Chapter 5. The role of big data for ICT monitoring and for development

5.1 Introduction

One of the key challenges in measuring the information society has been the lack of up-to-date and reliable data, in particular from developing countries. The information and communication technologies (ICT) sector is evolving rapidly, as are the types of service and application that are driving the information society, all of which makes identifying and tracking new trends even more challenging. As the key global source for internationally comparable ICT statistics, ITU is continuously working to improve the availability and quality of those statistics and identify new data sources. In this context, the emergence of big data holds great promise, and there is an opportunity to explore their use in order to complement the existing, but often limited, ICT data.

There is no unique definition of the relatively new phenomenon known as big data. At the most basic level it is understood as being data sets whose volume, velocity or variety is very high compared to the kinds of datasets that have traditionally been used. The emergence of big data is closely linked to advances in ICTs. In today's hyper-connected digital world, people and things leave digital footprints in many different forms and through ever-increasing data flows originating from, among other things, commercial transactions, private and public records that companies and governments collect and store about their clients and citizens, user-generated online content such as photos, videos, tweets and other messages, but also traces left by the Internet of Things (IoT), i.e. by those uniquely identifiable objects that can be tracked.

Big data have great potential to help produce new and insightful information, and there is a growing debate on how businesses, governments and citizens can maximize the benefits of big data. Although it was the private sector that first used big data to enhance efficiency and increase revenues, the practice has expanded to the global statistical community. The United Nations Statistical Commission (UNSC) and national statistical organizations (NSOs) are looking into ways of using big data sources to complement official statistics and better meet their objectives for providing timely and accurate evidence for policy-making.[1]

So far, there is limited evidence as to the value added by big data in the context of monitoring of the information society, and

there is a need to explore its potential as a new data source. While existing data can provide a relatively accurate picture of the spread of telecommunication networks and services, there are significant data gaps when it comes to understanding the development of the information society. Relatively little information, for example, is available on the demand side. While an increasing number of countries currently collect data on the individual use of ICTs, many developing countries do not produce such information (collected through household surveys or national population and housing censuses) on a regular basis. Consequently, not enough data are available about the types of activity that the Internet is used for, and little is known about the Internet user in terms of age, gender, educational or income level, and so on.

In other areas, such as education, health or public services, even fewer data are available to show developments over time and enable informed policy decisions. The recently published *Final WSIS Targets Review* report (Partnership, 2014), which attempts to assess developments in the information society between 2003 and 2013/14, shows that little information is available to track progress over time. It is obvious that greater efforts must be made to overcome the lack of reliable, timely and relevant statistics on the information society, and that big data have the potential to help realize those efforts.

In addition to the data produced and held by telecommunication operators, the broader ICT sector, which includes not just telecommunication companies but also over-the-top (OTT) service providers such as Google, Twitter, Facebook, WhatsApp, Netflix, Amazon and many others, captures a wide array of behavioural data. Together, these data sources hold great promise for ICT monitoring, and this chapter will explore the potential of today's hyper-connected digital world to expand on existing access and infrastructure indicators and move towards indicators on use, quality and equality of use.

At the same time, there is a growing debate on the role and potential of big data when it comes to providing new insights for broader social and economic development. Big data are already being leveraged to understand socio-economic well-being, forecast unemployment and analyse societal ties. Big data from the ICT industry play a particularly important role because they are the only stream of big data with global socio-economic coverage. In particular, mobile telephone access is quasi-ubiquitous, and ITU estimates that by the end of 2014 the number of global mobile subscriptions will be approaching 7 billion. At the same time, almost 3 billion people – 40 per cent of the world's population – will be using the Internet. In recent years, moreover, the strongest growth in telecommunication access and use has been recorded in the developing economies, where ICT penetration levels have increased and where big data hold great promise for development. However, while there are a growing number of research collaborations and promising proof-of-concept studies, no significant project has yet been brought to a replicable scale in the development sphere. Future efforts will have to overcome a number of barriers, including the development of models to protect user privacy while at the same time allowing for the extraction of insights that can improve service delivery to low-income populations. To this end, this chapter will contribute to the debate on big data for development, highlight advances, point to some best practices and identify challenges, including in regard to the production and sharing of big data for development.

The chapter will first (in Section 5.2) describe some of the current big data trends and definitions, highlight the technological developments that have facilitated the emergence of big data, and identify the main sources and uses of big data, including the use of big data for development and ICT monitoring. Section 5.3 will examine the range and type of data that telecommunication companies, in particular mobile-cellular operators, produce, and how those data are

currently being used to track ICT developments and improve their business. Section 5.4 looks at the ways in which telecom big data may be used to complement official ICT statistics and assist in the provision of new evidence for a host of policy domains, while Section 5.5 discusses the challenges of leveraging big data for ICT monitoring and broader development, including in terms of standardization and privacy. It will also make some recommendations for mainstreaming and fully exploiting telecom big data for monitoring and for social and economic development, in particular with regard to the different stakeholders involved in the area of big data from the ICT industry.

5.2 Big data sources, trends and analytics

With the origins of the term "big data" being shared between academic circles, industry and the media, the term itself is amorphous, with no single definition (Ward and Barker, 2013). At the most basic level of understanding, it usually refers to large and complex datasets, and reflects advances in technology that make it possible to capture, store and process increasing amounts of data from different data sources. Indeed, one of the key trends fostering the emergence of big data is the massive "datafication" and digitization, including of human activity, into digital "breadcrumbs" or "footprints".

In an increasingly digitized world, big data are generated in digital form from a number of sources. They include administrative records (for example, bank or electronic medical records), commercial transactions between two entities (such as online purchases or credit card transactions), sensors and tracking devices (for example, mobile phones and GPS devices), and activities carried out by users on the Internet (including searches and social media content) (Table 5.1).

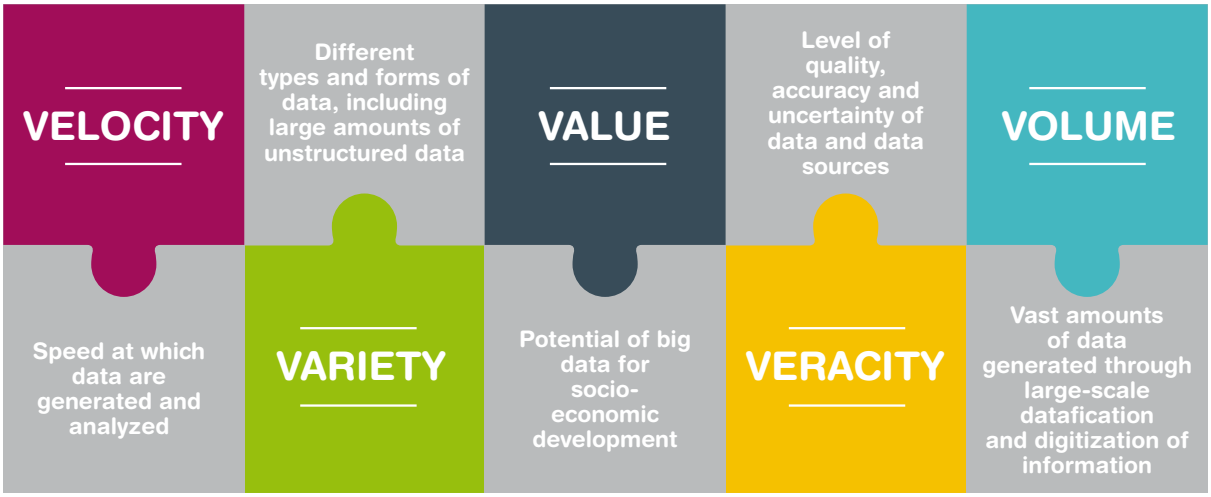
Big data is not just about the volume of the data. One of the earliest definitions, introduced by the Gartner consultancy firm, describes big data characteristics such as velocity and variety, in addition to volume (Laney, 2001). "Velocity" refers to the speed at which data are generated, assessed and analysed, while the

Table 5.1: Sources of big data

Sources	Some examples
Administrative data	• Electronic medical records • Insurance records • Tax records
Commercial transactions	• Bank transactions (inter-bank as well as personal) • Credit card transactions • Supermarket purchases • Online purchases
Sensors and tracking devices	• Road and traffic sensors • Climate sensors • Equipment and infrastructure sensors • Mobile phones • Satellite/GPS devices
Online activities/social media	• Online search activities • Online page views • Blogs and posts and other authored and unauthored online content and social media activities • Audio/images/videos

Source: ITU, adapted from UNSC (2013).

Figure 5.1: The five Vs of big data



Source: ITU.

term "variety" encompasses the fact that data can exist as different media (text, audio and video) and come in different formats (structured and unstructured). The three-Vs definition has caught on and been expanded upon. A fourth V – veracity – was introduced to capture aspects relating to data quality and provenance, and the uncertainty that may exist in their analysis (IBM, 2013). A fifth V – value – is included by some to acknowledge the potentially high socio-economic value that may be generated by big data (Jones, 2012) (Figure 5.1).

Included within the scope of big data is the category of transaction-generated data (TGD),[2] also sometimes described as "data exhaust" or "trace data". These are digital records or traces that have been generated as by-products of doing things (such as processing payments, making a phone call and so on) that leave behind bits of information. The value of this subset of big data is that it is directly connected to human behaviour and its accuracy is generally high. Most of the data captured by telecommunication companies can be classified as TGD.

As is often the case with technological innovation, it is the private sector that has been at the forefront of extracting value from this data deluge. Encouraged by promising results but also reduced budgets, the public sector is turning towards big data to improve its service delivery and increase operational efficiency. In addition, there are uses for big data in broader development and monitoring, and there is an increasing focus on big data's role in producing timely (even real-time) information, as well as new insights that can be used to drive social and economic well-being.

Big data uses by the private and public sectors

Marketing professionals, whose constant aim is to understand their customers, are now increasingly shifting from conventional methods, such as surveys, to the extraction of customer preferences from the analysis of big data. Walmart, the world's biggest retailer, has been one of the largest and earliest users of big data. In 2004, it discovered that the snack food known as Pop Tarts was heavily purchased by United States citizens preparing for serious weather events such as hurricanes. The correlation analysis revealed a behaviour associated with

a specific condition that then led Walmart to improve its production chain – in this case, by increasing the supply of Pop Tarts to areas likely to be affected by a disaster. Walmart has also made use of predictive analytics, which uses personal information and purchasing patterns to extrapolate to a likely future behaviour, and to better target and address customer needs. Together, large-scale automated correlation analysis and predictive analytics are two of the key techniques that have helped unleash the value of big data.

Nor is the private sector's use of big data techniques restricted solely to market research. Companies and whole industries (healthcare, energy and utilities, transport, etc.) are using such techniques to optimize supply chains and production (see Box 5.1 for an example from the energy industry). New value is extracted by being able to link new information on customers to the production process in a way that enables companies to tailor and segment their products at low cost. Firms that are highly proficient in their use of data-driven decision making have been found to have productivity levels up to 6 per cent higher than firms making minimal to no use of data for decision-making (Brynjolfsson, Hitt and Kim, 2011). Significantly, industries now have the ability to conduct controlled experiments at a scale and with a speed that are unprecedented. Google, for example, is running about a thousand experiments at any given point in time (Varian, 2013a). Telecom network operators make extensive use of such techniques when rolling out new services, among other things for the purpose of pricing. Telecom operators also use big data techniques to understand and control churn, optimize their management of customer relations and manage their network quality and performance.

These fundamental shifts in data exploitation to generate new socio-economic value, coupled with the simultaneous emergence of new rich data sources that can potentially be linked together and analysed with ease, have also sparked the interest of governments, researchers and development agencies. Encouraged by the potential of big data to produce new insights and slimmer budgets, governments (at all levels) are now looking to exploit big data and increase the application of data analytics to a range of activities, including monitoring and improvement of tax compliance and revenues, crime detection and prediction, and improvement of public service delivery (Giles, 2012; Lazer et al., 2009).

To this end, governments, in addition to the data they collect and generate themselves,

Box 5.1: How big data saves energy – Vestas Wind Systems improves turbine performance

Vestas, a global energy company dedicated to wind energy, with installations in over 70 countries, has used big data platforms to improve the modelling of wind energy production and identify the optimal placement for turbines.

Wind turbines represent a major investment and have a typical lifespan of 20 to 30 years. To determine the optimal placement for a turbine, a large number of location-dependent factors must be considered, including temperature, precipitation, wind velocity, humidity and atmospheric pressure. By using big data techniques based on a large set of factors and an extended set of structured and unstructured data, Vestas was able to significantly improve customer turbine location models and optimize turbine performance.

