Big data, the future of statistics

Experiences from Statistics Netherlands

Dr. Piet J.H. Daas
Senior-Methodologist, Big Data research coordinator
and Marco Puts, Martijn Tennekes, Alex Priem, Edwin de Jonge ….

Statistics
Netherlands

9 Sept, Stockholm

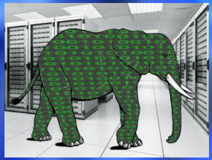
Overview

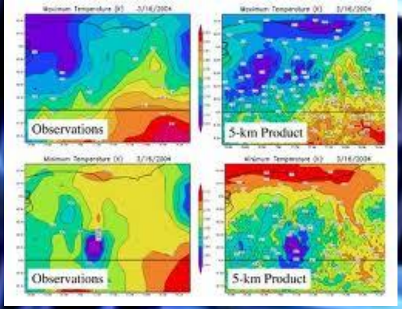
- **Big Data**
  - One of 8 research themes at our office
  - Which skills do you need?
  - Examples of our work
  - Lessons learned (so far)

Data, data everywhere

Information has gone from scarce to superabundant.

The Economist





Two kinds of data

Statistics Netherlands

Primary data

Our ‘own’ questionnaires

Secondary data

Data of ‘others’

- Administrative sources
- Big Data

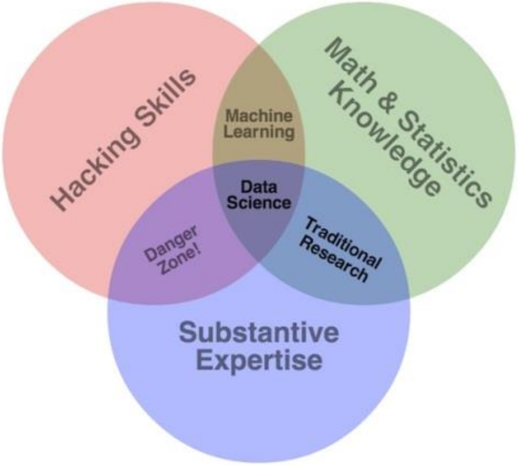
Big Data research at Stat. Netherlands

- Explorative, ‘data driven’
  - Case studies: Road sensors, Mobile phone data, Social media
  - There is now Big Data methodology yet (we are working on it)

- Combination of IT, methodology and Content (Data Science)

- Important topics for official statistics
  - Accessing Big Data in a structural way
  - Selectivity (‘representativity’)
  - Checking and editing large amounts of data
  - Reducing data size (without information loss)

Big Data skills



- We need new skills
  - At the border of IT and methodology
    - Data *Scientists* (a group)
    - High Performance *Analytics*

  - Available in a number of research areas ‘outside’ traditional statistics research
    - Computer sciences, artificial intelligence
    - Machine learning (Statistical learning)
    - Ability to extract information from new sources
      - Such as: texts and pictures/video’s

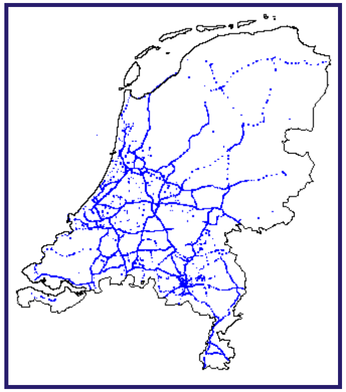
Case studie resultaten

Example 1: Roads sensors

Road sensor (traffic loop) data

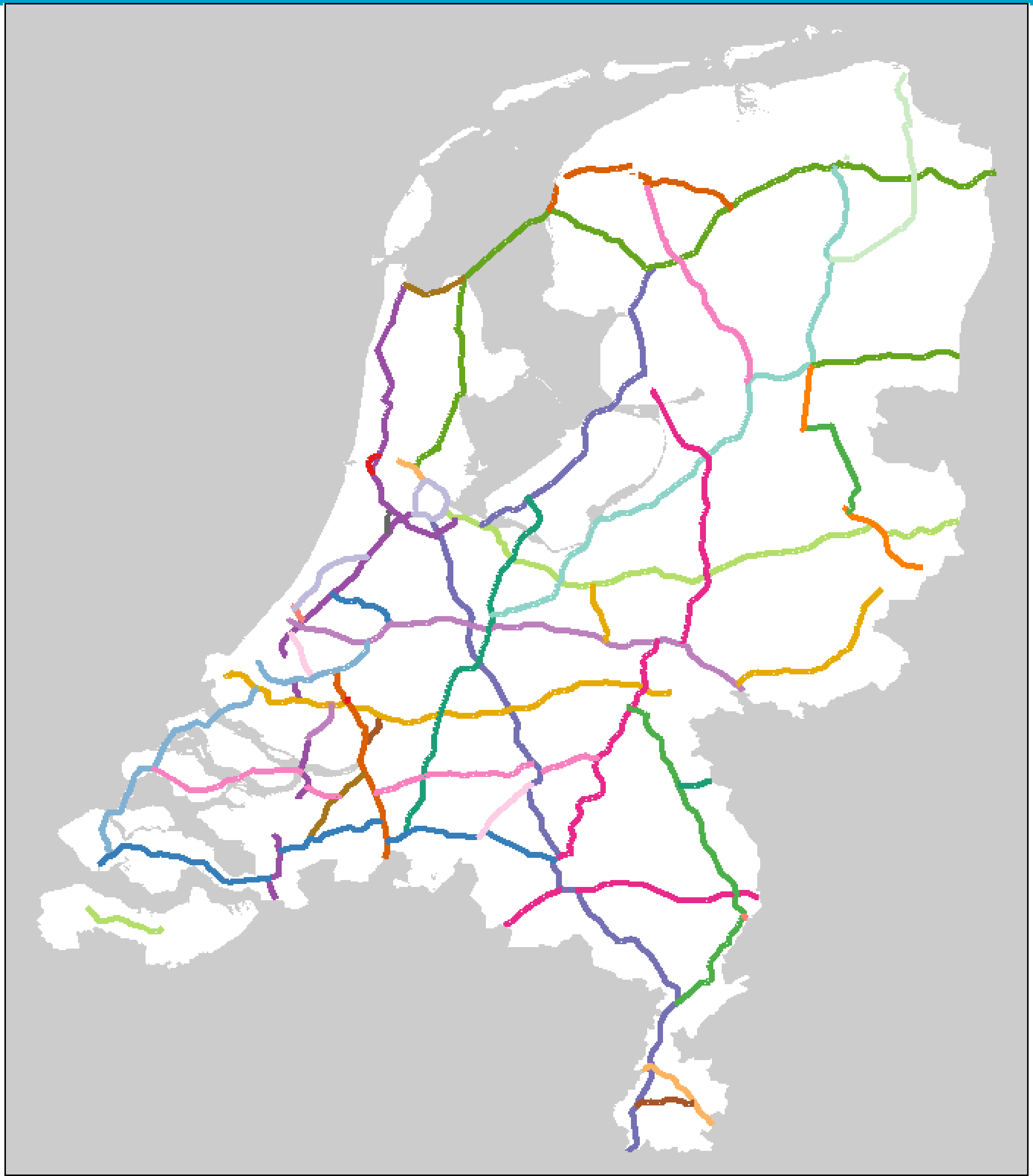
- Each minute (24/7) the number of passing vehicles is counted in around 20.000 'loops' in the Netherlands
  - Total and in different length classes

- Nice data source for transport and traffic statistics (and more)
  - A lot of data, around 230 million records a day

