

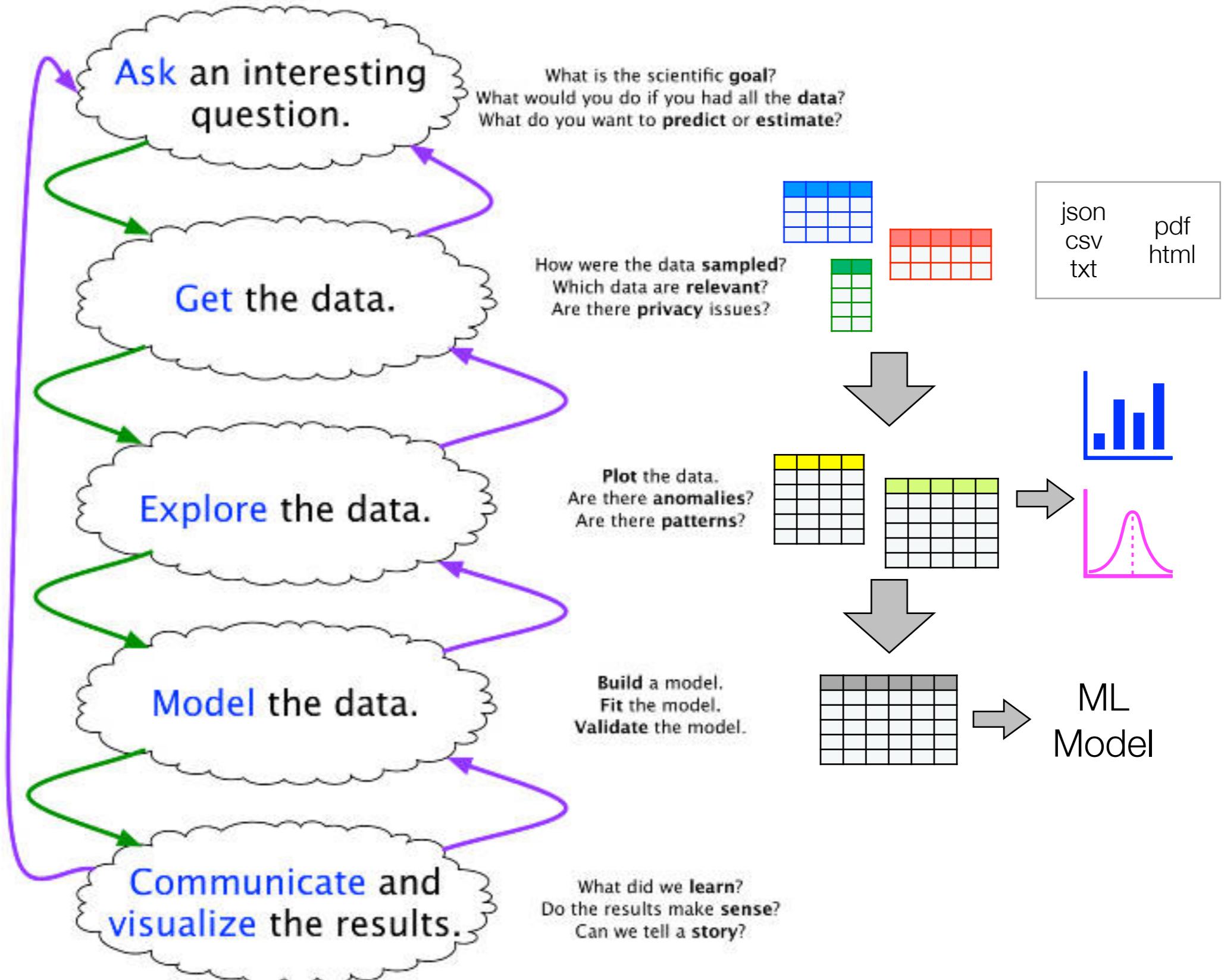


2110531 - Data Science and Data Engineering Tools

## Data Extraction

**Asst.Prof. Natawut Nupairoj, Ph.D.**  
Department of Computer Engineering  
Chulalongkorn University  
[natawut.n@chula.ac.th](mailto:natawut.n@chula.ac.th)

## The Data Science Process



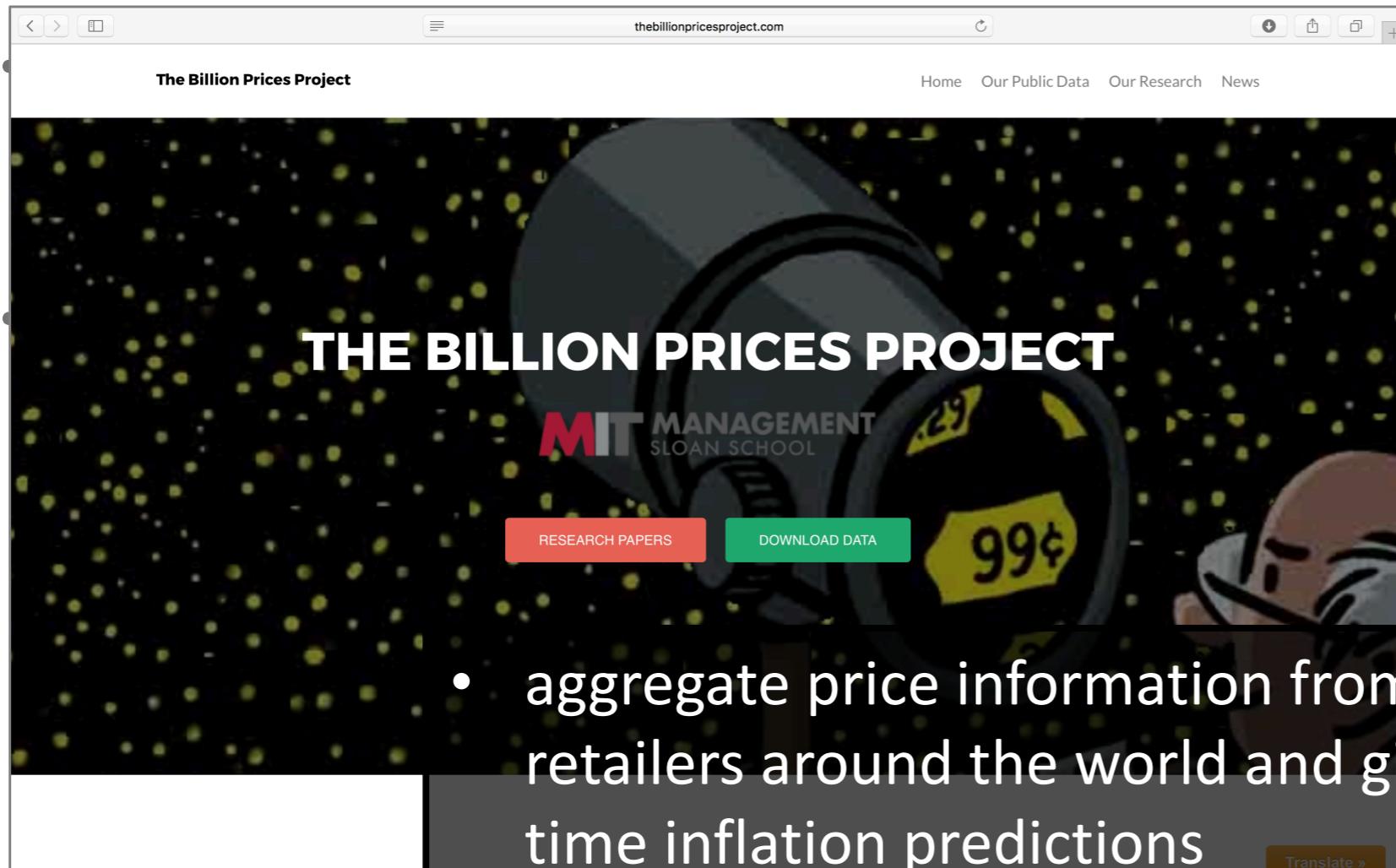
Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course <http://cs109.org/>.

# Internet Data Sources

---

- Internet provides lots of data sources
  - Financial
  - Weather
  - Sports
  - News
- Google creates business by utilizing Internet information
- Most data sources are human-readable, it is also possible to extract information systematically with program

# MIT's Billion Prices Project



data structures and

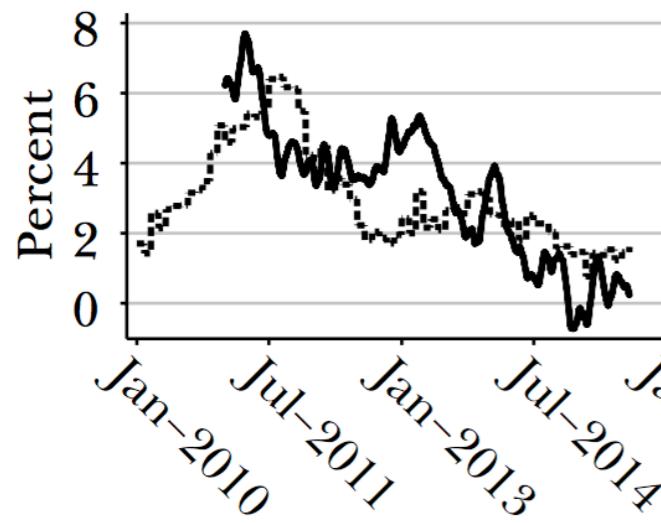
data with

- aggregate price information from multitude of online retailers around the world and gives real time inflation predictions
- Well integrated monitor daily price fluctuations of ~5 million items sold by ~300 online retailers in more than 70 countries

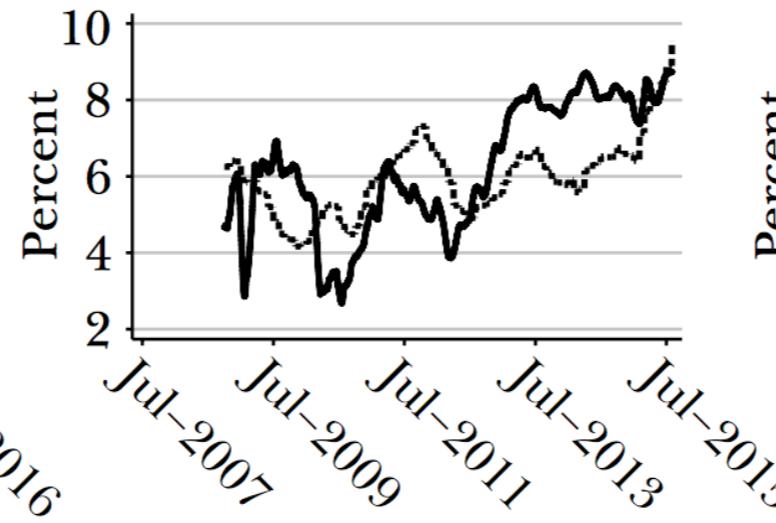
Cavallo, A., & Rigobon, R. (2016). The billion prices project: Using online prices for measurement and research. *The Journal of Economic Perspectives*, 30(2), 151-178.

