DRAFT

Mapping Big Data Solutions for the Sustainable Development Goals

Sriganesh Lokanathan, Thavisha Gomez, Shazna Zuhyle

March 2017

LIRNE*asia*
info@lirneasia.net | www.lirneasia.net

LIRNE*asia* is a pro-poor, pro-market think tank whose mission is *Catalyzing policy change through research to improve people"s lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology.*

Contact: 12 Balcombe Place, Colombo 00800, Sri Lanka. +94 11 267 1160. info@lirneasia.net
www.lirneasia.net

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Canada.

Authorship and acknowledgements

The report was prepared by a three-person team from LIRNEasia led by Sriganesh Lokanathan and comprised of Thavisha Gomez and Shazna Zuhyle. Comments by Prof. Rohan Samarajiva (Founding Chair, LIRNEasia) are gratefully acknowledged.

This research was carried out with the aid of a grant from the International Development Research Centre (IDRC), Canada.

Table of Contents

List of Tables

Executive Summary

The thrust to utilize big data for official statistics underscores the potential for generating new insights and complement existing measures. Big data can potentially revolutionize current official statistical systems in one of several ways[1]: a) Entirely replace existing statistical sources such as surveys; b) Partially replace existing statistical sources such as surveys; c) Provide complementary statistical information in the same statistical domain but from other perspectives; d) Improve estimates from statistical sources; and e) Provide completely new statistical information in a particular statistical domain. At the moment, complementing existing data is what offers the greatest potential for big data sources.

Research conducted to date has explored the utilization of different data sources for specific developmental applications. For example, the analysis of satellite imagery has typically been used to monitoring changes in topography including crop/yield estimation, drought monitoring, deforestation and carbon stock mapping among others. Mobile network big data has proven to be useful in understanding mobility patterns of the population, creditworthiness of its users, socioeconomic status of the population among others, while social media data is well positioned for sentiment analysis. There have been literature reviews (for example, Williams, 2016, UK"s Office of National Statistics; Lokanathan and Gunaratne, 2014) that have captured the statistical applications of big data, in particular mobile phone data.

The big data for development (BD4D) landscape hosts a range of players spanning government, industry, academia, and civil society among others. These players operate as policy actors, researchers, funders as well as intermediaries. Understanding the various actors at play offers greater opportunities for strategic partnership and collaborations. Thus, Goal 17 has been viewed through the lens of big data for development, stressing on the need for collaboration between various big data actors, and providing a snapshot of the BD4D landscape.

It is important to note that while insights derived from big data can be used to measure some SDG indicators, the potential for big data lies in being able to help achieve specific targets. For instance, while big data may not be particularly useful in calculating the death rate due to road traffic injuries (indicator 3.6.1), it can help identify hotspots for traffic accidents enabling authorities to take preventive action, thus contributing to the achievement of target 3.6 which is "by 2020, halve the number of global deaths and injuries from road traffic accidents."

Leveraging new as well as existing data sources (from both the public as well as private sectors) for the purposes of monitoring the progress towards the SDGs as well as for achieving them is not without challenges and requires the confluence of several factors. Developing economies in particular have much lower levels of "datafication" than developed economies, which means some of the most interesting and relevant data exists amongst the private sector as shown in this report. Accessing such data will not be without challenges, not least because in competitive industries such as the telecom sector, there would be competitive implications to sharing data. Even then there are costs and considerations associated with extracting and analyzing the data from private sector industries so as to sufficiently protect aspects related customer privacy/ confidentiality as well as protecting commercially sensitive business intelligence in competitive industries (e.g. the telecom sector). Innovation will have to unleash appropriate business models for leveraging these data sources. But innovation in this space will also require the confluence of a variety of actors such as

[1] See for example http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf

state, private, academia, and importantly non-governmental researchers and practitioners. New forms of partnerships between these actors are required not just because the true value comes from assembling different types of data from different sources, but also because of the inherent capacity challenges to make full use of the data. This is inherently a multi-disciplinary effort requiring computer scientists, statisticians, and domain/ subject-matter experts along with government officials.

As this report shows, It is important to remember that despite the rich body of literature and applications that already exist, the state of the art in innovative development focused applications of these new data sources is still very much in its embryonic stages. A new technique successfully applied for a specific context in a specific country or region doesn"t necessarily translate across time and space, especially when the underlying data that is being leveraged is behavioral data. A principal concern of the data revolution is about "counting the uncounted" and as such we need to pay particular attention to "representativity" of these new data sources that are being leveraged i.e. how accurately it reflects the population. Marginalization in the real world can often result in marginalization in the digital world beyond just issues of access to technologies, or being represented in the digitized data. As such localized testing of these new techniques with local experts (aware of ground truth and local context) are very important.

A consensus is yet to arise in addressing the privacy and ethical dilemmas that may arise from these new applications of data, but will continue to be important. It will be imperative that that laws, regulations, and guidelines are rooted in practical examples of harms rather than from imaginations so as to ensure that there public benefits are these data innovations are not greatly diminished. As datafication increases in society and digital platforms come to the fore, a related critical concern for the economy is the nature of competition in sectors. This may affect the type of data that is collected and shared and in particular the longer term success of the "data philanthropy" concept.

Introduction

The adoption of the 2030 sustainable development agenda, which seeks to "ensure that no one is left behind", places a strain on countries to report data that can be compared over time and are available at relatively frequent intervals. The 17 Sustainable Development Goals (SDGs) which will be monitored based on 169 targets, will in turn be measured by more than 230[2] targets that have been classified into three tiers based on methodology, standards and data available to measure/monitor them:

- **"Tier 1**: *Indicator conceptually clear, established methodology and standards available and data regularly produced by countries*
- **Tier 2**: *Indicator conceptually clear, established methodology and standards available but data are not regularly produced by countries*
- **Tier 3:** *Indicator for which there are no established methodology and standards or methodology/standards are being developed/tested."*

As of 21 December 2016, there were 83 Tier III indicators.[3]

This lack of data underscores the opportunities for non-traditional data sources to complement traditional statistics in the interval between official surveys, and develop new ways of monitoring the SDG targets. There is opportunity to capitalize on the use of non-traditional data sources such as mobile phone data, satellite-based technology and social media data among others to generate insights across a range of development issues.

This report has sought to capture the applications of new data sources, specifically big data sources to measure the sustainable development goals and relevant targets by reviewing relevant literature (both peer-reviewed and grey literature), and reports. Given the size and scope of research on the big data for development space, the references used in this report represent a sample of the research conducted.

Additionally, the report outlines the current concerns with the use of big data (privacy, marginalization, competition, etc.) and provide a discussion of the interplay of these issues so as to facilitate a conducive and sustainable environment for leveraging big data to achieve the global goals.

[2] While the number of indicators listed in the final proposal is 244, there are a few that are repeated across different targets. Hence, the number of unique targets is 232. http://unstats.un.org/sdgs/indicators/indicators-list/

[3] Tier Classification for Global SDG Indicators 21 December PDF. from http://docplayer.net/27317561-Tier-classification-for-global-sdg-indicators-21-december-2016.html

Big Applications for the SDGs

Goal 1: No Poverty

The United Nations" "No Poverty" goals seeks to "eradicate poverty in all its forms" through the achievement of seven key targets, each monitored by various indicators, making it imperative for countries to report poverty data that can be compared over time, and are available at relatively frequent intervals.

However, the dearth of poverty data in many countries presents significant challenges; for example, according to Serajuddin et al. (2015), of the 155 countries whose poverty data is monitored by the World Bank, more than 57 countries had either just one or no poverty estimates during the period 2002-2011. Moreover, with countries that did publish two poverty data points, 20 countries had more than a five-year interval between the estimates. Serajuddin et al. (2015) attribute the lack of household surveys to the lack of poverty estimates.

This lack of data underscores the opportunities for non-traditional data sources to complement traditional statistics in the interval between official surveys, and develop new ways of monitoring the SDG targets. Insights derived from big data sources may help to fill the gaps relating to poverty data, rather than replace official surveys. Mobile phone data for instance, can support this goal by helping to identify the poor--determine the socioeconomic status of the population, identify pockets of urban poverty, and estimate poverty rates--and by creating opportunities for greater financial inclusion among the poor.

TABLE 1: No Poverty

Possible Targets	Theme	Application	Data Source	References
1.1; 1.2	Socio economic status and wellbeing	Human mobility and socioeconomic levels	Mobile Phone Data	Frias-Martinez et al. (2012)
		Estimating poverty and wealth		Blumenstock et al. (2015)
		Socioeconomic status		Gutierrez et al. (2013)
	Poverty Mapping	Identifying the poor	Satellite Data	Elvidge et al. (2009); Jean et al. (2016)
		Urban poverty		Kohli et al. (2012)
1.4	Financial Inclusion	Creditworthiness of the unbanked	Mobile Phone Data	Kumar and Mohta (2012)
1.5	Disaster response	Human mobility after disasters		Lu et al. (2016); Wilson et al. (2016); Lu et al. (2012)

Key Big Data Sources

Mobile Phone Data, Satellite Data, Postal Data

Socioeconomic Status and Economic Well-being

Mobile Phone Data

Numerous studies (Frias-Martinez et al. 2012; Blumenstock et al. 2015) have sought to map variables derived from mobile phone data to make observations on the socioeconomic levels of populations. For instance, research conducted by Frias-Martinez et al. (2012) indicated that populations that had higher socioeconomic levels had a stronger linkage with larger ranges of mobility compared to populations that had lower socioeconomic levels. The mathematical model developed by Frias-Martinez et al. (2012) mapped mobility variables derived from information on Call Detail Records (CDRs) to socioeconomic levels. Among others, the results showed a higher correlation between socioeconomic levels and increases in the radius of gyration.[4]

Instead of identifying wealth and poverty at an aggregate level, Blumenstock et al. (2015) sought to understand how well mobile phone data could identify variables on an individual level by leveraging data from that individual"s digital footprint. A composite wealth index was developed based on the results of a phone survey with randomly selected subscribers (gathering information on the housing characteristics, asset ownership and other indicators) were combined with their mobile phone transaction data (after obtaining consent). The merged data of the respondents was used to show that it was possible to predict the wealth of mobile phone subscriber by the individual"s historical phone interaction data. This was then leveraged to make out-of-sample predictions for the rest of the subscribers (i.e. those who had not participated in the survey). While the model"s predictions (out-of-sample) at a district level have a strong correlation with data from the demographic and health survey for households owning a mobile phone (r=0.917) as well as all the households in the survey (r=0.916), a key caveat of this is the challenge in verifying the accuracy of data in micro regions given the lack of appropriate comparative data.

Moreover, CDR data combined with remote sensing data can also be leveraged to support the modeling of traditional measures of poverty at a more granular level and at more frequent intervals, and provide new insights on the distribution of poverty. For example, Steele et al. (2017) sought to model three traditional indicators of poverty using the two big data sources. CDR data included user mobility metrics, reload patterns as well as patterns of phone usage. Remote sensing data included metrics such as access to roads, weather, nighttime luminosity and other possible welfare-linked metrics. Results showed that models developed using both data sources had a stronger predictive power, with the most successful model being the reconstruction of the Demographic and Health Survey Wealth Index to predict poverty (r^2 = 0.76)

In addition to mobility data from CDRs, other research leverage mobile credit data to infer wealth patterns of users. For instance, Gutierrez et al. (2013), leveraged airtime credit purchase records and communication data of subscribers to understand socioeconomic levels based on the hypothesis that mobile users who make larger purchases of airtime credit would be more affluent compared to those who made multiple smaller purchases. However, the researchers emphasized that there was a lack of data to develop a predictive model that leveraged both variables, size and frequency.

[4] The radius of gyration Is a measure of how far an object travels from its center of gravity. In the case of humans, the radius of gyration roughly measures the typical range of a mobility of user in space.

Satellite Data

It must be noted that the use of nighttime light data proxies pre-dates the current big data phenomenon. Studies that leverage satellite image analysis for poverty estimation cut across key features that include nighttime and daytime imagery, regional and national level data, the time period of observation, as well as the resolution of images among others. In addition to estimating GDP (Elvidge, et al. 1997; Sutton, et al. 2007; Ebener, et al. 2005), studies have also sought to estimate poverty in general (Elvidge et al. 2009) or urban poverty (Kohli et al. 2012).

For instance, Elvidge et al. (2009) developed a poverty map by utilizing a poverty index calculated using population count and the brightness of nighttime lights as observed through satellite images. The model took into account outlines of human settlement, topography as well as land cover. The results that were derived were calibrated with national level poverty data. The estimate derived was a little lower than estimates from the World Development Indicators. The study suggested that stronger reference data for calibration purposes and improvements in satellite observations would enhance the results of the poverty map.

Kohli, et al. (2012) developed a generic framework that leverages information across three levels of the constructed environment to support the satellite image-classification of slums: object level, (road/building characteristics) settlement level (planned vs. unplanned) and the environ level (disaster prone regions).

However, a key caveat of using nighttime satellite imagery to estimate poverty is the challenge of differentiating regions that are already at the lower end of income distribution since satellite imagery in these areas would be dark (Jean et al. 2016). Jean et al. (2016) used both daytime and nighttime satellite imagery as well as survey data and machine learning to identify varying economic well-being levels in Malawi, Nigeria, Rwanda, Tanzania and Uganda. The computer model was trained to identify features of daytime satellite images that were predictive of poverty. Imagery from countries that already had survey data was used to validate the findings of the model.

Financial Inclusion

Mobile Phone Data

In 2012, the Consultative Group to Assist the Poor (CGAP) suggested that analyzing mobile phone consumption data (Kumar and Muhota, 2012) could identify the creditworthiness of the unbanked. For example, it was suggested that purchasing airtime credit in a consistent and frequent manner showed income predictability as well as being able to plan ahead, both positive indicators of loan repayment ability as opposed to those with inactive prepaid accounts or those that would run out of credit regularly.

Similarly, with a presence in over 20 countries, Tiaxa serves the unbanked in emerging markets, offering 10 million small loans daily in the form of cash (through Mobile Money) and as airtime credit (prepaid mobile phone subscriptions) Tiaxa"s Balance Advance feature which relies on numerous process including user behavior analytics of mobile phone users to determine amounts to advance to each user, is currently being utilized by 8 operators in Latin America.

Other

Big data startup, DemystData leverages numerous sources of big data—including telecommunications information, social, ID, fraud, websites, text, news, and logs—to assess the creditworthiness of customers. Its software integrates telecommunication data and social and

corporate data to develop risk profiles for small businesses or individuals, enabling banks to get a better sense of lending risks.

Disaster Response

Mobile Phone Data

There is a growing body of research that leverages mobile phone data to understand disaster response that include understanding human mobility patterns after disasters such as cyclones (Lu et al. 2016) and earthquakes (Wilson et al 2016; Lu et al. 2012). This has been described in greater detail in Goal 11.

Insights from Big Data vs. Traditional Indicators

Target 1.1 *By 2030, eradicate extreme poverty for all people everywhere, currently measured as people living on less than $1.25 a day.*

There is one indicator to measure this target:

- 1.1.1 Proportion of population below the international poverty line, by sex, age, employment status and geographical location (urban/rural).

Target 1.2 *By 2030, reduce at least by half the proportion of men, women and children of all ages living in poverty in all its dimensions according to national definitions*

There are two indicators to measure the achievement of this target

- 1.2.1 Proportion of population living below the national poverty line, by sex and age
- 1.2.2 Proportion of men, women and children of all ages living in poverty in all its dimensions according to national definitions

As discussed above big data sources such as mobile data and satellite data can be used to infer the socio economic status of population enabling authorities to identify pockets of poverty at more granular levels and at more frequent intervals, than is possible using traditional methods. Thus, insights derived from big data sources can feed into the achievement of target 1.1 and 1.2.

Target 1.4 *By 2030, ensure that all men and women, in particular the poor and the vulnerable, have equal rights to economic resources, as well as access to basic services, ownership and control over land and other forms of property, inheritance, natural resources, appropriate new technology and financial services, including microfinance*

Two indicators measure this target:

- 1.4.1 Proportion of population living in households with access to basic services
- 1.4.2 Proportion of total adult population with secure tenure rights to land, with legally recognized documentation and who perceive their rights to land as secure, by sex and by type of tenure

While the traditional indicators focus on access to basic services and tenure rights to land, big data sources such as mobile phone data have the potential to facilitate financial inclusion by helping to assess the credit worthiness of the unbanked.

Target 1.5 *By 2030, build the resilience of the poor and those in vulnerable situations and reduce their exposure and vulnerability to climate-related extreme events and other economic, social and environmental shocks and disasters*

Three indicators are set to measure this target:

- 1.5.1 Number of deaths, missing persons and persons affected by disaster per 100,000 people
- 1.5.2 Direct disaster economic loss in relation to global gross domestic product (GDP)
- 1.5.3 Number of countries with national and local disaster risk reduction strategies

Of these indicators, mobile phone data can contribute towards indicator 1.5.1, which is multi-purpose indicator with linkages to other SDG targets including 11.5, 13.2, 1.3, 14.2, 15.3, 3.9, and 3.6 among others[5]. While mobile phone data may not be able to directly measure this indicator, it can help contribute towards the measurement) by helping identify the movement of people after a disaster. Moreover, satellite data, analyzed quickly, can contribute to the measurement of 1.5.2.

Goal 2: Zero Hunger

With the UN projecting world population to rise to 8.5 billion by 2030 and to 9.7 billion by 2050[6], the second SDG aims to "end hunger, achieve food security and improved nutrition and promote sustainable agriculture." However, agricultural regions are susceptible to extreme weather conditions such as droughts, which severely affects their ability to contribute towards food supply. This makes it crucial for decision makers to be aware of the location and potential of drought conditions to provide targeted assistance. Analysis, both spatial and temporal, of satellite data could help determine both the severity and the extent of the drought in near real-time, as well as support the estimation of crop production.

TABLE 02: Zero Hunger

Possible Targets	Theme	Application	Data Source	References
2.1	Expenditure on Food	Proxy indicator for food expenditure	Mobile Phone Data	Decuyper et al. (2014)
2.1	Drought monitoring	Severity and extent of drought conditions	Satellite Data	Berhan et al. (2011); Tucker & Choudhury (1987); Henricksen & Durkin (1986)
2.4	Early crop yield assessment	Developing vegetation health indices		Kogen et al. (2011)
2.c.	Price Indexes	Constructing consumer price index	Online Prices	Cavallo & Rogobon (2016)

Big Data Sources
Satellite Data, Mobile Phone Data, Web Scraped Data

Proxy Indicator for Food Expenditure

Mobile Phone Data
Recent studies have sought to understand how suited mobile phone data derived indicators were to act a proxy for indicators of food security (Decuyper et al. 2014). Researchers compared mobile

[5] https://unstats.un.org/sdgs/files/metadata-compilation/Metadata-Goal-13.pdf
[6] UN projects world population to reach 8.5 billion by 2030, driven by growth in developing countries. (n.d.). Retrieved from http://www.un.org/sustainabledevelopment/blog/2015/07/un-projects-world-population-to-reach-8-5-billion-by-2030-driven-by-growth-in-developing-countries/

phone (activity and purchase of airtime credit) data from an East African country with a nationwide household survey that was conducted by the World Food Programme. The researchers found a high correlation (r>0.7) between purchases of airtime credit and the results from the survey for several food items. Given that this high correlation was witnessed for food items that were mainly market-dependent and not for those items grown at home, it can serve as a proxy indicator of food expenditure of "market-dependent" households.

Social Media Data

Similarly, the results of a study conducted by UN Global Pulse (2014) indicated a relationship between past statistics on food inflation and the volume of tweets regarding increases in food prices. Researchers analysed tweets prices relating to food prices in Indonesia over March 2011 to April 2013 by developing taxonomies relating to the issue, relying on a classification algorithm to categorize the results. A time series analysis was then run to identify the correlation between official food inflation statistics and food-related tweets as well.

