TECHNICAL BRIEF

USING A DATA SCIENCE APPROACH TO BUILD TIMELY, SUSTAINABLE, REPEATABLE AND USER-CENTERED ANALYSIS TO DRIVE ACTIONS

ABBREVIATIONS

LMIS – Logistics Management Information System
CRISP-DM – Cross Industry Standard Process for Data Mining
DevOps – Development Operations
M&E – Monitoring and Evaluation
MIS – Management Information System
GHSC-PSM – Global Health Supply Chain Program-Procurement and Supply Management
ROI – Return on Investment

BACKGROUND

Every day public health supply chain operators around the world make thousands of decisions and take related actions. The outcomes for some decisions are seen the next day, while others aren't visible for months. Thus, one of the most important resources for any supply chain decision maker—from the facility staff placing an order to the supply chain director signing for a year's procurement—is the data used to make decisions. Beyond just having data visible, they need to have insights into the story the data is telling to support the decisions that need to be made. Critically, the timing of the decision is just as important as making it in the first place. Not only do decision makers need timely data visibility, but they also need timely analysis of the trends and patterns in the data to inform those decisions. This necessitates having timely data analytics tools that can support that decision making.

This technical brief explores how to strengthen monitoring and evaluation (M&E) processes through a data science approach to analytics that enables supply chain decision makers to act based on timely, transparent, and repeatable analysis. Data science approaches use code to gain and automate insights into a dataset or multiple data sets that identify information to aid in decision making with more consistent and timely insights than many other approaches.

SUPPLY CHAIN MONITORING AND EVALUATION

Supply chain M&E helps to better understand a problem and its root causes and to target actions toward its resolution; for example, this could be understanding why certain warehouses or health facilities are stocking out of key products, or why products are expiring, and preventing these negative outcomes. It can be used to improve supply chain functioning by tracking progress against a benchmark or toward a measurable goal, such as maintaining an order fulfillment rate of more than 95 percent.

Supply chain M&E aims to collect and analyze data in a way that benefits the specific needs of the system's users, whether it be health facility pharmacists, regional health program managers, or donor organizations monitoring trends in project performance. Each stakeholder group may have different data needs; therefore, the types of measurements and analysis should be targeted accordingly. This entails understanding the nature of the data and the frequency with which the data is produced and made available for analysis. Crucially, it also means that data must be available to users in a timeframe that meets the needs of decision makers.

Some of the key challenges to making supply chain M&E operational for different user groups are the timeliness of the analysis for decision making, and the level of effort needed to conduct the analysis, which also impacts the timeliness. With limited resources available for M&E, labor-intensive manual calculation processes and the use of static reports and dashboards could mean that M&E systems must prioritize the longer-term monitoring needs of higher levels of the supply chain over the ability to address time-sensitive problems at the district or facility levels to tackle root causes of supply chain underperformance. A data science approach targets this need for timely, repeatable, and transparent analysis, providing supply chain decision makers with insights to support their decisions today, as they cannot wait for tomorrow.

A DATA SCIENCE APPROACH

Using a data science approach has two impacts: making analysis of supply chain data timely for decision makers and freeing up resources so that M&E can be tailored to the needs of stakeholders. This allows the analysis to match the timeframes when decisions must be made and enables analysts to focus on interpretation of the data rather than data collection and reporting. This can also bridge the rift often seen between M&E analysis and supply chain operations, so that both teams can work together to interpret data and make the connection on how to use the data to improve supply chain efficiency.

With a data science approach, a wide range of tools, both open source and proprietary, can be used to clean, manipulate, and perform calculations on datasets to produce meaningful indicators. The key feature of data science approaches is that they are written in code (e.g., Python, R, JS, Julia) that can be used repeatably and are transparent in the steps being undertaken by the analysis. The advantage is that they can be designed and set up before the data is available (using historical data) and they can then, once the data is available, be run and results made available as soon as decision makers need them.

Making the results available has two steps: the automated process to get the data and run the analysis and how the results are then made available to be incorporated with existing business processes and updating SOPs. Thus, it is advantageous to design a data science approach in collaboration with other teams in the supply chain, specifically the interaction between the data scientist (M&E teams) and the development operations team (management information system, or MIS, team), also known as DevOps. DevOps provides a foundation for the data behind the data science, and data science provides insights behind the data being used. Without a strong connection between the two, the potential of both is not realized.

