LECTURE 5 MINING WEB CONTENT II

LEK HSIANG HUI

OUTLINE

Document Object Model (DOM)

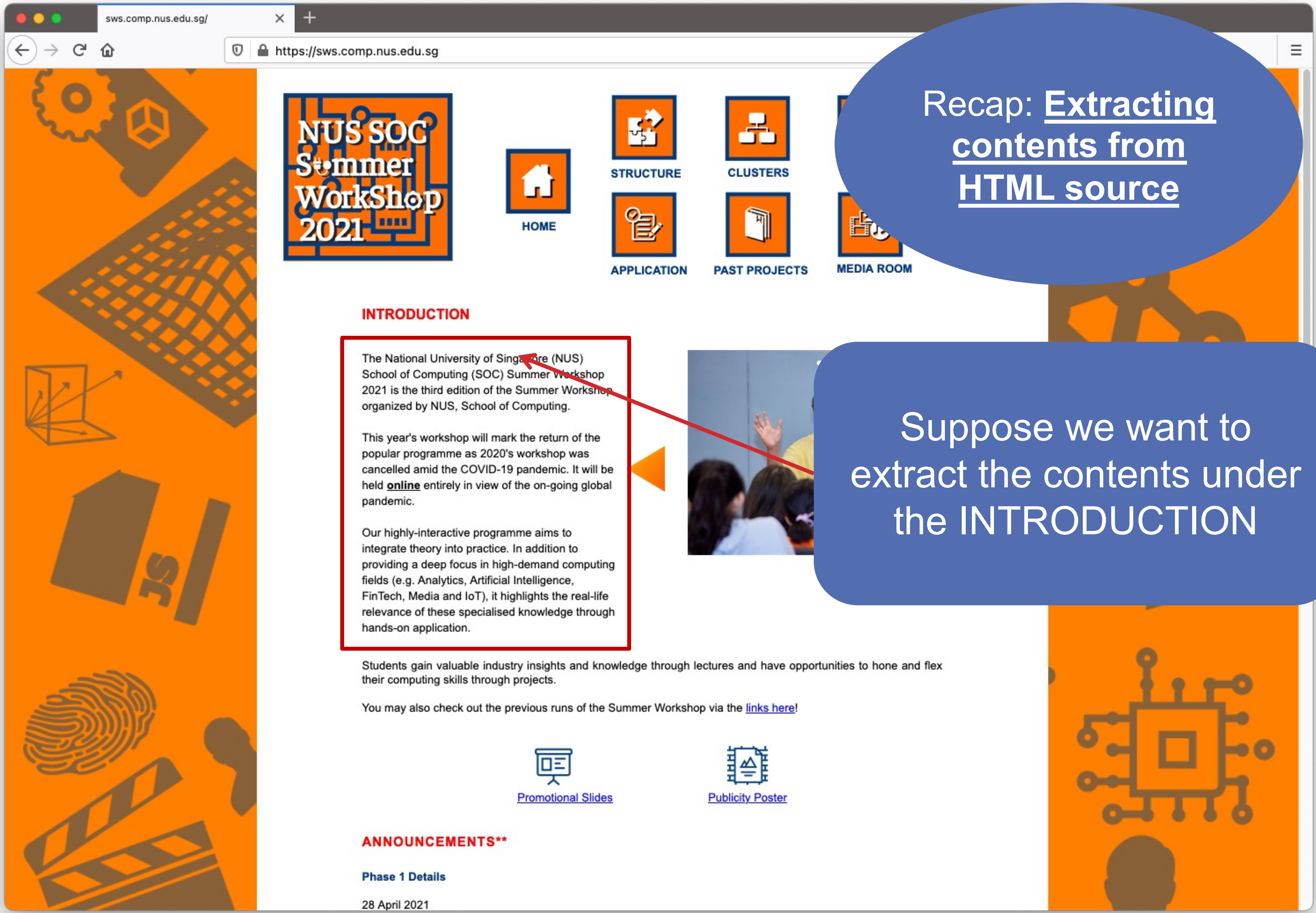
XPath

CSS Selectors

Extracting Content using HTML Parser

Recap: **Extracting contents from HTML source**

Suppose we want to extract the contents under the INTRODUCTION



INTRODUCTION

The National University of Singapore (NUS) School of Computing (SOC) Summer Workshop 2021 is the third edition of the Summer Workshop organized by NUS, School of Computing.

This year's workshop will mark the return of the popular programme as 2020's workshop was cancelled amid the COVID-19 pandemic. It will be held **online** entirely in view of the on-going global pandemic.

Our highly-interactive programme aims to integrate theory into practice. In addition to providing a deep focus in high-demand computing fields (e.g. Analytics, Artificial Intelligence, FinTech, Media and IoT), it highlights the real-life relevance of these specialised knowledge through hands-on application.

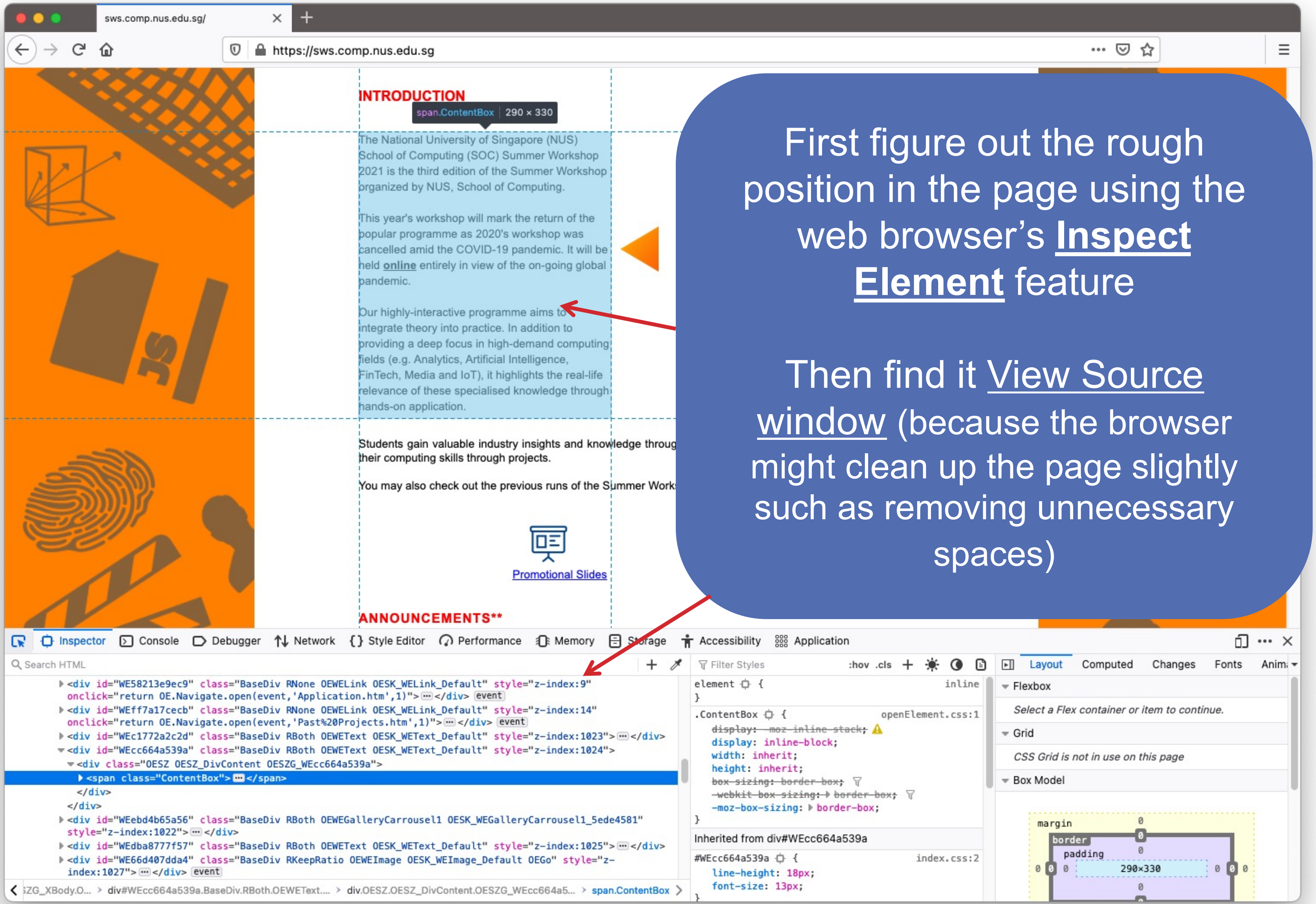
Students gain valuable industry insights and knowledge through lectures and have opportunities to hone and flex their computing skills through projects.

You may also check out the previous runs of the Summer Workshop via the links here!

ANNOUNCEMENTS**

Phase 1 Details

28 April 2021



INTRODUCTION

The National University of Singapore (NUS) School of Computing (SOC) Summer Workshop 2021 is the third edition of the Summer Workshop organized by NUS, School of Computing.

This year's workshop will mark the return of the popular programme as 2020's workshop was cancelled amid the COVID-19 pandemic. It will be held **online** entirely in view of the on-going global pandemic.

Our highly-interactive programme aims to integrate theory into practice. In addition to providing a deep focus in high-demand computing fields (e.g. Analytics, Artificial Intelligence, FinTech, Media and IoT), it highlights the real-life relevance of these specialised knowledge through hands-on application.

Students gain valuable industry insights and knowledge throug their computing skills through projects.

