Introduction to Big Data and Machine Learning

Lecture 9 Web Data Collection

Wenzheng Li
06/13/2024

OUTLINE

- Web-scraping using *Selenium*
- Interactive Mapping
- Google Map API
- OSMnx

Web-scraping

0

HTML (Hypertext Markup Language)

- HTML is the code read by a browser and defines the overall page structure

- A web page (or an html file) has two parts: a **“head”** and a **“body”**.

- **‘head’** contains
  - the page title
  - meta tags
  - links to CSS
  - Javascript
  - other referenced items

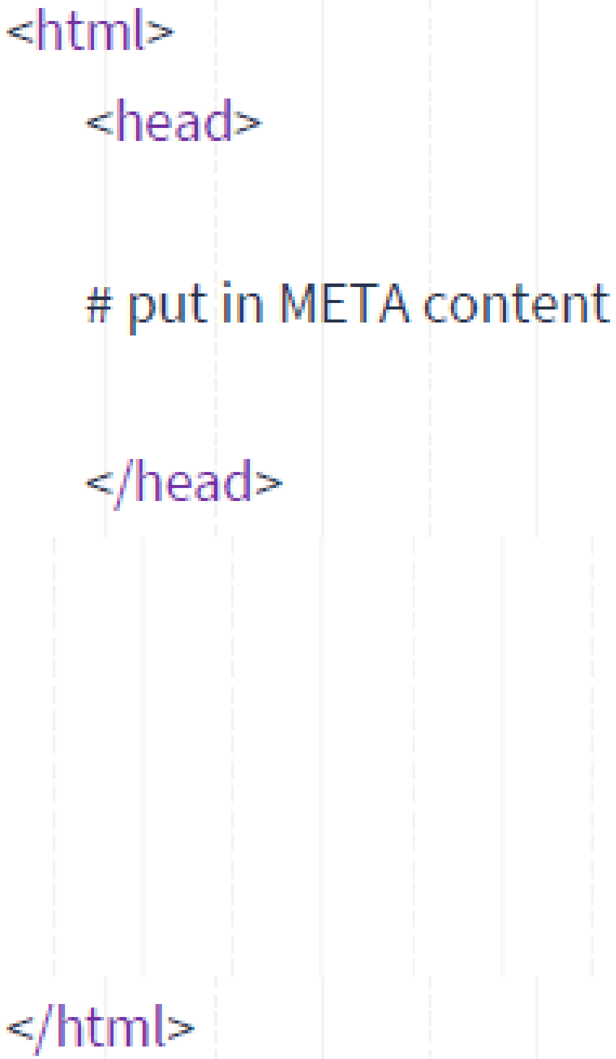
- **‘body’** represents the browser window itself and contains all of the elements that make up the contents of the page.

Basic

```
<!DOCTYPE html>
<html>










</html>
```

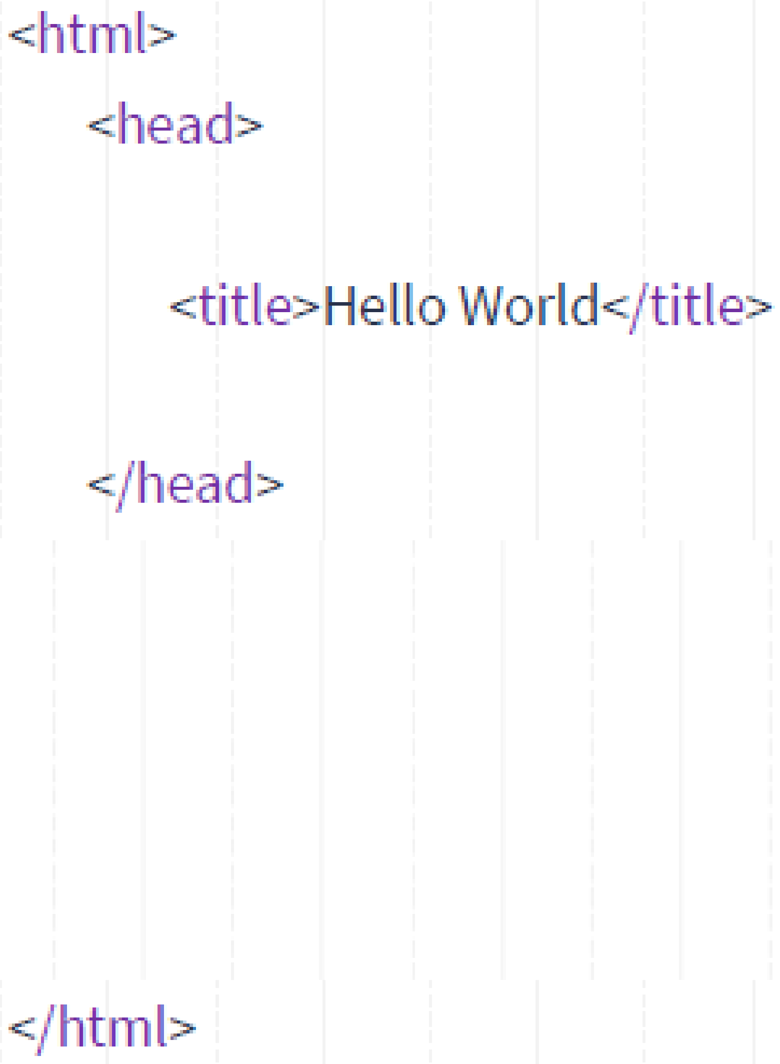
Basic

```
<!DOCTYPE html>
```



```
<html>
     <head>

     # put in META content

     </head>



</html>
```

