DATA POP ALLIANCE

WHITE PAPERS SERIES

THE LAW, POLITICS AND ETHICS OF CELL PHONE DATA ANALYTICS

Written by
Emmanuel Letouzé
Patrick Vinck
With Lanah Kammourieh

April 2015

HARVARD HUMANITARIAN INITIATIVE

mit media lab

ODI

IN PARTNERSHIP AND WITH FUNDING FROM

WORLD BANK GROUP
Leadership, Learning and Innovation

ABOUT DATA-POP ALLIANCE

Data-Pop Alliance is a research, policy and capacity building coalition on Big Data and development, 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.

ABOUT THE AUTHORS

Emmanuel Letouzé and Patrick Vinck contributed equally to the writing of this paper and are its authors. Lanah Kammourieh provided contribution on the legal framework.
Emmanuel Letouzé is the Director and co-Founder of Data-Pop Alliance. He is a Fellow at the Harvard Humanitarian Initiative, a Visiting Scholar at MIT Media Lab, a Senior Research Associate at ODI, a PhD Candidate (ABD) at UC Berkeley, and the author of UN Global Pulse's White Paper "Big Data for Development: Challenges and Opportunities" (2012).
Patrick Vinck is the co-Director and co-Founder of Data-Pop Alliance. He is Director of the Peacebuilding and Human Rights Data Program at the Harvard Humanitarian Initiative. He holds appointments at the Brigham and women's Hospital, Harvard School of Public Health, and Harvard Medical School. He also serves as a member on the Committee on Scientific Freedom and Responsibility of the American Association for the Advancement of Science.
Lanah Kammourieh is a lawyer, scholar, and journalist specialized in the law of new technologies and international law. She is a Research Affiliate of D ata-Pop Alliance.

ACKNOWLEDGEMENT

This paper is part of Data-Pop Alliance's 'White Papers Series'. It originated in discussions between members of the 'D4D team' involved in the first Orange D4D challenge in 2013—including representatives of Orange France Telecom Group, World Economic Forum USA, Université catholique de Louvain, MIT, UN Global Pulse, GSMA, and the authors of this paper. The report benefited from financial support from the World Bank Leadership, Learning and Innovation group, and the Harvard Humanitarian Initiative, which are gratefully acknowledged.
The views presented in this paper are those of the authors and do not represent those of their institutions and partners. This version benefited from suggestions and ideas from members of the D4D team, notably William Hoffman, Nicolas de Cordes, Kyla Reid, and Ed Barker, as well as Data-Pop Alliance directors and affiliates, notably Alex 'Sandy' Pentland, Yves-Alexandre de Montjoye, Nuria Oliver, Gabriel Pestre and Natalie Shoup. It underwent a peer review process at the World Bank led by Isabelle Huynh, who also provided extensive comments, under the authority of Randeep Sudan. Formal reviews were provided by Cordula Rastogi, Tim Kelly, and Bruno Lepri. Other comments from Alla Morrison, David Satola, Masatake Yamamichi, Jos Berens, Gabriel Pestre and Natalie Shoup were received.

Design by Liora Yuklea (Data-Pop Alliance)

Letouzé E, Vinck P. "The Law, Politics and Ethics of Cell Phone Data Analytics." Data-Pop Alliance White Paper Series. Data-Pop Alliance, World Bank Group, Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute. April 2015.

DATA–POP ALLIANCE
WHITE PAPERS SERIES

THE LAW, POLITICS AND ETHICS OF CELL PHONE DATA ANALYTICS

April 2015

Written by
Emmanuel Letouzé
Patrick Vinck
With Lanah Kammourieh

Data-Pop Alliance
The World Bank Group

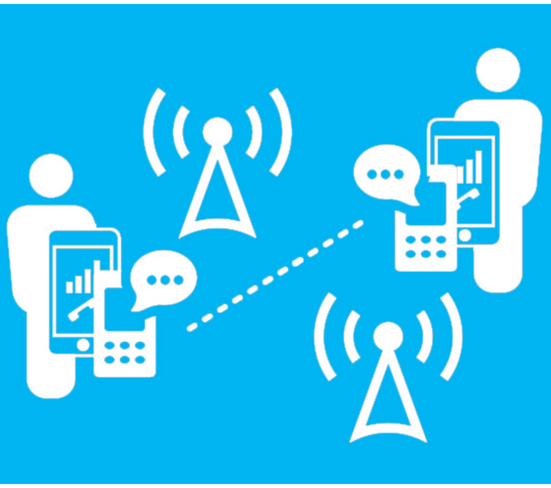
TABLE OF CONTENTS

INTRODUCTION

Mobile phones—including smartphones—are spreading across the world, possibly faster than any other technology.[1] Cell phones also increasingly drive Internet penetration, especially in low-income countries.[2] Global mobile data traffic is expected to multiply by more than 100 between 2009 and 2017. As users connect to mobile networks (and landlines), they generate data about each transaction. One particularly useful piece of data is contained in Call Detail Records (CDRs)—a record that includes among other data, the starting time of the call (or message), its duration, the originating and receiving phone numbers. Most mobile data can be geo-located thought various means.[3] CDRs carry the approximate location of the users based on cell phone towers activated, which sometimes can be made even more precise based on tower triangulation and wifi network connections. Large mobile phone service providers (hereafter telephone companies, or 'Telcos') may handle over 6 billion CDRs a day, a number growing exponentially and unlikely to abate in the foreseeable future.[4] While landlines also produce CDRs, the focus of this white paper is on mobile phone CDRs. Mobile phone subscription outweighs landline subscriptions by 2:1, and is greater than 50:1 in Africa, for example.[5]

EXAMPLE OF SELECTED CALL DETAIL RECORDS DATA

- Unique record identifier
- Pseudonymized phone number of caller
- Pseudonymized phone number of receiver
- Starting date and time, duration
- Caller cell-tower location
- Receiver cell-tower location
- Call type (voice, SMS, etc.)



CDRs are essential for billing, monitoring voice and data usage, and understanding and targeting customers based on their cell phone consumption patterns.[6] They may, for example, be used to target promotional events based on service subscription, or identify areas where investments in services are needed. Increasingly, however, CDRs are also recognized for the insight they provide into human behavior, movements, and social interactions. Proposed applications include mobility analysis to monitor movements ranging from daily commutes to forced displacement, social network analysis generating better understanding of socio-economic ties and interactions, or economic analysis as a proxy of economic activity and poverty levels. Academics and practitioners have called for the release of more data to ever growing numbers of researchers to leverage the technology and show the value that can be extracted from it—announcing, for example, that such data would help stop the spread of epidemics like Ebola.[7] CDRs can also be combined with other data sets, creating promising avenues for research and policy. Yet the benefits of CDR analytics remain largely hypothetical, and their potential remains largely locked and under-explored for privacy, technical, commercial, and ethical reasons.

This white paper examines ways to enhance the responsible sharing and use of CDRs, creating the right environment to explore their real potential for social good. Much of the discussion may apply to other forms of mobile data (e.g. generated through sensors and applications installed on smartphones, or from the use of social media). This paper, however, focuses specifically on CDRs and their analysis as a rapidly growing field of practice that relies on one of the largest, richest, and most dynamic sets of data available.

Sharing CDRs as individual and aggregated records is certainly difficult, for technical but also primarily legal and 'institutional' reasons. Telcos and governments are cautious about privacy and security implications of CDR analytics, and about public perception in the post-Snowden era. Telcos may also be concerned about granting access to commercially valuable data, as well as costs associated with data anonymization and access control.

CONSTRAINTS ON DATA SHARING

According to the 2015 World Economic Forum's whitepaper on 'Pathways for Progress":

"The pervasive culture of not sharing data retards development. The private sector's reluctance to share data is due largely to a utilitarian calculus of proprietary and competitive concerns that pervade market-based economies. They are not alone. Even within the UN system or among non-profits, a proprietary default position can make it difficult to get agencies to share programme data. Many concerns about sharing data are based on a lack of trust, a fear of incurring liabilities or a loss of institutional information control and arbitrage advantages (which create and maintain power differentials both within and between organizations)."

Source: 'Pathways for Progress'; WEF Data-Driven Development whitepaper, page 9

The concept of 'data philanthropy'[8] emerged around 2011 as an argument and modality for sharing and accessing proprietary data, including in the context of humanitarian crises.[9] But the assumptions behind the concept—the idea that the data collected by Telcos are "their data"— may be at odds with the notion that CDRs are essentially people's own data. Ethical considerations to guide CDR analytics are no less challenging. As the proposed use of CDR expands well beyond their original purpose, what are the choices given to the users? Do they understand and agree to what can be done with their data? Do they stand to reasonably benefit from CDR analytics, or are the benefits held by Telcos only? These are just a few of the critical questions that must be answered as CDR analytics quickly expands in response to the growing availability of data and progress in data warehousing, management, and computing, notably advances in algorithmic analysis. These questions must be addressed to avoid CDR analytics becoming an extractive industry subject to elite capture and creating new threats to personal and group privacy and security.

This white paper seeks to contribute to these questions that undermine efforts at generating societal benefits from CDR analytics. It explores the key considerations that must inform and frame current discussions and attempts to craft/sketch the legal, technical, and institutional architecture of CDR analytics as a growing field of practice. Importantly, it approaches the issue with an ethics lens, and considers the legal implications of the proposed ethical principles whilst bearing in mind political considerations. After providing additional contextual elements (Part 1), the paper summarizes current legal frameworks (part 2) before exploring structural socio-political parameters and incentives structuring the sharing of CDRs (part 3), proposing guiding ethical principles (Part 4) and discussing operational options and requirements (part 5).

