frontier data study

releasing the power of digital data for development

a guide to new opportunities

OCTOBER 2020

FRONTIER TECHNOLOGIES

UKaid from the British people

NIRAS DIGITAL FUTURES HUB

Commissioned by the UK Department for International Development's (DFID)
Frontier Technologies Hub in June 2019.

Data is an essential ingredient of effective decision-making in international development and wider society. New digital data sources, such as so-called 'big data', new technologies such as drones, and new techniques such as machine learning, have already generated significant interest and investment.

This is not least because traditional data sources, such as surveys and administrative records, have not been able to meet some pressing data needs on their own and can be very expensive. But it is also because new data sources can provide completely new kinds of valuable insights.

The Frontier Data Study investigated the potential of this new data frontier for international development: what's already there and what's coming, what are the lessons so far, and how to release the potential within the context of the UK's Department for International Development (DFID).

Rather than just digital data itself, new techniques and technologies were also investigated as they either generate new data with specific qualities and/or they allow for complex datasets to be analysed in new ways.

The following conclusions were based on a broad analysis of existing research, a widely advertised global stakeholder survey with a wide range of respondents, and tailored direct evidence gathering from a wide range of global experts and stakeholders within DFID, including those leading data innovation and staff who use data to help in decision-making.

This is an interactive document and we encourage you to explore all the links and case studies throughout.

There are 8 conclusions we discuss in this report.

1

There is **justified excitement and proven benefits in the use of new digital data sources**, particularly where timeliness of data is important or there are persistent gaps in traditional data sources.

This might include data from fragile and conflict-affected states, data supporting decision-making about marginalised population groups, or in finding solutions to address persistent ethical issues where traditional sources have not proved adequate.

2

In many cases, improvements in and greater access to traditional data sources could be more effective than just new data alone, including developing traditional data in tandem with new data sources. This includes innovations in digitising traditional data sources, supporting the sharing of data between and within organisations, and integrating the use of new data sources with traditional data.

3

Decision-making around the use of new data sources should be highly devolved by empowering individual staff and be focused on multiple dimensions of data quality, not least because there are no "one size fits all" rules that determine how new digital data sources fit to specific needs, subject matters or geographies. This could be supported by ensuring:

a. Research, innovation, and technical support are highly demand-led, driven by specific data user needs in specific contexts; and

b. Staff have accessible guidance that demystifies the complexities of new data sources, clarifies the benefits and risks that need to be managed, and allows them to be 'data brokers' confident in navigating the new data landscape, innovating in it, and coordinating the technical expertise of others.

The main report includes a description of the evidence and conclusions in a way that supports these aims, including a set of guides for staff about the most promising new data sources.

4

Where traditional data sources are failing to provide the detailed data needed, **most new data sources provide a potential route to helping with the Agenda 2030 goal to 'leave no-one behind,'** as often they can provide additional granularity on population sub-groups.

But, to avoid harming the interests of marginalised groups, strong ethical frameworks are needed, and affected people should be involved in decision-making about how data is processed and used. Action is also required to ensure strong data protection environments according to each type of new data and the contexts of its use.

5

New data sources with the highest potential added value for exploitation now, especially when combined with each other or traditional data sources, were found to be:

a. data from **Earth Observation (EO) platforms (including satellites and drones)**

b. **passive location data from mobile phones**

While there are specific limitations and risks in different circumstances, each of these data sources provide for significant gains in certain dimensions of data quality compared to some traditional sources and other new data sources.

6

The use of **Artificial Intelligence (AI)** techniques, such as through machine learning, has high potential to add value to digital datasets in terms of improving aspects of data quality from many different sources, such as social media data, and particularly with large complex datasets and across multiple data sources.

7

Beyond the current time horizon, the most potential for emerging data sources is likely to come from:

- **The next generation of Artificial Intelligence**
- **The next generation of Earth Observation platforms**
- **Privacy Preserving Data Sharing (PPDS) via the Cloud** and
- the **Internet of Things (IoT).**

No significant other data sources, technologies or techniques were found with high potential to benefit FCDO's work, which seems to be in line with its current research agenda and innovative activities.

Some longer-term data prospects have been identified and these could be monitored to observe increases in their potential in the future.

8

Several other factors are relevant to the optimal use of digital data sources which should be investigated and/or work in these areas maintained.

These include important internal and external corporate developments, importantly including continued support to Open Data/ data sharing and enhanced data security systems to underpin it, learning across disciplinary boundaries with official statistics principles at the core, and continued support to capacity-building of national statistical systems in developing countries in traditional data and data innovation.

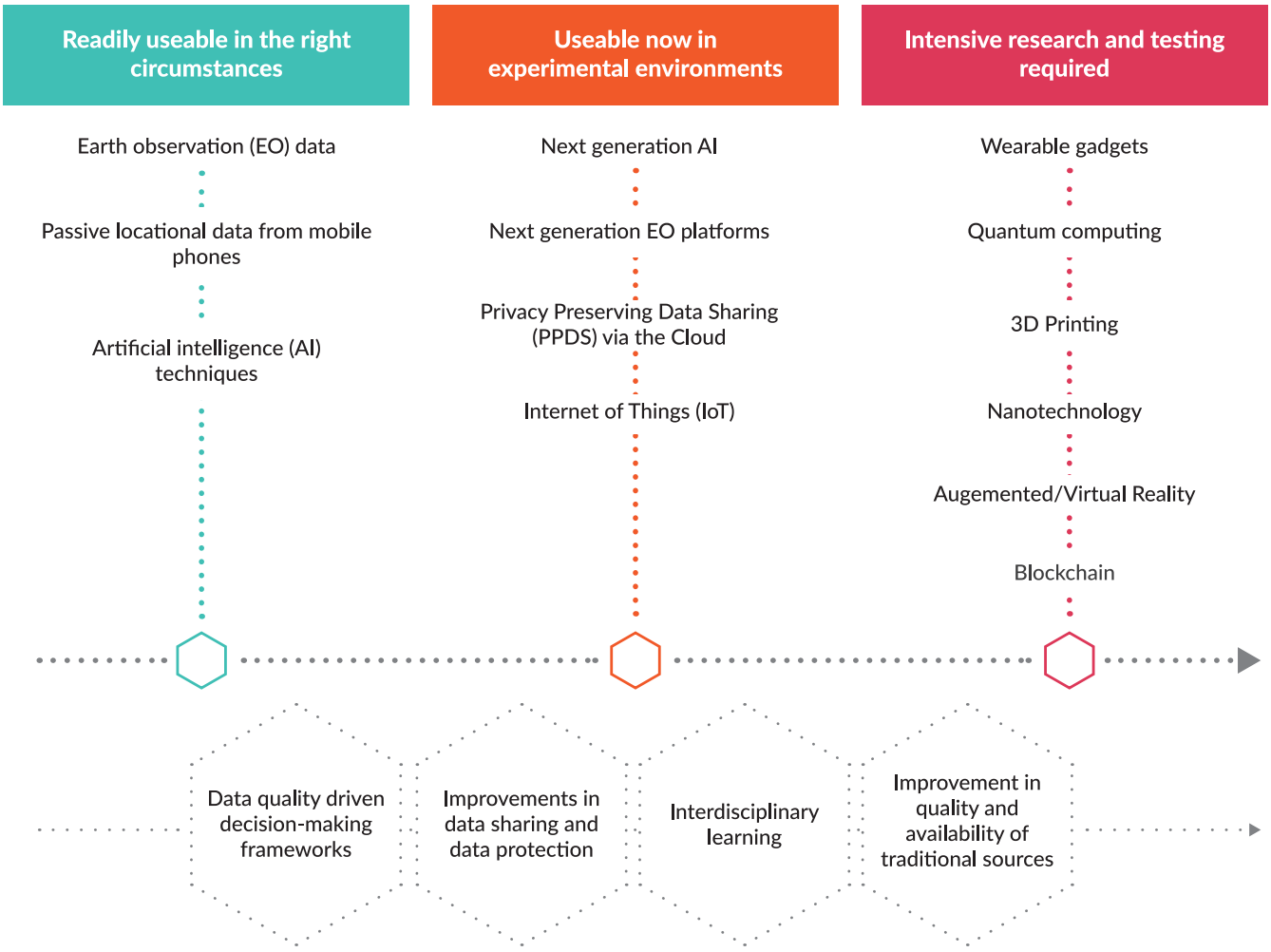


Figure 1: Summary of emerging opportunities in the digital data landscape

FOREWORD: POTENTIAL FOR NEW DATA SOURCES TO TACKLE THE COVID-19 PANDEMIC

Innovating with new data sources from global digital footprints has become an important part of responding to the pandemic in 2020. Traditional academic research and official statistics data are not often set up to give real-time or speedy insights that may be needed for rapid response decision-making, while new data can provide genuinely new kinds of insights not available from traditional methods.

This foreword sets out relevant lessons from the Frontier Data Study.

Passive locational data from mobile phones (automatically recorded by Mobile Network Operators or from apps on smartphones) has huge potential. This includes helping reduce the spread of COVID-19 and understanding the economic impacts of these actions, particularly in combination with EO data. It has not yet been used extensively, not least because of privacy concerns.

But where there are existing frameworks for data protection and anonymisation, such as between the Indonesian National Statistics Office (BPS) and a major national mobile phone operator, this can facilitate more rapid responses.

But there is innovation in other areas, such as Google working with researchers to understand the links between travel patterns and virus transmission rates across several countries using anonymised location data from mobile phones.

Investments now in legal and other agreements with Mobile Phone Operators or smartphone app owners may be of high value, not just in opening opportunities to respond to the current pandemic governments but setting a basis for moving more quickly in response to future similar events.

There is evidence that Governments and data owners are showing some willing to more speedily relax privacy rules for a range of digital data, with the aim of releasing data for research.

A useful blog by Positium looks at the technical and analytical capabilities of mobile positioning to control the spread of infectious diseases, including COVID-19, the pros and cons of different techniques, and an example of real data from Estonia. (Terminology may vary from that used in the Frontier Data Study)

But while this may help in making data available, the ethical risks need close management and, as with other data sources, especially where events are happening at speed.

Artificial Intelligence (AI)-derived data, such as through machine learning, has already been widely used during the pandemic and has provided some useful insights to inform decision-making. As opposed to passive location data, there is also huge scope to harvest the data that individuals actively input to smartphones and other devices, such as gathering opinions shared on social media and even in targeting information tailored to different population groups. But AI-derived data need to be used with caution for a range of significant data quality and ethical reasons.

Table 1 contains some examples of how the new data sources and techniques identified by the Frontier Data Study could be used: Earth Observation (EO) (including satellites and drones), passive location data from mobile phones, and the use of Artificial Intelligence (AI) with digital data.

Intervention Area	Some Solutions	Existing examples that could beadapted in developing countries
Prevention and surveillance	AI-derived data can be used to detect early possible outbreaks and track spread using new and traditional data sources (social media, news reports, air-line ticketing, animal and plant disease networks and official proclamations). Location data from mobile phones can give crucial data on risk profiling locations, identifying actions that will reduce disease spread, monitoring social distancing compliance and influencing factors, and tracing at-risk people exposed to infected Covid-19 individuals. EO data is crucial for understanding the location of transport networks and commercial ports, their real-time use and where closures need to be made to restrict spread.	A Canadian health monitoring platform BlueDot using AI (natural language processing) was able to generate the first alerts of the COVID-19 outbreak in Wuhan (before the WHO). A Lancet publication used AI on social media and news reports from DXY.cn to reconstruct the progression of the Covid-19 outbreak at the patient-level. An American company, UNACAST launched a Social Distancing Scoreboard that grades, county by county in the US, which residents are changing behaviour at the urging of health officials, using reduction in the total distance travelled as a proxy.
Diagnosis and treatment	AI-derived data can speed-up patient diagnosis of COVID-19 reducing pressure on overwhelmed hospitals. AI is being used to identify possible drugs targets.	Chinese hospitals used AI software to read CT lung scans and look for signs of pneumonia caused by coronavirus.
Mitigating economic impacts	Location data from mobile phones can show the social and economic consequences of movement-restriction measures on different sectors, by monitoring changes in movement within different work areas. EO data can monitor changes in key outcomes related to economic activity and human well-being. AI-derived data from Earth observation, MPPD and combined with other data sources can be used to predict economic impacts and explore relationships.	Location data from mobile phones has not yet been widely used for this application but examples exist from other sectors, e.g. assessing the employment impact of auto-mobile factory closures. EO data can monitor changes in trade (by tracking cargo moving through ports), economic activity within commercial and industrial areas (monitoring road traffic and air pollution).

Table 1: Examples of the use of new digital data opportunities in the COVID-19 pandemic

The **best insights for informing decision-making, in terms of data quality, are likely to come from combinations of new data sources and with traditional data** such as surveys and administrative data (such as hospital records). Ethical risks are however potentially increased with the increase in the volume and variety of data held about individuals and its traceability to individuals.

The long term of legacy of ill-thought action in this area needs to be carefully considered. Not least in terms of human rights and people's long-term willingness to share any personal data in the future.

There is no magic wand in using digital data sources. **The benefits, costs, and risks need to be weighed up for each data need and context. Risks also need to be managed**, in particular those related to the cultural, legal, and political context in which the data is used, and the characteristics of each data source.

The NIRAS Digital Futures Hub provides a list of potentially useful data sources and innovations here:

Corona Virus Knowledge Hub

ICT Works has set out some relevant key overall data lessons from the Ebola crisis in developing sound data interventions in times of crisis:

1. efforts need to be informed by and embedded in the local context if they are to succeed
2. look for ways that digital data can amplify local efforts (use a community-driven approach)
3. weak infrastructure can hinder real-time information sharing
4. official statistics can have a 3-week time lag from collection to digitisation
5. use existing digital platforms wherever possible, and
6. spend time building local country capacity to ensure sustainable use.

Despite potential time lags in official statistics, the Frontier Data Study shows that **data innovation should always be seen in the context of official statistics**. Current weaknesses in traditional official data sources in developing countries should not discount them from consideration.

These statistics may have inherently better data quality in numerous aspects and are often already integrated or more ready for integration into decision-making by local/national actors. It may be better to focus innovation on improving the timeliness and other quality aspects of some official statistics, rather than working with new data sources. The relative priorities of each must be weighed up in each case.

However, as the examples above illustrate, there are many cases, where **new data sources could add significant value in complementing official statistics**, particularly where there are gaps that need to be filled quickly. Moreover, innovation in this area could inform longer term sustainable improvements in national data infrastructures.

Open Data Watch have compiled a useful list of sources of information about data availability for new and traditional sources and the challenges to be addressed.

SECTION 2:

INTRODUCTION

2.1 SCOPE

The UK's Department for International Development (DFID), through its Frontier Technologies Hub, commissioned the Frontier Data Study to help it develop an effective approach to the new 'data landscape' in international development. The Study was carried out by the NIRAS Digital Futures Hub.

A DFID staff survey in 2019 showed that so-called "big data" was the 'technology' staff believed had the promise to improve their work. But there was a lack of awareness of new data sources and how to best harness them safely to support their work.

The emphasis of the final Study is on headline conclusions and advice to support strategy with respect to new data sources and to enable staff who are not data experts to engage effectively with a broad range of technical experts to make the best use of them.

The details of all the possible issues that will be encountered by data users cannot be included in a study of this length and within its main purposes.

The limitations on the amount of evidence that could be analysed also meant that conclusions could not be made in some areas of interest to DFID. For example, the need for internal corporate developments to optimise capability to tap into the potential of new data sources.

However, where the Study team picked up some relevant evidence in these areas this has been summarised in Section 5 on further considerations.

2.2 GUIDE TO TERMINOLOGY

The current digital data landscape is awash with terminology that is often ambiguous and not always used consistently.

This can create challenges for both technical and lay audiences. One of the Study's main aims is to give readers a sense of the key factors to be considered in using new data sources in an accessible way. This limits the amount of technical terminology introduced in the report, but a basic level of data and statistical literacy is assumed.

There will also inevitably be technical terms that are ambiguous or are absent in this report, mainly due to the requirement of brevity. DFID or external technical experts should advise DFID staff further where clarity is required.

The following explanations may be useful:

Categorisation of data sources

In some cases, the evidence gathered in this Study led to categorisations of data types that may differ from other categorisations used elsewhere - for instance, some people see 'spatial data' as one type of data, whereas this Study distinguishes between Earth Observation data and location data using mobile phones. This is based on the evidence collected by the Study about common strengths or weaknesses and/or applying generic approaches to data within these categories.

'New Data Sources'

The Study refers to new data sources as its focus of enquiry. This Study was initially commissioned to examine the potential of 'big data'. But this term is not widely used by the Study as it was found to be too limiting in understanding the full range of opportunities or likely developments in the current and future data landscape. Section 3 of the Study describes the types of data included in the Study's scope.

These are broadly data which exist in a digital format but are generated in a variety of ways, 'big' or 'small'. In fact, it was found that the biggest opportunity from new sources of data come from their digitisation, and not necessarily from whether the data is 'big' or something else. The digitisation of traditional sources is also a powerful new opportunity, alongside the wider ability to combine difference types of datasets that digitisation brings.

It is also important that, rather than just digital data itself, new techniques and technologies have been included in the Study as new data sources. This is because they either generate new data with specific qualities and/or they allow for complex datasets to be analysed in new ways. Data flows throughout the digital landscape, changing and taking on new qualities as it journeys through a process or a technology.

Ethics

Ethics are a significant issue with new data sources. The concept is referred to widely in the Study. It encompasses a wide variety of implications from legal to IT security concerns, from consent to cultural sensitivity, to inclusive design and discrimination.

