Unit 3: Introduction to Big Data

Marco Puts & Piet Daas

Centraal Bureau voor de Statistiek

Overview

- What is Big Data?
- Properties of Big Data
- Diversity of Big Data
- Statistical uses and examples
- General remarks
  - *Questions*

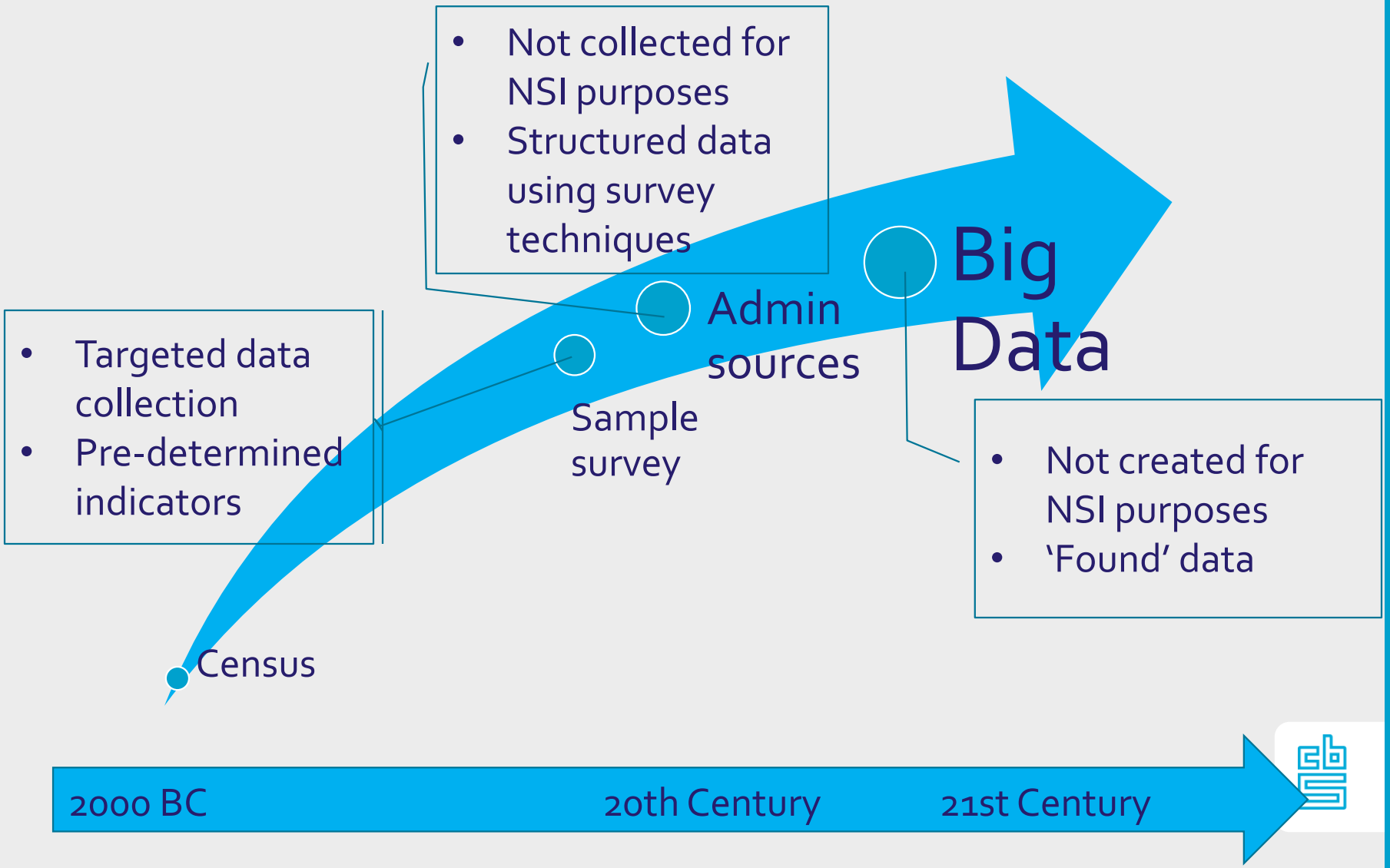
What is Big Data?

- In our modern world, more and more electronic devices are being used that continuously produce data which remains to be stored.
- This results in a data ‘deluge’, hence the term Big Data
- This data may have very interesting potential (statistical) uses!

What is Big Data? (2)

- Definitions
  - IT-view (Wikipedia)
    - Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them.
  - Gartner (~UNECE/official statistics)
    - Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.
  - A statistician
    - Big data is neither a survey nor an administrative data source, its something different

From primary to secondary data



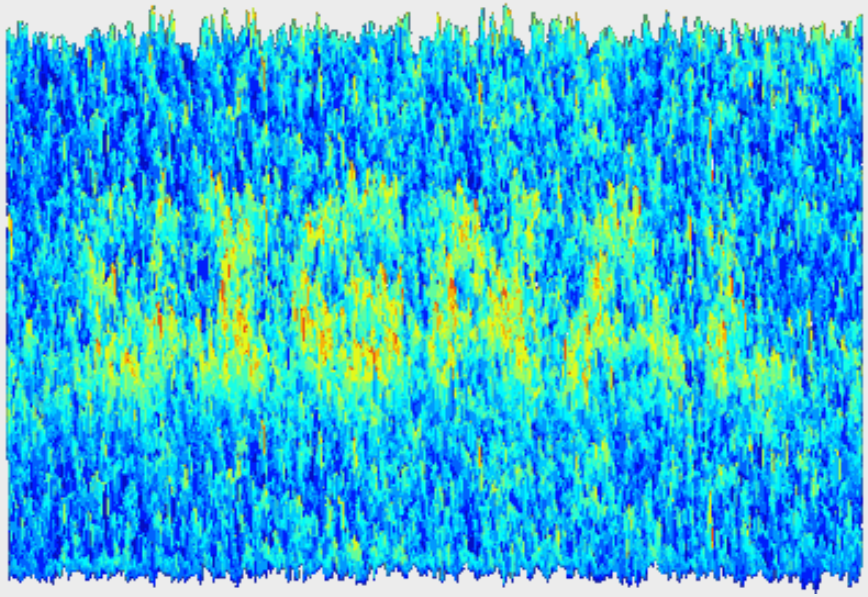
Properties of Big Data

- Big Data is a source of data that is:
  - Rapidly available
  - Usually available in large amounts
  - Often generated by an unknown population
  - May have poor quality metadata
  - Usually has low information content
  - May contains lot's of noise
  - Requires processing prior to use
  - Unknown design

- The V's and Big Data
  - Volume, Velocity, Veracity, Variety, Value, Variability, ….

Signal and Noise

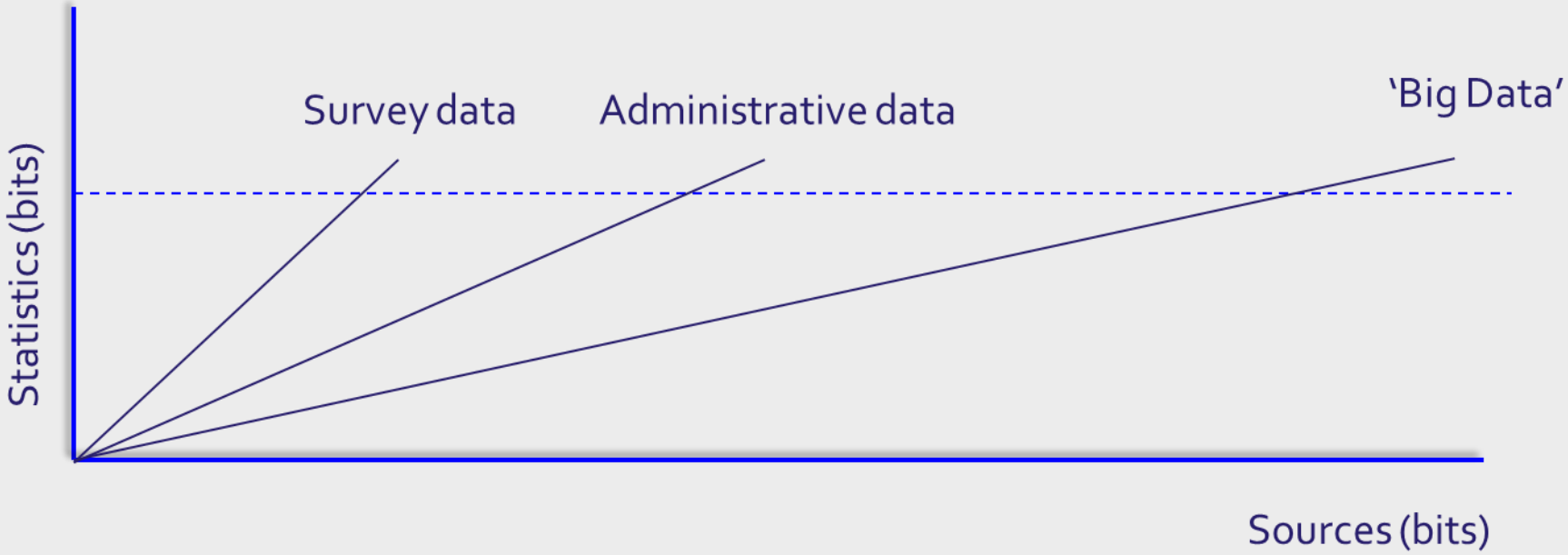
– Big Data has a lot of data, but often:
  - Most of the data are ‘noise’, only a limited part is ‘signal’
  - Signal is that part of the data the researcher is interested in
  - Goal is finding the signal
  - However: one man’s signal is another man’s noise



Data ≠ Information

Information content

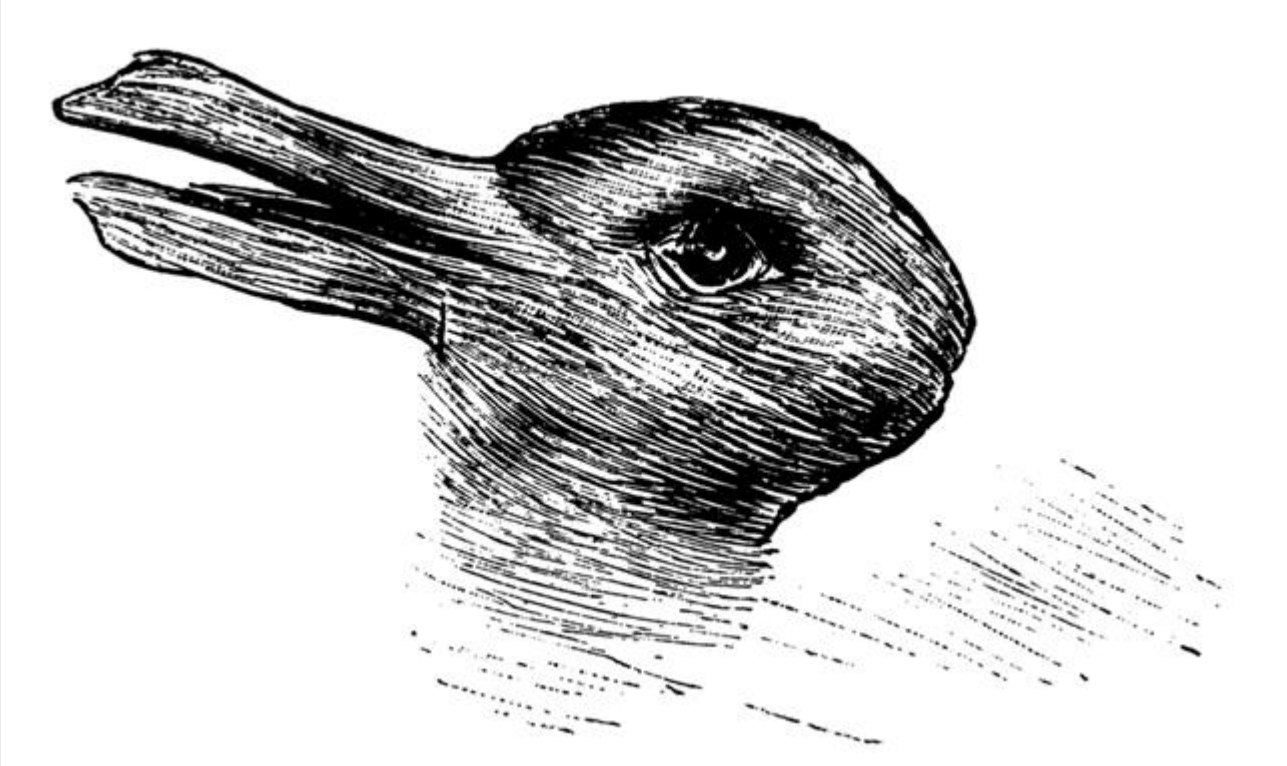
- Compared to the other data sources
  - Big Data has a low information content



