CHALLENGES AND SOLUTIONS TO THE USE OF INTERNET DATA IN THE DUTCH CPI

Robert Griffioen, Olav ten Bosch and Els Hoogteijling

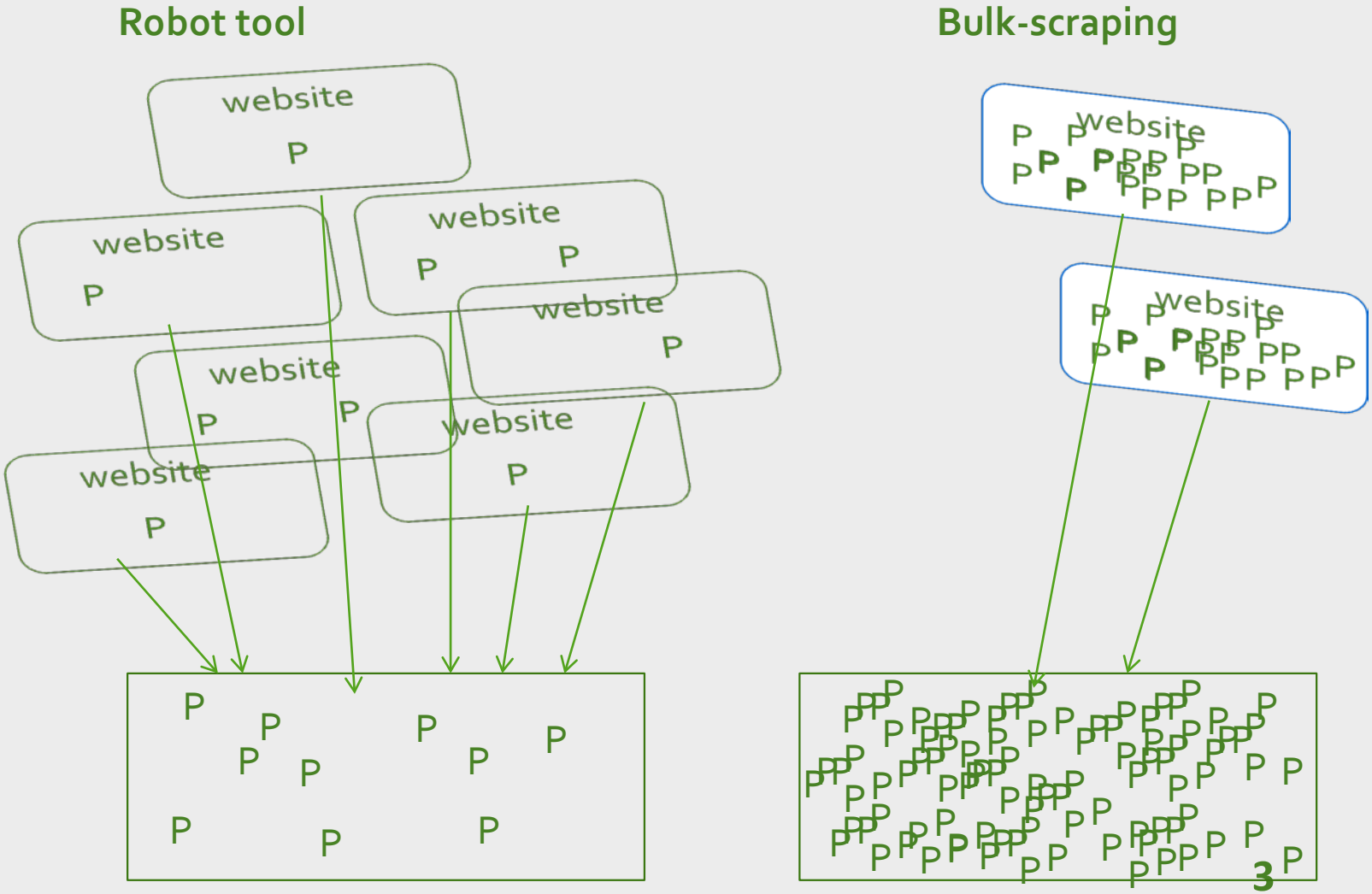
Workshop on Statistical Data Collection, The Hague, The Netherlands, 3-5 October 2016

Statistics Netherlands

Overview

- Introduction CPI and internet data collection
- Core value of using internet data in the CPI
- Technical, organizational and legal challenges of using internet data in the CPI
- Other challenges
- Conclusions and discussions

