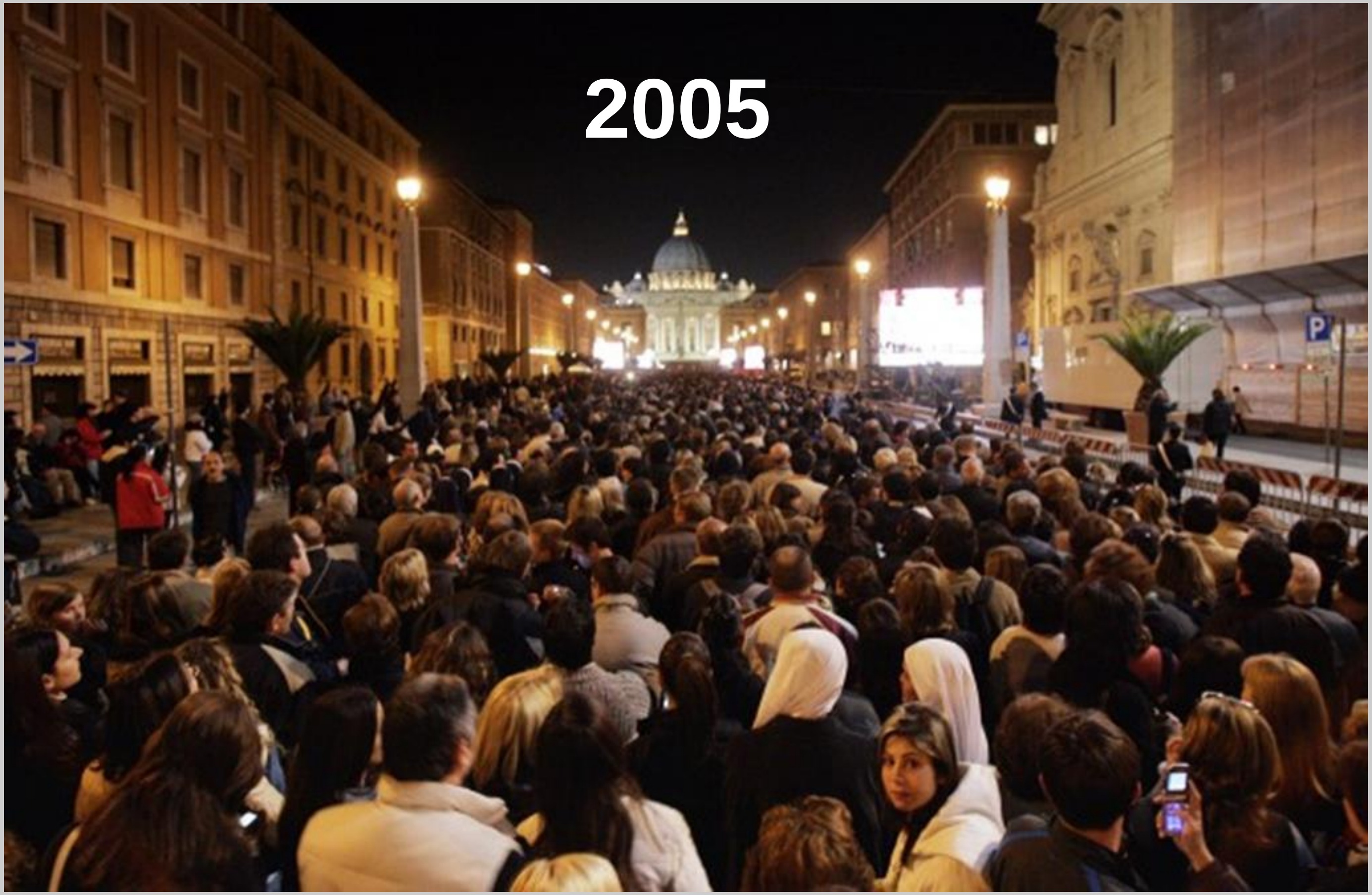
DESTATIS
wissen. nutzen.

Introduction to Big Data in Official Statistics

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

There was something different in the Vatican crowd in 2005…



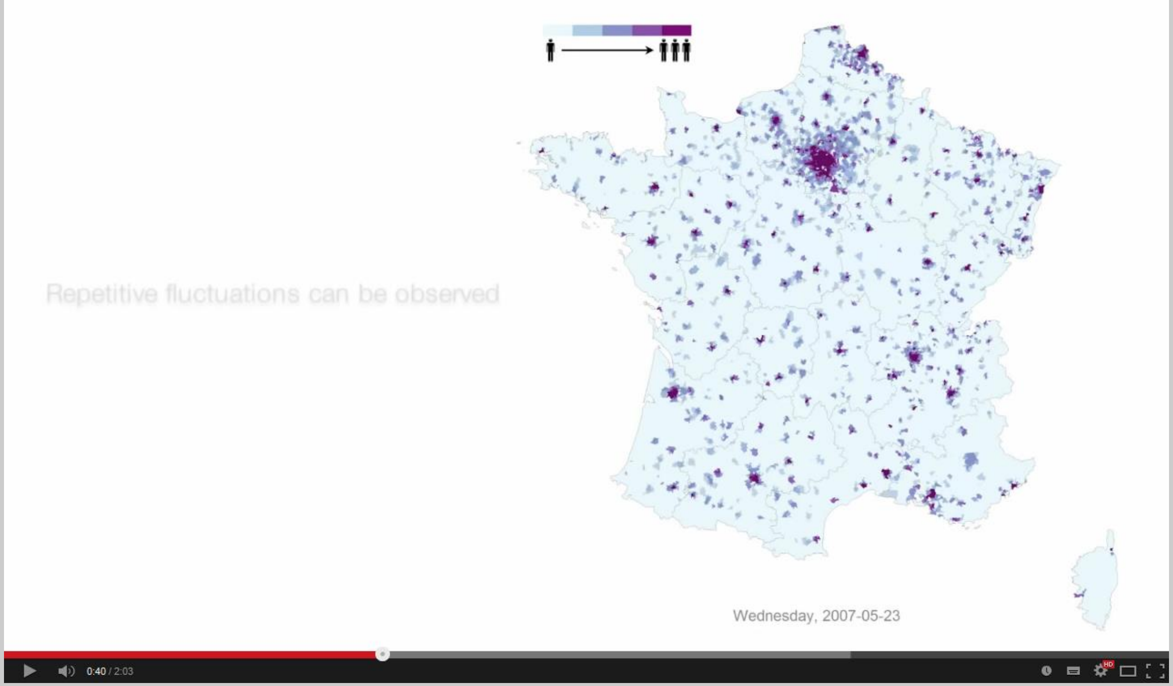
http://www.businessinsider.com/vatican-square-2005-and-2013-2013-3?IR=T

