Is it possible to use Big Data in the 2020 Census Round?

UNECE-Eurostat Expert Group Meeting on Population and Housing Censuses
Geneva
28-30 September 2016

Janusz Dygaszewicz
Central Statistical Office of Poland
Director of Programming and Coordination of Statistical Surveys Department

Pope Benedict election

Pope Francis election

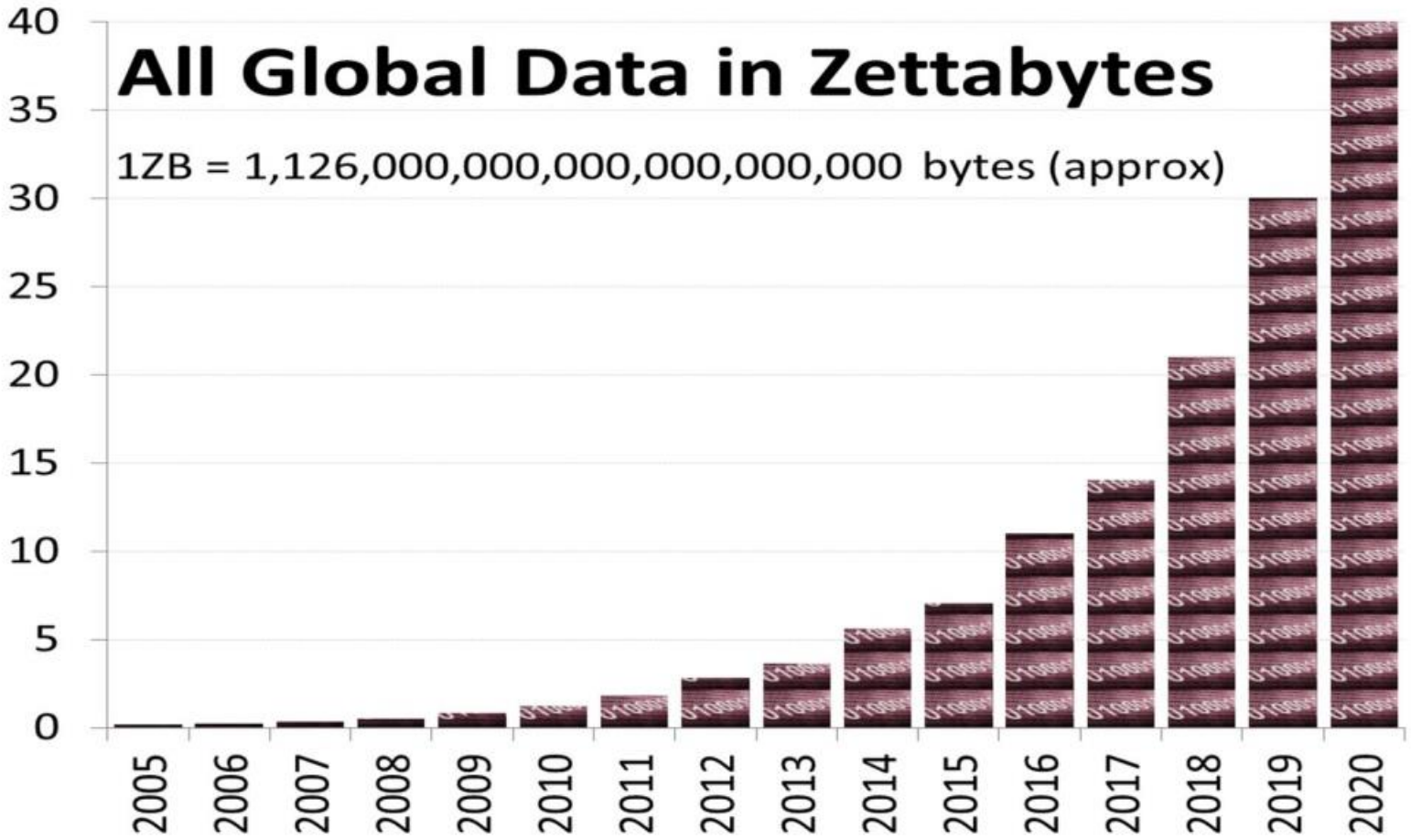
Eurostat

Using Big Data in Official Statistics

Currently official statistics are based on data from state registers and information obtained from surveys, respondents or acquired on the basis of interviewers and experts observations.

However, the world is continually changing, there are new phenomena which also require describing with statistical processes. Therefore it can't be limited only to the known data sources, you need to constantly seek new paths and solutions. The global trend in this field is Big Data.

