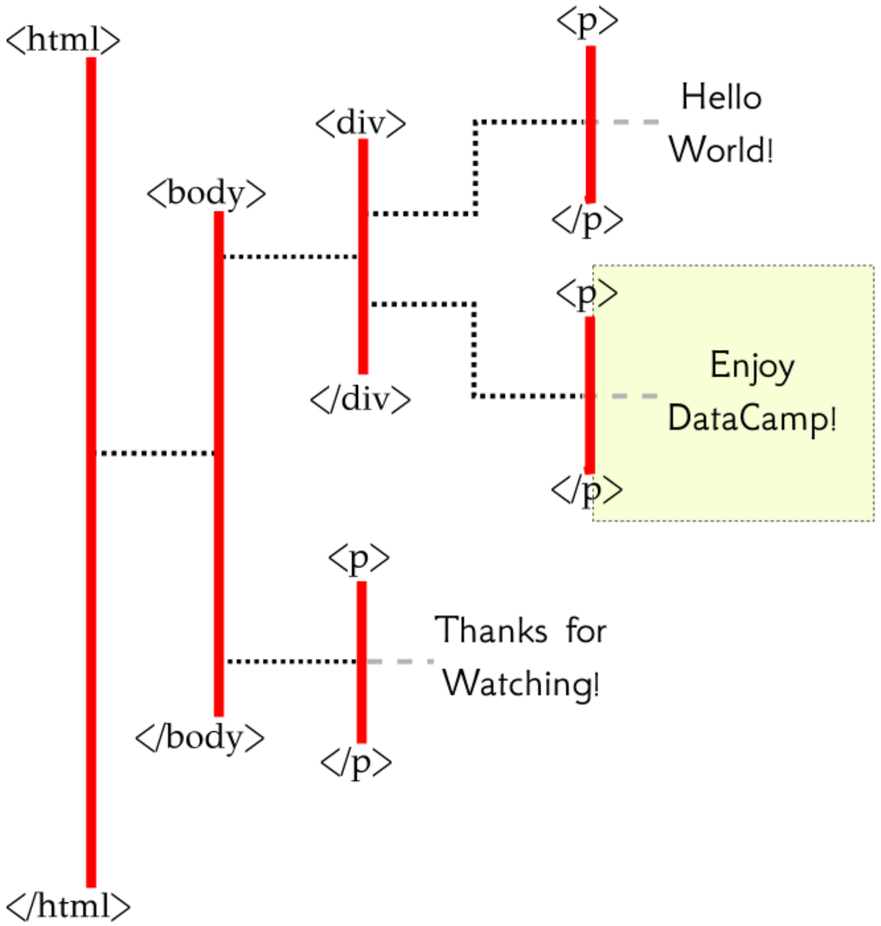


The Birds and the Ps

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

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

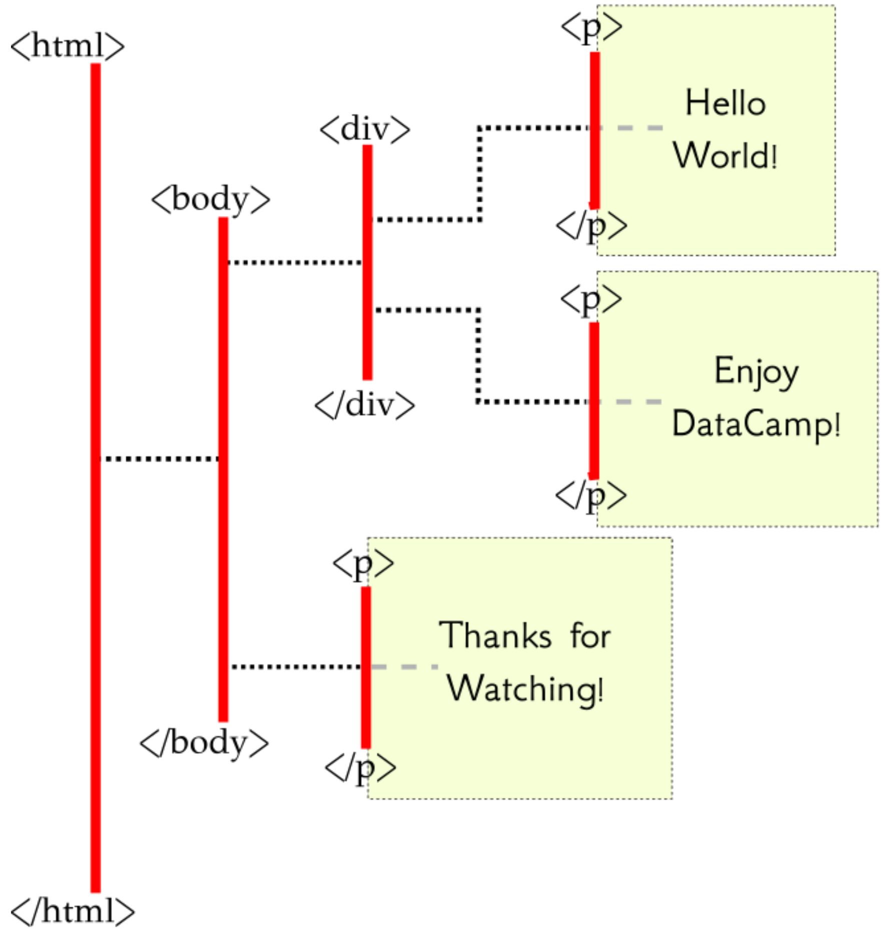


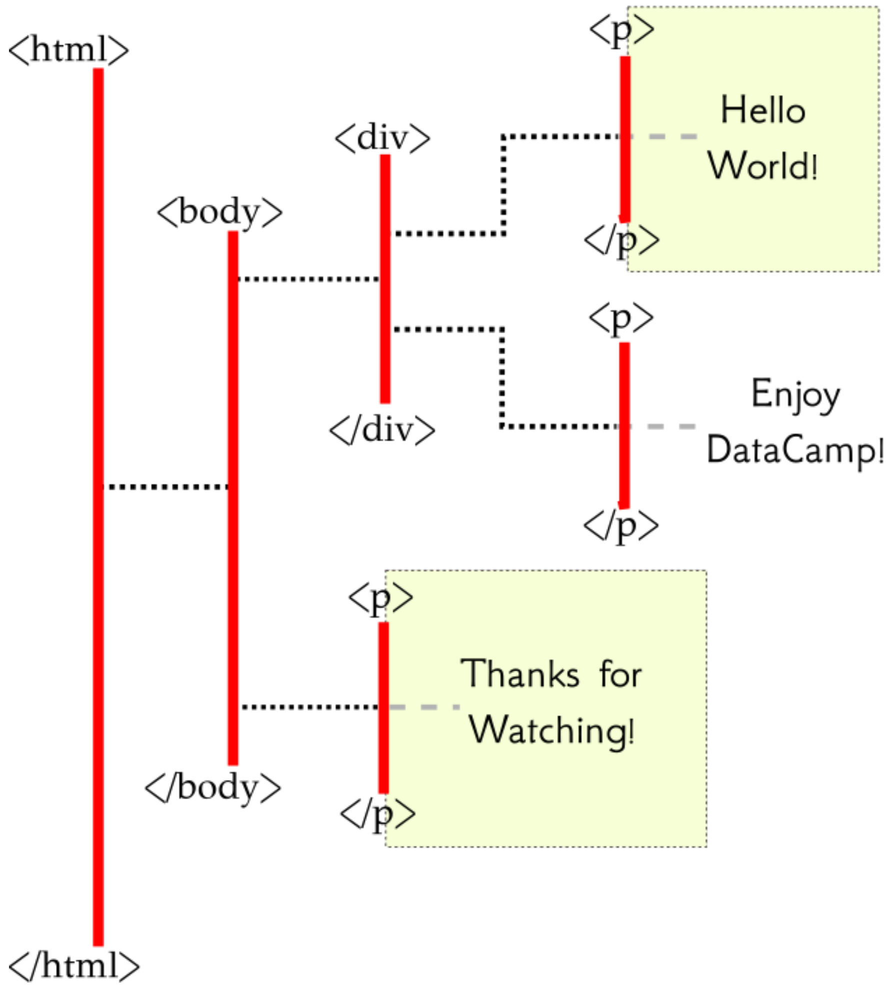


Double Slashing the Brackets

