

# Web Scraping

## Hey, web scraping!

You may want to...

- download all videos from a website;
- download all news articles from a media platform;
- download all academic papers from a journal;
- download all tweets / weibo of a specific person.

You may need to spend days and nights downloading these data manually, and you can easily make a lot of mistakes.

# Hey, web scraping!

In today's class, we are going to learn about webscraping.  
In Chinese, it is called 网络爬虫.

## What is webscraping?

Using tools to gather data you can see on a webpage.  
Almost anything you see on a website can be scraped.

It can be done with python, R,... We are doing it on R.

<https://www.youtube.com/embed/Ct8Gxo8StBU?enablejsapi=1>

# Learning about HTML

HTML stands for “HyperText Markup Language.”

Websites are written on the HTML language, and web scraping is based on reading and interpreting the HTML of a webpage.

**But how to find the HTML of a webpage?**

# Learning about HTML

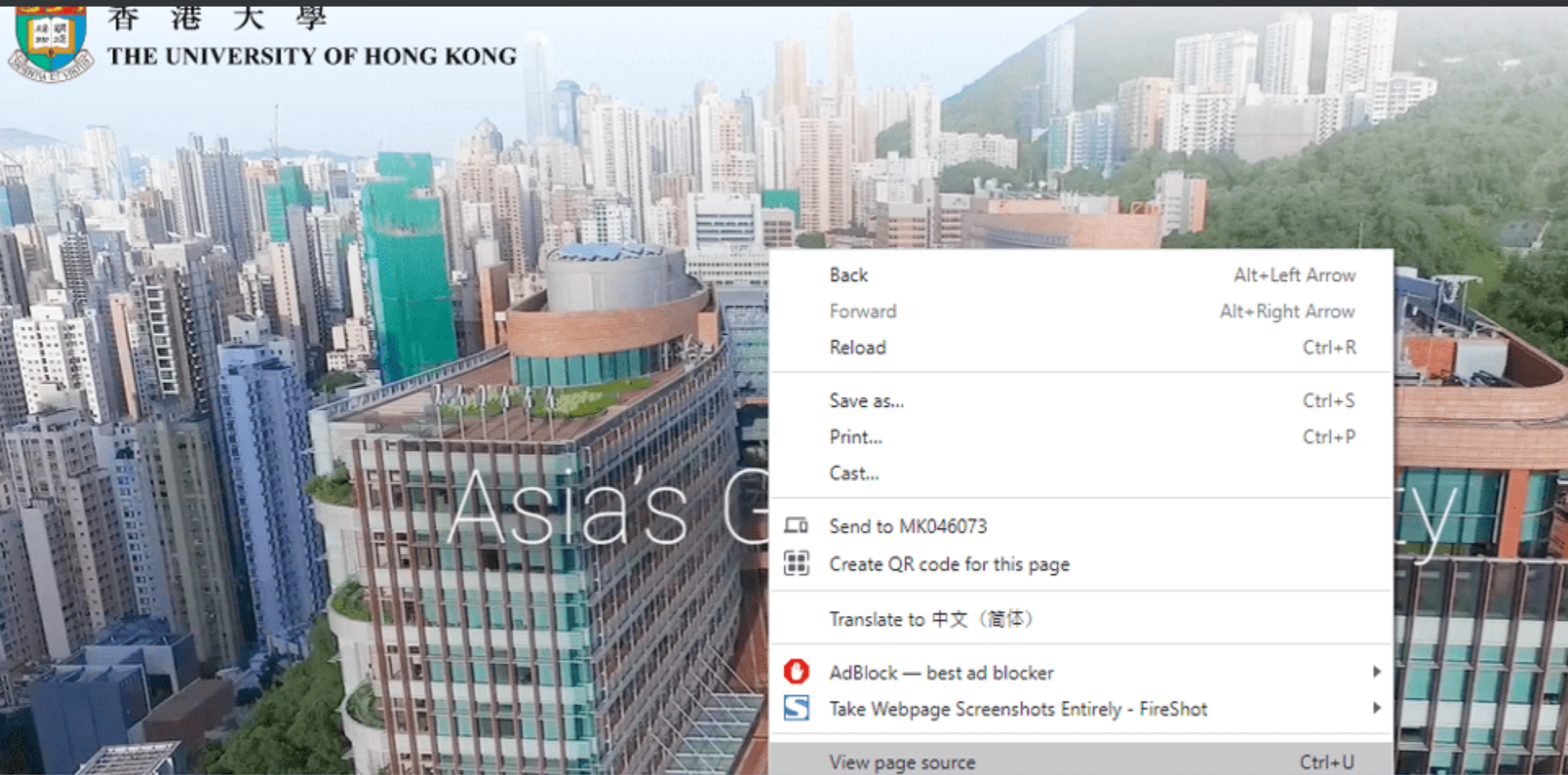
Please use Google Chrome as your browser.

If you are not a current user of Google Chrome, download and install one on your laptop. Google Chrome is particularly helpful for analyzing webpages for scraping.



