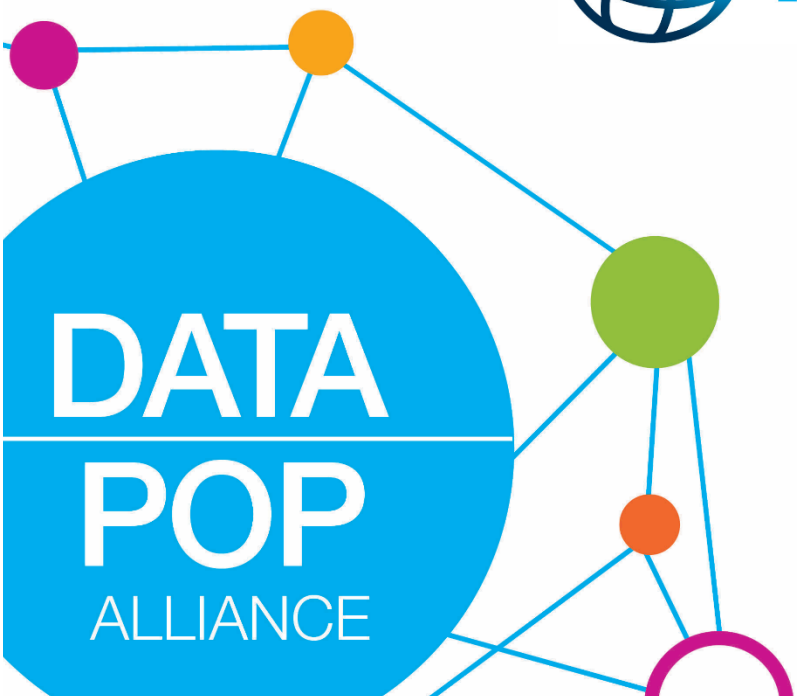
DATA-POP ALLIANCE AND WORLD BANK 'BIG DATA AND DEVELOPMENT' PRIMERS SERIES

BIG DATA AND MOBILITY —MIGRATION AND TRANSPORTATION

December 2015



WORLD BANK GROUP
Leadership, Learning and Innovation

HARVARD HUMANITARIAN INITIATIVE

mit media lab

ODI

About Data-Pop Alliance

Data-Pop Alliance is the first think-tank on Big Data and development, jointly created by the Harvard Humanitarian Initiative (HHI), the MIT Media Lab, and the Overseas Development Institute (ODI) to promote a people-centered Big Data revolution.

Acknowledgements

This paper is part of Data-Pop Alliance's Primers Series on Big Data and Development. The report benefited from financial support from the World Bank Leadership, Learning and Innovation group (Trevor Monroe and Adarsh Desai), which is gratefully acknowledged.

Its contributors are:

Ana Arieas
Bruno Lepri
Emmanuel Letouzé (Corresponding author: eletouze@datapopalliance.org),
Gabriel Pestre
Natalie Shoup

with additional contributions from:

Diego Canales Salas (World Bank)
Yves-Alexandre de Montjoye (MIT Media Lab and Data-Pop Alliance)
Patrick Vinck (Harvard University and Data-Pop Alliance)
Emilio Zagheni (U. Washington and Data-Pop Alliance)

It benefited from comments from Isabelle Huynh (World Bank).

The views presented in this paper are those of the authors and do not represent those of their institutions and partners.

DATA-POP ALLIANCE AND WORLD BANK 'BIG DATA AND DEVELOPMENT' PRIMERS SERIES

BIG DATA AND MOBILITY —MIGRATION AND TRANSPORTATION

Ana Areias
Bruno Lepri
Emmanuel Letouzé
Gabriel Pestre
Natalie Shoup

December 2015

1. Background and Rationale

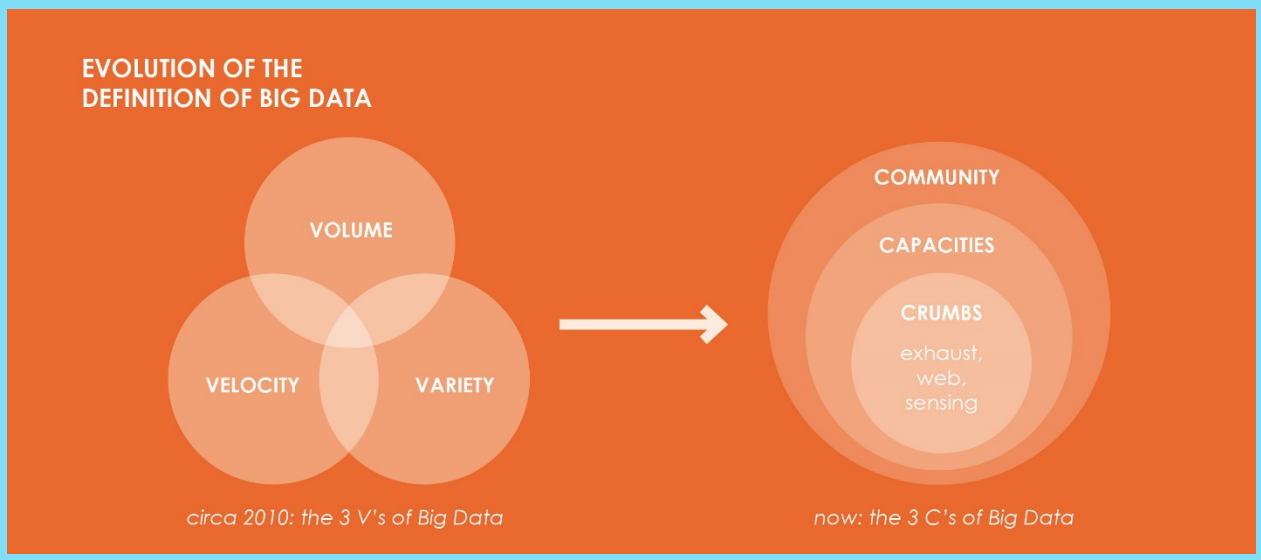
Big Data as an Ecosystem

This primer discusses the linkages between Big Data and mobility—specifically migration and transportation. Its main objective is to give its readers—World Bank staff, policymakers, researchers, development project managers, and other professionals—an overview of the main features and parameters of this nexus, as well as provide examples and discuss key considerations—technical, ethical, institutional, etc.—for developing projects, programs and other activities in the field.

In this primer, as well as in previous publications by the corresponding author, we understand Big Data as more than simply "big data" sources commonly characterized in the early 2010s by the 3 V's (Volume, Velocity and Variety); rather we conceptualize Big Data as an ecosystem which can instead be described using the "3 C's" (as illustrated in Figure 1). This ecosystem has at its core new sources of data, referred to as "digital crumbs,"[1] which come from sensor data (whether remote or physical), web-based data, and exhaust data. At the center of the Big Data ecosystem, these *crumbs* act as passively-emitted "digital translations of human actions and interactions."[2] Around these new data sources are new computing and analytical *capacities*, well exemplified by algorithms and other machine learning tools that are capable of "turning mess into decisions," but also enable visualization techniques. Finally, this ecosystem is animated by a Big Data *community*—individual and institutional producers, analysts, and people who make use of these crumbs through the aforementioned capacities.

This conceptualization of the 3 C's (Big Data crumbs, capacities, and community) helps better interpret the full applications and implications of Big Data, particularly in its illustration of the existence of feedback loops between its different parts (i.e. new data creating new capacities and new actors, and vice-versa, for example) and the interplay of complex dynamic processes with other sectors and actors. This contrasts with an exclusive focus on the "big data," which has led to a narrow and static techno-scientific approach. In this primer, Big Data refers to this ecosystem while big data refers to these new kinds of data (see Box 1).

