



Images and visualisation

Statistics Netherlands (CBS)

Edwin de Jonge

December 17, 2019

Images and visualisation

Sight is the largest and fastest perceptual input channel to our brain.

This makes consuming and generating visual images interesting as:

- **a data source**
- **medium for analysis and communication.**



Image as data source

3 cases at CBS / Statistics Netherlands

Train CNN / Deeplearning model for:

- CPI article classification of Web Shop
- Land Use Statistics
- Solar Energy production



Nieuwe items

Kleding
Schoenen & accessoires
Beauty

Trending now

The Holiday Gift Shop
Most wanted

Shop op item


Bekijk alle items
Jacks & Jassen
Knitwear
Vesten & Truien
Jurken
Tops
Overhemden & Blouses
Blazers
Basics
Broeken
Jeans
Jumpsuits & Playsuits

Rokken

Korte rokken
Midirokken
Maxiokken
Kokerrok
Spijkerrok

DAMES ROKKEN

Upgrade je rokken met rechte basics, schattige plooitjes of gestroomlijnde nietjes. Of je nu zin hebt in iets brutals als klederdracht, onze selectie damesrokken heeft wat je nodig hebt.

SORTEREN OP   FILTEREN & SORTEREN

359 items Model Product  



Midirok met pailletten
€ 39,99



Jacquardgeweven rok
€ 39,99



A-linerok
€ 19,99

Case 1: CPI



CPI image classification

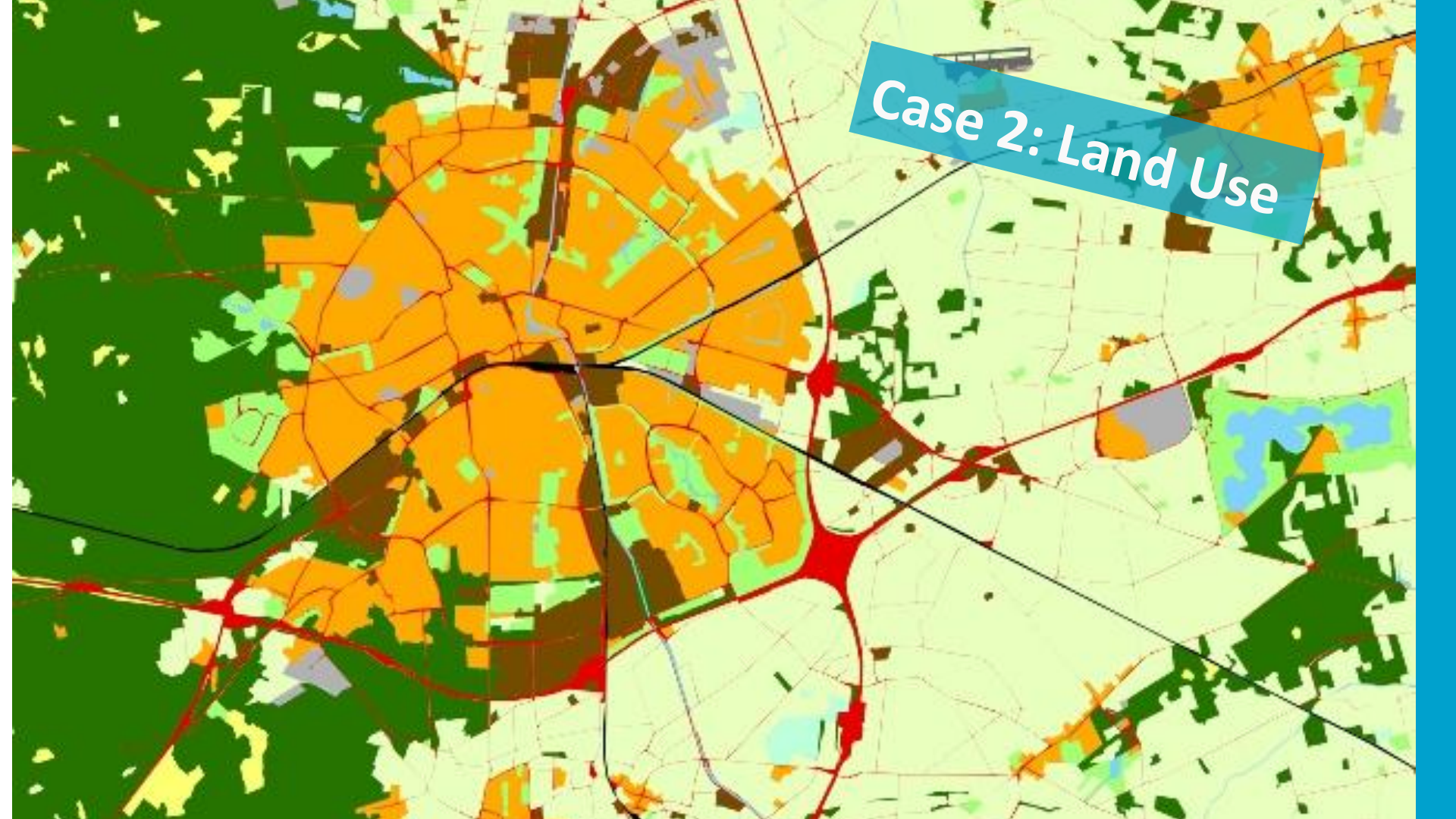
Statistics Netherlands uses web scraping for CPI:

- Clothing Web Shops contains > 100.000 articles
- Use Text to classify articles
- Experimented with classification image to improve classifier.

Outcome: Text classification is good (enough). Image by itself is worse, combination would give (slight) improvement



Case 2: Land Use



Case 2: Land Use Classification

- Land Use Statistics use aerial photo's to manually classify/derive land use (> 40 categories). Idea: use deeplearning to speed up the process. (currently 3 years...)
- Automatic Classifier (CNN) has accuracy > 90% for large categories, but not good enough to do everything automatically
- Current research: detect land use changes, so manual task takes much less time.



CASE 3: Solar Panels



Energy from Solar Panels

Use aerial photo's to detect solar panels, as input for solar energy production estimation.

Current status:

- Basic classification working (CNN), improving labelling of dataset by creating annotation tool.



Data Science en Visualisation?



Excellent tool for both analysis and communication:

Numerical quantities focus on expected values, graphical summaries on unexpected values.