1 CONTEXTUAL ELEMENTS

1.1 THREE TYPES OF ANALYTICS

The analysis of CDRs has drawn significant attention in the past few years as part of a broader interest in extracting social value from the analysis of "traces of human actions picked up by *digital devices*"[10] —or Big Data's applications and implications for society[11] and development.[12] Initial applications have ranged from health, crime, finance, and banking, to marketing and advertising. As discussed further below, major Telcos themselves have contributed to the phenomenon since at least 2012, when Orange, a French multinational telecommunications corporation operating mainly in Europe and Africa, along with several partners, organized the first 'Data for Development Challenge' (D4D).[13] Other groups and organizations in academia and the non-profit sector, notably Flowminder, which pioneered the analysis of CDR analytics in Haiti after the 2010 earthquake, have also generated a large body of literature and evidence on the potential of CDR analytics to yield unique behavioral 'insights' on human ecosystems.[14]

The fast growing space of 'CDR analytics for social good' takes advantage of how mobile carriers see the world. With CDRs, it has become possible to 'follow' and map the movements, actions, and interactions of an individual—or, rather, of a phone or SIM card—to look for patterns and trends in the data, especially in conjunction with other datasets, and attempt to model, understand, and affect human ecosystems.

A simple way to think about how CDR analytics can be leveraged for development and programming purposes is to distinguish three main types—using a taxonomy proposed for Big Data more generally[15]:

1. One is a *descriptive* function—via maps, descriptive statistics etc.—this may include, for example, the visualization of migration roads and movement patterns or the spread of epidemics;

2. Another is a *predictive* function, probabilistic by definition, in two senses of the term:
a. The first sense refers to predicting as inferring, or 'proxying', where CDR-based variables are used instead of another variable. One example is the use of CDRs as a proxy for measures of socio-economic levels;
b. The second sense is 'forecasting' where the goal is to assess the likelihood of some event(s) in a near or distant future. his may include for example applications related to early warning systems, which look at patterns associated with past events, such as an epidemic, to estimate the likelihood of a new event such as a new epidemic. Another example is to build on past forced population movement to forecast future displacement routes in case of a new crisis.

3. The third and least developed to date, is a prescriptive function, in which the predictive function of CDR is enhanced to examine the possible consequences of different choices of action, resulting in recommendations on the best course of action. The prescriptive function can be associated with 'future analysis' building on simulation, game theory and decision-analysis methods. In practice, this may lead, for example, to examining multiple likely patterns of forced displacement under various conditions to assist policy choices.

These types of CDR analytics have been used to study the spread of infectious diseases, internal migration and mobility, spatial dynamics in urban slums, reciprocal giving in the aftermath of a natural disaster, poverty, socioeconomic levels, and transportation.[16] Other areas of applicability have also been studied and discussed—including conflict and crime.[17]

While often lauded as revolutionary, Big Data efforts have had mixed success. Google Flu Trends (relying on Google search, not CDRs) is a well-known example of an application initially presented as holding the potential to make public health systems irrelevant, before it turned out to greatly overestimate actual flu cases[18] for complex reasons that have now been analyzed in depth.[19] The story of Google Flu Trends and other such initiatives show the importance for models and research to evolve and adapt, which in turn requires the ability to test and experiment using both historical and current data.

CDRs released to date, however, represent only a tiny fraction of those that exist and many more opportunities exist to use CDRs in a responsible way. This paper seeks to contribute to the discussion on how to make this possible.

1.2 THE EMERGENCE OF A FIELD OF PRACTICE

The idea for this paper came out of the 2012-13 D4D 'Ivory Coast' Challenge. For that challenge, researchers were given controlled access to four datasets derived from anonymized CDRs of phone calls and SMS messages between 5 million Orange customers in Côte d'Ivoire from December 1, 2011 to April 28, 2012. This was followed by a second challenge in 2014-15. The first challenge was covered extensively in the press.[20] The four datasets included aggregate data about antenna-to-antenna traffic on an hourly basis, fine resolution movement information about 50,000 phone numbers over a two week period, coarse resolution movement information about 500,000 phone numbers for the entire period, and communication subgraphs for 5,000 customers.

Several measures were taken to guarantee the anonymity of the data and control access to it. A three-tier anonymization process was designed based on feedback from French and British academic "Friendly Test teams" who tried to crack the data. This included:

a) the generation of random caller IDs for each dataset;
b) slight blurring (by 5%) of all 1,200 antenna positions and
c) the selected exclusion of off-network communications, first calls, and data from 'extreme' users who may be too easy to identify.

Access was controlled through:
(i) requirement of a project submission to the D4D Challenge signed by a representative of a research institution;
(i) commitment to follow Terms & Conditions controlling the use of data and the publication of results.

2012-13 'IVORY COAST' D4D: DATASETS MADE AVAILABLE UNDER CONTROLLED ACCESS

1

AGGREGATE DATA
Antenna-to-antenna traffic on an hourly basis for the entire period.

2

FINE RESOLUTION MOBILITY TRACES
Individual trajectories for 50,000 customers for two-week time windows with antenna location information.

3

COARSE RESOLUTION MOBILITY TRACES
Individual trajectories for 500,000 customers over the entire observation period with sub-prefecture location information.

4

COMMUNICATION SUB-GRAPHS
For 5,000 random customers up to 2 degrees of separation aggregated by two-week time window over five months.

Organizers were overwhelmed by the response. As such, the D4D revealed and spurred huge interest in the emerging field of CDR analytics. A total of 250 teams submitted proposals and received the data, of which 83 submitted papers. It culminated at the Third NetMob Conference with a 1-day event at MIT in May 2013.[21] The challenge generated papers on a wide range of submissions addressing questions about migration, poverty, public health, urban development and transportation, crisis response, demographic and economic statistics, and more. Four winners were identified, and 30 teams were granted permission to keep the data for further collaborative work. The success of the first edition led Orange to organize a second challenge in 2014-15, focusing on Senegal.[22]

Other data challenges organized by Telcos merit mention. One is Telefónica's 'Datathon for Social Good' co-organized with the Open Data Institute and MIT in September 2013 in London as part of the Campus Party Europe; another is Telecom Italia's Big Data Challenge.[23] Telefónica also worked with the Government of Mexico, the World Food Program and United Nations Global Pulse on understanding the value of cell tower activity to detect flooding in Mexico.[24]

Telefónica has also undertaken several research projects and initiatives 'in-house—i.e. without the data being released; for example, its researchers used aggregated anonymized CDRs to analyze which of three public policy interventions was most effective at curbing population movement in Mexico City during the H1N1 epidemic; in another 2011 paper using 2010 data from a "main city in Latin America" (Mexico City), the research team used CDRs in conjunction with survey data on socio-economic levels (SELs) to build a predictive model to 'assign' SELs to various areas on the basis of their digital signatures in CDRs—with a predictive power of 80%. The rationale for building such models is to either help estimate changes in SELs over time, or apply them to other locations.

Another attempt was that of The World Bank in Egypt. In 2012, the World Bank partnered with Vodafone and IBM Research after the 'Cairo Transport App Challenge' to analyze Cairo's traffic congestion. Under the initial terms of the agreement, Vodaphone Vodafone was to release historical anonymized CDRs, while the Dublin-based IBM research team would use its AllAboard solution (developed for the D4D Challenge mentioned above) to conduct the analysis. The project was put to a halt and eventually died after the

National Telecom Regulatory Authority made requests that the partners were unable or unwilling to meet—including the conditions that CDRs stayed on the Egyptian territory and that only Egyptian researchers should have access to them. IBM Research did not wish to install its AllAboard solution on Egyptian servers and their key research employees were indeed foreign nationals. Since then, in 2014, the World Bank Group organized an internal Big Data Innovation Challenge that resulted in far more submissions than expected, with many focusing on CDRs.

In addition to pointing to the opportunities of CDR analytics, these examples also give a sense of the shortcomings of the current state of the field, explored further in the next section.

1.3 LIMITATIONS AND GAPS

The promise of CDR analytics to advance our collective understanding of human dynamics is hard to deny. But many uncertainties and challenges remain that too often tend to be obscured and sidelined by at times narrow and short-sighted views. The projects and challenges discussed so far were unique in that they were made possible by the release of CDRs. Access to these data, however, remains a key limitation for researchers. There is therefore a need to provide a framework guiding the responsible sharing and use of CDRs.[25] Part of what this framework looks like must be informed by previous efforts at using CDRs. And for all its success, the first D4D challenge and several subsequent challenges also raised controversies and concerns.

1. First, there were no submissions from an Ivorian team—this issue of ownership of analytics capacities is increasingly recognized and requires investments in human capital, new technology, infrastructure, geospatial data and management systems to bridge the Big Data digital divide.[26]
2. Second, none of the results has yet led to any concrete implementation in Côte d'Ivoire to benefit the people whose data were used. This highlights the difficulty and importance of linking data and analysis to action, and the persisting 'response gap'—so that what is lacking often is not data but the willingness and ability to implement data-driven solutions.
3. Third, relatedly and critically for our purpose, the project also raised questions about issues of justice and ethics beyond aspects of consent and privacy.

The second D4D 'Senegal' Challenge was designed to partially address these criticisms—notably through greater involvement and engagement of Senegalese authorities and of a prize for the best 'ethical' project—although entries remained overwhelmingly from U.S. and European teams.

Many challenges remain, however, and in general, the practice and use of CDR analytics has been and remains characterized and hampered by biases and gaps—notably a focus on 'getting data and papers out', technical obstacles, institutional fragmentation and the absence of a clear ethical and regulatory framework.

