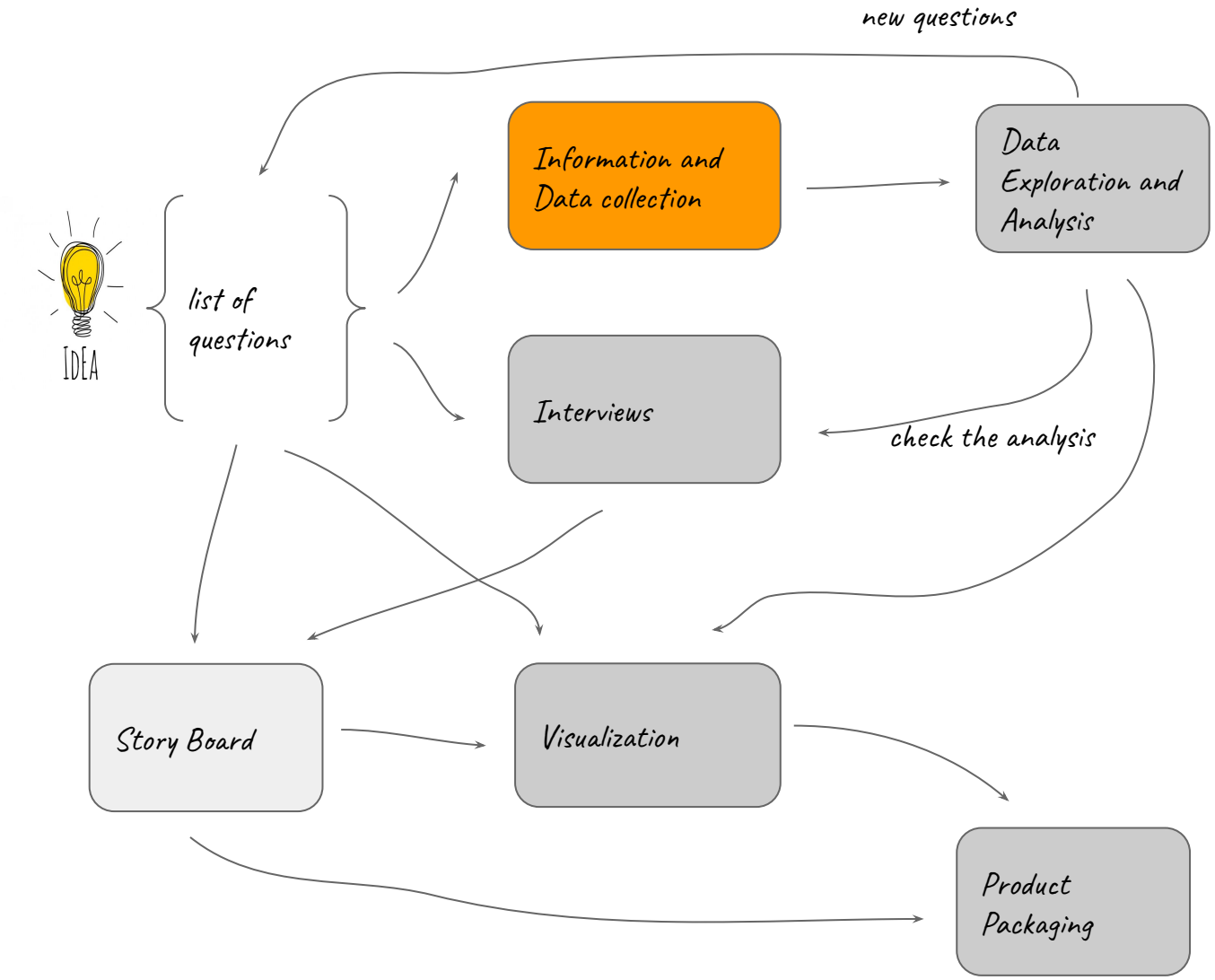
Data Journalism

Data Collection - Web Scraping

InfoUma 2019-20 ***Andrea Marchetti***

Data Journalism Workflow



Data Collection

1. **Open Data**
2. **Web Scraping**
3. Web Crawling through API
4. Scraping from PDF

Open Data Sources

- **User generated data**: Wikipedia, DBPedia, Wikidata
- **Open Government Data**: data.gov, data.gov.uk, data.gov.it
- **Statistics Institutes**: istat, eurostat, worldbank, oecd
- **Open Data aggregators**: awesome dataset, cooldataset
- **Open Data search engines:** Google Data Search, Google trends

