Harnessing Big Data for Statistical Purposes in Uganda

Bernard Justus MUHWEZI, Manager, Geo-Information Services, Uganda Bureau of Statistics

Kirsten Van CAMP, Programme Officer, UN Pulse Lab Kampala

Martin MUBANGIZI, Data Scientist, UN Pulse Lab Kampala

Content

- Introduction
- UN Pulse Labs
  - UN Pulse Lab Kampala
- Use Case Pilot Projects
  - Monitoring poverty
  - Real Time Analysis of Radio Content
- Data Privacy
- Other pilot areas identified
- Challenges
- Collaborating partners

Introduction

- Uganda recognizes and is keen on the importance of Big Data in data production
  - Mobile phone data
  - Satellite imagery or aerial imagery data
  - Social media –radio data
  - Credit card data
  - Smart meter electricity data
  - Road sensor data
  - Public transport usage data
  - Ships identification data
  - Health records data
  - Web scraping data
  - Scanner data

Introduction…

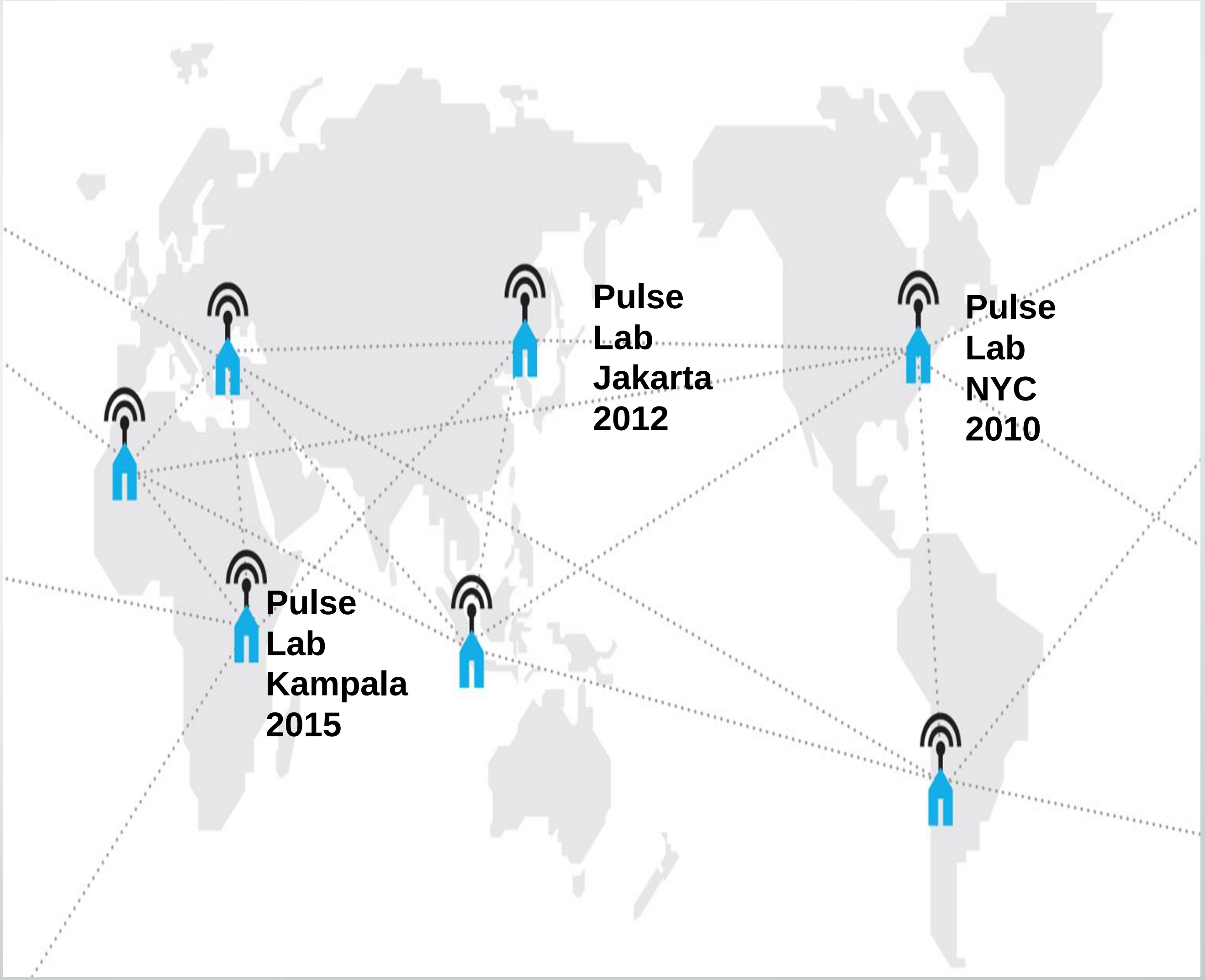
- Big data can be described by the following characteristics (Wiki definition):
  - Volume
    - The quantity of generated and stored data.
    - The size of the data determines the value and potential insight, and
    - Whether it can be considered Big Data or not
  - Variety
    - The type and nature of the data.
    - This helps people who analyze it to effectively use the resulting insight.
    - Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion
  - Velocity
    - In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. Big data is often available in real-time
  - Variability
    - Inconsistency of the data set can hamper processes to handle and manage it
  - Veracity
    - The data quality of captured data can vary greatly, affecting the accurate analysis

