# WEB-SCRAPING WITH PYTHON

# Overview

- Web Scraping 101
- Basic HTML
- String Detection with Regex
- Python: ***beautifulsoup***
  - *Example 1) Jeopardy Archive*
  - *Example 2) HackerNews Blog (If time allows)*
- Q&A / DataFest Ideas (If time allows)

# WEB-SCRAPING 101

# Web Scraping 101

*What is Web Scraping ?*

- Process of extracting data from websites by fetching them and extracting it
- Many different software options to use
- Fundamental process to creating datasets from online resources

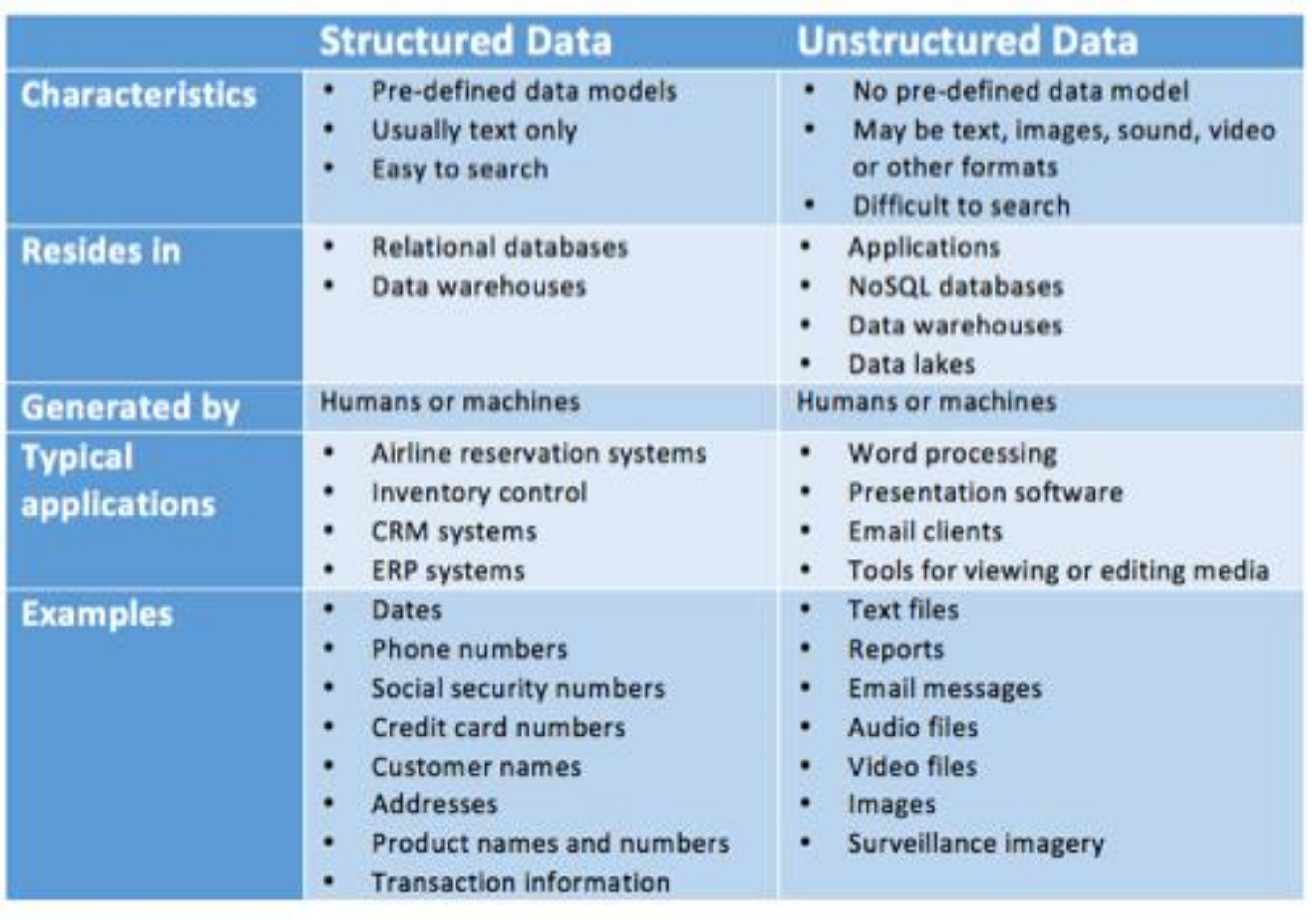
*How does it work ?*

- Take unstructured data (HTML tags) → structured format (data frame)
- Unstructured data is estimated to make up 65%-85% of the World Wide Web (!)

# Web Scraping 101

*Examples of Unstructured Data:*

- **Human Generated:** Social Media, Websites, Photo Sharing, Court Reports, etc.
- **Machine Generated:** Weather Data, Surveillance Photos, Sensors, etc.



| | Structured Data | Unstructured Data |
|---|---|---|
| **Characteristics** | • Pre-defined data models<br>• Usually text only<br>• Easy to search | • No pre-defined data model<br>• May be text, images, sound, video or other formats<br>• Difficult to search |
| **Resides in** | • Relational databases<br>• Data warehouses | • Applications<br>• NoSQL databases<br>• Data warehouses<br>• Data lakes |
| **Generated by** | Humans or machines | Humans or machines |
| **Typical applications** | • Airline reservation systems<br>• Inventory control<br>• CRM systems<br>• ERP systems | • Word processing<br>• Presentation software<br>• Email clients<br>• Tools for viewing or editing media |
| **Examples** | • Dates<br>• Phone numbers<br>• Social security numbers<br>• Credit card numbers<br>• Customer names<br>• Addresses<br>• Product names and numbers<br>• Transaction information | • Text files<br>• Reports<br>• Email messages<br>• Audio files<br>• Video files<br>• Images<br>• Surveillance imagery |

