Title
Applications and Implications of Big Data for Demo-Economic Analysis: The Case of Call-Detail Records

Author
Letouzé, Emmanuel Francis

Publication Date
2016

Applications and Implications of Big Data for Demo-Economic Analysis: The Case of Call-Detail Records

by

Emmanuel Francis Letouzé

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Demography

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Ronald Lee, Chair
Professor Jennifer Johnson-Hanks
Professor Edward Miguel

Summer 2016

**Applications and Implications of Big Data for Demo-Economic Analysis:
The Case of Call-Detail Records**

Abstract

Applications and Implications of Big Data for Demo-Economic Analysis:
The Case of Call-Detail Records

by

Emmanuel Francis Letouzé

Doctor of Philosophy in Demography

University of California, Berkeley

Professor Ronald Lee, Chair

This dissertation analyzes and discusses various applications and implications of Big Data for demo-economic analysis, focusing on the analysis of cell-phone data collected by telecommunication operators for billing purposes, commonly referred to as call-detail records, or CDRs, which include the time and duration of calls, the location of the emitter and receiver, etc. This is done by placing the resulting opportunities and challenges within the broader context of the 'Data Revolution', presented in Chapter 1. In this context, *applications* refer to ways in which CDR analytics can be used for research and policymaking purposes by leveraging the information contained in these data on human behaviors, for example to predict criminality (Chapter 2), and estimate income levels (Chapter 3), mobility patterns (Chapter 4), or population density (Chapter 5). *Implications* refer to ways in which CDR analytics has and can be expected to affect and be affected by ethical, political, legal, and institutional factors and processes (Chapter 6).

At the heart of Big Data are new kinds of passively emitted digital data or ‘crumbs’ that are the by-product of the fast growing and already near-ubiquitous use of digital devices and services by humans around the globe. These 'crumbs' leave digital traces of most of their actions that are collected and can be analyzed through powerful methods and machines by new types of stakeholders, including multi-disciplinary teams.

Chapter 1 analyzes the advent of the Big Data phenomenon over the past decade, with particular attention to its observed and possible effects on social science research and policymaking. In addition to providing an historical overview, it proposes taxonomies and concepts to clarify the nature and significance of the change brought about by Big Data. An important point of the chapter is the distinction it introduces between big data as new kinds of large datasets and Big Data as an ecosystem of 3Cs of Big Data: its crumbs, its capacities, and its communities. The chapter ends by questioning whether these kinds of digital data may replace traditional data and whether Big Data may render the scientific method obsolete, answering by the negative but arguing that social science research will dramatically evolve in contact with Big Data, while in turn shaping Big Data.

The subsequent four empirical chapters focus on different applications of CDR analytics to social problems:

Chapter 2 uses CDRs from Telefónica in London in conjunction with other socio-demographic data including police records to attempt to predict future crime hotspots, and presents a model with a predictive power of close to 70%. This chapter offers an example of one of the major functions of Big Data introduced in Chapter 1, its predictive function here understood as forecasting, alongside its descriptive, prescriptive, and discursive functions. It also provides an opportunity to discuss some key tools and concepts commonly used in machine-learning as well as merits and limits of these approaches to crime prediction for public policy.

Chapter 3 uses CDRs from Orange in Côte d'Ivoire made available as part of the 2013 Data for Development challenge—a modality that has been the hallmark of the field and contributed to developing Big Data communities over time—alongside meteorological data, with the goal of estimating whether weather could impact human mobility in ways that may violate the exclusion restriction in research using rainfalls as an instrument from economic conditions to assess the causal link between economic conditions and conflict. It presents a statistically significant relationship, suggesting that weather could affect conflict through other channels than economic conditions and casting doubt on the use of rainfalls as an instrument for economic conditions in these settings.

Chapter 4 uses the same dataset and attempts to predict, here in the sense of inferring or now-casting, the multi-poverty index based on DHS data to assess whether and how these kinds of data available at high levels of temporal and geographical granularities may help some of the data gaps that characterize and may impede the development of some of the poorest countries in the world, showing promising results.

Chapter 5 uses similar data as in Côte d'Ivoire but for Senegal, in conjunction with census data, to address the central issue of sample bias in big data by correlating estimated population size through cell-phone activity and census data. It proposes a novel approach to estimating biases in the data as a function of key demographic variables including age at different geographic levels.

Chapter 6 finally focuses on political economy implications of Big Data as an ecosystem and socio-technological phenomenon, with a focus on its prospects and requirements, including institutional, legal, ethical, and political. It shows that Big Data in general and CDR analytics in particular raise complex and contentious questions for social science research, policymaking, and societies at large—including around power dynamics, informed consent, fairness, and civic participation etc., which will require significant investments in developing adequate responses, including to human awareness and capacities.

It also argues, as does the overall conclusion, that Big Data and open algorithms notably can provide an entry and anchor point to challenge and improve the current state of the world by giving data emitters—citizens—greater control over the use of the data they generate in ways that could revive democratic ideals and principles and make it a potentially truly revolutionary force.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This dissertation is the direct result of about seven years of work during which I have benefited from the generosity, trust, and guidance of many people; it also builds on other past experiences and relations that I want to acknowledge.

First, I wish to thank my dissertation committee at Berkeley for their guidance, and their patience. I am particular indebted to my dissertation chair Ronald Lee, who encouraged and advised me along an untraditional graduate journey. As anyone who has had the privilege of being one of Ron's students knows, his generosity, humility, dedication, kindness, and knowledge are limitless. Jenna Johnson-Hanks showed an early interest in the topic and encouraged me to keep a focused eye on the ultimate objective of my research—this document—while also giving me the impetuous, through the example of her own research, to consider socio-cultural dimensions in my work; Ted Miguel also showed an early interest in the broad topic and was instrumental in turning my attention to the issue of sample bias, in addition to being a fantastic instructor in his Development Economics class.

I also want to thank Josh Goldstein for his encouragements during the last two years of this journey, Mike Hout for agreeing to be on my Qualifying examination committee, as well as John Wilmoth and David Reher, whose Historical Demography seminar, passion, and humor I enjoyed a lot. I would also like to thank Kenneth Wachter, especially for his words of personal support in January 2015 after the Charlie Hebdo attacks and his interest in my activity as a cartoonist.

At Berkeley too, I will forever be grateful to Monique Verrier and Liz Ozselcuk for their support. I would not have made it without Monique's help on (too) many occasions. Magali Barbieri, Robert Chung, Carl Mason, Carl Boe, and Gretchen Donehower were also supportive. I will fondly remember my cohort 31, especially Maia Sieverding and Sara Lopus and the inhabitants of the attic, Sarah Cowan, Reid Hamel, Maggie Frye, and Amal Harrati, and, in and out of the attic, my friend Jérémie Cohen-Setton. I also want to mention George Tapinos, who I know was a dear friend of the Department's and got me interested in economic demography while my Professor at Sciences Po in 1999 and 2000 before his premature passing in 2000.

A few people in particular gave me the opportunity to enter the 'Big Data space' and establish a base from which I was able to explore it. A heartfelt thanks and gratitude go to my fellow Data-Pop Alliance directors and co-directors, Patrick Vinck at Harvard, Alex 'Sandy' Pentland at MIT, Phuong Pham at Harvard, Claire Melamed and Emma Samman at ODI, and Linus Bengtsson at the Flowminder Foundation, as well as Erik Wetter and Andrew Tatem. I also want to thank our Research Affiliates, many of whom are also collaborators, co-authors and friends, especially Espen Beer Prydz, Yves-Alexandre de Montjoye, Beth Tellman, Fredrik Sjoberg, Romesh Silva, Thomas Roca, Lanah Kammourieh, Julia Manske, Bruno Lepri, Linnet Taylor, Sabrina Juran, Jacopo Staiano, Nuria Oliver, Simone Sala, Cathy O'Neil, Bessie Schwarz, Robert Darko Osei, Jay Ulfelder, and Emilio Zagheni, whose work is an inspiration. I am also especially thankful for the help of Gabriel Pestre over these past few years, as well as the rest of the Data-Pop Alliance team, Natalie Shoup, David

Sangokoya, Andrés Clavijo, Carson Martinez, Ana Lucia Martinez, Rodrigo Lara Molina, Fernanda Villaseñor, and Anna Swenson.

I will also remember the support I have benefited from other colleagues, professors, and friends since coming to the US in 2004, including, in no particular order, Paul Ladd, Antoine Heuty, Rathin Roy, Andrea Bubula, Adam Smith, Francisco Rodriguez, Ricardo Fuentes-Nieva, Haishan Fu, Nicolas de Cordes, Kenneth Cukier, Isabelle Huynh, William Hoffman, Alessia Lefébure, Amparo Ballivian, Tariq Khokhar, Judith Karl, John Ohiorhenuan, Zazie Schafer, Eva Kaplan, Katell Le Goulven, Lea Shanley, Karen Ballentine, Florent Peyre, Francesco Mancini, Jared Hatch, and Sara-Jayne Terp.

On a personal note, I want to especially thank my closest friends of nearly three decades Valéry Pratt, Christophe Moreau, and Michael Contet for their unflinching support and friendship, as well as many other friends in France, the US, Vietnam, and other countries, who have provided support in various ways over the past seven years and more, in the form of encouragements, questions, trips, laughs, and meals.

I of course owe a lot to my family, especially my parents, who—in addition to making the trip all the way from France for the graduation ceremony in May 2016—have backed and trusted me in everything I have ever decided and done, no matter how unusual or stress inducing—and are wonderful grand-parents. Many thanks also go to my older sister Anne, whose determination and dedication in her many academic and personal endeavors are an example.

I want to stress the central, defining role of my maternal grandparents from Brittany, both teachers who dedicated their lives to education, in shaping my personality and values. There are few days that go by where I don't find myself thinking about them and what they would have done in any given circumstance. I want to include in my thoughts the memory of my uncle Yannick, with whom I shared a lot, and acknowledge the role and support of my three cousins from Brittany, Nolwenn, Erwann, and Yann. I have been very lucky to have such a tight inner network of friends and family.

At the center of this network, I could not have completed my PhD without my partner of now more than twenty years, Julie, who has been supporting me for two decades (in both the English and French senses of the term) and gave birth to our twin daughters, Fleur and Louise, who could not have a better mum. I know the past few years have been rough, and I hope those to come will be easier now that this dissertation is out of the way.

Fleur and Louise, members of this first "data generation" I talk about, these few pages that took me away from you on so many occasions these past seven years—even before you were born—are dedicated to you. I love you more than I could imagine and hope you will live long and happy lives and follow your own paths driven by a fierce desire to learn, listen, speak out, share, challenge, create, and innovate to shape and improve the world you and future generations will grow up in.

Introduction: Origins and overview of this dissertation

I entered the 'Big Data' space in the fall of 2010, when it was then only emerging, by chance, but I would not say by accident.

Back in New York City on parental leave from my PhD after the birth of my twin daughters a few months earlier, I was about to be hired as a consultant for the United Nations Development Programme (UNDP), where I had previously worked for four years as a researcher before arriving at Berkeley, working and writing on fiscal policy for poverty reduction, post-conflict economic recovery, and migration. From 2000 to 2004, after graduating with an MA in Economic Demography and a BA in Political Science from Sciences Po Paris, I had worked for the French Ministries of Finance and Foreign Affairs in Hanoi, Vietnam, as a Technical Assistant on official statistics, public finance, and macroeconomics. In addition to being highly (trans)formative on a personal level, this experience shaped and fostered many topical and methodological interests that run through this dissertation. I was, for example, exposed to the notion of data mining for the first time in 2001, by a French economics professor involved in our projects. I also worked on official statistics, migration, and post-conflict recovery issues, with and in Vietnamese institutions. The political economy features of Vietnam were also of course fascinating. Once at Columbia University School of International and Public Affairs in 2004-05, I wrote my main MA research paper on the origins and effects of the 'household registration system' known as *Ho Khau* in Vietnam, and more generally focused on quantitative classes.

My assignment with UNDP in 2010 would have involved writing an econometrics and policy paper broadly focused on developing countries' vulnerability to the (then called) 2008-09 financial crisis (*i.e.*, ex-post). But I thought that vulnerability was best assessed at the micro level and the wealth of new kinds of fine-grained individual and community-level data now becoming more 'available'—or so I thought—would make this increasingly possible and desirable. I tried to argue that the paper should look beyond or below country-level data, but the response was no; we were going to compare the vulnerability of Mali vs. Madagascar as a whole.

I decided not to take the assignment, but said I would write a note about my views on vulnerability in this new context. A few weeks later I was approached by the Deputy Director of UN Global Pulse[1], a new unit in the Executive Office of the UN Secretary-General (set up in the aftermath of the 2008-09 financial and economic crisis), who had come across the note through a common friend in the UN system and asked if I would be interested in joining the team as a consultant. For the next 12 months, I worked on the unit's nascent research team, splitting my time between developing different 'proof-of-concept' projects (*e.g.*, using Twitter data to study perceptions of economic crisis-related stress the price of rice in Indonesia[2]), which were presented at the UN General Assembly in the fall of 2011, and writing the first report on "Big Data for Development," which was published in May 2012.[3]

This report helped me obtain opportunities in this new field, but more importantly it got me hooked. I became and remain fascinated by Big Data because it encapsulates almost everything I was ever interested in: social science questions, quantitative puzzles, methodological problems, as well as contentious political economy issues—on ethics, politics, power, civic engagement, and even art and design with ever powerful and creative data visualization techniques. I have since then dedicated my academic research and work more generally to understanding and advancing this new field of research and practice and hope to continue in the foreseeable future—especially as Director of Data-Pop Alliance[4], as well as a post-doctoral Visiting Scholar at MIT Media Lab and Visiting Fellow at Harvard, starting in the fall of 2016.

Importantly, I use the singular ('it') to talk about Big Data, written with capital letters B and D. As I will explain in Chapter 1, I distinguish Big Data and big data. In my opinion the latter—these new kinds of large, messy, and connected digital data sets described through the "3 V's" of big data for Volume, Velocity, and Variety in the early years—are simply the core and fuel of the former, which I refer to as an ecosystem of the "3 Cs" of Big Data for Crumbs, Capacities and Communities.[5] In my opinion, Big Data conceptualized as being beyond large datasets provides the foundation and is the expression of a new socio-technical phenomenon that warrants being discussed in revolutionary terms—as when people talk or write about the "Big Data Revolution"[6] or the "Data Revolution"[7] more broadly.

Over time, I have found myself in an interesting place: having been trained in traditional social science—I studied political science, macroeconomics, and economic demography at Sciences Po Paris then international affairs and development economics at Columbia before starting my PhD in the Department of Demography at Berkeley—I began writing reports

[1] www.unglobalpulse.org

[2] *Monitoring perceptions of crisis-related stress using social media data* 2011.

[3] Letouzé 2012.

[4] Data-Pop Alliance is a coalition on Big Data and Development I co-created with Patrick Vinck, Assistant Professor at Harvard, and have been heading since 2013. It was co-founded by the Harvard Humanitarian Initiative, the MIT Media Lab, and the Overseas Development Institute, recently joined by the Flowminder Foundation as its fourth core member. For more information see www.datapopalliance.org as well as an interview for the website KD Nuggets: http://www.kdnuggets.com/2015/04/interview-emmanuel-letouze-data-pop-alliance-human-rights.html

[5] MIT Media Lab (@medialab) 2016.

[6] Mayer-Schönberger and Cukier 2013.

[7] *A World That Counts: Mobilising The Data Revolution for Sustainable Development* 2014.

and papers, siting on panels, giving speeches, and answering interviews about ‘Big Data,’ a field where computer scientists, physicists, engineers, and a new breed of ‘data scientists’ reign. I never lied: I always stressed my very limited technical knowledge in all things ‘data science,’ but I picked up a few skills while working with people who mastered them, and we put our skills together.

Over the past six years, I have had the privilege and pleasure to write and talk about Big Data’s applications and implications for a variety of topics, including official statistics, ethics, law and politics, poverty, migration, conflict and crime, data literacy, individual and group privacy, climate change and resilience, algorithmic decision-making and open algorithms, the Sustainable Development Goals, monitoring and evaluation, among others. Some of this work has been mentioned in the *MIT Technology Review*, *The Economist*, and *The Wall Street Journal*, for example. I got to meet great researchers, practitioners, public officials, community organizers, and activists in close to 30 countries on six continents.

Often, I try to serve as a sort of connector and translator between two ‘worlds.’ I understand much more the world of the ‘bean counters,’ by which I refer to as traditional development research and policy—that of the UN, statisticians, demographers, economists, etc.—where I have now spent about twenty years, but I also increasingly understand the perspectives, incentives, and toolkits of the ‘geeks,’ who live and work in computer science departments, engineering schools, or large companie, including social media companies like Facebook and also telecom operators and banks, which are deemed to become ‘data companies.’ I try to tell the former: “there is really something big happening, it’s not just a fad,” while warning the latter to not “think you will solve global poverty by crunching numbers alone.”

Of course, eminent researchers and academics are firmly established in this space—the authors of the 2009 piece “Computational Social Science,”[8] such as Alex ‘Sandy’ Pentland, David Lazer, Gary King, Nicholas Christakis, and Deb Roy—make up the bulk of the fathers of this field. Younger academics such as Emilio Zagheni, Ingmar Weber, Yves-Alexandre de Montjoye, Augustin Chaintreau, Bogdan State, Daniela Witten, Tapan Parikh, Jenna Burrell, among others, are and will be central figures of its future. I do not have the pretension to be part of this academic league. I foresee my contribution as being more applied; I will want to have a foot in academia and the other with my toes spread across political forums, NGOs, civil society, and humanitarian action, between New York, Cambridge, Bogotá, Dakar, Kigali, Nairobi, Barcelona, Istanbul, Bangkok, Singapore, Sydney etc. But before I do that, I must first submit this dissertation.

This dissertation contains and aims to connect into a coherent piece my thoughts and writings of the past seven years focusing on Big Data’s applications and implications for the analysis of demo-economic processes. Interestingly, demographers have been relatively slow in entering the space, presumably because Big Data was and sometimes remains seen as a fad, relying on messy, unrepresentative data, when demographers perhaps more than any of their social scientists peers rely on solid data. At the same time, throughout its history—most evidently in the vast historical and formal demography literature with the

[8] Lazer, A. Pentland, et al. 2009.

contributions of Coale and Trussel,[9] Keyftiz,[10] Lee,[11] Wringley and Schofield,[12] Livi Bacci,[13] etc.—demography has been at the forefront of developing innovative methods to deal with new sources of data and social problems, while keeping its role as custodian of scientific soundness in population analysis. Over the past two years, things have changed, with an increasing number of demographers working on or in Big Data, and computer scientists realizing they lack and need the skills that demographers bring to the table.[14]

In the field itself, this dissertation falls at the intersection of—or at least has connections with—several disciplines in the social and computer sciences, but also humanities (demography of course, statistics, machine learning, but also politics and ethics) and touches on a range of themes (poverty, criminality, conflict, population density). A defining feature of Big Data is its inter-disciplinary or, as some say, anti-disciplinary, nature. Big Data blurs boundaries: both because the world has become increasingly complex and interconnected, such that all social issues are somehow related, not least under the influence of digital technology, and because making sense of these processes requires leveraging skill sets that no single person or even discipline can realistically claim to have.

As such, in this dissertation, I draw on several co-authored papers, many of which I was involved in a lead author capacity. I do not see this as a weakness or a challenge, but rather as a necessity and an opportunity. Working with these co-authors over the years has exposed me to new ways of thinking about and approaching problems, and to touch on themes that fall outside the sphere of 'traditional' demography—even economic demography. Reciprocally, I have typically been tasked with framing the research questions and empirical strategies of these papers. One of the contributions of these papers and this dissertation, I hope, is also to provide frameworks and parameters to think about these issues in a coherent and complex manner—for instance by proposing a conceptualization of Big Data that goes beyond a sole focus on the data, and a taxonomy of applications of Big Data.

This dissertation is organized as follows. Chapter 1 synthesizes the current state of my own thinking about Big Data for, or and, development and some key research strands in Big Data for population research. It aims to highlight the genesis, key features, and main sectorial applications as well as identify risks and requirements. The 4 subsequent chapters consist of empirical analyses in 3 main areas, relying on cell-phone data analysis: predicting crime in London using cell-phone data from Telefónica (Chapter 2), analyzing the relationship between weather and population movement in Ivory Coast (Chapter 3), and measuring poverty in Côte d'Ivoire (Chapter 4) and population density in Senegal (Chapter 5) using cell-phone data from Orange, with an attempt at estimating and correcting for sample bias. The final chapter discusses legal, ethical, political, and institutional considerations and requirements for the future of cell-phone data analysis and Big Data more broadly (Chapter 6).

[9] Coale and Trussell 1974.
[10] Keyfitz 1975.
[11] Lee 1974; Lee 1973.
[12] Wrigley and Schofield 1981.
[13] Livi Bacci 2003.
[14] Letouzé 2015.

Chapter 1

Genesis, Features, and Prospects of a Social Phenomenon

In less than a decade, the phenomenon of ‘big data’—or rather, ‘Big Data,’ as mentioned in the Introduction and discussed further below—has affected industries and activities from marketing and advertising to intelligence gathering and law enforcement, and, increasingly, academic research and policy–making. In this short period of time, it has stirred much excitement and skepticism. Is Big Data, what Andreas Weigend, former chief Scientist at Amazon, and others called a few years ago the “new oil,”[1] poised to be either a blessing or a curse for research and policy–making, and more broadly human development and social progress?

Let me get the semantics clear first. Throughout this chapter and dissertation, unless in direct quotes, I use ‘big data’ as a plural term to talk about big data as data or datasets and ‘Big Data’ as a singular term to talk about the phenomenon, or ecosystem, or industry—as discussed further below in some level of details.

In a best-selling book from 2013, optimists such as Kenneth Cukier and Viktor Mayer-Schönberger have called Big Data “a revolution that will change”—mostly for the better, according to them—“how we live, think and work,” as per the book’s sub-title.[2] Two senior World Bank officials expressed hope that big data may partly fix what they referred to as “Africa’s statistical tragedy,” and more broadly the dearth of reliable official statistics in some of the world’s poorest places.[3] But skeptics and critics have been more circumspect, and some plainly antagonistic: referring to Big Data as a “big ruse”[4] for example, as well as ‘big brother’ (or Big Brother), especially in the wake of the revelations by former US National Security Agency (NSA) contractor Edward Snowden of the nature and extent of the agency’s surveillance apparatus and practices.[5] Traditional social scientists and demographers

[1] Weigend 2013.

[2] Cukier 2013.

[3] Giugale 2012.

[4] Few 2012.

[5] MacAskill and Dance 2013.

in particular have focused attention on the many methodological questions and challenges that come with big data—chief of which is their statistical non-representativeness, because, as I have heard and been told many times, "not everybody has a cell-phone."

For any researcher or citizen interested in understanding and addressing social problems, Big Data does raise many hard questions. Some of the most critical ones cited for the intent and purposes of this chapter and dissertation include: What exactly is Big Data and where is it coming from? What are big data? How is Big Data expected to help advance knowledge, yield better 'insights,' inform public policies and programs—and decisions more generally? What does it mean for social science research in general and population science in particular? Will big data replace traditional sources of data and will Big Data replace traditional social science research in the foreseeable future?

1.1 Foundational years and facts of an eight year old phenomenon

In April 2014, in an overview article that Chapter 1 draws upon, I wrote that "Big Data, especially as applied to development and public policy issues, [was] in its intellectual and operational infancy."[6] Two-and-a-half years later, I tend to think, say, and write that Big Data is entering the age of reason—because it is roughly seven to eight years old—or a fast maturation phase. The body of research has expanded quickly and, importantly "the more radical positions of the early years have coalesced towards more consensual and balanced perspectives," which I am trying to summarize in this chapter.[7]

It is interesting and useful to start by providing a short timeline of the birth and rise of Big Data, especially in the realm of social science research and policy, and, in particular, as it pertains to economic development issues. One of the earliest mentions of the advent of this new phenomenon can be credited to computer scientist Joseph Hellerstein, now Professor at the University of California, Berkeley, who wrote about the "Industrial Revolution of Data" in November 2008.[8] In February 2009, as mentioned in the Introduction of this dissertation, a group of prominent American academics published an article titled "Computational Social Science,"[9] describing an emerging field that "leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors." The graphic accompanying the piece (see 1.1) captured the novelty of these new resources and approaches, particularly because it looks quite different from traditional bar charts, trend lines, and scatterplots to which we have long been accustomed.

Exactly a year later, *The Economist* ran in its printed version an article[10] about "The data deluge," with the sub-title "Businesses, governments and society are only starting to tap

[6] Letouzé 2014.

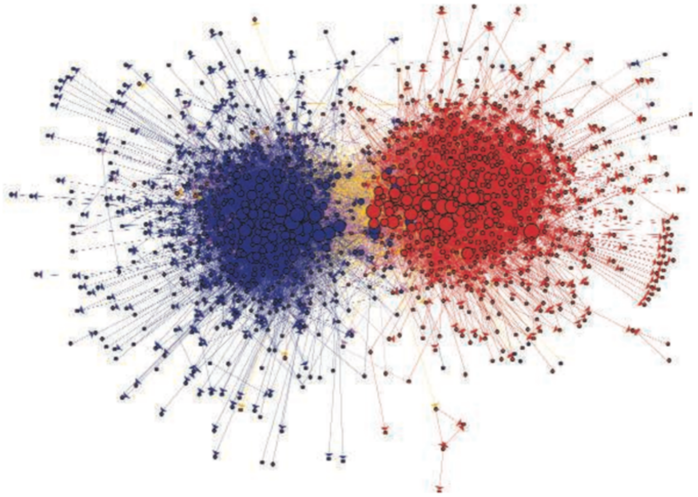
[7] Letouzé 2014.

[8] Hellerstein 2008.

[9] Lazer, A. Pentland, et al. 2009.

[10] "The data deluge" 2010.

Figure 1.1: Data from the Blogosphere



Shown is a link structure within a community of political blogs (from 2004), where red notes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it.

Source: David Lazer, Alex Pentland, et al. "Computational Social Science." In: *Science* 323.5915 (Feb. 2009), pp. 721–723. ISSN: 0036-8075. DOI: 10.1126/science.1167742. URL: http://science.sciencemag.org/content/323/5915/721.

its vast potential." The accompanying illustration (See 1.2)—a man in suit filtering diluvial data through an umbrella to provide only the necessary amount and kind of resources to a plant (possibly representing the filtering of certain types of information into an economy, company, or society)—has become near-iconic.

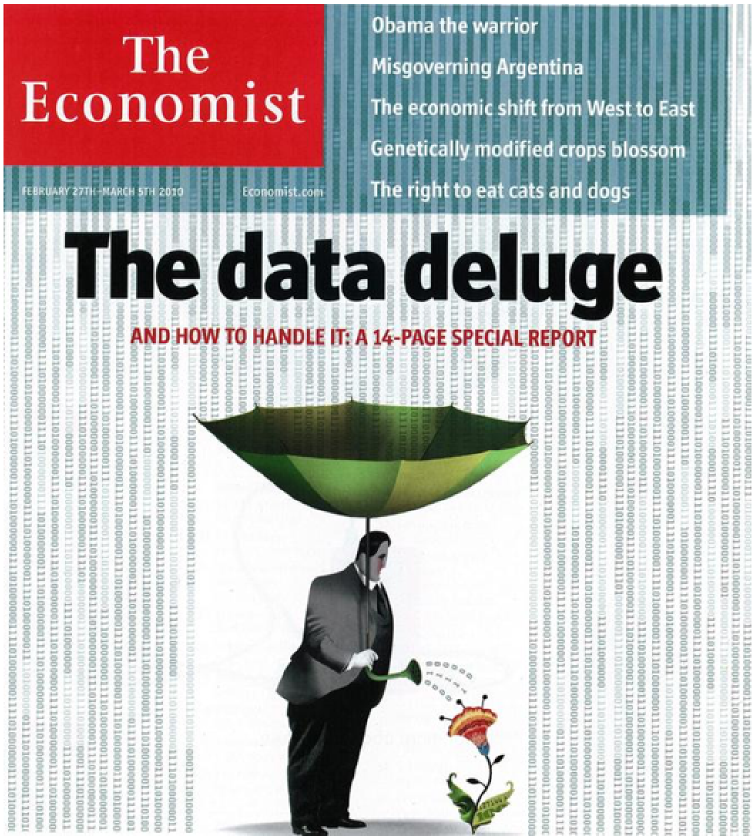
In 2011, Harvard Professor Gary King—a co-author of the "Computational Social Science"[11] piece mentioned in the introduction—published another seminal article in *Science* titled "Ensuring the data-rich future of social science",[12] with its sub-title containing the following lines:

> Massive increases in the availability of informative social science data are making dramatic progress possible in analyzing, understanding, and addressing many major societal problems. Yet the same forces pose severe challenges to the scientific infrastructure supporting data sharing, data management, informatics, statistical methodology, and research ethics and policy, and these are collectively holding back progress.

[11] Lazer, A. Pentland, et al. 2009.

[12] King 2011.

Figure 1.2: “The data deluge” cover from *The Economist* (2010)


The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
FEBRUARY 27TH–MARCH 5TH 2010
Economist.com
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

The policy world—especially the circles working on economic development—also started paying attention during those years. Three major reports came out in 2011 and 2012: in May 2011 by the McKinsey Global Institute, titled: “Big data: the next frontier for innovation, competition and productivity”;[13] in January 2012 by the World Economic Forum, titled “Big Data, Big Impact: New Possibilities for International Development;”[14] and in May 2012, by UN Global Pulse, titled “Big Data for Development: Challenges and Opportunities”,[15] which I authored while working there in 2011 as a Senior Development Economist. The UN Global Pulse report seems to have been the document actually coining the phrase “Big Data for Development,” which has since then been widely used—and which, over time, I started having issues with, as discussed further below.[16]

During and since those foundational years, publications and initiatives about ‘Big Data for development’ or ‘data science for social good,’ by their number and diversity, have become

[13] Manyika et al. 2011.

[14] World Economic Forum (WEF) 2012.

[15] Letouzé 2012.

[16] In a nutshell because the term “for” suggests a one-way, rather mechanistic, relationship between both parts of the phrase, when I think what is at play and needed is an iterative process and dialogue between them.

a source of big data themselves, such that it is impossible to mention and reference all or even most of them. Many will be referenced in the rest of this dissertation, but two merit an early mention. The first one is an article by danah boyd[17] and Kate Crawford titled “Six Provocations for Big Data,”[18] a critical review of Big Data’s implications and risks for social science research and societies at large. These “six provocations,” which I broadly agree with as I will attempt to illustrate in the rest of this chapter and dissertation, are:

1. “Automating Research Changes the Definition of Knowledge”
2. “Claims to Objectivity and Accuracy are Misleading”
3. “Bigger Data are Not Always Better Data”
4. “Not All Data Are Equivalent”
5. “Just Because it is Accessible Doesn’t Make it Ethical”
6. “Limited Access to Big Data Creates New Digital Divides”

The second piece worth mentioning here is an article published in Wired magazine in June 2008—a few months before Joseph Hellerstein’s quote cited above—written by journalist and commentator Chris Anderson, with the provocative title: “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.”[19] The following paragraph from Anderson’s article in particular has received much attention and continues to generate contentious debates some eight years later:

> This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.[20]

Further according to Anderson:

> [...] faced with massive data, this approach to science—hypothesize, model, test—is becoming obsolete. [...] There is now a better way. Petabytes allow us to say: “Correlation is enough.” We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.[21]

[17] Writing “danah boyd” without capital letters is not a misspelling; this is how she has decided and asked to write her name for several years.

[18] boyd and Crawford 2011.

[19] Anderson 2008.

[20] Anderson 2008.

[21] Anderson 2008.

It is hard to say whether Anderson really meant that "[w]ith enough data, the numbers speak for themselves"—near blasphemy for traditional social scientists trying to understand the world by unveiling causal processes backed by theoretical models rooted in observation—or what he meant exactly. But his arguments definitely get to some critical issues and questions raised by Big Data. My personal sense has long been that he stretched the envelope to spark a debate. I will come back to the issue of correlations versus causation and related questions about the role of theory in the future of social science research at the end of this chapter, and of this dissertation.

This attention and appeal to Big Data stemmed from two sets of factors that can be simply referred to as supply and demand factors. The supply side is evidently the analog-to-digital transition and early evidence that Big Data could help fill some of the gaps in current knowledge—including provide alternative measures of poverty and welfare;[22] this is in part due to a much cited 2009 paper that found that light emissions picked up by satellites could track economic activity and suggested they could supplement national accounting in data-poor places.[23] In recent years, the literature on call detail records (CDRs) has been especially rich, with new findings on the potential of analyzing CDRs to study migration patterns, socioeconomic levels, and disease spread, among others.[24]

The demand side, or the pull factor, has been the thirst for data on development issues—with the basic rationale being that economies and societies should be steered by policymakers relying on better navigation instruments and indicators that let them design and implement more agile and better targeted policies and programmes—and ever more so when so much more of them seem to be available. Both aspects will be covered in greater length and depth in the rest of this dissertation, especially Chapter 4.

Nevertheless, Big Data could still be a buzz or a bubble; old news marketed as a something revolutionary. So what exactly are we talking about and dealing with? How new and how big are these datasets?

1.2 How new is big data—and how big is it?

There is still no single agreed upon definition of big data or Big Data, and there may never be one. Although some argue that "to define is also to limit,"[25] I think it is useful to have some broad common conceptualization of what we mean by it. Critically, my own perspective has been and is that neither big data nor, even more so, Big Data, should be reduced to large datasets. Instead of focusing on the amount of raw material, I find it more useful to think in qualitative terms about the nature of the material, the entire ecosystem around it, and the larger socio-technological phenomenon created by as well as fueling this

[22] Letouzé 2013.

[23] Henderson, Storeygard, and Weil 2009.

[24] For the most comprehensive review of the relevant literature to date, see Blondel, Decuyper, and Krings 2015

[25] Cukier 2013.

ecosystem. In doing so, I highlight the key essential features of big data and distinguish big data from Big Data.

In the early years described above, much attention was paid in the mainstream press and the common discourse to the so-called "3 Vs of big data"[26]—which stand for Volume, Variety, and Velocity. These are indeed characteristics of the kinds of data and datasets that most people refer to as big data. As demographers, thinking of the these datasets as a human population, we understand that these features are correlates of one another: these datasets, generated from many different sources at high frequency, are large and growing. Other "Vs" were also added, including Value, Variability, Verification, Virality, and even Viscosity.[27]

Deeper thinking around the nature and novelty of these data has happened in the broad field of social science research and policy, some of which building on the 3 Vs,[28] and others moving away from this conceptualization. Alex 'Sandy' Pentland, for instance, differentiates social media data from credit card or cell-phone transaction data. He refers to the latter as "the little data breadcrumbs that you leave behind you as you move around in the world," as opposed to Facebook posts for instance, which can be "edited according to the standards of the day."[29] By this he means, notably, that while you can claim on Facebook to be boycotting Amazon because it has put your local bookseller out of business, your credit card transaction may be saying otherwise.[30]

Still considering the nature of these data, in UN Global Pulse's White Paper, I proposed two taxonomies according to which big data could be conceptualized and described: one taxonomy based on five key, namely "Digitally generated," "Passively produced," "Automatically collected," "Geographically or temporally trackable," and "Continuously analysed features;"[31] and another on four main sources, namely "Data exhaust," "Online Information," "Physical Sensors" and "Citizen Reporting or Crowd-sourced Data." [32]

[26] Soubra 2012.

[27] For discussion, see notably: Popescu 2011; R. ' Wang 2012; Beulke 2011.

[28] Burrell 2012.

[29] A. Pentland 2012.

[30] Both facts are potentially interesting from a research perspective—including because it may not even be you using your credit card to buy books on Amazon; the same way it may not be you using your phone.

[31] "(1) Digitally generated—*i.e.*, the data are created digitally (as opposed to being digitised manually), and can be stored using a series of ones and zeros, and thus can be manipulated by computers; (2) Passively produced—a by product of our daily lives or interaction with digital services; (3) Automatically collected—*i.e.*, there is a system in place that extracts and stores the relevant data as it is generated; (4) Geographically or temporally trackable—*e.g.*, mobile phone location data or call duration time; (5) Continuously analysed—*i.e.*, information is relevant to human well-being and development and can be analysed in real-time" A. Pentland 2012, p. 15.

[32] "Data Exhaust—passively collected transactional data from people's use of digital services like mobile phones, purchases, web searches, etc., and/or operational metrics and other real-time data collected by UN agencies, NGOs, and other aid organisations to monitor their projects and programmes (*e.g.*, stock levels, school attendance); these digital services create networked sensors of human behaviour; (2) Online Information—web content such as news media and social media interactions (*e.g.*, blogs, Twitter), news articles obituaries, e-commerce, job postings; this approach considers web usage and content as a sensor of human intent, sentiments, perceptions, and want; (3) Physical Sensors—satellite or infrared imagery of

Over time, and this is the position taken in this dissertation, I have stopped considering the fourth category of "Citizen Reporting or Crowd-sourced Data" as a kind of big data, because these are typically not very large, and, much more fundamentally, because they are actively generated.The active versus passive dichotomy with respect to crowdsourced data has been discussed a great deal in the geography literature. The distinction made is whether people are actively providing information about themselves for a specific purpose, or are at least aware of the data collection and don't object, or whether they are unaware of and may object to being observed/tracked and have their information collected and used for purposes other than they originally intended.[33] I now prefer to focus on the first three, as discussed in greater depth below. But this gets to a fundamental feature of big data. One of their main defining features and what accounts to large extent for their novelty, potential, and challenges, is to be, as put by a senior official from the US Bureau of Labor Statistics, "nonsampled data, characterized by the creation of databases from electronic sources whose primary purpose is something other than statistical inference."[34]

Fundamentally, the bulk of what makes up big data are digital, machine-readable, passively generated data about and by people, as the by-product of their use of digital devices and web-supported tools and platforms—most of which were unavailable five or ten years ago. It is also worth noting that a significant portion is unstructured data—such as YouTube videos and the text of a Facebook post—that do not fit in rows and columns. Structured data, on the other hand, fit in rows and columns; this means they are answers to questions the data collectors want answered. The implications of this fact will be addressed later in this chapter.

However, and critically, the distinction between passively versus actively generated data and consequently what data may versus may not 'qualify' as big data will become increasingly blurry. One case is Twitter data, as the social platform has been increasingly used for purposefully sharing real-time information in crisis contexts—for example, with dedicated hashtags after terrorist attacks to help people find shelter. Electronic medical records are another kind of data that are arguably 'actively' created and collected for some kind of analysis and it would seem odd to exclude them from the universe of big data. Actually, as hinted above through the reference to structured data and explained further below, even CDRs are collected 'actively' by telecommunication (telecom) operators, usually for billing purposes. But these data have secondary uses that were often not considered at the data collection stage. Other data may join big data all of sudden, as when large corpuses of analog data such as books are digitized[35] to allow for computational analysis of semantic and cultural

changing landscapes, traffic patterns, light emissions, urban development and topographic changes, etc; this approach focuses on remote sensing of changes in human activity;(4) Citizen Reporting or Crowd-sourced Data—Information actively produced or submitted by citizens through mobile phone-based surveys, hotlines, user generated maps, etc; While not passively produced, this is a key information source for verification and feedback." Letouzé 2012, p. 16

[33] I am grateful to Lea Shanley for having pointed this out.

[34] Horrigan 2013.

[35] Jean-Baptiste Michel et al. 2010.

trends as reflected historical writings.[36]

In a chapter for a report on technology and conflict prevention published in 2013, I and other co-authors defined big data "as the traces of human actions passively picked up by digital devices, or as the digital translation, understood in its literal sense, of human actions."[37] By "human actions" we meant and I mean here actions such as moving around, making a purchase or a phone call, researching a word online, having chopped down a hundred trees (which can be captured from satellite imagery), publishing a blog post or a piece of news online, typing and sending a tweet, or updating a Facebook status—the true intent or reliability of which may not be known. The essential features of the resulting digital traces in the digital age are that they are left as actions unfold in real-time,[38] are about what people do,[39] and allow connecting numerous smaller heterogeneous, unstructured and structured data streams.[40]

So the term 'big data' is actually largely misleading—perhaps a misnomer: size is not their defining feature: even a small 'big data' dataset is big data because it doe not stem from controlled processes like surveys and statistical imputations undertaken by official bodies or researchers that design them for a specific purpose. In contrast, the entire set of all censuses ever conducted in the world would not qualify as big data—no matter how large it may be. The true novelty is qualitative more than quantitative, and political economy, ethical, and legal issues—about informed consent and control notably—arise from this qualitative shift. But still, big data are pretty big, because every interaction with a digital device or service leaves a digital trace. Interestingly though, it is very hard to assess how much data are produced and how much are stored. According to a much-cited statistics from 2012,*"90% of the data available today had been produced in the past two years alone."*[41] There were and remain issues with the statement. For one, it is difficult to source—such that a Quora entry is actually "Who said 90% of data ever created was created in the last 2 years?"[42] Some sources mention "information" instead of "data" in the quote. Second, and relatedly, the methodology behind the assertion is unknown. Third, it says little about the nature and content of the data considered; most likely digital data, which is hard to compare with analog information such as traditional (paper) books. For example, a ten-minute YouTube video of people falling in different ways may be quite heavy in terms of data bytes but arguably contains less information about the human condition than John Steinbeck's Grapes of Wrath. Fourth, for anyone interested in development issues and the broader implications of the phenomenon, the figure doe not distinguish between geographic 'sources' of these data. Could it be the

[36] See notably Grimes 2013 and Aiden and J.-B. Michel 2013.

[37] Letouzé, Meier, and Vinck 2013.

[38] However, it must be noted that "for the purposes of global development, 'real-time' does not always mean occurring immediately. Rather, 'real-time' can be understood as information which is produced and made available in a relatively short and relevant period of time, and information which is made available within a timeframe that allows action to be taken in response *i.e.* creating a "feedback loop." Letouzé 2012.

[39] DeAngelis 2013.

[40] This last point was also highlighted by Lea Shanley.

[41] Frank 2012; Dragland 2013.

[42] Srinivasamurthy 2013.

case that 99% of these data are generated in and about developed countries? And finally, knowing that this may result from our collective production of "2.5 quintillion bytes of data being produced every day" in 2012 according to the same sources says little to the layman about the magnitude and significance of the phenomenon.

According to my own research of primary sources, I estimated in 2014 that between 2012 and 2014, over 1.2 zettabytes of data had been produced—1021 bytes, enough to fill 80 billion 16GB iPhones that would circle the earth more than 100 times. As demographers, we understand the relationship between the growth and 'age structure' of the global population of data. We can use standard demographic techniques to relate the rate of growth to the age structure—something that Emilio Zagheni has done for Twitter data for example.[43]

What is to the best of my knowledge the most scientifically sound method to address the question of the 'size' of big data has been developed by Hilbert and López.[44] In a co-authored paper published in 2015, we used this methodology to estimate and characterize the size of big data in terms of the world's *global storage capacity and the telecom capacity to access this storage* ('the cloud').[45] We estimated the world's technological capacity to store information had increased with a compound annual growth rate (CAGR) of 31% during the three decades between 1986 and 2014 (from 2.6 exabytes to 4.6 zettabytes), while the world's installed telecom capacity had grown with a CAGR of 35% during the same period, from 7.5 petabites to 25 exabits)—meaning they doubled roughly every two years over the period.

