Module 3 – The use of big data in official statistics

Summer school in public auditing and accountability
Data mining and analytics: what implications for auditing?
23-27 July 2018

Fernando Reis

Eurostat Task Force on Big Data

Summary

- Context about the use of big data in official statistics
- The big data phenomenon
- A statistical definition of big data
- Methodological challenges of big data
- Examples of applications

Initiatives at European and global level, the big data action plan and the strategy;

CONTEXT ABOUT THE USE OF BIG DATA IN OFFICIAL STATISTICS

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

Big data strategy

- **Start with concrete pilots**
- **3 time-frames**
  - Short-term
  - Medium-term
  - Long-term
- **Review the roadmap**

Big Data Action Plan and Roadmap @ a glance

Governance

Policy	Quality	Skills
Experience sharing	Legislation	IT Infrastructures
Methods	Ethics / Communication	Big data sources

Pilots

Implementation tools

- Study on legal, communication and skills issues related to the use of big data
- Consortion of National Statistical Institutes (ESSnet)
- Internal Eurostat activities

Big Data Study

- Ethical review and guidelines
- Communication Strategy
- Legal Review
- Development of a training strategy to bridge the Big Data skills gap in European official statistics
- Workshop on ESS Big Data, 13-14 Oct 2016
  - http://ec.europa.eu/eurostat/cros/content/ess-big-data-workshop-2016_en

- Start 1 January 2016
- End 1 December 2017

European Commission

Big Data Study – Ethical issues

- Ethical implications of the use of Big Data for production of official statistics
  - the data is held by the private sector
  - access to the data sets
  - professional independence
  - Transparency and quality (inputs, processing, outputs)
  - Impartiality
  - Privacy / confidentiality
  - reputation of official statistics
- Review of Code of Practice
- Ethical guidelines

Big Data Study – Legal aspects

Legal Review

– A comprehensive survey of current and upcoming legislation

- **within Member States and at EU level** that could have an impact on use of big data for statistical purposes

- (1) to identify cases where using special sources for specific statistical purposes or for official statistics in general would not be consistent with relevant legislation and
- (2) identify clauses that would allow access and use of special sources explicitly or implicitly.

•Report on legal review covering basic statistical laws and framework legislations. - Feb 2017

•Report on legal review covering other legislations. - Oct 2017

European Commission

Big Data Study – Legal aspects

Legal Review

- Personal data
- Copyright
- Databases
- Confidentiality

First results

- Very few legal obstacles in using big data by NSIs
  - Internet data
  - Limited retention period
- Data protection legislation
  - Exceptions for secondary use of personal data for statistical purposes
  - Exemption from duty to inform subjects
- Sector legislation
  - no evidence for restrictions in use

European Commission

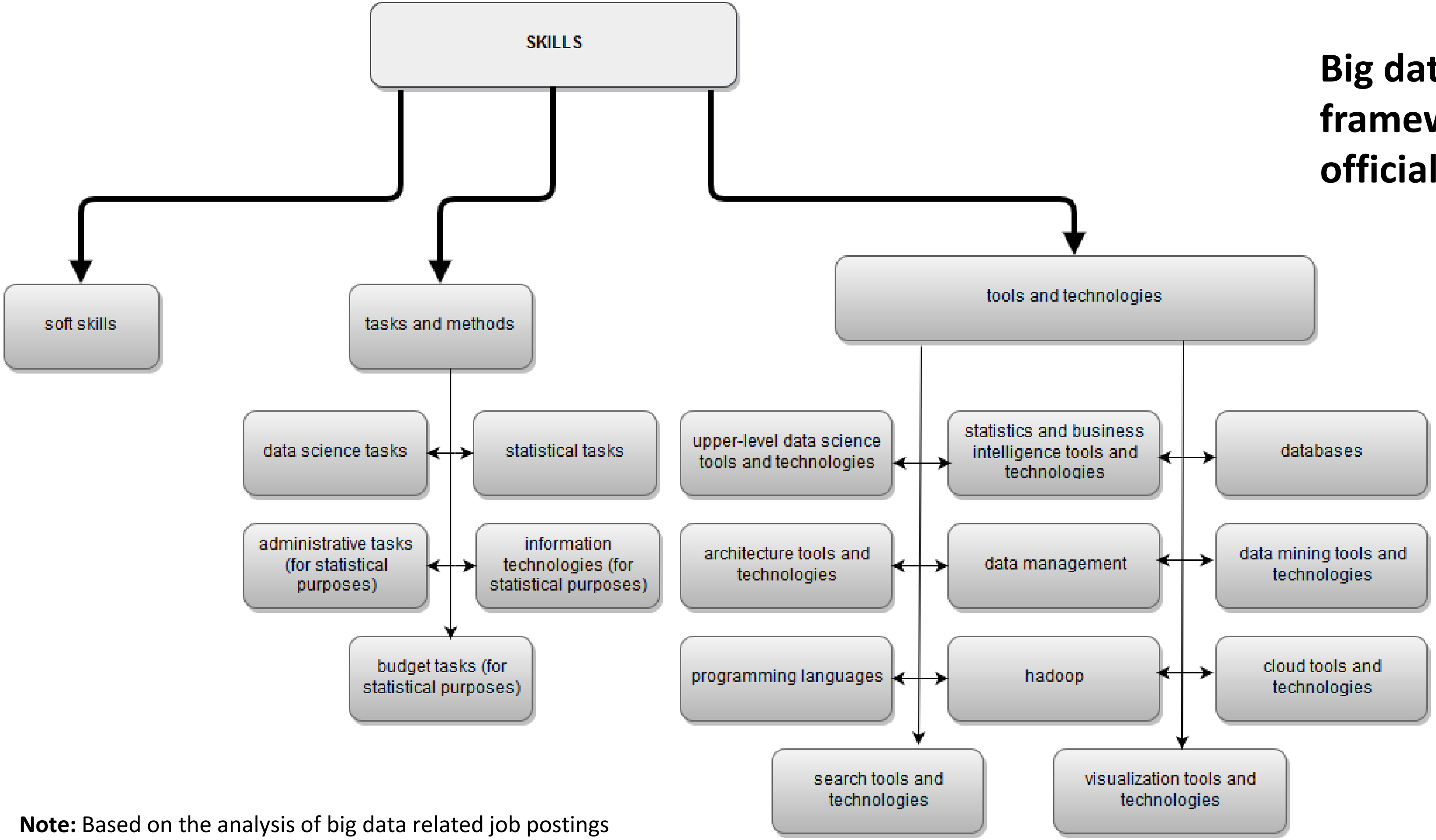
Big Data Sudy - Skills

Development of a training strategy to bridge the Big Data skills gap in European official statistics

- Identification of skills required for the use of big data sources (Feb 2017)
  - Skills framework
- Inventory of existing skills in Eurostat and in the NSIs in Europe (Jun 2017)
  - Questionnaire to NSIs and Eurostat
- Analysis of the big data training needs (Jul 2017)
- To define the training objectives and content (Aug 2017)
  - Competency-Based Education approach
- Develop a training provision strategy to bridge the skill gap (Oct 2017)
  - e.g. ESTP courses

European Commission

Big data skills framework for official statistics



Note: Based on the analysis of big data related job postings

Source: Eurostat (2018) Development of a training strategy to bridge the big data skills gap in European official statistics

Big data skills for official statistics

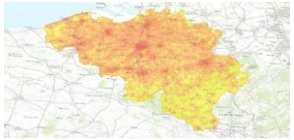
SOFT SKILLS	Communication,	CLOUD TECHNOLOGIES	Cloud computing
	Innovation and contextual awareness,	HADOOP	Hadoop
	Teamwork,	UPPER-LEVEL DATA SCIENCE TOOLS AND TECHNOLOGIES	Machine learning,
	Creative problem solving,		Databases,
	Negotiation,		Understanding algorithms,
	Leadership,		Data mining,
	Delivery of results,		Deep learning,
	Information privacy,		Artificial intelligence,
	Coordination		Natural language processing,
STATISTICAL AND DATA SCIENCE TASKS	Nowcasting and projections,		Stream processing and analysis,
	Nonresponse adjustment and weighting,		IoT (Internet of Things),
	Analysis of aggregated data,		Multimedia analysis,
	Multivariate analysis,		Web technologies (Web scrapping)
	Time series and seasonal adjustment,	STATISTICS AND BUSINESS INTELLIGENCE	SAS,
	Quality assessment,		Apache Spark,
	Data visualization,		SPSS
	Data resource management,	DATA MANAGEMENT	Apache Hive,
	Setting up data warehouses,		Apache HBase,
	Data storage,		Apache Sqoop,
	Data processing,		Apache Pig,
	Data conversion		Cloudera Impala
ADMINISTRATIVE TASKS (FOR STATISTICAL PURPOSES)	Quality assurance and compliance,	DATABASES	Passively parallel-processing databases (MPP),
	Project management		DBMS,
INFORMATION TECHNOLOGIES TASKS (FOR STATISTICAL PURPOSES)	System Architecture,		SQL,
	Hardware and infrastructure,		NoSQL,
	Developing software,		MongoDB
	Systems and software maintenance	SEARCH TECHNOLOGIES	Search based applications
PROGRAMMING LANGUAGES	R,	VISUALIZATION TECHNOLOGIES	D3,
	Python,		Shiny,
	Scala		Bokeh
ARCHITECTURE TOOLS AND TECHNOLOGIES	Distributed Parallel architecture,		
	Distributed computing,		
	Distributed filesystems		

Note: Based on survey of ESS big data experts
Source: Eurostat (2018) Development of a training strategy to bridge the big data skills gap in European official statistics

Internal Eurostat activities

- 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
  - Wikipedia use
  - Mobile phone for urban statistics
  - Web evidence for nowcasting

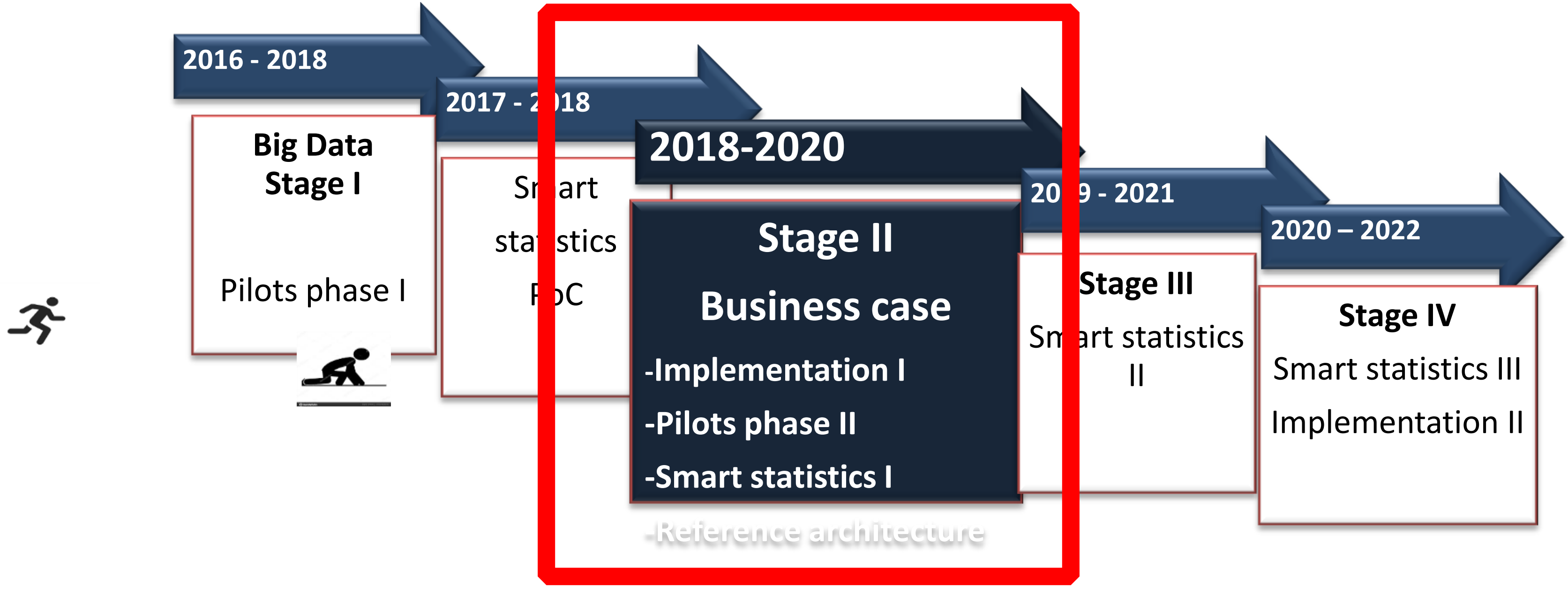
ESSnet Big Data (2016-2018)



- Online job vacancies (web scraping)
- Enterprise characteristics (web scraping)
- Energy consumption via Smart meters
- Maritime transport via AIS data
- Mobile network data
- Early estimates for key indicators
- Integration of multiple sources for different statistical domains (tourism, population, agriculture
- Methodology, Quality, IT
- **Final workshop and reports available at ESSnet wiki**

European Commission

Smart Statistics and Big Data



Business case 2018 - 2020: Contents 1/3

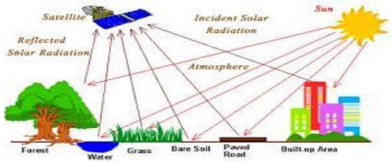
- **First implementation phase of <u>successful pilots</u> of stage I, limited to 3 countries, setting full-fledged implementation requirements**

- Webscraping job vacancies
- Webscraping enterprise characteristics
- Smart meters
- Automatic vessel identification system

Business case 2018 - 2020: Contents 2/3

New pilot projects – exploring data sources beyond stage I

- Use of financial transactions data
- Remote sensing



- Online platforms such as social media and sharing economy platforms

- Mobile network operator data (continued)

- Innovative sources and methods for tourism statistics

Business case 2018 - 2020: Contents 3/3

- **Extend the work on <u>smart statistics (trusted smart statistics)</u>**
  - Access to data sources, standards, …
  - Support initiatives such as TUS, HBS

- Use of citizen science data for individuals' well-being (wearables, smart devices, … **potential extended to tourism, TUS and HBS**)

- Citizen science data and smart cities
- Smart cities and connected vehicles

- Smart farming

What next?