A first basic technical challenge—that is not central to our investigation but is nonetheless related—is differential ownership of cell phones. The most important word in the phrase 'near ubiquitous' may well be 'near', and we must remain aware of the fact that penetration rates close or above 100% do not ensure representativeness. One issue is that as mobile data usage increases, traditional mobile cell phone usage may decrease, so that CDRs will no longer be as representative or relevant as other forms of mobile data. Furthermore, CDRs and subsequent analytics reflect the market share of the company whose CDRs are being analyzed, unless information from different operators can be obtained. A few papers have attempted to estimate and correct for sample bias in CDRs; for example, a paper relying on Kenyan data used 'ground truth' survey data to estimate the impact of differential cell phone ownership on the predictive power of CDR-inferred models of human mobility, finding the CDRs-based models to be surprisingly robust.[27]

Other researchers are currently building on previous work on email and IP data to develop sample bias correction methods.[28] But more research is needed to build solid sample bias correction methods to ensure that CDR analytics does not amplify basic inequities. Furthermore, a more systematic way to document and assess the quality of data (metadata) must be develop as well as standards and methods of data audits so that the validity and accuracy of CDR analytics can be documented. Differential ownership of analytics capacities is another major factor that may contribute to creating and widening a new digital divide between and within countries—as noted above, bridging this digital divide will require investments in human capital, new technology, infrastructure, geospatial data and management systems.[29]

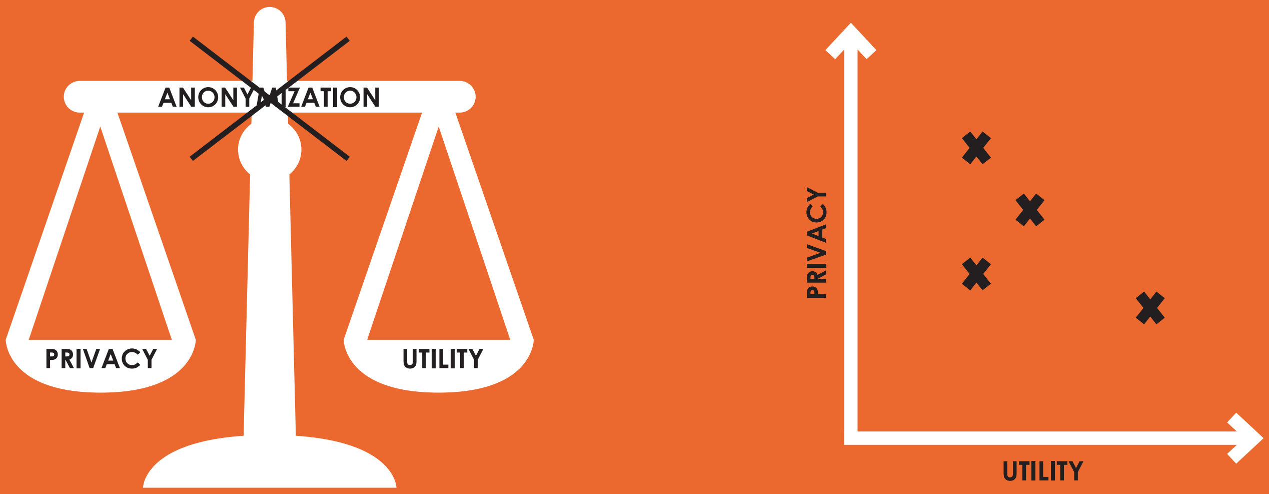
A second set of challenges relates to individual and group privacy[30] and security. These issues have become especially salient since Edward Snowden's revelations on the use of CDRs as part of the US National Security Agency (NSA) surveillance program. Most of the literature is based on carefully 'anonymized' and often aggregated data. But that may not necessarily suffice to alleviate privacy and security concerns. The possibility of de-anonymization and re-identification of previously anonymized or aggregated datasets has been known for years (when multiple datasets are combined, one of which contains an ID).[31]

The predictability and uniqueness of individual human behavior is what makes CDRs valuable, but also what creates risks of 're-identification'. For instance, a paper attempting to derive the 'maximum predictability' in human movement confirmed that human mobility was indeed highly predictable,[32] such that "in a dataset where the location of an individual is specified hourly, and with a spatial resolution equal to that given by the carrier's antennas, four spatio-temporal points are enough to uniquely identify 95% of the individuals".[33] Another recent paper using credit-card data showed that coarsening the data—deleting transaction features in the dataset such as purchase amount and location—only required additional data points to re-identify people out of aggregated datasets, meaning that "even coarse datasets provide little anonymity".[34] Furthermore, individuals' belonging to specific social groups—in terms of their gender, ethnicity, sexual orientation, etc.—tend to show in Big Data including CDRs, and may be used for targeting purposes—whether or not the individual's identities are known—which have raised concerns over 'group privacy'. In other words, 'anonymizing' datasets is in effect an uncertain endeavor; we may protect data and scrap off identifiers but that does not guarantee the privacy or even safety of individuals.

This realization has come with, if not spurred, 'privacy-preserving' innovation. For instance, researchers have developed a methodology that injects 'noise' in CDRs to make re-identification more difficult,[35] although, as mentioned above, this only means it requires a few more data points to single out individuals. Another example is the development, by MIT Media Lab researchers, of OpenPDS and SafeAnswers, respectively "a field-tested, personal metadata management framework which allows individuals to collect, store, and give fine-grained access to their metadata to third parties" and "a new and practical way of protecting the privacy of metadata at an individual level".[36]

But the fact remains that there is and will remain for the foreseeable future potential risks associated with CDR analytics, or more precisely a trade-off between utility derived from granularity and security. Even aggregated data may pose a risk as groups can still be identified based on locations—especially where spatial distribution is associated with ethnic or socio-economic characteristics.

UTILITY-PRIVACY TRADE-OFF WITH BIG DATA



Source: Source: Yves-Alexandre de Montjoye, 2115, lecture slide reproduced with the author's permission

Another set of challenges is legal and institutional. Right now, there is simply no coherent and comprehensive set of regulations or guidelines that govern the field of CDR analytics. Responses to growing demands for CDRs from researchers have typically been ad hoc, granted by Telcos on the basis on personal connections and other arrangements—or for data challenges at their will. Although the concept of 'data philanthropy' has received some traction and may be tactically fruitful, it assumes that CDRs and their rights effectively belong to Telcos—which is disputable. Current practice and legal and policy arrangements across the globe are not suited to the opportunities and risks ahead—and require serious rethinking and reframing.

"DATA PHILANTHROPY": BENEFITS AND LIMITS

"Data philanthropy" refers to the concept and practice of sharing data held by private corporations for purposes of analysis intended to have positive social impact. Although typically framed as a modern form of corporate social responsibility or charity, it has also been described as being "good for business" - by benefitting consumers and economies.[37]

A problem with data philanthropy is that it may reinforce the commonly-held assumption that the data recorded by private corporations effectively belong to them-that holding means owning, and that they may be altruistic or self-interested enough or both to 'give back' some of them. The issue is that there is a much stronger argument to be made that these data do not belong to private corporations, but rather to their individual emitters; the success of the term may make it hard to bring the argument home in the public domain.

Data philanthropy may also be a pragmatic approach convey ing the idea of data being a public good. But there is a risk that what even its proponents describe as a temporary tactical pragmatic move may turn be misconstrued as into a paradigmatic shift where holdership implies ownership, and that we too swiftly forget in the process the characteristics of public good —and the fact that knowledge, not data, is a public good.

Last, and fundamentally, we argue that this is the case because of ill-suited legal frameworks, poorly recognized political parameters and above all a lack of clearly defined ethical principles in which to ground these discussions. For instance, most discussions contrast "opportunities" with "challenges" (or "risks"), or the "promise" of CDR analytics with its "perils"—with little explicit recognition of the roles and rights of different actors, of their competing priorities, and the importance of context. Similarly, everybody agrees that CDR analytics must be 'responsible' or 'ethical'—but it is largely unclear what ethical framework ought to be used to inform action. In addition, calls for new ethical standards and norms appear to be made without considering the lessons from decades of research. This suggests the need to address emerging opportunities and concerns in the field of CDR analytics by identifying political parameters and ethical principles that will help formalize and expand it along clear pragmatic and paradigmatic lines.

2 LEGAL FRAMEWORKS

The use of CDRs must be grounded in respect for the applicable laws regarding data processing and privacy protection. At the same time, existing rules (whether international or domestic) appear increasingly ill-suited to the current fast technological advances and provide paltry privacy protection to users. The potential uses of CDRs, going beyond mere commercial profit or governmental spying and reaching into the realms of disaster relief, development, and health, must be made better known in order to inform the public debate on data privacy. Law and technology must evolve in a dynamic relationship to one another, guided by policy goals and ethical considerations.

A quick overview of the key legal environments for CDR collection and processing will give us a clearer picture of the rules currently binding Telcos, while evincing the need for publicly debated legal reform. We focus on international law as well as U.S. and EU data privacy protections, as those bodies of law have played the greatest part in shaping the behavior of communications companies with regards to user data. It must be noted that this focus alone reveals how much work is still required to build a basic legal framework for CDR analysis: indeed, the U.N. has recently pointed out the existence of a large gap in the adoption of "cyberlegislation" across the world. While privacy and data protection laws are generally strong in developed countries, the United Nations Conference on Trade and Development underlines that it remains "inadequate" in other parts of the world.[38]

2.1 INTERNATIONAL LAW ON PRIVACY

Although international law is short on privacy, it is not silent altogether. Several international instruments affirm and protect the right to a private life, in addition to the fact that the practice of many states is to recognize a constitutional right to privacy. These elements form a workable basis that should be built upon to create a stronger, clearer international base norm to inform CDR use.