You may also check out the previous runs of the Summer Work

Promotional Slides

ANNOUNCEMENTS**

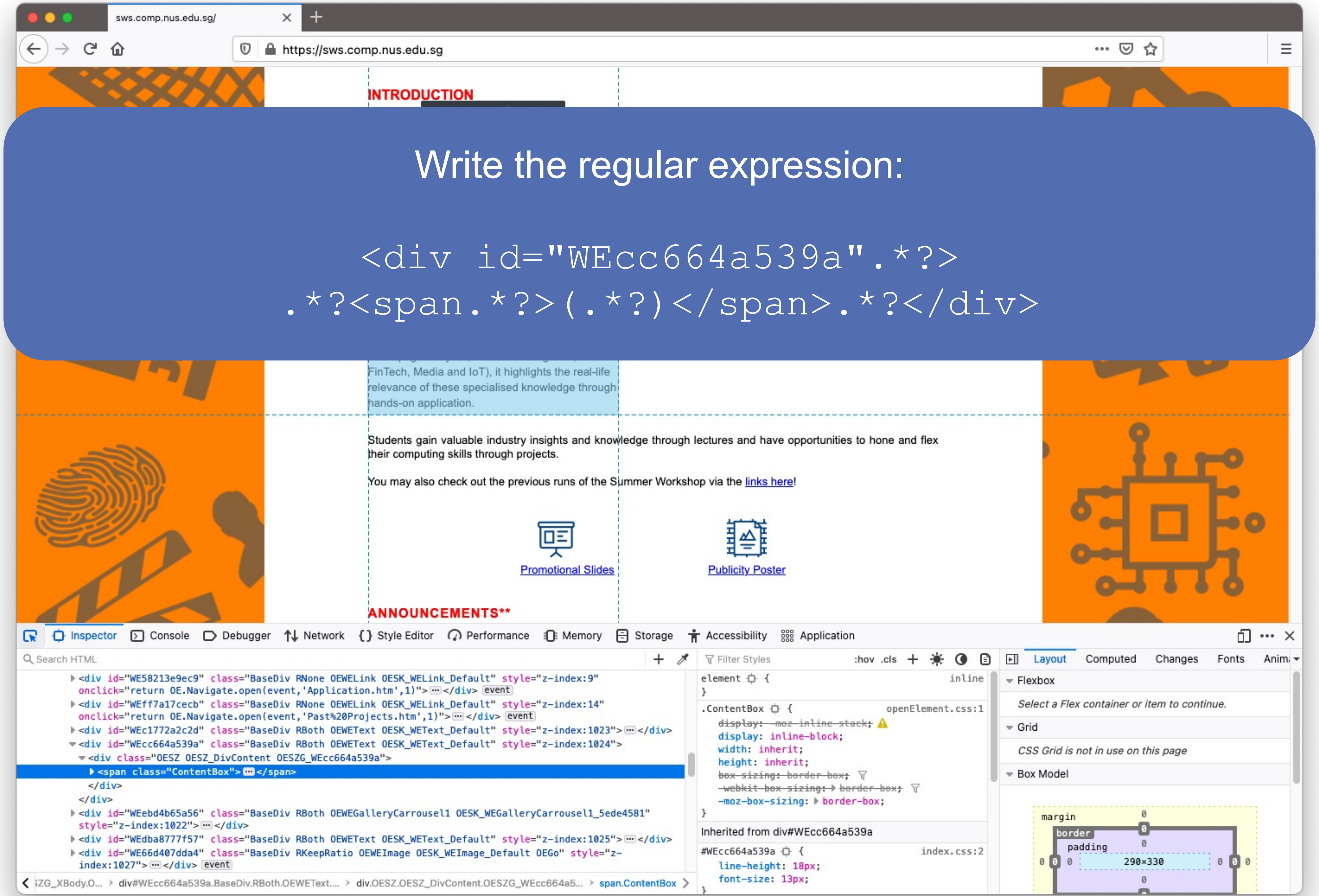
```
▸<div id="WE58213e9ec9" class="BaseDiv RNone OEWELink OESK_WELink_Default" style="z-index:9"
 onclick="return OE.Navigate.open(event,'Application.htm',1)">⋯</div> event
▸<div id="WEff7a17cecb" class="BaseDiv RNone OEWELink OESK_WELink_Default" style="z-index:14"
 onclick="return OE.Navigate.open(event,'Past%20Projects.htm',1)">⋯</div> event
▸<div id="WEc1772a2c2d" class="BaseDiv RBoth OEWEText OESK_WEText_Default" style="z-index:1023">⋯</div>
▾<div id="WEcc664a539a" class="BaseDiv RBoth OEWEText OESK_WEText_Default" style="z-index:1024">
  ▾<div class="OESZ OESZ_DivContent OESZG_WEcc664a539a">
    ▸<span class="ContentBox">⋯</span>
   </div>
 </div>
▸<div id="WEebd4b65a56" class="BaseDiv RBoth OEWEGalleryCarrousel1 OESK_WEGalleryCarrousel1_5ede4581"
 style="z-index:1022">⋯</div>
▸<div id="WEdba8777f57" class="BaseDiv RBoth OEWEText OESK_WEText_Default" style="z-index:1025">⋯</div>
▸<div id="WE66d407dda4" class="BaseDiv RKeepRatio OEWEImage OESK_WEImage_Default OEGo" style="z-
 index:1027">⋯</div> event
```

ZG_XBody.O... > div#WEcc664a539a.BaseDiv.RBoth.OEWEText.... > div.OESZ.OESZ_DivContent.OESZG_WEcc664a5... > span.ContentBox

```
element {                                   inline
}
.ContentBox {                    openElement.css:1
    ~~display: -moz-inline-stack;~~
    display: inline-block;
    width: inherit;
    height: inherit;
    ~~box-sizing: border-box;~~
    ~~-webkit-box-sizing: border-box;~~
    -moz-box-sizing: border-box;
}
Inherited from div#WEcc664a539a
#WEcc664a539a {                      index.css:2
    line-height: 18px;
    font-size: 13px;
}
```

First figure out the rough position in the page using the web browser's **Inspect Element** feature

Then find it View Source window (because the browser might clean up the page slightly such as removing unnecessary spaces)



sws.comp.nus.edu.sg/

https://sws.comp.nus.edu.sg

INTRODUCTION

FinTech, Media and IoT), it highlights the real-life relevance of these specialised knowledge through hands-on application.

Students gain valuable industry insights and knowledge through lectures and have opportunities to hone and flex their computing skills through projects.

You may also check out the previous runs of the Summer Workshop via the links here!

Promotional Slides

Publicity Poster

ANNOUNCEMENTS**

```
<div id="WE58213e9ec9" class="BaseDiv RNone OEWELink OESK_WELink_Default" style="z-index:9"
onclick="return OE.Navigate.open(event,'Application.htm',1)">⋯</div> event
<div id="WEff7a17cecb" class="BaseDiv RNone OEWELink OESK_WELink_Default" style="z-index:14"
onclick="return OE.Navigate.open(event,'Past%20Projects.htm',1)">⋯</div> event
<div id="WEc1772a2c2d" class="BaseDiv RBoth OEWEText OESK_WEText_Default" style="z-index:1023">⋯</div>
<div id="WEcc664a539a" class="BaseDiv RBoth OEWEText OESK_WEText_Default" style="z-index:1024">
  <div class="OESZ OESZ_DivContent OESZG_WEcc664a539a">
    <span class="ContentBox">⋯</span>
  </div>
</div>
<div id="WEebd4b65a56" class="BaseDiv RBoth OEWEGalleryCarrousel1 OESK_WEGalleryCarrousel1_5ede4581"
style="z-index:1022">⋯</div>
<div id="WEdba8777f57" class="BaseDiv RBoth OEWEText OESK_WEText_Default" style="z-index:1025">⋯</div>
<div id="WE66d407dda4" class="BaseDiv RKeepRatio OEWEImage OESK_WEImage_Default OEGo" style="z-
index:1027">⋯</div> event
```

ZG_XBody.O... > div#WEcc664a539a.BaseDiv.RBoth.OEWEText.... > div.OESZ.OESZ_DivContent.OESZG_WEcc664a5... > span.ContentBox

```
element {                                   inline
}
.ContentBox {                    openElement.css:1
  ~~display: -moz-inline-stack;~~
  display: inline-block;
  width: inherit;
  height: inherit;
  ~~box-sizing: border-box;~~
  ~~-webkit-box-sizing: border-box;~~
  -moz-box-sizing: border-box;
}
Inherited from div#WEcc664a539a
#WEcc664a539a {                       index.css:2
  line-height: 18px;
  font-size: 13px;
}
```

Flexbox

Select a Flex container or item to continue.

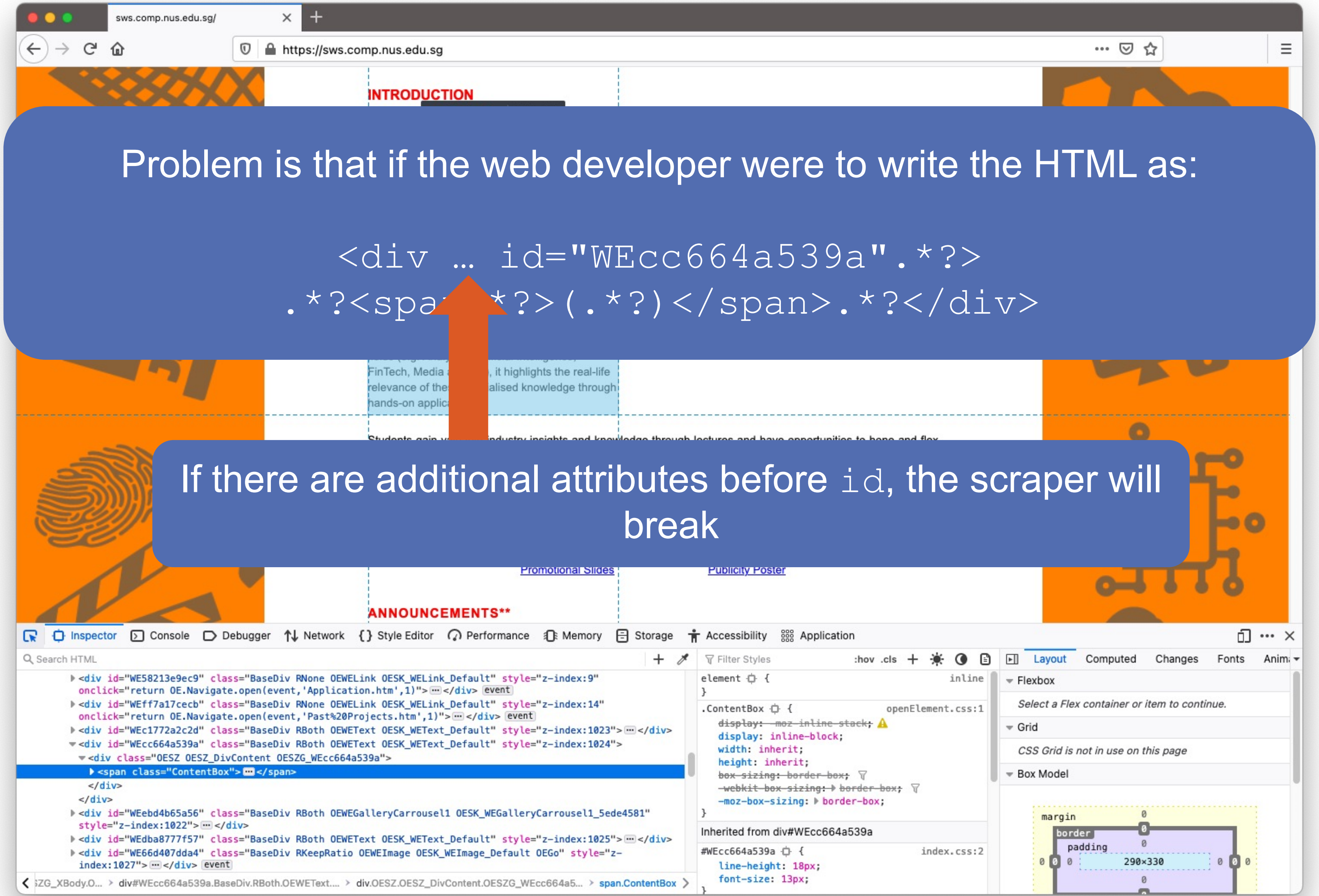
Grid

CSS Grid is not in use on this page

Box Model

Write the regular expression:

```
<div id="WEcc664a539a".*?>
.*?<span.*?>(.*?)</span>.*?</div>
```



Problem is that if the web developer were to write the HTML as:

```
<div … id="WEcc664a539a".*?>
.*?<span.*?>(.*?)</span>.*?</div>
```

If there are additional attributes before `id`, the scraper will break

FinTech, Media … it highlights the real-life relevance of the … alised knowledge through hands-on applic…

```
<div id="WE58213e9ec9" class="BaseDiv RNone OEWELink OESK_WELink_Default" style="z-index:9"
onclick="return OE.Navigate.open(event,'Application.htm',1)">⋯</div> event
<div id="WEff7a17cecb" class="BaseDiv RNone OEWELink OESK_WELink_Default" style="z-index:14"
onclick="return OE.Navigate.open(event,'Past%20Projects.htm',1)">⋯</div> event
<div id="WEc1772a2c2d" class="BaseDiv RBoth OEWEText OESK_WEText_Default" style="z-index:1023">⋯</div>
<div id="WEcc664a539a" class="BaseDiv RBoth OEWEText OESK_WEText_Default" style="z-index:1024">
  <div class="OESZ OESZ_DivContent OESZG_WEcc664a539a">
    <span class="ContentBox">⋯</span>
  </div>
</div>
<div id="WEebd4b65a56" class="BaseDiv RBoth OEWEGalleryCarrousel1 OESK_WEGalleryCarrousel1_5ede4581"
style="z-index:1022">⋯</div>
<div id="WEdba8777f57" class="BaseDiv RBoth OEWEText OESK_WEText_Default" style="z-index:1025">⋯</div>
<div id="WE66d407dda4" class="BaseDiv RKeepRatio OEWEImage OESK_WEImage_Default OEGo" style="z-
index:1027">⋯</div> event
```