TRUSTED

SMART

STATISTICS

The new model

- Push computation out
- Use data without sharing data
- Privacy by design
- Embed statistics in smart systems
- Trust by design

The new model

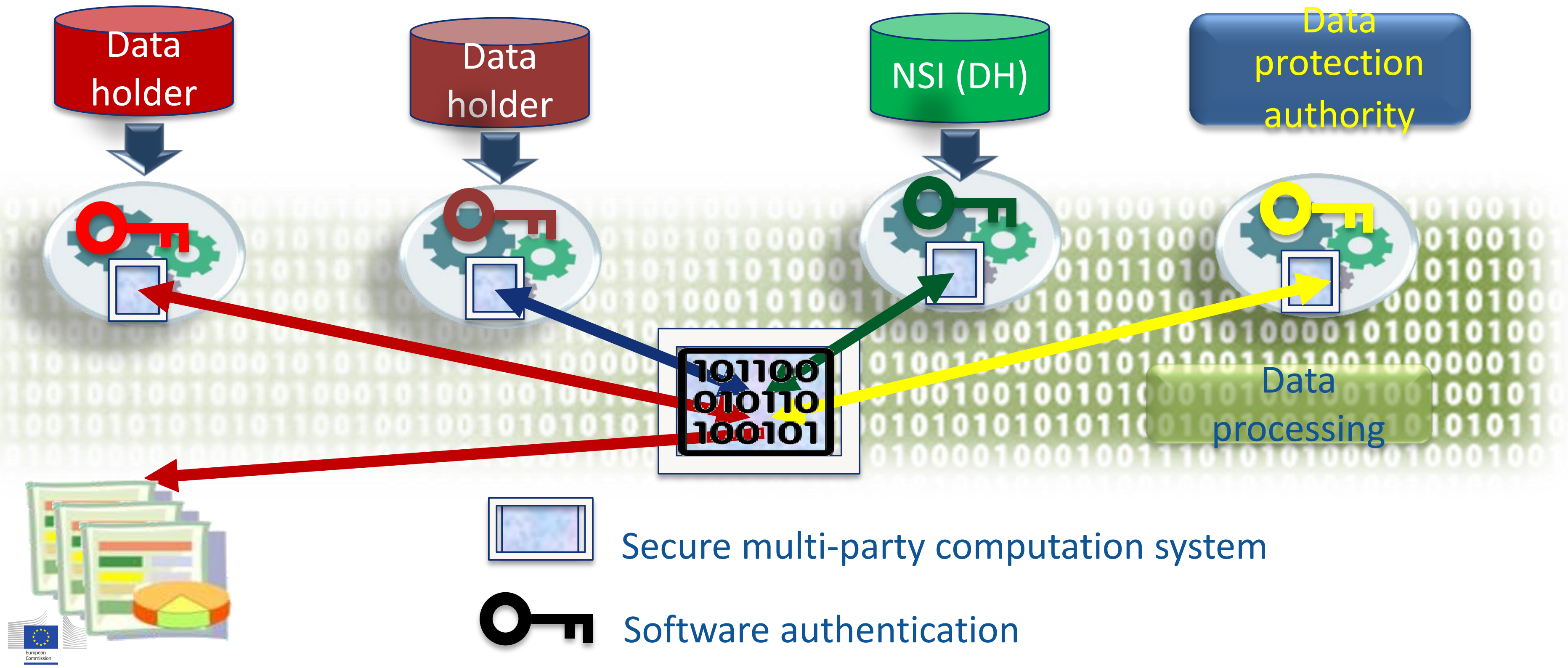
More input data

More analytics and AI

More statistics produced

Data use without sharing

Secure multiparty computation – setup example



Trust in smart statistics

Trust in Input Data

- Data is veracious
- Data provision is continuous, stable

Trust in Processing

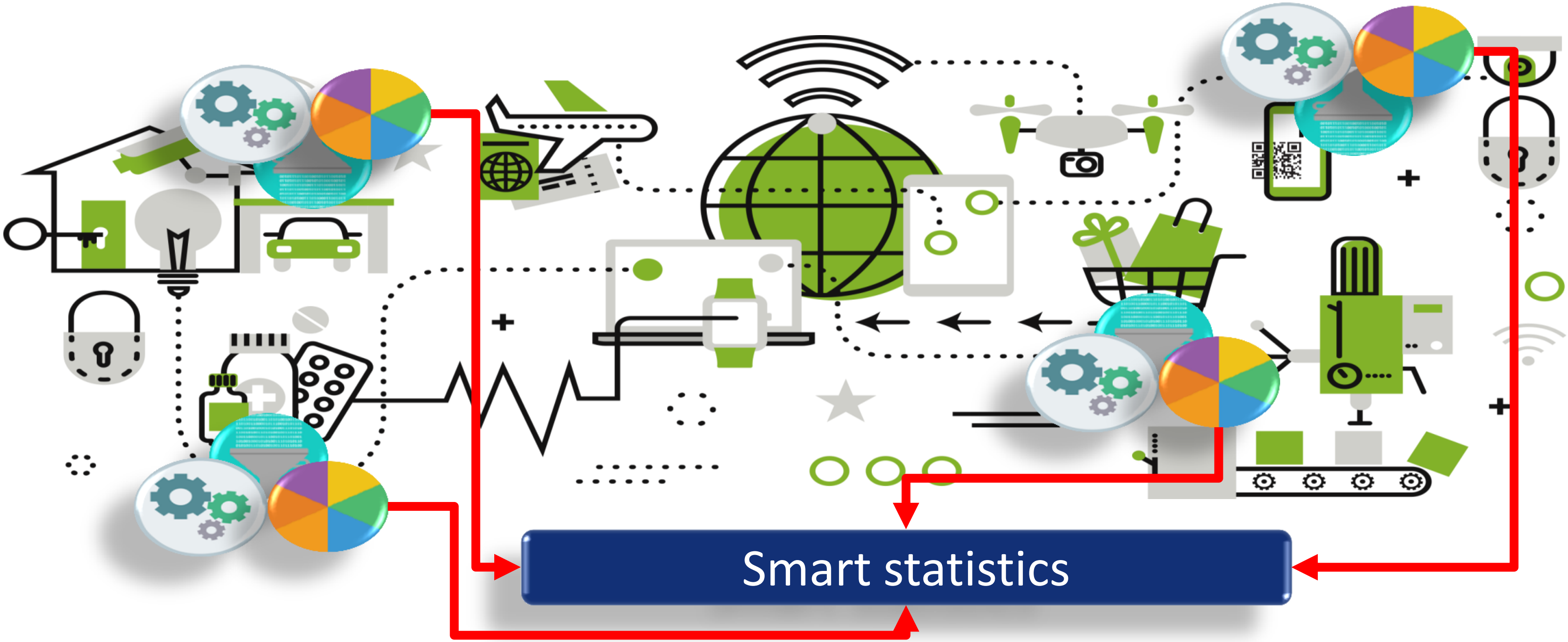
Data are processed only
- for agreed purpose
- by agreed method
- by authenticated software
- privacy & confidentiality guaranteed

Trust in Output

- Quality information provided
- Information on methodology
- Statistics corresponds to perceived situation