Another source of data about big data is Cisco,[46] which found that global mobile data traffic had grew almost 70% percent in 2014, reaching 2.5 exabytes per month at the end of 2014, up from 1.5 exabytes per month at the end of 2013. In 2012, for the first time, more than 50% of the traffic was due to mobile video traffic. Forecasting the mobile network through 2019, Cisco estimates that mobile data traffic will reach the following milestones within the next five years (the choice of highlights and wording are mine):

1. Global mobile data traffic will increase nearly tenfold at a compound annual growth rate of 57 percent from 2014 to 2019 to close to 25 exabytes per month by 2019;

2. By 2019 there will be nearly 1.6 mobile-connected devices per capita (11.5 billion) by 2019, (including M2M modules);

3. By 2019, more than half of all devices connected to the mobile network will be "smart" devices (up from 26 percent in 2014) and 97 percent of mobile data traffic will originate from these smart devices by 2019, up from 88 percent in 2014;

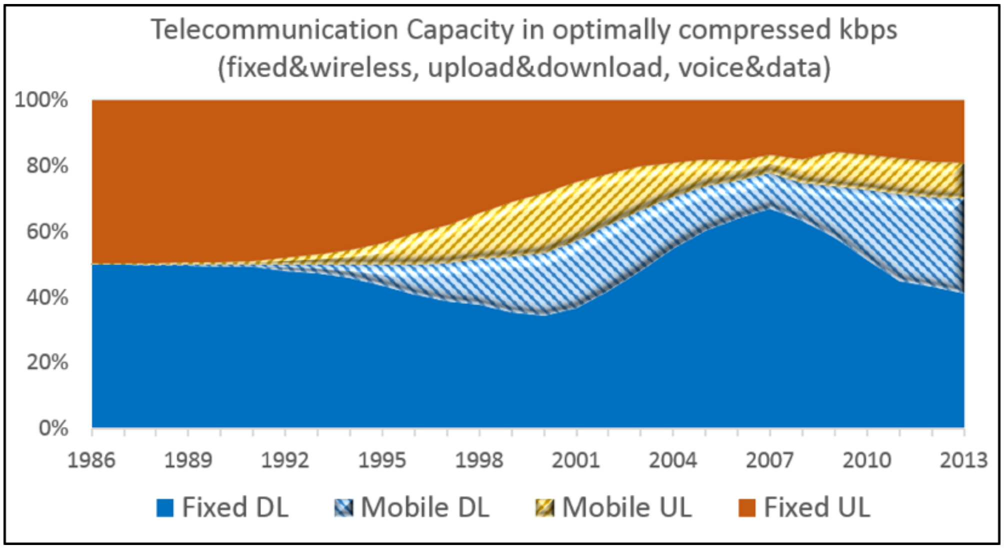
[43] During a workshop on Social Media and Population Research at the Population Association of America meeting in 2016.

[44] Hilbert and Lopez 2011; this methodology has been adopted by others, including: International Telecommunication Union (ITU) 2012.

[45] Actually, according to Hilbert and Lopez (2011) also the installed capacity to compute information, but quantifying computational capacity is rather tricky (given lack of agreement on useful metrics) and unfortunately no updated numbers exist.

[46] Cisco 2016.

Figure 1.3: Telecommunication capacity in optimally compressed kbps per uplink and downlink (with Martin Hilbert)



4. Tablets will exceed 10 percent of global mobile data traffic by 2016;

5. 4G traffic will be more than half of the total mobile traffic by 2017;

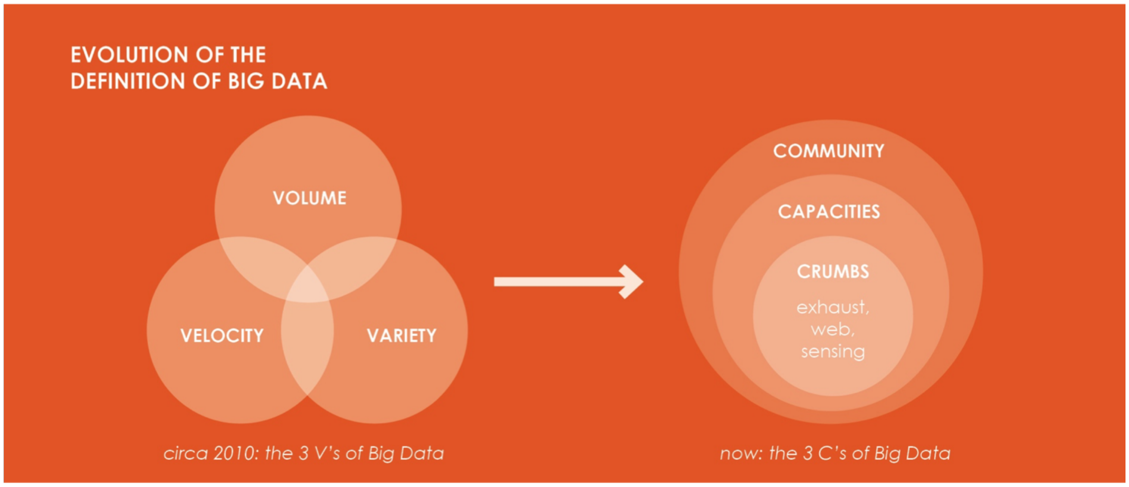
6. Mobile network connection speeds will increase more than twofold by 2019, reaching 4.0 Mbps by 2019;

7. By 2019, a 4G connection will generate 10 times more traffic on average than a non—4G connection;

8. Nearly three–fourths of the world's mobile data traffic will be video by 2019. Mobile video will increase 13–fold between 2014 and 2019, accounting for 72 percent of total mobile data traffic by the end of the forecast period; and

9. The Middle East and Africa will have the strongest mobile data traffic growth of any region with a 72% compounded annual growth rate, followed by Central and Eastern Europe (71% and Latin America and the Caribbean at 59%).

Figure 1.4: The Evolution of the definition of Big Data



1.3 Beyond big data: features and function of Big Data as an ecosystem of 3Cs

In recent years I have proposed a framework called "the 3 Cs of Big Data" to aim to capture and convey why and how I think of Big Data as an ecosystem[47] rather than merely as large datasets with specific features. The 3 Cs of Big Data stand for Big Data 'Crumbs,' 'Capacities,' and 'Communities,' represented by 3 concentric circles of increasing sizes. I think the framework is richer and more useful than the 3 Vs of Big Data of Volume, Velocity, and Variety; one limitation of the 3Vs is their exclusive focus on Big Data as data. Another is their emphasis on Big Data being essentially a quantitative, rather than a qualitative, shift.

First, 'Crumbs,' refers to "digital breadcrumbs" used by Pentland to describe "the little data breadcrumbs that you leave behind you as you move around in the world." A. Pentland 2012 In other words, the crumbs of Big Data are 'the' big data. In my taxonomy, the crumbs come in 3 main different types. One kind is small, 'hard,' structured data that can be easily quantified and organized in columns and rows for instance for systematic analysis. CDRs are the most widely known and used type of 'crumbs.' CDRs are metadata (data about data) collected by telecom operators that capture subscribers' use of their phones — including an identification code and, at a minimum, in the case of a cell—phone, the location of the phone tower that routed the call for both caller and receiver — and the time and duration of call. Large operators collect over six billion CDRs per day.[48] Strictly speaking, this first kind of data is what Pentland calls "digital breadcrumbs;" but I choose to use the term 'crumbs' to refer to the other two main kinds of data I consider as constituting most of the universe of

[47] MIT Media Lab (@medialab). *"Big data [is] an ecosystem," says @ManuLetouze of @datapopalliance, a global coalition that includes the @medialab http://mitsha.re/BOuA302mdVf.* [Tweet]. July 2016. URL: `https://twitter.com/medialab/status/755071681559949313`

[48] Letouzé, Vinck, and Kammourieh 2015.

big data.

A second kind of data ‘crumbs’ in my taxonomy includes videos, documents, blog posts, and other social media content. As mentioned above, these are unstructured data that are harder to analyze in an automated fashion (as they do not come in rows and columns), and they are also more subject to their authors’ editorial choices (they can be “edited according to the standards of the day”[50]). This may limit their research potential or lead to wrong conclusions if taken at face value, but these kinds of data have become key to studying people’s sentiments, desires, perceptions, beliefs, etc. Importantly, they require being turned into structured data to be subjected to automated analysis–for instance the number of pixels in a photo, or the number of times a key word appears in a book. This process is one where questions are asked to and about the initial data that reflect human intentions the same way CDRs reflect what telecom operators need and want to know. An interesting consequence is that unstructured data lend themselves to many types of research questions, including the possibility of challenging the kinds of structured data that are created out of them.[51]

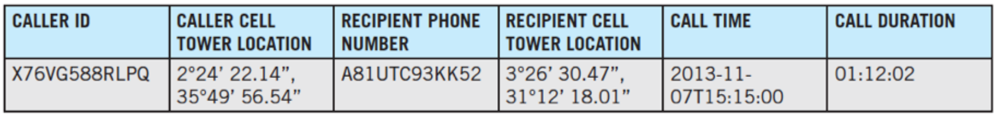
A third kind is gathered by digital sensors, either physical or remote. The former are devices physically installed or plugged to pick up human actions—such as electric meters, or, increasingly, wearable devices. Again, the data that is collected by these devices are collected for a specific purpose, but they may end up being used for other purposes as well—for which they were not initially collected; making them part of big data in the qualitative sense. Another kind comes from remote sensing—from video cameras or satellites for example. The use of geospatial data by hydrologists and climatologists may have been one of the earliest cases of Big Data use.

Rainfall or temperature data, or price data, are certainly big, high frequency, low granularity data. And they do have important bearing on human lives and ecosystems, and, as such, are integral parts of attempts at modeling and understanding these ecosystems. But, for definitional coherence, I generally choose to call these *contextual*(big) data not big data simply because they are not “digital traces of human actions,” even if and when human actions impact their patterns and trends—as in the case of prices or even rainfalls or temperatures. To be clear, they are not entirely exogenous to human actions—prices much less so than climate data—but they are not direct digital translations of human actions. One can certainly see the qualitative difference here, which has bearing on how much policy can affect their underlying determinants. There is also a quantitative difference in how ‘big’ either type can become: the growth, actual and potential, of big data as we define it is in

[50] A. Pentland 2012.

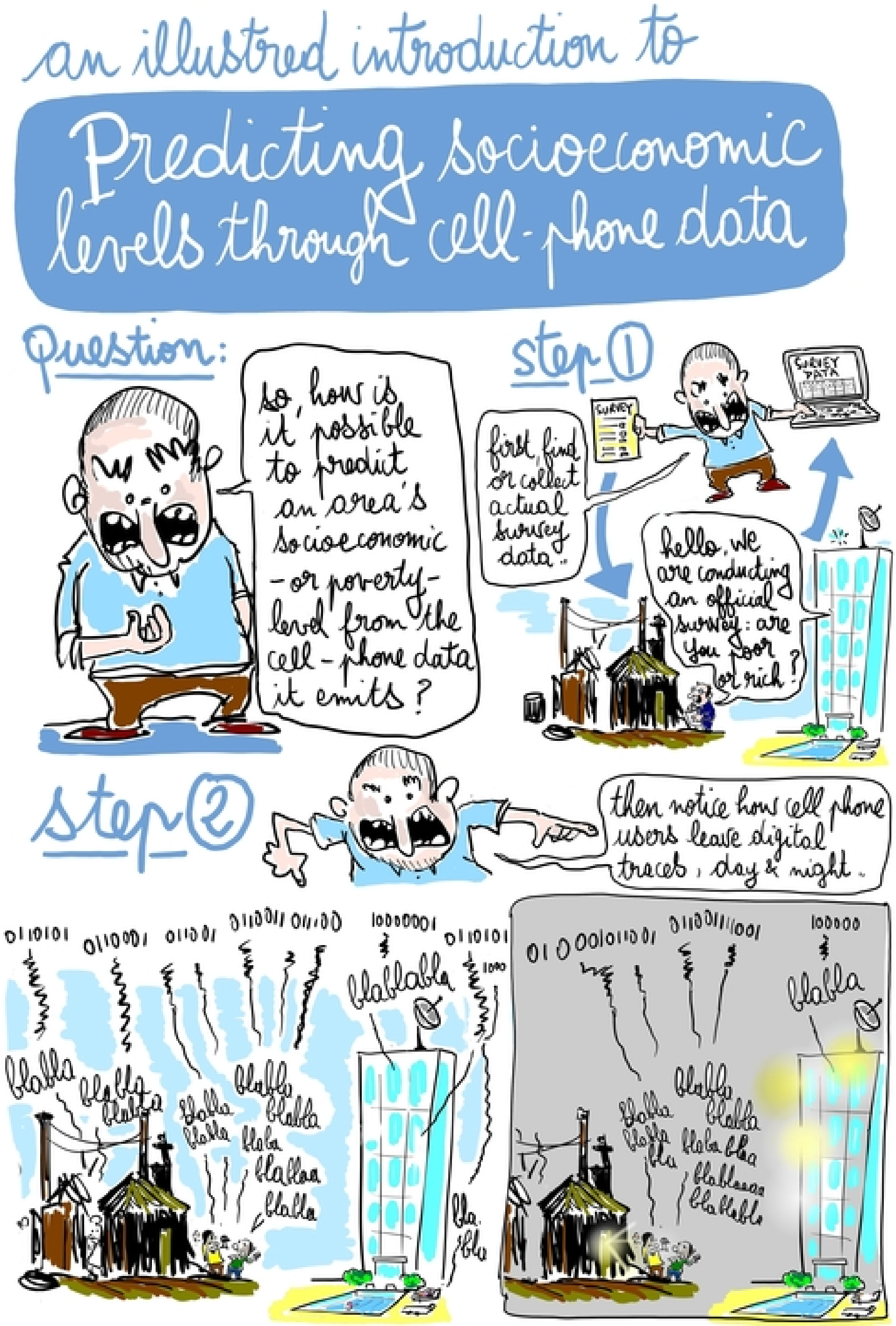
[51] I am grateful to Alex ‘Sandy’ Pentland for several discussions on this topic.

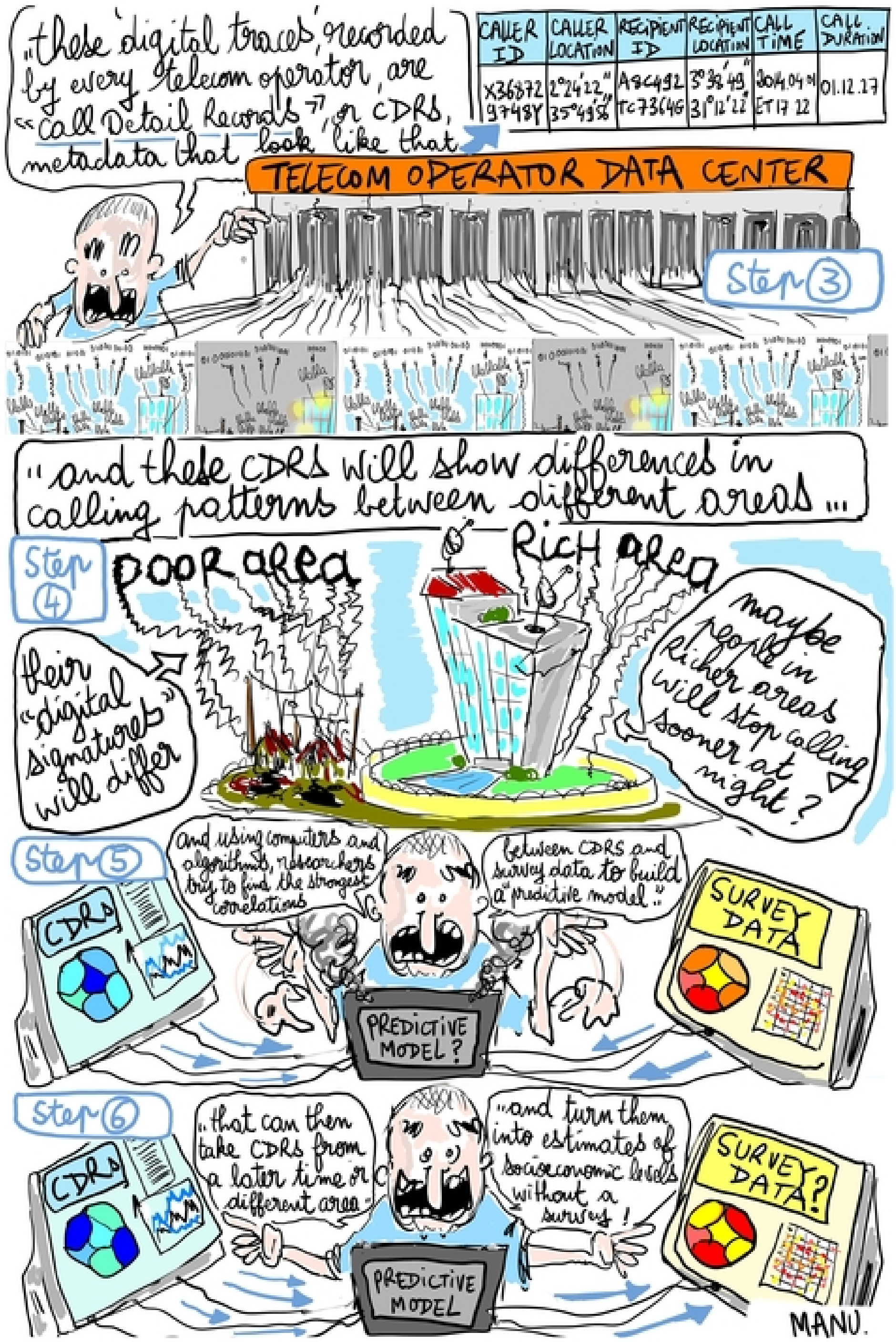
Figure 1.5: What Call Detail records (CDRs) look like



CALLER ID	CALLER CELL TOWER LOCATION	RECIPIENT PHONE NUMBER	RECIPIENT CELL TOWER LOCATION	CALL TIME	CALL DURATION
X76VG588RLPQ	2°24’ 22.14”, 35°49’ 56.54”	A81UTC93KK52	3°26’ 30.47”, 31°12’ 18.01”	2013-11-07T15:15:00	01:12:02

Figure 1.6: An illustrated introduction to predicting socio-economic levels through cell-phone data


an illustred introduction to
Predicting socioeconomic levels through cell-phone data
Question:
so, how is it possible to predict an area's socioeconomic -or poverty- level from the cell-phone data it emits?
step 1
first, find or collect actual survey data..
SURVEY
SURVEY DATA
hello, we are conducting an official survey: are you poor or rich?
step 2
then notice how cell phone users leave digital traces, day & night.
0110101
0110001
011001
01100110
0110
10000001
0110101
1000
blablabla
blabla
blabla
blabla
blabla
blabla
blabla
blabla
bla
blablaa
blabla
bla
bla
0100010110 01
01100111001
100000
blabla
blabla
blabla
blabla
blabla
bla
blabla
blaa
blablaaa
blablabla


"these 'digital traces' recorded by every telecom operator, are «Call Detail Records», or CDRS, metadata that look like that
CALLER ID
CALLER LOCATION
RECIPIENT ID
RECIPIENT LOCATION
CALL TIME
CALL DURATION
X36872 97487
2°24'22" 35°49'58"
A8C492 TC73646
3°38'49" 31°12'22"
2014.04.11 ET17 22
01.12.17
TELECOM OPERATOR DATA CENTER
Step 3
"and these CDRS will show differences in calling patterns between different areas...
Step 4
POOR area
RICH area
their "digital signatures" will differ
maybe people in Richer areas will stop calling sooner at night?
Step 5
and using computers and algorithms, researchers try to find the strongest correlations
between CDRS and survey data to build a "predictive model"
CDRs
SURVEY DATA
PREDICTIVE MODEL?
Step 6
.. that can then take CDRS from a later time or different area
"and turn them into estimates of socioeconomic levels without a survey!
CDRs
SURVEY DATA?
PREDICTIVE MODEL
MANU.

First published in: *Big data for development: Facts and figures* (2014).[49]

all likelihood, significantly greater than that of these contextual big data, thus leading, in conjunction with their qualitative difference, to greater opportunities and challenges to affect human ecosystems.

I hope that what broadly characterizes and qualifies as 'big data' is reasonably clear at this point—with the caveat that there is no one hard criterion—expect perhaps the fact of being machine-readable. Over time, indeed, boundaries will blur, with among other trends, more and more and soon all administrative records and personal health data being collected digitally and falling under the big data umbrella[52]. Ultimately, my hunch is that the qualifier 'big' will disappear and what will remain and keep people and researchers interested and talking will just be data.

The second C stands for Capacities—tools and methods, software and hardware, human, technical and institutional capacities. This is what Gary King meant when he wrote "Big Data is not about the data."[53] These capacities include powerful computers, parallel computing infrastructures, as well as visualisation techniques, families of algorithms, machine-learning, and deep learning techniques that are able to look for and unveil patterns and trends in vast amounts of complex data. Soon, with artificial intelligence advances we are told, we may be facing nothing less than "the end of code," when "we won't program computers. We'll train them like dogs."[54]

In the 3C model, the crumbs—the data—are only the raw material; it would be inert if it were not for those capacities to not only collect and store them, but also to clean, prepare, and analyze them. To use a historical analogy, the data are the coal, while the capacities are the machinery and factory. Reducing Big Data to big data is akin to equating the Industrial Revolution with coal; it is largely missing the point. But while technological capacities are progressing at a impressive pace, many other types of capacities will need to be developed in the next few years; including methods to better account for sample bias, and human capacities broadly. The latter are generally discussed under the concept of 'data literacy,' an important topic which I will discuss in the final chapter of this dissertation.[55]

The third C stands for Communities, as the outer encompassing circle. It brings up the human, societal, and political economy. Big Data is also constituted of a 'movement' of individual and institutional stakeholders that operate largely outside of traditional policy and research spheres, multidisciplinary teams of social and computer scientists. These individual and groups have incentives, objectives, skills, and constraints, that ought to be taken into consideration to understand what Big Data is, what it can do, and how it could be shaped to yield positive outcomes. The political economy of Big Data can simply not be adequately captured without recognizing its political and economic stakeholders.

The most visible members of the Big Data community can be found at data science

[52] See in particular The Kavli Foundation and New York University's Institute for the Interdisciplinary Study of Decision Making n.d.

[53] King 2013.

[54] Tanz 2016.

[55] Data-Pop Alliance 2015a.

'meetups' around the globe,[56] organize and/or participate in hackathons in Nairobi[57] or virtual 'data challenges' organized by telecom operators such as Orange,[58] Telefónica,[59] Telecom Italia,[60] or banks such as BBVA.[61] But the Big Data community more broadly includes all data emitters and users—*i.e.*, potentially the entire world population. There are very few and will be increasingly few human beings that should not be considered as being part of the Big Data community by virtue of being represented in a dataset. This has major political implications that I will partially address at various points in the rest of this dissertation, especially its last chapter, but that go far beyond its scope.

Now that we have a clearer and richer conceptualization of what Big Data is—this ecosystem of 'crumbs,' 'capacities,' and 'communities'—we can ask what good (or bad) it can do in general, and then in the more specific case of social science, including population research.

The first taxonomy I proposed in the UN Global Pulse paper distinguished the 'early warning' uses from 'real-time awareness,' or from 'real-time monitoring.' Later I and other co-authors proposed a taxonomy contrasting a descriptive function (such as a real-time maps), from predictive and prescriptive applications.[62] Importantly, the term predictive recoups both a forecasting function and an inference function. In the former case it is about estimating what may happen next, while the later is about what is happening now—which has been referred to as now-casting for many years. One could imagine predicting births this month from sales of baby formula.

Recently, with co-authors, we added a fourth function—a *discursive* function.[63]

To summarize, I think Big Data can serve the following four main functions:

1. One is a *descriptive* function—via maps, descriptive statistics, etc —– this may include, for example, the visualization of cell phone activity;

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, for instance, CDR–based variables are used alone or in combination with others to estimate the likely concomitant level of another variable. One example is the use of CDRs as a proxy for 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 —– this may include, for example, applications related to early warning systems, which look at patterns associated with past

[56] Meetup n.d.
[57] Borders n.d.
[58] Orange n.d.
[59] Grill 2013.
[60] Italia 2015.
[61] BBVA Innovation Center n.d.
[62] Letouzé, Meier, and Vinck 2013.
[63] Data-Pop Alliance 2015c.

Table 1.1: Data ‘inflation’

Unit	**Size**	**What it means**
Bit (b)	1 or 0	Short for “binary digit”, after the binary code (1 or 0) computers use to store and process data—including text, numbers, images, videos, etc.
Byte (B)	8 bits	Enough information to create a number or an English letter in computer code. It is the basic unit of computing.
Kilobyte (KB)	1,000 B or 210^{10} bytes	From “thousand” in Greek. One page of typed text is 2 KB.
Megabyte (MB)	1,000 KB or 220^{20} bytes	From “large” in Greek. The MP3 file ofa typical song is about 4 MB.
Gigabytes (GB)	1,000 MB or 230^{30} bytes	From “giant” in Greek. A two-hour film can be compressed into 1-2 GB. A 1 GB text file contains over 1 billion characters, or roughly 290 copies of Shakespeare’s complete works.
Terabyte (TB)	1,000 GB or 240^{40} bytes	From “monster” in Greek. All the catalogued books in America’s Library of Congress total 15 TB. All the tweets sent before the end of 2013 would approximately fill an 18.5 TB text file. Printing such a file (at a rate of 15 A4-sized pages per minute) would take over 1200 years.
Petabyte (PB)	1,000 TB or 250^{50} bytes	The NSA is reportedly analyzing 1.6% of global Internet traffic, or about 30 PB, per day. Continuously playing 30 PB of music would take over 60,000 years, which corresponds to the time that has elapsed since the first Homo Sapiens left Africa.
Exabyte (EB)	1,000 PB or 260^{60} bytes	1 EB of data corresponds to the storage capacity of 33,554,432 iPhone 5 devices with a 32 GB memory. By 2018, the total volume of monthly mobile data traffic is forecast to be about 0.5 EB. If this volume of data were stored on 32 GB iPhone 5 devices stacked one on top of the other, the pile would be over 283 times the height of the Empire State Building.
Zettabyte (ZB)	1,000 EB or 270^{70} bytes	It is estimated that in 2013, humanity generated 4-5 ZB of data, which exceeds the quantity of data in 46 trillion print issues of The Economist. If that many magazines were laid out sheet by sheet on the ground, they would cover the total land surface area of the Earth.
Yottabyte (YB)	1,000 ZB or 280^{80} bytes	The contents of one human’s genetic code can be stored in less than 1.5 GB, meaning that 1 YB of storage could contain the genome of over 800 trillion people, or roughly that of 100,000 times the entire world population.

The prefixes are set by the International Bureau of Weights and Measures.

Sources

Adapted and updated from The Economist by Emmanuel Letouzé and Gabriel Pestre, using data from Cisco, the Daily Mail, Twitter (via quora.com), SEC Archives (via expandedramblings.com), BistesizeBio.com, and "Uncharted: Big Data as a Lens on Human Culture" (2013) by Erez Aiden and Jean-Baptiste Michel.

events like an epidemic, to estimate the likelihood of a new event (*e.g.* a new epidemic). Another example is to build on past forced population movement to forecast future displacement route in case of a new crisis. Google Map is a good example of a service built on that function.

3. The third and not very much developed to date, is a *prescriptive* function, in which the predictive function of CDRs is enhanced to examine the possible consequences of different choices of action, resulting in recommendations on the best course of action. This function is the closest to finding causal relationships to prescribe a course of action on their basis. This function can also 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.

4. The fourth, in some ways encompassing, function of Big data is the *discursive* function. In the policy and social worlds, it is about spurring and shaping dialogues within and between Big Data communities, through and about Big Data. It can take the form of events about informed consent in the age of Big Data for example, where the collection and analysis of big data are used as an anchor, or entry point, to discuss legal frameworks, power dynamics and so forth.[64]
   An important question is whether, or rather, how people as emitters of data may want to gain greater control over their data, and what are the legal and political implications and requirements of such change.[65] In the academic world, it can take the form of using Big Data to reassess commonly held views and past conclusions, as in the case of chapter 4 for example. In other words, it is about leveraging the opportunities and challenges brought about by Big Data to discuss the state of the world and our collective knowledge about its determinants.

Other functional taxonomies have been developed. For example, in a paper titled "The Data revolution and Economic Analysis,"[66] two Stanford researchers discussed Big Data's opportunities for economic policy on the one hand, and for economic research on the other hand.

Regarding the former—economic policy—they identified four "potential uses":

1. "Making Use of Government Administrative Data"

2. (Developing) "New Measures of Private Sector Economic Activity"

3. "Improving Government Operations and Services"

4. (Improving) "Information Products or Services"

[64] Data-Pop Alliance 2015b.

[65] A. ' Pentland 2014; A. ' Pentland 2015.

[66] Einav and Levin 2013.

Regarding the latter—economic research—they identified three functions, or "areas of relevance":

1. "Novel Measurement and Research Designs"
2. "Statistical Learning and Economic Research"
3. "Embracing Heterogeneity"

I think each of these uses or areas of relevance builds on at least one of the four functions of Big Data I described above. There are certainly many ways in which one can think about how Big Data can affect the world; whether through its impact on industries, research, policies, people's lives directly (*e.g.* Google Map), etc. but I have found the four-tier taxonomy proposed above to be a useful frame of reference.

1.4 Will big data replace traditional data and / or Big Data render the scientific method obsolete?

At this point I turn to addressing the specific question of how Big Data may challenge, complement, or render obsolete traditional social science research, as well as a few other considerations that are more specific to social science research in general and population science in particular in the final section of this opening chapter, with specific references to the following chapters—notably Chapter 6 on political, ethical, and legal aspects.

Will big data replace traditional data and Big Data render the scientific method obsolete? Largely no. But Big Data—and big data as part of it (the "crumbs")—will definitely shake things up more and more.

Let me start by the first part of the question, focusing on official statistics, which I have thought, written, and talked about quite extensively, first while working on the UN Global Pulse report[67] and then especially since I was first asked to address the topic at the 2013 International Statistical Institute's World Statistical Congress in Hong Kong and in other public events and papers later.[68] The standard, easy, answer to the hard, legitimate question of the implications of both big data and Big Data for official statistics has been and remains that "big data will not replace but complement official statistics." The statement is partly valid, but largely flawed, not least because of its lack of conceptual and definitional clarity.

In defining official statistics as data produced by specific entities according to the Fundamental Principles of Official Statistics on the basis of some raw data, nothing—including in the Principles—prevents future official statistics to be based on big data sources.[69] In that sense, big data would neither replace nor complement official statistics: big data would

[67] See especially pages 36-38.

[68] For an overview, see: Giovannini 2010.

[69] United Nations Statistics Division 2014.

become a source of and for official statistics—[70] where 'official statistics' refers here both to some particular kinds of data and to the industry that produces them. These big data sources would then be used as raw material to produce both existing and new kinds of indicators, including alternative measures of welfare or new ways of capturing the state and trajectory of economies and societies in the future.[71] Indeed, it is puzzling that with so much more information available, as crude—and in great part as bad—an indicator of economy dynamism as GDP, developed in a data-poor industrial era, continues to be the number one target of public policies in the 21st Century. This has of course been a longstanding debate since at least Amartya Sen and the development of the Human Development Index, followed by the recommendations of the Stiglitz commission.[72] But I believe that the day when the increasing availability of fine-grained complex datasets about human actions and ever more powerful and ingenious methods to make sense of them will make GDP obsolete is not in a very distant future.

What many researchers and public officials wonder or worry more about is actually a slightly different question: it is whether some day there will be no more surveys because they will no longer be needed—or because there won't be any money or people left to conduct them, if governments stop investing in official statistics for instance. I doubt it. The basic reason is that surveys provide what are considered 'ground truth' data, without which all models based on big data are essentially blind, or wild guesses. This provides a powerful reason to not discard surveys. All four empirical chapters in this dissertation exemplify this fact; they combine CDRs and official data—in the sense of being collected or at least vetted by official entities as part of their mandate: from police records (Chapter 2), DHS (Chapter 4), and censuses (Chapter 5). In all these cases, and in many others in the literature, these official data are considered as the closest possible quantitative encapsulation of reality.

Of course, as discussed above and in greater length in Chapter 3, a non-trivial fraction of official data is bad, because of flawed methods, lack of timeliness that require questionable imputations, etc. Furthermore, as in the case of GDP, all official data are results of world views as to what ought to be measured. In that sense they are not objective and shrink the complexities of human ecosystems into quantifiable metrics that are all partial; they "confine and tame the personal and subjective."[73] If anything, better surveys will be needed to build more accurate models to more deeply understand and capture the human experience. Additionally, in some cases the outcomes of analyses that combine big data and survey data might be at odds with official figures, acting as a healthy check on their reliability—as has already been the case for both inflation and population figures.[74] As such, big data can help improve and challenge traditional data, through a dialogue.

This does not mean that censuses will continue to involve lengthy and highly costly door-to-door data collection processes; in fact in some OECD countries no longer do; they

[70] For an overview, see: Giovannini 2010.

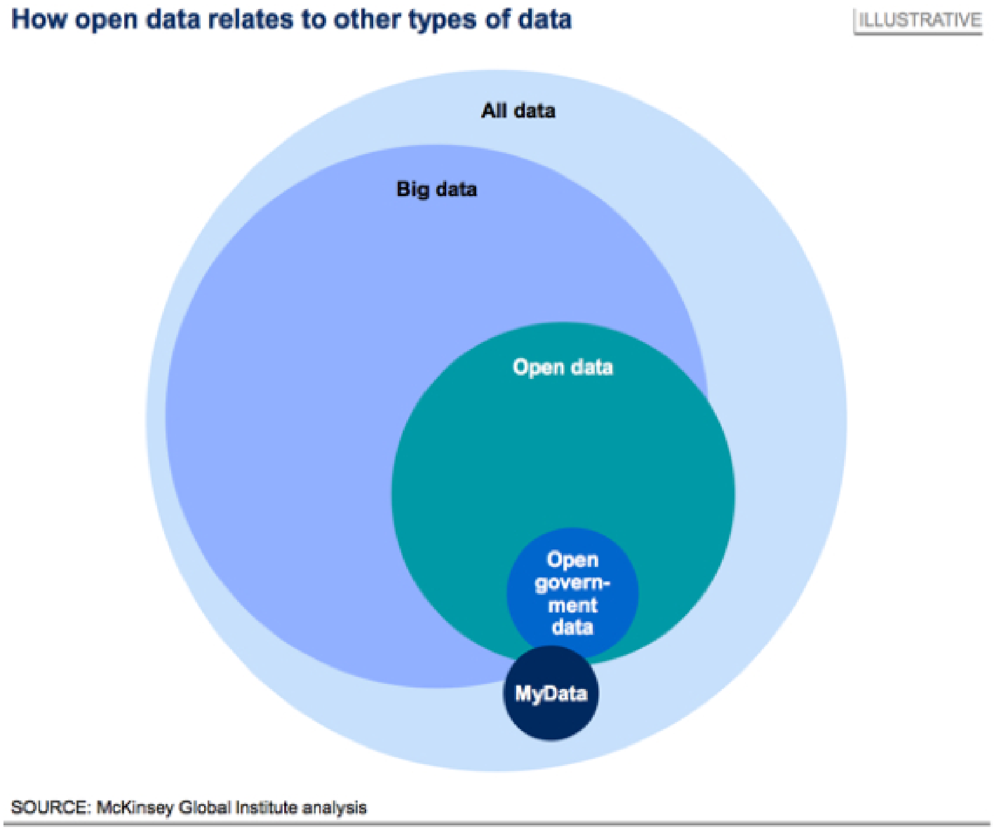
[71] Giovannini 2010; Letouzé 2013; Letouzé and Cohen-Setton 2014.

[72] Stiglitz, Sen, and Fitoussi 2009.

[73] Porter n.d.

[74] Lawhorn 2013.

Figure 1.7: How open data relates to other types of data



How open data relates to other types of data

SOURCE: McKinsey Global Institute analysis

use a mix of administrative data and sampled surveys. My sense, based on recent advances in the field, is that in the near future high resolution satellite imagery, CDRs and machine-learning techniques combined with surveys to create sampling frames will be widely used to dramatically reduce the time and cost of conducting censuses.[75]

What we are likely to see in the near future is an ever growing and complex universe of 'all data' from different sources—administrative records, 'straight' big data sources, survey data, official statistics, citizen-generated data, open data in the sense of public data made easily available for wide use, etc.—with some based on others and all interacting with each other in different ways (*e.g.* official statistics based on survey of big data and big data estimates calibrated using survey data). The respective 'communities' producing and using these data are also increasingly interacting and will most likely continue to do so, each bringing their tools and ontologies with them.

This leads me to the second part of the question: whether 'the scientific method' has or will become obsolete—and more realistically and interestingly how Big Data—with a focus here on the 2nd and 3rd of its Cs—is and may be changing social science research.

[75] Stevens et al. 2015.

The conclusion of Einav and Levin's paper touches on this question—although unfortunately rather superficially:

> There is little doubt, at least in our own minds, that over the next decades "big data" will change the landscape of economic policy and economic research. As we emphasized throughout, we don't think that big data will substitute for common sense, economic theory, or the need for careful research designs. Rather, it will complement them. How exactly remains to be seen.

[76]

I can only offer a few additional thoughts.

A lot of attention and tensions revolve around the issue of correlation versus causation. Many Big Data use cases do rely on finding correlations, as well as more broadly systematic patterns in the data. However, the immense majority of computers scientists and 'data scientists' I have worked with understand that finding a correlation between some X and some Y over long periods of time is not sufficient to conclude they are linked by a causal relationship such that acting on one would affect the other. In many cases however, this is a useful finding for researchers and policymakers. As attributed to Hal Varian, "even if all you have got is a contemporaneous correlation, you've got a 6-week lead on the reported values. The hope is that as you take the economic pulse in real time, you will be able to respond to anomalies more quickly."[77]

Relatedly, many also question whether Big Data is a-theoretical. Data mining refers to a process whereby some answers are asked to existing data, whereas—as noted by Chris Anderson—traditional social science research would follow the typical steps of hypothesis: data collection, testing, and theorization. In reality, the difference is more a matter of degree than a matter of nature. Theories do not come out of thin air but are rooted in observations.

What is indeed fundamentally different in the age of Big Data is that the notion of data minimization—*i.e.*, "the practice of limiting the collection of personal information to that which is directly relevant and necessary to accomplish a specified purpose"[78]—no longer operates, at least at the moment. These large streams and stocks of data are there, possibly containing answers to questions that may not have been asked yet. I do not think it is possible and desirable to go back to enforcing strict data minimization; I do think is that the practice of collecting all data indiscriminately and storing them indefinitely is a recipe for disasters—not to mention its economic and environmental costs. Critically though, the focus ought to be on consent, control, and use of the data—as I will discuss in Chapter 6.

There is however a risk with and indeed a tendency within Big Data as a community to focus too much on prediction at the expense of prescription. An example is the field of 'predictive policing'—central to Chapter 2. As discussed in Chapter 2 and other publications,[79] police

[76] Einav and Levin 2013.

[77] Bollier 2010.

[78] Marr 2016.

[79] Letouzé, Meier, and Vinck 2013.

and law enforcement in some US and UK cities have for years now analyzed data to assess the likelihood of increased crime in certain areas to dispatch resources accordingly. There is no doubt here about the fact that these are not, in most cases, causal relationships. But the ease and appeal of building good predictive models of criminality may divert attention and effort from understanding complex causal processes. And of course, unless there is knowledge of *why* crime is rising, it is difficult to put in place policies that tackle contributing factors. There is a risk to see more and more researchers developing good predictive models because they can, at the expense of putting in place research designs and addressing causal processes and root causes.

Validity considerations—both external and internal—ought to remain central in the data-rich future of social and population science. It may be worth stressing that even with very large samples, the people generating these data have typically selected themselves as data generators through their activity. The conditions of internal validity must also be kept in mind. A sharp drop in the volume of CDRs from an area might be interpreted, based on past patterns, as being an early indication of a looming conflict, but it could actually be caused by something different, such as a mobile phone tower having gone down in the area. Some of these issues—notably the issue of selection bias—are addressed in chapters Chapter 4 and, in particular Chapter 5.

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[80] for complex reasons that have now been analyzed in depth.[81] 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.

Social scientists and demographers in particular are well equipped not to fall in these traps. I believe much more work must be done to use Big Data at scale in a sound scientific way, and that the issue of sample bias in particular will receive the attention it deserves in the next few years. What this means and implies more broadly is that increasingly, academic research will be conducted by interdisciplinary teams of individuals coming from different perspectives and backgrounds to complement and challenge each other. Demography may also rebrand itself as population science (and demographers as population scientists) continuing to adapt to advances and changes in its environment as it has done since its birth.

Last, one of the most urgent avenues for academic research and public debate in the field—that also calls for multidisciplinary work, including involving philosophers, ethicists, and legal scholars—are risks to individual and group rights, privacy, identity, and security. In addition to surveillance activities and issues around their legality and legitimacy, there are important questions about 'data anonymization'—and its meaning and limits. As social science researchers find ourselves working with big data, we must be aware that the very notion of 'anonymized' data is being questioned. As early as 2008, a study of movie rentals

[80] Butler 2013.

[81] Lazer, R. Kennedy, et al. 2014.

showed that even ‘anonymized’ data could be re-identified and then ‘de-anonymized’— linked to a known individual by correlating rental dates of as few as three movies with the dates of posts on an online movie platform.[82] More recent research by de Montjoye et al. all but killed the notion that deleting personally identifiable information in a large dataset ensured anonymity;[83] in two papers looking at CDRs and credit card data, the authors showed that only four data points were theoretically sufficient to uniquely single out and re-identify individuals out of the whole dataset with as high as 95 per cent accuracy.

This is especially consequential because in most cases the people represented in these data sets have limited to no knowledge that they are producing data, or do not fully comprehend the scope of what can and is being done with their data. Often, even those collecting the data do not know what will become possible to do with these data in the future. There is no easy fix, and addressing the full implications of these perspectives would take an entire dissertation, or more. As mentioned, I will discuss issues of consent and control, of privacy and protection, in Chapter 6; right now it may suffice to say the following: first, in an increasingly digitally interconnected world, threats to privacy will grow—ceteris paribus. But that does not necessarily mean that “privacy is dead”—either because society would stop valuing it, or because the benefits of using the data would also trump the risks and costs. Nor does it mean that there are not, or will not, be technologies that will both ensure privacy and ensure that useful research is conducted with these data—as is the goal of a project I am involved in called OPAL, for “Open Algorithms”, presented in Chapter 6.

[82] Blondel, Esch, et al. 2013.

[83] de Montjoye, Hidalgo, et al. 2013; de Montjoye, Radaelli, et al. 2015.