ZG_XBody.O… › div#WEcc664a539a.BaseDiv.RBoth.OEWEText.… › div.OESZ.OESZ_DivContent.OESZG_WEcc664a5… › span.ContentBox

```
element {                                   inline
}
.ContentBox {                     openElement.css:1
  ~~display: -moz-inline-stack;~~
  display: inline-block;
  width: inherit;
  height: inherit;
  ~~box-sizing: border-box;~~
  ~~-webkit-box-sizing: border-box;~~
  -moz-box-sizing: border-box;
}
Inherited from div#WEcc664a539a
#WEcc664a539a {                       index.css:2
  line-height: 18px;
  font-size: 13px;
}
```

```
import re
import requests

page = requests.get("https://sw

html3 = page.content.decode("utf-8")

#. by default does not match for newline characters
#re.DOTALL - will make . match even for newline
pattern3 = '<div
 id="WEcc664a539a".*?>.*?<span.*?>(.*?)</span>.*?</div>'
results = re.search(pattern3, html3, re.DOTALL)

extracted_text = results.group(1)

#remove some html
#\\1 - refers to group 1
extracted_text = re.sub("<br />", "\n", extracted_text)
extracted_text = re.sub("&nbsp;", " ", extracted_text)
extracted_text = re.sub("<.*?>(.*?)</.*>", "\\1",
 extracted_text)
```

The page will look exactly same as before but this scraper will no longer work!

FLAW OF STRING-BASED APPROACH OF WEB SCRAPING

Too easily affected by the way how the HTML is written

- Minor changes (e.g. newlines, spaces, capitalization, shifting of attributes ordering, etc) might break the scraper
- Even when the page is still a totally valid page and might look exactly the same

RECAP: TECHNIQUES FOR WEB SCRAPING

The following are some of the techniques for doing web scraping:

- Extracting content from HTML source
- **Extracting content using a HTML parser**
- Web Scraping using APIs
- Scraping using an actual browser/headless browser

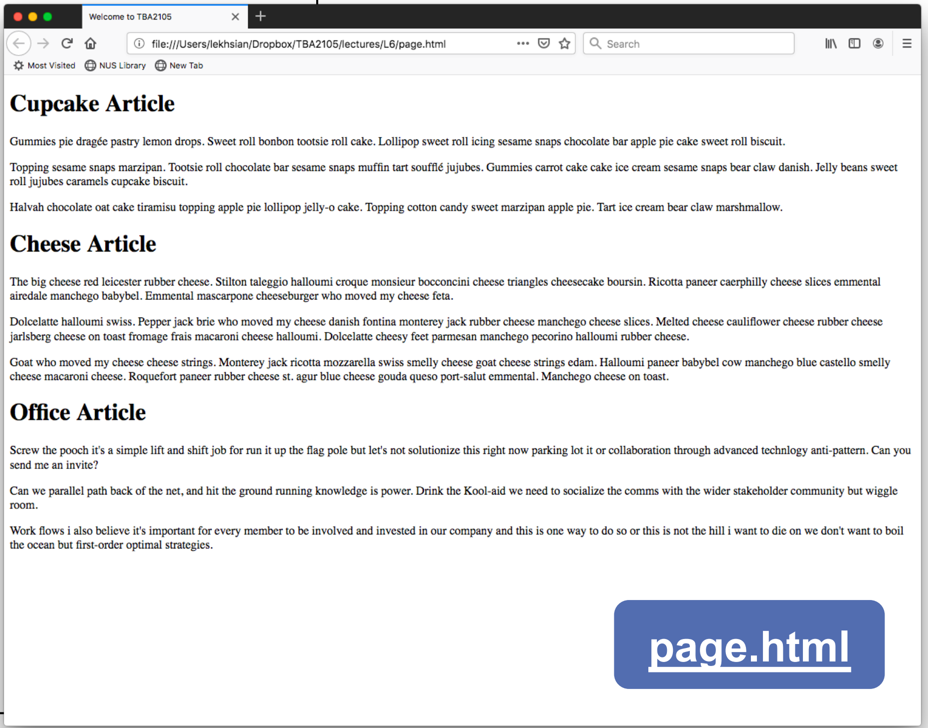
We will look at another approach which is more robust against this situation

DOCUMENT OBJECT MODEL (DOM)

Document Object Model (DOM)
XPath
CSS Selectors
Extracting Content using HTML Parser

```html
<html>
 <head>
  <meta charset="UTF-8" />
  <meta name="description" content="…"/>
  <meta http-equiv="X-UA-Compatible" content="ie=edge" />
  <title>Welcome to SWS3023</title>
 </head>
 <body>
  <div class="article" id="a0042">
   <h1>Cupcake Article</h1>
   <div class="header">…</div>
   <p>…</p>
   <p>…</p>
  </div>
  <div class="article" id="a0043">
   <h1>Cheese Article</h1>
   <div class="header">…</div>
   <p>…</p>
   <p>…</p>
  </div>
  <div class="article special" id="b0051">
   <h1>Office Article</h1>
   <div class="snippet">…</div>
   <p>…</p>
   <p>…</p>
  </div>
 </body>
</html>
```

Recall: HTML defines the content and the layout of the page


Welcome to TBA2105
file:///Users/lekhsian/Dropbox/TBA2105/lectures/L6/page.html
Search
Most Visited
NUS Library
New Tab

Cupcake Article

Gummies pie dragée pastry lemon drops. Sweet roll bonbon tootsie roll cake. Lollipop sweet roll icing sesame snaps chocolate bar apple pie cake sweet roll biscuit.

Topping sesame snaps marzipan. Tootsie roll chocolate bar sesame snaps muffin tart soufflé jujubes. Gummies carrot cake cake ice cream sesame snaps bear claw danish. Jelly beans sweet roll jujubes caramels cupcake biscuit.

Halvah chocolate oat cake tiramisu topping apple pie lollipop jelly-o cake. Topping cotton candy sweet marzipan apple pie. Tart ice cream bear claw marshmallow.

Cheese Article

The big cheese red leicester rubber cheese. Stilton taleggio halloumi croque monsieur bocconcini cheese triangles cheesecake boursin. Ricotta paneer caerphilly cheese slices emmental airedale manchego babybel. Emmental mascarpone cheeseburger who moved my cheese feta.

Dolcelatte halloumi swiss. Pepper jack brie who moved my cheese danish fontina monterey jack rubber cheese manchego cheese slices. Melted cheese cauliflower cheese rubber cheese jarlsberg cheese on toast fromage frais macaroni cheese halloumi. Dolcelatte cheesy feet parmesan manchego pecorino halloumi rubber cheese.

Goat who moved my cheese cheese strings. Monterey jack ricotta mozzarella swiss smelly cheese goat cheese strings edam. Halloumi paneer babybel cow manchego blue castello smelly cheese macaroni cheese. Roquefort paneer rubber cheese st. agur blue cheese gouda queso port-salut emmental. Manchego cheese on toast.

Office Article

Screw the pooch it's a simple lift and shift job for run it up the flag pole but let's not solutionize this right now parking lot it or collaboration through advanced technlogy anti-pattern. Can you send me an invite?

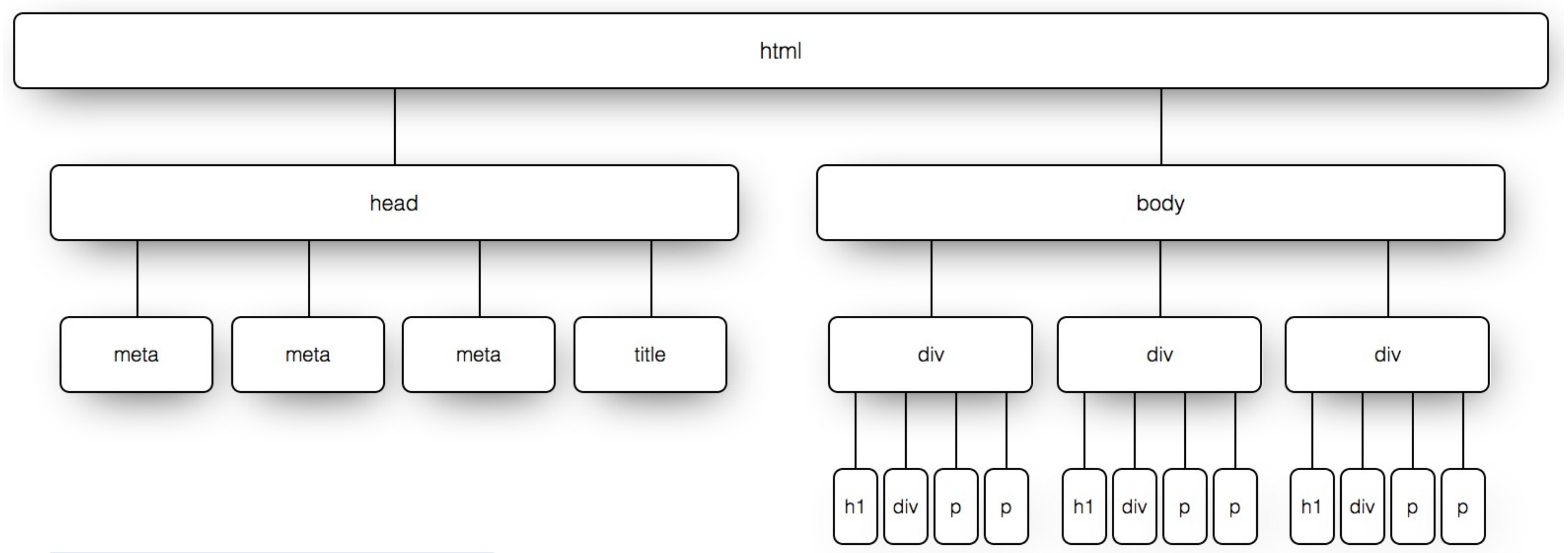
Can we parallel path back of the net, and hit the ground running knowledge is power. Drink the Kool-aid we need to socialize the comms with the wider stakeholder community but wiggle room.

