Center for Big Data Statistics

Big Data and Official Statistics:

Applications and how to use Machine Learning

Piet Daas, Statistics Netherlands &
Eindhoven University of Technology

Towards smart statistics

Main focus of this presentation

- Big Data in Official Statistics
  - How to use ML in Official Statistics?
    - Use of website texts for official statistics production
  - Applications of Big Data for official statistics
    - A number of successes

Using web sites text to detect Innovative Companies

Detecting innovation

- Web pages of companies provide information
  - The pages can be collected fairly easy
  - The text can be extracted fairly easy

- Here we look at:
  - The potential of *web pages* to provide information on the *innovative* character of a company
    - *Can text be used to detect innovation*?
  - For both *large* and *small* companies

Center for Big Data Statistics

Start Inductively !

The Community Innovation Survey

- The Community Innovation Survey (CIS)
  - Focusses on the innovativeness of companies
  - Is a European standardized survey
  - Send to 10,000 companies (with 10 or more Working Persons)
    - *No info on small companies (such as start-ups!)*
  - Need link to website (URL) for each company
    - Used URLsearch and lots and lots of manual checking

- This is a representative dataset for large (WP >= 10) companies
- The link between de unit and its website is completely checked

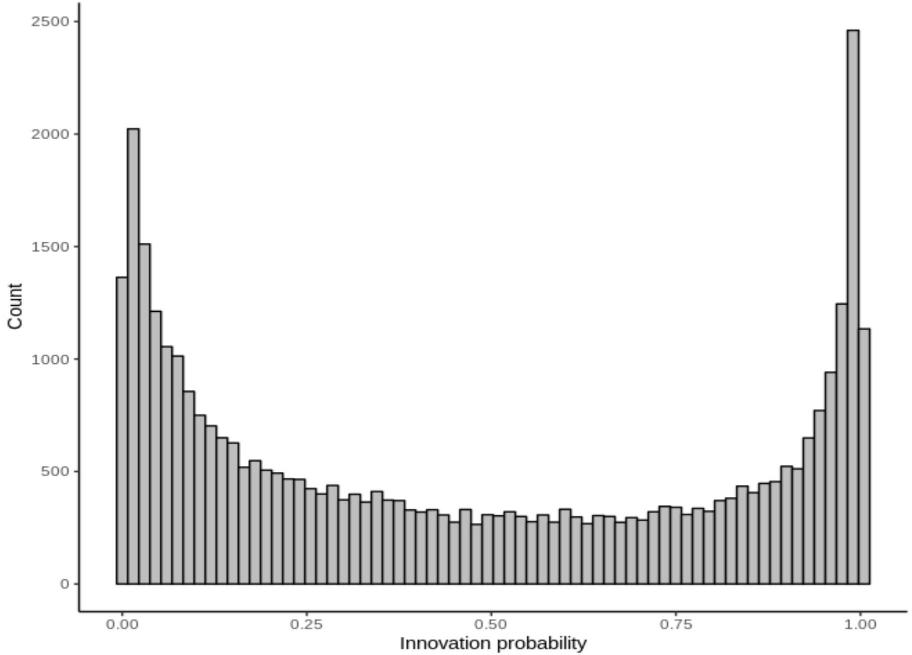
Validate model (internal)

- Created ML-based model with an accuracy of 88%
  - Trained on a 80% random sample, tested on remaining 20%
  - Ratio pos. and neg. examples comparable with CIS-survey
  - Only used web site texts with 10 or more words
  - Removed stop words, numbers, punctuation marks and words <= 2 characters, several other words later on

- Positive words with high weights in model:
  - com, system, inspiration, data, technology, *do*, analysis (and Language)
- Negative words very divers:
  - sale, buy, *create*, powered, shape, exclusive, …

Validate model (external)

- Tested model on large amounts of large company websites (not included in training & test set)
  - Around 37,500 companies
  - Used **chance of being innovative**



Hypothesis

- Can the model be used to detect small innovative companies (in the Netherlands)?

-We are forming a theory here!

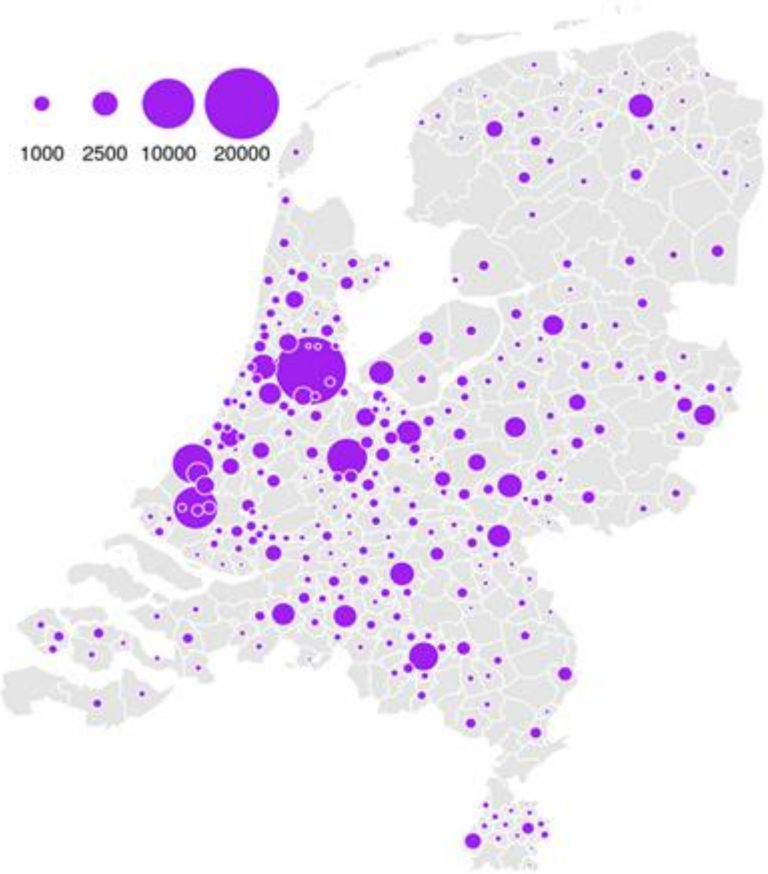
Deductive reasoning kicks in

Validate model (external 2)

