Q 2016
European Conference on Quality in Official Statistics

Using huge amounts of road sensor data for official statistics

Session 20

2.6.2016

Marco Puts, Piet Daas

Martijn Tennekes, Chris de Blois

Statistics Netherlands

INE Instituto Nacional de Estadística eurostat

Madrid, May 31 - June 3

Information value of Big Data

CBS

Sample Survey

Questionairs

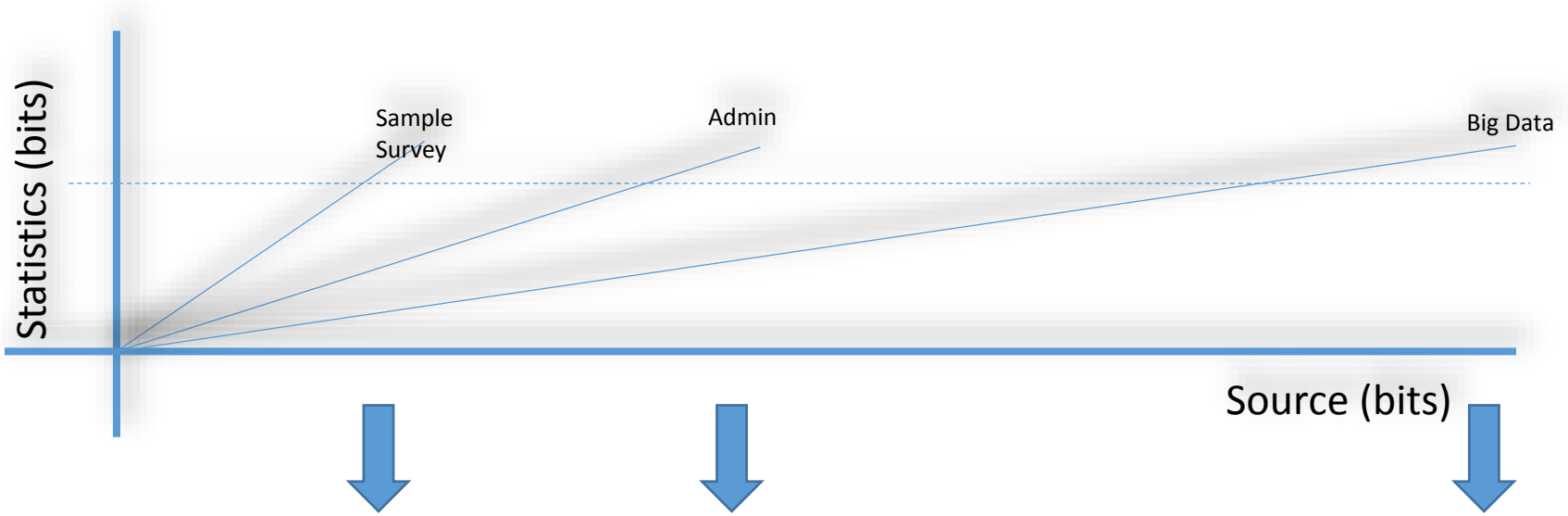
Secondary data

Data of ‘others’

- Administrative sources
- Big Data

Information value of Big Data

How much source data do you need?



