

# **Data Collecting and Pre-processing**

## **Economics with Data Science**

**J.H 10th, March, 2024**

[bristol.ac.uk](https://bristol.ac.uk)

# Content

- 1 Overview
  - Data structure
  - Data Pre-processing
- 3 Webpage Crawl
  - API scraping demo (UK job)
  - HTML scraping demo (Amazon)
- 4 Further Discussion

Data Collecting and Pre-processing

[bristol.ac.uk](http://bristol.ac.uk)

# 1 Overview

## Data Structure

- Different types of data: Structure, semi-structure, unstructured data

**Structure  
Data**

**Semi-structure  
Data**

**Unstructured  
Data**

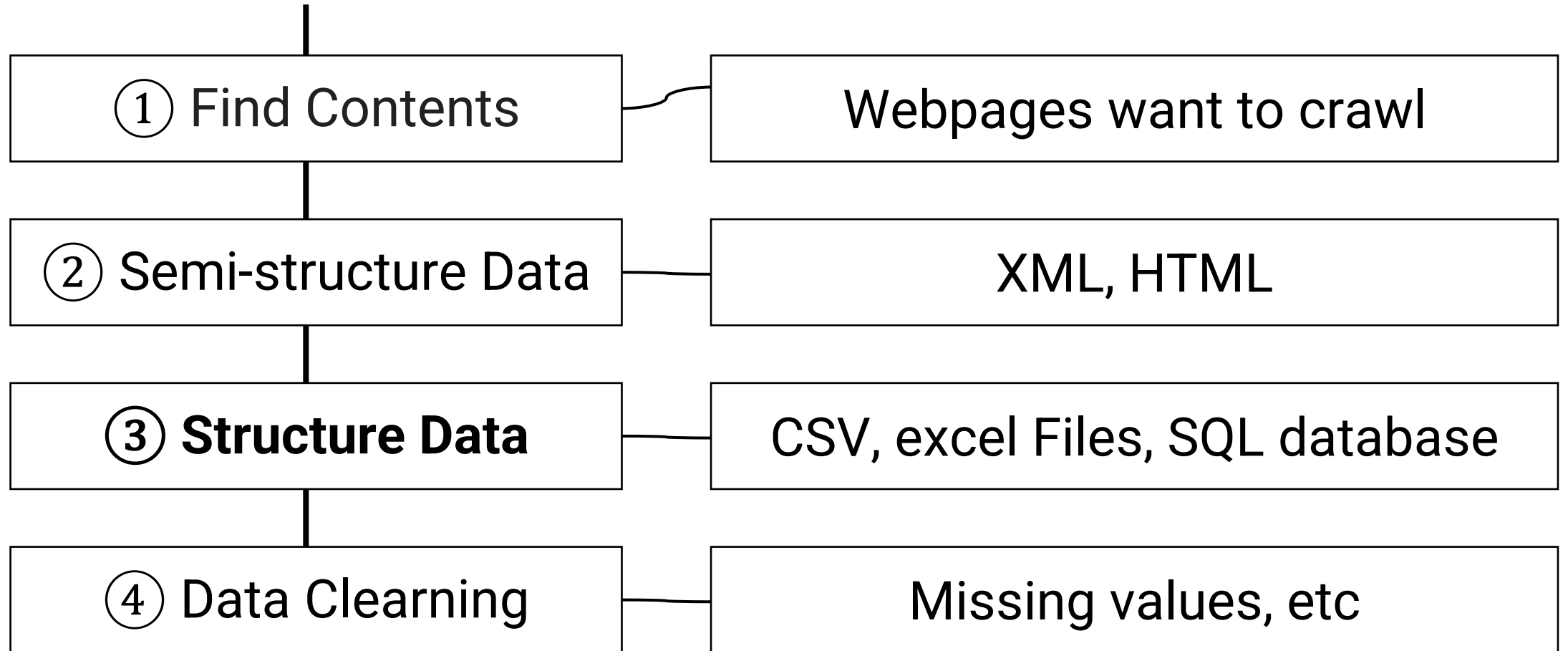
CSV, Excel Files,  
SQL database

XML, HTML  
files

Graph, Videos,  
No-SQL

# 1 Overview

## Data Pre-processing Steps



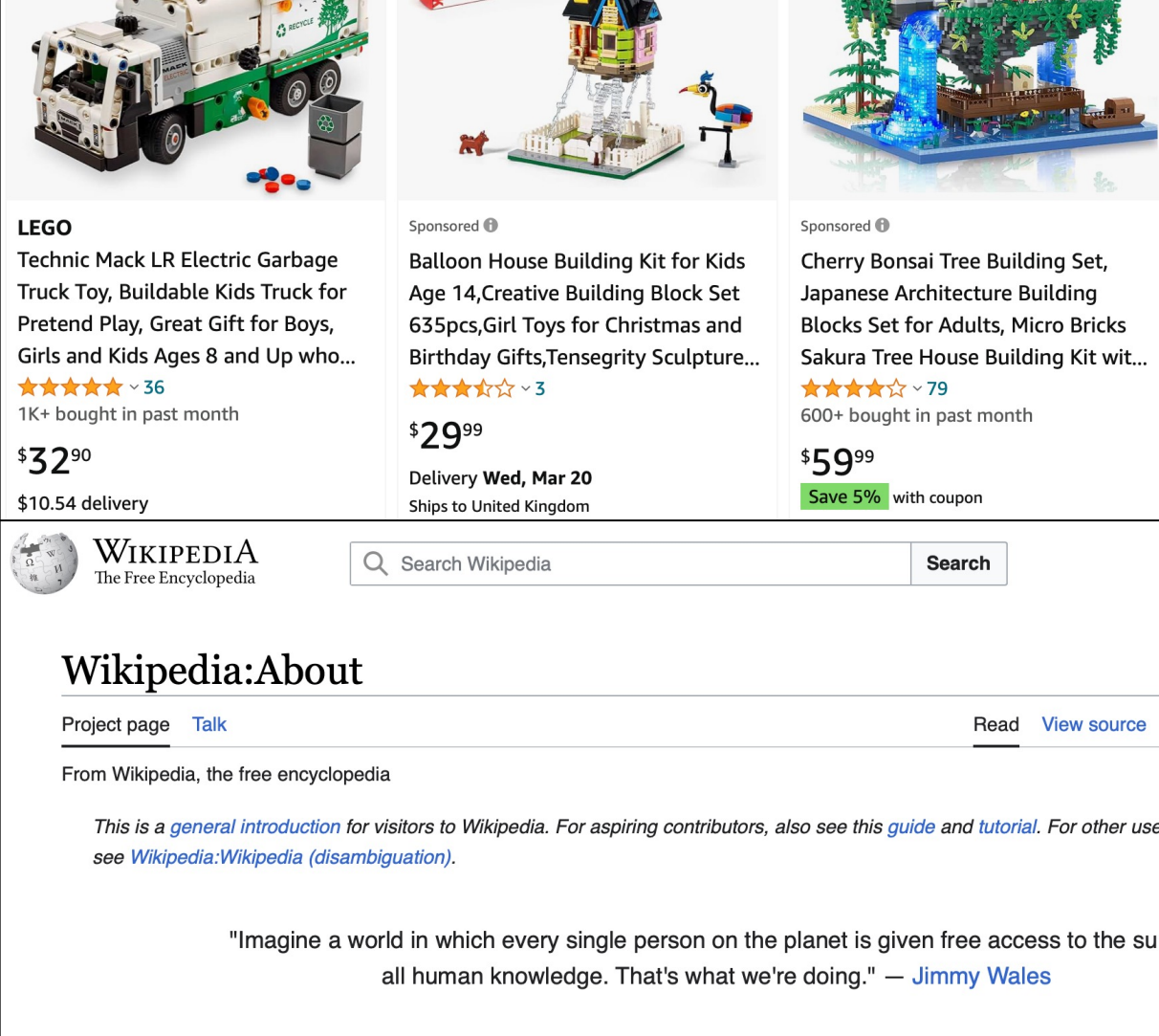
# 2 Webpage Crawl

## 2.1 Find Contents

Find and list all websites where the desired content can be obtained

First, check if these websites provide **API** services (**recommended**).  
If not, **crawl the HTML**.

bristol.ac.uk




The screenshot displays two distinct web pages. The top portion shows three Amazon product listings for LEGO sets. The first listing is for a 'LEGO Technic Mack LR Electric Garbage Truck Toy', priced at \$32.90 with a \$10.54 delivery fee. The second is a 'Sponsored' listing for a 'Balloon House Building Kit for Kids', priced at \$29.99. The third is another 'Sponsored' listing for a 'Cherry Bonsai Tree Building Set', priced at \$59.99 with a 5% discount coupon. The bottom portion of the screenshot shows the Wikipedia homepage, specifically the 'Wikipedia:About' page. It features the Wikipedia logo, a search bar, and introductory text about the encyclopedia, including a quote from Jimmy Wales: "Imagine a world in which every single person on the planet is given free access to the sum of all human knowledge. That's what we're doing."

