

# Lecture 18: Intro to web scraping

# Motivation: Taskmaster



# Show information

- 16 full series
- Each series involves 5 contestants, competing over 5-10 episodes
- Each episode involves approximately 5 tasks
- Contestants are scored from 1-5 (roughly) on each task

# Taskmaster data

1 [https://taskmaster.fandom.com/wiki/Series\\_11](https://taskmaster.fandom.com/wiki/Series_11)

The screenshot shows a table from the Taskmaster Wiki for Episode 1: It's not your fault. (18 March 2021). The table lists four tasks with their descriptions and scores for five contestants: Charlotte, Jamali, Lee, Mike, and Sarah. The scores are: Task 1 (Prize: Best thing you can carry, but only just) - 1, 2, 4, 5, 3; Task 2 (Do the most impressive thing under the table with one hand. You must be looking at the camera and waving with your other hand throughout your impressive thing.) - 2, 3<sup>[1]</sup>, 3, 5, 4; Task 3 (Catch the rat. You must be at least three meters from the rat when you catch it. The rat will run across the red green in 30 minutes.) - DQ, 1, 5, 3<sup>[2]</sup>, 3<sup>[2]</sup>; Task 4 (Deliver all the plates to Alex. When carrying the plates you must be traveling by scooter, or bicycle or hoverboard. For every plate fallen) - 2, 1, 5, 3<sup>[2]</sup>, 4<sup>[2]</sup>. A 'Spoilers!' warning box is visible above the table.

Task	Description	Charlotte	Jamali	Lee	Mike	Sarah
		Ritchie	Maddix	Mack	Wozniak	Kendall
Episode 1: It's not your fault. (18 March 2021)						
1	Prize: Best thing you can carry, but only just.	1	2	4	5	3
2	Do the most impressive thing under the table with one hand. You must be looking at the camera and waving with your other hand throughout your impressive thing.	2	3 <sup>[1]</sup>	3	5	4
3	Catch the rat. You must be at least three meters from the rat when you catch it. The rat will run across the red green in 30 minutes.	DQ	1	5	3 <sup>[2]</sup>	3 <sup>[2]</sup>
4	Deliver all the plates to Alex. When carrying the plates you must be traveling by scooter, or bicycle or hoverboard. For every plate fallen	2	1	5	3 <sup>[2]</sup>	4 <sup>[2]</sup>

# Goal and required steps

**Goal:** Explore the Taskmaster data across all completed series. Which contestants did worst? Which contestants did best? Did the scoring change over the series?

- Scrape data from each series from the website
- Combine, clean, and transform
- Do statistics!

# Scraping the data

```
1 library(rvest)
2 library(tidyverse)    reads HTML info from URL into R
3
4 tm <- read_html("https://taskmaster.fandom.com/wiki/Series_11")
5 tm
```

```
{html_document}
<html class="client-nojs" lang="en" dir="ltr">
[1] <head>\n<meta http-equiv="Content-Type" content="text/html;
charset=UTF-8 ...
[2] <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-
subject ...
```

# HTML basics

Here is a basic HTML page:

```
1 <html> ← start of HTML
2 <head>   ↗ title element
3 <title>Page title</title>
4 </head>   ↗ content
5 <body>   ↗ start tag
6 <h1 id='first'>A heading</h1>   ↗ heading
7 <p>Some text &amp; some bold text.</b></p>   ↗ bold
8 <img src='myimg.png' width='100' height='100'>
9 </body>   ↗ end tag
```

Start tag  
paragraph element

end tag  
content  
start tag  
heading  
bold  
for Source, width, height  
attributes

nl biggest  
then h2, etc..  
img: image element

<head> : metadata about page

<body> : what is displayed on the page

A heading

Some text



Some bold text.

# Some HTML elements

- <html>: start of the HTML page
- <head>: header information (metadata about the page)
- <body>: everything that is on the page
- <p>: paragraphs
- <b>: bold
- <table>: table

# Extracting HTML elements

The Taskmaster data we want looks like it is stored in a table. How can we extract it?

```
1 tm |>
2   html_elements("table") ← return all of the "table" elements
{xml_nodeset (4)}           in our html file
[1] <table style="width: 100%; text-align: center; border: 1px solid
#891100; ...
[2] <table class="toccoours" align="center" style="background: #891100;
colo ...
[3] <table class="pi-horizontal-group">\n<caption class="pi-header pi-
seconda ...
[4] <table class="tmtable"><tbody>\n<tr
class="tmtableheader">\n<th>Task</th> ...
```

html\_elements returns all the elements matching the selector.