Sample Issues

Source issues
Conversion issues

Source issues
Conversion issues
Selectivity issues
Data issues

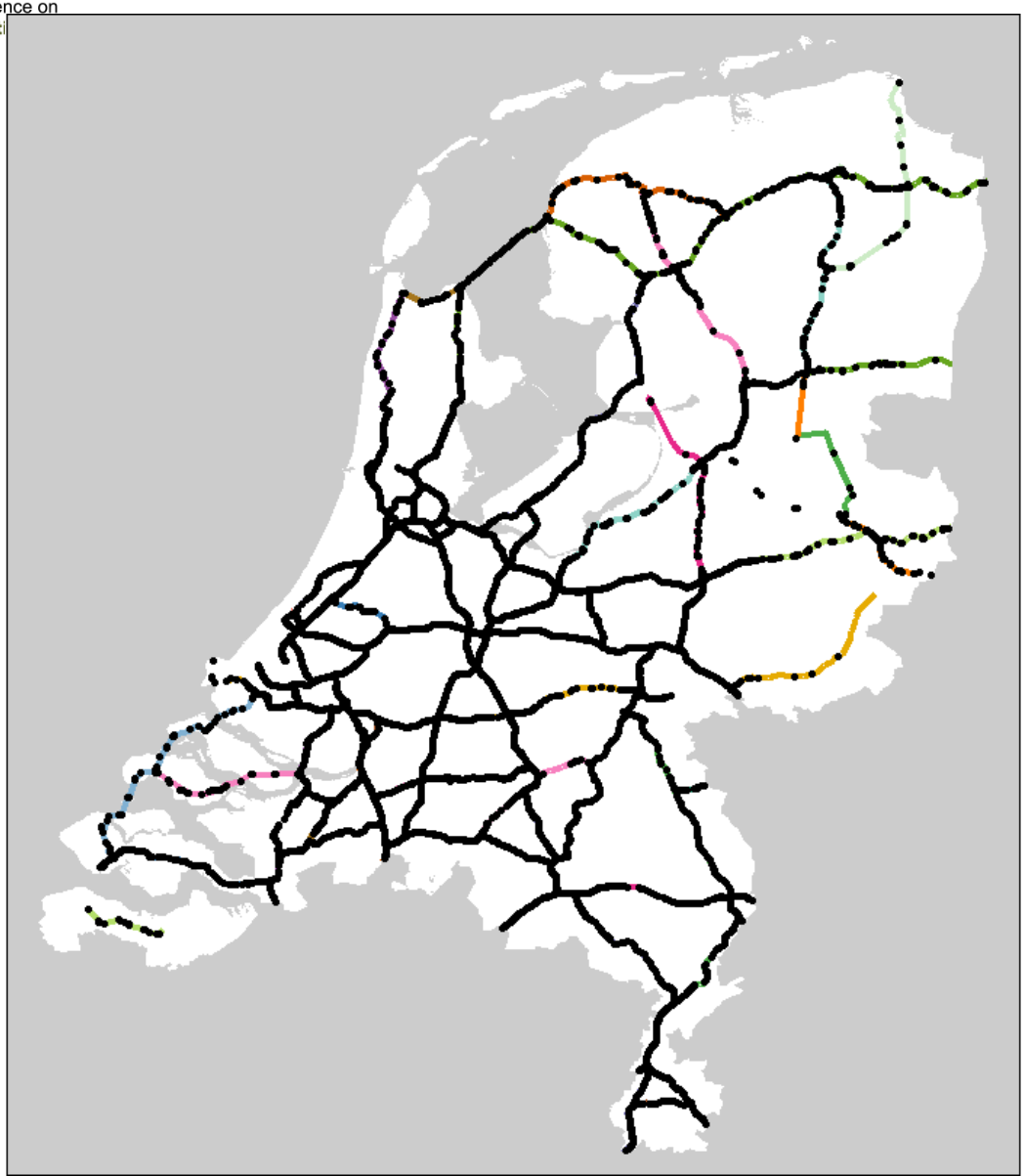
Road sensors

Road sensor data

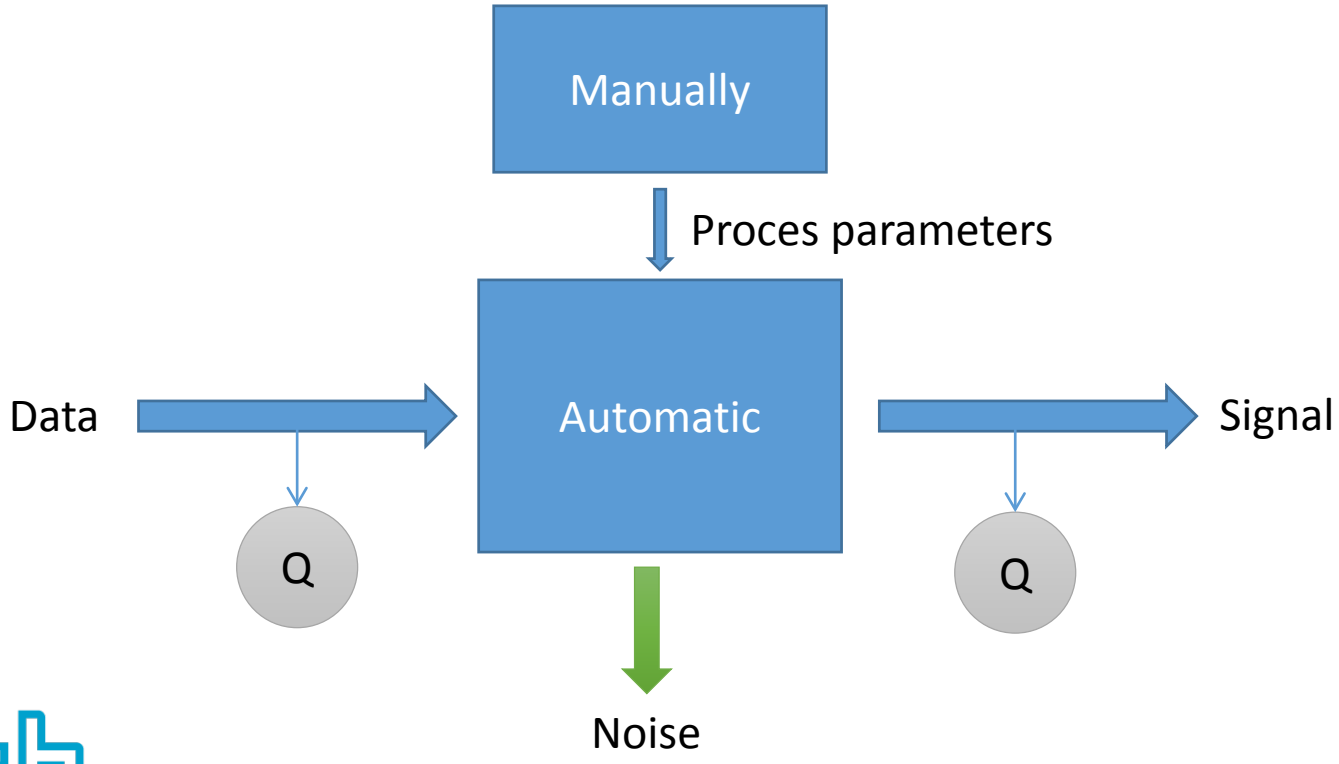
- Passing vehicle counts for each minute (24/7) at about 60.000 sensors in the Netherlands
- Types of sensors:
  - Induction loop
  - Camera
  - Bluetooth
- Length categories (e.g. small, medium, long vehicles)
- Large volume: approx. 230 mln records/day