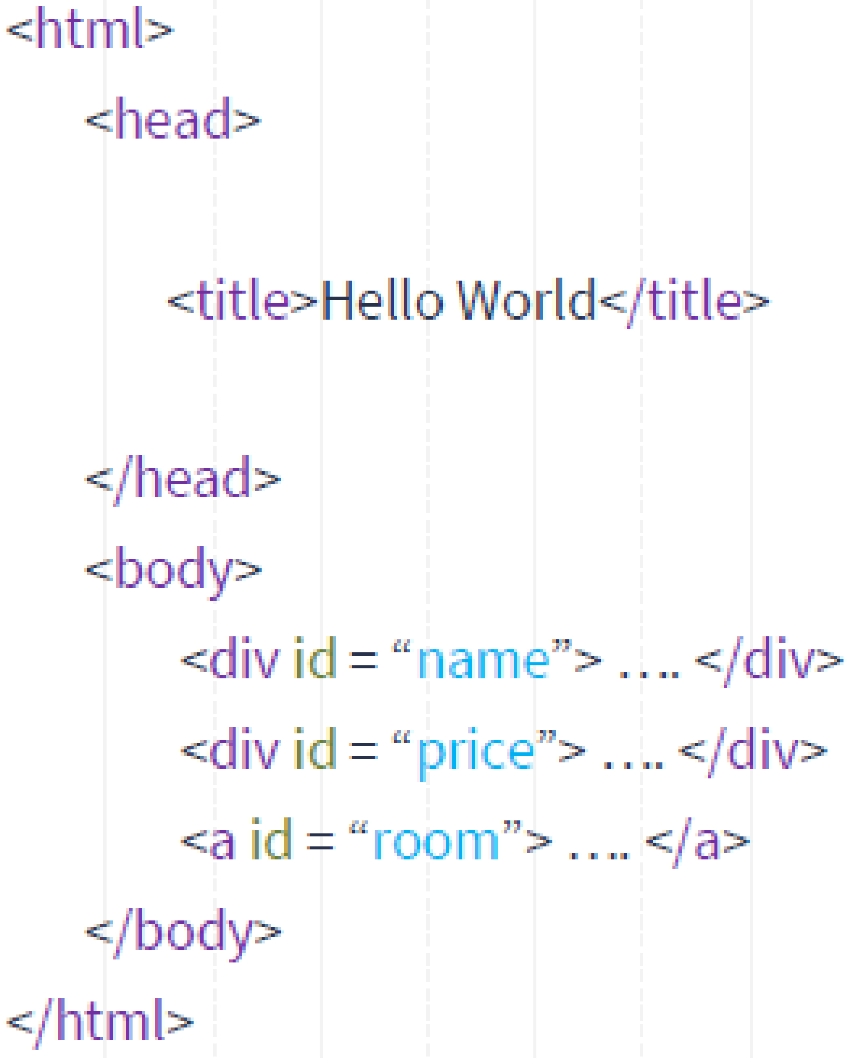
Basic



```
<!DOCTYPE html>
<html>
    <head>

        <title>Hello World</title>

    </head>




</html>
```