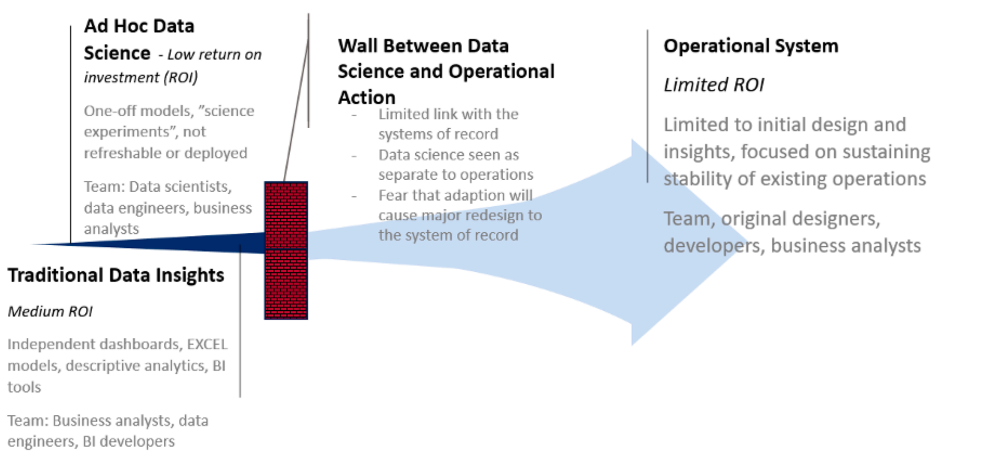

FIGURE 1. Separation of data science and development operations

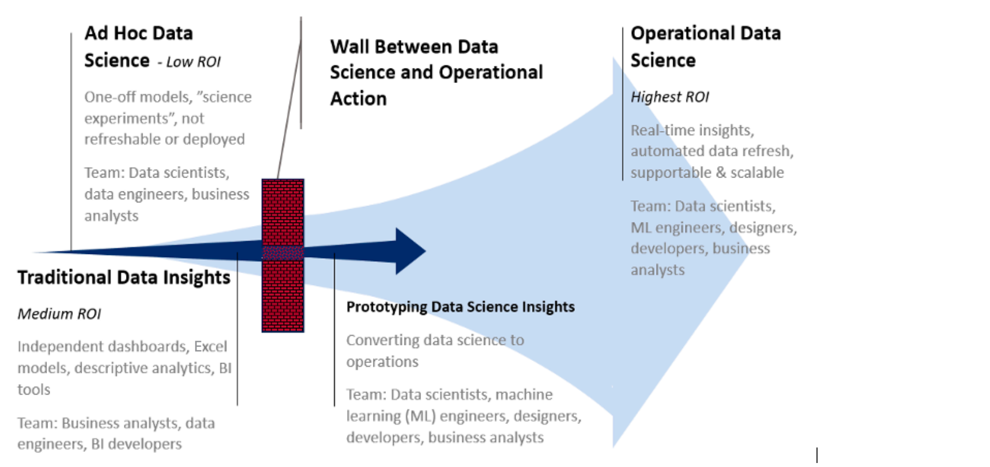

FIGURE 2. Connections of data science and development operations

Once an effective data science method is established, its connection with existing systems and management processes is essential to its adoption and scaling within the supply chain. Therefore, when analytics is conducted with a strong connection to DevOps, it facilitates turning data science approaches into operational approaches that can be used for timely decision making. How this connection is built and maintained depends on the decision that is being supported. In some cases, it is a scheduling of standard data reporting from the information system that is then used in the analysis; in other cases, it might be the exchange of data—both input and output—with the system, or even integrating the analysis tool within the system. When looking to provide analysis in a repeatable and scalable manner, the analysis is designed to work with information systems—strengthening them by using their data—and, where appropriate, identifying a path to greater integration within the existing system and processes. GHSC-PSM has focused the development of data science approaches with three fundamental aspects to enable the tools to be both repeatable and scalable:

- **Design for flexibility:** Data inputs from existing information systems are to be flexible to accept different data input sources. It is assumed that an information system is stable, however it may evolve, thus it needs to allow for changes to be made.
- **Design with open source:** For transparency and repeatability of the analysis, its code needs to be accessible so that users can review how the analysis is being done. Additionally, the advent of cloud computing opens the opportunity to build open-source cloud applications that make it possible to generate an interactive application for decision makers to use.
- **Design with decisions at the center of the design:** To promote successful use of the analysis, the outputs should be designed with the end user and the decision/action being supported. This requires understanding the characteristics of the decision at hand, then designing the analysis to match those characteristics.

