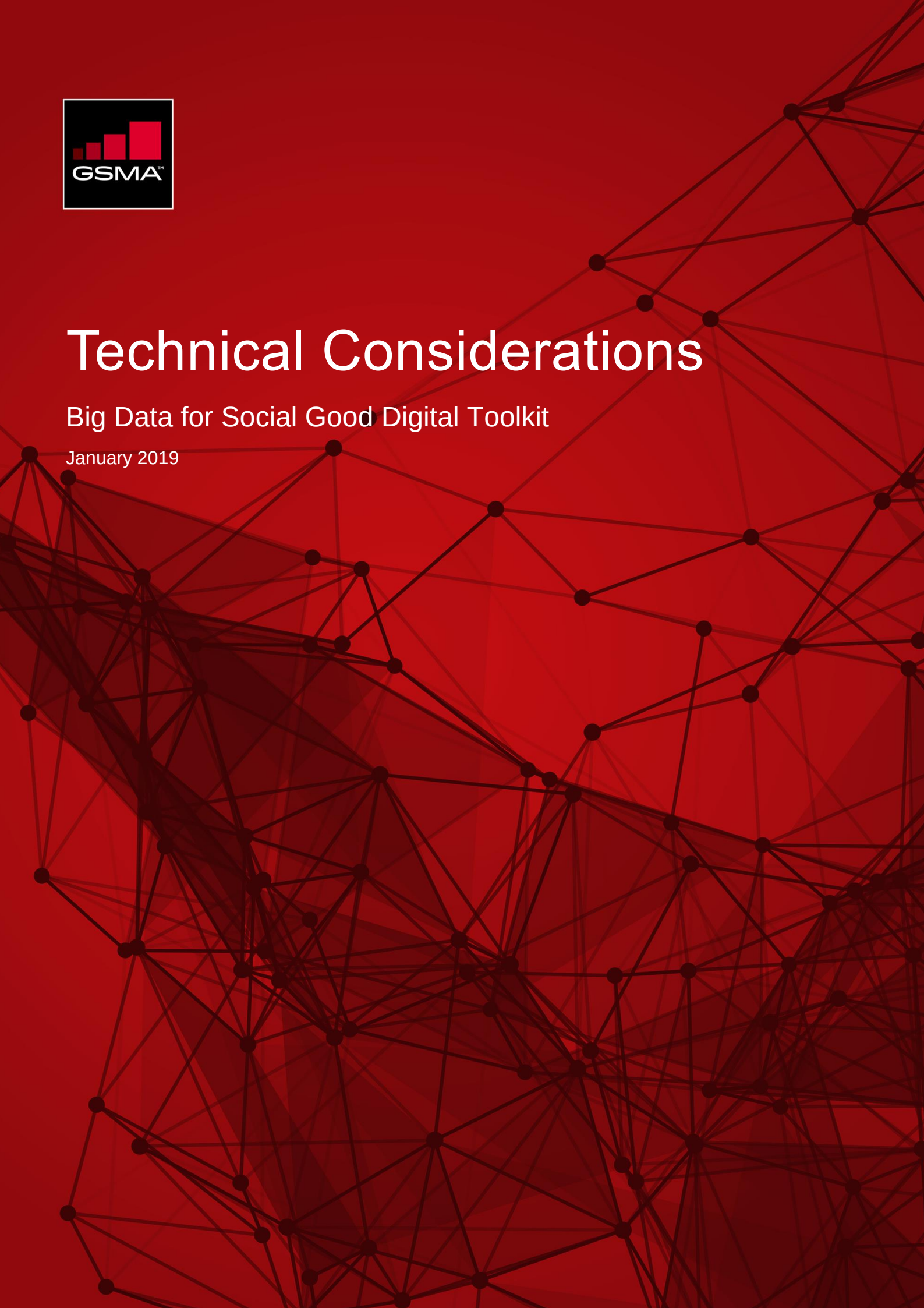


Technical Considerations

Big Data for Social Good Digital Toolkit

January 2019

Table of Contents

Introduction

This document is intended to outline the key technical considerations that must be made when working with Mobile Big Data. It is intended primarily for a non-technical audience from any current or interested stakeholder in a project involving Mobile Big Data. This could include government agencies, non-governmental agencies, charities or institutions working in the development sector, or commercial third parties. The document aims to build a general understanding of the use of Mobile Big Data to create actionable insights, given increasing stakeholder participation and demand for Mobile Big Data in social good projects.

Please note that this document forms part of the Big Data for Social Good Digital Toolkit. It is recommended to view other sections of this toolkit. Of particular relevance to this section is the information that relates to privacy and ethical considerations, which is to be found in the Policy and Regulation section of the toolkit. The principles outlined there govern all activities in this space.

Types of Mobile Big Data

Over the last 15 years, a significant body of research has proven that Mobile Big Data can offer new insights into otherwise invisible phenomena. There are three main types of information that can be derived from Mobile Big Data.

Location data

The most well-established use of Mobile Big Data is in field of human mobility and population mapping. Location data derived from network operations can shed light on this in near-real-time and at high granularity. For an overview of the study of human mobility, see (Barbosa-Filho et al., 2017). For the application of mobile phone data for population mapping, see (Deville et al., 2014). For a range of applications, see (Alexander, Jiang, Murga, & González, 2015; Astarita & Florian, 2001; Bengtsson et al., 2015; Iqbal, Choudhury, Wang, & González, 2014; Naboulsi, Fiore, Ribot, & Stanica, 2016; Wesolowski et al., 2013).

