Ghana Statistical Service

Data Science Roadmap

November 18, 2022

Ghana Statistical Service (GSS)

A five-year roadmap on the implementation of a data science strategy at the Ghana Statistical Service

Address

Ghana Statistical Service
Head Office,
P. O. Box GP 1098,
Head Office Building,
Location: Finance Close, Accra, Ghana.

Phone & Fax

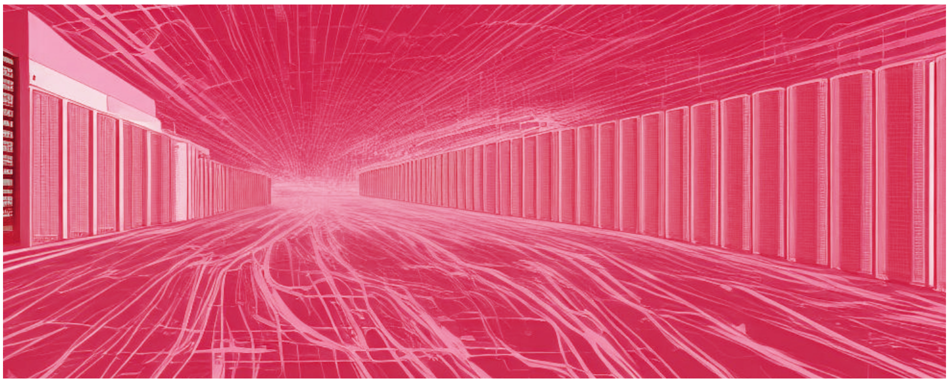
Phone: + 233-30-266-4304
Fax: +233-302-664304

Online

Email: info@statsghana.gov.gh
Website: www.statsghana.gov.gh

Table of Contents
Data Science Roadmap

01. Introduction
Data science & the roadmap.



01

This document serves as the roadmap to the Ghana Statistical Service (GSS) efforts in developing a data science policy. It first outlines both the need and the justification for such a document and then gives the principles this data science policy at GSS is built upon. Later chapters cover specific use-cases and the steps GSS will take in the upcoming 5 years to realize this data science vision.

1.1: The Need for Data Science at GSS

Data science has the potential to improve, streamline, and automate the production of currently-released statistics as well as serve as an enabler in the creation of new statistics. In a world of growing demand for real- or near real-time data among modern data users, statistical organisations have to modernize to satisfy the demand of these users. Furthermore, the ability to produce data for certain indicators, i.e., humanitarian related indicators, resides in new data sources which require modern data science techniques. Development of data science within the organisation is, therefore, imperative for the Service to release relevant and desired statistics. Data science will oversee the implementation of projects making use of a variety of different data sources and data science techniques, including analysis of large, smaller, unstructured datasets from other MMDAs, through interactive visualizations to improve the dissemination of GSS data and automation to improve the productivity of the

01. Introduction
Data science & the roadmap.

Act 1003

Act 1003 establishes the Ghana Statistical Service as the central statistics producing and coordinating institution for the National Statistical System and to strengthen the production of quality, relevant, accurate and timely statistical information for the purpose of national development.

operations of the National Statistical System (NSS).

Ghana Statistical Service, with the mandate provided by Act 1003 to provide leadership and coordination of the National Statistical System (NSS), undertook a national capacity assessment. This led to the organisation of a national data roadmap that proposed three key priorities for action in strengthening the NSS in Ghana; namely: addressing data gaps, encouraging data use, and strengthening the data ecosystem.

Through the application of data science processes and methodologies in all activities and projects, GSS aims to make the release of statistics more timely and frequent, while also applying more rigorous and quality controls on these statistics. At the same time, GSS aims to take full advantage of the use of data science to use modern and complex data sources, whether it is web-scraped data, satellite imagery, or mobile phone data, to produce relevant statistics which can currently not be produced, using traditional statistical methods.

1.2: The Need for Data Science at GSS

Implementing data science in the production of statistics at GSS will require cooperation both within GSS – between different directorates, units and staff members — and externally with private sector data providers, academia, ministries, and international organisations. It is imperative that all these internal and external stakeholders understand the data science vision of GSS to be able to cooperate on data science projects and develop new use-cases. A roadmap would serve as a method to keep the data science team on track in the development of use-cases, capacity building and innovations needed.

To harness the potential of data science, GSS will need to build data science capacity and develop new skill sets to be able to work with a combination of both traditional and non-traditional data sources and analytical methods. A roadmap highlights the skills that will be required and can make a start in the process of developing them.

02. History of Data Science at GSS

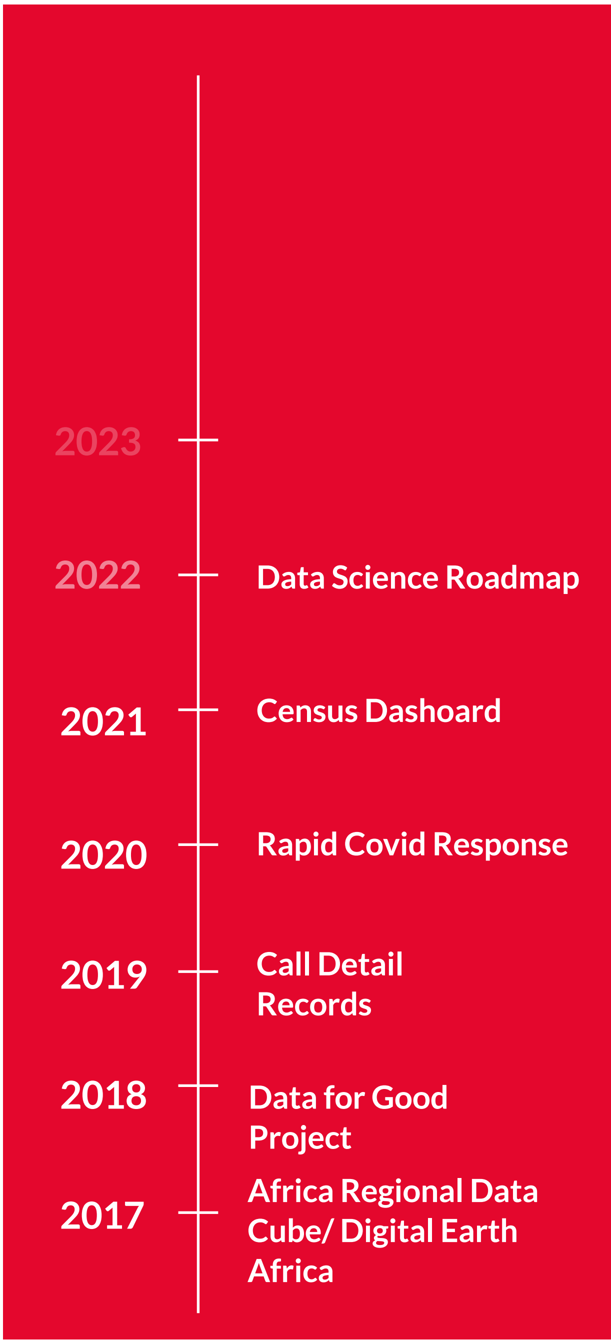
How has data science been used at GSS.

02

Data science is not new to GSS. The first projects that employed modern data science techniques were started back in 2017. This chapter will highlight some of the past uses of data science at GSS.

The data revolution, coupled with the adoption of the Sustainable Development Goals (SDGs) in September 2015, provided the context for National Statistical Offices globally to act. The demand for more timely, disaggregated and quality data is unprecedented, and traditional data systems are unable to keep pace.