Introduction: a two tier approach of web scraping

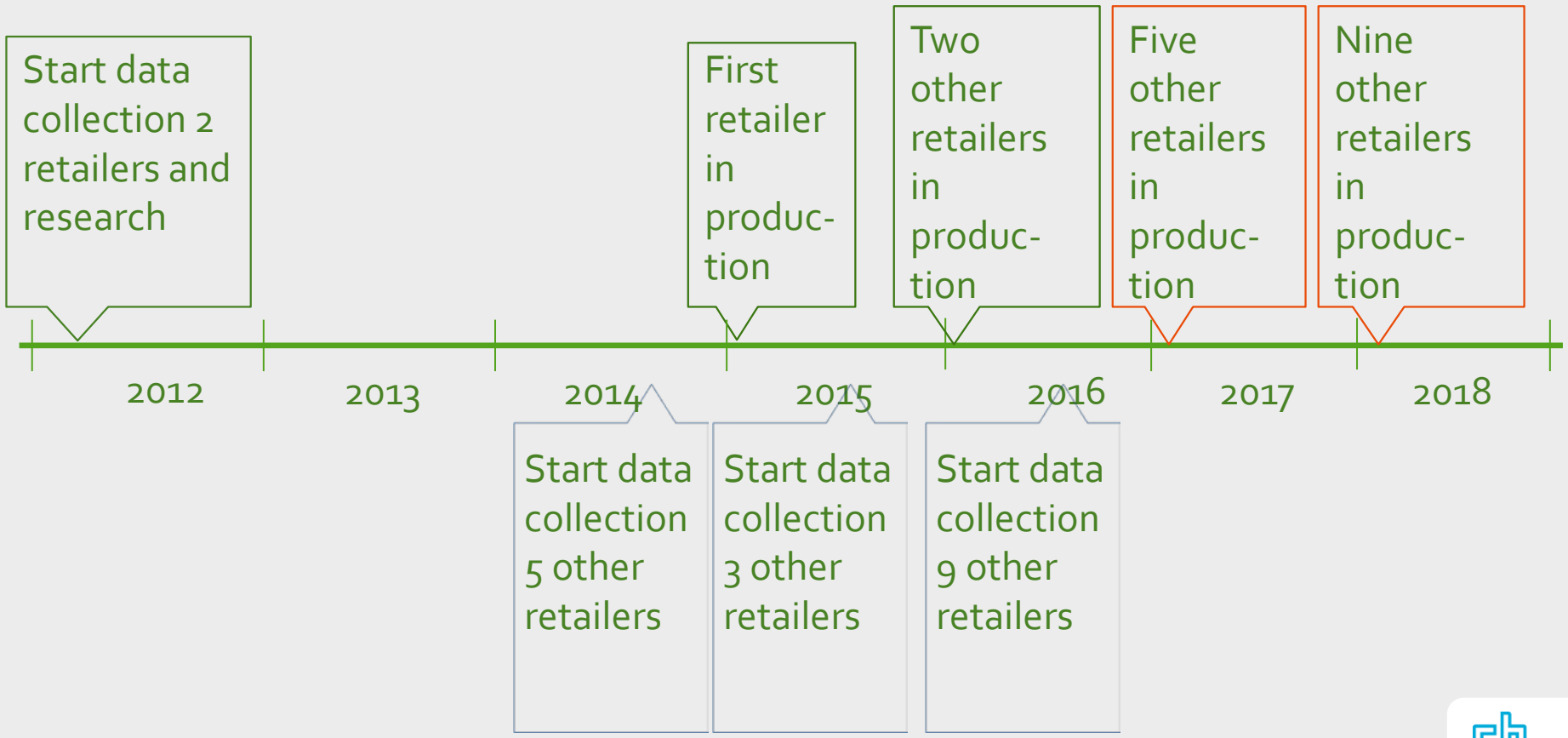


Introduction: history and plans robot tool

Robot tool: started at 2012, 2014 in production

ECOICOP	
31410	Laundry and dry cleaner's (service)
62390	Paramedical services (like acupuncturist, chiropractors and physiotherapist
72300	Services for maintenance and repair of private cars
72430	Driver license lessons, driver licence examinations and cost of a driver license.
73220	Cost of a taxi
94120	Cost of participation in recreation and sports
94210	Cinema, theatre and concerts
94220	Museums, libraries and zoos
94250	Photography services (for instance printing)
111110	Catering services
111120	Accommodations

