- DA 2009 -

DATA COLLECTION

<//METHODS - II//>

Week 1: Introduction to Web-Based Data Collection

</// ABOUT THIS COURSE – HOUSEKEEPING //>

Objectives:

- CLO1: **demonstrate** awareness of fundamental concepts in data collection from the web
- CLO2: **identify** and **apply** suitable tools and techniques to extract data from the web
- CLO3: **practice** ethical standards and professional integrity in collecting data from the web

No. of credits:

2 (10 weeks, 20 L – 20 P – 60 SS)

Assessment:

Quizzes - 20% (CLO 1, 2) ← Week 3
Case Studies - 40% (CLO 1, 2, 3) ← Week 6 & 10
Final Assessment - 40%

DA 2009 - Course Plan

<// SOME GROUND RULES //>

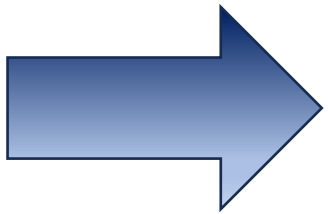
- **Don’t hesitate to ask questions** - no one knows 100% of anything
- **If your code works perfectly on the first try, check again**... something’s probably off!
- **It’s okay to feel stuck** - understanding comes from trial, error, and curiosity.
- **take a moment to think**: What do I need, and what am I trying to do?
- **Treat this like an adventure** - explore and have fun along the way!

<// TIME TO BREAK THE ICE //>

Click the link /

scan the code

and join the poll!



https://app.sli.do/event/qDdHSs3wLbBtvXfmsWw7oQ

</ DATA COLLECTION IN THE DIGITAL AGE />

In today's digital-first world, vast amounts of information are constantly being generated, shared, and stored online. Data collection has become fundamental to decision-making, personalization, business intelligence, and scientific research.

With data driving innovation in industries such as healthcare, marketing, finance, and education, the ability to access and analyze web-based data has become a key skill.

</> EXAMPLE – COVID 19 DASHBOARDS //>

During the COVID-19 pandemic, online data sources such as case dashboards, mobility reports, and social media sentiment helped governments and researchers make informed public health decisions.

- Worldometers: https://www.worldometers.info/coronavirus/
- WHO: https://data.who.int/dashboards/covid19/cases

</// SURVEYS VS WEB-BASED DATA COLLECTION //>

Aspect	Surveys (Primary Data Collection)	Web-Based Data Collection
Source	Often directly from respondents	Existing online data (websites, social media)
Cost & Speed	Higher cost and slower speed	Low cost and fast, often real-time
Control	Full control over questions and data format	Limited control since data is pre-generated
Ethics	Requires informed consent and ethical approval	Privacy and scraping legality as declared by the website
Use Cases	In-depth insights, causal analysis	Trend monitoring, large-scale behavior analysis

<// TYPES OF DATA AVAILABLE IN THE WEB //>

Web data can be broadly classified into three categories:

1. **Structured data**: This includes data that is highly organized and stored in tabular formats, such as databases or spreadsheets. Structured data is easy to search and analyze using standard query languages like SQL.
   Ex: A table of product prices in an online store's backend database.

2. **Unstructured data:** This includes data without a predefined format, such as plain text, images, videos, and audio files.
   Ex: A blog article, YouTube video, or customer review.

3. **Semi-structured data:** This type of data has some organizational properties but does not reside in a strict tabular format. It is commonly found in formats like JSON, XML, or CSV.
   Ex: API responses from weather data services or tweets in JSON format.

<// TYPES OF DATA – STRUCTURED DATA //>

- Structured data organized and stored in tabular formats, such as databases or spreadsheets.
- The rows and columns of the data should be related to one another.
- Structured data are stored in Relational Database Management Systems (RDBMS).
- Structured data is easy to search and analyze using standard query languages like SQL.

