Introduction to Data Scraping

Brought to you by:

smartcademy

Singapore’s leading training provider for in-demand tech skills & career transformation

Courses We Offer

1. Data Analytics
2. Digital Marketing
3. User Experience (UX) Design
4. Web/Mobile App Development

Companies that **hire** our graduates

Impacted

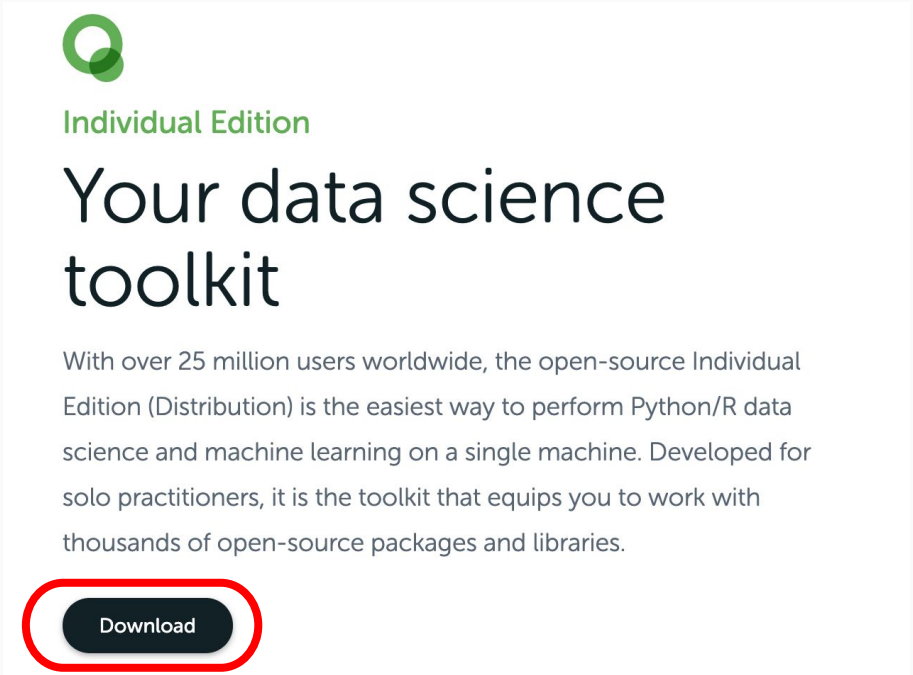
> 6000+

Introduction

Agenda

1. Introduction to AI and Data Scraping
2. Hands-on!

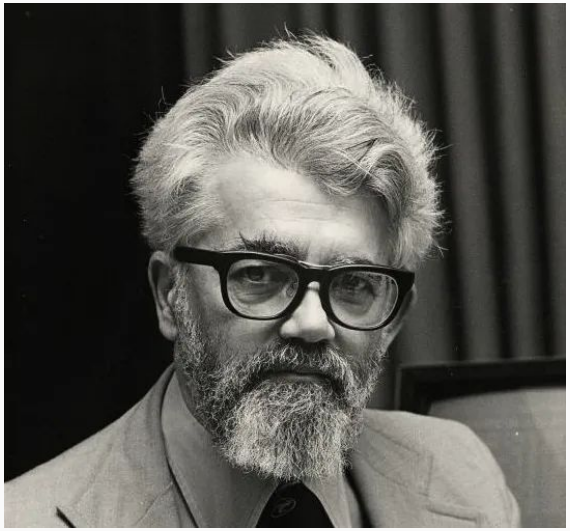
Using Python



Visit https://www.anaconda.com/products/individual
to download a Python distribution / data science “platform”

What is AI?