© FontShop/image100/Consumers/10021019

© Adam Gryko - Fotolia.com

© ponsulak - Fotolia.com

Population Mapping Using Mobile Phone Data



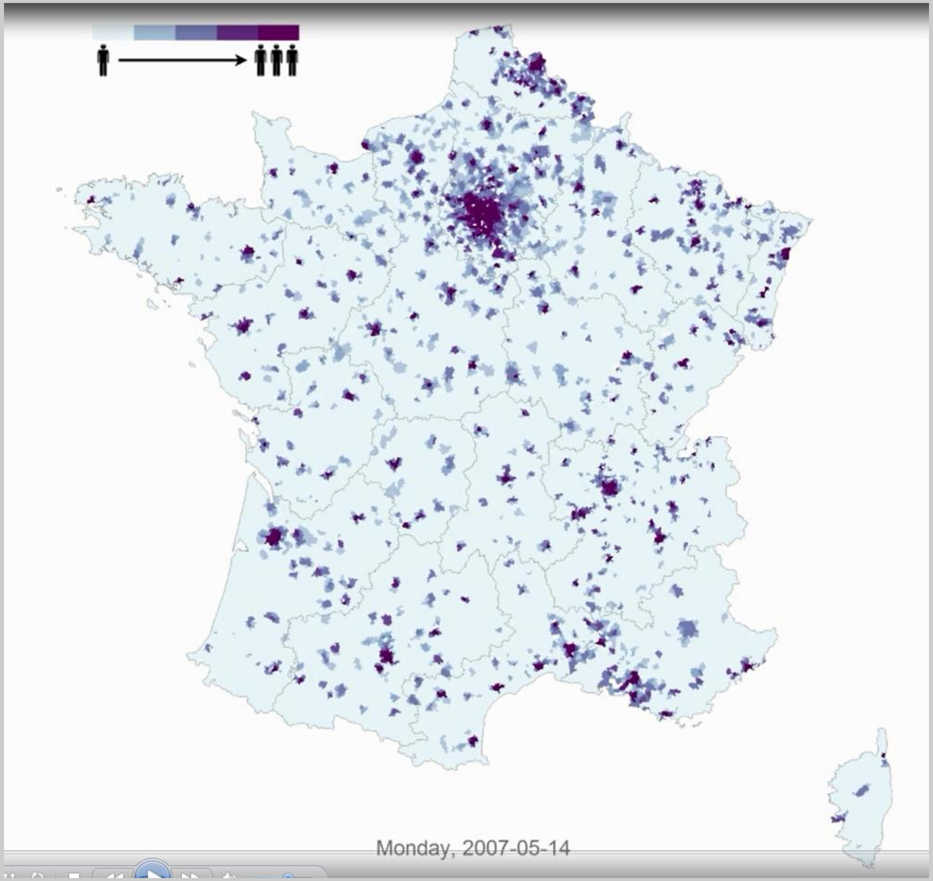
Repetitive fluctuations can be observed

Wednesday, 2007-05-23

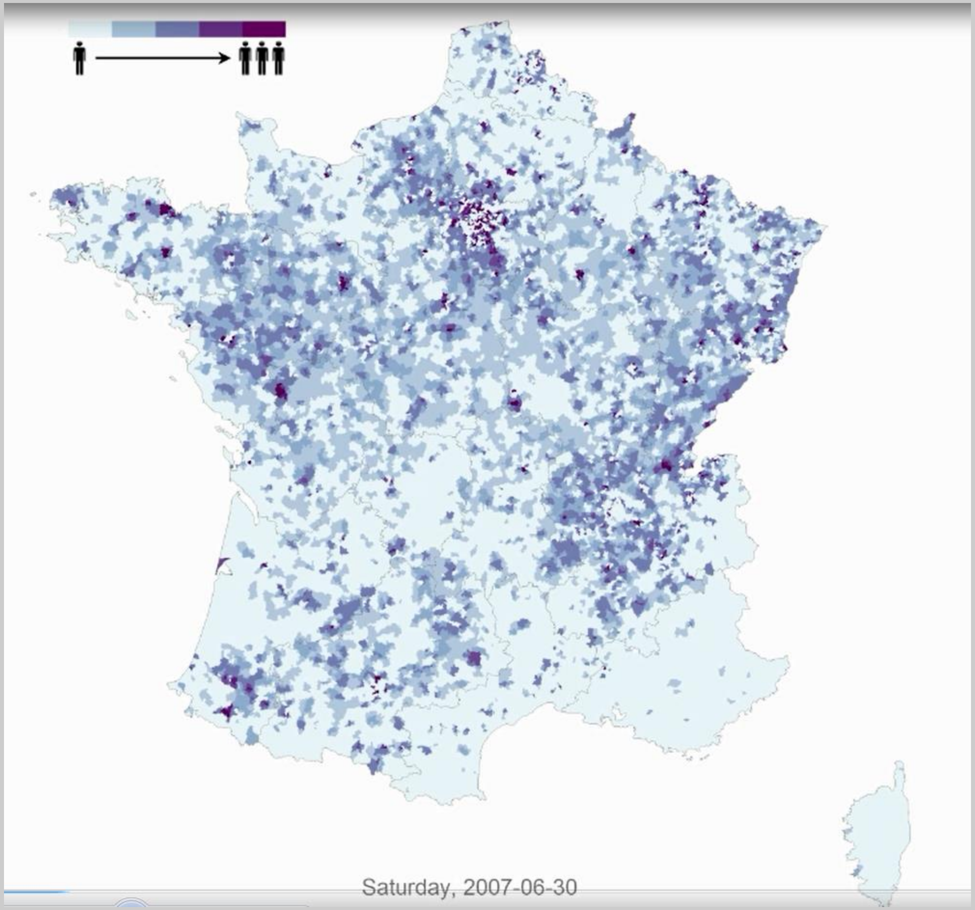
https://www.youtube.com/watch?v=qsUDH5dUnvY

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

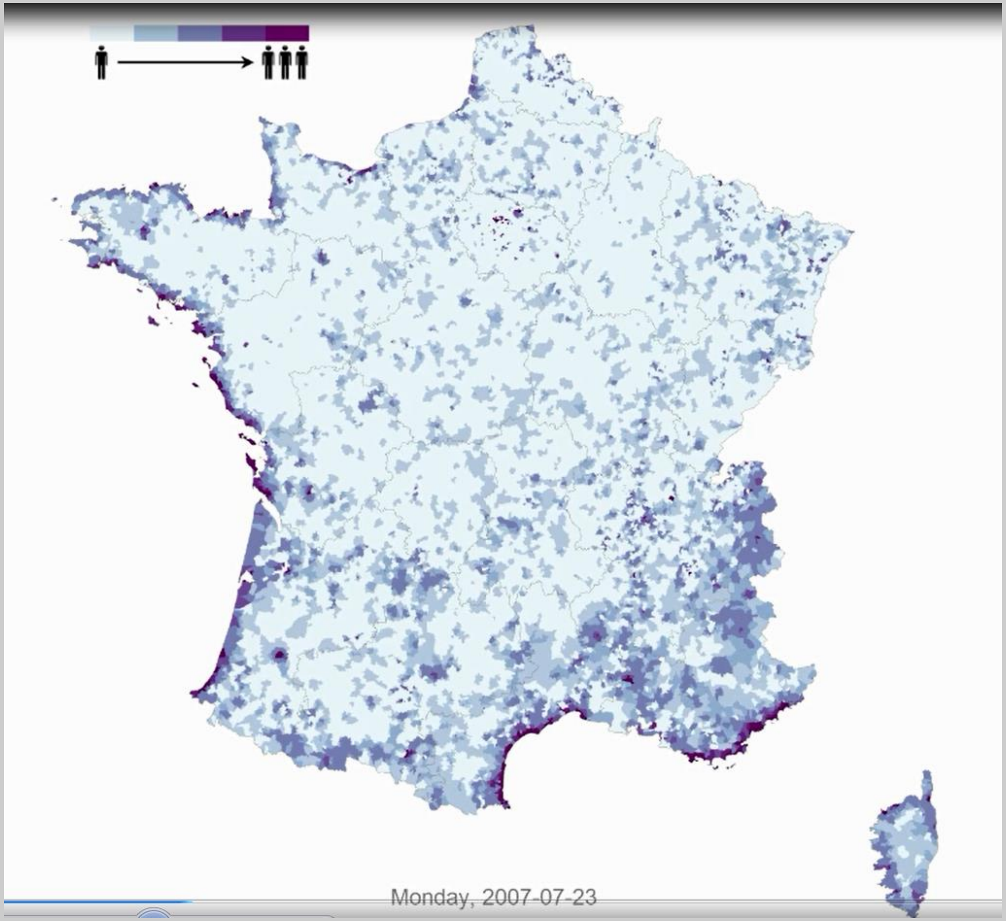
Population Mapping Using Mobile Phone Data



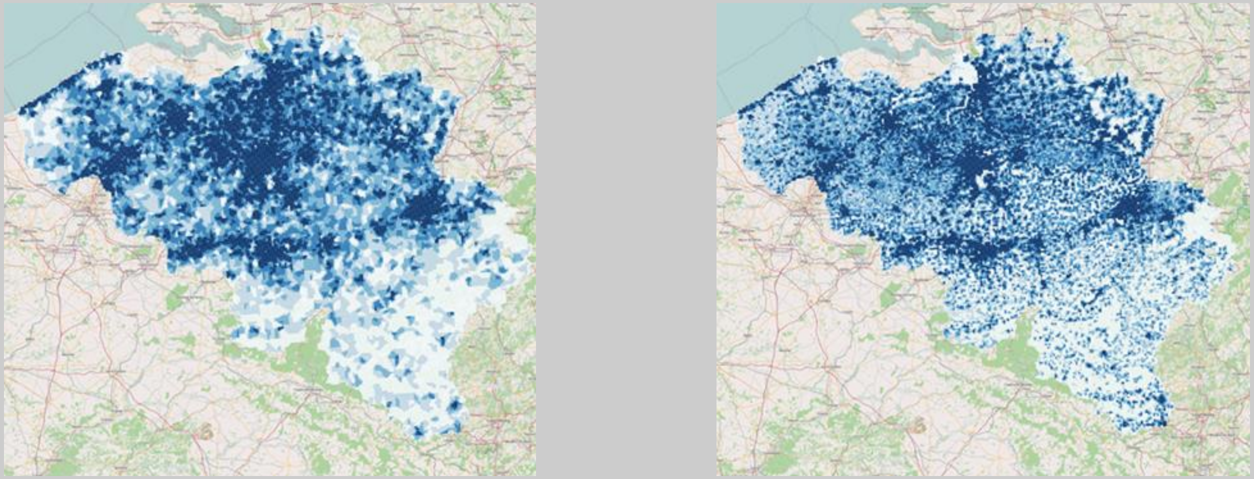
Population Mapping Using Mobile Phone Data



