TALL
SOM FORTELLER

Web scraping in the Norwegian CPI

DATA FOR NOW WORKSHOP

7 APRIL 2022

GUNNAR LARSSON

Statistisk sentralbyrå
Statistics Norway

Web scraping

- The extraction and transformation of unstructured data from the web into structured data

- Scraping tools – external software tools online, import.io, in-house solutions based on software programs like R, Python

- Increase of efficiency and productivity
  - Collecting more data at the same web page/online store

- Shift of tasks: less resources needed for the collection of data, more resources needed for the maintenance of programs and analysis of retrieved data

Web scraping II

- Many countries develop their own technical solution
  - Increased control compared to external software
  - More and more NSOs make their own solutions in R or Python

- Ethics of scraping
  - Minimization of harms
  - Check website's Terms of Use

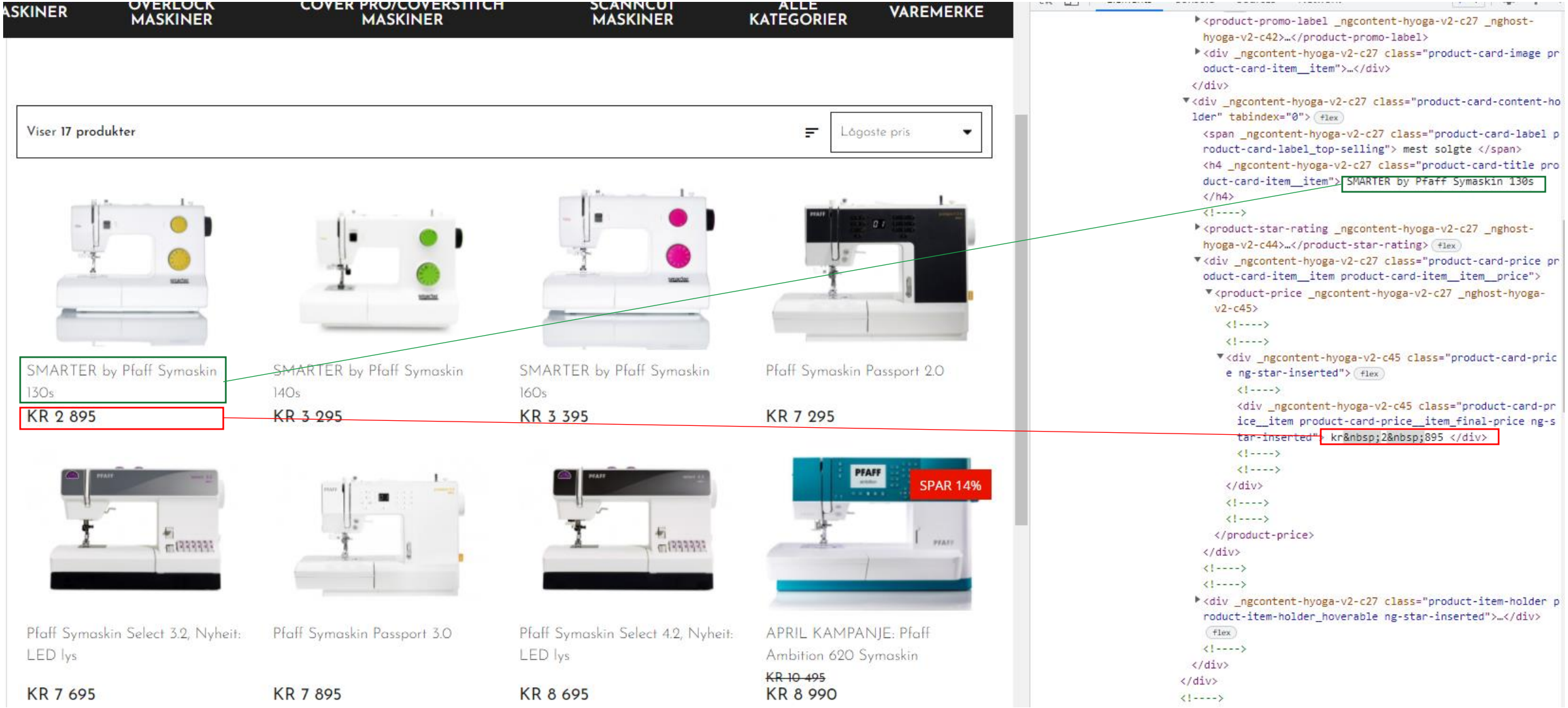
Generalized web scraper

- Able to web scrape a range of online stores covering multiple areas for the consumer price index
  - Gone from specific tool for each web site to generalized approach
- Built using Python, output stored in Excel for further use
  - Data must often be adjusted and cleaned before use
  - Mainly used to compile price indices for the CPI

Pros and cons

- Potential for improved price indices
  - Larger sample
  - Extended time dimension
- Challenging to find common traits among different web sites
  - Requires testing the web scraping tool on different types of web sites
- Web sites change over time
  - Can lead to maintenance

Demonstration



TALL SOM FORTELLER

Thank you!

Statistisk sentralbyrå
Statistics Norway