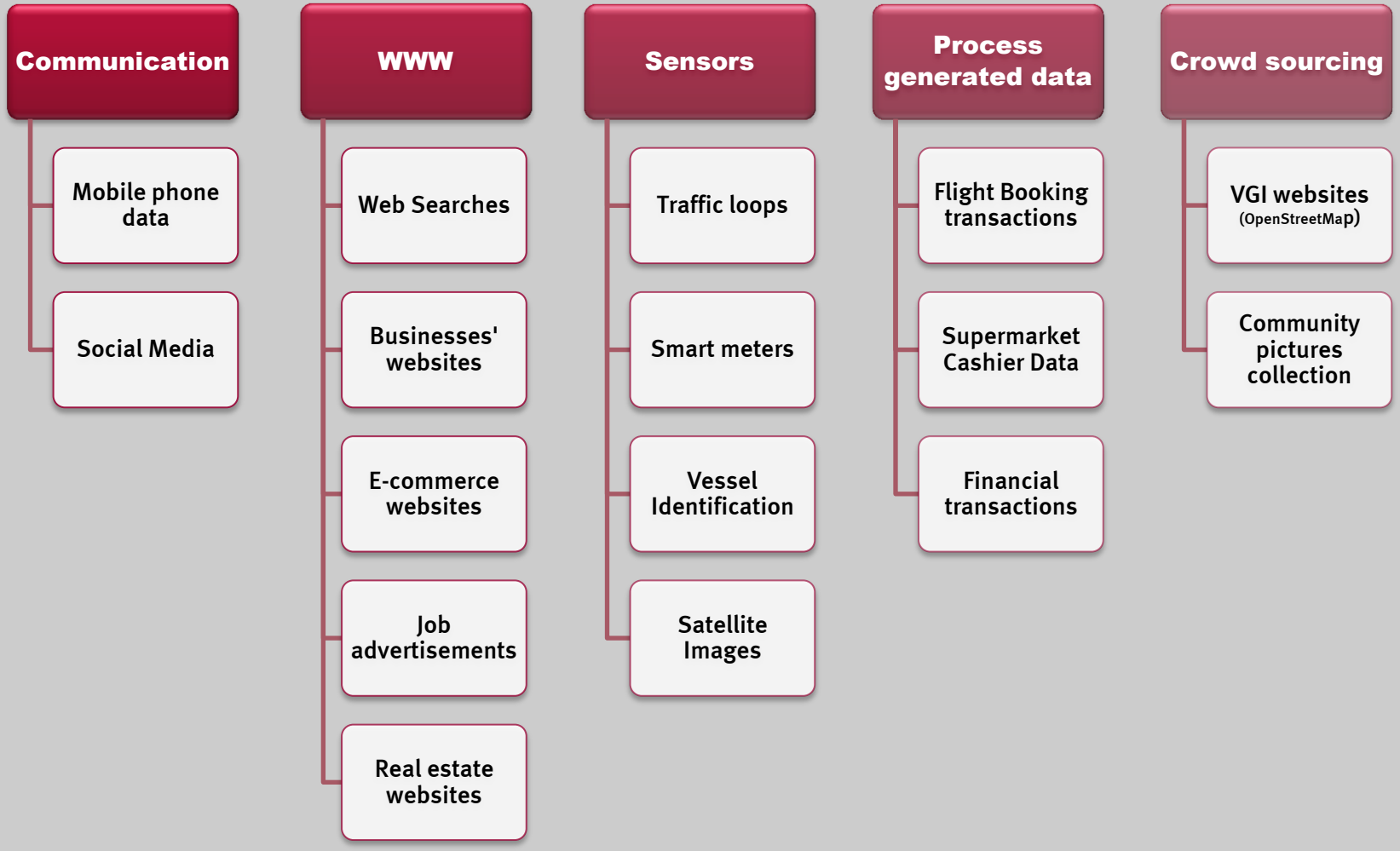
EMOS WEBINAR BIG DATA II: EXAMPLES FROM NATIONAL STATISTICAL INSTITUTES

Markus Zwick
Federal Statistical Office Germany,
Institute for Research and Development
in Official Statistics
14 June 2017

The data deluge



Agenda for today

- ESSnet Big Data
- Mobil Phone Data
- Satellite Data
- Enterprise data from the internet
- Statistical education in times of Big data

European Statistical System (ESS)

ESS Steering Group Big Data

- Oversees the implementation of the ESS Big Data Action Plan and Roadmap (BDAR)
- Identifies priorities from Member States BDAR at national level

ESS Task Force Big Data

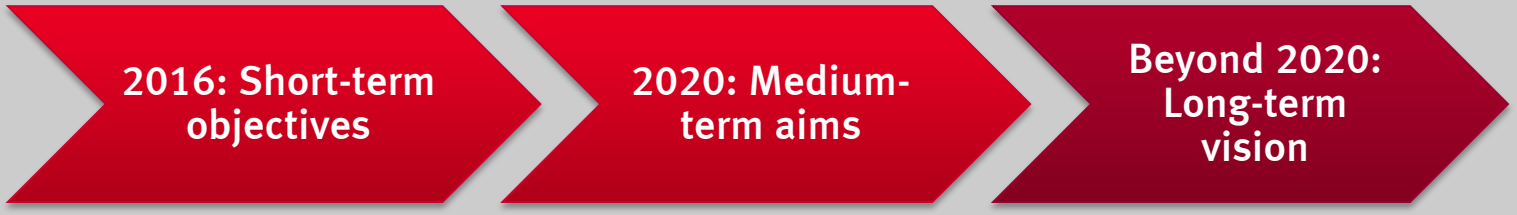
- Identifies priority actions and formulates a project proposal
- Manages and co-ordinate the implementation of the ESS Big Data Action Plan and Roadmap

ESSnet Big Data

- 22 national statistical institutes and organizations working together on different digital data projects

ESS Big Data Action Plan and Roadmap

- Vision: integration of big data sources into statistical production process beyond 2020
- Long to short-term objectives
- Implementation via procurement contracts
- ESSnet Big Data Project: a network of several organisations from the ESS working together on pilot studies in the Big Data field



ESSnet Big Data Project

02/2016 – 04/2018

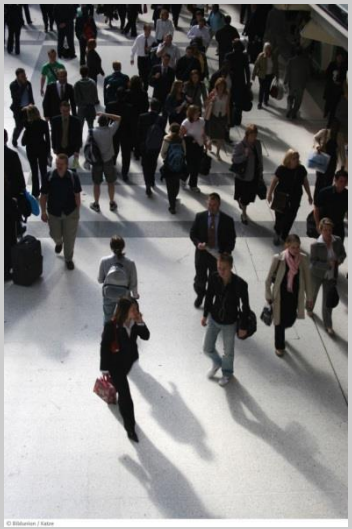
Work packages

(1) Web scraping of job vacancies
(2) Web scraping of enterprise characteristics
(3) Smart meters
(4) AIS Vessel Identification Data
(5) Mobile phone data
(6) Early estimates
(7) Multiple domains
(8) Methodology

https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/Main_Page

WP1: WEB SCRAPING OF JOB VACANCIES

- Mix of sources including job portals, job adverts on enterprise websites and job vacancy data from third party sources
- Data access
  - identifying the most important national job websites
  - legal aspects and copyright: who owns the data?
  - implementation of tools and technical infrastructure
- Data handling
  - removing duplicates
  - excluding the records which are not eligible
  - classification of job vacancies
  - quality assessment



WP2:WEB SCRAPING OF ENTERPRISE CHARACTERISTICS

- **Purpose:**

  investigate whether webscraping, text mining and inference techniques can be used to collect, process and improve general information about enterprises

- **Aim:**

  to demonstrate whether business registers can be improved by using webscraping techniques and by applying model-based approaches in order to predict the values of some key variablesfor each enterprise

  to verify the possibility of producing statistical outputs using predicted data, in combination or not with other sources of data

- **WP1 coordinates its work with WP2**

WP3: SMART METERS

- Smart meters = electricity meters which can be read from a distance and measure electricity consumption at a high frequency
- Possible usage:
  - for the production of energy statistics,
  - as an additional source at calculating:
    - census housing statistics
    - household costs
    - impact on environment

WP4: AIS DATA

- Real-time measurement data of ship positions, measured by the so-called AIS (Automatic Identification System)
- Possible usage
  - to improve the quality and internal comparability of existing statistics
  - for new statistical products relevant for the ESS

- Example
  - Developing a reference frame of ships and their travels in European waters and then linking this reference frame, by ship number, to register-based data about marine transport from port authorities.

WP5: MOBILE PHONE DATA

- Potential of mobile phone data as a data source for official statistics

Objectives:

- To gather the current experiences about data access to mobile phone data in the ESS
- To obtain access to the mobile phone data for their exploitation
- To propose some specific complementary potential outputs for official statistics based on this source of information

WP6: EARLY ESTIMATES

- Aim: to investigate multiple big data, administrative and other existing sources in order to produce early estimates for statistical purposes
- Expected output: guidelines and recommendations regarding usage of big data sources in the area of early estimates
- Focus: two concrete domains which have potential of getting “quick wins”:
  - social media data, newsfeeds and survey data for the aim of consumer confidence index
  - web-based sales inquiries for the aim of nowcasts of turnover indices

WP7: MULTIPLE DOMAINS

- Aim:
  - to investigate how a combination of big and administrative data as well as survey data can be used to improve current statistics and create new statistics in statistical domains.
- Challenges: representatively issues, record linkage and statistical matching, metadata, international comparability.

WP8: METHODOLOGY

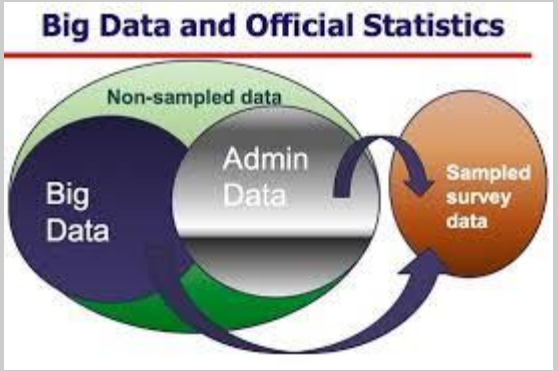
- Focus: Methodological aspects of the big data project
- Methodological aspects can only be addressed when the other work packages have laid the necessary foundations
- First meeting: April 2017

Questions

- What do you think are the most burning questions concerning Big Data and Official Statistics?
- What are the components of Big Data and Official Statistics?