The Data Roadmap in 2017 has led to the development and implementation of several projects including Data for Good. The Data for Good Project (D4G) is a flagship public-private partnership between GSS, Vodafone Ghana and Flowminder Foundation, a UK based NGO with expertise in the analysis of CDR to support humanitarian interventions. The primary objective of the project was to equip GSS with the capacity to access and analyse aggregated anonymized mobile phone metadata, also known as Call Detail Records (CDRs) to fill some of the data gaps identified in the production of national statistics to support evidence-based decision-making. The Data for Good project has since 2018 been building the technical and legal infrastructure



2023

2022 Data Science Roadmap

2021 Census Dashoard

2020 Rapid Covid Response

2019 Call Detail Records

2018 Data for Good Project

2017 Africa Regional Data Cube/ Digital Earth Africa

02. History of Data Science at GSS

How has data science been used at GSS

2017

GSS has been running data science projects within the SDG secretariat since 2017.

to enable access to and analyse anonymous and aggregated telecommunications data. During the first phase of the project, aside from the establishment of the organisational, legal, and technical frameworks, the capacity of six GSS staff was developed in the application of data science techniques with analysis of CDR as the area of application. The D4G partnership drew technical support from Flowminder. This partnership afforded the Service the opportunity to provide relevant statistics in a cost effective and timely manner. Notable are the three mobility analysis reports produced to support government non-pharmaceutical interventions on COVID-19 (GSS' website).The D4G project which is currently in its second phase, chalked some achievements in the first phase. In 2019-2021, it was expected to last until the end of 2023 and continue to strive forward based on a sustainability plan being developed.

At the same time, the Sustainable Development Goal (SDG) unit of GSS, in cooperation with Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) undertook two projects to use citizen generated data and data science to estimate novel statistical indicators around the topics of gender-based violence and solid waste management. Also, in 2017, the SDG unit started working with Digital Earth Africa (then named The Africa Regional Data Cube) on the use of satellite imagery for SDG monitoring.

The 2021 Population and Housing Census (PHC) saw increased use of digitized methods and data science. The fact that this was the first digital census allowed for automated daily quality checks and the launch of a dashboard to monitor progress and data quality in near real time. Latter surveys, such as the tourism survey and the Annual Household Income and Expenditure Survey (AHIES), continued to employ these dashboards.

Due to the collection of GPS estimates of structures in the PHC, geographic information system (GIS) technologies were leveraged. A Python-based data cleaning routine was implemented to sanitize the data and remove duplicates before processing. Map creation and the computation of complexity indicators were semi-automated in both R and Python. After the census, the matching of PHC results to the Post Enumeration Survey (PES) was automated, using a fuzzy matching algorithm together with a web interface for human validation to assist in this task. Traditionally, matching PHC to PES records had been a very cumbersome and time-consuming process in which paper questionnaires from both the PHC and PES were compared by a group of matchers. The digital alternative reduced the time needed to match individuals in 498 enumeration areas to 10 days.

03. Framework of the Data Science

What GSS Means by Data Science

03

The Data Science roadmap and how GSS understands the need and use of big data. Data science is built upon the framework of “8 Vs” of big data.

Volume

This relates to the size of data that projects using data science and big data will bring. GSS needs to build both the human and technological capacity to store and analyse data. With the increase in both size and types of data, data science will be progressively important to produce meaningful statistics from large amounts of data. For example, instead of having data on a couple thousand households a few times a year, new non-traditional data sources might contain the data of every transaction made at a supermarket (scanner data) and volume will be measured in terabytes instead of gigabytes. Another new field of innovation in official statistics is the use of administrative data. Administrative data is data that is collected by (government) institutions, such as tax authorities, police, cadaster, birth-and-death registries, universities, national identification authorities, etc., for administrative reasons. Administrative data is collected for either registration, transactions, record-keeping or another operational purpose and their statistical use is secondary. Tapping into this secondary use, administrative data can reduce costs, and allow for near real-time creation of statistics for more levels of disaggregation. However, to fully leverage administrative data, data science skills are required. Data from different registries and databases need to be extracted and matched to each other while guaranteeing confidentiality. By design, administrative data tends to be big data and suffers from different quality issues than traditional survey data.,

Velocity

Velocity relates to the speed at which new data is generated. Big data projects create data at a much faster rate than traditional surveys do. Satellite data is updated weekly, while, for example, telecom data can be generated near real time. Although most data are warehoused before analysis, there is an increasing need for real-time monitoring and processing of these enormous volumes of data. This means that pipelines need to be set up to be able to store, analyse and disseminate data at the same rate.

Variety

Variety relates to the different types and format data may come in. Instead of tabulating well-structured questionnaire data, big data often comes in different, unstructured data format. These might be images, videos, text messages, JSON files from APIs, or any other format. Some of these new data sources are passive data sources that are generated as the byproduct of some other process. Examples include cargo data from port authorities, sensor data (Global Positioning System [GPS], gyroscope and other sensors) from mobile phones, and administrative data from schools, hospitals, tax authorities and government agencies. A variety of data affects the inferences that will be made from the data and, therefore, an algorithm is required to ensure the integration and harmonization of data without repetitive manual intervention. GSS will need to develop the skills to be able to analyse all of these different data formats and find ways to make meaningful statistics from them.

Value

This relates to the value of collected data and the value of derived data. Data offers an enormous source of value for actionable insights, but only if interpreted and disseminated in the right way. GSS needs to be able to communicate to the bodies it cooperates with and to the users of the statistical products what the value of the generated data is. The research and data science directorate also seeks to ensure that the statistics or figures released are quality assured. At the same time, GSS as the custodian of much sensitive data needs to continuously work on fostering public confidence in the fact that GSS protects the privacy and confidentiality of those data.

Variability

This relates to the fluctuations and differences in data over time. Big data sources tend to be noisy and might produce the same data in different formats. These variations and inconsistencies in data over time necessitate the need for some adjustments or harmonization to be done according to acceptable standards. GSS needs to build the skills and software to recognize variable data and harmonize it in such a way that coherent statistics can still be estimated.

Veracity

This relates to the quality of the data. The quality of big data may depend on the source, the mode of collection, and the type of data. Recognizing inaccuracies or incompleteness in unstructured data requires both domain knowledge of the data as well as technical data science skills. The cleaning, quality assurance and interpretation of data of different levels of completeness and quality require skills and software that will have to be built at GSS.

Validity

Validity refers to how accurate big data is for a certain purpose. A sizable amount of large data is still useless and is referred to as "black data". Prior to analysis, the remaining portion of the gathered unstructured data needs to be validated, checked for relevance and cleaned. As the national statistical office, GSS will take a leading role in generating high quality and validated statistics using data science. GSS will need to generate statistics that are reliable and can be verified from other sources. To achieve this, statistics must be contextualized and the same statistics should be generated using different possible methods. To ensure data validity for all these different sources, GSS will generate statistics that are reproducible and verifiable from other sources. Where there are limitations to the statistics coming from non-traditional data sources, GSS will carefully explain the limitations of these new statistics and the disseminated data.

Visualization

Visualization relates to the process of representing abstract concepts visually. Data processed must be transformed into something easily comprehended and actionable. GSS needs to mature the skills to make high quality data visualizations and develop a service-wide style for these visualizations. The graphs and charts need to communicate and also be accurate. The organisation seeks to employ software and techniques that will be used to create powerful visuals and create dashboards for easy accessibility. Visualizations also play a crucial role in timely data quality monitoring. Often a visual of the data can tell you more about the structure, quality and gaps of the data than any report or summary statistic can. It is, therefore, important to incorporate data visualizations at all steps of the creation of statistics.