Privacy is guaranteed by Article 8 of the European Convention on Human Rights (ECHR), which cites "*the right to respect of (…) private and family life,*" and says this cannot be curtailed except in a manner consistent with the law and necessary for a finite number of legitimate social objectives. This is enforceable before the European Court of Human Rights, which has consistently held that the interception of the content of telephone, fax, and email communications falls within the purview of Article 8.[39] Challenges currently pending before it could have the result of further strengthening the right to privacy among signatories. Of course, the Convention and the interpretation of its provisions by the Strasbourg court are binding only for those signatories. The Universal Declaration of Human Rights of 1948 (UDHR) protects the right to privacy in its Article 12, citing "family, home, and correspondence," and prohibiting arbitrary interference with such privacy. Lastly, the International Covenant on Civil and Political Rights of 1966 (ICCPR) includes the right to privacy in its Article 17, here too citing "family, home, and correspondence" and using much the same language as the UDHR, prohibiting unlawful or arbitrary interference with this right. But these provisions remain very general; the UDHR is not directly binding; and the ICCPR is difficult to enforce.

2.2 U.S. LAW ON PRIVACY

The U.S.' first and broadest privacy protection lies in the Fourth Amendment, which prohibits unreasonable searches and seizures. However, courts have interpreted the Fourth Amendment in a way that excludes CDRs from its scope. A key line of Supreme Court cases has held that an individual has no reasonable expectation of privacy in information he or she has disclosed to third parties. In the case of telephone communications, this includes CDRs: in *Smith v. Maryland*, the Supreme Court held that dialing a telephone number to make a call eliminated any reasonable expectation of privacy in the number dialed, since it had to be conveyed to the telephone company.[40]

Other privacy protections are scattered across several statutes. The first is Title III of the Omnibus Crime Control and Safe Streets Act, adopted by Congress in 1968 and also known as the Federal Wiretap Act. It was adopted in the wake of a series of cases examining the constitutionality of wiretaps. It marked the first clear recognition by American lawmakers that technological developments were enabling the interception of communications, and that this ability should be limited by law. The willful interception of wire or oral communications was prohibited, except with a warrant issued by a judge upon showing of probable cause by law enforcement authorities; each interception order must be specific and limited in time. However, courts have agreed that "pen register" information, which we now call CDR, is not covered by the Federal Wiretap Act.[41]

In 1986, Congress adopted the Electronic Communications Privacy Act (ECPA), creating a private right of action against anyone who "intentionally intercepts, endeavors to intercept, or procures any other person to intercept or endeavor to intercept any wire, oral, or electronic communication."[42] However, in order to make a showing under Title I of ECPA that a conversation was illegally intercepted, a plaintiff must prove five elements: that the defendant (1) intentionally (2) intercepted, endeavored to intercept, or procured someone to intercept or endeavor to intercept (3) the contents of (4) an electronic communication (5) using a device. This showing is subject to statutory exceptions; the most important among which may be consent. Subsequent jurisprudence held that such consent could be explicit or implied. In addition, under ECPA, placing a pen register to collect call details records does not require a search warrant but only a court order. To obtain it, officials need only certify that the information likely to be obtained is "relevant to an ongoing criminal investigation."[44]

After 2001, user privacy was further curtailed in the name of the fight against terrorism. The PATRIOT Act amended ECPA as well as the legal framework for foreign intelligence surveillance. It expanded the definition of a pen register to include the collection of IP addresses and email metadata. It also loosened the restraints on data collection relating to foreign intelligence,[45] a tendency confirmed by further amendments in 2008 and laid bare by Edward Snowden's revelations on the dragnet surveillance conducted by the NSA. In June 2015, the USA FREEDOM Act was adopted. The bill imposes some new restrictions on the bulk collection of U.S. citizens' telephone records, one of the practices revealed by Edward Snowden. However, the government retains strong surveillance powers: the bill does not limit the collection of foreign intelligence; it also extends three PATRIOT Act provisions that had just expired.[46]

2.3 E.U. LAW ON PRIVACY

By contrast to the U.S.' piecemeal approach to data privacy, the E.U. has adopted legislation that provides blanket protection cutting across all sectors of public and business life—perhaps as a result of its view of privacy as dignity.[47] Despite this, its privacy protections come with important loopholes.

The first relevant document is not an E.U. act, but rather the 1981 Council of Europe Convention on the Protection of Individuals With Regard to Automatic Processing of Personal Data, ratified by 45 countries.[48] It is broad not only geographically, but also substantively: it covers all types of data processing, be it by government or business actors. It lays down general principles rather than specific requirements. In particular, Articles 5 and 7 stipulate that any data processed should be done so lawfully and reasonably, and should also be accurate, stored for specific and legitimate purposes, and secured against accidental or unauthorized destruction. Article 6 rules out special categories of sensitive data (racial, religious, and such) that cannot be processed without further safeguards.[49]

But it is in the 1990s that privacy legislation truly began to develop. The Data Protection Directive was adopted in 1995. This text, too, benefits from a broad definition of "personal data" as any information relating to a natural person, whether that person be identified in the data or identifiable from it. It also adopted a broad definition of "data processing" that covers collection, storage, retrieval, blocking, altering, and more. The main principle laid down by the text is that personal data should not be processed at all, except when certain conditions are met: for example, if the subject has given "unambiguous" consent; if the processing is necessary to the performance of a contract; if it is necessary for compliance with a legal obligation; if it is necessary to protect the vital interests of the subject…[50] Data can only be processed for the purposes specified.[51] Sensitive data (religious, racial, political, sexual, and health-related) benefits from additional protection.[52] Furthermore, the data subject must be informed of the processing[53] and has a right to access the data and, in some cases, to rectify or erase it.[54] There is, however, one crucial caveat. Article 3 of the E.U. Directive explicitly excludes law enforcement and security-related data processing from the scope of the act.[55]

The 2006 Data Retention Directive later fleshed out specific rules applying to data retention for law enforcement and other security purposes. Adopted in the wake of terrorist attacks in London and Madrid, it clearly leaned further than previous texts in the direction of security over privacy. In fact, where law enforcement and security concerns are at play, it adopted the reverse principle to the 1995 directive: it created an obligation to retain user data. The directive required states to ensure that service providers retained certain categories of data for purposes of investigation, detection, and prosecution of serious crime. This did not include the content of conversations, but rather CDRs including the dates, times, and duration of communications, as well as user IDs and telephone numbers. The directive limited the people who could access the stored data (article 4). The period of retention was to last at least six months and at most two years (article 6).[56] The Data Retention Directive provoked a backlash from constitutional courts and the European Court of Justice (ECJ): in April 2014, the ECJ ruled that the directive was invalid and void *ab initio* for being incompatible with the right to private life.[57]

2.4 REFORMING THE LEGAL FRAMEWORK FOR DATA COLLECTION AND PROCESSING

The Snowden revelations sparked a flurry of judicial activity and a jumble of reform proposals, of which the USA FREEDOM Act is the latest illustration. In April 2015, Amnesty International and other human rights groups brought a claim against the government of the United Kingdom for indiscriminate surveillance practices.[58] Lastly, the EU Parliament is working on a new Data Protection package including one directive and one regulation. However, lawmakers have yet to agree on the new rules; the package has been held up in Parliament for more than three years, perhaps a reflection of the many conflicting commercial, political, and ethical interests at play.[59]

This only underscores the need for a more transparent public debate over the ownership and use of data, over the balance between privacy and security, and between socially beneficial uses of data and individual and group privacy rights over those data. Existing laws were adopted at a time when the current mass collection and potential uses of data were unimaginable.

In order to avoid built-in obsolescence, stakeholders should heed the lessons of past attempts at legislating on data: instead of focusing on specific technologies and their existing uses, rules and regulations should take root in broad, strong principles regarding users' rights and protections and clear guidelines on the ethical and security considerations to shape any processing of their information. This must be done, however, on the basis of a sound understanding of the political parameters and incentives that have and will come into play when designing systems.

3 POLITICAL PARAMETERS

3.1 FIVE SETS OF CONSIDERATIONS

To understand the politics of CDR analytics, including collection and sharing, it is useful to consider three extreme positions determined by data collection and data sharing arrangements. These corner solutions delineate a space of possible scenarios and their underlying political forces.

1. A first position where no data is collected and therefore no data is shared or analyzed—this is a case that reflects an exclusive concern for individuals' data ownership and rights to privacy, confidentiality and security. For this reason, it is defined here as the '*extreme individual privacy case*', which can be loosely associated with individual actors.

2. A second position where all data are collected at all time, but no data is shared—the data are not public and rather remain in the hands of a limited number of actors who use them for commercial purposes, and refrain from sharing them because it could provide valuable information to competitors. Because this case appears to be primarily about commercial considerations, it is defined here as the extreme business interests case. This may arguably reflect corporate actors. However, similar views may be held, for example, by government when considering intelligence data gathering.

3. A third position where all data are collected and made public at all times, reflecting the idea that social 'public good' value can be yielded (obtained?) by opening and analyzing Big Data, including CDRs. Because this position is primarily about social 'public good' value—such as averting the next cholera outbreak or cutting transportation time—it is defined here as the extreme social good case. Arguably this may reflect government actors.

None of the three positions described above are realistic, nor are they desirable: critically, these are *ideal-typical* (or *stereotypical*) categories meant to facilitate the exposition of *kinds*, and not actor-specific concerns. All stakeholders, including users, corporations and governments have an interest in finding a right balance between how much is collected, how much is shared, and in what way—in terms of temporal and geographical aggregation, time lag etc. For example, individuals will likely agree to some amount of data being collected about them if it helps improve their experience or increase specific benefits. Users may perceive the value of having some of their personal data collected, shared, and analyzed—even as they may insist on strong anonymization, aggregation, or 'expiration dates'—if doing so can help save a life. At the same time it is clear that widespread data collection and data sharing is not supported under privacy, confidentiality and security considerations.