**LEGO**  
Technic Mack LR Electric Garbage Truck Toy, Buildable Kids Truck for Pretend Play, Great Gift for Boys, Girls and Kids Ages 8 and Up who...  
★★★★★ ~ 36  
1K+ bought in past month  
\$32<sup>90</sup>  
\$10.54 delivery

Sponsored ⓘ  
Balloon House Building Kit for Kids Age 14, Creative Building Block Set 635pcs, Girl Toys for Christmas and Birthday Gifts, Tensegrity Sculpture...  
★★★★☆ ~ 3  
\$29<sup>99</sup>  
Delivery **Wed, Mar 20**  
Ships to United Kingdom

Sponsored ⓘ  
Cherry Bonsai Tree Building Set, Japanese Architecture Building Blocks Set for Adults, Micro Bricks Sakura Tree House Building Kit wit...  
★★★★☆ ~ 79  
600+ bought in past month  
\$59<sup>99</sup>  
Save 5% with coupon

 **WIKIPEDIA**  
The Free Encyclopedia

### Wikipedia:About

[Project page](#) [Talk](#) [Read](#) [View source](#)

From Wikipedia, the free encyclopedia

*This is a [general introduction](#) for visitors to Wikipedia. For aspiring contributors, also see this [guide](#) and [tutorial](#). For other use see [Wikipedia:Wikipedia \(disambiguation\)](#).*

"Imagine a world in which every single person on the planet is given free access to the sum of all human knowledge. That's what we're doing." — [Jimmy Wales](#)

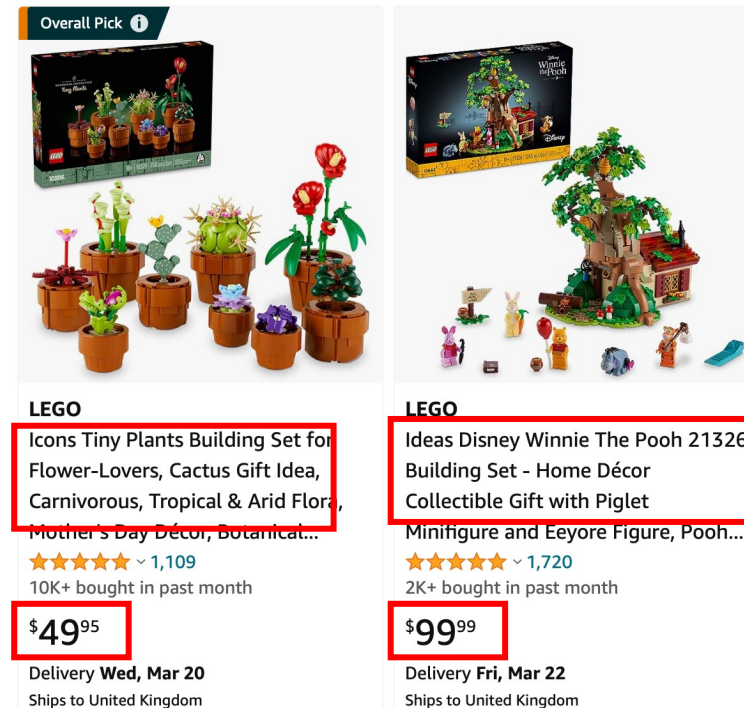
# 2 Webpage Crawl

## 2.2 Get Semi-structure Data

### Method 1: API

- The official data extraction service.
- Directly queries content to retrieve desired information

### Method 2: Webpage Crawl



Overall Pick 1

**LEGO**  
Icons Tiny Plants Building Set for Flower-Lovers, Cactus Gift Idea, Carnivorous, Tropical & Arid Flora, Mother's Day Décor, Botanical...

★★★★★ ~ 1,109  
10K+ bought in past month

**\$49<sup>95</sup>**

Delivery **Wed, Mar 20**  
Ships to United Kingdom

**LEGO**  
Ideas Disney Winnie The Pooh 21326 Building Set - Home Décor Collectible Gift with Piglet Minifigure and Eeyore Figure, Pooh...

