September 2017 SDN/17/06

IMF STAFF DISCUSSION NOTE

Big Data: Potential, Challenges, and Statistical Implications

Cornelia L. Hammer, Diane C. Kostroch,

Gabriel Quirós, and STA Internal Group

September 2017 SDN/17/06

IMF STAFF DISCUSSION NOTE

Big Data: Potential, Challenges, and Statistical Implications

Cornelia L. Hammer, Diane C. Kostroch,

Gabriel Quirós, and STA Internal Group

Statistics Department

Big Data: Potential, Challenges, and Statistical Implications

Prepared by Cornelia L. Hammer, Diane C. Kostroch, Gabriel Quirós, and STA Internal Group[1,2]

Authorized for distribution by Louis Marc Ducharme

JEL Classification Numbers:	E0, Y2, Z0
Keywords:	Big Data, Macroeconomic and Financial Statistics, Official Statistics, Data Quality, Surveillance
Authors' E-mail Addresses:	CHammer@imf.org, DKostroch@imf.org, GQuiros@imf.org

[1]The Internal Group on Big Data of the Statistics Department (STA) was established in August 2016 to investigate opportunities and challenges of big data for macroeconomic and financial statistics. The Group is led by Gabriel Quirós, STA Deputy Director, and composed of Serkan Arslanalp, Jose Maria Cartas, Aimee Cheung Kai Suet, Daniela Comini, Saurabh Gupta, Andreas Hake, Cornelia L. Hammer, Gary Steven Jones, Venkateswarlu Josyula, Diane C. Kostroch, Stephanie Medina Cas, Edgardo Ruggiero, and Patrizia Tumbarello.

[2] The authors would like to thank, in particular, for valuable comments received on previous drafts Louis Marc Ducharme, El Bachir Boukherouaa, and ITD colleagues for their excellent IT-related contribution; Claudia Dziobek and Mark Van Wersch for review and editing; and Manuela Goretti for her valuable comments. We would like to thank IMF Multimedia Services for graphics and our COM colleagues Jim Beardow, Linda Long, and Lucy Scott Morales for their editorial support. The paper has also greatly benefited from the review and comments received from IMF area departments (AFR, APD, EUR, MCD, WHD) and functional departments (COM, FAD, ICD, ITD, LEG, MCM, RES, SPR), and from the analytical input and administrative support provided by James Chan, Liliya Nigmatullina, and Zula Oimandakh (all STA).

CONTENTS

EXECUTIVE SUMMARY

This Staff Discussion Note reflects on the potential, challenges, and implications of big data for macroeconomic and financial statistics. It addresses the wide range of stakeholders of "official" data and statistics and covers interested users and producers. Good data and statistics, strategic elements for any society and economy, are essential for sound policy decision making in both the private and public sector. By now, many private companies as well as national and international organizations see that "big data" is no mere buzzword, but a medium-term concept that requires a long-term vision.

Big data is evolutionary and can provide innovative, real-time, and more granular insight for economic and financial analysis. Yet **big data opportunities for individual countries will be asymmetric and will depend on the country's characteristics and the availability of the systems and networks generating big data.** Big data offers opportunities, challenges, and implications for official statistics that compilers and users of statistics need to be aware of when they start to incorporate big data into their work plan to the extent relevant.

Numerous individual applications of big data are already being carried out, either by users or compilers of data and statistics. However, a systematic and structured discussion is lacking: this SDN attempts to offer such, emphasizing implications for macroeconomic and financial statistics. Further research and detailed analyses are essential to understanding if and how big data can directly and indirectly support IMF surveillance work.

What does big data mean? Although there is no agreed-on definition, the term is often characterized by the 3Vs—high-volume, high-velocity, and high-variety. More Vs have been added over time, such as veracity and volatility. Unlike statistical data compiled for specific purposes, big data is made up of byproducts found in business and administrative systems, social networks, and the internet of things. To structure the discussion, the paper presents a big data classification that is relevant for macroeconomic and financial statistics.

What is the potential of big data? Big data can benefit macroeconomic and financial statistics and ultimately policymaking through at least three features:

1. By answering new questions and producing new indicators
2. By bridging time lags in the availability of official statistics and supporting the timelier forecasting of existing indicators
3. As an innovative data source in the production of official statistics

What challenges come with big data? Data quality concerns, difficulties with access, and new required skills and technologies are the main challenges of big data. And while big data mainly measure insights, correlations, trends, and sentiments, detailed country-by-country time series in

accordance with internationally agreed standards remain crucial for measuring and monitoring countries' economic performance and policies over time.

Statistical Implications: Moving forward, international statistical cooperation is key to overcoming big data challenges and to building lasting partnerships among national and international statistical agencies, users, and data owners. The implications include the need for interested parties to build up the required skills and technologies in their organizations. Opportunities, challenges, and potential implications are particularly high for national statistical agencies: **the incorporation of big data as new data sources, either supplementing or substituting for traditional data sources, will not be exempt from methodological, organizational, and budgetary challenges.**

The success of big data projects lies not in implementing a particular piece of technology, but rather in establishing an environment of people and processes that take big data innovations forward and put them to work. **Given the diverse skills needed to deal with big data, it also provides an opportunity for organizations to break their internal silos, including between users and producers of data and statistics.**

From individual applications of big data to its incorporation into the systematic, regular, and large-scale production of statistics—before engaging in costly and time-consuming investments, organizations should begin with a proof of concept and should operationalize the project only after findings have proved valuable and feasible from an organizational point of view. Statistical agencies should decide on a case-by-case basis and select the most promising big data projects to complement existing statistics. Moreover, to keep abreast of developments, agencies should proactively search for big data sources to address the most urgent research needs. Selected big data projects may also be incorporated into capacity development activities to support the membership in building their capacity to benefit from available big data sources. Going forward, research in and compilation of best practices—for statistical techniques and methodologies that address veracity and volatility, specifically—need to be at the top of the statistical community's agenda.

Given that big data is not static but dynamic, the systems and networks generating big data continue to evolve, and with them the possibilities, challenges, and limitations of big data for statistics. Consequently, **the overall assessment made in this paper will need to be revisited as the worlds of big data and official statistics evolve.**

I. INTRODUCTION

1. Big data is no mere buzzword—big data is here to stay. Organizations that bypassed the initial hype now see the need to make decisions as to whether to intertwine big data with their future organizational culture. Others are experiencing pressure from the first movers. Organizations can implement big data and benefit from data science by learning from best practices and adapting those that have the deepest potential for meaningful insight. They can also proactively and innovatively search for big data sources that could help answer the most urgent research needs. IMF teams are already successfully implementing this practice.

2. Many companies in the private sector have implemented innovative tools to compete on analytics (Davenport 2006). Sophisticated information systems, rigorous analysis, and masterful exploitation of data are among the last remaining points of differentiation. Companies are mining the accumulated data and turning them into marketing analytics: *dynamic content* adapts ads, websites, or e-mail bodies based on the interests or past behavior of the viewer; *marketing segmentation* optimizes companies' promotions based on past purchases; and *cross-channel personalization* supports companies to retarget customers by planting ads in their social media feeds. Marketers (Chain Store Age 2015) believe that individualizing customers' e-shopping experience has led to higher engagement rates, consumer loyalty, and better brand recognition. Big data applications have leveraged interest in many organizations, from customer-facing companies in retail and e-commerce to banking, health care, and manufacturing. Simultaneously, the decreasing cost of storage and computing, as well as the flexibility, scale, reliability, and user-friendliness of the cloud, has changed the competitive landscape—basically democratizing this space. Data analysis is becoming a competitive advantage for different businesses of all sizes.

3. Public sector institutions are also interested in using big data and new technologies to deliver their mandate more effectively and efficiently. In national and international agencies, increasingly nontraditional methods are being used to improve living conditions. Big data can help predict food shortages through a combination of data on drought, weather conditions, migration, prices, and previous production levels (Data Floq). The crossing of real-time GPS data with information on car accidents and traffic jams can enhance the flow of public transportation and the velocity of cars in congested cities around the world (Data Revolution Group 2014). Long before big data was a term, satellite imagery had been used in weather forecasting and the enabling of geographic positioning data. Nowadays, foresters, urban planners, and agricultural managers—but also federal and development agencies—use satellite imagery or "big geospatial data" to combine and interpret various types of data to maximize the effectiveness of physical assets in the field. To use big data, government and international agencies will need to set up partnerships with the firms and other entities generating the data of interest.

4. The potential that the data revolution has in helping to address development challenges is widely recognized. The widespread use of mobile phone technology, internet

traffic, and social networking allows individuals in developing economies to gain access to banking services, employment information, medical services, and markets. Simultaneously, the large volume of data generated as a byproduct of those innovations presents new opportunities to gain wider and deeper insight into human behavior and social patterns. Such insight can complement and expand indicators that are being "traditionally" collected (UNGP 2012), including social development indicators. Through the *Innovations in Big Data Analytics* program, the World Bank (World Bank 2016) has set up initiatives to support developing economies to operationalize capabilities in big data. The World Bank sees great potential for big data to add value in the developing world—however, it acknowledges that capacity development and the willingness of the official sector and the private sector to come together will be key (ODI 2015). The voluntary data sharing of the French telecom company Orange with researchers for the Data for Development (D4D) Challenge (Challenge 4 Development 2013) in Africa is a good example of how to foster the necessary partnership between the private and public sector.

5. **Central banks are showing strong interest in using big data and new related technologies (BIS 2015).** Central banks see merits in the use of big data as a potentially effective forecasting tool to support macroeconomic and financial stability analyses. Monetary policy analysis (Central Banking 2016) can benefit from better and timelier nowcasts of macroeconomic variables; macro- and microprudential policies might benefit as well. Some pilot studies are being conducted to explore the conditions for making a more regular use of big data as part of a data toolkit.

6. **Incorporating big data into macroeconomic and financial statistics can support decision making.** There is potential for big data to produce new indicators, bridge time lags, and support the forecasting of existing data sets, as well as serve as a new innovative data source in the production of official statistics. To set the foundation for long-term success, a holistic understanding of the opportunities and challenges that come with big data is essential.

7. **This paper aims to contribute to the discussion of the potential and challenges of** big data and its applications for macroeconomic and financial statistics fit for analytical and policy use.

II. WHAT DOES BIG DATA MEAN?

8. **The term big data is not entirely new.** In a 2006 issue of the *Harvard Business Review*, Tom Davenport notes one method organizations such as Amazon, Capital One, and the Boston Red Sox use to dominate in their fields: analytics as a competitive differentiator—"businesses were awash in data and data crunchers" (Davenport 2006). In 2010, Hal Varian discussed computer-mediated transactions, whereby economic transactions involve a computer such as a point of sale terminal, a cash register, and more recently electronic commerce. Although the authors do not explicitly use the term "big data," the phenomenon and information they refer to will subsequently fall into the big data discussion.