Introduction…

- Big data and statistics
  - Complementarity should be the main interest, not replacement: this will allow us to preserve accuracy and improve efficiency, whilst demonstrating actual efficacy (Use statistics to demonstrate which Big Data analyses / methods are reliable and which might not be)
  - Adding now-casting and forecasting to traditional descriptive statistics
  - Potential for statistical modelling (if data revolution continues to grow exponentially and inclusively) through the use of Big Data
  - Collaborate to remove biases and create reliable Big Data initiatives
- Big data and scientific standards - How do data analysts aspire for quality and reliable research?
  - Data curation
    - the process of turning independently created data sources (structured and semi-structured data) into unified data sets ready for analytics, using domain experts to guide the process. It involves: Identifying data sources of interest (whether from inside or outside the enterprise)
  - Use of the data

UN Pulse Lab

2010-2015



About UN Pulse Lab Kampala

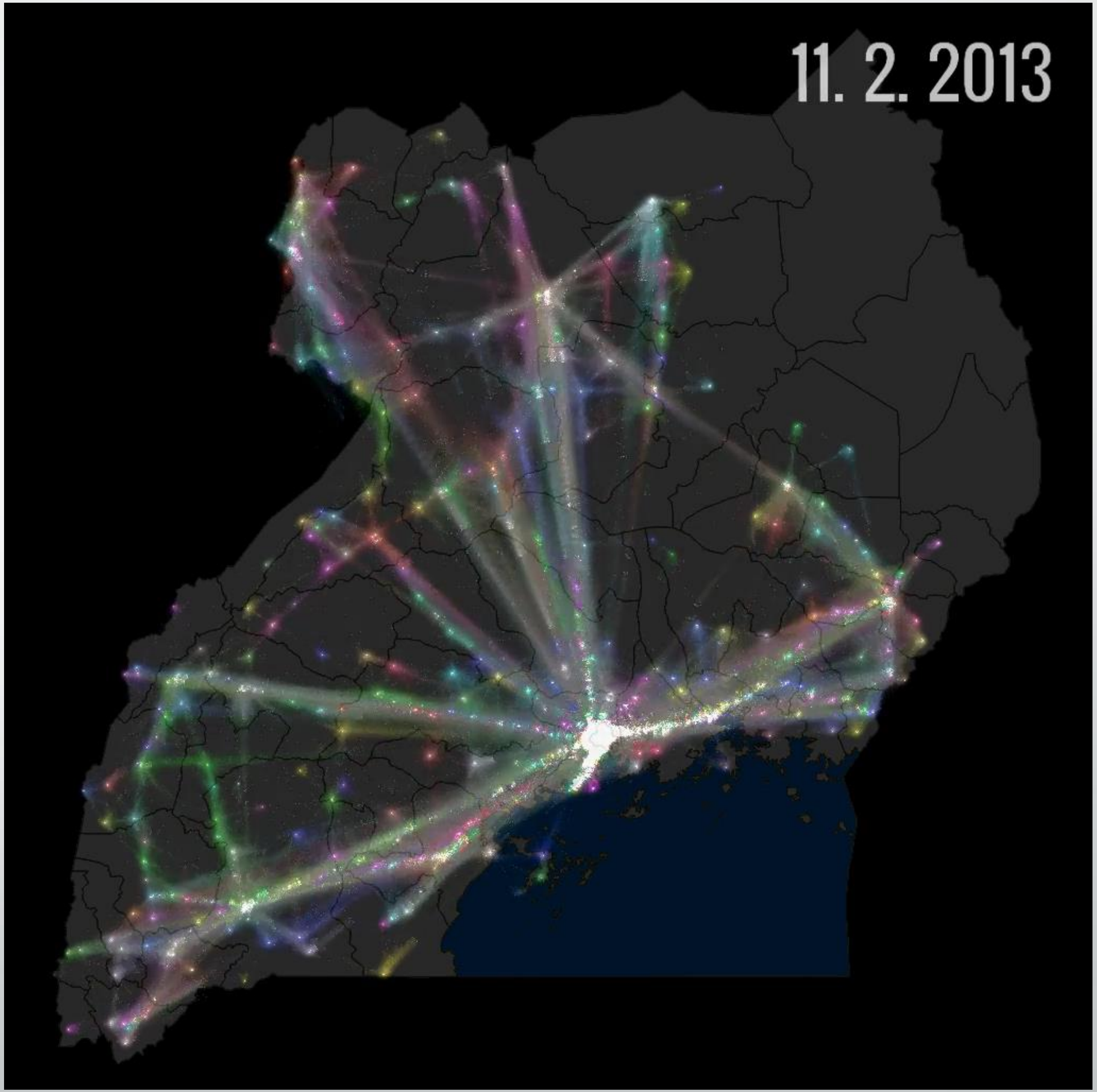
- Pulse Lab Kampala is the third Lab in the UN Global Pulse Network with other labs in New York and Jarkata.
- UN Global Pulse Lab is an initiative under the Executive Office of the UN Secretary General