Big data have enabled the creation of a new information environment and allowed the company to manage and analyse weather and location data in ways that were previously not possible. These new insights have led to improved decisions relating not only to wind turbine placement and operation, but also to more accurate power-production forecasts, not to mention greater business-case certainty, speedier results, and increased predictability and reliability. This reduces the cost to customers per kilowatt-hour produced, while increasing the accuracy of the customer's return-on-investment estimates.

Source: ITU, based on IBM (2012).

complement their official statistics by leveraging data from new sources, including crowd-sourced data generated by the public. In the United States, for example, Boston City Hall released the mobile app "Street Bump", which uses a phone's accelerometer to detect potholes while the app user is driving around Boston and notifies City Hall.[3] Some of the richest data sources for enabling governments and development agencies to improve service delivery are actually external. Such external data include those captured and/or collected by the private sector, as well as the digital breadcrumbs left behind by citizens as they go about their daily lives.

According to a recently published White House report, United States government agencies can make use of public and private databases and big data analytics to improve public administration, from land management to the administration of benefits. The Department of the Treasury has set up a "Do Not Pay" portal, which links various databases and identifies ineligible recipients to avoid wrong payments and reduce waste and fraud[4] (The White House, 2014).

Big data for development and ICT monitoring

One of the richest sources of big data is the data captured by the use of ICTs. This broadly includes data captured directly by telecommunication operators as well as by Internet companies and by content providers such as Google, Facebook, Twitter, etc. Big data from the ICT services industry are already helping to produce large-scale development insights of relevance to public policy. Collectively, they can provide rich and potentially real-time insights to a host of policy domains. It should be noted that in some countries and regions the use of big data, including big data from the ICT industry, is subject to national regulation. In the EU, for example, a number of directives require data producers to obtain users' consent before gathering any of their personal data.[5]

One of the best-known examples of leveraging the online population's digital breadcrumbs for development purposes is Google Flu Trends (GFT). Following its launch in 2008, GFT was remarkably accurate in tracking the spread of influenza in the United States, doing so more rapidly than the Centers for Disease Control and Prevention (CDC), with a lag time of only one day as opposed to one week. Although it has since been subject to criticism (see Section 5.5), GFT was held up as an outstanding example of big data in action and of the great potential of big data for broader development and monitoring (Mayer-Schönberger and Cukier, 2013; McAfee and Brynjolfsson, 2012). GFT worked by monitoring health-seeking behaviour expressed through online searches, with the search terms being correlated wherever they related to flu-like symptoms (Ginsberg et al., 2009). This proved to be so successful that it spawned similar efforts focusing on the use of search-engine data to understand dengue fever outbreaks,[6] monitor prescription drug use (Simmering, Polgreen and Polgreen, 2014), predict unemployment claims in the United States (Choi and Varian, 2009) and Germany (Askitas and Zimmermann, 2009), and forecast near-term values for economic indicators such as car and home sales and international visitor statistics (Choi and Varian, 2012).

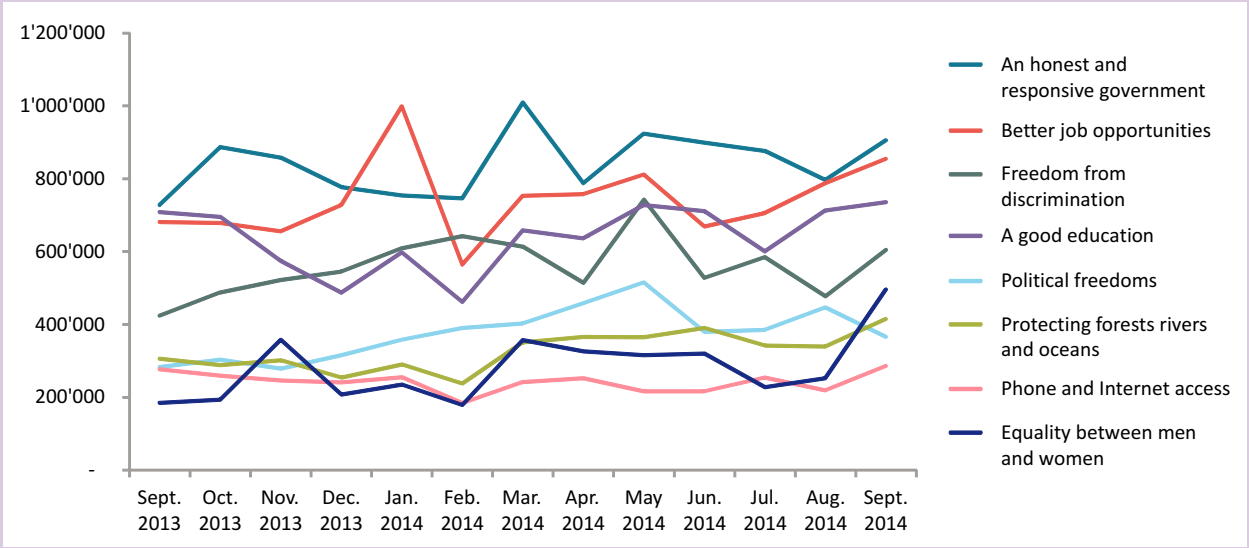
The Internet has also been a rich source of big data beyond the realm of user search terms. Online job-posting data are being used to supplement traditional labour statistics in the United States[7] and other countries. In another effort, an academic project at MIT known as the Billion Prices Project collects high-frequency price data from hundreds of online retailers.[8] The data are then used by researchers to understand a whole host of macroeconomic questions relating to, among other things, pricing behaviour, daily inflation and asset-price movements. This has the advantage of providing near real-time inflation statistics that are traditionally published monthly.

Box 5.2: How Twitter helps understand key post-2015 development concerns

As the process of formulating the post-2015 development agenda continues, UN Global Pulse and the Millennium Campaign are using big data and visual analytics to identify the most pressing development topics that people around the world are concerned about and consider a priority.

Their interactive visualization tool shows the 16 topics that people have tweeted the most about. Users can select a country to see the number of tweets generated by its Twitter users in regard to the highlighted topic, as compared to tweets about all the other topics. This information provides insight as to which of the various post-2015 issues are being talked about the most. In September 2014, at the global ("all countries") level, *An honest and responsive government* was the key priority, followed by *Better job opportunities* and *Freedom from discrimination*. Also highly ranked, in 7th position, was phone and Internet access. By clicking on any of the data points in the chart, the application provides information on the number of tweets (per month) for each topic. It also lists the top words that those tweets contained.

Chart Box 5.2: Using Twitter to visualize trends in global development topics



Source: UN Global Pulse, see: http://post2015.unglobalpulse.net/#.

UN Global Pulse, a UN initiative to use big data for sustainable development and humanitarian action, has been mining Twitter data from Indonesia (where Twitter usage is high)[9] to understand food price crises. Global Pulse was able to identify a consistent pattern among specific food-related tweets and the daily food price index. In fact, it was able to use predictive analytics on the Twitter data to forecast the consumer price index several weeks in advance (Byrne, 2013). As discussions on the post-2015 development agenda continue, UN Global Pulse is also using Twitter data to understand and compare the relevance of different development topics among countries (Box 5.2).

In fact, the ICT sector is itself using the Internet as a source of big data for monitoring purposes. Regulators and others are now using the Internet to crowd-source quality of service (QoS) data on broadband quality. For example, the United States Federal Communications Commission (FCC) has released mobile apps that enable consumers to check their broadband quality. The test results, which are anonymous, are then used by FCC to understand and address coverage and quality issues in different areas.[10]

Mobile data

Despite the rapid growth in Internet access, 60 per cent of the world's population is still not using the Internet. Household Internet penetration in developing economies is expected to reach 31 per cent by the end of 2014, as against almost 80 per cent in developed economies. In addition, as Internet penetration rates remain limited, Internet users are not (yet) representative of the population at large. For example, Internet users tend to be younger, relatively well educated, with men still more likely to be online than women, especially in developing countries[11] (ITU, 2013).

Depending on the source of Internet data, results may also be more or less biased. A 2013 study into the characteristics and behaviour of Facebook users, for example, revealed that while in many ways Facebook users have real-life behaviour and characteristics, in many ways the social network fails as a representation of society. On the one hand, for example, the American Facebook user's relationship status of "married" on Facebook is very similar to real life (census) data on the average age when American people get married. On the other hand, however, the average American Facebook user is much younger than the average citizen.[12] This is just one example, but it highlights the need to take account of particular characteristics and the limitations of producing representative results when extracting information from online users' behaviour.

Given the popularity of mobile-cellular services, non-Internet-related mobile-network big data seems to have the widest socioeconomic coverage in the near term, and the greatest potential to produce relatively representative information globally, particularly in developing countries. By the end of 2014, the number of mobile-cellular subscriptions is expected to be nearing 7 billion, and the number of mobile-cellular subscriptions per 100 inhabitants is expected to reach 90 per cent. Mobile data are already being utilized for research and policy-making, not only in developed but also in developing economies.

There are various examples of how mobile phone records have been used to identify socio-economic patterns and migration patterns, describe local, national and international societal ties, and forecast economic developments.[13] Data are also being used to improve responsiveness in the event of natural disasters or disease outbreaks. Lu, Bengtsson and Holme (2012) used mobile call records to study the population displacements following Haiti's 2010 earthquake, with a view to using such methods to improve the effectiveness of humanitarian relief operations immediately after a disaster. Call records have also been merged with epidemiological data to understand the spread of malaria in Kenya (Pindolia et al., 2012; Wesolowski et al., 2012a), and of cholera in Haiti after the 2010 earthquake (Bengtsson, Lu, Thorson, Garfield and von Schreeb, 2011) and in Côte d'Ivoire (Azman, Urquhart, Zaitchik and Lessler, 2013). Mobile network big data have been utilized to great effect in the area of transportation, helping to measure and model people's movements (even in real time) and understand traffic flows (Wu et al., 2013).

It is evident from the examples given that big data from the ICT sector, and especially those available to telecommunication operators, have wide applicability for informing multiple public policy domains. Leveraging such data to complement official statistics and facilitate broader development will enable governments as well as development agencies to better serve their citizens and beneficiaries. Less use has thus far been made of telecommunication big data with a view to understanding its potential for producing additional information and statistics on the information society. In assessing that potential, including the potential for providing complementary

information on the development of the information society, it is first important to better understand the type of data that can be made available.

5.3 Telecommunication data and their potential for big data analytics

Fixed and mobile telecommunication network operators, including Internet service providers (ISPs), are an important source of data and for the purpose of this chapter, all forms of telecommunication big data (either volume, velocity or variety) are being considered. Most telecommunication data can be considered as TGD,[14] that is, the result of an action undertaken (such as making a call, sending an SMS, accessing the Internet or recharging a prepaid card).

Since the service with the widest coverage and greatest uptake and popularity is the mobile-cellular service, data from mobile operators have the greatest potential to produce representative results and reveal developmental insights on the population, including in developing countries and, increasingly, low-income areas. Not surprisingly, the big data for development initiatives (outlined in Section 2.2) have mainly drawn on mobile-network big data rather than on those from fixed-telephone operators or ISPs. Figure 5.2 illustrates some of the similarities and differences in the type of information that mobile-network operators, as opposed to fixed-telephone operators and ISPs, produce, and shows some of the additional insights, in particular in terms of the location and mobility information that mobile networks and services generate.

Telecommunication data

The mobile telecommunication data that operators possess can be classified into different types, depending on the nature of the information they produce. They include traffic data, service access detail records, location and movement data, device characteristics, customer details and tariff data. For a more detailed overview of these types of data, see Chapter 5 Annex.

