Introduction to Big Data in Official Statistics

Prof. Dr. Markus Zwick
Federal Statistical Office Germany,
Institute for Research and Development in Official Statistics

© Federal Statistical Office Germany,
Institute for Research and Development in Official Statistics
There was something different in the Vatican crowd in 2005...
Population Mapping Using Mobile Phone Data


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

© Federal Statistical Office Germany, Institute for Research and Development in Official Statistics
Population Mapping Using Mobile Phone Data
Population Mapping Using Mobile Phone Data
Population Mapping Using Mobile Phone Data
The graph shows population densities derived from mobile phone counts at 4 am on Thursday 8 October (left) and the 2011 population census (right). The Pearson correlation between these two datasets is 0.85, a clear indication that mobile phone data are able to provide a valid and accurate measure of population density.

Google Skybox/Terra Bella

Quelle: https://terrabella.google.com

© Federal Statistical Office Germany,
Institute for Research and Development in Official Statistics
Google Skybox/Terra Bella

Quelle: https://terrabella.google.com

© Federal Statistical Office Germany,
Institute for Research and Development in Official Statistics
Google Skybox/Terra Bella

Quelle: http://blogs.wsj.com/digits/2014/06/10/google-getting-more-than-pretty-pictures-with-skybox
In 2013, four zettabytes of data were created by digital devices. In 2017, it is expected that the number of connected devices will reach three times the number of people on earth.

What do you think are the main characteristics of Big Data?
Features Big Data ("3 Vs")

- **Volume** – Large amount of Data

- **Velocity** - Speed in which new data:
  - Arise
  - Are available
  - Can be processed
  - Can change fundamentally

- **Variety** - Unstructured / heterogeneous data
Big Data Is More Than 3 Vs*

*2001 (Meta) / 2012 (Gartner) Definition of Big Data

**Volume**
- IDC Report 2011
- 8 billion TB in 2015
- 40 billion TB in 2020
- 90% of all data < 2 years
- storage • transport processing

**Variety**
- relational, graph time series, sensor, audio, video, text, geo, scientific, ...
- 80% unstructured

**Velocity**
- facebook 500 TB/day
- Large Hadron 35 GB/sec
twitter 300K tweets/min
- real time • stream


© Federal Statistical Office Germany,
Institute for Research and Development in Official Statistics
'Five Vs' of Big Data

- 'Volume', 'Velocity' and 'Variety' are the 'essential' characteristics of big data
- 'Veracity' and 'Value' are the 'qualification for use' characteristics of big data

nach Diego Kuonen, Universität Genua

© Federal Statistical Office Germany,
Institute for Research and Development in Official Statistics
The data deluge

- **Communication**
  - Mobile phone data
  - Social Media

- **WWW**
  - Web Searches
  - Businesses' Websites
  - E-commerce websites
  - Job advertisements
  - Real estate websites

- **Sensors**
  - Traffic loops
  - Smart meters
  - Vessel Identification
  - Satellite Images

- **Process generated data**
  - Flight Booking transactions
  - Supermarket Cashier Data
  - Financial transactions

- **Crowd sourcing**
  - VGI websites (OpenStreetMap)
  - Community pictures collection

© Federal Statistical Office Germany, Institute for Research and Development in Official Statistics
UNECE - Classification of Types of Big Data

1. Social Networks (human-sourced information)
   - Social Networks: Facebook, Twitter, Tumblr etc.
   - Blogs and comments
   - Pictures: Instagram, Flickr, etc.
   - Videos: Youtube etc.
   - Internet searches
   - Mobile data content: text messages
   - User-generated maps

2. Traditional Business systems (process-mediated data)
   - Data produced by Public Agencies
   - Medical records
   - Data produced by businesses
   - Commercial transactions
   - Banking/stock records
   - E-commerce
   - Credit cards

3. Internet of Things (machine-generated data)
   - Data from sensors
   - Home automation
   - Weather sensors
   - Traffic sensors
   - Mobile phone location
   - Cars
   - Satellite Images

http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data

© Federal Statistical Office Germany, Institute for Research and Development in Official Statistics
What do you think about the influence of Big Data on Official Statistics?