”

8 Vs

While eight is an arbitrary number, they do fit the vision GSS has on the use of data science and big data well. The original 3 Vs of big data were coined by Laney in 2001, these were later expanded to

- 4 Vs (https://www.oracle.com/technetwork/database/bigdata-appliance/overview/bigdatarefarchitecture-2297765.pdf)
- 5Vs (Anuradha, J. "A brief introduction on Big Data 5Vs characteristics and Hadoop technology." Procedia computer science 48 (2015): 319-324)
- 7 Vs (https://datafloq.com/read/3vs-sufficient-describe-big-data/)
- 10 Vs (https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx)
- 17 Vs (https://www.irjet.net/archives/V4/i9/IRJET-V4I957.pdf)
- 29 Vs (https://ijarsct.co.in/Paper936.pdf)
- 42 Vs (https://www.elderresearch.com/blog/the-42-vs-of-big-data-and-data-science/)

04. Use of Data Science at GSS
How Data Science will be used at GSS.

04

Ghana Statistical Service (GSS), as part of its legal mandate as the National Statistical Office of Ghana as enshrined in the Statistical Service Act, 2019 (Act, 1003), is responsible for the development, collection and reporting of statistics in Ghana. Data science will provide technical oversight and quality assurance of GSS data science projects to ensure that they meet customers' expectations and are delivered in line with technical and analytical standards from across the data science ecosystem. To construct architect data processing pipelines for the GSS and drive the collection and implementation of new data as well as enhancement of existing business data sources in both GSS and the NSS. This chapter also highlights the 5-year vision on the use of data science of the various directorates at GSS and attempts to answer the question: How will this directorate use data science over the next five years.

01 ///

Automate processes
-

Data science and reproducible methods will automate processes. By automating certain steps of the data collection, quality control, and statistical estimation, staff can make time to explore other data sources to improve the methodology and estimates on aggregates. This also means that statistics can be produced and released more often and timely. Concretely, this means that processes that are now done manually in Excel, or on paper will be automated. This can include data formatting, data extracting and file uploading but also table creation and the visualization in a standardized way.

02 ///

Use of new data and methods
-

New sources of non-traditional (big) data and appropriate data science methods to analyse this new data will enable staff to produce new or experimental statistics. This means that GSS can produce more value for its users. To do so, subject-matter experts and data scientists will have to work together, to derive meaning from these new statistics.

03 ///

Standardize
-

By standardizing analytical processes, code, and dashboards between different units of GSS, products and processes will become more uniform and easier to implement. This will make sharing data in an analysable format and setting up quality assurance processes, including dashboards more straightforward.

Economics Statistics Directorate

Using data science and better data warehousing techniques, the Economics Statistics Directorate wants to centralize the storage of all economic statistics data on a server system. Data and code would no longer be on just individual's laptops, but instead accessible to all members of the Directorate. This would facilitate data sharing and reproducibility of code.

By building R coding capabilities and by investing in data server systems, the Economics Statistics Directorate wants to introduce the use of Reproducible Analytical Pipelines (RAPs) for CPI, PPI, trade statistics, services statistics, and agricultural statistics, but also the Annual Household Income and Expenditure Survey (AHIES) and Integrated Business Establishment Survey (IBES) surveys. This would speed up the data collection and production of these statistics and streamline their releases.

As a third point, the Directorate aims to increase the use of digitized administrative data from the GRA for the computation of GDP and customer information from hotels and tourist sights to create timely tourism statistics. Fourth, the Directorate will set up online platforms for businesses that are part of the PPI, CPI and GDP computation frame to upload data on a monthly basis to the GSS servers, without GSS staff having to interview these businesses every month. Finally, and in addition to the use of administrative data, more sources of non-traditional data will be tapped into. The Directorate will strengthen collaboration with large shops and supermarkets to get access to scanner data to be able to estimate both prices as well as changes in consumption patterns and consumer demand.

Social and Demographic Statistics Directorate

In order to close the data gap, the Social and Demographic Statistics Directorate intends to use data science to automatically compile administrative data at GSS as well as other Ministries, Departments, and Agencies (MDAs). This automatic data compilation has already begun with the Ghana Police Service and will be used as a case study for other MDAs to understand how this technology can make their work more efficient and timely in terms of quality data

In addition, they would also like to explore the use of data science skills to automatically preprocess population data before feeding it into projection software seamlessly for further analysis. They have also considered how to disseminate the data generated through other platforms, such as the Databank, and National Reporting platform (data that can be used to report on the SDGs). Finally, the Directorate will like to incorporate non-traditional methods of collecting data such as data from satellites, and the financial sector leveraging technology and innovation.

Regional Offices Directorate

The regional offices directorate mainly supports the various units in the head office in survey implementation and data collection of administrative data. Firstly, the directorate aims to use web scraping to collect prices of goods or commodities from the websites of shops, supermarkets or malls for the CPI computation. This would eliminate human errors during data entry for the computation of the CPI figures. The directorate also aims to develop more templates using CAPIs for collecting administrative data from other institutions as well as a script (using R or python) that would check for errors in data. In addition, scripts can be developed using R or python to verify data that comes to the directorate. The directorate will integrate Natural Language Processing (NLP), a transcription application with telephone calls to capture interviews. This would speed up the data processing cycle and enable automated validation of data as well as ensure quality data.

Finance Directorate

In terms of the recording of financial transactions at GSS, most recordings and reports are done in Excel and are not accessible to all, except a few government-approved projects like the Harmonizing and Improving Statistics in West Africa Project (HISWAP) where financial transactions are recorded in the Ghana Integrated Financial Management Information

System (GIFMIS). The finance directorate wants to use data science to create a database and automate some financial transactions to enable near real-time monitoring of the day-to-day financial transactions in surveys or projects via interactive dashboards, ensuring visibility at all levels to make communication with management and stakeholders easier and on time. This is essential to enable the directorate to have some quality checks on records in the database to prevent errors in reports and memos.

Incorporating data science in the activities of the finance directorate would go a long way to help the finance and account processes of the directorate. A system that automates surveys' and projects' finances can be built with the help of data science to ensure that payment processes are not duplicated. GSS can build a database which would house all financial records and payment vouchers such that payments record of each person can be traced without difficulty. Also, an asset management system can be automated to give prompt feedback on asset stock depletion to aid asset restocking. In addition, with the help of data science, income generating Activities can be developed and monitored. This can be done by bringing on board the private sector, internally generated funds, consultancy services and exploiting the benefit of the training school. Therefore, a system that incorporates all these income generating avenues can be developed. The idea of linking output of staff to incentives can be made possible by building systems that benchmark one's work done to the incentives they receive. And all this can be achieved with the help of data science.

Survey Organisation and Census Directorate

The use of data science in the Survey organization and Censuses Directorate would make activities in the directorate much simpler than it is currently. Data science would help automate sampling design and weighting such that a robust master sample frame can be built. The computation of sample weights would be made easier and improve then precision of estimates once a system is built and fed with population characteristics. Also, with the help of data science, a dashboard that would monitor enumerator specific progress and completion rate of work can be automated. A similar dashboard could monitor all surveys at a time, so that management can access the progress of work of all ongoing surveys. Again, Surveys and Censuses work plans can be automated to help monitor the extent of completion of the preparatory activities. A database that would contain information of all enumerators who have worked on projects for the service can be built and automated such that the deletion of blacklisted field officers is made possible. Also, using algorithms, quality checks on survey and Censuses data can be built and automated using common variables and updated to suit a specific survey. This would enhance prompt data cleaning and timely release of Censuses and survey results.