There are also lots of terminologies used by different organisations in respect of ethical data issues such as 'data governance', 'data stewardship', 'data responsibility' and 'data protection'.

The Study was limited in its ability to analyse the validity of different categorisations and definitions in these areas, and can only offer a generic set of conclusions and advice which will need careful consideration in each case when new data sources are being used, alongside the DFID ethics guidelines for research, evaluation and monitoring,

Interoperability

Data interoperability addresses the ability of systems and services that create, exchange, and consume data to have clear, shared expectations for the contents, context, and meaning of that data[1]. It is an important concept for the optimisation of new data sources, particularly when combining data sources.

The Study does not refer to it widely as it is considered a core aspect of the IT environments required to support the transparency and data analysis that affects a broad range of data quality and ethical issues.

The preferred language used in the Study is designed to appeal to a lay audience. It is expected that IT advisors and technical experts who need to support staff will provide tailored advice in this area. Specifically, where data combinations are referred to in this Study, this implies interoperability is required.

2.3 METHODOLOGY

The study used a multi-disciplinary approach drawing on the perspectives of official statistics, data science, innovation in technology and data, and international development practice.

Data collection methods were as exhaustive as possible according to the budget, including an extensive review of the latest relevant global literature and research, a global survey of a wide range of stakeholders and experts in different sectors, the collection and analysis of global case studies focusing on the practical application of new data sources, key informant interviews, and focus group discussions with cross-sections of DFID staff.

The main means of evidence collection in the Study are listed below (further details are in Appendix 3). Data collection took place between July 2019 and April 2020.

DFID staff focus groups

4 focus groups with a cross-section of 26 DFID staff in the UK and a variety of Country Offices (including one focus group in Nepal), representing data decision-makers, data users, those driving data innovation and management in DFID, and those designing data strategies for Portfolios and Programmes.

This included many who were not data experts.

DFID working group

Ongoing inputs on methodological developments and the development of conclusions were given by a small working group of 6 DFID staff which included data/statistics experts, data scientists, and a DFID Country Office representative.

The Digital Futures Hub study team worked closely with DFID's Frontier Technologies Hub.

Literature review

Analysis of a wide range of research and literature related to the key questions. A short-list recommended bibliography is provided in Appendix 2 along with a more detailed list of key reference documents which informed the conclusions about new data sources.

New data sources pilots review

Analysis of key lessons from reports on the experimental use of new data sources in international development. From this research, the Study developed an interactive world map of case studies, provided on the NIRAS Digital Futures Hub webpage.

Global Stakeholder Survey

An online survey targeted different stakeholder groups; with sampling targeted from the following groups: statisticians, donor and implementing organisations, digital entrepreneurs/ conduits, and knowledge drivers/ data scientists) via widespread advertising through social media and direct requests to over 50 organisations. Questions were tailored around the key areas of enquiry of the study.

Respondents included representatives from: UNICEF (Data and Analytics), GIZ, Integrity, Data2X, independent M&E/data experts, Asian Development Bank (ADB), Land Equity, The University of Edinburgh, Practical Participation, ODW Consulting, UNICEF Libya, UNESCAP, UNICEF, European Space Agency (ESA), the Development Cafe, Statistics Netherlands, Palestine Central Bureau of Statistics, Indonesian Central Bureau of Statistics.

Key Informants

Senior Advisor Panel – 8 selected globally eminent experts from a mix of relevant backgrounds (official statistics, data science in international development and local policymaking, cyber-security, monitoring and evaluation, 'tech4good', from a range of public and private sectors) provided a range of pro bono inputs on technical questions throughout the Study[2].

For the list of those interviewed directly see Appendix 3:

Group discussions:

- **Jakarta Frontier Data Study Workshop –** selected experts from public and private sector data innovation organisations based in Jakarta
- World Bank data team for the DFID funded **Partnership for Knowledge based Poverty Reduction and Shared Prosperity in Nepal**

Six bilateral Key Informant Interviews (KIIs) including the Founder of Craig's List, the CEO of Blockchain quantum impact/fintech worldwide, a Program Manager from Global Compact: Data and AI at Facebook, a Senior Data Scientist at International Rice Research Institute (IRRI), the Head of DFID Data Science Hub, and DFID's Deputy Chief Scientific Advisor

2.4 STRUCTURE OF THE REPORT

Section 3 looks at how to best understand and use the new data landscape, namely:

- the new potential benefits it generates
- which data sources have the most potential for use right now and for what
- the key considerations when using those data sources
- which new data sources are going to present the best opportunities in the future
- the key challenges for opening the potential of those data sources in the future, and
- additional focus is given to ethical issues and how new data can help with the 'leaving no-one behind' agenda of the Sustainable Development Goals (SDGs)

Section 4 looks at what is required to release the full potential of new data sources in terms of empowering and inspiring a wide range of staff to integrate new data sources in their work. This includes dedicated user guides to the new data sources with the best potential to be used now.

Section 5 summarises further considerations for possible investigation and action.

Section 6 recaps on the Study's conclusions.

The Appendixes set out more detail on benefits and challenges of some of the data sources discussed in the Study, more detail on the evidence used for all conclusions, and descriptions of the evidence collection activities.

SECTION 3:

UNDERSTANDING AND NAVIGATING THE NEW DATA LANDSCAPE

3.1 OVERVIEW

The conclusions drawn out in this section represent strong lines of consensus that were identified among global technical experts and consistent lessons drawn from case studies and data innovation experiments across all continents. The detailed evidence and references that have informed these conclusions can be found in Appendix 2.

The table below sets out the purposes of each of the sub-sections in this Section.

Section	Purpose
3.1	Developing a useful definition of the data landscape to help orientate and anchor the conclusions.
3.2	Overview of the wide variety of uses for new data sources
3.3	Sets out the basic framework of analysis for this Study and what is required by staff to select what new data sources to use and how.
3.4	Concludes which data sources have the most potential for use right now and for what, and the key considerations when using them.
3.5	Concludes which data sources have lots of potential for the future, their benefits, and key remaining challenges
3.6	Focuses on the implications of new data sources for ethics
3.7	Focuses on the implications of new data sources for the 'Leaving No-One behind agenda'

Table 2: Purpose of each part of Section 3

3.2 WHAT IS THE NEW DIGITAL DATA LANDSCAPE?

The Study found that there is significant scope for new types of data to provide valuable insights to broad areas of work and to fill some significant data gaps. At the very least, these new data could serve as best available proxies in fragile states, and potentially replace some traditional data sources in other contexts at lower cost and better quality.

The survey of global stakeholders showed high levels of interest in the potential of new data sources to support international development, despite a consistent concern with data quality and ethical challenges.

Understanding the different types of digital data

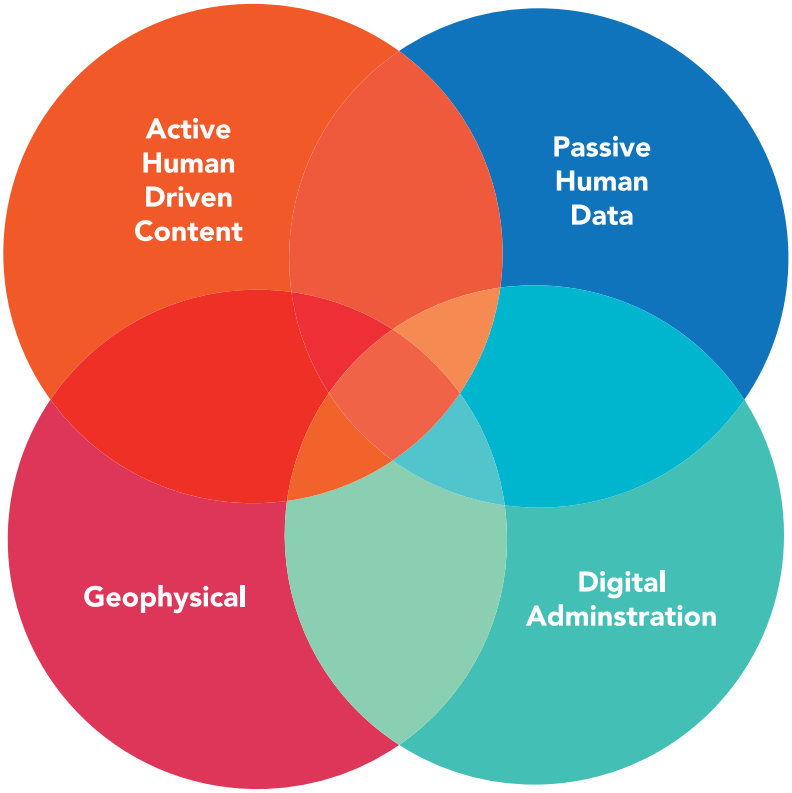
A first step to understanding the potential of new data sources is understanding what makes them different. There is a lot of discussion about so-called 'big data'. But that concept could be too limiting in understanding the full range of new opportunities and risks. There are many different forms of data becoming available. What is most important is the increasing digitisation of this wide variety of types of data.

One way of thinking about the data feeding into the current digital landscape is from four new types of origin (see Figure 1), each with its own strengths and weaknesses.

But boundaries are becoming blurred and there are certain data having the characteristics of more than one type, hence the overlap in the diagram in Figure 1 (right).

Social media
'News' media
Other online 'information'
Communication records

Mobile phone positioning data
Mobile application position tracking
Visual / audio monitoring
Financial / online transactions
Medical devices / Healthcare
Internet searches / click trails
Transport data



EO / remotely sensed data
Terrestrial measurement

Electronic voting
Government records
AI driven decision making

Figure 2: Types of Digital Data

3.3 WHAT ARE THE BEST USES OF NEW DATA SOURCES?

There is a vast array of potential applications of new digital data sources. For example, in improving understanding, implementation, and monitoring of emerging priorities in areas such as climate change mitigation, measuring the Sustainable Development Goals (SDGs), and supporting the 'Leaving no-one Behind' agenda.

Importantly, once any records are digitised this is an important enabler for data sharing and data combination. A simple example with traditional data, which is more and more common, is the allocation of geolocation codes to survey data. Moreover, where admin records such as tax records could be digitised in developing countries, there is an enormous potential to measure a range of phenomena such as mapping poverty, schooling rates, and health issues.

Support for this kind of development should also have long-term benefits for the quality of official statistics and policymaking by national and local governments.

Of course, there is a natural linkage between the types of data set out in Figure 1 above and their suitability to measure certain types of phenomena, such as geophysical data to measure geographical changes and 'active human content' to measure and manage opinions, and so on.

But no one data source or another was found to be particularly good for any specific sector or location. The Study found that choices about data sources and how to use them need to be:

- **driven by the needs of data users**
- **determined by assessments of specific needs and contexts, including how to support better long-term data infrastructures in developing countries, and**
- **patiently and cautiously developed within robust frameworks of data quality and ethics.**

The UN Development Group (UNDG) sets out just some of the possibilities of the use of new data sources in relation the SDGs in the Figure below.

1 NO POVERTY
Spending patterns on mobile phone services can provide proxy indicators of income levels

2 ZERO HUNGER
Crowdsourcing or tracking of food prices listed online can help monitor food security in near real-time

3 GOOD HEALTH AND WELL-BEING
Mapping the movement of mobile phone users can help predict the spread of infectious diseases

4 QUALITY EDUCATION
Citizen reporting can reveal reasons for student drop-out rates

5 GENDER EQUALITY
Analysis of financial transactions can reveal the spending patterns and different impacts of economic shocks on men and women

6 CLEAN WATER AND SANITATION
Sensors connected to water pumps can track access to clean water

7 AFFORDABLE AND CLEAN ENERGY
Smart metering allows utility companies to increase or restrict the flow of electricity, gas or water to reduce waste and ensure adequate supply at peak periods

8 DECENT WORK AND ECONOMIC GROWTH
Patterns in global postal traffic can provide indicators such as economic growth, remittances, trade and GDP

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
Data from GPS devices can be used for traffic control and to improve public transport

10 REDUCED INEQUALITY
Speech-to-text analytics on local radio content can reveal discrimination concerns and support policy response

11 SUSTAINABLE CITIES AND COMMUNITIES
Satellite remote sensing can track encroachment on public land or spaces such as parks and forests

12 RESPONSIBLE CONSUMPTION AND PRODUCTION
Online search patterns or e-commerce transactions can reveal the pace of transition to energy efficient products

13 CLIMATE ACTION
Combining satellite imagery, crowd-sourced witness accounts and open data can help track deforestation

14 LIFE BELOW WATER
Maritime vessel tracking data can reveal illegal, unregulated and unreported fishing activities

15 LIFE ON LAND
Social media monitoring can support disaster management with real-time information on victim location, effects and strength of forest fires or haze

16 PEACE, JUSTICE AND STRONG INSTITUTIONS
Sentiment analysis of social media can reveal public opinion on effective governance, public service delivery or human rights

17 PARTNERSHIPS FOR THE GOALS
Partnerships to enable the combining of statistics, mobile and internet data can provide a better and real-time understanding of today's hyper-connected world

Figure 3: UNDG examples of how data science and analytics can contribute to the SDGS[3]

3.4 HOW TO NAVIGATE THE NEW LANDSCAPE AND MAKE THE RIGHT CHOICES?

Decision-making around whether to innovate with new data sources and the choice of data source(s) should always depend on weighing up the benefits, risks, and costs of a range of options, including using new and/or traditional sources, in any specific context. But the potential is too good to ignore.

Figure 3 below shows one great example of how new data sources have been used, including how combining new data sources can yield even more effective results.

It has taken hundreds of years for official statistics to develop methodologies and quality management regimes for the use of traditional data sources. But even then, there is still a long way to go in terms of obtaining high levels of public trust. The excitement around new data sources should be seen in this light.

More robust methodologies should be developed, and some lessons still need to be learned on-the-ground. Fortunately, already existing frameworks and lessons from official statistics can be used and from which cautious steps can be taken forward. There are also lots of examples of challenges with new data already being quickly and successfully addressed to great benefit, many of which are included in this Study and its user guides.

Figure 4: Using drones for assessing flood risks in Sri Lanka

The World Bank in partnership with the Sri Lankan Government Disaster Management Center, as well as the Survey Department, assessed flood risk in the Mundeni Aru and Attanagalu Oya River Basins. The World Bank is using a combination of satellite and drone imagery to improve OpenStreetMap data to better assess financial and humanitarian risk from recurring floods.

The drone imagery composited on top of satellite imagery provides extremely detailed imagery and terrain models for vulnerable areas.

The portability of the drones, matched with the high spatial resolution of imagery they provide make them an excellent tool to aid in mapping where aerial or walking surveys are out of date, non-existent, cost prohibitive or otherwise hard to obtain.

Using data quality dimensions to guide data choices and analysis

A set of **Data Quality Dimensions**, as set out in the table below, were adapted by the Study from international standards in official statistics and used as the framework for analysis. These could also provide an essential future refence point for staff in DFID in deciding how to use new data sources and how to interpret the results.

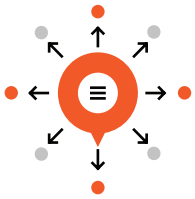
The dimension of **Relevance** is worth underlining as user needs should always drive data choices rather than the supply of new data or innovative ideas. It is useful to consider available data options. But **the starting point must be in examining their relevance, not in finding in a way to use something just because it is new or fashionable.**

Quality Dimension	Main Components for Assessment
Relevance – how far the information meets the needs of the development intervention in question	What are the uses of the information and how well do the data meet those needs?
Reliability/Accuracy – how far the information correctly describes what was intended to be measured and at enough level of detail for decision making	How are the main sources of error in the data quantified? How stable and consistent are the processes and analyses conducted? Is information disaggregated appropriately? Are there any population groups that could be better identified or are being left behind (for new data sources, especially those who are not plugged into the digital data landscape)? Are any modelling assumptions accurate and transparent?
Timeliness – how far the information is current and provided to a timetable that fits with decision making	How long does it take to process and publish the information? Is information available frequently enough to inform programme management decisions? What is the possibility or ongoing cost of getting the information at the right time?
Integrity and Ethics – maintaining the confidence, trust and objectivity of those using the information	Is measurement undertaken cost-effectively and ethically, not placing undue burden on respondents? Is deliberate falsification of the data possible/likely? What safeguards are in place to ensure anonymity and data security now and a long way into the future?
Coherence & Comparability – the degree to which different data sources use the same methods, standards and concepts, and can be compared over time and domain	Are data produced using documented and harmonised methods, standards and concepts across time and/or across different similar sources? Are data alike over different domains (i.e. spatial, sector, population, household type)? Can outputs be compared over time?

Table 3 - Data Quality Dimensions for decision-making in international development[4]

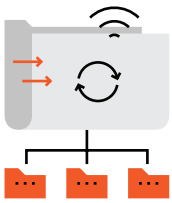
3.5 WHICH NEW DATA SOURCES HAVE THE HIGHEST POTENTIAL NOW?

Based on consistent evidence available from the analysis of existing research (see evidence base in Appendix 2) and the views of stakeholders (see Appendix 3), the Study found that:

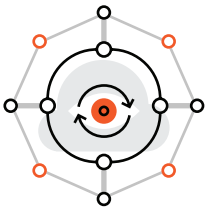


Often **the most promising uses of any of new data types are in combination with other data sources, new or traditional.** One of the reasons for this is that, unlike more traditional sources, most digital data is not being created for the purposes for which it might be used in international development.

Therefore cross-tabulation and sense-checking across several sources is often necessary to provide assurance about what the data is telling us. But it is the digital nature of this data which makes it combinable and therefore even more powerful.



In general, according to the categories in Figure 1, **geophysical and passive human data sources had the most potential;** while **digital administration records,** where they are available, had a lot of potential but these are not common yet in the developing world.