Figure 1: Big Data's 3 Cs as an ecosystem



Source: Letouzé, 2015

Box 1: Big Data vs. big data

'Big Data' refers to the ecosystem created by the concomitant emergence of the 3 Cs of Big Data where:

- the 1st C stands for digital bread *Crumbs*—these pieces of data passively emitted and collected as by-products of people's interactions with and uses of digital devices that provide unique insights about their behaviors and beliefs;
- the 2nd C stands for Big Data *Capacities*—which has also been referred to Big Data Analytics, that is the set of tools and methods, hardware and software, know-how and skills, necessary to process and analyze these new kinds of data—including visualization techniques, statistical machine learning algorithms, etc.;
- the 3rd C stands for Big Data *Community* or *Communities*, which describe the various actors involved in the Big Data ecosystem, from the generators of data to their analysts and end-users—i.e. potentially the whole population.

This ecosystem can be described and analyzed as a *complex system*, i.e. one where feedback loops exist between its different parts. At the most basic level, new companies (e.g. Twitter or its future competitor) help generate new kinds of data that in turn lead to the development of new kinds of analytical tools, generating new kinds of data, attracting new actors to take advantage of these new data and tools. It is possible that this new ecosystem may turn into or become part of a larger social phenomenon.

In contrast, 'big data' refers to the 1st C above, i.e. the streams and sets resulting from humans leaving digital traces when using cellphones (call detail records), credit cards (transaction records), transportation (subway or bus records, EZ pass logs), social media and search engines, or having their actions picked up by sensors, whether physical (electrical meters, weight sensors on a truck) or remote (satellites, cameras).

Source: Letouzé, E. "Big Data and Development Overview Primer". Data-Pop Alliance, SciDev.Net and the World Bank (2015)

This conceptualization of Big Data helps us think more broadly and deeply about its interactions with and relevance for mobility—the movements of people, goods, and means of transportation—a field that has a long history of requiring and spurring quantitative investigations and innovations through modeling, optimization, and visualization. As will be discussed, the emergence of big data related to mobility has been highly significant given the paucity of data on human movement.

On the one hand, as will be discussed below, the vast majority of these movements leave basic machine-readable digital traces (crumbs) of their points of departure and arrival that are amenable to analysis, individually and collectively. Sensors can also yield additional data points—about speed, cargo weight, etc. Big Data, in turn, affects mobility, by suggesting better transportation routes for example. The use of Google Maps by individuals is another example of Big Data both shaping and being shaped by mobility patterns. These are usually referred to as uses or *applications* of Big Data—in the case of mobility as well as in other sectors such as public health, energy, socioeconomic analysis, and more—although it is clear that there is a bidirectional or, more accurately, complex relationship between Big Data and these sectors. As will be discussed, Big Data can also help shed light on the relationship between mobility trends and patterns and other data streams—for example, public health and/or socioeconomic data.

At least two taxonomies of these applications have been proposed in recent years, one in a 2012 UN Global Pulse report,[3] i.e.

(i) real-time awareness,
(ii) real-time feedback, and
(iii) early warning;

and another one in a 2013 report on Big Data and Conflict Prevention,[4] expanded upon in a 2015 synthesis report on Big Data for Climate Change and Disaster Resilience, [5] i.e.

(i) descriptive,
(ii) predictive (i.e. forecasting future indicators and processes or inferring current indicators and processes), and
(iii) prescriptive (i.e. establishing causal relationships between indicator and processes),

to which the authors of a 2015 synthesis report on Big Data for Climate Change and Disaster Resilience[6] add a fourth function:

(iv) discursive (i.e. spurring and shaping dialogue within and between communities and with key stakeholders).

The examples provided below will be discussed within the latter taxonomy, as appropriate.

Another dimension to consider, which is often overlooked, is Big Data's *implications,* which refer to the socio-political, institutional, methodological, legal, and ethical questions, considerations, and requirements that continue to arise as a result of the emergence of Big Data as an ecosystem.

It is Big Data's implications as much as its applications—which are of course not entirely independent—that have the potential to turn Big Data into a social phenomenon with deep and long-lasting consequences, in ways that have led a few observers to compare it to the various phases of the Industrial Revolution,[7] with data replacing coal or other sources of economic surplus and social value. In the case of the Big Data and mobility nexus, in an age where cellphones and other digital devices act notably as sensors of our movements but also of other behaviors and beliefs, privacy considerations are paramount, and will require devising new ethic-legal frameworks and principles.[8]

We now turn to discussing applications of Big Data, starting with a summary of relevant sources of big data (section 2), before discussing key implications (section 3), and relevance for current and future World Bank projects and programs (section 4).

2. Opportunities and Applications of Big Data for Mobility Analysis

The deficiency of existing mobility data

People (and their goods) have always been on the move. From city to city, from country to country, and across continents, human mobility—spanning from short-term seasonal movements to long-term migration—has remained a fundamental part of human life and social development. Globalization and new possibilities in transnational transportation and infrastructure during the past hundred years have increased the scope, scale, and frequency of these movements. What we describe in this primer as mobility represents the interplay between human movements across space and place, and the transportation infrastructure that facilitates these movements.

Perhaps more than any other aspect of human life, human mobility in particular has long been poorly understood due to the lack of appropriate data. For instance, to date, the best source of data on bilateral migration is the so-called 'Sussex Matrix', which provides very crude and at times questionable estimates of bilateral stocks of migrants around the world.[9] In some cases, in the absence of reliable data on flows, data from the cells of the Sussex Matrix have incorrectly been

interpreted as 'movements' of people. In addition to inconsistencies in indicators and a lack of interoperability across systems, traditional data sources on mobility, such as census information and surveys, are often expensive and limited in frequency, leading to inaccurate projections of movements and inefficient use of infrastructure. Despite advances in data on remittances, human trafficking, and stocks of migrants, and their attributes from two decades of migration data collection, the 2009 report by the Commission on Migration Data for Development Research highlights the lack of "detailed, comparable, disaggregated data on migrant stocks and flows" as the main obstacle preventing the "formulation of evidence based policies to maximize the benefits of migration for economic development around the world."[10]

The paucity of appropriate mobility data leads to both ineffective policies and inefficient use of resources. Gaps in existing data on migration often lead to misperceptions on migrants and their impact, making it difficult for policy-makers to develop and promote effective policies. The opportunity costs of such policies are high: almost 1 billion migrants—roughly 214 million international migrants and 740 million internal migrants worldwide—produce trillions of dollars in global economic impact from "the money migrants send home, the taxes they pay, the funds they invest, the trade they stimulate and the knowledge and technology transfer they stimulate."[11] Additionally, as opportunities for infrastructure expansion decreases in rapidly urbanizing environments, urban planners are looking for more continuous data collection strategies and capacities beyond traditional data sources in order to maximize efficiency of current infrastructure, relieve congestion, reduce environmental impact, and reflect the multi-modality of human mobility.

Sources of big data for mobility analysis

Big data as an ecosystem provides new sources of data, which can both fill the gaps in existing data repositories and provide new pathways towards information and insight. Furthermore, big data of various kinds—exhaust data, web-based data, sensing data—is becoming available at levels of granularity that were not possible before.

In the case of mobility analysis, data generated by devices and users can be grouped into the four following sources: websites and social media services; mobile devices; automatic data collection systems; and physical sensors (see Table 1). Collection occurs in both the public and private settings, as well as in the more complex intersection of the two—i.e. public Tweets or Foursquare check-ins—which raises the question of whether use of such data requires explicit consent from the people it pertains to.