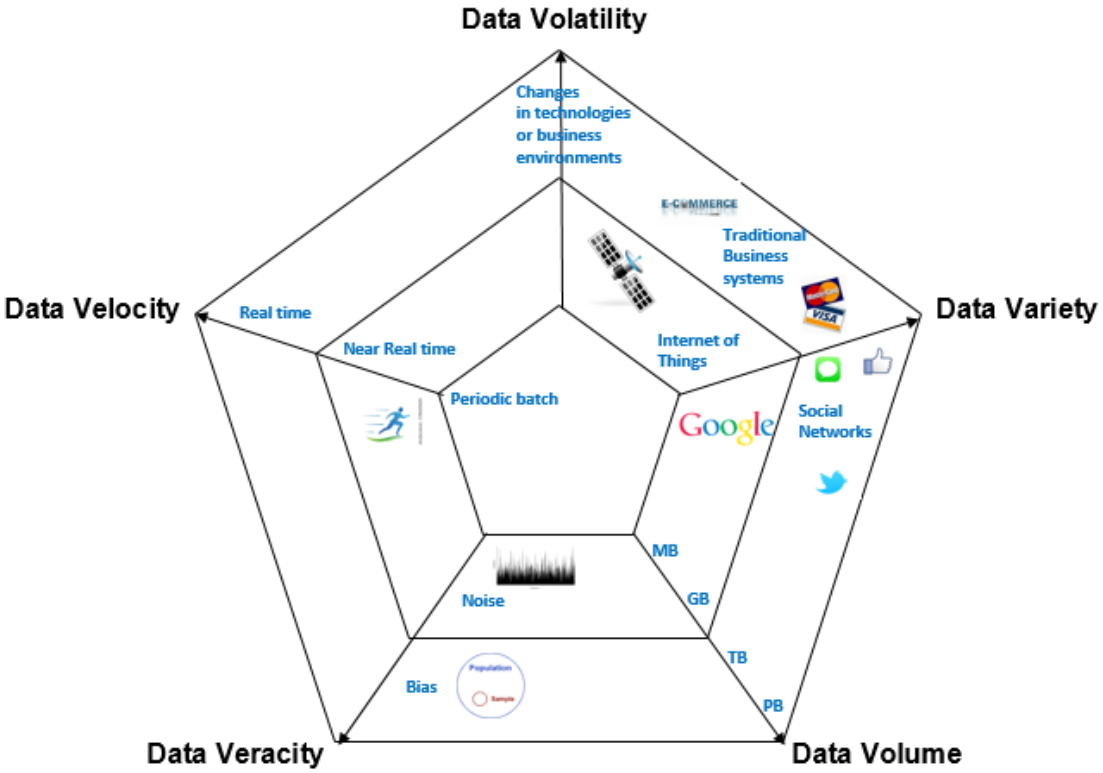
9. Although no agreed-on definition of big data currently exists, the term is often characterized by the 3Vs—high-volume, high-velocity, and high-variety.[3] High-volume refers to increasing exabyte data generated by machines, networks, and human interaction; high-velocity refers to the speed at which data are created, processed, and stored; and high-variety relates to the range and complexity of data types and sources. Data sets are so large and complex that traditional data-processing applications become insufficient to capture, store, and analyze the data. Instead, a network of human skills, advanced technologies, and data access infrastructure are essential to handle big data. This is a key challenge for statisticians and policymaking organizations seeking to incorporate big data in their toolkits.

10. Unlike statistical ("made") data that are compiled for specific purposes, big data is a byproduct "found" in business and administrative systems, social networks, and the internet of things. Social networks are online platforms that help people build social relations with others having similar interests (Facebook, Twitter, LinkedIn). Users create blogs and profiles, share pictures, and exchange messages and thereby provide human-sourced information that is digitalized and stored. Data in social networks are often ungoverned and unstructured. In its big data classification, the United Nations Economic Commission for Europe (UNECE) (see Appendix X) also includes in social networks internet searches and mobile data that can be more widely understood as human-sourced information. Traditional business systems are processes and procedures defined by businesses to provide value to their customers and generate process-mediated data, including administrative records. Business systems record well-governed, structured information on transactions, positions, and metadata related to business events (commercial transactions such as registering a customer or receiving an order) stored in relational database systems. The internet of things is a system of data-producing interrelated computing devices with embedded sensors and internet connectivity that measure and record events and situations in the physical world. Their output is structured machine-generated data (sensor records, computer logs, webcam, mobile phone location/GPS).

11. The list of Vs has grown over time, emphasizing both opportunities and challenges that companies and organizations face when incorporating big data into their existing business operations (Figure 1). Veracity refers to the noise and bias in the data as one of the biggest challenges to bringing value and validity to big data. Volatility refers to changing technology or business environments in which big data are produced, which could lead to invalid analyses and results, as well as to fragility in big data as a data source.

[3] Gartner analyst Doug Laney came up with the famous three Vs back in 2001.

Figure 1. The 5Vs of Big Data—Volatility, Variety, Velocity, Veracity, and Volume



Based on Doug Laney, 2001

12. The big data classification increasingly relevant for macroeconomic and financial statistics is presented in Box 1. While the categories of networks, systems, and machines generating big data as a byproduct are based on the UNECE big data classification (see Appendix I), this paper includes additional subcategories (in *italics*), such as administrative data, business websites, and online news. Based on outcomes of use cases, these additional subcategories appear to also have the potential to feed into macroeconomic and financial statistics and finally surveillance. Yet categories such as videos, medical records, and pictures are excluded as they are not currently applicable to macroeconomic and financial statistics, although they may become so at a later stage.

Box 1. Adapted UNECE Big Data Classification

1. Social Networks (human-sourced information)
- 1100. Social Networks: Facebook, Twitter, *LinkedIn*
- 1200. Blogs and comments
- 1600. Internet searches *on search engines (Google)*
- 1700. Mobile data content: text messages, *Call Detail Record, Data Detail Record, Location update, Radio coverage updates*
- *Online news*

2. Traditional Business systems (process-mediated data)

21. Data produced by public agencies
- *Administrative data*

22. Data produced by businesses
- 2210. Commercial transactions
- 2220. Banking/stock records
- 2230. E-commerce
- 2240. Credit cards
- *Business websites*
- *Scanner data*

3. Internet of Things (machine-generated data)

31. Data from sensors
- 311. Fixed sensors
  - 3111. Home automation
  - 3112. Weather/pollution sensors
  - 3113. Traffic sensors/webcam
  - 3114. Scientific sensors
- 312. Mobile sensors (tracking)
  - 3121. Mobile phone location *(GPS)*
  - 3122. Cars
  - 3123. Satellite images

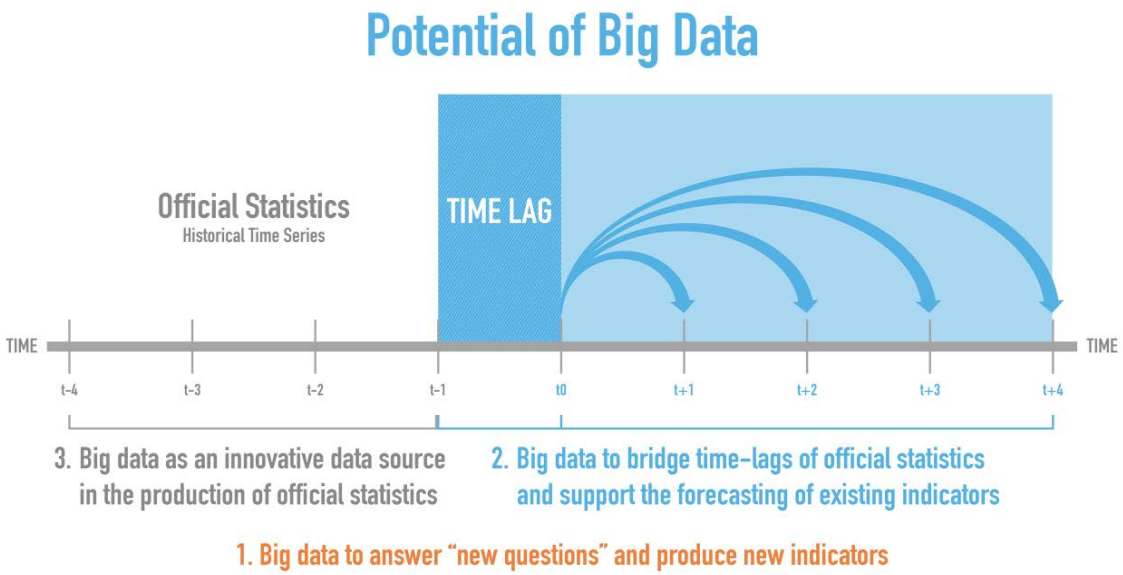
13. While the potential of big data may be large, the big data opportunities for individual countries will depend on the country's characteristics. The availability of social networks, traditional business, administrative systems, and the internet of things generating big data will vary. Consequently, to assess big data opportunities and their potential application for policymaking, one must take into consideration the availability of data and related tools, the users' capabilities, and privacy and security issues, as well as legal and technological systems.

III. POTENTIAL OF BIG DATA

14. The discussion in Section III is organized around three features through which big data can directly or indirectly benefit macroeconomic and financial statistics, and finally policymaking (Figure 2):

1. By answering new questions and producing new indicators
2. By bridging time lags in the availability of official statistics and supporting the timelier forecasting of existing indicators
3. By providing an innovative data source in the production of official statistics.

Figure 2. The Potential of Big Data



This approach allows for structuring the broad discussion about big data as well as distilling the most important elements for macroeconomic and financial analyses.

15. Although the three features may directly (1 and 2) or indirectly (3) provide valuable data for policy analysis, they are interlinked and as such cannot be completely separated. This paper provides examples but does not intend to give a comprehensive big data project inventory.[4] It rather seeks to illustrate how big data projects can potentially enhance policy analysis. However, moving forward, further exploration and proof of concepts are required to ensure their feasibility.

A. Big Data to Answer New Questions and Produce New Indicators

16. Big data breaks with the traditional method to search for causality. Working with big data implies seeking patterns and correlations that may not tell us *why* something is happening, but rather alert us *that* it is happening (Mayer-Schönberger and Cukier 2014). In this vein, new indicators can be developed to obtain real-time correlations and to establish a more comprehensive early-warning system (Kitchin 2015) that can monitor the buildup of country-specific as well as systemic risks in the real, external, fiscal, and financial sectors. Google web searches and Facebook posts are already used to predict stock market liquidity (Arouri and others 2014) and to construct sentiment indices that predict stock market activity (Karabulut 2013).