Active human driven content (such as from social media) had less potential. But Artificial Intelligence **(AI) could be used with such data to provide valuable insights** and to discover new and powerful insights across all different types of digital data.

The figure below shows a good example from the UN Pulse Lab in Kampala of how Artificial Intelligence (AI) is being used to analyse social media and radio show content.

Figure 5: Using 'Big Data' and AI to support peace and security in Uganda and Somalia

The UN Pulse Lab in Kampala and partners have explored the utility of analysing data from social media and public radio broadcasts to extract insights to feed early warning systems and inform peace and security processes.

The first test case used data extracted from social media, namely posts from public Facebook pages and groups, to analyse how influencers and fake news might be shaping discussions among online users in Somalia and to identify trending topics relevant to SDG16 – peace, justice and strong institutions.

The second case study analysed public discussions on radio shows to detect instances of rumours and misconceptions, and of social tensions as reported by listeners in Uganda.

These initial studies showed that analysis of 'big data' sources from social media and radio shows can provide rich and timely information for effective conflict mitigation by identifying trends as they emerge and monitoring contexts as they evolve.

The Study considered how best to categorise the most promising opportunities available and which types of data to focus on. Considering the balance of strengths and weaknesses of different data sources and the commonalities between, it was concluded that:

Earth Observation

Mobile Data

Artificial Intelligence

Earth Observation (EO) data (e.g. from satellites, drones, and other 'remote sensing' data), and **passive location data from mobile phones** are currently the most technically robust options from digital data sources as they tend to have more consistently robust characteristics across a number of the Data Quality Dimensions (as per Table 3).

Artificial Intelligence (AI) techniques, including things like mining of social media data but also as a very powerful technique to combine different data sources, can produce benefits in one or more dimensions of quality, such as in improving timeliness or helping to check the accuracy of other data sources.

AI is a technique that can release the potential of most digital data sources. For instance, satellite images which can now identify individual trees are being combined with AI to show the degree and prediction of deforestation caused by palm oil farming, soy plantations, and illegal logging[5].

In terms of the Data Quality Dimensions, the Study determined the following general points about these data sources:

- **EO (including drones) and passive location data from mobile phones**, while having important differences, both have a high potential to provide:
    - high quality data across a wide range of scenarios in terms of **Reliability/ Accuracy** and **Relevance**, given their proven precision and if the data covers enough of what is needed to be measured (e.g. population groups or land areas)
    - potentially good **Timeliness** depending on how and when data can be accessed and who it is owned by (in the short and long term)

- **AI**, if done well, can give powerful insights (**Relevance**) and can help in combining large and dynamic datasets (**Coherence and Comparability**). But its **Reliability/Accuracy** is unproven in many fields, particularly with certain datasets such as social media, and a lack of transparency can hinder data quality in all areas. Particular attention is needed to ensure:
    - transparency in strengths and weaknesses of an application
    - at least minimal accuracy, and,
    - optimal ethical protections.

Figure 5 (right) shows an example from DFID's Frontier Technologies Hub of how AI is being used in medical diagnoses in South Africa. Figure 6 how mobile phone data can be used to map poverty levels.

Figure 6: AI4TB - Testing and learning from the use of AI derived data in medical diagnoses

Vast numbers of people from across southern Africa who came to work in South Africa's gold mines were left with devastating occupational lung disease.

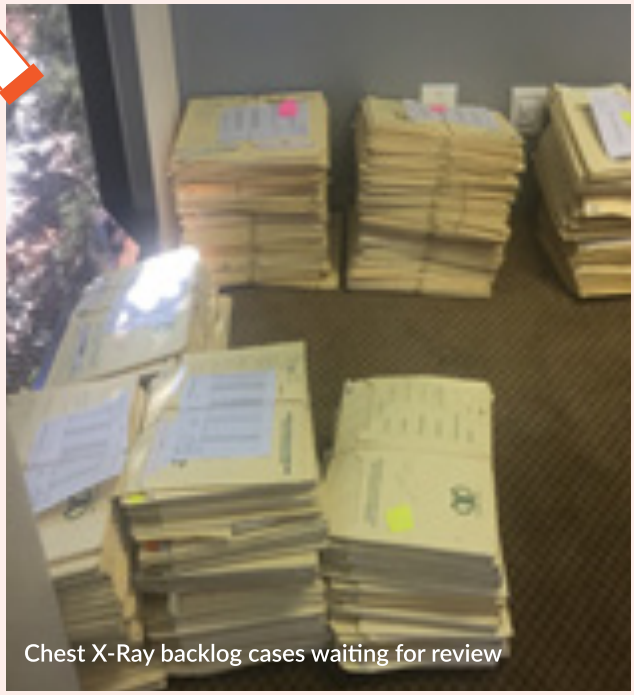
Hundreds of thousands of active and former miners must wait on average 5 years to get assessed for the health and social benefits to which they are entitled. Solutions were sought for using AI to process data to improve the efficiency of finding and assessing these miners.

Demographic and exposure information were used to determine the likelihood of a compensable lung disease, so this could be used to both prioritize clinical assessment as well as to potentially improve the accuracy of clinical diagnosis.

Computer-assisted detection (CAD) of TB and silicosis on chest X-rays was used to determine a preliminary diagnosis so this could be used to triage as part of a more efficient process requiring less scarce medical expertise.

Reflections from the researchers are important. The elements that lean towards AI are:

- the sheer magnitude of the task at hand, and the need for technological assistance to meet the challenge of reducing a 500-day backlog
- the potential to create tools that ex-miners can themselves use to assess their own likelihood of having a compensable disease, and
- potentially more consistent adjudication than when decisions are left to diverse practitioners.



Chest X-Ray backlog cases waiting for review

But they pointed out various stakeholders have considerable concerns that need to be addressed before any widespread implementation can be pursued:

- potential impact on de-skilling in existing medical personnel as well as dampening the emphasis on the development of more occupational lung disease experts to meet ongoing and future diagnostic needs
- concerns about transparency and accountability for clinical decisions made
- issues related to data security
- impact of private ownership of the technology on public sector funding, and
- biases and reliability of the AI.

Figure 7: Using passive location data and other call data records (CDR) to estimate poverty levels- Cote D'Ivoire – UN Pulse Lab

No full survey of the country's population has been published since a civil war in the 1990s. Researchers (Smith et al., 2012) used anonymised CDRs of five million Orange customers to assess both the level of activity among subscribers and locations where calls were made.

On the basis that higher levels of mobile communication and a wider range of calls could be used as a proxy indicator for prosperity, poverty levels of eleven regions of Côte d'Ivoire were quantified.

The estimate was validated when compared with a multidimensional poverty index created by University of Oxford, which uses indicators such as poor health, lack of education, inadequate living standard and threat from violence among other factors.

Combined with the passive location data from mobile phones, this allowed for new insights around the geographical distribution of different poverty levels.

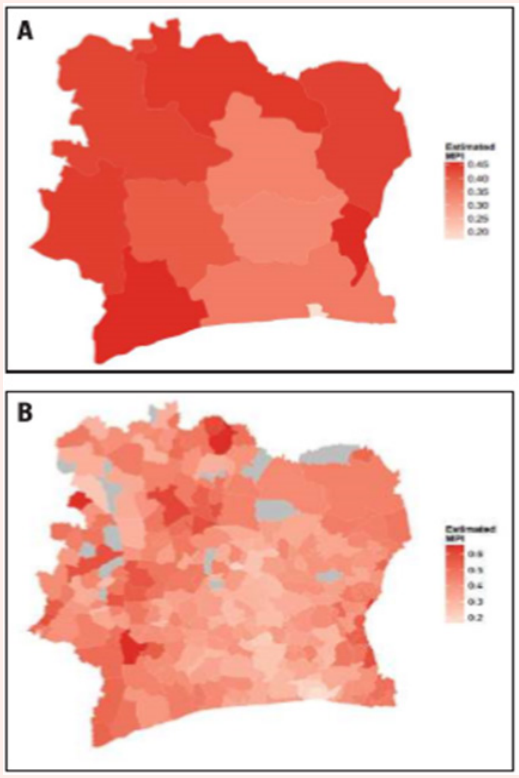


Diagram A shows poverty map estimated based on the antennas in the eleven major regions of Cote d'Ivoire, where the darker areas indicate higher estimated poverty level. Diagram B shows the Department poverty levels as approximated by the model used on regional level indicating the finer granularity possible when using CDRs. Source: Smith et al., 2012.

The quality dimensions of Coherence and Comparability and Integrity and Ethics are highly context specific dimensions of data quality. This underpins the conclusion that there is no one-size-fits all solution for deciding on which digital data source to use.

Coherence and Comparability will depend on how different data sources are combined or if the data source being used changes in nature over time or domain. For EO data, Coherence & Comparability is a very strong factor if the intention is to compare results among different countries.

On the other hand, ethics is an issue that needs close management across all new data sources in any specific context, even though new data sources can give significant ethical benefits (see later section).

Understanding the benefits and challenges of high potential digital data sources

The tables that follow set out the main benefits and challenges to be addressed of the new data sources this Study found to have the highest potential.

The Study has also developed tailored guidance for staff to inform decision-making. Click on the Frontier Data Study Guide links to access these short non-technical online guides, which include detailed case studies, information on how to optimise the benefits and manage risks, guidance on ethics, and checklists on the most important things to do when thinking about using these types of data.

Earth Observation (EO)

EO data generally come from satellites, and airborne and in-situ sensors (such as in oceans).

Main benefits:

EO data can provide accurate and reliable information and insights on both the physical and human environment. Recently EO data has become particularly valuable when combined with other data sources, such as data from smartphone apps to verify EO data in providing evidence of illegal logging.

The full range of applications is potentially huge across all areas of development as it has high frequency, high geographical coverage, and can often be free.

EO technology of course includes photography, but extends to the collection of other data such as through radar, hyperspectral imaging (basically the analysis of a wide spectrum of colours), LIDAR (light detection and ranging – basically using light like a radar does with sound, normally through lasers, to develop 3D models), and about physical and chemical features of the atmosphere (temperature / pressure /chemical composition etc.). Such data can be used to help monitor the state of a number of earth systems, such as oceans, coasts, rivers, soil, crops, forests, ecosystems, and human phenomena like built infrastructure, transport, and daily or long-term migration patterns.

Main challenges:

'Ground-truthing' – especially if used on their own, these data sources risk being misleading as the view they give relies on a one-dimensional perspective. At least some 'ground-truthing' is required to help calibrate and validate; and they are most effective when combined with other types of data in terms of overall analysis.

Privacy, security and surveillance issues relating to data held about private property and individuals' location can be a significant risk. This includes both data security issues and gaining consent, particularly in relation to how data might be stored or shared. There may also be legal challenges about the certainty of personal consent when using data from third-party providers.

Ensuring data can be accessed at an economical cost and over the time period required - EO data are produced by a range of providers, including governments/military agencies, space agencies, and the private sector. Due to the variety and their intended purpose, access and availability also varies greatly between these providers in terms of costs, coverage, resolution and timeliness. Although space agencies generally have long term and sustainable data commitments at free or restrained costs (for certain purposes only, e.g. research and non-commercial models), in general the standardised reliability of data availability cannot be guaranteed.

EO data via Drones

Sometimes called Unmanned Aerial Vehicles (UAVs)

Main benefits:

Data from drones (UAVs) may be considered part of the same family as data from Earth Observation (EO), and most of the guidance above on EO data applies. But there are some important differences.

Drones take high-quality aerial photographs and video and collect vast amounts of imaging data in places which are inaccessible or are too large for traditional methods. For example, after natural disasters to aid in security and recovery efforts. These high-resolution images can be used to create 3-D maps and interactive 3-D models for a range of purposes. Since they use GPS (Global Positioning System), they can also move to precise locations. This is especially helpful in such things as identifying weed infestations and monitoring crop health. They have a greater range of movement than manned aircraft or satellites, can fly lower and in more directions, allowing them to easily navigate traditionally hard-to-access areas.

Main challenges:

There are similar **privacy and ethical issues** to other EO data. Unlike other EO data, which is mostly taken without people's knowledge, drones will often draw the attention of the communities in which they are operating. Perceptions may therefore be negative unless there is good communication and community buy-in. In conflict environments they may be seen as a threat by the public. There are also limitations since the widespread use of drones is relatively new, and legislation is still catching up. Laws are different in each country, so the opportunities and risks are highly determined by location.

The **technology must also be robust enough** to avoid mid-air collisions, avoid falling onto people or crashing into buildings, including by being adaptable to local conditions such as different fuel types/ quality.

Passive location data from mobile phones

Geographical data is possible from almost any mobile network by (a) tracking the location of mobile phones in real time through a network of antennas or Global Positioning System (GPS) or (b) using location information from historical log files stored by mobile service providers (Call Data Records), or from geo-location data automatically recorded on smartphone apps. This study uses the term 'mobile phones' to include smartphones.

Globally other terms which relate to this category are Mobile Phone Positioning Data (MPPD) and Mobile Phone Data (MPD). But these do not clearly include smartphone apps and there is some confusion as to whether they refer to other types of Call Data Records (with significantly different benefits and challenges) or just those that relate to geolocation. Mobile phone data, such as general call or SMS records, 'active' location data (ie self-reported location), and non-location data from apps on smart phones, can be analysed using AI and other techniques.

As data sources they should therefore be considered separately, as there are very different quality dimensions at play. See AI section below.

Main benefits:

Mobile Phone Operators' systems and smartphone apps on mobile phones generate a large amount of phone usage data with location information attached to it, which allows for understanding when and where people are moving, and mobile phone ownership can be very high in many developing countries[6].

These data can allow for highly accurate tracking of population movements over space and time, such as migration across uncontrolled borders[7], tourism and commuting patterns, population movements during disasters and so on.

The location of a mobile phones is a reasonably reliable proxy for the location of people. Even when the data must be purchased, the cost-benefit ratios can be transformative compared to other available data sources, such as with poor quality population registers or highly expensive household surveys/ censuses, especially due to the accuracy of the positioning, the scale of the data available, and the ability to link it with other types of data generated by mobile phones (such as understanding the reason why someone is moving from their social media posts/SMSs etc.).

Main challenges:

One of the most important challenges is having good knowledge of the representativeness of the data; understanding who owns the phones being tracked and their characteristics is crucial to releasing the benefits (see user guide).

While in some limited cases, sampling may be improved on surveys as there is the possibility of tracking most people in a population group where there are high levels of phone ownership, this would require a highly regular pattern of phone ownership and use across different population groups. But the distribution and patterns of use of mobile phones can vary significantly across different population groups, different age, social, and income groups, nationalities and tourist types use mobile phones differently.

This issue appears to be a greater challenge when using geolocation data from smartphone applications, not least because smartphones in developing countries may be a small proportion of the overall total of mobile phones and apps can be used on multiple devices, some of them more static such as laptops.

The other main limitations on what can be achieved are due to:

- local legal frameworks
- general ethical concerns, including regarding the consent of phone owners and ongoing data protection issues
- levels of monopoly of Mobile Phone Operators (MPOs) or app owners; data can better and more easily managed, including sampling issues, where one Operator or app dominates the national or local market
- the willingness of Operators and application owners to cooperate on data sharing and over time, and
- in some cases, national boundaries can inhibit cross-border monitoring as people may have to switch their mobile phone provider, or do not use their phone. Tracking data from smartphone apps has a potential advantage here. But it is also possible to use mobile phones to a certain extent, for instance given that antennae coverage sometimes extends some way across borders, it is possible to monitor when and where people switch SIM cards, or via cooperation between two national MPOs

Artificial Intelligence (AI) techniques (such as Data Mining and Machine Learning)

Based on algorithms, computer systems and programs can perform tasks that previously required human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages. Currently, data mining and machine learning are being widely tested and used to draw previously hidden patterns from large and complex datasets, including related predictive capabilities.

At the moment, the most promise is in using data mining or machine learning in terms of human digital footprints, through information that is given in digital forms either passively or actively, or in finding patterns across these sources and other digital datasets such as satellite data and mobile phone data. In the future, the use of AI in the Internet of Things (i.e. objects sharing information with the internet and between themselves via the internet) may also give some reliable insights that are useful in international development.

Main benefits:

AI has the potential to both provide data to inform decision-making and to help implement objectives though automated decision-making, not least to fill some data gaps or provide insights related to monitoring and implementing the Sustainable Development Goals (SDGs). Although there are lots of limitations and challenges, the potential to add value is huge across many areas.

AI opens the possibility of finding patterns across all types of digital data, and the possible application of AI techniques are virtually endless, including things like analysis of social media posts and smart phone usage data.

Most promising is perhaps that AI can facilitate the combining of different types of data. For example, social media data and EO data can be combined to track post-disaster migration and the reasons for it. This could be transformative across a multitude of development needs. As well as providing insights and helping monitor developments, AI can also help implement development objectives, such as by analysing digital footprints to identify individuals with social care needs or to tailoring messages about health care practices to suit an individual's way of thinking, or by providing basic services via so-called 'chatbots'.

Main challenges:

Accuracy and ethical risks are very high in most cases, and sometimes the two issues are inter-linked. This stems from a wide range of challenges that will vary from case-to-case. Generally, the following may have to be addressed:

- biases in algorithms that may not initially be transparent - these will be specific to each case, but often will include incorrectly making assumptions about the way people think, act, or their physical characteristics, importantly in failing to take account of cultural/linguistic/biological differences between different population groups
- biases from using data that is only available for certain population groups, e.g. the digitally active, and
- leakage or hacking of highly personal data and its misuse by external parties.