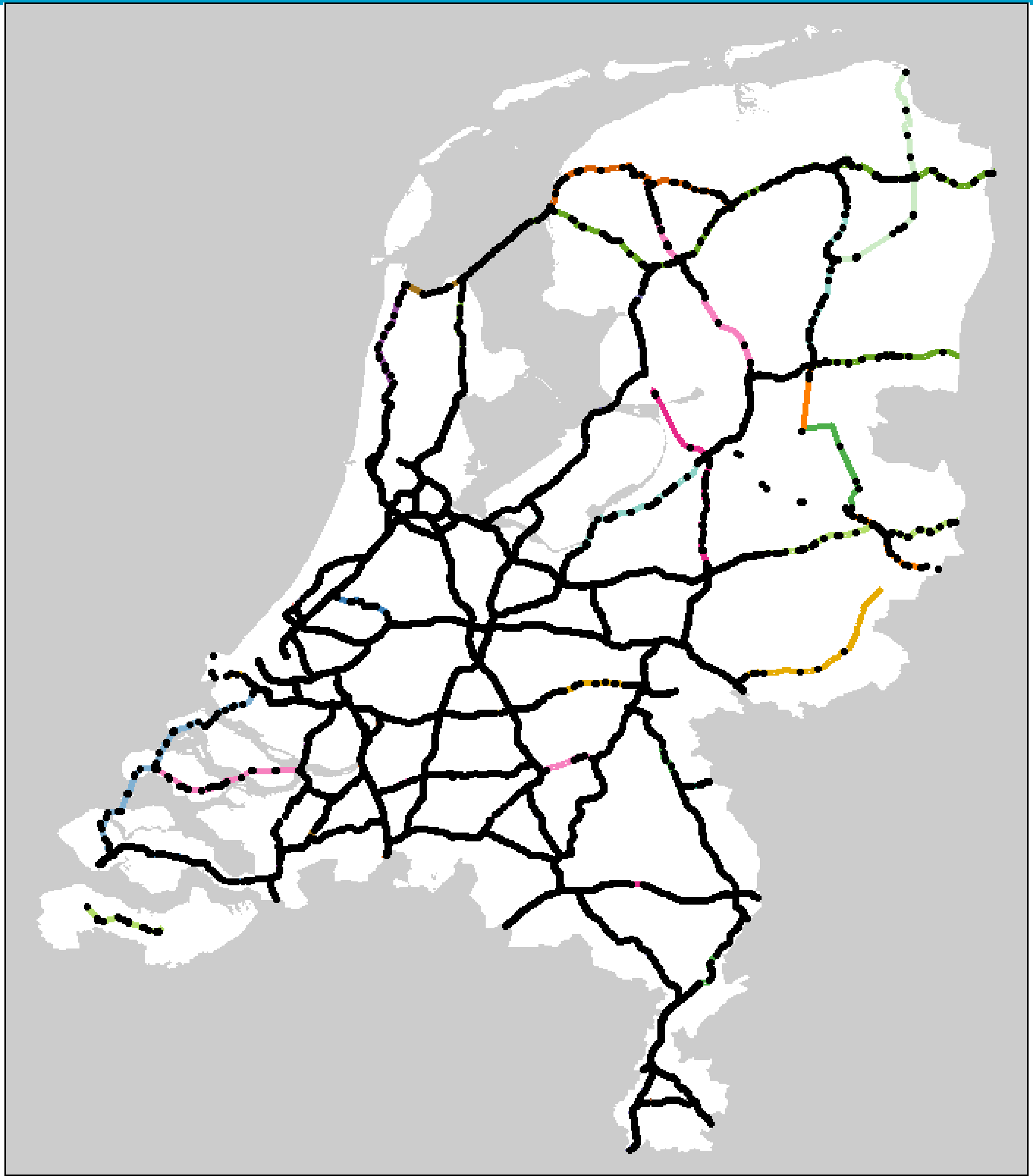


Locations

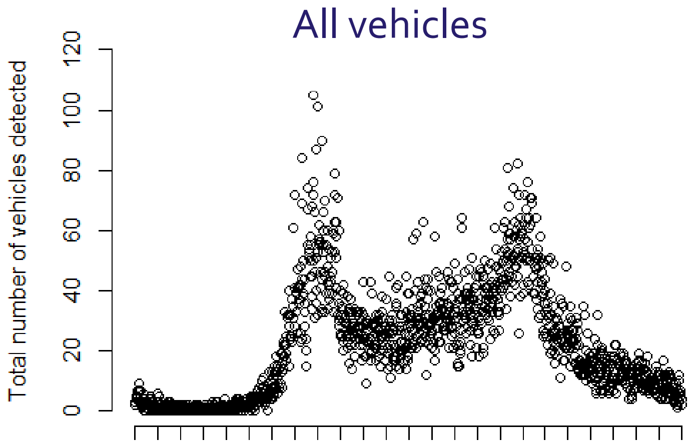
High ways in the Netherlands

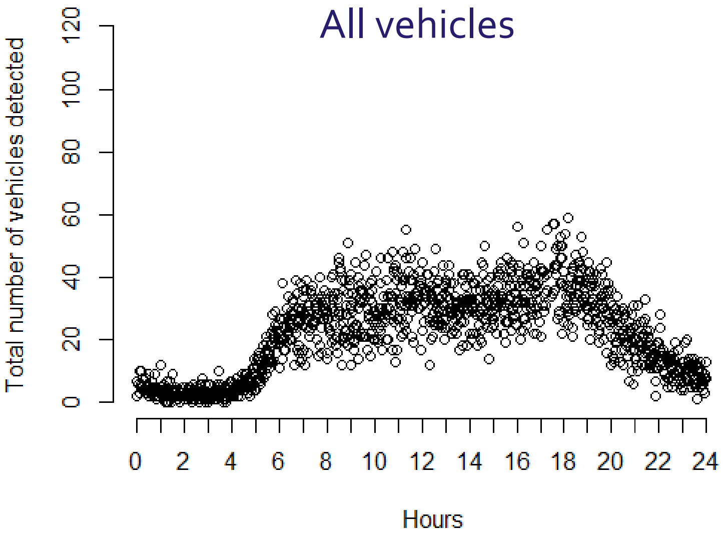


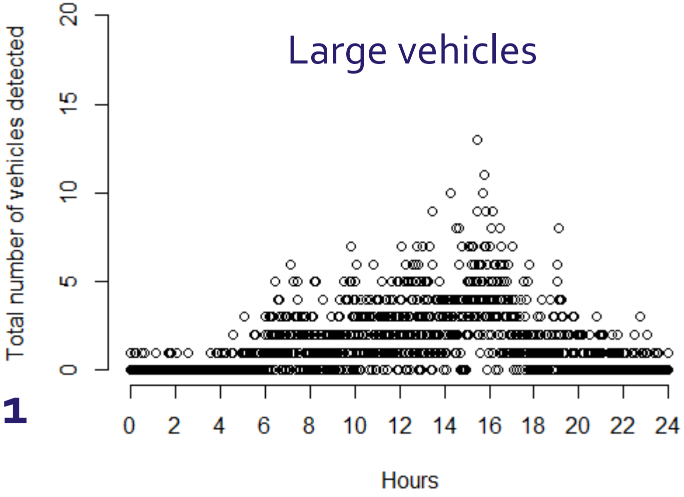
Coverage by road sensors

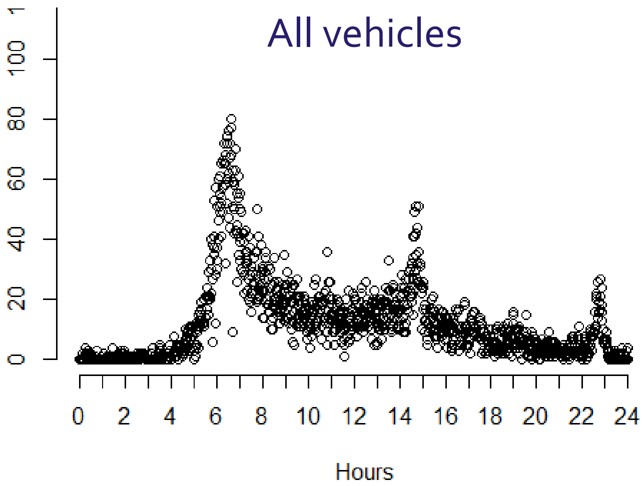


Road sensor data

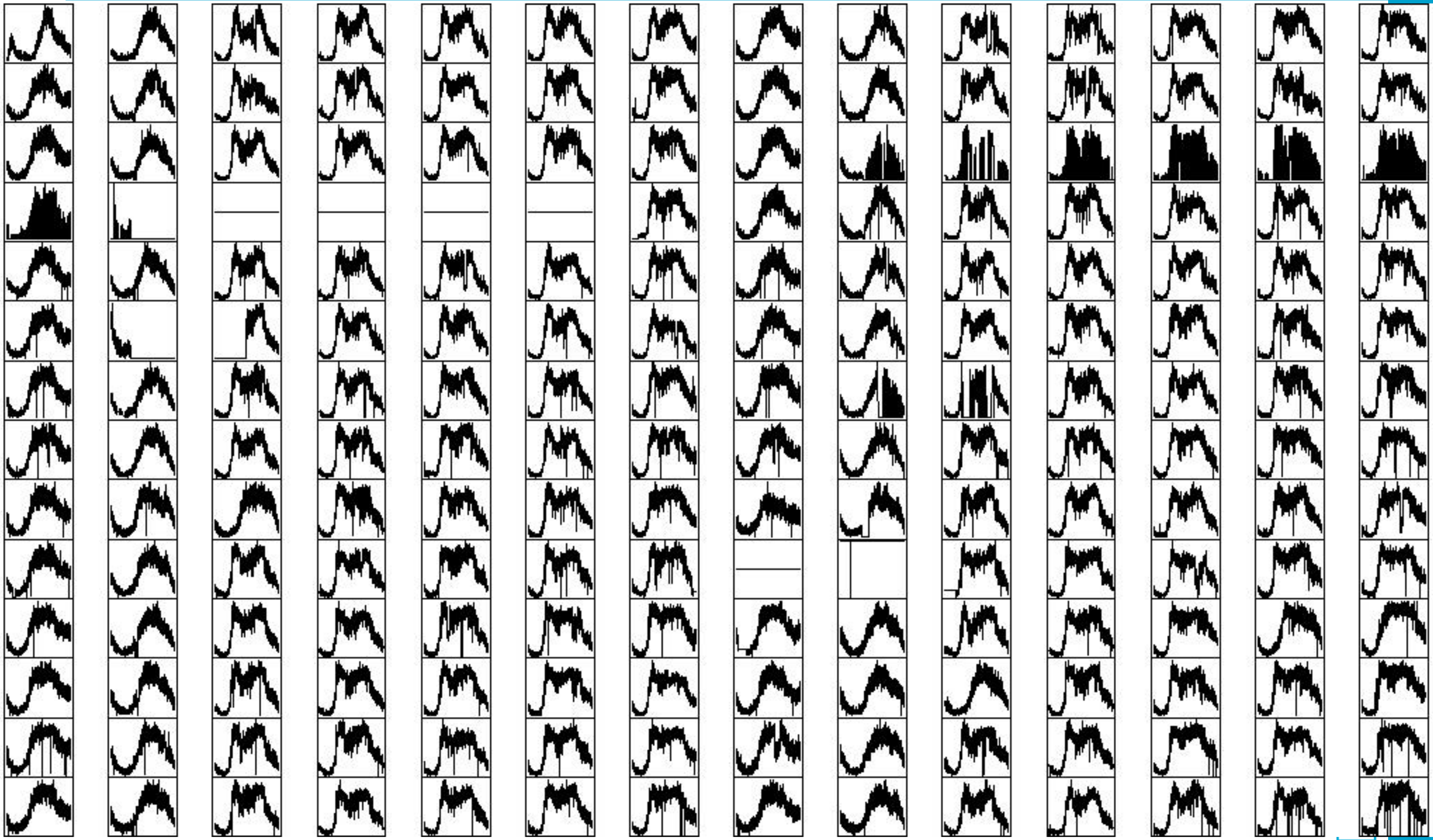




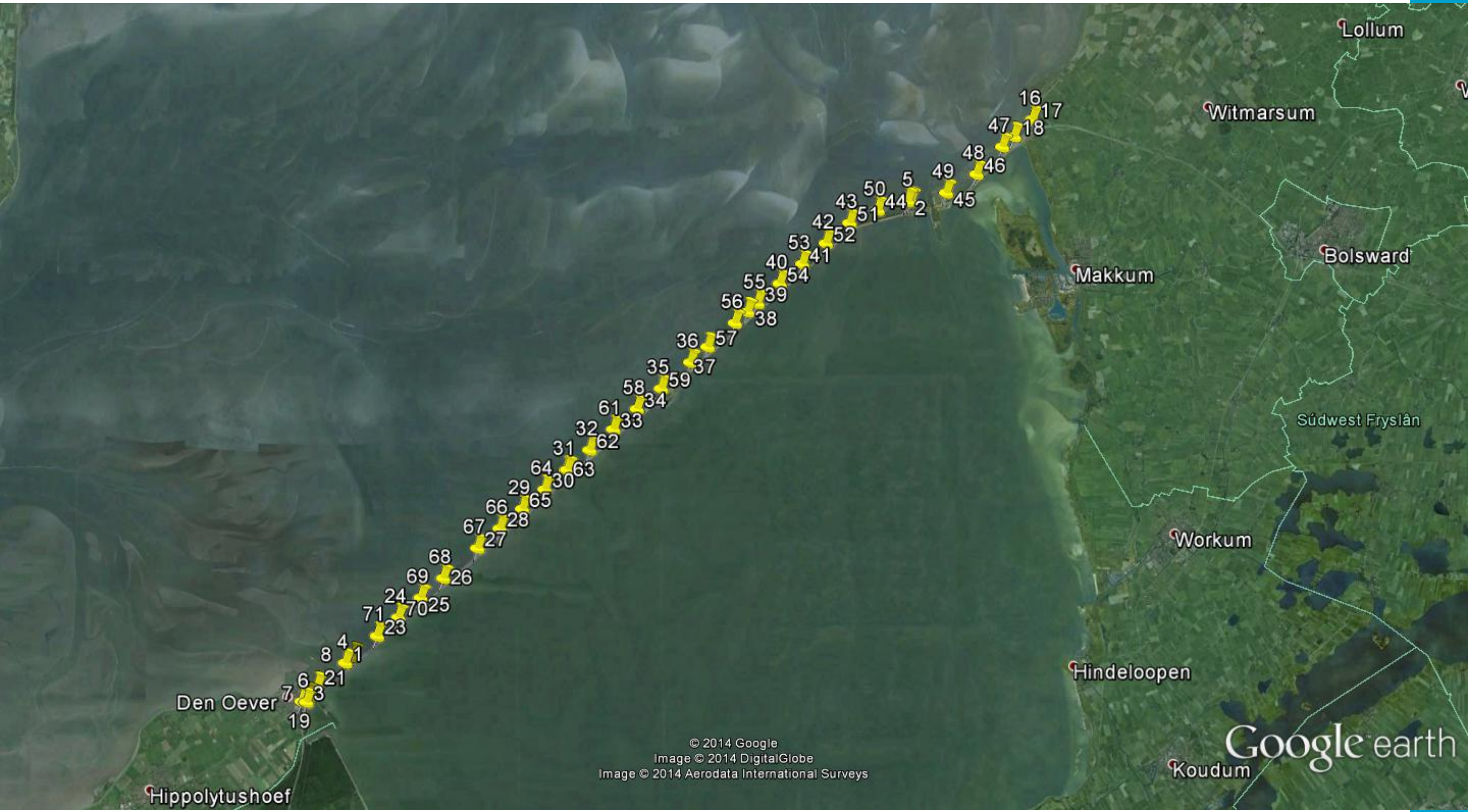




Minute data of 1 sensor for 196 days



‘Afsluitdijk’ (IJsselmeer dam)



