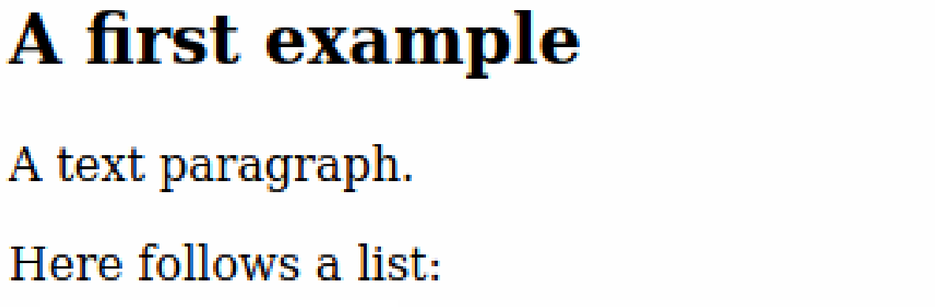
Hypertext Markup Language (HTML)

```
<html>
  <body>
    <h2>A first example</h2>
    <p>A text paragraph.</p>
    <p>
      Here follows a list:
    </p>
  </body>
</html>
```



HTML is organized hierarchically

A first example

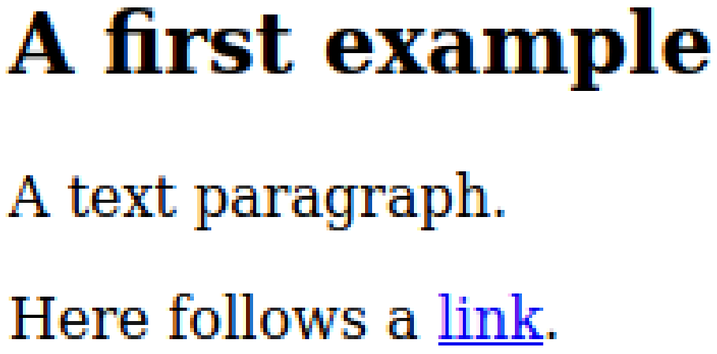
A text paragraph.

Here follows a list:

- Bullet 1
- Bullet 2
- Bullet 3

```
...
    <p>
      Here follows a list:
      <ul>
        <li>Bullet 1</li>
        <li>Bullet 2</li>
        <li>Bullet 3</li>
      </ul>
    </p>
...
```

HTML tags can have attributes




```
...
    <p>
      Here follows a
      <a href="https://google.com">link</a>.
    </p>
...
```

Reading HTML with R

```
library(rvest)
```

```
html <- read_html(html_document)
html
```

```
{html_document}
<html>
[1] <body> \n    <h2>A first example</h2>\n    <p>A text paragraph.</p>\n   ...
```

```
class(html)
```

```
"xml_document" "xml_node"
```

```
xml_structure(html)
```

```
<html>
  <body>
    {text}
    <h2>
      {text}
    {text}
    <p>
      {text}
    {text}
    <p>
      {text}
      <a [href]>
        {text}
      {text}
    {text}
```

Let's parse HTML!