Currency Code	Currency Name	T/T Buying	O/D Buying	T/T Selling
USD	U.S. Dollar	295.75	294.1295	303.75
GBP	U.K. Pound	405.1551	404.2541	419.1301
EUR	Euro	343.8748	343.0948	357.5081
JPY	Japanese Yen	2.0366	2.0315	2.117
AUD	Australian Dollar	192.0672	190.9776	200.8935
NZD	New Zealand Dollar	178.4491	177.562	185.5681
CHF	Swiss Franc	367.0689	366.1931	381.7145
SEK	Swedish Krona	30.8889	30.8125	32.1175
DKK	Danish Krone	46.0937	45.9692	47.9118
CAD	Canadian Dollar	215.5319	214.9878	224.0851
SGD	Singapore Dollar	230.6492	230.1851	239.8449
HKD	Hong Kong Dollar	37.4518	37.3698	38.9271

<//TYPES OF DATA – UNSTRUCTURED DATA//>

- This data does not have a predefined format, such as plain text, images, videos, and audio files.
- Social media comments and online reviews are also considered as a type of unstructured data.
- They are irregular, and ambiguous (uncertain) – stored in a data lake.
- Almost 80%-90% of data are unstructured, hence easily available & very useful!
- Most of the ML / AI related analytics are performed on unstructured data.

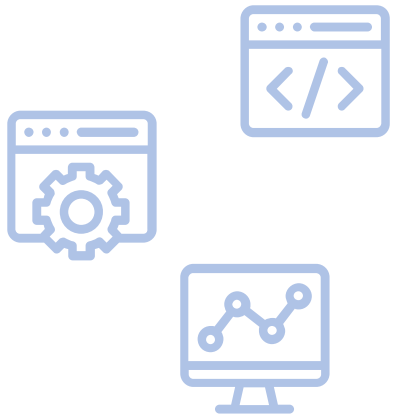
  Ex: Social media recommendation systems, Netflix, Supermarkets.

</// TYPES OF DATA – SEMI-STRUCTURED DATA //>

- This type of data has some organizational properties but does not reside in a strict tabular format.
- Emails are a type of unstructured data – since the metadata can fit in a database.
- Falls between structured and unstructured data.
- It is commonly found in formats like JSON (JavaScript Object Notation), XML (Extensible Markup Language), or CSV (Comma Separated Values).

Ex: API responses from data services or tweets in JSON format.

```
{"employees":[
  { "firstName":"John", "lastName":"Doe" },
  { "firstName":"Anna", "lastName":"Smith" },
  { "firstName":"Peter", "lastName":"Jones" }
]}
```



</// LOADING WEB DATA USING MS EXCEL //>

MS Excel allows you to easily extract tabular data from the web:

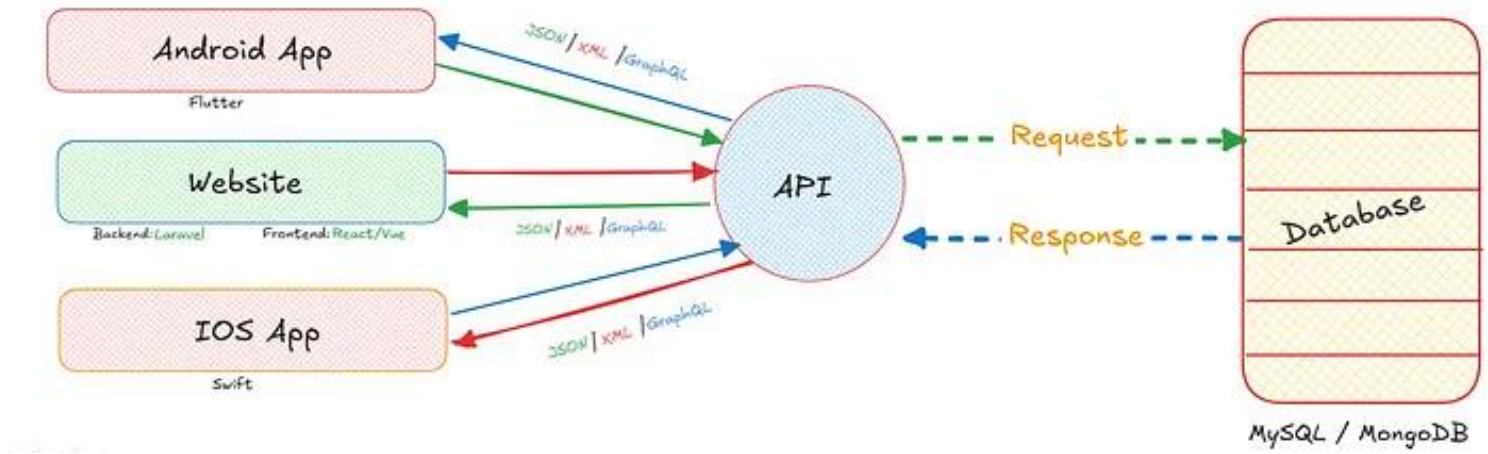
1. **Select Data → Get Data → From Web**
2. **Enter the Web URL**: Provide the link to the webpage you want to extract data from.
3. **Preview** the Data: Excel connects and displays available tables from the page.
4. **Select & Transform**: Choose the desired table, clean or shape the data as needed.
5. **Load into Excel**: Load the final data directly into your spreadsheet for analysis.