香港大學

THE UNIVERSITY OF HONG KONG



Back	Alt+Left Arrow
Forward	Alt+Right Arrow
Reload	Ctrl+R
Save as...	Ctrl+S
Print...	Ctrl+P
Cast...	
Send to MK046073	
Create QR code for this page	
Translate to 中文 (简体)	
AdBlock — best ad blocker	▶
Take Webpage Screenshots Entirely - FireShot	▶
View page source	Ctrl+U

```
1
2 <!DOCTYPE html>
3 <!--[if lt IE 9]><html class="no-js lte-ie9 lt-ie9lang-en" lang="en"><![endif]-->
4 <!--[if IE 9]><html class="no-js lte-ie9 ie9lang-en" lang="en"><![endif]-->
5 <!--[if gt IE 9]><!-->
6 <html class="no-js" xmlns="http://www.w3.org/1999/xhtml"
7   xml:lang="en" lang="en">
8
9 <head>
10   <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
11   <meta http-equiv="X-UA-Compatible" content="IE=edge" />
12   <meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">
13   <meta http-equiv="Content-Style-Type" content="text/css" />
14   <meta http-equiv="Content-Script-Type" content="text/javascript" />
15   <link rel="apple-touch-icon" sizes="180x180" href="/assets/img/apple-touch-icon.png">
16   <link rel="icon" type="image/png" href="/assets/img/favicon-32x32.png" sizes="32x32">
17   <link rel="icon" type="image/png" href="/assets/img/favicon-16x16.png" sizes="16x16">
18   <link rel="manifest" href="/assets/img/manifest.json">
19   <link rel="shortcut icon" href="/assets/img/favicon.ico">
20   <meta name="msapplication-config" content="/assets/img/browserconfig.xml">
21   <meta name="theme-color" content="#ffffff">
22   <noscript><style>
```



# Learning about HTML

The data you want to scrape appears in certain place of the HTML. For example, suppose that you want to scrape data from the HKU marketing faculty [webpage](#):



# Learning about HTML

You can find the name and images of the professors from the HTML file:

```
▼ <div class="people-card fadeInUp animated" data-animate="fadeInUp"> flex
  ▼ <a href="https://www.hkubs.hku.hk/people/jinzhao-du/" class="el-processed">
    ▶ <noscript> ⋮ </noscript>
      
    ▼ <div class="people-info">
      <div class="h5">Prof. Jinzhao DU</div> == $0
    </div>
  </a>
</div>
▶ <div class="wgl_col-3 people-item"> ⋮ </div>
▶ <div class="wgl_col-3 people-item"> ⋮ </div>
```

# Learning about HTML

And even the link to their photos (see [this link](#) for example).

```
▼ <div class="people-card fadeInUp animated" data-animate="fadeInUp"> flex
  ▼ <a href="https://www.hkubs.hku.hk/people/jinzhao-du/" class="el-processed">
    ▶ <noscript>...</noscript>
    
    ▼ <div class="people-info">
      <div class="h5">Prof. Jinzhao DU</div> == $0
    </div>
  </a>
</div>
▶ <div class="wgl_col-3 people-item">...</div>
▶ <div class="wgl_col-3 people-item">...</div>
```

Secret: Changing the webpage

# Webscraping

Suppose that you want to download the names of each individual marketing faculty, what should you do?

First, you need to get the HTML for the webpage.

Second, you need to analyze the HTML to get the desired information --- **this is much more difficult.**

# Webscraping

```
1 install.packages("rvest")
2 library(rvest)
3 url = "https://www.hkubs.hku.hk/people/faculty?
  pg=1&staff_type=faculty&subject_area=marketing&track=all"
4 webpage = read_html(url, encoding = "UTF-8")
5 print(webpage)
```

# Webscraping

Now, you get the HTML source file here. The next thing you need to do it to understand the HTML file, which is very challenging.

```
> print(webpage)
{html_document}
<html dir="ltr" lang="en-US" prefix="og: https://ogp.me/ns#">
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">\n<meta name="viewp ...
[2] <body class="page-template page-template-people-listing page-template-people-listing-php page ...
```

# Webscraping

To better understand the HTML code, you are strongly recommended to use Chrome as your browser.

