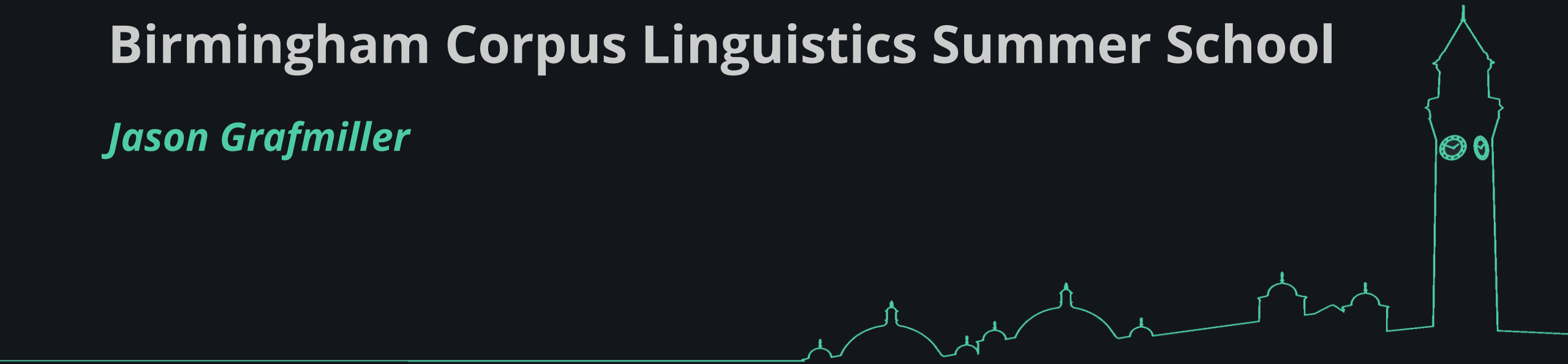
UNIVERSITY OF BIRMINGHAM

Web scraping with R

Part 1: What is web scraping?

Birmingham Corpus Linguistics Summer School

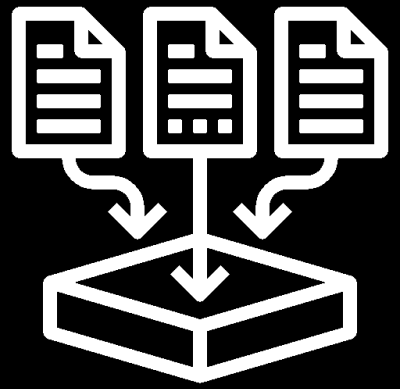
Jason Grafmiller



A basic workflow

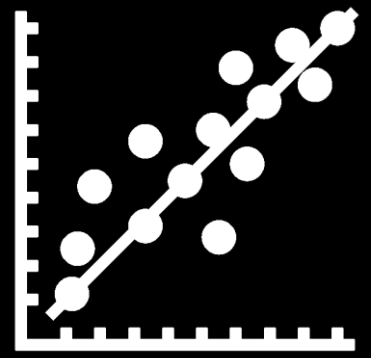
data source

data collection



data cleaning

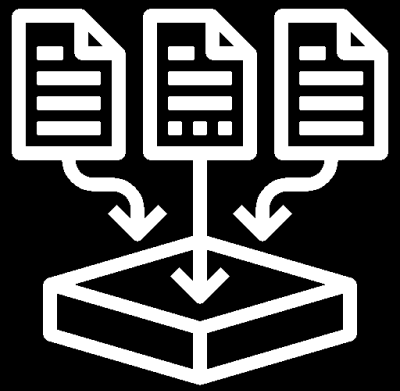
data analysis



A basic workflow

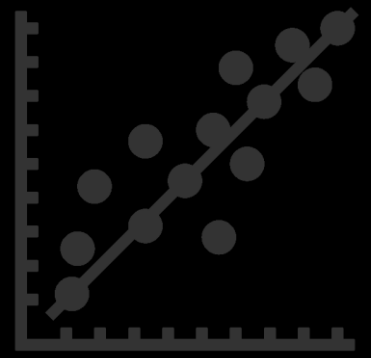
data source

data collection



data cleaning

data analysis



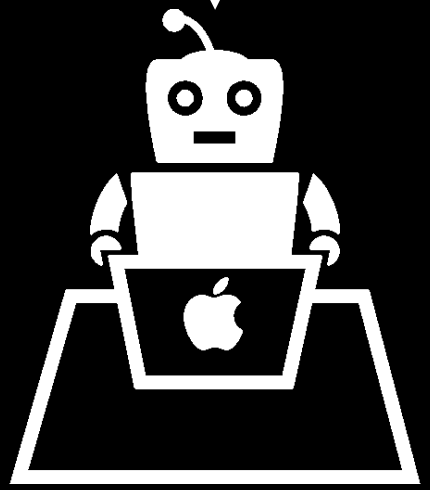
This session

“Webscraping is a term for various methods used to collect information from across the Internet. Generally, this is done with software that simulates human Web surfing to collect specified bits of information from different websites.”