Let's try it out!
https://www.sampath.lk/rates-and-charges?activeTab=exchange-rates

Now let's try this,
https://www.espncricinfo.com/series/bangladesh-in-sri-lanka-2025-1485499/sri-lanka-vs-bangladesh-2nd-test-1485504/full-scorecard

<// HOW WEBSITES TALK TO A DATABASE //>

- ✓ The frontend sends a request (to fetch, update or delete some data).
- ✓ The API receives this request and talks to the database on your behalf.
- ✓ The database processes the query and responds with the required data.
- ✓ The API sends that data back to the frontend, usually in: JSON, XML or GraphQL format.



<// WEB DATA BACKEND SOURCES //>

- Web data ultimately originates from backend storage systems but is accessed through user-facing interfaces like webpages or APIs.

Relational Databases	Non-Relational Databases (NoSQL)	Data Warehouses	Data Lakes
Structured, table-based systems using SQL.	Schema-less, flexible storage for varied data types.	Centralized storage optimized for analytics and querying.	Store volumes of raw, unstructured or semi-structured data.
Ex: MySQL, PostgreSQL, Oracle	*Ex:* MongoDB, Firebase, Cassandra	*Ex:* Google BigQuery, Snowflake	*Ex:* Hadoop HDFS, AWS S3, Azure Data Lake

</// WEB DATA DELIVERY LAYERS //>

- Web data ultimately originates from backend storage systems but is accessed through user-facing interfaces like webpages or APIs.

APIs (Application Programming Interfaces)

Defines a standard syntax for one software to communicate with another, returning structured formats like JSON or XML.

Ex: REST APIs, Twitter API, OpenWeather API

Web Pages (HTML / JavaScript)

The visible user interface of a website. Users extract content from static or dynamically rendered pages.

Ex: Parse-able content in any web page

Files and Documents:

Web pages often link to downloadable resources such as PDFs, spreadsheets, and CSV files.

Ex: Downloadable links available on webpages

<// STATIC VS DYNAMIC CONTENT //>

Static Content

- This refers to content that **does not change unless manually updated** by a webmaster (by modifying the underlying HTML code). It is directly available in the page source.
- Load faster but has a minimalistic design.
- Choose if: The content remains relatively same over time.

  Ex: The list of departments at a university.

Dynamic Content

- Content that is **generated or updated on the fly** using JavaScript or AJAX calls through the underlying database. Often requires interaction or asynchronous loading.
- Interactive and more appealing but might take time to load.
- Choose if: Content is updated regularly and requires users to login.

  Ex: A real-time sports scoreboard.

  Facebook feed content.