The targeted decisions drive analysis design. Therefore, there is a need to understand the characteristics of those decisions:

- What are the decisions?
- Where are the decisions made?
- When are the decisions made?
- Who is making the decisions?
- What are the actions that will occur as a result of the decisions?

Before analysis design, these fundamental aspects need to be understood, as they will drive many of the analytics decisions that will occur in the design process. For example, if the decision is made on a weekly basis, then data collection cannot be burdensome as it will prevent the timely analysis. This information needs to be factored into the analysis design.

The importance of designing analysis to meet the needs of decision makers is not a new concept. The Cross Industry Standard Process for Data Mining (CRISP-DM) was developed in 1996 as an open standard process model that describes common approaches used by data mining experts. This process was designed agnostic to any specific tools in mind and before many of the data science tools that are commonly used today. The ideas behind CRISP-DM are still as relevant today as they were in 1996, as they focus on designing data analysis to meet the needs of the user and creating a feedback mechanism to improve the analysis. This aligns with the key components of M&E in that it evaluates the data and analysis processes to continually improve the system and demonstrate evidence of such improvement.

THE CRISP-DM METHODOLOGY

In this section we delve into the components of the CRISP-DM methodology for data mining models for supply chain indicators. The six components are: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. While each of the components are listed in sequence of progression, in actual development, an earlier stage may be repeated to refine the analysis.

FIGURE 3. Connections of data science and development operations

CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING

For example, additional input parameters may be deemed necessary at the evaluation stage, thus requiring the data preparation and modeling steps to be repeated.

BUSINESS UNDERSTANDING

Understanding the business problem at hand is critical to ensure that analysis is actionable. Decisions stemming from the business problem will drive the design and goal of the data analysis, so it must be clearly defined in order to develop an effective action plan. Defining the business problem may include identifying bottlenecks, defining priorities, or implementing process improvement.

Several analytics tools created under GHSC-PSM serve as examples of addressing commonly seen business problems.

- The Zambia Consumption Anomaly Detection tool had an analytical bottleneck whereby new data was only being processed by a home office analyst in the United States, increasing the lag between data availability and analysis availability. After understanding this business process and identifying the bottleneck, it was resolved by updating the process so it could be run completely in-country, reducing the lag time as a result.
- The Guinea Inventory Supervisory Dashboard had a business problem that involved identifying facilities for supervisory intervention, but it required a way to give certain actions priority over others. Understanding the business process enabled the implementation of a scoring system that raises facilities with the greatest need for supervision to the top of the list. Business understanding of an analytics process can also highlight areas for process improvement.
- Like the tools mentioned above, the Ethiopia Inventory Analysis tool identifies anomalies and flags them for prioritized action and includes an added process for combining data from multiple sources. Data is spread across several different datasets which may have misalignments; for example, facility names may differ between sources. These misalignments can be a barrier to performing data analysis. The tool is an example of process improvement in that it reduces the time and effort required to link the data across sources, removing the barrier. The process is automated where possible, and where human input is required, includes information for review by the user, complete with suggestions to aid them in the task.
- One important function of M&E is to enable data users to identify risks to performance so that they can be mitigated. This includes identifying where, when, who, and what actions are likely needed. This aligns with the business understanding phase, as it means understanding the problems to be addressed, the business processes involved, and channels by which to target mitigation strategies.

- In the supply chain context, M&E provides the foundation of this business problem in context, defined through one or more indicators, such as stockout rate, reporting rate, rate of product loss, order fulfillment rate, etc., that can be routinely measured using readily available data. Defining the problem using common supply chain indicators can then help understand the extent of the matter.