John Tukey



Anscombe's quartet

Dataset 1

<i>x</i>	<i>y</i>
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

Dataset 2

<i>x</i>	<i>y</i>
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.13
7	7.26
5	4.74

Dataset 3

<i>x</i>	<i>y</i>
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

Dataset 4

<i>x</i>	<i>y</i>
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89



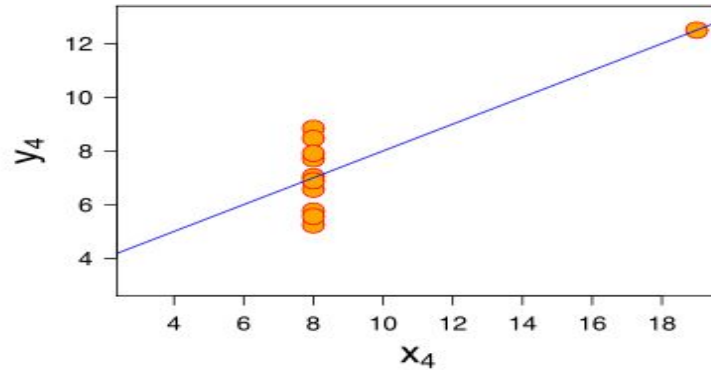
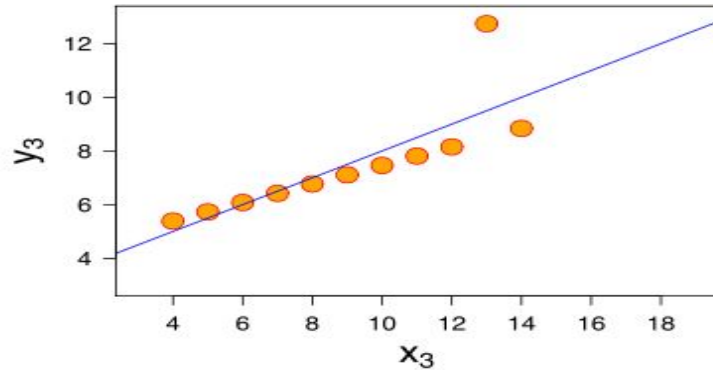
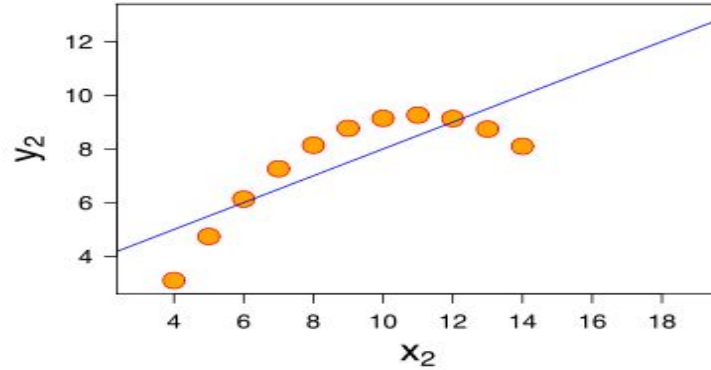
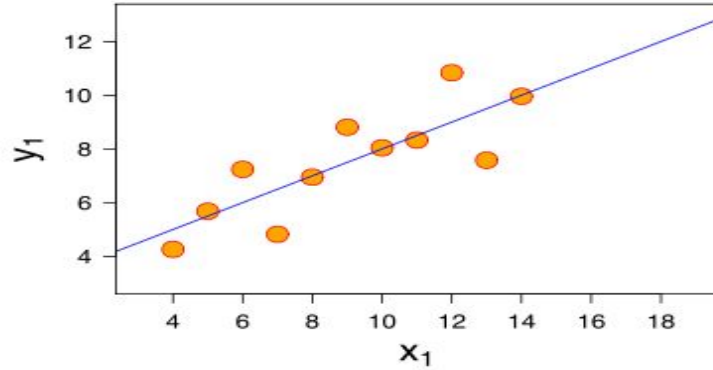
Anscombe's quartet

Statistical measure	value
Mean of x_1, x_2, x_3, x_4	Same: 9
Variance of x_1, x_2, x_3, x_4	Same: 11
Mean of y_1, y_2, y_3, y_4	Same: 7.50
Variance of y_1, y_2, y_3, y_4	Same: 4.1
Correlatie of ds_1, ds_2, ds_3, ds_4	Same 0.816
Linear regression ds_1, ds_2, ds_3, ds_4	Same: $y = 3.00 + 0.500x$

Looks same?



Let's plot!



Uncertainty visualisation

What is not surrounded by uncertainty cannot be the truth,

Richard Feynman

For official statistics, at least two reasons useful:

- Communicating accuracy
- Statistical/stochastic uncertainty

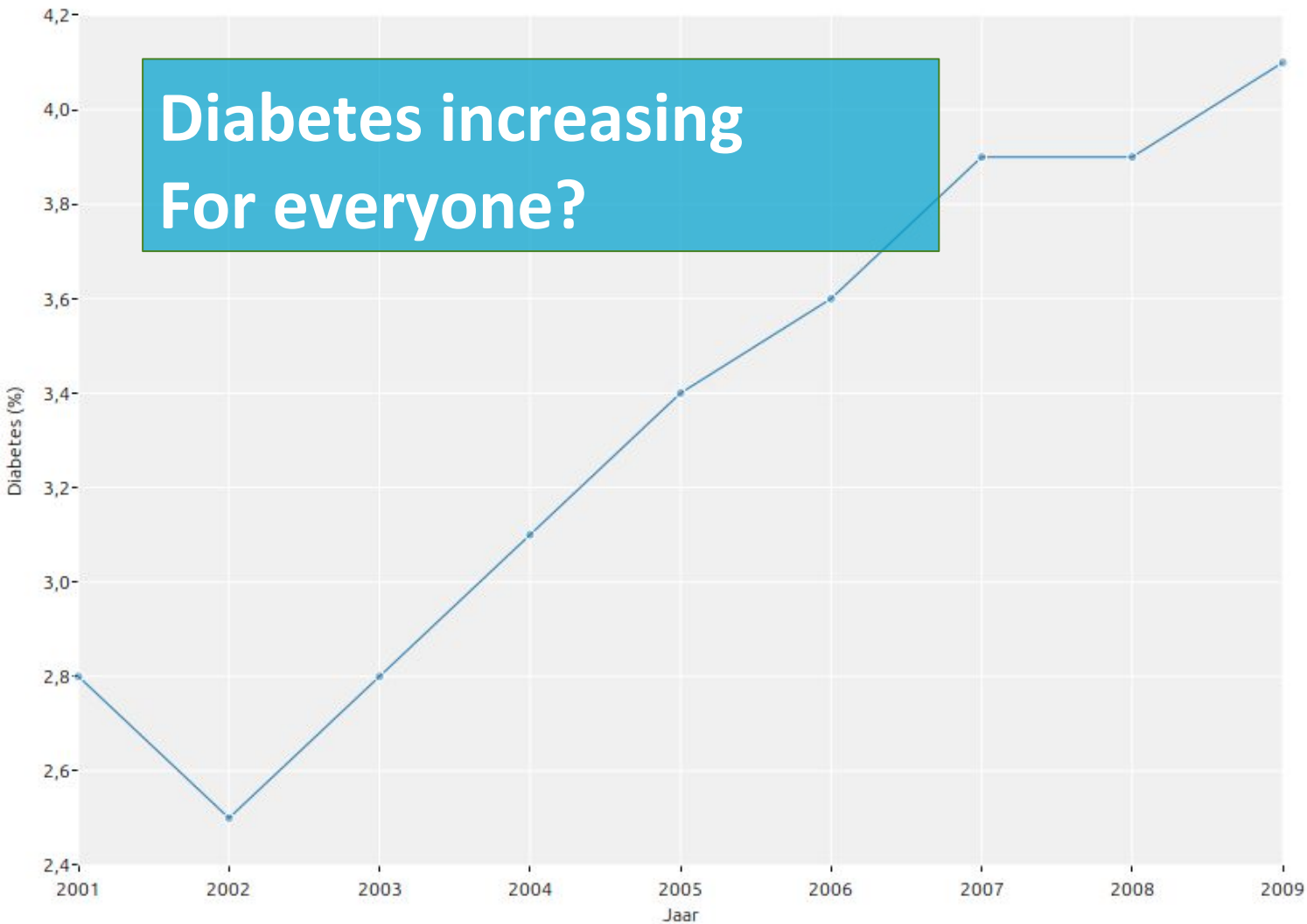
Let's view two cases of stats NL (CBS)

