Filling Information Gaps in Big Data Modelling of Sustainable Development Goals

KASSIM S. MWITONDI (PHD)
Sheffield Hallam University
College of Business, Technology & Engineering (BTE)
Industry & Innovation Research Institute (IIRI)
9410 Cantor Building, City Campus
153 Arundel Street
Sheffield, S1 2NU
United Kingdom

k.mwitondi@shu.ac.uk
https://www.shu.ac.uk/about-us/our-people/staff-profiles/kassim-mwitondi
https://uk.linkedin.com/in/kassim-mwitondi-9602091b
Twitter: @Mwitondi

Sustainable Development Goals

The 17 Sustainable Development Goals (SDGs) were signed up by 193 United Nations member states in 2015, as the blueprint for achieving a better and more sustainable future for mankind and planet earth span across various aspects of life. They are defined with 169 Targets and to be tracked by 232 unique Indicators.

https://sdgs.un.org/goals

- The document of indicators state that the indicators *"…should be disaggregated, where relevant, by income, sex, age, race, ethnicity, migratory status, disability and geographic location, or other characteristics…"*
- That aggregation is not exhaustive. Practically, it widens the scope of the indicators, creating new variables, many of which can only be uncovered within specific problem space and require interdisciplinary approach to unravel.
- Viewing each SDG as a source of Big Data is the way forward, but success will depend on our levels of data sharing data, interdisciplinary skills development, resources and making open science the norm.

EXISTING TOOLS FOR MONITORING SDGs

The SDG Atlas by the WBGroup

Millenium Institute another tool that plays a role in informing us of SDG attainment levels.

Our World in Data presents data across all available indicators from using official statistics from the UN SDG repository. It acknowledges that major data gaps still remain.

https://datatopics.worldbank.org/sdgatlas/ https://www.millennium-institute.org/isdg https://sdg-tracker.org/

While it is possible to capture key metrics on indicators and their dynamics in time, triggers of their spatio-temporal variations remain buried in data. Identifying them, in an interdisciplinary context, is a major step towards attaining agenda 2030.

Big Data Modelling of Sustainable Development Goals

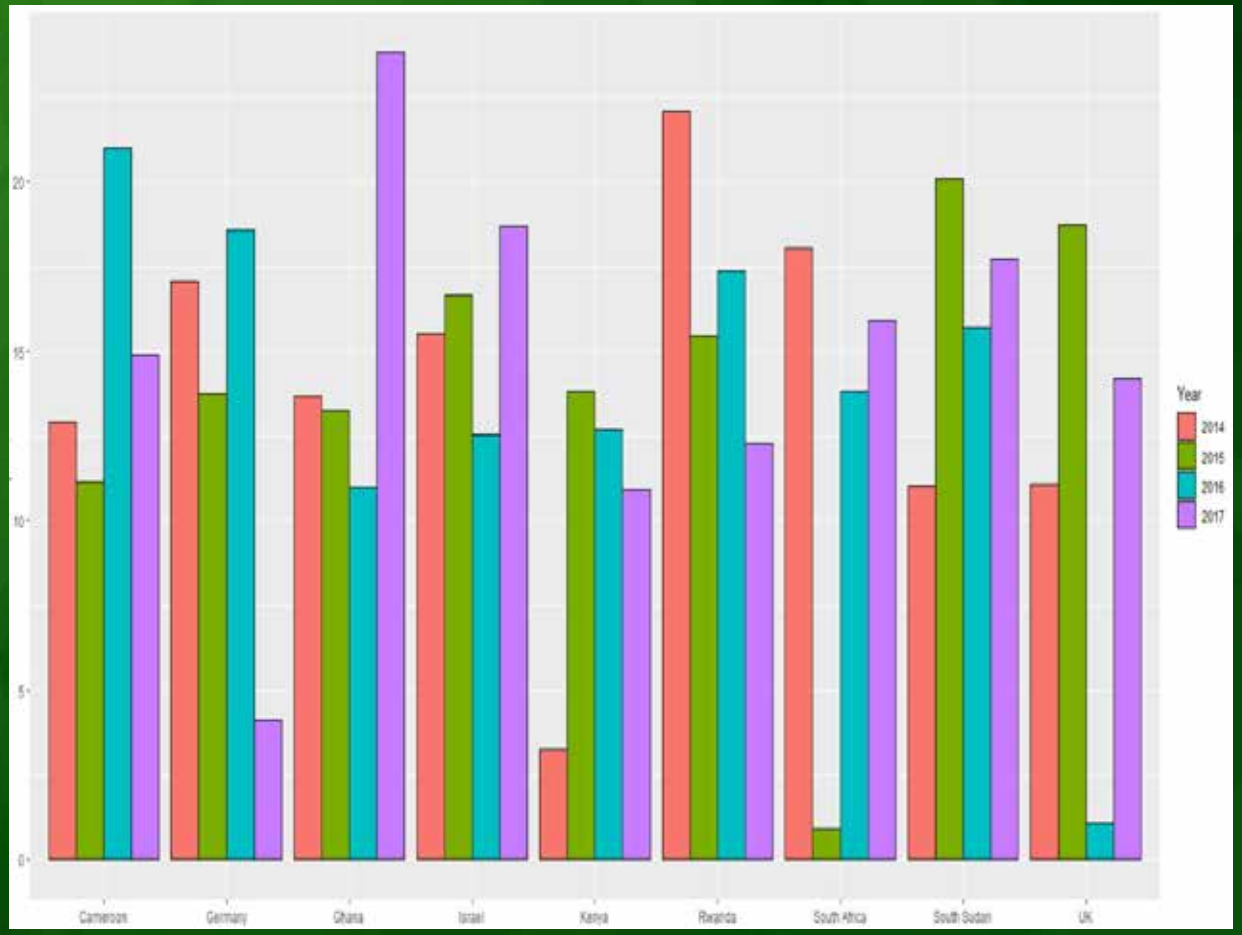
We have been researching into extracting knowledge from multi–faceted SDG data for different countries via an interdisciplinary approach involving Data Science and a range of fields across the United Nation's Sustainable Development Goals spectrum.