↑ nest function

# Extracting HTML elements

```
1 tm |>
2 html_elements("table")

{xml_nodeset (4)}
[1] <table style="width: 100%; text-align: center; border: 1px solid
#891100; ...
[2] <table class="toccoours" align="center" style="background: #891100;
colo ...
[3] <table class="pi-horizontal-group">\n<caption class="pi-header pi-
seconda ...
[4] <table class="tmtable"><tbody>\n<tr
class="tmtableheader">\n<th>Task\n</t ...
```

How do we know which table we want?

# Finding the right selectors

1. Open the webpage in Chrome
2. Right-click on the element you want, and click “Inspect”

The screenshot shows a webpage with a table of tasks. The table has columns for Task, Description, and five judges' names: Charlotte, Jamali, Lee, Mike, and Sarah. The first task is "Prize: Best thing you can carry, but only just." The second task is "Do the most impressive thing under the table with one hand. You must be looking at the camera and waving with your other hand throughout your impressive thing." The third task is "Catch the rat. You must be at least three meters from the rat when you catch it. The rat will run across the red green in 30 minutes." The developer tools Elements tab is open, highlighting the table element with a black oval. The table's class is "tmttable". The table structure is shown in the tree view, including the header row and individual cells.

Task	Description	Charlotte Ritchie	Jamali Maddix	Lee Mack	Mike Wozniak	Sarah Kendall
Episode 1: It's not your fault. (18 March 2021)						
1	Prize: Best thing you can carry, but only just.	1	2	4	5	3
2	Do the most impressive thing under the table with one hand. You must be looking at the camera and waving with your other hand throughout your impressive thing.	2	3 <sup>[1]</sup>	3	5	4
3	Catch the rat. You must be at least three meters from the rat when you catch it. The rat will run across the red green in 30 minutes.	DQ	1	5	3 <sup>[2]</sup>	3 <sup>[2]</sup>

<table      class = "tmttable" >

# Finding the right selector

```
1 tm |>  
2   html_element("[class='tmttable ']") |>  
3   html_table()  
  
# A tibble: 75 × 7  
  Task           Description `Charlotte Ritchie` `Jamali Maddix`  
  <chr>          <chr>          <chr>          <chr>  
  1 Lee Mack  
<chr>  
    1 Episode 1: It's n... Episode 1:... Episode 1: It's no... Episode 1: It'...  
    Episode 1...  
    2 1             Prize: Bes... 1               2               4  
    3 2             Do the mos... 2               3[1]            3  
    4 3             Catch the ... DQ            1               5  
    5 4             Deliver al... 2               1               5  
    6 5             Live: Stac... 0               0               0  
    7 Total         Total           5               7               17  
    8 Episode 2: The Lu... Episode 2:... Episode 2: The Lur... Episode 2: The...
```

import the table element  
(with class = "tmttable")  
into R

formatting as a data frame

# Finding the right selector

```
1 tm |> . for classes
2   html_element(".tmtable") |> .tmtable means [class="tmtable"]
3   html_table()

# A tibble: 75 × 7
  Task           Description `Charlotte Ritchie` `Jamali Maddix`  

  <chr>          <chr>       <chr>          <chr>
  `Lee Mack`  

  <chr>          <chr>       <chr>          <chr>
  <chr>
  1 Episode 1: It's n... Episode 1:... Episode 1: It's no... Episode 1: It'...
  Episode 1...
  2 1             Prize: Bes... 1           2           4
  3 2             Do the mos... 2           3[1]        3
  4 3             Catch the ... DQ        1           5
  5 4             Deliver al... 2           1           5
  6 5             Live: Stac... 0           0           0
  7 Total         Total            5           7           17
  8 Episode 2: The Lu... Episode 2:... Episode 2: The Lur... Episode 2: The...
```

# Extracting non-tabular data

1 [https://taskmaster.fandom.com/wiki/Charlotte\\_Ritchie](https://taskmaster.fandom.com/wiki/Charlotte_Ritchie)

**Charlotte Ritchie** is an English actress and singer-songwriter. She is a member of the classical crossover group [All Angels](#). She is also known for her roles in [BBC One](#) sitcom [Ghosts](#), [Channel 4](#) comedy-drama [Fresh Meat](#), [BBC Three](#) sitcom [siblings](#), [E4](#) sitcom [Dead Pixels](#) and BBC One period drama [Call the Midwife](#).

In 2021, Ritchie appeared as a contestant on [Series 11](#) of [Taskmaster](#), finishing last with 125 points. She was the guest on episodes 30 and 129 of [Taskmaster The Podcast](#).

**Contents [hide]**

- 1. Performance
- 2. Trivia
- 3. Video
- 4. External Links

**Charlotte Ritchie**

**Series 11**

<b>Score</b>	125
<b>Place</b>	5th
<b>Profile</b>	
<b>Born</b>	29 August 1989
<b>Twitter</b>	@Charitchie

**Contestant Guide**

How would we scrape Charlotte Ritchie's birthday?