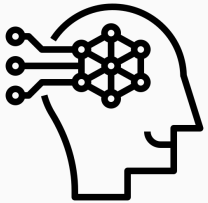
Artificial Intelligence is the science and engineering of making intelligent machines



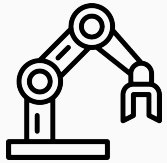
John Mccarthy

Natural Language Processing

Computer Vision

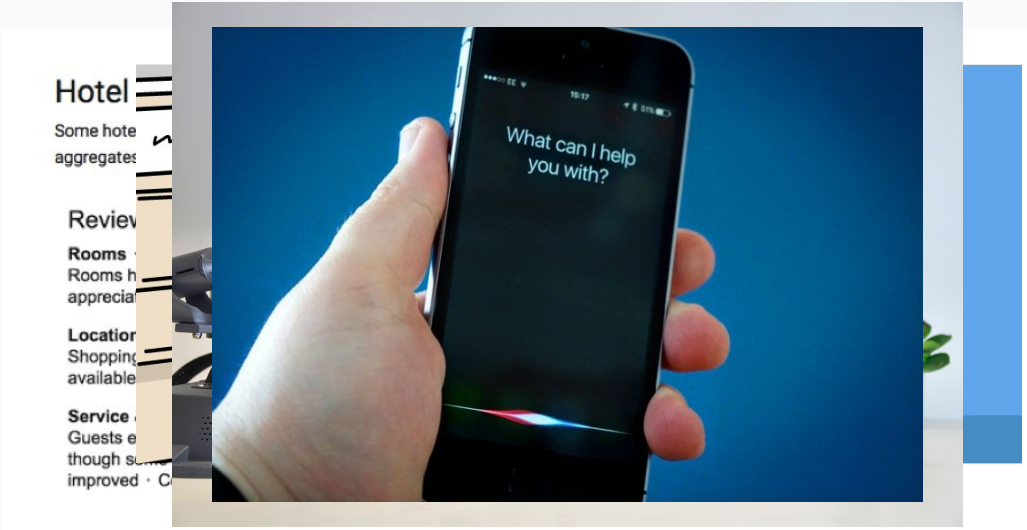


Machine Learning

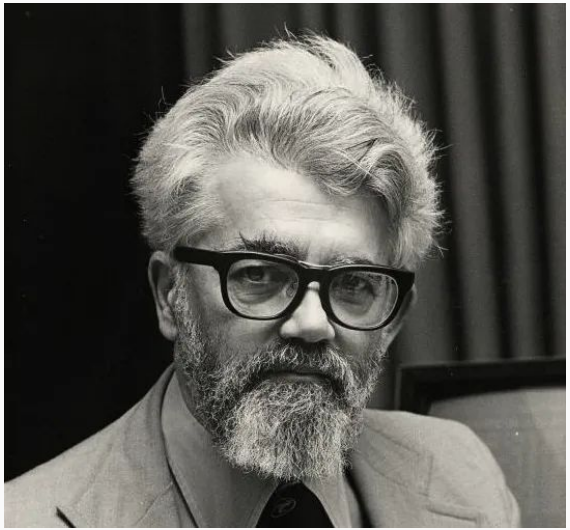


Automation & Robotics

Speech



AI is everywhere



John Mccarthy

Natural Language Processing

Computer Vision

Machine Learning

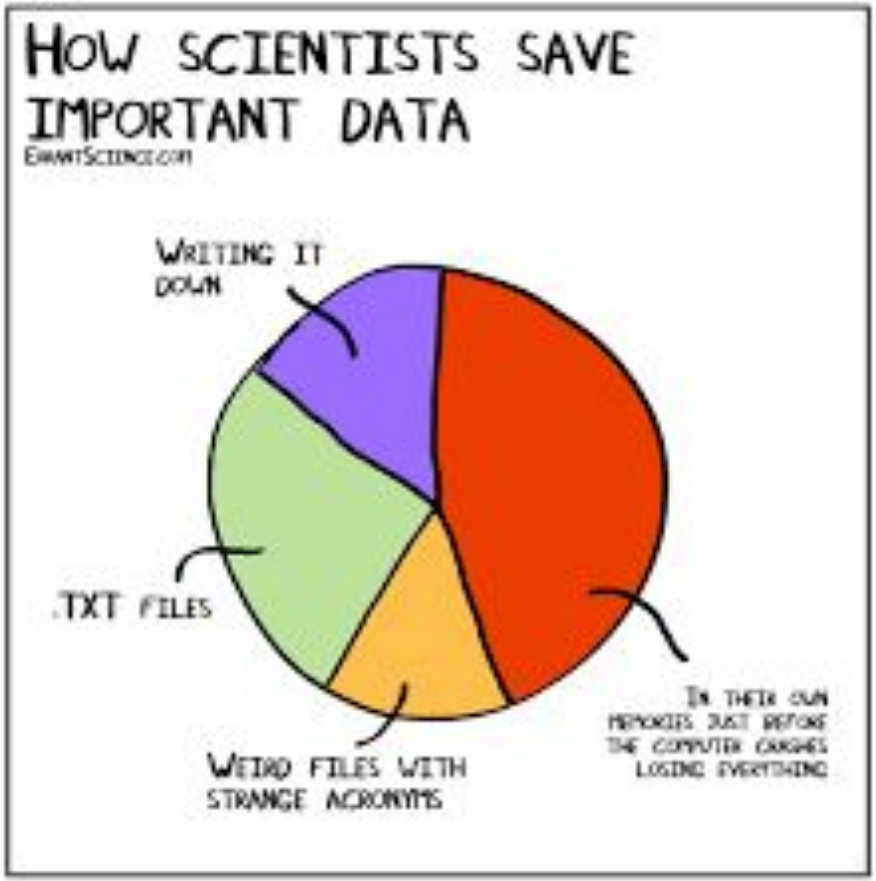
Automation & Robotics

Speech

Getting Data

What are the possible sources of data?

Data is key


HOW SCIENTISTS SAVE IMPORTANT DATA
WRITING IT DOWN
.TXT FILES
WEIRD FILES WITH STRANGE ACRONYMS
IN THEIR OWN MEMORIES JUST BEFORE THE COMPUTER CRASHES LOSING EVERYTHING

Possible Sources of Data

1. Primary Sources
   - Interviews
   - Surveys
   - Focus Groups
   - Sensors
     - Machines
     - Wearables
     - IoT
   - ...

2. Secondary Sources
   - Online Reviews
   - Online Comments
   - Social Media
   - Websites
   - ...

Ctrl C Ctrl V

Small scale web scraping

WIKIPEDIA
The Free Encyclopedia

Not logged in Talk Contributions Create account Log in

Article Talk | Read View source View history | Search Wikipedia

NASA

From Wikipedia, the free encyclopedia

Coordinates: 38°52′59″N 77°0′59″W

For other uses, see NASA (disambiguation).

The **National Aeronautics and Space Administration** (**NASA** /ˈnæsə/) is an independent agency of the U.S. federal government responsible for the civilian space program, as well as aeronautics and space research.[note 1]

NASA was established in 1958, succeeding the National Advisory Committee for Aeronautics (NACA). The new agency was to have a distinctly civilian orientation, encouraging peaceful applications in space science.[7][8][9] Since its establishment, most US space exploration efforts have been led by NASA, including the Apollo Moon landing missions, the Skylab space station, and later the Space Shuttle. NASA is supporting the International Space Station and is overseeing the development of the Orion spacecraft, the Space Launch System, Commercial Crew vehicles, and the planned Lunar Gateway space station. The agency is also responsible for the Launch Services Program, which provides oversight of launch operations and countdown management for uncrewed NASA launches.

NASA's science is focused on better understanding Earth through the Earth Observing System;[10] advancing heliophysics through the efforts of the Science Mission Directorate's Heliophysics Research Program;[11] exploring bodies throughout the Solar System with advanced robotic spacecraft such as *New Horizons*;[12] and researching astrophysics topics, such as the Big Bang, through the Great Observatories and associated programs.[13]

Contents [hide]