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

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

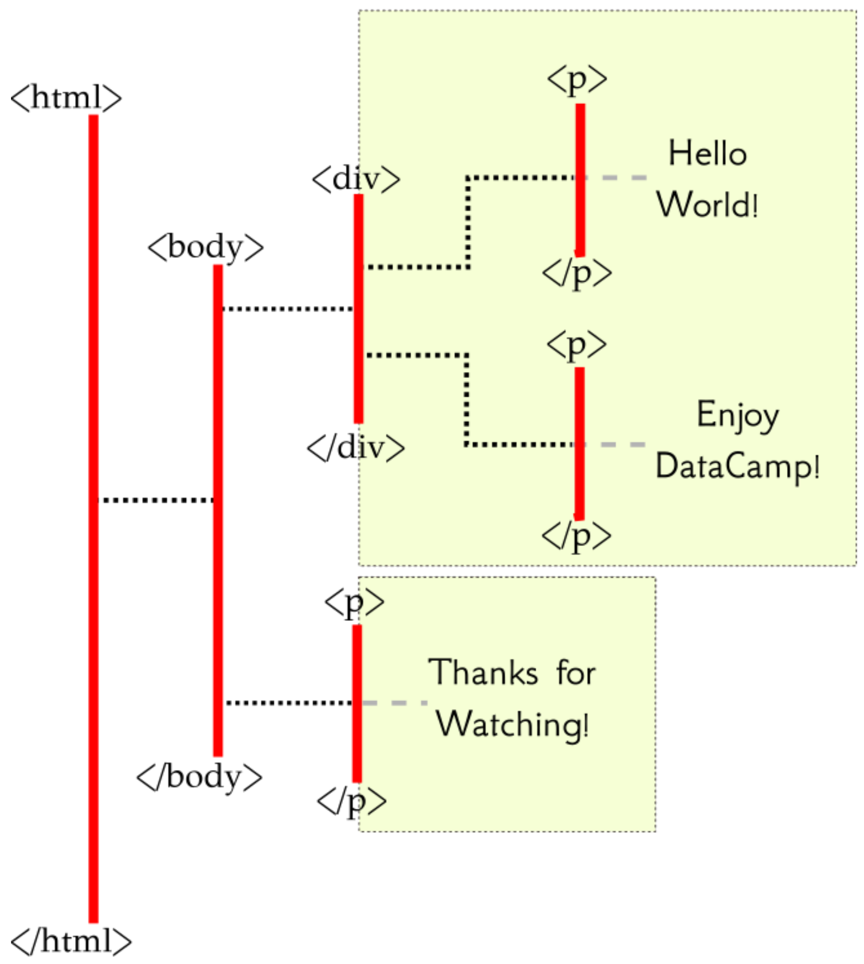




The Wildcard

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

- The asterisks * is the "wildcard"