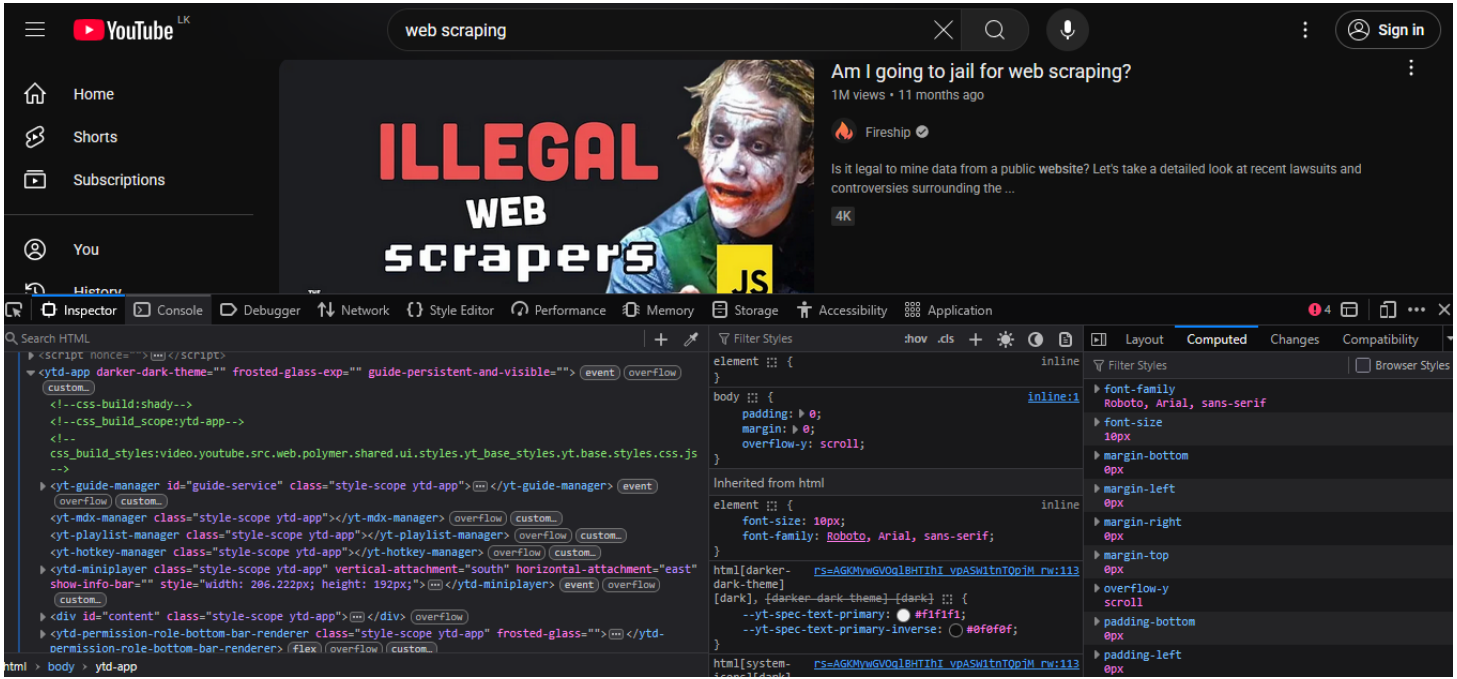
</> ACTIVITY - INSPECTING A WEB PAGE </>

Inspecting a web page involves using browser developer tools to examine the underlying HTML, CSS, and JavaScript. It helps identify the structure of the page, locate specific elements (like text, images, or data tables), and understand how content is loaded. This process is essential for web scraping, as it guides how to target and extract the desired data accurately.

To inspect a page:

- ✓ Right-click on the webpage and select “Inspect” (or press F12).
- ✓ Use the Elements tab to browse the HTML structure.
- ✓ Use the Network tab to see data requests and API calls.
- ✓ Identify relevant tags, classes, or IDs for scraping.