European Commission

Smart systems

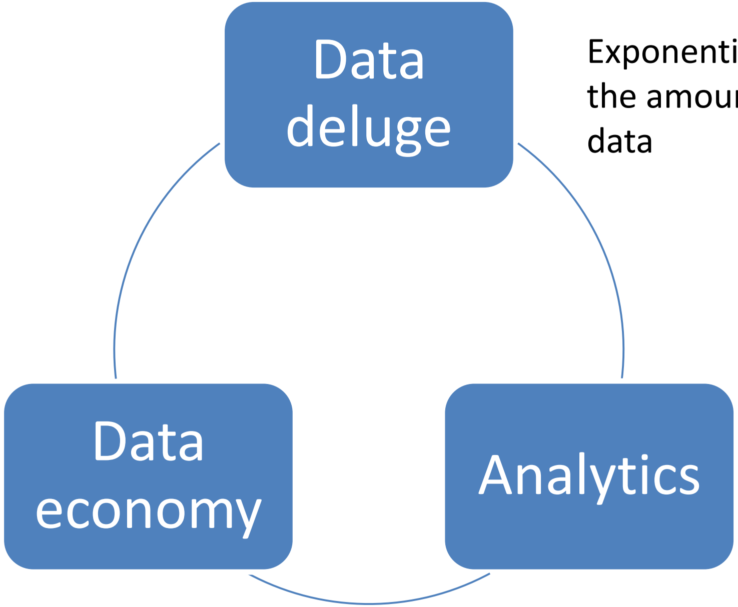


Smart statistics

THE BIG DATA PHENOMENON

The big data phenomenon

Exponential increase in the amount of available data



Multitude of data collectors and applications producers

Purely data driven applications

The big data phenomenon

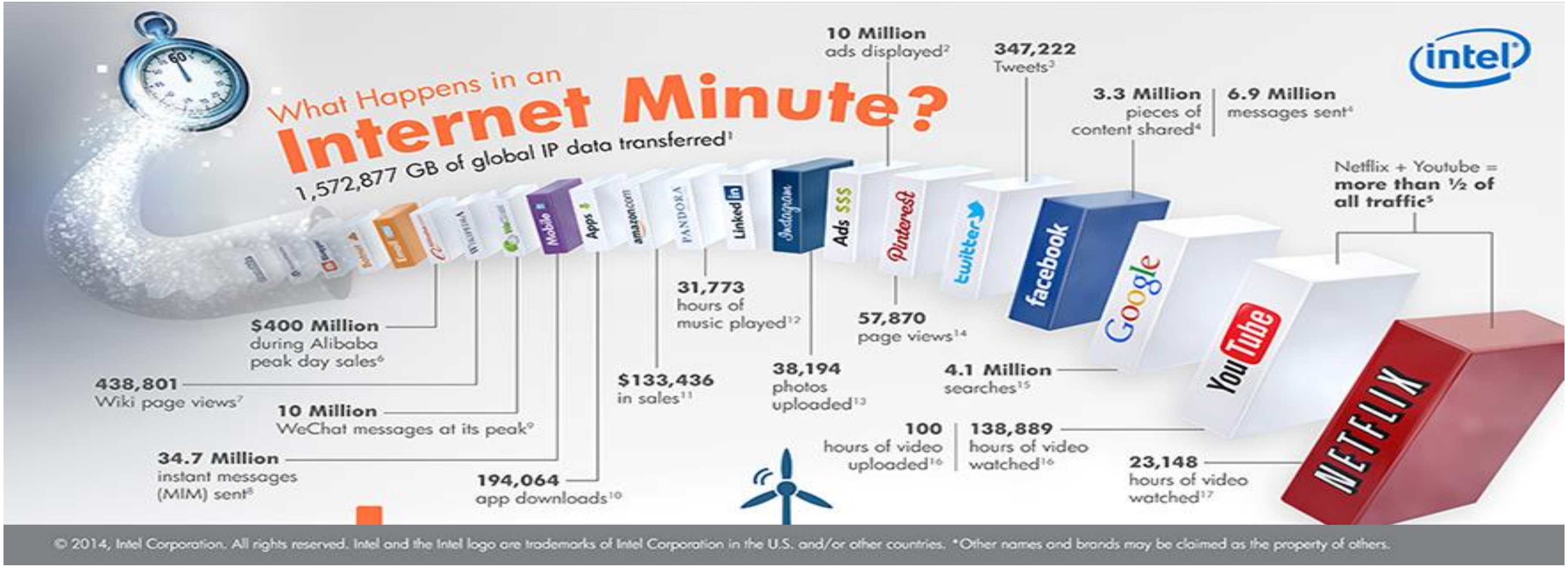
1. The data deluge

Proclamation of pope Benedict **2005**

Proclamation of pope Francis
2013

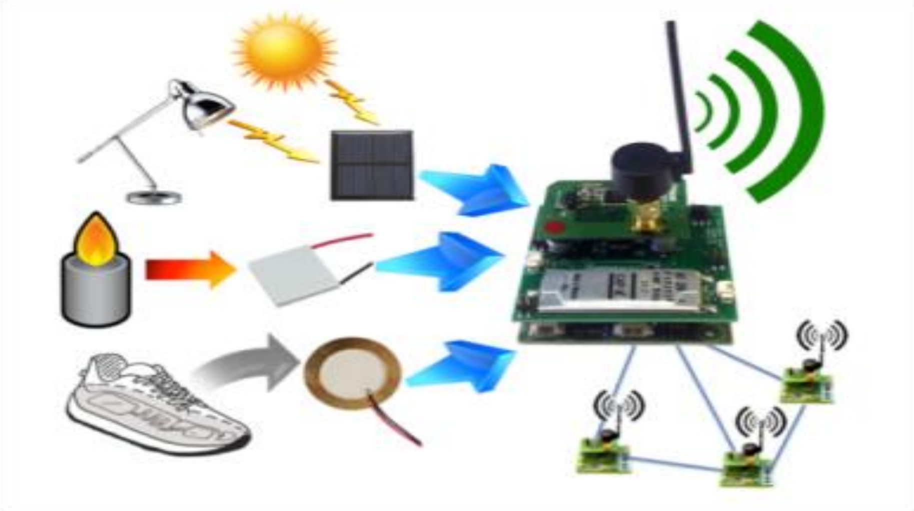
1. The data deluge

- The drivers
  - Datafication of people's lives



1. The data deluge

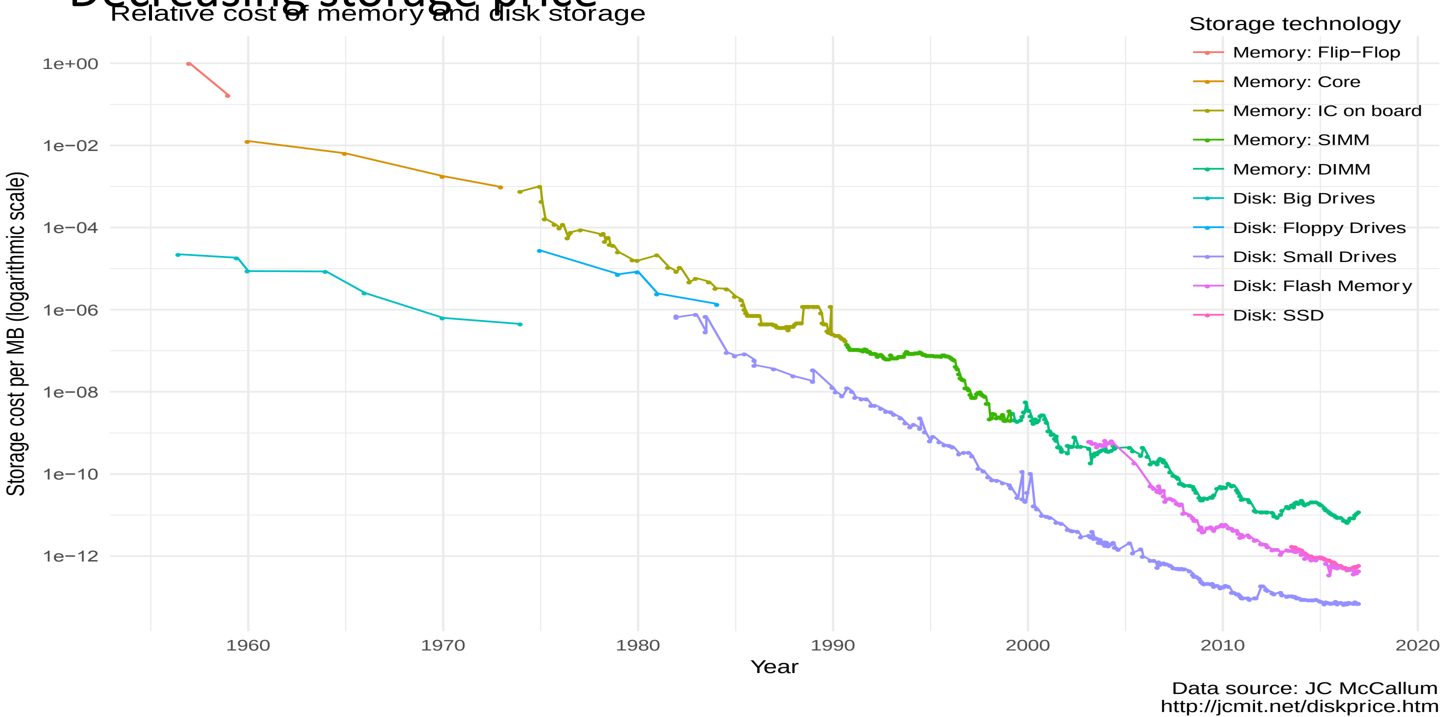
- The drivers
  - Datafication of people's lives



Temperature sensors	Proximity sensors
Acoustic sensors	Pressure sensors
Chemical/smoke and gas sensors	Level sensors
Tactile sensors	Cameras

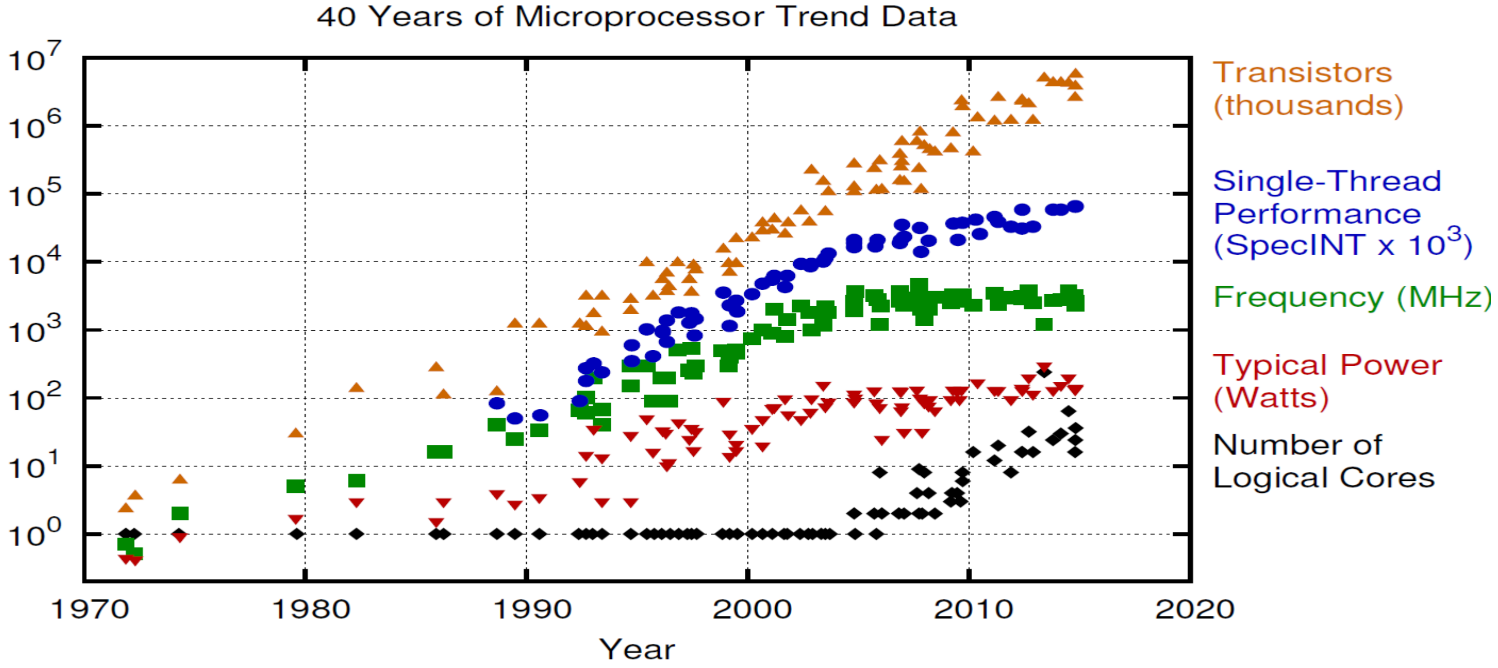
1. The data deluge

- The drivers
  - Datafication of people's lives
  - Decreasing storage price



1. The data deluge

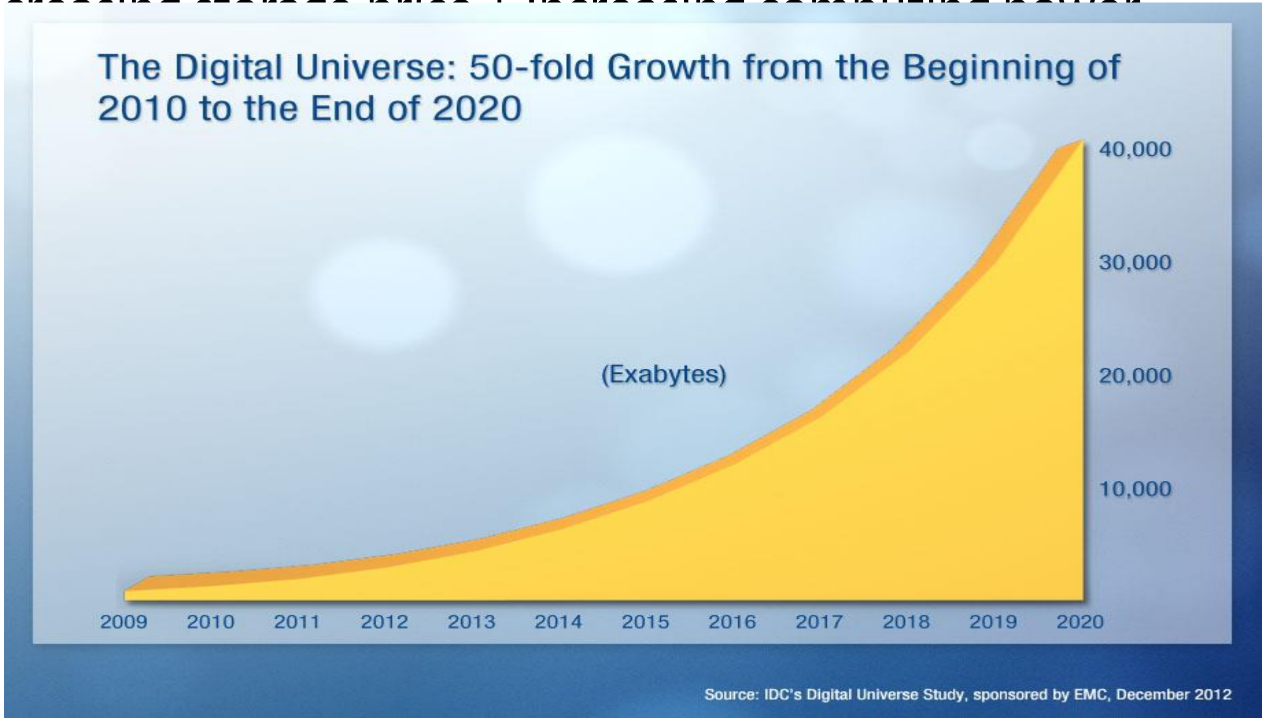
- The drivers
  - Datafication of people's lives
  - Decreasing storage price + Increasing computing power