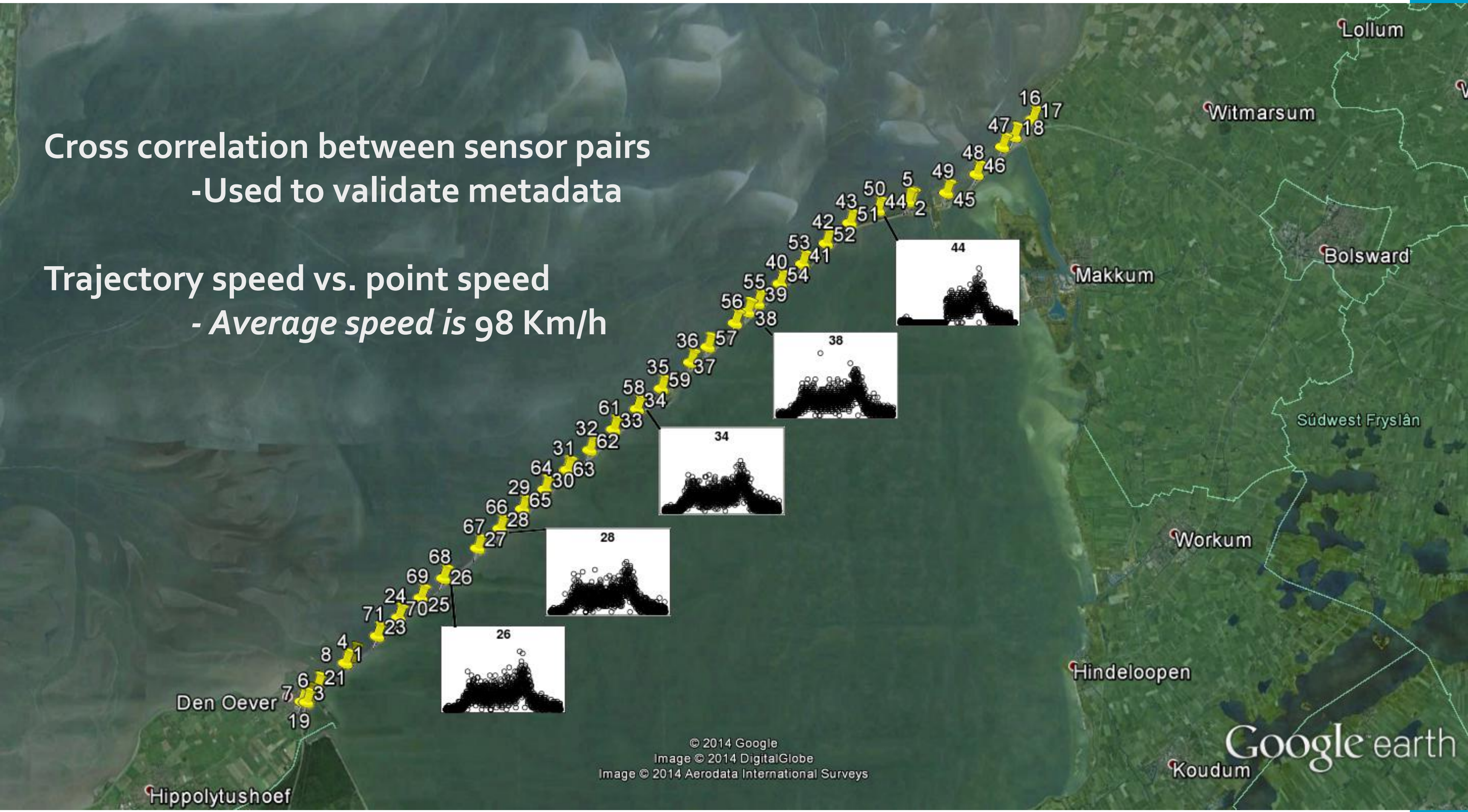
‘Afsluitdijk’ (IJsselmeer dam) (2)

Cross correlation between sensor pairs

-Used to validate metadata

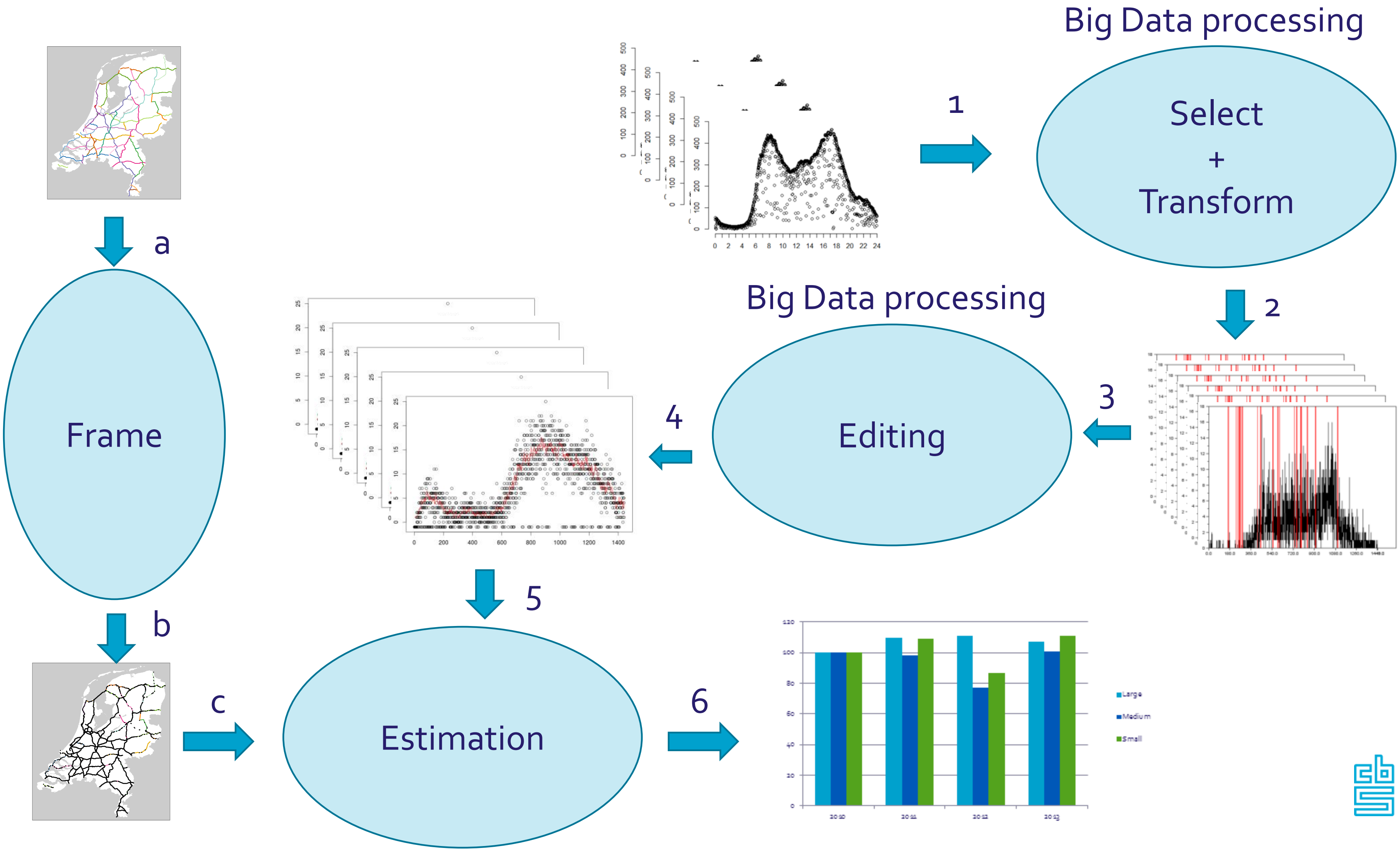
Trajectory speed vs. point speed

- Average speed is 98 Km/h



© 2014 Google
Image © 2014 DigitalGlobe
Image © 2014 Aerodata International Surveys