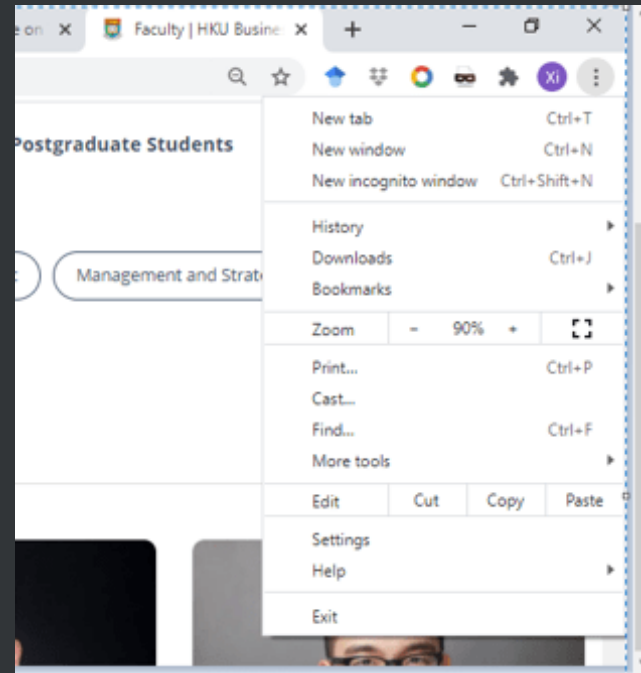
Chrome allows you to check the HTML code in a much more convenient matter.



# Check HTML with Chrome

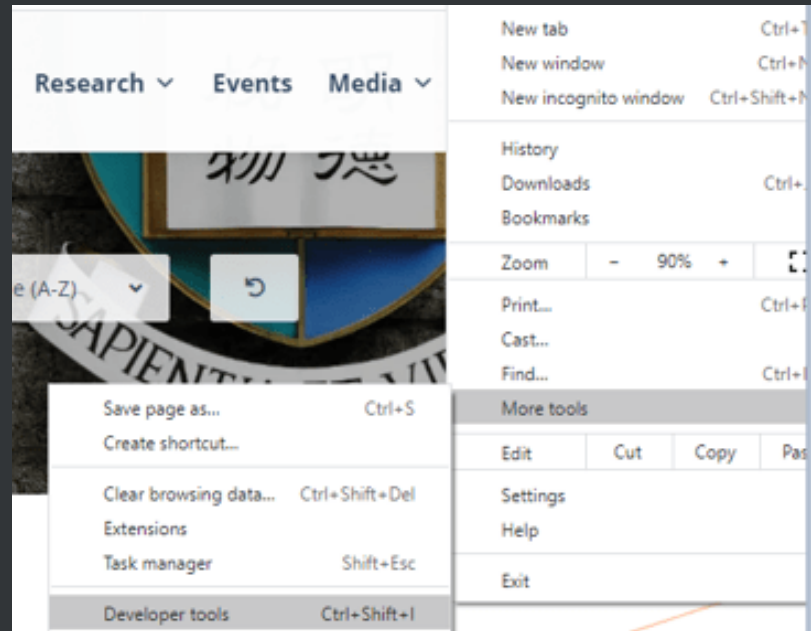
Open the webpage in your Chrome browser.

Click the upper right Chrome setting button of your browser and you will be directed here.




# Check HTML with Chrome


Choose “More tools” ...  
Choose “Developer tools” ...

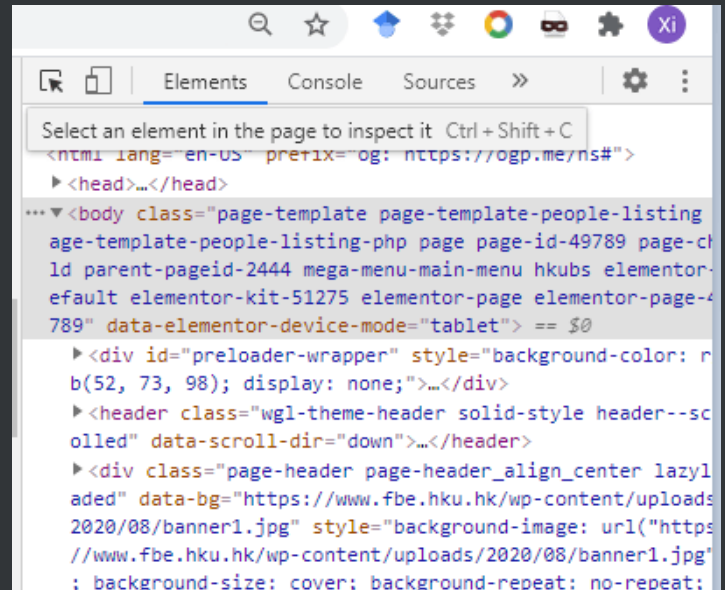


## Check HTML with Chrome

Click the  button and you will get to “select an element in the page to inspect it”.

Alternatively, use “Ctrl + Shift + C.”

If needed, also click the  button to switch between desktop and mobile version.