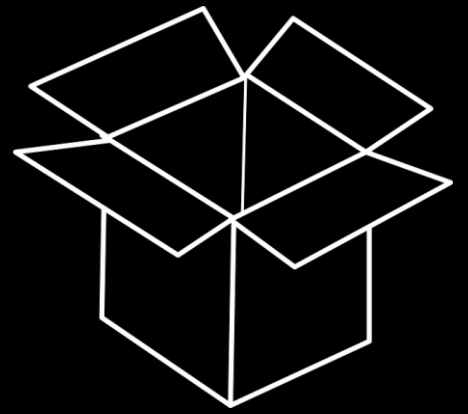
From: https://www.techopedia.com/definition/5212/web-scraping

Getting data from the Web

Copy/Paste



Scrape

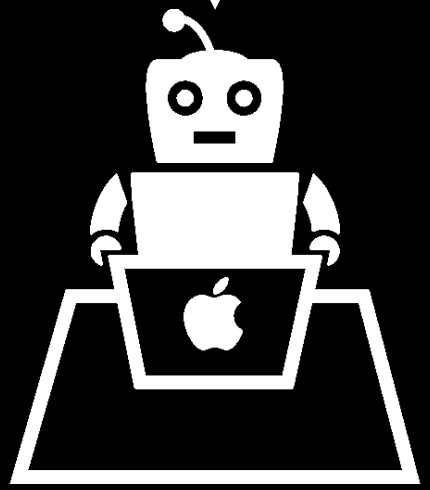


API

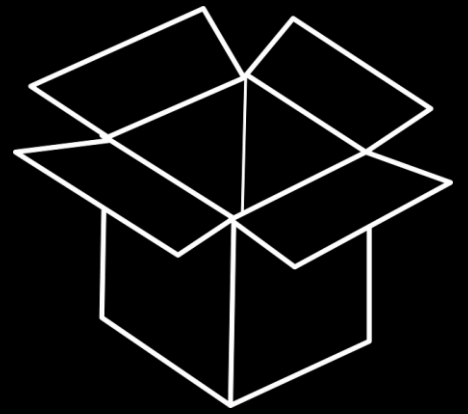
Getting data from the Web

Easy and accurate, but VERY slow

Copy/Paste



Scrape

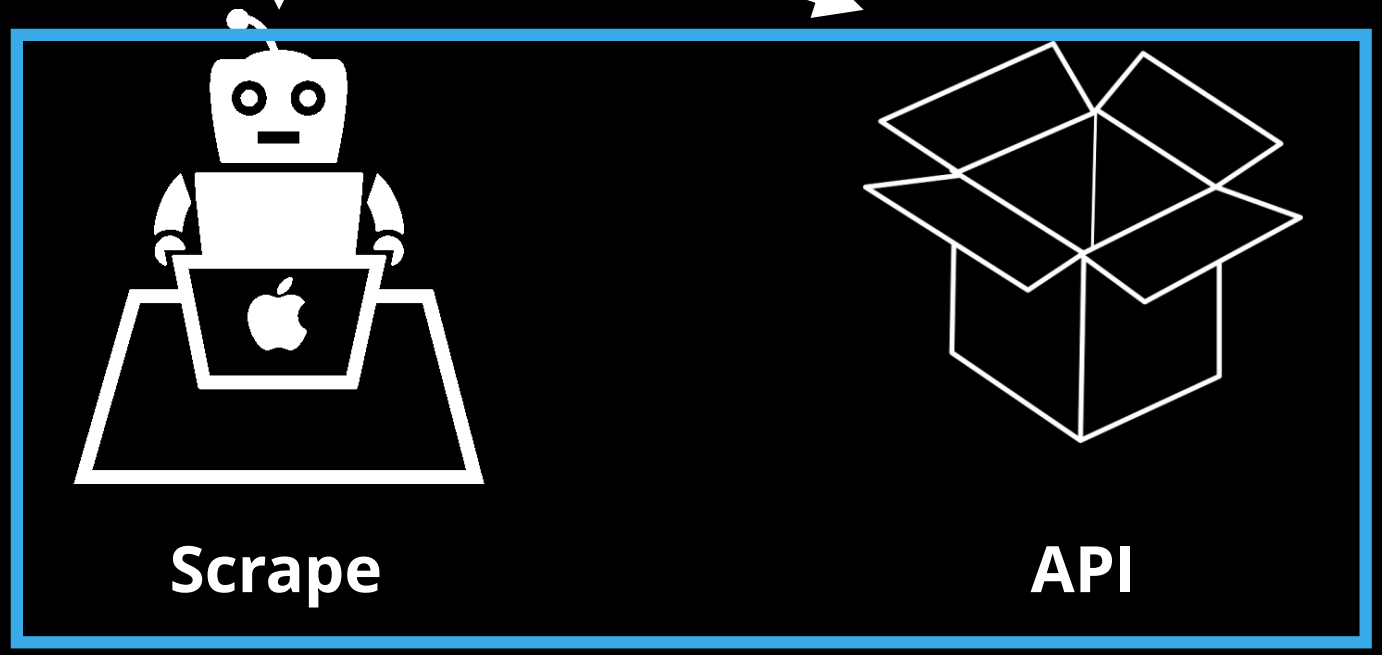


API

Getting data from the Web

Easy and accurate, but VERY slow

Fast, but a bit more complex to use



Getting data from the Web automatically

Scraping involves copying information on a webpage and storing it in another format for later use

- Basically mimics what a person might do with copy/paste

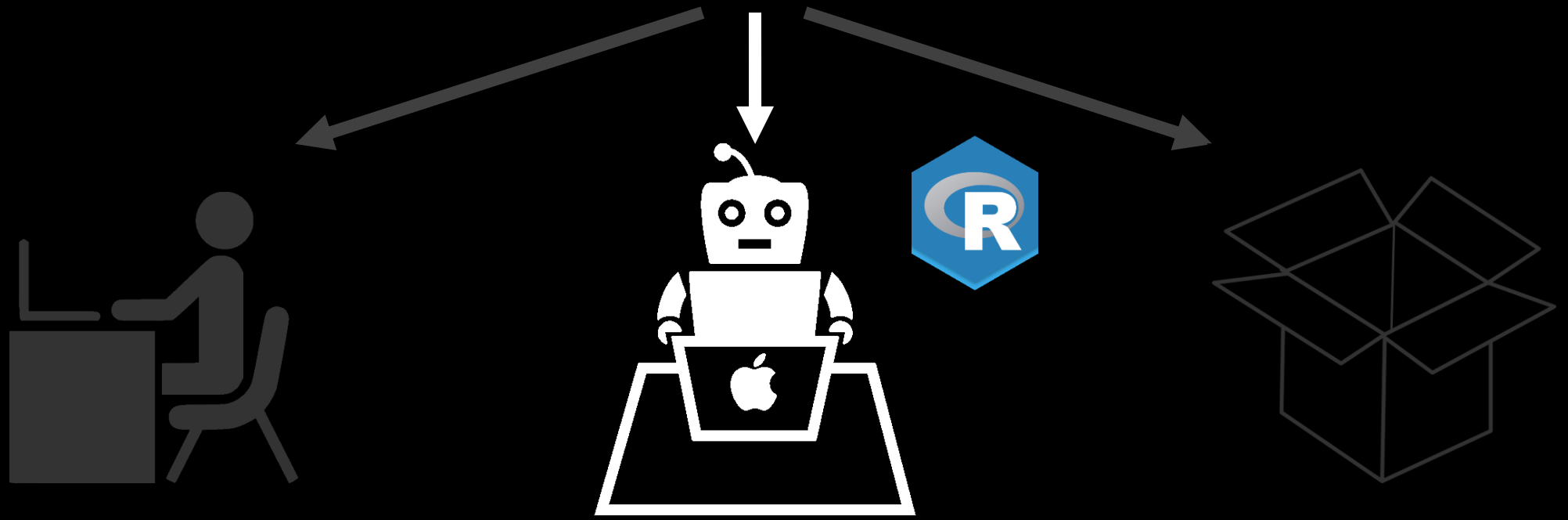
Getting data from the Web automatically

Scraping involves copying information on a webpage and storing it in another format for later use

- Basically mimics what a person might do with copy/paste

Application Programming Interfaces (APIs) provide direct access to data that is "prepackaged" for use by others

Getting data from the Web



Copy/Paste

Scrape

API

At a very minimum, our scraper needs to know...

1. Where to find the information, i.e. the website or domain we want to scrape
2. How to identify the relevant information on the page (or pages)
3. What to do with that information, i.e. where/how to store it

At a very minimum, our scraper needs to know...

1. Where to find the information, i.e. the website or domain we want to scrape
2. **How to identify the relevant information on the page**
3. What to do with that information

the hard part