- What are the benefits?
- What are the efforts and difficulties?
Official Statistics

- Professional Independence
- Mandate for Data Collection
- Commitment to Quality
- Impartiality and Objectivity
- Sound Methodology
- Non-excessive Burden on respondents
- Timeliness and Punctuality

European Statistics Code of Practice
and Official Statistics are public goods

Should Official Statistics go into a competition with new data producers like Google?
Possible benefits of Big Data for Official Statistics

- Faster results
- Lower cost
- Higher precision
  - For small groups like a freelancer at the country side
  - For small areas like the next street behind the corner
- Completeness
- Less burden for the respondents
Challenges of Big Data for Official Statistics

- Quality issues
- Privacy and legal constraints
- Permanent access to the data
- Competition at the information market
- Competition for the best brains
What does big data mean for Official Statistics?

- From finite population sampling methodology to additional statistical modelling and machine learning
- From designers of data collection processes to designers of statistical products
- Accreditation and certification may become core tasks of NSIs
Dimension of data

1 to 1 relation

Survey → Production → Statistical results

m to k relation

Multiple source → Datawarehouse → Population statistics → Enterprise statistics → National Accounts

© Federal Statistical Office Germany, Institute for Research and Development in Official Statistics
Data warehouse for Official Statistics

Data for Official Statistics:

- Generated by digital and survey data sources
- Often not generated primarily for the statistical purposes
- Statistical modelling will be a main activity
Privacy

- Federal Data Protection Act
- Federal Law on Statistics
- European Data Protection Act (Regulation (EU) 2016/679)

"In the age of Big Data all data is personal. Big Data is therefore the end of anonymity."

(Boehme-Neßler (2016), The End of Anonymity, DuD - Data Protection and Data Security, No. 7, p. 423)
International level of Official Statistics

UN
- UNSC Global Working Group on Big Data
- UNECE Big Data Project Inventory
- UNECE Sandbox

ESS
- Big Data Roadmap and Action Plan
- ESSnet Big Data
- Steering Group Big Data
- Task Force Big Data

© Federal Statistical Office Germany,
Institute for Research and Development in Official Statistics
Big Data and Official Statistics

Big Data UN Global Working Group

- members: countries and international organisations (OECD, Eurostat, World Bank)
- different task teams
- ‘Big Data project inventory’

UN Economic Commission for Europe (UNECE)

- Big Data projects on partnerships, quality, skills
- ‘Big Data Inventory’: open access online platform with detailed information about Big Data projects
European Statistical System (ESS)

Scheveningen Memorandum, 2013
- Big Data presents new challenges and opportunities for official statistics

ESS Task Force Big Data
- Identify priority actions and formulate a project proposal
- Manage and co-ordinate the implementation of the ESS Big Data Action Plan and Roadmap

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

ESSnet Big Data
- 22 national statistical institutes and organizations working together on different digital data sources

© Federal Statistical Office Germany, Institute for Research and Development in Official Statistics
DGINS: Scheveningen Memorandum

2. It is essential to develop an „Official Statistics Big Data Strategy“ at national and EU-level.
3. The implications of BiG data for legislation with regard to data protection and personal rights should be properly addressed.
4. Several NSIs are currently initiating or considering different uses of Big Data, with a momentum to share experiences and to collaborate.
5. Developing the necessary skills and capabilities to effectively explore Big Data is essential for their integration into the ESS.
6. Multidisciplinary character of Big Data requires synergies and partnerships to be effectively built.
8. The DGNIs agree on importance of following up the implementation of this memorandum by adopting an ESS action plan and roadmap by mid-2014.
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

© Federal Statistical Office Germany, Institute for Research and Development in Official Statistics
ESS Big Data Roadmap und Action Plan

Long-term vision (beyond 2020)
- Integrating digital data sources into statistics production
- Adapting national and European legal frameworks

Medium-term goals (until 2020)
- Carrying out feasibility studies
- Developing implementation recommendations

Short-term goals (by the end of 2016)
✓ Verifying data sources regarding availability, quality and legal framework for different applications
ESSnet Big Data Project

02/2016 – 04/2018

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

2016: Section ‘Co-operation with the scientific community, Microsimulation, Big Data’ is created