# Web Scraping 101

*Web Scraping Tools Available (For Free!)*

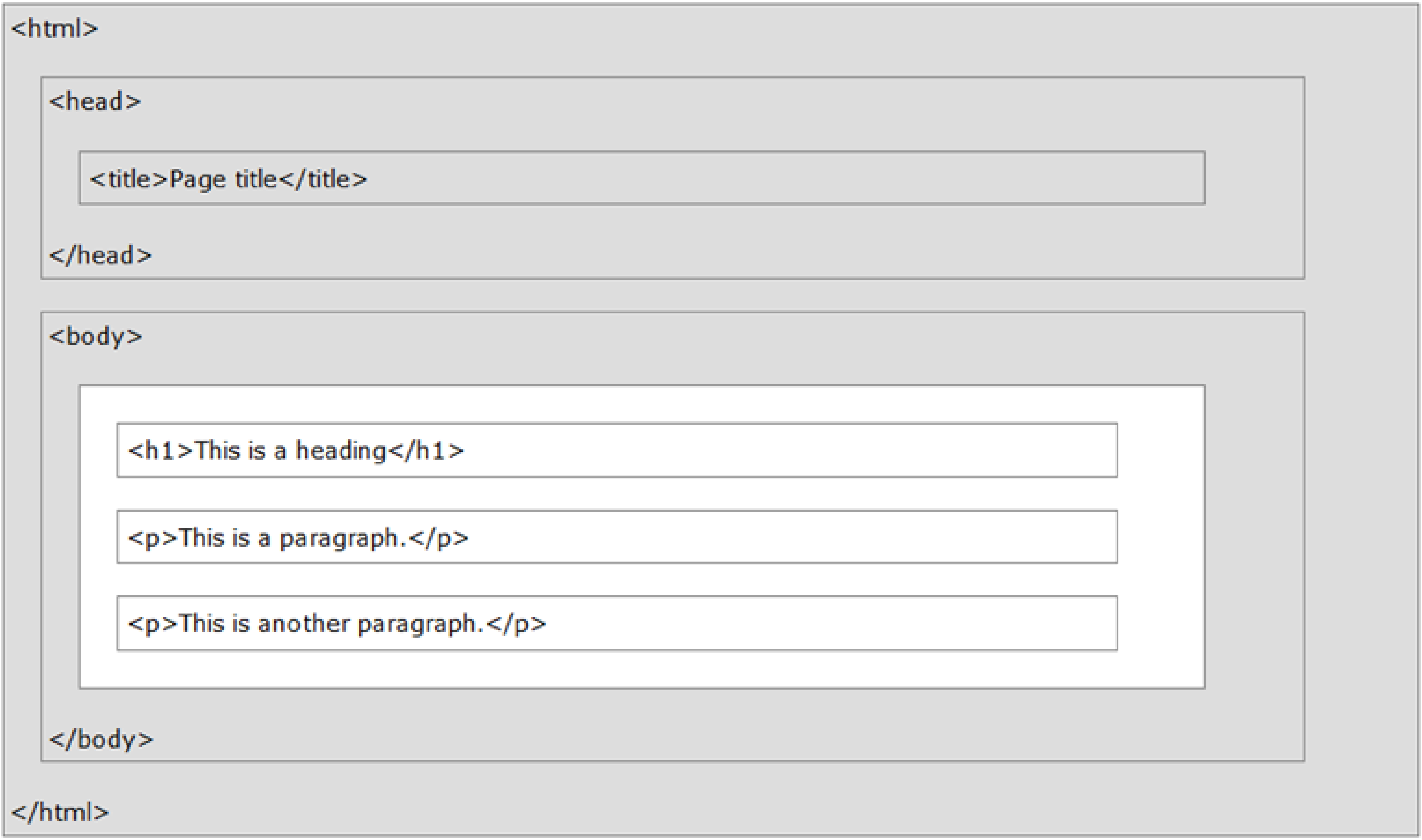
- Python (Beautifulsoup Library)
  - *https://pypi.org/project/beautifulsoup4/*
- Selector Gadget
  - *https://selectorgadget.com*
- R (Rvest Package)
  - *https://www.rdocumentation.org/packages/rvest/versions/0.3.5*

# BASIC HTML

And I mean basic…

# Basic HTML

- Websites are composed of a series of descriptive HTML tags
- **<tagname> *content goes here * </tagname>**
  - **<head> tags** contain meta information describing the site
  - **<title> tags** specify the document title
  - **<body> tags** contain page content that’s visible
  - **<h1> tags** define large headings
  - **<p> tags** define paragraphs