17. Although the effective and smart use of big data is evolving, the potential to transform the way we look at information is not controversial. Big data exploration is an exercise in asking new and better questions as well as an opportunity to challenge our

[4] A big data inventory can be found here: https://unstats.un.org/bigdata/inventory/.

conventional thinking about the collection and production of statistics. This view is reflected in international initiatives, such as the IMF's Big Data and Analytics Symposium or the European Big Data Hackathon, which challenge participants (professionals in the field, academics, and the private sector) to find innovative uses of big data to aid policy and decision making. Undoubtedly, more ideas will arise, some of which could broaden the range of traditional statistics and respond to research needs, after passing the proof of concept.

18. In some cases, big data can allow policy analysis to move beyond aggregates and look at what lies beneath to better inform policy responses. Mobile phones, call detail records, satellite imagery,[5] and specifically developed apps have the potential to predict socioeconomic patterns, population movements, credit-risk profiles, and climate-smart agriculture and to detect the buildup of crisis-related stress. More granular information from big data may help clarify policy interlinkages and thereby more quickly identify the effect of policy recommendations on firms and households, highlighting political tensions and/or inequality dynamics. Big data have the potential to help address development challenges and meet demands for compiling Sustainable Development Goals (SDG) indicators, such as gender equality.[6] The private company LinkedIn is already using its granular data to publish gender diversity statistics and provide training on gender statistics (Karani 2017).

19. Following the IMF Big Data and Analytics Symposium, the IMF in-house Big Data Innovation Challenge paved the way for innovative ways to leverage big data in future work of the IMF. The top six ideas were approved for proof of concept development: (1) Using SWIFT data to monitor global financial flows, (2) a sentiment-based early-warning system, (3) nowcasting GDP using Google trends data, (4) automating and expanding the Week @ the Beach Index, (5) pooling government cash flow data to enhance surveillance and policy analysis, and (6) applying analytics for better tax and customs administration. Other examples are the use of big administrative data for the IMF Fiscal Affairs Department Revenue Administration Gap Analysis Program to determine the value-added tax compliance gap and the IMF Research Department's use of big geophysical data to measure effects on low-income countries' economic activity. Such ideas may demonstrate sufficient robustness and accuracy to augment or supplement existing practices over time.

20. Big data may allow better measurement of the effects of financial inclusion, access to financial services, and economic growth. Electronic money systems, such as M-Pesa (Box 2), are growing rapidly in developing economies (Donovan 2012)[7] and with them, opportunities to measure the effects of financial inclusion on poverty reduction, gender

[5] Australia, China, Colombia, and Mexico have discovered the potential of satellite data for agricultural statistics.

[6] The IMF Statistics Department is involved in the following SDGs: (1) indicator 8.10.1.a, Number of commercial bank branches per 100,000 adults, and indicator 8.10.1.b, Number of ATMs per 100,000 adults (Tier I); (2) indicator 10.5.1, Financial soundness indicators (Tier III); (3) indicator 17.1.1, Total government revenue as a proportion of GDP, by source (Tier I); and (4) indicator 17.1.2, Proportion of domestic budget funded by domestic taxes (Tier I).

[7] More than 110 money mobile systems servicing more than 40 million customers.

inequality, and economic growth. Data from mobile transfer systems allow closer tracking of peer-to-peer transactions, which in turn may help produce more accurate estimates of remittances, regional disposable income, consumption patterns, and financial inclusion.

Box 2. M-Pesa Using Data Stored in Mobile Transfer Systems for Economic Policy Formulation

M-Pesa is a small-value mobile-phone-based money transfer system established in 2007 in Kenya. Since then, the system has expanded to developing economies in Africa, eastern Europe, the Middle East, and south Asia. Users deposit money through an agent in an account stored in an app on their cell phones. They can transfer money using personal-identification-secured text messages to any cell phone user. The transferred money may be picked up in cash by the beneficiary at any M-Pesa agent. Users are charged a fee for sending and withdrawing money. Originally devised to make money transfers easier, services were expanded to cover payment of salaries, microfinancing, and purchases of goods and services. In Kenya, more than 85 percent of households use M-Pesa. Some 78,000 agents are distributed everywhere in the country, reaching almost every village.

Potential Uses of Big Data

Measuring current transfers between M-Pesa users: Households pay bills and monthly insurance premiums or receive pension or social welfare payments. Many rural households rely on remittances from urban centers for survival. M-Pesa replaces a number of sometimes inefficient or limited options that were used for remittance payments (money transfer operators, foreign exchange bureaus, bus companies, friends and family) and gives households in rural areas access to financial services. M-Pesa data can produce more accurate estimates of indicators such as remittances, disposable income, and financial inclusion—all relevant information needed for real financial external sector indicators.

Estimating consumption patterns: Consumers use M-Pesa for cash in/cash out in retail stores, many of which are located outside urban centers. M-Pesa data can produce a clearer picture of regional demands by industry and serve as an early signal of changes in demand patterns.

21. Big data can support surveillance in low-income countries (LICs), where data problems are more acute. As illustrated by the case of M-Pesa, big data may become particularly supportive for monitoring macroeconomic and financial trends in LICs, where data are often scarce and outdated. A good example is household data, which are usually hard to collect. Human-sourced information from social networks, mobile data content, and counterparty data produced by businesses may be a good starting point to overcome data availability problems for households. Big data may also provide opportunities to improve data collection practices, compile new statistics, and reduce possible inefficiencies such as overlapping data collection (ADB 2013). The IMF Week @ the Beach Index (Box 3) is a valuable tool for the Caribbean region, where tourism accounts for a large part of GDP and regular cost and price indices are not well developed.

Box 3. Week @ the Beach Index

Inspired by *The Economist's* Big Mac Index, the Week @ the Beach Index constructs a price index comparing the cost of beach holidays around the world. The index incorporates in-country costs for a basket of goods typically consumed on a beach holiday, namely hotel rates, taxi fares, and price of meals and beverages (water, coffee, beer). The data for the index are drawn from Expedia and TripAdvisor (for hotel rates), Worldcabfares (for taxi fares), and Numbeo (for prices of meals and beverages).

Examples of Applications

The index is being recorded quarterly and has several uses: (1) as an indicator of competitiveness and a complement to external sector assessments, particularly useful in small states and some countries where tourism represents a large share of the economy; (2) as an alternative means to measure equilibrium real exchange rates; and (3) as an explanatory variable in empirical work, particularly as a longer time series is compiled. For example, the index was used as a variable to estimate price elasticities and individual variables, such as the number of hotels included as supply factors, when estimating the determinants of tourism flows. In addition, the sharing economy (for example Airbnb), which is not captured in official statistics, could be incorporated to improve market and price data.

Source: Laframboise and others 2014.

22. The IMF and other organizations could use big data to cross-check indicators used for surveillance that are currently produced by third parties and help assess the methods, data sources, and indicators' quality and characteristics. Big data are used by private sector companies and organizations to produce third-party indicators (TPIs) that are frequently used in IMF official staff reports. Alternative indicators could be produced by sourcing them through big data initiatives and could serve as a benchmark to assess TPIs' quality, characteristics, and methodological soundness.

23. Further big data analysis could shed light on the discussion when simple models and more data outperform more elaborate models based on less data (Mayer-Schönberger and Cukier 2014). Some potential gains from big data come not just from faster processing or better algorithms but simply because there is more and more varied data. The "datafication of everything" unlocks the latent value of a wider range of information. Thus, big data may also be an opportunity to rethink and reflect on how economic modeling can be conducted: Are we open to considering in some cases adaptation of analytics to the (new) data available, or do we prefer to stick to the traditional approach of adapting data to economic analysis?[8]

[8] Big data drive the broader adoption of existing analytic techniques at the IMF. Machine learning methods with algorithms that have been around for several decades have benefited tremendously from the availability of more data. The IMF working paper "Seeing in the Dark: A Machine-Learning Approach to Nowcasting in Lebanon" is one example of how these methods are being used at the Fund. In addition, "natural language processing" has generated a great deal of interest. The IMF recently hosted a seminar—"Text Mining, Inform and Connect," at which external and internal speakers presented their ideas and works in progress to infer meaning, context, and sentiment from unstructured content.

B. Big Data to Bridge Time Lags of Official Statistics and Support Forecasting of Existing Indicators

24. Faster insights are one of the biggest promises of big data, because key variables—for example, financial and price data—can be observed almost instantaneously. To monitor economic and financial developments and to provide early-warning signals of stability risks, timely data are essential to statistics fit for policy use. The winning IMF big data innovation project advocated the use of SWIFT data on transaction quantities and financial market prices to monitor global financial flows (Box 4).[9] SWIFT data are a promising example to obtain timely insights into possible contagion and spillover effects. Other studies are using SWIFT data to assess network concentration and cross-border transactions.[10]

Box 4. Using SWIFT to Monitor Global Financial Flows

SWIFT (Society for Worldwide Interbank Financial Telecommunication) is a standardized secure messaging system used by about 11,000 financial institutions in more than 200 countries. More than 25 million SWIFT messages are sent daily. SWIFT has more than 100 different message types providing information on the nature of the underlying transaction. For each type of message, the total number of messages sent and the total value of these payments are available.

Example of Applications

Global financial flows monitoring: SWIFT data are used to build indicators capturing financial flows around the world, broken down by region and currency. The *SWIFT Index* is one example that uses SWIFT data to provide early GDP growth estimates by using SWIFT traffic (volume) as a mirror for economic activity. When reviewing the SDR in 2015, the IMF used data from SWIFT as an indicator of the growing role of the Chinese renminbi in international transactions.[11] In the context of the IMF innovation challenge in 2015, a proposal was made to use SWIFT data to monitor global financial flows. An IMF staff team undertook a proof of concept exercise on using SWIFT data during 2016. In this context, SWIFT data have been used to nowcast international trade statistics to enhance timeliness.

Benefits

Currency compositions: SWIFT data could be used to monitor trends in the use of different currencies in financial transactions both globally and regionally. **Export/import indicators:** Different SWIFT message types may be used to distinguish trade from financial transactions, providing information to nowcast trade flows before official data are available. **Withdrawal correspondent banking relationships:** SWIFT data may support assessing and monitoring risks to correspondent banking relationships. In addition, access to SWIFT data by private financial institutions could support the banks' due diligence measures, such as anti-money-laundering principles.

[9] Limitations associated with the use of SWIFT data, among others, are **Access restrictions:** Institutions must purchase a subscription to SWIFT data. Although the interface is user friendly, there are limits on the size of a single download. **Publishing controls (data confidentiality):** SWIFT needs to approve external publication that uses their data to ensure confidentially is maintained. **Data processing:** Processing capacities must be built up to handle large data volumes. **Staff cost:** Some training is useful to become familiar with the interface and message types. Familiar tools such as Excel will not be sufficient to work with the data.