Once problems have been identified and prioritized, the business understanding should specify the mechanisms or channels by which that analyzed data can be used to target actions to improve the system, such as supervision visits, trainings, or adjusting order or distribution processes. Understanding the business processes themselves could mean reviewing the frequency of commodity distributions to various levels of the supply chain, order frequency, or inventory management and reporting procedures. The outcome is then determined by when, where, and by whom decisions need to be made. For example, analysis may be used to support determining where and when to prioritize supervision visits. Reviewing these processes can help determine what kind of data analysis is needed to support actions that can be targeted.

The role of M&E in the business understanding phase is working with key decision makers and data scientists to align the analysis to meet the business need and the operations need to drive timely decisions. It is also to achieve impact through the analysis by setting measurable objectives such as reducing stockouts or stock shortages, reducing product expiry, increasing the order fulfillment rate, or increasing on-time delivery to health facilities. Thus, the feedback from the analysis improves the business understanding and overall outcomes.

Data and business understanding together are critical for setting any benchmarks that will be used in the analysis.

Business understanding could mean recognizing that there are too many sites to supervise with the available resources, and that those most at risk must be prioritized.

For example, in Ghana, facilities face a range of challenges, all of which affect consistency and reliability of supply. Understanding inventory flow was a challenge, as evidenced by continuous stockouts of some products and overstocks of others. A review of the information flow suggested that a timelier analysis to identify patterns within site stocking practices would enable better-targeted supervision to high-priority sites at the right time. Previously, sites visited after being identified as having a stockout responded that they had since re-stocked (or that a delivery was on the way). Stockouts are a lagging indicator, and thus more timely indicators are needed to identify repeatedly poor-performing sites. Supervisors need to be targeted in their actions and prioritize sites for supervision. Additionally, the actions required at the targeted sites need to be identified. For example, for sites with stockout risk, actions may include facility inventory and order management. Conversely, if there is overstocking, an action could be exploring opportunities for redistribution between sites.

DATA UNDERSTANDING

In the data understanding stage, the goal is to identify what sources are available and necessary as input for the analysis model and how the data will be used to perform analysis for decision making to address the business problem. This stage of the process typically involves communicating cross-functionally to gather information. It is important to understand the systems that house the data, the frequencies at which they are updated, as well as the relationships and hierarchies between the features in the data sources that are considered. Beyond this, it is important to understand the existing decisions the data is supporting, and the processes associated with those decisions.

Data and business understanding together are critical for setting any benchmarks that will be used in the analysis. For example, if the analysis is targeted toward understanding the flow of inventory, then the business understanding will inform desired inventory turnover ratio targets. The data understanding identifies the data sources and flows that can support an analysis that will meet the needs of decision makers. It is worthwhile to note that while the purpose of the analysis is to target actions, it will be built on the foundation of the existing data used within the information systems. Unless the analysis is specifically designed for helping improve data quality, it will either assume the data is accurate and ready to use or look to identify and exclude data that does not meet user-set requirements for accuracy. The data mining model should only be responsible for validating, not correcting, the input data; the source of data should remain the single source of truth.

As within the business understanding context, M&E supports the understanding of data being used to generate existing metrics such as stockout rates. For example, the same data sources used to generate stockout rates can be analyzed using different approaches to provide additional insights. Inventory turns and consumption stability, for example, can be calculated from basic stock data routinely reported in most logistics management information systems (LMIS), and together these two metrics can help predict the likelihood of stockouts or overstocks. Thus, the same data can be used in multiple ways to provide greater value in meeting the decision makers' needs.

In the Ghana example, supply chain managers had good visibility into a large volume of data reported every other month to the LMIS but lacked the tools to be able to quickly interpret the patterns in the data for everyday use. The Ghana Supervision Tool combined the inventory turn rate by product type, and the coefficient of variance, which measures the stability of the consumption. The interaction of these two indicators together signals the type of stock risks (e.g., overstocking or stockouts) faced by a site.

For example, a site with a high rate of inventory turns and unstable consumption for a particular product is likely to experience stockouts. GHSC-PSM and USAID Ghana established a scoring system that identified sites that were at elevated risk of overstocking or stockouts, with a target score range where risks of both outcomes are minimized. As stockouts are a greater immediate risk to disrupting services than overstocks, those sites with the greatest risk of stockouts (the highest rates of both turnover and instability) should be in the top priority group.