Basic

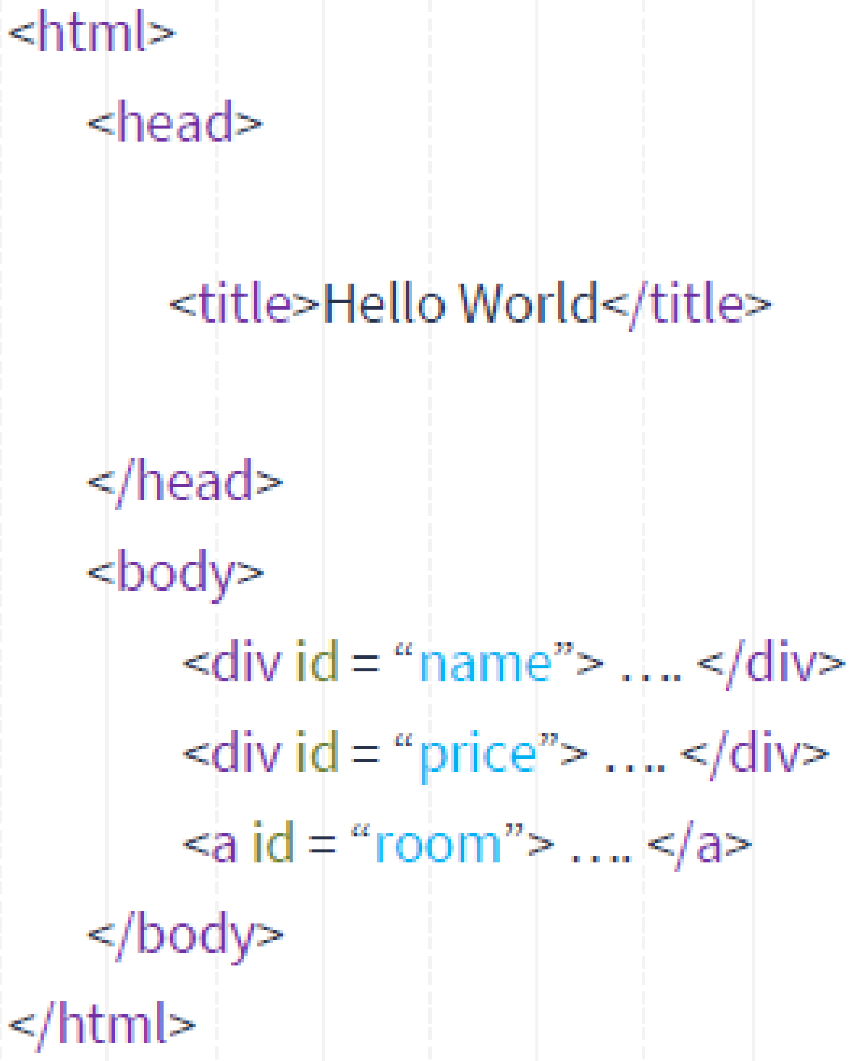
```
<!DOCTYPE html>
```



```
<html>
     <head>


          <title>Hello World</title>


     </head>
     <body>
           <div id = “name”> …. </div>
           <div id = “price”> …. </div>
           <a id = “room”> …. </a>
     </body>
</html>
```

Basic

```
<!DOCTYPE html>
```

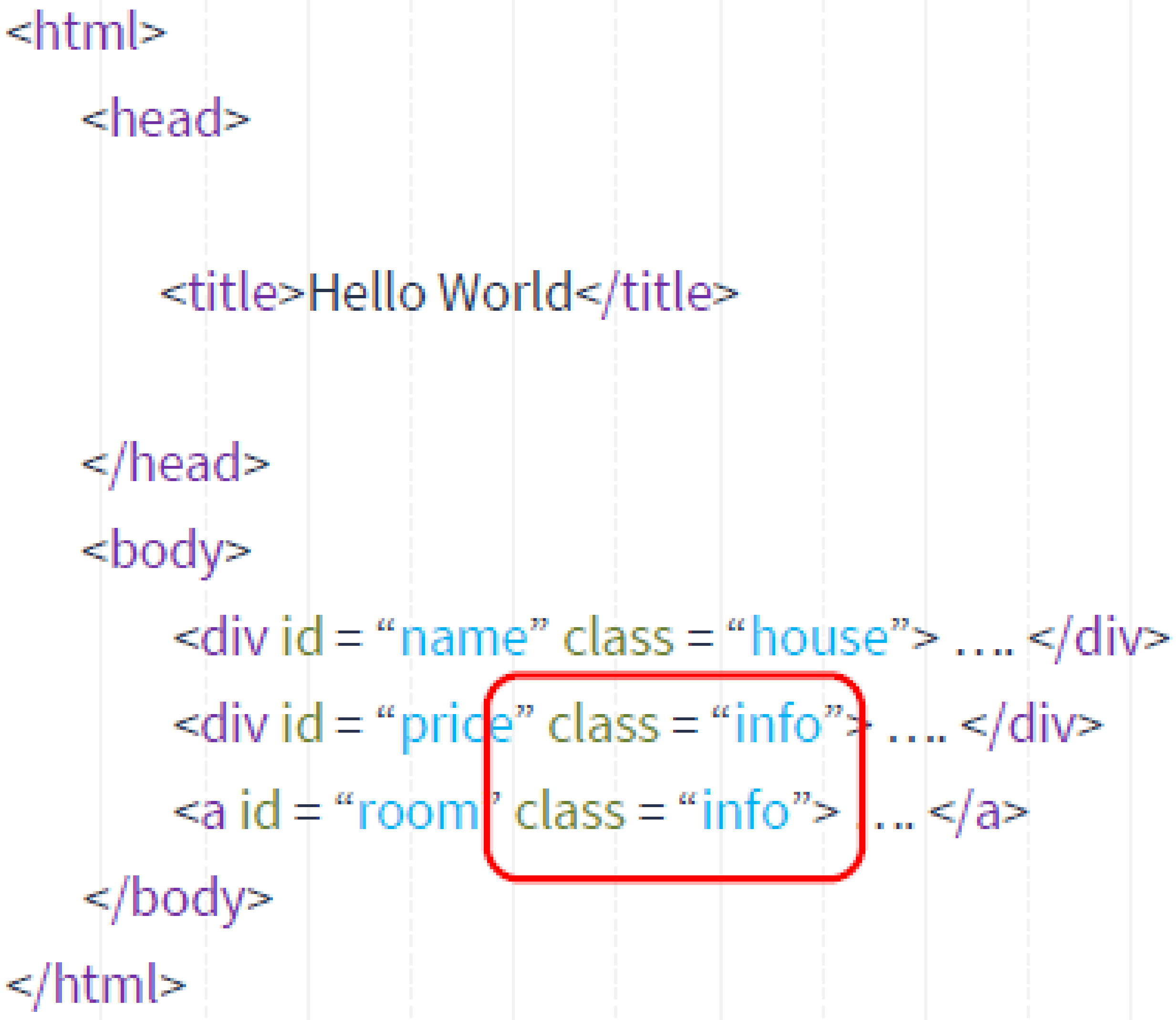


```
<html>
    <head>

        <title>Hello World</title>

    </head>
    <body>
        <div id = “name”> …. </div>
        <div id = “price”> …. </div>
        <a id = “room”> …. </a>
    </body>
</html>
```