Data



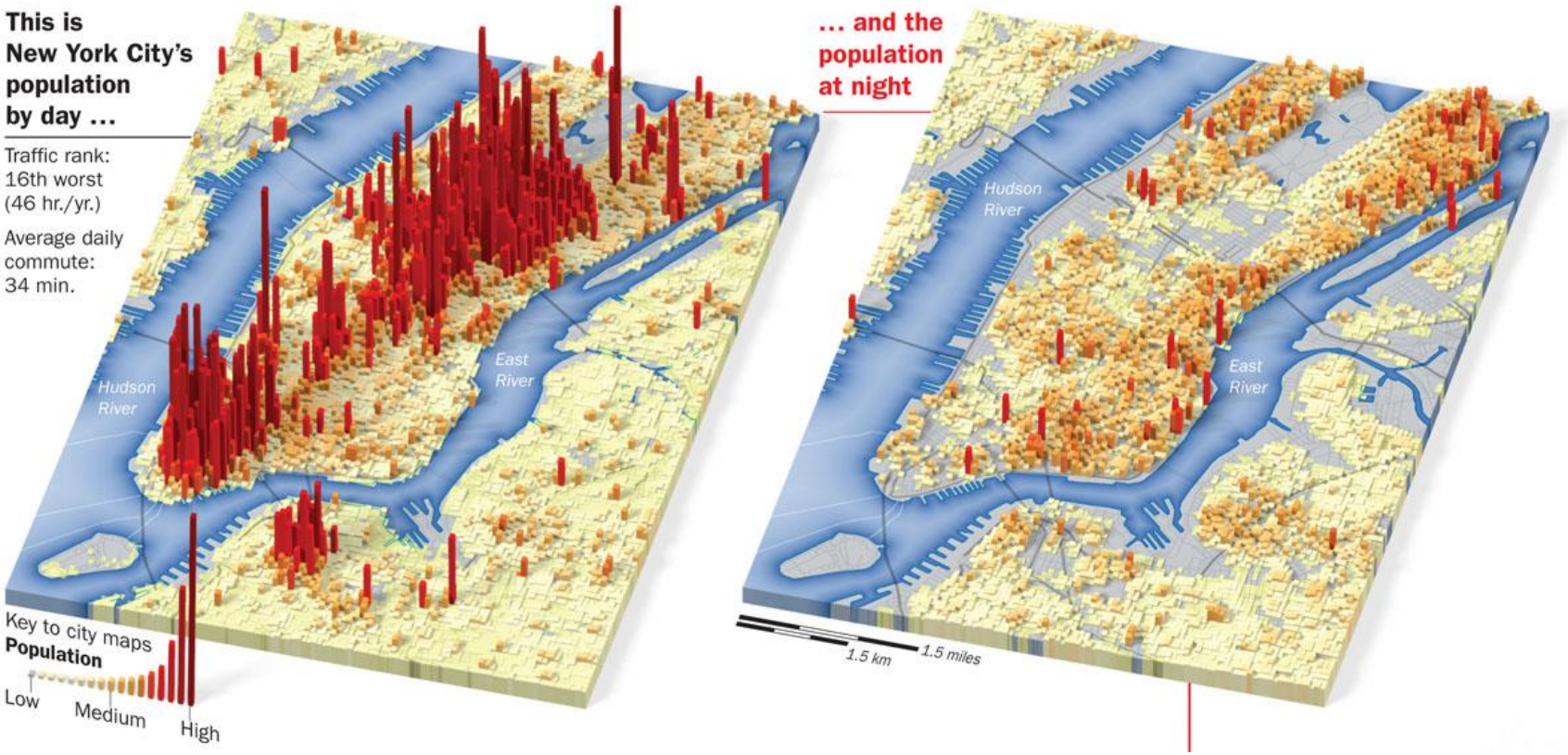
Source: UNECE

Sensors in mobile phones

Main culprit:

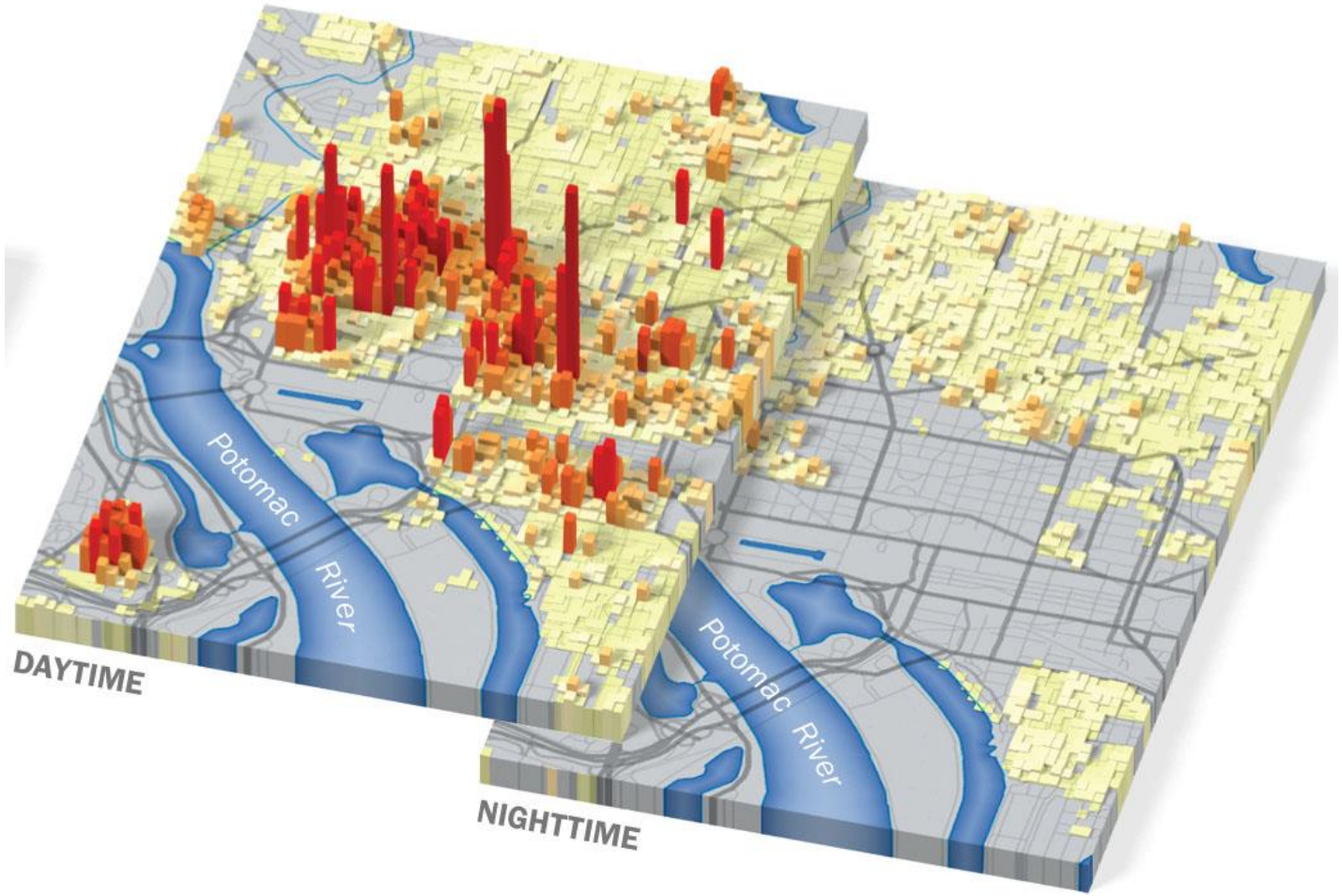
Source: http://rjacquez.com

Location data from mobile phones



Source: Time magazine Nov. 26, 2007

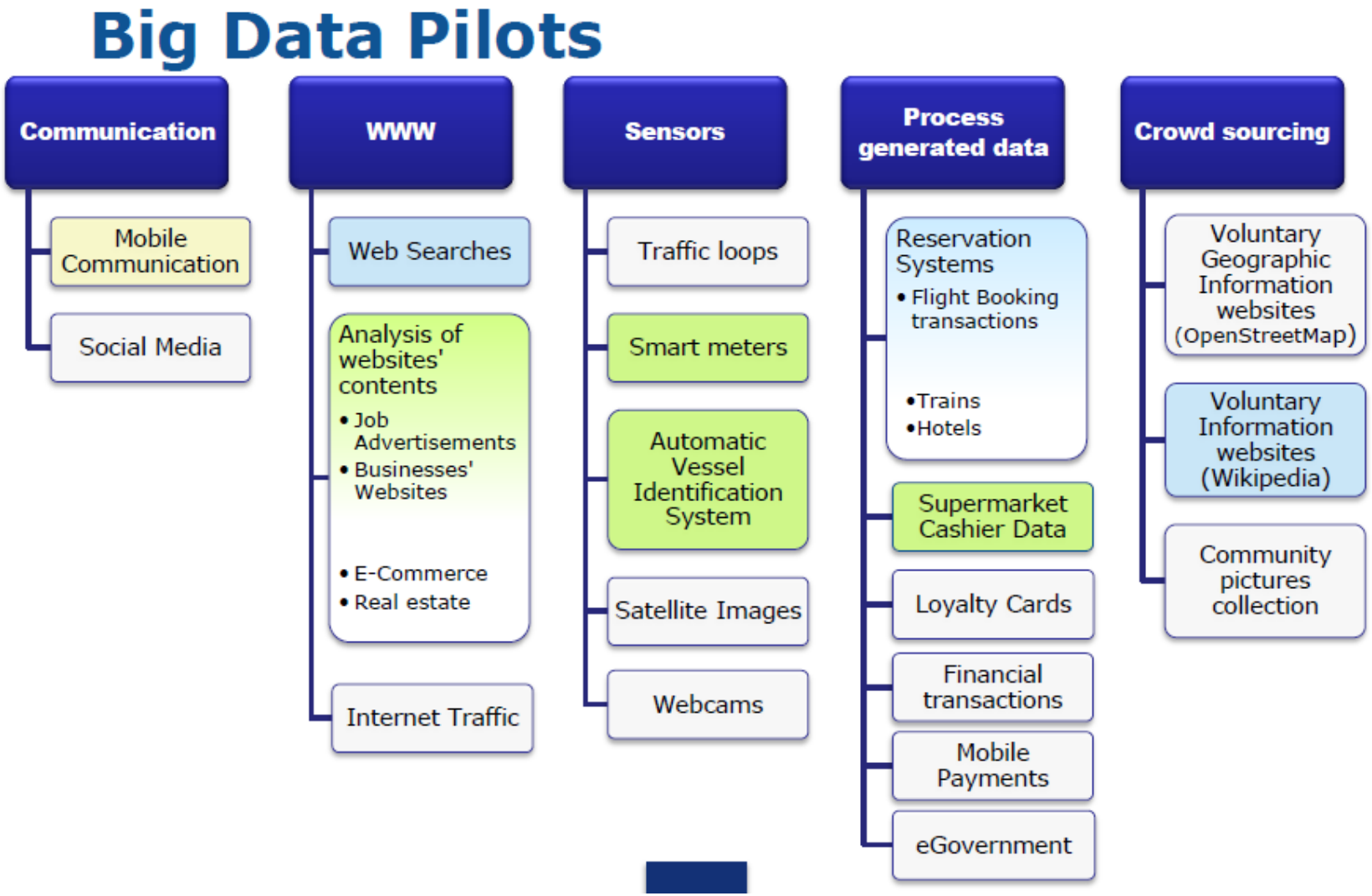
Day vs. night city population



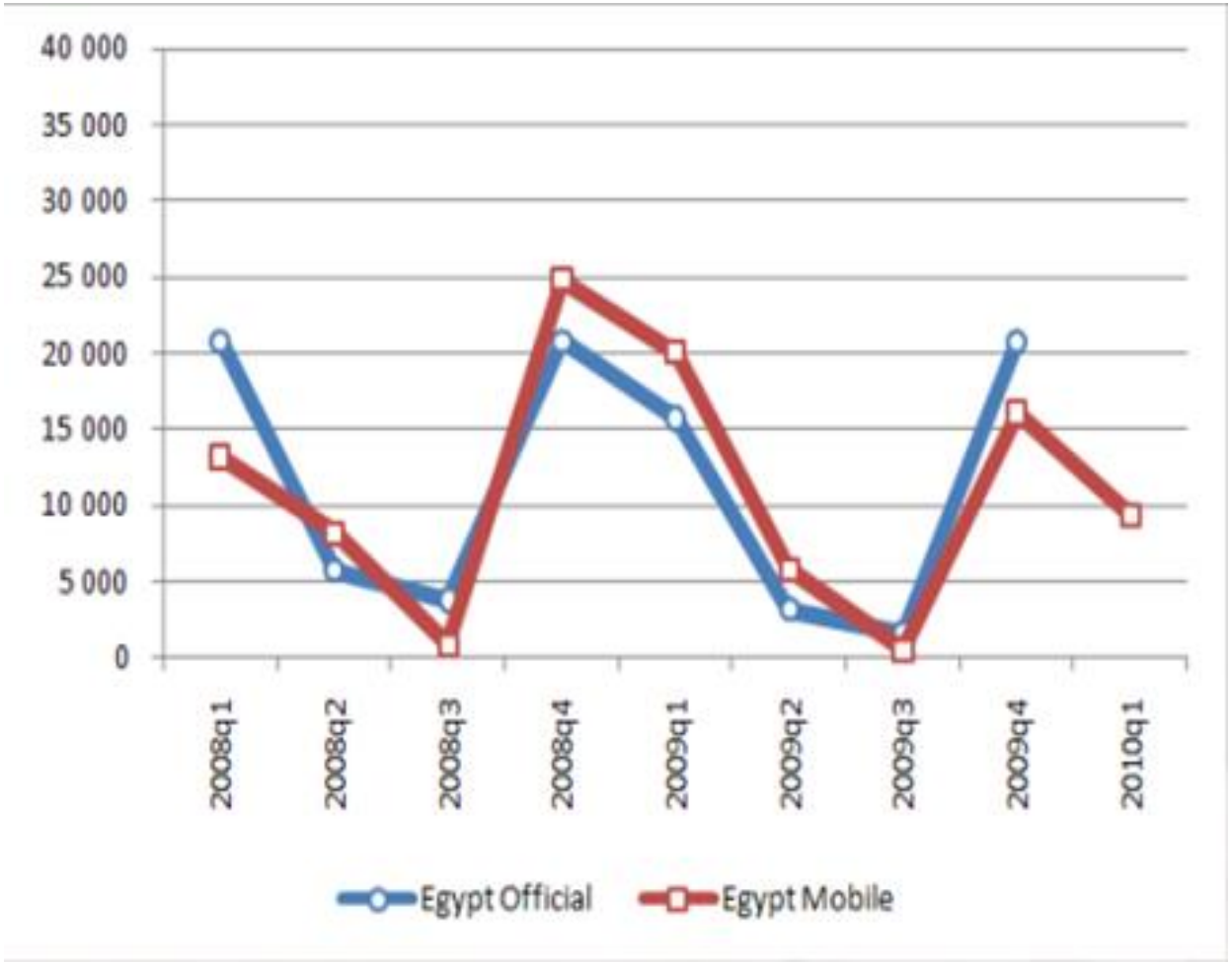