Large amounts of data are required to extract information from Big Data

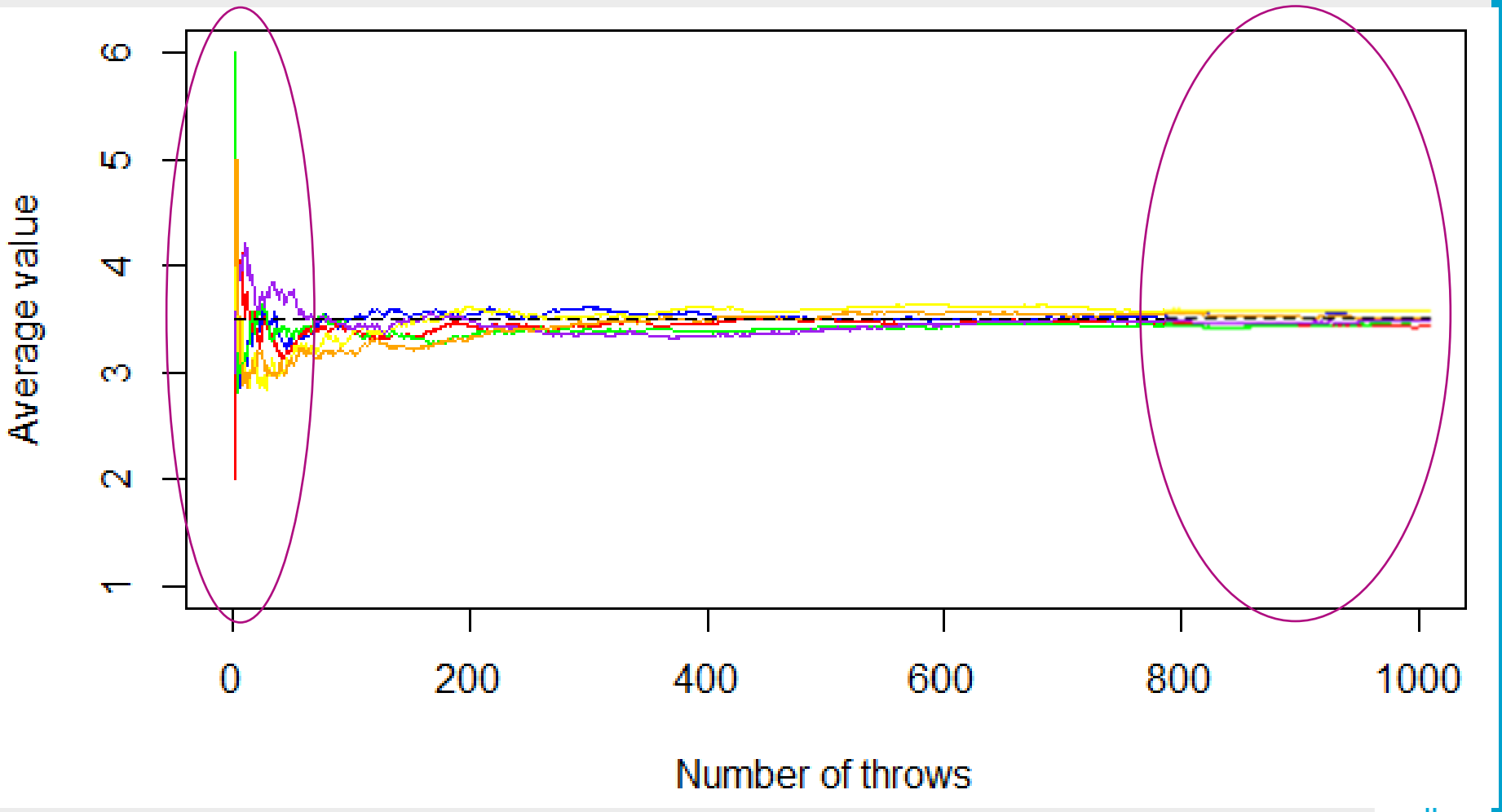
Paradigm shift: Different mindset

– Need for change in the way statisticians look at data



https://en.wikipedia.org/wiki/Rabbit%E2%80%93duck_illusion#/media/File:Kaninchen_und_Ente.png

Mindset: Small (survey) vs. Big



Development of the average value of 1,000 subsequent single dice throws for six different runs. The dotted line represents the expected value (3.5)

Diversity of Big Data

- There are many types of Big Data
  - Human generated, Machine generated, Admin data like, Image containing, Text containing, …

- There are many potential uses of Big Data
  - To produce rapidly available indicators (‘real-time’/day/week)
  - For new phenomenon (never measured)
  - As a supplemental data source for an already produced statistic (more detail, faster, cheaper, reduce admin burden, …)
  - …

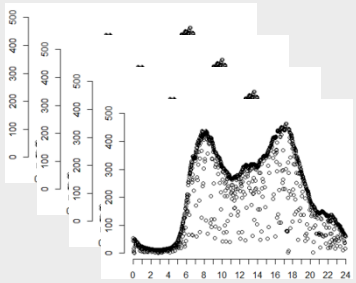
- Big Data can be used to:
  - Measure a phenomenon *Directly* or *Indirectly*.

Overview of Big Data based statistics

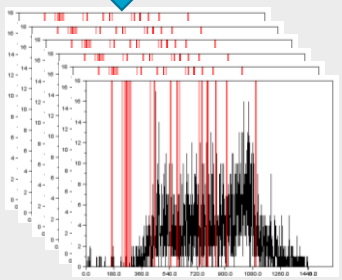
Nr	Name	Status, country	Sources used
1	Consumer Price Index	in production, multiple countries	Scanner data & web prices
2	Traffic intensities	in production, NL	Road sensors
3	Online job vacancies	towards implementation, ESSnet BD	Web portals & company
4	Enterprise characteristics	towards implementation, ESSnet BD	Company websites
5	Electricity/energy consumption	towards implementation, ESSnet BD	Smart meter data
6	Maritime and Inland waterway statistics	towards implementation, ESSnet BD	AIS data
7	Financial transaction based statistics	exploratory, ESSnet BD, NO	Bank transaction data
8	Earth observation derived statistics	towards implementation, ESSnet BD	Satellite / aerial pictures
9	Mobile network derived statistics	towards implementation, ESSnet BD	MNO data
10	Innovative tourism statistics	exploratory, ESSnet BD	Various data sources
11	Innovative company websites	towards implementation, NL	Company websites
12	Social mood on economy index	published experimental, IT	Social media (Twitter)
13	Mobile phone derived outbound tourism	experimental, AU/FI/Estonia	MNO data

Road sensors use: show steps involved

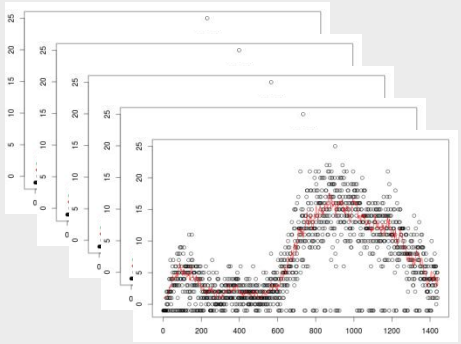
(A) Frame



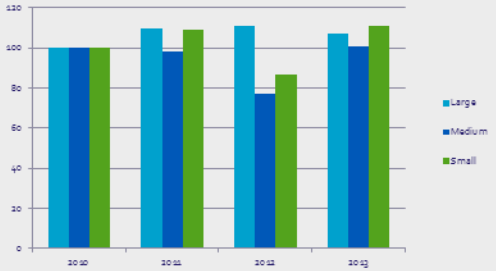
(1)
Transform
+ Select



(2)
Cleaning



(3)
Estimation



-*Regional estimates*
-Month/quarter/year

1. Consumer Price Index

- Traditionally collected by ‘interviewers’ visiting shops
  - Sample of shops
  - Sample of products (10-100 per shop)

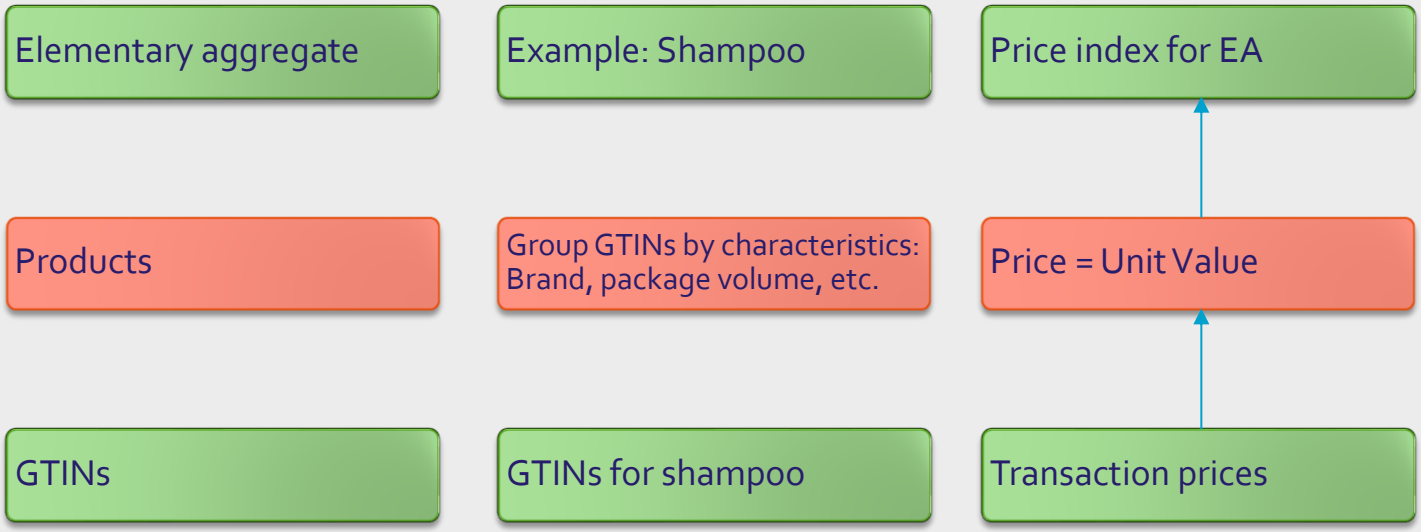
- More and more offices use alternative data sources
  - Scanner data (mainly retailers)
  - Web scraped prices (usually for specific products/services)

- Start of Big Data Methodology

Sources for CPI

Data dimension	Traditional (survey)	Scanner data	Web scraping
Data collection	Manually	Automatised	Automatised
Completeness/scope	Samples	All transactions	Bulk or sample
Metadata	Item description	Item description + attributes	Any website info
Price data	Offer prices	Transaction prices	Offer prices
Quantity data	None	Quantities sold	None

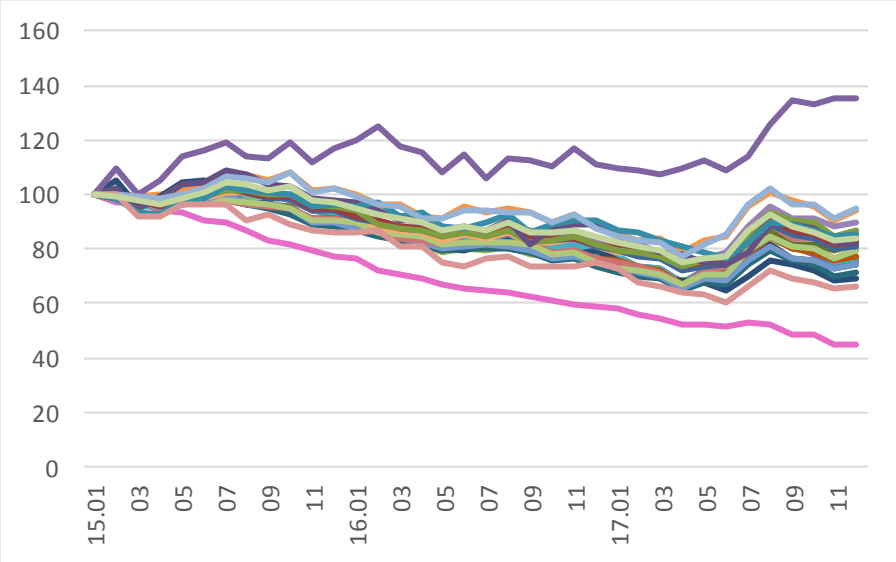
Elementary Aggregate



Elementary Aggregate

Impact of product definition on index: TV's

Pick the 'right' index …



Finding the right Index on micro level

Index is made up of:

- Quantity
- Price

2002-2009: Laspeyeres index:

Price compared to t_0 but with current quantities

2010-2017: Jevons Index

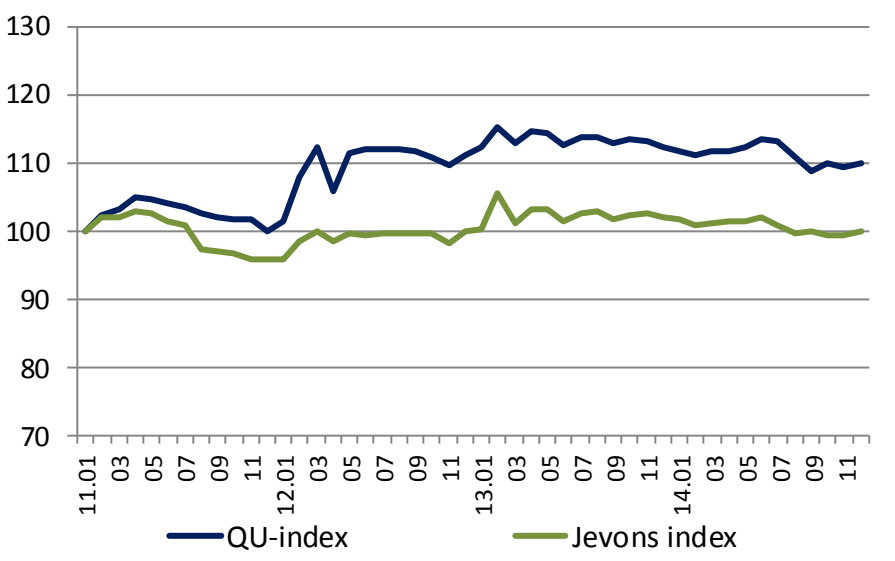
Only comparison of prices

2018- : Geary-Khamis Index (QU method)

An index based on both prices and quantities

QU-GK vs Jevons: Price index for sugar

The difference is mainly caused by the weighting (composition of the 'basket' changed)



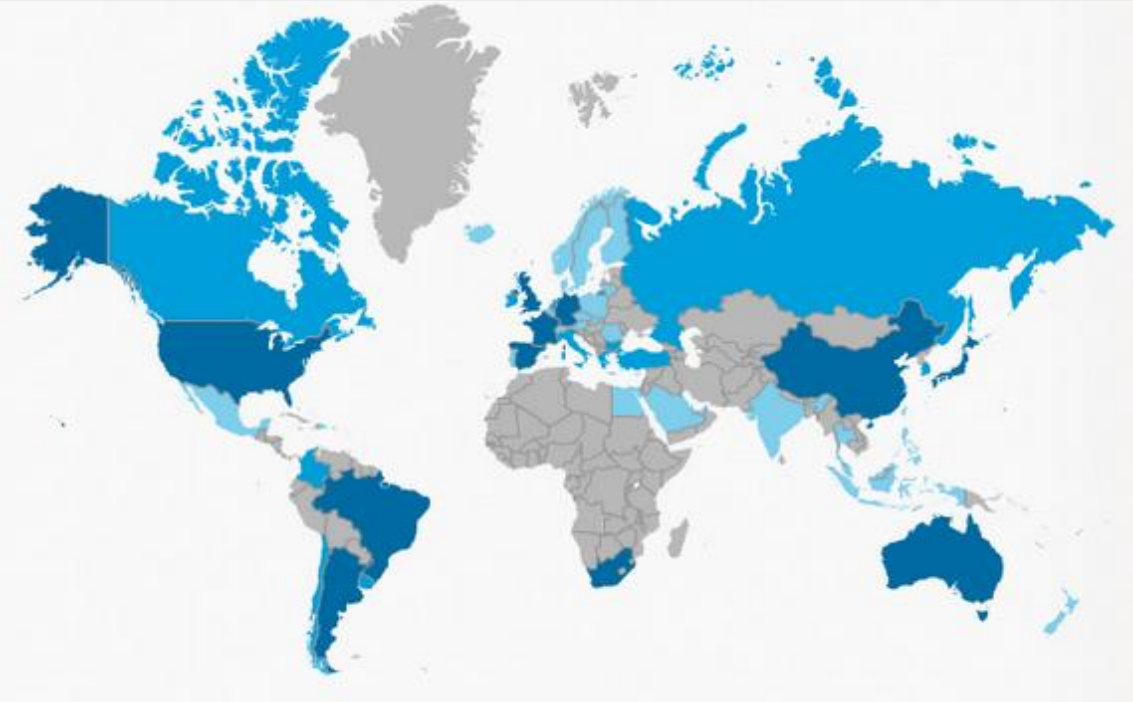
New CPI Methodology

Paradigm shift:

It took until 2018 before a method, appreciating all aspects of transaction data was used!

Consumer Price Index

- PriceStats produces worldwide daily CPI's based on web scraping alone
  - Every day retailers web sites are scraped worldwide to obtain the price of thousands of online products



2. Web based statistics

- Direct use
  - Statistics based on prices and company characteristics
  - “The data is what the data is”

- Indirect use
  - Statistics based on models extracting ‘information’ from a Big Data source
    - Detecting innovation from text on web page
    - Deriving job vacancies from online job portals

Detecting innovative companies

- Web pages of companies provide information
  - Can this be used to substitute the information collected by the Community Innovation Survey?
  - A survey on Innovation send very other year to a sample of 10.000 companies (WP >= 10)

- In the study we looked at:
  - The potential of *web pages* to provide information on the *innovative* character of a company
  - For both *large* (WP >= 10) and *small* (WP < 10) companies
  - The CIS survey data of 2016 was used for model development (as ground truth)

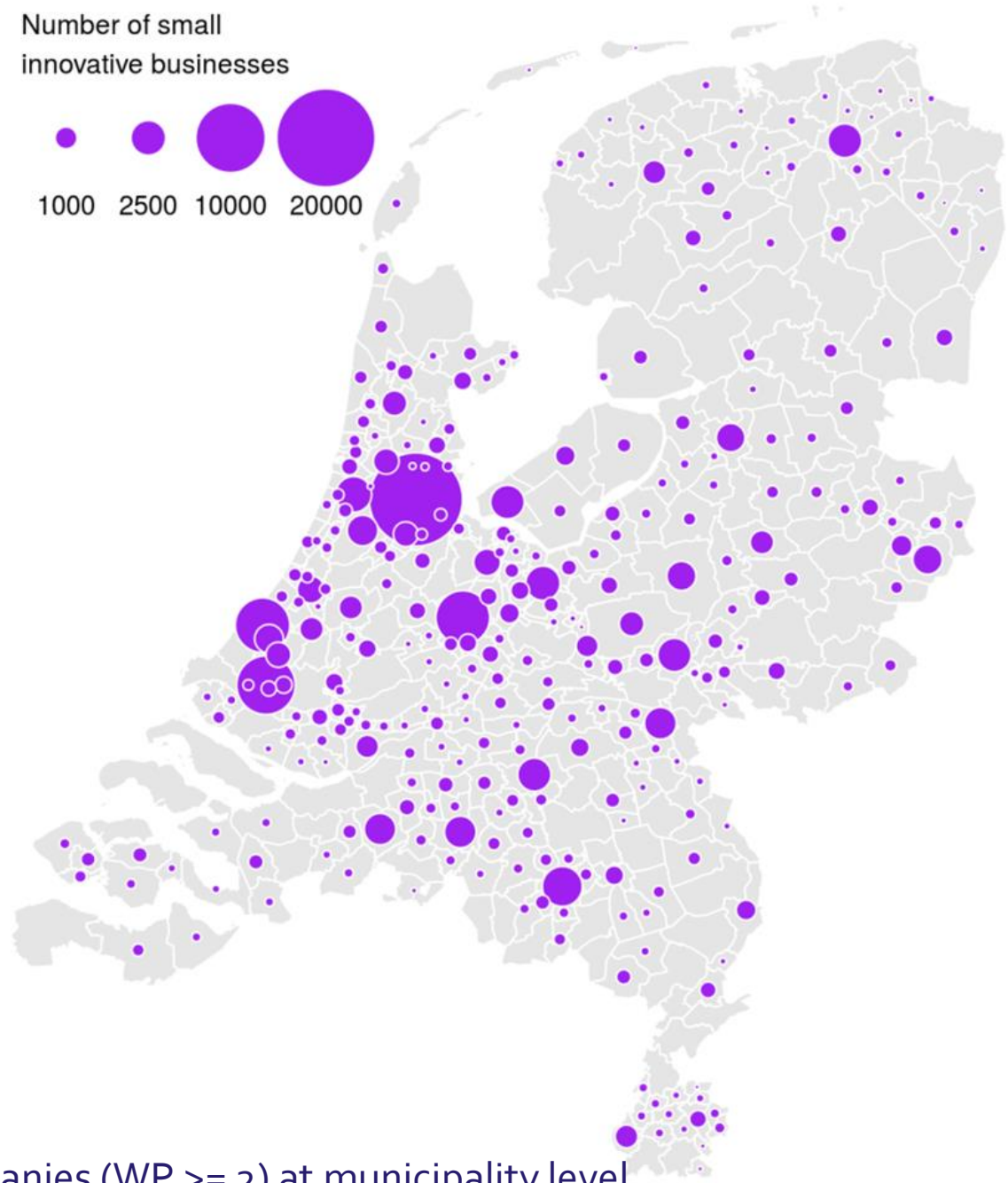
Detecting innovative companies (2)

- Relation Company – website
  - Used URLfinder, lists of companies url's, did rigorous manual checking
- Examples of Innovative and Non-Innovative companies
  - The CIS survey of 2016 detected 3,340 innovative companies. Used a similar sized sample of non-innovative companies
- Web scraper
  - Written our own program in Python
- Using Text as data
  - Different area of expertise, experimented a lot.
  - For large documents there are a number of standard processing steps
- Classification method/algoritm
  - Tried everything included in scikit-learn (Python library)
- Be critical
  - Checked a lot of the findings manually (essential).

Detecting innovative companies (3)

