Q 2016
European Conference on Quality in Official Statistics

Assessment of risks in the use of big data sources for producing official statistics – Results of a stakeholder survey

20 - Big Data Oriented Systems
2 June 2016

Albrecht Wirthmann,
Martin Karlberg,
Bogomil Kovachev,
Fernando Reis,
Loredana Di Consiglio

Eurostat, Luxembourg

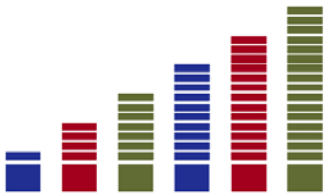
INE Instituto Nacional de Estadística
eurostat

Madrid, May 31 - June 3

Q 2016
European Conference on Quality in Official Statistics

Big Data & Risk Analysis

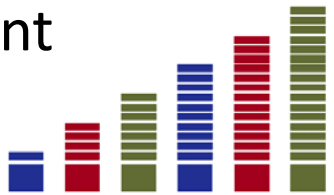
- Statistical Community has started to explore Big Data
  - ESSproject Big Data
- Identification and analysis of related risks is necessary
- What are inherent risks in using Big Data for Official Statistics?
  - During development phase
  - During production phase
  - Related to specific data sources
- Likelihood and Impact of events
- Taking the appropriate response
  - Ignore
  - Prevent
  - Mitigate
  - Terminate



Big Data Risks – Influencing Factors

Influencing factors

- New data / emerging market for data and derived services
  - Potential value of data
- Secondary data sources (design versus model based)
- From asking persons to analysis of observations
- Persons generating data (Social media, Wikipedia)
- Persons are data subjects
  - Sensitivity of data
- Third party data sources
  - Custodians are mostly private entities
- Competition between data custodians
- Volatility of data and data generation environment
- Unclear situation in terms of legal environment



