Unit 4:
Using Big Data

Piet Daas & Marco Puts

Centraal Bureau
voor de Statistiek

Overview

- Google flu
  - To illustrate the pros and cons of Big Data
- Extracting information from Big Data
  - Dealing with new sources of data: text and images
- IT aspects of Big Data
  - Dealing with large amounts of data
  - How to efficiently process large amounts of data?
- Big Data methodology
  - What are the steps needed to correctly use Big Data?
  - Quality: What is signal, what is noise?
  - Advantages of using visualizations
- *Assignment*

Introduction

- Using Big Data will -in general- involve developing a model
  - Explore the data
  - Check data quality
  - Check the need to pre-process (‘filter’) the data
  - Check any assumptions made
  - Build a model, test, etc.

- If a model is not required
  - ‘Do you simply add up the numbers?’

Google Flu detection

Google flu trends

- A very famous use of Big Data
- Published in 2009 (in Nature)
- On-line Google flu trends 2009-2013
- A counter reaction in 2014 (in Science)
- Google flu trends no longer available since 2015
  - https://www.google.org/flutrends/about/
- The future: new approaches by others

Google flu prediction

- Google flu prediction
- In 2009 a paper was published in Nature stating that the search behaviour of people on Google could be used to predict the occurrence of flu in the USA
  - Detecting influenza epidemics using search engine query data
  - Jeremy Ginsberg[1] , Matthew H. Mohebbi[1] , Rajan S. Patel[1] , Lynnette Brammer[2] , Mark S. Smolinski[1] & Larry Brilliant[1]
    [1]Google Inc., [2]Centers for Disease Control and Prevention
    Nature 457, pp. 1012-1014.

https://static.googleusercontent.com/media/research.google.com/nl//archive/papers/detecting-influenza-epidemics.pdf

Google flu prediction (2)

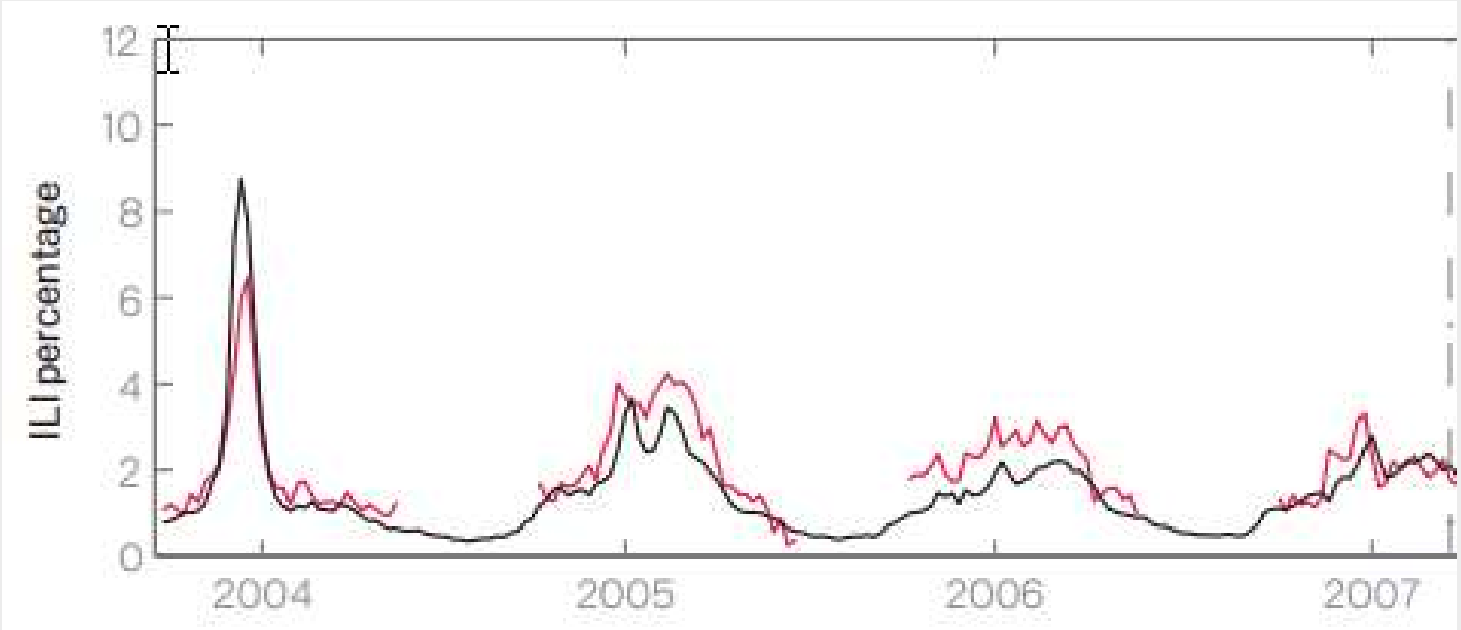
- Google flu prediction

- A high correlation was found between certain combinations of (Google) search terms and the number of patients with influenza-like symptoms visiting physicians (registered by the Center for Disease Control; CDC).
- A method was presented in which large numbers of Google search queries were used to track influenza-like illness in a population at the US regional level with a ***1 day*** lag.
- The official flu figures are produced by the CDC with a **1-2 week** lag.

https://static.googleusercontent.com/media/research.google.com/nl//archive/papers/detecting-influenza-epidemics.pdf

Google flu prediction (3)

- Google flu prediction findings
- CDC data in red, Search term prediction in black
- 45 Search ‘terms’ are used, correlation ~0.9



ILI: influenza-like illness

Google flu prediction (4)

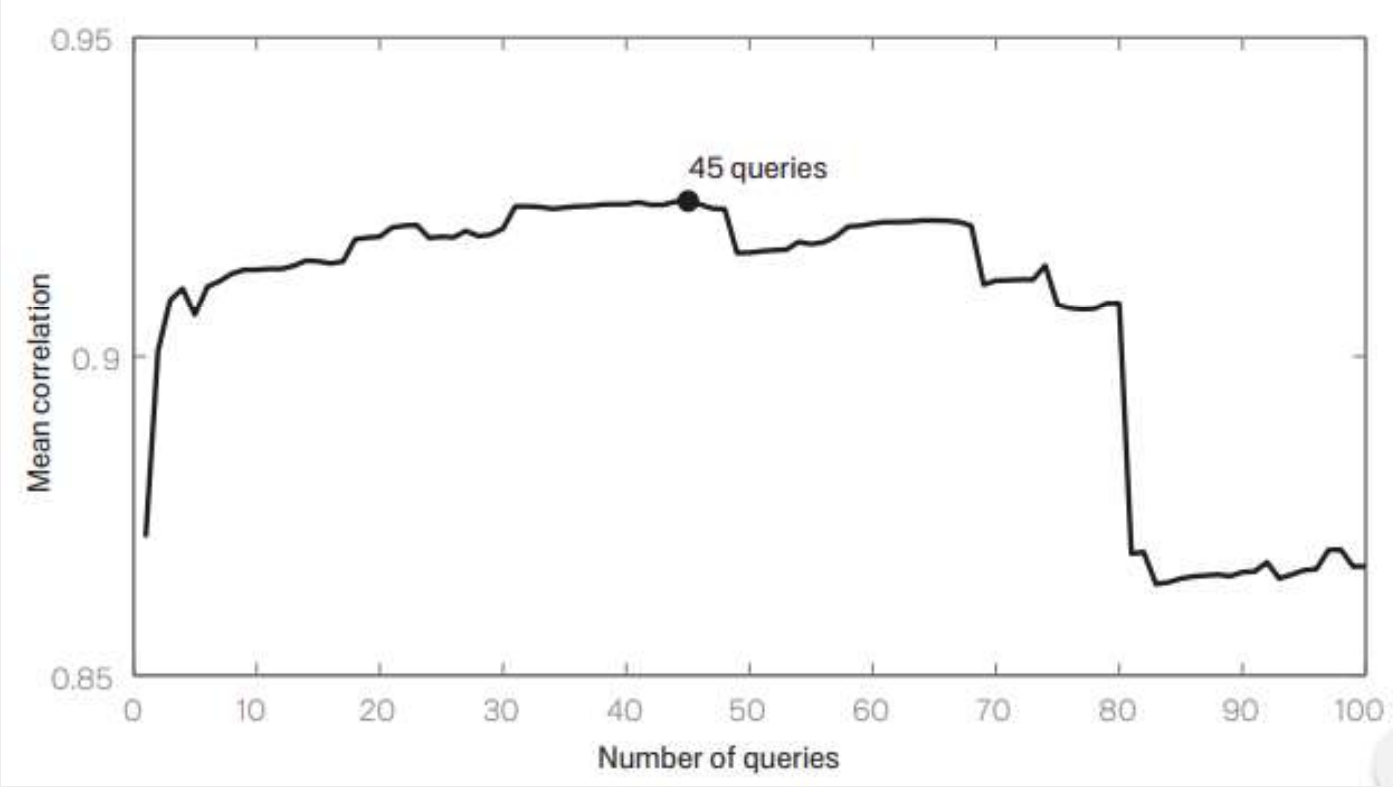
- 45 Search ‘terms’ are used
- An ‘automatic’ selection procedure was used to find the highest correlating terms
  - Terms: single words or combination of words
  - Linear model: $\text{logit}(P) = \beta_0 + \beta_1 \times \text{logit}(Q) + \varepsilon$
  - P = physician visits, Q = search query; $\text{logit}(P) = \ln(P/(1-P))$
- What terms are included?
- How are they selected? Fully automatic????

Google flu prediction (5)

- Selection of terms
  - 50 million search queries in total for 1152 CDC data points
- Each term was tested
  - Individually, by selected the highest scoring (correlation) across US regions
  - In combination, by selecting the best combinations of terms
  - AND, some terms were excluded (domain knowledge)
    - Such as: "*high school basketball*" and other related terms which coincide (no others were mentioned)
  - A total of 450 million models (term combinations) were evaluated

Google flu prediction (6)

– Looking for the optimal combination of terms



A drop-off after term 81 = '*oscar nominations*'

Google flu prediction (7)

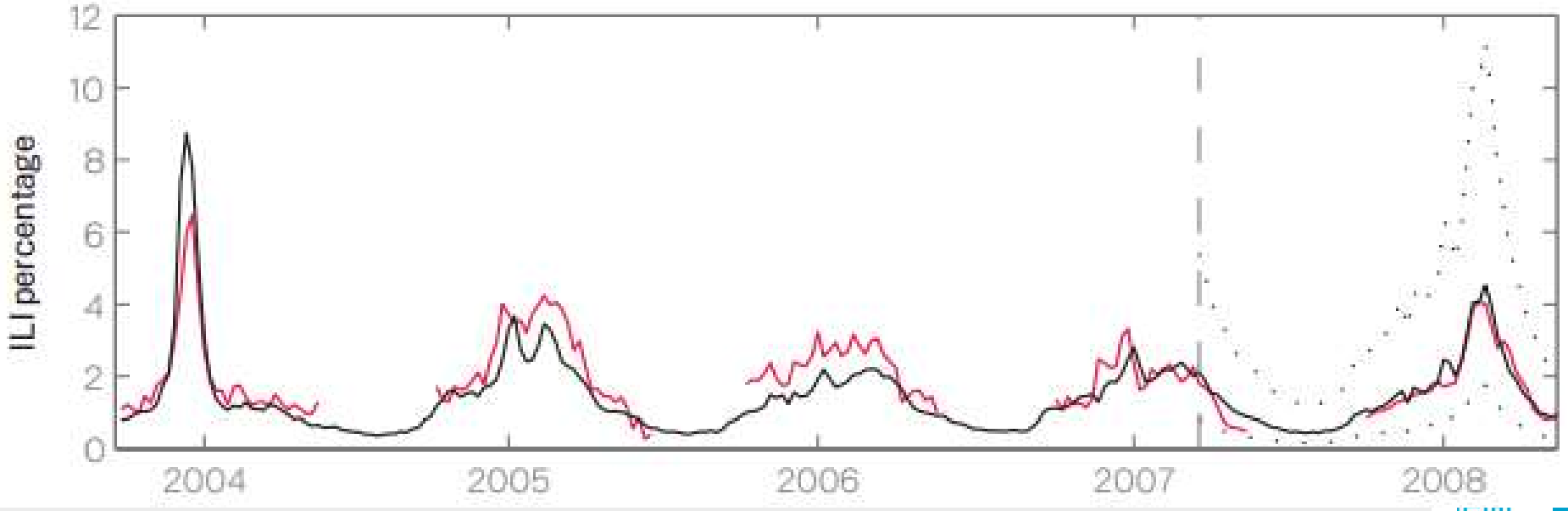
– Tops scoring ‘terms’

