




# Data Sources

SGA07\_DATASCI

21<sup>st</sup> January 2020



# Module Overview

- Web Crawlers
- Social Media
- Field Research Methods



# Book Keeping

- We have a TA (Joseph Oladokun)
- Past work review should begin earnestly
- 3 Practice Labs in this module
- Expect a long session on Thursday (Wk5 Catch up)



# Outcome

After this Module, you will;

- Understand the formal concepts to source for information
- Understand the principle and architecture of a web crawler
- Get an overview of social media in the evolution of web technologies
- Understand the importance of field research to gain better context on projects



# Web Crawler

- Index keywords and phrases for effective search
- designed to have speed, completeness, accuracy and scalability

“

A web crawler is an agent (software program) designed to search a website(s) so as to scrape, clean and organise its data for efficient analysis.

”

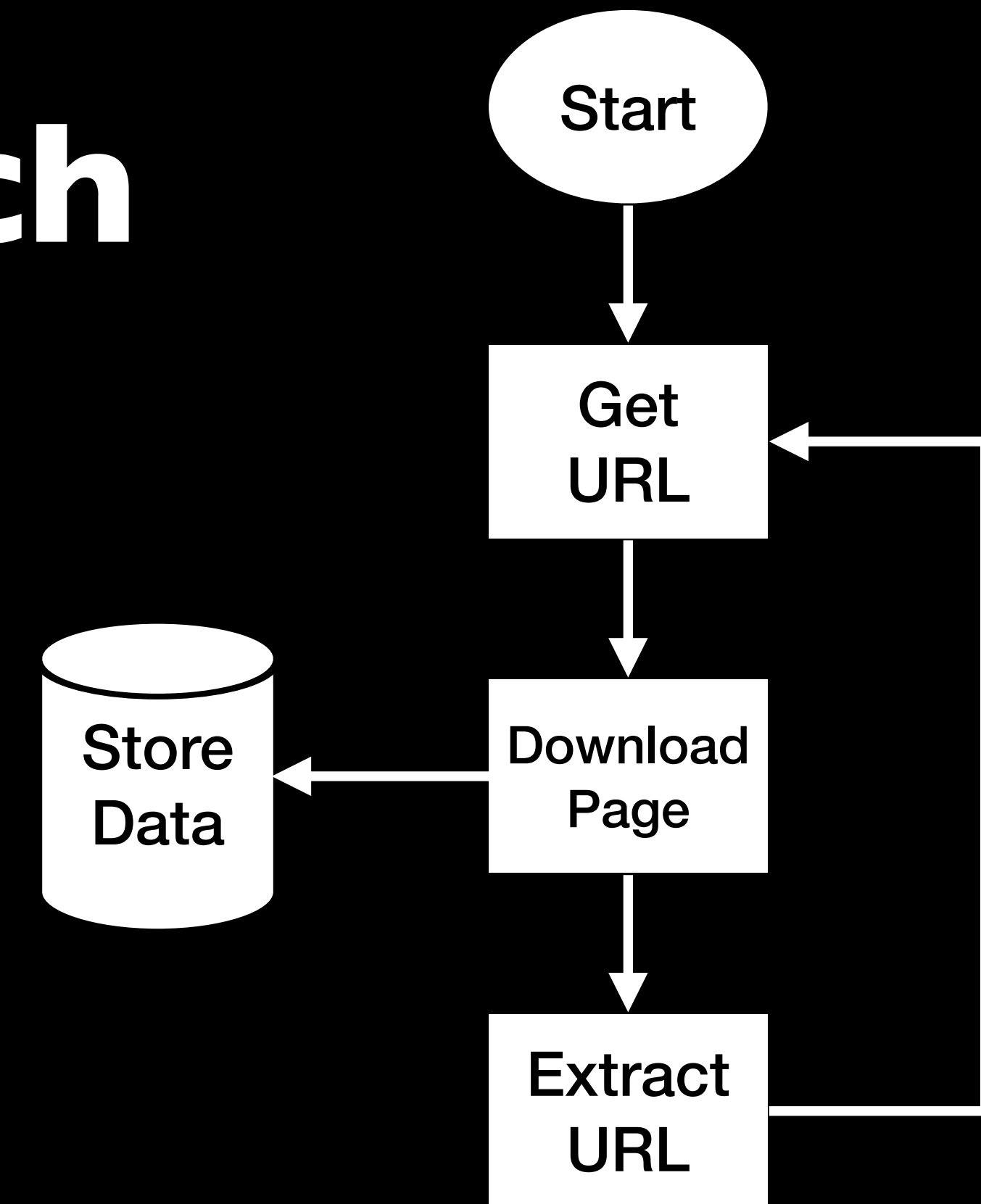


# Web Crawling Techniques

Crawlers	Search Pattern	Drawbacks
Breadth-first Crawling	Scans neighbor node from root level, if result not achieved then go to next level.	When the branches or tree is very deep then goes into infinite.
Depth-first Crawling	Scans from the root node and traverse next to its child leftmost node	Takes more time when the child node is large.
Targeted Crawling	Uses random (heuristics) crawling process	Takes more time when specific topics are very large.