3.6 WHICH NEW DATA SOURCES HAVE HIGH POTENTIAL BUT NEED MORE TESTING AND RESEARCH?

This Study found that developments in the four areas set out in this section are the most promising in terms of improved data availability or analysis techniques but there is a need for more testing and research, including for related technologies to become more advanced and robust. This assessment is based on:

- the technological developments anticipated by existing studies
- the level of existing evidence available on each technology's potential
- their current applicability to developing country contexts, and
- a high-level assessment of their generic characteristics against the Data Quality Dimensions

For each of these four high potential emerging data sources, the **Table in Appendix 1 gives further detail of the potential data applications and the challenges.**

Importantly, although there are some developments happening in some of these areas already and with some success, there are some key issues that need to be further tested and resolved before data from these technologies can be effectively mainstreamed, as per below. Some of the main issues withing the Data Quality Dimensions analysis framework (see Section 3.4) are also highlighted, but it must be remembered that analysis against these dimensions is always context specific.

1

Internet of Things (IoT) – Depending on levels of technological penetration in the developing world, near instant data with the use of technological infrastructures/ devices that interact with each other digitally may increasingly find application in different domains, such as home and industrial automation, medical aids, mobile healthcare, intelligent energy management, automotive, traffic management. 5G and future Gs connection will allow for an even more rapid transfer of data and information and interaction of devices.

This could allow for high levels of **Relevance** and **Timeliness** in getting information about or to inform related human activity, but, especially where large amounts of data are processed, **Accuracy** and **Integrity and Ethics** challenges remain as per those already known about with the use of AI to process data. Data from this technology is already being tested out, such as in Kenya where Internet of Things (IoT) stations across the country provide information on weather forecasts to inform more transparent, efficient, and scalable crop insurance platforms for smallholder farmers[8].

2

Privacy Preserving Data Sharing (PPDS) via the Cloud – The increasing use of Cloud computing capabilities will support Open Data (see section on ethics below), allowing anyone to access scalable and on demand processing power from anywhere in the world. Combined with the right solutions, this provides for platforms to make use of data in aggregate without sharing the underlying data (eg through encryption).

There could be some improvements on existing sources in terms of **Relevance** (getting access to the data required), in **Timeliness** (getting access more quickly), and in terms of **Integrity and Ethics** this could provide solutions to many of the ethical and privacy challenges with digital data sources. However, the challenges in terms of the risks of data sharing need further exploration, so the more security-robust techniques will hold the most promise. In connection to this, there is a balance to be struck with trying to achieve improved Accuracy – a key question to be answered is how one accesses the highly disaggregated data needed to support the Leaving No-one Behind agenda while avoiding the risks of disclosure.

There are a range of applications that need to be further tested for practical needs and implications including Secure Multiparty Computing (MPC) such as Zero Knowledge initiatives (such as qed-it.com, as infosum and endor) and Google's Private Join and Compute). Existing applications such as PowerBI, Tableau, R Shiny Web Applications, and interactive reporting using R Markdown, also provide interesting opportunities for a form of data sharing and analysis through the Cloud.

3

Next generation Artificial Intelligence (AI) – Humanity's ability to create data is still, overall, ahead of its ability to solve complex problems by using the data. Up to now machines have mainly carried out well defined tasks such as robotic assembly, or data analysis using pre-defined criteria. But an age is arriving where machine learning, via such things as Advanced Neural Networks (ANNs), will allow machines to interact with their environment in more flexible and adaptive ways. The data collected and analysed by such machines could lead to much improved automated decision-making and provide humans with insights that were previously impossible.

Currently, there is what is called "narrow AI"—single-task applications AI such as image recognition, language translation, and autonomous vehicles. In the future, researchers hope to achieve "artificial general intelligence" (AGI). This would involve systems that exhibit intelligent behaviour across a range of cognitive tasks. However, these capabilities are not estimated to be achieved for decades[9].

In terms of the Data Quality Dimensions, the more sophisticated the AI being used then the greater the amplification of potential benefits but also of the risks (as in Section 3.5 above). Significant learning from the use of current AI techniques will need to be applied and tested in each individual circumstance.

4

Next generation Earth Observation (EO) – There are plans for significantly increasing the number of satellites in lower orbit and for mid-altitude observation devices, including 'pico'/nano satellites.

Lower orbit devices can provide for greater **Timeliness**, as they circle the Earth more quickly and can provide more frequent data. In terms of pico/nano satellites, one of the main benefits is that, flying in constellation and working in tandem, they could also provide for better **Relevance and Accuracy** with wider coverage, quicker repeat times over locations, and multiple measurements taken from different linked instruments.

Video data from satellites may also become a more mainstream data source given their potential high levels of **Accuracy**. A few companies are already provide video data commercially, eg Earth-I (https://earthi.space/). But this is not widely available and often focussed on non-development needs so **Relevance** is an issue. This sort of data is also very sensitive, so **Integrity and Ethics** issues need to be thoroughly examined.

Broader trends

There are some overall trends to anticipate that will inform approaches to emerging data sources. These imply obligations on the international development community to explore the potential of the new data sources, such as the ones listed above, and develop effective techniques to manage risks.

More extensive use of Earth Observation and passive location data from mobile phones can be predicted with some certainty. Both are widely subject to ongoing successful experimentation and learning and seem likely to increasingly prove their potential.

As noted above, there is also evidence of their steadily increasing capability to help accurately monitor at least geographical phenomena, in a way that that can add a lot of value to the interpretation of traditional data.

There may also be a growing political pressure to use these more reliable new data sources, specifically in terms of supporting policy priorities such as a focus on Climate Change.

AI and Deep Learning (see Appendix 1) can both be expected to expand the boundaries of the possible that should be explored.

There will be at least a political pressure to examine how they can be used and managed within the international development context, given their increasing prevalence in human fields of experience and potential cost savings on traditional data sources.

In all areas, but particularly with AI and Deep Learning, the nature of these techniques means that **managing ethical risks and public perceptions will likely be enduring challenges**, particularly in ensuring that data security and legal provisions maintain pace with the implication of rapid evolutions of technology.

Figure 7 (right) sets out recommended reading for better understanding the technological developments that are likely to impact on the future of how UK development assistance is carried out.

Figure 8: Recommended reading on future data opportunities

UNESCAP Frontier Technologies Report

Gartner Hype Cycle Report on Emerging Technologies

Future Trends in Geospatial Information Management: the five to ten-year vision (UN Committee of Experts on Global Geospatial Information Management)

Institute of Development Studies (IDS) Frontier Technologies Report

Oxfam: Global Megatrends – Mapping the forces that affect us all

Other data sources

The Study also found there are other emerging technologies that are of interest in terms of seeing if they become better suited or can be better steered towards international development data needs in a longer time-frame in terms of the data they would generate:

- Wearable gadgets/tech
- Quantum computing
- 3-D printing
- Nanotechnology
- Augmented/virtual reality

Unlike the other higher potential areas, there is little existing evidence about the development data needs that these technologies might meet.

Also, in most cases, the widespread use of these technologies in developing country contexts is likely to be minimal for some time to come. Some potential benefits and challenges are set out in Section 5 on further considerations.

3.7 DEVELOPING A STRONG APPROACH TO ETHICS

Figure 9: Frontier Data Study: selected expert views on ethics from the global stakeholder survey

Some quotes from the global stakeholder survey results are given in here and show the strong emphasis given to ethical considerations by global stakeholders.

"Technology is often pushed without testing its limits or its effect on the population. Big data has good potential, however it can backfire, if not applied through ethical standards and data quality principles."

"Issues around data leaks and breaches of privacy are paramount. Exposing vulnerable populations can be a risk, for example providing locations on those fleeing persecution."

"On a larger scale, issues around data ownership, power, and profit risks further eroding public trust in institutions if new models giving citizens a voice are not found."

"Anonymised data can be less anonymous that we think. Ownership rights can be given away too casually."

"We are talking about fast and large data which means the ethical issues of traditional data can come at us fast and big too, so we need to be better prepared."

Developing good approaches to ethics is one of the essential enabling factors in using new digital data sources effectively.

The Study found that while new data sources bring new and often enhanced traditional ethical risks, these can be managed well, and risk aversion should not get in the way of finding the right ways to use new data sources.

Notwithstanding an already ethical imperative to investigate how new data sources can also help improve lives, the digitisation of any data can also help to tackle ethical issues.

Many of the ethical benefits of digital data are expressed through the **Open Data** movement, strongly supported by DFID. This seeks to make traditional and new data freely available, which is largely facilitated by the digital nature of new and some traditional data sources. It is also fuelled by the need to make vast amounts of data open for innovation and combination, so opportunities are not lost. Evidence shows that, combined with adequate data protection mechanisms, open data can have transformative positive effects that were not previously possible.

For example, in Namibia, the government turned to the country's largest telecommunications provider to identify where citizens were at high risk of contracting malaria, using open mobile data combined with satellite data, enabling the Ministry of Health to target and distribute 1.2 million bed nets to the most vulnerable communities[10].

The GovLab have identified six features that indicate why there is a compelling case to invest in opening up digital data sources to support ethical gains:

Participation - by facilitating citizen participation and mobilization, open data can allow a wider range of expertise and knowledge to address and potentially solve complex problems.

Trust - because open data increases transparency and avenues for citizen oversight, unlocking data can lead to higher levels of trust throughout societies and countries.

Equality - open data can lead to more equitable and democratic distribution of information and knowledge — though, several observers have also pointed out that just releasing open data can play a role in further entrenching power asymmetries related to access to technology and data literacy.

Scrutiny - because open data is subject to greater scrutiny and exposure than inaccessible institutional data, there is potential for enhanced review and improvement in the quality of government data by actors outside government.

Value Amplifier - opening government datasets in a flexible and equal manner can amplify the value of data by filling — and identifying — important data gaps in society.

Flexibility - when released in an interoperable, machine-readable manner, open data is easier to repurpose and combine with other pieces of information, which in turn means that it is more flexible, with secondary data potentially yielding innovative insights.

However, this Study also found a consensus among stakeholders and literature that there is an enhanced level of ethical risks with new data sources. There is a need to ramp up the safeguards used with traditional data and to find new ways to deal with new risks.

This is not least because, by its nature, much of the data will be highly personal and harvested without the knowledge and/or consent of individuals, particularly in terms of its intended or unintended end use.

Data systems which rely on AI appear to be particularly risky in terms of ethics.

Notwithstanding future developments in legal and other frameworks to mitigate these risks, enhanced risks stem from:

- International development typically involves multiple layers of accountability and responsibility. This can result in many potential layers of influence and data handling, such as with the funder, a foreign implementing organisation, a local implementing organisation, the implementing country government, and a software vendor. With AI this is compounded by the fact that data systems are typically built and managed by external partners. This trend is likely to continue because AI talent is in short supply.
- There can be an inherent lack of transparency in a machine learning system and it may be unclear about when, how, and about whom automated decisions are being made. People who use the system often did not design it and their understanding of how it works is likely to be limited.
- AI inherently has more risk because legal frameworks have not caught up with the implications and because of the complexity of different legal issues associated with data sets and organisations often working across countries.
- A perceived objectivity of these systems and overconfidence in the results they produce. This often eliminates direct responsibility as human operators mentally distance themselves from the outputs the system produces. A point that is applicable to a range of new data sources not just from AI.

In general, across all new data sources the Study found that the following factors appear to be the most important for consideration:

- **Accountability** – transparency is required about who or what is making decisions based on the data at any point
- **Data ownership and security** – establish legal rights as to who can do what with the data and establish data management systems which are subject to regular data Risk Assessments (see Figure 10 right) to ensure they are highly secure from hacking or leakage over long periods of time. Personal data combined across different sources can be highly dangerous
- **Inclusive design and decision-making** – vulnerable and marginalised groups should be involved
- **Addressing bias** – need a thorough level of challenge around possible biases in the data (including the data used to 'train' the program) and methodology, particularly if an algorithm is involved, and
- **Human involvement** – it is advisable to involve humans in vetting or making decisions at key points in a machine learning cycle.

There is however no one-size-fits-all solution for managing enhanced ethical risks with new data sources. These risks will vary depending on the type of data source or combination of data sources which are used, and most importantly on the specific context of the development intervention in question. Section 3.5 above provides guidance on the basic issues for each of the high potential data sources and includes links to further guidance.

Figure 10: Recommended reading on ethics in big data

Principles for Digital Development Ethics and practical experiences are built into these 9 useful Principles which are constantly updated and have been endorsed by DFID – good food-for-thought in thinking about the right approach to ethics in any specific case.

Data Privacy, Ethics and Protection – Guidance Note on Big Data for Agenda 2030, UN Development Group

Open data in developing economies – The GovLab sets out a range benefits and examples of where ethical risks with open data have been worth taking and what are the remaining challenges.

DFID ethics guidelines for Research, evaluation and monitoring

A highly important part of a strategy to address ethical issues is the need for data security. Figure 11 below sets out advice about what is needed.

Figure 11: Recommended approaches to data security

Fortunately, to a large extent data is data and the principles for its good management are universal. The Principles of Good Data Management (2005) provide a good lay person's guide to the detail of what needs to be done for any type of data.

A key component in addressing ethical and other risks, is carrying out regular **Data Risk Assessments** of the environment that the data resides in and transits across. Based on the outcome of the risk assessment, proportionate measures can be adopted. This may be particularly important when using new data sources to ensure that existing data management systems, often set up for traditional data sources, are able to adapt to the detailed risks of your new data source(s) in a specific context.

Data users and innovators, in partnership with 'data custodians' such as IT specialists, would ideally ensure that data systems comply with ISO standards, such as by establishing:

- An **Information Security Management System** (ISMS) – consider using a framework such as ISO 27001 or NIST CSF (800-53)
- An **Information and/or Data Governance Framework** – consider using ISO 38505

3.8 HOW CAN NEW DATA SOURCES HELP WITH THE 'LEAVING NO-ONE BEHIND' AGENDA?

Data2X has provided a list of inspiring examples around gender. Often solutions in one area can release benefits for addressing gender or other policy areas.

For instance, using the radio content analysis tool it developed (as per Figure 4 above) for peace and security issues, Pulse Lab Kampala created a real-time gender perceptions dashboard to unearth discussions and topics regarding sexual and gender-based violence across Uganda.

With the right approach that embeds ethical and data quality frameworks, the use of new digital data sources can improve inclusivity, as:

- the necessary management of ethical risks should inspire more inclusion of vulnerable groups in data collection methods and decision-making, and
- the potentially high granularity and accuracy of some data could allow for the better identification of marginalised groups and targeting interventions more efficiently.

The availability of new digital data sources gives an unprecedented opportunity to gain new insights and information to support decision-making by and about different population groups in support of the Agenda 2030 objective of 'Leaving No-One behind'.

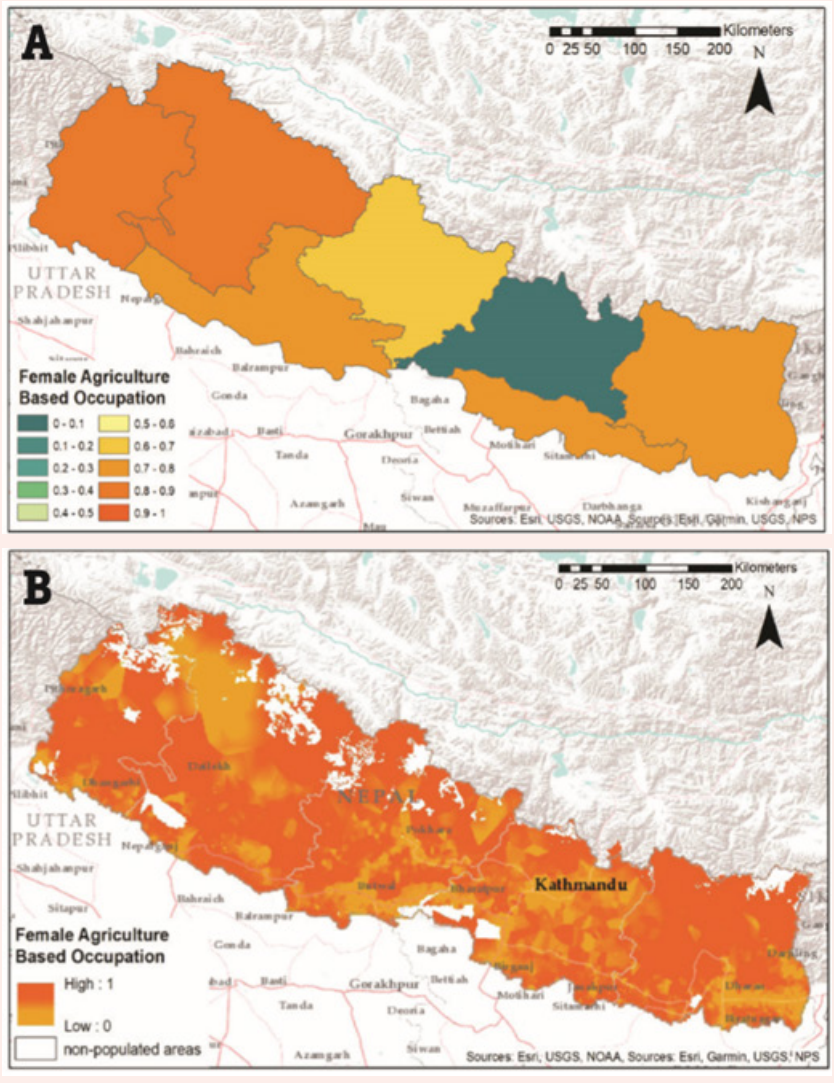
There are currently enormous data gaps in traditional data sources around marginalised groups, especially in terms of monitoring the SDGs. But with this opportunity comes ethical risks. These need to be managed well, particularly where population sub-groups are explicitly identified in or excluded from a dataset.