Dutch highways

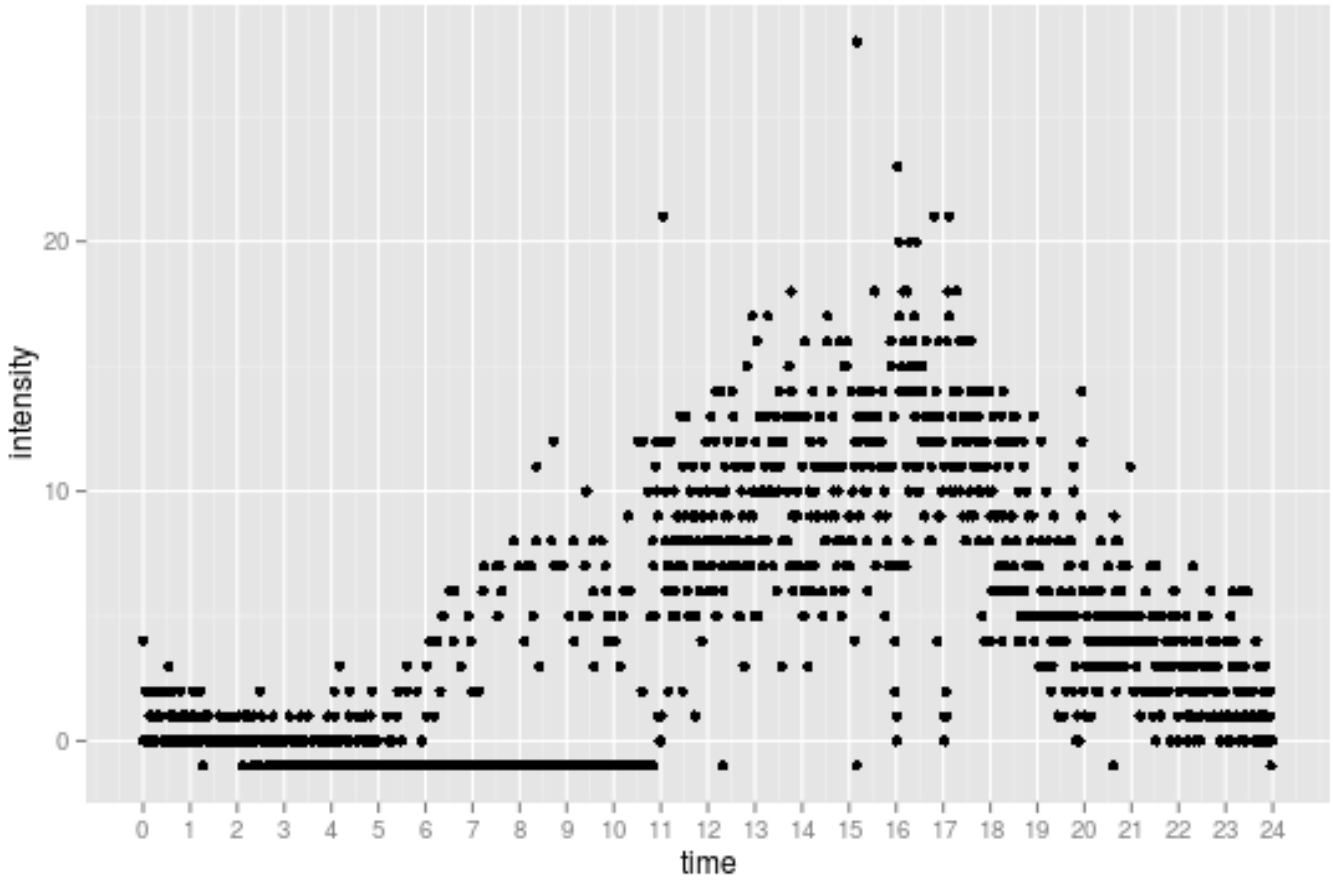
Dutch highways



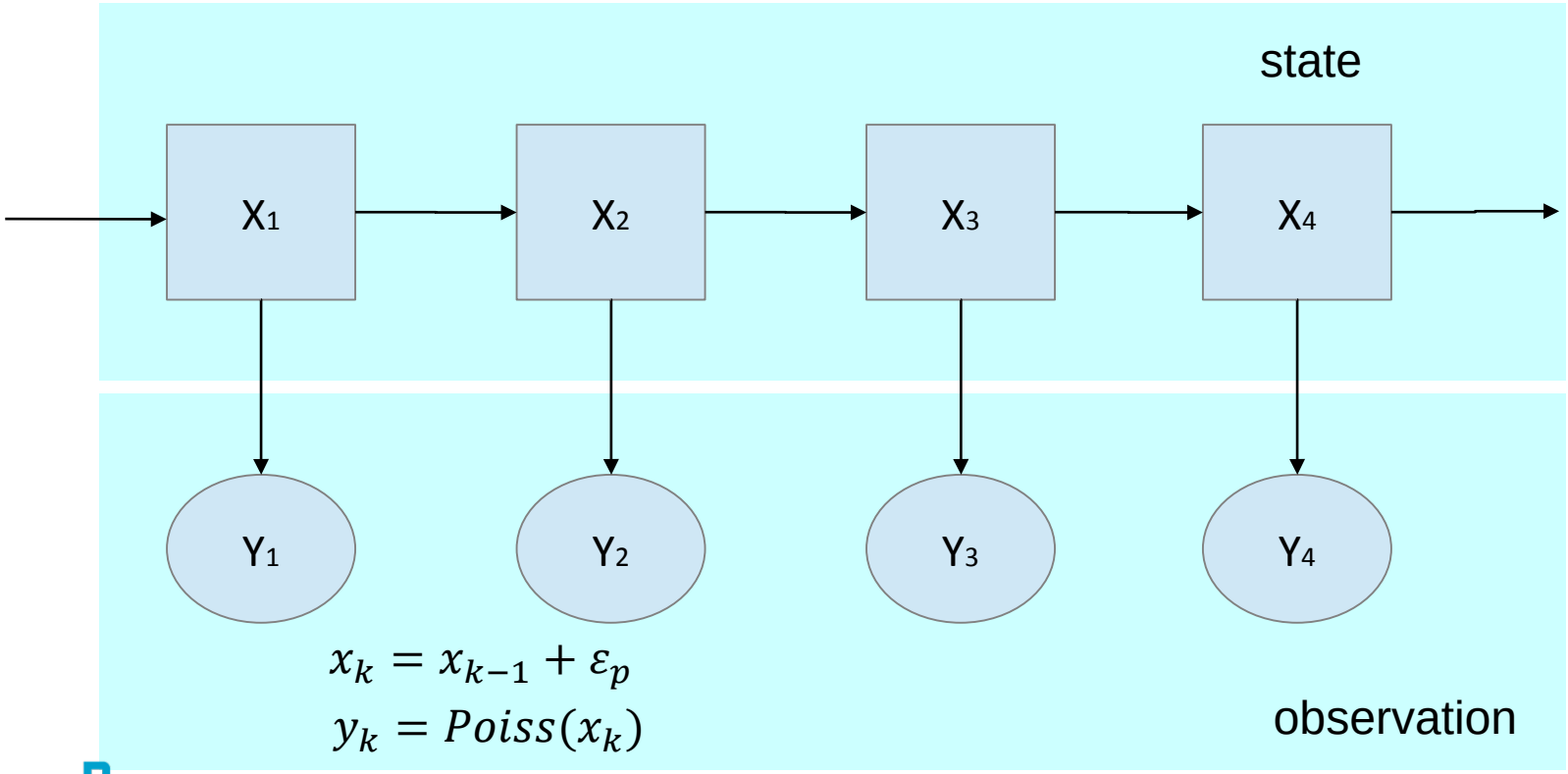
The Signal and the Data