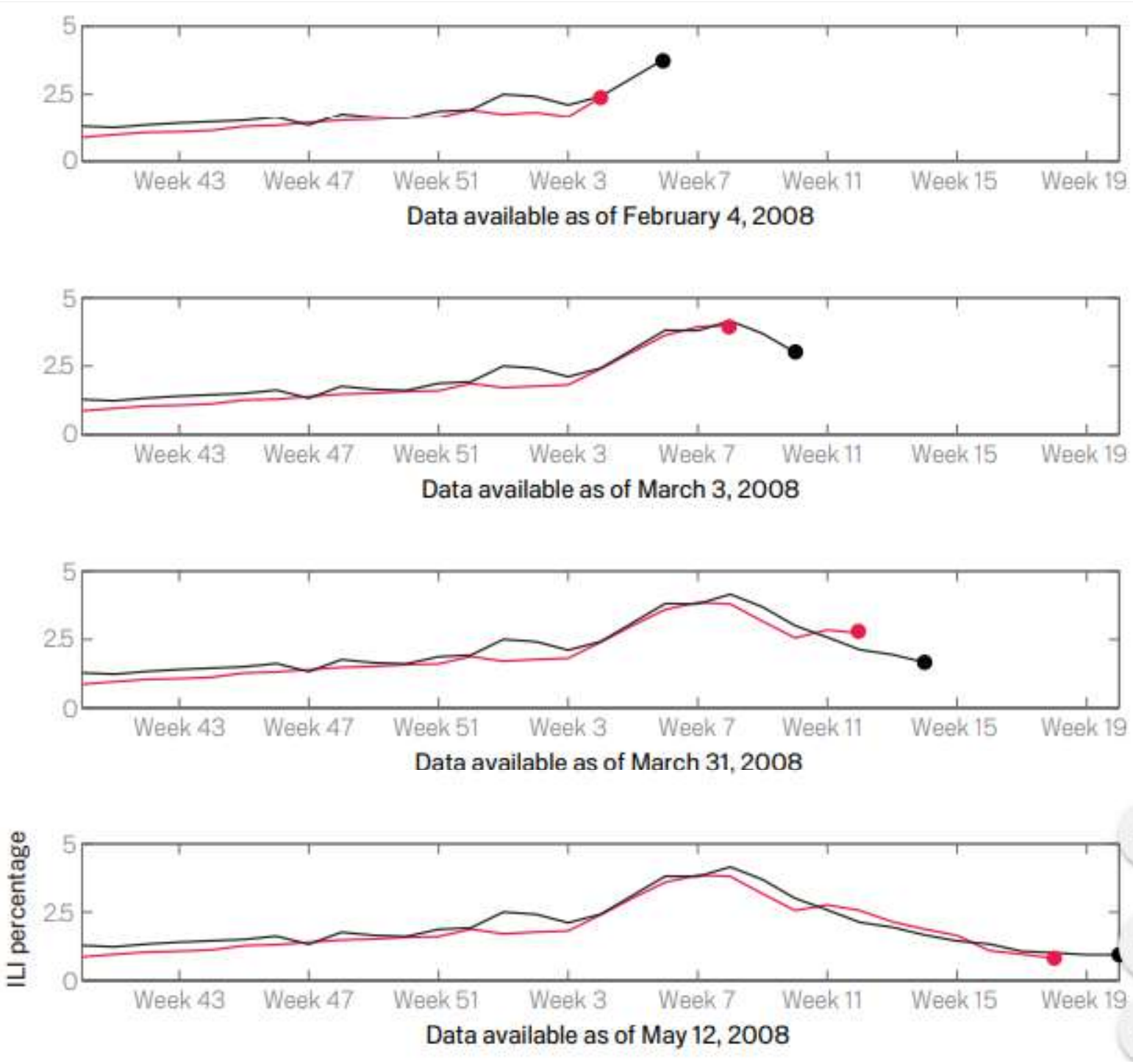
	Top 45 Queries	
Search Query Topic	N	Weighted
Influenza Complication	11	18.15
Cold/Flu Remedy	8	5.05
General Influenza Symptoms	5	2.60
Term for Influenza	4	3.74
Specific Influenza Symptom	4	2.54
Symptoms of an Influenza Complication	4	2.21
Antibiotic Medication	3	6.23
General Influenza Remedies	2	0.18
Symptoms of a Related Disease	2	1.66
Antiviral Medication	1	0.39
Related Disease	1	6.66
Unrelated to Influenza	0	0.00
	45	**49.40**

This not a very clear and transparent table!! (no more detailed info available)

Google flu prediction (8)

- Google flu prediction findings
- CDC data in red, Search term prediction in black
- Prediction in 2008 with 95% pred. intervals





Prediction in more detail (for a particular US region)

Model in black
CDC in red

Google flu prediction (10)

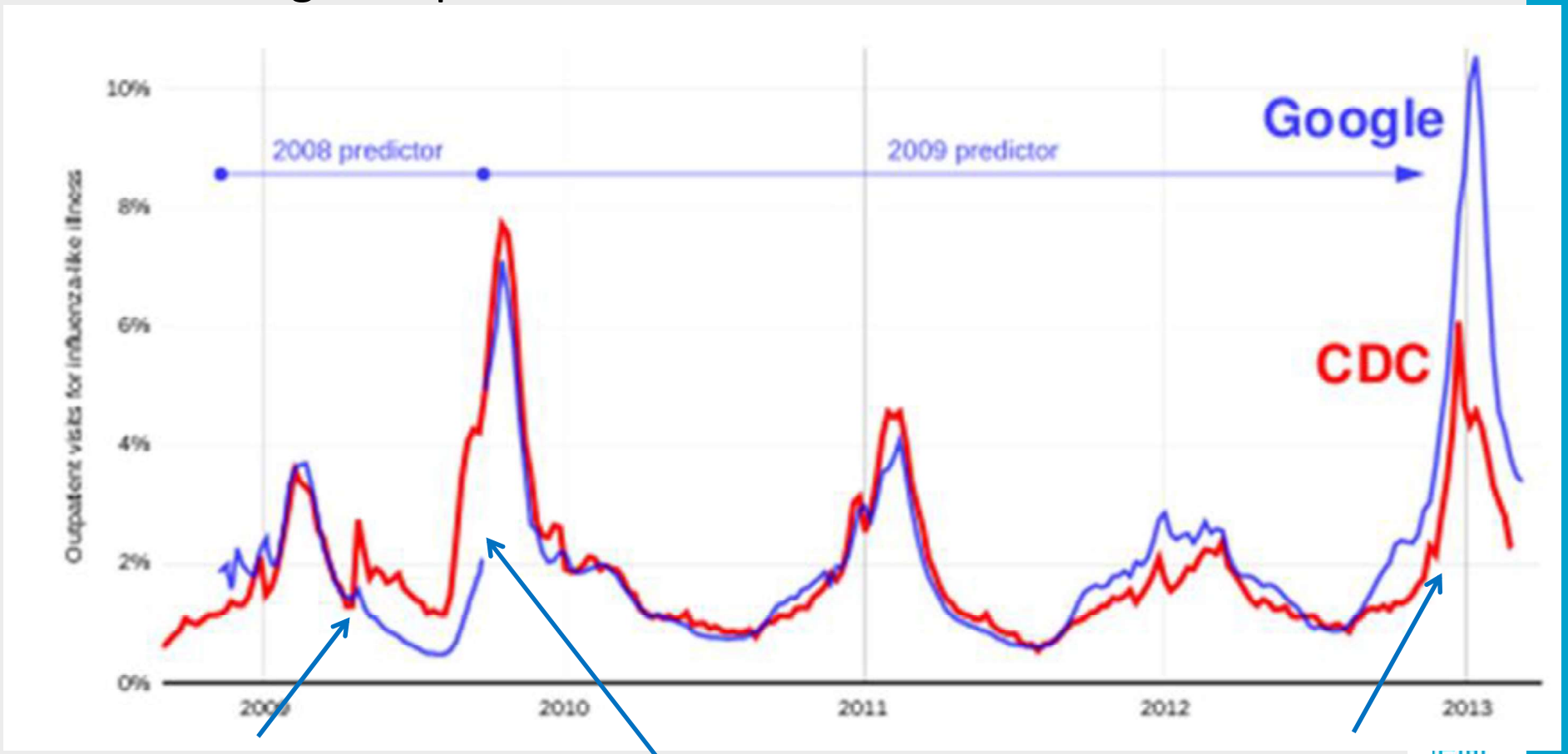
- Google flu prediction findings
- Results look very promising (for 2008 season)
- Correlation is high (~0.9) for search terms used

- But its unclear which search terms are exactly used
- Also selection process is not very transparent

- Lett's see what the future brings!

Google flu future

– Google flu predictions 2009 - 2013



Missed non-seasonal Flu (2009)

Updated model

OEPS!!
2x more predicted

Google flu issues

- Google flu prediction problems
- In 2014 a paper was published in Science on the errors in Google flu prediction
- The Parable of Google Flu: Traps in Big Data Analysis
  - David Lazer[1,2], Ryan Kennedy[1,2], Gary King[2], Alessandro Vespignani[1]

    [1]Northeastern univ. Boston, [2]Harvard Univ.

    Science 343, pp. 2003-2005.

https://gking.harvard.edu/files/gking/files/0314policyforumff.pdf

Google flu issues (2)

- Google flu prediction problems
- Identifies and discusses problems with 'Google search like' approaches
  - Google flu study is interesting but:
    - No replacement for more traditional methodologies
    - More work is needed (its not ready yet)
- "*Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data*"

https://gking.harvard.edu/files/gking/files/0314policyforumff.pdf

Google flu issues (3)

– The following issues were discussed

– 1) *Big Data Hubris*

- Big Data may not be a substitute for, but can be a supplement to traditional data
- Be aware of validity, reliability and dependency issues
  - The need to measure more than just correlation

- The initial 2008 Google flu model was more a 'seasonal detection engine' (hence missing the non-seasonal 2009 second-flu peak)
- Approach can be improved by including other, more reliable, near-real time data

Google flu issues (4)

– The following issues were discussed

– 2) *Algorithm dynamics*

- Stability of relation between terms and google flu results
- The data generation process is affected by
  - Google engineers adjusting/improving the search engine
  - A change in the use of Google search by people

- Purely based on correlation
  - Although some terms were removed for logical reasons

- The publication of the original Google flu paper may even have (negatively) affect this relation

Google flu issues (5)

– The following issues were discussed

– 3) *Transparency, Granularity and All-data*
  - *Transparency and replicability*: work could not be reproduced because the exact terms used were unknown and the search data was not publically available
  - *Use Big Data to Understand the Unknown*: The focus should have been much more on producing high quality, *very local* flu predictions. Less on improving speed.
  - *Study the Algorithm*: A need to focus on chances to the underlying algorithm to better understand the findings.
  - *It's Not Just About Size of the Data*: Focus should also be on the statistical properties of Big Data.