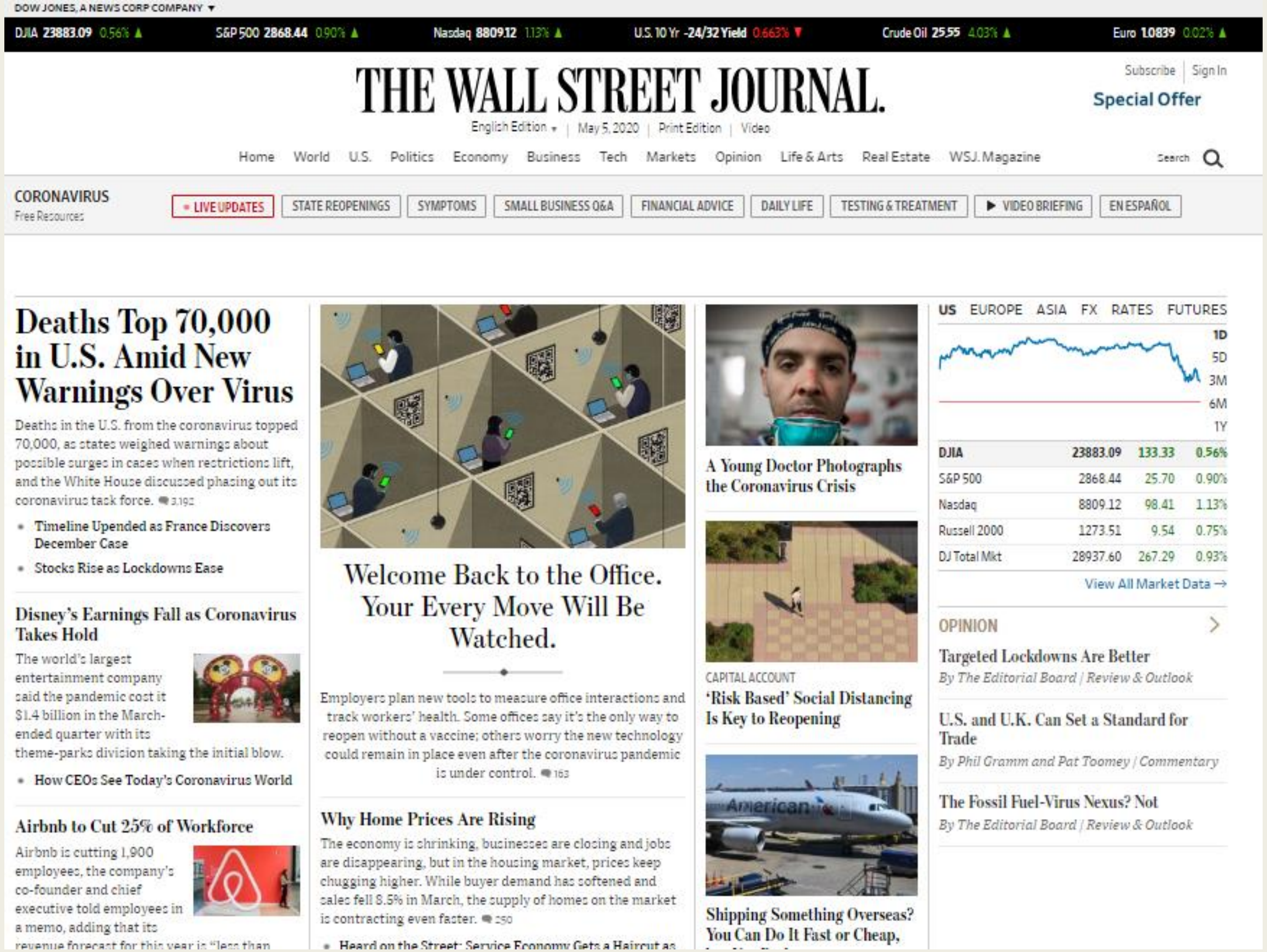


```
<html>
  <head>
    <title>Page title</title>
  </head>
  <body>
      <h1>This is a heading</h1>
      <p>This is a paragraph.</p>
      <p>This is another paragraph.</p>
  </body>
</html>
```

# Basic HTML Continued

- **<a> tags** define hyperlinks to URLs in the document
  - Important for iterating through pages
- **<td> tags** define a standard cell in an HTML table
  - Often where you will find data points
- **<img> tags** define images
- **<table> tags** define tables
- **<strong> tags** give text a strong emphasis

# But its never that simple…



**Deaths Top 70,000 in U.S. Amid New Warnings Over Virus**

Deaths in the U.S. from the coronavirus topped 70,000, as states weighed warnings about possible surges in cases when restrictions lift, and the White House discussed phasing out its coronavirus task force. 3,192

- Timeline Upended as France Discovers December Case
- Stocks Rise as Lockdowns Ease

**Disney's Earnings Fall as Coronavirus Takes Hold**

The world's largest entertainment company said the pandemic cost it $1.4 billion in the March-ended quarter with its theme-parks division taking the initial blow.