<h1> to <h6>: Heading tags
<p>: Paragraph tag
<a>: Anchor tag for hyperlinks
<img>: Image tag
<div>: Division tag for sections
<span>: Inline container tag

Basic

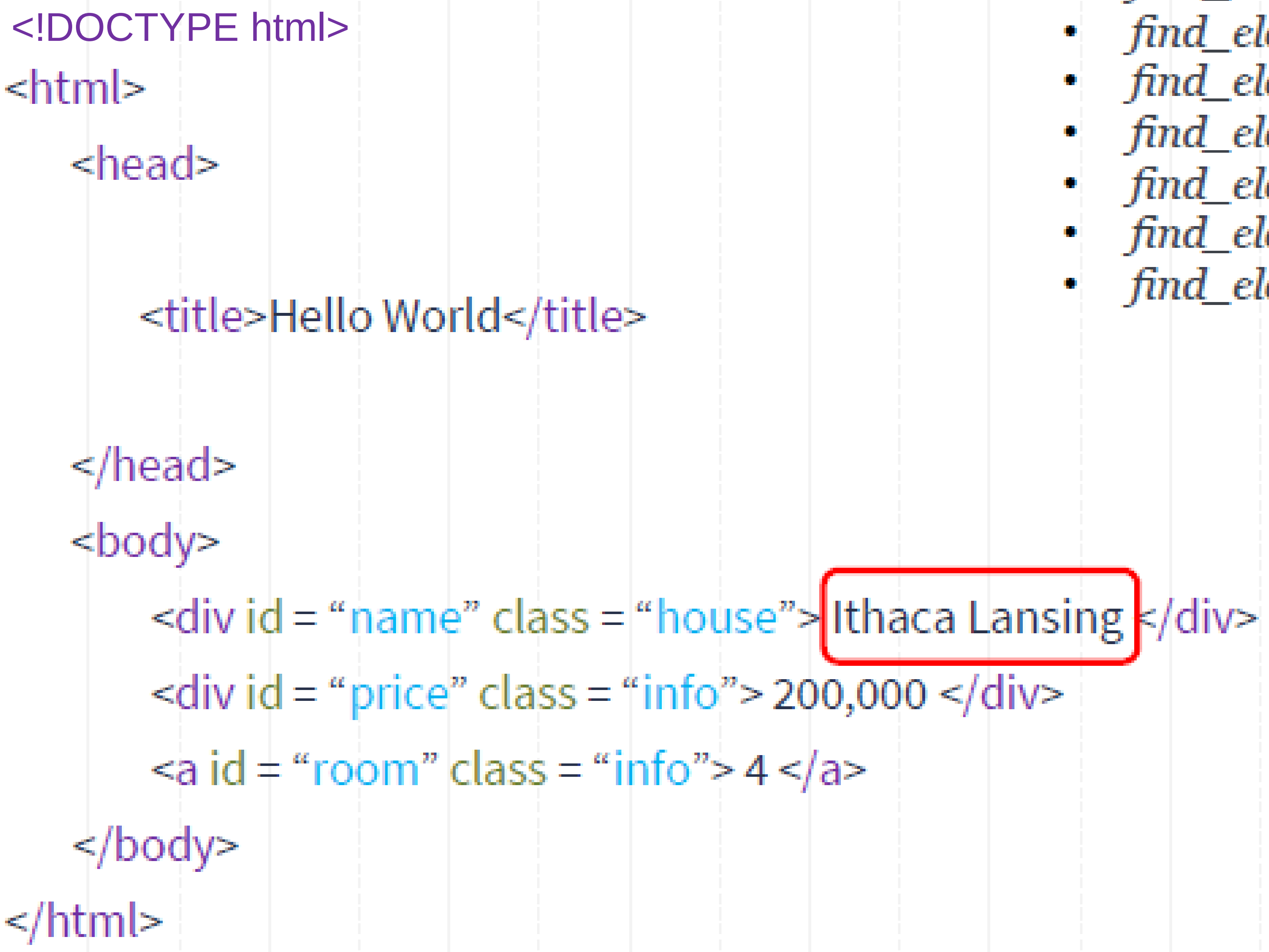
```
<!DOCTYPE html>
```



```
<html>
     <head>

          <title>Hello World</title>

     </head>
     <body>
           <div id = “name” class = “house”> …. </div>
           <div id = “price” class = “info”> …. </div>
           <a id = “room” class = “info”> …. </a>
     </body>
</html>
```