Q 2016
European Conference on Quality in Official Statistics

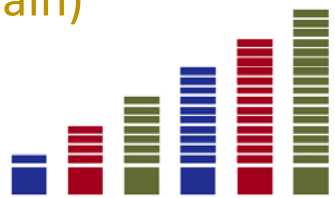
Big Data Risks

Development

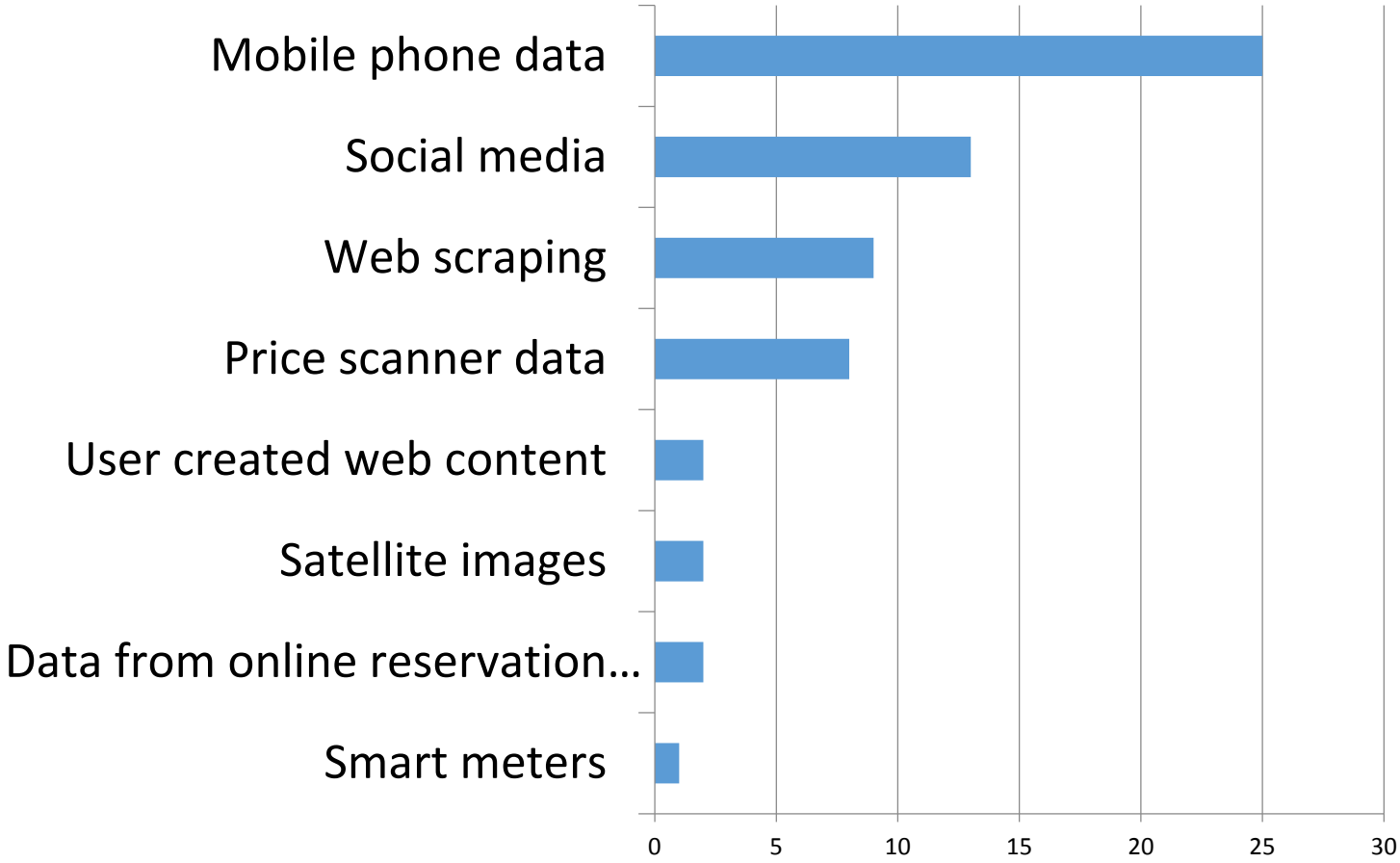
- Lack of access to data
- Non-compliance with relevant legislation
- Loss of credibility of statistical offices
- Lack of availability of experts

Production