Communication and Dissemination Directorate

Communication and Dissemination within the national statistics office has as it role to ensure effective efforts reach and engage target audience(s) of statistical products to educate and provide data for policy development and further research. Within the statistical service, this directorate has the cross-functional role of generating, archiving and making statistical data and information available to the public through data requests, press releases and a library system. For the next five years, the directorate envisages the use of data science to develop strategy to enhance tracking, cataloguing and data production. This would include for Data request; documentation and tracking,Generating monthly and quarterly reports in response rate, Identifying follow-ups within 2 weeks of request, Distribution of requests to functional directories. Sales and Publication would need data science for accounting and estimating charges based on requested content, financial record harmonisation between publications and finance directory, payment tracking. Library needs to assess product availability and stock determination and improve identification of shelving systemTracking of borrowed books and records.Warehouse would used data science to enhance archiving and documentation of the survey process and data, Creating a referencing system to identify data and records available. Production can also use data science to improving data request production, Modelling the pricing system for data request. Protocol would need a system to Tracking and monitoring of statistical service releases in the media space and identifying key messages and misreporting, Creating a system to monitor and track travel arrangements

Coordination and Programme Management Directorate (CPMD)

The Coordination and Project Management Directorate (CPMD) is responsible for the planning, monitoring, and implementing of projects with the Metropolitan, Municipal and District Assemblies (MMDA) and at Ghana Statistical Service. The directorate would like to use project management tools such as Trello or Click Up to monitor projects in the service and ensure effective communication within GSS and among stakeholders. The Coordination and Project Management Directorate seeks to use data science to create a database for the concepts of the Harmonizing and Improving Statistics in West Africa Project (HISWAP) and another that captures the records of their budgets. Data from these databases would be used to create a user-friendly dashboard and shared with management and local and international stakeholders for visibility and monitoring of projects at each level.

05. Model of Integrating Data Science

How Data Science will be developed at GSS.

Orga-nogram

The organogram can be found here: https://www.statsghana.gov.gh/profile_page.php?ProfileCategoryID=MTA1NTY1NjgxLjUwNg

200, 000 USD

For the next five years, GSS envisions an annual financial investment of USD 200,000 or Ghana Cedi equivalent for the integration of data science into its operations. This will cover the labour costs, equipment and production and launching of various use-cases.

The new provisional organogram of GSS establishes a new Data Science unit under the Research and Data Science directorate. This directorate also encompasses the GIS and Remote Sensing unit; Data Processing, Analysis and Archiving unit; and the Research unit. Furthermore, GSS recognizes the importance of data science by giving one of the two deputy government statisticians the title of Deputy Government Statistician of Economics and Data Science.

The Data Science Unit will be headed by a lead data scientist and will at the outset consist of nine data scientists. Within a year a total of 15 data scientists will be part of the data science unit. The data science unit will serve as an internal consultancy within the service to other units and directorates It is not the intention that the data science unit is running its own projects, but instead plays a demand-driven supporting and capacity building role within different units and directorates. To this end, the data scientist will be embedded in different units/departments of GSS. This will allow them to examine the current methodology of the units and propose new ideas or changes that are needed. The ideas will be discussed among members of the various units. Data scientists will work 3 days with the departments at the main office and 2 days at the data science unit. During these 2 days, they will update each other on the projects they work on and receive relevant training.

While trying to attract talent and people with an aptitude to learn different coding languages, most of the training of the data scientists will happen in-house. With the exception of the lead data scientist and the head of the data science unit, recruits are early-career economists and statisticians. New recruits will be put through a six-month training course consisting of internally developed materials as well as external training. In this training they will be taught different data science skills that they will then apply in different units. In the beginning this data science work will focus on 'low hanging' fruits, such as the automating of data formatting procedures, improving quality assurance, and developing dashboards. Later more complex methods and the analyses of non-traditional data sources will be incorporated into the work of different GSS departments and units. GSS recognizes that this training of new recruits will cost time and to be successful, data science recruits need to be 1) given time to build their capacity and 2) be incentivized to keep working at GSS after their initial training.

06. Skillsets and Training

What Capacity needs to be build.

06

This chapter expands on the various skills that are needed and/or currently lacking at GSS. Training in these skills will happen both internally, through self-study, and through cooperation with external institutions.

R

R is the most commonly used statistical coding language, well suited for data analyses, visualization, and machine learning. Data scientists at GSS would be expected to be familiar with these R Packages: dplyr, data.table, shiny, ggplot2, tidyr, sf, and RCrawler.

Python

Python is a general purpose computer language that can be used to work with a wide variety of data sources. Data scientists at GSS would be expected to be familiar with these Python libraries Pandas, Numpy, Geopandas, Matplotlib, Seaborn, Plotly, Flask, Re, and BeautifulSoup.

Dashboard Software

There is a clear demand for the development of dashboards at GSS. These will be used both internally for the monitoring of project implementation or survey monitoring and externally to communicate statistical products. The technical skills needed to develop and host these dashboards include rmarkdown, shiny, Power BI, GitHub pages, and Netlify.

Big Data Management Software

As data size increases and more sources of data will be used, it will no longer suffice to store data in .csv files. Instead GSS needs to develop both the server hardware and other IT architecture and know-how to store data in databases and use of SQL, NoSQL, and Apache Spark to interact with this data. Setting up and maintaining these different types of databases to host GSS data fall under the responsibility of the Information Technology Directorate. This directorate will have to cooperate closely with the Data Science Unit in the hosting of database systems. Not all data is suitable to be hosted on cloud infrastructure and GSS needs to invest in on-premise data warehousing infrastructure.

GitHub

Code sharing and version control are essential tools in a data scientist's toolbox. GSS envisions the use of git and the online extension of that in the form of Github, to host code and to collaborate on code.

Cloud Computing Software

Cloud computing infrastructure is an online computer system resource that stores large data and allows for powerful computations on that data regardless of the size of the data. The 3 top cloud providers include Amazon Web Services (a publicly available cloud platform), Microsoft Azure (a hybrid-cloud environment), and Google Cloud (a publicly available cloud platform). As GSS strives to be more innovative and conscious in offering real-time data or analytics to meet user needs, moving some of its services to the cloud will be essential to facilitate the timely production and publication of relevant statistics. Not all the data GSS analyses will be owned by GSS nor can all data be stored on-premise, so cloud computing solution will be employed. The technological skills needed to develop and host this cloud approach include the use of Microsoft Azure, Apache Spark, Google BigQuery, DataBricks. Moving to online storage and analyses of data means that in addition to support with hardware issues, the GSS Information Technology Directorate will need to build capacity in the use of and maintenance of virtual machines to support the data science unit's work. Related to cloud computing is the use of High-Performance Computing (HPC). By using specialized processors and data centuries, HPC allows for computations that are not possible on most laptops or desktops.

Data Security

New data sources will need new ways of protecting this data. Whether it is survey data or administrative data, working with data that belongs to individuals or companies requires knowledge on information security. This means that GSS will need both the technical capacity to protect data as well as the statistical skills to disseminate data in such a way that aggregates cannot be linked back to individuals. Hosting and protecting large and diverse sources of data will require secure hardware as well as the management of these servers.