Google flu issues (6)

- Lessons learned
  - Searching for flu ≠ having the flu
  - Peoples behavior and the Google search engine change over time
  - Be transparent in the approach used (method and data)

- But: there is room for improvement
  - Use search results as additional data for CDC-based findings
  - Update model more frequently

Google flu future

- Google flu trends is no longer available
  - Since 2015 the website is off-line
  - Historic series are still available
  - https://www.google.org/flutrends/about/
- Any progress in this area?
  - Use social media (as an additional data source)
    - Work of Prof. Gordon Pipa (Univ. Osnabrück) et al.
      - Combine Google search and social media (US data)
      - Include causal models
    - Work of Dr. M. Puts on Corona
      - Bayesian approach for using social media messages
      - Focus on detecting symptoms (discussed previously)

https://www.ikw.uni-osnabrueck.de/fileadmin/user_upload/ikw_cognitive_computing.pdf

Extracting Information

Extracting information

- Big data does not always directly contain what is needed
  - Need to ‘extract’ this from data/text/picture available
  - For example:
    - Derive ‘turnover’ from tax data available on company
    - Extract ‘innovation’ from the text on company web site
    - Extract sentiment from words in a message
    - Extract price from a picture
    - Extract type of clothing from a picture on the web site
    - Extract background characteristics of units from any ‘data’ available
    - …
  - Especially relevant for non-numeric data (text, images)

Extracting information (2)

- Find ways to obtain/derive background characteristics of units in Big Data sources
  - Make use of the massive amount of data to find clues indicative of important background characteristics of units
  - Use AI/machine learning approaches

- Example: Determine gender of social media users
  - How can we find and extract information relevant for this?

Extracting information (3)

– Obtain background characteristics of units in Big data

- For example: Dutch Twitter users
  - Only a part of the Dutch are active on Twitter
  - But which part?
  - Traditionally so-called ‘background’ characteristics play an important role in deriving this
  - Such as: gender, age, income, level of education etc.

- Is this possible for Big data?
  - Lett’s try it for *gender*

Example

– In the Twitter profile (shown below), there are at least 4 clues indicative for the gender of that person.

TWEETS 1,665 | FOLLOWING 74 | FOLLOWERS 175 | FAVORITES 81 | LISTS 1

Piet Daas

@pietdaas

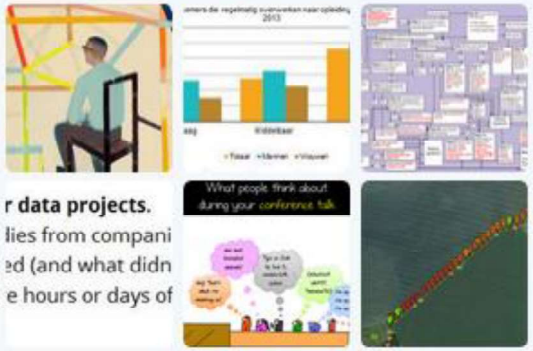
Researcher, Big Data scientist and father of 3.

Eindhoven

about.me/pietdaas

Joined February 2010

40 Photos and videos



Tweets | Tweets & replies | Photos & videos

Piet Daas retweeted

Big Data Network @BigDataNetwork · 16h

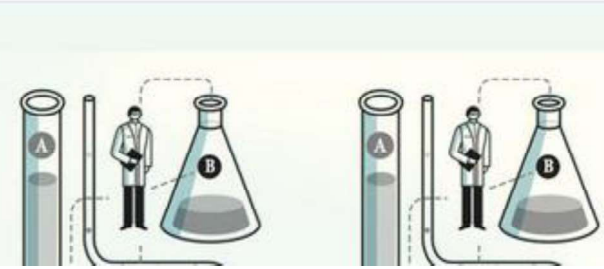
#BigData News » Apache Spark jumps on the R bandwagon: Apache Spark, the big data processi... bit.ly/1w2IqJo via @BigDataNetwork

5 2

Piet Daas retweeted

Dr. Diego Kuonen @DiegoKuonen · Feb 21

Is redoing scientific research the best way to find truth?
sciencenews.org/article/redoin…
#Science #BigData #Reproducibility



4) Picture

Piet Daas
@pietdaas

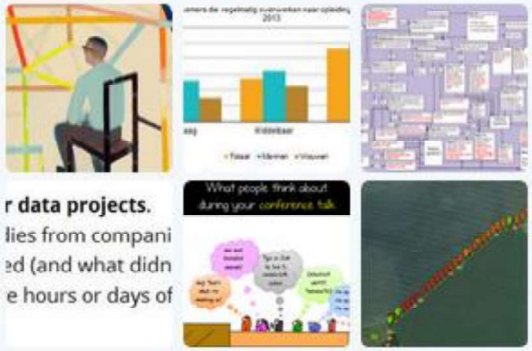
1)Name

Researcher, Big Data scientist and father of 3.

Eindhoven
about.me/pietdaas
Joined February 2010

2) Short bio

40 Photos and videos



r data projects.
lies from compani
ed (and what didn
e hours or days of

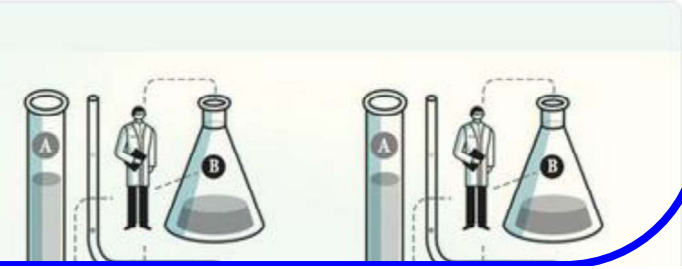
TWEETS 1,665 | FOLLOWING 74 | FOLLOWERS 175 | FAVORITES 81 | LISTS 1

3) Messages content

Tweets | Tweets & replies | Photos & videos

Piet Daas retweeted
Big Data Network @BigDataNetwork · 16h
#BigData News » Apache Spark jumps on the R bandwagon: Apache Spark, the big data processi... bit.ly/1w2IqJo via @BigDataNetwork
5 2

Piet Daas retweeted
Dr. Diego Kuonen @DiegoKuonen · Feb 21
Is redoing scientific research the best way to find truth?
sciencenews.org/article/redoin...
#Science #BigData #Reproducibility



Gender: overall results

	Diagnostic Odds Ratio (log)
First name	4.33
Short bio	-2.70
Tweet content	-1.19
Picture (faces)	0.57*

Diagnostic Odds Ratio =
(TP/FN) / (FP/TN)

Random guessing
log(DOR) = 0

- Multi-agent findings
  - Need 'clever' ways to combine these
  - Take processing efficiency of the 'agent' into consideration

* We did NOT use Deep Learning here

Extracting Information (4)

- From texts (~ text mining)
  - Find the relation between ‘features’ (words, characters) and the phenomenon of interest
  - Usually Machine Learning methods are applied
  - Relation found should make sense
    - For example: Innovation detection on web site texts (discussed in the previous lecture)

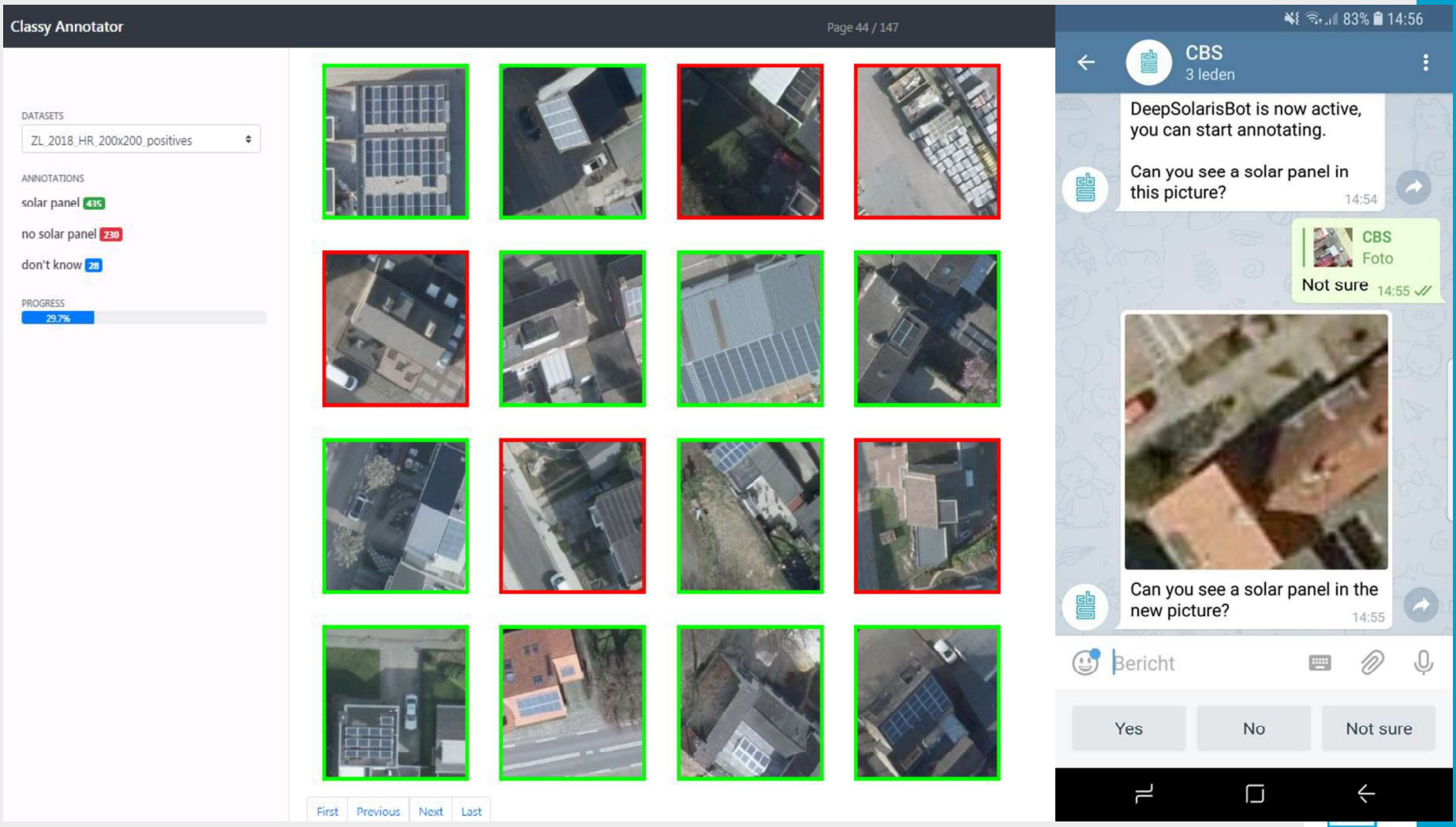
- From images (~ image mining)
  - Find the relation between pixels/image ‘features’ and the phenomenon of interest
  - Nowadays, usually Deep learning is used
  - Transparency of relation found is an issue
    - For example: Detect solar panels on aerial pictures

Detecting solar panels

- Deep Solaris project
- There is a register on addresses that have solar panels in our country (but its incomplete)
- Is it possible to detect solar panels on roofs in aerial pictures?
  - Need a training and test set

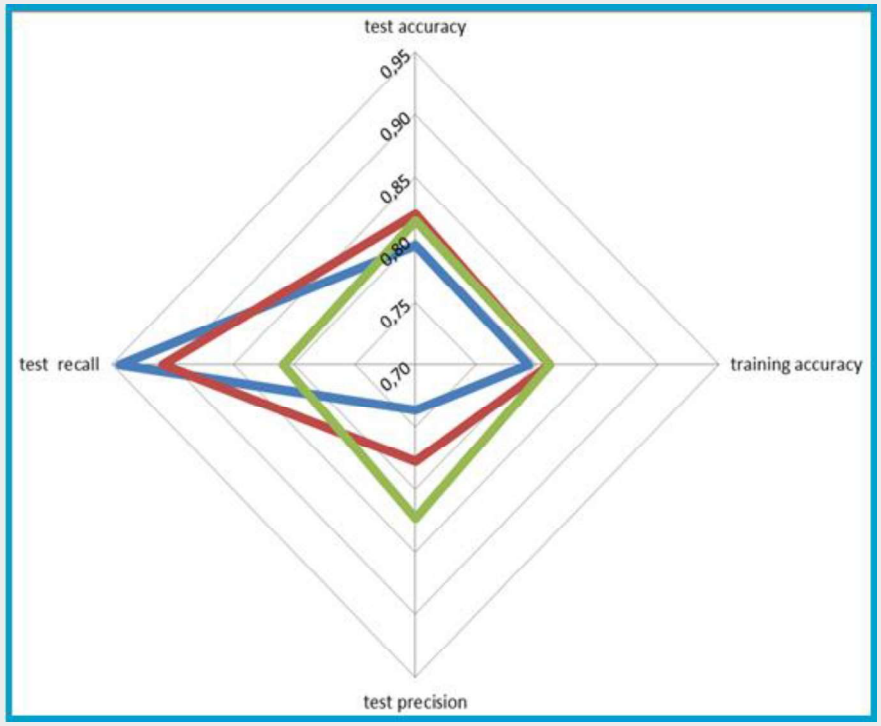
Example of register data: Tax office

Creating a training and test set



Findings

- The Deep Learning model is able to detect solar panels on roofs of houses in city of ‘Heerlen’ very well
- However: the model performs poorly on rooftops in other Dutch cities and also in other countries