Population Mapping Using Mobile Phone Data



Population Mapping Using Mobile Phone Data



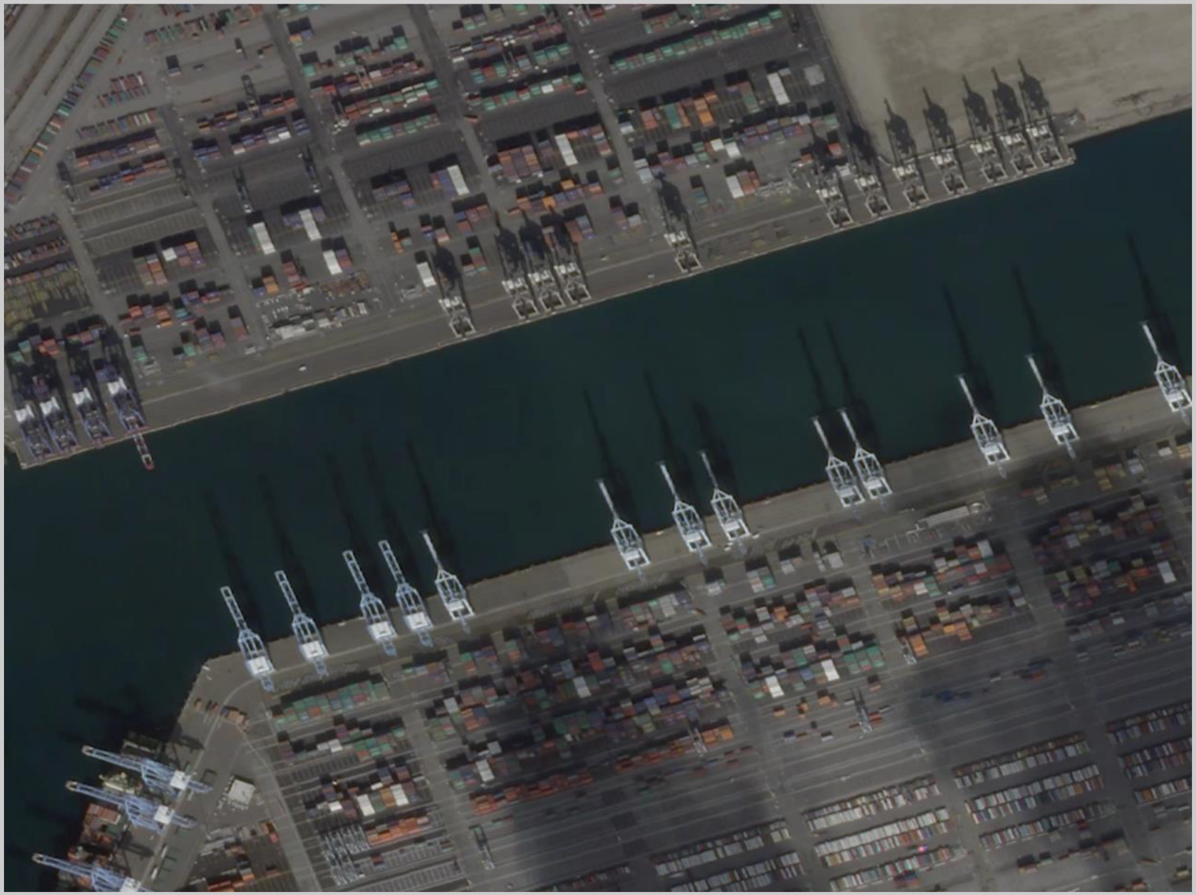
The graph shows population densities derived from mobile phone counts at 4 am on Thursday 8 October (left) and the 2011 population census (right). The Pearson correlation between these two datasets is 0.85, a clear indication that mobile phone data are able to provide a valid and accurate measure of population density.

de Meersman, F., G. Seynaeve, M. Debusschere, P. Lusyne, P. Dewitte, Y. Baeyens, A. Wirthmann, C. Demunter, F. Reis, H.I. Reuter (2016). Assessing the Quality of Mobile Phone Data as a Source of Statistics, Q2016 Conference paper.