When designing the data preparation, it is not highly tailored to one specific data source at a certain point in time, as data sources evolve over time. The intent is for the data preparation to be more robust to be able to handle such changes in data sources without the need to update code.

Local context such as the number of sites reporting into the system, those able to receive supervision visits, and the number of personnel available to conduct supervisions, will help determine how to set the thresholds for prioritizing site visits. Understanding the data, together with the business understanding, enables the analysis to provide recommended actions, such as increasing order quantities or distribution frequency for sites with high inventory turnover. Awareness of high consumption instability, for example, could more quickly identify sites for close monitoring of their consumption patterns.

DATA PREPARATION

Once the input sources and data process and flow are identified, the data wrangling informed by discovery and data exploration analysis can begin. The initial data understanding informs what variables need to be derived, removed, or transformed. This may include decisions on whether to fill in blank values with zero, exclude outliers, or normalize variables. For text variables, it could mean transformations like removing extra whitespace and normalizing capitalization.

When designing the data preparation, it is not highly tailored to one specific data source at a certain point in time, as data sources evolve over time. The intent is for the data preparation to be more robust to be able to handle such changes in data sources without the need to update code. Therefore, these preparations are designed to be independent of the data source. These changes and checks are applied universally—not ad hoc depending on the specific, known nuances of any one dataset. The goal is to make the process as agnostic to variation across datasets as possible to promote broader use, ease of use, and repeatability.

To achieve this agnosticism, however, the process requires some user input on a per-dataset basis that will account for the differences between the potential datasets the process could be run with. Figure 4 shows the graphical user interface that the user interacts with upon startup of an inventory analysis tool. Here the user must select which column headers from their dataset correspond to the tool's required attributes. For example, product consumption can have many names (e.g., quantity dispensed, amount issued) and it requires user knowledge to point the tool to the correct column containing consumption. These user-specified values are then assigned to generalized variables in the analysis which allow it to remain flexible to different input data.

For example, activities for Ghana and Nepal had a similar need for an analysis of facility inventory that prioritized sites for supervision based on the consistency of their consumption patterns, and the rate of inventory turnover of each product at each site. Initially, these analyses were developed separately, and each solution was tailored to the specific data inputs of each country's LMIS data. The data science approach was to build upon Ghana's and Nepal's analytical codes by refactoring them to create a generalized solution that could be used with any LMIS data extract that contained the necessary attributes for performing the inventory analysis. This meant creating the solution to configure different data sources to enable the inventory analysis to be done without adjusting any code.

A key component of this was shifting relevant column names, preferences, and variables from being "hard-coded" in the solution's script to being user-specified inputs. Re-use and transparency therefore become more robust, allowing the entire data solution to be packaged up and run in-country with no need for the user to open or edit the solution.

Next, creating descriptive statistics and visualizing the statistical relationship between variables will inform of any additional feature engineering needed.

