Week 12 | Hard-to-get data

What do you do when data doesn’t come with a download button?

- Scrape it! (from websites)
- Crack it! (pdfs)

Web scraping

A script that allows you to automatically collect useful data from a webpage.

Scraping with Google Spreadhseets

Two functions will help you extract the data you want from a webpage.

=IMPORTHTML()

=IMPORTXML()

=IMPORTHTML(URL, ”table/list”, 1)

Can look for either a table or a list tag.

Number dictates which table or list you're talking about (useful when there's more than one!).

=IMPORTXML(URL, ”//div”)

This is a more powerful tool. Goes beyond tables and lists!

Pass in the URL and a XPath query to run on the webpage.

XPath

A query language for selecting nodes from an XML (or HTML) document.

XPath nodes = elements or attributes.

Tabula

For getting tables out of PDFs.

http://tabula.technology/

BREAK!

Today’s exercises:

http://bit.ly/week-12-exercises