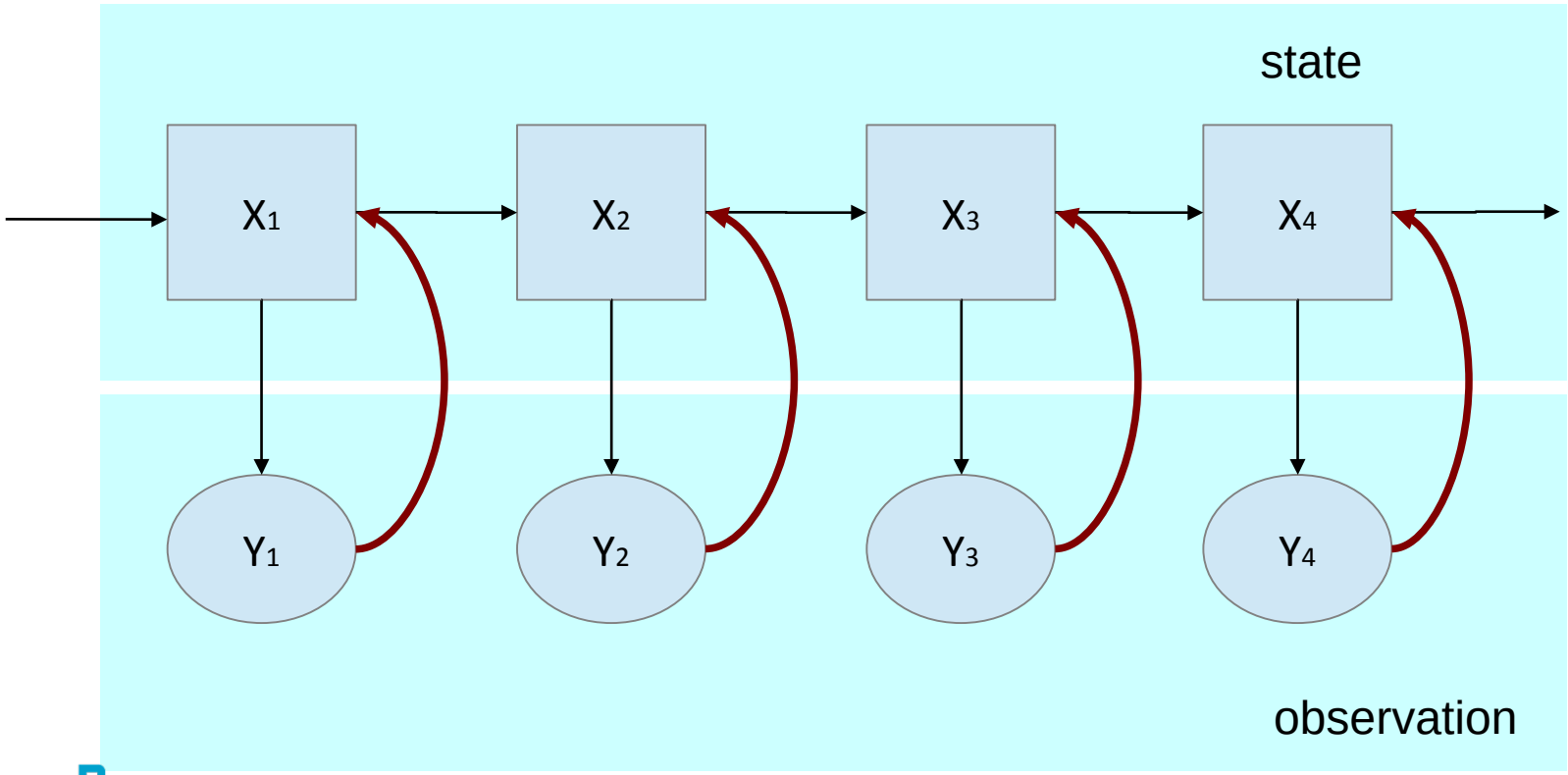
Quality of the Data



Cleaning the Data
Recursive Bayesian Estimation

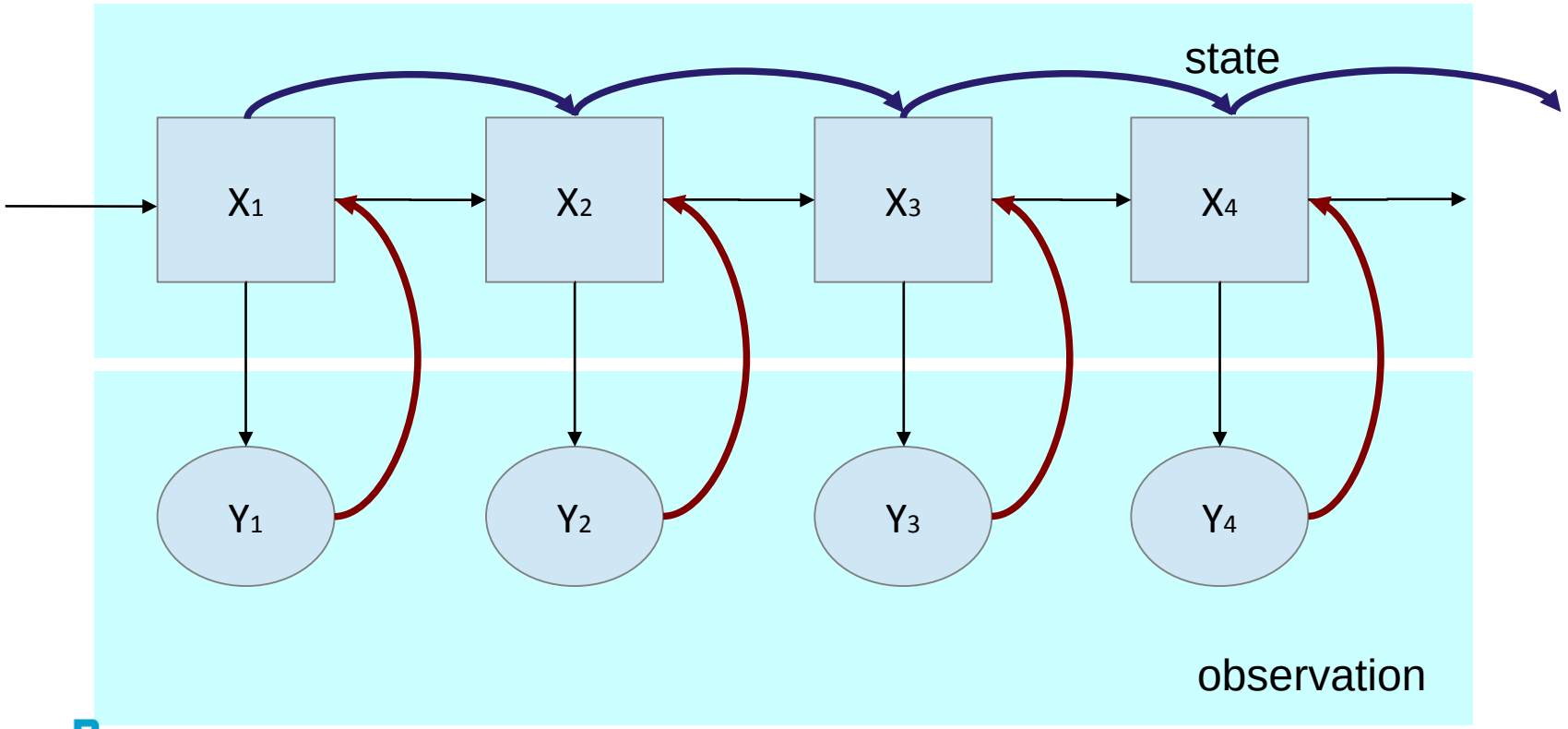