1 History

National Aeronautics and Space Administration

NASA seal

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

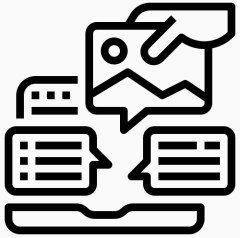
Contribute

Help
Learn to edit
Community portal
Recent changes
Upload file

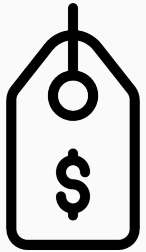
Tools

What links here
Related changes
Special pages
Permanent link

Different people scrape data for all sorts of (malicious) purposes

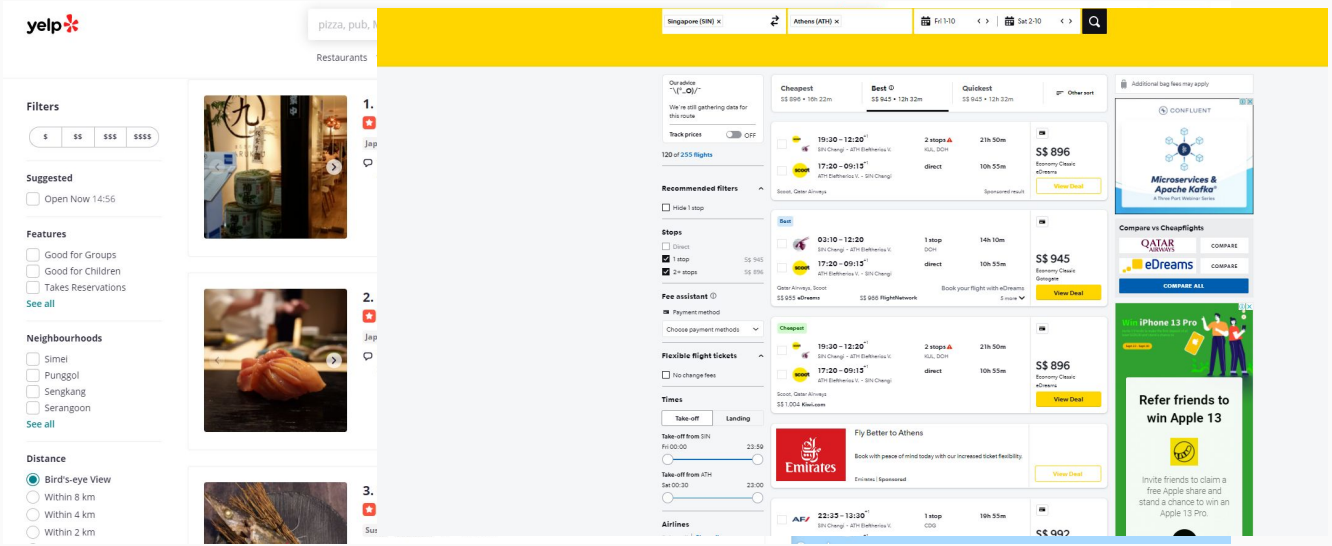


Content Scraping

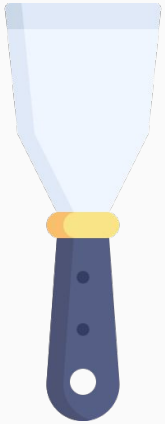


Price Scraping

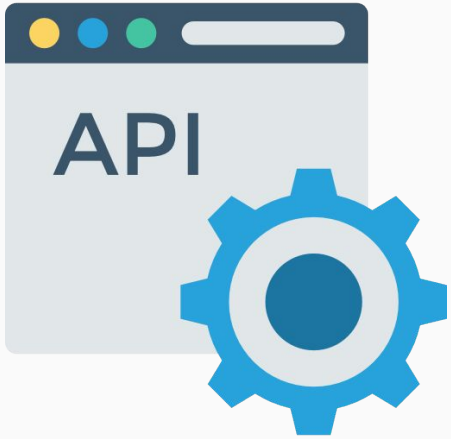
Contacts Scraping



Obtaining data from a webpage