## Check HTML with Chrome

Take Prof. Du's information as an example. You can see his name appears here in the HTML code. But what does this mean?

```
▼ <a href="https://www.hkubs.hku.hk/people/jinzhao-du/" class="el-processed">
  ▶ <noscript>...</noscript>
  
    <div class="h5">Prof. Jinzhao DU</div> == $0
  </div>
</a>
</div>
</div>
```

```

▼ <div class="listing">
  ▼ <div class="row"> flex
    ▶ <div class="wgl_col-3 people-item">...</div>
    ▶ <div class="wgl_col-3 people-item">...</div>
    ▶ <div class="wgl_col-3 people-item">...</div>
    ▶ <div class="wgl_col-3 people-item">...</div>
    ▶ <div class="wgl_col-3 people-item">...</div>
    ▼ <div class="wgl_col-3 people-item">
      ▼ <div class="people-card fadeInUp animated" data-animate="fadeInUp"> flex
        ▼ <a href="https://www.hkubs.hku.hk/people/jinzhao-du/" class="el-processed">
          ▶ <noscript>...</noscript>
          
          ▼ <div class="people-info">
            <div class="h5">Prof. Jinzhao DU</div> == $0
            </div>
          </a>
        </div>
      </div>
    ▶ <div class="wgl_col-3 people-item">...</div>

```

<https://www.youtube.com/embed/u0OeZfIfBRI?enablejsapi=1>  
<https://www.youtube.com/embed/u0OeZfIfBRI?enablejsapi=1>

# Understanding HTML

Here, the name information is within a “div” node.

And this node belongs to a “div” node.

This “div” node further belongs to another “a” node.

And so on....

We call this is “path”: ...div / div / div / a / div / div

# Understanding HTML

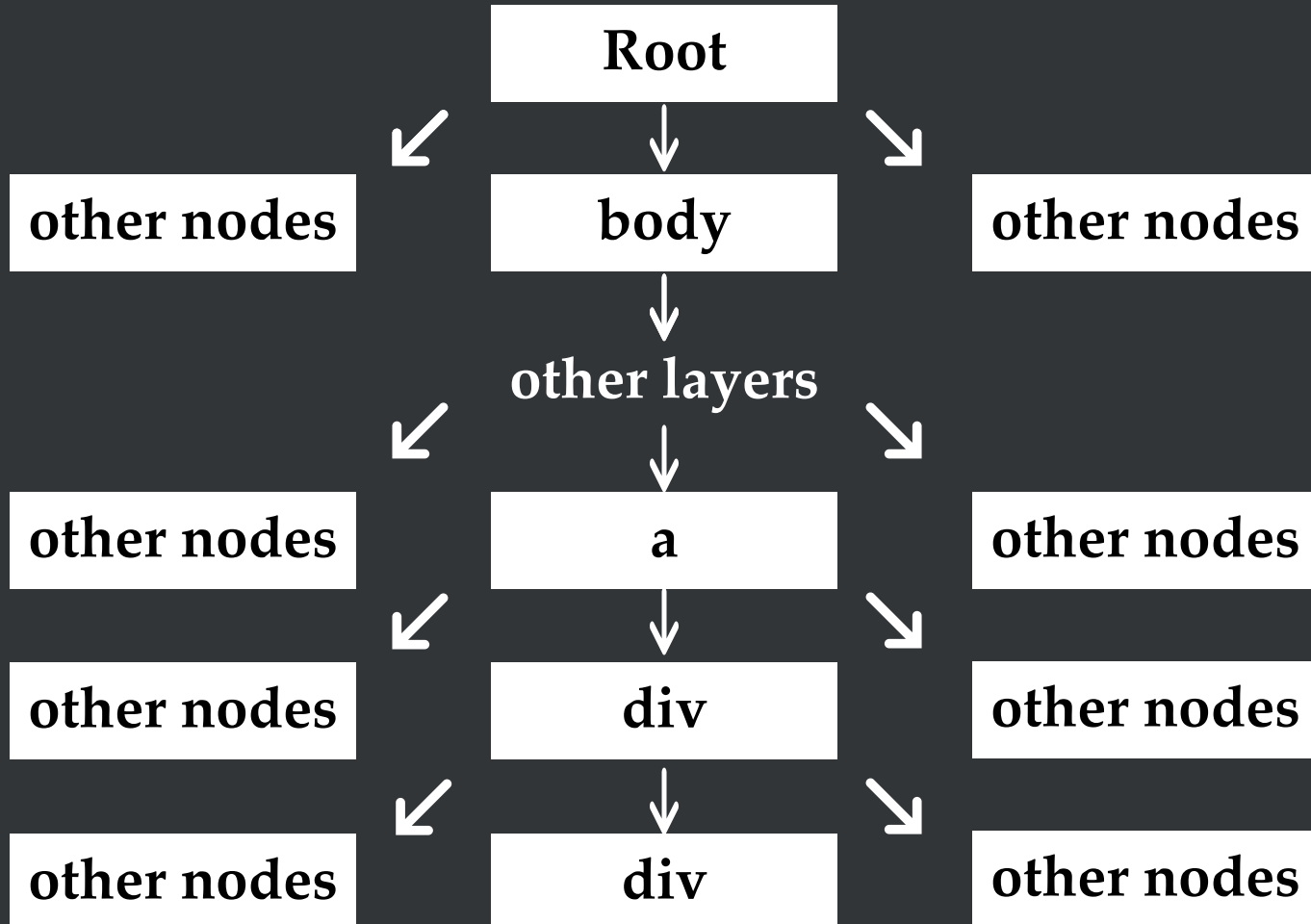
You can see that we have various types of nodes, including “div”, “a”, and “img”. You may wonder, “what do these types mean?”