Process overview



Big Data processing

Select + Transform

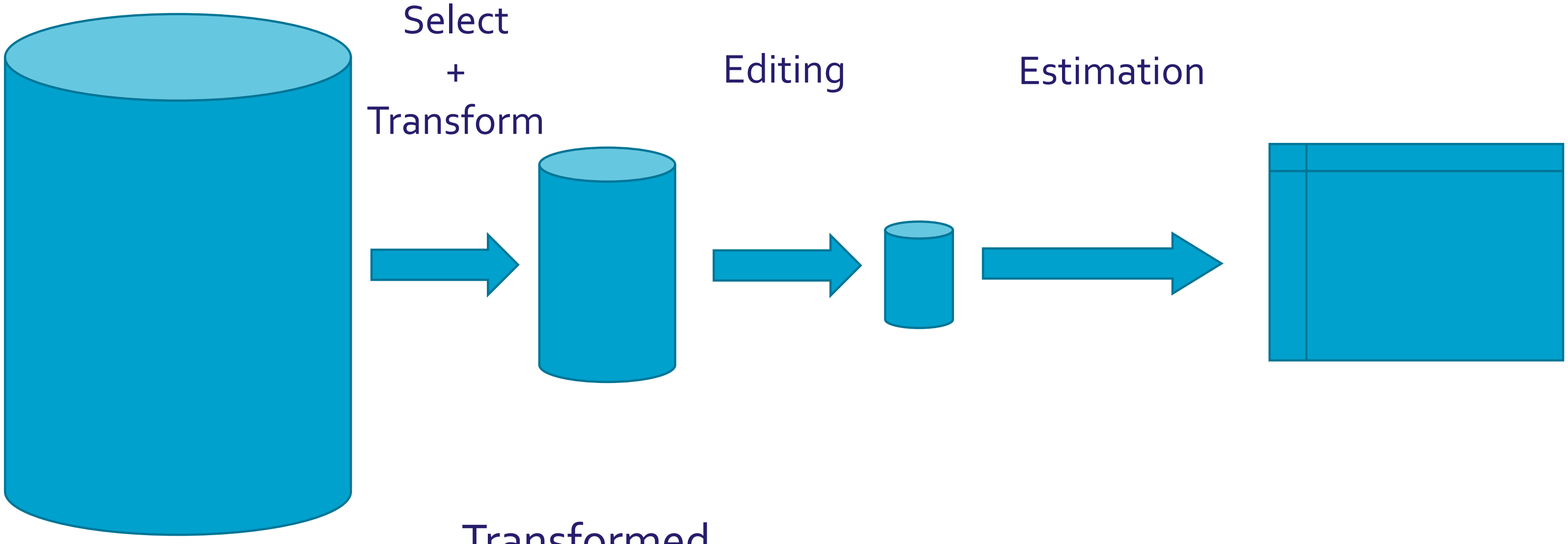
Big Data processing

Editing

Frame

Estimation

Data in process



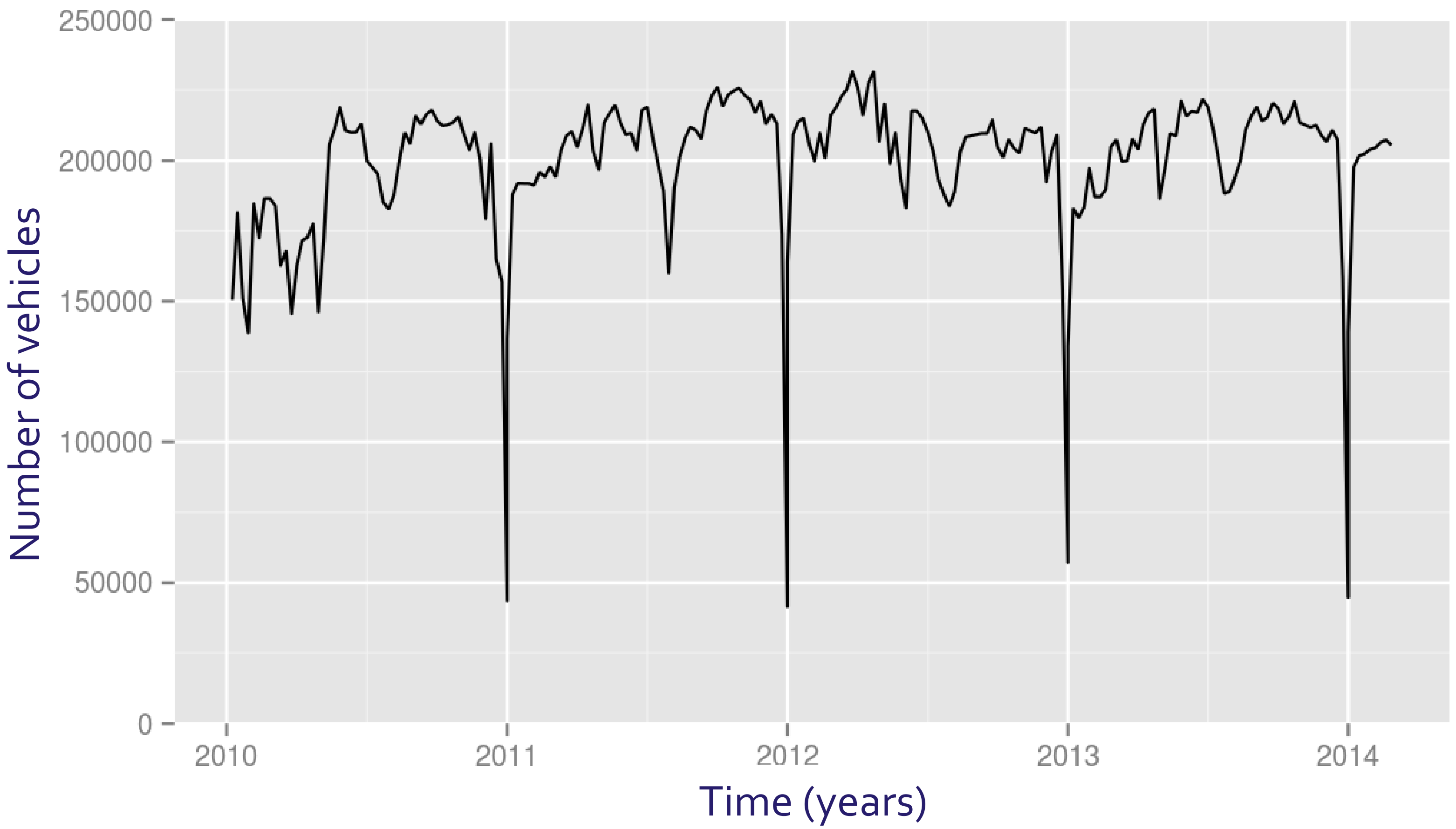
Raw data
80 TB
105 billion records
105 billion data points
2010 - 2014

Transformed data
70 GB
13 million records
15 million data points

Micro data
500 MB
13 million records
13 million data points

Traffic index figures
6 KB

Estimations for the whole country



This is our first Big Data based official publication !!
Findings: tiny.cc/6tno2x   Description: tiny.cc/skno2x

Example 2: Mobile phones

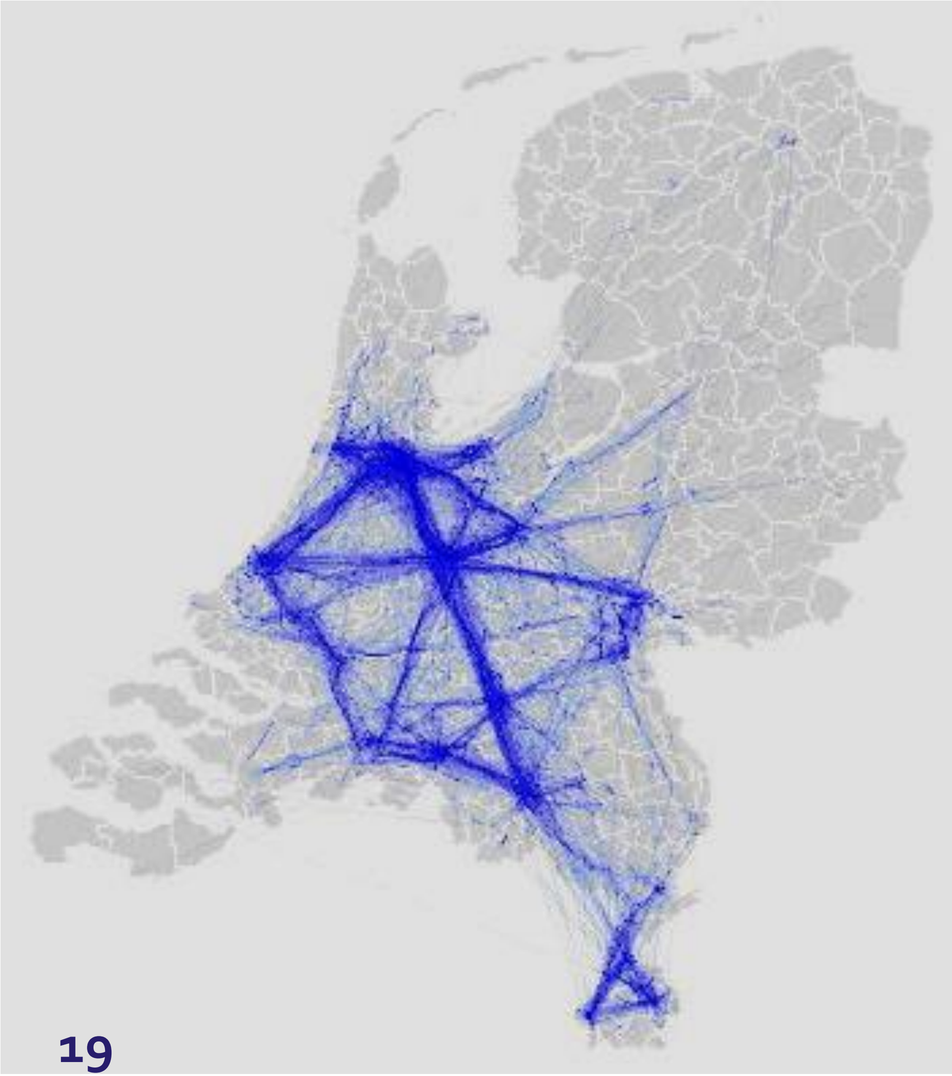
Mobile phone activity as a data source

- Nearly every person in the Netherlands has a mobile phone
  - Usually on them and almost always switched on!
  - Many people are very active during the day

- Can data of mobile phones be used for statistics?
  - *Travel behaviour* (of active phones)
  - '*Day time population*' (of active phones)
  - *Tourism* (new phones that register to network)

- Data of a single mobile company was used
  - Hourly *aggregates* per area (only when > 15 events)
  - Especially important for roaming data (foreign visitors)

Travel behaviour



Mobility of active users

- Anonymized data
- During a fourteen day period

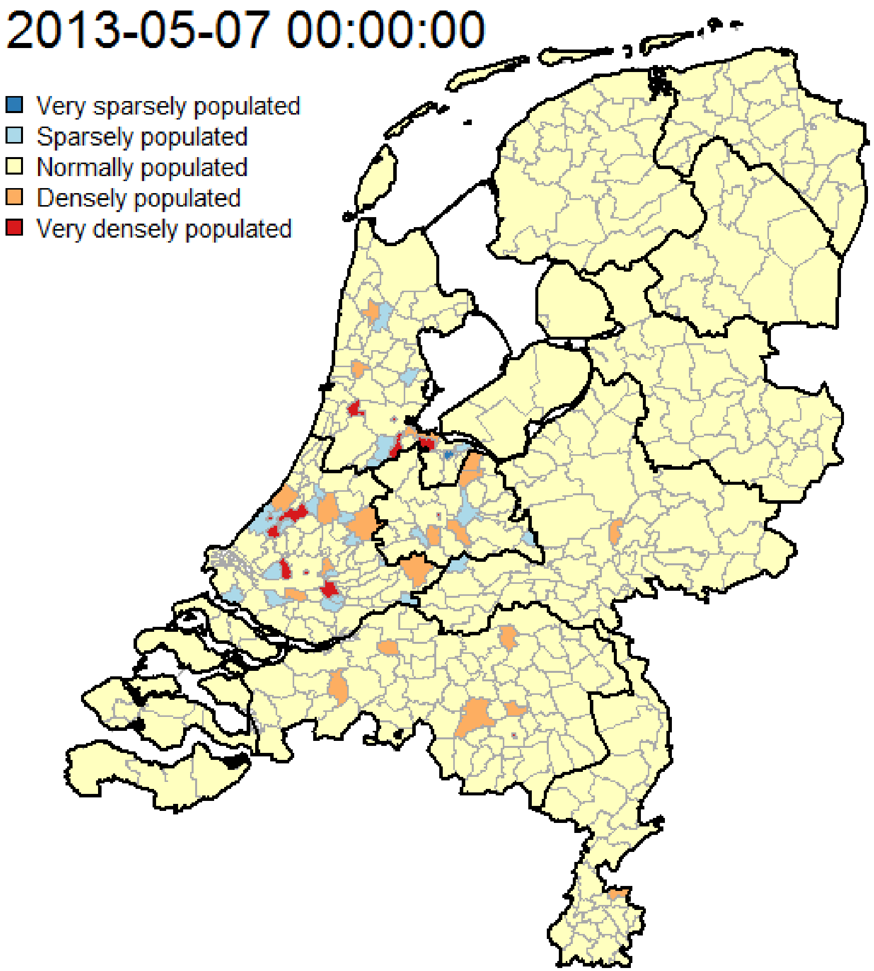
Based on:

- Mobile phone use
- Location of phone mast

What is observed:

- Large Dutch cities
- Hardly any activity in North-East and South-West of the country

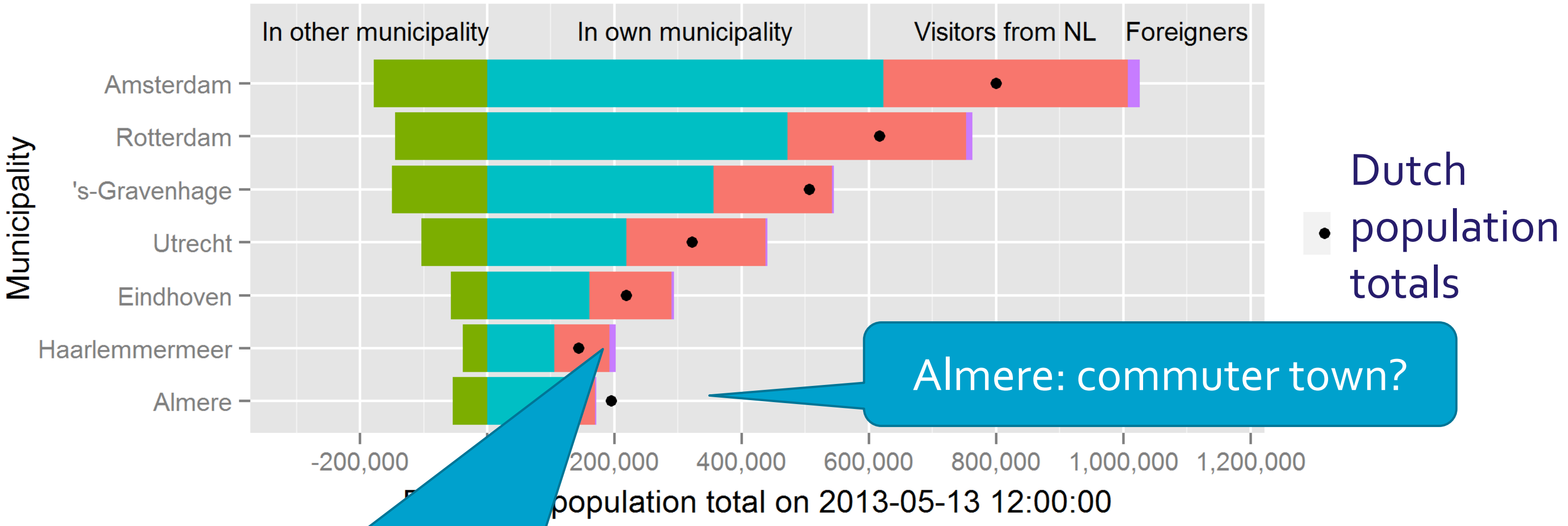
‘Day time population’



2013-05-07 00:00:00

- Hourly changes of mobile phone activity
- 7 & 8 May 2013
- Per area distinguished
- Only data for areas with > 15 events per hour

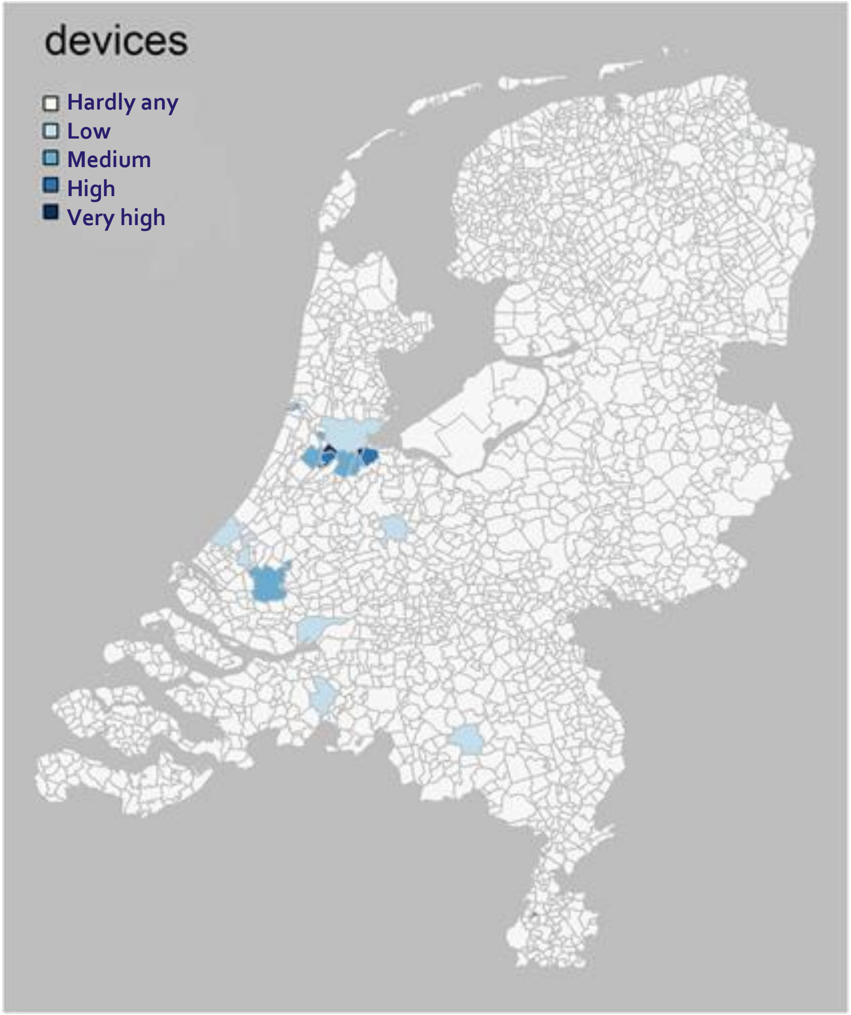
Daytime population results

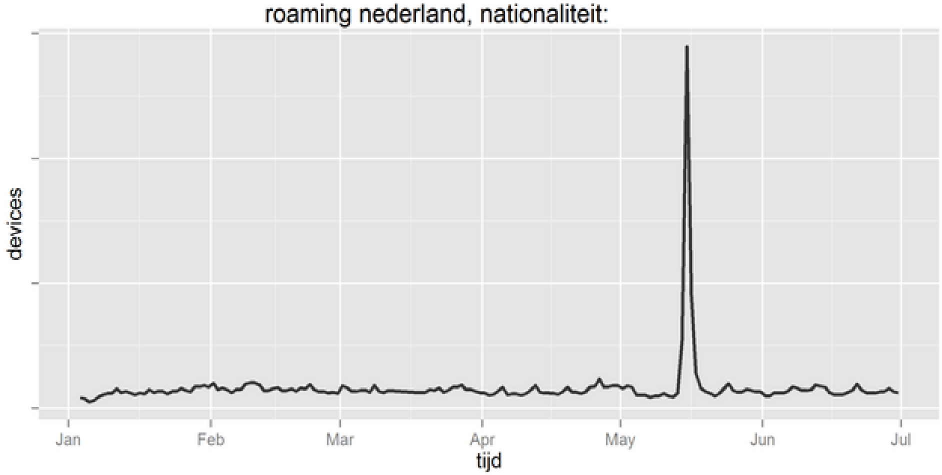


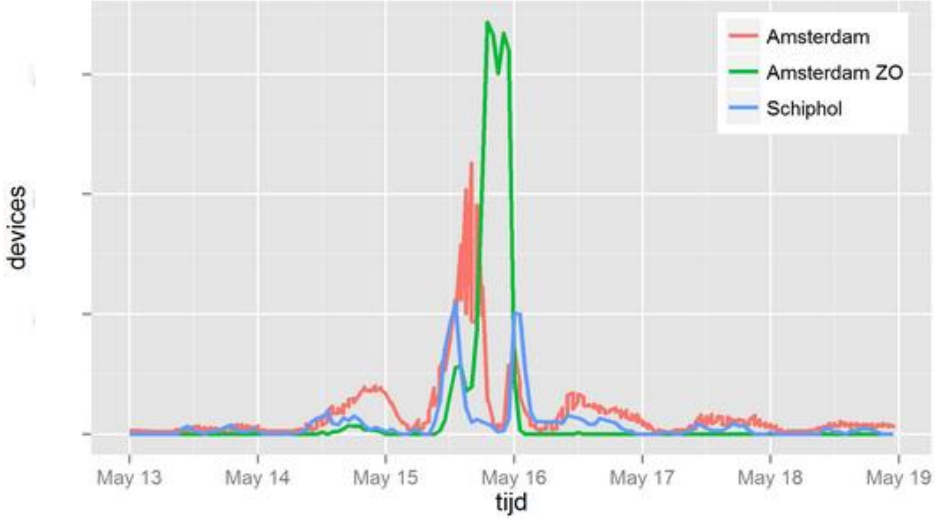
Almere: commuter town?

