# Web scraping in the Norwegian CPI

DATA FOR NOW WORKSHOP

7 APRIL 2022

**GUNNAR LARSSON** 



## 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

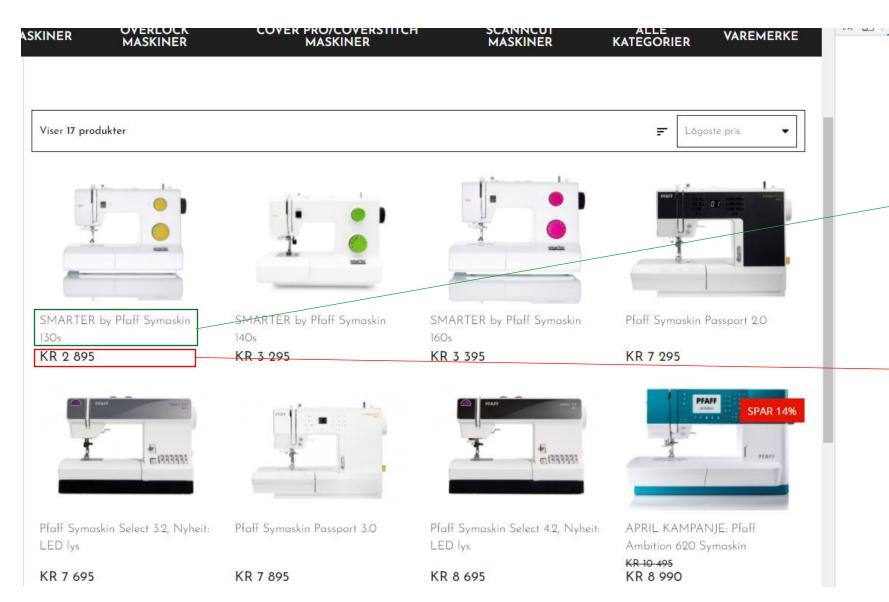


#### Pros and cons

- 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**



```
typroduct-promo-label _ngcontent-hyoga-v2-c27 _nghost-
   hyoga-v2-c42>...</product-promo-label>
  <div ngcontent-hyoga-v2-c27 class="product-card-image pr</p>
   oduct-card-item item">...</div>
  </div>
▼ <div _ngcontent-hyoga-v2-c27 class="product-card-content-ho
 lder" tabindex="0"> flex
   <span _ngcontent-hyoga-v2-c27 class="product-card-label p</pre>
   roduct-card-label_top-selling"> mest solgte </span>
   <h4 _ngcontent-hyoga-v2-c27 class="product-card-title pro</pre>
   duct-card-item_item"> SMARTER by Pfaff Symaskin 130s
    <!--->
  cproduct-star-rating _ngcontent-hyoga-v2-c27 _nghost-
   hyoga-v2-c44>...</product-star-rating> (+lex)
  ▼ <div _ngcontent-hyoga-v2-c27 class="product-card-price pr
   oduct-card-item_item_product-card-item_item_price">
    ▼▼duct-price _ngcontent-hyoga-v2-c27 _nghost-hyoga-
     v2-c45>
       <!--->
       <!--->
      ▼ <div _ngcontent-hyoga-v2-c45 class="product-card-pric
       e ng-star-inserted"> (flex)
         <!--->
         <div _ngcontent-hyoga-v2-c45 class="product-card-pr</pre>
         ice__item product-card-price__item_final-price ng-s
         tar-inserted"> kr  2  895 </div>
         <!--->
         <!--->
       </div>
       <!--->
       <!--->
     </product-price>
   </div>
   <!--->
  div _ngcontent-hyoga-v2-c27 class="product-item-holder p
   roduct-item-holder_hoverable ng-star-inserted">...</div>
   <!--->
 </div>
</div>
<!--->
```

## Thank you!