Social data

Patterns of communications relate to social networks, which can reveal international and interregional relationships. This has been shown to correlate with economic resilience and prosperity (Eagle, 2010), and has been used to monitor the aftermath of natural disasters (Lu, Wrathall, & Sundsøy, 2016). Monitoring the volume of activity in areas spreading out from the epicentre of a natural disaster can inform responders about the spread and scope of the event in its immediate aftermath.

Economic data

In some geographies, spending on mobile airtime is very closely linked to disposable income, and is thus a good measure of economic prosperity and food security (Smith, Mashhadi, & Capra, 2013). This can be used for building regional income maps, and dynamically for forming early responses to famine conditions.

Of these three areas, Location data is the best established, best understood, and most generalisable, and is the focus of this document.

The Big Data Process Framework

Mobile Big Data shares commonalities with other Big Data sources in the steps required to obtain valuable, actionable insights. There are broadly three stages, which are illustrated at a high level in Figure 1.

Processing

Like all real-world Big Data sets, Mobile Big Data is noisy. It requires care and expertise to produce accurate, reliable results, perform quality assurance and validation, and correct or control for potential sources of bias. These capabilities lie within operators, where the data is generated and thus best understood. Cleansing, filtering, and validating is covered in the Processing phase.

Analytics

The cleansed and processed data is combined with other sources and subjected to analysis around the research questions or objectives. This is the Analytics phase. The specifics of this stage are heavily dependent on the project and the desired outcomes, but it will almost certainly involve data from other, non-mobile sources, such as satellite or survey data or infrastructure maps, and the application of one or more specialised tools and techniques, such as geographical information systems (GIS), machine learning or statistical analysis. The expertise for this phase may sit within operators, agencies, governments or third parties.

Packaging

In the final Packaging phase, the results are compiled in a useable format, so that they can be incorporated into existing decision-making processes. There are several factors to consider when determining the optimal output format: key requirements to address are ongoing vs one-off, visual vs quantitative, and interactive vs static.

This document walks through some of the expert considerations that operators apply to ensure a successful project. The sections are structured around the three phases.

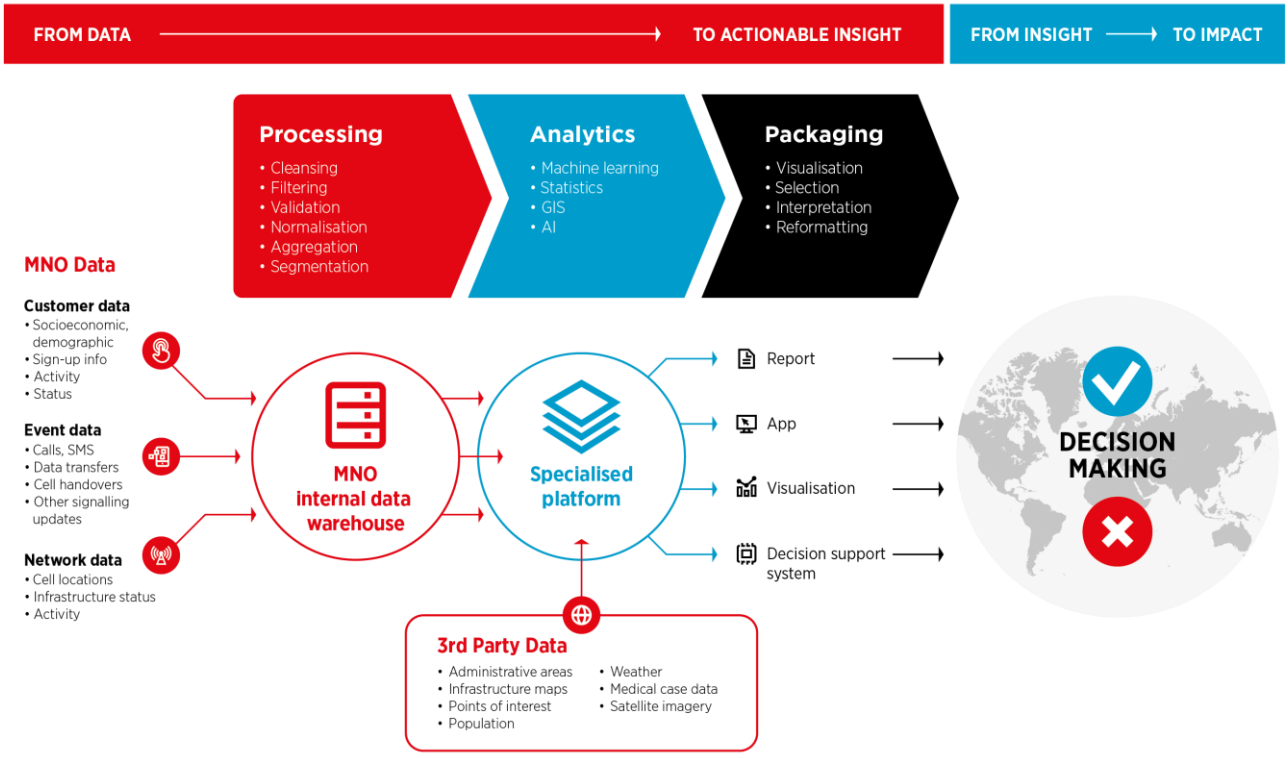


Figure 1 The Big Data Processing Framework. High-level illustration of the stages of Big Data processing

Processing

Raw Data

Mobile Big Data in its raw form consists of records of **events** where a SIM card has connected to a specific part of the network at a specific time. This gives an approximate location at that time for the SIM card, and thus – probably – the customer that owns the SIM card. Different network systems generate event data in different formats with different collections of fields, so the precise processing pipeline has to be designed by the network owners. Fields can include the sending and receiving numbers, identifiers for the part of the network that handled the origin and receipt of the event, and the start and end time of the event.