Five areas of Big Data use in mobility

Current applications of Big Data for mobility analysis have generally fallen into the following categories:

1. Designing intelligent transport systems
2. Studying human mobility patterns
3. Tracking human mobility in crisis contexts
4. Using mobility data to study crime
5. Monitoring disease outbreaks

Designing intelligent transport systems

The wealth of data has unlocked the potential for intelligent transport systems that leverage communication technology to address and mitigate transportation problems. Applications range from 'descriptive'—i.e. gathering real time information on traffic volumes and congestion levels, to 'predictive'—e.g. predicting traffic conditions and travel time. Historical data can also be used for prescriptive applications, such as helping with urban planning, infrastructure planning, and transport system optimizations.[12]

Table 1: Sources of Big Data for mobility analysis			
Big data sources	**Systems**	**Data generators**	
		Public settings	**Private settings**
Websites and social media services	Cookies, log-ins	–	Log in data Twitter data Facebook data
Mobile devices	Global Positioning System (GPS)	Automatic vehicle location (AVL) on buses GPS fleet tracking on taxis	E-hailing of taxis Ridesharing Apps
		Floating vehicle data Dedicated probing	In-vehicle navigation systems (Tom Tom, Inrix, Garmin)
	Mobile phone networks	Tweets (Twitter API) Check-ins (Foursquare)	Call detail records (CDR) from mobile phones Geo-location data Tweets, Micro-blogs, Check-ins, and photos from smartphones
Automatic data collection systems	Fare collection systems	Public transport ticketing Smart Cards Contactless payments Road Electronic Tolling	Public transport ticketing Contactless payments
	Automated Passenger Counters (APC)	In-vehicle (buses) Stations (subway, BRT)	–
Physical Sensors	Network Sensors	CCTV Cameras Surveillance cameras Traffic cameras Automatic number plate recognition Bluetooth detection Loop detectors Environment sensors (temperature, noise, air quality) Parking Meters Rail track sensors	–
	In-vehicle sensors	Bus Monitoring Sensors Driving performance Fuel consumption Engine temperature	–

Source: Authors' elaboration

Box 1 provides examples of projects that approach these transportation issues from the data science standpoint and as such employ multiple tools, frameworks, and programming languages for tackling issues such as: fusing and integrating hundreds of gigabytes per day of data from multiple real-time feeds produced by emergency operation centers, sensors, CCTV cameras; or providing performance analysis by identifying patterns, relationships between variables, and historic responses to events, and presenting them in a compelling manner through visualization tools and user-interfaces for these large-scale systems. Cellphone and GPS data can also be used to study and raise awareness about individuals' environmental footprint (see Box 2).

Box 1: Intelligent transport systems

Predicting traffic in Singapore: By combining data from video cameras, GPS devices in taxis, and sensors embedded in streets, IBM, together with Singapore's Land Transport Authority have developed an hourly predictive model of traffic conditions geared towards city administrators and commuters. The model predicts traffic volume and speed for 500 locations up to 60 minutes in advance with accuracy ranging between 85-95%.[13]

Optimizing bus routes in Abidjan, Ivory Coast: Modeling mobility patterns based on 2.5 billion CDRs as part of the Orange D4D Challenge, IBM Ireland developed the AllAboard Project in which the team sought to maximize ridership and minimize waiting and travel time under budget and existing fleet constraints. The analysis yielded alternative transit routes that reduced journey times by 10%.[14]

Optimizing infrastructure investment in Moscow: The city of Moscow used home-work commuting patterns gleamed from CDRs to decide whether to expand the rail network. The city ultimately decided not to invest in railway expansion but rather redraw 100 bus routes. This solution allowed Moscow to forego more than $1 billion in upfront investment costs and reduced commute time by 3 minutes per trip, saving hours per year for commuters.[15]

Improving road transportation in the US: The Center for Advanced Transportation Technology Laboratory at the University of Maryland has been developing transport technology solutions and tools to enhance performance, reduce congestion, improve safety, and facilitate communications across the transportation industry for the US at the local, state, or national level.[16]

Optimizing taxi ride sharing in New York: Researchers at MIT Senseable City Lab, Cornell University and the Italian National Research Council's Institute for Informatics and Telematics present a new technique that enabled them to exhaustively analyze 150 million trip records collected from more than 13,000 New York City cabs over the course of a year. Moreover, they also introduced the novel concept of "shareability networks" that allows for efficient modeling and optimization of the trip-sharing opportunities.[17]

Box 2: Effect of human travel on the environment

Estimating the carbon emissions of Los Angeles, San Francisco, and New York City commuters: Researchers affiliated with AT&T Labs, used CDRs to calculate home-to-work travel distances and estimate the median carbon emission per home-to-work commute of hundreds of thousands of people living in the Los Angeles, San Francisco, and New York metropolitan areas. Their estimates showed that in New York, living further from the center was correlated with an increased carbon footprint since most people commute into central Manhattan for work. However, this is not the case for Los Angeles, since there is no single geographical concentration of jobs. The result for San Francisco is somewhere in the middle of the two cases.[18]

Inferring Gas Consumption and Pollution Emission of Vehicles throughout a City: A team of researchers from Microsoft Asia used GPS trajectories from taxicabs to infer the gas consumption and pollution emission of these vehicles in the city of Beijing.[19]

Studying human mobility patterns

A number of research initiatives and publications have demonstrated the opportunities for characterizing and analyzing human mobility patterns through Big Data, including those of regular city-dwellers, commuters, visitors and tourists. For example, the Real Time Rome Project, developed by the MIT Senseable City Lab in partnership with Telecom Italia, has helped dynamically monitor Rome's complex network of small streets by merging anonymized CDRs with other sources of real time information such as the locational data from Rome's taxis and buses. By creating several information layers over a map of Rome, the project helped identify where people were converging and which touristic landmarks were most attractive on a given day.[20]

Using 18 million Foursquare check-ins, a research team from Carnegie Mellon University working on the Livehoods Project determined which places got visited by the same people and grouped those venues into areas, drawing the boundaries of so-called "Livehoods" or areas of similar character within a city. If the set of visitors to a given location is largely the same as the set of visitors to another location, those two places will most likely be part of the same Livehood. By publishing interactive maps that reflect the current activity of its citizens it aims to reflect the social dynamics of a city.[21]

Marta Gonzalez has used CDRs from mobile phones to analyze commuting patterns in Boston and the Bay Area in order to identify areas of the city that contribute most to congestion. Her work has been a key pillar for the study and analysis of mapping commuting patterns, which is continuously evolving as more researchers are granted access to such datasets. The World Bank is now working with her to analyze CDR datasets from Rio de Janeiro and Mexico City for similar purposes.[22]

Mobility can also be studied on a wider geograhic scale. Twitter data, for instance, has been used by the Mexican ministry of tourism to track the number of tourists travelling to certain parts of the country on a holiday weekend (see Box 3).