# Annual Inflation Rate: Online vs. CPI

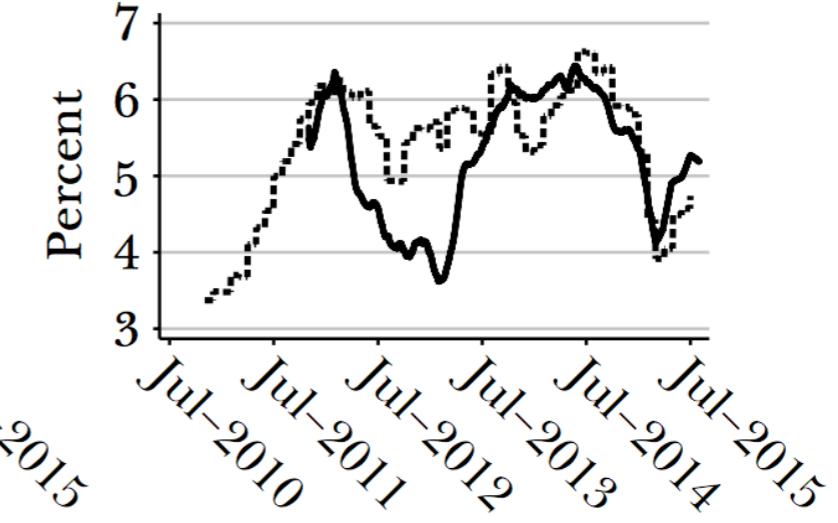
A: China



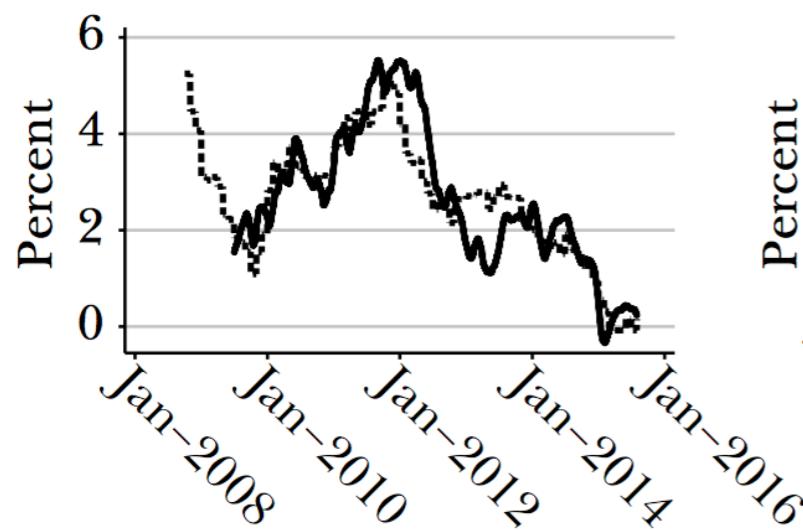
B: Brazil



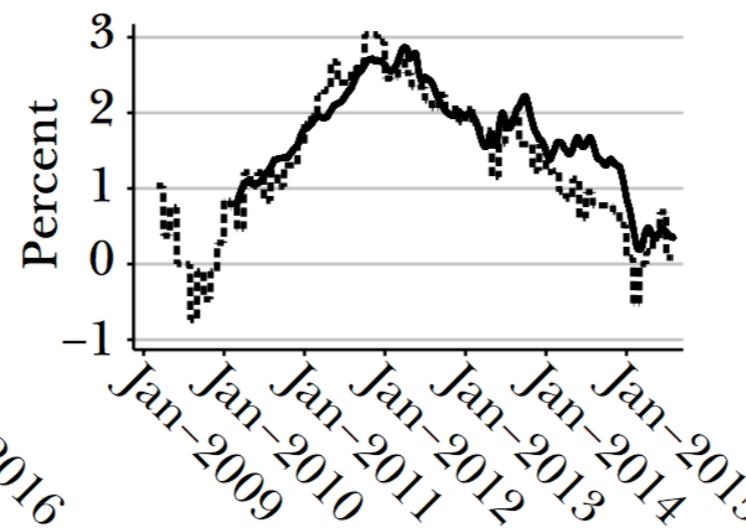
C: South Africa



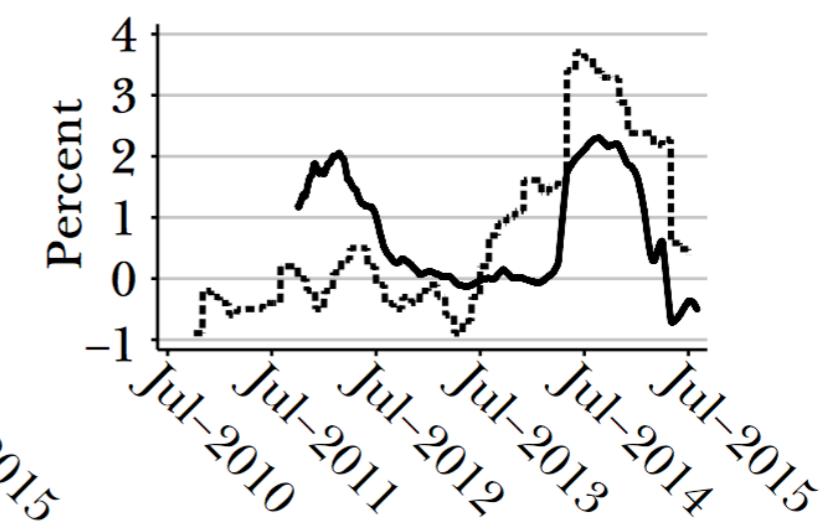
D: United Kingdom



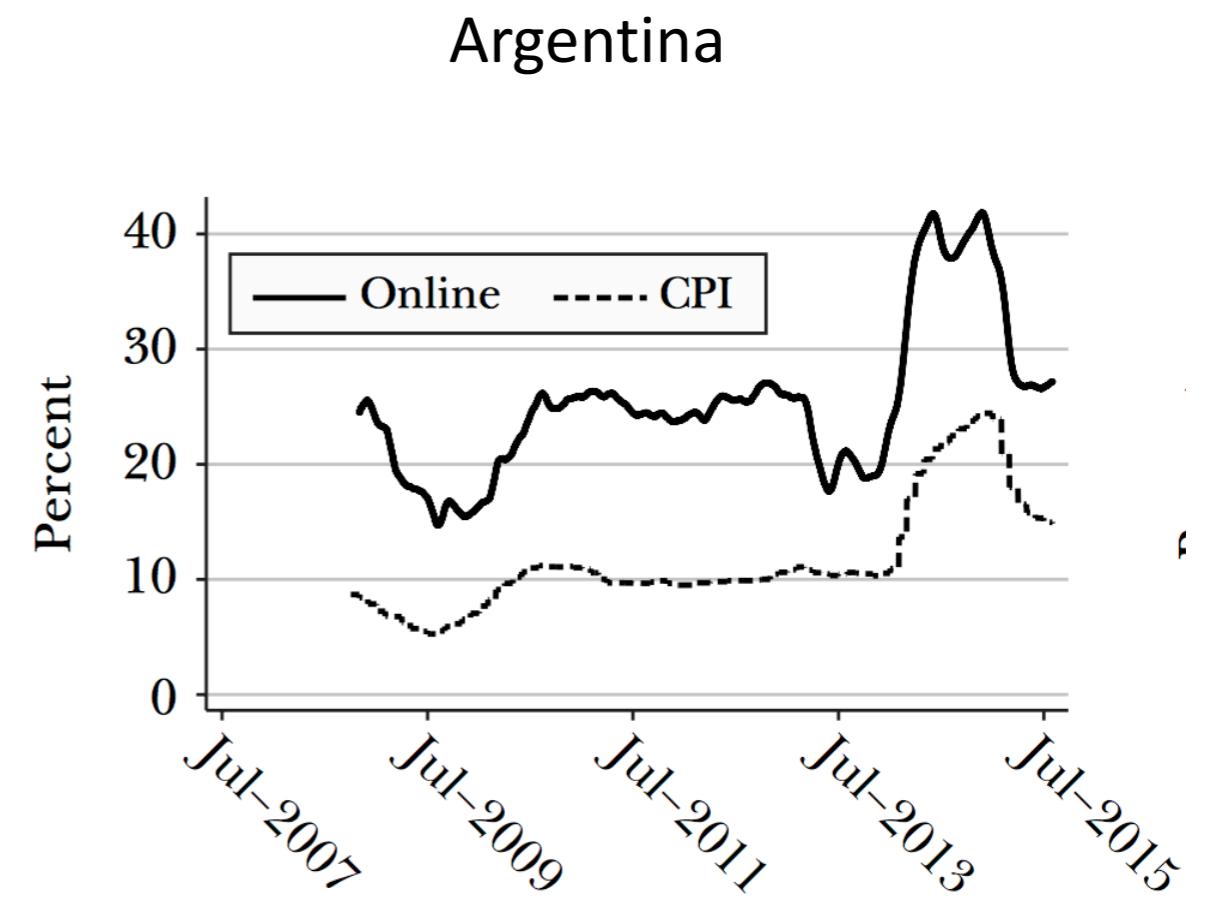
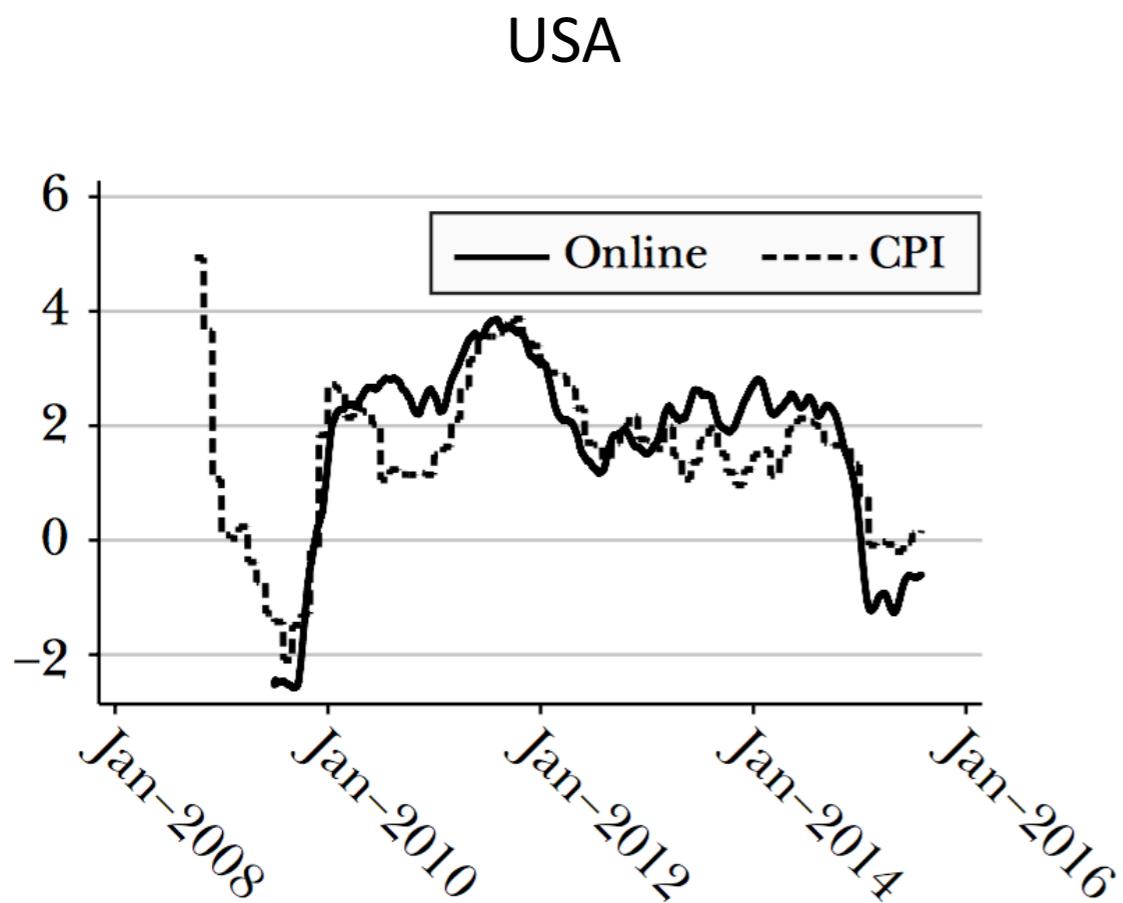
E: Germany