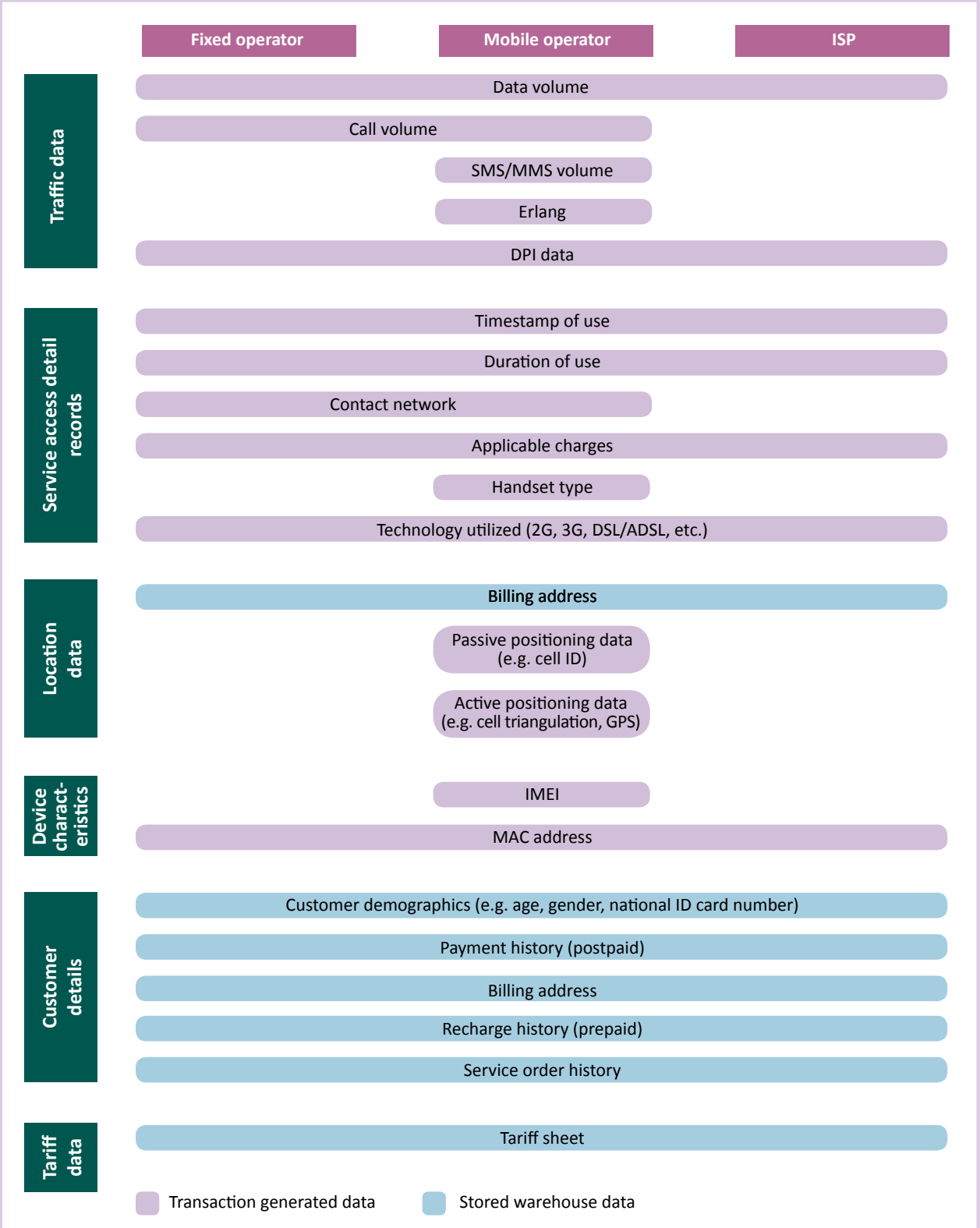
To collect **traffic data**, operators use a range of metrics to understand and manage the traffic flowing through their networks, including the measurement of Internet data volumes, call, SMS and MMS volumes, and value-added service (VAS) volumes. Internet service providers can also use deep packet inspection (DPI),[15] which is a special process for scanning data packages transiting the network.

Service access detail records, including call detail records (CDRs), are collected by operators whenever clients use a service. They are used to manage the infrastructure and for billing purposes, and include information on the time and duration of services used and the technology used, for example, for the mobile network (2G, 3G, etc.). These data are potentially also very useful for building a rich profile of customers, as outlined in this section.

Mobile networks capture a range of **movement and location variables** to identify user location and movement patterns. The degree of accuracy of this information depends on a number of factors, including the network used and device generation, and can be broadly classified into two different types: passive and active positioning data, with the latter providing more detailed and precise location information.

Since mobile user devices used to access mobile telecommunication services come with an international mobile station equipment identity (IMEI) number, operators can identify some **device characteristics**, including the handset make and model and type of technology (2G, 3G, LTE) employed. Mobile network operators can use the IMEI number to identify the specific mobile handset being used by a subscriber, which in turn can provide some insight as to that

Figure 5.2: An overview of telecom network data



Source: ITU, adapted from Naef et al. (2014).

Box 5.3: How mobile operators currently use data to track service uptake, business performance and revenues

Operators use their TGD to monitor the uptake and penetration of particular services, identify market shares and monitor their business performance, as well as for reporting purposes. They can also track the extent to which different technologies are used, not just in terms of handset capabilities but also actual usage, enabling them, for example, to determine the number of active mobile-cellular and active mobile-broadband subscriptions.

On the basis of the detailed service-usage data collected, telecommunication operators can produce a range of detailed indicators relating to service consumption. For each customer, it is possible to determine the minutes of use (MoU), number of originating and terminating calls, SMS and MMS usage, data upload volumes, data download volumes, level of use of different VAS, and level of use of different OTT services. These data can be reported as averages (over time or for different categories of user), as well as at various levels of aggregation (again over time or for different categories of user). These measures are often key performance indicators (KPIs), tracked and used in particular by operators, but also by regulators and at the international level.

Finally, service consumption data are used to produce revenue data and projections at various levels of disaggregation or aggregation. For example, the average revenue per user (ARPU) is a KPI for operators, which identify their most important customers on the basis of the revenue they generate for the company. Similarly, revenue projections are made not just at the level of a particular service,[16] but also to identify the most important network elements. For example, mobile operators collect indicators on the revenue being generated at the base station level, often in real time. Collectively, these revenue-based metrics can also be used by the operator to ensure a higher QoS and higher bandwidth at those locations generating the most revenue. Mobile operators will, for instance, often associate revenue data with resource allocation to ensure that QoS at the base stations used by their premium customers is maintained at the highest possible level.

Source: ITU.

subscriber's purchasing power (see below for more details).

In addition, telecommunication operators hold various **customer details** that were captured during the customer registration process. These can include the customer's name, age, gender, billing address and, in some cases, national identity card number. Customer details may also include a history of the services accessed, service option preferences as well as other details (as referred to in Chapter 5 Annex).

Finally, operators maintain **tariff data** in the form of billing records for their current and past services, from which information on a customer's usage patterns and preferences can be extracted.

The information outlined above is used at the aggregate level to derive a range of indicators to provide operators with information on the uptake of different services and on their business performance and revenues (Box 5.3). The information is also an important tool for regulatory authorities and policy-makers as they evaluate existing policies and establish new ones. At the international level, organizations such as ITU, but also consultancy firms and others, use aggregated revenue data to track and benchmark countries' ICT developments, monitor the evolution of the information society and identify digital divides.

The telecom industry's use of big data

Telecommunication companies are actively seeking to intensify their use of big data analytics in order to improve existing services and create new ones. For operators, big data open up opportunities for better understanding of their customers, which in turn leads to improved sales and marketing opportunities. At the same time, big data can help optimize network operations and create new revenue streams and business lines, for example when selling data.

Customer profiling

Telecom operators capture a range of behavioural data about their customers.

Customer profiles include details about customers' mobility patterns, social networks and consumption preferences. Collectively, these digital breadcrumbs enable operators to profile and segment their customers based on a variety of metrics (Figure 5.3). Depending on the country or region, there may be different privacy and data regulations governing the manner in which operators may keep and/or use such data. This being the case, the extent to which behavioural profiling is used by operators may vary greatly.

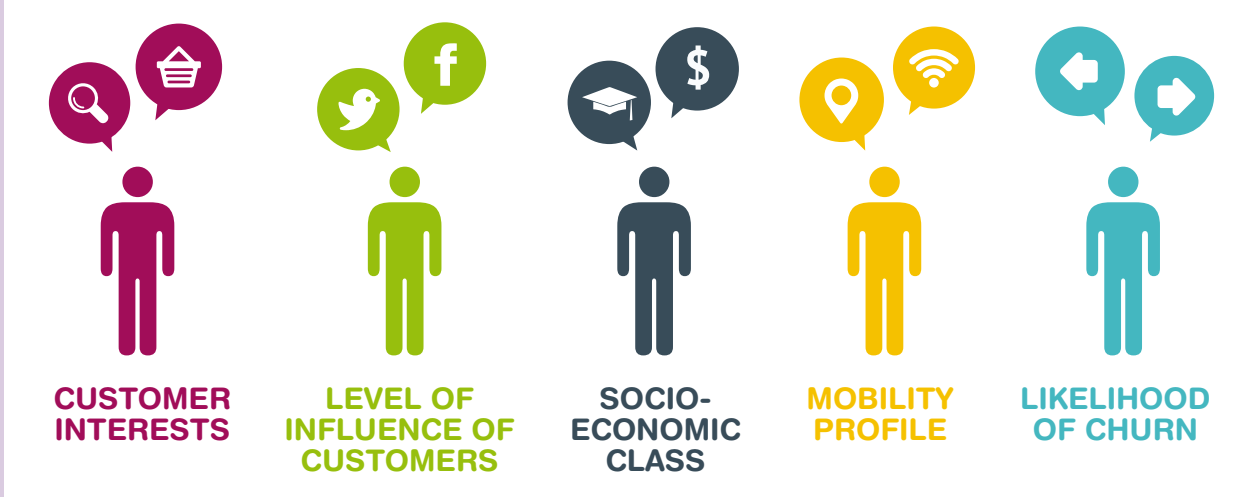
- Customer interests: these can be captured, or in some cases inferred, on the basis of usage levels (time spent and/or volume) for different VAS and OTT services. DPI can also be used to categorize interests based on sites visited (as opposed to content accessed). Sophisticated clickstream analyses from DPI data[17] can also generate more finely-grained interest classifications.

- Socio-economic class: While customer details will often enable operators to classify their customers' socio-economic status, such details are not always very reliable. Big data, on the other hand, can help to enhance that classification by enabling analysis of the levels of consumption of different services, including on the basis of spending (often in relation to other services), types of device used, frequency of change of handset, and so on.

- Likelihood of churn: The churn rate is a measure of the number of customers leaving the network or a particular service offered by an operator. Understanding churn is crucial to operators for obvious reasons. Big data techniques can help operators understand churn better by enabling them to model the likelihood of customers leaving the network (or opting out of a given service) by focusing on the customer's existing service usage behaviour.

- Level of influence of customers: Operators are keen to leverage service and technology diffusion among their subscribers with a view to marketing additional, customized services. This often calls for an understanding of the level of influence of each subscriber's social networks, both on-network (i.e. within the same operator) as well as off-network (i.e.

Figure 5.3: Customer profiling using telecom big data



Source: ITU.

in competitor networks). By identifying a large number of off-net users in a customer's network, operators may target the subscriber and/or the off-net users with promotions and incentives aimed at converting off-net connections into on-net users.

- Mobility profile: Mobile operators accord a high priority to identifying the locations most frequented by their customers, in order not only to ensure a high QoS in those areas, but also, more recently, to build mobility profiles of their customers that can be leveraged for location-based services.

Sophisticated customer profiling enables operators to personalize and market new services more effectively. For example, by understanding their customers' relationships to their social networks (and their relative importance within them), operators are able to model the diffusion of services and create targeted promotions. Furthermore, social network insights can be used by an operator to market its services to the off-network contacts that are connected to its customers, or to reduce churn rates. In the Republic of Korea, for example, SK Planet, a subsidiary of SK Telecom, uses big data to help its parent company to cut churn and generate new revenue, and has used data mining to achieve a fourfold improvement in churn forecasting. The operator found that customers planning to quit their current package tend to use specific search phrases, such as "data plan" or "operator benefits", at least three to seven days before taking action. When operators suspect that customers may be looking elsewhere, they may try to keep them by providing them with a tailored offer.[18]

Network planning and management

By analysing their networks in real time, operators can optimize routing and ensure QoS. The use of real-time DPI enables them to optimize traffic routes and details of traffic volumes, including the geospatial distribution of demand, and to plan and manage their networks more effectively through optimal resource allocation. By utilizing geospatial information about their most active customers and high-revenue regions, operators can adapt their resource allocation to ensure that more resources are channelled into active locations. This is an area of great significance to operators as they seek to understand the demands placed on their networks by the use of popular OTT services.

New business lines

To increase revenue streams, operators may also seek to monetize the data they hold. The simplest way of doing this is to sell (anonymized) data to third parties. The customer insights obtained through the analysis of usage data can also help create new business lines, either through innovation (e.g. new types of VAS) or by partnering with other businesses, including credit-scoring and related financial services. One example is the US-based big data startup Cignifi,[19] which obtains data from mobile operators and financial institutions to build credit profiles and evaluate customer creditworthiness (see Box 5.8). Cross-promotions with brick-and-mortar businesses are a potentially high-growth area in which the detailed mobility profiles available to operators are leveraged.