Basic

- *find_element_by_id*
- *find_element_by_name*
- *find_element_by_xpath*
- *find_element_by_link_text*
- *find_element_by_partial_link_text*
- *find_element_by_tag_name*
- *find_element_by_class_name*
- *find_element_by_css_selector*



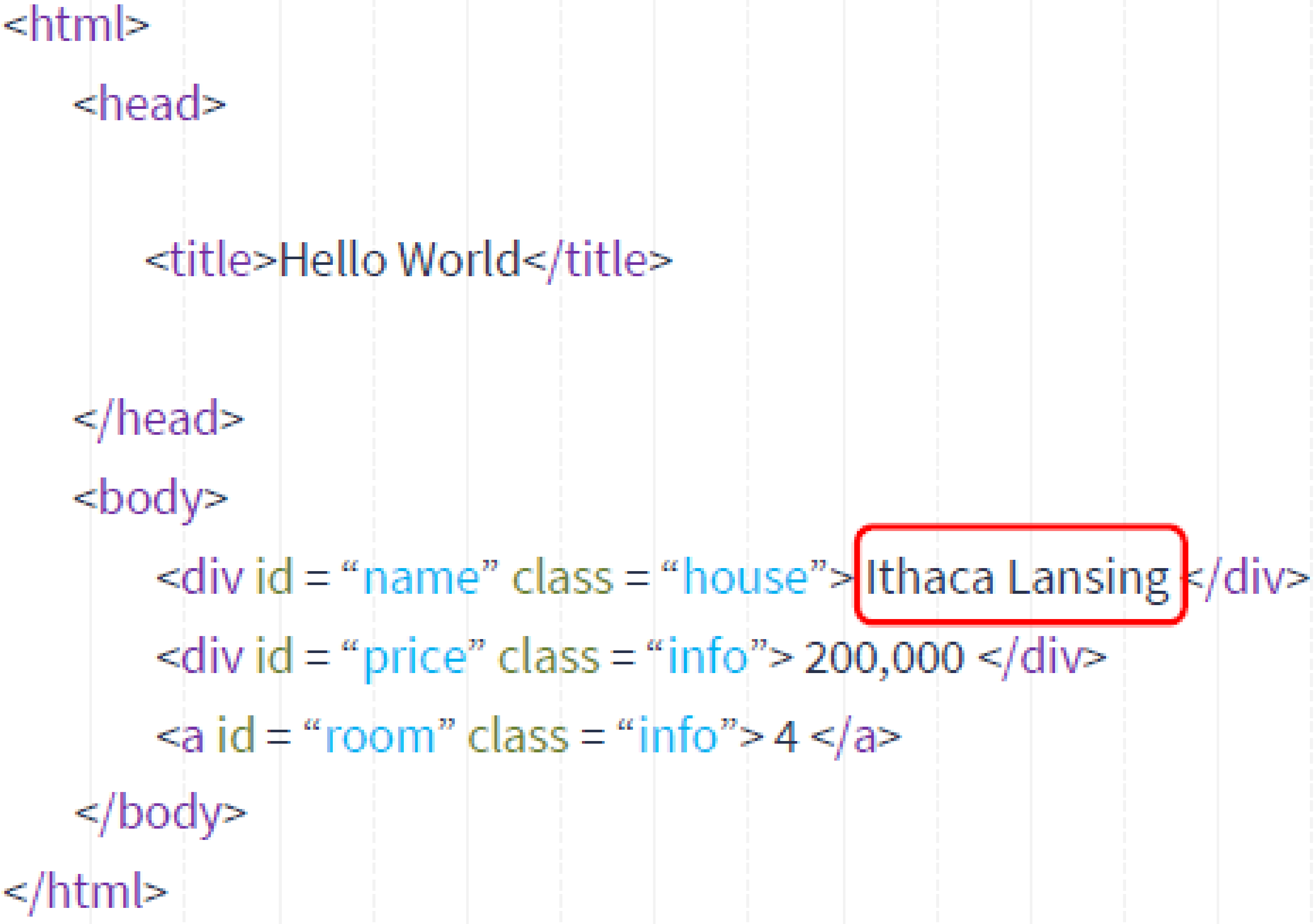
```
<!DOCTYPE html>
<html>
    <head>

        <title>Hello World</title>

    </head>
    <body>
        <div id = “name” class = “house”> Ithaca Lansing </div>
        <div id = “price” class = “info”> 200,000 </div>
        <a id = “room” class = “info”> 4 </a>
    </body>
</html>
```