Google Skybox/Terra Bella

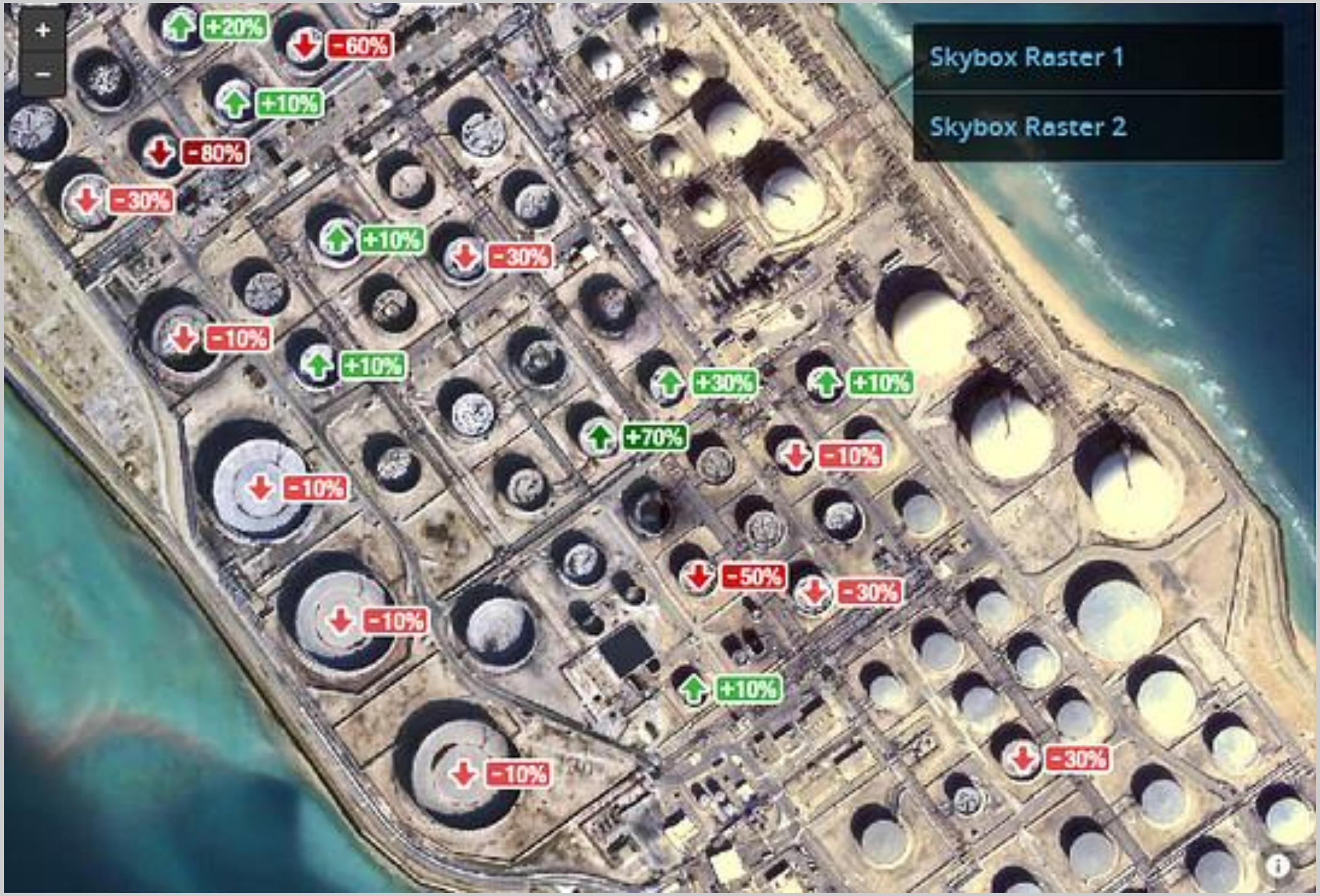
Quelle: https://terrabella.google.com

Google Skybox/Terra Bella

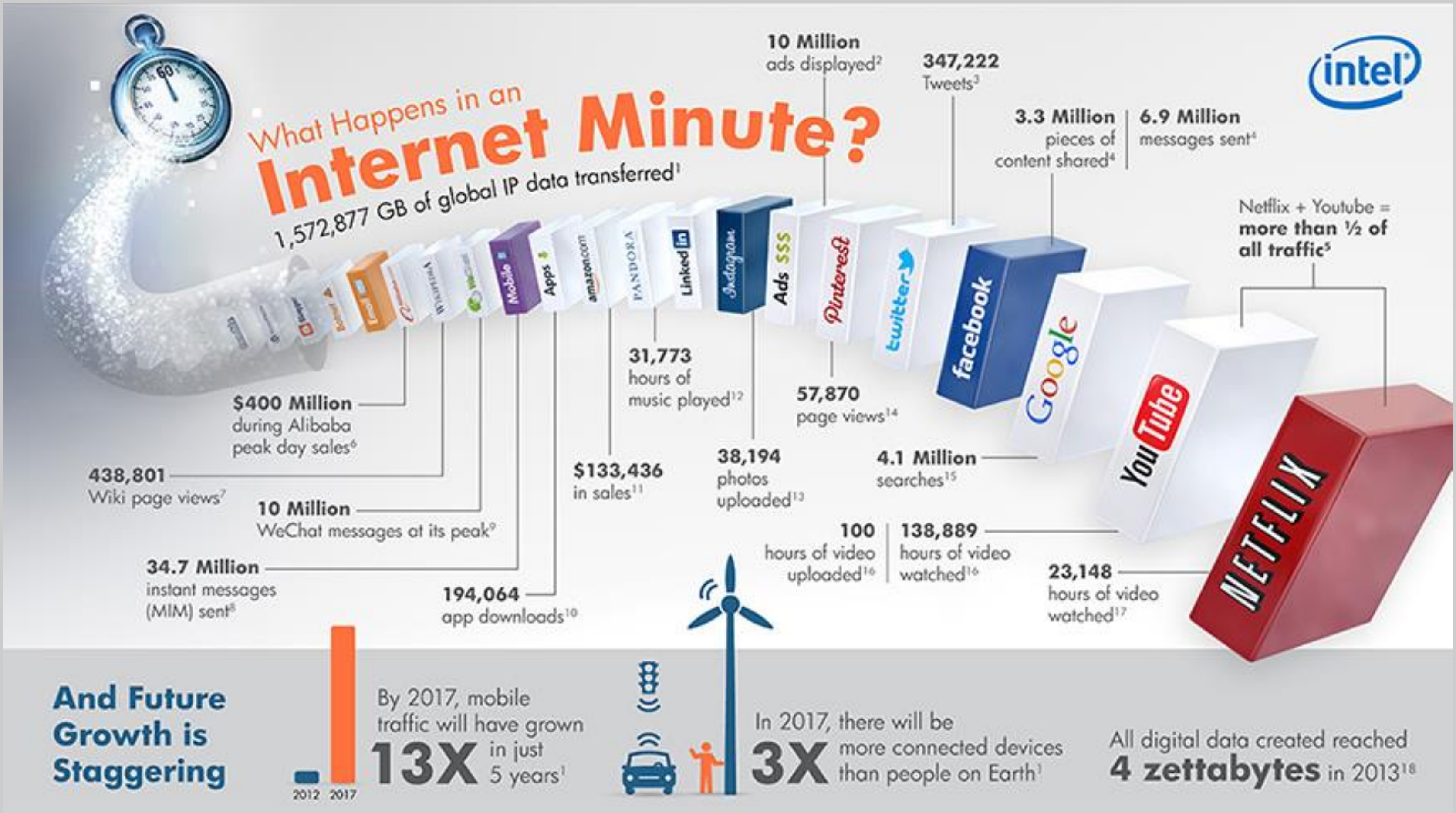


Quelle: https://terrabella.google.com

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.
http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html

What do you think are the main characteristics of Big Data?

Features Big Data (“3 Vs“)

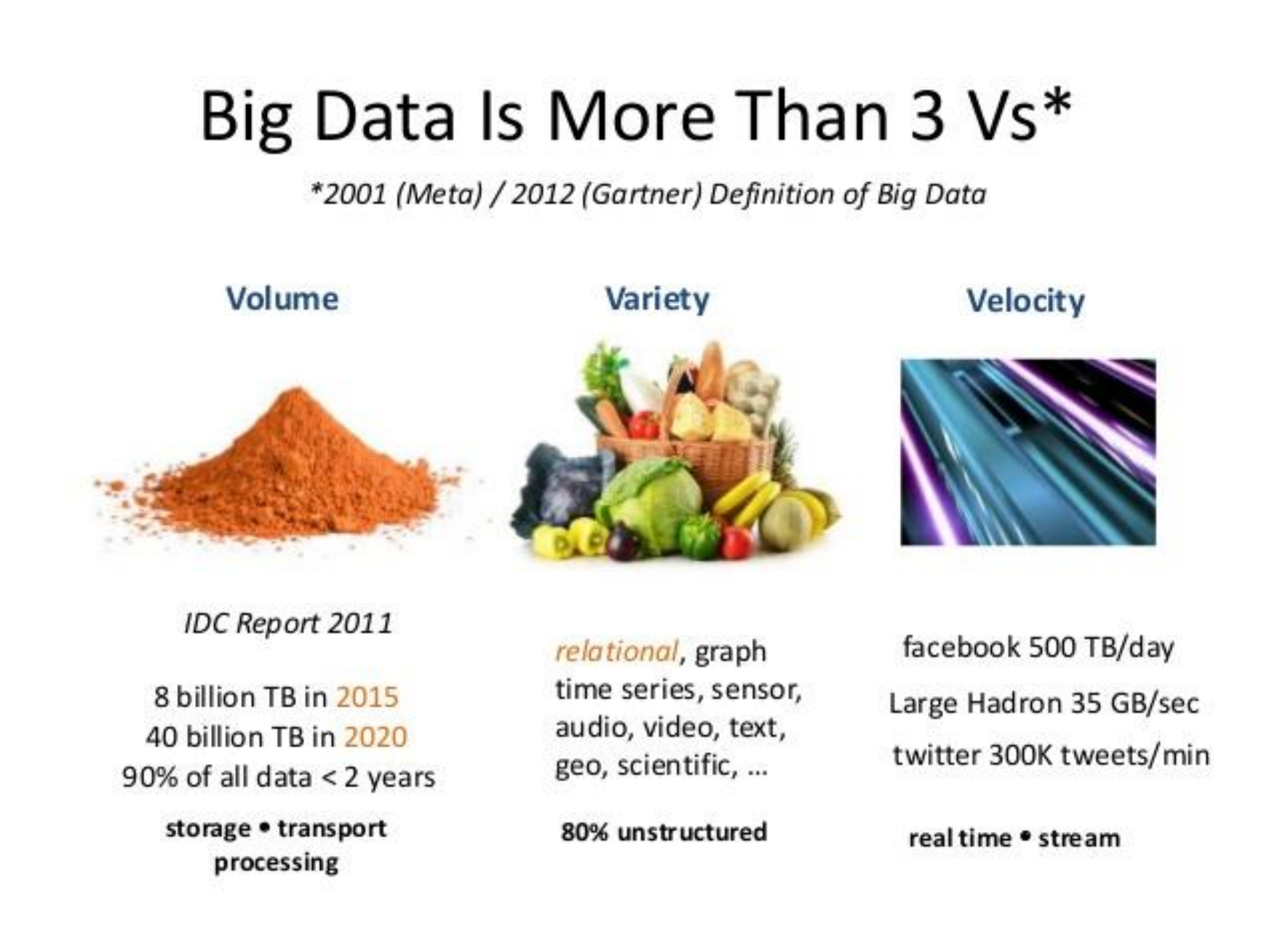
- **Volume – Large amount of Data**

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

© niyazz - Fotolia.com

© iStockphoto.com / olaser

© igor - Fotolia.com



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

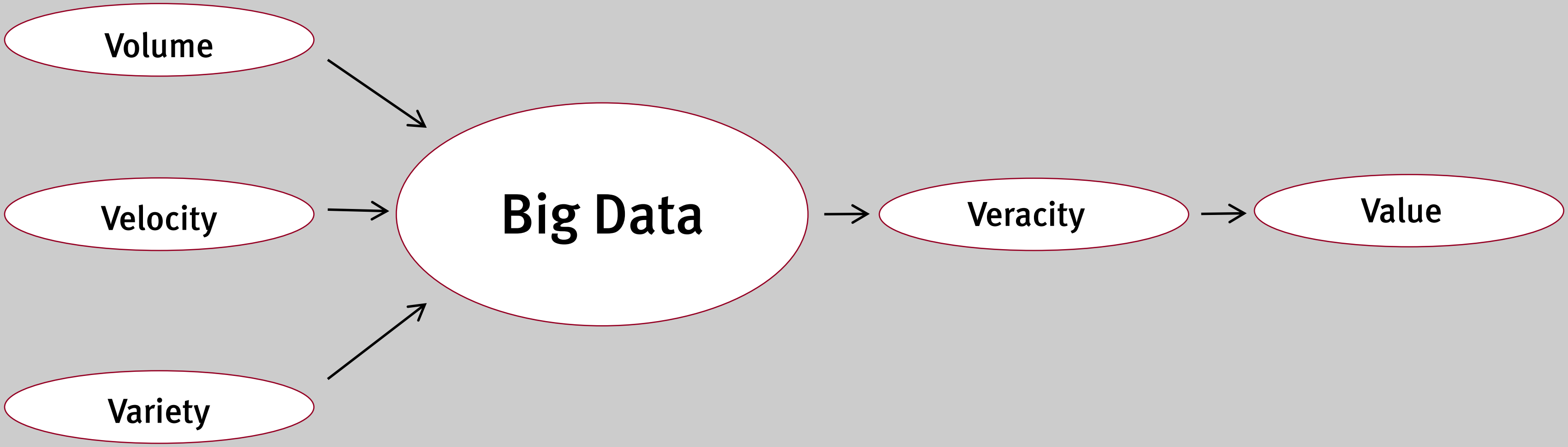
Large Hadron 35 GB/sec

twitter 300K tweets/min

real time • stream

Source: https://www.slideshare.net/andrewgardner5811/big-data-and-the-art-of-data-science **(IDC – International Data Corporation)**