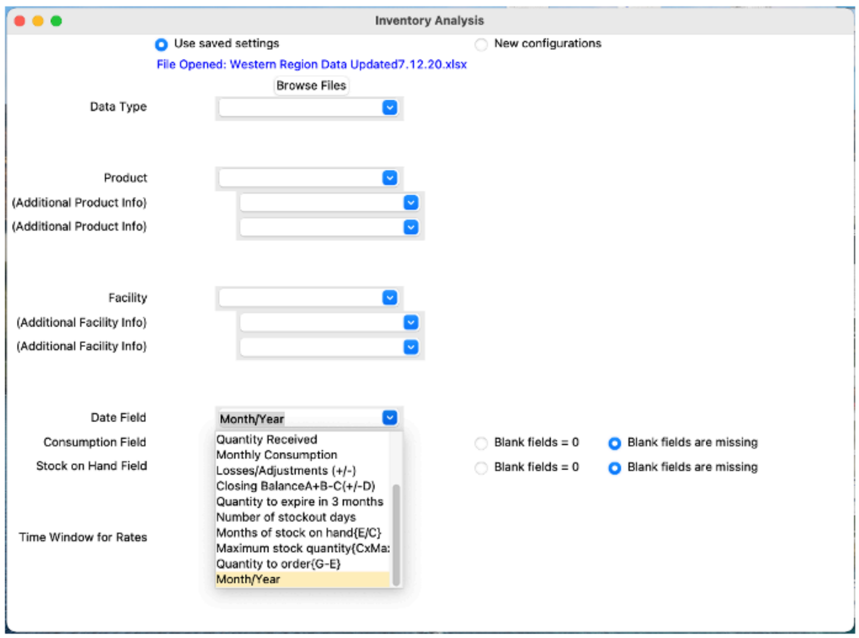

FIGURE 4. Inventory analysis. The user specifies key columns and information about their dataset in this input window. The user has loaded their data file (top) and has used the field names from the file to populate the key values. These values are assigned to the generalized variables within the solution, allowing it to run on any dataset with the relevant attributes.

MODELING

Having robust and standardized data inputs that are contained as their own module allows the model to perform robust and repeated analysis. This is known as the modeling stage or modeling module. The modeling module can employ a wide variety of analytics tools, the selection of which depends on the data (data understanding) and decisions being supported (business understanding). Tools can range from the calculation of set formulas and metrics to those such as optimization, sampling, and simulation models. The type of tool and packages used to develop the model are selected based on the type of model and what fits the business need (e.g., analysis performed by a Python script with results reported and shared in Excel).

Standard data inputs for the model enable stability in the analysis, and when building the model, it allows for more robust testing and for greater trust in the model's outputs. Specifically, it allows for easy de-bugging of issues because reliable norms about the state of the data are known since they are set in the data input module (for example it is known how blank records are handled). This ensures that, when building the analysis using any variety of analytics tools, the focus is on the design and functionality of the analysis itself, and not clouded by having to handle issues with input data.

Providing analysis to support decisions can have a cascading effect. For example, inventory turnover analysis identifies facilities at risk of stockout in the future. Once they are identified, the next question is which of them has the highest risk today? The value of this approach to analysis is that it is easy to build and expand on the original analysis, as it has already created a sound and repeatable analytic foundation that can be built upon. Thus, the development of additional analytic functions can be separated from the original analysis function, particularly when shared with other data scientists. For example, someone can develop an additional priority score function completely apart from someone else developing another analytics function. These additions will not affect each other's functioning or the original analytic functions. The key to this is that the sound foundation is already built, allowing for rapid expansions of functionality by the original team and by others after it has been shared.

In the example of Ghana and Nepal, the initial business need was to have an independent tool (not within the existing information systems) that was able to ingest data from an LMIS system to be run monthly or

Electronic LMIS Stock Analysis

Select Start Date	5/1/19
Month	May
Year	2019
Select Product	MICROLUT
Total Number of Sites Stocking	417
Number of sites included	231

Refresh Visuals

Total Number of Facilities

CONSUMPTION CoV		INVENTORY TURN RATE: LOW	MEDIUM	HIGH	VERY HIGH
		Monitor consumption and expiries closely	No action. Distribution as planned consumption and expiries closely	Need to increase flow of commodity to facility	Consistent pattern of stockout (nNot shown in chart) consumption and expiries closely
NO CONSUMPTION IN THE LAST 12 MONTHS	High risk of expiries if facility has stock (not shown in chart)				
HIGH	Supervision recommended. Target reasons for variability in consumption	High Risk of Overstock = 77	20	High Risk of Stockout = 17	Multiple Stockouts = 6
MEDIUM	Monitor variability in consumption	High Risk of Overstock=40	20	High Risk of Stockout = 27	Multiple Stockouts = 4
LOW		High Risk of Overstock=3	Target Zone = 5	High Risk of Stockout = 11	Multiple Stockouts = 1

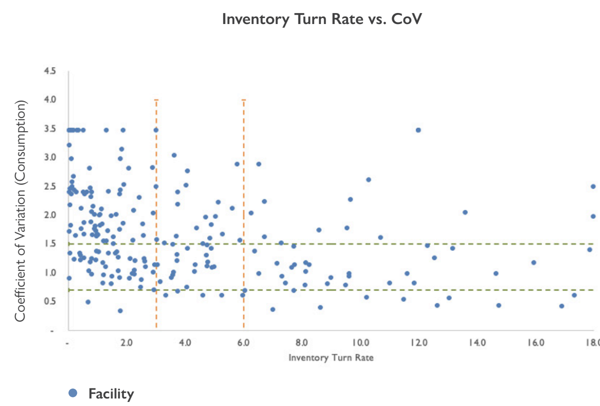

FIGURE 5. Electronic LMIS Stock Analysis

as needed. Additionally, the request was for the output to be in a familiar offline tool (e.g., Excel) to make it easier for adoption by end users.