Figure 12 - Combining data sources in Nepal to better understand gender issues

Flowminder and Data2X have highlighted the potential of combining geo-located survey, satellite imagery, and mobile phone data for creating well-being monitoring systems with high resolution in both space and time in Nepal. Data was analysed according to three key gendered indicators — literacy, agriculture-based occupations, and births in health facilities — at very high spatial resolution. Then they used de-identified mobile phone data to produce frequently updatable information on gendered mobility and migration patterns.

A key challenge was the need to predict gender patterns among a population of mobile phone subscribers. SIM sharing is an important complicating factor in predicting gender and inferring individual well-being. More work is needed on understanding the characteristics of mobile phone ownership patterns.

The maps below show how understanding of female participation in agriculture-based occupations was improved by the Flowminder approach (map B), allowing for much higher resolution than before (map A)



SECTION 4:

WHAT IS NEEDED TO RELEASE THE POTENTIAL?

“What we found is that you can get carried away with new methodologies and producing really snazzy platforms and dashboards that look amazing, but they are not actually maintained and used.

Data science is not something that can be done very easily, or quickly. I think that’s a misconception about data science in general.

Context matters.”

Derval Usher,
former Head UN Pulse Lab,
Jakarta

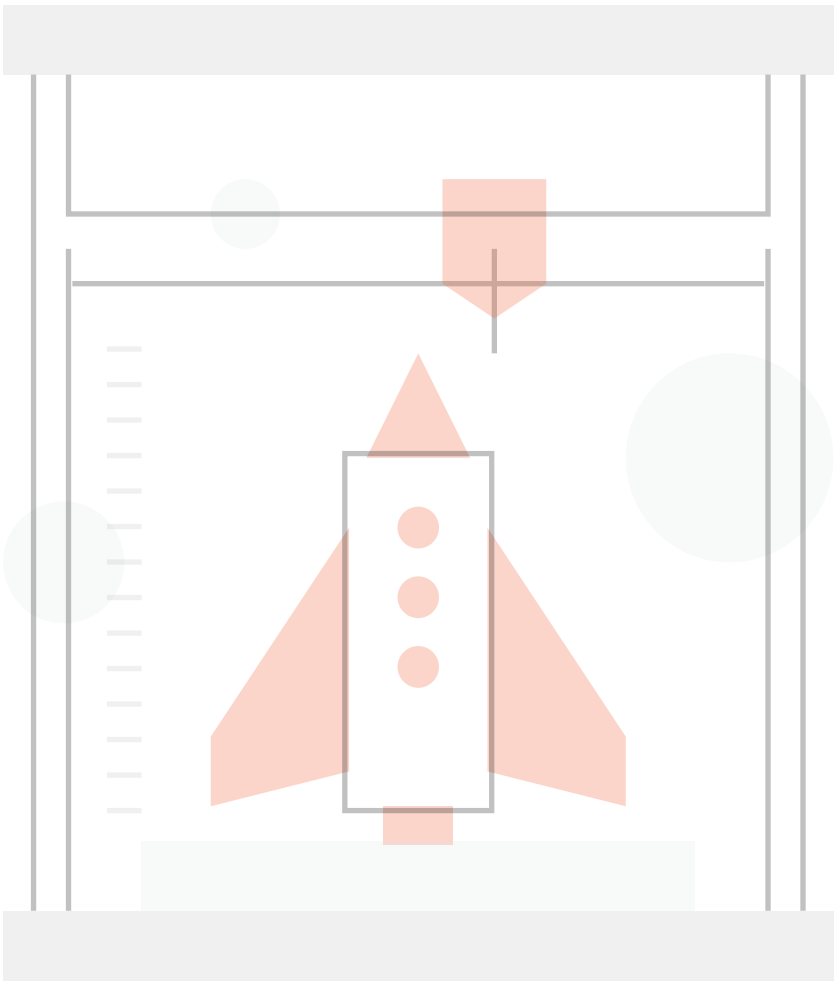
Interview with the Centre for Humanitarian Data, June 2019

4.1 OVERVIEW

There is already lots of research available to help understand the exciting things that can be done with new digital data sources. Section 3 set out the technical and ethical benefits and challenges.

The wider corporate context is however crucial. A major focus of this study was in understanding DFID staff's data needs. This helped in both understanding the potential of new data sources and the help staff would need to innovate with them.

This Section sets out the most important findings in terms of empowering staff to release the full potential of new data sources. Section 5 then looks at some potential further corporate considerations that were uncovered during the Study but were not part of its scope.



4.2 PUTTING STAFF IN THE DRIVING SEAT

The Study gathered evidence from focus groups with a cross-section of DFID staff in the UK and Country Offices, representing a cross-section of types of data decision-makers, from data users, to those driving data innovation and management in DFID, to those designing data strategies for Portfolios and Programmes.

There was some significant enthusiasm around the potential of new data sources to assist in a wide range of outstanding data needs.

Understanding staff's needs for data

Notwithstanding that the focus groups were not fully representative of all sectors and geographies, there was widespread enthusiasm around the potential of new data sources to assist in a wide range of outstanding data needs. Some of the most recognised opportunities for new data sources were in:

- challenging existing assumptions and sense-checking the interpretation of existing data
- filling current gaps in knowledge around the forming of social norms and possibly helping to shape them
- the agricultural sector
- Climate Change sector
- disaggregation at the sub-national and sub-population group level, and
- using new data sources as the only available/ most reliable data in fragile and conflict states.

However, there was also common feeling that existing data innovations were not necessarily being driven by staff needs, which were by and large highly specific to their area of work. This is in line with other evidence that indicates that to optimise the potential of new data sources, decision-making around their use should be highly devolved by empowering individual staff. This is not least because the nature of new data sources means that decisions around their use need to be highly context specific.

While new data sources have been much celebrated for their ability to provide 'real-time' data, there were mixed views as to whether that was needed at all in most DFID contexts.

In many cases accuracy and relevance were more valuable than speed, as there are few occasions where staff use data even as often as a weekly basis, except in disaster mitigation scenarios.

It was underlined by staff that, while the digital era means interesting opportunities through digital data, there is also no shortage of better-quality traditional data that could be accessed via other donors, governments, and the private sector etc. Investment in getting access to existing sources could often be more valuable than innovating around new sources.

Understanding staff's needs for support

Decision-makers on the ground wanted to be empowered to work with the right experts to unpick the pros and cons of any data choices or combinations of them. There was little appetite for developing detailed technical skills or knowledge.

But there was a notable appetite for systems and capabilities being developed in DFID, and between DFID and external organisations, which meant that they could work with others to efficiently take advantage of the opportunities in the new data landscape.

The following were specific requests:

- help in identifying and interacting with internal technical experts within DFID over the full range of technical and ethical issues
- simple checklists to help decision-making
- help in navigating heavy/numerous documentation
- plain non-technical language in any guidance
- help in understanding technological requirements and investment costs, and
- understanding and getting access to the data and knowledge of relevant innovations within DFID and other donors.

Without being able to carry out an assessment of the strengths and weaknesses of current corporate systems within DFID, based on discussions with staff and wider global evidence about the inherent needs of data users, the Study concludes that staff need to be supported in understanding and using the opportunities around them.

This needs to draw on the specific mix of expertise required to innovate with specific data sources in a way that provides:

- tailored, flexible, and responsive support which is tailored to individual needs and contexts
- clear channels for providing requests for technical support and/or research
- strong central safeguards/advice on ethics and data quality while minimising bureaucracy
- basic guidance and training on data/statistical literacy as relevant across the full data landscape - it is unrealistic to expect all DFID staff to gain the wide range of highly technical knowledge required to successfully innovate
- data and information management systems that allow for the ready sharing of innovations and data across cadres, Programmes, and Country Offices
- support to keep up to date and communicate about what is relevant from the enormous amount of and multitude of types of global research and learning around new data sources
- help in accessing data from external organisations, and
- access to easily absorbed guidance on the basics of a range of technical data issues.

Section 4.3 and the user guides provide examples of the type of overall guidance that could be developed.

4.3 DEVELOPING GUIDANCE FOR STAFF

In terms of the starting point for the development of appropriate guidance, the focus groups with DFID staff revealed perceptions that there is not enough help available or that too much of what is available is too generic or given in the wrong way. For example, there was a perception that a lot of the advice was available via voluminous and numerous technical written documents.

On the other hand, a cautiousness among DFID staff was consistent in respect of new data sources in terms of them:

- carrying significant and unknown ethical risks
- risking the exclusion/bias against key population groups
- needing a prohibitive amount of technical support or IT investment, and
- not being compatible with wider DFID HR and IT systems.

To help address some of this, the Study has synthesised and summarised available guidance across a number of relevant disciplines within the body of this report and in a generic checklist below.

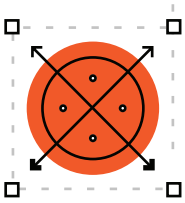
The Study has also produced a set of **user guides** (see Section 4.4) relevant to each of the high potential new data sources that are already ready for operationalisation.

The checklist on the following pages provides an **overall framework for staff** who are considering innovating with any new data sources to fill a data gap or to provide insights needed to improve decision-making.

The supplementary user guides (Section 4.4) should help in un-picking some of the core issues with each of the high potential data sources.

These guides and checklist can be used by staff now, but they could be developed further based on feedback from staff and reflecting ongoing developments with internal systems, and the data sources they related to.

CHECKLIST FOR NAVINGATING THE NEW DATA LANDSCAPE



☑ Define your data needs

You should clearly articulate **what it is your need and at what frequency you need it**. Then identify any gaps in the coverage or quality of data that may already be available.

Also consider how your need might connect to the **wider data needs of your partner country** and how any innovation can help support long term improvements in the overall data ecosystem.

☑ Establish support from colleagues and develop an implementation plan

Consider if you can carry out the work on your own or what support you might need internally. Bear in mind that if you are going to work with external partners, you may need to have effective oversight of highly technical issues.

- Statistics advisers
- Data Science Campus
- EPIC
- Data for Development team (D4D)
- Heads of Professions
- Policy teams

You should consider a range of options from traditional and new sources or combinations thereof. This includes scoping out what relevant data is available in specific countries, regionally, or globally.

For working with external partners, as a first step, you should identify relevant external experts who can advise what data is or could be made available – liaising with national bodies such as National Statistical Offices (NSOs) would be good practice, particularly as this supports identification of long-term benefits.

Working with your internal advisors and/or external experts, guide the design of a data collection and analysis plan that fits in with the Core Questions for Data Innovation of the Frontier Data Study (see above) and the key considerations from the User Guides on specific data sources.

The design of your new data system may require some initial research that tests out the specific feasibility of methodology and ethical issues etc in the context you are working in.

Develop an Implementation Plan that reflects the outcomes of this research and the items in this checklist.

The UN Pulse Lab provides an easy to follow detailed template for moving from an idea to proof-of-concept.

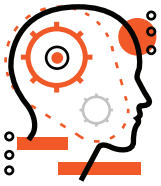
Analyse your data options against the Data Quality Dimensions

Using the **Data Quality Dimensions** set out in Section 3.4 of the Frontier Data Study, analyse and consider the **strengths and weaknesses of your various data options**. Log any specific weaknesses if you decide to proceed and include this in an Implementation Plan.

Of particular importance for all new data sources will be analysing any costs of the data, now or in the future, and whether or not the data will continue to be available when you need it.

The **DFID Smart Guide on Data Quality** will also help if your data need is around Monitoring and Evaluation.

The Digital Futures Hub has also developed a short introductory snapshot guide to data quality in general, as a key tool for any decision-making about using data in international development implementation.



Test your thinking against core DFID guidelines and strategies

Validate your thinking against:

- The Frontier Data Study
- Cadre/Portfolio strategies
- DFID Ethics Guidelines
- DFID Digital Strategy

Establish a robust strategy to deal with ethical risks and data security

Within your Implementation Plan, develop your initial ethical considerations in a **dedicated section on the core ethical risks to be managed**, remembering that it is highly beneficial to involve beneficiaries and potential marginalised groups in your thinking, and to carry out a **Data Risk Assessment** involving data security experts (see Section 3.7 of the Frontier Data Study).

The main considerations for ethics and data security are set out in the Frontier Data Study and its user guides.

Review and evaluate

As well as renewing your Data Risk Assessment on a regular basis, it is recommended to review the effectiveness of your plans after one complete cycle of data production and analysis

Below is a list of **core questions** that staff could seek to answer when commissioning and/or driving technical support providers to help them innovate in the new data landscape.

1. **What are the specific data problems that can be solved?**
2. **How can solutions be provided that fit in with relevant ethical considerations?**
3. **What are the technological and institutional requirements? (such as IT software, inter-institutional agreements for data access)**
4. **How can populations affected be involved in optimising positive outcomes?**
5. **What is the carbon footprint, compared to other forms of data?**
6. **How can the innovation ensure that populations who are socially or digitally excluded are not under-represented in the results?**
7. **How can the innovation provide sustainable improvements in local data ecosystems and decision-making?**
8. **What are the strengths and weaknesses of the data in terms of the Data Quality Dimensions (see section 3.4), particularly compared to traditional data sources? How can weaknesses be improved?**

The UN has developed the model below to inform the design of any data process. Key elements include starting with understanding the data user requirements (Specify Needs) and then the reviewing (Evaluate) how well the process has fulfilled those needs, which may of course change over time.

It is important throughout the process to understand data quality issues (including ethics), as per the Data Quality Dimensions set out in this Study, and recording 'metadata' (information about the process which helps to manage data quality issues).

Figure 13: The Generic Statistical Business Processing Model (UN Economic Commission for Europe)

The online Frontier Data Study User guides set out for the high potential new data sources provide more detail around:

- What is it?
- What can it be used for?
- How can it help with the Leaving No-One Behind Agenda?
- What are the main ethical considerations?
- Checklist of the most important things to consider
- Case studies and examples

The Study, in collaboration with The Development Café, also developed an interactive map of innovations with new data sources. It is open source and contributions are invited through this link.

The map includes references from another useful source of inspiration and learning about new data sources, but a little more connected to macro-scale policy-making: the UN Big Data Project Inventory.

SECTION 5:

FURTHER CONSIDERATIONS

5.1 OTHER INTERNAL CORPORATE DEVELOPMENTS

The Study has indicated that, to optimise the efficient use of new data sources, a number of key things are needed on a corporate level. These include supporting or continue in supporting:

- research and investigation into improved methodology and ways to decrease risks
- improvements in traditional data sources
- improvements in the availability of existing new and traditional data
- the development of ways to combine new and traditional data effectively, including robust data security, and
- the development of staff guidance including keeping these up to date and relevant.

The evidence gathered in the Study also found indications of other important areas of focus for any ambitious organisation in terms of optimising the use of new data sources, but which were outside the scope of detailed investigation in this Study:

- promoting an enabling culture and **Change Management regimes that allow for flexible responses to new data opportunities**
- developing **appropriate IT and other systems** (not least for strong internal data sharing and security), and
- establishing **strong inter-disciplinary connections** for sharing knowledge and data, and collaborating in developing data solutions, and
- developing **appropriate skill sets** within both the general competences of staff and those which might need to be centralised – i.e. those which can be made available to support devolved decision-making and data use as required in specific circumstances

In terms of getting right balance of skills needed either at the data user level or in terms of the support staff, the Study did not collect direct evidence about mapping the new landscape and staff skill sets. But it is seems worthwhile to review lessons from models developed by the private sector. Notwithstanding different objectives and priorities, the private sector have moved relatively quickly in mainstreaming new data opportunities.

On the other hand, there is very clear evidence that a good understanding of the principles at play in official statistics is highly beneficial across data users and support staff. There is a distinct risk globally that 'data science' becomes detached from this methodological safety net that is full of learning and insights for guiding safe and effective ways forward.

5.2 SUPPORTING EXTERNAL CHANGE

The Study collected evidence that indicates acting externally in several ways would be beneficial in helping to release the potential of new data sources.

It should remain important for FCDO to cooperate closely across government and with other relevant bodies in the UK and globally in sharing learning about new data sources, and, importantly, developing agreed standards in methodologies, data quality, and ethical controls.

There would be benefits from working with other institutions to achieve broader changes. The following issues have been consistently identified by global stakeholders as a high priority for inter-institutional cooperation:

- **development of semi-formal partnerships between different data owners and users** to support Open Data and open source solutions, supporting sustainable mutually beneficial data sharing arrangement
- **cooperation with the institutions that provide national or local level statistical infrastructure in beneficiary countries** - to both enhance approaches to data quality and ethics, and to leave a legacy that will support the ongoing development of their statistical systems
- **development of international and national legislation** - in a way that both allows for data to be shared and for it to be protected, and
- **cooperative frameworks for the provision of demand-led ideas from data users** - to efficiently meet the seeming high capacity of data innovators on the supply-side, pushing their research and innovations in a direction that is more likely to be adopted and useful across development partners' activities.

However, for many years, while the sharing of lessons and knowledge about new data sources has been commonplace internationally, some stakeholders pointed out an apparent blockage in converting this into practice or making major breakthroughs with the above challenges. This Study did not gather enough evidence to conclude about why this is so. But the issue of **a lack of transparency in institutional incentives in the public sector** was often identified by stakeholders as a barrier to progress.

5.3 DATA SOURCES WITH LONG TERM POTENTIAL

There are some emerging technologies on the horizon that the Study identified that could usefully be observed in terms of their potential to provide data in the longer term. Currently they seem to have lower potential to transform the data landscape.

Not least because there is little evidence about what specific data needs they could provide solutions for, while their use in developing country contexts is likely to be minimal for some time to come.

The table below sets out the key findings in respect of those technologies. It is recommended that these are tracked for their emerging potential and the challenges that would need to be addressed.