★★★★★ ~ 1,720  
2K+ bought in past month

**\$99<sup>99</sup>**

Delivery **Fri, Mar 22**  
Ships to United Kingdom

# 2.2 Get Semi-structure Data

## Method 1: API Introduction

### What is an API?

A web-based Application Programming Interface (API).  
A contract between a server and a user.

### Usages

User: send a specific request to website server.  
Server: return requested information.

### Features

Simple and easy to use, but sometimes not free when exceeded limits/ for business use/ etc.

# API Demo

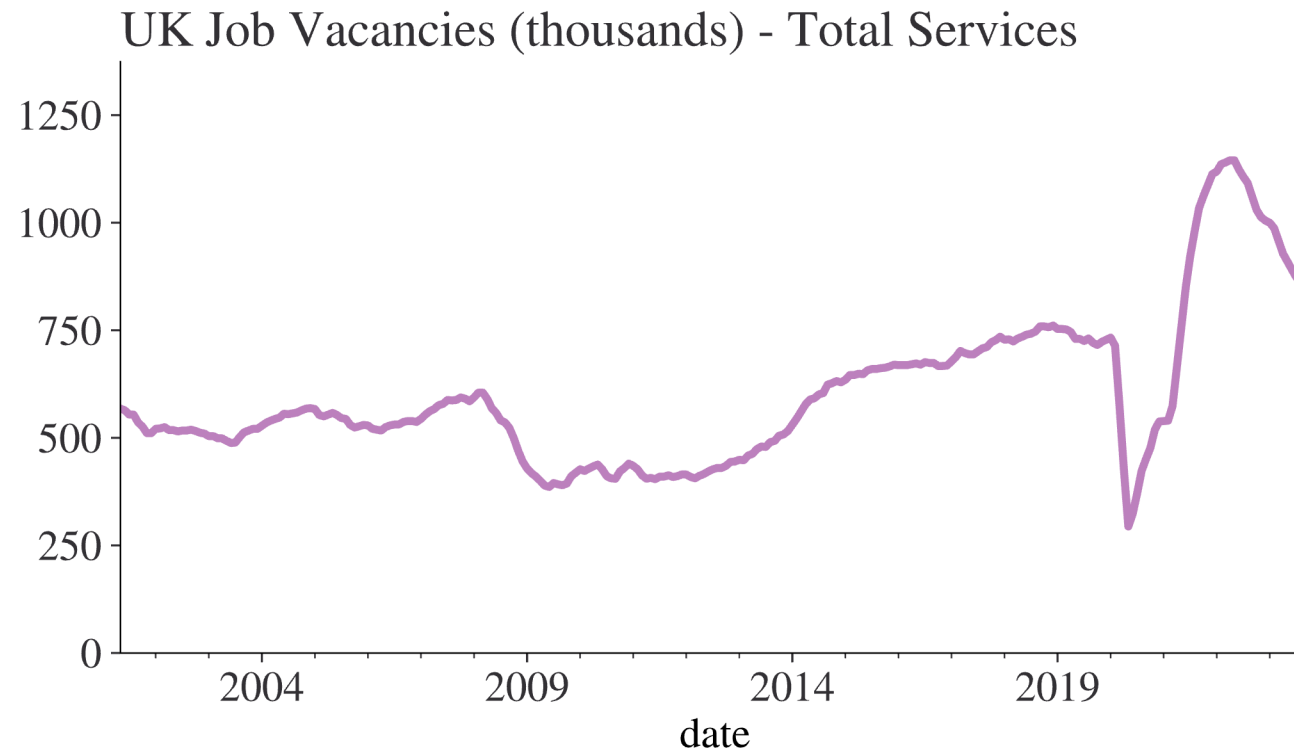
## Extract Data from UK public sector

```
# Get the data from the ONS API:  
import requests  
url = "https://api.ons.gov.uk/timeseries/JP9Z/dataset/UNEM/data"  
json_data = requests.get(url).json()
```

UK job vaccinations  
See & try to run full  
codes:

[Ref. tutorial link](#)

bristol.ac.uk





# API Demo

## Other Useful APIs

- Economics: APIs from [Berkeley Library](#)
- List of public APIs [Repository](#)
- NASDAQ Data Link <https://docs.data.nasdaq.com>
- Publicly-available economic data provided by [DBnomics](#)

