Big Data for Official Statistics

The UN perspective

Karoly Kovacs

Data Innovation and Capacity Branch, United Nations Statistics Division

Overview

- Big data – definition, data sources
- GWG on Big Data for Official Statistics
- Statistical Data Infrastructure
- Quality framework for Big Data

Big data
Definition – data sources -

Big Data

Wikipedia:

The term has been in use since the 1990s, with some giving credit to John Mashey for popularizing the term. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

(https://en.wikipedia.org/wiki/Big_data)

Big Data

Wikipedia:

"**Big data**" is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software. Data with many cases (rows) offer greater statistical power, while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate.

(https://en.wikipedia.org/wiki/Big_data)

Big Data

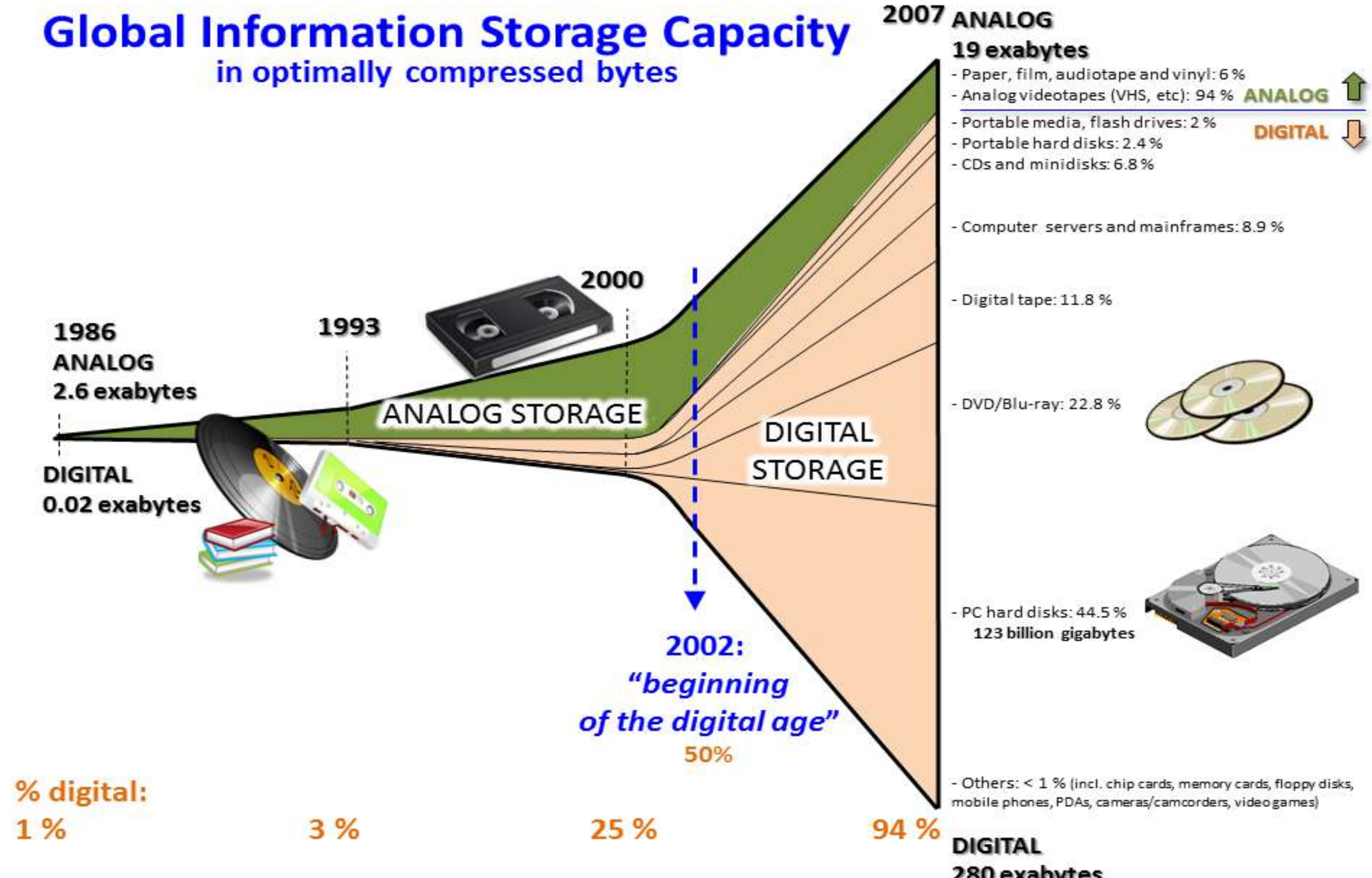
Wikipedia:

Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy, and data source. Big data was originally associated with three key concepts: *volume*, *variety*, and *velocity*. Other concepts later attributed with big data are *veracity (i.e., how much noise is in the data)* and *value*.

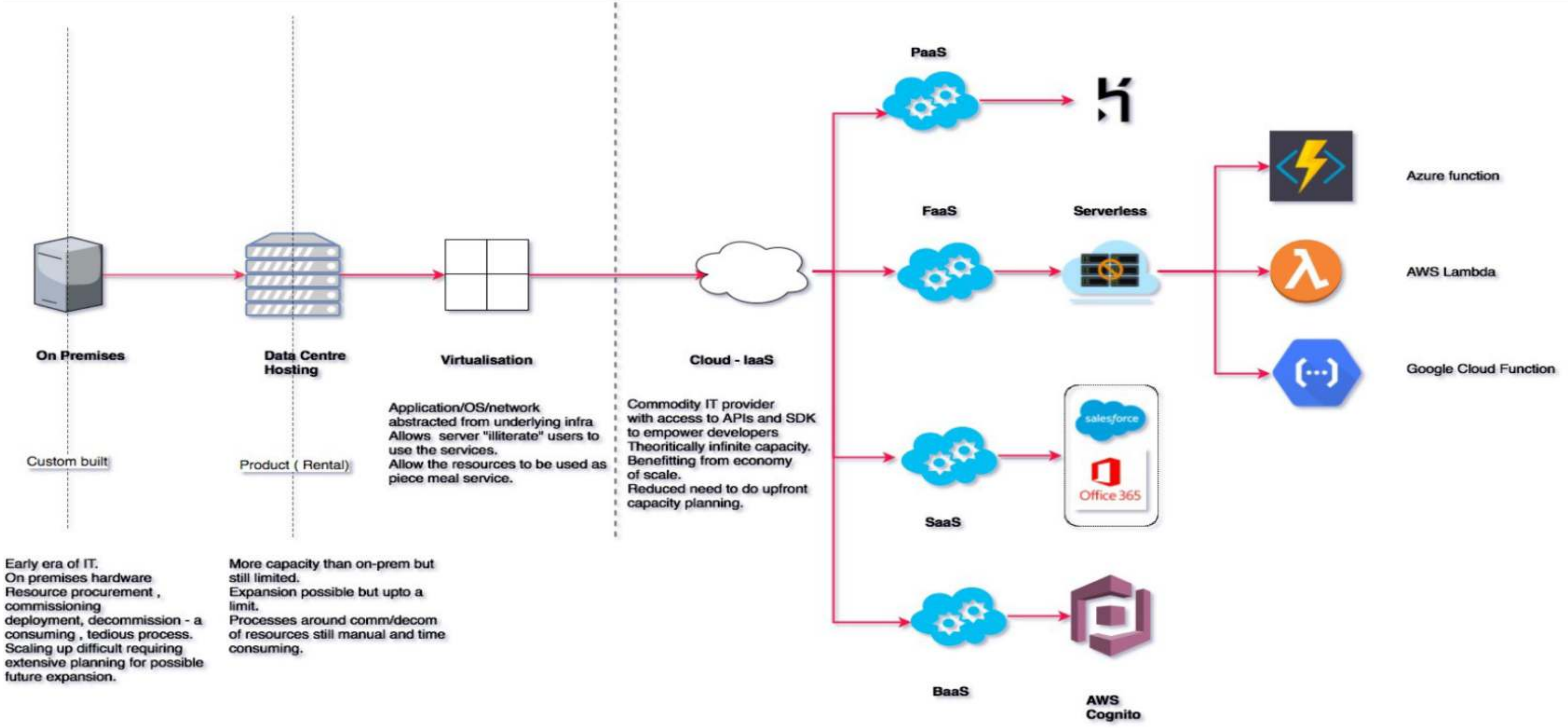
(https://en.wikipedia.org/wiki/Big_data)

Big Data

Wikipedia:



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html



Application/OS/network abstracted from underlying infra
Allows server "illiterate" users to use the services.
Allow the resources to be used as piece meal service.

Commodity IT provider with access to APIs and SDK to empower developers
Theoritically infinite capacity.
Benefitting from economy of scale.
Reduced need to do upfront capacity planning.

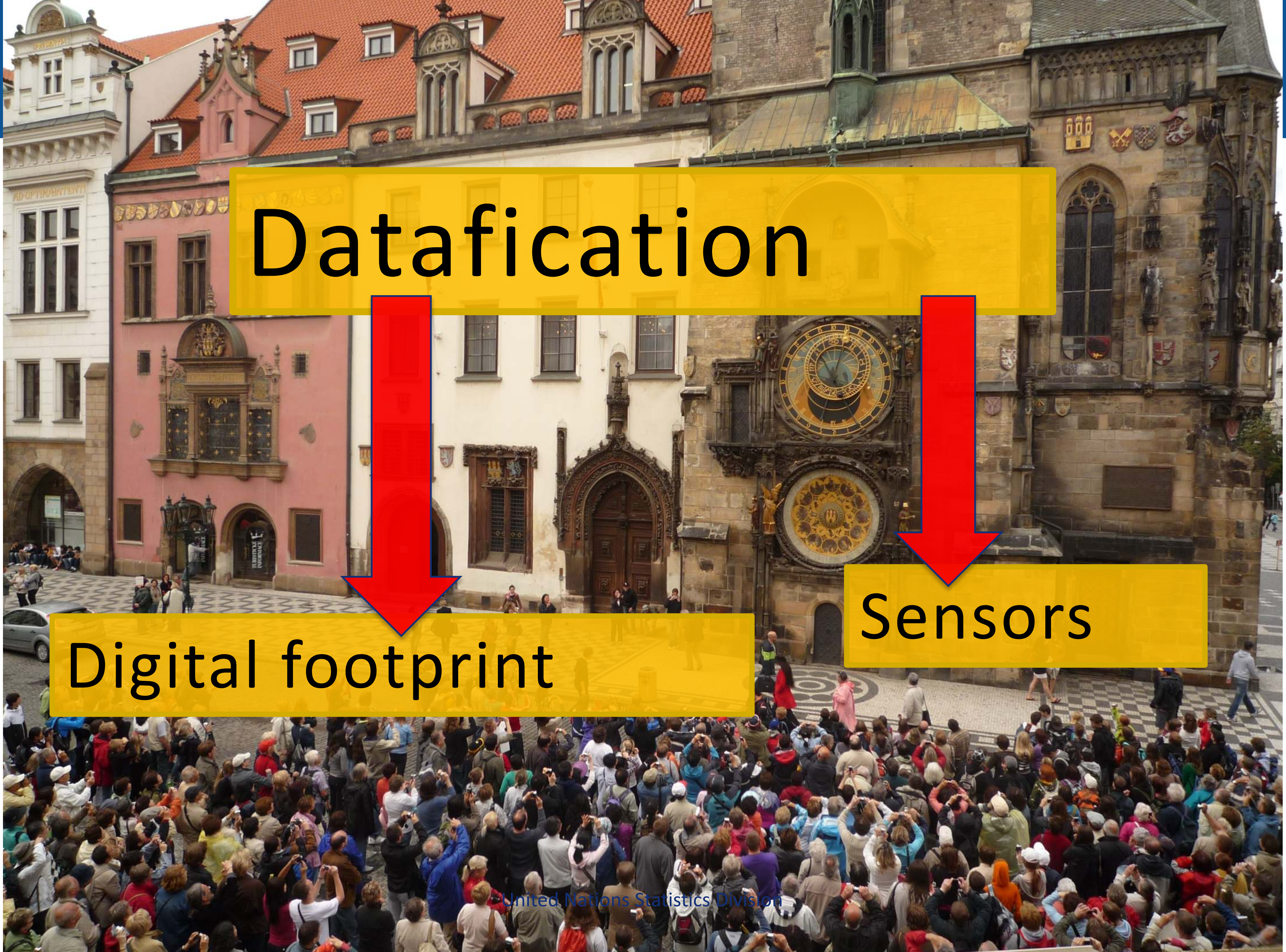
Early era of IT.
On premises hardware
Resource procurement , commissioning deployment, decommission - a consuming , tedious process.
Scaling up difficult requiring extensive planning for possible future expansion.