Topics concerning Big Data and Official Statistics

- Legal and ethical issues
- Quality
- Methodology
- Data owner ship
- Skills of the future statistician
- Smart statistics



MOBILE PHONE DATA

France: Dynamic Population Mapping

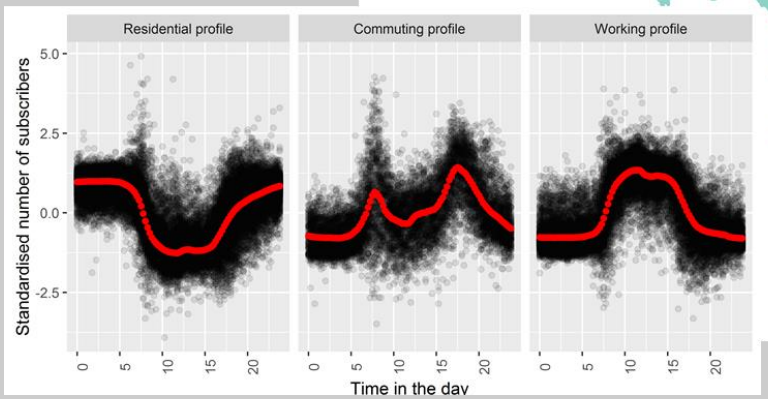
https://www.youtube.com/watch?v=qsUDH5dUnvY

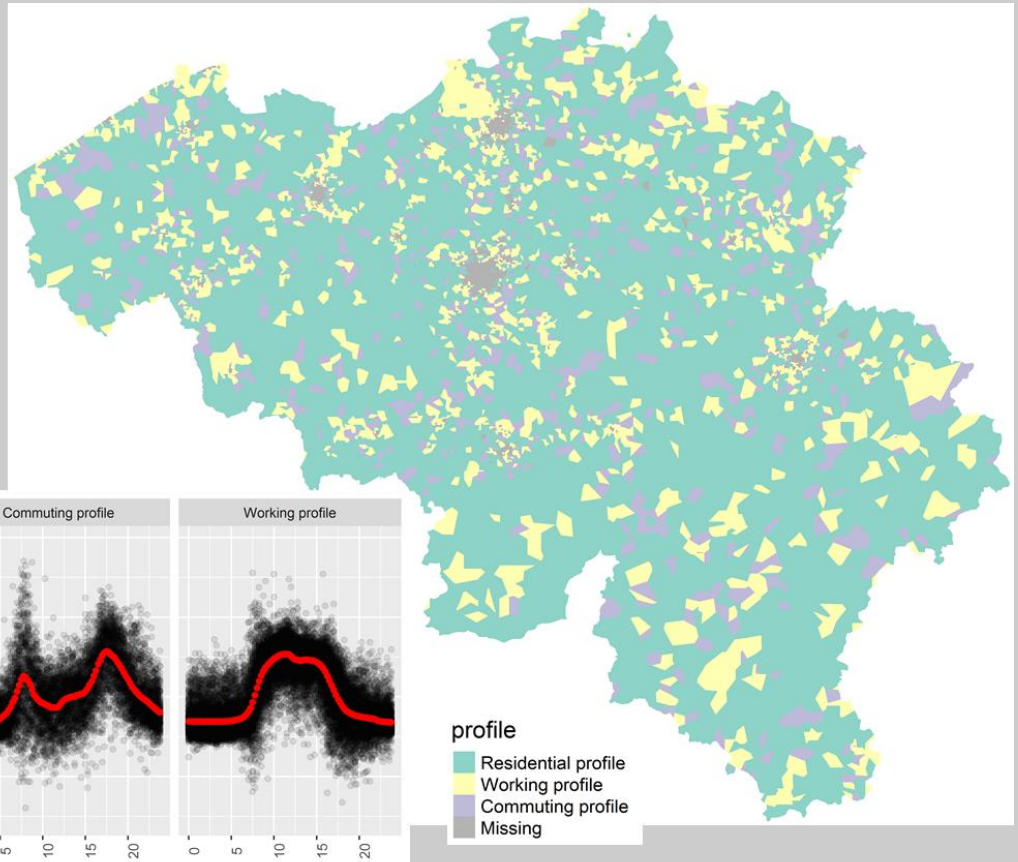
Deville, Pierre, et al. "Dynamic population mapping using mobile phone data." Proceedings of the National Academy of Sciences 111.45 (2014): 15888-15893.

Daytime population in Belgium

Down left: Standardized number of mobile phones at different times for three profiles "living", "commuting" and "work"

Right: Map with radio cells in Belgium assigned to the three profiles





Land use classification based on present population daily profiles from a big data source
Fernando Reis et al (2016)
http://nt17.pg2.at/data/x_abstracts/x_abstract_172.docx

Other projects

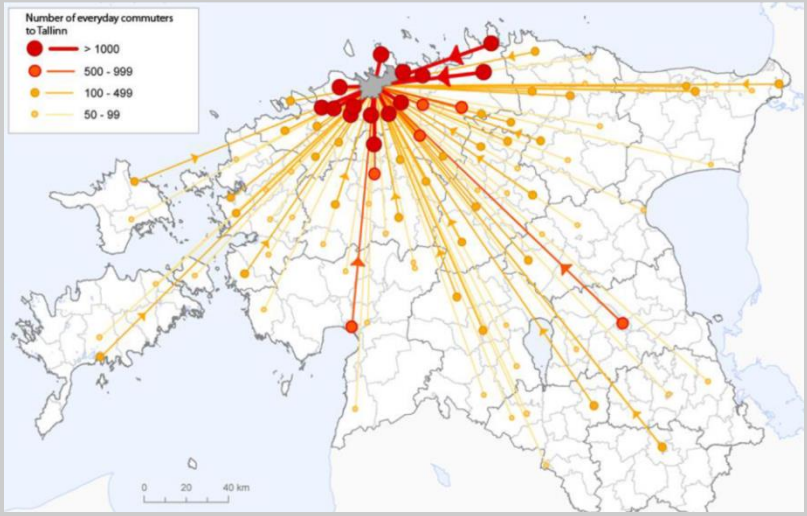
Tourism (Statistics Spain)

- Number of regional trips to destination (inhabitants)
- Number of night-stays (inhabitants and foreign visitors)

Quality implications of the use of big data in tourism statistics: three exploratory examples
F. C. García et al (2016), http://www.ine.es/q2016/docs/q2016Final00015.pdf

Commuter (Statistics Estonia)

- Hourly start and destination matrices
- Border-crossing
- Travel time and distance



Quelle: Positium (2014), http://unstats.un.org/unsd/trade/events/2014/Beijing/Margus%20Tiru%20-%20Mobile%20Positioning%20Data%20Paper.pdf

Mobile phone data

Cell Identification (CID)

CID is a generally unique number used to identify each base transceiver station (BTS) or sector of a BTS within a location area code (LAC)

Call detail record (CDR)

CDR is a data record produced by a telephone exchange or other telecommunications equipment that documents the details of a telephone call or other telecommunications transaction (e.g. text message)

Mobile phone data

International Mobile Subscriber Identity (IMSI)

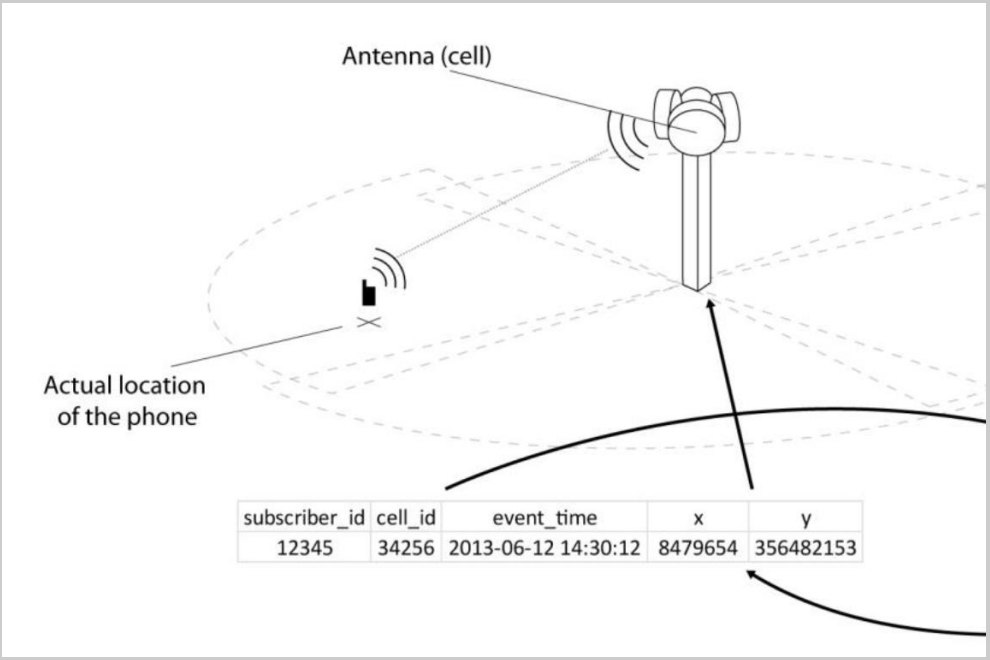
IMSI is used to identify the user of a cellular network and is a unique identification associated with all cellular networks.

International Mobile Equipment Identity (IMEI)

IMEI is a number, usually unique to identify 3GPP (i.e., GSM, UMTS and LTE) and iDEN mobile phones, as well as some satellite phones.

