Big Data Modelling of SDGs: Project Concept Note

Kassim S. Mwitondi

Sheffield Hallam University, Faculty of Science, Technology and Arts

Abstract

The proposed setting–**Development Science Framework (DSF)**, views each **SDG** as a source of highly correlated Big Data attributes and seeks to explore their interactions in a predictive, spatio-temporal context. For example, the relationships among **SDGs** 13, 14 & 15 on human livelihood, reflected by **SDGs** 1 through 7, are particularly intriguing and modelling their dynamics can provide answers to general questions in the form of **What Works Across Countries and Sectors?** The relationships among attributes describe the **invariant** conditions that epitomise the **meaningfulness** of the **concept** which **DSF** learns from available data. **DSF** adopts first principles approach, treating **SDGs** as sub-projects, driven by the well-documented relationship between knowledge & development in time and space. It focuses on trust, interdisciplinarity, authoritative data, knowledge transfer, resources, tools & techniques. Analogous to Eleanor Roosevelt's view on where **universal rights begin,** it takes a **bottom-up** approach to the universal agenda 2030, seeking real knowledge on indicators from *"...small places, close to home–so close and so small that they cannot be seen on"* the **World Bank SDG Atlas.** Its main objectives are therefore to map and deliver knowledge about clusters of individual societies on that global map, their disparate neighborhoods, activities, achievements and influencing factors, based on indicators deriving from relevant and authoritative data from various sources, including grey literature and citizen science. Its implementation strategy is for three pilot **DSF SDG** monitoring nodes to be established in three African countries with known socio-economic heterogeneity, in order to provide a clear baseline for the project. To effectively support agenda 2030 at the country level, the project requires a guarantee that policy makers affecting each monitoring node, are in full agreement with **Open Data Initiative (ODI)** agenda.

Key Words: Big Data, Citizen Science, Concept Learning, Culture, Grey Literature, ICT, IoT, SDGs

1 Project Description

Agenda 2030 will only become a reality if we have common and meaningful definitions of concepts and parameters that the **SDG** indicators ultimately lead to. That **meaningfulness** is the **invariant** that forms a concept, the rules of which the proposed **Development Science Framework (DSF)** learns from data attributes. Our main idea is to treat each of the 17 **SDGs** and all related indicators as highly correlated Big Data nodes, each being amenable to multi-variate modelling, with the potential of uncovering unknown attribute relationships. **DSF** has been discussed at several data science events, in three continents so far, and the idea now is to convert it to practical implementation. The rationale of the project unfolds from **SDG** challenges on the one hand, and opportunities on the other.

1.1 Challenges

The framework favours locally harnessed and utilised authoritative data, without marginalising multilateral institutions that host global data on their servers. This narrative has been raised as an issue by one of our informal reviewers from the UK **office of National Statistics (ONS),** arguing that we should be able to *...find a way to deliver shared, trusted data sources and common methods for the SDGs.* **DSF** is in line with this line of thinking, particularly with the four pillars of **ONS'** work on the UN Global Platform–**Trusted Partners, Trusted Data, Trusted Methods & Trusted Learning. DSF** is also in agreement with **ONS'** position on comparability of methods across nations and it is for this reason that **DSF** stipulates that we have common and meaningful definitions of concepts and parameters that the **SDG** indicators ultimately lead to. Now, this is precisely the main challenge, with a number of issues, as exemplified below

1. Data provenance plays a crucial role on inferences. Hence, data ownership and repeatability are fundamental.
2. Data attributes in each **SDG** will typically differ across countries and sectors. We need a common understanding.
3. The meaningfulness, scale and impact of **innovation** in South Korea are not the same as those in South Sudan.
4. Determining the usefulness of the information uncovered from data attributes–establishing what is interesting. This challenge can be tackled by adopting an interdisciplinary approach to project implementation.

1.2 Opportunities

Recent technological advancements are driving us into logistical situations in which we generate more data than we can make sense of. At the same time, we are witnessing increasingly great innovations in data acquisition and modelling, through **Machine Learning (ML) & Artificial Intelligence (AI)** techniques. Viewing each **SDG** as a Big Data node, and the entire set of 17 as a multi-disciplinary data fabric, underline the potential for Big Data modelling of **SDGs**, the subject of the proposed project. Tools like https://www.millennium-institute.org/isdg, developed by the Millenium Institute, simulates patterns based on alterations of some key **SDG** metrics. The tool will typically give an infinitely large number of patterns, depending on the perturbations made, which is only part of what Big Data is all about. Further, we can only simulate patterns based on some assumptions, parameters, environment etc, which vary invariably in a spatio-temporal context. In addressing real-world challenges, computational, statistical and analytical skills are mere tools that we need to address practical problems. Across **SDGs,** are infinitely many practical problems that require data-driven solutions and **DSF** is designed to add a predictive power to simulation tools such as this one.