Xposé

WEB SCRAPING IN PYTHON

Off the Beaten XPath

WEB SCRAPING IN PYTHON

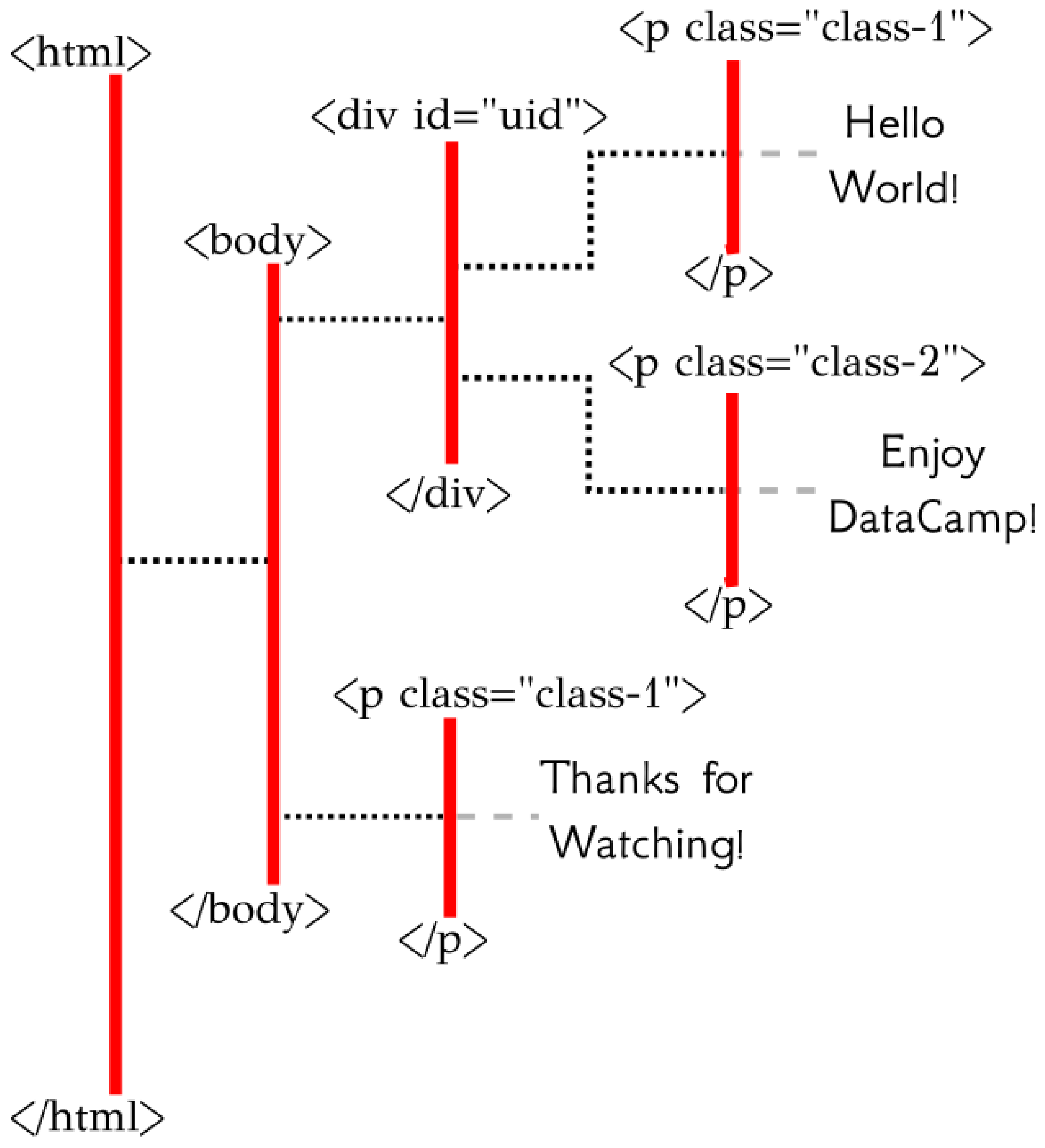
Thomas Laetsch
Data Scientist, NYU

datacamp

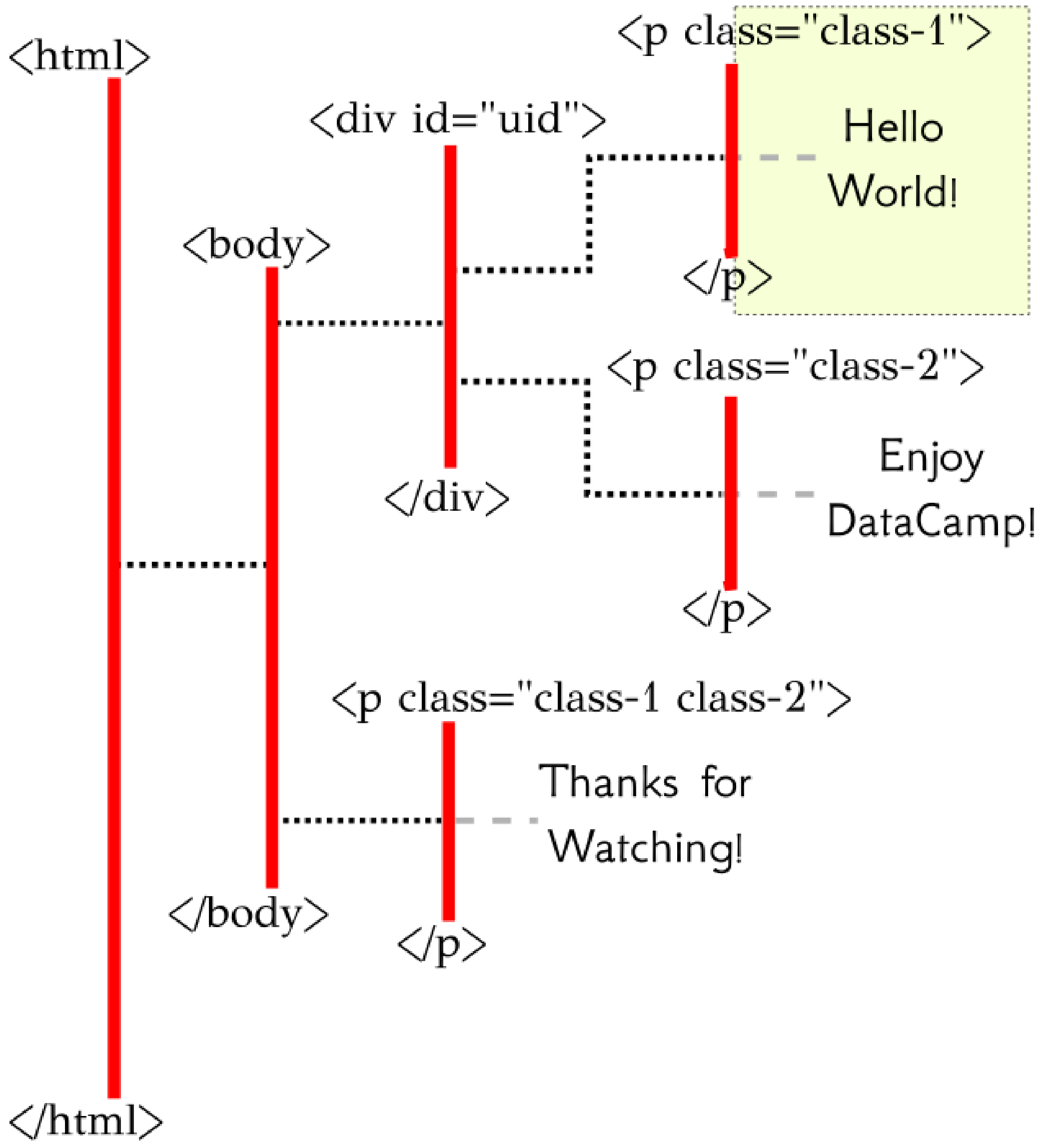
(At)tribute

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

Brackets and Attributes

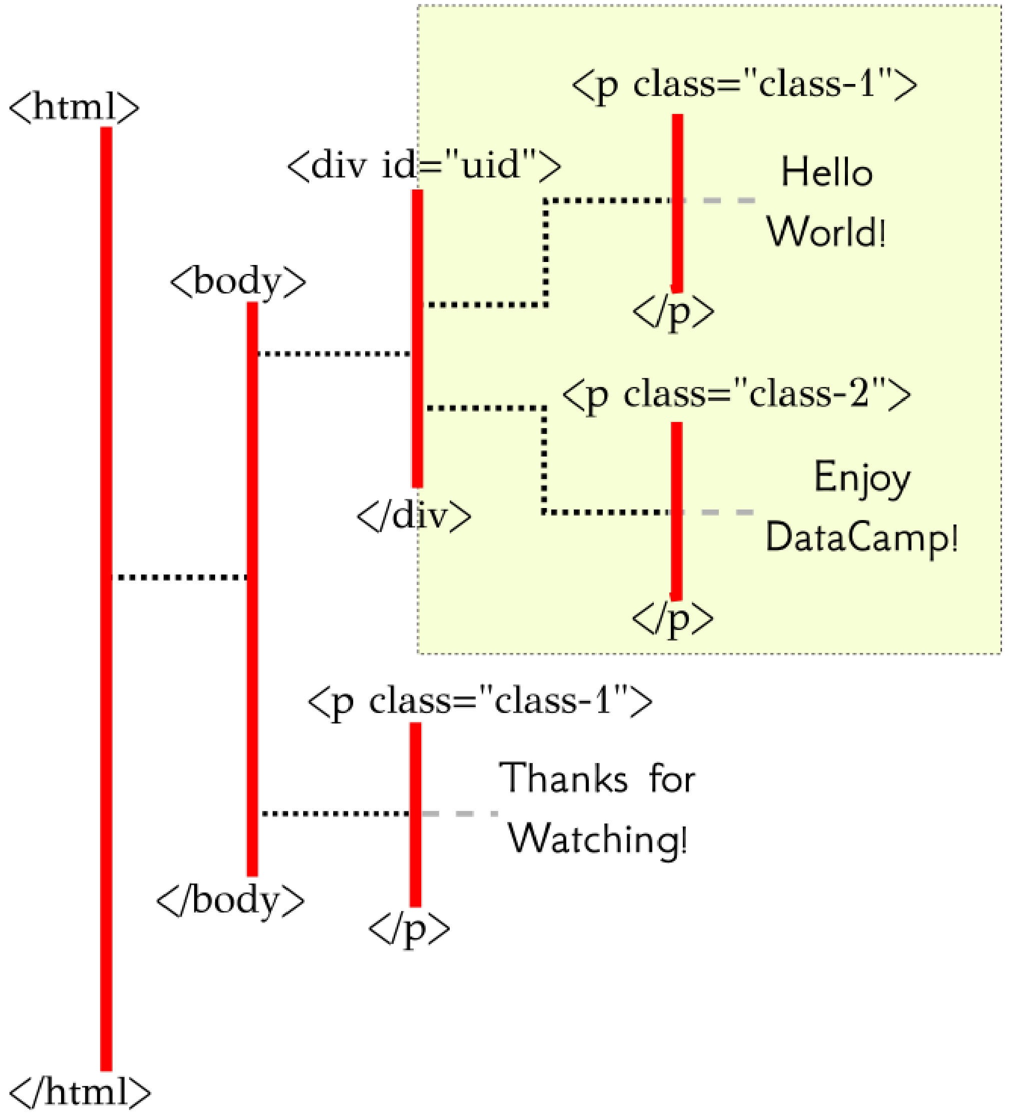