[10] Cook and Soramäki 2014 and Sy and Wang 2016 use the SWIFT message type MT103 to show regional network concentrations worldwide, with the United States being the network core. The Bank for International Settlements (BIS) in 2009 emphasized the opportunities to use SWIFT data to make cross-border transactions more transparent, especially if an intermediary bank, located in an economy other than that of the originator, is involved. In 2015 the BIS reiterated the usefulness of SWIFT data to better understand correspondent banking relationships and to allow for the monitoring of associated risks of withdrawal and pressures on correspondent banking relationships.

[11] IMF November 2015 Policy Paper "Review of the Method of Valuation of the SDR."

25. The use of big data creates an opportunity to extract economic signals in almost real time and to nowcast economic series before official figures are published. Timely information derived from big data will help assess the financial and economic state of individual economies and the world economy. It could support surveillance in maintaining stability, by providing almost-real-time information on fast-moving financial, commodity, and other markets, hence helping prevent or at least mitigate crises in the international monetary system. Information obtained through nowcasting exercises will further strengthen policymaking by extracting global prospects in such domains as financial markets, public finance developments, and regional economic outlooks.

26. Nowcasting (or real-time forecasting) is already widely used in the private and public sector for many indicators (FRBNY 2017; Banbura and others 2013).[12] The Billion Prices Project at MIT uses web scraping to collect price data from online retailers to monitor price changes and inflation at a daily frequency as well as turning points in inflation trends. Twitter was used to nowcast food prices in Indonesia by filtering and modeling tweets with price information (UNGP 2014). The private firm Predata (2016) condenses data sources from around the web into signals for geopolitical risk by monitoring digital conversations across open-source social and collaborative media. Subsequently, machine learning algorithms are applied to these signals to anticipate a variety of developments from asset price changes to civil protests, labor strikes, electoral results, and national security outcomes.

27. Other prominent examples of nowcasting are economic series on tourism, unemployment, retail trade, and trade flows.[13] Statistical compilers are venturing into the world of social networks and new applications of satellite imagery.[14] The Netherlands uses Facebook and Twitter data to estimate consumer sentiment, and China and Italy approximate job vacancy rates through web scraping. Other well-known examples are the use of publicly available Google Trends data to nowcast unemployment and consumer sentiment, as well as car and property sales (Pavlicek and Kristoufem 2015).

28. While leading indicators (OECD 1987)[15] have traditionally been used to anticipate financial and economic developments, big data may in some cases provide better predictors and improve forecasting accuracy thanks to the amount of available data. As a first step, forecasting exercises look at the range of information available to assess how the recent past developed in comparison with what was expected. Several studies have found that through more detailed and granular information provided by big data, the quality of estimated economic series can be improved (Galbraith and Tkacz 2013). Forecasting exercises may benefit

[12] Companies use forecasting techniques to streamline supply chain management.

[13] The Ministry of Finance in Colombia uses short-term trends to monitor economic activity derived from Google Trends.

[14] United Nations Statistics Division initiatives on Big Data.

[15] Leading indicators have been used to provide early-warning signals of turning points in economic activity (for example, the Organisation for Economic Co-operation and Development's composite leading indicators).

from big data by complementing existing statistical series with more granular, higher-frequency, and socioeconomic data. Through learning to harness big data sources and increasing the number of observations, big data may improve forecasting accuracy.

29. Even if new data sources serve as an input to forecasting, consistent and harmonized historical time series are still needed. While the time lag and frequency issues pertaining to official statistics spur the development of indicators from big data, the two should be seen as complementary. Both outputs should be compared to ensure robustness of new indicators vis-à-vis existing time series that can serve as a benchmark. Big data mainly measures "insights" and correlations (for example, movements, trends, sentiments) in contrast to "actual information" (for example, positions of outstanding debt at the end of the year).

C. Big Data as Data Source and Innovation in the Production of Official Statistics

30. Worldwide discussions of how official statistics should evolve in the age of big data show that statistical agencies are starting to undergo significant changes. At its 45th session in 2014, the UN Statistical Commission officially recognized that "Big Data constitute a source of information that cannot be ignored" (UNGWG 2017). In Europe, in 2013, all heads of the European national statistical offices signed the "Scheveningen Memorandum on Big Data and Official Statistics" (EC 2014) and committed to integrating an official statistics big data strategy action plan and road map with the wider governance strategy. Eurostat stressed that the typical use of big data in the EU member states will not be in isolation but in combination with already existing data sources (EC 2017). Indeed, the global statistical community is working on demystifying big data and embracing this novel data source as both a challenge and an opportunity going forward.

31. Strategies for integrating big data into official statistics could range from partially to entirely replacing existing statistical sources and from providing improved to complementary or completely new statistical outputs (Florescu and others 2014). In an era of limited budgets and declining responses to surveys (for example, Meyer, Mok, and Sullivan 2015) official statistics must explore new data sources, such as big data, that have the potential to be as relevant, yet more timely and more cost-effective, than traditional data collection methods. National authorities see [16] promising merits in producing faster and more frequent statistics, along with reducing the response burden and modernizing statistical production processes.

32. Compilers are piloting big data projects mostly to refine and complement traditional data sources. Some well-documented examples of big data use by official compilers are[17] (1) mobile phone data for tourism, transportation, and urban statistics (for example, Eurostat, Belgium, Brazil, Indonesia, Israel, Italy, World Bank in Nigeria, Poland); (2) web scraping

[16] UNSD GWG Survey and Project Inventory, https://unstats.un.org/bigdata/inventory/.

[17] UNSD Big Data Project Inventory https://unstats.un.org/bigdata/inventory/.

for price indices, labor market indicators, and enterprise profiling (Eurostat, China, Ecuador, Finland, Germany, Hungary, Japan); (3) smart meters for energy and environmental statistics (Eurostat, Belgium, Canada); and (4) credit card, cash register, and scanner data for price and other economic statistics. Table 1 in Section V presents big data sources with the potential to feed in to the official statistical domains. It is based on the UN classification (see Box 1) and tailored to the IMF statistical needs for surveillance.

33. In many countries, big data could be a low-cost, high-quality data source alternative for compiling official statistics. A promising example is Estonia's use of mobile positioning to collect data on travel services for balance of payments statistics as an alternative to border surveys (Box 5). In this case, the use of big data not only significantly reduced costs, it also improved the quality, accuracy, data availability, and comprehensiveness of the estimates compared with the traditional survey based estimates. For countries that largely depend on tourism, this is an opportunity to be considered.

Box 5. Mobile Positioning Data as a Data Source for International Travel Service Statistics

Background: In 2010 Statistics Estonia stopped using border surveys, which forced Eesti Pank (Central Bank of Estonia) to explore other ways to collect travel service estimates. Among the various alternatives explored (road sensors, credit card information, accommodation-based data), mobile positioning data proved to be the simplest, least costly, and timeliest. Moreover, there are no legal obstacles to accessing and using these data in Estonia.

Estimating travel services exports/imports. Mobile positioning data can be used to estimate trip duration, inbound and outbound international travelers, and monthly travel services exports/imports.

Cell phone use and location data provide information such as SIM card ID, date and time, location, and country code. Residency of a customer is assumed to be the SIM card residency. Methodology is based on the analysis of roaming patterns of SIM cards of anonymized data: (1) Nonresidents in the network of resident mobile operators provide estimates of inbound travelers and country. (2) Roaming activity reports received by resident mobile operators from nonresident mobile partner companies are used to estimate outbound travelers and country. Algorithms consider short and long visits and adjust for transit travel (harbors, airports, transit roads) and look at permanent workers from other countries and border noise (ship traffic and random switching). These data also provide the length of stay for travelers: for inbound travelers—number of days between first and last roaming activity; for outbound travelers—same as for inbound travelers, plus mapping of more than one country for a single trip.

Benefits: There are some significant gains in using these data compared with the data from conventional surveys and administrative sources. **Enhanced quality:** Accuracy, quality, and comprehensiveness of the estimates improved compared with traditional survey-based estimates. Travelers with captured nonpaid (staying with relatives or friends) and nonregistered hotel and other accommodations made the estimates more comprehensive. Travel services could be compiled with a more detailed breakdown by region or country of residence. **Timeliness:** Data are available almost to real time, so monthly and quarterly estimates could be generated with practically no time lag. **Cost**: Data are readily available; hence the costs are much lower than collection and processing of data from surveys and administrative and other sources.

Source: http://www.oecd.org/trade/its/46287481.ppt.

34. To stay ahead of the curve, statistical agencies should seize new opportunities that come with big data by providing additional services to their users. These could include the creation of flash estimates to bridge the timeliness of survey results and the production of innovative short-term indicators of economic and social interest. Furthermore, statistical agencies could consider new tasks, such as the accreditation or certification of data sets created by third parties or public or private sectors. By widening its mandate, it would help keep control of quality and limit the risk of private big data producers and users fabricating data sets that fail the test of transparency, proper quality, and sound methodology (MacFeely 2016).

Box 6. Administrative Data and Big Data

Administrative data can be considered a distinctive form of big data, originated as a byproduct of (usually) large-scale administrative systems and usually generated for purposes other than official statistics. Most statistics agencies around the world make use of administrative data. Reliance on data collected for nonstatistical purposes ("secondary data") with potentially extensive coverage instead of surveys ("primary data") has been conducive to addressing the growing demand for more and better statistics despite limited resources and declining response rates. Tax data, public finance transaction data, and social security records are prime examples.

Years of experience with administrative data may have paved the way for statistical agencies in their use of big data. Statistical agencies are mindful that they can take advantage of innovative data sources only when they ensure data quality, methodological soundness, and the protection of privacy and confidentiality. These prerequisites are important to maintaining public trust in the compiling authorities' ability to properly manage primary and secondary data sources and in the disseminated official statistics. Most statistical agencies have limited control or influence over the design and operation of the data collection process in administrative data—with exceptions such as the European Nordic countries (Denmark, Finland, Iceland, Norway, Sweden). Yet contrary to most types of big data, administrative records originate from the same public sector as official statistics offices. That said, the use of big data may prove more challenging than the integration of administrative data.

As with any data collection program, consideration of the use of source data is a matter of balancing the costs and benefits. The use of secondary data from administrative sources reduces collection costs and diminishes the often criticized respondent burden, giving the statistics agencies some leeway when negotiating new data demands. Administrative records have a variety of statistical uses: (1) as a sole source for some indicators (for example, customs data for merchandise trade, International Transactions Reporting Systems for balance of payments statistics); (2) combined with other sources (for example, tax records to measure production, use of taxation data for small businesses); (3) for verification and calibration (for example, comparison of survey estimates with estimates from a related administrative program); (4) for indirect estimation (benchmarking); and (5) for survey design (taking advantage of the richness of details).

In the **Nordic countries,** the use of registers for statistics is well established. Benchmark examples are the Population and Housing censuses, which are totally based on registers in Denmark—first in the world (1980)—Finland (1990), and Norway and Sweden (2011). Population structures were compiled based on registers in the 1970s in all Nordic countries.