Here, these types are called “tag”. For example, an “img” tag is used to mark up an image in the HTML language.

For detailed information, check [here](#).



# Understanding HTML



# Understanding HTML

This is something like your home address:

We have something like...

**Country / Province / City / District / Street / Building / Floor / Room**

The path helps us locate nodes and find the content of the nodes.

# Understanding HTML

However, unlike your home address, here each node does not have its name.

For example, we know it is a “**div**” node (not an “a” node) but there may be multiple “**div**” nodes.

My building is in a **street** (not an **avenue** or **road**) but there may be multiple **streets** here.

# Understanding HTML

Let's get all "div" nodes. This can be done by running this:

```
1 nodes <- html_nodes(webpage, xpath = '//div')
```

You can see that in total we have 262 "div" nodes.

```
1 print(length(nodes))
```

# Understanding HTML

We want to make the path more accurate to pin down to the “div” nodes that we are interested in. That is, we want to remove other unrelated “div” nodes.

We can do this by putting more restrictions on the path.

<https://www.youtube.com/embed/vNOyRZIkC7o?enablejsapi=1>

# Understanding HTML

Consider the following code:

```
1 nodes <- html_nodes(webpage, xpath = '//div/div')
2 print(length(nodes))
```

Here we restrict the parent of the “div” node must also be a “div” node. Now, we have 208 nodes --- still too many.

```
▼ <a href="https://www.hkubs.hku.hk/people/jinzhao-du/" class="el-processed">
  ▶ <noscript></noscript>
  
    <div class="h5">Prof. Jinzhao DU</div> == $0
  </div>
</a>
</div>
</div>
```

# Understanding HTML

Consider the following code:

```
1 nodes <- html_nodes(webpage, xpath = '//div[@class="people-info"]/div')
2 print(length(nodes))
```

Here we restrict the parent of the “div” node must also be a “div” node. Moreover, the its parent node must have a class attribute will is called “people-info.”

```
▼ <a href="https://www.hkubs.hku.hk/people/jinzhao-du/" class="el-processed">
  ▶ <noscript>...</noscript>
  
    <div class="h5">Prof. Jinzhao DU</div> == $0
  </div>
</a>
</div>
</div>
```



# Understanding HTML

Now, we only have 16 div nodes selected. These are actually all HKU marketing faculties. Let us print their names:

```
1 nodes <- html_nodes(webpage, xpath = '//div[@class="people-info"]/div')
2 for (node in nodes)
3   print(html_text(node))
```

# Understanding HTML

You can also use other refinement to select the nodes that you are looking for. For example, the following codes work as well:

```
1 nodes <- html_nodes(webpage, xpath = '//div[@class="h5"]')
2 for (node in nodes)
3   print(html_text(node))
```

The complete code is here.

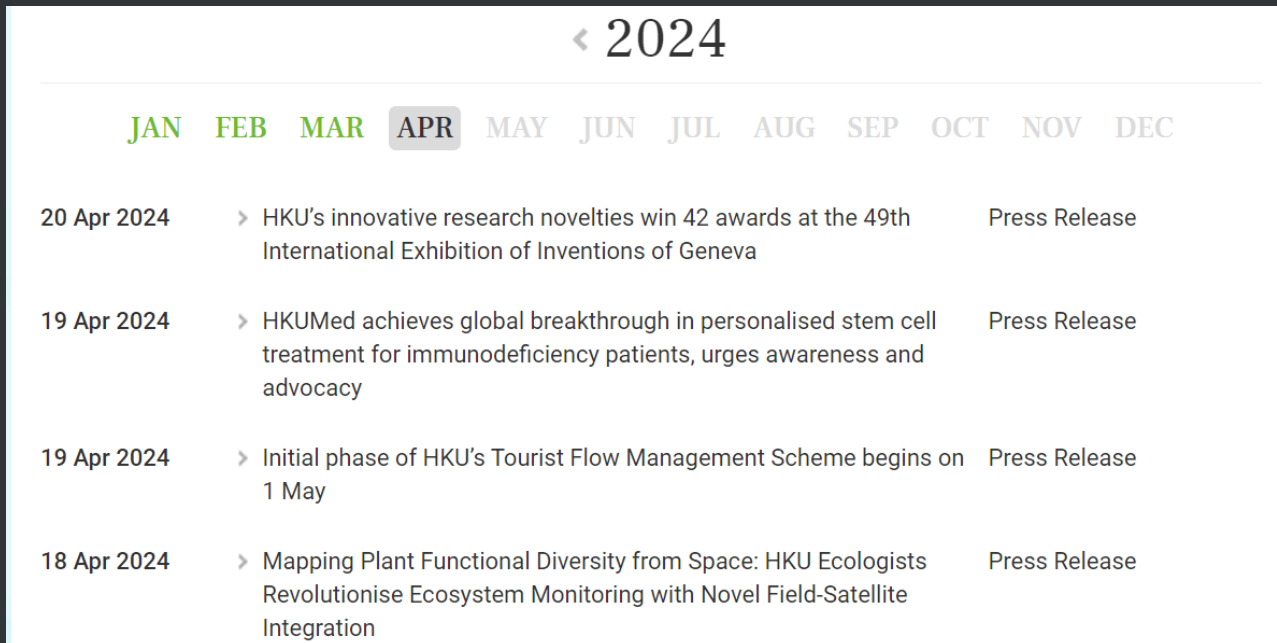
```
1 library(rvest)
2 url = "https://www.hkubs.hku.hk/people/faculty?
  pg=1&staff_type=faculty&subject_area=marketing&track=all"
3 webpage = read_html(url, encoding = "UTF-8")
4 nodes <- html_nodes(webpage, xpath = '//div[@class="h5"]')
5 for (node in nodes)
6   print(html_text(node))
```