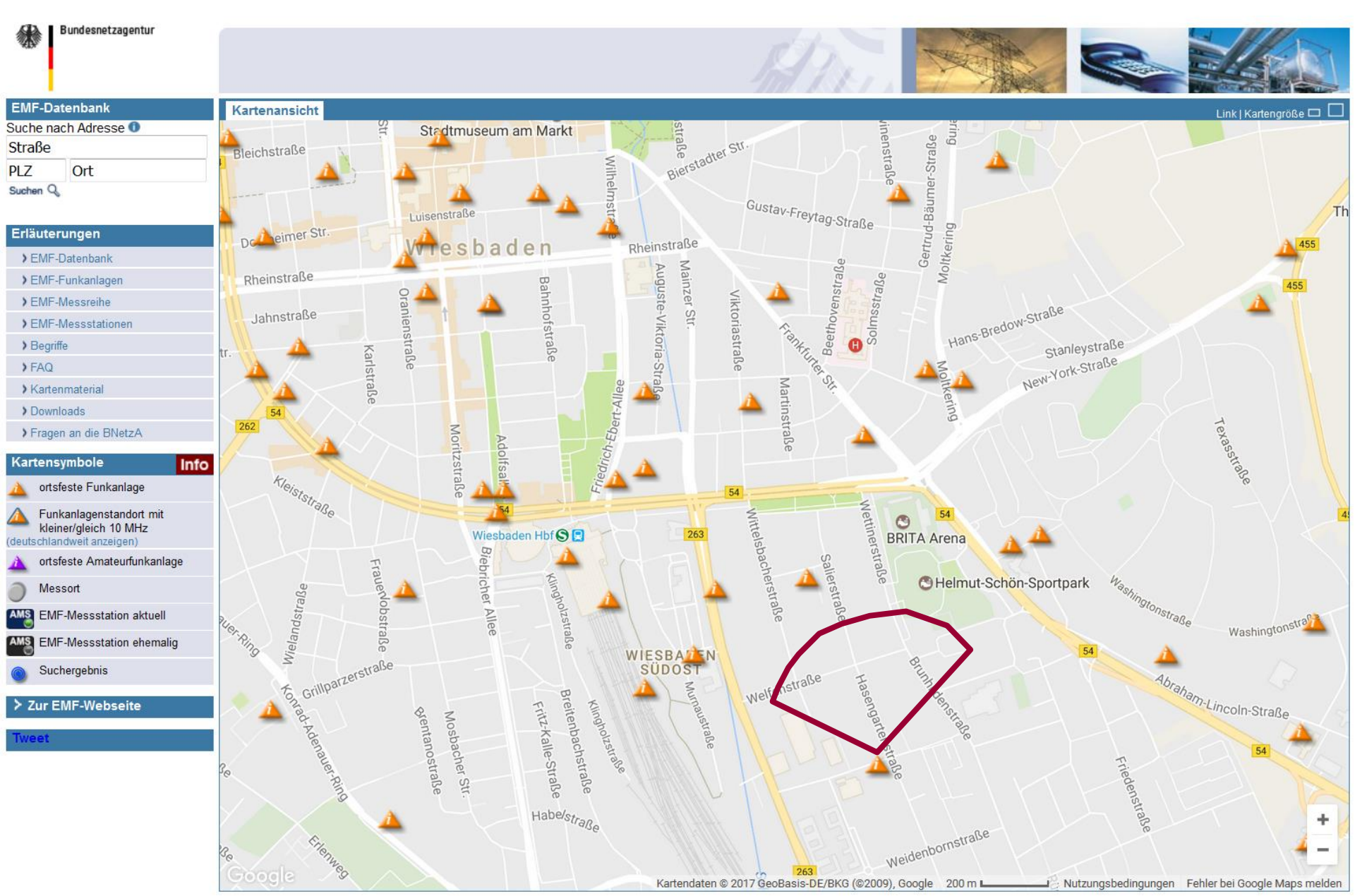
Mobile phone Data

- Anonymous analysis with aggregated location data



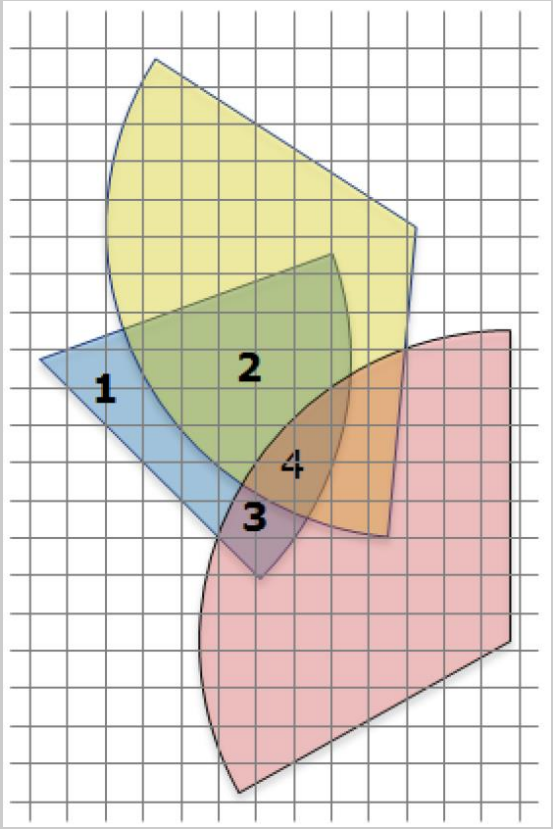
Source: Eurostat (2014), http://epp.eurostat.ec.europa.eu/portal/page/portal/tourism/methodology/projects_and_studies

Cell-ID	Date	Time	Number of mobile devices
12345	2017-01-10	04:00	XX
“	“	…	…
“	“	10:00	XXX



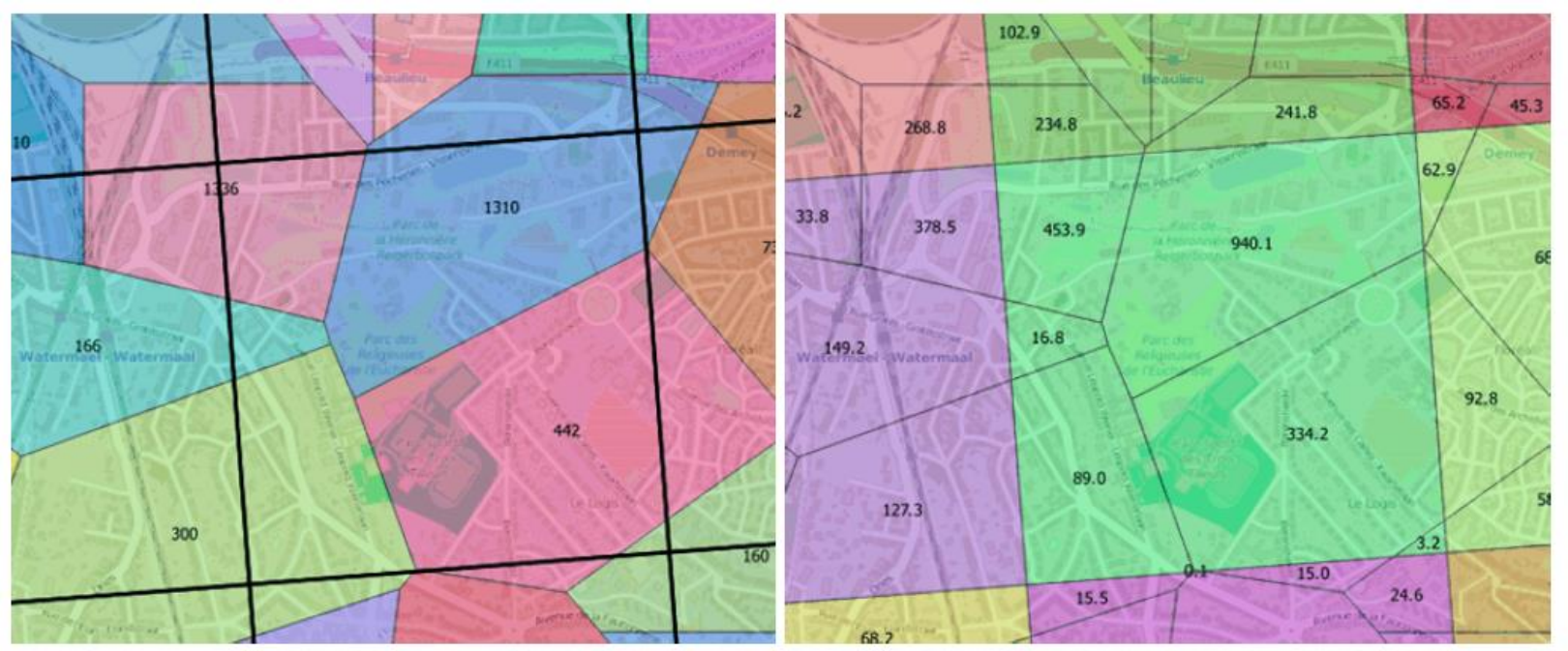
Quelle: http://emf2.bundesnetzagentur.de/karte/?lat=50.11201830731937&lon=8.679846801757778&zoom=15

Where is the mobil phone?



Scholtus, S. (2015) Aantekeningen over het toewijzingsalgoritme voor Daytime Population. Internal CBS note (in Dutch), Statistics Netherlands

From the cell to a grid



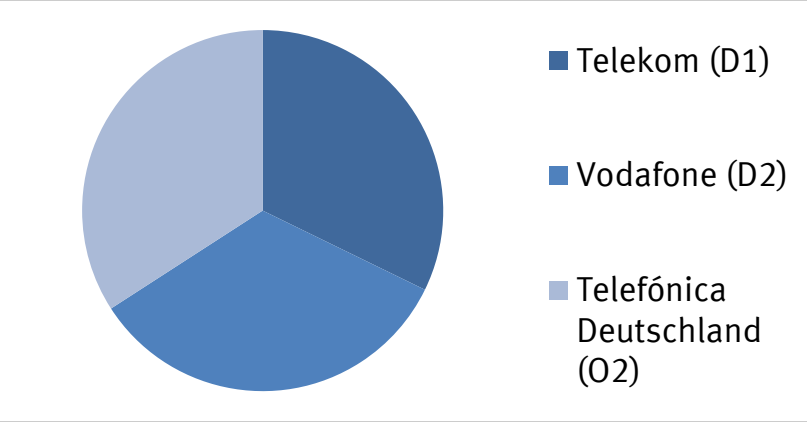
Assessing the Quality of Mobile Phone Data as a Source of Statistics,
Freddy De Meersman et al (2016)
https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP5_Documentation

Market share in Germany

Network operator	Market share	Number of users
Telekom (D1)	32%	41.849.000
Vodafone (D2)	34%	43.700.000
Telefónica Deutschland (O2)	34%	44.321.000
Total		129.870.000

Shares of network operators in the fourth quarter of 2016

Source: Bundesnetzagentur, https://www.bundesnetzagentur.de/DE/Sachgebiete/Telekommunikation/Unternehmen_Institutionen/Marktbeobachtung/Deutschland/Mobilfunkteilnehmer/Mobilfunkteilnehmer_node.html

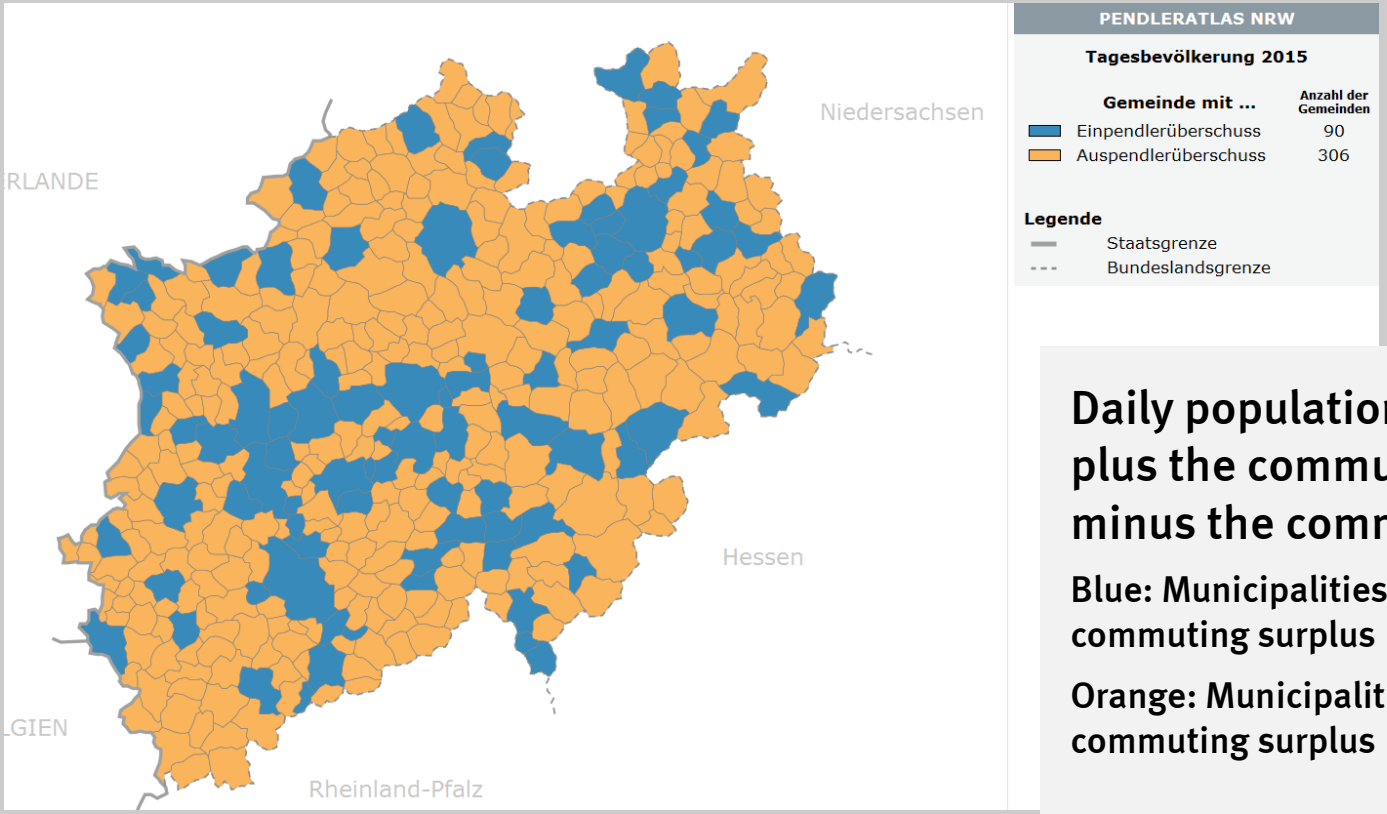


Analysis of Mobile phone data

- Negotiations with Telekom and Vodafone
- Feasibility studies:

  Daytime population and commuter flow on regional level compared with mobile phone data
- Planned cooperation with IT.NRW and Statistics Berlin/Brandenburg

Pendleratlas IT.NRW



Daily population 2015: Population plus the commuters (incoming) minus the commuters (outgoing)

Blue: Municipalities with incoming commuting surplus

Orange: Municipalities with outgoing commuting surplus

Quelle: https://www.pendleratlas.nrw.de

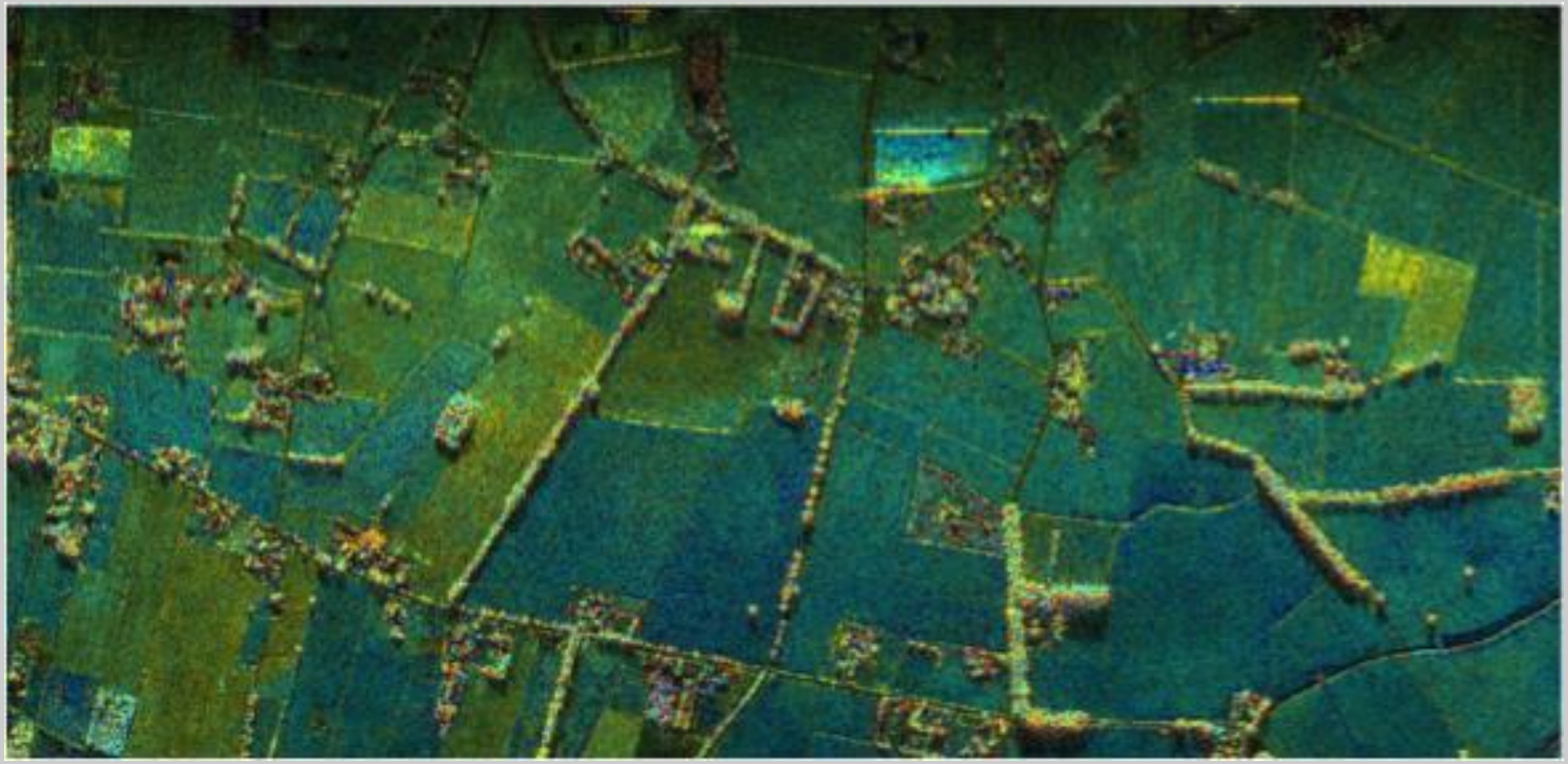
SATELLITE DATA

Quelle: http://www.esa.int/spaceinimages/Images/2016/12/Agricultural_monitoring_in_Spain

Leaf Area Index

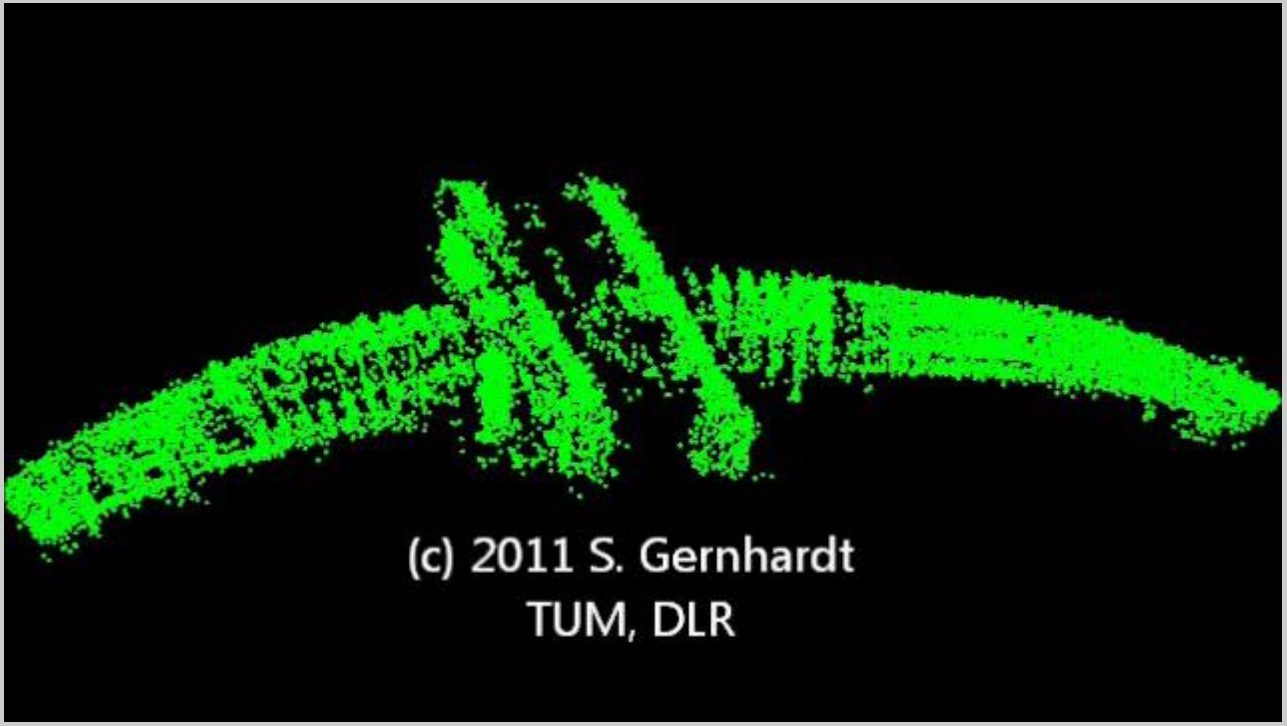
Leaf area index (LAI) is a dimensionless quantity that characterizes plant canopies. It is defined as the one-sided green leaf area per unit ground surface area (*LAI = leaf area / ground area, m^2 / m^2*) in broadleaf canopies.
Source: http://www.esa.int/spaceinimages/Images/2016/05/Czech_Republic_Leaf_Area_Index