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

1. The data deluge

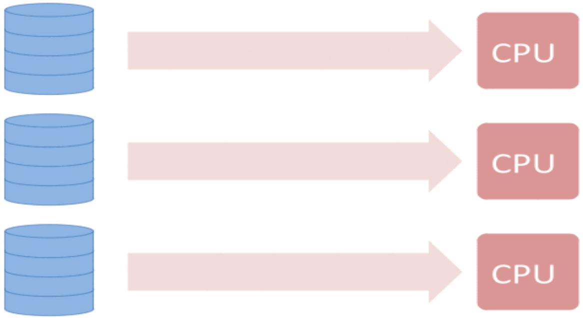
- The drivers
  - Datafication of people's lives
  - Decreasing storage price + Increasing computing power



1. The data deluge

- The drivers
  - Datafication of people's lives
  - Decreasing storage price + Increasing computing power
  - Development of distributed computing paradigm

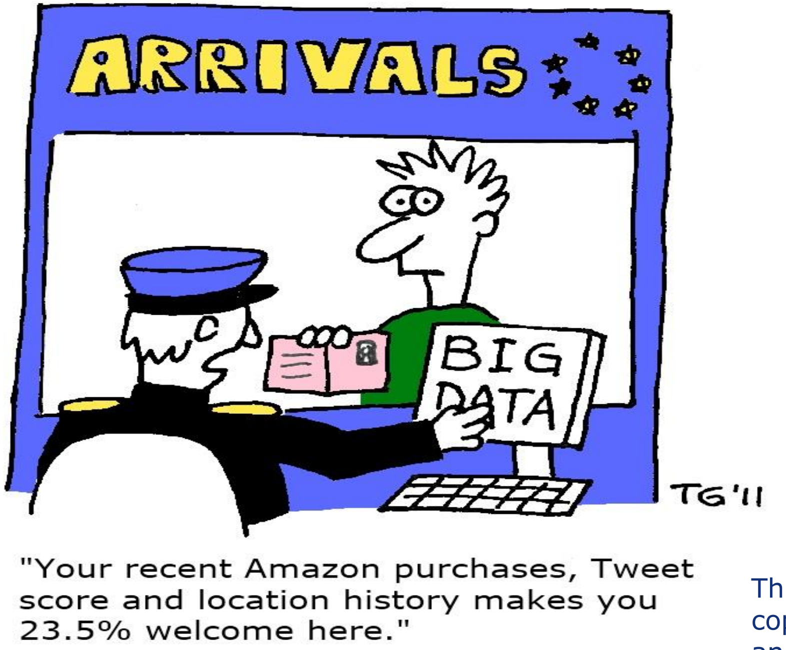
From ***serial computing*** to ***distributed computing***



Distributed data and processing

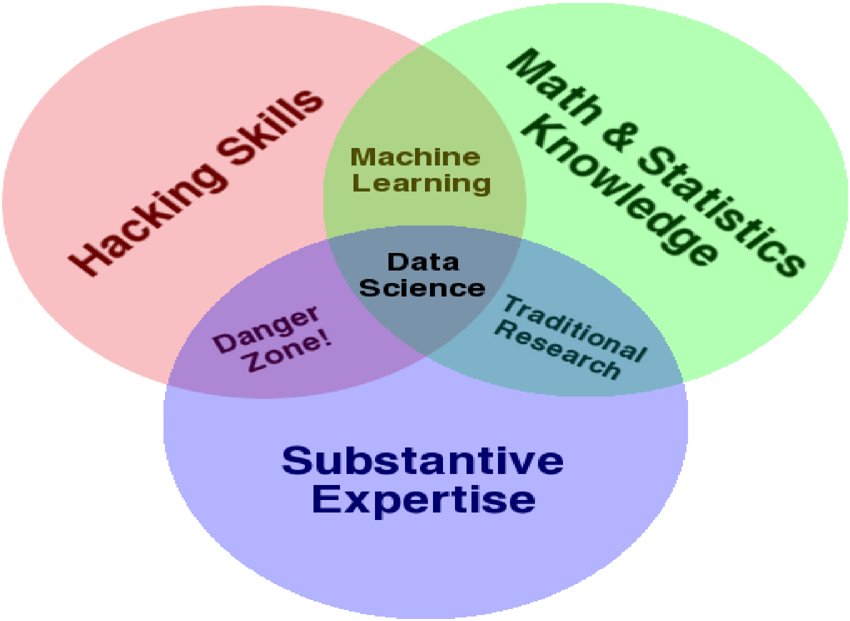
The big data phenomenon

2. Analytics



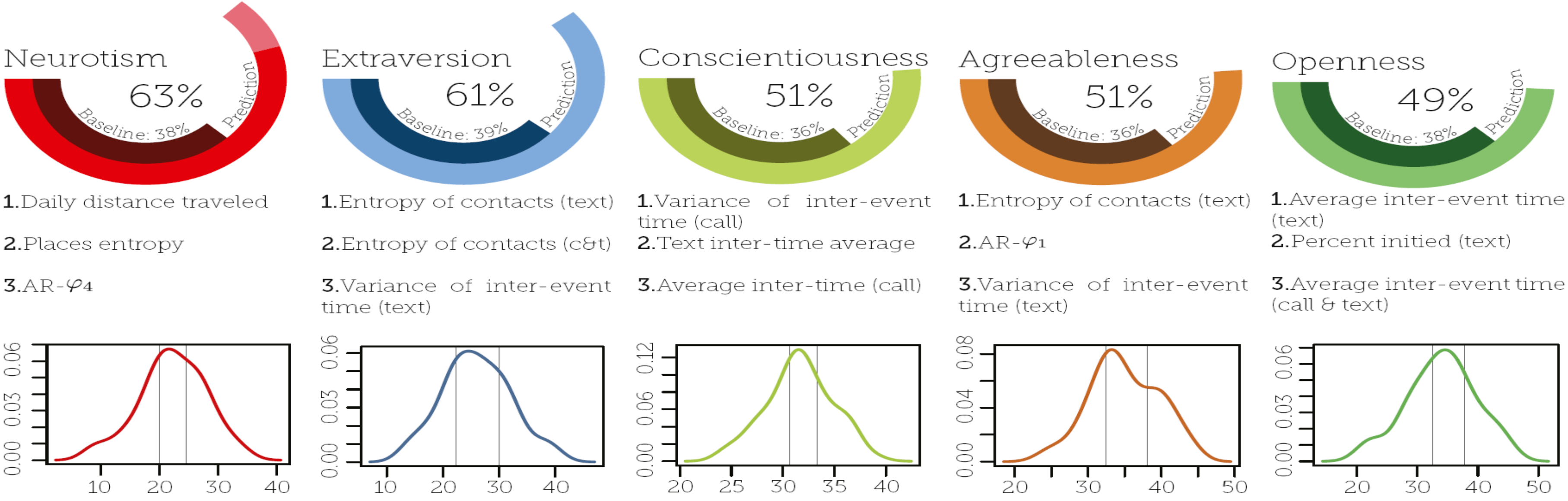
2. Analytics

Data Science: a new discipline?



2. Analytics

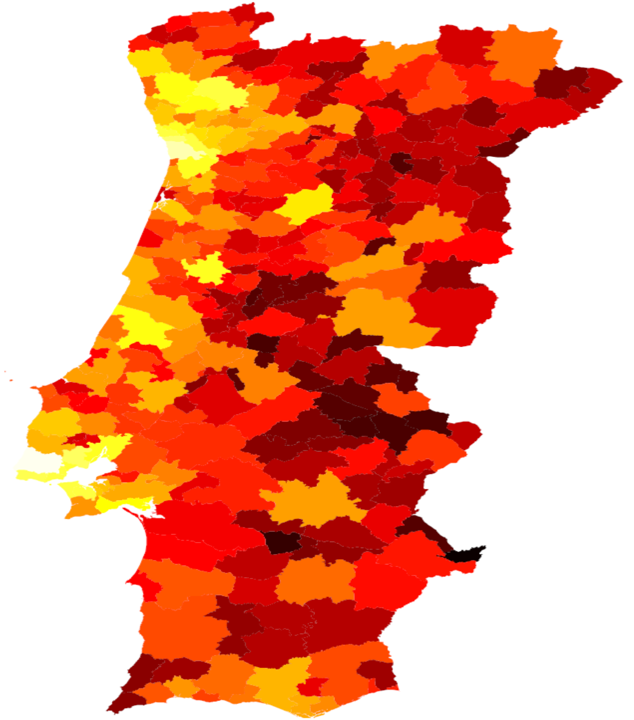
Predicting Personality Using Mobile Phone-Based Metrics



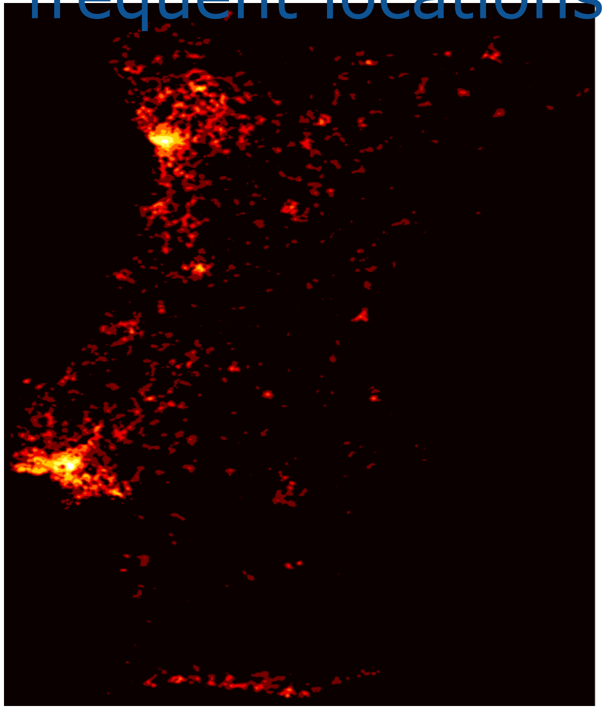
de Montjoye, Yves-Alexandre, et al. "Predicting personality using novel mobile phone-based metrics." Social Computing, Behavioral-Cultural Modeling and Prediction. Springer Berlin Heidelberg, 2013. 48-55.

2. Analytics

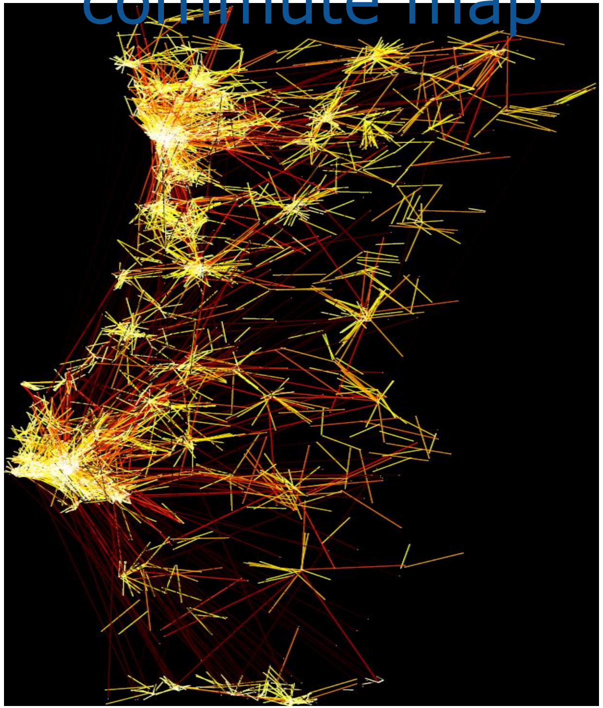
Population statistics



Mobile phone frequent locations



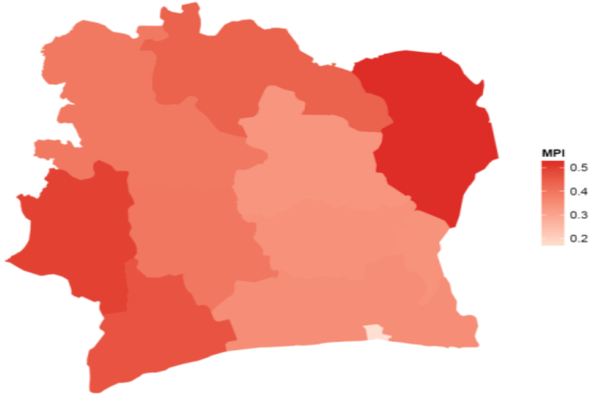
Mobile phone commute map



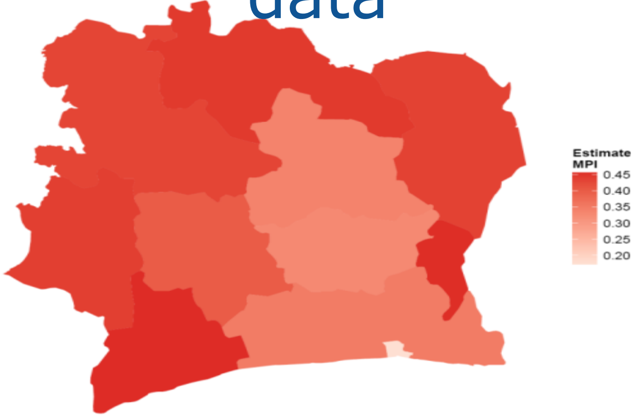
Csáji, Balázs Cs, et al. "Exploring the mobility of mobile phone users." Physica A: Statistical Mechanics and its Applications 392.6 (2013): 1459-1473.