- How CEOs See Today's Coronavirus World

**Airbnb to Cut 25% of Workforce**

Airbnb is cutting 1,900 employees, the company's co-founder and chief executive told employees in a memo, adding that its revenue forecast for this year is "less than

**Welcome Back to the Office. Your Every Move Will Be Watched.**

Employers plan new tools to measure office interactions and track workers' health. Some offices say it's the only way to reopen without a vaccine; others worry the new technology could remain in place even after the coronavirus pandemic is under control. 163

**Why Home Prices Are Rising**

The economy is shrinking, businesses are closing and jobs are disappearing, but in the housing market, prices keep chugging higher. While buyer demand has softened and sales fell 8.5% in March, the supply of homes on the market is contracting even faster. 250

- Heard on the Street: Service Economy Gets a Haircut as

**A Young Doctor Photographs the Coronavirus Crisis**

CAPITAL ACCOUNT

**'Risk Based' Social Distancing Is Key to Reopening**

**Shipping Something Overseas? You Can Do It Fast or Cheap,**

| | | | |
|---|---|---|---|
| DJIA | 23883.09 | 133.33 | 0.56% |
| S&P 500 | 2868.44 | 25.70 | 0.90% |
| Nasdaq | 8809.12 | 98.41 | 1.13% |
| Russell 2000 | 1273.51 | 9.54 | 0.75% |
| DJ Total Mkt | 28937.60 | 267.29 | 0.93% |

View All Market Data →

OPINION

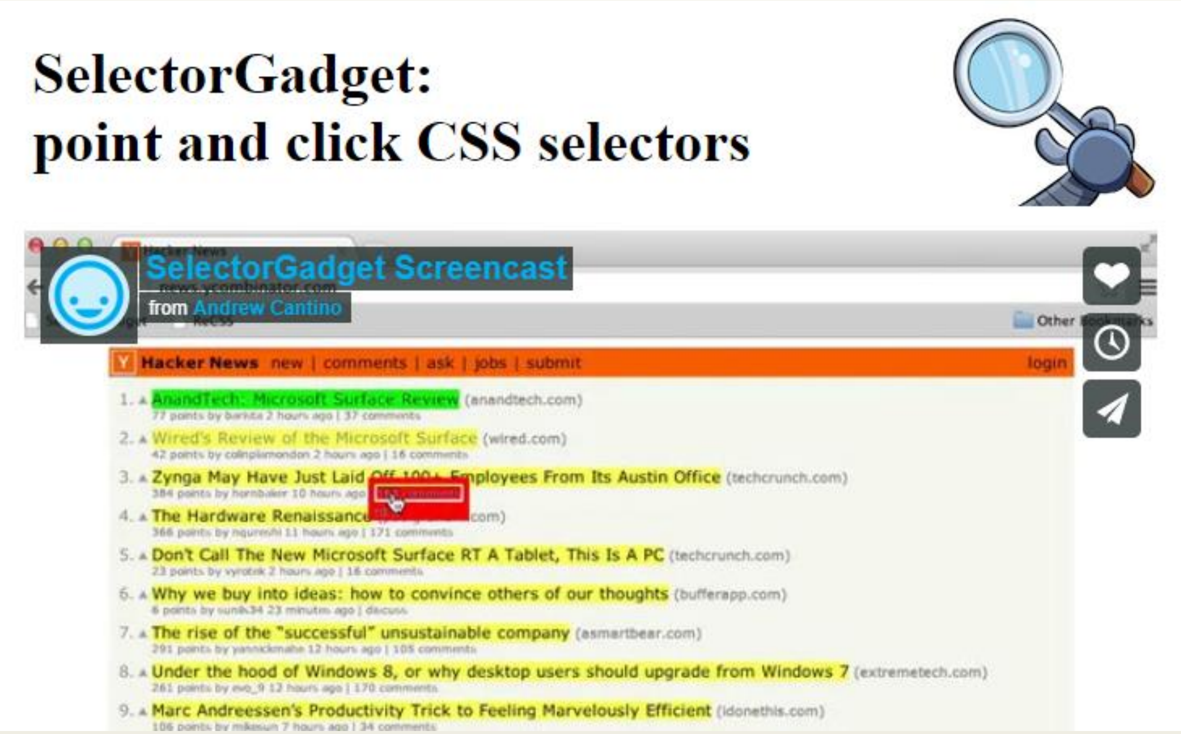
**Targeted Lockdowns Are Better**
*By The Editorial Board | Review & Outlook*

**U.S. and U.K. Can Set a Standard for Trade**
*By Phil Gramm and Pat Toomey | Commentary*

**The Fossil Fuel-Virus Nexus? Not**
*By The Editorial Board | Review & Outlook*

# The Workaround: Selector Gadget

Allows you to directly extract the html elements of the contents on a web page you desire!