‚Five Vs‘ of Big Data



Volume

Velocity

Variety

Big Data

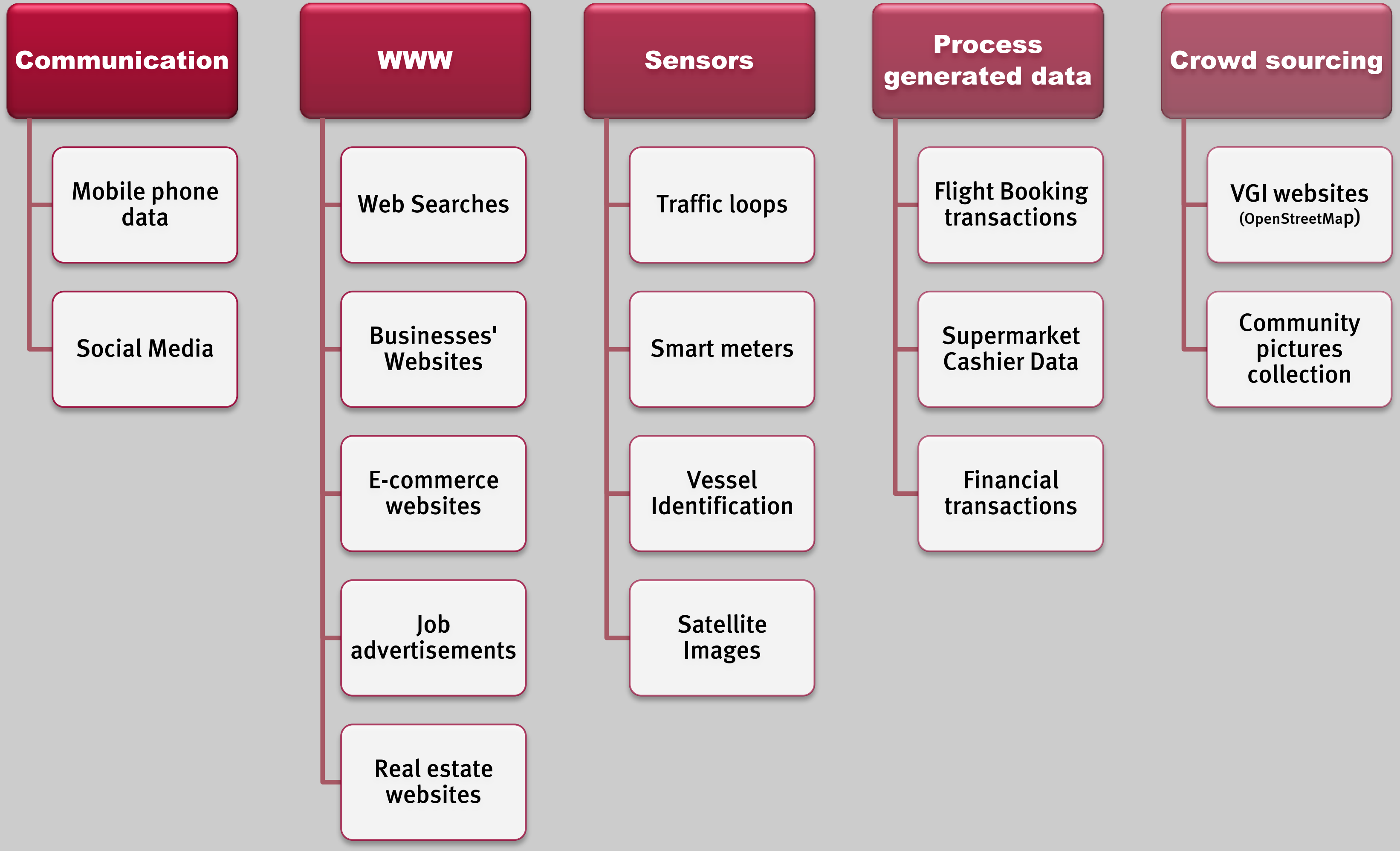
Veracity

Value

- ‚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
https://www.slideshare.net/kuonen/the-power-of-data-insights-big-data-as-the-fuel-and-analytics-as-the-engine-of-the-digital-transformation

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

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

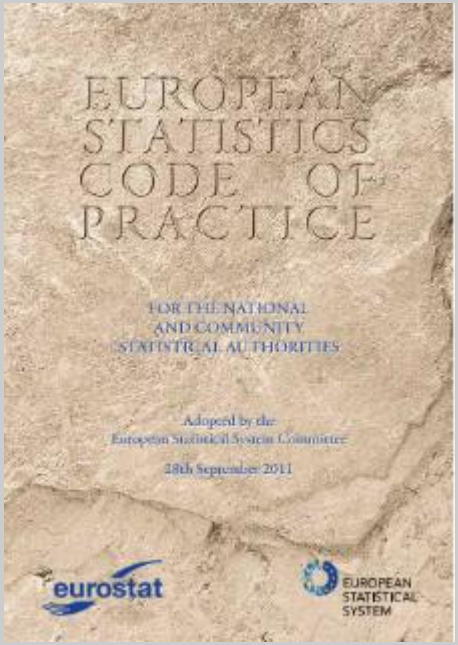
http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data

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

FOR THE NATIONAL AND COMMUNITY STATISTICAL AUTHORITIES

Adopted by the European Statistical System Committee

28th September 2011

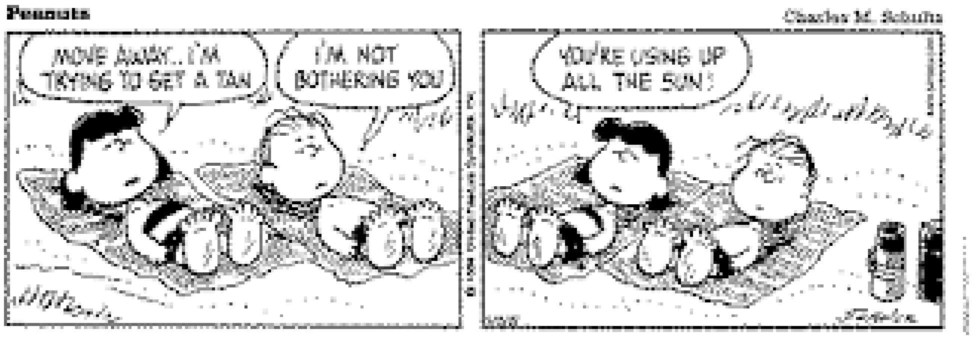
eurostat

EUROPEAN STATISTICAL SYSTEM

European Statistics
Code of Practice

and Official Statistics are public goods

Should Official Statistics go into a competition with new data producers like Google?


Peanuts
Charles M. Schulz
MOVE AWAY...I'M TRYING TO GET A TAN
I'M NOT BOTHERING YOU
YOU'RE USING UP ALL THE SUN!

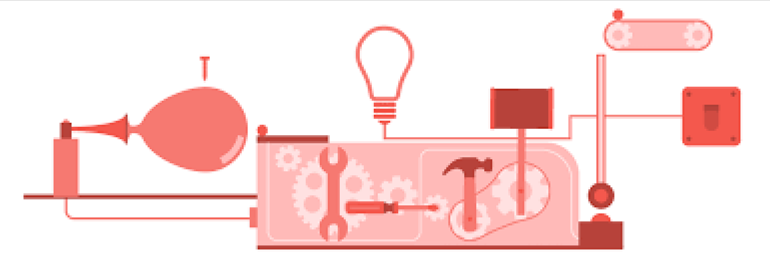
Common Good? Rival or Non-Rival?

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

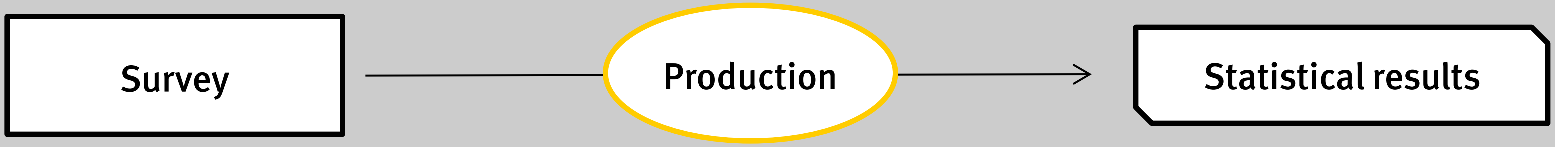


What does big data mean for Official Statistics?