- Tested Model on datasets of small companies (WP < 10)
  - Set of 900 scraped web sites of startups
    - 92% innovative
  - Large numbers of websites of small companies included in Dutch Business Registers (with a linked web site)
    - Of these 33% are innovative

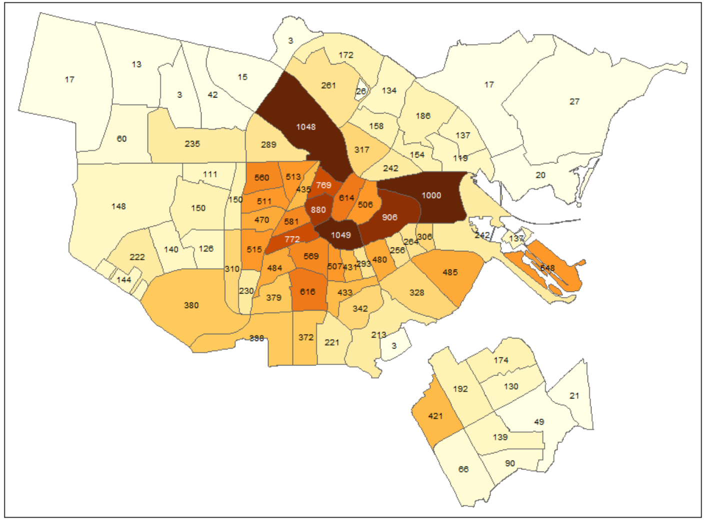
We are **Generalizing** here!

Validate model (external 3)



web sites at municipality and

Amsterdam



Hypothesis

- Can the model be used to estimate the number of **large** and small companies (in the Netherlands)?

Validate model (4)

- Used the classified linked websites (to the Business Register) to estimate the number of large and small innovative companies (~850.000 websites)
- After model bias correction, include an estimation of innov. companies with website < 10 words, and number of innovative companies without a website (0,01%)
  - Number of large innovative companies:

CIS-survey	19.916 ± 680
Web texts	19,276 ± 190

  - Number of small innovative companies:

Web text	33,599 ± 773 (WP 2 - 9)

An issue with texts of WP 0.1 – 1 companies: (semi)-self-employed

Hypothesis

- Can this approach be used to detect innovative companies in other European countries?

Validate model (extern 5)

– CIS survey is a European standardised survey

– Approach tested successfully in:

  - Germany (Kinne en Lenz, 2019) - Paper
  - Flanders (Belgium) (Reusens, 2021) - Intern

– - *Not in Sweden* (?) - Intern

However be aware

- We are measuring the association between the presence of certain (combinations of) words and the classification of innovative and non-innovative companies used in Europe.
  - This relation can be used to detect small innovative companies (WP 2-9)

- The relation depends on how a website is used by a company!

Remember what Marr said:

‘Try to understand the nature of the problem being solved’

Other examples of using Big Data in Official Statistics

2. Detecting Platform economy websites

- Statistics Netherlands wants to produce statistics on platform economy companies
  - "An online platform is a website or application that mediates or supports the exchange of goods, services or information between individuals, companies or organizations. The platform contains either offers from other parties, or, in addition to own offers, also offers from other parties"
  - Examples are: AirBnB, Uber, Amazon, …

- Here we look at:
  - Using the *text on web pages* to detect platform economy websites
  - To *pre-screen* the population of Dutch companies

Model development

- Statistics Netherlands experts provided a set of examples of platform economy websites (680 positive examples)
- Only a few negative examples were given
  - Added a random sample of non-platform econ. from Business Register
  - In the end 50% positive, 50% negative (remember this!)