Geospatial analysis

Geospatial and data science provide powerful tools for GSS in relating data to geography. By connecting data with location, GSS can gain greater insight, leading to better decisions and smarter outcomes. The importance of the Geospatial element to data science arises in part due to its ability to help us make sense of the patterns and results that emerge from the analysis. For example, new sources of data like satellite data, new technologies, and new analytical approaches, if applied responsibly, possess the potential to enable more agile, efficient and evidence-based decision-making, and can better measure progress on the SDGs in a way that is both inclusive and fair. The transformation of GIS professionals in GSS to become geospatial data scientists requires capacity building in the following areas, namely, Data Processing/Extract, Transform and Load (ETL), Spatial Structured Query Language (SQL), Geospatial Python, Spatial Data Science, Machine learning and AI. The core geospatial data engineering toolkit is GDAL, which can be run locally, via Python, or through common libraries like Fiona, Rasterio, or Geopandas. The data scientistsmust also develop R and Python programming skills and be familiar with Integrated Development Environments (IDEs) like VS Code, Pycharm, Jupyter Notebook, RStudio and Rmarkdown, but also be able to leverage Google Earth Engine for HPC geospatial analyses.

07. Data Science Use-cases

Potential use-cases and deliverables.

07

This chapter presents a list of some of the envisioned use-cases of data science and big data projects within GSS. This is by no means an exhaustive list of all possible uses of data science but should serve instead as an overview of possible uses of data science that can be implemented in the short to medium term. These use-cases were selected after consultations with various departments and units within GSS to see which projects would most benefit from data science input.

01 ///
Completion of CPI RAP
-

Currently, Consumer Price Index (CPI) data is collected using paper questionnaires, which are then keyed in and analysed by the head office. An effort is being made to enhance the data collection using computer-assisted personal interview (CAPI) application and to set up a reproducible analytical pipeline (RAP) for the computation of indices from raw price data, the quality checks of the data, and the dissemination of the data. Part of this RAP will be expanding the R code for the CPI calculation from 10 to 16 regions and adapting this code to make it more accessible for future modification by GSS.

Furthermore, GSS will work on the development of the CPI dashboard. By introducing a live dashboard showing detailed breakdowns of inflation, GSS will be able to inform both the public and policy makers better about the drivers and trends of inflation. This dashboard will also provide options for improving the accessibility of the reporting outputs and statistics for the public (for example, by ensuring visuals are more accessible for audiences with visual impairments and colour blindness and exploring a range of output formats). This platform will also include a personal inflation calculator for the general public to estimate person specific inflation. This project would be led by the Prices Unit.

Deliverables	**1) Functional CSPro CAPI Application** **2) R Code to quality check and analyse price data** **3) Online dashboard with inflation statistics**
Expected completion	**Q2 - 2023**

02 ///
Census Results Dashboard
-

During the dissemination stage of the 2021 Ghana PHC Census, there arose the need to share the regional and district data tables with stakeholders, institutions, and the general public. The data tables were uploaded onto the GSS website in downloadable Microsoft Excel format. In addition, the regional data tables were included in the published volumes of the general reports. It was agreed to expand access to these valuable data by developing a user-friendly and dynamic user interface to make searching for data easier and more efficient. To this end, the Census Dashboard was developed. It is hosted on GitHub and Heroku platforms (https://www.ghanastat.com).

The Census Dashboard provides results of data search for population, social, demographic, economic, housing, structures, and other indicators based on data from the 2021 PHC census. Moreover, it includes visual highlights culled from the respective report volumes, and a dynamic search feature based on region, district, and other variables. It also contains age pyramids generated for all the respective districts in Ghana. This dashboard will be further developed in the final months of 2022.

Deliverables	**1) Functional Online Dashboard with 2021 PHC Results** **2) Python code to automate the production of tables for the Census dashboard** **3) Code for back-end intergration with PHC results**
Expected completion	**Q4 - 2022**

03 ///
Databank for Census Data and R Package for Census Results
-

Inspired by the US Tidycensus R package and the rKenyaCensus package, GSS will release an R package that contains relevant census summary statistics in a user friendly way and easy to analyse format. As a back-end of this package, GSS will use an online databank that is currently being developed by Statistics Denmark together with GSS. Data from this databank can also serve as the backend to other interfaces for the census data.

Deliverables	**1) R package released on CRAN** **2) Functional online databank based on the StatsDenmark databank with summary tables from the 2021 PHC**
Expected completion	**Q2 - 2023**

As mentioned in the section on the history of data science at GSS, the CDR project has been one of the drivers of data science at GSS. In addition to the mobility analyses that have been released, GSS is working on a further two use-cases of the CDR data.

04 ///
New Use-cases Call Detail Records (CDR)
-

1. The National Disaster Management Organisation's (NADMO) multi-hazard Early Warning System (EWS) currently overlays static census projections to risk layers to estimate the number of people present at incident sites. This does not account for seasonal or even daily travel patterns, and thus provides unrealistic information which consequently suggests deficiencies in pre-disaster and post-disaster planning. CDR data will be able to give hour-by-hour population density estimates at sub-district level across the country, critical for efficient and proportionate resourcing decisions to be made on the ground responses. CDRs provide an opportunity to estimate the number of people at well-defined locations and the volume of traffic/people due to be travelling into that area in subsequent hours, thus being able to anticipate if a situation will escalate.
2. The use of CDRs and other data may provide the opportunity to prepare poverty estimates at the lower level with the possibility of annual updates. Poverty estimates are critical for the effective execution and monitoring of social intervention programs, especially poverty reduction programmes. Until recently, when the Annual Household Income and Expenditure Survey (AHIES) was rolled-out to provide expenditure aggregates for poverty estimation, the Ghana Living Standard Survey (GLSS) which is conducted every 5 years had remained the main source of data for poverty estimation. Both surveys have geographical coverage limitations, as sample representation is at the regional level and does not provide district level estimates. The limitation is largely due to the high financial and logistical cost associated with large data collection programmes such as a census. This situation has often left district actors to use regional poverty estimates in their planning, which may not be realistic. Secondly, the irregularity of these estimates also makes it difficult for districts to evaluate interventions within the usual 4-year planning cycle. Computation and publication of annual and district poverty estimates will greatly enhance the ability of districts to incorporate some local level estimates in their work annual programmes. The methodology for generating poverty maps will adopt the combination of proxy variables; the amount of airtime and the amount of MOMO on subscribers' accounts combined with housing characteristics from satellite data, survey and census data.

Deliverables	**1) Completed use-case with NADMO** **2) Released local poverty estimation**
Expected completion	**Ongoing**

05 ///
Additional Use of Satellite Data for Ecological Statistics

The use of image processing to provide relevant information is one of the data science use-cases at GSS. Ghana fights illegal mining (galamsey) that harms water bodies, forests, and agricultural farmland. GSS can use machine learning technology to analyse satellite images in areas where these activities are taking place. Because machine learning will be able to see patterns that the naked eye cannot, it will aid in the fight against illegal mining. Initial steps to this use-case have been made in a project between GSS, Centre for Remote Sensing and Geographic Information Services (CERGIS), NASA, and MIT.
https://doi.org/10.1016/j.scitotenv.2021.146644.

Deliverables	1)	**Explore new cases for the use of remote-sensing based statistics**
	2)	**Release at least one new remote sensing based report**
Expected completion	Q1 - 2024	

06 ///
District Administrative Data

Statistical data is only impactful if it is accessible to policy makers. To localize access to administrative data, GSS is working on a project to collate and summarize yearly administrative data from various Ministries Departments and Agencies (MDA's) at the district level. Data Science techniques will help with the streamlining and quality assuring of this data collection. Currently, this administrative data is collected using a paper template in all MDAs which are keyed and analyzed at the head office. An effort is being made to enhance administrative data within the National Statistical System (NSS) using computer-assisted data collection applications with the help of a tablet or online portals instead of the use of paper. Plans to set up data pipelines to link data from the various MDAs to the National Statistics Office (GSS). The following will be embedded in the data collection application, quality checks application, and automation of data analysis for the graphical display to avoid the manual generation of these data. This project would be led by the Demographic section of the Social and Demographic Statistics Directorate Section (SDSD).