https://sdgs.un.org/goals

- The 17 SDGs span across various aspects of life, each goal, defined with measurable aims for improving our quality of life, based on set targets and indicators.
- Although SDG indicators and targets are given, their triggers remain buried in data. We need to uncover them.
- The triggers vary in accordance with a range of factors- socio–economic, cultural, geo–political etc.
- The sheer number of indicators, their interactions and spatio-temporal variations make the SDG project a potential Big Data problem, with a highly complex modelling setting requiring interdisciplinary attention.

TRACING HIDDEN INFORMATION IN SDGs

Advances in computing power and explosions in data generation, have triggered data–intensive research across disciplines, through, *inter–alia*, different applications aimed at addressing our challenges and opportunities.



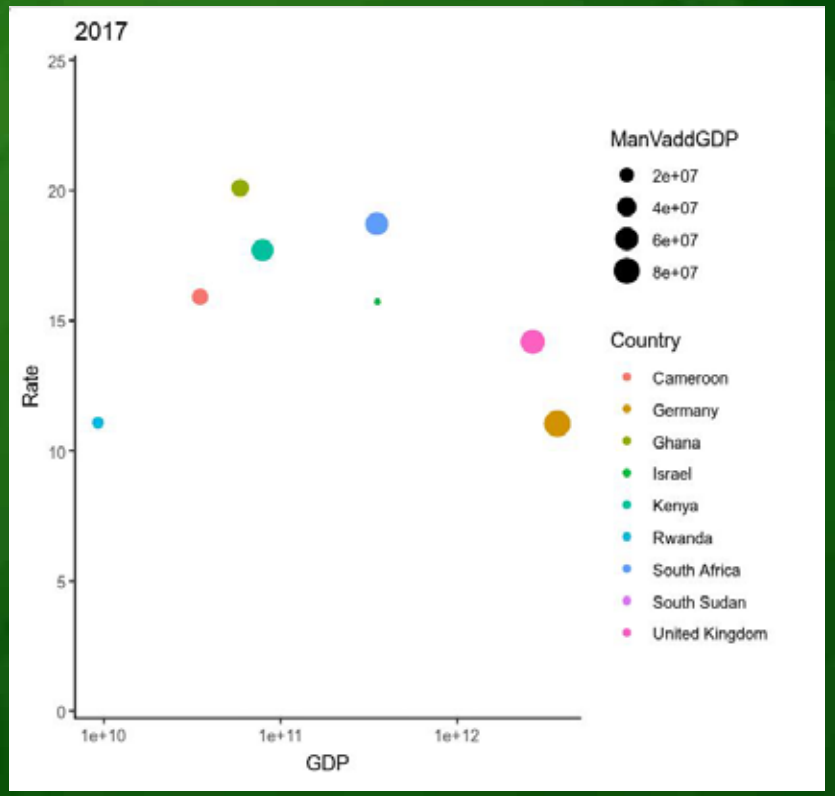
- **Example:** *Expenditure on core educational goods and services, such as teaching staff, school buildings, or school books and teaching materials, and peripheral educational goods and services such as ancillary services, general administration and other activities.*
- Taken as a percentage of total government expenditure creates other variables of interest such as GDP, GDP per capita, population, existing infrastructure etc.
- Variations across countries, can be downscaled to variations within with economic regions of countries.
- Relevant data attributes can come from different sources. These example use data from the UNESCO Institute for Statistics, Worldometers website and from the UN SDG Indicators data repository.
- Always let problem space determine your data sources.

