

# Intro to Web Scrapping

@vicohyeh | vicyeh.me

# Overview

- What is Data?
- Web Scraping
- Python Beautiful Soup

# What is Data?

“Information used to calculate, analyze or plan something” ~ Merriam-Webster



# Data

- ex. survey results, purchase history

#▲	Purchase Code ↕	Supplier ↕	Product ↕	Quantity ↕	Unit Price ↕	Payment ↕	Method ↕	Date ↕
1	bb6b7e1b3d46bfe	<a href="#">Chantale Stone ↗</a>	<a href="#">Iphone 6 ↗</a>	5	500	2500	cash	29-01-2015
2	4df9d5ab3a97733	<a href="#">Colette Gill ↗</a>	<a href="#">Iphone 6 plus ↗</a>	2	700	1400	cheque	21-01-2015
3	5699fc129839967	<a href="#">Chantale Stone ↗</a>	<a href="#">Samsung Galaxy Note 4 ↗</a>	3	300	900	cash	29-01-2015
4	4f89260ee9d8854	<a href="#">Colette Gill ↗</a>	<a href="#">Casio CTK ↗</a>	3	1200	3600	cheque	27-01-2015
5	f8c487618a53f05	<a href="#">Keely Rowland ↗</a>	<a href="#">Macbook Pro ↗</a>	3	1100	3300	cash	29-01-2015
6	2b0f206835ebbb6	<a href="#">Keely Rowland ↗</a>	<a href="#">Ibanez Guitar ↗</a>	5	2000	10000	cheque	29-01-2015
7	0345b62c8fa32c8	<a href="#">Joana Ethen ↗</a>	<a href="#">Rayban Clubmaster ↗</a>	20	500	10000	card	29-01-2015
8	5ef83d75fe0db97	<a href="#">Joana Ethen ↗</a>	<a href="#">Casio CTK ↗</a>	3	500	1500	cheque	29-01-2015
9	02bae24b0b7ed98	<a href="#">Joana Ethen ↗</a>	<a href="#">Samsung Galaxy Note 4 ↗</a>	4	300	1200	cheque	13-01-2015
10	6dee340a05a1145	<a href="#">Joana Ethen ↗</a>	<a href="#">Yamaha Electric Guitar ↗</a>	4	500	2000	cash	29-01-2015

Showing 1 to 10 of 10 entries

Previous 1 Next

# Raw Data

- Data that have not been collected but not formatted/analyzed.

# Data format

## Comma-separated values (CSV)

```
1 John,3108298585,john@gmail.com
2 Melissa,2930592839,mh2910@gmail.com
3 Roger,2930309902,yourroger@hotmail.com
4 Nathan,5039201102,ilovecoco@yahoo.com
5 Calvin,3102940580,cooldude@yahoo.com
6 Lisa,3102949492,lisadavis@gmail.com
7 Smith,2302901928,ssman29@gmail.com
```

## JSON

```
{ "users": [
  {
    "firstName": "Ray",
    "lastName": "Villalobos",
    "joined": {
      "month": "January",
      "day": 12,
      "year": 2012
    }
  },
  {
    "firstName": "John",
    "lastName": "Jones",
    "joined": {
      "month": "April",
      "day": 28,
      "year": 2010
    }
  }
] }
```

# Web scraping?

Pulling data straight out of HTML



**"HACKERS"**

# Rule of thumb

**Any content that can be viewed  
on a webpage can be scraped.**



# Pros vs Cons

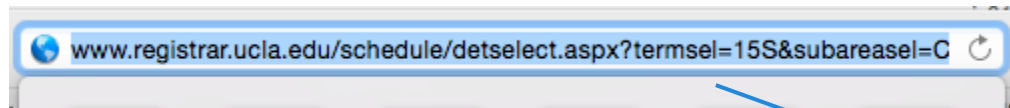
## Pros

- No need to read lengthy API Docs
- Usually able to retrieve more data than an official API provides
- Make you sound smart