## 2.2 Get Semi-structure Data

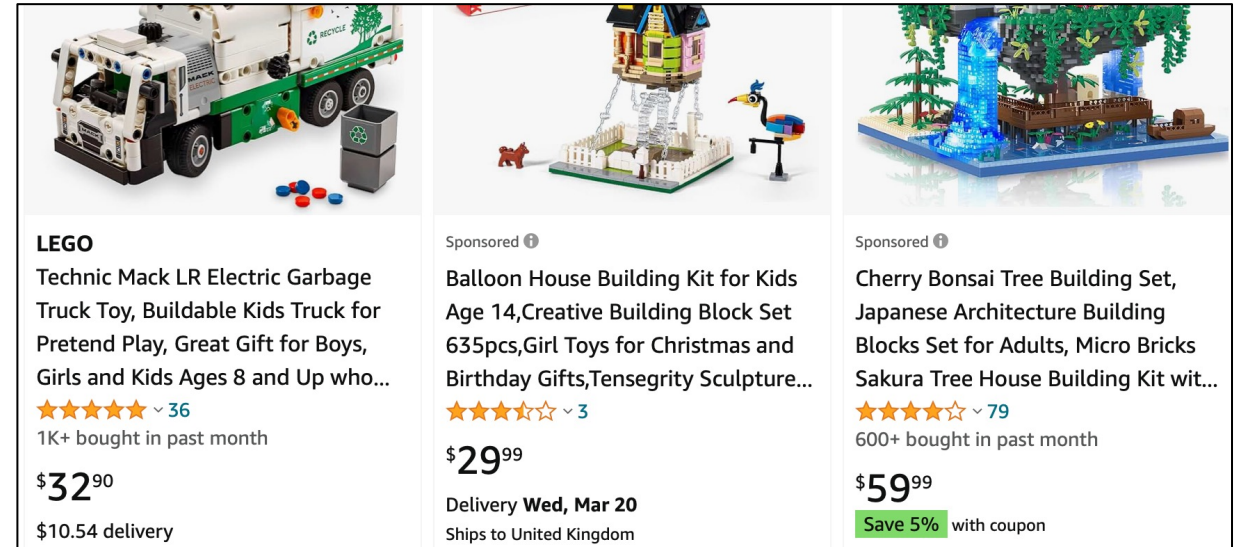
### Method 2: HTML Webpage Crawl

If there's no API service,  
we can do HTML-based scraping.

# HTML Webpage Crawl

## Amazon Demo

First, summarize the patterns/rules of the website to be crawled



<https://www.amazon.co.uk/s?k=lego>



<https://www.amazon.co.uk/s?k=keywords>

# HTML Webpage Crawl

## Amazon Demo: Extract contents

url="https://www.amazon.co.uk/s?k=manga"  
results = requests.get(url)

```
n:window,o=0,n="addEventListener",f="ABCDEFGHIJKLMNOPQRSTUVWXYZabcdefghijklmnopqrstuvwxyz0123456789-._~:;'/<br>[[],i=t.rx||{,r=i.c||{,e=r.rxp||"/rd/uedata",c=r.fi||5e3,a={},d={},w=  
[],v=0,x=0;x<64;x++)u[f[x]]=x;function h(n,i){return function(){try{return  
n.apply(this,arguments)}catch(n){l(n.message||n,n)}}}function l(n,i){n=(""+  
(n||"")).substring(0,100),w.push(o),w.push(n.length);for(var  
t=0;t<n.length;t++)w.push(n.charCodeAt(t));if(r.DEBUG)throw i||n;p()}function s(n,i){i=h(i),n in  
(d[n]=[]),d[n].push(i),n in a&&i()}function y(n,i){n in a||(a[n]=i,(d[n]||[]).forEach(function(n)  
{n(i)}))}function g(n){for(var  
i=0,t=0,o="",r=0;r<n.length;r+=1)for(t+=8,i=i<<8|n[r];6<=t;)o+=f[i>>t-6],i&=255>>8-(t-=6);return  
o<t&&(o+=f[i<<6-t]),o}function m(n){for(var i=0,t=0,o=  
[],r=0;r<n.length&&"!="n[r];r+=1)for(t+=6,i=i<<6|u[n[r]];8<=t;)o.push(i>>t-8),i&=255>>8-(t-  
=8);return new Uint8Array(o)}function p(){v||(setTimeout(h(A),c),v=1)}function A()  
{v=0,rx.ep&&0<w.length&&rx.ep(w,U),w=[]}function U(n){n=g(new Uint8Array(n));n=e+"?  
rid="+rx.rid+"&sid="+rx.sid+"&rx="+n;(new Image).src=n}function b(n){y("load",n)}function E(n)  
{b(n),y("unload",n),A()}(t.rx=i).err=l,i.r=h(s),i.e=h(y),i.exec=h,i.p=h(function(n,i)  
{w.push(255&n),w=w.concat(i),p()}),i.e64=h(function(i,n){s(n||"init",function(){var n;t.RXVM&&  
(n=m(i),t.$RX||(t.$RX=new t.RXVM),$RX.execute(n,t)})),i.e64=h(g),i.d64=h(m),y("init",{,n in t  
(t[n]("load",h(b)),t[n]("beforeunload",h(E)),t[n]("pagehide",h(E)))}(window);  
rx.ex64("UlgBKTUnV10vcExLUR1kV1dEXCNJQEtCUU0hUU1ASy9KS0ZKSfVJQFFALUZESUlhREZOJlZAUSNWUEdRSUAIqEtG  
rx.ex64("UlgBKSAhQUpLQCBTRElQQCBDSUpKVYdXXSFAXUBGJCQVKSFoRFFNBSVkJ5YleVNTUGJeVD43PDUm0nJTU1FDUlnW
```