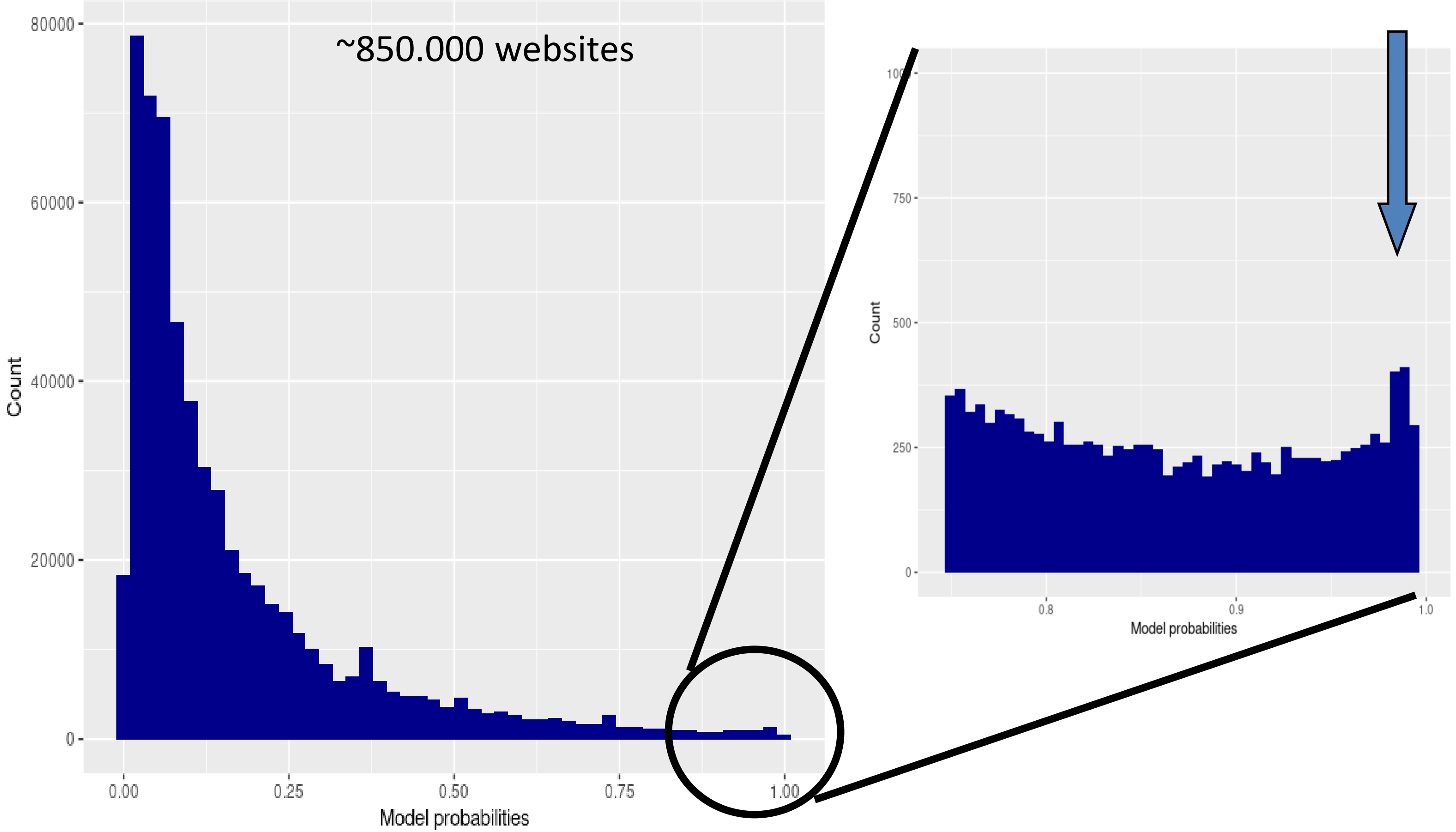
- Is there a difference between the text of platform and non-platform economy websites?
  - Developed a ML-model to detect platform economy websites based on the text
  - Standard preprocessing, combined multiple webpages per website (up to 200)
  - Support Vector Machine model with an accuracy of 82% (best option)

Model evaluation

- Model provided the *chance* of being a platform economy web site
  - Value between 0 and 1
  - U-shape distribution of test set (rel. small set)

- Words positively associated with platform economy:
  - **Register, com, login, platform, invest, sign up, …**
  - Negative associated words are indicative for other type of websites

Applying model to all websites linked to BR



~850.000 websites

Model evaluation (2)

First finding:

- The model indicated 9.802 websites as potential platform economy (all with p > 0.5)
- After manual checking websites with p > 0.8 were found the most interesting. Adult sites and low text were removed.
  - A total of 5.734 websites/4.170 companies remained.
- A total of 3.522 companies received a survey. 2.232 companies responded (63%) of which 537 were identified as platform economy companies.
  - Based on this it was found that platform economy companies all had websites with p >= 0.95 !
- Model has been used again this year to preselect companies
  - - Model was checked and found stable

3. Scanner data

- Many countries increasingly use ‘Big Data’ for the Consumer Price Index (CPI)
  - Main sources in NL: Scanner data and web scraped prices

  - Advantage: large amounts of Scanner data enable new types of analysis
    - To obtain quick insights