Basic

- *find_element_by_id*

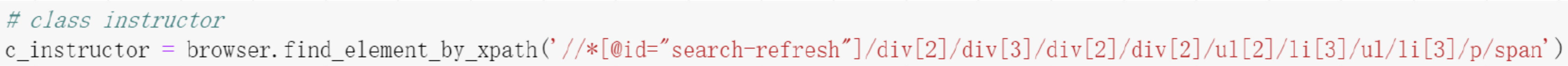
```
brower.find_element_by_id('name')
```

```
<!DOCTYPE html>
```

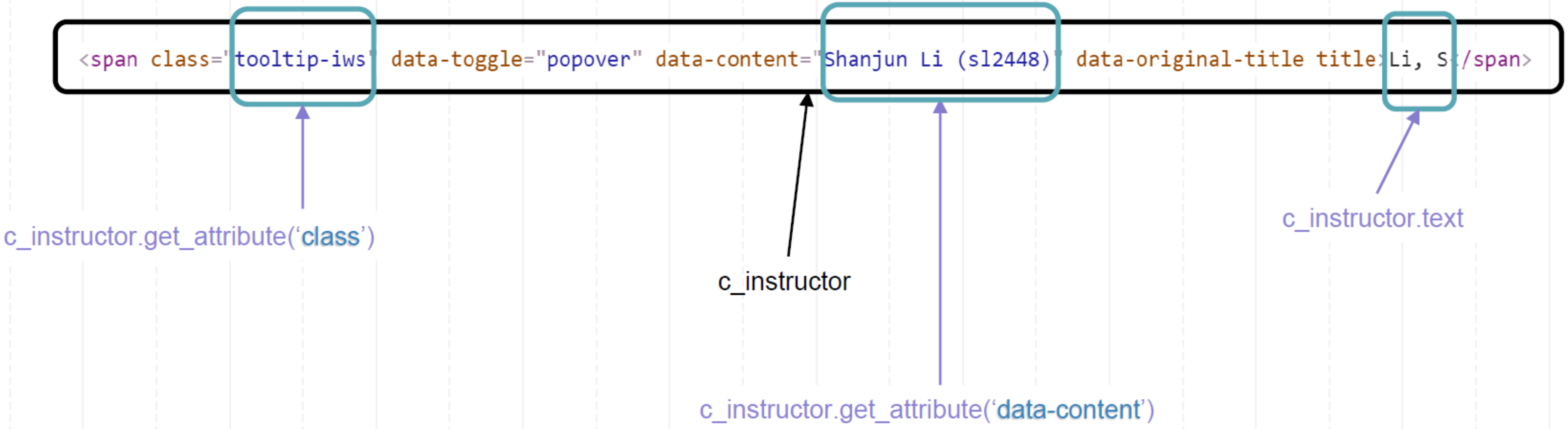


```
<html>
    <head>

        <title>Hello World</title>

    </head>
    <body>
        <div id = "name" class = "house"> Ithaca Lansing </div>
        <div id = "price" class = "info"> 200,000 </div>
        <a id = "room" class = "info"> 4 </a>
    </body>
</html>
```

Get Attribute from An Element



```
# class instructor
c_instructor = browser.find_element_by_xpath(’//*[@id=”search-refresh”]/div[2]/div[3]/div[2]/div[2]/ul[2]/li[3]/ul/li[3]/p/span’)
```



```
<span class="tooltip-iws" data-toggle="popover" data-content="Shanjun Li (sl2448)" data-original-title title>Li, S</span>
```

c_instructor.get_attribute(‘class’)

c_instructor

c_instructor.text

c_instructor.get_attribute(‘data-content’)