Registers are central to the use of administrative data, in different ways:

- Statistics based on one register (population structure, vital statistics)
- Combining several registers together (population and housing census)
- Combining registers with survey data
- Sampling frames
- Quality control

Preconditions for effective and efficient use of administrative data are

- National statistical office access rights to administrative sources to produce the official statistics
- Unique identification for all individuals and businesses, widely used in all administrative registers to combine data from different sources
- Adherence to confidentiality; that is, used only in aggregated form

- Influence of the statistical agencies in the design of the generation and collection of administrative data

Source: UNWDF 2017.

IV. WHAT CHALLENGES COME WITH BIG DATA?

A. Data Quality

35. For policymaking, the quality assessment of indicators derived from big data will be crucial to minimize governance, political, and reputational risks. While there is a strong demand for timelier and more granular data, the quality of indicators and underlying data sources must be assessed. During an experimental phase, newly produced indicators need to be benchmarked to existing ones for evaluation. This is to ensure that the new indicators meet the minimum data quality standards defined in long-established quality frameworks for real, external, fiscal, monetary, and financial statistics.

36. The suitability of any new data source will need to be assessed against a number of core features, such as accuracy, sustainability, and methodological soundness, while metadata[18] will be key to interpreting and assessing new data sources.[19] There is no assurance that the frameworks and interaction that generate big data will be around in the future because big data are mostly produced by the private sector as a byproduct of their business models and technologies. which are subject to change in the context of competitive markets. Consequently, the availability, comparability, and coherence of time series are at risk (Kitchin 2015).

37. Many types of big data do not represent random samples of the population. With social media services, for instance, population subgroups that do not use the technology will be underrepresented, if not otherwise captured or adjusted for. The users of big data must determine whether the specific big data source is methodologically sound and represents the population being analyzed. Metadata about population, units and events, the methods and processes applied, and their suitability and completeness are exceedingly important (UNECE 2014).

38. Indicators based on big data have a short time span and contain outliers, and their continuity cannot be guaranteed. Accessing this byproduct has only started recently; this limits comparability over time. Big data is often unstructured and therefore requires a proper transformation into time series observations and cleaning variables, such as replacing outliers

[18] The term "metadata" refers to metadata that provide information on every aspect of the data production cycle, such as data access, statistical concepts, compilation practices, and methodologies, as well as agencies assuming responsibility for the production of data.

[19] These factors are at the center of the IMF's Data Quality Assessment Framework, which is used for comprehensive assessments of countries' data quality (OECD 2012). Measure data quality across seven features: relevance, accuracy, credibility, timeliness, accessibility, interpretability, and coherence.

and missing observations with estimates (Eurostat 2016). Continuity of data provision cannot be guaranteed by regulatory frameworks but would become relevant for the use of big data in official statistics. Overall, instability can be due to institutional changes and discontinuity in data provision, but also because of improved technology and subsequent consumer behavioral changes.

39. Quality is a central concern in the production of data by official statistical agencies. Official statistics are a public good, and their professional standards and norms are reflected in the UN Fundamental Principles of Official Statistics (UNSD 2014). Innovative sources or methods need to be assessed for their suitability for producing official statistics. As expressed by the 5V characteristics, big data is complex, incomplete, and noisy and can contain outliers and extreme events. A lack of experience, best practices, comprehensive analysis, methodology, and research of quality standards could undermine the quality standards that statistics producers adhere to. Many Organisation for Economic Co-operation and Development countries that are experimenting with big data are hesitant, mainly due to the lack of a methodological framework related to data sources used (UNSD 2015). Going forward, research in and compilation of statistical techniques and methodologies best practices that address veracity and volatility, specifically, need to be at the top of the agenda of the statistical community.

40. Putting long-term uses of big data aside, "good enough" information that is available "now" can be used "now" for specific actions (Meyer and others 2013). Big data may not adhere to comprehensive data quality standards, but could still uncover meaningful insights by alerting us that something is happening. One example is sentiment analyses that are used to mine various sources of unstructured data for opinions or trends.

B. Access to Big Data

41. The origin and creation of big data are mostly outside the control of national or international institutions, with the notable exception of administrative data. Access to big data generally means access to proprietary data held by the private sector. To access, experiment with, and use big data effectively, users need to enter into agreements with private data owners, while maintaining their independence, and ensure a legal environment that addresses both privacy and confidentiality. When access is granted, users of big data sources should be transparent about the data source, the legitimacy, and the purpose of the use of the data. Traditional techniques used to ensure confidentiality will reach their limits, and users will need to find technologies and methods that safeguard their independence and reputation and the public trust. For instance, when governments agree that their compiling agencies cannot be left out of the big data revolution, an important role of the legislative body is to negotiate access to big data.

42. Privacy, confidentiality, and cybersecurity risks are a major concern when using big data. Big data contains large amounts of sensitive and personal information that may be exposed to privacy, confidentiality, and cybersecurity risks. If not sufficiently protected this

information can be vulnerable to cyberattacks, used to profile individuals, and sold to third parties. Information, depending on the national legal framework, may not be used legally without a person's prior knowledge. The potential loss of data can lead to reputational damages as well as loss of consumers' trust. Thus, international institutions and public agencies will need to ensure that data sources and indicators used were obtained without any violation of privacy or confidentiality regimes.

43. Privacy protection procedures and information technology techniques will be key to minimizing privacy, confidentiality, and cybersecurity risks when using granular big data sources. Companies, public agencies, and third-party data users will have to establish thorough privacy protection procedures. They must invest in security layers and adapt traditional information technology (IT) techniques such as cryptography, anonymization, and user access control to big data characteristics (Moreno, Serrano, and Fernández-Medina 2016) to safeguard privacy and ensure that data are not being reconstructed and traced back to the individual. Moreover, consumer rights may have to be strengthened through regulatory intervention, detailing what companies can do with the collected information (Smith and others 2012). The former will be essential to minimize privacy, confidentiality, and cybersecurity risks.

44. Several countries (UNSD 2015) have begun establishing public-private partnerships, but access rights are sometimes unaddressed and unclear. To help overcome these problems, the UN Global Working Group is currently preparing "access recommendations" and best practices for building lasting partnerships between official statistics producers and data owners. What is needed are standardized, ethical, and stable data sharing tools and protocols with private companies (Letouzé and Jütting 2014). However, the risk of volatility persists. There is no assurance that a company and its data will exist in the future—possibly leaving analytical and surveillance applications at risk in terms of continuity.

45. The price of big data will not necessarily be low or negligible. Where there is demand, there is a market—while the costs of producing big data may be marginal for the companies, the licensing or propriety costs for accessing this information may be significant. These costs will come on top of acquiring the processing and storage technology for in-house use or access through external vendors.

46. Big data is not just a byproduct but is becoming a major asset for companies. New companies that have emerged from the "big data revolution" range from data brokers, consultants,[20] and companies installing on-site technical platforms to so-called cloud vendors that offer big data as a service for a fee either on a public or a proprietary cloud. Increasingly, private companies sell their data for a profit; over time the generation of big data may become a major objective, and not just a byproduct, of their activity. The statistical community and official users, as well as national and international policymakers, will confront many decisions, including complex negotiation processes.

[20] www.bigdatapartnership.com.

47. Despite a multitude of technical options and platforms to choose from, the selection processes are quite complex. A typical approach that companies follow in setting up a big data practice is establishment of a center of competence focusing on technology options. They also set up an exploratory environment accessible to experts and practitioners. During the first year, an institution should aim to (1) identify standard technology platforms that can be provisioned by teams working on various projects; (2) establish the governance around the use, monitoring, and costs associated with the environment; (3) establish a network that allows them to tap into experts with specific skills; and (4) make business and budgetary provisions to fund the costs for the future.

48. Cloud technology offers flexibility to scale the technical infrastructure up or down depending on the needs. It is the preferred approach in the industry and typically can require a budgetary allocation of about $250,000 for the first year. The handling of costs for subsequent years are based on "pay for what you use" and requires strong governance to ensure efficient use of funds. It is expected that cloud-based systems will see wider adoption by comparator organizations and that the cost of storage and computing power will continue to drop. The IMF has started provisioning the cloud service and expects that, over the next few years, associated costs will grow incrementally with the incorporation of big data into analytic practices. The cost of data will be more difficult to predict, as some sources that could be useful may not yet exist as discrete products or may not even be available. In the future, some data may be in the public domain, while some may require licensing from commercial sources. Both cloud-based computing services and commercial data are typically licensed on a recurring basis, which has implications for the administrative budget.

C. New Skill Profiles and Technologies

49. Multidisciplinary teams will be needed to make big data speak. Statistical agencies engaged in big data projects know the importance of collaborating with abundant human and technical resources to proficiently exploit big data. The UN Global Working Group concluded that a multidisciplinary project team from different professional backgrounds is necessary to adapt the analytics to the data—rather than relying on traditional collection means such as carefully designed questionnaires. Statistical agencies, central banks, public agencies, and international organizations will not only have to train and develop existing staff to deal with big data but also must compete with the private sector to recruit or contract new staff. These new staff members must be familiar with big data—for example, data scientists (see Figure 3), IT architecture specialists, and data visualization specialists who work alongside subject matter professionals. On average, a big data practice in the initial stages starts with a core team of three to four staff members with a mix of technical and business skills.

Figure 3. Data Scientist

Box 7. Rethinking of Information Technology & IT Governance

Peter Lyman and Hal R. Varian's 2000 study "How Much Information?" (Lyman and Varian 2000) was the first to quantify the total amount of information stored on physical media—stating that the "world produces between 1 and 2 exabytes of unique information per year, which is roughly 250 megabytes for every man, woman, and child on earth." Also in 2000, Francis X. Diebold presented to the Eighth World Congress of the Econometric Society a paper titled " 'Big Data' Dynamic Factor Models for Macroeconomic Measurement and Forecasting," in which he states that "Big Data refers to the explosion in the quantity … of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology." In 2001, Doug Laney described the 3Vs that are now the generally accepted characteristics of big data.

Big data is not just big; it's expansive—it demands rethinking of information technology and IT governance. Long-standing business practices and legacy technology will need to be reconsidered as new data sources and analytic techniques are introduced to the institution.

Information Security and Privacy: Understanding the provenance of data from multiple sources, when there are transaction- or observation-level data, and ensuring that national and international standards for data privacy are observed are essential to the credibility of public sector initiatives.

Governance of Research Projects: Data-management practices at the IMF have relied upon structured processes around the publication of board papers and production of the *World Economic Outlook* to ensure that data of institutional value are preserved and documented. Big data projects are less likely to be tied to such processes and as such will need guidelines that address such issues as data retention, data destruction, the cost of running large models, and the sharing of software code and algorithms.

Managing Research Projects through Their Life Cycle: Technology groups are organized and optimized to deliver software and systems to support them in a production setting. Big data research projects are more likely to be highly iterative and may not result in deployment of a system but rather of new data sets, models, and visualizations. Such efforts will likely be led by the specific projects, with technology playing a subsidiary role.