Source: Time magazine Nov. 26, 2007

The potential of Big Data

Big Data sources - EUROSTAT

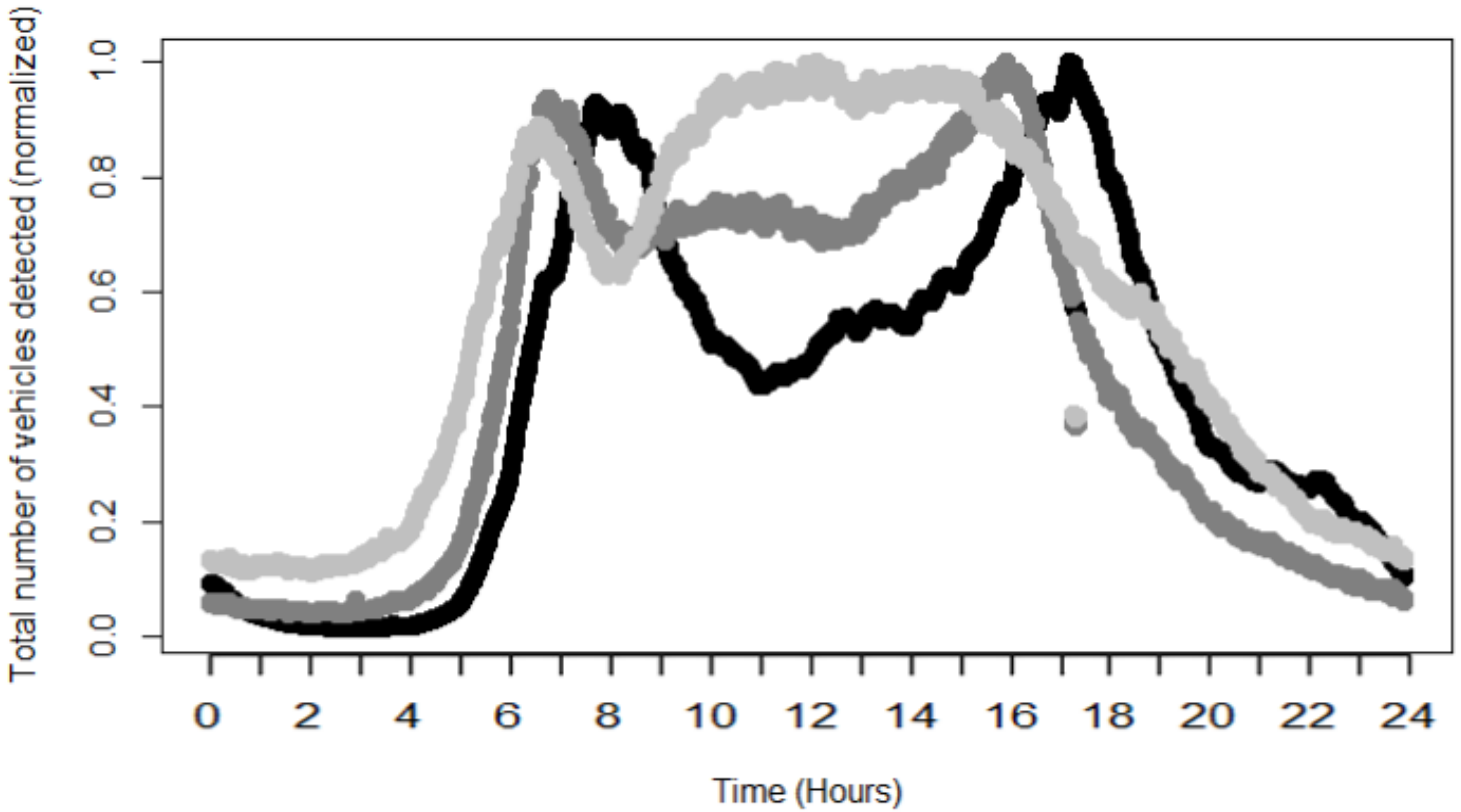


Estonia - tourist traffic to Egypt based on the roaming mobile phones



Source: Eurostat

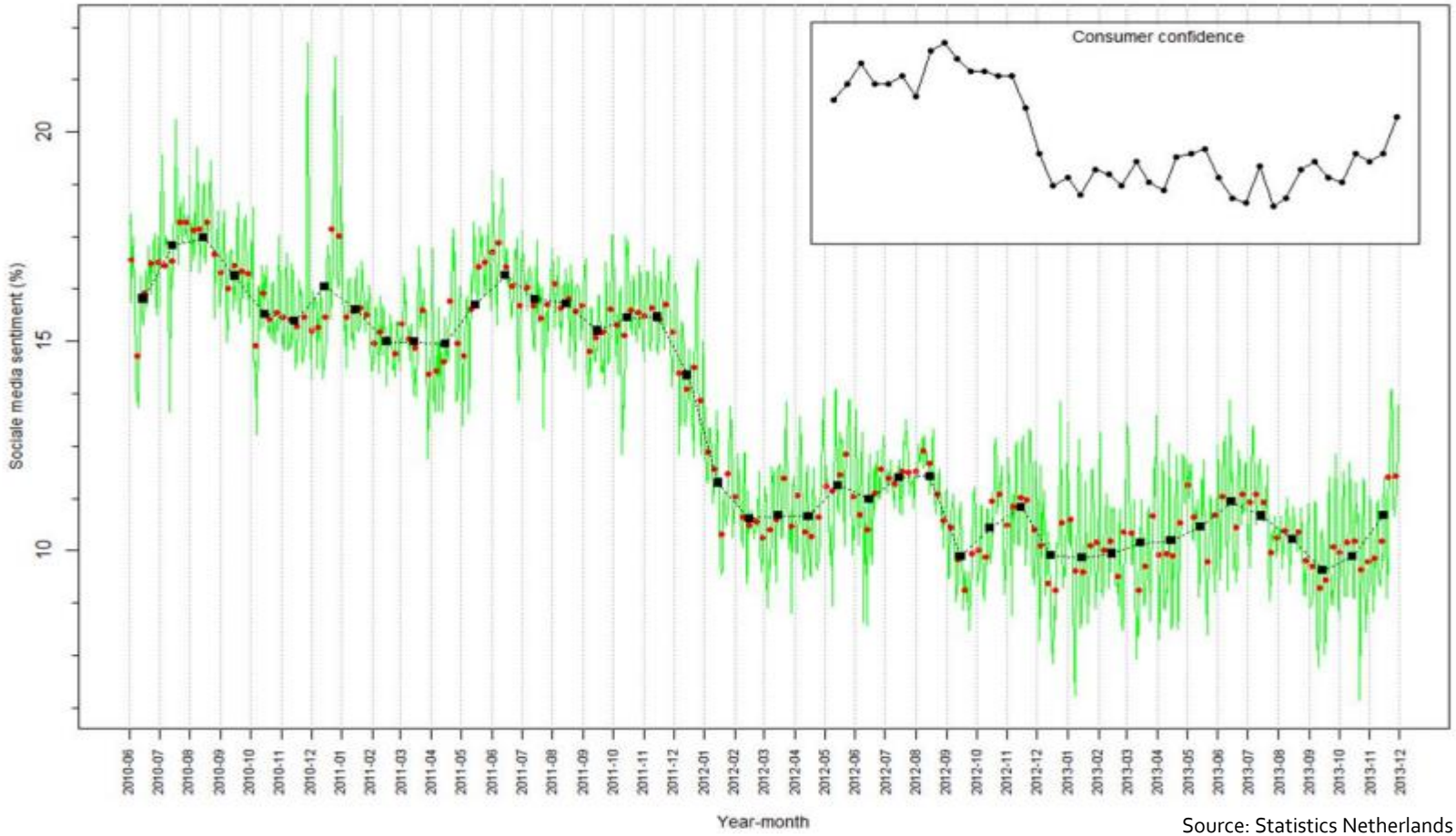
Sensors traffic



Normalized number of vehicles detected in three length categories on December 1st, 2011 after correcting for missing data. Small (<= 5.6 meter), medium-sized (>5.6 and <= 12.2 meter) and large vehicles (> 12.2 meter) are shown in black, dark grey and grey, respectively. Profiles are normalized to clearly reveal the differences in driving behaviour.

Social media

Study of sentiment and consumer confidence in the Netherlands on the basis of messages from social media.