Emerging Technology	Things to look out for
Wearable gadgets / tech	Could be particularly useful for interventions/preventative treatment. But in terms of data outputs there are high risks of false positives and false negatives, and many potential new ethical and data security challenges
Quantam computing	Still in an early development stage with lots to be learnt about the potential benefits and the risks
3D-Printing	Possibly some good potential data sources from its uses, such as by using the data it generates to support small-medium business in stimulating economic growth and in the mapping of natural resources.
Nanotechnology	Some potential data gains in terms of real-time and strategic monitoring where data about its use is available. For example, in data from equipment for cooking or medical sterilisation in remote areas.
Augmented / Virtual Reality	Many potential uses are likely to occur in certain sectors, for example in health in helping hearing/visually challenged people navigate the world. The resulting data may become helpful in assessing the design of implantation methods and tracking development interventions in similar areas.

Table 7 - Forthcoming technological evolutions with longer-term potential to benefit international development

The Study considered **Blockchain** as an emerging technology that has been around and tested for some years now – hence not included in the table above. It was found that it is being highlighted by some bodies as important for data sciences, particularly in predictive analysis, real time data analysis and secure data sharing.

The promise of blockchain is in a whole new way of managing and operating with data — no longer in a central perspective where all data should be brought together, but a decentralised manner where data may be analysed right off the edges of individual devices.

The potential benefits of Blockchain come from:

- it integrates with other advanced technologies, like cloud solutions, Artificial intelligence AI) and the Internet of Things (IoT)
- validated data generated via blockchain technology comes structured and complete, and it is unchangeable, and
- blockchain generated data may improve data integrity since it ascertains the origin of data though its linked chains.

Blockchain real world applications, however, are still in nascent stages - though it may not appear so due to the notoriety the technology has gained in a short period. This Study did not find enough valid evidence to be confident of its promotion as a viable data tool for DFID in the near future. None of the global stakeholders surveyed indicated it to have any significant potential at present.

One would expect that as the blockchain technology matures and there are more innovations around it, more concrete use cases will be identified and explored. That being said, a few serious challenges have been raised about its impact in data science where exceptionally large amounts of data need to be handled.

One concern is that blockchain applications will therefore be very expensive to pursue. This is because data storage on a blockchain is expensive compared to traditional means and 'Blocks' deal with relatively small amounts of data compared to the potentially large volumes of data collected frequently via many digital platforms.

How blockchain evolves to address this concerns and proceeds to disrupt the data science space will be particularly interesting to observe.

5.4 DEALING WITH A VAST AND EVOLVING DATA LANDSCAPE

The recent rapid growth in data availability provides new insights on the complexity of the challenges in international development. Indeed, the growth in data availability seems to be in some way a product of, or at least correlates to, an increasing complexity of the human experience.

Some of the most pressing challenges in current times are inherently complex and are possibly becoming increasingly complex, such as epidemics and climate change. **This implies that theories of what is happening in a specific context will need to be adapted/ adaptable to new, often rapidly developing, knowledge that new data can bring.**

Including addressing some known general weaknesses in current approaches to project design and Theories of Change, the development of more dynamic rather than static conceptual models would seem to be more appropriate to deal with the complex feedback loops that can now be observed[11].

This includes addressing changes in stakeholder perceptions and needs as they react to the data landscape. This Study has shown that these models should also be able to recognise and integrate the strengths and weaknesses of various data types of data sources in a specific context.

More broadly, the Study, within the limitations of its resources, reached out to all types of relevant stakeholders across the globe, was as rigorous as possible in its selection of research and case studies, and interacted with a wide cross-section of DFID staff. But there may still be other developments not included in the evidence which could be important in influencing how the Study's conclusions are interpreted.

This is a rapidly evolving field of knowledge so it may be beneficial to make this Study widely available to global stakeholders and **maintaining an ongoing mechanism to gather and process relevant evidence**. Similar research to this Study might be beneficial at least annually.

CONCLUSIONS

The Study made **8 main conclusions** based on:

- evidence from a wide variety of stakeholders and research
- statistical principles
- emerging innovative disciplines in data science, and
- emerging priorities and practicalities in international development

1

There is **justified excitement and proven benefits in the use of new digital data sources**, particularly where timeliness of data is important or there are persistent gaps in traditional data sources, such as in fragile and conflict states, or in supporting decision-making about marginalised population groups, or in addressing persistent ethical issues where traditional sources have not proved adequate.

2

In many cases, improvements in and greater access to traditional data sources could be more effective than just new data alone, including developments in tandem with new data sources.

This includes innovations in digitising traditional data sources, supporting the sharing of data between and within organisations, and integrating the use of new data sources with traditional data.

3

Decision-making around the use of new data sources should be highly devolved by empowering individual staff and be focused on multiple dimensions of data quality.

This is not least because there are no "one size fits all" rules that determine how new digital data sources fit to specific needs, subject matters or geographies. This could be supported by ensuring:

- Research, innovation, and technical support are highly demand-led, driven by specific data user needs in specific contexts
- Staff have accessible guidance that demystifies the complexities of new data sources, clarifies the benefits and risks that need to be managed, and allows them to be 'data brokers' confident in navigating the new data landscape, innovating in it, and coordinating the technical expertise of others. The Study itself aimed to provide a contribution to this.

4

Where traditional data sources are failing to provide the detailed data needed, **most new data sources provide a potential route to helping with the Agenda 2030 goal to 'leave no-one behind,'** as often they can provide additional granularity on population sub-groups.

But, to avoid harming the interests of marginalised groups, strong ethical frameworks are needed, affected people should be involved in decision-making about how data is processed and used, and assurances are needed that the data is representative of those marginalised groups.

Action is also required to ensure strong data protection environments according to each type of new data and the contexts of its use.

5

New data sources with the highest potential added value for exploitation now, especially when combined with each other or traditional data sources, were found to be:

- data from **Earth Observation (EO) platforms (including satellites and drones)**
- **passive location data from mobile phones**

While there are specific limitations and risks in different circumstances, each of these data sources provide for significant gains in certain dimensions of data quality compared to some traditional sources and other new data sources.

6

The use of **Articial Intelligence (AI)** techniques, such as through machine learning, has high potential to add value to digital datasets in terms of improving aspects of data quality from many different sources, such as social media data, and particularly with large complex datasets and across multiple sources.

7

Beyond the current time horizon, the most potential for emerging data sources is likely to come from the **next generation** of existing technologies and techniques in:

- **Artificial Intelligence**
- **Earth Observation platforms**
- **Privacy Preserving Data Sharing (PPDS) via the Cloud**
- **the Internet of Things (IoT).**

No significant other data sources, technologies or techniques were found with high potential to benefit FCDO's work, which seems to be in line with its current research agenda and innovative activities.

Some longer-term data prospects have been identified and these could be monitored to observe increases in their potential in the future.

8

Several other factors are relevant to the optimal use of digital data sources which should be investigated and/or work in these areas maintained.

These include important internal and external corporate developments, importantly including:

- continued support to Open Data/ data sharing and enhanced data security systems to underpin it
- learning across disciplinary boundaries with official statistics principles at the core and
- continued support to capacity-building of national statistical systems in developing countries in traditional data and data innovation.

APPENDIXES

APPENDIX 1: DATA OPPORTUNITIES POTENTIALLY USEFUL NOW IN TESTING ENVIRONMENTS

The tables below sets out the possible applications and challenges for the data opportunities set out in Section 3.6.

Internet of Things (IoT)

Examples of benefits within the data value chain	Examples of data quality, ethical, and other research challenges
IoT Sensor & Device Data was a previously expensive solution but should become relatively cheap, allowing for near-real-time data collection across the value chain or macro-based environment that can address current gaps in traditional data sources, in areas such as: • Monitoring environmental changes (e.g. linking in across IoT technologies to create 'Smart Cities'; water monitoring: quality, water table, flow and use; pollution mapping and warning systems) • Monitoring people's attitudes (eg basic digitally recorded satisfaction ratings) and monitoring people's behaviour (eg responses to tasks) • Early Warning Systems in areas with scarce infrastructure • Connected marketplaces for fair trading. • Wildlife tracking and anti-poaching efforts • Electricity monitoring • In-situ crop monitoring • A range of applications of the increased interconnectivity of new energy related systems such as electric vehicles, storage devices, or small-scale renewable energy systems at household level	**Sensors need verifying/calibrating** otherwise data will be very poor **Security vulnerabilities** in data systems due to: • companies developing affordable data systems • their very nature makes them susceptible to cyber-attacks as people have physical access to devices In many cases, **needs to be combined with geospatial information** as location provides a vital link between the sensors that will generate the IoT and the Uniform Resource Identifier (URI) assigned to a thing or object within that connected world **Strong reliable internet connections are needed** **The architecture of the internet needs to be developed from human orientated towards machine learning.** If not the IoT will need to take into account devices which are to all intents and purposes autonomous and act independently whether or not any person, or any system, is actively using them.

Privacy Preserving Data Sharing (PPDS) via the Cloud

Examples of benefits within the data value chain	Examples of data quality, ethical, and other research challenges
Enhanced capabilities to share data via large cloud-based data portals, particularly without risking disclosure or commercial disadvantage between two data owners could facilitate access to data held across the public and private sectors. This could mean: • New knowledge through combined data • Cross country data sharing: eg water quality, drought and pest issues • Access to private sector data held by mobile phone companies, insurance companies etc • Wider access to traditional data sources, such as official statistics or surveys carried out by other donors/ NGOs: which would plug knowledge gaps and minimise the need to take risks and costs in collecting new data of poorer quality	**Data collaboratives** – institutional arrangements to facilitate data sharing are still in their infancy and significant advances are still required before Cloud technology can be used effectively, not least in legal frameworks and 'soft' issues such as trust between and within public and private sectors and between international development organisations. **Establishing Data Trusts** to oversee data privacy issues may be part of the solution. Some of the potential to leverage a willingness to share data is via the partial hiding of details in the datasets; without access to such dis-aggregated data there will be lot less value for the Leaving No-One Behind Agenda, although high level summary data for sub-populations may be possible. Most Cloud providers now provide i**n-country data warehouses.** This is mostly for European countries reluctant to store data in the US. There is a challenge applying this in developing continents, and between donors and across relevant private sector datasets. Due to the **costs associated with creating a public Cloud service**, it is possible that not all countries will have access to them. Therefore, there's a risk that the technological gap will grow.

Next generation Artificial Intelligence (AI)

Examples of benefits within the data value chain	Examples of data quality, ethical, and other research challenges
The next generation of AI could lead to new insights about the ways in which objects or their properties are related, with applications in health, crime, agriculture, environment and so on. **Combinations with satellite technology are promising.** Processes based on the learning of geospatial concepts (locational accuracy, precision, proximity etc.), can be expected to improve the interpretation of aerial and satellite imagery, by improving the accuracy with which geospatial features can be identified. **Deep Learning (next generation of machine learning)** might be applied to improvements in the use of AI in areas such as Financial Market Segmentation, Smart Insurance and loans, Early Warning Systems for health, decision Support and recommendation systems. E.g. around what crops to grow; much better Chatbots for depression detection and disease diagnosis or real time speech analytics such as real time translation and response could be useful in areas where human help isn't readily available, but a need is urgent.	**Future challenges are likely to be those which currently observed with AI,** especially in terms of bias and the lack of human perspective – eg psychological/ personality profiling is likely to remain an ethical and data quality risk; and language detection/ analysis of its use is a problem especially where programs fail to account for linguistic and cultural diversity. Privacy issues are likely to increase further as the ability (or the perception of being able) to accurately identify people and information about them increases. These issues can hardly be said to be well addressed in the current era, so extreme caution is required.

Next generation Earth Observation (EO)	
Examples of benefits within the data value chain	**Examples of data quality, ethical, and other research challenges**
The applications for improved accuracy, granularity, and layering of data are manifold, including in disaster monitoring, animal and pest tracking, migration, and environmental monitoring.	**Future challenges are likely to be those which are currently observed with EO technology,** especially in terms of ethical/ security risks around the combination of geographical positioning data with data on resources or individuals – particularly in conflict zones. There will be an enhanced risk of abuses with the commercialisation of data and where technologies are only available to government/military agencies. A fundamental tension needs to be managed to ensure EO data is shared effectively for international development and the need to ensure the data is secure and not used to harm against population groups.

APPENDIX 2: BIBLIOGRAPHY AND FURTHER READING

Frontier Data Study recommended sources for an overview of the benefits, risks, key issues to manage, and case studies

The Frontier Data Study reviewed an extensive array of relevant literature. The global attention to better understanding the use of the emerging data landscape in international development has generated an amount of research that is difficult for any individual to digest and analyse in full.

This part of the bibliography highlights those sources that have either had a significant influence on the Study's conclusions and/or are, in the view of the Study team, sources which provide particularly useful insights for development practitioners/strategists or statisticians (traditional or those using new 'data science' techniques).

It is fully recognised that this list is not comprehensive in terms of relevant or useful literature. Further detailed references are given in the next section. Other references and sources are linked within the body of the report and the **Frontier Data Study online user guides.**

Author	Title / Link	Why should i read it?
United Nations	Big Data for Sustainable Development website	The United Nations in terms of both its official statisticians and data scientists (at the UN Pulse Lab) are at the forefront of relevant research and plain English Advice. The website gives a very concise summary of the benefits and risks of big data, and links to case studies and further reading
United Nations Global Pulse	Big Data for Development: Challenges & Opportunities Big Data for development: A Primer Integrating big data into the M&E of Development Programmes A guide to data innovation from idea to proof-of-concept The state of mobile data for social good report Mobile Phone Network Data for Development – Primer	This range of publications from The UN Pulse Labs in Kampala, Jakarta, and New York provides well-written introductions to the topic while giving handy detail for the layperson, inspiring examples, and lessons learned.
UN Women	Gender equality and big data: making gender data visible	Presents the benefits, risks, and policy implications of new data sources for gender issues
Data 2X	Big Data and Gender briefs	Learning from a wide range of application of new data and techniques to gender issues
World Economic Forum	Big Data, Big Impact: New Possibilities for International Development	Learning from a wide range of application of new data and techniques to gender issues
McKinsey Digital	Big data: The next frontier for innovation, competition, and productivity	Short summary to help understand how the private sector is approaching the emerging data landscape
IDRC / LIRNIE Asia	Mapping big data sources for the SDGs	Maps ideas for new data sources to each SDG, giving short practical examples, including how one could work with the private sector
ICT Works	12 Artificial Intelligence Initiatives in Health, Education, Human Rights	Quick inspiring list of examples of how AI can be used

Frontier Data Study recommended sources for understanding future possibilities in the emerging data landscape

General Overivew
UN-ESCAP Frontier Technologies Report
Gartner Hype Cycle Report on Emerging Technologies
Institute of Development Studies (IDS) Frontier Technologies Report
Oxfam: Global Megatrends – Mapping the forces that affect us all
Specific technologies
Future Trends in Geospatial Information Management: the five to ten-year vision (UN Committee of Experts on Global Geospatial Information Management)
Artificial Intelligence and Life in 2030 (Stanford University)
Microsoft Quantum Development Kit
The Future Can't Wait – Over the Horizon Views on Development (U.S. Department of State U.S. Agency for International Development National Defense University Woodrow Wilson International Center for Scholars)

Frontier Data Study detailed references

The evidence for the specific conclusions around new data sources and data techniques in Section 3 of this Study and expressed in the user Guides is drawn from a wide range of literature - including case studies, research reports, and cross-cutting literature (such as those highlighted above), Key Informant Interviews, and the technical experience of the Study team in discussing these issues over recent years with peers and other experts in relevant fields. Conclusions drawn are a professional judgement based on this wide field of evidence.

The following references set out some of the most significant literature sources in the drawing together of specific conclusions that are not already included in:

- the interactive map of individual case studies that were reviewed during the Study
- those references that are already highlighted in the main body of the report as recommended reading around the topic,
- and/ or in the table above.

The list below is not comprehensive, and it excludes those references already mentioned in the Report and user guides and include in the Frontier Data Study interactive map of innovations with new data sources. The available literature around the topics the Study covered is vast and there may be other sources not mentioned below which have influenced the Study's conclusions or that could provide contradictory evidence.

Selected sources of evidence for high potential data sources which can be used now in the right circumstances (Section 3.5 of the Report)

Cross-cutting ethical issues for future technologies:

- https://www.forbes.com/sites/jessicabaron/2018/12/27/tech-ethics-issues-we-should-all-be-thinking-about-in-2019/#58acda414b21
- https://aea365.org/blog/ite-tig-week-the-role-of-evaluators-in-the-fourth-industrial-revolution-by-valentine-j-gandhi/
- https://www.researchgate.net/publication/50247418_The_role_of_Ethics_in_the_process_of_Technology_Transfer_and_Development_of_206_Peugeot
- https://www.scientificamerican.com/article/the-many-ethical-implications-of-emerging-technologies/
- Bamberger, M., Raftree, L. & Olazabal, V. (2016) The role of new information and communication technologies in equity–focused evaluation: opportunities and challenges. Evaluation. Vol 22(2) 228–244.