A simplified sample of raw mobile event data might look like this:

```
4abc557e-31a7-4317-b685-577417d83cf0|0088412322|1717|9072918240|6170|2016-03-02T04:29:05.390+05:30|2016-03-02T04:29:05.390+05:30|SMS|0.5534686|ANSWERED
4f0d7102-ae4e-41a9-aba9-1d51a873a6ac|5385434610|6649|2813059711|7325|2016-01-31T22:47:52.922+05:30|2016-01-31T22:47:52.922+05:30|SMS|0.23320162|ANSWERED
2eff3cab-2d05-4e00-b85a-7e7734b01c1a|9169498166|9914|9667558137|9410|2016-02-20T05:56:10.169+05:30|2016-02-20T05:58:35.541+05:30|VOICE|0.96220714|ANSWERED
e36a70d2-0cd1-4d3c-976a-3fb153fa77b1|9006790545|4290|3871719770|5947|2016-02-23T17:14:22.157+05:30|2016-02-23T17:16:32.837+05:30|VOICE|0.94779676|ANSWERED
4a9ab89e-bf68-4a97-8210-f4a1851d0818|8220136935|9475|8937965555|2433|2016-01-29T02:57:27.196+05:30|2016-01-29T02:59:57.101+05:30|VOICE|0.5948102|ANSWERED
a0b02d9f-adda-4a6d-8fc6-8c3598bacd23|5181682740|7483|4756301014|3051|2016-02-17T21:10:24.848+05:30|2016-02-17T21:13:10.373+05:30|VOICE|0.094239|ANSWERED
043775f2-39c2-4c88-a2ac-6297c273ab83|0539876732|4123|2424901009|8636|2016-02-22T11:28:09.438+05:30|2016-02-22T11:28:09.438+05:30|SMS|0.3759588|ANSWERED
29f16e0e-8d22-4f20-a016-ff32d65d7697|8079051745|6165|7424830931|8237|2016-02-12T19:22:15.275+05:30|2016-02-12T19:24:26.418+05:30|VOICE|0.22994667|ANSWERED
66ab2f37-3a2d-4ac9-b09e-2ff93d87c324|8542342283|9124|9938689656|9819|2016-01-20T09:20:08.554+05:30|2016-01-20T09:20:08.554+05:30|SMS|0.6758641|ANSWERED
c3b29a0b-8354-46d3-8e99-b236a0eb52d5|8792207050|7969|6782021797|6366|2016-02-26T06:12:25.353+05:30|2016-02-26T06:15:20.571+05:30|VOICE|0.5348344|ANSWERED
3785aa77-5dff-43d8-be8a-69b40fe85e4e|2720995984|8707|6688624826|3964|2016-02-04T02:29:30.445+05:30|2016-02-04T02:29:30.445+05:30|VOICE|0.50171614|BUSY
5fa9ca7e-ef70-43c0-aa31-b19c7e6f8eba|3003772083|2602|5064298733|4606|2016-02-17T10:25:44.443+05:30|2016-02-17T10:28:28.512+05:30|VOICE|0.40635967|ANSWERED
ac2a9f75-312d-440c-bb71-b802c5b79c9d|9934602193|7244|1663352395|8535|2016-02-19T20:54:05.370+05:30|2016-02-19T20:54:05.370+05:30|SMS|0.31655604|ANSWERED
69308480-716e-48ad-addd-e8e612b949b7|9241496799|1973|9695010582|1820|2016-03-01T00:02:10.805+05:30|2016-03-01T00:04:54.925+05:30|VOICE|0.40509063|ANSWERED
929a6d2d-13de-4e62-b1b5-69c90db80c79|6306112673|1433|1530483937|6294|2016-02-08T09:47:05.350+05:30|2016-02-08T09:47:05.350+05:30|SMS|0.26306754|ANSWERED
27d3ee44-83c6-4113-b433-d1c244171716|2731118089|6232|5367530118|2624|2016-02-14T16:26:59.216+05:30|2016-02-14T16:29:03.040+05:30|VOICE|0.479172|ANSWERED
4b170b67-02c1-4b1c-a1aa-f285a3860f3c|2595452748|5126|3993041501|5750|2016-02-12T03:44:49.451+05:30|2016-02-12T03:44:49.451+05:30|SMS|0.6606054|ANSWERED
44c300c0-17cf-4789-b53e-859abc64a5a0|6490876954|7840|4209844062|7741|2016-03-04T20:46:34.694+05:30|2016-03-04T20:46:34.694+05:30|SMS|0.69000804|ANSWERED
55060046-e6ef-483a-b2c6-ad64a5bff38d|8296649958|5030|0844472010|9415|2016-02-02T03:14:47.881+05:30|2016-02-02T03:14:47.881+05:30|SMS|0.39084607|ANSWERED
e74e8b13-ecec-459d-8485-5720a0c47cf1|5211450746|8303|2384067870|7353|2016-02-06T03:49:00.538+05:30|2016-02-06T03:51:53.954+05:30|VOICE|0.11167997|ANSWERED
6250d2ba-fbfa-4d1c-a26c-280241e1ee84|7790387215|5831|2135458592|5649|2016-01-21T16:52:36.641+05:30|2016-01-21T16:52:36.641+05:30|VOICE|0.19077021|BUSY
422aa5bd-0308-44ed-ba65-9abfa4751282|4264572099|1189|0702817413|4487|2016-02-11T05:16:04.589+05:30|2016-02-11T05:18:17.164+05:30|VOICE|0.18305779|ANSWERED
1f9f55a8-1b4b-4b28-9e83-4cf3c27c71e6|2817504805|1638|3945575896|2816|2016-03-04T10:00:16.024+05:30|2016-03-04T10:02:32.430+05:30|VOICE|0.93125194|ANSWERED
ba9a2e49-25f7-4db3-981d-cdd43f825371|6462149426|7956|2263154466|2178|2016-01-18T09:22:06.538+05:30|2016-01-18T09:22:06.538+05:30|SMS|0.55426353|ANSWERED
1e1f9f53-58d8-4b90-a234-c2ef2cae2e95|5846025857|1635|0328956567|9967|2016-03-05T05:40:46.010+05:30|2016-03-05T05:40:46.010+05:30|SMS|0.675355|ANSWERED
```