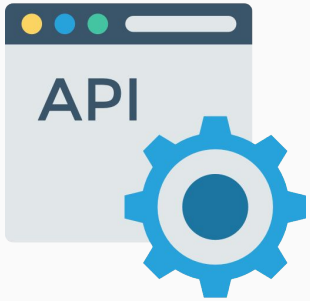


Web Scraper
Extract data from the website using a web scraping tool or software



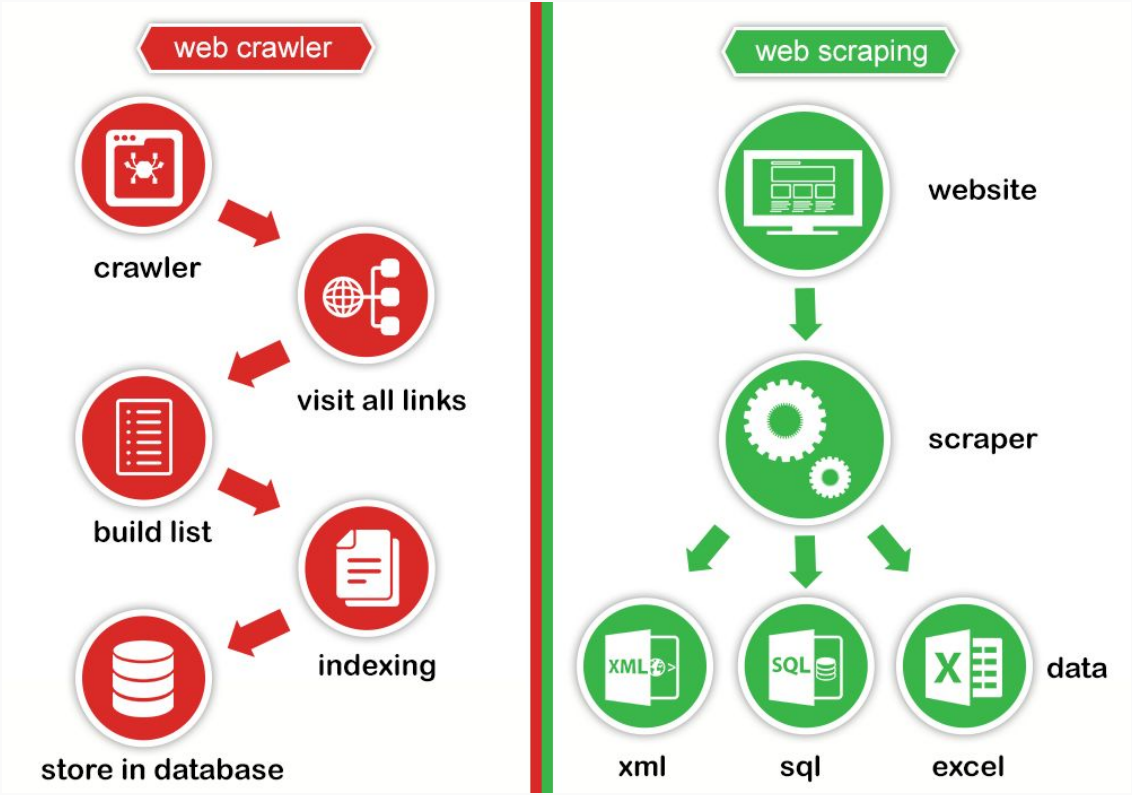
Application Programming Interface (API)
Website gives you direct access to specific data

Webscraper Vs API

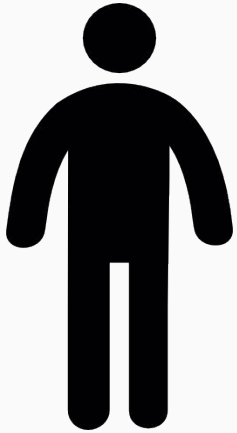


Webscraper	API
Done manually or by using software tools	Website needs to provide the API
May be illegal on some sites	Legal, but not all sites may provide
Can be programmed to collect the data you want	May not have all the data
F.O.C if done manually, but software tools may have a license	May require payment