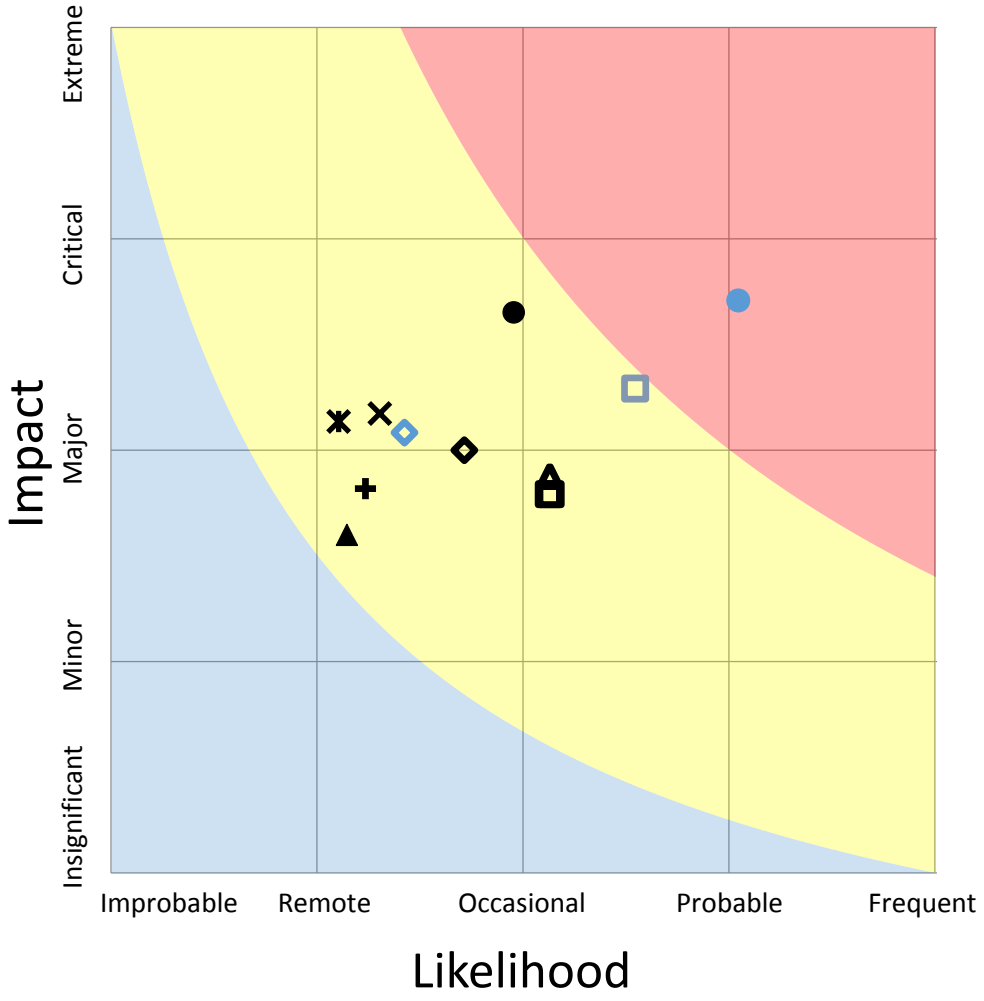
- Loss of access to data
- Changes in legal environment
- Data security breaches
- Data confidentiality breaches
- Data source manipulations
- Adverse public perception of big data usage by official statistics
- Loss of credibility of statistical offices
- Loss of experts (brain drain)