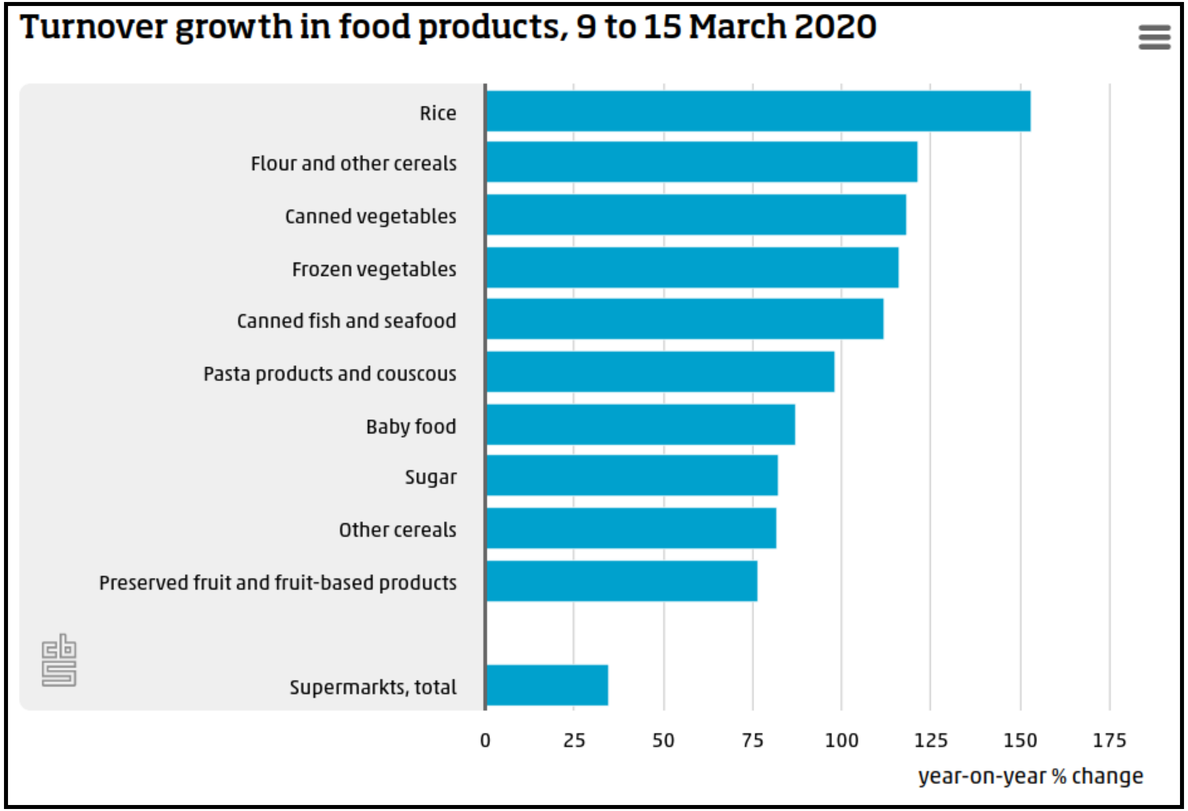
Hoarding during pandemic

Run on:

Rice,

Hand soap,

Toilet paper



Turnover growth in food products, 9 to 15 March 2020

Center for Big Data Statistics

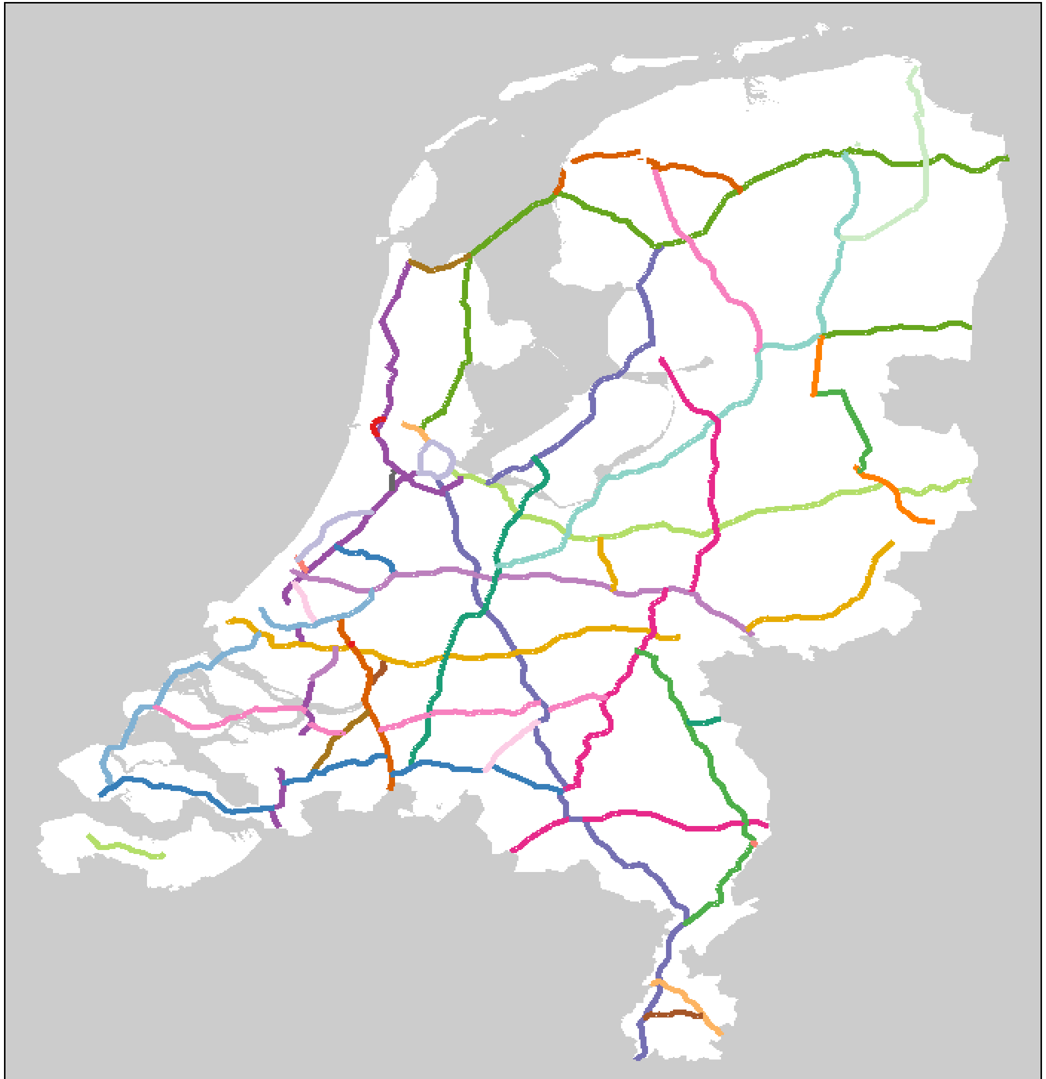
4. Road sensors

Road sensors in the Netherlands

- Passing vehicle counts for each minute (24/7) by about 60.000 sensors
- 20.000 on the Dutch highways
- Types of sensors:
  - Induction loop
  - Camera
  - Bluetooth