Web Scraping

Il **web scraping** (detto anche **web harvesting** o **web data extraction**) è una tecnica di estrazione di dati da un sito web per mezzo di programmi software che appartengono alla famiglia dei **bot.**

Un esempio di web scraping è strettamente correlato **all'indicizzazione** dei siti Internet effettuato dai motori di ricerca - crawler

Il web scraping si concentra nell’estrarre **dati non strutturati** presenti nella pagine HTML e immagazzinarli in **database**

Principali siti attaccati

- agenzie immobiliari (immobiliare.it)
- agenzie di viaggio (booking.com)
- commercio elettronico (amazon.it)
- motori di ricerca (google.com)
- siti di scommesse (bet.com)

...

Eterna lotta tra web scraper e web master

Metodi per prevenire il web scraping

Metodi per evitare di essere “banned”

- Utilizzare Robots Exclusion Standard (Googlebot è un esempio) per bloccare i bot che dichiarano la loro identità (a volte lo fanno usando stringhe degli user agent). I bot che non dichiarano la loro identità non permettono di essere distinti da un essere umano.
- Monitorare l'eccesso di traffico.
- Utilizzare tool come CAPTCHA che permettono di verificare se è stata una persona reale ad accedere ad un sito web. Se questo non fosse vero si tratterebbe quindi di un bot e CAPTCHA lo bloccherebbe. A volte però i bot sono codificati in modo tale da bloccare CAPTCHA o utilizzare servizi di terze parti che sfruttano il lavoro umano per leggere e rispondere in tempo reale alle sfide di CAPTCHA.
- Individuare i bot tramite gli honeypot o attraverso un altro metodo di identificazione di indirizzi IP dei crawler automatici.
- Aggiungere piccole variazioni di HTML/CSS per circondare dati importanti ed elementi di navigazione. Facendo ciò sarà necessario richiedere maggior coinvolgimento umano per la configurazione iniziale di un bot, questo perché essi si affidano alla consistenza del codice front-end del sito di destinazione. Se eseguito in maniera corretta si potrebbe rendere il sito web di destinazione troppo difficile da "raschiare" a causa della ridotta capacità di automatizzazione del processo di web scraping.

Eterna lotta tra web scraper e web master

Metodi per prevenire il web scraping	Metodi per continuare a fare web scraping
Utilizzare Robots Exclusion Standard (Googlebot è un esempio) per bloccare i bot che dichiarano la loro identità (a volte lo fanno usando stringhe degli user agent)	Impostare un user agent accettato
Monitorare l'eccesso di traffico.	Inserire dei ritardi magari random nel codice
Utilizzare tool come CAPTCHA	Utilizzare servizi di terze parti che manualmente risolvono i quiz esposti
Aggiungere piccole variazioni di HTML/CSS per circondare dati importanti ed elementi di navigazione	Usare espressioni regolari per bypassare l'HTML

Aspetti legali ed economici

Ubiquity and danger: The web scraping economy
(31/08/2016)

“**If your content can be viewed on the web, it can be scraped**,” said Rami Essaid, CEO of Distil Networks

58.000$ annui la paga media di un web scraper ma può raggiungere i 128.000$

Metodi di Web Scraping

- Alcuni siti web sono popolati con chiamate web api a Database e i risultati sono wrapped in html
  - Sfruttare le web api e ottenere i dati in json o xml
  - Tecniche di de-wrapping
- **Dom Parsing**
- **Text Pattern matching**
- Computer Vision + Machine Learning

★ in ordine di complessità

Web Scraping Tools Bibliography

The 10 Best Data Scraping Tools and Web Scraping Tools (31/12/2019)

Top 7 Python Web Scraping Tolls for Data Scientists (12/11/2019)

5 Tasty Python Web Scraping Libraries

Selenium Vs Scrapy (15/12/2018)

Web Scraping Tools

- Librerie
  - Beautifulsoap (python)
  - Scrapy (python)
  - Selenium (python)
  - Puppeteer (javascript)
- Browser Extensions (semplici)
  - Web Scraper
  - Grepsr
- Servizi a pagamento (puntano sulla semplicità a sull'efficienza)
  - Scraper api
  - ScrapeSimple
  - Octoparse
  - ParseHub