STATISTIEK
VLAANDEREN

A better statistic on innovative companies in Flanders using web scraping and machine learning

Vlaanderen
is sterk in cijfers

Presented by: Michael Reusens
Data Science Coordinator @ Statistics Flanders, Belgium
michael.reusens@vlaanderen.be

Collaboration

- Replication of work done by Piet Daas et al. at CBS, Netherlands (https://www.cbs.nl/nl-nl/over-ons/innovatie/project/innovatieve-hotspots)
- Collaboration with other NSI's is kickstarting our data science for official statistics journey
- Work presented is being done by a student as his master thesis and internship

Current innovative companies statistic

- Survey based (CIS survey)
  - Low frequency (every 2 years)
  - Information lag due to data processing
  - Small sample size
  - Only includes companies >10 employees
  - Survey burden

New approach

- Scrape homepages of websites Flemish companies
- Train a machine learning model to classify homepage as innovative/not innovative
  - Using CIS survey results as training data
- Aggregate into a statistic

New approach

- Web scraping + text classification
  - High frequency
  - Little information lag
  - Large coverage
  - No survey burden
  - --> Same data can be used to further analyse innovation in Flanders

Challenges

- Will the work done in the Netherlands be nicely replicable in Flanders?
- Technical challenges
  - Create a complete list of all Flemish companies' websites
  - Limitation of homepage-info and heterogeneity of website structures
  - Multiple languages: Dutch, English, French,…
  - …

Thank you for your attention

- Contact: michael.reusens@vlaanderen.be