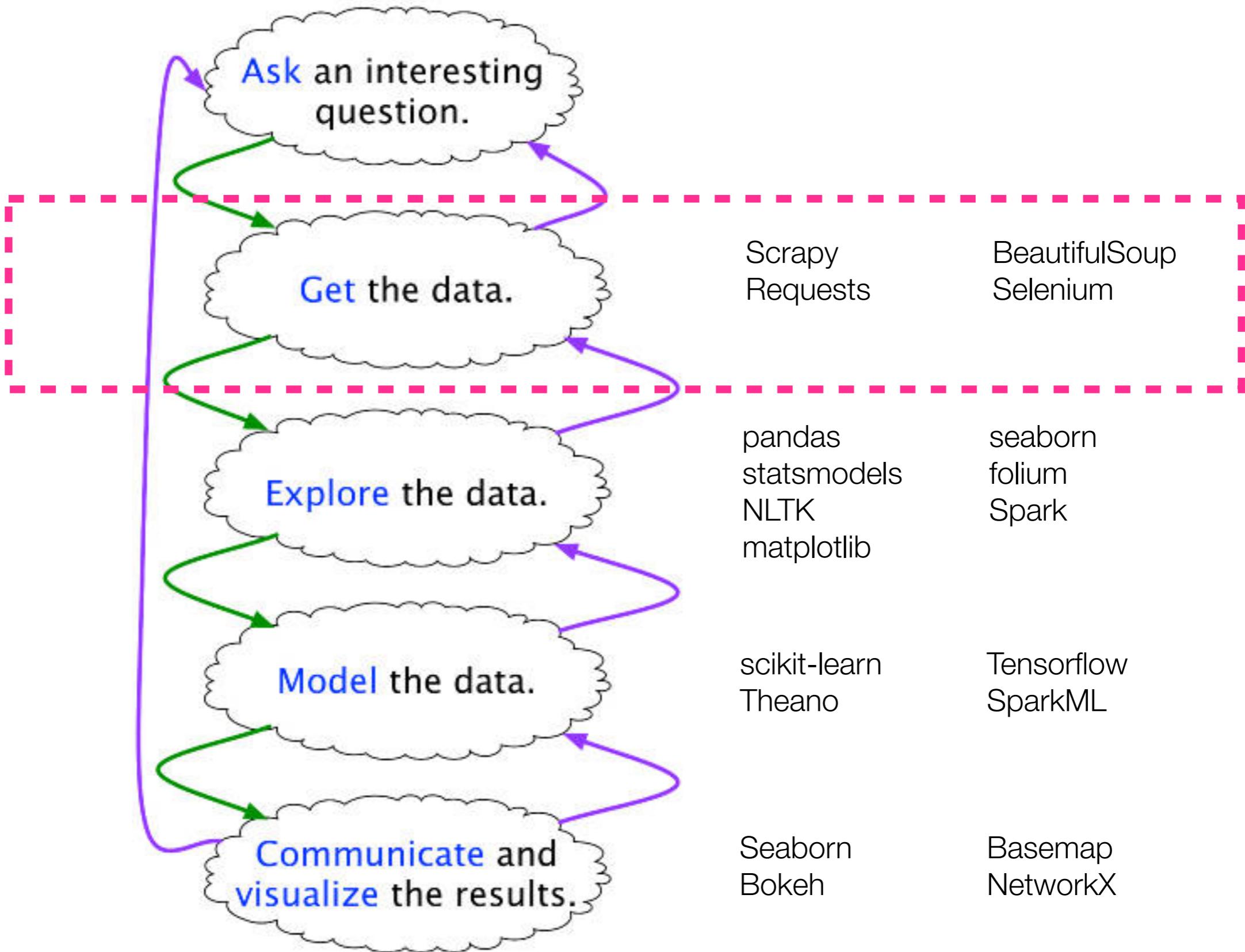


F: Japan



# Annual Inflation Rate: Online vs. CPI





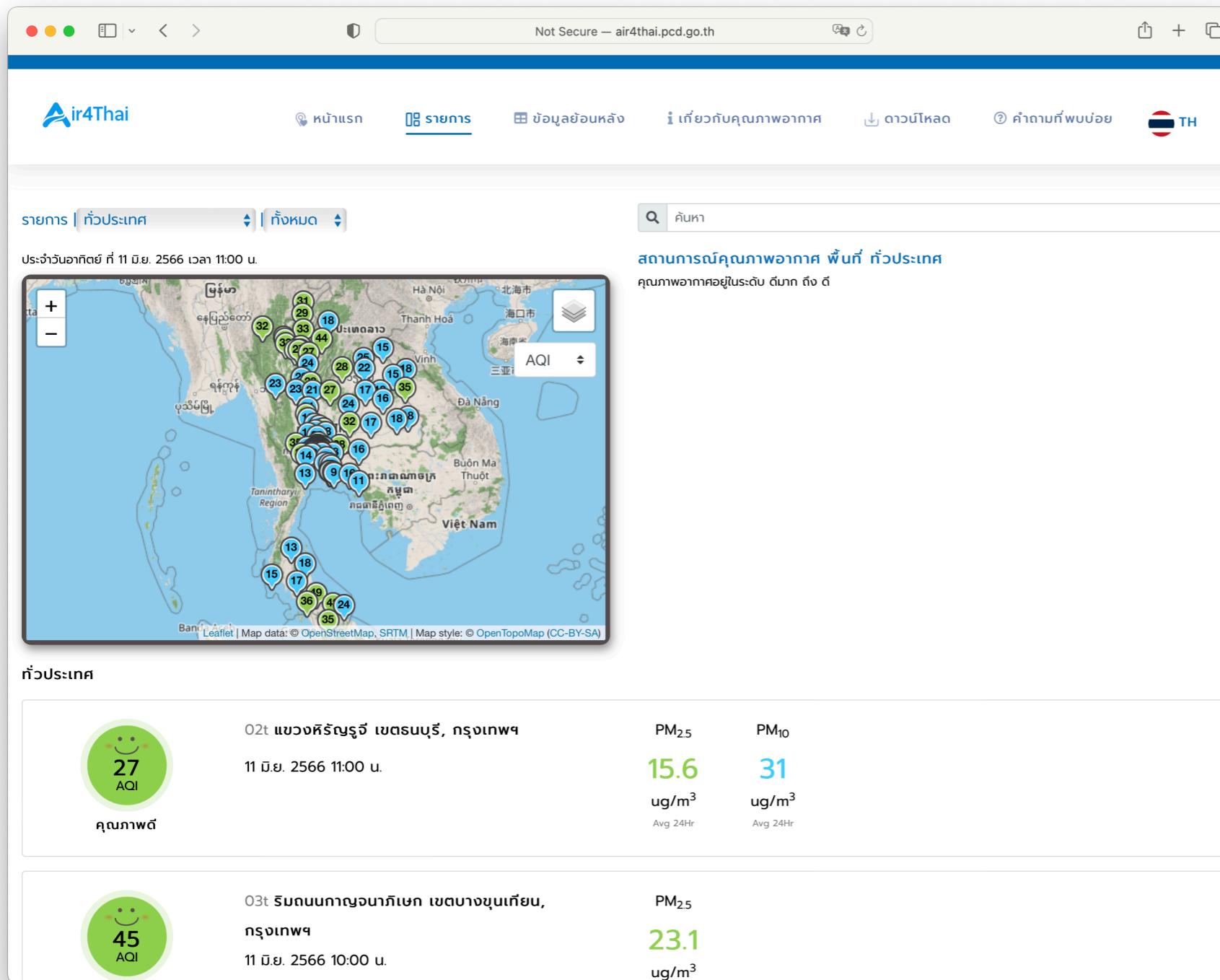
# Data Extraction

---

- Data extraction is the process of getting data from a source for further data processing, storage or analysis elsewhere
- Many data sources are in unstructured or human-readable format
  - Gather organization structure from company's website
  - Extract accounting information from PDF financial reports
  - Collect relationship information from social media platform

# Example Data Source: Air4Thai

<http://air4thai.pcd.go.th/webV3/#/Report>



# Data Extraction Steps

---

- Data Accessing
  - Reading a file
  - Issue a HTTP request
  - Call REST API using Request
  - Call Third-Party API using compatible library
  - Collect complicated web page with Selenium

# Data Extraction Steps

---

- Data Parsing
  - Parsing HTML page using BeautifulSoup
  - Parsing text using regular expression
  - Parsing docx using python-docx
  - Parsing excel using Pandas
  - Parsing JSON string using JSON

# Data Extraction Steps

---

- Data Loading
  - Save to database
  - Save to data file
  - Save to dataframe

# Scraping a Web Page

---

Understand HTML and how HTTP works



The screenshot shows the Gmail inbox interface on a Mac OS X system. The window title is "mail.google.com". The left sidebar includes "Compose", "Mail" (with 2 notifications), "Inbox" (selected, 2 notifications), "Starred", "Snoozed", "Sent", "Drafts", and "More". The main area displays two emails from "ทีมงาน Gmail". The top email is titled "เคล็ดลับในการใช้กล่องจดหมายโฉมใหม่ - สวัสดิ์ Natawut" and the bottom one is "พีเจอร์ที่ดีที่สุดของ Gmail จากทุกที่ - สวัสดิ์ Natawut". A "Get started with Gmail" card offers customization options like "Customize your inbox", "Set a signature", and "Enable desktop notifications". The bottom of the screen shows usage statistics ("Using 0.08 GB"), program policies ("Program Policies, Powered by Google"), and account activity ("Last account activity: 0 minutes ago, Open in 1 other location · Details").

Compose

Mail

Inbox 2

Starred

Snoozed

Sent

Drafts

More

Labels

Using 0.08 GB

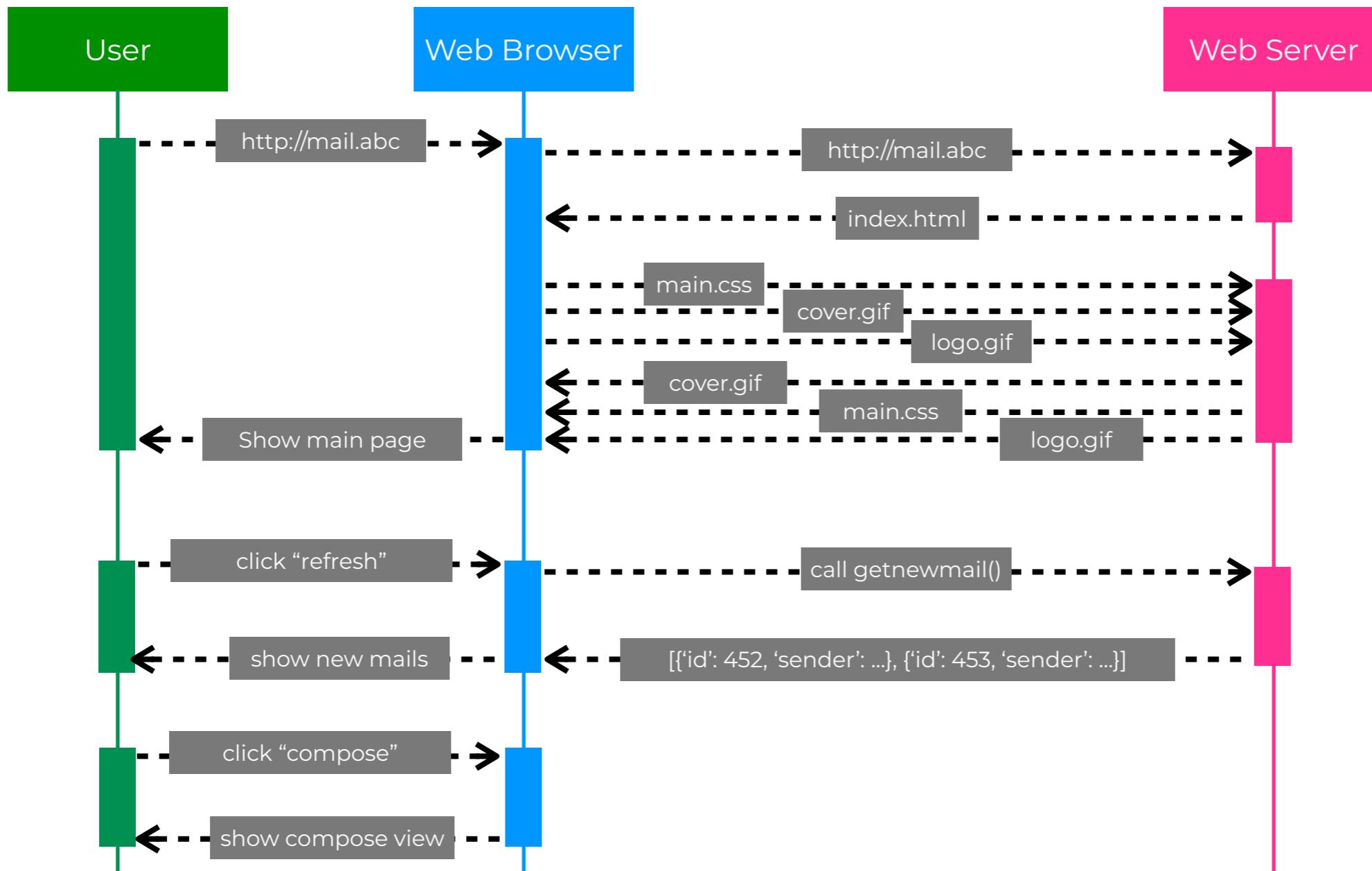
Program Policies

Powered by Google

Last account activity: 0 minutes ago

Open in 1 other location · Details

# Understand Web Page Mechanism



# HTML

---

- HTML is a markup language that defines the structure of a web page using series of elements to wrap different parts of the content to make it appear a certain way, or act a certain way
- For example, <p> is a tag to indicate a paragraph

```
<p>Hello, my name is Natawut Nupairoj. Nice to meet you!</p>
```

- Furthermore, HTML element can be nested

```
<p>Hello, my name is <b>Natawut Nupairoj</b>. Nice to meet you!</p>
```

- A web browser will show this snippet as followed

Hello, my name is **Natawut Nupairoj**. Nice to meet you!

# More on HTML Element

---

- HTML element can have an attribute to provide extra information regarding to this element

```
<p class="greeting">Hello, my name is <b>Natawut Nupairoj</b>. Nice to meet you!</p>
```

Note that “class” is an attribute that can be used in conjunction with CSS

- Some elements have no content and are called void elements

```

```

# Anatomy of HTML Page (my\_page.html)

---

```
<!DOCTYPE html>
<html lang="en-US">
  <head>
    <meta charset="utf-8" />
    <meta name="viewport" content="width=device-width" />
    <title>My test page</title>
  </head>
  <body>
    <p class="greeting">Hello, my name is <b>Natawut Nupairoj</b>. Nice to meet you!
  </p>
    
  </body>
</html>
```

---

## Greeting

Hello, my name is **Natawut Nupairoj**. Nice to meet you!



# CSS

---

- CSS (Cascading Style Sheets) is a declarative language that controls how webpages look in the browser
  - Font, size, color, background color, spacing
  - Alignment, number of columns
- The browser applies CSS style declarations to selected elements to display them properly
- A style declaration contains the properties and their values, which determine how a webpage looks

# CSS Rules and Selectors

---

- A CSS rule is a set of **properties** associated with a **selector**
- Here is an example that makes every HTML paragraph yellow against a black background:

```
/* The selector "p" indicates that all paragraphs in the document will be
affected by that rule */
p {
    /* The "color" property defines the text color, in this case yellow. */
    color: yellow;

    /* The "background-color" property defines the background color, in this case
black. */
    background-color: black;
}
```

- "Cascading" refers to the rules that govern how selectors are prioritized to change a page's appearance

```
<!DOCTYPE html>
<html lang="en-US">
  <head>
    <meta charset="utf-8" />
    <meta name="viewport" content="width=device-width" />
    <style>
      p {
        color: yellow;
        background-color: black;
      }
    </style>
    <title>My test page</title>
  </head>
  <body>
    <p class="greeting">Hello, my name is <b>Natawut Nupairoj</b>. Nice to meet you!</p>
    
  </body>
</html>
```

Note that CSS can be embedded in an HTML page or in a separated file

## Greeting

Hello, my name is **Natawut Nupairoj**. Nice to meet you!