Cleaning the Data
Recursive Bayesian Estimation



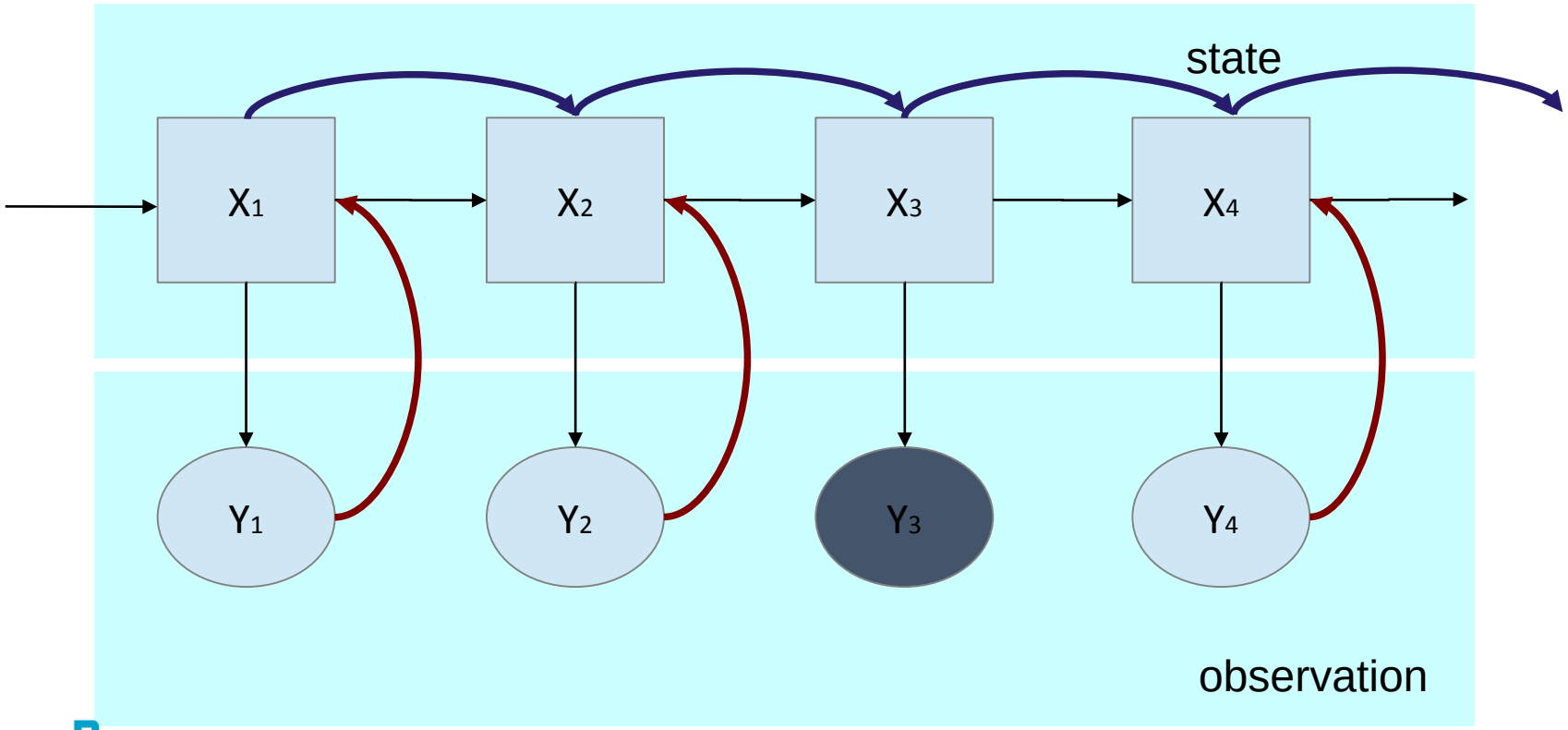
Update: $P(x_k|y_k) \propto P(x_k|y_{1..k-1})\, P(y_k|x_k)$