PageRank Algorithms	works on the importance of the web pages. It calculates inlinks or backlinks to that page.	Difficult to manage and update page index repository.

# Web Crawler Approach

- Pull: Proactive search
- Push: Content Aggregator





# Practice Lab

Write a python program to crawl, parse and store data from a website

Use the following Instructions:

- Create a new file web-crawler.py in your SGA07\_DATASCI directory
- Copy the code provided in the link [https://github.com/akinlabiceo/SGA07\\_DATASCI.git/hotel-web-crawler.py](https://github.com/akinlabiceo/SGA07_DATASCI.git/hotel-web-crawler.py)
- Install the following packages: requests and BeautifulSoup4
- Run the code `python3 web-crawler.py` in your terminal



# Social Media

“

Social media are platforms (dynamic websites) that leverages on the technologies of Web 2.0 to curate and share information through virtual communities and networks.

”

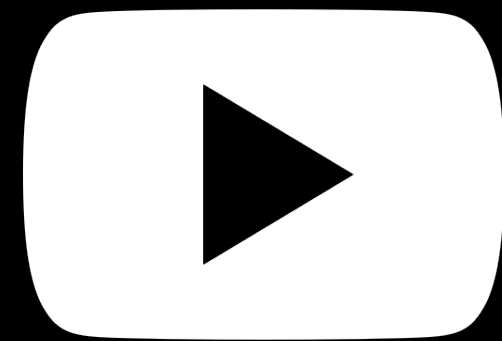
Web 1.0	Web 2.0	Web 3.0
Content- destination sites and personal portals.	Speedy- more timely information and more efficient tools to find information.	Ubiquitous- available at any time, anywhere and through any channel or device.
Search- critical mass of content derives need for search engines.	Collaborative- actions of users a mass, police, and prioritize content.	Efficient- relevant and contextual information find-able instantly.
Commerce- goes mainstream; digital good rise.	Trust Worthy- users establish trust networks and home trust radars.	Individualized- filtered and shared by friends or trust networks.