Elevation Data



Quelle: http://www.esa.int/spaceinimages/Images/2017/02/Fields_in_3D

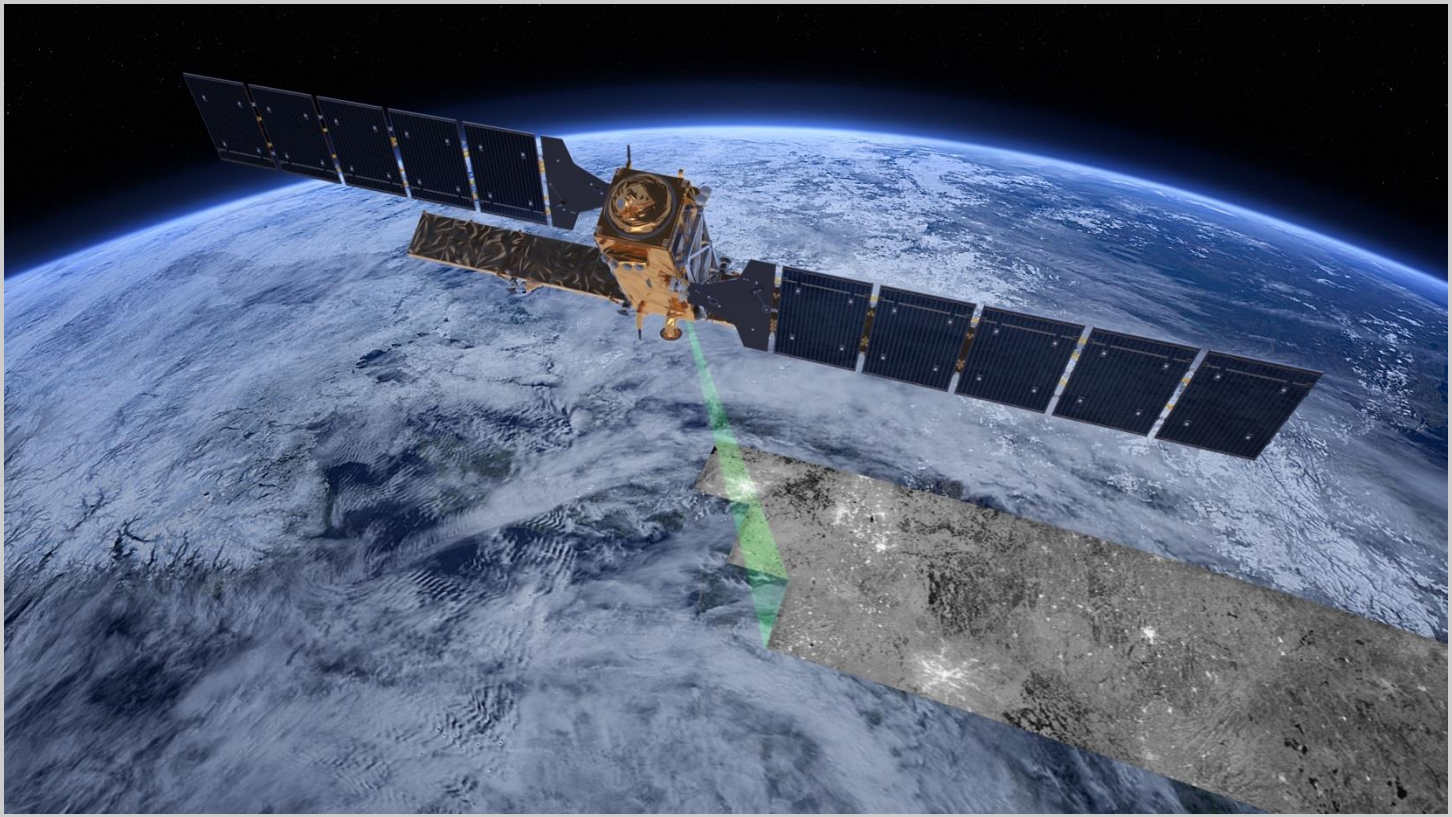
Radar data



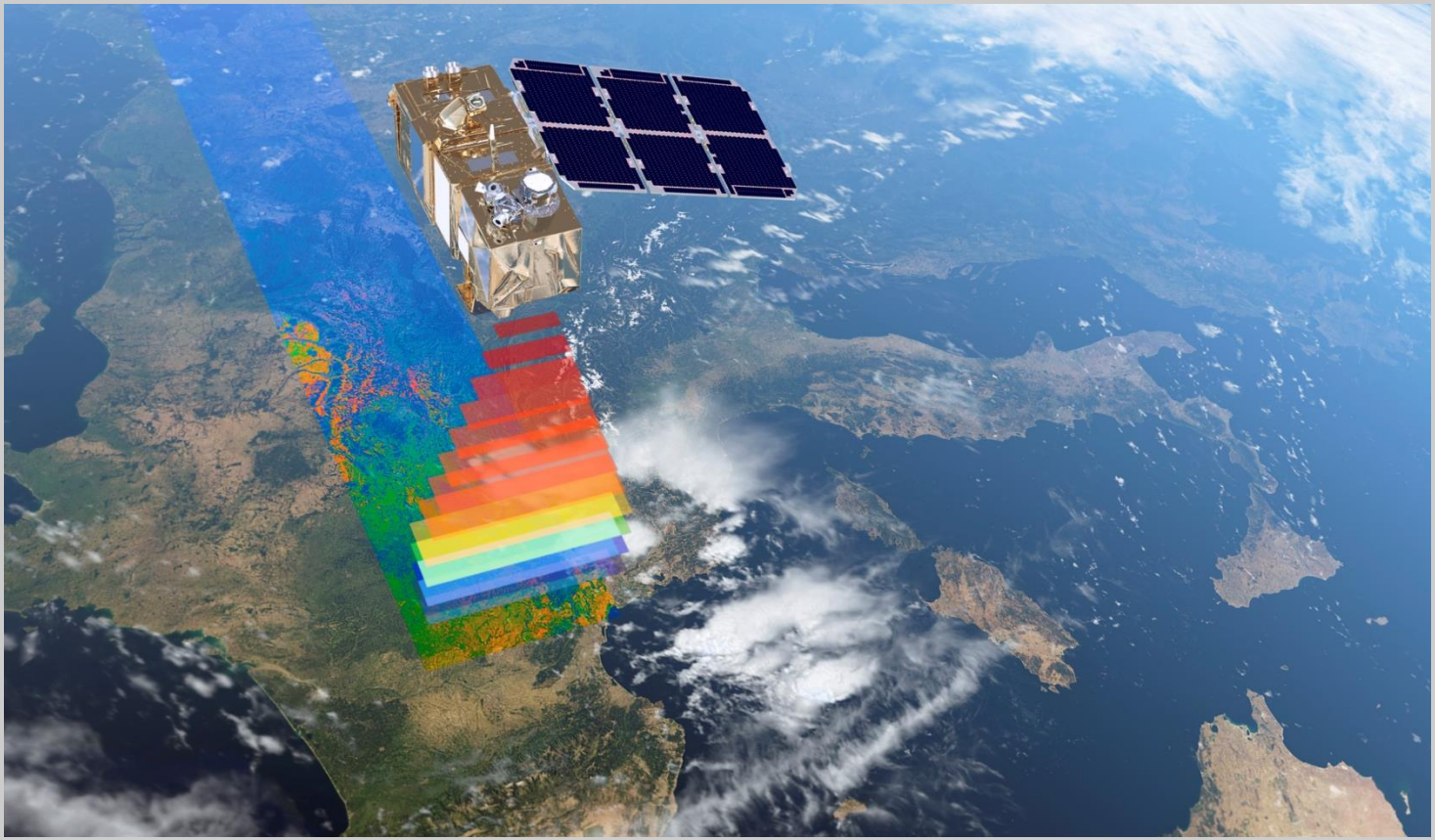
Copernicus programme

- European Earth Observation Program
- Measurements of satellites (sentinel missions), aircraft, ground or sea-based observation infrastructures
- Various Copernicus services for monitoring land, sea environment, atmosphere, etc.
  - e.g. High-resolution images for land cover
  - Data on land and ocean surfaces, regardless of lighting and weather (imaging radar data)
- Data are freely available

Sentinel-1: Radar Data



Sentinel-2



http://www.esa.int/esatv/Videos/2016/02/Sentinels_for_Copernicus

COP4STAT_2015plus

- Improvement of land use statistics (ESS land use and land cover survey, LUCAS)
- Further development of agriculture and farming statistics
- Cooperation with the Federal Agency for Cartography and Geodesy (BKG)
- Evaluation of various Copernicus products: optical image data of the Sentinel-2 sensor and other data (e.g. radar or elevation data)

MAKSWELL

MAKing Sustainable development and WELL-being frameworks work for policy analysis

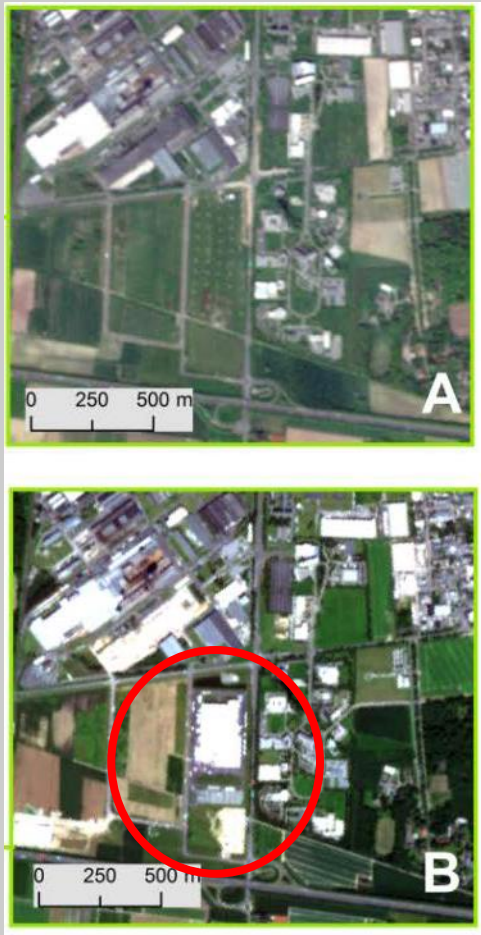
- MAKSWELL proposes to extend and harmonize indicators able to capture the main characteristics of the beyond-GDP approach
- MAKSWELL should improve i.a. the database both in relation to the timeliness, the integration with big data measures and the methodologies
- Partners: ISTAT, CBS, HCSO, Destatis and the universities from Pisa, Trier and Southampton

MONO

- Call: Merging Statistics and geospatial information
- Monitoring spatial sustainable development; (Semi-) automated analysis of satellite and aerial images for energy transition and sustainability indicators
- In cooperation with Statistics Netherlands, Statistics Belgium and IT.NRW
- Eurostat Grant: Project application submitted in May 2017

GebäuDE-21 (Building-21)

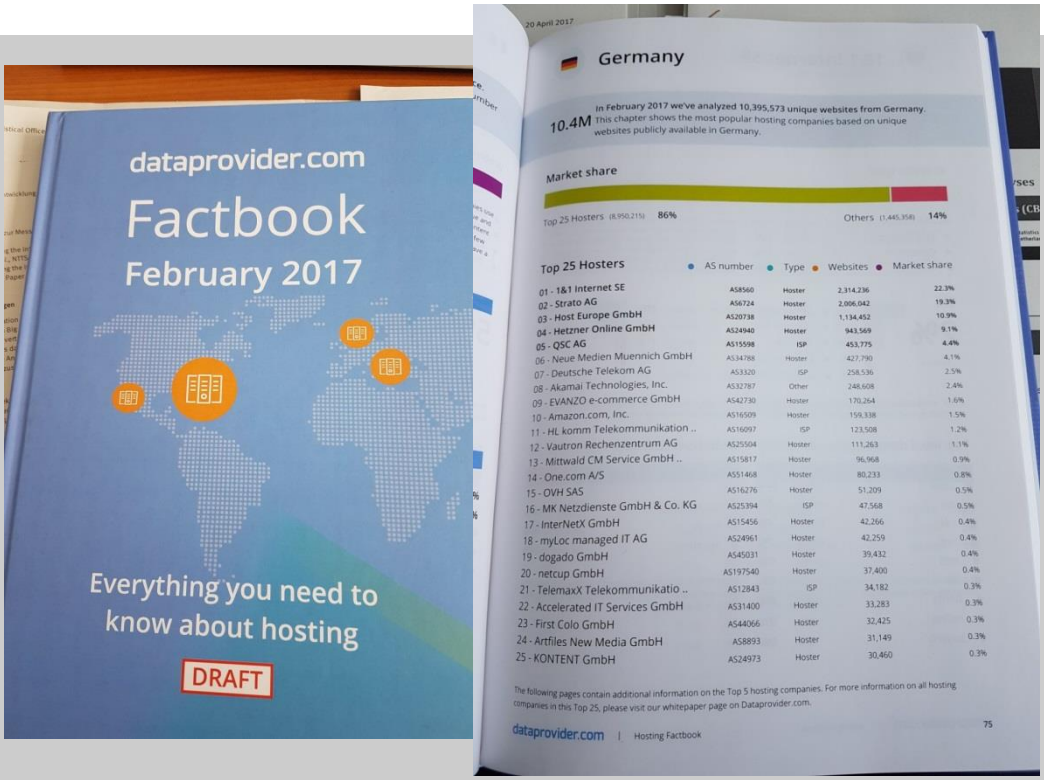
- Cooperation with DLR and the Leibniz Institute for Ecological Spatial Development
- Update of the address and building register for the time after Census 2021
- Identification of potential residential buildings and elimination of buildings without housing
- Change Analysis: Identification of newly built and demolished buildings



Source: Feasibility study for GebäuDE-21

DESTATIS
wissen.nutzen.