WEB SCRAPING IN R

Navigating HTML

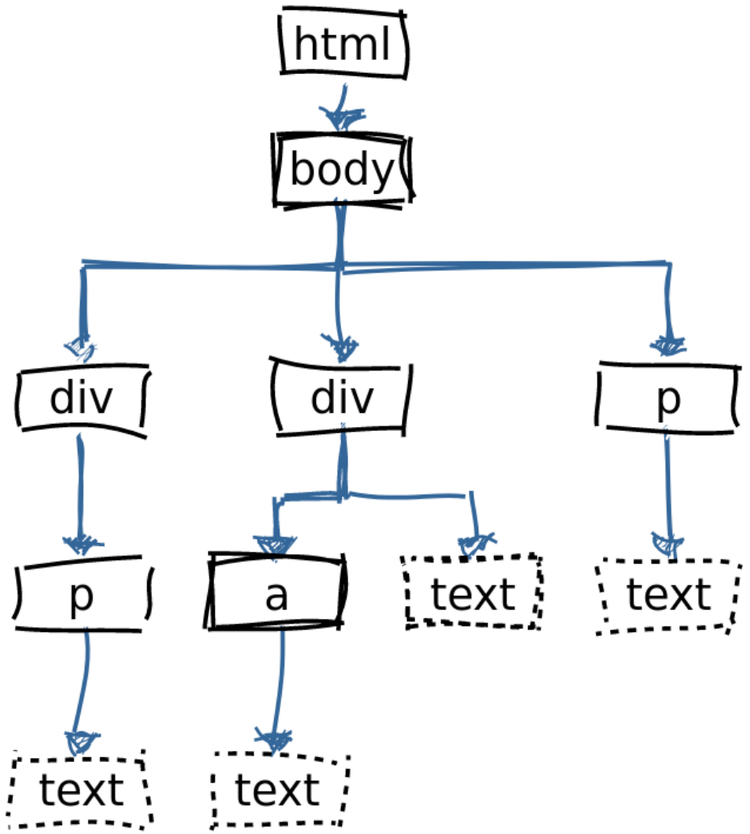
WEB SCRAPING IN R

Timo Grossenbacher
Instructor

datacamp

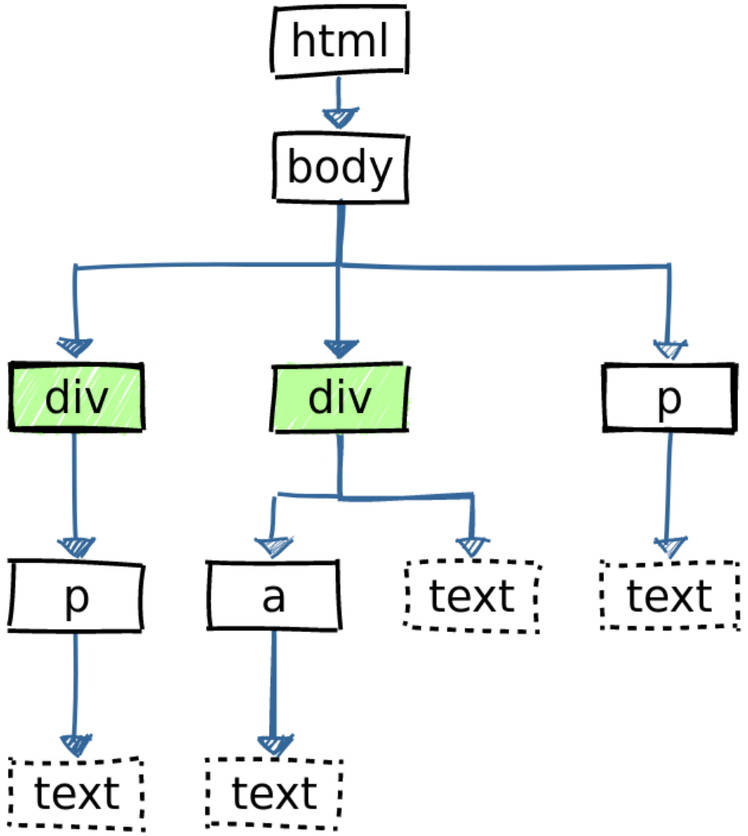
HTML is like a tree

```
<html>
  <body>
    <div>
      <p>The first paragraph.</p>
    </div>
    <div>
      Not an actual paragraph,
      but with a <a href="#">link</a>.
    </div>
    <p>A paragraph without an
      enclosing div.</p>
  </body>
</html>
```