Deliverables	1)	**Develop a platform for providers of regional administrative data to upload data**
	2)	**Develop code to create automated district level reports**
Expected completion	Q1 - 2024	

07 ///
Launch of Updated National Reporting Platform SDG -

The National Reporting Platform (NRP) for the Sustainable Development Goals represents an incomparably unique dashboard tool that enables the GSS to exhibit the data on Ghana's progress towards the SDGs and other development indicators. Furthermore, this dashboard for the SDGs helps in accomplishing the GSS 2020 – 2024 Corporate Plan, specifically Goal 5: Improve the production and use of Official Statistics for national development and planning. Ghana's NRP was initially launched in November 2018 during the African Statistics Day Celebration with approximately 80 indicators. However, due to emerging global practices, Ghana opted to adopt the Open SDGs Platform developed by both the USA and the UK.

Currently, with support from the Harmonizing and Improving Statistics in West Africa Project (HISWAP) and the Office for National Statistics (ONS), UK, work is ongoing to ensure that the NRP is relaunched as an Open SDGs dashboard during the week of the 2022 African Statistics Day commemoration, with a target of 120 indicators. This is vis-à-vis Ghana's Voluntary National Review Report on the nation's progress in successfully meeting 102 indicators across the 17 goals, which was released in June 2022 in New York, USA. The Voluntary National Review Report is important because it documents data on how Ghana is progressing in accomplishing the SDGs. However, in comparison and contrast, the National Reporting Platform is even more potent because of its unique visualization features and potential.

One of the most impactful ways to disseminate data is through visualizations. The National Reporting Platform (NRP) for disseminating Ghana's Sustainable Development Goals (SDG) data provides such an opportunity because it is an open source (GitHub) that presents data in disaggregation, time series, open, transparent and re-usable formats. Other vital features include interactive maps and metadata that describe the data provided. Another key feature is the Reporting Status tab which provides information on the reporting status of each of the seventeen (17) Sustainable Development Goals.

Ultimately, the SDG interactive dashboard (i.e. NRP) is a data-driven initiative that enables the tracking and monitoring of the United Nation's global goals. Furthermore, it gives a better understanding of the SDGs, and the data for achieving them form the basis for informed, evidence-based, effective planning and decision-making.

Data Science is not only highly essential to the NRP, but also virtually indispensable to its viability, development, and sustainability in several notable ways. As an illustration, the aforementioned key features of the NRP, namely: an instant visualization dashboard tool in terms of data and map display, among others, and its interactive nature, are all premised on the principles of data science. This implies that data science is indispensable to the present and future of the NRP platform in several ways such as:

- exploring novel data sources,
- developing new methods and tools for measuring indicators
- improving statistics that support SDG-relevant policies

In another instance, data science is very instrumental in developing disaggregated indicators, to ensure that people who are potentially at risk of being disadvantaged because of their demographic characteristics such as location, social status, or socio-economic status, among others, are recognized. This would help in making the NRP's visualization properties even more impactful.

Also, data science would enable Ghana Statistical Service to utilize modern methods which include machine learning and distributed data processing, among others, to exploit new and alternative data sources. Such new data sources include social media, mobile phone data such as the Call Detail Records, and satellite imagery data, among others. All these aforementioned products would make the NRP dashboard more meaningful, useful and purposeful.

Deliverables	**1) Launch of new NRP with 120 Indicators** **2) Yearly updates on indicators on platform**
Expected completion	**Q4 - 2022**

08 ///
Automate Quality Checks of Trade Data
-

The external trade statistics is produced using customs declaration from the Integrated Customs Management Systems (ICUMS). Currently, import and exports records are manually downloaded for every month, processed, and imported into the EUROTRACE software for processing and reporting external trade statistics. Due to the size and the format of the data, manual validation and formatting the data in Microsoft Excel is a time consuming task. Data science can be used to automate the formatting of trade records into the appropriate data format and to perform data quality assurance on the key variables for trade statistics reporting before the processing into the EUROTRACE.

Deliverables

1) **R code to automatically create a quality report on monthly trade data**
2) **R code to format monthly trade data to EUROTRACE format**

Expected completion **Q1 - 2023**

09 ///
Web Scraping for Labour Force and CPI Statistics
-

The GSS over the years was using only the paper-assisted personal interview (PAPI) to collect data for its use. It has however adopted the computer-assisted personal interview (CAPI) for its recent surveys. The CAPI mode of data collection is far superior to the PAPI in terms of data velocity and ability to monitor data in near real time, identify errors in data inputting and get errors corrected in real time. Both PAPI and CAPI are modes of collecting data from primary sources. There is however a chunk of data or information on the cloud or web that are rich in content and large in scope. Example of information on the web includes data on employment, sales, business, education, and so on. Given the limitation of PAPI and CAPI tools in collecting this data, the GSS has not been able to incorporate data on the web in the production of statistics. Data science will be used to break this gap and thus make it possible for the GSS to compliment statistics using information from the web.

The Data Science Unit will write computer programmes that will search for and fetch some targeted data from the web for analysis. Data that will be scrapped from the web can be used to compliment statistics on CPI, and some other labour force and marketing statistics. The GSS will therefore need the support of institutions or firms hosting targeted websites to realize this objective. The GSS will leverage its partnership with the UN African Regional Hub and the Office of United Kingdom National Statistics for the 'Web scraping for CPI project' to build capacity for this task.

Deliverables

1) **Join ONS/UN project on webscraping**
2) **Produce at least one report on webscraped statistics**

Expected completion **Q1 - 2024**

10 ///
Develop a Standardized Reporting Dashboard for Census/Surveys Fieldwork Monitoring
-

For the first time in the history of censuses and surveys in GSS, a data science tool was deployed to track, monitor and report on data collection and field activities in real-time during the 2021 Ghana PHC. The Census Dashboard was hugely instrumental in providing insights into the quality of data collected, the extent of coverage, performance and progress of field activities at the EA, district, regional and national levels. It also served as an effective tool for tracking the movement of field workers and teams. Data figures were monitored in real-time from the commencement of the census to its completion.

Providing diversities of views of the data collection process using charts and visualizations, the census dashboard greatly assisted in the triggering of timely interventions, where possible, in the data collection process to ensure data quality and completeness of coverage.
Based on the success of the census dashboard, a similar dashboard was also deployed for the Post-Enumeration Survey. The project implementation team used the PES dashboard heavily in their daily briefing in the evenings and assessed the daily progress reports submitted to the field supervisors.

The data science unit, based on the successes of the aforementioned data science tools, will help work with different directorates to develop a standardized tracking and monitoring dashboard for all censuses and surveys in the Service. Dashboards will seek to incorporate all the standardized tools, features and technologies of each of the above dashboards to develop a reusable, user-friendly, effective and high-performance fieldwork tracking and monitoring dashboard for all data collection activities in GSS.
The fundamental purpose of the dashboard will be to monitor key vital statistics in the datasets and flag problem areas, where necessary, suggesting interventions in those respective 'problem spots'.

Deliverables	**1) R code with a template for production of a dashboard for monitoring of survey field work.** **2) A pipeline to upload data from field work tablets to an online database which feeds data into a dashboards.**
Expected completion	**Q4 - 2022**

Once operational, the data science unit will present a monthly update on its work and inform the public of a new statistical release. The primary objective of the monthly release is to offer actionable insights to the public whilst promoting projects that will encourage more collaborations with other stakeholders. In addition to the release, the unit will be writing and publishing manuscripts, journal articles or policy briefs.