Brackets and Attributes



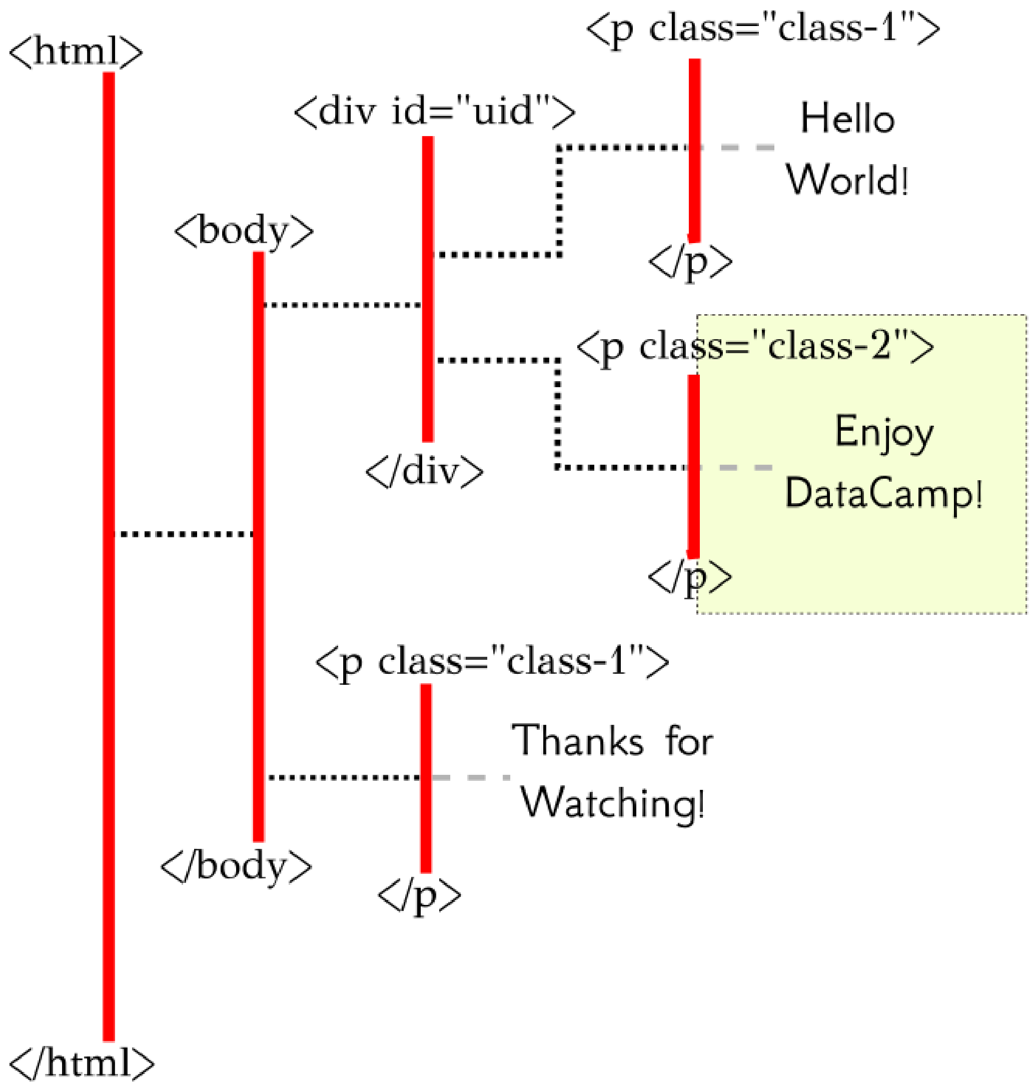
```
xpath = '//p[@class="class-1"]'
```

Brackets and Attributes



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

Brackets and Attributes



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

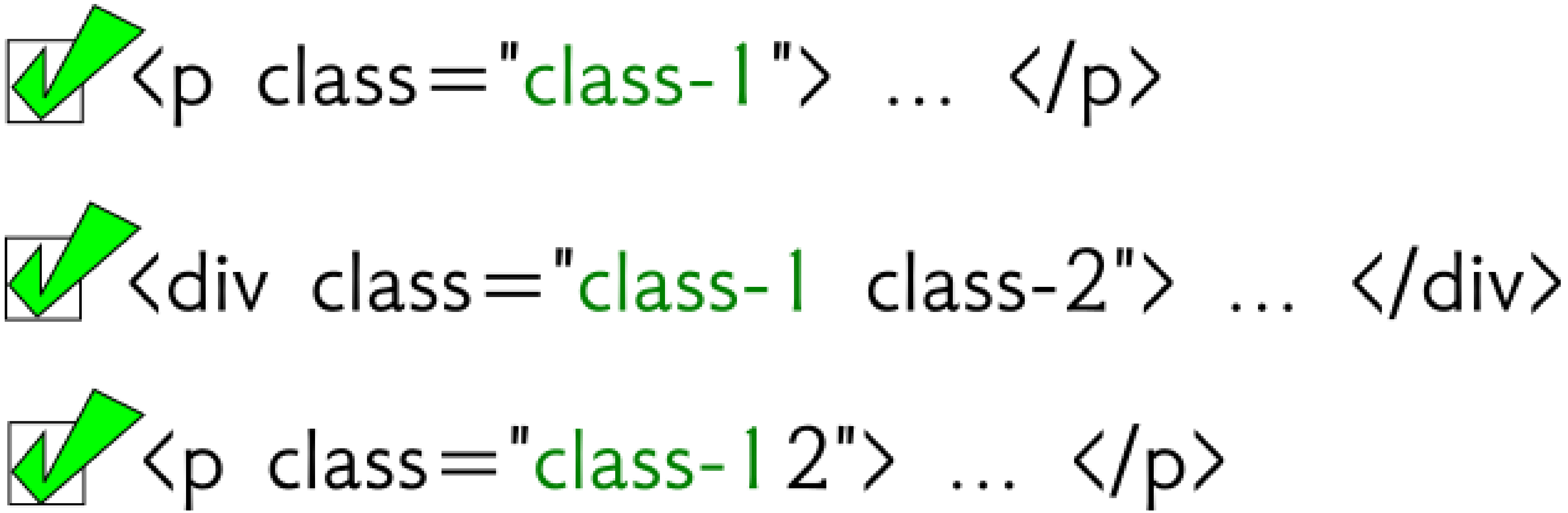
Content with Contains

Xpath Contains Notation:

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

Contain This

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



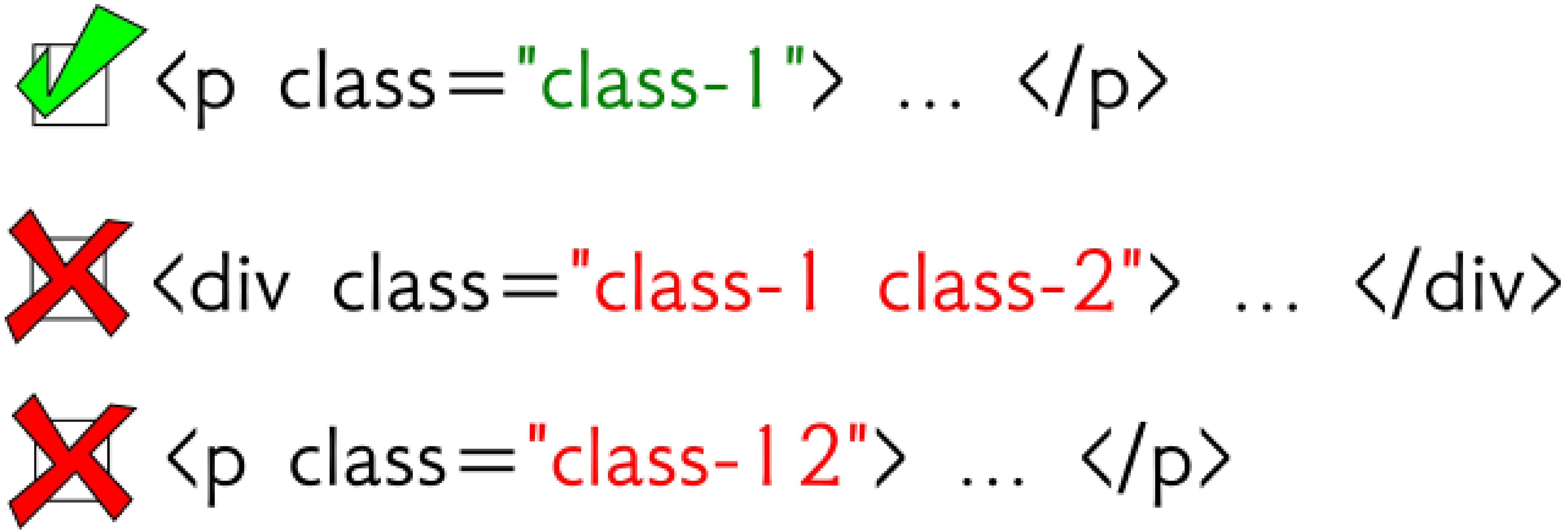
☑ `<p class="class-1"> ... </p>`

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

☑ `<p class="class-12"> ... </p>`

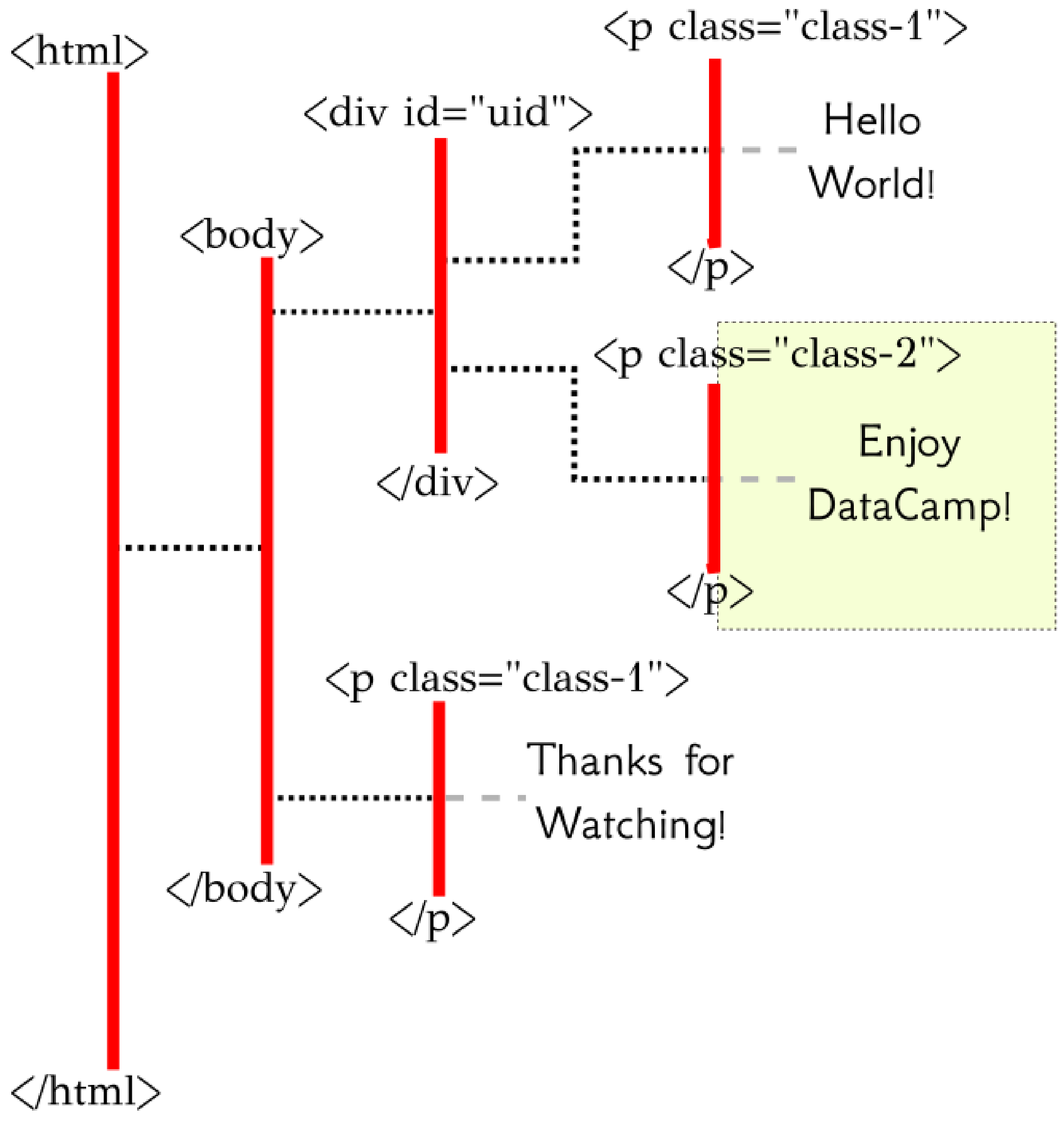
Contain This

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



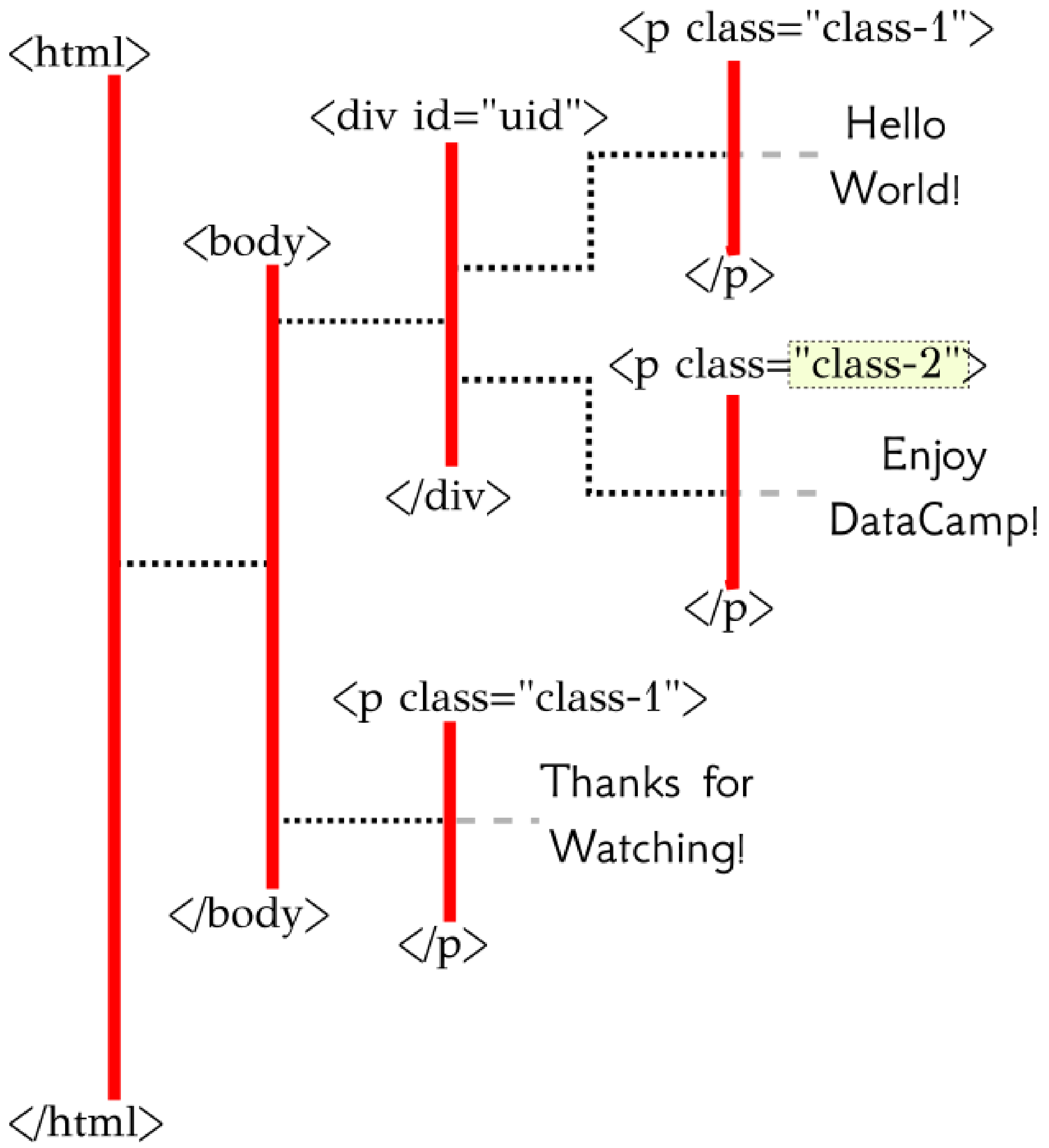
```
✓ <p class="class-1"> ... </p>

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

✗ <p class="class-12"> ... </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>'
```