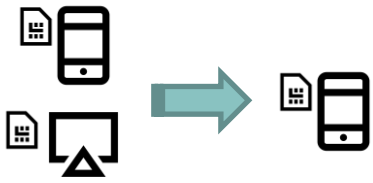
Each row of the above represents a voice call or SMS event. Other, passive events, such as network area updates, can also produce useable location data. However, these are significantly harder to work with. Data volumes are typically 2 orders of magnitude larger, and there is much more noise as a static device lying in an area covered by two or more cell signals can generate multiple handover events even though it is not moving. This means that processing requires more storage and computing power, and filtering steps have to be carefully tuned to the network topology.

The very first step is to irreversibly mask the identifying numbers of the caller and receiver by replacing them with a unique identifier. This is known as pseudonymisation, and is critical to protecting user privacy. Pseudonymisation will not remove the possibility altogether that someone could be identified from the dataset, but it does reduce the level of risk significantly throughout stages of processing in which the data cannot yet be aggregated or fully anonymised. Pseudonymisation is performed deep within operator systems to ensure that access to personal information is minimised.

Pseudonymised event data remains a long way from any real analytic power or value. It must undergo considerable processing to produce useable outputs. Some of these processing steps are described in the next section.

Location Data Processing

This section describes some necessary pre-processing steps required to obtain accurate location data.

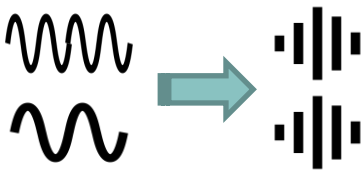


Filtering

Filtering refers to removing irrelevant records from the full set of data generated by the network. It requires detailed knowledge of the customer base as well as behaviours within the specific market. For example, in some markets it is common to share SIM cards between families or communities, whereas in others customers regularly use multiple SIMs. Some network systems create symmetrical data records for call termination, while others do not. Machine connections from IoT devices, which might generate many events, likely do not represent reliable mobility patterns and so need to be removed. Inactive SIMs and fraudulent usage need to be accounted for. Neglecting or mishandling this stage will produce inaccurate results.

Enrichment is the addition of other important data sources to the pseudonymised event data, to give context and meaning. This can include geospatial data describing network topology, operational data on infrastructure status, and system logs. The net effect is to add some structure to the data to make it more informative. Other external datasets are incorporated or layered later in the Analytics phase, but enrichment within operator systems is an important step.

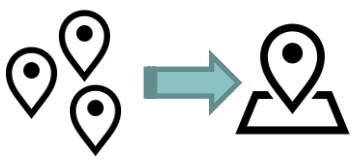
Enrichment



Regularisation

Regularisation is the process of mapping events to a unified framework in space and time. This requires selecting appropriate scales. If these scales are not selected correctly, important patterns will be obscured. For example, from a dataset that shows travel between regions from one week to the next it will not be possible to identify regular daily patterns, whereas a dataset of day-to-day travel will not show slower, long-distance travel. The type of event from which the data is derived sets lower-bound constraints due to market-specific usage patterns. Other relevant datasets will need to be mapped against later in the analysis, so the scales should be selected for compatibility. Official data is often based around administrative boundaries, but more some applications require more granular frameworks. The most appropriate scales for regularisation can only be identified once the question is understood and other data sources have been identified.

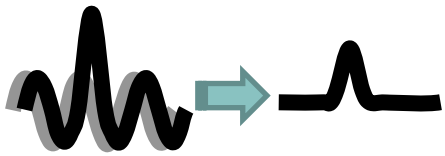
Location Assignment involves selecting or designing an algorithm to determine the actual location of a subscriber over a time period, based on the location events in that time period. Different algorithms will give different results, and the most appropriate solution will depend on the use case. For example, selecting the most frequent location during the hours of 9am –



Location assignment

5pm will give different results to using events between midnight and 4am. For some use-cases it may be necessary to count the number of visits made between areas, for others it is important to only count unique visitors, or to ensure that each person is counted only once. Some providers have sophisticated proprietary methods for detecting favourite locations or journeys. It is worth highlighting again here that the location data is not directly linked to the subscriber generating it, so privacy is well-protected.

Normalisation is comparison against some historical baseline. This helps establish the significance of a pattern. As an example, knowing that there are half a million people in the City of London borough is less informative than knowing that this is 10% less than on the average weekday, or two standard deviations less than the mean. As with other stages, the most appropriate normalisation approach will depend on the specifics of the question, such as relevant seasonal features or baselines.