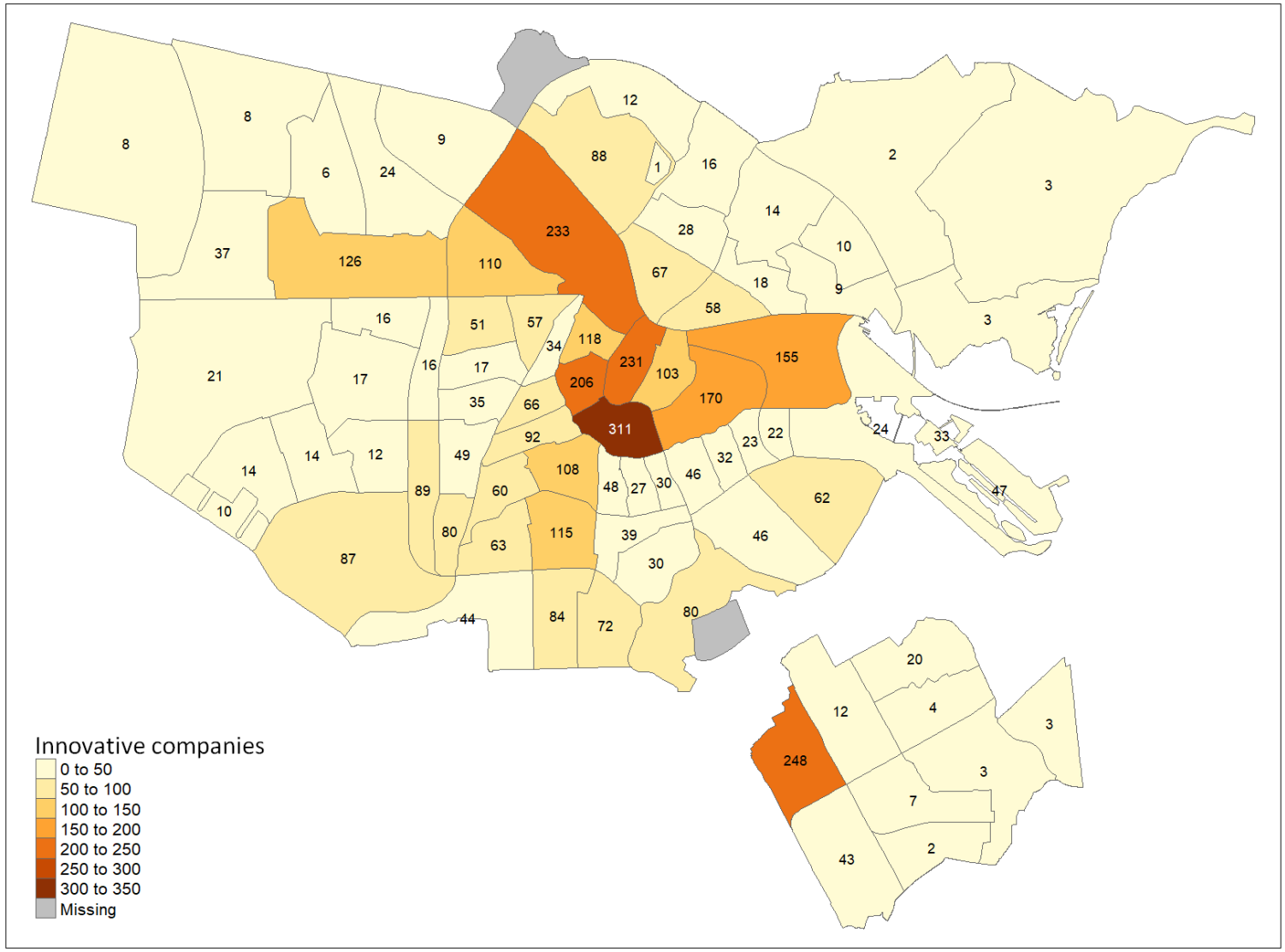
- The data
  - The main page of the web site of the companies included in the CIS survey were scraped (Innovative and non-innovative).
  - The text on that page was extracted, processed and used in model development
- Supervised learning
  - Various classification algorithms were tried and findings compared.
  - Logistic regression (L1-norm) won, Accuracy of 88%
  - Logical relation between the most important positive and negative words in the model (incl. manual checking).
- Model validation
  - Model could be used to detect both large and small innovative companies (such as startups)
  - Approach can be applied in NL and Germany (Kinne and Lenz 2019)
  - New method is able to reproduce the CIS-survey results
    - Survey 19.916 ± 680, Web based 19.276 ± 190 innovative **large** companies
  - Correct for model bias (FP), websites with words < 10, and innovative companies without a website (0.1%)

Dete)



Innovative companies (WP >= 2) at municipality level

Detecting Innovative companies (5)



Innovative companies (WP >= 2) at zip code level for Amsterdam

3. Social media based indicators

Popularity varies per country

For Europe:

1) Facebook (70%)
2) Twitter (11%)
3) Pinterest (8%)
4) Instagram (7%)
5) YouTube (2%)

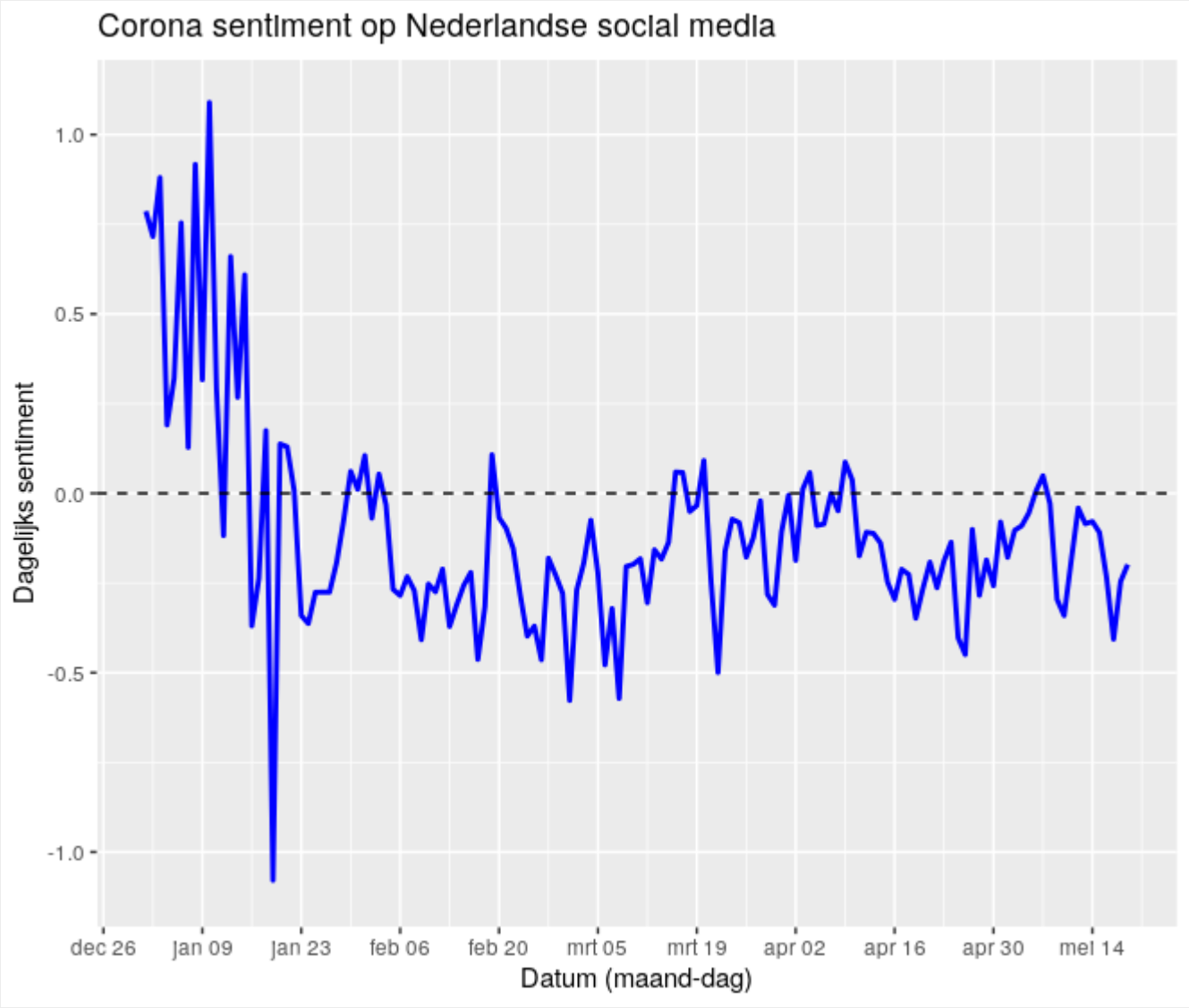
Getting the data can be challenging (relation with GDPR)

- Cooperate with the owner of the platform
- Use public access (API's or scraping) or buy access
- There are commercial companies that scrape it for you!

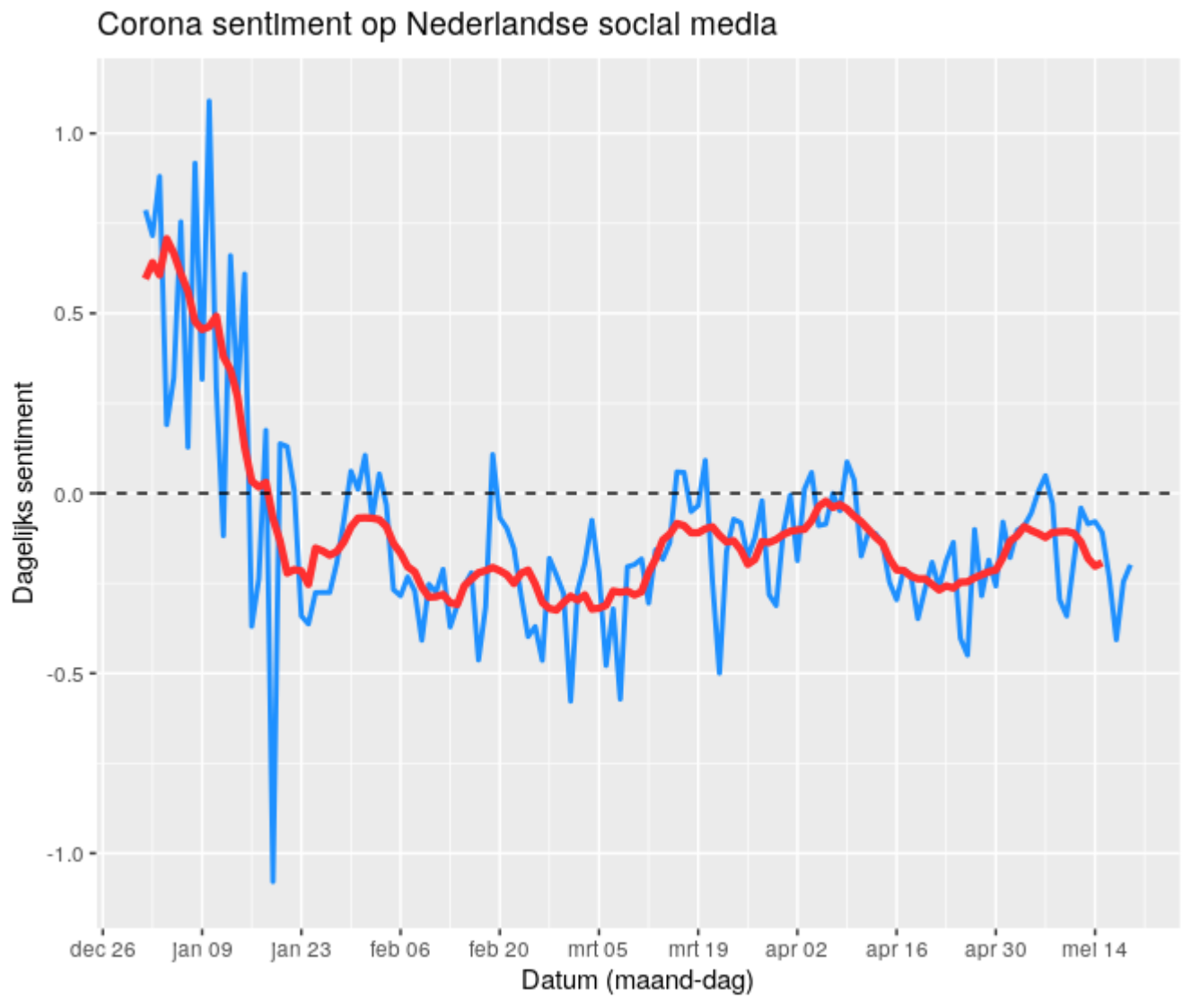
Corona indicators: sentiment indicator

- Creating a Corona-based sentiment indicator
  - What is the sentiment of the Dutch towards Corona?
- Select Corona containing messages
  - ‘Corona’, ‘Coronavirus’, ‘COVID-19’, ‘COVID19’, ‘’COROVID19’, SARS-COV-2’ and ‘pandemic’
  - Predominantly on Twitter and Facebook
- Determine sentiment of these messages
  - (#pos - #neg) / #total
  - Development over time