Work flows i also believe it's important for every member to be involved and invested in our company and this is one way to do so or this is not the hill i want to die on we don't want to boil the ocean but first-order optimal strategies.

page.html

DOCUMENT OBJECT MODEL (DOM)

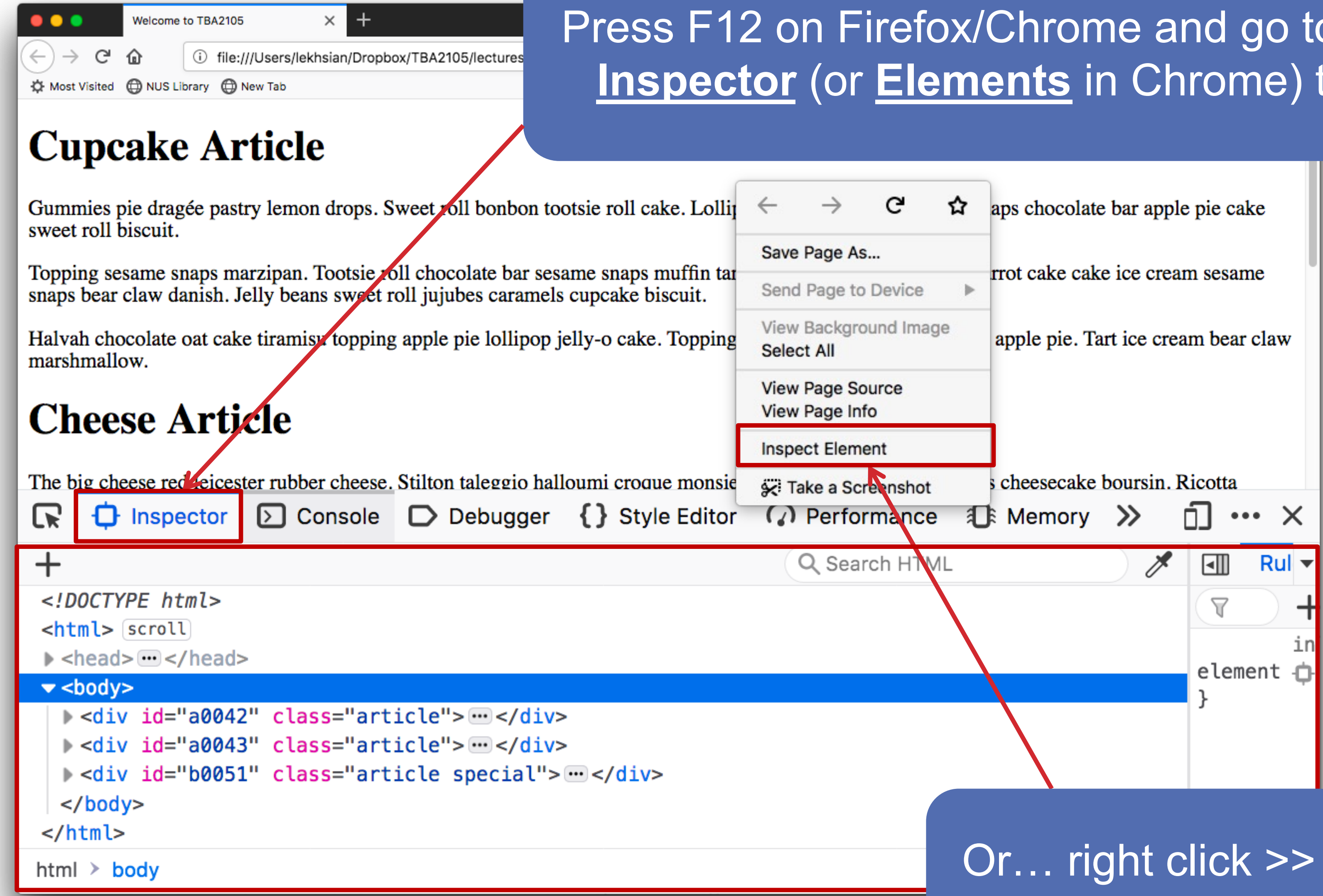
When the browser loads a page, it creates a DOM of the page and use it to render the page



DOM tree of **page.html**

VIEWING THE DOM ON WEB BROWSER

Press F12 on Firefox/Chrome and go to the **Inspector** (or **Elements** in Chrome) tab



Cupcake Article

Gummies pie dragée pastry lemon drops. Sweet roll bonbon tootsie roll cake. Lolli… …aps chocolate bar apple pie cake sweet roll biscuit.

Topping sesame snaps marzipan. Tootsie roll chocolate bar sesame snaps muffin tar… …rrot cake cake ice cream sesame snaps bear claw danish. Jelly beans sweet roll jujubes caramels cupcake biscuit.

Halvah chocolate oat cake tiramisu topping apple pie lollipop jelly-o cake. Topping… …apple pie. Tart ice cream bear claw marshmallow.

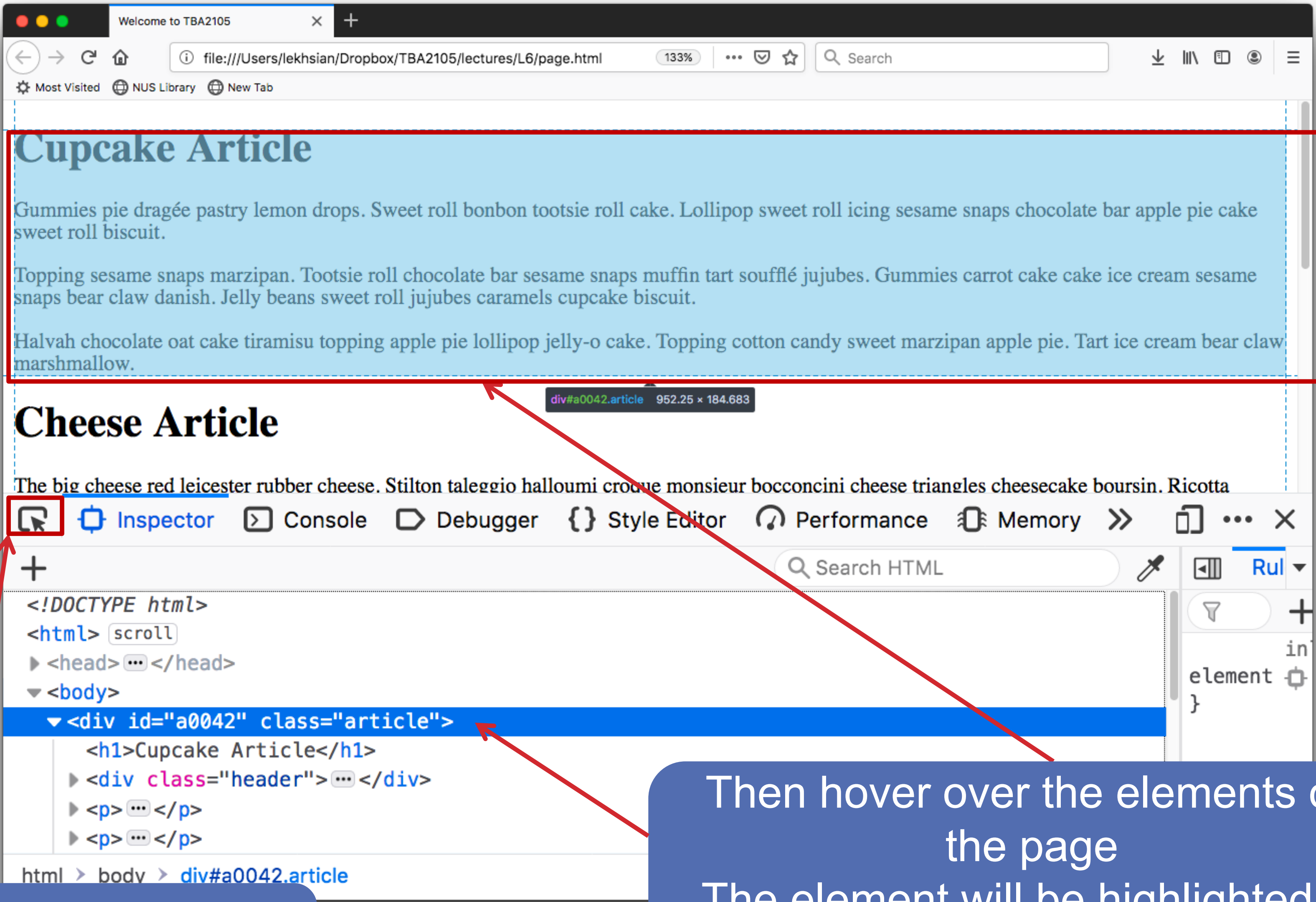
Cheese Article

The big cheese red leicester rubber cheese. Stilton taleggio halloumi croque monsie… …s cheesecake boursin. Ricotta

```
<!DOCTYPE html>
<html> scroll
 <head>…</head>
 <body>
   <div id="a0042" class="article">…</div>
   <div id="a0043" class="article">…</div>
   <div id="b0051" class="article special">…</div>
 </body>
</html>
```

Or… right click >> Inspect Element

VIEWING THE DOM ON WEB BROWSER



Cupcake Article

Gummies pie dragée pastry lemon drops. Sweet roll bonbon tootsie roll cake. Lollipop sweet roll icing sesame snaps chocolate bar apple pie cake sweet roll biscuit.

Topping sesame snaps marzipan. Tootsie roll chocolate bar sesame snaps muffin tart soufflé jujubes. Gummies carrot cake cake ice cream sesame snaps bear claw danish. Jelly beans sweet roll jujubes caramels cupcake biscuit.

Halvah chocolate oat cake tiramisu topping apple pie lollipop jelly-o cake. Topping cotton candy sweet marzipan apple pie. Tart ice cream bear claw marshmallow.

Cheese Article

The big cheese red leicester rubber cheese. Stilton taleggio halloumi croque monsieur bocconcini cheese triangles cheesecake boursin. Ricotta

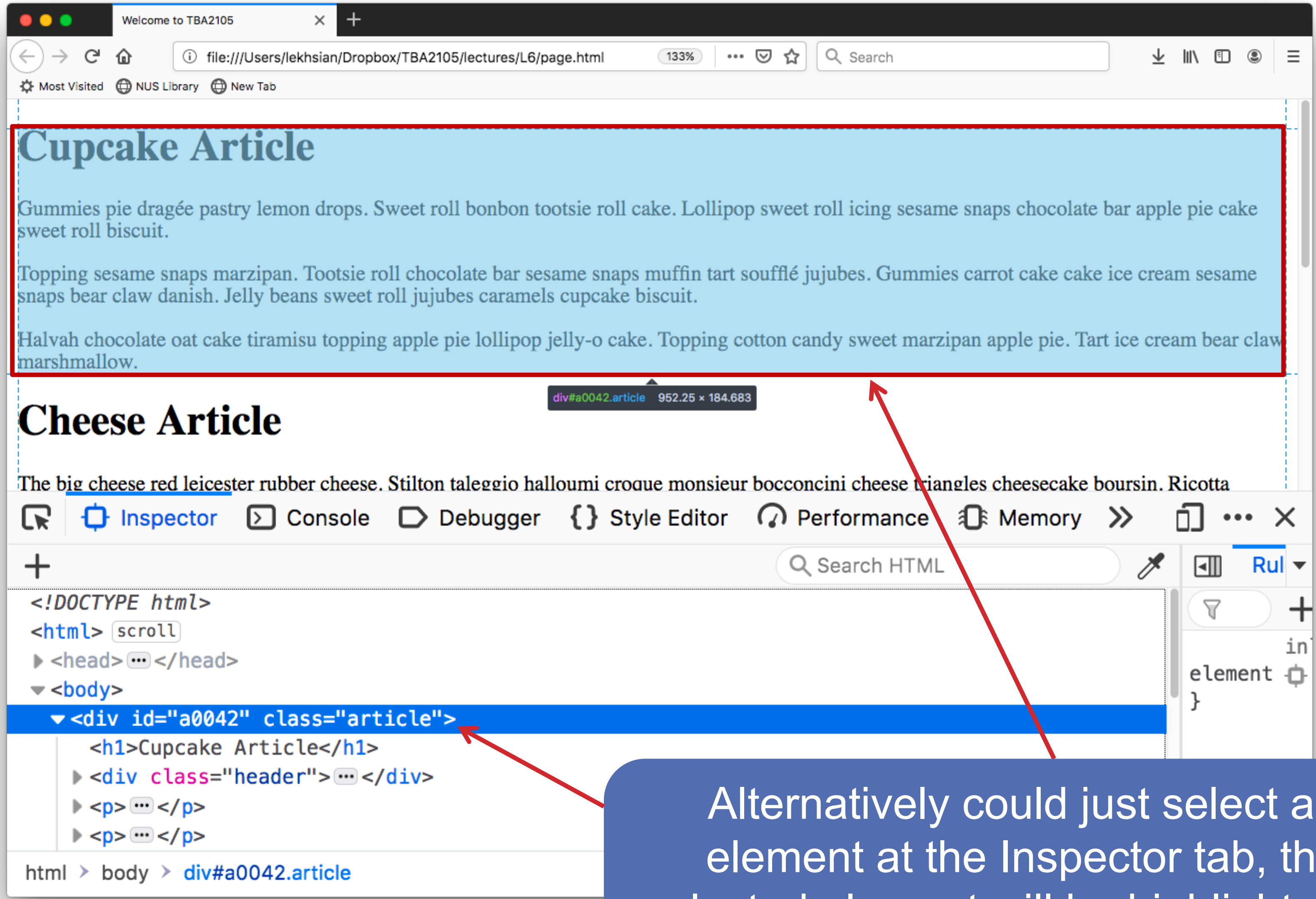
```
<!DOCTYPE html>
<html> scroll
 <head>…</head>
 <body>
  <div id="a0042" class="article">
    <h1>Cupcake Article</h1>
   <div class="header">…</div>
   <p>…</p>
   <p>…</p>
```

html > body > div#a0042.article

First Choose the Pick Element Icon

Then hover over the elements on the page
The element will be highlighted at the Inspector tab