2. Analytics

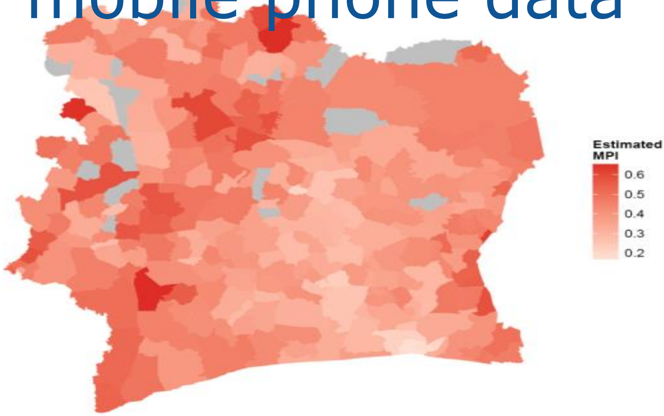
Multidimensional Poverty Index
(Lighter colour indicates higher poverty)



Poverty map estimated based on mobile phone data

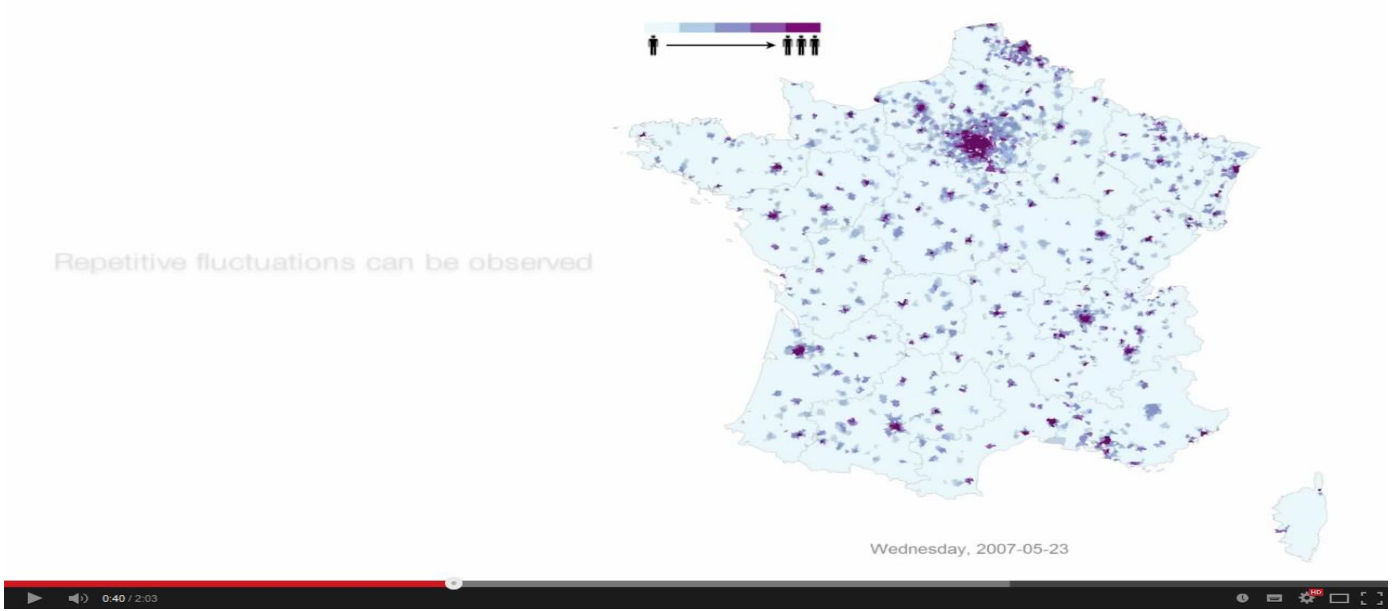


Poverty map in finer granularity estimated based on mobile phone data



Smith, Christopher, Afra Mashhadi, and Licia Capra. "Ubiquitous sensing for mapping poverty in developing countries." Paper submitted to the Orange D4D Challenge (2013).

2. Analytics



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

Population Mapping Using Mobile Phone Data

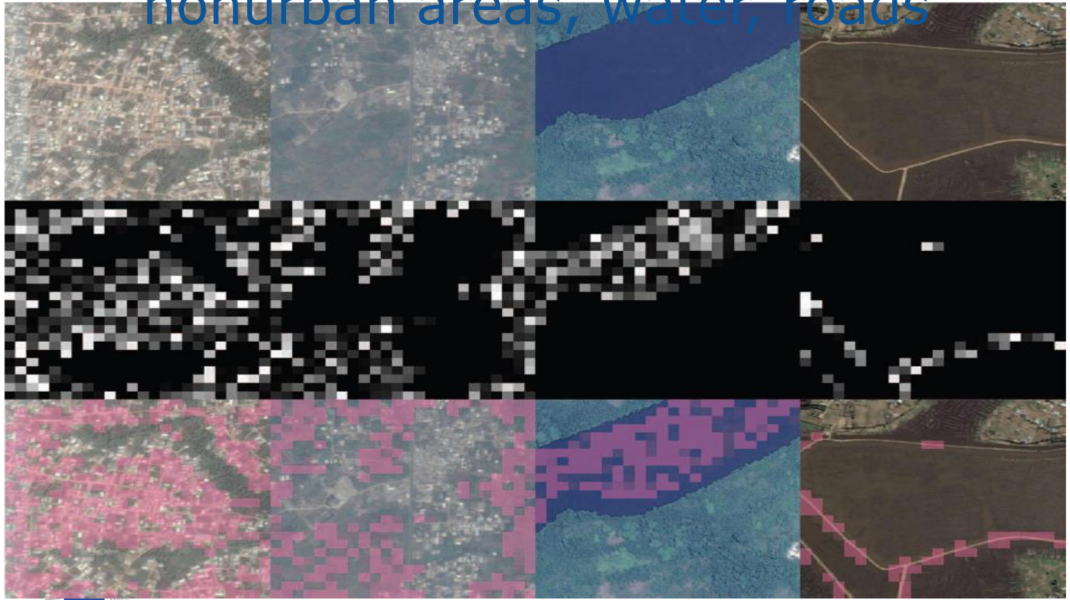
Deville, Pierre, et al. "Dynamic population mapping using mobile phone data." Proceedings of the National Academy of Sciences 111.45 (2014): 15888-15893.

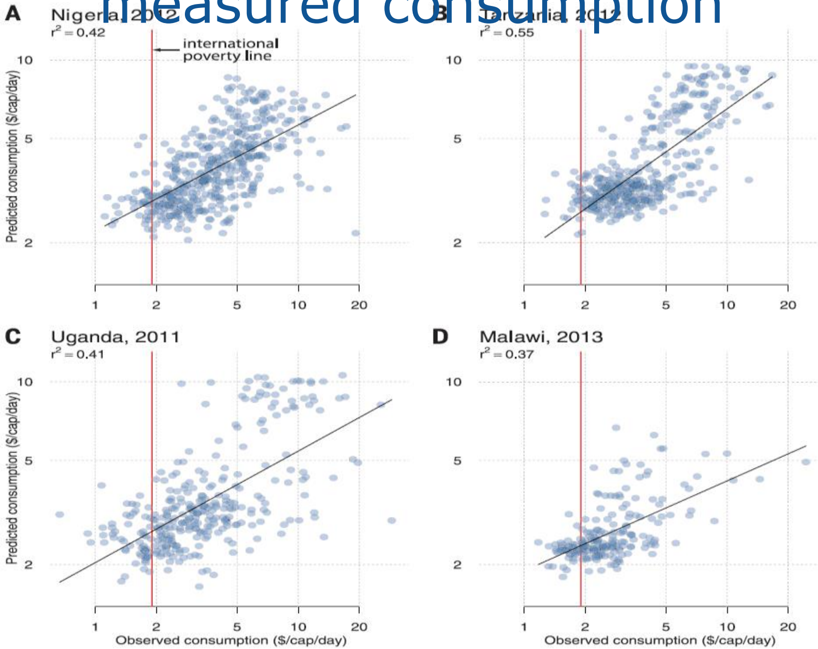
2. Analytics

Satellite images for predicting poverty

Extracting features: urban areas, nonurban areas, water, roads

Predicted vs survey measured consumption





[Jean, Neal, et al. "Combining satellite imagery and machine learning to predict poverty." Science 353.6301 (2016): 790-794.]

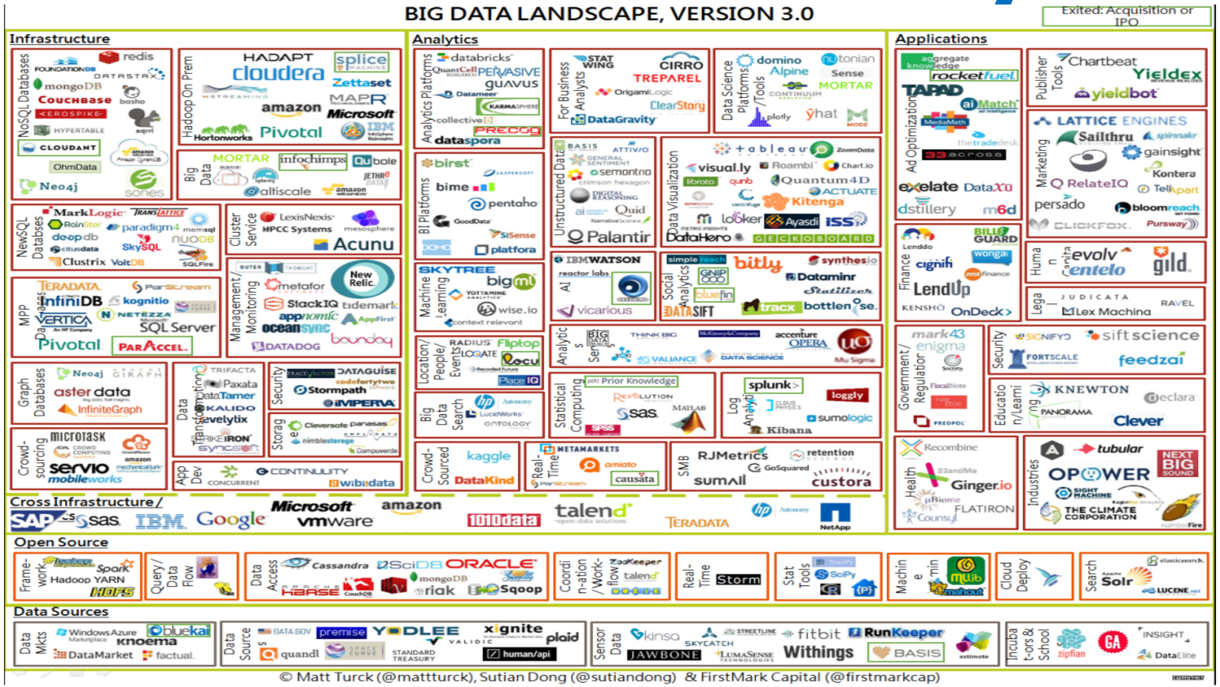
2. Analytics

So, is Data Science a new discipline?