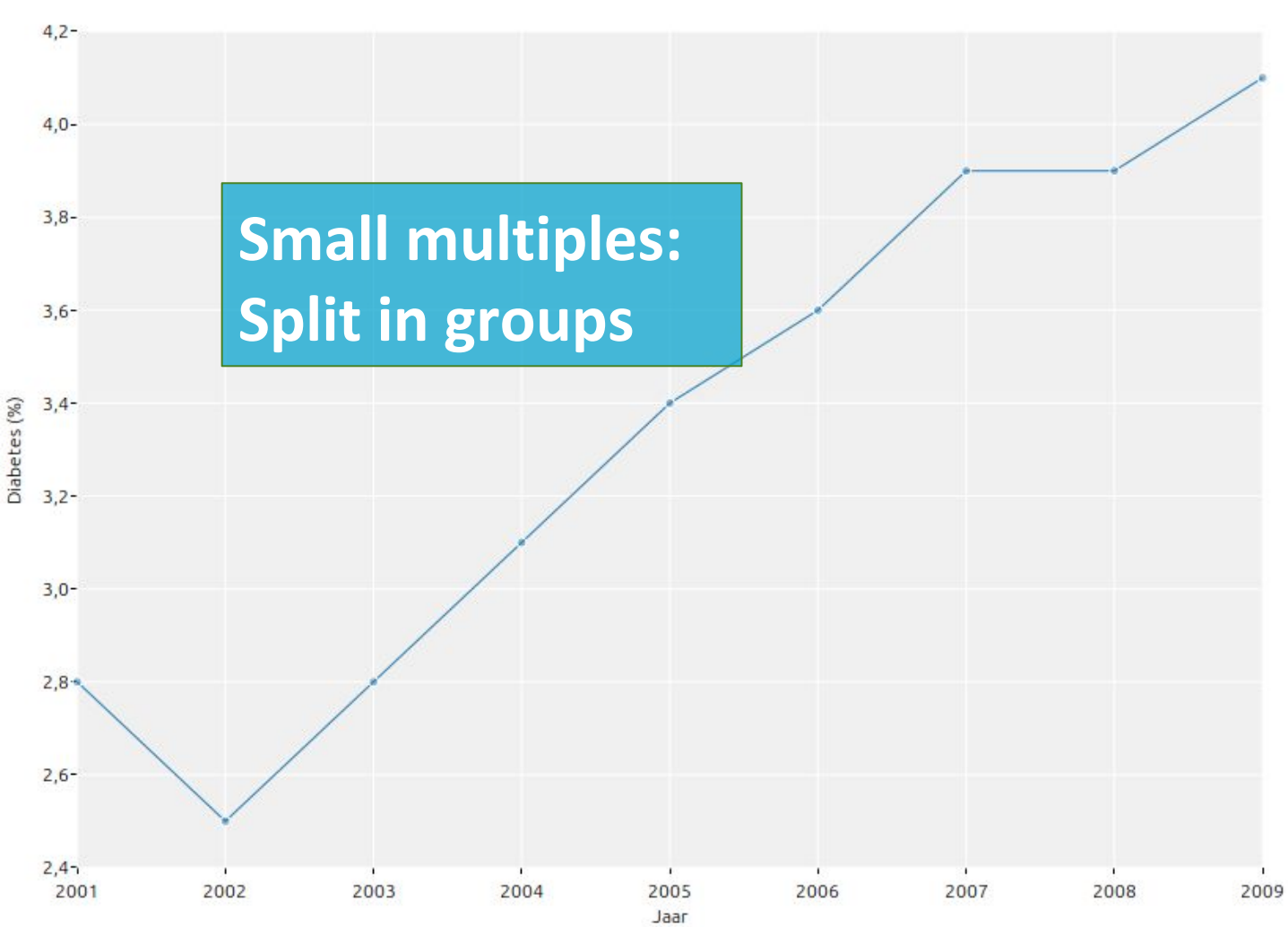


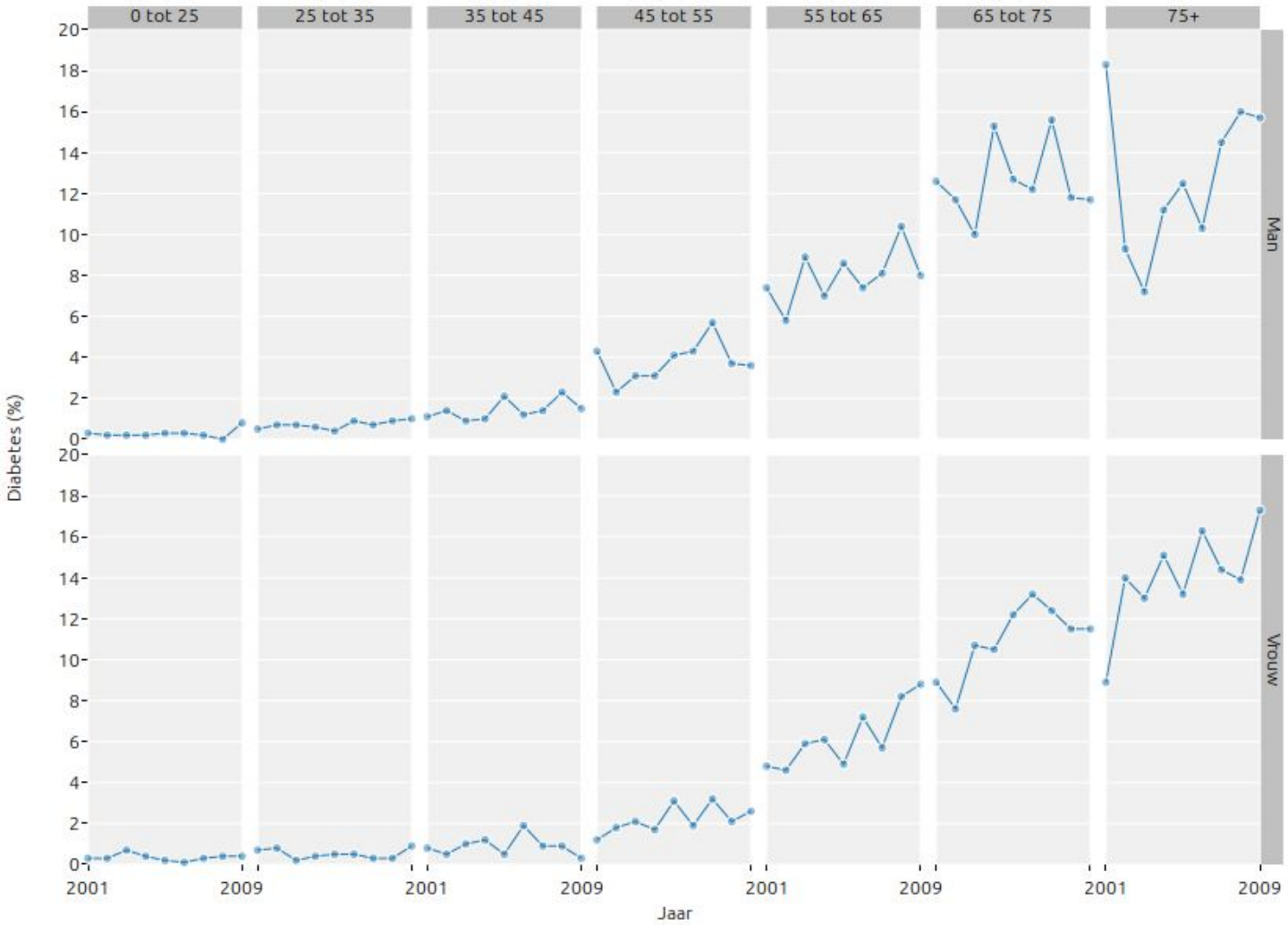
Diabetes incidence