# Identifying the right selector

```
▼<div class="pi-item pi-data pi-item-spacing pi-border-color"  
      data-source="born"> flex == $0  
      <h3 class="pi-data-label pi-secondary-font">Born</h3>  
      <div class="pi-data-value pi-font">29 August 1989</div>
```

```
1 read_html("https://taskmaster.fandom.com/wiki/Charlotte_Ritchie") |>  
2 html_element("[data-source='born']")
```

```
{html_node}  
<div class="pi-item pi-data pi-item-spacing pi-border-color" data-  
source="born">  
[1] <h3 class="pi-data-label pi-secondary-font">Born</h3>  
[2] <div class="pi-data-value pi-font">29 August 1989</div>
```

# Identifying the right selector

```
▼<div class="pi-item pi-data pi-item-spacing pi-border-color"  
      data-source="born"> flex == $0  
      <h3 class="pi-data-label pi-secondary-font">Born</h3>  
      <div class="pi-data-value pi-font">29 August 1989</div>
```

```
1 read_html("https://taskmaster.fandom.com/wiki/Charlotte_Ritchie") |>  
2   html_element("[data-source='born'] > .pi-font")
```

```
{html_node}      element where  
<div class="pi-data-value pi-font">  
    data-source = 'born'  
    ↑  
    class = 'pi-font'  
  
nesting  
  
element with      class      pi-font      nested inside  
an element       with      data-source = 'born'
```

# Identifying the right selector

```
▼<div class="pi-item pi-data pi-item-spacing pi-border-color"  
      data-source="born"> flex == $0  
      <h3 class="pi-data-label pi-secondary-font">Born</h3>  
      <div class="pi-data-value pi-font">29 August 1989</div>
```

```
1 read_html("https://taskmaster.fandom.com/wiki/Charlotte_Ritchie") |>  
2   html_element("[data-source='born'] > .pi-font") |>  
3   html_text2() ← turn HTML element into text in R
```

[1] "29 August 1989"

(extracting just the text)

(call then convert to date-time format if desired)

# Accessing HTML attributes

## Series 11



EDIT



Prev · 1 · 2 · 3 · 4 · 5 · CoC · 6 · 7 · 8 · 9 · 10 · NYT · **11** · 12 · NYT II · 13 · CoC II · 14 · NYT III · 15  
· 16 · Next

It's not your fault. • The Lure of the Treacle Puppies. • Run up a tree to the moon. •  
Premature conker. • Slap and tong. • Absolute casserole. • You've got no chutzpah. • An  
orderly species. • Mr Octopus and Pottyhands. • Activate Jamali.

```
▼<td align="center">
  ▼<span style="font-family: Veteran Typewriter;">
    <a href="/wiki/It%27s_not_your_fault." title="It's not your fault.">It's not your fault.</a>
  </span>
  " • "
  ▼<span style="font-family: Veteran Typewriter;"> == $0
    <a href="/wiki/The_Lure_of_the_Treacle_Puppies." title="The Lure of the Treacle Puppies.">The Lure of the Treacle Puppies.</a>
  </span>
  " • "
  ▶
```

# Accessing HTML attributes

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>
2   html_elements("td")

{xml_nodeset (436)}
[1] <td>\n<table class="toccoours" align="center" style="background:
#89110 ...
[2] <td>\n<a href="/wiki/Series_10" title="Series 10"><span
style="color: #F ...
[3] <td align="center">\n<span style="font-family: Veteran
Typewriter;"><a h ...
[4] <td class="pi-horizontal-group-item pi-data-value pi-font pi-
border-colo ...
[5] <td class="pi-horizontal-group-item pi-data-value pi-font pi-
border-colo ...
[6] <td colspan="7">Episode 1: <span style="font-family: Veteran
Typewriter; ...
[7] <td>\n<a href="/wiki/Best_thing_you_can_carry,_but_only_just"
```

# Accessing HTML attributes

```
▼<td align="center">
  ▼<span style="font-family: Veteran Typewriter;">
    <a href="/wiki/It%27s not your fault." title="It's not y
      our fault.">It's not your fault.</a>
  </span>
  " • "
  ▼<span style="font-family: Veteran Typewriter;"> == $0
    <a href="/wiki/The Lure of the Treacle Puppies." title=
      "The Lure of the Treacle Puppies.">The Lure of the
      Treacle Puppies.</a>
  </span>
  " • "
  ▼<span style="font-family: Veteran Typewriter;">
```

# Accessing HTML attributes

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>  
2 html_elements("td[align='center']")  
  
{xml_nodeset (1)}  
[1] <td align="center">\n<span style="font-family: Veteran Typewriter;">  
<a hr ...
```