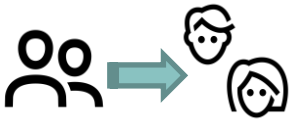
Normalisation

Validation is vital to ensure that the processed data reflects reality as accurately as possible. There are numerous possible sources of bias and inaccuracy, and it takes skill and an intimate understanding of the system that produced the data to mitigate these and validate the output. Operators are able to tune and refine their algorithms and establish appropriate scaling factors by checking outputs against a set of consenting users or the attendance at large public events. They can leverage network monitoring to ensure that outputs include all relevant data, and can use historical norms to establish expected bounds and flag potential errors.

Validation

Biases can emerge from the composition of the customer base compared with the population at large. This composition is best understood by the operators themselves.

In some markets it may be possible to separate data outputs along demographic lines, like gender, age, or income. This is known as "disaggregation" or "**Segmentation**", and depends very much on how customer data is collected, either by legal mandate or through business operations.

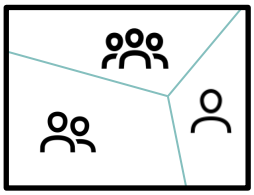


Segmentation

It is important to bear in mind that any of the above activities could result in a greater likelihood of individuals or groups of individuals being identifiable from the data. This could raise ethical or legal considerations that may vary from country to country, so an excellent understanding of the context is critical. It may be necessary to work with operators to ensure that they are able to produce a representative output or give an overview of the composition of the dataset as a whole.

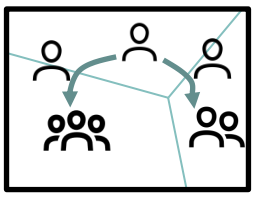
Aggregation is the process of collecting together individual trajectories, which are privacy-sensitive even in their pseudonymised state due because of the risk of reidentification, and building an indicator of an overall patterns. In many, but not all, cases, this an important pre-cursor to

analytics, and is a pre-requisite to sharing any data externally to operators' systems. There are two main types of aggregated indicator: occupancy, and mobility.



Occupancy

Occupancy refers to the number of people in an area at a specific time. This is useful for building population maps, or quantifying area risks in a disaster scenario.



Mobility

Mobility is the number of people moving between areas. This can be framed in many different ways to reveal patterns of habitual or occasional journeys, over different time scales and spatial resolutions. Applications include modelling the spread of infectious diseases and quantifying traffic levels.

Analytics

The analytics phase is the most difficult to discuss in general terms, because the techniques and components are very specific to the desired outcome. This section introduces some commonly used data sources, and describes at a high level some of the most important types of analysis and what they can do.

Data sources

Geography is an intuitive and critical anchoring point for most humanitarian and development work, and many relevant data sources have a strong geographical component.

Mapping data might include administrative boundaries, and Mobile Big Data outputs are often framed against these so that they can be aligned with other official statistics. Infrastructure and land usage maps can add useful pointers, and can improve the accuracy of journey analysis and population mapping. Population maps obtained from other sources can be used to calibrate estimates of mobility.

In epidemiological applications, the locations of medical cases are invaluable. This can be used to map the spread of specific diseases or disease strains, and when combined with mobility data allows prediction of future outbreaks.

Weather or climate data add an important dimension, for example when predicting the vulnerability to climatic extremes or predicting air quality.

Satellite imagery can provide a wealth of information to enrich other data sources. Image processing techniques have enabled settlement detection, car counting, and air quality and atmospheric measurements. These can be used to complement and calibrate more frequently updated data sources like mobile.

Techniques

Statistical analysis in the generic sense is the exploration of statistical properties within data. This might involve testing hypotheses, describing trends or searching for patterns.

Geospatial analysis is the analysis of data with a geographic or location component. As such it is central to many projects involving Mobile Big Data. The spatial relationships between points and regions impose special statistical considerations, and specialised systems (called GIS) exist to store and analyse the data structures that emerge.

Machine learning (ML) and **artificial intelligence** (AI) involve applying algorithms to historical data in order to make decisions or predictions. Rather than building a statistical model manually using domain knowledge and careful assumptions, ML automates the process of model-building by discovering patterns in relevant data. There are four broad categories of ML problem:

Classification is the assignment of a new data point to one or more discrete categories. Classification algorithms need to be trained on historical data with known classes. Predicting the gender of a mobile subscriber, detecting spam email, and identifying which customers are likely to leave in the near future are all examples of classification tasks.

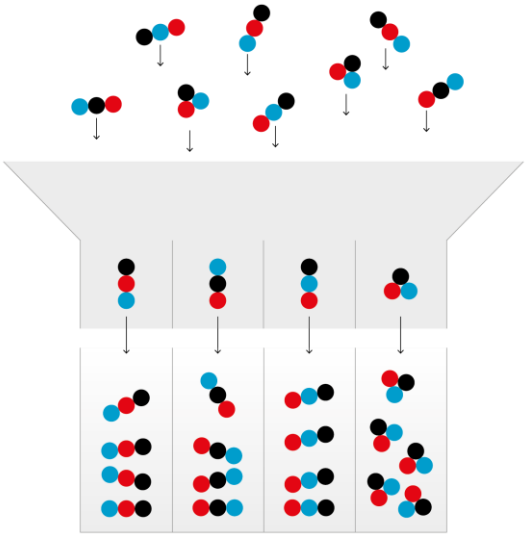


Figure 2 Classification

Regression is the prediction of a continuous quantity, like temperature or the level of air pollution. This can be used for fitting trends to make predictions about future events, or determining the strength of effect of one variable on another. Like classification, regression requires historical data from which to learn.