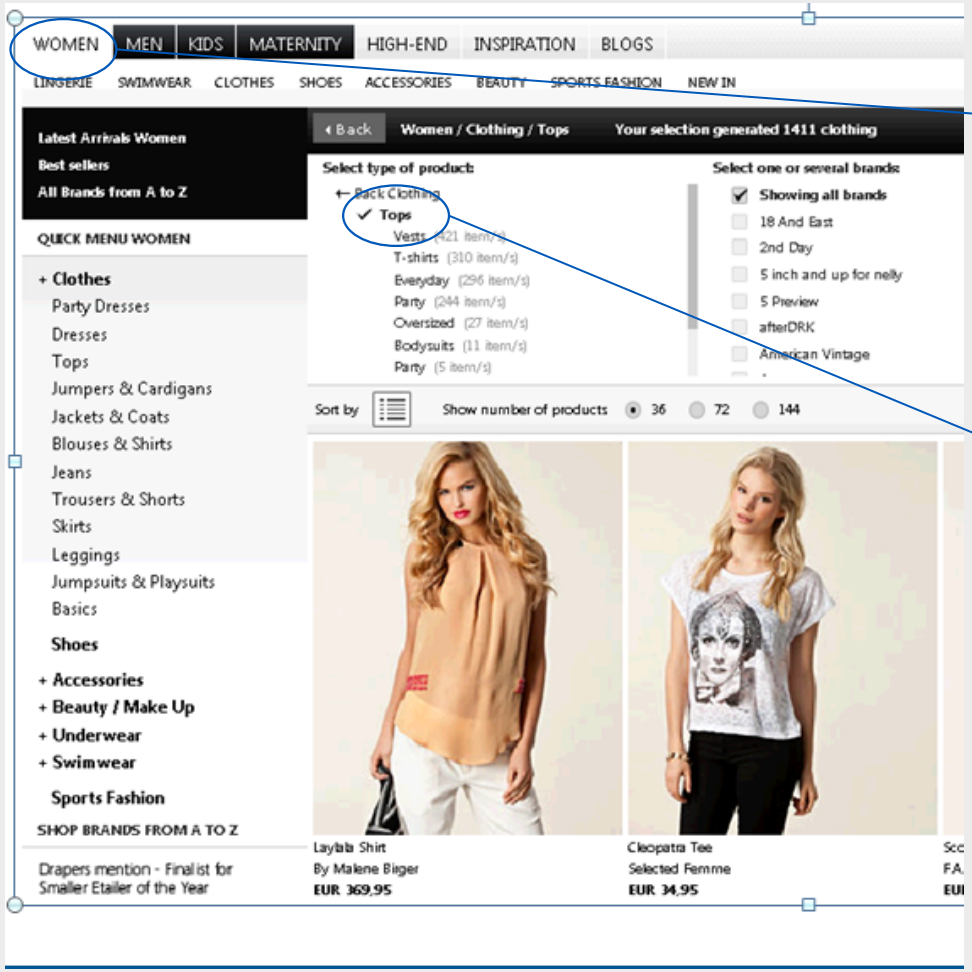
Introduction: history and plans bulk scraping



Core value internet data collection

- Internet is a very important medium in our daily lifes:
  - Communication
  - Information/consulting
  - Entertainment
  - Education
  - Shopping
  - Et cetera … internet is becoming a natural part of our daily lifes
  - → If official statistics does not use it, we may develop a bias towards traditional channels
- Cost reduction