IT and Big Data

(short)

IT aspects of Big Data

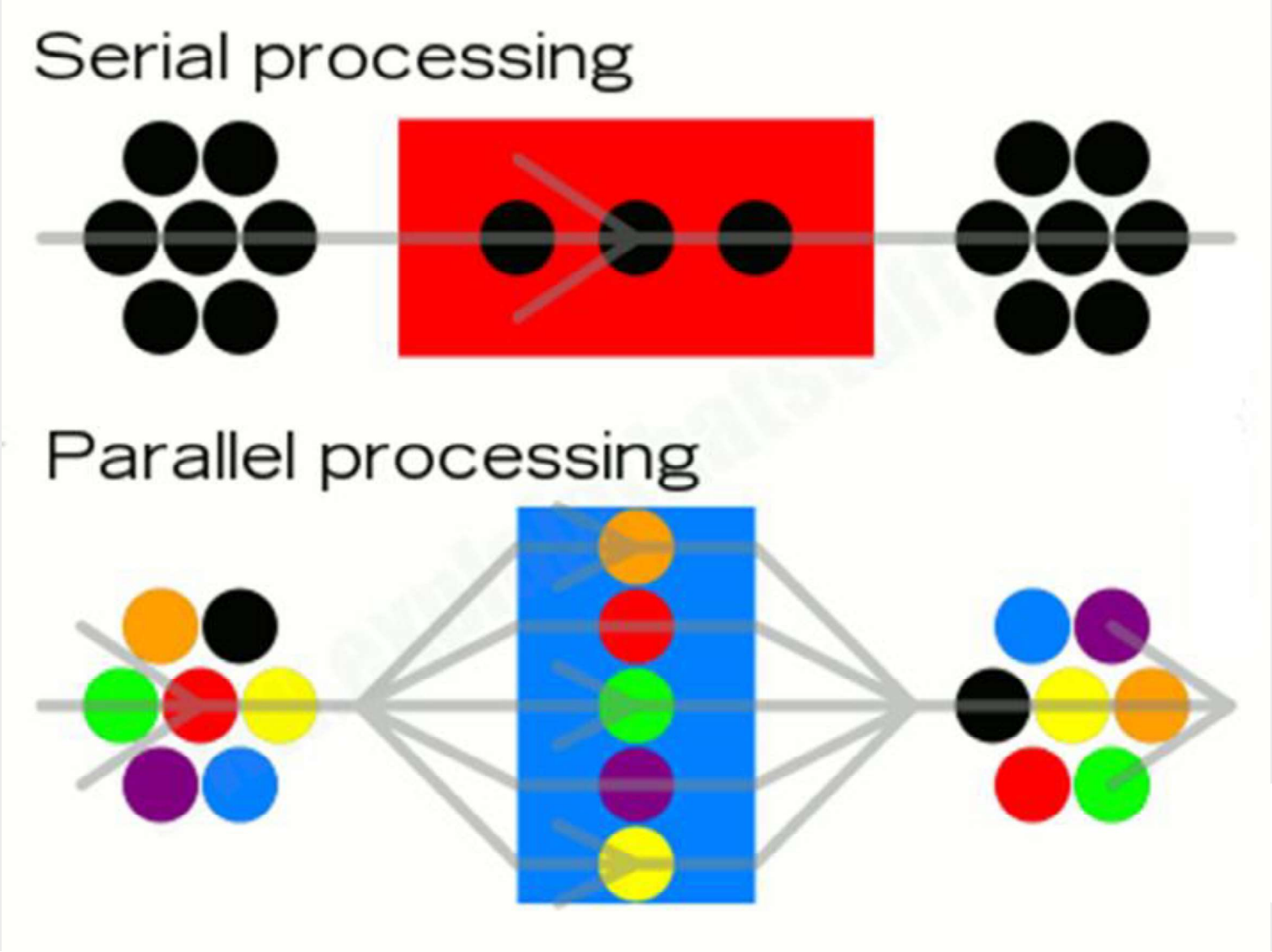
– When studying or, more general, using Big Data there is a need to know more about:
  - Ways to deal with large data sets
  - Ways to speed up analysis/processing
  - Ways to deal with unstructured data and data with a variable structure
  - Infrastructures suited for Big data analysis
  - Software to study Big Data sources
  - …

  - This is the IT-part of Big Data !!
    - We will provide a short overview

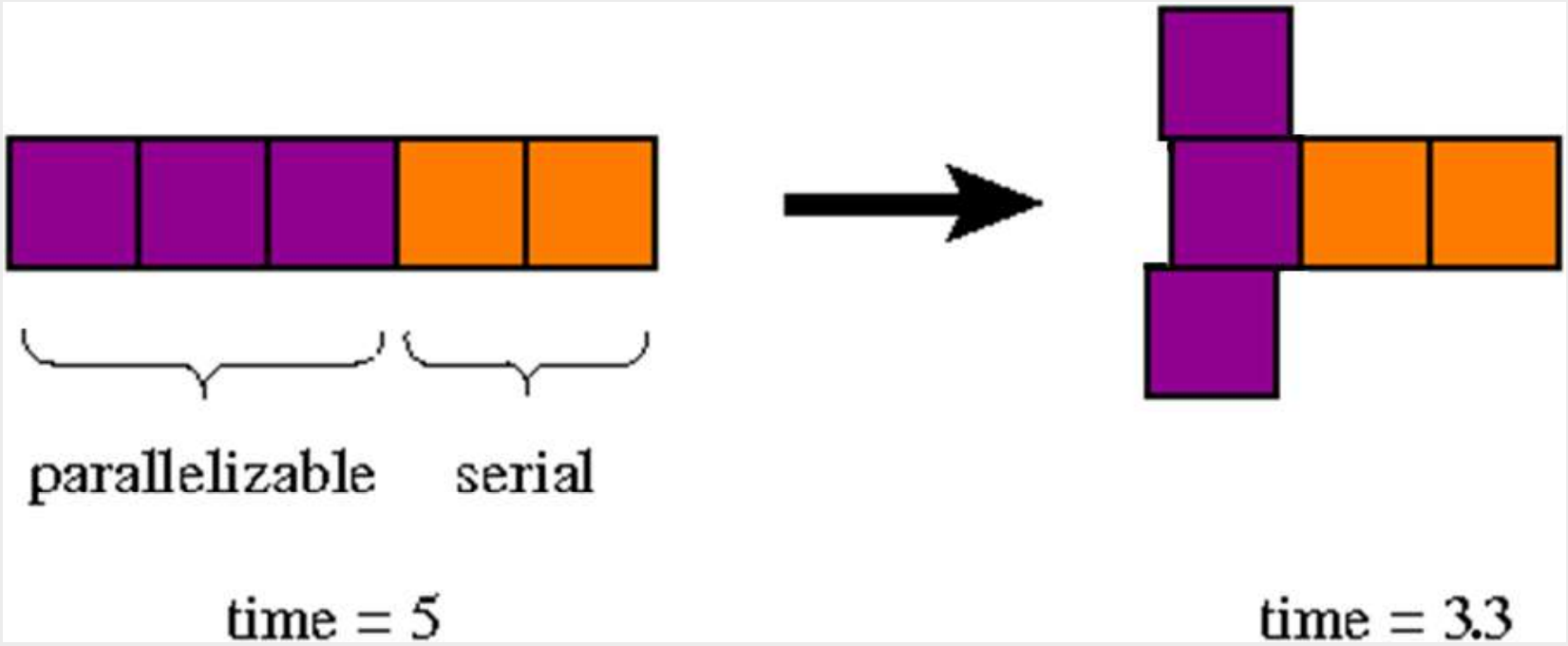
Big Data and speed

- How to speed up the analysis of large amounts of data?
- Analyse data in parallel
  - Use a larger computer
    - Vertical scaling
  - Use more computers
    - Horizontal scaling
- The most simple vertical scaling solution is using all cores available on the system used
  - Most computers used nowadays have 4-8 cores available of which (by default) only 1 is used by R or Python

Single vs. Multicores



When parallization works



Amdahl's Law: a law governing the speedup of using parallel processors on a problem, versus using only one serial processor.

$S(n) = T(1)/T(n)$ $(5/3.3 = 1.52)$

In practise

- Many things can be done in parallel
- Such as:
  - Bootstrapping
  - Reading files
  - Cross validations
  - Checking and editing data
  - Fitting models
  - ….
  - But!
    - There should (preferably) be no interdependencies between those tasks

Big Data Analytics Platforms

Scaling type	Platforms (Communication Scheme)	System/Platform			Application/Algorithm		
		Scalability	Data I/O performance	Fault tolerance	Real-time processing	Data size supported	Iterative task support
Horizontal scaling	Peer-to-Peer (TCP/IP)	★★★★★	★	★	★	★★★★★	★★
	Virtual clusters (MapRedce/MPI)	★★★★★	★★	★★★★★	★★	★★★★	★★
	Virtual clusters (Spark)	★★★★★	★★★	★★★★★	★★	★★★★	★★★
Vertical scaling	HPC clusters (MPI/Mapreduce)	★★★	★★★★	★★★★	★★★	★★★★	★★★★
	Multicore (Multithreading)	★★	★★★★	★★★★	★★★	★★	★★★★
	GPU (CUDA)	★★	★★★★★	★★★★	★★★★★	★★	★★★★
	FPGA (HDL)	★	★★★★★	★★★★	★★★★★	★★	★★★★

Sing, D., Reddy, C. (2014). A survey on Platforms for Big Data Analytics. Journal of Big Data 2014, 1:8.