- 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

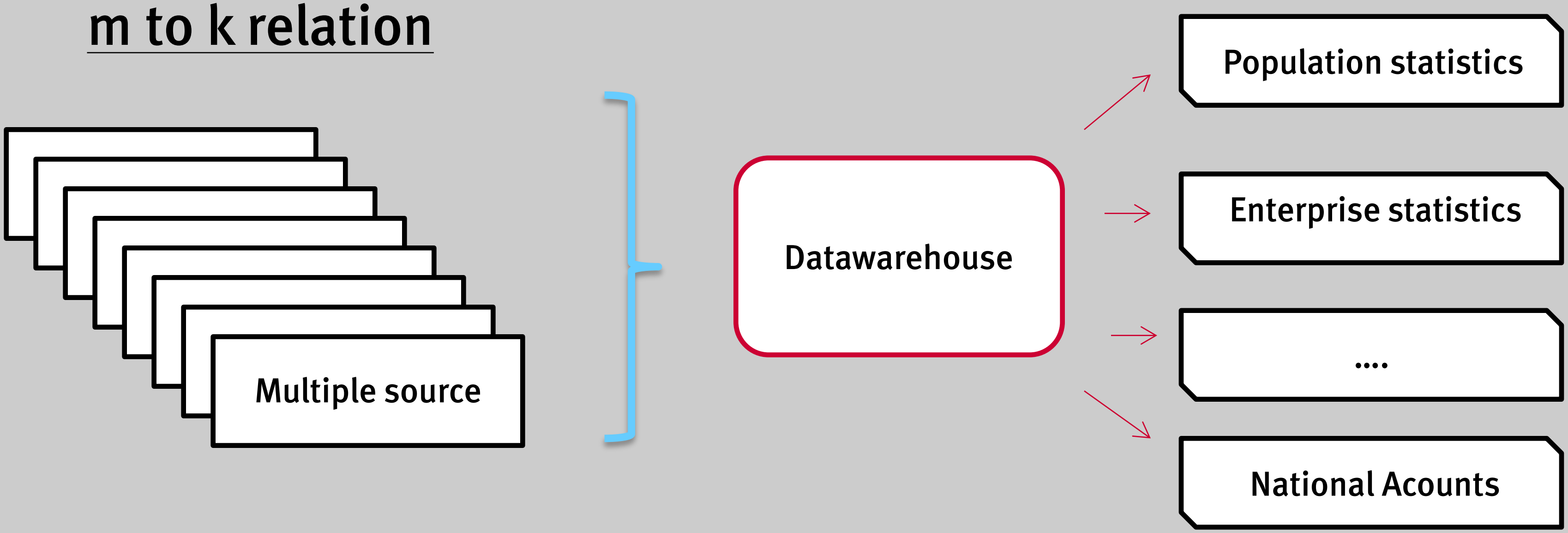
1 to 1 relation


Survey

Production

Statistical results

m to k relation


Multiple source

Datawarehouse

Population statistics

Enterprise statistics

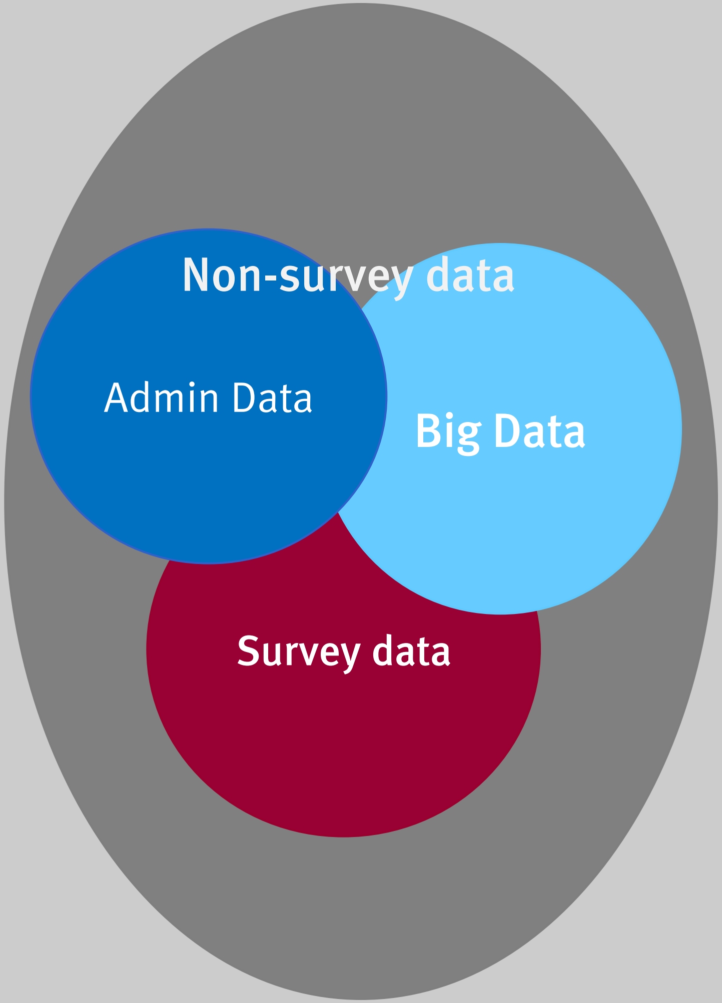
....

National Acounts

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)

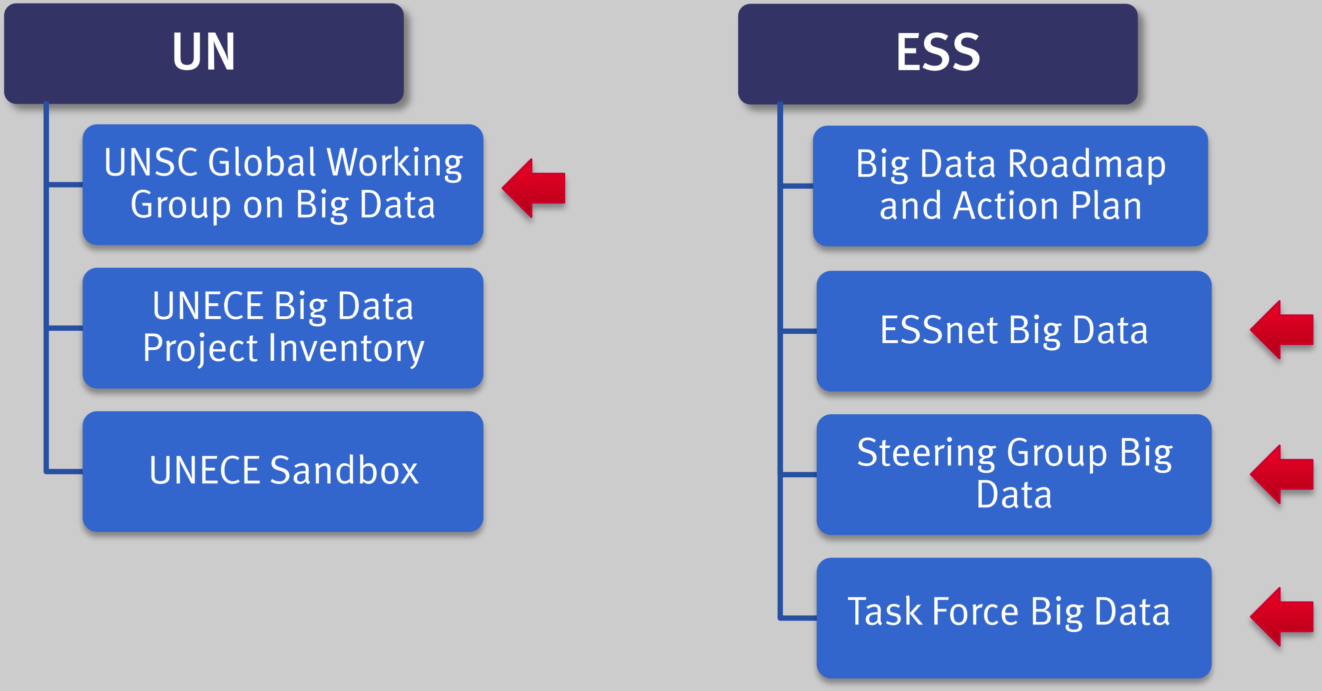
"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

"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

EUROPEAN
STATISTICAL
SYSTEM



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

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

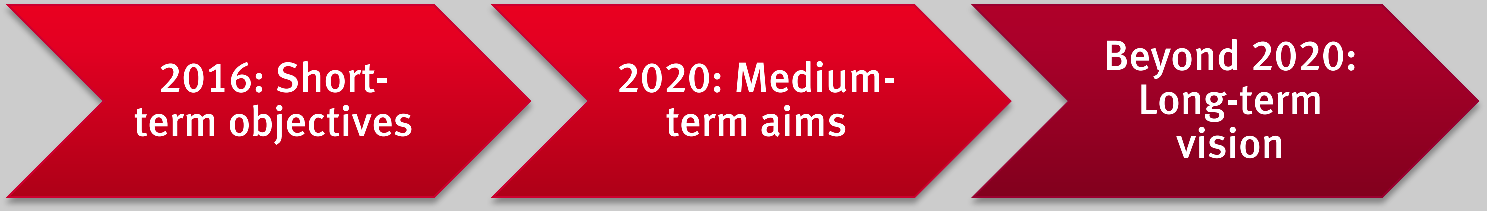
- 22 national statistical institutes and organizations working together on different digital data sources

DGINS: Scheveningen Memorandum

1. Big Data represent new opportunities and challenges for Official Statistics
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 adressed.
4. Several NSIs are currently initiating or considering different uses od 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.
7. The use of Big Data in the context of official statistics requires new developments in methodology, quality assessment and IT related issues.
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


2016: Short-term objectives

2020: Medium-term aims

Beyond 2020: Long-term vision

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

EUROPEAN
STATISTICAL
SYSTEM

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

© Franz Pfluegl - Fotolia.com

© ponsulak - Fotolia.com

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

Geographical information

Mobil data

Prices

Labour market

© Fancy by Veer/The World from Above/FAN2043678

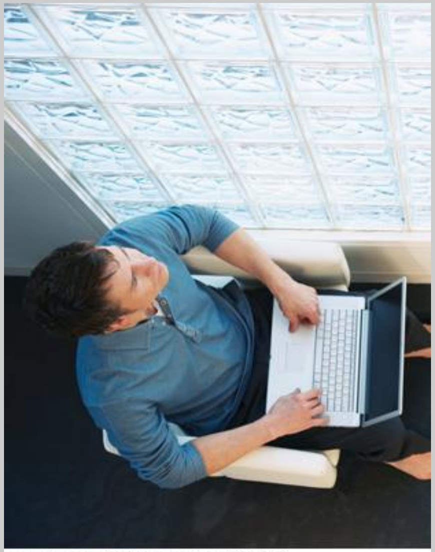
Web Scraping of Job Vacancies

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



© Fancy by Veer/Flex Workers/2923552

Web Scraping for Price Statistics

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

© Adam Gryko - Fotolia.com

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

© ponsulak - Fotolia.com

© Fancy by Veer/The World from Above/FAN2043631

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

© peerayot - Fotolia.com

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?