More capacity than on-prem but still limited.
Expansion possible but upto a limit.
Processes around comm/decom of resources still manual and time consuming.

Source: https://medium.freecodecamp.org/a-brief-history-of-serverless-or-how-i-learned-to-stop-worrying-and-start-loving-the-cloud-7e2fc633310d

What are common sources of Big Data?

- **Automatically generated data** in electronic format, **such as mobile phone data, social media data, electronic commercial transactions, sensor networks, smart meters, GPS tracking device, or satellite images**
- **High frequency, and/or fine granularity, and/or wide coverage**

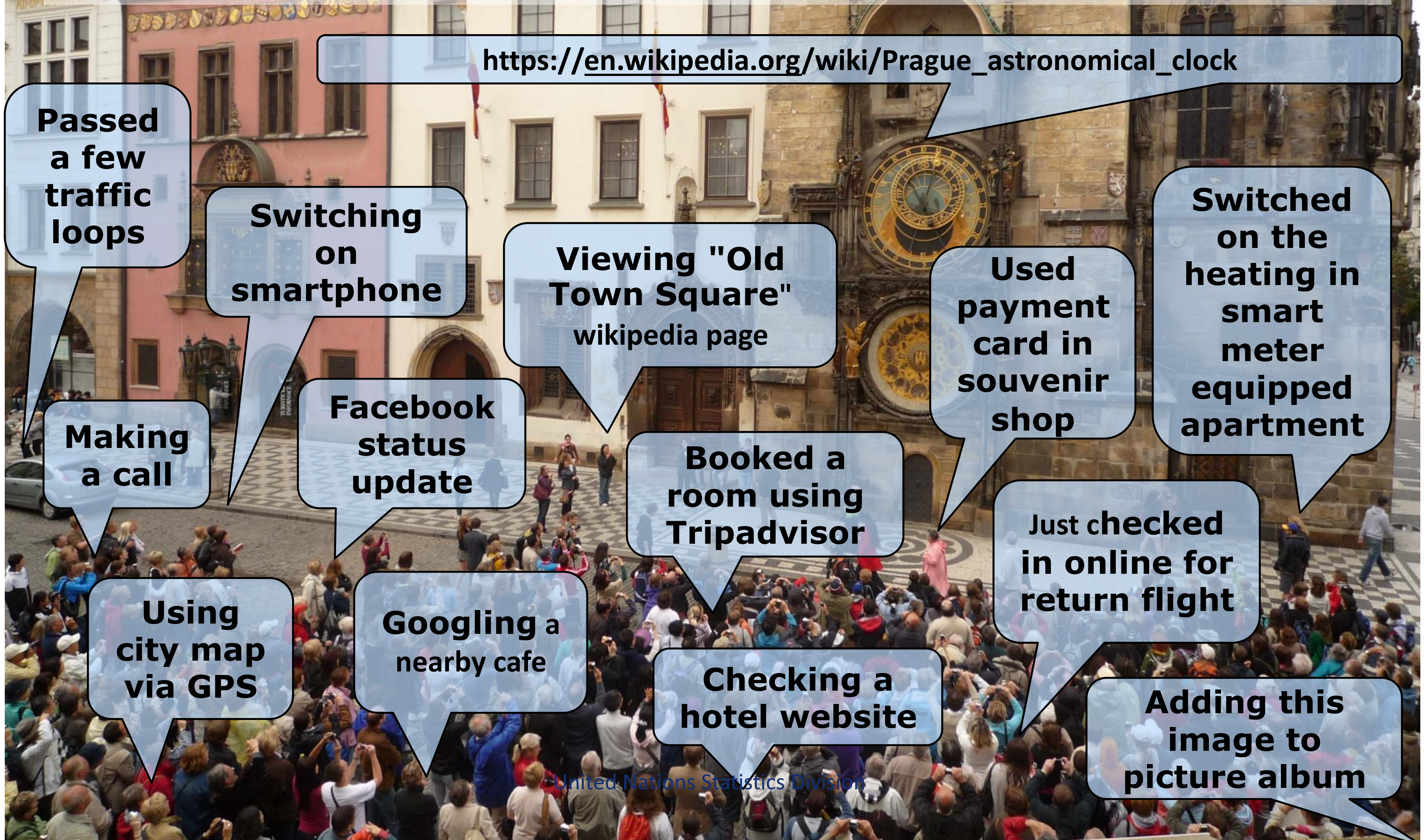


Datafication

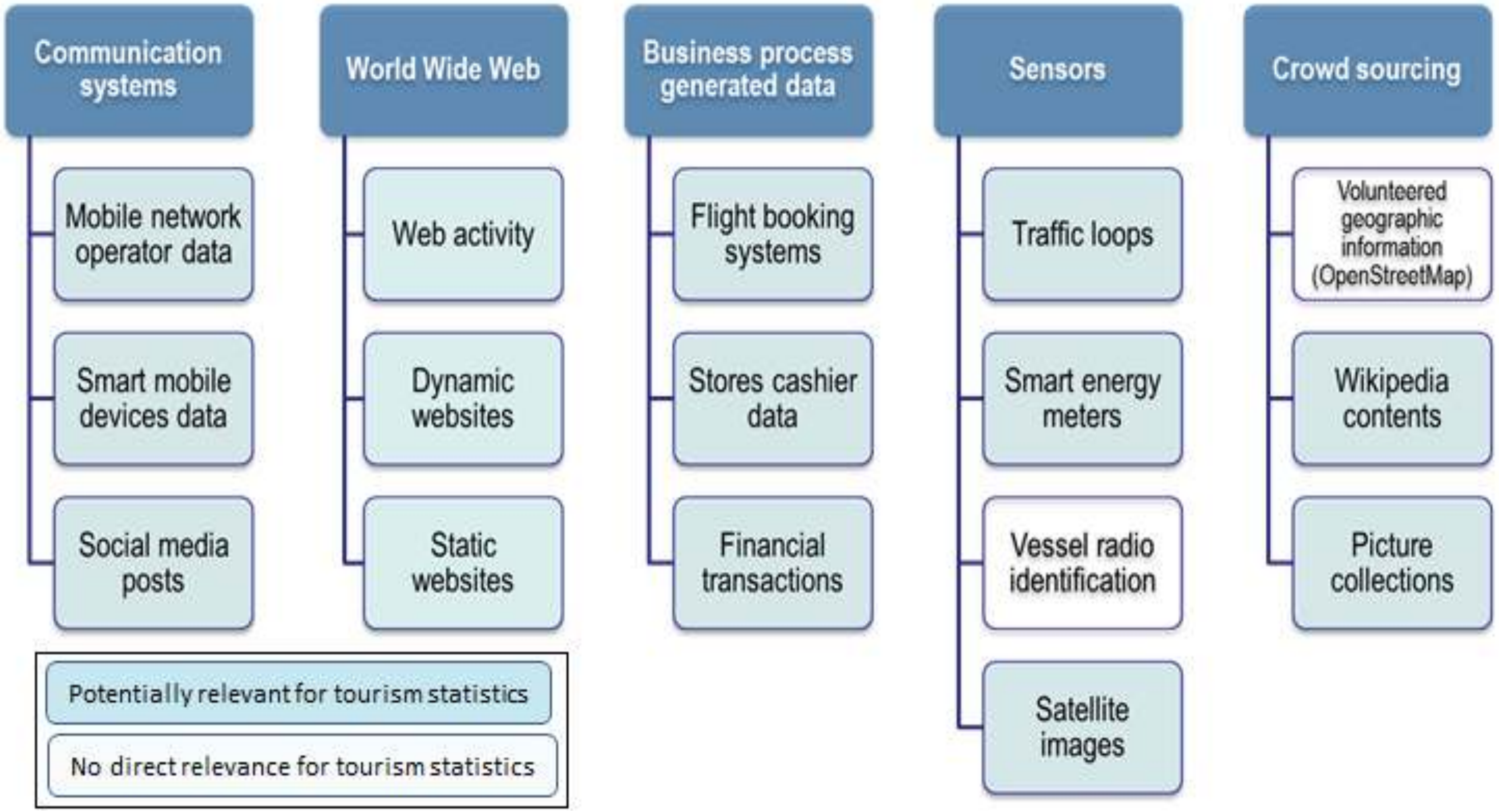
Digital footprint

Sensors

As a "special case" of human mobility, tourism is a human activity that leaves multiple traces, as a digital footprint or captured by sensors


https://en.wikipedia.org/wiki/Prague_astronomical_clock
Passed a few traffic loops
Switching on smartphone
Viewing "Old Town Square" wikipedia page
Used payment card in souvenir shop
Switched on the heating in smart meter equipped apartment
Making a call
Facebook status update
Booked a room using Tripadvisor
Just checked in online for return flight
Using city map via GPS
Googling a nearby cafe
Checking a hotel website
Adding this image to picture album

Taxonomy of big data sources (Eurostat 2017)