Cleaning the Data
Recursive Bayesian Estimation

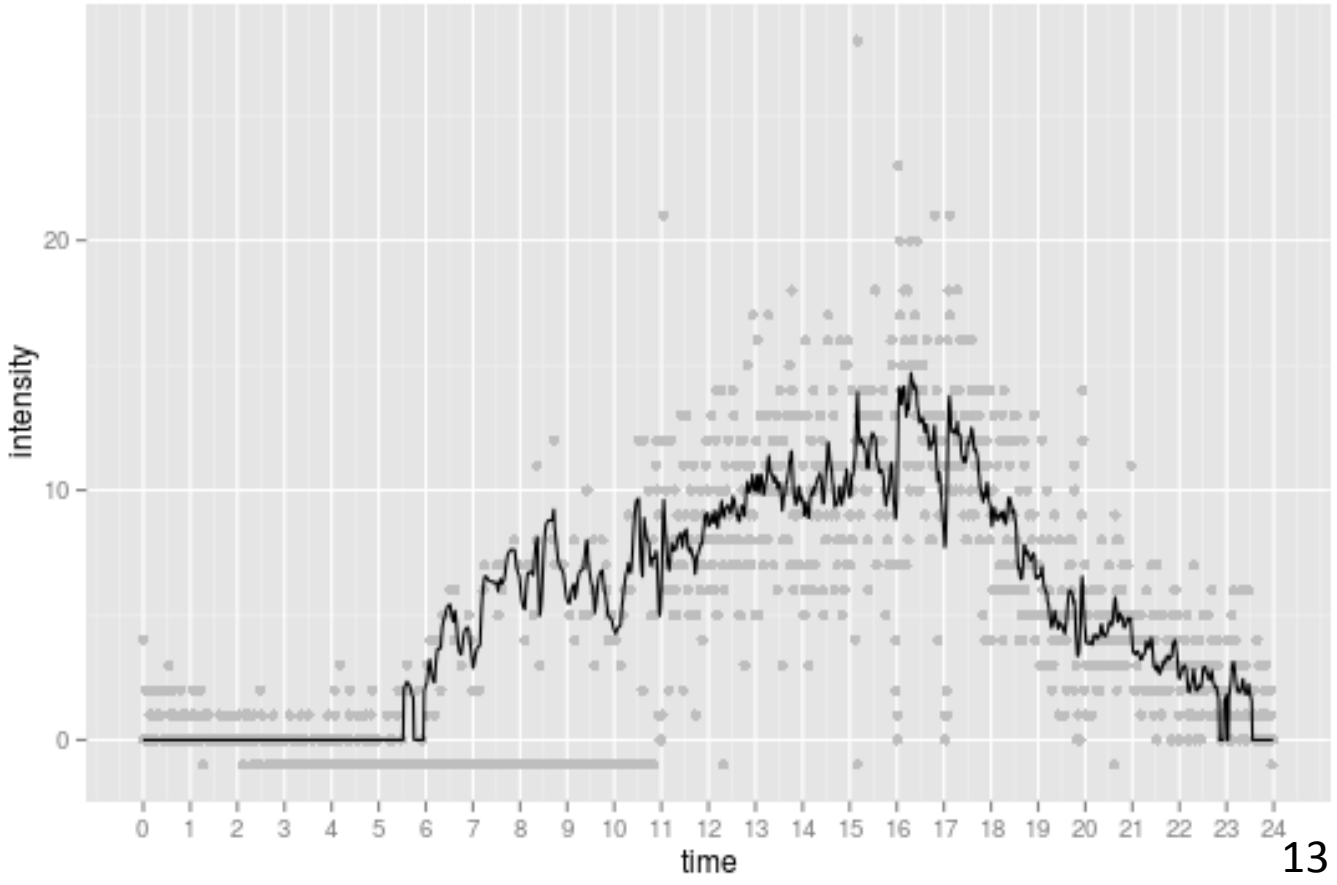


Prediction: $P(x_{k+1}|\, y_{1..k}) = \int_{-\infty}^{\infty} P(x_k|y_{1..k}) P(x_{k+1}|x_k)\, dx_k$

Cleaning the Data
Recursive Bayesian Estimation



Results of the filter



Monitoring Quality

- Number of minutes for which data is available varies per day per sensor
- Filter fills in blocks of missing values. For large blocks, the estimation of missing values is less accurate.
- Minimal deviation between original non-missing values and resulting signal.
- Smoothness of resulting signal