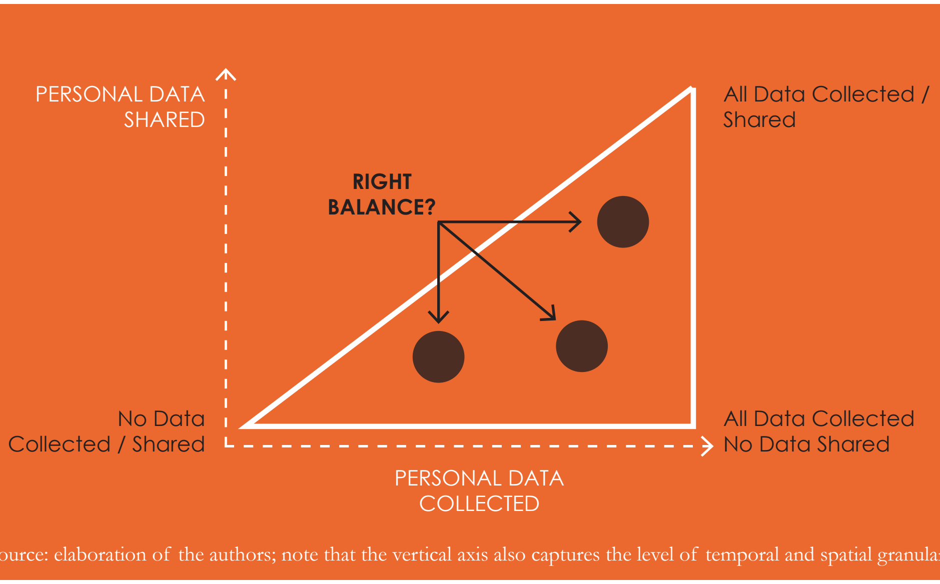
Corporations on the other hand may have an interest in ensuring that some of the data they collect and hold is made public if it can be used for the benefit of their users. They also aim to contribute to the development of the economies where they operate, for both commercial and societal considerations. However, having all data shared is not an acceptable position either given the commercial value and financial and technical barriers this would raise. Likewise, having all data collected and shared at all time is unlikely to be desirable in any circumstances.

These three positions delineate a bounded "data collection and sharing" triangle within which the right balance can be achieved. It helps assess and discuss the pros and cons of each coordinate in the triangle, all else equal, in a structured and systematic way. It further allows greater depth and complexity than when relying on straight axis ranging from 'promises' to 'perils', or by considering individual considerations as a mandatory but essentially secondary part in the dialogue between commercial and societal considerations.

Where and what the right balance is remains to be determined and will be influenced by two additional factors:
1. The features and characteristics of the data being shared, including the risk of re-identification, level of aggregation, and perception of sensitivity. For example, how much data can be collected and shared may depend on whether the data is seen as sensitive, touching on perceptions or feelings, as opposed to economic data such as consumption patterns;

2. Contextual characteristics, which are further discussed in the next section.

DATA COLLECTION AND SHARING SPACE
DATA CHARACTERISTICS AND CONTEXTUAL ENVIRONMENT


PERSONAL DATA SHARED
All Data Collected / Shared
RIGHT BALANCE?
No Data Collected / Shared
All Data Collected No Data Shared
PERSONAL DATA COLLECTED

Source: elaboration of the authors; note that the vertical axis also captures the level of temporal and spatial granularity

3.2 THE IMPORTANCE OF CONTEXTUAL FACTORS

The ‘right’ balance is influenced by contextual factors that are identified here as *systematic* and *idiosyncratic*.

- Systematic factors refer to the main prevailing features of the human ecosystems considered. For instance, there are inherent risks of security breaches through the entire ‘data chain’—from acquisition, storage, sharing of data, analysis, and sharing of results. But the problem will be especially salient where and when the operating environment of a company is weak at restricting access or use of CDRs, or where mobile phone companies are faced with oppressive regimes who may seek to gain access to sensitive data. So it may be that the appropriate balance in a given country would be ill advised in another, or that the ‘right balance’ in a given country may change over time with political and technical progress. The ‘social value’ argument and thus the case for opening up CDRs for analysis will be stronger where and as researchers and policymakers are better at using and relying on CDR analytics such that significant additional societal value is created and can be shared—in the form of greater political stability or higher economic growth. Also, how ‘commercial’ considerations play out and affect the choice of the ‘right’ balance for a given Telco—which are all faced with these questions—depends on the decision of others: if all participate, then the strength of the argument of a loss of comparative competitive advantage is lessened. The point is that although inconsistency of the legal or regulatory environment guiding opening and use of CDRs across countries can be problematic, it seems implausible and undesirable to settle on global standards and norms.

- In addition, idiosyncratic factors—i.e. fast changes in prevailing circumstances—matter. Just as much as the sensitivity of a malfunction detection system designed for an alarm clock needs to be enhanced if repurposed to monitor a nuclear plant, the right balance between the three sets of considerations, holding systemic parameters fixed, ought to change if prevailing conditions change dramatically—for instance in the case of an acute public health crisis. This does not mean that individual considerations—by which we mean privacy—are no longer relevant, but their weight must be reassessed against the expected benefits or opportunity costs of opening up the CDRs vs. keeping them locked—in ways that may not be straightforward. The Ebola crisis offers an interesting case to discuss these points and tensions concretely. Several commentators argued that the crisis made opening up CDRs a near moral imperative, and blamed poor coordination for the absence of effective action in that respect.[60] At the same time, and to play devil’s advocate, one could also argue that these countries’ political, economic and historical characteristics raise significant concerns as to the potential misuse of CDR analytics, especially in such volatile times and in their aftermath; it also largely remains to be seen if and how CDR analytics could effectively be used to improve response on the ground.

What we have established so far is that discussions about CDR analytics would benefit from being framed by the aforementioned political parameters, which can be distilled as follows:

1. There exist three distinct sets of legitimate considerations that all agents face to varying degrees in different places and at different times;

2. The appropriate balance depends on the type of data and

their characteristics, including level of aggregation;

3. The appropriate balance depends on slow changing characteristics (systemic factors) but can and probably should be altered by sudden crisis or events (idiosyncratic factors);

The discussion above further suggests that the position, modalities and movement of the ‘right balance’ depend critically on ethical principles that need to be spelled out. This is the focus of the next section.

4 ETHICAL PRINCIPLES

4.1 FRAMEWORK

The discussion of ethical principles, dilemmas and risks in collecting and sharing CDRs must build on several decades of progress in understanding and defining principles for ethical research. Similar eEthical principles historically have historically been developed with primarilyin the biomedical and behavioral sciences. in mind. The practice of Big Data analytics, and specifically the use of CDRs, closely resembles research cycles and processes, and the insights sought are relevant to behavioral science. While arguably corporations are not research institutions lessons can be learned and modeled from these more developed ethical frameworks and applied to these new emerging fields. This white paper proposes that existing ethical principles provide a valuable and possibly sufficient framework to guide the emerging field of CDR analytics. The research ethics frame is the most appropriate to highlight the most consequential and problematic issues as well as opportunities raised by CDR analytics broadly considered.

There are a number of landmark guides for ethical research principles as laid out in the Nuremberg Code, Declaration of Helsinki, and Belmont report. A more recent initiative by the US Department of Homeland Security, Science and Technology, Cyber Security Division revised and adapted established ethical principles in the context of the ICT and data revolutions. The result was published as The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research.” This effort is the most closely relevant to the concerns raised by CDR analytics and is used here as primary ethical framework.[61] The Menlo report, first published in December 2011 and amended in 2012, identified four key ethical principles for computer and information security research, reflecting exiting principles:

1. Beneficence;
2. Respect for Persons;
3. Justice;
4. Respect for Law and Public Interest.

The following table describes these four key ethical principles, which we then unpack and comment in the case of CDR analytics.

THE MENLO REPORT ETHICAL PRINCIPLES GUIDING ICT RESEARCH

1

BENEFICENCE

- *Do not harm;*
- *Maximize probable benefits and minimize probable harms;*
- *Systematically assess both risk of harm and benefit.*

2

RESPECT FOR PERSONS

- *Participation as a research subject is voluntary, and follows from informed consent;*
- *Treat individuals as autonomous agents and respect their right to determine their own best interests;*
- *Respect individuals who are not targets of research yet are impacted;*
- *Individuals with diminished autonomy, who are incapable of deciding for themselves, are entitled to protection.*

3

JUSTICE

- *Each person deserves equal consideration in how to be treated, and the benefits of research should be fairly distributed according to individual need, effort, societal contribution, and merit;*
- *Selection of subjects should be fair, and burdens should be allocated equitably across impacted subjects.*

4

RESPECT FOR LAW & PUBLIC INTEREST

- *Engage in legal due diligence;*
- *Be transparent in methods and results;*
- *Be accountable for actions.*

4.2 UNPACKING ETHICAL PRINCIPLES FOR CDR ANALYTICS

1. BENEFICENCE: UNDERSTANDING RISKS AND BENEFITS

The principle of beneficence refers to “a moral obligation to act for the others’ benefit, helping them to further their important and legitimate interests, often by preventing or removing possible harms.”[62] Under this principle, researchers must maximize the probability and magnitude of benefits to individual research subjects as well as to society. The recognized benefits are what transform CDRs into valuable assets whose potential should be unlocked.

However, what constitutes a benefit or a risk is not always straightforward or consensual—and, as discussed above, depends to a large degree on the actors considered. CDRs are largely stored and handled by private companies, which are the ones investing in transmission and storage infrastructures. Commercial considerations must therefore be taken into account in framing the risks and benefits of using and sharing CDRs.