Chapter 2

Predicting Crime Hotspots using Aggregated Cell-Phone and Socio-Demographic Data in London[1]

This chapter is interesting for two main reasons: first, because the topic it tackles—criminality—is embedded in longstanding debates about the root causes, structural determinants or simply correlates of criminality for public policy interventions; second, because it relies on standard methods and concepts in the machine-learning and data science literature—building predictive models through algorithms, distinguishing training and test sets, etc. with which most demographers and social scientists are not yet familiar with.

This chapter is organized as follows. Section 2.1 introduces the body of research within which this investigation takes place; Section 2.2 describes the data sources used for the analysis; Section 2.3 explains the research objective and empirical strategy; Section 2.4 presents the results; Section 2.5 discusses the findings and implications of this research; and Section 2.6 serves as a conclusion.

2.1 Big Data and the analysis of crime and violence

2.1.1 The rise and promise of Big Data to tackle social problems

As a social phenomenon, Big Data—characterized through its 3 Cs of Crumbs, Capacities, and Communities (see Chapter 1)—is rapidly changing the world we live in, challenging and often subverting long lasting paradigms in a broad range of domains. The almost

[1] The bulk of this chapter draws on a paper I co-authored in 2015, published in the peer-reviewed academic journal "Big Data." (Bogomolov et al. 2014) I particularly focused on framing the issue as part of the larger ongoing debates in the field, contributed to the demographic analysis and had a large role in its writing. It also builds on several other contributions I wrote as lead author, including the UN Global Pulse paper from 2012 and the chapter on Big Data and conflict prevention for UNDP and USAID, referenced in the text and other parts of the dissertation.

universal adoption of the mobile phone and the exponential growth of Internet-based services in particular have created unprecedented amounts of passively emitted data about human behavior and beliefs.

Although still "in its intellectual and operational infancy," as I put it, this field has gone through a rapid phase of expansion and maturation in a short period of time, driven by and spurring key studies on mapping the propagation of diseases (such as malaria and H1N1 flu), monitoring socio-economic deprivation, predicting human emergency behavior, detecting the impact of natural disasters such as floods, and inferring pollution emissions of vehicles. Additional examples are discussed in other chapters of this dissertation.

Interestingly, this double movement of expansion and maturation has been driven in large part by non-academic institutions, especially in its very early years (*e.g.*, United Nations Global Pulse, Flowminder.org, Data-Pop Alliance, DataKind), and initiatives (*e.g.*, Orange Data for Development, Telefonica Datathon for Social Good, Telecom Italia Big Data Challenge, Chicago Data Science for Social Good Fellowship). The UN Global Pulse report discussed the challenges and opportunities of using Big Data for societal challenges using a three-tier taxonomy of potential applications, namely "real-time awareness," "early warning," and "real-time feedback." A subsequent paper on the specific case of Big Data for conflict prevention I wrote as lead author in 2013 distinguished its "descriptive," "predictive," and "prescriptive" functions. Later, as mentioned above, I and co-authors added a fourth "discursive" function. These papers and taxonomies have been widely cited and used since then, and I think they provide a good base to discuss applications and implications of Big Data for research and policymaking. In the next paragraphs I will briefly summarize again these functions.

The prescriptive function corresponds to the golden standard of traditional academic research; it is the realm of causal inference, where a causal link is credibly established between two variables, which can then be used to change policy. Interestingly, Big Data is not a natural fit for this purpose, for a host of reasons ranging from sample bias to feedback loops. For example, it is difficult to see how a randomized control trial could be designed using big data. However, research in this area is poised to grow, and indeed Bayesian models are being developed to try and establish causality in time series.[2]

The descriptive function of Big Data is the most straightforward to grasp; it is the area of real-time traffic maps and word clouds, where visualizations play a significant role in challenging and improving standard descriptive statistics methods. The predictive function encompasses two different cases: inference and forecasting. The former focuses on attempting to gain insights on some variable of interest—poverty levels or population density, for example—on the basis of another data source—such as cell-phone use or light emissions. Anomaly detection systems would also fall in that category, in that they attempt to infer some outcome (whether there is a fire or not) based on current data streams. This has also been called 'nowcasting.' The latter, in contrast, is about forecasting what may happen in the future given current data features on the basis of past trends and patterns. It is different from the

[2] Shiffrin 2016.

prescriptive use, in that no causal link is required, although it may trigger action. To date, most Big Data applications have relied on this predictive function. Amazon and Facebook algorithms are well known examples of such applications. The discursive function entails assessing or reassessing the state of our world (or social processes less emphatically) using Big Data as a primary lever or entry point. In this chapter, the predictive function dominates, although the discursive function is not absent.

2.1.2 The case of violence and crime prediction

In the realm of public policy and counterterrorism, intelligence and law-enforcement agencies have also long used predictive models. Big Data is used for police work and public safety purposes. New York City Police Department (NYPD) collaborates with Microsoft to aggregate and analyze existing public safety data streams in real time for investigators and analysts.[3] With a focus on high-risks sites, the result of this collaboration, the Domain Awareness System, pulls data from a network of 3,000 close circuits camera along with license-plate readers, 911 calls, previous crime reports, and radiation detectors to identify threats.

Although the trend started in the 1990's when police forces started to systematically gathering and analyzing data from high-crime areas, the advent and adoption of Big Data analytics—by allowing the search for patterns and correlations in vast quantities of high frequency data—are leading to the development of a radically new form of "predictive policing"[4] (or "predictive analytics"[5]) to "predict,"[6] "sense,"[7] "stop,"[8] or "fight"[9] crime "before it happens,"[10] triggering references to the 2002 film *Minority Report.*[11] Beyond New York City, such programs are already in use in other U.S. cities,[12] notably Los Angeles, as well cities in the UK.[13]

Critics of these approaches have pointed to their inability to tackle root causes and the risks of profiling and harassment they may create, while questioning their efficiency.[14] But others have argued that curbing violence preemptively may have lasting structural impacts on communities.[15]

[3] New York City Office of the Mayor 2012.

[4] Beck and McCue 2009.

[5] Agence France Presse 2012.

[6] Beam 2011.

[7] Main 2011.

[8] Bailey 2012.

[9] IBM 2011.

[10] IBM 2011.

[11] Spielberg 2002.

[12] Rochester, Minnesota: "Rochester Police to fignt crime with big data mining technology" 2012.

[13] Mackie 2012.

[14] Morozov 2014.

[15] Bock and Lederach 2012.

Crime has not yet been widely covered in the Big Data literature apart from a few examples.[16] However, it provides fertile ground to advance our collective understanding of crime and to derive implications along two lines of work: first, we can validate the predictive power of *place-based crime models* built from 'anonymized' and aggregated human behavioral data with limited to no privacy risks; and second, we can revisit the distinctions and potential complementarities between the predictive and prescriptive functions of Big Data and those of short-term *vs.* long-term effects of policies on societies.

The main objective in the investigation is to propose and evaluate a Big Data approach to the problem of crime hot-spot forecasting, *i.e.* the identification, based on past data, of geographic locations which are likely to become the scene of a crime. In particular, we combine demographics with anonymized and aggregated *people dynamics* features, derived from mobile network activity, in order to predict whether specific locations are more or less likely to become crime hot-spots in the near future. No inference can be made about specific individuals.

2.1.3 A tale of two theories

Crime is a well-known social problem affecting the quality of life and the economic development of a society. Several works have shown that crime tends to be associated with slower economic growth at both the national level[17] and the local level, such as cities and metropolitan areas.[18] Dating back to the beginning of the 20th century, studies have focused on the behavioral evolution of criminals and its relations with specific characteristics of the neighborhoods in which they grew up, lived, and acted. Existing works tend to mainly explore relationships between criminal activity and socioeconomic variables such as education,[19] ethnicity,[20] income level,[21] and unemployment.[22]

Urbanists and architects have also investigated the relationships between people dynamics, urban environment, and crime.[23] Urban activist Jane Jacobs in particular[24] has emphasized *natural surveillance* as a key deterrent for crime: as people are moving around an area, they will act as "eyes on the street" able to observe what is going on around them. According to her, "A well-used city street is apt to be a safe street and a deserted city street is apt to be unsafe."[25] Hence, Jacobs suggested that high diversity among the population and high

[16] Toole, Eagle, and Plotkin 2011; T. Wang et al. 2013; Ferrara et al. 2014; Traunmueller, Quattrone, and Capra 2014.

[17] Mehlum, Moene, and Torvik 2005.

[18] Cullen and Levitt 1999.

[19] Ehrlich 1975.

[20] Braithwaite 1989.

[21] Patterson 1991.

[22] Patterson 1991.

[23] Jacobs 1961; Newman 1973.

[24] Jacobs 1961.

[25] Jacobs 1961.

number of visitors contribute to the safety of a given area and lead to less crime; in particular, Jane Jacobs proposed that what four key features contributed to a city's safety:

1. *mixed land uses* to attract people who have different purposes
2. *small blocks* that promote contact opportunities among people
3. *building diversity*, with a mix high-rent and low-rent tenants
4. *high density of population* that promotes high levels of concentration and interactions

In contrast, Newman's theory[26] argues that a high mix of people creates the anonymity needed for crime. Thus, according to the latter, low population diversity, a low visitor ratio, and a high ratio of residents are the features contributing to an area's safety. Several studies have tried to shed light onto these conflicting theories. Felson and Clarke[27] have proposed the Routine Activity Theory, which investigates how specific situations and variations in life-style affect the opportunities for crime. Specifically, they found that some places such as bars and pubs attract crime.

Criminologists have also started to investigate in detail significant concentrations of crime at micro levels of geography, regardless of the specific unit of analysis.[28] Research has shown that in what are generally seen as good parts of town there are often streets with strong crime concentrations, and in what are often defined as bad neighborhoods, there are locations relatively free of crime.[29] In 2008, criminologist David Weisburd proposed to switch the popular people-centric paradigm of police practices to a place-centric paradigm.[30]

Based on these findings, this problem is framed the problem of crime prediction with a *place-centric* and *data-driven* approach: specifically we investigate the predictive power of *people dynamics*—derived from a combination of mobile network activity and demographic information—to determine whether a *specific geographic area* is likely to become a scene of the crime.

2.2 Datasets

2.2.1 The 2013 'Datathon for Social Good' competition

The main datasets used were provided during a public competition—the Datathon for Social Good[31]—organized by Telefónica Digital, The Open Data Institute and the MIT Human Dynamics Group. This Datathon took place in the context of the Campus Party

[26] Newman 1973.

[27] Felson and Clarke 1998.

[28] P. L. Brantingham and P. J. Brantingham 1999.

[29] P. L. Brantingham and P. J. Brantingham 1999.

[30] Weisburd 2008.

[31] Grill 2013.

Europe 2013 at the O2 Arena in London in September 2013. As mentioned in Chapter 1, it has become fairly standard these past few years for 'big data' datasets to be made available to researchers in such a manner, anonymized and aggregated, in a controlled environment with tight rules and non-disclosure agreements.

Participants were provided access to the following data, among others:

1. *Anonymized and aggregated human behavioral and demographics data* computed from mobile network activity and demographics information in the London Metropolitan Area. We shall refer to this dataset as the Smartsteps dataset, because it was derived from Telefónica's Smartsteps product.

2. *Geo-localised Open Data*, a collection of openly available datasets with varying temporal granularity. This includes reported criminal cases, residential property sales, transportation, weather and London borough profiles related to homelessness, households, housing market, local government finance, and societal well-being (a total of 68 variables or 'metrics').

2.2.2 Datasets used

This section describes each of the datasets used in our investigation.

2.2.2.1 Criminal Cases Dataset

The criminal cases dataset includes the geo-location of all reported crimes in the UK but does not specify their exact date, just the month and year. The data provided in the public competition included the criminal cases for December 2012 and January 2013.

The dataset includes: the crime ID, the month and year when the crime was committed, its location with the longitude, latitude, and address where the crime took place, the police department involved, the Lower Layer Super Output Area (LSOA) code, the LSOA name and the crime type out of 11 possible types, including anti-social behavior, burglary, violent crime, shoplifting, etc.

LSOAs are small geographical areas defined by the United Kingdom Office for National Statistics following the 2001 census; they have a mean population of 1,500 and minimum population threshold of 1,000, Their aim here is to define areas, based on population levels, with time-invariant boundaries.

2.2.2.2 Smartsteps Dataset

Smarsets is a commercial analytical product developed and offered by Telefónica that "uses anonymous and aggregated mobile data to help organizations make better business decisions based on actual behaviour."[32] Smartsteps dataset consists of a geographic division of

[32] "Smart Steps" n.d.

Table 2.1: Smartsteps data provided by the challenge organizers

Type	**Data**
Origin based	Total no. of people No. of residents No. of workers No. of visitors
Gender based	No. of males No. of females
Age based	No. of people aged up to 20 No. of people aged 21–30 No. of people aged 31–40 No. of people aged 41–50 No. of people aged 51–60 No. of people aged over 60

All the demographic variables refer to
1 hour intervals and to each Smartsteps cell.

the London Metropolitan Area into cells whose precise location—latitude and longitude—and surface area were provided; whereas the actual shape of the cell was not. In total, there were 124,119 such cells. For each of the Smartsteps cells, a variety of demographic variables were provided, computed every hour for a 3-week period, from December 9th to 15th, 2012 and from December 23rd, 2012 to January 5th, 2013; in particular:

1. An estimation of *the number of people within each cell*, referred to as *footfall*, or. This estimation is derived from the mobile network activity by aggregating every hour the total number of unique phone calls in each cell tower, mapping the cell tower coverage areas to the Smartsteps cells, and extrapolating to the general population by taking into account the market share of the network in each cell location;

2. An estimation of *gender, age, and home/work/visitor group splits.* That is, for each Smartsteps cell and for each hour, the dataset contains not only an estimation of how many people are in the cell, but also of the percentage of these people who are at home, at work or just visiting the cell, as well their gender and age splits in the following brackets: 0-20 years, 21-30 years, 31-40 years, etc. (Table 2.1). This information is not directly available from the activity in the phone network infrastructure but was provided by GFK, a market research firm partnering with the event.

2.2.2.3 London Borough Profiles Dataset

The London borough profiles dataset is an official open dataset containing 68 different metrics about the population of a particular geographic area. The spatial granularity of the borough profiles data is at the LSOA level.

The information includes statistics about the population, households (derived from the census), demographic information (*e.g.* proportion of population aged 0-15 in 2011, proportion of working age population in 2011, proportion of population aged 65 or over in 2011, etc.), migrant population (*e.g.* proportion of largest, second largest, and third largest migrant population by country of birth in 2011), ethnicity (*e.g.* proportion of population from Black, Asian, and Minority Ethnic groups), language (*e.g.* proportion of people aged 3+ whose main language is not English), employment (*e.g.* female, male, and total employment rate in 2012), NEET (Not in Education, Employment, and Training) people, benefits (*e.g.* proportion of the working age population who claim out work benefits in 2012), qualifications (*e.g.* proportion of the working age population with no qualifications in 2012), earnings (*e.g.* male, female, and general gross annual pay in 2012), volunteering, jobs density, business survival, crime, fires, house prices, new homes, tenure, greenspace, recycling, carbon emissions, cars (*e.g.* number of cars and number of cars per household in 2011), indices of multiple deprivation, General Certificate of Secondary Education (GCSE) results, children in out-of-work families, life expectancy, teenage conceptions, happiness levels, political control (*e.g.* proportion of seats won by Labour, LibDem, and Conservatives), and election turnout.

2.3 Research objective and empirical strategy: predicting crime hotspots

2.3.1 Set up as a binary classification task

The problem is cast as one of crime hotspot classification; in machine-learning parlance it is referred to as a *binary classification task*. More specifically, for each Smartsteps cell in the dataset, we classify—in other words, we try to predict—whether it will experience a *high* or *low* crime level the following month. In a nutshell, this performed by using Smartsteps features—selected variables or metrics, as described further below—computed on December data and crime data from December, then predicted based on Smartsteps features for January whether these cells will experience high or low crime in January, and then checked against crime ground-truth observations from January. Furthermore, one of the main goals of our investigation is to assess whether using a predictive model relying on behavioral data would perform better than one that one based on Borough Profiles data alone.

We choose to formulate the problem as a binary classification task for two main reasons. First, on policy grounds. An important advantage of dichotomizing the ground-truth variable is to simplify the presentation of the results, making them easily understandable to a wide audience. Dichotomization is often used in criminology studies where one of the goals is

Figure 2.1: Spatial distribution of crime events

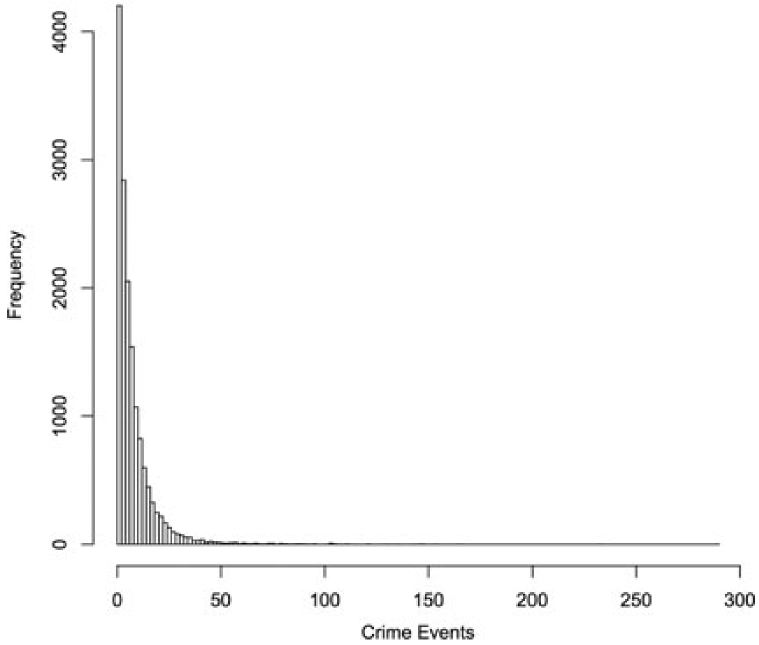


Table 2.2: Number of crime cases in January

Min.	Q1	Median	Mean	Q3	Max.
1	2	5	8.2	10	289

to present results to policy makers and police departments.[33] Indeed, given the fixed and finite resources available, policy makers and police departments are mainly interested in having a simple tool to decide where allocate "more" versus "less" resources while leaving the quantification of these resources to the decision maker.

Second, on methodological grounds. Dichotomizing a continuous variable—and in particular dichotomizing using the median—is statistically convenient when dealing with highly skewed distributions, as is the case here.[34] As depicted in Figure 2.1, the majority of Smartsteps cells in our dataset have few crimes (*e.g.*, only 1 crime event) while in a small proportion of the cells a high number of crimes is observed. The spatial distribution of the criminal cases for the month of January is summarized in Table 2.2. Given the high skewness of the distribution (skewness = 5.88, kurtosis = 72.5, mean = 8.2, median = 5), we opt to split the criminal dataset with respect to its median into two classes: *a low crime* (class '0') when the number of crimes in the given cell was less or equal to the median, and *a high crime* (class '1') when the number of crimes in a given cell was larger than the median.

[33] Boggs 1965; Farrington and Loeber 2000.

[34] Streiner 2002.

Hence, a cell with a number of crimes strictly higher than the median value of crimes for that particular month is considered as experiencing high crime, and is referred to as a *crime hotspot*. Following the empirical distribution, the two resulting classes are approximately balanced, with 53.15% of observations falling under the high crime class. Using another threshold—say 80-20 instead of 50-50—when using the median would make the classification task significantly more difficult, as is the case of predicting rare events.

2.3.2 Steps towards predicting crime hotspots

This section describes the steps to build the predictive model.

2.3.2.1 Training vs. test sets

First, we chose, as is standard in the literature, to separate the entire datasets into a training set—on which the predictive model is built— and a test set—on which its performance is evaluated. As is also customary, the training set contains 80% of the observations and the test set 20%. The split between the training and testing sets is done spatially, with 80% of the cells used for training and 20% of the cells used for testing. Importantly, no classification of January Smartsteps variables was possible since Smartsteps features are computed from December data while crime ground-truth variables are computed exclusively for January. In the following subsections we provide details of the experimental setup that we followed.

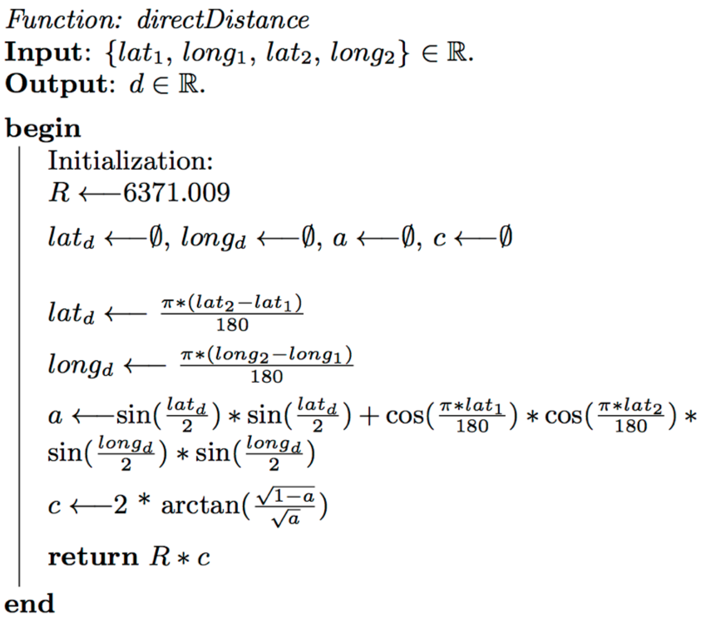
2.3.2.2 Referencing geo-tagged data to Smartsteps cells

Second, we need to link each crime event to the borough profile information were linked to one of the Smartsteps cells. Indeed, the Smartsteps cell IDs, the borough profiles and the crime event locations are not spatially linked in the datasets provided, and we did not have access to the actual shape of the Smartsteps cells. We geo-referenced each crime event by identifying the Smartsteps cell centroid closest to the location of the crime. We carry out a similar process for the borough profiles dataset. This is achieved by implementing the approximation Algorithm 1 (see Figure 2.2). Accounting for the curvature of the earth we introduced Algorithm 2 (see Figure 2.3) to calculate the direct spatial distance, given the *Fédération Aéronautique Internationale* Earth model, such that the Earth is treated as a three-dimensional ellipse, defined by two radii, a major axis (the radius at the equator) and a minor axis (the radius at the poles). The major axis is set to a constant equal to 6371.009 km.

2.3.2.3 Feature extraction

This step involves creating variables—referred to as features—from the available data that will be used in the predictive model; but only a subset of the features extracted—or variables created—will actually be used, following the feature selection step described below.

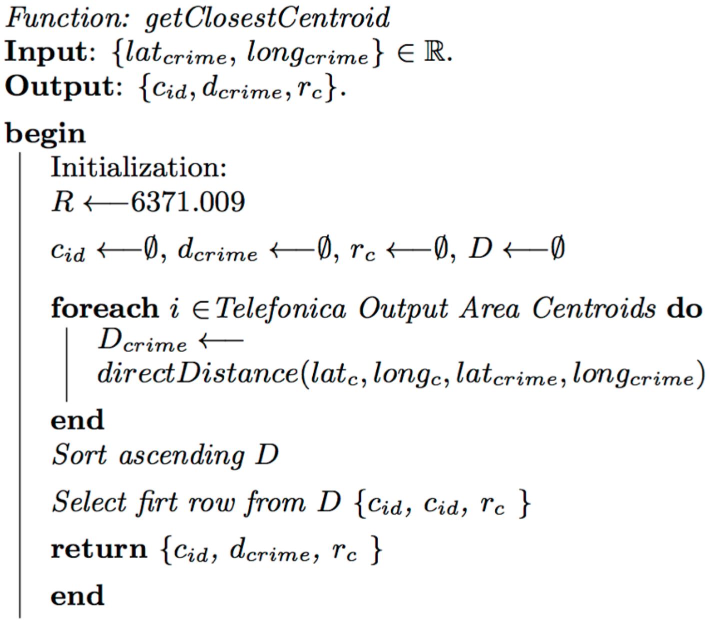
Figure 2.2: Algorithm 1: Approximating closest Telefonica's output area centroid for each crime event



```
Function: directDistance
Input: $\{lat_1, long_1, lat_2, long_2\} \in \mathbb{R}$.
Output: $d \in \mathbb{R}$.
begin
    Initialization:
    $R \longleftarrow 6371.009$
    $lat_d \longleftarrow \emptyset$, $long_d \longleftarrow \emptyset$, $a \longleftarrow \emptyset$, $c \longleftarrow \emptyset$

    $lat_d \longleftarrow \frac{\pi*(lat_2 - lat_1)}{180}$
    $long_d \longleftarrow \frac{\pi*(long_2 - long_1)}{180}$
    $a \longleftarrow \sin(\frac{lat_d}{2}) * \sin(\frac{lat_d}{2}) + \cos(\frac{\pi*lat_1}{180}) * \cos(\frac{\pi*lat_2}{180}) * \sin(\frac{long_d}{2}) * \sin(\frac{long_d}{2})$
    $c \longleftarrow 2 * \arctan(\frac{\sqrt{1-a}}{\sqrt{a}})$
    return $R * c$
end
```

Figure 2.3: Algorithm 2: Estimating direct distance



```
Function: getClosestCentroid
Input: $\{lat_{crime}, long_{crime}\} \in \mathbb{R}$.
Output: $\{c_{id}, d_{crime}, r_c\}$.
begin
    Initialization:
    $R \longleftarrow 6371.009$
    $c_{id} \longleftarrow \emptyset$, $d_{crime} \longleftarrow \emptyset$, $r_c \longleftarrow \emptyset$, $D \longleftarrow \emptyset$

    foreach $i \in$ Telefonica Output Area Centroids do
        $D_{crime} \longleftarrow directDistance(lat_c, long_c, lat_{crime}, long_{crime})$
    end
    Sort ascending $D$
    Select firt row from $D$ $\{c_{id}, c_{id}, r_c\}$
    return $\{c_{id}, d_{crime}, r_c\}$
    end
```

Diversity and *regularity* have been shown to be important in the characterization of different facets of human behavior. In particular, the concept of entropy[35]—roughly the level of disorder or uncertainty in a system—has been applied to assess the socio-economic characteristics of places and cities,[36] the predictability of mobility,[37] and spending patterns.[38] Hence, for each Smartsteps variable (see Table 2.1) we computed the mathematical functions which characterize its distribution and information theoretic properties, *e.g.*, mean, median, standard deviation, min, and max values and Shannon entropy,[39] in reference to Claude Shannon, one of the godfathers of information theory. The main idea is to estimate how much information on the whole system each piece of data contains, which provides a sense of how predictable the system is.

Furthermore, in order to be able to also account for temporal relationships within the Smartsteps data, the same computations described above were repeated on sliding windows of variable length (1-hour, 4-hour and 1-day), producing second-order feature—that is, features derived from pre-computed or first-order features—that help reduce computational complexity and the feature space itself, while preserving useful data properties.

In contrast, no data pre-processing was needed for the London borough profiles. Hence, we used the original 68 London borough profile features.

2.3.2.4 Feature selection

Remember that one of the goals is to provide a comparison between a predictive approach including behavioral data contained such as the Smartsteps data and a traditional one based on Borough Profiles data alone. Subsequently, we need to limit the number of features used by our model to 68, to match the maximum number of Borough Profiles variables that we were granted access to. Moreover, the limitation of the number of features actually used in the model reduces training times and enhances generalization performance by reducing the risk of overfitting.[40] In other words: a predictive model based on millions of features will likely yield a very high predictive power, but it has two main drawbacks. One, it imposes very high data requirements to be used (replicated, generalized) in other settings. Two it makes it prone to overfitting, which has been called "the cardinal sin of data mining."[41]

To select those 68 features from the set of all first- and second-order extracted features, we follow a so-called bootstrap aggregating (or bagging) procedure using exclusively data from the training set. Bagging is a machine-learning procedure, whereby predictors are constructed using bootstrapped samples from the training set, then aggregated to form a "bagged predictor." Each bootstrapped sample is formed using a drop-out strategy, leaving

[35] Shannon 1948.
[36] Eagle, Macy, and Claxton 2010.
[37] Song et al. 2010.
[38] Krumme et al. 2013.
[39] Shannon 1948.
[40] Guyon and Elisseeff 2003.
[41] Piatesky and Rajpurohit 2014.

out one third of the training examples. These left-out examples are used to form accurate estimates of important measurements for local optimization decisions (*e.g.*, to give better estimates of node probabilities and node error rates in decision trees).[42]

The metric we use for feature ranking and thus select feature section is the *mean decrease in the Gini coefficient of inequality*.[43] This choice was motivated because it outperformed other metrics such as mutual information, information gain, and chi square statistic.[44] In this context, the Gini coefficient, as the standard Gini coefficient capturing inequality widely used in economics, ranges between 0 and 1. Here, 0 means that all features have the same predictive power, suggesting maximum equality in predictive power, while 1 suggests that one given feature has all the predictive power; *i.e.* expressing maximal inequality in predictive power.

The feature with maximum mean decrease in Gini coefficient is expected to have the maximum influence in minimizing the out-of-the-bag error, namely the misclassification error rate which is estimated on the dropped-out samples during the bagging procedure. It is known in the literature that minimizing the out-of-the bag error results in maximizing common performance metrics used to evaluate models.[45] In other words, and intuitively, we are interested in finding those features, which when dropped from the model, cause the Gini coefficient to drop the most, *i.e.* to move from a state with greater predictive power inequality to a state with lower predictive power inequality, as this suggests that they carry strong predictive power themselves.

The top 20 features selected by the model are included in Table 2.3.

2.3.2.5 Model building

The classification is then performed with Random Forests (RF) ensemble classifiers, one of the most widely used and accurate learning algorithms available.[46] Random Forests also satisfy the max-margin property, do not require parameter tuning, and, more importantly, do not require the specification of a feature-space, as Support Vector Machines (SVMs) do through the kernels. Running the same experiments using SVMs with linear and RBF kernels yields less stable and less accurate results. Hence, we report the performance results only for the best model, based on RF.

Decision trees are an intuitive method to tackle classification and regression problems. In the case of binary classification, a tree assigns features by creating a control structure on a feature middle point for a decision of splitting either left or right through nodes of the tree depending on the value of a given point of the variable. A binary tree, by definition, ensures that each case of independent variable is assigned to a unique terminal node. The value of the terminal node is a predicted outcome and defines the classification decision. That means

[42] Breiman 1996b.
[43] Singh, Murthy, and Gonsalves 2010.
[44] Singh, Murthy, and Gonsalves 2010.
[45] Tuv et al. 2009.
[46] Biau 2012; Caruana, Karampatziakis, and Yessenalina 2008.

Table 2.3: Top 20 selected features ranked by mean in decrease accuracy

Base feature	Temporal resolution	1st order	2nd order	0	1	Mean decrease accuracy	Mean decrease gini
Age >60	Daily	Entropy.empirical	Entropy.empirical	4.48	5.43	9.02	18.75
At home	Daily	Mean	SD	3.20	7.60	8.91	27.13
Age <20	Daily	SD	Entropy.empirical	5.69	3.97	8.85	16.88
Age <20	Daily	Mean	Entropy.empirical	3.09	5.88	8.85	17.26
Age <20	Daily	Mean	SD	4.50	5.27	8.65	16.03
At home	Daily	Min	Entropy.empirical	6.39	2.32	8.61	15.99
At home	Daily	SD	SD	3.22	8.58	8.60	45.82
At home	Daily	SD	Mean	3.35	5.83	8.57	24.93
Age >60	Daily	Entropy.empirical	SD	4.62	4.95	8.56	20.45
At home	Daily	SD	Median	5.41	5.04	8.50	26.48
Age 31–40	Daily	Entropy.empirical	Max	2.33	5.79	8.44	16.24
Age 31–40	Daily	Min	SD	6.81	4.06	8.31	36.52
At home	Daily	Min	SD	4.36	6.85	8.29	34.26
At home	Daily	SD	Max	4.13	6.87	8.27	34.89
At home	Monthly	Max	–	3.92	5.42	8.26	29.86
At home	Monthly	SD	–	4.43	4.17	8.21	39.70
Age 51–60	Daily	Entropy.empirical	Entropy.empirical	4.74	4.11	8.13	16.64
Age <20	Daily	SD	SD	3.67	5.88	8.12	16.86
At home	Daily	Entropy.empirical	Entropy.empirical	5.13	4.82	8.08	18.55
At home	Daily	Max	SD	2.83	6.29	8.07	26.85

that the decision rule is a path down the tree to its terminal node. The decision boundary is estimated by an ensemble set of decision rules. RF algorithms produce a combination of trees, such that each one is dependent on the values of a random vector sampled independently with the same distribution for all the classification trees in the forest.[47]

The decision boundary is formed according to the margin function as follows. Given an ensemble of tree classifiers $h_1(x), h_2(x), ..., h_k(k)$, and if the training set is drawn at random from the empirical distribution of the random vector Y, then the margin function X is defined as:

$$mg(X,Y) = avg_k I\left(h_k(X) = Y\right) - max_{j!=Y} avg_k I\left(h_k(X) = j\right) \tag{2.1}$$

where $I(\cdot)$ is the characteristic function. The margin function measures the distance between the average vote at for the right class and the average vote at (X, Y) for any other class. For this model the generalization error function is given by:

$$PE^* = P_{X,Y}\left(mg(X,Y) < 0\right) \tag{2.2}$$

[47] Breiman 1996a.

where $P_{X,Y}$ is the probability over $\langle X, Y \rangle$ space. For any event $A \subset \Omega$ of the feature space the characteristic function $I(\cdot)$ of A is:

$$I_A(x) = \left\{ \begin{matrix} 1 \Leftrightarrow (x \subset A) \\ 0 \; otherwise \end{matrix} \right\} \left\{ \begin{matrix} 1 \,\exists\, x \\ 0 \; otherwise \end{matrix} \right\} \tag{2.3}$$

The procedure takes advantage of the performance improvements that are obtained by growing an ensemble of trees and voting for the most frequent class. Random vectors were generated before the growth of each tree in the ensemble, and a random selection without replacement was performed.[48]

2.4 Results

In this section the experimental results reported were obtained by the RF trained on different subsets of the selected features and tested on the test set. As is common in these settings, the performance metrics used to evaluate our approach are: (i) accuracy, (ii) F1 score, the harmonic mean between precision and recall, and (iii) area under the ROC (AUC) score (where ROC stands for Receiver Operating Characteristics)..[49] In simple terms, a ROC curve is a plot of the true positives (here, accurately predicting a cell as high crime) on the y-axis versus the false positives (incorrectly predicting a high crime cell as low crime) on the x-axis in a classification task.

In order to understand the value added by the Smartsteps data, we compare the performance of the RF[50] using all features (Smarsteps + Borough) versus:

1. a baseline majority classifier, which always returns the majority class ("high crime") as its prediction (accuracy = 53.15%), as well as two additional models trained with

2. (only the subset of selected features derived from the borough profiles dataset (Borough-only) and

3. only the subset of selected features derived from the Smartsteps dataset (Smartsteps-only).

Table 2.4 reports accuracy, F1-score, and the AUC metric for each of the models. First, the Smartsteps+Borough and the Smartsteps models significantly outperform the baseline majority classifier, with an increase of about 15% of accuracy.

Interestingly, the addition of the borough profiles features does not yield any significant improvement to the Smartsteps-only model (Smartsteps+Borough model accuracy = 68.83% vs. Smartsteps-only model accuracy = 68.04%). Moreover, the Borough-only model yields a competitive but significantly lower accuracy than the Smartsteps model: 62.18%, over 6%

[48] Breiman 1996a.

[49] Powers 2007; Provost and Fawcett 2013.

[50] Breiman 1996a.

Table 2.4: Metrics comparison

Model	**Accuracy**	**Accuracy, 95% CI**	**F1 score**	**AUC**
Smartsteps + borough profiles	68.83	(0.67 , 0.70)	68.52	0.63
Smartsteps	68.04	(0.66 , 0.69)	67.66	0.63
Borough profiles	62.18	(0.60 , 0.64)	61.72	0.58
Majority Classifiers (Baseline)	53.15	(0.53 , 0.53)	0	0.50

AUC, area under the ROC

lower than the accuracies obtained with the Smartsteps-only model (68.37%) and with the Smartsteps+Borough models (68.83%) while using the same number of variables.

In Table 2.4 we also report the F1-score for each model. This metric is the harmonic mean of the precision (the number of correct positive results divided by the number of all positive results) and the recall (the number of correct positive results divided by the number of positive results that should have been returned), where an F1 score reaches its best value at 1 and worst score at 0.[51]

Looking more in detail the performances of the different models on the “high crime” class, we focus on the *true positive rate*, namely the proportion of actual high crime cells which are correctly identified, and on the *true negative rate*, namely the proportion of actual low crime cells which are correctly identified. The Smartsteps-only and Smartsteps+Borough models obtain a good *true positive rate* performance of 74.20% and 73.90% respectively. Instead, the only-Borough model reaches a true positive rate of 68.81%, over 5% less than the models based on Smartsteps data.

When looking at the *true negative rate* performances, all the models obtain a worse performance: 63.07% for Smartsteps+Borough, 61.06% for only-Smartsteps, and 54.66% for only-Borough. Interestingly, our approach obtains a true positive rate about 10% higher than the true negative rate. Thus, our approach is performing better in correctly identifying high crime cells. It is worth emphasizing that, in our scenario, it seems more important to obtain good results on the “high crime” group: in fact, mistakenly assigning “high crime” to a cell (a false positive) seems less detrimental, from a social policy perspective, than erroneously classifying it as “low crime” (a false negative). As the results show, our proposed approach brings significant advantages for the task of hot-spot prediction.

For detailed analyses, Tables 2.5a, 2.5b and 2.5c report the confusion matrices of the only-Borough model, the only-Smartsteps-model, and the Smartsteps+Borough model, respectively.

Finally, a visual comparison of the ROC curves for each of the models is provided in Figure 2.4.

[51] Powers 2007; Provost and Fawcett 2013.

Table 2.5: Confusion matrices

(a) only-borough model

		Predicted	
		0	1
Actual	0	786	652
	1	509	1123

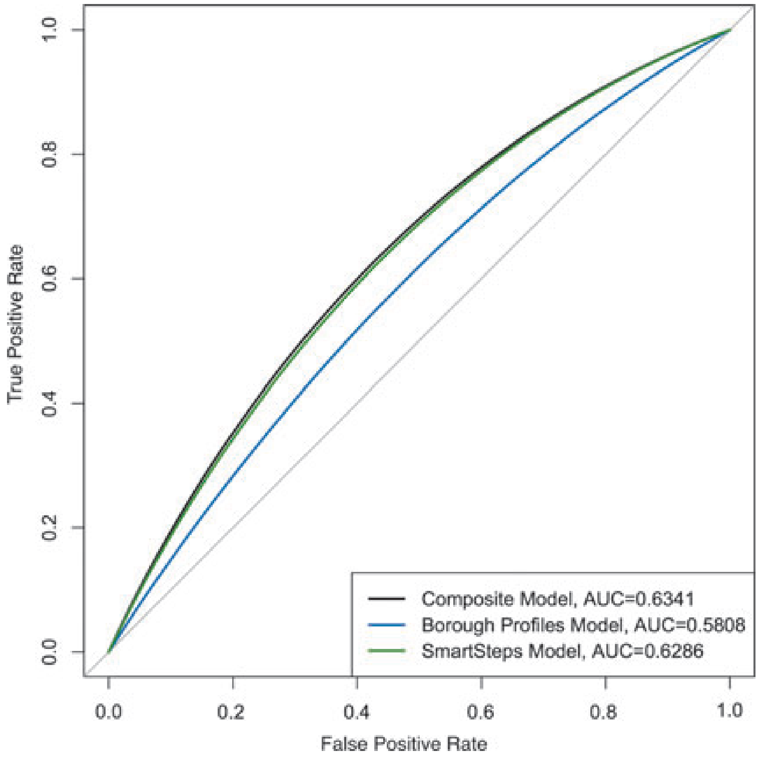
(b) only-Smartsteps model

		Predicted	
		0	1
Actual	0	878	560
	1	421	1211

(c) Smartsteps+borough model

		Predicted	
		0	1
Actual	0	907	531
	1	426	1206

Figure 2.4: ROC curves for Smartsteps + borough, only-Smartsteps, and only-borough profiles



2.5 Discussion and Implications

The results discussed in the previous section show that human behavioral and demographic data at a daily and monthly scale significantly can improve the prediction accuracy when compared to using rich statistical data about a borough's population (households census, demographics, ethnicity, employment, etc.). The borough profiles data does provide a fairly detailed view of the living conditions of a particular area in a city, yet this data is expensive and time-consuming to collect. Hence, this type of data is typically updated with low frequency (*e.g.*, every few years) making it difficult to predict potential changes in related outcomes or correlates.

Human behavioral data derived from mobile network activity combined with demographics, though less comprehensive than borough profiles, provides significantly finer temporal and spatial resolution.

Next, we focus on the most relevant predictors of crime level taking a look at the top-20 variables in our model, which are sorted by their mean reduction in accuracy in Table 2.3. The Smartsteps features have more predictive power than official statistics coming from borough profiles: no features listed in the top-20 are obtained using borough profiles.

Moreover, higher-level features extracted over a sequence of days from variables encoding the daily dynamics have more predictive power than features extracted on a monthly basis. Critically, this finding points out at the importance of capturing the temporal dynamics of a geographical area in order to predict its levels of crime.

Interestingly in light of the theoretical discussion above, features derived from the percentage of people in a certain cell who are at home both at a daily and monthly basis seem to be of extreme importance. In fact, 11 of the top 20 features are related to the *at home* variable. Newman's approach of "defensible space"[52] postulates the relevance of a high number of residents in an area to reduce crime. The predictive power of home variables seems to confirm their relevance. However, we found positive associations between the home variables and crime. Hence, our findings do *not* support Newman's thesis[53] suggesting that an increased ratio of residents is linked to less crime and higher urban safety. Similar results were found in recent work done by Traunmueller et al.[54] In their work, the researchers focus only on testing some hypotheses about people dynamics and crime using correlational analyses between footfall counts recorded by the mobile network activity and crime activities.

It is also interesting to note the role played by the unpredictability of the variables, captured by Shannon entropy features.[55] The entropy-based features in fact seem useful for predicting the crime level of places (8 features out of the top 20 are entropy-based features).

Here, the Shannon entropy captures the predictable structure of a place in terms of the types of people that are in that area over the course of a day. A place with high entropy would have a lot of variety in the types of people visiting it on a daily basis, whereas a place

[52] Newman 1973.

[53] Newman 1973.

[54] Traunmueller, Quattrone, and Capra 2014.

[55] Shannon 1948.

with low entropy would be characterized by regular patterns over time. In this case, the daily unpredictability in patterns related to different age groups, different use (home vs work) and different genders seems to be a good predictor for the crime level in a given area. In line with our results, Traunmueller et al.[56] found significant negative correlations between areas with higher age diversity and crime. Both our findings and those of Traunmueller et al.[57] support Jacobs' theory[58] of natural surveillance: high diversity of functions in a area and high diversity of people (gender-diversity and age-diversity) act as "eyes on the street" decreasing the number of crimes.

Interestingly, Eagle et al.[59] found that Shannon entropy used to capture the social and spatial diversity of communication ties within an individual's social network was strongly and positively correlated with economic development. Hence, high diversity areas seem to emerge as safer and more economically developed.