1.3 Research Question, Aims and Objectives

The intricacy of relationships among data attributes across **SDGs** and their impact on human livelihood, are evident, and well-documented. The relationships among attributes describe the **invariant** conditions that epitomise the **meaningfulness** of the **concept** which **DSF** learns from available data. The project shall be seeking to provide answers to general questions in the form of **Taming the SDG Big Data Potential: What Works Across Countries and Sectors?** The project aim will be to identify influencing factors of phenomena within disparate neighborhoods, activities, achievements, based on indicator deriving from relevant and authoritative data. For a pilot project, general and specific objectives will differ, depending on selected **SDGs** and spatio-temporal attributes. Its general objectives are as follows

1. To identify key data attributes within each **SDG**.
2. To determine the statistical relationships among attributes within **SDGs** of interest.
3. To carry out unsupervised modelling of **SDGs** of interest.
4. To carry out supervised modelling of **SDGs** of interest.
5. To map and deliver knowledge about clusters of individual societies on the global map.
6. To embed aspects of **SDG** modelling in all curricula across the spectrum–data analysis, data science, Big Data.
7. To dig up and utilise data from grey literature and citizen science.
8. To engage policy makers, development stakeholders and the general public in converting uncovered knowledge to good practice aimed at enhancing **SDG** attainment and the **invariant**.
9. To disseminate and present core findings at regular world-class forums.

As stated above, specific objectives can be defined, depending on selected **SDGs** and the nature of the underlying problem. For instance, focusing on **SDG-5**, in which the main aim is to empower women and girls in understanding the environment they live in and their potential, the following specific objectives may be defined

1. To engage women and girls with professional researchers in understanding their habitat through citizen science.
2. To promote interest of Science, Technology, Engineering & Mathematics (STEM) among girls (future mothers).

3. To raise awareness of the universality of science, and the coherence of knowledge, culture and technology.

Recognising that **SDGs** are multi-faceted, highly dynamic and highly correlated data sources, we present the methodology as a collection of projects, relating to cause-effect relationship between knowledge & development in time and space. The methodology, described below, focuses on improving the fabric of **SDG** indicators, using identifiable **SDGs** data attributes as drivers, to learn the concept via unsupervised, supervised & association models.

DSF aligns well with **ONS'** international engagement, which, *inter-alia*, involves working international partners to develop new data sources and methods as well as attain best practices and harmonise statistics to improve aggregation and comparison of country data. It will also contribute to **ONS' SDGs** development roadmap for official statistics, an implementation strategy that can be adapted on the pilot scheme.

The Department for International Development (DFID), Japanese International Co-operation Agency (JICA), United States Agency for International Development (USAID) and many other international development agencies have been pouring in money into developing countries for years. The impact of these funds has hardly been ratified. The recent report by Lord Ashcroft https://www.conservativehome.com/platform/2018/11/lord-ashcroft-tanzania-a-case-study-in-the-bulging-textbook-of-aid-failure.html is good case in support of Big Data Modelling of **SDGs.**

2 Methodology

To achieve the foregoing objectives, the project requires interdisciplinarity, authoritative and unified data sources & frameworks, techniques, skills & consensus on learning rules from data. Its successful implementation requires incorporating related activities into each country's learning, research and data collection practices. It can start as a pilot scheme, involving several schools, preferably starting with students with their focus on Science, Technology, Engineering and Mathematics (STEM); vocational training and taming citizen science and up to higher education. Scaling up would be easier, if there are proven success stories. The **DSF** is graphically illustrated in Figure 1.

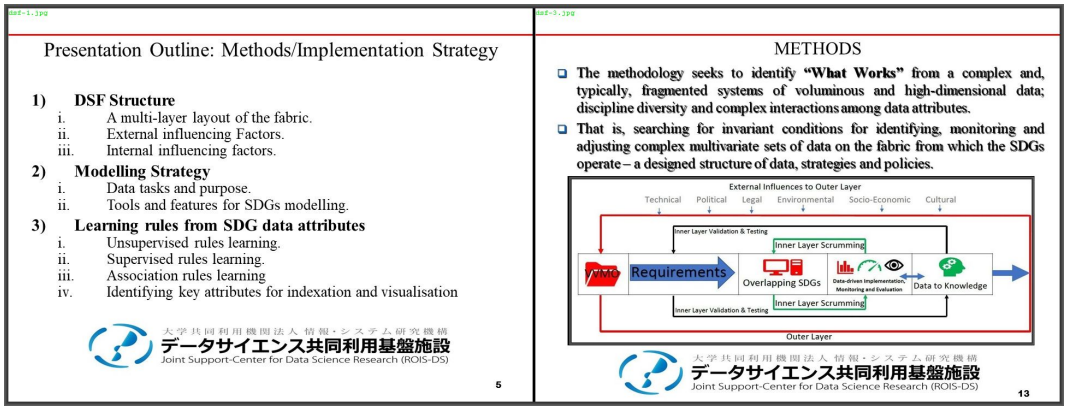

Figure 1: Graphical illustration of the **DSF** (Source: Mwitondi et al., 2018)

3 Expected Outcomes

Through **DSF** we can create a sustainable environment for unification of communities via shared values based on a combination of Big Data, citizen science, culture and technology. For example, focusing on women and girls, **SDG-5** may use information and communication technologies to facilitate opportunities for closer understanding of aspects of natural and social sciences that would ultimately enhance their self-esteem and confidence. Its design consists of sequentially planned transient activities each yielding measurable outcomes, and spanning across other **SDGs** based on specific socio-economic indicators. Groups of women and girls, in rural and urban centres, provides them with

opportunities to open up to issues that directly affect them. They become part and parcel of a diagnostic process to uncover the dysfunctional systems that hold them down and learn to prescribe and adopt solutions, through co-ordinated citizen science plans. Ultimately, we are able to uncover **what works** for them and to bring about **change**, a novel environment that allows girls to grow up as independent and confident members of an integrated community.

References