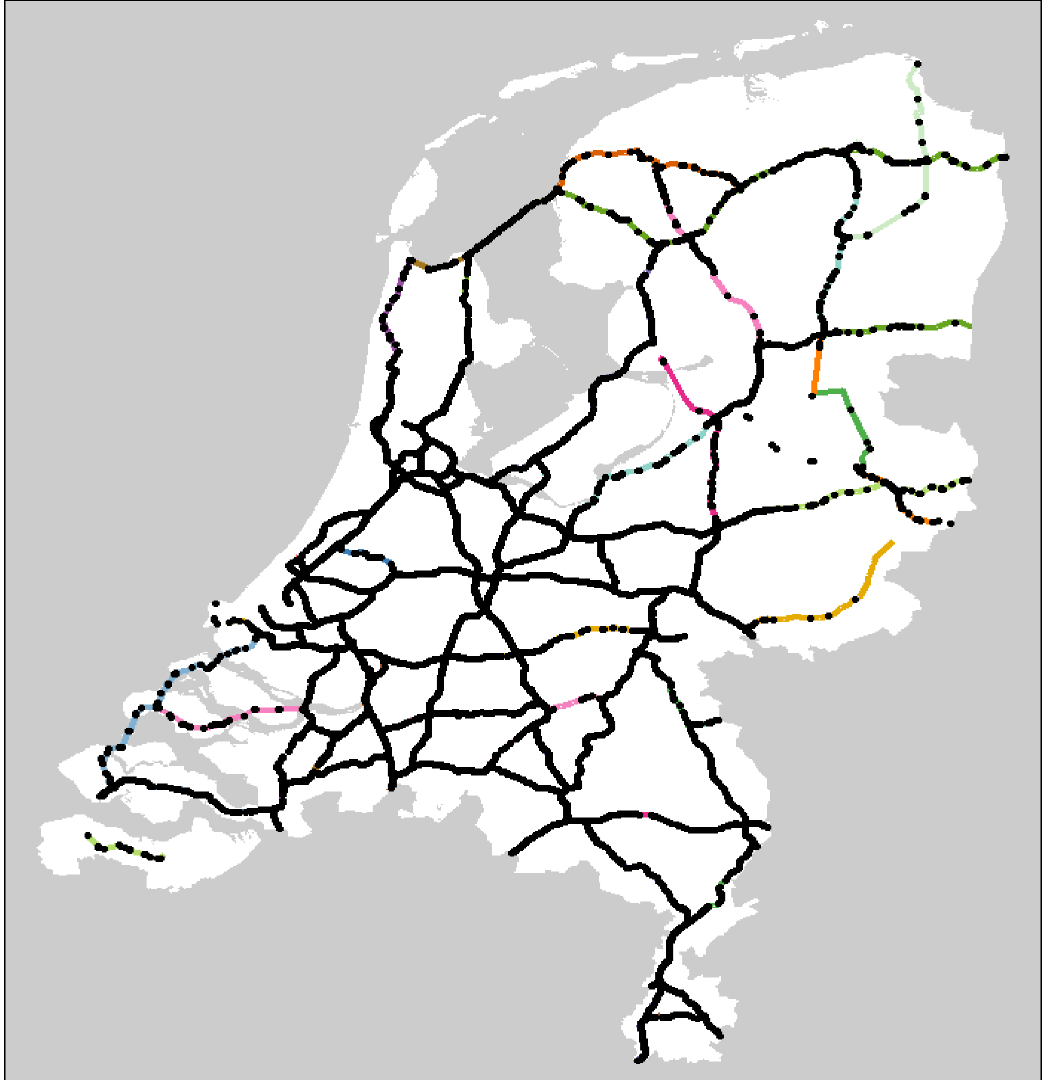
Center for Big Data Statistics

Dutch highways



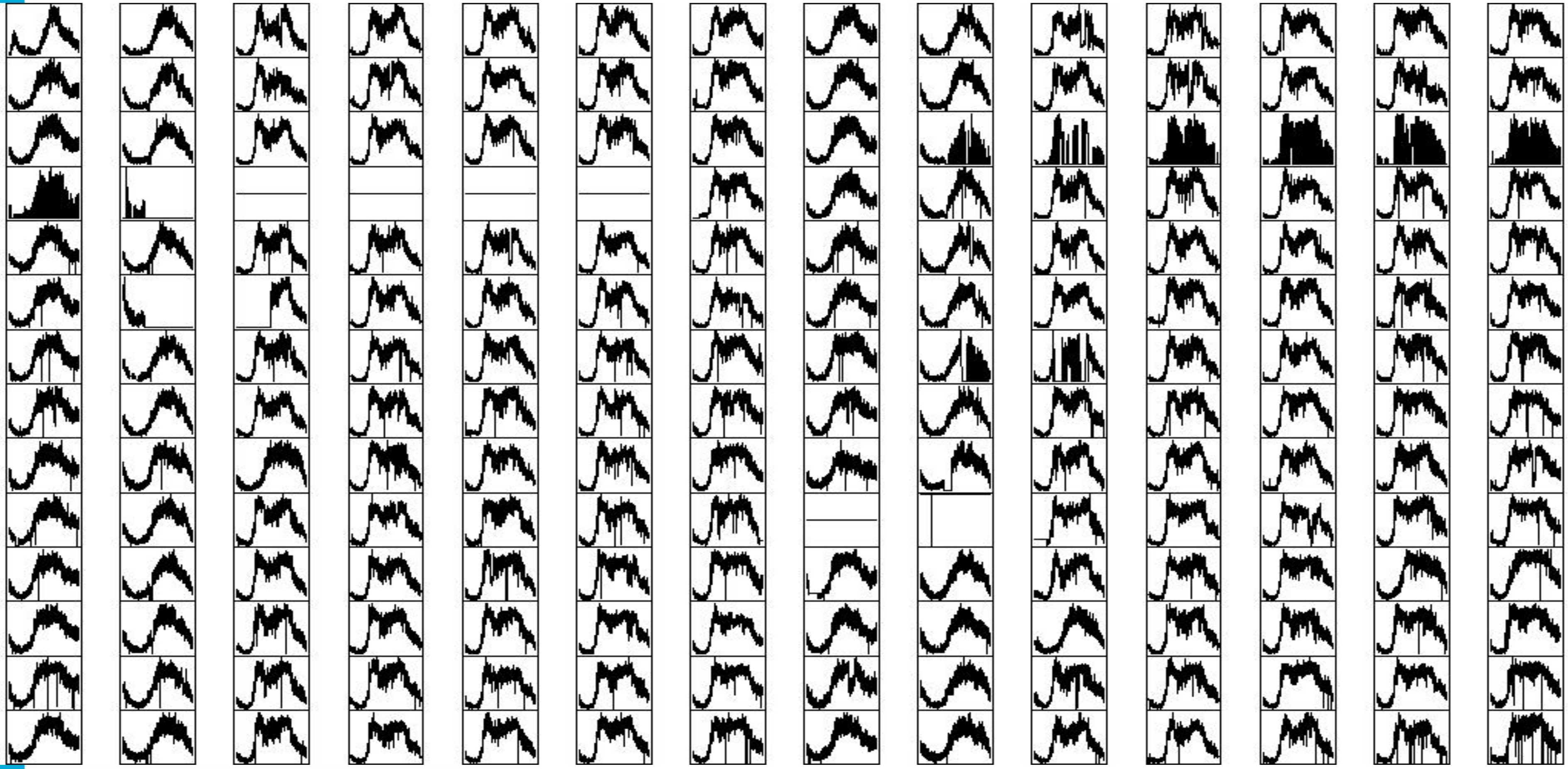
Dutch highways
+ road sensors

20.000 sensors
on highways



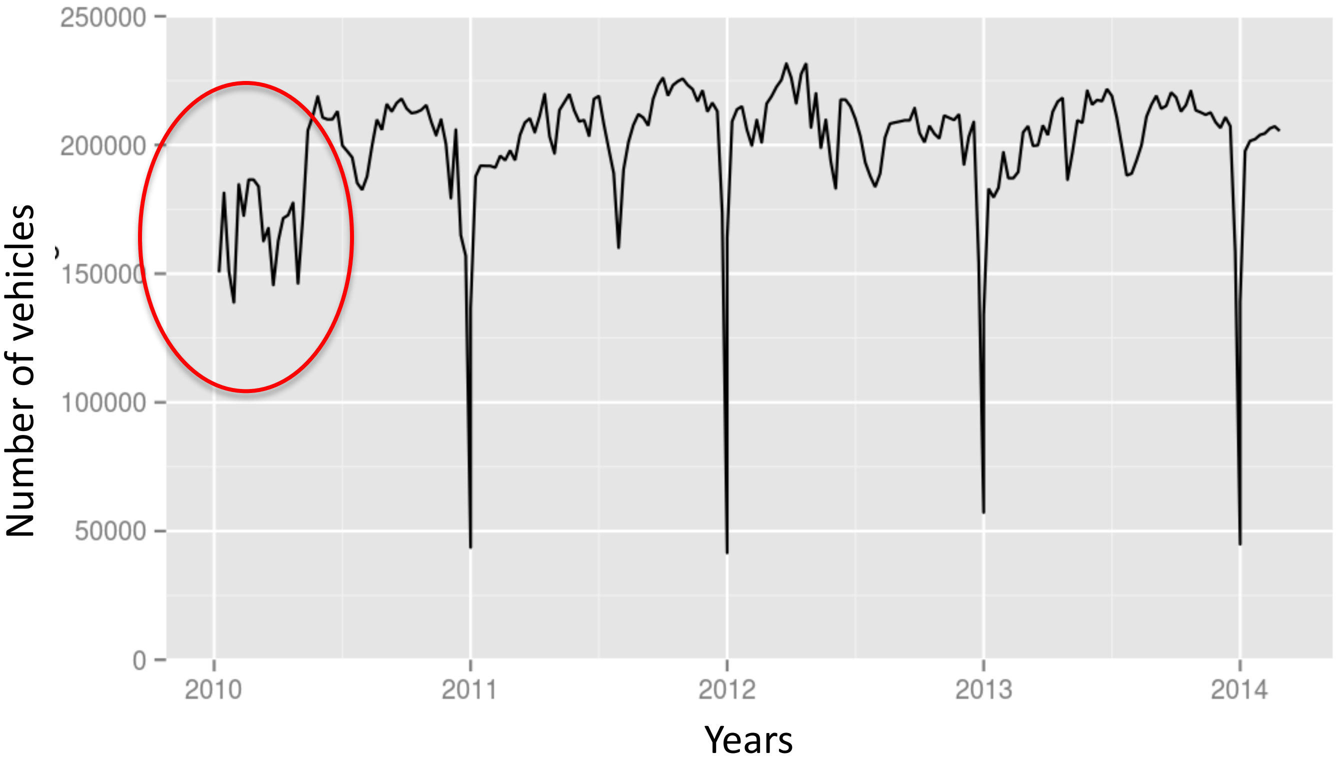
Center for Big Data Statistics

Minute data of 1 sensor for 196 days



Traffic intensity statistics

- Daily number of vehicles on Dutch highways



Center for Bi

5. Social media data

- Social media is a very interesting data source
  - But: only a small part of the population is active
  - We can only use publicly available messages

  - Performed a number of studies, some produced very interesting findings
    - General social media sentiment
    - Social tension indicator*
    - COVID-19 studies*