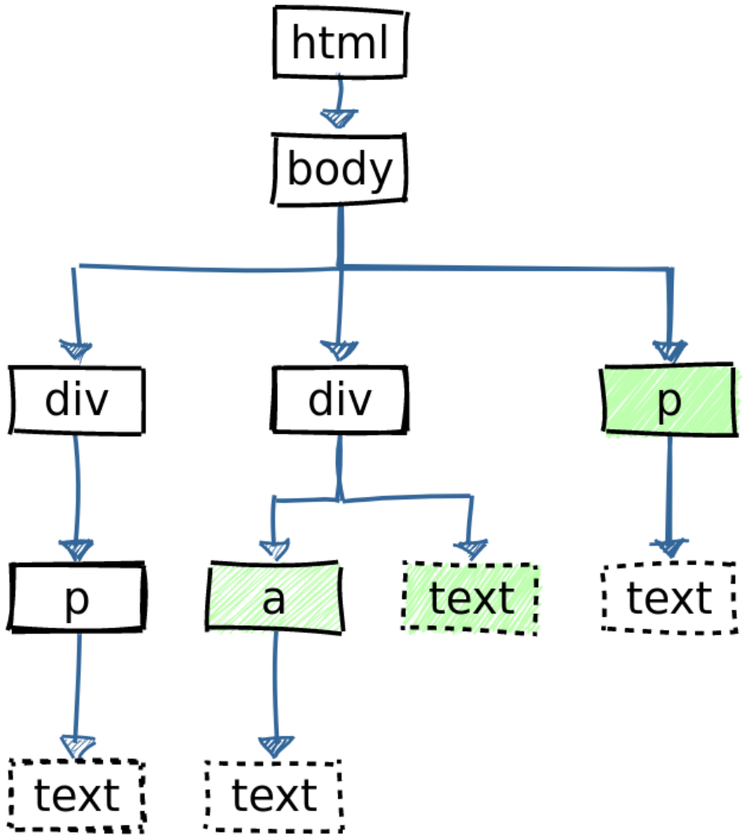
HTML is like a tree

```
<html>
  <body>
    <div>
      <p>The first paragraph.</p>
    </div>
    <div>
      Not an actual paragraph,
      but with a <a href="#">link</a>.
    </div>
    <p>A paragraph without an
      enclosing div.</p>
  </body>
</html>
```



HTML is like a tree

```
<html>
  <body>
    <div>
      <p>The first paragraph.</p>
    </div>
    <div>
      Not an actual paragraph,
      but with a <a href="#">link</a>.
    </div>
    <p>A paragraph without an
      enclosing div.</p>
  </body>
</html>
```



Navigating the tree with rvest

```
<html>
  <body>
    <div>
      <p>The first paragraph.</p>
    </div>
    <div>
      Not an actual paragraph,
      but with a <a href="#">link</a>.
    </div>
    <p>A paragraph without an
      enclosing div.</p>
  </body>
</html>
```

```
html <- read_html(html_document)
html_children(html)
```

```
{xml_nodeset (1)}
[1] <body>\n    <div>\n  < ...
```

```
html %>% html_children()
```

```
html %>% html_children() %>% html_text()
```

```
[1] "\n    \n      The first paragraph.\n
\n    \n      Not an actual paragraph, \n
but with a link.\n    \n    A paragraph ...
```

Navigating to nodes with selectors

```
<html>
  <body>
    <div>
      <p>The first paragraph.</p>
    </div>
    <div>
      Not an actual paragraph,
      but with a <a href="#">link</a>.
    </div>
    <p>A paragraph without an
      enclosing div.</p>
  </body>
</html>
```

```
html <- read_html(html_document)
html %>% html_node('body')
```

```
{xml_nodeset (1)}
[1] <body>\n    <div>\n  < ...
```

```
html %>% html_nodes('div p')
```

```
{xml_nodeset (1)}
[1] <p>The first paragraph.</p>
```