# HTML Webpage Crawl

## Amazon Demo: Parsing

Convert extracted semi-structure contents to structured data

One Piece (3-in-1 Edition) Volume 1: Inc  
USD 83.51  
Spy x Family Vol 1: Volume 1  
USD 42.58  
I Want to Eat Your Pancreas (Manga): The  
USD 15.02  
Wotakoi: Love Is Hard for Otaku Complete  
USD 9.25  
Assassination Classroom Complete Box Set  
USD 5.13  
Berserk Deluxe Volume 1  
USD 20.56  
Tokyo Ghoul: re Complete Box Set: Includ  
USD 16.44

Jujutsu Kaisen Vol 1: Volume 1  
USD 9.39  
Chainsaw Man Box Set: Includes Volumes 1  
USD 5.13  
Demon Slayer Complete Box Set: Includes  
USD 62.87  
Exzact Chopsticks Gift Set – 5 Pairs of  
USD 134.91  
Jujutsu Kaisen 0  
USD 6.41  
Tokyo Ghoul Complete Box Set: Includes v  
USD 8.85  
Death Note (All-in-One Edition)  
USD 5.13

There are other useful tools for web scraping,  
feel free to explore.

Scrapy  
BeautifulSoup  
Selenium

.....

# 3 Further Discussion

## Access Denied Error

Frequent visits, authorizations, etc., can trigger **anti-scraping** mechanisms.

```
<title>503 - Service Unavailable Error</title>
</head>
<body bgcolor="#FFFFFF" text="#000000">
<center>
<a href="https://www.amazon.co.uk/ref=cs_503_logo/">
</a>
<p align="center">
<font face="Verdana,Arial,Helvetica">
<font color="#CC6600" size="+2"><b>We're sorry</b></font><br/>
<b>An error occurred when we tried to process your request.<br/>We're working on the problem and e
Please note that if you were trying to place an order, it will not have been processed at this tim
<br/><br/>We apologise for the inconvenience.</b><p>
```

# 3 Further Discussion

## Access Denied Error

### (1) Switch to official API service

To discuss automated access to Amazon data please contact [api-services-support@amazon.com](mailto:api-services-support@amazon.com).

For information about migrating to our APIs refer to our Marketplace APIs at [https://developer.amazonservices.co.uk/ref=rm\\_5\\_sv](https://developer.amazonservices.co.uk/ref=rm_5_sv), or our Product Advertising API at [https://affiliate-program.amazon.co.uk/gp/advertising/api/detail/main.html/ref=rm\\_5\\_ac](https://affiliate-program.amazon.co.uk/gp/advertising/api/detail/main.html/ref=rm_5_ac) for advertising use cases.

### (2) Slow down request frequency and try again.

### (3) Use agent



# References

- Software Development: Programming and Algorithms
- Structured, Semi-structured and Unstructured data:  
<https://www.tutorialspoint.com/difference-between-structured-semi-structured-and-unstructured-data>
- Data scraping tutorials: <https://aeturrell.github.io/coding-for-economists/data-extraction.html>

# Thank you

[bristol.ac.uk](http://bristol.ac.uk)