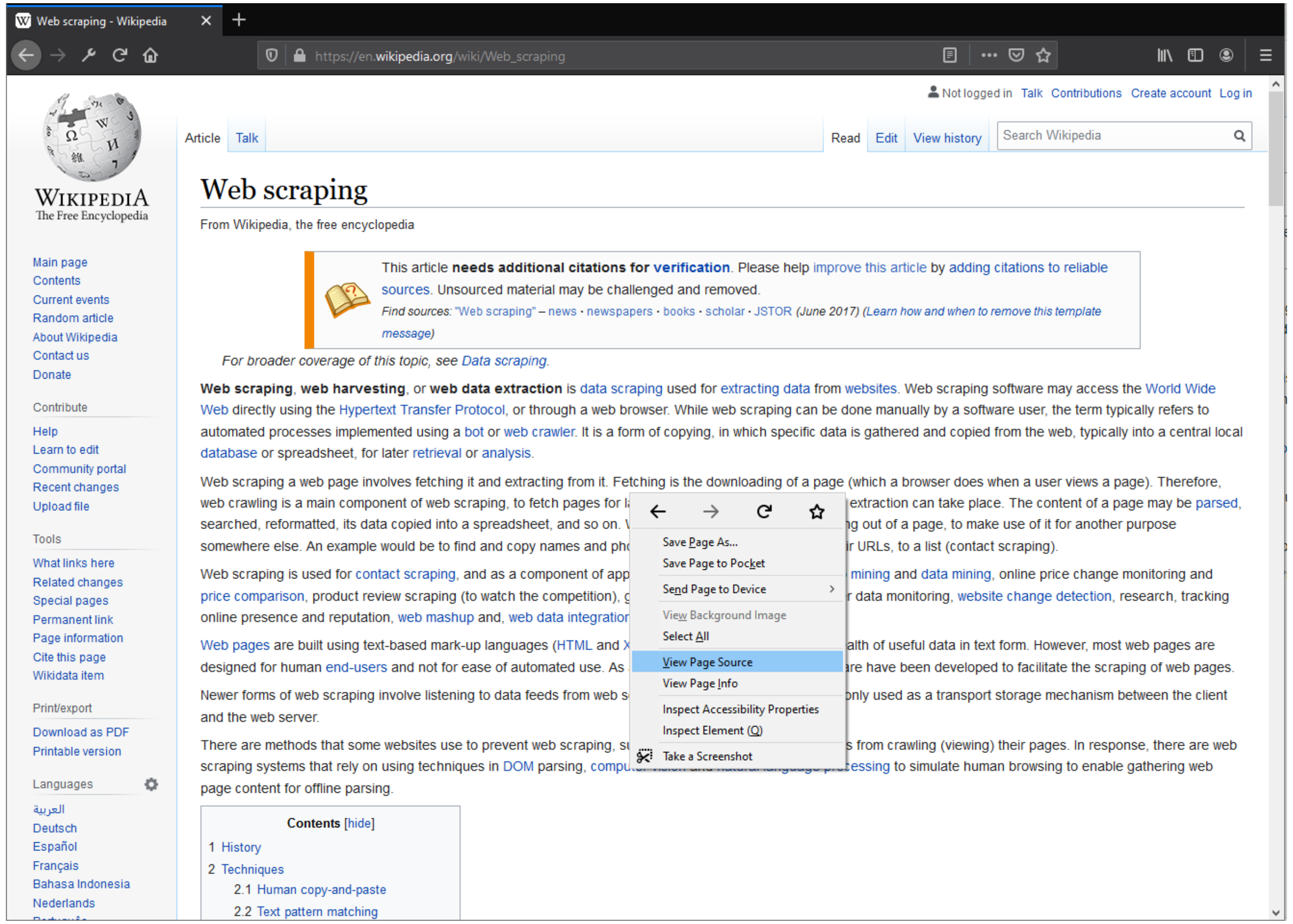
Select This Course!

"Inspecting the HTML"

WEB SCRAPING IN PYTHON

Thomas Laetsch, PhD
Data Scientist, NYU

"Source" = HTML Code



Web scraping

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed.
Find sources: "Web scraping" – news · newspapers · books · scholar · JSTOR (June 2017) (Learn how and when to remove this template message)

For broader coverage of this topic, see Data scraping.

Web scraping, **web harvesting**, or **web data extraction** is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

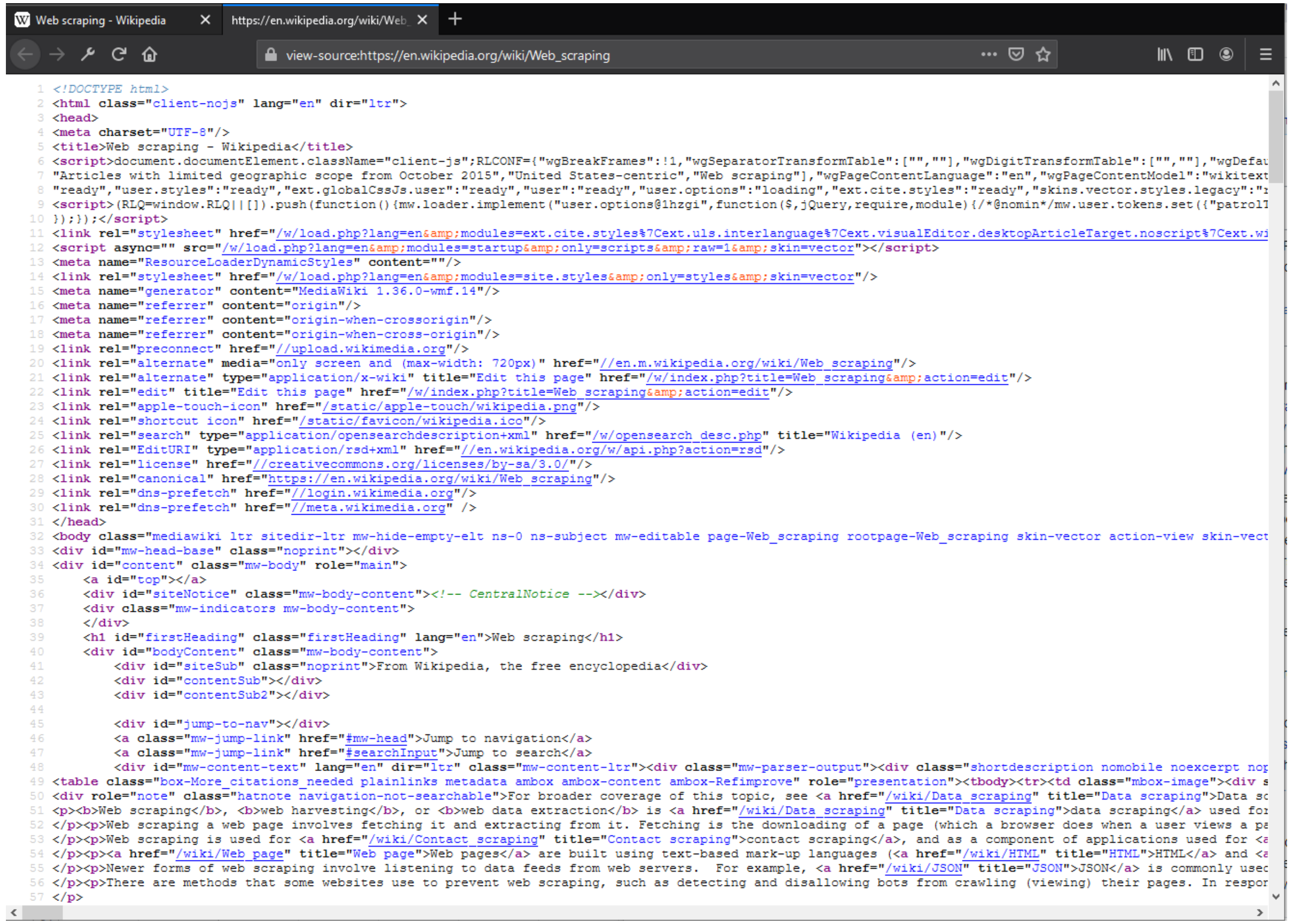
Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when a user views a page). Therefore, web crawling is a main component of web scraping, to fetch pages for l extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet, and so on. V g out of a page, to make use of it for another purpose somewhere else. An example would be to find and copy names and ph r URLs, to a list (contact scraping).

Web scraping is used for contact scraping, and as a component of app mining and data mining, online price change monitoring and price comparison, product review scraping (to watch the competition), g r data monitoring, website change detection, research, tracking online presence and reputation, web mashup and, web data integration

Web pages are built using text-based mark-up languages (HTML and X alth of useful data in text form. However, most web pages are designed for human end-users and not for ease of automated use. As re have been developed to facilitate the scraping of web pages.