VIEWING THE DOM ON WEB BROWSER



Welcome to TBA2105

file:///Users/lekhsian/Dropbox/TBA2105/lectures/L6/page.html

Cupcake Article

Gummies pie dragée pastry lemon drops. Sweet roll bonbon tootsie roll cake. Lollipop sweet roll icing sesame snaps chocolate bar apple pie cake sweet roll biscuit.

Topping sesame snaps marzipan. Tootsie roll chocolate bar sesame snaps muffin tart soufflé jujubes. Gummies carrot cake cake ice cream sesame snaps bear claw danish. Jelly beans sweet roll jujubes caramels cupcake biscuit.

Halvah chocolate oat cake tiramisu topping apple pie lollipop jelly-o cake. Topping cotton candy sweet marzipan apple pie. Tart ice cream bear claw marshmallow.

div#a0042.article 952.25 × 184.683

Cheese Article

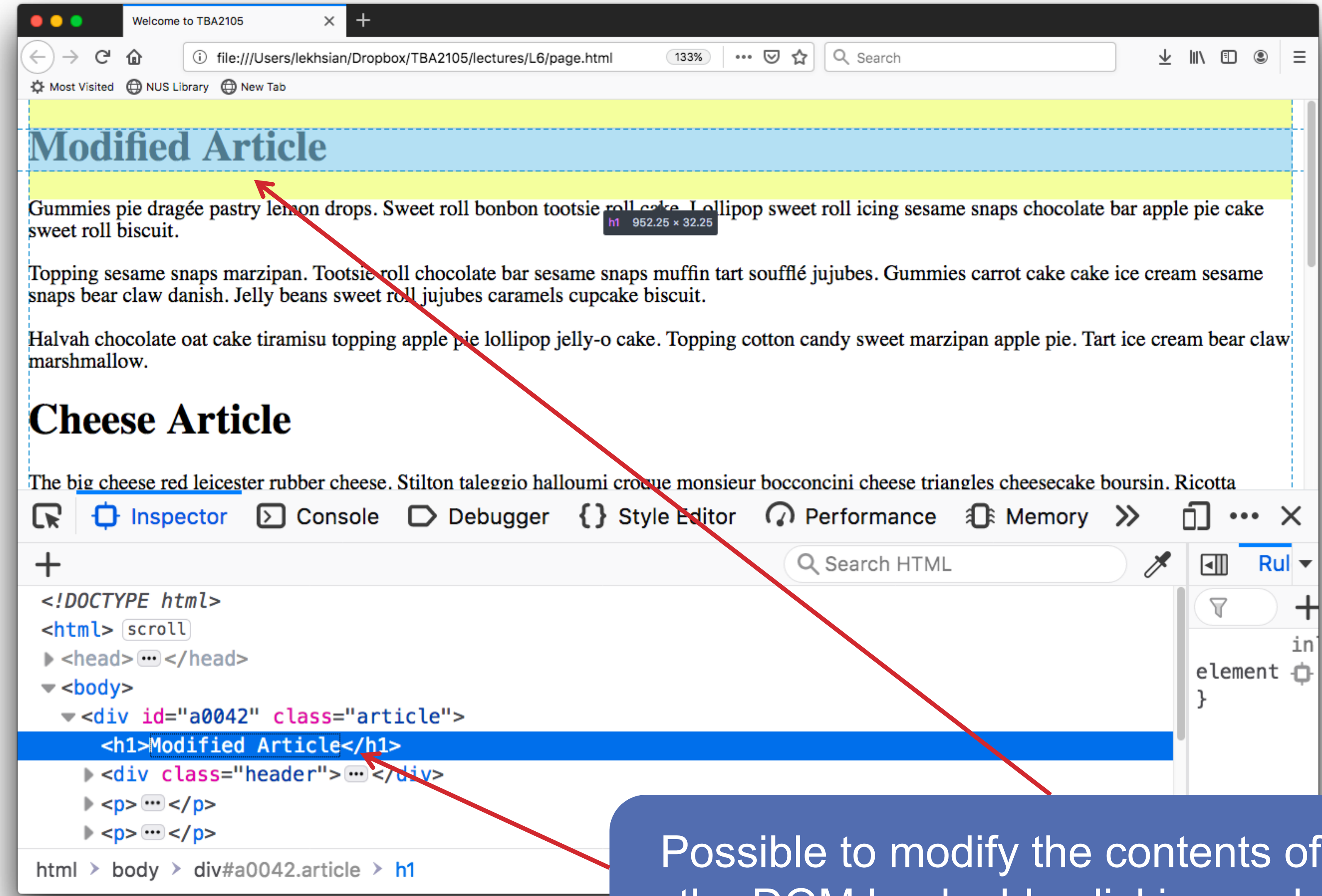
The big cheese red leicester rubber cheese. Stilton taleggio halloumi croque monsieur bocconcini cheese triangles cheesecake boursin. Ricotta

```
<!DOCTYPE html>
<html> scroll
  <head>…</head>
  <body>
    <div id="a0042" class="article">
      <h1>Cupcake Article</h1>
      <div class="header">…</div>
      <p>…</p>
      <p>…</p>
```

html > body > div#a0042.article

Alternatively could just select an element at the Inspector tab, the selected element will be highlighted on the page

MODIFYING THE DOM ON WEB BROWSER



Modified Article

Gummies pie dragée pastry lemon drops. Sweet roll bonbon tootsie roll cake. Lollipop sweet roll icing sesame snaps chocolate bar apple pie cake sweet roll biscuit.

Topping sesame snaps marzipan. Tootsie roll chocolate bar sesame snaps muffin tart soufflé jujubes. Gummies carrot cake cake ice cream sesame snaps bear claw danish. Jelly beans sweet roll jujubes caramels cupcake biscuit.

Halvah chocolate oat cake tiramisu topping apple pie lollipop jelly-o cake. Topping cotton candy sweet marzipan apple pie. Tart ice cream bear claw marshmallow.

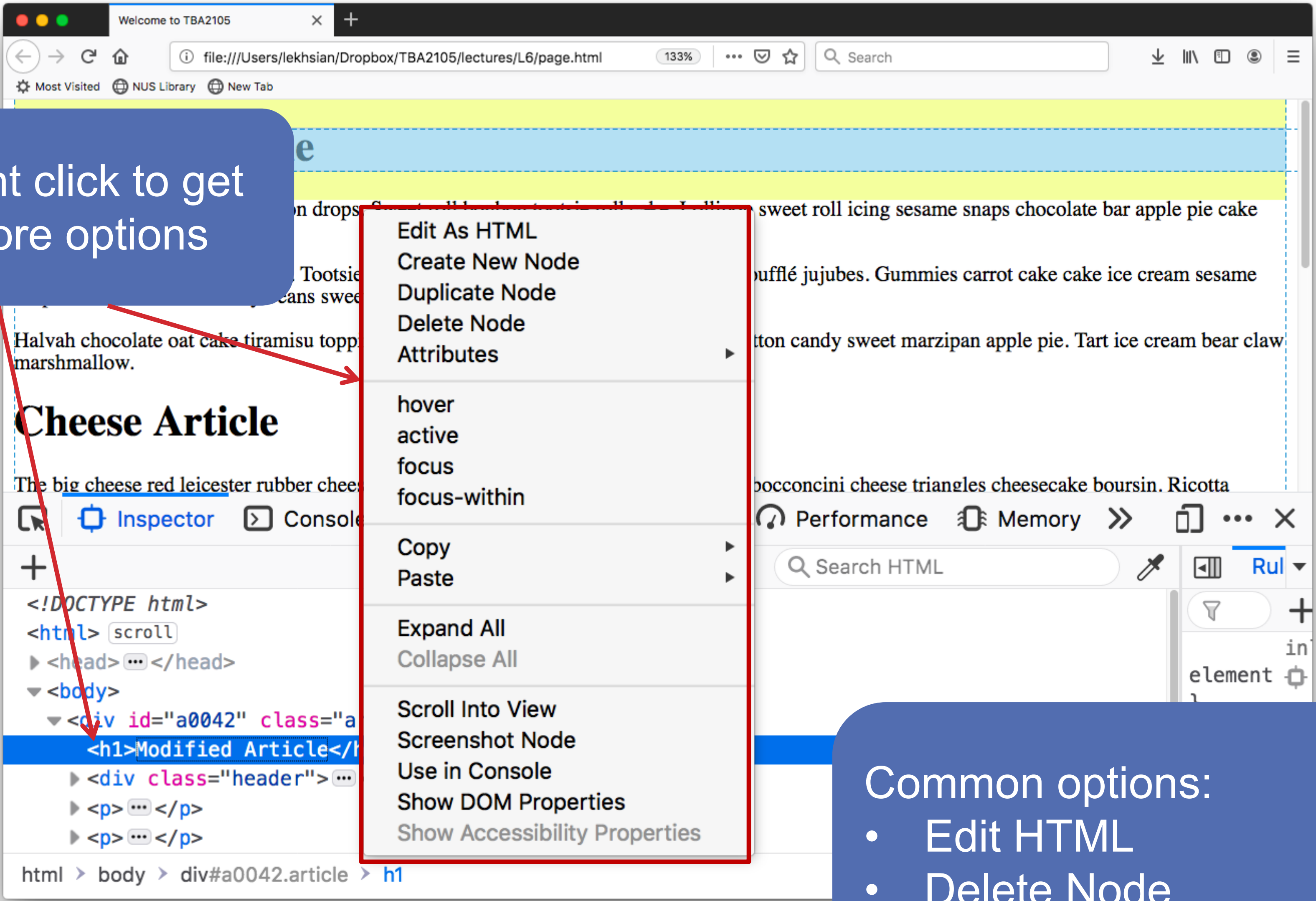
Cheese Article

The big cheese red leicester rubber cheese. Stilton taleggio halloumi croque monsieur bocconcini cheese triangles cheesecake boursin. Ricotta

```
<!DOCTYPE html>
<html> scroll
 ▸ <head>…</head>
 ▾ <body>
   ▾ <div id="a0042" class="article">
      <h1>Modified Article</h1>
    ▸ <div class="header">…</div>
    ▸ <p>…</p>
    ▸ <p>…</p>
```