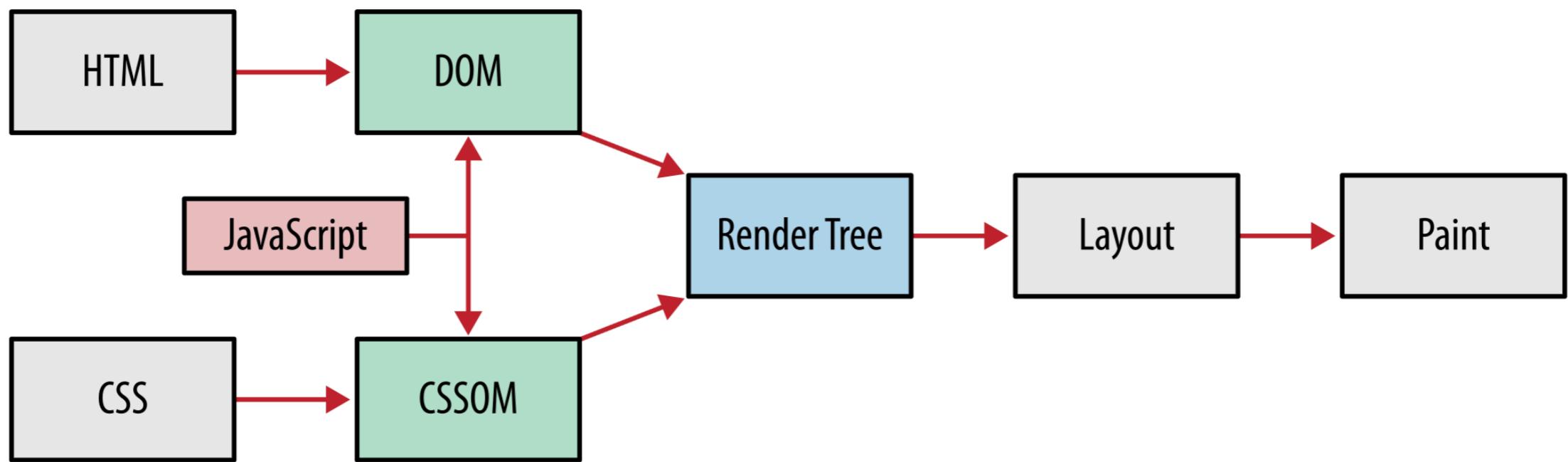
# CSS Selector

---

- CSS Selector is a pattern of elements and other terms that tell the browser which HTML elements should be selected to have the CSS property values inside the rule applied to them
  - Elements with a specific tag e.g. all H1 elements
  - Elements with a specific class / id / attribute e.g. all elements with class = “special”
  - Combination of selectors e.g.
    - all H1 elements with class = “special”
    - all H1 elements or elements with class = “special”
- A selector may select multiple elements at once

# Understand Web Page Mechanism

---



[https://www.w3schools.com/howto/tryit.asp?filename=tryhow\\_css\\_example\\_website](https://www.w3schools.com/howto/tryit.asp?filename=tryhow_css_example_website)

```
<!DOCTYPE html>
<html lang="en">
<head>
<title>Page Title</title>
<meta charset="UTF-8">
<meta name="viewport" content="width=device-width, initial-scale=1">
<style>
* {
  box-sizing: border-box;
}

/* Style the body */
body {
  font-family: Arial, Helvetica, sans-serif;
  margin: 0;
}

/* Header/logo Title */
.header {
  padding: 80px;
  text-align: center;
  background: #1abc9c;
  color: white;
}

/* Increase the font size of the heading */
.header h1 {
  font-size: 40px;
}

/* Sticky navbar - toggles between relative and fixed, depending on the scroll position. It is positioned relative until a given offset position is met in the viewport - then it "sticks" in place (like position:fixed). The sticky value is not supported in IE or Edge 15 and earlier versions. However, for these versions the navbar will inherit default position */
.navbar {
  overflow: hidden;
  background-color: #333;
  position: sticky;
  position: -webkit-sticky;
  top: 0;
}

/* Style the navigation bar links */
.navbar a {
  float: left;
  display: block;
  color: white;
  text-align: center;
  padding: 14px 20px;
  text-decoration: none;
}
```

My Website

A responsive website created by me.

Home    Link    Link    Link

**About Me**

Photo of me:

Image

Some text about me in culpa qui officia deserunt mollit anim..

**More Text**

Lore ipsum dolor sit ame.

Image

Image

**TITLE HEADING**

Title description, Dec 7, 2017

Image

Some text..

Sunt in culpa qui officia deserunt mollit anim id est laborum consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco.

**TITLE HEADING**

Title description, Sep 2, 2017

Image

```
<!DOCTYPE html>
<html lang="en">
<head>
<title>Page Title</title>
<meta charset="UTF-8">
<meta name="viewport" content="width=device-width, initial-scale=1">
<style>
* {
  box-sizing: border-box;
}

/* Style the body */
body {
  font-family: Arial, Helvetica, sans-serif;
  margin: 0;
}

/* Header/logo Title */
.header {
  padding: 80px;
  text-align: center;
  background: #1abc9c;
  color: white;
}

/* Increase the font size of the heading */
.header h1 {
  font-size: 40px;
}

/* Sticky navbar – toggles between relative and fixed, depending on the scroll position. It is positioned relative until a given offset position is met in the viewport – then it "sticks" in place (like position:fixed). The sticky value is not supported in IE or Edge 15 and earlier versions. However, for these versions the navbar will inherit default position */
.navbar {
  overflow: hidden;
  background-color: #333;
  position: sticky;
  position: -webkit-sticky;
  top: 0;
}
```

```
<body>

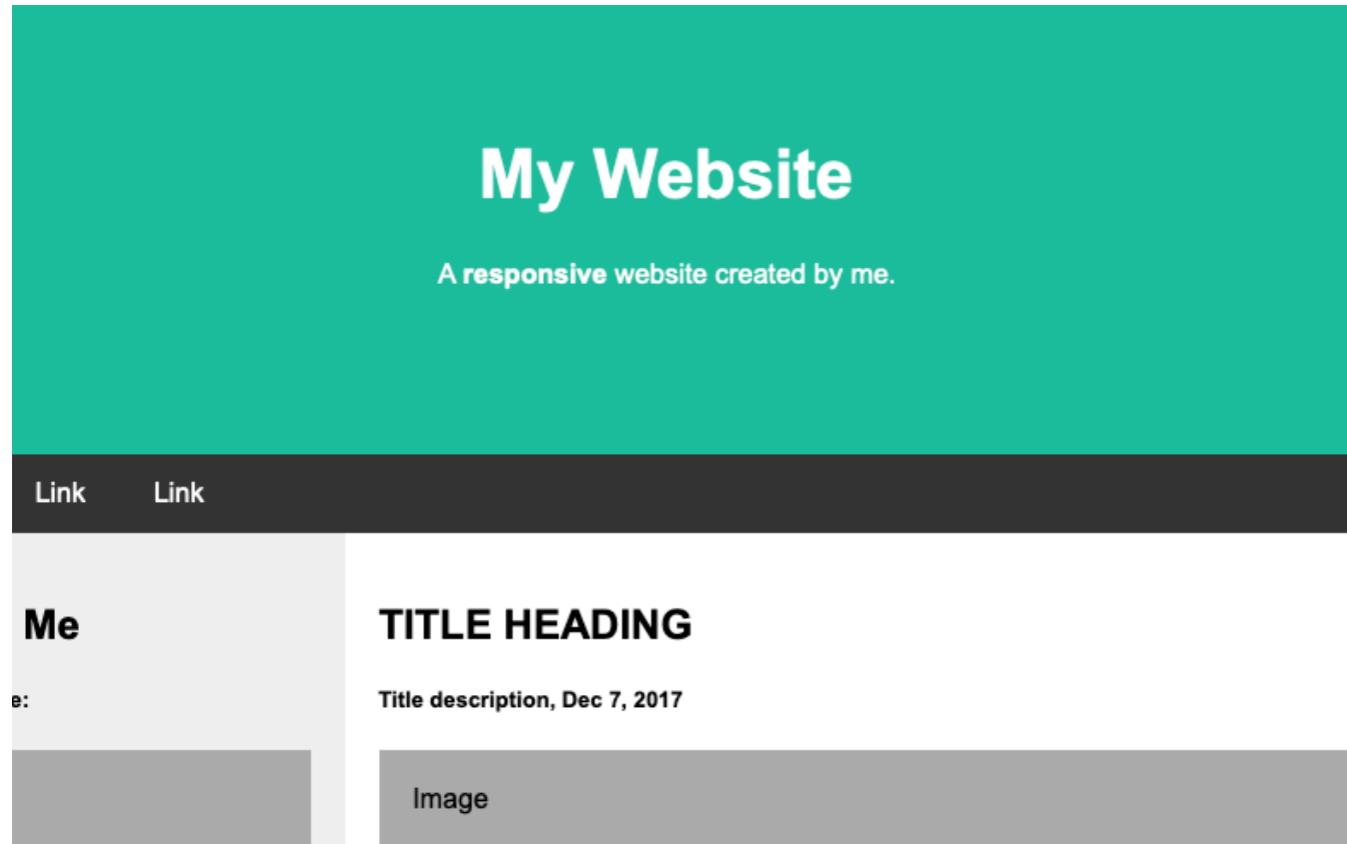
<div class="header">
  <h1>My Website</h1>
  <p>A <b>responsive</b> website created by me.</p>
</div>

<div class="navbar">
  <a href="#" class="active">Home</a>
  <a href="#">Link</a>
  <a href="#">Link</a>
  <a href="#" class="right">Link</a>
</div>

<div class="row">
  <div class="side">
    <h2>About Me</h2>
    <h5>Photo of me:</h5>
    <div class="fakeimg" style="height:200px;">Image</div>
    <p>Some text about me in culpa qui officia deserunt mollit anim..</p>
    <h3>More Text</h3>
    <p>Lorem ipsum dolor sit ame.</p>
    <div class="fakeimg" style="height:60px;">Image</div><br>
    <div class="fakeimg" style="height:60px;">Image</div><br>
    <div class="fakeimg" style="height:60px;">Image</div>
  </div>
  <div class="main">
    <h2>TITLE HEADING</h2>
    <h5>Title description, Dec 7, 2017</h5>
    <div class="fakeimg" style="height:200px;">Image</div>
    <p>Some text..</p>
    <p>Sunt in culpa qui officia deserunt mollit anim id est laborum consectetur adipisciing elit, sed
do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud
exercitation ullamco.</p>
    <br>
    <h2>TITLE HEADING</h2>
    <h5>Title description, Sep 2, 2017</h5>
    <div class="fakeimg" style="height:200px;">Image</div>
    <p>Some text..</p>
    <p>Sunt in culpa qui officia deserunt mollit anim id est laborum consectetur adipisciing elit, sed
do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud
exercitation ullamco.</p>
  </div>
</div>

<div class="footer">
  <h2>Footer</h2>
</div>

</body>
```



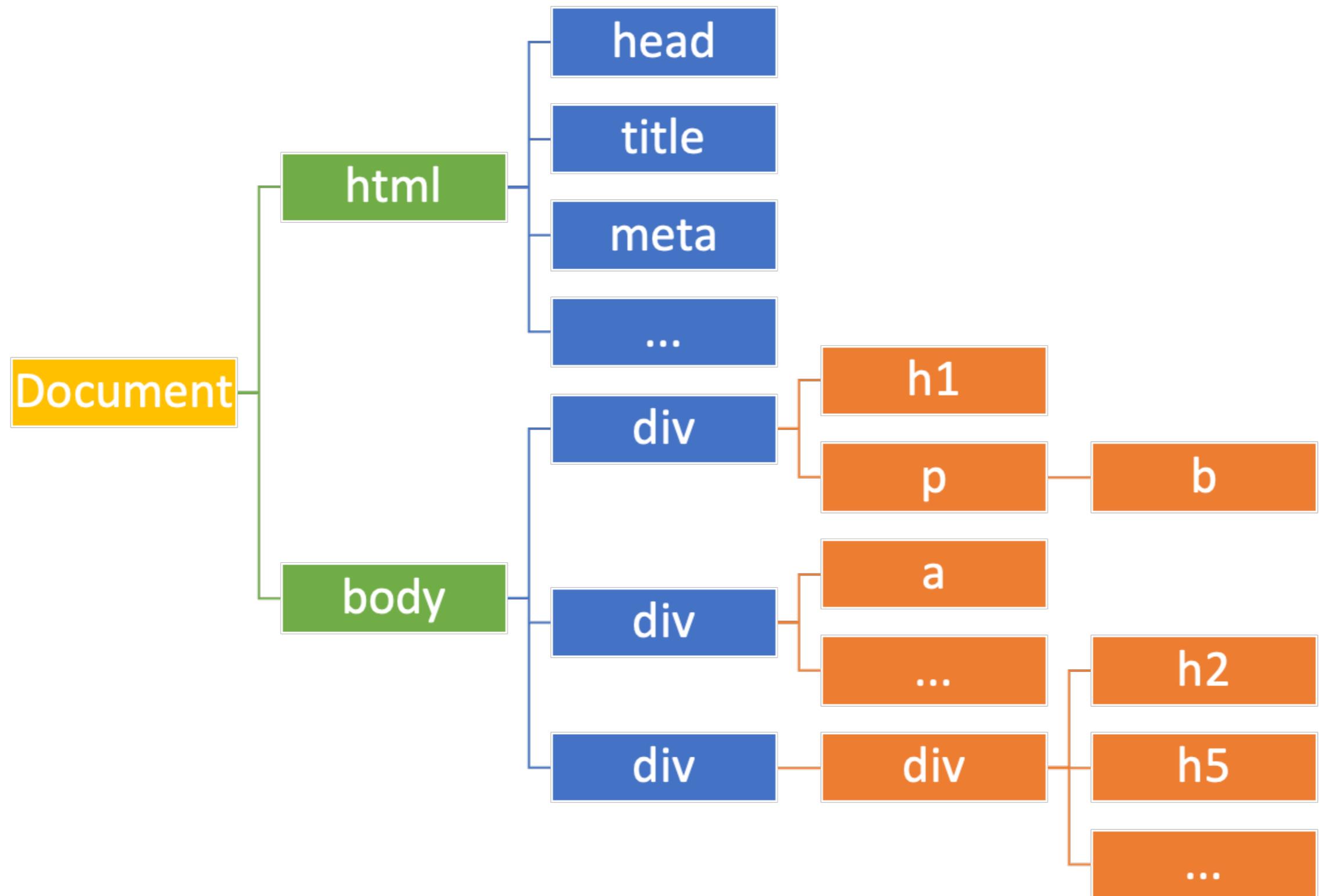
css

```
/* Header/logo Title */
.header {
  padding: 80px;
  text-align: center;
  background: #1abc9c;
  color: white;
}

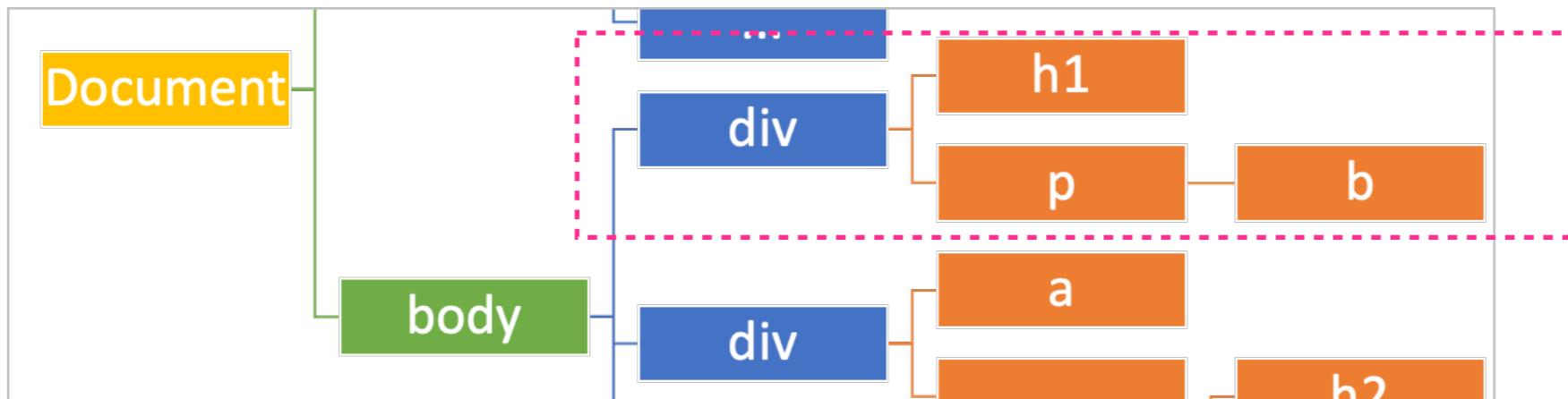
/* Increase the font size of the heading */
.header h1 {
  font-size: 40px;
}
```

body

```
<div class="header">
  <h1>My Website</h1>
  <p>A <b>responsive</b> website created by me.</p>
</div>
```