5.4 Big data from mobile telecommunications for development and for better monitoring

In 2013, the United Nations High-Level Panel of Eminent Persons on the Post-2015 Development Agenda called for a "data revolution" that draws on existing and new sources of data for the post-2015 development agenda (United Nations, 2013). In March 2014, the forty-fifth session of UNSC, the highest decision-making body for international statistical activities, presented a report on "big data and modernization of

statistical systems", and proposed the creation of a big data working group at the global level (UNSC, 2013).[20] Current uses of big data to complement official statistics are still exploratory, but there is a growing interest in this topic, as evidenced by the numerous initiatives being pursued by the United Nations, as well as by others, including the World Bank, OECD, Paris21 and NSOs.

There are many big data sources that can be used to monitor and assess development results. In a world where mobile telephony is increasingly ubiquitous, it is not surprising that mobile telecommunication big data have unique potential as a new data source, with high mobile-cellular penetration levels and the increasing use of mobile phones, even among the poorest and most deprived, making them particularly valuable by comparison with other types of telecommunication data. Indeed, when referring to the data revolution, the United Nations High-Level Panel cited the example of "mobile technology and other advances to enable real-time monitoring of development results".

This section will present some of the existing (and growing) evidence for the role of mobile big data in achieving development goals in various policy areas, including disaster management and sustainable and economic development.

In addition to their use for development, telecommunication big data have potential as a source to enable monitoring of the information society, although they have yet to assume a critical role in complementing the official ICT statistics that are collected and used for that purpose. As the lead agency on global telecommunication and ICT statistics, however, ITU is exploring the potential of big data to complement its existing, and often limited, set of ICT statistics. This section presents a first attempt to help identify some of the areas in which mobile telecommunication big data could complement existing ICT indicators to provide a more complete, comprehensive and up-to-date picture of the state of today's information society.

Mobile phone big data for development

Mobile data offer a view of an individual's behaviour in a low-cost, high-resolution, real-time manner. Each time a user interacts with a mobile operator, many details of the interaction are captured, creating a rich dataset relating to the consumer. Topping up airtime, making calls and sending SMSs, downloading applications or using value-added services are all examples of interactions for which the time, location, device, user and other detailed information are captured in the operator's system. From these interactions, information about identity, movement patterns, social relationships, finances and even ambient environmental conditions can be extracted. In addition to the fact that these data are uniquely detailed and tractable, the information captured cannot easily be derived from other sources on such a scale. The fact that the format of the data is relatively similar across different operators and countries creates a huge potential for the global scaling of any application found to have significant benefits. Box 5.4 illustrates the potential of mobile data for development in a number of different areas.

There have been a number of interesting research collaborations and some promising proof-of-concept studies, for example in the areas of disaster management and transportation planning, and for understanding socio-economic developments and societal structures.

Big data for disaster management and syndromic surveillance[21]

Mobility data collected immediately after a disaster can in many cases help emergency responders to locate affected populations and enable relief agencies to direct aid to the right locations (Lu, Bengtsson and Holme, 2012b).

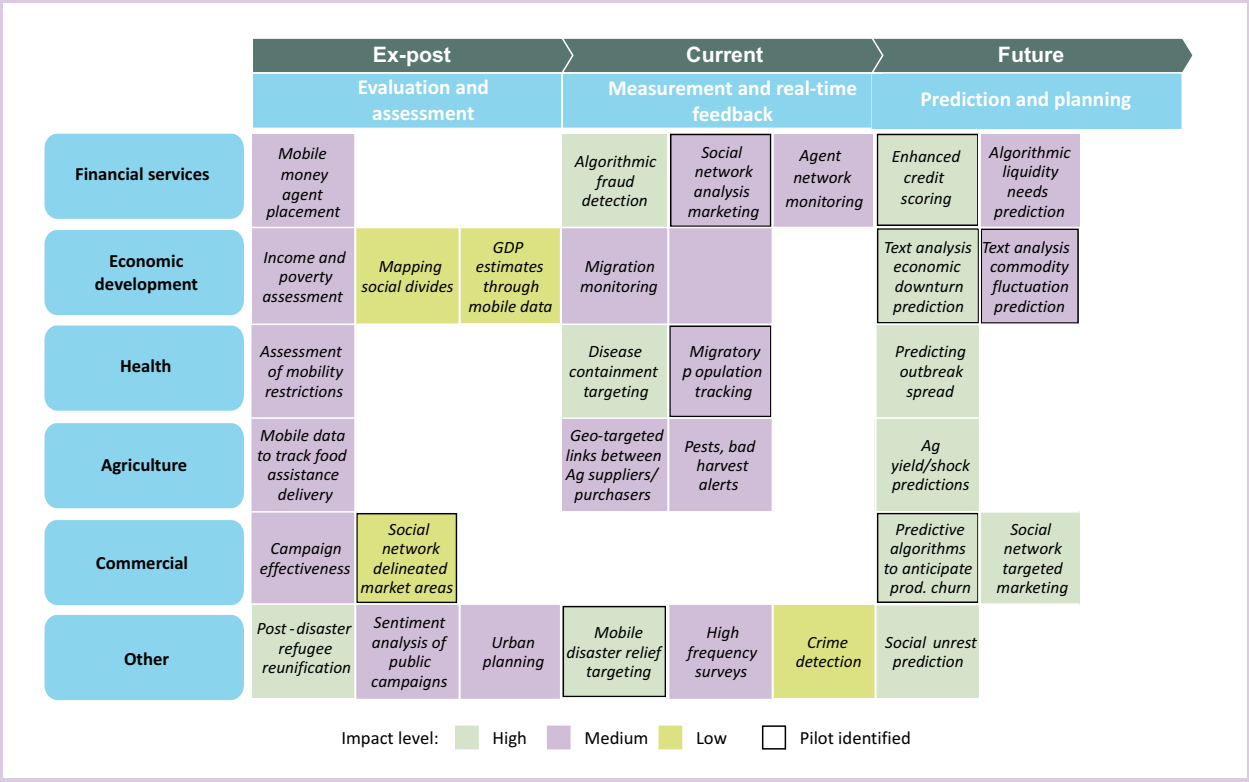
One application of such mobility data is for syndromic surveillance, especially to model the spread of vector-borne[22] and

Box 5.4: Using mobile data for development

A recently published report (Cartesian, 2014) explores the potential of mobile data for development. It points to three primary types of analysis - ex-post evaluation, real-time measurement, and future predictions and planning - in a number of areas (including health, agriculture and economic development), and suggests that *"the more predictive the analytics can be, the higher impact the analysis will have through the ability to anticipate future events or trends"* (Figure Box 5.4).

The report further highlights that while there have been a number of interesting research collaborations and some promising proof-of-concept studies, no significant programme has yet been brought to a replicable scale. Future efforts will have to overcome a number of barriers to scale, including the development of models which protect user privacy while still allowing for the extraction of insights that can serve development purposes, particularly where those in most need, including low-income populations, are concerned.

Figure Box 5.4: Areas of highest potential impact across sectors



Source: Naef et al. (2014).

other communicable diseases. Pioneering research in Kenya combined passive mobile positioning data with malaria prevalence data to identify the source and spread of infections (Wesolowski et al., 2012b). Similar work in Haiti showed how mobile phone data was used to track the spread of cholera after the 2010 earthquake (Bengtsson et al., 2011, see Box 5.5).

The integration of mobility data from mobile networks with geographic information frameworks,[23] supplemented with additional information, shows great potential for tracking the spread of vector-borne and other communicable diseases. This highlights the need to ensure that the response plan implemented after any disaster includes ensuring that any damaged mobile-network infrastructure is repaired as rapidly as possible.

Big data for better transportation planning

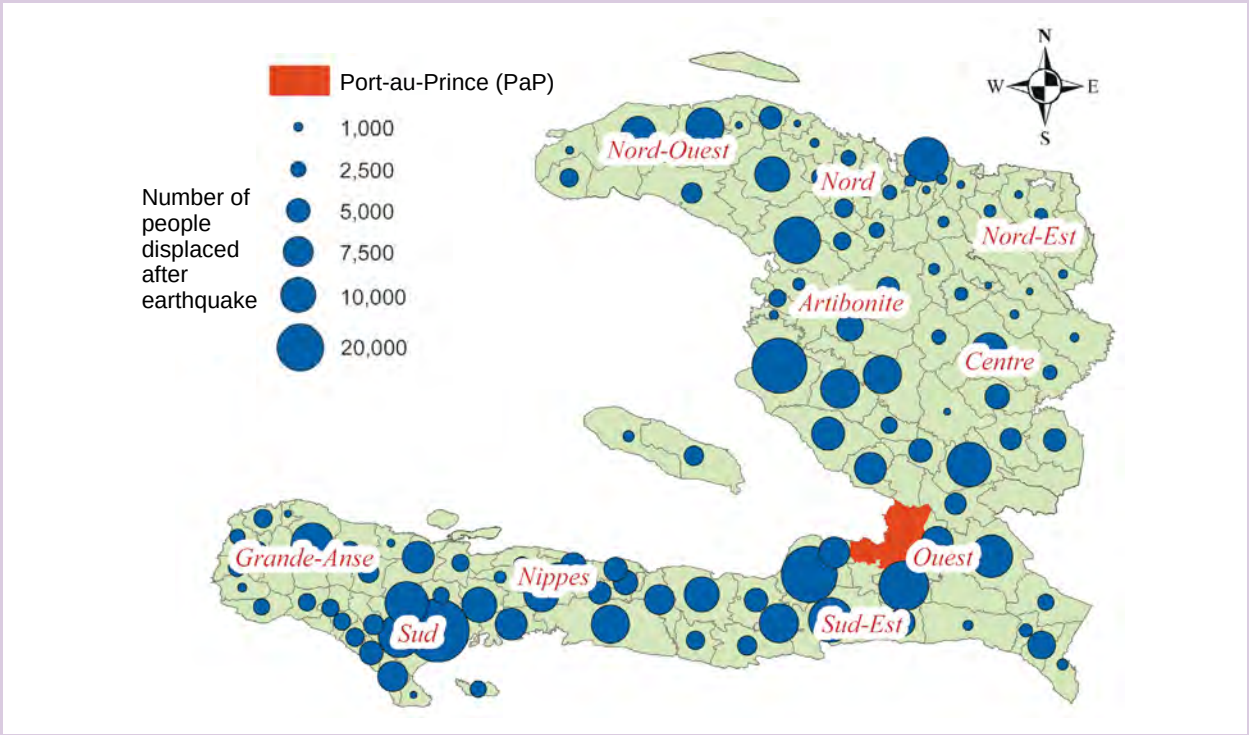
A data-centric approach to transportation management is already a reality in many developed economies. Transportation systems are being fed with sensor data from a multitude of sources such as loop detectors, axle counters, parking occupancy monitors, CCTV, integrated public transport card readers and GPS data derived not only from phones but also from public transport and private vehicles (Amini, Bouillet, Calabrese, Gasparini and Verscheure, 2011).

One advantage of mobile networks is that even the least developed mobile-network infrastructure generates passive positioning data, which, despite its limited spatial accuracy (cell ID), has great potential for transportation planning. For example, IBM researchers used CDR data from mobile operator Orange to map out citizens' travel routes in Abidjan, the largest city in Côte d'Ivoire, and show how data-driven insights could be used to improve the planning and management of transportation services, thereby reducing congestion (Berlingerio et al., 2013). By simply extending one bus route and adding four new ones, overall travel time was reduced by ten per cent. Passive mobile positioning data has also been used for transportation planning and management in Estonia (Ahas and Mark, 2005), and has provided reliable results in Sri Lanka (Lokanathan et al., 2014, see Box 5.6).

Box 5.5: How mobile-network data can track population displacements – an example from the 2010 Haiti earthquake

The Figure below shows the number of people estimated to have been in Port-au-Prince (PaP) on the day of the 2010 Haiti earthquake, but *outside* the capital 19 days later. The circles represent the numbers of people who were displaced. This map was produced on the basis of mobile-network data to show the potential of big data in tracking population movements.