11 /// Monthly Data Science Updates -

Deliverables	**1)**	**Monthly release on what the GSS Data Science Unit has been working on**
Expected completion	**Q1 - 2023**	

The Agriculture and Environmental Unit has generated the national and district tables from the 2018 Ghana Census of Agriculture (GCA). The tables have been generated at the national, regional and district levels.

In order to provide public access to the data, the unit intends to have a dashboard developed to provide an interactive interface for searching and viewing the data.

12 /// Ghana Census of Agriculture (GCA) Data Dashboard -

Deliverables	**1)**	**Dashboard to be released during African Statistics Day 2022**
Expected completion	**Q4 - 2022**	

The Agriculture and Environment Unit receives data on wholesale and resale prices of a variety of agricultural commodities on monthly basis from Ministry of Agriculture (MoFa). The MoFa obtains these data through a survey and forwards a set of tables generated from the data to Ghana Statistical Service.

The Data Science Unit intends to have a dashboard that will automate the receiving of the data from MoFa through a pipeline and the display of the various statistics on the dashboard.

13 /// Dashboard on Agriculture Commodity Prices -

Deliverables	**1)**	**Dashboard to released during African Statistics Day 2022**
Expected completion	**Q4 - 2022**	

14 ///
Continuous Updating of Housing Stock
-

Government's intention to improve revenue mobilization by widening its coverage to include collection of property rates necessitates regular update of data on housing stock, which GSS estimated from recent 2021 Population and Housing Census. GSS aims to deploy non-traditional data sources such as satellite imagery and drone footage, to frequently benchmark estimated housing stock between decennial censuses. Frequently updated housing stock serves multiple purposes, for example:

1. The drone and satellite imagery technology would provide knowledge of new settlements, (especially rural area) which would consequently be included to improve coverage and estimates of sampling frames for other survey work.
2. Assisting Ghana Post in the assigning of digital addresses. Ghana Post GPS is a national digital address system which covers every property within the borders of Ghana. It has become a requirement to provide this address system in many registration processes, including accessing loans from financial institutions and other registration processes such as Ghana Card which has now become the basic identification card to start other registration processes. This has made it almost compulsory for everyone living in Ghana to have the Ghana Post GPS address irrespective of their location in Ghana.
3. Related to the assigning of digital addresses, better estimates of the housing stocks can help with revenue realization through property taxes. Better estimates of both the locations of properties and information such as the type of properties, average the number of rooms and other characteristics that are considered in evaluating the property rates can provide information on the expected property rates to be collected per administrative district to the Ghana Revenue Authority (GRA).

Deliverables	**1) Completed pipeline using satellite data and other non-tradional data to update housing stock on a yearly basis** **2) Dashboard with updated housing stock information**
Expected completion	**Q1 - 2024**

15 ///
Marine Litter Estimation with EPA

-

Marine litter is defined as any solid waste dumped in the sea or on beaches, either intentionally or unintentionally, by agents such as humans, animals, wind, or other bodies of water. Plastic waste is considered enormous among all marine waste, accounting for 80% of all marine waste. As a result, GSS should play a role in the monitoring and creation of statistics around marine waste. However, due to the vastness of oceans and coastlines, traditional methods of gathering marine data, such surveys, are not always feasible. Especially not if data collection needs to be temporal as well as spatially accurate. As a result, non-traditional methods such as citizen science projects and satellite data are being used to monitor marine waste. This use of citizen science, via the "Citizen Science for the SDGs in Ghana (CS4SDGs)" project made Ghana the first country to officially report on SDG indicator 14.1.1b on marine litter using citizen science data. The project has also helped to bridge local data collection efforts of the public and volunteers with global monitoring processes by leveraging the SDG framework. Using such non-traditional data sources to estimate official statistics requires statistical skills that go beyond the analyses of survey results and requires the use of data science techniques. In the coming years, GSS, together with the Enviromental Protection Agency (EPA), as country leads on this indicator will continue to employ citizen science and other non-traditional data sources to estimate various environmental SDG indicators including 14.1.1b.

Deliverables	**1) Continued stewardship at the country co-lead on the use of citizen science to estimate environmental SDG indicators**
Expected completion	**Q1 - 2024**

08. External Stakeholders
Who will GSS be working with.

08

It is important to recognize that setting up a data science policy and the creation of statistics using big data requires cooperation both within GSS and externally with other stakeholders. This chapter summarizes the envisioned external stakeholders GSS will cooperate with. Some of these stakeholders are needed to get access to data, while others will assist in the analyses of data or the creation of new methodologies

01
African Union

Africa is crowded with a number of social and environmental problems such as poverty, inequality, limited access to quality education, unemployment, inadequate access to energy, hunger, pollution, climate change, etc. Finding collective solutions to these problems is core to the existence of the African Union. Big data is needed to get a better understanding of the issues and strategies to solve these problems. Member countries could, therefore, consolidate efforts to create the habit of storing data in a usable manner and making them accessible to African statistical offices for use. This will enable national statistical institutions to produce new statistics aimed at guiding policies to solve some of these African problems.

02
Statistical Commission for Africa

The Economic and Social Council United Nations Statistical Commission for Africa is currently developing a roadmap for the transformation and modernization of official statistics in Africa, which is expected to be released in March 2023. This roadmap aims “to support the development of national road maps and facilitate action planning. It provides guidance on ways in which countries can develop and implement national strategies for the development of statistics and annual work plans to optimize transformation and modernization. It also provides guidance on how to determine the desired state of a transformed and modernized national statistical system.”

source: https://papersmart.uneca.org/download/4693

03
Office of United Kingdom National Statistics (ONS)

ONS is a key development partner for capacity building at GSS. Through capacity building training and mentoring programmes, they can help GSS in developing technical skills as well as assist in the work on the different use-cases. Furthermore, lessons can be learned from the way ONS set up its own data science campus and assisted in setting up a data science hub in Rwanda for the National Institute of Statistics of Rwanda (NISR).

04
Statistics Denmark

Statistics Denmark is a key partner in capacity building for GSS. Furthermore Statistics Denmark is assisting GSS with the set up of a data bank system for the distribution of census data.

05
United Nations Big Data Regional Hub for Africa

Based in Kigali, Rwanda this is the African hub for the United Nations BigData programme, which aims to educate, collaborate and develop new technologies to work with new Big Data sources and methodologies. It will foster "international collaboration in the development of Official Statistics using new data sources and innovative methods and to help countries measure the Sustainable Development Goals (SDGs) to deliver the 2030 Sustainable Development Agenda." GSS will leverage its existence to collaborate with sister African countries that buy into the need for the use of data science in national statistical offices. This will create a platform for sister statistical offices to support each other in capacity building of data scientists, share methodologies, share success stories on new and efficient ways of working, and collaborate in developing and improving methods of generating official statistics efficiently,

06
Mobile Telecom Providers

As explained in the section on the history of data science, GSS has a running cooperation with Vodafone to provide aggregated and anonymized call detail records for the computation of mobility statistics. GSS also ran a pilot with MTN and Dahlberg on the use of CDR data to estimate labour participation statistics. CDR Data spans from the type of services clients use, frequency of use of service, duration of use of service, some business information, mobile money usage, etc. This goes to show that call detail records are an important source of data for a variety of statistics. Therefore, GSS will continuously strengthen cooperation with all telecom providers and the Ghana Chamber of Telecommunications.