Example: speed-up obtained

Table 2. Speed of various ways of processing NDW-data in R

Script nr.	Read-stage	Processing-stage	Write-stage	Time (min)	Speed-up
Script1	read.table	data.frame	write.table	110.9	-
Script2	read.table	data.table	write.table	108.2	1.02x (2.4%)
Script3	fread	data.table	write.table	20.8	5.3x (81%)
Script4	fread	data.table	fwrite	21.1	5.3x (81%)
Script5	fread (parallel 8)	data.table	fwrite	12.8	8.7x (88%)
Script6	fread (parallel 8)	data.table (parallel 8)	fwrite	12.2	9.1x (89%)
Script7	fread (parallel 8)	data.table (parallel 8)	fwrite (parallel 8)	8.6	13.0x (92%)
Script8	zip-> fread (parallel 8)	data.table (parallel 8)	fwrite (parallel 8)	6.7	16.6x (94%)
Script9	zip-> fread ---------->	data.table (parallel 8)	fwrite (parallel 8)	6.1	18.2x (94%)
Script10	zip-> fread ---------->	data.table (parallel 16)	fwrite (parallel 16)	5.7	19.4x (95%)

Multicore part: Script 4 vs. Script 10, speed diff. 21.1/5.7 = 3.7x

Big Data Methodology
(work in progress)

Big Data Methodology

- Work in progress, gradually being obtained from the many use cases

- Important topics
  - *Information extraction from text and images*
  - Big Data exploration (visualization methods)
  - Combining sources
  - Dealing with Errors
  - Inference

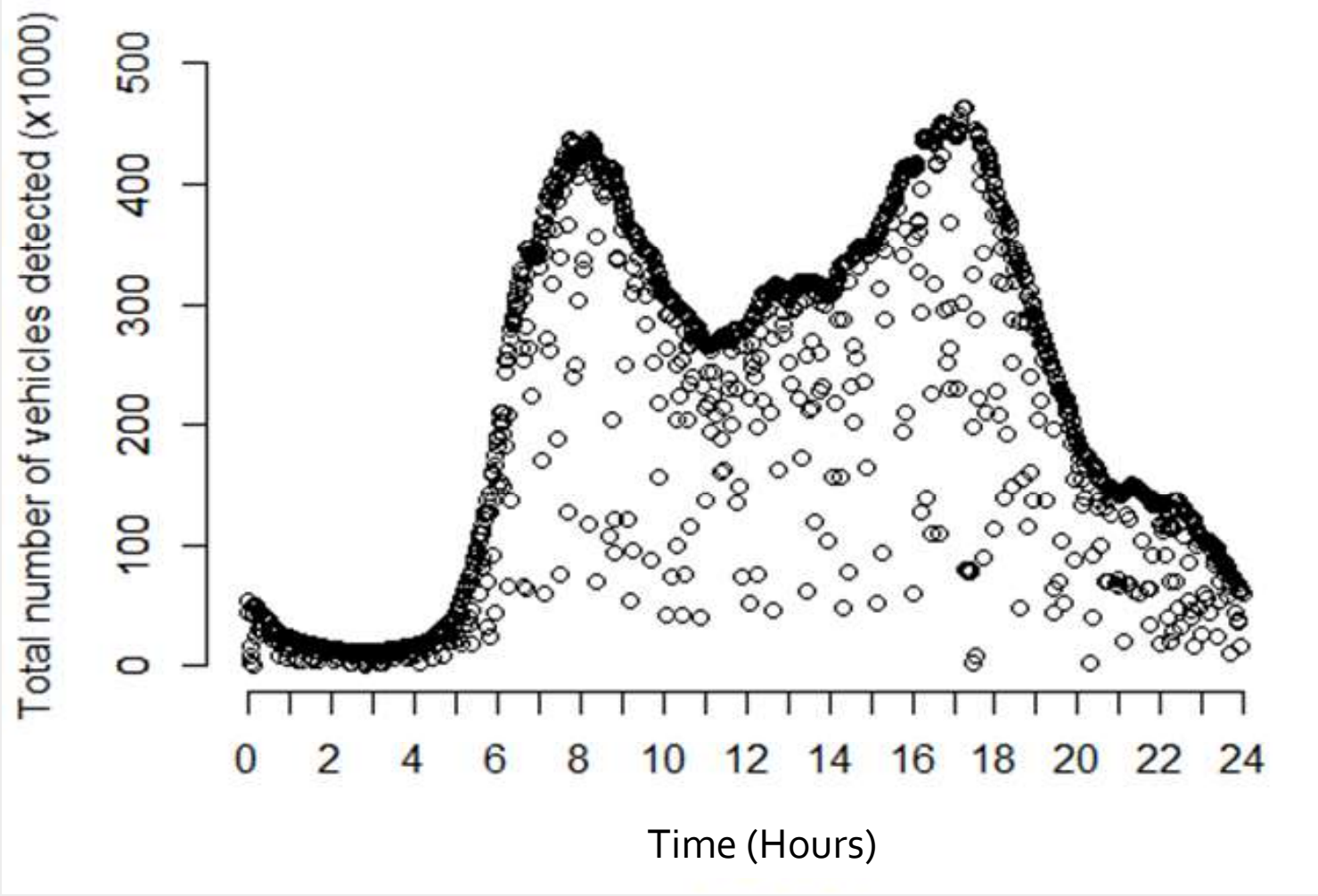
Big Data explorations

- Visualization methods really help when working with large data files
  - A good IT environment also helps

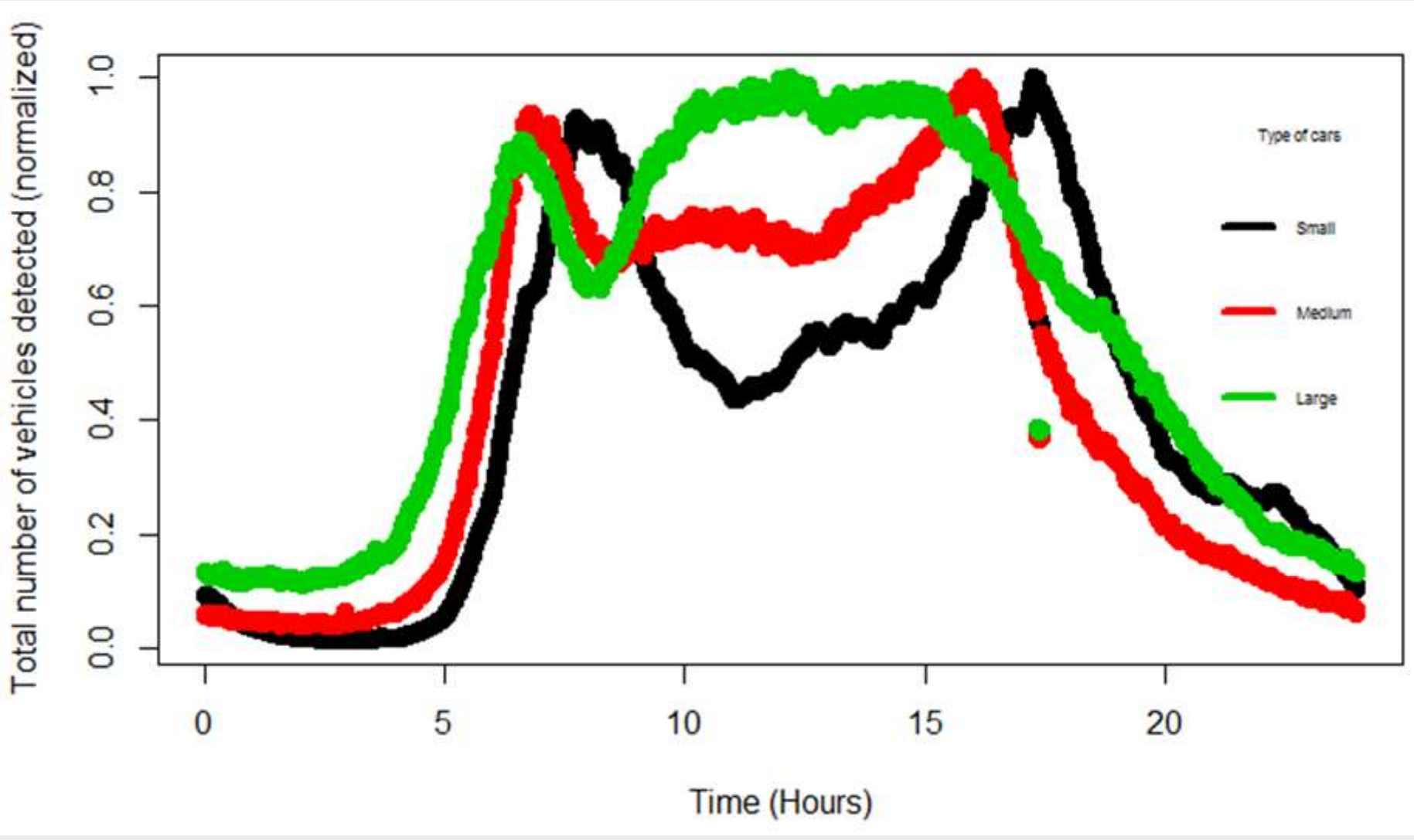
- Some examples of visualization that provide insights

Road sensors: all vehicles during a day

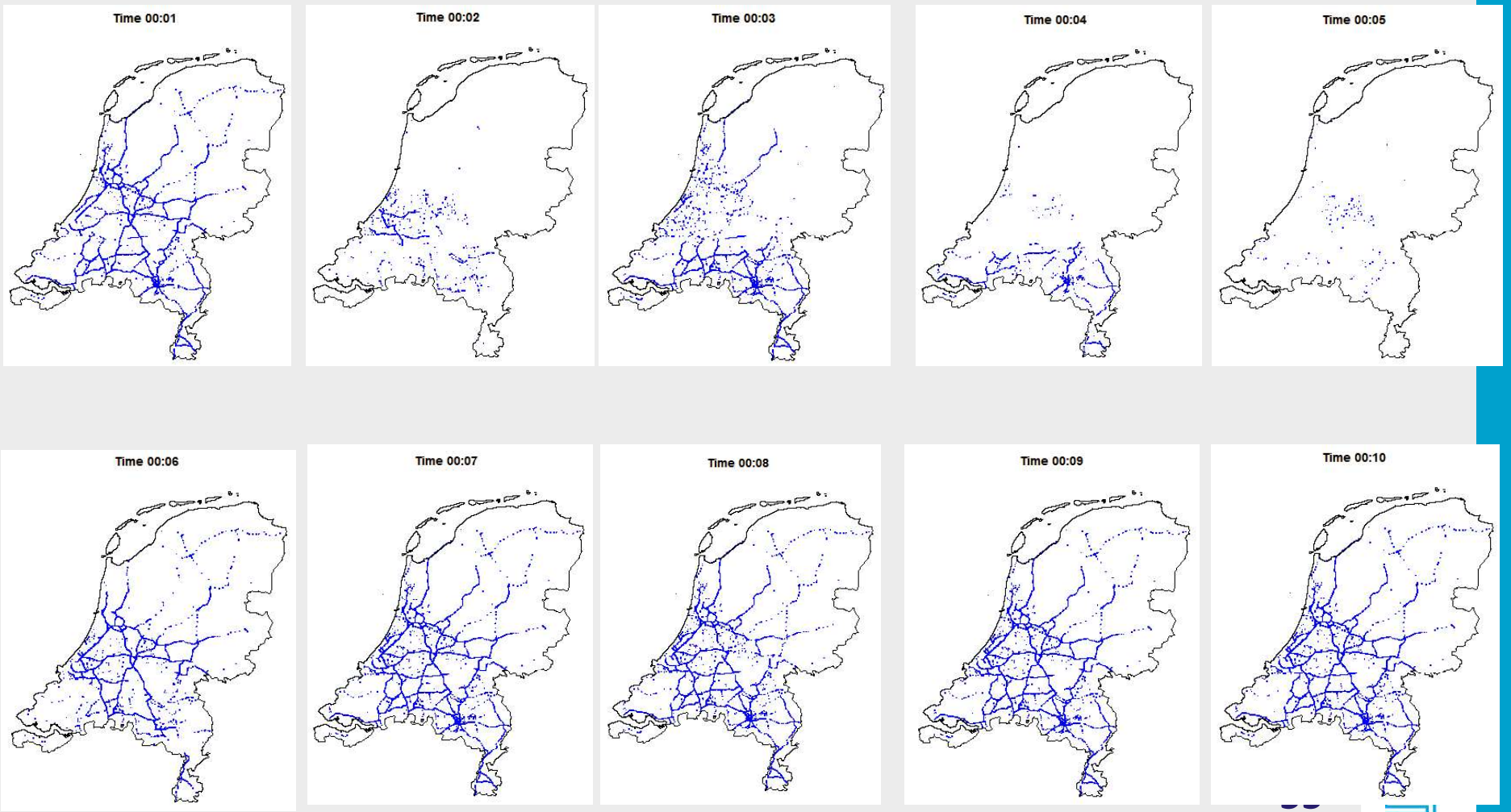
Total number of vehicles (for the whole country) detected by road sensors