# DOM Node Example



```
<div class="header">
  <h1>My Website</h1>
  <p>A <b>responsive</b> website created by me.</p>
</div>
```

- Node 'div' has 2 children (h1 and p) and 1 attribute (key='class', value='header')

# Parsing A Simple Webpage

- Applicable for a simple page only
  - No content rendered by javascript
  - Parse HTML with BeautifulSoup
- BeautifulSoup allows
  - Parsing HTML
  - Accessing and Navigating the DOM tree
  - Getting information from nodes

## Beautiful Soup Documentation

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

These instructions illustrate all major features of Beautiful Soup 4, with examples. I show you what the library is good for, how it works, how to use it, how to make it do what you want, and what to do when it violates your expectations.

This document covers Beautiful Soup version 4.9.1. The examples in this documentation should work the same way in Python 2.7 and Python 3.2.

You might be looking for the documentation for [Beautiful Soup 3](#). If so, you should know that Beautiful Soup 3 is no longer being developed and that support for it will end on December 31, 2020. If you want to learn about the differences between Beautiful Soup 3 and Beautiful Soup 4, see [Porting code to BS4](#).

This documentation has been translated into other languages by Beautiful Soup users:

- 这篇文档当然还有中文版.
- このページは日本語で利用できます([外部リンク](#))
- 이 문서는 한국어 번역도 가능합니다.
- Este documento também está disponível em Português do Brasil.
- Эта документация доступна на русском языке.

## Getting help

If you have questions about Beautiful Soup, or run into problems, [send mail to the discussion list](#). If your problem involves parsing an HTML document, be sure to mention [what the diagnose\(\)](#) function finds in your document.

## Quick Start

Here's an HTML document I'll be using as an example throughout this document. It's part of Lewis Carroll's *Alice's Adventures in Wonderland*:

```
html_doc = """
<html><head><title>The Dormouse's story</title></head>
<body>
<p class="title"><b>The Dormouse's story</b></p>

<p class="story">Once upon a time there were three little sisters; and their names
<a href="http://example.com/elsie" class="sister" id="link1">Elsie</a>,
<a href="http://example.com/lacie" class="sister" id="link2">Lacie</a> and
<a href="http://example.com/tillie" class="sister" id="link3">Tillie</a>;
and they lived at the bottom of a well.</p>

<p class="story">...</p>
"""
```

Running the "three sisters" document through Beautiful Soup gives us a `BeautifulSoup` object containing the document as a nested data structure:



jupyter 1 - Basic Webpage Scraping (unsaved changes)  Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

# Basic Webpage Scraping

Webpage scraping consists of two steps: crawling and parsing. In this tutorial, we focus on parsing HTML data. BeautifulSoup is a powerful tool to process static HTML. More details can be found at <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

To simplify our learning, we will use a simple example from W3Schools:

[https://www.w3schools.com/howto/tryit.asp?filename=tryhow\\_css\\_example\\_website](https://www.w3schools.com/howto/tryit.asp?filename=tryhow_css_example_website)

```
In [ ]: import sys
IN_COLAB = 'google.colab' in sys.modules
if IN_COLAB:
    !wget https://www.dropbox.com/s/w5khgpro1ym3icg/simple_page.html?dl=1
```

```
In [ ]: with open('simple_page.html') as f:
    html = f.read()
```

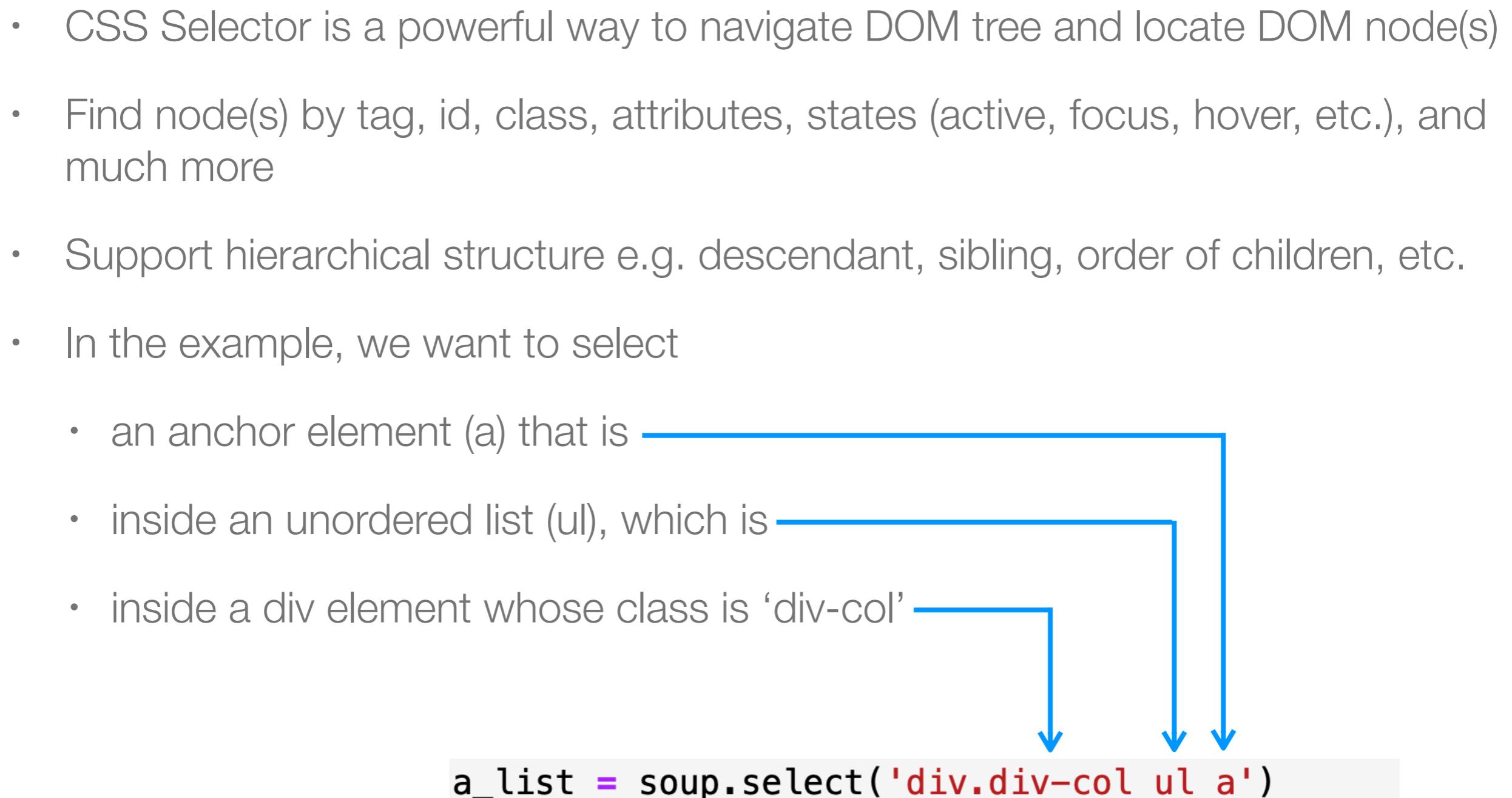
```
In [ ]: html
```

```
In [ ]: from bs4 import BeautifulSoup
from bs4.element import Tag
from IPython.core.display import HTML
```

```
In [ ]: soup = BeautifulSoup(html, "lxml")
print(soup.prettify())
```

# Understand CSS Selector Syntax

- CSS Selector is a powerful way to navigate DOM tree and locate DOM node(s)
- Find node(s) by tag, id, class, attributes, states (active, focus, hover, etc.), and much more
- Support hierarchical structure e.g. descendant, sibling, order of children, etc.
- In the example, we want to select
  - an anchor element (a) that is
  - inside an unordered list (ul), which is
  - inside a div element whose class is ‘div-col’



```
a_list = soup.select('div.div-col ul a')
```

Selector	Example	Example description
<u>.class</u>	.intro	Selects all elements with class="intro"
<u>.class1.class2</u>	.name1.name2	Selects all elements with both <i>name1</i> and <i>name2</i> set within its class attribute
<u>.class1 .class2</u>	.name1 .name2	Selects all elements with <i>name2</i> that is a descendant of an element with <i>name1</i>
<u>#id</u>	#firstname	Selects the element with id="firstname"
<u>*</u>	*	Selects all elements
<u>element</u>	p	Selects all <p> elements
<u>element.class</u>	p.intro	Selects all <p> elements with class="intro"
<u>element,element</u>	div, p	Selects all <div> elements and all <p> elements
<u>element element</u>	div p	Selects all <p> elements inside <div> elements
<u>element&gt;element</u>	div > p	Selects all <p> elements where the parent is a <div> element
<u>element+element</u>	div + p	Selects the first <p> element that is placed immediately after <div> elements
<u>element1~element2</u>	p ~ ul	Selects every <ul> element that is preceded by a <p> element
<u>[attribute]</u>	[target]	Selects all elements with a target attribute
<u>[attribute=value]</u>	[target=_blank]	Selects all elements with target="_blank"
<u>[attribute~=value]</u>	[title~=flower]	Selects all elements with a title attribute containing the word "flower"
<u>[attribute =value]</u>	[lang =en]	Selects all elements with a lang attribute value equal to "en" or starting with "en-"
<u>[attribute^=value]</u>	a[href^="https"]	Selects every <a> element whose href attribute value begins with "https"

# Example: Extracting references from a wiki page

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

## Big data

From Wikipedia, the free encyclopedia

*This article is about large collections of data. For the band, see [Big Data \(band\)](#). For buying and selling of personal and consumer data, see [Surveillance capitalism](#).*

This article **may contain an excessive number of citations**. Please consider removing references to unnecessary or disreputable sources, merging citations where possible, or, if necessary, flagging the content for deletion. In particular many references are "spammed" here for promotional purposes. These need to be removed.. (November 2019) ([Learn how and when to remove this template message](#))

**Big data** is a field that treats ways to analyze, systematically extract information from, or otherwise deal with **data sets** that are too large or complex to be dealt with by traditional **data-processing application software**. Data with many cases (rows) offer greater **statistical power**, while data with higher complexity (more attributes or columns) may lead to a higher **false discovery rate**.<sup>[2]</sup> Big data challenges include capturing data, **data storage**, **data analysis**, search, sharing, transfer, **visualization**, **querying**, updating, **information privacy** and data source. Big data was

Year	Storage Type	Capacity (exabytes)
1986	ANALOG	2.6 exabytes
1986	DIGITAL	0.02 exabytes
2000	ANALOG STORAGE	~1 exabyte
2002	DIGITAL STORAGE	~19 exabytes
2007	ANALOG	19 exabytes
2007	DIGITAL	123 billion gigabytes

Contents [[hide](#)]

(Top)

> [Definition](#)

[Characteristics](#)

[Architecture](#)

[Technologies](#)

> [Applications](#)

> [Case studies](#)

> [Research activities](#)

> [Critique](#)

[See also](#)

[References](#)

[Further reading](#)

[External links](#)

- Encouraging members of society to abandon interactions with institutions that would create a digital trace, thus creating obstacles to social inclusion

If these potential problems are not corrected or regulated, the effects of big data policing may continue to shape societal hierarchies. Conscientious usage of big data policing could prevent individual level biases from becoming institutional biases, Brayne also notes.