---

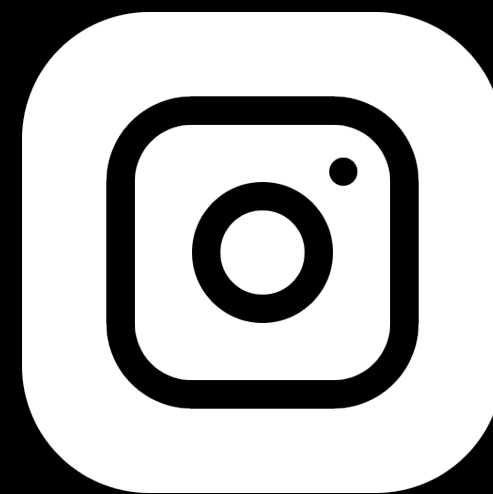
# Popular Social Networks



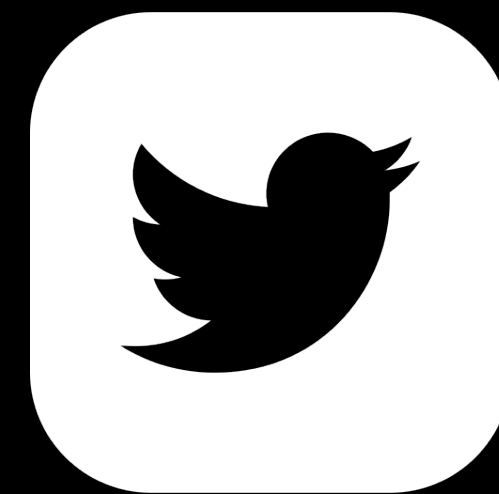
**Facebook**



**YouTube**



**Instagram**



**Twitter**



**LinkedIn**



**WhatsApp**



# Practice Lab

Write a python program to crawl, parse and store data from a twitter

Use the following Instructions:

- Create a new file `twitter-web-crawler.py` in your `SGA07_DATASCI` directory
- Create a Twitter developer account
- Copy the code provided in the link [https://github.com/akinlabiceo/SGA07\\_DATASCI.git/twitter-web-crawler.py](https://github.com/akinlabiceo/SGA07_DATASCI.git/twitter-web-crawler.py)
- Install the following packages: `tweepy`
- Run the code `python3 web-crawler.py` in your terminal



# Field Research

- techniques as observation, interviews and survey
- gather information that helps challenge your assumptions
- better understand people and context

“

The application of both qualitative and quantitative methods of data collection that aims to observe, interact and understand people while they are in a natural environment.

”