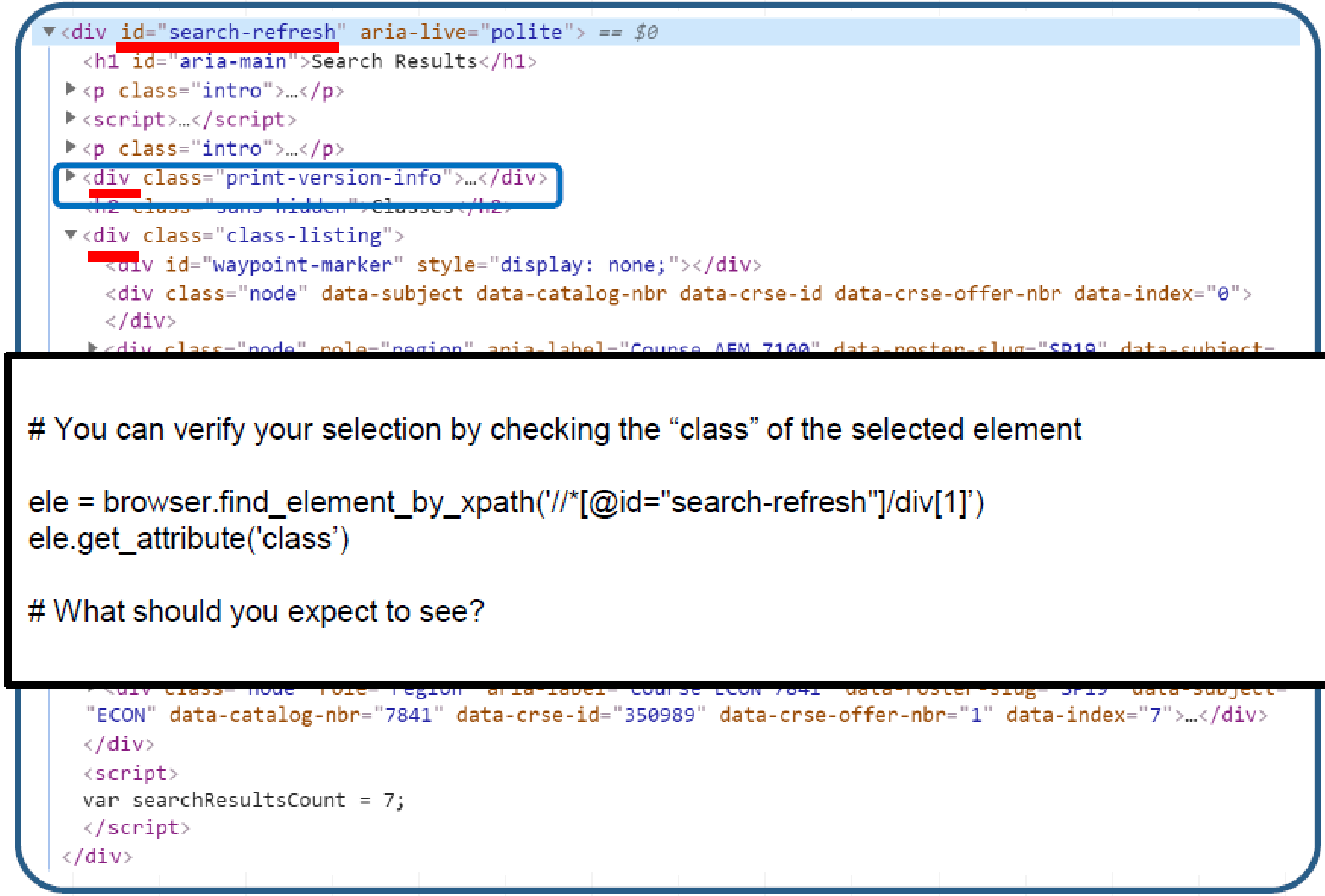
Web Scraping in a Loop

Loop Design

- Understand the XPath
- Using XPath to loop over results
- Replace absolute XPath (using direct “id”) with relative location XPath

XPath

'//*[@id="search-refresh"]/div[1]'



```
▼<div id="search-refresh" aria-live="polite"> == $0
   <h1 id="aria-main">Search Results</h1>
  ▶<p class="intro">…</p>
  ▶<script>…</script>
  ▶<p class="intro">…</p>
  ▶<div class="print-version-info">…</div>
   <h2 class="sans-hidden">Classes</h2>
  ▼<div class="class-listing">
     <div id="waypoint-marker" style="display: none;"></div>
     <div class="node" data-subject data-catalog-nbr data-crse-id data-crse-offer-nbr data-index="0">
     </div>
    ▶<div class="node" role="region" aria-label="Course AEM 7100" data-roster-slug="SP18" data-subject=
    ▶<div class="node" role="region" aria-label="Course ECON 7841" data-roster-slug="SP19" data-subject=
   "ECON" data-catalog-nbr="7841" data-crse-id="350989" data-crse-offer-nbr="1" data-index="7">…</div>
   </div>
   <script>
   var searchResultsCount = 7;
   </script>
</div>
```

You can verify your selection by checking the "class" of the selected element

ele = browser.find_element_by_xpath('//*[@id="search-refresh"]/div[1]')
ele.get_attribute('class')

What should you expect to see?

Loop Design

- Understand the XPath
- Using XPath to loop over results
- Replace absolute XPath (using direct “id”) with relative location XPath

XPath Comparison (#class weekday)

- First Result

```
//*[@id="search-refresh"]/div[2]/div[3]/div[2]/div[2]/ul[2]/li[3]/ul/li[1]/span/span/span
```

- Second Result

```
//*[@id="search-refresh"]/div[2]/div[4]/div[2]/div[2]/ul[2]/li[3]/ul/li[1]/span/span/span
```

- Third Result

```
//*[@id="search-refresh"]/div[2]/div[5]/div[2]/div[2]/ul[2]/li[3]/ul/li[1]/span/span/span
```