Foreigners at Schiphol Airport

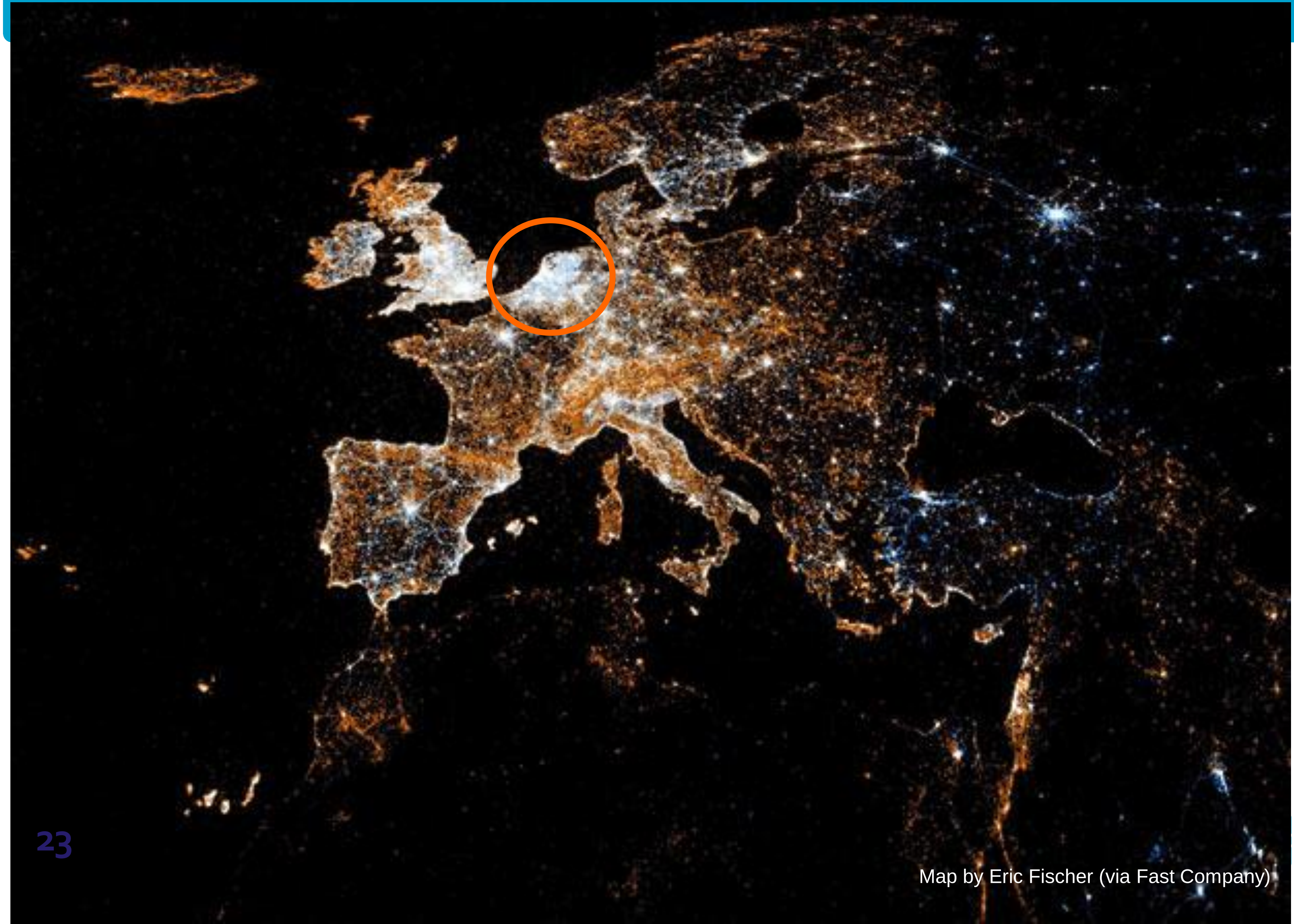
Tourism: Roaming during European league final





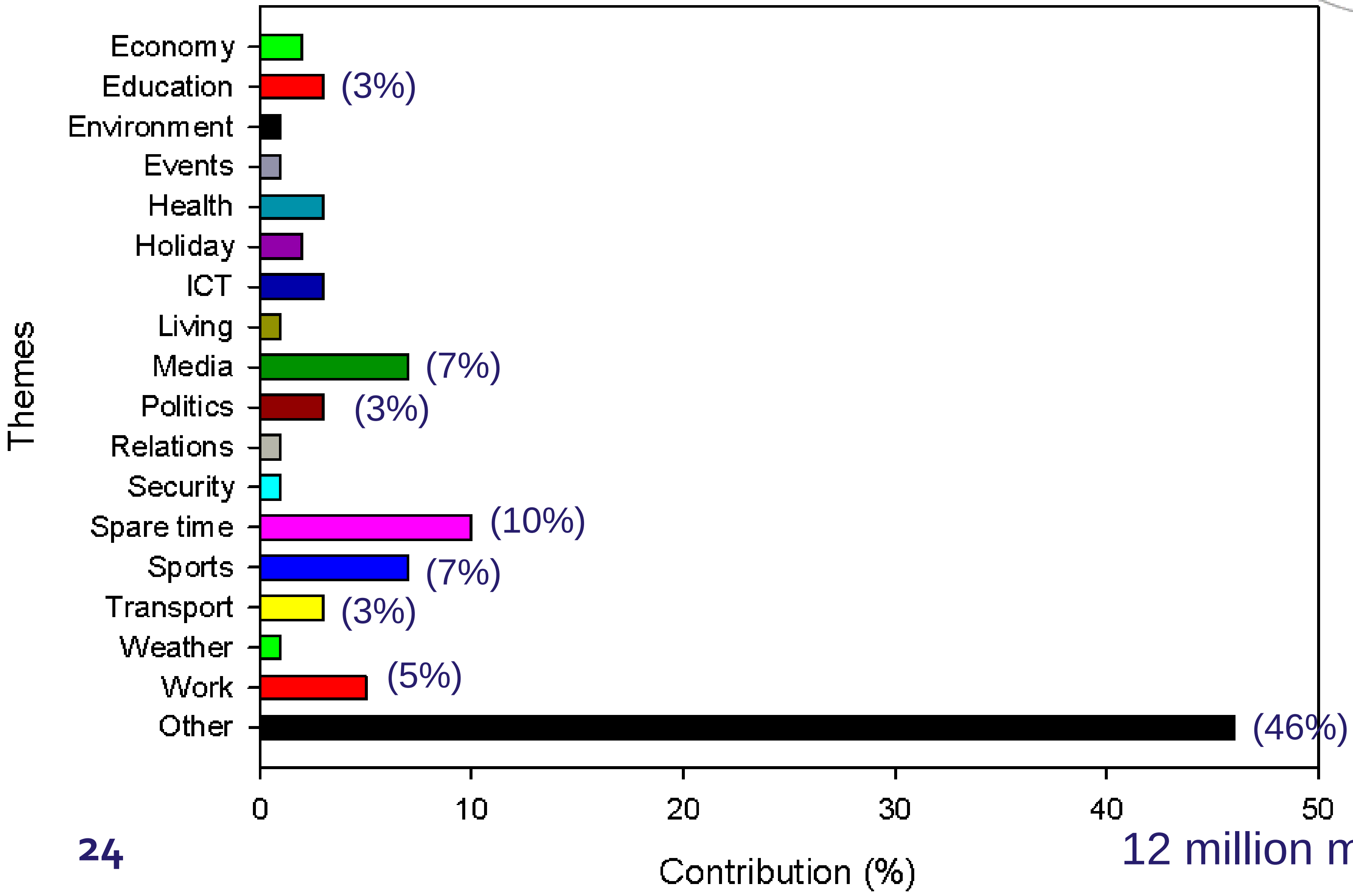


Example 3: Social media



Map by Eric Fischer (via Fast Company)