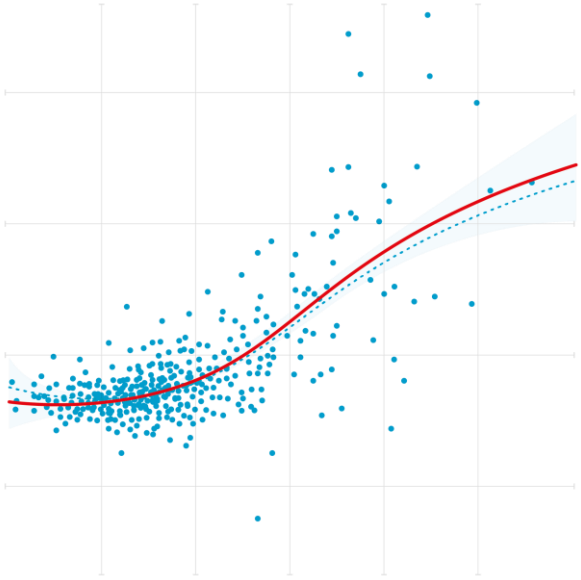


Figure 3 Regression

Clustering is the grouping of data into partitions according to common properties. Clustering algorithms are used when the structure of the data is not known. The task of the algorithm is to infer the structure in the data and divide it into logical groups. Examples are detecting topics in a set of documents, or identifying types of migratory movement pattern.

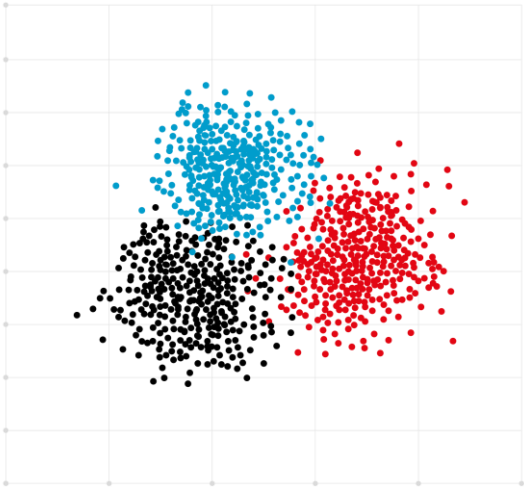


Figure 4 Clustering

Reinforcement learning is the optimal selection of a series of interactions with an environment, based on the predicted total benefit or optimal final outcome. An agent is fed information about its current state, and decides upon its next action. This action results in a reward, and a change to the state. Examples include piloting drones or self-driving cars, and playing chess.

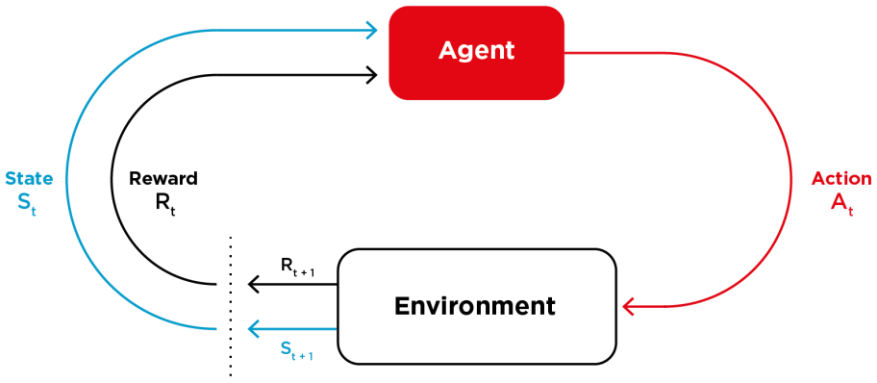


Figure 5 Reinforcement learning

These techniques can be very powerful, but also dangerous, as the automation can easily lead to the inclusion of hidden biases. Transparency and ethical considerations become especially important when using automated techniques.

References

For a thorough treatment of the application of machine learning in development, see https://www.usaid.gov/sites/default/files/documents/15396/AI-ML-in-Development.pdf

For a high-level introduction to machine learning and AI applications in the business world, see: https://www.mckinsey.com/~/media/McKinsey/Industries/Advanced%20Electronics/Our%20Insights/How%20artificial%20intelligence%20can%20deliver%20real%20value%20to%20companies/MGI-Artificial-Intelligence-Discussion-paper.ashx

Packaging

The processing steps described above are necessary precursors to a suitable and high-quality data product. What exactly that product should look like, how it should be published or presented, whether and how often it should be updated, what additional data it should incorporate, and what functionality it should offer depends on the end-user, the intended application, and ultimately what decisions or actions will be taken as a result.

It's important for all partners to work together to establish the requirements and ensure that the product will support the intended use. Mobile operators have numerous tools at their disposal for refinining, analysing and visualising data and will be able to help select and develop the most appropriate output. This might include one-off or regular reports or visualisations. Alternatively, an operator might offer a full analytic service, or build an internal application to support decision makers. The most appropriate option depends on the capacity and capabilities of the partners involved.

Key Requirements

When identifying the most appropriate output, it is important to consider the requirements. Some things to consider are:

- **Visual vs Quantitative**. Is the output intended to give an overview of a phenomena? Will it be necessary to drill down into the details of the phenomena?

- **One-off vs Updated**. Is the study historical, or single-use? Will it be necessary to obtain updated results as time progresses?