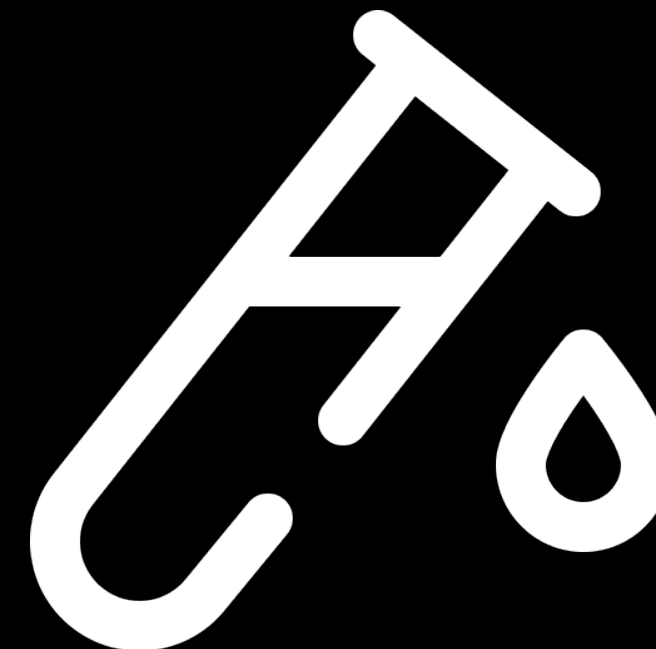
# Field Research Approach



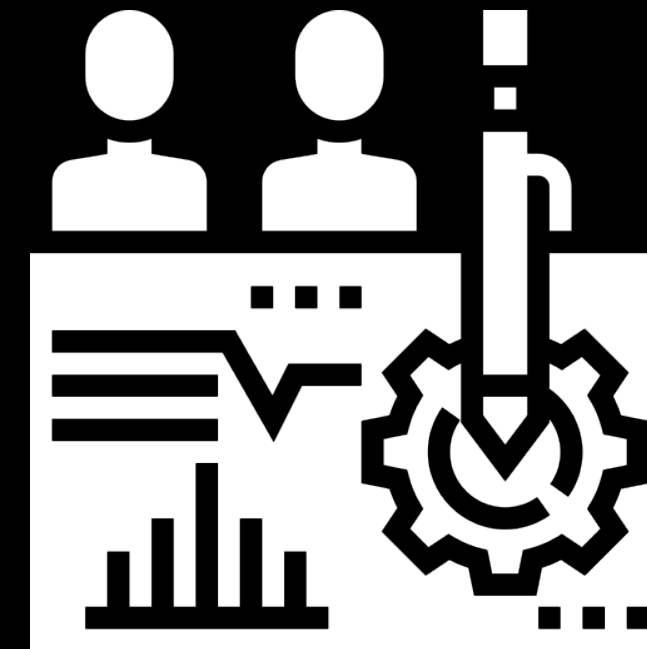
**Formulate  
Problem  
Statement**



**Effective  
Research  
Plan**



**Sample &  
Context  
Selection**



**Data  
Collection**



# Field Research Methods

Methods	Description
Direct Observation	In this method, the data is collected via an observational method or subjects in a natural environment. In this method, the behaviour or outcome of situation is not interfered in any way by the researcher.
Qualitative Interviews	Qualitative interviews are close-ended questions that are asked directly to the research subjects. The qualitative interviews could be either informal and conversational, semi-structured, standardised and open-ended or a mix of all the above three.
Survey Questionnaire	A questionnaire is a research instrument that consists of a set of questions or other types of prompts that aims to collect information from a respondent. A research questionnaire is typically a mix of close-ended questions and open-ended questions.



# Practice Lab

Use field research to build intuition on your final project

Use the following Instructions:

- Develop problem statement that encompasses your final project
- Develop a project plan with tasks, timeline and milestones for the project (constraint to at least 6 weeks)
- Compile a list of interview questions you would like to ask an expert in the domain of your project
- Develop a survey questionnaire (you can use survey monkey) that you share through your social media to get a general perspective for your project



# Recap/Summary

At the end of this Module, you should understand;

- Introduce you to web crawlers (techniques & approach)
- Overview of web technology evolutions
- Brief review of popular social networks
- Introduce you to field research (approach and methods)





# Suggested Material

- <https://www.webfx.com/blog/internet/what-is-a-web-crawler/>
- <https://computer.howstuffworks.com/internet/basics/search-engine1.htm>
- <https://www.octoparse.com/blog/web-crawling-how-to-build-a-crawler-to-extract-web-data>
- Analysing Different Web Crawling Methods by Bhavin M. Jasani, International Journal of Computer Applications (0975 - 8887) Volume 107 - No 5, December 2014
- Foundations and Trend R in Information Retrieval Vol. 4, No. 3 (2010) 175–246 2010  
C. Olston and M. Najork DOI: 10.1561/15000000017



# Suggested Material

- <https://www.analyticsvidhya.com/blog/2019/10/web-scraping-hands-on-introduction-python/>
- [https://en.wikipedia.org/wiki/Social\\_media](https://en.wikipedia.org/wiki/Social_media)
- <https://ourworldindata.org/rise-of-social-media>
- <http://www-public.imtbs-tsp.eu/~gibson/Teaching/Teaching-ReadingMaterial/OReilly07.pdf>
- <https://www.znetlive.com/blog/web-2-0/>



# Suggested Material

- <https://www.promptcloud.com/blog/scrape-twitter-data-using-python-r/>
- <https://www.questionpro.com/blog/field-research/>
- This is Service Design Doing: Applying Service Design Thinking in the Real World by Marc Stickdorn, Markus Edgar Hormess, Adam Lawrence and Jakob Schneider:  
Chapter 5 Research