Visualisation Examples

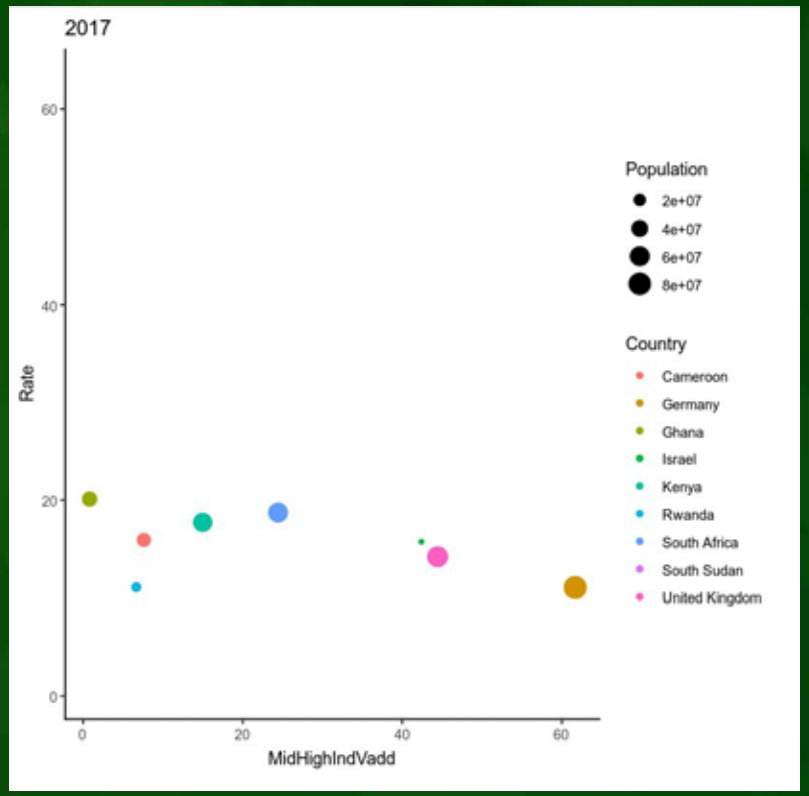
Preparing Data

VISUALISATION EXAMPLES

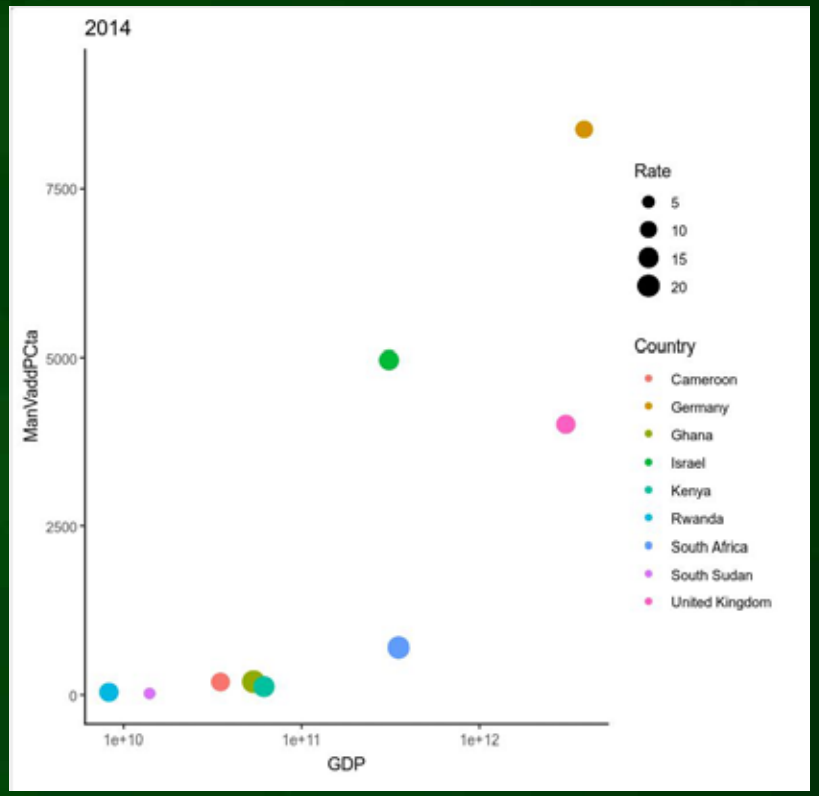
The screenshots below are animated multiple data attributes per country over the period 2014-2017. The animations can be updated as new data arrive and/or new needs arise



GDP vs Rate: Manufacturing Value Added as % of GDP



Mid and High Industrial Value Added vs Rate: Population



GDP vs Manufacturing Value Added Per Capita: Rate

Big Data – Challenges & Opportunities

Advances in computing power and explosions in data generation, have triggered data–intensive research across disciplines, through, *inter–alia*, different applications aimed at addressing our challenges and opportunities.