With a more long-term perspective**,** international migration rates and patterns have recently been studied using e-mail data for 43 million anonymous *Yahoo!* account holders between September 2009 and June 2011. Researchers at Qatar Computing Research Institute (QCRI) and the Max Planck Institute for Demographic Research (MPIDR) inferred location of email senders through the user's IP address. Estimates of migration flows were adjusted for the fact that some subsets of the population, such as older people, are underrepresented in email flows. This approach could be used to complement official migration data, which is often outdated and inconsistent.[23] Again, Emilio Zagheni and a team of researchers from QCRI and Stanford University used geo-located Twitter data to infer internal and international migration patterns.[24]

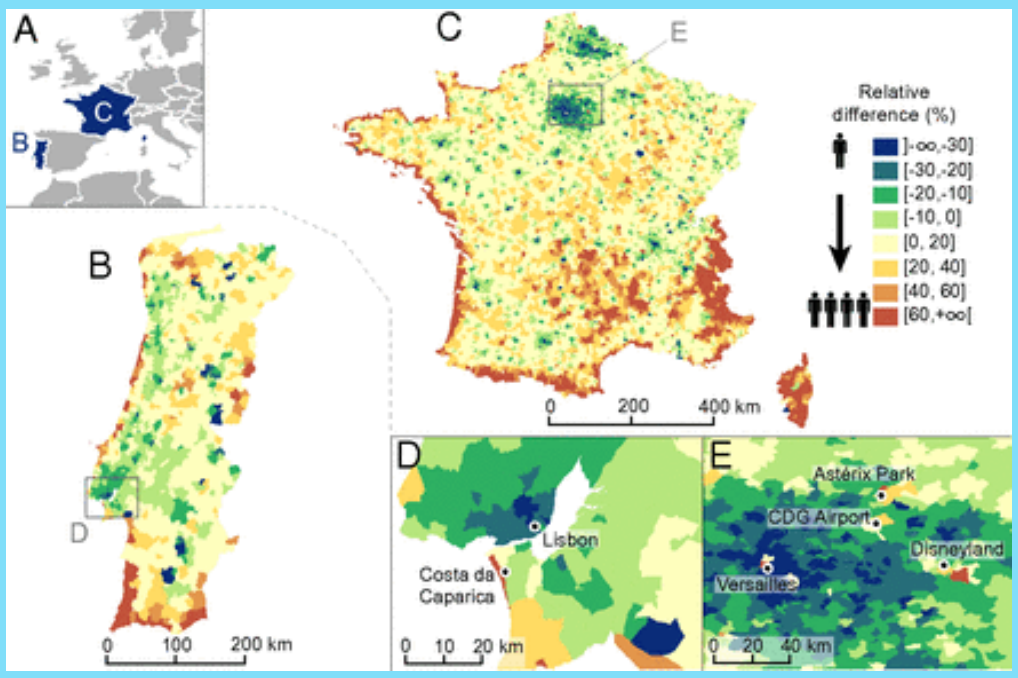
Box 3: Twitter for Tourism Monitoring in Mexico

In 2014, a working group on Big Data at INEGI conducted a pilot study to track domestic tourism from Twitter data, in order to contribute to the empirical modelling of individual tourist behavior. The objective of this pilot program was to identify the characteristics of an average Tweeting tourist in order to identify how many people travelled to Puebla and Guanajuato during the holiday weekend of February 1-3, 2014. The team of researchers from INEGI, in collaboration with the Mexican Ministry of Tourism, analysed 60 million Tweets from January to July 2014, from the continuous 1% georeferenced sample that Twitter makes available for free.[25] From this data, INEGI collected Tweets from the 7,955 Twitter users who Tweeted in Guanajuato (48%) and Puebla (52%) during the holiday. They then gathered all the Tweets sent by those users in the remainder of the target period (amounting to 827,424 total Tweets), and identified which users Tweeted from another state (presumably their homestate) after being in Guanajuato or Puebla, in order to map the origin of domestic tourism to those two areas during the holiday.[26] The resulting estimates of domestic tourism to Guanajuato and Puebla were compared to estimates made by the respective offices of tourism of those two states.

Source: Manske, Julia, et al. (2015) "Opportunities and Requirements for Leveraging Big Data for Official Statistics and the Sustainable Development Goals in Latin America."

Recently, Deville et al.[27] used CDRs from 3 major operators to study seasonal mobility patterns in France and Portugal, and compared their estimates of the resulting population densities against official census data (see Figure 2).

Figure 2: Seasonal changes in population distribution in Portugal and France



Source: Deville et al (2014), PNAS, vol. 111 no. 45

Tracking human mobility in crisis contexts

Big data has also been used to study population movements in crisis contexts. Important data sources include CDR data, Twitter data, and user account and IP address information from email providers.

In general, the focus of projects that use Big Data to track mobility in relation to crisis has been to complement official statistics and develop real-time indicators. Many pre-existing data sources, such as censes, are designed to track long term trends over larger areas (such as countries or administrative regions), and therefore lack the spatial and temporal granularity necessary to help predict or study the effects of more isolated occurrences like natural disasters, epidemics, or terrorist attacks.

Flowminder analyzed mobile phone location information provided by Digicel for all calls placed in the months before and after the 2010 Haiti earthquake and subsequent cholera outbreak to study trends of population movement after the disaster to inform relief assistance efforts. They estimated that, 19 days post-earthquake, 20% of the Port-au-Prince population left the capital[28]. The results were found to be consistent with a subsequent retrospective UN population-based survey. Mobility patterns around an identified cholera outbreak area were available and helped identify areas at risk of outbreak. Movement of people before and after floods is also visible in cellphone data (see Box 4).

Box 4: Changes in cellphone activity during the Tabasco floods

After the 2009 floods in the Mexican state of Tabasco, Pastor-Escuredo et al. used CDRs to show that flooding was visibly linked to detectable and unusual mobile phone activity in and around that area. By constructing a baseline of call activity for the affected areas using data from the weeks prior to the flood announcement, then measuring changes in location and duration of calls after the first predictions of heavy rainfall, and after severe flooding began, they were able to study how people moved around in preparation for the potential floods and in response to the actual floods. These reconstructions of the flood's impact were validated against the assessment of the flood area from Landsat-7 images, as well as figure kept by the state's authorities on the number of displaced people.[29]

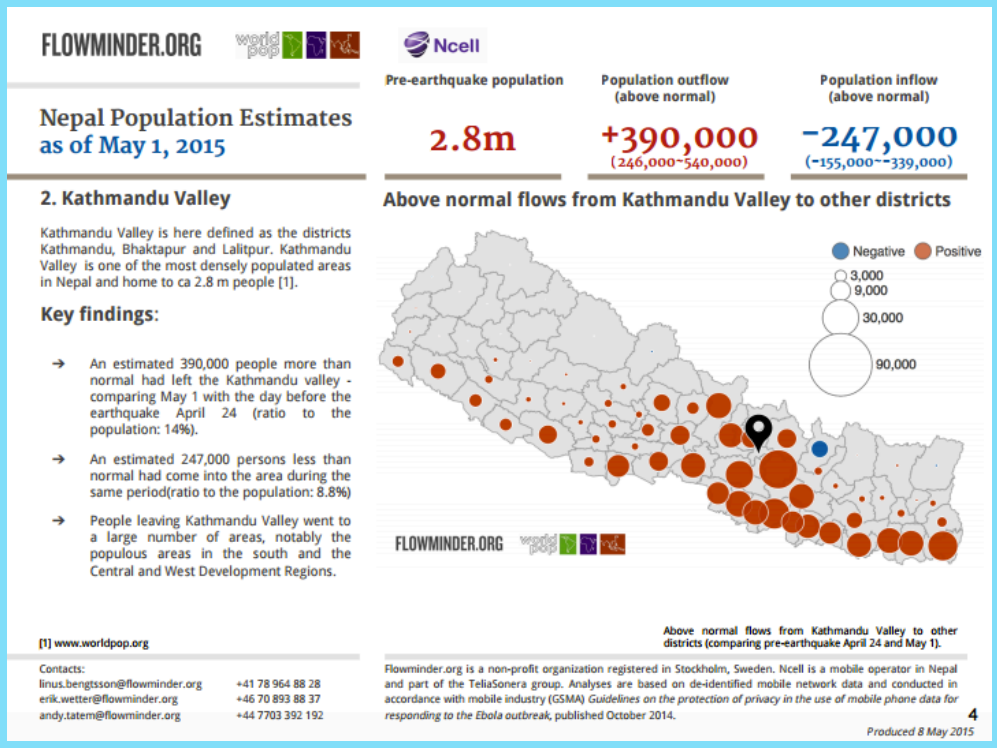
New Zealand has also used cellphone data to monitor population movements following the February 2011 earthquake in Christchurch City.[30] Flowminder has also produced estimates of population densities resulting from movements following the recent earthquake in Nepal (see Figure 3)—although their accuracy is unknown given the absence of ground truth data.