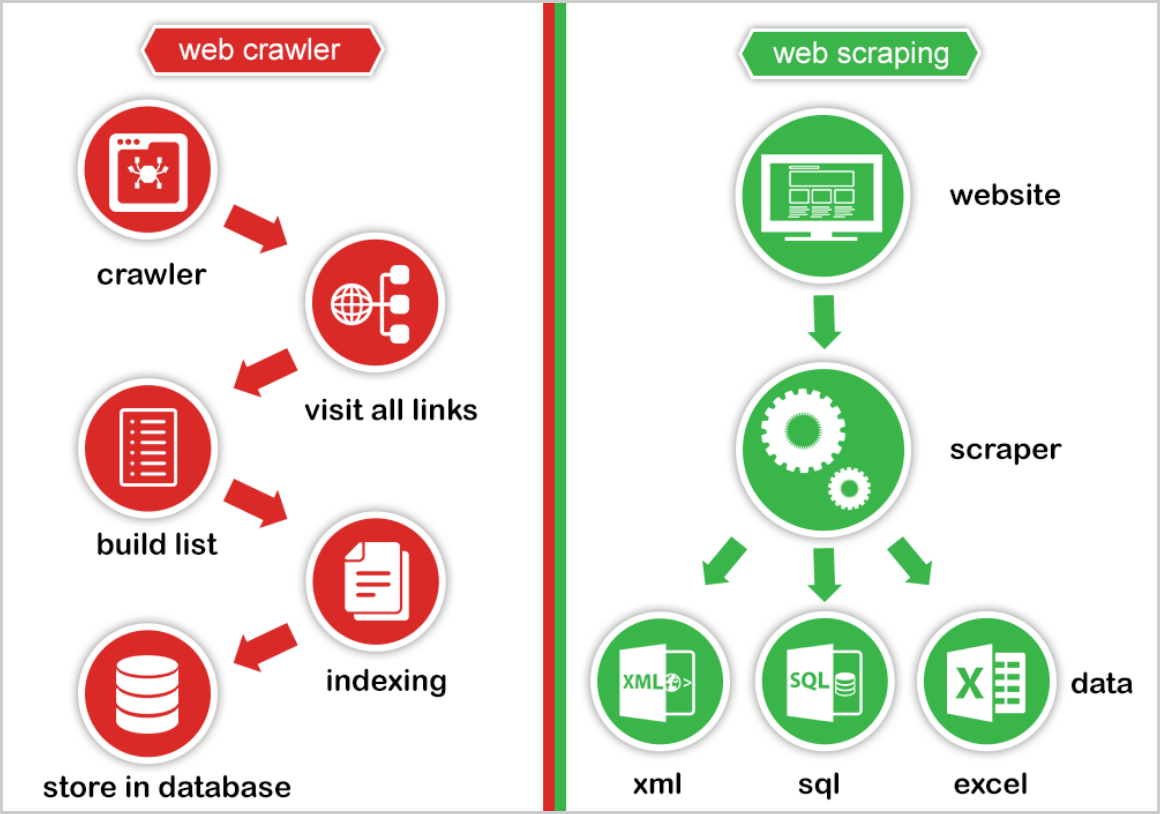
ENTERPRISE DATA FROM THE INTERNET



Dataprovider: We index the web and structure the data

Dataprovider transforms the internet into a structured database to help you gain insights about companies. We index more than 280 million domains each month and can tell you all about those (www.dataprovider.com).

Excursus: Web scraping vs Web crawling



DESTATIS
wissen.nutzen.

Measuring the internet economy in The Netherlands: a big data analysis

Lotte Oostrom, Adam N. Walker, Bart Staats, Magda Slootbeek-Van Laar, Shirley Ortega Azurduy, Bastiaan Rooijakkers

- https://www.cbs.nl/nl-nl/achtergrond/2016/41/measuring-the-internet-economy-in-the-netherlands
- https://www.conference-service.com/NTTS2017/documents/agenda/data/abstracts/abstract_122.html

Statistics Netherlands

Aim of the study

Main research question:

"What is the importance of the internet economy to the Dutch economy?"

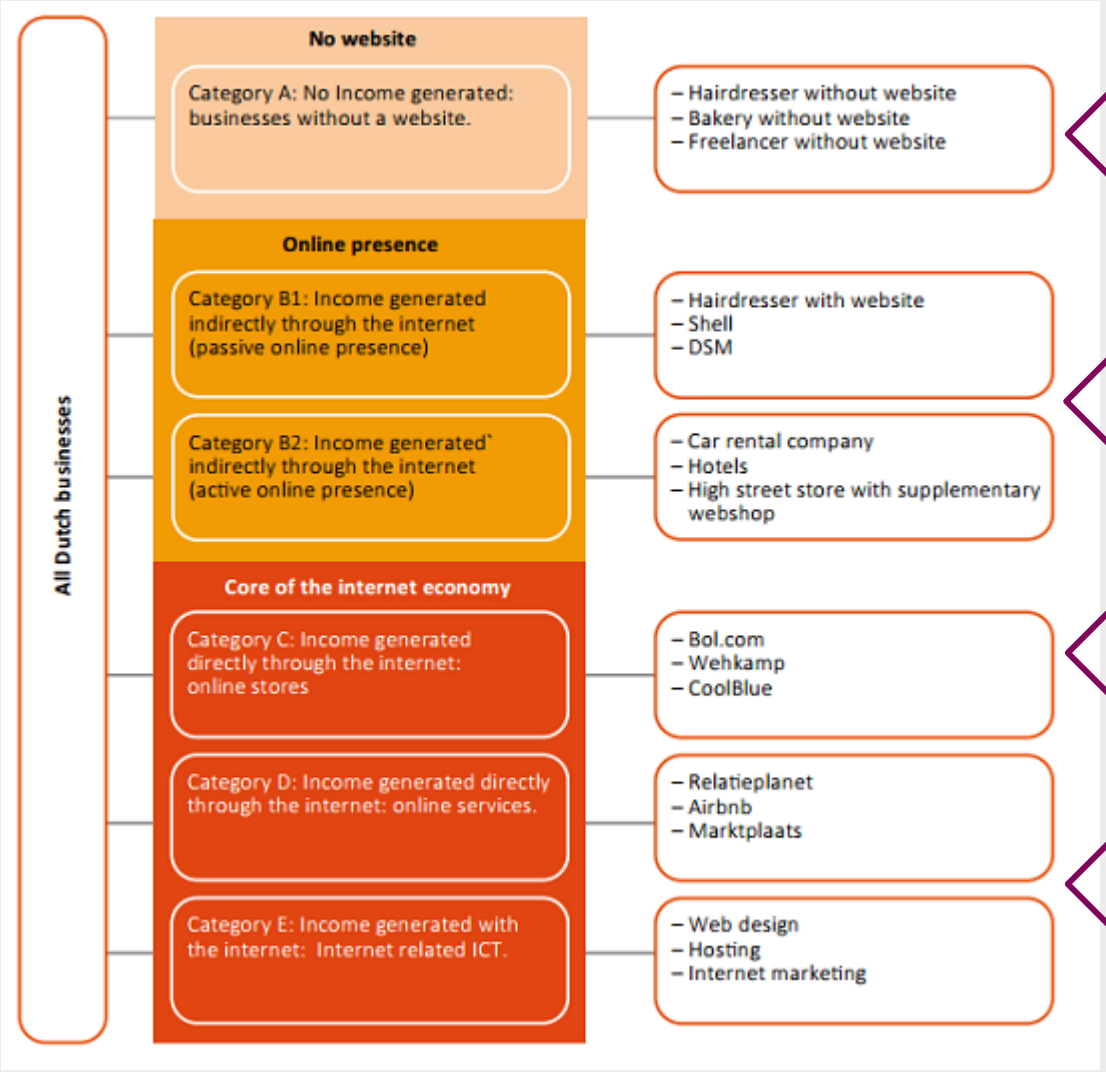
The aim of the research project was fourfold:

1. Determine a pragmatic definition of "the internet economy"
2. Show the importance and size of the internet economy in NL
3. Show the possibilities of new measurement methods with big data
4. Explain differences from regular statistics/concepts

Dataprovider dataset: 2,5 million Dutch websites

Business information	• Country, address, company name, Chamber of Commerce number, taks number, phone number, e-mail, ...
eCommerce	• eCommerce probability, shopping cart software, delivery services, payment methods, products, prices,...
Content	• Title, description, keywords, category, language, author....
Other	• Marketing, social media, links, technical and hosting information, ...

Definition of the internet economy



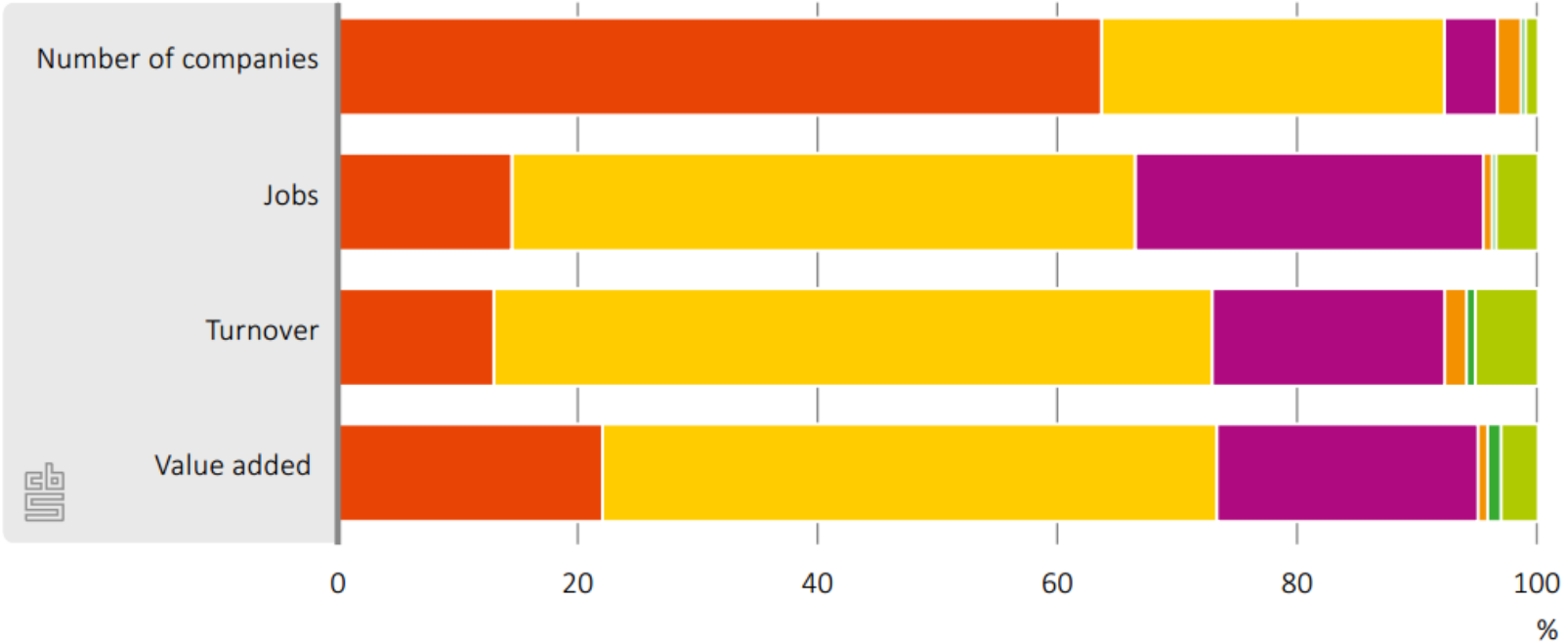
Businesses without a website

Businesses with a website that do not belong to category C, D or E

Online stores according to eCommerce variables

Online services or internet related ICT according to keyword analysis

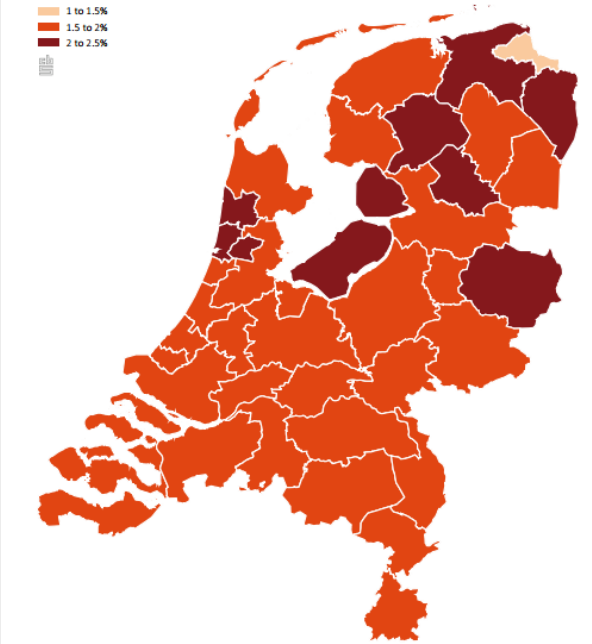
Relative distribution of number of companies, jobs, turnover and value added by Internet categories, 2015



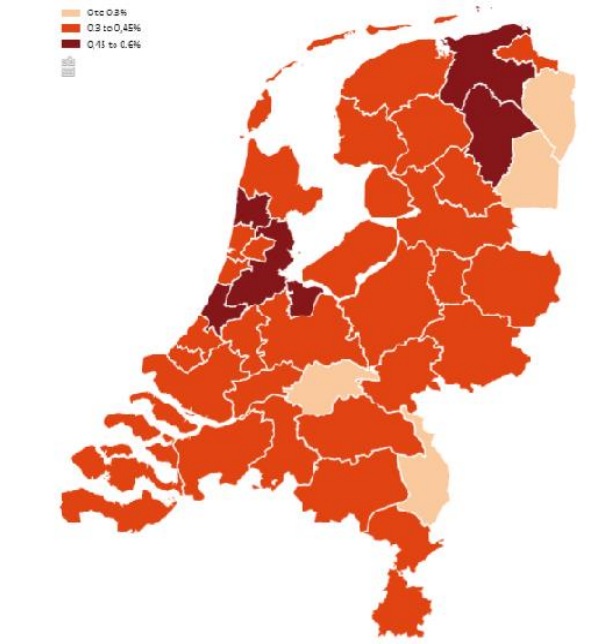
The core of the internet economy (online stores, online services and internet related ICT) consists of 50,000 businesses and provides 345,000 jobs (4.4% of the total) and a turnover of € 104 billion (7.7% of the total).