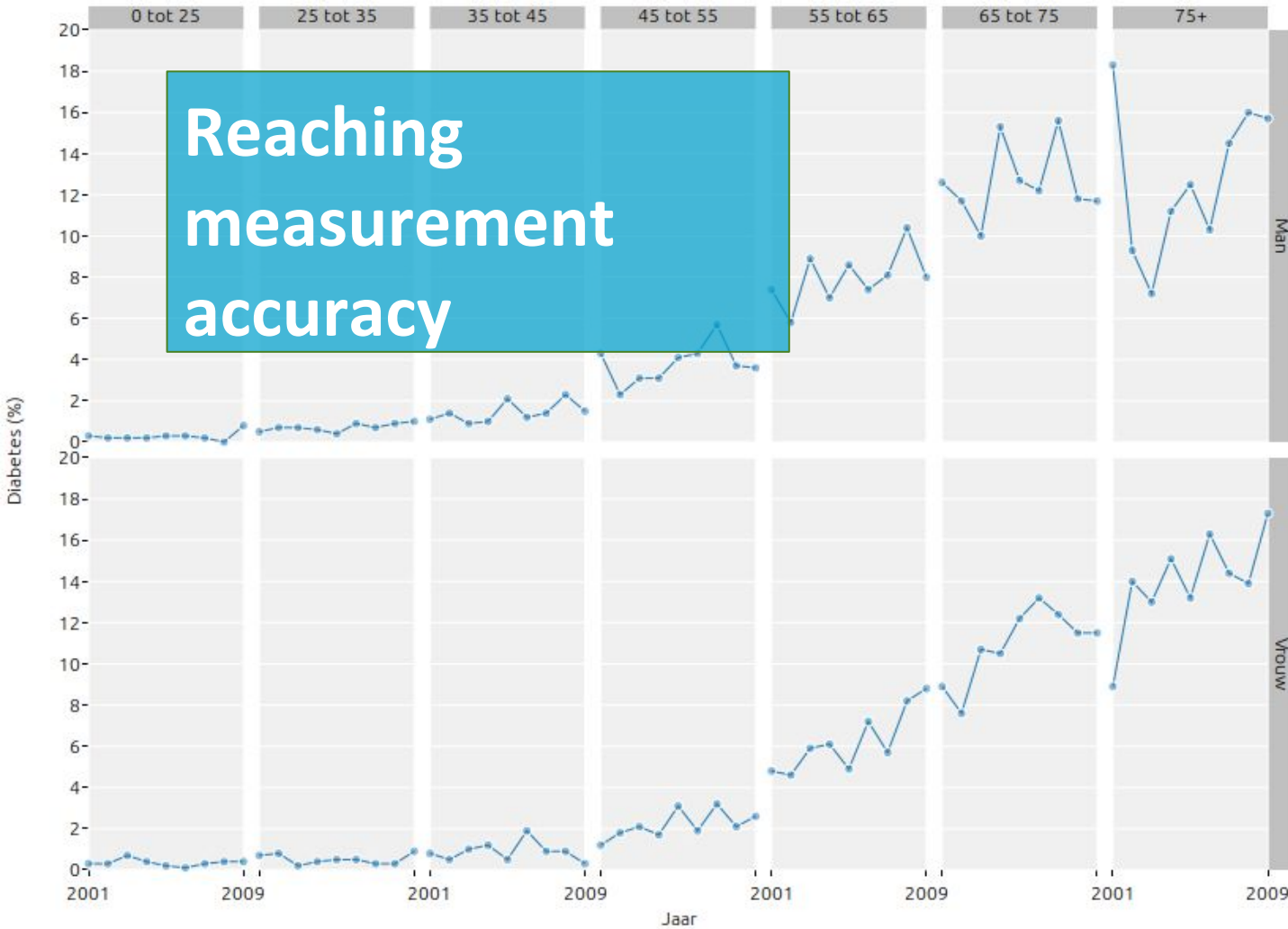
- Based on a (large) health survey of statistics netherlands (CBS)

