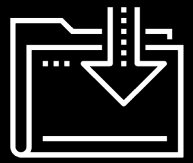
Scraping HTML

Data Boot Camp

Lesson 11.1

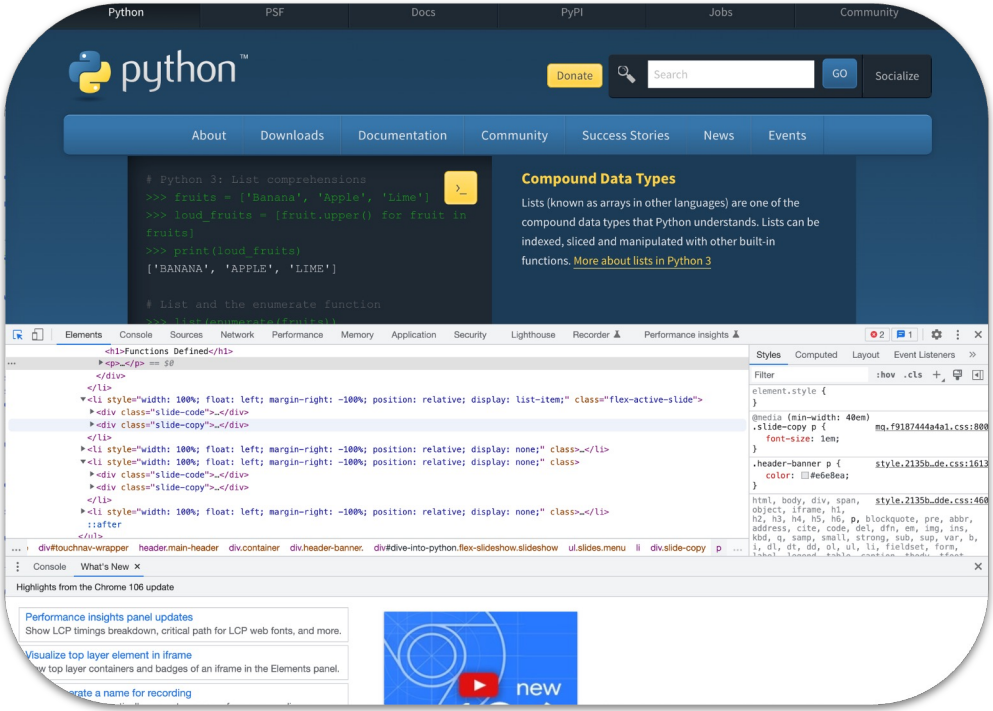



WELCOME

Data Collection…

…via web scraping!

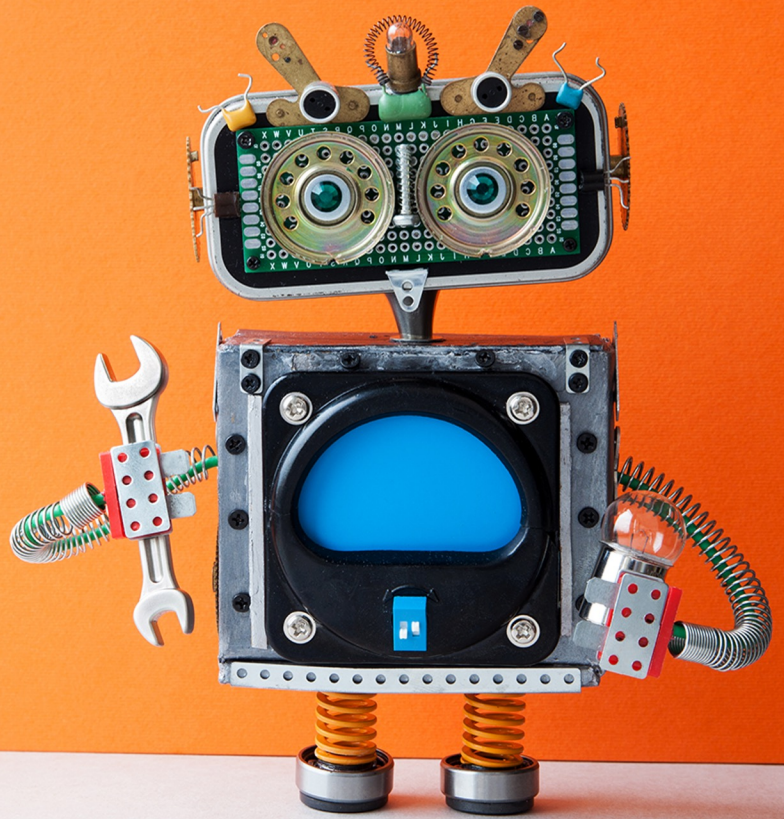
- **Data collection:** the process of gathering specific information, typically for a targeted analytical purpose
- **Web scraping:** a technique for collecting data from public websites by using knowledge of web design



All the Technologies!