- Communication systems
  - Mobile network operator data
  - Smart mobile devices data
  - Social media posts
- World Wide Web
  - Web activity
  - Dynamic websites
  - Static websites
- Business process generated data
  - Flight booking systems
  - Stores cashier data
  - Financial transactions
- Sensors
  - Traffic loops
  - Smart energy meters
  - Vessel radio identification
  - Satellite images
- Crowd sourcing
  - Volunteered geographic information (OpenStreetMap)
  - Wikipedia contents
  - Picture collections

Potentially relevant for tourism statistics

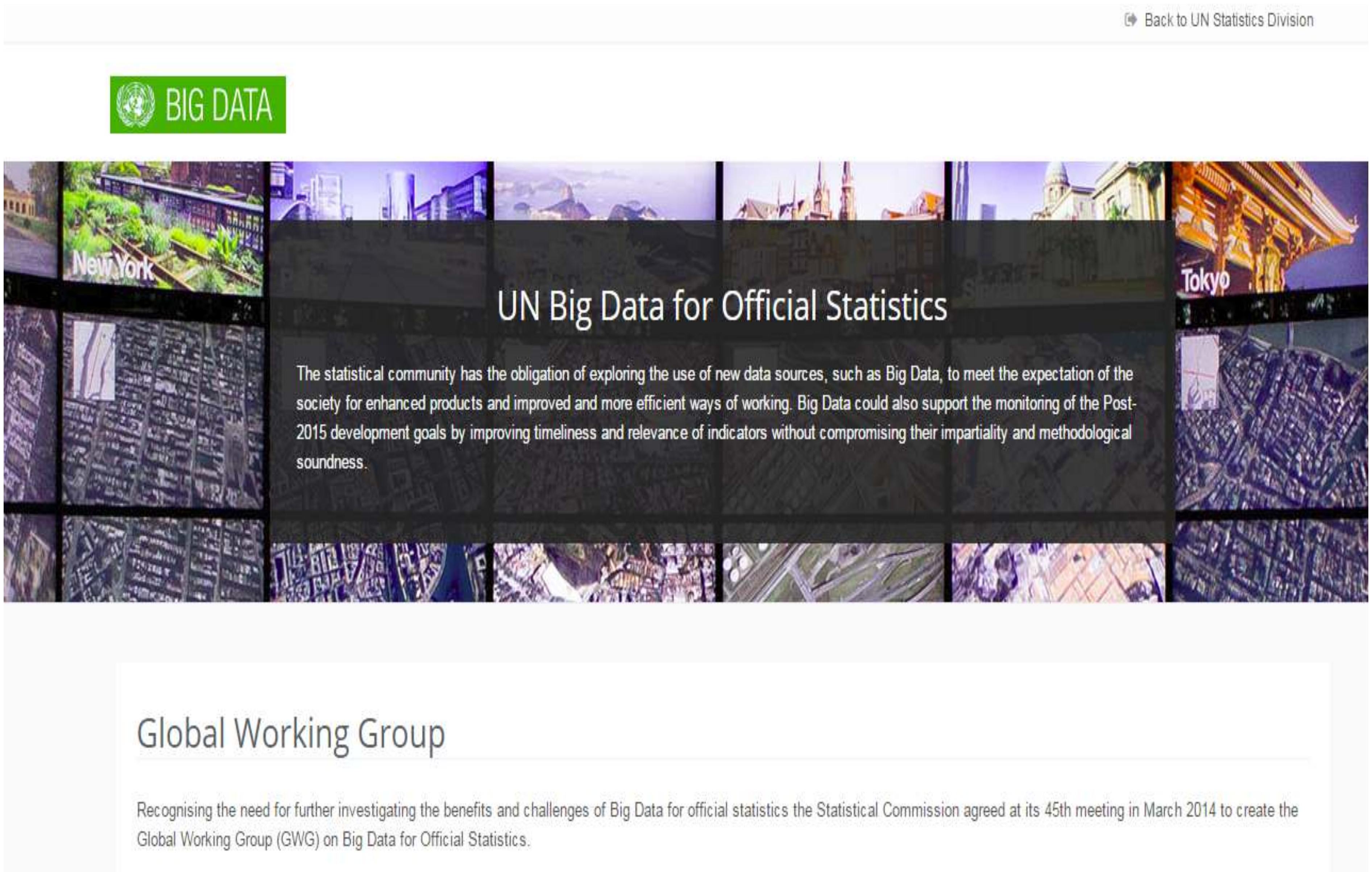
No direct relevance for tourism statistics

Why are Big Data important?

- ✓ **Big Data can keep official statistics relevant** – private sector moves fast
- ✓ **Big Data are part of modernization of statistical systems** – new production processes and partnerships
- ✓ **Big Data can help core national statistics** – for integrated economic, social and environmental policies
- ✓ **Big Data can help meeting the data demand of the 2030 agenda** – monitoring policies – "leave no one behind"
- ✓ **Big Data are needed for agile statistics** – for emergency issues

UN Global Working Group on Big Data for Official Statistics

BigData
UN Global Working Group



UN Big Data for Official Statistics

The statistical community has the obligation of exploring the use of new data sources, such as Big Data, to meet the expectation of the society for enhanced products and improved and more efficient ways of working. Big Data could also support the monitoring of the Post-2015 development goals by improving timeliness and relevance of indicators without compromising their impartiality and methodological soundness.

Global Working Group

Recognising the need for further investigating the benefits and challenges of Big Data for official statistics the Statistical Commission agreed at its 45th meeting in March 2014 to create the Global Working Group (GWG) on Big Data for Official Statistics.

Big Data for Official Statistics

Drivers:

- **Availability of automatically generated data** in electronic format, such as mobile phone, social media, electronic commercial transactions, sensor networks, smart meters, GPS tracking device, or satellite images
- **Higher frequency, more granularity, wider coverage, lower cost for data collection**
- **Modernisation of statistical production and services & the 2030 Agenda for sustainable development**

United Nations Global Working Group on Big Data for Official Statistics

- Created in March 2014
- 44 members (28 countries and 16 international agencies)
- 8 Task Teams
- Coordination of the work of the TTs
- Preparation of meetings, including international conferences
- UN Global Platform
- Reporting to the United Nations Statistical Commission

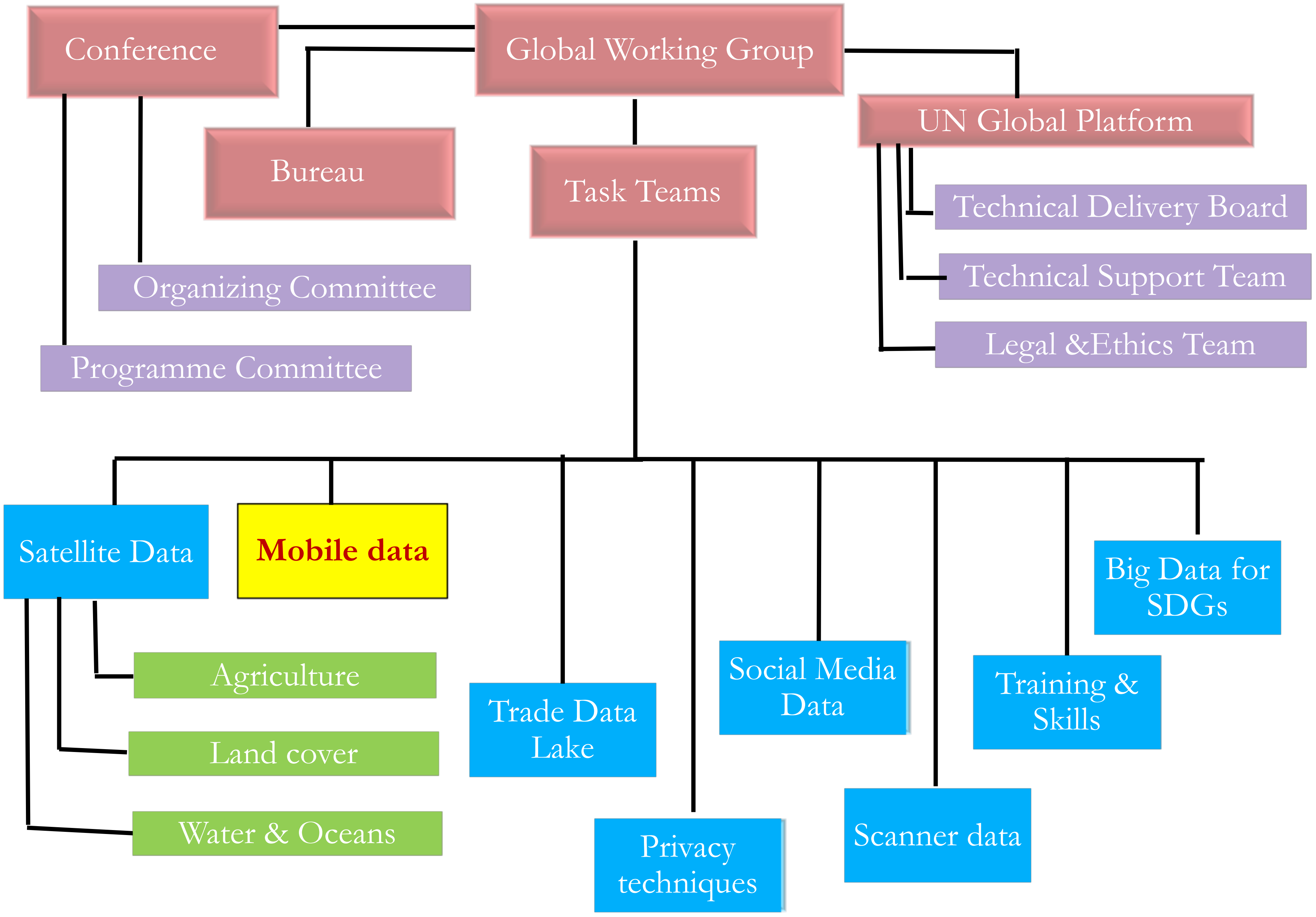
Composition of the GWG

Countries:

- Australia, Bangladesh, Brazil, Cameroon, Canada, China, Colombia, Denmark, Egypt, Georgia, Germany, Indonesia, Ireland, Italy, Mexico, Morocco, Netherlands, Oman, Pakistan, Philippines, Poland, Republic of Korea, Saudi Arabia, Switzerland, UAE, UK, Tanzania, US

Organizations:

- AfDB, CARICOM, Eurostat, FAO, IMF, OECD, GCC-Stat, ITU, UN GP, UNECA, UNECE, UNESCAP, UN SIAP, UNSD, UPU, WB



Conference

Global Working Group

UN Global Platform

Bureau

Task Teams

Organizing Committee

Programme Committee

Technical Delivery Board

Technical Support Team

Legal &Ethics Team

Satellite Data

Mobile data

Agriculture

Land cover

Water & Oceans

Trade Data Lake

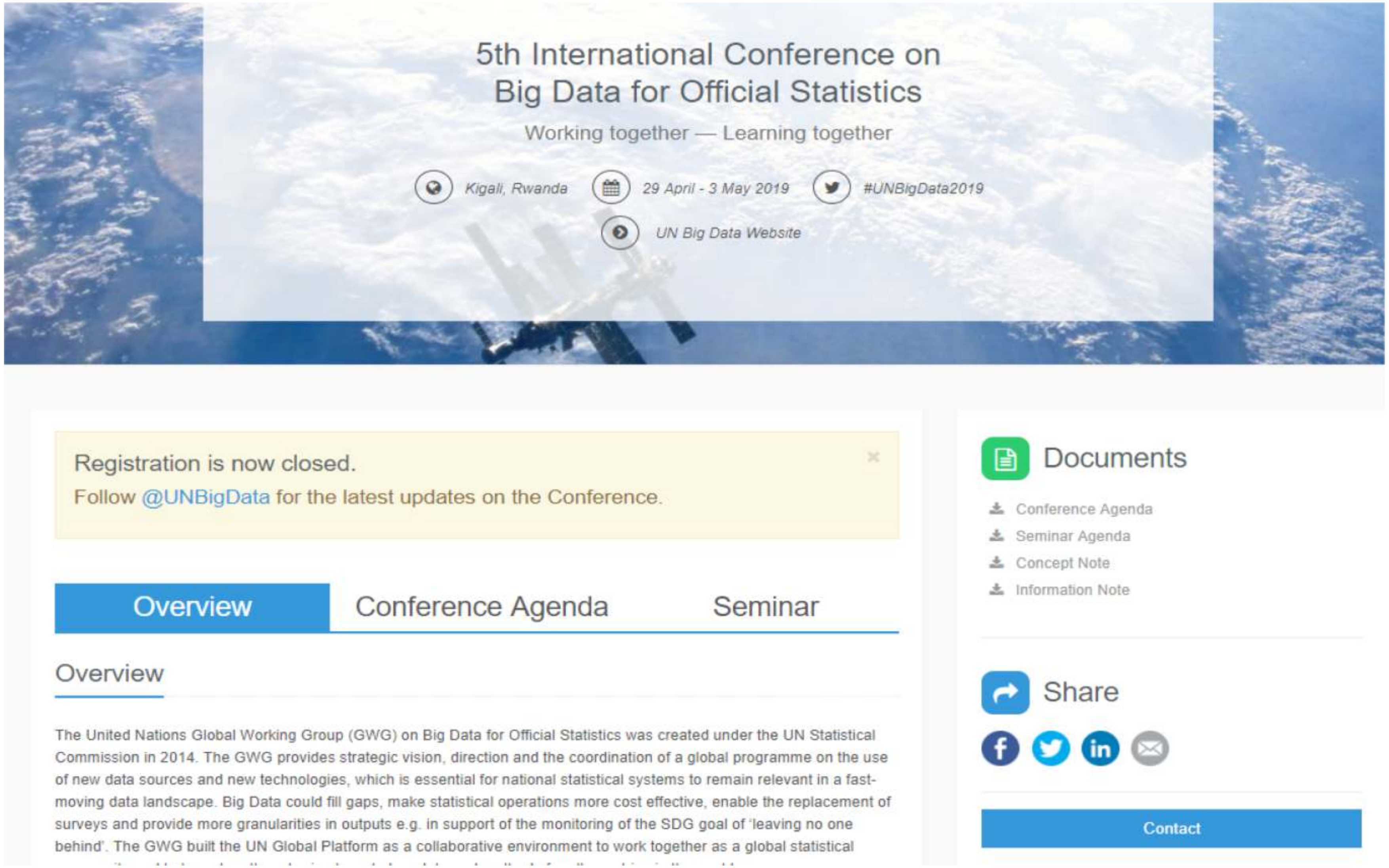
Privacy techniques

Social Media Data

Scanner data

Training & Skills

Big Data for SDGs



5th International Conference on Big Data for Official Statistics

Working together — Learning together

Kigali, Rwanda · 29 April - 3 May 2019 · #UNBigData2019

UN Big Data Website

Registration is now closed.
Follow @UNBigData for the latest updates on the Conference.

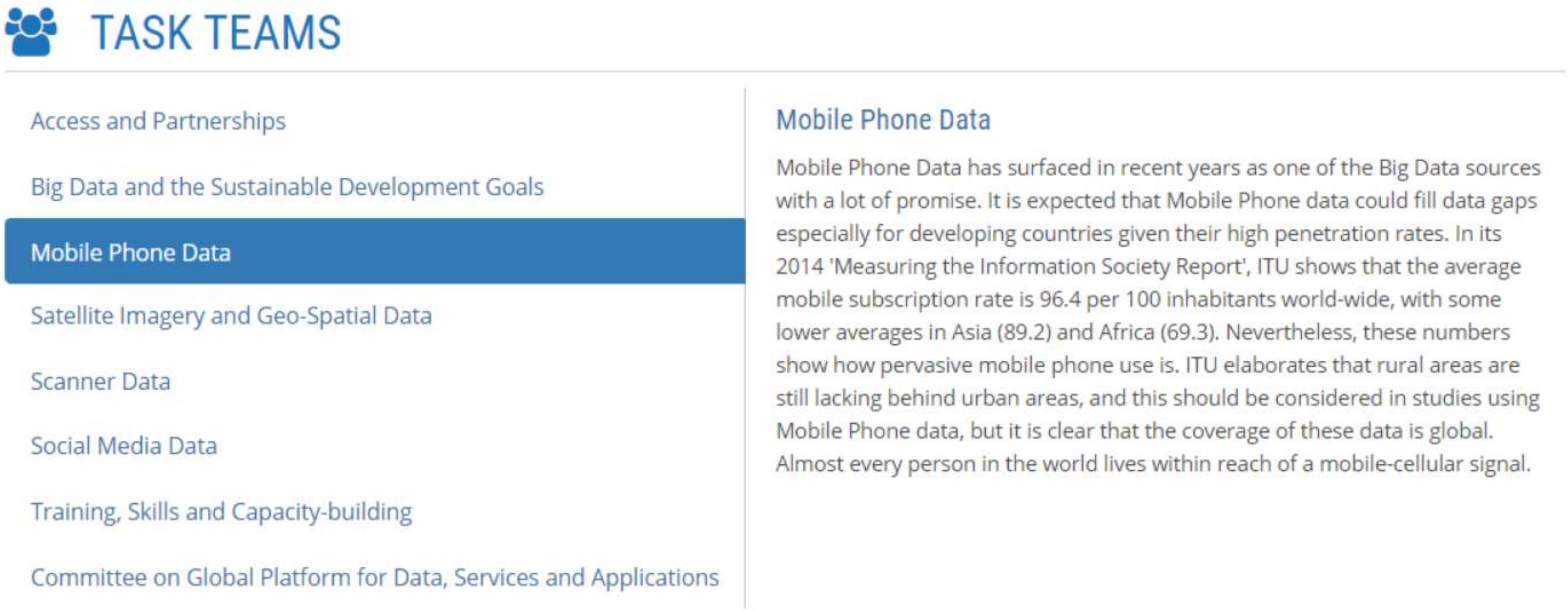
Overview

The United Nations Global Working Group (GWG) on Big Data for Official Statistics was created under the UN Statistical Commission in 2014. The GWG provides strategic vision, direction and the coordination of a global programme on the use of new data sources and new technologies, which is essential for national statistical systems to remain relevant in a fast-moving data landscape. Big Data could fill gaps, make statistical operations more cost effective, enable the replacement of surveys and provide more granularities in outputs e.g. in support of the monitoring of the SDG goal of 'leaving no one behind'. The GWG built the UN Global Platform as a collaborative environment to work together as a global statistical

Documents

- Conference Agenda
- Seminar Agenda
- Concept Note
- Information Note

Share



TASK TEAMS

- Access and Partnerships
- Big Data and the Sustainable Development Goals
- **Mobile Phone Data**
- Satellite Imagery and Geo-Spatial Data
- Scanner Data
- Social Media Data
- Training, Skills and Capacity-building
- Committee on Global Platform for Data, Services and Applications

Mobile Phone Data

Mobile Phone Data has surfaced in recent years as one of the Big Data sources with a lot of promise. It is expected that Mobile Phone data could fill data gaps especially for developing countries given their high penetration rates. In its 2014 'Measuring the Information Society Report', ITU shows that the average mobile subscription rate is 96.4 per 100 inhabitants world-wide, with some lower averages in Asia (89.2) and Africa (69.3). Nevertheless, these numbers show how pervasive mobile phone use is. ITU elaborates that rural areas are still lacking behind urban areas, and this should be considered in studies using Mobile Phone data, but it is clear that the coverage of these data is global. Almost every person in the world lives within reach of a mobile-cellular signal.

Handbook on the use of mobile phone data for official statistics – draft version is available at:

https://unstats.un.org/bigdata/taskteams/mobilephone/Handbook%20on%20Mobile%20Phone%20Data%20for%20official%20statistics%20-%20Draft%20Nov%202017.pdf