## See also [edit]

*For a list of companies, and tools, see also: [Category:Big data](#)*

- |  |   |   |
|--|---|---|
| <ul style="list-style-type: none"> <li>• <a href="#">Big data ethics</a> – Ethics of mass data analytics</li> <li>• <a href="#">Big data maturity model</a> – Aspect of computer science</li> <li>• <a href="#">Big memory</a> – A large amount of random-access memory</li> <li>• <a href="#">Data curation</a> – work performed to ensure meaningful and enduring access to data</li> <li>• <a href="#">Data defined storage</a> – Marketing term for managing data by combining application, information and storage tiers</li> </ul> | <ul style="list-style-type: none"> <li>• <a href="#">Data engineering</a> – Software engineering approach to designing and developing information systems</li> <li>• <a href="#">Data lineage</a> – Origins and events of data</li> <li>• <a href="#">Data philanthropy</a> – Aspect of culture</li> <li>• <a href="#">Data science</a> – Interdisciplinary field of study on deriving knowledge and insights from data</li> <li>• <a href="#">Datafication</a> – Technological trend</li> <li>• <a href="#">Document-oriented database</a> – Type of computer program</li> </ul> | <ul style="list-style-type: none"> <li>• <a href="#">List of big data companies</a></li> <li>• <a href="#">Very large database</a> – type of database containing a very large amount of data, so much that it can require specialized architectural, management, processing and maintenance methodologies</li> <li>• <a href="#">XLDB</a> – annual conference series on databases, data management and analytics</li> </ul> |
|--|---|---|

## References [edit]

1. ^ Hilbert, Martin; López, Priscila (2011). "The World's Technological Capacity to Store, Communicate, and Compute Information". *Science*. **332** (6025): 60–65. Bibcode:2011Sci...332...60H. doi:10.1126/science.1200970. PMID 21310967. S2CID 206531385. Archived from the original on 14 April 2016. Retrieved 13 April 2016.
2. ^ Breur, Tom (July 2016). "Statistical Power Analysis and the contemporary "crisis" in social sciences". *Journal of Marketing Analytics*. London, England: Palgrave Macmillan. **4** (2–3): 61–65. doi:10.1057/s411270-016-0001\_2. ISSN 2050-2218.
104. ^ Josh Rogin (2 August 2018). "Ethnic cleansing makes a comeback – in China". No. Washington Post. Archived from the original on 31 March 2019. Retrieved 4 August 2018. "Add to that the unprecedented security and surveillance state in Xinjiang, which includes all-encompassing monitoring based on identity cards, checkpoints, facial recognition and the collection of DNA from millions of individuals. The authorities feed all this data into an artificial-intelligence machine that rates people's loyalty to the Communist Party in order to control every aspect of their lives."

merging. Conscious use of big data mining can also lead to better decision making.

Brayne also notes.

## See also [edit]

- a** 91.84×16 panies, and tools, see also: [Category:Big data](#)
- [Big data ethics](#) – Ethics of mass data analytics
- [Big data maturity model](#) – Aspect of computer science
- [Big memory](#) – A large amount of random-access memory
- [Data curation](#) – work performed to ensure meaningful and enduring access to data
- [Data defined storage](#) – Marketing term

- [Data engineering](#) – engineering and design of data systems
- [Data lineage](#) – data provenance
- [Data processing](#) – study of data analysis, insight and inference
- [Data science](#) – Data driven insight
- [Datafication](#) –

```

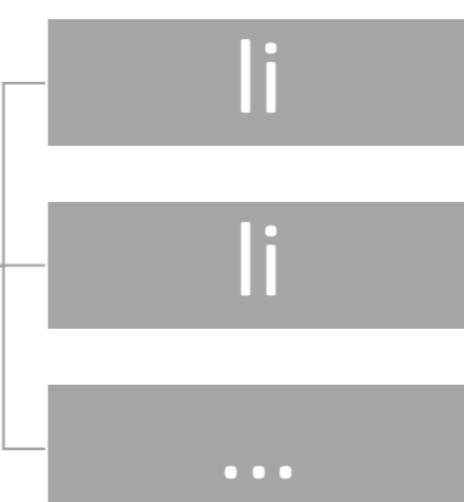
<h2> == $0
  <span class="mw-headline" id="See_also">See also</span>
  <span class="mw-editsection">...</span>
</h2>
<link rel="mw-deduplicated-inline-style" href="mw-data:TemplateStyle096">
<div role="note" class="hatnote navigation-not-searchable selfref">
<style data-mw-duplicate="TemplateStyles:r1147244281">...</style>
<div class="div-col" style="column-width: 15em;">
  <ul>
    <li>
      ::marker
      <a href="/wiki/Big_data_ethics" title="Big data ethics">Big
        data
        &nbsp;– Ethics of mass data a https://en.wikipedia.org/wiki/Big_data
      </a>
    </li>
    <li>...</li>
    <li>...</li>
    <li>...</li>
  </ul>
</div>

```

class = “div-col”

div

ul



Find all these ‘a’

jupyter 2 - Wikipedia page data extraction (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Markdown

## Wikipedia page data extraction

In this tutorial, we will learn how to extract a static page and convert it into useful information.

We first get a wikipedia page using requests.

```
In [ ]: import requests  
import re  
from bs4 import BeautifulSoup
```

```
In [ ]: bigdata = requests.get('https://en.wikipedia.org/wiki/Big_data')
```

```
In [ ]: len(bigdata.text)
```

```
In [ ]: bigdata.text
```

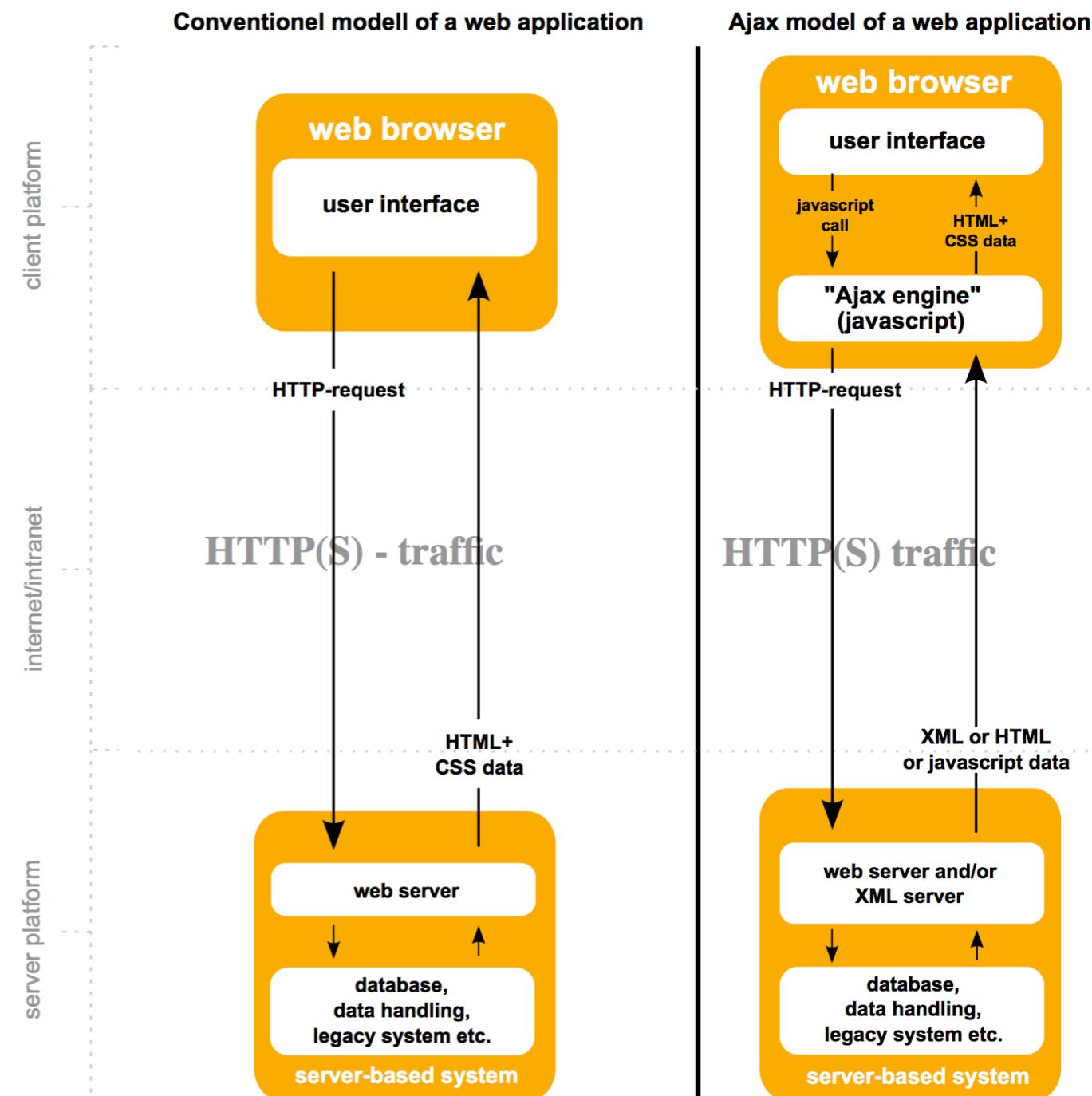
## Parsing a wikipedia page

```
In [ ]: soup = BeautifulSoup(bigdata.text, "lxml")  
print(soup.prettify())
```

```
In [ ]: soup.title.string
```

## Extracting AJAX-based Webpage

- Modern web applications are AJAX-based e.g. SPA
- Typical web request approach does not work



Source: [https://en.wikipedia.org/wiki/Ajax\\_\(programming\)](https://en.wikipedia.org/wiki/Ajax_(programming))

# https://www.settrade.com/settrade/home

หน้าหลัก - SETTRADE.COM - +

settrade.com/th/home

SET : 1,555.11 เปลี่ยนแปลง : - (-) มูลค่าการซื้อขาย (ล้านบาท) : - ปริมาณการซื้อขาย ('000 หน่วย) : - สถานะ : Closed ข้อมูลล่าสุด 10 มิ.ย. 2566 03:20:11

**settrade** www.settrade.com หุ้น อุปทาน กองทุนรวม บกอคراج หัวสาร/บกความ บริการ/เครื่องมือ

จุดประกายไอเดียกี๊ใช่เพื่อการลงทุนในแบบของคุณ

Get Quote ใส่ชื่อย่อหักทรัพย์ (หุ้น, อุปทาน, กองทุนรวม)

Most Active Volume

SET	mai	ปริมาณ AOM ('000 หุ้น)
STARKE	1,115,837	▼ 0.11 -0.03 (-21.43%)
TTB	403,423	▼ 1.66 -0.01 (-0.60%)
WAVE	271,751	▼ 0.19 -0.02 (-9.52%)
SIRI	138,190	▲ 1.90 +0.02 (+1.06%)
BANPU	104,832	▲ 8.75 +0.10 (+1.16%)

SET 1,555.11 - (-) สรุปภาวะตลาด สรุปซื้อขายสินวัน ข้อมูลสถิติ

SET 1,555.11 - (-) 1,559.43 1,549.57 1,560 1,555 1,550 1,545

10:00 12:00 14:00 16:00

Gainers - Unchanges - Losers -

หมายเหตุ • ข้อมูลเฉพาะหักทรัพย์ที่มีการซื้อขายในรูปแบบ AOM • ข้อมูลรวมโดยทั่วไป เน้นวันหลังเวลา 17:30 น. เป็นข้อมูลที่รวม

รายชื่อหักทรัพย์ SET ดูเพิ่มเติม →

กู้บุญธรรม DRx

กู้บุญธรรม DRx

เพื่อช่วยเพิ่มประสิทธิภาพการใช้งาน ท่านสามารถศึกษารายละเอียดเพิ่มเติมได้ที่ [นโยบายการใช้คุกกี้](#)

ยอมรับ

DevTools - www.settrade.com/th/home

Elements Console Sources Network Performance Memory Application Security Lighthouse Recorder ▾ Performance insights ▾ AngularJS

✖ 8 ⚠ 1 ⚡ 6 ⚙ ⚧

Filter Invert Hide data URLs All Fetch/XHR JS CSS Img Media Font Doc WS Wasm Manifest Other Has blocked cookies Blocked Requests 3rd-party requests