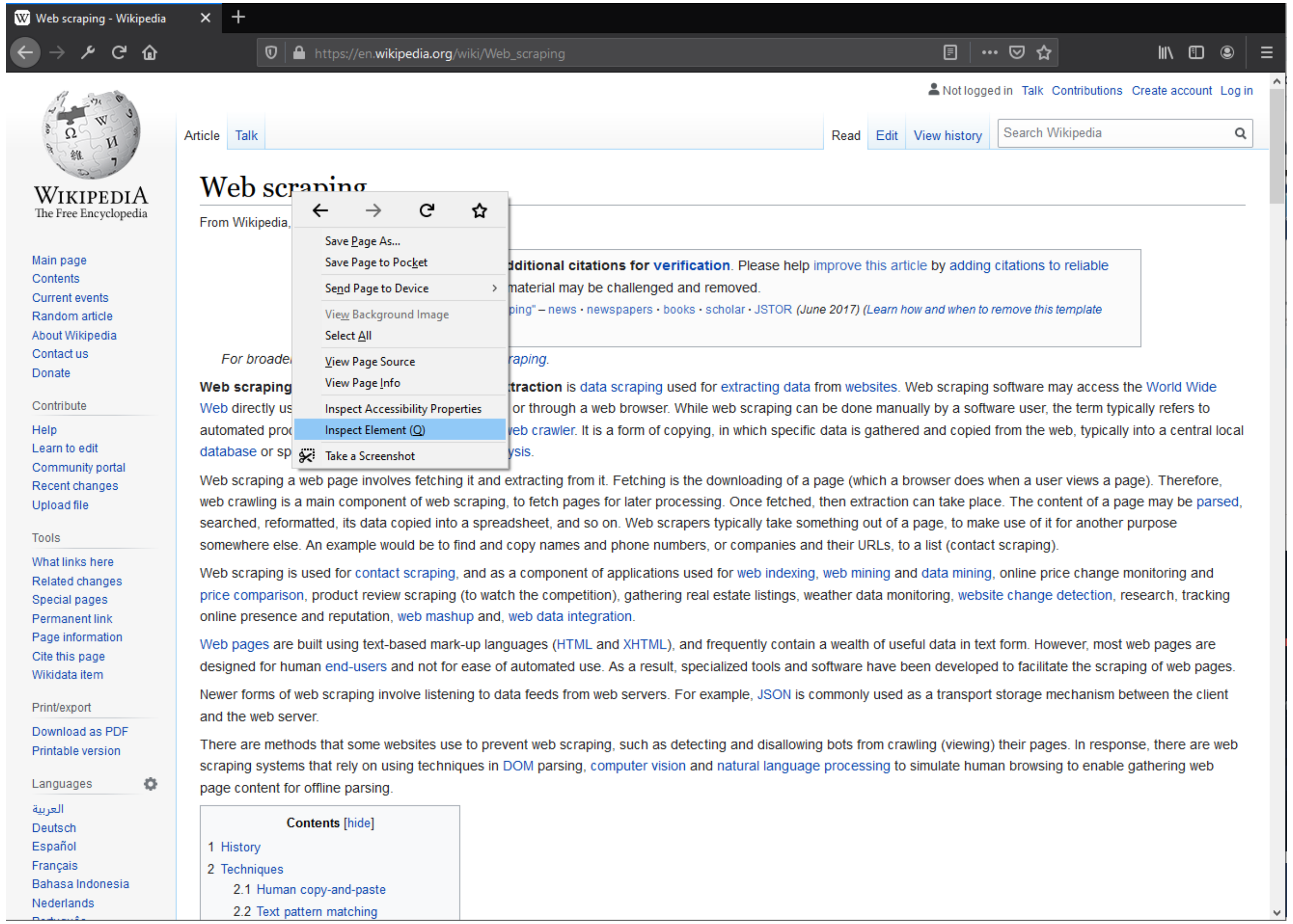
Newer forms of web scraping involve listening to data feeds from web s only used as a transport storage mechanism between the client and the web server.

There are methods that some websites use to prevent web scraping, su s from crawling (viewing) their pages. In response, there are web scraping systems that rely on using techniques in DOM parsing, compu essing to simulate human browsing to enable gathering web page content for offline parsing.


Web scraping - Wikipedia
https://en.wikipedia.org/wiki/Web_
view-source:https://en.wikipedia.org/wiki/Web_scraping

```
<!DOCTYPE html>
<html class="client-nojs" lang="en" dir="ltr">
<head>
<meta charset="UTF-8"/>
<title>Web scraping - Wikipedia</title>
<script>document.documentElement.className="client-js";RLCONF={"wgBreakFrames":!1,"wgSeparatorTransformTable":["",""],"wgDigitTransformTable":["",""],"wgDefau
"Articles with limited geographic scope from October 2015","United States-centric","Web scraping"],"wgPageContentLanguage":"en","wgPageContentModel":"wikitext
"ready","user.styles":"ready","ext.globalCssJs.user":"ready","user":"ready","user.options":"loading","ext.cite.styles":"ready","skins.vector.styles.legacy":"r
<script>(RLQ=window.RLQ||[]).push(function(){mw.loader.implement("user.options@1hzgi",function($,jQuery,require,module){/*@nomin*/mw.user.tokens.set({"patrolT
});});</script>
<link rel="stylesheet" href="/w/load.php?lang=en&amp;modules=ext.cite.styles%7Cext.uls.interlanguage%7Cext.visualEditor.desktopArticleTarget.noscript%7Cext.wi
<script async="" src="/w/load.php?lang=en&amp;modules=startup&amp;only=scripts&amp;raw=1&amp;skin=vector"></script>
<meta name="ResourceLoaderDynamicStyles" content=""/>
<link rel="stylesheet" href="/w/load.php?lang=en&amp;modules=site.styles&amp;only=styles&amp;skin=vector"/>
<meta name="generator" content="MediaWiki 1.36.0-wmf.14"/>
<meta name="referrer" content="origin"/>
<meta name="referrer" content="origin-when-crossorigin"/>
<meta name="referrer" content="origin-when-cross-origin"/>
<link rel="preconnect" href="//upload.wikimedia.org"/>
<link rel="alternate" media="only screen and (max-width: 720px)" href="//en.m.wikipedia.org/wiki/Web_scraping"/>
<link rel="alternate" type="application/x-wiki" title="Edit this page" href="/w/index.php?title=Web_scraping&amp;action=edit"/>
<link rel="edit" title="Edit this page" href="/w/index.php?title=Web_scraping&amp;action=edit"/>
<link rel="apple-touch-icon" href="/static/apple-touch/wikipedia.png"/>
<link rel="shortcut icon" href="/static/favicon/wikipedia.ico"/>
<link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikipedia (en)"/>
<link rel="EditURI" type="application/rsd+xml" href="//en.wikipedia.org/w/api.php?action=rsd"/>
<link rel="license" href="//creativecommons.org/licenses/by-sa/3.0/"/>
<link rel="canonical" href="https://en.wikipedia.org/wiki/Web_scraping"/>
<link rel="dns-prefetch" href="//login.wikimedia.org"/>
<link rel="dns-prefetch" href="//meta.wikimedia.org" />
</head>
<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject mw-editable page-Web_scraping rootpage-Web_scraping skin-vector action-view skin-vect
<div id="mw-head-base" class="noprint"></div>
<div id="content" class="mw-body" role="main">
    <a id="top"></a>
    <div id="siteNotice" class="mw-body-content"><!-- CentralNotice --></div>
    <div class="mw-indicators mw-body-content">
    </div>
    <h1 id="firstHeading" class="firstHeading" lang="en">Web scraping</h1>
    <div id="bodyContent" class="mw-body-content">
        <div id="siteSub" class="noprint">From Wikipedia, the free encyclopedia</div>
        <div id="contentSub"></div>
        <div id="contentSub2"></div>

        <div id="jump-to-nav"></div>
        <a class="mw-jump-link" href="#mw-head">Jump to navigation</a>
        <a class="mw-jump-link" href="#searchInput">Jump to search</a>
        <div id="mw-content-text" lang="en" dir="ltr" class="mw-content-ltr"><div class="mw-parser-output"><div class="shortdescription nomobile noexcerpt nop
<table class="box-More_citations_needed plainlinks metadata ambox ambox-content ambox-Refimprove" role="presentation"><tbody><tr><td class="mbox-image"><div s
<div role="note" class="hatnote navigation-not-searchable">For broader coverage of this topic, see <a href="/wiki/Data_scraping" title="Data scraping">Data sc
<p><b>Web scraping</b>, <b>web harvesting</b>, or <b>web data extraction</b> is <a href="/wiki/Data_scraping" title="Data scraping">data scraping</a> used for
</p><p>Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when a user views a pa
</p><p>Web scraping is used for <a href="/wiki/Contact_scraping" title="Contact scraping">contact scraping</a>, and as a component of applications used for <a
</p><p><a href="/wiki/Web_page" title="Web page">Web pages</a> are built using text-based mark-up languages (<a href="/wiki/HTML" title="HTML">HTML</a> and <a
</p><p>Newer forms of web scraping involve listening to data feeds from web servers.  For example, <a href="/wiki/JSON" title="JSON">JSON</a> is commonly usec
</p><p>There are methods that some websites use to prevent web scraping, such as detecting and disallowing bots from crawling (viewing) their pages. In respor
</p>
```

Inspecting Elements



Web scraping

From Wikipedia,

dditional citations for verification. Please help improve this article by adding citations to reliable material may be challenged and removed.

ping" – news · newspapers · books · scholar · JSTOR (June 2017) (Learn how and when to remove this template

For broade raping.