Drought Monitoring

Satellite Data

There are many studies (for example, Unganai & Kogan, 1998; Peters et al. 2002; Wan, Wang & Lee, 2004; Rhee & Carbone, 2010) that have been conducted on the use of remote sensing technology for detecting drought conditions. For instance, a study conducted by Henricksen, & Durkin (1986) leveraging radiometer data (Advanced Very High Resolution Radiometer - AVHRR) derived from meteorological satellites found a strong correlation (r=0.99) between the "rates of change of the normalized difference vegetation index (NDVI) derived from the AVHRR data and threshold values of a soil moisture index at the beginning and ends of growing periods". Tucker & Choudhury (1987) concluded that "multi-temporal satellite data have application in the detection and quantification of drought through the ability of these data to estimate the photosynthetic capacity of the terrestrial surface and record microwave surface brightness." Moreover, Wang & Qu (2007) proposed a new Normalized Multi-band Drought Index to monitor vegetation and soil moisture based on satellite sensors.

Crop Management/Yield Estimation

Satellite Data

In addition to monitoring topographical changes, satellite data can also be used to estimate crop yields. For example, Unganai & Kogan (1998) leveraged a previously developed vegetation condition index based on AVHRR data and a temperature condition index to assess drought conditions in Southern Africa with results validated through on the ground data. The results suggested that the two indices could be used to develop scenarios of corn yield 6-13 weeks ahead of harvesting.

Similarly, studies such as those conducted by Kogan et al. (2011) used vegetation health indices based on AVHRR sensors to provide cumulative estimates on a weekly basis for a range of indicators including thermal and moisture conditions of canopy for the relevant season. Calculated for 1981–2010 period, the indices were compared with yields in 24 countries and demonstrated a strong correlation between the vegetation health indices and corn, wheat, sorghum and soybean yields during the period indicating its potential to acts as a proxy for early crop yield assessment.

Moreover, the Group on Earth Observation Global Agricultural Monitoring (GEOGLAM)[7] uses a combination of ground-data and satellite images to provide agricultural production forecasts at global, regional as well as national levels. The Group on Earth Observations (GEO) developed the GEOGLAM initiative, which was endorsed by the G20 in June 2011. GEOGLAM seeks to "coordinate satellite monitoring observation systems in different regions of the world in order to enhance crop production projections and weather forecasting data," and to that extent, developed a crop monitor to provide the Agricultural Market Information System with an assessment of global crop production conditions for four key crops: wheat, maize, rice, and soy for the G20 countries plus Spain and 7 other countries. The Earth Observation data is used to assess evapotranspiration, rainfall, soil moisture, temperature and the Normalized Difference Vegetation Index.

Private Sector

US-based Climate Corporation[8] offers farmers a mobile software-as-a-solution that provides weather simulation information and provides farmers with estimates for crop yields based on daily weather data over the past months. Weather measurement information obtained from around 2.5 million locations daily is processed with 150 billion soil observations to provide farmers with 24-hour and seven-day forecasts of rain, wind and temperature for specific areas (200 acres), enabling farmers to make more informed decisions regarding planting and harvesting crops.

Insights from Big Data vs. Traditional Indicators

Target 2.1 *By 2030, end hunger and ensure access by all people, in particular the poor and people in vulnerable situations, including infants, to safe, nutritious and sufficient food all year round*

Big data sources may not be able to directly measure the two proposed indicators to measure the progress towards the achievement of this target

- 2.1.1 Prevalence of undernourishment
- 2.1.2 Prevalence of moderate or severe food insecurity in the population, based on the Food Insecurity Experience Scale (FIES)

However, the analysis of big data sources can provide insights on food security that could be vital when serving populations in vulnerable situations. For instance, crop yield estimation enables governments to better prepare for any projected shortages.

Goal 3: Good Health and Wellbeing

The third goal of the SDGs seeks to “ensure healthy lives and promote well-being for all at all ages.” In particular there is a focus on ending epidemics such as malaria, neglected tropical diseases and communicable diseases.

The use of mobile phone data, particularly CDR, has been used in many cases to study the spread of mosquito-borne diseases such as dengue and malaria. Analysis of mobility data for mosquito-borne disease propagation relies on the premise that the disease could be spread in one of three ways: (1) Infected parasites moving to a region, (2) Infected humans visiting a region and (3) residents visiting a region with the disease and becoming infected and returning to the region. Given that mosquitos fly within a small radius, the principal means of epidemic spread is through human carriage. The underlying theory is the fact that social interaction and thereby mobility facilitate the spread diseases. By combining movement patterns with data of reported cases (e.g. malaria) areas of

[7] See: https://www.earthobservations.org/geoglam.php

[8] The company is a subsidiary of Monsanto Company. See: https://www.forbes.com/sites/bruceupbin/2013/10/02/monsanto-buys-climate-corp-for-930-million/#55db1d1d177a

transmission risk can be prioritized. Such risk maps can be used to aid targeted interventions and to curtail the spread of such diseases.

Numerous studies (Tatem et al. 2009; Bengstton et al. 2011; Wesolowski et al. 2014; Ruktanonchai et al. 2016) have leveraged mobility variables derived from mobile phone data to understand the spread of diseases such as Malaria (Tatem et al. 2009; Tatem et al. 2014; Ruktanonchai et al. 2016), Ebola (Wesolowski et al. 2014), Rubella (Wesolowski et al. 2015), Dengue (Wesolowski et al. 2015), and Cholera (Bengtsson et al. 2011) among others.

TABLE 03: Good Health and Well-being

Potential Target	Theme	Application	Data Source	References
3.3	Disease Propagation	Mobility from regions of disease outbreak	Mobile Phone Data	Wesolowski, et al. (2015); Bengston et al. (2015); Wesolowski et al. (2014)
		Sources and sinks for diseases		Ruktanonchai et al. (2016); Tatem et al. (2014)
		Disease Importation rate		Tatem et al. (2009)
		Seasonal trends of diseases	Search Engine Data	Schuster et al. 2010; Yang et al. 2010; Xu et al. 2010

Big Data Sources
Social Media Data, Search engine query Data, Mobile Phone Data, Sensor Data

Disease Propagation

Mobile Phone Data
Tatem et al. (2009) leveraged anonymized CDR data from mobile operator Zantel to gauge the malaria importation rate in Zanzibar. The researchers studied the mobility patterns of population in Zanzibar travelling to mainland Tanzania, extracting information such as length of stay and locations visited by analyzing CDRs over a three-month period. This was combined with malaria risk data (based on where and how long travellers stayed) to develop a mathematical model to estimate a malaria importation rate.

Tatem et al. (2014) demonstrated the manner in which human mobility connected the areas of malaria transmission risks. Researchers leveraged satellite data, CDRs and surveillance data, integrating movement data with case-based risk maps. Researchers leveraged anonymised call detail records by a leading mobile service operator, aggregated at a cell tower level to estimate the movement of two groups of travellers: "returning residents" – those who had visited an area of malaria risk and returned home, and "visitors" – residents of areas that had malaria risk visiting other locations. The researchers were able to identify area that were exporters of the diseases and other areas that served as sinks.

Wesolowski et al. (2014) further studied the use of CDRs in the Ebola outbreak context. They developed models estimate travel between locations based on the distance between locations and the size of the population at the locations, using CDRs to understand national mobility patterns in Kenya, Cote d'Ivoire and Senegal (in the absence of mobile data from the countries that were affected by Ebola at the time). The models would help identify the relative amount of traffic between the different populations as well as identify the more popular routes of travel. Given that mobility of

people is a key driver of the disease, understanding the movement of the population within the country would enable the delivery of targeted control policies by identifying the possible routes taken by infected individuals.

Similarly, Wesolowski et al. (2015) inferred mobility patterns of population from mobile phone data from almost 15 million subscribers in Kenya quantifying the patterns of seasonal travel and combined this with rubella incidence data to predict the transmission of the disease. Moreover, the combination of spatial and seasonal travel data enables the characterization of risk fluctuations seasonally.

Wesolowski, et al. (2015) developed an epidemiological model for transmission of dengue in travelers by leveraging climate information and mobile data from around 40 million subscribers to help predict the propagation of dengue in Pakistan. They analysed historical (2013) dengue case data that was spatially explicit and compared the epidemic dynamics with epidemiological model. The results derived showed that mobility estimates based on mobile data could be used to predict the timing and the geographic spread of diseases, not just in recently epidemic locations, but also in emerging areas. Moreover, the researchers were able to generate dynamic risk maps that could be leveraged to contain the disease.

Bengtsson et al. (2011) leveraged mobile phone data from Digicel Haiti to assess the movement of mobile phones and thereby estimate the movement of population from regions with a Cholera outbreak. This information was shared with relief agencies that were mobilizing in response to the outbreak. The results from the analysis were later compared with actual cases at a district level showing that mobility patterns derived from mobile phone data were able to predict the spread of the epidemic better than standard population mobility models. Moreover, it was also revealed that there was correlation between mobile phone data and the magnitude of an outbreak.

Similarly, Ruktanonchai et al. (2016) developed a mathematical model that leveraged call detail records of mobile phone users in Namibia and malaria parasite rate maps to estimate regions of "self-sustaining malaria transmission". The researchers modified a model developed by Cosner et al. (2009) in a manner that would enable them to parameterize human mobility based on mobile data. The researchers were able to identify areas that acted as sources and others that acted as Malaria sinks enabling authorities to conducted targeted elimination programs.

Search Engine Data

There is also evidence (Schuster et al. 2010; Yang et al. 2010; Xu et al. 2010) to show that search engine queries on symptoms and pharmaceuticals are related to seasonal trends of a vast spectrum of health related issues from specific diseases to conditions such as sleep disorders. The studies referenced in this report all used Google trends as their data primary data source leveraging other data sources such as climate related data (Yang et al., 2010), revenue of pharmaceuticals (Schuster et al. 2010) to determine correlations. The studies show that the analysis of search engine data along with other relevant data sources can be used to map spatial and geographical spread of, for example, communicable diseases, which in turn can be used for better risk management.

Hotspots for Road Traffic Accidents

Target 3.6 focuses on the reduction of deaths and injuries from road traffic accidents. The use of historical accident data coupled with other variables such as social events and weather can be used to identify potential areas of accidents, enabling authorities to take preventative action. For instance, Tennessee Highway Patrol (THP) leveraged IBM analytics to reduce road traffic accidents

by developing a model to predict future accidents based on historical data: weather data, geotagged crash data, geo-tagged data on driving under influence (DUI) as well as data around special events such as parades and games. The data was used to identify correlations between accidents and DUI arrests, and external influences such as events, weather, location, time of day/year and day of week etc. Based on this analysis, the model could extrapolate to predict hotspots for future incidents.

Insights from Big Data vs. Traditional Indicators

Target 3.3 *By 2030, end the epidemics of AIDS, tuberculosis, malaria and neglected tropical diseases and combat hepatitis, water-borne diseases and other communicable diseases*

There are five indicators under this target that measure the incidence of various epidemics offering a static picture.

- 3.3.1 Number of new HIV infections per 1,000 uninfected population, by sex, age and key populations
- 3.3.2 Tuberculosis incidence per 1,000 population
- 3.3.3 Malaria incidence per 1,000 population
- 3.3.4 Hepatitis B incidence per 100,000 population
- 3.3.5 Number of people requiring interventions against neglected tropical diseases

Meanwhile the analysis of big data such as mobile phone data can contribute to the achievement of target 3.3 by helping to understand the spread of mosquito borne diseases such as malaria enabling the provision of targeted care and preventive measures.

Target 3.6 *By 2020, halve the number of global deaths and injuries from road traffic accidents*

While big data may not be useful to compute the indicator for this target (3.6.1 Death rate due to road traffic injuries) it can help identify hotspots for traffic accidents enabling authorities to take preventive action, thus contributing to the achievement of target 3.6

Goal 4: Quality Education

The achievement of Goal 4 is the key to achieving many of the other SDGs.[9] While there has been progress towards improving literacy skills and increasing school enrolment rates, universal education goals are yet to be met. For instance, over 100 million youth do not possess basic literacy skills[10].

TABLE 04: Quality Education

Possible Targets	Theme	Application	Data Source	References
4.6	Illiteracy Prediction	Areas of low literacy	Mobile Phone Data	Sundsøy, P. (2016)

Big Data Sources Used

Mobile Phone Data

Literacy Prediction

[9] See: QUALIT Y EDUCATION: WHY IT MATTERS. (n.d.). Retrieved from http://www.un.org/sustainabledevelopment/wpcontent/uploads/2017/02/ENGLISH_Why_it_Matters_Goal_4_QualityEducation.pdf

[10] Education - United Nations Sustainable Development. (n.d.). Retrieved from http://www.un.org/sustainabledevelopment/education/

Mobile Phone Data

Sundsøy, P. (2016)[11] developed a machine learning algorithm to predict illiteracy by analyzing mobile phone records in a developing country in Asia to derive indicators from mobile phone data that could reflect the socioeconomic and mobility patterns of users. A survey of 76,000 mobile phone users was conducted to validate the study and gather information on literacy. The model showed 70% accuracy with the study also showing that individual illiteracy can be aggregated at a cell tower level. The results indicate predictors of illiteracy: social networks (there is less diversity among social contacts for those who were illiterate), location – areas of poor development could be identified.

Trends in Education

Massive Open Online Courses

The advent of Massive Open Online Courses (MOOC) offer opportunities to glean insights on the vast network of users registered on their sites including information on demographics, geography, subjects of engagement and effectiveness of various modes of learning that have the potential to inform policy relating to education. HarvardX and MITx released findings (Ho et al, 2014; Ho et al. 2015) from their open courses launched through the edX platform.

Insights from Big Data vs. Traditional Indicators

.

Target 4.6 *By 2030, ensure that all youth and a substantial proportion of adults, both men and women, achieve literacy and numeracy*

This target is proposed to be measured by one indicator:

- 4.6.1 Percentage of population in a given age group achieving at least a fixed level of proficiency in functional (a) literacy and (b) numeracy skills, by sex.

The targets for this goal largely revolve around access to education, formal and non-formal spanning pre-primary to tertiary education While there have been efforts to predict illiteracy using mobile phone data, there appears to be limited research on the use of big data to measure these targets.

Goal 5: Achieve Gender Equality and Empower all Women and Girls

Gender inequality poses significant ramifications to all areas of society including reducing poverty, to improving health and education among others. In order to ensure that "no one is left behind", gender disaggregated data are required so that all population groups are addressed.

TABLE 05: Gender Equality

Possible Targets	Theme	Application	Data Source	References
5.1	Gender Prediction	Gender prediction	Mobile Phone Data	Sundsøy et al. (2015); Blumenstock & Eagle (2010)

Big Data Sources Used
Mobile Phone Data, Social Media

Gender Prediction

Mobile Phone Data
The development of models that could predict the gender of anonymous mobile phone users presents significant opportunities for decision-makers to leverage existing models that assess socioeconomic status and geography of mobile phone users to conduct targeted gender-specific development initiatives and develop informed, evidence-based policies.

Sundsøy et al. (2015)[12] have utilized mobile phone data and machine learning to derive mobile phone user's gender based on their phone usage relying on indicators that included average delay in response and the radius of gyration. This information, combined with the ability to identify socio-economic status of phone users would enable decision-makers to identify key characteristics of women in impoverished areas. Moreover, in their analysis of mobile phone usage and access patterns in Rwanda, Blumenstock & Eagle (2010) noted differences by gender that were statistically different, for example in terms of ownership it was more likely that women used shared phones.

Pilot Projects – Mobile Phone Data; Twitter Data
Data2x has been supporting numerous pilot projects to answer gender questions leveraging big data sources such as social media, call detail records and satellite data. For instance, it is supporting the Flowminder Foundation to improve the spatial resolution of information that already exists from standard surveys (for example, indicators on freedom of movement, access to contraceptives, malnutrition etc.) by leveraging satellite data.

Similarly, its partner UN Global Pulse sought to sex-disaggregate tweets on development topics between the period May 2012 to July 2015. The 350 million development-related tweets were yielded over the course of this period by filtering for 25,000 keywords (in Spanish, Portuguese, French, and English) related to 16 development areas and relying on an algorithm a combination of name, url and twitter us/id to identify the gender of the user.

Insights from Big Data vs. Traditional Indicators
There appears to be limited research on the generation of gender-disaggregated insights from big data sources. This is possibly explainable in terms of pseudonymization of data that is caused by concerns over privacy. Usually ability to identify gender is also removed when personally identifiable information (PIP) are masked.

Goal 6: Clean Water and Sanitation

While access to water and sanitation is a human right, according to the UN, over 40 percent of the global population is affected by water scarcity and over 600 million do not have access to improved sources of drinking water.[13] In addition to the provision of universal access to drinking water, it is

[13] Water and Sanitation - United Nations Sustainable Development. (n.d.). Retrieved from http://www.un.org/sustainabledevelopment/water-and-sanitation

imperative for countries to be proactive in conserving their water-related ecosystems and improve water-use efficiency.

Numerous studies (Haas et al. 2009; Lu et al. 2011; Pekel et al. 2014; Rokni et al. 2014; Mueller et al. 2016) have been conducted that revolves around mapping surface water using remote sensing data. Insights derived from such studies could feed into understanding the changes in the global water-related ecosystem, assess the level of water stress and help drive evidence-based policymaking to protect and restore ecosystems.

Moreover, in relation to improving water use efficiency, the advent of smart water meters creates opportunities to regulate the supply of water by leveraging predictive analytics, generating water consumption patterns, detecting suspicious behaviors such as leaks and identifying peak demand.

TABLE 06: Clean Water and Sanitation

Possible Targets	Theme	Application	Data Source	References
6.6	Changes in water-related ecosystem	Change in surface water	Satellite Data	Mueller et al. (2016); Haas et al. (2009); Rokni et al. (2014); Pekel et al. (2014);

Big Data Sources
Smart Meter Data, Satellite Data

Mapping changes in water-related ecosystem

Satellite Data
There have been numerous studies that have sought to map changes in surface water using various techniques. Rokni et al. (2014) relied on satellite data and applied the Normalized Difference Water Index (NDWI) to model the Lake Urma's (Iran) spatiotemporal changes over the period between 2000 and 2013. The results of the study showed a trend of decreasing surface water over the period analysed. Similarly, Pekel et al. (2014) conducted an experiment of the African continent and proposed an approach to provide continental dynamic information on water surface detection using a " generic multi-temporal and multi-spectral image analysis method" and the product yielded an accuracy of 91.5%.

Mueller et al. (2016) relied on satellite data spanning 27 years to map surface water in Australia. The researchers used a decision tree classifier-based algorithm for water detection as well as a comparison methodology. A model on a pixel-by-pixel basis was utilized whereby each pixel was classified as water or not-water based on regression tree that was trained on sample of water and non-water tiles that represented the various landscapes across Australia. The results enabled to differentiate areas where water was persistent from areas where water was collected for a short time (such as areas under flood).

Researchers at the European Joint Research Center are utilizing Google Earth Engine [Include reference] to map 30 years of surface water occurrence globally. The study analyses over 2.8 million landstat images spanning the period from March 1985 to March 2015 using sensor neutral methodology. They are relying on a pixel-based approach with spatiotemporal validation based on 20,000 validation pixels with overall accuracy above 90 percent.

Similarly, NASA's Gravity Recovery and Climate Experiment (GRACE) provides the means to identify locations of freshwater availability via two satellites that identify water gain and loss. It has been used to generate maps that show groundwater loss in California for a particular period of time and the scarcity of freshwater in other regions of the United States.

Insights from Big Data vs. Traditional Indicators

Target 6.6 By 2020, protect and restore water-related ecosystems, including mountains, forests, wetlands, rivers, aquifers and lakes

- 6.6.1 Change in the extent of water-related ecosystems over time

The one indicator to measure this target proposed the use of earth observation data to compute the indicator. Thus, satellite data can be used alongside ground data to estimate change in extent of freshwater systems[14].

Goal 7: Access to Clean and Affordable Energy

With around 20 per cent of the world's population lacking access to modern energy services, this goal is impacts numerous sectors including business, agriculture, information technology and infrastructure among others[15]. This is further complicated by the fact that energy remains the key contributor to global climate change.[16] This underscores the need for access to clean energy.