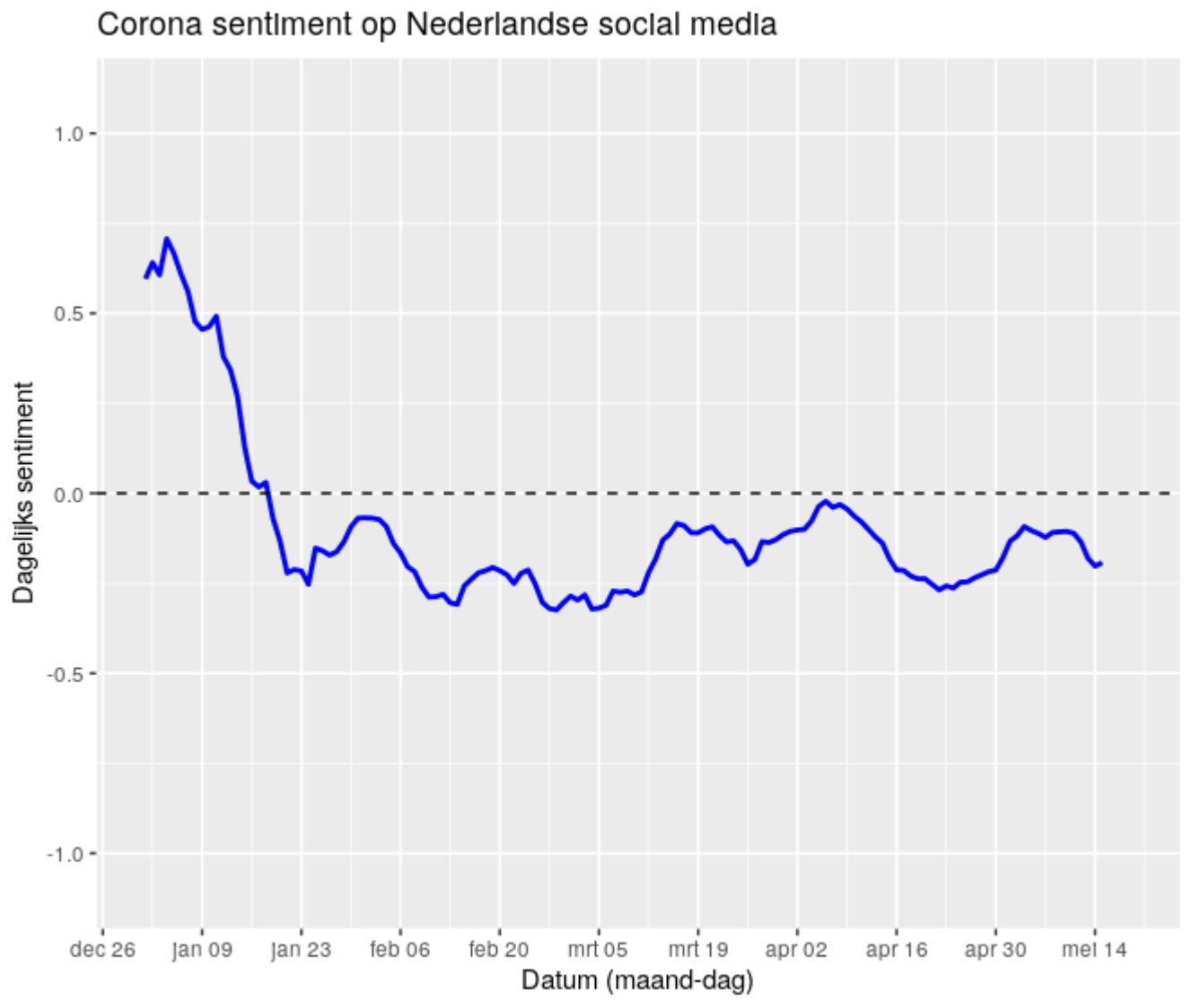
Raw daily signal



Raw daily signal + filter (moving average)



Filtered signal: Corona sentiment



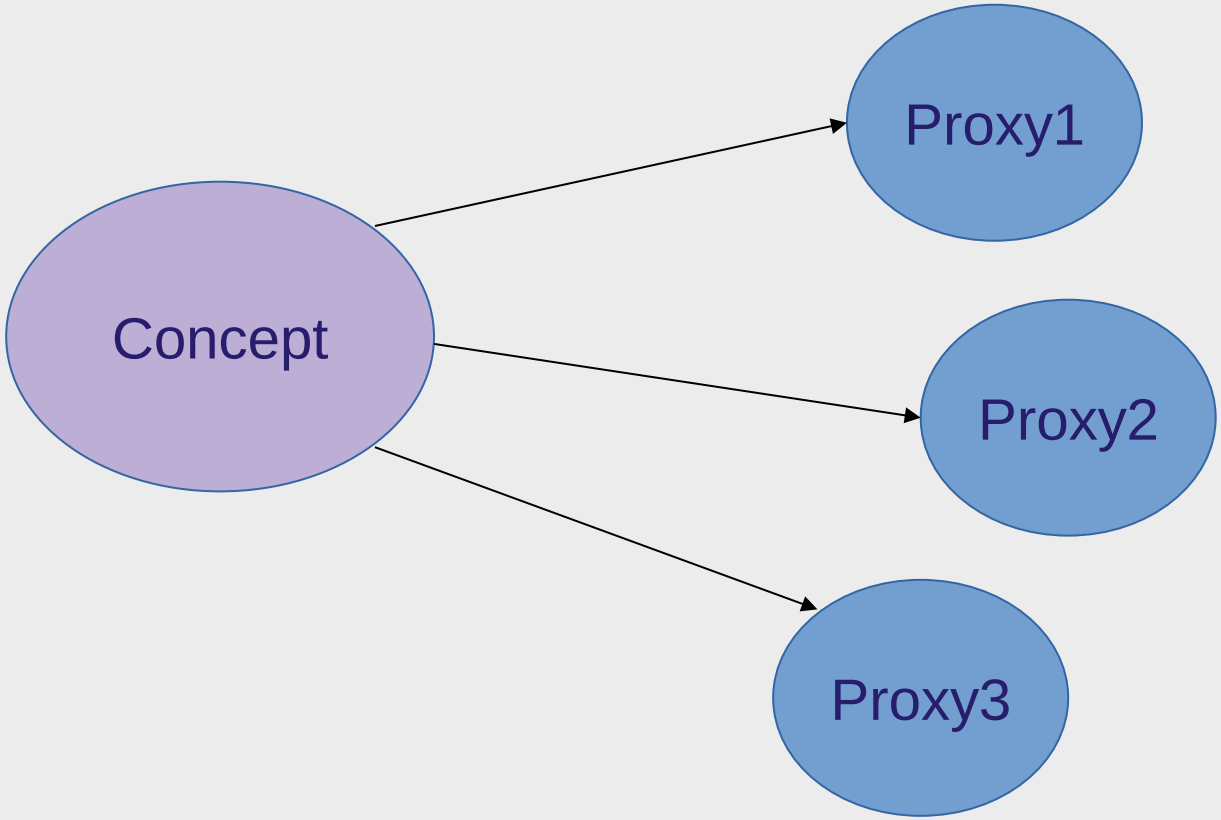
Findings

- Starts positive
  - Mainly because everybody hopes Corona stays in China
- Becomes negative on Jan 20th
  - Corona spreads from China to Japan and South-Korea
- Lowest value on Feb 27-28th
  - Date that first Corona patient was detected in NL
- Stays negative after Jan 20th
  - Moves up and down. Is affected by lockdown rules, politicians, remarks of famous Dutch people etc.

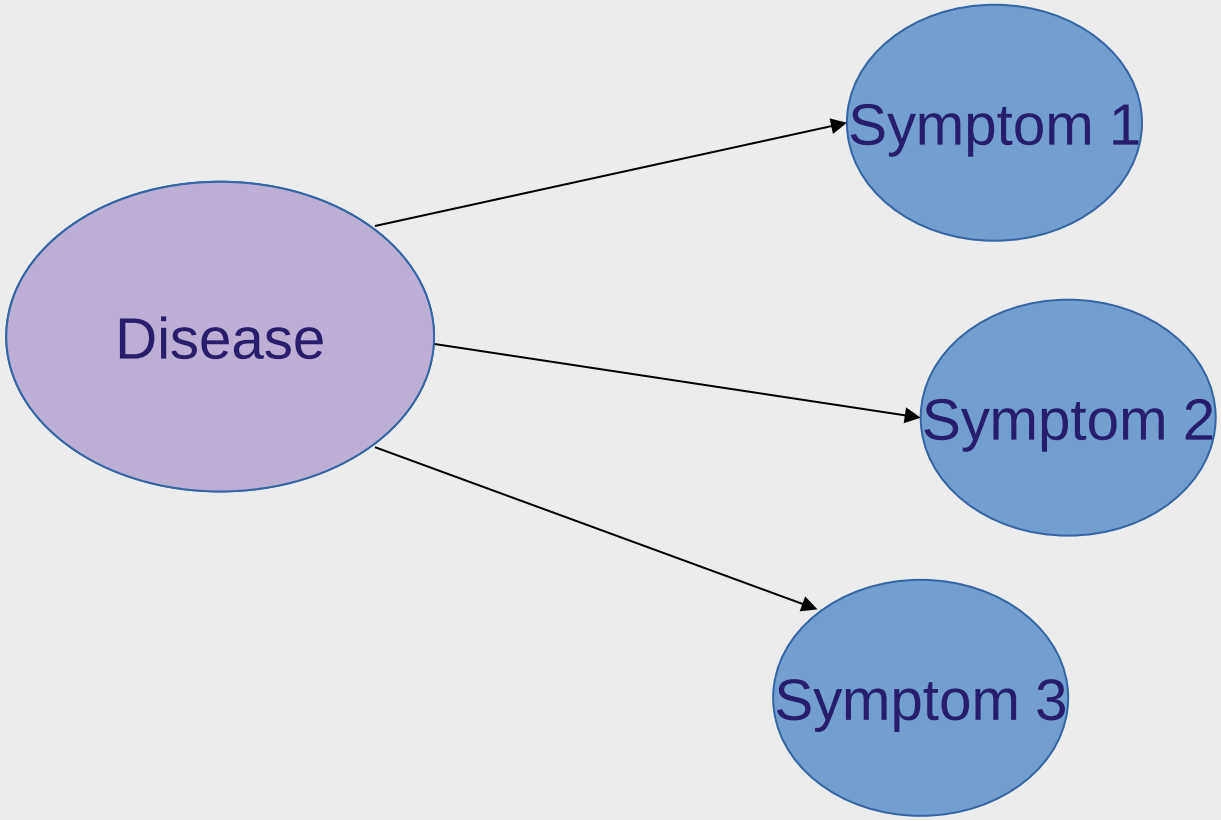
Corona indicators: infection indicator

- Creating a Corona infection indicator using social media
  - Are there messages that report corona related symptoms?
  - Develop a model to derive the number of infected persons
  - Study the development over time

Concepts and operationalizations



Concepts and operationalizations



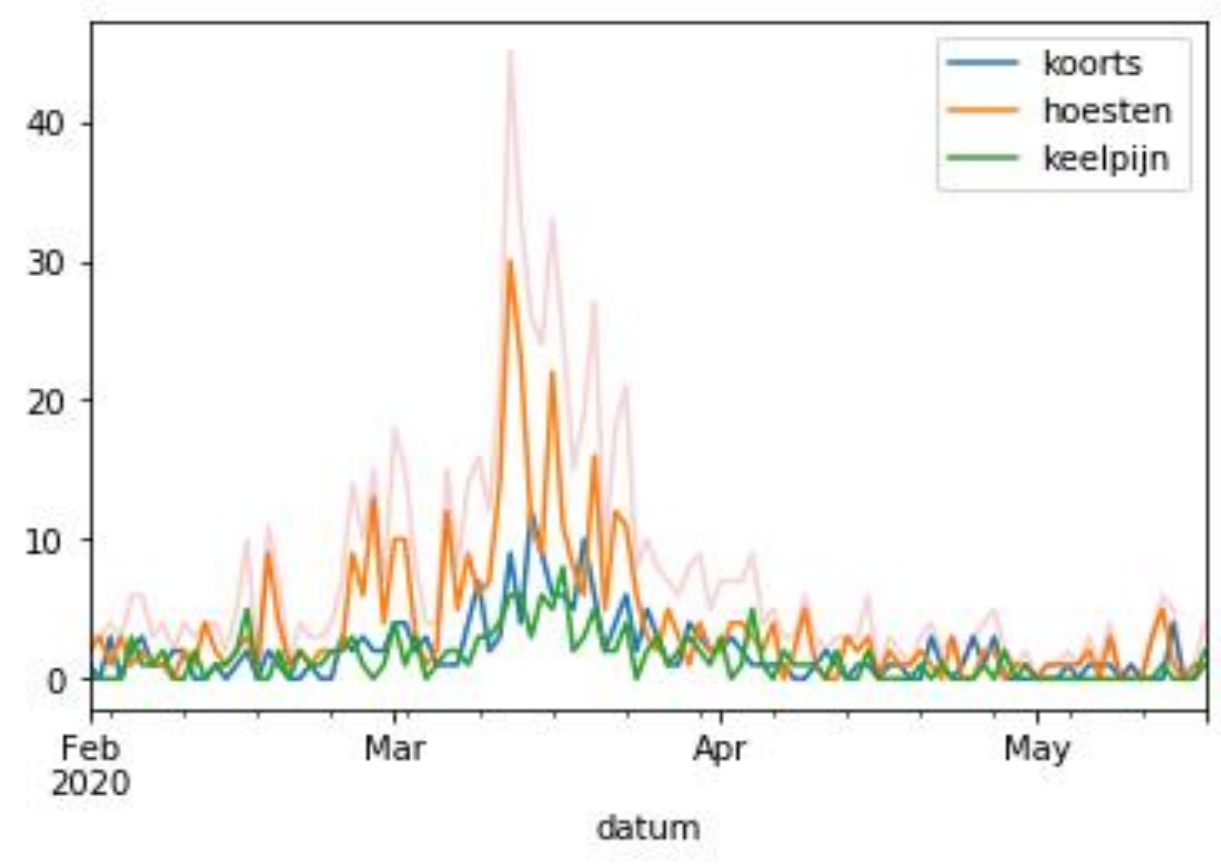
Selection of Features

- Three symptoms for COVID-19 (preselected using Coosto):
  - Soar throat
  - Cough
  - Fever
- Over a period of 2 years, 1700 messages where collected mentioning these symptoms
- Annotated by hand
- Logistic regression