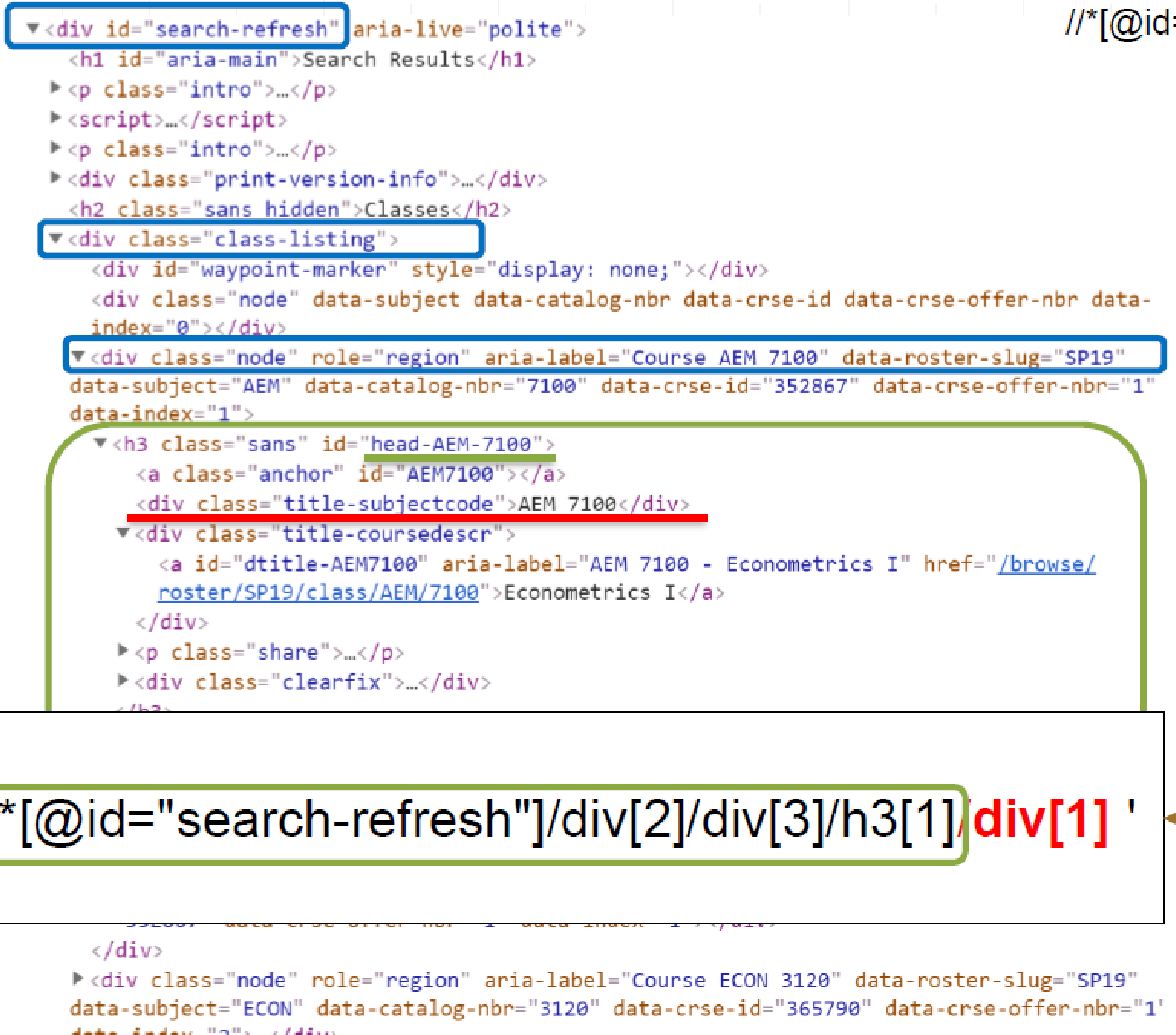
Loop Design

- Understand the XPath
- Using XPath to loop over results
- Replace absolute XPath (using direct “id”) with relative location XPath

XPath

'//*[@id="head-AEM-7100"]/div[1] '

XPath



```
▼<div id="search-refresh" aria-live="polite">
   <h1 id="aria-main">Search Results</h1>
  ▶<p class="intro">…</p>
  ▶<script>…</script>
  ▶<p class="intro">…</p>
  ▶<div class="print-version-info">…</div>
   <h2 class="sans hidden">Classes</h2>
  ▼<div class="class-listing">
     <div id="waypoint-marker" style="display: none;"></div>
     <div class="node" data-subject data-catalog-nbr data-crse-id data-crse-offer-nbr data-
     index="0"></div>
   ▼<div class="node" role="region" aria-label="Course AEM 7100" data-roster-slug="SP19"
   data-subject="AEM" data-catalog-nbr="7100" data-crse-id="352867" data-crse-offer-nbr="1"
   data-index="1">
     ▼<h3 class="sans" id="head-AEM-7100">
        <a class="anchor" id="AEM7100"></a>
        <div class="title-subjectcode">AEM 7100</div>
       ▼<div class="title-coursedescr">
          <a id="dtitle-AEM7100" aria-label="AEM 7100 - Econometrics I" href="/browse/
          roster/SP19/class/AEM/7100">Econometrics I</a>
        </div>
       ▶<p class="share">…</p>
       ▶<div class="clearfix">…</div>
```

```
     </div>
   ▶<div class="node" role="region" aria-label="Course ECON 3120" data-roster-slug="SP19"
   data-subject="ECON" data-catalog-nbr="3120" data-crse-id="365790" data-crse-offer-nbr="1"
```

//*[@id="search-refresh"]

/div[2]

/div[3]

/h3[1]

'//*[@id="search-refresh"]/div[2]/div[3]/h3[1]/div[1] '

Full XPath