Big data can be particularly useful in understanding electrification (and thus also urban/ rural differences). For example, there have been numerous studies regarding the use of nighttime satellite data for various measures of social and economic indicators including electrification rates (Elvidge et al. 1997). Similarly, the use of big data and machine learning can help address the problem of intermittency problems of renewable energy sources such as wind and solar.

The deployment of smart meters offers utilities so many data points that can potentially offer insights into consumer demand, particularly as it captures information such as meter status, quality of power and electricity consumption. In addition to providing demand-side data, smart meters are important in effective demand-side management, which allow for more efficient use of generation capabilities, thereby reducing costs of energy enabling provision to more people.

TABLE 07: Affordable and Clean Energy

Possible Targets	Theme	Application	Data Source	References
7.1	Access to electricity	Nighttime luminosity	Satellite Data	Elvidge et al. (1997); Elvidge et al. (2009); Townsend & Bruce (2010); Doll & Pachauri (2010); Chen & Nordhaus (2011)

[14] https://unstats.un.org/sdgs/files/metadata-compilation/Metadata-Goal-6.pdf
[15] http://www.un.org/sustainabledevelopment/wp-content/uploads/2016/08/7_Why-it-Matters_Goal-7_CleanEnergy_2p.pdf
[16] http://www.un.org/sustainabledevelopment/energy/

7.1	Residential electricity consumption	Determinants of electricity consumption	Smart Meter Data	Kavousian et al. (2013)

Big Data Sources
Smart Meter data, Satellite Data

Access to Electricity
Grid electricity is supplied by electricity distribution companies who maintain records for billing purposes. Therefore, supply-side data is available. However, in some countries there are illegal connections, which may be estimated by other means. Households may have electricity from non-grid sources, which can be estimated from satellite and other data.

Satellite Data
There have been many examples of leveraging satellite-based technology to understand electrification rates and have been discussed in detail in Goal one. For instance, Townsend & Bruce (2010) sought to use satellite imagery data between 1997 and 2002 to estimate the geographic distribution of electricity usage across Australia. The results yielded a strong correlation ($R^2 = 0.9346$) between a state's consumption of electricity and nighttime luminosity. The researchers developed a model to address the overglow effect (i.e. the spread of light to neighbouring areas) by leveraging the relationship between the strength of the light source and distance of dispersion from the source to estimate electricity consumption at a more granular level enabling the estimation of statistical local area (greater than $10km^2$) electricity consumption.

Moreover, Doll & Pachauri (2010) sought to identify areas in developing countries with no access to electricity using nighttime satellite imagery and population datasets. The study compared access to electricity between 1990 and 2000 to derive estimates of population without access to electricity based on satellite data.

Electricity Consumption

Smart Meter Data
Kavousian et al. (2013) use smart meter data to examine the determinants of residential consumption of electricity. The researchers collected data from over 1600 households in the US. The researchers developed models for daily minimum and maximum electricity consumption. The study found that the key determinants of electricity use included floor area, weather and location with little correlation found between consumption of electricity and the level of income. An important caveat to note is that the participants of this study were employees of a Silicon Valley based technology company, with over half reporting incomes over USD 150,000.

Private Sector
Big data solutions such as the one offered by IBM (Hybrid Renewable Energy forecasting solution - HyRef) that leverages sensors, analytics technology, cameras etc. to generate weather forecasts in areas of wind farms even 30 days in advance could assist in better forecasting the energy from the wind farms that can be directed towards the power grids.

Similarly, US-based startup Autogrid offers a solution that analyses energy data such as data collected from generators and transformers, electricity consumption to offer power and utility

companies services such as trends in energy usages, predictions and grid device performance management.

Insights from Big Data vs. Traditional Indicators

Target 7.1 By 2030, ensure universal access to affordable, reliable and modern energy services
There are two indicators to measure this target:

- 7.1.1 Proportion of population with access to electricity
- 7.1.2 Proportion of population with primary reliance on clean fuels and technology

Of these, satellite data can be used to measure indicator 7.1.1.

Goal 8: Decent Work and Economic Growth

The availability of stable jobs is a vital component in helping the nearly 2.2 billion people around the world who live below the poverty line (of USD 2)[17]. Thus SDG 8 is linked with the overall objective of eradicating poverty and re-aligning work, economic and social policies to increase job creation, economic productivity etc. One of the key objectives is to create the necessary conditions for sustainable economic growth and to increase access to decent working conditions for the whole age range of the working population and for equal opportunities for men and women, persons with disabilities, inter alia. Big data, particularly mobile phone data can help assess the socio economic status of populations (Frias-Martinez et al. 2012; Blumenstock et al. 2015) enabling governments to provide targeted initiatives for the provision of jobs. Similarly, there have been studies that have leveraged search engine query data to glean unemployment trends (Xu et al. 2013). Access to finance is also a key concern for those at the bottom of the pyramid and mobile phone data can play a role here in helping to assess the creditworthiness of the unbanked (Kumar and Mohta, 2012).

TABLE 08: Decent Work and Economic Growth

Possible Targets	Theme	Application	Data Source	References
8.1	GDP	GDP and Human Development	Postal Data	Hristova et al. (2016)
8.5	Unemployment	Unemployment trends	Search Engine Data	Xu et al., (2013)
8.9	Tourism	Destination of tourists	Mobile Phone Data	Ahas et al. (2008)
8.9	Tourism	Seasonal tourism	Mobile Phone Data	Ahas et al. (2007)
8.1	GDP	Economic development	Satellite Data	Elvidge et al. (1997); Sutton and Constanza (2002); Ebener et al. (2005); Chen and Nordhaus (2011);

[17] http://www.un.org/sustainabledevelopment/economic-growth/

Big Data Sources
Mobile Phone Data, Web Scraped Data, Search Engine Data, Postal Data

Price Indexes

Web Scraped Data
The Billion Prices Project, an academic initiative of the Massachusetts Institute of Technology, conducts economic research that leverages price data collected daily from online retailers globally. They have published numerous research papers based on their analyses that could contribute towards measuring/monitoring goal 08. For instance, in their study, Cavallo and Rogobon (2016) work with the Billion Prices Project to "construct daily price indexes" using web-scraping software to collect online prices from large multichannel retailers (with both an online and offline presence) focusing on products that were included in a typical basket for official consumer price index and had consumer expenditure weights. Similarly Cavallo (2013) calculated the annual inflation rate for Argentina for the period (2007-2011) using online pricing data from key retailers' websites.

Economic Development

Satellite Data
Nighttime luminosity for instance has been used as a proxy for economic development at a country level (Elvidge et al. 1997) with studies attempting to estimate GDP at a sub-national level (Sutton et al. 2007; Ebener et al. 2005).

For instance, a study by Elvidge et al. (1997) found a correlation (R^2=0.97) between luminosity and GDP across 21 countries. The study also revealed significant outliers in relation to the luminosity and population indicating that the population distribution across geographies should also factor in economic development at the local level. Similarly, Doll et al. (2000) attempted to develop global maps using the relationship between GDP and area lit analyzing nighttime satellite imagery for 6 months at a "global 1km composite." However, Doll et al. (2000) noted that the spatial resolution of the data obtained limited the effectiveness of the technique. Similarly, Sutton and Constanza (2002) sought to capture conventional GDP as well as non-marketed economic value, with conventional GDP being measured based on nighttime luminosity derived from satellite imagery to develop a global map of economic activity at a 1-km2 resolution.

Ebener et al. (2005) adopted a different approach to that used by Doll et al. (2000) to estimate per-capita income distribution at a sub-national level, trying to identify if other components/combinations of light data (such as mean frequency of observation and number of cells with light) and other parameters could be used to develop a model for country level prediction of per capita income. Sutton et al. (2007) built on the work by Ebener et al. (2005) contrasting two methods for estimating 'sub-national GDP levels' of Turkey, India, China and the United States in 2000. The researchers leveraged population data, GDP at a sub-national level (for the four countries) as well as city light data for the period 1992-93 and 2000 to developed regression models to estimate GDP at a state level for the four countries. The first method, a reproduction of the ideas by Ebener et al. (2005), summed up the values of intensity of light (based on nighttime satellite images) whilst the second used a spatial analytic approach on the imagery patterns as per the density of the population.

Moreover, studies have differed in terms of the period of observations with some studies focusing on shorter time frames (Doll et al. 2000) whilst others such as Chen and Nordhaus (2011) have looked at a longer time frame when conducting their analysis, examining the use of nighttime lights

as a proxy for country-level output at a 1° latitude × 1° longitude grid-cell level, comparing output and luminosity between 1992 and 2008.

Postal data

Hristova et al. (2016) sought to examine the use of global postal flows (from 187 countries) to estimate socioeconomic indicators that can be used to gauge national wellbeing. The study leveraged aggregated postal data from 2010 to 2014 that was collected by the Universal Postal Union. The records were used as a proxy indicator and were 'correlated to fourteen socioeconomic indicators. The study revealed close correlation between postal weighted outflows and Gross Domestic Product (r=0.79) and weighted outflow of postal data and the Human Development Index (r = 0.77) and weighted outflow of postal data.

Mobile Phone Data

Kreindler and Miyauchi (2015) sought to quantify the relationship between commuting flows (as derived from anonymized CDRs) and urban economic activity in Sri Lanka. This proxy indicator for economic activity has a key advantage in that it would enable the quantification of the informal economy which is missing from official statistics. This is a particular problem in developing economies.

Unemployment

Mobile Phone Data

Creating sustainable economic growth has a focus on ensuring job creation. However, job retention is also as vital. There is research (Toole et al. 2015) on using mobile phone data to identify shocks in the work force such as large-scale job losses, identifying individuals affected by such shocks and predicting changes in aggregate unemployment rates. Such insights are usually hard to capture and really highlight the potential use of big data to aid in the measurements and enrichment of achieving SDGs. Toole, et al (2015) infers changes in the macro-economy by using mobile phone data for two specific cases in Europe. The authors identify mobile phone users affected by a large-scale lay-off by assigning Bayesian probability weights on the changes of mobile phone activity. The analysis is carried out at multiple levels; individual (looking at behavioral changes associated with job loss) and thereafter at the community and province level (where the relationship inferred at the individual level can be used for prediction of unemployment.

Search Engine Data

Xu et al., (2013) use neural networks and support vector regressions to model the relationship between unemployment rates and search engine query data with data mining methods used thereafter to forecast unemployment trends.

Tourism

Target 8.9 deals with promoting sustainable tourism, specifically those that protect local culture whilst creating jobs.

Mobile Phone Data

Ahas et al. (2008) evaluated the use of passive mobile phone data to understand tourism in Estonia leveraging data on the call activities of phones on roaming (i.e. foreign) in the network cells. The information captured included the location, country of origin as well as time. For popular tourist regions, the study found a strong correlation (R nearly 0.99) with traditional statistics on accommodation, but lower correlations in regions where tourism infrastructure was not as strong

and where there were many transit tourists. Similarly, a previous study Ahas et al. (2007) used the social positioning method to analyse mobile phone data to understand the distribution of tourists by country of origin at a network cell level. The study revealed that coastal areas were more popular among tourists in the summer, with inland regions preferred in winter.

The European Commission – Eurostat conducted a study that among other things, sought to assess the feasibility of leveraging mobile phone data to generate statistics relating to tourism as well as highlighting the various challenges relating to the use of the data source. In addition, the study highlights the needs for a "central framework" that would facilitate the process of legally accessing data for national statistics offices as well as other stakeholders. Moreover the need for approved methodology for the generation of tourism statistics using phone data was also underscored. The study concluded that mobile positioning data could supplement official statistics at present.

NSOs have begun exploring the use of big data for tourism statistics as per the UN big data project inventory. For example, countries such as Belgium and the Czech Republic are exploring the feasibility of leveraging mobile phone data for the generation of tourism statistics. Similarly Hungary, seeks to leverage road sensor data to estimate traffic at its borders to improve the accuracy of its tourism statistics. Moreover, during 2012-2013 Statistics Netherland carried out research on the use of big data for tourism, which included using web scraped data – to identify tourist accommodation and characteristics, and anonymised call detail records to understand inbound tourism.

Insights from Big Data vs. Traditional Indicators

Target 8.1 Sustain per capita economic growth in accordance with national circumstances and, in particular, at least 7 per cent gross domestic product growth per annum in the least developed countries

There is one traditional indicator to measure this:

- 8.1.1 Annual growth rate of real GDP per capita

Satellite data and postal data have both been used to develop proxy indicators for economic development and can complement traditional indicators.

Target 8.5 By 2030, achieve full and productive employment and decent work for all women and men, including for young people and persons with disabilities, and equal pay for work of equal value

There are two indicators to measure this target:

- 8.5.1 Average hourly earnings of female and male employees, by occupation, age and persons with disabilities
- 8.5.2 Unemployment rate, by sex, age and persons with disabilities

By leveraging search engine query data, there is opportunity to identify unemployment trends with mobile phone data helping identify shocks in workforce – insights that would be hard to capture using the indicators above.

Target 8.9 By 2030, devise and implement policies to promote sustainable tourism that creates jobs and promotes local culture and products

There are two indicators to measure this target:

- 8.9.1 Tourism direct GDP as a proportion of total GDP and in growth rate
- 8.9.2 Number of jobs in tourism industries as a proportion of total jobs and growth rate of jobs, by sex

While big data sources may not be able to directly measure the two indicators above, there are able to provide valuable insights on tourism as a whole that could influence the policies developed to promote sustainable tourism.

Goal 9: Industry, Innovation and Infrastructure

Many developing countries still grapple with a lack of such basic infrastructure as roads, electrical power and access to water and sanitation. SDG 9 seeks to ensure that there is equitable access to sustainable infrastructure that is not just resilient but is also of high quality and reliability[18]. Satellite based technology coupled with machine learning can help identify infrastructure such as roads (Jean et al. 2016) and be leveraged to asses rural populations' access to such. Similarly, mobile phone data, which can be used to estimate mobility patterns, can contribute towards urban planning. Big data sources have greater applicability to target 9.1.

TABLE 09: Industry, Innovation and Infrastructure

Possible Targets	Theme	Application	Data Source	References
9.1	Predictors of poverty	Road access and rural population	Satellite Data	Mena & Malpica (2005); Jean et al. (2016)
9.1	Transport Planning	Real-time traffic monitoring	Sensor data	Shi & Abdel-Aty (2015)
9.1	Transport Planning	Road usage patterns	Mobile Phone Data	Toole et al. (2014)
9.1	Transport Planning	Origin-destination flows	Mobile Phone Data	Calabrese et al. (2011); Samarajiva et al (2015)
9.1	Transport Planning	Traffic monitoring	GPS data	Google Traffic

Big Data Sources
Satellite Data, Sensor Data, Mobile Phone Data

Road Access and Rural Population

Satellite Data
Mena & Malpica (2005) developed a model for automatic road network extraction based on satellite and aerial color imagery of high resolution. The model leverages four components: "data preprocessing; binary segmentation based on three levels of texture statistical evaluation; automatic vectorization by means of skeletal extraction; and finally a module for system evaluation."

Jean et al. (2016) used both daytime and nighttime satellite imagery as well as survey data and machine learning to identify varying economic well-being levels in Malawi, Nigeria, Rwanda, Tanzania and Uganda. The computer model was trained to identify features of daytime satellite images that were predictive of poverty including the identification of roads.

Mobility Patterns and Urban Planning

[18] http://www.un.org/sustainabledevelopment/infrastructure-industrialization/

Mobile Phone Data

Calabrese et al. (2011) sought to identify origin-destination flows of a population in order to estimate travel demand by analyzing location data from mobile phone users in the Boston Metropolitan area. Researchers show that the results correlate well with official data at the county level as well as the more granular census tract level. Similarly, Samarajiva et al. (2015) analysed CDRs from multiple mobile phone operators in Sri Lanka to understand mobility patterns in Colombo. This included peak and off-peak travel patterns, traffic as well as congregations of people. Specifically, the researchers sought to measure identifying changes in population density at a particular time relative to midnight, with the findings closely matching the results of a transportation survey conducted by the government.

Also related to mobility, but using a different method is the work of Hayano, & Adachi. (2013) who use GPS data (based on an subscription-based service offered by the service provider) to analyze movement to a particular area. The limitation of course is the useable sample of users who had to have subscribed to the GPS tracking service offered by the mobile operator.

Moreover, in order to map out relationship between people's mobility and their social networks, Phithakkitnukoon, S. et al (2012) conducted a study in Portugal analyzing a year's worth of mobile phone data for over a million mobile users. Among the key findings, the study found that around four-fifths of places visited by users were within a 20km radius of the location of the closest (in terms of geography) social ties. The study also revealed that population density impacted the radius, with areas that had a denser population having a shorter geo-social radius than compared to less dense areas.

Similarly, Calabrese et al. (2010) sought to understand the relationship between social events and the home areas of the attendees based on the premise that throughout time people tend to follow given preference patterns. The study analysed close to one million cell phone traces to show that there is a strong correlation between the home areas (origins) of the attendees and the type of social event held.

Sensor Data

A study conducted by Shi & Abdel-Aty (2015) deployed a Microwave Vehicle Detection System (MVDS) over 75 miles of the Orlando expressway network to explore the "viability of a proactive real-time traffic monitoring strategy evaluating operation and safety simultaneously." The study, which showed that traffic congestion tended to be localized and varied with time, further used Beysian Inference and random forest data mining techniques to develop models for crash prediction.

Estimating Traffic Flow

Mobile Phone Data

A study by Toole at al (2014) developed a model that leveraged call detail records to identify road usage patterns. An algorithm to mine calls and identify the phone users location transition probabilities (TP) was developed and these TP were then used in conjunction with demographic data to estimate "origin-destination flows of residents between any two intersections of a city" – congestion for each roadway was then estimated with the help of an algorithm.

Satellite Data

Similarly, a study by Necular (2015) sought to identify traffic patterns and traffic flows by utilizing a GPS data from about 3,600 drivers accounting for around 10,000 traces of vehicle GPS to identify

"contiguous set of road segments and time intervals which have the largest statistically significant relevance in forming traffic patterns."

Google Traffic
Google Traffic leverages crowd sourced traffic data from mobile users using Google Maps (on iphones) or android users with location turned on to identify traffic speed and congestion. The data provided by these phone users is used to analyse the speed and number of cars on the road. Furthermore, its history of traffic patterns is also leveraged to predict changes in traffic.

Insights from Big Data vs. Traditional Indicators
Target 9.1 Develop quality, reliable, sustainable and resilient infrastructure, including regional and trans-border infrastructure, to support economic development and human well-being, with a focus on affordable and equitable access for all
There are two traditional indicators to measure target 9.1:

- 9.1.1 Proportion of the rural population who live within 2 km of an all-season road
- 9.1.2 Passenger and freight volumes, by mode of transport

Big data sources can help understand patterns of road usage, pockets of congestion, and generally determine mobility patterns of population, all key factors in the development of infrastructure

Goal 10: Reduced Inequalities

According to the UN, income inequality within countries has risen, and while there has been inroads into reducing poverty, there has been rising consensus regarding the need for inclusive development, that is taking into account economic, social and environmental aspects of development.[19] With regards to specific targets, mobile phone data which can be leveraged to assess the socio economic status of populations at granular levels can contribute towards identifying the bottom 40 percent of population enabling governments to provide targeted assistance to promote income growth, helping to achieve target 10.1 and a portion of 10.2 that focuses on economic inclusion.

TABLE 10: Reduced Inequalities

Possible Targets	Theme	Applicationn	Data Source	References
10.1	Socio economic status and wellbeing	Socioeconomic status	Mobile Phone Data	Frias-Martinez et al. (2012); Blumenstock et al. (2015)

Big Data Sources
Mobile Phone Data

Socioeconomic Status and Economic Well-being
Goal 1 deals with this topic in detail

Insights from Big Data vs. Traditional Indicators

[19] http://www.un.org/sustainabledevelopment/inequality/

Target 10.1 By 2030, progressively achieve and sustain income growth of the bottom 40 per cent of the population at a rate higher than the national average.
This is to be measured using one indicator:

- 10.1.1 Growth rates of household expenditure or income per capita among the bottom 40 per cent of the population and the total population

While mobile phone data would not be able to provide growth rates per capita, it could contribute towards assessing changes in the socioeconomic status of populations.