Across sectors & nations, Big Data brings technical and application challenges and opportunities.

- Technically, they are pathways to addressing data sharing, modelling, infrastructure, security, governance and security issues.
- Application-wise, they relate to influential policies for improving decision making at institutional, national, regional and global levels, creating a Development Continuum.
- They present potential knowledge for unlocking our understanding of the mutual impact–positive and negative, resulting from our interaction with our environment.

The indicators were not written on stone, and so a simple unscientific way to assess progress would be to think of how often they have been reviewed!

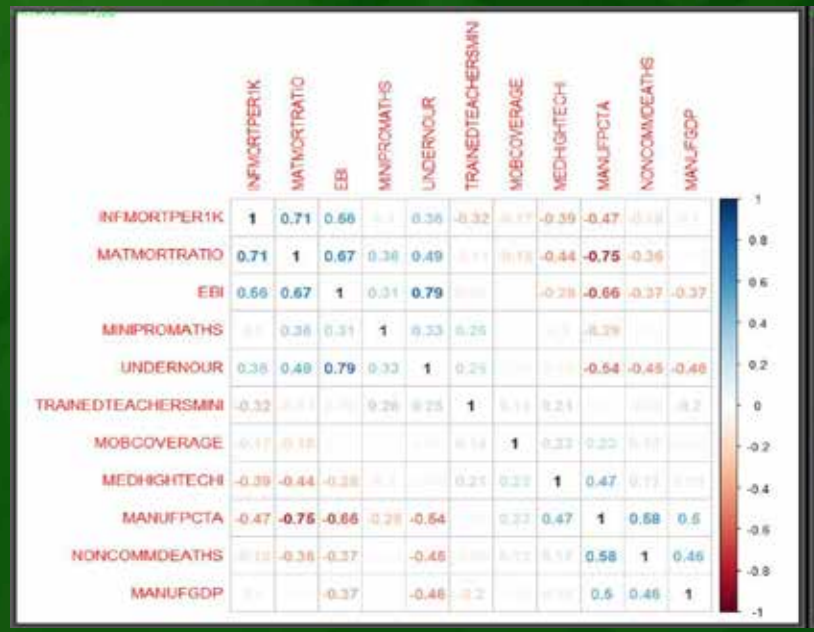
Detecting Naturally Arising Structures in Data