Figure Box 5.5: Tracking mobility through mobile phones



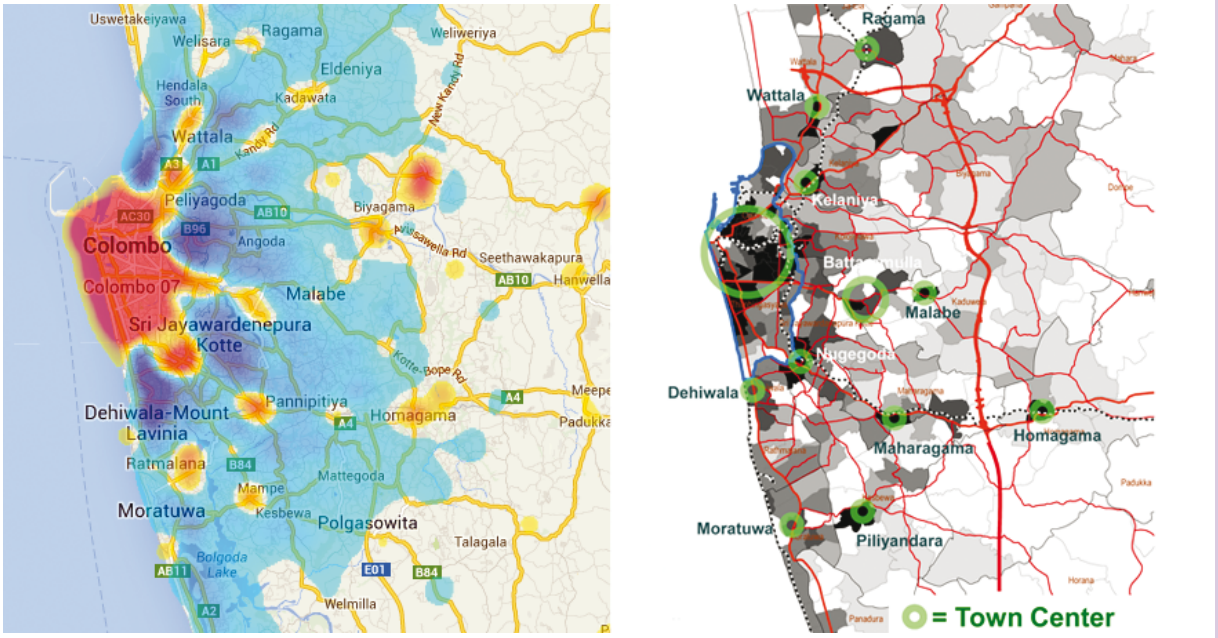
Source: Bengtsson et al. (2011).

Box 5.6: Leveraging mobile-network data for transportation and urban planning in Sri Lanka

Very similar findings between the results of an official household survey assessing mobility patterns (right-hand map) with the results of a big data analysis using mobile-phone data (left-hand map) underscore the merits of big data. The image on the left, based on mobile-phone data, depicts the relative population density in Colombo city and its surrounding regions at 1300 hours on a weekday in 2013, compared to midnight the previous day. While the yellow to red colouring shows areas in which the density has increased relative to midnight, the blue colouring depicts areas in which the density has decreased relative to midnight (the darker the blue, the greater the loss in density). The clear areas are those in which the overall density has not changed. The image reflects the movement of people from the outskirts of the city to its centre during the day.

An almost identical finding is to be seen in the map on the right, which depicts the major transportation transit points identified using a costly survey of 40 000 households to understand mobility patterns.

Figure Box 5.6 : Mobile big data (left) versus official survey data (right)



Source: Lokanathan et al. (2014).

Both passive and active positioning data are used to analyse traffic conditions, particularly in urban areas with higher base-station density. Active positioning data (especially GPS) produce higher precision in location data and are therefore the most useful. Operators may offer such specialized services (based on passive or active location data) either directly, or by providing data to third parties. Mobile network data are less expensive, are in real time and are less time-consuming to produce than survey data, particularly in urban and peri-urban areas where base-station density tends to be high.

In another example, the analysis of mobility flows between two Spanish cities derived from three different data sources – mobile-phone data, geolocated Twitter messages and the census – showed very similar results, and although the representativeness of the Twitter geolocated data was lower than the (real-time) mobile-phone and census data, the degrees of consistency between the population density profiles and mobility patterns detected by means of the three datasets were significant (Lenormand et al, 2014).

Big data for socio-economic analysis

Data from mobile operators can provide insights in the areas of economic development and socio-economic status, often in near real time. Big data techniques can therefore complement official statistics in the intervals between official surveys, which are usually relatively expensive and time-consuming and therefore carried out infrequently. In many cases, insights derived from big data sources may help to fill in the gaps, rather than replace official surveys. It should also be noted that mobile-network big data are one of the few big data sources (and often the only one) in developing economies that contain behavioural information on low-income population groups.

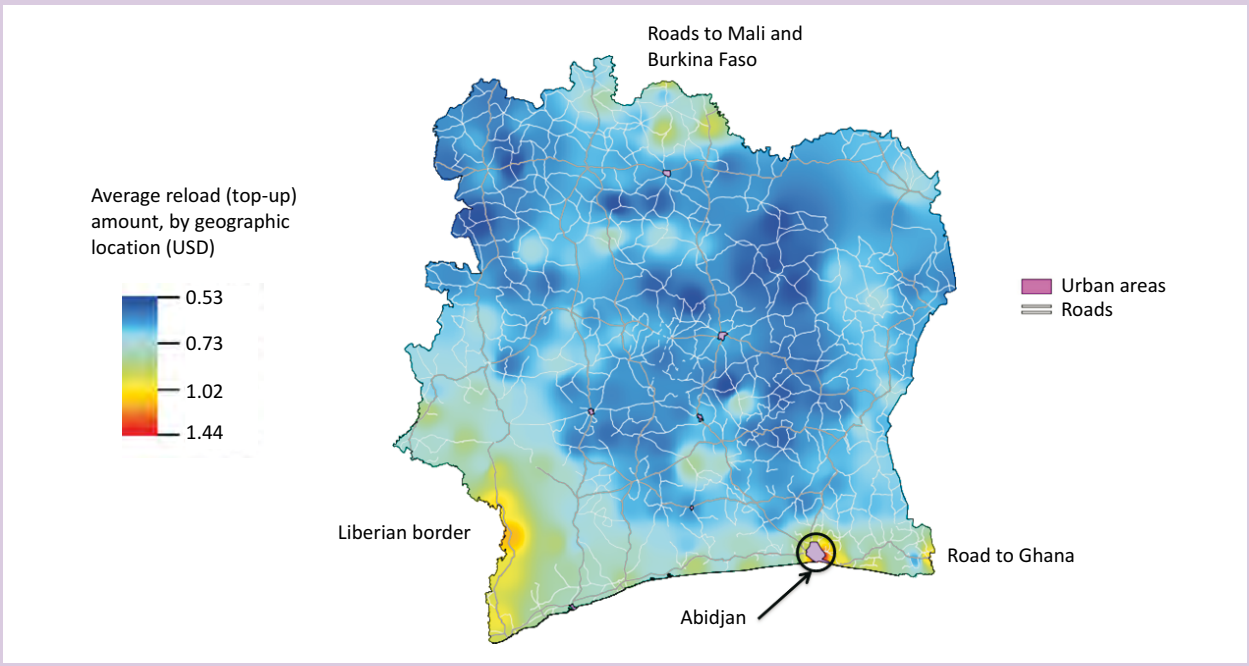
Frias-Martinez et al. (2012) developed a mathematical model to map human mobility variables derived from mobile-network data to people's socio-economic and income levels. The model took into account existing socio-economic and income-related data derived from official household surveys, and the results showed that populations with higher socio-economic levels are more strongly associated with larger mobility ranges than populations from lower socio-economic levels. By extending this method, the study suggested that it was possible to create a model to estimate income levels based on data from mobile-network operators.

Another study, by Gutierrez, Krings and Blondel (2013), used two types of mobile-network data, namely subscriber communication data and airtime credit purchase records, to assess socio-economic and income levels. The authors used airtime purchase records based

Box 5.7: Poverty mapping in Côte d'Ivoire using mobile-network data

In Côte d'Ivoire, researchers used mobile-network data (specifically communication patterns, but also airtime credit purchase records) from Orange to estimate the relative income of individuals, as well as the diversity and inequality of income levels. The research helped to understand socio-economic segregation at a fine-grained level for Côte d'Ivoire, with the following map showing poor areas (in blue) in relation to the areas of high economic activity (yellow to red areas).

Figure Box 5.7: High-and low-income areas in Côte d'Ivoire



Source: Gutierrez et al. (2013).

on the assumption that users who make large purchases are more affluent than those who make multiple smaller purchases, as people with lower incomes will not have enough ready cash to make single purchases of large amounts of airtime credit. They combined this analysis with a study of users' social networks, with two users being considered as connected if they communicated with each other at least once a month. Results showed that people tend to socialize with those who have a similar purchasing power (i.e. a similar income level).

Research suggests that operators have access to potentially valuable information that could help improve poverty mapping (See Box 5.7) and identify sudden events that depress the local economy (i.e. economic shocks). One of the challenges has to do with operator sensitivity regarding revenue data and the difficulty this poses for outside parties wishing to obtain such data.

The use of mobile-operator TGD can also foster financial inclusion by facilitating the provision of credit to the unbanked. In 2012, the Consultative Group to Assist the Poor (CGAP) and GSM Association (GSMA) estimated that close to 2 billion people had a mobile phone but no bank account. After analysing mobile consumption variables, CGAP suggested that it was possible to identify the creditworthiness of the unbanked (Kumar and Muhota, 2012). For example, people who purchase airtime frequently and in a consistent manner demonstrate income predictability and an ability to plan ahead, which may be a positive indication of their ability to repay a loan. Conversely, people with prepaid accounts that are inactive or which regularly run out of credit would perhaps not repay a loan in a timely manner. A compelling example of how mobile big data can be used for the unbanked is Cignifi, a big data startup that uses the mobile phone records of poor people to assess their creditworthiness when they apply for a loan (Box 5.8).

Big data for understanding societal structures

Social-network studies relying on self-reporting relational data typically involve both a limited number of people and limited number of time points (usually one). As a result, social-network analysis has generally been confined to the examination of small population groups through a small number of snapshots of interaction patterns. By examining social communication patterns based on telecommunication data, it has become possible to obtain insights into societal structures on a scale that was previously unavailable. Mobile-phone records

Box 5.8: Using mobile-phone data to track the creditworthiness of the unbanked

Cignifi, a big data startup, has developed an analytic platform to provide credit and marketing scores for consumers, based on their mobile-phone data. The Cignifi business model is founded on the idea that "Mobile phone usage is not random – it is highly predictive of an individual consumer's lifestyle and risk".

Based on the behavioural analysis of each mobile-phone user – phone calls, text messages, data usage and, extrapolating from these, lifestyles – the company identifies patterns and uses them to generate individual credit risk profiles. This information could help many of the world's unbanked to have access to insurance, credit cards and loans. Scores are dynamic and respond to changes in customer activity as the data are refreshed, usually every two weeks. In addition to updating a person's creditworthiness, the system also helps to identify a customer's appetite for different products and inclination to churn.

The credit-scoring model is being tested in Tanzania and Brazil. In Brazil, Cignifi constructed 50 behavioural variables from 2.3 million prepaid mobile subscriptions and verified the findings from the model against historical lending data from approximately 40 000 borrowers using the mobile operator Oi's lending business, Oi Paggo. The test showed that the model was an accurate predictor of default, with the score proving to be a useful complement to, although not a replacement for, the credit underwriting effort. Experian Microanalytics ran a similar trial in the Philippines (Kumar and Muhota, 2012).

Source: ITU, based on Cignifi.

have been used to study the geographic dispersion and cohesion of societies in relation to socio-economic boundaries by examining the geospatial distribution of societal ties in both developed and developing economies (Sobolevsky et al., 2013).

However, telecommunication data are also revolutionizing the study of societal structures at the micro level. Eagle et al. (2009) show that it is possible to assess friendship using data from mobile-network operators, and that the accuracy is high when compared with self-reported data. Leveraging these behavioural signatures to obtain an accurate characterization of relationships in the absence of survey data could also enable the quantification and prediction of macro and micro social-network structures that have thus far been unobservable.