Goal 11: Sustainable Cities and Communities

With almost 60 percent[20] of the world's population projected to live in urban areas by 2030, governments will face increased urban challenges including managing congestion, adequate housing, addressing urban poverty, providing sound infrastructure and ensuring access to basic services in a sustainable manner. Among other things, urban planning requires information on the mobility of population, areas of congregation, areas of residence, their social networks as well as general economic conditions. To that extent, understanding the dynamics between mobility, the physical distance of movements along with corresponding social networks, relationships that may or may not result in movement within a geographic space, can feed in to many aspects of urban planning. Big data sources, such as mobile phone data and satellite-based technology are well-positioned to provide insights of relevance to urban planning.

TABLE 11: Sustainable Cities and Communities

Possible Targets	Theme	Application	Data Source	References
11.1	Urban Poverty	Identifying slums	Satellite Data	Kohli et al. (2012)
11.1	Poverty Mapping	Identifying Poverty	Satellite Data	Jean el al. (2016)
11.2	Transport Planning	Geo-social Radius	Mobile Phone Data	Phithakkitnukoon et al. (2012)
11.2	Transport Planning	Origin-Destination Flows	Mobile Phone Data	Calabrese et al. (2011); Samarajiva et al (2015)
11.2	Transport Planning	Population Hotspots	Mobile Phone Data	Louail et al. (2014)
11.2	Transport Planning	Social events and home locations	Mobile Phone Data	Calabrese et al. (2010)
11.3	Land use	Land cover/land use changes	Remote Sensing Data	Tso & Mather (2001); Lu & Weng, (2007); Thomas et al. (2011)
11.5	Disaster	Human mobility	Mobile Phone	Lu et al. (2012); Lu et al.

[20] http://www.un.org/sustainabledevelopment/cities/

	response	after disasters	Data	(2016); Wilson et al. (2016)

Big Data Sources
Satellite Data, Mobile Phone Data

Identifying Urban Poverty
The identification of slums has been discussed under Goal 1.

Mobility Patterns

Mobile Phone Data
The use of mobile phone data to estimate mobility patterns has been described under Goal 9.

Land Use

Satellite Data
Satellite-based technology can be leveraged to identify temporal change that is invaluable for the study and understanding urban settlements. In particular, previous literature (Lu et al. 2012) has captured numerous studies (including Tso & Mather, 2001; Lu & Weng, 2007; Thomas et al. 2011) have leveraged remote sensing data to map changes in land cover/land use.

Disaster Response

Mobile Phone Data
Given the impact of extreme weather conditions, knowing the movement of displaced population will help steer disaster response facilitating the provision of targeted relief. There is a growing body of research that leverages mobile phone data to understand disaster response that include understanding human mobility patterns after disasters such as cyclones (Lu et al. 2016) and earthquakes (Wilson et al 2016; Lu et al. 2012). It is possible to leveraged mobile phone data to describe "short-term features (hours–weeks) of human mobility during and after extreme weather events, which are extremely hard to quantify using standard survey based research" (Lu et al. 2016). By using two de-identified datasets of mobile phone users in Bangladesh, researchers conducted analyses on climate stressed regions in Bangladesh in the periods before and after Cyclone Mahasen This dataset included the location of the mobile phone tower nearest to the caller. Analyses were also conducted on long-term migration patterns at a national level), where this dataset includes the location of the phone user's most frequently used mobile phone tower for each month. The results enabled the researchers to "quantify incidence, direction and duration of migration episodes enabling characterization of previously undocumented features of long-term migration patterns in climate stressed areas."

Analysis conducted by Wilson et al. (2016) on call detail records of mobile phone users showed the changes in mobility patterns of the population after the 2015 earthquake in Nepal at a granular detail including the movement of an estimated 390,000 people from the valley to surrounding areas. A key feature was the rapid deployment of analytical capacity and computational architecture that enabled the detailed estimation of displacement after the earthquake.

Flowminder, in collaboration with mobile phone operator Digicel analysed the mobility of two million de-identified mobile phones in Haiti after the 2010 earthquake, showing that results from their analyses closely corresponded to actual movement of the population after the earthquake (Lu et al.,

2012). The analysis showed that an estimated 630,000 individuals left Port-au-Prince in the days after the earthquake. Moreover, the findings also revealed that those who left Port-au-Prince during Christmas and New Year (before the earthquake) went to those same places after the earthquake, indicating that if travel and communicating patterns can be used to identify social networks, it could be used to predict movement after crises.

Insights from Big Data vs. Traditional Indicators

Target 11.1 *By 2030, ensure access for all to adequate, safe and affordable housing and basic services and upgrade slums*.

There is one indicator to measure the progress of this target

- 11.1.1 Proportion of urban population living in slums, informal settlements or inadequate housing

There have been studies conducted using satellite data that can be leveraged to assess urban poverty by identifying slum dwellings, thereby helping identify households that need access to housing.

Target 11.2 *By 2030, provide access to safe, affordable, accessible and sustainable transport systems for all, improving road safety, notably by expanding public transport, with special attention to the needs of those in vulnerable situations, women, children, persons with disabilities and older persons*

There is only one indicator to measure the achievement of target:

- 11.2.1 Proportion of population that has convenient access to public transport, by sex, age and persons with disabilities.

Big data sources can help identify the movement of population, areas of congestion, population hotspots all factors that affect the demand for transport systems.

Target 11.5 *By 2030, significantly reduce the number of deaths and the number of people affected and substantially decrease the direct economic losses relative to global gross domestic product caused by disasters, including water-related disasters, with a focus on protecting the poor and people in vulnerable situations.*

There are two indicators to measure this indicator:

- 11.5.1 Number of deaths, missing persons and persons affected by disaster per 100,000 people
- 11.5.2 Direct disaster economic loss in relation to global GDP, including disaster damage to critical infrastructure and disruption of basic services

Of the two indicators to measure target 11.5, mobile phone data can contribute towards indicator 11.5.1, which is multi-purpose indicator with linkages to other SDG targets including 1.5, 13.2, 1.3, 14.2, 15.3, 3.9, and 3.6 among others[21]. While mobile phone data may not be able to directly measure this indicator, it can help contribute towards the measurement) by helping identify the movement of people after a disaster.

Goal 12: Responsible Consumption and Production

According to the UN, Goal 12 deals with "promoting resource and energy efficiency, sustainable infrastructure, and providing access to basic services, green and decent jobs and a better quality of life for all."[22] This has ramifications in terms of the world's consumption of energy, food and water to ensure that resources are used efficiently to gain more with less.

[21] https://unstats.un.org/sdgs/files/metadata-compilation/Metadata-Goal-13.pdf
[22] http://www.un.org/sustainabledevelopment/sustainable-consumption-production/

Reducing Food Waste

Satellite Data

Insights derived from targets in some of the other goals that have been discussed, can contribute towards the achievement of Goal 12. For instance, given that only two-thirds of all produced is eventually consumed[23], food waste remains a key concern. Goal 2 discusses the role of big data sources in crop yield estimation. The ability to forecast production of agricultural crop will enable decision makers to prepare for potential shortages/surpluses and take corrective action, if needed, ensuring that food waste is minimized – this will contribute towards the achievement of target 12.3 (By 2030, halve per capita global food waste at the retail and consumer levels and reduce food losses along production and supply chains, including post-harvest losses), although it would not be able to measure the proposed indicator for this (12.3.1 food loss index).

Improving Energy Use

Goal 7 deals with the use of smart meters for electricity usage analysis and sensor data for renewable energy use supply estimation.

Goal 13: Climate Action

SDG Goal 13 is of significant importance to the global community as evidenced by the Paris Agreement of 2015, where 195 countries adopted a global climate deal that set out a plan to curb global warming to below 2^{o} C.[24] The Paris Agreement, which provides a roadmap to strengthen climate resilience, is crucial to achieve the SDGs. This further strengthens the case for accurate and timely data to measure progress towards the achievement of these targets. Big data, particularly satellite-based technology are well poised to aid the monitoring indicators of climate change.

TABLE 13: Climate Action

Possible Targets	Theme	Applications	Data Source	References
13.1	Disaster response	Human mobility after disasters	Mobile Phone Data	Lu et al. (2012); Lu et al. (2016); Wilson et al. (2016)
13.3	Changes in water-related ecosystem	Change in surface water	Satellite Data	Haas et al. (2009); Rokni et al. (2014); Pekel et al. (2014); Mueller et al. (2016)
13.3	Drought monitoring	Severity and extent of drought conditions		Henricksen, & Durkin (1986); Tucker & Choudhury (1987); Berhan et al. (2011)

[23] http://www.un.org/sustainabledevelopment/sustainable-consumption-production/
[24] https://ec.europa.eu/clima/policies/international/negotiations/paris_en

Big Data Sources
Mobile Phone Data, Satellite Data

Disaster Response

This has been covered in detail under Goal 11

Assessing Global Surface Water

This has been covered in detail under Goal 6

Drought Monitoring

Satellite Data
Many agricultural regions in the world are susceptible to extreme weather conditions such as droughts, which severely affects their ability to contribute towards food supply. This makes it crucial for decision makers to be aware of the location and potential of drought conditions to provide targeted assistance. Analysis, both spatial and temporal, of satellite images could help determine both the severity and the extent of the drought usually in near real-time.

For example, a study conducted by Berhan et al. (2011) in Ethiopia relied on satellite data to derive a spatial distribution of drought-affected areas in the region of focus. Drought condition data was generated by comparing the Normalized Difference Vegetation Index (NDV) for the first ten-day period of October 2009 with long-term mean NDVI. The results derived aligned with the rainfall recorded in the country during this period.

Insights from Big Data vs. Traditional Indicators
Target 13.1 *Strengthen resilience and adaptive capacity to climate-related hazards and natural disasters in all countries*

- 13.1.1 Number of countries with national and local disaster risk reduction strategies
- 13.1.2 Number of deaths, missing persons and persons affected by disaster per 100,000 people

Of the two indicators to measure target 13.1, mobile phone data can contribute towards indicator 13.1.2 (Number of deaths, missing persons and persons affected by disaster per 100,000 people) Indicator 13.1.2 is a multi-purpose indicator with linkages to other SDG targets including 1.5, 11.5, 1.3, 14.2, 15.3, 3.9, and 3.6 among others[25]. Big data sources may not be able to directly measure this indicator, it can help contribute towards the measurement) by helping identify the movement of people after a disaster.

Target *13.3 Improve education, awareness-raising and human and institutional capacity on climate change mitigation, adaptation, impact reduction and early warning*

- 13.3.1 Number of countries that have integrated mitigation, adaptation, impact reduction and early warning into primary, secondary and tertiary curricula
- 13.3.2 Number of countries that have communicated the strengthening of institutional, systemic and individual capacity-building to implement adaptation, mitigation and technology transfer, and development actions

[25] https://unstats.un.org/sdgs/files/metadata-compilation/Metadata-Goal-13.pdf

Similarly, while the indicators for target 13.3 focus on the number of countries that have implemented select initiatives/policies to strengthen national capacity (institutional, individual and systematic) to combat climate change, big data can be used to provide early warning of the effects of climate change.

Goal 14: Life Below Water

The achievement of SDG 14, which aims to "conserve and sustainable use the world's oceans, seas and marine resources,"[26] would be a key contributor to a sustainable future. However, human activities have severely impacted nearly 40 percent of oceans[27]. The use of satellite-based technology offers significant opportunity to ensure the protection of the marine ecosystem.[28]

For instance, the use of satellite data can be used to track the movement of marine vessels, understand their trajectory, identify illegal fishing and identify if ships cross marine protect areas. In particular, public data from the Automatic Ship Identification Systems (AIS) that is used my marine vessels worldwide can be used to observe marine activity (McCauley et al. 2016). The increasing amounts of AIS data require analysis, which requires the building up of national capacity. Initiatives include Global Fishing Watch, a collaboration between Google, Oceana and SkyTruth, that leverages data generated from AIS to assess fishing activity in near real-time. Other private companies that offer such marine fleet monitoring solutions include Big Ocean Data and Marine Traffic. Marine Traffic's service includes that provision of mobile apps that enable users to track vessels on their phones.

Big Data Sources
Satellite Data

Insights from Big Data vs. Traditional Indicators
Although it doesn't appear to be able to measure specific indicators under this goal, by tracking the movement of marine vessels, satellite data can help monitor levels of pollution, better regulate fishing practices, and ensure the protection of the marine ecosystem, contributing towards target 14.1 (which deals with the reduction of marine pollution of all kinds), 14.2 (which seeks to ensure the sustainable management and protection of marine ecosystems) and 14.4 ("effectively regulate harvesting and end overfishing, illegal, unreported and unregulated fishing and destructive fishing practices")

Goal 15: Life on Land

Deforestation and desertification creates significant ramifications for sustainable development. With an annual forest loss of around 13 million hectares, there is an increasing need to combat deforestation.[29] Remote sensing data has been instrumental in assessing in changes in topography over time and space and there are numerous studies that have sought to measure changes in surface water (Sawaya et al. 2003; Feyisa et al. 2014, Pekel et al. 2016), monitor drought conditions

[26] http://www.un.org/sustainabledevelopment/wp-content/uploads/2016/08/14_Why-it-Matters_Goal-14_Life-Below-Water_3p.pdf
[27] http://www.un.org/sustainabledevelopment/oceans/
[28]
[29] See: http://www.un.org/sustainabledevelopment/biodiversity/

(Henricksen, & Durkin, 1986; Tucker & Choudhury, 1987) and assess changes in forest cover (Hansen et al. 2013).

TABLE 15: Life on Land

Possible Targets	Theme	Application	Data Source	References
15.1	Identify Deforestation	Forest mapping	Satellite Data	Hansen et al. (2014); Ohmann et al. (2014)
15.3	Combat desertification	Changes in vegetation		Hutchinson et al. (2015)

Forest Mapping

Satellite Data

Satellite data has generally been used to access changes in forest cover across regions. One particular study (Hansen et al. 2014) by Google and researchers at the University of Maryland leveraged satellite data (from NASA/USGS Landsat 7) to map the gain and loss of forest cover globally (the analysis excluded Antarctica and select Arctic Islands) over the period 2000 to 2012, with Google Earth Engine processing 20 terapixels of Landsat data. The study found that over the 12-year period, 2.3 million square kilometers of forest area was lost globally, with a new forest gain of 800,000 square kilometers. Moreover, Global Forest Watch uses Google Earth Engine to analyze and create datasets including annual tree cover gain/loss data from the University of Maryland.[30]

Ohmann et al. (2014) integrate forest inventory plot data with satellite image data to map forest vegetation at regional levels. Their objective was to quantify the effects of scale related methods on map accuracy, as methodological choices (scale, resolution etc.) often affect the maps that are developed from satellite and plot data.

Similarly, monitoring land use in terms of vegetation change can be useful in ensuring that land is being used in a sustainable manner. For instance, Hutchinson et al. (2015) analysed 10 year's worth of MODIS NDVI data (available at 16 day intervals) for army training lands in northeast Kansas to NDVI trends to identify changes in vegetation. Potter (2014) explores the use of the Landsat difference index methodology as a low-cost mode of monitoring change caused by climate and biological factors. This method proves to be feasible in comparison with high-resolution aerial images and ground-based survey data.

Insights from Big Data vs. Traditional Indicators

Target 15.1 *By 2020, ensure the conservation, restoration and sustainable use of terrestrial and inland freshwater ecosystems and their services, in particular forests, wetlands, mountains and drylands, in line with obligations under international agreements*

- 15.1.1 Forest area as a proportion of total land area
- 15.1.2 Proportion of important sites for terrestrial and freshwater biodiversity that are covered by protected areas, by ecosystem type

Target 15.3 *By 2030, combat desertification, restore degraded land and soil, including land affected by desertification, drought and floods, and strive to achieve a land degradation-neutral world*

- 15.3.1 Proportion of land that is degraded over total land area

[30] https://developers.google.com/earth-engine/tutorial_gfw_01

In relation to the above targets listed above, satellite data in combination with other sources of data could potentially be leveraged to measure some of the indicators, specifically, 15.1.1 (Forest area as a proportion of total land area) and 15.3.1 (Proportion of land that is degraded over total land area).

Goal 16: Peace, Justice and Strong Institutions

One of the targets under this goal deals with the reduction of violence and related deaths and big data sources can play an important role in fighting crime through the analysis of historic crime data coupled with other data such as mobility patterns and demographics. According to Chen et al. (2012), in addition to crime analysis and prediction, big data can play a role in cyber security, open source intelligence, terrorism informatics and computational criminology among others. Big data insights have the greater applicability to target 16.1 (Significantly reduce all forms of violence and related death rates everywhere).

TABLE 16: Peace, Justice and Strong Institutions

Possible Targets	Theme	Application	Data Source	References
16.1	Predictive policing	Crime prediction	Mobile Phone Data	Bogomolov et al. (2014)
		Support crime prediction	Social Media Data	Gerber (2014)

Big Data Sources Used
Mobile Phone Data, Social Media Data

Predictive Policing

Mobile Phone Data
New data sources have the potential to better equip the police force to identify areas of potential crime. For instance, aggregated anonymized mobile phone data, which can be used to derive information on human mobility, can also be used to identify crime hotspots when used in conjunction with demographic data. For example, a study conducted by Bogomolov et al. (2014) in London sought to identify potential locations for future crimes based on aggregated people dynamics features from the analysis of past mobile phone data, as well as open data that included sales of residential property, weather data, transportation data, and data on criminal cases reported. When compared against real crime data, the results showed an accuracy of nearly 70% in predicting whether or not an area would be a crime spot in the next month.

Similarly, authorities have also been using crime prediction software solutions such as Predpol and Palantir. Predpol for instance, relies on information on crimes such as time, place type of crime, and proprietary algorithms to generate predictions on when and where crimes are more likely to happen.

Social Media Data
Studies have also investigated the use of social media data for crime incident prediction. For instance, Gerber (2014) conducted linguistic analysis coupled with topic modeling to identify discussion points in a city in the United States, which was then fed into a crime prediction model. The study found that the prediction of crime improved for more than three-fourths of crime types when twitter data was included in the crime prediction model.

Insights from Big Data vs. Traditional Indicators

Target 16.1 *Significantly reduce all forms of violence and related death rates everywhere*
The progress of this target would be measured using four indicators:

- 16.1.1 Number of victims of intentional homicide per 100,000 population, by sex and age
- 16.1.2 Conflict-related deaths per 100,000 population, by sex, age and cause
- 16.1.3 Proportion of population subjected to physical, psychological or sexual violence in the previous 12 months
- 16.1.4 Proportion of population that feel safe walking alone around the area they live

While big data sources may not be able to measure the four specific indicators for this target, insights derived from big data sources can help identify hotspots for crime enabling authorities to more effectively allocate resources and ultimately contributing to the achievement of target 16.1.

Goal 17: Partnerships for the Goals

There are numerous players that are active in the BD4D space and this section provides a high level assessment of the key players spanning government, multilateral organizations, civil society, industry, donor organizations and academia, among others. One of the key elements of discourse around the SDGs overall is the need for strategic partnerships, which is mimicked in the BD4D space as well. The need for multi-stakeholder engagement was also emphasized by Mr. Wu Hongbo, United Nations Under Secretary General for Economic and Social Affairs at the recently concluded UN World Data Forum in Cape Town (Jan 15-18, 2017):

> "Countries all around the world are mobilizing to carry out the 2030 Agenda for Sustainable Development. To do so, it is essential to have accurate, reliable, timely and disaggregated data…. This will require everyone in the statistics and data community – from governments, the private sector, the scientific and academic communities and civil society -- to find ways to work across different domains and create partnerships and synergies."[31]

The full potential of big data for development may only be realized if there is collaboration between various actors, including but not limited to government, academia, funding agencies, private sector, multilateral/international institutions, civil society, media and other non-governmental organizations (NGOs).