# Accessing HTML attributes

```
▼<td align="center">
  ▼<span style="font-family: Veteran Typewriter;">
    <a href="/wiki/It%27s not your fault." title="It's not y
      our fault.">It's not your fault.</a>
  </span>
  " • "
▼<span style="font-family: Veteran Typewriter;"> == $0
  <a href="/wiki/The Lure of the Treacle Puppies." title=
    "The Lure of the Treacle Puppies.">The Lure of the
    Treacle Puppies.</a>
  </span>
  " • "
▶<span style="font-family: Veteran Typewriter;"></span>
```

# Accessing HTML attributes

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>
2   html_elements("td[align='center'] > span")  
  
{xml_nodeset (10)}  
[1] <span style="font-family: Veteran Typewriter;"><a  
href="/wiki/It%27s_not ...  
[2] <span style="font-family: Veteran Typewriter;"><a  
href="/wiki/The_Lure_o ...  
[3] <span style="font-family: Veteran Typewriter;"><a  
href="/wiki/Run_up_a_t ...  
[4] <span style="font-family: Veteran Typewriter;"><a  
href="/wiki/Premature_ ...  
[5] <span style="font-family: Veteran Typewriter;"><a  
href="/wiki/Slap_and_t ...  
[6] <span style="font-family: Veteran Typewriter;"><a  
href="/wiki/Absolute_c ...  
[7] <span style="font-family: Veteran Typewriter;"><a
```

# Accessing HTML attributes

```
▼<td align="center">
  ▼<span style="font-family: Veteran Typewriter;">
    <a href="/wiki/It%27s not your fault." title="It's not y
      our fault.">It's not your fault.</a>
  </span>
  " • "
  ▼<span style="font-family: Veteran Typewriter;"> == $0
    <a href="/wiki/The Lure of the Treacle Puppies." title=
      "The Lure of the Treacle Puppies.">The Lure of the
      Treacle Puppies.</a>
  </span>
  " • "
  ▼<span style="font-family: Veteran Typewriter;">
```

# Accessing HTML attributes

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>
2   html_elements("td[align='center'] > span") |>
3   html_element("a")  
  
{xml_nodeset (10)}  
[1] <a href="/wiki/It%27s_not_your_fault." title="It's not your  
fault.">It's ...  
[2] <a href="/wiki/The_Lure_of_the_Treacle_Puppies." title="The Lure of  
the ...  
[3] <a href="/wiki/Run_up_a_tree_to_the_moon." title="Run up a tree to  
the m ...  
[4] <a href="/wiki/Premature_conker." title="Premature  
conker.">Premature co ...  
[5] <a href="/wiki/Slap_and_tong." title="Slap and tong.">Slap and  
tong.</a>  
[6] <a href="/wiki/Absolute_casserole." title="Absolute  
casserole.">Absolute ...  
[7] <a href="/wiki/You%27ve_got_no_chutzpah." title="You've got no
```

# Accessing HTML attributes

```
▼<td align="center">
  ▼<span style="font-family: Veteran Typewriter;">
    <a href="/wiki/It%27s not your fault." title="It's not y
      our fault.">It's not your fault.</a>
  </span>
  " • "
▼<span style="font-family: Veteran Typewriter;"> == $0
  <a href="/wiki/The Lure of the Treacle Puppies." title=
    "The Lure of the Treacle Puppies.">The Lure of the
    Treacle Puppies.</a>
  </span>
  " • "
▶<span style="font-family: Veteran Typewriter;"></span>
```

# Accessing HTML attributes

```
1 read_html("https://taskmaster.fandom.com/wiki/Series_11") |>
2   html_elements("td[align='center'] > span") |>
3   html_element("a") |>
4   html_attr("href")  
  
[1] "/wiki/It%27s_not_your_fault."  
[2] "/wiki/The_Lure_of_the_Treacle_Puppies."  
[3] "/wiki/Run_up_a_tree_to_the_moon."  
[4] "/wiki/Premature_conker."  
[5] "/wiki/Slap_and_tong."  
[6] "/wiki/Absolute_casserole."  
[7] "/wiki/You%27ve_got_no_chutzpah."  
[8] "/wiki/An_orderly_species."  
[9] "/wiki/Mr_Octopus_and_Pottyhands."  
[10] "/wiki/Activate_Jamali."
```

# Class activity

[https://sta279-s24.github.io/class\\_activities/ca\\_lecture\\_18.html](https://sta279-s24.github.io/class_activities/ca_lecture_18.html)