Earth Observation and Next generation Earth Observation (EO)

- Flood mapping tools for disaster preparedness and emergency response using satellite data and hydrodynamic models: A case study of Bagmathi basin, India- G. Amarnath, K. Matheswaran, P. Pandey, N. Alahacoon and S. Yoshi-moto, International Water Management Institute (IWMI), Colombo, Sri Lanka.
- Modeling the Present and Future Desertification Risk State: A case study in Kolli hill, Eastern Ghats of Tamil Nadu, India- G. Saji and S. Jayakumar, Environmental Informatics and Spatial Modeling Lab (EISML), Department of Ecology and Environmental Sciences, School of Life Sciences, Pondicherry University, Puducherry, India.
- Abuelgasim, A.A., Ross, W., Gopal, S., Woodcock, C.E., 1999. Change detection using adaptive fuzzy neuralnetworks: environmental damage assessment after the Gulf War. Remote Sensing of Environment 70,208–223.
- Anderson, J.R., Hardy, E.E., Roach, J.T., Witmer, R.E., 1976. A land-use and land-cover classification system foruse with remote sensor data. US geological Survey Professional Paper 964, Washington, DC.
- Carlotto, M.J., 1999. Reducing the effects of space-varying, wavelength-dependent scattering in multispectralimagery. International Journal of Remote Sensing 20, 3333 –3344.
- Chan, J.C.-W., Chan, K.-P., Yeh, A.G.-O., 2001. Detecting the nature of change in an urban environment—a comparison of machine learning algorithms. Photogrammetric Engineering and Remote Sensing 67,213–225.
- Chavez, P.S., 1996. Image-based atmospheric corrections revisited and improved. Photogrammetric Engineering and Remote Sensing 62, 1025–1036.
- Chen, D., Stow, D.A., Tucker, L., Daeschner, S., 2001. Detecting and enumerating new building structuresutilizing very-high resolution image data and image processing. Geocarto International 16, 69– 82
- Su, Z.B. Troch P.A. (2003). Applications of quantitative remote sensing to hydrology. Physics and Chemistry of the Earth, 28(1-3): 1-2.

- Sui, D.Z. and Maggio, R.C. (1999). Integrating GIS with hydrological modeling: practices, problems, and prospects. Computers, Environment and Urban Systems, 23(1): 33-51.
- Voss, C.I. and Provost, A.M. (2002). SUTRA: a model for saturated-unsaturated variable-density ground-water flow with solute or energy transport. US. Geological Survey Water Resources Investigation Report WRIR 02-4231, 250 pp
- Burke, M. and Lobell, D.B. 2017. Satellite-based assessment of yield variation and its determinants in smallhoder African systems. Proceedings for the National Academy of Sciences. 114(9): 2189-2194.
- Goldblatt, R.; You, W.; Hanson, G.; Khandelwal, A.K. Detecting the Boundaries of Urban Areas in India: A Dataset for Pixel-Based Image Classification in Google Earth Engine. Remote Sens. 2016, 8, 634.
- Engstrom, Ryan; Hersh, Jonathan; Newhouse, David. 2017. Poverty from Space: Using High-Resolution Satellite Imagery for Estimating Economic Well-Being. Policy Research Working Paper;No. 8284. World Bank, Washington, DC.
- Group on Earth Observations Global Agricultural Monitoring (GEOGLAM), Whitcraft (GEOGLAM), F. Kerblat (CSIRO)
- Algal Bloom Early Warning Alert System; T. Malthus (CSIRO)
- Flood Prediction System Using the Global Satellite Map of Precipitation (GSMaP); R. Oki and M. Kachi (JAXA)
- Global Mangrove Watch – Mapping Extent and Annual Changes in the Global Mangrove Cover; A. Rosenqvist (SoloEO)
- Earth Observation for Water-related Ecosystem Monitoring; C. Giardino and M. Bresciani (CNR-IREA)
- Mapping Urban Growth, M. Paganini (ESA), T. Esch and M. Marconcini (DLR)
- Air Pollution Monitoring for Sustainable Cities and Human Settlements M. Kikuchi, A. Kuze and S. Sobue (JAXA)
- Using Remote Sensing for Water Quality Monitoring of the Great Barrier Reef; T. Schroeder (CSIRO)
- Mapping Forest Cover Extent and Change, and Progressing Sustainable Forest Management; A. Kavvada (NASA, GEO) and M. Hansen (University of Maryland)
- The Global Forest Observations Initiative and Space Agency Support to Forest Monitoring; M. Steventon and S. Ward (Symbios)
- Efforts Targeting Land Degradation to Achieve Neutrality; M. Paganini (ESA), A. Held (CSIRO), M. Cherlet (JRC), S. Minelli (UNCCD Sec.), S. Walter (UNCCD Global Mechanism)
- https://info.alen.space/predictions-about-the-future-of-small-satellites-and-new-space
- Big Earth Data: disruptive changes in EO data management and analysis: Sudmans et al (2019), International Journal of Digital Earth
- https://www.shapingtomorrow.com/home/alert/3769106-Future-of-Satellites
- NASA Earth Observatory Orbits Catalogue https://earthobservatory.nasa.gov/features/OrbitsCatalog

Earth Observation with Drones

- Gonzalez, Luis F., et al. "Unmanned aerial vehicles (UAVs) and artificial intelligence revolutionizing wildlife monitoring and conservation." Sensors 16.1 (2016): 97.
- Ammour, N.; Alhichri, H.; Bazi, Y.; Benjdira, B.; Alajlan, N.; Zuair, M. Deep Learning Approach for Car Detection in UAV Imagery. Remote Sens. 2017, 9, 312.
- Martini T., Lynch M., Weaver A., van Vuuren T. (2016) The Humanitarian Use of Drones as an Emerging Technology for Emerging Needs. In: Custers B. (eds) The Future of Drone Use. Information Technology and Law Series, vol 27. T.M.C. Asser Press, The Hague

Artificial Intelligence

- https://www.idiainnovation.org/s/AI-and-international-Development_FNL.pdf
- A. Halevy, P. Norvig and F. Pereira, "The Unreasonable Effectiveness of Data," in IEEE Intelligent Systems, vol. 24, no. 2, pp. 8-12, March-April 2009.
- Chen Sun, Abhinav Shrivastava, Saurabh Singh, Abhinav Gupta; The IEEE International Conference on Computer Vision (ICCV), 2017, pp. 843-852
- LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015). https://doi.org/10.1038/nature14539
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.
- Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, 2019.
- Najafabadi, Maryam M., et al. "Deep learning applications and challenges in big data analytics." Journal of Big Data 2.1 (2015): 1.
- Müller, Vincent C., and Nick Bostrom. "Future progress in artificial intelligence: A survey of expert opinion." Fundamental issues of artificial intelligence. Springer, Cham, 2016. 555-572.
- Pearl, J. & Mackenzie, D. (2018) The Book of Why: The New Science of Cause and Effect. New York, NY. Basic Books, Inc.
- McKenzie, D. How can machine learning and artificial intelligence be used in development interventions and impact evaluations? Accessed: https://blogs.worldbank.org/impactevaluations/how-can-machine-learning-and-artificial-intelligence-be-used-development-interventions-and-impact
- Strusani, D. and Vivien Houngbonon, G. The Role of Articial Intelligence in Supporting Development in Emerging Markets. EM Compas World Bank. Accessed: https://openknowledge.worldbank.org/bitstream/handle/10986/32365/The-Role-of-Artificial-Intelligence-in-Supporting-Development-in-Emerging-Markets.pdf?sequence=1&isAllowed=y
- Subjective happiness index based on twitter in Indonesia - Asita Sekar Asri, Siti Mariyah (2019)

Location data from mobile phones

References are mainly covered in the main sources of literature mentioned in the Report, interactive map of case studies, and the user guides, and all articles on this subject available via the UN Pulse Lab websites. They also include:

- Candia J, González M, Wang P, Schoenharl T, Madey G, Barabási AL. (2008) Uncovering individual and collective human dynamics from mobile phone records. Journal of Physics A: Mathematical and Theoretical 41
- Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A tale of many cities: universal patterns in human urban mobility. PloS one 7
- Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C. (2014) Geo-located twitter as proxy for global mobility patterns. Cartography and Geographic Information Science 41
- Lessons for effective public-private partnerships from the use of mobile phone data in Indonesian tourism statistics - Siim Eskom, Titi Kanti Lestari (2019)
- Measuring commuting statistics in Indonesia using mobile positioning data - Amanda Pratama Putra, Ignatius Aditya Setyadi, Siim Esko, Titi Kanti Lestari (2019)
- The use of big data as leading indicators of tourism demand - Titi Kanti Lestari, Siim Esko, Alexander Rayner, Amalia A. Widyasanti (2019)
- The use of mobile positioning data to obtain accommodation statistics: Case study of Indonesia - Agus Ruslani, Wa Ode Zuhayeni Madjida, Amin Rois Sinung Nugroho
- https://www.geopoll.com/blog/mobile-phone-penetration-africa/
- https://www.pewresearch.org/internet/2019/03/07/use-of-smartphones-and-social-media-is-common-across-most-emerging-economies/
- http://www.15th-tourism-stats-forum.com/pdf/Papers/S3/3_2_Indonesia%27s_Experience_of_using_Signaling_MPD_for_Official_Tourism_Statistics.pdf
- Mobile Phone Call Data as a Regional Socio-Economic Proxy Indicator (2015) - Sanja Šćepanović, Igor Mishkovski, Pan Hui, Jukka K. Nurminen, and Antti Ylä-Jääski; https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4405276/
- Ranjan G, Zang H, Zhang ZL, Bolot J (2012) Are call detail records biased for sampling human mobility? SIGMOBILE Mob Comput Commun Rev 16
- 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
- On the privacy-conscientious use of mobile phone data: https://www.nature.com/articles/sdata2018286

Selected sources of evidence of data sources and techniques with longer term potential (Section 3.6 of the Report)

General

- http://www.inveneo.org/wp-content/uploads/2014/07/FINALTop-ICT-Hardware-Challenges-White-Paper.pdf
- https://www.itu.int/en/action/broadband/Documents/Harnessing-IoT-Global-Development.pdf
- Gerster, R. & Zimmerman, S. (2003). Information and Communication Technologies (ICTs) For Poverty Reduction? Discussion Paper. Swiss Agency for Development and Cooperation. Available at http://www.gersterconsulting.ch/docs/ict_for_poverty_reduction.pdf
- Gerster, R., & Zimmerman, S. (2005). Upscaling Pro-Poor ICT Policies and Practices: A Review of Experience with Emphasis on Low-Income Countries in Asia and Africa. Swiss Agency for Development and Cooperation. Available at: http://www.itu.int/wsis/docs2/pc2/parallel/up-scalingict-policies.pdf
- Honan, M. (2014, Feb. 24). Facebook's Plan to Conquer the World with Crappy Phones and Bad Networks. Wired. Available at: https://www.wired.com/2014/02/facebook-plans-conquer-world-slew-low-end-handsets/
- Hosman, L. & Baikie, B. (2013). Solar Powered Cloud Computing Datacenters. IEEE IT Professional, March-April, 15-21.
- Kraemer, K., Dedrick, J. and Sharma, P. (2009). "One Laptop Per Child: Vision vs. Reality," Comm. ACM, vol. 52, no. 6, pp. 66–73.
- London, T. & Hart, S.L. (2004). Reinventing Strategies for Emerging Markets: Beyond the Transnational Model. Journal of International Business Studies, 35 (5), pp. 350-370.
- Waugamon, A., (2014). Using Technology for Social Good: An Exploration of Best Practice in the Use of Information and Communication Technologies (ICTs) for Development. Nashville: TN: United Methodist Communications. Available at: http://www.umcom.org/site/c.mrLZJ9PFKmG/b.9031619/k.4677/Using_Technology_for_Social_Good.htm

The Internet of Things (IoT)

- http://www.inveneo.org/wp-content/uploads/2014/07/FINALTop-ICT-Hardware-Challenges-White-Paper.pdf
- https://www.itu.int/en/action/broadband/Documents/Harnessing-IoT-Global-Development.pdf
- Gerster, R. & Zimmerman, S. (2003). Information and Communication Technologies (ICTs) For Poverty Reduction? Discussion Paper. Swiss Agency for Development and Cooperation. Available at http://www.gersterconsulting.ch/docs/ict_for_poverty_reduction.pdf
- Gerster, R., & Zimmerman, S. (2005). Upscaling Pro-Poor ICT Policies and Practices: A Review of Experience with Emphasis on Low-Income Countries in Asia and Africa. Swiss Agency for Development and Cooperation. Available at: http://www.itu.int/wsis/docs2/pc2/parallel/up-scalingict-policies.pdf

- Honan, M. (2014, Feb. 24). Facebook's Plan to Conquer the World with Crappy Phones and Bad Networks. Wired. Available at: http://www.wired.com/gadgetlab/2014/02/facebook-plans-conquerworld-slew-low-end-handsets/
- Hosman, L. & Baikie, B. (2013). Solar Powered Cloud Computing Datacenters. IEEE IT Professional, March-April, 15-21.
- Kraemer, K., Dedrick, J. and Sharma, P. (2009). "One Laptop Per Child: Vision vs. Reality," Comm. ACM, vol. 52, no. 6, pp. 66–73.
- London, T. & Hart, S.L. (2004). Reinventing Strategies for Emerging Markets: Beyond the Transnational Model. Journal of International Business Studies, 35 (5), pp. 350-370.
- Waugamon, A., (2014). Using Technology for Social Good: An Exploration of Best Practice in the Use of Information and Communication Technologies (ICTs) for Development. Nashville: TN: United Methodist Communications. Available at: http://www.umcom.org/site/c.mrLZJ9PFKmG/b.9031619/k.4677/Using_Technology_for_Social_Good.htm

Privacy Preserving Data Sharing (PPDS) via the Cloud

- M. Kallahalla, E. Riedel, R. Swaminathan et al., "Scalable secure file sharing on untrusted storage," in Proceedings of the FAST'03 Proceedings of the 2nd USENIX Conference on File and Storage Technologies, pp. 29–42, 2003. View at: Google Scholar
- S. Yu, C. Wang, K. Ren, and W. Lou, "Achieving secure, scalable, and fine-grained data access control in cloud computing," in Proceedings of the IEEE INFOCOM, pp. 1–9, March 2010.View at: Publisher Site | Google Scholar
- V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in Proceedings of the 13th ACM Conference on Computer and Communications Security (CCS '06), pp. 89–98, November 2006. View at: Publisher Site | Google Scholar
- X. Dong, J. Yu, Y. Luo, Y. Chen, G. Xue, and M. Li, "Achieving an effective, scalable and privacy-preserving data sharing service in cloud computing," Computers & Security, vol. 42, pp. 151–164, 2014.View at: Publisher Site | Google Scholar
- J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-policy attribute-based encryption," in Proceedings of the IEEE Symposium on Security and Privacy (SP '07), pp. 321–334, May 2007.View at: Publisher Site | Google Scholar
- B. Waters, "Ciphertext-policy attribute-based encryption: an expressive, efficient, and provably secure realization," in Public Key Cryptography (PKC '11), pp. 53–70, Springer, Berlin, Germany, 2011.View at: Publisher Site | Google Scholar | MathSciNet
- X. Boyen and B. Waters, "Anonymous hierarchical identity-based encryption (without random oracles)," in Advances in Cryptology—CRYPTO 2006, vol. 4117 of Lecture Notes in Computer Science, pp. 209–307, Springer, Berlin, Germany, 2006.View at: Publisher Site | Google Scholar
- Beimel., Secure schemes for secret sharing and key distribution [Ph.D. thesis], Israel Institute of Technology, Technion, Haifa, Israel, 1996.
- M. Ito, A. Saito, and T. Nishizeki, "Secret sharing scheme realizing general access structure," Electronics & Communications in Japan, vol. 72, no. 9, pp. 56–64, 1989.View at: Google Scholar

- J. Benaloh and J. Leichter, “Generalized secret sharing and monotone functions,” On Advances in Cryptology, vol. 403, pp. 27–36, 1988.View at: Google Scholar
- M. Karchmer and A. Wigderson, “On span programs,” The Eighth Annual Structure in Complexity Theory, pp. 102–111, 1993.View at: Google Scholar
- B. H. Bloom, “Space/time trade-offs in hash coding with allowable errors,” Communications of the ACM, vol. 13, no. 7, pp. 422–426, 1970. View at: Publisher Site | Google Scholar
- T. Nishide, K. Yoneyama, and K. Ohta, “Attribute-based encryption with partially hidden encryptor-specified access structures,” in Proceedings of the International Conference on Applied Cryptography & Network Security, vol. 5037, pp. 111–129, 2008.View at: Google Scholar
- J. Lai, R. H. Deng, and Y. Li, “Expressive CP-ABE with partially hidden access structures,” in Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security, ASIACCS 2012, pp. 18-19, Republic of Korea, May 2012. View at: Google Scholar
- S. Jiang, X. Zhu, and L. Wang, “EPPS: Efficient and privacy-preserving personal health information sharing in mobile healthcare social networks,” Sensors, vol. 15, no. 9, pp. 22419–22438, 2015.View at: Publisher Site | Google Scholar
- R. J. McEliece and D. V. Sarwate, “On sharing secrets and Reed-Solomon codes,” Communications of the ACM, vol. 24, no. 9, pp. 583-584, 1981.View at: Publisher Site | Google Scholar | MathSciNet
- S. Obana, “Almost optimum t-cheater identifiable secret sharing schemes,” in EUROCRYPT, vol. 6632, pp. 284–302, 2011.View at: Google Scholar
- H. Hoshino and S. Obana, “Cheating detectable secret sharing scheme suitable for implementation,” in Proceedings of the 4th International Symposium on Computing and Networking, CANDAR 2016, pp. 623–628, Japan, November 2016.View at: Google Scholar