Big data to monitor the information society

There is a case to be made for analysing data captured by telecommunication operators in the interests of improving the current range of indicators used for monitoring the information society. An internationally-accepted and widely-adopted list of indicators is the core list of ICT indicators developed by the Partnership on Measuring ICT for Development, a multi-stakeholder initiative launched in 2004.[24] This list includes, among others, the key-infrastructure, access and individual-use indicators that ITU collects and disseminates. Some of these indicators are amenable for augmentation using big data analytics.[25]

The core indicators on ICT infrastructure and access include indicators on mobile-cellular and mobile-broadband subscriptions, which remain some of the most widely used and internationally comparable telecommunication indicators produced for tracking the information society. One of the main issues with mobile-cellular and mobile-broadband subscription data is that they do not refer to unique subscriptions, or mobile users. Since one person can have multiple subscriptions, or share a subscription with another person, it is not possible to determine how many individuals subscribe to, or use, the mobile service. It is often the case that countries with large numbers of prepaid subscriptions display relatively high penetration rates since prepaid cards can often be purchased at no initial cost and do not involve a recurring (monthly) cost. In countries with high interconnection charges, for example, multiple prepaid cards can help avoid high costs when making off-net calls. In addition, prepaid consumers do not usually cancel their account after ceasing to use a given network, making it important for operators to monitor the time during which a SIM card remains inactive. In June 2014, for example, GSMA estimated that, globally, the number of unique mobile subscriptions was just over 50 percent, whereas the number of connections per 100 population far exceeded 100 per cent.[26]

Survey-based data, for example on Internet users and mobile-phone users, do not entail the same issues as subscription data. They are collected through household surveys, directly from citizens, and their level of reliability is relatively high. The advantage of surveys is that they can go into more depth on the use of ICTs. For example, one of the core indicators reflects the types of online activity pursued by Internet users, and includes response categories such as seeking health information, obtaining information from government entities or participating in social networks. Survey-based data can also be broken down by individual characteristics, including gender, age, educational level and occupation, which substantially increase the data's added value. However, the main challenge with these data is that they are not widely available (in particular, many developing countries do not yet collect data on individual use of, or household access to, ICTs), are relatively expensive to produce, and are much less timely than subscription data (often with a time lag of one year). Consequently, data on users of the information society and the types of online service they consume are limited, and in many

cases outdated. Against this background, mobile networks and mobile big data could be used to identify alternative, less costly and faster ways of carrying out representative surveys (Box 5.9).

Given the shortcomings of existing administrative data from operators and survey data collected by NSOs, it is particularly interesting to assess some of the ways in which big data can be used to overcome the shortcomings of existing key ICT indicators and to provide additional insights into ICT access and use, user behaviour, activities and also the individual user. Big data could help in obtaining more granular information in several areas, and big data techniques could be applied to existing data to produce new insights. In particular, operators' big data could produce information in the following areas:

Individual subscriber characteristics: Additional categorization across both time and space are possible for subscription indicators, and big data could provide additional information on gender, socio-economic status and user location. Information on gender or age, for example, could be derived from customer registration information (notwithstanding a number of challenges and privacy issues, as discussed later in Section 5.5). The socio-economic status of the person linked to a subscription could be derived from big data techniques applied to users' consumption information, as well as other data contained in customer registration information. In addition, the analysis of customers' mobility patterns will often allow for an understanding of important locations (work and home being the two most important) and of the use of mobile services in rural versus urban areas. It would thus be possible to gain a more reliable and more granular understanding of service penetration across space on the basis of actual behaviour/activity, rather than of what customers may report through a survey.

Service activity and use: All subscription data could provide information as to location. In the case of fixed-telephone and fixed-broadband subscriptions, which are linked to an address through the billing information, it is possible to obtain information on the administrative division of subscribers, distinguish between urban and rural areas, and provide breakdowns by any kind of geographic categorization. Operators could track the types of service that a subscriber uses and the frequency and intensity of use of each of them. Similarly, when it comes to broadband access, operators can also potentially produce detailed information on the technology being used by the subscriber, as well as the associated speed and quality of access for each of the technologies used. Based on the household/individual details provided during the registration process, it could also be possible to provide more information on individual/household characteristics, or to link those characteristics to other (administrative) databases in order to

Box 5.9: Using mobile big data and mobile networks for implementing surveys

An important measurement for assessing the development of the information society is the extent to which households have access to ICTs. Given the need for continued recourse to surveys for collecting the corresponding data, and the declining response rates where traditional surveys are concerned (Groves, 2011), mobile operators could develop platforms to facilitate the collection of survey data. This could include targeting a wide variety of respondents covering the full spectrum of appropriate demographic profiles, followed by a process of extrapolation using big data analytics. In 2011, for example, UN Global Pulse partnered with Jana,[27] a mobile-technology company, to explore the feasibility of using mobile phones for the deployment of rapid global surveys on well-being.[28] This requires, however, that the mobile users targeted for the survey match the requisite survey profile. For instance, if one of the requirements was for the survey to be representative of women, there would have to be some way to determine/infer the demographics of the targeted mobile-phone user. To that end, the World Bank has experimented with the use of mobile phones to conduct statistically representative monthly household surveys in Latin America and the Caribbean.[29]

Source: ITU.

create new information. Consumption patterns could also deliver additional information on the socio-economic status of the person/household linked to a subscription.

Particularly rich possibilities exist where data from mobile-cellular and mobile-broadband subscriptions are concerned, since they are linked to mobility profiles. The indicators for such subscriptions could be further broken down to understand the utilization of services – including voice, data and VAS – over time, and the intensity of use. Mobile operators are able to provide information not only on the different technologies (3G, LTE-Advanced, etc.), but also on the types of service that subscribers are using, and the frequency and intensity of that use. They could, therefore, potentially identify Internet and VAS usage patterns between rural and urban areas, and identify the kinds of application or webpage that mobile-Internet users access. Combined with individual subscriber characteristics, this information could provide new and rich insights into the digital divide and help understand usage patterns, including intensity of use, by gender, socio-economic status and also location.

Greater utilization of DPI could lead to additional insights that can help to classify access and intensity of use with respect to different Internet activities carried out by individuals. This information is currently collected only by countries that carry out household ICT surveys. In addition, mobile-operator data could be combined with customer information from popular online services, such as Facebook, Google or other, local (financial, social etc.) services to provide additional insights. This could be done by using probabilistic analyses to match the profiles developed using data from online services with customer profiles generated from analyses of mobile-operator data. This would require telecommunication operators, OTT providers and other Internet content providers to work together and share information.

This technique is, currently, probably the least developed one, also because of the lack of a good ontology and of privacy issues. However, as operators seek to gain a better understanding of their customers in terms of the type of content they consume (as revealed through clickstream analyses), DPI may provide greater insights for measuring the information society. In addition, if websites could be individually classified in terms of the information they provide, then Internet-user activities, including their frequency and intensity, could be much better understood.

By applying big data techniques to survey data and administrative data from operators, new insights could be derived, in particular, in respect of the following:

Subscriptions versus subscribers: Big data techniques could help extrapolate the actual number of unique mobile subscribers or users, rather than just subscriptions, by comparing subscription numbers to user numbers derived from household surveys, and by taking into account usage patterns or data from popular Internet companies such as Google or Facebook. By linking data collected from different sources and combining subscription data and usage patterns, a correlation algorithm could be developed to reverse engineer approximate values for these indicators, in order to estimate user numbers in between surveys, and possibly in real time. This could be pursued in a similar way to the work done by Frias-Martinez and Virseda (2012) on estimating socio-economic variables using mobile-phone usage data, as described in greater detail at the beginning of this section. It is important to note here that, depending on such correlation techniques, big data methods only complement existing surveys rather than replacing them completely (see Section 5.5 for a further discussion of this).

In sum, relatively simple big data techniques can help analyse and provide complementary information on existing ICT data, and provide new insights into the measurement of the information society. This includes information on the use of different services and applications, intensity, frequency, and the geographic

locations from which subscribers access ICT services and applications. All of these insights on subscribers could potentially be further disaggregated by different demographic and socio-economic profiles. However, all of them relate to subscriptions. Given multiple SIM usage and the fact that users will in many cases be using ICT services from more than one operator or device, additional techniques need to be leveraged if the insights articulated for subscribers are to be extended to unique individuals. Such techniques will often include combining data from surveys with big data to build new correlation and predictive analytic techniques.

Finally, it should be noted that the methods that could help improve the indicators on individual and household access and use could also be used to complement information on the use of ICTs in businesses, as well as the health and education sectors. In all cases, and for other big data for development projects, big data analysis cannot replace survey data, which is needed to build and test correlations and to validate big data results.

While the opportunities discussed above present what is analytically possible, data access and privacy considerations are complex and nuanced, and therefore place constraints on what is practically feasible or advisable. Section 5.5 discusses these challenges in greater detail.

5.5 Challenges and the way forward

Attempting to extract value from an exponentially growing data deluge of varying structure and variety comes with its share of challenges. The most pressing concerns are those associated with the standardization and interoperability of big data analytics, as well as with privacy and security. Addressing such privacy and other concerns with respect to data sharing and use is critical, and it is important for big data producers and users to collaborate closely in that regard. This includes raising awareness about the importance and potential of producing new insights, and the establishment of public-private partnerships to exploit fully the potential of big data for development.

Data curation, standardization and continuity

Data curation and data preparation help to structure, archive, document and preserve data in a framework that will facilitate human understanding and decision-making. Traditional curation approaches do not scale with big data and require automation, especially since 85 per cent of big data are estimated to be unstructured (TechAmerica Foundation, 2012). Dealing with large heterogeneous data sets calls for algorithms that can understand the data shape while also providing analysts with some understanding of what the curation is doing to the data (Weber, Palmer and Chao, 2012).

Telecom network operators themselves have to contend with interoperability issues arising from the different systems (often from different vendors) they employ. It is not uncommon for operators to write customized mediation software to overcome potential inter-comparability issues among data from different systems. The problems are compounded when one has to take account of secondary third-party users that may seek to leverage the data. The framework used by an NSO to organize data would be different from that used by a network engineer or a marketing or business intelligence specialist. Naturally, telecom network operators have curated their data based on their needs. To be able to use telecom big data for development and monitoring, and to guarantee its continuity, the creation of a semantic framework would require greater consensus among the many diverse stakeholders involved (telecom operators, network equipment manufacturers, system developers, developmental practitioners and researchers, NSOs, etc.).

Accessing and storing data, and data philanthropy

Big data for development is still in its nascent stages and, as such, comes with its share of challenges, not least of which is obtaining access to what is essentially private data. Private corporations would hesitate to share information on their clients and their business processes in case such sharing is illegal, precipitates a loss of user confidence and/or accidentally reveals competitive business processes. More importantly, companies will not share until there are incentives to do so. Until holders of big data become more comfortable about their release, it is going to be difficult for third-party research entities to gain access.

Researchers (mainly from developed countries, with some exceptions such as LIRNE*asia*) have recently succeeded in obtaining mobile-network big data, but it has taken them considerable time to build and leverage the necessary relationships with operators. Such privileged access is for the most part conditioned by lengthy legal agreements whose preparation requires major investments of time. All the parties to such agreements have to address the necessary parameters as to how data are to be used, including the manner in which they are to be anonymized and extracted, and with regard to time periods for access, etc. Even once agreements are in place, both researchers and operators face costs arising from the technical challenges associated with extraction of the data, on account of different curation approaches and problems relating to the interoperability of different systems.

Some mobile operators are taking tepid steps towards sharing data more publicly. Orange, for example, hosted a "Data for Development Challenge," releasing an aggregated anonymized mobile dataset from Côte d'Ivoire to researchers and convening a conference at MIT in early 2013, where 84 papers from different researchers were presented. A follow-up conference, this time using Orange data from Senegal, is planned for 2015.[30] In 2014, Telecom Italia initiated a similar challenge, making data from the territories of Milan and the Autonomous Province of Trento available to researchers for analysis.[31] It has, however, gone one step further: in addition to releasing some of its own telecom datasets, it partnered with other data providers to curate and release additional big datasets containing weather, public and private transport, energy, event and social network data. In both the Orange and Telecom Italia cases, researchers had to go through an approval process in order to gain access.

Organizations such as UN Global Pulse are seeking to popularize the concept of "data philanthropy", aimed at systematizing the regular and safe sharing of data by building on the precedents being created by the ad hoc activities outlined earlier. Such efforts by UN Global Pulse, as well as by other organizations such as LIRNE*asia*, that seek to bring different stakeholders to the same table, remain critical to the efforts being made to open up private-data stores in order to obtain actionable development insights.

