

**Webscraper by  
Moop Studios**

# Webscrapping Process

1. Visual inspection: Figure out what to extract
2. Make HTTP request to page
3. Parse HTTP response
4. Persist/Utilize the relevant data

# Obstacles

- Authentication
  - Hidden values (Username/Password)
  - Setting headers
  - Possible HTTP status codes: 401, 403, 407
- Server blacklisting
  - Request rate analyzation
  - Header inspection
  - Honeypots
  - Pattern detection
- Redirects/Captchas
- Infinite scrolling (Facebook, Instagram, etc.)
- 

## Links

Google query

syntax: <https://stackoverflow.com/questions/38619478/google-search-web-scraping-with-python>

Elite proxy list: <http://www.gatherproxy.com/proxylist/anonymity/?t=Elite>

## Overall process