- The vision of UN Global Pulse is a future in which Big Data is harnessed safely and responsibly as a public good.
- Its mission is to accelerate discovery, development and scaled adoption of Big Data innovation for sustainable development and humanitarian action
    - Big data has a lot of potential in supporting the work of government's statistical offices

About UN Pulse Lab Kampala…

- Pulse Lab Kampala promotes capacity building for youth in African through several initiatives including Data Science Africa, in which they are trained in data science skills applied to Big Data
    - Machine learning
    - Supervised learning
    - Bayesian techniques
    - Artificial intelligence
        - Neural networks
        - Decision trees
    - Data visualization methods
    - Traditional statistical methods*
- Promotes use of Big Data for official statistics, and appeals for collaboration across NSO in Africa for partnership

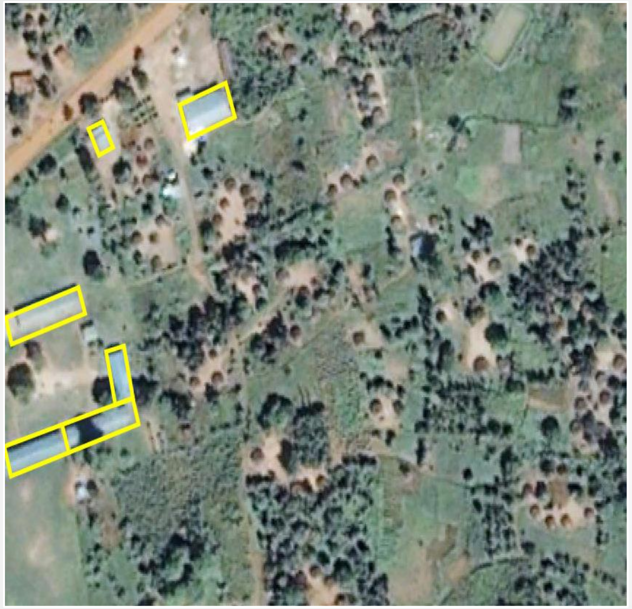
Telecoms Real Time Data

Population mobility matrix, generated from millions or billions of call details records (CDR) and presented in an easy way like in the slides, the lights representing the movement of people in and out of Kampala through out the Month of Feb 2013