Our proposed approach could have clear practical implications by informing police departments and city governments on how and where to invest their efforts and on how to react to criminal events with quicker response times. From a proactive perspective, the ability to predict the safety of a geographical area may provide information on explanatory variables that can be used to identify underlying causes of these crime occurrence areas and hence enable officers to intervene in very narrowly defined geographic areas.

The distinctive characteristic of our approach lies in the use of features computed from aggregated and anonymized mobile network activity data in combination with some demographic information. Previous research efforts in criminology have tackled similar problems using background historical knowledge about crime events in specific areas,[60] criminals' profiling,[61] or wide description of areas using socio-economic and demographic indicators.[62] Our findings provide evidence that aggregated and anonymized data collected by the mobile infrastructure, combined with demographic information, contains relevant information to describe a geographical area in order to predict its crime level.

The first advantage of our approach is its predictive ability. Our method predicts crime level using variables that capture the dynamics and characteristics of the demographics and nature of a place rather than only making extrapolations from previous crime histories. Operationally, this means that the proposed model could be used to predict new crime occurrence areas that are of similar nature to other well known occurrence areas.

Even though the newly predicted areas may not have seen recent crimes, if they are similar enough to prior ones, they could be considered to be high-risk areas to monitor closely. This is an important advantage given that in some areas people are less inclined to report crimes.[63]

[56] Traunmueller, Quattrone, and Capra 2014.

[57] Traunmueller, Quattrone, and Capra 2014.

[58] Jacobs 1961.

[59] Eagle, Macy, and Claxton 2010.

[60] Eck et al. 2005; Mohler et al. 2011.

[61] Turvey 1999.

[62] Ellis, Beaver, and Wright 2009.

[63] Tarling and Morris 2010.

Moreover, our approach provides new ways of describing geographical areas. Recently, some criminologists have started to use risk terrain modeling[64] to identify geographic features that contribute to crime risk, *e.g.*, the presence of liquor stores, certain types of major stores, bars, etc. Our approach can identify novel risk-inducing or risk-reducing features of geographical areas. In particular, the features used in our approach are dynamic and related to human activities.

Further, as suggested in the introduction section, this study is relevant to two related debates that will shape to a great extent the future expansion and maturation of Big Data for Social Good as an intellectual and operational field: first, on the differences and complementarities between the predictive and prescriptive uses of Big Data, and second about the potential trade-offs between short and medium to long term policy interventions.

By design, predictive approaches are not meant to identify and thus address the complex processes that contribute to criminal behaviors in human societies. However, they can, as in the case of our study, shed interesting light on correlates of crime that are not out of the reach of public policies and community-based programs (*e.g.*, specific people dynamics, characteristics of places, etc.). Unveiling such correlates may inform subsequent academic research and policy pilots that may lead to crime reduction in the long run (*e.g.*, crime prevention through environmental design[65]). In other words: insights from predictive models may inform prescriptive approaches (and vice-versa).

In addition, short-term effects can have, cumulatively, structural impacts. In the area of conflict prevention for example, 'operational' (or direct) prevention efforts in general and 'preventive diplomacy' in particular have been increasingly recognized as critical to longer-term 'strategic' interventions intended to address the root causes of conflicts—economic, political, etc. One argument is that short-term, targeted interventions that avoid conflict escalations allow socio-political adjustment mechanisms to take place, gradually gearing societies away from oscillations around various stages of violence.[66] Community-based early warning systems of conflict have also been found to be more efficient than their previous top-down counterparts,[67] and it remains to be seen whether and how 'at-risk' communities could be empowered to make the most of Big Data for Social Good to reduce crime.

Note that the case study described in the paper suffers from a number of limitations due to the constraints of the datasets used. First of all, we had access only to 3 weeks of Smartsteps data collected between December 2012 and the first week of January 2013. In addition, the crime data provided was aggregated on a monthly basis. As previous studies have shown different crime types follow different temporal patterns.[68] Furthermore, having access to crime events aggregated on a weekly, daily, or hourly basis would enable us to validate the described approach with finer time granularity, predicting crime in the next week, day or hour. Finally, the human behavioral data used in this study is derived from the

[64] Caplan and L. W. Kennedy 2010.

[65] Jeffery 1977.

[66] Bock and Lederach 2012.

[67] OECD 2009.

[68] Felson and Poulsen 2003.

mobile network infrastructure of a mobile operator. There are many other sources of human behavioral data that could also be included in our analysis (*e.g.*, geo-tagged social media, public transport logs, etc.) and that could add complementary and valuable information for the task at hand. We leave this exploration to future work.

However, despite these limitations, the proposed approach illustrates the value of large-scale human dynamics data—which is actually available in an existing product (Smartsteps)—to classify crime levels.

2.6 Conclusion

This chapter presents a novel approach to predicting crime hotspots from human behavioral data derived from mobile network activity, in combination with demographic information. It shows how this type of data outperforms the use of traditional census data to classify areas of high and low crime. It also underlines the role of temporal variables in predicting crime. The work presented in this chapter opens the way to a new understanding of the role of large—scale human behavioral features on place—centric approaches to model crime and illustrates an interesting use case of how Big Data can be used to shed light on a major social issue.

Chapter 3

Weather, Mobility Patterns, and Conflict in Côte d'Ivoire[1]

This chapter uses CDRs from Côte d'Ivoire to study the relationship between weather conditions, population movements, and conflict events. Its main objective is to study whether weather conditions affect population movements. Finding a statistically significant relationship would cast doubts on the exclusion restriction in Miguel, Satyanath and Sergenti's seminal paper from 2004. In that paper, these authors used changes in rainfalls as an instrument for economic growth to study the causal effect between changes in income and the likelihood of conflict, which rested on the assumption that "weather shocks should affect civil conflict only through economic growth."[2]

Here, I find a strongly statistically significant correlation between rainfalls (and temperature) and population movements, captured by total distance covered around all weather stations in Côte d'Ivoire. Although this result does not in itself prove that the aforementioned exclusion restriction was violated, it suggests that future research on causal pathways to conflict, especially those using instrumental variable approaches relying on weather variables, should pay greater attention to the population movement channel. To the best of my knowledge, this paper—especially its version as a term paper in May 2013—also provided the first econometric evidence of a direct relationship between weather and human mobility.

Before discussing in greater depth the significance of these questions and answers, it may be useful to clarify their terms. First, in this chapter, "weather conditions" are captured by precipitation (or rainfalls, used interchangeably) and temperature levels, collected from all weather stations across Côte d'Ivoire. Second, "population movements" (or "mobility patterns") refer to the estimated sum of daily distances traveled by individuals in the immediate vicinity of these weather stations, inferred on the basis of cell-phone data indicating their location each time their phone was used during the day. Rephrased with these elements in mind, the

[1] This chapter draws on a term paper submitted in May 2013 for the 275 Economic Demography graduate class. It also builds on my previous research and interest in conflict and migration in Vietnam between 2000 and 2004 and at the United Nations between 2006 and 2009

[2] Edward Miguel, Satyanath, and Sergenti 2004.

question is: do people seem to be moving more or less depending on the amount of rain and/or level of temperature they experience? The short answer emerging from this analysis is that indeed people do seem to move more as precipitation and temperature levels increase. Both weather variables, jointly and separately, explain only a small fraction of changes in population movements, but the relationships are all highly statistically significant.

The chapter also suggests another potential application of Big Data in the field of development and social science research[3] beyond or building on its descriptive, predictive, and prescriptive functions—a 'revisiting role;' *i.e.*, the possibility of revisiting previous academic theories and models on the basis of new high frequency and low granularity data streams not available when these theories and models were first developed.

Substantively, the question has both theoretical and also potentially policy implications. I think that the question is in itself intellectually interesting. Thinking of individuals as electrons in an electric field, one could for instance be curious to find out wonder whether "shocking" the field (by increasing the prevailing temperature or precipitation level) may lead them to freeze or be more active.[4] In the case of weather conditions, it is easy to come up with stories and arguments that would support either hypothesis. For instance, one could hypothesize that people are more likely to stay indoor under heavy rain, or hot temperatures. In contrast, these same conditions could induce people to leave their home if it gets flooded, or to have to travel longer distances that usual if roads become impracticable, for example. Arguments for greater distances traveled in hotter days are not as straightforward to come up with, but there is certainly no theoretical impossibility.

Testing both hypotheses is hard, and would be very costly, with traditional data, because it implies being able to observe individual movements very precisely. To the best of my knowledge, the question has actually never been empirically investigated, although there appears to be anecdotal evidence or a wide belief that changes in rainfalls do affect migration patterns in the long run, as discussed below. With CDRs, though, that may be possible at a fraction of the cost a typical survey would cost. This is because CDRs contain information on where and when a cell-phone was used, in real time. In this case it is the phone (actually, the *SIM-card*) that travels, but in this chapter it is assumed that the observations pick up the movements of specific and unchanging individuals, although this limitation is reassessed in the concluding section.

In addition, there is a deeper potential implication, as mentioned above. Finding that changes in rainfalls and/or temperatures may be correlated with changes in human mobility patterns would be directly relevant to the literature on the drivers of conflict, notably. Indeed, finding a correlation between rainfalls and population movements may cast doubt on the restriction exclusion assumed by Miguel, Satyanath, and Sergenti in their 2004 seminal paper[5] (MSS 2004 hereafter) on economic shocks and civil war, as explained further below. Given Côte d'Ivoire's economic structure, climate, and history of conflict, discussed below, using it

[3] Letouzé 2012.

[4] See, Cameron, and Schwartz 2013.

[5] Edward Miguel, Satyanath, and Sergenti 2004.

to test the hypothesis seemed interesting.

The rest of this chapter is organized as follows. Section 3.1 provides additional theoretical and conceptual elements on the question at hand; Section 3.2 provides an contextual overview of Côte d'Ivoire where the analysis is grounded, providing a quick visual representation of the relationship between cases of violence and cell-phone activity; Section 3.3 presents the data and strategy used; Section 3.5 presents the results, and Section 3.6 concludes with a short discussion of the emerging implications, limitations and future leads.

3.1 Question: Weather, income, mobility, conflict, and an exclusion restriction

Certain countries and communities appear to be trapped in a continuous and seemingly close to inescapable cycle of poverty and violence. The contemporary history of several parts of Africa in particular—the Great Lakes region, the Horn of Africa, Western Africa to a somewhat lesser extent, for instance—has given credit to the existence of what has been described in the literature as a "conflict trap."[6] Cross-sectionally, the correlation between conflict—especially civil war—and economic development is evidently very strongly negative. Using extreme cases in point, the Democratic Republic of the Congo (DRC) has experienced conflict for decades, Norway has not. [7] But whereas the DRC and Norway are both resource-rich countries, they differ in many obvious ways beyond income per capita and overall human development, including in terms of their political regime, history, climate, and so on. This begs an old question: is income the only factor at play?

A large body of research has attempted to identify causal determinants of pathways to conflict. Perhaps the most widely held conviction in academic and policy circles remains that low economic development—and perhaps even more so *negative economic shocks* at low levels of development (or perhaps *runs* of negative shocks)—increases the likelihood of conflict. To simplify somewhat, the literature has theorized this link as a 'grievance' factor.[8] At the same time, it is possible that the negative correlation between the direction of an economic shock and the likelihood of conflict may turn positive depending on the context and the nature of the shock, in equally simple terms: when there is suddenly a larger cake to fight for. This has been referred to as the 'greed' argument.

Of course, as the very notion of a conflict-poverty-conflict 'trap' suggests, endogeneity poses a major identification challenge: the overall detrimental socio-economic effects of conflict are undisputed, such that the correlation found may simply reflect reverse causation, which obviously would not be dealt with with lags alone. It is also theoretically possible that both poor economic performance and conflict could actually result from a third common set of factors (climatological, political, 'cultural', for instance). The notion that democracies

[6] Collier, Elliott, et al. 2003.

[7] When faced with a rare case of large-scale gun violence in Utøya in the summer of 2011, Norway has reacted with impressive restraint.

[8] Collier and Hoeffler 2000.

don't fight each other makes no mention of income, but the fact is that the large majority of democracies are developed countries.

A number of papers have used instrumental variables to try and assess the existence, direction, and magnitude of any causal relationship between economic conditions and conflict dynamics. Dube and Vargas (2007, 2008, 2013) relied on changes in the world market prices of oil and coffee to identify the effects of changing economic conditions on conflict dynamics in Colombia. They found a new, intuitively powerful, result: the effect of changes in economic conditions differs whether the change—say an increase—affects a labor- or capital- intensive sector of the economy. In the former case, the grievance argument dominates, and conflict is less likely (in good times), because unemployment rises, and with it the opportunity cost of conflict, etc. In the latter, the greed argument dominates, because the incentive to fight for the pie grows with its size.[9]

A couple of years earlier, in 2004, MSS had published their paper, using rainfall fluctuations as an instrument for changes in income. They identified a causal relationship between economic conditions and the likelihood of conflict. The robustness of their main finding rested largely on the necessary assumption that an exclusion restriction was not violated. Specifically, in order for the exclusion restriction, and therefore for the claim, to hold, there has to have been solid evidence that the instrument—rainfalls—affected conflict *only* through its effect on economic conditions. Any evidence of a causal effect of rainfalls on conflict running through a different channel would weaken the central claim of the paper.

In the words of the authors:

> While it is intuitively plausible that the rainfall instruments are exogenous, they must also satisfy the exclusion restriction: weather shocks should affect civil conflict only through economic growth.[10]

So, the substantive question is: what if it turned out that people's movements were affected by weather? The underlying hypothesis here is that weather shocks could induce people to change their regular mobility patterns, which may increase the likelihood of conflict irrespective of the impact of weather on income. This could happen for example by leading people whose paths usually do not cross to actually come in contact, etc.

This requires stretching the evidence presented in this chapter a bit, but the general idea is clear. As hinted above, there is actually anecdotal evidence that real world processes consistent with the hypothesis may have actually played out. For example, in the case of Darfur, it has been claimed—and contested—that "general alterations in rain patterns have affected migration patterns of Darfuri nomadic tribes who breed cattle and camels. These changes subsequently led to increasing clashes between nomadic and sedentary farmers about the traditional land-tenure system."[11]

[9] Dube and Vargas 2007; Dube and Vargas 2008; Dube and Vargas 2013.

[10] Edward Miguel, Satyanath, and Sergenti 2004, p. 275.

[11] "The war in Darfur: Nate Barton" 2012.

Two additional points are worth adding. First, MSS 2004 did of course investigate whether the violation was likely or not. In the final version of the paper, the authors assessed "the possibility that high levels of rainfall might directly affect civil conflict independently of economic conditions." Reviewing a number of ways this could happen, MSS "acknowledge[d] that [they] [were] unable to definitively rule out the possibility that rainfall could have some independent impact on the incidence of civil conflict beyond its impact working through economic growth" although they "believe[d] that these other effects [were] likely to be minor."[12] Interestingly, though, changes in mobility patterns was not among the direct potential channels investigated, perhaps because there were simply no data to test it. Nonetheless, it must be noted that according to MSS results, "higher levels of rainfall are empirically associated with significantly *less* conflict in the reduced-form regressions," (their emphasis on *less*).

Second, the present chapter is not the first one to try and test the exclusion restriction in MSS 2004: in a recent paper, Sarsons (2011) evaluated it by identifying districts downstream from dams in India, where income is said to be much less sensitive to rainfall fluctuations. The author found that rain shocks remained equally strong predictors of riot incidence in these districts, suggesting that "rainfall affects rioting through a channel other than income," which "cast[s] doubt on the conclusion that income shocks incite riots.".[13]

The contribution of this chapter to the formalization of the field is arguably limited; however, it does suggest, as mentioned, a potential application of these new kinds of data to social problems, namely the possibility of testing and discussing the robustness of previous theories and findings on the basis of new sources.

3.2 Context: Conflict and Cell-Phone in Côte d'Ivoire

Côte d'Ivoire offers interesting grounds for analyzing whether CDR could help gain insights into socio-demographic processes, especially as they relate and shed light on the literature on drivers of conflict. The first reason is the history and drivers of conflict in the country—with large-scale violence that took place until 2011. The civil war (2002-07) and the post-electoral crisis (2010-11) in Côte d'Ivoire were both rooted in three intertwining drivers of conflict that are well known to other African nations.

One is the legacy of the colonial past and the bumpy road to democracy. In the thirty years following independence from France (1960), Côte d'Ivoire was a single party state. During this period, President Henri Bedié shared the benefits of rapid economic growth with different ethnic groups within the country, but concentrated political power within his own ethnic group, the Baoulé. In the 1990s, after opposition parties were legalized, the competition amongst political parties scrambling to rally support resulted partly in the politicization of citizenship and ethnicity.[14]

[12] Edward Miguel, Satyanath, and Sergenti 2004, p. 276.
[13] Sarsons 2011.
[14] Cederman, Gleditsch, and Hug 2012.

The second driver of conflict was fierce competition over resources in a largely agricultural economy. President Bedié introduced a controversial land policy in 1963 that gave rights to land ownership to "those who make it productive," an initiative designed to fuel the boom in cocoa and coffee, the two sectors driving the economy.[15]

The policy spurred significant migration from the northern half of Côte d'Ivoire and neighboring countries. When commodity prices plummeted in the 1980s, competition for jobs and land increased, as did anti-migrant and anti-immigrant sentiment. People who had moved to urban areas during the economic boom moved back to rural areas where immigrant and migrant workers were working on cocoa and coffee plantations.[16] The grievances over economic opportunities quickly evolved into ethnic grievances between Southern and Northern ethnicities and tribes. These features suggest that population distribution and movement may remain a significant element when studying drivers of conflict in the country.

Contested citizenship claims constituted a third driver. A key source of tension was the dispute over citizenship rights, notably access to land tenure rights and political voice. The concept of *Ivoirité* introduced in 1993 enabled the government to deny rights to Ivorians of immigrant ancestry, but was disguised as a patriotic doctrine to identify the national identity.[17] It created an excessive administrative burden for proving nationality and required that applicants provide proof of two parents of Ivorian nationality. Individuals perceived to be foreign or Northern, on the basis of ethnicity, were harassed at checkpoints, denied access to public services, and subject to land seizures.[18]

Further, a number of recent violent events were recorded precisely over the period for which CDR data are available, in areas that are the focus of the subsequent analysis given their localization and those of weather stations and cell phone towers, as described in the subsequent section. This is only meant to suggest that the question and this chapter's overall strategy appear sound.

The three events described in Figure 3.1 from the ACLED dataset[19] identify 10 instances of violence during the timeframe for which we have CDR data, all time-stamped, geocoded, and succinctly described (actors, number of casualties if applicable).[20]

[15] Mitchell 2012.

[16] Bah 2010.

[17] Bah 2010.

[18] Bah 2010.

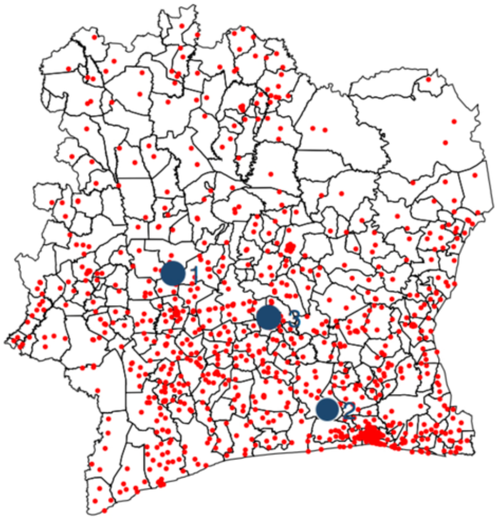
[19] Armed Conflict Location & Event Data Project 2012.

[20] Another major event occurred on April 24th, 2012 in Sakre by the Liberian border, but the fact that it happened only 3 days after the end of the timeframe for which we have data made it difficult to include it in the analysis, although it may have yielded very interesting results given the nature and scope of the event. A unidentified armed group attacked the village of Sakre on 24 April 2012, killing eight, burning ten homes, and prompting internal displacement within the region. According to a spokesman for the UN mission in Côte d'Ivoire, at least 250 people fled the village, and roughly 3,000 left villages in the region. Most sought refuge in the city of Tai, some 30 km away. A military source claimed the attackers were Liberia—based sympathizers of Laurent Gbagbo, the former Ivorian President currently held in custody by the International Criminal Court in The Hague for crimes against humanity. Four of the attackers were arrested.

Table 3.1: Côte d'Ivoire's recent history

Year	**Event**
1960	Independence from France
60's- 70's	Economic heyday; 7% GDP growth on av. (coffee & cocoa)
1963	Immigration policy grants land to "those who make it productive" – favors migrant coffee and cocoa farmers
Late 1980's	Commodity prices fall
	Migrant farmers perceived as responsible
1990	Opposition parties legalized – FPI only real opposition (Gbagbo)
	Gbagbo accuses Ouattara of being anti-Ivorian and foreign
1993	Boigny dies
	Ouattara's supporters form RDR
	President Konan Bedié introduces *Ivoirité*
1995	Uneasy alliance between PDCI and FPI; *Ivoirité* used to disqualify Ouattara
	Strong ethnicization of politicis
1998	New land code discriminates against foreigners and migrants from North
1999	Bedié overthrown in bloodless military coup led by retired General Rober Guei, who perpetuated ivoirité – Guei wished to run in next election as PDCI candidate but the party refuses (as he overthrew their group). He founds his own group.
	Debate continues over eligibility of Ouatarra.
2000	Guei declares himself president after election, but popular uprising forces him to leave
	Laurent Gbagbo declares himself new president after Guei forced out
2002	Ouattara issued proof by judge of his Ivorian citizenship
	September – coup attempt – rebels identify themselves later as representing MPCI
	Start of first civil war between the North (RDR – against Ivoirité) and South (FPI – pro-ivoirité) – PDCI lies low, uneasily allied with FPI
2003	>700,000 displaced by civil war
2005	New alliance forms – PDCI/Bédié ally with Ouattara and northern rebels against Gbagbo/FPI
2007	Ouagadougou Accord finally brings end to civil war power sharing deal with Gbagbo as Pres & Guillaume Soro as PM – representing the Mouvement patriotique de Côte d'Ivoire (MPCI)
2010	First election after end of conflict in '07 (had been delayed 6 times) Election dispute – Ouattara and Gbagbo : 2nd civil war.
	Ouattara assumes presidency
2012	An HRW report (Jan 2012) claims most refugees have not returned and that hundred thousand remain internally displaced
	Nov – Ouattara dissolved gov't after PDCI voted against new law that would recognize women as joint heads of household

Figure 3.1: Three violent events



Event #1: Vavoua, December 17th and 18th, 2011: On 17 December 2011, the Republican Forces of Côte d'Ivoire (FRCI) severely beat a young man attempting to avoid a road block in Vavoua, a city in the northern half of Côte d'Ivoire which had been under rebel control since 2002. The man died the same day in hospital. In retaliation, a group of youth armed with clubs and rifles attempted to storm the military camp on 18 December 2011. The government forces fired warning shots, but then fired directly on the advancing group, killing five of the young people.

Event #2: Lagunes / Sikensi, December 26th, 2011: A series of violent clashes in the Sikensi department of southern Côte d'Ivoire killed four and injured 15 on 26 December 2011. The violence began when an Abidji youth clashed with a soldier of the Republican Forces of Côte d'Ivoire (FRCI) in the village of Sikensi. The Abidji youth was killed, and news of his death served as a trigger for more violence. Residents of Katadji, a nearby village, attacked an FRCI post, resulting in the death of two soldiers. A subsequent clash on the same day took a turn toward inter-ethnic conflict when an Abidji youth knifed and killed a young Malinke man in the village of Sikensi. The Malinke are often viewed by local Abidji as synonymous with the FRCI.

Event #3: Yamoussoukro, March 15th, 2012: On 14 March 2012, national newspapers reported on rumors that pro-Soro supporters and the Ivorian Popular Front (FPI)[a] were collaborating to undermine President Ouattara's administration and destabilize the country. In a front page article, Le Nouveau Reveil claimed the FPI was taking a defiant posture with the Ouattara administration, and that "pro-FPI hardliners" within the military were sabotaging government plans. On the same day, Fraternité Matin reported on the Defense Minister's to areas in Abidjan to verify the security situation regarding rumors of plans to destabilize the national government.

Sources

Armed Conflict Location & Event Data Project. *ACLED Database for Côte d'Ivoire.* [Online]. 2012. URL: `http://www.acleddata.com/data/africa/`

Agence France Presse. "Violence in western Ivory Coast kills six: official." In: *Modern Ghana* (Dec. 2011). URL: `http://www.modernghana.com/news/367663/violence-in-western-ivory-coast-kills-six-official.html`

Agence France Presse (AFP). "Four dead in southern Ivory Coast clashes." In: *Modern Ghana* (Dec. 2011). URL: `http://www.modernghana.com/news/369010/0/four-dead-in-southern-ivory-coast-clashes.html`

Information Section of the Public Affairs Office of the American Embassy in Abidjan, Cote d'Ivoire. *National Daily Press Review.* [Online]. May 2012. URL: `http://photos.state.gov/libraries/cotedivoire/231771/Pdfs/NationalDailyPressReview_%20may2012_001.pdf`

[a] Guillaume Soro was the Ivorian Prime Minister from 2007 to 2012. The FPI is the political party of former President Laurent Gbagbo.

The second main reason why Côte d'Ivoire makes an especially promising case for such an investigation lies in the country's large cell-phone penetration. Mobile phone subscriptions rapidly multiplied in Côte d'Ivoire over the decade preceding the analysis, from roughly 473,000 in 2000 to 15.6 million in 2011. With a market penetration of 86%, mobile phones dominated the communication market; only 2.2% of the population uses the Internet, and the number of fixed line subscriptions was on the decline.[21]

Amongst the five major mobile operators, Orange and MTN led the market with 6.1 and 5.8 million subscribers respectively.[22] Most subscriptions are pre-paid, as opposed to post-paid contracts, which require, longer-term contractual agreements. Orange has four times as many post-paid mobile phone contracts as MTN, indicating a potentially wealthier customer base. However, with only 43,579 subscribers, this contract type represents a minor fraction of Orange's customer base.[23]

As in many developing countries, Ivorians are heavily reliant on mobile phones to stay abreast of security-related news and to touch base with family members when crisis occurs. Information spread by word of mouth is not always accurate, and a number of NGOs and UNHCR are working to improve access to accurate information amongst refugees, who are particularly vulnerable, *e.g.*, by disseminating crisis-specific information via SMS and radio.[24]

3.3 Data and Strategy

3.3.1 Data sources

The primary source of data consists of four datasets based on anonymized CDRs of phone calls and SMS exchanges between 5 million of Orange's clients in Côte d'Ivoire between December 1, 2011 and April 28, 2012. The datasets were provided as part of a research challenge,[25] and are described as follows:

1. Antenna-to-antenna traffic on an hourly basis;
2. Individual trajectories for 50,000 customers for two-week time windows with antenna location information;
3. Individual trajectories for 50,000 customers over the entire observation period with sub-prefecture location information; and

[21] Statistics drawn from: International Telecommunication Union (ITU) 2013.

[22] Number of subscribers as of 30 June 2012, as published in: Agence des Telecommunications de Côte d'Ivoire (ATCI) 2012.

[23] Atlantique Télécom dominates the post-paid market with roughly 906,000 subscribers in 2011, roughly one-third of their customer base, according to: Agence des Telecommunications de Côte d'Ivoire (ATCI) 2012.

[24] IRIN 2011.

[25] Orange 2012.

4. A sample of communication graphs for 5,000 customers.

The second source of data is the Climate Data Online (CDO) database of the National Climatic Data Center (NCDC).[26] Côte d'Ivoire has 10 weather stations, located in Figure 3.2 along with all cell towers located in neighboring areas—as described below. Weather observations are daily records of temperature in Fahrenheit at the weather station and precipitation in inches.

The third, somewhat less central, data source is conflict and violence data from ACLED.[27] The dataset identifies 10 instances of violence during the timeframe for which we have CDR data, all time-stamped, geocoded, and succinctly described (actors, number of casualties if applicable). We chose to focus on the largest 3 events, noted Event #1, Event #2 and Event #3, described in Figure 3.1.

The reason for looking at these events was to make sure that there was no complete geographic disconnect between the areas where the correlation between weather and movement was analyzed and the areas where recent violence took place. The 4th event on the Liberian border to the West was dropped because it occurred 2 days before the end of the observation period and did not lend itself to the same kind of analysis.

These three events are also located in Figure 3.2, which shows that these events occurred in areas very close to where the weather-mobility analysis is conducted. In other words, even if we cannot make direct inferences or references relating the weather-mobility relationship and these violent events, the underlying assumption is nonetheless that the weather-mobility discussion can be assumed to be relevant to an analysis centered on testing the MSS 2004 exclusion restriction, because violent events did happen in these areas.

3.3.2 Data treatment

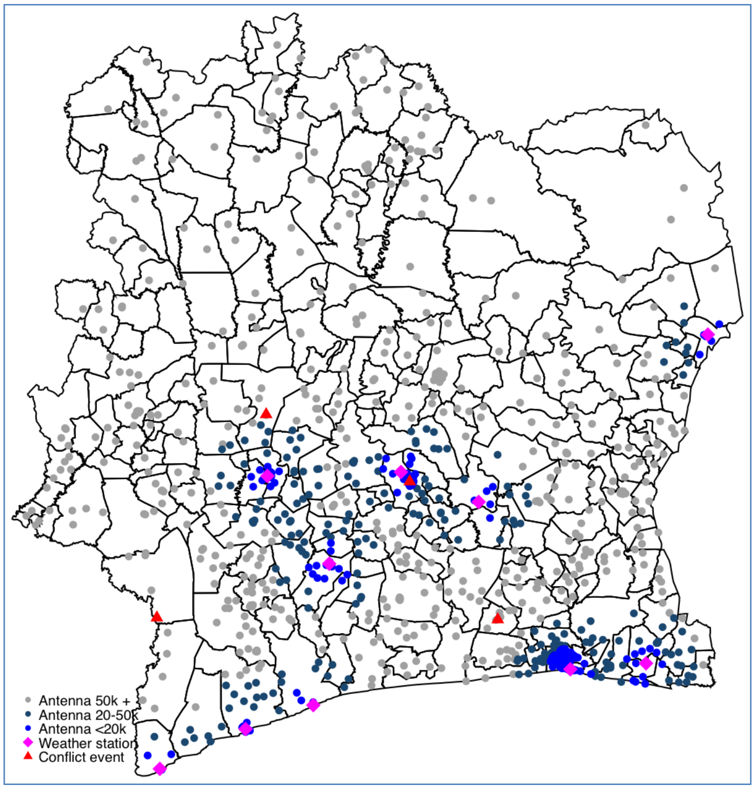
Cell phone data at the tower level are used, calculating daily aggregates movements around each relevant tower. To ensure that weather patterns affect areas for which corresponding CDR is available, we restrict the analysis to CDR data from cell towers located within a 50 km radius around any given weather station. This leaves us with 691 antennas, with observations over 140 days for a total of 75,636 antenna-day observations. All antennas found within 50 km of a given weather station are assumed to be affected by the same weather conditions. Note that there is no overlap between radiuses: in other words no antenna is located within 50 km of two different weather stations.

Movements are inferred from the distance (in kilometers) between subsequent connections to different cellphone towers for a sample of 20,000 individuals across the country using the haversine formula to estimate great-circle distances between the cell phone towers on a sphere from their longitudes and latitudes. Specifically, movements are calculated and 'assigned' to cell towers and thus weather stations in the following way:

[26] National Climatic Data Center (NCDC) 2011.

[27] Armed Conflict Location & Event Data Project 2012.

Figure 3.2: Weather stations, cell towers and recent conflict events in Côte d'Ivoire, 2011-2012



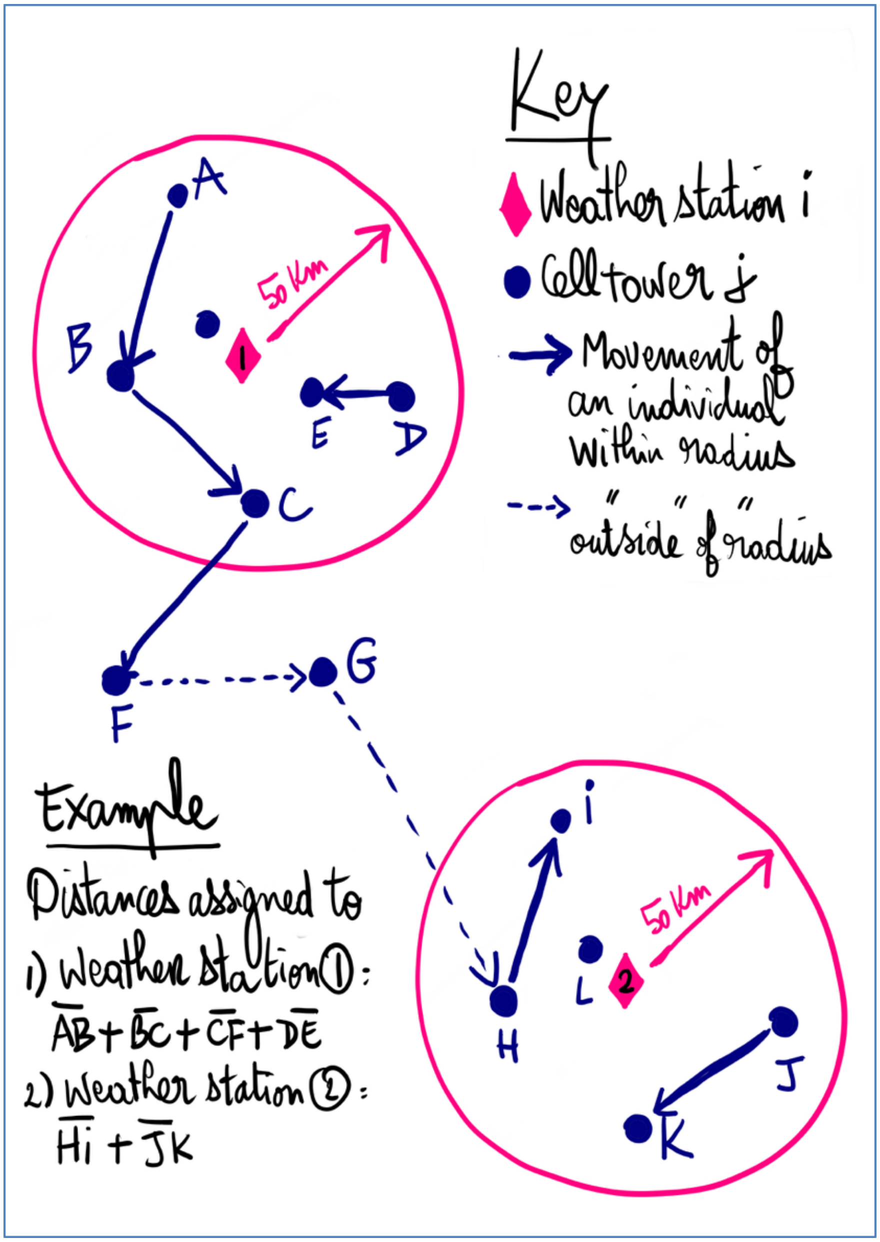
> *If a cell-phone activates tower A, then tower B (assume that B is 3 km away from A) and then tower C (assume that C is 2 km away from B), the initial move of 3 km is assigned to tower A, while the subsequent 2 km move is assigned to tower B.*

On a cautionary note, there are many other ways one could assign these movements, but since I am primarily interested in deviation from the 'normal' pattern, the specific choice of a mobility metric is unlikely to greatly affect the result. This is however something that may warrant further investigation. One drawback of this way of measuring mobility is that movements outside the 50 km circle that originated within it get unrecorded. For example, if a person moves from A to D, then D to E, and D is outside the 50 km radius, only A to D is considered to 'belong' to (taking place in) the surrounding area. Still, A to D, registered at A, may constitute a deviation from the normal pattern and is taken into account. The cumulative distances of all towers located within 50 km of a given weather station are then summed up and 'assigned' to that weather station.

This set up summarized in Figure 3.3.

Call duration and volume (*i.e.*, number of calls) are also added to check whether they are also affected by weather and add some depth to the story. For instance, it would be reassuring to find that call volume and duration are also affected, such that any effect on

Figure 3.3: Setup of the movement analysis



movement may look like part of a more general behavioral response to changes in weather.[28] And indeed they are.

Because movements are consistently varying across cellphone towers, and since we are mainly interested in deviation from the 'normal' trends, standardized values are calculated using a z-score approach for each antenna, giving them mean of 0 and standard deviation of 1.[29] Call volume and duration are also standardized using z-score method, to remove tower specific features and trends in call patterns. Weather data are also 'detrended' using weekly (7-day) moving averages centered around the observation day, drawing on the short-run economic fluctuations and demographic response literature.[30]

Descriptive statistics of all variables are provided in Table 3.2.

3.3.3 Summary Statistics

3.4 Strategy

First, plots of all relationships between weather and CDR variables were drawn, to get a sense of their features, intensity, direction, etc. Selected plots are shown below.

Then, to conduct econometric analysis and test whether there is any statistically significant correlation between weather variables and CDR variables—especially so population movements inferred on their basis—we use normal OLS with standardized measure of daily movement, number of calls, and duration of calls at the antenna level as dependent variables, calculated as described above.

The independent variables are temperature and precipitation. Controls for day of the week are also added.

3.5 Results

3.5.1 Visual representations

First, selected visual representations of the relationship between weather variables and population movement are provided, as well as of the relationship between weather variables and call volume and duration. The goal is simply to give a better sense of the data and relationships between the different variables.

[28] It is also theoretically possible, although less likely, that the fact that changes in weather conditions may affect call duration and volume could also directly affect the likelihood of conflict in a causal manner—or be indicative of changes in moods that may. But this is highly speculative and impossible to test with these data.

[29] z-score formula: $z = (x - m)/sd$ where x is the value for a day, m is the mean and sd is the standard deviation of that tower.

[30] See in particular: Wrigley and Schofield 1981.

Table 3.2: Summary Statistics

Variable	**N**	**Mean**	**SD**	**Min**	**Max**	**Sum**
Number of voice calls for antenna	75636	3932.92	3397.54	1.00	36180.00	297,000,000
Duration of voice calls for antenna	75636	496162.10	412286.70	14.00	3607277.00	3,750,000,000
Number of voice calls (z-score)	75636	0.00	0.99	-4.51	5.86	..
Duration of voice calls (z-score)	75636	0.00	0.99	-4.50	5.32	..
Distance moved out from antenna (km)	75636	470.75	427.09	0.00	8562.72	35,600,000
Standardized distance moved (z-score)	75636	0.00	1.00	-3.55	9.01	(62)
Temperature (mean daily)	75636	81.14	2.44	70.50	92.10	6,137,475
Precipitation (total daily)	75636	0.07	0.25	0.00	2.87	5,312
Distance from conflict event						
Event 1	75636	287.66	109.73	11.00	463.53	21,800,000
Event 2	75636	124.35	75.00	35.74	348.20	9,405,292
Event 3	75636	185.31	74.48	0.19	352.10	14,000,000
Event 4	75636	320.75	104.61	89.69	586.53	24,300,000
Conflict-date dummy						
Event 1	75636	0.00	0.01	0.00	1.00	2
Event 2	75636	0.00	0.00	0.00	1.00	1
Event 3	75636	0.00	0.02	0.00	1.00	29
Event 4	75636	0.00	0.00	0.00	0.00	-
Dummies for day of the week						
- Sunday	75636	0.14	0.35	0.00	1.00	10,570
- Monday	75636	0.14	0.34	0.00	1.00	10,407
- Tuesday	75636	0.15	0.35	0.00	1.00	11,013
- Wednesday	75636	0.15	0.35	0.00	1.00	11,103
- Thursday	75636	0.14	0.35	0.00	1.00	10,730
- Friday	75636	0.15	0.35	0.00	1.00	11,058
- Saturday	75636	0.14	0.35	0.00	1.00	10,755

Figures 3.4 and 3.5 provide visual representations of the relationship between weather variables and population movement for selected weather station; Figures 3.6, 3.7 and 3.8 do the exact same thing for call volume and duration.

3.5.2 OLS regression results

Table 3.3 provides the main finding of the chapter. It shows that population movement ($dist\_z$ variable) increases with both rainfalls and temperature. More specifically, it shows that higher rain and temperature levels are associated with greater than usual aggregate movements, in all specifications.[31] In other words, population movement is found to increase from their normal levels on wetter days, hotter days, and hotter and wetter days. All coefficients are highly statistically significant.

However, the R-squared are extremely low for all specifications. In other words, weather can be said to explain part of the variations in movement, but a very small part.

Tables 3.4 and 3.5 provide the results for call volume (no_z) and duration (dur_z). Overall, similar conclusions emerge.

[31] Note that all specifications include days of the week as controls.