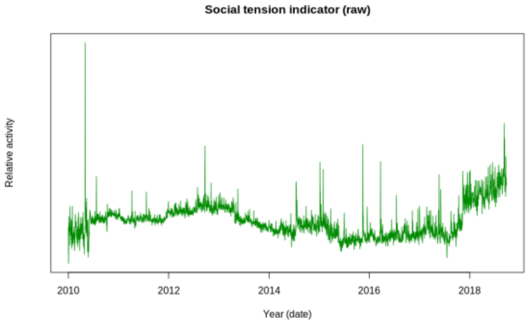
Social tension indicator

- Develop a timely social media based indicator

- How does a fast (real-time) statistic look like?
  - Based on all previous experiences with social media
  - Making use of the typical strengths of social media

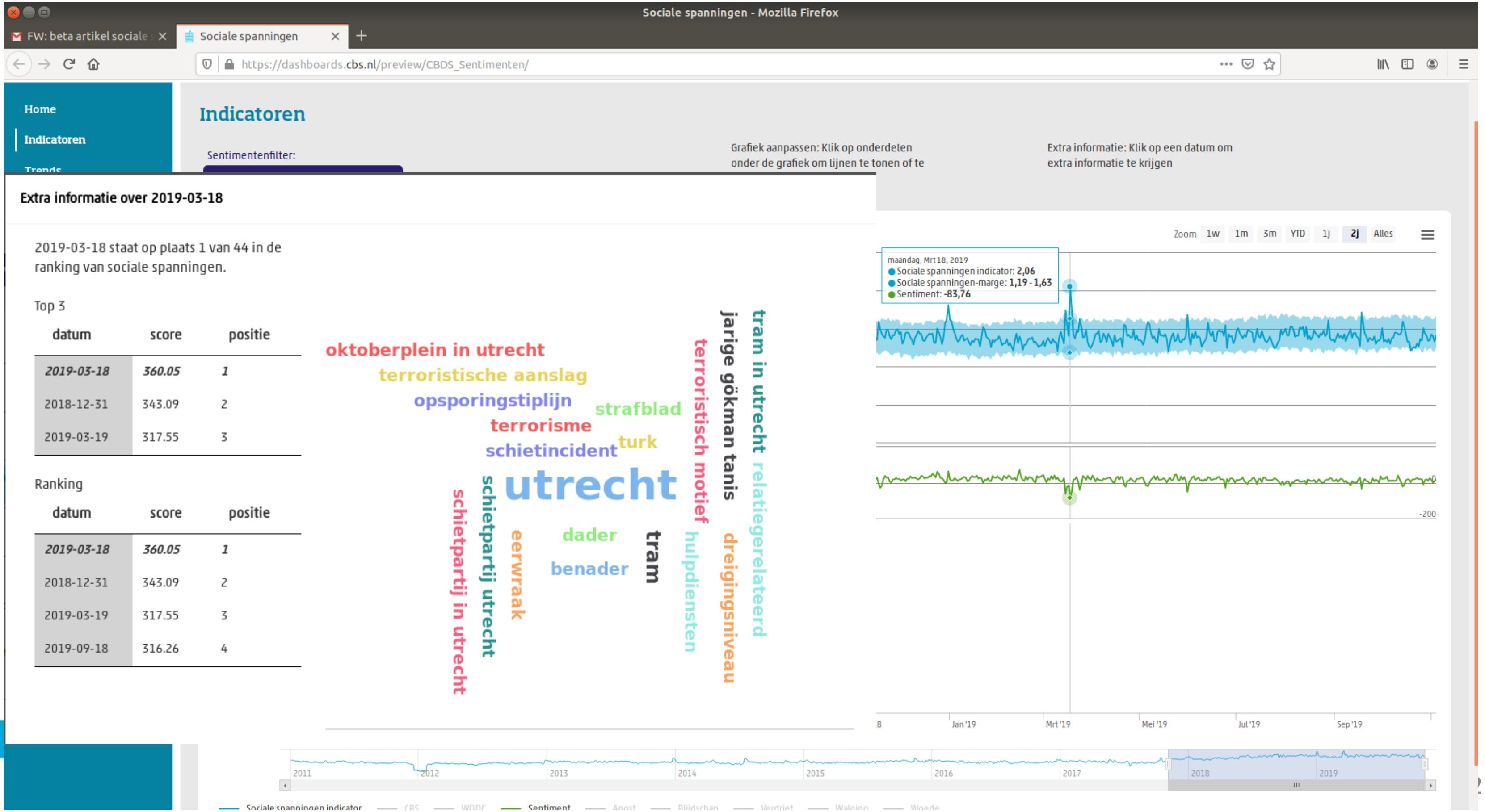
- On Twitter people really want to be the first to report interesting news.
- Started with the idea of a ‘real-time’ safety monitor
  - Can we measure the feeling of safety/unsafety online?

Social tension indicator (2)



- Interviewed people on the type of words use safety/unsafety
  - Ended up with a list of ~350 words
- Checked how often these words were used on Twitter
  - Used Coosto access, a nearly complete Dutch Twitter Database
  - Only ~150 words are used frequently enough online
  - Removed messages of people from Flanders as good as possible
- An interesting profile occurred but:
  - Are we measuring safety/unsafety feeling?
  - Check what the peaks represent → *Social unrest*

Social tension indicator: dashboard



Sociale spanningen - Mozilla Firefox

FW: beta artikel sociale

Sociale spanningen

https://dashboards.cbs.nl/preview/CBDS_Sentimenten/

Home

Indicatoren

Trends

Indicatoren

Sentimentenfilter:

Grafiek aanpassen: Klik op onderdelen onder de grafiek om lijnen te tonen of te

Extra informatie: Klik op een datum om extra informatie te krijgen

Extra informatie over 2019-03-18

2019-03-18 staat op plaats 1 van 44 in de ranking van sociale spanningen.