This week we will cover the following:

- HTML
- CSS
- Beautiful Soup
- Splinter
- Chrome DevTools
- And more!



Class Objectives

By the end of today's class you will be able to:

Identify HTML components in a website.

Create a basic HTML document.

Scrape data from a website by using BeautifulSoup.

Style HTML elements by using CSS.

Questions?

Activity: Getting Started

In this activity, you will make sure that all of the libraries you need for this week have been installed.

Suggested Time:

15 Minutes

Activity: Installing ChromeDriver

- Go to the download page on Selenium project
- Choose “ChromeDriver server for win”.
- Your browser will download a zip file
- Extract the folder and add the .exe file to your PATH.

- If you do not have **Homebrew,** run the following command:

```
/bin/bash -c "$(curl -fsSL
https://raw.githubusercontent.com/Homebrew/in
stall/HEAD/install.sh)"
```

- Once you have **Homebrew**, run the following command:

```
brew install chromedriver
```

Activity: Installing Packages

Instructions:

Open up a terminal window and run the following commands:

```
pip install "splinter[selenium4]"
pip install bs4
pip install html5lib
pip install lxml
```

Instructor Demonstration

How Websites are Made

Languages of the Web

HTML

HyperText Markup Language

HTML

```
<div style="background:#eeeeee; border: none; padding:
10px; margin: 10px; font-family: 'Merriweather', serif;">

<h1>Sharing Your Work?</h1>

<h2>Sharing with Creative Commons Licences</h2>

<p>As you create designs for people on the internet to
see and interact with, you have many options for sharing
your work. Visit
<a href="https://creativecommons.org/">Creative
Commons</a> to learn more about the different licenses
you can apply to your original creations.</p>
```

CSS

Cascading Style Sheets

CSS

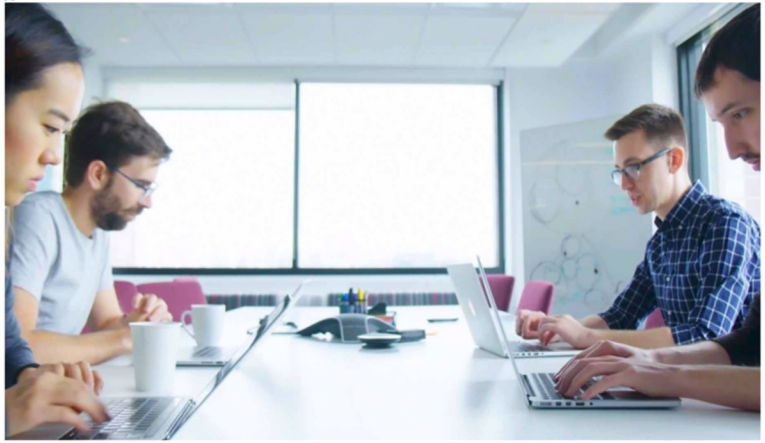
```
@import
url('https://fonts.googleapis.com/css2?family=MuseoModerno:wght@500&display=swap');

h1 {
    font-family: 'MuseoModerno', cursive;
    font-size: 48px;
    color: #a01047;
  }
p {font-family: 'Alegreya Sans', sans-serif;
    size: 12px;
    color: #2c0003;
    }
```

Without CSS

Skip to content

- About 2U
- Our Approach
- Our Partners
- Careers
- Latest

- Contact Us
- Investors
- Press
- GetSmarter
- Trilogy



Edtech with a human touch.

At 2U, it's a mix of proprietary technology and passionate people that truly powers our world-class online learning experience. And while we started with graduate programs, we're evolving to meet the needs of learners across their lifetimes.