**Diabetes increasing
For everyone?**



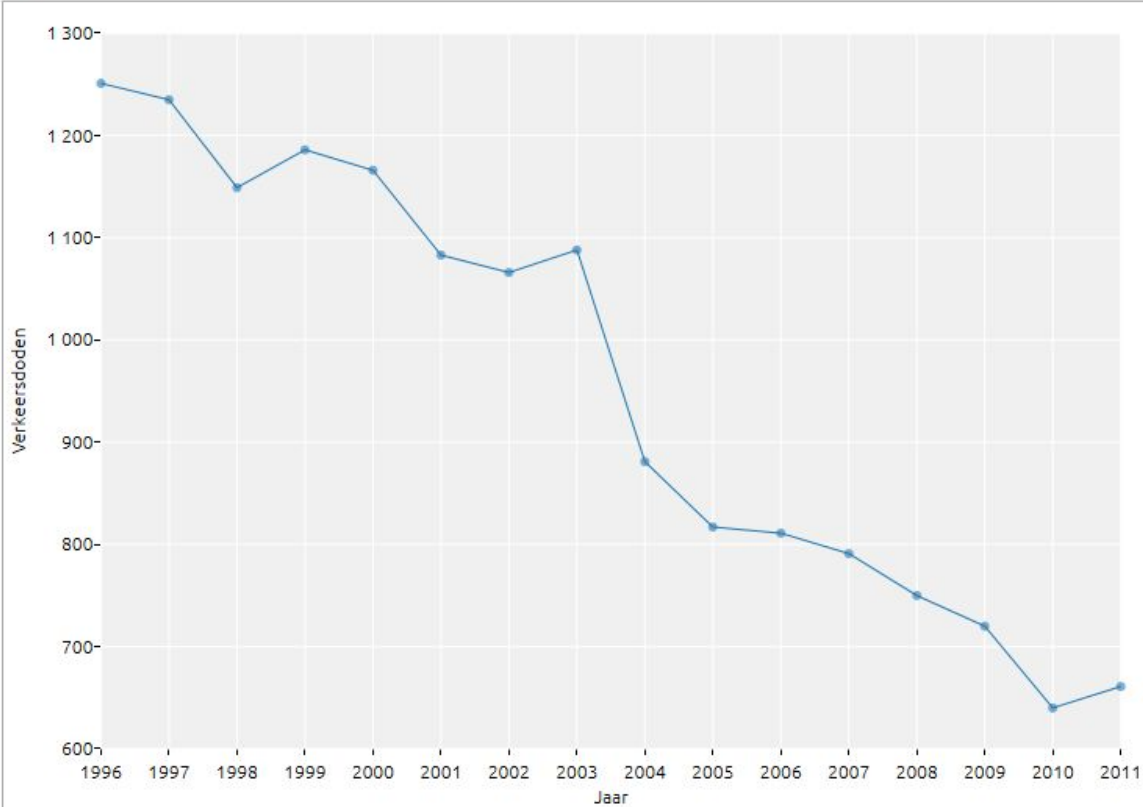




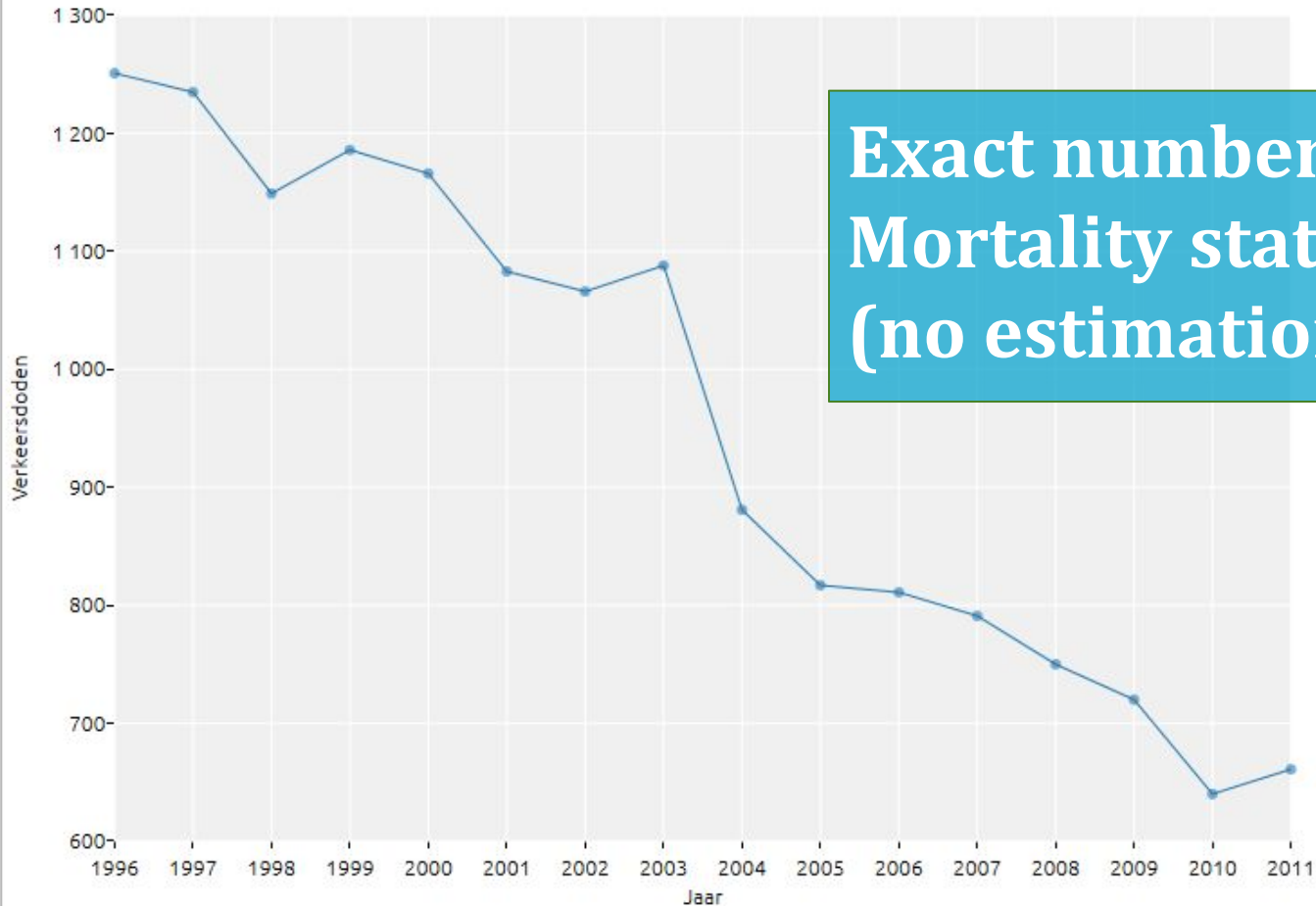


Traffic casualties

Verkeersdoden

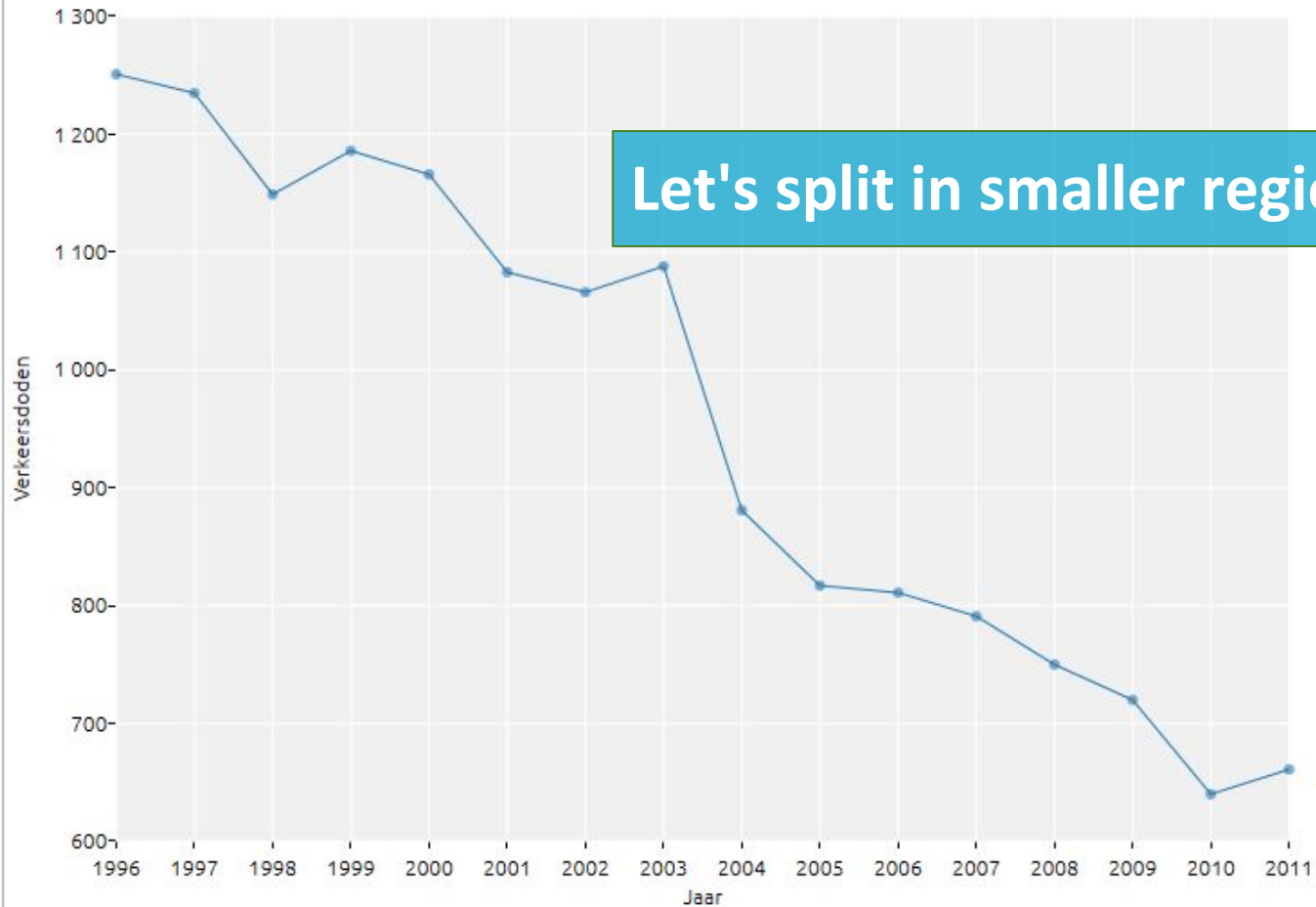


Traffic casualties

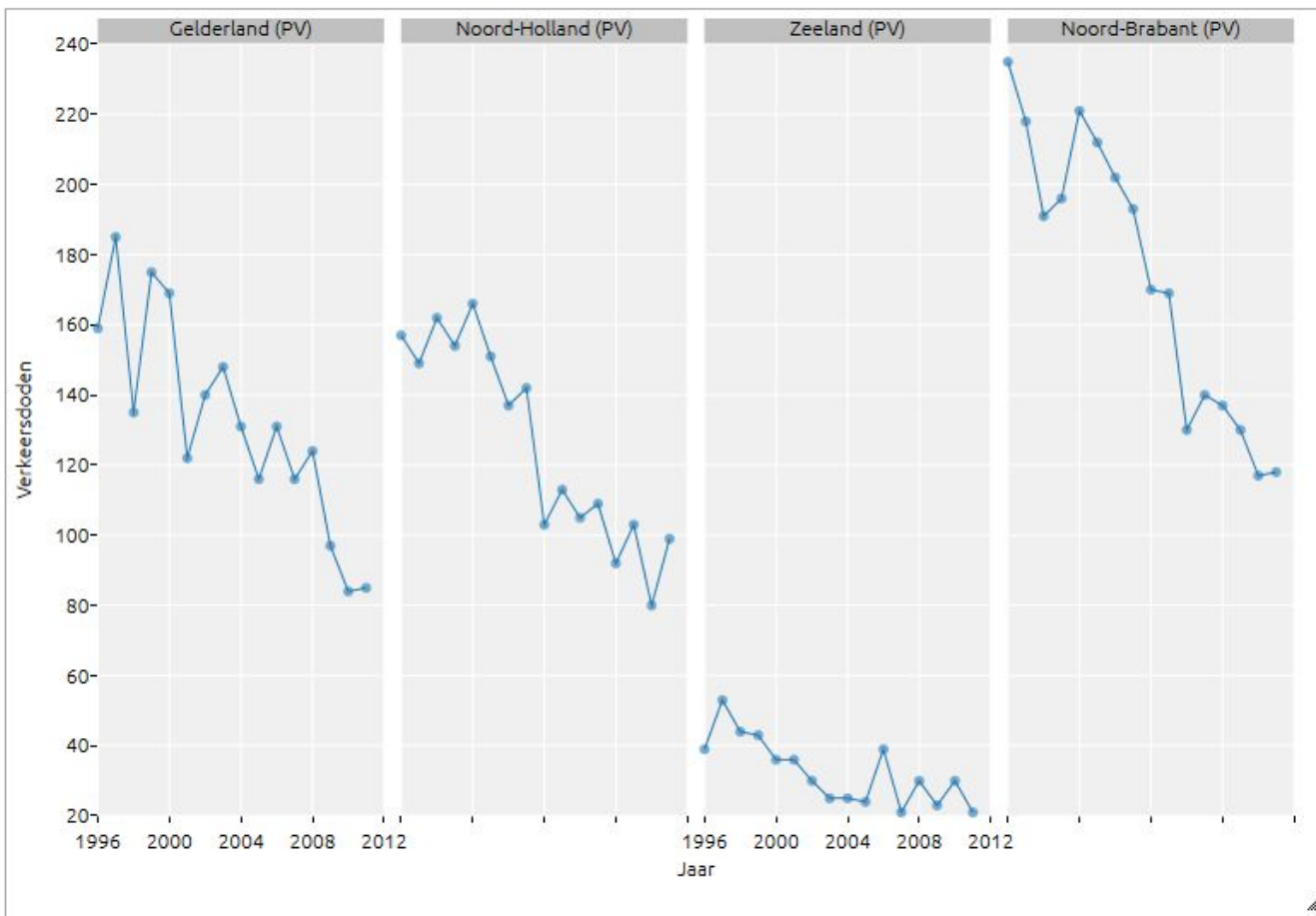