Using mobility data to study crime

A joint team from Fondazione Bruno Kessler, MIT Human Dynamics group and Telecom Italia won the "Datathon for Social Good" competition in London, organized by Telefonica and Open Data Institute (ODI) at Campus party 2013 with a project entitled "Predicting Next Month Crime from Mobile Network Activity."[31] In this project, researchers presented a novel approach able to predict crime levels in a geographic space from mobile phone data. Specifically, the team of researchers used mobile phone data to get insights on the presence and on the hourly mobility patterns of specific classes of subjects (e.g. young vs. old, residents vs. tourists) and then they use this information as predictors.

A follow-up of this project will be launched in Bogotá in collaboration with World Bank, Telefonica, DANE and with a focus not only on the mobility detected by mobile phone data but also the mobility detected by transportation data.

Figure 3: Movement in Nepal after the 2015 earthquake



Source: Flowminder

Monitoring disease outbreaks

Another well-know application is epidemiology; notably the quantification of the impact of mobility on malaria spread. A group of researchers from Harvard, Carnegie Mellon, University of Oxford used mobile phone data to analyze the travel patterns of nearly 15 million individuals over the course of a year in Kenya, combining these data with a simple transmission model of malaria based on highly spatially resolved malaria infection prevalence data to map routes of parasite dispersal. Their analysis identified specific importation routes that contribute to malaria epidemiology on regional spatial scales. Importantly, their model attempted to correct for sample selection bias.

3. Implications and Requirements of Big Data use

Data accessibility

Most of the data that the selected examples have been based on is collected by private entities and is not readily accessible by researchers. This means that it is necessary to develop data-sharing standards, agreements and infrastructure to share data, especially when it comes to facilitating access to mobile phone data while limiting privacy risks. A number of organizations are working on devising such arrangements, notably a tripartite project of the Leiden University, the World Economic Forum and NYU's GovLab.[32]

Recently, research challenges that provide access to a large number of teams to the same dataset are becoming a truly valuable framework to advance the Big Data field (especially, the fields based on using CDRs). Examples are offered by Orange's "Data for Development" (D4D) initiative in 2013 and 2014-2015 and by Telecom Italia Big Data Challenge 2014 and 2015.

However, in specific contexts the difficulty in accessing data is still evident; e.g. Ebola outbreak. The outbreak took place in one of the most highly connected and densely populated regions of Africa and hence accurate information on population movements was of paramount relevance for monitoring the progression of the outbreak, predicting the spread, and facilitating the design of interventions and surveillance and containment strategies. Vital questions include how the affected regions are connected by population flows, which areas are major mobility hubs, what types of movement typologies exist in the region, and how all of these factors are changing as people react to the outbreak and movement restrictions are put in place. Just a decade ago, obtaining detailed data to answer such questions over this huge region would have been impossible.

Today, such valuable data exist and are collected in real-time. However, during the Ebola outbreak they remain unused—in part because the appropriate legal and ethical frameworks and institutional processes are not in place, and need to be built in a way that is sustainable and takes into account a wide range of factors, notably the risk of Big Data becoming an extractive industry.[33]

Ethical and political aspects

One of the greatest risks with Big Data is the notion that it could provide exogenous, techno-scientific and technocratic solutions to some of the world's most intractable problems—the now old notion of Big Data being "the new oil". A good example are 'smart cities,' which some commentators worry will not be participatory: in the words of Anthony Townsend, director of the Institute of the Future and author of *Smart Cities: Big Data, Civic Hackers, and the Quest for a New Utopia*, "Some people want to fine tune a city like you do a race car but they are leaving citizens out of the process."[34]

More generally, Big Data applications must also seek the right balance between access to data and preserving user privacy. Location data in particular, such as mobility traces from CDRs, are unique enough that coarsening them spatially and temporally is not enough to guarantee anonymity[35].

Group privacy is also likely to become a major concern; even when individual privacy is protected, i.e. where identification is near or effectively impossible, it may be possible to track the movements of groups and types of individuals—for example to identify skin color and height and thus age in camera surveillance data. In such cases, the use and value of the insights derived from the resulting analysis must be very high to justify the inherent and potential harms and detrimental affects for these communities and neighborhoods.[36]

Box 5: Tackling the Privacy Challenges of Location Data

Tackling the Privacy Challenges of Location Data: While mobile phone data has great potential for good, its use raises privacy issues that need to be addressed. The lack of names, home addresses, phone numbers or other obvious identifiers does not make a dataset anonymous. Indeed, recent research[37] compromised the privacy of mobile phone location data by showing that as few as four spatio-temporal points are enough to re-identify 95% of individuals in a dataset of 1.5M people. The same paper also shows that, contrarily to popular beliefs, even coarsened or noisy location datasets provide little anonymity. Moreover, several studies have shown that people evaluate as highly risky sharing data on their locations. Recently, some researchers found that location is the most valued piece of information in a study designed to investigate the monetary value that people assign to their personal information as it is collected by their mobile phone.

Acknowledging these risks and developing an appropriate regulatory framework is essential to making mobile phone data broadly available and used. While there are no known methods (and will probably never be[38]) to de-identify individual location data with complete certainty, promising data-driven approaches to make re-identification significantly harder have begun to develop. These range from sampling the data or limiting its longitudinality[39] up to mobile phone operators sharing "SafeAnswers"[40] such as behavioral indicators or summary statistics with third-parties.

Analytical hurdles

Last are basic but severe analytical issues, starting with the sample selection bias. Differential mobile phone ownership and internet access among age groups, socio-economic strata and geographical areas mean that mobility analyses that rely on CDR data may generate biased estimates that are not representative of the general population and are only applicable on subsets of the population.

A paper relying on Kenyan data survey data as 'ground truth' has shown that mobility estimates derived from CDRs are surprisingly robust to the substantial biases in phone ownership across different geographical and socioeconomic groups[41]. In general, however sample selection bias can compromise the validity of results if it is not properly accounted for. Sample (or sampling) bias refers to a situation where data is drawn from a subset of a population where different members of the population do not have the same likelihood of being represented and where these differences affect their behaviors—in other words different members of the population exhibit characteristics that affect both their behavior and their likelihood of being represented in the sample—such that any inference that can be made within the sample at hand may not be generalized outside of the sample.[42] In most cases indeed differences in individual features such as income, age, gender, etc., are correlated with differences in outcomes (preferences or behaviors). A major source of sample selection bias is self-selection into using a given device or service, such that it is particularly important to develop methods for understanding and correcting the associated biases to ensure that outcomes of models relying on these inputs are useable.