Data Sources - Responses

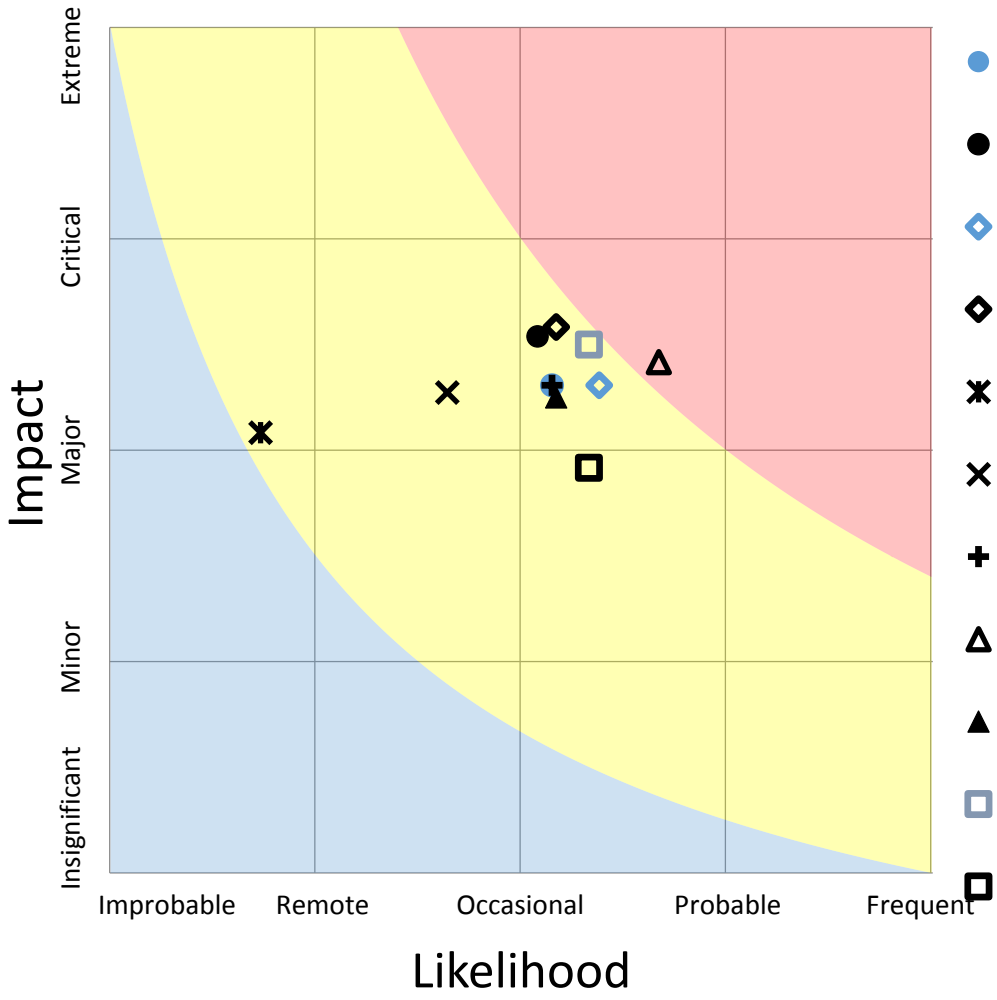


Mobile Phone Data



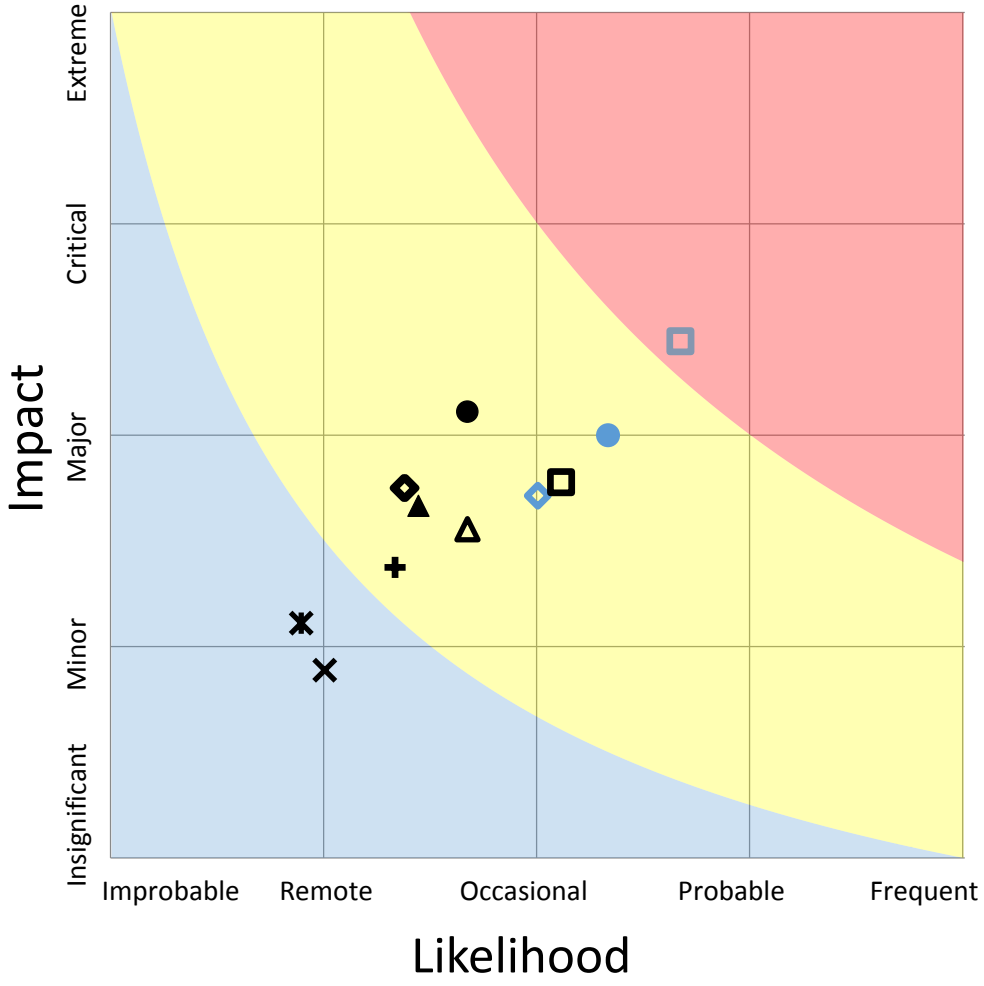
- **1 - Lack of access to data (Dev.)**
- **2 - Loss of access to data (Prod.)**
- 3 - Non-compliance with relevant legislation (Dev.)
- 4 - Neg. Changes in the legal environment (Prod.)
- 5 - Data security breaches (Prod.)
- 6 - Data confidentiality breaches (Prod.)
- 7 - Data source manipulation (Prod.)
- 8 - Adverse Public Perception (Prod.)
- 9 - Loss of credibility (Prod.)
- **10 - Lack of experts (Dev.)**
- 11 - Loss of experts (Prod.)