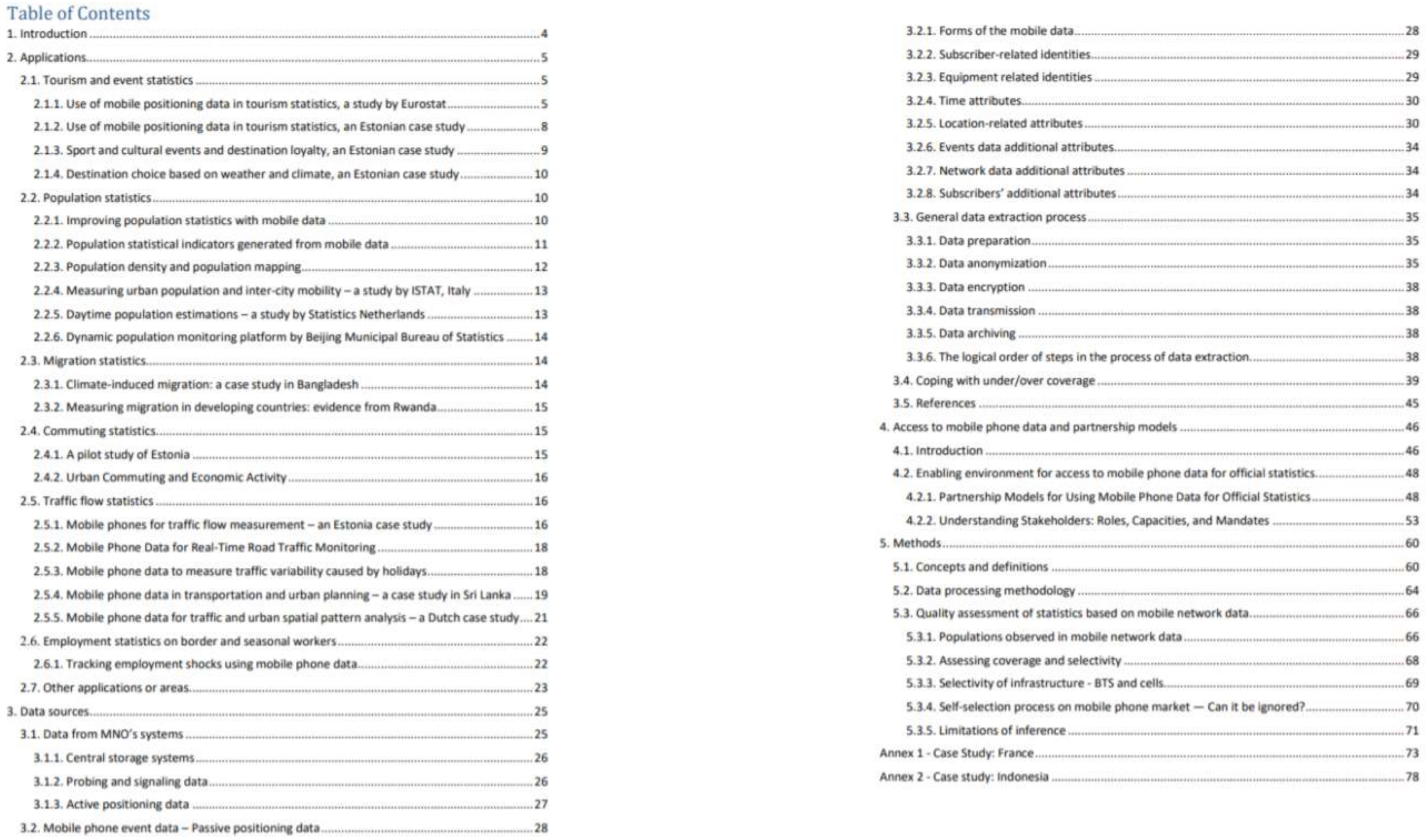
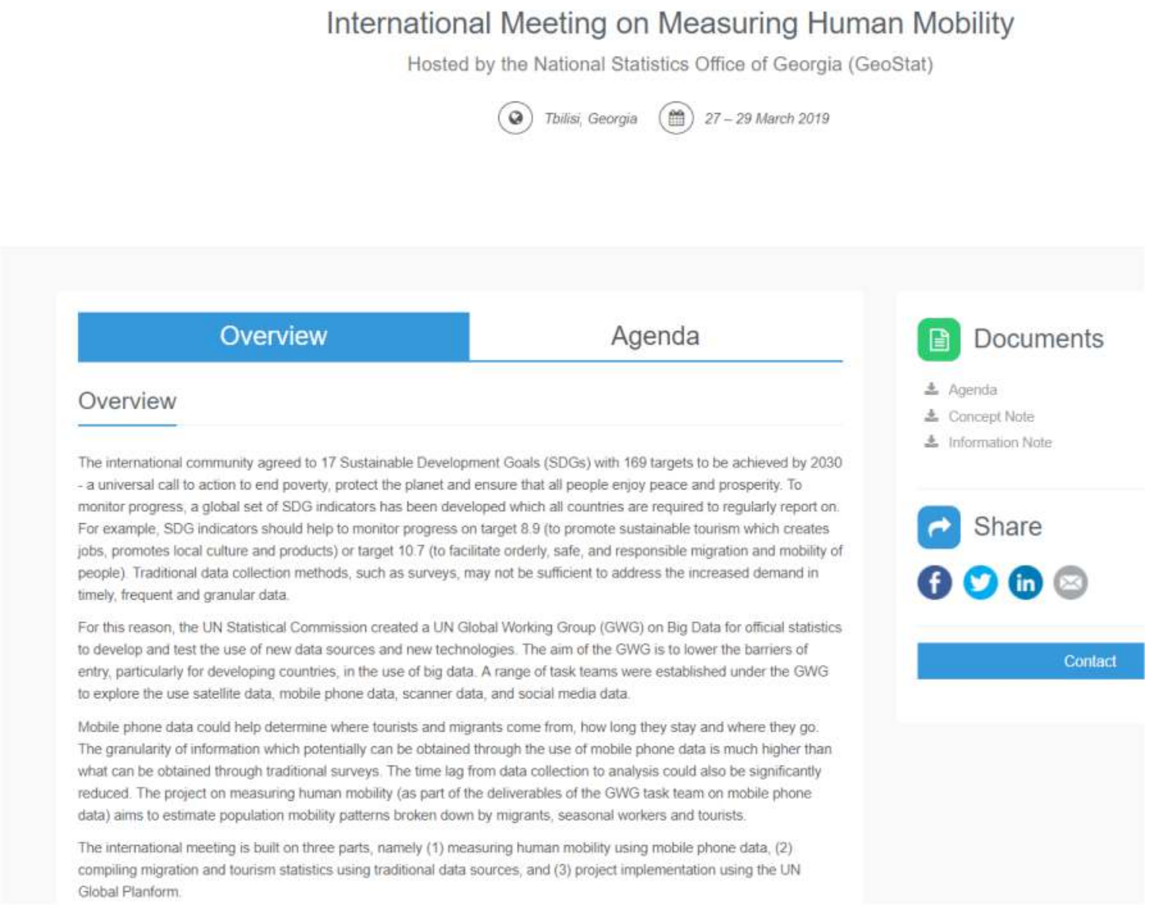


Table of Contents

UNSD project on measuring human mobility with using mobile phone data

https://unstats.un.org/bigdata/events/2019/tbilisi/default.asp



International Meeting on Measuring Human Mobility

Hosted by the National Statistics Office of Georgia (GeoStat)

Tbilisi, Georgia *27 – 29 March 2019*

Overview

The international community agreed to 17 Sustainable Development Goals (SDGs) with 169 targets to be achieved by 2030 - a universal call to action to end poverty, protect the planet and ensure that all people enjoy peace and prosperity. To monitor progress, a global set of SDG indicators has been developed which all countries are required to regularly report on. For example, SDG indicators should help to monitor progress on target 8.9 (to promote sustainable tourism which creates jobs, promotes local culture and products) or target 10.7 (to facilitate orderly, safe, and responsible migration and mobility of people). Traditional data collection methods, such as surveys, may not be sufficient to address the increased demand in timely, frequent and granular data.

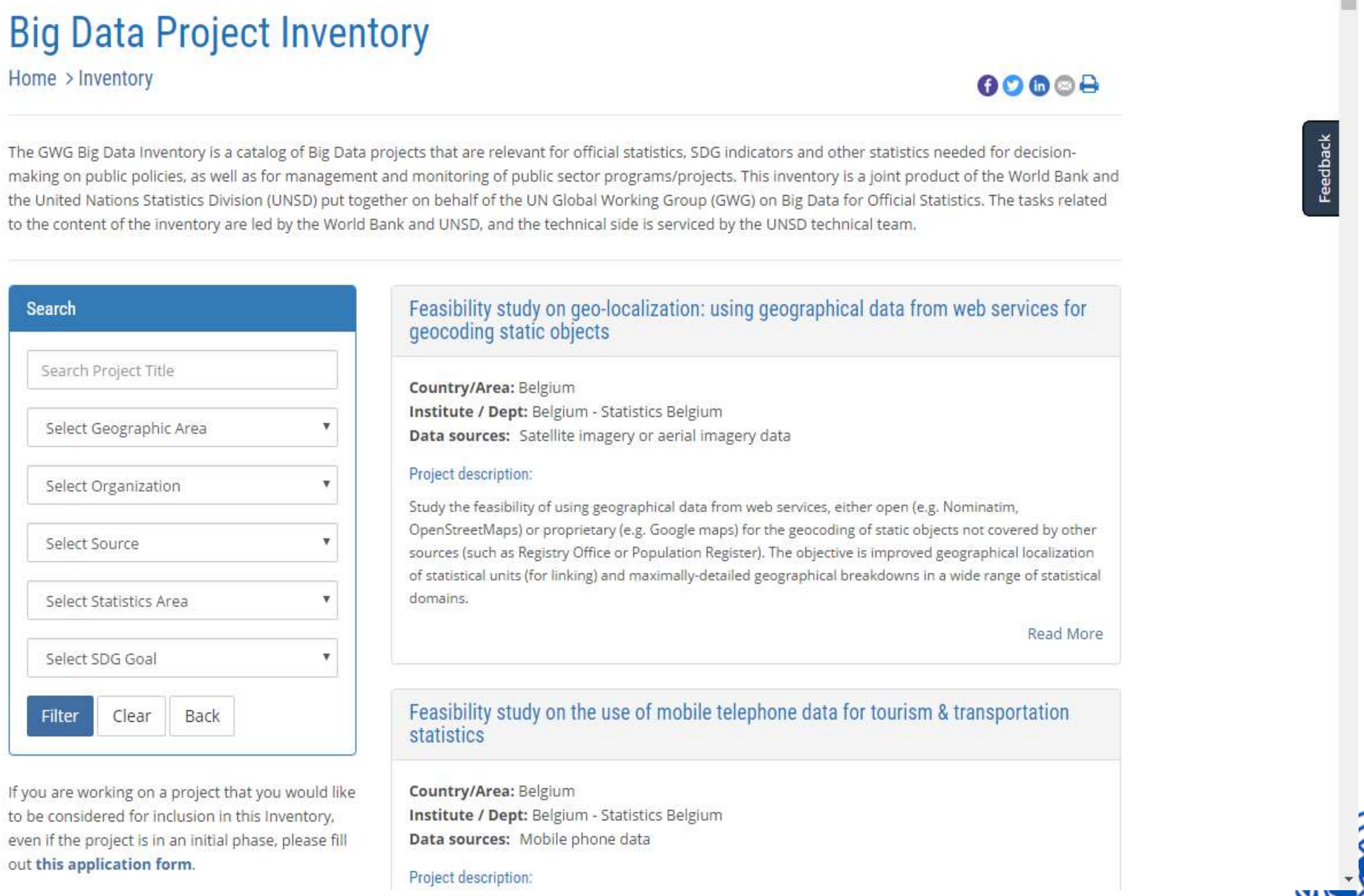
For this reason, the UN Statistical Commission created a UN Global Working Group (GWG) on Big Data for official statistics to develop and test the use of new data sources and new technologies. The aim of the GWG is to lower the barriers of entry, particularly for developing countries, in the use of big data. A range of task teams were established under the GWG to explore the use satellite data, mobile phone data, scanner data, and social media data.

Mobile phone data could help determine where tourists and migrants come from, how long they stay and where they go. The granularity of information which potentially can be obtained through the use of mobile phone data is much higher than what can be obtained through traditional surveys. The time lag from data collection to analysis could also be significantly reduced. The project on measuring human mobility (as part of the deliverables of the GWG task team on mobile phone data) aims to estimate population mobility patterns broken down by migrants, seasonal workers and tourists.

The international meeting is built on three parts, namely (1) measuring human mobility using mobile phone data, (2) compiling migration and tourism statistics using traditional data sources, and (3) project implementation using the UN Global Planform.