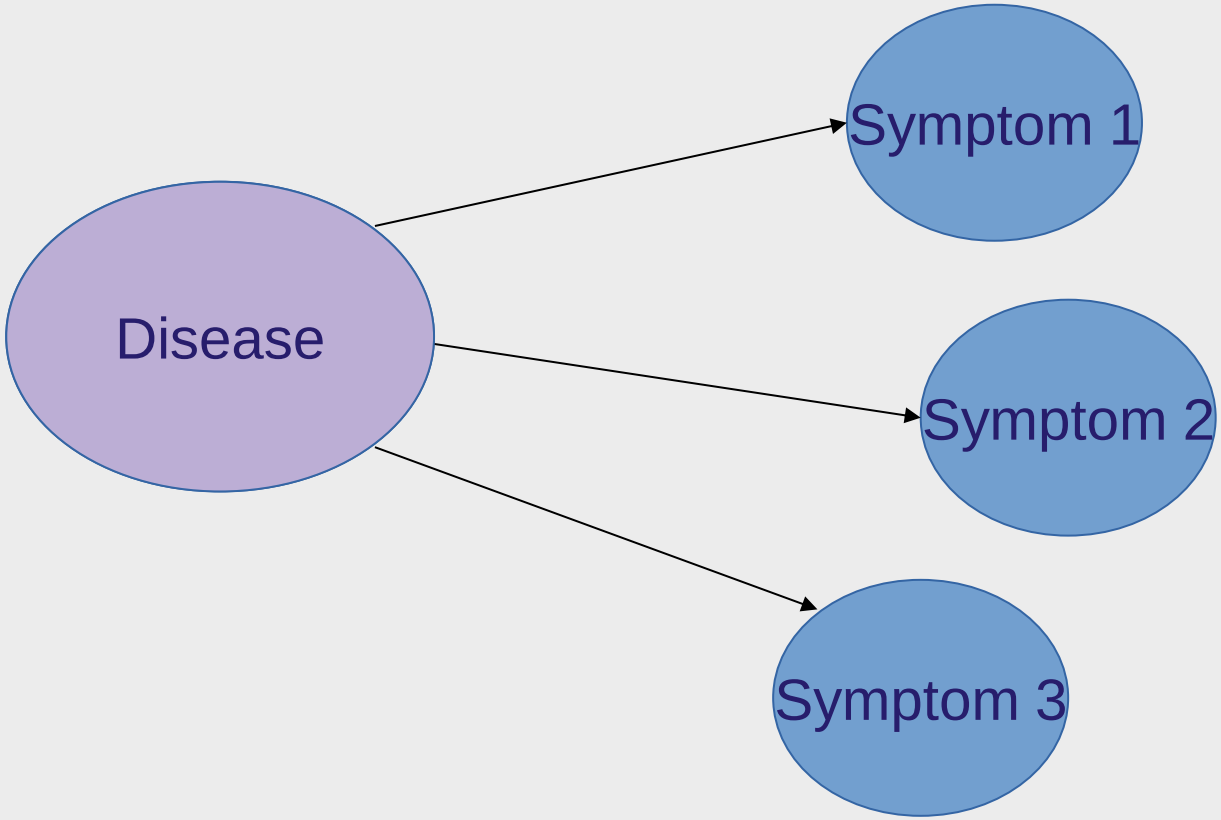
Training/test set

Symptoom	# Messages	% symptoms in dataset	accuracy
Hoesten (cough)	500	36 %	0.92
Keelpijn (sour throat)	700	78 %	0.94
Koorts (fever)	500	29 %	0.83
Totaal	**1700**		

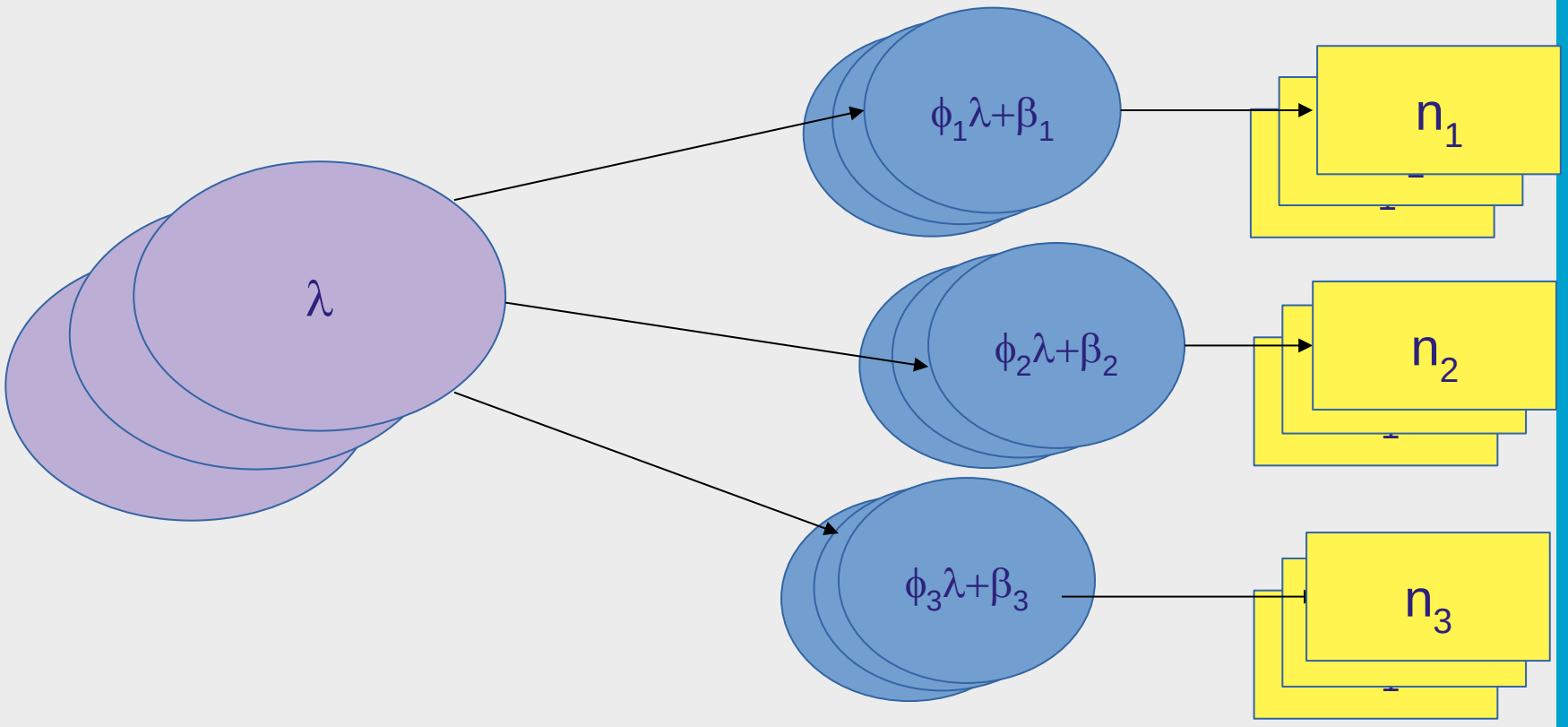
Tweets


koorts
hoesten
keelpijn
40
30
20
10
0
Feb
2020
Mar
Apr
May
datum

Bayesian Approach



Bayesian Approach



Bayesian Model

Random Walk: $log\lambda_t = log\lambda_{t-1} + \epsilon_t,$ where $\epsilon_t \sim N(0, \sigma^2)$

Observations: $n_{s,t} \sim NB(\phi_s\lambda_t + \beta_s, \alpha_s)$ Semi Poisson Process

Priors:

$$\frac{1}{\sigma^2} \sim Gamma(a_\tau, b_\tau)$$

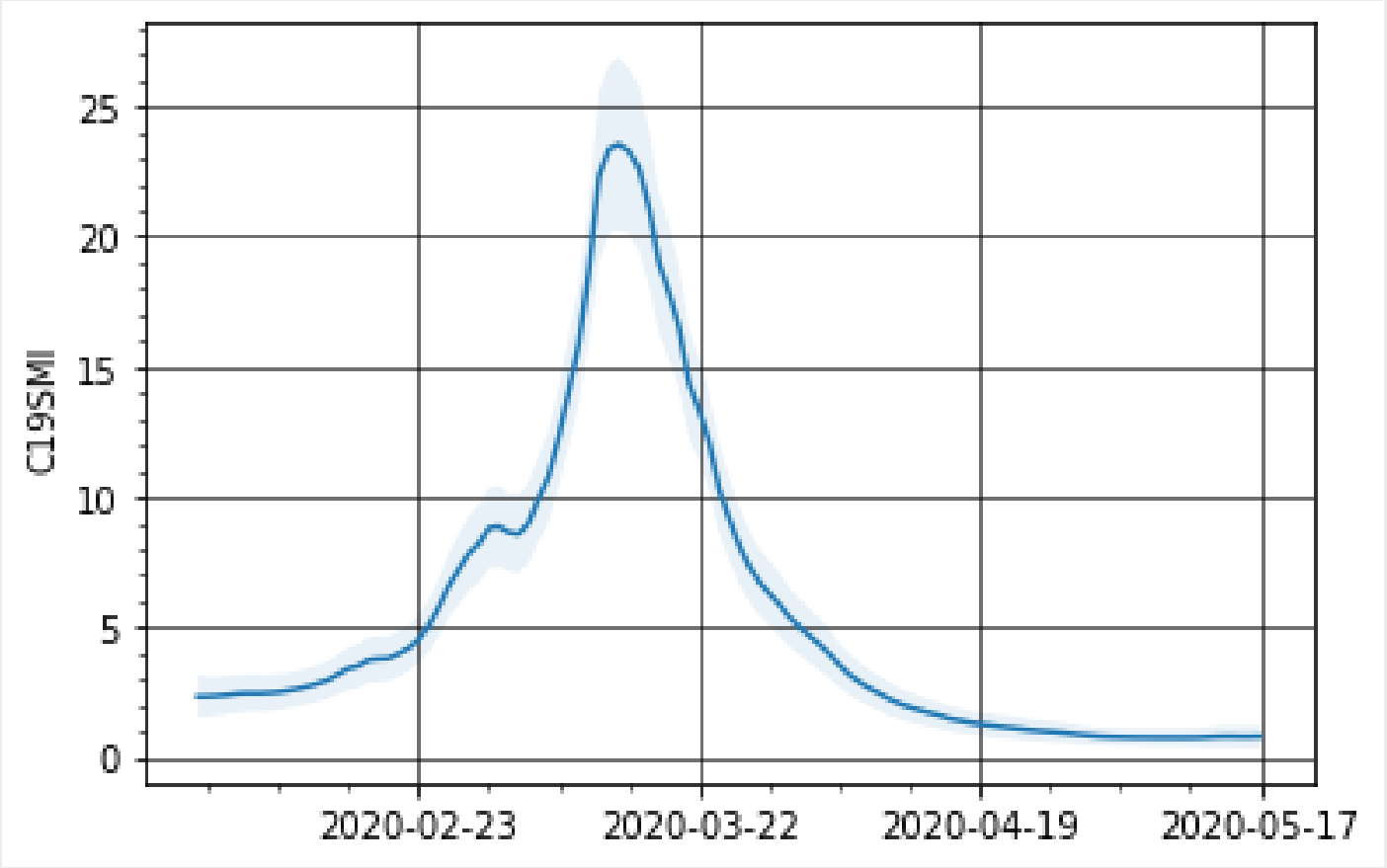
$$\phi \sim Dirichlet(a_\phi)$$

$$\alpha \sim Gamma(a_\alpha, b_\alpha)$$

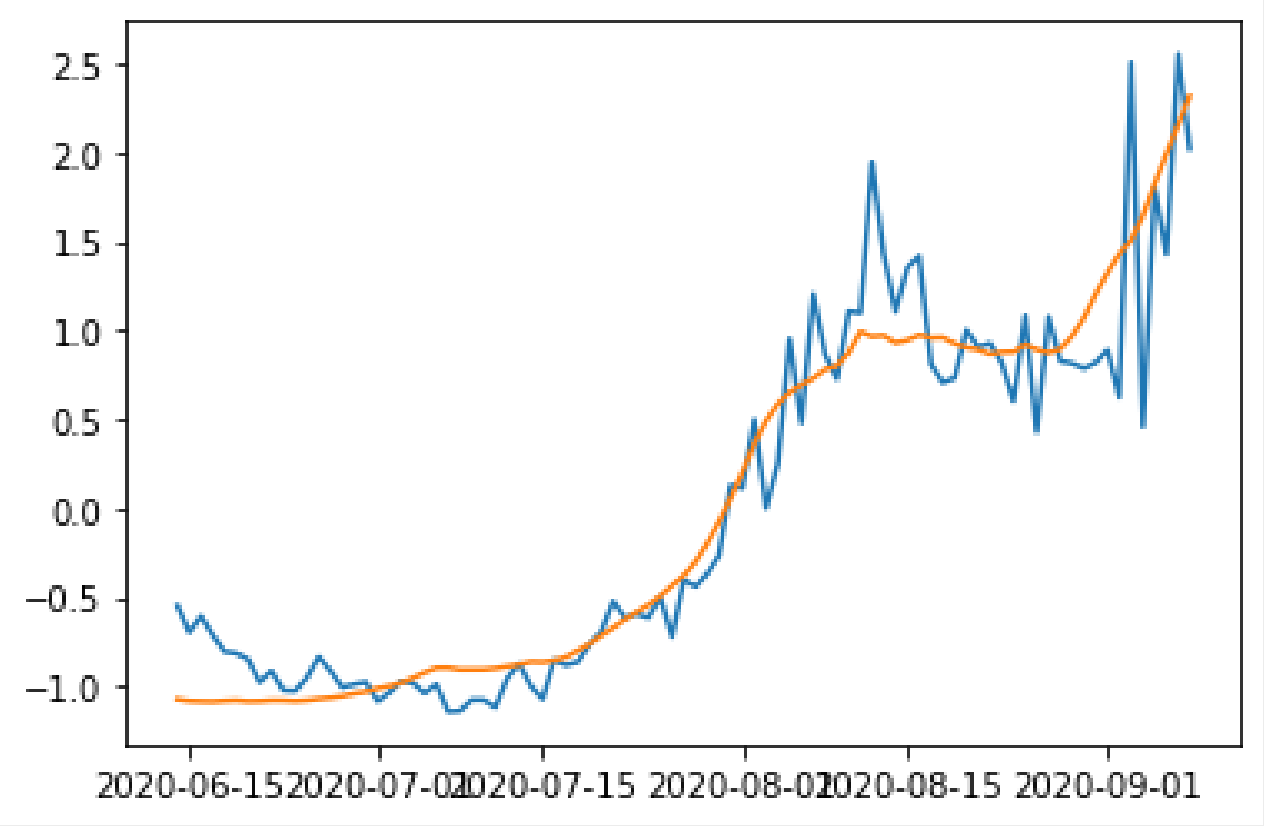
$$\beta_s \sim Gamma(a_\beta, b_\beta)$$

- Implemented in pyMC3
- No U-Turn Sampler (Hamiltonian Monte Carlo)
- Burn in: 4000 iterations
- Sample size: 200 samples

Applying the model

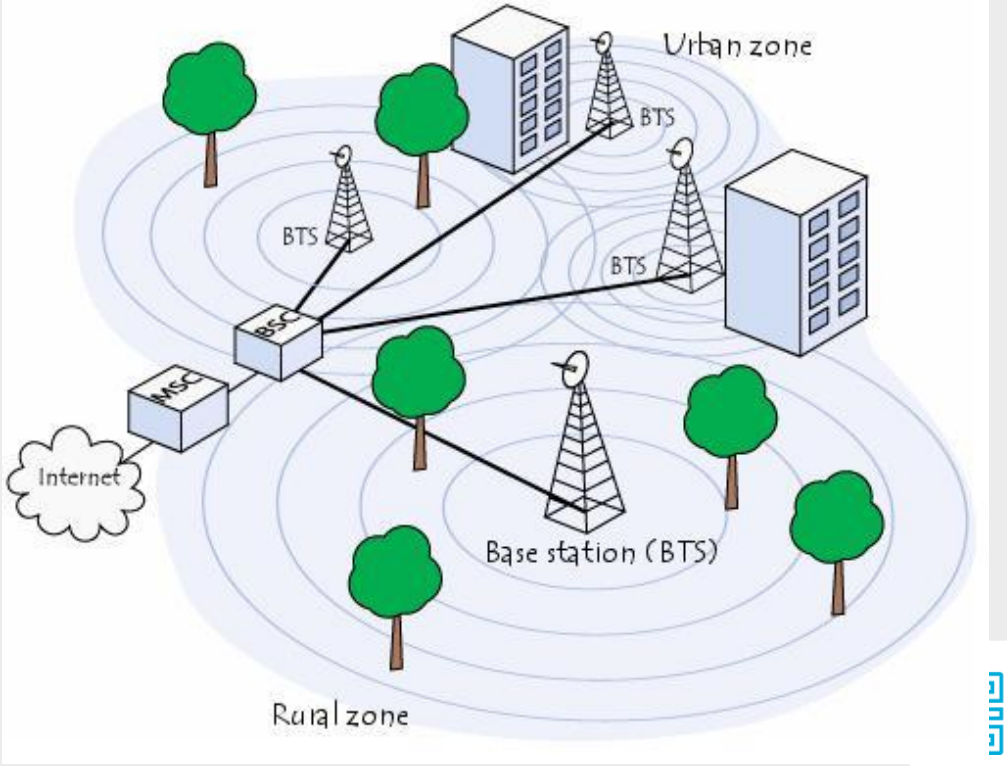


Comparison with national data



4. Mobile Network Operator (MNO) data

– To provide optimal use of mobile phones, mobile phone operators must have a well operating infrastructure of phone masts and base stations that cover the country well

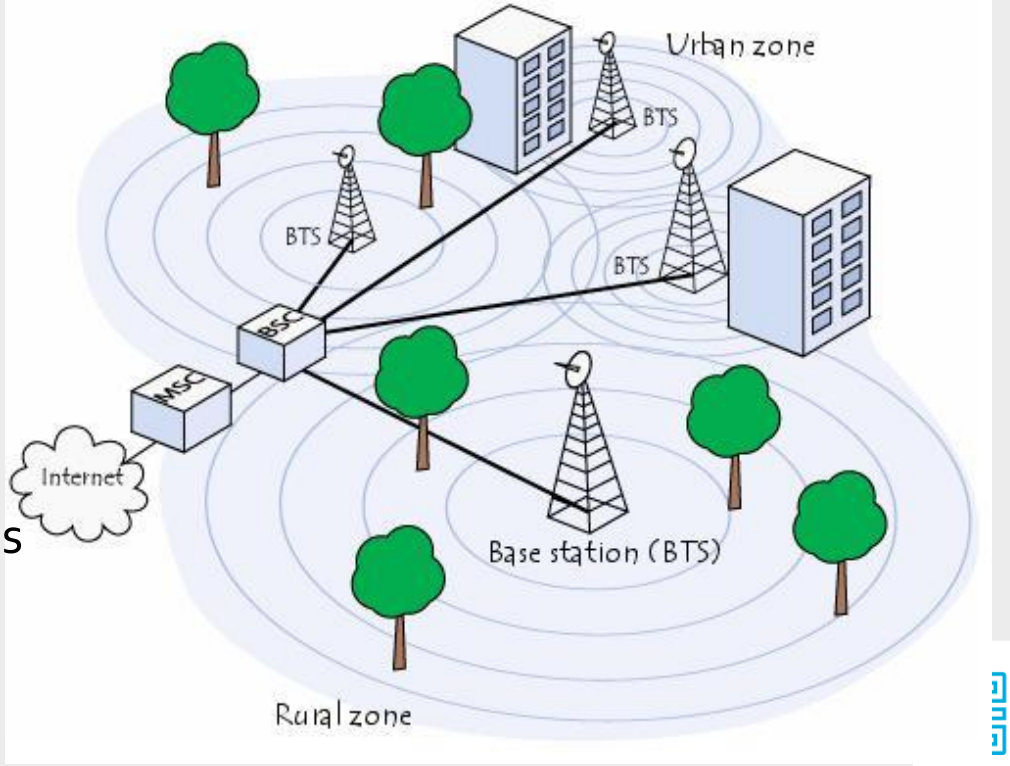


Mobile Network Operator (MNO) data

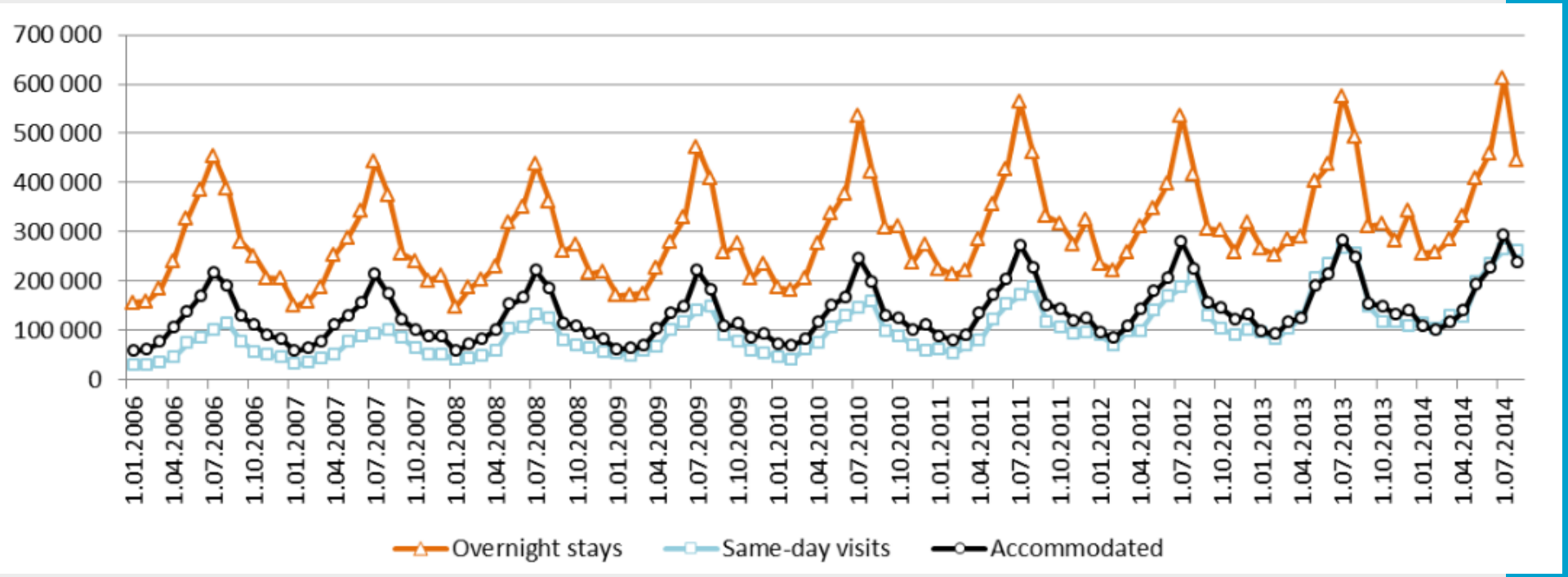
– To enable proper use, the Base station to which a mobile phone is connected must be known, at each point in time