**SelectorGadget: point and click CSS selectors**

SelectorGadget Screencast

from Andrew Cantino

Hacker News new | comments | ask | jobs | submit login

1. AnandTech: Microsoft Surface Review (anandtech.com)
   77 points by barista 2 hours ago | 37 comments
2. Wired's Review of the Microsoft Surface (wired.com)
   42 points by colinplamondon 2 hours ago | 16 comments
3. Zynga May Have Just Laid Off 100+ Employees From Its Austin Office (techcrunch.com)
   384 points by hornbaker 10 hours ago
4. The Hardware Renaissance com)
   366 points by nqureshi 11 hours ago | 171 comments
5. Don't Call The New Microsoft Surface RT A Tablet, This Is A PC (techcrunch.com)
   23 points by vyrotek 2 hours ago | 16 comments
6. Why we buy into ideas: how to convince others of our thoughts (bufferapp.com)
   6 points by sunils34 23 minutes ago | discuss
7. The rise of the "successful" unsustainable company (asmartbear.com)
   291 points by yannickmahe 12 hours ago | 105 comments
8. Under the hood of Windows 8, or why desktop users should upgrade from Windows 7 (extremetech.com)
   261 points by evo_9 12 hours ago | 170 comments
9. Marc Andreessen's Productivity Trick to Feeling Marvelously Efficient (idonethis.com)
   106 points by mikesun 7 hours ago | 34 comments

# Additional HTML tips …

- Two of very common HTML attributes are **class** and **id**. They are used for grouping and identifying HTML tags.

# EXAMPLES & TIPS

Using Beautiful Soup to scrape the *Jeopardy* Archive and Hacker News

# Tip 1: Inspect your data source:

- Explore the website
  - *Which data points do you want to gather?*
  - *How should they be organized ?*
- Decipher the information of URLS!
  - *Base URL:*
    - Represent path to the website itself
  - *Query Parameters:*
    - Represent the additional values that can be declared on a page
      - *Starter parameters (start with ?)*
      - *Information (key = value)*
    - Parameters are often separated by an '&'

# Example URLs

What is the Base URL?

What are the Query Parameters?

- *http://www.j-archive.com/showgame.php?game_id=6527*
- https://news.ycombinator.com/submitted?id=whoishiring
- *https://laist.com/search_query.php?q=homelessness*

# Example URLs

What is the Base URL?

What are the Query Parameters?

- *http://www.j-archive.com/showgame.php?game_id=6527*
- https://news.ycombinator.com/submitted?id=whoishiring
- *https://laist.com/search_query.php?q=homelessness*

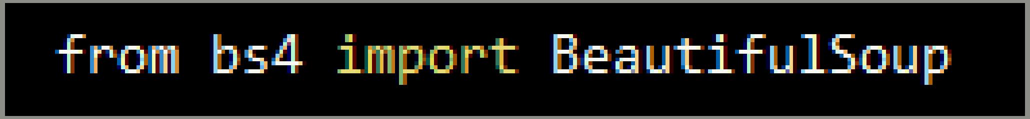
# Example URLs

What is the Base URL?

What are the Query Parameters?