As mentioned, early work on sample selection bias correction with Big Data streams has been done by Zagheni and Weber using data provided by *Yahoo!* about email user accounts. In their 2012 paper *"You are where you E-mail"* [43], they proposed a method for studying human migration patterns based on geographic information for a large sample of *Yahoo!* e-mail messages, self-reported demographic information of *Yahoo!* users, migration rates for 11 European countries gathered by Eurostat from national statistical agencies, and international statistics on Internet penetration rates by age and gender. Based on IP address, they determine the country from which a user sends the most emails, then study how that location changes over time across all users.

Other correction methodologies for accounting for biases in CDRs use have also been proposed and research is on-going in this area.[44] A proposed model to correct for sample selection bias in estimating population density and size that can be applied to estimations of movements uses cellphone penetration from Demographics and Health (DHS) surveys as a predictor of the bias (see Box 6).

Box 6: Accounting for sample selection bias using mobile phone penetration rates in Senegal

Using call detail records (CDRs) from Senegal provided by the telecommunications company Orange in the context of the Data for Development (D4D) Challenge 2014, Letouzé, Zagheni and Pestre propose a methodology for estimating population density at the *arrondissement* level based on the number of callers from each location.[45] The CDRs give the user ID, location of the caller and the exact timestamp of each call made in the 2013 calendar year for a random sample of about 150,000 subscribers. Based on these trajectories, it is possible to observe how the number of users in each location changes from month to month and from weekday to weekend.

However, in order to generalize these results to the population as a whole, it is necessary to account for differences in cellphone ownership between locations. Mobile phone penetration rates are expected to be uniformly high in developed countries, but there may be greater heterogeneity within and across developing countries, and factors such as income can affect how and how much a given cellphone user makes calls or sends texts.

The authors therefore rescale their estimates using measures of cellphone ownership obtained from the 2012-2013 Demographics and Health (DHS) survey for Senegal, which covers a time period fairly similar to that of the CDRs. The goal of their analysis is to understand which groups and regions are under- or over-represented in the sample, and then to inflate or deflate the population estimates according, thus demonstrating the possibility of drawing valid estimates from non-representative samples.

Another avenue is to use machine learning to predict population movement from ground truth and big data sources but the downside is the lack of external validity over time and space of such approaches.

These analytical challenges are especially salient and relevant since the social sector there is a shortage of individuals with the specific capabilities necessary for these analytical positions. Big Data applications demand interdisciplinary teams working together to analyze data, including computer scientists, statisticians, social scientists, urban planners, etc. More generally, there is a huge dearth of understanding and basic awareness around Big Data (its conceptualization and capacities) that will need to be filled if Big Data is to have a positive impact in the area of mobility as in others.

4. The World Bank and the Big Data-mobility nexus: projects and prospects

Ongoing World Bank projects on Big Data use

The World Bank's transport strategy aims to ensure safe, clean and affordable transport for all. The primary areas of focus are: urban transport logistics; environmental footprint of the sector, infrastructure; greener, more efficient and cost-effective transportation; and safety of informal transport.

How can Big Data contribute to these goals? The Bank is actively working on various initiatives utilizing big data, where the underlying motivation to do so is encased on substituting data that is expensive to collect or that we were not able to collect before this 'revolution.'

The following is a non-exhaustive selection of relevant Big Data projects the World Bank is actively pursuing:

Developing origin-destination surveys from cellphone data to improve transportation planning and optimizing public systems in Mexico City and Rio de Janeiro:

This project aims to (1) develop efficient, precise, and low-cost strategic transport planning tools, using new sources of massive information from cellphone data, (2) provide a reasonable substitute or complement to traditional urban transportation data sources (e.g., Origin-Destination travel surveys), and (3) streamline an internal methodology within Bank-related transport projects in need of travel surveys and related transport planning tools to produce data and tools at a much lower cost and within a shorter lifecycle.

Using Big Data to improve freight transportation flows and environmental sustainability of supply chains in Indonesia:

The World Bank and the MIT (logistics lab) propose to apply Big Data to the problem of freight transport. In simple terms the CDR trace of a truck is distinctively different from that of a pedestrian or a taxi driver to be automatically classified. Hence, such critical transport flows can be estimated from Big Data, instead of from costly field surveys that cannot be replicated on a regular basis. This approach is potentially a game changer in applied transport economics, and could help improve the performance and the environmental sustainability of supply chains, especially in congested environments such as port cities.

Providing transparency and accountability in the transport sector using open transport data in Sao Paulo, Brazil:

With vendor lock-in with proprietary systems and lack of access to bus fleet automatic vehicle location (AVL) data, Brazil's transport sector considerably lacks transparency. This project was able to (1) grant access to General Transit Feed Specification (GTFS) schedules and Automatic Vehicle Location (AVL) data to citizens through a hackathon event, (2) set up an API[46] for Olho Vivo, their real-time bus location service, for developers to consume and create apps from it, (3) analyze historical data and produce performance indicators using business intelligence to respond to citizens' protests and clamor by providing transparency and accountability in the transport sector, and (4) allow the creation of a new Urban Mobility Lab, which won the 2014 Enterprising City/State Award, 'MobiPrize'[47] presented at the World Congress on Intelligent Transport Systems event in Detroit, for 'taking bold decisions" to change institutional culture (proprietary data and formal procurement processes) in exchange for open standards and protocols, non-proprietary technologies and open data in order to create solutions that address the city's congestion challenges and its lagging bus system.

Using automatic vehicle location (AVL) and fare card data to create a bus performance dashboard in Sao Paulo, Brazil:

In response to a lack of performance indicators and operational oversight of the concessions managing the bus fleet, research using AVL and fare card data is underway towards the creation of a bus performance dashboard. The dashboard aims to provide transparency on bus location, route optimization, and safety, as well as commuter transit use, demand bottlenecks, and feedback on performance.

Understanding the relationship between urban infrastructure and crime in Bogotá, Colombia:

This project aims to understand the relationships between urban infrastructure characteristics and six different types of crime (Homicide, Assaults, Theft to persons, Automobile thefts, Motorcycle theft, Residential property burglaries, Commercial property burglaries). The research aims to (1) map crime in space-time and correlating it with urban characteristics, land use, city equipment, and social variables, (2) evaluate the effect of the development of Bus Rapid Transit System routes on crime, and (3) evaluate survey-collected social variables and BRT use and its effect on crime.

Strategic Entry-Points

The World Bank is uniquely positioned to play a significant role in shaping the future of the Big Data and mobility nexus via different entry points and along different lines.

First is certainly the research and methodological aspects of the issue; more work could be done or funded in the area of estimations and sample selection bias correction, which may require the development and implementation or funding of survey to collect ground truth data.

Research can and should be done on par with innovation and pilots, but the latter need to more systematically include a learning component—i.e. incorporating feedback loops and going beyond proofs of concepts in order to build methods that may allow reproducibility and scaling up. The post-2015 agenda and the inclusion of variables and targets related to human mobility and climate sustainability provide a political window for doing work in this area.

A third entry point is the Bank's lending mechanisms that give its leverage to advance objectives in the area of Big Data and mobility—by fostering projects that have a greater potential to create a healthy environment for Big Data to make a positive contribution to traffic decongestion for instance, informed by sound ethical principles.

In addition, the Bank can play a role in weighing in on these ethical principles and legal frameworks through knowledge products and convening, including highlighting the risks and challenges of using Big Data for tracking human mobility as much as its potential; in that respect the SDG agenda also offers a good political window.