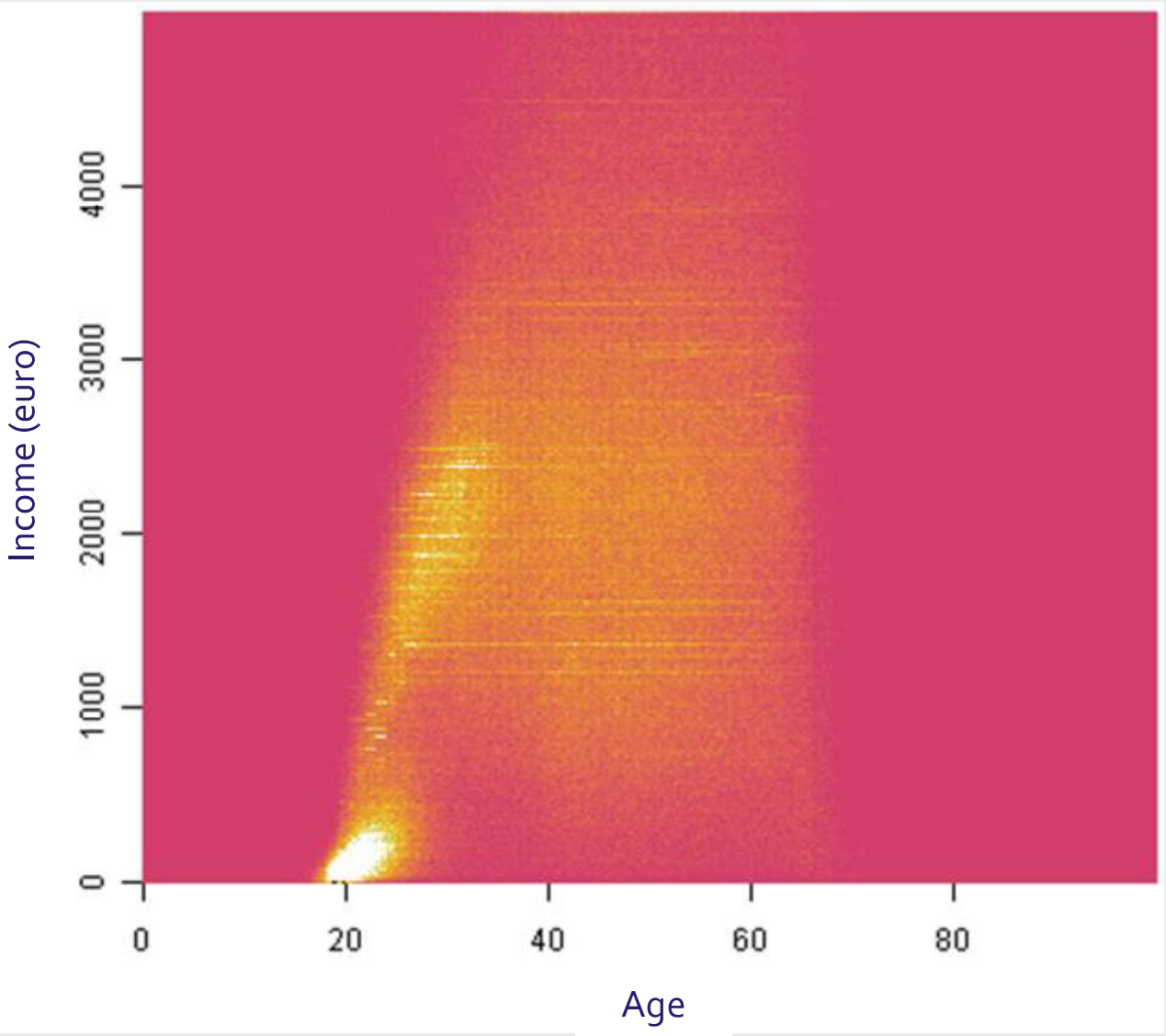
Road sensors: small, medium, large

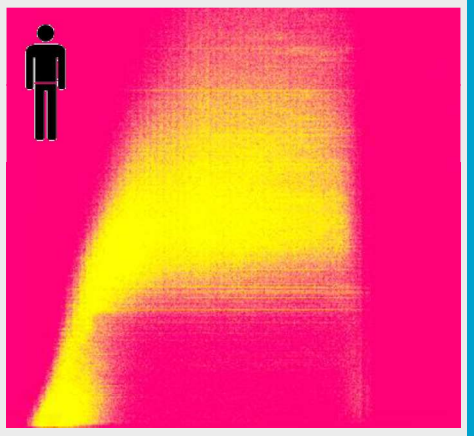


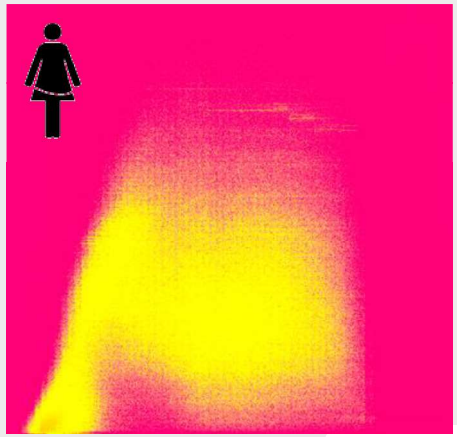
Small multiples: sensor activity



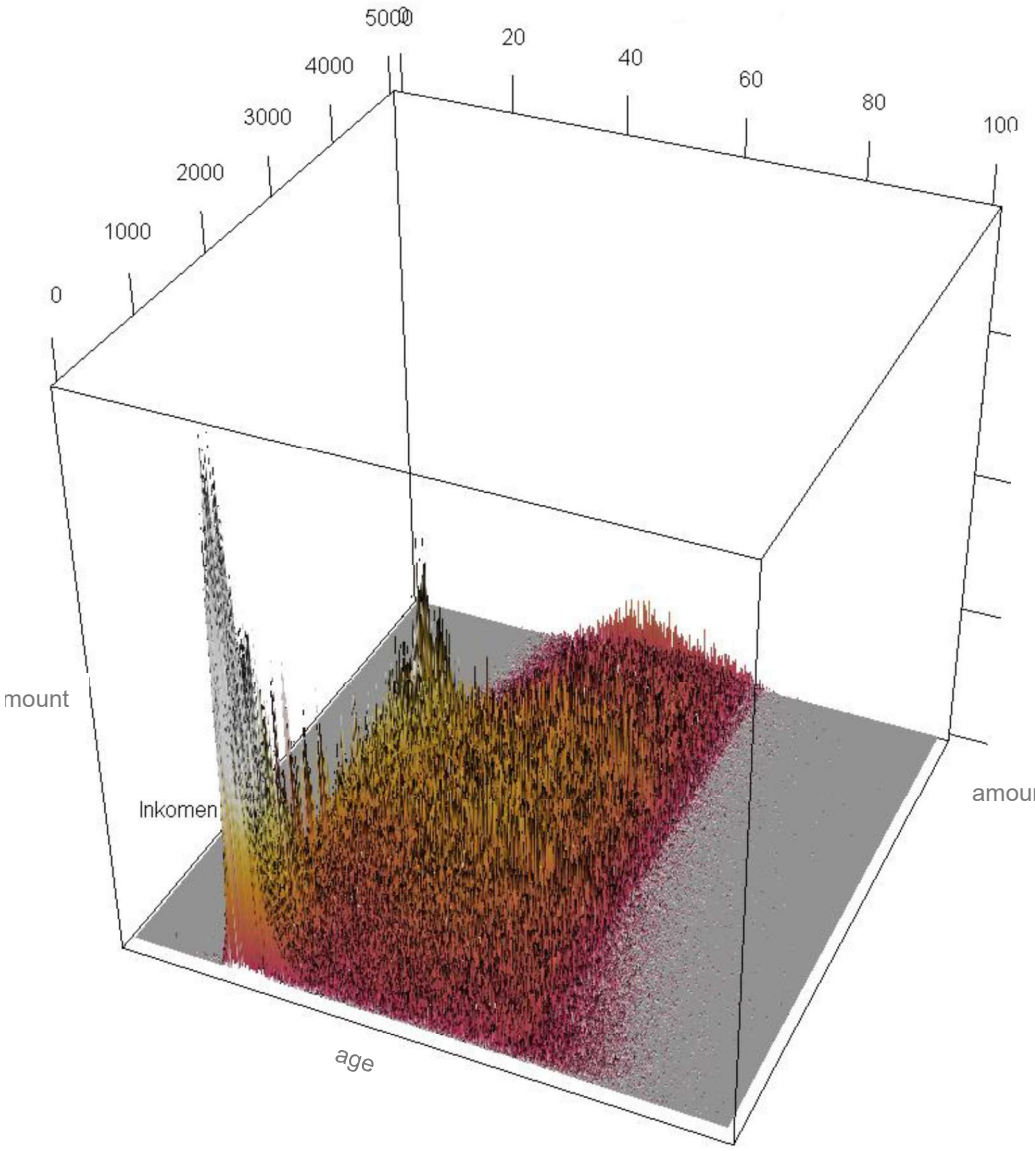
Heatmap: Age vs. income



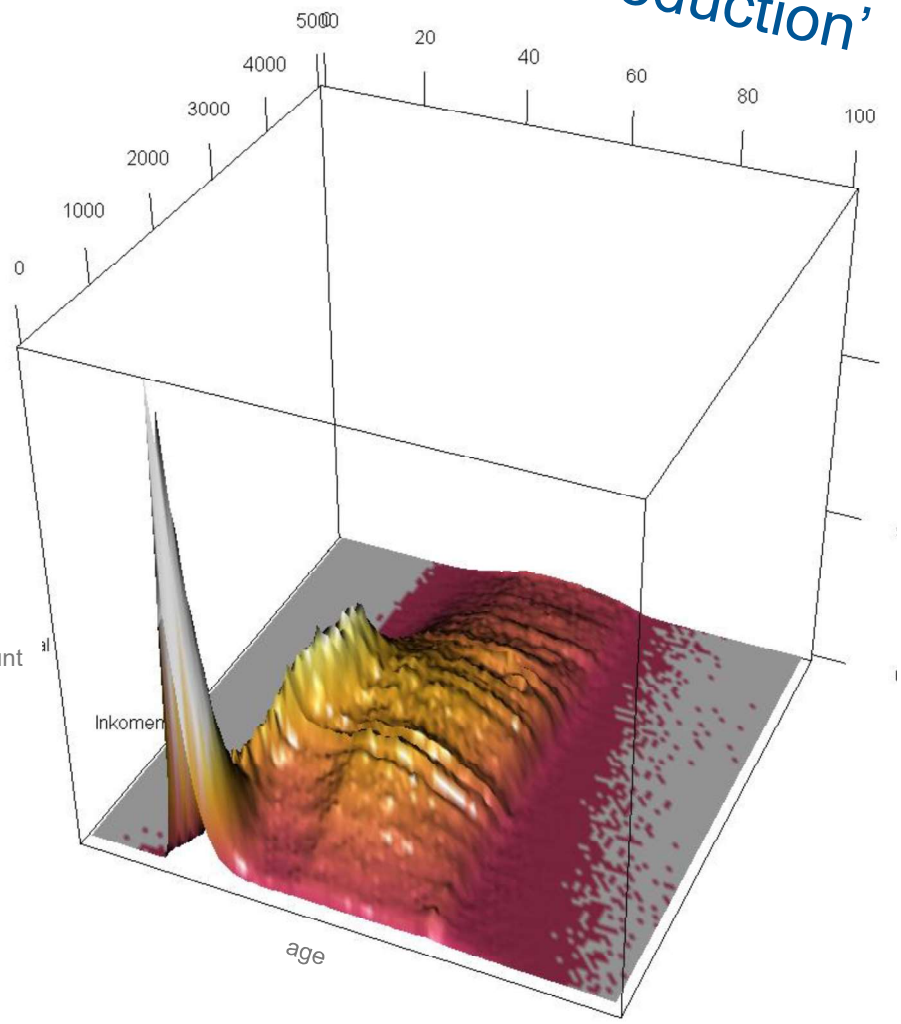




Heatmap 3D: Age vs. Income



After 'data reduction'



Combining sources

- In many Big Data studies, data sources are combined with other sources (BD, admin data, survey data)
  - At the unit-level
  - At the geo-location level (areas)
  - At higher aggregated levels