Approach

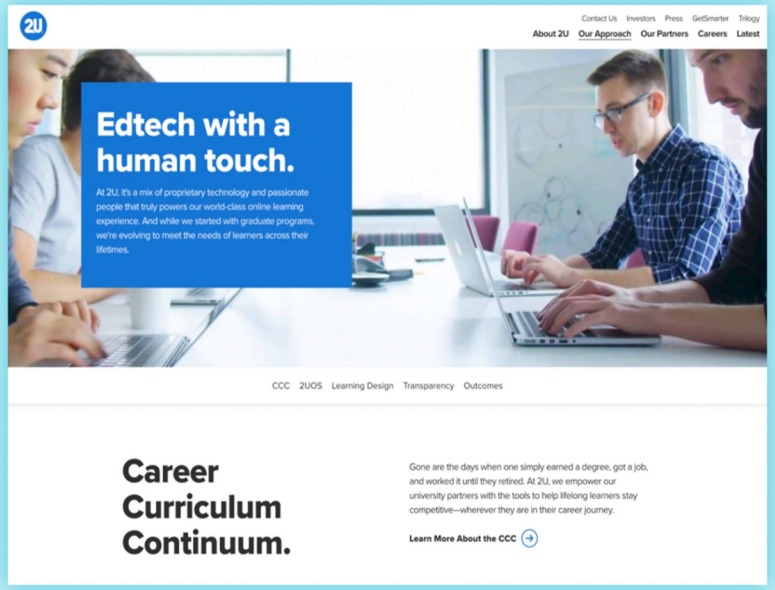
- CCC
- 2UOS
- Learning Design
- Transparency
- Outcomes

Career Curriculum Continuum.

Gone are the days when one simply earned a degree, got a job, and worked it until they retired. At 2U, we empower our university partners with the tools to help lifelong learners stay competitive—wherever they are in their career journey.

Learn More About the CCC

With CSS



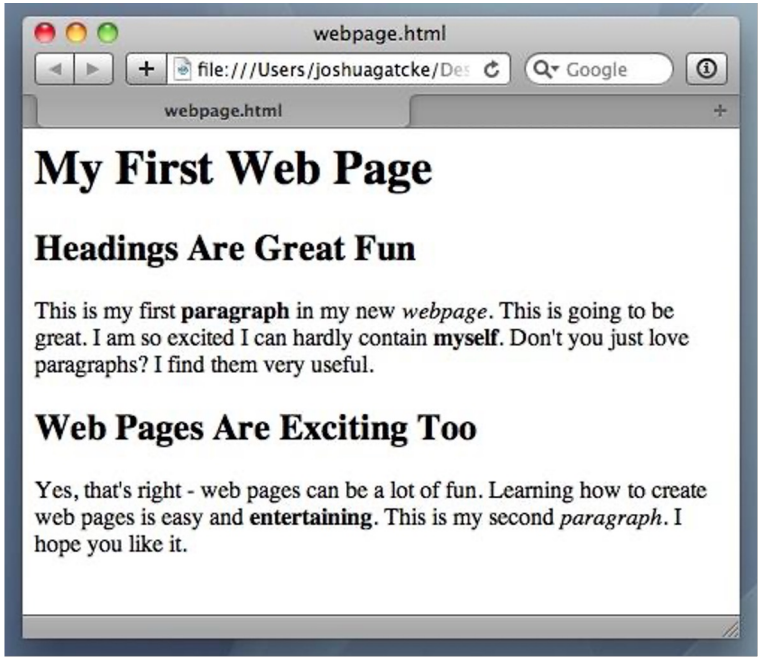
Instructor Demonstration

Hello HTML

Hello HTML

HTML5

- HTML is one of the three base languages behind every website.
- It defines all the basic content and a bit of formatting.



Hello HTML

HTML elements are rendered by the browser as visible parts of a webpage

Hello HTML

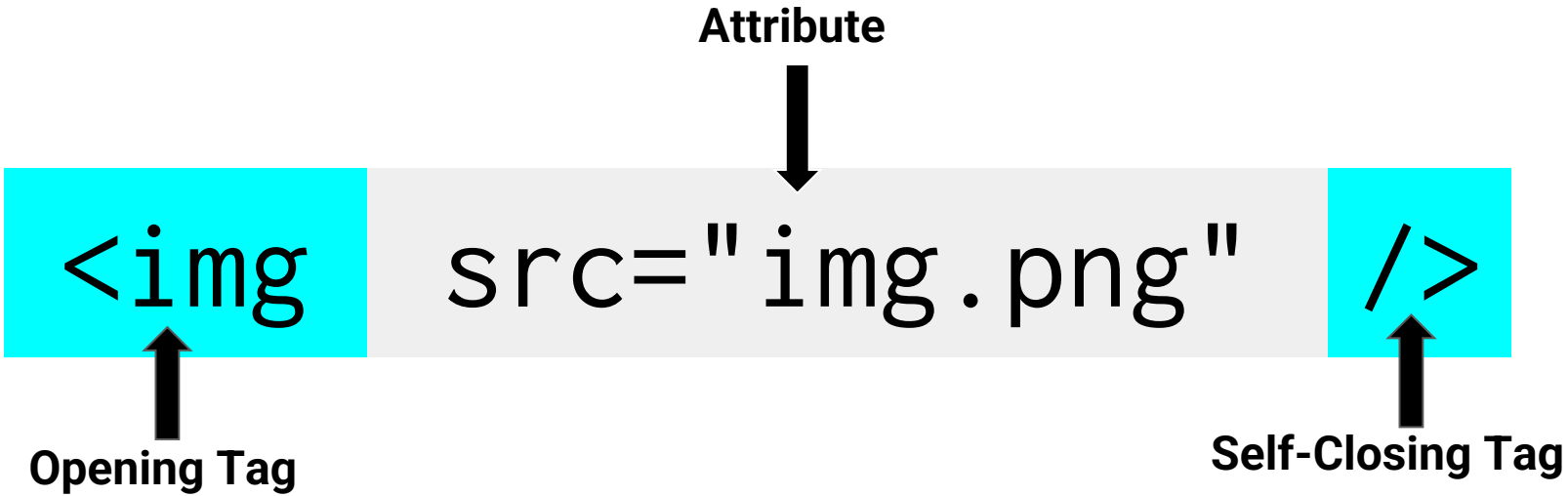
HTML Syntax (Basic)