Possible to modify the contents of the DOM by double clicking and editing

MODIFYING THE DOM ON WEB BROWSER



```
<!DOCTYPE html>
<html> scroll
▸ <head>…</head>
▾ <body>
  ▾ <div id="a0042" class="a
      <h1>Modified Article</h
    ▸ <div class="header">…
    ▸ <p>…</p>
    ▸ <p>…</p>
```

Right click to get more options

Common options:
- Edit HTML
- Delete Node
- Copy (to be elaborated later)

NAVIGATING THE DOM TREE

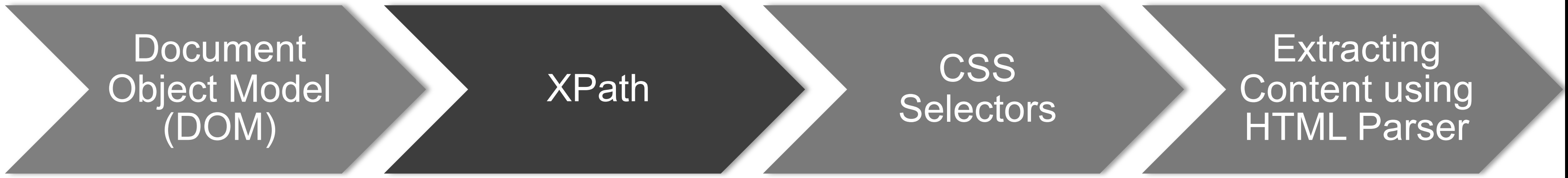
Possible to navigate the DOM programmatically:

- e.g. <u>extract content</u>, add/remove elements, etc

2 ways to navigate the DOM tree

- Using **XPath**
- Using **CSS Selectors**

XPATH


Document Object Model (DOM)
XPath
CSS Selectors
Extracting Content using HTML Parser

XPATH

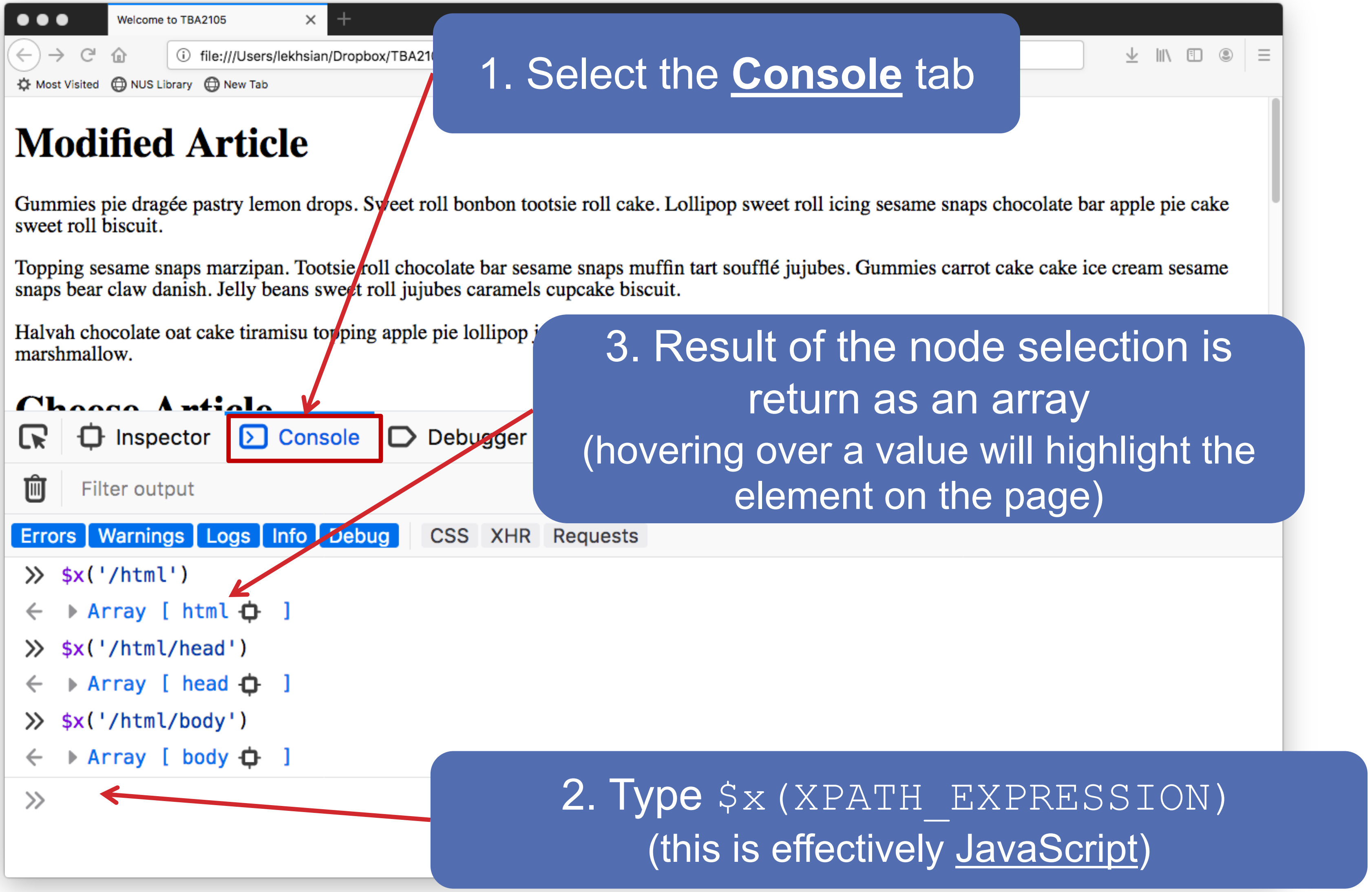
XPath is a way to navigate through elements and attributes in an eXtensible Markup Language (XML) document

- HTML is very similar to a XML document
- Thus, XPath can be used to navigate through HTML documents also

XPATH EXPRESSIONS

XPath Expression	Description
`/html`	Select the **html** node
`/html/head`	Select the **head** node (notice the navigation path)
`/html/body/div`	Select all the **div** directly under **body** node
`/html/body/div[1]`	Select the first **div** directly under **body** node
`//div`	Select all **div** in the document (regardless of its ancestor)
`//div[1]`	Select all the first **div** in the document (originating from an ancestor). Note that this can select multiple div
`//div[last()]`	Select all the last **div** in the document (originating from an ancestor). Note that this can select multiple div
`//body/div`	Select all the **div** nodes that is a **direct child** of the **body** node
`//body//div`	Select all the **div** nodes that is a **descendent** of the **body** node
`//body//*`	Select all the nodes under the **body** node (`*` is a wildcard)

TRYING XPATH ON THE BROWSER



Modified Article

Gummies pie dragée pastry lemon drops. Sweet roll bonbon tootsie roll cake. Lollipop sweet roll icing sesame snaps chocolate bar apple pie cake sweet roll biscuit.

Topping sesame snaps marzipan. Tootsie roll chocolate bar sesame snaps muffin tart soufflé jujubes. Gummies carrot cake cake ice cream sesame snaps bear claw danish. Jelly beans sweet roll jujubes caramels cupcake biscuit.

Halvah chocolate oat cake tiramisu topping apple pie lollipop j marshmallow.

```
>> $x('/html')
<- ▸ Array [ html ]
>> $x('/html/head')
<- ▸ Array [ head ]
>> $x('/html/body')
<- ▸ Array [ body ]
>>
```

1. Select the Console tab

2. Type `$x(XPATH_EXPRESSION)`
(this is effectively JavaScript)

3. Result of the node selection is return as an array
(hovering over a value will highlight the element on the page)

XPATH EXPRESSIONS

Exercise: Try out these expression using the browser **Console**

XPath Expression	Description
`/html`	Select the **html** node
`/html/head`	Select the **head** node (notice the navigation path)
`/html/body/div`	Select all the **div** directly under **body** node
`/html/body/div[1]`	Select the first **div** directly under **body** node
`//div`	Select all **div** in the document (regardless of its ancestor)
`//div[1]`	Select all the first **div** in the document (originating from an ancestor). Note that this can select multiple div
`//div[last()]`	Select all the last **div** in the document (originating from an ancestor). Note that this can select multiple div
`//body/div`	Select all the **div** nodes that is a **direct child** of the **body** node
`//body//div`	Select all the **div** nodes that is a **descendent** of the **body** node
`//body//*`	Select all the nodes under the **body** node (`*` is a wildcard)

XPATH EXPRESSIONS

XPath Expression	Description
`//div[@id]`	Select all **div** node with **id** attribute
`//div[@class='article']`	Select all **div** node with **class** attribute with the value **article** (must be **exact match**, i.e. even with extra spaces will not work)
`//*[@class='article']`	Select **any** node with **class** attribute with the value **article**
`//div[starts-with(@id,'a004')]`	Select all **div** node with **id** attribute having value starting with **a004**
`//div[contains(@id,'00')]`	Select all **div** node with **id** attribute having value containing **00**
`//*[contains(@class,'article')]/..`	Select **parent** node of **any** node with **class** attribute having value containing **article**
`//div | //p`	Select all **div** and **p** nodes

CASE SENSITIVITY

XPath Expression	Case Sensitive?

`//div[@id]`
`//DIV[@id]`
`//div[@ID]`
`//DiV[@ID]`
	All these are equivalent. Both the node name and the attribute names are **not case sensitive**

`//div[@class='article']`
`//div[@class='Article']`
	These are not equivalent. The attribute value is **case sensitive**
	
	Same principle applies to `starts-with()` and `contains()`

XPATH REFERENCES

https://www.w3schools.com/xml/xpath_syntax.asp

CSS SELECTORS

Document Object Model (DOM)
XPath
CSS Selectors
Extracting Content using HTML Parser

CSS SELECTORS

CSS selectors is another way to select elements in HTML

- Used mainly for selecting HTML elements in order to apply styling into the webpage
- But could also be used for referencing to elements in JavaScript or for doing web scraping
- More commonly used compared to XPath
- Tends to be shorter compared to XPath

CSS SELECTORS

CSS selectors is another way to select elements in HTML

- …
- For the purpose of web scraping, these are the most useful type of selectors
    - **Element** selector
    - **Class** selector
    - **Id** selector
    - **Attribute** selector
    - **Pseudo-Classes** selector
    - **Relationship** selector