Figure 3.4: Rainfalls vs. mobility for weather station 6 and all towers with 50 km

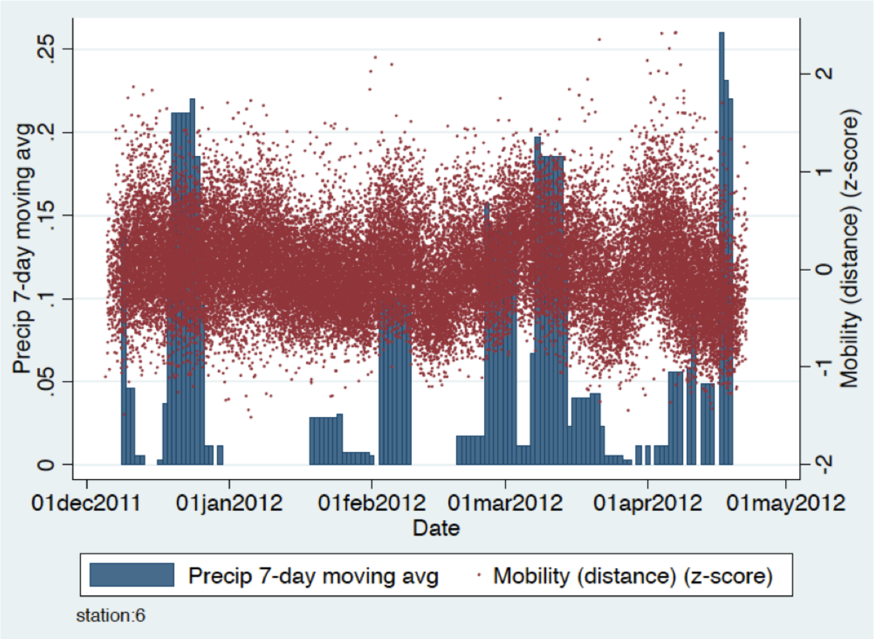


Figure 3.5: Temperature vs. mobility for weather station 9 and all towers with 50 km

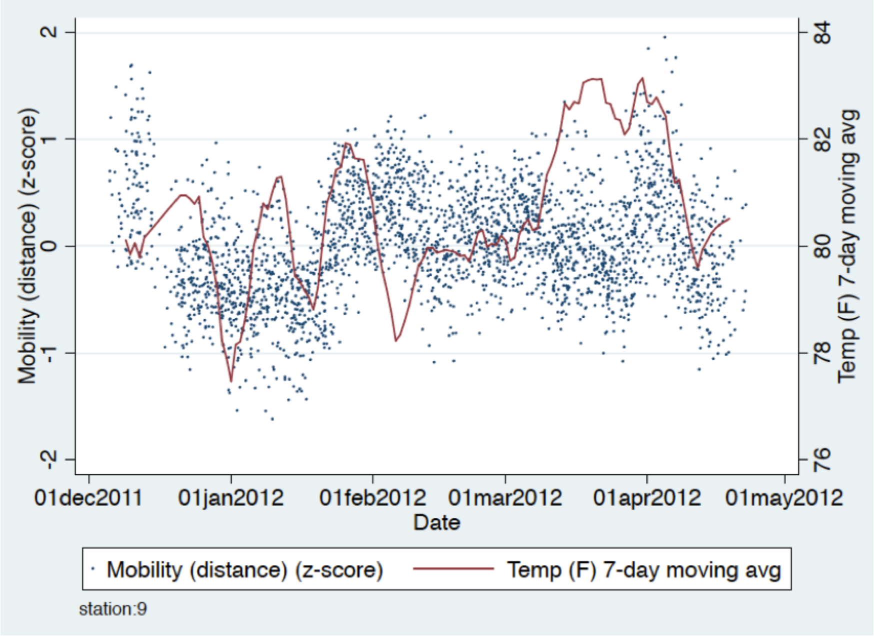


Figure 3.6: Temperature vs. call volume for weather station 1 and all towers with 50 km

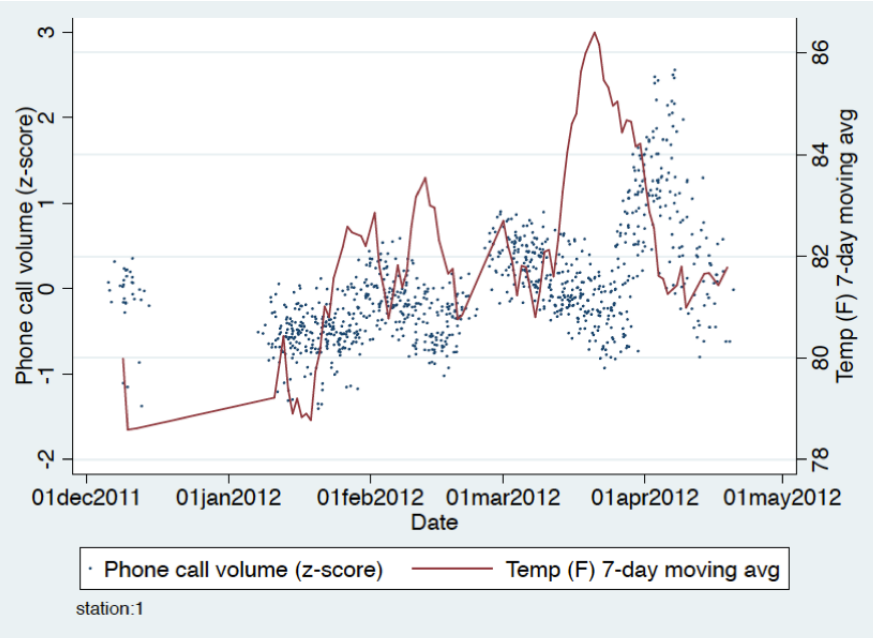


Figure 3.7: Rainfalls vs. call duration for weather station 6 and all towers with 50 km

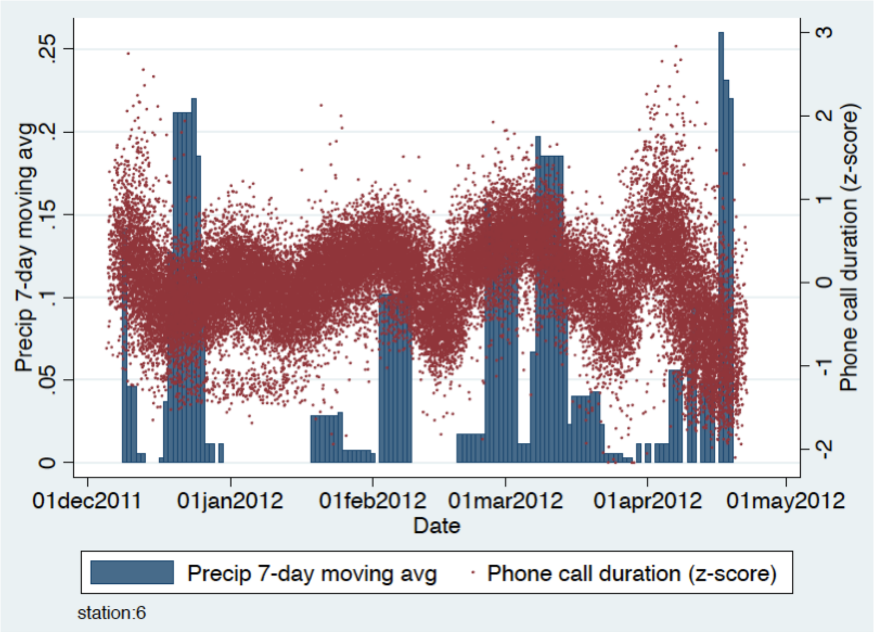


Figure 3.8: Temperature vs. call volume for weather station 6 and all towers with 50 km

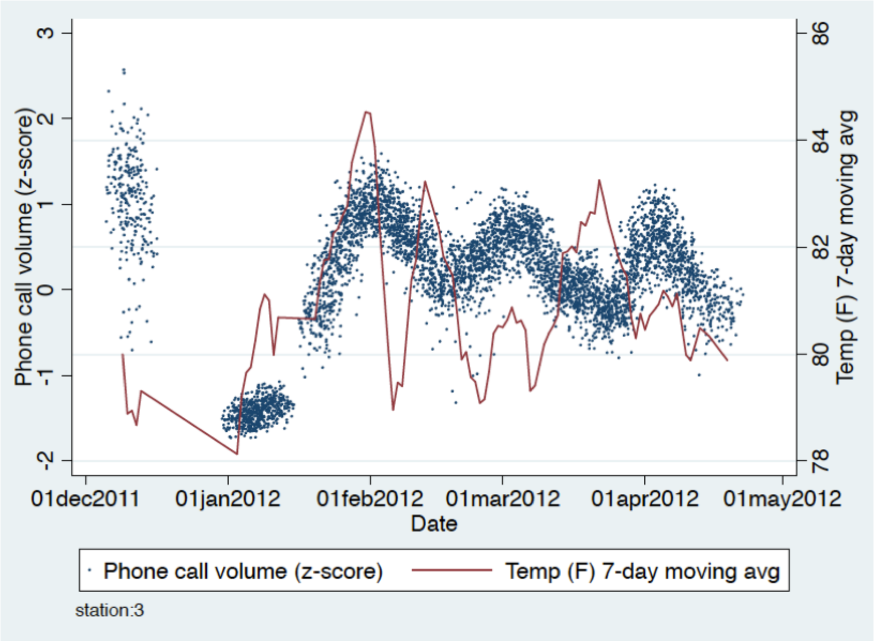


Table 3.3: Weather's impact on population movement

	(1)	(2)	(3)	(4)	(5)
VARIABLES	dist_z	dist_z	dist_z	dist_z	dist_z
Rainfall	0.105 *** (0.0159)	0.125 *** (0.0162)		0.150 *** (0.0166)	
Temperature	0.00437 *** (0.00148)		0.0100 *** (0.00146)	0.0132 *** (0.00149)	
Constant	-0.547 *** (0.121)	0.00957 ** (0.00376)	-0.813 *** (0.118)	-1.080 *** (0.121)	-0.186 *** (0.00994)
Observations	75,636	75,636	75,636	75,636	75,636
R-squared	0.026	0.001	0.001	0.002	0.026

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.4: Weather's impact on call volume

	(1)	**(2)**	**(3)**	**(4)**
VARIABLES	no_z	no_z	no_z	no_z
Rainfall	0.0725 *** (0.0140)	0.100 *** (0.0135)		0.113 *** (0.0142)
Temperature	0.000857 (0.00159) (0.0157)		0.00396 *** (0.00153)	0.00632 *** (0.00159)
Constant	-0.165 (0.132)	-0.00705 * (0.00376)	-0.321 *** (0.124)	-0.521 *** (0.130)
Weekday controls	YES	No	No	No
Observations	75,636	75,636	75,636	75,636
R-squared	0.018	0.001	0.000	0.001

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3.5: Weather's impact on call duration

	(1)	**(2)**	**(3)**	**(4)**	**(5)**
VARIABLES	dur_z	dur_z	dur_z	dur_z	dur_z
Rainfall	0.226 *** (0.0142)	0.165 *** (0.0138)		0.254 *** (0.0145)	
Temperature	0.0420 *** (0.00146)		0.0408 *** (0.00144)	0.0461 *** (0.00148)	
Constant	-3.370 *** (0.120)	-0.0116 *** (0.00376)	-3.312 *** (0.117)	-3.762 *** (0.120)	0.0569 *** (0.0119)
Observations	75,636	75,636	75,636	75,636	75,636
R-squared	0.026	0.002	0.010	0.014	0.015

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

3.6 Discussion and Implications

The key finding of this chapter is the strongly statistically significant and positive correlation between weather variables—precipitation and temperature—and population movements.

In light of the opening discussion, this finding seems to weaken the exclusion restriction in MSS 2004 that the only causal channel through which rainfalls affect the likelihood of conflict is income. Here, I do *not* find that rainfalls (and temperature) increase the likelihood of conflict through population movement. But I *do* find that weather conditions—both rainfalls and temperature—seems to induce, or are correlated with, changes in the intensity of population movements. In turn, there are various plausible stories one can think of as to why and how greater population movement may increase the likelihood of violence. Then, we may wonder whether this potential violation of the exclusion restriction should be reassessed in future research.

Many caveats and uncertainties remain. One, as was just mentioned, even taken at face value, this finding should not be readily interpreted as proving with any level of certainty that the exclusion restriction in MSS 2004 was indeed violated and that their main results are flawed. This analysis was conducted for a given country at a given period of time, many years after MSS's own analysis. For all we know and don't know, the relationships between rainfalls, income and conflict may just have changed in the past decade. Furthermore, I cannot claim that the relationships found in this chapter should hold for the entire country of Côte d'Ivoire; we have data for subscribers of Orange only, who may react differently from other subscribers and other people in the country, and we take a subset—hopefully representative, though—of these clients. In addition, there are many reasons why someone may not be convinced by the way population movements are measured in the chapter. First and foremost, what we observe are *cell-phone* movements, or even *SIM-card* movements, not directly individuals moving around. It is well known that people in developing countries often share phones and SIM-cards. So, people may not change their movement patterns; only the way they share phones and SIM-cards may.

However, interestingly, Miguel and co-authors have for the past few years explicitly abandoned using rainfalls as an instrument, precisely on the ground that the exclusion restriction was unlikely to hold.[32] The main finding of this chapter is consistent with this decision.

Further, improvements that could be made to the analysis would include a deeper look into autocorrelation and time series causality. Weather and conflict variables have a high likelihood of being autocorrelated as an event at time t for either of these variables is likely to have strong correlation to the event at time $t+n$. In improving on the current analysis, it would be necessary to select the best functional form for the data by first looking at the shape of the data. The effects of variable t on $t+n$ most likely diminish as n increases to infinity or perhaps there is a pattern on an annual basis for a variable such as weather.

A distributed lag model could be considered to reflect the changing effects over time.

[32] S. Hsiang, Burke, and E. Miguel 2013; Burke, S. M. Hsiang, and E. Miguel 2015.

Possible lag-models include a polynomial lag or a geometric lag model. In geometrically declining weights on past rainfall there would be two parameters estimated: the rate of decline and the size of the effect. A polynomial lag accounts well for collinearity. Going even further in selecting the functional form, it would also be possible to select the specifications more robustly through Wald tests to look at significance of the variables or Lagrange-Multiplier tests to see if variables may be missing from this assumption.

Despite these caveats and limitations, I think the present chapter is valuable for various reasons. To the best of my knowledge, this paper provides the first econometric evidence of a direct relationship between weather and human mobility. The paper also demonstrates how Big Data can be used to revisit theories and models developed using more traditional data.

It is also valuable for the possible leads and necessary next steps to which it points. First, to call for further research in both fields at the intersection of which it falls—the conflict literature and the Big Data literature—with perhaps a specific focus on mobility analysis. It also shows that much more work is needed on the methodological front when using big data streams, to build tools and frameworks than may help better understand the insights contained in these new kinds of data and 'validate' them. Because "Big Data is not about the data,"[33] only if we are collectively willing to invest time and efforts in formalizing the field will we contribute to "ensuring the data-rich future of the social sciences."[34]

[33] King 2013.

[34] King 2011.

Chapter 4

Poverty Estimations in Côte d'Ivoire[1]

Socio-economic and demographic data even on basic indicators as population size and income levels are notoriously lacking or unreliable in poor countries, where bad data and low development reinforce each other. The intersection of Big Data and official development statistics has received significant attention, with discussions around Big Data's potential to partly fill some key data gaps—for instance, to 'leapfrog' statistical systems in developing countries[2]—but relatively few empirical analyses have addressed this theme to date.

Drawing and expanding on a rapidly growing body of literature, this chapters focuses on predicting multidimensional poverty at the sous-prefecture and sub-national levels in Côte d'Ivoire using CDRs from Orange in conjunction with data from the 2013 DHS.

Despite limitations, this chapter adds value to the existing literature in at least two respects:

1. It validates measures derived from CDRs against up-to-date socio-economic (DHS) ground-truth data collected during the same period—the first half of 2012—which to my knowledge, has not been done before;
2. It constructs a multidimensional poverty index (MPI) at the household level that can be aggregated to the level of cell phone towers, therefore providing much more spatially granular poverty estimates than have been obtained to date for Côte d'Ivoire.

This chapter is structured as follows. It begins with an deeper diver than in Chapter 1 on the use of CDRs for official statistics, including a concise literature review of the state and limits of knowledge in the specific case of poverty estimations using CDRs (Section 4.1). It then describes empirical attempts at predicting poverty levels (Section 4.2) in Côte d'Ivoire, after which it concludes by commenting on some policy implications and limitations (Section 4.3).

[1] The chapter draws on an unpublished research paper I wrote as lead author with co-author Emma Samman (Overseas Development Institute) and inputs from Espen Beer Prydz (World Bank).

[2] Giugale 2012; Others have made similar remarks: Fengler 2013.

4.1 Socio-economic and poverty analysis: Context and concepts

As mentioned above, many development experts have lamented—explicitly or implicitly—the "statistical tragedy"[3] affecting developing countries in general and in Sub-Saharan Africa in particular. The term 'statistical tragedy'—a reference to the continent's "growth tragedy" of the 1990s[4]—describes the dearth of recent and reliable development data that is believed to hamper development policy and programming, according to the saying that you can't manage what you don't measure. Marcelo Guigale, then Director of Economic Policy and Poverty Reduction Programs for Africa for the World Bank Group, developed the following analogy:

> How would you feel if you were on an airplane and the pilot made the following announcement: "This is your captain speaking. I'm happy to report that all of our engines checked fine, we have just climbed to 36,000 feet, will soon reach our cruising speed, and should get to our destination right on time.... I think. You see, the airline has not invested enough in our flight instruments over the past 40 years. Some of them are obsolete, some are inaccurate and some are just plain broken. So, to be honest with you, I'm not sure how good the engines really are. And I can only estimate our altitude, speed and location. Apart from that, sit back, relax and enjoy the ride." This is, in a nutshell, the story of statistics in Africa.[5]

Overall, a good indicator of a region's poverty continues to be the absence of recent, reliable poverty indicators—placing them essentially "off the map."[6] The availability of reliable, up-to-date, disaggregated data covering a wider range of human development dimensions has improved over time—notably following the introduction of the Demographic and Health Surveys (DHS) programs in the mid-1980s and its subsequent scaling up, and in the light of systems set up to monitor the Millennium Development Goals (MDG) since 2000, but many gaps remain.

A statistic often missing is population size. Only 12 of 49 countries in sub-Saharan Africa have held a census since 2004,[7] and several African countries haven't had a census in three decades. Their population structure and distribution—educated guesses—are particularly problematic given that a small difference in estimated vs. actual demographic growth rates can make a big difference to estimates in a short amount of time.

According to Claire Melamed, from the Overseas Development Institute:

[3] Devarajan 2011.

[4] Easterly and Levine 1997.

[5] Giugale 2012.

[6] "Off the map" 2014.

[7] United Nations Statistics Division 2016.

> Most of what we think of as facts in development, are actually estimates. We have actual numbers on maternal mortality for just 16% of all births, and on malaria for about 15% of all deaths. For six countries in Africa, there is basically no information at all.[8]

Some of those data gaps have received considerable attention from the media and others. For instance it is now well-known that Ghana's GDP jumped by 40% in 2010 and Nigeria's by 60% in 2014—'overnight,' following a standard rebasing exercise. The gaps in poverty data appear to have been subject to less scrutiny—this is despite the fact that we can only track the evolution of $1.25 a day poverty in some 60 countries.[9]

These realizations have notably presided over the 2013 call for a "Data Revolution" by the High Level Panel on the Post-2015 Development Agenda established by United Nations Secretary-General Ban Ki-moon (although the reference to a 'data revolution' predated this call by as many as 7 years, as mentioned above), described in the following terms:

> [...] a new international initiative to improve the quality of statistics and information available to citizens. We should actively take advantage of new technology, crowd sourcing, and improved connectivity to empower people with information on the progress towards the targets.[10]

This UN process culminated in the publication of a report by a UN-appointed group of independent experts in November 2014 titled "A Word That Counts," which contains the following lines:

> [...] National capacity for data science must be developed to leverage opportunities in big data, to complement high-quality official statistics. [...] Applications of big data for the public good must be developed and scaled up transparently, demonstrating full compliance with applicable laws." Further, according to the report, "data gathered will need to be disaggregated by gender, geography, income, disability, and other categories, to make sure that no group is being left behind.[11]

As previously noted, frustration over the current state of the data landscape has also been fueled by the 'supply side'—namely, the growing availability of new kinds of data, new tools, methods, and evidence that Big Data in general and CDR analytics in particular may be able to alleviate in part the statistical tragedy. As mentioned in Chapter 1 too, the distinctive feature of CDRs—their ability to trace human behavior spatially and over time—has spawned both examples of how they can be used to improve policy and programming, as well as controversies over technical issues such as sample bias and broader political economy issues.

[8] Melamed 2014.

[9] Rodriguez Takeuchi and Samman 2015.

[10] *A New Global Partnership: Eradicate Poverty and Transform Economies through Sustainable Development* 2015, p. 21.

[11] *A World That Counts: Mobilising The Data Revolution for Sustainable Development* 2014, p. 23.

Some ‘use cases’ are cited in almost every single paper—how CDR analysis has been used to study malaria spread in Africa,[12] or socioeconomic levels in a Latin American city[13]—but there exist literally dozens of papers relying on CDR data covering various sectors, from public health to crime.[14]

As discussed in the preceding chapters, the various ways Big Data in general and CDR analytics notably may benefit development can be described through taxonomy contrasting four functions: descriptive; predictive—either for inferring (or nowcasting or ‘proxying’) one variable based on another, or forecasting a future event; prescriptive; and discursive. The analysis presented in this chapter sits primarily in the ‘proxying’ category, although other functions are present. Before presenting the analysis, the following paragraphs give a brief summary of the literature relevant to income levels.

The idea of using Big Data to predict in the sense of inferring poverty levels can be traced to a paper first released in 2009 which found that light emissions picked up by satellites could track GDP growth and proposed that they could supplement national accounting in data-poor countries.[15] This finding has been validated elsewhere[16], but there is also evidence that this relationship can fade once the penetration of electric lighting approaches saturation.[17]

Given the near ubiquity of cell phones around the world and the information they pick up about the collective behavior of users embedded in CDRs, they are a natural data source to provide granular proxies of poverty at a low cost and in a timely manner.[18] Of particular interest are methods that can analyze data aggregated to the level of cell towers rather than individual records, which somewhat reduce privacy concerns, and methods that are easy to interpret, so as to heighten the confidence of data users.[19] Experiments have been conducted using individual level cell phone credit purchases and calling patterns[20]. Among previous attempts, two papers have explored aggregated CDRs in relation to poverty in Côte d’Ivoire, which provide solid points of reference.

Smith-Clarke et al. (2014) report on two experiments they conducted in Côte d’Ivoire on

[12] Wesolowski et al. 2014.

[13] Soto et al. 2011.

[14] For a review, see in particular: Blondel, Decuyper, and Krings 2015.

[15] Henderson, Storeygard, and Weil 2009.

[16] See for example: Chen and Nordhaus 2011 and Olivia et al. 2014 (who use ‘gold standard’ data on electrification and economic growth for 5,000 sub-districts in Indonesia between 1992 and 2008).

[17] See: Kulkarni et al. 2011; Smith-Clarke, Mashhadi, and Capra 2014; McClellan et al. 2013.

[18] Smith-Clarke, Mashhadi, and Capra 2014; Jean et al. 2016; J. Blumenstock, Cadamuro, and On 2015.

[19] Smith-Clarke, Mashhadi, and Capra 2014.

[20] For example, Soto et al. 2011 use CDRs to predict poverty at the level of cell tower areas in a Latin American city with about 500,000 citizens, and compare their findings with official estimates – using information about the aggregated behavioral, social network and mobility of users, this approach predicted the socio-economic status of 80% of areas correctly. Gutierrez, Krings, and Blondel 2013 derived a proxy wealth indicator for Côte d’Ivoire on the basis of information on phone credit top-ups. These authors hypothesized that poorer people would be likely to top up their phone credit in smaller amounts and with greater frequency but their results have not yet been validated against any established wealth indicator (cited in (Smith-Clarke, Mashhadi, and Capra 2014)).

a so-called ‘Region B.’[21] For the former, they use CDRs on total traffic between cell phone towers for over 5 million phone users to construct geographically detailed income poverty maps, and to ‘ground-truth’ these data using 2008 poverty estimates from the International Monetary Fund (IMF) available nationally and for 11 subnational regions. For the latter, they use call records of around 928,000 mobile phone subscribers from early 2012. They followed a similar method in both cases:

First, they define Voronoi[22] areas associated with each cell phone tower[23], then aggregate CDR data to the level of each of the eleven subnational regions. Using the CDR data, they construct variables reflecting cell phone activity (call volume and duration), gravity residuals (as the difference between observed and expected flows between two areas, based on their population and the distance between them), network advantage (an entropy measure that captures call diversity between each two areas, and a measure of degree distribution), and introversion (the volume of traffic to other areas compared to traffic within each area). They found strong negative correlations between activity and poverty (around −.83 for Côte d’Ivoire, slightly lower in Region B), but as in the case of electrification noted above, cautioned that such a relationship may erode as the mobile telecom market becomes more mature.

They also found strong positive correlations between the mean negative residual of the gravity model and poverty levels (around .83 in Côte d’Ivoire, and slightly lower in Region B), and mostly strong negative correlations using the network advantage measures. The introversion measure was positively correlated in Côte d’Ivoire and negatively correlated in Region B—which they attributed either to distinct calling patterns in the latter or because of differences in the representativeness of the datasets (Region B having a smaller proportion of the total number of phone users).

Smith-Clarke et al. (2013) undertook a similar analysis but, interestingly, used the Multidimensional Poverty Index (MPI)[24] derived from 2005 DHS data for the purpose of ‘ground-truthing,’ as discussed in Chapter 1. A weakness is naturally the fact that the gap between the CDR data and the DHS data was about 7 years. They computed the same variables, and found that they had high predictive power:

- *Activity*: call volume and duration display strong negative correlations with poverty.
- *Gravity*: the negative residual of the gravity measure is strongly negatively correlated with poverty.
- *Diversity*: both measures display a strong negative relationship with poverty.
- *Introversion*: strongly positively related to poverty.

[21] To preserve confidentiality, Region B was described simply as ‘another developing region.’

[22] Weisstein 2016.

[23] The authors note that “the number of Voronoi cells intersecting regional boundaries is negligible with respect to the number of those fully contained in them”: Smith-Clarke, Mashhadi, and Capra 2014, p. 5.

[24] Oxford Poverty & Human Development Initiative (OPHI) n.d.(b).

The authors concluded by suggesting that "a valuable extension to our work would [...] be to obtain more up to date socioeconomic data."[25] Here, I undertake to predict poverty using CDR data in conjunction with socioeconomic data from the 2012 DHS collected roughly at the same time.

4.2 Empirical Application to Poverty Monitoring in Côte d'Ivoire

4.2.1 Data and Strategy

The analysis uses anonymized CDRs of phone calls and SMS exchanges between five million of Orange's customers in Côte d'Ivoire between December 1, 2011, and April 28, 2012. As described in Chapter 3, this dataset was released as part of Orange's 2012 Data for Development Challenge; the analysis uses two of the 4 datasets that were available:[26]

1. Antenna-to-antenna traffic on an hourly basis;

2. Individual trajectories for 50,000 customers.

The same data has been used to produce a wide range of socio-economic analyses. Some research focused on poverty, as discussed above. As noted, a common feature of these studies is a lack of up-to-date poverty data to test the predicted relationship between call patterns and poverty at a granular level. Our analysis is able to test the relationship between the CDRs and predicted poverty alongside observed welfare data and therefore 'ground-truths' the poverty estimates derived from CDRs.

This is made possible by the fact that the 2011/2012 Demographic and Health Survey (DHS) was conducted from December 2011 to May 2012,[27] almost exactly the same period as the CDR dataset[28]. Although the DHS does not collect consumption or income data, its measures of asset ownership, health, and education are commonly used in calculating the Multidimensional Poverty Index (MPI), originally developed by Foster and Alkire[29] and later adopted by the UNDP for their Human Development Reports.[30]

The multidimensional poverty index (MPI) offers insights into the extent and intensity of multidimensional poverty by combining ten indicators intended to reflect the poor's experience of deprivation. This chapter focuses on the multidimensional poverty headcount, the share of the population within an area considered poor (on the basis of being deprived in at least

[25] Smith-Clarke, Mashhadi, and Capra 2014, p. 6.

[26] Blondel, Esch, et al. 2013.

[27] Demographics and Health Surveys (DHS) Program 2013.

[28] The CDR dataset covers the period from 1 Dec. 2011 to 28 Apr. 2012: Blondel, Esch, et al. 2013.

[29] Oxford Poverty & Human Development Initiative (OPHI) n.d.(a). See also: Alkire and Santos 2013.

[30] Oxford Poverty & Human Development Initiative (OPHI) n.d.(b). We replicate UNDP's estimation of the MPI.

3 of 10 indicators, in other words having a weighted deprivation score of at least 33%). It also uses the measure for intensity of multidimensional poverty, which is the average share of deprivations experienced by people classified as living multidimensional poverty.

In order to normalize the call variables—*i.e.*, to look at their number and duration per capita—I used a raster dataset for population counts for 2010 at 2.5 arc-minutes resolution from the Center for International Earth Science Information Network (CIESIN), Columbia University.[31] I also used administrative population estimates at the sous-prefecture level obtained from UNOCHA.[32]

Voronoi polygons are formed by partitioning the map of Côte d'Ivoire into 1,214 cells with each cell containing an antenna and all the points that are closest to that antenna.[33] Population counts were then calculated for each polygon and used to create per capita call variables. Finally, MPI headcount data from DHS clusters were matched to their respective polygons, based on the closest antenna to the DHS cluster. The DHS clusters are therefore the primary unit of analysis, but I also aggregate the results up to more comprehensive domains. The DHS clusters have data from about 30 randomly selected households. Four sources of imprecision may skew the results.

First, for anonymization purposes, the geocoded location of the DHS cluster is displaced by up to 5 kilometers in the dataset. Second, a cellphone tower may not be located close to the DHS cluster—the median distance to the distorted cluster position to the closest cellphone tower is 3 kilometers, with 5% of the clusters being located more than 15 kilometers from the closest antenna. Third, I only have access to CDRs from one operator, whose penetration may be skewed across the country—a potential source of bias that cannot be controlled and corrected for in the poverty analysis with the available data. Fourth, MPI variables are estimated at the cluster level, which is not necessarily representative of the welfare level of the area covered by the closest cellphone tower from which we have CDRs. For simplicity, it is assumed they represent the same (or similar) populations (Figure 4.1).

To establish the relationship between the CDR data and the MPI score we run a set of simple linear regressions of the form:

$$MPI = \alpha + X\beta + \epsilon \tag{4.1}$$

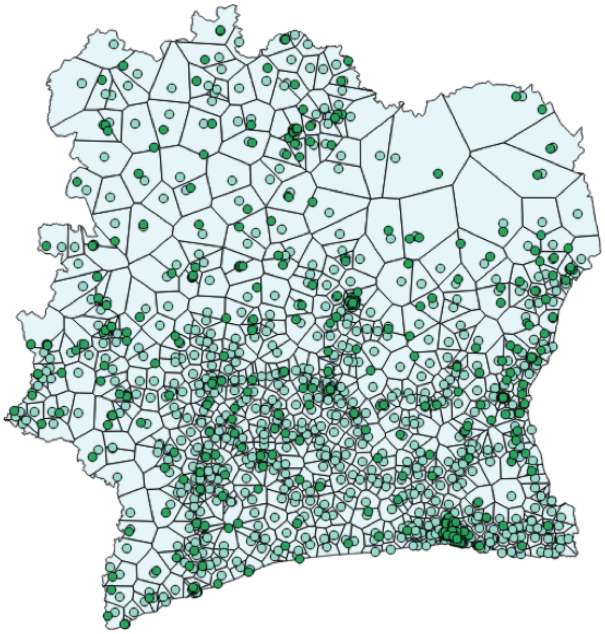
where MPI is a multi-dimensional poverty measure at the enumeration area cluster level (primarily headcount ratio at 33% deprivation) and X is a vector of (normalized) call variables from the CDR dataset at the cellphone tower closest to the cluster, taking various forms. Further, the CDR variables used pertain to outgoing calls, at the cellphone tower level.

[31] Center for International Earth Science Information Network (CIESIN), Columbia University, United Nations Food and Agriculture Programme (FAO), and Centro Internacional de Agricultura Tropical (CIAT) 2005.

[32] GeoHive 2014.

[33] Weisstein 2016.

Figure 4.1: Location of Voronoi cells, cell phone antennas (light circles) and DHS clusters (dark circles) in Côte d'Ivoire



4.2.2 Results

Tables 4.1 and 4.2 show the main regression results, primarily a negative relationship between outgoing call volume and the cluster level estimates of poverty headcount and poverty intensity. While the coefficients are strongly significant, the R-squared is around 0.3 for the model with headcount as the dependent variable and 0.15 for the models with intensity as the dependent model. Corresponding scatter plots in Figure 4.2 illustrate the results of column (1) in Tables 4.1 and 4.2. The somewhat limited predictive power of actual poverty levels at the cluster level is not unexpected (Figure 4.3).

Of key interest to policy makers is the extent to which such granular data can be used to predict the spatial distribution of poverty.

Recall that Smith et al.,[34] who used MPI measures derived from survey data for the year 2005 at the level of Côte d'Ivoire's eleven regions found strong negative correlations between the total outgoing volume ($r = -.774$, *p-value* $= .005$) and duration ($r = .791$, *p-value* $= .004$) of calls within a region and its MPI score. Their sub-regional estimates used diversity of call duration primarily, as this had the strongest correlation with poverty. They first derived a linear model using ordinary least squares regression on the MPI level for the eleven large regions of Côte d'Ivoire. This model was then used to estimate poverty levels at the sub-prefecture level, of which there are 255 in Côte d'Ivoire. The data at smaller levels of aggregation are noisier and thus result in weaker correlations.

[34] Smith-Clarke, Mashhadi, and Capra 2014.

Table 4.1: Regression results for MPI headcount and call volumes

	(1)	**(2)**	**(3)**
VARIABLES	MPI H33	MPI H33	MPI H33
Call duration per capita (log)	-0.106 *** (0.00913)		-0.421 *** (0.0991)
Call volume per capita (log)		-0.104 *** (0.00930)	0.318 *** (0.0997)
Constant	0.948 *** (0.0338)	0.440 *** (0.0216)	2.474 *** (0.479)
Observations	290	290	290
R-squared	0.319	0.301	0.342

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

H33 is Headcount assuming deprivation score of at least 33%

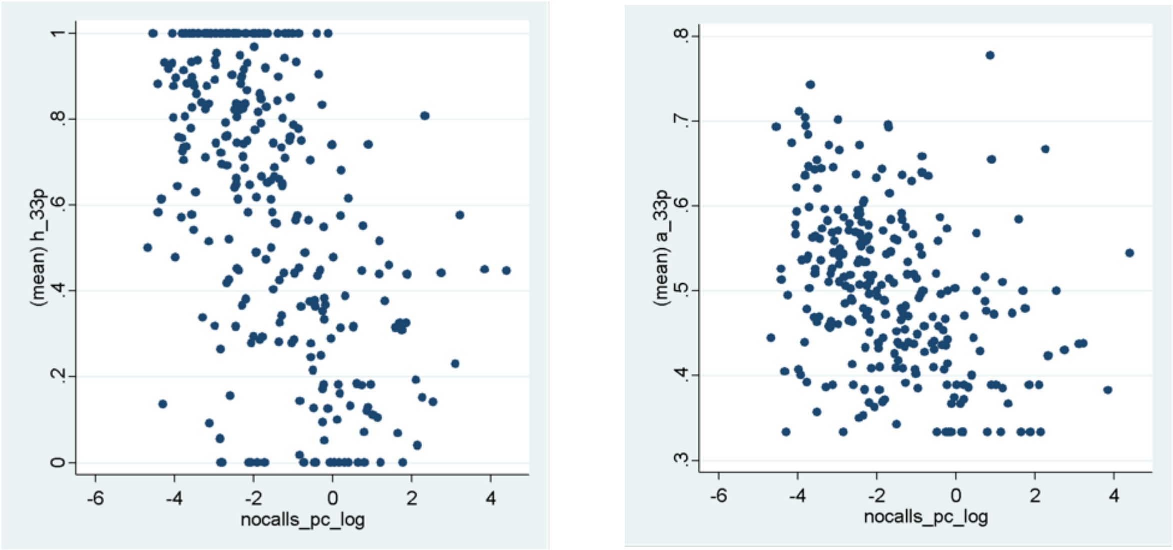
Table 4.2: Regression results for MPI intensity and call volumes

	(1)	**(2)**	**(3)**
VARIABLES	MPI A33	MPI A33	MPI A33
Call duration per capita (log)		-0.0802 ** (0.0345)	-0.0802 ** (0.0345)
Call volume per capita (log)	-0.0197 *** (0.00312)	0.0606 * (0.0346)	0.0606 * (0.0346)
Constant	0.467 *** (0.00743)	0.854 *** (0.166)	0.854 *** (0.166)
Observations	270	270	270
R-squared	0.130	0.147	0.147

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

A33 is Intensity assuming deprivation score of at least 33%

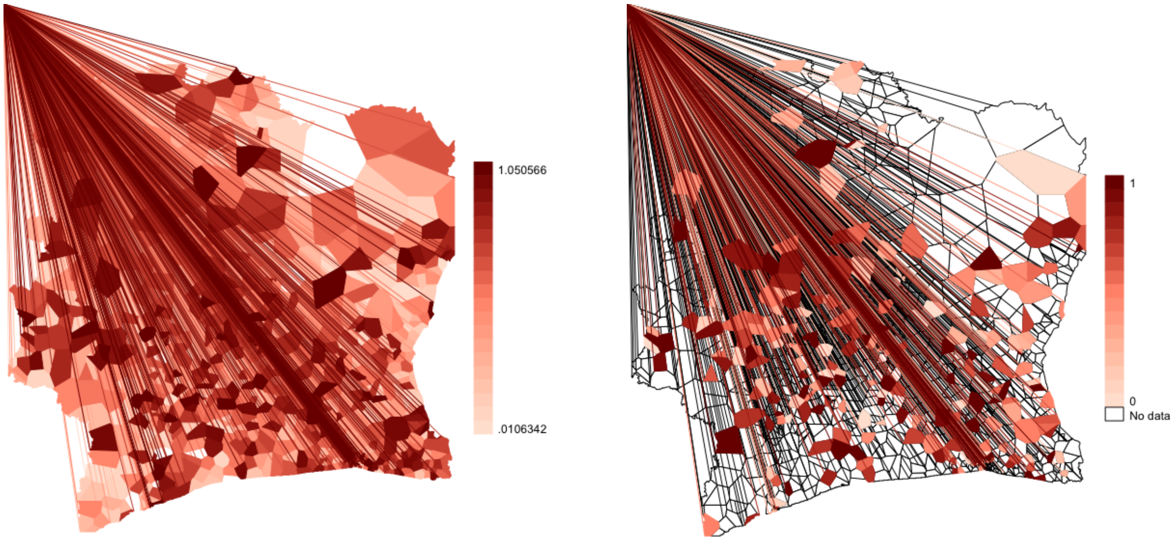
Figure 4.2: Scatter plot of call volume (number of calls per capita) and MPI poverty headcount



(a) MPI Headcount, assuming deprivation score of at least 33%

(b) MPI Intensity, assuming deprivation score of at least 33%

Figure 4.3: Observed and predicted poverty level at the cluster level: matched at the antenna level



(a) observed

(b) predicted

4.3 Discussion and Implications of the Côte d'Ivoire Application

The analysis points to several results as well as interesting avenues that will warrant further investigation.

This chapter provides additional support to a nascent body of evidence exploring potential linkages between CDR data and poverty measures including the MPI. In particular, it points to a negative and statistically significant but fairly weak relationship between call volume and the MPI (headcount and intensity) at the level of the 255 sous-prefectures. Further analysis will be needed to establish the relationship between other types of variables that can be derived from CDRs at the cell tower level and the MPI at the sous-prefecture level, and to analyse these relationships at higher spatial levels, for the 11 regions of Côte d'Ivoire. Its innovation is to show that DHS data can indeed be used to predict population at the sous-prefecture (and regional) level in Côte d'Ivoire despite uneven rates of cell phone penetration across regions.

Some other limitations must be noted. First, the paper does not address criticisms made to others of its kind, covered more extensively in Chapter 6—that these data were aggregated and used with little if any consent of the population that generated them. A risk that has not received the attention it merits is Big Data's potential to create a 'new digital divide' that may widen rather than close existing gaps in income and power worldwide. One of the 'three paradoxes' of Big Data is that because it requires analytical capacities and access to data that only a fraction of institutions, corporations, and individuals have, a 'data revolution' may disempower the very communities and countries it promises to serve. People with the most data and capacities will be in the best position to exploit Big Data for economic advantage, even as they claim to use them to benefit others. I discuss this at greater length in Chapter 6 of this dissertation.

Another challenge is that of putting the data to use. Most discussions about the 'data revolution' and the statistical tragedy assume that 'data matter,' *i.e.*, that poor data are to blame for poor policies. But one could argue that data in the hands of policymakers has historically played only a marginal role in the decisions leading to bad policies and poor outcomes. In short, I remain to be convinced that making even the best, most accurate, near-real time data available to governments would necessarily make a difference to the reduction of poverty and inequality. I want to stress that the challenge may lie more in the area of politics. Nonetheless, there are clear gaps and inequalities associated with the availability of data in different parts of the world, and these gaps need redressing, as one small part of the changes that are needed.

Chapter 5

The ABCDE of Big Data: Assessing Biases in using CDRs for Demographic Estimates—The Case of Population Density in Senegal[1]

This final empirical chapter zooms on a critical yet widely under-researched question for using 'Big Data' to produce socio-economic and demographic estimates: as mentioned at various points in the preceding chapters, the bulk of big data have not been produced for research purposes and the sample is typically not representative of the underlying population. Against this background, this investigation uses CDRs made available as part of the 2014 Orange Data for Development (D4D) Senegal Challenge—to estimate population densities in Senegal and contrast the results with the latest Census to understand the nature and size of the biases.

This chapter builds on and feeds into the now large body of research that has leveraged Big Data—understood both as new kinds of passively emitted data about people's actions and interactions and as powerful computing techniques—in general, and CDR analytics in particular, to infer a wide range of social, economic, and demographic indicators. Many papers have come out of the D4D Senegal Challenge—and indeed the winning paper was precisely about using these digital breadcrumbs to infer socio-demographic indicators.[2] However, to my knowledge, this is the fist time that Senegal's 2013 Census has been used for analytical purposes in conjunction with CDRs, to improve our collective understanding of sample selection bias and how to address it.

[1] This chapter draws on an ongoing paper written with Emilio Zagheni and Gabriel Pestre; an earlier version was submitted and accepted at the World Bank annual 2016 'ABCDE' Development Economics conference in June, which I presented in June. The title is a direct reference to the conference's acronym. See http://www.worldbank.org/en/events/2015/11/10/annual-bank-conference-on-development-economics-2016-data-and-development

[2] Bruckschen, Schmid, and Zbiranski 2014.

As discussed in Chapters 1 and 3 in particular, the attention paid to Big Data by the development policy and research communities in recent years has sprung from two main sets of factors that can be lumped under the broad categories of supply and demand. On the demand side is the dearth of good—*i.e.*, reliable, timely, disaggregated—data on development processes and outcomes that is believed to constrain development policy and programming. This has been compared to a "statistical tragedy" in the case of Africa—in reference to the continent's "growth tragedy" of the 1990s.[3] The availability of reliable, up-to-date, disaggregated data covering a wider range of human development dimensions has improved over time—notably thanks to the Demographic and Health Surveys (DHS) programs in the mid-1980s and the Millennium Development Goals (MDG) monitoring framework since 2000, but many data gaps remain as the world faces the uphill tasks of achieving the Sustainable Development Goals (SDGs).

The supply side refers to the expectations raised by the "Data Revolution"—and of "Big Data for development" as a general and fast emerging field of practice. In less than a decade since the early years when it was shown that GDP could be inferred "from outer space" using light emissions, Big Data's potential to address some of the world's most acute challenges has spurred a great deal of both excitement and skepticism. The debates surrounding the topic have many facets that go beyond the scope of this chapter—including political, ethical, and legal[4]—and with time a growing consensus is found on some of the most contentious issues.

But excessive claims about the real state of knowledge[5] and too little attention to the central issue of statistical representativeness have continued to fuel skepticism amongst many traditional social scientists—chief of which is demographers. I hope this particular piece of research will show promising avenues for better assessing and addressing sample selection bias in Big Data sources and help spur the interest of demographers and other social scientists in fulfilling Big Data's potential and producing methods to generate development estimates.

The rest of this chapter is organized as follows. Section 5.1 presents the context and rationale for the chapter in more depth; Section 5.2 introduces the data sources that we used: CDRs and Census data for Senegal in 2013; Section 5.3 presents the results that we obtained in terms of improving our understanding of the biases in CDRs for estimating population density in the context of the developing world; Section 5.4 discusses the possibility of applying these results to lower levels of disaggregation; Section 5.5 presents a difference-in-difference approach to reduce biases when estimating relative changes in population size over time. An illustrative example for the region of Dakar is offered. Finally, Section 5.6 provides a short discussion of our work, the implications, and the next steps.

[3] Easterly and Levine 1997.

[4] Letouzé, Vinck, and Kammourieh 2015; McDonald 2016.