National Big Data Roadmap

- Big Data Lab: build up a stock of large digital datasets
- development of a Big Data training programme
- intensified collaboration with the Federal States on Big Data
- set up co-operations with governmental departments
- develop a communication strategy
Big Data Projects at Destatis

- European Space Agency programme: Copernicus
- GebäuDE-21
- Web Scraping and Scanner data
- Contacts to Vodafone, Telekom, DLR, Bitkom, UNECE
- ESSnet Big Data

© Federal Statistical Office Germany, Institute for Research and Development in Official Statistics
Web Scraping of Job Vacancies

Data access
- identify the most important national job websites
- legal aspects and copyright: who owns the data?
- implementation of tools and technical infrastructure

Data handling
- remove duplicates
- exclude the records which are not eligible
- classification of job vacancies
- quality assessment

© Federal Statistical Office Germany, Institute for Research and Development in Official Statistics
Web Scraping for Price Statistics

- replace manual price collection on the internet
- increase number of collected prices

Collected prices
- flights, hotels, online retailers, online pharmacies, car rentals, rail fares, coach transports, city trips, package holidays

Outlook
- automation of further price collections
- development of a working environment for the permanent use of automatized price collections
Challenges and risks

Before data acquisition:
- legislation
- data confidentiality and security
- adverse public perception

After data acquisition:
- representativity
- duplicates
- stable access
- data source manipulations
Future Projects

- scanner data: retail prices
- mobile phone data: population, tourism, commuting statistics
- remote sensing data: statistics about buildings, energy consumption, emissions

© Federal Statistical Office Germany, Institute for Research and Development in Official Statistics
What do you think does have to change in statistical education in the times of Big Data?
Statistical Education in times of Big Data

**Context**

- Permanent growth of accessible digital data
- Arising needs for development of statistical education
- Data Scientists and iStatisticians wanted/needed
- EMOS - common solution to the new challenges

© Federal Statistical Office Germany, Institute for Research and Development in Official Statistics
What are the target groups for statistical education?

- Pupils, students, professionals
- Data producer, data manager
- Data users, analysts, decision maker
- Data scientists, iStatistician
What are the instruments to teach statistics?

- Traditional methods like front teaching
- E-learning
- Webinars
- Massive Open Online Courses (MOOC)
- Blended learning

© Federal Statistical Office Germany, Institute for Research and Development in Official Statistics
Transformation of the statistician’s profile: the Data Scientist

Competences

- Analytical and computing skills
- Delivering quality and ethical analyses
- Process management skills
- Communication skills
- Developing analytical expertise
The Data Scientist will be a team with specialized member skills
Big Data Team Level Competency

- Team work
- Interpersonal and communication
- Delivery of results
- Innovation and contextual awareness
- Specialist knowledge and expertise
- Statistical/IT skills
- Data Analytical/Visualisation skills

Competency profiles have been created by the UNECE High-Level Group for the Modernisation of Official Statistics

http://www1.unece.org/stat/platform/display/bigdata/Competency+Profiles
Big Data Team Leader Level Competency

- Leadership and Strategic Direction
- Judgement and decision-making
- Management and delivery of results
- Building relationships and communication
- Specialist knowledge and expertise
- Statistical/IT skills
- Data Analytical/Visualisation skills

Competency profiles have been created by the UNECE High-Level Group for the Modernisation of Official Statistics

http://www1.unece.org/stat/platform/display/bigdata/Competency+Profiles
NSIs and universities

- Educating the next generation of statisticians
- Improving 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

Introduction to big data and its tools
Hands-on immersion on big data tools
The use of R in official statistics: model based estimates
Can a statistician become a data scientist?
Automated collection of online prices
Big Data sources: Web, Social Media and text analytics
Advanced big data sources – Mobile phone and other sensors

Course books: https://circabc.europa.eu/faces/jsp/extension/wai/navigation/container.jsp

© Federal Statistical Office Germany, Institute for Research and Development in Official Statistics
Thank you for the attention!

markus.zwick@destatis.de
https://de.linkedin.com/in/markus-zwick-72393213

© Federal Statistical Office Germany,
Institute for Research and Development in Official Statistics