Multilateral Organizations

United Nations

The Global Working Group on Big Data for Official Statistics
The United Nations has been a key player within the big data for official statistics space. In particular, through the establishment of a Global Working Group (GWG) on Big Data for Official Statistics, the United Nations Statistics Division (UNSD) sought to explore the challenges and benefits involved in utilizing big data for official statistics purposes while taking into consideration issues related to

[31] Briefing on the opening of the UN World Data Forum. (2017, January 15). Retrieved from http://undataforum.org/WorldDataForum/wp-content/uploads/2017/01/Opening-remarks-Mr.-Wu-UNWDF-press-briefing-15-Jan.pdf

privacy, methodology, technology and legislation among others. One of the six task teams within this GWG is focused on big data and the SDGs with members of this team including a cross-section players such as World Bank, United Nations Economic and Social Commission for Asia and the Pacific, International Telecommunication Union, University of Pennsylvania, Orange, Data-pop Alliance and Paris21.[32]

It was the UNSD, in partnership with CSO Ireland, that organized the third international conference on Big data for Official statistics in Dublin in August/September 2016 where issues of access to big data and the need for partnerships, related privacy issues, the need for capacity building in NSOs and the use of big data for measuring SDGs were explored.

Global Platform for Big Data for Official Statistics
Furthermore, the GWG has proposed the development of a Global Platform for Big Data for Official Statistics to bring together ""public and private big data networks at the national, regional and global level both held by public and private agencies to make data, services and applications accessible and accelerate their synergies for research and capacity building." [33]

UN Big Data Project Inventory
Moreover, the UN Big Data Project Inventory, a joint effort of the World Bank and the UNSD for the GWG, showcases big data projects (a majority of which are initiatives by National Statistics Offices around the world) that are at various stages of implementation, ranging from exploratory research to projects being implemented, that support SDG measurement and/or relate to official statistics. The Inventory, which lists projects categorized by statistical area, data source would also provide categorization of projects by SDG goal.

UN Global Pulse
A big data initiative by the United Nations Secretary-General, UN Global Pulse conducts big data for development research through a network of innovation labs. The Pulse labs in Jakarta, Kampala and New York seek to develop new frameworks and methods to leverage big data for development and they also collaborate with a range of stakeholders including academia, public sector, governments and other UN agencies. Areas of study include climate & resilience, privacy & ethics, economic wellbeing, food & agriculture, and gender among others. UN Global Pulse has also been active in scoping out big data solutions to address the SDGs with a list of projects uploaded on the UN Big Data Project Inventory.

World Bank

The World Bank collaborated with SecondMuse Associates to publish a report, 'Big Data in Action for Development'[34] that sought to better understand how big data could be utilized in the development

[32] Using Big Data for the Sustainable Development Goals — UN GWG for Big Data. (n.d.). Retrieved from https://unstats.un.org/bigdata/taskteams/sdgs/
[33] Business Model for Global Platform for Big Data for Official Statistics in support of the 2030 Agenda for Sustainable Development. (2016, August 20). Retrieved from https://unstats.un.org/unsd/bigdata/conferences/2016/gwg/Item%203%20-%20Business%20Model%20for%20Global%20Platform%20for%20Big%20Data%20-%2016%20August%202016.pdf; Committee on Global Platform for Data, Services and Applications — UN GWG for Big Data https://unstats.un.org/bigdata/taskteams/globalplatform/
[34] Big Data in Action for Development. (n.d.). Retrieved from http://live.worldbank.org/sites/default/files/Big%20Data%20for%20Development%20Report_final%20version.pdf

sector. Moreover, in September 2016 they launched the World Bank Big Data Innovation Challenge[35] that invited applicants from around the world to provide big data innovations that could address issues related to climate change. Furthermore, the World Bank is involved in numerous big data related projects around the world spanning the prediction of poverty using mobile phone data in Guatamela to tracking poverty using Satellite Data in Pakistan.[36]

World Food Programme

The World Food Programme has also been exploring the use of data innovations to better position themselves in the fight against global hunger and the achievement of SDG 2.[37] For example, it teamed up with UN Global Pulse to develop a new method to estimate household food expenditure in Africa based on the analysis of mobile phone data.[38]

Non-Governmental Organizations/Non-Profits/Civil Society

Name of Organization	Description
data2x	An initiative of the United Nations Foundation, Data2x seeks to improve the quality of gender data along with the availability and use of it and to that end, has been supporting research pilots to understand how various big data collection and analysis methods could address gender gaps.
DATA POP ALLIANCE; HARVARD HUMANITARIAN INITIATIVE; mit media lab; ODI	Data-pop Alliance, whose core members include Harvard Development Initiative, MIT Media Lab, the Overseas Development Institute and the Flowminder Foundation, draws together various actors in the BD4D space. It focuses on themes around big data and development including politics and governance, climate change and resilience, data ethics and literacy among others.
FLOWMINDER.ORG	Sweden-based non-profit Flowminder seeks to leverage insights derived from the analysis of big data sources such as satellite data and mobile operator data to serve middle and low-income countries through various application areas that include socioeconomic analysis, precision epidemiology and disaster response.

Fay, M. (2016, October 20). The data revolution continues with the latest World Bank Innovation challenge. Retrieved from http://blogs.worldbank.org/voices/data-revolution-continues-latest-world-bank-innovation-challenge

[36] Big Data Project Inventory — UN GWG for Big Data. (n.d.). Retrieved from https://unstats.un.org/bigdata/inventory/

[37] Innovation at the World Food Programme. (2016, May). Retrieved from http://documents.wfp.org/stellent/groups/public/documents/communications/wfp284025.pdf

[38] WFP And UN Global Pulse Show How Big Data Can Save Lives And Fight Hunger. (2015, April 08). Retrieved from https://www.wfp.org/news/news-release/wfp-and-un-global-pulse-show-how-big-data-can-save-lives-and-fight-hunger

Sri Lanka-based non-profit, LIRNEasia has been engaged in big data for development research since 2012, with solutions that contribute to urban and transportation planning, and land use classification among others, and is also involved in conducting exploratory research on the potential for leveraging big data for other public purposes (such as dengue propagation). LIRNE*asia* has also been active in raising addressing issues related to privacy, security and methodology.

Since 2015, the Centre for Internet and Society, an India-based non-profit has been undertaking research into Big Data both from the perspective of human rights and development. Their big data and rights work looks into the harms and benefits of big data through case studies on digital identity, credit scoring, predictive policing, transportation, and smart grids in the Indian context. Their big data and development work has focused on the ways in which big data can be used to support the achievement of the global goals in the Indian context.

Academia/Research Labs

Academic Researchers

There is a growing body of academic researchers involved in big data research. For instance, some of researchers captured in this report include:

***Mobile Phone Data*:** Joshua Blumenstock (University of California, Berkeley), Vanessa-Frias Martinez (University of Maryland), Rein Ahas (University of Tartu), Ryousuke Shibasaki (University of Tokyo), Adeline Decuyper (University of Louvain) and Santi Phithakkitnukoon (Chiang Mai University)

***Satellite/Remote Sensing Data*:** Neal Jean (Stanford University), Marshall Burke (Stanford University), Komeil Rokni (Birjand University of Technology), J.M. Shawn Hutchinson (Kansas State University), and Alexander Brooker (University of Ottawa).

***Online Data*:** Alberto Cavallo and Roberto Rigobon (Massachusetts Institute of Technology), Albert C. Yang (Harvard Medical School), Wei Xu (Remin University of China)

Research Labs

HuMNet Lab - MIT

The Human Mobility and Networks Lab at the Massachusetts Institute of Technology seeks to explore social networks, mobility and cities by leveraging both machine learning and statistical physics.

Shiabasaki & Sekimoto Lab

A part of the Center for Spatial Information Science & Institute of Industrial Science, the University of Tokyo, the lab leverages satellite data, sensors, GPS and mobile phone data. Their big data research includes emergency management, people flow, dynamic census, mapping (roads, buildings, economic activities), urban and region analysis

Cambridge Big Data Strategic Research Initiative
The research initiative draws multi-disciplinary expertise from all six schools at the University of Cambridge to conduct big data research across five themes that include "theoretical foundations, imaging, data management and processing, ethics access and impact, and making big data work."[39]

Private Sector

Big Data Analytics/Solutions Providers

Real Impact Analytics
Through its Data for Good segment, Belgium-based Real Impact Analytics seeks to leverage mobile phone data to help meet the SDGs. In addition to the provision of operational apps for development agencies, the company is also involved in conducting analytical research in areas such as food security and financial inclusion.

Google Earth Engine
Google Earth Engine offers a platform for large-scale analysis of geospatial data. Their archive includes satellite imagery that spans over 40 decades (new images added daily) and enables data mining at a global scale. In addition to business users, the platform is also available to government users as well as for public purposes. Numerous developmental initiatives have leveraged Google Earth Engine's infrastructure including for malaria risk mapping as well as for assessing changes in global surface water.[40]

Data Providers

Telecom Operators
Telenor is among the mobile operators whose data has been used for development purposes. For example, in 2013, Telenor research, Oxford University, Harvard T.H. Chan School of Public Health and the University of Peshawa partnered to understand the impact of human mobility on the spread of dengue. The study leveraged data from 30 million Telenor Pakistan subscribers.
Similarly, in 2013, together with Grameen Phone and the Flowminder Foundation, the Telenor Group leveraged mobile phone data to understand migration patterns in Bangladesh in response to Cyclone Mahasen.

Orange Group and Sonatel have also shared anonymized Mobile Phone Data in Senegal with research laboratories as part of the data for development challenge that focused on the themes of agriculture, energy, health, official statistics and transport/urban planning. Moreover, South Korea's leading telephone company, KT Corporation sought to understand the spread of avian influenza by leveraging telecommunication big data.

Through its Telefonica Dynamics Insights division, Telefonica provides big data analytics solutions in the areas of transport, cities & regions, retail and media serving a client base that spans across private sectors operators to governments. Telefonica has also been active in the big data for social good space, having collaborated with MIT, GSMA, DataPop Alliance and the UN Global Pulse among others.

[39] http://www.bigdata.cam.ac.uk/research/BIGDATA/directory/research-themes
[40] Case Studies. (n.d.). Retrieved from https://earthengine.google.com/case_studies/

In addition to this. telecom operators have also provided their data to third party analytics providers under strict non-disclosure agreements as in the case of LIRNE*asia* in Sri Lanka with non-disclosure agreements restricting the organization from disclosing the names and the number of operators they have partnered with them.

Social Media Platforms
Social media platforms have also been utilizing their data, technology and expertise for development purposes. For example, in September 2016, Twitter and UN Global Pulse entered into a strategic data partnership offering UN Global Pulse access to the platform's data tools to generate insights that could help address the SDGs.

Facebook[41] worked with Columbia University and the World Bank to develop population maps of the world. Its Connectivity Labs divisions teamed up with their data science division, 'machine learning and Artificial Intelligence groups' and the infrastructure unit to analyse satellite imagery across 20 countries by leveraging techniques such as Facebook's image-recognition engine.

Search Engines
Although it no longer publishes Flu Trends and Dengue Trends, Google attempted make predictions of flu activity based on the analyses of search queries.

Infrastructure Service Providers

Big data analytics brings with it the challenges of data storage with many organizations lacking the necessary storage capacity to house such large volumes of data. Third party providers such as Amazon Web services come in to the picture offering infrastructure as-a-service solutions enabling companies to house their data outside their physical premises whilst making it accessible remotely. Other providers include Google Cloud Platform and Microsoft Azure.

Government

Governments hold large volumes of data regarding its citizens and analysis of such information can create insights that have implications for development.

National Statistics Offices

National Statistics Offices (NSOs) have been exploring the use of big data in official statistics. For example, the ONS big data project by the UK's Office for National Statistics aims to develop a strategy to leverage big data in its official statistics. In addition to analytical research using big data sources such as twitter data and smart electricity meters, the project has also conducted a literature review on mobile phone data's statistical uses. In 2015, the GWG on big data for official statistics conducted a global survey to assess the ground realities of the usage of big data by NSOs. Of the 93 countries that responded, around half indicated at least one big data-related project, reporting a total of 115 big data projects. Big data sources such as satellite data, scanner data, mobile phone data, web-scraping data, mobile phone data were among the sources of big data that were of interest to NSOs with big data being used for various statistical applications including transport, tourism, population and price. Moreover, based on the results of the survey developing countries were more likely to view meeting new data requirements such as measuring the SDGs as a benefit of

[41] Conditt, J. (2016, November 15). This is how the world looks on Facebook"s population maps. Retrieved from https://www.engadget.com/2016/11/15/facebook-population-maps-open-source/

big data, compared to their OECD counterparts[42]. The UN Big Data Project Inventory captures NSO big data initiatives.

Funders and Donor Agencies

Foundations that have provided funding for big data for development initiatives include International Development Research Centre, The World Bank, the Gates Foundation, The William and Flora Hewlett Foundation,[43] among others.

[42] See: https://unstats.un.org/unsd/statcom/47th-session/documents/2016-6-Big-data-for-official-statistics-E.pdf
[43] See: http://datapopalliance.org/; http://www.hewlett.org/grants/united-nations-foundation-for-continuation-of-data2x-including-big-data-pilots-0/

Challenges

Analytical challenges

While there is a considerable body of literature and work that has been built up leveraging these new data sources for developmental purposes, overall these are still the early stages. Much of what has been done can be considered proof-of-concepts and there has been limited scaling up. Hence there are still analytical challenges that have to be overcome. When dealing with large quantities of data, often unstructured and often from multiple sources, there is an implicit assumption that data will be "messy." The belief is that "what we lose in accuracy at the micro-level, we gain in insight at the macro level" (pp. 36, Mayer-Schönberger & Cukier, 2013). This is misleading. Data quality and its provenance do matter and the question is important in establishing generalizability of the Big Data findings. Knowing such ground context is important not just in understanding the base raw data, but also when interpreting results. For example Nathan Eagle, a pioneer in using big data for development, upon discovering low population mobility following a flood when analyzing CDR data from Rwanda, theorized the cause to be the outbreak of cholera. However a quick ground survey revealed that the real was washed out roads (David, 2013).

Whilst the large data sizes may make questions regarding the sampling rate irrelevant, knowing the representativeness of the data is still important. Even as mobile subscriptions in many developing countries nears 100%, it still doesn't mean that every person in the country owns a mobile phone. Depending on the research being pursued, questions such as the extant of coverage of the poor, or the levels of gender representation amongst mobile phone users can be very important. Street Bump, a mobile app that notifies Boston City Hall whenever app users drive over a pothole on Boston roads, suffers from a selection bias. This is because the app is biased towards the demographics of the app users, who often hail from affluent areas with greater smartphone ownership (Harford, 2014). Theoretically it could be possible that so long as resources are allocated on the basis of insights from such apps, the poorer parts of the city will be become further marginalized. Hence even in the big data paradigm, understanding and accounting for measurement bias, ensuring internal and external validity, and understanding inter-dependencies in the data, all remain important. These are foundational issues not just for "small data" but also for "big data" (Boyd & Crawford, 2012). Understanding data biases will be important, not least because one of the tag lines for the data revolution is "counting the uncounted" and "leaving no one behind."

The confusion of correlation with causation becomes more pronounced in the big data paradigm. By being observational, big data can measure only correlation and not causality. The techniques of data mining and machine learning, which underpin much the big data analytics, are primarily about correlation and predictions. When big data started to gain popularity, evangelists were quick to proclaim the end of theory and hypothesis testing, with correlation being all that mattered (see for example Anderson, 2008 and even Mayer-Schönberger & Cukier, 2013). While it is true often correlations can be enough to make decisions, the evangelistic proclamations have not been come to bear. The noted behavioral economist Sendhil Mullainathan argues that inductive science (i.e. algorithmically mining big data sources) will not drown out traditional deductive science (i.e. hypothesis testing) even in a Big Data paradigm (Mullainathan, 2013). 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 behavior (i.e. deductive science) rather than just predicting it.

All this is not to say that causal modeling is not possible in the big data paradigm. In fact this is achievable by conducting experiments (Varian, 2013). Telecom network operators themselves use such techniques when rolling out new services or, for that matter, for pricing purposes. But third-

party researchers will not easily have access to such experimentation possibilities since these are proprietary systems.

Another misconception the rise of big data will mean that surveys will go the way of the dodo. Even when leveraging MNBD for development, surveys and supplemental datasets will remain important to sharpen the analyses and especially to verify the underlying assumptions. For example Blumenstock & Eagle (2010) ran a basic household survey against a randomized set of phone numbers prior to data anonymization to build a training dataset. This allowed them to understand variations in mobility, social networks and consumption amongst men and women, and between different socio-economic groups that wouldn't have been possible using just the call records.

The broader analytical challenge is that new scientific knowledge from big data especially when this is using data from private sources is difficult to verify. Transparency and replicability are critical if the underlying methods and resultant insights are to be honed and improved. This is particularly important given extant embryonic stages of computational social science. This underscores how important it will be to open up the private data sources (in a manner that addresses potential privacy concerns) so as to be able to avail of the benefits of proper peer-review.

Accessing Data

The promise of the data revolution for sustainable development has been premised on the continually increasing levels of "datafication" of the economy and human behavior. The sources of big data are both in the private as well as public sectors. But in developing economies more comprehensive "datafication" (at least in terms of coverage of the population) has taken place in the private sector.

As this report has shown, a considerable and growing body of work showcase the potential of mobile network big data as one of the most potent sources for insights of relevance to developmental policy. However mobile phone operators are mostly private enterprises and the sector itself is highly competitive which makes data access difficult. There are several reasons why accessing this data even for developmental purposes is difficult. Firstly given that operators possess behavioral data about their customers, there are privacy concerns even with datasets being pseudonymized. Furthermore regulations and laws may be in place that prevent or limit data sharing. More often than not the laws have yet to catch up and there is often a regulatory and legal vacuum in developing economies. Operators would often prefer to tread carefully, if at all, for fear of attracting greater regulatory attention, and especially more so if the business case itself isn't very clear. Thirdly the mobile phone services sector is often highly competitive. As such, operators are wary that competitors could glean commercially sensitive information from data sharing with third parties. With the exception of LIRNEasia in Sri Lanka, nearly all instances of development-focused use of mobile network data involved the use of data from only one operator.

There are several countervailing forces in play. In the early days of the rise of big data in the public discourse, there was some recognition by operators that the data they possessed could have commercial uses that could help them generate new revenue sources. But there was less clarity on the business use cases. This created opportunities for academics and researchers to negotiate data access, which enabled the operators to also identify potential business cases, while allowing researchers and practitioners to explore developmental uses cases. This facilitated innovation with operators making their data available (under agreements) for challenges and competitions.[44] While

[44] For example the D4D challenges initiated in partnership with Orange, made historical psuedononymized data available for researchers under agreements for challenges in 2013 (data from Cote d"Ivoire) and 2014 (data from Senegal). For more information refer to http://www.d4d.orange.com/en/Accueil

it can be argued that the state of the art is still in its embryonic stages, by now there are some reasonable fleshed out development use cases. After MNBD gathered renewed interest in the public discourse during the Ebola crises in West Africa, GSMA (the global lobbying group for most GSM operators in the world) came up with recommendations that suggested in-house analyses rather than data sharing.[45] The current trend is for custom hardware and software infrastructure that sits behind the operator's firewall. The data are analyzed, results aggregated and then provided to outside parties. This is what happened in Nepal after the recent earthquake. Ncell, the largest mobile operator (based on subscriber numbers) in Nepal worked with Flowminder to utilize their data to map out the population displacements that had occurred as a result of the earthquake. The GSMA's suggested approach is gaining traction, with Flowminder, and the Data Pop Alliance proposing this method as the way forward. There are some clear benefits. Operator's concerns in relation to privacy, competition, and even regulatory attention would be better assuaged. But for this to work, there are several things to consider to have this working smoothly: (a) the developmental use case and the associated techniques and related code should be sufficiently robust to allow for a "plug and play" deployment; (b) there should be minimal need for other third-party "sensitive" data; (c) there must be some knowledge of the level of representativity of the operator's data. It is difficult to see how this approach could cover all different use cases since the type and complexity of the development insights needed would vary from context to context.