[5] Sustainable Development Goals Blog 2016.

5.1 Contextualization and motivation: the promise and pitfalls of Big Data-based development estimates

As mentioned in the introduction, the idea of using Big Data to produce development estimates has been traced back to a frequently-cited paper which found that light emissions picked up by satellites could track GDP growth and proposed that they could supplement national accounting in data-poor countries.[6] This finding has been validated in other sources,[7] but there is also evidence that this relationship can fade once the penetration of electric lighting approaches saturation.[8]

In more recent years, the high rate of growth of cell-phones penetration and use around the world, as well as the richness of information about the individual and collective behavior of users embedded in CDRs, have made them the focus of a large number of scientific articles and debates. In particular, the fact that people move with their cell-phones has given rise to a whole strand of research attempting to infer population movement and distribution, both in crisis and non-crisis contexts.[9]

In particular, the value or having estimations of population movement, distribution, and potentially structure by age and gender, before, during, and after a natural hazard, is very high. Promising applications have been developed. For example, Pastor-Escuredo et al. studied how people moved before and after the major 2009 floods in the Mexican state of Tabasco. These reconstructions of the flood's impact were validated against the assessment of the flood area from Landsat-7 images, as well as official figures on the number of displaced people.[10] Other examples in the cases of post-earthquake mobility analysis using CDRs include the case of Rwanda[11] and Flowminder's work in Haiti and in Nepal.[12]

However, inferring population-wide estimates from a sample of the population—*i.e.*, the subset of people who own and use a cell-phone at any given point in time—requires an understanding of how well that sample represents the population from which is drawn. This kind of problem with digital data was first exposed in some length for the specific case of relying on crowdsourced data by Patrick Ball, Jeff Klingner, and Kristian Lum in March 2011.[13]

In this context, data from volunteers were used to infer building damage in Haiti after the 2010 earthquake. But the signals actually painted a misleading picture because proximity to damage correlated strongly (negatively) with people's willingness and ability to report it. The

[6] Henderson, Storeygard, and Weil 2009.

[7] See for example: Chen and Nordhaus 2011 and Olivia et al. 2014 (who use 'gold standard' data on electrification and economic growth for 5,000 sub-districts in Indonesia between 1992 and 2008).

[8] Kulkarni et al. 2011; Smith-Clarke, Mashhadi, and Capra 2014.

[9] Deville et al. 2014; J. E. Blumenstock 2012; Flowminder 2015.

[10] Pastor-Escuredo et al. 2014.

[11] J. E. Blumenstock 2012.

[12] Flowminder 2015.

[13] Ball, Klingner, and Lum 2011.

bulk of reports, controlling for building location, indeed came from *less* damaged zones.[14] A similar problem was raised after Hurricane Sandy hit the New York and New Jersey areas of the United States in 2012, when most of the Tweets about Sandy came from Manhattan—an area which was hit much less than other parts of New York and New Jersey.[15]

The issue is of course not limited to crisis contexts where the event itself affects the sample and its behavior. The possibility of making population-wide inferences from data from digital devices and services in the absence of additional information is limited by the simple fact that not everybody has access to these devices and services. Moreover, and critically, access and usage are not randomly distributed.

Overall, three main sources of biases can be identified:

1. *Selection bias*: people who use a cell-phone or who sign up with a specific carrier are not necessarily representative of the underlying population; some people text instead of calling, so they don't show up in call lists.

2. *Compositional changes*: the characteristics of the people in the sample change over time: some people start using their phone with the provider, some stop using a phone.

3. *Behavioral changes*: people change the way they use cell-phones over time for various reasons, and users may use their cell-phones in different ways during the week or during the weekend, when on vacation or at work, etc.

Valid estimates could be drawn from non-representative samples, without any correction, in some specific cases. For example, if the underlying population thinks in the same way about a specific issue. For instance, asking the micro-subset of billionaires whether they prefer dining with friends or plowing a field will likely yield similar results to those found in the population at large. But in most cases the 'signals' coming from the non-representative fraction of the population will not provide a good picture of the experience or perspectives of its whole because different people think and act differently.

Well-understood bias can nonetheless be accounted for and corrected to some extent by understanding how sample selection bias may skew the representation of each group in the sample (*i.e.*, the dataset) and 'unskewing' the data by giving more weight to certain observations (*i.e.*, entries in the dataset). This is commonly referred to as sample bias correction. The key is to bring in other variables (such as the proportion of people in each age group who use the technology in question, in the case of most digital data) and use them to account for the under- or over-representation of certain groups within the sample, in order to get a better picture of the whole population.

Early work in this area 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," the authors propose 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

[14] Ball, Klingner, and Lum 2011.

[15] Crawford 2013.

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.[16]

In order for these observations to provide a reliable estimate of global migration, Zagheni and Weber applied a sample bias correction method based on the following assumptions:

- "When Internet penetration is very high, then the population of Yahoo! users is highly representative of the entire population;"[17]
- There is an "over-representation of more educated and mobile people in groups for which the Internet penetration is low."[18]

They subsequently divided the observations according to gender, age group, and country, and applied a different correction factor to each group—the lower Internet penetration, the greater the expected bias, and the more correction is needed. They then calibrated their preliminary emigration estimates for 11 European countries against data on age-specific emigration rates published by Eurostat for those countries in 2009, in order to estimate a shape parameter for each subgroup. From these shape parameters, they then estimate emigration rates, by age and gender, for a large number of countries and discuss their results for two cases, the United States and the Philippines.[19]

The reliability of their correction factor depends a lot on the availability of ground truth data to calibrate the shape parameter. As the authors point out, using data for European countries with relatively high Internet penetration rates means that there is a larger uncertainty for developing countries in their model:

> Estimates for countries with high Internet penetration rates are not very sensitive to changes in [the shape parameter] k, whereas estimates for countries with low Internet penetration rates are. Since we only have statistics from European countries, the likelihood function with respect to the parameter k tends to be fairly flat, meaning that we have rather high uncertainty.[20]

A similar observation was made by Deville et al. regarding their estimates of population densities in France and Portugal: "applying the method to low-income countries where penetration rates are increasing rapidly but still exclude an important fraction of the population would require further sensitivity analyses of the impact of phone use inequalities, especially as marginalized populations also are the most vulnerable to disasters, outbreaks, and conflicts."[21]

[16] Zagheni and Weber 2012.

[17] Zagheni and Weber 2012.

[18] Zagheni and Weber 2012.

[19] Zagheni and Weber 2012.

[20] Zagheni and Weber 2012.

[21] Deville et al. 2014.

In other words, as mentioned in Chapter 1, correcting for sample selection bias is essential and sample bias correction requires solid ground truth data for calibration—making traditional representative datasets absolutely critical. The next sections introduce the data that was available and present the methodology used to address the issue of sample bias in the specific case of population density in Senegal.

5.2 Data

5.2.1 Call Detail Records from Orange

The analysis uses anonymized CDRs of phone calls and SMS exchanges between more than 9 million Orange customers in Senegal between 1 January 2013 to 31 December 2013. These CDRs were released as part of Orange's 2014 Data for Development (D4D) Challenge.[22] The D4D Challenge provided three different datasets:

- Dataset 1: One year of site-to-site traffic for 1666 sites on an hourly basis;
- Dataset 2: Fine-grained mobility data (site level) on a rolling 2-week basis with bandicoot behavioral indicators at individual level for about 300,000 randomly sampled users;
- Dataset 3: One year of coarse-grained (3rd administrative level) mobility data with bandicoot behavioral indicators at individual level for about 150,000 randomly sampled users.

The methodology in this chapter employs Dataset 3, which provides the complete call list for the 2013 calendar year for 146,352 users meeting both of the following criteria:

1. Users having interactions on more than 75% of days in the given period.
2. Users having had an average of less than 1000 interactions per week (since users with more than 1000 interactions per week were presumed to be machines or shared phones).

The CDRs provide the point of origin of the call at country's 3rd administrative level (*i.e.*, the *arrondissement*, of which there are 123 in this dataset), which allows us to estimate how many Orange callers were present in each *arrondissement* on a given day (given that the sample in Dataset 3 is representative of Orange subscribers).

5.2.2 Census data from ANSD Sénégal

Our population and demographic data comes from the *Recensement Général de la Population et de l'Habitat, de l'Agriculture et de l'Elevage* (RGPHAE), Senegal's official Census,

[22] de Montjoye, Smoreda, et al. 2014.

which is carried out by the *Agence Nationale de la Statistique et de la Démographie du Sénégal* (ANSD). The 2013 edition of the RGPHAE was conducted over the 21 day period from November 19 to December 14 of that year.[23]

The ANSD provided us with a one-tenth random sample of the Census, which contains datasets on individuals, households, emigration, deaths, agriculture, and livestock. The focus is mainly on variables from the individual and household datasets.

These data are used as a ground truth, to calibrate CDR-based estimates, which is made possible by the fact that, similar to the case of Côte d'Ivoire in Chapter 4, the Census and CDR data cover roughly the same period. The goal of this analysis is, on one hand, to develop a methodology for using CDRs to calculate census-like indicators of population and demographic characteristics; and on the other hand to study in depth how the CDRs paint a biased picture of the general population, and how some of those biases can be accounted for to make good use of the CDRs as a complement to the census.

5.2.3 Geolocation of the CDR and Census data

Senegal is administratively divided into 4 levels:

- *région* (region)
- *département* (department)
- *commune / arrondissement / ville* (CAV)
- *commune d'arrondissement / commune rural* (CACR)

It should be noted that the 3rd administrative level, generally referred to as *CAV*, includes *communes* and *villes* (generally large towns and cities, respectively), which are administered separately from communes. However, for historical reasons, each *commune* and *ville* generally lies within the geographic boundaries of a single *arrondissement*.

The CDR dataset assumed that the country was neatly partitioned into arrondissements at the 3rd administrative level, and used a scheme with 14 regions, 45 departments, and 123 arrondissements. A number of *communes* and *villes* were therefore grouped with their nearest arrondissement in the data we received from Orange. The Census data is geolocated to Senegal's 4th administrative level (*commune d'arrondissement / commune rural*), and is divided into 547 such areas.

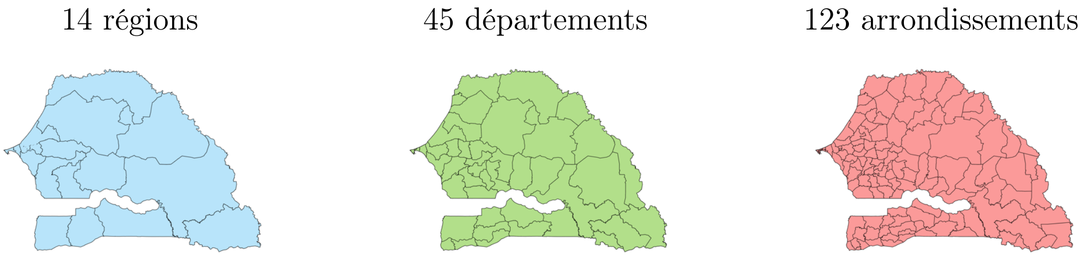
Using a combination of spatial merges in GIS, in-depth consultation of Senegal's laws pertaining to changes in the administrative division of the country,[24] and tables of administrative areas from the GADM database of Global Administrative Areas,[25] we were able to map all 547 areas in the Census data to exactly one of the 123 areas in the CDR data.

[23] Agence Nationale de la Statistique et de la Démographie du Sénégal (ANSD) 2013.

[24] République du Sénégal 1996; République du Sénégal 2013.

[25] GADM database of Global Administrative Areas 2015.

Figure 5.1: Administrative breakdown of Senegal used by the CDR dataset.



14 régions

45 départements

123 arrondissements

5.3 Estimating Population Density using Call Detail Records

5.3.1 The standard approach for evaluating population density

In the literature about modeling population density using CDRs, the standard approach relies on the following model:

$$\log(P) = \alpha + \beta\, log(U) + \epsilon \tag{5.1}$$

Where P is population size for a specific geographic area and time; U is the number of cell-phone users for the respective geographic area and time; α is a scale ratio parameter; β is the parameter that describes the superlinear effect in the relationship between users and population size; ϵ is a random error. The parameters are typically estimated using a regression model (see Deville et al.[26]).

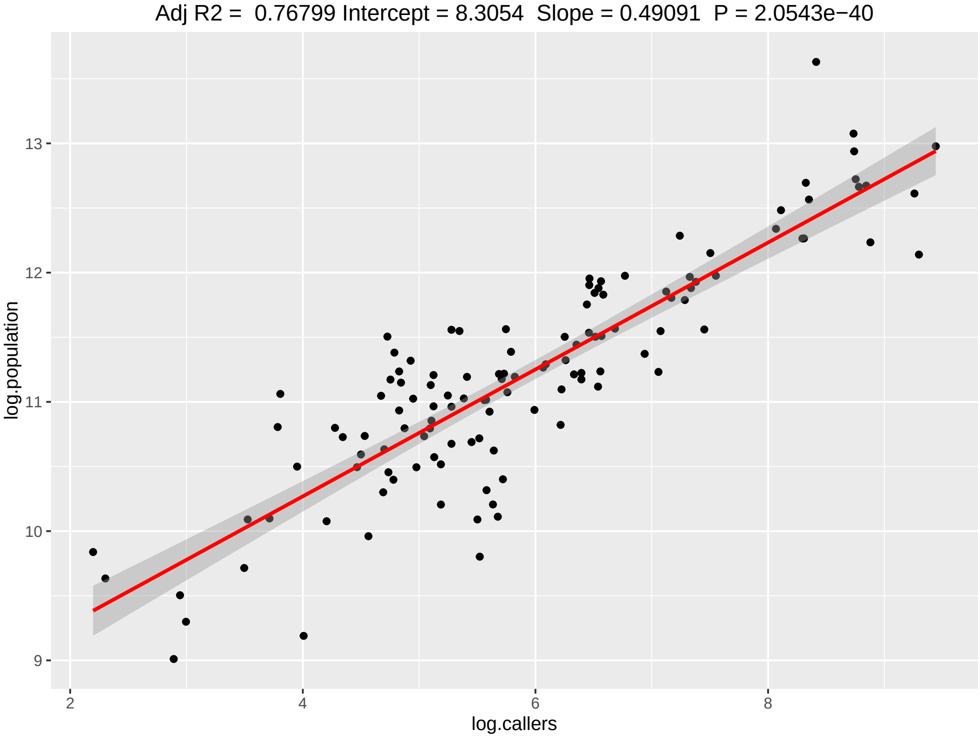
The model described in equation (5.1) has proven useful in the context of high-income countries, with rather uniform cell-phone penetration rates. We chose to use the same model for Senegal, although another functional form may be better suited, which should be tested in future revisions of this research.

That same baseline model performs quite well with our data for Senegal. Figure 5.2 shows an example of model fit for equation (5.1) using Census and CDR data for Senegal (2013). With an R^2 equal to 0.768, the relationship seems to hold fairly well in the Senegalese context.

However, when we looked into errors and spatial correlations, we observed systematic patterns. For example, Figure 5.3 shows ratios of population size and number of callers at the arrondissement level in Senegal. These values can be interpreted as the inverse of cell-phone penetration rates. An initial exploratory analysis indicates that there are some patterns in the distribution of cell-phone penetration rates, with clusters that form for areas that are geographically close, or that have the same type of urban vs rural setting, or with similar

[26] Deville et al. 2014.

Figure 5.2: Fit for the regression model of population size on number of callers (log-log scale) for Senegal (2013), following equation (5.1).



demographic characteristics. In the next section, we will discuss an illustrative example and some analyses related to this issue.

5.3.2 Using Census data to identify patterns of bias

Census data provide valuable information about socio-demographic and economic characteristics of each *arrondissement* in Senegal. This information can be leveraged to understand whether there are systematic biases in the relationship between population size and number of callers:

$$\log(P) = \alpha + \beta\ \log(U) + bias + \epsilon \tag{5.2}$$

Figure 5.4 shows the relationship between the number of callers and the actual resident population at the *arrondissement* level in Senegal. Figure 5.5 and Figure 5.6 show the corresponding relationships at the *département* and *région* levels, respectively. The data points are color-coded to indicate the average age of the population in each *arrondissement*, based on Census information. It is relevant to observe that there are systematic differences across age groups. The red data points, for areas with younger populations, lie mostly above the regression line. Conversely, the green data points, for areas with older populations, lie

Figure 5.3: Ratios of population size and number of callers in our CDR sample, for arrondissements in Senegal (2013).

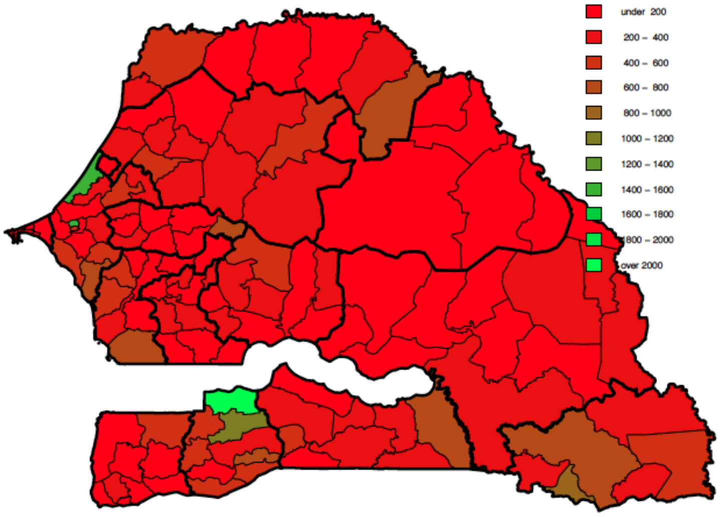


Table 5.1: Regression coefficients and associated standard errors for the linear model where the dependent variable is the logarithm of population size for each *arrondissement* in Senegal.

Intercept	10.644^{***}
	(0.393)
log(callers)	0.597^{***}
	(0.027)
mean population age	-0.135^{***}
	(0.021)

mostly below the regression line. In other words, using the standard model of equation (5.1) leads us to underestimate population size/density in regions with younger populations, and overestimate it in regions with older population age structure. This holds true at the *arrondissement*, *département*, and *région* levels.

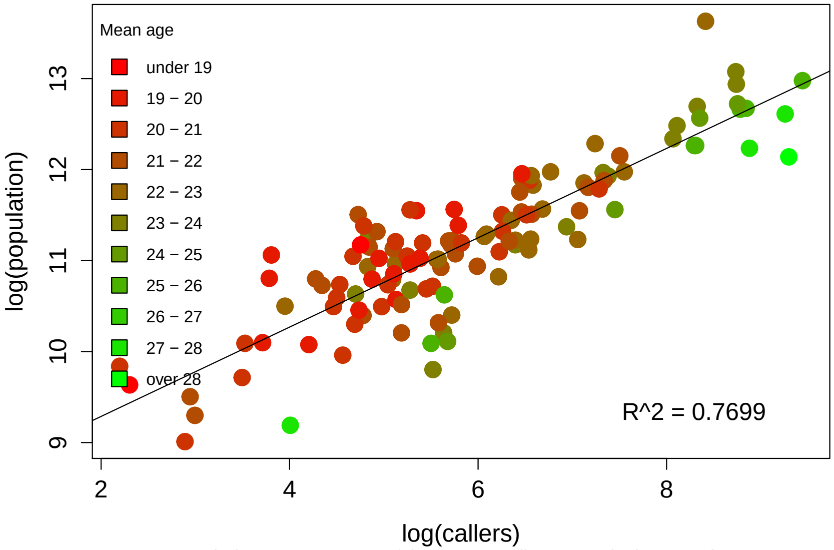
Including mean population age in the standard model at the arrondissement level, as specified in equation (5.3), significantly improves the fit ($R^2 = 0.827$):

$$\log(P) = \alpha + \beta \ \log(U) + \gamma \ mean.age + \epsilon \tag{5.3}$$

As Table 5.1 shows, the coefficient associated to average population age is negative and highly significant, indicating that, all else being equal, estimates of population size based on number of callers for regions with older population structure would be adjusted downwards as expected from the visualization in Figure 5.4.

The example that we discussed indicates that there are systematic differences in the relationship between callers and population size. Patterns emerge when rich data sets, like

Figure 5.4: Relationship between the number of callers (from CDR data) and the actual population (from Census data). The data points are color-coded to include the mean population age for each *arrondissement*.



Each dot represents one of the 123 arrondissements in the CDR data

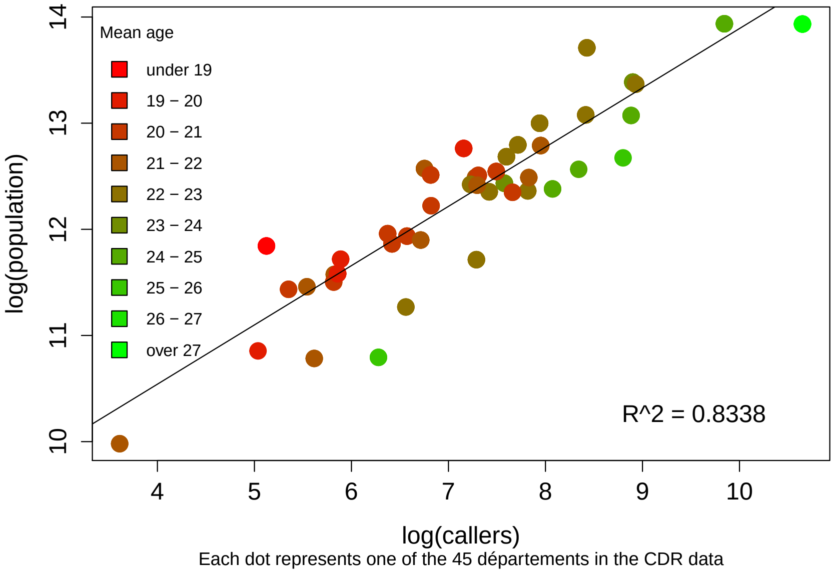
census data, can be used to complement CDRs. Age is obviously one of the most important demographic characteristics of a population. We showed that variations in age structure distort the relationship between number of callers and population size. Understanding the size and direction of distortions allows us to improve estimates of specific indicators of interest.

So far we have discussed an illustrative example using age structure. Other potential confounding factors that are related to socio-economic characteristics of users (*e.g.*, educational attainment) or behavioral differences in cell-phone use (*e.g.*, differences between weekdays and weekends or between different months of the year) may also be considered. In our case, we do not have access to demographic information about the cell-phone users themselves (as this would have to be provided by the carrier), but controlling for the socio-economic and behavioral characteristics of each administrative area can be leveraged to improve estimates of population density.

5.4 Projecting the regression coefficients down to lower administrative levels

In this section, we explore the possibility of applying the regression results from the previous section across various levels of disaggregation.

Figure 5.5: Relationship between the number of callers (from CDR data) and the actual population (from Census data). The data points are color-coded to include the mean population age for each *département*.

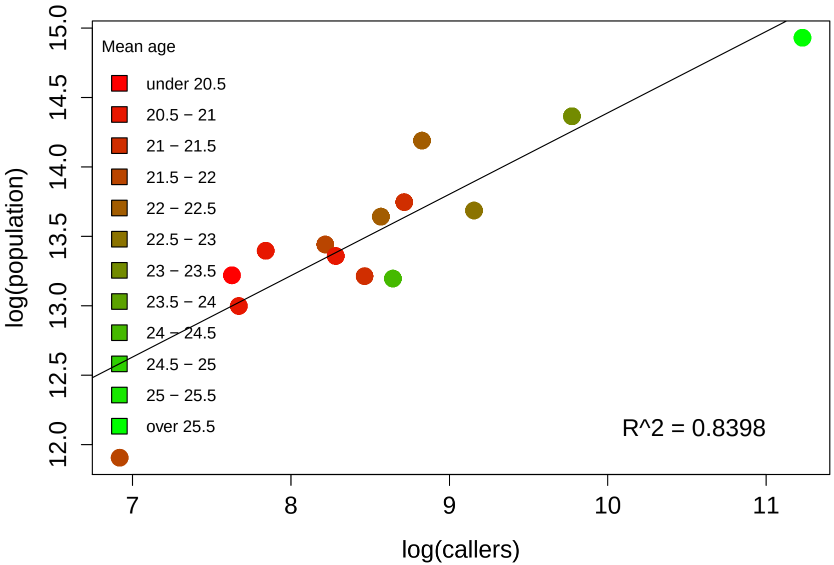


In order to see whether the coefficients calculated at a given administrative level could be projected down to smaller geographic areas, we calculated regression coefficients and fitted values at the *région* level, then used those coefficients to estimate population at the *département* level and *arrondissement* level. We also calculated the coefficients and fitted values at the *département* level and used them to estimate populations at the *arrondissement* level. Finally, we calculated the coefficients and fitted values at the *arrondissement* level. This was done for both the standard model in equation (5.1) and the model improved with mean age in equation (5.3).

We compared these population estimates to the Census populations for the corresponding administrative level, and calculated the mean absolute percentage error (MAPE) in each case. The results are presented in Tables 5.3 and 5.2.

We notice that with the standard model used in the literature, it seems hard to down-project. For instance, when estimating populations at the *arrondissement* level, the MAPE more the doubles, from 2.80% to 7.21%, when using fits from the *région* level (two levels up) instead of the *arrondissement* level itself. However, once we use information about mean age to improve the regression, as in equation (5.3), the marginal error still increases as we move to smaller geographic areas, but not at the same rate as it does without the additional information. Thus for the second regression, in the same situation, the MAPE only increases from 2.52% to 3.53%.

Figure 5.6: Relationship between the number of callers (from CDR data) and the actual population (from Census data). The data points are color-coded to include the mean population age for each *région*.



Each dot represents one of the 14 régions in the CDR data

These results suggest that by adding certain explanatory variables such as age to the regression, we make our ability to extrapolate populations at smaller geographic levels much more robust than in the standard model.

5.5 Estimating Population change over time using a difference-in-differences approach

In this section we use a difference-in-differences approach to evaluate the extent to which the population density in certain geographic areas changes relative to other areas, and over time.

We chose regions within the Dakar area that have very similar cell-phone penetration rates. Figure 5.8 shows trends in the average number of cell-phone users for the *arrondissements* Grand Dakar and Parcelles Assainies, over the course of a year. The two regions show trends that are almost perfectly parallel.

Figure 5.9 shows trends in average number of cell-phone users for the *arrondissements* Grand Dakar and Dakar Plateau. These two *arrondissements* are also part of the greater Dakar area. Dakar Plateau is an important center for commercial activity and tourism.

The trend lines are parallel for the period between August and January. During this period,

Figure 5.7: Fit for the regression model of population size on number of callers and age (log-log scale) for Senegal (2013), following equation (5.3).

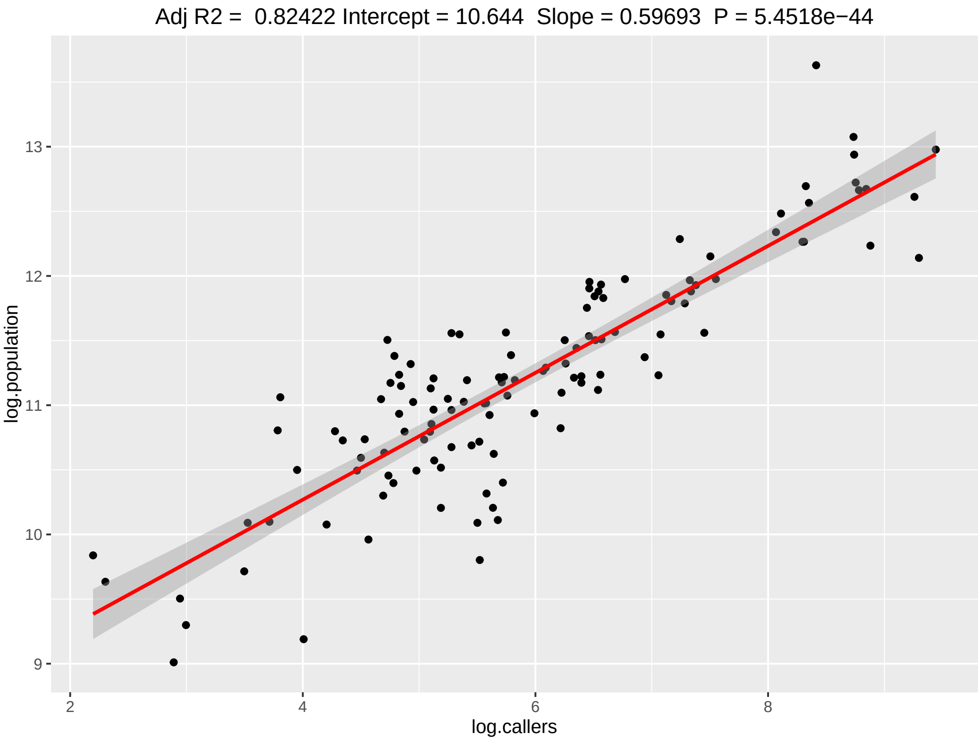


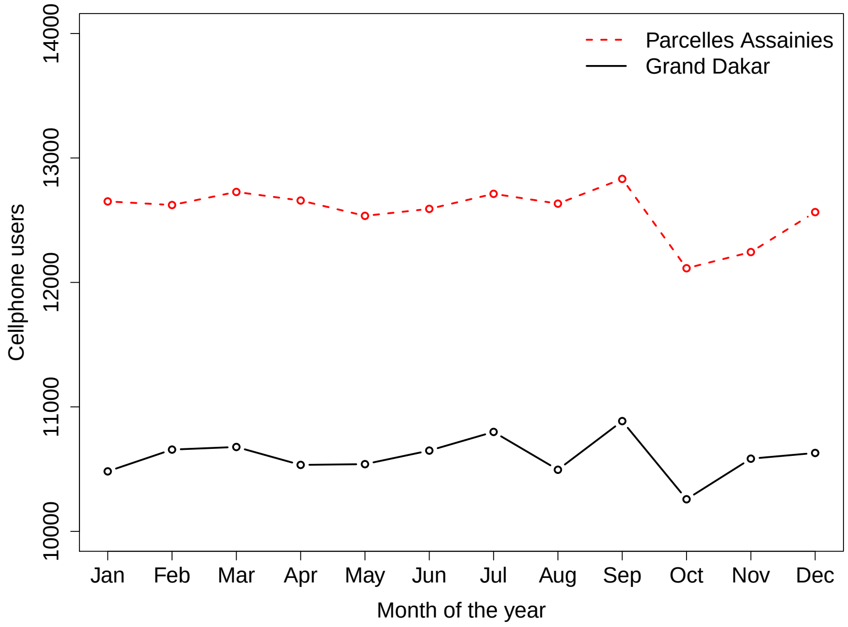
Table 5.2: Mean absolute percentage errors (MAPE) for the regression in equation (5.3).

	Estimates of log(population) at ...		
	arrondissement level	*département* level	*région level*
using coefficients fitted at ...			
arrondissement level	2.80%	–	–
département level	4.35%	1.90%	–
région level	7.21%	3.75%	1.51%

Table 5.3: Mean absolute percentage errors (MAPE) for the regression in equation (5.1).

	Estimates of log(population) at ...		
	arrondissement level	*département* level	*région level*
using coefficients fitted at ...			
arrondissement level	2.52%	–	–
département level	3.16%	1.68%	–
région level	3.53%	1.98%	1.21%

Figure 5.8: Average number of cell-phone users for the *arrondissements* Grand Dakar and Parcelles Assainies over the course of the year.



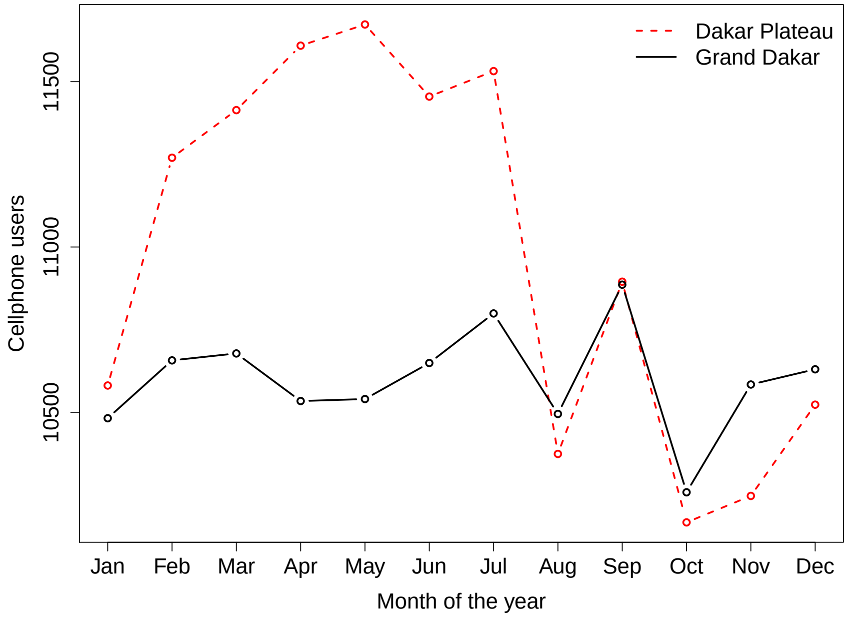
the trend is very similar to the one observed in Figure 5.8 for Parcelles Assainies. Between February and July, the number of cell-phone users in the Dakar Plateau rapidly increases, suggesting that there might be a seasonal pattern that differentially affects the Dakar Plateau. In order to evaluate the size of the effect, we estimated the following difference-in-differences model:

$$U_i^t = \beta_0 + \beta_1 G_i + \beta_2 T_t + \beta_3 G_i T_t + e_{it} \tag{5.4}$$

where U_i^t is the number of cell-phone users for the regions of Dakar Plateau and Grand Dakar, over time. G_i is an indicator variable that is equal to 1 if the observation is for the region Dakar Plateau, 0 otherwise. T_t is an indicator variable that takes the value 1 during the period from February to July, 0 otherwise. The difference in difference estimator $\hat{\delta}$ is equal to the estimate for the parameter β_3. The estimate for β_3 is 940.67 (s.e. $=$ 154.12) and is highly significant. This is a large change in population size: based on the results form the regression models estimated in the previous section, the change in population size would be in the order of about 100 thousand people.

Although further investigation would be required to determine the reasons behind these changes in population size, we expect two main factors contribute to these differentials. The first one is temporary migration of workers, who spend part of the year in the city, during the dry months, and part of the year in the countryside, during the rain seasons, as they help family members with agricultural production. As second potential explanation is related to

Figure 5.9: Average number of cell-phone users for the *arrondissements* Grand Dakar and Dakar Plateau over the course of the year.



flows of tourists. In any case, this comparison of Grand Dakar, Dakar Plateau, and Parcelles Assainies demonstrates that, even in the absence of ground truth, CDRs can be used to infer relative changes over time in population size at the subnational level.

5.6 Conclusions and Discussion

This chapter builds on and feeds into a growing body of research on how Big Data can be leveraged to produce socio-economic and demographic estimates. Using CDR data from Orange's 2014 Data for Development Senegal Challenge and Census data from Senegal's 2013 *Recensement Général de la Population et de l'Habitat, de l'Agriculture et de l'Elevage*, we investigated the possibility of combining traditional and new data sources to understand patters of bias in CDRs and to improve estimates of demographic indicators based on CDR data.

Our results demonstrate that many of the potential sources of bias in a CDR dataset can be better understood and accounted for, given sufficient ground truth. Starting from a simple log-log model relating number of callers to population data from the Census, we looked for other variables in the Census that had similar values across areas where the model consistently over- of under-estimated the population size.

We observed, for instance, significant differences in the relationship between number of callers and population density for *arrondissements* with different age structures, and used this to improve the predictive power of our model. We also explored how well the regression fits from a given geographic level can be projected down to a lower geographic level. Although our model is fairly simple—in this case, we take the log-log model traditionally found in the literature, and add only age—it appears to be fairly robust to this sort of extrapolation. We then developed an approach based on a difference-in-differences regression to evaluate relative changes over time at the subnational level, demonstrating that some inference about population size is still possible in the absence of ground truth.

This model is a first step to show that accounting for sample selection bias is possible, given sufficient ground truth data, and it could be extended in multiple directions depending on context. We hope that this work will show promising avenues for better assessing and addressing sample selection bias in Big Data sources and help spur the interest of demographers and other social scientists in fulfilling Big Data's potential and producing methods to generate development estimates.

Chapter 6

Legal, Ethical, Political and Institutional Implications and Requirements of Big Data and Demo-Economic Research and Policy in the Future[1]

While the bulk of this dissertation so far has focused on applications of CDR analytics, this final chapter will discuss what I think to be some of the key *implications* of CDR analytics for social science research, policymaking, and societies more broadly. As mentioned in earlier parts of this dissertation, by implications I mean ways in which CDR analytics has and can be expected to affect and be affected by a host of legal, ethical, political, and institutional factors and processes. In other words, this chapter touches on the political-economy of CDR analytics in a broad sense.

The potential of CDR analytics to advance our collective understanding of human dynamics, including population dynamics, and to positively affect human development and social progress, is hard to deny. But many uncertainties and challenges remain that for quite some time tended to be obscured and side–lined by narrow techno–scientific and short–sighted views. The development of CDR analytics for research and policy ends has indeed not been without creating controversies and experiencing backlashes that give a good sense of the many contentious questions and concerns it raises.

This chapter 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 at crafting the legal, technical, and institutional architecture of CDR analytics as a field of growing practice. After providing additional contextual elements (Section 6.1), the chapter summarizes current legal frameworks (Section 6.2) before exploring structural socio–political parameters and incentives structuring the sharing

[1] This chapter draws on papers co-authored with Patrick Vinck, Lanah Kanmourieh and David Sangokoya, all as lead author, as well as my own personal writings and various interactions and discussions over the past three years in particular.

of CDRs (Section 6.3), discussing possible guiding ethical principles (Section 6.4), and putting forward operational options and requirements for the future (Section 6.5).

6.1 Reflecting on selected CDR analytics challenges

6.1.1 It all started in Africa

I use in this section and its title the word ‘challenges’ in two senses of the word—’challenge’ as ‘competition’ and as ‘difficulties.’ A first example of a data challenge in the first sense is the first Orange ‘D4D’ challenge, the data of which I used in Chapter 3 and Chapter 4. This data challenge was organized in a few months in 2012 and culminated at the Third NetMob Conference with a 1–day dedicated event at MIT in May 2013.[2] Its organizers were overwhelmed by the response, which revealed and spurred huge interest in the field of CDR analytics. A total of 250 teams submitted proposals and received the data, of which 83 submitted papers. 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—including myself. The success of the first edition led Orange to organize a second challenge in 2014–15, focusing on Senegal.[3]

For all its success, the first D4D challenge also raised controversies and concerns:

1. First, there was not a single submission from an Ivorian team, exemplifying vividly the underlying issue of differential capacities and awareness that led boyd and Crawford, among others, to stress, as mentioned in Chapter 1, the risk of “a new digital divide.”[4]

2. Second, when the organizers asked by a show of hands who in the audience, among those who had submitted a paper, had ever been to Côte d’Ivoire, I was one of fewer than half a dozen. This exemplified the technocratic tendency to think that human problems can be solved remotely through VPN (virtual private network) access, with no deep contextual and local knowledge.

3. Third, none of the results led to any concrete implementation in Côte d’Ivoire to benefit the people whose data were used, highlighting the persistence of what is referred to a ‘response gap’ in the humanitarian sector in particular—the disconnect between information and action.

4. Fourth, and critically, the initiative also raised questions about issues of justice and ethics including aspects of consent and privacy. In fact, none of the people represented in the data ever *truly* gave their informed consent for these data to be used.

[2] *NetMob 2013* 2013.

[3] Orange n.d.

[4] boyd and Crawford 2011.

The fiercest critics denounced it, privately, as a neocolonial initiative by well-meaning Westerners who did not understand nor care about Africa and were simply using an imbalance in power to exploit local resources—here, data—with little involvement of and benefits to local populations.[5] 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. But this did not stop some critics from pointing to many unanswered questions:

1. In the age of data abundance, when secondary and future uses of the data are unknown, how do we assess the risks and benefits of releasing data—and who should govern this assessment?
2. What is 'informed consent' when most users sign forms and check boxes without ever reading the terms and conditions they 'consent' to—particularly in developing countries when access to such a critical tool as a cell phone hinges on consenting?
3. Who owns the data, or the rights to the data?
4. How can populations participate and weigh in on this research, and further, benefit from its outcomes?

The Ebola crisis brought these tensions to a whole new level in a sequence of events that had a lasting impact on debates about the future of CDR analytics. When the Ebola epidemics broke out in Sierra Leone, Liberia, and Guinea in 2014, several pundits and media—chief of which The Economist—reckoned that the crisis made opening up CDRs a near moral imperative, and blamed poor coordination for the absence of effective action.[6] Kenneth Cukier in a public event in January 2016 stated that under such conditions, "not using the data was the moral equivalent of burning books."[7] Others felt differently, arguing that these countries' political, economic, and historical characteristics raised significant concerns as to the potential misuse of the data, especially in such volatile times and in their aftermath. Experts in humanitarian assistance questioned the usefulness of the insights and maps that may result from the analysis to improve response on the ground when these countries' already fragile public health systems and public institutions more broadly, as well as international responders, were already overwhelmed by many competing demands.

In the end, after dozens of conference calls over many months eventually involving over fifty participating organizations—including several UN agencies—none of the research teams and institutions that sought to gain access to the data got permission from the relevant local authorities. This episode and its lessons led to the publication of a report titled "Ebola, a

[5] An MIT researcher whom I talked to during the event told me bluntly "You are French. Think of what would have happened if these were data on French users. There would be riots in Paris."

[6] "Ebola and big data: Waiting on hold" 2014.

[7] Data-Pop Alliance 2016.

Big Data Disaster" in March 2016, which received significant attention and partly reignited lingering tensions within this relatively small community.[8]

Other much less publicized failed attempts at accessing CDRs offer other valuable lessons, particularly on political economy aspects. 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, 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 Egyptian National Telecom Regulatory Authority made requests that the partners were unable or unwilling to meet; these included 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 foreign nationals.[9]

These cases give a sense of the shortcomings of the current state of the field and challenges moving ahead, summarized below.

6.1.2 Summary of key challenges

A first 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 telecom operators on the basis of personal connections and other arrangements—or for data challenges at their will. 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.

Some have introduced and described the general notion of "data philanthropy,"[10] which 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 benefiting consumers and economies.[11]

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 be a pragmatic approach conveying the idea of data being a public good, but there is a risk that what even its proponents describe as a temporary tactical move

[8] McDonald 2016.

[9] This is based on a direct written account of the World Bank manager in charge of the project.

[10] Kirkpatrick 2011.

[11] For more on data philanthropy, see for example: "Sharing Data As Corporate Philanthropy" 2014

may turn into a paradigmatic shift. We too swiftly forget the characteristics of public good, and the fact that knowledge, not data, is itself a public good.

Another set of challenges relates to individual (and group) privacy 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' of previously anonymized or aggregated datasets has been known for years (when multiple datasets are combined, one of which contains an ID).[12]

The high degree of predictability of human behaviour is what makes CDRs valuable, but also what creates risks of 'de-identification.' For instance, a paper attempting to derive the 'maximum predictability' in human movement confirmed that human mobility was indeed highly predictable,[13] 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."[14] This study showed that coarsening the data–aggregation further only required additional data points to re–identify people out of aggregated datasets, meaning that "even coarse datasets provide little anonymity." 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 endeavour.