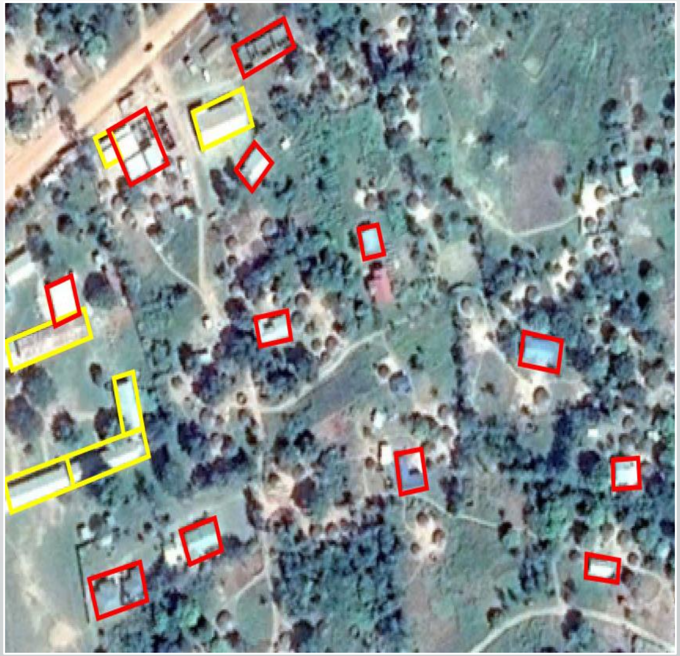


Monitoring Poverty

Using “machine counting” technique, Satellite imagery from 2012, with metal-roofed buildings highlighted in yellow., image from 2014, with new metal-roofed buildings highlighted in red.