Within the next 18 months, the Task Team on the use of mobile phone data would like to achieve the following:

- Develop handbook, training materials, e-learning course and update guidelines on using mobile phone data for official statistics
- Document and further develop methodologies and algorithms on using mobile phone data for statistical applications (Tourism statistics, Migration statistics, Population density statistics)
- Develop methodologies on using mobile phone data for quality checks and getting complementary information on SDG indicators
- Organize project meeting on the use of mobile phone data to measure human mobility, Tbilisi, Georgia, March 2018
- Organize regional workshop in Indonesia, June 2019



Big Data Project Inventory

Home > Inventory

The GWG Big Data Inventory is a catalog of Big Data projects that are relevant for official statistics, SDG indicators and other statistics needed for decision-making on public policies, as well as for management and monitoring of public sector programs/projects. This inventory is a joint product of the World Bank and the United Nations Statistics Division (UNSD) put together on behalf of the UN Global Working Group (GWG) on Big Data for Official Statistics. The tasks related to the content of the inventory are led by the World Bank and UNSD, and the technical side is serviced by the UNSD technical team.

If you are working on a project that you would like to be considered for inclusion in this Inventory, even if the project is in an initial phase, please fill out **this application form**.

Feasibility study on geo-localization: using geographical data from web services for geocoding static objects

Country/Area: Belgium
Institute / Dept: Belgium - Statistics Belgium
Data sources: Satellite imagery or aerial imagery data

Project description:

Study the feasibility of using geographical data from web services, either open (e.g. Nominatim, OpenStreetMaps) or proprietary (e.g. Google maps) for the geocoding of static objects not covered by other sources (such as Registry Office or Population Register). The objective is improved geographical localization of statistical units (for linking) and maximally-detailed geographical breakdowns in a wide range of statistical domains.

Read More

Feasibility study on the use of mobile telephone data for tourism & transportation statistics

Country/Area: Belgium
Institute / Dept: Belgium - Statistics Belgium
Data sources: Mobile phone data

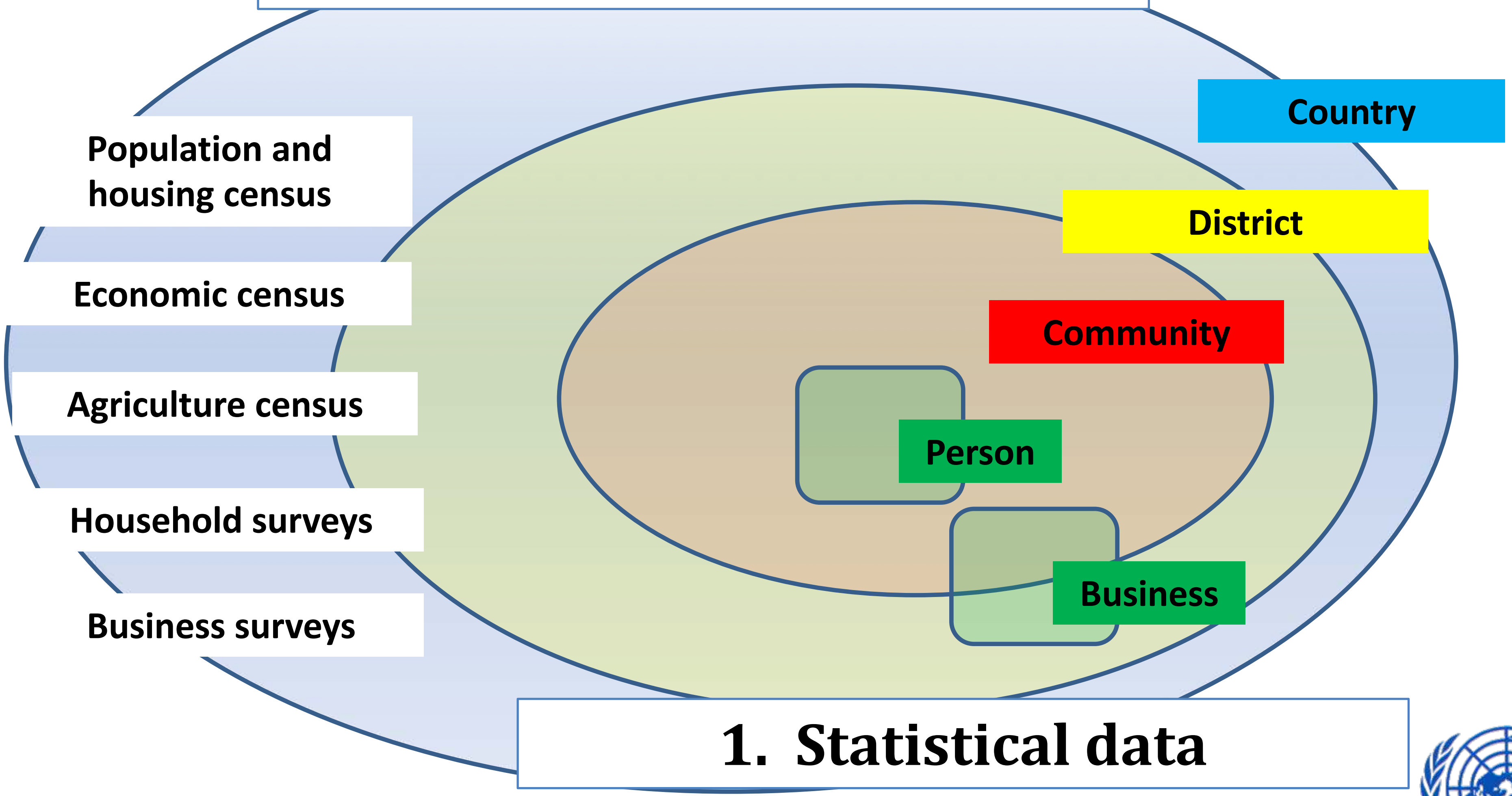
Project description:

Strong “outside” participation

- Positium,
- Telenor,
- IBM
- Google,
- Data Pop,
- World Pop,
- Flowminder,
- Orange,
- UNU-EHS,
- World Economic Forum,
- NASA,
- Harvard