This realization has gone on par 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,[15] although as mentioned above it only 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."[16]

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 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.

Differential ownership of analytic *capacities* is another major challenge and factor that may contribute to creating and widening a new digital divide between and within countries.

[12] Narayanan and Shmatikov 2008.

[13] Lu et al. 2013; Song et al. 2010.

[14] de Montjoye, Hidalgo, et al. 2013.

[15] Becker et al. 2013.

[16] de Montjoye, Shmueli, et al. 2014.

As noted above, bridging this digital divide will require investments in human capital, new technology, infrastructure, geospatial data, and management systems.[17]

Lastly, and fundamentally, another challenge is the lack of agreed–upon political parameters and 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 frameworks and principles 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—in recognition of and informing current legal frameworks.

6.2 Legal frameworks to protect privacy: status, limits and prospects

As the following section argues, the use of CDRs must be grounded in respect for the applicable prevailing 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. The technology and the law must evolve in a dynamic relationship to one another, guided by the policy goals and ethical considerations described above. A quick overview of the key legal environments for CDR collection and processing will give us a clearer picture of the rules currently binding telecom operators, while evincing the need for publicly debated legal reform. This section focuses 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.

6.2.1 International law on privacy

Although international law is short on privacy, privacy is not absent altogether from international law; 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

[17] *A World That Counts: Mobilising The Data Revolution for Sustainable Development* 2014.

stronger, clearer international norms 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,"[18] 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 of course 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.[19] Challenges currently pending before it could have the result of further strengthening the right to privacy among signatories.

Of course, this is binding only for parties to the ECHR. 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, and the UDHR and ICCPR are difficult to enforce.

6.2.2 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.[20]

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 CDRs, is not covered by the Federal Wiretap Act.[21]

[18] European Court of Human Rights 1953.

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

[20] *Smith v. Maryland (1979) 442 U.S. 735, at 745-746* n.d.

[21] Parrish 1977; Following Robert Pikowsky's definition, a pen register is a device that can be attached to

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."[22] 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.[23] In addition, under ECPA, placing a pen register 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."[24]

6.2.3 EU law on privacy

By contrast to the U.S.' piecemeal approach to data privacy, the EU 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.[25] Despite this, its privacy protections come with important loopholes.

The first relevant document is not an EU act, but rather the 1981 Council of Europe Convention on the Protection of Individuals With Regard to Automatic Processing of Personal Data,[26] ratified by 45 countries as of 2016.[27] It is not only broad 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, 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.

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

a specific phone line for the purpose of covertly recording the outgoing telephone numbers dialed, see: Pikowsky 2003.

[22] United States Congress 1986.

[23] United States Congress 1986.

[24] United States Congress 1986.

[25] Whitman 2004.

[26] Council of Europe 1981.

[27] Council of Europe 2016.

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...[28] Data can only be processed for the purposes specified.[29] Sensitive data (religious, racial, political, sexual, and health-related) benefit from additional protection.[30] Furthermore, the data subject must be informed of the processing[31] and has a right to access the data and, in some cases, to rectify or erase it.[32] There is, however, one crucial caveat. Article 3 of the EU Directive explicitly excludes law enforcement and security-related data processing from the scope of the act.[33]

The 2006 Data Retention Directive later fleshed out the rules that apply to data retention for law enforcement and other security purposes. Adopted in the wake of terrorist attacks in London and Madrid, it clearly leans further than previous texts in the direction of security over privacy. In fact, where law enforcement and security concerns are at play, it adopts the reverse principle to the 1995 directive: it creates an obligation to retain user data. The directive requires states to ensure that service providers retain certain categories of data for purposes of investigation, detection, and prosecution of serious crime. This does 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 limits the people who can access the stored data (article 4). The period of retention must last at least six months and at most two years (article 6).[34]

6.2.4 Reforming the legal framework for data collection and processing

The Snowden revelations sparked a jumble of reform proposals and a flurry of judicial activity. In April 2015, Amnesty International and other human rights groups brought a claim against the government of the United Kingdom for indiscriminate surveillance practices.[35] The EU Data Retention Directive has also provoked a backlash from constitutional courts and the European Court of Justice (ECJ) that makes its future uncertain. In April 2014, the ECJ ruled that the directive was invalid for being incompatible with the right to private life. In October 2015 in Schrems v Data Protection Authority—the case challenging Facebook's use of a user's personal information—the court invalidated the longstanding Safe Harbour agreement, nullifying a fifteen-year agreement between US companies and the EU. Additionally, after several years of debate and work on a new Data Protection package following a call to reform in 2012, the European Commission adopted a new data protection reform package, including

[28] European Parliament and Council of the European Union 1995, article 2 (a) and article 7 (a) (f).

[29] European Parliament and Council of the European Union 1995, article 6 (c).

[30] European Parliament and Council of the European Union 1995, article 8.

[31] European Parliament and Council of the European Union 1995, article 10 and article 11.

[32] European Parliament and Council of the European Union 1995, article 12.

[33] European Parliament and Council of the European Union 1995, article 3(2) and article 13 (1) (a)-(d).

[34] European Parliament and Council of the European Union 2006.

[35] Howell 2015.

the General Data Protection Regulation (GDPR) and the Data Protection Directive for Police and Criminal Justice Authorities. This new legislation serves as an update to the data protection rules based on the 1995 Data Protection Directive and the 2008 Framework Decision for the Police and Criminal Justice Sector. Despite these recent developments, the GDPR remains vague across several elements, leaving room for interpretation and not always providing sufficient guidance for businesses and supervisory authorities. For instance, the rules applying to profiling based on big data are only vaguely articulated. It is yet to be seen how businesses and the public will adapt to these new standards, and several recent discussions reflect the many conflicting commercial, political, and ethical interests still at play.

The underlying principles of the new regulation involve informing the data subject and, at times, obtaining their consent, purpose limitation, and data minimization. Critics note the limits of these policy tools in a big data world where the use cases often evolve after the actual collection of the data. Also, in view of the amount of data sources and networked nature of data processing, it seems inevitable that, while recognizing the need for transparency and reasonable choice, the emphasis will have to move from the actual collection of data into ensuring responsible and accountable use of data.

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. With CDR analysis now fast-growing, it is indispensable that its potential applications be considered in setting this balance.

However, in order to avoid built-in obsolescence, stakeholders should heed the lessons of past attempts at legislating on data: instead of focusing rules and regulations on specific technologies and uses, the conversation should take root in broad, strong principles regarding users' rights and protections and clear guidelines on the ethical and security considerations to shape any use of their information.

6.3 Political Parameters

Understanding the politics or political economy of sharing CDRs as individual or aggregate records can usefully be informed by broader considerations on data collection and data sharing in the era of Big Data. More specifically, it can be informed by three extreme positions that are bounded by data collection and data sharing considerations:

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 times, but no data are 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 governments when considering intelligence data gathering.

3. A third position where all data are collected and made public at all times, reflects the idea that social 'public good' value can be yielded 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.

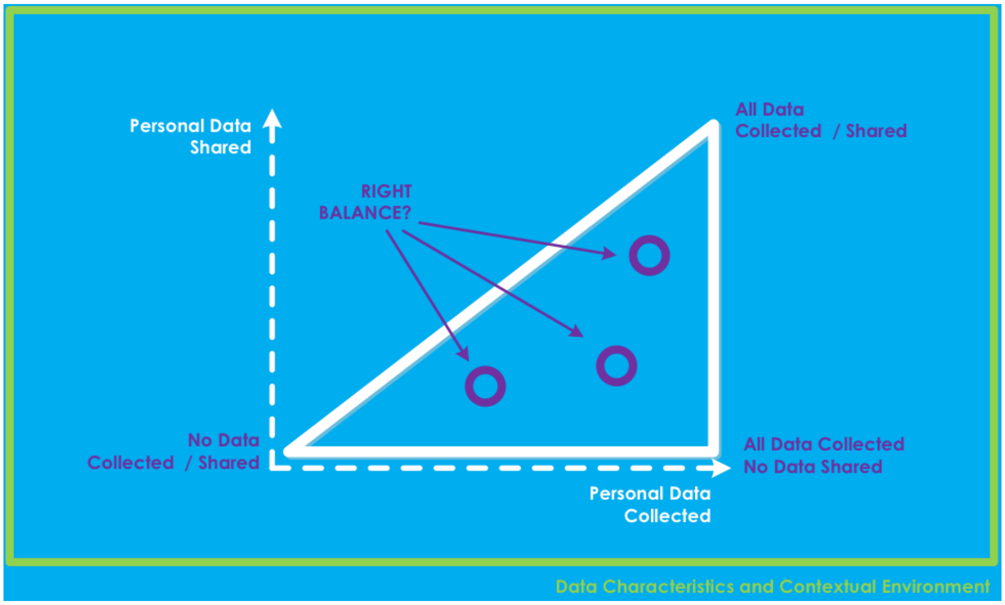
None of the three positions described above are realistic, nor are they desirable: critically, these are *ideal–typical* (or *stereotypical*) categories meant to facilitate the exposition of kinds, and not actor–specific, concerns. All stakeholders, including users, corporations, and governments have an interest in finding a right balance on how much is collected, how much is shared, and how—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, given the commercial value and financial and technical barriers this would raise. Likewise, having all data collected and shared at all times is unlikely to be desirable in any circumstances.

These three positions delineate a bounded "data collection and sharing" triangular space 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:

Figure 6.1: Data Collection and Sharing Space


Personal Data Shared
All Data Collected / Shared
RIGHT BALANCE?
No Data Collected / Shared
All Data Collected No Data Shared
Personal Data Collected
Data Characteristics and Contextual Environment

Note that the vertical axis also captures the level of temporal and spatial granularity.

Source: Elaboration of the author.

1. The features and characteristics of the data being shared, including the risk of de–identification, level of aggregation, and perception of sensitivity. For example, how much data can be collected and shared may depend on whether the data are seen as sensitive, touching on perceptions or feelings, as opposed to economic data such as consumption patterns;

2. Contextual characteristics, identified here as *systemic* and *idiosyncratic.*

- Systematics 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 regime 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 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 telecom operator—which are all faced with these questions—depend 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 versus 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.[36] At the same time, and to play the 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 their 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.

[36] “Ebola and big data: Waiting on hold” 2014.

6.4 Ethical principles

6.4.1 Framework: The Menlo Report

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. Arguably corporations may not be research institutions when using CDRs, and ethical principles were primarily developed with biomedical and behavioral sciences in mind. However, the practice of Big Data analytics, and specifically CDRs, closely resembles research cycles and processes, and the insight sought are relevant to behavioral science. This dissertation proposes that existing ethical principles provide a valuable and possibly sufficient framework to guide the emerging field of CDR analytics. The research ethic frame is the most appropriate to highlight the most consequential and problematic issues as well as opportunities raised by CDRs 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 the Belmont Report. A more recent initiative by the US Department of Homeland Security, Science and Technology, Cyber Security Division revised and adapted 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[37]

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
   a) Do not harm;
   b) Maximize probable benefits and minimize probable harms;
   c) Systematically assess both risk of harm and benefit.
2. Respect for Persons:
   a) Participation as a research subject is voluntary, and follows from informed consent;
   b) Treat individuals as autonomous agents and respect their right to determine their own best interests;
   c) Respect individuals who are not targets of research yet are impacted;
   d) Individuals with diminished autonomy, who are incapable of deciding for themselves, are entitled to protection.
3. Justice:

[37] Dittrich and Kenneally 2012.

   a) 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;
   b) Selection of subjects should be fair, and burdens should be allocated equitably across impacted subjects.

4. Respect for Law and Public Interest:
   a) Engage in legal due diligence;
   b) Be transparent in methods and results;
   c) Be accountable for actions.

6.4.2 Unpacking ethical principles for CDR analytics

These principles play out in a variety of ways in the case of CDR analytics:

6.4.2.1 Beneficence

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.”[38] 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.

6.4.2.2 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, it is often with little to no choice, since not consenting would limit their access to the technology. Furthermore, the choice given to consumers is typically to either dissent or fully consent regardless of what use

[38] *Stanford Encyclopedia of Philosophy* n.d.

of the data can be done several years later, or by a third party should it be accessed. 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 dependent 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 used despite strong opposition on moral and ethical grounds. An even more advanced model proposes that individuals would permanently 'carry' a set of permissions that they grant to algorithm seeking to use their data—no matter what data—enabling them to modify access and permissions at any time.

6.4.3 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 analysis), even when unequal access to technology exists. Yet, whose data is considered in CDRs analysis is inherently affected by unequal access and use of mobile phones, creating inherent biases and violating the principle of justice.

This creates yet another tension in CDR analysis: 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 countries: 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 hinder 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).

As discussed in Chapter 5, biases may be unproblematic as long as they are well understood and corrected for, which must become a research priority. Furthermore, it is likely that as cell phone penetration and patterns of use change, there will be a need to constantly adapt methods and develop algorithms 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 discriminations, 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—credit card transactions, social media data, etc, and in all likelihood CDRs 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. Group privacy will be come an increasingly prominent topic of research and debate.

6.4.4 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, inconsistencies of the legal or regulatory environment guiding 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 use of CDRs, what is the recourse for that individual?).

Telecom operators 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. Telecom operators may be confronted to local, legal requirements that may be at odds with international law and could potentially be held liable if their data were used in mass atrocities, something not entirely impossible. In repressive environments, telecom operators should consider their first priority to protect the sources of information (their customers) and place sensitive data beyond the reach of authorities, even though this may be against their financial and commercial interests. At the same time, telecom operators which have access to potentially life–saving information may be morally, if not legally, required to make that information available.

6.5 Way forward

6.5.1 General considerations

Having used core ethical principles to frame the key challenges emerging in the rapidly growing practice of CDRs analysis, 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 operators contribute—as socio-economic agents—to enhancing the welfare of societies where they operate. Telecom operators should not mimic the most negative aspects of extractive industries were valuable resources are exported with no or little 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 telecom operators. 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.

Telecom operators 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 force) access to data for public good. 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 telecom operators to provide data must be coordinated. Data requirements must also be better defined to avoid demands that seeks 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. Information contained in 'old' CDRs are interesting for research, and their result do not depend too much on the timing of extraction of the data (say within the last 2 or 3 years). Some applications may need 'fresh' data, even real time, but these are technically and ethically more difficult. So for these data, we need to be clear about the benefits we expect for individuals and society 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 analysis 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 re-highlight that CDRs alone offer only limited insight, and that their richness is 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; for example, tracking poverty or socioeconomic levels using CDRs requires having poverty or socioeconomic data with which to start, and a CDR analysis alone would be very sensitive to sample bias. ‘Historical data,’ even aggregated data, are extremely valuable or perhaps in a way even more than ‘real-time’ raw data because they not just allow but indeed limit/compel ‘us’ to focus on building methods and tools under greater constraints, above and beyond (*i.e.*, before) attempting to do nowcasting of current populations/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 setting are absolutely essential ingredients and objectives for the expansion of CDR analysis. 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 ICT and data revolutions, as during the Open Government Partnership Summit sessions on whistle blowing, privacy, and safeguarding civic space—especially in light of the Snowden case—or at the Technology Salon on Participatory Mapping.[39] 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.

It remains to be seen what governance and technical arrangement 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.

[39] “Ethics and risk in open development” 2013.

A number of pointers can be discussed describing some of the minimum requirements for such arrangements and partnerships between mobile phone companies, the research community, governments, and other civil society actors.

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 (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 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 is lost 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 use, seeking to answer questions like 'is real time needed?' 'If not, what type of past data?' '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 telecom operators maintain their contractual relation with their customers, the respect of their privacy, and control critical information that may help direct competitors (local marketshare, 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 "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 partnership and data processing accreditation are likely to be necessary. A centralized system (real institutional CDR sharing and clearinghouse 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 analysis (*e.g.*, the use of background maps), some elements of data and indicators sharing will be effective as well.

6.5.2 Data literacy or literacy in the age of data

In recent months, the notion of 'data literacy' has received significant attention and been framed as an enabler of the "Data Revolution." But what data literacy is, and subsequently how and for what purposes it ought to be promoted, has remained rather blurry.

Claude Lévi-Strauss wrote in his 1955 masterpiece "Tristes Tropiques" that literacy was a *"strange thing"*;[40] so is data literacy, which should not be promoted without specifying what is meant and expected from it. Co-authors and I have put forth two main counter proposals with strong historical and political undertones. One is to talk about—or at least think in terms of— 'literacy in the age of data' as a much more useful concept, defined as "the desire and ability to constructively engage in society through and about data."[41] The second is to promote it via and for social inclusion.

To date, the appeal and success of 'data literacy' in the public discourse and psyche reflect and fuel a relatively narrow conceptualization of the Data Revolution itself rooted in a simplistic diagnostic of the world's problems and what data can do about them. For too many champions of data literacy, the main solutions focus on technical capacity gaps that need to be filled and fixed, so that more people can become better at analyzing data. This is of course partly accurate: to varying degrees, the vast majority of the world's population, including those crafting and implementing public policies or other public service functions, are ill-equipped to deal with and take advantage of the new world of data. There are massive and pressing needs to strengthen technical capacities for the positive transformative power of data to be unleashed in sectors and regions where lack of relevant and timely information has been a real impediment to social progress. Priority constituencies will include national statistical officers, elected officials, journalists, communities, and individuals. Building those key skills will require significant investment over many years if they are to remain relevant in the age of data.

However, the current data literacy narrative overlooks many more complex and controversial questions. History has repeatedly shown how technology could entrench rather than challenge power structures that perpetuate detrimental outcomes—for instance inequity, poverty, corruption, and environmental degradation. This is obviously because technology is often invented and used first and primarily, when not exclusively, by those in power. The promotion and diffusion of technology to the masses is not necessarily at odds with this model, as Lévi-Strauss argued about literacy promotion. This is old news, but history has a tendency to repeat itself as its lessons are forgotten.

[40] Lévi-Strauss 1955.

[41] Data-Pop Alliance 2015a.

History sheds light on how defining and promoting literacy—who was literate and who was not—has been often entrenched with the constructs and perpetuation of power structures within societies—at odds with the notion of literacy as a necessarily empowering and enlightenment force. There is a risk that the same processes may play out in the age of data, at a speed and scope commensurable with those of the spread of data as a social phenomenon.

I with co-authors define data literacy as “the desire and ability to constructively engage in society through and about data.” Five observations emerge from this definition:

1. “Desire and ability” highlights technology as a magnifier of human intent and capacity.

2. “Ability” underlines literacy as a continuum, moving away from the dichotomy of literate and illiterate.

3. “Data” is understood broadly as “individual facts, statistics, or items of information.”

4. “Constructively engage in society” suggests an active purpose driving the desire and ability”.

5. And “through or about data” offers the possibility for individuals to engage as data literate individuals without being able to conduct advanced analytics.

This definition—as well as the nature of data itself—encompasses elements and principles from each of these sub-kinds of literacy (such as media, statistical, scientific computational, information and digital literacies), moving away from medium-centred definitions of literacy towards a more encompassing one.

Data literacy is not primarily about enabling individuals to master a particular skill or to become proficient in a certain technology platform. Rather it is about equipping individuals to understand the underlying principles and challenges of data. This understanding will in turn empower people to comprehend, interpret, and use the data they encounter, and even to produce and analyze their own data. This can only be achieved by considering data literacy as a means toward a necessary reinvention of community engagement and empowerment—towards what I term data inclusion.

6.5.3 Open algorithms as a new paradigm and priority for research and development?

Algorithms have bad press. They seem to be everywhere yet often hidden and generally poorly understood by the general public. They are referred to as ‘black boxes’ concealing sophisticated and insidious mechanisms that crunch citizen-consumers’ data to make predictions that turn into prescriptions, and lock these subjects into their condition. Some argue they widen inequality and may threaten democracy.

There is of course partial truths and needed caution about the risks posed by a growing reliance on algorithms in various aspects and activities of our lives, particularly those created ‘on our behalf’ by corporations and governments.

However, the rise of algorithms and other technological developments may provide a historical opportunity, both a practical way and moral obligation, to re–engineer current power structures and decision-making processes within data-infused societies in positive ways.

What I have come to term 'algovernance' is both old and new—societies are governed by codified and standardized rules and mechanisms that use and produce information towards prediction and prescription. In today's digital world, we think and talk about different kinds of algorithms; those of the Big Data revolution. They feed on different inputs, seem less 'human,' and have more immediate and possibly more powerful effects. At the same time, our contemporary world is highly unequal, unfair, and unstable. What role can algorithms play to make this world a better place? Can algorithms of the Big Data era and the opportunities, risks, and questions they raise, be leveraged as forces of positive disruption? In the age of data, could algorithms be less fallible due to human error or bias and can be designed in a more robust manner to improve the human condition. Could they offer opportunities to question outcomes? I believe the answers to all these questions is yes. And I believe that open algorithms could constitute "a new paradigm for using private data for social good."

Public goods algorithms, accountability, and transparency are fundamental in government and corporate use of such powerful decision support tools, both in validating their utility toward the public interest as well as redressing corrupt or unjust harms they may generate. Further, even when or as these algorithms may be open and transparent, access and use of these data remain legitimately constrained by ethical, political, legal, and commercial considerations, especially when the notion of 'anonymized' data is being tested by recent research.

The various initiatives described above such as the Orange D4D challenges or 'hackathons' organized by other telecom operators such as Telefónica or Telecom Italia have allowed researchers to work with mobile data in a tightly controlled environment; in other cases, academic teams have had access to more granular data through specific agreements—which often rely on personal relationships. But these modalities do not allow for the scalability, systemization, and involvement of a wide range of stakeholders that are necessary for harnessing the full social potential of these data and the algorithms that swift through them to solve complex human problems. Further, they are locked in the privacy-utility dilemma.

There have been attempts at finding technical and even more so governance 'solutions,' including through discussions and initiatives about data standards and general calls for the "responsible" use of data, but deep and hard questions about whether and how data—as well as the algorithms that swift through them to solve complex problems—may spur development and democracy are yet to be answered or even asked correctly.

Algorithmic transparency is a necessary but insufficient condition to ensure the democratic and socially desirable deployment and diffusion of such tools; attention also ought to be paid to the rules and modalities presiding over access and control of their raw material—the data—as well as the use and learning from decisions based on their results. The crux of the issue is that most people do not want their personal data to be out in the open, even as they may contain insights that could help curb poverty or conflict, which most also deem desirable.

Against this background, I along with colleagues from Orange, MIT, the World Economic

Forum, and Imperial College have initiated the Open Algorithms (OPAL) project, a multi-partner initiative that aims to open—without exposing—data collected and stored by private companies by "sending the code to the data" rather than the other way around, to enhance the design and monitoring of development policies and programs, accountability of government action, and citizen engagement.[42] OPAL's core, which is being developed with funding from the French Development Agency and the World Bank, will consist of an open platform allowing open algorithms to run on the servers of partner companies, behind their firewalls, to extract key development indicators and operational data of relevance for a wide range of potential users. Request for predetermined indicators by third parties—*e.g.*, mobility matrices, poverty maps, population densities—will be sent to them via the platform; pre-developed algorithms will run on the data in a multiple privacy-preserving manner, and results will be made available via an API.

The platform will also be used as a hook and agora for civic engagement of a broad range of social constituents—academic institutions, private sector companies, official institutions, non-governmental and civil society organizations. Technology alone will not revive democracy but technical interventions can be used to shape and spur value systems of trust and collaboration across ecosystems on and via data and algorithms. Focusing on opening other parts of the process beyond just the data can revive democratic principles of deliberative decision-making by leveraging new data, new tools and, new actors for more accountable and agile policies and decisions.

In other words, the vision of the OPAL project is to design, develop, and deploy an open algorithms platform and connected activities to:

1. Improve long-term decision-making and support tactical operations
2. Address multiple privacy and security challenges
3. Strengthen accountability and accuracy for the ethical use of data
4. Catalyze a vibrant and inclusive local data ecosystem
5. Leverage the value of open-source, agile, affordable and scalable technologies

6.6 Conclusion

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 CDRs analysis, or derived indicators, a relatively standard set of tools for researchers by the end of the decade.

[42] Roca and Letouzé 2016.

A broad array of societal opportunities of Big Data in emerging countries are real and there are ways to develop the tools, process and policies to covers both the society 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.

However, if this Data Revolution is to bring about positive change, it has to be an evolution towards social inclusion in the age of data – towards data inclusion. If a 'business-as-usual' framing for the Data Revolution continues unabated, efforts toward greater data literacy may reinforce existing power dynamics that promote social exclusion. This transitional period is the opportune time to create a path towards empowerment. In particular, data literacy focused on building data inclusion and awareness offers a doorway to fostering and managing data-driven discussions and decisions for all people, backed on principles and practices such as open algorithms.

Concluding remarks from Rwanda

In August 2016, while in Rwanda during the final stretch and sprint to finish and file this dissertation, I drove back from Lake Kivu, one of the African Great Lakes located between the western part of the country and the Democratic Republic of the Congo, to Kigali, the capital. I was there for both professional and personal reasons, spending the first few days of my stay with my friend Valéry and a friend of his, a fellow Rwandan philosopher. The 3-hour drive was rather typical of many trips in Sub-Saharan Africa, with dusty roads and their occasional potholes, packed minibuses driving at high speed, great landscapes of fields and hills, and busy villages and towns, with stalls and stores selling SIM cards from the telecom operators MTN and Tigo every few hundred meters.

Perhaps even more than in other places, I was struck during the entire trip by the number of people walking alongside the road, for what seemed like very long distances on often steep segments, carrying large containers, pieces of wood, etc. One obvious reason is that despite major improvements and achievements in the past 2 decades, Rwanda remains one of the poorest countries on the planet, with a Human Development Index slightly above 0.48 that places it in the 163rd position in the world for this indicator, just below Haiti. With very high gas prices and massive import taxes on cars, owning a motorbike or a car is unaffordable to the immense majority of its population. Another reason is the country's population density, close to 1,000 people per square mile, the 5th highest in the world for countries with areas greater than 10,000 square miles.[43]

In the global public psyche, but also in its daily life for the foreseeable future, Rwanda is indissociable from the genocide that took place over 3 months between April and June 1994, when about 20% of its population, most of them Tutsi, and most of the Tutsi, were slaughtered by Hutu militias and extremists, but also people next door, typically with machetes. Also next door were UN 'Blue Helmets,' who were neither given the mandate nor the means by their Headquarters to prevent as many as 1,000,000 people from being killed in 100 days in a country the size of Maryland. The genocide ended when the forces of the Rwandan Patriotic Front led by Paul Kagame took control of the country. At that point there were only a dozen thousands of Tutsi alive. Since then the conflict has echoed in all countries in the region. Several governments have since offered apologies for their inaction (not that of France). Back in 1994, I was in a French "preparatory class" in Paris; I remember hearing

[43] That is excluding small countries and territories like Macao, Monaco, Singapore, etc. These statistics and more can be found or computed at: `http://hdr.undp.org/en/countries/profiles/RWA`.

about massacres being committed in Rwanda, but it seemed both distant and common. Along the road and across the country, memorials commemorate the genocide—for which a word had to be invented in the national language, kinyarwanda: "jenocide."

Half-way through the trip we made a halt to take a few pictures of the valley, talked to a group of boys who literally came down from the cliff above us, the youngest being about the age of my daughters, wearing a torn t-shirt and no shoes, with a steady and serious look; we then had lunch at a local Franciscan parish in Gitarama at the invitation of the priest—rice, beans, chicken, and salad from their garden, followed by a coffee that I knew would keep me awake late into the night. There I found myself chatting with a short slender Scottish man in his 60s, who was volunteering for a year to teach carpentry to about 50 teenagers. He had lived, traveled, and worked in many countries—El Salvador, Colombia, England, Spain, the U.S.—since leaving Glasgow to escape violence and unemployment in the mid-1970s. His face and hands showed a life of hard work and tough conditions; his eyes glittered curiosity and compassion. He said he loved teaching the kids, loved the people in the community, and even enjoyed the local dark beer, called "Turbo King." At 42, he started university and got a degree in sociology, focusing on local development. He wanted to start a doctorate, but it did not work out. He asked me what I did for a living; I told him I was finishing my PhD in Demography and running an organization focusing on "Big Data's application and implications for research and development." He had never heard the term; I turned to the people in the room with their cell-phones, most of them basic smart phones, and explained that these acted as sensors of almost everything we did as humans.

I talked about how it could help map mobility movements, predict poverty and crime; I described some of my research and training projects in Colombia, Côte d'Ivoire, Senegal, and Rwanda. We also mentioned Snowden and the NSA, and spoke about how Paul Kagame most certainly had developed advanced data analytic capacities to monitor its few opponents since becoming Rwanda's President in 2000. We shared our disappointment with the Brexit (although he looked forward to a new referendum on Scottish independence from the UK that he thought was now inevitable), our astonishment about the state of the U.S. and European political landscapes, our dismay at the recent waves of terrorist attacks around the globe, and our disapproval of lies and fear-mongering. We exchanged our email addresses and he gave me the contact of a close friend of his, who writes for various English and Irish newspapers on the peace process in Colombia.

We then left and arrived in Kigali where I had to wrap up my dissertation. About an hour into working on the document in LaTex online, a power outage shut down the Internet connection. It took about 30 minutes to get it back up, during which I feared the Gods of technology would once again fail me, and thought it would have been safer to cancel my trip and finish from Brooklyn. Once the electricity and connection were running again, I went back to work. A couple more outages occurred in the following hours.

This is our world, or at least this is in many ways the world in which I live and work. It is a big, complex, and interconnected world, with sharp inequalities, still widespread poverty, lots of cowardly politicians and conservative bureaucrats, too rigid institutional systems, mostly bad governments, cell-phones almost everywhere, wonderful dedicated people, Internet connection feeling as vital as running water, and data flowing around. And I find myself wondering, every single day: What can I do about it? What can data do about it? How can I play a positive role in improving the prospects of what I have called the fist "data generation" of humans?[44], the one my daughters are fully part of? I am not naive enough to think that data 'alone' can fix the world. "Information is power," we are told, but many atrocities or simply bad decisions happen with full information. The notion that the world would be a much better place to live in if only policymakers had better data is not a very serious one. But I do believe or want to believe that we are living in or entering a rather extraordinary era, where many of the old power structures and systems that preside over such bad outcomes can be challenged and changed to make way for new forms of societal governance that, somewhat emphatically, revive democratic ideals and principles, and in which, data can play an significant role.

How?

In this dissertation, I attempted to cover some options and requirements that I think need to be tackled. One is enhancing our collective knowledge about observable outcomes and their underlying causal processes; another is to challenge common wisdom or straight lies on the basis of facts—to instill a data culture where it will become harder to make things up without being called on it, with consequences—be they legal, reputational, or electoral. As I was quoted saying in 2014: "it is perfectly possible to lie with data [...], but the growing expectation that policies be based on figures means that it is getting harder to tell the truth without them."[45]

A key requirement is to build data literacy conceptualized as literacy in the age of data; going far beyond, above, and below the—important—provision of technical skills. A literate citizenry will probably want to seize data as a lever of power; chief of which by demanding and I believe getting more control over their data,[46] in ways that respect their privacy and security. I believe that a project like OPAL has the potential to trigger some of these changes. Significant efforts and resources will be needed to build capacities, systems, partnerships, and trust, within and between all actors of the Big Data communities, through investments in joint research, training and education, dialogue, and advocacy, and so forth. As an example of how I think I can help, in a few days, I will be meeting with faculty members of the University of Rwanda in Kigali to develop joint research and training projects; from there, I will go to Nairobi to do the same.

For demographers, Big Data should not be seen as a threat nor a fad; already many have embraced and engaged with Big Data. My sense is that the future should and will be one where boundaries between disciplines, perspectives, and skills blur to make them part of a community of scientists interested in understanding and solving human problems. Demographers have a lot to contribute to the future of Big Data—bringing their toolkits and methods, eclectic interests ranging from formal to social via historical and economic

[44] See notably: MasterCard Center for Inclusive Growth 2015.

[45] "Off the map" 2014.

[46] Pentland 2016

demography, and their concerns for sound statistical and ethical principles and standards.

Despite the tragedies that make the news every day, I am overall confident about the prospects of the next generations and hope to make a small contribution through my current and future work, starting this fall as a post-doctoral Visiting Scholar at MIT Media Lab while continuing to run Data-Pop Alliance.

Biblography

A New Global Partnership: Eradicate Poverty and Transform Economies through Sustainable Development. High-Level Panel of Eminent Persons on the Post-2015 Development Agenda, May 2015. URL: `http://www.post2015hlp.org/the-report/`.

A World That Counts: Mobilising The Data Revolution for Sustainable Development. United Nations Secretary-General's Independent Expert Advisory Group on a Data Revolution for Sustainable Development (IEAG), Nov. 2014. URL: `http://www.undatarevolution.org/report/`.

Agence des Telecommunications de Côte d'Ivoire (ATCI). *Abonnés téléphone mobile.* [Online]. June 2012. URL: `http://www.atci.ci/index.php/Service-mobile/abonnes-service-mobile.html`.

Agence France Presse. "Violence in western Ivory Coast kills six: official." In: *Modern Ghana* (Dec. 2011). URL: `http://www.modernghana.com/news/367663/violence-in-western-ivory-coast-kills-six-official.html`.

— "Police using 'predictive analytics' to prevent crimes before they happen." In: *Raw Story* (July 2012). URL: `http://www.rawstory.com/2012/07/police-using-predictive-analytics-to-prevents-crimes-before-they-happen/`.

Agence France Presse (AFP). "Four dead in southern Ivory Coast clashes." In: *Modern Ghana* (Dec. 2011). URL: `http://www.modernghana.com/news/369010/0/four-dead-in-southern-ivory-coast-clashes.html`.

Agence Nationale de la Statistique et de la Démographie du Sénégal (ANSD). *Recensement Général de la Population et de l'Habitat, de l'Agriculture et de l'Elevage.* [Online]. 2013.

Aiden, Erez and J.-B. Michel. *Uncharted: Big Data as a Lens on Human Culture.* Ed. by Riverhead. 2013. URL: `https://www.amazon.es/Uncharted-Data-Lens-Human-Culture/dp/1594487456`.

Alkire, Sabina and Maria Emma Santos. *Measuring Acute Poverty in the Developing World: Robustness and Scope of the Multidimensional Poverty Index.* OPHI Working Paper 59. Oxford Poverty and Human Development Initiative, University of Oxford, Mar. 2013. URL: `http://www.ophi.org.uk/wp-content/uploads/ophi-wp-59.pdf`.

Anderson, Chris. *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. Online. Dec. 2008. URL: `http://www.wired.com/2008/06/pb-theory/`.

Armed Conflict Location & Event Data Project. *ACLED Database for Côte d'Ivoire*. [Online]. 2012. URL: `http://www.acleddata.com/data/africa/`.

Bah, Abu Bakarr. "Democracy and civil war: Citizenship and peacemaking in Côte d'Ivoire." In: *African Affairs* 109.437 (Oct. 2010), pp. 597–615. ISSN: 0001-9909. DOI: `10.1093/afraf/adq046`. URL: `http://afraf.oxfordjournals.org/content/109/437/597`.

Bailey, Ronald. "Stopping Crime Before It Starts." In: *Reason.com* (July 2012). URL: `http://reason.com/archives/2012/07/10/predictive-policing-criminals-crime`.

Ball, Patrick, Jeff Klingner, and Kristian Lum. *Beneblog: Technology Meets Society: Crowdsourced data is not a substitute for real statistics*. [Online]. Mar. 2011. URL: `http://benetech.blogspot.com/2011/03/crowdsourced-data-is-not-substitute-for.html`.

BBVA Innovation Center. *Innova Challenge*. URL: `http://www.centrodeinnovacionbbva.com/en/innovachallenge/what-innova-challenge`.

Beam, Christopher. "Time Cops: Can police really predict crime before it happens?" In: *Slate* (Jan. 2011). URL: `http://www.slate.com/articles/news_and_politics/crime/2011/01/time_cops.html`.

Beck, Charlie and Colleen McCue. "Predictive Policing: What Can We Learn from Wal-Mart and Amazon about Fighting Crime in a Recession?" In: *The Police Chief* LXXVI.11 (Nov. 2009). URL: `http://www.rawstory.com/rs/2012/07/29/police-using-predictive-analytics-to-prevents-crimes-before-they-happen/`.

Becker, Richard, Chris Volinsky, Ramón Cáceres, Karrie Hanson, Sibren Isaacman, Ji Meng Loh, Margaret Martonosi, James Rowland, Simon Urbanek, and Alexander Varshavsky. "Human Mobility Characterization from Cellular Network Data." In: *Communications of the ACM* 56.1 (Jan. 2013), pp. 74–82. ISSN: 00010782. DOI: `10.1145/2398356.2398375`. URL: `http://dl.acm.org/citation.cfm?doid=2398356.2398375`.

Beulke, Dave. *Big Data Impacts Data Management: The 5 Vs of Big Data*. Online. Nov. 2011. URL: `http://davebeulke.com/big-data-impacts-data-management-the-five-vs-of-big-data/`.

Biau, Gérard. "Analysis of a Random Forests Model." In: *Journal of Machine Learning Research* 13 (apr 2012), pp. 1063–1095. ISSN: 1533-7928. URL: `http://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf`.

Blondel, Vincent D., Adeline Decuyper, and Gautier Krings. "A survey of results on mobile phone datasets analysis." In: *EPJ Data Science* (2015). DOI: 10.1140/epjds/s13688-015-0046-0. URL: http://epjds.epj.org/images/stories/news/2015/10.1140--epjds--s13688-015-0046-0.pdf.

Blondel, Vincent D., Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. "Data for Development: the D4D Challenge on Mobile Phone Data." In: *arXiv.org* (Jan. 2013). version 2. DOI: arXiv:1210.0137v2[cs.CY]. URL: http://arxiv.org/abs/1210.0137.

Blumenstock, J.E., G. Cadamuro, and R. On. "Predicting Poverty and Wealth from Mobile Phone Metadata." In: *Science* 250.6264 (Nov. 2015). URL: http://www.uvm.edu/~cdanfort/csc-reading-group/blumenstock-science-2015.pdf.

Blumenstock, Joshua E. "Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda." en. In: *Information Technology for Development* 18.2 (Apr. 2012), pp. 107–125. ISSN: 0268-1102. DOI: 10.1080/02681102.2011.643209. URL: http://www.jblumenstock.com/files/papers/jblumenstock_itd2012_wp.pdf.

Bock, Joseph G. and John Paul Lederach. *The Technology of Nonviolence: Social Media and Violence Prevention.* Cambridge, MA: MIT Press, July 2012. 288 pp. ISBN: 978-0-262-01762-6. URL: https://mitpress.mit.edu/books/technology-nonviolence.

Boggs, Sarah L. "Urban Crime Patterns." In: *American Sociological Review* 30.6 (1965), pp. 899–908. ISSN: 0003-1224. DOI: 10.2307/2090968.

Bogomolov, Andrey, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, and Alex Pentland. "Once Upon a Crime: Towards Crime Prediction from Demographics and Mobile Data." In: *Proceedings of the 16th International Conference on Multimodal Interaction.* New York: ACM, 2014, pp. 427–434. URL: http://arxiv.org/abs/1409.2983.

Bollier, David. *The Promise and Peril of Big Data.* The Aspen Institute, 2010. URL: https://www.emc.com/collateral/analyst-reports/10334-ar-promise-peril-of-big-data.pdf.

Borders, Devs Without. *<Br/eak>Poverty Hackathons.* URL: http://devswithoutborders.org/breakpoverty/.

boyd, danah and Kate Crawford. "Six Provocations for Big Data." In: *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society.* Sept. 2011. DOI: 10.2139/ssrn.1926431. URL: http://ssrn.com/abstract=1926431.

Braithwaite, John. *Crime, shame and reintegration.* United Kingdom: Cambridge University Press, 1989. ISBN: 978-0-521-35668-8.

Brantingham, Patricia L. and Paul J. Brantingham. "A Theoretical Model of Crime Hot Spot Generation." In: *Studies on Crime and Crime Prevention* 8.1 (1999), pp. 7–26. URL: https://www.ncjrs.gov/App/publications/abstract.aspx?ID=177811.

Breiman, Leo. "Bagging predictors." In: *Machine Learning* 24.2 (Aug. 1996), pp. 123–140. ISSN: 0885-6125. DOI: 10.1007/BF00058655.

— *Out-of-Bag Estimation.* University of California Statistics Department, 1996. URL: https://www.stat.berkeley.edu/~breiman/OOBestimation.pdf.

Bruckschen, Fabian, Timo Schmid, and Till Zbiranski. "Cookbook for a socio-demographic basket: Constructing key performance indicators with digital breadcrumbs." In: *Data for Development Challenge Senegal, Book of Abstracts: Scientific Papers.* MIT Media Lab, Cambridge, MA, 2014, pp. 122–131.

Burke, M., S. M. Hsiang, and E. Miguel. "Climate and Conflict." In: *Annual Review of Economics* (2015). DOI: 10.1146/annurev-economics-080614-115430.

Burrell, Jenna. *The Ethnographer's Complete Guide to Big Data: Answers (part 2 of 3).* June 2012. URL: http://ethnographymatters.net/blog/2012/06/11/the-ethnographers-complete-guide-to-big-data-part-ii-answers/.

Butler, Declan. "When Google got flu wrong." In: *Nature* 494 (Feb. 2013), pp. 155–156. URL: http://www.nature.com/news/when-google-got-flu-wrong-1.12413.

Caplan, Joel M. and Leslie W. Kennedy. *Risk Terrain Modeling Manual: Theoretical Framework and Technical Steps of Spatial Risk Assessment for Crime Analysis.* Newark, NJ: Rutgers Center on Public Security, 2010. 122 pp.

Caruana, Rich, Nikos Karampatziakis, and Ainur Yessenalina. "An Empirical Evaluation of Supervised Learning in High Dimensions." In: *Proceedings of the 25th International Conference on Machine Learning.* ICML '08. New York, NY, USA: ACM, 2008, pp. 96–103. ISBN: 978-1-60558-205-4. DOI: 10.1145/1390156.1390169.

Cederman, Lars-Erik, Kristian Skrede Gleditsch, and Simon Hug. "Elections and Ethnic Civil War." In: *Comparative Political Studies* (Sept. 2012). ISSN: 0010-4140. DOI: 10.1177/0010414012453697. URL: http://cps.sagepub.com/content/early/2012/09/05/0010414012453697.

Center for International Earth Science Information Network (CIESIN), Columbia University, United Nations Food and Agriculture Programme (FAO), and Centro Internacional de Agricultura Tropical (CIAT). *Gridded Population of the World: Future Estimates (GPWFE).* Socioeconomic Data and Applications Center (SEDAC), Columbia University, 2005. URL: http://sedac.ciesin.columbia.edu/gpw.

Chen, Xi and William D. Nordhaus. "Using luminosity data as a proxy for economic statistics." In: *Proceedings of the National Academy of Sciences* 108.21 (May 2011), pp. 8589–8594. ISSN: 0027-8424. DOI: 10.1073/pnas.1017031108. URL: http://www.pnas.org/content/108/21/8589.

Cisco. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015–2020 White Pape*. Online. Feb. 2016. URL: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html.