Resulting Indicators

Number of Measurements

$|M|$

Block indicator

For each block: $\frac{N(N+1)}{2}$

Difference between data and signal

$$D = \frac{\sum_{k \in M} x_k}{\sum_{k \in M} y_k} - 1$$

Smoothness of the signal

$$S = \frac{1}{K} \sum_{k=1}^{K} \frac{(y_k - y_{k-1})^2}{(y_k + y_{k-1})^2}$$

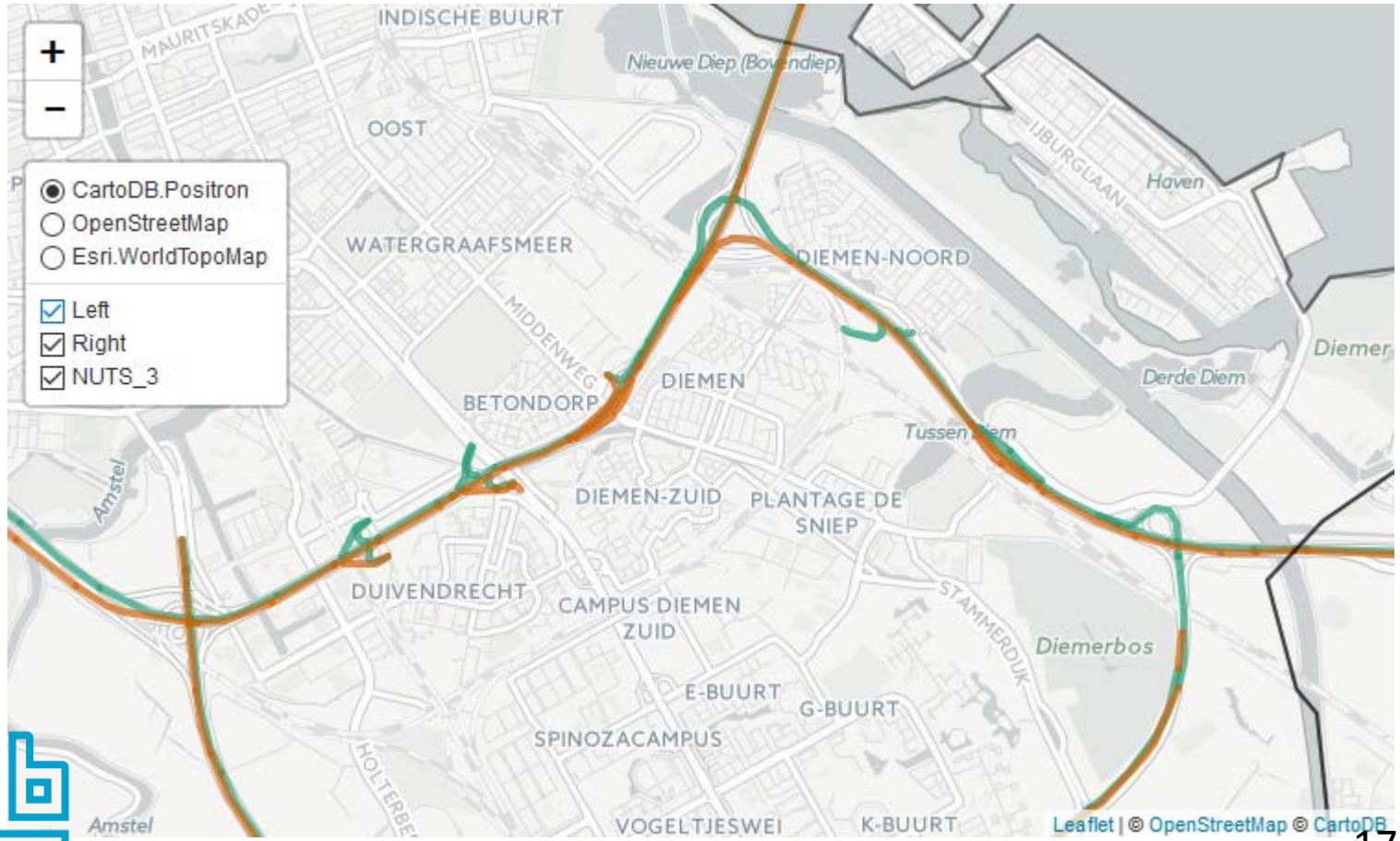
Precision/Accuracy

The filter does not introduce extra errors:

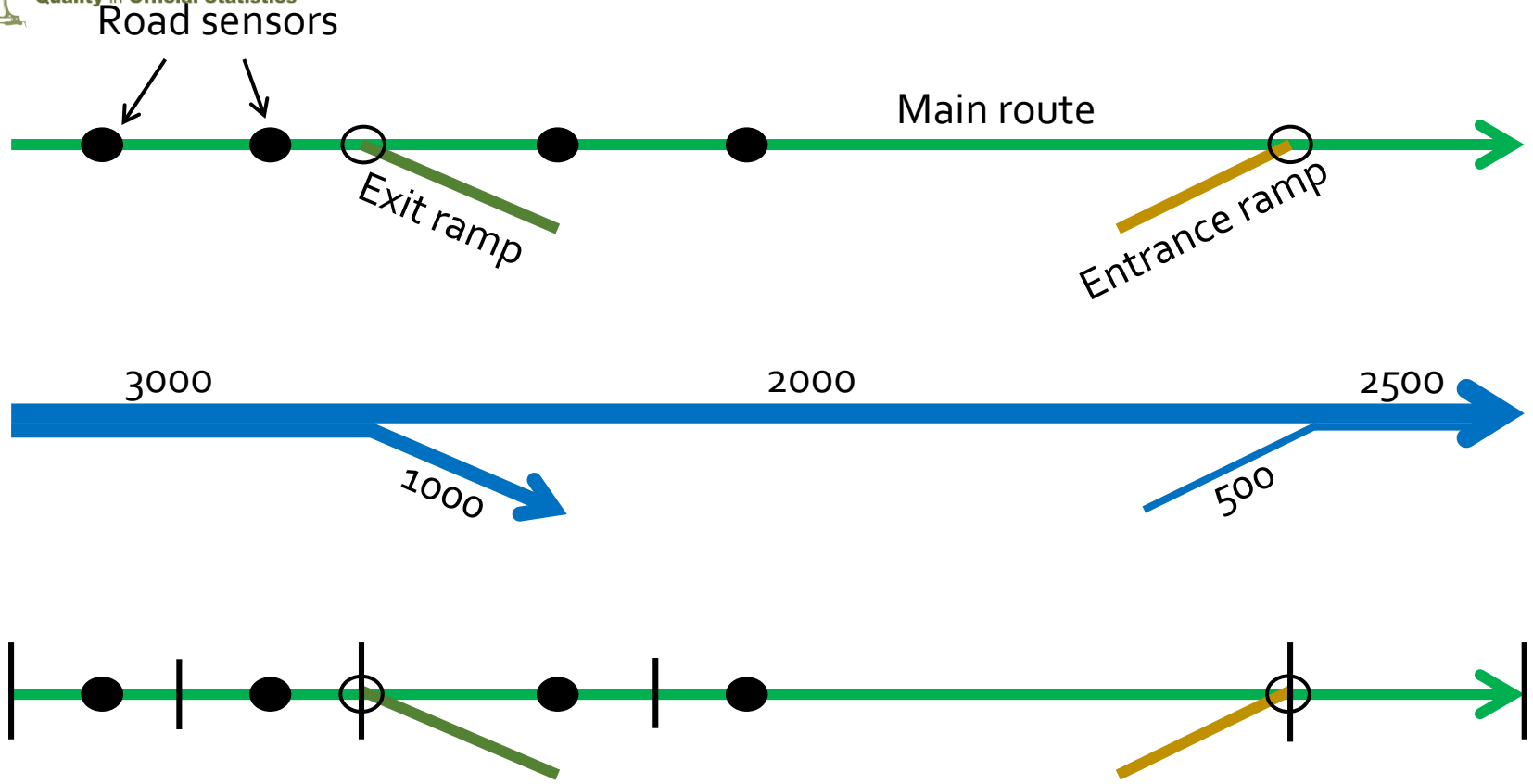
Precision: 3.6%

Accuracy:+0.13%

Card material

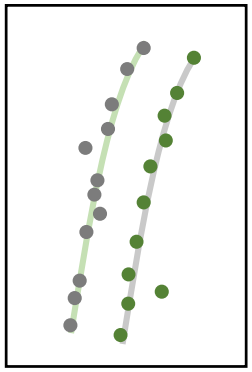


Calibration of the road sensors

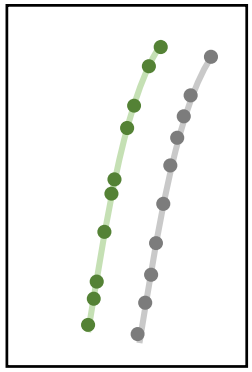


Quality of locations of sensors

- Check and (if necessary) correct traffic flow direction
- Projection of road sensors on roads
- Group sensors with unique location
- Remove outliers