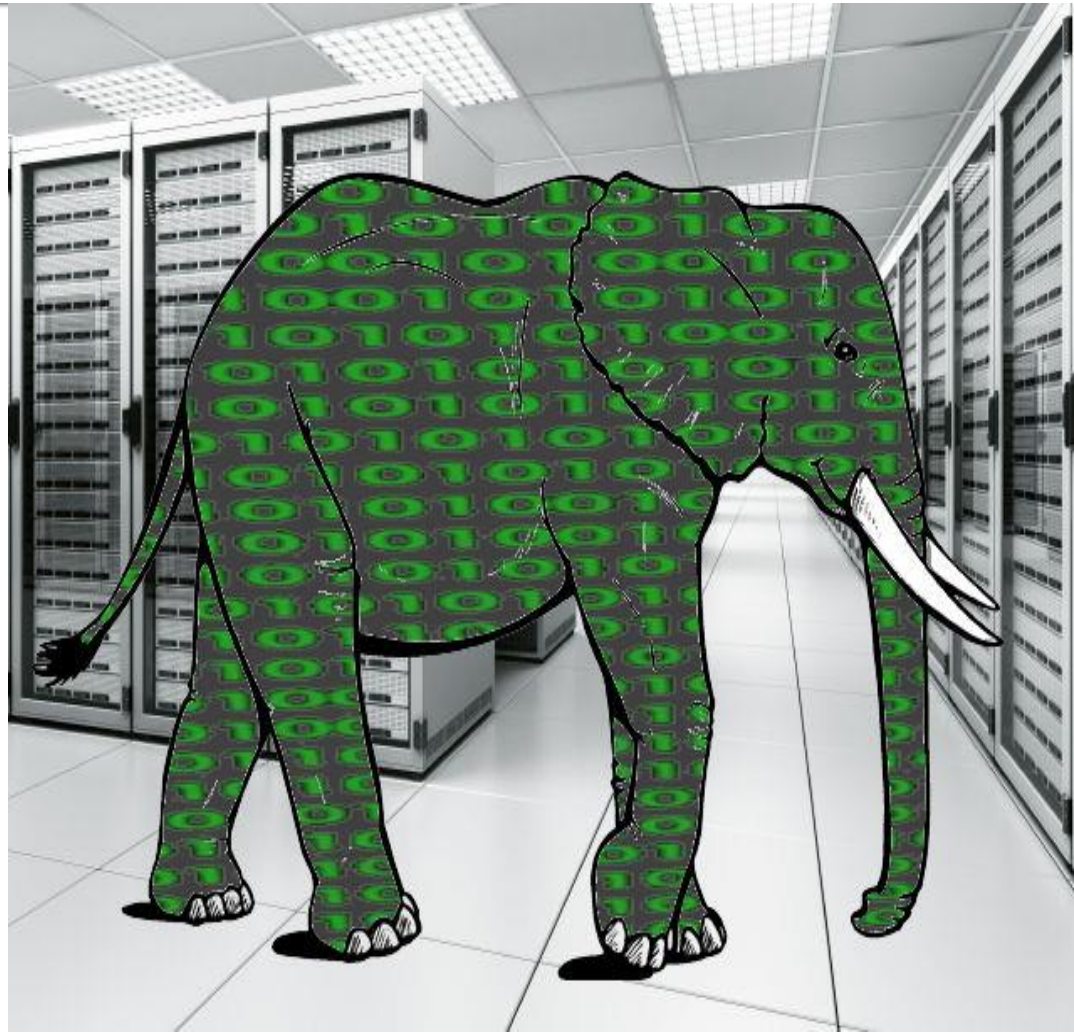


Source: Statistics Netherlands

Using Big Data in Official Statistics

The use of Big Data, not only would **complement the data available to the statistics**, but in the distant future would enable the **replacement of part of the currently existing conventional surveys.**

As a result, it would enable to **reduce the burden of citizens** in filling in questionnaires **and statistical interviewers** in collecting these questionnaires, while maintaining the existing high-quality data.



Big Data – Big Obstacles?

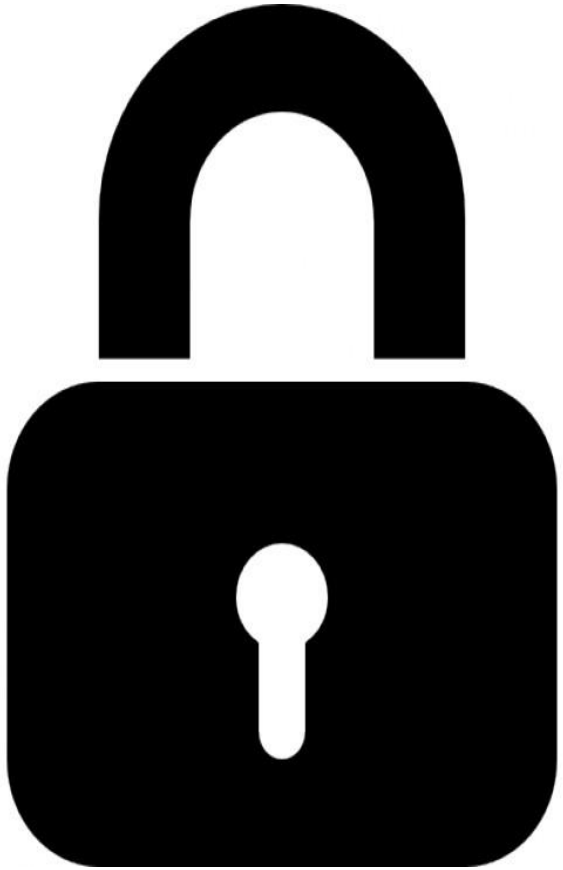
Law	Data safety	Privacy
Ethics	Competence	Methods
Technologies	Quality	Access to the data

Law

- The lack of sufficient legal basis for collection, analysis and storage of Big Data;
- Law can be (and is) very restrictive.

Data safety

- Physical protection ;
- Legal protection;
- Properly trained employees.



Privacy

Not only law

We ourselves are the best protection, controlling data, which we publish

Ethics

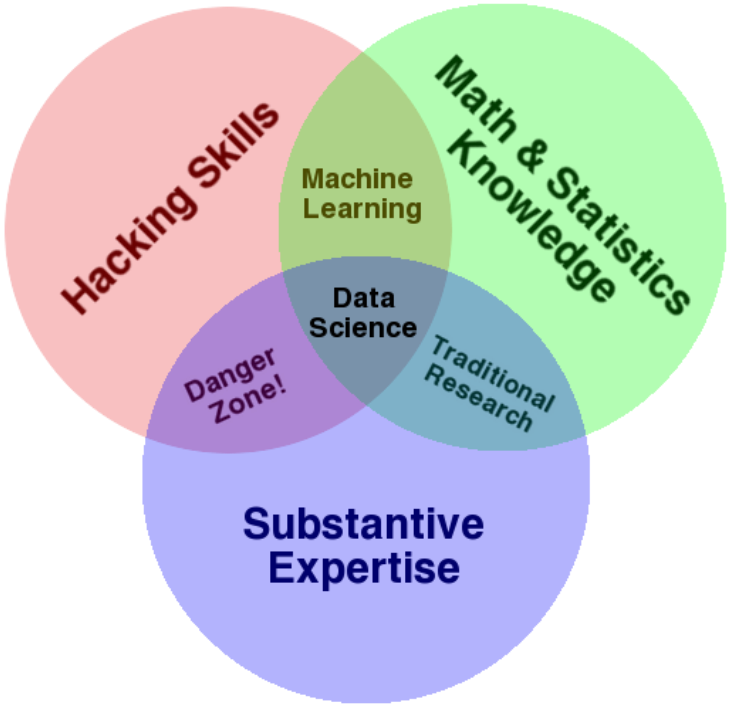
- There are many sources of data containing confidential information about individuals or households. The combination of different data sources, relating to the same object of observation, increases the sensitivity of the data
- Is the law keep up with the changes?
- Why do we need these data? Whether "the end justifies the means"?