The question then is how to facilitate greater data sharing by service providers for public purposes, while addressing concerns related to privacy and competition. As a regulated industry, mobile network operators (MNOs) operate under license, which could theoretically be argued as a form of concessionary contract to deliver a government service, and as such, licenses could theoretically include provisions for data sharing. 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. While this philosophy is gaining traction, it would have to be built on research that articulates the concrete developmental potential for this data, whilst at the same time generating mechanisms to both protect privacy as well as the business interests of the private sector actor sharing the data, and thirdly also develop a clear business case for the private sector actor to share the data.

Capacity

Data science is a frontier field and will require broad expertise in a variety of fields. These include a combination of data mining, statistics, domain expertise, and also skills in data preparation, cleaning, and visualization. NSOs may have deep in-house statistical skills, but this is not enough to work with the large volumes of big data, which call for computer science, and decision analysis skills, which are not emphasized in traditional statistical courses (McAfee & Brynjolfsson, 2012). Currently there is a mismatch between the supply and demand for talent with the needed broader skill-sets i.e. data scientists. McKinsey predicts that by 2018 the demand for data-savvy managers and analysts in the United States would be 450,000, yet the supply will fall far short at only 160,000 (Manyika et al., 2011). In the short-term, the work, especially those for public-purposes will have to be done using collaborative teams, which can draw on a variety of skill-sets to collectively analyze and make sense of the data.

[45] The complete GSMA guidelines on the use of mobile data for responding to Ebola can be found at http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2014/11/GSMA-Guidelines-on-protecting-privacy-in-the-use-of-mobile-phone-data-for-responding-to-the-Ebola-outbreak-_October-2014.pdf

But capacity constraints are not just in conducting data science. We do not expect policy makers to necessarily become data scientists, but we need them to be informed consumers of big data outputs. As such paying particular attention to enlightening policy makers on how big data can be leveraged will be important. People who can connect these different disciplines and domains will have an important role.

Privacy

Privacy issues have taken center stage with the rise of big data. The conversation on privacy involves not just academics, but also state and private sector, and the general public who often are the primary producers of such data through their activities. But a consensus on how to address the attendant privacy challenges is yet to be reached. Privacy, as commonly understood, "is a sweeping concept, encompassing (among other things) freedom of thought, control over one's body, solitude in one's home, control over personal information, freedom from surveillance, protection of one's reputation, and protection from searches and interrogations" (Solove, 2008, p. 1). Attempts to define it in terms of boundary control by individuals (e.g., Samarajiva, 1994: 90) are difficult to translate into practical policy. The International Telecommunication Union (ITU) for example, defines individual privacy as "the right of individuals to control or influence what information related to them may be disclosed" (ITU, 2006). But it is difficult to clearly demarcate what an individual has authority over in the case of data generated as a by-product of a transaction, where the data are co-produced and held by one party. Current practices related to data privacy utilize a rights-based approach. This approach is best illustrated by the "inform and consent" policy, whereby companies inform users of what data is being collected about them and how it will be utilized by them and potentially also by the companies" affiliates and partners. But as Mayer-Schönberger & Cukier (2013) rightly point out, the "inform and consent" model is impractical. Most current user privacy policies are lengthy and written in legalese that makes understanding them difficult for the layperson. They also do not deal effectively with the secondary uses for the data, which often only manifest long after the original data was collected. It becomes impractical then for companies to know in advance all the potential uses or continuously seek permission for each new use.

The lines between personal and non-personal information are further blurred, when data is mashed up. Previously non-personal data, when mixed, could at times reveal insights that can easily be linked to an actual individual (Ohm, 2010). A recent study showed how personal attributes such as ethnicity, religious and political views, and even sexual orientation can be inferred from Facebook likes (Kosinski, Stillwell, & Graepel, 2013)

Even as computational social scientists utilize anonymization techniques address privacy concerns, the methods themselves are being called into question (Narayanan & Shmatikov, 2008). De Montjoye, Hidalgo, Verleysen, & Blondel (2013) were able to identify 90% of the people using just 4 data points from an anonymized set of 1.5 million CDRs. Even though the data itself have any identity information, the authors showed that the real-world identities could be found by cross-referencing their data with other public data. The emergent state of the art in technical solutions to limit re-identification may hold promise but it is still in the early stages. One particular promising approach is differential privacy, which seeks to ensure that the results that are derived from a dataset are virtually the same whether a particular individual was in it or not. This is accomplished in principle by adding noise to the dataset in such a manner that it does not affect the overall statistical robustness of the results within a certain level of sensitivity.[46]

[46] For a survey of differential privacy techniques see Ji, Lipton, & Elkan (2014)

One can hope that new privacy-preserving techniques will be sufficiently advanced by the time developing economies become more "datafied." In the meantime, these new data sets can be leveraged for public purposes under controlled situations backed by legal agreements and approval processes. It is such research that is currently advancing the state of the art in understanding the privacy bounds for specific use cases when data with confidential information are released to third parties and/or to the public.

While the means to address privacy concerns are far from clear, there is general agreement that there must be some safeguards. These safeguards may be technological, conceptual, legal or even a combination of all three.

Ethics

While the broader data for development movement is showing the usefulness of leveraging new data sources for developmental purposes, there are challenging questions to understand the ethical dilemmas that emerge from using data about people's behaviors. These applications often commonly fall under the secondary-use category, with the data being leveraged for purposes other than what it may have originally been intended for. Practical ways of addressing privacy concerns around secondary use and permissions for secondary use are difficult as was mentioned earlier. An appropriate legal framework is still emergent, with little consensus. Hence the onus on understanding and addressing the ethical dilemmas will remain on researchers and practitioners. There will be a potential need for the development of professional standards to ensure that practitioners determine legal and ethical issues and address them. These will have to be done on context-by-context basis. There is some movement towards this. The UN Office for the Coordination of Humanitarian Affairs (OCHA) recently published a policy brief on how to build data responsibility into humanitarian data ecosystem (OCHA, 2016a). OCHA also issued a guidance note for data collection and disaggregation based on principles around participation, data disaggregation, self-identification, transparency, privacy, and accountability (OCHA 2016b). But the conflict will be in addressing ethical dilemmas in a manner that does not reduce the societal and/or individual utility for the innovations that are currently emerging in using data for developmental purposes.

Competition

Competition is seen as a good that ensures a "level playing field" to give all competitors equal opportunities, though not identical or equal outcomes. In terms of competition, the relevant issues to consider in the explosion of data as well as in the techniques for leveraging them are the effects of mergers and acquisitions across distinctly different markets on aggregation of data. For example, it has been argued that Google's acquisition of Nest, a supplier of home thermostats and Carbon Monoxide detectors, operating in a distinctly different market, should still have attracted greater regulatory scrutiny because of the potential of data aggregation (Stucke & Grunes, 2016: 89-92).

Even when the entities controlling data are not regulated monopolies, they may approach monopoly status are certain markets (Rosoff, 2014). In such instances as well as in cases of mergers or acquisitions that would increase market share, it is likely that competition authorities will pay attention to the effects of data in addition to conventional competition issues. This appears to be at least part of the justification for the attention being paid to Google by European competition authorities. The issue here is whether traditional conceptions of market definitions continue to be relevant.

Whilst this may not initially seem to affect issues related to public uses of private sector data for developmental purposes, how the emergent competition issues get resolved will affect the type of data that is collected and shared and in particular the longer term success of the "data philanthropy" concept.

Conclusion

The thrust to utilize big data for official statistics underscores the potential for generating new insights and complement existing measures. Big data can potentially revolutionize current official statistical systems in one of several ways[47]: a) Entirely replace existing statistical sources such as surveys; b) Partially replace existing statistical sources such as surveys; c) Provide complementary statistical information in the same statistical domain but from other perspectives; d) Improve estimates from statistical sources; and e) Provide completely new statistical information in a particular statistical domain. At the moment, complementing existing data is what offers the greatest potential for big data sources.

Research conducted to date has explored the utilization of different data sources for specific developmental applications. For example, the analysis of satellite imagery has typically been used to monitoring changes in topography including crop/yield estimation, drought monitoring, deforestation and carbon stock mapping among others. Mobile network big data has proven to be useful in understanding mobility patterns of the population, creditworthiness of its users, socioeconomic status of the population among others, while social media data is well positioned for sentiment analysis. There have been literature reviews (for example, Williams, 2016, UK's Office of National Statistics; Lokanathan and Gunaratne, 2014) that have captured the statistical applications of big data, in particular mobile phone data.

The big data for development (BD4D) landscape hosts a range of players spanning government, industry, academia, and civil society among others. These players operate as policy actors, researchers, funders as well as intermediaries. Understanding the various actors at play offers greater opportunities for strategic partnership and collaborations. Thus, Goal 17 has been viewed through the lens of big data for development, stressing on the need for collaboration between various big data actors, and providing a snapshot of the BD4D landscape.

However it is essential to be cognizant of the fact that the state of the art is still developing and there are analytical and technological challenges to be aware of. We need to ensure that incorrect conclusions are not drawn from a blind application of big data techniques. The current techniques including surveys will remain important not only to bootstrap some of the big data techniques with training data but also to fine-tune models to ground realities. Hence big data will not completely replace traditional surveys but rather complement them. Similarly given the analytical challenges it will be very important that there is transparency and replicability in the analyses. A principal concern of the data revolution is about "counting the uncounted" and as such we need to pay particular attention to "representativity" of these new data sources that are being leveraged i.e. how accurately it reflects the population. Marginalization in the real world can often result in marginalization in the digital world beyond just issues of access to technologies, or being represented in the digitized data. As such localized testing of these new techniques with local experts (aware of ground truth and local context) are very important. Being mindful of these analytical challenges does not negate the potential of big data, but rather helps refine and improve the overall process of leveraging big data for monitoring and achieving the SDGS.

Developing economies in particular have much lower levels of "datafication" than developed economies, which means some of the most interesting and relevant data exists amongst the private sector as shown in this report. Accessing such data will not be without challenges, not least because in competitive industries such as the telecom sector, there would be competitive implications to sharing data. Even then there are costs and considerations associated with extracting and analyzing

[47] See for example http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf

the data from private sector industries so as to sufficiently protect aspects related customer privacy/ confidentiality as well as protecting commercially sensitive business intelligence in competitive industries (e.g. the telecom sector). Innovation will have to unleash appropriate business models for leveraging these data sources. But innovation in this space will also require the confluence of a variety of actors such as state, private, academia, and importantly non-governmental researchers and practitioners. New forms of partnerships between these actors are required not just because the true value comes from assembling different types of data from different sources, but also because of the inherent capacity challenges to make full use of the data. This is inherently a multi-disciplinary effort requiring computer scientists, statisticians, and domain/ subject-matter experts along with government officials.

There are also challenges in understanding and addressing the privacy implications of Big Data. It can be seen that the sharing (subject to appropriate privacy protocols) of privately held data such as mobile phone records can be mutually beneficial to both government as well as the private sector. For example, emerging research in Africa shows how the reduction in airtime top-ups forecast declines in income among the poor. This can allow for targeted and timely policy actions by government to address the underlying problems, which would not be possible with the lagged, and often limited, insights revealed by traditional statistics. Such a collaborative early warning and early action system shows how such sharing could be considered a business risk mitigation strategy for operators in emerging markets. But such cooperation is predicated on opening up the currently privileged access that a few researchers and organizations have been given to mobile operator datasets.

ANNEX I: SDG Targets by Big Data Source

Possible Targets	Theme	Statistical application	Data Source	References
16.1	Predictive policing	Crime prediction	Mobile Phone Data	Bogomolov et al. (2014)
5.1	Gender Prediction	Gender prediction	Mobile Phone Data	Sundsøy et al. (2015); Blumenstock & Eagle (2010)
11.2	Transport Planning	Geo-social Radius	Mobile Phone Data	Phithakkitnukoon et al. (2012)
13.1	Disaster response	Human mobility after disasters	Mobile Phone Data	Lu et al. (2012); Lu et al. (2016); Wilson et al. (2016)
1.1; 1.2	Socio economic status and wellbeing	Human mobility and socioeconomic levels	Mobile Phone Data	Frias-Martinez et al. (2012)
1.1; 1.2	Socio economic status and wellbeing	Estimating poverty and wealth	Mobile Phone Data	Blumenstock et al. (2015)
1.1; 1.2	Socio economic status and wellbeing	Socioeconomic status	Mobile Phone Data	Gutierrez et al. (2013)
1.4	Financial Inclusion	Creditworthiness of the unbanked	Mobile Phone Data	Kumar and Mohta (2012)
1.5	Disaster response	Human mobility after disasters	Mobile Phone Data	Lu et al. (2016); Wilson et al. (2016);Lu et al. (2012)
2.1	Expenditure on Food	Proxy indicator for food expenditure	Mobile Phone Data	Decuyper et al. (2014)
3.3	Disease Propagation	Mobility from regions of disease outbreak	Mobile Phone Data	Wesolowski, et al. (2015); Bengston et al. (2015); Wesolowski et al. (2014)
3.3	Disease Propagation	Sources and sinks for diseases	Mobile Phone Data	Ruktanonchai et al. (2016); Tatem et al. (2014)
3.3	Disease Propagation	Disease Importation rate	Mobile Phone Data	Tatem et al. (2009)
4.6	Illiteracy Prediction	Areas of low literacy	Mobile Phone Data	Sundsøy, P. (2016)

8.9	Tourism	Destination of tourists	Mobile Phone Data	Ahas et al. (2008)
8.9	Tourism	Seasonal tourism	Mobile Phone Data	Ahas et al. (2007)
9.1	Transport Planning	Road usage patterns	Mobile Phone Data	Toole et al. (2014)
9.1	Transport Planning	Origin-destination flows	Mobile Phone Data	Calabrese et al. (2011); Samarajiva et al (2015)
10.1	Socio economic status and wellbeing	Socioeconomic status	Mobile Phone Data	Frias-Martinez et al. (2012); Blumenstock et al. (2015)
11.2	Transport Planning	Origin-Destination Flows	Mobile Phone Data	Calabrese et al. (2011); Samarajiva et al (2015)
11.2	Transport Planning	Population Hotspots	Mobile Phone Data	Louail et al. (2014)
11.2	Transport Planning	Social events and home locations	Mobile Phone Data	Calabrese et al. (2010)
11.5	Disaster response	Human mobility after disasters	Mobile Phone Data	Lu et al. (2012); Lu et al. (2016); Wilson et al. (2016)
9.1	Transport Planning	Traffic monitoring	GPS data	Google Traffic
2.c.	Price Indexes	Constructing consumer price index	Online Prices	Cavallo and Rogobon (2016)
8.1	GDP	GDP and Human Development	Postal Data	Hristova et al. (2016)
11.3	Land use	Land cover/land use changes	Remote Sensing Data	Tso & Mather (2001); Lu & Weng, (2007); Thomas et al. (2011)
1.1; 1.2	Poverty Mapping	Identifying the poor	Satellite Data	Elvidge et al. (2009); Jean et al. (2016)
1.1; 1.2	Poverty Mapping	Urban poverty	Satellite Data	Kohli et al. (2012)

2.1	Drought monitoring	Severity and extent of drought conditions	Satellite Data	Berhan et al. (2011); Tucker & Choudhury (1987); Henricksen, & Durkin (1986);
2.4	Early crop yield assessment	Developing vegetation health indices	Satellite Data	Kogen et al. (2011)
6.6	Changes in water-related ecosystem	Change in surface water	Satellite Data	Mueller et al. (2016); Haas et al. (2009); Rokni et al. (2014); Pekel et al. (2014);
7.1	Access to electricity	Nighttime luminosity	Satellite Data	Elvidge et al. (1997); Elvidge et al. (2009); Townsend & Bruce (2010); Doll & Pachauri (2010); Chen & Nordhaus (2011)
8.1	GDP	Economic development	Satellite Data	Elvidge et al. (1997); Sutton and Constanza (2002); Ebener et al. (2005); Chen and Nordhaus (2011);
9.1	Predictors of poverty	Road access and rural population	Satellite Data	Mena & Malpica (2005); Jean et al. (2016)
11.1	Urban Poverty	Identifying slums	Satellite Data	Kohli et al. (2012)
11.1	Poverty Mapping	Identifying Poverty	Satellite Data	Jean el al. (2016)
13.3	Changes in water-related ecosystem	Change in surface water	Satellite Data	Haas et al. (2009); Rokni et al. (2014); Pekel et al. (2014); Mueller et al. (2016)
13.3	Drought monitoring	Severity and extent of drought conditions	Satellite Data	Henricksen, & Durkin (1986); Tucker & Choudhury (1987); Berhan et al. (2011)
15.1	Identify Deforestation	Forest mapping	Satellite Data	Hansen et al. (2014); Ohmann et al. (2014)
15.3	Combat desertification	Changes in vegetation	Satellite Data	Hutchinson et al. (2015)
3.3	Disease Propagation	Seasonal trends of diseases	Search Engine Data	Schuster et al. 2010; Yang et al. 2010; Xu et al. 2010

8.5	Unemployment	Unemployment trends	Search Engine Data	Xu et al., (2013)
9.1	Transport Planning	Real-time traffic monitoring	Sensor data	Shi & Abdel-Aty (2015)
7.1	Residential electricity consumption	Determinants of electricity consumption	Smart Meter Data	Kavousian et al. (2013)
16.1	Predictive policing	Support crime prediction	Social Media Data	Gerber (2014)

ANNEX II: Big Data Sources by Key Applications

Theme	Possible Targets	Statistical application	Data Source	References
Access to electricity	7.1	Nighttime luminosity	Satellite Data	Elvidge et al. (1997); Elvidge et al. (2009); Townsend & Bruce (2010); Doll & Pachauri (2010); Chen & Nordhaus (2011)
Changes in water-related ecosystem	6.6	Change in surface water	Satellite Data	Mueller et al. (2016); Haas et al. (2009); Rokni et al. (2014); Pekel et al. (2014);
Changes in water-related ecosystem	13.3	Change in surface water	Satellite Data	Haas et al. (2009); Rokni et al. (2014); Pekel et al. (2014); Mueller et al. (2016)
Combat desertification	15.3	Changes in vegetation	Satellite Data	Hutchinson et al. (2015)
Disaster response	13.1	Human mobility after disasters	Mobile Phone Data	Lu et al. (2012); Lu et al. (2016); Wilson et al. (2016)
Disaster response	1.5	Human mobility after disasters	Mobile Phone Data	Lu et al. (2016); Wilson et al. (2016);Lu et al. (2012)
Disaster response	11.5	Human mobility after disasters	Mobile Phone Data	Lu et al. (2012); Lu et al. (2016); Wilson et al. (2016)
Disease Propagation	3.3	Disease Importation rate	Mobile Phone Data	Tatem et al. (2009)
Disease Propagation	3.3	Mobility from regions of disease outbreak	Mobile Phone Data	Wesolowski, et al. (2015); Bengston et al. (2015); Wesolowski et al. (2014)
Disease Propagation	3.3	Seasonal trends of diseases	Search Engine Data	Schuster et al. 2010; Yang et al. 2010; Xu et al. 2010
Disease Propagation	3.3	Sources and sinks	Mobile Phone	Ruktanonchai et al. (2016);

		for diseases	Data	Tatem et al. (2014)
Drought monitoring	2.1	Severity and extent of drought conditions	Satellite Data	Berhan et al. (2011); Tucker & Choudhury (1987); Henricksen, & Durkin (1986);
Drought monitoring	13.3	Severity and extent of drought conditions	Satellite Data	Henricksen, & Durkin (1986); Tucker & Choudhury (1987); Berhan et al. (2011)
Early crop yield assessment	2.4	Developing vegetation health indices	Satellite Data	Kogen et al. (2011)
Expenditure on Food	2.1	Proxy indicator for food expenditure	Mobile Phone Data	Decuyper et al. (2014)
Financial Inclusion	1.4	Creditworthiness of the unbanked	Mobile Phone Data	Kumar and Mohta (2012)
GDP	8.1	Economic development	Satellite Data	Elvidge et al. (1997); Sutton and Constanza (2002); Ebener et al. (2005); Chen and Nordhaus (2011);
GDP	8.1	GDP and Human Development	Postal Data	Hristova et al. (2016)
Gender Prediction	5.1	Gender prediction	Mobile Phone Data	Sundsøy et al. (2015); Blumenstock & Eagle (2010)
Identify Deforestation	15.1	Forest mapping	Satellite Data	Hansen et al. (2014); Ohmann et al. (2014)
Illiteracy Prediction	4.6	Areas of low literacy	Mobile Phone Data	Sundsøy, P. (2016)
Land use	11.3	Land cover/land use changes	Remote Sensing Data	Tso & Mather (2001); Lu & Weng, (2007); Thomas et al. (2011)
Poverty Mapping	11.1	Identifying Poverty	Satellite Data	Jean el al. (2016)
Poverty Mapping	1.1; 1.2	Identifying the poor	Satellite Data	Elvidge et al. (2009); Jean et al. (2016)
Poverty Mapping	1.1; 1.2	Urban poverty	Satellite Data	Kohli et al. (2012)
Predictive policing	16.1	Crime prediction	Mobile Phone Data	Bogomolov et al. (2014)
Predictive policing	16.1	Support crime prediction	Social Media Data	Gerber (2014)
Predictors of poverty	9.1	Road access and rural population	Satellite Data	Mena & Malpica (2005); Jean et al. (2016)
Price Indexes	2.c.	Constructing consumer price index	Online Prices	Cavallo and Rogobon (2016)
Residential electricity consumption	7.1	Determinants of electricity consumption	Smart Meter Data	Kavousian et al. (2013)

