

Introduction to web scraping

Warmup

Work on the warmup activity (handout), then we will discuss as a class.

```
read_html("https://fake-web-page.html") |>
```

```
  html_element("table.wikitable2") |>
```

↑ or: "table[class =
 'wikitable2']"

```
  html_table()
```

Motivation: Great British Bake Off



Accessing HTML attributes

https://en.wikipedia.org/wiki/The_Great_British_Bake_Off

Series 1 (2010) [[edit](#)]

Main article: [The Great British Bake Off \(series 1\)](#)

Accessing HTML attributes

https://en.wikipedia.org/wiki/The_Great_British_Bake_Off

Series 1 (2010) [\[edit\]](#)

Main article: [The Great British Bake Off \(series 1\)](#)

s 1 (2010)  258.47 × 18.5

in article: [The Great British Bake Off \(series 1\)](#)

```
"Main article: "  
...  
<a href="/wiki/The_Great_British_Bake_Off/Series_1" title="The Great British Bake Off Series 1"> == $0
```

<a> : hyperlink element

href : url for the hyperlink

Accessing HTML attributes

(Demo)

Accessing HTML attributes

```
1 gbbo <- read_html("https://en.wikipedia.org/wiki/The_Great_British_Bake_Off")
2 gbbo |>
3   html_elements("div.hatnote.navigation-not-searchable > a") |>
4   html_attr("href")
```

```
[1] "/wiki/Great_Basin_Bird_Observatory"
[2] "/wiki/List_of_The_Great_British_Bake_Off_contestants"
[3]
"/wiki/List_of_The_Great_British_Bake_Off_finalists_(series_1%E2%80%937)"
[4]
"/wiki/List_of_The_Great_British_Bake_Off_finalists_(series_8%E2%80%93present)"
[5] "/wiki/The_Great_British_Bake_Off_series_1"
[6] "/wiki/The_Great_British_Bake_Off_series_2"
[7] "/wiki/The_Great_British_Bake_Off_series_3"
[8] "/wiki/The_Great_British_Bake_Off_series_4"
[9] "/wiki/The_Great_British_Bake_Off_series_5"
[10] "/wiki/The_Great_British_Bake_Off_series_6"
[11] "/wiki/The_Great_British_Bake_Off_series_7"
```

Accessing HTML attributes

```
1 gbbo |>  
2   html_elements("div.hatnote.navigation-not-searchable > a") |>  
3   html_text2()
```

```
[1] "Great Basin Bird Observatory"  
[2] "List of contestants"  
[3] "List of finalists (series 1–7)"  
[4] "List of finalists (series 8–present)"  
[5] "The Great British Bake Off (series 1)"  
[6] "The Great British Bake Off (series 2)"  
[7] "The Great British Bake Off (series 3)"  
[8] "The Great British Bake Off (series 4)"  
[9] "The Great British Bake Off (series 5)"  
[10] "The Great British Bake Off (series 6)"  
[11] "The Great British Bake Off (series 7)"  
[12] "The Great British Bake Off (series 8)"  
[13] "The Great British Bake Off (series 9)"
```


Class activity

Spend the remainder of the class period on the class activity (CSS Diner: practice with CSS selectors). You do not need to submit anything.