The potential benefits and harm of any project making use of CDRs certainly depend on that specific project’s objectives. Furthermore, unlocking the benefits of CDRs will require experimentation and practice that may not have direct value or benefits besides learning—akin to fundamental science which ultimately leads to broader benefits.

2. RESPECT FOR PERSONS

The issue of consent is gaining attention and is central to the privacy concerns relating to the use and sharing of CDRs. Specifically, users of mobile phone handsets rarely grant formal permission for their personal data to be used and shared. If they do so, it is often with little to no choice, since not consenting would greatly or fully limit restrict their access to the technology and associated services. Furthermore, the choice given to consumers is typically to either dissent or fully consent , regardless of what use may be made of the data several years later, or by a third party should it be accessed by them. There is little to no way for consumers to exclude specific usage of their data that they do not want, raising major questions around the secondary use of data.

The issue of consent is not purely “informational”—i.e. does a user agree or not with proposed uses of data about themselves. Ultimately it is about potential risks and enabling users to make decisions for themselves. Granting usage of their data may expose individuals to various harms and risks, especially as increased data sharing increases the risk of confidentiality breaches or misuse of the data. The use of their data may also go against their cultural or religious values. Much of the discussion has been focused on “opt-in / opt-out” which requires the user to either actively consent to terms of use that include data sharing, or to actively dissent, the default setting being that of consent. More advanced models being discussed include a more flexible process where permissions can be granted in a variety of ways and depends upon the context of use—either through explicit consent or implicitly through compatible action.

For secondary use, it is generally agreed that uses that are consistent with the original context can carry the permission granted in that context, but that new uses should require new consent. Broad (unlimited) consent remains widely use despite strong opposition on moral, ethical, and legal grounds. An even more advanced model proposes that individuals would permanently “carry” a set of permissions that they grant to algorithms seeking to use their data—no matter what data, enabling them to modify access and permissions at any time.

3. JUSTICE: BIAS AND INEQUALITIES

The principle of justice highlights issues of fairness and equal distribution of risks and benefits. Arguably one of its key aspects is that everyone must have an opportunity to contribute and benefit (e.g. from CDR analytics) even when unequal access to technology exists. Yet, whose data is considered in CDR analytics is inherently affected by unequal access to and use of mobile phones, creating inherent biases and violating the principle of justice.

This creates yet another tension in CDR analytics: it is especially relevant in otherwise data-poor environments, but it is precisely in these environments that access to technology is most unequal, which implies that CDRs are non-representative data. The underlying challenge is that CDRs will typically reflect structural inequalities in any given country: owning a cell phone is strongly correlated with socio-economic status, and even in countries with high mobile phone penetration, CDRs may be analyzed along criteria that would single out more affluent individuals or areas. These biases hinders the external validity of findings based on CDRs and may potentially reinforce structural inequalities (if, for instance, programs are based on data from areas with high cell phone usage).

Biases may be unproblematic as long as they are well understood and corrected for. Besides Buckee et al. (2013), correction methodologies for biases in e-mail data have already been proposed, although validation is difficult for lack of reliable ‘ground-truthing’ data.[63] No similar efforts exist to this day with regards to CDRs. Furthermore, it is likely that as cell phone penetration and patterns of use change, there will be a need to constantly adapt methods and algorithms developed to correct biases. This is clearly a challenge and a priority for future research.

Beside the issue of bias in the data, the analysis of CDRs may also lead to unequal targeting of individuals or groups based on their ethnicity, gender, religion, and sexual orientation. The notion of group privacy recoups the rights to groups and their members not to be identified and targeted as such; this concept is likely to gain traction as it is intrinsically related to discrimination, targeting, etc. It is indeed possible to predict group-level characteristics—for instance, distinguishing a 40-year old gay male from a 20 year old heterosexual female using various big data streams—using credit card transactions, social media data, etc., and in all likelihood CDR may also reflect similar characteristics. In such a case, having anonymized, even aggregated, data, may be insufficient to avoid discriminations and negative targeting. These concerns however may be at odds with the increased popularity of the concepts of "hyper-personalization" of marketing, under which individual characteristics are defined so well that they enable corporations to offer highly customized offers and services.

4. RESPECT FOR LAW AND PUBLIC INTEREST

The fourth and last principle framing our discussion highlights the need to engage in legal due diligence; be transparent in methods and results; and be accountable for actions. However, inconsistency of the legal or regulatory environment guiding the opening and use of CDRs across countries is problematic where legal protections are insufficient to protect the individual, and where cross-border accountability is difficult to enforce (e.g. if an individual is put at risk because of a foreign organization's use of his or her CDRs, what recourse is available to that individual?).

Telcos are especially concerned about their legal exposure if CDRs were to be used to identify, target and/or discriminate against specific individuals or groups. For example, participants in protests can easily be identified through CDRs. Telcos may be confronted to local legal requirements that may be at odd with international law and could potentially be held liable if their data were used in mass atrocities, something not entirely impossible. In repressive environments, Telcos should consider as their first priority to protect the sources of information (their customers) and place-sensitive data beyond the reach of authorities, even where this may be against their financial and commercial interests. At the same time, Telcos which have access to potentially life-saving information may be morally, if not legally, required to make that information available.

5 OPERATIONAL IMPLICATIONS AND POLICY OPTIONS

5.1 OPERATIONAL IMPLICATIONS

Having used core ethical principles to frame the key challenges emerging in the rapidly growing practice of CDR analytics, this paper also serves as a call to renew commitment to these principles. Putting these principles into practice requires agreeing on a number of operational implications without which they will remain dead slogans.

One is to recognize the plurality of actors, the legitimacy of all and the responsibilities of each, which calls for a collegial and coordinated approach to the problem. Telecom companies contribute—as socio-economic agents—to enhancing the welfare of societies where they operate. They should not mimic the most negative aspects of extractive industries were valuable resources are exported with little or no benefits locally. At the same time, local government, researchers, and organizations are unlikely to have the ability to take advantage of CDRs, including the necessary financial resources and local expertise, without assistance from Telcos. This will require new public-private partnerships that leverage private sector data for public policy. It will also require new collaborations with researchers and investment in research capacities to develop skills and research in cloud and high performance computing, for example through North-South and South-South PhD program development.

Telcos may participate in such partnerships if regulators and legislators ensure that investments by telecom operators are fairly rewarded and incentivized. These same legislators must at the same time ensure that the rights of their citizens be fully upheld and will need the appropriate regulatory frameworks to enable (and at times compel) access to data for public good while remaining mindful of privacy rights. Researchers may also be stakeholders in the public-private sharing of CDRs, but they too must have defined roles and responsibilities. Researchers should engage in and support efforts to find standardized data sharing tools and protocols. At the same time, the multiple and sometimes competing demands on Telcos to provide data must be coordinated.

Data requirements must also be better defined to avoid demands that seek to capture anything and everything in near real-time, especially when and where historical data may be sufficient for the proposed work. Indeed, the notion that CDRs will help spur 'agile' development in the near future—which would justify getting real-time data for instance—is largely unsubstantiated. Informations contained in 'old' CDRs are interesting for research, and their result does not depend too much on the timing of extraction of the data (let's say within the last 2 or 3 years). Some applications may well need 'fresh' data, even real-time data, but these are technically and ethically more difficult to obtain. As a result, when such data are used, we must be clear about the expected individual and societal benefits that justify these efforts/risks. Additionally, aspects of capacity development and participation of researchers from countries whose data are being used should also become standard practice.

These considerations show that the responsible development of CDR analytics will require the involvement, support and good will of all actors involved. Too often the questions raised in this paper are discussed in isolation by a select group of actors, with the individual perspective being the least represented. The recent set up of the UN group is illustrative of the visibility given to corporate and societal perspectives (government) at the expense the individual perspective. Similarly, calls for open data and data philanthropy are largely framed around corporate and societal benefits, with insufficient attention paid to individual considerations.

In addition, it is important to highlight that CDRs alone offer only limited insight, and that their richness is only unlocked when combined with other data streams. There is therefore a need to create better integration and access across data streams. More traditional forms of data are needed. E.g. tracking poverty or socioeconomic levels using CDRs requires having poverty or socioeconomic data to start with (and a CDR analysis alone would be very sensitive to sample bias). 'Historical data', even aggregated data, are extremely valuable, and in a way even more so than 'real-time' raw data, because they do not just allow us, but indeed compel us to focus on building methods and tools under greater constraints, before attempting to do now-casting of current variables in any way.

Another key operational principle is to think and act strategically, with a longer-term horizon than the next paper or quarterly report. Changing the overall timeframe—thinking and planning for the next five to ten years—does change short-term decisions and priorities. Capacity-building and standard- and norm-setting are absolutely essential ingredients and objectives for the expansion of CDR analytics. This refers to the need to build on existing models and norms, as well as ongoing work. An example of such an attempt is the WEF's 'personal data initiative' examining among other issues how the process of granting permissions for personal data use and exchange (consent) must be updated for a Big Data (CDRs) world.

Ethical concerns are not exclusive to CDRs: similar debates and questions are regularly raised in the context of the ICT and data revolutions, as during the Open Government Partnership Summit sessions on whistleblowing, privacy, and safeguarding civic space—especially in light of the Snowden case—or at the Technology Salon on Participatory Mapping.[64] The fact that similar issues are being discussed by a wide range of actors with a wide range of perspectives suggests a high potential for cross-discipline learning.

A last operational principle is context-sensitivity and appropriateness—which we shall illustrate by discussing the value of and case for using non-anonymized data in crisis contexts. The critical use of non-anonymized data offers a good illustration of the need to find a right balance between various interests, but also identify the appropriate mechanisms and principles for the responsible sharing of data.

During a disaster, access to identifiable data from mobile phone operators may be critical to assist with the reunification of families separated by the disaster, or to assist the identification of body remains. Mobile phone data may also be associated with identifiable data for the purpose of tracking services and benefits used by disaster-affected individuals.