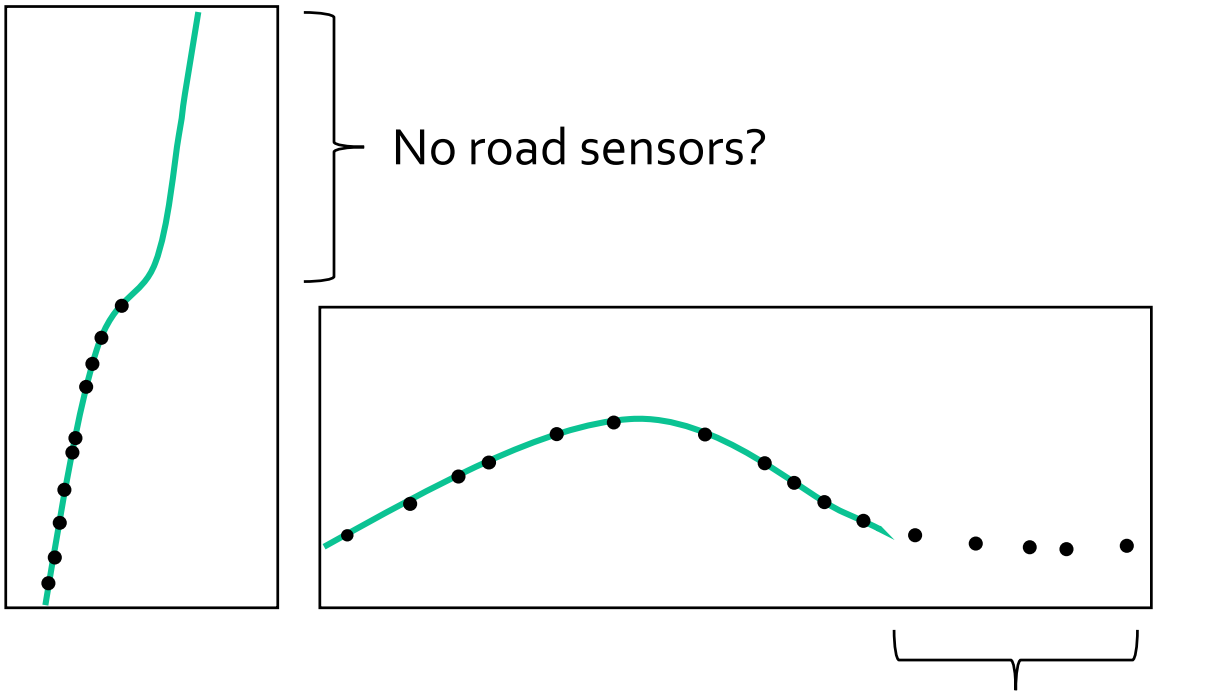


Raw sensor metadata



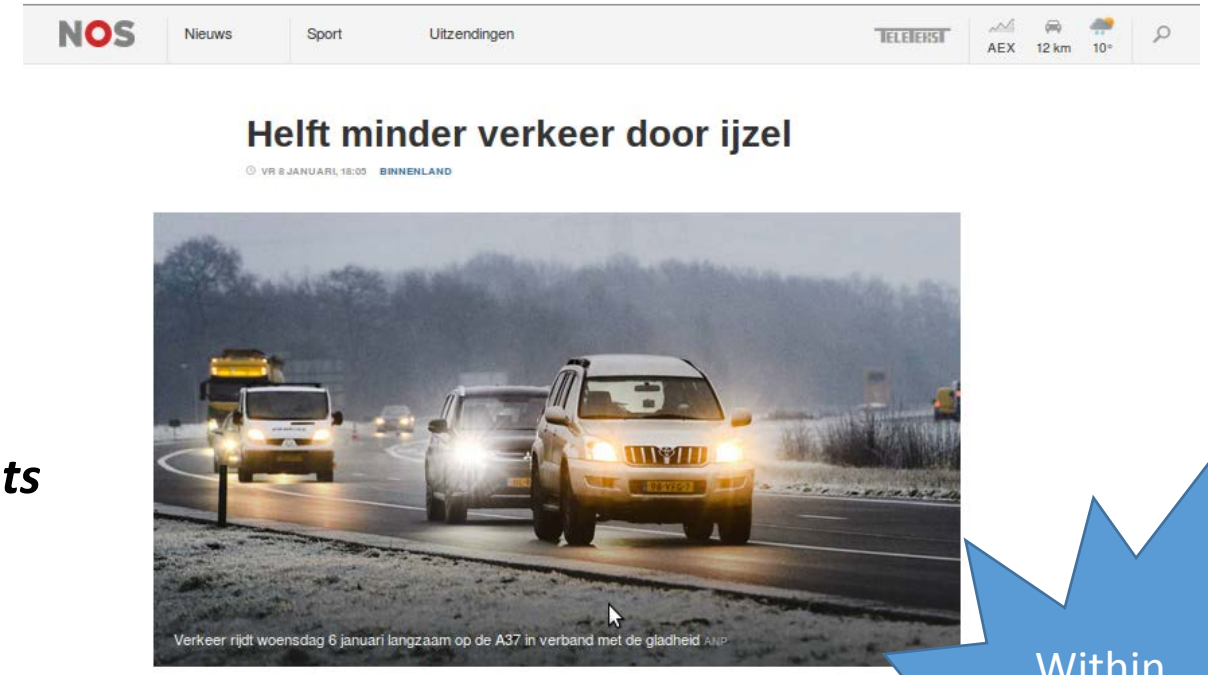
Edited sensor metadata

Metadata synchronization



Data journalism and (almost) real time statistics

Respond to *current events*



NOS Nieuws Sport Uitzendingen TELETEKST AEX 12 km 10°

Helft minder verkeer door ijzel

VR 8 JANUARI, 18:05 BINNENLAND

Verkeer rijdt woensdag 6 januari langzaam op de A37 in verband met de gladheid ANP

Veel mensen hebben de afgelopen dagen in Noord-Nederland het advies opgevolgd om vanwege de ijzel niet de weg op te gaan. De gladde wegen leidden tot een halvering van het verkeer op de rijkswegen.

Het CBS becijfert dat in de eerste drie werkdagen van 2016 gemiddeld 600 voertuigen per uur reden op de zes rijkswegen in Friesland, Drenthe en Groningen. In de afgelopen vier jaar waren dat er in diezelfde dagen gemiddeld 1200.

Op de N33, van Assen naar Eemshaven, was de invloed van de ijzel het grootst. Daar was 75 procent minder verkeer dan gemiddeld. Er reden slechts 115 voertuigen per uur.

De N33 is de rustigste rijksweg van Nederland. Het drukst is de A13 tussen Den Haag en Rotterdam, met in 2014 gemiddeld 5800 voertuigen per uur.

Within two days!