Top 3

datum	score	positie
2019-03-18	*360.05*	*1*
2018-12-31	343.09	2
2019-03-19	317.55	3

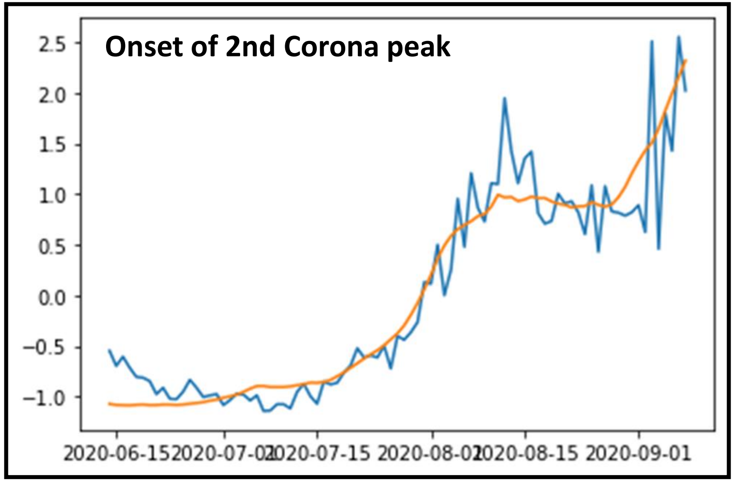
Ranking

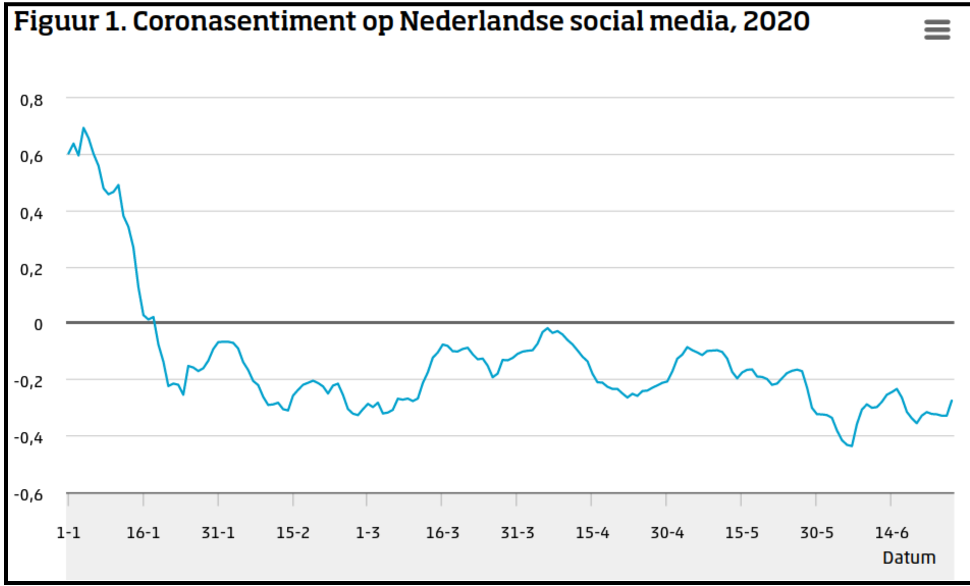
datum	score	positie
2019-03-18	*360.05*	*1*
2018-12-31	343.09	2
2019-03-19	317.55	3
2019-09-18	316.26	4

maandag, Mrt 18, 2019
● Sociale spanningen indicator: **2,06**
● Sociale spanningen-marge: **1,19 - 1,63**
● Sentiment: **-83,76**

COVID related work

- Mainly based on Twitter
  - Daily insight on sentiment towards COVID-19
  - COVID-19 symptoms spread




Figuur 1. Coronasentiment op Nederlandse social media, 2020

https://www.cbs.nl/nl-nl/over-ons/innovatie/project/corona-sentimentsindicator

https://coms.events/NTTS2021/data/sessions/en/session_48.html

6. Overview of Big Data uses and their status

1. **Consumer Price Index (in production, multiple countries)**
2. **Traffic intensities (in production, NL)**
3. *Social Tension indicator (will go into production, NL)*

4. Online job vacancies (towards implementation, ESSnet)
5. Enterprise characteristics (towards implementation, ESSnet)
6. Electricity/energy consumption (towards implementation, ESSnet)
7. Maritime and Inland waterway statistics (towards implementation, ESSnet)
8. Financial transaction based statistics (exploratory, ESSnet)
9. Earth observation derived statistics (towards implementation, ESSnet)
10. Mobile network derived statistics (towards implementation, ESSnet)
11. Innovative tourism statistics (exploratory, ESSnet)
12. Innovative company websites (towards implementation, NL)
13. Social mood on economy index (published experimental, IT)
14. Mobile phone derived outbound tourism (experimental, AU/FI/Estonia)

Conclusions

- Big data is a very interesting data source for statistics
  - Gain new insights
  - Produce fast and more detailed statistics

- Lessons learned:
  - ML is great to extract information from texts and images
  - Make sure to include a statistics view on ML (and all Big Data work)
  - Start early to obtain access to new data sources
  - Prepare for lots of discussions with more ‘traditional statisticians’
  - Communicate with the general public and be aware of privacy

Big Data

Thank you for your attention !!

Center for
Big Data Statistics

Towards **Smart Statistics**

Prof. Dr. Piet J.H. Daas

- Statistics Netherlands (CBS)
  - Senior-methodologist (20+ years)
  - Big Data research leader (since 2011)

- Eindhoven Univ. of Technology
  - Part-time Professor “Big Data in official statistics” (since 2019)

- I focus is on the methods needed to (re-)use ‘already existing’ data for official statistics

Center for Big Data Statistics