Socio economic status and wellbeing	1.1; 1.2	Estimating poverty and wealth	Mobile Phone Data	Blumenstock et al. (2015)
Socio economic status and wellbeing	1.1; 1.2	Human mobility and socioeconomic levels	Mobile Phone Data	Frias-Martinez et al. (2012)
Socio economic status and wellbeing	1.1; 1.2	Socioeconomic status	Mobile Phone Data	Gutierrez et al. (2013)
Socio economic status and wellbeing	10.1	Socioeconomic status	Mobile Phone Data	Frias-Martinez et al. (2012); Blumenstock et al. (2015)
Tourism	8.9	Destination of tourists	Mobile Phone Data	Ahas et al. (2008)
Tourism	8.9	Seasonal tourism	Mobile Phone Data	Ahas et al. (2007)
Transport Planning	11.2	Geo-social Radius	Mobile Phone Data	Phithakkitnukoon et al. (2012)
Transport Planning	9.1	Origin-destination flows	Mobile Phone Data	Calabrese et al. (2011); Samarajiva et al (2015)
Transport Planning	11.2	Origin-Destination Flows	Mobile Phone Data	Calabrese et al. (2011); Samarajiva et al (2015)
Transport Planning	11.2	Population Hotspots	Mobile Phone Data	Louail et al. (2014)
Transport Planning	9.1	Real-time traffic monitoring	Sensor data	Shi & Abdel-Aty (2015)
Transport Planning	9.1	Road usage patterns	Mobile Phone Data	Toole et al. (2014)
Transport Planning	11.2	Social events and home locations	Mobile Phone Data	Calabrese et al. (2010)
Transport Planning	9.1	Traffic monitoring	GPS data	Google Traffic
Unemployment	8.5	Unemployment trends	Search Engine Data	Xu et al., (2013)
Urban Poverty	11.1	Identifying slums	Satellite Data	Kohli et al. (2012)

References

Ahas, R., Aasa, A., Mark, Ü., Pae, T., & Kull, A. (2007). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. Tourism Management, 28(3), 898-910.

Ahas, R., Aasa, A., Roose, A., Mark, Ü., & Silm, S. (2008). Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. Tourism Management, 29(3), 469-486.

Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. *Wired Magazine*, *16*(7)

Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., & Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. Proceedings of the National Academy of Sciences, 106(51), 21484-21489.

Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., ... & Piarroux, R. (2015). Using mobile phone data to predict the spatial spread of cholera. Scientific reports, 5.

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & Von Schreeb, J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. PLoS Med, 8(8), e1001083.

Berhan, G., Tadesse, T., Atnafu, S., & Hill, S. (2011). Drought Monitoring in Food-Insecure Areas of Ethiopia by Using Satellite Technologies. In Experiences of Climate Change Adaptation in Africa (pp. 183-200). Springer Berlin Heidelberg.

Blumenstock, J. E. (2016). Fighting poverty with data. Science, 353(6301), 753-754.

Blumenstock, J., & Eagle, N. (2010, December). Mobile divides: gender, socioeconomic status, and mobile phone use in Rwanda. In Proceedings of the 4th ACM/IEEE International Conference on Information and Communication Technologies and Development (p. 6). ACM.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. Science, 350(6264), 1073-1076.

Bogomolov, A., Lepri, B., Staiano, J., Letouzé, E., Oliver, N., Pianesi, F., & Pentland, A. (2015). Moves on the street: Classifying crime hotspots using aggregated anonymized data on people dynamics. Big Data, 3(3), 148-158.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014, November). Once upon a crime: towards crime prediction from demographics and mobile data. In Proceedings of the 16th international conference on multimodal interaction (pp. 427-434). ACM.

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Brdar, S., Gavrić, K., Ćulibrk, D., & Crnojević, V. (2016). Unveiling Spatial Epidemiology of HIV with Mobile Phone Data. Scientific reports, 6.

Brooker, A., Fraser, R. H., Olthof, I., Kokelj, S. V., & Lacelle, D. (2014). Mapping the activity and evolution of retrogressive thaw slumps by tasselled cap trend analysis of a Landsat satellite image stack. Permafrost and Periglacial Processes, 25(4), 243-256.

Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C. (2011). Estimating Origin-Destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area. IEEE Pervasive Computing, 99.

Calabrese, F., Pereira, F. C., Di Lorenzo, G., Liu, L., & Ratti, C. (2010, May). The geography of taste: analyzing cell-phone mobility and social events. In International Conference on Pervasive Computing (pp. 22-37). Springer Berlin Heidelberg.

Cambridge Big Data. (n.d.). Retrieved from http://www.bigdata.cam.ac.uk/research

Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. Journal of Monetary Economics, 60(2), 152-165.

Cavallo, A., & Rigobon, R. (2016). The Billion Prices Project: Using online prices for measurement and research. The Journal of Economic Perspectives, 30(2), 151-178.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS quarterly, 36(4), 1165-1188.

Chen, X., & Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. Proceedings of the National Academy of Sciences,108(21), 8589-8594.

Coombes, P. J., & Barry, M. E. (2014). A systems framework of big data driving policy making–Melbourne"s water future. In OzWater2014 Conference. Australian Water Association. Brisbane Australia.

Cosner, C., Beier, J. C., Cantrell, R. S., Impoinvil, D., Kapitanski, L., Potts, M. D., ... & Ruan, S. (2009). The effects of human movement on the persistence of vector-borne diseases. Journal of theoretical biology, 258(4), 550-560.

Costăchioiu, T., & Datcu, M. (2010, June). Land cover dynamics classification using multi-temporal spectral indices from satellite image time series. InCommunications (COMM), 2010 8th International Conference on (pp. 157-160). IEEE.

D4D. (n.d.). Orange Data for Development Challenge in SenegalRetrieved from http://d4d.orange.com/content/download/43330/405662/version/3/file/D4Dchallenge_leaflet_A4_V2Eweblite.pdf

Data 2x. (n.d.).Big Data and Gender Data Gaps. Retrieved from http://data2x.org/wp-content/uploads/2014/08/Big-Data-Projects.pdf

Data Kind. (n.d.). Retrieved from http://www.datakind.org/

David, T. (2013). Big Data from Cheap Phones. *Technology Review*, *116*(3), 50–54.

Decuyper, A., Rutherford, A., Wadhwa, A., Bauer, J. M., Krings, G., Gutierrez, T., ... & Luengo-Oroz, M. A. (2014). Estimating food consumption and poverty indices with mobile phone data. arXiv preprint arXiv:1412.2595.

De Montjoye, Y.-A., Hidalgo, C. a, Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, *3*, 1376. doi:10.1038/srep01376

Doll, C. H., Muller, J. P., & Elvidge, C. D. (2000). Night-time imagery as a tool for global mapping of socioeconomic parameters and greenhouse gas emissions. AMBIO: a Journal of the Human Environment, 29(3), 157-162.

Doll, C. N., & Pachauri, S. (2010). Estimating rural populations without access to electricity in developing countries through night-time light satellite imagery. Energy Policy, 38(10), 5661-5670.

Eagle, N., Macy, M., & Claxton, R. (2010). Network diversity and economic development. Science, 328(5981), 1029-1031.

Ebener, S., Murray, C., Tandon, A., & Elvidge, C. C. (2005). From wealth to health: modelling the distribution of income per capita at the sub-national level using night-time light imagery. international Journal of health geographics, 4(1), 1.

Elvidge, C. D., Baugh, K. E., Kihn, E. A., Kroehl, H. W., Davis, E. R., & Davis, C. W. (1997). Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption. International Journal of Remote Sensing, 18(6), 1373-1379.

Elvidge, C. D., Safran, J., Tuttle, B., Sutton, P., Cinzano, P., Pettit, D., ... & Small, C. (2007). Potential for global mapping of development via a nightsat mission. GeoJournal, 69(1-2), 45-53.

Elvidge, C. D., Sutton, P. C., Ghosh, T., Tuttle, B. T., Baugh, K. E., Bhaduri, B., & Bright, E. (2009). A global poverty map derived from satellite data.Computers & Geosciences, 35(8), 1652-1660.

Eurostat. (2014). Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics. Retrieved from http://ec.europa.eu/eurostat/documents/747990/6225717/MP-Consolidated-report.pdf

FAO. (n.d.).The Global Partnership on Sustainable Development Data. Working together towards the Global Sustainable Development Goals. Retrieved from http://aims.fao.org/activity/blog/global-partnership-sustainable-development-data-working-together-towards-global

Feyisa, G. L., Meilby, H., Fensholt, R., & Proud, S. R. (2014). Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. Remote Sensing of Environment, 140, 23-35.

Finger, F., Genolet, T., Mari, L., de Magny, G. C., Manga, N. M., Rinaldo, A., & Bertuzzo, E. (2016). Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. Proceedings of the National Academy of Sciences, 201522305.

Flowminder Foundation. (n.d.). The First Application of Mobile Operator Data in Low- or Middle-Income Countries — Modelling Malaria Parasite Movements Between Zanzibar and the Tanzanian Mainland. Retrieved from http://www.flowminder.org/case-studies/the-first-application-of-mobile-operator-data-in-low-or-middle-income-countries-modelling-malaria-parasite-movements-between-zanzibar-and-the-tanzanian-mainland

Flowminder Foundation. (n.d.).Haiti Cholera Outbreak 2010 . Retrieved from http://www.flowminder.org/case-studies/haiti-cholera-outbreak-2010

Fortuna, C. (2016, December 19). 5 Vertical Integration Opportunities For Smart Water & Smart Cities. Retrieved from https://cleantechnica.com/2016/12/19/five-vertical-integration-opportunities-smart-water-smart-cities/

Frias-Martinez, V., & Virseda, J. (2012, March). On the relationship between socio-economic factors and cell phone usage. In Proceedings of the fifth international conference on information and communication technologies and development (pp. 76-84). ACM.

Frias-Martinez, V., Virseda-Jerez, J., & Frias-Martinez, E. (2012). On the relation between socio-economic status and physical mobility. Information Technology for Development, 18(2), 91-106.

Gerber, M. S. (2014). Predicting crime using Twitter and kernel density estimation. Decision Support Systems, 61, 115-125.

Google Earth Engine. (n.d.). Case Studies. Retrieved from https://earthengine.google.com/case_studies/

Gutierrez, T., Krings, G., & Blondel, V. D. (2013). Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. arXiv preprint arXiv:1309.4496.

Haas, E. M., Bartholomé, E., & Combal, B. (2009). Time series analysis of optical remote sensing data for the mapping of temporary surface water bodies in sub-Saharan western Africa. Journal of Hydrology, 370(1), 52-63.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., ... & Kommareddy, A. (2013). High-resolution global maps of 21st-century forest cover change. science, 342(6160), 850-853.

Hansen, M., Potapov, P., Moore, R., & Hancher, M. (2013, November 14). The first detailed maps of global forest change. Retrieved from https://research.googleblog.com/2013/11/the-first-detailed-maps-of-global.html

Hayano, R. S., & Adachi, R. (2013). Estimation of the total population moving into and out of the 20 km evacuation zone during the Fukushima NPP accident as calculated using "Auto-GPS" mobile phone data. Proceedings of the Japan Academy. Series B, Physical and biological sciences, 89(5), 196.

Heerschap, N., Ortega, S., Priem, A., & Offermans, M. (2014, May). Innovation of tourism statistics through the use of new big data sources. In 12th Global Forum on Tourism Statistics, Prague, CZ. http://tsf2014prague. cz/assets/downloads/Paper (Vol. 201).

Henricksen, B. L., & Durkin, J. W. (1986). Growing period and drought early warning in Africa using satellite data. International Journal of Remote Sensing, 7(11), 1583-1608.

Hess, D., & Savitz, J. (n.d.). Global Fishing Watch. Retrieved from http://oceana.org/sites/default/files/global_fishing_watch_report_final.pdf

Ho, A. D., Reich, J., Nesterko, S. O., Seaton, D. T., Mullaney, T., Waldo, J., & Chuang, I. (2014). HarvardX and MITx: The first year of open online courses, fall 2012-summer 2013. Ho, AD, Reich, J., Nesterko, S., Seaton, DT, Mullaney, T., Waldo, J., & Chuang, I.(2014). HarvardX and MITx: The first year of open online courses (HarvardX and MITx Working Paper No. 1).

Hristova, D., Rutherford, A., Anson, J., Luengo-Oroz, M., & Mascolo, C. (2016). The International Postal Network and Other Global Flows as Proxies for National Wellbeing. PloS one, 11(6), e0155976.

Humnet Lab. (n.d.). Retrieved from http://humnetlab.mit.edu/wordpress/news/

Hutchinson, J. M. S., Jacquin, A., Hutchinson, S. L., & Verbesselt, J. (2015). Monitoring vegetation change and dynamics on US Army training lands using satellite image time series analysis. Journal of environmental management,150, 355-366.

Ingram, D. G., Matthews, C. K., & Plante, D. T. (2015). Seasonal trends in sleep-disordered breathing: evidence from Internet search engine query data. Sleep and Breathing, 19(1), 79-84.

ITU. (n.d.). Goal 14, Oceans. Retrieved from http://www.itu.int/en/sustainable-world/Pages/goal14.aspx

ITU. (2006). *Security in Telecommunications and Information Technology: An overview of issues and the deployment of existing ITU-T Recommendations for secure telecommunications*. Retrieved from http://www.itu.int/dms_pub/itu-t/opb/hdb/T-HDB-SEC.03-2006-PDF-E.pdf

Jabari, S., & Zhang, Y. (2014, March). Building detection in very high resolution satellite image using HIS model. In Proceedings of ASPRS 2014 Annual Conference, Louisville, Kentucky.

Jahani, E., Sundsoy, P. R., Bjelland, J., Iqbal, A., Pentland, A., & de Montjoye, Y. A. (2015). Predicting gender from mobile phone metadata. In NetMob Conference in Cambridge, MA.

Janecek, A., Valerio, D., Hummel, K. A., Ricciato, F., & Hlavacs, H. (2015). The Cellular Network as a Sensor: From Mobile Phone Data to Real-Time Road Traffic Monitoring. IEEE Transactions on Intelligent Transportation Systems, 16(5), 2551-2572.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. Science, 353(6301), 790-794.

Ji, Z., Lipton, Z. C., & Elkan, C. (2014). Differential Privacy and Machine Learning: a Survey and Review, 1–30. Learning; Cryptography and Security; Databases. Retrieved from http://arxiv.org/abs/1412.7584

Jiang, S., Ferreira Jr, J., & González, M. C. (2015). Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. In Int. Workshop on Urban Computing.

Kavousian, A., Rajagopal, R., & Fischer, M. (2013). Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants" behavior. Energy, 55, 184-194.

Kearns, J. (2015, July 08). Satellite Images Show Economies Growing and Shrinking in Real Time. Retrieved from http://www.bloomberg.com/news/features/2015-07-08/satellite-images-show-economies-growing-and-shrinking-in-real-time

Kogan, F., Adamenko, T., & Kulbida, M. (2011). Satellite-based crop production monitoring in Ukraine and regional food security. In Use of Satellite and In-Situ Data to Improve Sustainability (pp. 99-104). Springer Netherlands.

Kohli, D., Sliuzas, R., Kerle, N., & Stein, A. (2012). An ontology of slums for image-based classification. Computers, Environment and Urban Systems,36(2), 154-163.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior, 2–5. doi:10.1073/pnas.1218772110

Kreindler, G. & Miyauchi, Y. (2015). Commuting and Productivity: Quantifying Urban Economic Activity using Cell Phone Data. LIRNEasia

Krsinich, F. (2015). Implementation of consumer electronics scanner data in the New Zealand CPI.

KT Corp. (2016, June 27). KT CEO Hwang Chang-Gyu Delivers Speech at UN Global Compact: "Join hands with the UN to prevent the spread of global diseases". Retrieved from http://www.prnewswire.com/news-releases/kt-ceo-hwang-chang-gyu-delivers-speech-at-un-global-compact-join-hands-with-the-un-to-prevent-the-spread-of-global-diseases-300290512.html

Kumar, K. and Muhota, K. (2012), Can Digital Footprints Lead to Greater Financial Inclusion? (pp. 1–4). Washington DC.

Lange, M., Alpert, P., & David, N. (2013, April). Utilizing Mobile-Phone-Link Data to Improve Rainfall Monitoring over Cyprus. In EGU General Assembly Conference Abstracts (Vol. 15, p. 7743).

Lokanathan, S., & Lucas Gunaratne, R. (2014). Behavioral insights for development from Mobile Network Big Data: enlightening policy makers on the State of the Art.

Lokanathan, S., DE SILVA, N., Kreindler, G., Miyauchi, Y., & Dhananjaya, D. (2014). Using Mobile Network Big Data for Informing Transportation and Urban Planning in Colombo. Available at SSRN.

Lokanathan, S., Kreindler, G. E., de Silva, N. N., Miyauchi, Y., Dhananjaya, D., & Samarajiva, R. (2016). The Potential of Mobile Network Big Data as a Tool in Colombo"s Transportation and Urban Planning. Information Technologies & International Development, 12(2), pp-63.

Louail, T., Lenormand, M., Cantú, O. G., Picornell, M., Herranz, R., Frias-Martinez, E., ... & Barthelemy, M. (2014). From mobile phone data to the spatial structure of cities. arXiv preprint arXiv:1401.4540.

Louail, T., Lenormand, M., Ros, O. G., Picornell, M., Herranz, R., Frias-Martinez, E., . . . Barthelemy, M. (2014, June 13). From mobile phone data to the spatial structure of cities. Retrieved from http://www.nature.com/articles/srep05276

Lu D, Weng Q. A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing. 2007;28(5):823–870.

Lu, D., Hetrick, S., Moran, E., & Li, G. (2012). Application of time series Landsat images to examining land-use/land-cover dynamic change. Photogrammetric engineering and remote sensing, 78(7), 747.

Lu, S., Wu, B., Yan, N., & Wang, H. (2011). Water body mapping method with HJ-1A/B satellite imagery. International Journal of Applied Earth Observation and Geoinformation, 13(3), 428-434.

Lu, X., Bengtsson, L., & Holme, P. (2012). Predictability of population displacement after the 2010 Haiti earthquake. Proceedings of the National Academy of Sciences, 109(29), 11576-11581.

