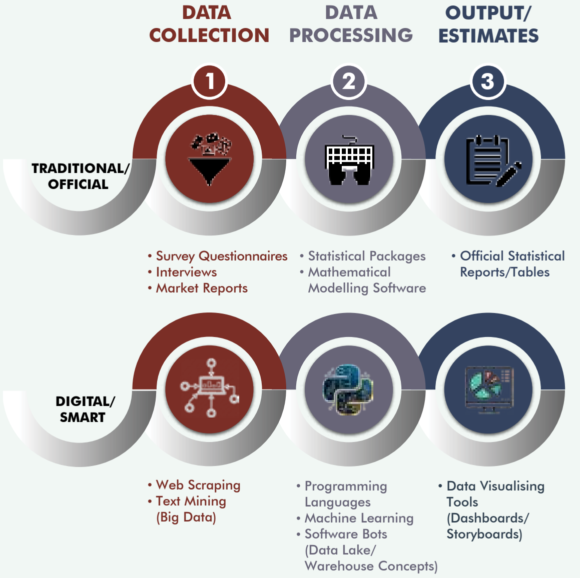
BOX 8
Breaking Stereotypes: How Digitalisation will Impact the Traditional Process of Statistics

The traditional role of statisticians has evolved over time gradually without major changes. The process of statistics generally includes stages of data collection, data processing and dissemination of final estimates/outcome (Figure B 8.1). Until recently, the most common advancements in the field of statistics focused mainly on the use of statistical software in data processing, while stages of data collection and dissemination of final results remained static. Rapid developments in information technology, however, challenged the traditional role of statisticians as certain statistical functions are becoming increasingly complex.

Figure B 8.1
Traditional vs. Smart Statistics



Meanwhile, data collection methods continued to evolve over time without being a limiting factor in the process of statistics. Traditional methods such as survey through questionnaires, one-to-one or phone interviews and secondary data from published sources are still being used as official sources for statistics produced by any organisation due to their reliability and customisability. However, over time, these traditional data collection methods experienced constraints not because of intrinsic features of such methods, but due to the widening horizons of digitalisation coupled with high frequency data generated through various means and platforms. These high frequency data get accumulated in various formats around cyber space, but, until recently, alternative uses of such data have not been adequately explored.

Data processing plays a vital role in obtaining a refined set of estimates, which is used for decision-making. Data collected using traditional methods and processed manually are currently being analysed using various statistical packages. Mathematical modelling software is commonly being used to build up statistical relationships among variables. 'Official Statistics', which are national level statistical outcomes, comprise key economic indicators and a series of estimates related to economic activities. Around the world, this process of compiling official statistics is followed under various limitations and assumptions in assessing the status of an economy. Despite its limitations, the traditional process of compiling statistics has gained validation and reliability over time as it is based on scientific methodology.

The term 'Big Data' came to light in this backdrop when the traditional statistical process was challenged by a wave of digitalisation, coupled with the rise of information and communication technologies. In simple terms, digitalisation of the economy refers to the use of digital technology for various economic activities and Big Data refers to the accumulation of high frequency data from non-traditional sources for the use of economic agents. Despite that these concepts originated in early 2000's, most of the terminology used in relation to Big Data lack universally accepted definitions[1]. Nevertheless, its usage seems to have encompassed over a wider spectrum as different online platforms on the World Wide Web generate a large part of Big Data while cashier terminals of supermarkets, automated vending machines, surveillance camera units, etc., also gather massive volumes of data on a daily basis. This data is originally recorded for a particular purpose and Big Data becomes a by-product of that process. Because of its recording frequency, such data becomes an incredible source, which outperforms most of the traditional data sources used in the statistical process in terms of volume and speed.

How digitalisation has influenced official statistics

With the emergence of Big Data, one of the concerns raised by users was its eligibility to replace official statistics. As most of the data used in the statistical processes are primarily collected for a designated purpose, relying on secondary data through unorthodox means is still a challenging move to be taken by statisticians. In the present context, statistics generated using these unconventional data is called 'Smart Statistics' and are used as an alternative stream to complement official statistics. One of the benefits of maintaining a supportive stream of statistics is that it can help to validate official statistics, and also over time smart statistics will contribute towards continuously improving the quality of the entire statistical process

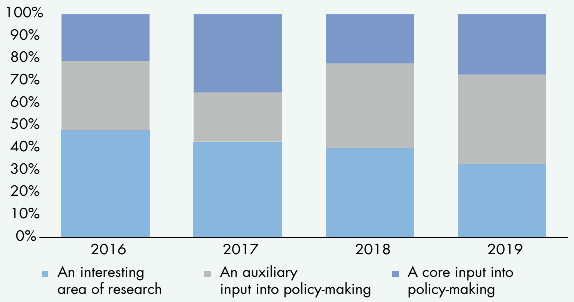
[1] History of Big Data - Available at: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

within an organisation. For instance, a survey on Big Data in central banks[2] carried out in 2019 shows that central banks around the world have identified Big Data as an auxiliary input into policy-making processes (Figure B 8.2).