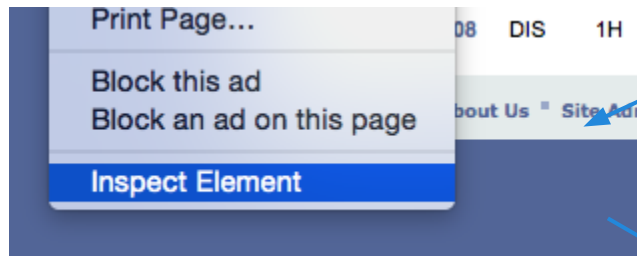
## Cons

- Less intuitive
- If HTML changes, web scraper has to change

# Strategy



Course Webpage	Library	Reserves	Textbooks												
ID Number	Type	Sec	Days	Start	Stop	Building	Room	Res't	#En	EnCp	#WL	WLCp	Statu:		
Crs Info	LEC	1	TR	2:00P	3:50P	WGYOUNG	CS50	Yes	242	280	0	0	Open		
187101201	DIS	1A	F	10:00A	11:50A	PUB AFF	1337	Yes	30	35	0	0	Open		
187101202	DIS	1B	F	10:00A	11:50A	BOELTER	3400	Yes	27	35	0	0	Open		
187101203	DIS	1C	F	10:00A	11:50A	HUMANTS	135	Yes	28	35	0	0	Open		
187101204	DIS	1D	F	10:00A	11:50A	PAB	1749	Yes	33	35	0	0	Open		



```
<td class="dgdClassDataColumnSpacer"> </td>
<td class="dgdClassDataEnrollTotal">
  <span id="ctl00_BodyContentPlaceHolder_detselect_ctl02_ctl02_EnrollTotal">
    <span class="bold">242</span>
  </span>
</td>
<td class="dgdClassDataColumnSpacer"> </td>
<td class="dgdClassDataEnrollCap">
```

# Python Beautiful Soup

Python library for pulling data out of HTML and XML

- Support popular HTML parsers
- Easy to navigate, search, and modify the parse tree

# Installation

## Linux

```
apt-get install python-bs4
```

```
apt-get install python-lxml
```

## Mac

```
pip install beautifulsoup4
```

```
pip install lxml
```

# Hack Time!



# Applications

ex. Class Scanner

If you've found ClassScanner helpful, please consider making a donation to help us pay for our server costs. Thanks!

[Donate](#) [Donate Bitcoin](#) [Donate Dogecoin](#) [Dropbox](#) [Uber](#) [Lyft](#)

Dashboard (503) 819-6919

[Dashboard](#) [+ Add New Class](#) [URSA](#) [UCLA Registrar](#) [Bruinwalk.com](#)

CS 174A PHILOS 6

Section	Days	Start	End	Enrollment	Waitlist	Status	Action
DIS 1A	F	2:00P	3:50P	35/40	0/0	Open	⋮
DIS 1B	F	4:00P	5:50P	36/40	0/0	Open	⋮
DIS 1C	F	4:00P	5:50P	32/40	0/0	Open	⋮

[Show All Sections](#)

### Computer Science 174A

Course Description	Lecture, four hours; discussion, two hours; outside study, six hours. Enforced requisite: course 32. Basic principles behind modern two- and three-dimensional computer graphics systems, including complete set of steps that modern graphics pipelines use to create realistic images in real time. How to position and manipulate objects in scene using geometric and camera transformations. How to create final image using perspective and orthographic transformations. Basics of modeling primitives such as polygonal models and implicit and parametric surfaces. Basic ideas behind color spaces, illumination models, shading, and texture mapping. Letter grading.
Units	4.0
Grading Detail	Letter grade Only

# Further Reading

- Beautiful Soup Documentation

<http://www.crummy.com/software/BeautifulSoup/bs4/doc/>

- “*Gaming the UCLA Enrollment Process*” by Shahid Chohan

<http://shahidchohan.com/gaming-the-ucla-enrollment-process/>

# Questions?

Contact me at [vic.yeh@ucla.edu](mailto:vic.yeh@ucla.edu)

Or tweet me @vicohyeh :)