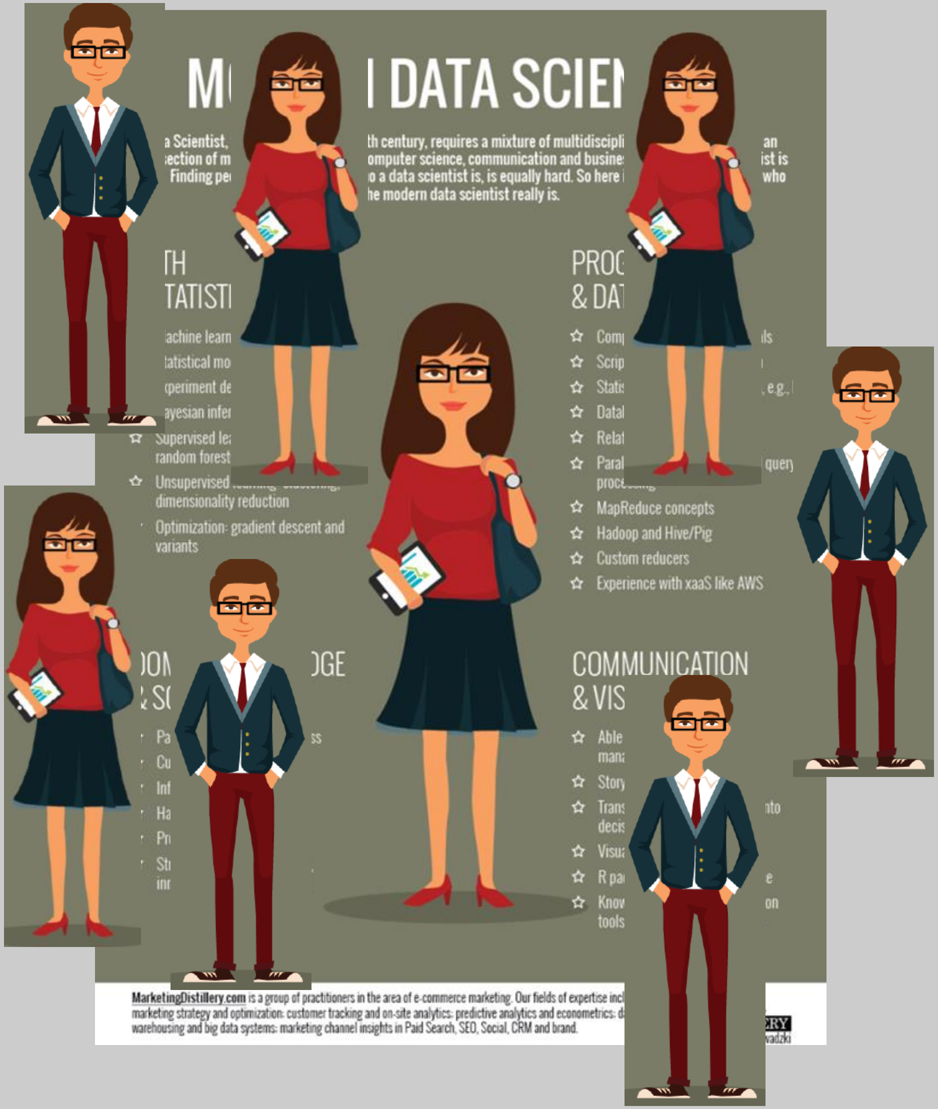
- Traditional methods like front teaching
- E-learning
- Webinars
- Massive Open Online Courses (MOOC)
- Blended learning

© Judywie / photocase.de

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


M
I DATA SCIEI
a Scientist,
th century, requires a mixture of multidiscipli
omputer science, communication and busine
o a data scientist is, is equally hard. So here i
he modern data scientist really is.
TH
TATISTI
achine learn
tatistical mo
periment de
ayesian infer
Supervised le
random forest
Unsupervised
dimensionality reduction
Optimization: gradient descent and variants
PROC
& DAT
Com
Scrip
Stati
Data
Rela
Paral
proc
MapReduce concepts
Hadoop and Hive/Pig
Custom reducers
Experience with xaaS like AWS
COMMUNICATION
& VIS
Able
man
Story
Trans
decis
Visu
R pa
Know
tools
MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise incl
marketing strategy and optimization: customer tracking and on-site analytics: predictive analytics and econometrics: d
warehousing and big data systems: marketing channel insights in Paid Search, SEO, Social, CRM and brand.

The Data Scientist will be a team with specialized member skills

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

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

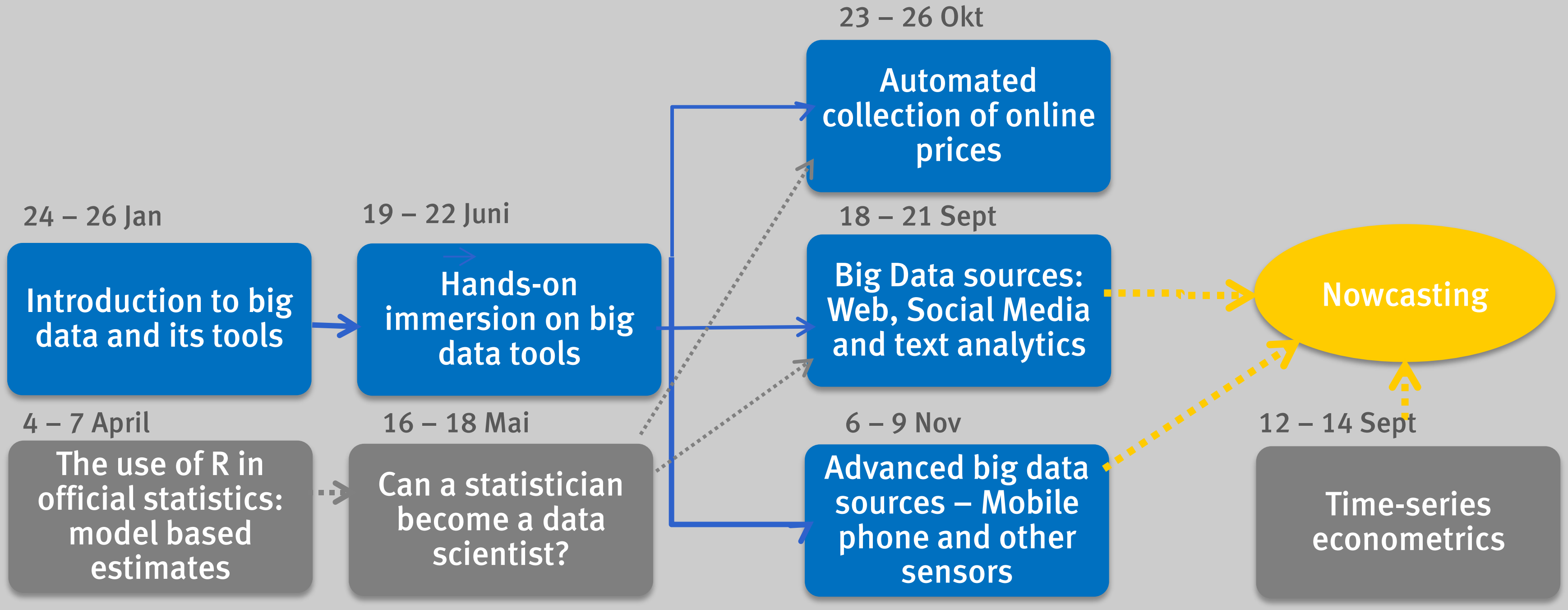
- Label for the Master study
- Network of master studies with the main focus on Official Statistics and data production at European level
- The aim: To strengthen cooperation between universities and producers of Official Statistics and to train professionals
- A practical mix of competences and knowledge 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