- Signal processing
- Predictive analytics – machine learning

The big data phenomenon

3. Data economy



3. Data economy

- Monetisation of data: Data is the new oil
  - *Data has become a key infrastructure for 21st century knowledge economies. Data are not the "new oil" as still too often proclaimed. They are rather an infrastructure and capital good that can be used across society for a theoretically unlimited range of productive purposes, without being depleted. (OECD, 2015)*
  - Data always had value, but monetisation was difficult. Now business models based on data have emerged (e.g. Google, Facebook)

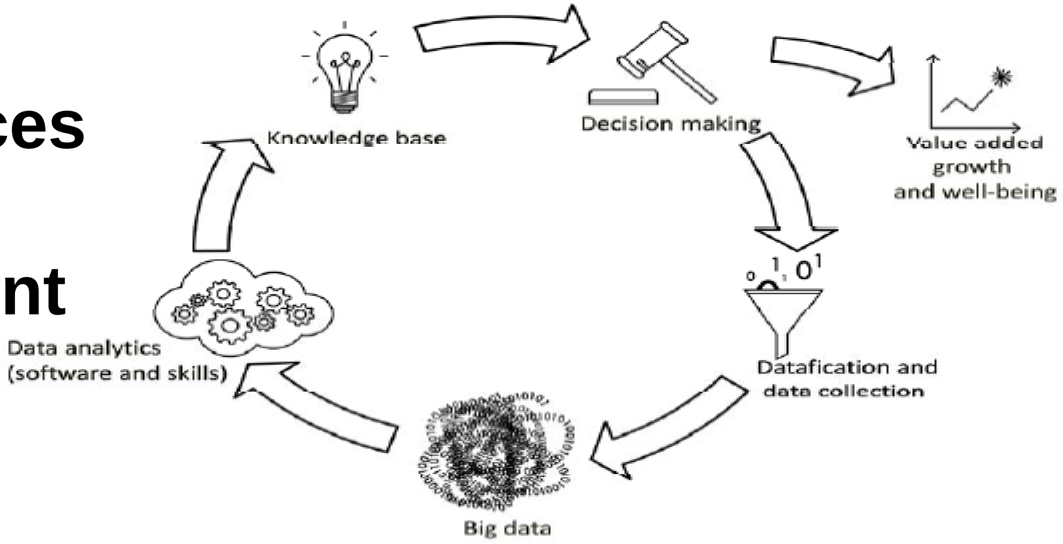
3. Data economy

- Monetisation of data: Data is the new oil
- Data as a new factor of production (competitive differentiating factor for businesses)

European Commission

3. Data economy

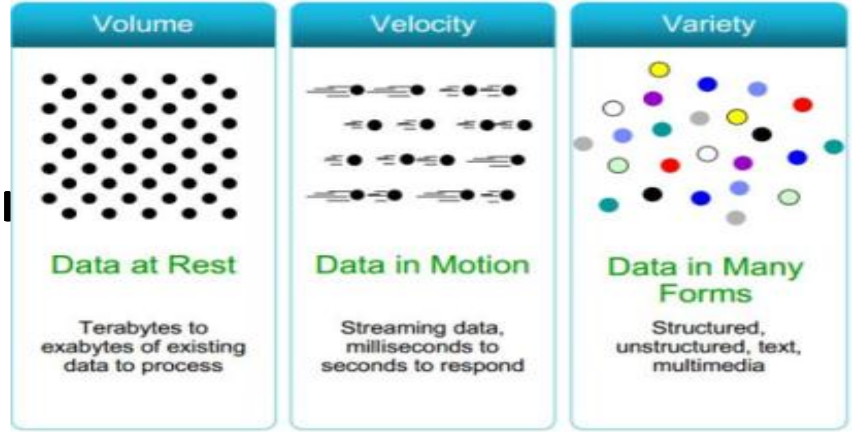
- Monetisation of data: Data is the new oil
- Data as a new factor of production (competitive diff businesses)
  - Data and analytics services exchanged between different organisations
- Data ecosystem



A STATISTICAL DEFINITION OF BIG DATA

Common definition of big data

- The 3 V’s of big data
  - Volume (large 'n' x 'p')
  - Velocity (data streaming)
  - Variety (different types of data & simultaneous use of several types of data)
- Sometimes other V's added:
  - 'Veracity'
- **This is a technological definitio**



A statistical definition of big data

- What do we mean by big data?
  - Natural language textual data
  - Network data
  - Multimedia (images, sound and video)
  - Positioning / location data
  - [Web activity]: Websites visited, …
  - [Sensors data]: Traffic sensors, …
  - [Process generated data]: Booking systems, bank transfers, …

A statistical definition of big data

- Big data vs. Big data sources
  - **Big data:** High dimensional data automatically captured during the use of IT systems or by sensors
  - **Big data sources:**
    - Non-designed ("found data") – online social networks, …
    - Designed – Satellite images, flying drones, traffic loops, …

A statistical definition of big data

- Consequences of high dimensionality
  - Curse (and bless) of dimensionality
  - Noise accumulation
  - Spurious correlations
  - Incidental endogeneity
- Methods to reduce dimensionality:
  - Generic
    - PCA
    - Prediction via machine learning (supervised learning)
    - Cluster analysis (unsupervised learning)
  - Specific
    - Natural language processing
    - Network analysis
    - Image recognition

Sources of big data

1. Human-sourced information (e.g. social networks)

- *Social Networks: Facebook, Twitter, Tumblr etc.*
- *Blogs and comments*
- *Personal documents*
- *Pictures: Instagram, Flickr, Picasa etc.*
- *Videos: Youtube etc.*
- *Internet searches*
- *Mobile data content: text messages*
- *User-generated maps*
- *E-Mail*

2. Process-mediated data (traditional business systems)

Data produced by Public agencies

- *Medical records*

Data produced by businesses

- *Commercial transactions*
- *Banking/stock records*
- *E-commerce*
- *Credit cards*

3. Machine-generated data

Data from sensors

Fixed sensors

- *Home automation*
- *Weather/pollution sensors*
- *Traffic sensors*
- *Traffic cameras*
- *Scientific sensors*
- *Security/surveillance videos/images*

Mobile sensors (tracking)

- *Mobile phone location*
- *Cars*
- *Satellite images*

Data from computer systems

- *Logs*
- *Web logs*

Adapted from [UNECE (2013) Classification of Types of Big Data]

Problems with non-designed sources of big data

- Coverage errors
- Selectivity bias
- Variables (of interest) not observed (directly)
- Unit identification problem (a.k.a. unit-error)

METHODOLOGICAL CHALLENGES OF BIG DATA

Methodological Challenges

- Indirect measurement of target and auxiliary variables
  - Variables of interest often are not directly measured (e.g. consumer confidence from social media posts)
  - Introduction of prediction – source of measurement error not to be underestimated
  - Model based estimates
    - Traditional linear regression models
    - Machine learning
  - Accuracy measures are very important

European Commission

Methodological Challenges

- Curse of dimensionality
  - When the dimensionality increases, the volume of the space increases so fast that the available data become sparse
  - To obtain a statistically reliable result, the amount of data needed often grows exponentially with the dimensionality
    - 10^2=100 evenly spaced sample points suffice to sample a 1-dimensional cube (a line) with no more than 10^{-2}=0.01 distance between points
    - an equivalent sampling (10^{-2}=0.01 distance between points) of a 10-dimensional hypercube would require $10^{20}[=(10^2)^{10}]$ sample points
  - An issue also because of use of highly non-linear models

Methodological Challenges

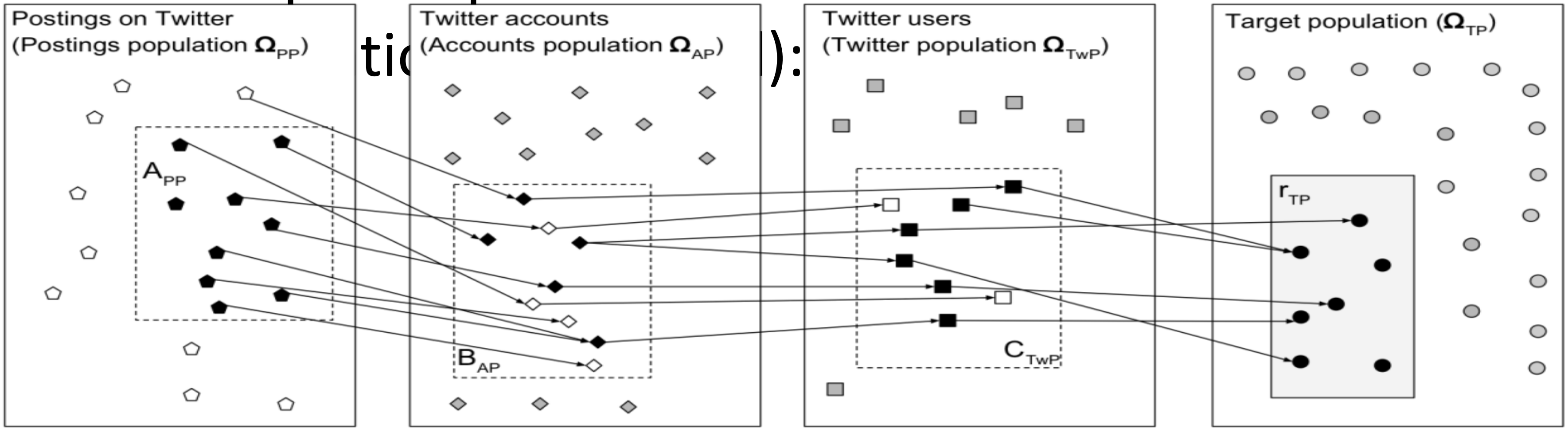
- Unit identification problem (aka unit error)
  - Multitude of populations in big data sources
  - Example: Populations observed in Twitter data (registration is required):
    - Twitter postings
    - Twitter accounts
    - Twitter users
  - Typical target populations in official statistics
    - Households
    - Persons resident in the country
    - Enterprises registered in a country

European Commission

Methodological Challenges

- Unit identification problem (aka unit error)
  - Example: Populations observed in Twitter data

tic … l):



Methodological Challenges

- Unit identification problem (aka unit error)
  - Consequences:
    - Introduces bias in point estimates at population and domain levels
    - It propagates to all statistics produced
  - A solution:
    - Unit error theory allows to study and measure the impact in terms of bias, variance, efficiency and consistency

European Commission

Methodological Challenges

- Selectivity
  - Error which results from:
    - individuals (unit specific; whether to tweet, use certain mobile provider)
    - data holder decisions (data holder specific; e.g. in terms of business concept, technical infrastructure)
  - includes coverage, measurement or non-response (or missingness) error
  - Consequence: potential bias in estimates

Methodological Challenges

- Selectivity
  - missingness (or non-response) and coverage components of selectivity can be represented as:
    - We define a response indicator variable R_i as

$$R_i = \begin{cases} 1 & if\ i \in r\ (element\ i\ responds, i.e. is\ not\ missing) \\ 0 & if\ i \notin r\ (element\ i\ does\ not\ respond, i.e. is\ missing) \end{cases}$$

    - The probability that a given unit will respond, i.e. will not be missing, (response propensity, propensity score) is given by

$$\rho_i = E(R_i = 1|\mathbf{x}, y) = P(R_i = 1|\mathbf{x}, y),$$

    - where x refer to auxiliary variables (e.g. demographics) and y is the target variable

Methodological Challenges

- Selectivity
  - 3 missingness mechanisms:
    - Missing Completely at Random (MCAR)

$$P(R_i = 1|\mathbf{x}, y) = P(R_i = 1).$$

    - missingness is due to random events, such as system failures and interruptions in the data collection process, which are not associated with neither x, v or y
    - the response probability (or non-missing probability) does not depend on x, v and y
    - Missingness is ignorable, as it does not have impact on bias

Methodological Challenges

- Selectivity
  - 3 missingness mechanisms:
    - Missing at Random (MAR)

$$P(R_i = 1|\mathbf{x}, y) = P(R_i = 1|\mathbf{x}),$$

    - missingness is connected with x, but not the target variable y.
    - the response probability is the expectation of the response indicator variable conditional on auxiliary variables, but not on target variable
    - Missingness is not ignorable, but can be corrected if information on auxiliary variables is available

Methodological Challenges

- Selectivity
  - 3 missingness mechanisms:
    - Missing Not at Random (MNAR)

$$P(R_i = 1|\mathbf{x}, y) \neq P(R_i = 1|\mathbf{x}).$$

    - missingness is not only related to auxiliary variables but also to the target variable y
    - response probability is expectation of response indicator variable conditional on auxiliary variables and target variable itself
    - Ignoring missingness in the presence of MNAR mechanism may result in large biases and erroneous inferences

Methodological Challenges

- Selectivity
  - Methods for correcting selectivity
    - Unit level approach
      - At individual (statistical units) level
    - Domain level approach
      - At aggregated level

Methodological Challenges

- Selectivity
  - Methods for correcting selectivity (unit level)
    - Reweighting
      - Methods that account for existing information about auxiliary variables -> if correlated to selectivity mechanism will correct it
        - » Generalized weight share method
        - » Calibration (model-free and model-assisted)
        - » Pseudo-empirical likelihood
      - Methods that address directly the selectivity mechanism
        - » Propensity weighting (model directly the propensity)
        - » Two-stage weighting method
        - » Lepkowski method (for under-coverage and self-selection)