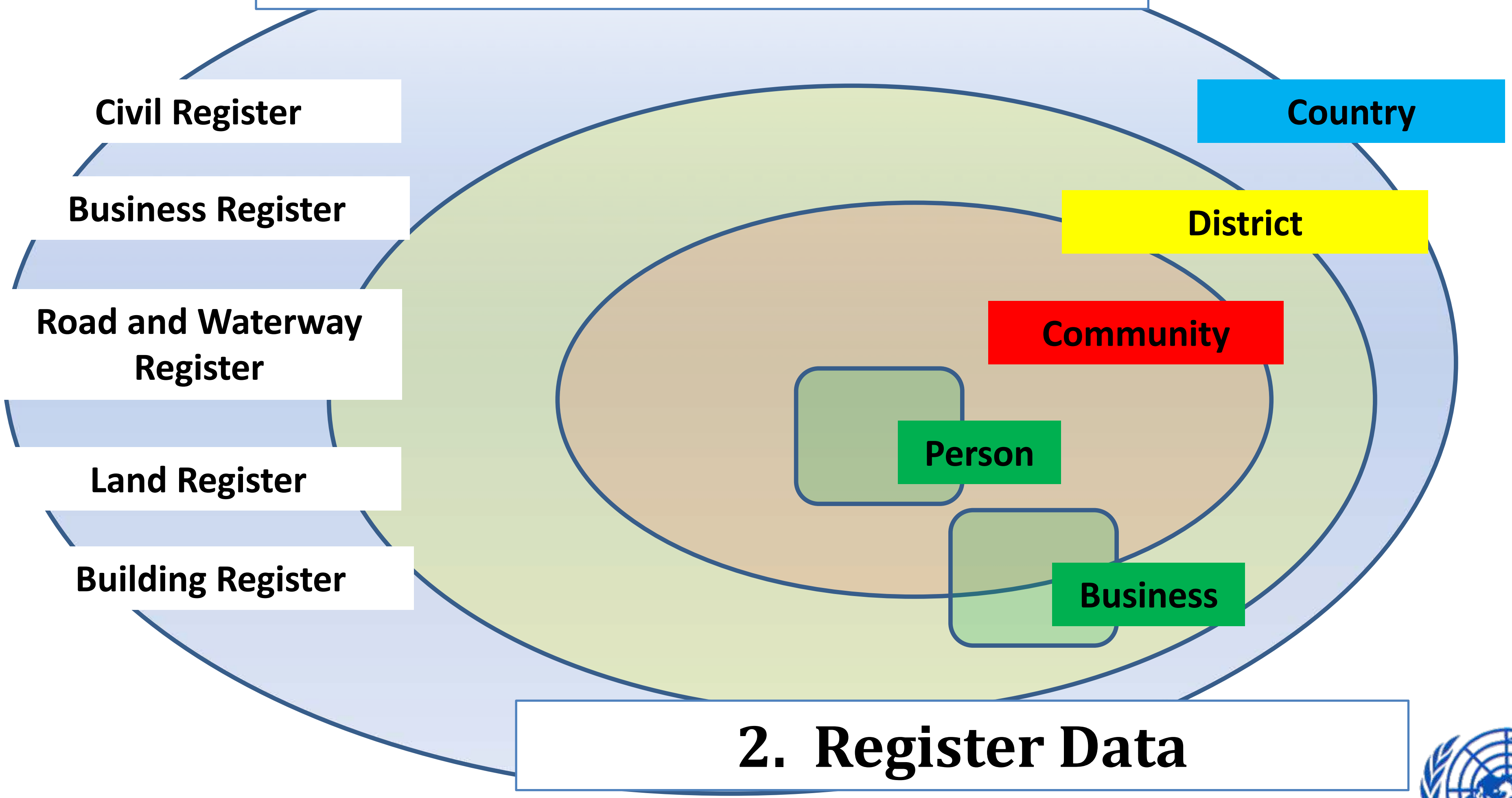
Statistical Data Infrastructure

Statistical Data Infrastructure



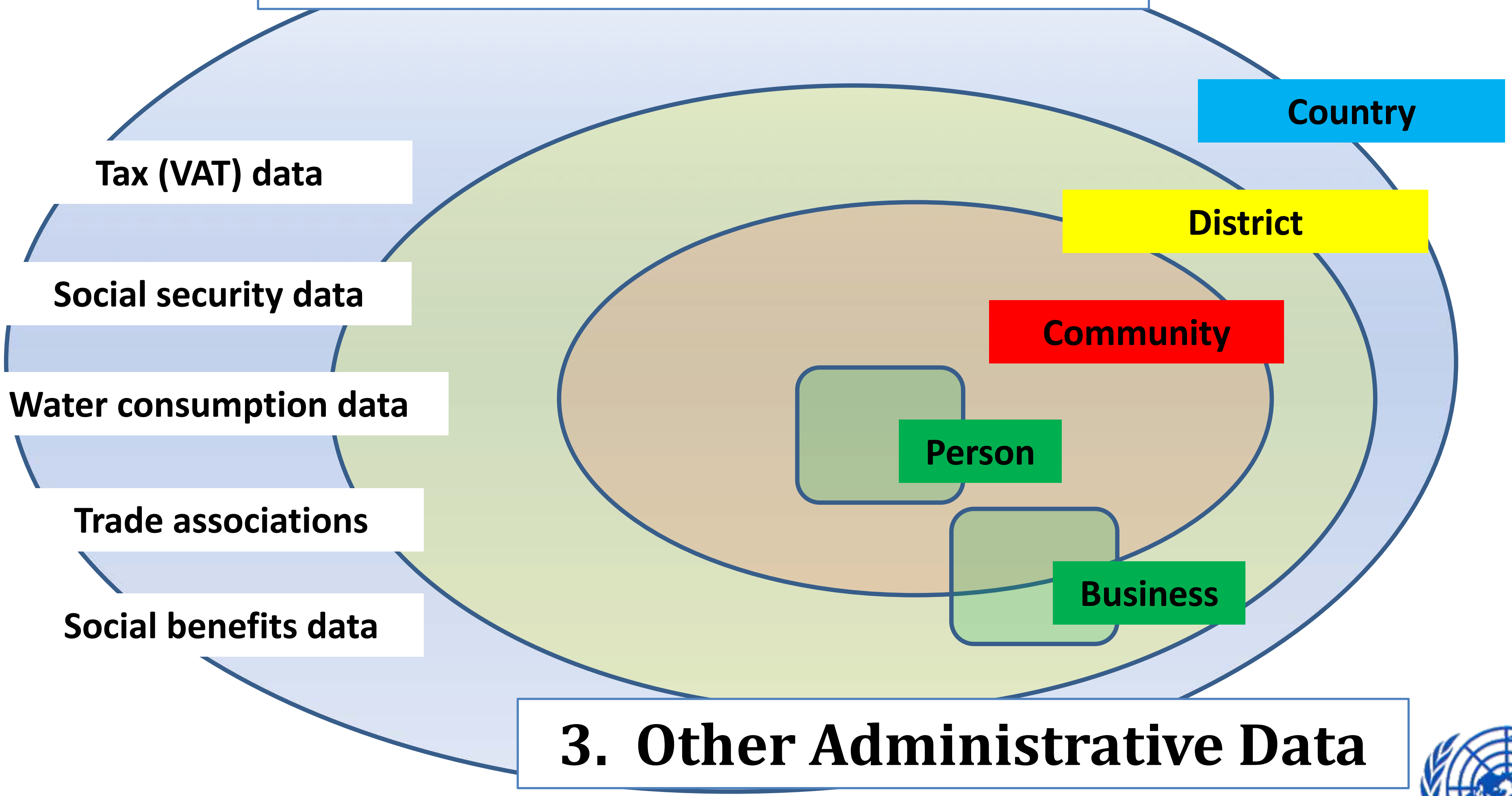
1. Statistical data

Statistical Data Infrastructure



2. Register Data

Statistical Data Infrastructure



Tax (VAT) data

Social security data

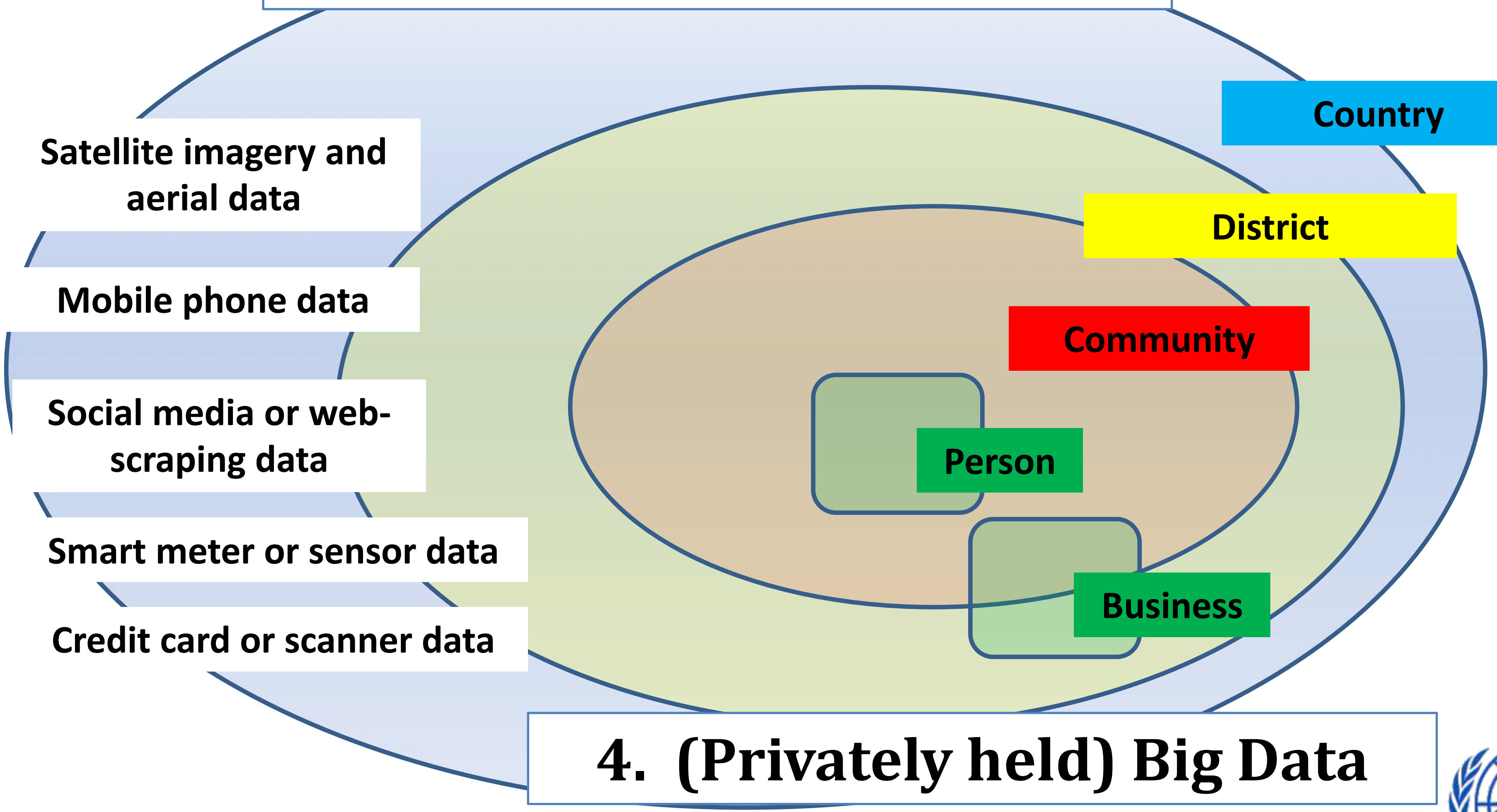
Water consumption data

Trade associations

Social benefits data

3. Other Administrative Data

Statistical Data Infrastructure

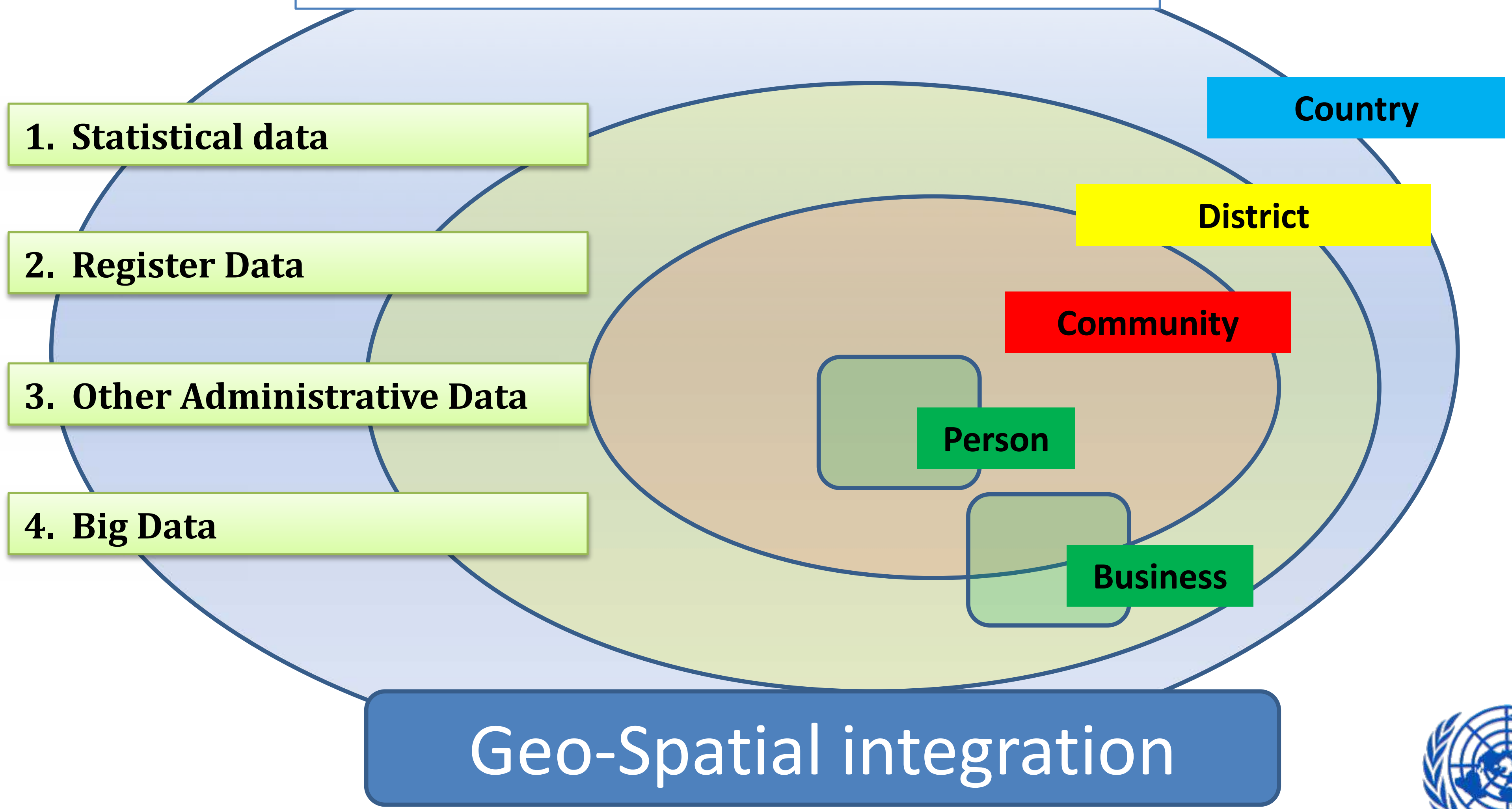


Satellite imagery and aerial data
Mobile phone data
Social media or web-scraping data
Smart meter or sensor data
Credit card or scanner data
Country
District
Community
Person
Business