Regional distribution

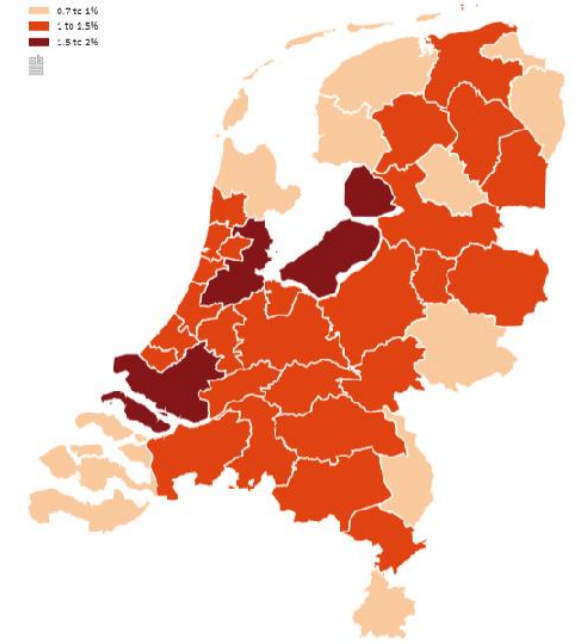
Certain regions are more prominent in the internet economy than others.



Online stores



Online services



Internet related ICT

Project ideas for Germany

- Internet economy in the Federal State Hessen
- Further development of the German business register
- Using of information and communication technology in German companies
- KombiFid II
  Combination of firm data coming from different data producer
- Big Data Hackathon as internal training

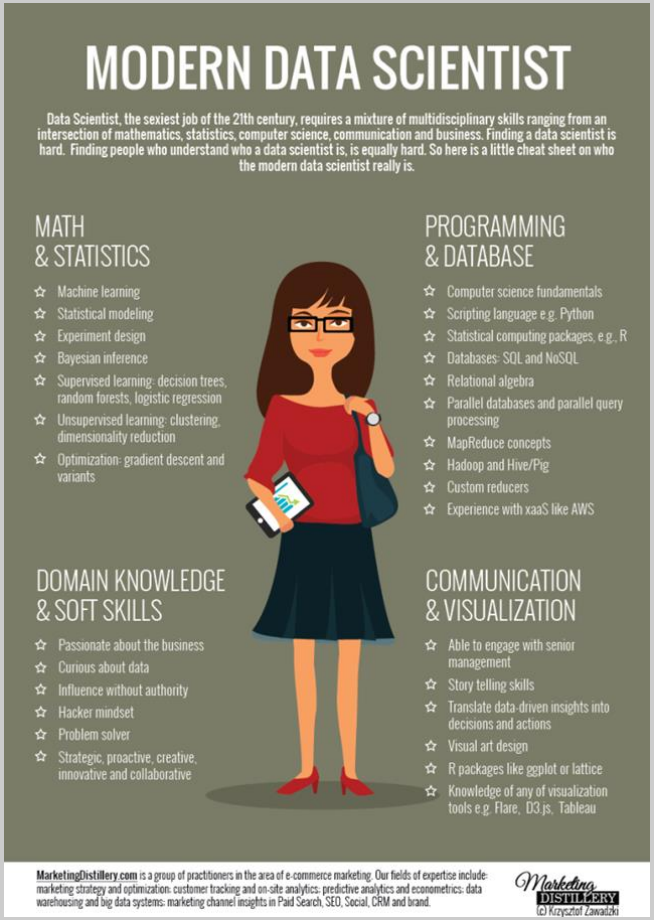
Automated Price Collection

- Reproduction of manual price collection on the Internet
- Increasing the number of collected prices
- Examples: flights, hotels, online retailers, online pharmacies, car rentals, train tariffs, coach trips, city trips, package holidays

Further plans

- Further development of automated collection
- Development of a work environment for the permanent use of automated price surveys

Statistical Education in times of Big Data

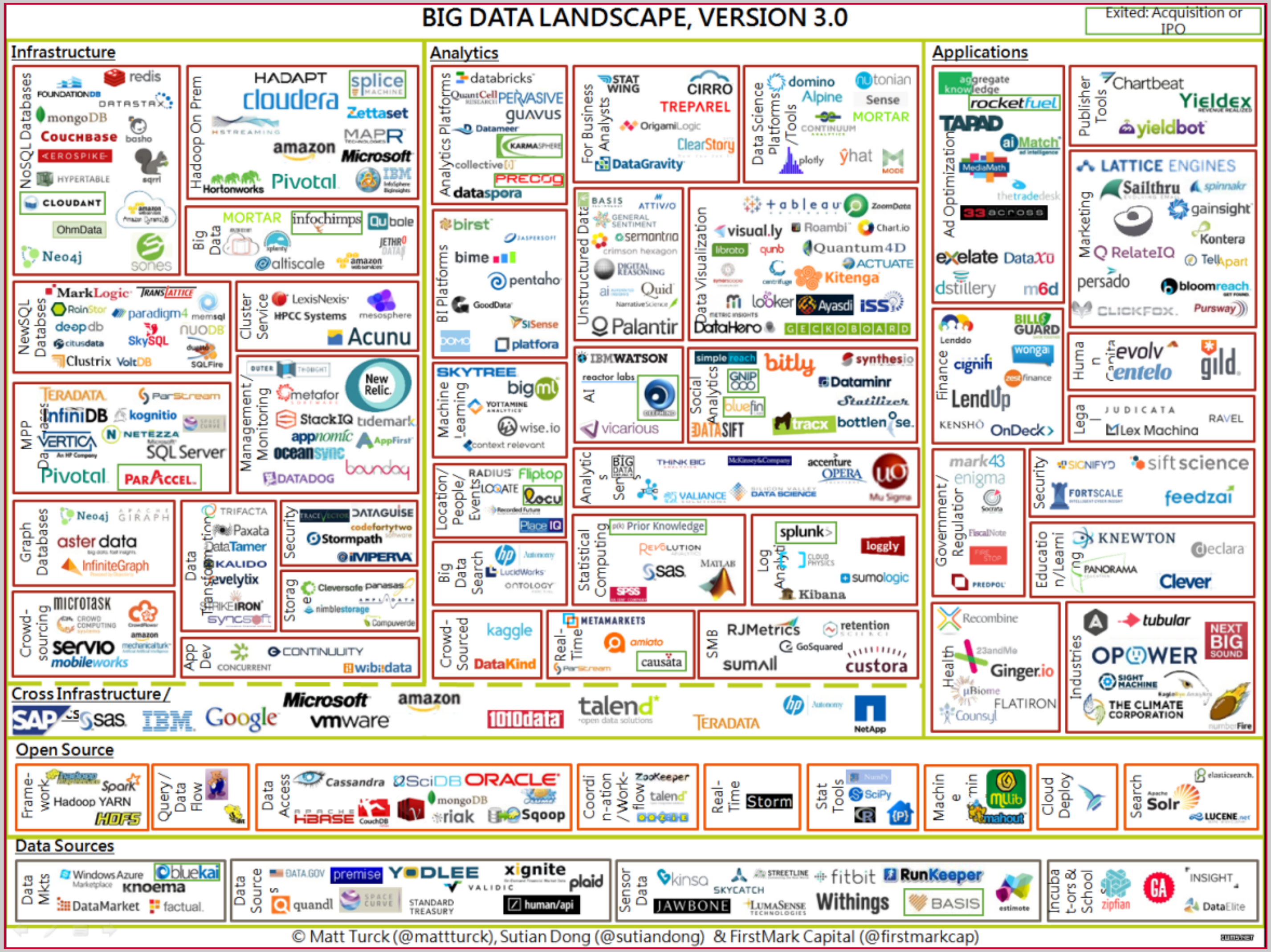


65th plenary session of the Conference of European Statisticians, Geneva, 19 - 21 June 2017

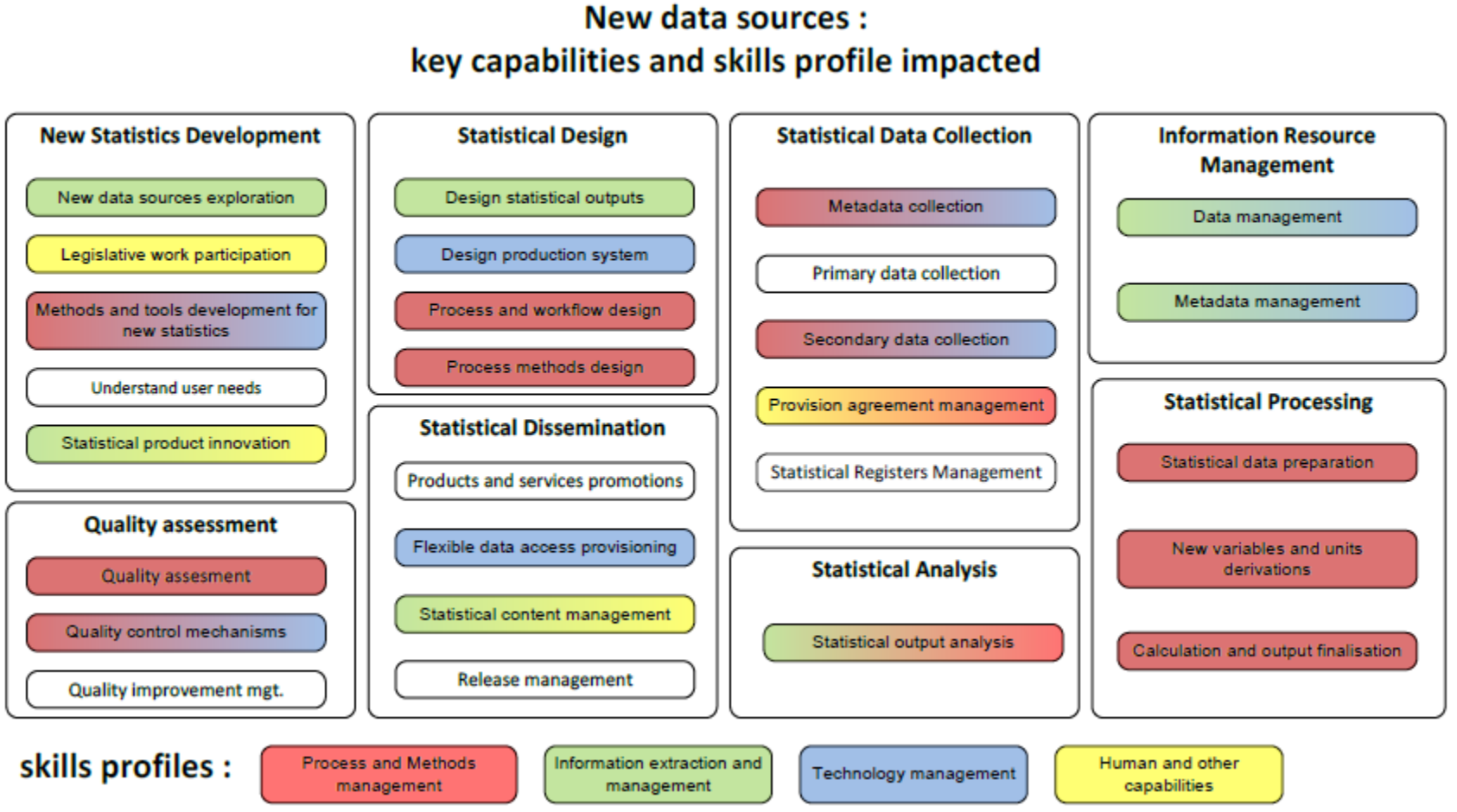
Seminar on the next generation of statisticians and data scientists