The biggest impact of digitalisation on the statistical universe is its capability to transform a highly fractionalised and disconnected organisation in terms of data sharing into a centralised data warehouse, which promotes a brand new 'data culture' among all the members. Digitalisation demands everyone who uses the data to know the entire process of statistics, which extends from data collection to final estimates. In the case of central banks, the analytical departments will need to equip their staff with necessary programming skills to exploit the Big Data available in their 'Data Lake'[3]. Anyone who has access to the Data Lake, irrespective of their area of expertise, will need the skill-set of a data analyst to get the maximum benefit from digitalisation.

Figure B 8.2

Big Data in central banks - Central Banking Survey (2019)



Source: Central Banking (2019)

What digitalisation has in offer for statisticians

Digitalisation contributes to the field of statistics at each stage of its traditional process, from data collection to the final outcome. The traditional data collection methods would be replaced/supported by data digitally collected using methods such as Web Scraping and Text Mining. Web Scraping, in simple terms, refers to the process of extracting specific data available on relevant web sites by using programming software. Central banks around the world have started projects on scraping price and other data recorded on various web sites to support their monetary policy-setting process, which include forecasting inflation using price data, labour force dynamics of the economy through various job market web sites, and identifying the unorganised sector contribution to the national output through web sites related to Barter/Sharing economy.

Text Mining is another digital technique, which is commonly used to evaluate the momentum of a community before or after a certain policy change takes place. This technique can identify a pool of key words related to people's sentiments and track such words on various social media platforms to identify the frequency of their use. This will provide an estimate on the direction of public expectations of a policy or any other targeted phenomenon.

Ongoing projects of central banks, which use web scraping and text mining are focused on different areas. The Sveriges Riksbank (Central Bank of Sweden) uses data scraped from grocery retailers' websites as a tool of error correction in inflation forecasting. The Central Bank of Armenia collects house prices displayed online by real estate agents to compile a house price index. The Deutsche Bundesbank (Central Bank of Germany) uses web data to capture depositors' sentiments by means of text mining the queries related to the term "deposit insurance", which provides them with an understanding about the depositors' confidence in the banking sector of Germany. The European Central Bank, the Bank of England (Central Bank of the United Kingdom), the Banque de France (Central Bank of France) and other regulatory institutions around the world have started various projects in the area of Big Data and digitalisation, as the use of Big Data collected through web scraping and text mining is often advantageous over the conventional survey data because of the high velocity and high volume.

The contribution of digitalisation to the data processing stage is the most sophisticated and complicated element of the entire process of smart statistics. One of the differences between the traditional job of statisticians and the digital process is that both data collection and processing run on software using programming languages thereby leading to limited manual intervention in the digital process. In an advanced setting, the human element would be further reduced with the use of 'Machine Learning' and 'Artificial Intelligence' that will capture algorithms in data and carryout analytical tasks autonomously.

Data visualisation is the final stage of the statistical process, which converts the statistical estimates to graphical and other means of presentation. Official statistics, which are traditionally presented through tables, reports and graphs will be generated using tools such as visualising dashboards, which graphically and illustratively present the key statistical estimates. These dashboards are interactive and could be tailored to the needs of individual users. Central banks in their transformation towards being digital are exploring the potential of using these visualising dashboards for managerial decision-making in the policy-setting process.

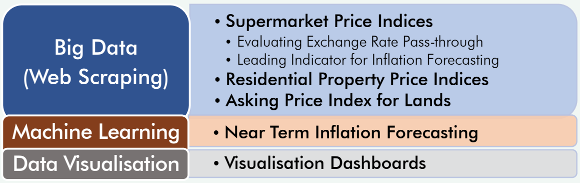
[2] Big data in central banks (2019 Survey results). Available at: https://www.centralbanking.com/central-banks/economics/data/4508326/big-data-in-central-banks-2019-survey-results

[3] Data Lake – A system or a repository of data stored in its natural/raw format. Available at: https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/

Digital footprint: Smart statistics in the Central Bank of Sri Lanka

The Central Bank of Sri Lanka (CBSL) also embarked on its digitalisation journey in the field of statistics with a few pilot projects (Figure B 8.3). The high exchange rate volatility observed during the latter part of 2018, underscored the need to assess the degree of exchange rate pass-through to the consumer items. In this regard, CBSL took initiatives to adopt web scraping to collect prices of imported food and beverage items posted online on a daily basis. With the success of this project, CBSL increased the number of food and beverage items, of which the prices are collected by web scraping daily, as a pilot project to develop a leading indicator to strengthen the inflation forecasting process.