ELEMENT, CLASS, ID SELECTOR

Element selector used to select all elements with a certain type of tag

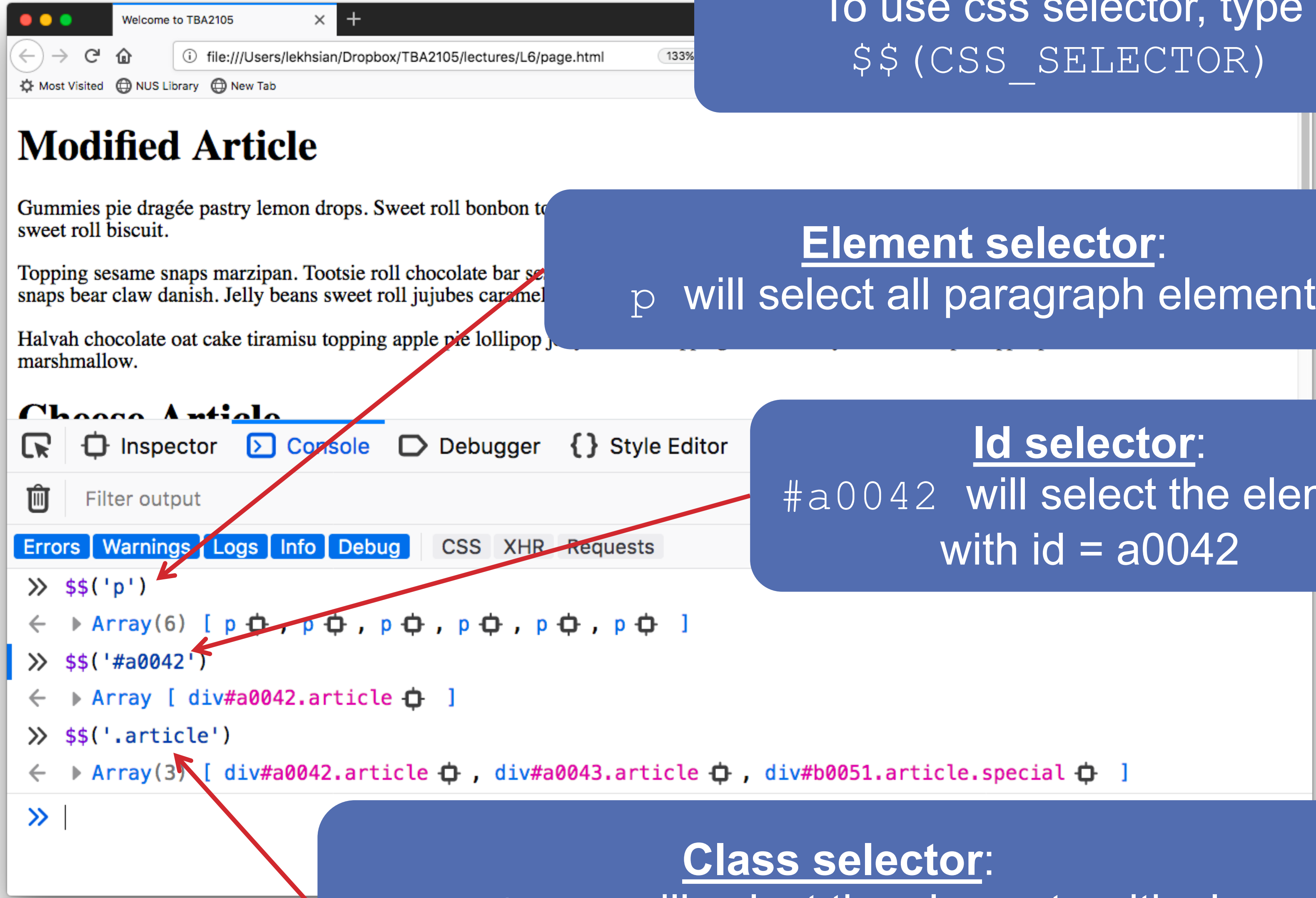
Id selector used to select the element with a certain id value

- HTML usually use `id` to uniquely identify an element

Class selector used to select the element with a certain class attribute value

- HTML elements usually use `class` to denote the style of the element

TRYING CSS SELECTORS



Welcome to TBA2105

file:///Users/lekhsian/Dropbox/TBA2105/lectures/L6/page.html

Most Visited NUS Library New Tab

Modified Article

Gummies pie dragée pastry lemon drops. Sweet roll bonbon to
sweet roll biscuit.

Topping sesame snaps marzipan. Tootsie roll chocolate bar se
snaps bear claw danish. Jelly beans sweet roll jujubes caramel

Halvah chocolate oat cake tiramisu topping apple pie lollipop j
marshmallow.

```
>> $$('p')
<- ▸ Array(6) [ p , p , p , p , p , p ]
>> $$('#a0042')
<- ▸ Array [ div#a0042.article ]
>> $$('.article')
<- ▸ Array(3) [ div#a0042.article , div#a0043.article , div#b0051.article.special ]
>> |
```

To use css selector, type
`$$(CSS_SELECTOR)`

Element selector:
`p` will select all paragraph elements

Id selector:
`#a0042` will select the element with id = a0042

Class selector:
`.article` will select the elements with class containing article class

CSS SELECTORS

Possible to multiple multiple type of selectors

CSS Selector	Description
`div#a0042`	This selects the **div** node with **id = a0042**
`div.article`	This selects all the **div** nodes with **class** containing **article** (need not be an exact match, the element can have multiple class values separated by spaces)
`.article#a0042`	This selects all nodes with **class** containing **article** and with **id = a0042**
`.article.special`	This selects all nodes with **class** containing both **article** and **special** (can be any ordering but be careful that no spaces between the 2 classes)

`div,p`
`div , p`
`div, p`
	This selects all the **div** and **p** nodes

ATTRIBUTE SELECTORS

CSS Selector	Description
`meta[name]`	This selects the **meta** node with a **name** attribute
`meta[name='description']`	This selects the **meta** node with a **name** attribute with the value **description** (exact match)
`meta[content~='page']`	This selects the **meta** node with a **content** attribute containing a **page** as a <u>whole word match</u>. Would not match somepage.
`meta[content*='learn']`	This selects the **meta** node with a **content** attribute containing **learn**. Matches as long as there is a **substring** of learn.

`[class = 'article']`
`*[class = 'article']`
	This selects all node with **class** attribute that is **article** (exact match)

`[id ^= 'a00']`	This selects all node with **id** attribute that <u>starts with</u> **a00**
`[id $= '51']`	This selects all node with **id** attribute that <u>ends with</u> **a00**

PSEUDO-CLASSES SELECTORS

CSS Selector	Description
`div:first-child`	This selects all **div** nodes that is the 1st child of its parent
`p:last-child`	This selects all **p** nodes that is the last child of its parent
`p:nth-child(3)`	This selects all **p** nodes that is the 3rd node of its parent

RELATIONSHIP SELECTORS

CSS Selector	Description
`body div`	This selects all **div** nodes that are descendents of **body**
`body > div`	This selects all **div** nodes that are direct child of **body**
`h1 + *`	This selects all nodes that is a next sibling of **h1**
`.article .snippet`	This will find all nodes having **class** containing **article**, and select all descendent nodes having **class** containing **snippet**. Notice that there is a space separating the 2 classes

Can mix with the other selectors discussed before

CASE SENSITIVITY

CSS Selector	Case Sensitive?

`DIV`
`Div`
`DiV`
	These are equivalent. Tag names are **not case sensitive**

`.article`
`.Article`
	These are not equivalent. Classes are **case sensitive**

`#a0042` `#A0042`	Likewise, these are not equivalent. Ids are **case sensitive**

`meta[name]`
`meta[Name]`
	These are equivalent. The attribute names are **not case sensitive**

`meta[name='description']`
`meta[name='Description']`
	These are not equivalent. Attribute values are **case sensitive** (as principles as class and id)

CSS SELECTOR REFERENCES

https://www.w3schools.com/cssref/css_selectors.asp

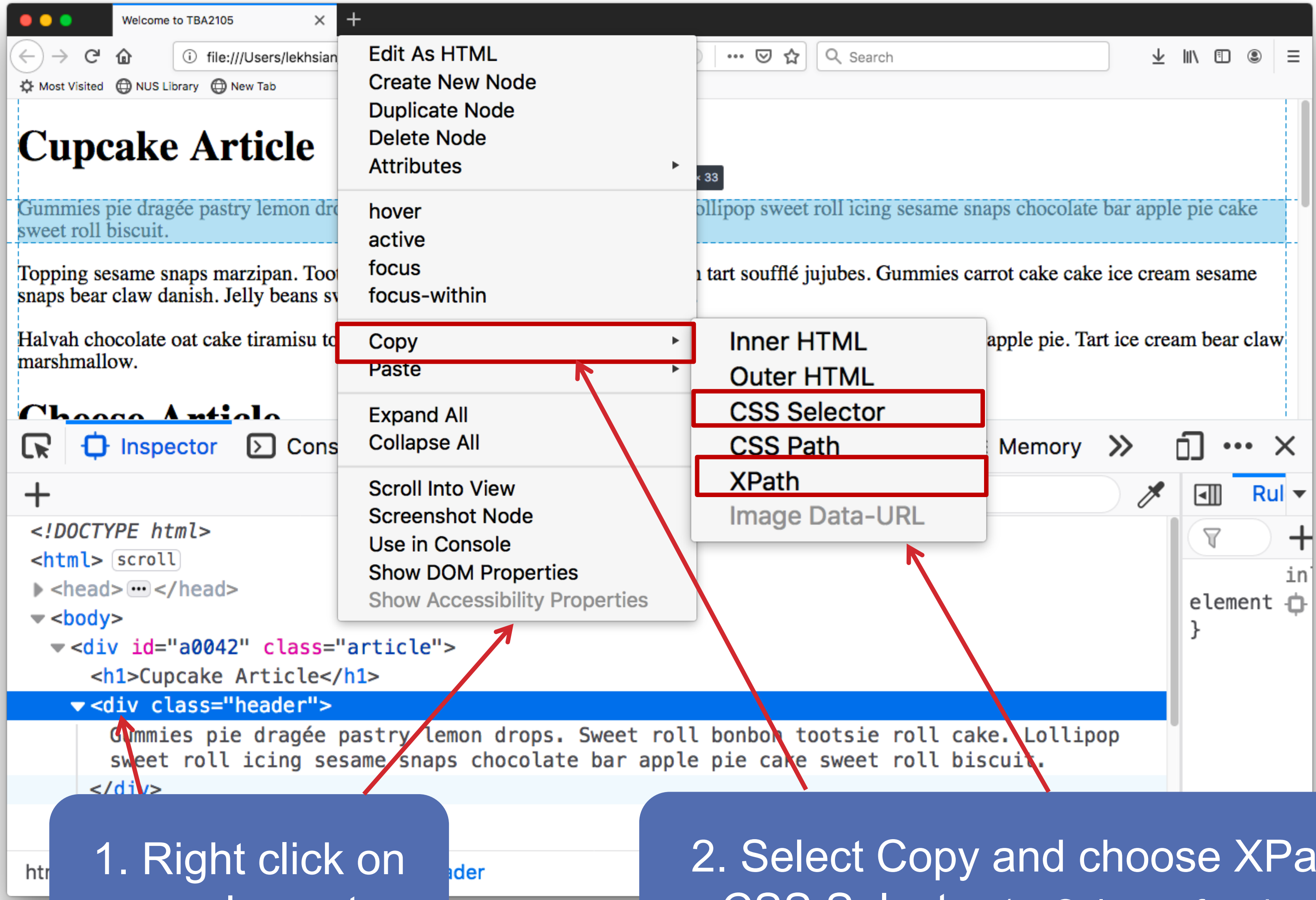
XPath vs CSS Selectors:

https://johnresig.com/blog/xpath-css-selectors/

Not confident of writing your own XPath/CSS Selectors?

- The browser also allows you to copy the XPath/CSS Selector expressions

COPY XPATH/CSS SELECTOR



Gummies pie dragée pastry lemon dr ollipop sweet roll icing sesame snaps chocolate bar apple pie cake sweet roll biscuit.

Topping sesame snaps marzipan. Too tart soufflé jujubes. Gummies carrot cake cake ice cream sesame snaps bear claw danish. Jelly beans sv

Halvah chocolate oat cake tiramisu to apple pie. Tart ice cream bear claw marshmallow.