Gulu district 2012



Gulu district 2014

Monitoring Poverty

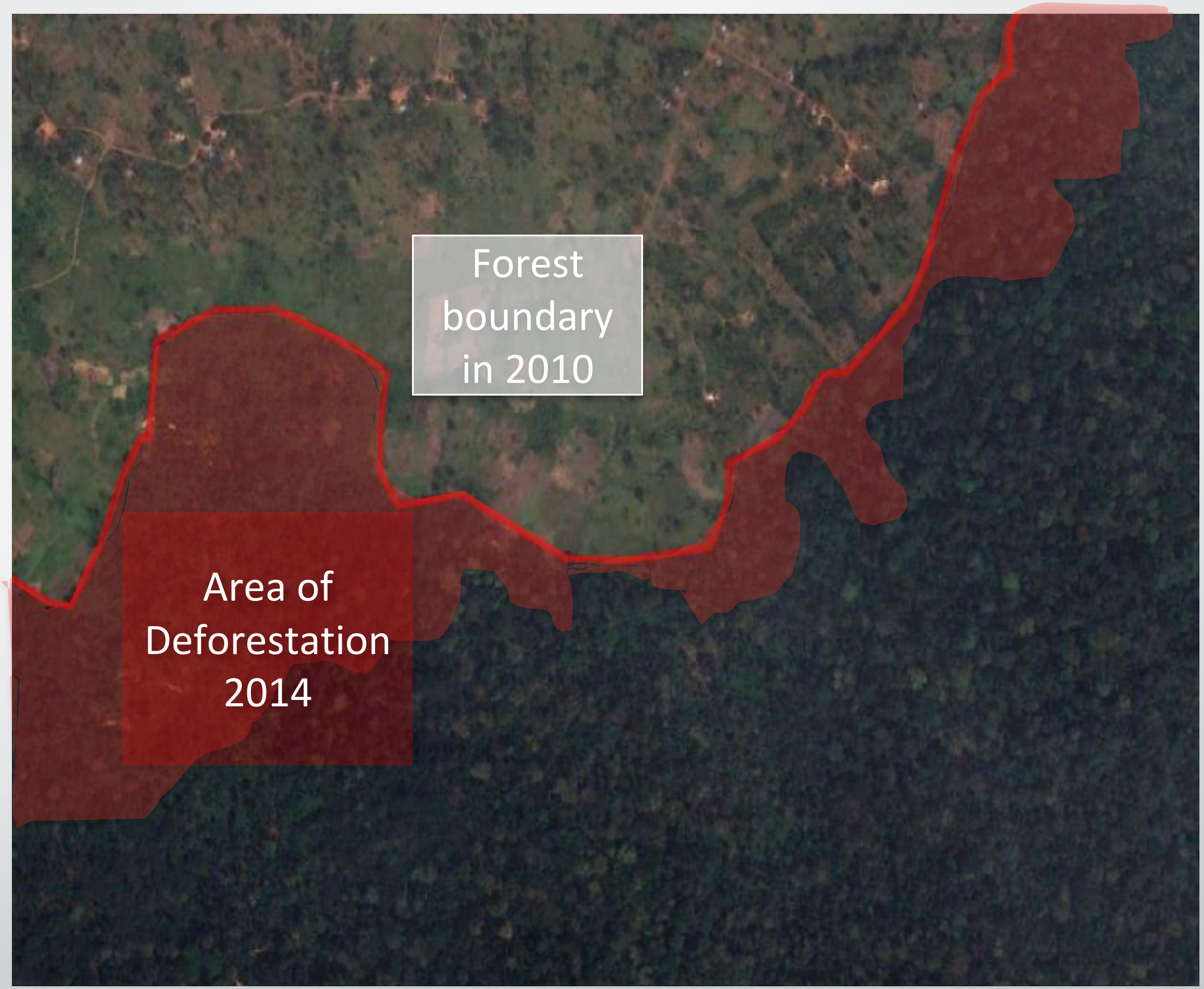
A tool through machine learning that automatically identifies and counts different types of roofs of households, as a proxy indicator of poverty in Uganda as in many other countries.

As the household economy improves, the thatched roof is changed to a metal one.