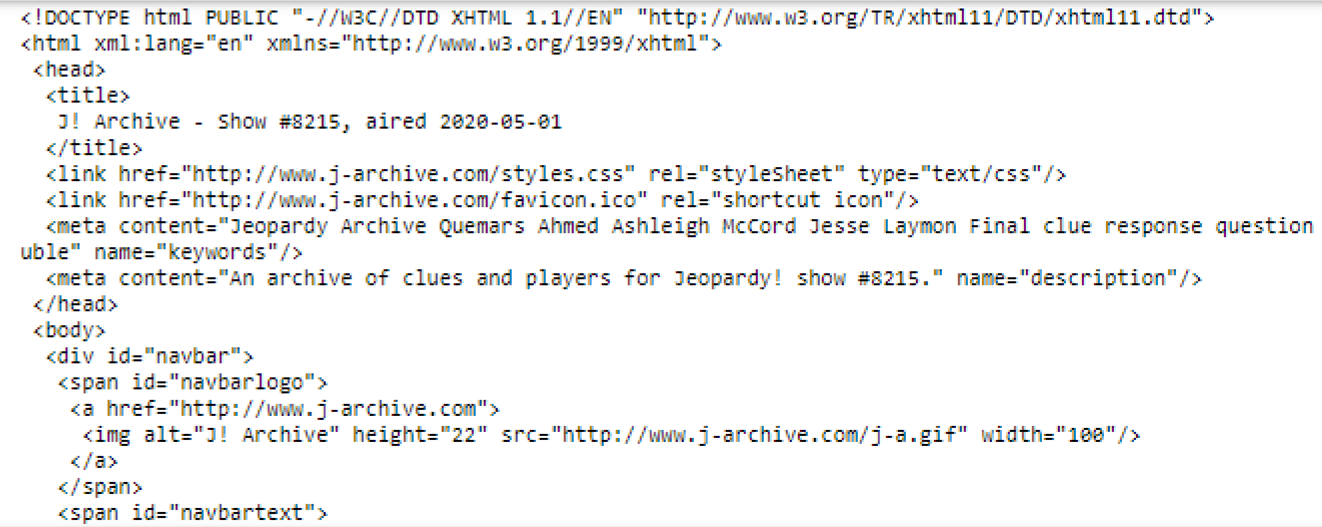
- *http://www.j-archive.com/showgame.php?game_id=6527*
- https://news.ycombinator.com/submitted?id=whoishiring
- *https://laist.com/search_query.php?q=homelessness*

# Tip 2:
# Fetch the Page & Review Contents



```
from bs4 import BeautifulSoup
```

- Using Beautiful Soup Package
  - *Creates a 'soup' of html tags*

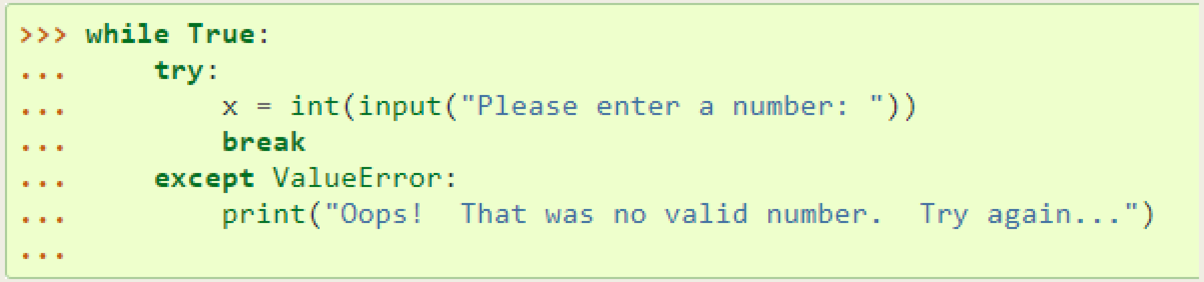


```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN" "http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd">
<html xml:lang="en" xmlns="http://www.w3.org/1999/xhtml">
 <head>
  <title>
   J! Archive - Show #8215, aired 2020-05-01
  </title>
  <link href="http://www.j-archive.com/styles.css" rel="styleSheet" type="text/css"/>
  <link href="http://www.j-archive.com/favicon.ico" rel="shortcut icon"/>
  <meta content="Jeopardy Archive Quemars Ahmed Ashleigh McCord Jesse Laymon Final clue response question
uble" name="keywords"/>
  <meta content="An archive of clues and players for Jeopardy! show #8215." name="description"/>
 </head>
 <body>
  <div id="navbar">
   <span id="navbarlogo">
    <a href="http://www.j-archive.com">
     <img alt="J! Archive" height="22" src="http://www.j-archive.com/j-a.gif" width="100"/>
    </a>
   </span>
   <span id="navbartext">
```

- Where is the content you want?
- How is it stored on the site?

