Big data (phenomenon) and official statistics
The data deluge
What Happens in an Internet Minute?

1,572,877 GB of global IP data transferred

- 31,773 hours of music played
- 57,870 page views
- 3,3 Million pieces of content shared
- 347,222 Tweets
- 6.9 Million messages sent
- Netflix + Youtube = more than ½ of all traffic

And Future Growth is Staggering

By 2017, mobile traffic will have grown 13X in just 5 years

In 2017, there will be 3X more connected devices than people on Earth

All digital data created reached 4 zettabytes in 2013
The data deluge

- Increasing use of ICT
- Decreasing storage price
- Data ubiquity: text, sound, images, video
- Organic data / exhaust data / digital footprint

The data deluge

• What type of data is big data?
  • **Volume** (large N and large P)
  • **Velocity** (data streaming and real-time statistics)
  • **Variety**
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
"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."
Analytics

- How to deal with exhaust data?
  - Dealt by machine learning / predictive analytics
- Massive datasets
  - Foster machine learning
- Data science: a new discipline?
  - Signal processing (audio, image, video)
  - Natural Language Processing (NLP)
  - Network data
  - Distributed computing
  - Multiple inference
  - Over-fitting
Analytics

- Visualisation
- Data-driven applications
  - The end of theory
- Predictive analytics
  - Ex: Google translate, voice recognition, suggestions systems, health applications
- New types of data products
  - Official stat: chances of getting a new job
Predicting Personality Using Mobile Phone-Based Metrics

1. Daily distance traveled
2. Places entropy
3. AR-ψ4

1. Entropy of contacts (text)
2. Entropy of contacts (c&c)
3. Variance of inter-event time (text)

1. Variance of inter-event time (call)
2. Text inter-time average
3. Average inter-time (call)

1. Entropy of contacts (text)
2. AR-ψ1
3. Variance of inter-event time (call)

1. Average inter-event time (text)
2. Percent initiated (text)
3. Average inter-event time (call & text)

---
Multidimensional Poverty Index (Lighter colour indicates higher poverty)

Poverty map estimated based on mobile phone data

Poverty map in finer granularity estimated based on mobile phone data

Population statistics

Mobile phone frequent locations

Mobile phone commute map

Population Mapping Using Mobile Phone Data


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

- Monetisation of data: Data is the new oil
- Data as a new factor of production (competitive differentiating factor for businesses)
- A threat to official statistics? (ex: Argentina)
- Data ecosystem
- The cases of Google and Facebook
What does big data mean for official statistics?

• Change of paradigm
  • From: finite population sampling methodology
  • To: additional statistical modelling and machine learning
  • from designers of data collection processes to designers of statistical products

• Privacy
  • Use of digital footprint
  • Data subject lack of control of data
  • High data detail and insight from analytics
Multisource statistics and multipurpose sources

- Tourism Statistics
- Mobile Phone Data
- Population Statistics
- Satellite Images
- Traffic Statistics
- Migration Statistics
- Mobile phone data
- Population Statistics
- Smart Meters
- VGI websites
- Multisource statistics and multipurpose sources
Scheveningen Memorandum on Big Data

- Examine the **potential** of Big Data sources for official statistics
- Official Statistics Big Data **strategy** as part of wider government strategy
- Address **privacy** and **data protection**
- **Collaboration** at European and global level
- Address need for **skills**
- **Partnerships** between different stakeholders (government, academics, private sector)
- Developments in **Methodology, quality** assessment and **IT**
- Adopt **action plan and roadmap** for the European Statistical System
ESS big data action plan

- Policy
- Quality
- Skills
- Experience sharing
- Legislation
- IT Infrastructures
- Methods
- Ethics / Communication
- Pilots
Eurostat big data pilots

• Contracts
  • Feasibility study on the use of mobile phone data for tourism statistics
  • Internet as a data source for information society statistics
  • Accreditation of big data sources

• Internal projects
  • Mobile phone for urban statistics
  • Web evidence for nowcasting
What do future official statisticians need to know in the big data age?

• Traditional: survey methodology and design based inference
  • Use of auxiliary information
  • Small area estimation
  • Confirmatory surveys
• Model based inference
  • Machine learning
  • Nowcasting
  • Predictive analytics
• Signal processing
### ESTP courses related to big data

#### Introduction to big data and its tools
- Big data and the several digital traces people leave behind them;
- Overview of big data sources: sensors and the IoT, process-mediated data; human-sourced data;
- The implications of big data for official statistics;
- Big data initiatives in official statistics at international level;
- Privacy and personal data protection;
- Examples of the use of big data for producing statistics (relevant for official statistics);
- Methodological challenges of big data for official statistics, e.g. over-fitting, multiple inference, and model-based inference.

#### Hands-on immersion on big data tools
- Design based to model based to algorithm based
- Statistical learning
  - model selection and regularization methods (ridge and lasso);
  - tree-based methods, random forests and boosting; support-vector machines.
- Posets
- Network analysis and web analytics
- Graph database and advanced data visualisation
- Geo-spatial statistics
- Perceived Price index
- Embedded experimental evaluation
- Multimode data collection

#### Big data courses
- Can a statistician become a data scientist?
  - The use of R in official statistics:
    - model based estimates
    - Time-series econometrics
- Statistical learning and model based estimates for official statistics
  - Social network analysis and web analytics
  - Methodology courses
    - Big data and the several digital traces people leave behind them;
    - Overview of big data sources: sensors and the IoT, process-mediated data; human-sourced data;
    - The implications of big data for official statistics;
    - Big data initiatives in official statistics at international level;
    - Privacy and personal data protection;
    - Examples of the use of big data for producing statistics (relevant for official statistics);
    - Methodological challenges of big data for official statistics, e.g. over-fitting, multiple inference, and model-based inference.
- Visualisation and its importance in the analysis of big data;
- Data science and its role in big data analytics;
  - Overview of big data tools, e.g. distributed computing:
    - Hadoop;
    - Map Reduce;
    - Pig and Hive;
    - Spark;
    - NoSQL databases;
    - RHadoop;
  - Web scrapping
    - Content and sentiment analysis on social media
    - Text mining
    - Mobile phone operators data;
    - Traffic loops data;
    - Satellite images;
    - Vessels and planes identification systems;
- Basic programming with R
  - Reading XML data
  - Reading SDMX format
  - R vs. RSTUDIO
  - Screening tools for database
  - Modelling and analytical capabilities of R
  - Generating Graphical Tables for Websites and Documents with R
  - R packages: practical applications for daily production
  - GitHub
- Linear and polynomial regression,
  - logistic regression and linear discriminant analysis;
  - cross-validation and the bootstrap,
  - model selection and regularization methods (ridge and lasso);
  - nonlinear models, splines and generalized additive models;
  - tree-based methods, random forests and boosting; support-vector machines
  - Some unsupervised learning methods
    - principal components
    - clustering (k-means and hierarchical)
  - Computing with R
  - Features of time series: trends, cycles, non-constant volatility
  - Basic methods to de-trend series (moving averages, smoothing, first differences, linear, segmented trend)
  - Stochastic processes
  - The Box-Jenkins approach (ARIMA) for non-seasonal models
  - Outliers: identification and treatment
  - Model misspecification checking
  - Detecting the type of non-stationarity (unit root tests)
  - Introduction to ARCH/GARCH
  - Time Series Regression
  - Uncertainty and forecasting

Thank you for your attention

Fernando Reis

Eurostat Task Force on Big Data

feraldo.reis@ec.europa.eu

https://github.com/reisfe/

https://twitter.com/reisfe/

https://linkedin.com/in/reisfe/