There is a gap that needs to be addressed if large-scale pooling and sharing of such data are to become a reality. Cross-sector and cross-domain collaboration would benefit greatly from facilitators or intermediaries capable of addressing issues related to standardization and data-curation practices when pooling data from multiple sources. This facilitatory role may even be played by a third-party organization able to subsume regulatory and privacy burdens faced by data providers, effectively acting as a gatekeeper to ensure that data are used transparently and in a way that contributes to overall scientific knowledge generation, while ensuring that any safeguards that may be applicable in respect of private information are applied. Such an approach was taken recently by the pharmaceutical company Johnson and Johnson, which decided to share all of its clinical trial data. To facilitate the process, they hired Yale University's Open Data Access (YODA) Project

to act as gatekeepers (Krumholz, 2014). YODA undertakes the necessary scientific review of any proposals (from scientists around the world) to make use of the data and ensures that necessary privacy and data usage guidelines are followed.[32]

The question remains as to who is best placed to act as gatekeepers and standard-bearers when it comes to telecom network big data. Some have argued that NSOs are well placed to ensure that best practices are followed in the collection and representation of big data, and to provide a stamp of trust for potential third-party data seekers. Telecom operators, for their part, are mostly regulated by sector-specific regulators who can also have purview and dictate terms governing the privacy and data-reporting responsibilities of operators. Ultimately, however, the decision as to who takes on the gatekeeper and standardization function requires the confluence of multiple actors. It is here that organizations such as ITU, UN Global Pulse and others have a greater role to play in building an institutional model for data sharing and collaboration, in consultation with all stakeholders.

The sharing (subject to appropriate privacy protocols) of privately held data such as mobile-phone records can be mutually beneficial to both government and private sector. For example, mobile-network operators monitor and forecast their revenue at the cell-tower level. Emerging research in Africa shows how reductions in revenues, including airtime top-ups, could presage declines in income in specific regions. This could allow for targeted and timely policy actions by government to address the underlying problems, which would not be possible with the delayed insights provided by traditional statistics. Such a collaborative early-warning and early-action system shows how data sharing could be considered a business risk mitigation strategy for operators in emerging markets. However, such cooperation is predicated on opening up the currently privileged access that a few researchers and organizations have been given to mobile-operator datasets.

Finally, it should be noted that the emergence of big data is closely linked to advances in the ICT sphere, including the falling cost of data storage. Depending on the data volume, storage can still be costly, especially where privacy considerations pre-empt the use of specialized third-party cloud-based services. But as storage prices continue to fall, they are expected to be less of an issue.

Privacy and security

As social scientists look towards private data sources, privacy and security concerns become paramount. To mitigate the potential risks, all stakeholders must see tangible benefits from such data sharing. These stakeholders include not just the public and private sectors, but also, significantly, the general public, who in many cases are the primary producers of such data through their activities. It is also the public that must ultimately decide on how the data they produce may be used. The World Economic Forum's "Rethinking Personal Data" project has identified key trust challenges facing the personal data economy, and hosts consultations to deepen understanding of what type of trust frameworks are needed between individuals and the private and public sectors in today's new data ecosystem.[33] Discussions must address the individual's privacy expectations, as well as those of private-sector stakeholders looking to protect their competitiveness. The most common approach to addressing this issue has been the rights-based approach. ITU, for example, has defined individual privacy as "the right of individuals to control or influence what information related to them may be disclosed" (ITU, 2006).[34] Central to the rights-based privacy framework is the implicit or explicit existence of personal data that needs to be protected. OECD, for example, defines personal data as "any information relating to an identified or identifiable individual (data subject)" (OECD, 2013). The result of such an approach has been the policy of "inform and consent" practised by most companies to inform users of what data are

being collected and how they will be used. It has been argued, however, that in a big data world the "inform and consent" approach is woefully inadequate and impractical, and that a new approach is needed (Mayer-Schönberger and Cukier, 2013; WEF, 2013).

Firstly, user-privacy policies have morphed into long documents written in 'legalese' that most users can hardly comprehend and have little patience for reading in full. Secondly, in the big data paradigm, the greatest potential often lies in secondary uses, which may well manifest long after the data was originally collected. It is thus impractical for companies to have *a priori* knowledge of all the potential uses and to seek permission from the user every time such a new use is found. Given the volumes of data that individuals are now generating, companies would find themselves struggling to maintain meaningful control.

Of greater concern is how to articulate the privacy issues that may arise when data from one source is combined with data from other sources to reveal/infer new data and insights. This blurs the lines between personal and non-personal information, allowing seemingly non-personal data to be linked to an actual individual (Ohm, 2010). Digitized behavioural data crumbs may in fact greatly diminish personal privacy. The use of DPI, for example can technically reveal all of a user's online activity. Going one step further, it is possible to understand a person's needs, behaviours and preferences by using data-mining techniques on the digital breadcrumbs. For instance, a recent study showed how Facebook "likes" could accurately predict a range of behavioural attributes such as, inter alia, sexual orientation, ethnicity, religious and political views, and use of addictive substances (Kosinski, Stillwell and Graepel, 2013).

Data anonymization[35] (i.e. methods designed to strip data of personal information), employed by computational social scientists, has been called into question (Narayanan and Shmatikov, 2008). A recent study of mobile CDRs for 1.5 million anonymized users covering a 15 month period showed how the authors were able to identify 90 per cent of the users with just four data points, and 50 per cent with just two points (de Montjoye, Hidalgo, Verleysen and Blondel, 2013). Although the actual real-world identities of the users were unknown, the authors point out that the data could in fact be de-anonymized completely by cross-referencing them with other data sources. The attendant privacy concerns about such cross-referencing are clear, and have to be taken seriously and addressed.

However such de-anonymization concerns remain, for the time being, somewhat premature for developing countries, mainly because the levels of 'datafication' in developing economies are still quite low. Where mobile-phone records are concerned, the large majority of connections in the developing world are prepaid, with minimal (if any) associated registration information. Security imperatives have increasingly prompted governments to require registration information, even for prepaid customers (GSMA, 2013b), but even with registrations becoming mandatory for prepaid connections, the registered user and the actual user may not be one and the same. Depending on the country, SIM resellers may pre-register the SIMs they sell under their own name, and SIMs that are registered by one family member may be used by other members of the same family. Sri Lankan operators, for example, see a great mismatch between the person registering a subscription and the person using it. The same may also be the case in many other developing countries.[36]

Irrespective, there is a consensus that there have to be safeguards in place, be they technological, conceptual, legal or, more likely, a combination of all three. These safeguards must also ensure that data are kept secure. Data breaches undermine consumer confidence and hinder efforts to exploit big data for the greater social good. Encryption, virtual private networks (VPNs), firewalls, threat monitoring and auditing are some potential technical solutions that are currently employed,

but they need to be mainstreamed (Adolph, 2013). The paradigmatic shift required to address privacy has started, but it will be some time before a consensus is achieved on the most appropriate method(s). In response to the growing trend to unlock socio-economic value from the rising tide of big data, the World Economic Forum (WEF) initiated a global multi-stakeholder dialogue on personal data that advocated a principle-based approach, with the principles arising from a new approach that shifts governance from the data per se to its use; acknowledges the importance of context rather that treating privacy as a binary concept; and acknowledges the need for new tools to actively engage users, enabling them to make clear choices based on an actual value exchange (WEF, 2013).

Given the complexity of the questions related to privacy and data protection in a big data world, the danger is that these questions may take too long to resolve and further delay the potential use of big data for broader development. Hence, a balanced risk-based approach may be required in the context of what is under discussion here, i.e. the use of telecom big data for monitoring and development. This does still require the confluence of appropriate stakeholders. But as UN Global Pulse suggests, research into the use of big data for development can be "sandboxed", with appropriate privacy protections imposed on researchers, while still ensuring that the broader privacy implications and solutions continue to be discussed and worked out.[37]

Veracity in data, analysis and results

"Garbage in, garbage out", or GIGO for short, is a computer science concept that refers to the fact that the veracity of the output of any logical process depends on the veracity of the input data. In the big data paradigm, it is easy to overlook that concept, given the expectation that when dealing with vast volumes of (often unstructured) data from a multitude of sources, "messiness" is to be expected. As Mayer-Schönberger and Cukier (2013) note, "What we lose in accuracy at the micro level we gain in insight at the macro level." This common conception can often be misleading. Data quality and their provenance do matter, and the question is important in establishing the generalizability of the big data findings.

Data provenance and data cleaning

Understanding data provenance involves tracing the pathways taken by data from the originating source through all the processes that may have mutated, replicated or combined the data that feed into the big data analyses. This is no simple feat. Nor, given the varied sources of data that are utilized, is it always as feasible as the scientific community would wish. However, at the very least it is important to understand some aspects of the origin of data. For example, the fact that some mobile-network operators choose to include the complete routing of a call that has been forwarded means that there may be multiple records in the CDRs for the same call. If that is not taken into consideration, the subsequent social network analysis could contain errors (overstating or understating tie strength, for example). While it may not be possible to establish data provenance as envisaged by scientists, it is at the very least important to understand the underlying processes that may have created the data.

Data cleaning remains a key part of the process to ensure data quality. It is important to verify that the quantitative and qualitative (i.e. categorical) variables have been recorded as expected. In a subsequent step, outliers must be removed, using decision-tree algorithms or other techniques. However, data cleaning itself is a subjective process (for example, one has to decide which variables to consider) and not a truly agnostic one as would be desirable, and is thus open to philosophical debate (Bollier, 2010).

Are the data representative?

Related to the question of data provenance is the issue of understanding the underlying

population whose behaviour has been captured. The large data sizes may make the sampling rate irrelevant, but they do not necessarily make it representative. Not everyone uses Twitter, Facebook or Google. For example, ITU estimates suggest that 40 per cent of the world's population uses the Internet. In other words, more than four billion people globally are not yet using the Internet, and 90 per cent of them are from the developing world. Of the world's three billion Internet users, two-thirds are from the developing countries. Even though mobile-cellular penetration is close to 100 per cent, this does not mean that every person in the world is using a mobile phone. This issue of representativeness is of high relevance when considering how telecommunication data may be used for monitoring and development. While the potential benefits to be gained from leveraging mobile-network operator data for monitoring and development purposes hinges on the large coverage, close to the actual population size, it is nevertheless not the whole population. Questions such as the extent of coverage of the poor, or the levels of gender representation among telecom users, are all valid considerations. While the registration information might provide answers, the reality is that the demographic information on telecom subscribers, for example, is not always accurate. With prepaid subscriptions being the norm in most of the developing world, the demographic information contained in mobile-operator records is practically useless, even with mandatory registration as discussed above.

The issue of sampling bias is best illustrated by the case of Street Bump, a mobile app developed by Boston City Hall. Street Bump uses a phone's accelerometer to detect potholes while users of the app are driving around Boston and notifies City Hall. The app, however, introduces a selection bias since it is slanted towards the demographics of app users, who often hail from affluent areas with greater smartphone ownership (Harford, 2014). Hence, the "big" in big data does not automatically mean that issues such as measurement bias and methodology, internal and external data validity and data interdependencies can be ignored. These are fundamental issues not just for "small data" but also for "big data" (Boyd and Crawford, 2012).

Behavioural change

Digitized online behaviour can be subject to self-censorship and the creation of multiple personas, so studying people's data exhaust may not always give us insights into real-world dynamics. This may be less of an issue with TGD, where in essence the data artefact is itself a by-product of another activity. Telecom network big data, which mostly fall under this category, may be less susceptible to self-censorship and persona development, but the possibility of these phenomena cannot be ruled out. Nor is it inconceivable that users may stop using their mobiles, or even turn them off, in areas where they do not wish their digital footprint to be left behind. In a way, big data analyses of behavioural data are subject to a form of the Heisenberg uncertainty principle, whereby as soon as the basic process of an analysis is known, there may be concerted efforts to exhibit different behaviour and/or actions to change the outcomes (Bollier, 2010). For example, the famous Google page-rank algorithm has spawned an entire industry of organizations that claim to enhance website page rankings, and search-engine optimization (SEO)[38] is now an established part of website development.

Changes in behaviour could also partially explain the declining veracity of Google Flu Trends (GFT), researchers having found influenza-like illness rates as reflected by Google searches to be no longer necessarily correlating with actual influenza virus infections (Ortiz et al., 2011). Recent research has shown that since 2009 (when GFT failed to reflect the non-seasonal influenza outbreak), infrequent updates have not improved the results and GFT has in fact persistently overestimated flu prevalence (Lazer, Kennedy, King and Vespignani, 2014). GFT does not and cannot know what factors contributed to the strong correlations found in its initial

work. The point is that the underlying real-world actions of the population that turned to Google with its health queries, and which contributed to the original correlations identified by GFT, may in fact have changed over time, diminishing the robustness of the original algorithm. For example, the enthusiasm surrounding GFT may well have created rebound effects, with more and more people turning to Google with their broader health questions, thereby introducing additional search terms (due to different cultural norms and/or ground conditions) and collectively introducing biases for which GFT has been unable to account. Such potential problems could have been foreseen and resolved had the GFT method been more transparent (see Section 5.2).