Web scraping **traction** is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly us or through a web browser. While web scraping can be done manually by a software user, the term typically refers to automated proc eb crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or sp ysis.

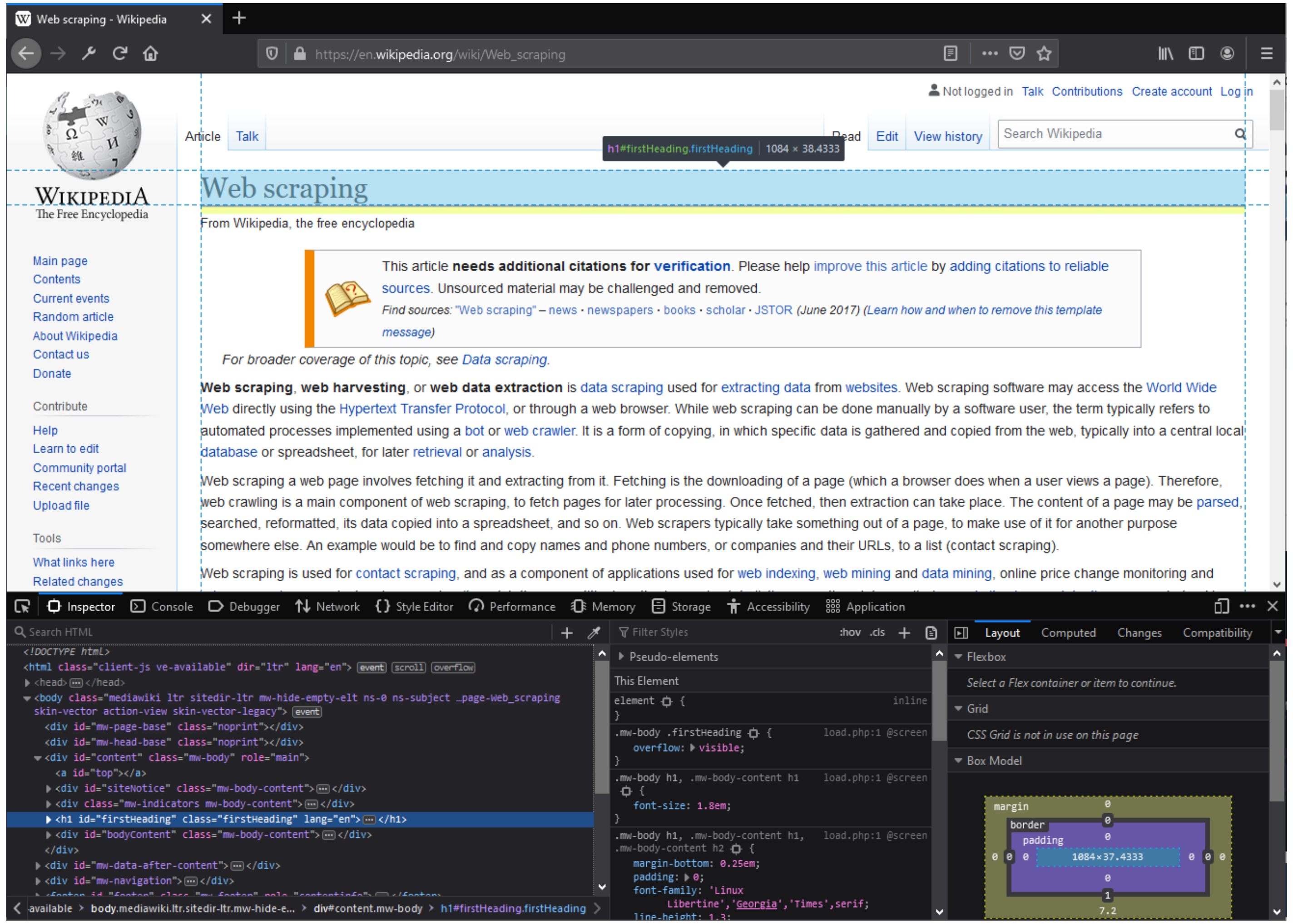
Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when a user views a page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet, and so on. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. An example would be to find and copy names and phone numbers, or companies and their URLs, to a list (contact scraping).

Web scraping is used for contact scraping, and as a component of applications used for web indexing, web mining and data mining, online price change monitoring and price comparison, product review scraping (to watch the competition), gathering real estate listings, weather data monitoring, website change detection, research, tracking online presence and reputation, web mashup and, web data integration.

Web pages are built using text-based mark-up languages (HTML and XHTML), and frequently contain a wealth of useful data in text form. However, most web pages are designed for human end-users and not for ease of automated use. As a result, specialized tools and software have been developed to facilitate the scraping of web pages.

Newer forms of web scraping involve listening to data feeds from web servers. For example, JSON is commonly used as a transport storage mechanism between the client and the web server.

There are methods that some websites use to prevent web scraping, such as detecting and disallowing bots from crawling (viewing) their pages. In response, there are web scraping systems that rely on using techniques in DOM parsing, computer vision and natural language processing to simulate human browsing to enable gathering web page content for offline parsing.



Web scraping

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed.
Find sources: "Web scraping" – news · newspapers · books · scholar · JSTOR (June 2017) (Learn how and when to remove this template message)

For broader coverage of this topic, see Data scraping.

Web scraping, **web harvesting**, or **web data extraction** is data scraping used for extracting data from websites. Web scraping software may access the World Wide Web directly using the Hypertext Transfer Protocol, or through a web browser. While web scraping can be done manually by a software user, the term typically refers to automated processes implemented using a bot or web crawler. It is a form of copying, in which specific data is gathered and copied from the web, typically into a central local database or spreadsheet, for later retrieval or analysis.

Web scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when a user views a page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet, and so on. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. An example would be to find and copy names and phone numbers, or companies and their URLs, to a list (contact scraping).

Web scraping is used for contact scraping, and as a component of applications used for web indexing, web mining and data mining, online price change monitoring and

```
<!DOCTYPE html>
<html class="client-js ve-available" dir="ltr" lang="en"> event scroll overflow
  <head> ··· </head>
  <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject _page-Web_scraping
  skin-vector action-view skin-vector-legacy"> event
    <div id="mw-page-base" class="noprint"></div>
    <div id="mw-head-base" class="noprint"></div>
    <div id="content" class="mw-body" role="main">
      <a id="top"></a>
      <div id="siteNotice" class="mw-body-content"> ··· </div>
      <div class="mw-indicators mw-body-content"> ··· </div>
      <h1 id="firstHeading" class="firstHeading" lang="en"> ··· </h1>
      <div id="bodyContent" class="mw-body-content"> ··· </div>
    </div>
    <div id="mw-data-after-content"> ··· </div>
    <div id="mw-navigation"> ··· </div>
```

```
available > body.mediawiki.ltr.sitedir-ltr.mw-hide-e... > div#content.mw-body > h1#firstHeading.firstHeading
```

```
element {                                          inline
}
.mw-body .firstHeading {             load.php:1 @screen
    overflow: visible;
}
.mw-body h1, .mw-body-content h1     load.php:1 @screen
  {
    font-size: 1.8em;
}
.mw-body h1, .mw-body-content h1,    load.php:1 @screen
.mw-body-content h2 {
    margin-bottom: 0.25em;
    padding: 0;
    font-family: 'Linux
        Libertine','Georgia','Times',serif;
    line-height: 1.3;
```

Select a Flex container or item to continue.

CSS Grid is not in use on this page

HTML text to Selector

```python
from scrapy import Selector
```

```python
import requests
url = 'https://en.wikipedia.org/wiki/Web_scraping'
html = requests.get( url ).content
```

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

You Know Our Secrets

WEB SCRAPING IN PYTHON