Session I	**How can an official statistician become a data scientist?**
Session II	**Are universities training students to work in statistical offices?**
Session III	**The future role of statisticians**

http://www.unece.org/stats/documents/2017.06.ces.html#/

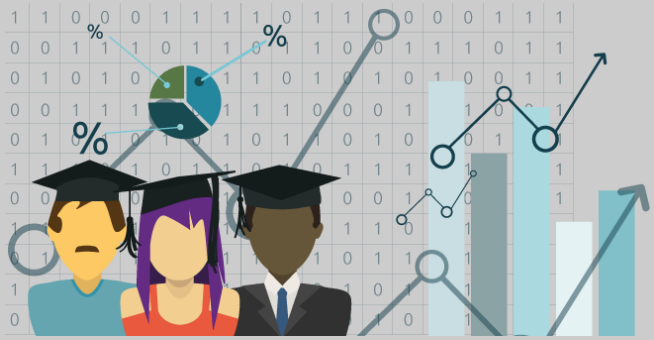

BIG DATA LANDSCAPE, VERSION 3.0
Exited: Acquisition or IPO
Infrastructure
NoSQL Databases
Hadoop On Prem
Big Data
NewSQL Databases
Cluster Service
MPP Databases
Management/ Monitoring
Graph Databases
Data Transformation
Security
Crowd-sourcing
Storage
App Dev
Analytics
Analytics Platforms
For Business Analysts
Data Science Platforms /Tools
BI Platforms
Unstructured Data
Data Visualization
Machine Learning
AI
Social Analytics
Location/ People/ Events
Analytics Services
Big Data Search
Statistical Computing
Log Analytics
Crowd-Sourced
Real-Time
SMB
Applications
Ad Optimization
Publisher Tools
Marketing
Finance
Human Capital
Legal
Government/ Regulation
Security
Education/Learning
Health
Industries
Cross Infrastructure/ Analytics
Open Source
Framework
Query/ Data Flow
Data Access
Coordination /Workflow
Real-Time
Stat Tools
Machine Learning
Cloud Deploy
Search
Data Sources
Data Mkts
Data Sources
Sensor Data
Incubators & Schools
© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)

New data sources – key capabilities and skills profile influenced



https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/2017/CES_23_E_Next_generation_Finland___Eurostat.pdf

NSIs and universities



- Educating the next generation of statisticians
- Further developing the curriculum by including more aspects of new digital sources also in introductory courses
- Improving the statistical literacy for the data user side

EMOS - European Master of Official Statistics

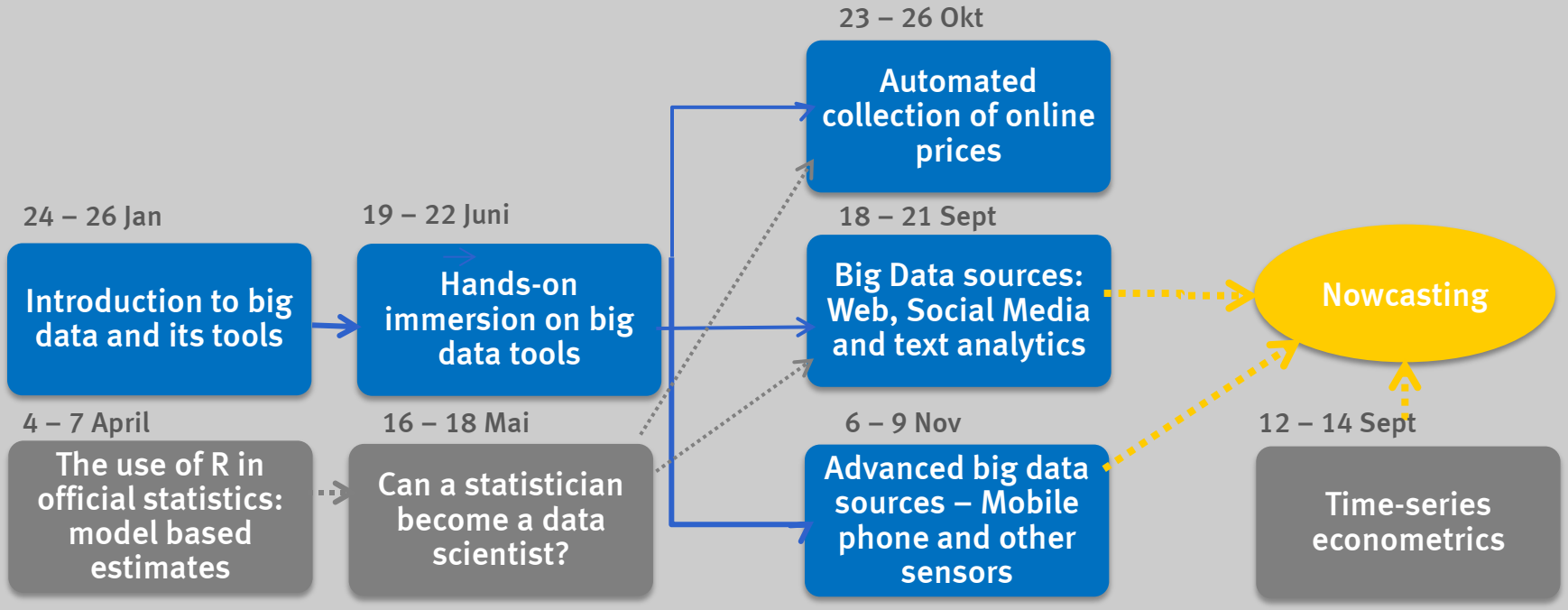
- Label for the Master study
- Network of master studies with the main focus on Official Statistics and data production at European level
- The aim: To strengthen cooperation between universities and producers of Official Statistics and to train professionals
- A practical mix of competences and knowledge as well as suggested topics for the master thesis, internships, EMOS workshops and webinars
- A way for an ongoing training inside the NSIs

UNECE Sandbox

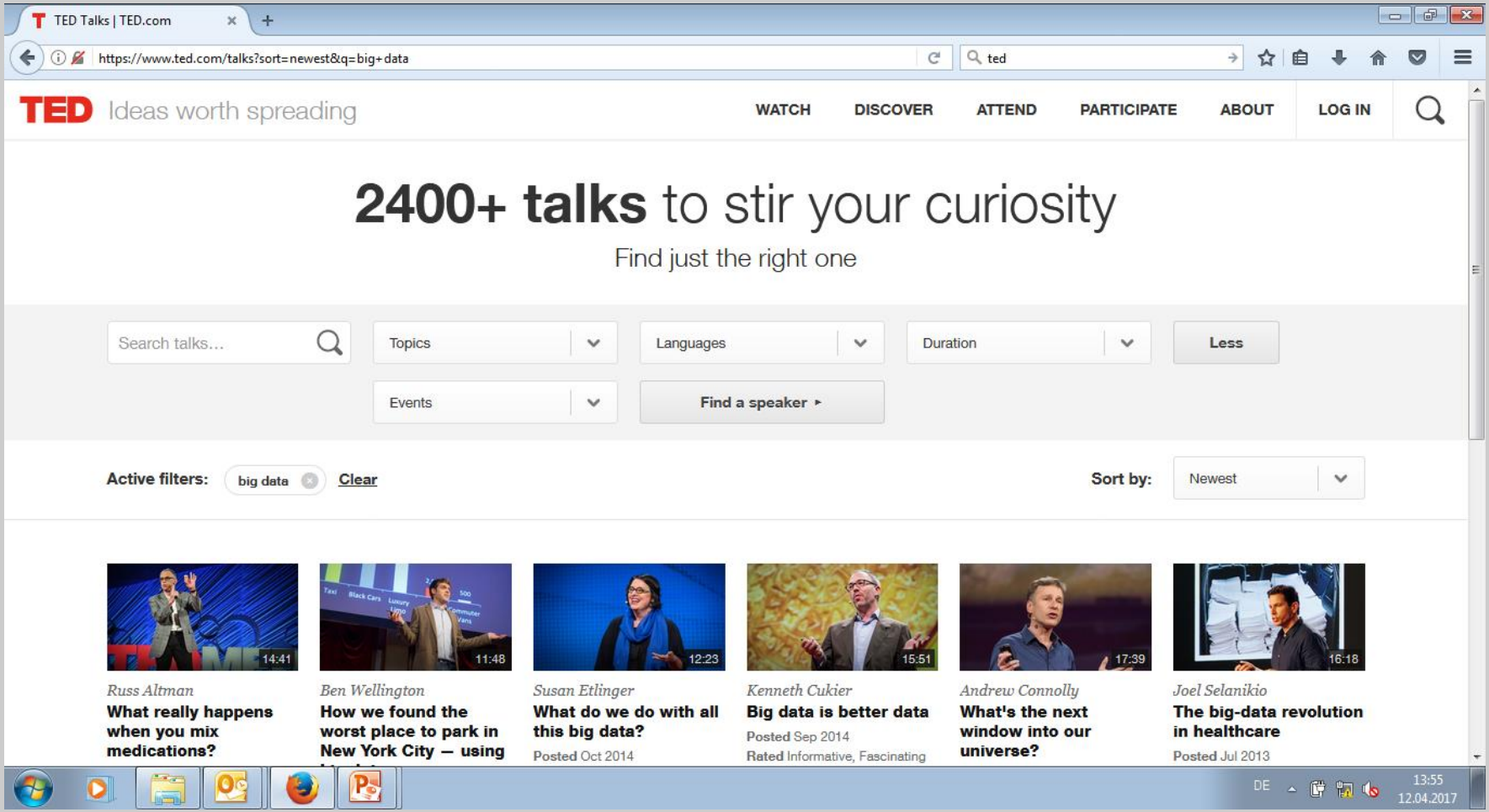
- Set up by the Irish Centre for High-end Computing and the Irish Central Statistics Office
- Contains data sets and tools for international experiments
- Remote access and processing of data
- Experiments with data from social media, mobile phones, smart meters, traffic loops

http://www1.unece.org/stat/platform/display/bigdata/Sandbox

European Statistical Training Programme ESTP - Big Data 2017



http://ec.europa.eu/eurostat/web/european-statistical-system/training-programme-estp



THANK YOU FOR THE ATTENTION!

Markus Zwick

www.destatis.de