Next generation Artificial Intelligence

- Preparing for the Future of Artificial Intelligence 2016, Executive Office of the President National Science and Technology Council Committee on Technology: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf
- https://www.forbes.com/sites/forbestechcouncil/2019/11/06/the-next-generation-applications-of-artificial-intelligence-and-machine-learning/#3961700347bc
- https://brighterion.com/next-generation-artificial-intelligence-machine-learning/
- https://www.usaid.gov/digital-development/machine-learning/AI-ML-in-development/summary
- https://www.usaid.gov/cii/ai-in-global-health
- Rahman MR, Lateh H (2017) Climate change in Bangladesh: a spatio-temporal analysis and simulation of recent temperature and rainfall data using GIS and time series analysis model. Theoret Appl Climatol 128(1-2):27–41CrossRefGoogle Scholar
- Schneider A, Friedl MA, Potere D (2009) A new map of global urban extent from MODIS satellite data. Environ Res Lett 4(4):044003CrossRefGoogle Scholar

- Levien LM, Roffers P, Maurizi B, Suero J, Fischer C, Huang X 1999 A machine-learning approach to change detection using multi-scale imagery. American Society of Photogrammetry and Remote Sensing. Annual conference, Portland, Oregon, May 20, 1999Google Scholar
- Tang Z, Tang H, He S, Mao T (2015) Object-based change detection model using correlation analysis and classification for VHR image. IEEE international geoscience and remote sensing symposium (IGARSS), pp. 4840–4843, July 2015Google Scholar
- Pal M (2006) Support vector machine-based feature selection for land cover classification: a case study with DAIS hyperspectral data. Int J Rem Sens 27(14):2877–2894CrossRefGoogle Scholar
- Zhao J, Gong M, Liu J, Jiao L (2014) Deep learning to classify difference image for image change detection. International joint conference on neural networks (IJCNN), pp 411-417Google Scholar
- Chan T, Jia K, Gao S, Lu J, Zeng Z, Ma Y (2015) PCANet: a simple deep learning baseline for image classification? IEEE Trans Image Process 24(12):5017–5032MathSciNetzbMATHCrossRefGoogle Scholar
- Kuwata K, Shibasaki R (2015) Estimating crop yields with deep learning and remotely sensed data. IEEE Intl Geosci Rem Sens Symp (IGARSS):858–861Google Scholar
- Zhu XX, Tuia D, Mou L, Xia G, Zhang L, Xu F, Fraundorfer F (2017) Deep learning in remote sensing: a comprehensive review and list of resources. IEEE Geosci Rem Sens Mag 5(4):8–36CrossRefGoogle Scholar

Blockchain

- https://blockchainomics.tech
- Search for Common Good Webinar: Blockchain for International Development https://www.youtube.com/watch?v=AHFA81QpMgk&t=580s
- https://www.eurasia.undp.org/content/rbec/en/home/presscenter/pressreleases/2018/blockchain-research-to-support-sustainable-development-goals.html
- https://www.gsb.stanford.edu/sites/gsb/files/publication-pdf/study-blockchain-impact-moving-beyond-hype.pdf
- Blockchain Climate Risk Crop Insurance – Global Innovation Lab for Climate Finance https://climatepolicyinitiative.org/wp-content/uploads/2019/10/Blockchain_instrument-analysis.pdf

APPENDIX 3: METHODOLOGICAL NOTES

The NIRAS Data Futures Hub Study team was led by **Matthew Shearing,** with **Dr Valentine Gandhi** as the lead Technical Advisor and supported by a small team of data experts: **Will Apted, Casper Samsø Fibæk, and Dr Matthew McConnachie.** The development of the methodology and drawing of conclusions was influenced by the team's experience and expertise in data science, official statistics, data for development, international development programme implementation, and prior experience of working with DFID in particular.

The overall methodology is set out in Section 2 of the Study and relied on gathering, analysing and synthesising relevant opinions and research from a wide range of relevant stakeholders. The methodology, especially the detail of the approach, evolved as the results of each phase of the Study were analysed.

This Appendix provides more details on some of the key activities and some of the high-level results which influenced the conclusions.

DFID working group

A small team of DFID staff were selected to provide guidance on the development of the detail of the methodology, the development of the structure of the report, and interpretation of the results.

- Paddy Brock, Senior Statistician and Team Leader, FCDO
- Rachael Beaven, Data Revolution Lead, FCDO
- Ian Coady, Geospatial Adviser, FCDO
- Benjamin Kumpf, Head of Innovation, DFID
- Alex Jones, Deputy Head of EPIC Department and Head of Emerging Futures, FCDO
- Paula McLeod, Head of Profession, Statistics Department, DFID
- Mandeep Samra, Digital Adviser, DFID
- Billie Selby, Innovation and Technology Policy and Programme Manager, FCDO
- Andrew Toft, Digital Policy Adviser, FCDO
- Tom Wilkinson, Head of Data Science at FCDO & Site Lead for Data Science Campus, Office for National Statistics

Senior Advisor Group

Eight selected globally eminent experts from a mix of relevant backgrounds (official statistics, data science in international development and local policymaking, cyber-security, monitoring and evaluation, 'tech4good', from a range of public and private sectors) provided a range of pro bono inputs on technical questions throughout the Study, including the design of the global stakeholder survey, interpretation of the evidence and support in drawing conclusions:

- Dr Michael Bamberger, big data for development expert
- Dr Rick Davies, Tech4Good/ development expert
- Titi Kanti Lestari, Indonesian Bureau of Statistics
- Derval Usher, UNICEF/ ex Head of UN Pulse Lab Jakarta
- Chris Sumner, Co-Founder/ data security expert at the Online Privacy Foundation, Associate CLODE Consultants
- Dr Brian King, big data for agriculture coordinator, CGIAR
- Setiaji, Head of ICT, West Java Government, Indonesia
- Dr Cora Mezger, Director of Statistical Consulting, Department of Statistics, Oxford University

Initial consultation with selected global experts

A first step for the Study was in consulting selected global stakeholders to help shape the details to be addressed in the next steps. As well as discussions with the DFID e-working group, this included remote consultation with the Senior Advisor Group and Key Informant Interviews with some of them and selected others, as per:

Name	Position	Date of interview
Dr Brian King	Coordinator CGIAR Big Data Forum	08th October
Dr Michael Bamberger	Independent Consultant, BIG Data for Evaluation Expert	11th October 2019
Dr Jane Thomason	CEO, Blockchain quantum impact/fintech worldwide	12th October 2019
Anna Lerner Nesbitt	Program Manager, Global Compact: Data and AI at Facebook	14th October 2019
Dr Sridhar Gummadi	Senior Data Scientist, IRRI	14TH October 2019
Craig Newmark	Founder, Craigslist	15th October 2019

These consultations focussed on unpicking how best to focus research in answering the following questions from the Study's Terms of Reference:

- What does Big Data mean in the development context?
- What are the most promising use cases DFID could support - in particular niche areas or at scale?
- What might be useful activities for DFID's forthcoming Data Science campus around big data?

- What has past experience of applying Big Data in development contexts revealed in terms of risks and pitfalls to avoid?
- What do we know about who is left behind when it comes to Big Data?
- What is not useful for developing countries in this area?
- What are good principles for investing in big data in a responsible way?

The following is summary of the emerging themes from the initial discussions:

Defining Big Data

General agreement on the 5 Vs (Volume, Velocity, Veracity, Value and Variety). The respondents agree that big data should be defined as distinct from traditional data sources at the same time included in the larger data landscape discussions, primarily to ensure the checks and balances of traditional data sources are also applied to big data.

Emerging operational definition

Big data refers to the large, diverse sets of information that grow at ever-increasing rates. It encompasses the volume of information, the velocity or speed at which it is created and collected, and the variety or scope of the data points being covered. Big data often comes from multiple sources and arrives in multiple formats.

Segmentation of Data Sources

The responses were similar, in that, typologies should depend on the end user's needs rather than the technological capacities, if one were to classify big data cases, it's easy to arrange them as sectors, however, in the context of development, if we follow a bottom up approach, based on the issue at hand that would be more useful.

Big data sources that are more useful:

- We could make fancy models, but how reliable are they, are they answering the questions? Should start with helping DFID programme managers understand if their field level interventions need big data, or what value is that adding.
- Sentiment analysis of social media, real time tracking of events, forecasting models can be useful.

Dealing with challenges

There is a cautious approach to using big data in development, and the thinking is emerging. One view is that research and evaluation, as a field, would be overwhelmed by big data / data science rhetoric. But there has been countervailing development, which is that evaluative thinking is pushing back against unthinking enthusiasm for the use of data science algorithms.

The emphasis is on "evaluative thinking" rather than "evaluators" as a category of people, because a lot of this pushback is coming from people who would not identify themselves as evaluators. There are different strands to this evaluative response.

One is a social justice perspective, reflected in recent books such as "Weapons of Math Destruction", "Automated Inequality", and "Algorithms of Oppression" which emphasise the human cost of poorly designed and or poorly supervised use of algorithms using large amounts of data to improve welfare and justice administration. Another strand is more like a form of exploratory philosophy and it has focused on how it might be possible to define "fairness" when designing and evaluating algorithms that have consequences for human welfare.

Another strand is perhaps more technical in focus, but still has a value concern. This is the literature on algorithmic transparency. Without transparency it is difficult to assess fairness neural networks are often seen as a particular challenge. Associated with this are discussions about "the right to explanation" and what this means in practice.

In parallel there is also some infiltration of data science thinking into mainstream evaluation practice. DFID is funding the World Bank's Strategic Impact Evaluation Fund (SIEF) latest call for "nimble evaluations". These are described as rapid and low cost and likely to take the form of an RCT, but ones which are focused on improving implementation rather than assessing overall impact. This type of RCT is directly equivalent to A/B testing used by the internet giants to improve the way their platforms engage with their users.

Hopefully these nimble approaches may bring a more immediate benefit to people's lives than RCTs, which have tried to assess the impact of a whole project and then inform the design of subsequent projects. In addition to these there are technical challenges to Big data as well, like problems of integrity, data quality, technical access, as well as inclusion related issues.

Consultation with DFID Data Science Hub

Discussions with the head of DFID Data Science Hub (DSH) on 31 October 2019, indicated:

- In terms of optimising the efficiency and utility of the Frontier Data Study's outputs in coordination with the activities of the DSH, the following were noted:
    - DSH are developing models for Use-Cases of different data sources
    - DSH are engaging in capacity-building activities for DFID staff

- There is interest from DSH in the Study providing:
    - Practical guidance and a blueprint for the development of user tools and integrating them into the design and delivery of DFID Programmes
    - Support in raising relevant awareness and capacity of DFID staff, driving a data-aware community
    - Identifying new projects for DSH to work on
    - Identifying potential skills needs/activities within the DSH to do the above effectively

Global stakeholder survey

Two parallel consultative activities were carried out, alongside a review of relevant global literature and case study research: a survey of global stakeholders and focus groups with DFID staff.

The survey was targeted at different sampling groups, using mass-marketing through social media and direct contacting by email of around a dozen individual organisations in each target group. The groups and resulting level of responses were:

Target Sample Group	Responding Organisations / individuals
Statisticians	Statistics Netherlands, Palestine Central Bureau of Statistics, Indonesian Central Bureau of Statistics, UNICEF (Data and Analytics)
Donor and implementing organisations	GIZ, Integrity (x2), Data2X, independent M&E experts (x2), Asian Development Bank (ADB), Land Equity, Practical Participation, ODW Consulting, UNICEF Libya, UNESCAP, UNICEF (x2)
European Space Agency (ESA)	European Space Agency (ESA)
Knowledge Drivers/ Data Scientists	Development Café, the University of Edinburgh

The Survey questions, which were all accompanied by some discursive text around the purpose of the questions, were:

Defining our challenges

1. What do you think is the value of defining big data separately to other digital or traditional sources of data?
2. How can we make a useful segmentation of big data sources that help us make the right choices in designing data processes that support development needs?
3. How do you think we can best use big data or other new data sources in ways that support long term improvements in decision-making capabilities across society, including ensuring equality of access to readily usable data?

Refining the evidence-base

1. What applications of new sources of data do you think are particularly ripe for exploitation and why? Which ones are particularly problematic?
2. Do you think big data or other new sources of data hold particular promise for impact for decision-making in particular sectors or geographies?

Dealing with the challenges

1. What do you think are the most pressing technical and cross-cutting challenges in optimising the use of emerging data sources?
2. Do you have any evidence of what can go ethically wrong in terms of the way we use new data sources?
3. What do you think are the most pressing needs in the way we use emerging data sources for ensuring we can effectively monitor issues across different population groups? ('leaving no-one behind')

The results of the survey were all analysed and are reflected in the overall conclusions of the Study. Summaries are provided below:

Segementation of Data Sources
Support for the definitions proposed in the Study
Need use-driven perspective
Active vs passive is most useful distinction (points towards inherent biases)
Too much literature on classification, not enough on usage
Defining data sources
Already 'well-defined' v too vague'
Categorisation allows for required different approaches
Maybe more about questions to ask about the data, and solutions will follow the answers
Focus on driving value through integration
Most promising new data sources
Spatial mapping/ remote sensing
Mobile phone positioning data
Tax data
Social media
Health records

Most promising Sectors/ Geographies
Climate
Health
Disaster forecasting
Population movements
Agriculture
Least promising new data sources
All of them
Social media
AI and algorithms
Supporting long-term improvements in decision-making
Treat new sources with scepticism
New sources most useful on testing assumptions
Same challenges as traditional sources: is this an opportunity to tackle them?
Open data for scrutiny of decisions [transparency challenge]
What can go wrong with ethics
Data access, security, and ownership is key
Conflict research with GIS
Machine Leaning bias
Traditional standards not applied
Most pressing needs for leaving no-one behind
Ethical and data quality frameworks are fundamental – inclusive of the target groups
Culture/language variations can have a deep impact on result

Cross-cutting challenges
Curtailing over-enthusiasm
Technical capabilities and inter-disciplinary working
Investment in technology
Organisational re-alignment
Public understanding
Equality of access
Data governance

DFID staff focus groups

Consultations with DFID staff as data users were focussed on interactive discussions around:

- Concerns and excitements about the possibilities of new data sources
- Priority needs for improved data
- What help is needed

Four focus groups were held:

DFID office location	Number of participants	Date
East Kilbride, UK (and remote from Country Offices)	7	31 October 2019
London, UK	4	14 January 2020
East Kilbride, UK (and remote from Country Offices)	10	16 January 2020
Kathmandu, Nepal	5	20 January 2020

Other Key Informant Interviews

Two group discussions were held with a range of data experts.

On 17 February 2020, a '**Frontier Data Study Workshop**' was held in Jakarta, Indonesia, as a centre of growing expertise in innovation with new data sources to connect with decision-making in the public and private sectors. The discussion focussed on sense-checking the emerging conclusions of the Study, including the potential of new data sources and what is needed to operationalise them.

Participant	Organisation
Titi Kani Lestari	Indonesian Central Burau of Statistics (BPS), Director leading on big data for official statistics
Ramda Yanurzha	VP Research and Insights, Gojek (Indonesian App for ride-hailing and other personal services) – attended in personal capacity
Endiyan Rakhmanda	Co-Founder IYKRA (education platform for technology focusing on developing professionals' and organizations' capability in business, data and technology) (and formerly of Jakarta Smart City)
Sriganesh Lokanathan	UN Pulse Lab Jakarta, Data Innovation and Policy Lead
An Nisa Tri Astuti	Tifa Foundation (works with Indonesian civil society to promote open society in Indonesia)

On 20 January 2020, a discussion was held in Kathmandu with 4 data scientists from the **World Bank data team for the DFID funded Partnership for Knowledge based Poverty Reduction and Shared Prosperity in Nepal.**

Other Key Informant Interviews (KIIs)

The DFID deputy Chief Scientific Advisor was interviewed on 21 February 2020 to sense-check emerging conclusions and inform how the Study could best set out its final conclusions.

Endnotes

1 Data Interoperability Standards Consortium https://datainteroperability.org/

2 Dr Michael Bamberger, Dr Rick Davies, Titi Kanti Lestari, Derval Usher, Chris Sumner, Dr Brian King, Setiaji, Dr Cora Mezger. See Appendix 3 for more details.

3 https://unsdg.un.org/sites/default/files/UNDG_BigData_final_web.pdf

4 Adapted from various in international standards in official statistics, and the DFID Smart Guide to Data Quality

5 https://spaceknow.com/case-studies/deforestation/SpaceKnow-Deforestation-Case-Study.pdf

6 https://www.pewresearch.org/internet/2019/03/07/use-of-smartphones-and-social-media-is-common-across-most-emerging-economies/

7 Even though mobile networks generally change in line with national borders, this is still possible by monitoring the switching of SIM cards and given that some tracking coverage extends over borders. Indonesian National Statistical Office has shown this is possible by monitoring people movements across the Indonesia-East Timor border and other areas such as PNG.

8 https://climatepolicyinitiative.org/wp-content/uploads/2019/10/Blockchain_instrument-analysis.pdf

9 National Science and Technology Council 2016, 7

10 https://opendatawatch.com/blog/is-open-data-at-odds-with-citizens-privacy/

11 This broad point was raised by Dr Rick Davies. Theory-improvement needs such as these are in a paper for CEDIL (funded by DFID) in 2018: Davies, Rick. 2018. 'Representing Theories of Change: Technical Challenges with Evaluation Consequences'. CEDIL.

FRONTIER
TECHNOLOGIES

NIRAS
DIGITAL FUTURES HUB

Frontier Technologies Hub: helping the UK Foreign, Commonwealth and Development Office harness the potential of frontier technologies to tackle some of the world's biggest challenges

The Foreign & Commonwealth Office (FCO) and the Department for International Development (DfID) merged on 1 September 2020 to form the Foreign, Commonwealth & Development Office (FCDO).

The FCDO pursues our national interests and projects the UK as a force for good in the world. We promote the interests of British citizens, safeguard the UK's security, defend our values, reduce poverty and tackle global challenges with our international partners.