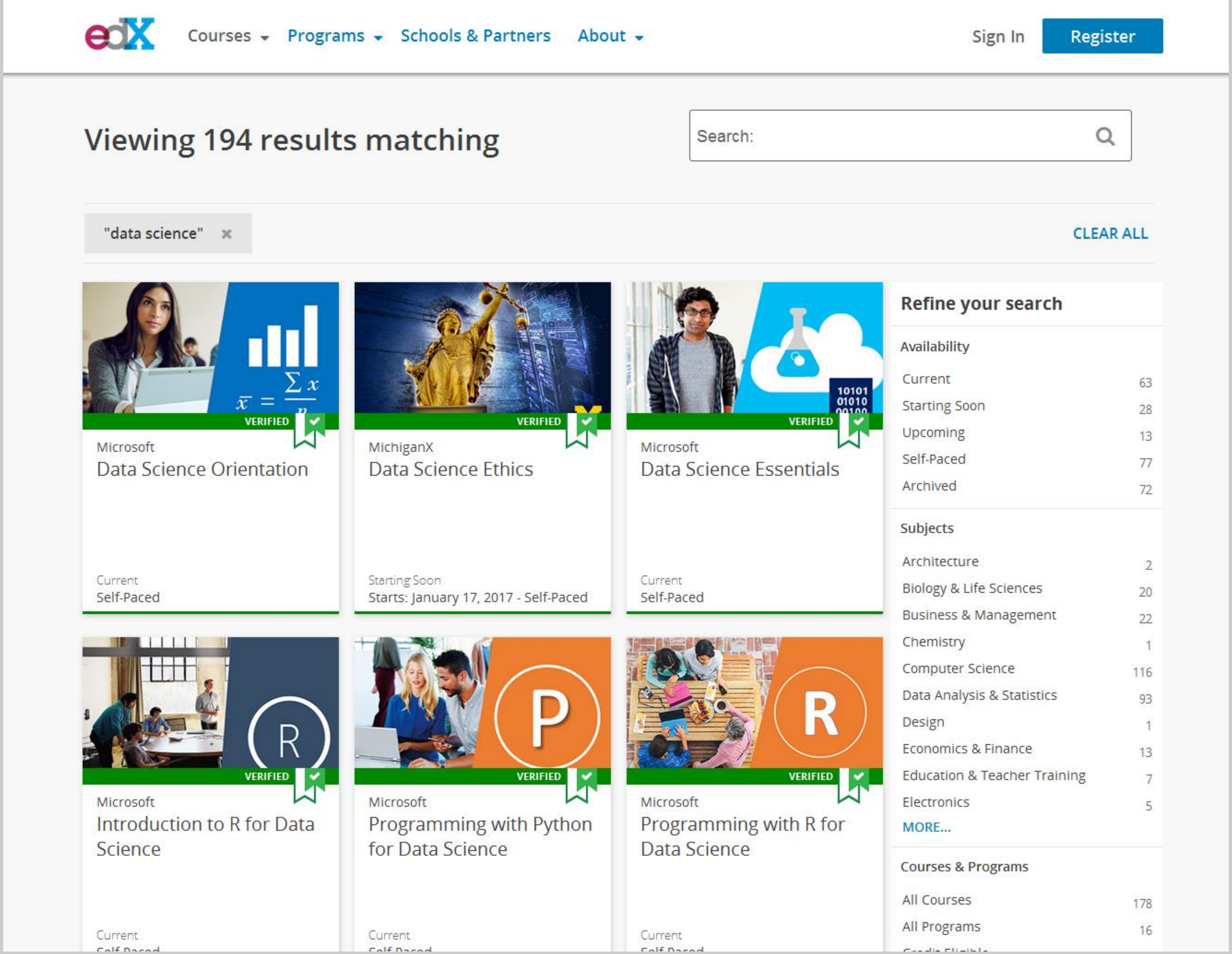
- Set up by the Irish Centre for High-end Computing and the Irish Central Statistics Office
- Contains data sets and tools for international experiments
- Remote access and processing of data
- Experiments with data from social media, mobile phones, smart meters, traffic loops

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

European Statistical Training Programme ESTP - Big Data 2017


24 – 26 Jan
Introduction to big data and its tools
19 – 22 Juni
Hands-on immersion on big data tools
23 – 26 Okt
Automated collection of online prices
18 – 21 Sept
Big Data sources: Web, Social Media and text analytics
6 – 9 Nov
Advanced big data sources – Mobile phone and other sensors
Nowcasting
4 – 7 April
The use of R in official statistics: model based estimates
16 – 18 Mai
Can a statistician become a data scientist?
12 – 14 Sept
Time-series econometrics

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



Viewing 194 results matching

Search:

"data science" ✖

CLEAR ALL

Microsoft
Data Science Orientation
Current
Self-Paced

MichiganX
Data Science Ethics
Starting Soon
Starts: January 17, 2017 - Self-Paced

Microsoft
Data Science Essentials
Current
Self-Paced

Microsoft
Introduction to R for Data Science
Current
Self-Paced

Microsoft
Programming with Python for Data Science
Current
Self-Paced

Microsoft
Programming with R for Data Science
Current
Self-Paced

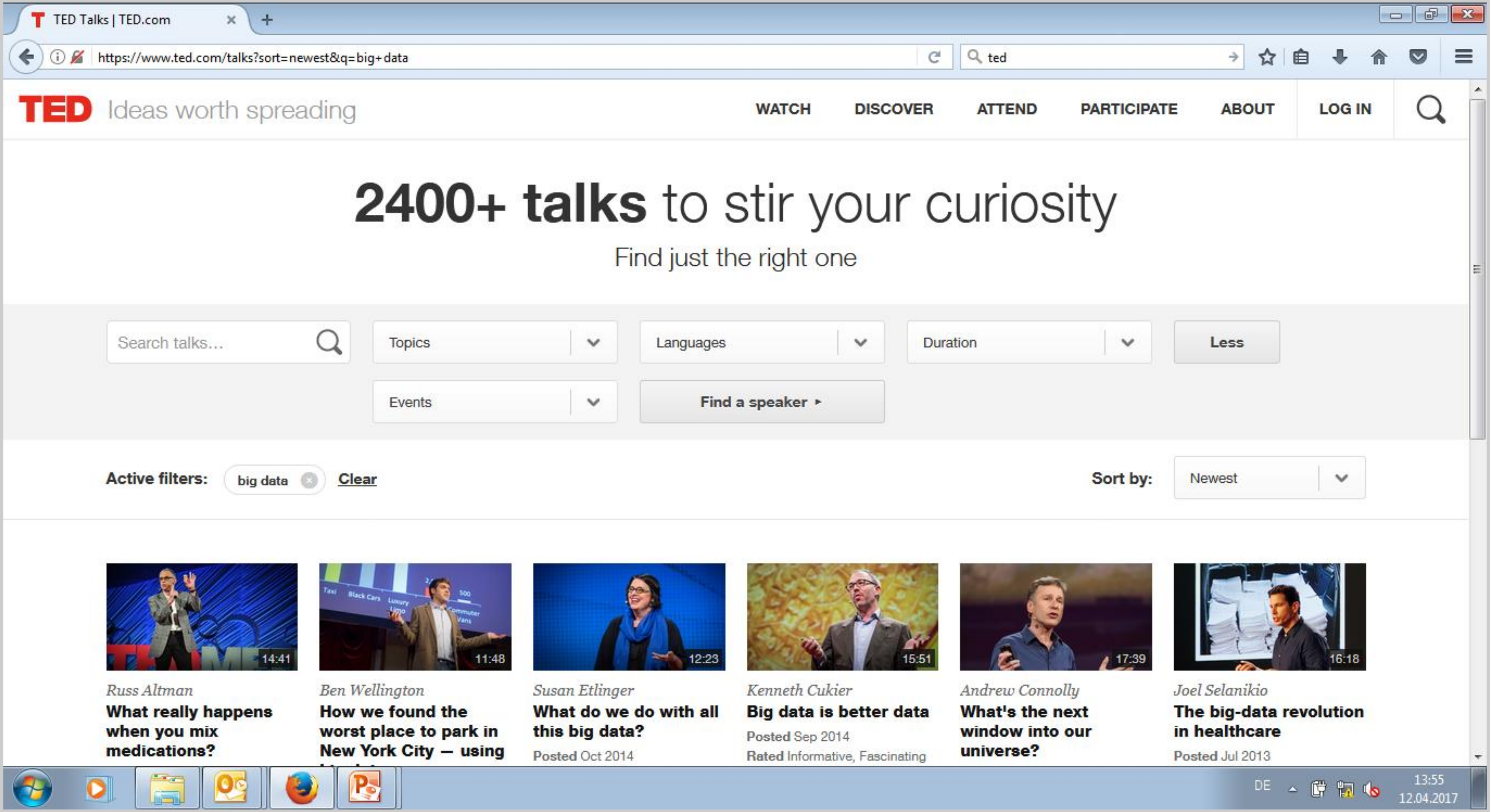
Refine your search

Availability	
Current	63
Starting Soon	28
Upcoming	13
Self-Paced	77
Archived	72

Subjects	
Architecture	2
Biology & Life Sciences	20
Business & Management	22
Chemistry	1
Computer Science	116
Data Analysis & Statistics	93
Design	1
Economics & Finance	13
Education & Teacher Training	7
Electronics	5

MORE...

Courses & Programs	
All Courses	178
All Programs	16
Credit Eligible	



https://www.ted.com/talks?sort=newest&q=big+data

ted

2400+ talks to stir your curiosity

Find just the right one

Active filters: big data ⊗ Clear

Sort by: Newest

Russ Altman
What really happens when you mix medications?

Ben Wellington
How we found the worst place to park in New York City — using

Susan Etlinger
What do we do with all this big data?
Posted Oct 2014

Kenneth Cukier
Big data is better data
Posted Sep 2014
Rated Informative, Fascinating

Andrew Connolly
What's the next window into our universe?

Joel Selanikio
The big-data revolution in healthcare
Posted Jul 2013

Thank you for the attention!

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