A Data Scraper's Toolbox

Alice Corona

Data Journalist &
Researcher

data-journalism-tools.silk.co

@alice_corona

alice-corona.nl

?

Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

“process of automatically collecting information from the Web”

~ Wikipedia on “Web Scraping”

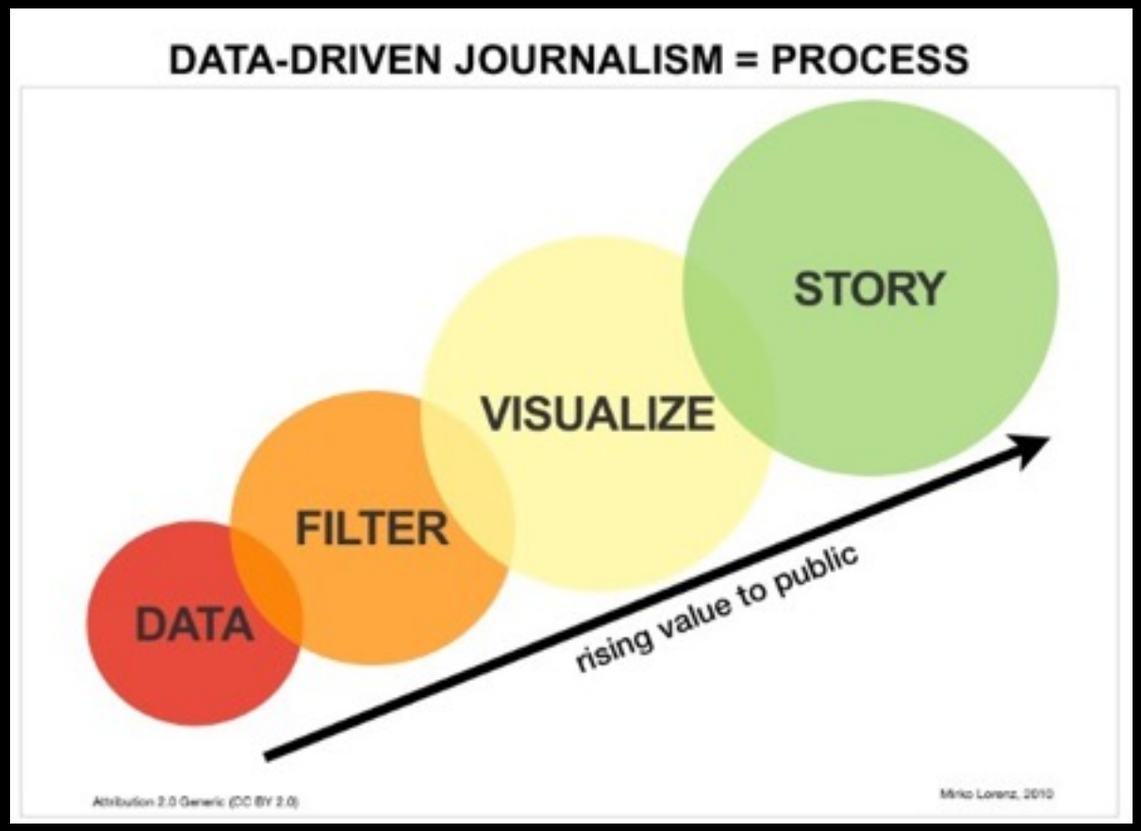
Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

- Manual copy-paste
- Text grepping / RegEx
- HTML parsing (CSS, XPATH…)
- Web Scraping Software

Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

Despite its fame, scraping is a highly creative process. It requires…

- Problem solving
- Lateral thinking
- Planning for efficiency
- Prediction and planning for possible outcomes and uses
- Methodological choices affect the whole outcome



Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

There cannot be a course on web scraping teaching you ‘solutions’

We can dive into the fine art of scraping by learning tools and techniques. Then each **individual** personality can adapt, mix and mash this knowledge according to each **individual** problem

Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

Copy-Paste

Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

Data Import in Google Sheets

Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

Chrome's *Scrape-Similar*

Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

X-Path Parsing in Google Sheets

Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

SearchLink

"SearchLink is a System Service for OS X which handles searching multiple sources and automatically generating Markdown links for text"

http://brettterpstra.com/projects/searchlink/

APIs

Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

Scripts

Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

Scraping Software

Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

“Hidden Secrets”

Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

is scraping legal?

Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

is scraping ethical?

Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

Consider…

- Brute force attacks and DoS (prefer APIs, when available)
- Your use of the data (copyright, privacy…)

Credits: Toolbox icons created by Ralf Schmitzer from the Noun Project

A Data Scraper's Toolbox

Alice Corona

Data Journalist &
Researcher

data-journalism-tools.silk.co

@alice_corona

alice-corona.nl