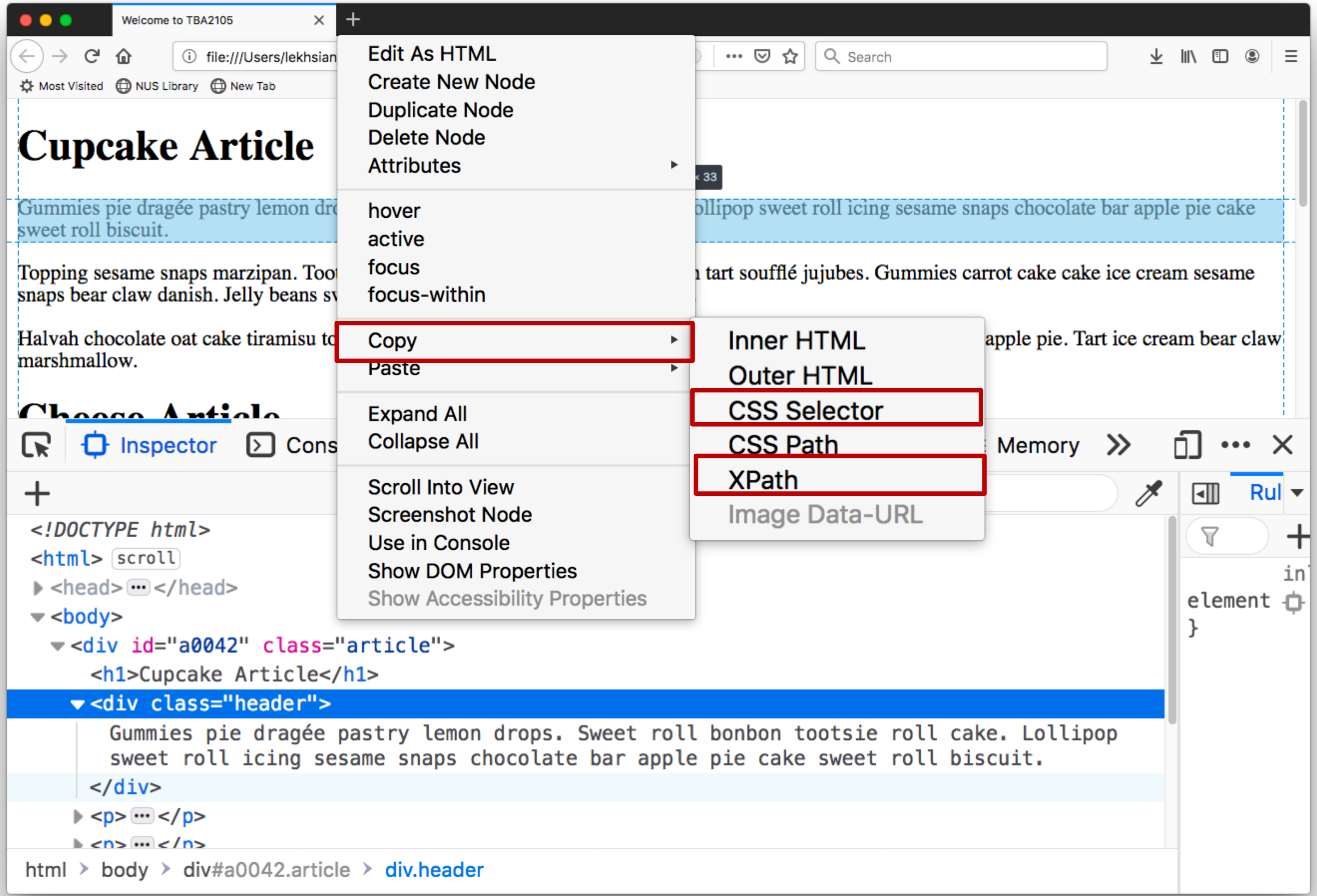
```
<!DOCTYPE html>
<html> scroll
  <head>…</head>
  <body>
    <div id="a0042" class="article">
      <h1>Cupcake Article</h1>
      <div class="header">
        Gummies pie dragée pastry lemon drops. Sweet roll bonbon tootsie roll cake. Lollipop
        sweet roll icing sesame snaps chocolate bar apple pie cake sweet roll biscuit.
      </div>
```

1. Right click on an element

2. Select Copy and choose XPath or CSS Selector (or Selector for chrome)

COPY XPA

Note that these expressions tend to be very long and overly specific (still better to write manually)



Cupcake Article

Gummies pie dragée pastry lemon dr… …ollipop sweet roll icing sesame snaps chocolate bar apple pie cake sweet roll biscuit.

Topping sesame snaps marzipan. Too… …n tart soufflé jujubes. Gummies carrot cake cake ice cream sesame snaps bear claw danish. Jelly beans s…

Halvah chocolate oat cake tiramisu t… …apple pie. Tart ice cream bear claw marshmallow.

Cheese Article

```
<!DOCTYPE html>
<html> scroll
 <head>…</head>
 <body>
  <div id="a0042" class="article">
    <h1>Cupcake Article</h1>
   <div class="header">
      Gummies pie dragée pastry lemon drops. Sweet roll bonbon tootsie roll cake. Lollipop
      sweet roll icing sesame snaps chocolate bar apple pie cake sweet roll biscuit.
    </div>
   <p>…</p>
   <p>…</p>
```

EXTRACTING CONTENT USING HTML PARSER

Document Object Model (DOM) → XPath → CSS Selectors → **Extracting Content using HTML Parser**

HTML PARSING

Different programming platform has its own library for doing HTML parsing

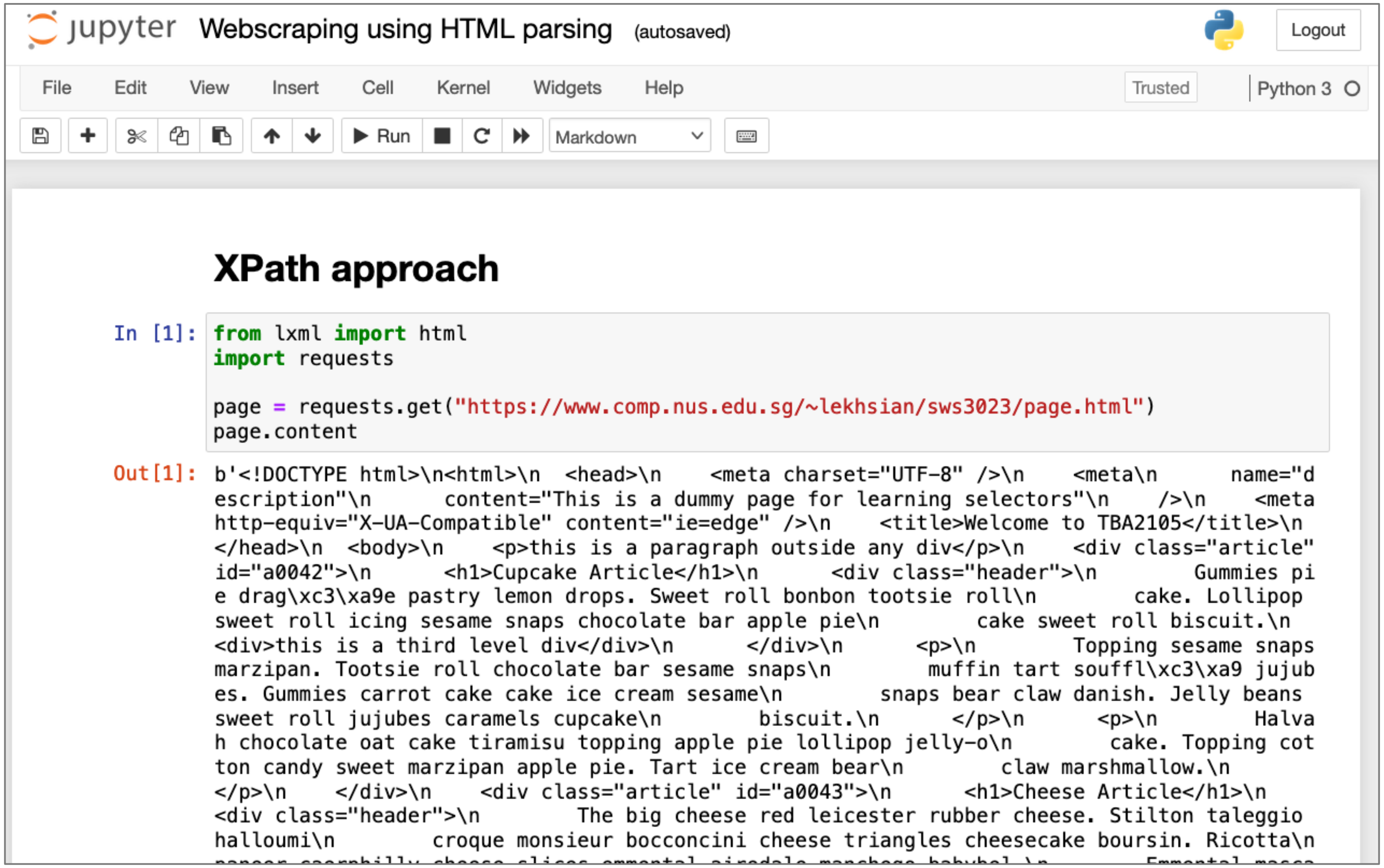
- After parsing, we are able to select the HTML elements using XPath and/or CSS selectors
- Can use the **lxml** (for XPath) and **BeautifulSoup** packages

Idea:

- Parse the HTML document
- Use XPath/CSS Selector to select the element(s)
- Extract the value (attribute/text/html) of the element(s)

HANDS-ON: WEB SCRAPING

Download and access:
Webscraping using HTML parsing.ipynb


jupyter
Webscraping using HTML parsing (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Markdown

XPath approach

In [1]:
```
from lxml import html
import requests

page = requests.get("https://www.comp.nus.edu.sg/~lekhsian/sws3023/page.html")
page.content
```

Out[1]:
```
b'<!DOCTYPE html>\n<html>\n  <head>\n    <meta charset="UTF-8" />\n    <meta\n      name="d
escription"\n      content="This is a dummy page for learning selectors"\n    />\n    <meta
http-equiv="X-UA-Compatible" content="ie=edge" />\n    <title>Welcome to TBA2105</title>\n
</head>\n  <body>\n    <p>this is a paragraph outside any div</p>\n    <div class="article"
id="a0042">\n      <h1>Cupcake Article</h1>\n      <div class="header">\n        Gummies pi
e drag\xc3\xa9e pastry lemon drops. Sweet roll bonbon tootsie roll\n        cake. Lollipop
sweet roll icing sesame snaps chocolate bar apple pie\n        cake sweet roll biscuit.\n
<div>this is a third level div</div>\n      </div>\n      <p>\n        Topping sesame snaps
marzipan. Tootsie roll chocolate bar sesame snaps\n        muffin tart souffl\xc3\xa9 jujub
es. Gummies carrot cake cake ice cream sesame\n        snaps bear claw danish. Jelly beans
sweet roll jujubes caramels cupcake\n        biscuit.\n      </p>\n      <p>\n        Halva
h chocolate oat cake tiramisu topping apple pie lollipop jelly-o\n        cake. Topping cot
ton candy sweet marzipan apple pie. Tart ice cream bear\n        claw marshmallow.\n
</p>\n    </div>\n    <div class="article" id="a0043">\n      <h1>Cheese Article</h1>\n
<div class="header">\n        The big cheese red leicester rubber cheese. Stilton taleggio
halloumi\n        croque monsieur bocconcini cheese triangles cheesecake boursin. Ricotta\n
```

SUMMARY

Document Object Model (DOM)

Navigating the DOM tree

- XPath and CSS Selectors

Extracting Content using HTML Parser

- lxml & BeautifulSoup packages

WHAT'S NEXT?

Mining Web Content III