## Exercise

Great! You now have a sense of how to scrape data from the web. It is very preliminary, and you will need a lot more exercises. Let us try the following exercise.

# Exercise

HKU makes press announcements on its official news webpage: <https://hku.hk/press/all/>



The image shows a screenshot of the HKU press releases page for April 2024. At the top, there is a navigation bar with the year '2024' and a left arrow. Below this, the months of the year are listed: JAN, FEB, MAR, APR, MAY, JUN, JUL, AUG, SEP, OCT, NOV, DEC. The 'APR' month is highlighted with a grey background. Below the navigation bar, there is a list of four press releases, each with a date, a brief description, and the type of release.

Month	Year	Press Release Title	Type
JAN	2024		
FEB	2024		
MAR	2024		
APR	2024		
MAY	2024		
JUN	2024		
JUL	2024		
AUG	2024		
SEP	2024		
OCT	2024		
NOV	2024		
DEC	2024		
20 Apr	2024	> HKU's innovative research novelties win 42 awards at the 49th International Exhibition of Inventions of Geneva	Press Release
19 Apr	2024	> HKUMed achieves global breakthrough in personalised stem cell treatment for immunodeficiency patients, urges awareness and advocacy	Press Release
19 Apr	2024	> Initial phase of HKU's Tourist Flow Management Scheme begins on 1 May	Press Release
18 Apr	2024	> Mapping Plant Functional Diversity from Space: HKU Ecologists Revolutionise Ecosystem Monitoring with Novel Field-Satellite Integration	Press Release

# Exercise

Try to download the titles of these press articles!

**URL:** <https://hku.hk/press/all/>

## Exercise

First, let us scrape the titles. We must understand the corresponding HTML code to scrape the data.



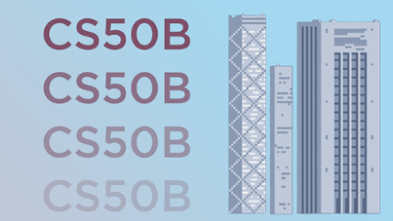
```
▼ <div class="press-item">
  <span class="date">17 Apr 2024</span>
  ▼ <span class="details">
    ▼ <a href="/press/news_detail_27225.html"> == $0
      "Jockey Club Community Elderly Mental Wellness Enhancement
      Project Explores an Inclusive Community with Senior Citizens
      in Their Everyday Life"
    </a>
  </span>
  <span class="news-type">Press Release</span>
</div>
```

```
1 library(rvest)
2 url = "https://hku.hk/press/all/"
3 webpage = read_html(url, encoding = "UTF-8")
4 nodes <- html_nodes(webpage, xpath = '//div[@class="press-
  item"]/span/a')
5 for (node in nodes)
6   print(html_text(node))
```



# Exercise

Now, let us visit the Harvard School of Professional Learning: <https://pll.harvard.edu/trending>

		
<p><b>HUMANITIES</b></p> <p>ONLINE</p> <p><b>The Path to Happiness: What Chinese Philosophy Teaches Us about the Good Life</b></p> <p>Why should we care about Confucius? Explore ancient Chinese philosophy, ethics, and political theory to challenge your assumptions of what it means to be happy, live a meaningful life, and change the world.</p> <p>FREE* 13 WEEKS LONG AVAILABLE NOW</p>	<p><b>COMPUTER SCIENCE</b></p> <p>ONLINE</p> <p><b>CS50's Introduction to Artificial Intelligence with Python</b></p> <p>Learn to use machine learning in Python in this introductory course on artificial intelligence.</p> <p>FREE* 7 WEEKS LONG AVAILABLE NOW</p>	<p><b>COMPUTER SCIENCE</b></p> <p>ONLINE</p> <p><b>CS50's Computer Science for Business Professionals</b></p> <p>This is CS50's introduction to computer science for business professionals.</p> <p>FREE* 6 WEEKS LONG AVAILABLE NOW</p>

# Exercise

In this exercise, we attempt to scrape the course titles, e.g.,  
“CS50's Introduction to Artificial Intelligence with Python”

Try this exercise yourself!