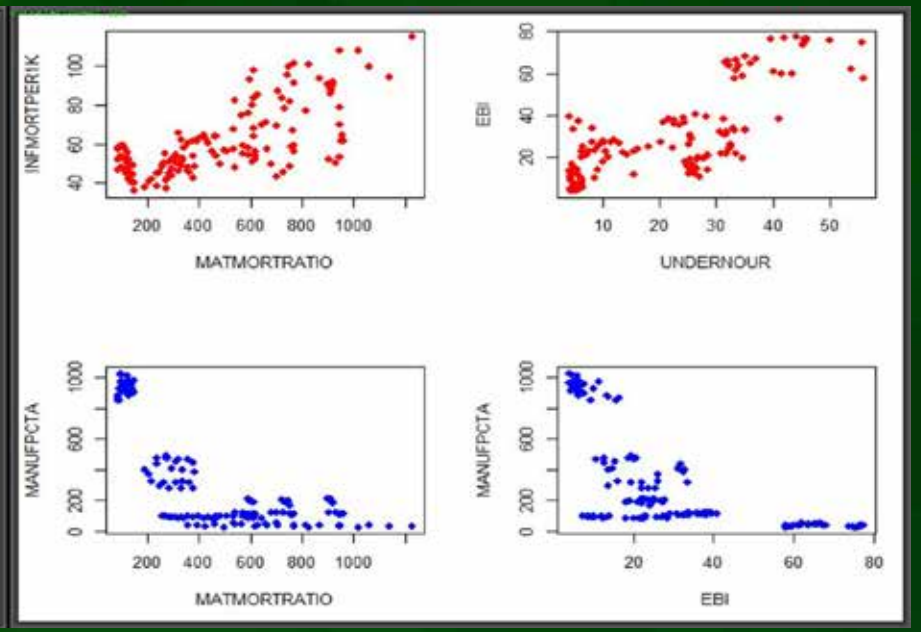
- Further examples focus on structured data–i.e., based on the standard Relational Database Management System
- Indicators for SDGs # 1, 2, 3, 4 and 9, obtained from the UN SDGI data repository to be used as data attributes for describing variations across SDGs, and six African countries–Kenya, Rwanda, Cameroon, Ghana, Botswana and South Africa over the period 2000-2017.

INDICATOR	VARIABLE NAME	SDG	CASES	PERIOD
Employed population below intern. poverty line	EBI	1	153	2000-16
Prevalence of undernourishment %	UNDERNOUR	2	102	2000-16
Infant mortality rate (deaths per 1,000 live births)	INFMORTPER1K	3	324	2000-17
Maternal mortality ratio	MATMORTRATIO	3	288	2000-15
Deaths attributed to non-communicable diseases	NONCOMMDEATHS	3	357	2000-16
Death due to road traffic injuries per 100K people	ROADACCIDEATHS100K	3	83	2000-13
Participation rate in organized learning	EARLYORGLEARN	4	83	2001-18
Prop. of teachers with at least min. pedag. train.	TRAINEDTEACHERSMINI	4	616	2000-18
Minimum proficiency in mathematics	MINIPROMATHS	4	99	2000-17
Minimum proficiency in reading	MINIPROREAD	4	80	2001-17
CO2 emissions from fuel in millions of tonnes	COEMISSFUEL	9	85	2000-16
Kg. of CO2 per unit of GDP in USD	COEMISSGDP	9	85	2000-16
Kg. of CO2 per manufact. unit in USD	COEMSSPUMAV	9	85	2000-16
Total official flows for infrastructure	INFRASFLOW	9	102	2000-16
Manufact. value added as a proportion of GDP	MANUFGDP	9	114	2000-18
ManufacT. value added per capita	MANUFPCTA	9	114	2000-18
Prop. of med. and high-tech industry value added	MEDHIGHTECHI	9	102	2000-16
Prop. of population covered by mobile network	MOBCOVERAGE	9	140	2000-17