Social Media



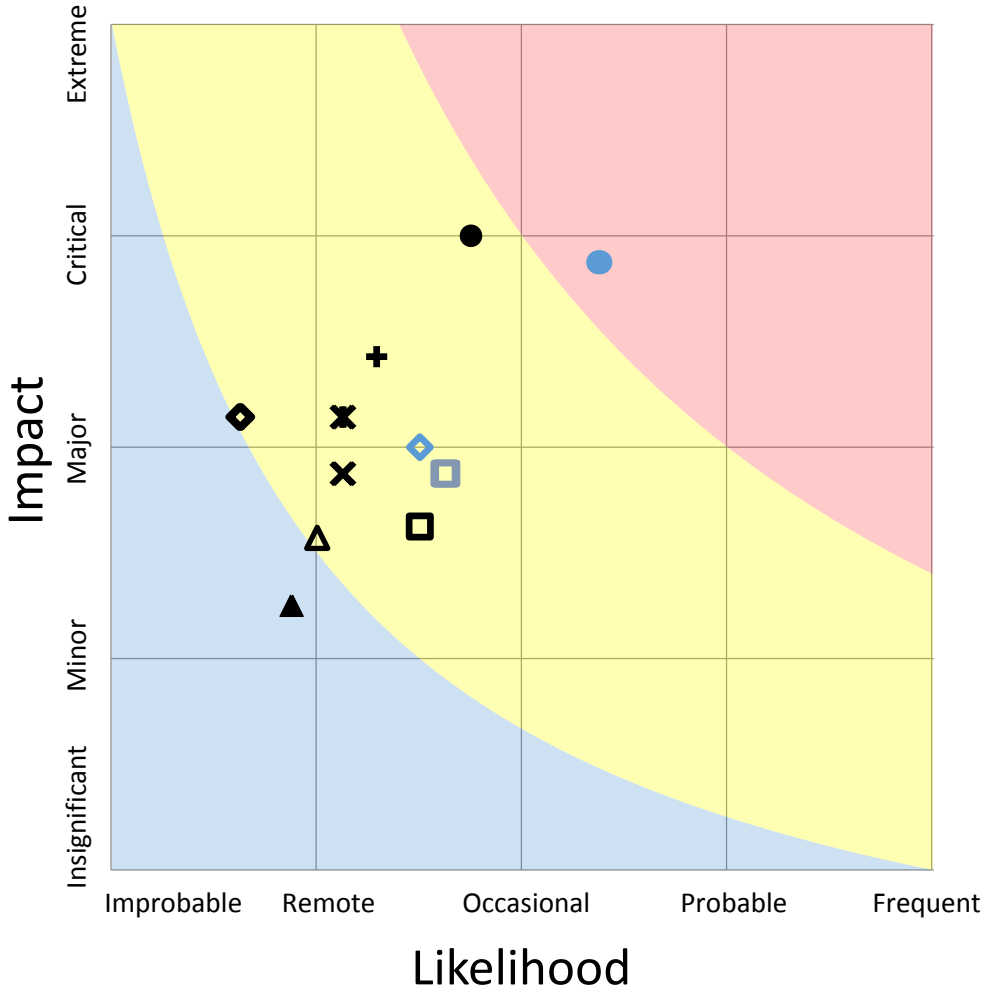
- 1 - Lack of access to data (Dev.)
- 2 - Loss of access to data (Prod.)
- **3 - Non-compliance with relevant legislation (Dev.)**
- **4 - Neg. Changes in the legal environment (Prod.)**
- 5 - Data security breaches (Prod.)
- 6 - Data confidentiality breaches (Prod.)
- **7 - Data source manipulation (Prod.)**
- **8 - Adverse Public Perception (Prod.)**
- **9 - Loss of credibility (Prod.)**
- **10 - Lack of experts (Dev.)**
- 11 - Loss of experts (Prod.)