Web Scraping Vs Web Crawling



Web Scraping Vs Web Crawling



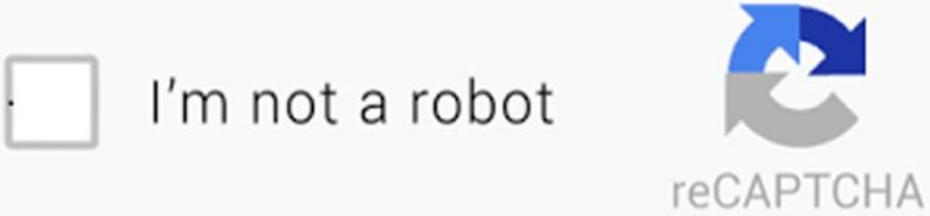
Scrapers will pretend to be web browsers, while a crawler bot will indicate its purpose and not attempt to trick a website into thinking it's something it is not

Scrapers will take advanced actions like filling out forms, or otherwise engaging in behaviors to reach a certain part of the website. Crawlers will not.

A scraper is designed to pull specific content, it may be designed to pull content explicitly marked to be ignored

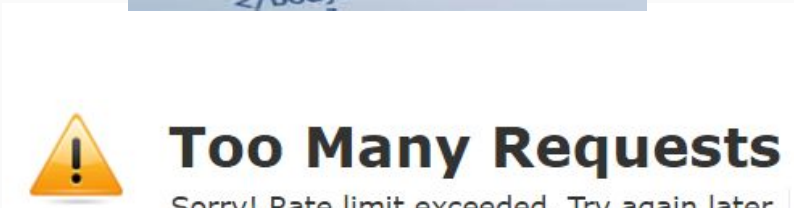
Some websites put in measures to STOP web scrapers from getting their data

CAPTCHAs



Change HTML markup at regular intervals

Rate limit requests



Is Web Scraping Legal?

Is Web Scraping and Crawling Legal in Singapore?

There is no specific law in Singapore that directly addresses whether web crawling or scraping is legal. However, crawling and scraping could possibly attract civil liability under existing contract law and copyright law, and even criminal liability under the Computer Misuse Act.

Breach of website terms of use

When accessing a website, the user generally agrees to access it in accordance with the website's "terms of use" as stated on the site, and this forms a legally binding agreement.

If the terms of use prohibit crawling and scraping, doing so on the website would be a breach of this agreement. The website owner may then sue the bot operator for breach of contract, and claim monetary compensation for any loss suffered in relation to any downtime or slowdown the website may have incurred.

For example, when property listing platform 99.co first started operations, it used a web scraper to scrape rental listings from competing platform PropertyGuru for listing on its own platform. PropertyGuru regarded this as potential breaches of its website's Terms of Service and Acceptable Use Policy and infringement of its copyright in the listings, and requested 99.co to stop doing so.

Both platforms eventually settled the matter out of court in 2015, with 99.co signing an agreement not to substantially reproduce the content in PropertyGuru's website without its consent.

https://singaporelegaladvice.com/law-articles/legal-scrape-crawl-websites-data-singapore/

Is Web Scraping Legal?

The People's Court of Suiyang District in Central Henan Province imprisoned the Chinese software developer and his employer for three years in prison and a $70,260 fine (450,000 Yuan). the marketer used web scraping software to access data that was not publicly available

Is Web Scraping Legal?

POLICY \ US & WORLD \ TECH \

Clearview AI hit with sweeping legal complaints over controversial face scraping in Europe

The privacy watchdogs believe Clearview's image-scraping methods violate European laws

By Ian Carlos Campbell | @soupsthename | May 27, 2021, 5:48am EDT

The complaints filed in France, Austria, Greece, Italy, and the United Kingdom say that the company's method of documenting and collecting data — including images of faces it automatically extracts from public websites — violates European privacy laws. New York-based Clearview claims to have built "the largest known database of 3+ billion facial images."

How do we scrape data?

1. Extract the HTML
2. Parse the HTML
3. Extract the relevant the necessary data
4. Store/Transform the data

Hands-on!