The model outputs the resulting data as a series of tables, which can be fed into various reporting tools such as Excel or PowerBI. For this component, the analyst re-purposed the existing Excel dashboard used for the Ghana and Nepal activities; however, the output tables can be easily connected to another tool where a user could create their own reporting visuals. This again is a key aspect of the approach—that the outputs are agnostic to where and how they are presented to users, thus not tying them to the original output tool.

EVALUATION

In the CRISP-DM approach, the evaluation has two components that work together. First, it assesses the outputs of the model to ensure that it will support the decision-making needs. Secondly, it may provide additional insights into the business understanding. Both can happen simultaneously. The path for release of the tool means ensuring an understanding of how the insights provided to the users support decisions. The deployment stage will include refining the original plan to ensure the outputs fit in with the existing business processes. The feedback in the business understanding is about validation of assumptions when first conducting the analysis, as the outcomes of the analysis might shed new light on the overall understanding of the business processes. These aspects happen concurrently because they must ensure analysis is used today by decision makers and that additional insights continue to feed an improved understanding of those insights being used to support decisions.

Ghana and Nepal provide examples of the feedback used to improve the business understanding. In both countries, when analyzing the stability of the consumption pattern for certain products at certain facilities, it is not as stable as assumed (Note: this pattern is now being seen in multiple countries and programs). The realization is that average monthly consumption is not as reliable as originally thought in representing future consumption. Thus, the outcome is that for these facilities with unstable consumption, using average monthly consumption and standard inventory order rules to determine order amount and safety stock will not be sufficient in preventing both future stockouts or overstocking. Instead, there is a need to look at strategies for managing site/product combinations with unstable consumption, such as maintaining higher safety stock levels or increasing delivery frequency. This insight was not part of the original analysis; however, it provides decision makers with greater information to help determine actions at sites that are at higher risk of stocking out and with unstable consumption, as compared to sites with higher risk of stocking out, but with stable consumption.

Additionally, this work needs to be linked with the overall M&E strategy for strengthening supply chain performance. As described in the business understanding, it will also need to be combined with standard indicators and setting realistic yet meaningful targets or benchmarks for clearly defined periods, which will help assess the extent to which objectives are being met. Progress toward targets is best reviewed at intervals that align with when the business process is being measured and when data is available, such as the frequency at which orders are delivered to facilities.

DEPLOYMENT AND INTEGRATION WITH EXISTING DECISION-MAKING PROCESSES

The business-understanding and data-understanding steps will identify how the analysis is to be incorporated into the existing systems and business process. Thus, when designing analysis to be repeatable and agnostic to data sources (and maintaining robustness of the modeling module to function even when there are changes in the data source), inputs and outputs need to be designed with flexibility to adapt to changes. This allows for the analysis to remain useful as business processes evolve and as it is used in different contexts. This strikes a balance between meeting the business needs of today and of a future environment.

This is particularly important when connected with existing information systems. Thus, as highlighted earlier, there is need for a strong connection between data science and DevOps teams. For example, if the analysis is providing feedback into the system, then a strong connection with DevOps is needed throughout the process to ensure the analysis can scale into the existing system. If the analysis relies on the existing system as a primary data source, engagement with DevOps is still required; however, the engagement is different, as the outputs from the analysis are used either in parallel or separate from the existing system. The advantage of this approach is two-fold. First, it allows for rapid iteration and development of new analytics that were not part of the original system design, generating proof of concept that if shown to be valuable, can be integrated into the existing system if desired. Secondly, it allows for linking and analyzing multiple data sources without interruption of the existing system(s). For example, the use and analysis of various demographic data with logistics data to better target actions. If analysis were done within the logistics system, it would require adjustment to the LMIS to allow this and support multiple integrations. When done in a separate analytics tool, it functions without affecting the logistics data systems and makes it transferrable to other countries and systems that do not have the same logistics information system. The key remains that communication between data science and DevOps is crucial to the sustainability and continued use of data within information systems.
In the Ghana and Nepal example, the use, or deployment of the tool was happening in parallel to the existing information systems for the following reasons:

- The desired analysis was not directly possible in the existing system. This approach provided a way of demonstrating and “tuning” the analysis without affecting the information system or its current development.
- The use case was targeted to specific users that needed the data presented differently to facilitate decision making.
- It enabled the analysis to be “stand-alone” as the timing of when the analysis was needed was not consistent and would be used by stakeholders that were not necessarily regular users of the existing system.