Web Scraping



- **1 - Lack of access to data (Dev.)**
- **2 - Loss of access to data (Prod.)**
- **3 - Non-compliance with relevant legislation (Dev.)**
- 4 - Neg. Changes in the legal environment (Prod.)
- 5 - Data security breaches (Prod.)
- 6 - Data confidentiality breaches (Prod.)
- 7 - Data source manipulation (Prod.)
- 8 - Adverse Public Perception (Prod.)
- 9 - Loss of credibility (Prod.)
- **10 - Lack of experts (Dev.)**
- **11 - Loss of experts (Prod.)**

Prices – Scanner Data



- **1 - Lack of access to data (Dev.)**
- **2 - Loss of access to data (Prod.)**
- 3 - Non-compliance with relevant legislation (Dev.)
- 4 - Neg. Changes in the legal environment (Prod.)
- 5 - Data security breaches (Prod.)
- 6 - Data confidentiality breaches (Prod.)
- **7 - Data source manipulation (Prod.)**
- 8 - Adverse Public Perception (Prod.)
- 9 - Loss of credibility (Prod.)
- 10 - Lack of experts (Dev.)
- 11 - Loss of experts (Prod.)

Big Data Risks – Prevention / Mitigation

General

- **Risk Management**
- Prior impact analysis / crisis strategy
- **Analysis of terms and conditions**
- **Communication Strategy**
- **Involvement of Stakeholders**

Data Access

- **Legislative Initiatives**
- Constant monitoring and observation (pro-active approach)
- Partnerships (win-win situation, long term engagement)
- **Early involvement of data protection authorities**
- Diversification, use of alternative sources
- **International initiatives, harmonisation**
- Coordination among statistical community
- **Communication Strategy**

Big Data Risks – Prevention / Mitigation

Security, Confidentiality, Reputation

- **Involve Privacy Commissions, Data Protection Authorities**
- IT security measures,
- **Raise staff awareness**
- Improve methods for confidentiality / anonymisation
- Communication

-> Build trust

Skills

- Cooperation with academia
- Collaboration within statistical community
- Invest in training
- **Attractive salaries**
- **Sub-contracting**
- **Smart pooling of resources**

Q 2016
European Conference on Quality in Official Statistics

Big Data Additional Risks

Volatility of Sources

- Due to events outside control of Statistical Offices
  - E.g. Change of default settings in Operating Systems

IT Infrastructure

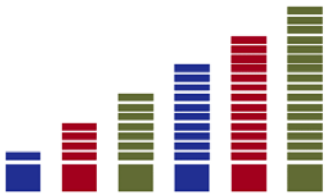
- Lack of adequate IT resources

Competition

- Statistical Offices as one player among other data collectors

Methodology

- How to measure precision of data
- Consistency and reliability
- Errors in matching concepts with data
- Changes in quality of products



Big Data Risks - Conclusions

More experience with Big Data as compared to the first publication

Most frequent replies for

- Mobile Phone Data
- Social Media
- Web Scraping
- Price scanner data

Replies show that proposed risks are relevant

Additional risks are put forward to integrate into framework

Assessment of risks differs by data source and by phase

Likelihood and Impact lower than is first publication

- Difference according to experience with data source?

Highest risks assigned to

- Access to data / Data Manipulation
- Legislation
- Lack of skills
- Public Opinion

-> concentrate actions for prevention and mitigation