The Bank can play an essential role in training, including capacity building and awareness raising; as mentioned Big Data is not only a technical issue; the technical requirements are very complex and require investments in physical and human capital that are beyond the possibilities of even the biggest donors alone; the Bank could play an active convening role in that area.

Lastly, the Bank can lead in convening within and between various constituencies of the Big Data ecosystem including civil society organizations, governments and private sector companies around common and principles to shape the future of Big Data using mobility as a fertile training ground.

Works Cited

Anand, Komal. "Enterprising City/State MobiPrize Winner for 2014." Mobi Platform. 7 April 2015. Avialable at: http://mobi-platform.com/drum-roll-please-enterprising-citystate-mobiprize-winner-for-2014-is/

Anderson, Chris. *The Long Tail: Why the Future of Business is Selling Less of More.* Hachette Books, 2008.

Anttila-Hughes, Jesse; Dumas, Marion; Jones, Lindsey; Pestre, Gabriel; Qiu, Yue; Levy, Marc; Letouzé, Emmanuel; Sala, Simon; Schwarz, Bessie; Shoup, Natalie; Tellman, Elizabeth; Vinck, Patrick. "Big Data for Climate Change and Disaster Resilience: Realising the Benefits for Developing Countries." September 2015. Available at: http://datapopalliance.org/item/dfid-big-data-for-resilience-synthesis-report/

Becker, Richard; Cáceres, Ramón; Hanson, Karrie; Isaacman, Sibren; Loh, Ji Meng; Martonosi, Margaret; Rowland, James; Urbanek, Simon; Varshavsky, Alexander; and Volinsky, Chris. "Human Mobility Characterization from Cellular Network Data." *Communications of the ACM*, vol. 56, no. 1 (January 2013): 74-82.

Bengtsson, Linus; Lu, Xin; Thorson, Anna; Garfield, Richard, and von Schreeb, Johan. "Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti." *PLOS Medicine* (30 August 2011). Available at: http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001083

Center for Advanced Transportation Technology. "Research." Available at: http://www.catt.umd.edu/research

Commission on International Migration Data for Development Research and Policy. "Migrants Count: Five Steps Toward Better Migration Data." 25 May 2009. Available at: http://www.cgdev.org/publication/migrants-count-five-steps-toward-better-migration-data

Cranshaw, Justin; Schwartz, Raz; Hong, Jason I.; and Sadeh, Norman. "The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City." *The 6th International AAAI Conference on Weblogs and Social Media* (Dublin, Ireland, 2012). Available at: http://livehoods.org/research

De Montjoye, Yves-Alexandre; Hidalgo, César A.; Verleysen, Michel; and Blondel, Vincent D. "Unique in the Crowd: The privacy bounds of human mobility." *Scientific Reports*, vol. 3, no. 1376 (25 March 2013). Available at: http://www.nature.com/articles/srep01376

De Montjoye, Yves-Alexandre; Kendall, Jake; and Kerry, Cameron F. "Enabling Humanitarian Use of Mobile Phone Data." *Center for Technology Innovation at Brookings*. 12 November 2014. Available at: http://www.brookings.edu/research/papers/2014/11/12-enabling-humanitarian-use-mobile-phone-data

De Montjoye, Yves-Alexandre; Shmueli, Erez; Wang, Samuel S.; and Pentland, Alex 'Sandy'. "openPDS: Protecting the Privacy of Metadata through SafeAnswers." PLos one, vol. 9, no. 7 (9 July 2014). Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098790

De Montjoye, Yves-Alexandre; Smoreda, Zbigniew; Trinquart, Romain; Ziemlicki, Cezary; and Blondel, Vincent D. "D4D-Senegal: The Second Mobile Phone Data for Development Challenge." 31 July 2014. Available at: http://arxiv.org/abs/1407.4885

Deville, Pierre; Linard, Catherine; Martin, Samuel; Gilbert, Marius; Stevens, Forrest R.; Gaughan, Andrea E.; Blondel, Vincent D.; and Tatem, Andrew J. "Dynamic population mapping using mobile phone data." *Proceedings of the National Academy of Sciences*, vol. 111, no. 45 (11 November 2014): 15888–15893. Available at: http://www.pnas.org/content/111/45/15888

Flowminder. "Nepal Population Estimates (and movements) as of 01-May-2015." Humanitarian Response. 2015. Available at: https://www.humanitarianresponse.info/en/operations/nepal/document/flowminder-nepal-population-estimates-and-movements-01-may-2015

Fiadino, Pierdomenico; Valerio, Danilo; Ricciato, Fabio; and Hummel, Karin. "Steps towards the Extraction of Vehicular Mobility Patterns from 3G Signaling Data." *Proceedings of the 4th International Workshop, Traffic Monitoring and Analysis 2012*, Vienna, Austria, 12 March 2012: 66-80. Available at: http://tma2012.ftw.at/TMA/papers/TMA2012paper8.pdf

Fondazione Bruno Kessler. "Predicting Next Month Crime from Mobile Network Activity – 'Datathon for Social Good' competition in London." 16 September 2013. Available at: http://www.fbk.eu/news/predicting-next-month-crime-mobile-network-activity-datathon-social-good-competition-london

Hellerstein, Joseph. "The Commoditization of Massive Data Analysis." *O'Reilly Radar*. 19 November 2008. Available at: http://radar.oreilly.com/2008/11/the-commoditization-of-massive.html

IBM Research. "AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data." 2013. Available at: http://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=4746

IBM Research. "IBM and Singapore's Land Transport Authority Pilot Innovative Traffic Prediction Tool." 1 August 2007. Available at: http://www-03.ibm.com/press/us/en/pressrelease/21971.wss

International Data Responsibility Group. "Data Governance Project." Available at: http://www.responsible-data.org/data-governance-project.html

Letouzé, Emmanuel; Meier, Patrick; and Vinck, Patrick. "Big Data for Conflict Prevention: New Oil and Old Fires," in "New Technology and the Prevention of Violence and Conflict." International Peace Institute, 2013: 4-27.

Letouzé, Emmanuel, and Vinck, Patrick. "The Law, Politics and Ethics of Cell Phone Data Analytics." April 2015. Available at: http://datapopalliance.org/item/white-paper-the-law-politics-and-ethics-of-cell-phone-data-analytics/

Letouzé, Emmanuel; Zagheni, Emilo; and Pestre, Gabriel. "Correcting for Sample Bias with Application to the Case of Senegal" *World Development Report 2016: Digital Dividends* (forthcoming).

Lichtle Fragoso, Pedro Manuel; and Sánchez Salinas, Juan Carlos. "Uso Productivo de Big Data y Redes Sociales en el Sector Turismo." Secretaría de Turismo (Sectur) de México. *Documentos de Investigación Estadística y Económica*, no. 2014-1 (October 2014). Available at: http://www.datatur.beta.sectur.gob.mx/Documentos%20Publicaciones/2014_1_DocInvs.pdf

Manske, Julia; Sangokoya, David; Barrett, Lauren; Pestre, Gabriel; and Letouzé, Emmanuel. "Opportunities and Requirements for Leveraging Big Data for Official Statistics and the Sustainable Development Goals in Latin America." 2015. Available at:

http://caribbean.eclac.org/content/opportunities-and-requirements-leveraging-big-data-official-statistics-and-sustainable

Max Planck Institute for Demographic Research (MPIDR). "You are where you email. Global migration trends discovered in e-mail data." 25 June 2012. Available at: http://www.demogr.mpg.de/en/news_press/press_releases_1916/you_are_where_you_e_mail_global_migration_trends_discovered_in_e_mail_data_2939.htm