The Challenges and Opportunities Presented by Open-Source Software: Much of the technology used in big data research and analysis is open-source software, evolves quickly, is governed by a loosely coordinated community of practitioners, and is supported via crowd-sourced collaboration spaces. Adopting a similar support framework at the IMF would initially be countercultural.

Cloud Computing: Many organizations appear slow and reluctant in the face of adopting the cloud as a platform for information technology. For big data, this reluctance will need to be addressed, owing to the low cost, speed of deploying new servers and software, and rapidly maturing service and security models available in the marketplace.

This box was prepared by El Bachir Boukherouaa, IMF IT Department.

V. STATISTICAL IMPLICATIONS

50. Statistical agencies should further step up their involvement in big data projects and learn about what has already been achieved in close collaboration with international organizations leading the methodological development. Statistical agencies could contribute, learn, and profit from projects and work already ongoing in the big data community. It will be essential to not only leverage but also contribute to existing big data networks, such as the United Nations Global Working Group on Big Data. Robust governance frameworks around big data that promote international and national cooperation will need to avoid overlaps and ensure effective collaboration.

51. Statistical agencies need to take stock of those projects that have potential to enrich official macroeconomic and financial statistics. Expertise is needed for the judgment of data quality issues and methodologies, but also in areas such as institutional change management, building of strategic partnerships, and communication with users. For the inventory of best practices, it is necessary to cultivate strong links to networks of big data experts in academia, the private sector, and the international statistical community and use the network to streamline international efforts on training, skills, and capacity development in member countries. Big data applications could introduce a paradigm shift for the statistical system, comparable to the integration of administrative data over a decade ago, but in a more revolutionary way. But as for administrative data, data availability, access, and, thus, progress will be uneven across countries and organizations.

52. The UNECE High-Level Group for the Modernization of Official Statistics (HLG-MOS) made the sensible decision to create the Sandbox,[21] before engaging in costly adventures. The Sandbox is a shared platform and data repository (subject to confidentiality constraints) that statistical agencies and other users can utilize to test and evaluate tools, techniques, workflows, and methodologies and to collaborate with their peers on the way forward. The bottom line for using big data is of course whether the benefits can outweigh the costs and whether a certain level of quality and international comparability can be ensured.[22] Participation in the UNECE Sandbox projects is recommended as the proper environment to stimulate discussions about approaches that could work in the compilation of official statistics, among other things (UNSD 2015). In particular, countries with less-developed statistical systems may benefit from the Sandbox.

53. The most promising big data sources for macroeconomic and financial statistics will be identified over time and with asymmetric opportunities for individual countries. Social

[21] http://www1.unece.org/stat/platform/display/bigdata/Sandbox. The current annual fee is €10,000. The platform is composed of a set of five servers that enable reliable storage of up to 56 terabytes of data and high-performance concurrent processing of these data across 80 CPU cores. Users access a unified login portal connected to the internet via a high-bandwidth network connection.

[22] Statistics Netherlands (CBS) has used the Big Data Sandbox as a prototype to create its own internal platform.

networks, traditional business systems, and the internet of things generate data types that may feed into standard statistical domains (national accounts, external sector, financial, government finance and price statistics) as well as additional fields. Taking into consideration the countries' characteristics and after passing the proof of concept, statisticians—jointly with stakeholders—will establish the most promising big data programs for macroeconomic and financial statistics. Given the cost and complexities of new required technologies, skills of big data users, and available source data in individual countries, the actual opportunities for individual countries may be asymmetric. In other words, not all countries will benefit from big data in the same fashion.

54. Table 1 (Appendix II) presents mapping between different types of big data and standard macroeconomic, financial, and other statistics. Organized in four columns, Table 1 uses the big data classification from Box 1, and links it to the traditional statistical domains, as well as to some additional more specific statistics that will potentially benefit from the new big data sources and types.

55. There are many promising big data applications relevant for macroeconomic and financial statistics, but where to start? While the most promising big data sources for macroeconomic and financial statistics will be identified over time, considering the countries' characteristics, there are already many big data applications being conducted that can serve as a basis for the medium term. Table 2 may serve as a starting point.[23]

56. Although there appears to be a preference for some big data sources, derived indicators span across all traditional statistical domains and beyond. For practical illustration, Table 2 (Appendix III) provides a broad overview of big data applications and their potential for macroeconomic, financial, and other statistics. Organized in four columns, Table 2 showcases these applications and creates the link between the big data sources used, indicators derived, and the statistical domains potentially benefiting. The motivation for using big data is structured along the three potential areas of big data defined in Section III: (1) to answer new questions and produce new indicators, (2) to bridge time lags in the availability of official statistics and support the more timely forecasting of existing indicators, and (3) as an innovative data source in the production of official statistics. While a preference for some big data sources can be observed, such as Google Trends, Twitter, Facebook, official registries, and scanner and price data, the indicators inferred cover all traditional statistical domains and allude to ample new applications and statistics.

57. Official statistics need to develop new data quality concepts and expand existing frameworks to incorporate the opportunities and challenges that come with big data. Extracting insights from unstructured data (Piatetsky-Shapiro 2012) as opposed to compiling factual accounting information (for example, using companies' balance sheets) creates a new landscape for official statistics that requires revisiting compilation processes and statistical deliverables. Data quality frameworks must become adjusted as well to the new data landscape.

[23] A big data inventory can also be found here: https://unstats.un.org/bigdata/inventory/.

At the same time, data comparability over time and across countries should be preserved to the extent possible, a goal that will be challenged by the very nature of big data. Both statisticians and users need to be mindful of the resulting trade-offs.

58. To assess and ascertain data quality of big data sources and methods, the official statistical community should join and coordinate efforts through the standing committees on statistics[24] and other expert groups. International statistical manuals and guides, such as the IMF *BPM6 Compilation Guide* and the UNECE *Guide to Measuring Global Production*, may need to be updated to keep providing a sound basis for the development and application of statistical practices and incorporating the new opportunities and challenges raised by big data. The standing committees' views, advice, and validation of the appropriate way to integrate big data in official statistics would provide reassurance on continued reliability, accuracy, and methodological soundness of the data sets. This could lead to the drafting of practical guidance notes and metadata outlines, building on what has already been done by national and international organizations.

59. Inside organizations such as central banks and the IMF and statistics and user departments should join forces to identify a core set of early-warning and snapshot indicators that can be used for policymaking. Statistical support will be key for the evaluation and validation of big data indicators for their suitability and methodological soundness. The use of big data for new indicators must be made transparent in terms of the applied methodology and the data origin; otherwise the value of policy advice and forecasting can be seriously weakened. Statistics departments will have a lead role in the drafting of guidance notes for best practices and robust concepts to support staff in using these indicators consciously. Information technology departments will be important stakeholders in this effort to provide the technical support.

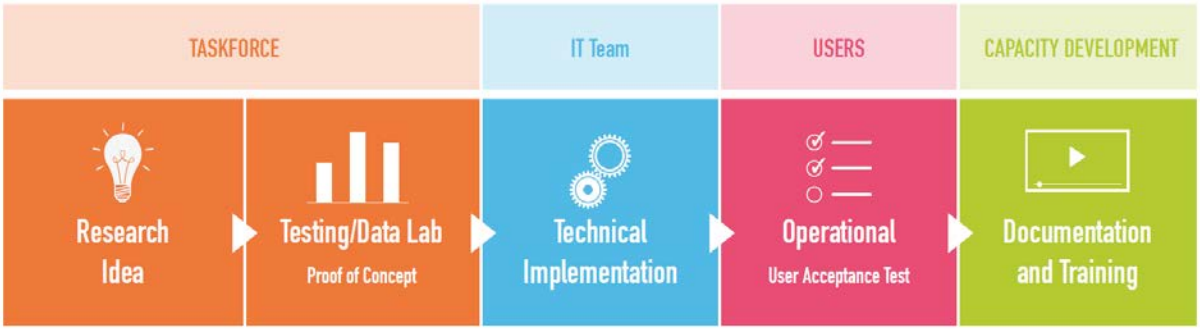
60. Specialized task forces could be established to break silos and concentrate on selected big data projects. In addition to cross-departmental groups, specialized task forces consisting of specialized topic-related and IT experts could be established. Task forces could be dedicated to pilot projects and would report to the cross-departmental group on the progress, challenges, and opportunities of the pilot projects under consideration. By employing multidisciplinary expertise in the task forces, big data will also be an opportunity for national and international institutions to break silos.

61. These specialized task forces could start with a brainstorming phase (Figure 4) to identify the desired outputs. Phases 1 and 2 could lie with the established task forces, working within the time targets set by the cross-departmental group. Phase 3 would be the technical implementation led by IT teams in collaborations with the task force experts. Simultaneously, the new products would have to be assessed by users for their "fit for purpose" (adequacy) for

[24] Such as the International Agency Group on Economic and Financial Statistics, the IMF Balance of Payments Statistics Committee (BOPCOM), Government Finance Statistics Advisory Committee (GFSC), and the Inter Secretariat Working Group on National Accounts (ISWGNA).

surveillance work. Statistics departments should provide the methodological guidance. If regarded as "fit for purpose" the new products could be piloted, incorporated into surveillance work, and expanded to a larger scale. In the last phase, institutions could decide whether to include the new products in capacity development activities (training, technical assistance).

Figure 4. Pilot Studies by ad hoc Taskforces



62. An outcome of the specialized task forces could be dynamic and interactive guidance notes. Guidance notes would entail assessments of pilot projects for methodological soundness, reliability of data source, and their potential to serve as an input to policy analysis for both international and national organizations. The guidance notes should be dynamic (timely and high-frequency updates) and interactive (online based documents, with interactive links and search functions). For some projects, new methodologies may have to be developed.

63. Capacity Development and Technical Assistance. For institutions such as the IMF—which are strongly engaged in technical assistance to developing economies—the results for statistics and for indicators directly usable in bilateral and multilateral surveillance could be incorporated into capacity development programs after having further explored their potential and challenges mentioned in this paper. Capacity development could encompass the provision of expertise in the judgment of data quality issues and methodologies, as well as the conveyance of best practices on institutional change management, building of strategic partnerships, and communication with users. For the inventory of best practices, international institutions need to cultivate strong links to networks of big data experts in academia, the private sector, and the international statistical community and use the network to streamline international efforts on training, skills, and capacity development in member countries. The UNSD Global Working Group, composed of nine other international and regional organizations and 22 countries, is an already established expert group on big data in the field of official statistics.