- **Static vs Interactive**. Does the end-user need to be able to interact with and adjust the output, perform functional queries, reframe the display? Is there a requirement for other, more sophisticated functionality? Is a combination of views required? What decisions will be made with it?

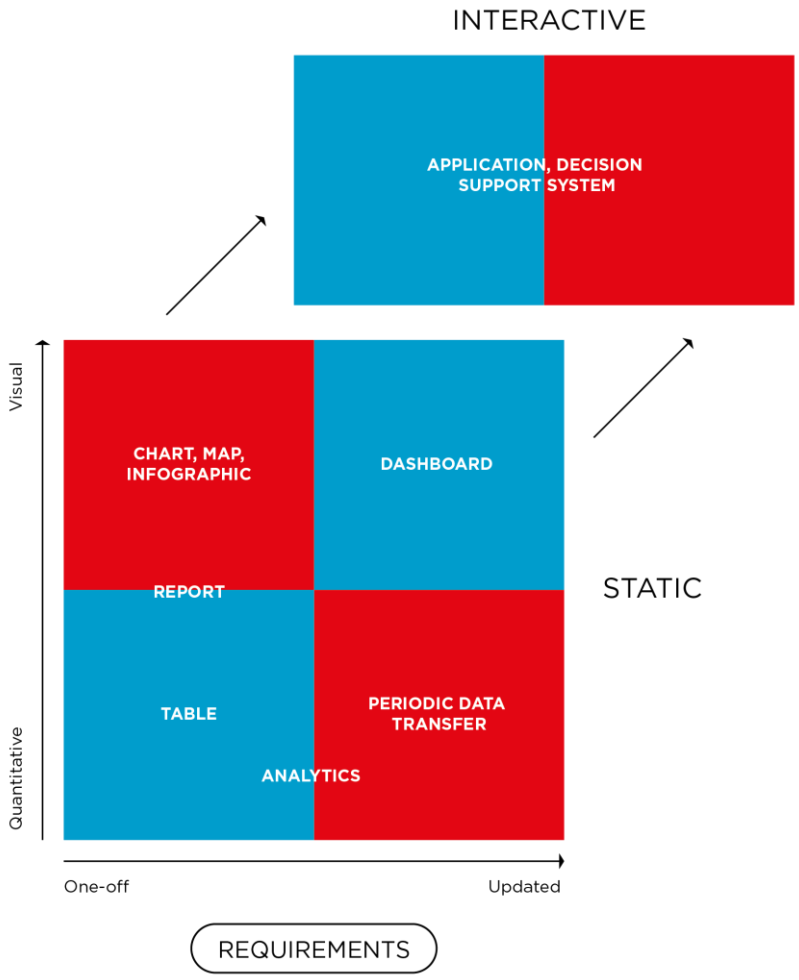


Figure 6 A set of possible outputs mapped against requirements

Conclusions

The preceding chapters present an overview of the steps involved to move from raw mobile network data to actionable insights that can help decision-makers to take relevant actions and realise impact. Mobile Big Data is a uniquely powerful input to advanced analytics models and decision support solutions, giving precise, up-to-date results, based on a large sample size.

However, the mobile data alone does not directly provide answers. Rather, it is a raw material that must be processed and shaped into a useful output. This requires skill, expertise, and facilities, as well as knowledge and understanding of the particular mobile network in question. Mobile network operators (MNOs) alone can fulfil this role.

Moreover, the output must be tailored to the use-case, to produce a solution that addresses the question or decisions to be made, and fits the specific requirements of the end-user. This should be done with scale and sustainability in mind, and with the utmost consideration and respect for the privacy and security of individuals and groups. Close working relationships between stakeholders on both supply and demand side are therefore key to successful projects.

Governments and agencies wishing to use Mobile Big Data could also benefit from building capacity and identifying relevant data sets internally, to fully capture the potential of data-driven solutions, and extend the possibilities for collaboration. Nurturing big data awareness and capabilities, allocating budget, and establishing partnerships will facilitate the unlocking of Mobile Big Data for the benefit of all.

Case Studies

This section discusses some case studies that have run under the BD4SG programme, and describes how the output was tailored to the use case. To read the complete case studies, visit our main website or click the links below.

Predicting air pollution levels 24-48 hours in advance in São Paulo, Brazil

Data and system parameters:

- **Mobile data set type**: Occupancy
- **Location algorithm**: traffic score derived from journey analysis
- **Spatial region**: 1 km square grid
- **Time period**: 10 months
- **Time window**: 1 hour
- **System update frequency**: No update in demonstration; production likely daily or hourly
- **Other datasets**: historical air pollution, weather, road maps
- **Analytical techniques**: machine learning (regression)
- **Output**: dashboard

Telefonica's analytics division LUCA used their internal Smart Steps platform to analyse journeys in the busy city of São Paulo. Their aim was to predict air quality at a number of sites across the city. To do this, LUCA had to quantify traffic levels in the areas surrounding the sites of interest. They used sophisticated algorithms to detect and split journeys according to dwell locations, and to assign these journeys to routes through the road network. They then aggregated these journeys to produce a mobility score over a 1 km square grid on an hourly basis. The resulting dataset was a form of occupancy, representing the number of people in motion in each of the grid squares. LUCA used this data to build a number of analytic models, and assembled the results into a dashboard for demonstrations. The demonstration dashboard is based on historical data from January-October 2017, but productised versions would likely be updated on a daily or hourly basis.