Narayanan, Arvind; and Felten, Edward W. "No silver bullet: De-identification still doesn't work." 9 July 2014. Available at: http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf

Pastor-Escuredo, David; Morales-Guzmán, Alfredo; Torres-Fernández, Yolanda; Bauer, Jean-Martin; Wadhwa, Amit; Castro-Correa, Carlos; Romanoff, Liudmyla; Lee, Jong Gun; Rutherford, Alex; Frias-Martinez, Vanessa; Oliver, Nuria; Frias-Martinez, Enrique; and Luengo-Oroz, Miguel. "Flooding through the lens of mobile phone activity." *IEEE Global Humanitarian Technology Conference (GHTC), 2014*: 279-286. Available at http://arxiv.org/abs/1411.6574

Pentland, Alex "Sandy". "Reinventing Society in the Wake of Big Data: A Conversation with Alex (Sandy) Pentland." *Edge*, 30 August 2012. Available at: https://edge.org/conversation/reinventing-society-in-the-wake-of-big-data

Sanders, Robert. "Cellphone, GPS data suggest new strategy for alleviating traffic tie-ups." *Berkeley News*, 20 December 2012. Available at: http://newscenter.berkeley.edu/2012/12/20/cellphone-gps-data-suggest-new-strategy-for-alleviating-traffic-tie-ups/

Santia, Paolo; Resta, Giovanni; Szell, Michael; Sobolevsky, Stanislav; Strogatz, Steven H.; and Ratti, Carlo. "Quantifying the benefits of vehicle pooling with shareability networks." Proceedings of the National Academy of Sciences, vol. 111, no. 37 (16 September 2014): 13290–13294. Available at: http://www.pnas.org/content/111/37/13290

São Paulo Transporte (SPTrans). "API Olho Vivo." Available at: http://www.sptrans.com.br/desenvolvedores/APIOlhoVivo.aspx

Shang, Jingbo; Zheng, Yu; Tong, Wenzhu; Chang, Eric; and Yu, Yong. "Inferring Gas Consumption and Pollution Emission of Vehicles throughout a City." *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (New York City, 24-27 August 2014): 1027-1036. Available at: http://research.microsoft.com/apps/pubs/?id=217455

Statistics New Zealand. *Using cellphone data to measure population movements*. Wellington: Statistics New Zealand, 2012. Available at: www.stats.govt.nz/~/media/Statistics/services/earthquake-info/using-cellphone-data-measure-pop-movement.pdf

Sutherland, Peter D. "Migration is development: How migration matters to the post-2015 debate." *Migration and Development*, vol. 2, no. 2 (December 2013): 151-156.

Transportation Research Board of the National Academies. "Open Data: Challenges and Opportunities for Transit Agencies." *Transit Cooperative Research Program (TCRP), Synthesis 115* (2015). Available at: http://onlinepubs.trb.org/Onlinepubs/tcrp/tcrp_syn_115.pdf

UN Global Pulse. "Big Data for Development: A Primer." June 2013. Available at: http://unglobalpulse.org/sites/default/files/Primer%202013_FINAL%20FOR%20PRINT.pdf

UNECE. "Mexico (INEGI) - Tweet Analysis." *Big Data Inventory*. Updated 20 March 2015. Available at: http://www1.unece.org/stat/platform/x/OBeDBg

Wakefield, Jane. "Tomorrow's cities: Do you want to live in a smart city?" *BBC News*, Technology. 19 August 2015. Available at: http://www.bbc.com/news/technology-22538561

Wesolowski, Amy; Eagle, Nathan; Noor, Abdisalan M.; Snow, Robert W.; and Buckee, Caroline O. "The Impact of Biases in Mobile Phone Ownership on Estimates of Human Mobility." *Journal of the Royal Society Interface*, vol. 10, no. 81 (April 2013). Available at: http://rsif.royalsocietypublishing.org/content/10/81/20120986

Wikipedia "Selection bias." Last updated 28 October 2015. Available at: https://en.wikipedia.org/wiki/Selection_bias

World Bank. "Bilateral Migration and Remittances datasets." *Data & Research Prospects*. Available at: http://go.worldbank.org/JITC7NYTT0

Zagheni, Emilio; Garimella, Venkata Rama Kiran; Weber, Ingmar; and State, Bogdan. "Inferring International and Internal Migration Patterns from Twitter Data." *Proceedings of the 23rd international conference on World wide web* (Seoul, Korea, 7-11 April 2014): 439-444. Available at: http://ingmarweber.de/wp-content/uploads/2014/02/Inferring-International-and-Internal-Migration-Patterns-from-Twitter-Data.pdf

Zagheni, Emilio; and Weber, Ingmar. "You are where you E-mail: Using E-mail Data to Estimate International Migration Rates" *Proceedings of the 4th Annual ACM Web Science Conference* (Evanston, IL, 22-24 June 2012): 348-351. Available at: http://download.repubblica.it/pdf/2012/tecnologia/email.pdf

Endnotes

[1] Pentland, 2012; and Letouzé, 2012, 2013, 2014 and 2015.
[2] Letouzé et al, 2013.
[3] Letouzé, 2012.
[4] Letouzé et al., 2013.
[5] Letouzé et al., 2015
[6] Antilla-Hughes et al., 2015.
[7] Hellerstein, 2008; and Anderson, 2008.
[8] Letouzé & Vinck, 2015.
[9] World Bank, "Bilateral Migration and Remittances datasets."
[10] Commission on International Migration Data for Development Research and Policy, 2009.
[11] Sutherland, 2013.
[12] Fiadino et al., 2012.
[13] IBM Research, 2007.
[14] IBM Research, 2013.
[15] Transportation Research Board of the National Academies, 2015.
[16] Center for Advanced Transportation Technology.
[17] Santia et al., 2014.
[18] Becker et at., 2013.
[19] Shang et al., 2014.
[20] Calabrese et al., 2010.
[21] Cranshaw, 2012.
[22] Sanders, 2012.
[23] Max Planck Institute for Demographic Research, 2012.
[24] Zagheni et al, 2014.
[25] UNECE, 2015.
[26] Lichtle Fragoso & Sánchez Salinas, 2014.
[27] Deville et al., 2014.
[28] Bengtsson, 2011.
[29] Pastor-Escuredo et al., 2014.
[30] Statistics New Zealand, 2012.
[31] Fondazione Bruno Kessler, 2013.
[32] International Data Responsibility Group; see also: UN Global Pulse, 2013.
[33] Letouzé & Vinck, 2015; de Montjoye, Kendall, et al., 2014.
[34] Wakefield, 2015.
[35] De Montjoye et al., 2013.
[36] Letouzé & Vinck, 2015.
[37] De Montjoye et al., 2013.
[38] Narayanan & Felten, 2014.
[39] De Montjoye, Smoreda, et al., 2014.
[40] De Montjoye, Shmueli, et al. 2014.
[41] Wesolowski, et al., 2013.
[42] "Sampling bias is systematic error due to a non-random sample of a population, causing some members of the population to be less likely to be included than others, resulting in a biased sample, defined as a statistical sample of a population (or non-human factors) in which all participants are not equally balanced or objectively represented. [Sampling bias] undermines the external validity of a test (the ability of its results to be generalized to the rest of the population)." Wikipedia, "Selection bias," accessed 11 December 2015: https://en.wikipedia.org/wiki/Selection_bias
[43] Zagheni & Weber, 2012.
[44] Letouzé, Zagheni, et al.
[45] Letouzé, Zagheni, et al.
[46] São Paulo Transporte (SPTrans).
[47] Anand 2015.