Core value internet data collection: good quality

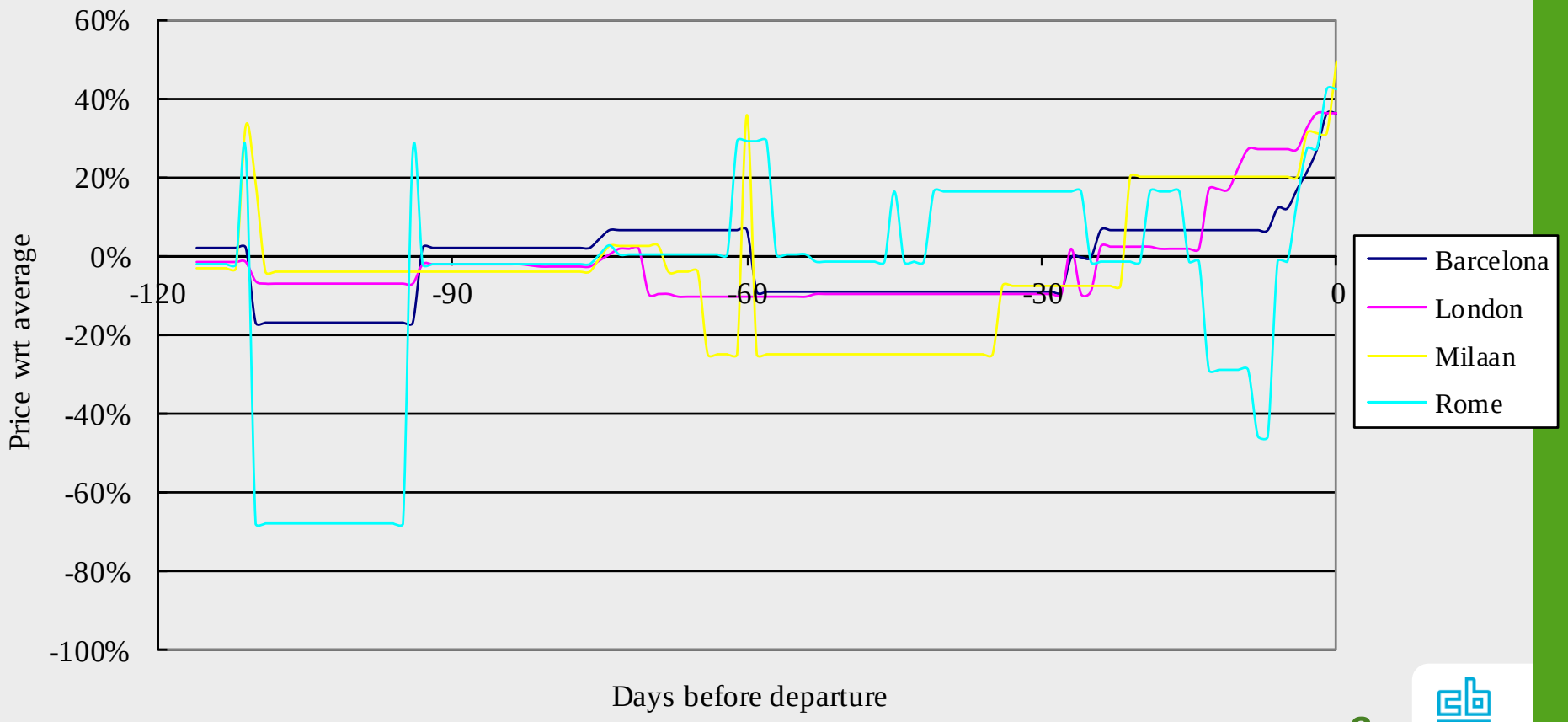


Text classification

- L-Coicop
  - Women
  - Men
  - Kids
- Clothing type
  - Jeans
  - Trousers
  - Top
  - Sweaters
  - …
- Classification rule:

“sweater” and not “jack”

Core value internet data collection: better insight in data



Challenges & solutions: organization

Team	Robot team - Research - IT	CPI team
Process	- Web scraper Development and Maintenance - Data Monitoring	Data monitoring / validation
	Monitoring • File size • Number of items	Checks: • Number of items after deduplication

Challenges & solutions: organisation

Team

DevOps
- IT
- CPI
- Research

Process

- Web scraper Development and Maintenance
- Data Monitoring

Monitoring
- File size
- Number of items

Data monitoring / validation

Checks:
- Number of items after deduplication

Challenges and solution: new skills and knowledge

- IT-skills for web scraping and data-analyses
- Machine learning skills and knowledge to process big data

- New employees
- Acquisition of new knowledge and skills
  - External courses
  - Internal knowledge transfer

Challenges & solutions: other

- Mental: shared vision, a common urgency, enthusiasm and courage
- Legal issue

Conclusions & discussion

- Internet is (becoming) a natural part of our daily life → Not using internet as a data source in official statistics might not only cause a bias in data sources but also in statistical publications.
- The business case for internet data will probably grow stronger over the years, as experience, knowledge and tools can also apply to other statistics.
- New opportunity's