- Increases both coverage and quality of measurements

- Linking Big data is challenging because
  - These sources usually lack information needed for linking
  - The units included may not be identical to the statistical units of the researcher

Combining sources (2)

- Linking Units
  - Preferably with a unique id
  - Deterministic, Probabilistic, Machine Learning based
  - Using addresses works well (when standardized)
  - Geolocation information also works well

- At higher aggregated levels
  - Groups, region, country
  - Reduces complexity of the linking process
  - Usually increases coverage

- Try to use as much of the information that is available
  - A multidisciplinary approach is highly recommended
  - Try to add information extracted in an indirect way

Machine Learning in Official Statistics

Marco Puts

Quality of Official Statistics

- Relevance
- Accuracy
- Accessibility
- Clarity
- Coherence
- Comparability

Accessability and clarity

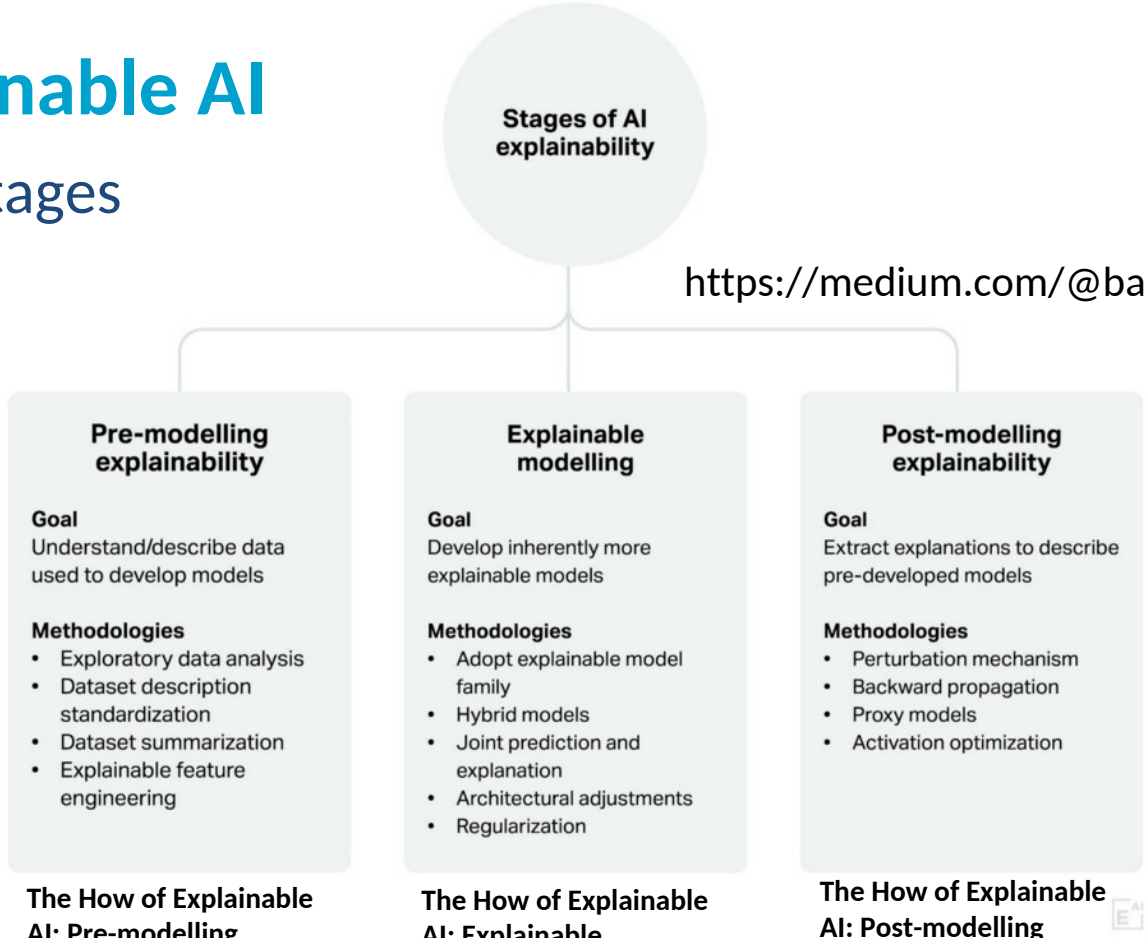
eXplainable AI

Validation

- The best way to validate a model is by understanding
- Marr (1982): Three levels at which an information -processing device should be described to be fully understood:
  - Computational Theory (How does the model relate to the reality?)
    - What is the goal?
    - Why is it appropriate?
    - Logic of the strategy?
  - Representation and algorithm (Design Pattern)
    - Input/output
    - Algorithm
  - (hardware) Implementation (How is it realized?)

eXplainable AI

Three Stages



https://medium.com/@bahador.khaleghi

The How of Explainable AI: Pre-modelling Explainability

The How of Explainable AI: Explainable modelling

The How of Explainable AI: Post-modelling Explainability

Coherence and Comparability

Coherence and Comparability

How well is the model able to give a stable result over time (and space)?

- Domain Specificity of the Model
- Correlation and Causation
- Concept drift

Domain Specificity of the Model

- Model is trained for a certain domain
- Translates in the distributions in the feature space

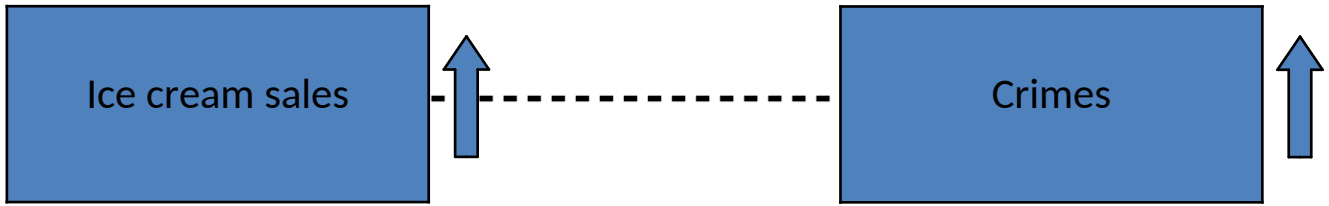
Domain Specificity of the Model

Domain Adaptation

- Divergence based Domain Adaptation
  Try to minimize the difference of distributions between features in two domains
- Adversarial based Domain Adaptation
  Create generator that makes source and target domain indiscriminable, which is detected by a trained discriminator. (GAN architecture)

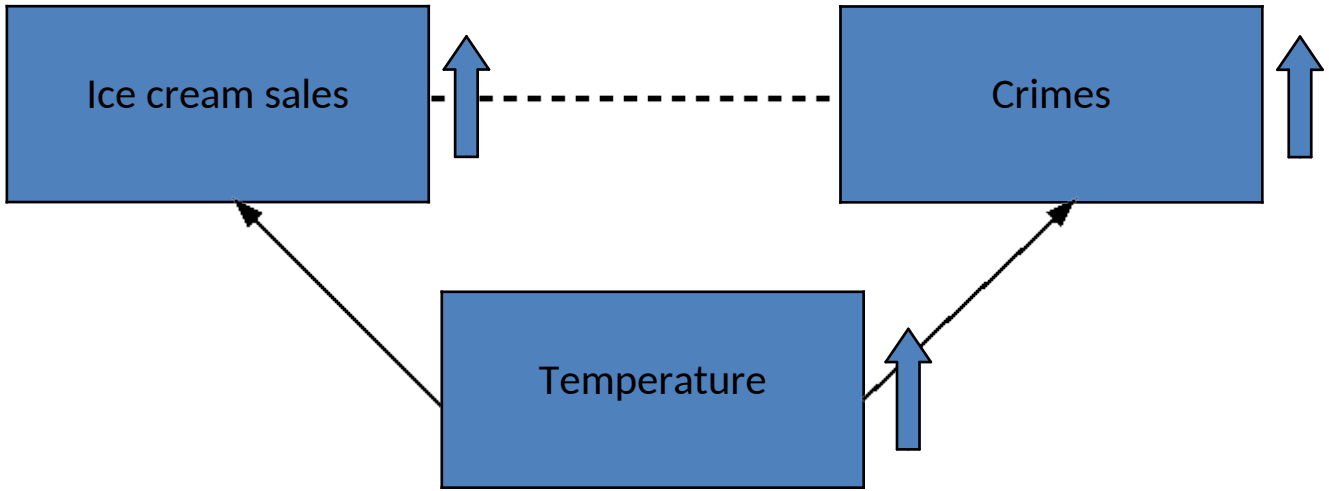
Correlation and Causation

- When applying Machine Learning, interpretation becomes even more important
- Be aware that: “Correlation does not imply Causation“
- Correlation merely checks the association between variables as do ML & AI
- The associations found do not have to be causal!

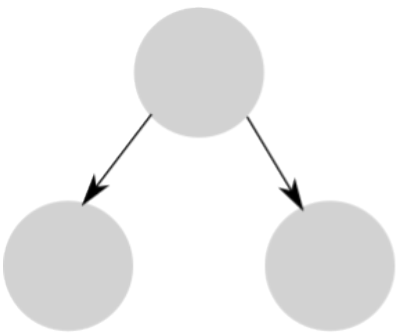


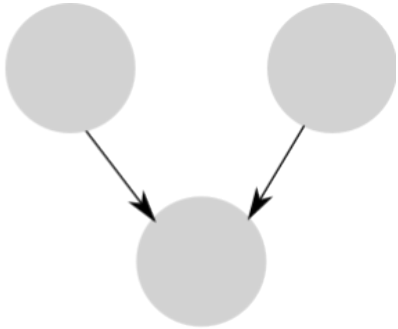
Correlation and Causation

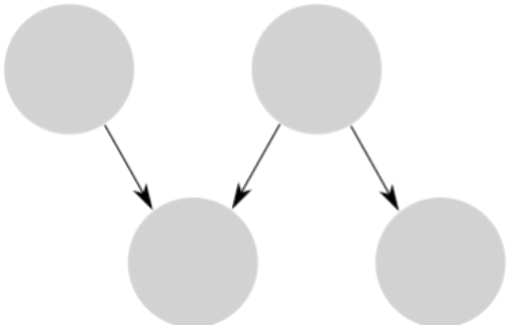
- When applying Machine Learning, interpretation becomes even more important
- Be aware that: “Correlation does not imply Causation“
- Correlation merely checks the association between variables as do ML & AI
- The associations found do not have to be causal!