Methodological Challenges

- Selectivity
  - Methods for correcting selectivity (unit level)
    - Modelling approach
      - The basic idea is that if the models include explanatory variables correlated to the selectivity mechanism then they can correct or mitigate selectivity bias
        - » Heckman selection model
        - » Hierarchical Bayes models
        - » Calibrated Bayes
        - » Pattern mixture model
        - » Machine learning (non-linear models)

Methodological Challenges

- Selectivity
  - Methods for correcting selectivity (unit level)
    - Data linking approach
      - Normally applied before reweighting or modelling methods are used
      - Purpose is to obtain from other datasets auxiliary variables not available in the big data
      - Methods:
        - » Record linkage (linking data from the same unit)
        - » Sample matching (linking data from normally different but <u>very similar</u> units)

Methodological Challenges

- Selectivity
  - Methods for correcting selectivity (domain level)
    - Reweighting
      - $\breve{\theta}_{cd}^{adj} = \breve{\theta}_{cd} \times a_d^{cover} \times a_d^{active} \times a_d^{share} \times a_{cd}^{cal}$
      - a^{cover} adjustment for coverage of technology
      - a^{active} adjustment for fraction of active users
      - a^{share} adjustment for market share data provider
      - a_{cd}^{cal} adjustment to known population totals (e.g. background variables)
      - Valid with MAR and corrects for coverage problems

Methodological Challenges

- Selectivity
  - Methods for correcting selectivity (domain level)
    - Modelling approach
      - Estimation of bias in big data source
        - » From a sample survey (or registers)
        - » Direct estimation + model for sub-domains based on a "ground truth"

European Commission

Methods to correct selection bias

- Unit-level methods to correct selectivity
  - Pseudo-design approach - reweighting
    - Generalized weight share method
    - Model-free calibration
    - Model-assisted calibration
    - Propensity weighting
    - Pseudo-empirical likelihood
    - Adjusted weights
    - Two-step weighting method

Methods to correct selection bias

- Unit-level methods to correct selectivity
  - Pseudo-design approach - reweighting
  - Modelling approach
    - Small area estimation approach
      - M-quantile models.
    - Bayesian Approach
      - Hierarchical Bayesian approach
      - Calibrated Bayes approach
      - Pattern-mixture models – MNAR case
    - Machine-learning approach
      - K-nearest neighbours
      - Artificial Neural networks
      - Classification and regression trees

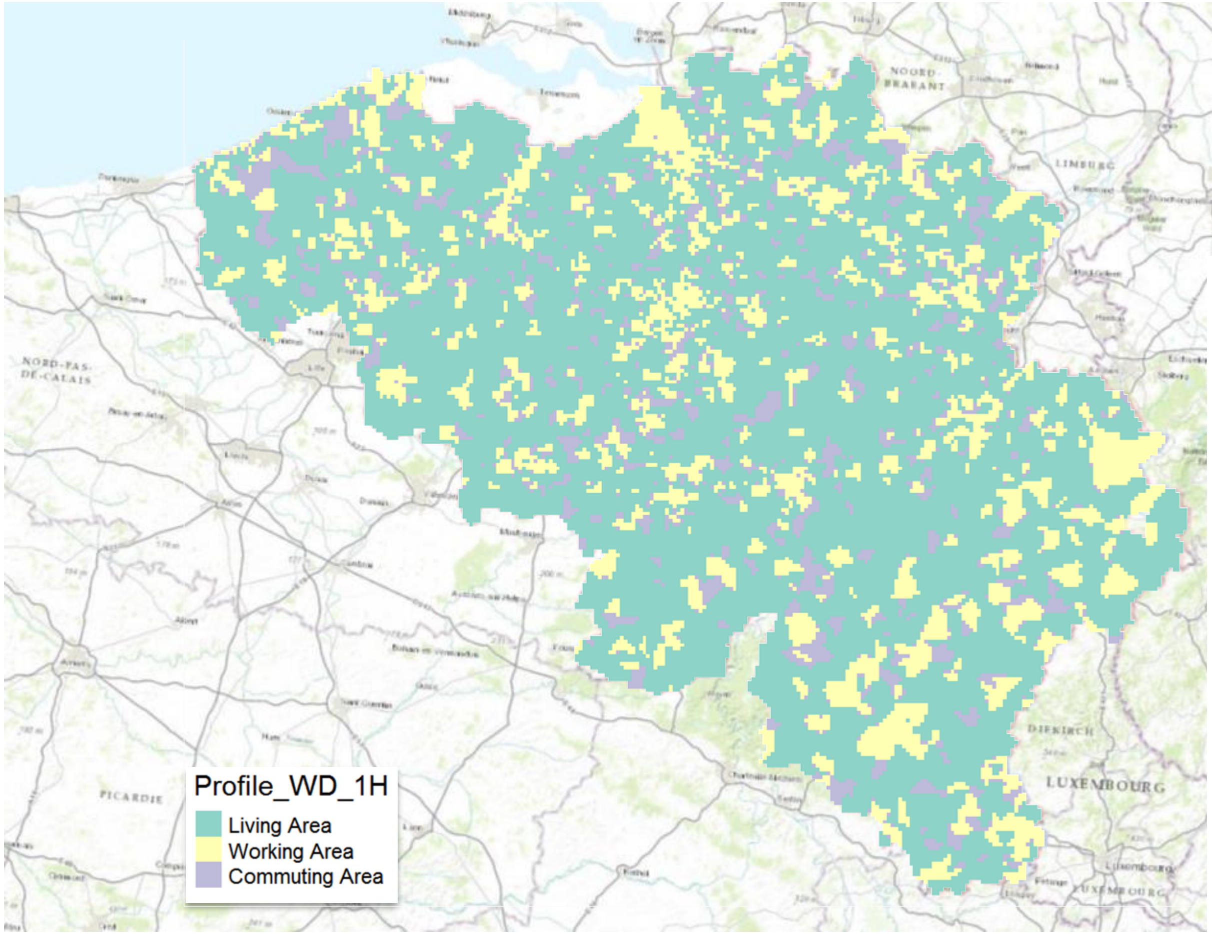
Methods to correct selection bias

- Unit-level methods to correct selectivity
  - Pseudo-design approach - reweighting
  - Modelling approach
  - Data linking approach
    - Record linkage
    - Sample Matching
    - Software

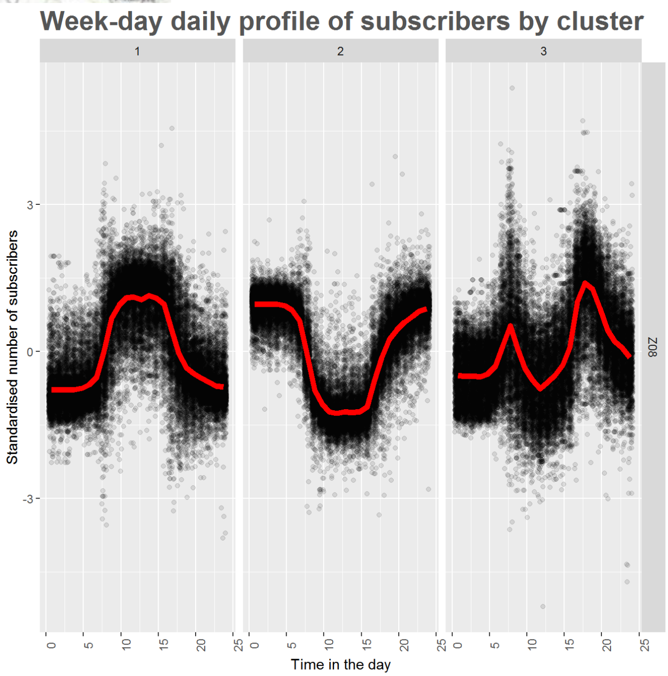
Methods to correct selection bias

- Unit-level methods to correct selectivity
  - Pseudo-design approach - reweighting
  - Modelling approach
  - Data linking approach
- Domain-level methods to correct selectivity
  - Pseudo-design methods – reweighting
  - Modelling approach
    - Direct estimation of bias
    - Blending of estimates

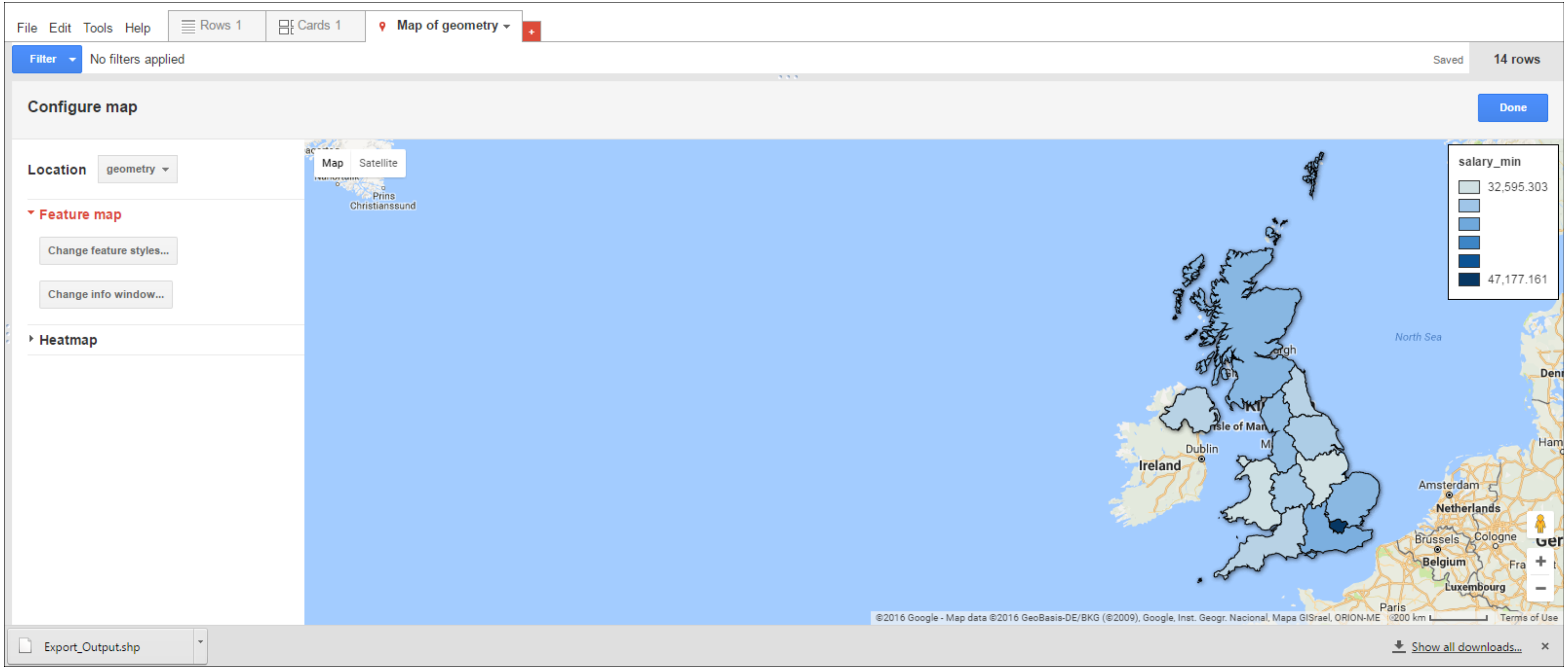
EXAMPLES OF APPLICATIONS



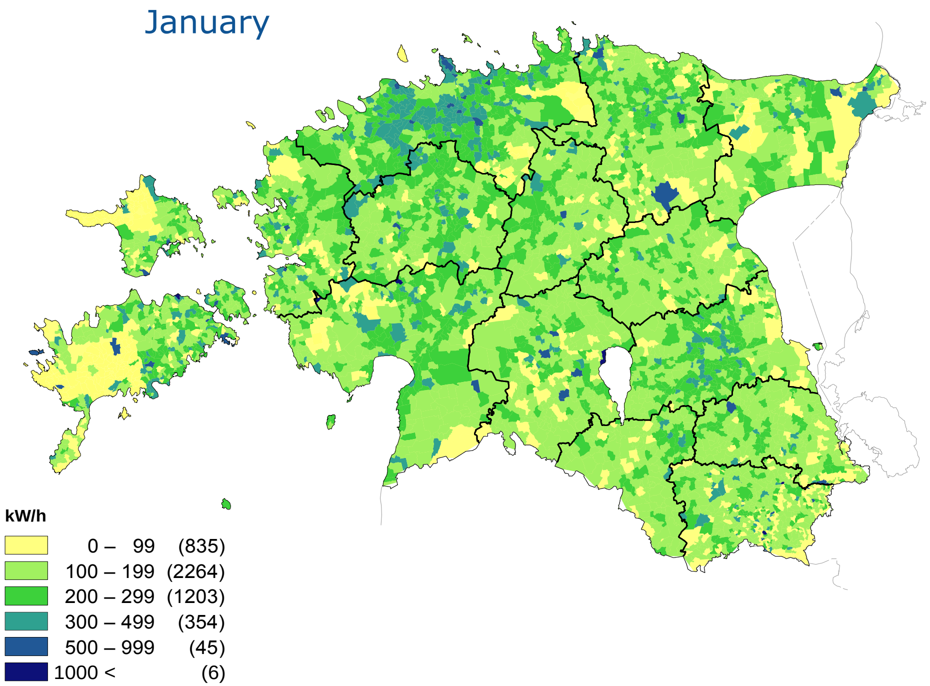
Mobile network data for automatic classification of territory