Table 1: Selected indicators used in the study

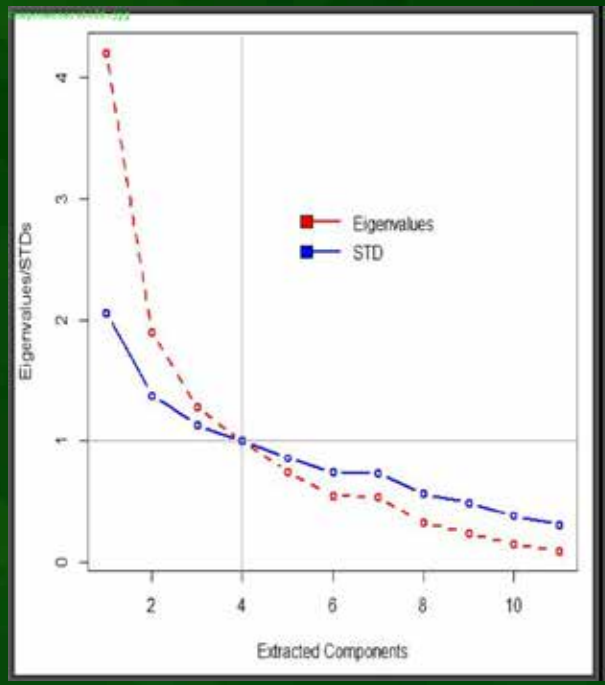
EDA - Correlations

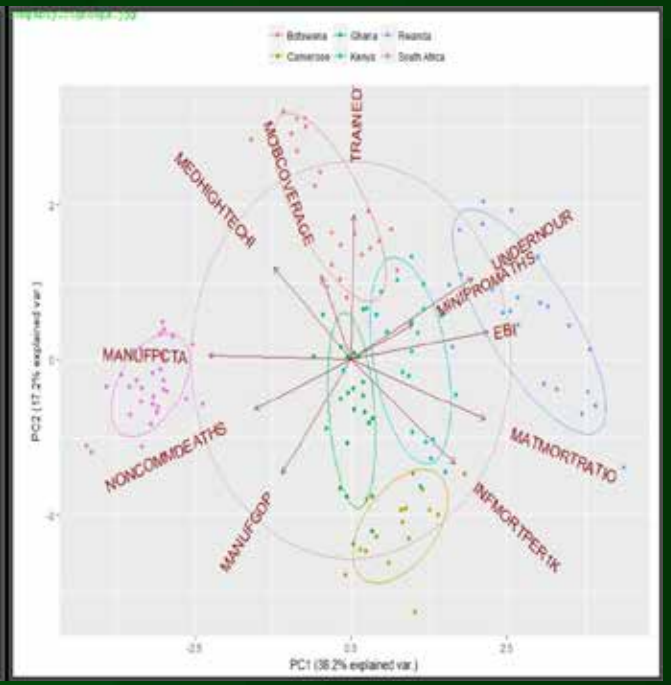




Unsupervised Modelling – PCA

- The left hand side panel exhibits the cut–off point of extracted components, 4 suggested in this case, based on the eigenvalue rule. What this suggests is that, due to multi–collinearity, the 11 indicators can realistically be explained in 4 main categories.
- In searching for triggers of SDG indicators, a thorough understanding of these four categories of variables is required. Hence the emphasis of interdisciplinarity.





- Country–specific patterns emerge–on manufacturing per capita, for instance, South Africa dominates, with deaths from non–communicable diseases and other aspects.
- Issues relating to poverty–maternal & infant mortality, undernourishment and the EBI. It is important to interpret findings with care, for the nature of the patterns.

Concluding Remarks and General Discussions

- There is plenty of evidence across the African continent that working in silos has failed to deliver, yet we continue to overlook the potential of Big Data and AI.
- Whenever we are called upon to address a complex issue–such as deteriorating performance in schools, we must fully appreciate the role of Big Data.
- If the problem we started with was classroom performance, it would be utterly wrong to pick up only a handful factors, as a thorough investigation might reveal quite subtle factors–not always about trained teachers, increased salaries or housing for teachers, as we still have parents who believe that a spiritual leader can bless children's text books and uniforms and so boost their performance; some will seek help from witch doctors and others will simply resort to buying exam papers.
- We need to bridge unseen or ignored gaps in knowledge, deriving from silos or fixed agenda. If we are particularly concerned about the link between school performance and brain growth from the embryo to early years, then nutritionists, neurologists, economists, environmentalists, educationists, psychologists, data scientists and other specialists must speak the same language–in time and space.
- Locals understand their terrain, they understand their culture and many know what their preferences are. They are a great source of **citizen science**, and by sustainably involving them and investing in interdisciplinary studies, porfolios of tools and skills will be built. Only then would we draw conclusions that are evidence-based.

References

[1] Kharrazi, A. Challenges and Opportunities of Urban Big-data for Sustainable Development. *Asia-Pacific Tech Monitor* **2017**, *34*, 17–21.

[2] Kruse, C. S.; Goswamy, R.; Raval, Y.; Marawi, S. Challenges and Opportunities of Big Data in Health Care: A Systematic Review. *JMIR Medical Informatics* **2016**, *4*, e38.