Lu, X., Wrathall, D. J., Sundsøy, P. R., Nadiruzzaman, M., Wetter, E., Iqbal, A., ... & Bengtsson, L. (2016). Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh. Global Environmental Change, 38, 1-7.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

Madhawa, K., Lokanathan, S., Maldeniya, D., & Samarajiva, R. (2015). Using mobile network big data for land use classification.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Retrieved from http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*, *90*(10), 60–6, 68, 128. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/23074865

McCauley, D., (2016, April 13). How Satellites and Big Data can Help to save the ocean. Retrieved from http://e360.yale.edu/features/how_satellites_and_big_data_can_help_to_save_the_oceans

Mena, J. B., & Malpica, J. A. (2005). An automatic method for road extraction in rural and semi-urban areas starting from high resolution satellite imagery. Pattern recognition letters, 26(9), 1201-1220.

Mueller, N., Lewis, A., Roberts, D., Ring, S., Melrose, R., Sixsmith, J., ... & Ip, A. (2016). Water observations from space: Mapping surface water from 25years of Landsat imagery across Australia. Remote Sensing of Environment, 174, 341-352.

Mullainathan, S. (2013). What Big Data Means For Social Science. *HeadCon "13 Part I*. Retrieved from http://edge.org/panel/headcon-13-part-i

Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)* (pp. 111–125). IEEE. doi:10.1109/SP.2008.33

Necula, E. (2015). Analyzing Traffic Patterns on Street Segments Based on GPS Data Using R. Transportation Research Procedia, 10, 276-285.

OECD. (n.d.). Retrieved from https://stats.oecd.org/glossary/detail.asp?ID=4344

Office for National Statistics. (n.d.). Retrieved from http://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/theonsbigdataproject

Ohmann, J. L., Gregory, M. J., & Roberts, H. M. (2014). Scale considerations for integrating forest inventory plot data and satellite image data for regional forest mapping. Remote sensing of environment, 151, 3-15.

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, *57*(6).

OCHA. (2016a). Building data responsibility into humanitarian action. *OCHA Policy and Studies Series*, (18). United Nations Office for the Coordination of Humanitarian Affairs. Retrieved from https://docs.unocha.org/sites/dms/Documents/TB18_Data Responsibility_Online.pdf

OCHA. (2016b). *A Human Rights-based Approach to Data: Leaving No One Behind in the 2030 Development Agenda*. United Nations Office for the Coordination of Humanitarian Affairs. Retrieved from http://www.ohchr.org/Documents/Issues/HRIndicators/GuidanceNoteonApproachtoData.pdf

Pekel, J. F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. Nature.

Pekel, J. F., Vancutsem, C., Bastin, L., Clerici, M., Vanbogaert, E., Bartholomé, E., & Defourny, P. (2014). A near real-time water surface detection method based on HSV transformation of MODIS multi-spectral time series data. Remote sensing of environment, 140, 704-716.

Peters, A. J., Walter-Shea, E. A., Ji, L., Vina, A., Hayes, M., & Svoboda, M. D. (2002). Drought monitoring with NDVI-based standardized vegetation index. Photogrammetric engineering and remote sensing, 68(1), 71-75.

Phithakkitnukoon, S., Smoreda, Z., & Olivier, P. (2012). Socio-geography of human mobility: A study using longitudinal mobile phone data. PloS one,7(6), e39253.

Pinkovskiy, M., & Sala-i-Martin, X. (2014). Lights, Camera,... Income!: Estimating Poverty Using National Accounts, Survey Means, and Lights (No. w19831). National Bureau of Economic Research.

Plante, D. T., & Ingram, D. G. (2015). Seasonal trends in tinnitus symptomatology: evidence from Internet search engine query data. European Archives of Oto-Rhino-Laryngology, 272(10), 2807-2813.

Potter, C. (2014). Ten years of forest cover change in the Sierra Nevada detected using Landsat satellite image analysis. International Journal of Remote Sensing, 35(20), 7136-7153.

Potter, C. (2014). Ten years of forest cover change in the Sierra Nevada detected using Landsat satellite image analysis. International Journal of Remote Sensing, 35(20), 7136-7153. (http://www.tandfonline.com/doi/full/10.1080/01431161.2014.968687)

Prasad, A. K., Chai, L., Singh, R. P., & Kafatos, M. (2006). Crop yield estimation model for Iowa using remote sensing and surface parameters. International Journal of Applied Earth Observation and Geoinformation, 8(1), 26-33.

Predpol. (2014, March 07). PredPol Partners LAPD-Foothill Records Day Without Crime!. Retrieved from http://www.predpol.com/predpol-partners-lapd-foothill-records-day-without-crime/

PredPol. (n.d.). About Us | Predictive Policing. Retrieved from http://www.predpol.com/about/

Rama Rao, N., Kapoor, M., Sharma, N., & Venkateswarlu, K. (2007). Yield prediction and waterlogging assessment for tea plantation land using satellite image‐based techniques. International Journal of Remote Sensing, 28(7), 1561-1576. (http://www.tandfonline.com/doi/full/10.1080/01431160600904980)

Real Impact Analytics. (n.d.). Retrieved from https://realimpactanalytics.com/en/data-for-good

Rhee, J., Im, J., & Carbone, G. J. (2010). Monitoring agricultural drought for arid and humid regions using multi-sensor remote sensing data. Remote Sensing of Environment, 114(12), 2875-2887.

Rincon, P. (2016, August 18). Satellite images used to predict poverty. Retrieved from http://www.bbc.com/news/science-environment-37122748

Roberts, Peter, K. C. Shyam, and Cordula Rastogi. 2006. "Rural Access Index: A Key Development Indicator." Transport Papers TP-10. The World Bank Group, Washington, DC.

Rokni, K., Ahmad, A., Selamat, A., & Hazini, S. (2014). Water feature extraction and change detection using multitemporal Landsat imagery. Remote Sensing, 6(5), 4173-4189.

Rosoff, M. (2014, November 24). Here"s how dominant Google is in Europe. Retrieved September 25, 2016, from http://www.businessinsider.com/heres-how-dominant-google-is-in-europe-2014-11

Ruktanonchai, N. W., DeLeenheer, P., Tatem, A. J., Alegana, V. A., Caughlin, T. T., zu Erbach-Schoenberg, E., ... & Smith, D. L. (2016). Identifying malaria transmission foci for elimination using human mobility data. PLoS Comput Biol, 12(4), e1004846.

Samarajiva, R., & Lokanathan, S. (2013). Using behavioral big data for public purposes: Exploring frontier issues of an emerging policy arena.

Samarajiva, R., Lokanathan, S., Madhawa, K., Kreindler, G., & Maldeniya, D. (2015). Big Data to Improve Urban Planning. Economic & Political Weekly, 50(22), 43.

Sannier, C., Gilliams, S., Ham, F., & Fillol, E. (2015). Use of Satellite Image Derived Products for Early Warning and Monitoring of the Impact of Drought on Food Security in Africa. In Time-Sensitive Remote Sensing (pp. 183-198). Springer New York.

Sawaya, K. E., Olmanson, L. G., Heinert, N. J., Brezonik, P. L., & Bauer, M. E. (2003). Extending satellite remote sensing to local scales: land and water resource monitoring using high-resolution imagery. Remote sensing of Environment, 88(1), 144-156.

Šćepanović S, Mishkovski I, Hui P, Nurminen JK, Ylä-Jääski A (2015) Mobile Phone Call Data as a Regional Socio-Economic Proxy Indicator. PLoS ONE 10(4): e0124160. doi:10.1371/journal.pone.0124160

Schuster, N. M., Rogers, M. A., & McMahon Jr, L. F. (2010). Using search engine query data to track pharmaceutical utilization: a study of statins. The American journal of managed care, 16(8), e215.

Serajuddin, U. (2015, April 30). Much of the world is deprived of poverty data. Let"s fix this. Retrieved from http://blogs.worldbank.org/developmenttalk/much-world-deprived-poverty-data-let-s-fix

Serajuddin, U., Uematsu, H., Wieser, C., Yoshida, N., & Dabalen, A. (2015). Data deprivation: another deprivation to end. World Bank Policy Research Working Paper, (7252).

Shi, Q., & Abdel-Aty, M. (2015). Big data applications in real-time traffic operation and safety monitoring and improvement on urban expressways.Transportation Research Part C: Emerging Technologies, 58, 380-394. http://www.sciencedirect.com/science/article/pii/S0968090X15000777

Shiabasaki Lab. (n.d.). Retrieved from http://shiba.iis.u-tokyo.ac.jp/home_en/overview/

Shiabasaki Lab. (n.d.). Retrieved from://shiba.iis.u-tokyo.ac.jp/home_en/research/

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., ... & Hadiuzzaman, K. N. (2017). Mapping poverty using mobile phone and satellite data. Journal of The Royal Society Interface, 14(127), 20160690.

Stucke, M E., Grunes, A. P. (2016). *Big data and competition policy*. Oxford: Oxford University Press

Sundsøy, P. (2016). Can mobile usage predict illiteracy in a developing country?. arXiv preprint arXiv:1607.01337.

Sundsøy, P. R., Bjelland, J., Iqbal, A., D, Pentland, A., Jahani, E., & De montjoye, Y. (2015, January). A cross country study of gender prediction using mobile phone metadata. Retrieved from https://www.researchgate.net/publication/270960813_A_cross_country_study_of_gender_prediction_using_mobile_phone_metadata

Sutton, P. C., & Costanza, R. (2002). Global estimates of market and non-market values derived from nighttime satellite imagery, land cover, and ecosystem service valuation. Ecological Economics, 41(3), 509-527.

Sutton, P. C., Elvidge, C. D., & Ghosh, T. (2007). Estimation of gross domestic product at sub-national scales using nighttime satellite imagery. International Journal of Ecological Economics & Statistics, 8(S07), 5-21.

Tatem, A. J., Huang, Z., Narib, C., Kumar, U., Kandula, D., Pindolia, D. K., ... & Lourenço, C. (2014). Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. Malaria journal, 13(1), 1.

Tatem, A. J., Qiu, Y., Smith, D. L., Sabot, O., Ali, A. S., & Moonen, B. (2009). The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. Malaria journal, 8(1), 1.

Telefonica. (2014, November 27). Big Data for Social Good: Opportunities and Challenges. Retrieved from https://www.telefonica.com/en/web/public-policy/blog/article/-/blogs/big-data-for-social-good-opportunities-and-challenges/

Telefonica. (n.d.). Retrieved from http://dynamicinsights.telefonica.com/smart-steps/our-sectors/for-transport/

Telenor. (2016, March 08). How to use Big Data for Social Good. Retrieved from https://www.telenor.com/media/articles/2016/how-to-use-big-data-for-social-good/

Telenor. (2016, March 09). Big Demand for Big Data: New Telenor study on Dengue Fever in Pakistan. Retrieved from http://www.gsma.com/mobilefordevelopment/programme/digital-identity/big-demand-for-big-data-new-telenor-study-on-dengue-fever-in-pakistan

Think-Asia. (n.d.). Retrieved from https://think-asia.org/bitstream/handle/11540/5146/Proceedings%20of%20the%20Expert%20Meeting%20on%20Crop%20Monitoring%20for%20Food%20Security.pdf?sequence=1#page=147

Thomas NE, Huang C, Goward SN, Powell S, Rishmawi K, Schleeweis K, Hinds A. Validation of North American forest disturbance dynamics derived from Landsat time series stacks. Remote Sensing of Environment. 2011;115:19–32.

Thompson, K., & Kadiyala, R. (2014). Leveraging Big Data to Improve Water System Operations. Procedia Engineering, 89, 467-472.

Tompkins, A. M., & McCreesh, N. (2016). Migration statistics relevant for malaria transmission in Senegal derived from mobile phone data and used in an agent-based migration model. Geospatial health, 11(1s).

Toole, J. L., Colak, S., Alhasoun, F., Evsukoff, A., & Gonzalez, M. C. (2014). The path most travelled: mining road usage patterns from massive call data. arXiv preprint arXiv:1403.0636.

Toole, J. L., Lin, Y. R., Muehlegger, E., Shoag, D., González, M. C., & Lazer, D. (2015). Tracking employment shocks using mobile phone data. Journal of The Royal Society Interface, 12(107), 20150185.

Townsend, A. C., & Bruce, D. A. (2010). The use of night-time lights satellite imagery as a measure of Australia"s regional electricity consumption and population distribution. International Journal of Remote Sensing, 31(16), 4459-4480.

Transport & ICT. 2016. Measuring Rural Access: Using New Technologies. Washington DC: World Bank, License: Creative Commons Attribution CC BY 3.0

Tso B, Mather PM. Classification Methods for Remotely Sensed Data. Taylor & Francis; London: 2001. p. 332.

Tucker, C. J., & Choudhury, B. J. (1987). Satellite remote sensing of drought conditions. Remote sensing of Environment, 23(2), 243-251.

UN Global Pulse (n.d.). Retrieved from http://www.unglobalpulse.org/blog/big-data-development-action-global-pulse-project-series

UN Global Pulse. (2014). Mining Indonesian Tweets to understand food price crises. Jakarta: UN Global Pulse.

UN Global Pulse. (2015). Using mobile phone data and airtime credit purchases to estimate food security. Retrieved from www.unglobalpulse.org/sites/default/files/UNGP_ProjectSeries_Airtimecredit_Food_2015.pdf

UN Global Pulse. (n.d.). A look at the gender distribution of tweets about global development. Retrieved from http://www.unglobalpulse.org/news/look-gender-distribution-post-2015-related-twitter-discussions

UN Global Pulse. (n.d.). Retrieved from http://www.unglobalpulse.org/about-new

UN Global Pulse. (n.d.). Retrieved from http://www.unglobalpulse.org/pulse-labs

Unganai, L. S., & Kogan, F. N. (1998). Drought monitoring and corn yield estimation in Southern Africa from AVHRR data. Remote Sensing of Environment, 63(3), 219-232.

United Nations Statistics Division. (2016, August 16). Business Model for Global Platform for Big Data for Official Statistics in support of the 2030 Agenda for Sustainable Development. Retrieved from https://unstats.un.org/unsd/bigdata/conferences/2016/gwg/Item%203%20-%20Business%20Model%20for%20Global%20Platform%20for%20Big%20Data%20-%2016%20August%202016.pdf

United Nations Statistics Division. (n.d.). Big Data for Official Statistics. Retrieved from http://unstats.un.org/unsd/bigdata/

United Nations Statistics Division. (n.d.). http://unstats.un.org/bigdata/inventory

United Nations Statistics Division. (n.d.). http://unstats.un.org/bigdata/inventory/?selectID=WB54

United Nations Statistics Division. (n.d.). http://unstats.un.org/unsd/bigdata/conferences/2016/presentations/day%202/Grant%20Cameron.pdf

United Nations Statistics Division. (n.d.). International Conference on Big Data for Official Statistics. Retrieved from http://unstats.un.org/unsd/bigdata/conferences/2016/default.asp

United Nations Statistics Division. (n.d.). Using Big Data for the Sustainable Development Goals — UN GWG for Big Data. Retrieved from http://unstats.un.org/bigdata/taskteams/sdgs/

United Nations Statistics Division. (n.d.).Expert Group Meeting on data disaggregation — SDG Indicators. Retrieved from http://unstats.un.org/sdgs/meetings/egm-data-dissaggregation

United Nations. (n.d.).UN projects world population to reach 8.5 billion by 2030, driven by growth in developing countries. Retrieved from http://www.un.org/sustainabledevelopment/blog/2015/07/un-projects-world-population-to-reach-8-5-billion-by-2030-driven-by-growth-in-developing-countries/

University of Columbia. (2016, February 22). Working with Facebook to Create Better Population Maps. Retrieved from http://blogs.ei.columbia.edu/2016/02/22/working-with-facebook-to-create-better-population-maps/

Unsalan, C., & Boyer, K. L. (2011). Multispectral Satellite Image Understanding: From Land Classification to Building and Road Detection. Springer Science & Business Media.

Vaitla, B. (2014). The Landscape of Big Data for Development.

Varian, H. R. (2013), Big Data: New Tricks for Econometrics

Wan, Z., Wang, P., & Li, X. (2004). Using MODIS land surface temperature and normalized difference vegetation index products for monitoring drought in the southern Great Plains, USA. International Journal of Remote Sensing, 25(1), 61-72.

Wang, L., & Qu, J. J. (2007). NMDI: A normalized multi‐band drought index for monitoring soil and vegetation moisture with satellite remote sensing. Geophysical Research Letters, 34(20).

Wästfelt, A., Tegenu, T., Nielsen, M. M., & Malmberg, B. (2012). Qualitative satellite image analysis: Mapping spatial distribution of farming types in Ethiopia. Applied Geography, 32(2), 465-476.

Wesolowski, A., Buckee, C. O., Bengtsson, L., Wetter, E., Lu, X., & Tatem, A. J. (2014). Commentary: Containing the Ebola outbreak–the potential and challenge of mobile network data. PLOS currents outbreaks.

Wesolowski, A., Metcalf, C. J. E., Eagle, N., Kombich, J., Grenfell, B. T., Bjørnstad, O. N., ... & Buckee, C. O. (2015). Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. Proceedings of the National Academy of Sciences, 112(35), 11114-11119.

Wesolowski, A., Qureshi, T., Boni, M. F., Sundsøy, P. R., Johansson, M. A., Rasheed, S. B., ... & Buckee, C. O. (2015). Impact of human mobility on the emergence of dengue epidemics in Pakistan. Proceedings of the National Academy of Sciences, 112(38), 11887-11892.

Wicaksono, P., Danoedoro, P., Hartono, H., Nehren, U., & Ribbe, L. (2011, October). Preliminary work of mangrove ecosystem carbon stock mapping in small island using remote sensing: above and below ground carbon stock mapping on medium resolution satellite image. In SPIE Remote Sensing (pp. 81741B-81741B). International Society for Optics and Photonics.

Williams, S. (2016). Statistical uses for mobile phone data: literature review. Retrieved from https://www.ons.gov.uk/file?uri=/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/literaturereviewonstatisticalusesofmobilephonedatafinal.pdf

Wilson, R., zu Erbach-Schoenberg, E., Albert, M., Power, D., Tudge, S., Gonzalez, M., ... & Pitonakova, L. (2016). Rapid and near real-time assessments of population displacement using mobile phone data following disasters: the 2015 Nepal earthquake. PLoS currents, 8.

World Pop. (n.d.). Retrieved from http://www.worldpop.org.uk/about_our_work/about_worldpop/

Xu, W., Han, Z. W., & Ma, J. (2010, July). A neural netwok based approach to detect influenza epidemics using search engine query data. In 2010 International Conference on Machine Learning and Cybernetics (Vol. 3, pp. 1408-1412). IEEE.

Xu, W., Li, Z., Cheng, C., & Zheng, T. (2013). Data mining for unemployment rate prediction using search engine query data. Service Oriented Computing and Applications, 7(1), 33-42.

Xu, W., Zheng, T., & Li, Z. (2011, October). A neural network based forecasting method for the unemployment rate prediction using the search engine query data. In e-Business Engineering (ICEBE), 2011 IEEE 8th International Conference on (pp. 9-15). IEEE.

Yang AC, Huang NE, Peng C-K, Tsai S-J (2010) Do Seasons Have an Influence on the Incidence of Depression? The Use of an Internet Search Engine Query Data as a Proxy of Human Affect. PLoS ONE 5(10): e13728. doi:10.1371/journal.pone.0013728

Yuan, L., Zhang, J., Shi, Y., Nie, C., Wei, L., & Wang, J. (2014). Damage mapping of powdery mildew in winter wheat with high-resolution satellite image. Remote sensing, 6(5), 3611-3623.

Zhang, R., Su, H., Tian, J., Chen, S., Zhan, J., Deng, X., ... & Wu, J. (2008, July). Drought monitoring in northern China based on remote sensing data and land surface modeling. In IGARSS 2008-2008 IEEE International Geoscience and Remote Sensing Symposium (Vol. 3, pp. III-860). IEEE.