4. (Privately held) Big Data

BigData
UN Global Working Group

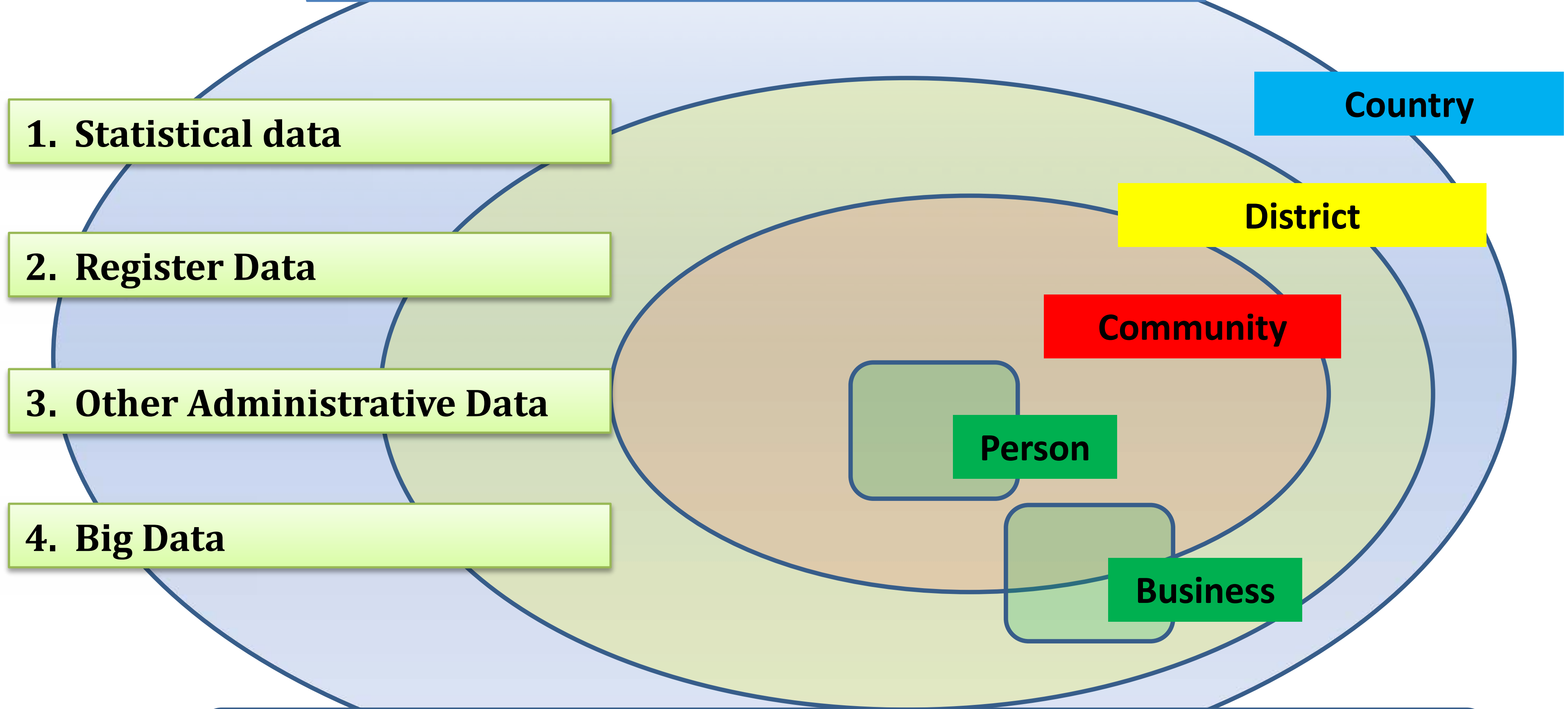
Statistical Data Infrastructure



1. Statistical data
2. Register Data
3. Other Administrative Data
4. Big Data

Geo-Spatial integration

Statistical Data Infrastructure



1. Statistical data
2. Register Data
3. Other Administrative Data
4. Big Data

Country

District

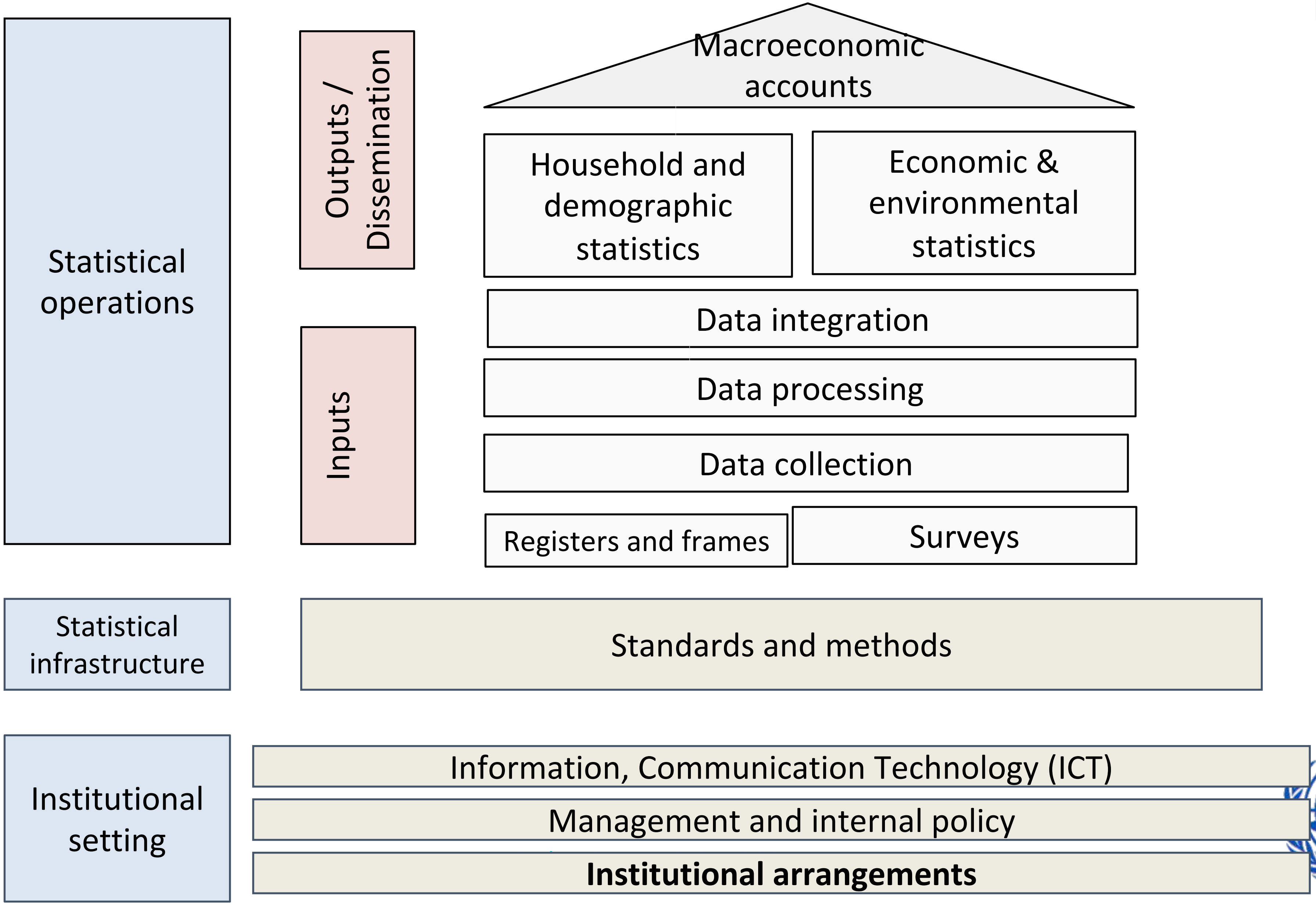
Community

Person

Business

Statistical integration: SNA and SEEA

Integrated statistics approach



Statistical operations

Outputs / Dissemination

Inputs

Macroeconomic accounts

Household and demographic statistics

Economic & environmental statistics

Data integration

Data processing

Data collection

Registers and frames

Surveys

Statistical infrastructure

Standards and methods

Institutional setting

Information, Communication Technology (ICT)

Management and internal policy

Institutional arrangements

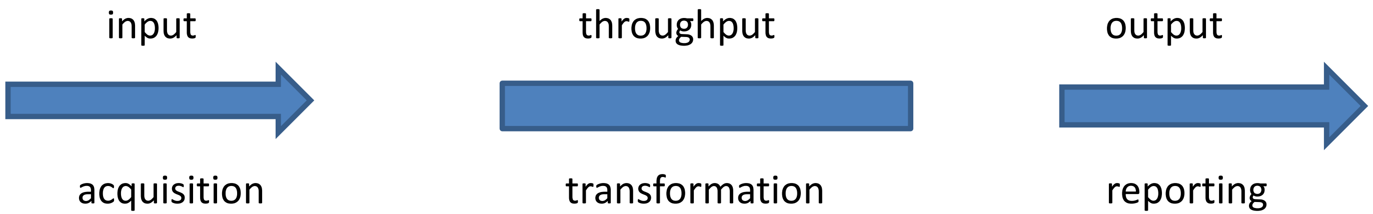
Quality framework for big data

Framework for NSO to assess quality of big data

General approach

Quality: to be evaluated in the of intended use (‘fitness for use’)

Generic statistical business process model:



input

acquisition

throughput

transformation

output

reporting

Framework: For each phase define appropriate quality dimensions and quality indicators

Framework for NSO to assess quality of big data

Hyperdimensions

The concept of hyperdimension was taken from the Netherland administrative data quality framework.

- Source: Related to the type of data, the entity from which the data is obtained, and how it is administered and regulated.
- Metadata: Description of concepts, file contents, and processes.
- Data: Related to quality of the data itself.

Framework for NSO to assess quality of big data

Quality dimensions

- Institutional/business environment
- Privacy and security, complexity
- Completeness, usability, time factor
- Accuracy
  - selectivity
- Coherence
  - linkability
- Validity
- Accessibility, clarity, relevance

Hyperdi-mension	Quality Dimension	Factors to consider
Source	**Institutional Environment**	Sustainability of the entity-data provider Reliability status, transparency, interpretability
	Privacy and Security	Legislation, Data Keeper vs. Data provider Restrictions, Perception
Metadata	**Complexity**	Technical constraints, Sructured or Unstructured Readability, Presence of hierarchies and nesting
	Completeness	Metadata is available, interpretable and complete
	Usability	Resources required to import and analyse Risk analysis
	Time-related	Timeliness, Periodicity, Changes through time
	Linkability	Presence and quality of linking variables
	Coherence	Use of standards
	Validity	Transparency of methods and processes Soundness of methods and processes

Thank you!
Murakoze!

Karoly Kovacs

United Nations Statistics Division | Department of Economic and Social Affairs

Email: bigdata@un.org

http://unstats.un.org/unsd/bigdata