Correlation and Causation



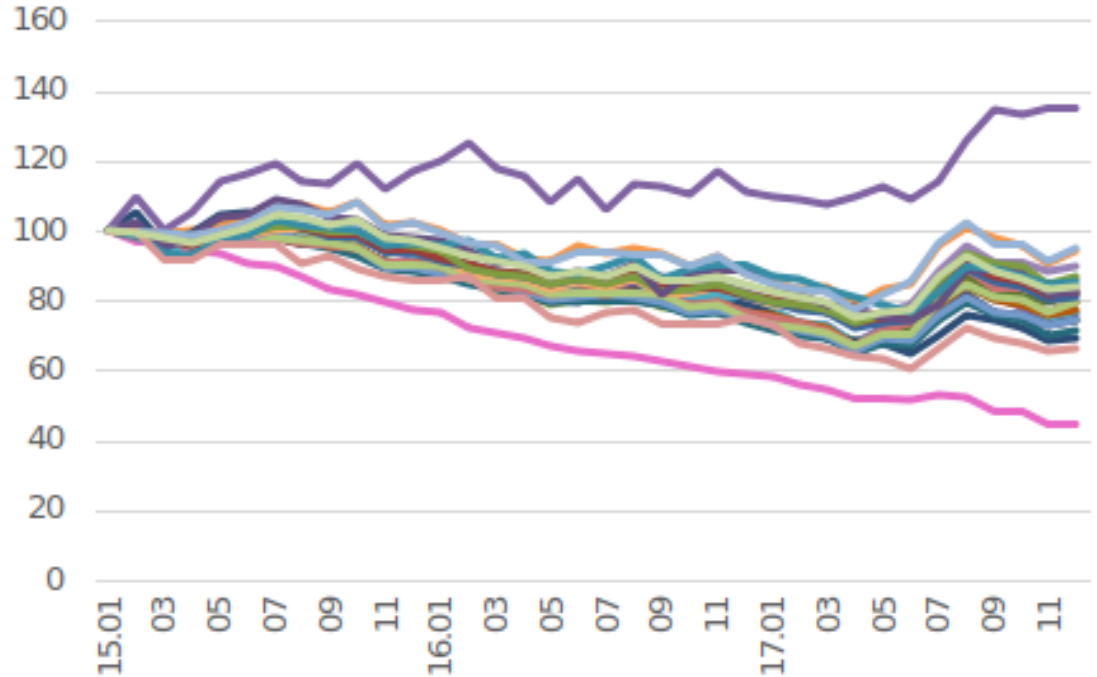




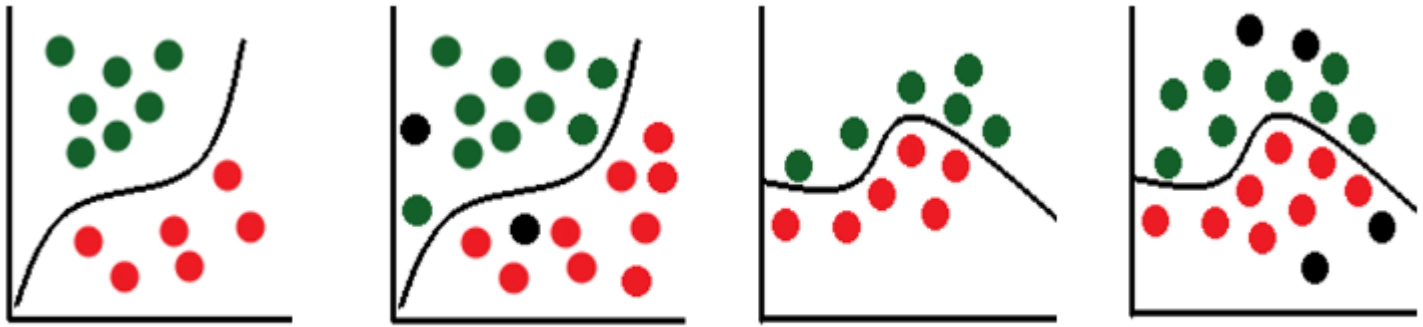
- Counterfactuals
- Granger Causality
- Colinearity

Concept drift: example from Dutch CPI

price of televisions



Concept drift



1) Original 2) Virtual Drift 3) Real Drift 4) Hybrid Drift

Source: Jameel et al., 2020

Concept drift

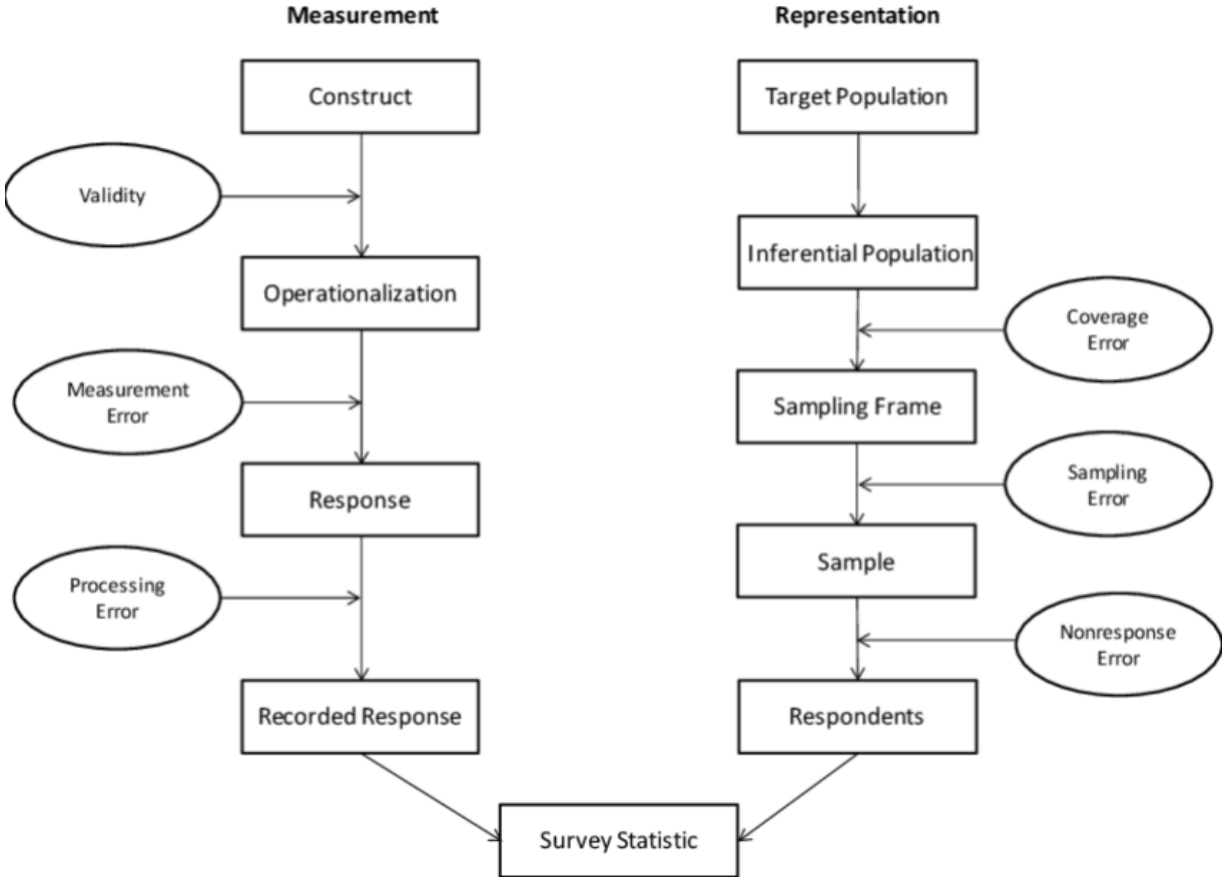
- refitting and updating the model
- Weigh data to age
- Create new models over time and learn
- Detect the phase of concept drift
- Remove trends and seasonal effects in the data

Accuracy and reliability

Accuracy and reliability

- Annotated data and representativity
- Misclassification

Annotated data and representativity



Representativity of training sets

get the right set of features

Sampling methodology is a valid way to overcome this:

- (Stratified) Random Sampling in the population

- Finding strata:
  - Clustering features
  - Using background information
  - Stratify and weighing or multiple models?

Annotated Data Set

To what extend is the “observed Ground Truth” real?

Ground Truth

- Underdetermination Problem
- most things cannot be measured directly

Three questions:

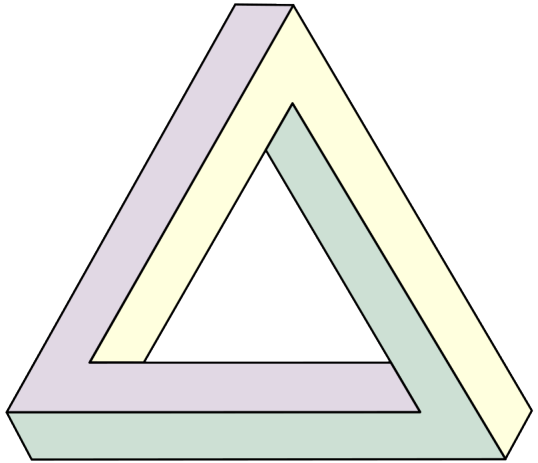
1. Does the GT exist?
2. Is there a procedure to find it?
3. Is the procedure applicable?

Often hard to answer these questions

personal comunication with Prof. Pim Haselager

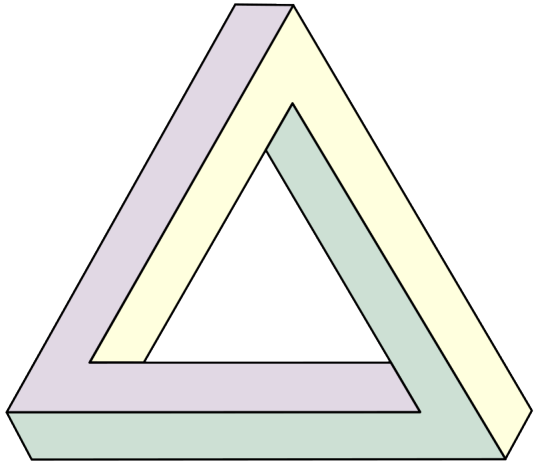
The asymptotical behavior towards annotated data

The ML algorithm can never outperform the annotator, since it will learn the mistakes of the annotator.



The asymptotical behavior towards annotated data

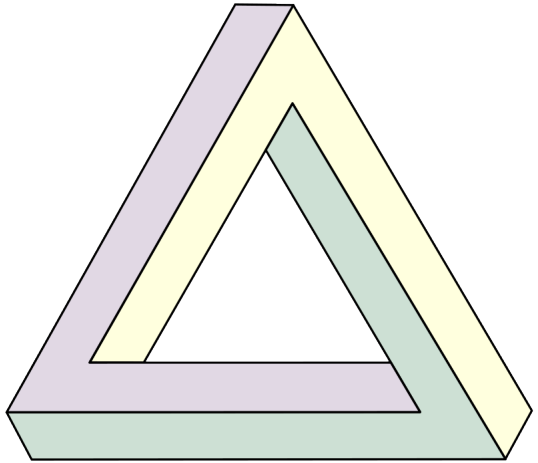
- Perceptual limits of the annotator
- Perception bias



The asymptotical behavior towards annotated data

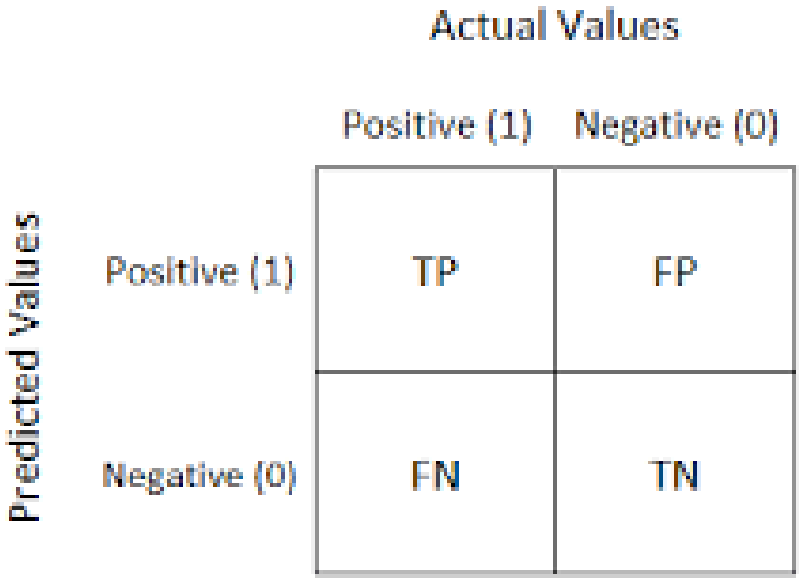
Mistakes in annotation are present in:

- Training set
- Test set
- Validation set



So how to detect these errors?

Missclassification



Conclusion

Quality of Official Statistics

- Relevance
- Accuracy
- Accessibility
- Clarity
- Coherence
- Comparability

- New research topics within machine learning appear due to applications in Official Statistics!

Facts that matter

SPONSORED BY THE

Federal Ministry of Education and Research

AUFSTIEG DURCH BILDUNG
OFFENE HOCHSCHULEN

These slides were created as part of IPSDS.
IPSDS is sponsored by the German
Federal Ministry of Education and Research (BMBF) within the framework of the program „Aufstieg durch Bildung: offene Hochschulen“

Uploaded as part of the BERD Academy offers within BERD@NFDI,
funded by the DFG (460037581).