**Exact numbers!
Mortality stats
(no estimation)**

Traffic casualties

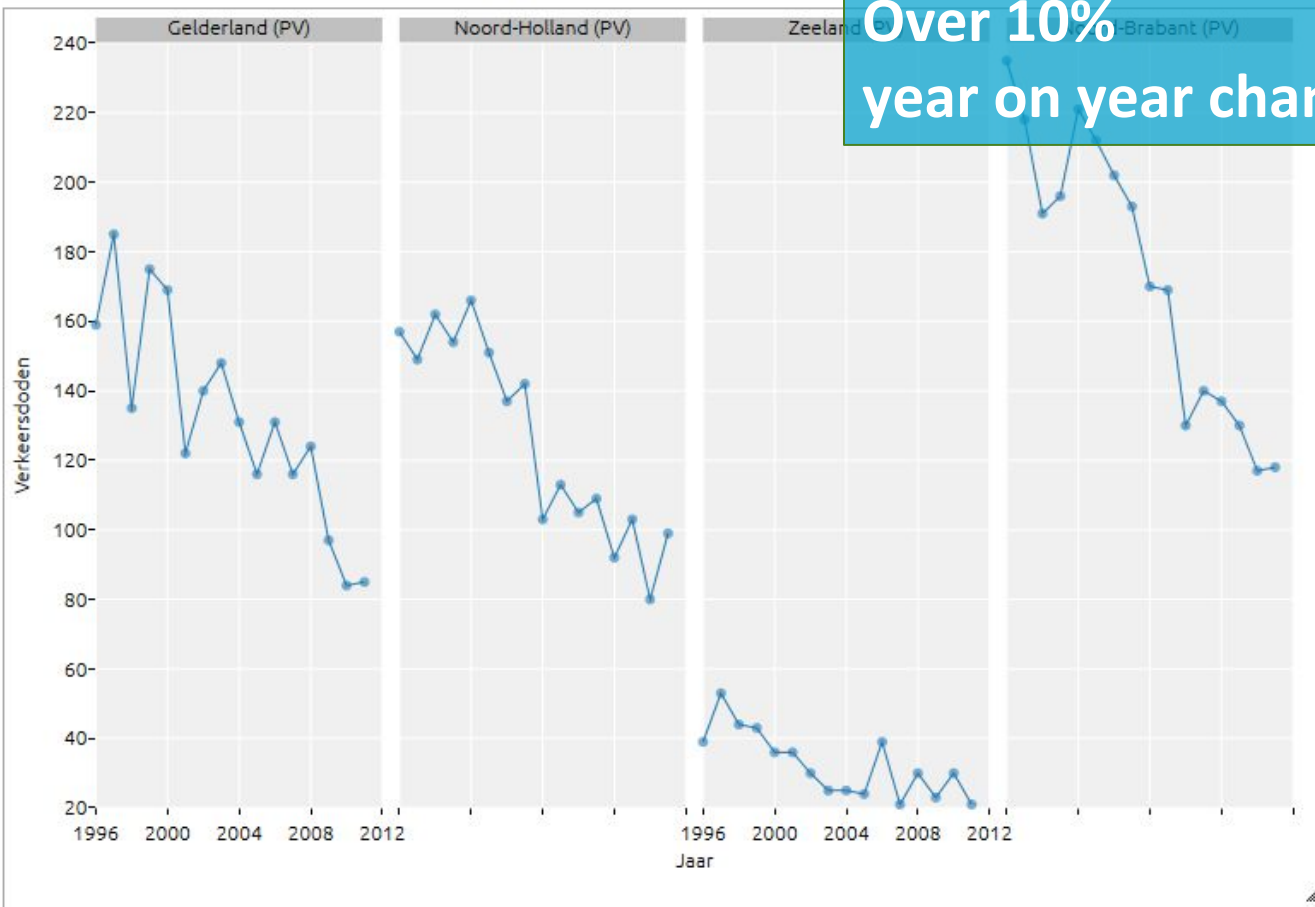


Traffic casualties



Traffic casualties

Over 10%
year on year changes!



Case 2: Stochastic uncertainty



User Studies show:

Non-expert users can read probability intervals!

The perception of visual uncertainty representation by non-experts

Tak, Toet, van Erp, *Transactions Visualisation and computer Graphics*, 2014

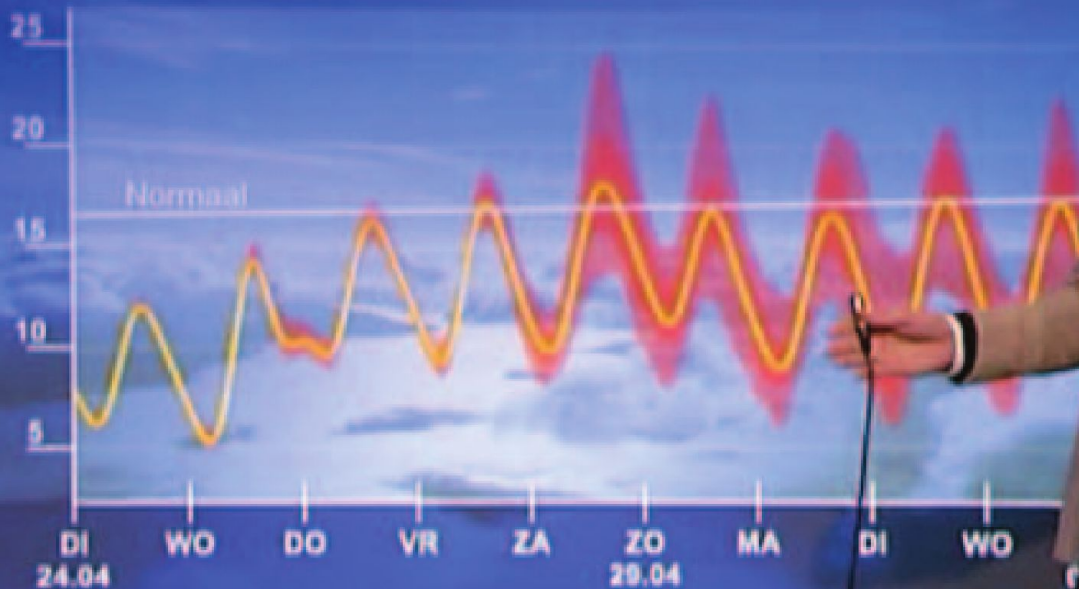
displaying uncertainty improves data assessment