Hello HTML

HTML Syntax (with Attribute)

Hello HTML

Tricky Tags (Self-Closing)



Activity: My first HTML

In this activity, you will create your first web page using HTML.

Suggested Time:

20 Minutes

Activity: My First HTML

Instructions

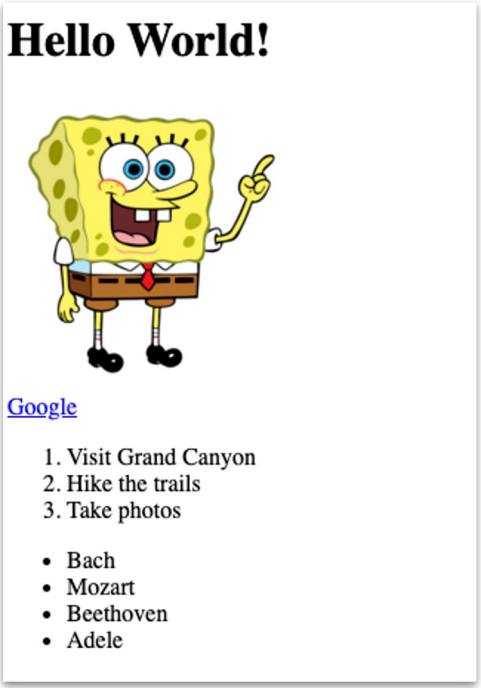
In a new HTML file, create the basic structure of an HTML document and include in it the following:

- `DOCTYPE` declaration
- `<head>` element with nested `<title>` element
- `<h1>` element with a title of your choice
- An image
- A link to an external page, such as google.com
- An ordered list of things to do on your next vacation
- An unordered list of four bands/musicians you like. To create an unordered list, use the `<ul>` tag.

You should be checking the rendered HTML in a web browser as you code to make sure you're going in the right direction.

My First HTML Example

Your HTML page will look similar to the following image.



Time’s Up! Let’s Review.


Break

Instructor Demonstration

Introduction to Beautiful Soup

Activity: From Soup to Nuts

In this activity, you will perform basic HTML scraping with BeautifulSoup.

Suggested Time:

15 Minutes

Activity: From Soup to Nuts

Instructions

In this activity, you will use BeautifulSoup to extract the following information from an HTML document:

- The `<head>` element
- The first `<h1>` element, then its text
- The first `<h2>` element, then its text
- The first anchor (`<a>`), then its `href` attribute.
- The first `<ul>` element, and its first list item (`<li>`), as well as the list item's text

Time’s Up! Let’s Review.

Instructor Demonstration

Styling HTML with CSS

Instructor Demonstration

CSS Selectors

Activity: CSS My List

In this activity, you will create your first web page using HTML and CSS.

Suggested Time:

15 Minutes

Activity: CSS My List

Instructions

In a new HTML file, create three ordered lists. Each list should have four items.

- The first should be a list of four cities. The entire list should have an id of "cities".
- The second should be a list of four food entrees. Two of them should contain meat, and two of them should be vegetarian. The meat list items should have a class called "meat". The vegetarian list items should have a class called "vegetarian".
- The third should be a list of four movies. Your favorite movie on that list should have an id called "favorite". Only that list item should have an id.

In the style section, use CSS selectors to color your target elements.

- Color the entire list of cities purple.
- Color the meat-containing dishes brown, and the vegetarian dishes green.
- Color your favorite movie orange.

CSS My List: Example

Your HTML page will look similar to the following image.

Time’s Up! Let’s Review.


Recap

Questions?


The End