We can count metal roofs and thus, monitor poverty trends

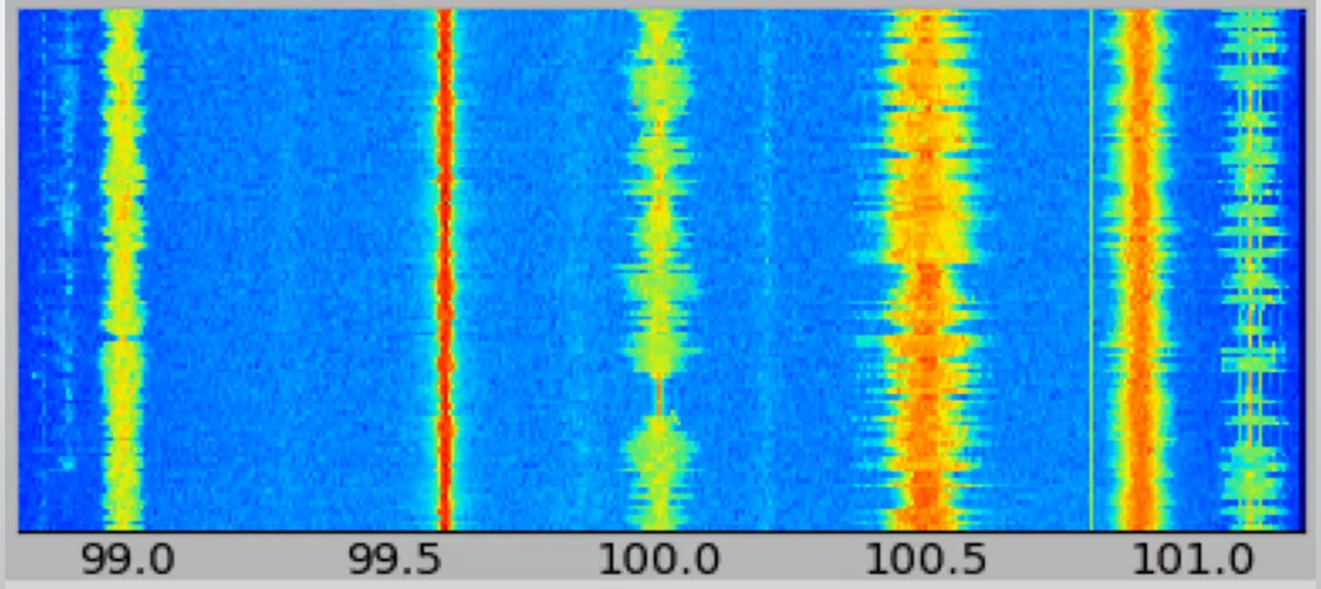
Monitoring Land Cover

A tool through machine learning that automatically identifies the areas depleted, usually requiring also some work of field verification to calibrate the tool



Real Time Analysis of Radio Content

The tool finds among all the public radio content the topics of interest relating to public service delivery for development



Data Privacy

Of major concern when using Big Data as they may be negative and positive impacts on individual or group of individuals.

There goal therefore is to minimize the risk to harm (negative impacts) and maximize the positive impacts of the project.

An assessment of these impact should be done before a Big Data project to insure that proper measures to mitigate the risk to the harms.

Other pilot areas identified

- Application for real-time monitoring of performance of ambulances
- Application to monitor, in real time, financing of the health sector
- Application for real-time analysis of radio content on health service delivery
- Application to inform development of policies and programme on Business Technical Vocational Education and Training
- Application for monitoring, in real time, budgeting and expenditure on education sector
- Application for real-time analysis of radio content to inform BTVET and teacher training programme

Challenges

- Inappropriate or limited legal framework
  - Not fully understandable and hard to document governing rules
- Human capacity
  - No appropriate skill to access and manage Big Data, meaning of limited analytics
- Methodological challenges
  - New areas of application of data science to be appreciated into main stream statistics
- Lack of adequate understanding of the analytical tools for specific data areas
- Access costs to datasets
  - Data not readily available
- Limited access to datasets
  - Data in the hands of other agencies that essentially private and bit confidential in nature
- Perceptions of Big Data by statisticians
  - Not convinced about its application for statistical purposes due to its non-traditional methods of use and analysis

Collaborating partners

- UN Pulse Lab, Kampala
- University of Edinburgh
- Uganda of Statistics
- Ministry of Health
- Ministry of Education
- National Planning Authority
- Belgian Technical Cooperation, Kampala Office
- National Information Technology Authority
- Telecommunication Companies in Uganda
- Uganda Communications Commission*

Other Referenced Examples

- Mobile phone data for tourism and transportation / for mobility and urban statistics
  - https://unstats.un.org/bigdata/inventory/?selectID=2015011
- Physical accessibility and remoteness
  - https://unstats.un.org/bigdata/inventory/?selectID=201422
  - http://www.data2x.org/wp-content/uploads/2017/03/Big-Data-and-the-Well-Being-of-Women-and-Girls.pdf
  - In Uganda Pulse Lab Kampala researched the access of populations to clean water, gaining insights on SDG10 with utility data (2017
- Use of satellite imagery to obtain geographical information
  - https://unstats.un.org/bigdata/inventory/?selectID=2015061
- Exploring the use of social media messages for economic indicators
  - https://unstats.un.org/bigdata/inventory/?selectID=2015064
  - In Uganda social media messages were used to analyze perceptions towards different types of contraception

2018 Data Science Africa

The Data Science Africa Workshop shall be conducted in Nyeri Town, Kenya, with a theme “Data for Change in Africa”

A week long training of trainers and two day workshop, Nyeri Kenya (31st May – 8th June 2018, and Data Science Africa 2018 Abuja with 3 days summer school and two day workshop, 12th -16th November 2018.

You are invited.