In such contexts, the societal value of identifiable CDR data is very high, with the crisis potentially justifying significantly downplaying individual and commercial concerns for some time. Similar uses of CDR data have already taken place, but without guiding principles, these are potentially creating liabilities and risks for affected communities.

5.2 POSSIBLE POLICY OPTIONS

It remains to be seen what governance and technical arrangements should dictate the sharing of data and indicators based on CDRs to increase the availability of datasets and the efficiency of data analysis, tool development, knowledge sharing and so on.

A number of pointers can be discussed describing some of the minimum requirements for such arrangements. This could serve as a starting point for mobile phone companies to work with the research community, governments and other civil society actors toward minimum principles or governance structures.

One critical path to explore is to learn from advances in the protection of human subjects in research to establish a systematic review process to validate when and where such data should be shared. This could, for example, be done under the supervision of neutral, internationally recognized organizations. Specific criteria to judge the benefits and risks should be established under very clear circumstances (e.g. sudden-onset disaster), and reviews should create a learning process to decrease the risks of inappropriate use of the data. Limits may also be established as to the type of analysis that is permitted (e.g. localization of people reported as missing…). For providers, this may require getting prior informed consent from subscribers with the delicate decision of making this mandatory, or as an opt-in or opt-out decision.

In non-acute crisis contexts, solutions should enable mixed usage with various levels of privacy setting / concerns / noise or quality degradation in the data depending on the ultimate usage and perhaps the actors involved. Access to anonymous CDRs might be granted to a research lab for a specific contract, while only access to aggregated indicators (volumes of calls per day per antenna, etc.) could be accessed by a larger community in a more open fashion. Furthermore, some data may need to be eliminated from public records (e.g. antennas at military sites, data from 'extreme' users). Specific terms and conditions must be developed to address this "data cleaning" process. The level of information loss due to CDRs detail reduction (e.g. how much do we lose by reducing the granularity from Antenna location, to the level up) must result from a systematic and balanced analysis of objectives, risks and benefits. This may require establishing minimal data requirements based on various research uses, seeking to answer questions like 'Is real time needed?', 'If not, what type of past data?', and 'Is there a minimum sample size for a particular analysis?'

A likely compromise on more systematic sharing of CDRs would enable both individuals and mobile phone companies to maintain and possibly enhance control over CDRs to respect individuals' agency, while Telcos maintain their contractual relation with their customers, the respect of their privacy, and control critical information that may help direct competitors (local market share, zone of customer acquisitions). Any solution should also enable CDRs to systematically carry standardized metadata that include any limits on the use of the metadata. In case of aggregation, use of CDRs should be restricted to the most restrictive use granted by any individual whose data are included in the aggregated data. One key aspect of the enhanced control of individuals and metadata that must necessarily accompany CDRs is the ability to maintain an "expiration date" to protect privacy and other individual rights in the long term, echoing ongoing current discussions on the "*erasable future of social media*".

Whichever approach is chosen should further enable greater participation and capacity development of local actors, while complying with local privacy protection regulations. Local partnerships and data processing accreditation are likely to be necessary. A centralized system (real institutional CDR sharing and clearing house for research) is, on the other hand, unlikely and undesirable. Rather, a more distributed model based on principles and standards is more likely to be implemented by various actors, both to enable a better control and develop specific areas of expertise. Due to the many commonalities between analyses (e.g. the use of background maps), some elements of data/indicators sharing will be effective as well.

In conclusion, the risks, constraints and challenges of enabling wider access to CDRs to support social good should not obscure the fact that the combination of exponential growth rates of mobile phone penetration and data production in low- and middle-income countries and intense interest and efforts from social scientists and policy-makers will, in all likelihood, make CDR analytics, or derived indicators, a standard tool for researchers by the end of the decade.

The broad array of societal opportunities presented by Big Data in emerging countries is real, and there are ways to develop the tools, process and policies to cover both the societal needs and the commercial development goals of local companies, often with a combination of the two on the same projects. Despite the inherent value of CDRs for mobile phone companies, these actors recognize that broad principles of open innovation or open data should apply with limits to guarantee that ethical principle are respected, and that the process results from a consensus between the views and interests of all stakeholders, including users.

END NOTES

[1] DeGusta, Michael, “Are smart phones spreading faster than any technology in human history?” *Massachusetts Institute of Technology Review*, 2012.

[2] “The Mobile Economy 2015,” London: GSM Association, 2015, available at: http://www.gsmamobileeconomy.com/GSMA_Global_Mobile_Economy_Report_2015.pdf.

[3] Haddad, Ryan, Tim Kelly, Teemu Leinonen, and Vesa Saarinen, “Using Locational Data from Mobile Phones to Enhance the Science of Delivery,” Washington, DC: World Bank, 2014, available at: https://openknowledge.worldbank.org/handle/10986/19316.

[4] Bouillet, Eric, et al., “Processing 6 billion CDRs/day: from research to production (experience report),” *Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems*, New York: ACM, 2012: p. 264-67.

[5] Based on computation of statistics available from the International Telecommunication Union, available at: http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx.

[6] See for example: Davenport, Thomas H., *Enterprise analytics: optimize performance, process, and decisions through big data*, Upper Saddle River: FT Press, 2012; or McAfee, Andrew and Erik Brynjolfsson, “Big data: The management revolution,” *Harvard Business Review*, vol. 90, no. 10 (October 2012): p. 60-68, available at: http://hbr.org/2012/10/big-data-the-management-revolution.

[7] Wall, Matthew, “Ebola: Can big data analytics help contain its spread?” BBC News, 15 October 2014, available at: http://www.bbc.com/news/business-29617831.

[8] Kirkpatrick, Robert, “Data Philanthropy is Good for Business,” Forbes Magazine, 20 September 2011, available at: http://www.forbes.com/sites/oreillymedia/2011/09/20/data-philanthropy-is-good-for-business/.

[9] de Montjoye, Yves-Alexandre, Jake Kendall, and Cameron F. Kerry, “Enabling Humanitarian Use of Mobile Phone Data,” *Issues in Technology Innovation*, Washington, DC: Center for Technology Innovation at Brookings, November 2014, available at: http://www.brookings.edu/~/media/research/files/papers/2014/11/12-enabling-humanitarian-mobile-phone-data/brookingstechmobilephonedataweb.pdf.

[10] Letouzé, Emmanuel, Patrick Vinck, and Patrick Meier, “Big Data for Conflict Prevention: When the New Oil Meets Old Fires,” New York: International Peace Institute, 2013.

[11] Pentland, Alex, “Reinventing Society in the Wake of Big Data,” Edge, 30 August 2012.

[12] Letouzé, Emmanuel, “Big Data for Development: Challenges & Opportunities,” New York: UN Global Pulse, 2012; and Letouzé, Emmanuel, “Big Data and Development Primer,” Data-Pop Alliance, 2015.

[13] de Cordes, Nicolas, “A ‘big data’ competition open to the scientific community,” Orange, available at: http://www.orange.com/en/about/Group/our-features/2013/D4D/Data-for-Development.

[14] For an overview, see “Mobile Phone Network Data for Development,” New York: UN Global Pulse, October 2013.

[15] Letouzé, Emmanuel, Patrick Vinck, and Patrick Meier, “Big Data for Conflict Prevention: When the New Oil Meets Old Fires,” New York: International Peace Institute, 2013.

[16] See Nurmi, Petteri, “Data Analysis from Mobile Networks,” Helsinki: University of Helsinki, 2012; Talbot, David, “African Bus Routes Redrawn Using Cell-Phone Data,” *MIT Technology Review*, 30 April 2013; Letouzé, Emmanuel, Patrick Vinck, and Patrick Meier, “Big Data for Conflict Prevention: When the New Oil Meets Old Fires,” New York: International Peace Institute, 2013.

[17] See for example: Perry, Chris, “Machine Learning and Conflict Prediction: A Use Case,” *Stability: International Journal of Security and Development*, vol. 2, no. 3 (31 October 2013): p. 56-73, available at: http://ssrn.com/abstract=2358117; Himelfarb, Sheldon, “Can Big Data Stop Wars Before They Happen?” *Foreign Policy*, 25 April 25 2014, available at: http://foreignpolicy.com/2014/04/25/can-big-data-stop-wars-before-they-happen/; Ulfelder, Jay, “A New Statistical Approach to Assessing Risks of State-Led Mass Killing,” *Dart-Throwing Chimp*, 22 January 2014, available at: https://dartthrowingchimp.wordpress.com/2014/01/22/a-new-statistical-approach-to-assessing-risks-of-state-led-mass-killing/; Bogomolov, Andrei, Bruno Lepri, et al., “Moves on the street: Predicting crime hotspots using aggregated anonymized data on people dynamics,” 2015; Bogomolov, Andrey, Bruno Lepri, et al., “Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data,” *Proceedings of the 16th International Conference on Multimodal Interaction*, New Work: ACM, 2014: p. 427-434, available at: http://arxiv.org/abs/1409.2983; Letouzé, Emmanuel, Patrick Vinck, and Patrick Meier, “Big Data for Conflict Prevention: When the New Oil Meets Old Fires,” New York: International Peace Institute, 2013; see also on CDR sharing for peace-building applications the Peace Radar project at Leiden University : http://datapool.zz-demos.net.

[18] Butler, Declan, “When Google got flu wrong,” *Nature*, no. 494 (14 February 2013): p. 155-156, available at: http://www.nature.com/news/when-google-got-flu-wrong-1.12413.

[19] Lazer, David, Ryan Kennedy, Gary King, and Alessandre Vespignani, “The parable of Google Flu: Traps in big data analysis,” Science, vol. 343, no. 6176 (March 2014): p. 1203-1205.