Name	Status	Type	Initiator	Size	Time	Waterfall
pm_match?https://simage2.p...	200	document	um.simpli.fi/pm_match?https://simage2.p...	53 B	38 ms	
Pug?vcode=bz0yJnR5cGU9MSZqc...	200	document	pm.w55c.net/ping_match.gif?ei=PUBMAT...	301 B	41 ms	
cs?pid=45&rndcb=7018091133	302	document / Redirect	sync.1rx.io/usersync2/pubmatic&gdpr=0...	435 B	535 ms	
4298321646126704324?dspret=1&gdpr=0&gdpr_consent=&us_privacy=	302	document / Redirect	ad.turn.com/r/cs?pid=45&rndcb=7018091...	664 B	68 ms	
Pug?vcode&piggybackCookie={viewer_token}	200	document	csync.loopme.me/?pubid=11331&redirect...	74 B	39 ms	
cksync.php?cs=1&ovsid=3d8264dc-3139-4be6-8991-58dedda99953&type=loop&gd...	200	gif	csync.loopme.me/_	473 B	97 ms	
RX-a9f8e578-bdd5-43f2-8a3a-4b62a5cdec0b-004?redir=...kie%3DRX-a9f8e578-bd...	302	document / Redirect	sync.1rx.io/usersync/turn/4298321646126...	527 B	206 ms	
Pug?vcode=bz0yJnR5cGU9MSZjb2RIPTMyMDMmdGw9NDMyMDA=...oookie=RX-a9...	200	document	sync.targeting.unrulymedia.com/csnc/RX...	336 B	41 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	18 ms	
SPug?partnerID=158497&gdpr=0&gdpr_consent=&us_privacy=	200	script	user_sync.html?kdntuid=1&p=158497:1	128 B	40 ms	
list?type=INDEX	304	xhr	xhr.js:177	123 B	17 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.0 kB	23 ms	
dc_oe=ChMIsNHvorq7_wlV4pdmAh2XCwV0EAAYACCox6dGQhMI...1;&timestamp=...	200	gif	express_html_inpage_rendering_lib_200_2...	108 B	97 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	23 ms	
PugMaster?sec=1&async=1&kdntuid=1&rnd=42620359&p=1...0&spug=1&coppa=0&...	200	script	user_sync.html?p=156011&s=165626&pr...	168 B	42 ms	
list?type=INDEX	304	xhr	xhr.js:177	125 B	19 ms	
ecm3?ex=pubmatic.com&id=PM_UID2271923A-F270-436E-BF37-15B167A758A2	200	document		479 B	272 ms	
list?type=INDEX	304	xhr	xhr.js:177	124 B	38 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.0 kB	26 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	18 ms	
list?type=INDEX	304	xhr	xhr.js:177	124 B	21 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	16 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	17 ms	
list?type=INDEX	304	xhr	xhr.js:177	124 B	22 ms	
list?type=INDEX	304	xhr	xhr.js:177	124 B	22 ms	
?ai=ChC2P1d2FZPrdMt-NssUPpuGv0A2S_af_cNSby8n5EcLlk...w_focus&gqid&qqid...	200	text/html	window_focus_fy2021.js:6	0 B	50 ms	
?ai=CGKJv1t2FZPTUE4yXoQPZ7quQD7CK7qJvrqf81ZUNZBAB...w_focus&gqid&q...	200	text/html	window_focus_fy2021.js:6	0 B	48 ms	
list?type=INDEX	304	xhr	xhr.js:177	126 B	16 ms	
list?type=INDEX	304	xhr	xhr.js:177	125 B	34 ms	
list?type=INDEX	304	xhr	xhr.js:177	125 B	18 ms	
dc_oe=ChMIsNHvorq7_wlV4pdmAh2XCwV0EAAYACCox6dGQhMI...1;&timestamp=...	200	gif	express_html_inpage_rendering_lib_200_2...	63 B	108 ms	
list?type=INDEX	304	xhr	xhr.js:177	125 B	23 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.0 kB	15 ms	
?ai=ChC2P1d2FZPrdMt-NssUPpuGv0A2S_af_cNSby8n5EcLlk...w_focus&gqid&qqid...	200	text/html	window_focus_fy2021.js:6	0 B	53 ms	
?ai=CGKJv1t2FZPTUE4yXoQPZ7quQD7CK7qJvrqf81ZUNZBAB...w_focus&gqid&q...	200	text/html	window_focus_fy2021.js:6	0 B	49 ms	
list?type=INDEX	304	xhr	xhr.js:177	124 B	21 ms	
list?type=INDEX	304	xhr	xhr.js:177	123 B	19 ms	
?ai=ChC2P1d2FZPrdMt-NssUPpuGv0A2S_af_cNSby8n5EcLlk...w_focus&gqid&qqid...	200	text/html	window_focus_fy2021.js:6	0 B	49 ms	
?ai=CGKJv1t2FZPTUE4yXoQPZ7quQD7CK7qJvrqf81ZUNZBAB...w_focus&gqid&q...	200	text/html	window_focus_fy2021.js:6	0 B	49 ms	

48 requests | 18.6 kB transferred | 77.7 kB resources

DevTools - www.settrade.com/th/home

Elements Console Sources Network **Performance** Memory Application Security Lighthouse Recorder ▾ Performance insights ▾ AngularJS

✖ 8 ⚠ 1 ⚡ 6 ⚙ ⚔

Filter  Invert  Hide data URLs All | Fetch/XHR JS CSS Img Media Font Doc WS Wasm Manifest Other  Has blocked cookies  Blocked Requests  3rd-party requests

20000 ms 40000 ms 60000 ms 80000 ms 100000 ms 120000 ms 140000 ms 160000 ms 180000 ms 200000 ms 220000 ms 240000 ms 260000 ms

Name	Status	Type	Initiator	Size	Time	Waterfall
list?type=INDEX	304	xhr	xhr.js:177	124 B	38 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.0 kB	26 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	18 ms	
list?type=INDEX	304	xhr	xhr.js:177	124 B	21 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	16 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	17 ms	
list?type=INDEX	304	xhr	xhr.js:177	124 B	22 ms	
list?type=INDEX	304	xhr	xhr.js:177	124 B	22 ms	
list?type=INDEX	304	xhr	xhr.js:177	126 B	16 ms	
list?type=INDEX	304	xhr	xhr.js:177	125 B	34 ms	
list?type=INDEX	304	xhr	xhr.js:177	125 B	18 ms	
list?type=INDEX	304	xhr	xhr.js:177	125 B	23 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.0 kB	15 ms	
list?type=INDEX	304	xhr	xhr.js:177	124 B	21 ms	
list?type=INDEX	304	xhr	xhr.js:177	123 B	19 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	23 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.0 kB	19 ms	
list?type=INDEX	304	xhr	xhr.js:177	123 B	20 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	27 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.0 kB	22 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	25 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	25 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	26 ms	
list?type=INDEX	304	xhr	xhr.js:177	125 B	26 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.0 kB	31 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	27 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.0 kB	27 ms	
list?type=INDEX	304	xhr	xhr.js:177	123 B	20 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	24 ms	
list?type=INDEX	200	xhr	xhr.js:177	(disk cache)	0 ms	
list?type=INDEX	304	xhr	xhr.js:177	124 B	20 ms	
list?type=INDEX	304	xhr	xhr.js:177	124 B	25 ms	
list?type=INDEX	304	xhr	xhr.js:177	125 B	17 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.0 kB	27 ms	
list?type=INDEX	304	xhr	xhr.js:177	124 B	27 ms	
list?type=INDEX	304	xhr	xhr.js:177	124 B	31 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	30 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.0 kB	20 ms	
list?type=INDEX	200	xhr	xhr.js:177	1.1 kB	19 ms	

45 / 79 requests | 27.8 kB / 35.5 kB transferred | 164 kB / 165 kB resources

DevTools - www.settrade.com/th/home

Elements Console Sources Network Performance Memory Application Security Lighthouse Recorder Performance insights AngularJS

Preserve log Disable cache No throttling Filter Invert Hide data URLs All Fetch/XHR JS CSS Img Media Font Doc WS Wasm Manifest Other Has blocked cookies Blocked Requests 3rd-party requests

20000 ms 40000 ms 60000 ms 80000 ms 100000 ms 120000 ms 140000 ms 160000 ms 180000 ms 200000 ms 220000 ms 240000 ms 260000 ms 280000 ms 300000 ms 320000 ms 340000 ms

Name Headers Payload Preview Response Initiator Timing Cookies

```

1   "indexIndustrySectors": [
      {
        "symbol": "SET",
        "nameEN": "SET",
        "nameTH": "SET",
        "prior": 1559.50000,
        "open": 1558.94000,
        "high": 1559.43000,
        "low": 1549.57000,
        "last": 1555.11000,
        "change": null,
        "percentChange": null,
        "volume": null,
        "value": null,
        "querySymbol": "SET",
        "marketStatus": "Closed",
        "marketDateTime": "2023-06-10T03:20:11.564264874+00:00",
        "marketName": "SET",
        "industryName": "",
        "sectorName": "",
        "level": "INDEX"
      },
      {
        "symbol": "SET50",
        "nameEN": "SET50",
        "nameTH": "SET50",
        "prior": 945.95000,
        "open": 944.80000,
        "high": 945.28000,
        "low": 938.26000,
        "last": 941.58000,
        "change": -4.37000,
        "percentChange": -0.46000,
        "volume": 1240610800,
        "value": 29122205677.81000,
        "querySymbol": "SET50",
        "marketStatus": "Closed",
        "marketDateTime": "2023-06-10T03:20:11.564264874+00:00",
        "marketName": "SET",
        "industryName": "",
        "sectorName": "",
        "level": "INDEX"
      },
      {
        "symbol": "SET100",
        "nameEN": "SET100",
        "nameTH": "SET100",
        "prior": 2104.43000,
        "open": 2102.36000,
        "high": 2103.88000,
        "low": 2088.55000,
        "last": 2096.69000,
        "change": -7.74000,
        "percentChange": -0.37000,
        "volume": 1822239600,
        "value": 34185571732.51000,
        "querySymbol": "SET100",
        "marketStatus": "Closed",
        "marketDateTime": "2023-06-10T03:20:11.564264874+07:00",
        "marketName": "SET",
        "industryName": "",
        "sectorName": ""
      }
    ]
  
```

The dashboard displays the following data:

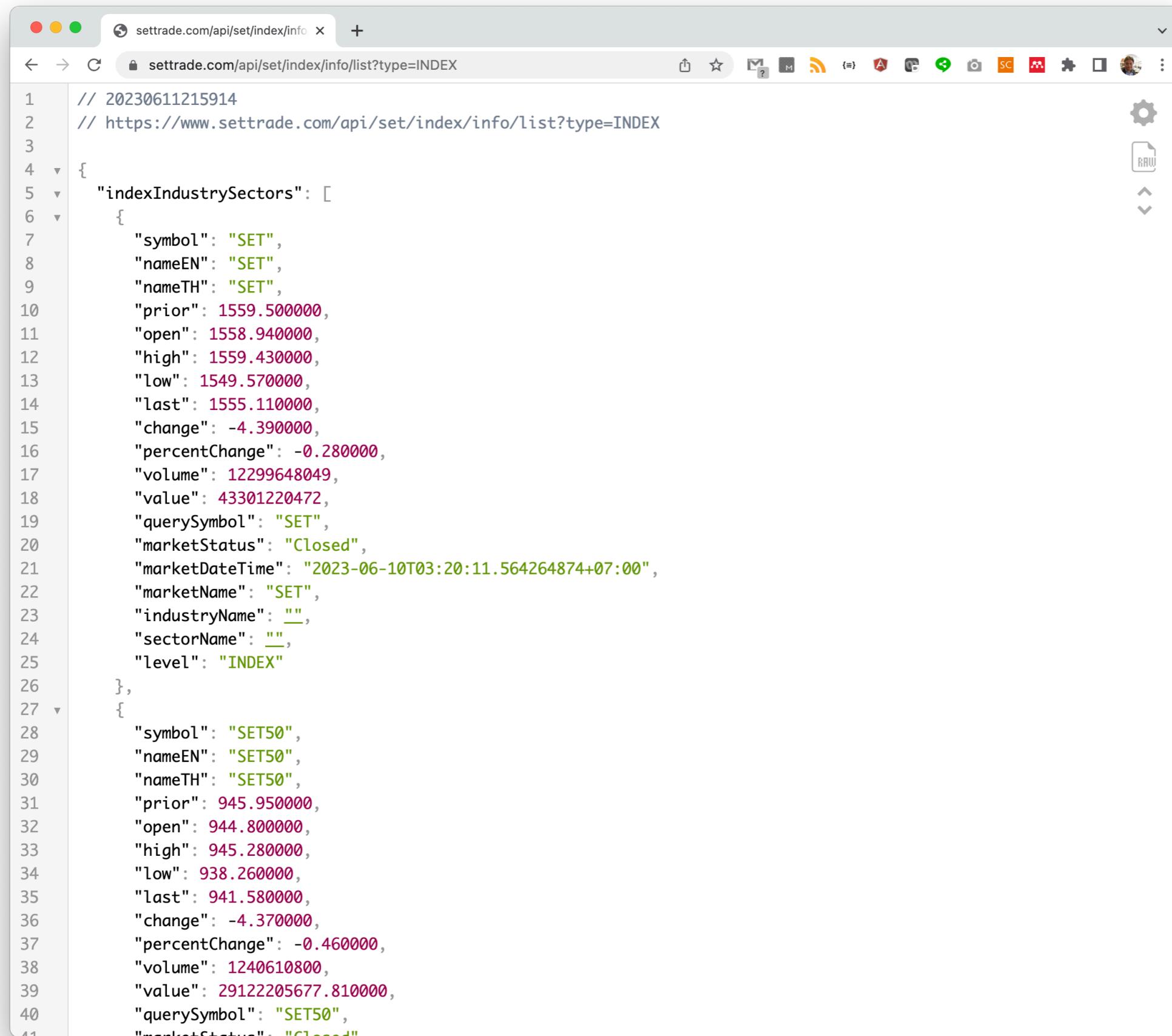
- SET**: Current price 1,555.11, down 4.39 (-0.28%)
- SET50**: Current price 941.58, down 4.37 (-0.46%)
- SET100**: Current price 2,096.69, down 7.74 (-0.37%)
- sSET**: Current price 1,001.14, up 2.23 (+0.22%)
- SETCLMV**: Current price 888.25, down 1.38 (-0.16%)

Other metrics shown:

- สุ่งสุด: 1,559.43
- ต่ำสุด: 1,549.57
- ปริมาณ ('000 หุ้น): 12,299,648
- มูลค่า (ล้านบาท): 43,301.22

Bottom status bar: 64 / 102 requests | 41.3 kB / 49.0 kB





The screenshot shows a web browser window displaying a JSON API response from [settrade.com/api/set/index/info/list?type=INDEX](https://www.settrade.com/api/set/index/info/list?type=INDEX). The response is a list of index industry sectors, with two entries shown in the main pane and more entries visible in the scrollable list.

```
// 20230611215914
// https://www.settrade.com/api/set/index/info/list?type=INDEX

{
  "indexIndustrySectors": [
    {
      "symbol": "SET",
      "nameEN": "SET",
      "nameTH": "SET",
      "prior": 1559.50000,
      "open": 1558.94000,
      "high": 1559.43000,
      "low": 1549.57000,
      "last": 1555.11000,
      "change": -4.39000,
      "percentChange": -0.28000,
      "volume": 12299648049,
      "value": 43301220472,
      "querySymbol": "SET",
      "marketStatus": "Closed",
      "marketDateTime": "2023-06-10T03:20:11.564264874+07:00",
      "marketName": "SET",
      "industryName": "",
      "sectorName": "",
      "level": "INDEX"
    },
    {
      "symbol": "SET50",
      "nameEN": "SET50",
      "nameTH": "SET50",
      "prior": 945.95000,
      "open": 944.80000,
      "high": 945.28000,
      "low": 938.26000,
      "last": 941.58000,
      "change": -4.37000,
      "percentChange": -0.46000,
      "volume": 1240610800,
      "value": 29122205677.81000,
      "querySymbol": "SET50",
      "marketStatus": "Closed"
    }
  ]
}
```

localhost

jupyter 3 - REST API Data Extraction (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Markdown

## REST API Data Extraction

Gathering data from a REST API is quite typical. Most Single-Page-Application (SPA) and AJAX dynamic pages rely on REST APIs. In addition, most vendor-specific APIs such as Facebook, Twitter, etc., base on REST.