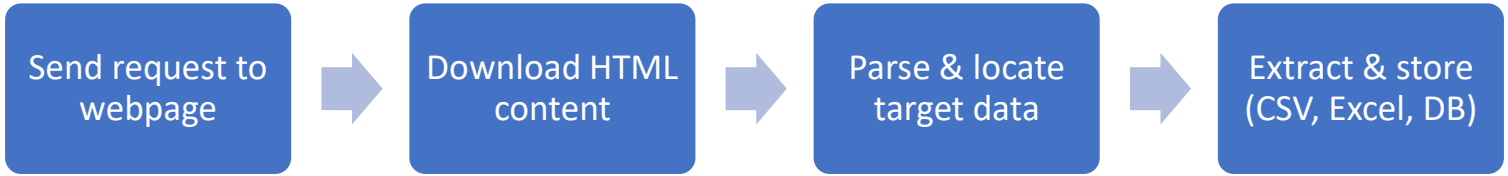
<// ACTIVITY - INSPECTING A WEB PAGE //>



<// WHAT IS WEB SCRAPING //>

Web scraping is the automated extraction of data from websites. It allows you to collect structured or semi-structured information for analysis, monitoring, or reporting.

Key Steps:



Types of Data Collected:

- Structured Data (tables, lists)
- Semi-Structured Data (JSON from APIs)
- Unstructured Data (text from articles or reviews)

</// USE CASES – PUBLIC HEALTH //>

Monitoring disease outbreaks, tracking health behavior, or identifying early warning signals through digital platforms:

- Google Trends (trends.google.com) – Tracks search queries related to symptoms (e.g., "fever", "cough"), useful for detecting outbreak patterns.
- HealthMap (www.healthmap.org) – Aggregates data from news, government sources, and social media to map global disease activity.
- Twitter/X – Analyzes real-time posts for syndromic surveillance, public sentiment on vaccines, or localized health concerns.

</> USE CASES – E-COMMERCE </>

Analyzing online product listings, consumer feedback, and competitor pricing for business intelligence.

- Amazon (www.amazon.com) – Used for collecting product reviews, ratings, seller rankings, and dynamic pricing data.
- eBay (www.ebay.com) – Tracks auction behavior, pricing trends, and customer demand across various product categories.
- PriceGrabber (www.pricegrabber.com) – Compares prices across multiple online retailers for consumer electronics, fashion, etc.

</// USE CASES – RETAIL //>

Scraping retail websites helps monitor product availability, dynamic pricing, user feedback, and promotional campaigns across multiple platforms for market intelligence and competitor analysis.

- Keells (www.keellssuper.com) – Offers access to product listings, promotions, category-based prices, and stock status.
- Walmart (www.walmart.com) – Useful for analyzing inventory levels, product pricing history, discounts, and delivery availability.
- Daraz (www.daraz.lk) – Tracks product listings, seller competition, discount strategies, and flash sale behavior across diverse categories.

<// USE CASES – NEWS MONITORING //>

Tracking news narratives, media coverage, misinformation trends, and public discourse over time.

- Google News (news.google.com) – Aggregates breaking news across regions and categories; useful for topic tracking.
- Reuters, BBC, CNN – Offer structured headline feeds and metadata that can be scraped to monitor reporting patterns.
- Media Cloud (www.mediacloud.org) – Analyzes news bias, media attention cycles, and source comparison in a research-friendly format.

<// USE CASES – FINANCE & INVESTMENT //>

Collecting financial data for quantitative analysis, portfolio decisions, or algorithmic trading.

- Yahoo Finance (https://finance.yahoo.com/) – Provides historical stock prices, financial statements, and analyst ratings.
- CoinMarketCap (www.coinmarketcap.com) – Offers real-time cryptocurrency pricing, market cap, and volume data.
- Investing.com – Supplies economic indicators, forecasts, exchange rates, and commodity prices.

</// USE CASES – JOB MARKET ANALYSIS //>

Studying job postings, hiring patterns, skill demands, and salary insights across sectors:

- LinkedIn (www.linkedin.com) – Tracks job ads, company hiring trends, and professional skill networks.

- Indeed (www.indeed.com) – Provides large-scale job listings with salary ranges and job description keywords.