# Tip 3:
# Error Catching

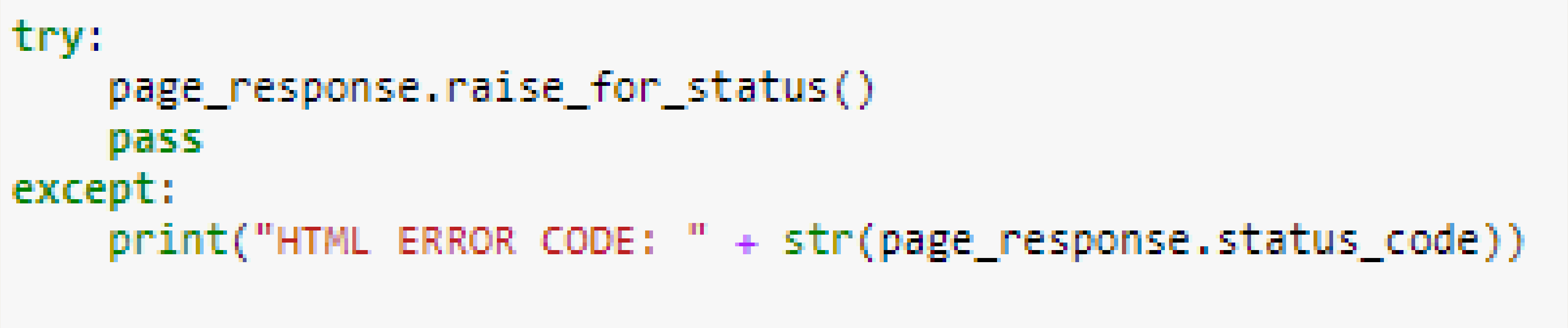
- This will happen!!
- Important for ensuring the quality of the data captured.



```
>>> while True:
...     try:
...         x = int(input("Please enter a number: "))
...         break
...     except ValueError:
...         print("Oops!  That was no valid number.  Try again...")
...
```

- Try clause will be executed as normal
- If and only if try clause breaks is the exception raised. This is where you handle error management.

# Error handling example



```
try:
    page_response.raise_for_status()
    pass
except:
    print("HTML ERROR CODE: " + str(page_response.status_code))
```

# ADDITIONAL TIPS / IDEAS

# R vs. Python for web scrapping?

- Use whatever you are comfortable with!

- Use the rvest package in R
  - *Similar to Beautiful Soup*

- At the end of the day, the Selector Gadget is the tool that will be the most help to you.

# A few websites that might be worth scraping!

- *LAist*
  - *https://laist.com/search_query.php?q=homelessness*
  - *How has reporting on homelessness changed since quarantine?*
- *Law360*
  - *https://www.law360.com/articles/1252836/coronavirus-the-latest-court-closures-and-restrictions*
  - *How fast did courts respond to quarantine measures?*
- *Craigslist*
  - *https://losangeles.craigslist.org/*
  - *What are people selling / buying during a pandemic ?*

# Some final thoughts

- When in doubt: Stack Overflow!
- Additional Materials on HTML: https://www.w3schools.com/html/default.asp
- Additional Materials on Regex: https://docs.python.org/3/library/re.html
- Start simple and then expand your data collection!
  - *Don't try to capture too much data all at once. Get a single page working then add more dynamic features.*
  - *Dynamic sites? Check out the Selenium library in Python for additional tools.*

# THANK YOU!

Add me on LinkedIn!