In this approach, as described earlier, the outputs from the model are designed not to be tied to one way of being presented or used. Thus, it can be flexibly deployed and integrated with a stand-alone dashboard, or into a suite of other dashboards in Power-BI, Tableau or DHIS2. It is recognized that the same data is not just used to support a single decision. Thus, the output from the analysis is designed with the possibility of use by other decision makers and other systems. This is critical when looking to share with other users, as they will have different processes and systems they want to integrate with the analysis.

SHARING THE APPROACH

A key attribute of a data science approach is sharing of the code for use and expanding upon it over time by numerous contributors. The rapid growth and breadth of data science tools is attributed to this sharing of code, enabling its growth, and sustaining a broad variety of tools across multiple industries. When a solution is fully functional and ready to share, it can then be posted on an internet file-hosting platform such as GitHub. Platforms like GitHub allow software and files to be shared within an organization, or publicly so that it is provided as open-source software with a release license. The tool is posted as a repository where a new user can download it, use it, and can also create a “clone” of it to have their own version of the solution. The user can modify or update their branch of the solution, without any risk of changing the original solution posted by the original author. They can then, in turn, share their updates via the platform as well. If a community of users decides to work together on the code, they can then collaborate and merge their modifications or updates together to make a new solution. Depending on the license agreement associated with the release of the code, it can, but may not necessarily include the original author’s code.

Hosting platforms like GitHub enable easy sharing of solutions, but the structure of the solution itself is key for facilitating new users’ ability to understand and build upon it. Refactoring the analytical process into different segments makes for easier tracing of logic flows and enables a developer to “plug in” a new module to the process.

A key attribute of a data science approach is sharing of the code for use and expanding upon it over time by numerous contributors.

CONCLUSIONS AND RECOMMENDATIONS

The data science approach aims to enhance M&E’s ability to respond effectively, enabling business processes to improve continuously. As described in the beginning of this technical brief, with limited resources available for M&E, tools and approaches like data science reduce labor-intensive manual calculation processes and the use of static reports. Building more automated and robust tools to prepare and present data to users enables greater value-add from M&E by strengthening the use of data within a system. Additionally, designing the analysis to be able to run immediately means the data is available for presenting outputs in a timely manner to decision makers, facilitating action. It allows for timely actions and a stronger connection between the analysis and use. When M&E is factored into everyday decisions, the potential of M&E becomes more often realized and valued across the supply chain.

Designing a data science approach and using code to implement analysis requires two key components: a design driven by the decisions being supported, and a strong connection with DevOps processes. The CRISP-DM methodology assists in ensuring that the decision and user remain at the center of the design, and that analysis builds on the existing information system to increase data use. The sustainability and scalability of the analytics tool depends on both. Analyses designed with this approach help realize the full potential held within information systems. Therefore, closer tying of these streams is recommended.

Finally, there are two parts to continue to leverage the growth and use of data science within the global health supply chain context. First, mechanisms for the sharing of data science code on platforms such as GitHub need to be established. GHSC-PSM will release the code used within this activity on GitHub to encourage transparency, and to promote the use of data science to support decision making. This, in turn, can enable others to build upon the original analytical work to then create new and exciting analytics tools for everyone to use. Second, beyond just documenting the code, the ability to share the code depends on how the analysis is designed. In many cases, analytic tools are designed to be tailored solely to very specific use cases and data sources that meet the direct needs of the initial stakeholders. Therefore, to facilitate the reuse and expansion of the tools beyond the original design, there is a need to refactor the code to provide this foundation when sharing code for these tools. In many ways, code is like a living entity. It needs a foundation that allows it to grow beyond its initial, specific context; otherwise, its potential will remain unrealized.