Real-world context

Knowing and understanding the real-world context therefore remains important when considering big data analyses for monitoring purposes. Dr Nathan Eagle, a pioneer in the use of cellphone records to understand phenomena related to social development and public health, stresses the importance of weeding out false assumptions by conducting an a priori survey of even a small number of people. For example, in one instance, when CDR data from Rwanda showed low mobility in the wake of flooding, he theorized that this was due to an outbreak of cholera. A ground survey, however, revealed the true cause of the low mobility to be washed-out roads (David, 2013). A knowledge of ground conditions and context is also relevant when it comes to the generalizability of telecom-data analyses based on big data. For example, prior research had established a power-law distribution between the frequency of airtime recharges and average recharge amount.[39] It was further found that the poor tended to top up more frequently but in smaller amounts by comparison with those higher up on the socio-economic ladder (UN Global Pulse, 2012). When researchers working with Sri Lankan mobile datasets attempted to use these findings to help them segregate their analyses for different socio-economic groups, they were unable to do so. A survey of local context based on interviews with operators provided the reason: almost two-thirds of prepaid customers generally chose to recharge using scratch cards. Higher denomination scratch cards were not as readily available as those with lower denominations. Hence, anyone wanting to reload a higher amount often bought multiples of lower-denomination cards. After recharging with one card, the rest were kept aside for when the need arose. A lack of awareness of this local context would have led researchers to assume, mistakenly, that differing airtime-credit purchasing patterns among different socio-economic groups were not prevalent within the Sri Lankan population.

Causation versus correlation

It is easy to confuse correlation with causation in the big data paradigm, leading to the discovery of misleading patterns. As Google's Chief Economist, Hal Varian, notes, "there are often more police in precincts with high crime, but that does not imply that increasing the number of police in a precinct would increase crime" (Varian, 2013b). Big data draws many of its techniques from machine learning, which is primarily about correlation and predictions.[40] Big data are by their very nature observational and can measure only correlation and not causality. Supporters of big data have predicted the end of theory and hypothesis-testing, with correlation trumping causality as the most relevant method (Anderson, 2008; Mayer-Schönberger and Cukier, 2013). However, such predictions may be premature. The behavioural economist Sendhil Mullainathan notes that inductive science (i.e. the algorithmic mining of big data sources) will not drown out traditional deductive science (i.e. hypothesis testing), even in a big data paradigm. Among the three Vs in the traditional big data definition, volume and variety produce countervailing forces. More volume makes big data induction techniques easier and more effective, while more variety makes them harder and less effective. It is this variety issue that will ensure the need for explaining behaviour (i.e. deductive science)

rather than merely predicting it (Mullainathan, 2013).

Causal modelling is possible in a big data paradigm by conducting experiments. Telecom network operators themselves use such techniques when rolling out new services or, for that matter, for pricing purposes. The question, then, is how third-party researchers will be able to leverage operators' systems in order to conduct such experiments. There is no simple answer to this, since these are proprietary systems and the issue will have to be addressed.

The role of traditional "small data" in verification

The documented failures of GFT also point to the importance of traditional statistics as corroborating evidence. For example, the true value of GFT is realized only through its pairing with "small data," in this case the statistics collected by the Centers for Disease Control and Prevention (CDC). In fact, as Lazer et al. (2014) note, when combined with small data, "Greater value can be obtained by combining GFT with other near–real time health data." Where data from mobile-network operators are used for syndromic surveillance, as in the case of malaria in Kenya (Wesolowski et al., 2012a), big data are most useful as a basis for encouraging timely investigation, rather than as a replacement for existing measures of disease activity. Even when engaging with the broader question of how telecommunication network data could be used for monitoring, surveys and supplemental datasets will remain important to sharpen the analyses and especially to verify the underlying assumptions. For instance, Blumenstock and Eagle (2012) ran a basic household survey against a randomized set of phone numbers prior to data anonymization to build a training dataset. This enabled them, for example, to understand variations in mobility, social networks and consumption among men and women, and between different socio-economic groups, which would not have been possible using only the call records. Similarly, Frias-Martinez and Virseda (2012) needed census data to build their algorithms and provide training data for their algorithms to reverse engineer approximate survey maps. Official statistics will thus continue to be important to building the big data models and for periodic benchmarking so that the models can be fine-tuned to reflect ground realities.

Transparency and replicability

The issues with GFT also illustrate transparency and replicability problems with big data research. The fact that the original private data may in many cases not be available to everyone underscores the importance of opening up such private-data sources (in a manner that addresses potential privacy concerns) or of peer reviews that can hone and improve the analyses. Instead, consumers of such research have no option but to take the analysis and the results on faith. In the case of GFT, for example, the researchers, in their original Nature paper (Ginsberg et al., 2009), did not publish the original 45 search terms that had been used to make the correlation, rendering replicability impossible. Indeed, where methods are transparent they can be updated more effectively when ground realities change – something that could have prevented the problems with GFT.

Skills

Engaging with and extracting value from big data calls for a combination of specialized skills in the areas of data mining, statistics and domain expertise, as well as data preparation, cleaning and visualization. NSOs may have deep statistical skills in house, but this is not enough when it comes to working with large volumes of big data calling for computer science and decision-analysis skills that are not emphasized in traditional statistical courses (McAfee and Brynjolfsson, 2012). NSOs recognize this shortcoming. In a recent global survey of NSOs from 200 economies, conducted by UNSC, respondents identified the development and

retention of staff with the necessary skills as one of their main challenges, and identified intensive training and capacity development of their staff as a prerequisite to being able to exploit new big data sources (UNSC, 2013). Currently, there is a mismatch between the supply of and demand for talented individuals with the requisite broader skill sets, i.e. data scientists. McKinsey predicts that by 2018 the demand for data-savvy managers and analysts in the United States will amount to 450 000, whereas the supply will fall far short of this, at only 160 000 (Manyika et al., 2011). This suggests that organizations wishing to leverage big data for development will face competition from the private sector when seeking to attract the right talent. Unfortunately, developing countries, which stand to benefit the most from the use of telecommunication big data to complement official statistics, have a shortage of advanced analytical skills by comparison with developed economies. Until such time as systematic capacity development yields proper rewards, it will remain essential to import skills from outside (both local and international), despite the difficulties of attracting individuals with the right skill profiles.

The way forward

Current research suggests that new big data sources have great potential to complement official statistics and produce insightful information to foster development. In particular, the private sector, but also a number of development organizations, and governments have started to exploit this potential.

At the same time, this chapter argues that while there have been a number of research collaborations and promising proof-of-concept studies, no significant programme has yet been brought to a replicable scale. Future efforts to mainstream and derive full benefit from the use of big data will have to overcome a number of barriers. This includes the development of models which protect user privacy while still allowing for the extraction of insights that can serve development purposes, in particular where developing countries are concerned. Very limited information is available on opportunities for using big data to complement official ICT statistics. Although this report highlights some of the big data sources and techniques that could be used, further research is needed to understand and confirm the usefulness of big data sources for monitoring the information society.

As with other official statistics, it is paramount for big data producers and big data users to collaborate and to initiate a dialogue to identify opportunities and understand needs and constraints. Since many of the big data sources lie within the private sector, close cooperation between NSOs, on the one hand, and telecommunication operators and Internet companies, including search engines and social networks, on the other, is necessary and could be institutionalized through public-private partnerships.

Operators and Internet companies

Business interests will naturally provide operators and Internet companies with the incentive to talk to commercial vendors of big data analytics. In addition, operators and Internet companies can benefit greatly from engagement with academia and researchers to understand how to leverage big data for different purposes. Such engagement will also broaden their understanding of the limitations and assist them in the development of new methodologies, algorithms and software techniques that can be repurposed for business-use cases. Indeed, where the applications of data use for development are concerned, operators also have an interest in maximizing the economic well-being of their customer base.

Operators and Internet companies need to take advantage of their existing customer relationships to elicit a greater understanding of consumer concerns and needs in relation to privacy. They are well placed to develop a

privacy framework, in consultation with other stakeholders.

Given their business concerns, operators and Internet companies may hesitate to pool and share their data with those from other sources (including from competitors), but this is something that is worth exploring. Combining big data sources has great potential to increase added value and produce new insights. There is scope for exploring established models for such pooling – for example, the sharing by banks of some of their customer data with credit bureaux.

Governments

Governments have different opportunities and different roles to play in the exploitation of big data for monitoring and development. They can use big data to identify areas where rapid intervention may be necessary, to track progress and make sure their decisions are evidence based, and to strengthen accountability. More and more governments are recognizing the importance of big data and have set up communities of practice and working groups to study their use and potential impact (UNSC 2013).

Governments should also facilitate the legislative changes that are required and take a lead in setting big data standards. To this end, national regulatory authorities (NRAs) and NSOs, in consultation with other national stakeholders, are best placed to lead the corresponding discussions and bring together the relevant stakeholders.

In particular, NSOs, given their legal mandate to collect and disseminate official statistics and set statistical standards, have an important role to play. They could become standards bodies and big data clearing houses that promote analytical best practices in relation to the use of big data for complementing official statistics and for development. Those standards, which NSOs are in the best position to enforce, would also have to encompass best practices in relation to data curation and metadata standards. To this end, NSOs must also prioritize the upgrading of the in-house technical skills they require in order to handle big data, while at the same time investing in the necessary computational infrastructure.

As the main regulatory interface to the telecom sector, NRAs are well placed to co-champion the national discussion on how telecommunication big data may be leveraged for social good. Regulators have a role to play in facilitating the introduction of legislation that addresses privacy concerns while encouraging data sharing in a secure manner. The following recommendations were made in a recently published ITU draft paper (ITU, 2014):

- **Establishing mechanisms to protect privacy**: Regulators could develop a regulatory mechanism that would shift the focus of privacy protection from informed consent at the point of collecting personal data to accountable and responsible uses of personal data. This mechanism would foresee a well-resourced privacy regulator with the expertise and power to enforce such a use-based privacy protection mechanism. In return, data users would be permitted to reuse personal data for novel purposes where a privacy assessment indicates minimal privacy risks.

- **Restricting the use of probabilistic predictions**: While the use of big data can help better decision-making through probabilistic predictions, this information should not be used against citizens. Regulators should restrict the ways in which government agencies and others can utilize big data predictions.

- **Fostering big data competition and openness**: Regulators could foster big data competition in increasingly concentrated big data markets, including by ensuring that data holders allow others to access their data under fair and reasonable terms.

International stakeholders

International stakeholders – including UN agencies and initiatives (such as ITU and UN Global Pulse), the Partnership on Measuring ICT for Development, ICT industry associations and producers of big data (Google, Facebook, etc.) – have an important role globally. More work is needed to understand fully the potential of big data and examine the challenges and opportunities related to big data in the ICT sector. To this end, the key international stakeholders have to work together to facilitate the global discussion on the use of big data.

UN Global Pulse, as one of the main UN initiatives exploring the use of big data, can do much to inform and motivate the discussion on global best practices and the use of big data for development.

Where using big data for monitoring the information society is concerned, new partnerships, including public-private partnerships between data providers and the ICT statistical community, including ITU, could be formed to explore new opportunities and address challenges, including in the area of international data comparability and standards.

As one of the main international bodies working on issues related to the telecommunication sector, ITU could leverage its position to facilitate global discussion on the use of telecom big data for monitoring the information society.

Together, ITU and UN Global Pulse could facilitate the work that needs to be done by NRAs and NSOs, through awareness raising and engagement on privacy frameworks, data sharing, and analytical global best practices. ITU could help reduce the transaction costs associated with obtaining telecommunication big data, for example by facilitating the standards-setting process. Standardized contracts for obtaining data access as well as standards on how the data are stored, collated and curated can collectively reduce the overall transaction costs of accessing and leveraging telecommunication big data for social good.

Academia, research institutes and development practitioners

The research into how telecom data may be used to aid broader development is being done mainly by academia, public and private research institutes and, to a lesser degree, development practitioners. This makes them important stakeholders in defining the state of the art with respect to leveraging big data for development. They, more than others, have been the first to engage with telecommunication operators with a view to using their data for development. They therefore understand the potential and challenges from multiple perspectives. Their collective experiences will be valuable as big data for development becomes mainstreamed.