# Exercise

First, we identify the root of each individual course. We need to inspect the HTML code first.

```
▼ <div class="field field--name-title field--type-string field--label-hidden field__items">
  ▼ <h3 class="field__item">
    <a href="/course/cs50s-introduction-artificial-intelligence-python" hreflang="en" data-
      once="dlInternalCampaignClickedViewCourses dlLinkClicked">CS50's Introduction to
      Artificial Intelligence with Python</a> == $0
  </h3>
</div>
```

```
1 library(rvest)
2 url = "https://pll.harvard.edu/trending"
3 webpage = read_html(url, encoding = "UTF-8")
4 nodes <- html_nodes(webpage, xpath = '//h3/a')
5 for (node in nodes)
6   print(html_text(node))
```

# Scraping Images

Previously, we have discussed how to scrape text information from a website using a web scraper.

Now, let us consider scraping images from the web.

# Scraping Images

Let us go back to the HKU marketing faculty [webpage](#):



# Scraping Images

You can find a link to each photo (in “src” or “data-src” attribute):

```
▼ <div class="people-card fadeInUp animated" data-animate="fadeInUp"> flex
  ▼ <a href="https://www.hkubs.hku.hk/people/chu-ivy-dang/" class="el-processed">
    ▶ <noscript>...</noscript>
     == $0
    ▶ <div class="people-info">...</div>
```

# Scraping Images

Once you get the [link](#), you will have access to the photo.

An example of link: [https://www.hkubs.hku.hk/wp-content/uploads/fly-images/11554/FBE\\_0712\\_web--scaled-800x800-ct.jpg](https://www.hkubs.hku.hk/wp-content/uploads/fly-images/11554/FBE_0712_web--scaled-800x800-ct.jpg)

So, our first step to get the link information.



# Scraping Images

```
1 url = "https://www.fbe.hku.hk/people/faculty?
  pg=1&staff_type=faculty&subject_area=marketing&track=all"
2 webpage = read_html(url, encoding = "UTF-8")
3 image_nodes <- html_nodes(webpage, xpath =
  '//div/a/img[@width="800"]')
4 print(length(image_nodes))
```

# Scraping Images

But that's not enough. We not only want to get the nodes, but also need the link to each of the nodes. The link appears in the "src" or "data-src" attribute.

```
▼ <div class="people-card fadeInUp animated" data-animate="fadeInUp"> flex
  ▼ <a href="https://www.hkubs.hku.hk/people/chu-ivy-dang/" class="el-processed">
    ▶ <noscript>⋮</noscript>
     == $0
    ▶ <div class="people-info">⋮</div>
```

# Scraping Images

But that's not enough. We not only want to get the nodes, but also need the link to each of the nodes. The link appears in the "src" or "data-src" attribute.

```
1 image_nodes <- html_nodes(webpage, xpath = '//div/a/img[@width="800"]')
2 for (image in image_nodes)
3 {
4   photourl <- html_attr(image, "data-src")
5   print(photourl)
6 }
```

# Scraping Images

```
1 for (image in image_nodes)
2 {
3   photourl <- html_attr(image, "data-src")
4   print(photourl)
5   download.file(photourl,
6                 paste0(toString(number), '_HKU_Photo.jpg'), mode = 'wb')
7   number = number + 1
8 }
```

The complete code is here.

```
1 url = "https://www.fbe.hku.hk/people/faculty?
pg=1&staff_type=faculty&subject_area=marketing&track=all"
2 webpage = read_html(url, encoding = "UTF-8")
3 image_nodes <- html_nodes(webpage,xpath = '//div/a/img[@width="800"]')
4 number = 1
5 for (image in image_nodes)
6 {
7   photourl <- html_attr(image, "data-src")
8   print(photourl)
9   download.file(photourl,
10                 paste0(toString(number), '_HKU_Photo.jpg'), mode = 'wb')
11   number = number + 1
12 }
```

# Static vs. Dynamic Websites

<https://www.youtube.com/embed/hlg6q6OFoxQ?enablejsapi=1>

## Dynamic Websites

What we learned in today's class works well for static websites. But it does not work equally well on dynamic websites. If you want to scrape data from a dynamic website, you may need to use some more advanced tools.



## Dynamic Websites

If you want to scrape data from a dynamic website, there is a tool called “selenium”. We also have a packaged called “RSelenium” in R.