Coale, Ansley J. and T. James Trussell. "Model fertility schedules: variations in the age structure of childbearing in human populations." In: *Population index* 40.2 (Apr. 1974), pp. 185–258. URL: https://www.jstor.org/stable/2733910.

Collier, Paul, V. L. Elliott, Håvard Hegre, Anke Hoeffler, Marta Reynal-Querol, and Nicholas Sambanis. *Breaking the Conflict Trap: Civil War and Development Policy*. World Bank, 2003. URL: https://openknowledge.worldbank.org/bitstream/handle/10986/13938/567930PUB0brea10Box353739B01PUBLIC1.pdf.

Collier, Paul and Anke Hoeffler. *Greed and Grievance in Civil War*. World Bank Policy Research, May 2000. DOI: http://dx.doi.org/10.1596/1813-9450-2355. URL: http://elibrary.worldbank.org/doi/abs/10.1596/1813-9450-2355.

Council of Europe. "Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data." In: *ETS no. 108* (Jan. 1981). URL: http://conventions.coe.int/Treaty/en/Treaties/Html/108.htm.

— "Chart of signatures and ratifications of Treaty 108: Status as of 06/08/2016." In: *ETS no. 108* (Aug. 2016). URL: http://conventions.coe.int/Treaty/Commun/ChercheSig.asp?NT=108&CM=1&DF=&CL=ENG.

Crawford, Kate. "The Hidden Biases in Big Data." In: *Harvard Business Review* (Apr. 2013). URL: https://hbr.org/2013/04/the-hidden-biases-in-big-data.

Cukier, Kenneth. "The data revolution." In: *The Economist* (2013). URL: http://www.economist.com/blogs/prospero/2013/05/kenneth-cukier-big-data.

Cullen, Julie Berry and Steven D. Levitt. "Crime, Urban Flight, and the Consequences for Cities." In: *Review of Economics and Statistics* 81.2 (May 1999), pp. 159–169. ISSN: 0034-6535. DOI: 10.1162/003465399558030. URL: http://www.mitpressjournals.org/doi/abs/10.1162/003465399558030.

Data-Pop Alliance. "Beyond Data Literacy: Reinventing Community Engagement and Empowerment in the Age of Data." In: (Sept. 2015). URL: http://datapopalliance.org/wp-content/uploads/2015/10/BeyondDataLiteracy_DataPopAlliance_Sept30.pdf.

Data-Pop Alliance. “Big Data and Privacy: understanding the possibilities and pitfalls of the data revolution in Germany.” In: *Digitising Europe Initiative*. Nov. 2015. URL: `http://datapopalliance.org/wp-content/uploads/2016/05/VFI_DataPopAlliance_Berlin.pdf`.

— “Big Data for Climate Change and Disaster Resilience: Realising the Benefits for Developing Countries.” In: (2015). URL: `http://datapopalliance.org/wp-content/uploads/2015/11/Big-Data-for-Resilience-2015-Report.pdf`.

— ““Not using data is the moral equivalent of burning books”.” In: *Digitising Europe Initiative*. Brussels, Belgium, Jan. 2016. URL: `http://www.vodafone-institut.de/event/not-using-data-is-the-moral-equivalent-of-burning-books/`.

de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen, and Vincent D. Blondel. *Unique in the Crowd: The privacy bounds of human mobility*. Research rep. Scientific Reports, 2013. DOI: `10.1038/srep01376`. URL: `http://www.nature.com/articles/srep01376`.

de Montjoye, Yves-Alexandre, Laura Radaelli, Vivek Kumar Singh, and Alex ‘Sandy’ Pentland. *Unique in the shopping mall: On the reidentifiability of credit card metadata*. Science, 2015. URL: `http://science.sciencemag.org/content/347/6221/536.full`.

de Montjoye, Yves-Alexandre, Erez Shmueli, Samuel S. Wang, and Alex ‘Sandy’ Pentland. “openPDS: Protecting the Privacy of Metadata through SafeAnswers.” In: *PLOS ONE* 9.7 (July 2014), p. 98790. ISSN: 1932-6203. DOI: `10.1371/journal.pone.0098790`. URL: `http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0098790`.

de Montjoye, Yves-Alexandre, Zbigniew Smoreda, Romain Trinquart, Cezary Ziemlicki, and Vincent D. Blondel. “D4D-Senegal: The Second Mobile Phone Data for Development Challenge.” In: *arXiv.org* abs/1407.4885 (2014). DOI: `arXiv:1407.4885v1[cs.CY]`.

DeAngelis, Steve. *Big Data is about People and Behavior*. Online. Feb. 2013. URL: `http://enterpriseresilienceblog.typepad.com/enterprise_resilience_man/2013/02/big-data-is-about-people-and-behavior.html`.

Demographics and Health Surveys (DHS) Program. *Côte d’Ivoire: Standard DHS, 2011-12*. [Online]. June 2013. URL: `http://dhsprogram.com/what-we-do/survey/survey-display-311.cfm`.

Devarajan, Shanta. “Africa’s statistical tragedy.” In: *Africa Can End Poverty (World Bank Blog)* (Oct. 2011). URL: `http://blogs.worldbank.org/africacan/africa-s-statistical-tragedy`.

Deville, Pierre, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. "Dynamic population mapping using mobile phone data." In: *Proceedings of the National Academy of Sciences* 111.45 (2014), pp. 15888–15893. DOI: `10.1073/pnas.1408439111/-/DCSupplemental`. URL: `http://www.pnas.org/content/111/45/15888.abstract`.

Dittrich, D. and E. Kenneally. *The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research*. Department of Homeland Security (USA), Aug. 2012. URL: `http://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/`.

Dragland, Åse. *Big Data, for better or worse: 90% of world's data generated over last two years*. Online. The post is reprinted from materials provided by SINTEF. The original item was written by Åse Dragland. May 2013. URL: `https://www.sciencedaily.com/releases/2013/05/130522085217.htm`.

Dube, Oeindrila and Juan Vargas. "Are All Resources Cursed? Coffee, Oil and Armed Conflict in Colombia." In: *Weatherhead Center for International Affairs, Harvard University, Cambridge, MA* (Jan. 2007). Working Paper Series, No. 07-01. URL: `http://projects.iq.harvard.edu/files/wcfia/files/2007_1_vargas.pdf`.

— "Commodity Price Shocks and Civil Conflict: Evidence from Colombia." In: *Mimeo, Harvard University* (2008). URL: `http://isites.harvard.edu/fs/docs/icb.topic470466.files/Commodity%20Price%20Shocks%20and%20Civil%20Conflict%20_%20Evidence%20from%20Colombia_Oeindrila%20Dube_October%2008.pdf`.

— "Commodity Price Shocks and Civil Conflict: Evidence from Colombia." In: *The Review of Economic Studies* 80.4 (2013), pp. 1384–1421. URL: `http://restud.oxfordjournals.org/content/early/2013/02/15/restud.rdt009`.

Eagle, Nathan, Michael Macy, and Robert Claxton. "Network diversity and economic development." In: *Science* 328.5981 (2010), pp. 1029–1031. DOI: `10.1126/science.1186605`. URL: `http://science.sciencemag.org/content/328/5981/1029`.

Easterly, William and Ross Levine. "Africa's Growth Tragedy: Policies and Ethnic Divisions." In: *The Quarterly Journal of Economics* 112.4 (Nov. 1997), pp. 1203–1250. ISSN: 0033-5533. DOI: `10.1162/003355300555466`. URL: `http://qje.oxfordjournals.org/cgi/doi/10.1162/003355300555466`.

"Ebola and big data: Waiting on hold." In: *The Economist* (Oct. 2014). From the print edition. ISSN: 0013-0613. URL: `http://www.economist.com/news/science-and-technology/21627557-mobile-phone-records-would-help-combat-ebola-epidemic-getting-look`.

Eck, John E., Spencer Chainey, James G. Cameron, Michael Leitner, and Ronald E. Wilson. *Mapping Crime: Understanding Hotspots.* Washington, D.C.: National Institute of Justice, Aug. 2005. URL: `https://www.ncjrs.gov/pdffiles1/nij/209393.pdf`.

Ehrlich, Isaac. "On the relation between education and crime." In: *Education, Income and Human Behavior* (1975). Ed. by F. Thomas Juster. ISBN : 0-07-010068-3, pp. 313–338. URL: `http://www.nber.org/chapters/c3702.pdf`.

Einav, Liran and Jonathan D. Levin. "The Data Revolution and Economic Analysis." In: *NBER Working Paper Series* (2013). Working Paper 19035. URL: `http://www.nber.org/papers/w19035.pdf`.

Ellis, Lee, Kevin M. Beaver, and John Paul Wright. *Handbook of Crime Correlates.* OCLC: 262892601. Amsterdam; Boston: Elsevier/Academic Press, 2009. ISBN: 978-0-12-373612-3.

"Ethics and risk in open development." In: *Open Knowledge International Blog* (2013). URL: `http://blog.okfn.org/2013/11/05/ethics-and-risk-in-open-development/`.

European Court of Human Rights. *European Convention on Human RIghts.* European Council, 1953. URL: `http://www.echr.coe.int/Documents/Convention_ENG.pdf`.

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." In: *Official Journal of the European Union* L 281 (1995). article 2 (a) and article 7 (a)-(f). URL: `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML`.

— "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." In: *Official Journal of the European Union* L 105 (2006), pp. 54–63. URL: `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:105:0054:0063:EN:PDF`.

Farrington, David P. and Rolf Loeber. "Some benefits of dichotomization in psychiatric and criminological research." In: *Criminal Behaviour and Mental Health* 10.2 (June 2000), pp. 100–122. ISSN: 1471-2857. DOI: `10.1002/cbm.349`. URL: `http://onlinelibrary.wiley.com/doi/10.1002/cbm.349/abstract`.

Felson, Marcus and R. V. G Clarke. *Opportunity makes the thief: practical theory for crime prevention.* OCLC: 41020072. London: Home Office, Policing, Reducing Crime Unit, Research, Development, and Statistics Directorate, 1998. ISBN: 978-1-84082-159-8.

Felson, Marcus and Erika Poulsen. “Simple indicators of crime by time of day.” In: *International Journal of Forecasting* 19.4 (Oct. 2003), pp. 595–601. ISSN: 0169-2070. DOI: 10.1016/S0169-2070(03)00093-1.

Fengler, Wolfgang. “Big data and development: “The second half of the chess board”.” In: *Africa Can End Poverty (World Bank Blog)* (Feb. 2013). URL: http://blogs.worldbank.org/africacan/big-data-and-development-the-second-half-of-the-chess-board.

Ferrara, Emilio, Pasquale De Meo, Salvatore Catanese, and Giacomo Fiumara. “Detecting criminal organizations in mobile phone networks.” In: *Expert Systems with Applications* 41.13 (Oct. 2014), pp. 5733–5750. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2014.03.024. arXiv: 1404.1295. URL: http://arxiv.org/abs/1404.1295.

Few, Stephen. “Big Data, Big Ruse.” In: *Visual Business Intelligence Newsletter* (2012). URL: http://www.perceptualedge.com/articles/visual_business_intelligence/big_data_big_ruse.pdf?goback=.gde_1814785_member_170461368.

Flowminder. *Nepal Population Estimates as of May 1, 2015.* [Online]. May 2015. URL: https://www.humanitarianresponse.info/en/operations/nepal/document/flowminder-nepal-population-estimates-and-movements-01-may-2015.

Frank, Christopher. *Improving Decision Making in the World of Big Data.* Mar. 2012. URL: http://www.forbes.com/sites/christopherfrank/2012/03/25/improving-decision-making-in-the-world-of-big-data/#60b67a7d4b4d.

GADM database of Global Administrative Areas. *Version 2.8.* [Online]. Nov. 2015. URL: http://www.gadm.org/country.

GeoHive. *Côte d’Ivoire: administrative units, extended.* [Online]. May 2014. URL: http://www.geohive.com/cntry/coteivoire_ext.aspx.

Giovannini, Enrico. “Statistics 2.0: The Next Level.” In: *10th National Conference of Statistics.* Rome, Italy, Dec. 2010. URL: http://en.istat.it/istat/eventi/2010/10_conferenza_statistica/Relazione_pres_10conf.pdf.

Giugale, Marcelo. “Fix Africa’s Statistics.” In: *The World Post* (Dec. 2012). URL: http://www.huffingtonpost.com/marcelo-giugale/fix-africas-statistics_b_2324936.html.

Grill, Andrew. “Smart Steps: Using Big data for social good.” In: *Telefonica Dynamic Insights* (Aug. 2013). [Online]. URL: http://dynamicinsights.telefonica.com/2013/08/29/using-big-data-for-social-good/.

Grimes, William. *Big Data Becomes a Mirror ‘Uncharted,’ by Erez Aiden and Jean-Baptiste Michel.* Online. Dec. 2013. URL:

http://www.nytimes.com/2013/12/25/books/uncharted-by-erez-aiden-and-jean-baptiste-michel.html?_r=0.

Gutierrez, Thoralf, Gautier Krings, and Vincent D. Blondel. "Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets." In: *arXiv.org* (Sept. 2013). DOI: arXiv:1309.4496v1[cs.CY]. URL: https://arxiv.org/abs/1309.4496v1.

Guyon, Isabelle and André Elisseeff. "An Introduction to Variable and Feature Selection." In: *Journal of Machine Learning Research* 3 (Mar. 2003), pp. 1157–1182. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944968.

Hellerstein, Joseph. *The Commoditization of Massive Data Analysis*. Online. Nov. 2008. URL: http://radar.oreilly.com/2008/11/the-commoditization-of-massive.html.

Henderson, J. Vernon, Adam Storeygard, and David N. Weil. "Measuring Economic Growth from Outer Space." In: *National Bureau of Economic Research* (July 2009). NBER Working Paper No. 15199. DOI: 10.3386/w15199. URL: http://www.nber.org/papers/w15199.

Hilbert, Martin and Priscila Lopez. "The World's Technological Capacity to Store, Communicate, and Compute Information." In: *Science* 332.60 (2011). DOI: 10.1126/science.1200970. URL: http://www.uvm.edu/~pdodds/files/papers/others/2011/hilbert2011a.pdf.

Horrigan, Michael W. *Big Data: A Perspective from the BLS*. Online. Jan. 2013. URL: http://magazine.amstat.org/blog/2013/01/01/sci-policy-jan2013/.

Howell, Valerie. "Rights groups take UK government to European Human Rights Court over mass surveillance." In: *jurist.org* (2015). URL: http://jurist.org/paperchase/2015/04/rights-groups-take-uk-government-to-european-court-of-human-rights-over-mass-surveillance.php.

Hsiang, S., M. Burke, and E. Miguel. "Quantifying the Influence of Climate on Human Conflict." In: *Science* 341 (2013). DOI: 10.1126/science.1235367. URL: http://users.clas.ufl.edu/prwaylen/geo3280articles/Climate%20Change%20and%20Conflict.pdf.

IBM. "Memphis PD: Keeping ahead of criminals by finding the "hot spots"." In: *Smarter Planet Leadership Series* (May 2011). URL: http://www-07.ibm.com/sg/clientstories/cases/memphis_police_department.html.

Information Section of the Public Affairs Office of the American Embassy in Abidjan, Cote d'Ivoire. *National Daily Press Review*. [Online]. May 2012. URL: http://photos.state.gov/libraries/cotedivoire/231771/Pdfs/NationalDailyPressReview_%20may2012_001.pdf.

International Telecommunication Union (ITU). "Chapter 5: Measuring communication capacity in bits and bytes." In: *Measuring the Information Society*. 2012. URL: `https://www.itu.int/en/ITU-D/Statistics/Documents/publications/mis2012/MIS2012_without_Annex_4.pdf`.

— *Country Statistics, 2000-2011.* [Online]. Jan. 2013. URL: `http://www.itu.int/ITU-D/ict/statistics/`.

IRIN. *Fighting rumours with fact.* [Online]. May 2011. URL: `http://www.irinnews.org/news/2011/05/13/fighting-rumours-fact`.

Italia, Telecom. *Big Data Challenge 2015.* 2015. URL: `http://www.telecomitalia.com/tit/en/bigdatachallenge.html`.

Jacobs, Jane. *The Death and Life of Great American Cities.* New York, NY: Random House, 1961.

Jean, N., M. Burke, M. Xie, M. Davis, D. Lobell, and S. Ermon. "Combining machine learning and satellite imagery to predict poverty." In: *Science* forthcoming (2016).

Jeffery, C. Ray. *Crime Prevention Through Environmental Design.* Sage Publications London, 1977. ISBN: 0-8039-0086-4.

Kennedy v. United Kingdom (2011), 52 EHRR 4, at 118.

Keyfitz, Nathan. "How Do We Know the Facts of Demography?" In: *Population and Development Review* 1.2 (Dec. 1975), pp. 267–288. URL: `http://www.jstor.org/stable/1972224`.

King, Gary. "Ensuring the Data-Rich Future of the Social Sciences." In: *Science* 331.6018 (Feb. 2011), pp. 719–721. URL: `http://gking.harvard.edu/files/gking/files/datarich.pdf`.

— "Big Data is Not About the Data!" In: *Golden Seeds Innovation Summit, New York City*. Jan. 2013. URL: `http://gking.harvard.edu/presentations/big-data-not-about-data`.

Kirkpatrick, Robert. "Data Philanthropy is Good for Business." In: *Forbes Magazine* (Sept. 2011). URL: `http://www.forbes.com/sites/oreillymedia/2011/09/20/data-philanthropy-is-good-for-business/`.

Krumme, Coco, Alejandro Llorente, Manuel Cebrian, Alex 'Sandy' Pentland, and Esteban Moro. "The predictability of consumer visitation patterns." In: *Scientific Reports* 3.1645 (Apr. 2013). ISSN: 2045-2322. DOI: `10.1038/srep01645`.

Kulkarni, Rajendra, Kingsley E. Haynes, Roger R. Stough, and James D. Riggle. "Light based growth indicator (LBGI): exploratory analysis of developing a proxy for local economic growth based on night lights." In: *Regional Science Policy & Practice* 3.2

(2011), pp. 101–113. URL: http://econpapers.repec.org/article/blargscpp/v_3a3_3ay_3a2011_3ai_3a2_3ap_3a101-113.htm.

Lawhorn, Chad. "Census rejects city's appeal of 2010 population totals; new Census numbers for Douglas County show growth slowed in 2012." In: *Lawrence Journal-World* (Mar. 2013). URL: http://www2.ljworld.com/weblogs/town_talk/2013/mar/14/census-rejects-citys-appeal-of-2010-popu/.

Lazer, David, Ryan Kennedy, Gary King, and Alessandre Vespignani. "The Parable of Google Flu: Traps in Big Data Analysis." In: *Science* 343.6176 (Mar. 2014), pp. 1203–1205.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. "Computational Social Science." In: *Science* 323.5915 (Feb. 2009), pp. 721–723. ISSN: 0036-8075. DOI: 10.1126/science.1167742. URL: http://science.sciencemag.org/content/323/5915/721.

Lee, Ronald. "Population in Preindustrial England: An Econometric Analysis." In: *Quarterly Journal of Economics* 87.4 (Nov. 1973), pp. 581–607. URL: http://qje.oxfordjournals.org/content/87/4/581.short.

— "Estimating Series of Vital Rates and Age Structures from Baptisms and Burials: A New Technique, with Applications to Preindustrial England." In: *Population Studies* 28.3 (Nov. 1974), pp. 495–512. URL: http://www.jstor.org/stable/2173642.

Letouzé, Emmanuel. *Big Data for Development: Challenges & Opportunities*. United Nations Global Pulse, May 2012. URL: http://www.unglobalpulse.org/projects/BigDataforDevelopment.

— *Could Big Data provide alternative measures of poverty and welfare?* Online. June 2013. URL: http://www.developmentprogress.org/blog/2013/06/11/could-big-data-provide-alternative-measures-poverty-and-welfare.

— *Big data for development: Facts and figures*. [Online]. Apr. 2014. URL: http://www.scidev.net/global/data/feature/big-data-for-development-facts-and-figures.html.

— "Demography, meet Big Data; Big Data, meet Demography: Reflections on the Data-Rich Future of Population Science." In: *United Nations Expert Group Meeting (UN EGM) on Strengthening the Demographic Evidence Base for the Post-2015 Development Agenda*. United Nations HQ, New York, Oct. 2015. URL: http://www.un.org/en/development/desa/population/events/pdf/expert/23/Presentations/EGM-S5-Letouze%20presentation.pdf.

Letouzé, Emmanuel and Jérémie Cohen-Setton. *Blogs review: GDP, welfare and the rise of data-driven activities*. Blog. Feb. 2014. URL: http://bruegel.org/2014/02/blogs-review-gdp-welfare-and-the-rise-of-data-driven-activities/.

Letouzé, Emmanuel, Patrick Meier, and Patrick Vinck. "Big Data for Conflict Prevention: New Oil and Old Fires." In: *New Technology and the Prevention of Violence and Conflict*. Ed. by Francesco Mancini. International Peace Institute, 2013, pp. 4–27. URL: https://www.ipinst.org/images/pdfs/IPI_Epub-New_Technology-final.pdf.

Letouzé, Emmanuel, Patrick Vinck, and Lanah Kammourieh. "The Law, Politics and Ethics of Cell Phone Data Analytics." In: (Apr. 2015). URL: http://datapopalliance.org/item/white-paper-the-law-politics-and-ethics-of-cell-phone-data-analytics/.

Lévi-Strauss, Claude. *Tristes Tropiques*. Terre humaine. Librairie Plon, 1955. ISBN: 2-266-11982-6.

Livi Bacci, Massimo. "Return To Hispaniola: Reassessing a Demographic Catastrophe." In: *Hispanic American Historical Review* 83.1 (2003), pp. 3–51. ISSN: 1527-1900. URL: https://muse.jhu.edu/article/38903.

Lu, Xin, Erik Wetter, Nita Bharti, Andrew J. Tatem, and Linus Bengtsson. "Approaching the Limit of Predictability in Human Mobility." In: *Scientific Reports* 3.2923 (2013). URL: http://www.nature.com/srep/2013/131011/srep02923/full/srep02923.html.

MacAskill, Ewen and Gabriel Dance. "NSA Files: Decoded. What The Revelations Mean for You." In: *The Guardian* (2013). URL: http://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded.

Mackie, Phil. "Could 'predictive policing' help prevent burglary?" In: *BBC News* (Sept. 2012). URL: http://www.bbc.com/news/uk-19623631.

Main, Frank. "Police sensing crime before it happens." In: *Chicago Sun-Times* (Jan. 2011). URL: http://www.suntimes.com/3295264-417/intelligence-crime-police-weis-department.html.

Malone v. United Kingdom (1985) 7 EHRR 14, at 64.

Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. "Big data: The next frontier for innovation, competition, and productivity." In: *McKinsey Global Institute* (2011). URL: http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation.

Marr, Bernard. "Why Data Minimization Is An Important Concept In The Age of Big Data." In: *Forbes* (2016). URL: http:

//www.forbes.com/sites/bernardmarr/2016/03/16/why-data-minimization-is-an-important-concept-in-the-age-of-big-data/#5d71f32f327f.

MasterCard Center for Inclusive Growth. *Building Literacy for the Data Generation.* [Online]. Dec. 2015. URL: http://mastercardcenter.org/insights/building-literacy-data-generation/.

Mayer-Schönberger, Viktor and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think.* Houghton Mifflin Harcourt, Mar. 2013. ISBN: 978-0-544-00293-7. URL: https://books.google.com/books?id=HpHcGAkFEjkC.

McClellan, Nick, Max Shron, Daniel Duckworth, Marc Georges, Alex Mentch, Thu Nguyen, Mindy Lee, Travis Korte, and Stephanie Kao. "Predicting Small-Scale Poverty Measures from Night Illumination." In: *World Bank Big Data Exploration.* Washington, D.C., Mar. 2013. URL: https://hackpad.com/Predicting-Small-Scale-Poverty-Measures-from-Night-Illumination-f6RoPTY6IWB.

McDonald, Sean. "Ebola: A Big Data Disaster – Privacy, Property, and the Law of Disaster Experimentation." English. In: *The Centre for Internet and Society* (Jan. 2016). URL: http://cis-india.org/papers/ebola-a-big-data-disaster (visited on 01/2016).

Meetup. *Data Science.* URL: http://www.meetup.com/topics/data-science/.

Mehlum, Halvor, Karl Moene, and Ragnar Torvik. "Crime induced poverty traps." In: *Journal of Development Economics* 77 (2005), pp. 325–340. URL: http://www.sciencedirect.com/science/article/pii/S0304387805000106.

Melamed, Claire. "Development data: how accurate are the figures?" In: *The Guardian* (Jan. 2014). ISSN: 0261-3077. URL: http://www.theguardian.com/global-development/poverty-matters/2014/jan/31/data-development-reliable-figures-numbers.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. "Quantitative Analysis of Culture Using Millions of Digitized Books." In: *Sciencexpress* (Dec. 2010).

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. "Economic shocks and civil conflict: An instrumental variables approach." In: *Journal of political Economy* 112.4 (2004), pp. 725–753. URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=562402.

MIT Media Lab (@medialab). *"Big data [is] an ecosystem," says @ManuLetouze of @datapopalliance, a global coalition that includes the @medialab http://mitsha.re/BOuA302mdVf.* [Tweet]. July 2016. URL: https://twitter.com/medialab/status/755071681559949313.

Mitchell, Matthew I. “Migration, citizenship and autochthony: Strategies and challenges for state-building in Côte d’Ivoire.” In: *Journal of Contemporary African Studies* 30.2 (Apr. 2012), pp. 267–287. DOI: `10.1080/02589001.2012.664415`. URL: `http://dx.doi.org/10.1080/02589001.2012.664415`.

Mohler, George O., Martin B. Short, P. Jeffrey Brantingham, Frederic (Rick) Paik Schoenberg, and George E. Tita. “Self-exciting point process modelling of crime.” In: *Journal of American Statistical Association* 106 (2011), pp. 100–108. URL: `http://www.math.ucla.edu/~mbshort/papers/crime3.pdf`.

Monitoring perceptions of crisis-related stress using social media data. United Nations Global Pulse, 2011.

Morozov, Evgeny. *To Save Everything, Click Here: The Folly of Technological Solutionism*. OCLC: 873934585. New York: PublicAffairs, 2014. ISBN: 978-1-61039-370-6.

Narayanan, Arvind and Vitaly Shmatikov. *Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)*. The University of Texas at Austin, 2008. URL: `https://arxiv.org/pdf/cs/0610105.pdf`.

National Climatic Data Center (NCDC). *Climate Data Online (CDO)*. [Online]. 2011. URL: `https://www.ncdc.noaa.gov/cdo-web/`.

NetMob 2013. Cambridge, MA, May 2013.

New York City Office of the Mayor. “Mayor Bloomberg, Police Commissioner Kelly and Microsoft Unveil New, State-of-the-Art Law Enforcement Technology that Aggregates and Analyzes Existing Public Safety Data in Real Time to Provide a Comprehensive View of Potential Threats and Criminal Activity.” In: *News from the Blue Room* (Aug. 2012). [Online]. URL: `http://www.nyc.gov/html/om/html/2012b/pr291-12.html`.

Newman, Oscar. *Defensible space: Crime prevention through urban design*. Macmillan Publishing, Oct. 1973. 264 pp. ISBN: 978-0-02-000750-0.

OECD. *Preventing Violence, War and State Collapse*. Paris: Organisation for Economic Co-operation and Development, Feb. 2009. ISBN: 978-92-64-05980-1.

“Off the map.” In: *The Economist* (Nov. 2014). From the print edition. ISSN: 0013-0613. URL: `http://www.economist.com/node/21632520`.

Olivia, Susan, John K. Gibson, Lars K. Brabyn, and Glen A. Stichbury. “Monitoring economic activity in Indonesia using night light detected from space.” In: *The 12th Indonesian Regional Science Association Conference*. Makassar, Indonesia, 2014.

Orange. *The D4D Challenge is a great success!* URL: `http://www.d4d.orange.com/en/Accueil`.

— *Data for Development (D4D) Challenge Côte d’Ivoire*. [Online]. 2012. URL: `http://www.d4d.orange.com/`.

Oxford Poverty & Human Development Initiative (OPHI). *Alkire Foster method: OPHI's method for multidimensional measurement.* [Online]. n.d. URL: http://www.ophi.org.uk/research/multidimensional-poverty/alkire-foster-method/.

— *Global Multidimensional Poverty Index.* [Online]. n.d. URL: http://www.ophi.org.uk/multidimensional-poverty-index/.

Parrish, R.B. "Circumventing Title III : the Use of Pen Register Surveillance in Law Enforcement." In: *Duke Law Journal* (1977).

Pastor-Escuredo, David, Alfredo Morales-Guzmán, Yolanda Torres-Fernández, Jean-Martin Bauer, Amit Wadhwa, Carlos Castro-Correa, Liudmyla Romanoff, Jong Gun Lee, Alex Rutherford, Vanessa Frias-Martinez, Nuria Oliver, Enrique Frias-Martinez, and Miguel Luengo-Oroz. "Flooding through the lens of mobile phone activity." In: *arXiv.org* (Oct. 2014), pp. 279–286. DOI: 10.1109/GHTC.2014.6970293. arXiv: 1411.6574. URL: http://arxiv.org/abs/1411.6574.

Patterson, E. Britt. "Poverty, Income Inequality, and Community Crime Rates." In: *Criminology* 29.4 (Nov. 1991), pp. 755–776. ISSN: 1745-9125. DOI: 10.1111/j.1745-9125.1991.tb01087.x.

Pentland, Alex. "Reinventing Society in the Wake of Big Data." In: *Edge* (Aug. 2012).

Pentland, Alex 'Sandy'. "Saving Big Data from Itself." In: *Scientific American* 311 (July 2014), pp. 64–67. DOI: 10.1038/scientificamerican0814-64. URL: http://www.nature.com/scientificamerican/journal/v311/n2/full/scientificamerican0814-64.html.

— "Who should we trust to manage our data?" In: (Oct. 2015). URL: https://www.weforum.org/agenda/2015/10/who-should-we-trust-manage-our-data.

Piatesky, Gregory and Anmol Rajpurohit. "The Cardinal Sin of Data Mining and Data Science: Overfitting." In: *KD Nuggets News* 14.15 (June 2014). URL: http://www.kdnuggets.com/2014/06/cardinal-sin-data-mining-data-science.html.

Pikowsky, Robert A. "The Need for Revisions to the Law of Wiretapping and Interception of Email." In: *Michigan Telecommunications and Technology Law Review* (2003). URL: http://www.mttlr.org/volten/pikowsky.pdf.

Popescu, Alex. *BigData: Volume, Velocity, Variability, Variety.* Online. June 2011. URL: http://nosql.mypopescu.com/post/6361838342/bigdata-volume-velocity-variability-variety.

Porter, Theodore M. *UCLA Department of History: Faculty.* [Online]. n.d. URL: http://www.history.ucla.edu/faculty/theodore-porter.

Powers, David Martin Ward. *Evaluation: from precision, recall and F-factor to ROC, informedness, markedness and correlation.* SIE-07-001. Flinders University, Adelaide, Australia, Dec. 2007. URL: `https://csem.flinders.edu.au/research/techreps/SIE07001.pdf`.

Provost, Foster and Tom Fawcett. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking.* 1st edition. Sebastopol, CA: O'Reilly Media, Aug. 2013. 414 pp. ISBN: 978-1-4493-6132-7.

République du Sénégal. *Loi no96-06 du 22 mars 1996 portant Code des Collectivités locales.* [Online]. 1996.

— *Loi no2013-10 du 28 décembre 2013 portant Code général des Collectivités locales.* [Online]. 2013.

Roca, Thomas and Emmanuel Letouzé. "Open algorithms: A new paradigm for using private data for social good." In: *Devex* (July 2016). URL: `https://www.devex.com/news/open-algorithms-a-new-paradigm-for-using-private-data-for-social-good-88434`.

"Rochester Police to fignt crime with big data mining technology." In: *weSRCH* (Jan. 2012). [Online]. URL: `http://www.wesrch.com/electronics/prEL11TZNW3LCEG`.

Rodriguez Takeuchi, Laura and Emma Samman. *Patterns of progress on the MDGs and implications for target setting post-2015.* Overseas Development Institute (ODI), Mar. 2015. URL: `https://www.odi.org/publications/9353-patterns-progress-mdgs-implications-target-setting-post-2015`.

Sarsons, Heather. "Rainfall and Conflict." In: *North East Universities Development Conference (NEUDC) 2011.* Oct. 2011. URL: `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.230.9257`.

See, Victor, Robert F. Cameron, and Stephen R. Schwartz. "Non-adiabatic electron behaviour due to short-scale electric field structures at collisionless shock waves." In: *arXiv.org* (2013). DOI: `arXiv:1304.4841v1[astro-ph.SR]`. URL: `http://arxiv.org/abs/1304.4841`.

Shannon, Claude Elwood. "A Mathematical Theory of Communication." In: *Bell System Technical Journal* 27.3 (July 1948), pp. 379–423. ISSN: 1538-7305. DOI: `10.1002/j.1538-7305.1948.tb01338.x`.

"Sharing Data As Corporate Philanthropy." In: *Markets for Good* (Sept. 2014). URL: `http://www.marketsforgood.org/sharing-data-as-corporate-philanthropy/`.

Shiffrin, Richard M. "Drawing causal inference from Big Data." In: *PNAS, Proceedings of the National Academy of Sciences.* Vol. 113. 27. 2016. DOI: `10.1073/pnas.1608845113`. URL: `http://www.pnas.org/content/113/27/7308.extract`.

Singh, Sanasam Ranbir, Hema A. Murthy, and Timothy A. Gonsalves. "Feature Selection for Text Classification Based on Gini Coefficient of Inequality." In: *JMLR Workshop and Conference Proceedings: Feature Selection in Data Mining.* Fourth International Workshop on Feature Selection in Data Mining. Red. by Huan Liu, Hiroshi Motoda, Rudy Setiono, and Zheng Zhao. Vol. 10. Hyderabad, India, June 2010, pp. 76–85. URL: http://www.jmlr.org/proceedings/papers/v10/sanasam10a/sanasam10a.pdf.

"Smart Steps." In: *Telefonica Dynamic Insights* (n.d.). [Online]. URL: http://dynamicinsights.telefonica.com/smart-steps/.

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

Smith-Clarke, Christopher, Afra Mashhadi, and Licia Capra. "Poverty on the Cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* CHI '14. New York, NY, USA: ACM, 2014, pp. 511–520. ISBN: 978-1-4503-2473-1. DOI: 10.1145/2556288.2557358. URL: http://doi.acm.org/10.1145/2556288.2557358.

Song, Chaoming, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. "Limits of predictability in human mobility." In: *Science* 327.5968 (Feb. 2010), pp. 1018–1021. DOI: 10.1126/science.1177170.

Soto, Victor, Vanessa Frias-Martinez, Jesus Virseda, and Enrique Frias-Martinez. "Prediction of Socioeconomic Levels Using Cell Phone Records." In: *User Modeling, Adaption and Personalization.* Ed. by Joseph A. Konstan, Ricardo Conejo, José L. Marzo, and Nuria Oliver. Lecture Notes in Computer Science 6787. Springer Berlin Heidelberg, July 2011, pp. 377–388. ISBN: 978-3-642-22361-7. DOI: 10.1007/978-3-642-22362-4_35. URL: http://link.springer.com/chapter/10.1007/978-3-642-22362-4_35.

Soubra, Diya. *The 3 Vs that define Big Data.* [Online]. 2012. URL: http://www.datasciencecentral.com/forum/topics/the-3vs-that-define-big-data.

Spielberg, Steven. *Minority Report.* [Film]. 2002.

Srinivasamurthy, Supreeth. "Who said "90% of data ever created was created in the last 2 years"?" In: *Quora* (Oct. 2013). [Online]. URL: https://www.quora.com/Who-said-90-of-data-ever-created-was-created-in-the-last-2-years.

Stanford Encyclopedia of Philosophy. [Online]. URL: http://plato.stanford.edu/.

Stevens, Forrest R., Andrea E. Gaughan, Catherine Linard, and Andrew J. Tatem. "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data." In: *PLOS ONE* (2015). URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0107042.

Stiglitz, Joseph E., Amartya Sen, and Jean-Paul Fitoussi. *Report by the Commission on the Measurement of Economic Performance and Social Progress.* Stiglitz-Sen-Fitoussi

Commission, Sept. 2009. URL: http://www.stiglitz-sen-fitoussi.fr/documents/rapport_anglais.pdf.

Streiner, David L. “Breaking up is hard to do: the heartbreak of dichotomizing continuous data.” In: *Canadian Journal of Psychiatry*. Research Methods in Psychiatry 47.3 (Apr. 2002), pp. 262–266. ISSN: 0706-7437. URL: https://ww1.cpa-apc.org/publications/archives/cjp/2002/april/streiner.PDF.

Sustainable Development Goals Blog. *UN Adviser underlines importance of partnership with mobile-communications industry to achieve Sustainable Development Goals*. [Online]. Feb. 2016. URL: http://www.un.org/sustainabledevelopment/blog/2016/02/un-adviser-calls-for-new-mobile-communications-industry-partnership-to-achieve-sustainable-development-goals/.

Tanz, Jason. *Soon We Won’t Program Computers. We’ll Train Them Like Dogs*. Online. May 2016. URL: http://www.wired.com/2016/05/the-end-of-code/.

Tarling, Roger and Katie Morris. “Reporting crime to the police.” In: *British Journal of Criminology* 50 (Mar. 2010), pp. 474–479. DOI: 10.1093/bjc/azq011. URL: http://bjc.oxfordjournals.org/content/50/3/474.short.

“The data deluge.” In: *The Economist* (Feb. 2010). From the print edition. ISSN: 0013-0613. URL: http://www.economist.com/node/15579717.

The Kavli Foundation and New York University’s Institute for the Interdisciplinary Study of Decision Making. *Kavli HUMAN Project*. [Online]. URL: http://kavlihumanproject.org/about/.

“The war in Darfur: Nate Barton.” In: *AP World Class Weebly* (2012). URL: http://apworldwiki2011-12.weebly.com/case-study-darfur.html.

Toole, Jameson L., Nathan Eagle, and Joshua B. Plotkin. “Spatiotemporal Correlations in Criminal Offense Records.” In: *ACM Transactions on Intelligent Systems and Technology* 2.4.38 (July 2011), pp. 1–18. ISSN: 2157-6904. DOI: 10.1145/1989734.1989742.

Traunmueller, Martin, Giovanni Quattrone, and Licia Capra. “Mining mobile phone data to investigate urban crime theories at scale.” In: *Lecture Notes in Computer Science*. 6th International Conference on Social Informatics. Vol. 8851. 2014, pp. 396–411. DOI: 10.1007/978-3-319-13734-6_29. URL: http://link.springer.com/chapter/10.1007%2F978-3-319-13734-6_29.

Turvey, Brent E. *Criminal profiling : an introduction to behavioral evidence analysis*. San Diego, CA: Academic Press, 1999. 462 pp. ISBN: 978-0-12-705040-9.

Tuv, Eugene, Alexander Borisov, George Runger, and Kari Torkkola. “Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination.” In: *Journal of Machine Learning Research* 10 (July 2009), pp. 1341–1366. ISSN: 1533-7928. URL: http://www.jmlr.org/papers/volume10/tuv09a/tuv09a.pdf.

United Nations Development Programme (UNDP). “Rwanda: Human Development Indicators.” In: *Human Development Reports* (2015). URL: `http://hdr.undp.org/en/countries/profiles/RWA`.

United Nations Statistics Division. *Fundamental Principles of Official StatisticsFundamental Principles of Official Statistics (A/RES/68/261 from 29 January 2014)*. Jan. 2014. URL: `http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx`.

— “Census dates for all countries.” In: *2020 World Population and Housing Census Programme* (May 2016). URL: `http://unstats.un.org/unsd/demographic/sources/census/censusdates.htm`.

United States Congress. *Electronic Communications Privacy Act*. 18 U.S.C. § 2510 et seq. 1986.

Wang, R. ‘Ray’. *Monday’s Musings: Beyond The Three V’s of Big Data – Viscosity and Virality*. Online. Feb. 2012. URL: `http://blog.softwareinsider.org/2012/02/27/mondays-musings-beyond-the-three-vs-of-big-data-viscosity-and-virality/`.

Wang, Tong, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. “Learning to Detect Patterns of Crime.” In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný. Lecture Notes in Computer Science 8190. Springer Berlin Heidelberg, Sept. 2013, pp. 515–530. ISBN: 978-3-642-40993-6. DOI: `10.1007/978-3-642-40994-3\_33`. URL: `http://www.site.uottawa.ca/~nat/Courses/csi5387_Winter2014/paper21.pdf`.

Weber v. Germany (2008) 46 EHRR SE5, at 77.

Weigend, Andreas. “Big Data, Social Data, and Marketing.” In: *World Marketing Forum*. Mexico City, June 2013. URL: `http://weigend.com/files/speaking/Weigend_WorldMarketingForum_MEX_2013.06.27.pdf`.

Weisburd, David. “Place-Based Policing.” In: *Ideas in American Policing*. Vol. 9. Police Foundation, Jan. 2008, pp. 1–16. URL: `https://www.policefoundation.org/publication/place-based-policing/`.

Weisstein, Eric W. *Voronoi Diagram*. [Online]. July 2016. URL: `http://mathworld.wolfram.com/VoronoiDiagram.html`.

Wesolowski, Amy, Gillian Stresman, Nathan Eagle, Jennifer Stevenson, Chrispin Owaga, Elizabeth Marube, Teun Bousema, Christopher Drakeley, Jonathan Cox, and Caroline O. Buckee. “Quantifying travel behavior for infectious disease research: a comparison of data from surveys and mobile phones.” In: *Scientific Reports* 4 (July 14, 2014). ISSN: 2045-2322. DOI: `10.1038/srep05678`. URL: `http://www.nature.com/articles/srep05678`.

Whitman, James Q. "Two Western Cultures of Privacy: Dignity Versus Liberty." In: *Yale Law Journal* 113.6 (Apr. 2004), p. 1160. URL: http://www.yalelawjournal.org/article/the-two-western-cultures-of-privacy-dignity-versus-liberty.

World Economic Forum (WEF). *Big Data, Big Impact: New Possibilities for International Development.* Cologny/Geneva, Switzerland, 2012. URL: https://www.weforum.org/reports/big-data-big-impact-new-possibilities-international-development/.

Wrigley, Edward Anthony and Roger S. Schofield. *The Population History of England, 1541-1871: A Reconstruction.* Harvard University Press, 1981. ISBN: 978-0-674-69007-3. URL: http://www.cambridge.org/au/academic/subjects/history/british-history-after-1450/population-history-england-15411871.

Zagheni, Emilio and Ingmar Weber. "You Are Where You e-Mail: Using e-Mail Data to Estimate International Migration Rates." In: *Proceedings of the 4th Annual ACM Web Science Conference.* WebSci '12. New York, NY, USA: ACM, 2012, pp. 348–351. ISBN: 978-1-4503-1228-8. DOI: 10.1145/2380718.2380764. URL: http://www.demogr.mpg.de/publications%5Cfiles%5C4598_1340471188_1_Zagheni&Weber_Websci12.pdf.