Accelerating the end of Tuberculosis in India

Data and system parameters:

- **Mobile data set type**: Mobility
- **Location algorithm**: predicted home and work locations, aggregated over space and time
- **Spatial region**: 1 km square grid; custom-defined catchment areas
- **Time period**: 2 weeks
- **Time window**: 2 weeks
- **System update frequency**: N/A
- **Other datasets**: TB case data, administrative boundaries, population maps
- **Analytical techniques**: statistical analysis, geospatial analysis
- **Output**: application

Airtel teamed up with the GSMA and the WHO and ITU's joint initiative Be He@lthy, Be Mobile, to explore how mobile data could offer relevant insights in the fight against TB. As this was a proof-of-concept, the requirements were for a one-off product, with high visual impact and interactivity.

The data selected was a form of mobility, connecting home and work locations over the same period. A relevant two-week period was selected to avoid overlapping with any major public holidays or other disruptive events. Two spatial aggregates were prepared. The first was based on a 1 km square grid. This was chosen to allow reasonable analytic flexibility whilst still protecting privacy. The other aggregate was designed to align with the other major dataset in the study, which was TB incidence data, giving the number of new cases per clinic. Spatial regions were designed around the locations of the clinics to represent catchment areas, and the mobile data was aggregated again at this level.

The application was based on a map. Interactivity included panning and zooming, and selection and toggling of different data layers and display modes. Although functional queries are not directly supported, because of the degree of interactivity, the mapping component, and the complexity of the underlying data model this output qualifies as an application.

Responding quickly and effectively to natural disasters in Japan

Data and system parameters:

- **Mobile data set type**: Occupancy
- **Location algorithm**: Last-known unique location, aggregated over space and time
- **Spatial region**: 250 m square grid
- **Time period**: variable
- **Time window**: Variable (depends on individual updates)
- **System update frequency**: 5 minutes
- **Other datasets**: infrastructure and topographical maps, points of interest, weather, IoT sensors, connected cars
- **Analytical techniques**: geospatial analysis, machine learning
- **Output**: application

In Japan, KDDI are working with cross-industry IoT partners to build an information platform for managing the response to disaster scenarios. GPS data from consenting users of KDDI's app gives the occupancy of threatened areas in near-real-time, allowing the emergency services to make informed and timely decisions. The output is an application, incorporating data visualisation with messaging, a report feed, and AI decision support. The data is updated frequently over the period covering the event.

Building communities resilient to climatic extremes

Data and system parameters:

- **Mobile data set type**: Individual trajectories
- **Location algorithm**: Daily model location
- **Spatial region**: Level 1 and 2 administrative regions (departments, municipalities)
- **Time period**: 1 year (2017)
- **Time window**: daily

- **System update frequency**: N/A
- **Other datasets**: administrative boundaries, climate data, hydric vulnerability, socioeconomic data (GDP, industries per region)
- **Analytical techniques**: geospatial analysis, machine learning (clustering)
- **Output**: report, visualisation

Telefónica are working with the United Nations Food and Agriculture Organisation (FAO) to identify patterns of migration and internal displacement due to climatic vulnerability, and to locate and quantify this otherwise silent phenomenon. This requires analysis of the locations of individual customers over long periods of time. Such a sensitive operation is only possible because Telefónica are able to conduct the analysis themselves, keeping data in-house and working closely with FAO to ensure that the interests of this vulnerable group is protected.

The project is based in Colombia. The first step in the analysis is to identify regions with high sensitivity to climatic variability, based on historic rainfall, topography, and industrial makeup. Regions with high reliance on farming or mining and low GDP are particularly vulnerable to the effects of drought. This involves geospatial and statistical analysis. Then filtering and clustering algorithms are applied to mobility data to identify groups of customers who left the affected regions during times of severe drought. The results are visualised and shared in the form of reports.

Bibliography

Alexander, L., Jiang, S., Murga, M., & González, M. C. (2015). Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, *58*, 240–250. https://doi.org/10.1016/j.trc.2015.02.018

Astarita, V., & Florian, M. (2001). The use of mobile phones in traffic management and control. *IEEE Intelligent Transportation Systems Conference Proceedings*, 10–15. https://doi.org/10.1109/ITSC.2001.948621

Barbosa-Filho, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., … Tomasini, M. (2017). Human Mobility: Models and Applications. Retrieved from http://arxiv.org/abs/1710.00004

Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., … Piarroux, R. (2015). Using Mobile Phone Data to Predict the Spatial Spread of Cholera. *Scientific Reports*, *5*, 1–5. https://doi.org/10.1038/srep08923

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., … Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, *111*(45), 15888–15893. https://doi.org/10.1073/pnas.1408439111

Eagle, N. (2010). Network Diversity and Economic Development. *Science*, *328*, 1029–1031. https://doi.org/10.1126/science.1186605

Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin-destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, *40*, 63–74. https://doi.org/10.1016/j.trc.2014.01.002

Lu, X., Wrathall, D. J., & Sundsøy, P. R. (2016). Detecting climate adaptation with mobile network data in Bangladesh: anomalies in communication , mobility and consumption patterns during cyclone Mahasen. *Climatic Change*, 505–519. https://doi.org/10.1007/s10584-016-1753-7

Naboulsi, D., Fiore, M., Ribot, S., & Stanica, R. (2016). Large-scale Mobile Traffic Analysis : a Survey.

Smith, C., Mashhadi, A., & Capra, L. (2013). Ubiquitous Sensing for Mapping Poverty in Developing Countries. *Proceedings of the 3rd International Conference on the Analysis of Mobile Phone Datasets*.

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2013). Quantifying the impact of human mobility on malaria, *338*(6104), 267–270. https://doi.org/10.1126/science.1223467.Quantifying