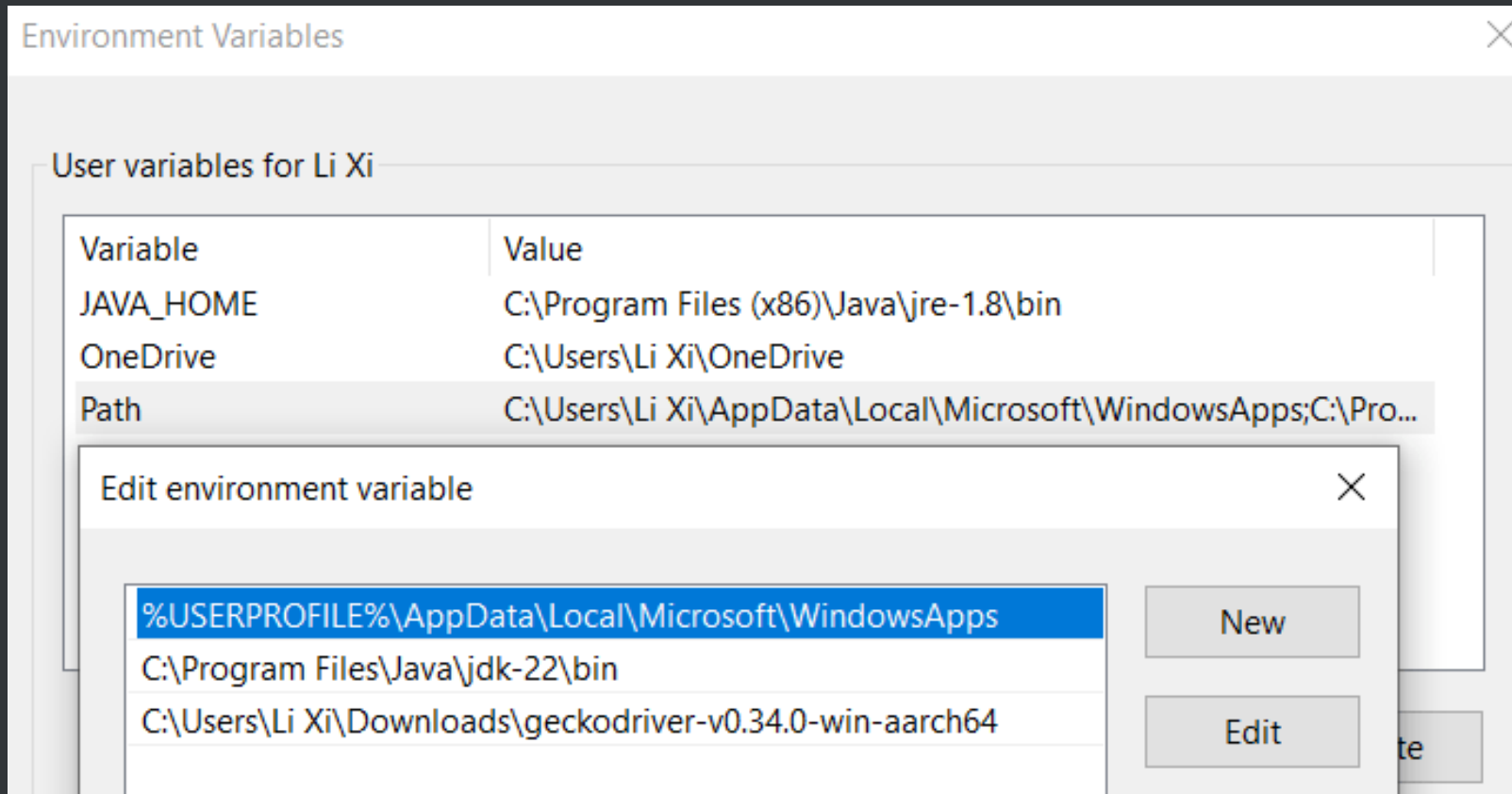
The selenium tool allows your scraper to visit a webpage like a human-being. That is, if you write a scraper with selenium, your scraper will also be able to scroll down your pages, click buttons, enter your password, etc.

## Dynamic Websites (Optional)

To scrape from a dynamic website, you need to:

- Install package “RSelenium”
- Install Firefox browser
- Download and install Java JDK and add its path to your system environment variable “PATH”
- Download and unzip Firefox Driver and add its path to your system environment variable “PATH”

# Dynamic Websites (Optional)



## Demonstration (Optional)

```
1 library(RSelenium)
2 driver <- rsDriver(browser="firefox", port=4990L, verbose=F, chromever =
  NULL)
3 remote_driver <- driver[["client"]]
4 remote_driver$open()
5 remote_driver$navigate("https://ximarketing.github.io")
6 element <- remote_driver$findElement(using = "xpath", value =
  '//a[contains(text(),"Teaching")]')
7 element$clickElement()
8 Sys.sleep(5)
9 element <- remote_driver$findElement(using = "xpath", value =
  '//a[contains(text(),"Algorithms")]')
10 element$clickElement()
11 Sys.sleep(5)
12 element <- remote_driver$findElement(using = "xpath", value = '//input')
13 element$sendKeysToElement(list("7025"))
14 Sys.sleep(5)
15 element <- remote_driver$findElement(using = "xpath", value = '//button')
16 element$clickElement()
17 Sys.sleep(5)
18 remote_driver$close()
```

Thank you!  
Enjoy scraping!