- If the data, especially sensitive, are easily accessible whether should we use them?
- EUROSTAT: As a basis for the ethical review the following two documents should be used: UN Fundamental Principles of Official Statistics; European Statistics Code of Practice (CoP).

Competence

- Another problem arising both on the international and national level is the lack of experts - so-called data scientists.
- The demand is for multi-disciplinary group of specialists.

Data Scientist - competences



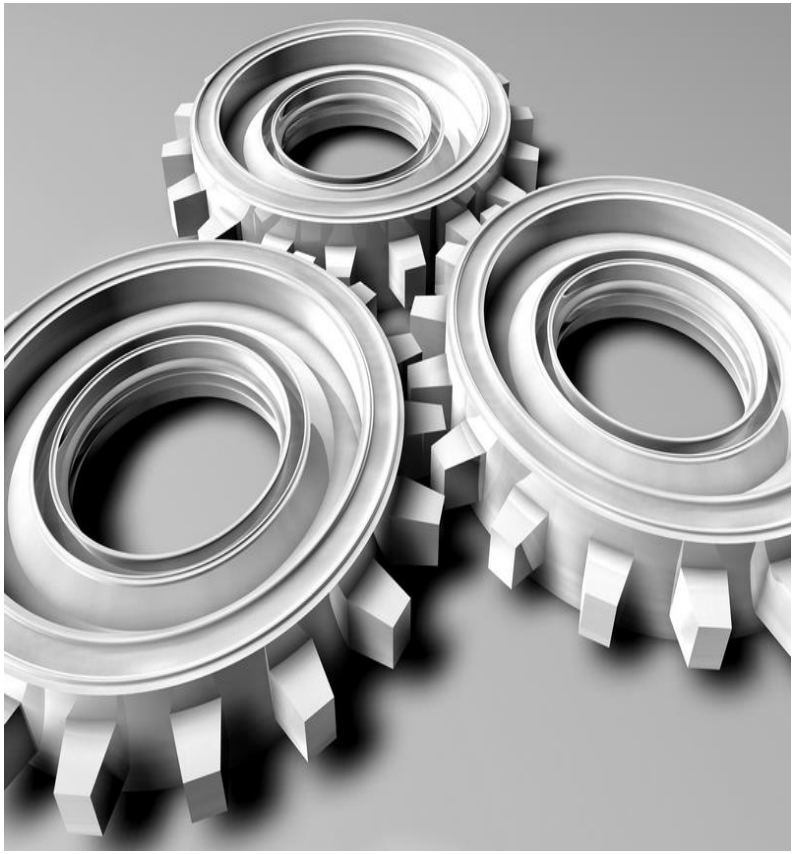
Specialized knowledge + Behavioural skills + ICT skills + Data analytical skills + Methodol ogist skills = Data Scientist

Methods

Due to the fact that Big Data is a relatively new phenomenon, many countries have not yet developed clear guidelines for collection, processing and sharing such data

Technologies

- Building infrastructure vs. the use of clouds or the environment of the source (real-time or near real-time);
- Developing methods of operation of this type of data;
- Competence building.



Quality

- 3,4, and even 5 „V”

volume, velocity, variety, value, veracity

- Analysis of unstructured volumes for achieving precise results (90% / 10%).
- Selection and matching of data sources to the needs.
- Ensuring quality:

clarity, relevance, accuracy, timeliness, coherence, transparency, accessibility, cost.

Access to the data

- The private sector usually uses its own data - easy access.
- The public sector needs data from the private sector - difficult access.
- The public sector has a well-defined and structured registration data

Conclusion

- **The official statistics cannot afford to rush and hasty action.**

It must uphold the quality and integrity of all data that is collected, in particular - data from censuses. It does not mean stopping the search for new paths.

- **Big Data is a good direction just a little too fresh.**

All initiatives will soon give the answer how great potential for statistics is Big Data.
After a more careful and exact analysis it will gradually turn out which, at first smaller and less complicated surveys, can be supplemented by Big Data.

Conclusion

- **With time, the circle of surveys using Big Data will be expanded.** There is a chance that in a few years we will be witnesses of the replacement of current collecting data methods by new opportunities.

- **Taking the above into consideration, the use of Big Data will not be possible in the 2020 census round but it will be possible a few years later.**


Thank You
For Your Attention