Navigating to nodes with selectors

```
<html>
  <body>
    <div>
      <p>The first paragraph.</p>
    </div>
    <div>
      Not an actual paragraph,
      but with a <a href="#">link</a>.
    </div>
    <p>A paragraph without an
      enclosing div.</p>
  </body>
</html>
```

```
html %>% html_nodes('p')
```

```
{xml_nodeset (2)}
[1] <p>The first paragraph.</p>
[2] <p>A paragraph without an enclosi...
```

```
html %>% html_nodes('div') %>%
    html_nodes('p')
```

```
{xml_nodeset (1)}
[1] <p>The first paragraph.</p>
```

Extracting attributes

```
html %>%
    html_node('a') %>%
    html_attr('href')
```

```
[1] #
```

```
html %>%
    html_node('a') %>%
    html_attrs()
```

```
href
 "#"
```

Let's do this!

WEB SCRAPING IN R

Scrape your first table

WEB SCRAPING IN R

Timo Grossenbacher
Instructor

datacamp

Name	Profession	Age	Country
Dillon Arroyo	Carpenter	54	UK
Rebecca Douglas	Developer	32	USA

```
<table>
    <tr>
      <td>Name</td><td>Profession</td><td>Age</td><td>Country</td>
    </tr>
    <tr>
      <td>Dillon Arroyo</td><td>Carpenter</td><td>54</td><td>UK</td>
    </tr>
    <tr>
      <td>Rebecca Douglas</td><td>Developer</td><td>32</td><td>USA</td>
    </tr>
</table>
```

Name	Profession	Age	Country
Dillon Arroyo	Carpenter	54	UK
Rebecca Douglas	Developer	32	USA

```
<table>
    <tr>
      <th>Name</th><th>Profession</th><th>Age</th><th>Country</th>
    </tr>
    <tr>
      <td>Dillon Arroyo</td><td>Carpenter</td><td>54</td><td>UK</td>
    </tr>
    <tr>
      <td>Rebecca Douglas</td><td>Developer</td><td>32</td><td>USA</td>
    </tr>
</table>
```

Scraping a table with rvest

```
html <- read_html(table_html) # table with <th> header cells
html %>%
    html_table()
```

```
[[1]]
             Name Profession Age Country
1   Dillon Arroyo  Carpenter  54      UK
2 Rebecca Douglas  Developer  32     USA
```

Scraping a table with rvest

```
html <- read_html(table_html) # table without <th> header cells
html %>%
    html_table(header = TRUE)
```

```
[[1]]
             Name Profession Age Country
1   Dillon Arroyo  Carpenter  54      UK
2 Rebecca Douglas  Developer  32     USA
```

Scraping a table with rvest

```
html <- read_html(table_html)
html %>%
    html_table(header = TRUE, fill = TRUE)
```

```
[[1]]
             Name Profession Age Country
1   Dillon Arroyo  Carpenter  54      UK
2 Rebecca Douglas  Developer  32    <NA>
```

Scraping a table with rvest

If a table has a header row (with th elements) and no gaps, scraping it is straightforward, as with the following table (having **ID "clean"**)

```
# Extract the "clean" table into a data frame
mountains <- mountains_html %>%
  html_node("table#clean") %>%
  html_table()
```

Scraping "tables" in reality

```
<div class="rTable">
     <div class="rTableRow">
       <div class="rTableHead"><strong>Name</strong></div>
       <div class="rTableHead"><span style="font-weight: bold;">Telephone</span></div>
       <div class="rTableHead"> </div>
     </div>
     <div class="rTableRow">
       <div class="rTableCell">John</div>
       <div class="rTableCell"><a href="tel:0123456785">0123 456 785</a></div>
       <div class="rTableCell"><img src="images/check.gif" alt="checked" /></div>
     </div>
     <div class="rTableRow">
         ...
     </div>
</div>
```

[1] Example taken from h ps://html-cleaner.com/features/replace-html-table-tags-with-divs/

Let's practice!

WEB SCRAPING IN R