VI. CONCLUSIONS AND WORK AHEAD

64. By now, many national and international statistical organizations have recognized that big data is not just a buzzword, but a potentially strategic asset that requires a vision and a plan. **Big data necessitates a strategy to select the most promising applications to complement official statistics and to bring added value, such as improved timeliness, supporting forecasting of existing data sets, and producing new indicators.**

58. Organizations need to go beyond the existing individual and scattered applications of big data. Organizations embarking on big data analytics require a strategic organizational plan to deliver measurable and high-scale results. Before engaging in costly and time-consuming investments, particularly in low-income countries, international and national organizations should begin with a proof of concept or pilot project, and the project should be operationalized only after the findings have proved valuable and feasible from an organizational point of view. Sound partnerships, legal issues, and the right skills and technologies are as important as statistical expertise, data representativeness and methodological accuracy, and effective collaboration between data scientists and subject matter economists. Especially within the realm of official statistics, international coordination efforts are key.

59. Best practices for the building of lasting partnerships between official statistical agencies and data owners are being developed and tested, legal questions clarified, and best-use cases field-tested. Organizations learn that big data success is not about implementing one piece of technology, but about putting together an environment of people and processes that take the big data innovations forward and put them to work. Infrastructure spaces and big data sources continue to evolve, and with them the possibilities, challenges, and limitations. Given the diverse skills and collaboration needed, big data projects are also an opportunity to break institutional silos.

65. The actual opportunities offered by big data to macroeconomic and financial statistics vary considerably across statistical domains. There are promising opportunities for statistics on flows and transactions, insights, correlations, trends, and sentiments, but, currently, less so on statistics on stocks or the breakdown of flows into transactions, revaluations, and other volume changes. Detailed country-by-country time series in accordance with internationally agreed standards remain crucial for measuring and monitoring countries' economic performance and policies over time.

66. International organizations responsible for official statistics should work in close cooperation with user departments. A new dimension for international statistical coordination and cooperation should be considered, including its incorporation into capacity development activities. Statistics departments of international organizations and the wider statistical community can contribute to the big data discussion in their respective area of expertise. The IMF's Statistical Department will continue "its mission to provide strong leadership for the development and application of sound statistical practices" and reach out to the rest of the IMF and the community at large to foster the use of big data for macroeconomic and financial statistics as input for policymaking.

67. Big data is not a static but a dynamic phenomenon, so the systems and networks generating it will continue to evolve, as well as the opportunities that big data offers, the challenges it poses, and its statistical implications.

VII. REFERENCES

Arouri M., A. Aouadi, P. Foulquier, and F. Teulon. 2014. *Can Information Demand Help to Predict Stock Market Liquidity? Google It!* https://www.ecb.europa.eu/events/pdf/conferences/140407/Aouadi_CanInformationDemandHelpToPredictStockMarketLiquidityGoogleIt.pdf?7dd64c397041aaf1086faf73b3eac25b.

Asian Development Bank (ADB). 2013. *Big Data: Vital Statistics for Development*. https://www.adb.org/features/big-data-vital-statistics-development.

Banbura, M., D. Giannone, M. Modugno, and L. Reichlin. 2013. "Now-Casting and the Real-Time Data Flow." Working paper series 1564, European Central Bank, Frankfurt. https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1564.pdf.

Bank for International Settlements (BIS). 2015. *Central Bank Use of and Interest in Big Data*. Irving Fisher Committee report. http://www.bis.org/ifc/publ/ifc-report-bigdata.pdf.

Central Banking. 2016. *Big Data in Central Banking: 2016 Survey*. www.centralbanking.com.

Chain Store Age. 2015. *Survey: Marketers Value Personalization*. http://www.chainstoreage.com/article/survey-marketers-value-personalization#.

Challenge 4 Development. http://www.d4d.orange.com.

Cook, S., and K. Soramäki. 2014. "The Global Network of Payment Flows." SWIFT Institute Working Paper No. 2012-006. https://www.swiftinstitute.org/wp-content/uploads/2014/09/SWIFT-Institute-Working-Paper-No.-2012-006-Network-Analysis-of-Global-Payment-Flows_v5-FINAL.pdf.

Data Floq. *How Big Data Can Help the Developing World Poverty*. https://datafloq.com/read/big-data-developing-world-beat-poverty/168.

Data Revolution Group. 2014. *Data Innovation: Big Data and New Technologies*. http://www.undatarevolution.org/data-innovation.

Davenport, T. 2006. "Competing on Analytics." *Harvard Business Review*. https://hbr.org/2006/01/competing-on-analytics.

Donovan, K. 2012. "Mobile Money for Financial Inclusion." In *Information and Communications for Development 2012: Maximizing Mobile*. Edited by the World Bank Group, Washington, DC. http://siteresources.worldbank.org/extinformationandcommunicationandtechnologies/resources/ic4d-2012-chapter-4.pdf.

European Commission (EC). 2014. *Scheveningen Memorandum*. https://ec.europa.eu/eurostat/cros/content/scheveningen-memorandum_en.

———. 2017. *Big Data.* https://ec.europa.eu/eurostat/cros/content/big-data.en.

European Commission, Eurostat. 2016. Big Data and Macroeconomic Nowcasting: From Data Access to Modelling. http://ec.europa.eu/eurostat/en/web/products-statistical-working-papers/-/KS-TC-16-024.

Federal Reserve Bank of New York (FRBNY). 2017. "Nowcasting Report." https://www.newyorkfed.org/research/policy/nowcast.

Florescu, D., M. Karlberg, F. Reis, P. R. D. Castillo, M. Scaliotis, and A. Wirthmann. 2014. *Will 'Big Data' Transform Official Statistics?* Eurostat. http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf.

Galbraith, J. W., and G. Tkacz. 2013. "Nowcasting GDP: Electronic Payments, Data Vintages and the Timing of Data Releases."

Karabulut, Y. 2013. *Can Facebook Predict Stock Market Activity?* https://www.ecb.europa.eu/events/pdf/conferences/140407/Karabulut_CanFacebookPredictStockMarketActivitiy.pdf?902eb04ceaa17187b7353be87992b83a.

Karani, K. 2017. *Training Course on Gender Statistics and Gender Budgeting*. https://www.linkedin.com/pulse/training-course-gender-statistics-budgeting-kennedy-karani-3.

Kitchin, R. 2015. "Big Data and Official Statistics: Opportunities, Challenges and Risks." *Statistical Journal of the International Association of Official Statistics* 31 (3) (9).

Laframboise, N., N. Mwase, J. Park, and Y. Zhou. 2014. "Revisiting Tourism Flows to the Caribbean: What Is Driving Arrivals?" IMF Working Paper 14/229, International Monetary Fund, Washington, DC.

Letouzé E., and J. Jütting. 2014. *Official Statistics, Big Data and Human Development: Towards a New Conceptual and Operational Approach*, Paris 21. https://www.odi.org/sites/odi.org.uk/files/odi-assets/events-documents/5161.pdf.

Lyman, P., and H. R. Varian. 2000. "How Much Information?" http://www2sims.berkeley.edu/research/projects/how-much-info/summary.html.

MacFeely, S. 2016. "The Continuing Evolution of Official Statistics: Some Challenges and Opportunities," *Journal of Official Statistics* 32 (4).

Mayer-Schönberger, V., and K. Cukier. 2014. *A Revolution That Will Transform How We Live, Work and Think: Big Data.* John Murray Publishers, Great Britain.

Meyer, C., T. McGuire, M. Masri, and A. Wahab Shaikh. 2013. "Four Steps to Turn Big Data into Action." https://www.forbes.com/sites/mckinsey/2013/10/22/four-steps-to-turn-big-data-into-action/#5dcac9094380.

Meyer, B. D., W. K. C. Mok, and J. X. Sullivan. 2015. "Household Surveys in Crisis." *Journal of Economic Perspectives* 29 (4): 199–226.

Moreno, J., M. A., Serrano, and E. Fernández-Medina. 2016. "Main Issue in Big Data Security." http://www.mdpi.com/1999-5903/8/3/44/pdf.

Nathan, M., and A. Rosso. 2013. "Measuring the UK's Digital Economy with Big Data." NIESR, July. http://www.niesr.ac.uk/publications/measuring-uk%E2%80%99s-digital-economy-big-data#.WKI5xGeQyos.

Oostrom, L., A. Walker, B. Staats, M. Slootbeek-Van Laar, S. Ortega Azurdy, and B. Rooijakkers. 2016. "Measuring the Internet Economy in the Netherlands: A Big Data Analysis." CBS Discussion Paper 2016/14. https://www.cbs.nl/-/media/_pdf/2016/40/measuring-the-internet-economy.pdf.

Organisation for Economic Cooperation and Development (OECD). 1987. https://search.oecd.org/eco/outlook/35252065.pdf.

———. 2015. *OECD Digital Economy Outlook 2015*. Paris. http://dx.doi.org/10.1787/9789264232440-en.

Overseas Development Institute (ODI). 2015. "What Is the Future of Official Statistics in the Big Data Era?" Public panel discussion. https://www.odi.org/events/4068-what-future-official-statistics-big-data-era.

Pavlicek, J., and L. Kristoufem. 2015. "Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries." PLOS One, May 22.

Piatetsky-Shapiro, G. 2012. "Managing Uncertainty: Big Data Hype (and Reality)." *Harvard Business Review,* October 18. https://hbr.org/2012/10/big-data-hype-and-reality.

Predata. 2016. http://www.predata.com.

Smith, M., C. Szongott, B. Henne, and G. Von Voight. 2012. "Big Data Privacy Issues in Public Social Media." In *Digital Ecosystems Technologies (DEST)*, 6th IEEE International Conference, 1–6.

Sy, A., and T. Wang. 2016. "De-risking, renminbi, internationalization, and regional integration." Africa Growth Initiative at the Brookings Institution.

United Nations Economic Commission for Europe (UNECE). 2013. *Classification of Types of Big Data*. http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data.

———. 2014. *A Suggested Framework for the Quality of Big Data*. http://www1.unece.org/stat/platform/download/attachments/108102944/Big%20Data%20Quality%20Framework%20-%20final-%20Jan08-2015.pdf?version=1&modificationDate=1420725063663&api=v2.

United Nations Global Pulse (UNGP). 2012. "Big Data for Development: Challenges and Opportunities."

———. 2014. Nowcasting Food Prices in Indonesia Using Social Media Signals (2014) http://www.unglobalpulse.org/nowcasting-food-prices.

United Nations Global Working Group (UNGWG). 2017. *Big Data*. http://unstats.un.org/bigdata.

———. Survey and Project Inventory, https://unstats.un.org/bigdata/inventory.

United Nations Statistics Division (UNSD). 2014. *Fundamental Principles of Official Statistics*. https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx.

———. 2015. *Report of the 2015 Big Data Survey*. https://unstats.un.org/unsd/statcom/47th-session/documents/BG-2016-6-Report-of-the-2015-Big-Data-Survey-E.pdf.

United Nations World Data Forum (UNWDF). 2017. *A series of presentations by Statistics Norway, Statistics Denmark, Statistics Sweden and Statistics Finland.* www.undataforum.org.