[3] Yan, M.; Haiping, W.; Lizhe, W.; Bormin, H.; Ranjan, R.; Zomaya, A.; Wei, J. Remote sensing big data computing: Challenges and opportunities. *Future Generation Computer Systems* **2015**, *51*, 47 – 60.

[4] IUCN, In the spirit of nature, everything is connected. 2018; https://www.iucn.org/news/europe/201801/spirit-nature-everything-connected.

[5] Mwitondi, K.; Munyakazi, I.; Gatsheni, B. An Interdisciplinary Data-Driven Framework for Development Science. *DIRISA National Research Data Workshop, CSIR ICC, 19-21 June 2018, Pretoria, RSA* **2018**.

[6] Mwitondi, K.; Munyakazi, I.; Gatsheni, B. Amenability of the United Nations Sustainable Development Goals to Big Data Modelling. *International Workshop on Data Science-Present and Future of Open Data and Open Science, 12-15 Nov 2018, Joint Support Centre for Data Science Research, Mishima Citizens Cultural Hall, Mishima, Shizuoka, Japan* **2018**.

[7] Ishikawa, K. *Guide to Quality Control*; Asian Productivity Organization, 1976.

[8] Primmer, E.; Furman, E. Operationalising ecosystem service approaches for governance: Do measuring, mapping and valuing integrate sector-specific knowledge systems? *Ecosystem Services* **2012**, *1*, 85 – 92.

[9] SDG, Sustainable Development Goals. 2015; https://www.un.org/sustainabledevelopment/sustainable-development-goals/.

[10] SDGI, Sustainable Development Goals Indicators. 2017; https://unstats.un.org/sdgs/indicators/database/.

[11] SDGCA, Sustainable Development Goals. 2015; https://sdgcafrica.org/.

[12] Mwitondi, K. S. *Tracking the Potential, Development, and Impact of Information and Communication Technologies in Sub-Saharan Africa*; International Council for Science (ICSU-ROA), 2018.

[13] Meusburger, P. In *Knowledge and the Economy*; Meusburger, P., Glückler, J., el Meskioui, M., Eds.; Springer Netherlands: Dordrecht, 2013; pp 15–42.

[14] Parr, M.; Musker, R.; Schaap, B. GODAN'S Impact 2014 to 2018 - Improving Agriculture, Food and Nutrition with Open Data. 2018.

[15] Ogundari, K.; Awokuse, T. Assessing the Contribution of Agricultural Productivity to Food Security levels in Sub-Saharan African countries. **2016**, 26.

[16] UN-Global-Pulse, Big Data for Development: Challenges and Opportunities.UN Global Pulse. **2012**.

[17] UN-Global-Pulse, Big Data for Development and Humanitarian Action: Towards Responsible Governance. 2016.

[18] Bamberger, M. Integrating Big Data Into the Monitoring and Evaluation of Development Programmes. **2016**.

[19] Roser, M.; Ortiz-Ospina, E.; Ritchie, H.; Hasell, J.; Gavrilov, D. Our World in data: Research and interactive data visualizations to understand the worlds largest problems. 2018.

[20] WBGroup, Atlas of Sustainable Development Goals From World Development Indicators. 2018.

[21] Downs, S. M.; Fraser, S. N.; Storey, K. E.; Forbes, L. E.; Spence, J. C.; Plotnikoff, R. C.; Raine, K. D.; Hanning, R. M.; McCargar, L. J. Geography Influences Dietary Intake, Physical Activity and Weight Status of Adolescents. *Journal of Nutrition and Metabolism* **2012**, *2012*, number of pages 6.

[22] Codd, E. F. A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM* **1970**, *13*, 377–387.

[23] Chapmann, J. *Machine Learning Algorithms*; CreateSpace Independent Publishing Platform, 2017.

[24] Kogan, J. *Introduction to Clustering Large and High-Dimensional Data*; Cambridge University Press, 2007.

[25] Agrawal, R.; Imieliński, T.; Swami, A. Mining Association Rules Between Sets of Items in Large Databases. *SIGMOD Rec.* **1993**, *22*, 207–216.

[26] Mwitondi, K. S.; Zargari, S. A. An iterative multiple sampling method for intrusion detection. *Information Security Journal: A Global Perspective* **2018**, *27*, 230–239.

Thank You for Your Attention