07
Academia

A mutually beneficial way of building data science and statistical skills is by fostering cooperation between GSS and university programmes that offer (masters) degrees in data science or statistics. GSS can familiarise students with the system of official statistics, production models, statistical methods and dissemination while benefiting from students' master's theses and graduate internships on new statistical and data science methodologies. GSS could provide certification to students who successfully complete such a programme. In addition, the GSS data science team and facilitators of data science programmes in academic institutions could support each other in the training of new recruits of data scientists in GSS and graduate students. GSS could again support academia in the development of more relevant data science curricula. This will help produce data scientists who are capable of feeding the GSS data science directorate and be relevant to other institutions in need of data scientists.

08
Private Sector

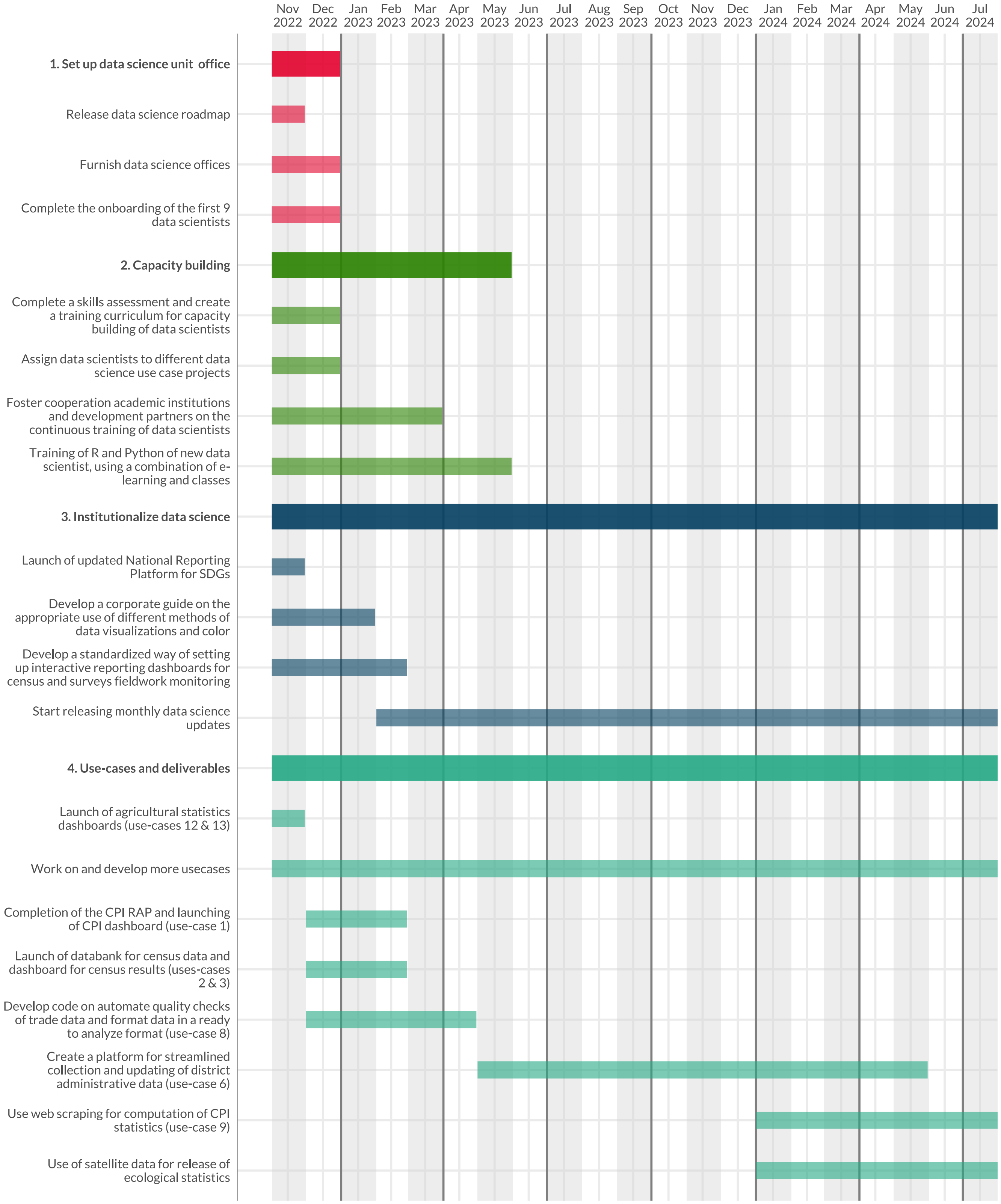
Collaboration with the private sector is key to GSS in its quest to produce relevant statistics which currently cannot be produced. The private sector (be it formal or informal) has large volumes of data that GSS currently does not have access to. Some of the private sector data might come in the form of well-structured survey data (for example, customer satisfaction surveys) or structured databases (such as air pollution sensor data), while other data might be in the form of administrative data (such as banking records), and finally, data might be unstructured, but still informative (examples could include traffic cameras, private satellite data, or drone data). Extracting statistics from and gaining access to any of these sources requires, in addition to technical know-how, partnership building with different holders of private sector data and the development of proper data sharing agreements. GSS will be able to build on its experience in developing data sharing agreements with different telecom providers (Vodafone and MTN), to be able to sign proper data sharing agreements with private sector actors.

09
Government Agencies

As the steward of the NSS and while moving towards the use of more and more administrative data (as explained in chapter 3), GSS needs to coordinate with different government institutions and departments. Institutions that hold valuable administrative data include the Ghana Revenue Authority (GRA), Ghana Immigration Service (GIS), Driver and Vehicle Licensing Agency (DVLA), Ghana Education Service (GES), and the Registrar General's Department (RGD). Where government data is not yet in a digitized form, GSS could train and encourage government agencies to store data in soft/digital form and make them accessible to GSS. Such data will enable GSS to realize its mandate of guiding government policies through data and statistics. Government agencies will thus benefit directly from statistics and draw insights into new ways of doing things. GSS can also help agencies automate some of their work.

08. Timeline and Actions

What are the next steps.


Nov 2022
Dec 2022
Jan 2023
Feb 2023
Mar 2023
Apr 2023
May 2023
Jun 2023
Jul 2023
Aug 2023
Sep 2023
Oct 2023
Nov 2023
Dec 2023
Jan 2024
Feb 2024
Mar 2024
Apr 2024
May 2024
Jun 2024
Jul 2024
1. Set up data science unit office
Release data science roadmap
Furnish data science offices
Complete the onboarding of the first 9 data scientists
2. Capacity building
Complete a skills assessment and create a training curriculum for capacity building of data scientists
Assign data scientists to different data science use case projects
Foster cooperation academic institutions and development partners on the continuous training of data scientists
Training of R and Python of new data scientist, using a combination of e-learning and classes
3. Institutionalize data science
Launch of updated National Reporting Platform for SDGs
Develop a corporate guide on the appropriate use of different methods of data visualizations and color
Develop a standardized way of setting up interactive reporting dashboards for census and surveys fieldwork monitoring
Start releasing monthly data science updates
4. Use-cases and deliverables
Launch of agricultural statistics dashboards (use-cases 12 & 13)
Work on and develop more usecases
Completion of the CPI RAP and launching of CPI dashboard (use-case 1)
Launch of databank for census data and dashboard for census results (uses-cases 2 & 3)
Develop code on automate quality checks of trade data and format data in a ready to analyze format (use-case 8)
Create a platform for streamlined collection and updating of district administrative data (use-case 6)
Use web scraping for computation of CPI statistics (use-case 9)
Use of satellite data for release of ecological statistics