This enables tracking of a mobile phone. Hence its location can be obtained

- Very interesting applications
- Privacy considerations

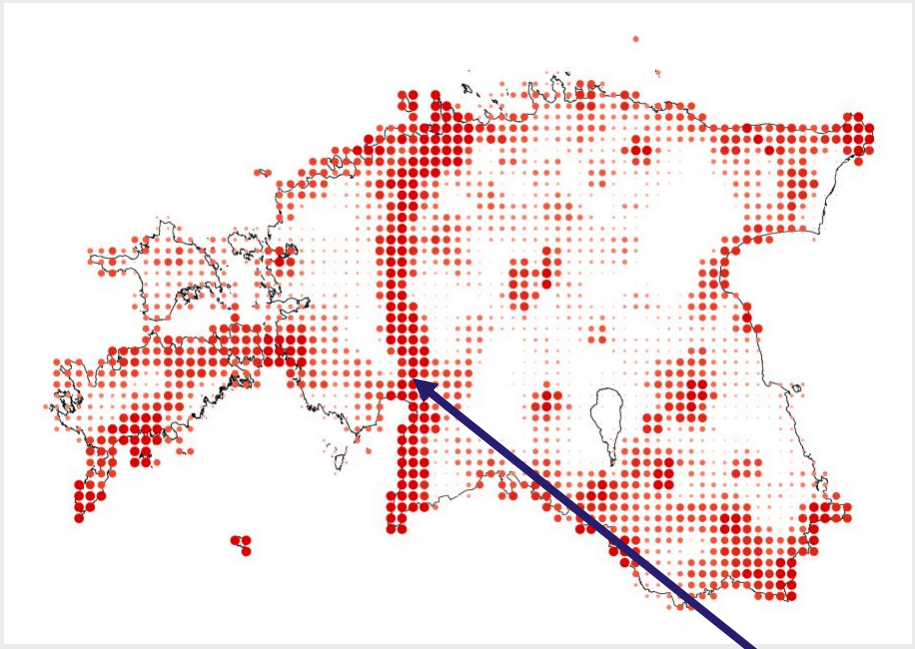


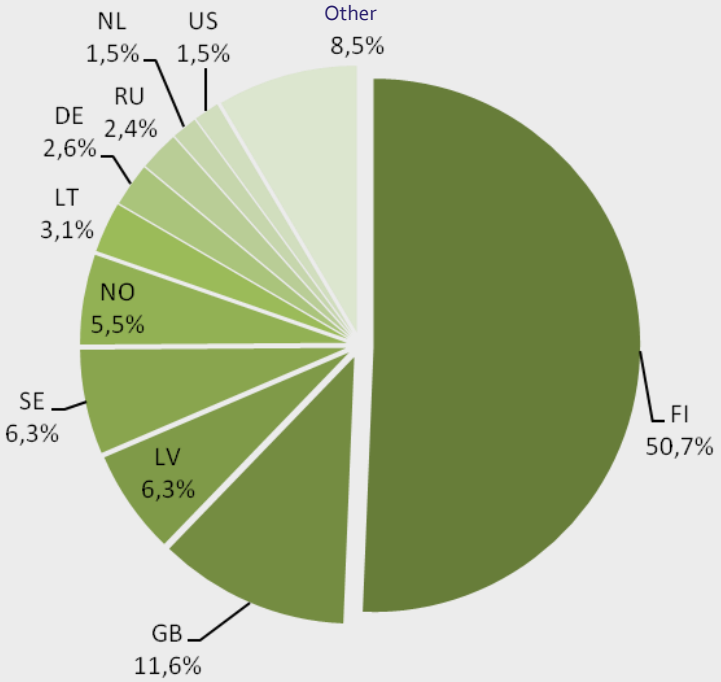
Inbound tourism: Estonia

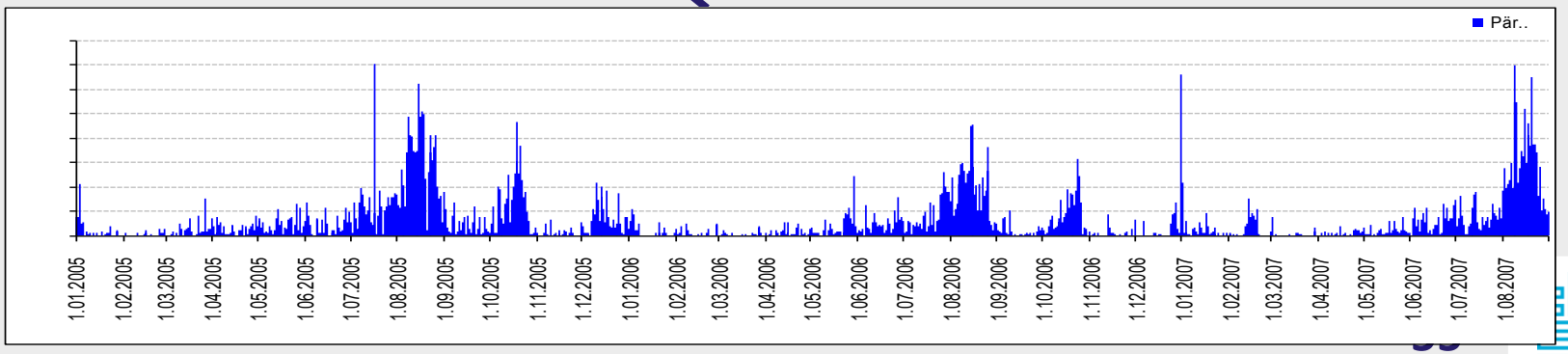


- Coherence between overnight stays and same-day visits of inbound tourists from mobile positioning statistics (estimation based on the data of two MNOs),and official accommodation statistics (Statistics Estonia 2014)

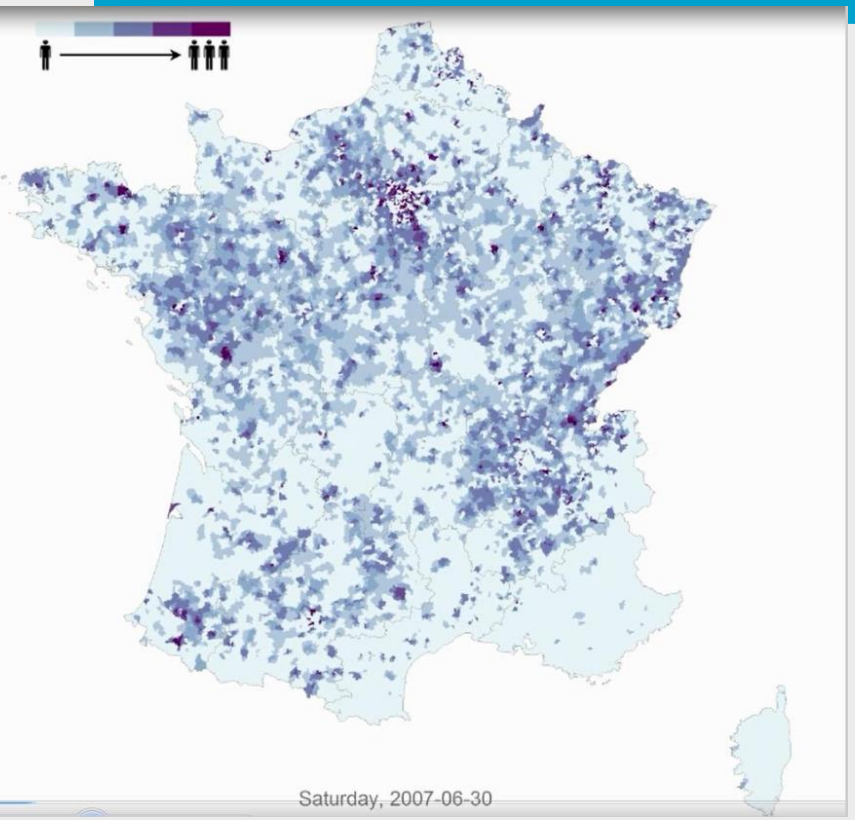
Inbound tourism: Estonia (2)

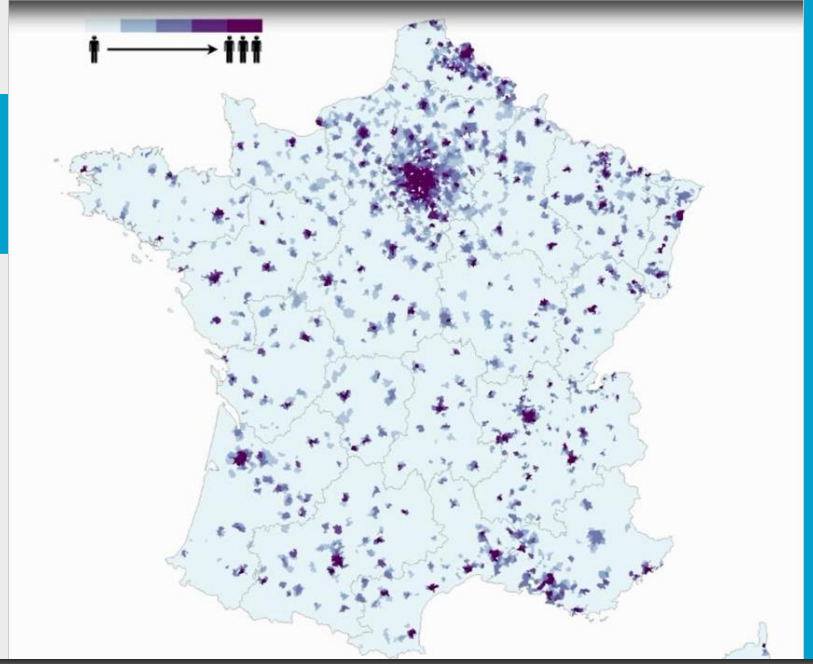


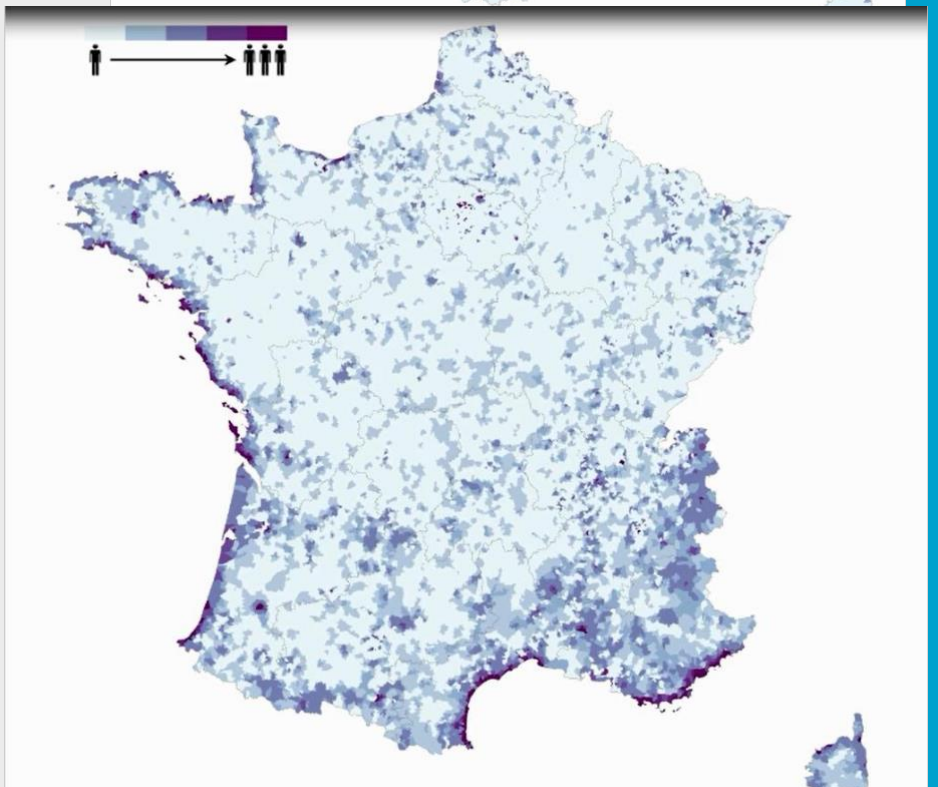




Mobility in France







https://www.youtube.com/watch?v=qsUDH5dUnvY

Deville, Pierre, et al. "Dynamic population mapping using mobile phone data." Proceedings of the National Academy of Sciences 111.45 (2014): 15888-15893.

General remarks

- We have seen many examples of Big data based studies
- These are very divers (both in data and methods used)

- Using Big Data brings new challenges for statistics
- Each data source has its own methods
- Data needs to be selected, cleaned, and interpreted
- Data generating process is important for lowering the volatility of the results
  - Models are used
    - Language models, Stochastic models etc.
  - Need to process large amounts of data

Need for knowledge on

- Extracting information from Big Data
  - Dealing with large amounts of data
  - Dealing with new sources of data: *text* and images
  - Dealing with errors in Big Data

- IT aspects of Big Data
  - How to efficiently process large amounts of data?

- Big Data methodology
  - What are the steps needed to correctly use Big Data?
  - What is signal, what is noise?
  - Advantages of using visualizations

More on this next week!!

Question 1

- What sources do you consider Big Data?
  - Social media messages
  - Product prices on web sites
  - Satellite pictures
  - Sensor data of cows
  - Persons register of China
  - Activity tracker data of 8 persons

Question 2

– What are the risks when using:
  - Scanner data
  - Social media messages
  - Web sites
  - MNO (‘mobile phone’) data

  - Select two sources and write down what you think is important!