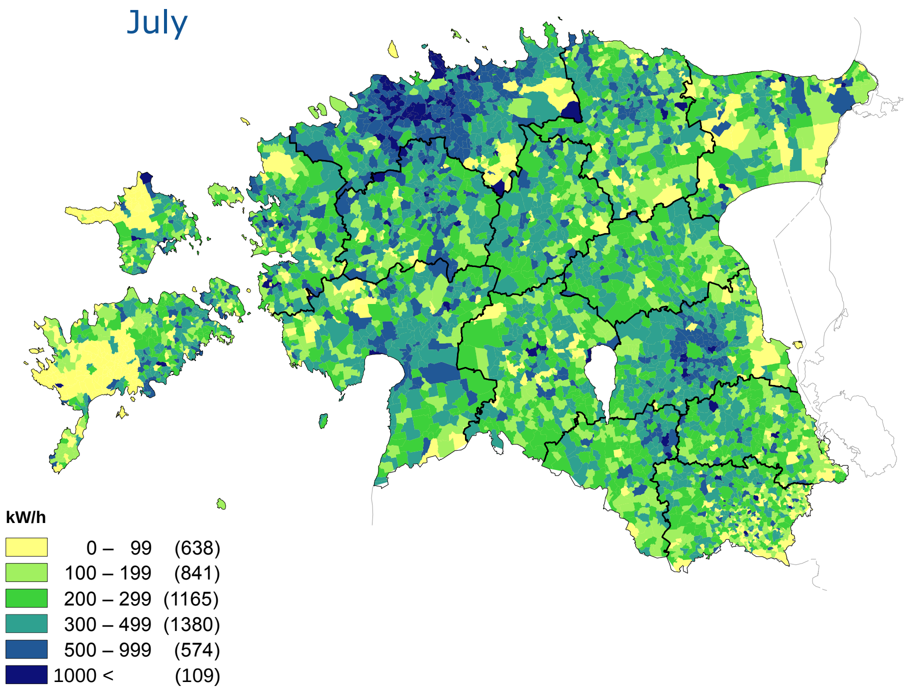


Distribution of offered salaries for job adverts from a web job portal on a specific

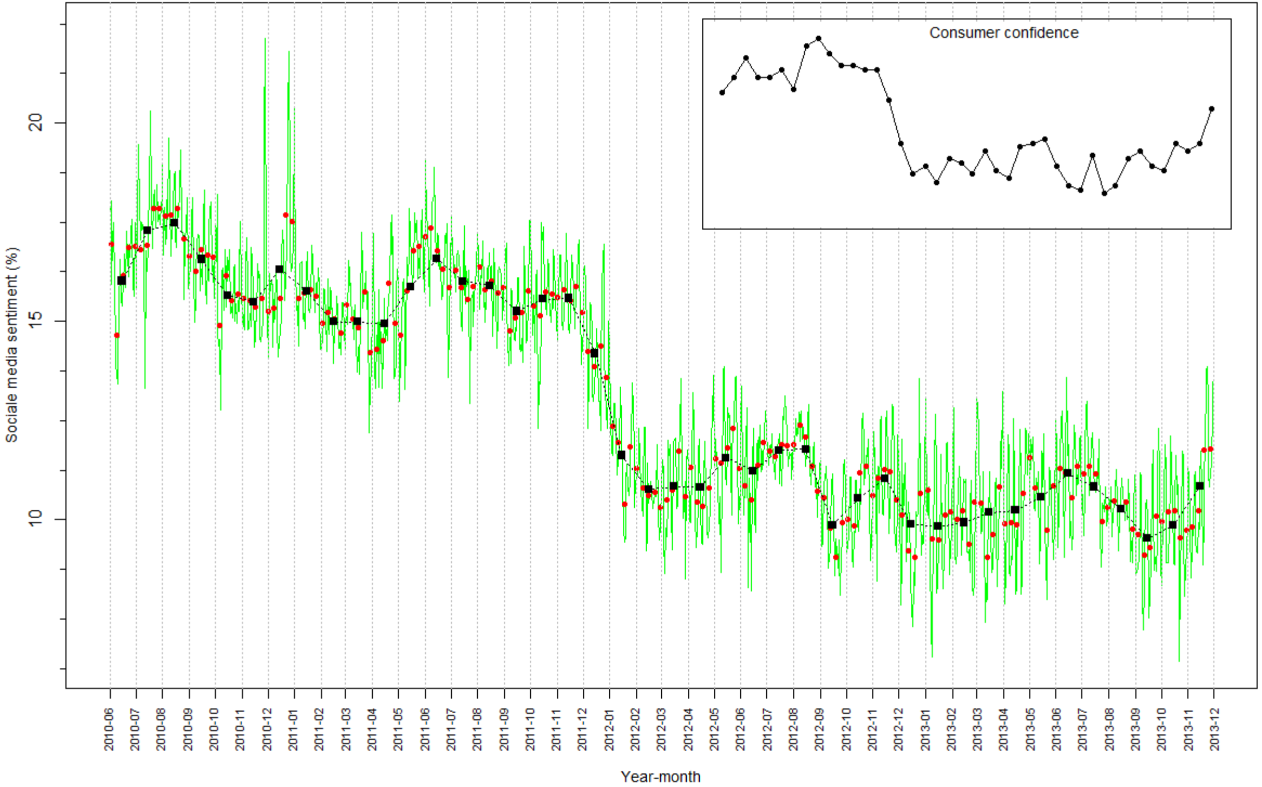


Average monthly electricity consumption of private persons (smart electricity meters)





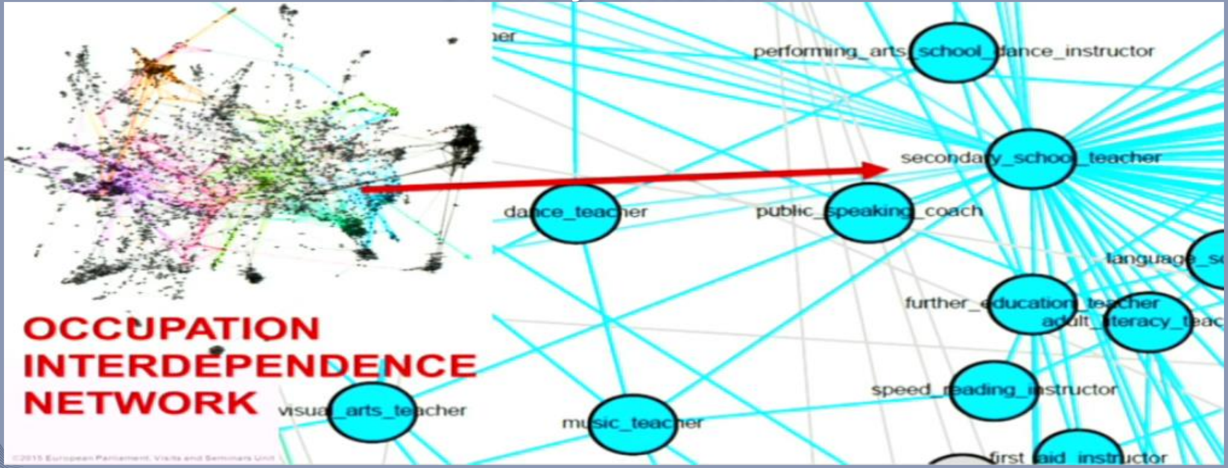
Sentiment in tweets



New phenomena, new statistics

- Technological innovation creates new phenomena to be measured with new statistical products
  - Platform / sharing economy, online job markets, cryptocurrencies, smart contracts, initial coin offerings, …

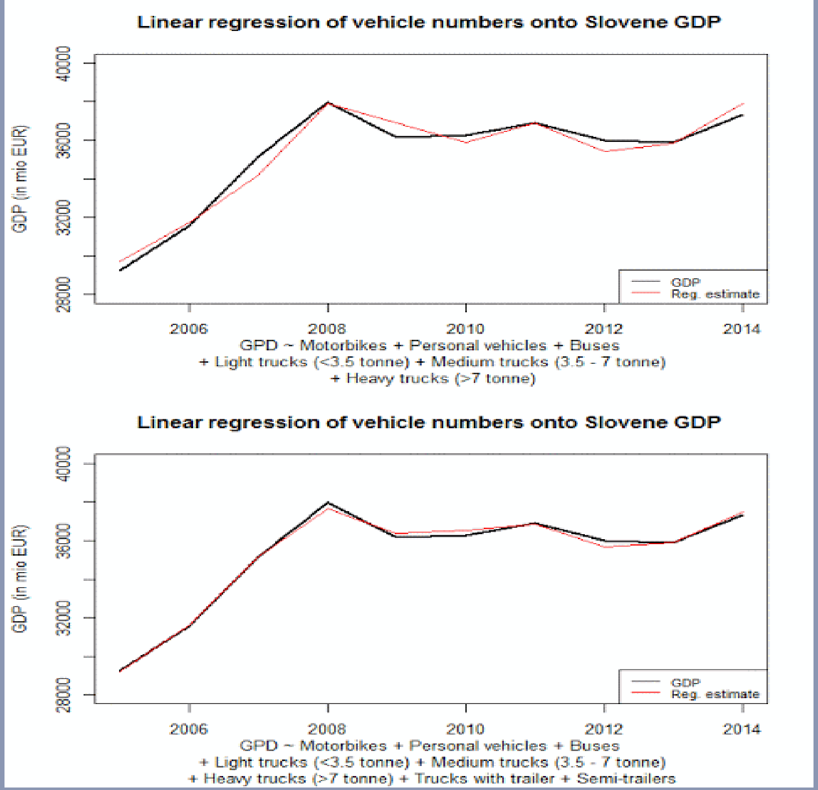
Online job market



New data sources

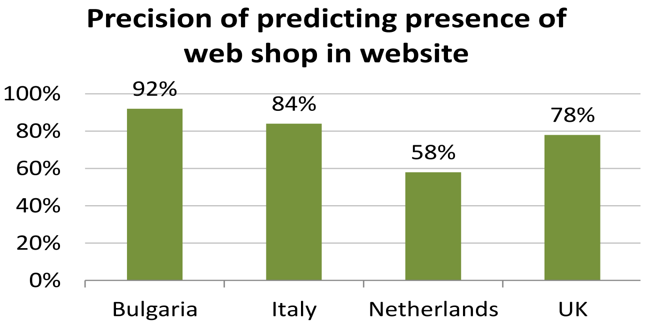
- Technological innovation creates new data sources
  that capture additional dimensions of phenomena, improve timeliness & relevance of statistics
  - IoT, smart vehicles, smart meters, smart houses, wearables, online social networks …

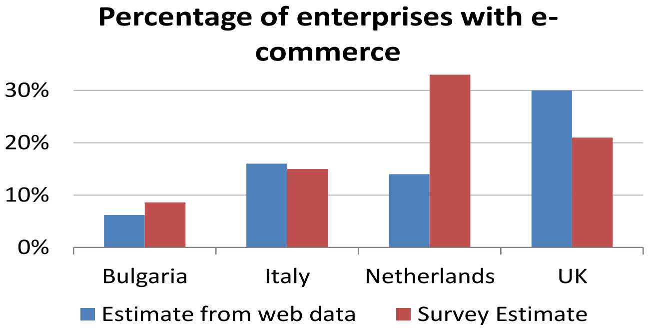
Nowcasting Slovenian GDP using traffic loops data



New processing opportunities

- Technological innovation creates <u>new processing opportunities</u> for existing data
  - automatic text interpretation, cognitive image processing, deep learning, AI
  - turning documents, images, videos, text messages etc. into mineable sources





European Commission

Thank you for your attention

Fernando Reis

Eurostat Task Force on Big Data

fernando.reis@ec.europa.eu

https://github.com/reisfe/

https://twitter.com/reisfe/

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