World Bank. *Big Data in Action for Development. Washington, DC.* http://live.worldbank.org/sites/default/files/Big%20Data%20for%20Development%20Report_final%20version.pdf.

———. 2016. *Delivering on Big Data.* Washington, DC.

Appendix I. Classification Developed by the UNECE Task Team on Big Data, (UNECE Wiki June 2013)[25]

1. Social Networks (human-sourced information): This information is the record of human experiences, previously recorded in books and works of art and later in photographs, audio, and video. Human-sourced information is now almost entirely digitized and stored everywhere from personal computers to social networks. Data are loosely structured and often ungoverned.
1100. Social Networks: Facebook, Twitter, Tumblr etc.
1200. Blogs and comments
1300. Personal documents
1400. Pictures: Instagram, Flickr, Picasa, etc.
1500. Videos: YouTube etc.
1600. Internet searches
1700. Mobile data content: text messages
1800. User-generated maps
1900. E-mail

2. Traditional Business Systems (process-mediated data): These processes record and monitor business events of interest, such as registering a customer, manufacturing a product, taking an order, etc. The process-mediated data thus collected are highly structured and include transactions, reference tables, and relationships, as well as the metadata that set its context. Traditional business data are the vast majority of what IT managed and processed, in both operational and business intelligence systems—usually structured and stored in relational database systems. (Some sources belonging to this class may fall into the category of "Administrative data.")

21. Data Produced by Public Agencies
2110. Medical records

22. Data Produced by Businesses
2210. Commercial transactions
2220. Banking/stock records
2230. E-commerce
2240. Credit cards

3. Internet of Things (machine-generated data): This information is derived from the phenomenal growth in the number of sensors and machines used to measure and record the events and situations in the physical world. The output of these sensors is machine-generated data, and from simple sensor records to complex computer logs, it is well structured. As sensors proliferate and data volumes grow, it is becoming an increasingly important component of the

[25] UNECE 2013.

information stored and processed by many businesses. Its well-structured nature is suitable for computer processing, but its size and speed are beyond traditional approaches.

31. Data from Sensors

- 311. Fixed sensors
  - 3111. Home automation
  - 3112. Weather/pollution sensors
  - 3113. Traffic sensors/webcam
  - 3114. Scientific sensors
  - 3115. Security/surveillance videos/images
- 312. Mobile sensors (tracking)
  - 3121. Mobile phone location
  - 3122. Cars
  - 3123. Satellite images

32. Data from Computer Systems

- 3210. Logs
- 3220. Web logs

Appendix II. Table 1 Linking Big Data and Statistical Domains

Data Source+	Data Type	Statistical Domains	Additional Statistical Domains*
Social Networks	1100. Social Networks: Facebook, Twitter, LinkedIn 1200. Blogs and comments 1600. Internet searches on search engines (Google) 1700. Mobile data content: text messages, Call Detail Record, Data Detail Record, Location update, Radio coverage updates Online news	National accounts External sector statistics Financial statistics Price statistics Government finance statistics (public debt statistics)	Sentiment indices (investor, consumer) Social statistics Labor statistics Migration statistics Tourism statistics Population statistics Household consumption statistics Sustainable Development Goals indicators Early-warning indicators Transportation statistics Urban statistics
Traditional Business Systems	**Data produced by public agencies** Administrative data	Government finance statistics National accounts Price statistics External sector statistics	Sustainable Development Goals indicators
	Data produced by businesses 2210. Commercial transactions 2220. Banking/stock records 2230. E-commerce 2240. Credit cards Business websites Scanner data	National accounts Price statistics External sector statistics Financial statistics	Social statistics Business registers Employment statistics Household consumption statistics Transport statistics Sustainable Development Goals indicators
Internet of Things (machine-generated data)	**Data from sensors** **311. Fixed sensors** 3111. Home automation 3112. Weather/pollution sensors 3113. Traffic sensors/webcam 3114. Scientific sensors **312. Mobile sensors (tracking)** 3121. Mobile phone location 3122. Cars 3123. Satellite images	National accounts Satellite accounts External sector statistics Government finance statistics Price statistics	Traffic/transport statistics Energy statistics Land use statistics Agricultural statistics Environment statistics Transport and emission statistics Air emission statistics Sustainable Development Goals indicators
Based on adapted UN big data classification+			*Based on European Statistical System Committee (2014)

Appendix III. Table 2: Current Applications of Big Data in Macroeconomic and Financial Statistics

Data Origin+	Data Type	Data Source and Techniques	Potential Indicators Derived	Statistical Domains	What May Be the Potential?*
Social Networks	Social networks, blogs and comments 1100. Social Networks: Facebook, Twitter, LinkedIn 1200. Blogs and comments 1600. Internet searches on search engines (Google) 1700. Mobile data content: text messages, Call Detail Record, Data Detail Record, Location update, Radio coverage updates Online news	**Google** trends and search data	nowcast GDP nowcast unemployment consumer sentiment car and property sales	National accounts External sector statistics Financial statistics price statistics	**2**
		Mobile phone system data (electronic money schemes, e.g., M-Pesa) Peer-to-peer transactions	financial inclusion indicators remittances, regional disposable income, consumption patterns poverty reduction SDG "Gender Equality" economic growth	National accounts External sector statistics Financial statistics Price statistics	**1,3**
		Twitter tweets	consumer confidence index border mobility, tourism, transitioning of migrants nowcast food prices sentiment and topic trend analysis	Mobility and urban statistics Price statistics Demographic and social statistics	**1,2,3**
		Web scraping of **Facebook posts, Wikipedia articles**	geopolitical risk indicators price changes civil protests/labor strikes and national security events consumer sentiment inclusive infrastructure for sustainable development	Price statistics National accounts Demographic and social statistics Labor statistics	**1,2**
		Call Detail Record data	SDG indicators, travel/tourism, transport, migration	Mobility and urban statistics	**1,3**
Traditional Business Systems	**Data produced by public agencies** Administrative data	**Taxation registers**	consumer spending small business income nonresident businesses controlled by resident parent corporations business profiling flight reservation system	National accounts Price statistics External sector statistics Labor statistics Tourism statistics Transportation statistics	**2, 3**
		Population/business registers	multisourcing to derive population and housing census population structure	National accounts Demographic and social statistics	**3**
	Data produced by businesses 2210. Commercial transactions	**SWIFT data** on transaction quantities and financial market prices	global financial flows network concentration cross-border transactions export/import indicators	National accounts Price statistics External sector statistics Financial statistics	**2,3**

Data Origin+	Data Type	Data Source and Techniques	Potential Indicators Derived	Statistical Domains	What May Be the Potential?*
	2220. Banking/stock records 2230. E-commerce 2240. Credit cards Business websites Scanner data		withdrawal of correspondent banking relationships trade financing		
		Web scraping to collect **price data** from **online retailers**	daily inflation turning points in inflation trends e-commerce index	Price statistics Financial statistics	**2,3**
		Web scraping **business websites**	enterprise profiling job vacancies	National accounts Financial statistics Labor statistics	**2,3**
		Scanner data Prices and quantities	national and regional consumer prices household income and expenditure	Price statistics National accounts Financial statistics	**2,3**
		Credit card data	consumer spending growth trends of retail sales	National accounts External sector statistics	
Internet of Things (machine-generated data)	**Data from sensors** **311. Fixed sensors** 3111. Home automation 3112. Weather/pollution sensors 3113. Traffic sensors/webcam 3114. Scientific sensors **312. Mobile sensors (tracking)** 3121. Mobile phone location 3122. Cars 3123. Satellite images	GPS **positioning/tracking data**	travel services exports/imports trip duration inbound/outbound international travelers remoteness index traffic intensity	National accounts External sector statistics Demographic statistics Transport statistics Urban statistics Tourism statistics Population statistics	**1,2,3**
		Traffic/road sensors	proxy of economic growth/health commuting time traffic intensity incoming/outgoing traffic travel/tourism	National accounts External sector statistics Transport statistics Tourism statistics Mobility statistics	**1,2,3**
		Satellite imagery Research and mapping of weather and climate data	improved geographical localization of statistical units and assets spatial sampling frame for output measurement land use and geostatistical cartography crop planting area, land use, and agricultural output population and asset location as proxy for SDG "Gender Equality"	National accounts Price statistics External sector statistics Demographic and social statistics Transport statistics Agricultural statistics Demographic and urban statistics	**1,3**
		Smart meters (energy consumption measures)	nonoccupancy rates household consumption electricity supply and consumption price differentials household structure and size	Environmental and energy statistics National accounts Price statistics Demographic and social statistics Transportation statistics Geospatial statistics Agricultural statistics Rural and population statistics	**1,2,3**

Data Origin+	Data Type	Data Source and Techniques	Potential Indicators Derived	Statistical Domains	What May Be the Potential?*
Based on adapted UN big data classification+		*1. Big data to answer "new questions" and produce new indicators 2. Big data to bridge time lags in the availability of official statistics and supporting the more timely forecasting of existing indicators 3. Big data as an innovative data source in the production of official statistics			

Appendix IV. Big Data and the Digital Economy

Big data is often described as the fuel of the digital economy. Data have become an important asset for the economy, comparable to human and financial resources. Most economic activities will rely on data within a few years, and this will provide opportunities to many economic sectors. The digital economy encompasses services delivered or intermediated through digital technology such as internet platforms and smartphones and information and communication technology hardware and software. Specialized service firms provide data storage, transfer, and mining services within and across borders, thus reducing transaction costs (OECD 2015).

Digitalization poses challenges for price and thus volume measures of GDP. Digitization makes price comparisons that control for quality differences hard. There is also a large variety of pricing models, especially for internet and mobile services. Moreover, digitalization has seen the creation of new business models, which may raise the quality of services offered and lead to switching between products. This is also common for software and information and communication technology products for which the concept of quality versus price change has been fast and challenging to distinguish.

Big data can help measure the digital economy. Big data can provide new insight into digital transactions. Statistical agencies can directly collect information from new digital intermediaries (such as eBay) on transactions and digital products from the internet via web scraping. The new data sourced in this way allow for higher-frequency collection and better control for quality and thus price changes.

These two examples show how big data can better quantify the digital sector:

A National Institute of Economic Social Research (NIESR) paper using growth intelligence (GI) data found that the digital economy in the United Kingdom was substantially larger than previous estimates (Nathan and Rosso 2013). GI used data from the web, social media, news feeds, patents, and other sources and layered these data on top of public data from the UK registrar of companies, Companies House.

Another study combined big data with regular statistics to study the size of the Dutch internet economy (Oostrom and others 2016). Statistics Netherlands, Google, and Dataprovider collaborated on a study that links businesses' website data to statistical information on the business through a complex matching algorithm. It found that the core of the internet economy is roughly the same size as the construction sector.