Effect of displaying uncertainty in Line and Bar charts, Van der

Laan, [de Jonge](#), Solcer, IVAPP, 2015



Weather forecast (Dutch television)

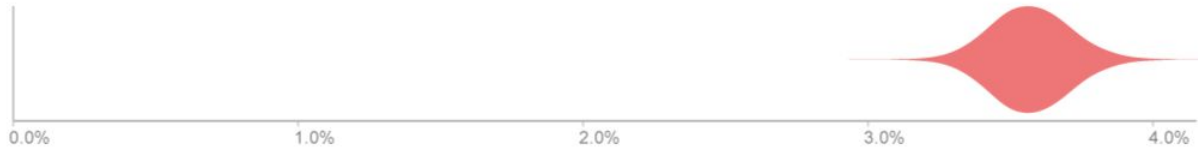


Uncertainty Viz (density)

Uncertainty in the US unemployment rate with the May 2019: Continuous encoding



Density plot



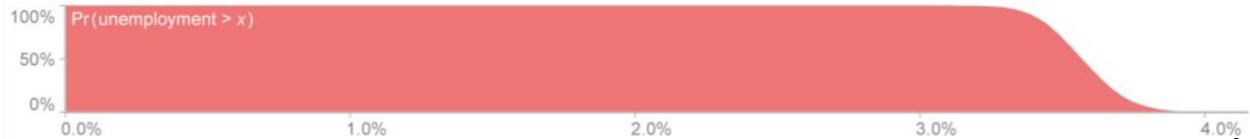
Violin plot



Gradient plot



Cumulative distribution function (CDF)

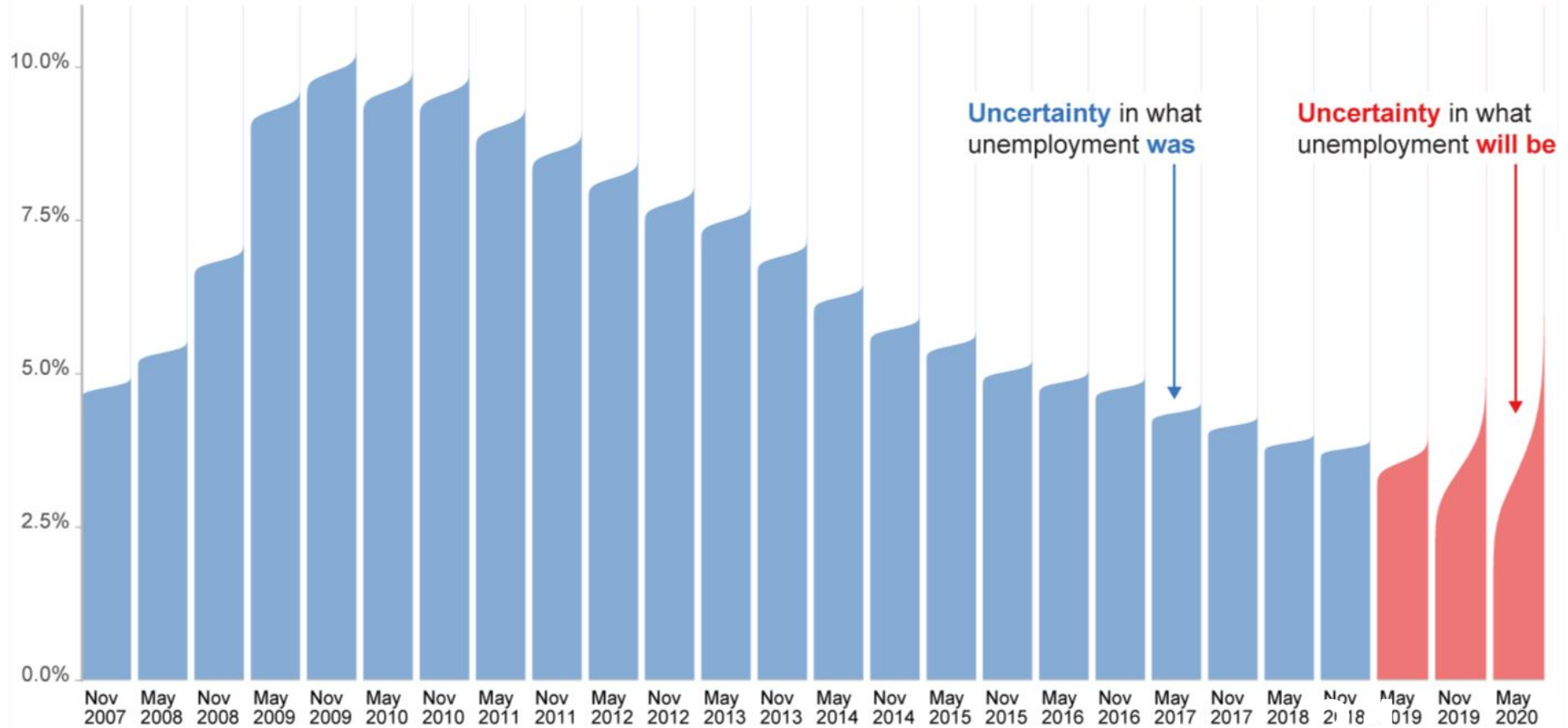


Complementary CDF (CCDF)



Uncertainty Viz (density)

US unemployment over time



communicate uncertainty in how well our statistical model describes the

Matthew Kay and Jessica Hullman (2019)

COMUNIKOS

COMUNIKOS: Eurostat project

- Goal: guidelines in **COM**municating **U**ncertain **K**nowledge in **O**fficial **S**tatistics

Tasks:

- Describe possible sources of uncertainty
- Visualisation Guidelines
- Methods for calculating uncertainty measures
- POC on Scanner Data