<// USE CASES – EDUCATION & RESEARCH //>

Accessing academic content, citation patterns, and online learning trends.

- Google Scholar (scholar.google.com) – Allows scraping of publication titles, citation counts, authors, and research topics.

- Coursera (www.coursera.org) – Contain course metadata, enrollment stats, and subject coverage across global universities.

</// USE CASES – TRANSPORTATION & TRAVEL //>

Tracking airfare, hotel pricing, travel routes, and user-generated reviews for tourism analytics:

- Skyscanner (www.skyscanner.com) – Used to compare airfare trends across time, routes, and airlines.
- Booking.com – Provides hotel availability, prices, user ratings, and seasonal demand data.
- TripAdvisor (www.tripadvisor.com) – Useful for collecting destination reviews, travel preferences, and tourist behavior insights.

</// USE CASES – SPORTS //>

Tracking player statistics, match schedules, ticket pricing, fan sentiment, and media trends in sports:

- ESPN Cricinfo (www.espncricinfo.com) – Provides cricketer statistics, team rankings, match schedules, historical performance data, and live score updates.

- Transfermarkt (www.transfermarkt.com) – Offers football data on player transfers, market values, team formations, and match history

</// USE CASES – ENTERTAINMENT //>

Tracking movie ratings, top music charts, ticket pricing, fan sentiment, and media trends in entertainment sectors like music and movies:

- Spotify (https://developer.spotify.com/) – Through its API, it allows access to streaming counts, artist popularity, playlist trends, and regional listening patterns. Helpful in understanding music consumption behavior.

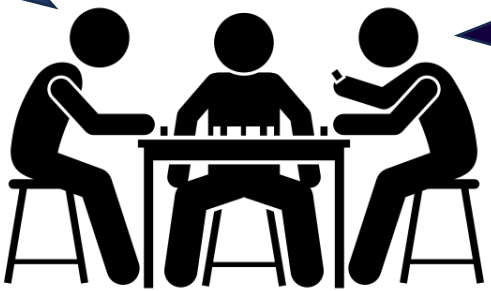
- IMDb (www.imdb.com) –> Provides movie ratings, cast info, box office trends, and user reviews

</// TRUE OR FALSE? //>

If data is visible on a website, it is always legal to scrape it.

Dynamic websites cannot be scraped using any tools.

Data obtained through scraping is immediately ready for analysis.



</// COMMON WEBSCRAPING MISCONCEPTIONS //>

×	Web scraping is always considered illegal	➡ Not always illegal
×	Ability to view data on a website means it is legal to scrape	➡ No, some website components are not allowed to be scraped
×	Web scraping is the same as automated copy-pasting	➡ It’s much more than that!
×	Any scraping tool can be used on any website	➡ There’s no “works-for-all” universal scraper
×	Dynamic websites cannot be scraped	➡ They can be scraped using the right tools
×	Data collected through scraping is ready for analysis	➡ No, data cleaning is essential after scraping

</> CHALLENGES OF WEB SCRAPING </>

1. **Website Structure Changes:**
   - Even minor updates to HTML can break scrapers
   - Requires frequent maintenance and updates
2. **Anti-Bot Measures:**
   - Captchas, login walls, and rate limits block scraping
   - Sites may use JavaScript-based detection or IP blocking
3. **Dynamic Content:**
   - Data loaded via JavaScript/AJAX may not appear in initial HTML
   - Requires tools like Selenium or headless browsers to capture
4. **Legal and Ethical Concerns:**
   - Not all scraping is allowed — always check terms of service
   - Avoid scraping personal or copyrighted content
5. **Data Quality Issues:**
   - Extracted data may be incomplete, duplicated, or messy
   - Needs cleaning, validation, and transformation

<// TO DO LIST BEFORE WEEK 2 //>

- Install Python 3.x
- Create a GitHub account
- Read more about the use cases of web scraping

https://app.sli.do/event/qDdHSs3wLbBtvXfmsWw7oQ


THANK YOU
<SEE YOU NEXT WEEK!>