The most important step of extracting data via REST API is to identify the endpoint.

```
In [ ]: import requests  
import json  
import pprint
```

## Call REST API

After we investigate the main page of settrade.com, we can figure out the endpoint of the market information using debugger in the browser.

```
In [ ]: api_url = 'http://api.settrade.com/api/market/SET/info'
```

```
In [ ]: data_info = requests.get(api_url)  
data_info.text
```

## Extract data

# Extract Complicated Webpage with Selenium

---

- Modern webpages are Javascript-based, which can be very difficult or impossible to use our previous techniques
  - Normal web applications usually base on session of multiple web pages using cookies and HTTP headers
  - Single Page Application (SPA) augments existing elements when being interacted and generates new elements on-the-fly
- Selenium is an open source framework that allow us to scrape data from all these web applications

# Selenium Architecture



# Selenium Capabilities

---

- Process any website similar to normal web browser
  - Access a web page using normal URL
  - Grab HTML and other resource files in a particular web page
  - Find components based on criteria e.g. CSS selector, etc.
  - Interacting with components in a web page e.g. click, input keyboard
  - Navigate within a web page and across multiple pages
- During operation, Selenium can be normal mode or headless mode
- Checkout <https://www.selenium.dev/documentation/> for more information

# Example: Selenium on Google

---

- Visit <https://www.google.com>
- Get HTML of this web page
- Find a query box
- Enter ‘ප්‍රසාද තුනුවේ’ in the query box and press ‘enter’
- Search result links ('a' elements under an element with id='search')
- Click the first link

The screenshot shows a Jupyter Notebook interface running on localhost. The title bar indicates the notebook is titled "jupyter 5 - Selenium (autosaved)". The top menu bar includes File, Edit, View, Insert, Cell, Kernel, Widgets, Help, Trusted, and Python 3 (ipykernel). Below the menu is a toolbar with various icons for file operations like save, new, cut, copy, paste, and run. The main content area features a large heading "Data Extraction with Selenium". A descriptive text follows, stating: "In this tutorial, we discuss how to use Selenium to extract data from the web. Please see <https://selenium-python.readthedocs.io> for more details." Below this, a section titled "Installation" provides instructions on how to install Selenium using pip. It lists two commands: "pip install chromedriver\_autoinstaller" and "pip install selenium". Two code cells are shown in the notebook:

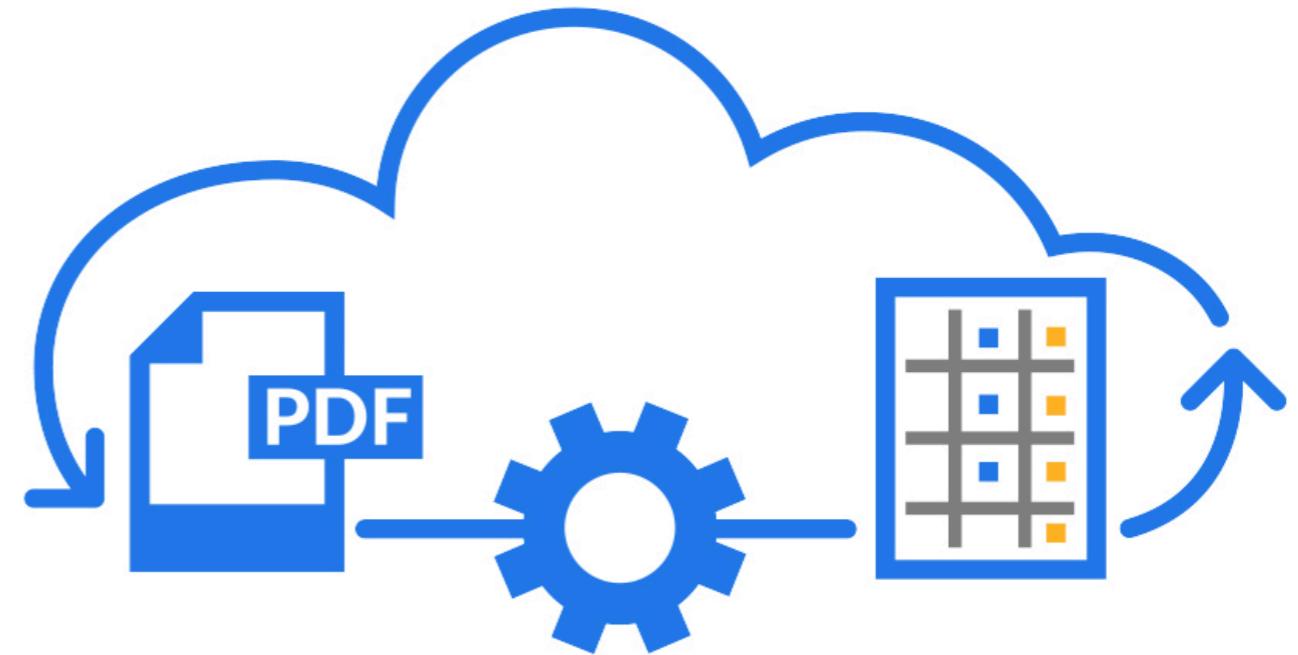
```
In [ ]: from selenium import webdriver  
import chromedriver_autoinstaller  
import time  
import os
```

```
In [ ]: chromedriver_autoinstaller.install()
```

# Document Extraction

---

How we can utilize data in PDF  
and Excel



# Excel Data Extraction

---

- Excel is one of the most popular and widely-used data tools; but Python is more powerful in many areas:
  - More powerful data importing and manipulation
  - Easier automation
  - Easier working with lots of data
  - Advanced statistics and machine learning capabilities
  - Advanced data visualization capabilities
- Pandas has excellent methods for reading all kinds of data from Excel files

# Example

- ข้อมูลภาวะเศรษฐกิจสังคมครัวเรือนและแรงงานเกษตร in open data thailand
- <https://data.go.th/dataset/dataoae2104>
- Excel file has been included in the shared file:  
agri\_landuse.xlsx

## ข้อมูลภาวะเศรษฐกิจสังคมครัวเรือนและแรงงานเกษตร

องค์กร : สำนักงานเศรษฐกิจการเกษตร

เผยแพร่ข้อมูลจากการสำรวจประจำปี ครอบคลุมด้วย การใช้กําลัง,รายได้,รายจ่าย,หนี้สิน,กรัพย์สิน,ข้อมูลพื้นฐานของ

	ดาวน์โหลด	การใช้กําลัง  498 downloads
	ดาวน์โหลด	ข้อมูลพื้นฐานครัวเรือน  443 downloads
	ดาวน์โหลด	กรัพย์สิน  212 downloads
	ดาวน์โหลด	รายได้รายจ่าย  408 downloads
	ดาวน์โหลด	หนี้สิน  272 downloads

Data source cannot be displayed

ข้อมูลทรัพยากร : การใช้

วันที่สร้างไฟล์/ลิงก์ข้อมูลใน catalog

รูปแบบของไฟล์ชุดข้อมูล

รหัสกรัพยากรข้อมูล

รายละเอียด

เงื่อนไขในการเข้าถึงข้อมูล

ปีที่เริ่มต้นจัดทำเนื้อหาข้อมูล

ปัลสุดที่จัดทำเนื้อหาข้อมูล

การจัดจำแนก

localhost

jupyter 6\_excel\_extraction (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel) ○

Excel Data Extraction

Pandas provides a `read_excel` method to read an excel file and convert it to dataframe automatically. It includes several parameters to customize how we read excel e.g. select sheet, rows to be used as column name, skipping rows, etc.

First we import pandas and name it `pd`

In [1]: `import pandas as pd`

Read an excel file without supplying other parameters. Notice that the column name is incorrect and the first row contains all column names

In [2]: `df = pd.read_excel('agri_landuse.xlsx')`

In [4]: `df`

Out[4]:

	การถือครอง และการใช้ประโยชน์ที่ดิน ปี	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4	Unnamed: 5
0	รายการ	ประเภท	ภาคเหนือ	ภาคตะว.ออก เฉียงเหนือ	ภาคกลาง	ภาคใต้
1	เนื้อที่รวมทั้งหมดของครัวเรือน (ไร่/ครัว เรือน)	24.15531	25.022066	21.901518	27.194832	19.729259

# Reading PDF

---

- PDF is the most popular document format; we can perform several operations on PDF using PyPDF2 e.g.
  - Extract Metadata
  - Extract Text from a PDF
  - Extract Images
  - Merging PDF files
  - Transforming PDFs
  - Interactions with PDF Forms
- Install PyPDF2 using:

```
pip install PyPDF2
```

# Example

- รายงานสถานการณ์การจำหน่ายไฟฟ้า ปี 2563 in open data thailand
- <https://data.go.th/dataset/sell-2563>
- PDF file has been included in the shared file: ee\_status.pdf

## รายงานสถานการณ์การจำหน่ายไฟฟ้า ปี 2563

องค์กร : การไฟฟ้านครหลวง



ดาวน์โหลดกันหมด

กพน. รายงานสถานการณ์การจำหน่ายไฟฟ้า ปี 2563

ดาวน์โหลด	รายงานสถานการณ์การจำหน่ายไฟฟ้า เดือนมกราคม 2563 <span style="color: orange;">↗ 28</span> <a href="#">downloads</a>
ดาวน์โหลด	รายงานสถานการณ์การจำหน่ายไฟฟ้า-เดือนกุมภาพันธ์-2563 <span style="color: green;">↗ 8</span> <a href="#">downloads</a>
ดาวน์โหลด	รายงานสถานการณ์การจำหน่ายไฟฟ้า-เดือนมีนาคม-2563 <span style="color: green;">↗ 3</span> <a href="#">downloads</a>
ดาวน์โหลด	รายงานสถานการณ์การจำหน่ายไฟฟ้า-เดือนเมษายน-2563 <span style="color: green;">↗ 6</span> <a href="#">downloads</a>
ดาวน์โหลด	รายงานสถานการณ์การจำหน่ายไฟฟ้า-เดือนพฤษภาคม-2563 <span style="color: green;">↗ 4</span> <a href="#">downloads</a>
ดาวน์โหลด	รายงานสถานการณ์การจำหน่ายไฟฟ้า-เดือนมิถุนายน-2563 <span style="color: green;">↗ 5</span> <a href="#">downloads</a>
ดาวน์โหลด	รายงานสถานการณ์การจำหน่ายไฟฟ้า-เดือนกรกฎาคม-2563 <span style="color: green;">↗ 2</span> <a href="#">downloads</a>

Data source cannot be di...

ข้อมูลกรัพยากร : ราย...

วันที่ปรับปรุงไฟล์/ลิงก์ข้อมูล catalog

วันที่สร้างไฟล์/ลิงก์ข้อมูลใน catalog

รูปแบบของไฟล์ชุดข้อมูล

รหัสกรัพยากรข้อมูล

รายละเอียด

The screenshot shows a Jupyter Notebook interface running on localhost. The title bar indicates the notebook is titled "jupyter 7\_pdf\_extraction (autosaved)". The toolbar includes standard file operations (File, Edit, View, Insert, Cell, Kernel, Widgets, Help), a status bar ("Not Trusted", "Python 3 (ipykernel)"), and a toolbar with icons for saving, running, and cell selection.

## PDF Extraction

PyPDF2 is an excellent PDF manipulating package in python. It provides several capabilities include:

- Read a PDF file
- Extract text and other information
- Manipulate PDF file e.g. rotate, merge, split, etc.

We must first install the package:

```
pip install PyPDF2
```

Once installed, we can import PyPDF2 into our program.

```
In [ ]: from PyPDF2 import PdfReader
```

Read a sample file, 'ee\_status.pdf', then count number of pages, and print out metadata

```
In [ ]: reader = PdfReader('ee_status.pdf')
```

```
In [ ]: len(reader.pages)
```

```
In [ ]: meta = reader.metadata
```

# Recommended Resources

---

- <http://python.gotrained.com/python-json-api-tutorial/>
- <https://www.analyticsvidhya.com/blog/2015/10/beginner-guide-web-scraping-beautiful-soup-python/>
- <https://www.dataquest.io/blog/web-scraping-tutorial-python/>
- <https://www.datacamp.com/courses/importing-data-in-python-part-2>
- <https://code.tutsplus.com/tutorials/the-30-css-selectors-you-must-memorize--net-16048>

# Assignment

---

- Extract information about “วันพระ” ในระยะเวลา 3 ปีจาก
  - <https://www.myhora.com/ปฏิทิน/วันพระ-พ.ศ.2565.aspx>
  - <https://www.myhora.com/ปฏิทิน/วันพระ-พ.ศ.2566.aspx>
  - <https://www.myhora.com/ปฏิทิน/วันพระ-พ.ศ.2567.aspx>
- Count the distribution of number of week days that are “วันพระ” for all three years and answer the following questions
  - How many วันพระ in total (of 3 years)?
  - How many days in total (of 3 years) that วันพระ is Monday?
  - Which day of the week that has the minimum number of วันพระ? Which has the maximum?
- Note that you can use dateparse package to parse Thai date. There is an example shown in the assignment template