Dutch Twitter topics

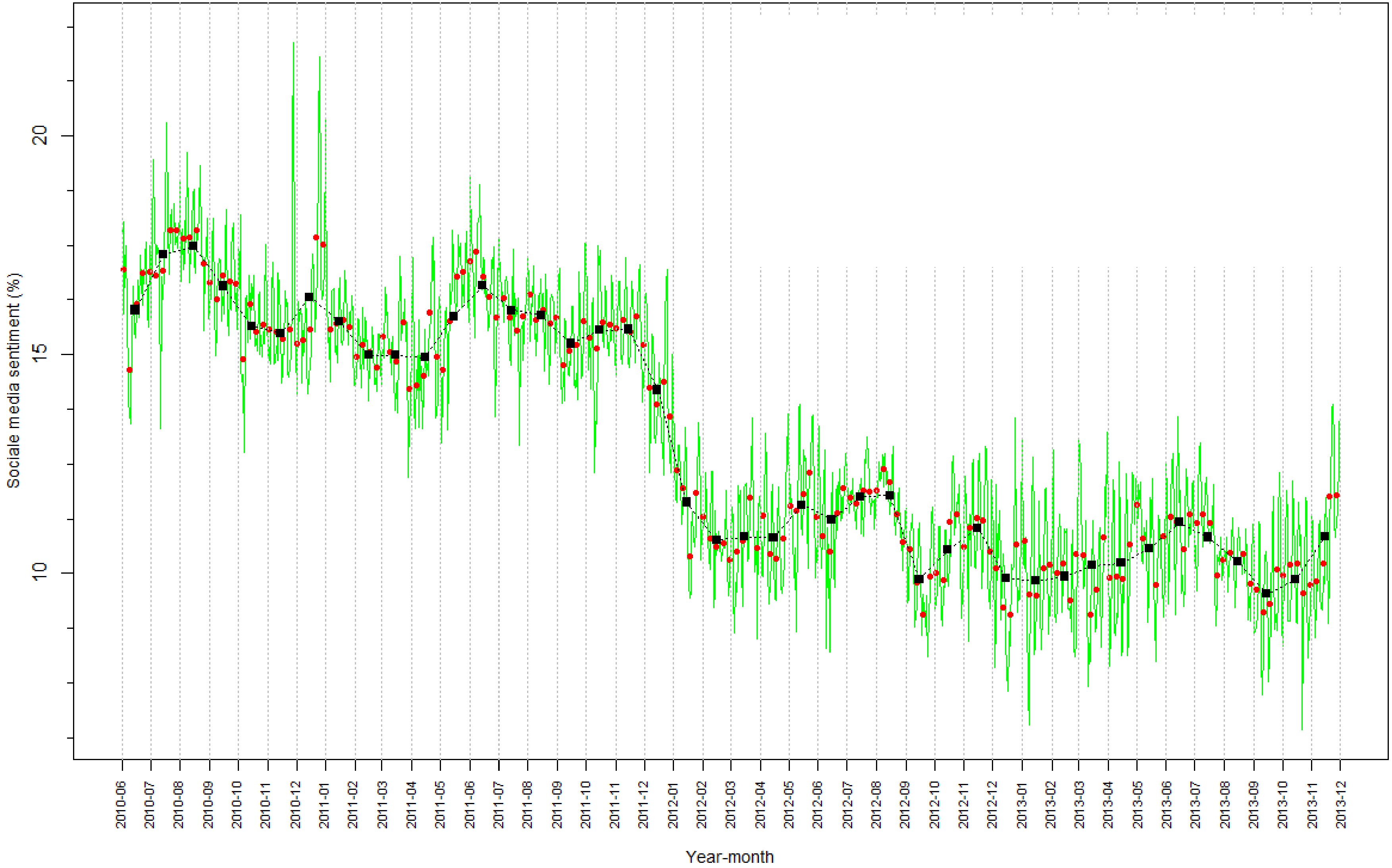


12 million messages

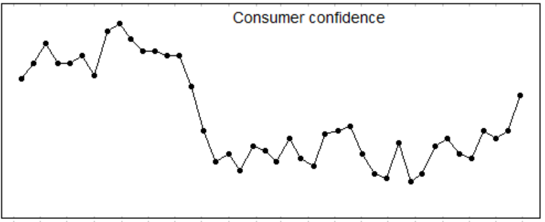
Sentiment in Dutch social media

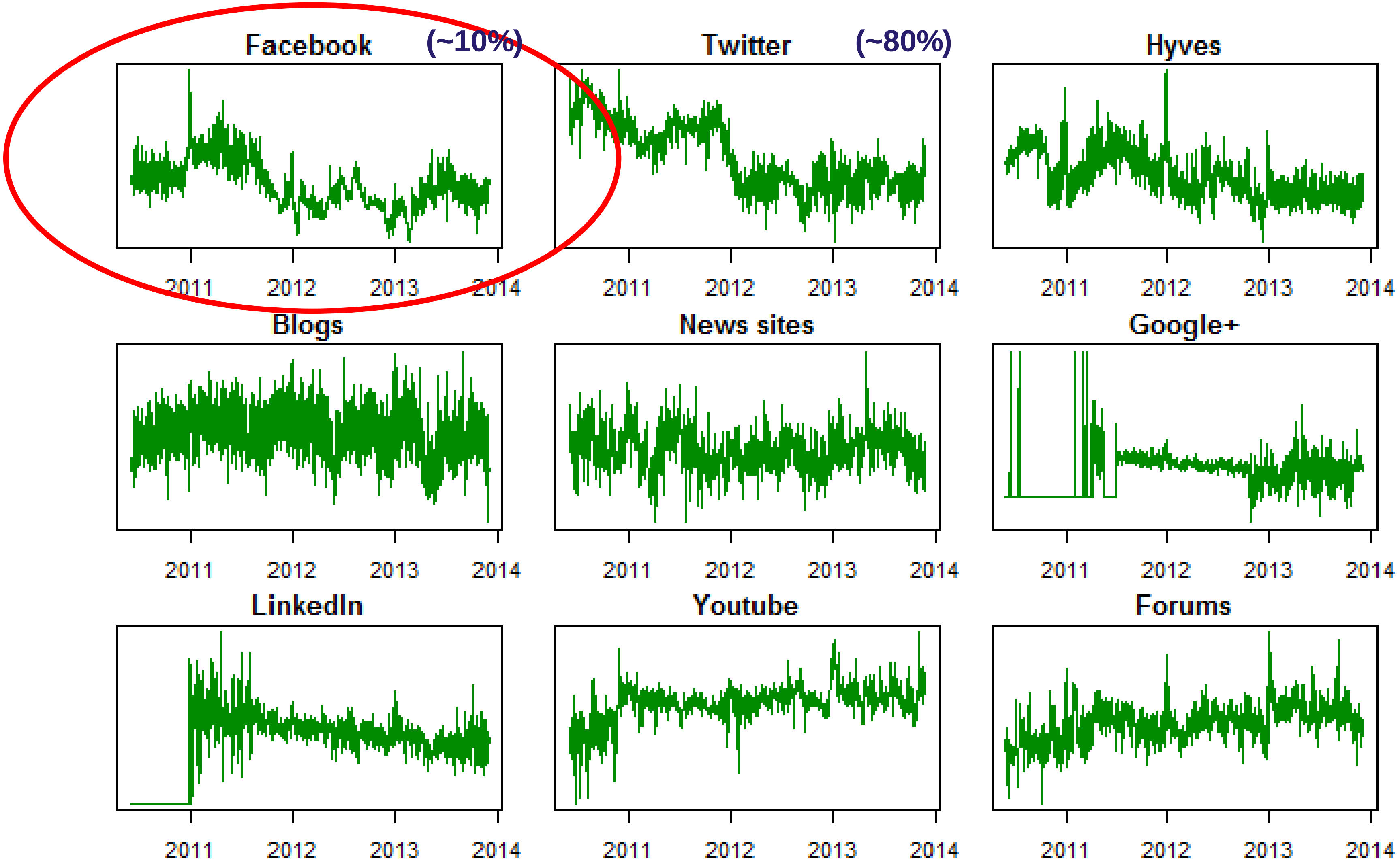
- About the data
    - Dutch firm that continuously collects ALL *public* social media messages written in Dutch
    - Dataset of more than 3.5 billion messages!
        - Covering June 2010 till the present
        - Between 3-4 million new messages are added per day
- About sentiment determination
    - 'Bag of words' approach
        - List of Dutch words with their associated sentiment
        - Added social media specific words ('FAIL', 'LOL', 'OMG' etc.)
    - Use overall score to determine sentiment
        - Is either positive, negative or **neutral**
    - Average sentiment per period (day / week / month)
        - (#positive - #negative)/#total * 100%

Daily, weekly, monthly sentiment



Sentiment per platform





Platform specific results

Table 1. Social media messages properties for various platforms and their correlation with consumer confidence

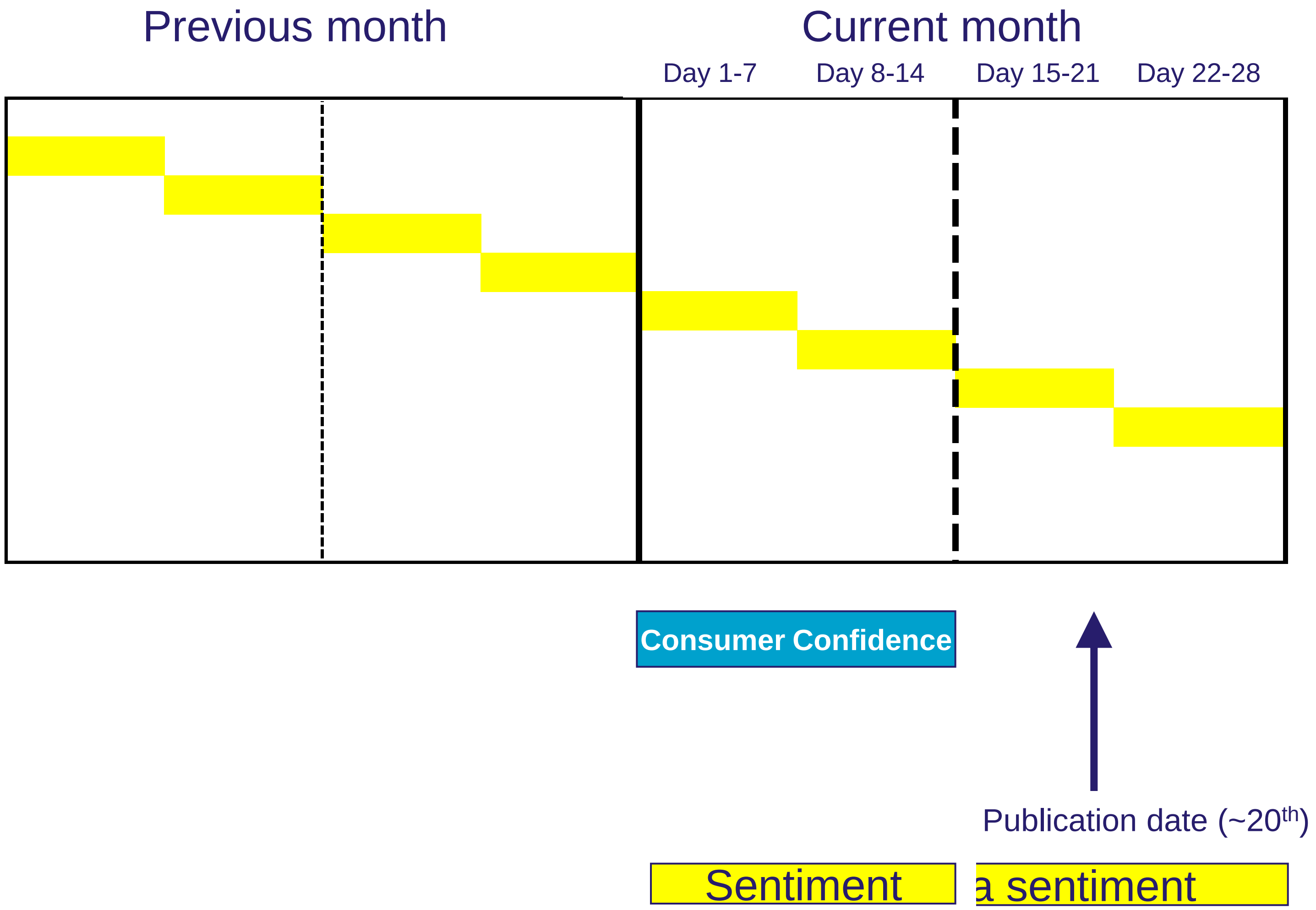
Social media platform	Number of social media messages[1]	Number of messages as percentage of total (%)	Correlation coefficient of monthly sentiment index and consumer confidence (r)[2]
All platforms combined	3,153,002,327	100	0.75
Facebook	334,854,088	10.6	0.81*
Twitter	2,526,481,479	80.1	0.68
Hyves	45,182,025	1.4	0.50
News sites	56,027,686	1.8	0.37
Blogs	48,600,987	1.5	0.25
Google+	644,039	0.02	-0.04
Linkedin	565,811	0.02	-0.23
Youtube	5,661,274	0.2	-0.37
Forums	134,98,938	4.3	-0.45

[1] period covered June 2010 untill November 2013

[2] confirmed by visual inspecting scatterplots and additional checks (see text)