Figure B 8.3
Smart Statistics Projects


Big Data (Web Scraping)
• Supermarket Price Indices
• Evaluating Exchange Rate Pass-through
• Leading Indicator for Inflation Forecasting
• Residential Property Price Indices
• Asking Price Index for Lands
Machine Learning
• Near Term Inflation Forecasting
Data Visualisation
• Visualisation Dashboards

Since early 2019, in order to overcome the bottlenecks in data collection for compilation of the Land Price Index, which was based on valuations obtained from the Government Valuation Department semi-annually for residential, commercial and industrial lands in Colombo District, CBSL launched a pilot project to collect the market prices of properties advertised online, through web scraping on a monthly basis. Accordingly, CBSL is developing an Asking Price Index for lands in Colombo District and intends to expand it to the entire country. Moreover, a House Price Index for residential housing properties covering the entire country and a Condominium Price Index covering properties in the Colombo District are being compiled using scraped data. Additionally, a House Rental Price Index is also expected to be constructed using web scraped data to understand the status of the rental market and its impact on household expenditure dynamics.

CBSL has initiated another pilot project for forecasting near-term inflation using Machine Learning by employing a multi-step time series with Long Short-Term Memory (LSTM). The bank further uses visualisation dashboards to present the national accounting estimates, which are annually compiled for the provincial level. The use of visualisation dashboards is becoming popular within CBSL as they complement the official statistical reports used for decision-making.

Challenges for central banks and Way forward

With the rise of digitalisation across economies, central banks are concerned with the use of data analytics and other technologies for the policy-setting process. One of the key challenges in digitalising the statistical processes is to develop the skill-set within organisations for economists, statisticians, mathematicians to use the specific software/ programming languages and the top management to be 'tech-savvy'. Digitalisation of statistics will demand individuals with a strong mathematical background to solve complex issues and open up opportunities for data scientists to be a part of statistical departments.

Digitalisation will require changes to the existing regulatory and infrastructure frameworks to facilitate adoption of new technologies. Access rights to privately owned web spaces in terms of legality and possible blockage imposed by the administrators of such web sites are also key concerns. Access rights to some domains are already given at a premium and the cost effectiveness of paying for such data is another concern for statisticians as traditional data collection methods could be relatively inexpensive in certain instances.

Another commonly discussed issue is the reliability of data extracted from secondary sources. Irrespective of the high volumes and velocity, regulators are reluctant to base policy decisions entirely on digital sources. It is identified that statisticians in public agencies have an agitation towards using certain web-based technologies mostly because of disclosure risks and cyber security threats. Most of these technologies demand 'Cloud-based' data warehousing, which may enhance exposure to data security risk.

Developing economies such as Sri Lanka will have to focus on establishing digital infrastructure platforms and regularising the use of digital technology in economic activities by introducing governance frameworks, which is a priority area even for CBSL at present. Considering these limitations and challenges faced by the central banks worldwide, digitalisation of statistical processes is still at its early stages of the life cycle, and central banks carry the view of using it as an alternative stream rather than fully migrating to smart statistics. However, the transformation is already happening at a faster rate. Therefore, conventional statistical processes will need to be geared to embrace the new challenges ahead, to ride the tide together with the rest of the world.

References

1. Hinge, D. and Karolina, S. (2019) 'Big Data in Central Banks: 2019 Survey Results', *Central Banking*.
2. Reimsbach-Kounatze, C. (2015) 'The Proliferation of "Big Data" and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis', *OECD Digital Economy Papers,* Paris: OECD Publishing.
3. Tissot, B. (2015) 'Big Data and Central Banking', *IFC Bulletin 44, March 2015.*
4. Weydert, N. (2019) 'Preparing the Future: The Impact of Digitalisation on Official Statistics', *SLS/STATEC joint event, 17 May 2019.*
5. Wuermeling, J. (2019) 'The Deutsche Bundesbank's digital transformation', *Central Banking.*