[20] Including the MIT Technology Review, Wall Street Journal, Wired Magazine, Le Monde, La Republica, Arte Futurs, and the BBC.

[21] See NetMob (May 1-3, 2013, MIT) webpage: http://perso.uclouvain.be/vincent.blondel/netmob/2013/.

[22] See Data for Development Senegal (2014) webpage: http://www.d4d.orange.com/en/home.

[23] For a description of these challenges, see for example: http://www.telecomitalia.com/tit/en/bigdatachallenge.html http://dynamicinsights.telefonica.com/674/the-details.

[24] See for example: Pastor-Escuredo, David, et al., "Flooding through the Lens of Mobile Phone Activity," *Proceedings of the IEEE Global Humanitarian Technology Conference*, Silicon Valley, CA, 2014: p. 279-286.

[25] In addition to the principles and suggestions put forth in the present paper, another ongoing initiative worth mentioning is the Data Governance Project: http://www.responsible-data.org/data-governance-project.html and the wok of Data & Society's Council for Big Data, Ethics, and Society: http://bdes.datasociety.net

[26] "A World That Counts: Mobilising The Data Revolution for Sustainable Development," Independent Expert Advisory Group on a Data Revolution for Sustainable Development, November 2014, available at: http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf.

[27] Buckee, Caroline O., et al., "The impact of biases in mobile phone ownership on estimates of human mobility," *Interface (Journal of the Royal Society)*, vol. 10, no. 81 (April 2013), available at: http://rsif.royalsocietypublishing.org/content/10/81/20120986.

[28] Zagheni, Emilio, and Ingmar Weber, "Demographic Research with Non-Representative Data," *International Journal of Manpower*, vol. 36, no. 1 (2015): p. 13-25.

[29] "A World That Counts: Mobilising The Data Revolution for Sustainable Development," Independent Expert Advisory Group on a Data Revolution for Sustainable Development, November 2014, available at: http://www.undatarevolution.org/wp-content/uploads/2014/11/A-World-That-Counts.pdf.

[30] See chapter on "Big Data and Group Privacy" by Lanah Kammourieh et al., L., Floridi, F. and van der Sloot, B., (Eds.) forthcoming book *Group Privacy: New Challenges of Data Technologies*, to be published in the fall of 2015.

[31] Narayanan, Arvind, and Vitaly Shmatikov, "Robust De-anonymization of Large Sparse Datasets," *Proceedings of the IEEE Symposium on Security and Privacy*, Oakland, CA, 18-22 May, 2008: p. 111-125. Available at: https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf.

[32] Lu, Xin, Erik Wetter, Nita Bharti, Andrew J. Tatem, and Linus Bengtsson, "Approaching the Limit of Predictability in Human Mobility," *Scientific Reports*, vol. 3, no. 2923 (2013), available at: http://www.nature.com/srep/2013/131011/srep02923/full/srep02923.html; Bagrow, James P., Dashun Wang, Albert-László Barabási, "Collective Response of Human Populations to Large Scale Emergencies," *PLoS ONE*, vol. 6, no. 3 (2010): e17680, available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0017680

[33] de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel," *Scientific Reports*, vol. 3, no. 1376 (2013), available at: http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html

[34] de Montjoye, Yves-Alexandre, Laura Radaelli, Vivek Kumar Singh, Alex "Sandy" Pentland , "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221 (30 January 2015): p. 536-539.

[35] Becker, Richard, et al., "Human Mobility Characterization from Cellular Network Data," *Commuinications of the ACM*, vol. 56, no. 1 (January 2013): p. 74-82, available at: http://dl.acm.org/citation.cfm?id=2398375

[36] See OpenPDS and SafeAnswers at: http://openpds.media.mit.edu/.

[37] For more on data philanthropy, see for example: "Sharing Data As Corporate Philanthropy," *Markets for Good*, 12 September 2014, available at: http://www.marketsforgood.org/sharing-data-as-corporate-philanthropy/; Kirkpatrick, Robert, "Data Philanthropy is Good for Business," Forbes, 20 September 2011, available at: http://www.forbes.com/fdc/welcome_mjx.shtml; Kirkpatrick, Robert, "Data Philanthropy: Public & Private Sector Data Sharing for Global Resilience," *UN Global Pulse Blog*, 16 September 2011, available at: http://www.unglobalpulse.org/blog/data-philanthropy-public-private-sector-data-sharing-global-resilience.

[38] UNCTAD, *Information Economy Report*, 24 March 2015, http://unctad.org/en/PublicationsLibrary/ier2015_en.pdf, pp. 64-65.

[39] See *Malone v. United Kingdom* (1985) 7 EHRR 14, at 64; *Weber v. Germany* (2008) 46 EHRR SE5, at 77; and *Kennedy v. United Kingdom* (2011), 52 EHRR 4, at 118.

[40] Smith v. Maryland, 442 U.S. 735 (1979), at 745-746.

[41] R. B. Parrish, "Circumventing Title III : the Use of Pen Register Surveillance in Law Enforcement", Duke Law Journal, 1977, at 751. Following Robert Pikowsky's definition, a pen register is a device that can be attached to a specific phone line for the purpose of covertly recording the outgoing telephone numbers dialed. See Robert A. Pikowsky, "The Need for Revisions to the Law of Wiretapping and Interception of Email," 10 *Michigan Telecommunications and Technology Law Review*, 1, 43 (Fall 2003), at 17, available at: http://www.mttlr.org/volten/pikowsky.pdf.

[42] 18 U.S.C. 2511(1)(a) (1994) and 18 U.S.C. 2520 (2005).

[43] *In re Pharmatrak, Inc.*, 329 F.3d 9 (1st Cir. 2003), at 23-24.

[44] 18 U.S.C. 3123.

[45] "Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act of 2001," Pub. L. No. 107-56, 107th Cong. §§ 201-225 (2001).

[46] USA FREEDOM Act, Pub. L. 114-23, https://www.congress.gov/114/bills/hr2048/BILLS-114hr2048enr.pdf. *See also* Human Rights

Watch, "US : Modest Step by Congress on NSA Reform", 8 May 2015, http://www.hrw.org/news/2014/05/08/us-modest-step-congress-nsa-reform.

[47] Whitman, James Q., "Two Western Cultures of Privacy: Dignity Versus Liberty," *Yale Law Journal*, vol. 113, no. 6, at 1160 (April 2004), available at: http://www.yalelawjournal.org/article/the-two-western-cultures-of-privacy-dignity-versus-liberty

[48] Council of Europe, "Convention for the protection of Individuals with regard to Automatic Processing of Personal Data," Status as of May 2nd, 2013, available at: http://conventions.coe.int/Treaty/Commun/ChercheSig.asp?NT=108&CM=1&DF=&CL=ENG.

[49] Council of Europe, "Convention for the protection of Individuals with regard to Automatic Processing of Personal Data," *ETS*, no. 108 (28 January 1981), available at: http://conventions.coe.int/Treaty/en/Treaties/Html/108.htm.

[50] European Parliament and Council of the European Union, "Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data," *Official Journal of the European Union*, vol. L 281 (23 November 1995), article 2 (a) and article 7 (a)-(f), available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML.

[51] *Id.*, article 6 (c).

[52] *Id.*, article 8.

[53] *Id.*, articles 10 and 11.

[54] *Id.*, article 12.

[55] *Id.*, article 3 (2) and article 13 (1) (a)-(d).

[56] European Parliament and Council of the European Union, "Directive 2006/24/EC of the European Parliament and of the Council of 15 March 2006 on the Retention of Data Generated or Processed in Connection With the Provision of Publicly Available Electronic Communications Services or of Public Communications Networks and Amending Directive 2002/58/EC," *Official Journal of the European Union*, vol. L 105 (13 April 2006): p. 54-63, available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:105:0054:0063:EN:PDF.

[57] Grand Chamber, Digital Rights Ireland Ltd. (C–293/12) v. Minister for Communications, Marine and Natural Resources, http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:62012CJ0293&rid=1.

[58] Howell, Valerie, "Rights groups take UK government to European Human Rights Court over mass surveillance," *jurist.org*, 10 April 2015, available at: http://jurist.org/paperchase/2015/04/rights-groups-take-uk-government-to-european-court-of-human-rights-over-mass-surveillance.php.

[59] The text of the proposed Data Protection Regulation is available at http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf. The proposed amendments are available at http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P7-TA-2014-0212+0+DOC+XML+V0//EN.

[60] "Ebola and big data: Waiting on hold," *The Economist (print edition)*, 25 October 2014, available at: http://www.economist.com/news/science-and-technology/21627557-mobile-phone-records-would-help-combat-ebola-epidemic-getting-look.

[61] Dittrich, David, Erin Kenneally, et al., "The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research," Department of Homeland Security (USA), 2012, available at: http://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/.

[62] Stanford Encyclopedia of Philosophy, accessed February 2015, available at: http://plato.stanford.edu/entries/principle-beneficence/.

[63] Zagheni, Emilio, and Ingmar Weber, "Demographic Research with Non-Representative Data," *International Journal of Manpower*, vol. 36, no. 1 (2015): p. 13-25.

[64] "Ethics and Risk in Development," Open knowledge blog, 2013, available at: http://blog.okfn.org/2013/11/05/ethics-and-risk-in-open-development/#sthash.xMA0K3wi.dpuf

DATA POP ALLIANCE

PROMOTING A PEOPLE-CENTERED BIG DATA REVOLUTION

HARVARD HUMANITARIAN INITIATIVE

mit media lab

ODI

DATA-POP ALLIANCE IS A COALITION 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.

WWW.DATAPOPALLIANCE.ORG

CONTACT@DATAPOPALLIANCE.ORG