*cointegrated

Schematic overview



Previous month

Current month

Day 1-7

Day 8-14

Day 15-21

Day 22-28

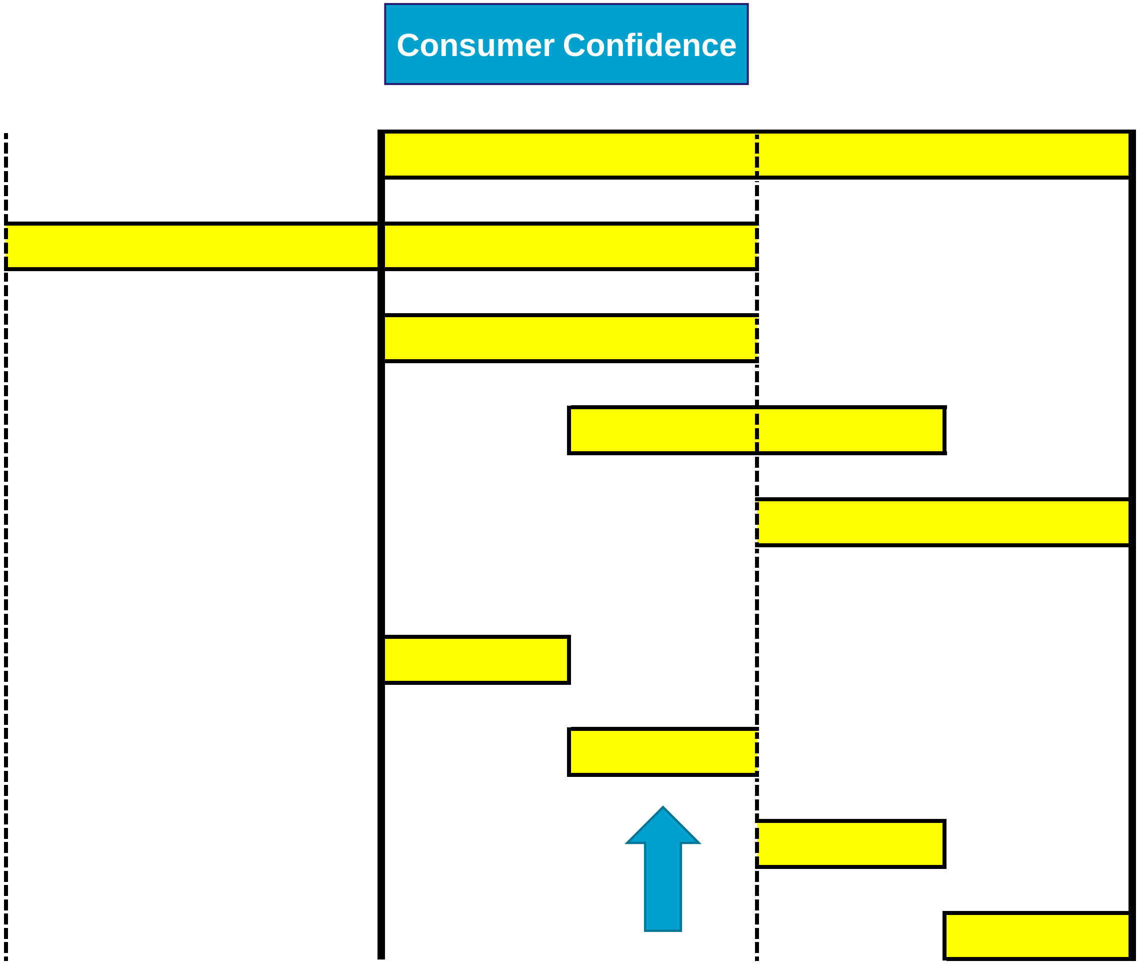
Consumer Confidence

Publication date (~20th)

Sentiment

a sentiment

Results of comparing various periods



Facebook

0.81*

0.85*

0.82

0.82*

0.79*

0.79

0.82*

0.79*

0.75*

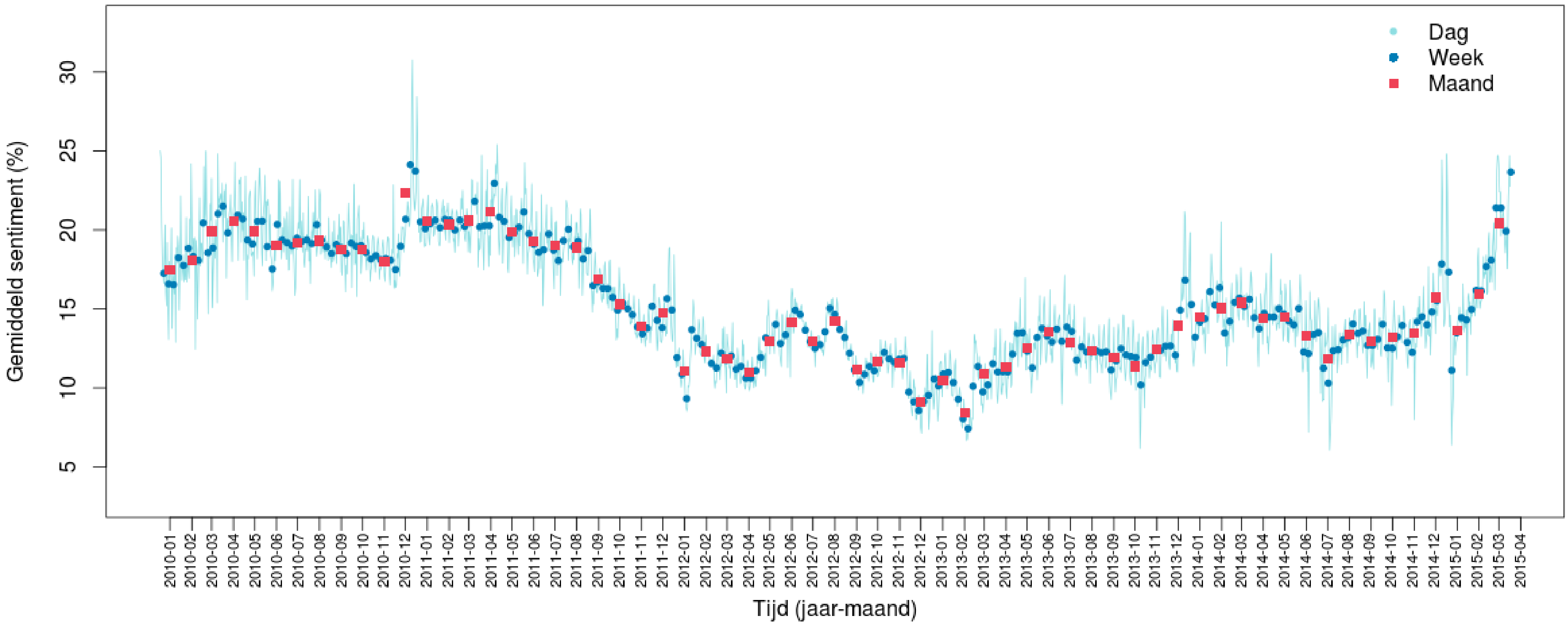
*cointegration

LOOCV results

Overall findings

- Correlation and cointegration
  - 1st 'week' of Consumer confidence usually has 70% response
  - Best correlation and cointegration with 2nd 'week' of the month
    - Highest correlation 0.93* (all Facebook * specific word filtered Twitter)
- Granger causality
  - Changes in Consumer confidence *precede* changes in Social media sentiment
  - For all combinations shown!
    - Only tried linear models (so far)
- Prediction
  - Slightly better than random chance
  - Best for the 4th 'week' of month

‘Sentiment’ indicator voor NL (beta-versie)



Gebaseerd op het gemiddelde sentiment van *publieke* NL-talige Facebook en Twitter berichten

Lessons learned

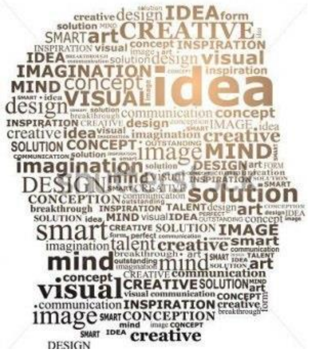
The most important ones:

1) There are many *types* of Big Data
2) Know how to access and analyse *large amounts* of data
3) Find ways to *deal with noisy* and unstructured data
4) *Mind-set* for Big Data ≠ Mind-set for survey data
5) Need to *beyond* correlation
6) Need people with the *right skills*, knowledge and mind-set
7) Solve *privacy* and security issues
8) Data management & *costs*

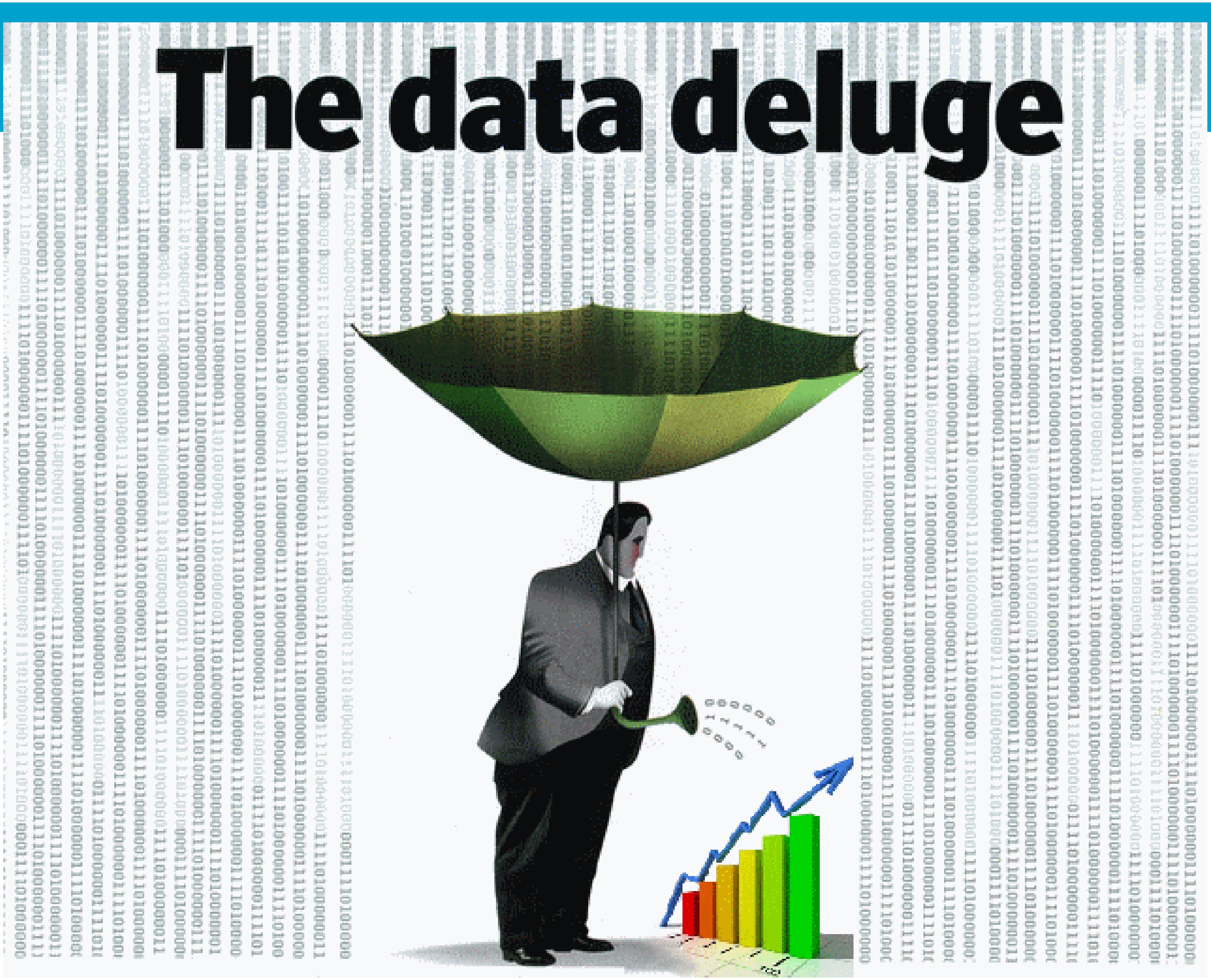
We are getting more and more grip on these topics
& we published our first Big Data based official statistics!!

The Future

The future of statistics looks



BIG



@pietdaas

Thank you for your attention !